SEPTEMBER 2023

Master Degree - Data Science
Prof. Nico Di Domenica
Prof. Giovanni Collini

# Marketing Analytics


Julia Lan Bui Xuan 882385
Michele Salvaterra 891109

02

# Marketing objectives

## Customer Focus

Prevent high-value customer churn through a customer retention marketing campaign utilizing churn and RFM analysis.

## Product Focus

Increase profit through a marketing campaign for product cross-selling, employing market basket analysis.

## Feedback Focus

Target both detractor and promoter customers with a loyalty-focused engagement marketing campaign to mitigate the adverse effects of detractors and encourage the positive impact of promoters, leveraging sentiment analysis.

03

## Data cleaning

During this stage, we primarily focused on defining the attributes' desired format, identifying missing data, and excluding problematic records.

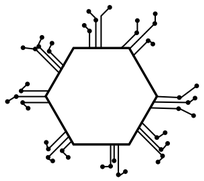
## Preprocessing

Subsequently, we structured the data into a format suitable for the data modeling phase.

## Data Exploration

We conducted various qualitative and quantitative analyses to identify and eliminate redundant attributes.

## Modelling



We developed and deployed different models.

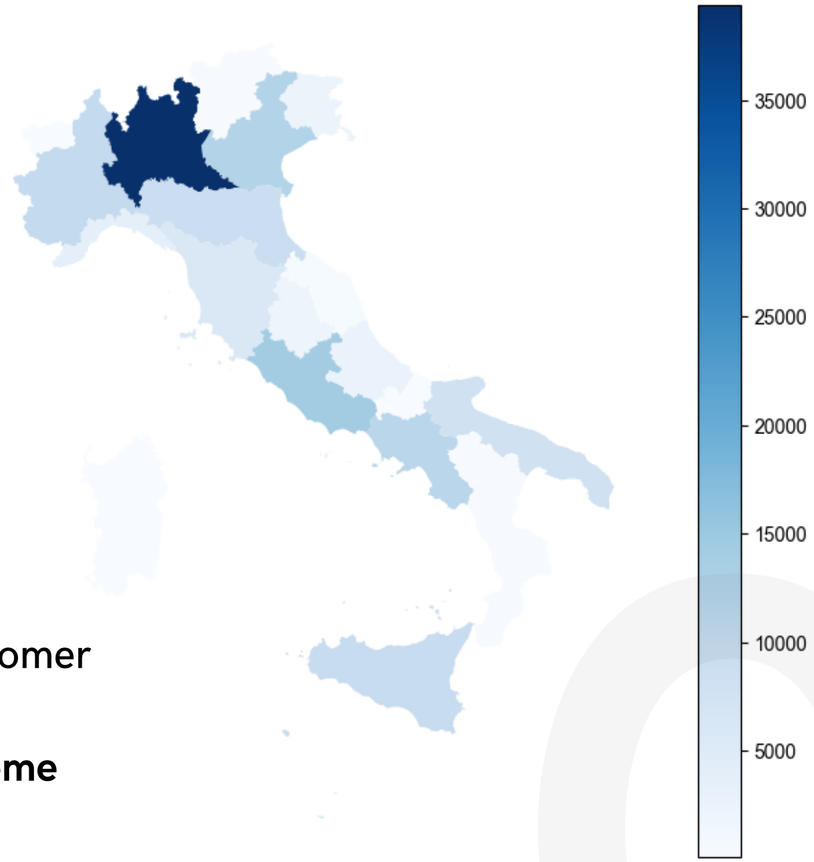
## Evaluation

We integrated the results from the different models to facilitate the optimal data-driven marketing actions.

# Data exploration

## Customer distribution

This map represents the number of customers per region. **Lombardy** is the region with most of the customers (29%), followed by **Lazio** (10%) and **Veneto** (9%). **Sardegna**, **Molise** and **Aosta** are the regions with less customers (<0.2%)




It is crucial to delve deeper into the drivers of customer distribution, considering **factors beyond mere population density** such as **economic** factors, **income** levels, **infrastructure** and **accessibility**.

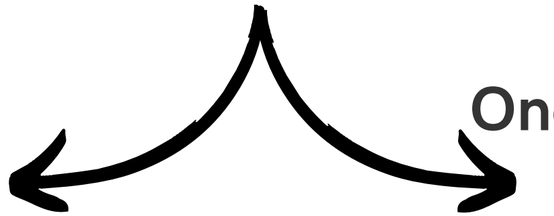
# Data exploration

## Inactive customers

**26%**

The customers who have not placed any orders in the time period from 1 May 2022 to 30 April 2023

