MM ENG

A.Y. 2024/2025

B.LAB

HANDOUTS

STRATEGIC MARKETING AND ANALYTICS (MODULE 1)

-GENERAL-

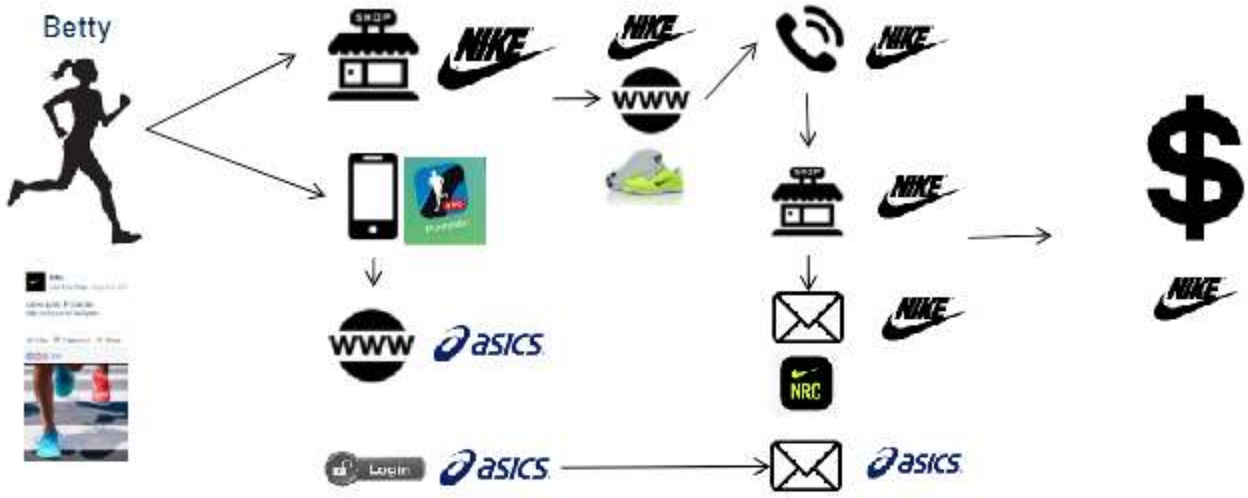
WRITTEN BY
CHIARA TUA

EDITED BY
VITTORIA NASONTE

TEACHING DIVISION

STRATEGIC MARKETING & ANALYTICS
DATA & ANALYTICS FOR STRATEGIC MARKETING DECISIONS

Setting the scene



Betty is a runner, and she might be exposed on some target adv, through a post as, for example, buy a new running pair of shoes.
The post triggers her attention because she is considering buying a new pair of running shoes, so she goes walk to the city centre, enters the Nike store and try's them on, but she is not sure if they are fit or not, so she postponed thinking about it.
In the meanwhile, she is using a free run activity tracker in which she is exposed to an Asis adv, which proposes to register to the newsletter in order to have a discount on the first purchase.
After a while she takes her decision opting for a Nike pair of shoes but not at the Nike store but online cause she can customize them by choosing the color and so on.
She receives the shoes, but the color is not what she ordered, so she phones the customer service saying that there is an issue with her order.
The customer service fixes the problem and asks Betty if her want to receive the new pair at home or if she prefers to collect it at the store. Betty decides to pick them to the store.
In the meanwhile, Asis keep emailing her with possible new offers, but at the end the transaction is for Nike.

What this example tells us?
This is the description of the decision process -> when we buy something new, we pass through different phases:
1) search for alternatives
2) purchase
3) evaluation of your purchase
-> this is called **customer journey** which is the decision process from the very beginning until the end.

Company can track each phase of the customer journey, even the searching initial phase. Customers can be tracked at the same time from different company. Acquire the ability to *analyze data* to *answer relevant* *marketing questions* and *support decision* is a fundamental part of marketing manager work. Data and analytics are used in marketing to try to create efficient and effective marketing strategies.

Which marketing question when analysing data?

Distribution

How do channels impact a brand's choice and customer decision-making?

Brand Choice

How can we increase the likelihood that brand x is chosen?

Consumer Analysis and Targeting

Who are our customers?
Who is most likely to purchase?
Which product variant are they most likely to buy?

Value of the Customer Base

Are our customers profitable? Are our customers loyal?

How can we increase the profitability of the customer base?

Marketing Actions

Can discounts effectively lead new customers to make their initial purchase?

SEPHORA

Objective: Encourage purchases from consumers who have not used the loyalty card in the last six months.

How are these evaluations made?

Which Strategy is more Effective?

Why do we need analytics in marketing?

Statistic helps managers in **dealing with uncertainty** -> managers can make smart decisions and lead staff more effectively.
Data and analytics reduce uncertainty -> if I'm a manager I don't want to risk, so if I can provide empirical evidence, it will be so much better; moreover, you need to test your idea -> in this way is highly recommend use data.

Prior: Managers' intuition Decision

Decisions based on managers intuition.
The management provides preliminary estimates (prior) about the probability that the event will occur following a strategy.

Prior = The management expects a 10% increase in
purchases by customers with loyalty cards through the email and smartphone free shipping campaign.
Pay attention to:

- Avaiability bias
- Overconfidence

In statistics (Bayesian theory) Prior.
Refers to the initial estimate or probability one has regarding an event
before considering new information or data

Posterior: Data

Decisions Based on Data Analysis.
We can gather information and use it to refine the preliminary (prior) estimates regarding the probability that an event will occur.
Collect Data > Posterior

Posterior= After analyzing data, we estimate that 5% of loyalty card customers will respond positively to the free shipping campaign promoted via email and smartphone notifications.

In statistics (Bayesian Theory) Posterior:
The updated probability of an event after considering new data or information.

Value of analytics in marketing

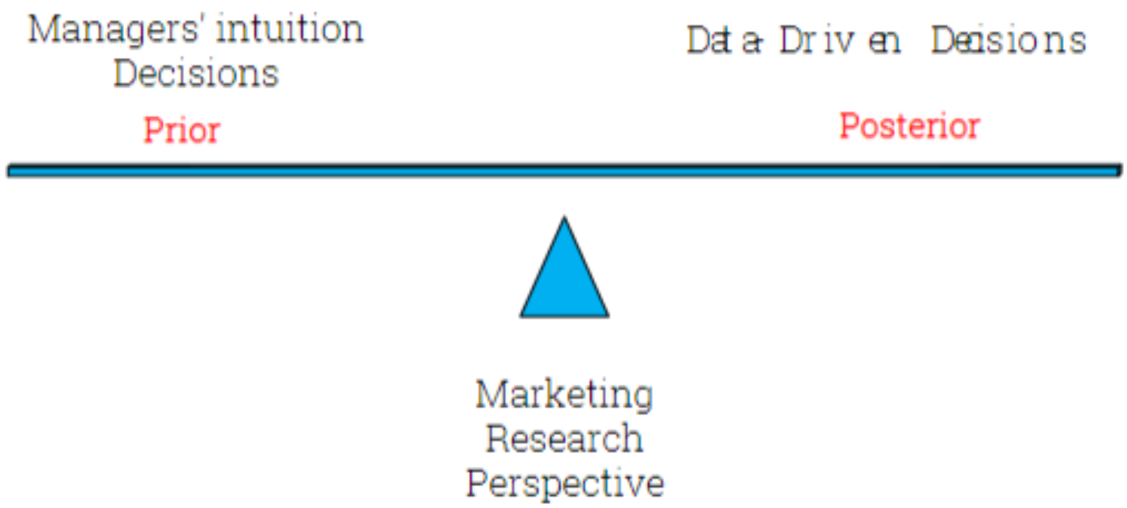
- You can collect information that is used to **refine** the prior estimate of the probability that an event occurs.

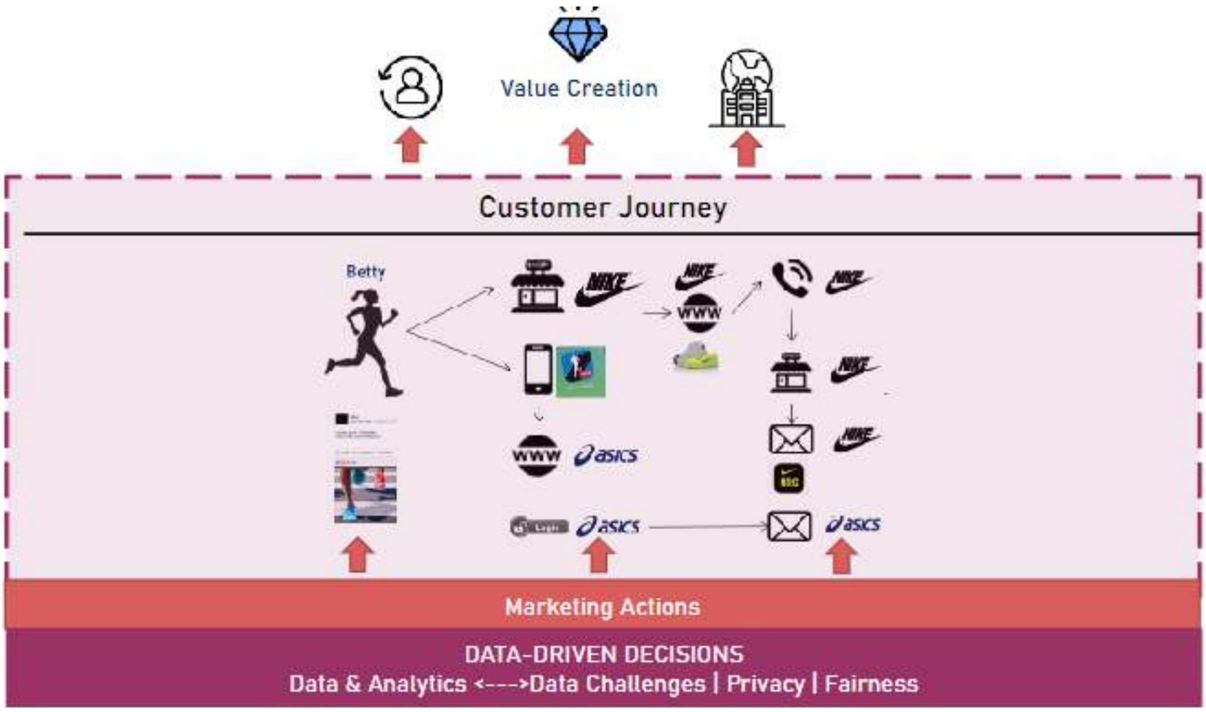
  Prior -> Collect Data> Posterior
  [Prior you get for "free"; and maybe it's enough]
- Statistics should reduce uncertainty associated with predicting future events.

  Prior(variance) > Posterior(Variance)
- Reliability of information is not ignored
- Statistic should help you to quantify the consequences of a planned business action.

Choose the action that maximizes some decision rule

Consider Both Approaches





<u>The final aim is to create value for both the company and the customers.</u>

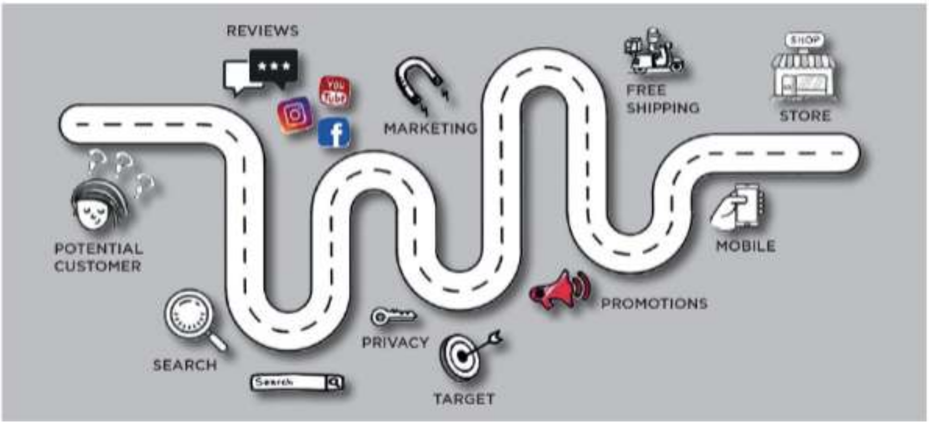
Big data & customer journey

Data in Marketing: What data?

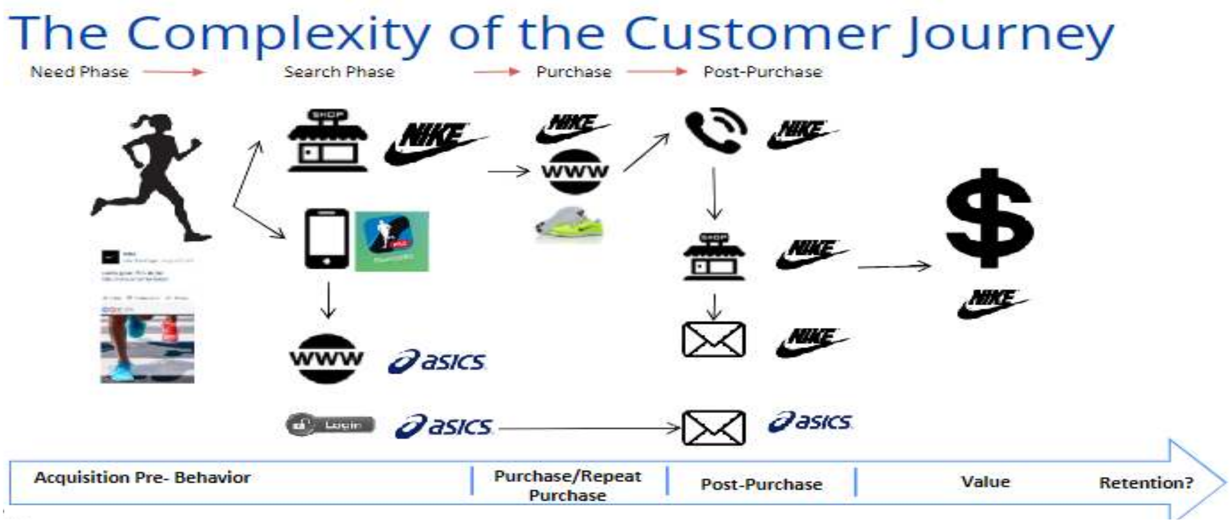
Making decisions about the types of data is a "measurement problem".
We can have 2 distinct macro classes of data:

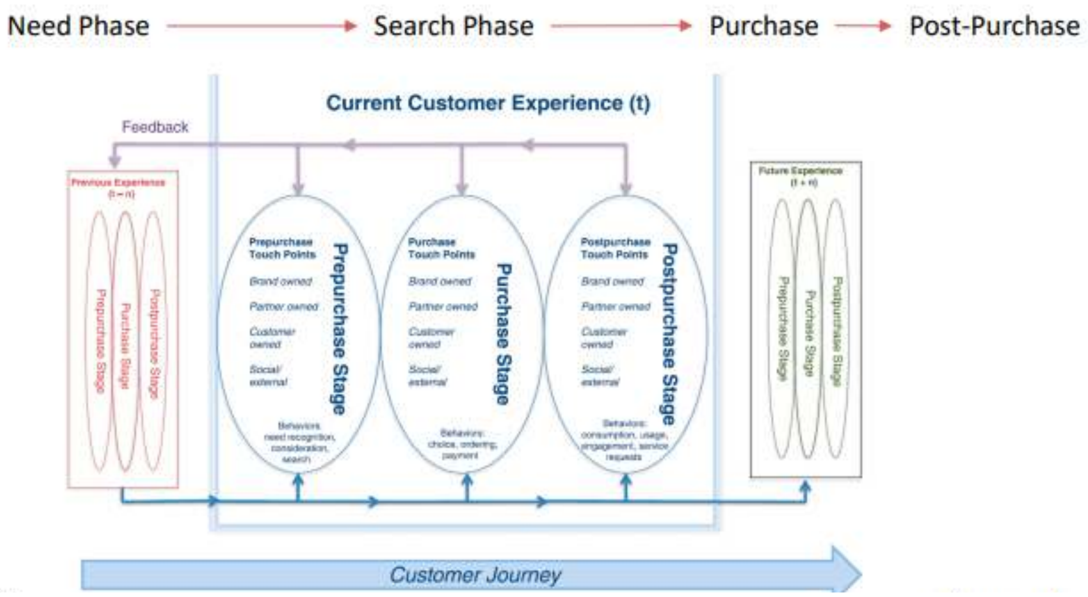
- **Primary data**: can be collected in 2 different ways and for a specific purpose (and this is the main difference with secondary data):
    - *Qualitative research* (focus group, observation, in-depth interviews)
    - *Quantitative research* (market surveys, lab experiment)
- **Secondary data**: refers to data that has been previously collected, so they are already available, and for some other purpose. These data are not collected specifically for a study at hand but can be utilized to gain insights.
  Secondary data are the data that can be derived from both internal and external sources such as social media pages, loyalty cards, website, corporate CMR, e-commerce, and physical stores. -> we don't have to pay for them, and it is here data marketing analytics was born
  Field experiment: experimental part combined with secondary data

The Customer journey



Remember the example of Betty





The relationship between customer journey and data analysis is really strong, because all the interaction between a firm and its clients can be tracked becoming data to analyzed.
More formally the kind of behavior that we can track is a **pre-acquisition behavior** -> we can have search data before purchasing, we can have data about the purchase occasion and then we can have data about post-purchase occasion, so what happened after.

We have a pre-purchase stage -> we can have data collected thanks to different touch points such as social media, the store and so on, and also it is interesting to understand if those touch points are owned by the brand of managed by a partner.

Example: Procter & Gamble
We decide to do a partnership with Amazon for a specific product, for example Gillette.
In this case, Amazon is a partner so the channel is the e-commerce (touch point) but it is not owned but the brand.

Thinking about the customer journey, which data can be considered as the most relevant?

- Time spent on the website
- The arrival channel (social media, corporate page, an influencer page...)
- Customers' opinion post-purchase
- Response to promotions
- If customers purchase or not
- What customer purchased
- Through which channel customers interacted with (social media, store, online website, call centre...)

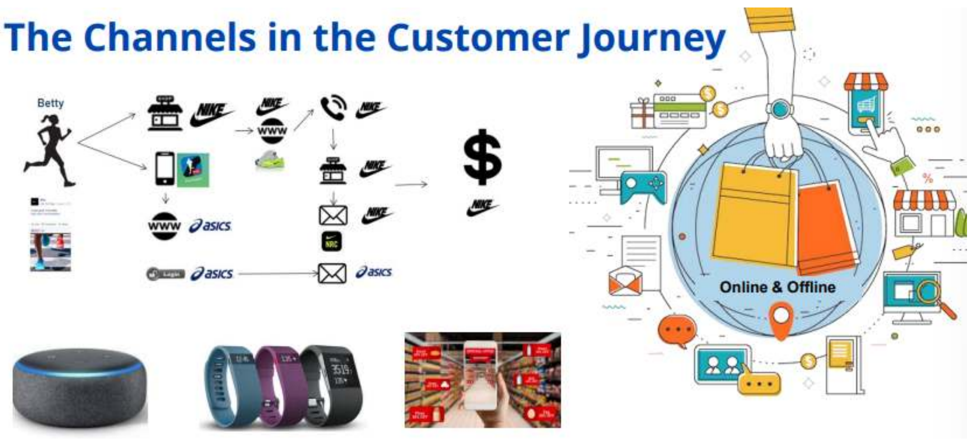
Key element of the Customer Journey

1. Digital **channels** & touchpoints and integration with physical channels (Brand-owned, Partner-owned or Customer-owned touchpoints)
   Omnichannel perspective
   - Social media

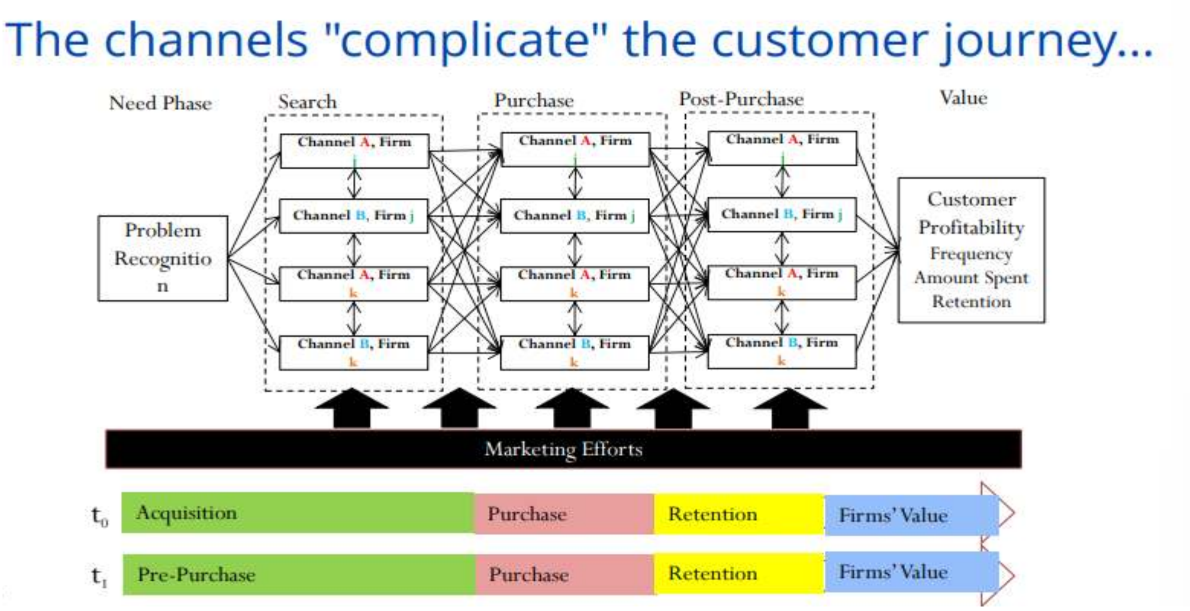
    - Physical store
    - Online channel
- These are just example of the possible touch points that we can track and analyzed give the firm relevant information

2. **Data**
3. **Analytics** & Better targeting

CHANNELS & TOUCHPOINTS



Track the presence within the store is very difficult -> for every single purchase occasion we can track HOW the customer interacts with the different touch points and channels not only for the first purchase occasion, but also for the cumulative purchase occasion or search sessions that the customer comulates interacting with us.



Example: Walmart retail channel strategy?

    - Store
    - Online store
    - Website
    - Pick-up service

- Front door delivery
- Smart door lockers
- Walmart + -> a subscription service similar to Amazon Prime that gives members express delivery, discounted petrol and other perks

Customers have more choices than ever before when it comes to how they get their groceries. They can shop with us in stores, order online for free pickup, or have groceries delivered to their front doors. Customers can also order fresh groceries and everyday essentials and have them delivered directly into the kitchen or garage fridge

Walmart partnered with TikTok to pioneer innovation for the fastest growing digital community. They brought a shoppable live stream experience to U.S.
TikTok users for the very first time. The TikTok community enjoyed shopping while engaging with their favorite creators. During the event, we netted 7X more views than anticipated and grew our TikTok followers by 25%.

March 9, 2021 William White, Chief Marketing Officer, Walmart U.S.

The company gave clues to its reasoning, saying TikTok's integration of ecommerce and advertising "was a clear benefit to creators and users" and would "provide Walmart with an important way for us to reach and serve omnichannel customers as well as grow our third-party marketplace and advertising businesses".

=> This is an example of channel segmentation
Example: Sephora channel strategy

- Catalog
- Community
- Store
- Smartphone

	NEED	SEARCH	PURCHASE	POST
STORE	I need something and I go to the store	I search my product, and, in the meanwhile, I find something else	It is easier finalize the purchase in store	They give you samples which incentive you to buy again
SOCIAL MEDIA	I see a post and I find new desire	I see something through the social and I want to buy it	Possibility to da an Adv which facilitate purchase actions	Newsletter, surveys, analyse the insights, broadcast channels
MAKE-UP ROOM	You need that service, or you	You see the products and	It facilitates the purchase after trying the	They usually ask you feedback

	want to learn something	want to learn more about them	products, moreover you paid for the experience itself	about the experience
WEBSITE	I need something but I don't have time to go to the store	While I'm searching a product, I'll also have correlated products	Possibility to buy online	Loyalty cards

How did all start? Omnichannel… Past 15 years

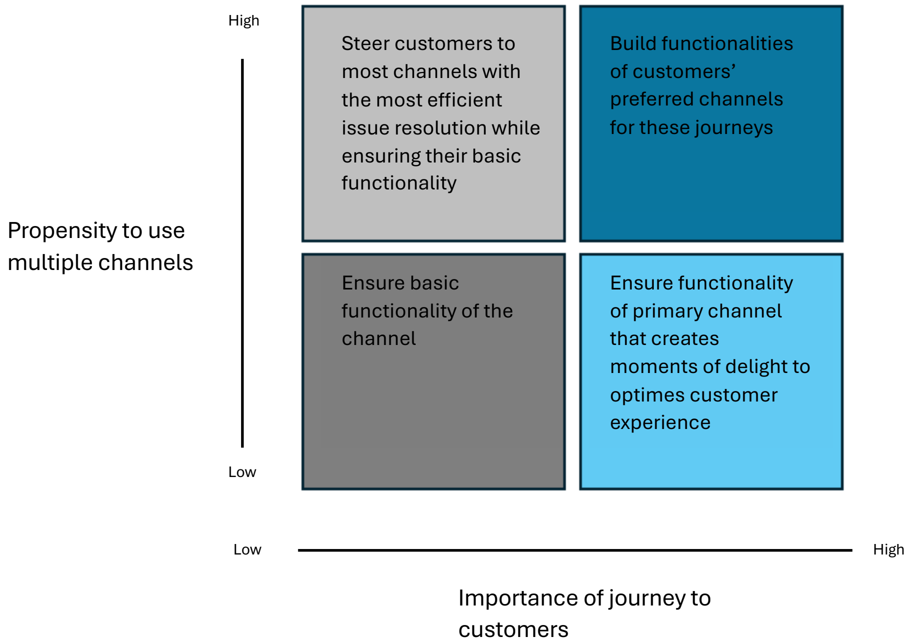
Why omnichannel?

As a company we can have multiple advantages:

1. Some channels can be more costly for some customers
2. It is possible to shape and orient the customer journey of the customers towards channels that provide more margin for the brand
3. To expand customer base (Sephora is active on social media even through the purchase button because probably they realise that they want to reach a specific target, the younger generation, that usually buy through this channel -> to reach them the company must adapt to their preferences)
   It can be a tool to increase customer satisfaction, because I can reach all the customer in different way, moment and occasion

4. It can be a tool to increase customer satisfaction, because I can reach all the customer in different way, moment and occasion

An organization can tailor its omnichannel approach by mapping each customer journey to a quadrant of the matrix and focusing on only two or three in the top-right corner.



Omnichannel Opportunities	Omnichannel Threats
Right channeling: high margin channels can produce a significant reduction in costs	**Cannibalization**: profits cannibalization across channels: increase the number, same profits, more channel
Customer satisfaction: provide customer a better service/experience	**Brand value erosion**: the brand value could be eroded if channels are not well managed
Expansion customer base: the % of online and mobile shopping has increased and continues to increase	**Channel coordination**: the company should be able to effectively manage different channels -> decrease in consumer satisfaction, customer retention…
Increase profits	

N.B. Use multiple channels can also be dangerous depending on the kind of products that the company is selling.

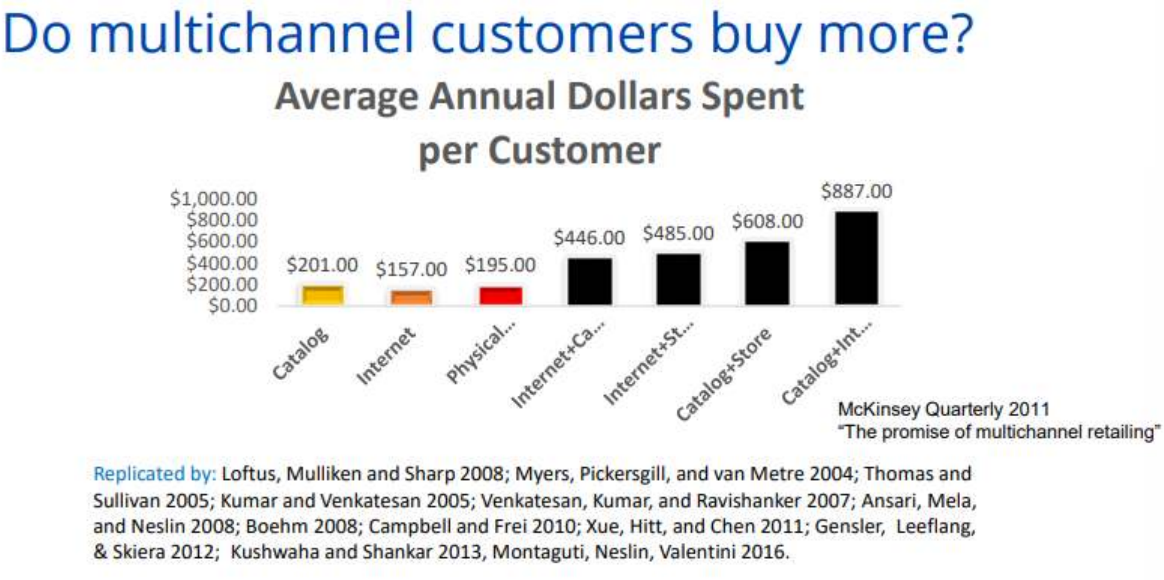
A growing interest in the omnichannel strategy…

70% of Saks customers purchasing online also purchase in stores

The multi-channel customers spend from 3-4 times more than the single channel

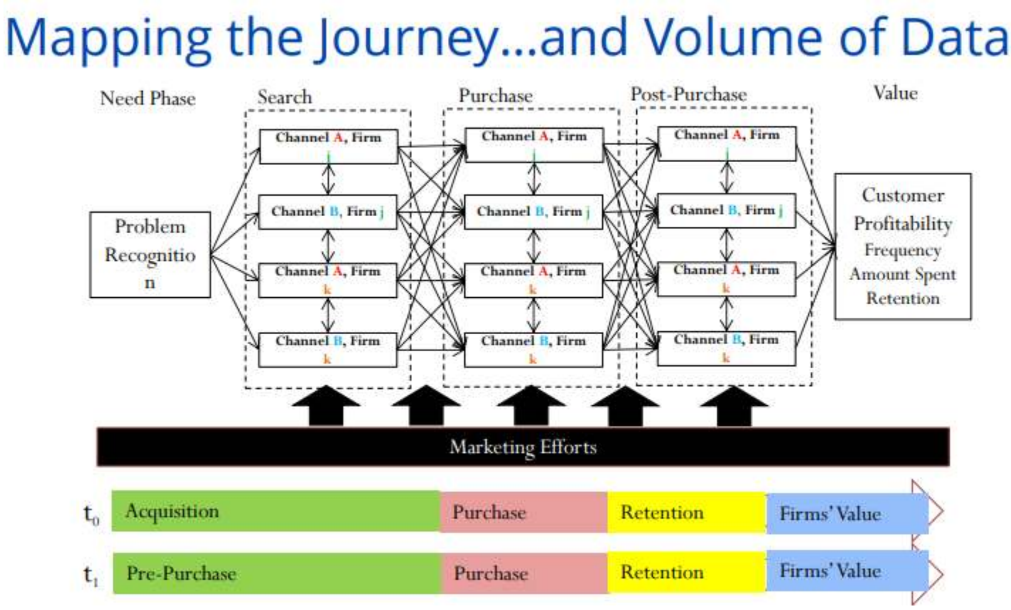
Several businesses decide to implement different channels because they find evidence that customers tend to be more profitable for the brands -> of course this is not a rule valid for every firm and industry, so it is important to verify this behavior before investing in this marketing activity.

Is there a relationship between channel choice and customer profitability?

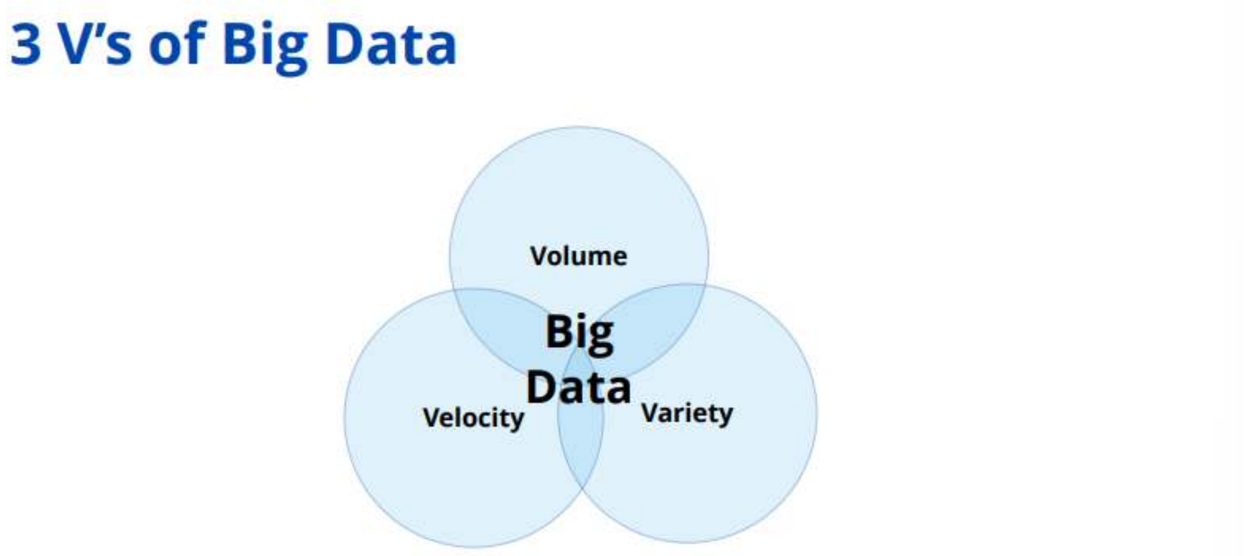


DATA

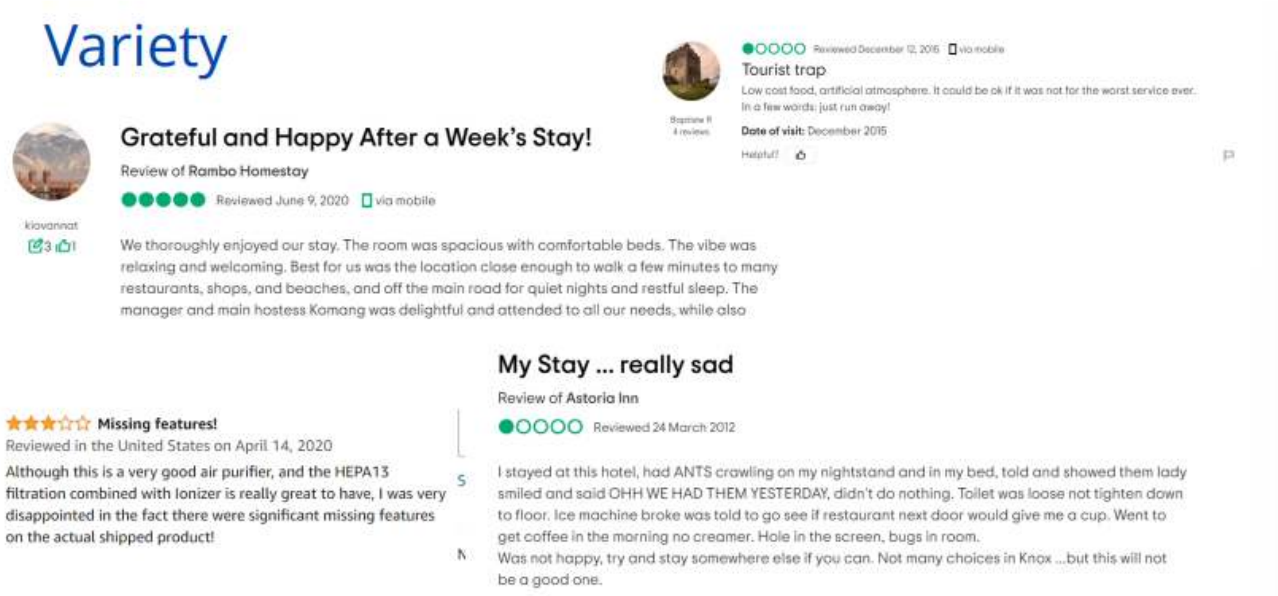
Data -> available data are big, that is why they are so interesting



Nowadays we have the possibility to enter in contact with several secondary data -> the volume of disposable data in huge, but remember that we don't necessarily need a lot of data, but what is more important is have good data available now!

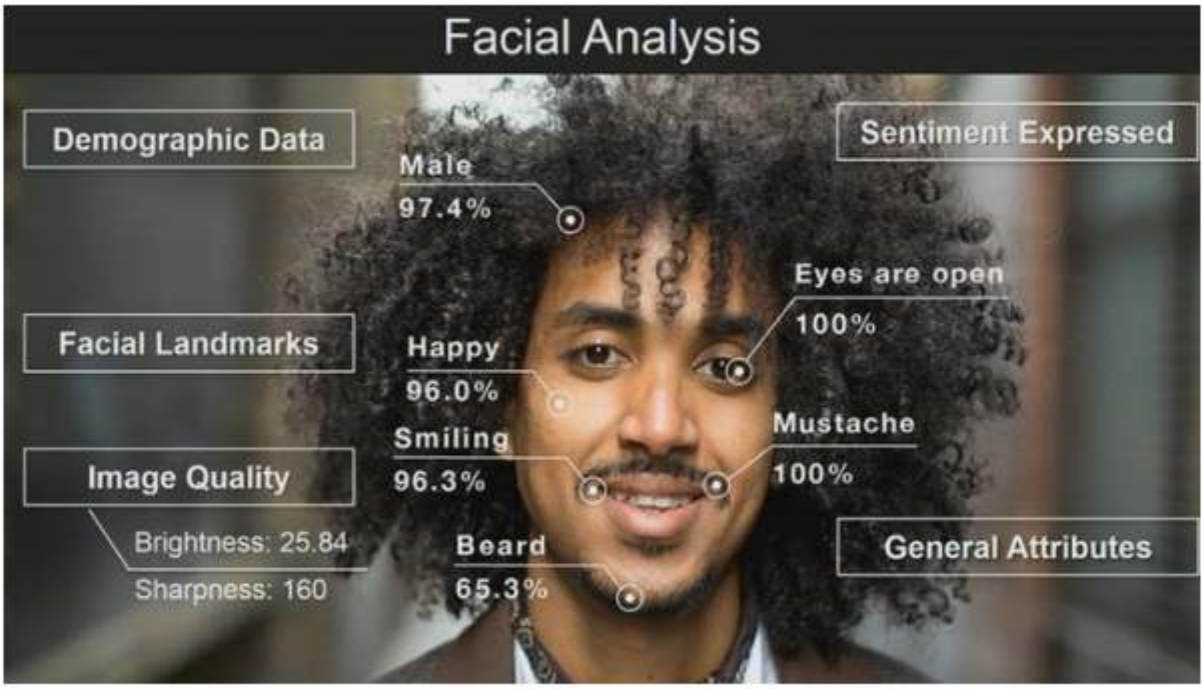


- **Volume**: data are big, we have a lot of information (less interesting: in principles we don't need a lot of data, but good data)
- **Velocity**: I can have the data now -> we don't need anymore to do surveys even though they are still useful -> we can now opt for an instant marketing strategy
- **Variety**: also, text are data (through Python I can transform a text into data) that can be analysed; even photos can be data, about our preferences, emotions, age, gender, vocal audios



Variety

Example: Booking.com, Tripadvisor can classity UGC with labels associated to the presence of specific elements (e.g. swimming pool, mountain

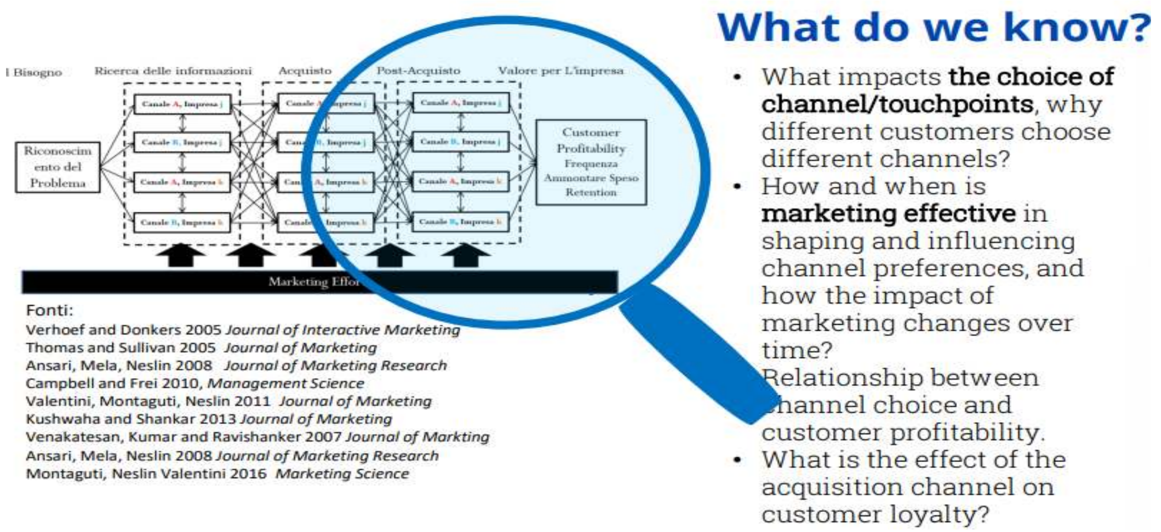


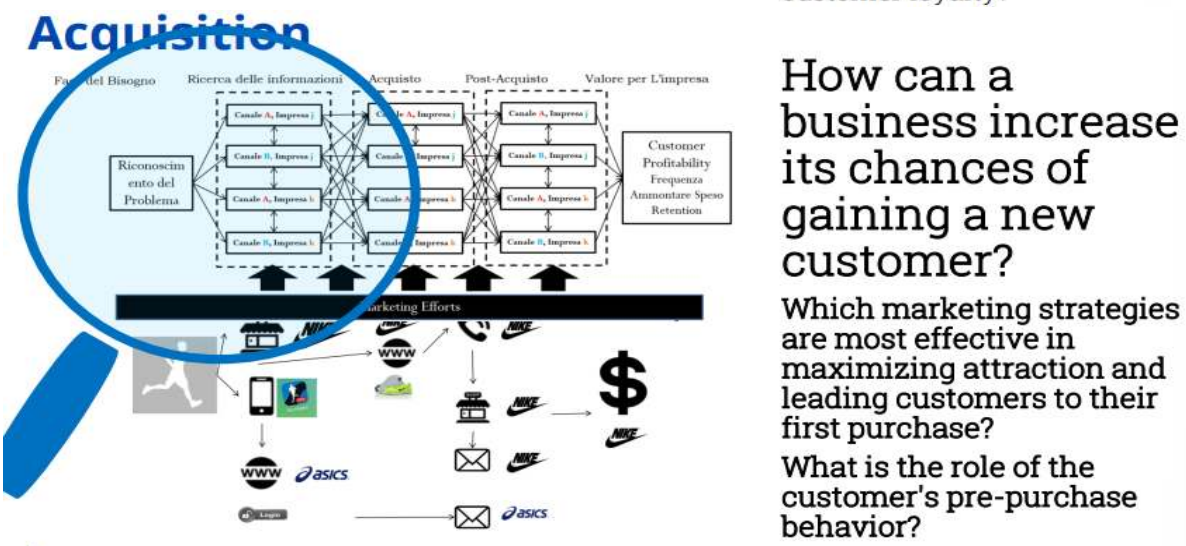
More info: https://www.youtube.com/watch?v=fk-TxySUAzw

[What Data are Voice Assistants Collecting]

ANALYTICS & DATA-DRIVEN DECISIONS

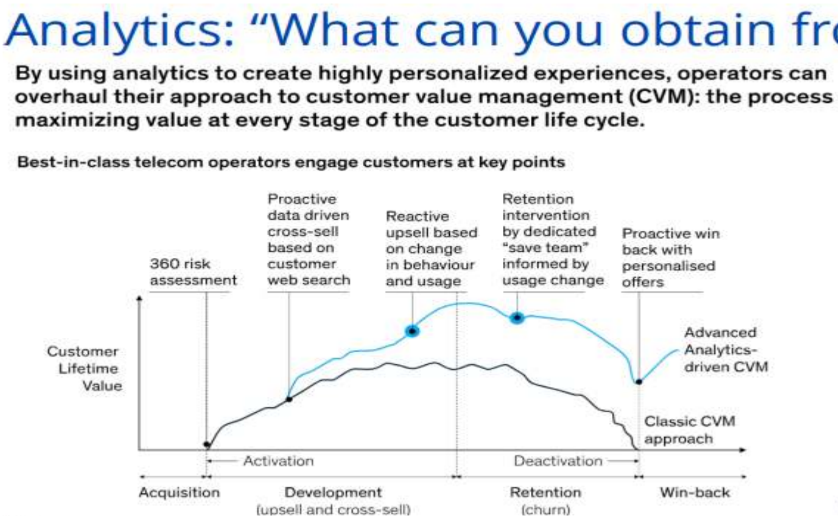
Big data and the information derived from them can be used anytime.
Big data and analytics work basing on
algorithms -> as a whole these practices work
How does it work?





The challenge... **Map and analyze the entire customer journey**.

- How can customers be segmented based on their journey?
- What are the predominant patterns?
- Which types of marketing activities work best?
- What leads to higher profitability?

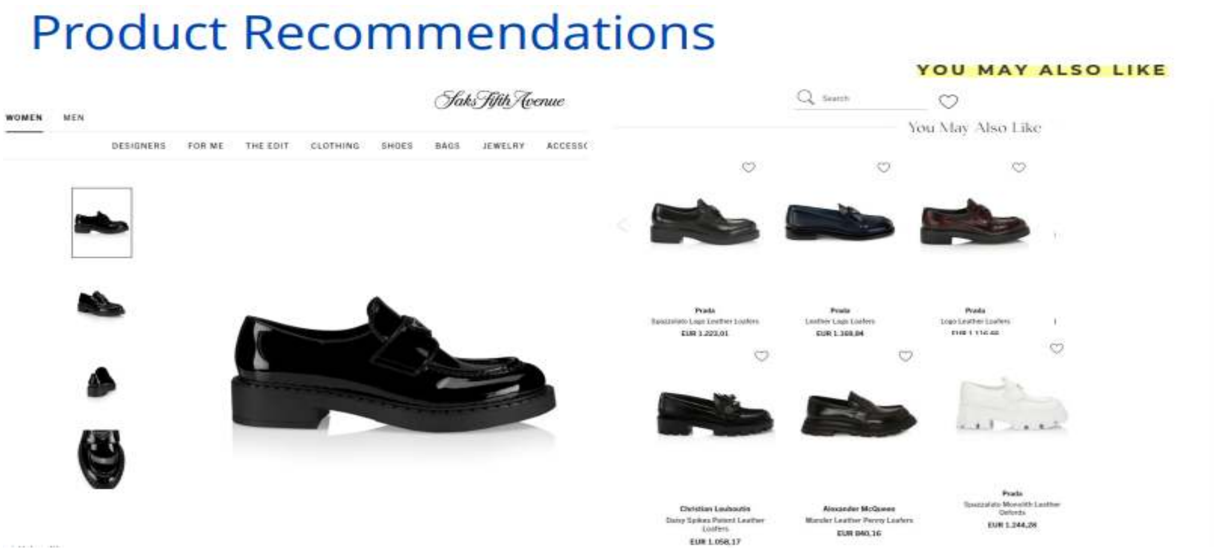


Souce: McKinsey Quarterly

Why Data Science and Analytics?

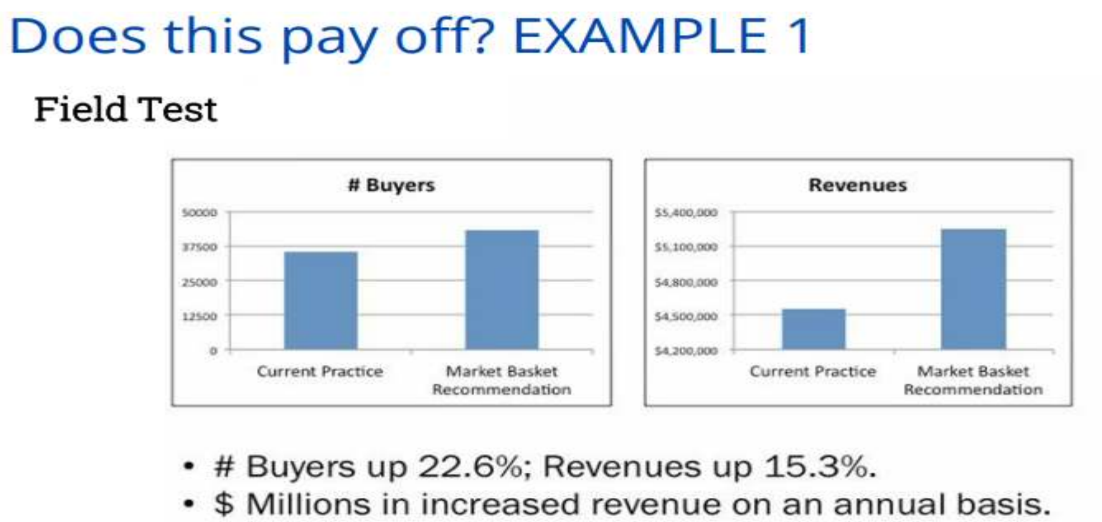
Data Analytics can be seen as Bottom Line, but which is the impact on the customer's value?

Overwiev: Why do I receive these?



How is this done? Market Basket Analysis

- Data: 1,000,000s page visit records
- Compute probability Pr(view product B | view product A)
- If the customer view product A, recommend product B's with maximum probability p(B|A)



Example 2 | NPTB (Direct Email)

Bank wants to cross-sell home improvement loan

- Cross-selling: loan ▪

Idea: use **next-product-to-buy model (NPTB)** to identify those customers with an high probability to "buy" a loan

Compute probability:

- **Pr(Buy Loan Next | Demos, Previous products bought)**
- Data: 100,000 customer purchase records for estimation

Does this pay off? EXAMPLE 2

Field Test

- NPTB model produced more responses and more revenues per customer.
- NPTB model produced 530% ROI, vs. -16% for current practice

How is this done? The basic intuition
A simple predictive model for targeting

SEPHORA

Number of offers mailed:	1,000,000
Profit contribution per response:	$80
Cost per mailing:	$.70
Response rate:	1%

Profit = 1,000,000 x .01 x $80 – 1,000,000 x $.70
= $800,000 - $700,000
= $100,000

The direct marketing campaign is effective!

Most of the investment in direct marketing is wasted!

Targeting analysis: lift-based approach

170000= (3%*100000*80)-(100000*0.7)

Decile	Number of Prospects	PredictResp Rate	Profit	Cumulative Profit
1	100,000	3.00%	$170,000	$170,000
2	100,000	2.00%	$ 90,000	$ 260,000
3	100,000	1.40%	$ 42,000	$ 302,000
4	100,000	1.15%	$ 22,000	$ 324,000
5	100,000	1.00%	$ 10,000	$ 334,000
6	100,000	0.60%	$ -22,000	$ 312,000
7	100,000	0.40%	$ -38,000	$ 274,000
8	100,000	0.30%	$ -46,000	$ 228,000
9	100,000	0.10%	$ -62,000	$ 166,000
10	100,000	0.05%	$ -66,000	$ 100,000

=> Profits improvement →$100,000 __>$334,000

We must stop at the fifth decile otherwise we will start losing profits.

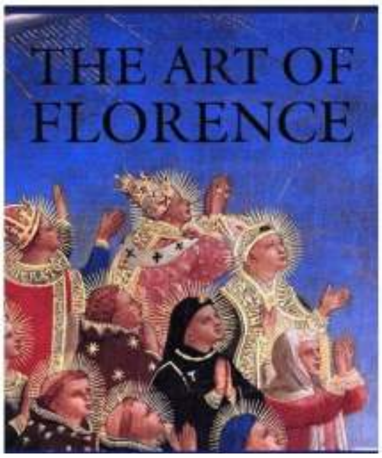
Return on marketing investment: ROI

ROI_1 = $100,000/$700,000 = 14.3%

ROI_2= $334,000/$350,000 = *95.4%*

How do it practically?

Barnes & Noble targeting strategy



- B&N want to send an offer for the purchase of the book "The art of Florence" to a sample of **1,000 customer** in their CMR
- The sample is randomly selected
- **DATA**: For these customers, they have information on previous purchase
  - Number of months since the last purchase (recency)
  - Number of art book previously purchased (art)

Objective: estimate the probability that a generic customer in the target buys the book "The Art of Florence".
Method: Regressive type - Regression (Logit)
Outcome variable of the model: dummy 0/1 (1=buys, 0=does not buy).
Independent variables: Recency, Art.

```
Logistic regression                             Number of obs   =       1000
                                                LR chi2(2)      =      69.14
                                                Prob > chi2     =     0.0000
Log likelihood = -251.46648                     Pseudo R2       =     0.1209

------------------------------------------------------------------------------
    purchase |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     recency |  -.0707172   .0192297    -3.68   0.000    -.1084067   -.0330277
         art |   .9890522   .1346605     7.34   0.000     .7251224    1.252982
       _cons |  -2.225636   .2389241    -9.32   0.000    -2.693918   -1.757353
------------------------------------------------------------------------------
```

INTERPRETATION:

a. Check p-value: is it relevant? -> As the fact that for both regency and art it is 0.000 the estimation is really reliable

b. Coefficients: regency is -0.707172, while art is 0.9890622 so they have different effects on purchases. => Regency is going to decrease the likelihood to purchase by 7%, while Art is going to increase the likelihood to purchase by 9.89%
c. Constant tells us that as a whole, it is really unlikely to purchase the products together.

We trust the model, so we can now predict the future and they way in which customer will behave.
Let's take two consumers: Mario (id.21) and Anna (id.145).
Mario made his last purchase 6 months ago. Also, Mario has previously bought an art book.
Anna made her last purchase 18 months ago, and she has never purchased from the product category (art books).
What is the difference in terms of the likelihood of purchasing the book "The Art of Florence" for the two customers?

Mario (X_1=6, X_2=1)
U_{mario}=-2.22+-0.07*6+0.98*1=
= -1.66

$$p_{Mario} = \frac{\exp(-1.66)}{[1+\exp(-1.66)]} = 0.16$$

Anna (X_1=18, X_2=0)
U_{Anna}=-2.22+-0.07*18+0.98*0=
= -3.48

$$p_{Anna} = \frac{\exp(-3.48)}{[1+\exp(-3.48)]} = 0.03$$

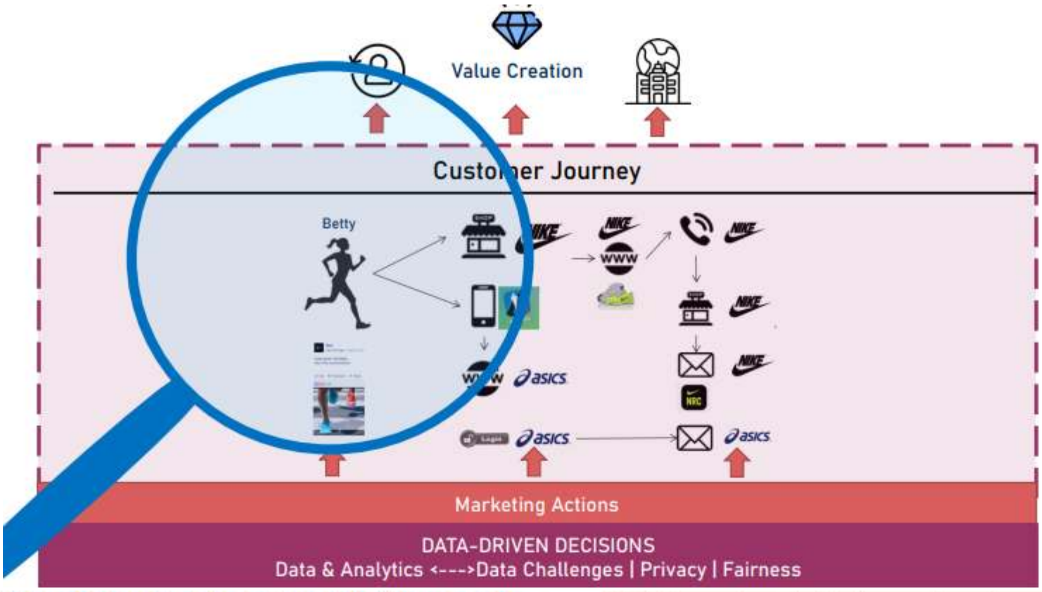
Mario
16%

Anna
3%

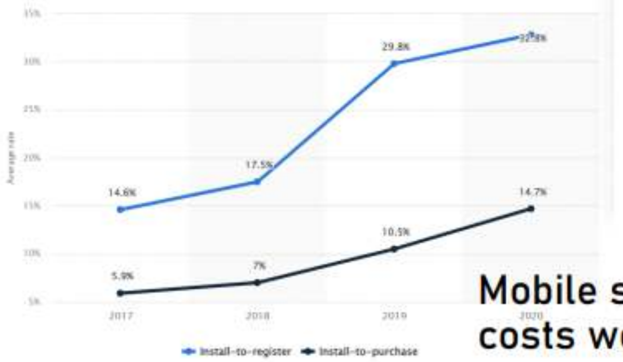
Who should be included in the target?
We know that: the cost of sending the offer is **$1 for each customer**.
The net profit generated from a book purchase by a customer is $6 (net of the cost spent on sending the offer by mail).

($6)*$p_{Ac}$ + (-$1)*(1-$p_{Ac}$) >0 => p_{Ac} > 1/7 => Purchase probability > 14%

Customer acquisition



Mobile shopping app user acquisition rate worldwide from 2017 to 2020



Vodafone cuts outlook after weak performance in Germany NOVEMBER 15, 2022

In Germany, which accounts for 30 per cent of group revenue, adjusted ebitda fell 7.4 per cent to €2.68bn, in part due to losses in broadband customers and higher customer acquisition costs.

Naked Wines shares dive on profitability concerns

"If we are spending £40mn-plus on customer acquisition each year then we have to be clear that we are going to get a satisfactory return on that investment," he said. Jonathan Eley SEPTEMBER 14 2022

Mobile shopping app user acquisition costs worldwide, by type

Characteristic	Cost-to-install	Cost-to-register	Cost-to-purchase
Shopping apps (NET)	2.87	8.76	19.47
Brand commerce apps	4.32	10.93	31.8
Marketplace apps	2.02	7.28	32.44
Coupon & reward apps	5.57	9.32	13.29

The average cost of customer acquisition varies by industry. Insurance customer acquisition rose to $900 per customer

Acquisition phase: our task as manager is to acquire new customers.
What can we do?

- we can use marketing tools in order to obtain the results
- data can be a tool depending on which data we have -> if our problem is an acquisition problem and we want to use secondary data probably we have information about the search pattern.

The acquisition phase is an extremely relevant aspect of the customer journey and a possibility to improve revenues and expand our audience with new individuals and customers. It is a very delicate phase -> it is the more expensive one because we need to convince someone who doesn't know us to purchase our products. So we need to calibrate our action very carefully because they have to be effective.

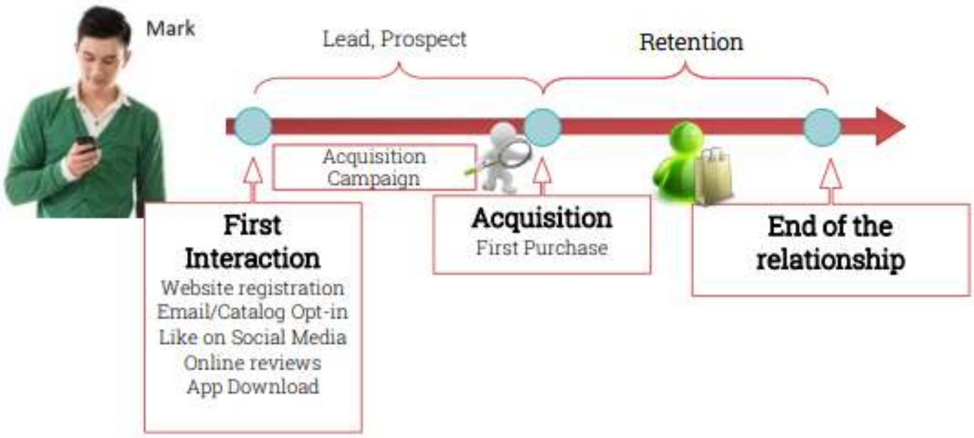
Customer acquisition: Background

Customer acquisition -> the first time a new customer purchases from a firm or subscribes to a service.

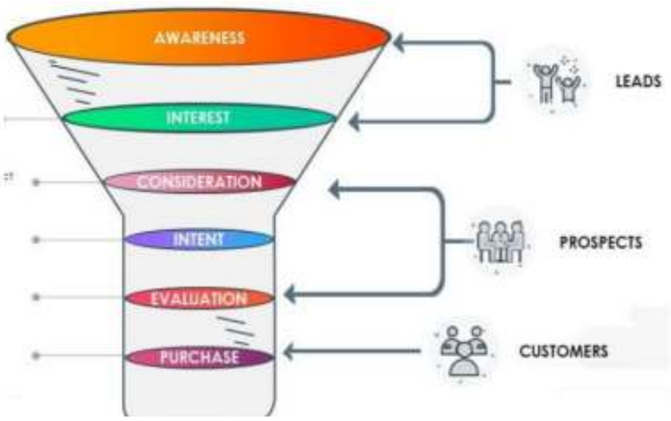
We can distinguish between to distinct worlds:

a. member/contract/subscription-based business (Netflix, Disney +, Spotify, ...) -> you know exactly when the customer was acquired because they sign a contract
b. you need something within your business in order to define if you have acquired a new customer -> a way to define it is to track the very first purchase as acquisition, but it is not possible for every business (ex. Sephora and in-store purchases -> that is why they usually try to convince you to subscribe to the loyalty card, so in this way they can track if you are doing repeated purchase or not)

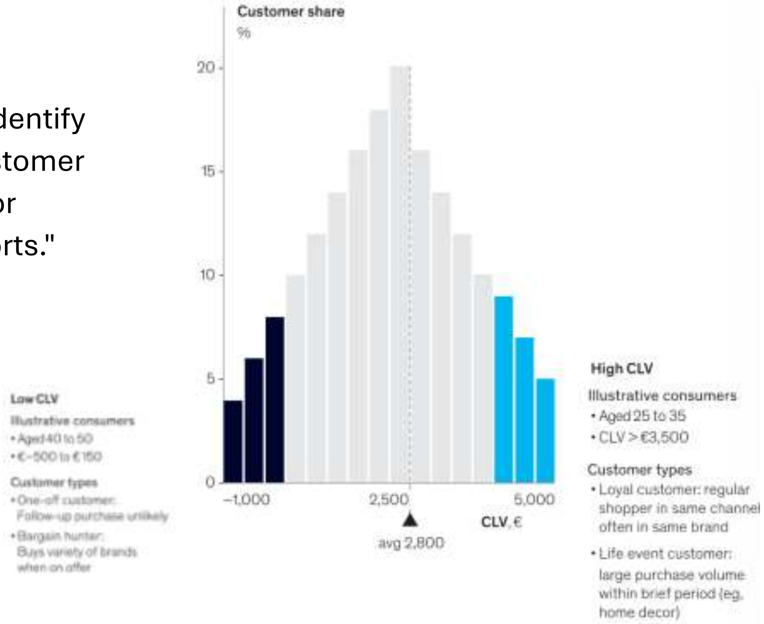
Does the first purchase really represent the first interaction with the brand? NO, but it represents the first engaged interactions with the brand.
In fact, the first purchase rarely represents the firm's first contact with a customer -> first interaction usually came from website registration, email/catalog opt-in like on social media online reviews app download
N.B. We call lead or prospects those individuals that are engaged with the firm (social media, newsletter) but never purchase, be as soon as they start to purchase they become acquire => this means that there is a per-acquisition and a post-acquisition phase with different data



Customer acquisition is a crucial phase of Customer Relationship Management (CRM) because, before focusing on retention and customer satisfaction, a company must identify the customers who are most likely to be acquired and who are worth acquiring



- "Data help marketing and sales teams identify indicators of high CLV and low CAC (customer acquisition costs) respectively, and tailor marketing campaigns to individual cohorts."
- Example "cohort analysis" →



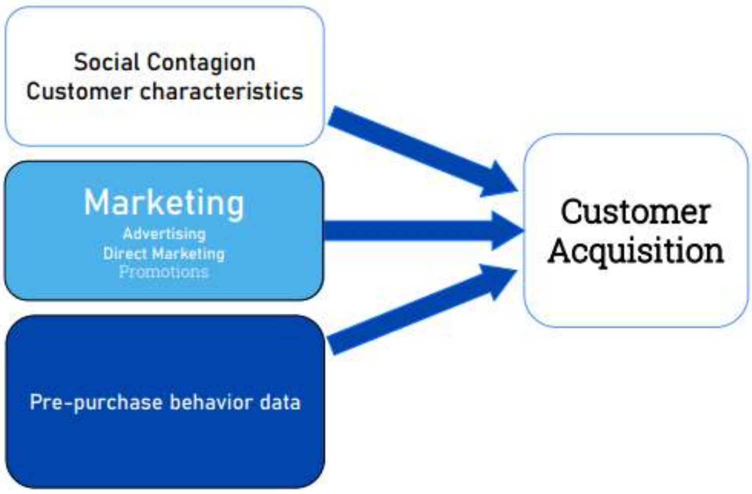
THINGS ARE DIFFERENT TODAY!

Most firms and organizations can:

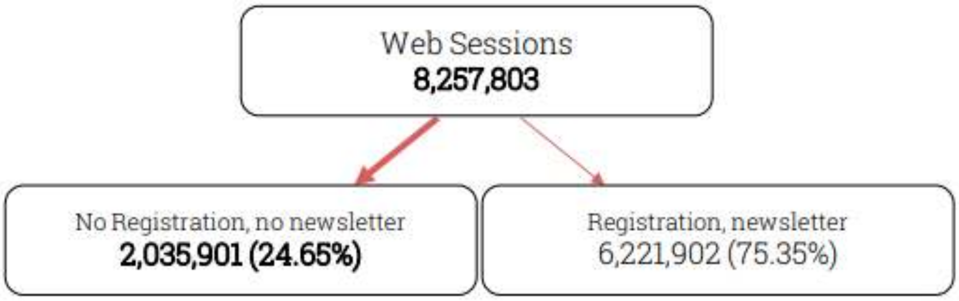
1. Identify the first purchase
2. Track pre-acquisition behavior (e.g. search activity);
3. Monitor marketing activity at customer-level

What impacts the likelihood of acquiring a new customer?

