Academic year 2023/2024

07/06/2024

Customer Segmentation: A Key to Unlocking Business Growth and Success

João Ferreira 20221912, Miguel Mendes 20221904, Rodrigo Maia 20221934

Faculty: Fernando Bação, Ivo Bernardo

1. Executive Summary

1.1. Overview of problem addressed

In this project, we were asked to analyze two datasets containing data related to customer demographics, behavior, and transactional history. Our task was, initially, to perform customer segmentation and identify distinct groups of customers based on their shared characteristics. After that, we develop targeted marketing strategies that are appropriate and appealing to customers' needs and preferences.

1.2. Summary of methodology and approach

Regarding the methodology adopted, we began with an exploratory data analysis to better understand our data, identify the distribution of our variables, and detect outliers and missing values. After that, we checked the correlation between variables, reconstructed them, and standardized them so that we could carry out the next steps. To finalize this process, we imputed the missing values.

Following this initial phase, we proceeded to the core of the project: customer segmentation. This complex process involved various algorithms, each contributing to the final result: KMeans, Hierarchical Clustering, DBScan, SOM, and UMAP visualizations.

The final solution resulted from an initial K-Means application with 8 clusters, followed by a second K-Means implementation to separate 3 clusters that were initially mixed. In the end, our final segmentation has 10 clusters.

1.3. Summary of key findings and results

In general, we concluded that different factors influence customer segmentation. In addition to the obvious differences in the different types of products preferred, which mark some segments, it should be noted that factors such as age and the percentage of products bought on promotion are very relevant in this division. In addition, the relationship between the number of complaints, the variety of stores visited, and the variety of goods purchased are also aspects that distinguish some consumer groups.

1.4. A recommendation based on your findings

Based on our segmentation using various algorithms and the analysis of association rules, we propose different marketing and promotional approaches adapted to the characteristics of each cluster formed. By focusing on the distinct preferences and tendencies of each customer group, we can optimize our strategy to have more satisfied customers and greater sales volume. These specific strategies include promotions, campaigns, and even simple physical modifications to the store.

2. Exploratory Data Analysis

Our analysis is built on two key datasets, customer_info and customer_basket, which provide a comprehensive view of our customers. The former offers valuable insights into each customer's demographic data and spending behavior, while the latter details the various baskets purchased by our customers.

Proceeding to the data cleaning and preprocessing phase, we identified whether any of our variables had missing values, followed by some visualizations to understand their distribution and identify outliers. Here, we detected that some variables had skewed distributions, but no values were seen as outliers that had to be removed, except the lifetime_spending_fish variable (Figure 1). After analysis, we realized that most of these customers had "Fishy" in their name, probably indicating that they would be disproportionate compared to the rest of the customers. Therefore, we decided to create a small data set with them so that they could be analyzed later without influencing the rest of the process (in the final solution, these clients will constitute a separate cluster). Also, the latitude and longitude variables, analyzed together, showed some points that were distant from the others (Figure 2), but we decided not to remove them and to interpret them later. Besides that, we rebuilt some variables: turned gender and loyalty_card_number (1 if there is an entry, 0 if it was a Nan) into dummies, and changed the format of variable customer_birth_date into age so that it was in a more straightforward format to interpret.

Following the data cleaning and preprocessing, we standardized our data using the MinMax scaler, which yielded the most favorable results after several tests. We also examined the correlation between variables and found that the lifetime_spend_fish variable had a relatively strong correlation with others. We tested removing it (Figure 3), but the best results were achieved by including it. To complete this part of the EDA, we imputed the few missing values in variables such as kids_home, teens_home, number_complaints, distinct_stores_visited, typical_hour, lifetime_spend_vegetables, and lifetime_spend_fish.

Below are some visualizations that allow you to check some of the abovementioned aspects. Additional visualizations can be found in the annex section at the end of this report.