## Active customers



### Repeaters

**68%**

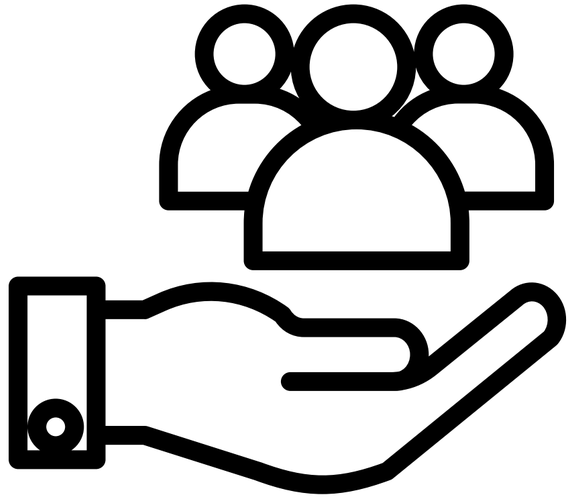
Customers have placed more than one order in the last year

### One-shooters

**32%**

Customer have placed only one order in the last year

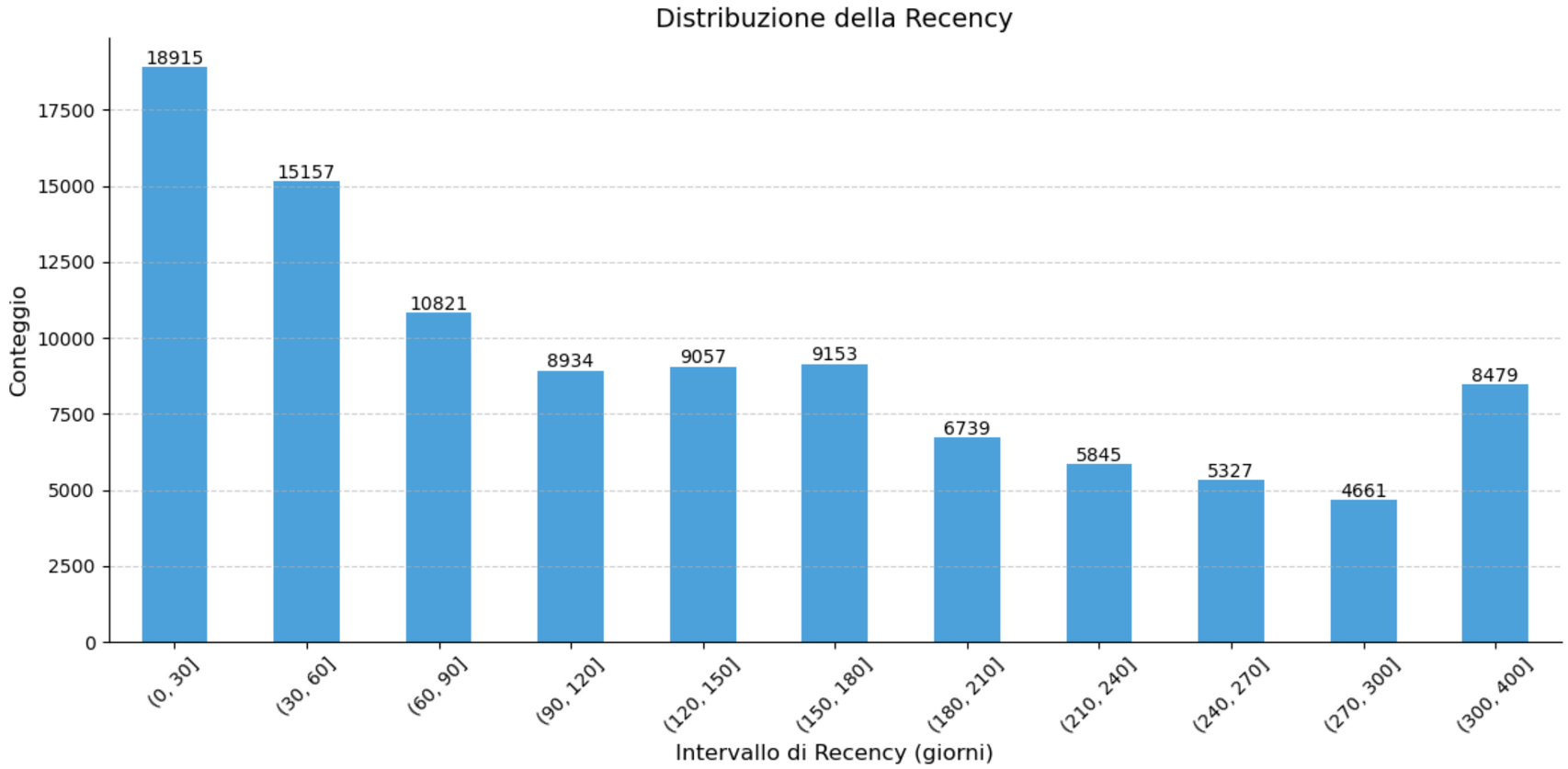
# Costumer focus



Customer focused analysis is a business approach that centers on prioritizing and addressing customer needs, desires, and expectations to enhance customer satisfaction and achieve business success.