Evidence from scientific literature

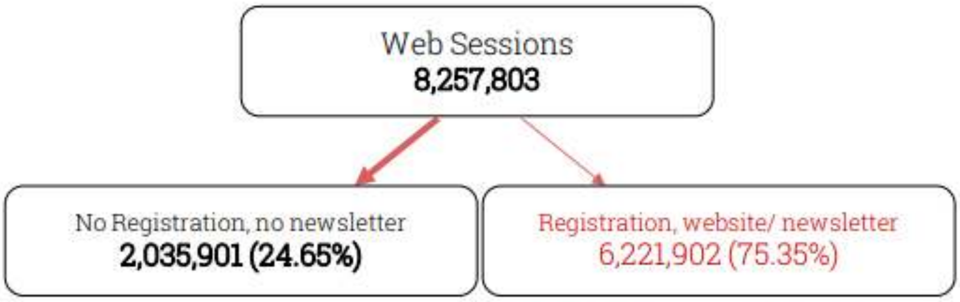


Example: pre-acquisition data



- User identification based on cookies and email ID

Dreambox Add	Number of products added to the Dreambox in each period
Dreambox Drop	Number of products deleted from the Dreambox in each period
Device	Number of devices used in each period
Website Sessions	Number of website sessions in each period
Website Session Length	Average website session time length in each period
m-site Sessions	Number of m-site sessions in each period
m-Site Session Length	Average m-site session time length in each period
Average Pages/screens	Average number of pages/screens seen in each session
Average Products	Average number of products seen in each session
Average Suggested Products	Average number of suggested products (similar/same designer) seen in each period
Average Clicks	Average number of clicks in each session
Average Filters	Average number of filters used in each searching session
Total Pages/screens	Total number of pages/screens seen in each period
Total Products	Total number of products seen in each period
Total Suggested Products	Total number of suggested products (similar/same designer) seen in each period
Total Clicks	Total number of clicks in each session
Total Filters	Total number of searching sessions with filters in each period
Ranking Type	Number of times the user has used each type of ranking in each period



For 75% of users, the firm begins monitoring their behavior (also prior to their acquisition)

- User identification based on cookies and email ID

SOURCE: Anonymous Company

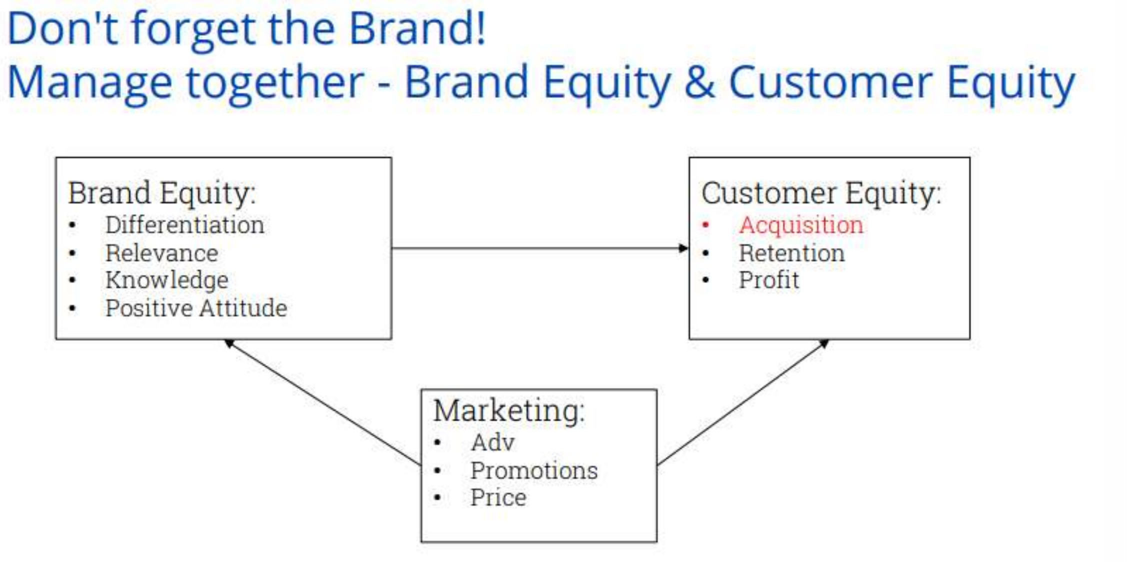
There is a phase in which for the company we are a cookie, and ID that is interacting with them -> in that phase, even before purchasing, the incentive of the business is try to incentivise some form of registration (social media registration, newsletter, download the app) -> they try to push us to do these actions, in order to start collecting data about us, our preferences and our habits.

- For 75% OF USERS the firm begins monitoring their behaviour (also prior to their acquisition) -> even if you share a fake email, or a mail that you never use the information that you are sharing with them is still relevant because they will be able to analyse your behavior and linked your device with an ID number -> we you register the company start to have the right to collect data about you because when you do the registration you always accept privacy terms!

Marketing Analytics to Acquire the Customer: what to DO and NOT to do

What TO DO (Do) and what NOT to do (don't):

- Do: Believe in the value of Data.
- Do: Invest in Data first, then in the Method.
- Do: Think "across channels."
- Don’t: Forget the BranD
- Don’t: Ignore Privacy
- Don't: Believe in the Big Data trend "without critical thinking." ▪ (Identifying the right marketing strategies and the right data is not easy.)



Exercise:

Try to register / understand how to register on the Website / Newsletter of two brands or companies ▪ Alternatively, check your email account to find requests for newsletter registration / renewal of privacy terms

- New balance:
  1) accept cookies
  2) do you want to receive new info before the others?
  3) insert the email address
- Dove:
  1) cookies

2) It is possible to register through the uk website (no italian):
   - Age >=16
   - Email address

The registration phase leads us, which is strictly related to customer acquisition, to three strategic decisions:

1. How to ask to register
2. Depending on the country we could have different regulation
3. What information ask

⇨ This is a key phase cause a first occasion to target our audience

What does PRIVACY have to do with marketing?

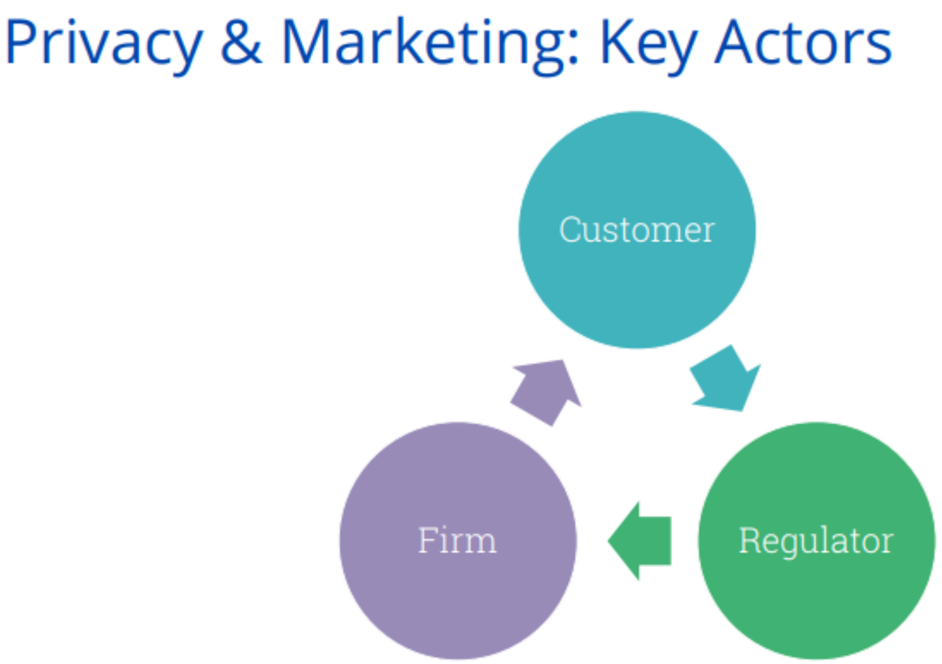
WHY?

- DATA → Better Targeting → Customized Marketing Strategy

CHALLENGES

- Profiling and targeting
- Data collection and retention
- How will the way of retaining and collecting data change

In Europe we have GDPR

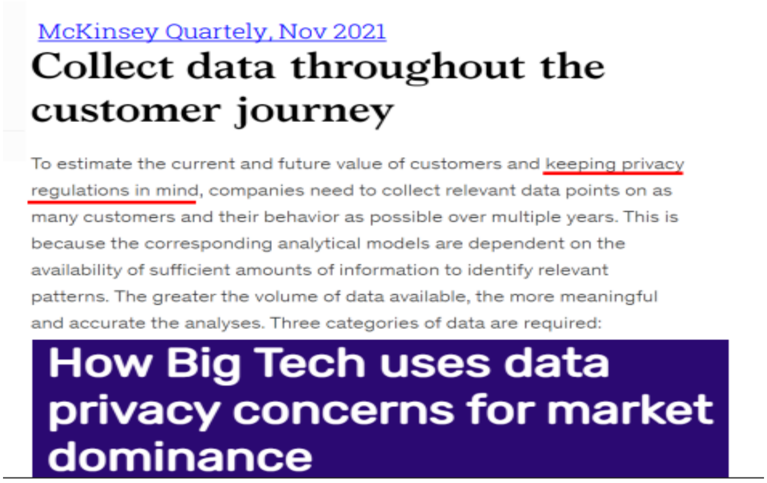


FIRMS: PRIVACY AS BUSINESS PROBLEM


Marketing In The New Era Of Data Privacy

Forbes

A customer-centric approach to marketing in a privacy-first world

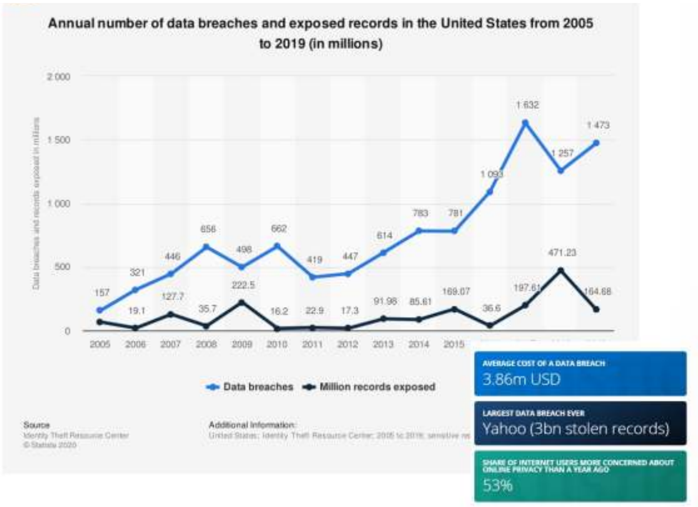

McKinsey Quartely, Nov 2021

Collect data throughout the customer journey

To estimate the current and future value of customers and keeping privacy regulations in mind, companies need to collect relevant data points on as many customers and their behavior as possible over multiple years. This is because the corresponding analytical models are dependent on the availability of sufficient amounts of information to identify relevant patterns. The greater the volume of data available, the more meaningful and accurate the analyses. Three categories of data are required:

How Big Tech uses data privacy concerns for market dominance

As consumers we are more aware about the importance of privacy -> we might be sensitive about the information that we want to share.



DATA BREACHES

Surprising increase in the number of Data Breach events since 2015.

Increase awareness value of data

If a brand is involved in a data breach than of course is a signal for us to be more aware about our data.


Netflix's Race-Based Marketing Shows Potential For Anticompetitive Data Abuses

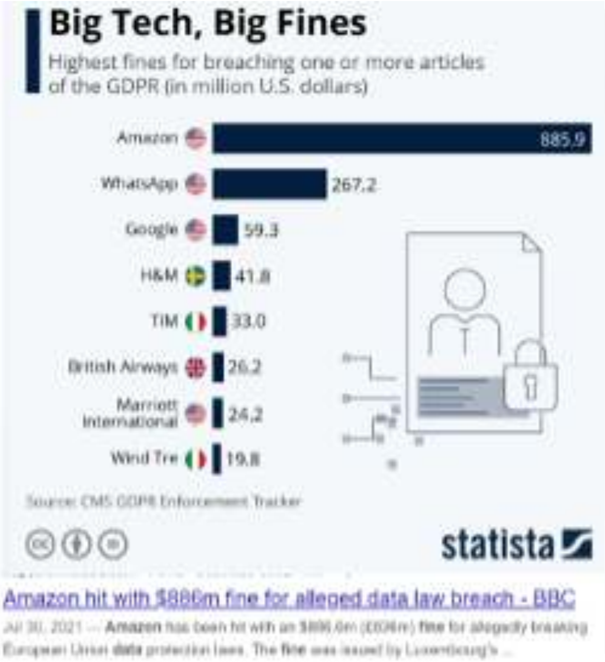
Adam Candeub Contributor
Washington Bytes Contributor
Policy

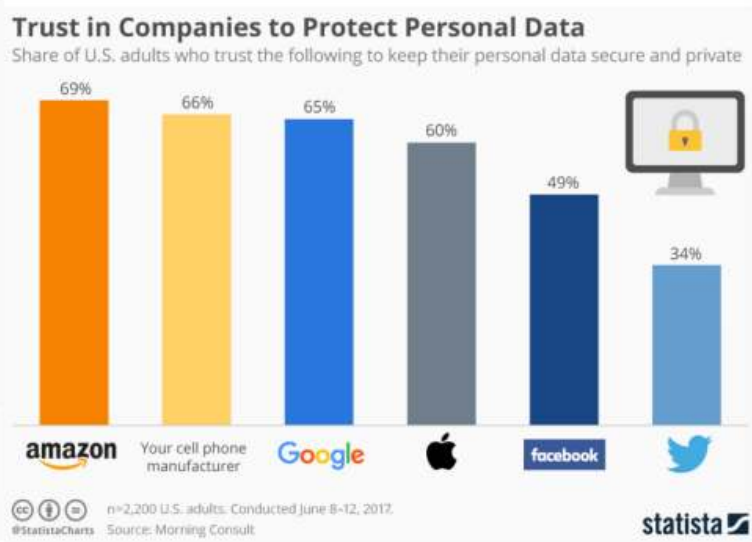
The most recent example of Netflix's abuse of personal data is its alleged promoting to African Americans of videos that show black characters. Without denying its discriminatory marketing, Netflix responded that, "We don't ask members for their race, gender or ethnicity, so we cannot use this information to personalize their individual Netflix experience. The only information we use is a member's viewing history."

Source: https://www.forbes.com/sites/washingtonbytes/2018/10/30/netflixs-race-based-marketing-shows-potential-for-anticompetitive-data-abuses/?sh=7ed48b173f48

In 2018, Netflix faced criticism when users discovered that the algorithm was sometimes categorizing content by race, leading to problematic recommendations

The answer of Netflix in that case was that they were just using data and algorithms -> the algortim maybe was biased, which means that data can also be problematic regarding to privacy; algortihms can be super useful but at the same time they can create problems



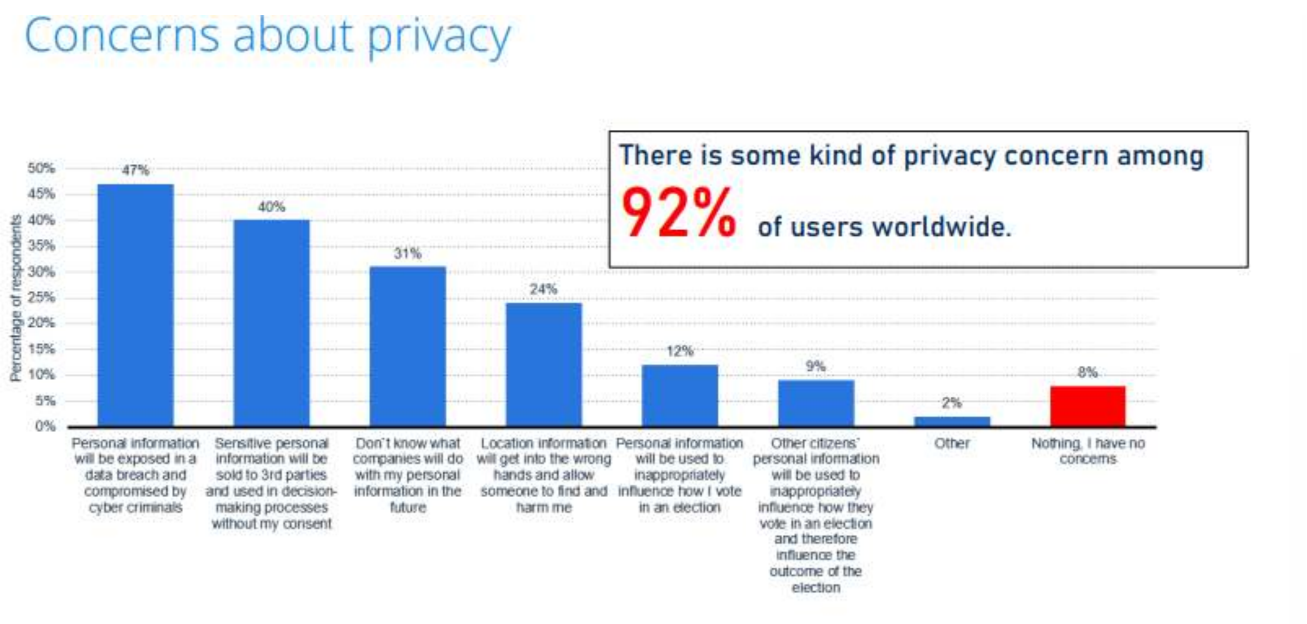


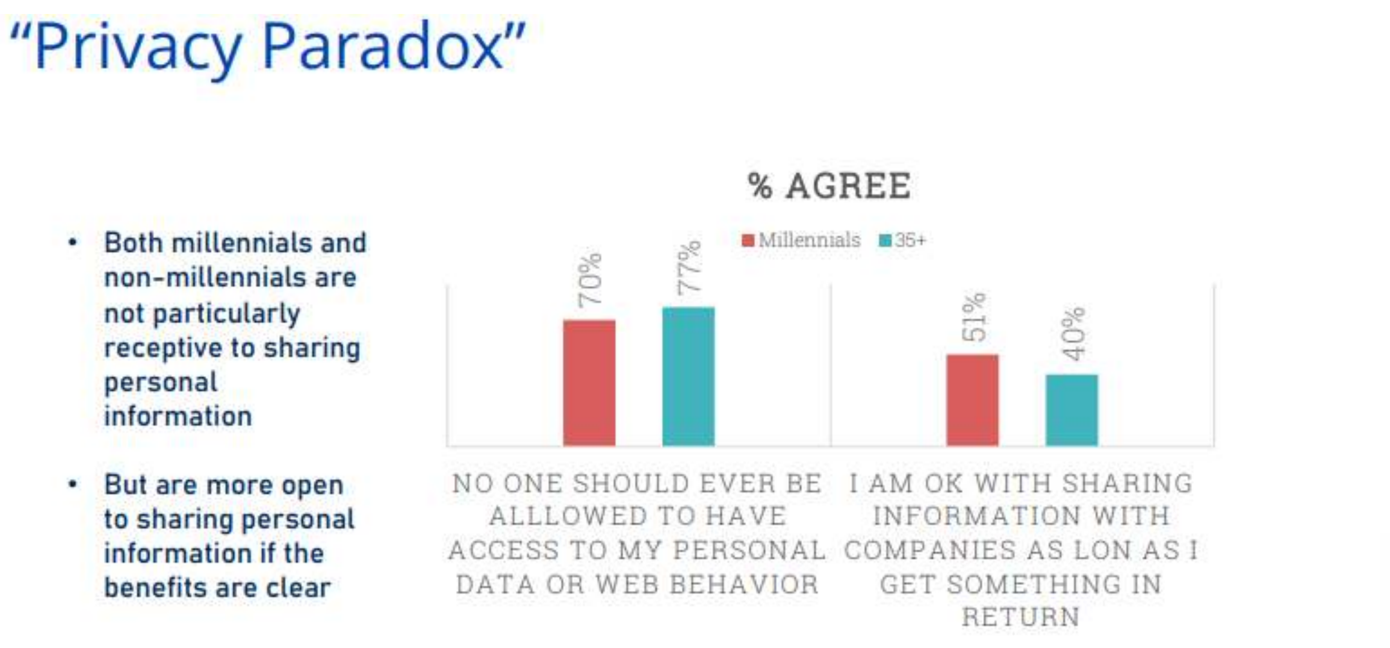
Amazon's Gender-Biased Algorithm

In 2018, it was reported that Amazon's recruiting tool, which used machine learning to review resumes and identify top candidates, showed a gender bias. The algorithm favored male candidates over female candidates, reflecting the gender disparities present in the tech industry.

Data-breaches and fines

Instagram: $403 million
In September 2022, Ireland's Data Protection Commissioner (DPC) fined Instagram for violating children's privacy under the terms of the GDPR.
T-Mobile: $350 million (July 2022)
WhatsApp: $255 million
Facebook-owned messaging service WhatsApp was fined €225 million ($255 million) in August 2021 for a series of GDPR cross-border data protection infringements in Ireland.





The value of personal information

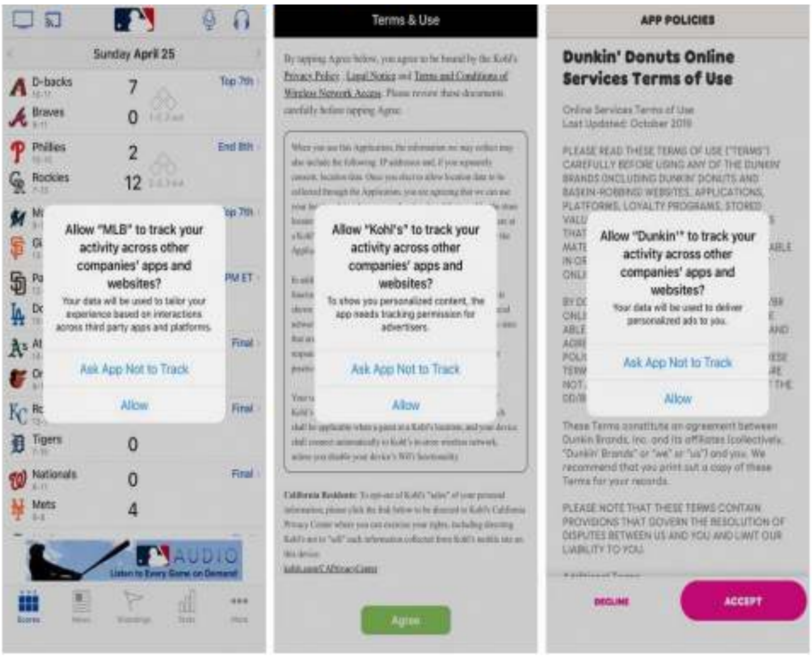
We are moving towards a world where consumers will have to allow the use of personal information: What will individuals do? Which individuals will be more inclined to give up personal information? Consumers are heterogeneous in terms of privacy preferences → The data and opt-ins collected might not represent the entire population (Lin 2021).

⇨ The 'privacy paradox' exists! → Businesses need to better understand consumer behavior and their intention to share personal information (Kim Barasz and John 2021) → difference between third party and first party

This explains why privacy has become a marketing strategy -> the way they ask you to accept the privacy conditions has a marketing strategy behind.
Particularly in Europe this is really relevant because if you don't accept the privacy they MUST stop tracking you -> the data about you are not available anymore, so for firms is fundamental to have customers that accept privacy conditions by structuring a successful marketing strategy linked to it.

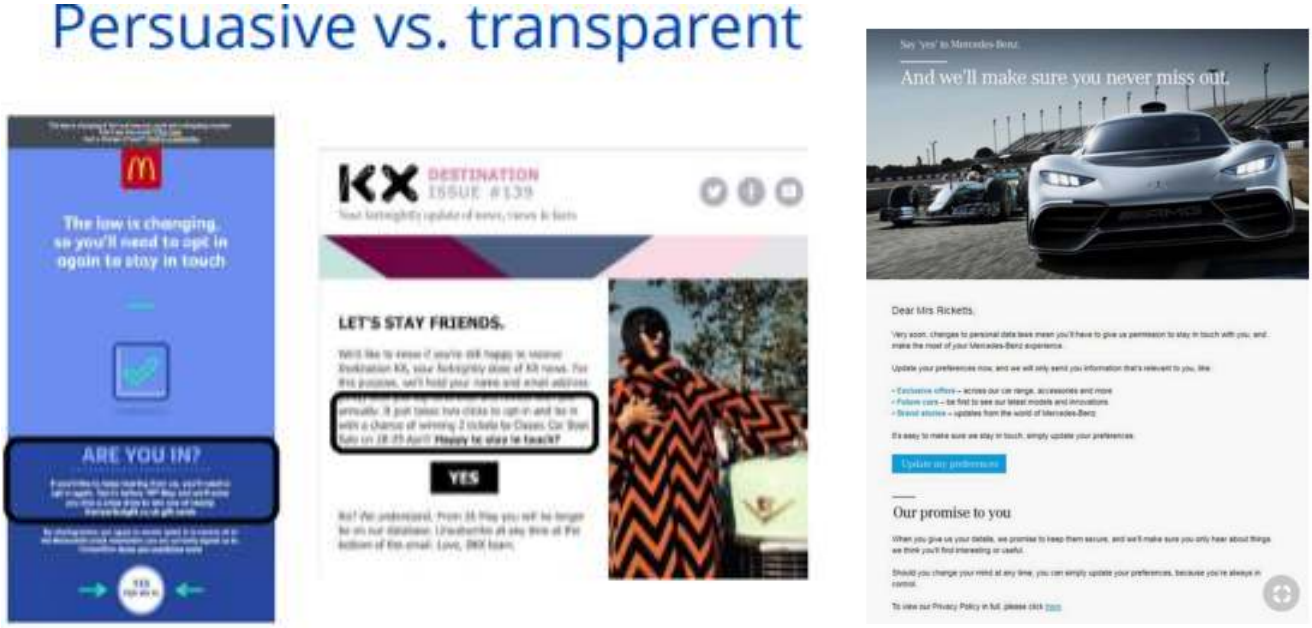
APP TRACKING TRASPARENCY: APPLE

Caso Apple



The interesting thing about this campaign is that Apple is informing you that we dealing with partner company you can choose whether share your personal information our not, but this is not valid for them who keep tracking you.

Persuasive vs. transparent



Example: Field Test – How to get Opt-ins

Informative High

Persuasive High

The image used in the field test removed

: Field Test Groups

	Persuasive		
Informative	*Low*	*Moderate* Sconto	*High* Sconto & Framing
Low	**G1 Generic Message** Total contacts=2339	**G3 Moderately Persuasive** Total contacts=2340	**G5 Highly Persuasive** Total contacts=2340
High	**G2 Informative** Total contacts=2340	**G4 Informative & Moderately Persuasive** Total contacts=2340	**G6 Informative & Highly Persuasive** Total contacts=2379

Field Test: Logistic Regression (n=14078) DV=accept privacy=1, 0 otherwise

	Coef.	z	p-value
Informative	0.35	1.48	0.138
Moderately Persuasive Discount	1.04	4.97	0.000
Highly Persuasive Discount + Framing	-0.17	-0.66	0.512
Moderately Persuasive * Informative	-0.19	-0.70	0.487
Highly Persuasive * Informative	0.68	2.09	0.037
Constant	-4.28	-24.03	0.000
Number of Observations=14,078			
LR $\chi^2(5) = 83.32$, p-value=0.000			

As a company you can not only use data but you can also think about seling data -> that is why a marketing strategy about privacy is fundamental!

ARTEA: DESINING TARGETING STRATEGIES HARVARD BUSINESS REVIEW – CASE

HBR Artea: a customer acquistion problem

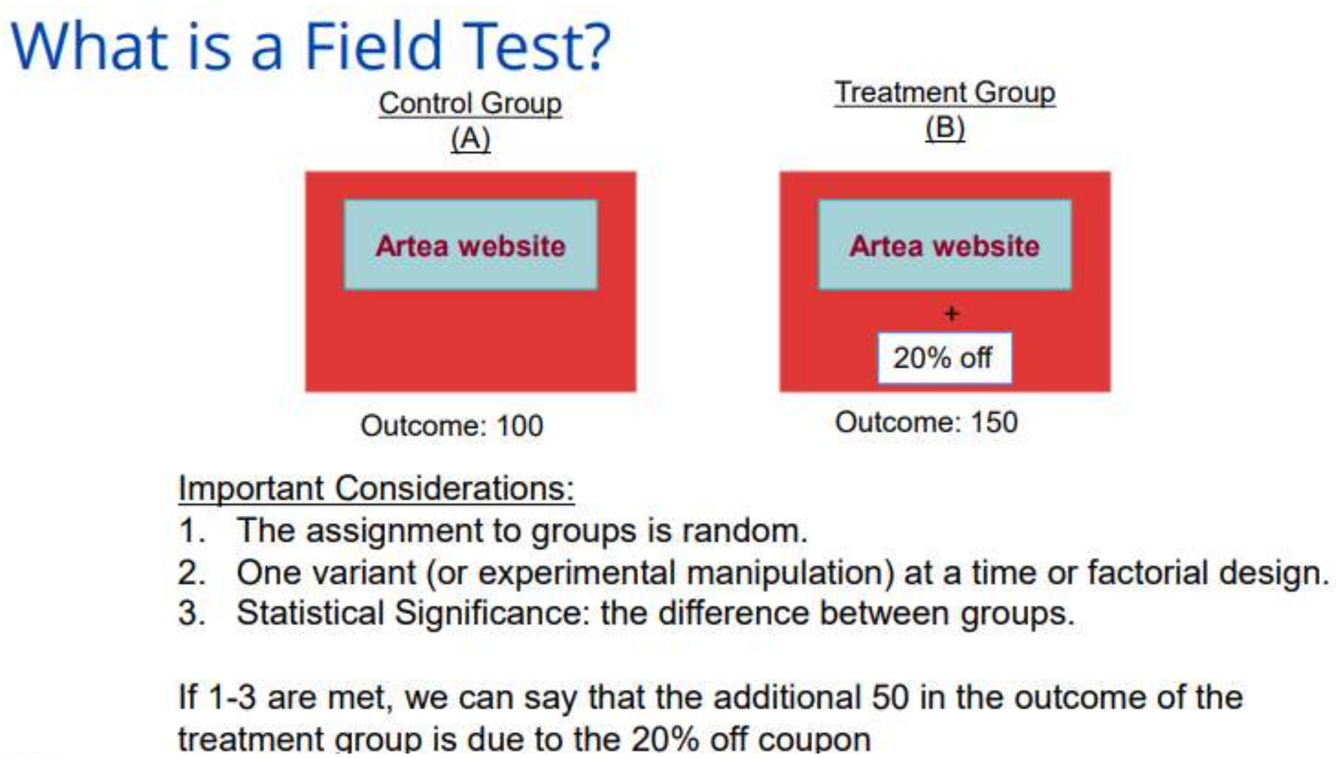
- Industry: Clothing and Accessories
- Data Driven Culture – Data Science Team – Customer Dashboard
- Business Problem:
    - 87% of those who visited the site have never made a transaction.
    - Engagement metrics (e.g. time spent on the site, reviews, etc.) are okay.
    - CEO Alex Campbel wants to increase sales by leveraging the data available from the pre-acquisition phase coming from the website.
    - What could Artea do to improve acquisition?
    - Wait for them to purchase or use strategies to encourage the first purchase?
- **Action**: Alex asked the data science team to explore the possibility of incentivizing purchases, and in this way also improving acquisition, by sending discount coupons to users registered on the site. He decided to run a FIELD test by sending out a 20% discount coupon

Dati field test (AKA A/B test)

5,000 website users
2,502 → treatment group (received the coupon)
2,498 → control group (did not receive the coupon)

Available information:

- dataset containing information on past behavior (purchase and browsing)



A Field test is an experiment, whiich is a procedure in which one or more **variables (treatments)** are **manipulated**, and OBSERVED data related to an "**outcome**" variable of interest (e.g., choice, amount spent, purchase frequency) are collected, while **controlling for other variables** that might distort the result (e.g., consumer characteristics, etc.).

⇨ Without experimentation, there is an association but not causation

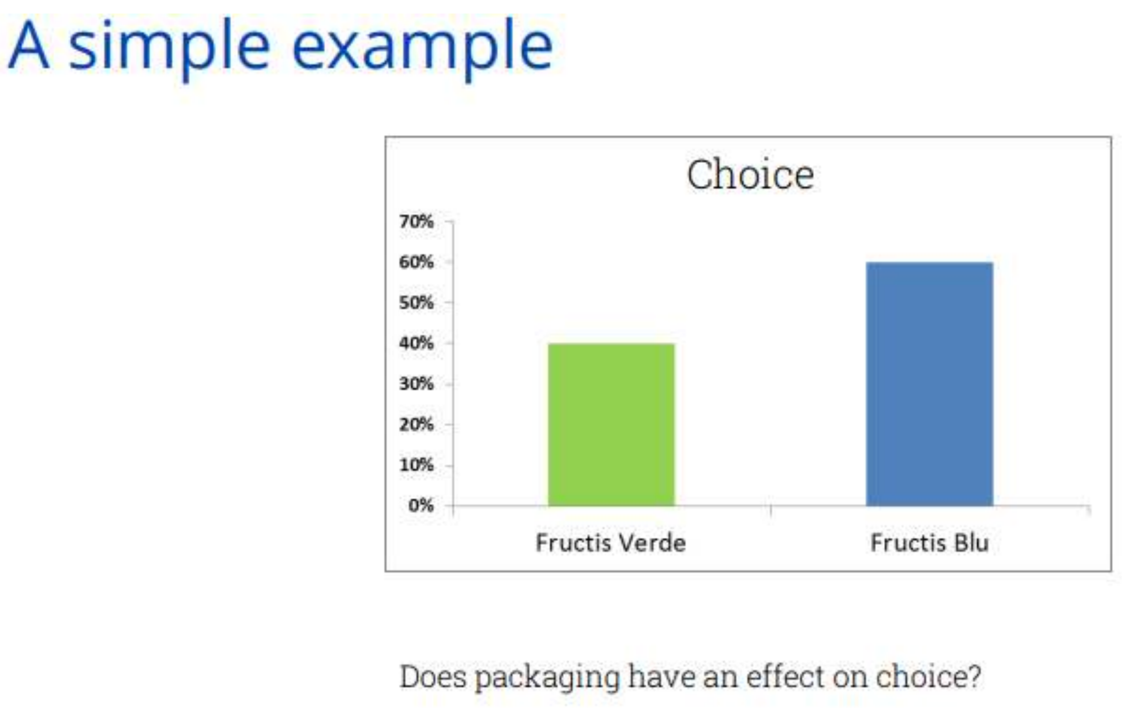
FIELD Test Logic: A simple example

FIELD Test Logic: Causation

Necessary requirements to say that X ----> Y: •

(1) X must occur before Y
(2) There must be evidence of an association between X and Y
(3) Control for other factors.
In an experiment:
(1) X is typically manipulated.
(2) The relationship between X and Y can be estimated by collecting data.
(3) This is typically done through randomization.
Example:

- Color Packaging ---------> Choice



Experiments:

- Manipulation: The practice related to the creation of different levels of a variable X. Jargon: the variable X is manipulated.
- Independent Variable (X): Variable X manipulated or altered by the "researcher." Example: X= color of the packaging.
- Dependent Variable (O): Variable for which the "experimenter" expects a change following the manipulation of X. The success of the experiment will be evaluated based on the level of O.
- Example: Choice of Brand X (e.g. Fructis)
- Experimental Group (EG): The group of individuals subjected to the experiment


Connecticut


- Control Group (CG): Group of individuals not subjected to the experiment


Florida


- Extraneous or Confounding: Variables that, in addition to X, can have an impact on O (e.g., competitors' reactions).
- Selection Bias: A problem that occurs if the experimental group is systematically different in relevant aspects from the control group. In other words, if subjects assigned to the experimental group systematically differ from subjects assigned to the control group

Connecticut



Florida


- Randomization (R): Procedure through which subjects/units are randomly assigned to groups (experimental and control)
- Treatment Effect: Result of the experiment
  (e.g. Blue Choice - Green Choice = 0.6 - 0.4 = 20%)
- Experimental Design: Set of procedures that guide the experimental study.
  6 Relevant Steps:
  1. Which variables do you want to manipulate/check (e.g. packaging)
  2. Which levels of variable X need to be manipulated (e.g. color)
  3. What is the dependent variable (e.g. choice)
  4. How to select the units to be tested
  5. How to control for selection bias
  6. How to minimize the influence of external factors.

Artea: Data

Nome	Descrizione
id	Unique identification code of the customer or potential customer
trans_after	Number of transactions after the experiment
revenue_after	Total revenues ($) after the experiment
test_coupon	Dummy variable that takes the value of 1 if the customer or potential customer received the coupon
num_past_purch	Number of previous purchases
spent_last_purchase	Total amount spent ($) in previous purchases
weeks_since_visit	Number of weeks since the last visit to the site
browsing_minutes	Total minutes spent on the site during the last visit
shopping_cart	Indicates if the user added a product to the cart during the last visit but did not make a purchase (1=yes, 0=no)
Channel of Acquisition	Refers to the channel corresponding to the first contact with Artea (registration on the site), therefore acquired as a prospect and not as a customer:
channel_Facebook	Indicates if the customer was acquired via Facebook (1=yes, 0=no)
channel_Instagram	Indicates if the customer was acquired via Instagram (1=yes, 0=no)
channel_Referral	Indicates if the customer was acquired via Referral (1=yes, 0=no)
channel_Other	Indicates if the customer was acquired through other channels (1=yes, 0=no)

Questions

1. Could Artea increase transactions due to this campaign? Can Artea increase the average spending ($) per customer? If so, by how much?
2. Who among the registered users should receive the coupon?

Artea: Designing Targeting Strategy

SUGGESTED STEPS IN THE ANALYSIS:

Step 0: Open the dataset 'Artea.xls'

```
#Step 0:Open dataset Artea.xls
db_artea = pd.read_excel('DataArtea.xlsx')
db_artea.head()
```

	id	trans_after	revenue_after	test_coupon	num_past_purch	spent_last_purchase	weeks_since_visit	browsing_minutes	shopping_cart	channel_Facebook	channel_I
0	6001	0	0.0	0	6	62.99	6	1	0	1	
1	6002	0	0.0	1	2	53.99	0	7	1	0	
2	6003	0	0.0	1	3	88.98	3	4	0	1	
3	6004	0	0.0	0	1	68.99	1	19	0	1	
4	6005	0	0.0	1	3	66.49	4	20	0	0	

```
db_artea.shape
```

```
(5000, 14)
```

Step 1: Verify the random assignment

Objective: We calculate the average values of the observable variables before the field test, comparing them between the control group and the treatment group. If the allocation is random, we should expect two nearly identical groups.

```
# Step 1: Check for correct random allocation in the Control vs. Experimental (or coupon) group
desc = ['num_past_purch', 'spent_last_purchase', 'weeks_since_visit', 'browsing_minutes', 'shopping_cart', 'channel_Facebook', 'channel_Instagram', 'cha
db_artea.groupby("test_coupon")[desc].mean()
```

	num_past_purch	spent_last_purchase	weeks_since_visit	browsing_minutes	shopping_cart	channel_Facebook	channel_Instagram	channel_Referral	channel_
test_coupon									
0	2.019616	56.689480	3.182946	13.707366	0.299840	0.206165	0.314251	0.045236	0.0
1	2.091527	58.155823	3.256994	13.670663	0.285771	0.217426	0.306555	0.049960	0.0

We compared the control group with the experimental group for observable variables before the field test experiment. We note that there are no evident differences in the averages. This seems to confirm that the groups were correctly selected randomly, but we can still test this formally. We could choose single t-tests that compare the coupon=1 group vs. coupon=0. (See the results in Table 2 as an example with one variable).

```
# Step 1: we can run a t-test for each variable

import statsmodels.stats.weightstats as smws

# Split the data into two groups based on 'test_coupon'
group1 = db_artea[db_artea['test_coupon'] == 0]['num_past_purch']
group2 = db_artea[db_artea['test_coupon'] == 1]['num_past_purch']

# Perform t-test
t_stat, p_value, df = smws.ttest_ind(group1, group2, alternative='two-sided', usevar='pooled')

print(f"T-statistic: {t_stat}")
print(f"P-value: {p_value}")
```

```
T-statistic: -0.9930013925760496
P-value: 0.3207573469477035
```

```
# step 1: try with channel facebook, which is a dummy so we need a t-test for proportions:

from statsmodels.stats.proportion import proportions_ztest

# Calculate the number of successes (channel_Facebook = 1) for each group
count_group1 = db_artea[db_artea['test_coupon'] == 0]['channel_Facebook'].sum()
count_group2 = db_artea[db_artea['test_coupon'] == 1]['channel_Facebook'].sum()

# Calculate the total number of observations for each group
nobs_group1 = len(db_artea[db_artea['test_coupon'] == 0])
nobs_group2 = len(db_artea[db_artea['test_coupon'] == 1])

# Perform the z-test for proportions
z_stat, p_value = proportions_ztest([count_group1, count_group2], [nobs_group1, nobs_group2])

print("Z-statistic:", z_stat)
print("P-value:", p_value)
```

```
Z-statistic: -0.9744409774178499
P-value: 0.3298376316653264
```

```
db_artea.groupby("test_coupon")[['trans_after', 'revenue_after']].mean()
```

	trans_after	revenue_after
test_coupon		
0	0.125701	7.780168
1	0.151878	7.538673

Note that the difference is not significant. This is the result we expected. Randomization ensures that the two groups are identical, meaning there are no significant differences in characteristics and behaviors prior to the field test between the treated group and the control group.

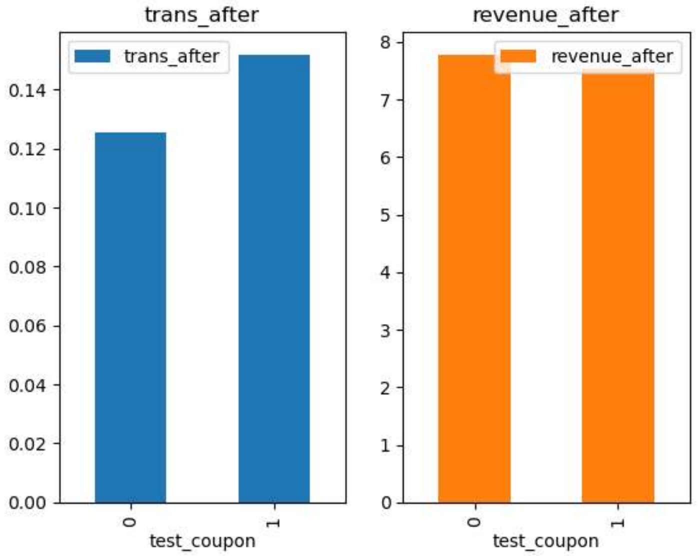
Step 2: Model Free Evidence: Check the effect of Coupon through descriptive statistics

Objective: We want to check if the variables observed after the experimental treatment, namely "trans_after" and "revenues_after", show different average values for the Treated vs. Control group.

Let's start with a descriptive analysis and observe the difference in means to determine if noticeable differences exist.

```
#Step 2: Model Free Evidence - Effectiveness of the COUPON
db_artea.groupby("test_coupon")[['trans_after', 'revenue_after']].mean().plot.bar(subplots=True, layout=(1,2))

#The parameter 'subplots=True' splits the result into two separate plots, one for 'trans_after' and the other for 'revenue_after'.
#Using 'layout=(1,2)' organizes the two plots in a single row and two side-by-side columns.
```

```
array([[<AxesSubplot:title={'center':'trans_after'}, xlabel='test_coupon'>,
        <AxesSubplot:title={'center':'revenue_after'}, xlabel='test_coupon'>]],
      dtype=object)
```

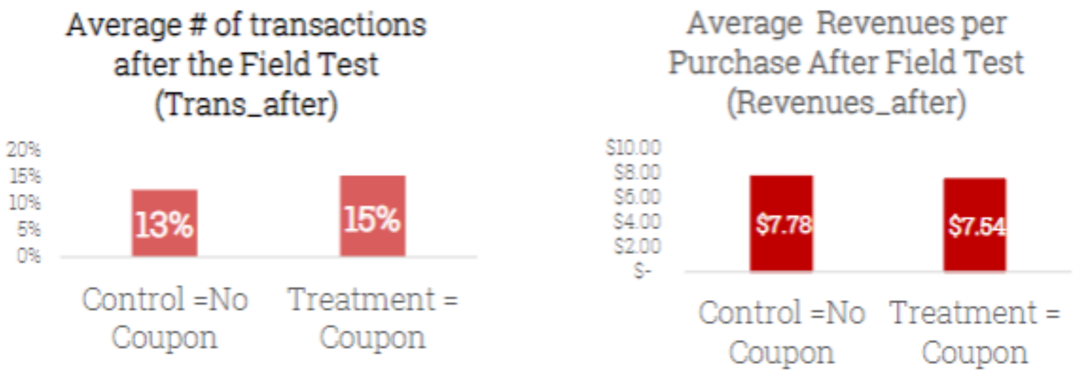


The difference between the average "trans_after" of the treatment group and the control group is 13% -10% = 3%.
3% represents the effect of the experimental treatment (also called TREATMENT EFFECT).
The difference between the average "revenues_after" of the experimental group and the control group is 7.78

- 7.53 = 0.24.
The differences appear very small. We will check in step 3 if they are statistically significant.

Step 2: Model Free Evidence – Efficacia del Coupon

Descriptive Statistics



+ 2% Are these differences significant? **- 0.24**

What is the error associated with the estimate?

Step 3: Test the effect of coupons on consumer behavior

We can then more "formally" test using a statistical test to compare means of independent samples if the difference in means is "statistically significant". In this case, given it's a randomized experiment, we can perform a t-test for independent samples for each outcome variable. Alternatively, two regressions can be conducted with dependent variables first being "trans_after" and then "revenues_after", and the independent variable being "test_coupon". The results will align.

```python
# STEP 3: TEST COUPON EFFECTIVENESS ON TRANS_AFTER USING A REGRESSION

import statsmodels.formula.api as smf

model = smf.ols(formula='trans_after ~ test_coupon', data=db_artea).fit()

print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            trans_after   R-squared:                       0.001
Model:                            OLS   Adj. R-squared:                  0.001
Method:                 Least Squares   F-statistic:                     4.872
Date:                Tue, 24 Sep 2024   Prob (F-statistic):             0.0273
Time:                        01:56:56   Log-Likelihood:                -2748.1
No. Observations:                5000   AIC:                             5500.
Df Residuals:                    4998   BIC:                             5513.
Df Model:                           1
Covariance Type:            nonrobust
===============================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------
Intercept       0.1257      0.008     14.982      0.000       0.109       0.142
test_coupon     0.0262      0.012      2.207      0.027       0.003       0.049
==============================================================================
Omnibus:                     3814.872   Durbin-Watson:                   1.963
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            66979.526
Skew:                           3.610   Prob(JB):                         0.00
Kurtosis:                      19.413   Cond. No.                         2.62
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

```python
# Step 3: we can run a t-test alternatively

import statsmodels.stats.weightstats as smws

# Split the data into two groups based on 'test_coupon'
group1 = db_artea[db_artea['test_coupon'] == 0]['trans_after']
group2 = db_artea[db_artea['test_coupon'] == 1]['trans_after']

# Perform t-test
t_stat, p_value, df = smws.ttest_ind(group1, group2, alternative='two-sided', usevar='pooled')

print(f"T-statistic: {t_stat}")
print(f"P-value: {p_value}")
```

```
T-statistic: -2.207188281608223
P-value: 0.027346188300061674
```

```
# STEP 3: TEST COUPON EFFECTIVENESS ON REVENUES_AFTER USING A REGRESSION

model = smf.ols(formula='revenue_after ~ test_coupon', data=db_artea).fit()

print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          revenue_after   R-squared:                       0.000
Model:                            OLS   Adj. R-squared:                 -0.000
Method:                 Least Squares   F-statistic:                    0.1306
Date:                Tue, 24 Sep 2024   Prob (F-statistic):              0.718
Time:                        02:00:55   Log-Likelihood:                -22906.
No. Observations:                5000   AIC:                         4.582e+04
Df Residuals:                    4998   BIC:                         4.583e+04
Df Model:                           1
Covariance Type:            nonrobust
===============================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------
Intercept       7.7802      0.473     16.456      0.000       6.853       8.707
test_coupon    -0.2415      0.668     -0.361      0.718      -1.552       1.069
==============================================================================
Omnibus:                     3946.015   Durbin-Watson:                   1.968
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            74926.257
Skew:                           3.764   Prob(JB):                         0.00
Kurtosis:                      20.406   Cond. No.                         2.62
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

STEP 3: Effect of Coupon # of Transactions

Y= # of transactions after Field Test

	Coefficients	Standard Error	t Stat	P-value
Intercept	0.13	0.01	14.98	0.00
test_coupon	0.03	0.01	2.21	0.03

We can include more variables to control the effect on Y, but since the coupon test is randomized, its impact will not change significantly.

# Transactions	Coef	SE	t Stat	P value
channel_Other	0.16	0.04	4.36	0.00
channel_Referral	0.13	0.03	4.96	0.00
channel_Instagram	0.11	0.01	8.39	0.00
channel_Facebook	0.11	0.01	7.83	0.00
shopping_cart	0.17	0.01	14.41	0.00
browsing_minutes	0.00	0.00	3.91	0.00
weeks_since_visit	-0.02	0.00	-8.56	0.00
spent_last_purchase	0.00	0.00	-3.43	0.00
num_past_purch	0.06	0.00	24.62	0.00
test_coupon	0.03	0.01	2.38	0.02
Intercept	-0.06	0.02	-3.56	0.00

STEP 3: Test effect of coupon

Y= # of transactions after Field Test

	Coefficients	Standard Error	t Stat	P-value
Intercept	0.13	0.01	14.98	0.00
test_coupon	0.03	0.01	2.21	0.03

Y= revenues after Field Test

	Coefficients	Standard Error	t Stat	P-value
Intercept	7.78	0.47	16.46	0.00
test_coupon	-0.24	0.67	-0.36	0.72

STEP 3: Effect of Coupon # of Transactions

Y= # of transactions after Field Test

	Coefficients	Standard Error	t Stat	P-value
Intercept	0.13	0.01	14.98	0.00
test_coupon	0.03	0.01	2.21	0.03

We can include more variables to control the effect on Y, but since the coupon test is randomized, its impact will not change significantly.

# Transactions	Coef	SE	t Stat	P value
channel_Other	0.16	0.04	4.36	0.00
channel_Referral	0.13	0.03	4.96	0.00
channel_Instagram	0.11	0.01	8.39	0.00
channel_Facebook	0.11	0.01	7.83	0.00
shopping_cart	0.17	0.01	14.41	0.00
browsing_minutes	0.00	0.00	3.91	0.00
weeks_since_visit	-0.02	0.00	-8.56	0.00
spent_last_purchase	0.00	0.00	-3.43	0.00
num_past_purch	0.06	0.00	24.62	0.00
test_coupon	0.03	0.01	2.38	0.02
Intercept	-0.06	0.02	-3.56	0.00

The coupon causes a 0.03 pp increase in transaction likelihood.
The coupon does not have a significant effect on revenues.

Step 4: Heterogeneity in responsiveness to coupon

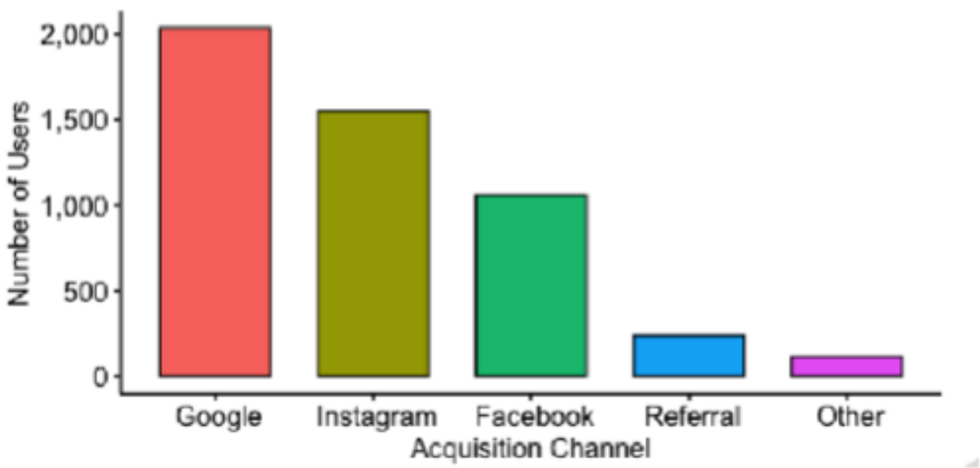
Purpose: understand if the Coupon is more/less effective for customers with certain observable characteristics. We can start by looking at the effect of the coupon by different channels of acquisition

We have variables in the database capturing whether the customer was acquired via Instagram, Facebook, Referral, etc.

```
#Step 4: Explore customer heterogeneity
# start with different channels of acquisition
db_artea.groupby(["channel", 'test_coupon'])[['trans_after', 'revenue_after']].mean()
```

channel	test_coupon	trans_after	revenue_after
Facebook	0	0.159223	9.814913
	1	0.220588	10.895294
Google	0	0.078870	4.922288
	1	0.058532	2.803897
Instagram	0	0.146497	9.171783
	1	0.204694	10.337346
Other	0	0.258621	14.859483
	1	0.241379	12.756690
Referral	0	0.185841	11.179558
	1	0.240000	11.518080

Channels of Acquisition



- Model Free- Evidence:
- **By comparing different sub-groups of the population across treatment and control we can learn which groups are most effected by the intervention**

	Acquired via Instagram		
	Control	Coupon	Difference
Trans_after	0.15	0.20	6%
Revenues_after	9.17	10.34	1.17
Total	1552		

	Acquired via Google		
	Control	Coupon	Difference
Trans_after	0.08	0.06	-2%
Revenues_after	4.92	2.80	-2.12
Tot Customers	2035		

```
#regression with interaction effects channels

model = smf.ols(formula='trans_after ~ test_coupon + num_past_purch + spent_last_purchase + weeks_since_visit + browsing_minutes+ shopping_cart + channel

print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            trans_after   R-squared:                       0.179
Model:                            OLS   Adj. R-squared:                  0.177
Method:                 Least Squares   F-statistic:                     77.66
Date:                Tue, 24 Sep 2024   Prob (F-statistic):          2.96e-201
Time:                        02:07:24   Log-Likelihood:                -2257.3
No. Observations:                5000   AIC:                             4545.
Df Residuals:                    4985   BIC:                             4642.
Df Model:                          14
Covariance Type:            nonrobust
==================================================================================================
                                     coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------------
Intercept                         -0.0461      0.019     -2.379      0.017      -0.084      -0.008
test_coupon                       -0.0112      0.017     -0.663      0.507      -0.044       0.022
num_past_purch                     0.0571      0.002     24.613      0.000       0.053       0.062
spent_last_purchase               -0.0004      0.000     -3.434      0.001      -0.001      -0.000
weeks_since_visit                 -0.0203      0.002     -8.527      0.000      -0.025      -0.016
browsing_minutes                   0.0030      0.001      3.859      0.000       0.001       0.005
shopping_cart                      0.1723      0.012     14.374      0.000       0.149       0.196
channel_Facebook                   0.0870      0.021      4.232      0.000       0.047       0.127
channel_Instagram                  0.0724      0.018      4.005      0.000       0.037       0.108
channel_Referral                   0.0926      0.038      2.452      0.014       0.019       0.167
channel_Other                      0.1627      0.051      3.166      0.002       0.062       0.263
test_coupon:channel_Facebook       0.0521      0.029      1.802      0.072      -0.005       0.109
test_coupon:channel_Instagram      0.0726      0.026      2.828      0.005       0.022       0.123
test_coupon:channel_Referral       0.0724      0.052      1.386      0.166      -0.030       0.175
test_coupon:channel_Other         -0.0076      0.073     -0.104      0.917      -0.150       0.135
==============================================================================
Omnibus:                     3189.104   Durbin-Watson:                   1.960
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            45914.220
Skew:                           2.847   Prob(JB):                         0.00
Kurtosis:                      16.710   Cond. No.                     1.26e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.26e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

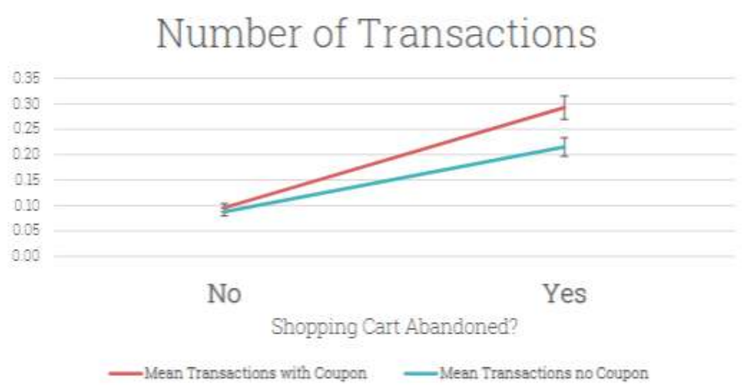
```
model = smf.ols(formula='revenue_after ~ test_coupon + num_past_purch + spent_last_purchase + weeks_since_visit + browsing_minutes+ shopping_cart + chan

print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           revenue_after   R-squared:                       0.172
Model:                             OLS   Adj. R-squared:                  0.170
Method:                  Least Squares   F-statistic:                     73.94
Date:                 Tue, 24 Sep 2024   Prob (F-statistic):          4.96e-192
Time:                         02:07:32   Log-Likelihood:                -22435.
No. Observations:                 5000   AIC:                         4.490e+04
Df Residuals:                     4985   BIC:                         4.500e+04
Df Model:                           14
Covariance Type:             nonrobust
===================================================================================================
                                      coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------------
Intercept                          -2.1951      1.095     -2.004      0.045      -4.343      -0.048
test_coupon                        -1.6230      0.955     -1.699      0.089      -3.495       0.249
num_past_purch                      3.1716      0.131     24.185      0.000       2.915       3.429
spent_last_purchase                -0.0144      0.006     -2.378      0.017      -0.026      -0.003
weeks_since_visit                  -1.1121      0.135     -8.241      0.000      -1.377      -0.848
browsing_minutes                    0.1586      0.045      3.551      0.000       0.071       0.246
shopping_cart                       9.2953      0.678     13.706      0.000       7.966      10.625
channel_Facebook                    5.2746      1.163      4.536      0.000       2.995       7.554
channel_Instagram                   4.5153      1.023      4.413      0.000       2.510       6.521
channel_Referral                    5.4818      2.136      2.566      0.010       1.294       9.669
channel_Other                       9.0358      2.906      3.109      0.002       3.338      14.734
test_coupon:channel_Facebook        1.5461      1.634      0.946      0.344      -1.657       4.749
test_coupon:channel_Instagram       2.9593      1.452      2.039      0.042       0.114       5.805
test_coupon:channel_Referral        2.3094      2.954      0.782      0.434      -3.481       8.100
test_coupon:channel_Other          -0.6657      4.112     -0.162      0.871      -8.726       7.395
==============================================================================
Omnibus:                     3307.920   Durbin-Watson:                   1.961
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            48295.917
Skew:                           2.995   Prob(JB):                         0.00
Kurtosis:                      16.998   Cond. No.                     1.26e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.26e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

By including in the regression interaction effects (i.e. 0 1 variables that identify whether the customer was acquired via a specific channel & received the coupon) we can estimate the differential impact.

To whom should the coupon be sent?

	Y= Number of transactions after the field test (trans_after)		Y= revenues after field test (revenues_after)	
	Coefficients	P-value	Coefficients	P-value
Intercept	-0.05	0.02	-2.20	0.05
test_coupon	-0.01	0.51	-1.62	0.09
num_past_purch	0.06	0.00	3.17	0.00
spent_last_purchase	0.00	0.00	-0.01	0.02
weeks_since_visit	-0.02	0.00	-1.11	0.00
browsing_minutes	0.00	0.00	0.16	0.00
shopping_cart	0.17	0.00	9.30	0.00
channel_Facebook	0.09	0.00	5.27	0.00
channel_Instagram	0.07	0.00	4.52	0.00
channel_Referral	0.09	0.01	5.48	0.01
channel_Other	0.16	0.00	9.04	0.00
Facebook & Coupon	0.05	0.07	1.55	0.34
Instagram & Coupon	0.07	0.00	2.96	0.04
Referral & Coupon	0.07	0.17	2.31	0.43
Other & Coupon	-0.01	0.92	-0.67	0.87



```
# INTERACTION EFFECTS WITH SHOPPING CART
formula = 'trans_after ~ test_coupon + num_past_purch + spent_last_purchase + weeks_since_visit + browsing_minutes + channel_Facebook + channel_Instagra
ols_model = smf.ols(formula, data=db_artea).fit()
print(ols_model.summary2())
```

```
                 Results: Ordinary least squares
==================================================================
Model:              OLS              Adj. R-squared:     0.178
Dependent Variable: trans_after      AIC:                4535.4102
Date:               2024-09-24 02:07 BIC:                4613.6165
No. Observations:   5000             Log-Likelihood:     -2255.7
Df Model:           11               F-statistic:        99.26
Df Residuals:       4988             Prob (F-statistic): 6.45e-205
R-squared:          0.180            Scale:              0.14469
------------------------------------------------------------------
                          Coef.  Std.Err.    t    P>|t|  [0.025  0.975]
------------------------------------------------------------------
Intercept                -0.0529  0.0185 -2.8553 0.0043 -0.0892 -0.0166
test_coupon               0.0009  0.0128  0.0687 0.9452 -0.0242  0.0260
num_past_purch            0.0572  0.0023 24.7021 0.0000  0.0527  0.0618
spent_last_purchase      -0.0004  0.0001 -3.3978 0.0007 -0.0006 -0.0002
weeks_since_visit        -0.0206  0.0024 -8.6255 0.0000 -0.0252 -0.0159
browsing_minutes          0.0032  0.0008  3.9951 0.0001  0.0016  0.0047
channel_Facebook          0.1116  0.0144  7.7301 0.0000  0.0833  0.1398
channel_Instagram         0.1080  0.0129  8.3752 0.0000  0.0827  0.1333
channel_Referral          0.1299  0.0261  4.9785 0.0000  0.0788  0.1811
channel_Other             0.1590  0.0363  4.3787 0.0000  0.0878  0.2302
shopping_cart             0.1312  0.0167  7.8593 0.0000  0.0985  0.1640
shopping_cart:test_coupon 0.0846  0.0237  3.5716 0.0004  0.0381  0.1310
------------------------------------------------------------------
Omnibus:            3183.367       Durbin-Watson:         1.958
Prob(Omnibus):      0.000          Jarque-Bera (JB):      45608.537
Skew:               2.842          Prob(JB):              0.000
Kurtosis:           16.661         Condition No.:         549
==================================================================
```

Customer development & retention

We are in the middle phase of the customer journey.

Wrap-up: Artea

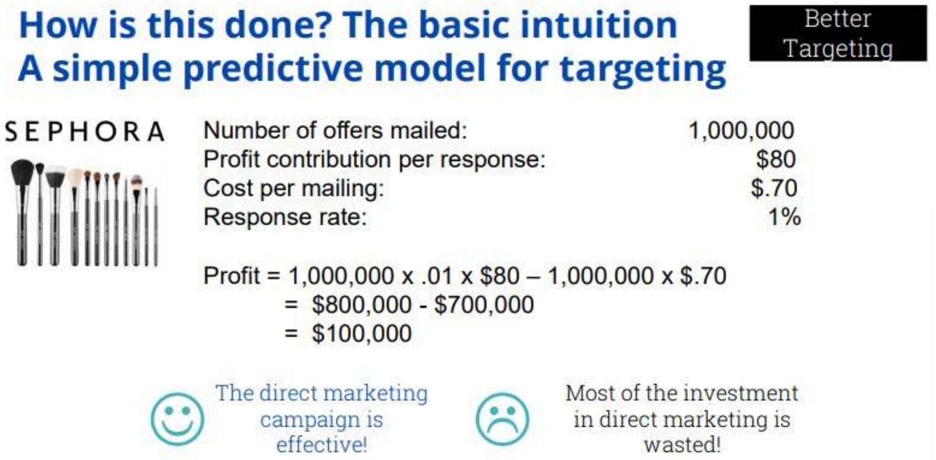
The purpose of the Artea case was to focusing on the beginning of the journey in order to find a strategy to acquire new customers who are not engaged with our firm.
What conclusions have you drawn from this case?

- Customer Acquisition -> find a way to increase acquisition and the number of new customers. In this case they implemented discounts (new adv campaign, pricing strategy, testimonial could be other ideas).

- Randomized Field Test can be used to evaluate marketing effectiveness -> They were creating two groups, one (Experimental group) was associated with the discount the other one (Control group) no, in order to understand the number of new customers in both groups and so verify is the strategy proposed was effective or not.

- Strategies for managing customer heterogeneity (diversity of the individuals) -> overall the discount, which was the strategy implemented to acquire new customers, but what we want to verify here is whether the strategy is significant for all the customers or not. Usually anything fit perfectly to everybody, so that is why we have to divide the dataset in groups in order to verify for which groups the strategy is effective basing on the groups' characteristics => BETTER TARGETING:

it is important to managing customer heterogeneity because the strategy is never successful for everybody, so we need to identify groups with different response to the discounts, to do so we have to run in Python interactive variables. The main goal is to better customize the experience for the customers according to their characteristics, behavior and expectat



The campaign has positive profits, and it is effective, but at the same time, in terms of marketing investments is not profitable because the amount of money spent is to high.
So how can we do better? -> if we are able to find those individuals who responds more and targetedthem, we will be able to reduce costs, while still increasing profits.

Targeting analysis: lift-based approach

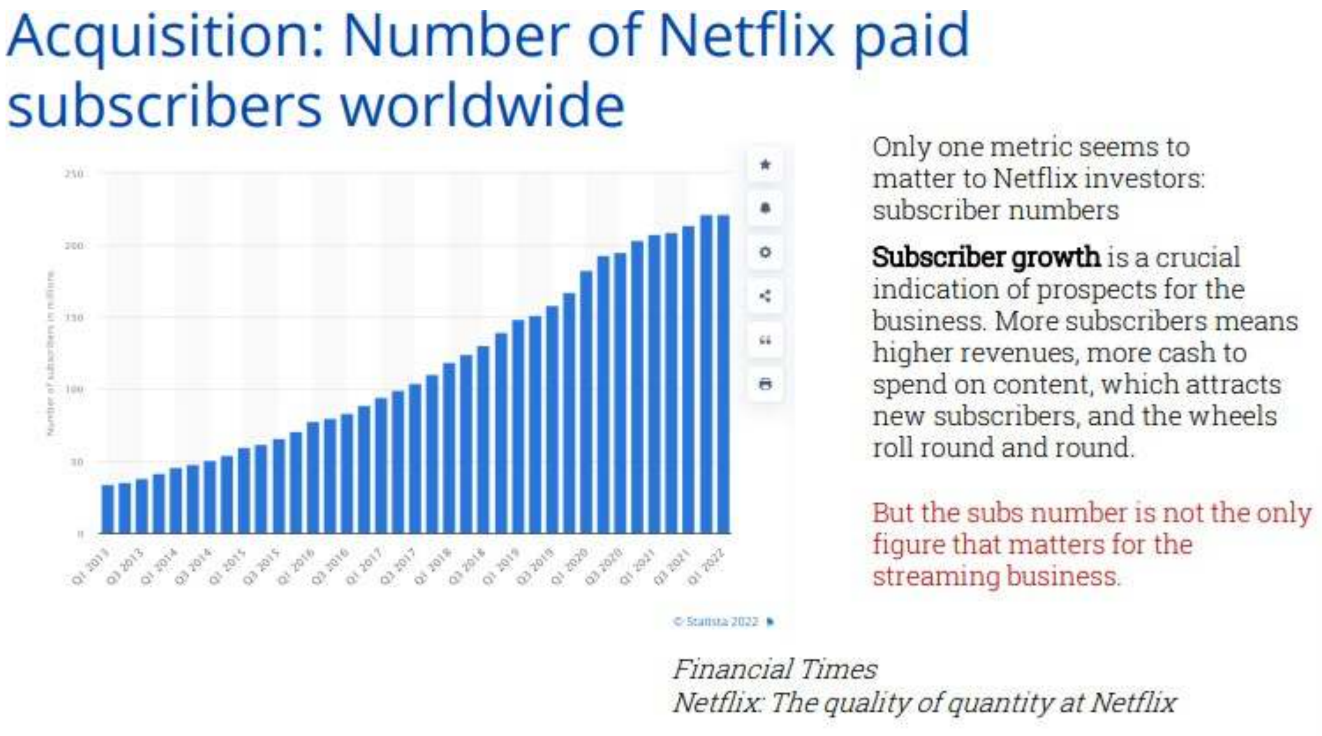
170000= (3%*100000*80)-(100000*0.7)

Decile	Number of Prospects	PredictResp Rate	Profit	Cumulative Profit
1	100,000	3.00%	$170,000	$170,000
2	100,000	2.00%	$ 90,000	$ 260,000
3	100,000	1.40%	$ 42,000	$ 302,000
4	100,000	1.15%	$ 22,000	$ 324,000
5	100,000	1.00%	$ 10,000	$ 334,000
6	100,000	0.60%	$ -22,000	$ 312,000
7	100,000	0.40%	$ -38,000	$ 274,000
8	100,000	0.30%	$ -46,000	$ 228,000
9	100,000	0.10%	$ -62,000	$ 166,000
10	100,000	0.05%	$ -66,000	$ 100,000

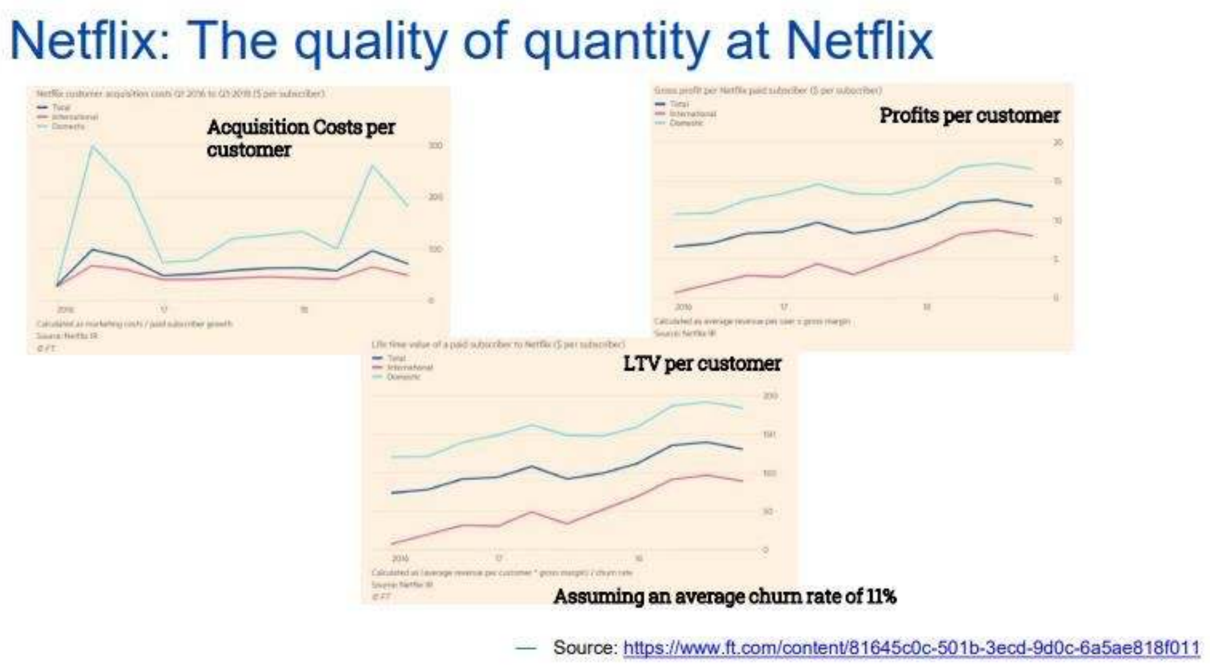
Better Targeting

→Profits Improvement → $100,000→ $334,000

From group 6 we start losing money, so if we stop at group 5 the cumulative profit is so much better than before simple because we are saving money while gaining 334,000$ more.