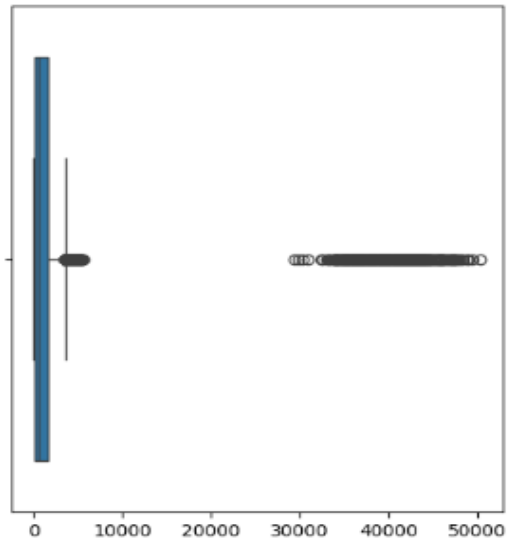


Figure 1: lifetime_spend_fish distribution

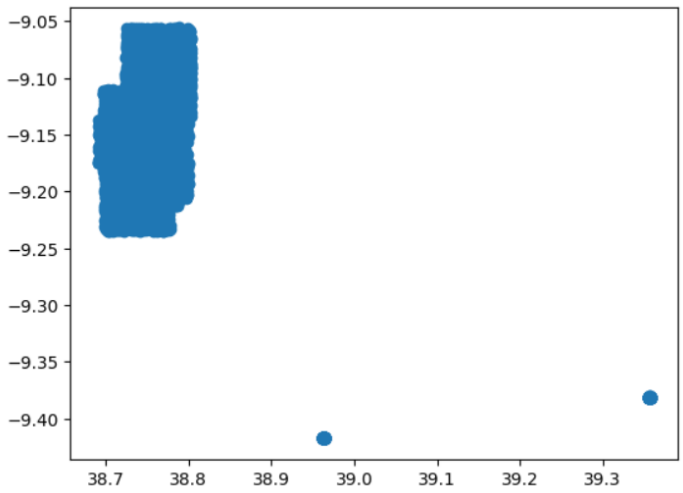


Figure 2: latitude and longitude

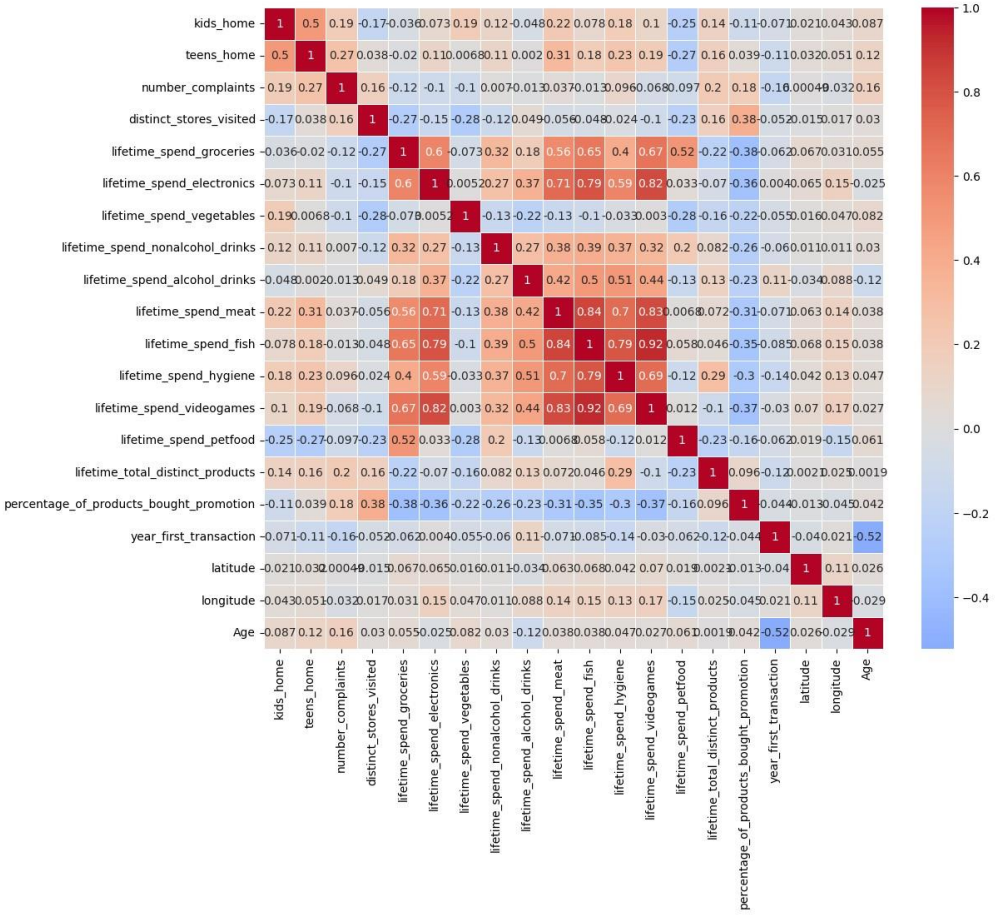


Figure 3: Correlation matrix

3. Customer Segmentation

The segmentation process began with using K-Means and the elbow method to find the optimal number of clusters (Figure 5). We continued with other methods to figure out the correct number of clusters. We used an auxiliary function to plot the silhouette scores (Figure 6). After that, we also tried a solution with hierarchical clustering, in which the input was the 100 centroids obtained through a K-Means with 100 segments (Figure 7). After all these steps and analyzing the results obtained, any number between 7 and 10 seemed reasonable. However, throughout our work and the different solutions we tested, 8 clusters yielded the best results; thus, we utilized the number in our final solution.