# R-Recency

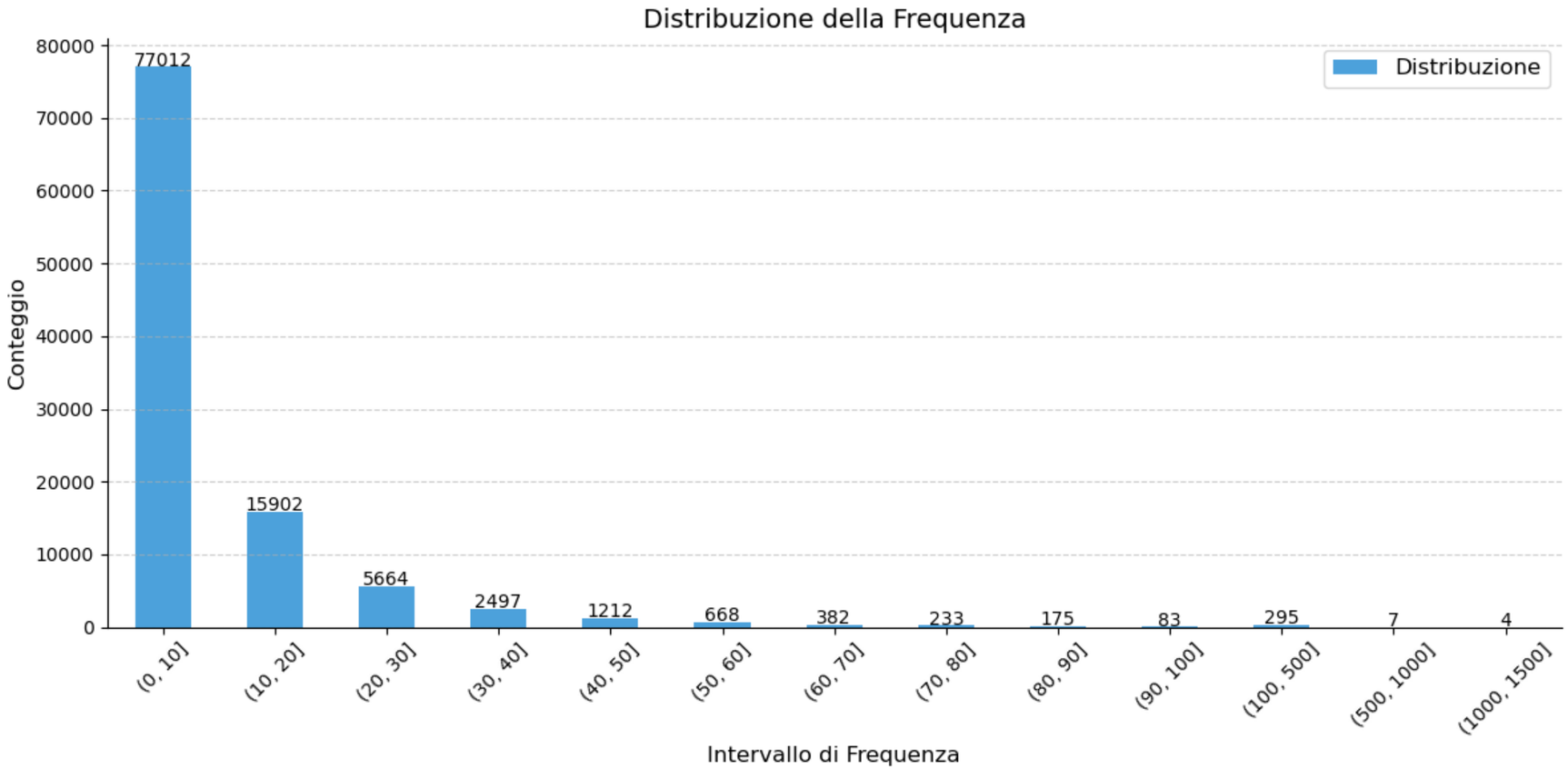
In RFM analysis, "recency" is one of the three key components used to segment and analyze customer data. It refers to how recently a customer has engaged with or made a purchase. It measures the time elapsed since the customer's last interaction or transactionency for customers who had made a purchase with a company.




The distribution peaks to the left, indicating that more customers have made their last purchase recently.