The company is advertising a lot, particularly to the investors who are growing in terms of numbers of subscribers.
But then they focused on a different element: "*Is this enough*?" Actually no! -> Acquisition is not the only important metric to take in account, for example when we talk about Netflix's subscriber we will not only want to know how many new customers we are to acquire but also the average profits provided by each of the new costumers is very important.



To acquire a new customer, they spend a lot of money.
Has we can see the profits for customer are not that great… they are declining.
Customer lifetime value = is a metric that takes into account not only the profit but also the likelihood of

customer to stay with the company alongside -> it is a very important metric, a key performance indicator able to tell the manager of a firm how much each customer that we have acquired is going to value for us in the future. As we can see also this metric has a declining trend.

We can also combine these two metrics to create a new metric, which is the ratio between the cost of acquisition and the customer lifetime value

What does this graph tell us?

Source: FT, The quality of quantity at Netflix, October 2018

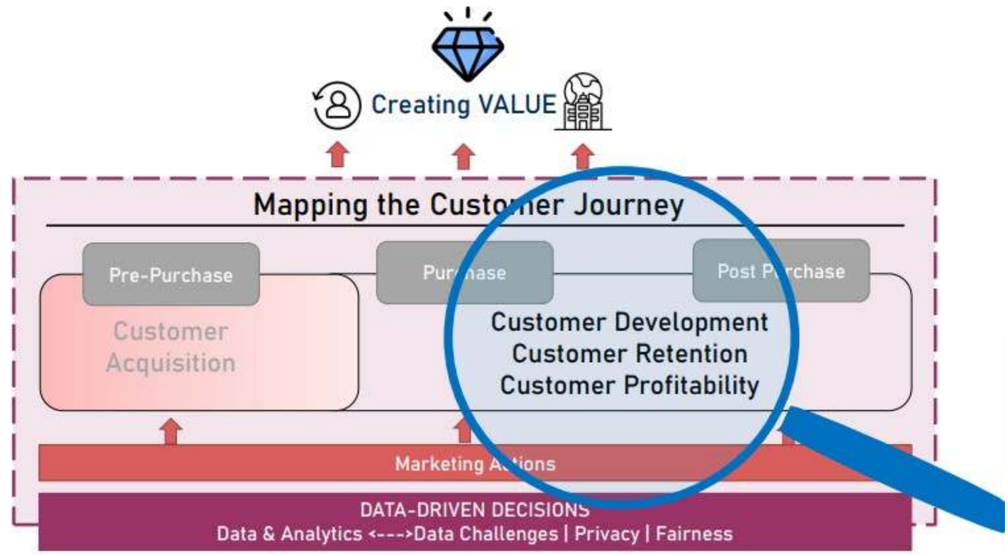
This value is pretty stable, but definitely not that great as the acquisition rate is increasing.

Acquisition & Retention: Key Metrics for Business Health

We need to acquire new customer if we want to grow, but at the same time we want to make sure that we are able to retain those customers -> is less expensive for a firm to work on the current customer base.

Low retention rates + high customer acquisition costs = significant financial impact

- In key sectors like telecom, TV streaming, finance/insurance, and health, customer retention is paramount
- Annual retention rates in wireless phone providers typically range from 70% to 85% (which is not super: the company loses, at the end of each year, 30% of their customer. This means that they have to acquire at least the same amount of customer that decide to leave the business. But even if the are able to acquire the same amount loss is not good, because acquire new customer is more expensive than retain current customer base)



CUSTOMER DEVELOPMENT: What do we mean?

Creating value for both the customer and the organization (increasing customer lifetime value): How?

> Customer lifetime value must increase -> how profitable is each customer for us

> In order to increase customer lifetime value we have to create value for the customer means that we are in a good shape in terms of retention -> if customer are satisfied, even the firm will be satisfy, so it is a win-win situation

CUSTOMER DEVELOPMENT: WHAT DO WE MEAN?

- Increase Customer Lifetime Value: aka CLV or LTV (lifetime value of a customer)
- Different ways to compute it:

The CLV calculation is an estimate of the expected value of the customer's value over time

$$\text{CLV} = \sum_{t=1}^{\infty} \frac{m_t r^{t-1}}{(1+\delta)^{t-1}}$$

m=Revenues - Costs
d=Discount rate
r= Retention
c=1-r=Churn

$$\text{CLV} = \sum_{t=0}^{\infty} \frac{m(1-c)^{t-1}}{(1+\delta)^{t-1}} = \frac{m(1+\delta)}{(\delta+c)}$$

churn

It's a forward-looking concept that focuses on the future instead of the past to evaluate or measure a customer's profitability.

$$CLV = \frac{m(1+d)}{1+d-r}$$

Blattberg et al (2008)

First, define each component of the metric:
Margin (m) = revenues – cost (include all kind of costs, so both fixed and variable one, as well as both production costs and marketing costs linked to the product)
Delta (d) = discount of rate (financial concept -> if I have this money now which will be their value inthe future -> time value of the money)
Retention rate (r) = retention (a value always between 0 and 1 -> likelihood for each consumer to beretained)
Churn rate (c) = 1-r = churn (likelihood for each consumer to live the business in the next year)
Time (t) = years

Some assumptions:

- margin must be constant so everything can be simplified in

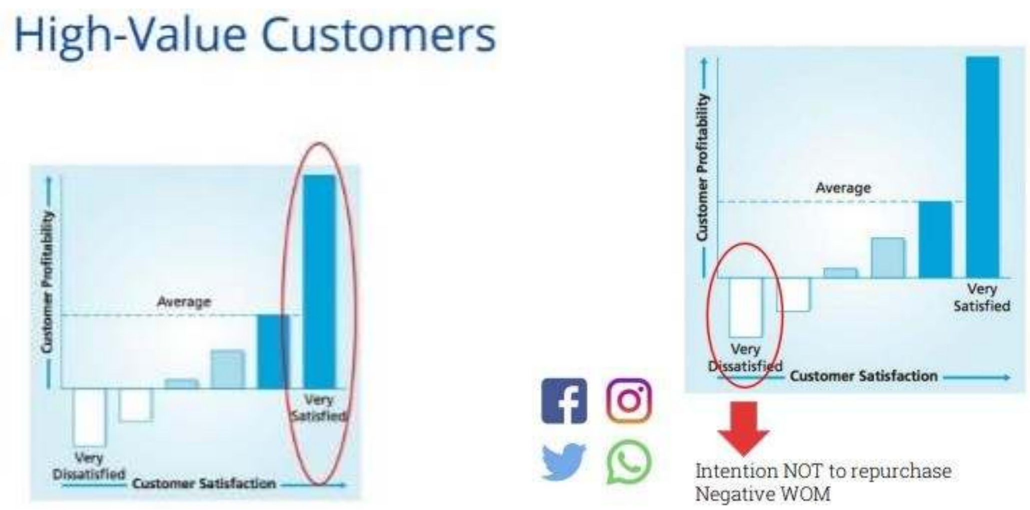
$$CLV = \frac{m(1+d)}{1+d-r}$$

We can impact positively the customer lifetime value by:

- **better targeting**: we know that the same marketing strategy can be better and more effectivefor certain customers rather than the others -> if we are able to cut marketing costs in a cleverway, we can *reduce cost in the parameter m* of customer lifetime value
- **increase customer satisfaction**: positive impact, even though it is a indirect effect, on the likelihood to be retained
- **increase frequency of purchase**: if customers are spending more, we can increase the spending component (m) of the customer lifetime value thanks to higher revenues
- **increase the volume of each purchase occasion** (instead of spending 100 spend now 200)
- **share of wallet**: increase the portion of purchases for our brand while reducing, in doing so, the amount spent at our competitors

We can try to hand up with the strategy that increase **upselling** and **cross selling**

Cross-selling: I have an iPhone but not a Mac; if the company can induce me to buy also the laptop this is a cross selling
Upselling is inducing the customer to buy a better version of the same item (iPhone 15 -> iPhone 16)



Very satisfied customers are more profitable customers, while low satisfied customers are not only less profitable, but they can also negatively impact on the brand by doing bad reviews, and so on.

High-value customers = Satisfied Customers

Customer Satisfaction	Customer Percent	CSI Score	Customer Revenue	Percent Margin	Gross Profit	Retention Cost	Customer Profit
Very Satisfied	25%	100	$1,200	60%	$720	$100	$620
Satisfied	35%	80	$800	50%	$400	$100	$300
Somewhat Satisfied	20%	60	$300	40%	$120	$100	$20
Somewhat Dissatisfied	15%	40	$80	40%	$32	$100	–$68
Dissatisfied	3%	20	$60	40%	$24	$100	–$76
Very Dissatisfied	2%	0	$50	40%	$20	$100	–$80
	100%	72	$655	49%	$350	$100	$250

Satisfaction & Customer Development→ Increase Value of the Customer

High-value customers = Satisfied Customers

On average acquiring new customers costs more than retaining them

N.B. Satisfaction is not equal to retention, so we will always lose also satisfied customer.

Dissatisfied customers who complain can be retain if we have a great customer service, but if they do not complain is so much harder for the company to implement strategies to retain them -> they stay only if searching and switching cost are too high for them to change business.
If we have customer that do complain that means that we still have possibilities to do better and fix these problem-> **IMPLEMENT A STRONG SERVICE RECOVERY:** give the customers, the possibility and the private space to complain, without using the web in order to then implement the best strategies to fix the problems raised.

RETENTION OR CHURN

Customer Lifetime=1/(1-RetentionRATE)

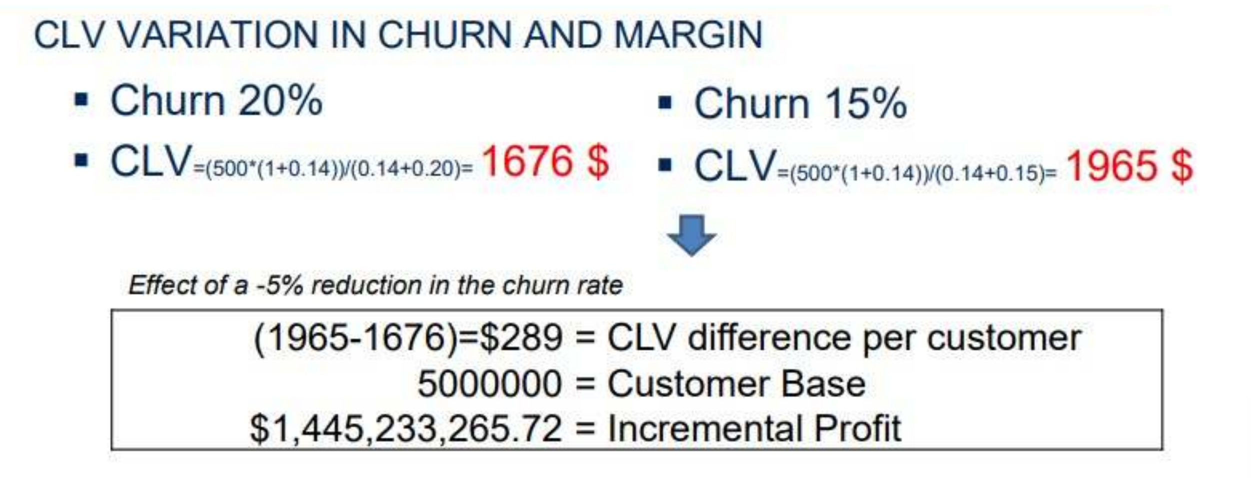
Churn Rate=1-Retention Rate

Year	Customer Retention	Customer Life
2002	72%	3.6
2003	77%	4.3
2004	80%	5.0
2005	82%	5.6
2006	85%	6.7

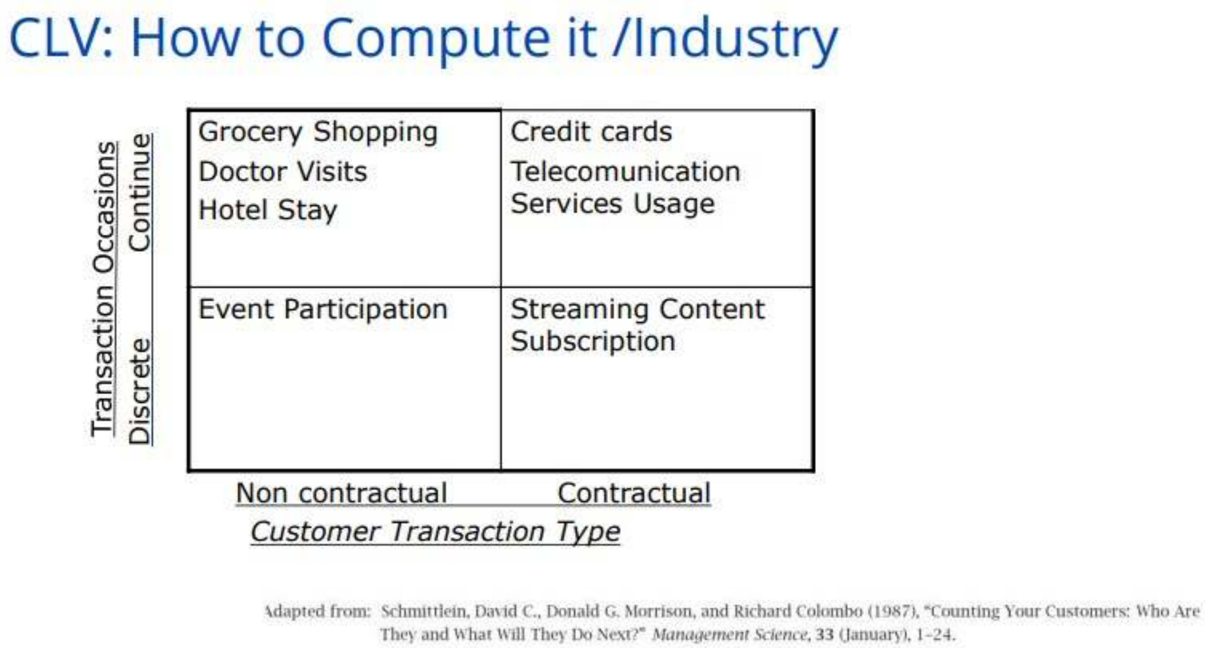
CLV & CHURN: CONVEX FUNCTION

$$CLV = \frac{m(1+d)}{1+d-r}$$

This function is convex, which means that if we are able to reduce the churn rate we can gain much more in terms of customer lifetime value -> A small increase in terms or retention is going to produce a great increase in terms of lifetime value.



The difference in terms of CLV is small when reducing the churn rate, if we consider the single customer, but when we compute the total incremental profit, we can see a huge improvement, even tough the churn rate reduction was really small.



To compute CLV we necessarily need the **churn rate** or otherwise the **retention rate**.

So, if we are in a contractual business, it is pretty easy to compute it, because we know the number of contracts each year -> at the end of each year the company perfectly know the number of customers that they were able to retain.
This is not true in not-contractual businesses because without a contract is difficult for the businessto understand and figure out when the customer leaves the business -> in this case the measure of CLV is not of course a precise value.

CUSTOMER DEVELOPMENT

e.g. Up-selling, Cross-Selling, Frequency of purchase, SOW, Customer Profitability

$$CLV \; LTV = \frac{m(1+d)}{1+d-r}$$

Reduce Churn, Work on Retention & Loyalty

Better targeting is another strategy that can reduce the costs and so increase *m*. Different strategies can have different impact on some key component of the CLV.

- How do companies react to having a 'churn problem' or a problematic LTV/CAC ratio

Churn Management

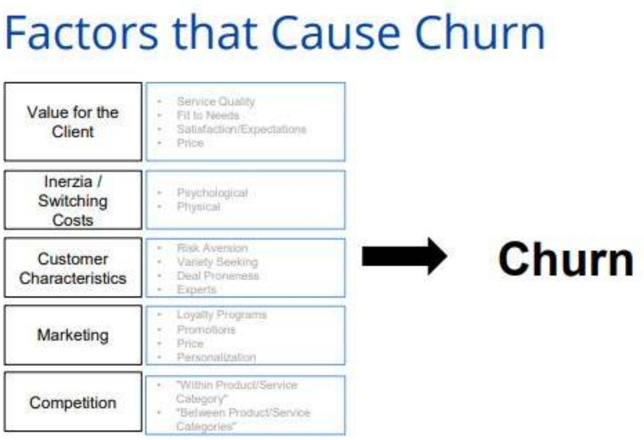
Objective: Focus on the 'Retention' Component of LTV
Unit of Analysis: Customer

- $Churn_i$ = Probability that an individual *i* will leave the company in a given period (ex. The end of the year)

Unit of Analysis: Firm

- Churn= percentage of the customer base that leaves the company in a given period
- Churn=c=1-Retention Rate=1-r

What leads to churn? What are the contributing factors?



Value for the client

- Service quality
- Fit to needs
- Satisfaction/expectations
- Price
- if services/products have a value for me, if I'm satisfied with them, if I find the price in line with my expectation about it.

Inertia/switching cost:

- physical) -> sometimes we have problems to change with a positive impact on retention
- psychological -> inertia

Competition:

- within product/service category
- between product/service categories
- the more competitors we have in a market, the more difficult, even for the best firm, is to retain customers

Marketing:

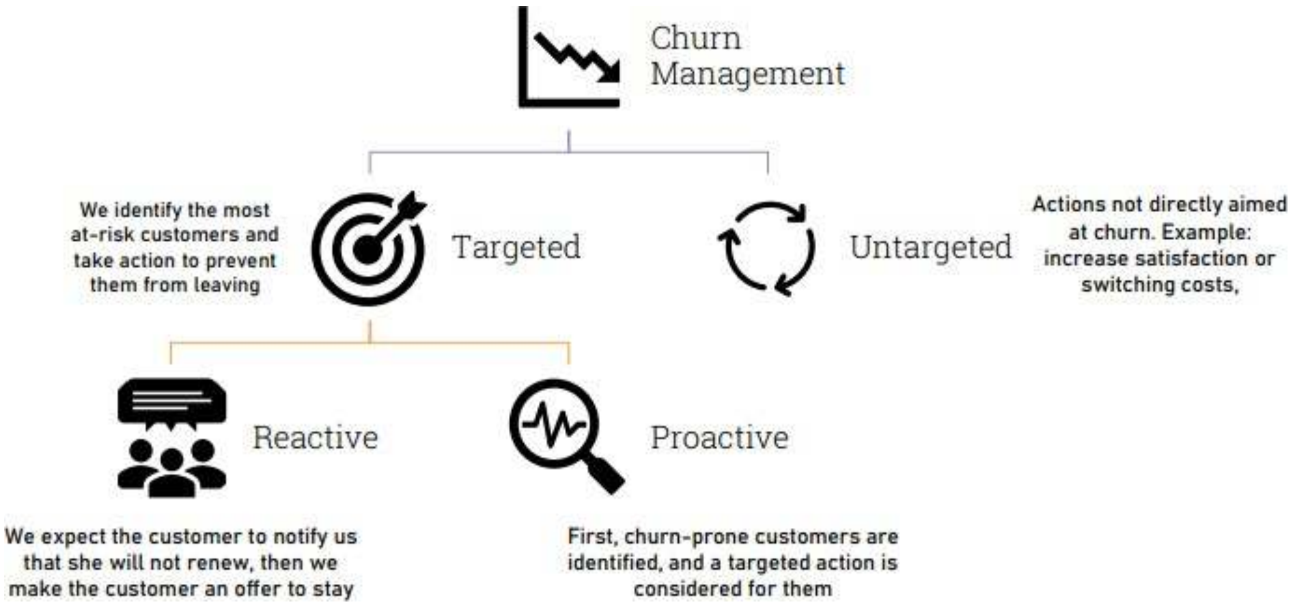
- loyalty
- promotions
- price
- personalization
- if the marketing actions are effective, they have a negative impact on churn by increasing retention.

Customer characteristic:

- risk aversion
- variety seeking
- deal proneness
- experts
    - if we understand that a group is more likely to churn, we have to understand how to acquire them bringing them in our business.

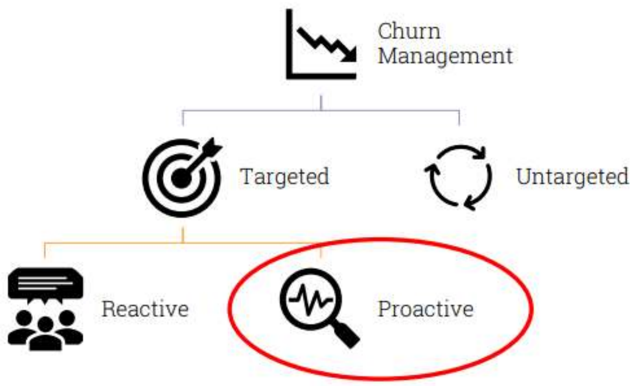
These are all the macro factor that we know for sure have an impact on retention. So as a strategy we can decide to focus on one or more of them.

How to Manage Churn?



There are two macro strategy that we can adopt:

1. Hand up with churn management **untargeted strategy** -> the key idea if that action are not directly aimed at churn -> I identify an action to do not directly aimed to reduce churn, but to increase something else to have a positive impact on churn (ex. Increase customer satisfaction because if I increase that indirectly I hope to have a positive in reducing churn; increase switching cost to indirectly have a positive impact on churn)
2. Hand up with churn management **targeted strategy** -> completely different mindset becausethe first purpose is to identify the consumers that are at risk of churn. We effectively create a strategy to reduce the risk of churn applying it to only those customers at risk of churn. I have two option to react:
    a. *Targeted reactive*: we don't do anything until the customers notify us that they want to leave, and in that moment, we apply our strategy -> ex. Sky
    b. *Targeted proactive*: I have a metric, so I know who at rick on churn, but before they contact me, I'm gonna do something because I fell that waiting is risking.



They key step is we want to apply this strategy is to identify **in advance** those who are at risk of churnbyusing a **predictive model**:

1. Predictive model: goal is to predict individual-level churn
2. Proactive churn management actions by using creative strategies

The 'process' of developing a predictive model

The development of a predictive model is primarily a **process** (here is where we can demonstrate our ability as manager), which extends beyond mere model estimation and consists of **several phases**:

a. problem definition
b. data preparation
c. model estimation
d. model evaluation
e. verification of predictive capability...
    - What is the **key question** we want to answer?
    - What **data** do I need to make these predictions? -> Now we have a dataset provided by the teachers but maybe it will not be the case in the future and have to provide evidence that the data the we want to buy are crucial for the purpose of retain customers
    - What **statistical technique** should I use?
    - Does it work/**Do I trust the model**? -> Develop a predictive model the works well is fundamental
    - Does it have any **managerial relevance**? Can I use the results?"

1. **DEFINING THE RPOBLEM**
    a. Defining the managerial problem
2. **PREPARING THE DATA**
    a. Identifying the behavior ton predict
    b. Constructing the dataset
    c. Preprocessing the data
3. **ESTIMATING THE MODEL**
    a. Selecting predictors
    b. Choosing the modeling technique
    c. Estimating the model
    d. Evaluating the model
4. **PREDCTIVE VALIDITY**
    a. Calibration and Validation
5. **TARGETING**
    a. Lift (e.g.>1) cutoff
    b. Individual scores

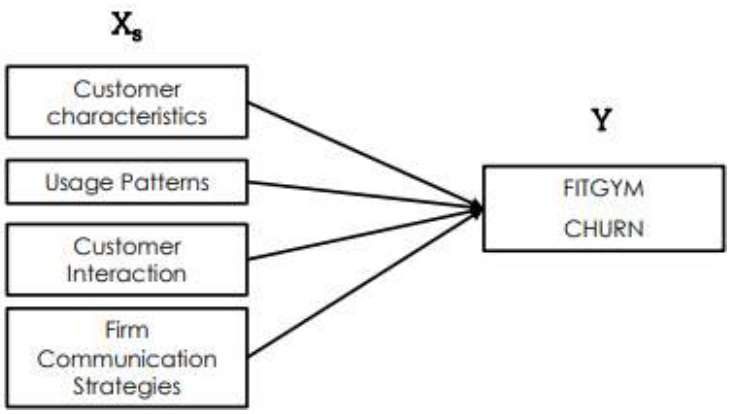
EXAMPLE - FitLife Gym predictive churn model: PREPARING THE DATA

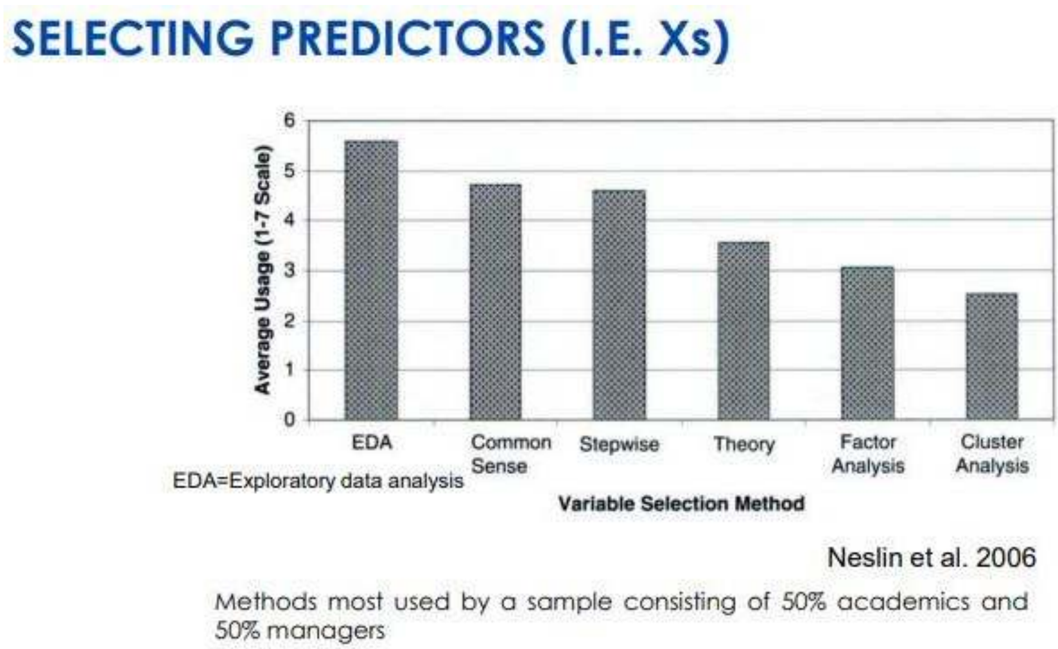
We have to prepare a predictive model for FitLine Gym, which means list all the data that we think could be useful to predict the risk of churn in this industry and for this firm.

1) **Usage Patterns**: Customers who haven't visited the gym in the past month and have a declining trend in their monthly attendance might be at risk of churning.
2) **Customer Demographics**: Young professionals or students might be more likely to churn due to changing schedules, location, or financial constraints (we can get these data by the subscription that they fill to subscribe to the gym).
3) **Customer Interaction**: Customers who have raised complaints or concerns through the gym's feedback system multiple times without satisfactory resolution might be dissatisfied and consideringleaving.
4) **Member Engagement**: Assessing the level of engagement, such as class attendance, participationin fitness challenges, or app usage for workout tracking, can provide insights into a member's likelihood to continue or churn
5) **Communication**: Evaluating the effectiveness of communication channels (e.g., email, SMS, app notifications) in engaging and retaining customers.
6) **Referral Source**: Analyzing how members were acquired (e.g., referral, marketing campaign, walk-ins). Members referred by existing satisfied customers may have lower churn rates

At the end of this phase, we will be able to create the model.

Question: Which factors can help predict churn?

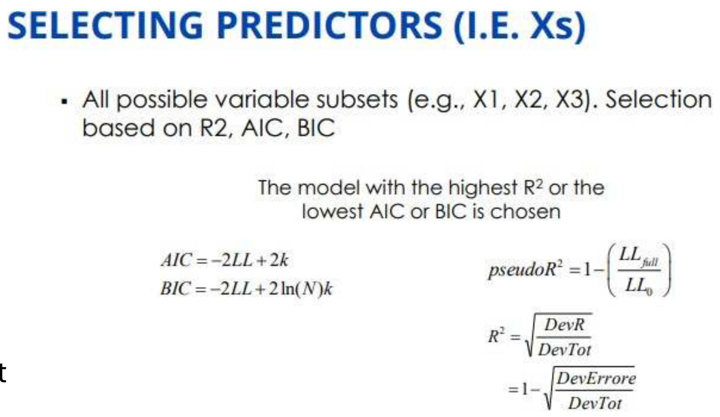




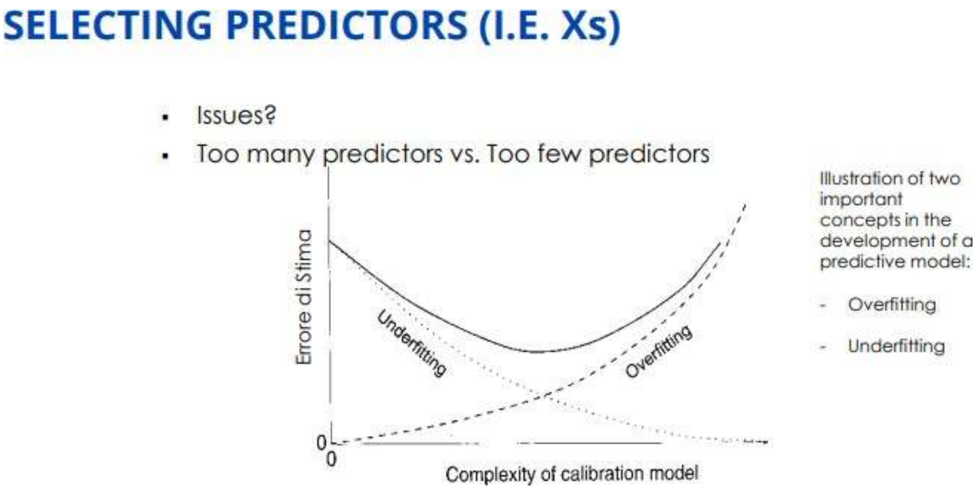
How do I select talking about the independent variables in my prediction model?
Both **statistical theory** and **intuition** could be use in order to select the Xs of the model.

Example: Among the independent variable we have *age*, but the only values registered in the dataset are 25-26 -> it is basically a constant, so I don't have to add this variable in my model.
Or again we have an apparently interesting information about the number of individuals acquired through referral (1% of the sample. It means that we have a dummy variable that is always 0 except for a tiny portion of 1 -> it is basically constant; it seems interesting but because it does not vary it is better not include them.
We might exclude some variable that don't change very much in order to have a better model with just relevant information.
Common sense could use as well -> I think that maybe some variables are useful, so I add them in the model.
Moreover, we have theory that can help us in choosing our variables.
When we have a huge number of Xs is typical to have problems when including all of them in the model, especially the multicollinearity one due to the possible correlation among variables ->
sometimes people want to reduce the dimension by creating common factors or **cluster analysis**.
Stepwise is another possible approach: the regression is done step by step, introducing all the variable in the list step-by-step trying to understand the best model with only significant variables by analysing the R^2 every time -> the teacher is skeptical to use this model as a first approach; if adopting this model we need to pay attention in order to add only what is significant without skiping information that could be us

Typically, we check the R^2 and then we select the solution with the varibale which optimizes it.
Remember that when we use Logit we have just a pseudo R^2 and not the real value associated to this parameter.
The value of the R^2 tent to increase when we have a lot of variables, and must be biased which means that have to many variables is not always good for our model, not that our model explains better the situation.



To avoid the biased R^2 that we can get when we have to many varibales we use AIC and BIC. LL is the maximum likelihood calculated through the Logit function.



Underfitting: not enough varibale in order to fit the model.
Overfitting: we include in the model too many varibales. The model runs, but it is not significant; some varibale are correlated to each other, or some of them are too similar.

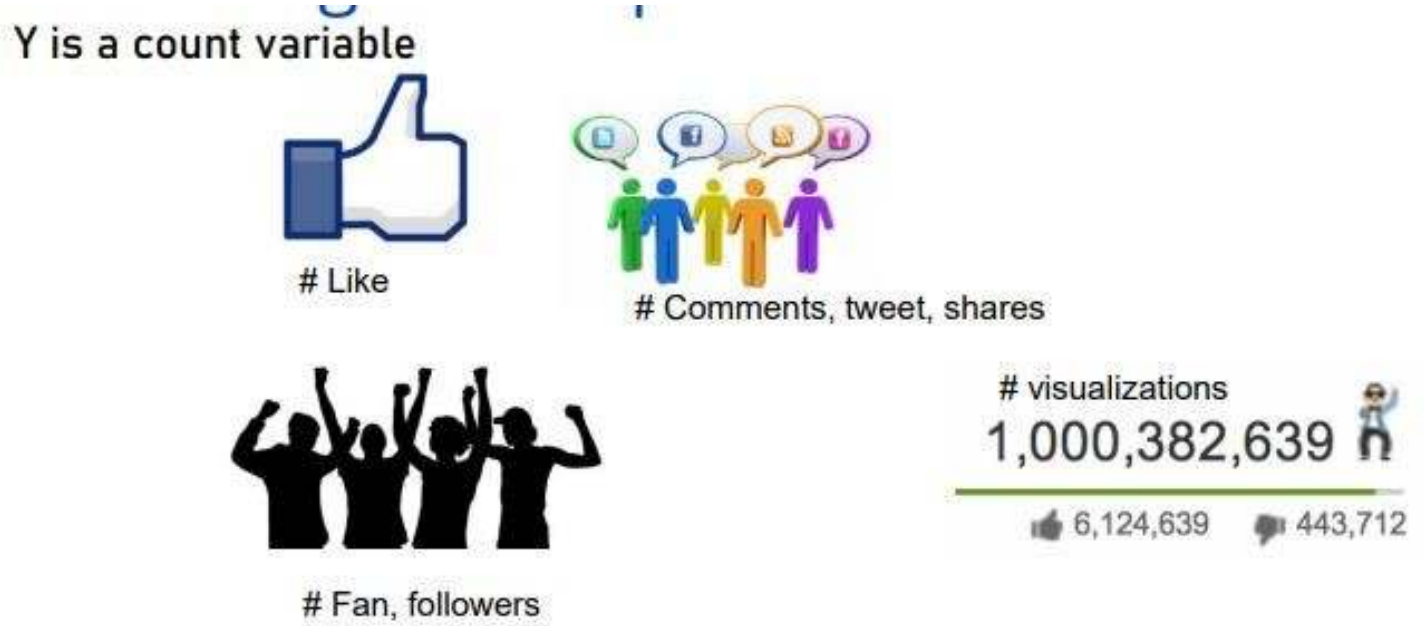
Estimating the model: choosing the modelling technique

We can choose among different possibilities:

- OLS REGRESSION
- Logit/Probit
- MNL/MNP/Conditional
- Multivariate Logit/Probit
- Tobit (I, II)
- Hazard—Poisson, NBD

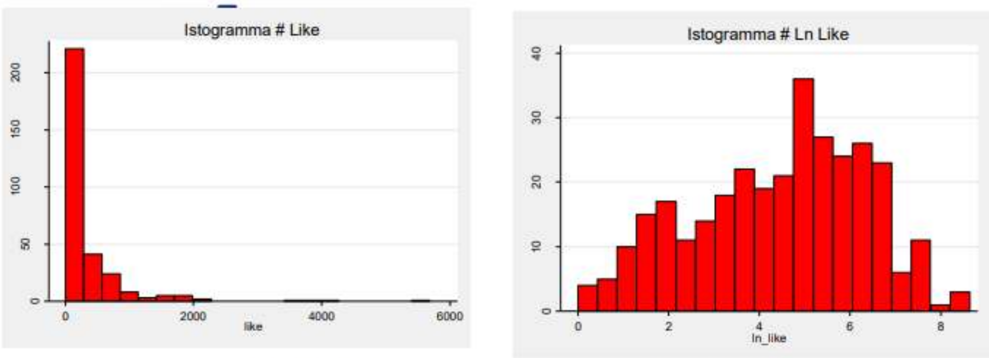
EVERYTHING (or almost everything) DEPENDS ON Y! -> We have to understand which type of variable is Y in order to understand which is the right model to use

We will use an *OLS* if the data are *normally distributed*.



In this case the best model to use could be *OLS* after a transformation of the data, or otherwise use *more advance models*.

Possible alternative when the dependent variable is a count variable



Advanced: Poisson, NBD, ZIP, ZINB

Ln(Like) OLS Regression

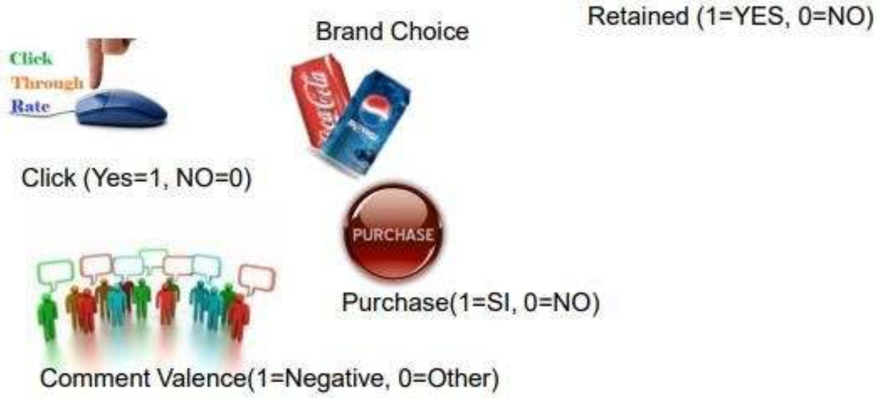
In this case where 0 and 1 are just labels, we can only apply a *logit model*.

⇨ Y tell us which model is the right one to use

Calibration & Validation

Our purpose is to develop a strategy to manage churn, more specifically targeted proactive which means identify in advance those that are at risk of churn -> we need to predict who is more likely to churn, so here the task is to create a great predictive model, and we need to test the predictability of the model.

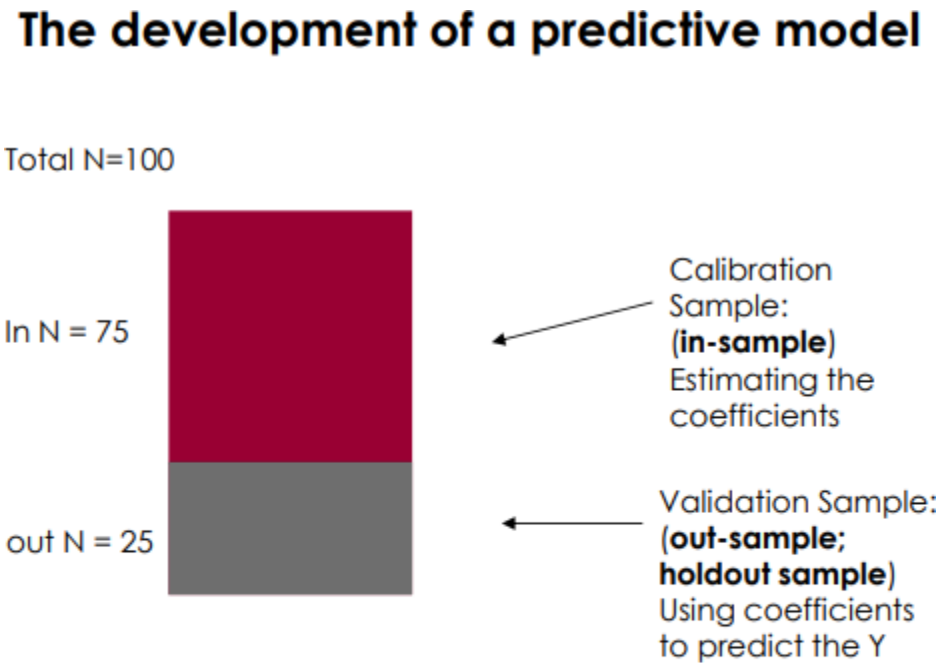
The dependent variable is a qualitative nominal variable with 2 categories



Models: Logit / Probit

How to do it?
Typically, we have a dataset; imagine a dataset of 100 customers. Of course, whatever we have in the dataset happened in the past which means that I can use previous data to develop a predictive model that maybe is able to predict the behavior of the customer this year. To do this I have to take the data, divide the dataset into two sub-samples -> one to estimate the model and the other one to check if the prediction model predicts well:
A – a «**calibration**» sample also known as in-sample

B – a «**validation**» sample also known as out-sample, on which testing the model



Step 1 - In-sample (Calibration): estimate coefficients (using only 75 individuals and not 100): $Y_{INi}=a+b_1X1_{INi}+b_2X2_{INi}+b_3X3_{INi}+e_{INi}$

Step 2 – Out-sample (Validation): using obtained parameters (a, b_1, b_2, b_3) at step 1 to predict Y_hat$_{OUTi}$ (the predicted y)
Step 3 – compare the estimated values with the observed ones and calculate the prediction error as=(Y_{OUTi} -$\hat{Y}_{OUTi}$)

Before doing this process there are to question to answer:

1. What should be the sample size?
   - the calibration sample must be higher than the validation one if the model is lower than 500 observations -> *Calibration sample = 75%* of the total number of the observations, while the *validation sample = 25%* of the observation
   - Otherwise, if the model has a large number of observations the calibration sample size must be select randomly because we want to create to groups that are identical in every dimension -> *Calibration sample= validation sample= 50%*
2. How much of the sample should be allocated to calibration?

- N< 500 (sample size is relatively small) calibration sample =75%, validation sample= 25%. Why?
- For sizes >=500, any 'reasonable' split works well (Steckel and Vanhonacker 1993).

How to choose the validation and calibration samples? -> RANDOM

Calibration and Validation: EXAMPLE

```
                                                                N=100
Y=_const+bX     if CALIBRATION==1

      Source |       SS       df       MS              Number of obs =      75
-------------+------------------------------           F(  1,    73) =  184.02
       Model |    740.7946     1    740.7946           Prob > F      =  0.0000
    Residual |  293.872067    73  4.02564476           R-squared     =  0.7160
-------------+------------------------------           Adj R-squared =  0.7121
       Total |  1034.66667    74   13.981982           Root MSE      =  2.0064

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           X |   1.605073   .1183215    13.57   0.000     1.369258    1.840887
       _cons |   15.35343   1.171641    13.10   0.000     13.01835     17.6885
------------------------------------------------------------------------------
```

Y= SALES
X= Amount Spent ADS ($)

Y=sales a=15.35

X=ADV b=1.6

I want to predict sales taking into consideration a specific level of advertising.
Sales_hat = 15.35 + 1.6*35(=x1) => here we will have the predicted value of advertising for this level of sales.

Predicting CHURN

Using the model parameters, you can estimate the churn rate at the individual customer level

The Lift Chart

- One way to evaluate predictive capability while also considering a proactive approach is to use lift logic
- Logic:
  - I identify different distinct groups based on the prediction of the dependent variable (Y_hat). Example: 2 groups, 5 groups, 10 groups (quintiles, deciles, etc.) -> the standard is 10 group, and in group number 10 we will have those that are very unlikely to churn while in group 1 there are those who are most likely to churn.
  - I check the mean of the observed variable (Y) in each of these groups.
  - I do this for both the calibration and validation models -> predict both the sample_in and the sampe_out.

- **Lift = average response of the (k-th) group, relative to the average response of the sample**
  r_k= Average response of the group k (k-le)
  r_0= Average response across the entire sample

$$\lambda_k = \frac{r_k}{r_0}$$

- **What does a very high lift indicate?**
- **it is possible to significantly increase the response probability by targeting the group that is more likely to respond compared to the entire sample.**

Decile	Observed Churn	Lift
1	35%	3.02
2	30%	2.59
3	25%	2.16
4	12%	1.03
5	8%	0.69
6	5%	0.43
7	1%	0.09
8	0%	0.00
9	0%	0.00
10	0%	0.00
Totale	12%	

This table is called lift table. In observed churn column we have the real variable while the decile column is what we predict.
If our model tells us that we have the ones who are more likely to change in group, we will expect a higher average than the last group where my model should put the ones at less risk of churn.

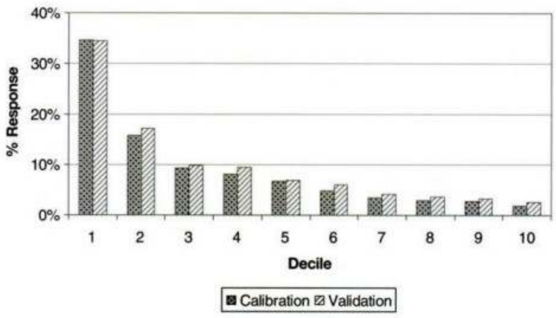
Example : Y=Churn Cliente i

Decile	Observed Churn	Lift
1	35%	3.02
2	30%	2.59
3	25%	2.16
4	12%	1.03
5	8%	0.69
6	5%	0.43
7	1%	0.09
8	0%	0.00
9	0%	0.00
10	0%	0.00
Totale	12%	

Example Lift del Top-Decile=3
It is about 3 times more likely for customers included in the top decile to churn compared to the average

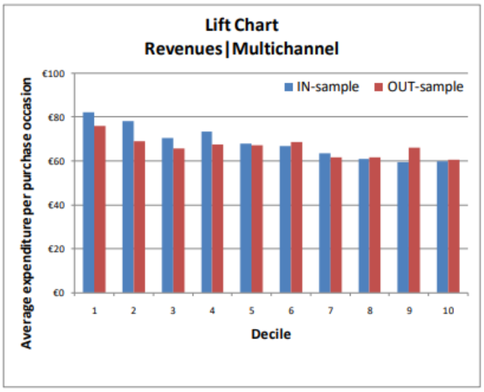
Whit the lift chart we can immediately understand if the model is good or not -> in fact, the model is good when in group 1 we have those who are at more risk with a higher average and, as we go up in terms of group, we can see how both the % of risk of churn and the average value decrease.

How do you interpret this LIFT CHART?



Fine. I would love to have this chart. We can use it to identify in advance those who are in risk of churn.

How do you interpret this LIFT CHART?



So and so because it is flat -> not good news because the model is not able to discriminate among groups. The same is when the model is neither increasing nor decreasing

How do you interpret this LIFT CHART?

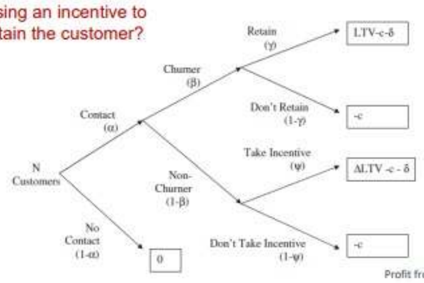
If our model gives this result we are fine and we can use to predict future trends and behavoir.

How do you identify customers who are much more likely to churn?
Targeting:

- Lift (e.g. >1) cutoff
- Individual Scores

Churn Management: Proactive Management

Using an incentive to retain the customer?



N= Total number of clients

α=Probability that the customer will be contacted as part of a churn program

β=Probability that the customer is a churner | has been contacted

Y=Probability the customer is retained| Churner

ψ=Probablity that a non-churner receives the incentives

Δ = Increase in LTV among non-churners who received the incentive

c= cost of contacting the customer

δ= cost of the incentive

Profit from a proactive action aimed at reducing churn

$$
\begin{aligned}
\Pi &= N\{\alpha\beta\gamma(LVC - c - \delta) + \alpha\beta(1-\gamma)(-c) + \alpha(1-\beta)\psi(\Delta LVC - c - \delta) \\
&\quad + \alpha(1-\beta)(1-\psi)(-c)\} \\
&= N\alpha\{(\beta\gamma + (1-\beta)\psi\Delta)LVC - \delta(\beta\gamma + (1-\beta)\psi) - c\} \qquad (24.8)
\end{aligned}
$$

Churn Management: Proactive

$$
\begin{aligned}
\Pi &= N\{\alpha\beta\gamma(LVC - c - \delta) + \alpha\beta(1-\gamma)(-c) + \alpha(1-\beta)\psi(\Delta LVC - c - \delta) \\
&\quad + \alpha(1-\beta)(1-\psi)(-c)\} \\
&= N\alpha\{(\underbrace{\boxed{\beta\gamma}}_{\text{Incremental profit from recovered customers}} + \underbrace{\boxed{(1-\beta)\psi\Delta}}_{\text{Incremental profit of delighted non-churners}})LVC - \underbrace{\boxed{\delta(\beta\gamma + (1-\beta)\psi)}}_{\text{Cost of incentives for churners and non-churners}} - \underbrace{\boxed{c}}_{\text{Cost of contact}}
\end{aligned}
$$

Maximum incentive cost

$$
\delta < \frac{(\beta\gamma + (1-\beta)\psi\Delta)LTV - c(1+\rho)}{(1+\rho)(\beta\gamma + \psi(1-\beta))}
$$

N= Total number of clients

α=Probability that the customer will be contacted as part of a churn program

β=Probability that the customer is a churner | has been contacted

Y=Probability the customer is retained| Churner

ψ=Probablity that a non-churner receives the incentives

Δ = Increase in LTV among non-churners who received the incentive

c= cost of contacting the customer

δ= cost of the incentive

These last two slides are not in the exam.

BOOKS R US: CASE STUDY – PREDICTING CHURN

Caso BooksRUs: Churn Management

- BookRus is a major multichannel European book retailer. The company sells books through stores, mail-order, phone and the Internet.
- Subscription base business. Membership renews at the end of each year.
- The firm mails a print catalog to the customer base five times per year
- The company opted for an omnichannel strategy providing members several possibilities to purchase: the physical stores, the website, mail/order or phone, the app

Step 1: Open the dataset DataCohort1.xlsx.

```
# Importing necessary libraries
import pandas as pd
import statsmodels.formula.api as smf
import matplotlib.pyplot as plt
import numpy as np
import os
```

```
#Step 0:Open dataset DataCohort1.xls
file_path = r'C:\Users\ValentiniS\DataCohort1.xlsx'
df = pd.read_excel(file_path)
df.head()
```

	id	churn	profits	multichannel	age	female	street_agent	north	early_email	bigcity	mean_city	franchisee	initialstorepromo	initialweb	initialstore
0	17051200	0	76.169998	1	33.0	1	1	1	1	0	0	0	0	1	0
1	17137573	0	123.320000	0	61.0	0	0	0	1	0	1	0	0	0	1
2	17071854	0	49.459999	0	23.0	1	0	0	1	0	0	0	0	1	0
3	17126628	1	42.689999	0	67.0	1	0	1	1	0	0	0	0	1	0
4	17126606	0	49.060001	0	46.0	1	0	1	1	0	0	1	0	1	0

```
df.columns
```

```
Index(['id', 'churn', 'profits', 'multichannel', 'age', 'female',
       'street_agent', 'north', 'early_email', 'bigcity', 'mean_city',
       'franchisee', 'initialstorepromo', 'initialweb', 'initialstore',
       'initialmobile', 'initialrevenues', 'initialreturns'],
      dtype='object')
```

Step 2: Estimate a logit model with the dependent variable (DV) = churn, and identify the factors that have the greatest impact on churn.

```
# Step 1: Run logit DV=churn
# create a quadratic effect for the variable "initialreturns,"
df["initialreturns2"] = df["initialreturns"]**2
```

Creating a **quadratic effect in a regression** can be useful for capturing non-linear relationships between the predictor variable and the response variable. In many real-world scenarios, the relationship between variables is not always linear and can exhibit curves or bends.

Here are a few reasons why using a quadratic effect (or higher-order polynomial terms) might be beneficial in a regression analysis:

- **Non-linear Relationships:** A quadratic effect allows the model to account for situations where the response variable does not change linearly with the predictor. For instance, some phenomena may show an initial increase at a decreasing rate or vice versa.
- **Capturing Curvature:** When there is a U-shaped or inverted U-shaped relationship between the predictor and the response, a quadratic term can help capture this curvature more accurately.
- **Better Prediction:** In cases where the true relationship is indeed quadratic, incorporating a quadratic term can lead to better predictions by aligning the model more closely with the underlying data distribution.

However, it's essential to exercise caution when using higher-order polynomial terms, as they can introduce overfitting, making the model too complex and less generalizable to new data.

```
# Step 2: Run Logit DV=churn
formula = ('churn ~ multichannel + age + female + street_agent + '
           'north + early_email + bigcity + mean_city + franchisee + '
           'initialstorepromo + initialweb + initialstore + initialmobile + '
           'initialrevenues + initialreturns + initialreturns2')

model = smf.logit(formula, data=df).fit()

print(model.summary2())
```

```
Optimization terminated successfully.
         Current function value: 0.313075
         Iterations 7
                          Results: Logit
=================================================================
Model:              Logit            Pseudo R-squared: 0.037
Dependent Variable: churn            AIC:              22194.0895
Date:               2023-10-09 17:28 BIC:              22338.1511
No. Observations:   35391            Log-Likelihood:   -11080.
Df Model:           16               LL-Null:          -11500.
Df Residuals:       35374            LLR p-value:      1.2991e-168
Converged:          1.0000           Scale:            1.0000
No. Iterations:     7.0000
-----------------------------------------------------------------
                   Coef.  Std.Err.    z     P>|z|   [0.025  0.975]
-----------------------------------------------------------------
Intercept         -1.4676   0.0738 -19.8916 0.0000 -1.6122 -1.3230
multichannel      -1.0227   0.1131  -9.0401 0.0000 -1.2445 -0.8010
age               -0.0126   0.0012 -10.5182 0.0000 -0.0150 -0.0103
female            -0.4504   0.0367 -12.2635 0.0000 -0.5224 -0.3784
street_agent      -0.3603   0.0390  -9.2463 0.0000 -0.4367 -0.2840
north              0.1753   0.0517   3.3918 0.0007  0.0740  0.2766
early_email       -0.0514   0.0388  -1.3263 0.1847 -0.1274  0.0246
bigcity            0.7864   0.0594  13.2274 0.0000  0.6698  0.9029
mean_city          0.1454   0.0701   2.0746 0.0380  0.0080  0.2828
franchisee        -0.0500   0.0387  -1.2920 0.1964 -0.1258  0.0258
initialstorepromo  0.4627   0.0823   5.6248 0.0000  0.3015  0.6240
initialweb        -0.1298   0.1900  -0.6833 0.4944 -0.5023  0.2426
initialstore      -0.5876   0.0833  -7.0549 0.0000 -0.7509 -0.4244
initialmobile     -0.3554   0.1405  -2.5296 0.0114 -0.6308 -0.0800
initialrevenues    0.0064   0.0021   3.0102 0.0026  0.0022  0.0105
initialreturns    -0.0229   0.0133  -1.7185 0.0857 -0.0490  0.0032
initialreturns2    0.0003   0.0001   1.7563 0.0790 -0.0000  0.0005
=================================================================
```

We included at first everything as we don't have to many varibles.
p-value of multichannel is 0.000 -> it is significant, which means that this varibale has some impact on churn. To understand the impact we have to check the sign of the coefficient; in this case is negative which is a good news because it has a negative impact on churn -> reducing churn is exactlywhat we are looking for.
On the other hand, initialstorepromo, even though it is significant due to a p-value equal to 0 has a positive impact on churn, which means that those who receives a promo at the beginnig are moe likelyto churn.

How to interpret the coefficient in a Logistic Regression? POSSIBILITY – COMPUTE OR (exponantial of the coefficients)

The model assumes that:

Y_i = 1 with probability p_i

0 with probability 1-p_i

- p_i represents the probability that the event occurs (e.g. Y_i=1="YES")
- If an event occurs with probability p_i, then the odds ratio (OR) will be $p_i/(1-p_i)$ -> average impact of thevariable. When OR=1 this is the benckmark of the indiffrence position

Take these two coefficients as examples:

Results: Logit

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
multichannel	-1.0227	0.1131	-9.0401	0.0000	-1.2445	-0.8010
initialstorepromo	0.4627	0.0823	5.6248	0.0000	0.3015	0.6240

$b_{multichannel}$=-1.002

Odds Ratio (OR): Exp(-1.002) ≈ 0.359 The odds ratio is calculated by taking the exponential of the coefficient. An odds ratio of 0.359 means that for a one-unit increase in the "multichannel" variable (i.e. being multichannel), the odds of churn decrease by approximately 64% (because 1 - 0.359 ≈ 0.641, or about 64%).

In simpler terms, being in the "multichannel" category decreases the odds of churn by approximately 64% compared to not being in the "multichannel" category.

$b_{Initial\ promo}$=0.4627

Odds Ratio (OR): Exp(0.4627) ≈ 1.588
The odds ratio is calculated by taking the exponential of the coefficient. An odds ratio of approximately 1.588 means taking advantage of a promotion in the initial period increases the odds of churn by about 1.588 times compared to not being in the "initial promo" category.

In simpler terms, taking advantage of a promo at the beginning of the relationship with the firm increases the odds of churn by approximately 58.8% compared to not taking advantage of a promo at the beginning.

Interpreting Coefficients in a Logistic Regression Model: odds ratio

Imagine x_k = age

- **If β_k is positive**→for example if β_k =1, e^1=2.7:
- We can say that a one-unit increase in age increases the probability of your binary variable y (e.g. retention) by a factor of 2.7 (or 170%), keeping all other factors constant

- **if β_k is negative**→ for example if β_k = -1, e^{-1}=.36,
- we can say that a one-unit increase in age decreases the probability of your binary variable y (e.g. retention) by a factor of 0.36 (or 64% less), keeping all other factors constant.

Step 3: Estimate a model with the DV = profits.

```
#Step 3: Run OLS reg DV=profits
formula = ('profits ~ multichannel + age + female + street_agent + '
           'north + early_email + bigcity + mean_city + franchisee + '
           'initialstorepromo + initialweb + initialstore + initialmobile + '
           'initialrevenues + initialreturns + initialreturns2')

model = smf.ols(formula, data=df).fit()

print(model.summary2())
```

```
                 Results: Ordinary least squares
====================================================================
Model:              OLS              Adj. R-squared:     0.336
Dependent Variable: profits          AIC:                287829.0750
Date:               2023-10-09 17:28 BIC:                287973.1367
No. Observations:   35391            Log-Likelihood:     -1.4390e+05
Df Model:           16               F-statistic:        1122.
Df Residuals:       35374            Prob (F-statistic): 0.00
R-squared:          0.337            Scale:              199.23
--------------------------------------------------------------------
                   Coef.  Std.Err.    t    P>|t|   [0.025  0.975]
--------------------------------------------------------------------
Intercept          4.6034   0.3116 14.7718 0.0000  3.9926  5.2142
multichannel      20.9517   0.3207 65.3373 0.0000 20.3231 21.5802
age                0.1509   0.0048 31.4807 0.0000  0.1415  0.1603
female             1.2085   0.1623  7.4482 0.0000  0.8905  1.5265
street_agent       3.1672   0.1585 19.9766 0.0000  2.8565  3.4780
north              1.9387   0.2022  9.5889 0.0000  1.5424  2.3349
early_email        1.8377   0.1644 11.1769 0.0000  1.5154  2.1599
bigcity            0.1535   0.2579  0.5952 0.5517 -0.3519  0.6589
mean_city          0.0237   0.2711  0.0873 0.9304 -0.5076  0.5550
franchisee         0.4522   0.1596  2.8340 0.0046  0.1394  0.7649
initialstorepromo -1.8296   0.3417 -5.3540 0.0000 -2.4994 -1.1598
initialweb        11.0348   0.7429 14.8539 0.0000  9.5788 12.4909
initialstore      -2.5538   0.3014 -8.4724 0.0000 -3.1446 -1.9630
initialmobile     10.0674   0.5492 18.3326 0.0000  8.9910 11.1438
initialrevenues    0.5308   0.0086 61.5983 0.0000  0.5139  0.5476
initialreturns    -0.5039   0.0541 -9.3126 0.0000 -0.6099 -0.3978
initialreturns2    0.0021   0.0008  2.7236 0.0065  0.0006  0.0036
--------------------------------------------------------------------
Omnibus:             11760.006      Durbin-Watson:         2.016
Prob(Omnibus):       0.000          Jarque-Bera (JB):      71137.918
Skew:                1.468          Prob(JB):              0.000
Kurtosis:            9.295          Condition No.:         1680
====================================================================
* The condition number is large (2e+03). This might indicate
strong multicollinearity or other numerical problems.
```

Step 4: Focus on churn and test if the model has predictive ability. Use:

- Calibration and Validation
- Lift Chart

Step 4 involves assessing the model's ability to predict churn events.

One of the evaluation techniques mentioned is the use of a **lift chart**, which visually represents the model's predictive performance. However, please note that there are alternative approaches to evaluate the model, besides the lift chart, that can also be effective.

The evaluation process utilizes both **in-sample** and **out-of-sample**. **In-sample** is the portion used to train or fit the model **out-of-sample** is held separately to evaluate how well the model performs on unseen data.

