Marketing Analytics

Samir Doghmi 897358 – Roberto Ferrari 852220 – Davide Prati 845926

OBJECTIVES

Explore available data, from **2022-01-01** to **2023-05-09**, to identify patterns, seasonality of sales **regardless refunded orders**, and propose new **data-driven strategies** for the marketing campaign.

Through **sentiment analysis**, enanche customer service by in-depth analysis of customer reviews and feedback

Development of a marketing strategy based on **RFM segmentation**

A **product association analysis** is performed to identify product combinations frequently co-purchased within orders

We want to reiterate how all the considerations and analyses made were done only on **purchased orders**; we will devote an additional discussion to **refunded orders** in Section 1

TABLE OF CONTENTS

1 **Business overview and visual highlights**:
- Customer Analysis
- Sales analysis
- Products and Favorite stores insight
- A Look at the Refunded Orders

2 **RFM analysis**:
- Active vs Inactive based on repurchase curve
- RFM-classes
- Suggestions for optimal marketing investmen

3 **Churn Analysis**:
- Churn distribution
- Model evaluation for optimal performances
- Recomended retention strategies

4 **MBA**:
- Identify product affinities in customer behavior
- Reccomendations to incentive purchasing and design promotions

5 **Sentiment Analysis**:
- Predict and examine emotional tone with Berta algorithms
- Featured strategies for retaining clients with negative feelings

6 **Extra Points**:
- **Extra (1)**: Exploit the re-training strategy of a large pretrained language model (Doc2Vec) for the Feedback-focused strategy
- **Extra (2)**: a detailed marketing campaign that combine customer RFM and product MBA results

ANALYSIS FLOW: A DATA-DRIVEN COMMUNICATION STRATEGY

● DATA CLEANING

in this phase, operations were performed to define the format of the attributes, handle inconsistencies in the data, identify missing values and generate new columns useful for later investigations

● PREPROCESSING

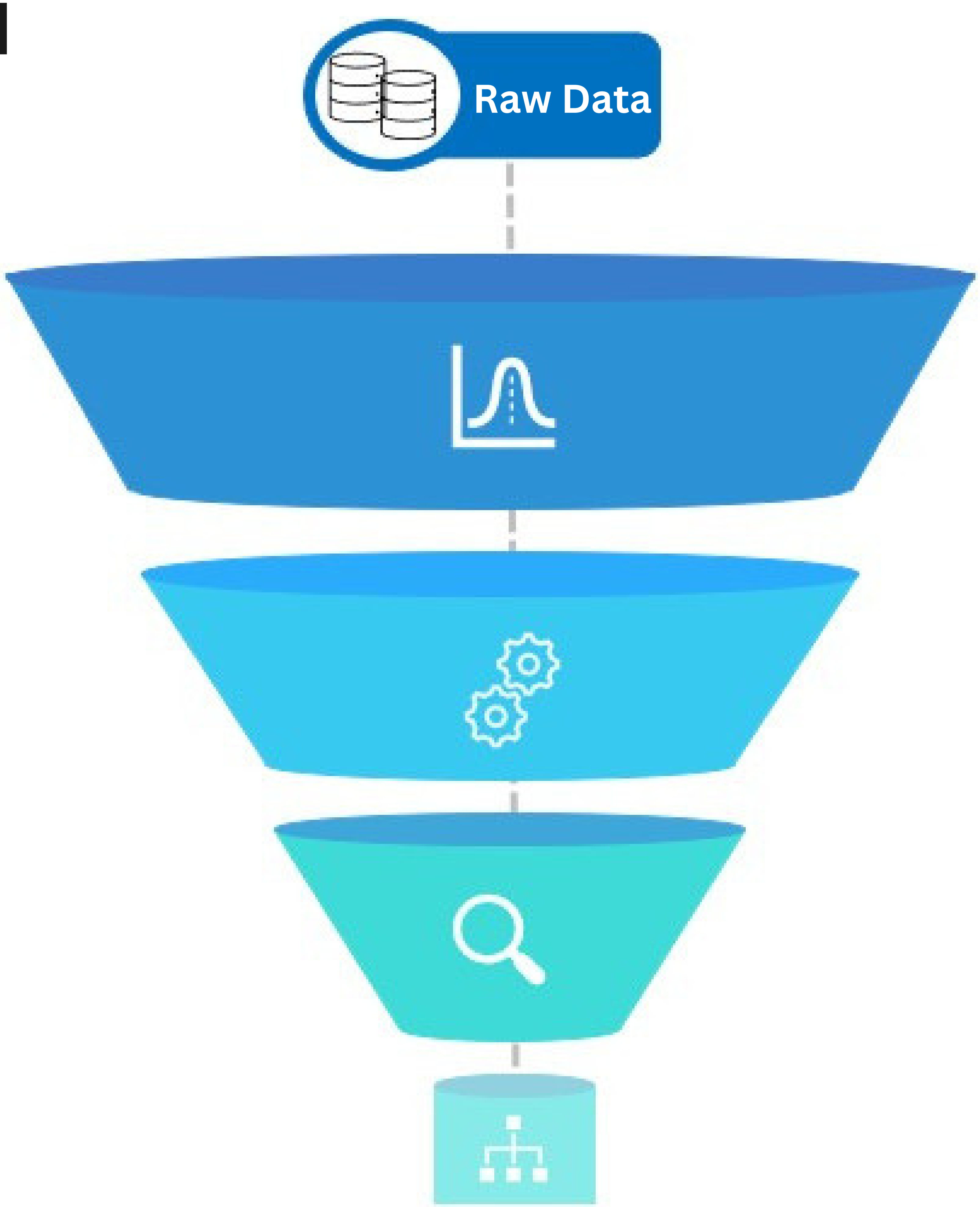
The objective of this phase was to merge the CSV files, remove duplicates and conduct an outlier analysis using ANOVA and t-Student tests.

● DATA EXPLORATION

In this phase, a descriptive analysis of the variables deemed most relevant was conducted to obtain a deeper understanding of the overall business.

● ADVANCED ANALYSIS:

RFM analysis and sentiment analysis were conducted. Subsequently, at this stage, the data were prepared to be used as input for various machine learning algorithms with the aim of developing a churn prediction model.