# F-Frequency

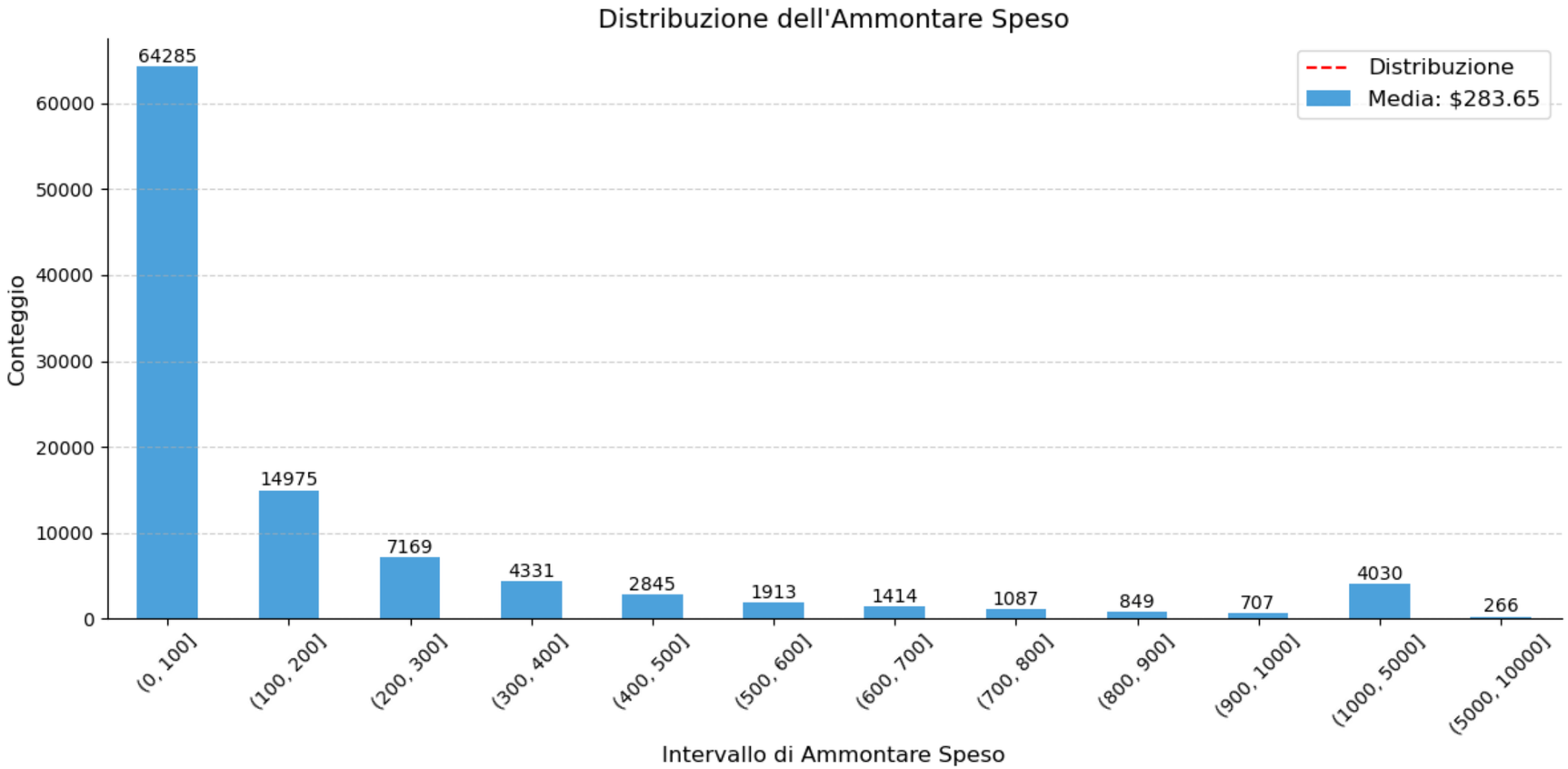
It measures the number of interactions or purchases made by a customer during the specified time period. Customers who make frequent purchases or interactions are typically considered more loyal and valuable to a business because their ongoing engagement often leads to higher revenue and a stronger customer relationship.




As can be observed, the vast majority of customers have placed orders ranging from **0 to 10.**

# M-Monetary

It represents the monetary value of a customer's transactions or purchases over a specific period. It quantifies how much money a customer has spent on products or services.




From the distribution we can see that customers on average spend small amounts of money, **0 to 100**.

# Customer Rating by RFM Score

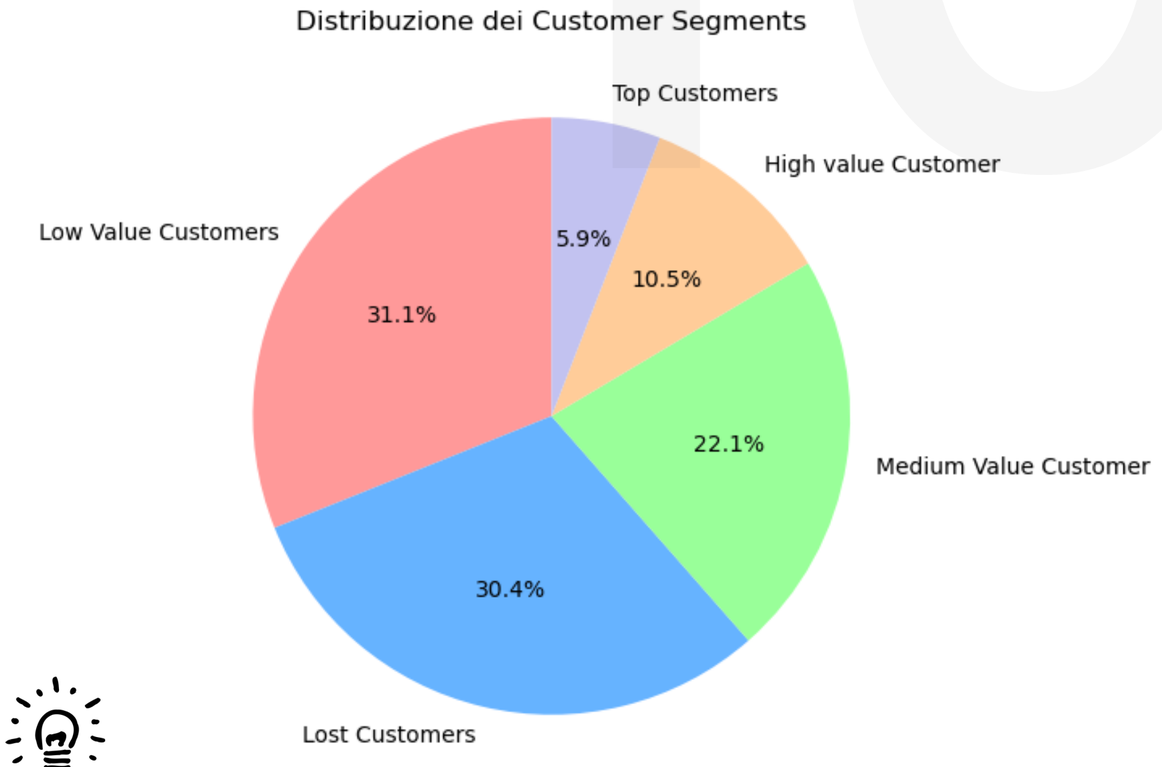
**Top Customer (RFM score > 4.5):**
High recency, frequency, and monetary value customers. Our most valuable segment.