The creation of in-sample and out-of-sample data involves a **random selection process**. This randomness ensures that the selected samples are representative and unbiased, providing a fair evaluation of the model's performance. Randomly selecting data for both in-sample and out-of-sample evaluations enhances the

robustness of the assessment.

```
# Step 4: Generate random data
df["random"] = np.random.uniform(size=df.shape[0])
df["random"].describe()
# Calculate the median as the cutoff value
cut_off = np.median(df["random"])

# Print the cutoff value
print("Cutoff value:", cut_off)
```

```
Cutoff value: 0.4963136276443664
```

```
# Create the 'out_sample' dummy variable
df["out_sample"] = np.where(df['random'] > cut_off, 1, 0)

# Calculate the frequency table for the 'out_sample' column
frequency_table = df['out_sample'].value_counts()
# Calculate the percentage for each value
percentage_table = (df['out_sample'].value_counts(normalize=True) * 100).round(1)
# Print the frequency & percentage tables
print(frequency_table)
print(percentage_table)
```

```
out_sample
0    17696
1    17695
Name: count, dtype: int64
out_sample
0    50.0
1    50.0
Name: proportion, dtype: float64
```

```
df.head()
```

	id	churn	profits	multichannel	age	female	street_agent	north	early_email	bigcity	...	franchisee	initialstorepromo	initialweb	initialstore	initialm
0	17051200	0	76.169998	1	33.0	1	1	1	1	0	...	0	0	1	0	
1	17137573	0	123.320000	0	61.0	0	0	0	1	0	...	0	0	0	1	
2	17071854	0	49.459999	0	23.0	1	0	0	1	0	...	0	0	1	0	
3	17126628	1	42.689999	0	67.0	1	0	1	1	0	...	0	0	1	0	
4	17126606	0	49.060001	0	46.0	1	0	1	1	0	...	1	0	1	0	

5 rows × 21 columns

```python
# I don't need the variable random anymore and i will drop it
df.drop(columns=['random'], inplace=True)
df.head()
```

	id	churn	profits	multichannel	age	female	street_agent	north	early_email	bigcity	mean_city	franchisee	initialstorepromo	initialweb	initialstore
0	17051200	0	76.169998	1	33.0	1	1	1	1	0	0	0	0	1	0
1	17137573	0	123.320000	0	61.0	0	0	0	1	0	1	0	0	0	1
2	17071854	0	49.459999	0	23.0	1	0	0	1	0	0	0	0	1	0
3	17126628	1	42.689999	0	67.0	1	0	1	1	0	0	0	0	1	0
4	17126606	0	49.060001	0	46.0	1	0	1	1	0	0	1	0	1	0

```python
# we re-estimate the churn model only for the in-sample

# Filter the DataFrame for in-sample
in_sample_df = df[df['out_sample'] == 0]
# Define your formula
formula = ('churn ~ multichannel + age + female + street_agent + '
           'north + early_email + bigcity + mean_city + franchisee + '
           'initialstorepromo + initialweb + initialstore + initialmobile + '
           'initialrevenues + initialreturns + initialreturns2')
# Estimate the model using the filtered DataFrame
model = smf.logit(formula, data=in_sample_df).fit()
# Print the model summary
print(model.summary2())
```

```
Optimization terminated successfully.
         Current function value: 0.321471
         Iterations 7
                          Results: Logit
=================================================================
Model:              Logit            Pseudo R-squared: 0.037
Dependent Variable: churn            AIC:              11411.5144
Date:               2023-10-09 17:32 BIC:              11543.7930
No. Observations:   17696            Log-Likelihood:   -5688.8
Df Model:           16               LL-Null:          -5905.8
Df Residuals:       17679            LLR p-value:      2.5165e-82
Converged:          1.0000           Scale:            1.0000
No. Iterations:     7.0000
-----------------------------------------------------------------
                   Coef.  Std.Err.    z     P>|z|   [0.025  0.975]
-----------------------------------------------------------------
Intercept         -1.4283   0.1036 -13.7920 0.0000 -1.6312 -1.2253
multichannel      -1.0698   0.1593  -6.7150 0.0000 -1.3820 -0.7575
age               -0.0127   0.0017  -7.6206 0.0000 -0.0160 -0.0095
female            -0.4564   0.0510  -8.9524 0.0000 -0.5563 -0.3564
street_agent      -0.3519   0.0541  -6.5016 0.0000 -0.4580 -0.2458
north              0.2048   0.0729   2.8075 0.0050  0.0618  0.3478
early_email       -0.0697   0.0541  -1.2880 0.1977 -0.1757  0.0364
bigcity            0.7713   0.0840   9.1809 0.0000  0.6067  0.9360
mean_city          0.1572   0.0980   1.6036 0.1088 -0.0349  0.3492
franchisee        -0.0509   0.0538  -0.9452 0.3446 -0.1564  0.0546
initialstorepromo  0.5114   0.1126   4.5431 0.0000  0.2908  0.7321
initialweb         0.0523   0.2454   0.2132 0.8311 -0.4286  0.5333
initialstore      -0.5728   0.1154  -4.9626 0.0000 -0.7991 -0.3466
initialmobile     -0.2108   0.1892  -1.1144 0.2651 -0.5815  0.1599
initialrevenues    0.0050   0.0030   1.6996 0.0892 -0.0008  0.0109
initialreturns    -0.0396   0.0210  -1.8817 0.0599 -0.0808  0.0016
initialreturns2    0.0003   0.0002   1.4655 0.1428 -0.0001  0.0007
=================================================================
```

```
# MY MODEL previously estimated logistic regression model

# Predict probabilities for both in-sample and out-sample
# Predict probabilities for the entire dataset
df['y_hat'] = model.predict(df)
df.head()
```

	id	churn	profits	multichannel	age	female	street_agent	north	early_email	bigcity	...	franchisee	initialstorepromo	initialweb	initialstore	initialm
0	17051200	0	76.169998	1	33.0	1	1	1	1	0	...	0	0	1	0	
1	17137573	0	123.320000	0	61.0	0	0	0	1	0	...	0	0	0	1	
2	17071854	0	49.459999	0	23.0	1	0	0	1	0	...	0	0	1	0	
3	17126628	1	42.689999	0	67.0	1	0	1	1	0	...	0	0	1	0	
4	17126606	0	49.060001	0	46.0	1	0	1	1	0	...	1	0	1	0	

5 rows × 21 columns

```
# create decide for OUT SAMPLE
# Filter the DataFrame for in-sample
out_sample_df = df[df['out_sample'] == 1]

df.loc[df['out_sample'] == 1, 'decile_out'] = pd.qcut(out_sample_df['y_hat'], 10, labels=False)
#CHECK IF DECILES ARE CREATED CORRECTLY
df[["decile_out",  "y_hat"]].groupby("decile_out").mean()
```

decile_out	y_hat
0.0	0.036700
1.0	0.057600
2.0	0.070239
3.0	0.080519
4.0	0.088854
5.0	0.099676
6.0	0.112738
7.0	0.127526
8.0	0.152104
9.0	0.206303

```
#DECILE=1 SHOULD HAVE THE HIGHEST Y_HAT AND 10 THE LOWEST
df.loc[df['out_sample'] == 1, 'decile_out'] = 10 - df.loc[df['out_sample'] == 1, 'decile_out']
#CHECK IF DECILES ARE CREATED CORRECTLY
df[["decile_out",  "y_hat"]].groupby("decile_out").mean()
```

decile_out	y_hat
1.0	0.206303
2.0	0.152104
3.0	0.127526
4.0	0.112738
5.0	0.099676
6.0	0.088854
7.0	0.080519
8.0	0.070239
9.0	0.057600
10.0	0.036700

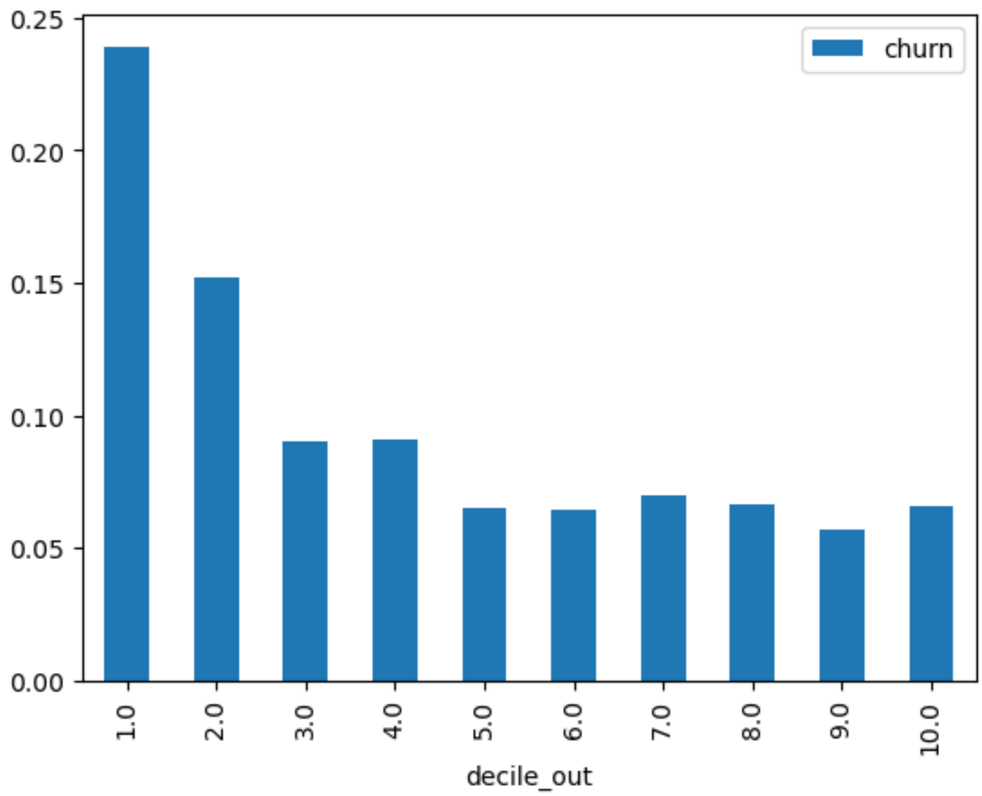
The first group has 20% probability to churn. Now we have to compare the model with the real y which is **churn.**

```
#LIFT TABLE WITH OBSERVED CHURN
df[["decile_out", "churn"]].groupby("decile_out").mean()
```

	churn
decile_out	
1.0	0.239389
2.0	0.151806
3.0	0.090234
4.0	0.090604
5.0	0.064699
6.0	0.064262
7.0	0.069977
8.0	0.066102
9.0	0.056529
10.0	0.065537

```
# create the lift chart

df[["decile_out",  "churn"]].groupby("decile_out").mean().plot.bar()
```

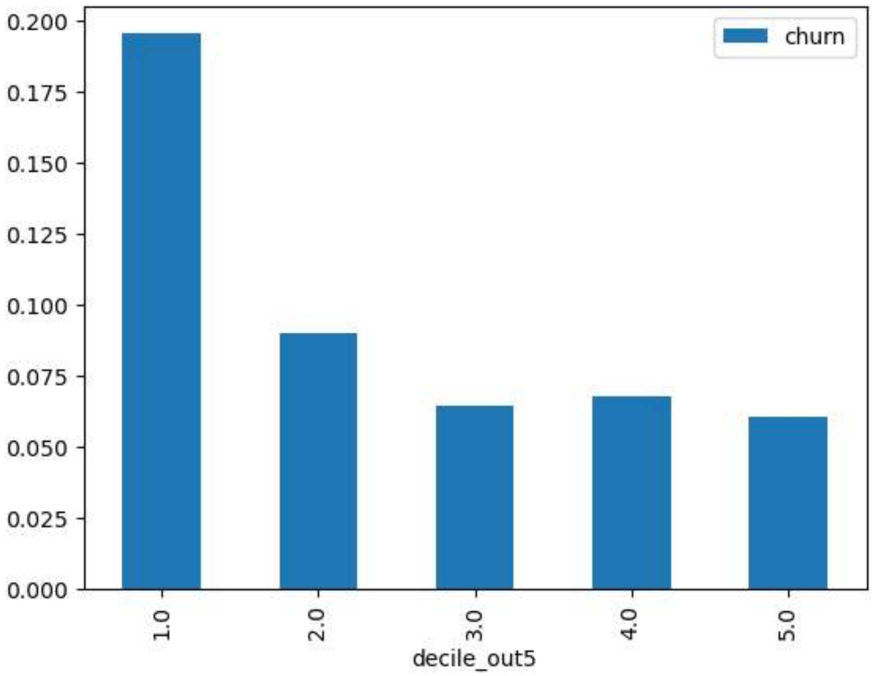
<AxesSubplot:xlabel='decile_out'>



This model is able to distinguish pretty well the first four/five groups, then the results are pretty much the same, so now we want to use just the significant groups.

```
# Lift with 5 groups instead of 10

# Assuming you want to assign 5 groups for 'decile_out5'
out_sample_df = df[df['out_sample'] == 1]

# Calculate deciles for the out-sample with 5 groups
df.loc[df['out_sample'] == 1, 'decile_out5'] = pd.qcut(out_sample_df['y_hat'], 5, labels=False)
df.loc[df['out_sample'] == 1, 'decile_out5'] = 5 - df.loc[df['out_sample'] == 1, 'decile_out5']

# CHART WITH 5 GROUPS

df[["decile_out5", "churn"]].groupby("decile_out5").mean().plot.bar()
```

```
<AxesSubplot:xlabel='decile_out5'>
```



```
#oper dataset cohort 2
df2_path = r'C:\Users\ValentiniS\2023_05_Marketing AXA\Lab 1 Session2_BooksRUS_Churn\DataCohort2.xlsx'
df2 = pd.read_excel(df2_path)
df2.head()
```

	id	profits	churn_observed	multichannel	age	female	north	bigcity	mean_city	early_email	franchisee	street_agent	initialweb	initialstore	initialmobil
0	1	39.040001	0	0	38.720001	1	1	0	0	1	0	0	0	0	
1	2	53.080002	0	1	38.720001	0	1	0	0	1	0	0	1	0	
2	3	0.000000	0	0	62.000000	1	1	0	0	1	1	0	1	0	
3	4	56.049999	1	1	34.000000	0	1	0	0	1	1	0	1	0	
4	5	9.290000	0	0	38.720001	1	1	0	0	1	1	0	0	0	

NOW I OPEN A NEW DATASET BECAUSE I WANT TO PREDICT CHURN IN THIS NEW DATASET USING THE PARAMETER OF THE MODEL JUST ESTIMATED

```
# Step4:
### Predict churn on a new dataset, df2
### using the parameters from the logistic regression model estimated on dataset db_c1.
# Add a new column "churn_hat" to df2 containing the churn predictions.
df2['churn_hat'] = model.predict(df2)
df2.head()
```

	id	profits	churn_observed	multichannel	age	female	north	bigcity	mean_city	early_email	...	street_agent	initialweb	initialstore	initialmobile	initial
0	1	39.040001	0	0	38.720001	1	1	0	0	1	...	0	0	0	1	
1	2	53.080002	0	1	38.720001	0	1	0	0	1	...	0	1	0	0	
2	3	0.000000	0	0	62.000000	1	1	0	0	1	...	0	1	0	0	
3	4	56.049999	1	1	34.000000	0	1	0	0	1	...	0	1	0	0	
4	5	9.290000	0	0	38.720001	1	1	0	0	1	...	0	0	0	0	

5 rows × 21 columns

```
# Step4:
# create DECILE
#CHECK IF DECILES ARE CREATED CORRECTLY
df2['decile'] = pd.qcut(df2['churn_hat'], 10, labels=False)
df2['decile'] = 10 - df2['decile']
df2.head()
```

	id	profits	churn_observed	multichannel	age	female	north	bigcity	mean_city	early_email	...	initialweb	initialstore	initialmobile	initialstorepromo	i
0	1	39.040001	0	0	38.720001	1	1	0	0	1	...	0	0	1	0	
1	2	53.080002	0	1	38.720001	0	1	0	0	1	...	1	0	0	0	
2	3	0.000000	0	0	62.000000	1	1	0	0	1	...	1	0	0	0	
3	4	56.049999	1	1	34.000000	0	1	0	0	1	...	1	0	0	0	
4	5	9.290000	0	0	38.720001	1	1	0	0	1	...	0	0	0	0	

5 rows × 22 columns

```
# Step4:

# LIFT TABLE
df2[["decile", "churn_observed"]].groupby("decile").mean()
```

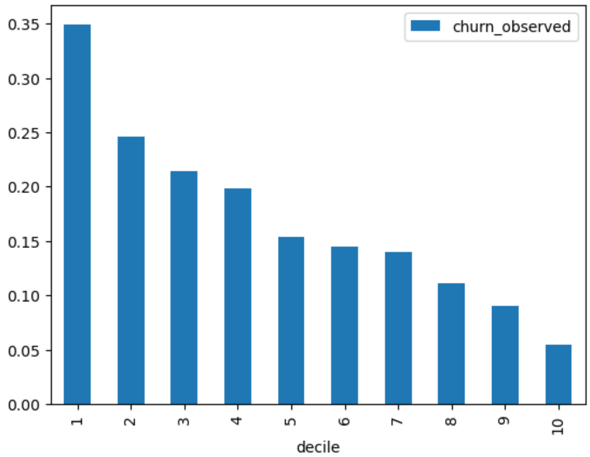
decile	churn_observed
1	0.349201
2	0.245854
3	0.214728
4	0.198307
5	0.154173
6	0.144524
7	0.140196
8	0.111003
9	0.090407
10	0.054054

```
# Step4:
df2[["decile", "churn_observed"]].groupby("decile").mean().plot.bar()
```

Step 6: Use the estimates to predict churn in a new dataset, DataCohort2.xlsx.

Whom would you include in the 'Targeted Proactive' action? Why?

<AxesSubplot:xlabel='decile'>

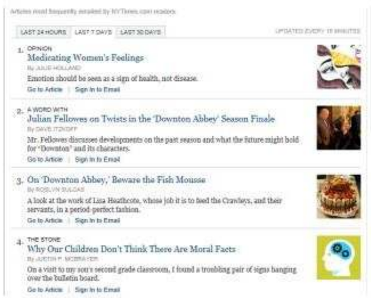


Evaluate the effectiveness of marketing actions

Our task is not only developing a marketing plan, but also be able to reassure the management that our marketing plan is going to be effective.

Overview: Example 1 New York Times

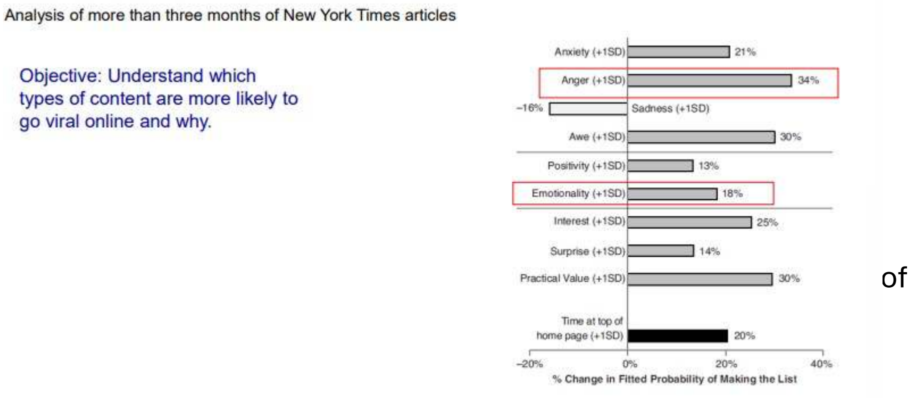
Why are some articles from The New York Times more shared than others and become viral? What are some potential factors that contribute to explaining their success?



Main argument: The emotional content of the article -> unboots the engagement, understanding which are the element which allows a content to become viral

1. Text analysis to quantify the article's emotional content
2. Text analysis to quantify the type of emotions evoked by the article (anger, anxiety, sadness, joy, humor etc.)
3. Experiment where they «manipulate» the emotional content of an ADV

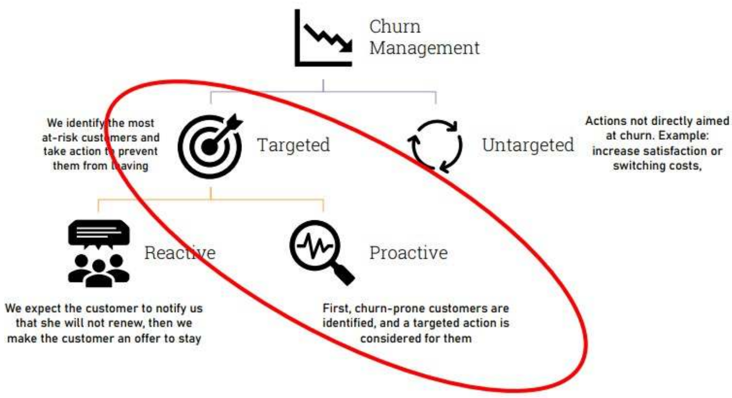
They have a brand, and their managerial task is to develop a social media marketing campaign for thatbrand with the purpose to increase engagement -> they can create a content that is able to generate emotions, because hopefully this kind of content would be more likely to be shared.



- *Experiment*-> to check if generating emotion could increase engagement; in principle we wantto link our brand with positive emotion, but it is not always the case
- Two versions of an ADV that differed on how much humor they evoked
- High- or low-amusement version of a story about a recent advertising campaign for Jimmy Dean sausages
- The two versions were pre-tested to verify that they are perceived differently in terms of humor levels
- Likelihood to share the ADV:
  - high-amusement (M = 3.97) versus low-amusement (M = 2.92; p <.005)
  -

Punch Line -> Predictive models and field experiments are commonly used also together to generateand test **the effectiveness of marketing actions and strategies**.

These methods can provide valuable insights for identifying possible marketing strategies andimproving overall effectiveness, before actually entering the market.
How can they be used in the context of Churn Management?



Lift chart is a perfect tool to predict if my customers are likely or not to leave.

Example

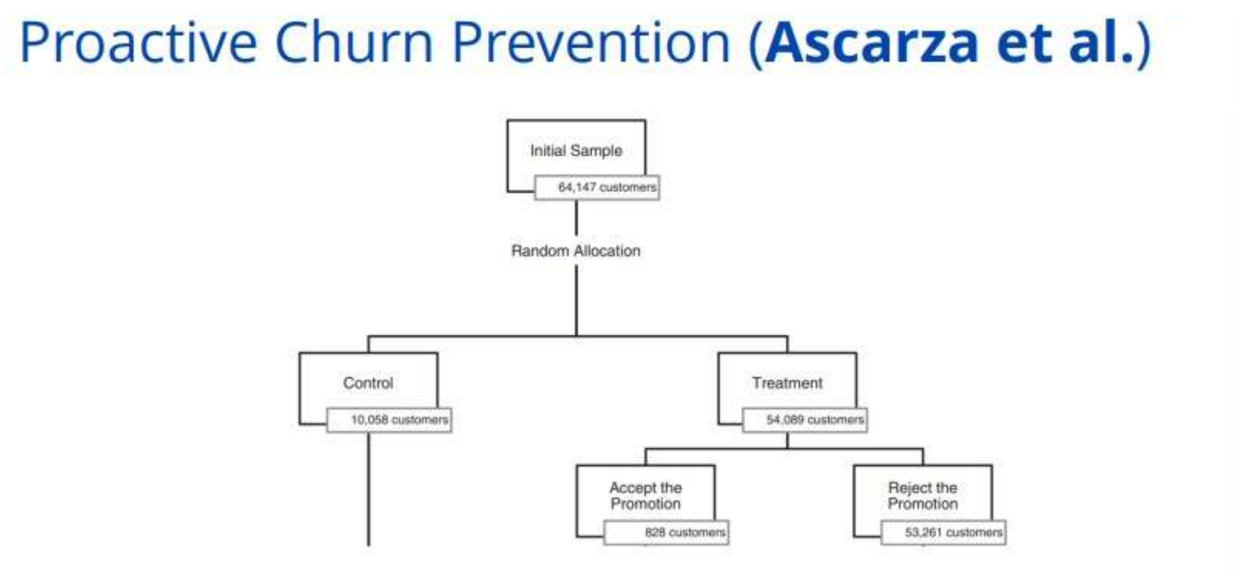
Ascarza, Iyengar and Schleicher (2016) "Proactive Churn Prevention Using Plan Recommendations: Evidence from a Field Experiment." Journal of Marketing Research

Proactive Churn Prevention (Ascarza et al.)

Field experiment (A/B test) with a telecom provider in South America
Intervention:

- Select ~65k customers active in Q1 2011
- 15% of customers left aside as a control group
- Treatment consists of calling the customers (high risk churn) and offering to upgrade to either $47 or $63 plan
  - No other commitments / No fine print
  - If accepted, extra $15 during the following 3 months

For all customers, including those in the control condition, one or both of the featured plans would have been better than their current plan. To incentivize customers to upgrade to the suggested plans, the company offered an additional credit of $15 for each of the following three months if they agreed to upgrade to one of the featured plans.



CASE 3 PILGRIM: TRAGETED PROACTIVE COMPAIGN

Context

Industry: Banking, Pilgrim
BankSample: 31,634

- For these customers, a churn model was estimated. Hence, each customer is associated with a churn probability, estimated at the time of acquisition through an internally consolidated model.
- Each customer is assigned a decile (1=customers with the highest churn risk, 10=lowest).
- The company has decided to focus a **targeted proactive strategy** on the customers in the **first 3 deciles**. They were the subject of a field test. Following one year after the field test, information on churn was collected for the entire sample of 31,634 customers.

Field Test

10 quantiles of churn_hat	Freq.	Percent	Cum.
1	3,163	10.00	10.00
2	3,163	10.00	20.00
3	3,164	10.00	30.00
4	3,163	10.00	40.00
5	3,164	10.00	50.00
6	3,163	10.00	60.00
7	3,163	10.00	70.00
8	3,164	10.00	80.00
9	3,163	10.00	90.00
10	3,164	10.00	100.00
Total	31,634	100.00	

Dataset: PilgrimTargetedField.xls

id	Customer identifier code
decile	Decile of belonging derived from the churn model
target_proactive	1 = included in the churn program (proactive), 0 = control (random selection)
retention	1=retained, 0=churn
MainlyOnline_bank_previous	1 indicates if the customer predominantly uses online banking, 0 if not
District	District of residence (USA)
Tenure	Number of years as a customer
AboveMedian_Tenure_Target	Variable indicating whether tenure is above the median of customers = 1, or below the median = 0
age	Age in classes, 0 indicates missing
agemiss	Variable that takes a value of 1 if age is missing, 0 otherwise
inc	Income, 0 indicates missing
incmiss	Variable that takes a value of 1 if income is missing, 0 otherwise
dist1100	Dummy variables identifying the 3 districts of residence for analysis
dist1200	Dummy variables identifying the 3 districts of residence for analysis
dist1300	Dummy variables identifying the 3 districts of residence for analysis
churn_hat	Prediction of churn from predictive model (made before observing actual churn)

Decile and churn hat columns were created **before** observing the actual retention.

Field test → Variable Targeted Proactive

Did not Receive Anything

Field Test	# of Customers
0=Control Group	4,394
1= Receive Incentive	5,096
Missing= Not included in the field	22,144
Total	31,634

Not subjected to testing

They receive a call and email message from customer care that proposes: If retained, Offer the customer a personalized financial consultation with a banking expert, a free travel insurance and benefit abroad and exclusive access to tailored events

Purpose

A field experiment designed to address the following key questions:

1. Is the target proactive strategy effective?
2. Yes, no, why? Carefully justify the response
3. **CASE PILGRIM: FIELD TEST**

Suggested Steps:

Step 0: Open the dataset PilgrimTargetedField.xls.

```
import pandas as pd
import statsmodels.formula.api as smf
import matplotlib.pyplot as plt
import numpy as np
import os
```

```
#Step 0:Open dataset PilgrimTargetedField.xls
file_path = r'C:\Users\ValentiniS\PilgrimTargetedField.xls'
db_pilgrim = pd.read_excel(file_path)
db_pilgrim.head()
#note variable DECILE is already created
```

	id	decile	target_proactive	retention	MainlyOnline_bank_previous	District	Tenure	AboveMedian_Tenure_Target	age	agemiss	inc	incmiss	dist1100	dist120
0	8197	1	0.0	1	1	1200	9	0	0	1	0	1	0	
1	4270	1	1.0	0	0	1200	2	0	0	1	0	1	0	
2	4598	8	NaN	1	0	1200	4	0	2	0	6	0	0	
3	30687	1	0.0	0	0	1200	5	0	0	1	0	1	0	
4	29042	3	1.0	1	0	1200	9	0	1	0	1	0	0	

```
db_pilgrim.shape
```

```
(31634, 16)
```

Step 1a: Check: the lift chart to verify if the predictive model worked using the already created decile variable and observed retention. Remember, deciles are created based on the predicted churn.

Step 1b: Check: verify the random assignment of the Targeted Proactive variable.

```
#I have the observed retention, therefore  compute the observe Churn
db_pilgrim['Churn'] = 1-db_pilgrim['retention']
db_pilgrim.head()
```

	id	decile	target_proactive	retention	MainlyOnline_bank_previous	District	Tenure	AboveMedian_Tenure_Target	age	agemiss	inc	incmiss	dist1100	dist1200
0	8197	1	0.0	1	1	1200	9	0	0	1	0	1	0	1
1	4270	1	1.0	0	0	1200	2	0	0	1	0	1	0	1
2	4598	8	NaN	1	0	1200	4	0	2	0	6	0	0	1
3	30687	1	0.0	0	0	1200	5	0	0	1	0	1	0	1
4	29042	3	1.0	1	0	1200	9	0	1	0	1	0	0	1

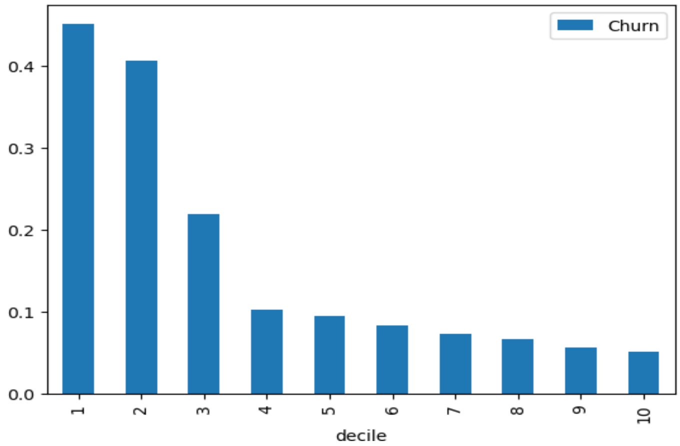
```
#descriptive table associated to the lift chart: Average Observed Churn per decile
db_pilgrim [["decile", "Churn"]].groupby("decile").mean()
```

decile	Churn
1	0.451470
2	0.406892
3	0.219027
4	0.102118
5	0.095449
6	0.083465
7	0.073032
8	0.066056
9	0.056908
10	0.050885

Churn= 1-retention.

```
# Create the lift chart: Churn
db_pilgrim [["decile", "Churn"]].groupby("decile").mean().plot.bar()
```

<AxesSubplot:xlabel='decile'>



This model is super in terms of prediction validity!

```
# Step 1b: Check Random Assignment
mask = db_pilgrim[db_pilgrim['decile']<4]
formula = ('target_proactive ~ MainlyOnline_bank_previous + Tenure +'
           'age + agemiss + inc + incmiss + dist1100 + dist1200')

model = smf.logit(formula, data = mask).fit()

print(model.summary2())
```

```
Optimization terminated successfully.
         Current function value: 0.690220
         Iterations 4
                          Results: Logit
=====================================================================
Model:                  Logit              Pseudo R-squared:  0.000
Dependent Variable:     target_proactive   AIC:               13118.3732
Date:                   2023-10-15 22:07   BIC:               13182.7951
No. Observations:       9490               Log-Likelihood:    -6550.2
Df Model:               8                  LL-Null:           -6552.0
Df Residuals:           9481               LLR p-value:       0.89257
Converged:              1.0000             Scale:             1.0000
No. Iterations:         4.0000
---------------------------------------------------------------------
                           Coef.  Std.Err.    z     P>|z|  [0.025 0.975]
---------------------------------------------------------------------
Intercept                  0.0759   0.1606  0.4723 0.6367 -0.2389 0.3906
MainlyOnline_bank_previous 0.0305   0.0669  0.4557 0.6486 -0.1007 0.1617
Tenure                    -0.0031   0.0042 -0.7531 0.4514 -0.0113 0.0050
age                        0.0077   0.0381  0.2029 0.8392 -0.0670 0.0825
agemiss                    0.0127   0.1326  0.0958 0.9237 -0.2471 0.2725
inc                        0.0168   0.0236  0.7134 0.4756 -0.0294 0.0630
incmiss                    0.0254   0.1183  0.2147 0.8300 -0.2064 0.2572
dist1100                   0.0216   0.0829  0.2607 0.7943 -0.1409 0.1841
dist1200                   0.0684   0.0603  1.1332 0.2571 -0.0499 0.1866
=====================================================================
```

We don't need to create a random variable, because it was created by the management in this case.The two group are identified by the value of the target_proactive variable (0 means control group, 1 experimental group). If my purposes is to randomize.
We run a logit model with as dependent variable target_proactive and as independent variables all the customer's characteristic in order to assign a p-value to verify if there is a likelihood to be included in the control group or in the experimental group.

Example

Age: p-value (which is the amount of tolerated error in the estimate)= 0.8392 -> it is not significant andwith are happy about that because we don't want that age is associated in any possible way to the likelihood to belong to the control group or the experimental group -> if it is not significant that meansthat age does not explain the likelihood to be included in the control group or in the experimental group. If the p-value in this test is significant will be a problem -> if age increases, it is more likely to beadd in the experimental group which means a bias and we have not randomized)

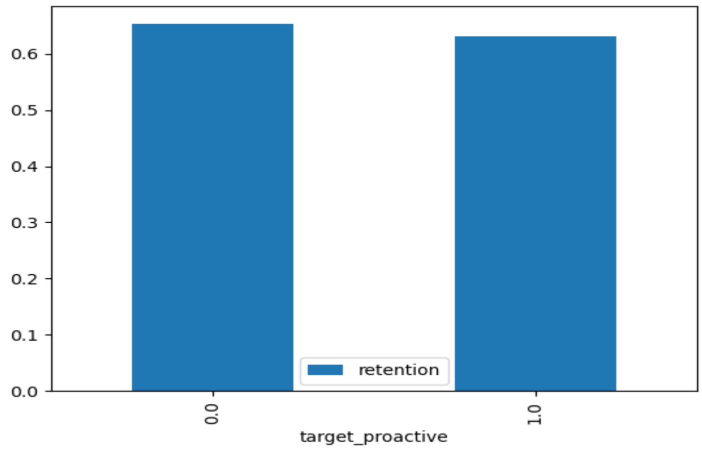
Step 2: Model Free Evidence: Verify if the targeted proactive campaign worked.

```
# Step 2: Results of the Field test: Model Free Evidence
db_pilgrim[['target_proactive', "retention"]].loc[db_pilgrim['decile']<4].groupby("target_proactive").mean()
```

target_proactive	retention
0.0	0.652025
1.0	0.631279

```
db_pilgrim[['target_proactive', "retention"]].loc[db_pilgrim['decile']<4].groupby("target_proactive").mean().plot.bar()
```

<AxesSubplot:xlabel='target_proactive'>



The two output are the same in terms of meaning, but we simply use two different tools.

We can say that the campaign is not effective because the results is pretty much the same in bothgroups, and above all in the experimental group is even lower than in the control group.

Step 3: Test: Test if the targeted proactive campaign worked and comment on the results.

The variable target_proactive reduces retention of about 10% -> OR= Exp(-0.10)=0.90→1-0.90=0.10

```
# Filter the DataFrame considering only the FIRST 3 DECILES
# not needed already done in previous step, but copy and paste here for
#illustrative purposes
mask = db_pilgrim[db_pilgrim['decile']<4]
```

```
# Define your formula
formula = ('retention ~ target_proactive + MainlyOnline_bank_previous +'
        'Tenure + age + agemiss + inc + incmiss + dist1100 + dist1200' )
# Estimate the model using the filtered DataFrame
model = smf.logit(formula, data=mask).fit()
# Print the model summary
print(model.summary2())
```

```
Optimization terminated successfully.
         Current function value: 0.604497
         Iterations 6
                             Results: Logit
=======================================================================
Model:                 Logit             Pseudo R-squared:   0.074
Dependent Variable:    retention         AIC:                11493.3463
Date:                  2023-10-15 22:11  BIC:                11564.9263
No. Observations:      9490              Log-Likelihood:     -5736.7
Df Model:              9                 LL-Null:            -6196.1
Df Residuals:          9480              LLR p-value:        5.4128e-192
Converged:             1.0000            Scale:              1.0000
No. Iterations:        6.0000
-----------------------------------------------------------------------
                            Coef.  Std.Err.    z    P>|z|   [0.025  0.975]
-----------------------------------------------------------------------
Intercept                  1.6036   0.2536  6.3242 0.0000  1.1066  2.1006
target_proactive          -0.1008   0.0449 -2.2458 0.0247 -0.1887 -0.0128
MainlyOnline_bank_previous 0.1994   0.0748  2.6646 0.0077  0.0527  0.3461
Tenure                     0.0594   0.0046 13.0154 0.0000  0.0505  0.0684
age                        0.0683   0.0659  1.0369 0.2998 -0.0608  0.1975
agemiss                   -1.0950   0.2096 -5.2248 0.0000 -1.5058 -0.6842
inc                        0.0730   0.0393  1.8573 0.0633 -0.0040  0.1500
incmiss                   -0.7889   0.1752 -4.5034 0.0000 -1.1323 -0.4456
dist1100                   0.0866   0.0899  0.9638 0.3352 -0.0895  0.2627
dist1200                   0.1210   0.0650  1.8594 0.0630 -0.0065  0.2484
=======================================================================
```

Step 4: Reflect on customer heterogeneity and try to explore.

Exploring customer heterogeneity means understand how different are the individuals within a sample -> whatever action we do will have a different impact on individuals because each of us is different from the other.
So this process means checking if there are groups of individuals that respond differently to our campaign.

```
[53]: # Step 4: Focus on Tenure [one can explore with other and more variables]
# Results of the Field test distinct by Tenure
# Descriptive Statistics
db_pilgrim[["AboveMedian_Tenure_Target", "Tenure"]].groupby("AboveMedian_Tenure_Target").agg({"Tenure": ["mean", "min", 'max', 'std', "count"]})
```

[53]:

	Tenure				
	mean	min	max	std	count
AboveMedian_Tenure_Target					
0	5.31706	0	9	2.382365	15117
1	13.85639	9	19	2.613972	16517

```
[60]: #Step 4
# Create sub-dataframe based on Tenure
mask_tenure_low = db_pilgrim[db_pilgrim['AboveMedian_Tenure_Target']==0]
mask_tenure_high = db_pilgrim[db_pilgrim['AboveMedian_Tenure_Target']==1]
```

```
[61]: #Logit only for those TENURE LOW
# Define your formula
formula = ('retention ~ target_proactive + MainlyOnline_bank_previous +'
        'Tenure + age + agemiss + inc + incmiss + dist1100 + dist1200' )
# Estimate the model using the filtered DataFrame
model = smf.logit(formula, data=mask_tenure_low).fit()
# Print the model summary
print(model.summary2())
```

```
Optimization terminated successfully.
         Current function value: 0.617853
         Iterations 7
                              Results: Logit
==========================================================================
Model:                   Logit              Pseudo R-squared:   0.081
Dependent Variable:      retention          AIC:                5883.4237
Date:                    2023-10-15 22:25   BIC:                5948.0721
No. Observations:        4745               Log-Likelihood:     -2931.7
Df Model:                9                  LL-Null:            -3190.6
Df Residuals:            4735               LLR p-value:        9.1360e-106
Converged:               1.0000             Scale:              1.0000
No. Iterations:          7.0000
--------------------------------------------------------------------------
                            Coef.  Std.Err.    z     P>|z|   [0.025  0.975]
--------------------------------------------------------------------------
Intercept                   1.2099   0.3563  3.3960 0.0007  0.5116  1.9081
target_proactive            0.1243   0.0625  1.9888 0.0467  0.0018  0.2468
MainlyOnline_bank_previous  0.0543   0.1014  0.5356 0.5923 -0.1444  0.2529
Tenure                      0.1115   0.0125  8.9393 0.0000  0.0870  0.1359
age                         0.1606   0.1015  1.5826 0.1135 -0.0383  0.3596
agemiss                    -1.0793   0.2966 -3.6388 0.0003 -1.6606 -0.4979
inc                         0.0753   0.0558  1.3479 0.1777 -0.0342  0.1847
incmiss                    -0.7588   0.2481 -3.0590 0.0022 -1.2450 -0.2726
dist1100                    0.1058   0.1264  0.8371 0.4025 -0.1419  0.3536
dist1200                    0.1710   0.0909  1.8815 0.0599 -0.0071  0.3492
==========================================================================
```

Tenure could explain a different response in terms of incentive because it divides the customer between short-term and long-term customers. For sure when talking about long-term customer we can consider inertia, which is something that we cannot consider when analysing short-term customers behavior. Moreover, some marketing actions cannot be proposed to short-term customers because it can be dangerous and vice versa.

To do this point we can act in three ways:

1. Descriptive statistic
2. Use interaction term
3. Divide the dataset in two sub-datasets: in this case low tenure (0<x<9) and high tenure (x>=9). And then create a regression in order to understand the impact of this variable on retention.

TENURE LOW: Positive impact of target_proactive of 13% on retention OR=exp(0.1209)=1.13
Before, when we analyzed the entire sample the effect was negative (reduction of 10%).

```
#Logit only for those TENURE HIGH
# Define your formula
formula = ('retention ~ target_proactive + MainlyOnline_bank_previous +'
           'Tenure + age + agemiss + inc + incmiss + dist1100 + dist1200' )
# Estimate the model using the filtered DataFrame
model = smf.logit(formula, data=mask_tenure_high).fit()
# Print the model summary
print(model.summary2())
```

```
Optimization terminated successfully.
         Current function value: 0.582142
         Iterations 6
                          Results: Logit
=====================================================================
Model:                 Logit             Pseudo R-squared:  0.071
Dependent Variable:    retention         AIC:               5544.5229
Date:                  2023-10-15 22:26  BIC:               5609.1714
No. Observations:      4745              Log-Likelihood:    -2762.3
Df Model:              9                 LL-Null:           -2973.4
Df Residuals:          4735              LLR p-value:       2.3145e-85
Converged:             1.0000            Scale:             1.0000
No. Iterations:        6.0000
---------------------------------------------------------------------
                            Coef.  Std.Err.    z    P>|z|   [0.025  0.975]
---------------------------------------------------------------------
Intercept                   0.9013   0.4030  2.2364 0.0253  0.1114  1.6912
target_proactive           -0.3431   0.0653 -5.2508 0.0000 -0.4712 -0.2150
MainlyOnline_bank_previous  0.3565   0.1136  3.1381 0.0017  0.1338  0.5791
Tenure                      0.1138   0.0130  8.7456 0.0000  0.0883  0.1393
age                         0.0127   0.0886  0.1436 0.8858 -0.1609  0.1863
agemiss                    -1.0005   0.3035 -3.2966 0.0010 -1.5953 -0.4056
inc                         0.0763   0.0552  1.3825 0.1668 -0.0319  0.1845
incmiss                    -0.8254   0.2480 -3.3284 0.0009 -1.3115 -0.3394
dist1100                    0.0542   0.1294  0.4185 0.6756 -0.1995  0.3078
dist1200                    0.0680   0.0944  0.7206 0.4712 -0.1170  0.2531
=====================================================================
```

TENURE HIGH: Negative Impact of target_proactive on Retention -29% Why? -> «Broken Inertia»

```
#how to ask odds ratio as output

# Access the model parameters (coefficients)
model_params = model.params

# Compute the odds ratios (exponentiate the coefficients)
odds_ratios = np.exp(model_params)

# Create a DataFrame to display the odds ratios with their respective names
odds_ratios_df = pd.DataFrame({'Odds Ratio': odds_ratios, 'Coefficient': model_params})

# Print the DataFrame with odds ratios
print(odds_ratios_df)
```

```
                            Odds Ratio  Coefficient
Intercept                     3.352995     1.209854
target_proactive              1.132372     0.124314
MainlyOnline_bank_previous    1.055781     0.054281
Tenure                        1.117935     0.111484
age                           1.174261     0.160639
agemiss                       0.339847    -1.079259
inc                           1.078157     0.075253
incmiss                       0.468218    -0.758821
dist1100                      1.111616     0.105815
dist1200                      1.186546     0.171046
```

Step 5: Comment on the results of your explorations.

WRAP UP:

- A targeted proactive action aimed at reducing churn actually has a negative effect on retention, reducing it overall by about 10%.
- By examining different response probabilities for various customer groups, it is noted that tenure moderates the effect; those with low tenure respond positively, while long-term customers respond negatively.
- The inertia effect may explain the result, in line with the literature by Ascarza, Iyengar, and Schleicher (2016) titled 'The Wrong Way to Reduce Churn' in Idea Watch, Harvard Business Review.

Take Home:

- The probability of churn and the probability of response to a specific marketing actionshould be considered in conjunction!

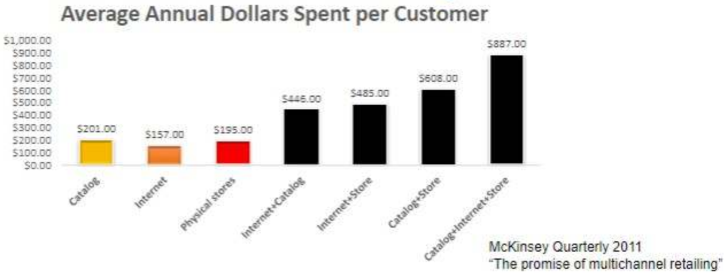
- *Randomized field tests if possible and/or the development of marketing response models combined with churn probabilities*

UNTARGETED CHURN REDUCTION STRATEGY – CASE 4: BOOKS R US -> FIELD TEST

In this case we just want to reduce the churn rate by applying actions that we do believe that couldhave a positive impact on the likelihood to churn.

Channels, touchpoints and profits...

Do multichannel customers buy more?



Replicated by: Loftus, Mulliken and Sharp 2008; Myers, Pickersgill, and van Metre 2004; Thomas and Sullivan 2005; Kumar and Venkatesan 2005; Venkatesan, Kumar, and Ravishanker 2007; Ansari, Mela, and Neslin 2008; Boehm 2008; Campbell and Frei 2010; Xue, Hitt, and Chen 2011; Gensler, Leeflang, & Skiera 2012; Kushwaha and Shankar 2013, Montaguti, Neslin, Valentini 2016.

MANAGERIAL IMPLICATION: firms should use marketing to create more multichannel customersLogic,

- Marketing induces more customers to become multichannel.
- Multichannel customers are more profitable than they would have been had they not been multichannel.
- As a result, average profit per customer should increase.

Is this **ACTIONABLE**?

Case 4 BooksRUs: Summary

Can marketing actions create more multichannel customers?
Will this increase the profitability of the customer base?
If yes, how?
—By reducing churn?
—By increasing spending volume?
—By increasing purchase frequency?

Case 4 BooksRUs:
Can Marketing Induce Multichannel Buying and More Profitable Customers?

"Untargeted strategy"

— BookRus is a major multichannel European book retailer. The company sells books through stores, mail-order, phone and the Internet.
— The firm mails a print catalog to the customer base five times per year
— The company opted for an omnichannel strategy providing members several possibilities to purchase: the physical stores, the website, mail/order or phone, the app
— Despite the efforts of providing more channels to purchase the core customer base was still strongly related to the physical store without exploring other available channels (70% of customers purchased mainly using only the Physical Store)
— Additionally. the management wasn't sure about how much effective an omnichannel strategy could be in increasing average customer profitability

In doing this, **Roberta is attempting to implement an "UNTARGETED" strategy** aimed at reducing churn and increasing customer base profitability. The approach is indirect; therefore, the campaignand field testneedtobecarefullydesigned.

CASE OBJECTIVES

BooksRUsimplementafieldexperimenttoaddressfourquestions:

1. Canamarketingcampaignbedesignedtocreatemoremultichannelcustomers?
2. Ifso,aremultichannelcustomersmoreprofitablethantheywouldhavebeenhadtheynotbeen multichannel?
3. Whatistheimpactonchurn?
4. What types of marketing campaigns work best, and why?

DATA

Thecompanyselectedon3cohortsofcustomerswholivedwithinatleastonestore'sserviceareaand were acquired in the last period of the year (September – December): Cohort 1, Cohort 2 and Cohort3. Allcustomersincludedinthesecohortwereobservedsincetheveryfirstpurchase:

Cohorts1and2wereusedtotestifmultichannelcustomersareassociatedwithhigherprofitsandless churn.Cohort3wasselectedtoconductthefieldtest

Before running the Field Test Roberta wanted to have a first correlational empirical evidence that the use of multiple channels was associated with more profits. She used data of Cohort 1 to conduct thisanalysis.

Phase 0 - Are Multichannel Customers More Profitable Customers?

Phase 0 - Are Multichannel Customers More Profitable Customers?

Case 2 Results

Results: Ordinary least squares

Model:	OLS	Adj. R-squared:	0.336
Dependent Variable:	profits	AIC:	287829.0750
Date:	2022-05-16 07:26	BIC:	287973.1367
No. Observations:	35391	Log-Likelihood:	-1.4390e+05
Df Model:	16	F-statistic:	1122.
Df Residuals:	35374	Prob (F-statistic):	0.00
R-squared:	0.337	Scale:	199.23

	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
multichannel	20.9517	0.3207	65.3373	0.0000	20.3231	21.5802
age	0.1509	0.0048	31.4807	0.0000	0.1415	0.1603
female	1.2085	0.1623	7.4482	0.0000	0.8905	1.5265
street_agent	3.1672	0.1585	19.9766	0.0000	2.8565	3.4780
north	1.9387	0.2022	9.5889	0.0000	1.5424	2.3349
early_email	1.8377	0.1644	11.1769	0.0000	1.5154	2.1599
bigcity	0.1535	0.2579	0.5952	0.5517	-0.3519	0.6589
mean_city	0.0237	0.2711	0.0873	0.9304	-0.5076	0.5550
franchisee	0.4522	0.1596	2.8340	0.0046	0.1394	0.7649
initialstorepromo	-1.8296	0.3417	-5.3540	0.0000	-2.4994	-1.1598
initialweb	11.0348	0.7429	14.8539	0.0000	9.5788	12.4909
initialstore	-2.5538	0.3014	-8.4724	0.0000	-3.1446	-1.9630
initialmobile	10.0674	0.5492	18.3326	0.0000	8.9910	11.1438
initialrevenues	0.5308	0.0086	61.5983	0.0000	0.5139	0.5476
initialreturns	-0.5039	0.0541	-9.3126	0.0000	-0.6099	-0.3978
initialreturns2	0.0021	0.0008	2.7236	0.0065	0.0006	0.0036
const	4.6034	0.3116	14.7718	0.0000	3.9926	5.2142

Phase 0 - Are Multichannel Customers More Profitable Customers?

Case 2 Results

Results: Ordinary least squares

Model:	OLS	Adj. R-squared:	0.336
Dependent Variable:	profits	AIC:	287829.0750
Date:	2022-05-16 07:26	BIC:	287973.1367
No. Observations:	35391	Log-Likelihood:	-1.4390e+05
Df Model:	16	F-statistic:	1122.
Df Residuals:	35374	Prob (F-statistic):	0.00
R-squared:	0.337	Scale:	199.23

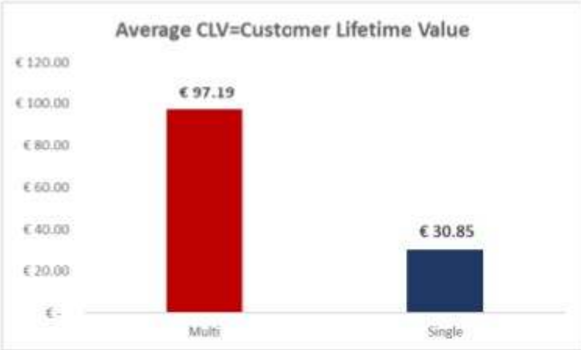
	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
multichannel	20.9517	0.3207	65.3373	0.0000	20.3231	21.5802
age	0.1509	0.0048	31.4807	0.0000	0.1415	0.1603
female	1.2085	0.1623	7.4482	0.0000	0.8905	1.5265
street_agent	3.1672	0.1585	19.9766	0.0000	2.8565	3.4780
north	1.9387	0.2022	9.5889	0.0000	1.5424	2.3349
early_email	1.8377	0.1644	11.1769	0.0000	1.5154	2.1599
bigcity	0.1535	0.2579	0.5952	0.5517	-0.3519	0.6589
mean_city	0.0237	0.2711	0.0873	0.9304	-0.5076	0.5550
franchisee	0.4522	0.1596	2.8340	0.0046	0.1394	0.7649
initialstorepromo	-1.8296	0.3417	-5.3540	0.0000	-2.4994	-1.1598
initialweb	11.0348	0.7429	14.8539	0.0000	9.5788	12.4909
initialstore	-2.5538	0.3014	-8.4724	0.0000	-3.1446	-1.9630
initialmobile	10.0674	0.5492	18.3326	0.0000	8.9910	11.1438
initialrevenues	0.5308	0.0086	61.5983	0.0000	0.5139	0.5476
initialreturns	-0.5039	0.0541	-9.3126	0.0000	-0.6099	-0.3978
initialreturns2	0.0021	0.0008	2.7236	0.0065	0.0006	0.0036
const	4.6034	0.3116	14.7718	0.0000	3.9926	5.2142

PHASE 0 - ARE MULTICHANNEL CUSTOMERS MORE PROFITABLE CUSTOMERS? CLV PERSPECTIVE

She also computed the CLV using this (Cohort 1) :

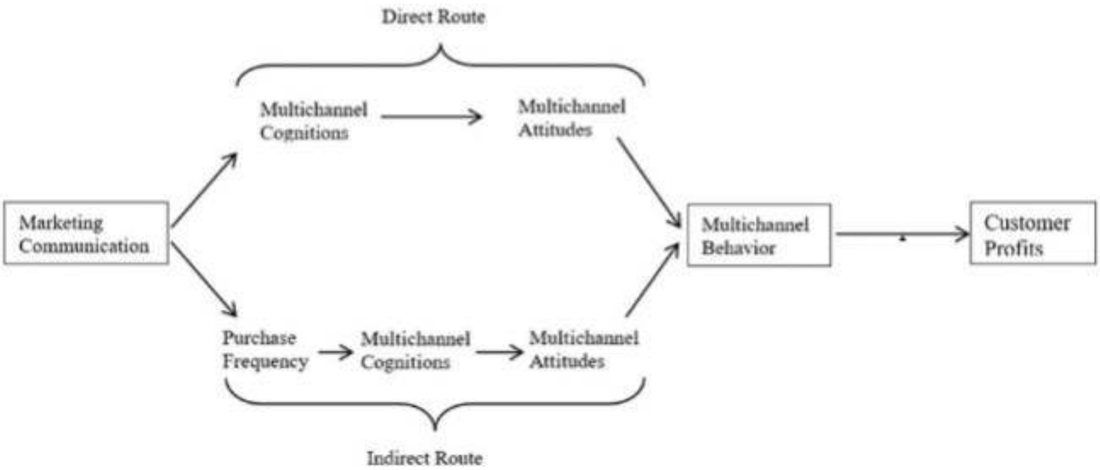
$$CLV = m\frac{(1+d)}{(1+d-r)} - AC$$

Discount rate=0.12
r=retention rate=1-churn rate
d=discount rate
AC=acquisition cost= 10 €



Design of the field test

This figure depicts how marketing communication can induce customers to become multichannel shoppers who in turn are more profitable.



ACTIONABILITY

FOUR CAMPAIGNS: IS MULTICHANNEL CUSTOMER STRATEGY ACTIONABLE?

IMPLEMENTATION OF THE FIELD TEST

Group	Number of customers
MNF	6810
MF	6831
VPNF	6829
VPF	6821
Control	3419
	30710

Caso BooksRUs: Field Test

Step 0: Open the dataset Cohort3.xls.

```
import pandas as pd
import statsmodels.api as sm
import statsmodels.formula.api as smf
import numpy as np
import researchpy as rp
import os
from statsmodels.stats.proportion import proportions_ztest
```

```
#Step 0:
#Open dataset  Cohort3BooksFieldTest.xls
file_path = r'C:\Users\ValentiniS\Cohort3BooksFieldTest.xls'
df = pd.read_excel(file_path)
df.head()
```

	id	groups	profits	churn_observed	multichannel	mf	mnf	vpf	vpnf	c	...	mean_city	early_email	franchisee	street_agent	initialweb	initialstore	initialmol
0	1	Control	39.040001	0	0	0	0	0	0	1	...	0	1	0	0	0	0	
1	2	mnf	53.080002	0	1	0	1	0	0	0	...	0	1	0	0	1	0	
2	3	vpnf	0.000000	0	0	0	0	0	1	0	...	0	1	1	0	1	0	
3	4	mnf	56.049999	1	1	0	1	0	0	0	...	0	1	1	0	1	0	
4	5	Control	9.290000	0	0	0	0	0	0	1	...	0	1	1	0	0	0	

5 rows × 24 columns

```
df.shape
```

```
(30710, 24)
```

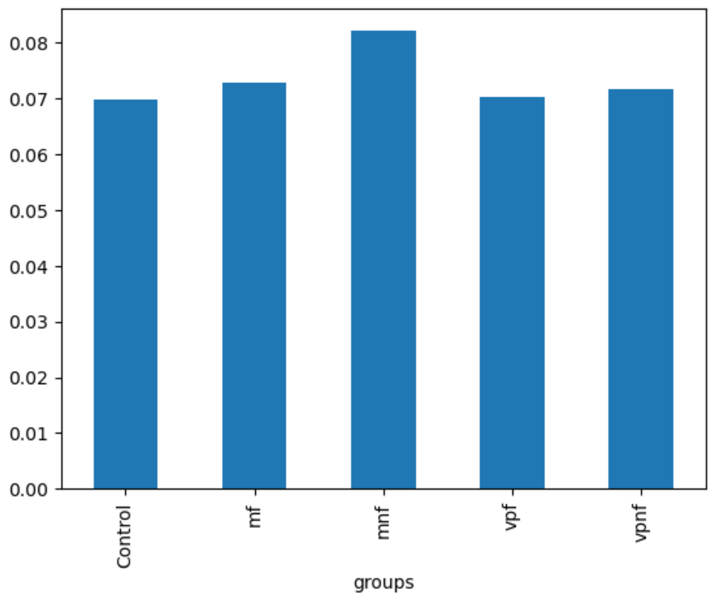
Step 1: Model Free Evidence: Check if and which of the campaigns

```
# Step 1: Model Free Evidence: Check if and which of the campaigns generated the most multichannel customers
df.groupby("groups")['multichannel'].mean()
```

```
groups
Control    0.069903
mf         0.072757
mnf        0.082085
vpf        0.070224
vpnf       0.071753
Name: multichannel, dtype: float64
```

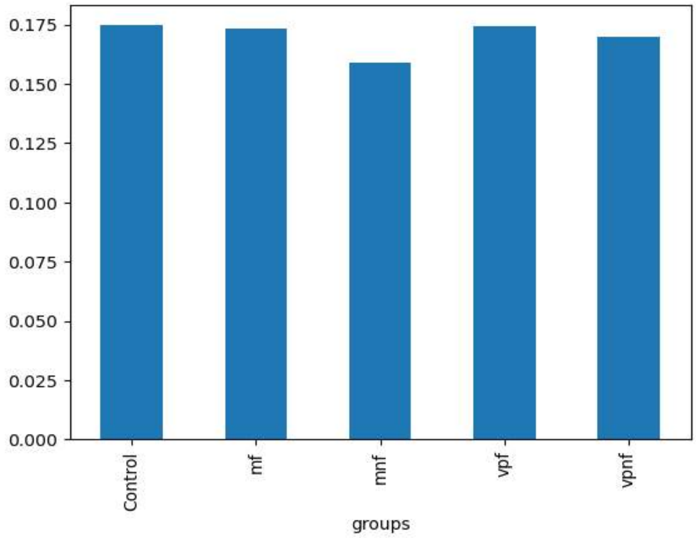
```
# Step 1: Model Free Evidence: chart
df.groupby("groups")['multichannel'].mean().plot.bar()
```

```
<AxesSubplot:xlabel='groups'>
```



```
#impact on churn: model-free evidence
df.groupby("groups")['churn_observed'].mean().plot.bar()
```

<AxesSubplot:xlabel='groups'>

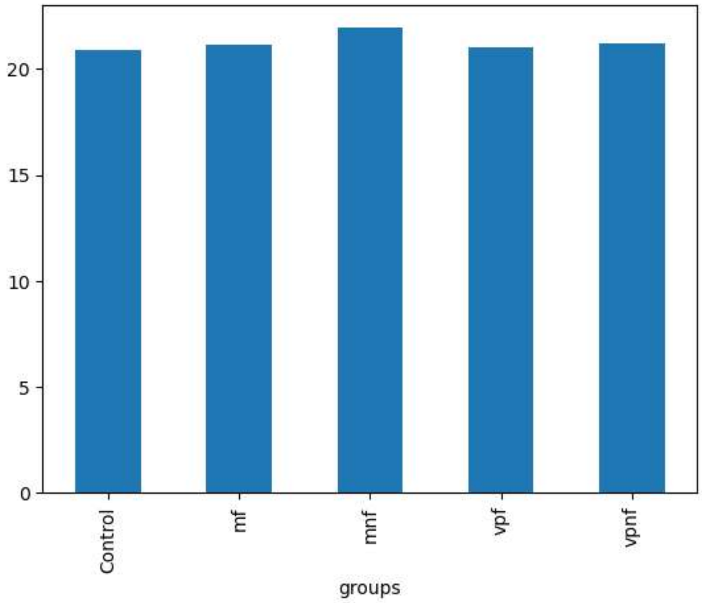



```
#impact on churn
df.groupby("groups")['churn_observed'].mean()
```

```
groups
Control    0.174612
mf         0.173327
mnf        0.158737
vpf        0.174168
vpnf       0.170010
Name: churn_observed, dtype: float64
```

```
#impact on profits
df.groupby("groups")['profits'].mean().plot.bar()
```

<AxesSubplot:xlabel='groups'>




```
#impact on profits
df.groupby("groups")['churn_observed'].mean()
```

```
groups
Control    0.174612
mf         0.173327
mnf        0.158737
vpf        0.174168
vpnf       0.170010
Name: churn_observed, dtype: float64
```

Step 2: Statistical Test: Test if and which of the campaigns generated the most multichannel customers.

```python
# Step 2: Results of the Field test: TEST
formula = ('multichannel ~ mf + mnf + vpf + vpnf')

model = smf.logit(formula, data = df).fit()

print(model.summary2())
```

```
Optimization terminated successfully.
         Current function value: 0.263012
         Iterations 7
                         Results: Logit
=================================================================
Model:              Logit            Pseudo R-squared: 0.001
Dependent Variable: multichannel     AIC:              16164.1787
Date:               2023-10-16 00:35 BIC:              16205.8404
No. Observations:   30710            Log-Likelihood:   -8077.1
Df Model:           4                LL-Null:          -8081.7
Df Residuals:       30705            LLR p-value:      0.056010
Converged:          1.0000           Scale:            1.0000
No. Iterations:     7.0000
------------------------------------------------------------------
              Coef.   Std.Err.     z      P>|z|   [0.025   0.975]
------------------------------------------------------------------
Intercept   -2.5882    0.0671  -38.5884  0.0000  -2.7196  -2.4567
mf           0.0431    0.0817    0.5275  0.5979  -0.1170   0.2031
mnf          0.1738    0.0803    2.1648  0.0304   0.0164   0.3312
vpf          0.0049    0.0821    0.0600  0.9522  -0.1560   0.1659
vpnf         0.0281    0.0818    0.3434  0.7313  -0.1323   0.1885
=================================================================
```

Remember to exclude the control group.

```python
# Step 2: Results of the Field test: test
###### CHURN
formula = ('churn_observed ~ mf + mnf + vpf + vpnf')

model = smf.logit(formula, data = df).fit()

print(model.summary2())
```

```
Optimization terminated successfully.
         Current function value: 0.455253
         Iterations 6
                         Results: Logit
=================================================================
Model:              Logit            Pseudo R-squared: 0.000
Dependent Variable: churn_observed   AIC:              27971.6567
Date:               2023-10-16 00:36 BIC:              28013.3185
No. Observations:   30710            Log-Likelihood:   -13981.
Df Model:           4                LL-Null:          -13985.
Df Residuals:       30705            LLR p-value:      0.088281
Converged:          1.0000           Scale:            1.0000
No. Iterations:     6.0000
------------------------------------------------------------------
              Coef.   Std.Err.     z      P>|z|   [0.025   0.975]
------------------------------------------------------------------
Intercept   -1.5533    0.0450  -34.4800  0.0000  -1.6416  -1.4650
mf          -0.0089    0.0552   -0.1619  0.8714  -0.1172   0.0993
mnf         -0.1144    0.0559   -2.0446  0.0409  -0.2240  -0.0047
vpf         -0.0031    0.0552   -0.0559  0.9554  -0.1113   0.1051
vpnf        -0.0323    0.0554   -0.5827  0.5601  -0.1408   0.0763
=================================================================
```

The effect should be indirect -> we want to observe if the newly created multichannel customer are now able to reduce churn and improve profits.

```python
# Step 2: Results of the Field test: Test
###### profits

formula = ('profits ~ mf + mnf + vpf + vpnf')

model = smf.ols(formula, data = df).fit()

print(model.summary2())
```

```
                 Results: Ordinary least squares
=================================================================
Model:              OLS              Adj. R-squared:     0.000
Dependent Variable: profits          AIC:                273197.3616
Date:               2023-10-16 00:36 BIC:                273239.0233
No. Observations:   30710            Log-Likelihood:     -1.3659e+05
Df Model:           4                F-statistic:        2.319
Df Residuals:       30705            Prob (F-statistic): 0.0546
R-squared:          0.000            Scale:              427.52
-------------------------------------------------------------------
              Coef.   Std.Err.     t      P>|t|    [0.025   0.975]
-------------------------------------------------------------------
Intercept    20.8947    0.3536  59.0891  0.0000  20.2016  21.5878
mf            0.2598    0.4332   0.5998  0.5486  -0.5892   1.1088
mnf           1.0367    0.4334   2.3921  0.0168   0.1872   1.8861
vpf           0.1476    0.4333   0.3407  0.7333  -0.7016   0.9968
vpnf          0.2800    0.4332   0.6464  0.5180  -0.5690   1.1291
-------------------------------------------------------------------
Omnibus:              5182.527      Durbin-Watson:         1.840
Prob(Omnibus):        0.000         Jarque-Bera (JB):      10331.001
Skew:                 1.029         Prob(JB):              0.000
Kurtosis:             4.959         Condition No.:         8
=================================================================
```

Step 3: Test if and which of the 4 campaigns had an effect on churn, and reflect on which variables to include in the test.

```
#check if multichannl impacts Churn

formula = ('churn_observed ~ multichannel')

model = smf.logit(formula, data = df).fit()

print(model.summary2())
```

```
Optimization terminated successfully.
         Current function value: 0.450560
         Iterations 7
                         Results: Logit
=================================================================
Model:              Logit            Pseudo R-squared: 0.011
Dependent Variable: churn_observed   AIC:              27677.4182
Date:               2023-10-16 00:38 BIC:              27694.0829
No. Observations:   30710            Log-Likelihood:   -13837.
Df Model:           1                LL-Null:          -13985.
Df Residuals:       30708            LLR p-value:      2.0765e-66
Converged:          1.0000           Scale:            1.0000
No. Iterations:     7.0000
-----------------------------------------------------------------
                Coef.  Std.Err.    z     P>|z|   [0.025  0.975]
-----------------------------------------------------------------
Intercept      -1.5238   0.0155 -98.5023 0.0000 -1.5541 -1.4935
multichannel   -1.3417   0.0944 -14.2192 0.0000 -1.5266 -1.1567
=================================================================
```

Being multichannel is associated with less churn and higher profits.

Step 4: Test if and which of the 4 campaigns had an effect on profits, and reflect on which variables to include in the test.

```
#check if multichannl impacts  profits

formula = ('profits ~ multichannel')

model = smf.ols(formula, data = df).fit()

print(model.summary2())
```

```
                 Results: Ordinary least squares
=================================================================
Model:              OLS              Adj. R-squared:     0.169
Dependent Variable: profits          AIC:                267496.2869
Date:               2023-10-16 00:38 BIC:                267512.9516
No. Observations:   30710            Log-Likelihood:     -1.3375e+05
Df Model:           1                F-statistic:        6268.
Df Residuals:       30708            Prob (F-statistic): 0.00
R-squared:          0.170            Scale:              355.12
------------------------------------------------------------------
                Coef.   Std.Err.     t      P>|t|   [0.025   0.975]
------------------------------------------------------------------
Intercept      18.8756   0.1117  168.9365  0.0000  18.6566  19.0946
multichannel   32.5797   0.4115   79.1714  0.0000  31.7731  33.3863
-----------------------------------------------------------------
Omnibus:             4764.087      Durbin-Watson:          1.840
Prob(Omnibus):       0.000         Jarque-Bera (JB):       8454.467
Skew:                1.006         Prob(JB):               0.000
Kurtosis:            4.599         Condition No.:          4
=================================================================
```

*** please note you can control also for other variables but result would not change dramatically, since it is a result of a randomized field test

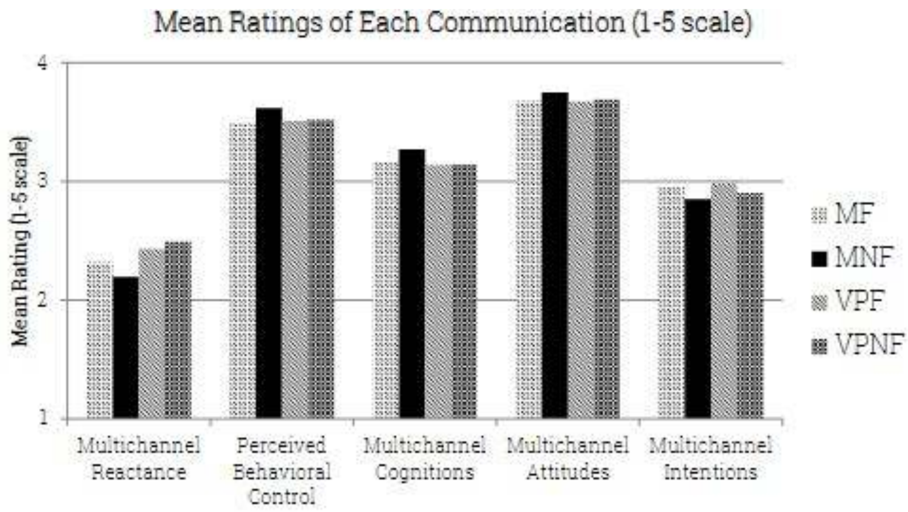
```
#check if multichannl impacts Churn and profits

formula = ('churn_observed ~ multichannel + age + '
           'north + female +'
           'bigcity + mean_city + early_email +'
           'franchisee + street_agent +'
           'initialweb + initialstore + initialmobile + initialstorepromo +'
           'initialreturns + initialrevenues' )

model = smf.logit(formula, data = df).fit()

print(model.summary2())
```

```
Optimization terminated successfully.
         Current function value: 0.430332
         Iterations 7
                         Results: Logit
=================================================================
Model:              Logit            Pseudo R-squared: 0.055
Dependent Variable: churn_observed   AIC:              26462.9731
Date:               2023-10-16 00:44 BIC:              26596.2906
No. Observations:   30710            Log-Likelihood:   -13215.
Df Model:           15               LL-Null:          -13985.
Df Residuals:       30694            LLR p-value:      0.0000
Converged:          1.0000           Scale:            1.0000
No. Iterations:     7.0000
-----------------------------------------------------------------
                   Coef.  Std.Err.    z     P>|z|   [0.025  0.975]
-----------------------------------------------------------------
Intercept         -0.6497   0.0683  -9.5120 0.0000 -0.7836 -0.5159
multichannel      -1.3793   0.0962 -14.3423 0.0000 -1.5678 -1.1908
age               -0.0163   0.0011 -14.9638 0.0000 -0.0184 -0.0142
north              0.2890   0.0413   6.9938 0.0000  0.2080  0.3700
female            -0.5703   0.0327 -17.4175 0.0000 -0.6345 -0.5061
bigcity            1.0623   0.0509  20.8806 0.0000  0.9626  1.1620
mean_city          0.2255   0.0734   3.0724 0.0021  0.0817  0.3694
early_email       -0.0667   0.0339  -1.9676 0.0491 -0.1332 -0.0003
franchisee        -0.0674   0.0333  -2.0232 0.0431 -0.1327 -0.0021
street_agent      -0.4389   0.0332 -13.2206 0.0000 -0.5040 -0.3739
initialweb        -0.1024   0.1239  -0.8260 0.4088 -0.3453  0.1406
initialstore      -1.2206   0.6032  -2.0234 0.0430 -2.4030 -0.0383
initialmobile     -0.3960   0.1319  -3.0022 0.0027 -0.6545 -0.1375
initialstorepromo  1.0329   0.6041   1.7100 0.0873 -0.1510  2.2169
initialreturns    -0.0080   0.0057  -1.3911 0.1642 -0.0192  0.0033
initialrevenues    0.0099   0.0011   9.2757 0.0000  0.0078  0.0120
=================================================================
```

Step 5: Which campaign worked better? Why?

