NOVA IMS Information Management School

PROJECT REPORT

CUSTOMER SEGMENTATION

2024

David Duarte 20221899
Marta Alves 20221890
Renato Bernardino 20221960

Index

1. Introduction

Humans naturally categorize things to enhance understanding and comprehension. Companies employ a similar approach with the goal of better understanding their customers and retaining them. In today's information age, this practice is crucial for companies to survive amidst increasing product offerings and intensifying competition.

To thrive, businesses must adopt marketing strategies that target specific customer segments without alienating others. By segmenting customers based on their preferences and needs, companies can focus on the most relevant aspects of their clientele, ensuring higher retention rates, maximized profits, and stronger customer loyalty.

In this context, our project focuses on customer segmentation to generate valuable business insights. Additionally, we developed targeted campaigns for the identified clusters, demonstrating the practical applications of clustering and how companies can leverage this approach to enhance their marketing strategies.

2. Methodology

In terms of methodology, our initial step involved thorough examination of the dataset to comprehend the available columns and understand the data's nature. Subsequently, we conducted feature transformations, followed by a meticulous check for inconsistencies such as missing values, which were promptly addressed. Once these preliminary steps were completed, we selected the most suitable scaler for our variables

Utilizing plots, we identified unidimensional outliers and subsequently pinpointed our initial cluster, determining it was optimal to isolate it from the rest of the data. Additionally, we removed multidimensional outliers, which were found to contain a distinct cluster. Our modeling efforts encompassed various techniques including DBSCAN, Mean-shift, KMeans, and Ward, implemented across two different approaches—one focused solely on economic variables, while the other incorporated both economic and behavioral variables.

Following model execution, we conducted a comprehensive comparison of results, culminating in the identification of ten distinct customer segments. Subsequently, we delved into each cluster, devising tailored marketing campaigns tailored to address the unique tastes and needs of each segment.

3. Exploratory Analysis

The dataset we received contains information on a company's customers, their spending habits, and some personal details. Our first step was to explore this data, examining the available features and extracting as much information as possible before applying our models.

3.1 Dataset Structure

Before doing any type of transformation, we started by looking at the general information about the dataset regarding customer information. Using the info() method we noticed the presence of missing values in some variables, and with describe() we were guaranteed with the presence of categorical features, as they dont have any statistical data. Moreover, some values were observed as weirdly high, such as the maximum value for lifetime_spend_fish, indicating the possibility of being highly requested by some customers.

3.2 Identification and Treatment of Data Inconsistencies

In this section, we did various checks and treatments to make sure the data is of good quality, including changing some variables, getting rid of ones that don't tell us much, and handling missing values by either imputing or removing them. These steps are important to keep the data accurate for future analysis and modeling.

3.2.1 Create and Transform Variables

After examining the dataset, we modified some variables. We created dummy variables for customer_gender, assigning a value of 1 if the customer is female and 0 otherwise. We also created a dummy variable for loyalty_card_number, where a value of 1 indicates the customer has a loyalty card and 0 indicates they do not. Missing values in the loyalty_card_number field were treated as 0. Furthermore, we created customer_birth_year, from the customer_birthdate.

We observed a significant number of variables representing lifetime spending on various products. To streamline our analysis, we opted to aggregate these variables to derive the overall total spent. To achieve this, we created a new column in our dataset, where missing values were handled by imputing the median of each respective column. It's important to note that this imputation process was performed solely on a copy of the dataset, ensuring the integrity of the original data remained intact.

3.2.2 Visualize

To get a clearer picture of our data and how customers are spread out across different geographic areas, we used data visualization techniques to create plots. The map allowed us to see where customers are located and gave us insights into how densely packed they are in certain areas and whether there are any clusters of customer locations.

The dataset included longitude and latitude columns, so we plotted these coordinates to check if there were any interesting patterns. The results were the following:

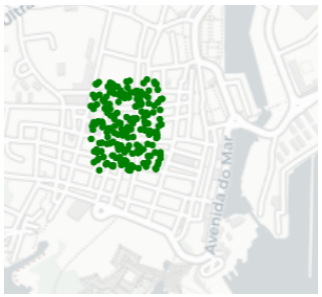


Fig.1

Fig.2

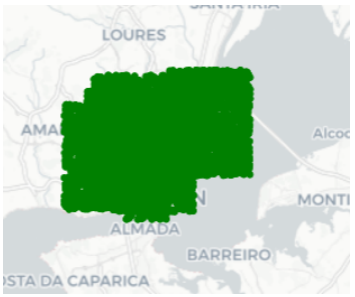


Fig.3

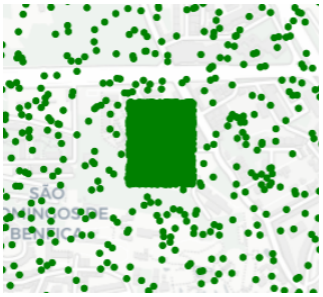


Fig.4

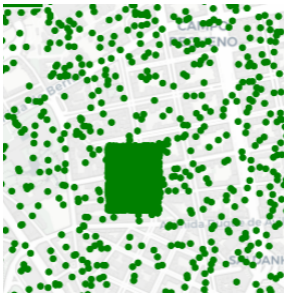


Fig.5

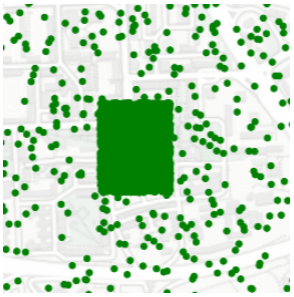
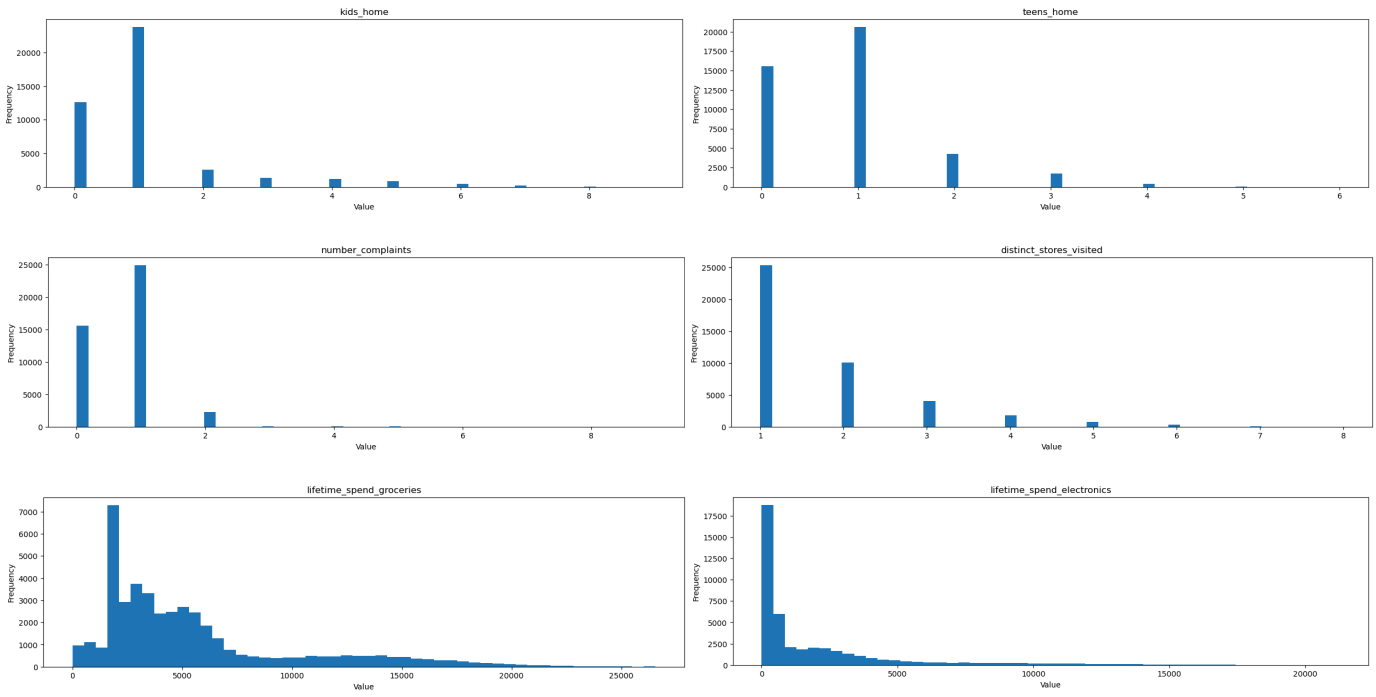