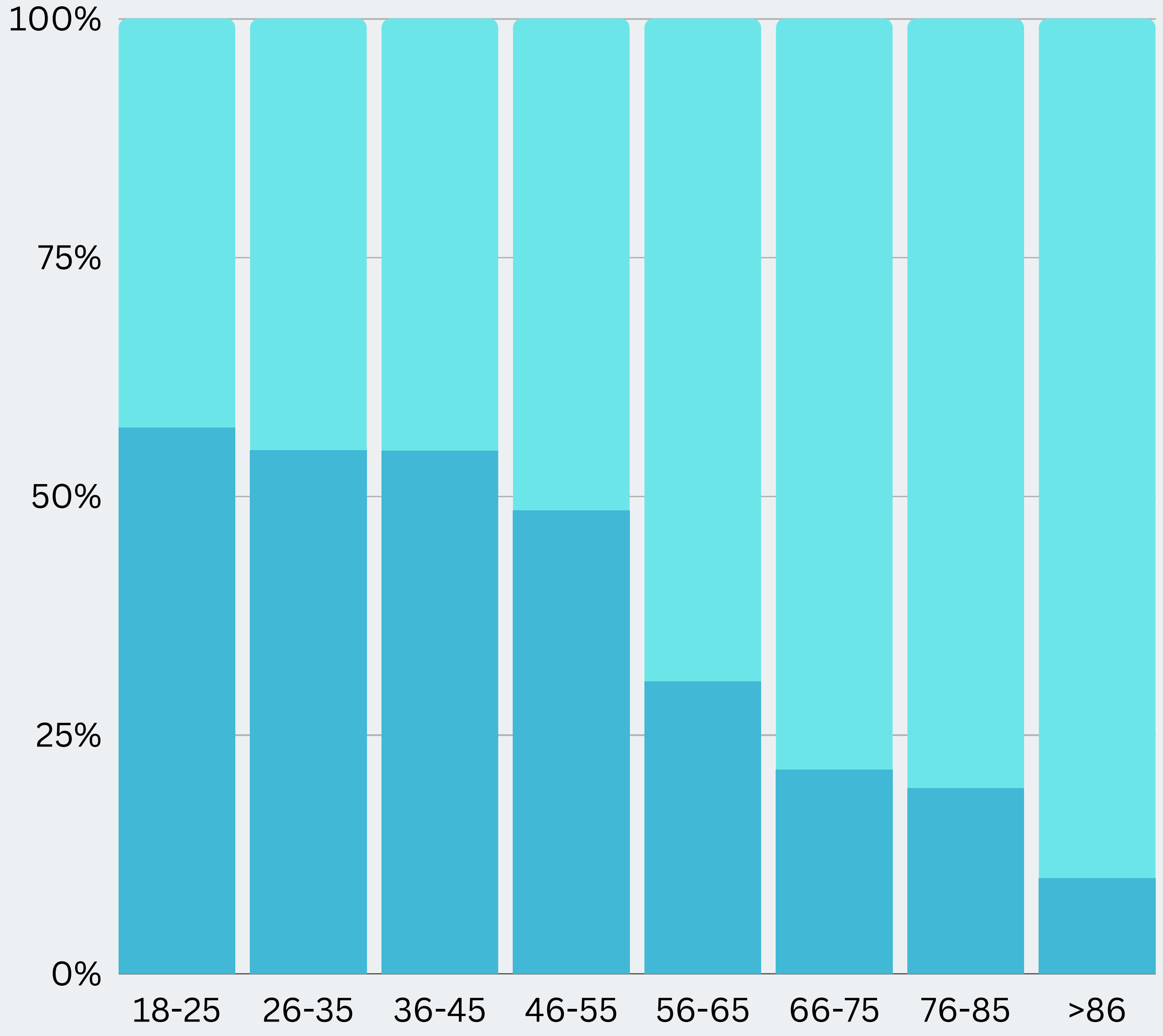


1 CUSTOMER ANALYSIS

Gender Distribution

50.05% 49.95%

Gender Distribution and Age



Macro-Regions

39% Northwest

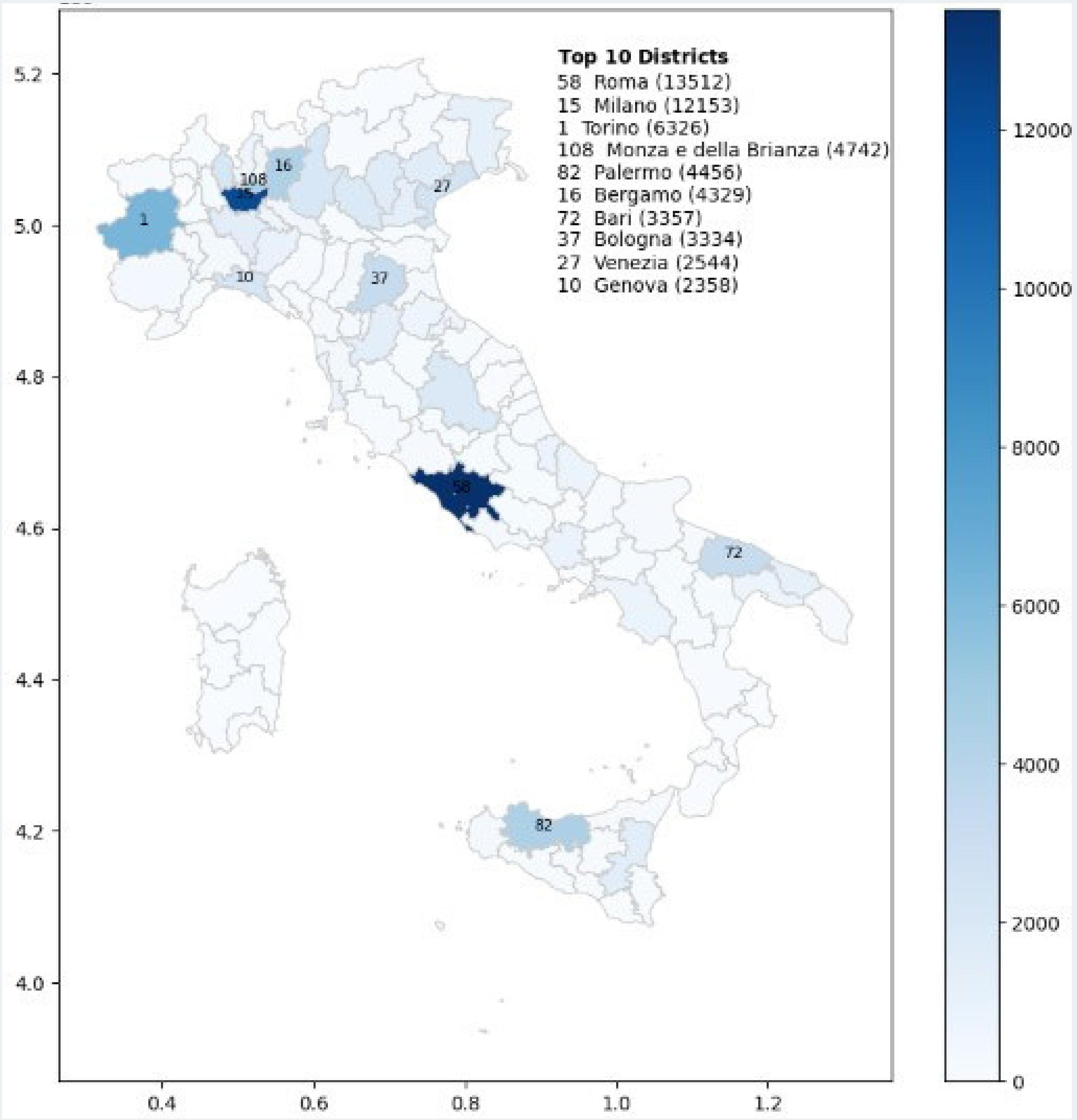
17% Northeast

20% Central

17% South

7% Islands

Aggregated Customer Accounts by District

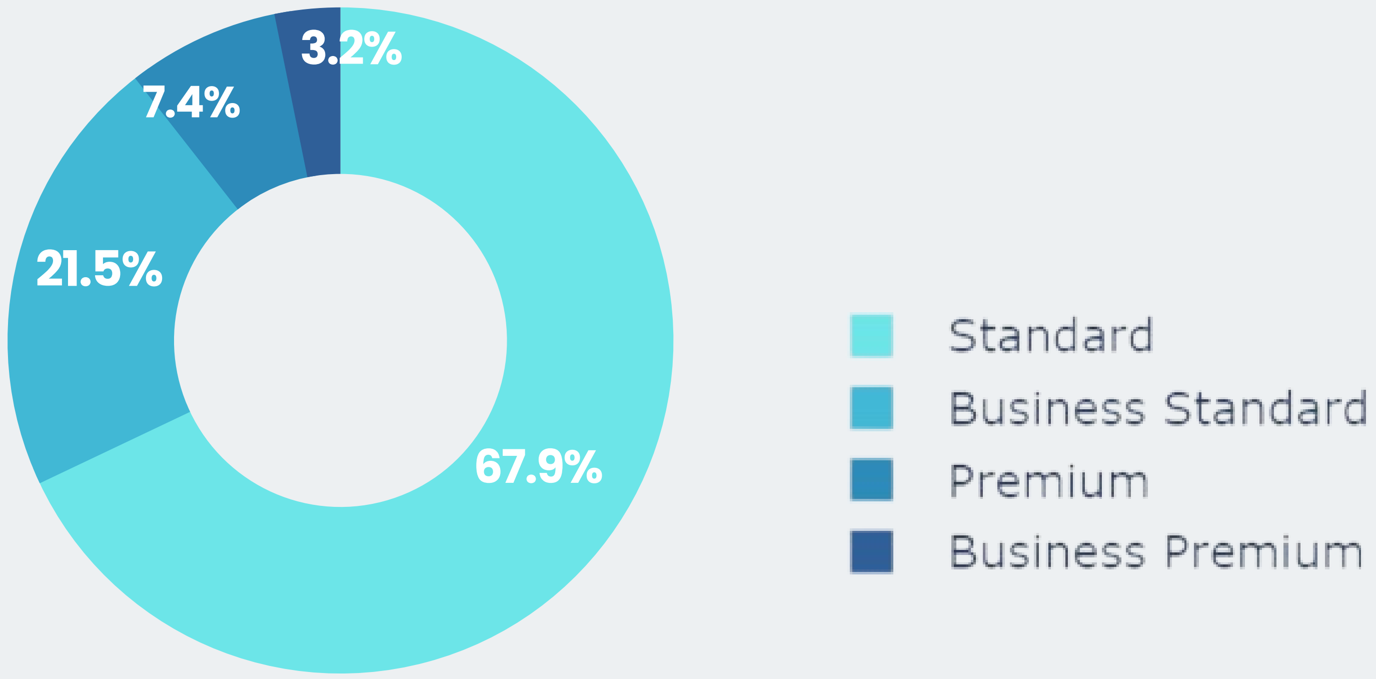


Top 10 Districts

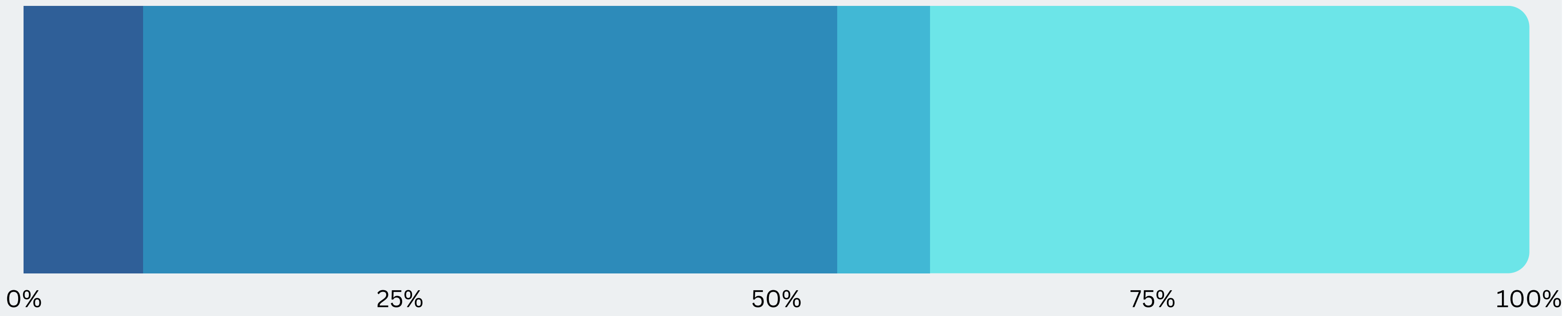
58	Roma (13512)
15	Milano (12153)
1	Torino (6326)
108	Monza e della Brianza (4742)
82	Palermo (4456)
16	Bergamo (4329)
72	Bari (3357)
37	Bologna (3334)
27	Venezia (2544)
10	Genova (2358)

1 CUSTOMER ANALYSIS

Custom Distribution per Loyalty Plan

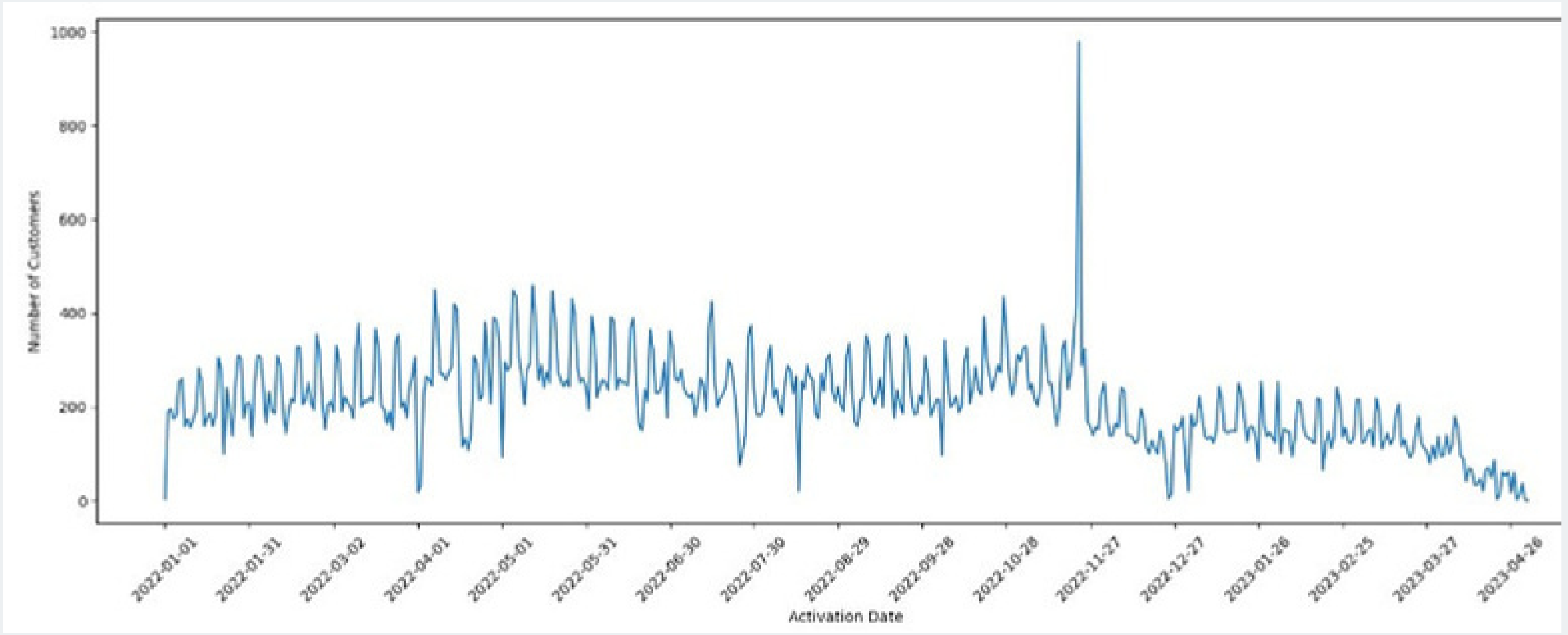


Total Purchases by Loyalty Plan



Although the standard plan is the most prevalent, the descriptivs analysis shows that customers with a **business plan** are the most **profit making** for the business

Distribution of Account Activation Dates Over Time



Maximum Number of Customers Sharing an Account

The total number of shared accounts is 372, of which the majority, 222, are related to the Business plan and 132 to the Standard plan. In terms of age combinations, there are 24 accounts shared between the **[36-45]** and **[46-55]** groups, followed by 22 accounts shared between the **[36-45]** and **[26-35]** groups.