**High-Value Customer (4.5 > RFM score > 4):**
Not top-tier but still valuable, contributing significantly.

**Medium-Value Customer (4 > RFM score > 3):**
Solid contributors to our business, occupying the middle ground.

**Low-Value Customer (3 > RFM score > 1.6):**
Customers who contribute moderately to our business.

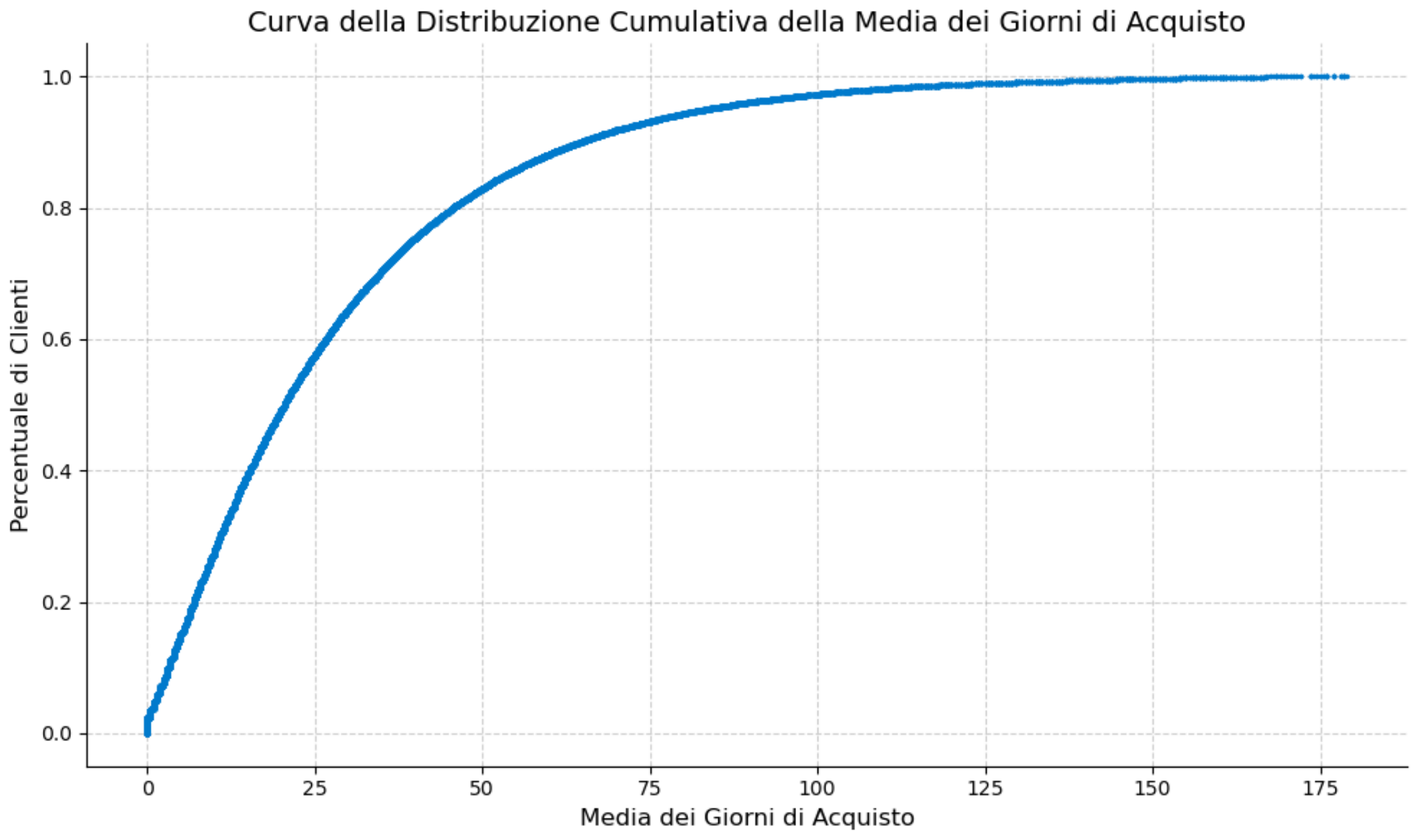
**Lost Customer (RFM score < 1.6):**
Less engaged or inactive customers, requiring re-engagement efforts.