Fig.6

We obtained a map with 3 distinct groups: Fig.1, located at Peniche, Fig.2 located at Ericeira and Fig.3, representing customers spread across Lisbon and its surroundings. Fig. 4, 5, and 6 are an approximation of Fig.3, revealing that there's condensed groups of customers near São Domingos de Benfica, Campo Pequeno and Campo Grande, respectively.

Having this information, we passed to each variable distribution to get an initial insight about outliers. In this step, we plotted the histogram for all variables contained in customer_info_num, a dataset containing all the numeric information of the original dataset.



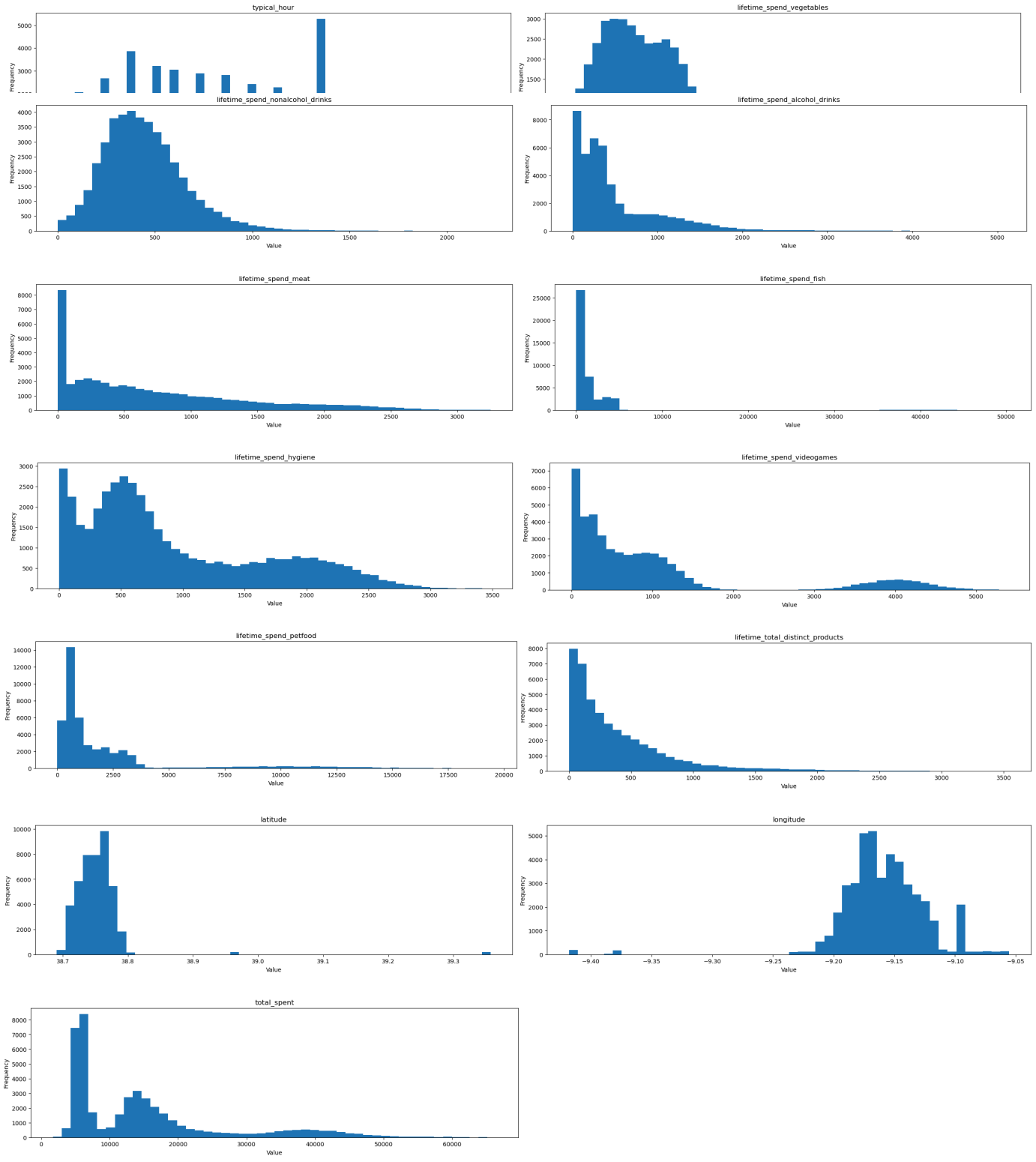
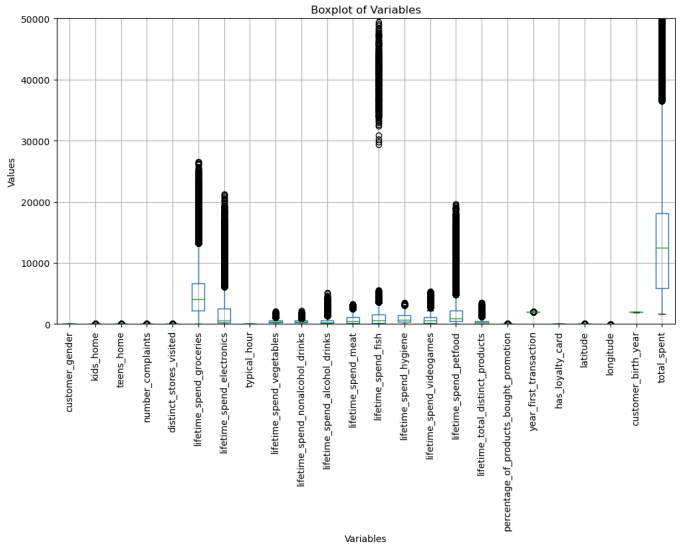


Fig.7

As we can see, all of the above variables contain outliers, contributing to the conclusion that these outliers are related to different clusters. For example, in latitude and longitude, we confirmed the existence of clusters related to location, as we speculated previously. Additionally, the majority of the features have a right-skewed distribution.