MNF (reactance theory) is the campaign which worked better; the value proposition options don't tell customers the presence of multichannel. Discount means that there is something for you, but in this case, MF force you to purchase, but also on which channels do the purchase -> it is really manipulating and it have a negative effect on customers.

Theory of planned behavior -> perceived behavioral control is associated to an increase in positive attitude toward an action. In this context this concept can be reassumed in the fact that just the
knowledge of the presence of multiple channels is helpful for our purpose.



- MNF produce less "reactance" and increased "perceived control"!

Multichannel customers are more profitable: further analyses

Industry: Books

	Average annual Profit per customer, distinct by frequency of purchase								
Number of Purchase occasions	2	3	4	5	6	7	8	9	>9
Single Channel	€20.28	€29.18	€36.27	€37.68	€34.14	€36.26	€38.03	€43.20	€50.11
Multichannel	€23.36	€37.05	€45.14	€52.92	€59.96	€60.37	€57.34	€67.97	€63.34
Difference	€3.08	€7.87	€8.87	€15.23	€25.83	€24.01	€19.31	€24.77	€13.23
p-value	0.006	0.000	0.000	0.000	0.000	0.000	0.003	0.105	0.173

margins...

Margins	Multichannel Purchases	Non-Multichannel Purchases
Store	31%	79%
Internet, Mail, Mobile	69%	21%

However, TT > 0 even for combinations of high margin channels

Multichannel	Single Channel	TT=Difference in Profits
Internet/Mobile	Internet	€8.52
Internet/Mobile	Mobile	€17.54
Internet/Mobile	Mail Order	€ 9.64
Internet/Mail Order	Internet	€12.31
Internet/Mail Order	Mail	€13.64
Internet/Mail Order	Mobile	€23.34
Phone/Mail Order	Mobile	€18.00
Phone/Mail Order	Mail	€8.00
Phone/Mail Order	Internet	€7.23

Post-test survey conducted on the company's customers: 2,068 respondents -> Lab Experiment: each respondent was exposed to only one of the four communications using an experimental logic

FOR DOUBTS OR SUGGESTIONS ON THE HANDOUTS

CHIARA TUA

chiara.tua@studbocconi.it
@chiara_tua
+39 3479789059

GABRIELE CARDINALE

vittoria.nasonte@studbocconi.it
@_vittorian_
+39 3274441476

FOR INFO ON THE TEACHING DIVISION

MARCO FORMISANO

marco.formisano@studbocconi.it
@marco_formisano__
+39 3313433934

ELENA CACIOLI

elena.cacioli@studbocconi.it
@elenaciolii_
+39 3928931605

TEACHING DIVISION

OUR PARTNERS

70C+ CLUB

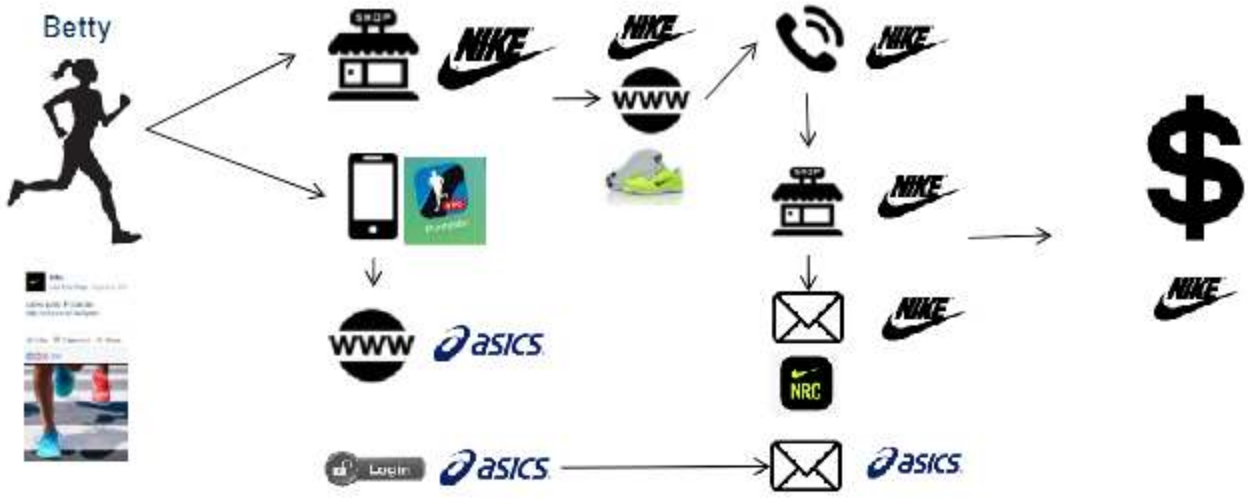
ETHAN SUSTAINABILITY

DELIVERY VALLEY
NO GENDER KITCHEN

LA PIADINERIA

STRATEGIC MARKETING & ANALYTICS
DATA & ANALYTICS FOR STRATEGIC MARKETING DECISIONS

Setting the scene



Betty is a runner, and she might be exposed on some target adv, through a post as, for example, buy a new running pair of shoes.
The post triggers her attention because she is considering buying a new pair of running shoes, so she goes walk to the city centre, enters the Nike store and try's them on, but she is not sure if they are fit or not, so she postponed thinking about it.
In the meanwhile, she is using a free run activity tracker in which she is exposed to an Asis adv, which proposes to register to the newsletter in order to have a discount on the first purchase.
After a while she takes her decision opting for a Nike pair of shoes but not at the Nike store but online cause she can customize them by choosing the color and so on.
She receives the shoes, but the color is not what she ordered, so she phones the customer service saying that there is an issue with her order.
The customer service fixes the problem and asks Betty if her want to receive the new pair at home or if she prefers to collect it at the store. Betty decides to pick them to the store.
In the meanwhile, Asis keep emailing her with possible new offers, but at the end the transaction is for Nike.

What this example tells us?
This is the description of the decision process -> when we buy something new, we pass through different phases:
1) search for alternatives
2) purchase
3) evaluation of your purchase
-> this is called **customer journey** which is the decision process from the very beginning until the end.

Company can track each phase of the customer journey, even the searching initial phase. Customers can be tracked at the same time from different company. Acquire the ability to *analyze data* to *answer relevant* *marketing questions* and *support decision* is a fundamental part of marketing manager work. Data and analytics are used in marketing to try to create efficient and effective marketing strategies.

Which marketing question when analysing data?

Distribution

How do channels impact a brand's choice and customer decision-making?

Brand Choice

How can we increase the likelihood that brand x is chosen?

Consumer Analysis and Targeting

Who are our customers?
Who is most likely to purchase?
Which product variant are they most likely to buy?

Value of the Customer Base

Are our customers profitable? Are our customers loyal?

How can we increase the profitability of the customer base?

Marketing Actions

Can discounts effectively lead new customers to make their initial purchase?

Why do we need analytics in marketing?

Statistic helps managers in **dealing with uncertainty** -> managers can make smart decisions and lead staff more effectively.
Data and analytics reduce uncertainty -> if I'm a manager I don't want to risk, so if I can provide empirical evidence, it will be so much better; moreover, you need to test your idea -> in this way is highly recommend use data.

Prior: Managers' intuition Decision

Decisions based on managers intuition.
The management provides preliminary estimates (prior) about the probability that the event will occur following a strategy.

Prior = The management expects a 10% increase in
purchases by customers with loyalty cards through the email and smartphone free shipping campaign.
Pay attention to:

- Avaiability bias
- Overconfidence

In statistics (Bayesian theory) Prior.
Refers to the initial estimate or probability one has regarding an event
before considering new information or data

Posterior: Data

Decisions Based on Data Analysis.
We can gather information and use it to refine the preliminary (prior) estimates regarding the probability that an event will occur.
Collect Data > Posterior

Posterior= After analyzing data, we estimate that 5% of loyalty card customers will respond positively to the free shipping campaign promoted via email and smartphone notifications.

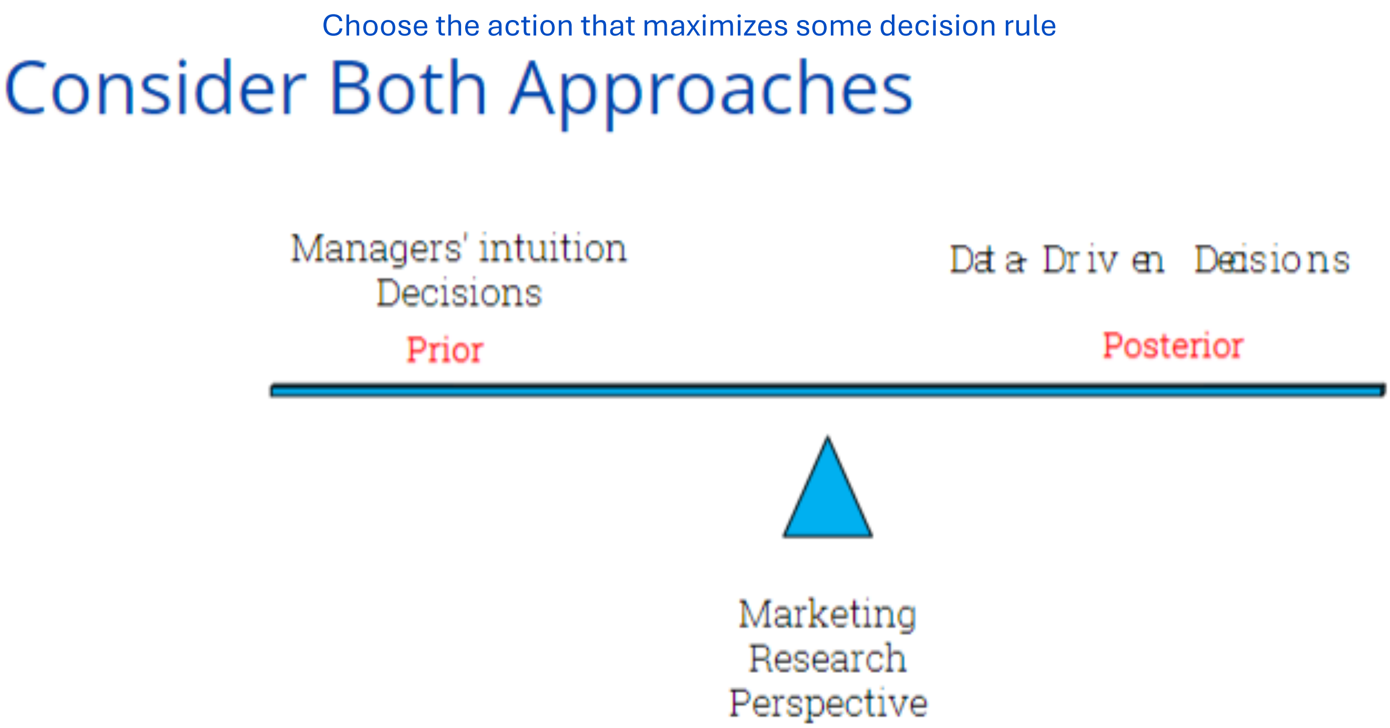
In statistics (Bayesian Theory) Posterior:
The updated probability of an event after considering new data or information.

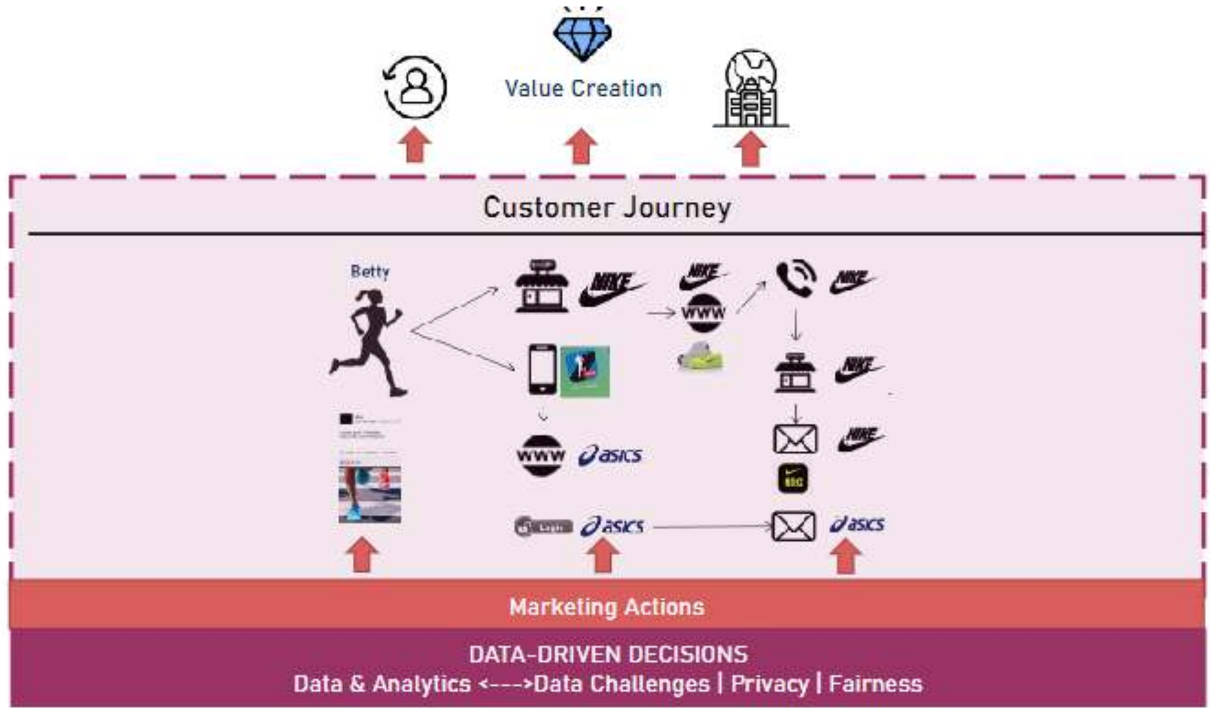
Value of analytics in marketing

- You can collect information that is used to **refine** the prior estimate of the probability that an event occurs.

  Prior -> Collect Data> Posterior
  [Prior you get for "free"; and maybe it's enough]
- Statistics should reduce uncertainty associated with predicting future events.

  Prior(variance) > Posterior(Variance)
- Reliability of information is not ignored
- Statistic should help you to quantify the consequences of a planned business action.





<u>The final aim is to create value for both the company and the customers.</u>

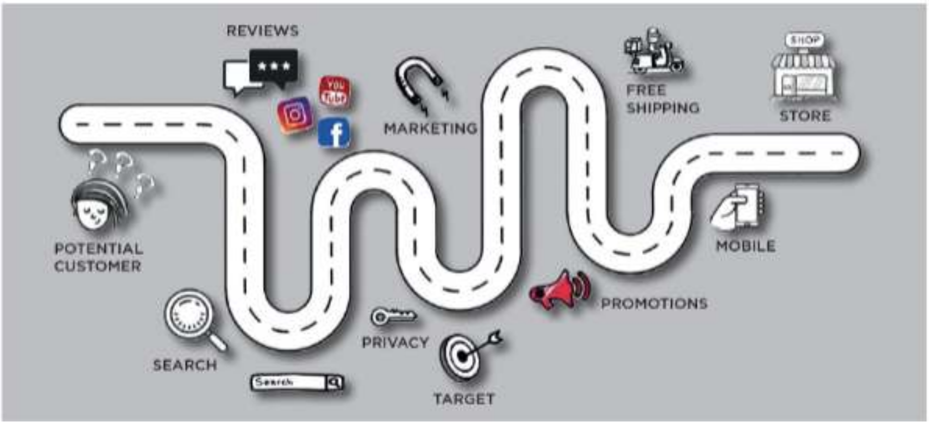
Big data & customer journey

Data in Marketing: What data?

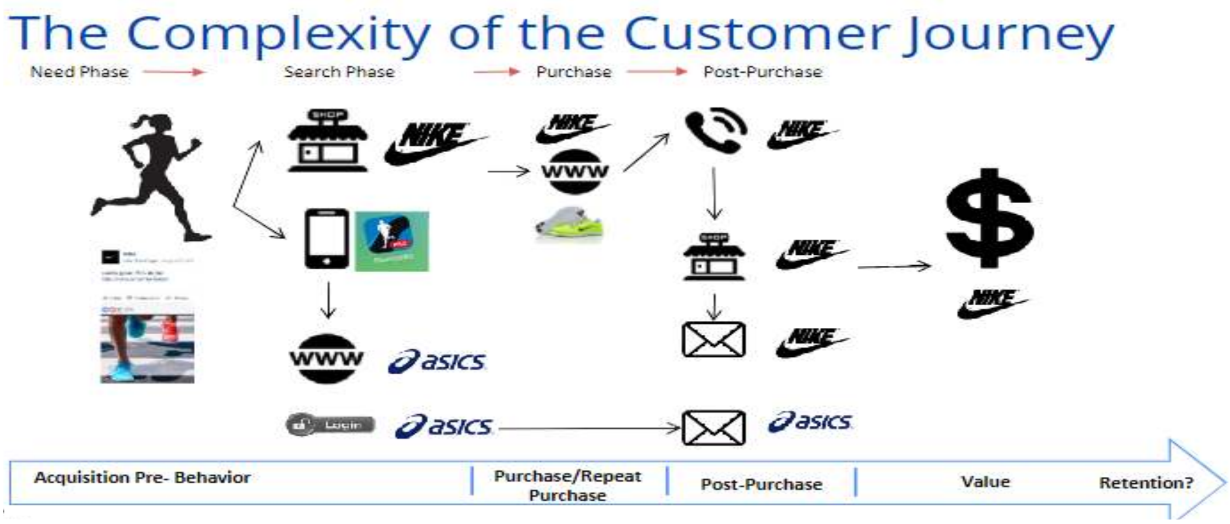
Making decisions about the types of data is a "measurement problem".
We can have 2 distinct macro classes of data:

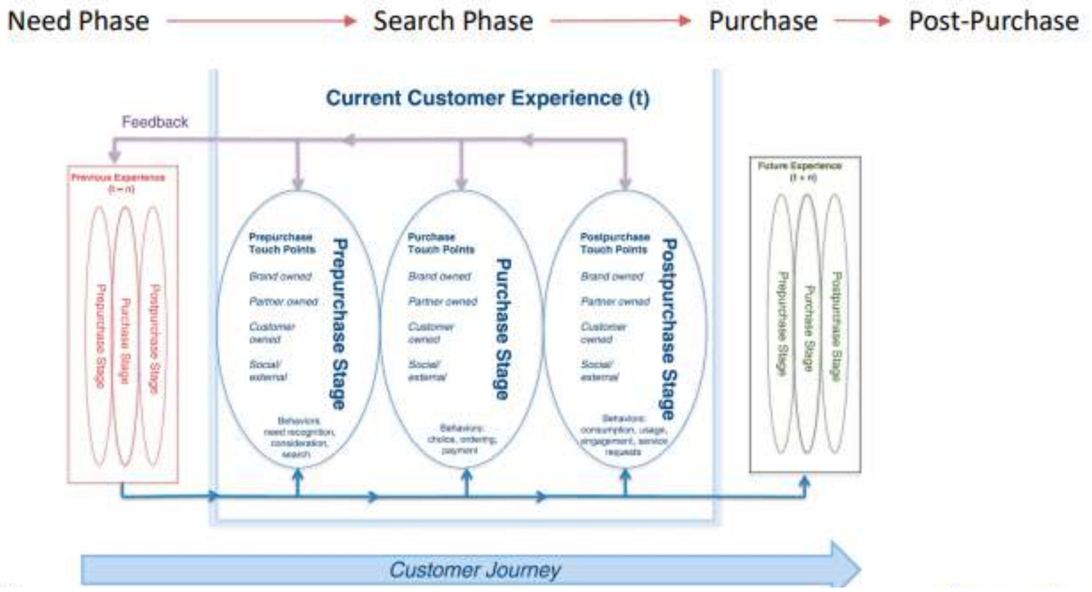
- **Primary data**: can be collected in 2 different ways and for a specific purpose (and this is the main difference with secondary data):
  - *Qualitative research* (focus group, observation, in-depth interviews)
  - *Quantitative research* (market surveys, lab experiment)
- **Secondary data**: refers to data that has been previously collected, so they are already available, and for some other purpose. These data are not collected specifically for a study at hand but can be utilized to gain insights.
  Secondary data are the data that can be derived from both internal and external sources such as social media pages, loyalty cards, website, corporate CMR, e-commerce, and physical stores. -> we don't have to pay for them, and it is here data marketing analytics was born
  Field experiment: experimental part combined with secondary data

The Customer journey



Remember the example of Betty





The relationship between customer journey and data analysis is really strong, because all the interaction between a firm and its clients can be tracked becoming data to analyzed.
More formally the kind of behavior that we can track is a **pre-acquisition behavior** -> we can have search data before purchasing, we can have data about the purchase occasion and then we can have data about post-purchase occasion, so what happened after.

We have a pre-purchase stage -> we can have data collected thanks to different touch points such as social media, the store and so on, and also it is interesting to understand if those touch points are owned by the brand of managed by a partner.

Example: Procter & Gamble
We decide to do a partnership with Amazon for a specific product, for example Gillette.
In this case, Amazon is a partner so the channel is the e-commerce (touch point) but it is not owned but the brand.

Thinking about the customer journey, which data can be considered as the most relevant?

- Time spent on the website
- The arrival channel (social media, corporate page, an influencer page...)
- Customers' opinion post-purchase
- Response to promotions
- If customers purchase or not
- What customer purchased
- Through which channel customers interacted with (social media, store, online website, call centre...)

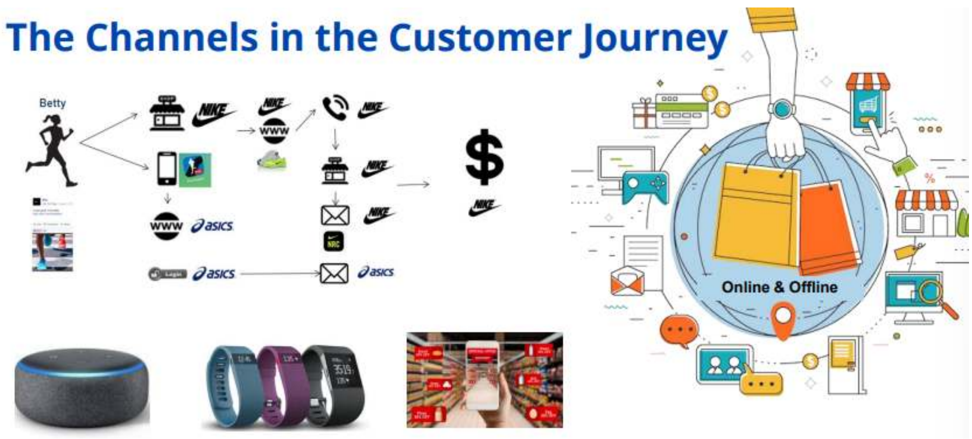
Key element of the Customer Journey

1. Digital **channels** & touchpoints and integration with physical channels (Brand-owned, Partner-owned or Customer-owned touchpoints)
   Omnichannel perspective
   - Social media

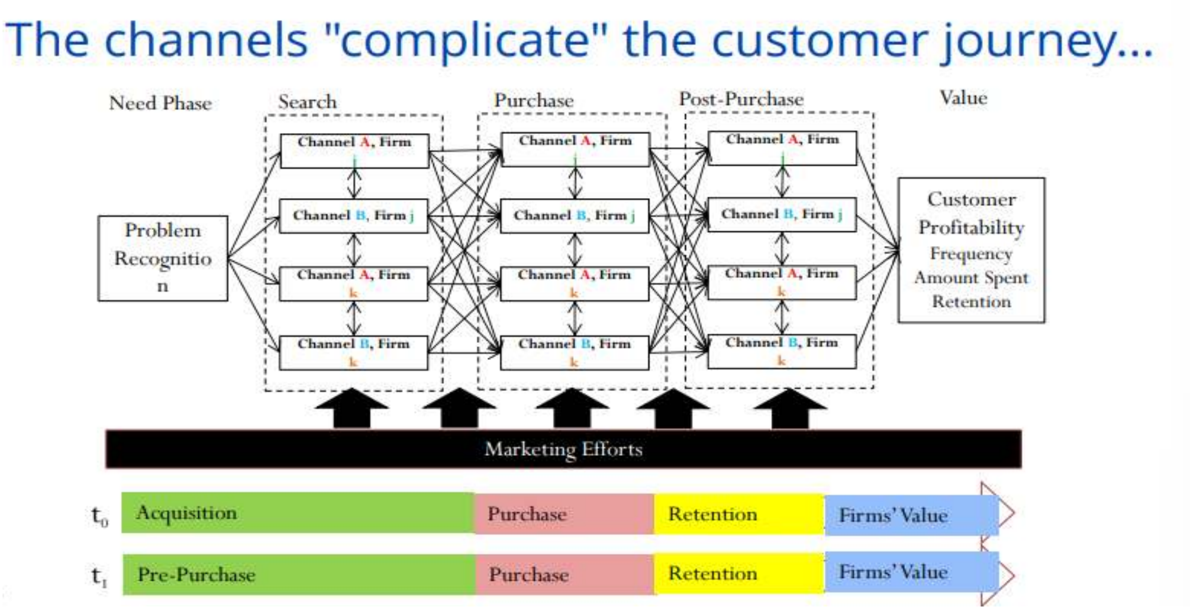
- Physical store
- Online channel
- These are just example of the possible touch points that we can track and analyzed give the firm relevant information

2. **Data**
3. **Analytics** & Better targeting

CHANNELS & TOUCHPOINTS



Track the presence within the store is very difficult -> for every single purchase occasion we can track HOW the customer interacts with the different touch points and channels not only for the first purchase occasion, but also for the cumulative purchase occasion or search sessions that the customer comulates interacting with us.



Example: Walmart retail channel strategy?

- Store
- Online store
- Website
- Pick-up service

- Front door delivery
- Smart door lockers
- Walmart + -> a subscription service similar to Amazon Prime that gives members express delivery, discounted petrol and other perks

Customers have more choices than ever before when it comes to how they get their groceries. They can shop with us in stores, order online for free pickup, or have groceries delivered to their front doors. Customers can also order fresh groceries and everyday essentials and have them delivered directly into the kitchen or garage fridge

Walmart partnered with TikTok to pioneer innovation for the fastest growing digital community. They brought a shoppable live stream experience to U.S.
TikTok users for the very first time. The TikTok community enjoyed shopping while engaging with their favorite creators. During the event, we netted 7X more views than anticipated and grew our TikTok followers by 25%.

March 9, 2021 William White, Chief Marketing Officer, Walmart U.S.

The company gave clues to its reasoning, saying TikTok's integration of ecommerce and advertising "was a clear benefit to creators and users" and would "provide Walmart with an important way for us to reach and serve omnichannel customers as well as grow our third-party marketplace and advertising businesses".

=> This is an example of channel segmentation
Example: Sephora channel strategy

- Catalog
- Community
- Store
- Smartphone

	NEED	SEARCH	PURCHASE	POST
STORE	I need something and I go to the store	I search my product, and, in the meanwhile, I find something else	It is easier finalize the purchase in store	They give you samples which incentive you to buy again
SOCIAL MEDIA	I see a post and I find new desire	I see something through the social and I want to buy it	Possibility to da an Adv which facilitate purchase actions	Newsletter, surveys, analyse the insights, broadcast channels
MAKE-UP ROOM	You need that service, or you	You see the products and	It facilitates the purchase after trying the	They usually ask you feedback

	want to learn something	want to learn more about them	products, moreover you paid for the experience itself	about the experience
WEBSITE	I need something but I don't have time to go to the store	While I'm searching a product, I'll also have correlated products	Possibility to buy online	Loyalty cards

How did all start? Omnichannel… Past 15 years

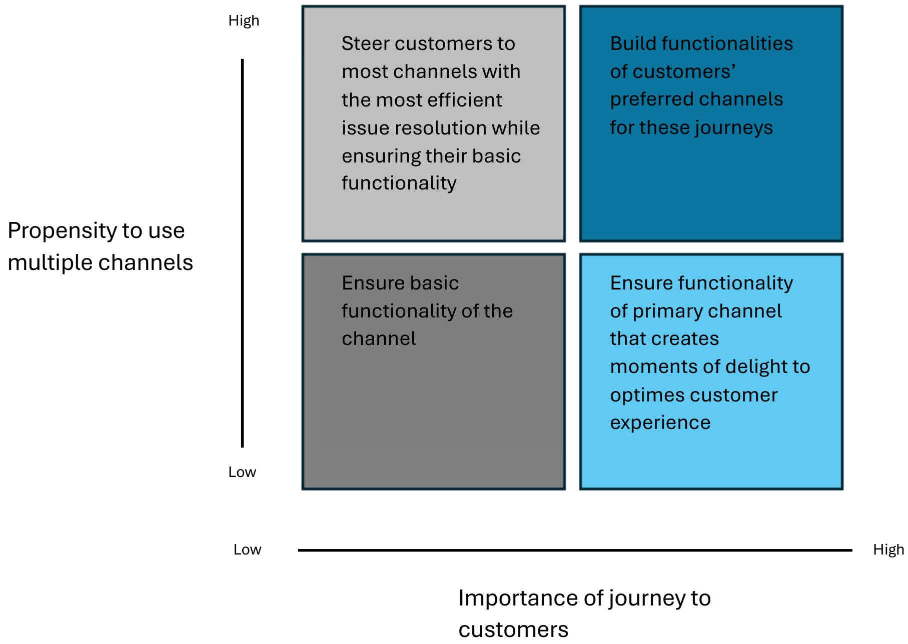
Why omnichannel?

As a company we can have multiple advantages:

1. Some channels can be more costly for some customers
2. It is possible to shape and orient the customer journey of the customers towards channels that provide more margin for the brand
3. To expand customer base (Sephora is active on social media even through the purchase button because probably they realise that they want to reach a specific target, the younger generation, that usually buy through this channel -> to reach them the company must adapt to their preferences)
   It can be a tool to increase customer satisfaction, because I can reach all the customer in different way, moment and occasion

4. It can be a tool to increase customer satisfaction, because I can reach all the customer in different way, moment and occasion

An organization can tailor its omnichannel approach by mapping each customer journey to a quadrant of the matrix and focusing on only two or three in the top-right corner.



Omnichannel Opportunities	Omnichannel Threats
Right channeling: high margin channels can produce a significant reduction in costs	**Cannibalization**: profits cannibalization across channels: increase the number, same profits, more channel
Customer satisfaction: provide customer a better service/experience	**Brand value erosion**: the brand value could be eroded if channels are not well managed
Expansion customer base: the % of online and mobile shopping has increased and continues to increase	**Channel coordination**: the company should be able to effectively manage different channels -> decrease in consumer satisfaction, customer retention…
Increase profits	

N.B. Use multiple channels can also be dangerous depending on the kind of products that the company is selling.

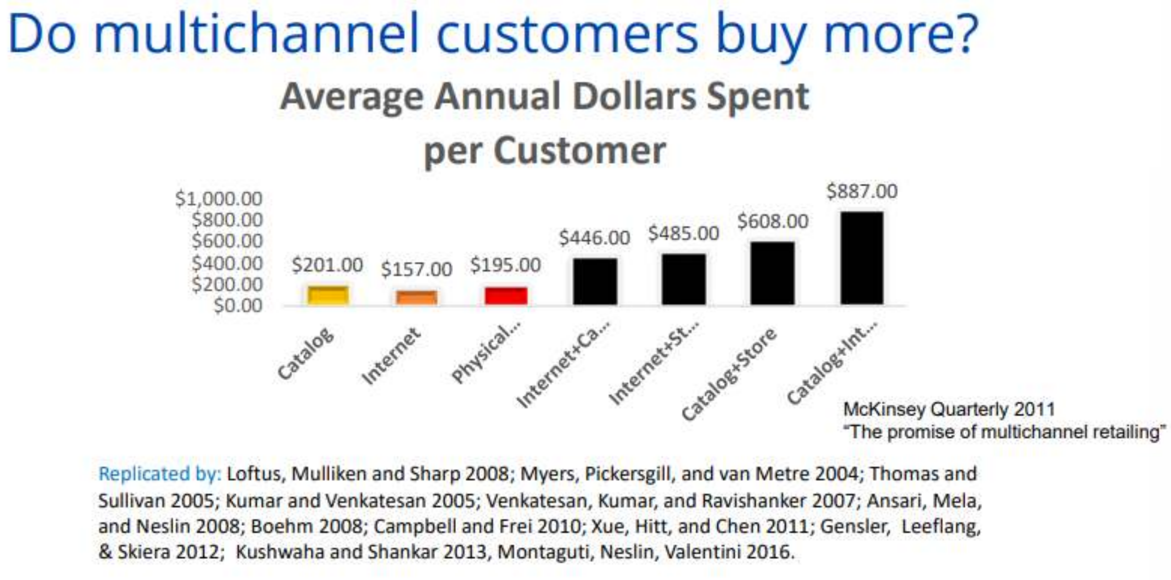
A growing interest in the omnichannel strategy…

70% of Saks customers purchasing online also purchase in stores

The multi-channel customers spend from 3-4 times more than the single channel

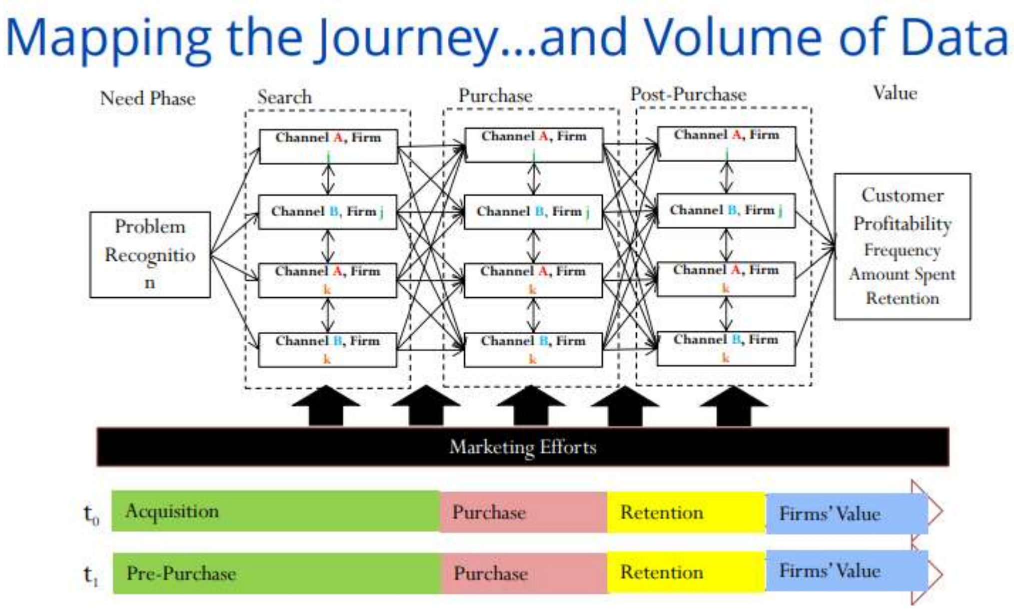
Several businesses decide to implement different channels because they find evidence that customers tend to be more profitable for the brands -> of course this is not a rule valid for every firm and industry, so it is important to verify this behavior before investing in this marketing activity.

Is there a relationship between channel choice and customer profitability?

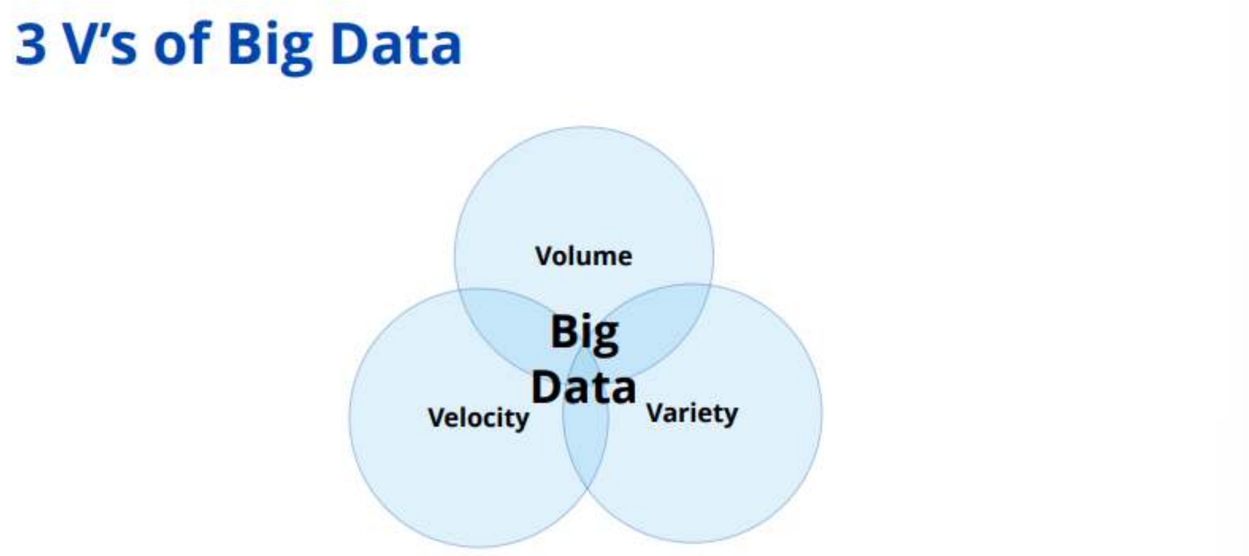


DATA

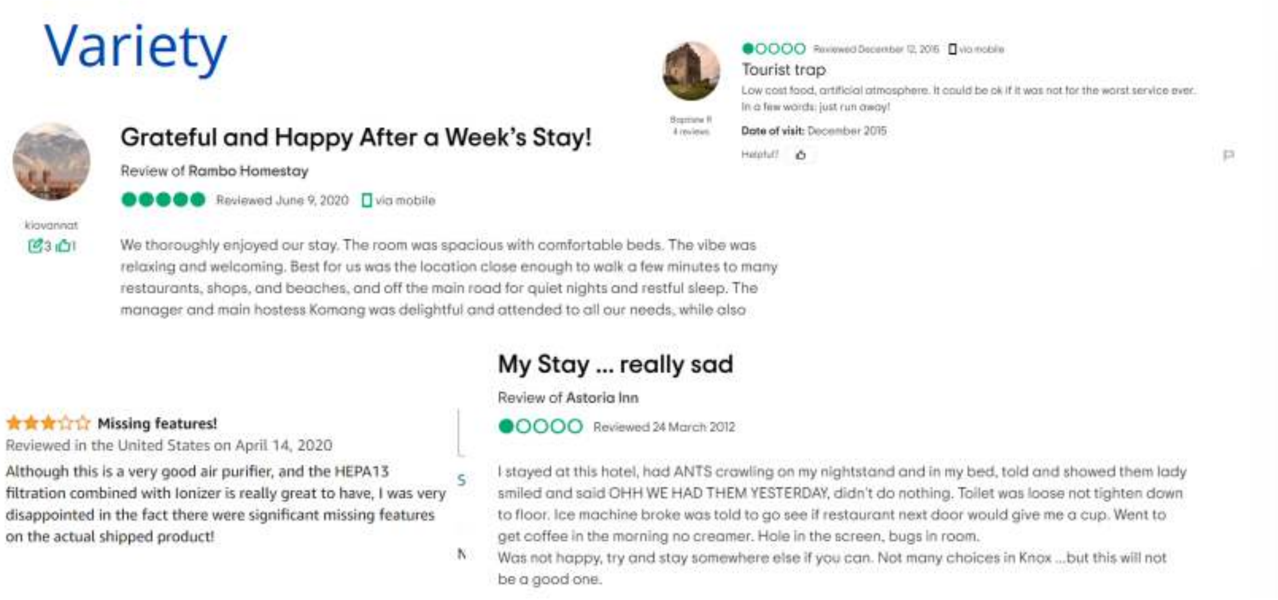
Data -> available data are big, that is why they are so interesting



Nowadays we have the possibility to enter in contact with several secondary data -> the volume of disposable data in huge, but remember that we don't necessarily need a lot of data, but what is more important is have good data available now!



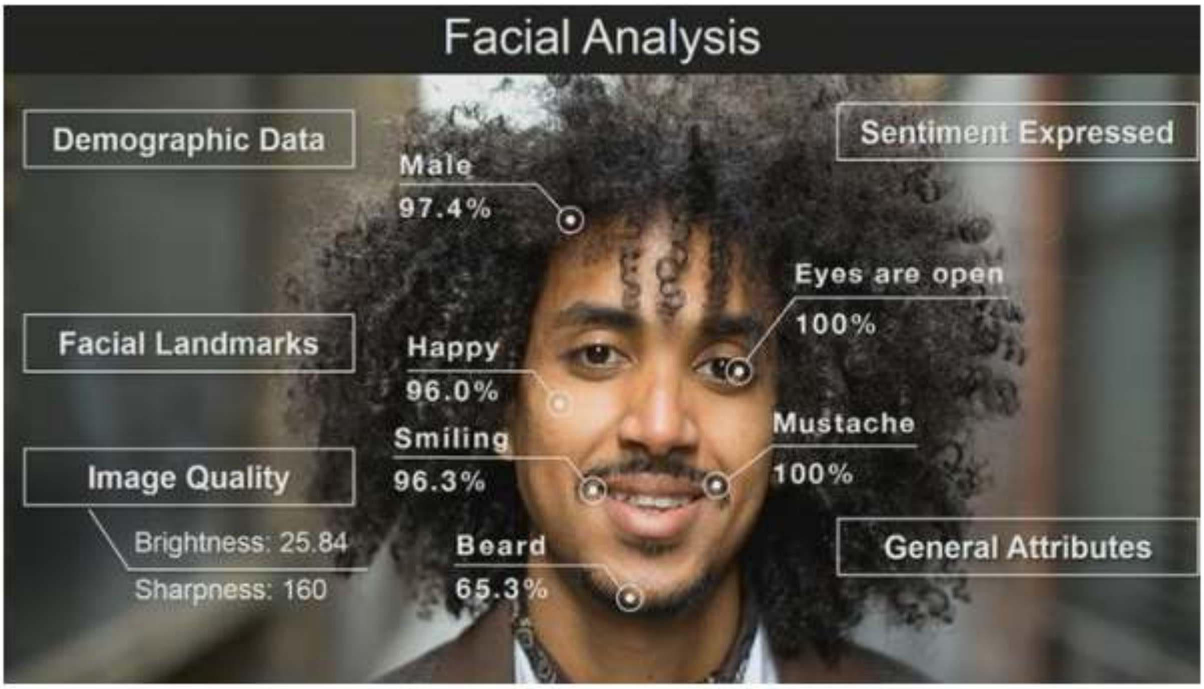
- **Volume**: data are big, we have a lot of information (less interesting: in principles we don't need a lot of data, but good data)
- **Velocity**: I can have the data now -> we don't need anymore to do surveys even though they are still useful -> we can now opt for an instant marketing strategy
- **Variety**: also, text are data (through Python I can transform a text into data) that can be analysed; even photos can be data, about our preferences, emotions, age, gender, vocal audios



amazon Rekognition

Variety

Example: Booking.com, Tripadvisor can classity UGC with labels associated to the presence of specific elements (e.g. swimming pool, mountain

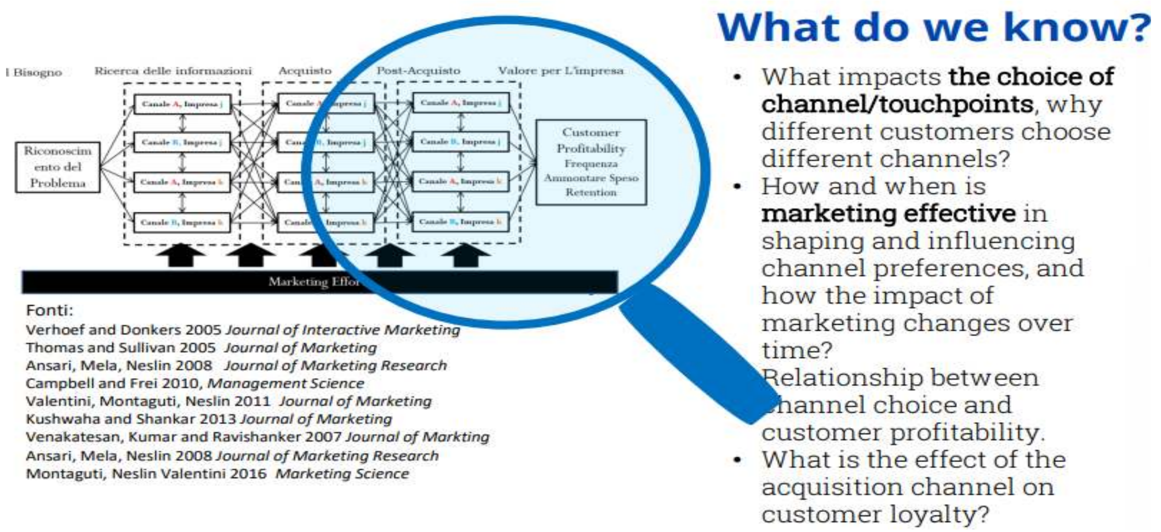


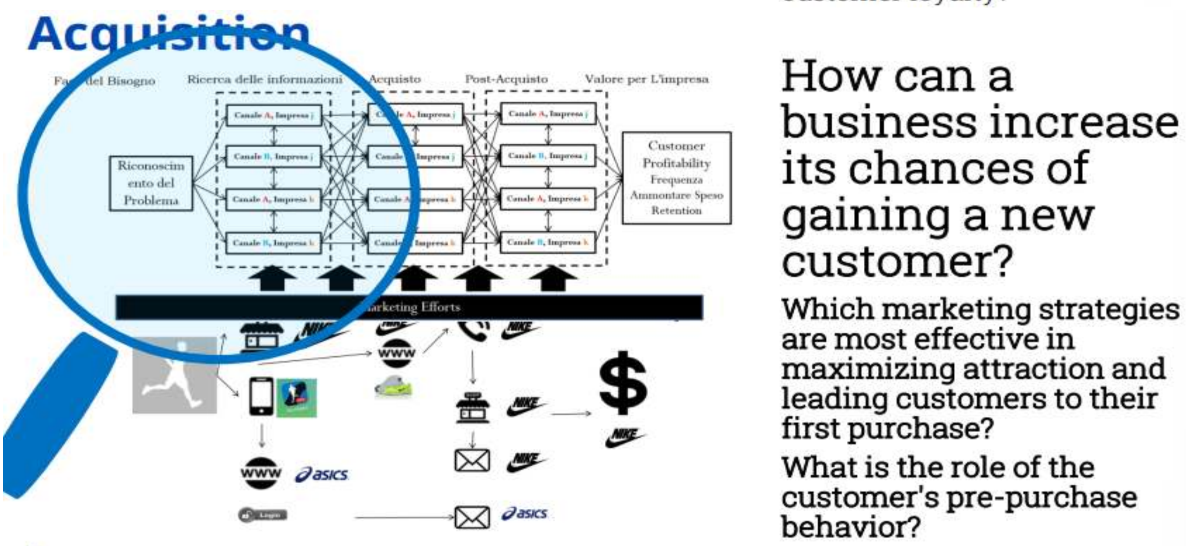
More info: https://www.youtube.com/watch?v=fk-TxySUAzw

[What Data are Voice Assistants Collecting]

ANALYTICS & DATA-DRIVEN DECISIONS

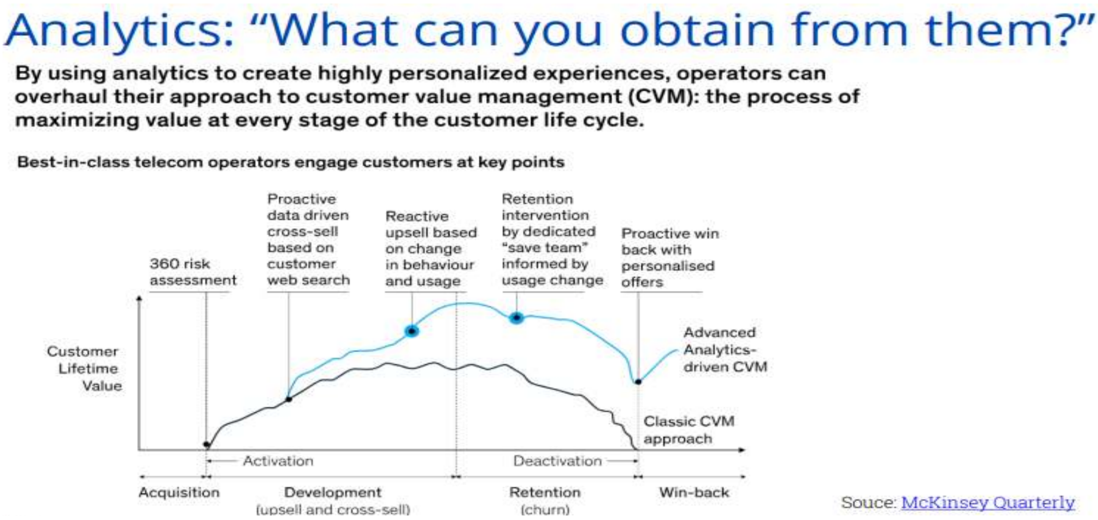
Big data and the information derived from them can be used anytime.
Big data and analytics work basing on
algorithms -> as a whole these practices work
How does it work?





The challenge… **Map and analyze the entire customer journey**.

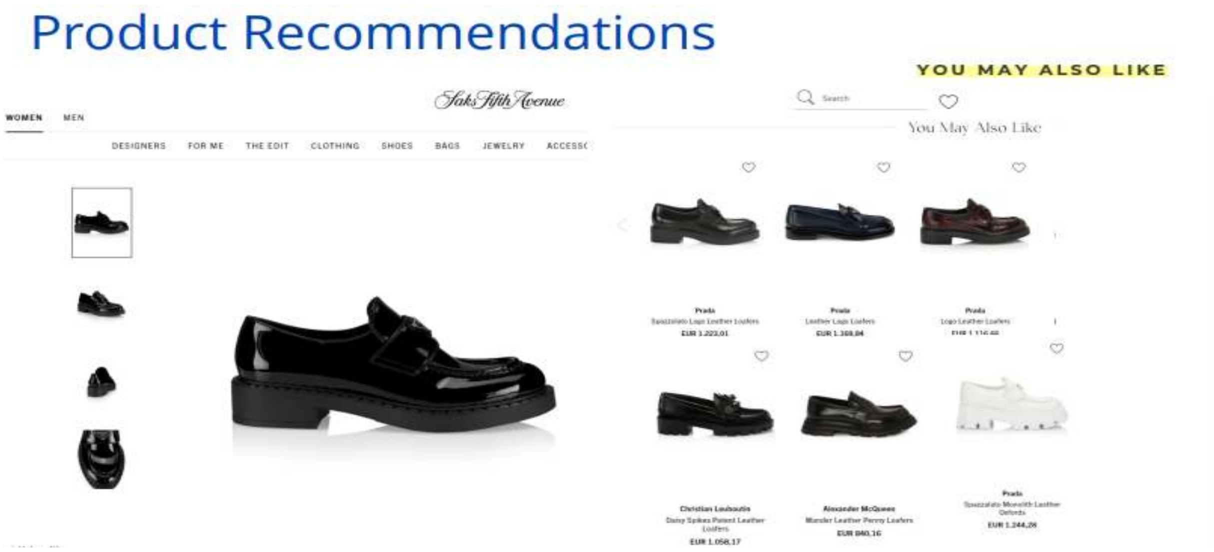
- How can customers be segmented based on their journey?
- What are the predominant patterns?
- Which types of marketing activities work best?
- What leads to higher profitability?



Why Data Science and Analytics?

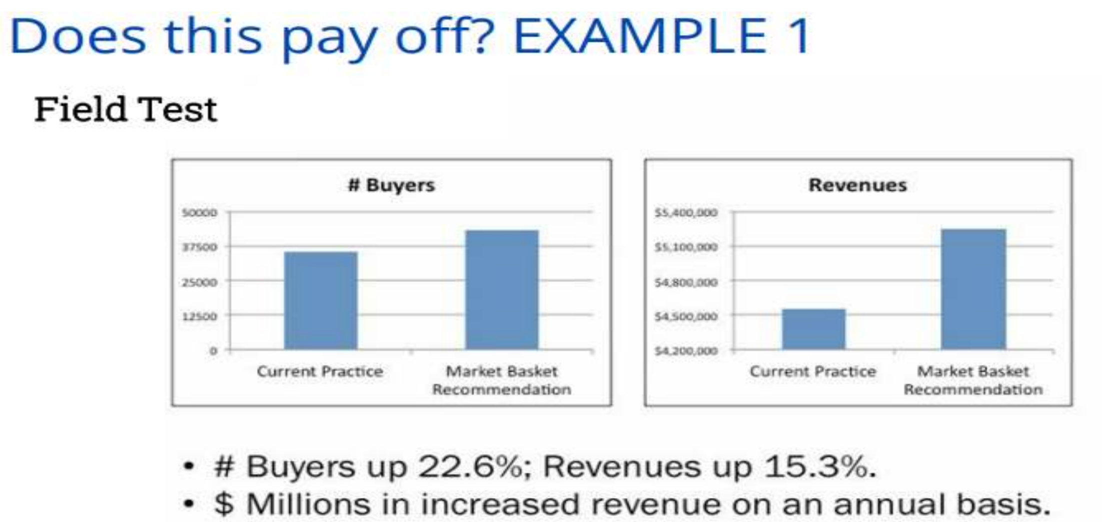
Data Analytics can be seen as Bottom Line, but which is the impact on the customer's value?

Overwiev: Why do I receive these?



How is this done? Market Basket Analysis

- Data: 1,000,000s page visit records
- Compute probability Pr(view product B | view product A)
- If the customer view product A, recommend product B's with maximum probability p(B|A)



Example 2 | NPTB (Direct Email)

Bank wants to cross-sell home improvement loan

- Cross-selling: loan ▪

Idea: use **next-product-to-buy model (NPTB)** to identify those customers with an high probability to "buy" a loan

Compute probability:

- **Pr(Buy Loan Next | Demos, Previous products bought)**
- Data: 100,000 customer purchase records for estimation

Does this pay off? EXAMPLE 2

Field Test

- NPTB model produced more responses and more revenues per customer.
- NPTB model produced 530% ROI, vs. -16% for current practice

How is this done? The basic intuition
A simple predictive model for targeting

SEPHORA

Number of offers mailed:	1,000,000
Profit contribution per response:	$80
Cost per mailing:	$.70
Response rate:	1%

Profit = 1,000,000 x .01 x $80 – 1,000,000 x $.70
= $800,000 - $700,000
= $100,000

The direct marketing campaign is effective!

Most of the investment in direct marketing is wasted!

Targeting analysis: lift-based approach

170000= (3%*100000*80)-(100000*0.7)

Decile	Number of Prospects	PredictResp Rate	Profit	Cumulative Profit
1	100,000	3.00%	$170,000	$170,000
2	100,000	2.00%	$ 90,000	$ 260,000
3	100,000	1.40%	$ 42,000	$ 302,000
4	100,000	1.15%	$ 22,000	$ 324,000
5	100,000	1.00%	$ 10,000	$ 334,000
6	100,000	0.60%	$ -22,000	$ 312,000
7	100,000	0.40%	$ -38,000	$ 274,000
8	100,000	0.30%	$ -46,000	$ 228,000
9	100,000	0.10%	$ -62,000	$ 166,000
10	100,000	0.05%	$ -66,000	$ 100,000

=> Profits improvement →$100,000 __>$334,000

We must stop at the fifth decile otherwise we will start losing profits.

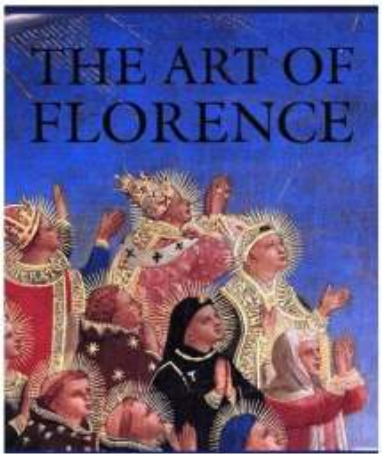
Return on marketing investment: ROI

ROI_1 = $100,000/$700,000 = 14.3%

ROI_2= $334,000/$350,000 = *95.4%*

How do it practically?

Barnes & Noble targeting strategy



- B&N want to send an offer for the purchase of the book “The art of Florence” to a sample of **1,000 customer** in their CMR
- The sample is randomly selected
- **DATA**: For these customers, they have information on previous purchase
  - Number of months since the last purchase (recency)
  - Number of art book previously purchased (art)

Objective: estimate the probability that a generic customer in the target buys the book "The Art of Florence".
Method: Regressive type - Regression (Logit)
Outcome variable of the model: dummy 0/1 (1=buys, 0=does not buy).
Independent variables: Recency, Art.

```
Logistic regression                              Number of obs   =       1000
                                                 LR chi2(2)      =      69.14
                                                 Prob > chi2     =     0.0000
Log likelihood = -251.46648                      Pseudo R2       =     0.1209

------------------------------------------------------------------------------
    purchase |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     recency |  -.0707172   .0192297    -3.68   0.000    -.1084067   -.0330277
         art |   .9890522   .1346605     7.34   0.000     .7251224    1.252982
       _cons |  -2.225636   .2389241    -9.32   0.000    -2.693918   -1.757353
------------------------------------------------------------------------------
```

INTERPRETATION:

a. Check p-value: is it relevant? -> As the fact that for both regency and art it is 0.000 the estimation is really reliable

b. Coefficients: regency is -0.707172, while art is 0.9890622 so they have different effects on purchases. => Regency is going to decrease the likelihood to purchase by 7%, while Art is going to increase the likelihood to purchase by 9.89%
c. Constant tells us that as a whole, it is really unlikely to purchase the products together.

We trust the model, so we can now predict the future and they way in which customer will behave.
Let's take two consumers: Mario (id.21) and Anna (id.145).
Mario made his last purchase 6 months ago. Also, Mario has previously bought an art book.
Anna made her last purchase 18 months ago, and she has never purchased from the product category (art books).
What is the difference in terms of the likelihood of purchasing the book "The Art of Florence" for the two customers?

Mario (X_1=6, X_2=1)
U_{mario}=-2.22+-0.07*6+0.98*1=
= -1.66

$$p_{Mario} = \frac{\exp(-1.66)}{[1+\exp(-1.66)]} = 0.16$$

Anna (X_1=18, X_2=0)
U_{Anna}=-2.22+-0.07*18+0.98*0=
= -3.48

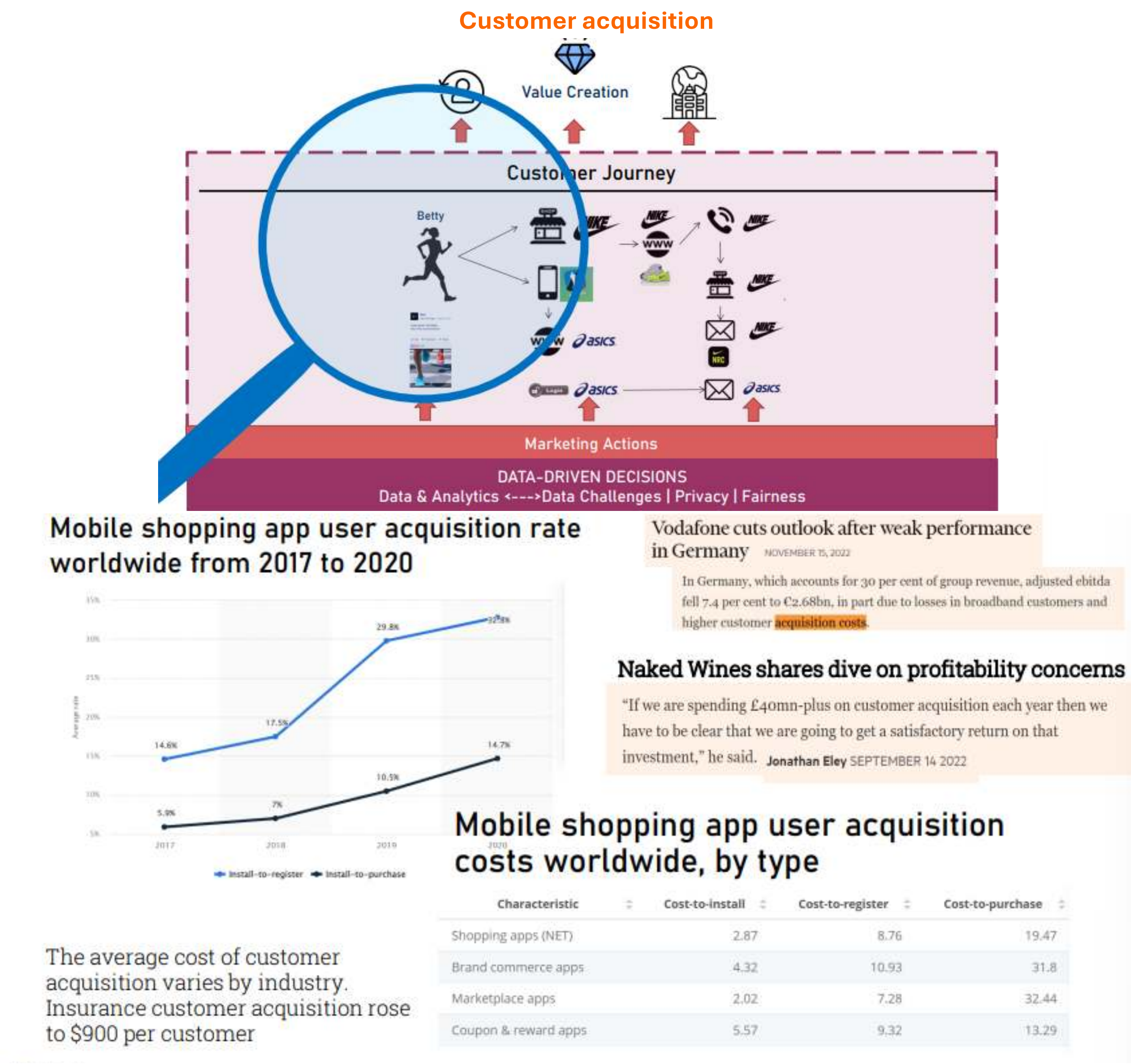
$$p_{Anna} = \frac{\exp(-3.48)}{[1+\exp(-3.48)]} = 0.03$$

Mario
16%

Anna
3%

Who should be included in the target?
We know that: the cost of sending the offer is **$1 for each customer**.
The net profit generated from a book purchase by a customer is $6 (net of the cost spent on sending the offer by mail).

(\$6)*p_{Ac} + (-\$1)*(1-p_{Ac}) >0 => p_{Ac} > 1/7 => Purchase probability > 14%



Characteristic	Cost-to-install	Cost-to-register	Cost-to-purchase
Shopping apps (NET)	2.87	8.76	19.47
Brand commerce apps	4.32	10.93	31.8
Marketplace apps	2.02	7.28	32.44
Coupon & reward apps	5.57	9.32	13.29

Acquisition phase: our task as manager is to acquire new customers.
What can we do?

- we can use marketing tools in order to obtain the results
- data can be a tool depending on which data we have -> if our problem is an acquisition problem and we want to use secondary data probably we have information about the search pattern.

The acquisition phase is an extremely relevant aspect of the customer journey and a possibility to improve revenues and expand our audience with new individuals and customers. It is a very delicate phase -> it is the more expensive one because we need to convince someone who doesn't know us to purchase our products. So we need to calibrate our action very carefully because they have to be effective.

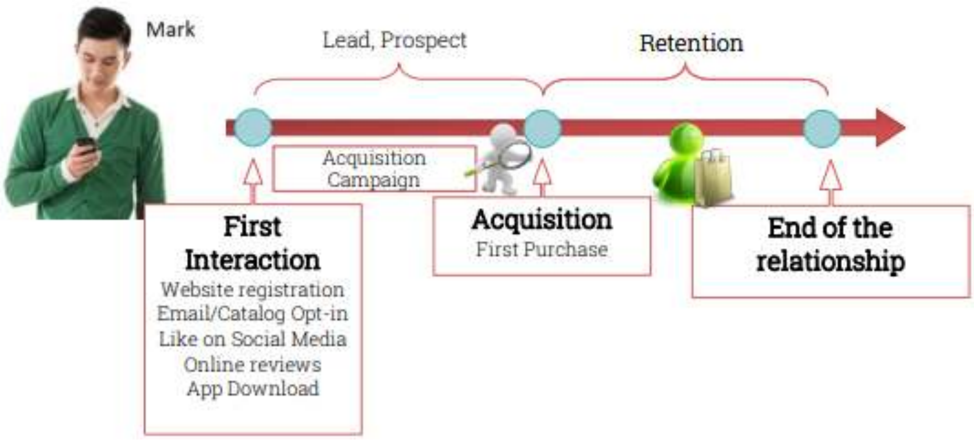
Customer acquisition: Background

Customer acquisition -> the first time a new customer purchases from a firm or subscribes to a service.

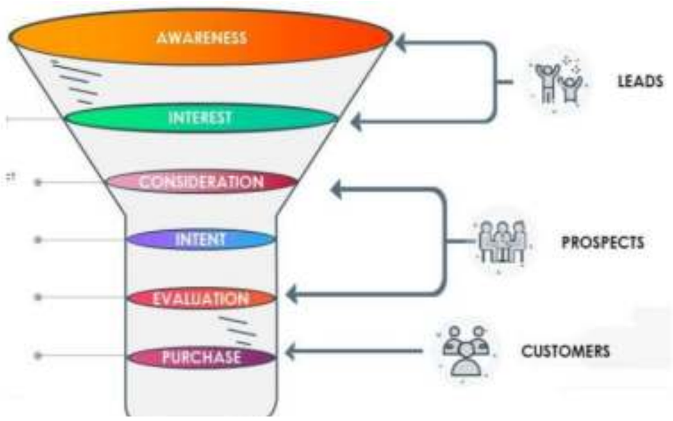
We can distinguish between to distinct worlds:

a. member/contract/subscription-based business (Netflix, Disney +, Spotify, ...) -> you know exactly when the customer was acquired because they sign a contract
b. you need something within your business in order to define if you have acquired a new customer -> a way to define it is to track the very first purchase as acquisition, but it is not possible for every business (ex. Sephora and in-store purchases -> that is why they usually try to convince you to subscribe to the loyalty card, so in this way they can track if you are doing repeated purchase or not)

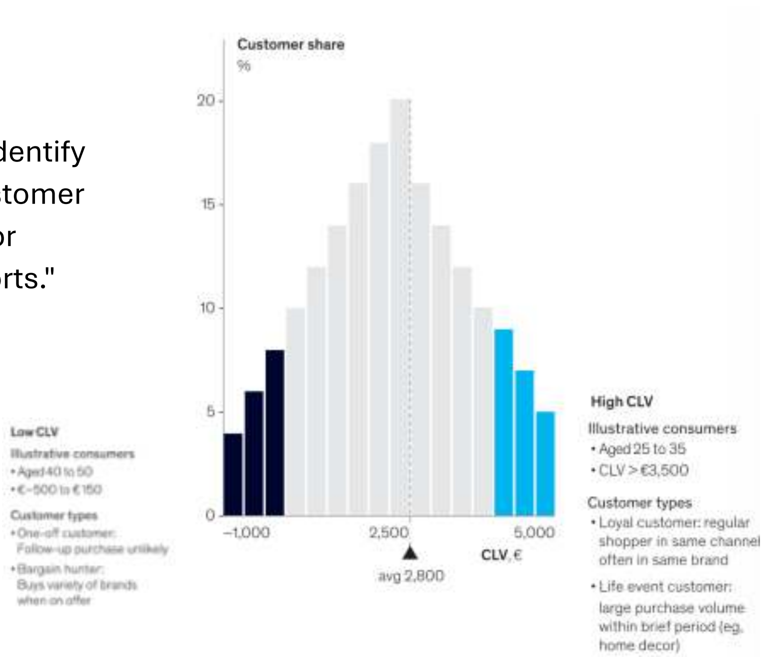
Does the first purchase really represent the first interaction with the brand? NO, but it represents the first engaged interactions with the brand.
In fact, the first purchase rarely represents the firm's first contact with a customer -> first interaction usually came from website registration, email/catalog opt-in like on social media online reviews app download
N.B. We call lead or prospects those individuals that are engaged with the firm (social media, newsletter) but never purchase, be as soon as they start to purchase they become acquire => this means that there is a per-acquisition and a post-acquisition phase with different data



Customer acquisition is a crucial phase of Customer Relationship Management (CRM) because, before focusing on retention and customer satisfaction, a company must identify the customers who are most likely to be acquired and who are worth acquiring



- "Data help marketing and sales teams identify indicators of high CLV and low CAC (customer acquisition costs) respectively, and tailor marketing campaigns to individual cohorts."
- Example "cohort analysis" →



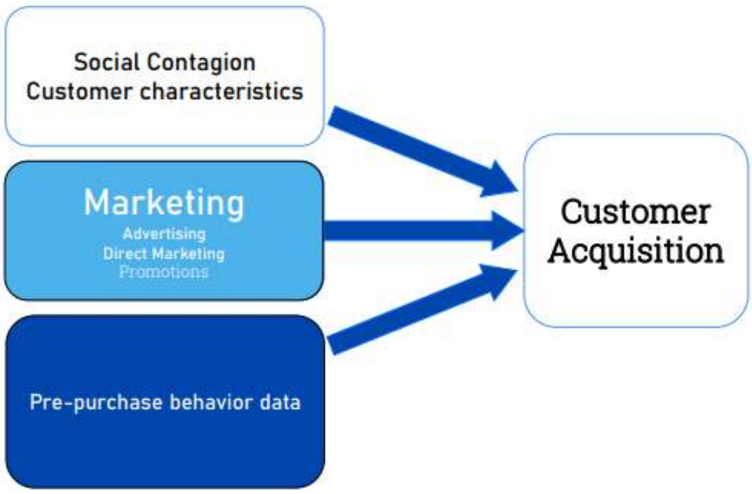
THINGS ARE DIFFERENT TODAY!