We need to investigate if there are any issues with customers who have **low RFM scores**. Additionally, it might be worth considering sending **exclusive email discounts** to encourage these customers to make purchases.

# Repurchase time scale

Considering the customers who made more than one order, the average number of days between transactions was computed in order to evaluate the repurchase time scale was evaluated.




90% of consumers repurchase within **65** days.
Based on the insights from this curve, we have determined the threshold value for segmenting the customer base for all the models we will explore.

# Churn Model

Churn Model is indeed concerned with predicting which customers are most likely to discontinue their relationship or stop using a company's products or services.

As for the explanatory variables we selected:

- **(Recency, Frequency, Monetary)**: we used the previously calculated RFM.
- **Età**: we incorporated customers age.
- **review_text**: If the customer has ever written a review, it has been transformed into a binary variable.
- **loyalty_type**: The type of loyalty account that the customer has.

We use the **Repurchase Time Scale** curve to find the target variable:

- **1:** when the customer has not purchased on average for 65 days
- **0:** when the customer has purchased before 65 days.

# Churn Model

We encountered an issue with imbalanced classes, with only 10% of customers in the churn category. Therefore, oversampling and undersampling techniques were applied, resulting in an approximately 50% balance between churn and non-churn customers. Finally, oversampling was chosen as it provided better results to the models.

- **Logistic Regression**: is a statistical method for binary classification, estimating the probability of one of two outcomes based on input features.
- **Decision Tree**: used in churn analysis to predict customer attrition by mapping decision rules based on customer data.
- **Random Forest**: a potent ensemble learning technique utilized in churn analysis to boost prediction accuracy by amalgamating multiple decision trees based on customer data.

# Churn Model

| Model | Accuracy | Precision | Recall | F1-score | AUC |
|---|---|---|---|---|---|
| **Logistic Regression** | 0.66 | 0.68 | 0.62 | 0.65 | 0.75 |
| **Decision Tree** | 0.96 | 0.93 | 1.00 | 0.96 | 0.96 |
| **Random Forest** | 0.97 | 0.95 | 1.00 | 0.97 | 0.99 |

The **Random Forest** model has shown superior performance, with a F1-score of 97%. This model can be valuable for identifying the variables that have the most significant impact on the likelihood of customer attrition.

Churn models identify at-risk customers and take steps to retain them, reducing customer loss. By retaining existing customers, companies preserve revenue, reduce the need for costly customer acquisition and optimize resource allocation.

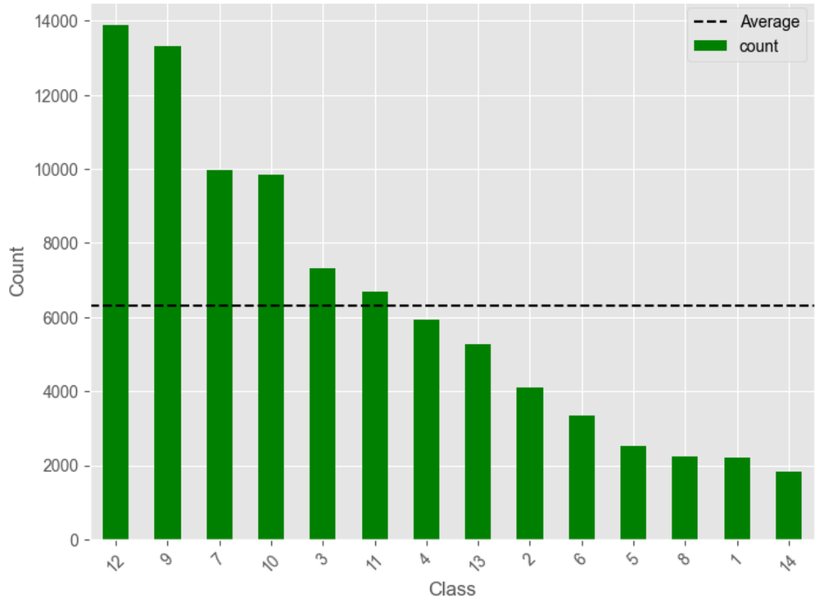
# Product focus

Product focused analysis can significantly enhance the profitability of a marketing campaign through cross-selling products by identifying **associations** and **patterns** among customer purchases.