Below are some plots that illustrate the aspects mentioned above:

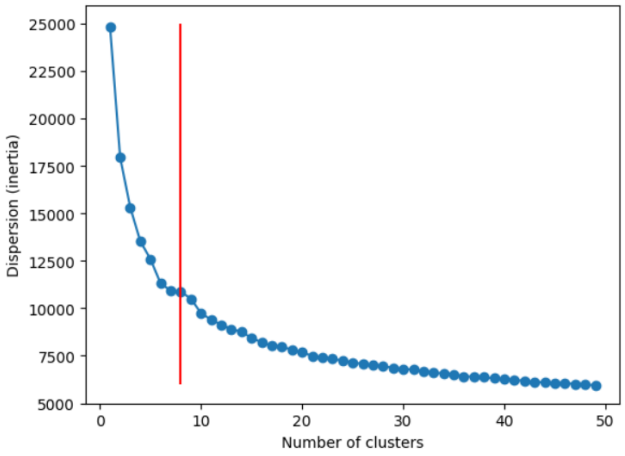


Figure 4: KMeans suggested nº of clusters

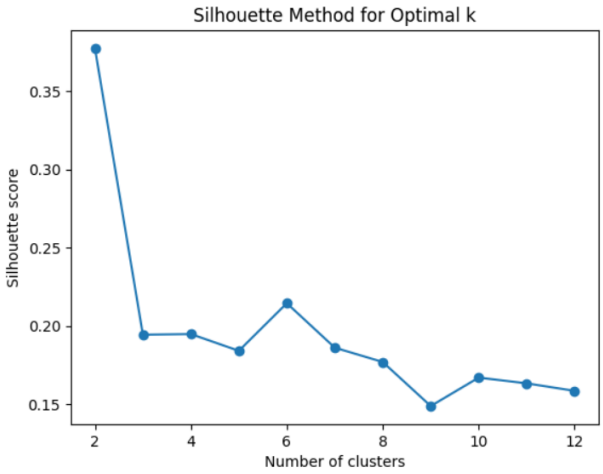


Figure 5: Sillhouette method

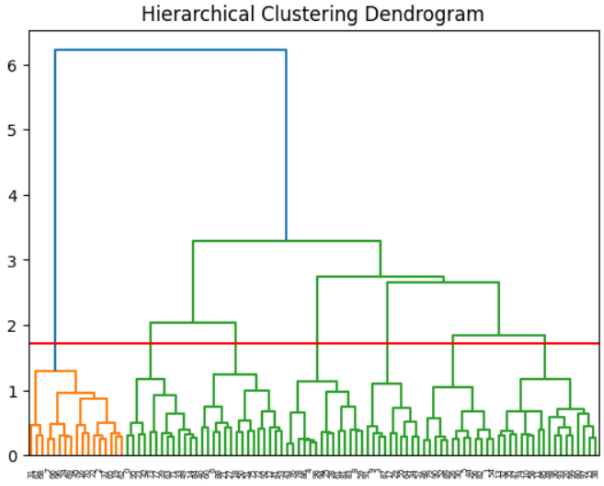


Figure 6: Hierarchical clustering with 100 initial centroids

As for the clustering process, we started by using KMeans (with the 8 clusters already mentioned). We then tried using a self-organizing map, which, after analyzing the quantization error, was trained in 500 iterations. To complete the initial phase of this process, we used the DBScan method. Here, we used an auxiliary function to optimize the epsilon hyperparameter, choosing eps=0.5.