1 SALES OVERVIEW

Tot amount spent

18.93 Millions €

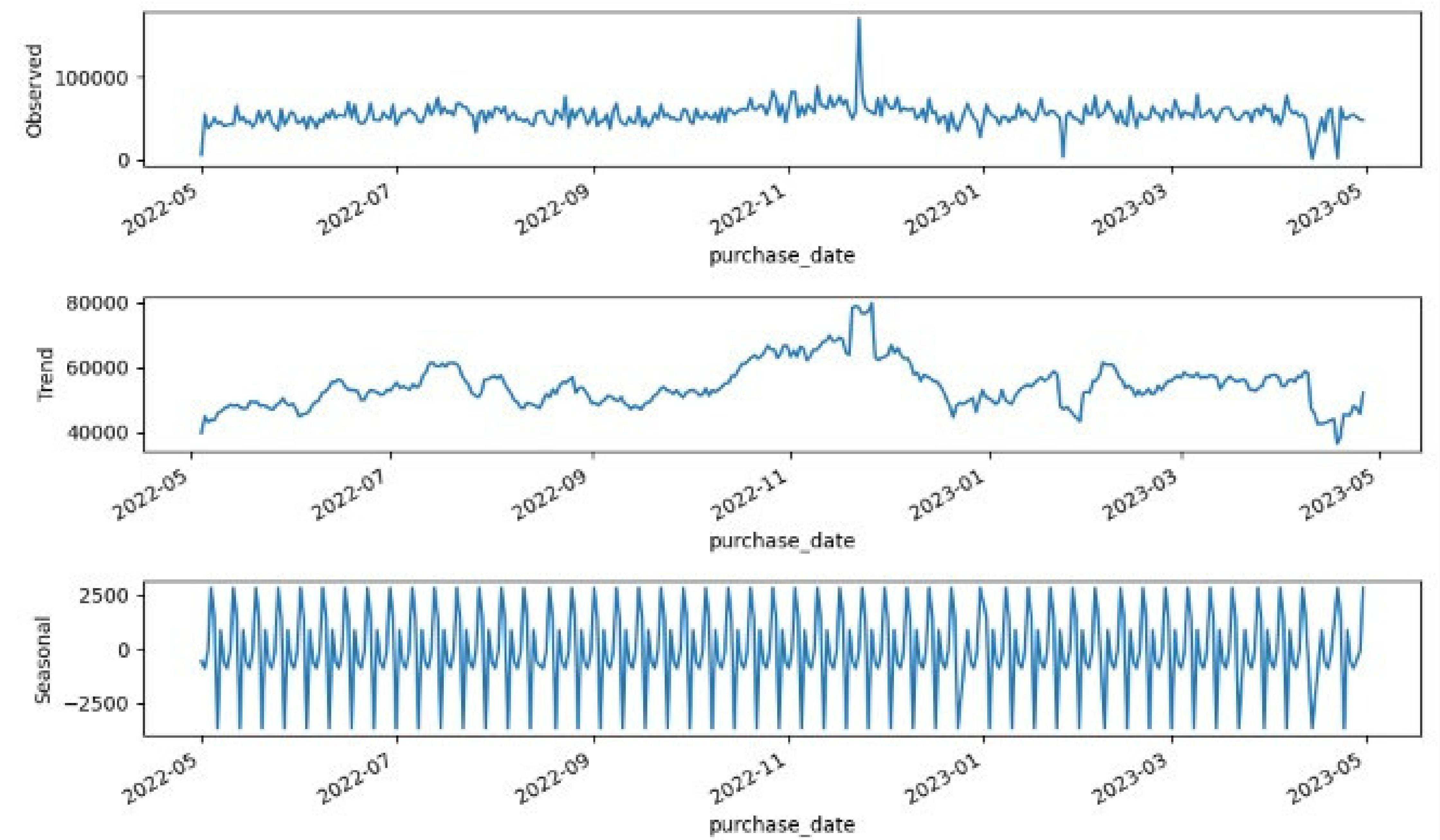
Tot orders

885,435

Orders with at least one discounted item

73,45%

Sales Insights



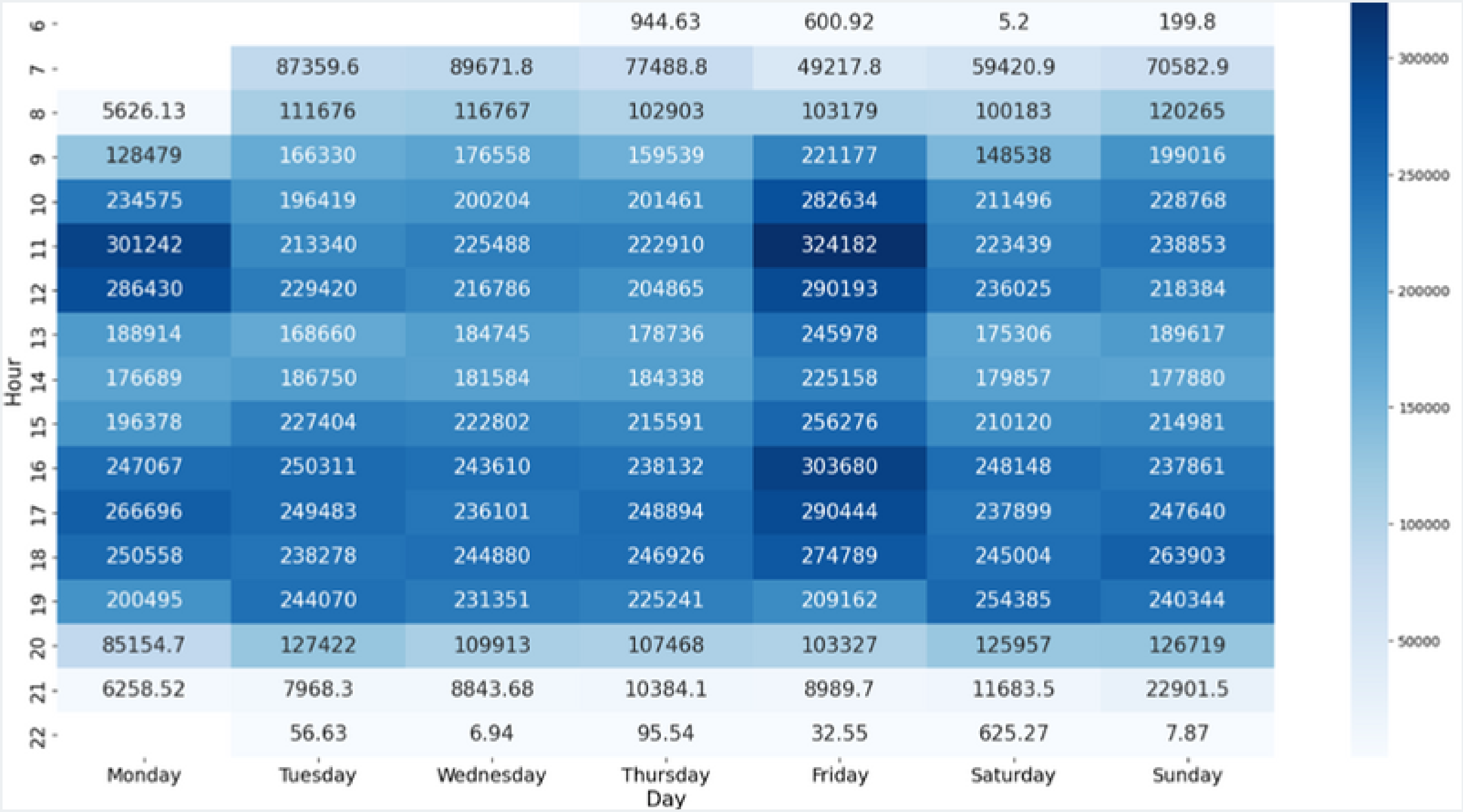
2022-11-23, the day with the highest number of account activations, also saw the highest number of orders (2249), with a sales volume of EUR 64967.66, making it the day with the highest volume of sales and orders.

1 SALES OVERVIEW

It is clear that **Friday** is the most **profitable** day, with peak shopping preferences concentrated during the time slots from **10 am to 12 pm,** and **4 pm to 7 pm**.

To **alleviate** congestion, an alternative approach could involve providing **vouchers** that incentivise customers to shop on **less frequented days and times.** This not only benefits customers by offering more pleasant and relaxed shopping conditions, but also optimises overall shop operations.

Sales heatmap: hour vs day



Hour	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
6				944.63	600.92	5.2	199.8
7		87359.6	89671.8	77488.8	49217.8	59420.9	70582.9
8	5626.13	111676	116767	102903	103179	100183	120265
9	128479	166330	176558	159539	221177	148538	199016
10	234575	196419	200204	201461	282634	211496	228768
11	301242	213340	225488	222910	324182	223439	238853
12	286430	229420	216786	204865	290193	236025	218384
13	188914	168660	184745	178736	245978	175306	189617
14	176689	186750	181584	184338	225158	179857	177880
15	196378	227404	222802	215591	256276	210120	214981
16	247067	250311	243610	238132	303680	248148	237861
17	266696	249483	236101	248894	290444	237899	247640
18	250558	238278	244880	246926	274789	245004	263903
19	200495	244070	231351	225241	209162	254385	240344
20	85154.7	127422	109913	107468	103327	125957	126719
21	6258.52	7968.3	8843.68	10384.1	8989.7	11683.5	22901.5
22		56.63	6.94	95.54	32.55	625.27	7.87

1 PRODUCTS AND STORES

n products	n class products	n stores
2000	14	49

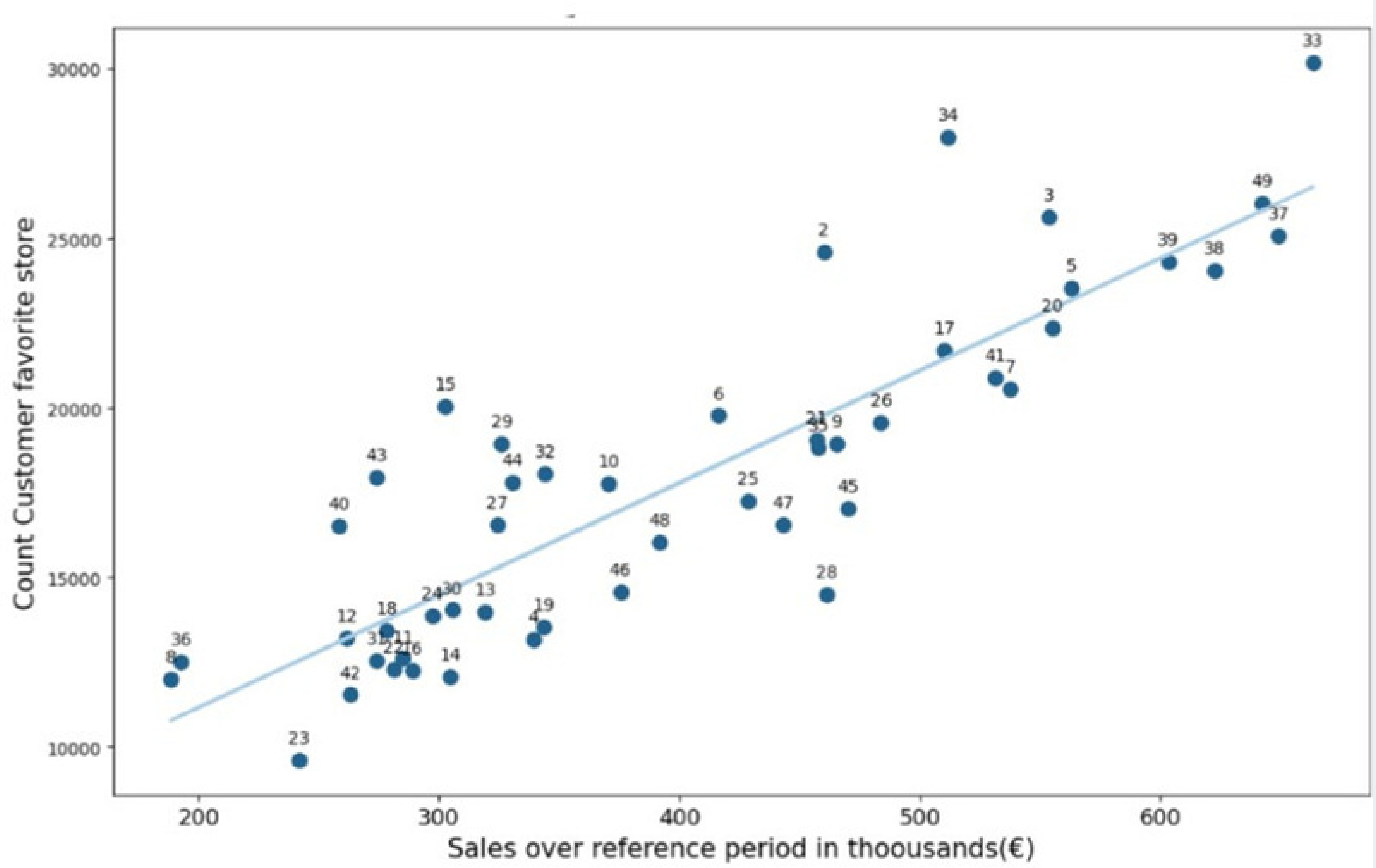
Top Sales by class product

Class	Value	Avg discount(%)
6	€4235471.90	4.23
7	€3546910.43	3.13
2	€3048921.70	5.15
3	€2399729.57	2.01
13	€1424491.40	2.21

Top Sales by products

Product	Value	Avg discount(%)
48020504	€588075.32	10.66
48011971	€574806.96	9.81
36483013	€172847.64	9.22
48020203	€164432.98	11.03
33883955	€158387.76	3.72