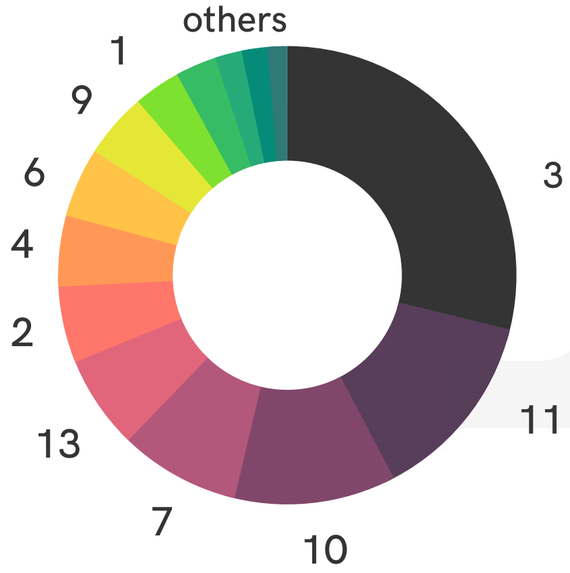
# Product class distribution

The majority of products belong to category **12, 9, 7.**
However, the majority of purchased products belong to category **3, 11, 10**, representing the **29%**, **13%** and **11%** of the purchases respectively.

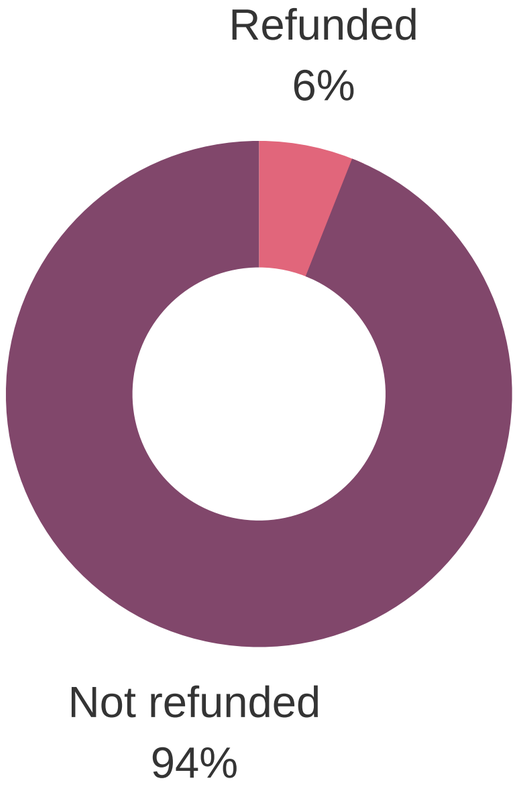
**Product class distribution**




**Purchased products** by class



# Refunded products




Among all the products sold, **6% are refunded.**

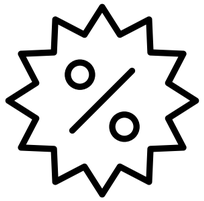
The **most refunded products belong to class 3** (35320 refunds) representing 56% of total refunds.
They are followed by categories 6 and 7, representing 11% and 7% of total refunds.

The most refunded product has the following id: 35662515. It was refunded 3492 times!

It is important to understand the different reasons behind refunds, such as products defects, quality issues or any  store-related factors.

# Profits and discounts

Products belonging to **category 2 provide the greatest profit** to the company, followed by categories 6 and 7. They represent respectively 39%, 15% and 13% of the total profits.



Price reduction in transactions was analysed to understand their potentials:
**42% of customers have purchased discounted products!**
The discounted products most frequently purchased are 48011971 and 48020504, both belonging to class 2.
Discount amount depends on the store.

By focusing on targeted discounts and refining the pricing strategy, it can be possible to leverage the profit potential of category 2 products while satisfying the customers' desire for discounts, ultimately improving business's overall performance.

# Market Basket Analysis

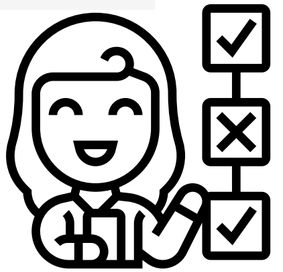
Without considering refunds and transactions with only one products, Market Basket Analysis is performed to understand **customer behavior** and identify **patterns of co-occurrence among items in shopping basket**.

A pivot table was created, with transaction as rows and products as columns.
To uncover frequent itemsets for association rules, the **Apriori algorithm** was employed, with a minimum support threshold of 0.003.

A set of rules is obtained, containing antecedent product and consequent product with support, confidence, lift, leverage and conviction evaluation metrics.

# Market Basket Analysis

The top rules were extracted by evaluating the **lift metric**.
Threshold set: 0.7 for confindence metric and 0.004 for leverage metric.
**The extracted rules contain products of category 3 as antecedent and consequent**.
The top 2 rules have a lift of 119.5 and they contain the same products, meaning that the antecedent of one is the consequent of the other and viceversa.