Besides plotting histograms, we also plotted a box-plot, to get a better statistical overview of how the data is distributed, obtaining the following plot:



Again, this reveals the presence of severe outliers in some variables, like lifetime_spend_fish.

Fig.8

3.2.3 Missing Values

Continuing from our previous discussion on missing values, this section addresses their treatment. We used feature selection techniques and methods like the holdout method and cross-validation to find the best imputation strategies.

By rigorously testing different imputation methods, we aimed to identify the most effective techniques for handling missing data.

After seeing how many missing values we had per category, we checked the correlation between variables through a heatmap, guiding effective imputation and further analysis.

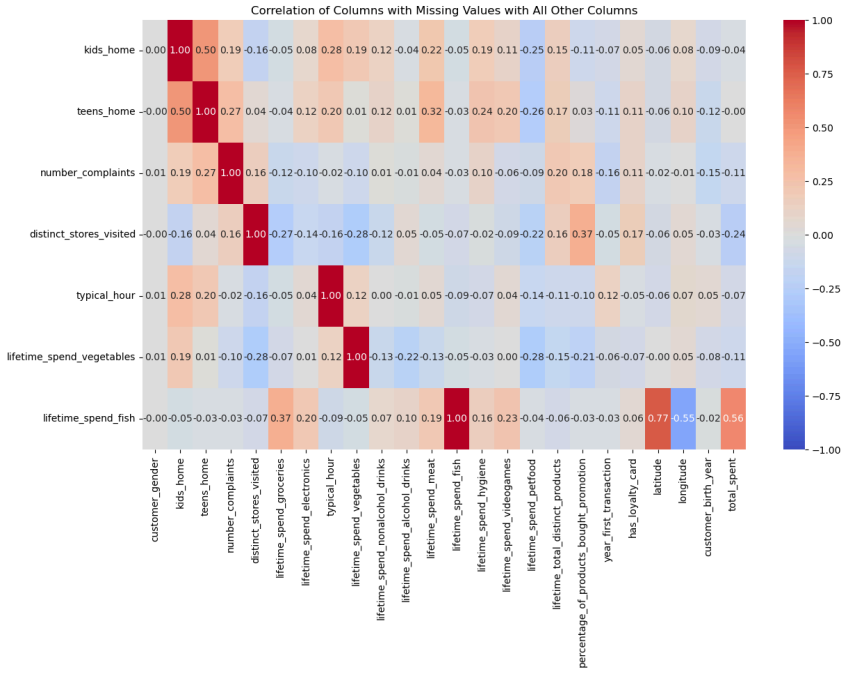


Fig.9

One of the most noticeable correlations happens between lifetime_spend_fish and latitude, having a strong positive correlation (0.77), suggesting that customers who spend more on fish tend to be located in higher latitude and longitude places like Peniche and Ericeira.

Following this, we scaled the data not only to use the techniques that we implemented, since they work with distances, but also to get better scores.

Given the outliers in our dataset, we chose RobustScaler. This method effectively handles outliers, minimizing their impact on our analyses and modeling. Using RobustScaler ensures that our results remain robust and reliable.

After applying the scaler, we separated the data with missing values from the data without missing values. We then created a target dataframe, with all the variables with missing values, and used Lasso to identify which features are most important for each target variable. We opted to eliminate variables with coefficients between -0.1 and 0.1. On the right side is an example of the output, which represents the algorithm for the variable teens_home, indicating that we should remove 12 variables like, for example, customer_gender, distinct_stores_visited, lifetime_spend_groceries, lifetime_spend_electronics, etc.

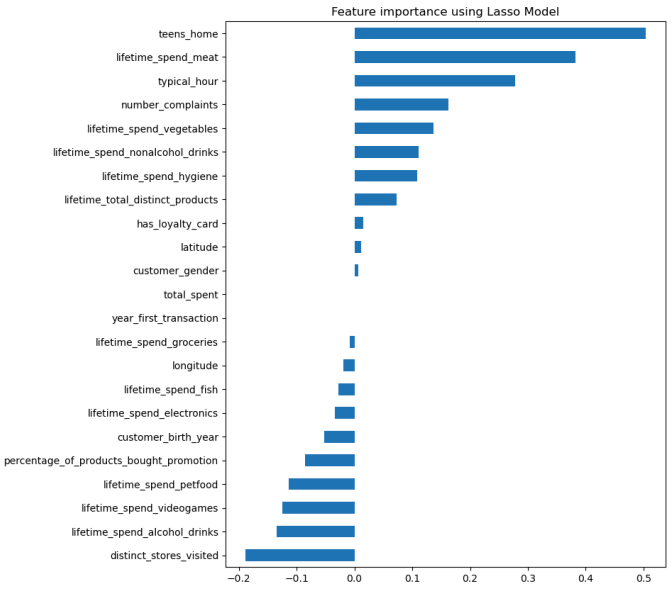


Fig.10

Afterward, we created a dictionary where the keys are the target variables and the values are the columns to be removed. The decision to eliminate these variables was based on Lasso regression results and correlation analysis.

To test KNN and random forest, we created three functions:

1. **Splits**: This function splits the dataset into training and validation sets, while removing the eliminated variables and the target variable.
2. **Mean_absolute_percentage_error**: This function calculates the mean absolute percentage error (MAPE) between true and predicted values.
3. **scoring_function**: This function creates a custom scorer that favors lower values of the error metric.

We tested KNN with different numbers of neighbors, but realized that the difference in score wasn't significant.

For random forest we used optuna because of the enomerous hyperparameters combination possible, but due to the extended runtime, we limited our demonstration to 3

trials, focusing on the most critical hyperparameters. The scores from the full run (with more hyperparameters and trials) and the demonstration run varied by only 5% to 20% from the demonstration scores.

Given the underwhelming performance of both KNN and RF in our scoring metrics (score higher than 30%), we've opted to employ median imputation to preserve the integrity and structure of our dataset. This method ensures that missing values are filled with a central tendency measure, thereby maintaining the distribution and characteristics of the original data.