Favorite store by customers vs Store sales



1 A LOOK AT REFUNDED ORDERS

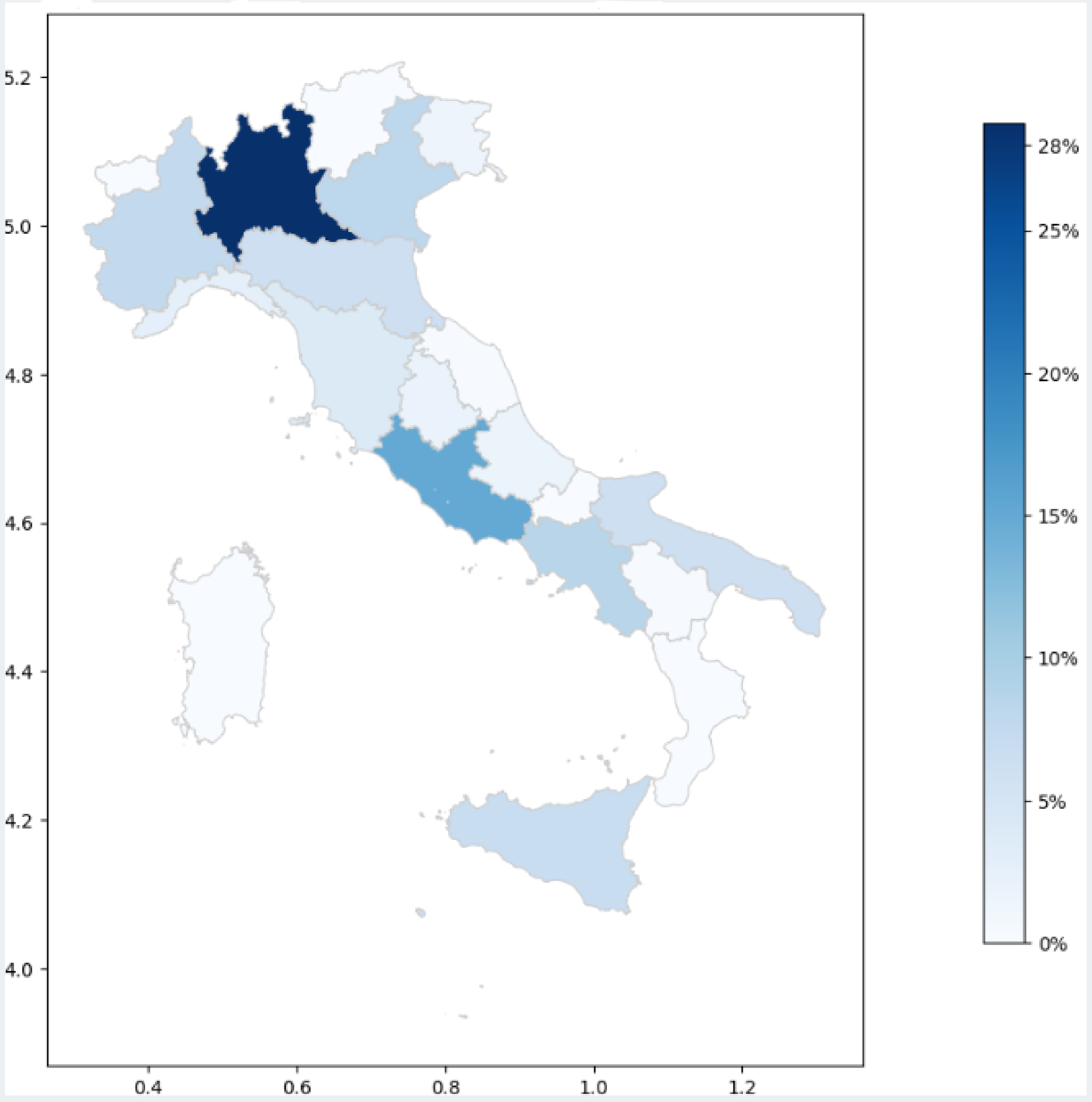
12,62%

Customers who have refunded at least one item

Most customers have a **standard or business plan**, respectively 53% and 37% of the total. The **age group [36-45]** is the most represented and exceeds the age group [46-55] in second place by a factor of three.

The peak of sales returns was observed on **2022-11-23**, with EUR 9,050 refunded. In addition, **November 2022** recorded the highest monthly sum of sales returns, with a total of EUR 129,435 and 3,673 items reimbursed.

Aggregated Customer Accounts by Region



1 A LOOK AT REFUNDED ORDERS

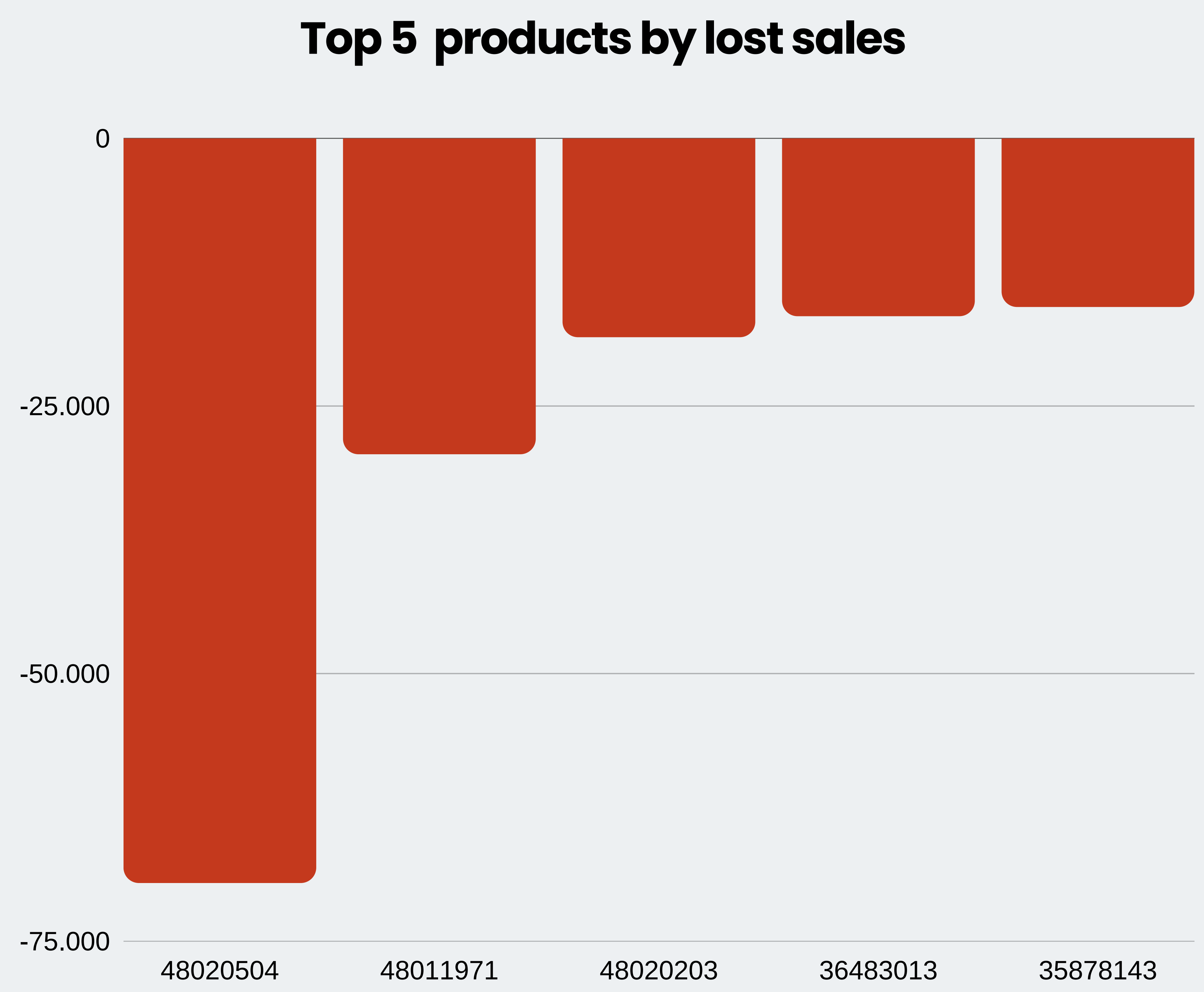
Tot amount refunded

6.3 million €

Tot orders

80495

Make a **Good First Impression** With Detailed Descriptions and High-Quality Visuals. **Improving product descriptions** and **visuals** is crucial for reducing returns and enhancing customer satisfaction. For example, high-quality images, videos, and customer reviews help customers make informed decisions. Furthermore, the return of items can be **discouraged** by implementing a **strategy of increasing the associated costs**.



2 RFM ANALYSIS

RFM analysis is a customer segmentation technique that stands for Recency, Frequency, and Monetary value. It is a marketing analysis tool that identifies an organization's best customer segment based on these three essential customer behavior traits:

- **Recency (R)**: How recently did the customer make a purchase?
- **Frequency (F)**: How often does the customer make purchases?
- **Monetary value (M)**: How much does the customer spend?

But before that, to distinguish between **inactive** and **active** customers, we establish a **threshold** of **106 days** that captures the period in which **90% customers repeatedly complete a purchase**.

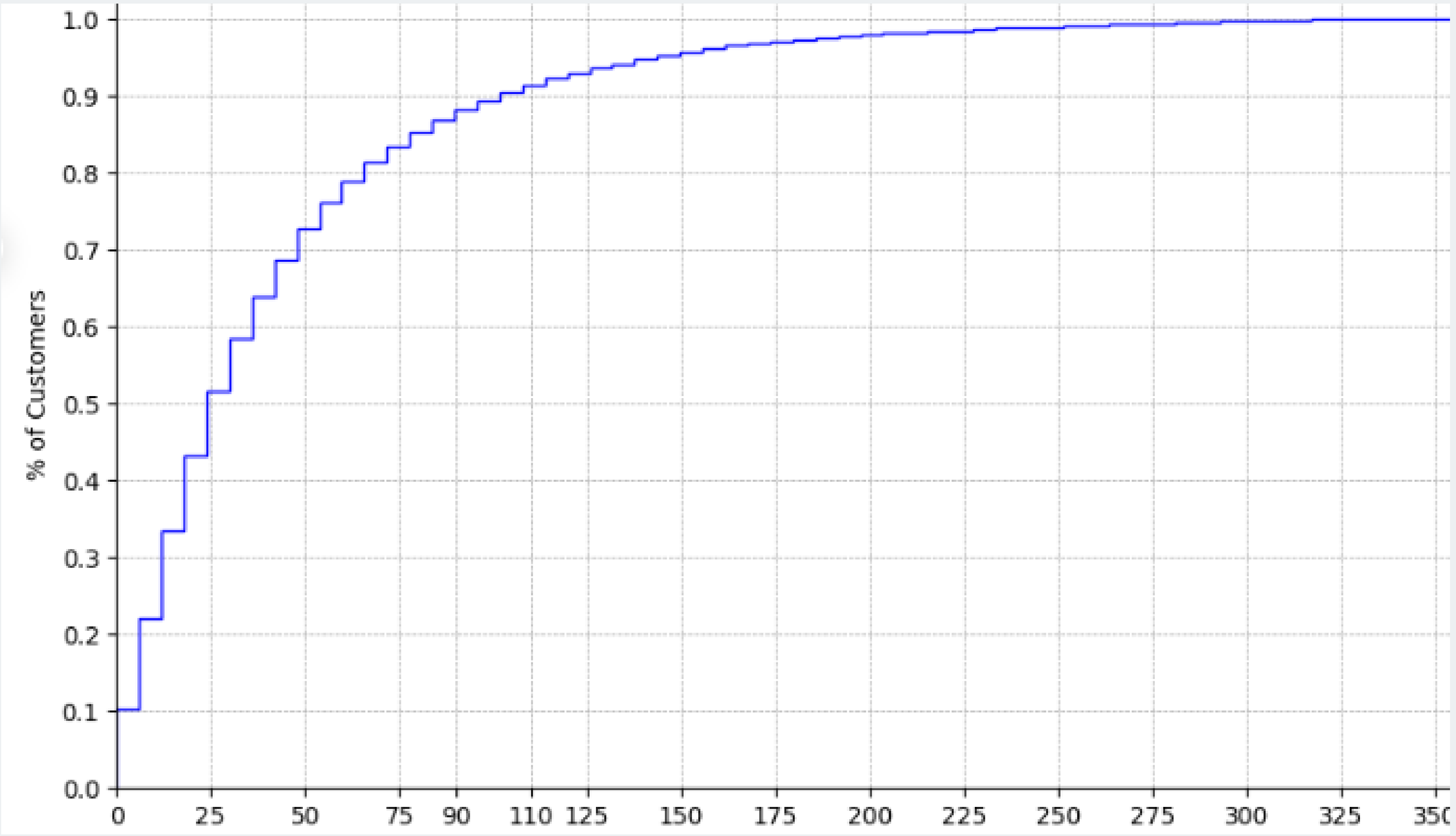
Notably, during the period from **May 1, 2022** to **April 30, 2023, 34,851** people made a single purchase.