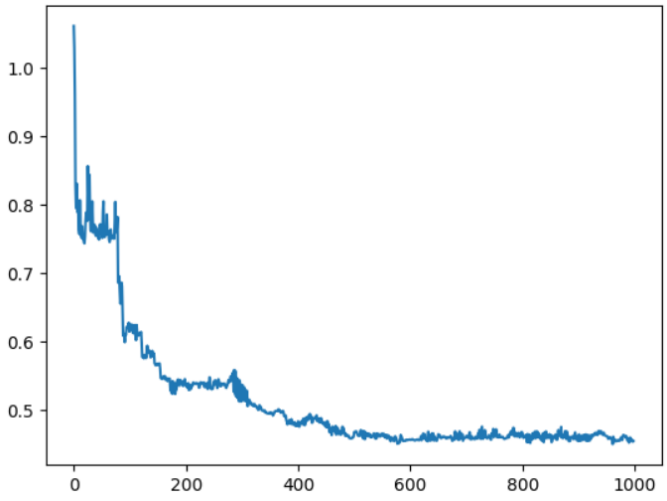


Fig 7: quantization error of SOM algorithm

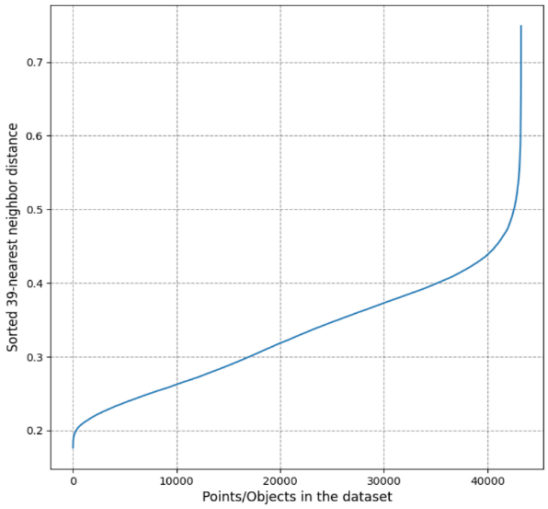


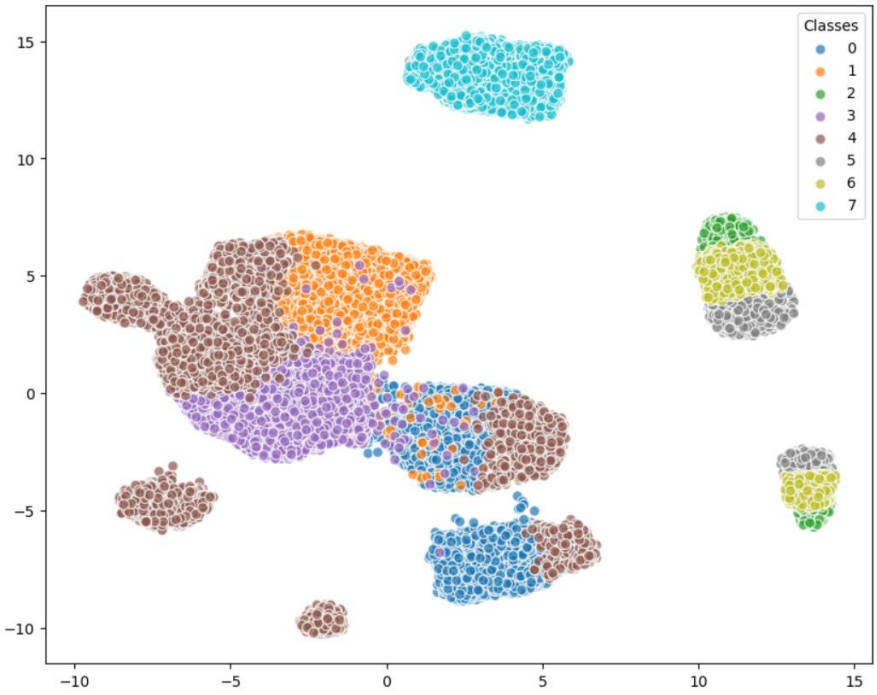
Fig 8: optimization of EPS for DBSCAN

4. Cluster General Analysis

We used the UMAP visualization for each of the three clustering algorithms used previously to analyze the segmentation solutions obtained in a general way.