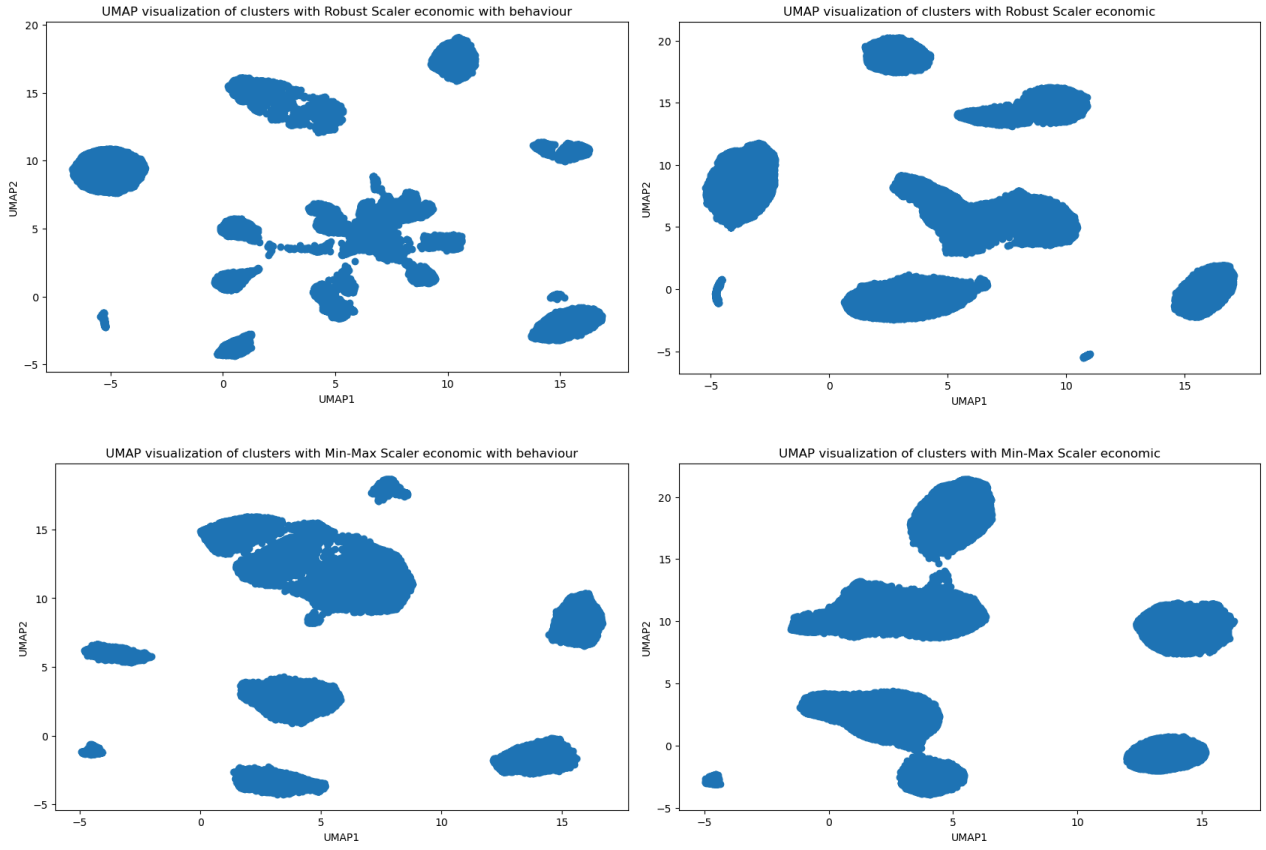
4. Customer segmentation

After treating the dataset, we moved on to segmenting our customers. We rigorously tested different models to find clear groups among our customers. This helped us better understand our customer segments, so we can tailor strategies and approaches specifically to them.

4.1 Scaler for Clustering

As most clustering models work based on distance, we needed to scale the data. To determine the most suitable scaler, we conducted tests using various scaling techniques and visualized their distributions with UMAP plots. After thorough experimentation, we kept only the scaler examples that most accurately reflected the clustering distribution observed across all our tests.

Fig.11



Economic variables: all lifetime_spend variables and total_spend

Economic with behavior: all variables except location and binary variables

With the previous maps, we concluded:

1. There are small clusters that need the behavioral variables to distinguish themselves from the big clusters;
2. It is important that clusters are clearly separated from each other;
3. The number of big clusters always stays the same, which can indicate that we only need business information to identify them.

So, we decided to use the Robust Scaler, as it already has properties that help with outliers. By examining the two UMAPs, we can infer that with business and behavioral variables, there are small clusters that grow separated, and we think they are worth studying. The Robust Scaler, even with only business variables, creates strong and distinct clusters. Therefore, we plan to mix these two approaches.

From this point, we analyzed clusters using the Robust Scaler with both business and behavioral variables, and identified the clusters that can only be obtained with this combination. Furthermore, we performed clustering using only the business variables to identify the big clusters. The last step was creating segmentations based on the behavioral profiles of the customers, allowing us to distinguish certain traits even among individuals with similar business traits, helping with a better customer marketing approach.

4.2 Unidimensional Outliers

As we begin our initial cluster definition, it's important to check for unidimensional outliers, because detecting these early on is vital as they can influence how clusters form and what we take from them. By spotting and handling these outliers at this stage, we sat ourselves up for a more accurate and reliable cluster analysis later on.

We initiated the process by plotting the variables from the dataset business_behaviour_rb. This dataset is a copy of robust_scaller1, which includes scaled numeric variables, excluding binary ones, latitude, and longitude.

The results showed that some variables contained outliers, as concluded before, but lifetime_spend_fish had an obvious group of observations that stood very far apart from the rest (Fig.12), leading to the decision of using thresholding to create the first cluster.

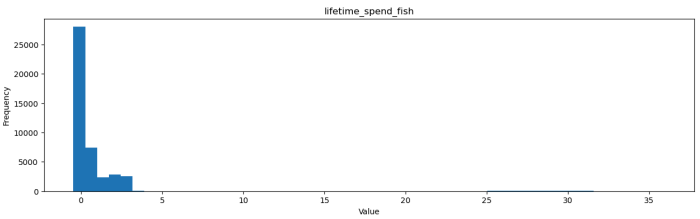


Fig.12

5. Multidimensional Outliers

Before continuing, it was essential to remove this type of outliers to guarantee the efficacy of our clustering techniques.

5.1 DBSCAN

As we learned in class, DBSCAN is a good algorithm to handle multidimensional outliers, as it marks low-density regions as outliers.

We began by identifying the optimal value for the epsilon parameter in DBSCAN using the k-NN distance plot. Epsilon determines the maximum distance between two points for them to be considered neighbors.

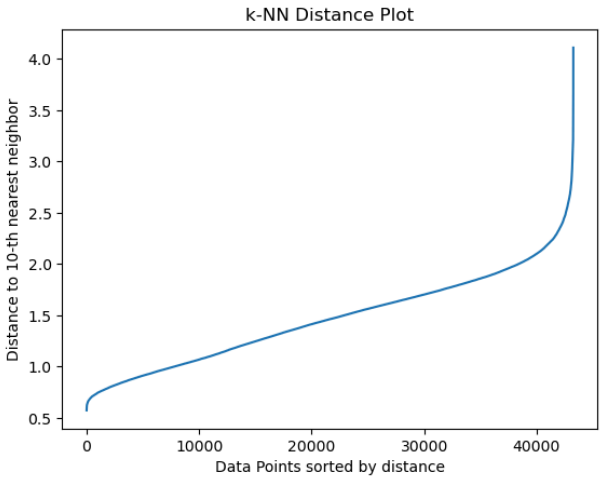


Fig.13

By analyzing the plot, we chose 2.25, since it's the "elbow" point, where the rate of distance change slows down significantly. This ensures that DBSCAN identifies clusters based on the data's density while also spotting outliers effectively.