Customer Base Distribution

39% Inactive

61% Active

Customers by days for next purchase



2 RFM ANALYSIS

We divide active customers into **three groups** using the percentiles of each of the RFM measures. Then, we combine the groups of **frequency** and **recency percentiles** to calculate new **customer loyalty classifications**.

Finally, we integrate these loyalty states with the **monetary value** categories to create complete **RFM classes:**

	Diamond n. 16537	Gold n. 13704	Silver n. 11375	Bronze n. 4721	Copper n. 3925	Tin n. 7707	Cheap n. 5390
Average spending	€643.14	€265.82	€63.84	€189.25	€249.47	€58.05	€24.12
Average n. of orders	8.29	5.67	4.21	3.03	2.00	2.00	2.00
Average n. of items per order	2.93	2.51	2.01	2.17	3.15	2.22	1.93
Days until next purchase	52	103	116	223	177	181	64

2 RFM ANALYSIS

Diamond and **Gold** classes are the most **cost-effective customers** and generate the **most value** for the company.They are the top-tier customers who contribute significantly to business.
The following are **suggested marketing strategies** to decrease costs, increase customer loyalty and ensure their satisfaction:

EXCLUSIVE LOYALTY PROGRAM

Design an exclusive loyalty programme, offering personalised benefits and early access to sales promotions or gifts after reaching certain spending levels. The objective is to strengthen the bond with clients, ensuring a lasting and mutually beneficial relationship.

PERSONALISED CUSTOMER SERVICE AND FEEDBACK MECHANISM

Provide a dedicated customer service line to promptly answer questions and effectively resolve any problems fostering higher customer satisfaction while extracting valuable insights for operational optimization. In addition, Employ a versatile communication strategy, including email, social media, and direct calls, to engage customers and collect valuable feedback

CUSTOMISED CUSTOMER EXPERIENCE

Leveraging customer data for personalised recommendations and discounts, encouraging them to share their positive experiences. Constantly reminding customers of the benefits of shopping in our shops, reinforcing the reasons for engagement to increase trust, as well as using the newsletter as a cost-effective loyalty tool.

INSPIRE WITH A MISSION OR VISION

A well-defined mission or vision can serve as a powerful competitive differentiator. A compelling mission aligns a brand not only with a purpose beyond profit, but also resonates with customers, showing the brand's commitment to a positive global impact. This strategy could encourage customers to engage with the brand on multiple levels, from repeat purchases to referrals to participation in loyalty programs and brand communities.

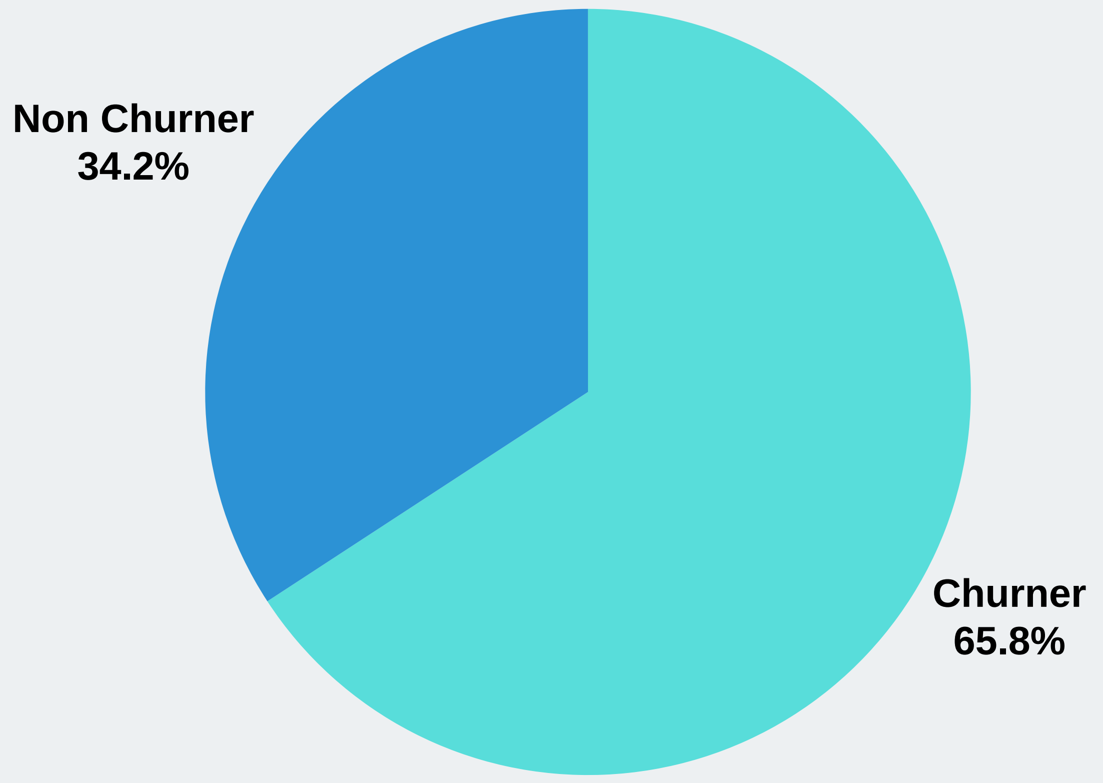
3 CHURN ANALYSIS

In the context of marketing analytics, churning refers to **the moment a client stop buying a company’s products or using its services**. The **goal** of churn analysis is to build, through data analysis, a model which will be capable of **predicting which customer is likely to churn**, and, from here, **develop a retention strategy** that can increment the value of our business.

Dataset preparation and churning customer definition

First, we merged the wo datasets: the one obtained through the preprocessing phase and the one containing the labelling of the customer base. We chose, in fact, to consider **only the active portion of our customer base**, defined through the repurchase curve analysis.

Second, we set a reference date, 14/01/2023, to define the **lookback period** and the **holdout period**, whose range is both 106 days, just as the repurchase interval, and imposed that a customer is a churner if he or she doesn’t make a purchase in both of these periods. If the customer is a churner than the result is 1, 0 otherwise.



3 CHURN ANALYSIS

Model	Accuracy	Precision	Recall	AUC
KNN	0.84	0.82	0.83	0.92
SVM	0.89	0.87	0.89	0.95
Random Forest	0.9	0.89	0.89	0.96
Logistic Regression	0.8	0.78	0.77	0.88
Gradient Boosting	0.92	0.9	0.93	0.97

Our Best Model

After some preprocessing operations (splitting the data between train and test sets, class balancing and value standardization), we trained the models and performed model evaluation to define the optimal choice: we found that the best model overall is the **Gradient Boosting**, which has a very high AUC, with a value of 0.97. This indicates that the classifier is **performing very well in distinguishing between the positive and negative classes.**

3 CHURN ANALYSIS

Looking at the final results and to sum up this churn analysis, we provided **three possible strategies** to prevent customer from abandoning our services:

FEEDBACK ANALYSIS AND BETTER SERVICES

It could be useful to collect and analyze feedbacks from customers that abandoned our services and **conduct a SWOT-like analysis** in order to understand the reason behind the abandoning and to find weaknesses in the offer: from here it could be possible to better our products and to create an overall better experience for our customers, in order to prevent churn.

OFFER CUSTOMIZATION AND TARGETED COMMUNICATION

Use advanced analytics to understand the behaviors of both churners and non churners in order to find trends and patterns and identify segments: from here **develop a customized offer** and **use targeted communication channels**, such as personalized email or in-app notifications, to provide relevant information and special offers to customers based on their past behaviors.

PROVIDING EXCELLENT CUSTOMER SERVICE

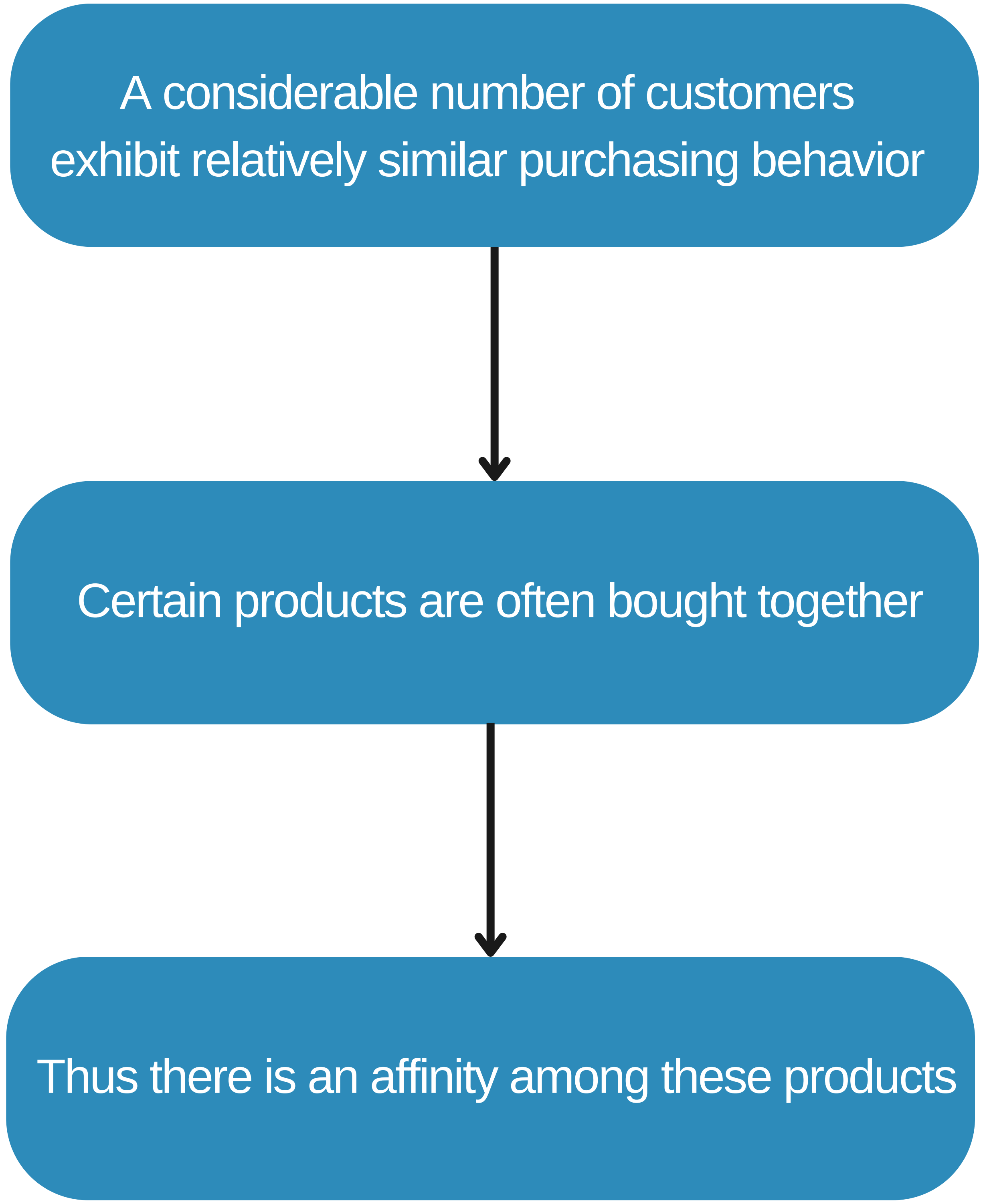
Use intelligent chatbots or virtual assistants to answer frequently asked questions and streamline the support process, or **implement loyalty programs** that reward loyal customers with discounts, gifts, or exclusive experiences. Customers are more likely to stay if they feel appreciated and rewarded for their loyalty.

4 MBA

Market Basket Analysis (MBA) is a model designed to identify, quantify, and measure **relationships between product purchases**. Specifically, it aims to find products that are frequently purchased together, based on customer transactions.