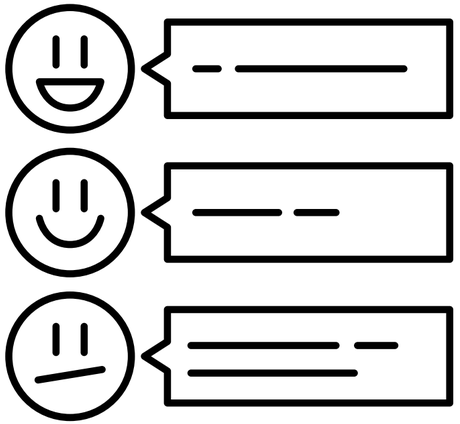
**The high lift values indicate a very strong association between the two products**.
Customers who purchase one of these items are highly likely to purchase the other as well. It's a strong indicator that customers perceive these **items as complementary** or have a **strong preference for buying them together**.

It is reasonable consider bundling these two products together in promotions or placing them in close proximity in their stores to encourage cross-selling. Offering discounts for antecedent products can be also viable strategy to boost sales and promote related products.

# Feedback focus

Feedback focused analysis is important to allow a loyal engagement marketing campaign to **reduce the negative impact** of detractors and to **incentive the positive effect** of promoters

# Sentiment Analysis

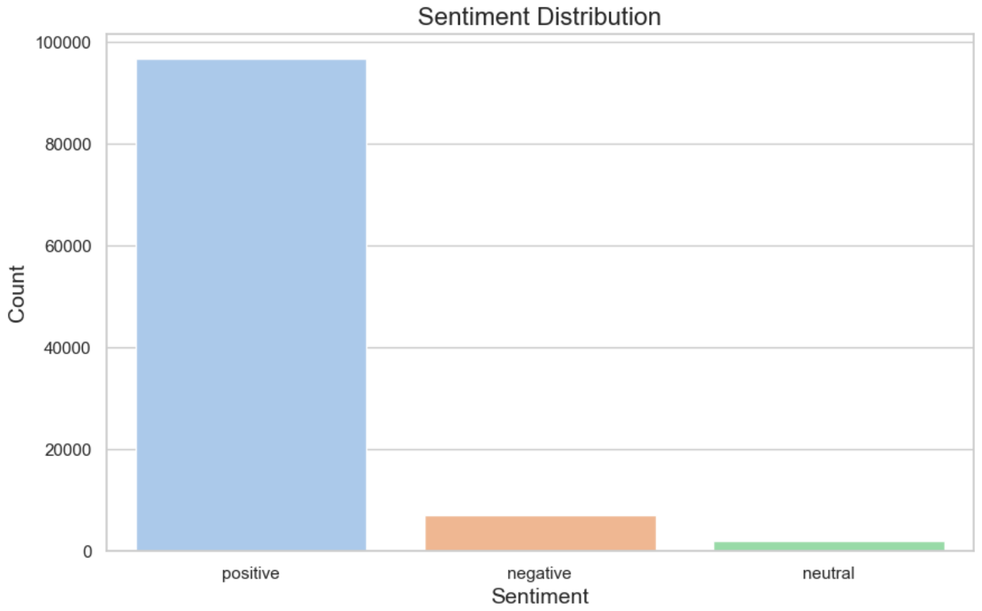


In order to extract the sentiment from the feedbacks, some **preprocessing steps** were applied: removal of punctuation, emojis, newlines, extra whitespaces, stopwords and links, followed by tokenization and lemmatization.

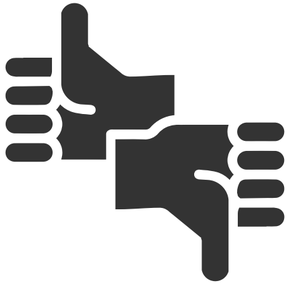
Consequently, **VADER sentiment analysis** was applied. It is a rule-based sentiment analyzer, that is sensitive to both **polarity** and **intensity** of **emotion**. It relies on a dictionary which maps lexical features to emotion intensities called sentiment scores.

Finally **polarity classification** is perfomed: the sentiment score is obtained as the sum of intensity of each word in the text and then classified into:
Positive, negative and neutral

# Sentiment Analysis




The majority of **feedbacks are positive**.



Positive: 91%
Negative: 7%
Neutral: 2%

70% of positive and negative feedbacks come from customers with a **standard loyalty type.**

Analyzing these customers and the products in depth is crucial, as it enables the **development of personalized marketing** campaigns and **product-focused initiatives**.

# Most frequently used words

Wordclouds are a valuable tool as they can help understanding the key elements that influence negative and positive feedback. They offer a quick and intuitive way to extract meaningful insights from unstructured text data, making it easier to make data-driven decisions and take action based on customer feedback.

**Positive feedbacks**


**Negative feedbacks**


The most used words for negative reviews are: **product, taste, coffee, time, bad.**

The most used words for positive products are: **love, good, look, want, try, think.**

SEPTEMBER 2023

Master Degree - Data Science
Prof. Nico Di Domenica
Prof. Giovanni Collini

# Thank you!

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

Julia Lan Bui Xuan 882385
Michele Salvaterra 891109