Having the Epsilon defied, we ran the algorithm on business_behaviour_rb, removing in the end 480 outliers.

5.2 Analyzing Outliers

To proceed, we opted to compare the 'original' dataset, robust_scaler1, with the outlier dataset. We accomplished this by generating histograms for each variable using a custom function. These histograms were overlaid to facilitate a direct comparison of their distributions, enabling us to draw meaningful conclusions.

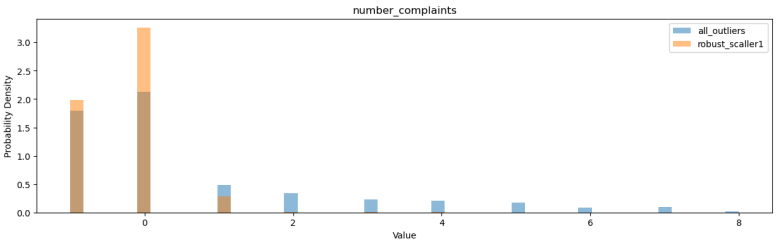


Fig.14

Number_complaints standed out as it presented way more outliers when compared to the 'original' dataset. This information allowed us to come to the conclusion that this was also a cluster, a group of people who like to complain a lot, or according to internet slang, 'Karens'.

Having this, we utilized K-means clustering to explore and uncover more intriguing behavioral patterns about these people. To determine the potential number of clusters among the outliers, we created the following plots:

Fig.15

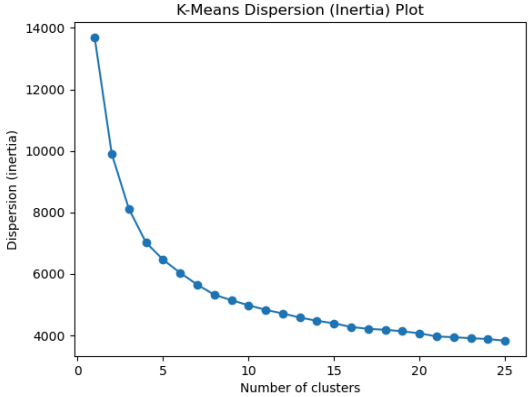


Fig.16

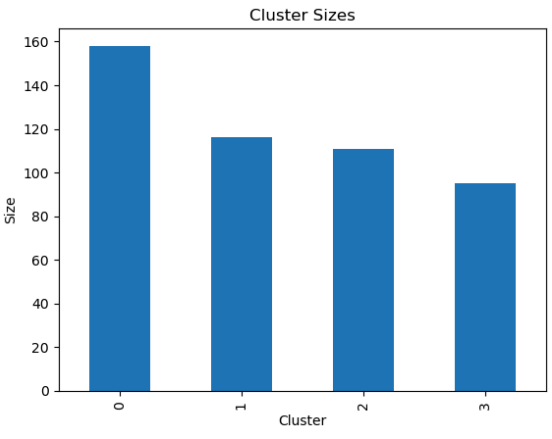


Fig.15 suggests that 3 to 5 clusters might be the most appropriate choice for this dataset, balancing the reduction in inertia with the simplicity of the model, and we ended up choosing 4 clusters whereas Fig.16 indicates that the outliers can be grouped into four distinct clusters with varying sizes, with one cluster being particularly dominant, the 'Karens'.

Our next step was to analyze their behavior in comparison to the rest of the customers. After plotting the four clusters, we found that only the 'Karens' cluster, characterized by a high number of complaints, stood out from the others. This distinction presented an opportunity to study this specific behavior further.

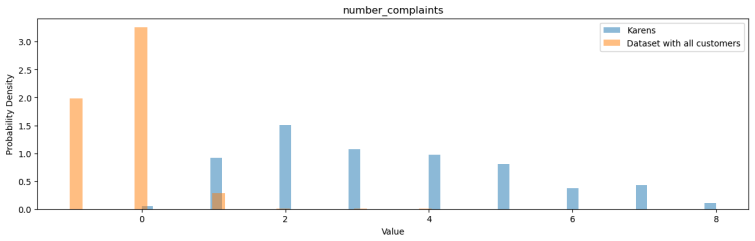


Fig.17

6.Clustering algorithms

To gain a deeper understanding of our remaining data and uncover additional hidden patterns, we employed various clustering algorithms to identify more clusters.

6.1 K-Means

The first clustering technique we implemented was K-Means. Using the elbow method, we determined that 7 clusters were optimal. We then manually adjusted the labels temporarily to explore and compare different techniques.

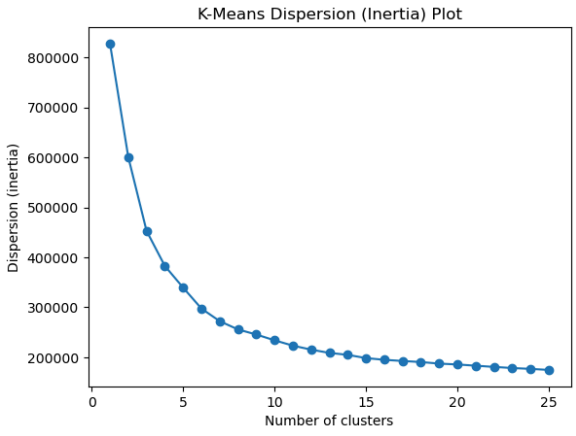


Fig.18

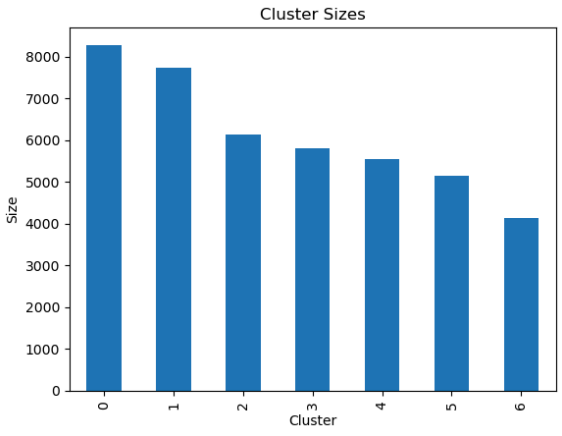


Fig.19

Another important information was the silhouette score, a statistical measure used for assessing clustering effectiveness, and served as a decisive factor in resolving ambiguity during the cluster definition process. Unfortunately, in this case, the score, 0.30581169981248885, indicated a low clustering performance.