The mathematical and probabilistic measures used to quantify these affinities are as follows:

- **Support**, which measures how frequent the products are purchased with other specific products
- **Confidence**, which provides a measurement of the likelihood of a specific association between products
- **Lift**, which provides a measurement of the strength of these associations



A considerable number of customers exhibit relatively similar purchasing behavior

Certain products are often bought together

Thus there is an affinity among these products

4 MBA

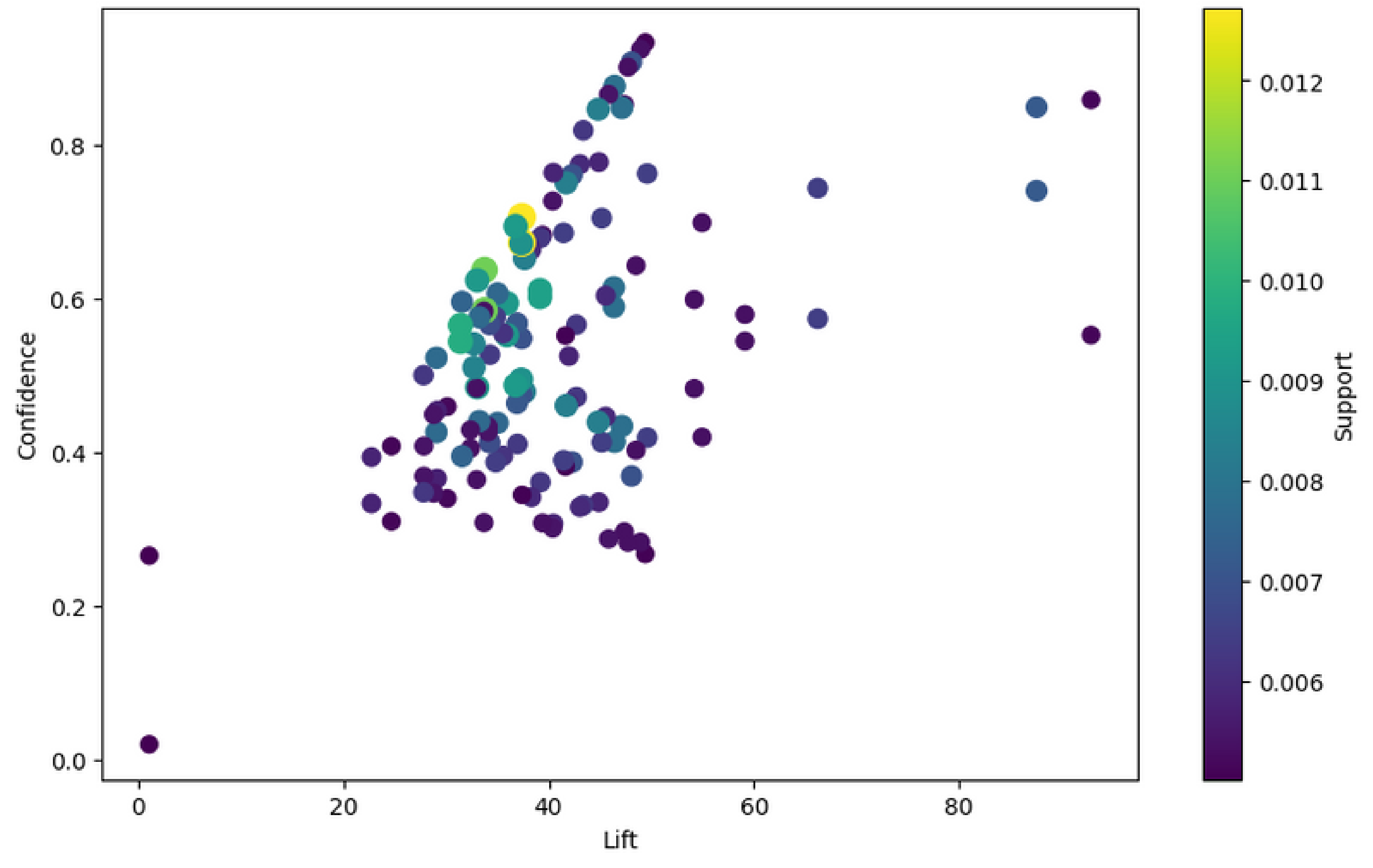
Top Rules by Lift

Antecedents	Consequents	Lift
36298206	36298122	92.85
36298381	36298353	87.55
31047485	31047464	66.22
32078795, 32079103, 32079082	32078935	59.13
32079082, 32078935, 32078795	32079103	59.13

Top Rules by Confidence

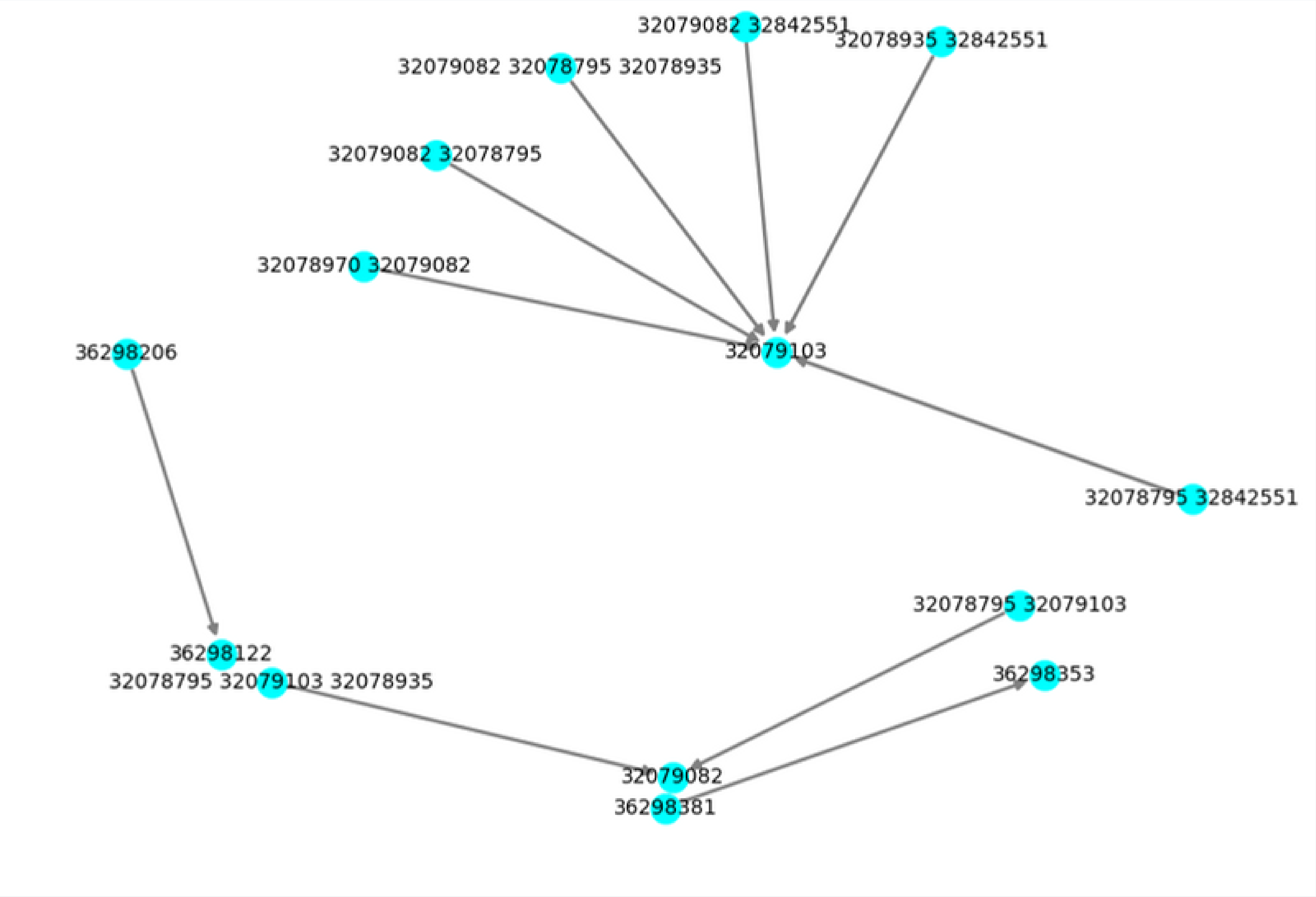
Antecedents	Consequents	Confidence
32078795, 32842551	32079103	0.933
32078935, 32842551	32079103	0.925
32079082, 32842551	32079103	0.909
32079082, 32078795, 32078935	32079103	0.902
32079082, 32078795	32079103	0.877

Correlation between lift and confidence



4 MBA

Top ten confidence network graph



The chart displays the top 10 pairs (antecedents, consequents) based on the **confidence** parameter. As observed, the product with ID 32079103 is frequently purchased following the acquisition of numerous other products, sometimes even in pairs. Additionally, there are instances of more linear relationships.

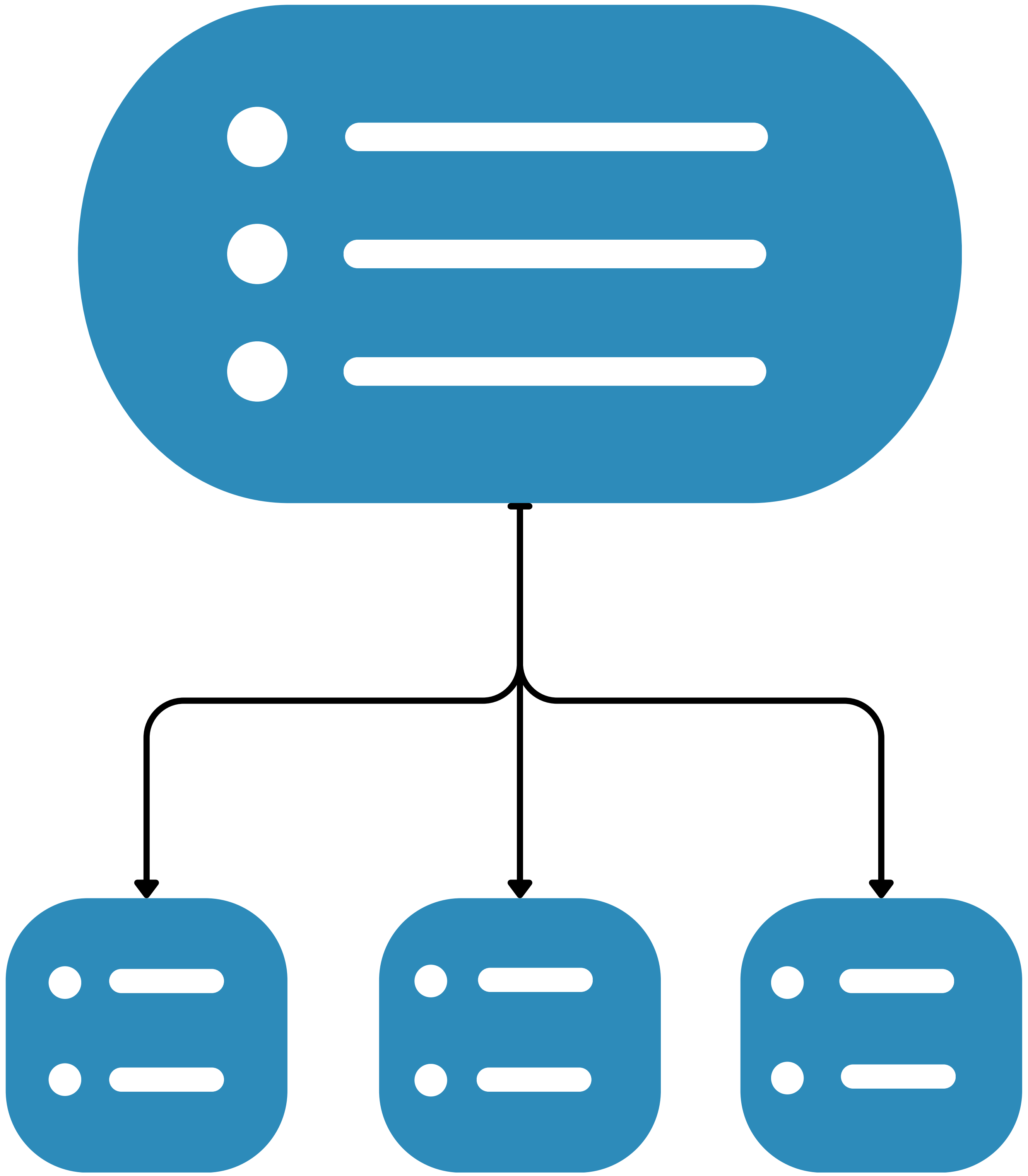
To implement an optimal strategy, understanding the nature of these products and the physical store layout is crucial. This knowledge would allow for revenue optimization, such as strategic placement of products on the shelves.