Most firms and organizations can:

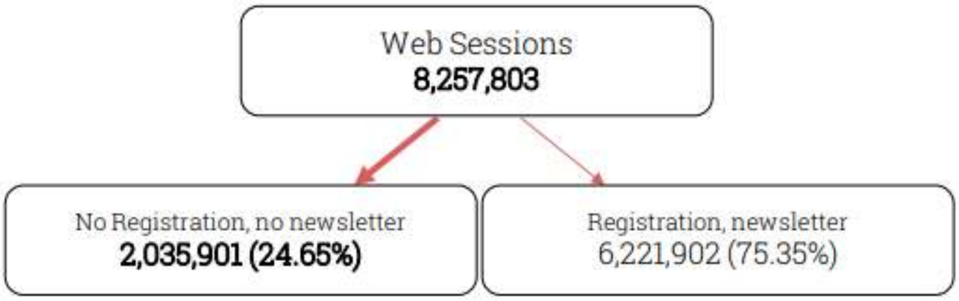
1. Identify the first purchase
2. Track pre-acquisition behavior (e.g. search activity);
3. Monitor marketing activity at customer-level

What impacts the likelihood of acquiring a new customer?

Evidence from scientific literature

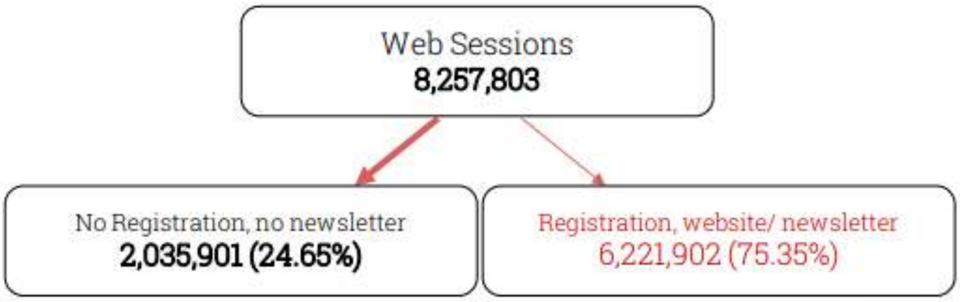


Example: pre-acquisition data



- User identification based on cookies and email ID

Dreambox Add	Number of products added to the Dreambox in each period
Dreambox Drop	Number of products deleted from the Dreambox in each period
Device	Number of devices used in each period
Website Sessions	Number of website sessions in each period
Website Session Length	Average website session time length in each period
m-site Sessions	Number of m-site sessions in each period
m-Site Session Length	Average m-site session time length in each period
Average Pages/screens	Average number of pages/screens seen in each session
Average Products	Average number of products seen in each session
Average Suggested Products	Average number of suggested products (similar/same designer) seen in each period
Average Clicks	Average number of clicks in each session
Average Filters	Average number of filters used in each searching session
Total Pages/screens	Total number of pages/screens seen in each period
Total Products	Total number of products seen in each period
Total Suggested Products	Total number of suggested products (similar/same designer) seen in each period
Total Clicks	Total number of clicks in each session
Total Filters	Total number of searching sessions with filters in each period
Ranking Type	Number of times the user has used each type of ranking in each period



For 75% of users, the firm begins monitoring their behavior (also prior to their acquisition)

- User identification based on cookies and email ID

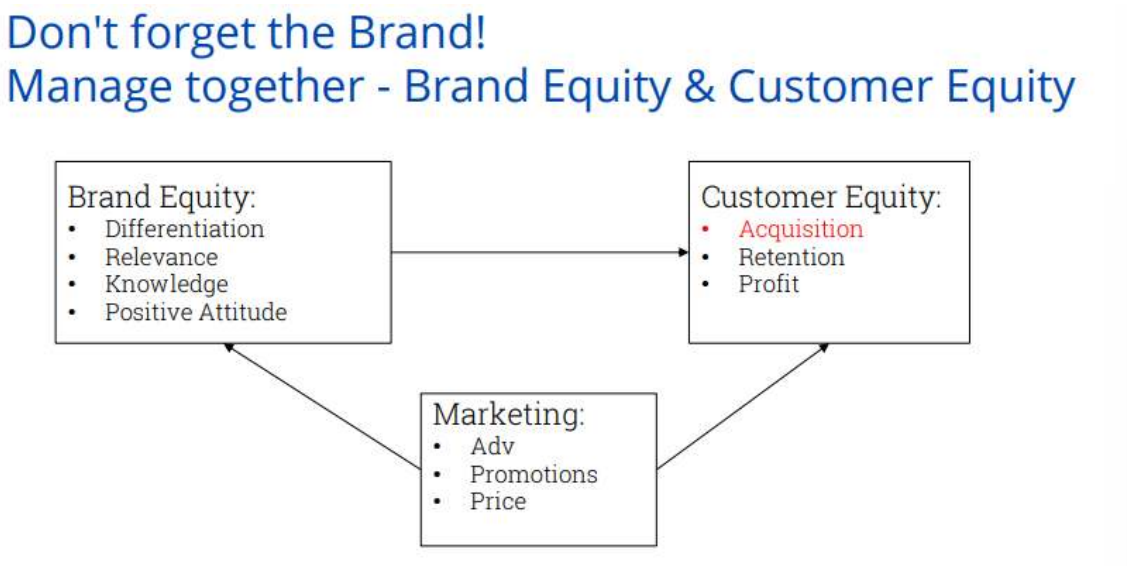
SOURCE: Anonymous Company

There is a phase in which for the company we are a cookie, and ID that is interacting with them -> in that phase, even before purchasing, the incentive of the business is try to incentivise some form of registration (social media registration, newsletter, download the app) -> they try to push us to do these actions, in order to start collecting data about us, our preferences and our habits.

- For 75% OF USERS the firm begins monitoring their behaviour (also prior to their acquisition) -> even if you share a fake email, or a mail that you never use the information that you are sharing with them is still relevant because they will be able to analyse your behavior and linked your device with an ID number -> we you register the company start to have the right to collect data about you because when you do the registration you always accept privacy terms!

Marketing Analytics to Acquire the Customer: what to DO and NOT to do

What TO DO (Do) and what NOT to do (don't):
- Do: Believe in the value of Data.
- Do: Invest in Data first, then in the Method.
- Do: Think "across channels."
- Don’t: Forget the BranD
- Don’t: Ignore Privacy
- Don't: Believe in the Big Data trend "without critical thinking." ▪ (Identifying the right marketing strategies and the right data is not easy.)



Exercise:

Try to register / understand how to register on the Website / Newsletter of two brands or companies ▪ Alternatively, check your email account to find requests for newsletter registration / renewal of privacy terms
- New balance:
  1) accept cookies
  2) do you want to receive new info before the others?
  3) insert the email address
- Dove:
  1) cookies

2) It is possible to register through the uk website (no italian):
    - Age >=16
    - Email address

The registration phase leads us, which is strictly related to customer acquisition, to three strategic decisions:

1. How to ask to register
2. Depending on the country we could have different regulation
3. What information ask

⇨ This is a key phase cause a first occasion to target our audience

What does PRIVACY have to do with marketing?

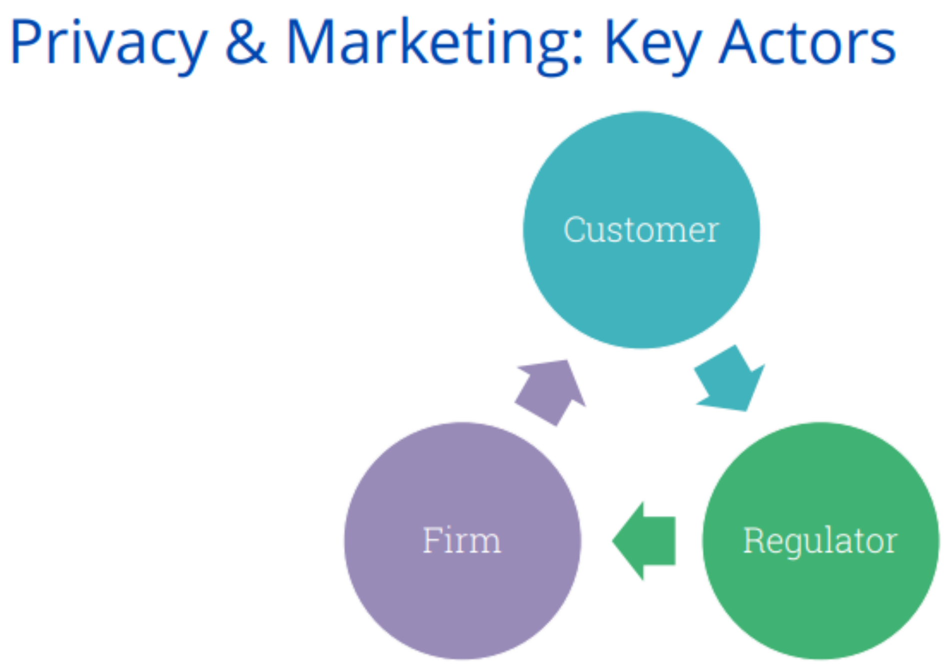
WHY?

- DATA → Better Targeting → Customized Marketing Strategy

CHALLENGES

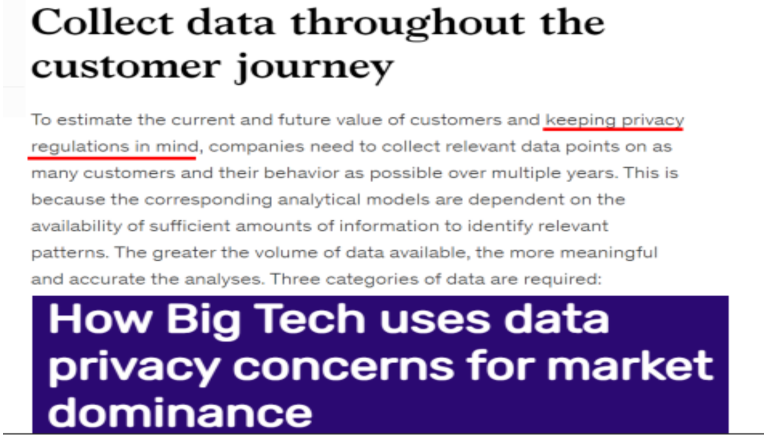
- Profiling and targeting
- Data collection and retention
- How will the way of retaining and collecting data change

In Europe we have GDPR



FIRMS: PRIVACY AS BUSINESS PROBLEM

McKinsey Quartely, Nov 2021

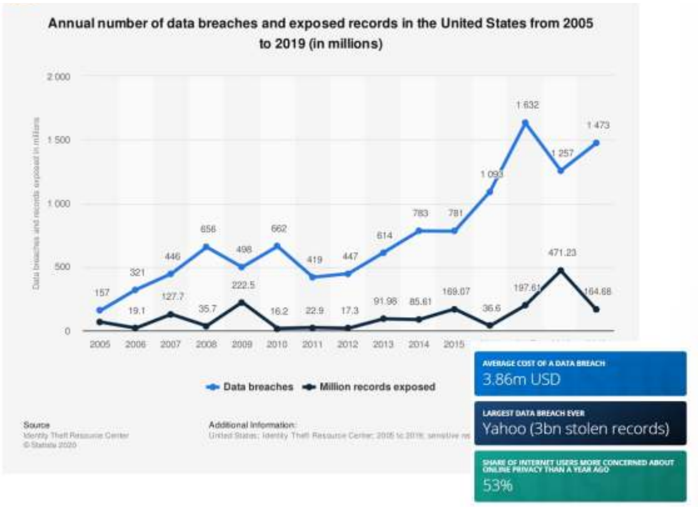


As consumers we are more aware about the importance of privacy -> we might be sensitive about the information that we want to share.

DATA BREACHES

Surprising increase in the number of Data Breach events since 2015.

Increase awareness value of data

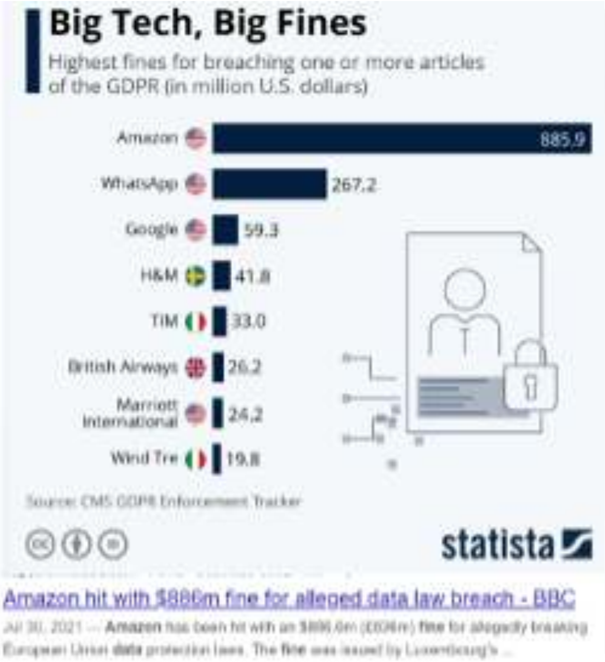


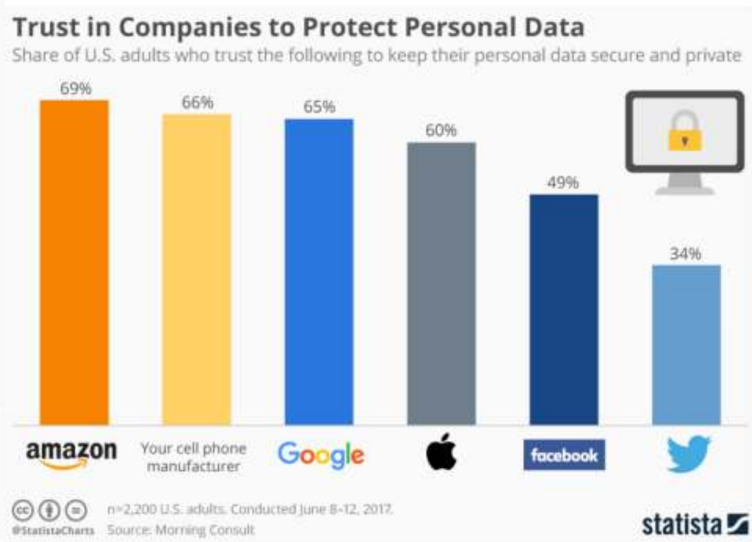
If a brand is involved in a data breach than of course is a signal for us to be more aware about our data.

Source: https://www.forbes.com/sites/washingtonbytes/2018/10/30/netflixs-race-based-marketing-shows-potential-for-anticompetitive-data-abuses/?sh=7ed48b173f48

In 2018, Netflix faced criticism when users discovered that the algorithm was sometimes categorizing content by race, leading to problematic recommendations

The answer of Netflix in that case was that they were just using data and algorithms -> the algortim maybe was biased, which means that data can also be problematic regarding to privacy; algortihms can be super useful but at the same time they can create problems



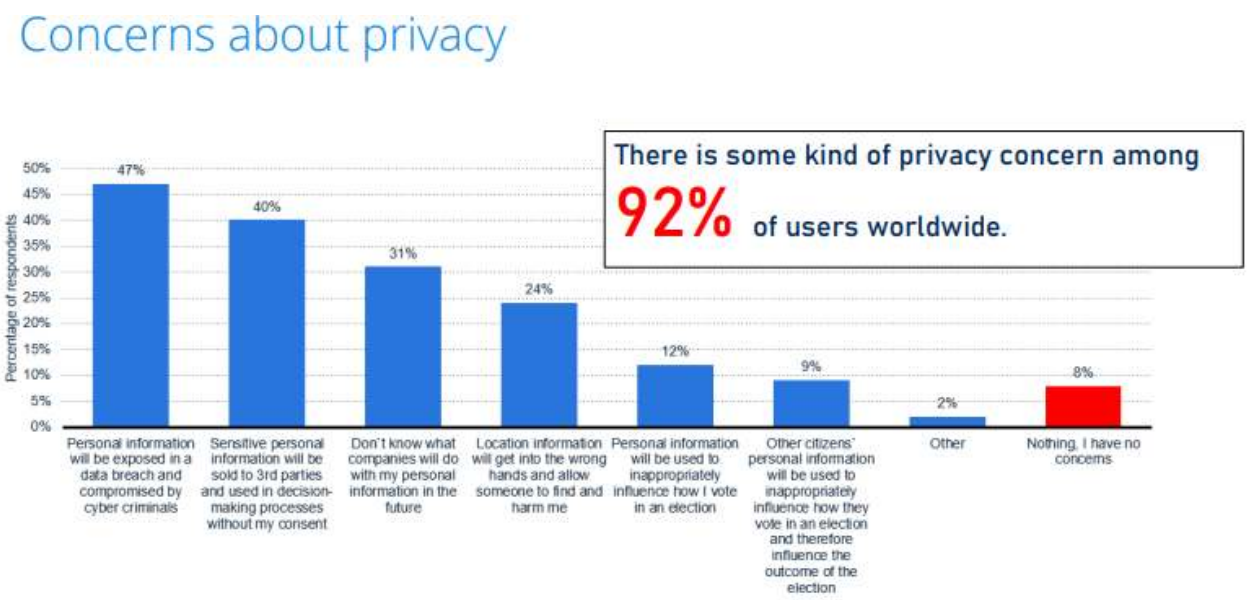


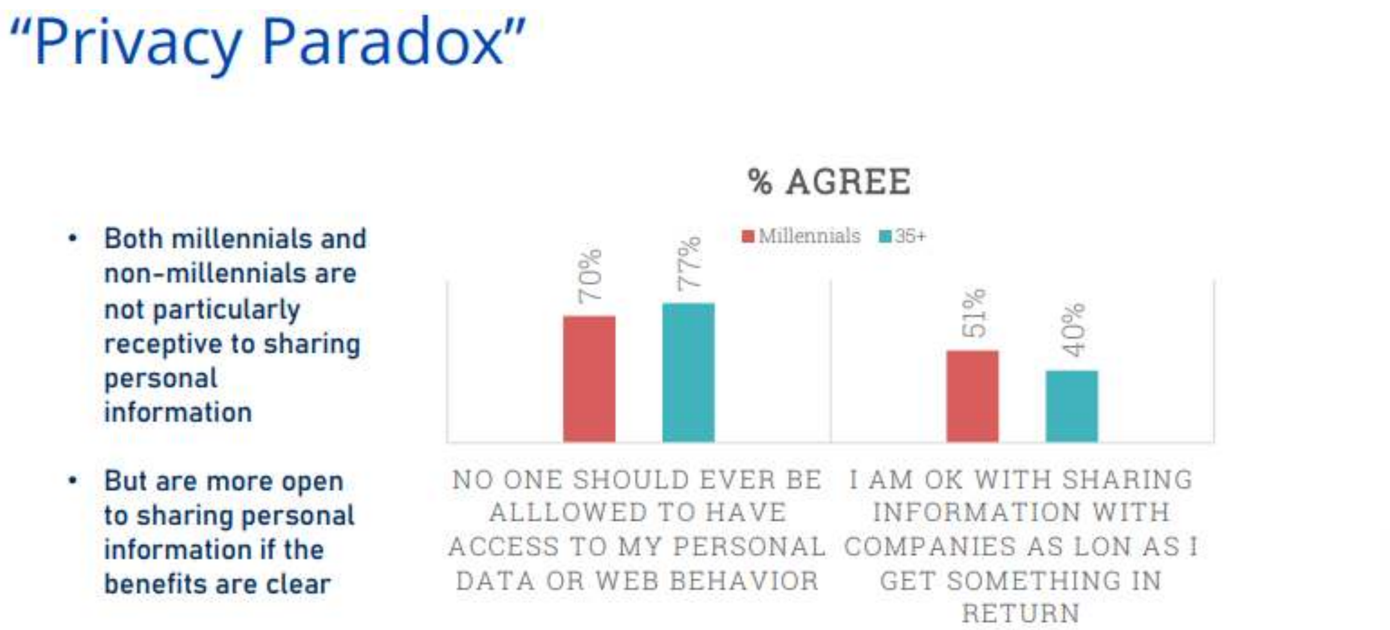
Amazon's Gender-Biased Algorithm

In 2018, it was reported that Amazon's recruiting tool, which used machine learning to review resumes and identify top candidates, showed a gender bias. The algorithm favored male candidates over female candidates, reflecting the gender disparities present in the tech industry.

Data-breaches and fines

Instagram: $403 million
In September 2022, Ireland's Data Protection Commissioner (DPC) fined Instagram for violating children's privacy under the terms of the GDPR.
T-Mobile: $350 million (July 2022)
WhatsApp: $255 million
Facebook-owned messaging service WhatsApp was fined €225 million ($255 million) in August 2021 for a series of GDPR cross-border data protection infringements in Ireland.





The value of personal information

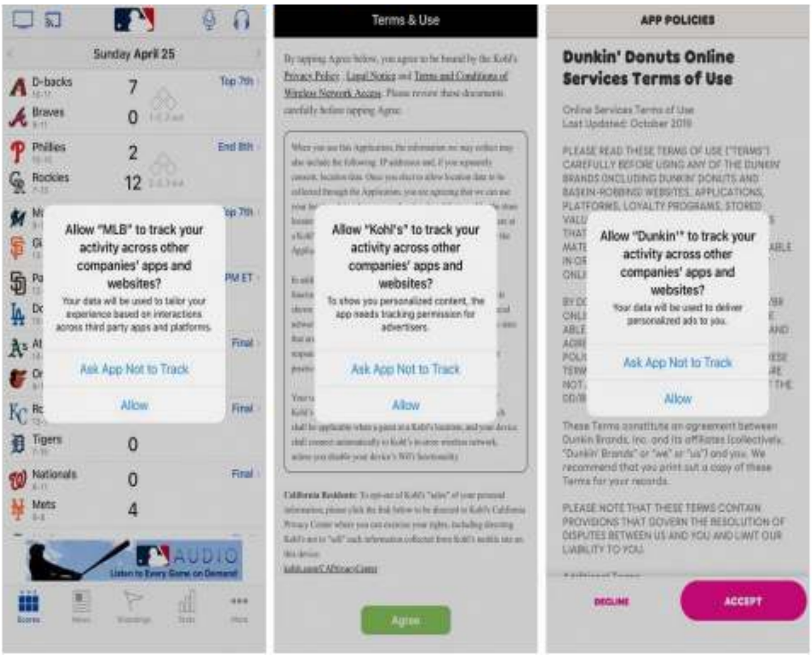
We are moving towards a world where consumers will have to allow the use of personal information: What will individuals do? Which individuals will be more inclined to give up personal information? Consumers are heterogeneous in terms of privacy preferences → The data and opt-ins collected might not represent the entire population (Lin 2021).

- ⇨ The 'privacy paradox' exists! → Businesses need to better understand consumer behavior and their intention to share personal information (Kim Barasz and John 2021) → difference between third party and first party

This explains why privacy has become a marketing strategy -> the way they ask you to accept the privacy conditions has a marketing strategy behind.
Particularly in Europe this is really relevant because if you don't accept the privacy they MUST stop tracking you -> the data about you are not available anymore, so for firms is fundamental to have customers that accept privacy conditions by structuring a successful marketing strategy linked to it.

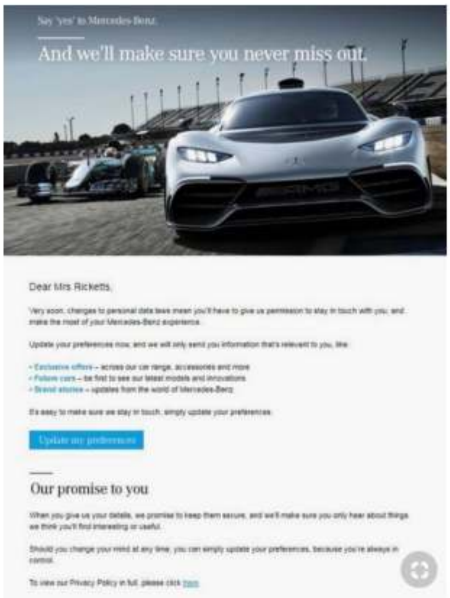
APP TRACKING TRASPARENCY: APPLE

Caso Apple



The interesting thing about this campaign is that Apple is informing you that we dealing with partner company you can choose whether share your personal information our not, but this is not valid for them who keep tracking you.

Persuasive vs. transparent



Example: Field Test – How to get Opt-ins

Informative High

Persuasive High

The image used in the field test removed

: Field Test Groups

	Persuasive		
Informative	*Low*	*Moderate* Sconto	*High* Sconto & Framing
Low	**G1 Generic Message** Total contacts=2339	**G3 Moderately Persuasive** Total contacts=2340	**G5 Highly Persuasive** Total contacts=2340
High	**G2 Informative** Total contacts=2340	**G4 Informative & Moderately Persuasive** Total contacts=2340	**G6 Informative & Highly Persuasive** Total contacts=2379

Field Test: Logistic Regression (n=14078) DV=accept privacy=1, 0 otherwise

	Coef.	z	p-value
Informative	0.35	1.48	0.138
Moderately Persuasive Discount	1.04	4.97	0.000
Highly Persuasive Discount + Framing	-0.17	-0.66	0.512
Moderately Persuasive * Informative	-0.19	-0.70	0.487
Highly Persuasive * Informative	0.68	2.09	0.037
Constant	-4.28	-24.03	0.000
Number of Observations=14,078			
LR $\chi^2(5) = 83.32$, p-value=0.000			

As a company you can not only use data but you can also think about seling data -> that is why a marketing strategy about privacy is fundamental!

ARTEA: DESINING TARGETING STRATEGIES HARVARD BUSINESS REVIEW – CASE

HBR Artea: a customer acquistion problem

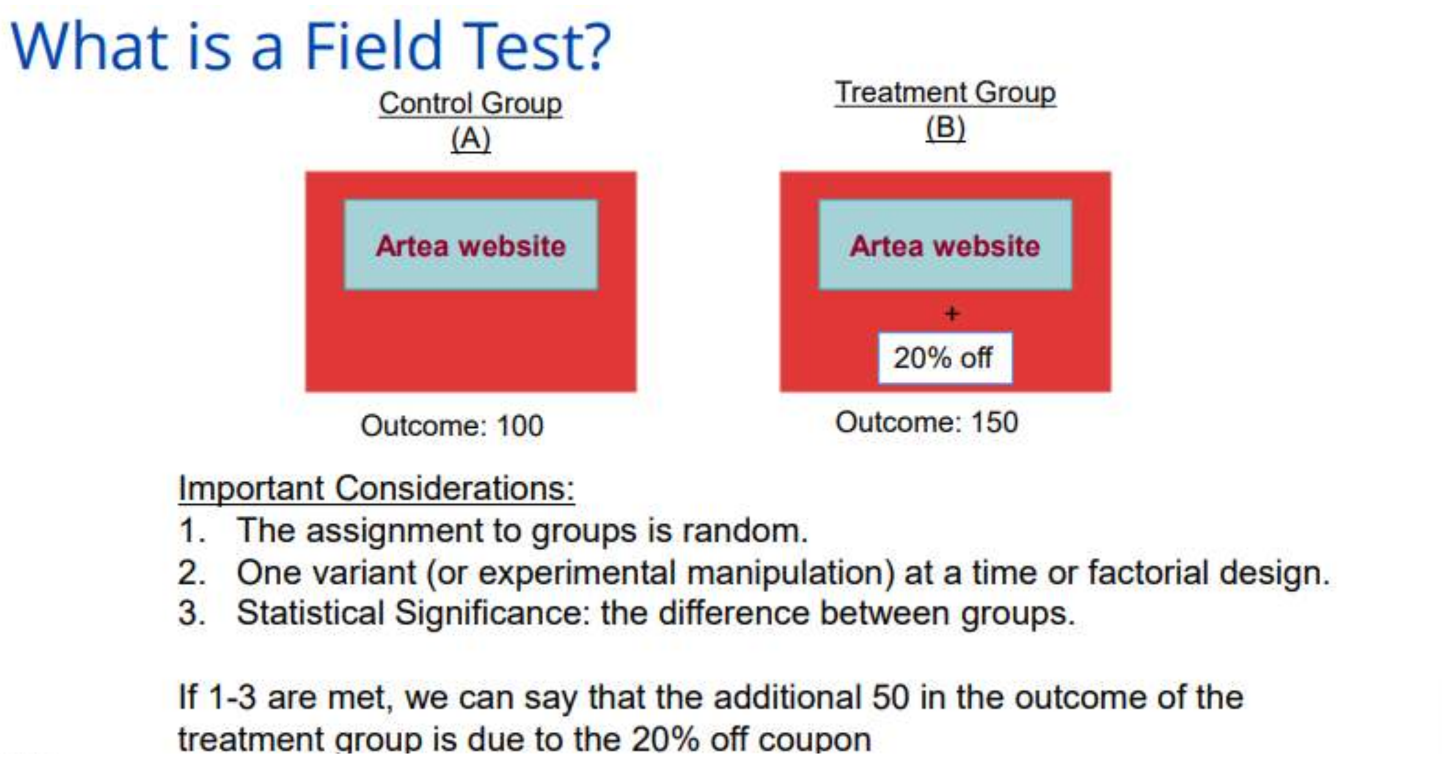
- Industry: Clothing and Accessories
- Data Driven Culture – Data Science Team – Customer Dashboard
- Business Problem:
    - 87% of those who visited the site have never made a transaction.
    - Engagement metrics (e.g. time spent on the site, reviews, etc.) are okay.
    - CEO Alex Campbel wants to increase sales by leveraging the data available from the pre-acquisition phase coming from the website.
    - What could Artea do to improve acquisition?
    - Wait for them to purchase or use strategies to encourage the first purchase?
- **Action**: Alex asked the data science team to explore the possibility of incentivizing purchases, and in this way also improving acquisition, by sending discount coupons to users registered on the site. He decided to run a FIELD test by sending out a 20% discount coupon

Dati field test (AKA A/B test)

5,000 website users
2,502 → treatment group (received the coupon)
2,498 → control group (did not receive the coupon)

Available information:

- dataset containing information on past behavior (purchase and browsing)



A Field test is an experiment, whiich is a procedure in which one or more **variables (treatments)** are **manipulated**, and OBSERVED data related to an "**outcome**" variable of interest (e.g., choice, amount spent, purchase frequency) are collected, while **controlling for other variables** that might distort the result (e.g., consumer characteristics, etc.).

⇨ Without experimentation, there is an association but not causation

FIELD Test Logic: A simple example

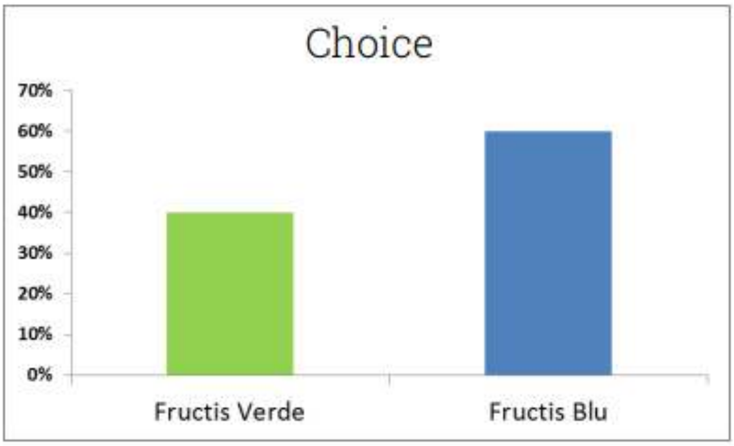
FIELD Test Logic: Causation

Necessary requirements to say that X ----> Y: •

(1) X must occur before Y
(2) There must be evidence of an association between X and Y
(3) Control for other factors.
In an experiment:
(1) X is typically manipulated.
(2) The relationship between X and Y can be estimated by collecting data.
(3) This is typically done through randomization.
Example:

- Color Packaging ---------> Choice

A simple example



Does packaging have an effect on choice?

Experiments:

- Manipulation: The practice related to the creation of different levels of a variable X. Jargon: the variable X is manipulated.
- Independent Variable (X): Variable X manipulated or altered by the "researcher." Example: X= color of the packaging.
- Dependent Variable (O): Variable for which the "experimenter" expects a change following the manipulation of X. The success of the experiment will be evaluated based on the level of O.
- Example: Choice of Brand X (e.g. Fructis)
- Experimental Group (EG): The group of individuals subjected to the experiment

- Control Group (CG): Group of individuals not subjected to the experiment

- Extraneous or Confounding: Variables that, in addition to X, can have an impact on O (e.g., competitors' reactions).
- Selection Bias: A problem that occurs if the experimental group is systematically different in relevant aspects from the control group. In other words, if subjects assigned to the experimental group systematically differ from subjects assigned to the control group

- Randomization (R): Procedure through which subjects/units are randomly assigned to groups (experimental and control)
- Treatment Effect: Result of the experiment
  (e.g. Blue Choice - Green Choice = 0.6 - 0.4 = 20%)
- Experimental Design: Set of procedures that guide the experimental study.
  6 Relevant Steps:
  1. Which variables do you want to manipulate/check (e.g. packaging)
  2. Which levels of variable X need to be manipulated (e.g. color)
  3. What is the dependent variable (e.g. choice)
  4. How to select the units to be tested
  5. How to control for selection bias
  6. How to minimize the influence of external factors.

Artea: Data

Nome	Descrizione
id	Unique identification code of the customer or potential customer
trans_after	Number of transactions after the experiment
revenue_after	Total revenues ($) after the experiment
test_coupon	Dummy variable that takes the value of 1 if the customer or potential customer received the coupon
num_past_purch	Number of previous purchases
spent_last_purchase	Total amount spent ($) in previous purchases
weeks_since_visit	Number of weeks since the last visit to the site
browsing_minutes	Total minutes spent on the site during the last visit
shopping_cart	Indicates if the user added a product to the cart during the last visit but did not make a purchase (1=yes, 0=no)
Channel of Acquisition	Refers to the channel corresponding to the first contact with Artea (registration on the site), therefore acquired as a prospect and not as a customer:
channel_Facebook	Indicates if the customer was acquired via Facebook (1=yes, 0=no)
channel_Instagram	Indicates if the customer was acquired via Instagram (1=yes, 0=no)
channel_Referral	Indicates if the customer was acquired via Referral (1=yes, 0=no)
channel_Other	Indicates if the customer was acquired through other channels (1=yes, 0=no)

Questions

1. Could Artea increase transactions due to this campaign? Can Artea increase the average spending ($) per customer? If so, by how much?
2. Who among the registered users should receive the coupon?

Artea: Designing Targeting Strategy

SUGGESTED STEPS IN THE ANALYSIS:

Step 0: Open the dataset 'Artea.xls'

```
#Step 0:Open dataset Artea.xls
db_artea = pd.read_excel('DataArtea.xlsx')
db_artea.head()
```

	id	trans_after	revenue_after	test_coupon	num_past_purch	spent_last_purchase	weeks_since_visit	browsing_minutes	shopping_cart	channel_Facebook	channel_I
0	6001	0	0.0	0	6	62.99	6	1	0	1	
1	6002	0	0.0	1	2	53.99	0	7	1	0	
2	6003	0	0.0	1	3	88.98	3	4	0	1	
3	6004	0	0.0	0	1	68.99	1	19	0	1	
4	6005	0	0.0	1	3	66.49	4	20	0	0	

```
db_artea.shape
```

```
(5000, 14)
```

Step 1: Verify the random assignment

Objective: We calculate the average values of the observable variables before the field test, comparing them between the control group and the treatment group. If the allocation is random, we should expect two nearly identical groups.

```
# Step 1: Check for correct random allocation in the Control vs. Experimental (or coupon) group
desc = ['num_past_purch', 'spent_last_purchase', 'weeks_since_visit', 'browsing_minutes', 'shopping_cart', 'channel_Facebook', 'channel_Instagram', 'cha
db_artea.groupby("test_coupon")[desc].mean()
```

	num_past_purch	spent_last_purchase	weeks_since_visit	browsing_minutes	shopping_cart	channel_Facebook	channel_Instagram	channel_Referral	channel_
test_coupon									
0	2.019616	56.689480	3.182946	13.707366	0.299840	0.206165	0.314251	0.045236	0.0
1	2.091527	58.155823	3.256994	13.670663	0.285771	0.217426	0.306555	0.049960	0.0

We compared the control group with the experimental group for observable variables before the field test experiment. We note that there are no evident differences in the averages. This seems to confirm that the groups were correctly selected randomly, but we can still test this formally. We could choose single t-tests that compare the coupon=1 group vs. coupon=0. (See the results in Table 2 as an example with one variable).

```
# Step 1: we can run a t-test for each variable

import statsmodels.stats.weightstats as smws

# Split the data into two groups based on 'test_coupon'
group1 = db_artea[db_artea['test_coupon'] == 0]['num_past_purch']
group2 = db_artea[db_artea['test_coupon'] == 1]['num_past_purch']

# Perform t-test
t_stat, p_value, df = smws.ttest_ind(group1, group2, alternative='two-sided', usevar='pooled')

print(f"T-statistic: {t_stat}")
print(f"P-value: {p_value}")
```

```
T-statistic: -0.9930013925760496
P-value: 0.3207573469477035
```

```
# step 1: try with channel facebook, which is a dummy so we need a t-test for proportions:

from statsmodels.stats.proportion import proportions_ztest

# Calculate the number of successes (channel_Facebook = 1) for each group
count_group1 = db_artea[db_artea['test_coupon'] == 0]['channel_Facebook'].sum()
count_group2 = db_artea[db_artea['test_coupon'] == 1]['channel_Facebook'].sum()

# Calculate the total number of observations for each group
nobs_group1 = len(db_artea[db_artea['test_coupon'] == 0])
nobs_group2 = len(db_artea[db_artea['test_coupon'] == 1])

# Perform the z-test for proportions
z_stat, p_value = proportions_ztest([count_group1, count_group2], [nobs_group1, nobs_group2])

print("Z-statistic:", z_stat)
print("P-value:", p_value)
```

```
Z-statistic: -0.9744409774178499
P-value: 0.3298376316653264
```

```
db_artea.groupby("test_coupon")[['trans_after', 'revenue_after']].mean()
```

	trans_after	revenue_after
test_coupon		
0	0.125701	7.780168
1	0.151878	7.538673

Note that the difference is not significant. This is the result we expected. Randomization ensures that the two groups are identical, meaning there are no significant differences in characteristics and behaviors prior to the field test between the treated group and the control group.

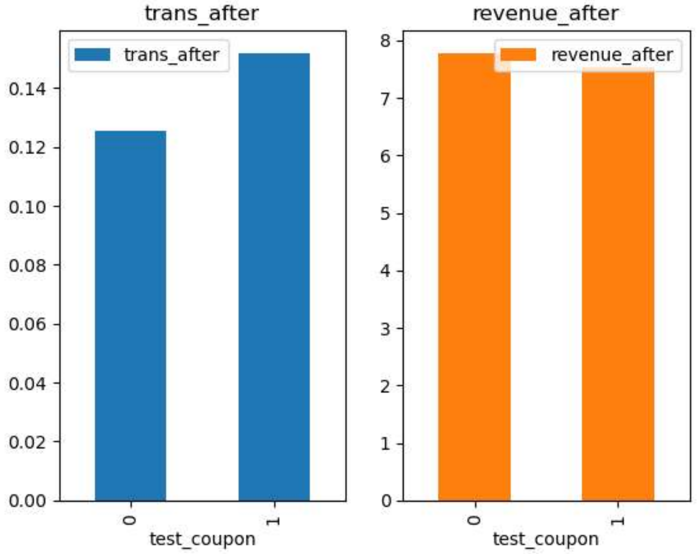
Step 2: Model Free Evidence: Check the effect of Coupon through descriptive statistics

Objective: We want to check if the variables observed after the experimental treatment, namely "trans_after" and "revenues_after", show different average values for the Treated vs. Control group.

Let's start with a descriptive analysis and observe the difference in means to determine if noticeable differences exist.

```
#Step 2: Model Free Evidence - Effectiveness of the COUPON
db_artea.groupby("test_coupon")[['trans_after', 'revenue_after']].mean().plot.bar(subplots=True, layout=(1,2))

#The parameter 'subplots=True' splits the result into two separate plots, one for 'trans_after' and the other for 'revenue_after'.
#Using 'layout=(1,2)' organizes the two plots in a single row and two side-by-side columns.
```

```
array([[<AxesSubplot:title={'center':'trans_after'}, xlabel='test_coupon'>,
        <AxesSubplot:title={'center':'revenue_after'}, xlabel='test_coupon'>]],
      dtype=object)
```

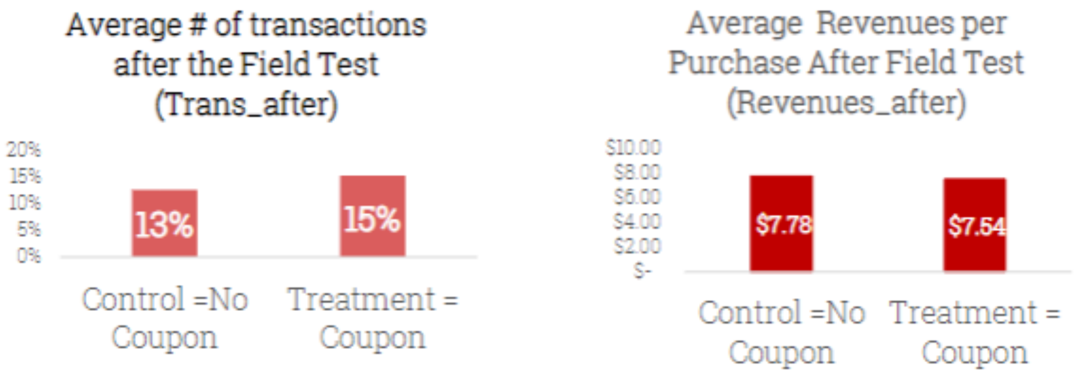


The difference between the average "trans_after" of the treatment group and the control group is 13% -10% = 3%.
3% represents the effect of the experimental treatment (also called TREATMENT EFFECT).
The difference between the average "revenues_after" of the experimental group and the control group is 7.78

- 7.53 = 0.24.
The differences appear very small. We will check in step 3 if they are statistically significant.

Step 2: Model Free Evidence – Efficacia del Coupon

Descriptive Statistics



+ 2% Are these differences significant? **- 0.24**

What is the error associated with the estimate?

Step 3: Test the effect of coupons on consumer behavior

We can then more "formally" test using a statistical test to compare means of independent samples if the difference in means is "statistically significant". In this case, given it's a randomized experiment, we can perform a t-test for independent samples for each outcome variable. Alternatively, two regressions can be conducted with dependent variables first being "trans_after" and then "revenues_after", and the independent variable being "test_coupon". The results will align.

```
# STEP 3: TEST COUPON EFFECTIVENESS ON TRANS_AFTER USING A REGRESSION

import statsmodels.formula.api as smf

model = smf.ols(formula='trans_after ~ test_coupon', data=db_artea).fit()

print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            trans_after   R-squared:                       0.001
Model:                            OLS   Adj. R-squared:                  0.001
Method:                 Least Squares   F-statistic:                     4.872
Date:                Tue, 24 Sep 2024   Prob (F-statistic):             0.0273
Time:                        01:56:56   Log-Likelihood:                -2748.1
No. Observations:                5000   AIC:                             5500.
Df Residuals:                    4998   BIC:                             5513.
Df Model:                           1
Covariance Type:            nonrobust
===============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------
Intercept      0.1257      0.008     14.982      0.000       0.109       0.142
test_coupon    0.0262      0.012      2.207      0.027       0.003       0.049
==============================================================================
Omnibus:                     3814.872   Durbin-Watson:                   1.963
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            66979.526
Skew:                           3.610   Prob(JB):                         0.00
Kurtosis:                      19.413   Cond. No.                         2.62
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

```
# Step 3: we can run a t-test alternatively

import statsmodels.stats.weightstats as smws

# Split the data into two groups based on 'test_coupon'
group1 = db_artea[db_artea['test_coupon'] == 0]['trans_after']
group2 = db_artea[db_artea['test_coupon'] == 1]['trans_after']

# Perform t-test
t_stat, p_value, df = smws.ttest_ind(group1, group2, alternative='two-sided', usevar='pooled')

print(f"T-statistic: {t_stat}")
print(f"P-value: {p_value}")
```

```
T-statistic: -2.207188281608223
P-value: 0.027346188300061674
```

```
# STEP 3: TEST COUPON EFFECTIVENESS ON REVENUES_AFTER USING A REGRESSION

model = smf.ols(formula='revenue_after ~ test_coupon', data=db_artea).fit()

print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          revenue_after   R-squared:                       0.000
Model:                            OLS   Adj. R-squared:                 -0.000
Method:                 Least Squares   F-statistic:                    0.1306
Date:                Tue, 24 Sep 2024   Prob (F-statistic):              0.718
Time:                        02:00:55   Log-Likelihood:                -22906.
No. Observations:                5000   AIC:                         4.582e+04
Df Residuals:                    4998   BIC:                         4.583e+04
Df Model:                           1
Covariance Type:            nonrobust
===============================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------
Intercept       7.7802      0.473     16.456      0.000       6.853       8.707
test_coupon    -0.2415      0.668     -0.361      0.718      -1.552       1.069
==============================================================================
Omnibus:                     3946.015   Durbin-Watson:                   1.968
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            74926.257
Skew:                           3.764   Prob(JB):                         0.00
Kurtosis:                      20.406   Cond. No.                         2.62
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

STEP 3: Effect of Coupon # of Transactions

Y= # of transactions after Field Test

	Coefficients	Standard Error	t Stat	P-value
Intercept	0.13	0.01	14.98	0.00
test_coupon	0.03	0.01	2.21	0.03

We can include more variables to control the effect on Y, but since the coupon test is randomized, its impact will not change significantly.

# Transactions	Coef	SE	t Stat	P value
channel_Other	0.16	0.04	4.36	0.00
channel_Referral	0.13	0.03	4.96	0.00
channel_Instagram	0.11	0.01	8.39	0.00
channel_Facebook	0.11	0.01	7.83	0.00
shopping_cart	0.17	0.01	14.41	0.00
browsing_minutes	0.00	0.00	3.91	0.00
weeks_since_visit	-0.02	0.00	-8.56	0.00
spent_last_purchase	0.00	0.00	-3.43	0.00
num_past_purch	0.06	0.00	24.62	0.00
test_coupon	0.03	0.01	2.38	0.02
Intercept	-0.06	0.02	-3.56	0.00

STEP 3: Test effect of coupon

Y= # of transactions after Field Test

	Coefficients	Standard Error	t Stat	P-value
Intercept	0.13	0.01	14.98	0.00
test_coupon	0.03	0.01	2.21	0.03

Y= revenues after Field Test

	Coefficients	Standard Error	t Stat	P-value
Intercept	7.78	0.47	16.46	0.00
test_coupon	-0.24	0.67	-0.36	0.72

STEP 3: Effect of Coupon # of Transactions

Y= # of transactions after Field Test

	Coefficients	Standard Error	t Stat	P-value
Intercept	0.13	0.01	14.98	0.00
test_coupon	0.03	0.01	2.21	0.03

We can include more variables to control the effect on Y, but since the coupon test is randomized, its impact will not change significantly.

# Transactions	Coef	SE	t Stat	P value
channel_Other	0.16	0.04	4.36	0.00
channel_Referral	0.13	0.03	4.96	0.00
channel_Instagram	0.11	0.01	8.39	0.00
channel_Facebook	0.11	0.01	7.83	0.00
shopping_cart	0.17	0.01	14.41	0.00
browsing_minutes	0.00	0.00	3.91	0.00
weeks_since_visit	-0.02	0.00	-8.56	0.00
spent_last_purchase	0.00	0.00	-3.43	0.00
num_past_purch	0.06	0.00	24.62	0.00
test_coupon	0.03	0.01	2.38	0.02
Intercept	-0.06	0.02	-3.56	0.00

The coupon causes a 0.03 pp increase in transaction likelihood.
The coupon does not have a significant effect on revenues.

Step 4: Heterogeneity in responsiveness to coupon

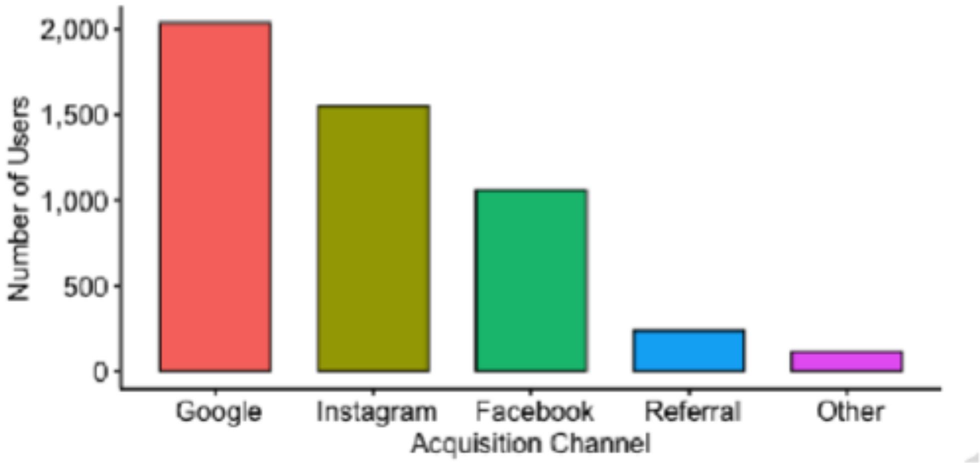
Purpose: understand if the Coupon is more/less effective for customers with certain observable characteristics. We can start by looking at the effect of the coupon by different channels of acquisition

We have variables in the database capturing whether the customer was acquired via Instagram, Facebook, Referral, etc.

```
#Step 4: Explore customer heterogeneity
# start with different channels of acquisition
db_artea.groupby(["channel", 'test_coupon'])[['trans_after', 'revenue_after']].mean()
```

channel	test_coupon	trans_after	revenue_after
Facebook	0	0.159223	9.814913
	1	0.220588	10.895294
Google	0	0.078870	4.922288
	1	0.058532	2.803897
Instagram	0	0.146497	9.171783
	1	0.204694	10.337346
Other	0	0.258621	14.859483
	1	0.241379	12.756690
Referral	0	0.185841	11.179558
	1	0.240000	11.518080

Channels of Acquisition



— Model Free- Evidence:

— **By comparing different sub-groups of the population across treatment and control we can learn which groups are most effected by the intervention**

	Acquired via Instagram		
	Control	Coupon	Difference
Trans_after	0.15	0.20	6%
Revenues_after	9.17	10.34	1.17
Total	1552		

	Acquired via Google		
	Control	Coupon	Difference
Trans_after	0.08	0.06	-2%
Revenues_after	4.92	2.80	-2.12
Tot Customers	2035		

```
#regression with interaction effects channels

model = smf.ols(formula='trans_after ~ test_coupon + num_past_purch + spent_last_purchase + weeks_since_visit + browsing_minutes+ shopping_cart + channel

print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            trans_after   R-squared:                       0.179
Model:                            OLS   Adj. R-squared:                  0.177
Method:                 Least Squares   F-statistic:                     77.66
Date:                Tue, 24 Sep 2024   Prob (F-statistic):          2.96e-201
Time:                        02:07:24   Log-Likelihood:                -2257.3
No. Observations:                5000   AIC:                             4545.
Df Residuals:                    4985   BIC:                             4642.
Df Model:                          14
Covariance Type:            nonrobust
==================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------------
Intercept                       -0.0461      0.019     -2.379      0.017      -0.084      -0.008
test_coupon                     -0.0112      0.017     -0.663      0.507      -0.044       0.022
num_past_purch                   0.0571      0.002     24.613      0.000       0.053       0.062
spent_last_purchase             -0.0004      0.000     -3.434      0.001      -0.001      -0.000
weeks_since_visit               -0.0203      0.002     -8.527      0.000      -0.025      -0.016
browsing_minutes                 0.0030      0.001      3.859      0.000       0.001       0.005
shopping_cart                    0.1723      0.012     14.374      0.000       0.149       0.196
channel_Facebook                 0.0870      0.021      4.232      0.000       0.047       0.127
channel_Instagram                0.0724      0.018      4.005      0.000       0.037       0.108
channel_Referral                 0.0926      0.038      2.452      0.014       0.019       0.167
channel_Other                    0.1627      0.051      3.166      0.002       0.062       0.263
test_coupon:channel_Facebook     0.0521      0.029      1.802      0.072      -0.005       0.109
test_coupon:channel_Instagram    0.0726      0.026      2.828      0.005       0.022       0.123
test_coupon:channel_Referral     0.0724      0.052      1.386      0.166      -0.030       0.175
test_coupon:channel_Other       -0.0076      0.073     -0.104      0.917      -0.150       0.135
==============================================================================
Omnibus:                     3189.104   Durbin-Watson:                   1.960
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            45914.220
Skew:                           2.847   Prob(JB):                         0.00
Kurtosis:                      16.710   Cond. No.                     1.26e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.26e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

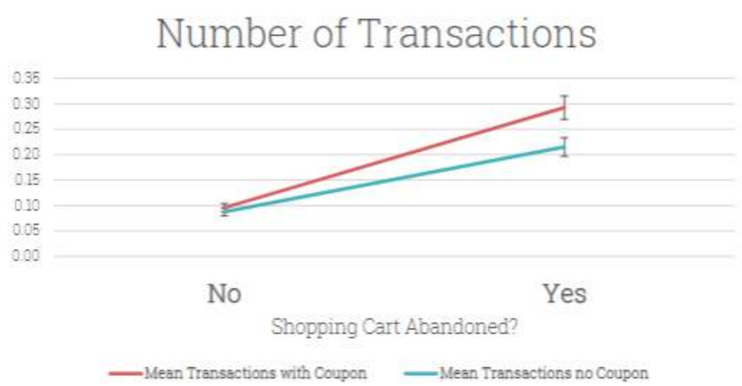
```
model = smf.ols(formula='revenue_after ~ test_coupon + num_past_purch + spent_last_purchase + weeks_since_visit + browsing_minutes+ shopping_cart + chan

print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           revenue_after   R-squared:                       0.172
Model:                             OLS   Adj. R-squared:                  0.170
Method:                  Least Squares   F-statistic:                     73.94
Date:                 Tue, 24 Sep 2024   Prob (F-statistic):          4.96e-192
Time:                         02:07:32   Log-Likelihood:                -22435.
No. Observations:                 5000   AIC:                         4.490e+04
Df Residuals:                     4985   BIC:                         4.500e+04
Df Model:                           14
Covariance Type:             nonrobust
===================================================================================================
                                      coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------------
Intercept                          -2.1951      1.095     -2.004      0.045      -4.343      -0.048
test_coupon                        -1.6230      0.955     -1.699      0.089      -3.495       0.249
num_past_purch                      3.1716      0.131     24.185      0.000       2.915       3.429
spent_last_purchase                -0.0144      0.006     -2.378      0.017      -0.026      -0.003
weeks_since_visit                  -1.1121      0.135     -8.241      0.000      -1.377      -0.848
browsing_minutes                    0.1586      0.045      3.551      0.000       0.071       0.246
shopping_cart                       9.2953      0.678     13.706      0.000       7.966      10.625
channel_Facebook                    5.2746      1.163      4.536      0.000       2.995       7.554
channel_Instagram                   4.5153      1.023      4.413      0.000       2.510       6.521
channel_Referral                    5.4818      2.136      2.566      0.010       1.294       9.669
channel_Other                       9.0358      2.906      3.109      0.002       3.338      14.734
test_coupon:channel_Facebook        1.5461      1.634      0.946      0.344      -1.657       4.749
test_coupon:channel_Instagram       2.9593      1.452      2.039      0.042       0.114       5.805
test_coupon:channel_Referral        2.3094      2.954      0.782      0.434      -3.481       8.100
test_coupon:channel_Other          -0.6657      4.112     -0.162      0.871      -8.726       7.395
==============================================================================
Omnibus:                     3307.920   Durbin-Watson:                   1.961
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            48295.917
Skew:                           2.995   Prob(JB):                         0.00
Kurtosis:                      16.998   Cond. No.                     1.26e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.26e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

By including in the regression interaction effects (i.e. 0 1 variables that identify whether the customer was acquired via a specific channel & received the coupon) we can estimate the differential impact.

To whom should the coupon be sent?

	Y= Number of transactions after the field test (trans_after)		Y= revenues after field test (revenues_after)	
	Coefficients	P-value	Coefficients	P-value
Intercept	-0.05	0.02	-2.20	0.05
test_coupon	-0.01	0.51	-1.62	0.09
num_past_purch	0.06	0.00	3.17	0.00
spent_last_purchase	0.00	0.00	-0.01	0.02
weeks_since_visit	-0.02	0.00	-1.11	0.00
browsing_minutes	0.00	0.00	0.16	0.00
shopping_cart	0.17	0.00	9.30	0.00
channel_Facebook	0.09	0.00	5.27	0.00
channel_Instagram	0.07	0.00	4.52	0.00
channel_Referral	0.09	0.01	5.48	0.01
channel_Other	0.16	0.00	9.04	0.00
Facebook & Coupon	0.05	0.07	1.55	0.34
Instagram & Coupon	0.07	0.00	2.96	0.04
Referral & Coupon	0.07	0.17	2.31	0.43
Other & Coupon	-0.01	0.92	-0.67	0.87



```python
# INTERACTION EFFECTS WITH SHOPPING CART
formula = 'trans_after ~ test_coupon + num_past_purch + spent_last_purchase + weeks_since_visit + browsing_minutes + channel_Facebook + channel_Instagra
ols_model = smf.ols(formula, data=db_artea).fit()
print(ols_model.summary2())
```

```
                 Results: Ordinary least squares
==========================================================================
Model:                  OLS                 Adj. R-squared:        0.178
Dependent Variable:     trans_after         AIC:                   4535.4102
Date:                   2024-09-24 02:07    BIC:                   4613.6165
No. Observations:       5000                Log-Likelihood:        -2255.7
Df Model:               11                  F-statistic:           99.26
Df Residuals:           4988                Prob (F-statistic):    6.45e-205
R-squared:              0.180               Scale:                 0.14469
--------------------------------------------------------------------------
                           Coef.  Std.Err.    t     P>|t|   [0.025  0.975]
--------------------------------------------------------------------------
Intercept                 -0.0529   0.0185 -2.8553 0.0043 -0.0892 -0.0166
test_coupon                0.0009   0.0128  0.0687 0.9452 -0.0242  0.0260
num_past_purch             0.0572   0.0023 24.7021 0.0000  0.0527  0.0618
spent_last_purchase       -0.0004   0.0001 -3.3978 0.0007 -0.0006 -0.0002
weeks_since_visit         -0.0206   0.0024 -8.6255 0.0000 -0.0252 -0.0159
browsing_minutes           0.0032   0.0008  3.9951 0.0001  0.0016  0.0047
channel_Facebook           0.1116   0.0144  7.7301 0.0000  0.0833  0.1398
channel_Instagram          0.1080   0.0129  8.3752 0.0000  0.0827  0.1333
channel_Referral           0.1299   0.0261  4.9785 0.0000  0.0788  0.1811
channel_Other              0.1590   0.0363  4.3787 0.0000  0.0878  0.2302
shopping_cart              0.1312   0.0167  7.8593 0.0000  0.0985  0.1640
shopping_cart:test_coupon  0.0846   0.0237  3.5716 0.0004  0.0381  0.1310
--------------------------------------------------------------------------
Omnibus:                3183.367         Durbin-Watson:            1.958
Prob(Omnibus):          0.000            Jarque-Bera (JB):         45608.537
Skew:                   2.842            Prob(JB):                 0.000
Kurtosis:               16.661           Condition No.:            549
==========================================================================
```

Customer development & retention

We are in the middle phase of the customer journey.

Wrap-up: Artea

The purpose of the Artea case was to focusing on the beginning of the journey in order to find astrategy to acquire new customers who are not engaged with our firm.
What conclusions have you drawn from this case?

- Customer Acquisition -> find a way to increase acquisition and the number of new customers. In this case they implemented discounts (new adv campaign, pricing strategy, testimonial could be other ideas).

- Randomized Field Test can be used to evaluate marketing effectiveness -> They were creating two groups, one (Experimental group) was associated with the discount the other one (Control group) no, in order to understand the number of new customers in both groups and so verify is the strategy proposed was effective or not.

- Strategies for managing customer heterogeneity (diversity of the individuals) -> overall the discount, which was the strategy implemented to acquire new customers, but what we want to verify here is whether the strategy is significant for all the customers or not. Usually anything fit perfectly to everybody, so that is why we have to divide the dataset in groups in order to verify for which groups the strategy is effective basing on the groups' characteristics => BETTER TARGETING:

it is important to managing customer heterogeneity because the strategy is never successful for everybody, so we need to identify groups with different response to the discounts, to do so we have to run in Python interactive variables. The main goal is to better customize the experience for the customers according to their characteristics, behavior and expectat

How is this done? The basic intuition
A simple predictive model for targeting

Better Targeting

SEPHORA

Number of offers mailed:	1,000,000
Profit contribution per response:	$80
Cost per mailing:	$.70
Response rate:	1%

Profit = 1,000,000 x .01 x $80 – 1,000,000 x $.70
= $800,000 - $700,000
= $100,000

The direct marketing campaign is effective!

Most of the investment in direct marketing is wasted!

The campaign has positive profits, and it is effective, but at the same time, in terms of marketing investments is not profitable because the amount of money spent is to high.
So how can we do better? -> if we are able to find those individuals who responds more and targetedthem, we will be able to reduce costs, while still increasing profits.

Targeting analysis: lift-based approach

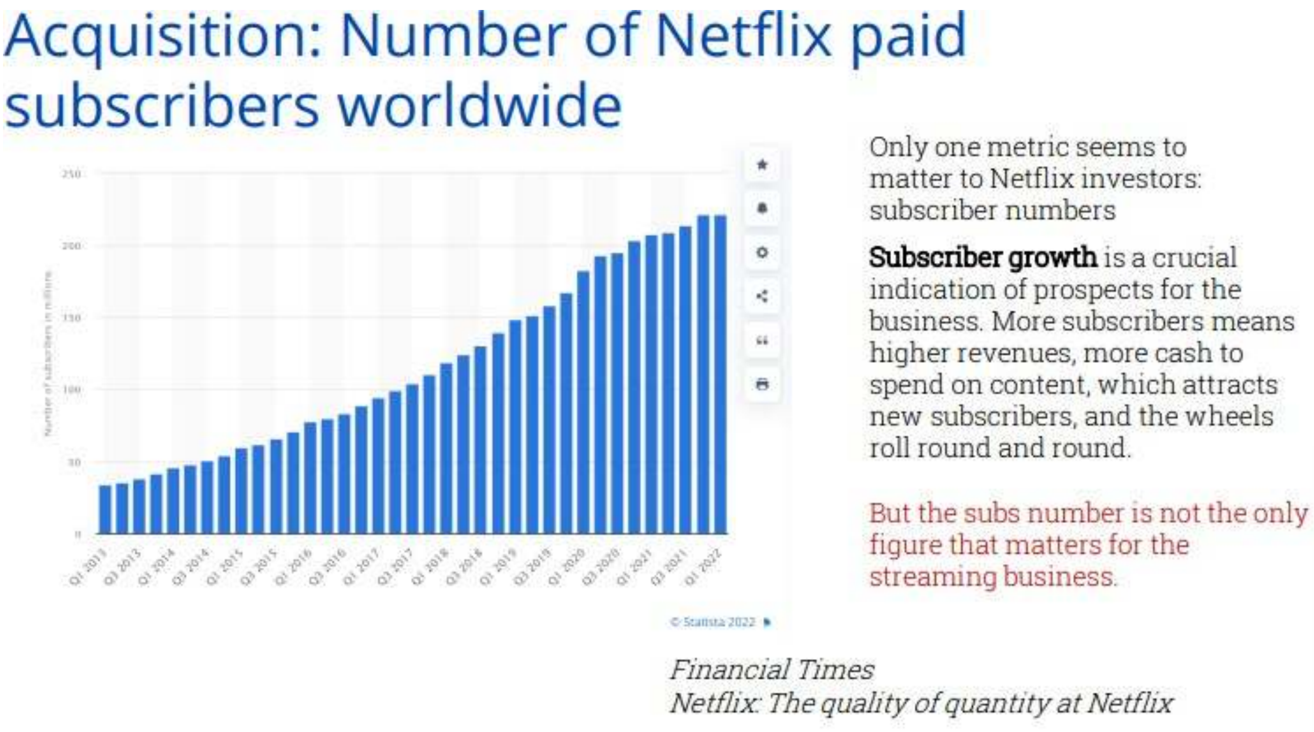
170000= (3%*100000*80)-(100000*0.7)

Decile	Number of Prospects	PredictResp Rate	Profit	Cumulative Profit
1	100,000	3.00%	$170,000	$170,000
2	100,000	2.00%	$ 90,000	$ 260,000
3	100,000	1.40%	$ 42,000	$ 302,000
4	100,000	1.15%	$ 22,000	$ 324,000
5	100,000	1.00%	$ 10,000	$ 334,000
6	100,000	0.60%	$ -22,000	$ 312,000
7	100,000	0.40%	$ -38,000	$ 274,000
8	100,000	0.30%	$ -46,000	$ 228,000
9	100,000	0.10%	$ -62,000	$ 166,000
10	100,000	0.05%	$ -66,000	$ 100,000

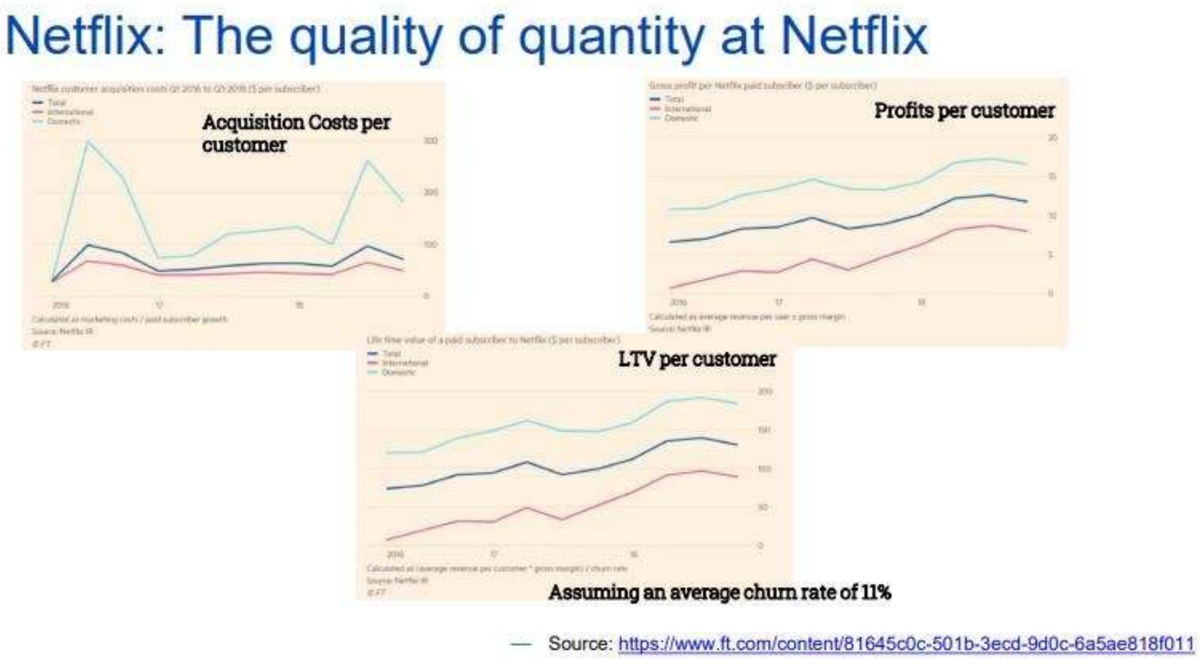
Better Targeting

→Profits Improvement → $100,000→ $334,000

From group 6 we start losing money, so if we stop at group 5 the cumulative profit is so much better than before simple because we are saving money while gaining 334,000$ more.