6.2 Mean-Shift

One challenge we faced with Mean-Shift clustering was finding the right bandwidth. Despite trying an estimator, it didn't work out. After a few tries, we settled on a bandwidth of 2. But, with this bandwidth, not all points got assigned to a cluster. So, to make comparison easier with other methods like the Confusion Matrix, we grouped clusters with fewer than 500 clients into one cluster.

```
0      7833
1      7410
4      5654
5      4808
6      4513
2      4140
111    3900
3      2772
7      1037
18      705
Name: Mean_shift, dtype: int64
```

Fig.20

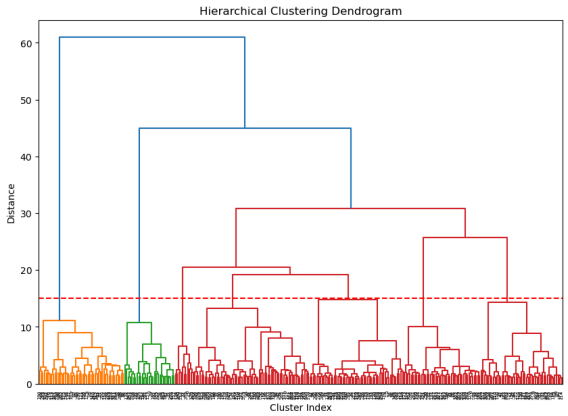
As it is demonstrated in the image beside, we obtained the labels of clusters, on the left side, and how many customers are distributed by them.

We reinforce that cluster 111 contains all clients that were not assigned to a cluster.

The silhouette score was once more evaluated, retrieving also a low result of 0.26434197142900107.

6.3 Hierarchical Clustering

The dendrogram plays a crucial role in this method, serving as a tree-like visualization of the algorithm, and provides insight into determining the optimal number of clusters.



We opted to utilize the Ward method for hierarchical clustering after comparing histograms generated by different methods. While these histograms exhibited remarkably similar distributions, the decisive factor in favor of the Ward method was its superior silhouette score. This score signifies more distinct and cohesive clusters, reinforcing the suitability of the Ward method for our analysis.

Utilizing the Ward Linkage method, we achieved optimal outcomes. Figure 21 displays the dendrogram generated through the Agglomerative method, employing Ward Linkage. Based on this visualization, we opted to implement 8 clusters. With a Silhouette Score: 0.2863458263340555

6.4 Comparison of Algorithms

To compare the performances of the previous algorithms, we performed two Confusion Matrices between K-means clusters and the Mean Shift ones, and Hierarchical clusters with Mean Shift ones, obtaining the following results:

	Mean_shift 0 Cluster	Mean_shift 1 Cluster	Mean_shift 2 Cluster	Mean_shift 3 Cluster	Mean_shift 4 Cluster	Mean_shift 5 Cluster	Mean_shift 6 Cluster	Mean_shift 7 Cluster	Mean_shift 8 Cluster	Mean_shift 9 Cluster
K-means 0 Cluster	7670	38	1	2	66	0	0	0	0	500
K-means 1 Cluster	0	7052	0	1	12	2	0	0	0	671
K-means 2 Cluster	163	320	8	2745	520	56	0	1037	0	1275
K-means 3 Cluster	0	0	0	0	4925	0	0	0	705	176
K-means 4 Cluster	0	0	0	0	0	0	4513	0	0	1034
K-means 5 Cluster	0	0	0	24	131	4750	0	0	0	244
K-means 6 Cluster	0	0	4131	0	0	0	0	0	0	0
K-means 7 Cluster	0	0	0	0	0	0	0	0	0	0
K-means 8 Cluster	0	0	0	0	0	0	0	0	0	0
K-means 9 Cluster	0	0	0	0	0	0	0	0	0	0

	Mean_shift 0 Cluster	Mean_shift 1 Cluster	Mean_shift 2 Cluster	Mean_shift 3 Cluster	Mean_shift 4 Cluster	Mean_shift 5 Cluster	Mean_shift 6 Cluster	Mean_shift 7 Cluster	Mean_shift 8 Cluster	Mean_shift 9 Cluster
hierarchical 0 Cluster	0	0	0	0	0	0	0	0	0	0
hierarchical 1 Cluster	0	0	0	0	0	0	4513	0	0	1034
hierarchical 2 Cluster	47	4	4140	47	0	0	0	0	0	10
hierarchical 3 Cluster	0	0	0	8	0	0	0	998	0	54
hierarchical 4 Cluster	369	7335	0	68	28	14	0	5	0	1857
hierarchical 5 Cluster	7386	63	0	2361	3	0	0	34	0	474
hierarchical 6 Cluster	0	0	0	66	5	4733	0	0	0	127
hierarchical 7 Cluster	31	8	0	222	5618	61	0	0	705	344
hierarchical 8 Cluster	0	0	0	0	0	0	0	0	0	0
hierarchical 9 Cluster	0	0	0	0	0	0	0	0	0	0

Fig.22

After analyzing the confusion matrix, histograms, and silhouette scores, we pinpointed two distinct clusters that emerge only when considering both economic and behavioral variables. These clusters were:

1. **University Students (uni_students):** Comprising predominantly young customers, this cluster exhibits a notable preference for alcohol and complementary items. Moreover, these clients tend to make purchases later in the day;
2. **Early loyal birds (Swans)/ Higiene:** This cluster includes longtime customers who visit the shop early in the day and purchase a substantial amount of distinct products, mostly hygiene and groceries products. Swans are a bird and often seen as symbols of love and fidelity, that were the reasons for the name.

Therefore, we removed those clusters from the dataset.

6.5 Separate Economic and Behavior Perspectives

On this stage, we focused on creating behavioral segmentation patterns for our clients. Following this, we conducted a segmentation analysis centered on the types of products clients purchase, which we referred to as economic segmentation.

For this purpose, we utilized behaviour_rb, leveraging robust_scaller1, which includes all clients, even those previously removed, for conducting a comprehensive behavioral segmentation of all customers. This approach facilitates later comparisons with clusters. Meanwhile, business_rb utilizes business_behaviour_rb, enabling economic segmentation of the remaining clients without the influence of behavioral characteristics.

6.5.1 Behavior_rb

Initially, we opted for 5 clusters using the elbow method. We then assessed the silhouette score and generated corresponding histograms, resulting in the subsequent findings:

- Cluster 0 - Base_customer (normal customers)
- Cluster 1- New Clients (young and recent clients)
- Cluster 2- Promotion Seekers (always look for promotions and even go to different stores if needed)
- Cluster 3- Big Families (have a lot of kids and teens)
- Cluster 4- Distinct / Curious (buy a lot of different products)

6.5.2 Business_rb

We applied the previous approach (KMeans, Mean-shift, Ward, Analysis) for business_rb and after careful consideration and numerous attempts, we finalized these clusters:

1. **Pet_lovers:** This cluster reflects a type of customer that uses our retail mainly to buy daily needs both for them and for their pets, showing a high consumption in categories such as groceries and pet food.
2. **Best_Customers:** The top spenders in our retail, show and high spending in all categories being characterized by their total amount spent.
3. **Vegan:** This cluster consists of clients which have a large amount of purchases in vegetables and tend to buy close to no meat. However, there are some purchases in fish.
4. **Low_Budget:** This cluster provides better insights about our cheap customers, later will find out they tend to buy most of their products in promotion
5. **Normal_customer:**This cluster shows the typical behavior of our average client staying within the mean in all categories

Despite the good silhouette score of 0.6245 the ward clustering was separated into three distinct clusters (vegan, low_budget, normal_customer). This decision was based on the clear differences observed in the histograms, indicating distinct segments.