5 SENTIMENT ANALYSIS

Sentiment analysis is a process that uses **natural language processing** (NLP) and machine learning techniques to identify and **extract** the emotional tone, attitude or sentiment expressed in a given text or message. In this study, each review is classified as **positive**, **negative** or **neutral**.

First, however, we process the textual data by converting it to **lower case**, **removing URLs**, **tags**, **punctuation**, **numbers**, and common **English stopwords**.

We have opted for **BERT-Base-Uncased** as a pre-trained model. It is a natural language processing model that captures context from both the left and right side of words. It consists of **12 levels**, **768 hidden units**, **12 attention heads** and a total of **110 million parameters**.



5 SENTIMENT ANALYSIS

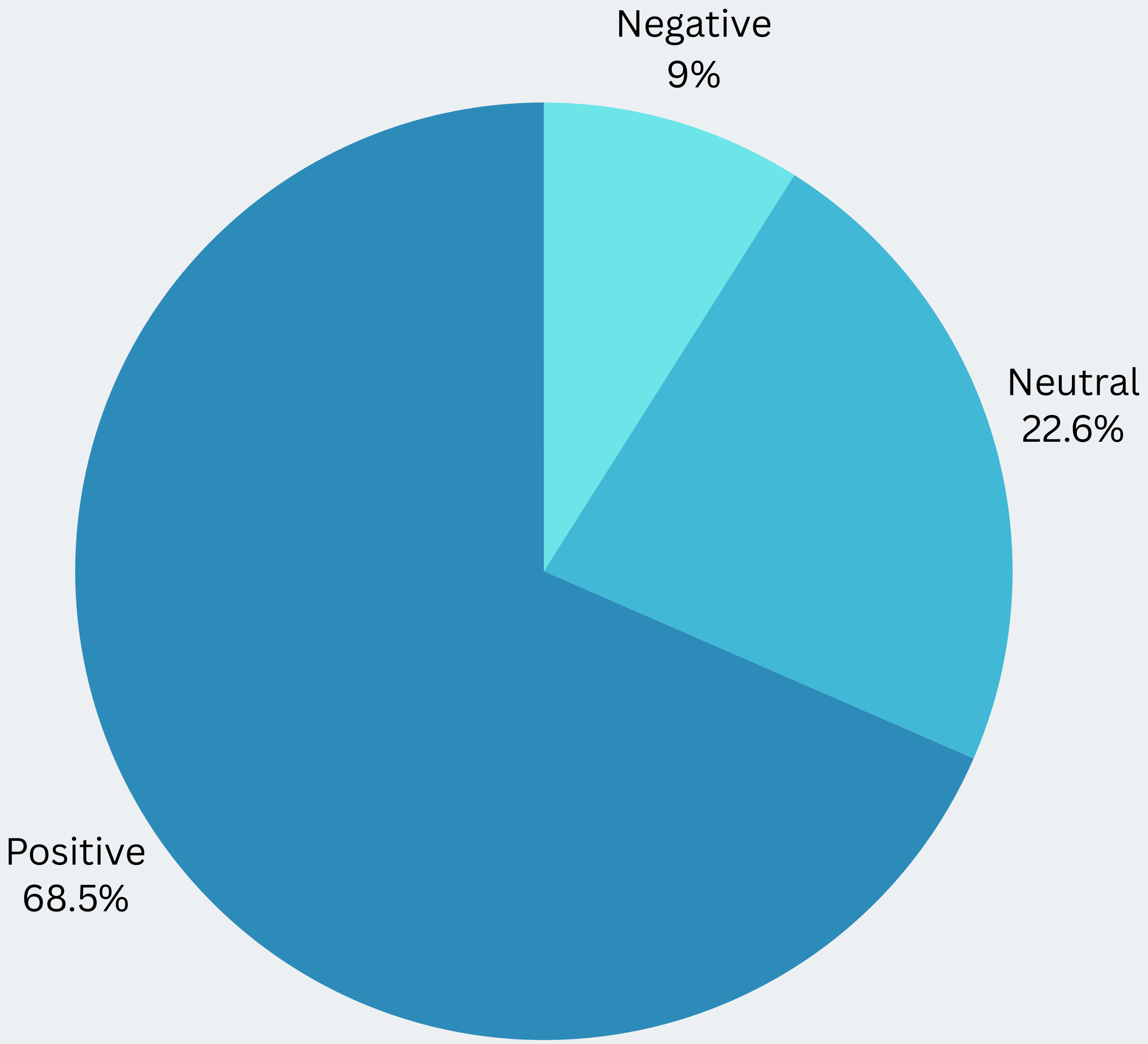
Our sentiment analysis using BERT-Base-Uncased achieved impressive results:

- **Test Accuracy: 0.86**, which indicates that our model correctly classified reviews as positive, negative, or neutral with a high degree of accuracy.

- **Test Loss: 0.38**, which is a measure of how well the model performed during training. A lower loss indicates better model performance

Since the reviews are only **related** to the customer_id and not to the order or product purchased, it is **not possible** to investigate the features of the reviews further.

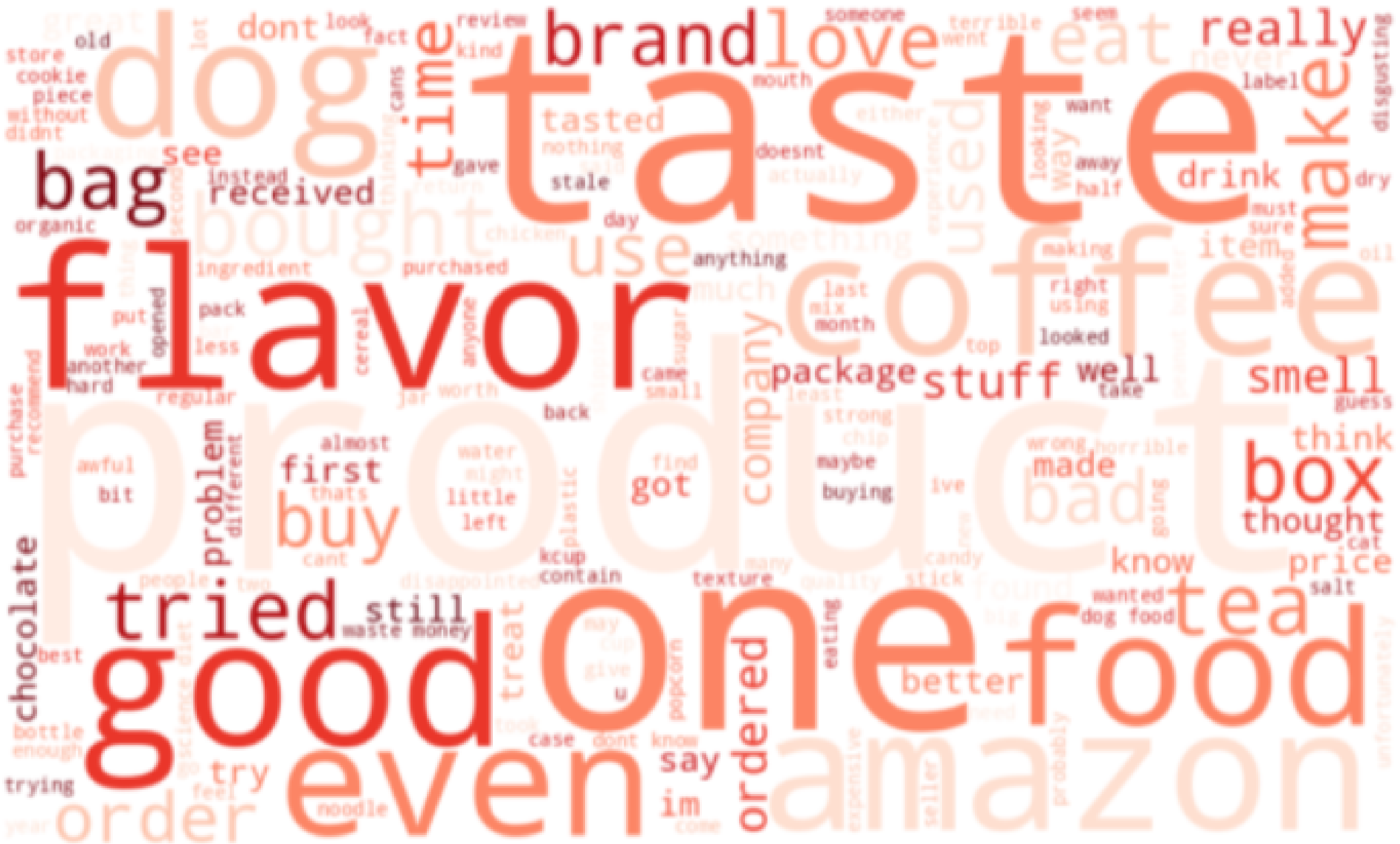
Distribution of Sentiment of Customer Reviews

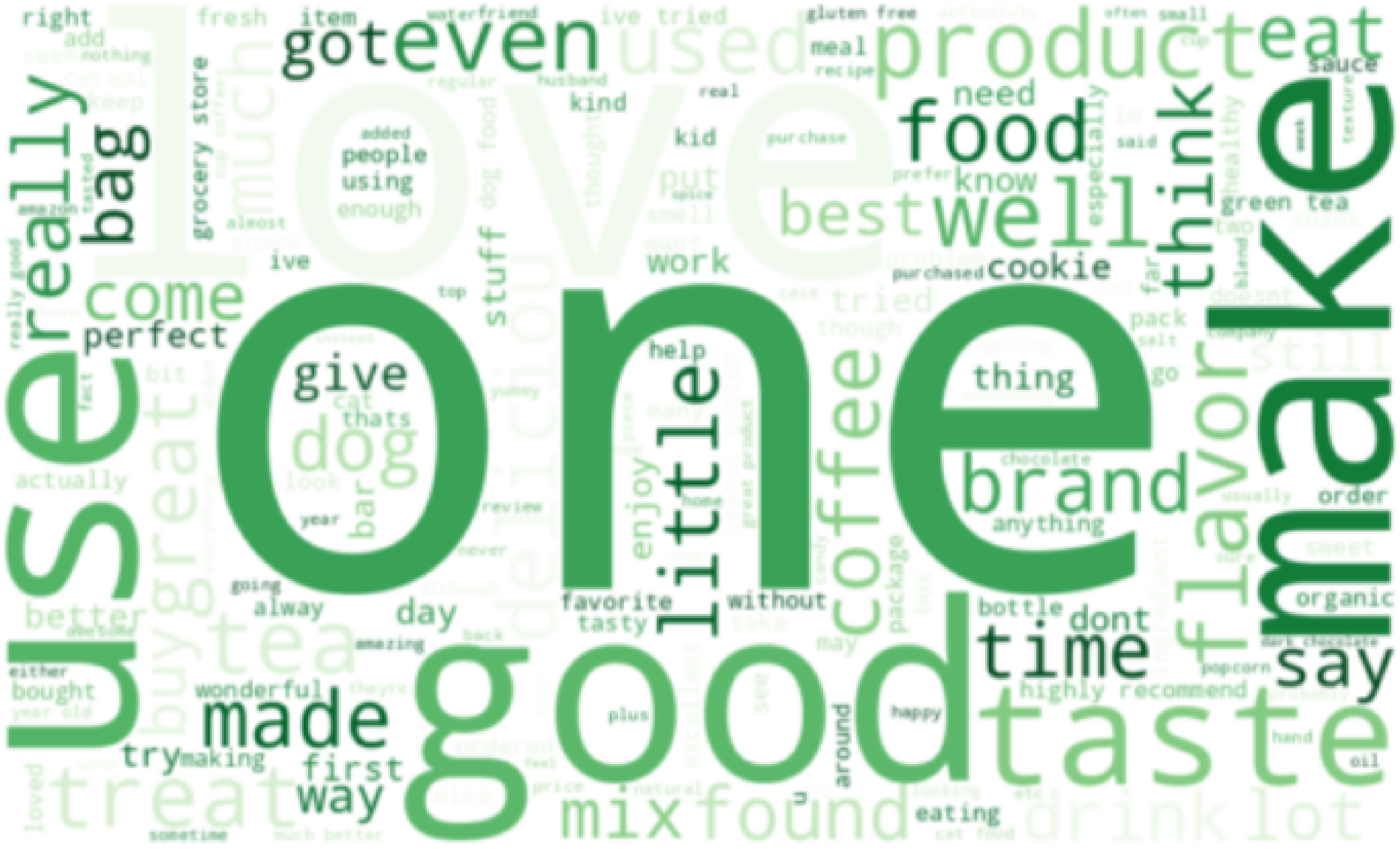


More than **75 %** of customers **actively** provided reviews. In particular, in the **18-55 age group**, females tend to write the most reviews, while after the **age of 66**, the trend reverses in favour of males.