UMAP for K-Means solution:

Looking at the visualization below, we can see that the light blue cluster is the most characteristic and distinctive. In addition, it is also noticeable that the yellow, green, and gray clusters stand out from the rest, appearing well-defined, as does the brown cluster, which, despite appearing in different regions of the space, is relatively well-delineated. In contrast, we see some dispersion between the purple, orange, and dark blue clusters. However, this issue was later resolved, as explained in this report.



Afterward, we made UMAP visualizations to understand the results obtained with the Self-Organizing Map and DBScan. However, as seen below, none of these techniques allowed us to obtain an interpretable segmentation.

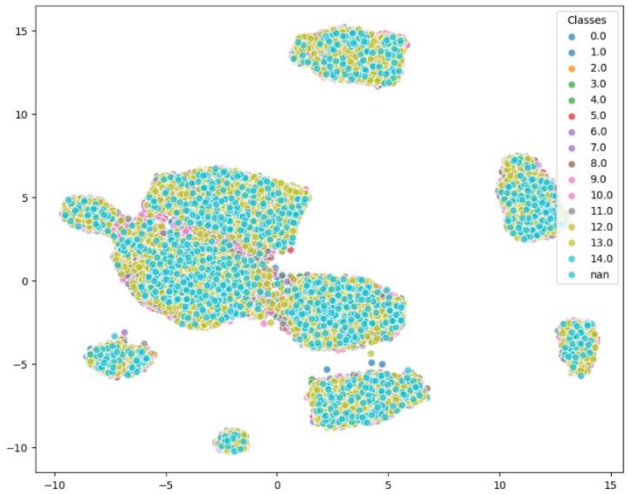


Fig 10: UMAP for SOM segmentation

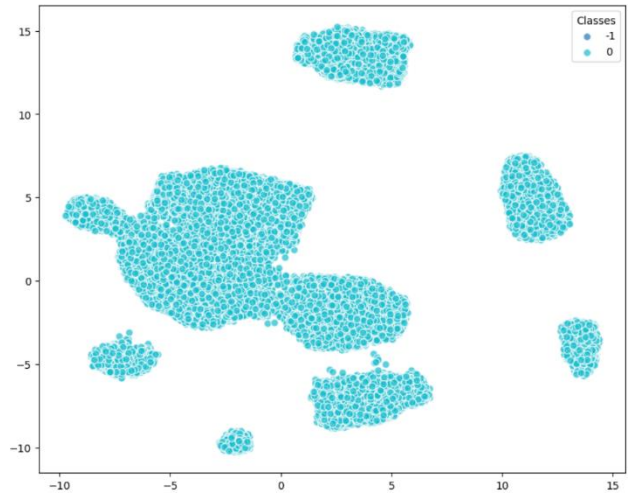


Fig 11: UMAP for DBSCAN segmentation

As expected, we kept the solution obtained with K-Means because, despite the problems already mentioned, it showed potential to reach the final goal. Therefore, we proceeded with our work, more specifically, by trying to improve the three clusters that were not possible to separate initially (dark blue, orange, and purple).

5. Cluster Improvement

We started this phase by following steps similar to those we used to identify the number of clusters. Again, we used K-Means and the elbow method, as well as hierarchical clustering with 100 initial centroids. Looking at the plots below, we concluded that four was the appropriate number of clusters to make.