6.6 Merging Datasets

To join all of the clusters (not the behavior), we developed two functions that merge the datasets effectively: merge_clusters (merges all the clusters of the same clustering method), combine_clusters (merges 2 clustering methods together).

6.7 Assign other Missing Values to Clusters

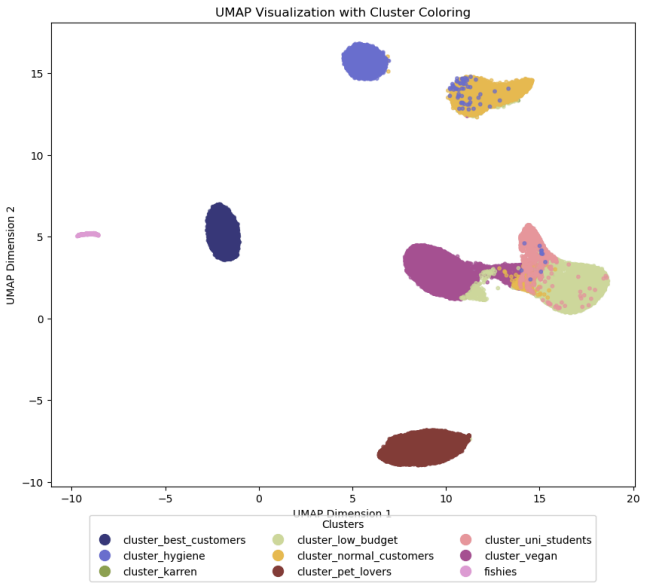
We have settled on an approach of assigning non-clustered customers (outliers) to the nearest cluster by calculating Euclidean distances to cluster means. This method offers several advantages, including consistency and simplicity.

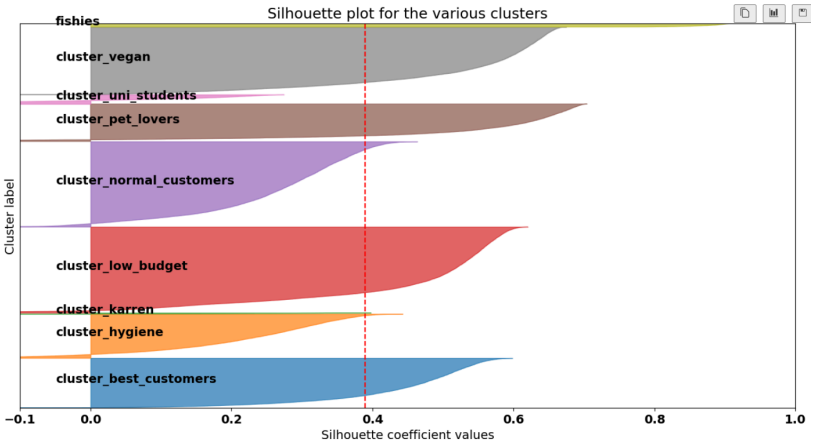
The last step was gathering all information into a common data frame.

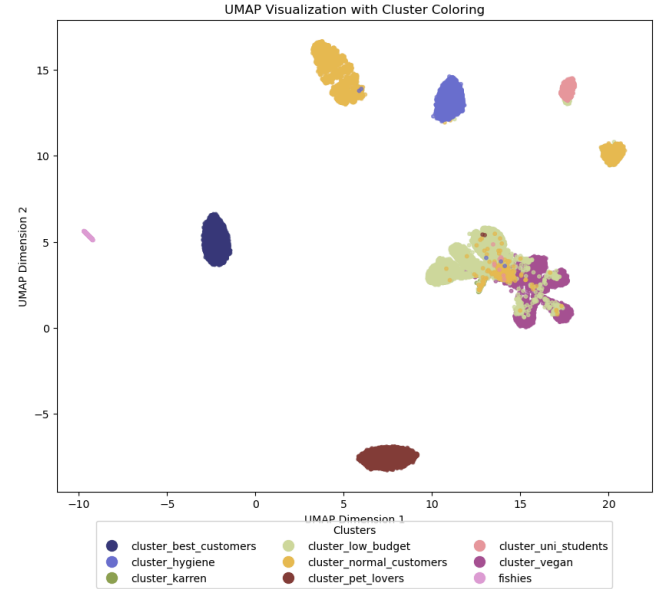
6.8 Visualizations

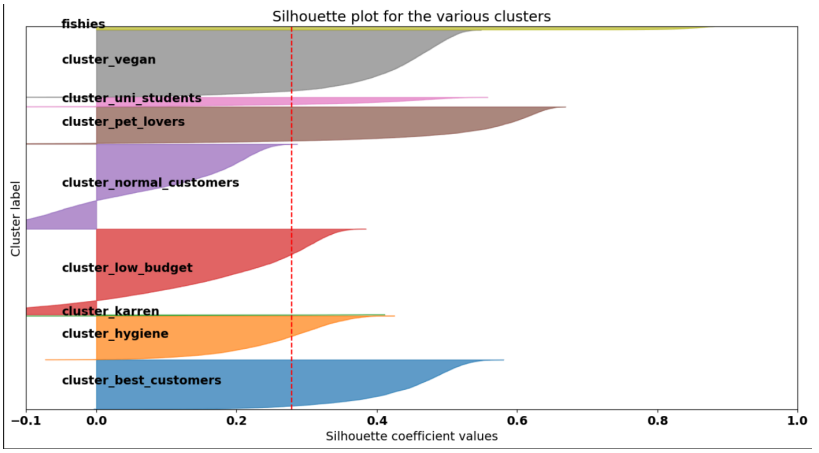
Having all the necessary information, we passed on to visualization, to get a better perspective of the calculated clusters. For this purpose, we chose UMAP to plot business-only and business with behavior.

Fig.23









It's evident that our clusters exhibit diverse characteristics:

- **Very Strong Clusters:** Examples include pet_lovers, best_customers, and fishies, which remain consistent across different approaches, showcasing robust patterns.
- **Almost Perfect Clusters:** For instance, cluster Swan/Higiene demonstrates minor variations in the business approach but emerges as a strong cluster in the other approach.
- **One-sided Clusters:** Uni_students exemplifies this category, as it only appears distinctly in the behavioral with economic approach.
- **Noisy Clusters:** Normal_customers can be considered a noisy cluster, as it appears almost perfect in the business approach but exhibits some noise, merging with another cluster, in the behavioral approach.
- **Messy Clusters:** Clusters like vegan and low_budget, although displaying differences in histograms, they do poorly in the UMAP visualization. Despite this, we retain them as they represent distinct behaviors.
- **Almost Invisible Clusters:** Represented by karren, these clusters don't form distinct clusters but still represent a discernible behavior within the data.