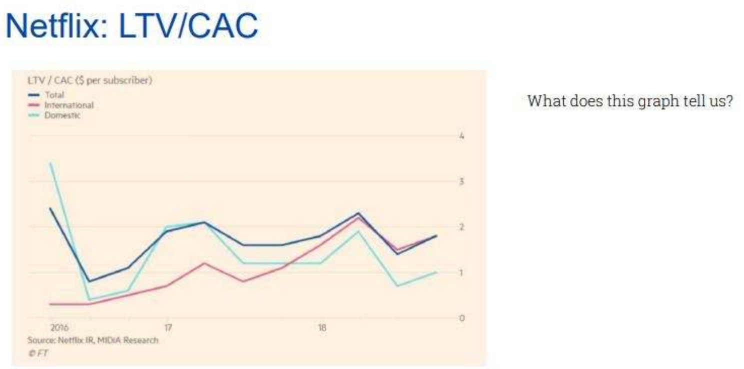
The company is advertising a lot, particularly to the investors who are growing in terms of numbers of subscribers.
But then they focused on a different element: "*Is this enough*?" Actually no! -> Acquisition is not the only important metric to take in account, for example when we talk about Netflix's subscriber we will not only want to know how many new customers we are to acquire but also the average profits provided by each of the new costumers is very important.



To acquire a new customer, they spend a lot of money.
Has we can see the profits for customer are not that great… they are declining.
Customer lifetime value = is a metric that takes into account not only the profit but also the likelihood of

customer to stay with the company alongside -> it is a very important metric, a key performance indicator able to tell the manager of a firm how much each customer that we have acquired is going to value for us in the future. As we can see also this metric has a declining trend.

We can also combine these two metrics to create a new metric, which is the ratio between the cost of acquisition and the customer lifetime value



Source: FT, The quality of quantity at Netflix, October 2018

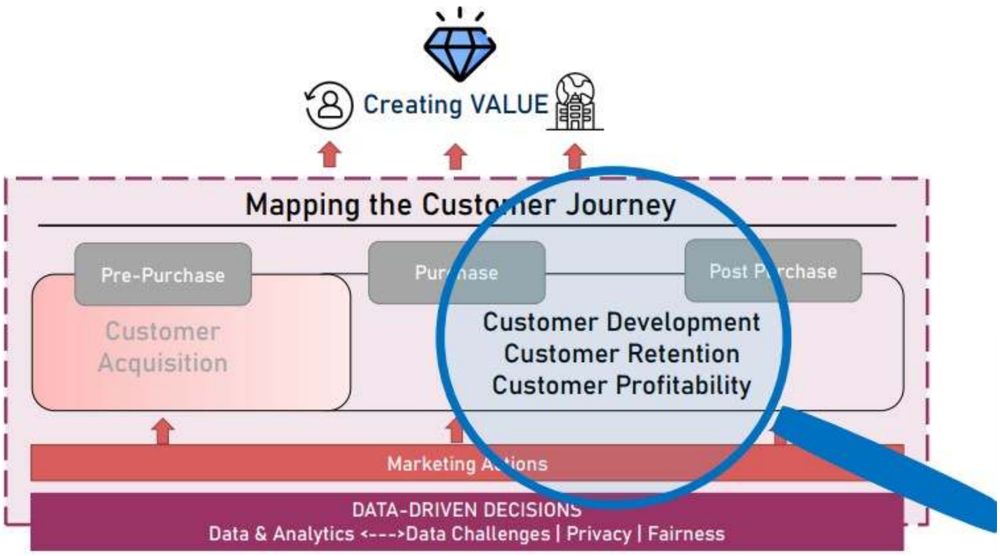
This value is pretty stable, but definitely not that great as the acquisition rate is increasing.

Acquisition & Retention: Key Metrics for Business Health

We need to acquire new customer if we want to grow, but at the same time we want to make sure that we are able to retain those customers -> is less expensive for a firm to work on the current customer base.

Low retention rates + high customer acquisition costs = significant financial impact

- In key sectors like telecom, TV streaming, finance/insurance, and health, customer retention is paramount
- Annual retention rates in wireless phone providers typically range from 70% to 85% (which is not super: the company loses, at the end of each year, 30% of their customer. This means that they have to acquire at least the same amount of customer that decide to leave the business. But even if the are able to acquire the same amount loss is not good, because acquire new customer is more expensive than retain current customer base)



CUSTOMER DEVELOPMENT: What do we mean?

Creating value for both the customer and the organization (increasing customer lifetime value): How?

> Customer lifetime value must increase -> how profitable is each customer for us

> In order to increase customer lifetime value we have to create value for the customer means that we are in a good shape in terms of retention -> if customer are satisfied, even the firm will be satisfy, so it is a win-win situation

CUSTOMER DEVELOPMENT: WHAT DO WE MEAN?

- Increase Customer Lifetime Value: aka CLV or LTV (lifetime value of a customer)
- Different ways to compute it:

The CLV calculation is an estimate of the expected value of the customer's value over time

$$CLV = \sum_{t=1}^{\infty} \frac{m_t r^{t-1}}{(1+\delta)^{t-1}}$$

$$CLV = \sum_{t=0}^{\infty} \frac{m(1-c)^{t-1}}{(1+\delta)^{t-1}} = \frac{m(1+\delta)}{(\delta+c)}$$

churn

It's a forward-looking concept that focuses on the future instead of the past to evaluate or measure a customer's profitability.

m=Revenues - Costs
d=Discount rate
r= Retention
c=1-r=Churn

$$CLV = \frac{m(1+d)}{1+d-r}$$

Blattberg et al (2008)

First, define each component of the metric:
Margin (m) = revenues – cost (include all kind of costs, so both fixed and variable one, as well as both production costs and marketing costs linked to the product)
Delta (d) = discount of rate (financial concept -> if I have this money now which will be their value in the future -> time value of the money)
Retention rate (r) = retention (a value always between 0 and 1 -> likelihood for each consumer to be retained)
Churn rate (c) = 1-r = churn (likelihood for each consumer to live the business in the next year)
Time (t) = years

Some assumptions:

- margin must be constant so everything can be simplified in

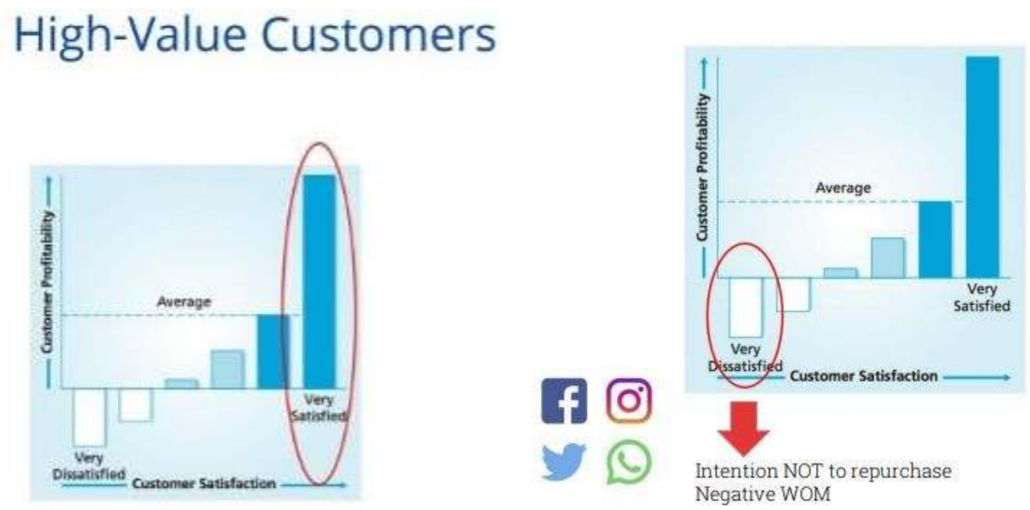
$$CLV = \frac{m(1+d)}{1+d-r}$$

We can impact positively the customer lifetime value by:

- **better targeting**: we know that the same marketing strategy can be better and more effectivefor certain customers rather than the others -> if we are able to cut marketing costs in a cleverway, we can *reduce cost in the parameter m* of customer lifetime value
- **increase customer satisfaction**: positive impact, even though it is a indirect effect, on the likelihood to be retained
- **increase frequency of purchase**: if customers are spending more, we can increase the spending component (m) of the customer lifetime value thanks to higher revenues
- **increase the volume of each purchase occasion** (instead of spending 100 spend now 200)
- **share of wallet**: increase the portion of purchases for our brand while reducing, in doing so, the amount spent at our competitors

We can try to hand up with the strategy that increase **upselling** and **cross selling**

Cross-selling: I have an iPhone but not a Mac; if the company can induce me to buy also the laptopthis is a cross selling
Upselling is inducing the customer to buy a better version of the same item (iPhone 15 -> iPhone 16)



Very satisfied customers are more profitable customers, while low satisfied customers are not only less profitable, but they can also negatively impact on the brand by doing bad reviews, and so on.

High-value customers = Satisfied Customers

Customer Satisfaction	Customer Percent	CSI Score	Customer Revenue	Percent Margin	Gross Profit	Retention Cost	Customer Profit
Very Satisfied	25%	100	$1,200	60%	$720	$100	$620
Satisfied	35%	80	$800	50%	$400	$100	$300
Somewhat Satisfied	20%	60	$300	40%	$120	$100	$20
Somewhat Dissatisfied	15%	40	$80	40%	$32	$100	–$68
Dissatisfied	3%	20	$60	40%	$24	$100	–$76
Very Dissatisfied	2%	0	$50	40%	$20	$100	–$80
	100%	72	$655	49%	$350	$100	$250

Satisfaction & Customer Development→ Increase Value of the Customer

High-value customers = Satisfied Customers

On average acquiring new customers costs more than retaining them

N.B. Satisfaction is not equal to retention, so we will always lose also satisfied customer.

Dissatisfied customers who complain can be retain if we have a great customer service, but if they do not complain is so much harder for the company to implement strategies to retain them -> they stay only if searching and switching cost are too high for them to change business.
If we have customer that do complain that means that we still have possibilities to do better and fix these problem-> **IMPLEMENT A STRONG SERVICE RECOVERY:** give the customers, the possibility and the private space to complain, without using the web in order to then implement the best strategies to fix the problems raised.

RETENTION OR CHURN

Customer Lifetime=1/(1-RetentionRATE)

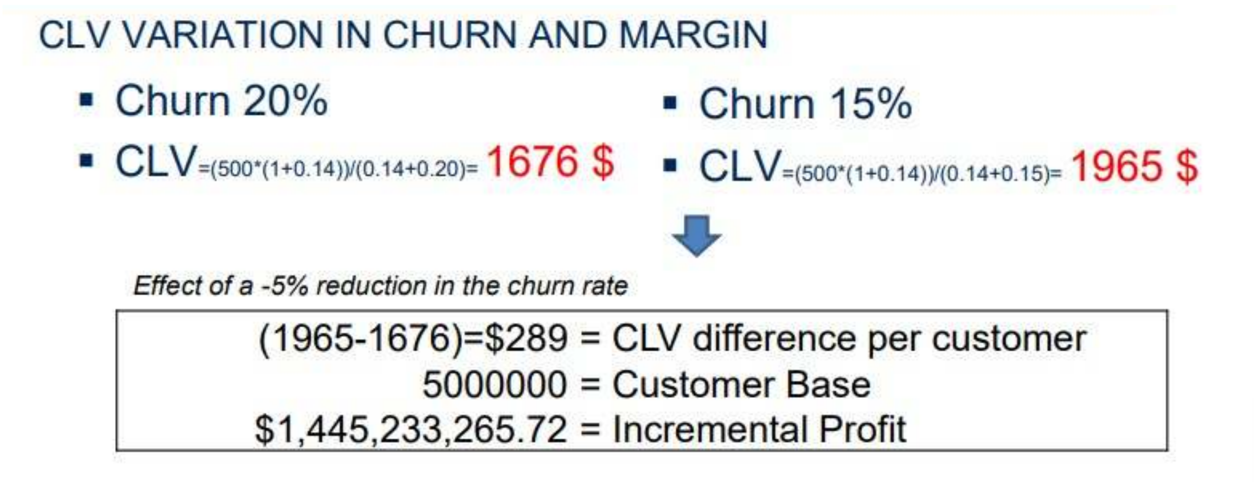
Churn Rate=1-Retention Rate

Year	Customer Retention	Customer Life
2002	72%	3.6
2003	77%	4.3
2004	80%	5.0
2005	82%	5.6
2006	85%	6.7

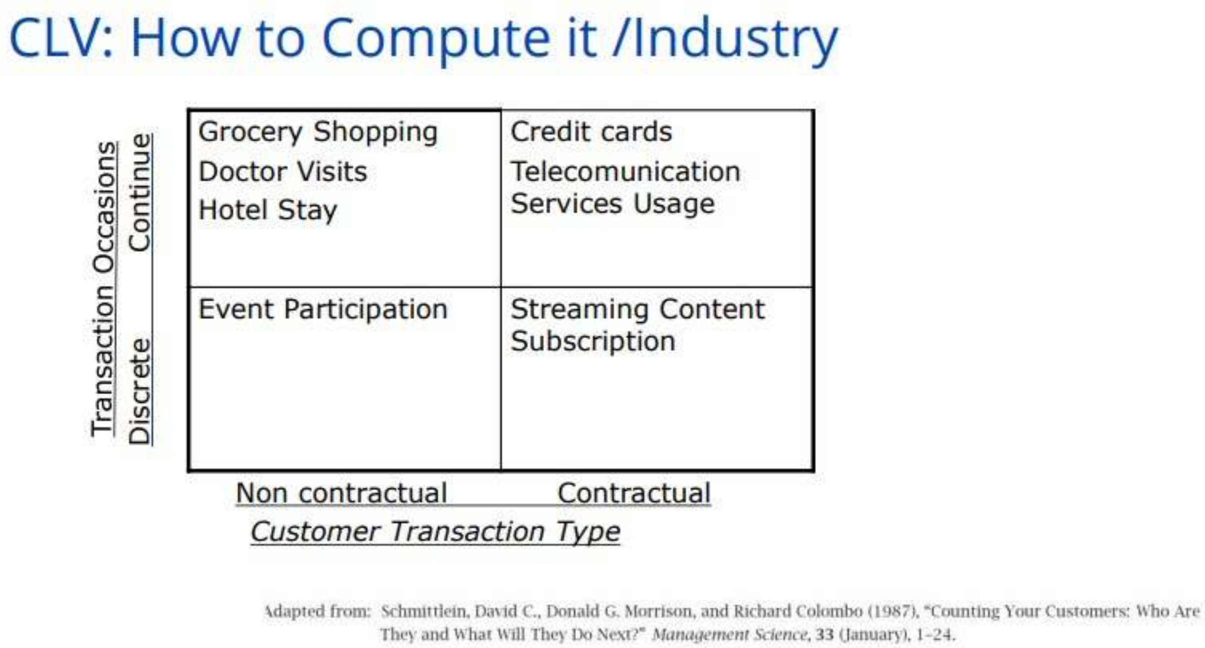
CLV & CHURN: CONVEX FUNCTION

$$CLV = \frac{m(1+d)}{1+d-r}$$

This function is convex, which means that if we are able to reduce the churn rate we can gain much more in terms of customer life time value -> A small increase in terms or retention is going to produce a great increase in terms of lifetime value.



The difference in terms of CLV is small when reducing the churn rate, if we consider the single customer, but when we compute the total incremental profit, we can see a huge improvement, eventough the churn rate reduction was really small.



To compute CLV we necessarily need the **churn rate** or otherwise the **retention rate**.

So, if we are in a contractual business, it is pretty easy to compute it, because we know the number of contracts each year -> at the end of each year the company perfectly know the number of customers that they were able to retain.
This is not true in not-contractual businesses because without a contract is difficult for the businessto understand and figure out when the customer leaves the business -> in this case the measure of CLV is not of course a precise value.

CUSTOMER DEVELOPMENT

e.g. Up-selling, Cross-Selling, Frequency of purchase, SOW, Customer Profitability

$$\underset{LTV}{CLV} = \frac{m(1+d)}{1+d-r}$$

Reduce Churn, Work on Retention & Loyalty

Better targeting is another strategy that can reduce the costs and so increase *m*. Different strategies can have different impact on some key component of the CLV.

- How do companies react to having a 'churn problem' or a problematic LTV/CAC ratio

Churn Management

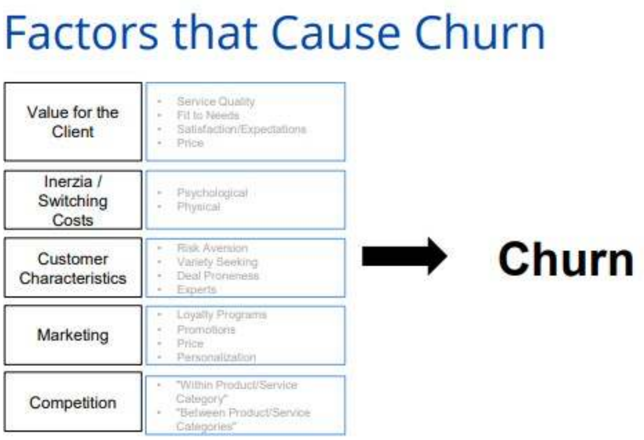
Objective: Focus on the 'Retention' Component of LTV
Unit of Analysis: Customer

- $Churn_i$ = Probability that an individual *i* will leave the company in a given period (ex. The end of the year)

Unit of Analysis: Firm

- Churn= percentage of the customer base that leaves the company in a given period
- Churn=c=1-Retention Rate=1-r

What leads to churn? What are the contributing factors?



Value for the client

- Service quality
- Fit to needs
- Satisfaction/expectations
- Price
- if services/products have a value for me, if I'm satisfied with them, if I find the price in line with my expectation about it.

Inertia/switching cost:

- physical) -> sometimes we have problems to change with a positive impact on retention
- psychological -> inertia

Competition:

- within product/service category
- between product/service categories
- the more competitors we have in a market, the more difficult, even for the best firm, is to retain customers

Marketing:

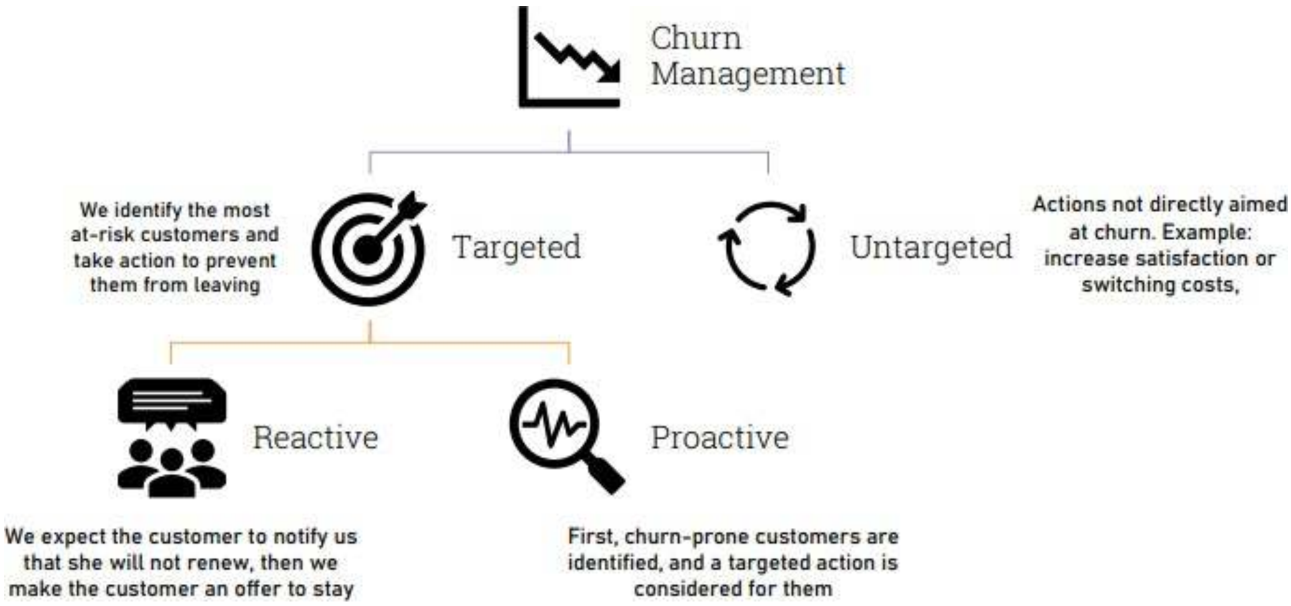
- loyalty
- promotions
- price
- personalization
- if the marketing actions are effective, they have a negative impact on churn by increasing retention.

Customer characteristic:

- risk aversion
- variety seeking
- deal proneness
- experts
    - if we understand that a group is more likely to churn, we have to understand how to acquire them bringing them in our business.

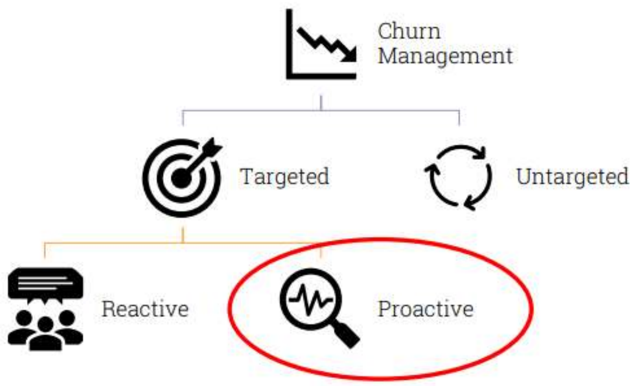
These are all the macro factor that we know for sure have an impact on retention. So as a strategy we can decide to focus on one or more of them.

How to Manage Churn?



There are two macro strategy that we can adopt:

1. Hand up with churn management **untargeted strategy** -> the key idea if that action are not directly aimed at churn -> I identify an action to do not directly aimed to reduce churn, but to increase something else to have a positive impact on churn (ex. Increase customer satisfaction because if I increase that indirectly I hope to have a positive in reducing churn; increase switching cost to indirectly have a positive impact on churn)
2. Hand up with churn management **targeted strategy** -> completely different mindset becausethe first purpose is to identify the consumers that are at risk of churn. We effectively create a strategy to reduce the risk of churn applying it to only those customers at risk of churn. I have two option to react:
    a. *Targeted reactive*: we don't do anything until the customers notify us that they want to leave, and in that moment, we apply our strategy -> ex. Sky
    b. *Targeted proactive*: I have a metric, so I know who at rick on churn, but before they contact me, I'm gonna do something because I fell that waiting is risking.



They key step is we want to apply this strategy is to identify **in advance** those who are at risk of churnbyusing a **predictive model**:

1. Predictive model: goal is to predict individual-level churn
2. Proactive churn management actions by using creative strategies

The ‘process’ of developing a predictive model

The development of a predictive model is primarily a **process** (here is where we can demonstrate our ability as manager), which extends beyond mere model estimation and consists of **several phases**:

a. problem definition
b. data preparation
c. model estimation
d. model evaluation
e. verification of predictive capability...
    - What is the **key question** we want to answer?
    - What **data** do I need to make these predictions? -> Now we have a dataset provided by the teachers but maybe it will not be the case in the future and have to provide evidence that the data the we want to buy are crucial for the purpose of retain customers
    - What **statistical technique** should I use?
    - Does it work/**Do I trust the model**? -> Develop a predictive model the works well is fundamental
    - Does it have any **managerial relevance**? Can I use the results?"

1. **DEFINING THE RPOBLEM**
    a. Defining the managerial problem
2. **PREPARING THE DATA**
    a. Identifying the behavior ton predict
    b. Constructing the dataset
    c. Preprocessing the data
3. **ESTIMATING THE MODEL**
    a. Selecting predictors
    b. Choosing the modeling technique
    c. Estimating the model
    d. Evaluating the model
4. **PREDCTIVE VALIDITY**
    a. Calibration and Validation
5. **TARGETING**
    a. Lift (e.g.>1) cutoff
    b. Individual scores

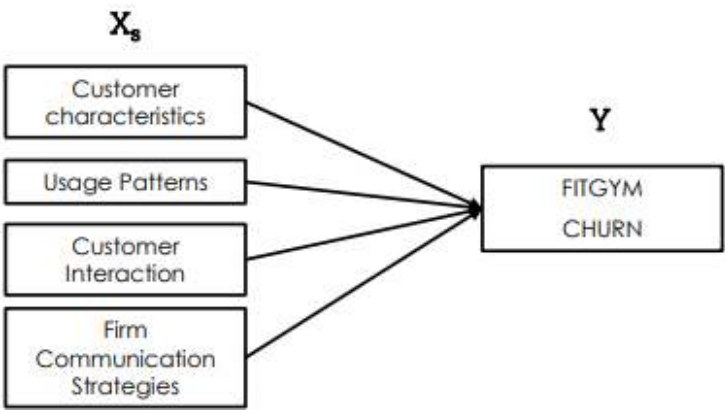
EXAMPLE - FitLife Gym predictive churn model: PREPARING THE DATA

We have to prepare a predictive model for FitLine Gym, which means list all the data that we think could be useful to predict the risk of churn in this industry and for this firm.

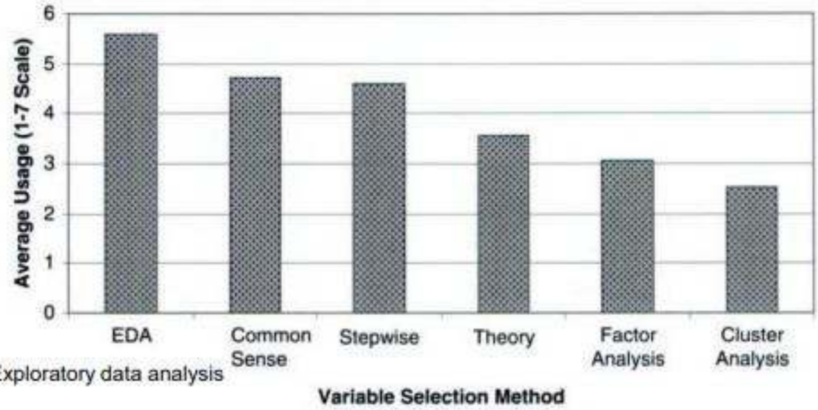
1) **Usage Patterns**: Customers who haven't visited the gym in the past month and have a declining trend in their monthly attendance might be at risk of churning.
2) **Customer Demographics**: Young professionals or students might be more likely to churn due to changing schedules, location, or financial constraints (we can get these data by the subscription that they fill to subscribe to the gym).
3) **Customer Interaction**: Customers who have raised complaints or concerns through the gym's feedback system multiple times without satisfactory resolution might be dissatisfied and consideringleaving.
4) **Member Engagement**: Assessing the level of engagement, such as class attendance, participationin fitness challenges, or app usage for workout tracking, can provide insights into a member's likelihood to continue or churn
5) **Communication**: Evaluating the effectiveness of communication channels (e.g., email, SMS, app notifications) in engaging and retaining customers.
6) **Referral Source**: Analyzing how members were acquired (e.g., referral, marketing campaign, walk-ins). Members referred by existing satisfied customers may have lower churn rates

At the end of this phase, we will be able to create the model.

Question: Which factors can help predict churn?



SELECTING PREDICTORS (I.E. Xs)



EDA=Exploratory data analysis

Neslin et al. 2006

Methods most used by a sample consisting of 50% academics and 50% managers

How do I select talking about the independent variables in my prediction model?
Both **statistical theory** and **intuition** could be use in order to select the Xs of the model.

Example: Among the independent variable we have *age*, but the only values registered in the datasetare 25-26 -> it is basically a constant, so I don't have to add this variable in my model.
Or again we have an apparently interesting information about the number of individuals acquired through referral (1% of the sample. It means that we have a dummy variable that is always 0 except fora tiny portion of 1 -> it is basically constant; it seems interesting but because it does not vary it is betternot include them.
We might exclude some variable that don't change very much in order to have a better model with justrelevant information.
Common sense could use as well -> I think that maybe some variables are useful, so I add them inthe model.
Moreover, we have theory that can help us in choosing our variables.
When we have a huge number of Xs is typical to have problems when including all of them in themodel, especially the multicollinearity one due to the possible correlation among variables -> sometimes people want to reduce the dimension by creating common factors or **cluster analysis**.
Stepwise is another possible approach: the regression is done step by step, introducing all the variable in the list step-by-step trying to understand the best model with only significant variables by analysing the R^2 every time -> the teacher is skeptical to use this model as a first approach; if adopting this model we need to pay attention in order to add only what is significant without skiping information that could be us

Typically, we check the R^2 and then we select the solution with the varibale which optimizes it.
Remember that when we use Logit we have just a pseudo R^2 and not the real value associated to this parameter.
The value of the R^2 tent to increase when we have a lot of variables, and must be biased which means that have to many variables is not always good for our model, not that our model explains better the situation.

SELECTING PREDICTORS (I.E. Xs)

- All possible variable subsets (e.g., X1, X2, X3). Selection based on R2, AIC, BIC

The model with the highest R^2 or the lowest AIC or BIC is chosen

$$AIC = -2LL + 2k$$
$$BIC = -2LL + 2\ln(N)k$$

$$pseudoR^2 = 1 - \left(\frac{LL_{full}}{LL_0}\right)$$

$$R^2 = \sqrt{\frac{DevR}{DevTot}} = 1 - \sqrt{\frac{DevErrore}{DevTot}}$$

To avoid the biased R^2 that we can get when we have to many varibales we use AIC and BIC. LL is the maximum likelihod calculated through the Logit function.

SELECTING PREDICTORS (I.E. Xs)

- Issues?
- Too many predictors vs. Too few predictors

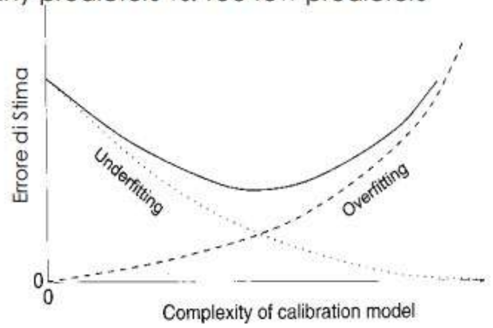


Illustration of two important concepts in the development of a predictive model:

- Overfitting
- Underfitting

Underfitting: not enough varibale in order to fit the model.
Overfitting: we include in the model too many varibales. The model runs, but it is not significant; some varibale are correlated to each other, or some of them are too similar.

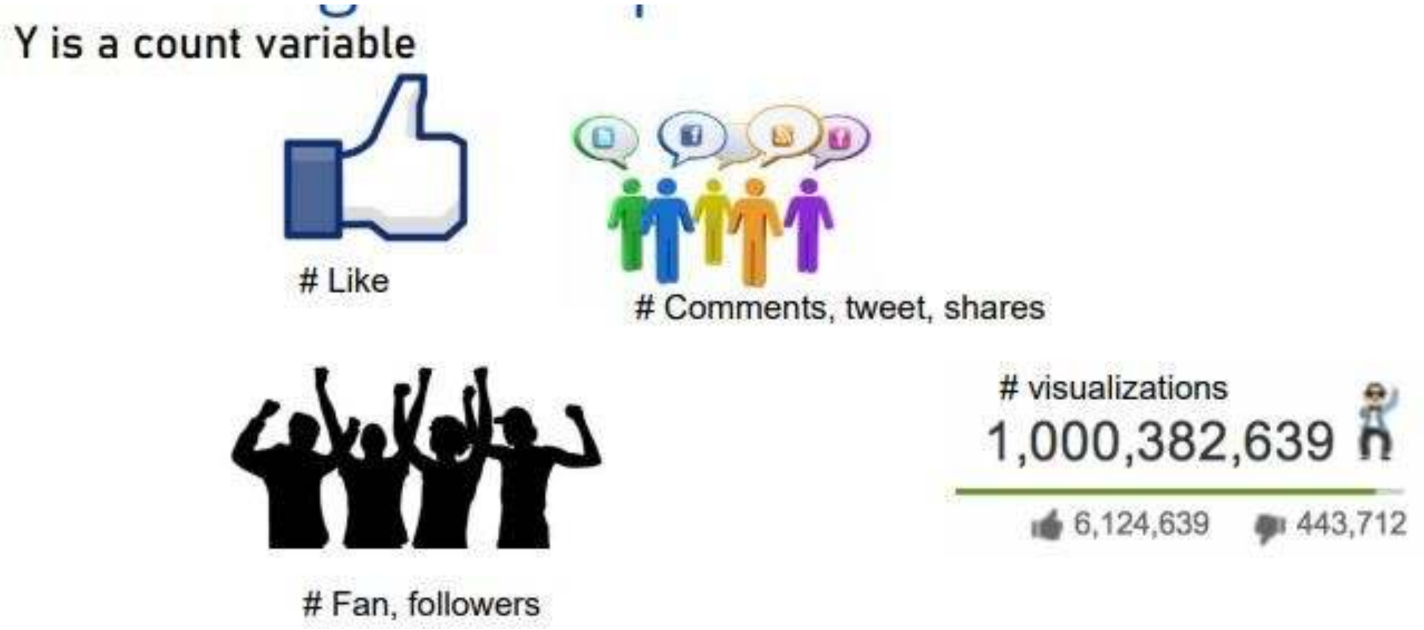
Estimating the model: choosing the modelling technique

We can choose among different possibilities:

- OLS REGRESSION
- Logit/Probit
- MNL/MNP/Conditional
- Multivariate Logit/Probit
- Tobit (I, II)
- Hazard—Poisson, NBD

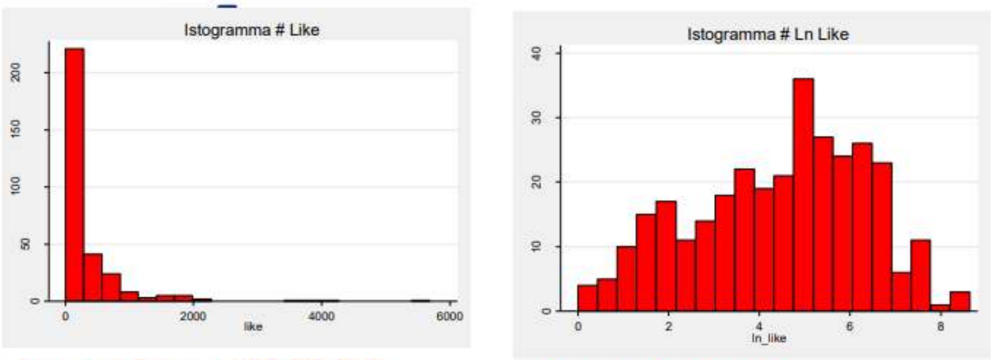
EVERYTHING (or almost everything) DEPENDS ON Y! -> We have to understand which type of variable is Y in order to understand which is the right model to use

We will use an *OLS* if the data are *normally distributed*.



In this case the best model to use could be *OLS* after a transformation of the data, or otherwise use *more advance models*.

Possible alternative when the dependent variable is a count variable



Advanced: Poisson, NBD, ZIP, ZINB **Ln(Like) OLS Regression**

In this case where 0 and 1 are just labels, we can only apply a *logit model.*

⇨ Y tell us which model is the right one to use

Calibration & Validation

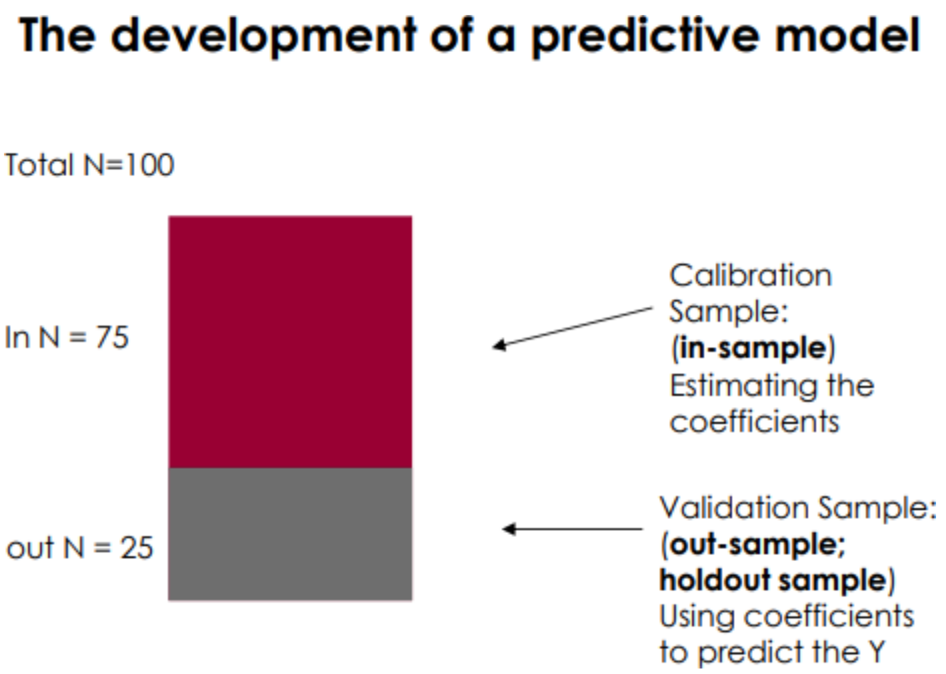
Our purpose is to develop a strategy to manage churn, more specifically targeted proactive which means identify in advance those that are at risk of churn -> we need to predict who is more likely to churn, so here the task is to create a great predictive model, and we need to test the predictability of the model.

The dependent variable is a qualitative nominal variable with 2 categories

Models: Logit / Probit

How to do it?
Typically, we have a dataset; imagine a dataset of 100 customers. Of course, whatever we have in the dataset happened in the past which means that I can use previous data to develop a predictive modelthat maybe is able to predict the behavior of the customer this year. To do this I have to take the data, divide the dataset into two sub-samples -> one to estimate the model and the other one to check if the prediction model predicts well:
A – a «**calibration**» sample also known as in-sample

B – a «**validation**» sample also known as out-sample, on which testing the model



Step 1 - In-sample (Calibration): estimate coefficients (using only 75 individuals and not 100): $Y_{INi}=a+b_1X1_{INi}+b_2X2_{INi}+b_3X3_{INi}+e_{INi}$

Step 2 – Out-sample (Validation): using obtained parameters (a, b_1, b_2, b_3) at step 1 to predict Y_hat_{OUTi} (the predicted y)
Step 3 – compare the estimated values with the observed ones and calculate the prediction error as=(Y_{OUTi} -Y_hat_{OUTi})

Before doing this process there are to question to answer:

1. What should be the sample size?
   - the calibration sample must be higher than the validation one if the model is lower than500 observations -> *Calibration sample = 75%* of the total number of the observations, while the *validation sample = 25%* of the observation
   - Otherwise, if the model has a large number of observations the calibration sample size must be select randomly because we want to create to groups that are identical in every dimension -> *Calibration sample= validation sample= 50%*
2. How much of the sample should be allocated to calibration?

- N< 500 (sample size is relatively small) calibration sample =75%, validation sample= 25%. Why?
- For sizes >=500, any 'reasonable' split works well (Steckel and Vanhonacker 1993).

How to choose the validation and calibration samples? -> RANDOM

Calibration and Validation: EXAMPLE

N=100

```
Y=_const+bX   if CALIBRATION==1

      Source |       SS       df       MS              Number of obs =      75
-------------+------------------------------           F(  1,    73) =  184.02
       Model |   740.7946      1   740.7946            Prob > F      =  0.0000
    Residual | 293.872067     73  4.02564476           R-squared     =  0.7160
-------------+------------------------------           Adj R-squared =  0.7121
       Total | 1034.66667     74   13.981982           Root MSE      =  2.0064

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           X |   1.605073   .1183215    13.57   0.000     1.369258    1.840887
       _cons |   15.35343   1.171641    13.10   0.000     13.01835     17.6885
------------------------------------------------------------------------------
```

Y= SALES
X= Amount Spent ADS ($)

Y= sales a=15.35

X= ADV b= 1.6

I want to predict sales taking into consideration a specific level of advertising.
Sales_hat = 15.35 + 1.6*35(=x1) => here we will have the predicted value of advertising for this level of sales.

Predicting CHURN

Using the model parameters, you can estimate the churn rate at the individual customer level

The Lift Chart

- One way to evaluate predictive capability while also considering a proactive approach is to use lift logic
- Logic:
  - I identify different distinct groups based on the prediction of the dependent variable (Y_hat). Example: 2 groups, 5 groups, 10 groups (quintiles, deciles, etc.) -> the standard is 10 group, and in group number 10 we will have those that are very unlikely to churn while in group 1 there are those who are most likely to churn.
  - I check the mean of the observed variable (Y) in each of these groups.
  - I do this for both the calibration and validation models -> predict both the sample_in and the sampe_out.

- Lift = average response of the (k-th) group, relative to the average response of the sample

  r_k= Average response of the group k (k-le)

  r_0= Average response across the entire sample

$$\lambda_k = \frac{r_k}{r_0}$$

- What does a very high lift indicate?
- it is possible to significantly increase the response probability by targeting the group that is more likely to respond compared to the entire sample.

Decile	Observed Churn	Lift
1	35%	3.02
2	30%	2.59
3	25%	2.16
4	12%	1.03
5	8%	0.69
6	5%	0.43
7	1%	0.09
8	0%	0.00
9	0%	0.00
10	0%	0.00
Totale	12%	

This table is called lift table. In observed churn column we have the real variable while the decile column is what we predict.
If our model tells us that we have the ones who are more likely to change in group, we will expect a higher average than the last group where my model should put the ones at less risk of churn.

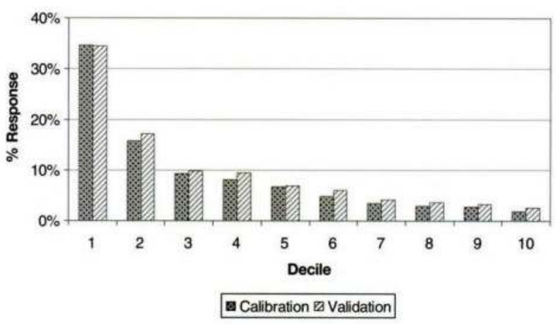
Example : Y=Churn Cliente i

Decile	Observed Churn	Lift
1	35%	3.02
2	30%	2.59
3	25%	2.16
4	12%	1.03
5	8%	0.69
6	5%	0.43
7	1%	0.09
8	0%	0.00
9	0%	0.00
10	0%	0.00
Totale	12%	

Example Lift del Top-Decile=3
It is about 3 times more likely for customers included in the top decile to churn compared to the average

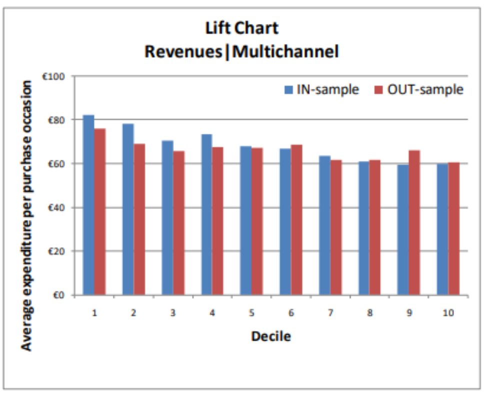
Whit the lift chart we can immediately understand if the model is good or not -> in fact, the model is good when in group 1 we have those who are at more risk with a higher average and, as we go up in terms of group, we can see how both the % of risk of churn and the average value decrease.

How do you interpret this LIFT CHART?



Fine. I would love to have this chart. We can use it to identify in advance those who are in risk of churn.

How do you interpret this LIFT CHART?



So and so because it is flat -> not good news because the model is not able to discriminate amonggroups. The same is when the model is neither increasing nor decreasing

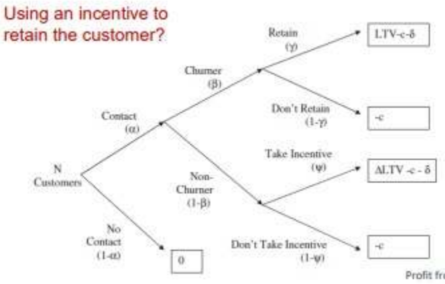
How do you interpret this LIFT CHART?

If our model gives this result we are fine and we can use to predict future trends and behavoir.

How do you identify customers who are much more likely to churn?
Targeting:

- Lift (e.g. >1) cutoff
- Individual Scores

Churn Management: Proactive Management



N= Total number of clients

α=Probability that the customer will be contacted as part of a churn program

β=Probability that the customer is a churner | has been contacted

Y=Probability the customer is retained| Churner

ψ=Probablity that a non-churner receives the incentives

Δ = Increase in LTV among non-churners who received the incentive

c= cost of contacting the customer

δ= cost of the incentive

Profit from a proactive action aimed at reducing churn

$$\begin{aligned}\Pi &= N\{\alpha\beta\gamma(LVC - c - \delta) + \alpha\beta(1-\gamma)(-c) + \alpha(1 \quad (\Delta LVC - c - \delta) \\ &\quad + \alpha(1-\beta)(1-\psi)(-c)\} \\ &= N\alpha\{(\beta\gamma + (1-\beta)\psi\Delta)LVC - \delta(\beta\gamma + (1-\beta)\psi) - c\} \qquad (24.8)\end{aligned}$$

Churn Management: Proactive

$$\begin{aligned}\Pi &= N\{\alpha\beta\gamma(LVC - c - \delta) + \alpha\beta(1-\gamma)(-c) + \alpha(1-\beta)\psi(\Delta LVC - c - \delta) \\ &\quad + \alpha(1-\beta)(1-\psi)(-c)\} \\ &= N\alpha\{(\underbrace{\boxed{\beta\gamma}}_{\text{Incremental profit from recovered customers}} + \underbrace{\boxed{(1-\beta)\psi\Delta}}_{\text{Incremental profit of delighted non-churners}})LVC - \underbrace{\boxed{\delta(\beta\gamma + (1-\beta)\psi)}}_{\text{Cost of incentives for churners and non-churners}} - \underbrace{\boxed{c}}_{\text{Cost of contact}}\}\end{aligned}$$

Maximum incentive cost

$$\boxed{\delta < \frac{(\beta\gamma + (1-\beta)\psi\Delta)LTV - c(1+\rho)}{(1+\rho)(\beta\gamma + \psi(1-\beta))}}$$

N= Total number of clients

α=Probability that the customer will be contacted as part of a churn program

β=Probability that the customer is a churner | has been contacted

Y=Probability the customer is retained| Churner

ψ=Probablity that a non-churner receives the incentives

Δ = Increase in LTV among non-churners who received the incentive

c= cost of contacting the customer

δ= cost of the incentive

These last two slides are not in the exam.

BOOKS R US: CASE STUDY – PREDICTING CHURN

Caso BooksRUs: Churn Management

- BookRus is a major multichannel European book retailer. The company sells books through stores, mail-order, phone and the Internet.

- Subscription base business. Membership renews at the end of each year.

- The firm mails a print catalog to the customer base five times per year

- The company opted for an omnichannel strategy providing members several possibilities to purchase: the physical stores, the website, mail/order or phone, the app

Step 1: Open the dataset DataCohort1.xlsx.

```
# Importing necessary libraries
import pandas as pd
import statsmodels.formula.api as smf
import matplotlib.pyplot as plt
import numpy as np
import os
```

```
#Step 0:Open dataset DataCohort1.xls
file_path = r'C:\Users\ValentiniS\DataCohort1.xlsx'
df = pd.read_excel(file_path)
df.head()
```

	id	churn	profits	multichannel	age	female	street_agent	north	early_email	bigcity	mean_city	franchisee	initialstorepromo	initialweb	initialstore
0	17051200	0	76.169998	1	33.0	1	1	1	1	0	0	0	0	1	0
1	17137573	0	123.320000	0	61.0	0	0	0	1	0	1	0	0	0	1
2	17071854	0	49.459999	0	23.0	1	0	0	1	0	0	0	0	1	0
3	17126628	1	42.689999	0	67.0	1	0	1	1	0	0	0	0	1	0
4	17126606	0	49.060001	0	46.0	1	0	1	1	0	0	1	0	1	0

```
df.columns
```

```
Index(['id', 'churn', 'profits', 'multichannel', 'age', 'female',
       'street_agent', 'north', 'early_email', 'bigcity', 'mean_city',
       'franchisee', 'initialstorepromo', 'initialweb', 'initialstore',
       'initialmobile', 'initialrevenues', 'initialreturns'],
      dtype='object')
```

Step 2: Estimate a logit model with the dependent variable (DV) = churn, and identify the factors that have the greatest impact on churn.

```
# Step 1: Run logit DV=churn
# create a quadratic effect for the variable "initialreturns,"
df["initialreturns2"] = df["initialreturns"]**2
```

Creating a **quadratic effect in a regression** can be useful for capturing non-linear relationships between the predictor variable and the response variable. In many real-world scenarios, the relationship between variables is not always linear and can exhibit curves or bends.

Here are a few reasons why using a quadratic effect (or higher-order polynomial terms) might be beneficial in a regression analysis:

- **Non-linear Relationships:** A quadratic effect allows the model to account for situations where the response variable does not change linearly with the predictor. For instance, some phenomena may show an initial increase at a decreasing rate or vice versa.
- **Capturing Curvature:** When there is a U-shaped or inverted U-shaped relationship between the predictor and the response, a quadratic term can help capture this curvature more accurately.
- **Better Prediction:** In cases where the true relationship is indeed quadratic, incorporating a quadratic term can lead to better predictions by aligning the model more closely with the underlying data distribution.

However, it's essential to exercise caution when using higher-order polynomial terms, as they can introduce overfitting, making the model too complex and less generalizable to new data.

```
# Step 2: Run Logit DV=churn
formula = ('churn ~ multichannel + age + female + street_agent + '
           'north + early_email + bigcity + mean_city + franchisee + '
           'initialstorepromo + initialweb + initialstore + initialmobile + '
           'initialrevenues + initialreturns + initialreturns2')

model = smf.logit(formula, data=df).fit()

print(model.summary2())
```

```
Optimization terminated successfully.
         Current function value: 0.313075
         Iterations 7
                          Results: Logit
=================================================================
Model:              Logit            Pseudo R-squared: 0.037
Dependent Variable: churn            AIC:              22194.0895
Date:               2023-10-09 17:28 BIC:              22338.1511
No. Observations:   35391            Log-Likelihood:   -11080.
Df Model:           16               LL-Null:          -11500.
Df Residuals:       35374            LLR p-value:      1.2991e-168
Converged:          1.0000           Scale:            1.0000
No. Iterations:     7.0000
-----------------------------------------------------------------
                   Coef.  Std.Err.    z     P>|z|   [0.025  0.975]
-----------------------------------------------------------------
Intercept         -1.4676   0.0738 -19.8916 0.0000 -1.6122 -1.3230
multichannel      -1.0227   0.1131  -9.0401 0.0000 -1.2445 -0.8010
age               -0.0126   0.0012 -10.5182 0.0000 -0.0150 -0.0103
female            -0.4504   0.0367 -12.2635 0.0000 -0.5224 -0.3784
street_agent      -0.3603   0.0390  -9.2463 0.0000 -0.4367 -0.2840
north              0.1753   0.0517   3.3918 0.0007  0.0740  0.2766
early_email       -0.0514   0.0388  -1.3263 0.1847 -0.1274  0.0246
bigcity            0.7864   0.0594  13.2274 0.0000  0.6698  0.9029
mean_city          0.1454   0.0701   2.0746 0.0380  0.0080  0.2828
franchisee        -0.0500   0.0387  -1.2920 0.1964 -0.1258  0.0258
initialstorepromo  0.4627   0.0823   5.6248 0.0000  0.3015  0.6240
initialweb        -0.1298   0.1900  -0.6833 0.4944 -0.5023  0.2426
initialstore      -0.5876   0.0833  -7.0549 0.0000 -0.7509 -0.4244
initialmobile     -0.3554   0.1405  -2.5296 0.0114 -0.6308 -0.0800
initialrevenues    0.0064   0.0021   3.0102 0.0026  0.0022  0.0105
initialreturns    -0.0229   0.0133  -1.7185 0.0857 -0.0490  0.0032
initialreturns2    0.0003   0.0001   1.7563 0.0790 -0.0000  0.0005
=================================================================
```

We included at first everything as we don't have to many varibles.
p-value of multichannel is 0.000 -> it is significant, which means that this varibale has some impact on churn. To understand the impact we have to check the sign of the coefficient; in this case is negative which is a good news because it has a negative impact on churn -> reducing churn is exactly what we are looking for.
On the other hand, initialstorepromo, even though it is significant due to a p-value equal to 0 has a positive impact on churn, which means that those who receives a promo at the beginnig are moe likely to churn.

How to interpret the coefficient in a Logistic Regression? POSSIBILITY – COMPUTE OR (exponantial of the coefficients)

The model assumes that:

Y_i = 1 with probability p_i

0 with probability 1-p_i

- p_i represents the probability that the event occurs (e.g. Y_i=1="YES")
- If an event occurs with probability pi, then the odds ratio (OR) will be $p_i/(1-p_i)$ -> average impact of the variable. When OR=1 this is the benckmark of the indiffrence position

Take these two coefficients as examples:

Results: Logit

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
multichannel	-1.0227	0.1131	-9.0401	0.0000	-1.2445	-0.8010
initialstorepromo	0.4627	0.0823	5.6248	0.0000	0.3015	0.6240

$b_{multichannel}$=-1.002

Odds Ratio (OR): Exp(-1.002) ≈ 0.359 The odds ratio is calculated by taking the exponential of the coefficient. An odds ratio of 0.359 means that for a one-unit increase in the "multichannel" variable (i.e. being multichannel), the odds of churn decrease by approximately 64% (because 1 - 0.359 ≈ 0.641, or about 64%).

In simpler terms, being in the "multichannel" category decreases the odds of churn by approximately 64% compared to not being in the "multichannel" category.

$b_{Initial\ promo}$=0.4627

Odds Ratio (OR): Exp(0.4627) ≈ 1.588
The odds ratio is calculated by taking the exponential of the coefficient. An odds ratio of approximately 1.588 means taking advantage of a promotion in the initial period increases the odds of churn by about 1.588 times compared to not being in the "initial promo" category.

In simpler terms, taking advantage of a promo at the beginning of the relationship with the firm increases the odds of churn by approximately 58.8% compared to not taking advantage of a promo at the beginning.

Interpreting Coefficients in a Logistic Regression Model: odds ratio

Imagine x_k = age

- **If β_k is positive**→for example if β_k =1, e^1=2.7:
- We can say that a one-unit increase in age increases the probability of your binary variable y (e.g. retention) by a factor of 2.7 (or 170%), keeping all other factors constant

- **if β_k is negative**→ for example if β_k = -1, e^{-1}=.36,
- we can say that a one-unit increase in age decreases the probability of your binary variable y (e.g. retention) by a factor of 0.36 (or 64% less), keeping all other factors constant.

Step 3: Estimate a model with the DV = profits.

```
#Step 3: Run OLS reg DV=profits
formula = ('profits ~ multichannel + age + female + street_agent + '
           'north + early_email + bigcity + mean_city + franchisee + '
           'initialstorepromo + initialweb + initialstore + initialmobile + '
           'initialrevenues + initialreturns + initialreturns2')

model = smf.ols(formula, data=df).fit()

print(model.summary2())
```

```
                 Results: Ordinary least squares
=====================================================================
Model:              OLS              Adj. R-squared:     0.336
Dependent Variable: profits          AIC:                287829.0750
Date:               2023-10-09 17:28 BIC:                287973.1367
No. Observations:   35391            Log-Likelihood:     -1.4390e+05
Df Model:           16               F-statistic:        1122.
Df Residuals:       35374            Prob (F-statistic): 0.00
R-squared:          0.337            Scale:              199.23
---------------------------------------------------------------------
                      Coef.  Std.Err.    t    P>|t|   [0.025  0.975]
---------------------------------------------------------------------
Intercept             4.6034   0.3116 14.7718 0.0000  3.9926  5.2142
multichannel         20.9517   0.3207 65.3373 0.0000 20.3231 21.5802
age                   0.1509   0.0048 31.4807 0.0000  0.1415  0.1603
female                1.2085   0.1623  7.4482 0.0000  0.8905  1.5265
street_agent          3.1672   0.1585 19.9766 0.0000  2.8565  3.4780
north                 1.9387   0.2022  9.5889 0.0000  1.5424  2.3349
early_email           1.8377   0.1644 11.1769 0.0000  1.5154  2.1599
bigcity               0.1535   0.2579  0.5952 0.5517 -0.3519  0.6589
mean_city             0.0237   0.2711  0.0873 0.9304 -0.5076  0.5550
franchisee            0.4522   0.1596  2.8340 0.0046  0.1394  0.7649
initialstorepromo    -1.8296   0.3417 -5.3540 0.0000 -2.4994 -1.1598
initialweb           11.0348   0.7429 14.8539 0.0000  9.5788 12.4909
initialstore         -2.5538   0.3014 -8.4724 0.0000 -3.1446 -1.9630
initialmobile        10.0674   0.5492 18.3326 0.0000  8.9910 11.1438
initialrevenues       0.5308   0.0086 61.5983 0.0000  0.5139  0.5476
initialreturns       -0.5039   0.0541 -9.3126 0.0000 -0.6099 -0.3978
initialreturns2       0.0021   0.0008  2.7236 0.0065  0.0006  0.0036
---------------------------------------------------------------------
Omnibus:             11760.006       Durbin-Watson:          2.016
Prob(Omnibus):       0.000           Jarque-Bera (JB):       71137.918
Skew:                1.468           Prob(JB):               0.000
Kurtosis:            9.295           Condition No.:          1680
=====================================================================
* The condition number is large (2e+03). This might indicate
strong multicollinearity or other numerical problems.
```

Step 4: Focus on churn and test if the model has predictive ability. Use:

- Calibration and Validation
- Lift Chart

Step 4 involves assessing the model's ability to predict churn events.

One of the evaluation techniques mentioned is the use of a **lift chart**, which visually represents the model's predictive performance. However, please note that there are alternative approaches to evaluate the model, besides the lift chart, that can also be effective.

The evaluation process utilizes both **in-sample** and **out-of-sample**. **In-sample** is the portion used to train or fit the model **out-of-sample** is held separately to evaluate how well the model performs on unseen data.