5 SENTIMENT ANALYSIS

We tried to examine the reviews to find out the **frequency** of words used. We then created a **word cloud image** for negative and positive reviews, as follows:





Looking at both aspects, it is **challenging** to make recommendations without considering the broader context and other useful factors. However, it is crucial to recognise that **addressing negative reviews** is a key step in effectively identifying and understanding root cause and attempting to effectively reduce them.

6 EXTRA (1)

In order to get **further insights on the sentiment expressed** by the customers and refine the feedback-focused strategy, we performed sentiment analysis on another natural language model: **Doc2Vec**

Doc2Vec is an extension of Word2Vec, a popular natural language processing technique. Unlike Word2Vec, which learns vector representations for individual words, Doc2Vec extends this concept to learn representations for entire documents, allowing it to **capture semantic relationships** between documents in addition to words. It's a powerful tool for tasks like document similarity, sentiment analysis, and content recommendation.

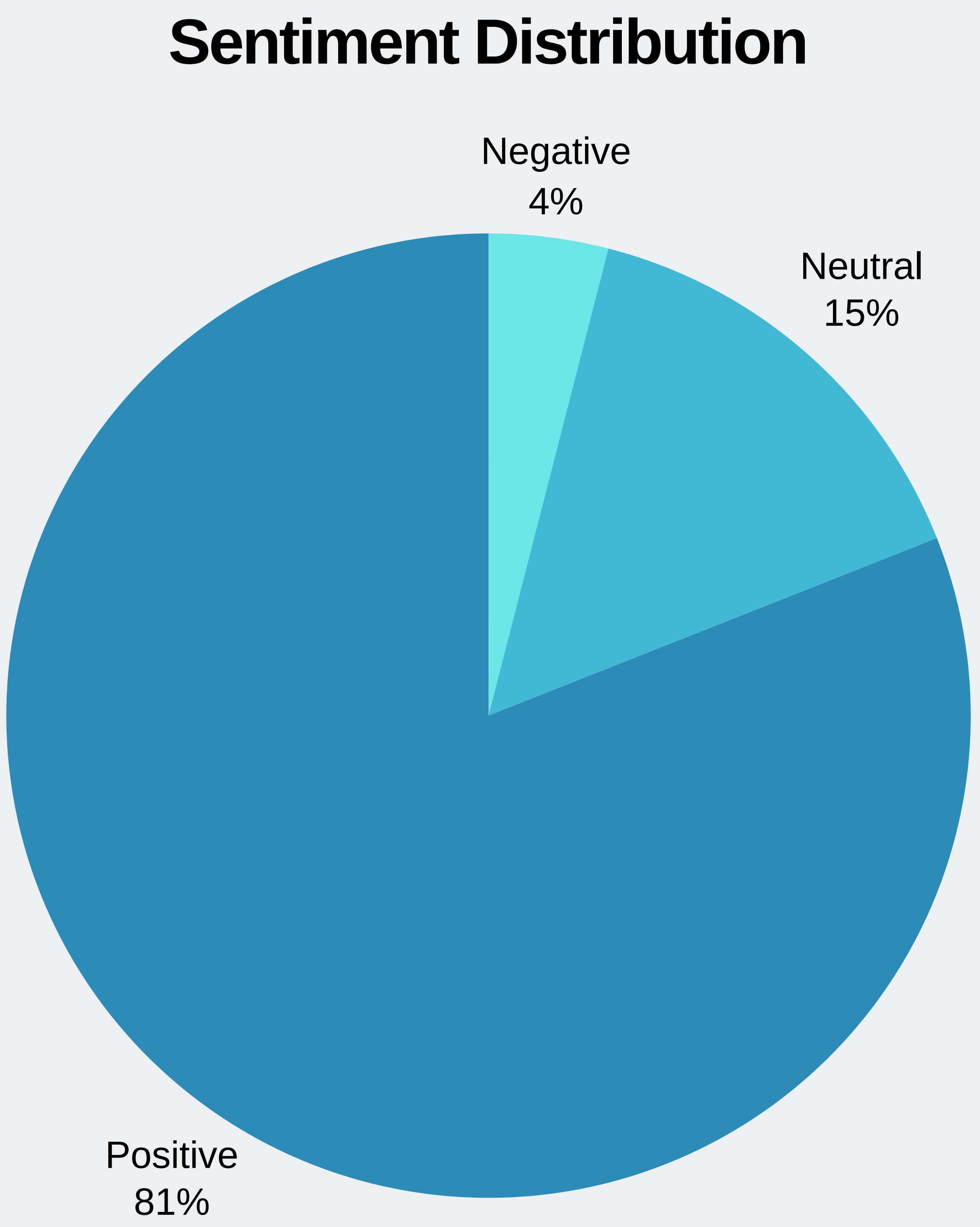
After some preprocessing operations (URLs, tags, punctuation, numbers and stop words removal), we performed tokenization of the text for the training phase and trained the model on the 'tbl_labelled_reviews' dataset. Then we made **predictions on the validation set with 70% accuracy** (using logistic regression) and finally used the trained model to make predictions on actual reviews from customers, using the 'tbl_customers_reviews' dataset.

6

EXTRA (1)

The results are the following: first, we obtained an **accuracy** score of **74%** on the **customer review table**, meaning that the model correctly predicts the actual sentiment label assigned to each customer review, with a success rate of 74%. Second we observed the count for each label: the model labels **4%** of the reviews as **negative**, **15%** as **neutral** and the last **81%** as **positive**.

A possible approach to manage the negative sentiment and improve the **feedback-focused strategy**, could be to **create a dedicate channel** to communicate with costumers after the purchase in order to engage in a constructive dialogue: first, responding promptly to the feedback to demonstrate attention and concern, then asking more details to better understand the problem and finally provide a plan to resolve the issue, offering a concrete solution to the costumer.



6 EXTRA (2)

With the aim of **leveraging the segmentation provided by RFM analysis in the context of market basket analysis**, we select only customers categorized as 'Diamond', the highest possible value, and focus on their purchasing trends.

Top frequent products purchased by diamond customers, with their consequent in MBA

Customer	Antecedent	Frequency	Consequent
111388	33700716	19	32079103
255502	33700716	16	32079103
526586	33700716	14	32079103
280767	33700716	14	32079103
530265	33700716	14	32079103

⚠ Since the consequent is obtained by MBA dataset, the pair (antecedent, consequent) is always the same and is not influenced by the customer

We decided to offer Diamond customers a **discount on items** that are generally **subsequent** to the items they purchase most frequently. However, the discount will not be the same for everyone and every product, but it will be **proportional to the purchase frequency**.
The purpose of the strategy is to encourage Diamond customers, who are already inclined to buy the preceding item, to **make an additional purchase on the subsequent item**. Obviously, **the discount will be temporary**: the belief is that once the customer acquires the habit of also buying the item subsequent to the preceding one, they will not give it up even if the discount no longer entices them.

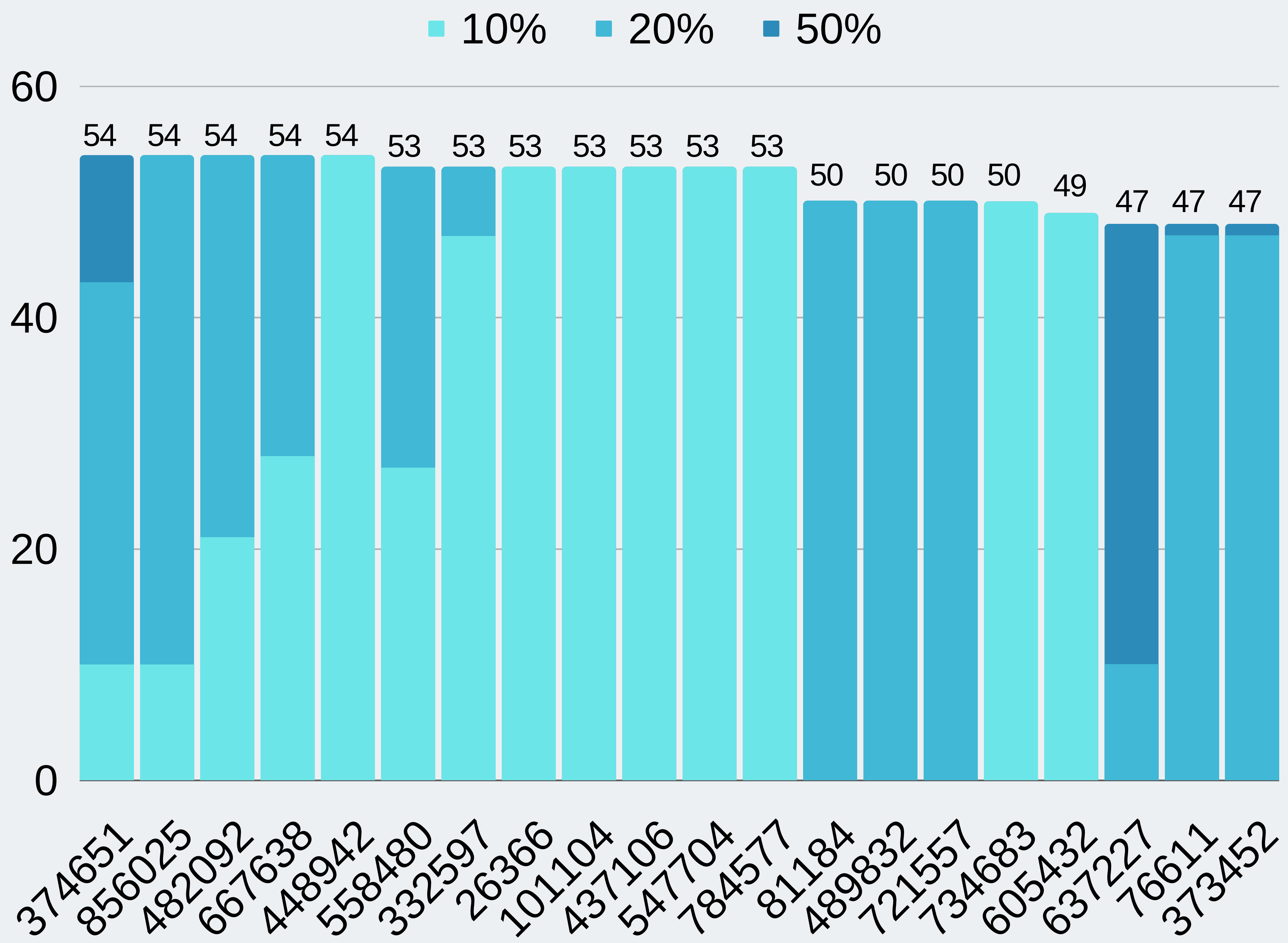
6 EXTRA (2)

We then define a discount that is proportional to the purchase frequency, specifically increasing as the frequency increases. By calculating the corresponding quantiles at 50% and 90% regarding the frequency, we offer the customer a **50%** discount on products resulting from items purchased at least 5 times, a **20%** discount if purchased 3 or 4 times, and a **10%** discount if the purchase occurred only 2 times.

Sample of the three different situations

Id	Antecedent	Freq	Consequent	Discount
507515	33700716	12	32079103	50%
798781	33700716	3	32079103	20%
51773	34129942	2	31618405	10%

Top 20 customers ordered by number of discounts, including the discount type highlighted



THANK YOU!