7. Analyzing the clusters with behavior profile

In this segment we try to understand a little more about the behavior of the customers as in a group. So we will do a confusion matrix of our 9 clusters and the customers profiles.

Fig.24

	cluster_best_customers	cluster_hygiene	\
regular	4892	1208	
new_clients	718	262	
promotions	32	181	
Big_family	0	4	
distinct products	0	3351	

	cluster_pet_lovers	cluster_uni_students	\
regular	3446	6	
new_clients	802	1063	
promotions	0	2	
Big_family	0	0	
distinct products	0	2	

	cluster_low_budget	cluster_normal_customers
regular	2660	1961
new_clients	762	2243
promotions	5986	123
Big_family	15	4699
distinct products	353	669

Fig.23

	cluster_vegan	fishies	cluster_karren
regular	6245	283	13
new_clients	1201	78	0
promotions	157	18	44
Big_family	0	0	6
distinct products	92	0	54

Best Customers: Typically regular customers, but there are 718 young clients within this group. This subcluster is particularly interesting and should be targeted in a marketing campaign to ensure their continued loyalty.

Swans/Hygiene: Customers in this cluster tend to purchase a variety of products.

Pet Lovers: Generally regular customers with consistent purchasing patterns.

University Students: As previously analyzed, these are young customers

who exhibit distinct shopping behaviors.

Low Budget: Primarily composed of customers who frequently buy items on promotion, as anticipated.

Normal Customers: Often large families with substantial purchasing needs.

Vegan: Usually regular customers with consistent buying habits.

Fishies: Regular customers, likely tied to coastal cities with fishing activities.

Karen: Prefer unique products and promotions, indicating a specific interest in distinctive offerings.

After this we created another cluster **Younglings**, that are the 718 new customers with big spends.

8. Geospatial analysis

Next, our goal was to identify where our clusters are located. So to plot this map, we merged the data to obtain the unscaled location information and then plotted the map according to clusters.

Customer Locations by Cluster Fig.25

After analyzing the plotted map, we observed that customers associated with the "fishies" cluster are predominantly located in Peniche and Ericeira. This finding is not surprising, given the heatmap output and the fact that these are coastal cities with robust fishing activities. This insight suggests that the retail store likely collaborates directly with the fish market, purchasing fish to resell to other markets. The significant amount of money spent by these customers indicates that they are probably not individual consumers but rather businesses involved in the fish trade.

9. Association Rules

In our effort to design effective promotions, we conducted a detailed analysis of our customer clusters and their associated rules, revealing strong product connections.

Association rules are crucial as they help identify frequently bought together items, enhance cross-selling opportunities, optimize inventory management, and personalize the customer experience. By combining cluster demographic analysis with these insights, we created targeted promotions tailored to each customer segment. For instance, recognizing that young professionals often buy yoga mats and protein bars together, we might offer a bundle discount. This strategy ensures our promotions are data-driven, relevant, and engaging, ultimately boosting customer satisfaction and sales.

Having said this after analyzing the invoices per cluster we have seen a peculiar event. As seen in the comparison between the number of clients and the number of invoices for each cluster, the "pet_lovers" cluster, which constitutes 10% of our customer base, only accounts for 25 invoices out of 89,952, representing a mere 0.03% of the total invoices. Conversely, the "uni_students" cluster, which comprises 2.5% of the clients, contributes 2.6% of the invoices. This outcome is advantageous in scenarios involving low data volumes, as it shows that the percentage of invoices closely matches the percentage of customers in the "uni_students" cluster, indicating a balanced and consistent invoicing pattern.

10. Marketing Campaign

Best Customer

You might also be interested! : In a purchase of an electronic device offer discounts on accessories such as chargers , cases , airpods.

Drink and Geek : In a large purchase of electronic devices (500€) offer a free champagne bottle

Karrens

Stacking Stonks : Coupons for everyday groceries which are allowed to be stacked (Gain coupon for each X€ spent)

TrickorTreat : Discounts in sweets (cakes, candy bars , gums , pastry)

Low Budget

'Tá Barato, leva mais!' : For clients which use promotions already (promotion seekers) but tend to buy only a few items in a lot of promotions apply take 5 pay 4 products in order to take advantage of the high frequency and boost the monetary

'Tá Barato, leva mais vezes!' : Create a campaign that for each transaction associated with a loyalty card gains X points, points will then be converted to instore balance with a "small" validation date and with a minimum total needed to be usable. We use this campaign to try to attract customers that can have a high/medium monetary but low frequency

An oily week : Create a promotion for the oil and cooking oil by using loyalty card

Normal Customers

Generic: Example: "Buy any two items from the electronics section and get 30% off on cooking supplies."

For all the family: Promotions for products related to kids and teens, as exemple baby food, diapers.

Not school Again: In September they receive special promotions ro buy books, pens, pencils …

Pet Lovers

Perfect Duo: When buying pet_food get 10% in your groceries

Uni Students

It's gonna be legendary : In the purchase of 2 bottles of wine, the third is free

What a Tuna: In the process of playing music in the store entrance, they receive a 30% discount on a bottle of beer.

Vegan

Staying Green: When buying vegetables/bio product get 20% in card if you get to a certain total.

Fishies

I Want the entire Ocean: When buying large volumes of fish get 10% in next purchases

Swans

Early Bird: If spending a X amount, get free breakfast on the shop

Generic: Example: Buy any two groceries items receive a coupon of 10% in a hygiene product

A little bit of everything: If a client buys X amount of different products, a pack of cooking oil, oil and a cake.

Younglings

You might also be interested! : In a purchase of an electronic device offer discounts on accessories such as chargers , cases , airpods.

Drink and Geek : In a large purchase of electronic devices (500€) offer a free champagne bottle

May the force be with you: Host themed sales events with time-limited discounts on a wide range of electronics and alcoholic beverages.

11. Conclusion

With this project, we present a clustering approach to a client segmentation problem. Throughout the project, we enhanced our machine learning skills and seized the opportunity to apply the knowledge acquired in our classes more effectively. By meticulously analyzing and processing the dataset, we identified distinct customer segments, each with unique characteristics and behaviors.

Our approach involved several steps, including data preprocessing, feature engineering, and the application of multiple clustering algorithms. We evaluated, used, and compared models such as KMeans, DBSCAN, Mean-Shift, and Ward's method, ultimately identifying ten distinct customer segments. This comprehensive comparison enabled us to select the most effective clustering strategy tailored to our dataset. However, we also acknowledge the limitations of our approach. Due to the project deadline, we were unable to fully optimize our clustering process.

Furthermore, we delved into the specific traits of each cluster, enabling us to propose targeted marketing campaigns aimed at addressing the specific tastes and needs of each segment. This project not only improved our technical proficiency with clustering techniques and dimensionality reduction methods but also demonstrated the practical applications of these methods in real-world business scenarios.

In conclusion, this project exemplifies the power of machine learning in customer segmentation, providing actionable insights that can drive more personalized and effective marketing strategies. It also underscores the continuous learning and application process inherent in the field of data science, bridging theoretical concepts with practical implementation.