The creation of in-sample and out-of-sample data involves a **random selection process**. This randomness ensures that the selected samples are representative and unbiased, providing a fair evaluation of the model's performance. Randomly selecting data for both in-sample and out-of-sample evaluations enhances the

robustness of the assessment.

```
# Step 4: Generate random data
df["random"] = np.random.uniform(size=df.shape[0])
df["random"].describe()
# Calculate the median as the cutoff value
cut_off = np.median(df["random"])

# Print the cutoff value
print("Cutoff value:", cut_off)
```

```
Cutoff value: 0.4963136276443664
```

```
# Create the 'out_sample' dummy variable
df["out_sample"] = np.where(df['random'] > cut_off, 1, 0)

# Calculate the frequency table for the 'out_sample' column
frequency_table = df['out_sample'].value_counts()
# Calculate the percentage for each value
percentage_table = (df['out_sample'].value_counts(normalize=True) * 100).round(1)
# Print the frequency & percentage tables
print(frequency_table)
print(percentage_table)
```

```
out_sample
0    17696
1    17695
Name: count, dtype: int64
out_sample
0    50.0
1    50.0
Name: proportion, dtype: float64
```

```
df.head()
```

	id	churn	profits	multichannel	age	female	street_agent	north	early_email	bigcity	...	franchisee	initialstorepromo	initialweb	initialstore	initialm
0	17051200	0	76.169998	1	33.0	1	1	1	1	0	...	0	0	1	0	
1	17137573	0	123.320000	0	61.0	0	0	0	1	0	...	0	0	0	1	
2	17071854	0	49.459999	0	23.0	1	0	0	1	0	...	0	0	1	0	
3	17126628	1	42.689999	0	67.0	1	0	1	1	0	...	0	0	1	0	
4	17126606	0	49.060001	0	46.0	1	0	1	1	0	...	1	0	1	0	

5 rows × 21 columns

```
# I don't need the variable random anymore and i will drop it
df.drop(columns=['random'], inplace=True)
df.head()
```

	id	churn	profits	multichannel	age	female	street_agent	north	early_email	bigcity	mean_city	franchisee	initialstorepromo	initialweb	initialstore
0	17051200	0	76.169998	1	33.0	1	1	1	1	0	0	0	0	1	0
1	17137573	0	123.320000	0	61.0	0	0	0	1	0	1	0	0	0	1
2	17071854	0	49.459999	0	23.0	1	0	0	1	0	0	0	0	1	0
3	17126628	1	42.689999	0	67.0	1	0	1	1	0	0	0	0	1	0
4	17126606	0	49.060001	0	46.0	1	0	1	1	0	0	1	0	1	0

```
# we re-estimate the churn model only for the in-sample

# Filter the DataFrame for in-sample
in_sample_df = df[df['out_sample'] == 0]
# Define your formula
formula = ('churn ~ multichannel + age + female + street_agent + '
           'north + early_email + bigcity + mean_city + franchisee + '
           'initialstorepromo + initialweb + initialstore + initialmobile + '
           'initialrevenues + initialreturns + initialreturns2')
# Estimate the model using the filtered DataFrame
model = smf.logit(formula, data=in_sample_df).fit()
# Print the model summary
print(model.summary2())
```

```
Optimization terminated successfully.
         Current function value: 0.321471
         Iterations 7
                         Results: Logit
=================================================================
Model:              Logit            Pseudo R-squared: 0.037
Dependent Variable: churn            AIC:              11411.5144
Date:               2023-10-09 17:32 BIC:              11543.7930
No. Observations:   17696            Log-Likelihood:   -5688.8
Df Model:           16               LL-Null:          -5905.8
Df Residuals:       17679            LLR p-value:      2.5165e-82
Converged:          1.0000           Scale:            1.0000
No. Iterations:     7.0000
-----------------------------------------------------------------
                   Coef.  Std.Err.    z     P>|z|   [0.025  0.975]
-----------------------------------------------------------------
Intercept         -1.4283   0.1036 -13.7920 0.0000 -1.6312 -1.2253
multichannel      -1.0698   0.1593  -6.7150 0.0000 -1.3820 -0.7575
age               -0.0127   0.0017  -7.6206 0.0000 -0.0160 -0.0095
female            -0.4564   0.0510  -8.9524 0.0000 -0.5563 -0.3564
street_agent      -0.3519   0.0541  -6.5016 0.0000 -0.4580 -0.2458
north              0.2048   0.0729   2.8075 0.0050  0.0618  0.3478
early_email       -0.0697   0.0541  -1.2880 0.1977 -0.1757  0.0364
bigcity            0.7713   0.0840   9.1809 0.0000  0.6067  0.9360
mean_city          0.1572   0.0980   1.6036 0.1088 -0.0349  0.3492
franchisee        -0.0509   0.0538  -0.9452 0.3446 -0.1564  0.0546
initialstorepromo  0.5114   0.1126   4.5431 0.0000  0.2908  0.7321
initialweb         0.0523   0.2454   0.2132 0.8311 -0.4286  0.5333
initialstore      -0.5728   0.1154  -4.9626 0.0000 -0.7991 -0.3466
initialmobile     -0.2108   0.1892  -1.1144 0.2651 -0.5815  0.1599
initialrevenues    0.0050   0.0030   1.6996 0.0892 -0.0008  0.0109
initialreturns    -0.0396   0.0210  -1.8817 0.0599 -0.0808  0.0016
initialreturns2    0.0003   0.0002   1.4655 0.1428 -0.0001  0.0007
=================================================================
```

```
# MY MODEL previously estimated logistic regression model

# Predict probabilities for both in-sample and out-sample
# Predict probabilities for the entire dataset
df['y_hat'] = model.predict(df)
df.head()
```

	id	churn	profits	multichannel	age	female	street_agent	north	early_email	bigcity	...	franchisee	initialstorepromo	initialweb	initialstore	initialm
0	17051200	0	76.169998	1	33.0	1	1	1	1	0	...	0	0	1	0	
1	17137573	0	123.320000	0	61.0	0	0	0	1	0	...	0	0	0	1	
2	17071854	0	49.459999	0	23.0	1	0	0	1	0	...	0	0	1	0	
3	17126628	1	42.689999	0	67.0	1	0	1	1	0	...	0	0	1	0	
4	17126606	0	49.060001	0	46.0	1	0	1	1	0	...	1	0	1	0	

5 rows × 21 columns

```
# create decide for OUT SAMPLE
# Filter the DataFrame for in-sample
out_sample_df = df[df['out_sample'] == 1]

df.loc[df['out_sample'] == 1, 'decile_out'] = pd.qcut(out_sample_df['y_hat'], 10, labels=False)
#CHECK IF DECILES ARE CREATED CORRECTLY
df[["decile_out",  "y_hat"]].groupby("decile_out").mean()
```

decile_out	y_hat
0.0	0.036700
1.0	0.057600
2.0	0.070239
3.0	0.080519
4.0	0.088854
5.0	0.099676
6.0	0.112738
7.0	0.127526
8.0	0.152104
9.0	0.206303

```
#DECILE=1 SHOULD HAVE THE HIGHEST Y_HAT AND 10 THE LOWEST
df.loc[df['out_sample'] == 1, 'decile_out'] = 10 - df.loc[df['out_sample'] == 1, 'decile_out']
#CHECK IF DECILES ARE CREATED CORRECTLY
df[["decile_out",  "y_hat"]].groupby("decile_out").mean()
```

decile_out	y_hat
1.0	0.206303
2.0	0.152104
3.0	0.127526
4.0	0.112738
5.0	0.099676
6.0	0.088854
7.0	0.080519
8.0	0.070239
9.0	0.057600
10.0	0.036700

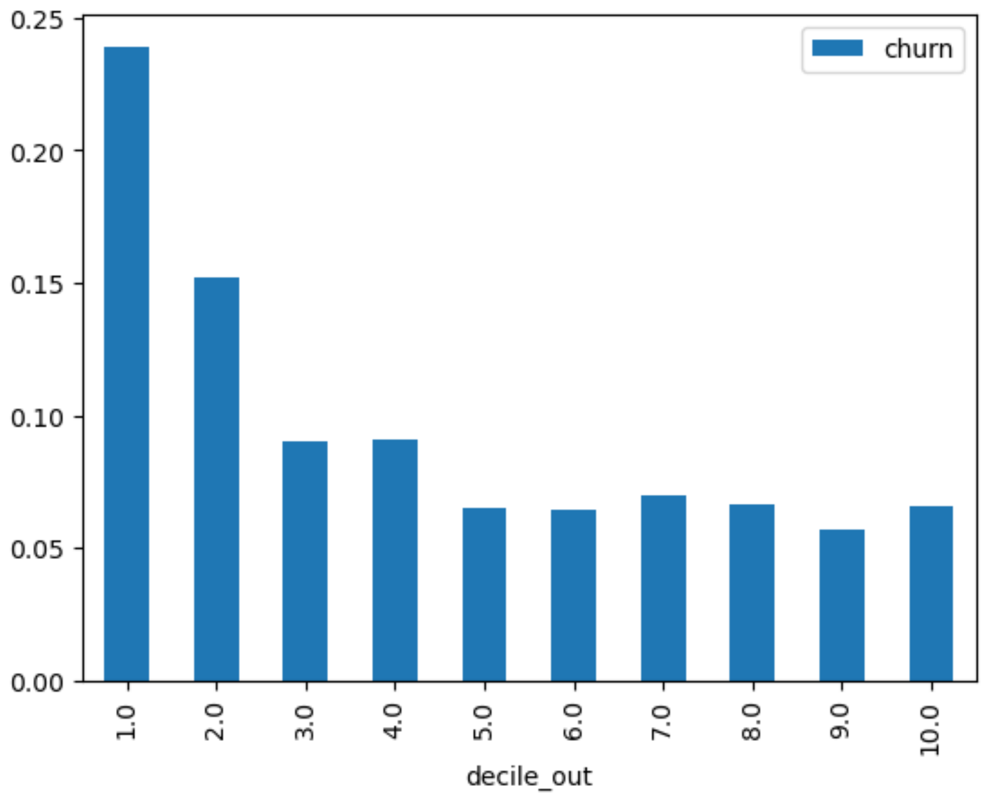
The first group has 20% probability to churn. Now we have to compare the model with the real y which is **churn.**

```
#LIFT TABLE WITH OBSERVED CHURN
df[["decile_out", "churn"]].groupby("decile_out").mean()
```

decile_out	churn
1.0	0.239389
2.0	0.151806
3.0	0.090234
4.0	0.090604
5.0	0.064699
6.0	0.064262
7.0	0.069977
8.0	0.066102
9.0	0.056529
10.0	0.065537

```
# create the lift chart

df[["decile_out",  "churn"]].groupby("decile_out").mean().plot.bar()
```

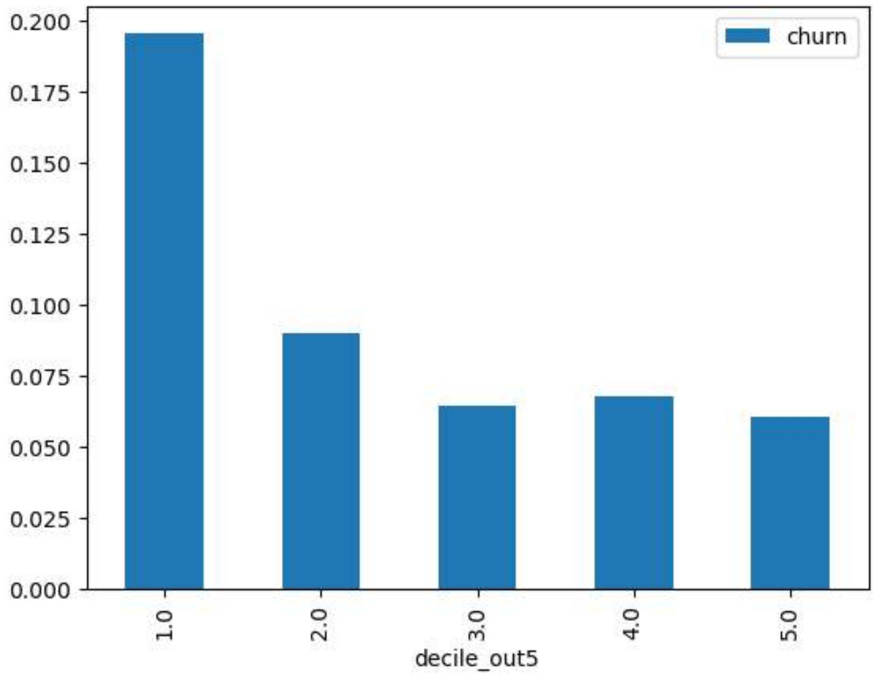
```
<AxesSubplot:xlabel='decile_out'>
```



This model is able to distinguish pretty well the first four/five groups, then the results are pretty much the same, so now we want to use just the significant groups.

```
# lift with 5 groups instead of 10

# Assuming you want to assign 5 groups for 'decile_out5'
out_sample_df = df[df['out_sample'] == 1]

# Calculate deciles for the out-sample with 5 groups
df.loc[df['out_sample'] == 1, 'decile_out5'] = pd.qcut(out_sample_df['y_hat'], 5, labels=False)
df.loc[df['out_sample'] == 1, 'decile_out5'] = 5 - df.loc[df['out_sample'] == 1, 'decile_out5']

# CHART WITH 5 GROUPS

df[["decile_out5", "churn"]].groupby("decile_out5").mean().plot.bar()
```

<AxesSubplot:xlabel='decile_out5'>



```
#oper dataset cohort 2
df2_path = r'C:\Users\ValentiniS\2023_05_Marketing AXA\Lab 1 Session2_BooksRUS_Churn\DataCohort2.xlsx'
df2 = pd.read_excel(df2_path)
df2.head()
```

	id	profits	churn_observed	multichannel	age	female	north	bigcity	mean_city	early_email	franchisee	street_agent	initialweb	initialstore	initialmobil
0	1	39.040001	0	0	38.720001	1	1	0	0	1	0	0	0	0	
1	2	53.080002	0	1	38.720001	0	1	0	0	1	0	0	1	0	
2	3	0.000000	0	0	62.000000	1	1	0	0	1	1	0	1	0	
3	4	56.049999	1	1	34.000000	0	1	0	0	1	1	0	1	0	
4	5	9.290000	0	0	38.720001	1	1	0	0	1	1	0	0	0	

NOW I OPEN A NEW DATASET BECAUSE I WANT TO PREDICT CHURN IN THIS NEW DATASET USING THE PARAMETER OF THE MODEL JUST ESTIMATED

```
# Step4:
### Predict churn on a new dataset, df2
### using the parameters from the logistic regression model estimated on dataset db_c1.
# Add a new column "churn_hat" to df2 containing the churn predictions.
df2['churn_hat'] = model.predict(df2)
df2.head()
```

	id	profits	churn_observed	multichannel	age	female	north	bigcity	mean_city	early_email	...	street_agent	initialweb	initialstore	initialmobile	initial
0	1	39.040001	0	0	38.720001	1	1	0	0	1	...	0	0	0	1	
1	2	53.080002	0	1	38.720001	0	1	0	0	1	...	0	1	0	0	
2	3	0.000000	0	0	62.000000	1	1	0	0	1	...	0	1	0	0	
3	4	56.049999	1	1	34.000000	0	1	0	0	1	...	0	1	0	0	
4	5	9.290000	0	0	38.720001	1	1	0	0	1	...	0	0	0	0	

5 rows × 21 columns

```
# Step4:
# create DECILE
#CHECK IF DECILES ARE CREATED CORRECTLY
df2['decile'] = pd.qcut(df2['churn_hat'], 10, labels=False)
df2['decile'] = 10 - df2['decile']
df2.head()
```

	id	profits	churn_observed	multichannel	age	female	north	bigcity	mean_city	early_email	...	initialweb	initialstore	initialmobile	initialstorepromo	i
0	1	39.040001	0	0	38.720001	1	1	0	0	1	...	0	0	1	0	
1	2	53.080002	0	1	38.720001	0	1	0	0	1	...	1	0	0	0	
2	3	0.000000	0	0	62.000000	1	1	0	0	1	...	1	0	0	0	
3	4	56.049999	1	1	34.000000	0	1	0	0	1	...	1	0	0	0	
4	5	9.290000	0	0	38.720001	1	1	0	0	1	...	0	0	0	0	

5 rows × 22 columns

```
# Step4:

# LIFT TABLE
df2[["decile", "churn_observed"]].groupby("decile").mean()
```

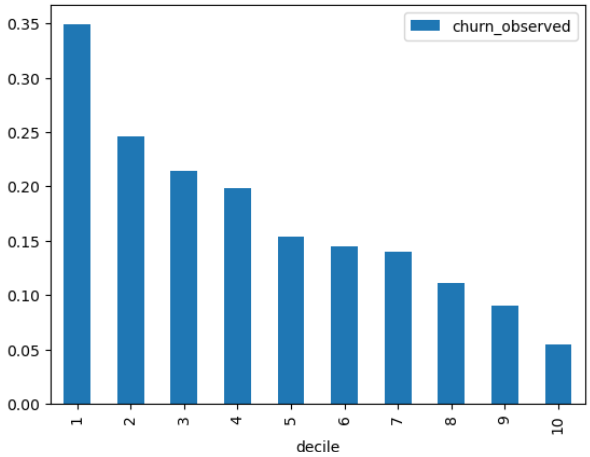
decile	churn_observed
1	0.349201
2	0.245854
3	0.214728
4	0.198307
5	0.154173
6	0.144524
7	0.140196
8	0.111003
9	0.090407
10	0.054054

```
# Step4:
df2[["decile", "churn_observed"]].groupby("decile").mean().plot.bar()
```

Step 6: Use the estimates to predict churn in a new dataset, DataCohort2.xlsx.

Whom would you include in the 'Targeted Proactive' action? Why?

<AxesSubplot:xlabel='decile'>



Evaluate the effectiveness of marketing actions

Our task is not only developing a marketing plan, but also be able to reassure the management that our marketing plan is going to be effective.

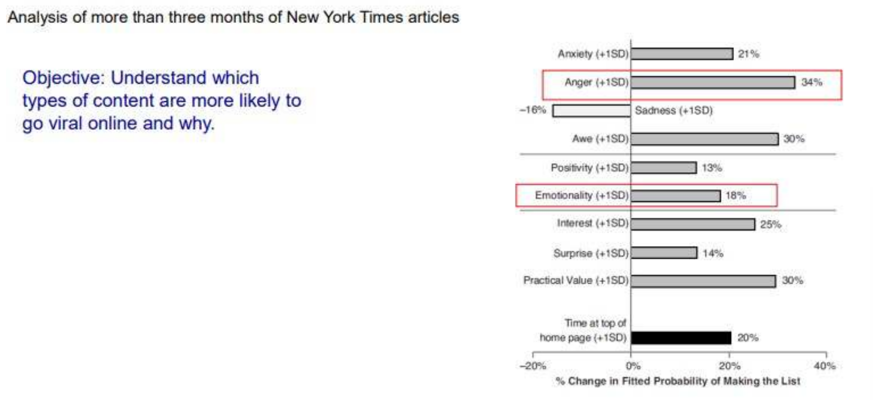
Overview: Example 1 New York Times

Why are some articles from The New York Times more shared than others and become viral? What are some potential factors that contribute to explaining their success?

Main argument: The emotional content of the article -> unboots the engagement, understanding which are the element which allows a content to become viral

1. Text analysis to quantify the article's emotional content
2. Text analysis to quantify the type of emotions evoked by the article (anger, anxiety, sadness, joy, humor etc.)
3. Experiment where they «manipulate» the emotional content of an ADV

They have a brand, and their managerial task is to develop a social media marketing campaign for thatbrand with the purpose to increase engagement -> they can create a content that is able to generate
emotions, because hopefully this kind content would be more likely to be shared.

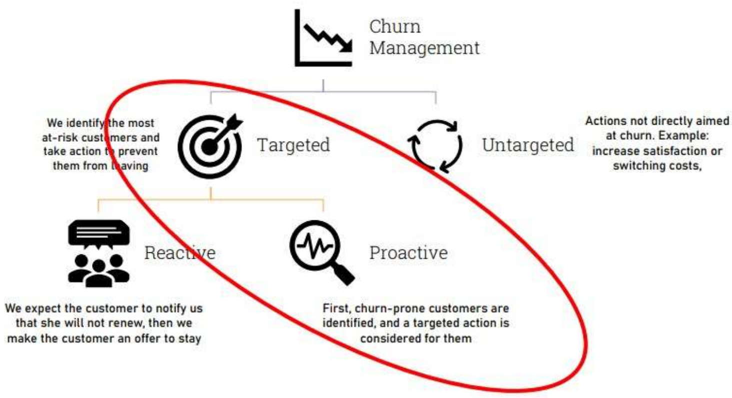


of

- *Experiment*-> to check if generating emotion could increase engagement; in principle we wantto link our brand with positive emotion, but it is not always the case
- Two versions of an ADV that differed on how much humor they evoked
- High- or low-amusement version of a story about a recent advertising campaign for Jimmy Dean sausages
- The two versions were pre-tested to verify that they are perceived differently in terms of humor levels
- Likelihood to share the ADV:
  - high-amusement (M = 3.97) versus low-amusement (M = 2.92; p <.005)
  -

Punch Line -> Predictive models and field experiments are commonly used also together to generateand test **the effectiveness of marketing actions and strategies**.

These methods can provide valuable insights for identifying possible marketing strategies andimproving overall effectiveness, before actually entering the market.
How can they be used in the context of Churn Management?



Lift chart is a perfect tool to predict if my customers are likely or not to leave.

Example

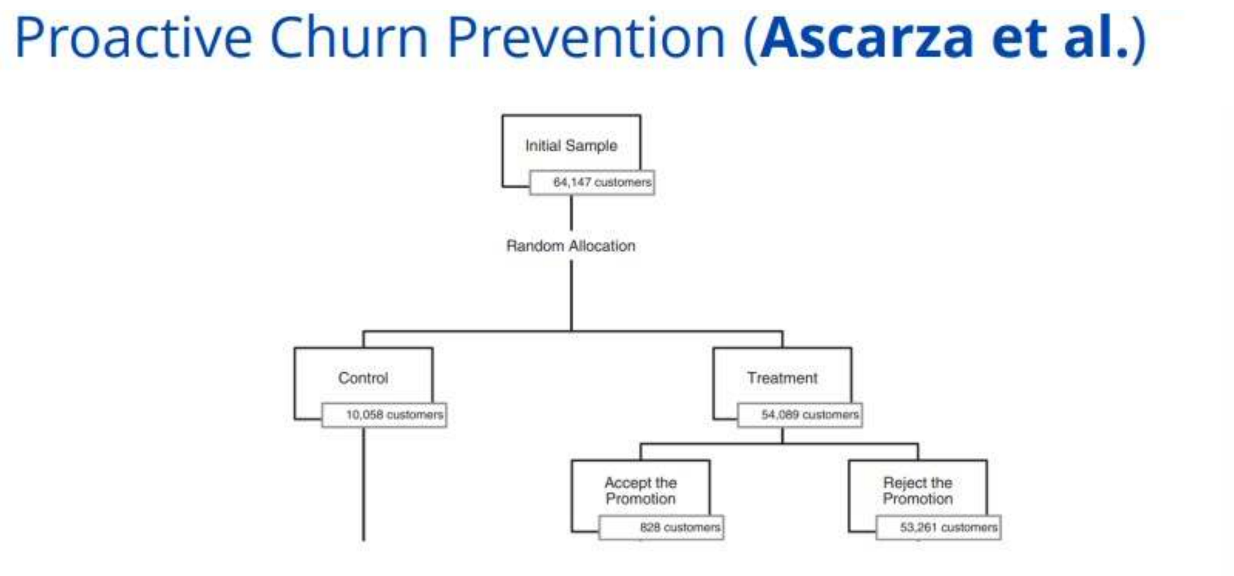
Ascarza, Iyengar and Schleicher (2016) "Proactive Churn Prevention Using Plan Recommendations: Evidence from a Field Experiment." Journal of Marketing Research

Proactive Churn Prevention (Ascarza et al.)

Field experiment (A/B test) with a telecom provider in South America
Intervention:

- Select ~65k customers active in Q1 2011
- 15% of customers left aside as a control group
- Treatment consists of calling the customers (high risk churn) and offering to upgrade to either $47 or $63 plan
    - No other commitments / No fine print
    - If accepted, extra $15 during the following 3 months

For all customers, including those in the control condition, one or both of the featured plans would have been better than their current plan. To incentivize customers to upgrade to the suggested plans,the company offered an additional credit of $15 for each of the following three months if they agreed to upgrade to one of the featured plans.



CASE 3 PILGRIM: TRAGETED PROACTIVE COMPAIGN

Context

Industry: Banking, Pilgrim
BankSample: 31,634

- For these customers, a churn model was estimated. Hence, each customer is associated with a churn probability, estimated at the time of acquisition through an internally consolidated model.
- Each customer is assigned a decile (1=customers with the highest churn risk, 10=lowest).
- The company has decided to focus a **targeted proactive strategy** on the customers in the **first 3 deciles**. They were the subject of a field test. Following one year after the field test, information on churn was collected for the entire sample of 31,634 customers.

Field Test

10 quantiles of churn_hat	Freq.	Percent	Cum.
1	3,163	10.00	10.00
2	3,163	10.00	20.00
3	3,164	10.00	30.00
4	3,163	10.00	40.00
5	3,164	10.00	50.00
6	3,163	10.00	60.00
7	3,163	10.00	70.00
8	3,164	10.00	80.00
9	3,163	10.00	90.00
10	3,164	10.00	100.00
Total	31,634	100.00	

Dataset: PilgrimTargetedField.xls

id	Customer identifier code
decile	Decile of belonging derived from the churn model
target_proactive	1 = included in the churn program (proactive), 0 = control (random selection)
retention	1=retained, 0=churn
MainlyOnline_bank_previous	1 indicates if the customer predominantly uses online banking, 0 if not
District	District of residence (USA)
Tenure	Number of years as a customer
AboveMedian_Tenure_Target	Variable indicating whether tenure is above the median of customers = 1, or below the median = 0
age	Age in classes, 0 indicates missing
agemiss	Variable that takes a value of 1 if age is missing, 0 otherwise
inc	Income, 0 indicates missing
incmiss	Variable that takes a value of 1 if income is missing, 0 otherwise
dist1100	Dummy variables identifying the 3 districts of residence for analysis
dist1200	Dummy variables identifying the 3 districts of residence for analysis
dist1300	Dummy variables identifying the 3 districts of residence for analysis
churn_hat	Prediction of churn from predictive model (made before observing actual churn)

Decile and churn hat columns were created **before** observing the actual retention.

Field test → Variable Targeted Proactive

Did not Receive Anything

Field Test	# of Customers
0=Control Group	4,394
1= Receive Incentive	5,096
Missing= Not included in the field	22,144
Total	31,634

Not subjected to testing

They receive a call and email message from customer care that proposes: If retained, Offer the customer a personalized financial consultation with a banking expert, a free travel insurance and benefit abroad and exclusive access to tailored events

Purpose

A field experiment designed to address the following key questions:

1. Is the target proactive strategy effective?
2. Yes, no, why? Carefully justify the response

3. CASE PILGRIM: FIELD TEST

Suggested Steps:

Step 0: Open the dataset PilgrimTargetedField.xls.

```
import pandas as pd
import statsmodels.formula.api as smf
import matplotlib.pyplot as plt
import numpy as np
import os
```

```python
#Step 0:Open dataset PilgrimTargetedField.xls
file_path = r'C:\Users\ValentiniS\PilgrimTargetedField.xls'
db_pilgrim = pd.read_excel(file_path)
db_pilgrim.head()
#note variable DECILE is already created
```

	id	decile	target_proactive	retention	MainlyOnline_bank_previous	District	Tenure	AboveMedian_Tenure_Target	age	agemiss	inc	incmiss	dist1100	dist120
0	8197	1	0.0	1	1	1200	9	0	0	1	0	1	0	
1	4270	1	1.0	0	0	1200	2	0	0	1	0	1	0	
2	4598	8	NaN	1	0	1200	4	0	2	0	6	0	0	
3	30687	1	0.0	0	0	1200	5	0	0	1	0	1	0	
4	29042	3	1.0	1	0	1200	9	0	1	0	1	0	0	

```python
db_pilgrim.shape
```

```
(31634, 16)
```

Step 1a: Check: the lift chart to verify if the predictive model worked using the already created decile variable and observed retention. Remember, deciles are created based on the predicted churn.

Step 1b: Check: verify the random assignment of the Targeted Proactive variable.

```python
#I have the observed retention, therefore  compute the observe Churn
db_pilgrim['Churn'] = 1-db_pilgrim['retention']
db_pilgrim.head()
```

	id	decile	target_proactive	retention	MainlyOnline_bank_previous	District	Tenure	AboveMedian_Tenure_Target	age	agemiss	inc	incmiss	dist1100	dist1200
0	8197	1	0.0	1	1	1200	9	0	0	1	0	1	0	1
1	4270	1	1.0	0	0	1200	2	0	0	1	0	1	0	1
2	4598	8	NaN	1	0	1200	4	0	2	0	6	0	0	1
3	30687	1	0.0	0	0	1200	5	0	0	1	0	1	0	1
4	29042	3	1.0	1	0	1200	9	0	1	0	1	0	0	1

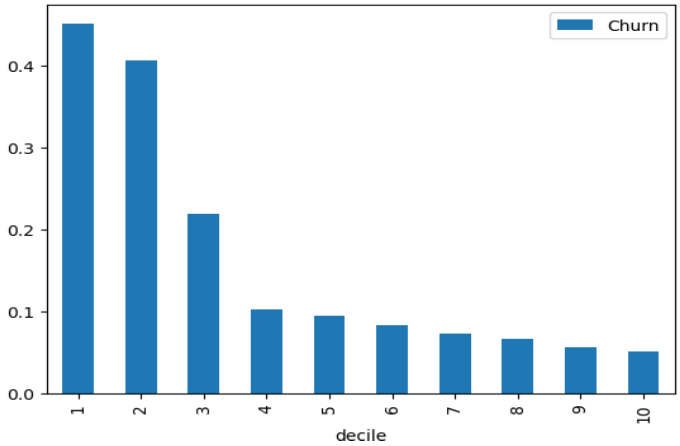
```python
#descriptive table associated to the lift chart: Average Observed Churn per decile
db_pilgrim [["decile", "Churn"]].groupby("decile").mean()
```

decile	Churn
1	0.451470
2	0.406892
3	0.219027
4	0.102118
5	0.095449
6	0.083465
7	0.073032
8	0.066056
9	0.056908
10	0.050885

Churn= 1-retention.

```
# Create the lift chart: Churn
db_pilgrim [["decile", "Churn"]].groupby("decile").mean().plot.bar()
```

```
<AxesSubplot:xlabel='decile'>
```



This model is super in terms of prediction validity!

```
# Step 1b: Check Random Assignment
mask = db_pilgrim[db_pilgrim['decile']<4]
formula = ('target_proactive ~ MainlyOnline_bank_previous + Tenure +'
           'age + agemiss + inc + incmiss + dist1100 + dist1200')

model = smf.logit(formula, data = mask).fit()

print(model.summary2())
```

```
Optimization terminated successfully.
         Current function value: 0.690220
         Iterations 4
                           Results: Logit
=====================================================================
Model:                 Logit               Pseudo R-squared:  0.000
Dependent Variable:    target_proactive    AIC:               13118.3732
Date:                  2023-10-15 22:07    BIC:               13182.7951
No. Observations:      9490                Log-Likelihood:    -6550.2
Df Model:              8                   LL-Null:           -6552.0
Df Residuals:          9481                LLR p-value:       0.89257
Converged:             1.0000              Scale:             1.0000
No. Iterations:        4.0000
---------------------------------------------------------------------
                           Coef.  Std.Err.    z    P>|z|  [0.025 0.975]
---------------------------------------------------------------------
Intercept                  0.0759   0.1606  0.4723 0.6367 -0.2389 0.3906
MainlyOnline_bank_previous 0.0305   0.0669  0.4557 0.6486 -0.1007 0.1617
Tenure                    -0.0031   0.0042 -0.7531 0.4514 -0.0113 0.0050
age                        0.0077   0.0381  0.2029 0.8392 -0.0670 0.0825
agemiss                    0.0127   0.1326  0.0958 0.9237 -0.2471 0.2725
inc                        0.0168   0.0236  0.7134 0.4756 -0.0294 0.0630
incmiss                    0.0254   0.1183  0.2147 0.8300 -0.2064 0.2572
dist1100                   0.0216   0.0829  0.2607 0.7943 -0.1409 0.1841
dist1200                   0.0684   0.0603  1.1332 0.2571 -0.0499 0.1866
=====================================================================
```

We don't need to create a random variable, because it was created by the management in this case. The two group are identified by the value of the target_proactive variable (0 means control group, 1 experimental group). If my purposes is to randomize.
We run a logit model with as dependent variable target_proactive and as independent variables all the customer's characteristic in order to assign a p-value to verify if there is a likelihood to be included in the control group or in the experimental group.

Example

Age: p-value (which is the amount of tolerated error in the estimate)= 0.8392 -> it is not significant andwith are happy about that because we don't want that age is associated in any possible way to the likelihood to belong to the control group or the experimental group -> if it is not significant that meansthat age does not explain the likelihood to be included in the control group or in the experimental group. If the p-value in this test is significant will be a problem -> if age increases, it is more likely to beadd in the experimental group which means a bias and we have not randomized)

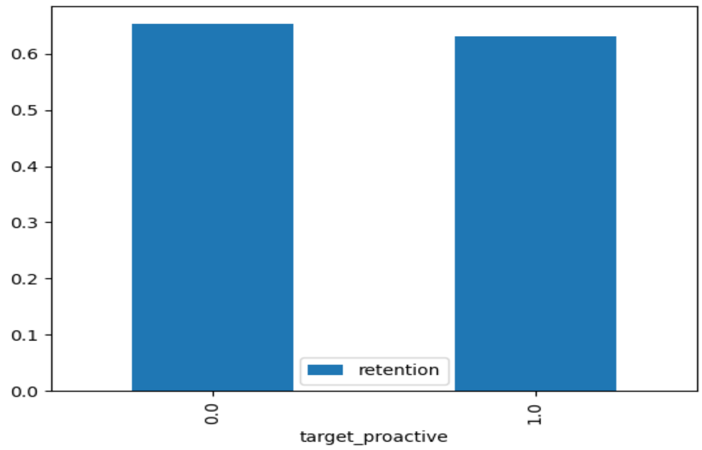
Step 2: Model Free Evidence: Verify if the targeted proactive campaign worked.

```
# Step 2: Results of the Field test: Model Free Evidence
db_pilgrim[['target_proactive', "retention"]].loc[db_pilgrim['decile']<4].groupby("target_proactive").mean()
```

target_proactive	retention
0.0	0.652025
1.0	0.631279

```
db_pilgrim[['target_proactive', "retention"]].loc[db_pilgrim['decile']<4].groupby("target_proactive").mean().plot.bar()
```

<AxesSubplot:xlabel='target_proactive'>



The two output are the same in terms of meaning, but we simply use two different tools.

We can say that the campaign is not effective because the results is pretty much the same in bothgroups, and above all in the experimental group is even lower than in the control group.

Step 3: Test: Test if the targeted proactive campaign worked and comment on the results.

The variable target_proactive reduces retention of about 10% -> OR= Exp(-0.10)=0.90→1-0.90=0.10

```
# Filter the DataFrame considering only the FIRST 3 DECILES
# not needed already done in previous step, but copy and paste here for
#illustrative purposes
mask = db_pilgrim[db_pilgrim['decile']<4]
```

```
# Define your formula
formula = ('retention ~ target_proactive + MainlyOnline_bank_previous +'
           'Tenure + age + agemiss + inc + incmiss + dist1100 + dist1200' )
# Estimate the model using the filtered DataFrame
model = smf.logit(formula, data=mask).fit()
# Print the model summary
print(model.summary2())
```

```
Optimization terminated successfully.
         Current function value: 0.604497
         Iterations 6
                          Results: Logit
==================================================================
Model:              Logit            Pseudo R-squared:  0.074
Dependent Variable: retention        AIC:               11493.3463
Date:               2023-10-15 22:11 BIC:               11564.9263
No. Observations:   9490             Log-Likelihood:    -5736.7
Df Model:           9                LL-Null:           -6196.1
Df Residuals:       9480             LLR p-value:       5.4128e-192
Converged:          1.0000           Scale:             1.0000
No. Iterations:     6.0000
------------------------------------------------------------------
                           Coef.  Std.Err.    z    P>|z|  [0.025  0.975]
------------------------------------------------------------------
Intercept                  1.6036   0.2536  6.3242 0.0000  1.1066  2.1006
target_proactive          -0.1008   0.0449 -2.2458 0.0247 -0.1887 -0.0128
MainlyOnline_bank_previous 0.1994   0.0748  2.6646 0.0077  0.0527  0.3461
Tenure                     0.0594   0.0046 13.0154 0.0000  0.0505  0.0684
age                        0.0683   0.0659  1.0369 0.2998 -0.0608  0.1975
agemiss                   -1.0950   0.2096 -5.2248 0.0000 -1.5058 -0.6842
inc                        0.0730   0.0393  1.8573 0.0633 -0.0040  0.1500
incmiss                   -0.7889   0.1752 -4.5034 0.0000 -1.1323 -0.4456
dist1100                   0.0866   0.0899  0.9638 0.3352 -0.0895  0.2627
dist1200                   0.1210   0.0650  1.8594 0.0630 -0.0065  0.2484
==================================================================
```

Step 4: Reflect on customer heterogeneity and try to explore.

Exploring customer heterogeneity means understand how different are the individuals within a sample ->whatever action we do will have a different impact on individuals because each of us is different from the other.
So this process means checking if there are groups of individuals that respond differently to our campaign.

```
[53]: # Step 4: Focus on Tenure [one can explore with other and more variables]
      # Results of the Field test distinct by Tenure
      # Descriptive Statistics
      db_pilgrim[["AboveMedian_Tenure_Target", "Tenure"]].groupby("AboveMedian_Tenure_Target").agg({"Tenure": ["mean", "min", 'max', 'std', "count"]})
```

[53]:

	Tenure				
	mean	min	max	std	count
AboveMedian_Tenure_Target					
0	5.31706	0	9	2.382365	15117
1	13.85639	9	19	2.613972	16517

```
[60]: #Step 4
      # Create sub-dataframe based on Tenure
      mask_tenure_low = db_pilgrim[db_pilgrim['AboveMedian_Tenure_Target']==0]
      mask_tenure_high = db_pilgrim[db_pilgrim['AboveMedian_Tenure_Target']==1]
```

```
[61]: #Logit only for those TENURE LOW
      # Define your formula
      formula = ('retention ~ target_proactive + MainlyOnline_bank_previous +'
                 'Tenure + age + agemiss + inc + incmiss + dist1100 + dist1200' )
      # Estimate the model using the filtered DataFrame
      model = smf.logit(formula, data=mask_tenure_low).fit()
      # Print the model summary
      print(model.summary2())
```

```
Optimization terminated successfully.
         Current function value: 0.617853
         Iterations 7
                              Results: Logit
=========================================================================
Model:                   Logit              Pseudo R-squared:   0.081
Dependent Variable:      retention          AIC:                5883.4237
Date:                    2023-10-15 22:25   BIC:                5948.0721
No. Observations:        4745               Log-Likelihood:     -2931.7
Df Model:                9                  LL-Null:            -3190.6
Df Residuals:            4735               LLR p-value:        9.1360e-106
Converged:               1.0000             Scale:              1.0000
No. Iterations:          7.0000
-------------------------------------------------------------------------
                            Coef.  Std.Err.    z     P>|z|   [0.025  0.975]
-------------------------------------------------------------------------
Intercept                   1.2099   0.3563  3.3960 0.0007  0.5116  1.9081
target_proactive            0.1243   0.0625  1.9888 0.0467  0.0018  0.2468
MainlyOnline_bank_previous  0.0543   0.1014  0.5356 0.5923 -0.1444  0.2529
Tenure                      0.1115   0.0125  8.9393 0.0000  0.0870  0.1359
age                         0.1606   0.1015  1.5826 0.1135 -0.0383  0.3596
agemiss                    -1.0793   0.2966 -3.6388 0.0003 -1.6606 -0.4979
inc                         0.0753   0.0558  1.3479 0.1777 -0.0342  0.1847
incmiss                    -0.7588   0.2481 -3.0590 0.0022 -1.2450 -0.2726
dist1100                    0.1058   0.1264  0.8371 0.4025 -0.1419  0.3536
dist1200                    0.1710   0.0909  1.8815 0.0599 -0.0071  0.3492
=========================================================================
```

Tenure could explain a different response in terms of incentive because it divides the customer between short-term and long-term customers. For sure when talking about long-term customer we can consider inertia, which is something that we cannot consider when analysing short-term customers behavior. Moreover, some marketing actions cannot be proposed to short-term customers because it can be dangerous and vice versa.

To do this point we can act in three ways:

1. Descriptive statistic
2. Use interaction term
3. Divide the dataset in two sub-datasets: in this case low tenure (0<x<9) and high tenure (x>=9). And then create a regression in order to understand the impact of this variable on retention.

TENURE LOW: Positive impact of target_proactive of 13% on retention OR=exp(0.1209)=1.13
Before, when we analyzed the entire sample the effect was negative (reduction of 10%).

```
#Logit only for those TENURE HIGH
# Define your formula
formula = ('retention ~ target_proactive + MainlyOnline_bank_previous +'
           'Tenure + age + agemiss + inc + incmiss + dist1100 + dist1200' )
# Estimate the model using the filtered DataFrame
model = smf.logit(formula, data=mask_tenure_high).fit()
# Print the model summary
print(model.summary2())
```

```
Optimization terminated successfully.
         Current function value: 0.582142
         Iterations 6
                          Results: Logit
=====================================================================
Model:              Logit            Pseudo R-squared: 0.071
Dependent Variable: retention        AIC:              5544.5229
Date:               2023-10-15 22:26 BIC:              5609.1714
No. Observations:   4745             Log-Likelihood:   -2762.3
Df Model:           9                LL-Null:          -2973.4
Df Residuals:       4735             LLR p-value:      2.3145e-85
Converged:          1.0000           Scale:            1.0000
No. Iterations:     6.0000
---------------------------------------------------------------------
                            Coef.  Std.Err.    z    P>|z|  [0.025  0.975]
---------------------------------------------------------------------
Intercept                   0.9013   0.4030  2.2364 0.0253  0.1114  1.6912
target_proactive           -0.3431   0.0653 -5.2508 0.0000 -0.4712 -0.2150
MainlyOnline_bank_previous  0.3565   0.1136  3.1381 0.0017  0.1338  0.5791
Tenure                      0.1138   0.0130  8.7456 0.0000  0.0883  0.1393
age                         0.0127   0.0886  0.1436 0.8858 -0.1609  0.1863
agemiss                    -1.0005   0.3035 -3.2966 0.0010 -1.5953 -0.4056
inc                         0.0763   0.0552  1.3825 0.1668 -0.0319  0.1845
incmiss                    -0.8254   0.2480 -3.3284 0.0009 -1.3115 -0.3394
dist1100                    0.0542   0.1294  0.4185 0.6756 -0.1995  0.3078
dist1200                    0.0680   0.0944  0.7206 0.4712 -0.1170  0.2531
=====================================================================
```

TENURE HIGH: Negative Impact of target_proactive on Retention -29% Why? -> «Broken Inertia»

```
#how to ask odds ratio as output

# Access the model parameters (coefficients)
model_params = model.params

# Compute the odds ratios (exponentiate the coefficients)
odds_ratios = np.exp(model_params)

# Create a DataFrame to display the odds ratios with their respective names
odds_ratios_df = pd.DataFrame({'Odds Ratio': odds_ratios, 'Coefficient': model_params})

# Print the DataFrame with odds ratios
print(odds_ratios_df)
```

```
                            Odds Ratio  Coefficient
Intercept                     3.352995     1.209854
target_proactive              1.132372     0.124314
MainlyOnline_bank_previous    1.055781     0.054281
Tenure                        1.117935     0.111484
age                           1.174261     0.160639
agemiss                       0.339847    -1.079259
inc                           1.078157     0.075253
incmiss                       0.468218    -0.758821
dist1100                      1.111616     0.105815
dist1200                      1.186546     0.171046
```

Step 5: Comment on the results of your explorations.

WRAP UP:

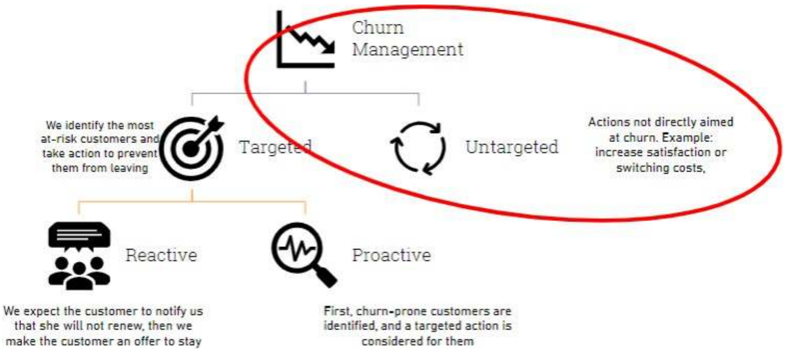
- A targeted proactive action aimed at reducing churn actually has a negative effect on retention, reducing it overall by about 10%.
- By examining different response probabilities for various customer groups, it is noted that tenure moderates the effect; those with low tenure respond positively, while long-term customers respond negatively.
- The inertia effect may explain the result, in line with the literature by Ascarza, Iyengar, and Schleicher (2016) titled 'The Wrong Way to Reduce Churn' in Idea Watch, Harvard Business Review.

Take Home:

- The probability of churn and the probability of response to a specific marketing actionshould be considered in conjunction!

- *Randomized field tests if possible and/or the development of marketing response models combined with churn probabilities*

UNTARGETED CHURN REDUCTION STRATEGY – CASE 4: BOOKS R US -> FIELD TEST

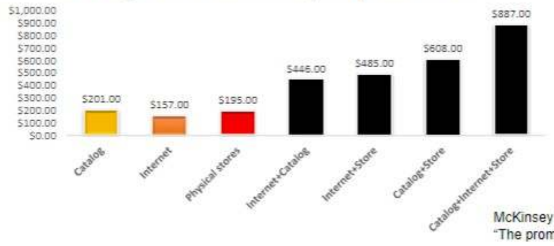


In this case we just want to reduce the churn rate by applying actions that we do believe that could have a positive impact on the likelihood to churn.

Channels, touchpoints and profits...

Do multichannel customers buy more?

Average Annual Dollars Spent per Customer



McKinsey Quarterly 2011
"The promise of multichannel retailing"

Replicated by: Loftus, Mulliken and Sharp 2008; Myers, Pickersgill, and van Metre 2004; Thomas and Sullivan 2005; Kumar and Venkatesan 2005; Venkatesan, Kumar, and Ravishanker 2007; Ansari, Mela, and Neslin 2008; Boehm 2008; Campbell and Frei 2010; Xue, Hitt, and Chen 2011; Gensler, Leeflang, & Skiera 2012; Kushwaha and Shankar 2013, Montaguti, Neslin, Valentini 2016.

MANAGERIAL IMPLICATION: firms should use marketing to create more multichannel customersLogic,

- Marketing induces more customers to become multichannel.
- Multichannel customers are more profitable than they would have been had they not been multichannel.
- As a result, average profit per customer should increase.

Is this **ACTIONABLE**?

Case 4 BooksRUs: Summary

Can marketing actions create more multichannel customers?
Will this increase the profitability of the customer base?
If yes, how?
—By reducing churn?
—By increasing spending volume?
—By increasing purchase frequency?

Case 4 BooksRUs:
Can Marketing Induce Multichannel Buying and More Profitable Customers?

"Untargeted strategy"

— BookRus is a major multichannel European book retailer. The company sells books through stores, mail-order, phone and the Internet.
— The firm mails a print catalog to the customer base five times per year
— The company opted for an omnichannel strategy providing members several possibilities to purchase: the physical stores, the website, mail/order or phone, the app

— Despite the efforts of providing more channels to purchase the core customer base was still strongly related to the physical store without exploring other available channels (70% of customers purchased mainly using only the Physical Store)

— Additionally. the management wasn't sure about how much effective an omnichannel strategy could be in increasing average customer profitability

In doing this, **Roberta is attempting to implement an "UNTARGETED" strategy** aimed at reducing churn and increasing customer base profitability. The approach is indirect; therefore, the campaignand field testneedtobecarefullydesigned.

CASE OBJECTIVES

BooksRUsimplementafieldexperimenttoaddressfourquestions:

1. Canamarketingcampaignbedesignedtocreatemoremultichannelcustomers?
2. Ifso,aremultichannelcustomersmoreprofitablethantheywouldhavebeenhadtheynotbeen multichannel?
3. Whatistheimpactonchurn?
4. What types of marketing campaigns work best, and why?

DATA

Thecompanyselectedon3cohortsofcustomerswholivedwithinatleastonestore's service areaand were acquired in the last period of the year (September – December): Cohort 1, Cohort 2 and Cohort 3. Allcustomersincludedinthesecohortwereobservedsincetheveryfirstpurchase:

Cohorts 1 and 2wereusedtotestifmultichannelcustomersareassociatedwithhigherprofitsandless churn. Cohort3wasselectedtoconductthefieldtest

Before running the Field Test Roberta wanted to have a first correlational empirical evidence that the use of multiple channels was associated with more profits. She used data of Cohort 1 to conduct thisanalysis.

Phase 0 - Are Multichannel Customers More Profitable Customers?

Phase 0 - Are Multichannel Customers More Profitable Customers?

Case 2 Results

Results: Ordinary least squares

Model:	OLS	Adj. R-squared:	0.336
Dependent Variable:	profits	AIC:	287829.0750
Date:	2022-05-16 07:26	BIC:	287973.1367
No. Observations:	35391	Log-Likelihood:	-1.4390e+05
Df Model:	16	F-statistic:	1122.
Df Residuals:	35374	Prob (F-statistic):	0.00
R-squared:	0.337	Scale:	199.23

	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
multichannel	20.9517	0.3207	65.3373	0.0000	20.3231	21.5802
age	0.1509	0.0048	31.4807	0.0000	0.1415	0.1603
female	1.2085	0.1623	7.4482	0.0000	0.8905	1.5265
street_agent	3.1672	0.1585	19.9766	0.0000	2.8565	3.4780
north	1.9387	0.2022	9.5889	0.0000	1.5424	2.3349
early_email	1.8377	0.1644	11.1769	0.0000	1.5154	2.1599
bigcity	0.1535	0.2579	0.5952	0.5517	-0.3519	0.6589
mean_city	0.0237	0.2711	0.0873	0.9304	-0.5076	0.5550
franchisee	0.4522	0.1596	2.8340	0.0046	0.1394	0.7649
initialstorepromo	-1.8296	0.3417	-5.3540	0.0000	-2.4994	-1.1598
initialweb	11.0348	0.7429	14.8539	0.0000	9.5788	12.4909
initialstore	-2.5538	0.3014	-8.4724	0.0000	-3.1446	-1.9630
initialmobile	10.0674	0.5492	18.3326	0.0000	8.9910	11.1438
initialrevenues	0.5308	0.0086	61.5983	0.0000	0.5139	0.5476
initialreturns	-0.5039	0.0541	-9.3126	0.0000	-0.6099	-0.3978
initialreturns2	0.0021	0.0008	2.7236	0.0065	0.0006	0.0036
const	4.6034	0.3116	14.7718	0.0000	3.9926	5.2142

Phase 0 - Are Multichannel Customers More Profitable Customers?

Case 2 Results

Results: Ordinary least squares

Model:	OLS	Adj. R-squared:	0.336
Dependent Variable:	profits	AIC:	287829.0750
Date:	2022-05-16 07:26	BIC:	287973.1367
No. Observations:	35391	Log-Likelihood:	-1.4390e+05
Df Model:	16	F-statistic:	1122.
Df Residuals:	35374	Prob (F-statistic):	0.00
R-squared:	0.337	Scale:	199.23

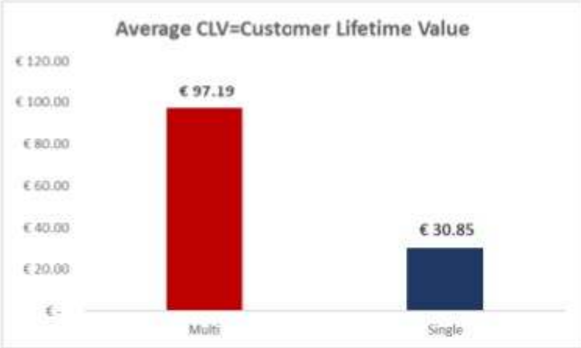
	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
multichannel	20.9517	0.3207	65.3373	0.0000	20.3231	21.5802
age	0.1509	0.0048	31.4807	0.0000	0.1415	0.1603
female	1.2085	0.1623	7.4482	0.0000	0.8905	1.5265
street_agent	3.1672	0.1585	19.9766	0.0000	2.8565	3.4780
north	1.9387	0.2022	9.5889	0.0000	1.5424	2.3349
early_email	1.8377	0.1644	11.1769	0.0000	1.5154	2.1599
bigcity	0.1535	0.2579	0.5952	0.5517	-0.3519	0.6589
mean_city	0.0237	0.2711	0.0873	0.9304	-0.5076	0.5550
franchisee	0.4522	0.1596	2.8340	0.0046	0.1394	0.7649
initialstorepromo	-1.8296	0.3417	-5.3540	0.0000	-2.4994	-1.1598
initialweb	11.0348	0.7429	14.8539	0.0000	9.5788	12.4909
initialstore	-2.5538	0.3014	-8.4724	0.0000	-3.1446	-1.9630
initialmobile	10.0674	0.5492	18.3326	0.0000	8.9910	11.1438
initialrevenues	0.5308	0.0086	61.5983	0.0000	0.5139	0.5476
initialreturns	-0.5039	0.0541	-9.3126	0.0000	-0.6099	-0.3978
initialreturns2	0.0021	0.0008	2.7236	0.0065	0.0006	0.0036
const	4.6034	0.3116	14.7718	0.0000	3.9926	5.2142

PHASE 0 - ARE MULTICHANNEL CUSTOMERS MORE PROFITABLE CUSTOMERS? CLV PERSPECTIVE

She also computed the CLV using this (Cohort 1) :

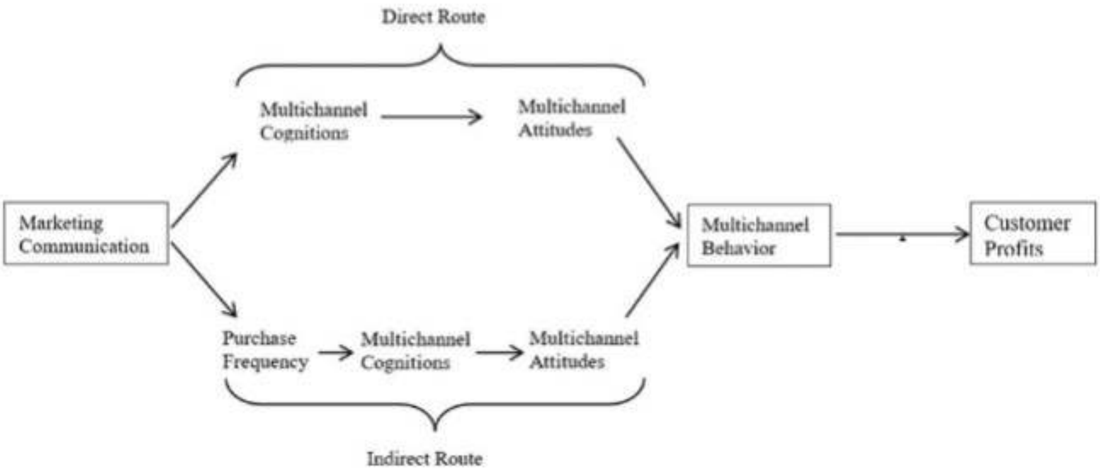
$$CLV = m\frac{(1+d)}{(1+d-r)} - AC$$

Discount rate=0.12
r=retention rate=1-churn rate
d=discount rate
AC=acquisition cost= 10 €



Design of the field test

This figure depicts how marketing communication can induce customers to become multichannel shoppers who in turn are more profitable.



ACTIONABILITY

FOUR CAMPAIGNS: IS MULTICHANNEL CUSTOMER STRATEGY ACTIONABLE?

«Multichannel Message» promoting multichannel shopping

«Value Proposition Message» promoting the company

IMPLEMENTATION OF THE FIELD TEST

Group	Number of customers
MNF	6810
MF	6831
VPNF	6829
VPF	6821
Control	3419
	30710

Caso BooksRUs: Field Test

Step 0: Open the dataset Cohort3.xls.

```
import pandas as pd
import statsmodels.api as sm
import statsmodels.formula.api as smf
import numpy as np
import researchpy as rp
import os
from statsmodels.stats.proportion import proportions_ztest
```

```
#Step 0:
#Open dataset  Cohort3BooksFieldTest.xls
file_path = r'C:\Users\ValentiniS\Cohort3BooksFieldTest.xls'
df = pd.read_excel(file_path)
df.head()
```

	id	groups	profits	churn_observed	multichannel	mf	mnf	vpf	vpnf	c	...	mean_city	early_email	franchisee	street_agent	initialweb	initialstore	initialmol
0	1	Control	39.040001	0	0	0	0	0	0	1	...	0	1	0	0	0	0	
1	2	mnf	53.080002	0	1	0	1	0	0	0	...	0	1	0	0	1	0	
2	3	vpnf	0.000000	0	0	0	0	0	1	0	...	0	1	1	0	1	0	
3	4	mnf	56.049999	1	1	0	1	0	0	0	...	0	1	1	0	1	0	
4	5	Control	9.290000	0	0	0	0	0	0	1	...	0	1	1	0	0	0	

5 rows × 24 columns

```
df.shape
```

(30710, 24)

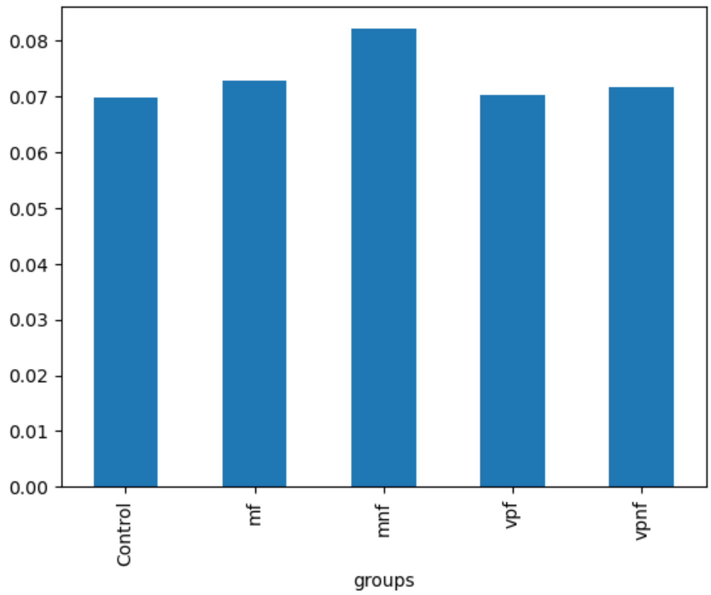
Step 1: Model Free Evidence: Check if and which of the campaigns

```
# Step 1: Model Free Evidence: Check if and which of the campaigns generated the most multichannel customers
df.groupby("groups")['multichannel'].mean()
```

```
groups
Control    0.069903
mf         0.072757
mnf        0.082085
vpf        0.070224
vpnf       0.071753
Name: multichannel, dtype: float64
```

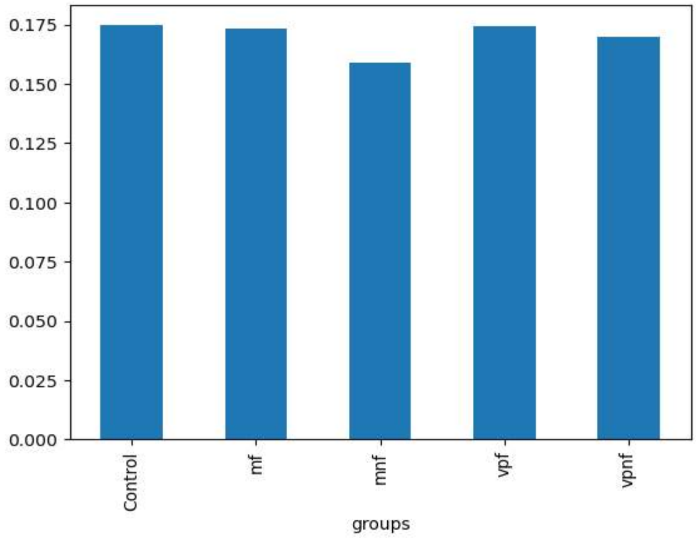
```
# Step 1: Model Free Evidence: chart
df.groupby("groups")['multichannel'].mean().plot.bar()
```

<AxesSubplot:xlabel='groups'>



```
#impact on churn: model-free evidence
df.groupby("groups")['churn_observed'].mean().plot.bar()
```

```
<AxesSubplot:xlabel='groups'>
```

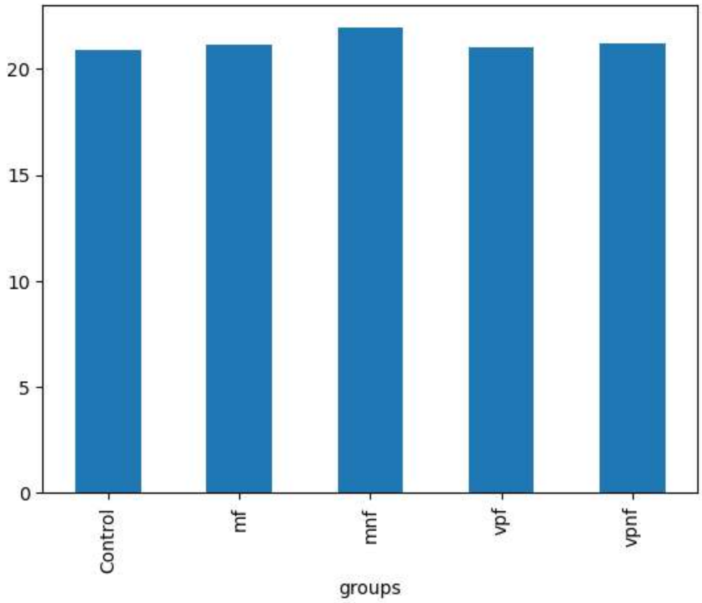



```
#impact on churn
df.groupby("groups")['churn_observed'].mean()
```

```
groups
Control    0.174612
mf         0.173327
mnf        0.158737
vpf        0.174168
vpnf       0.170010
Name: churn_observed, dtype: float64
```

```
#impact on profits
df.groupby("groups")['profits'].mean().plot.bar()
```

```
<AxesSubplot:xlabel='groups'>
```




```
#impact on profits
df.groupby("groups")['churn_observed'].mean()
```

```
groups
Control    0.174612
mf         0.173327
mnf        0.158737
vpf        0.174168
vpnf       0.170010
Name: churn_observed, dtype: float64
```

Step 2: Statistical Test: Test if and which of the campaigns generated the most multichannel customers.

```
# Step 2: Results of the Field test: TEST
formula = ('multichannel ~ mf + mnf + vpf + vpnf')

model = smf.logit(formula, data = df).fit()

print(model.summary2())
```

```
Optimization terminated successfully.
         Current function value: 0.263012
         Iterations 7
                         Results: Logit
=================================================================
Model:              Logit            Pseudo R-squared: 0.001
Dependent Variable: multichannel     AIC:              16164.1787
Date:               2023-10-16 00:35 BIC:              16205.8404
No. Observations:   30710            Log-Likelihood:   -8077.1
Df Model:           4                LL-Null:          -8081.7
Df Residuals:       30705            LLR p-value:      0.056010
Converged:          1.0000           Scale:            1.0000
No. Iterations:     7.0000
------------------------------------------------------------------
              Coef.   Std.Err.     z      P>|z|   [0.025   0.975]
------------------------------------------------------------------
Intercept   -2.5882    0.0671  -38.5884  0.0000  -2.7196  -2.4567
mf           0.0431    0.0817    0.5275  0.5979  -0.1170   0.2031
mnf          0.1738    0.0803    2.1648  0.0304   0.0164   0.3312
vpf          0.0049    0.0821    0.0600  0.9522  -0.1560   0.1659
vpnf         0.0281    0.0818    0.3434  0.7313  -0.1323   0.1885
=================================================================
```

Remember to exclude the control group.

```
# Step 2: Results of the Field test: test
###### CHURN
formula = ('churn_observed ~ mf + mnf + vpf + vpnf')

model = smf.logit(formula, data = df).fit()

print(model.summary2())
```

```
Optimization terminated successfully.
         Current function value: 0.455253
         Iterations 6
                         Results: Logit
==============================================================
Model:              Logit            Pseudo R-squared: 0.000
Dependent Variable: churn_observed   AIC:              27971.6567
Date:               2023-10-16 00:36 BIC:              28013.3185
No. Observations:   30710            Log-Likelihood:   -13981.
Df Model:           4                LL-Null:          -13985.
Df Residuals:       30705            LLR p-value:      0.088281
Converged:          1.0000           Scale:            1.0000
No. Iterations:     6.0000
---------------------------------------------------------------
              Coef.   Std.Err.     z      P>|z|   [0.025   0.975]
---------------------------------------------------------------
Intercept   -1.5533    0.0450  -34.4800  0.0000  -1.6416  -1.4650
mf          -0.0089    0.0552   -0.1619  0.8714  -0.1172   0.0993
mnf         -0.1144    0.0559   -2.0446  0.0409  -0.2240  -0.0047
vpf         -0.0031    0.0552   -0.0559  0.9554  -0.1113   0.1051
vpnf        -0.0323    0.0554   -0.5827  0.5601  -0.1408   0.0763
==============================================================
```

The effect should be indirect -> we want to observe if the newly created multichannel customer are now able to reduce churn and improve profits.

```
# Step 2: Results of the Field test: Test
###### profits

formula = ('profits ~ mf + mnf + vpf + vpnf')

model = smf.ols(formula, data = df).fit()

print(model.summary2())
```

```
                 Results: Ordinary least squares
=============================================================
Model:              OLS              Adj. R-squared:     0.000
Dependent Variable: profits          AIC:                273197.3616
Date:               2023-10-16 00:36 BIC:                273239.0233
No. Observations:   30710            Log-Likelihood:     -1.3659e+05
Df Model:           4                F-statistic:        2.319
Df Residuals:       30705            Prob (F-statistic): 0.0546
R-squared:          0.000            Scale:              427.52
--------------------------------------------------------------
             Coef.   Std.Err.     t      P>|t|   [0.025   0.975]
--------------------------------------------------------------
Intercept   20.8947   0.3536   59.0891  0.0000  20.2016  21.5878
mf           0.2598   0.4332    0.5998  0.5486  -0.5892   1.1088
mnf          1.0367   0.4334    2.3921  0.0168   0.1872   1.8861
vpf          0.1476   0.4333    0.3407  0.7333  -0.7016   0.9968
vpnf         0.2800   0.4332    0.6464  0.5180  -0.5690   1.1291
--------------------------------------------------------------
Omnibus:              5182.527      Durbin-Watson:         1.840
Prob(Omnibus):        0.000         Jarque-Bera (JB):      10331.001
Skew:                 1.029         Prob(JB):              0.000
Kurtosis:             4.959         Condition No.:         8
=============================================================
```

Step 3: Test if and which of the 4 campaigns had an effect on churn, and reflect on which variables to include in the test.

```
#check if multichannl impacts Churn

formula = ('churn_observed ~ multichannel')

model = smf.logit(formula, data = df).fit()

print(model.summary2())
```

```
Optimization terminated successfully.
         Current function value: 0.450560
         Iterations 7
                         Results: Logit
=================================================================
Model:              Logit            Pseudo R-squared: 0.011
Dependent Variable: churn_observed   AIC:              27677.4182
Date:               2023-10-16 00:38 BIC:              27694.0829
No. Observations:   30710            Log-Likelihood:   -13837.
Df Model:           1                LL-Null:          -13985.
Df Residuals:       30708            LLR p-value:      2.0765e-66
Converged:          1.0000           Scale:            1.0000
No. Iterations:     7.0000
-----------------------------------------------------------------
                 Coef.  Std.Err.    z     P>|z|   [0.025  0.975]
-----------------------------------------------------------------
Intercept       -1.5238   0.0155 -98.5023 0.0000 -1.5541 -1.4935
multichannel    -1.3417   0.0944 -14.2192 0.0000 -1.5266 -1.1567
=================================================================
```

Being multichannel is associated with less churn and higher profits.

Step 4: Test if and which of the 4 campaigns had an effect on profits, and reflect on which variables to include in the test.

```
#check if multichannl impacts profits

formula = ('profits ~ multichannel')

model = smf.ols(formula, data = df).fit()

print(model.summary2())
```

```
                 Results: Ordinary least squares
==================================================================
Model:              OLS              Adj. R-squared:     0.169
Dependent Variable: profits          AIC:                267496.2869
Date:               2023-10-16 00:38 BIC:                267512.9516
No. Observations:   30710            Log-Likelihood:     -1.3375e+05
Df Model:           1                F-statistic:        6268.
Df Residuals:       30708            Prob (F-statistic): 0.00
R-squared:          0.170            Scale:              355.12
-------------------------------------------------------------------
                Coef.   Std.Err.    t      P>|t|    [0.025   0.975]
-------------------------------------------------------------------
Intercept      18.8756    0.1117 168.9365 0.0000 18.6566 19.0946
multichannel   32.5797    0.4115  79.1714 0.0000 31.7731 33.3863
------------------------------------------------------------------
Omnibus:             4764.087      Durbin-Watson:         1.840
Prob(Omnibus):       0.000         Jarque-Bera (JB):      8454.467
Skew:                1.006         Prob(JB):              0.000
Kurtosis:            4.599         Condition No.:         4
==================================================================
```

*** please note you can control also for other variables but result would not change dramatically, since it is a result of a randomized field test

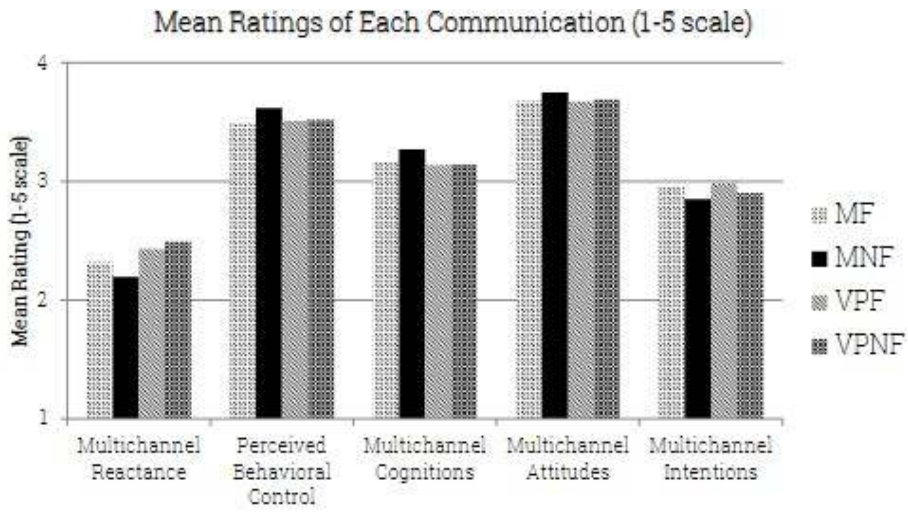
```
#check if multichannl impacts Churn and profits

formula = ('churn_observed ~ multichannel + age + '
           'north + female +'
           'bigcity + mean_city + early_email +'
           'franchisee + street_agent +'
           'initialweb + initialstore + initialmobile + initialstorepromo +'
           'initialreturns + initialrevenues' )

model = smf.logit(formula, data = df).fit()

print(model.summary2())
```

```
Optimization terminated successfully.
         Current function value: 0.430332
         Iterations 7
                         Results: Logit
==================================================================
Model:              Logit            Pseudo R-squared: 0.055
Dependent Variable: churn_observed   AIC:              26462.9731
Date:               2023-10-16 00:44 BIC:              26596.2906
No. Observations:   30710            Log-Likelihood:   -13215.
Df Model:           15               LL-Null:          -13985.
Df Residuals:       30694            LLR p-value:      0.0000
Converged:          1.0000           Scale:            1.0000
No. Iterations:     7.0000
------------------------------------------------------------------
                   Coef.  Std.Err.    z     P>|z|   [0.025  0.975]
------------------------------------------------------------------
Intercept         -0.6497   0.0683  -9.5120 0.0000 -0.7836 -0.5159
multichannel      -1.3793   0.0962 -14.3423 0.0000 -1.5678 -1.1908
age               -0.0163   0.0011 -14.9638 0.0000 -0.0184 -0.0142
north              0.2890   0.0413   6.9938 0.0000  0.2080  0.3700
female            -0.5703   0.0327 -17.4175 0.0000 -0.6345 -0.5061
bigcity            1.0623   0.0509  20.8806 0.0000  0.9626  1.1620
mean_city          0.2255   0.0734   3.0724 0.0021  0.0817  0.3694
early_email       -0.0667   0.0339  -1.9676 0.0491 -0.1332 -0.0003
franchisee        -0.0674   0.0333  -2.0232 0.0431 -0.1327 -0.0021
street_agent      -0.4389   0.0332 -13.2206 0.0000 -0.5040 -0.3739
initialweb        -0.1024   0.1239  -0.8260 0.4088 -0.3453  0.1406
initialstore      -1.2206   0.6032  -2.0234 0.0430 -2.4030 -0.0383
initialmobile     -0.3960   0.1319  -3.0022 0.0027 -0.6545 -0.1375
initialstorepromo  1.0329   0.6041   1.7100 0.0873 -0.1510  2.2169
initialreturns    -0.0080   0.0057  -1.3911 0.1642 -0.0192  0.0033
initialrevenues    0.0099   0.0011   9.2757 0.0000  0.0078  0.0120
==================================================================
```

Step 5: Which campaign worked better? Why?

MNF (reactance theory) is the campaign which worked better; the value proposition options don't tell customers the presence of multichannel. Discount means that there is something for you, but in this case, MF force you to purchase, but also on which channels do the purchase -> it is really manipulating and it have a negative effect on customers.

Theory of planned behavior -> perceived behavioral control is associated to an increase in positive attitude toward an action. In this context this concept can be reassumed in the fact that just the
knowledge of the presence of multiple channels is helpful for our purpose.



- MNF produce less "reactance" and increased "perceived control"!

Multichannel customers are more profitable:
further analyses

Industry: Books

	Average annual Profit per customer, distinct by frequency of purchase								
Number of Purchase occasions	2	3	4	5	6	7	8	9	>9
Single Channel	€20.28	€29.18	€36.27	€37.68	€34.14	€36.26	€38.03	€43.20	€50.11
Multichannel	€23.36	€37.05	€45.14	€52.92	€59.96	€60.37	€57.34	€67.97	€63.34
Difference	€3.08	€7.87	€8.87	€15.23	€25.83	€24.01	€19.31	€24.77	€13.23
p-value	0.006	0.000	0.000	0.000	0.000	0.000	0.003	0.105	0.173

margins...

Margins	Multichannel Purchases	Non-Multichannel Purchases
Store	31%	79%
Internet, Mail, Mobile	69%	21%

However, TT > 0 even for combinations of high margin channels

Multichannel	Single Channel	TT=Difference in Profits
Internet/Mobile	Internet	€8.52
Internet/Mobile	Mobile	€17.54
Internet/Mobile	Mail Order	€ 9.64
Internet/Mail Order	Internet	€12.31
Internet/Mail Order	Mail	€13.64
Internet/Mail Order	Mobile	€23.34
Phone/Mail Order	Mobile	€18.00
Phone/Mail Order	Mail	€8.00
Phone/Mail Order	Internet	€7.23

Post-test survey conducted on the company's customers: 2,068 respondents -> Lab Experiment: each respondent was exposed to only one of the four communications using an experimental logic

FOR DOUBTS OR SUGGESTIONS ON THE HANDOUTS

CHIARA TUA
chiara.tua@studbocconi.it
@chiara_tua
+39 3479789059

VITTORIA NASONTE
vittoria.nasonte@studbocconi.it
@_vittorian_
+39 3274441476

FOR INFO ON THE TEACHING DIVISION

MARCO FORMISANO
marco.formisano@studbocconi.it
@marco_formisano__
+39 3313433934

ELENA CACIOLI
elena.cacioli@studbocconi.it
@elenacaciolii_
+39 3928931605

TEACHING DIVISION

OUR PARTNERS

700+ CLUB

ETHAN SUSTAINABILITY

DELIVERY VALLEY
NO GENDER KITCHEN

LA PIADINERIA