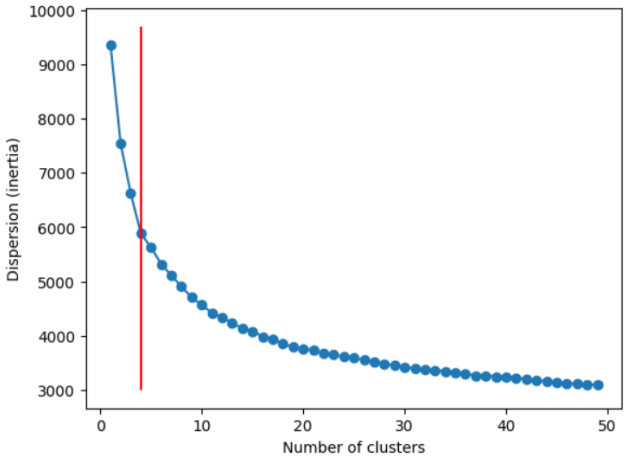


Fig 12: K-means – Elbow method

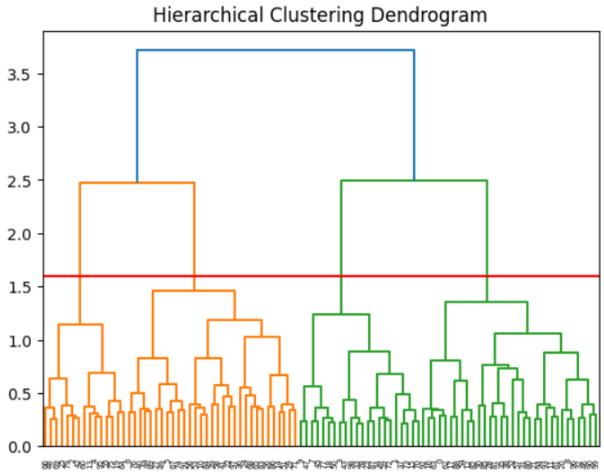


Fig 13: Hierarchical clustering – Dendogram

Then, just as before, we performed a UMAP visualization, which can be seen below:

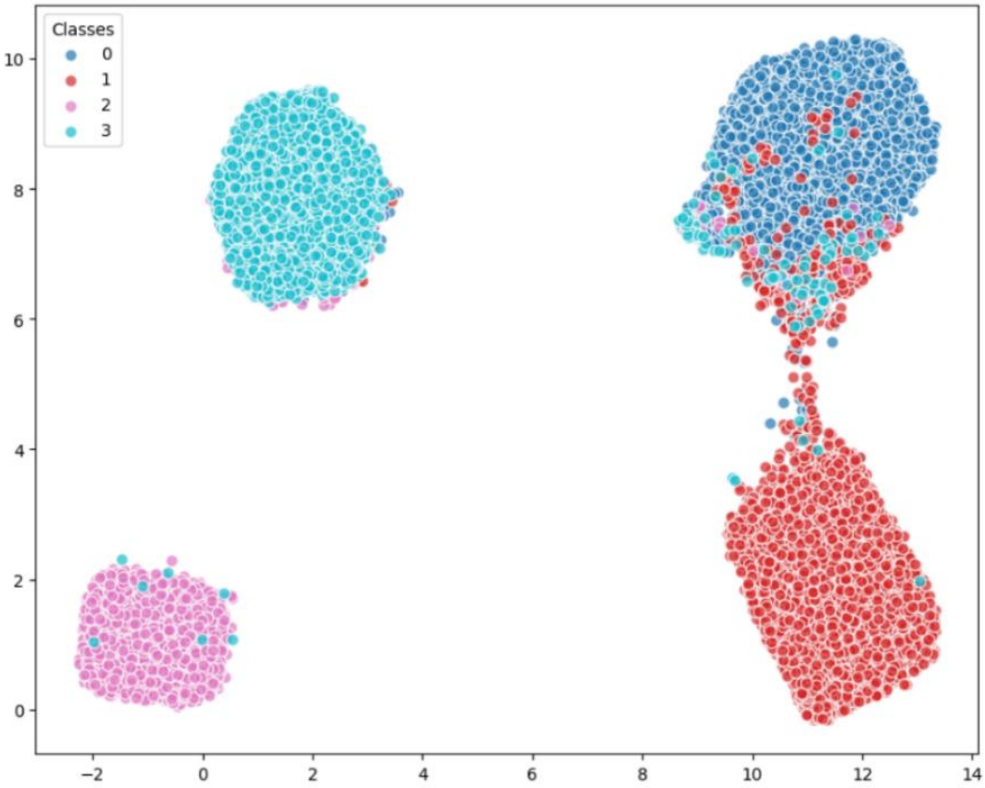


Fig 14: UMAP visualization of separation of clusters

Upon closer examination, the three initially ambiguous clusters have significantly improved in definition. Notably, the (now) light blue and pink clusters have become much more distinct. While the red and blue clusters still exhibit some mixing, the degree is incomparably reduced compared to the previous solution.

To summarize this part of the project, we started by testing different clustering techniques: K-Means, Self-Organizing Maps, and DBScan. After performing numerous tests and looking at the UMAP visualizations for each technique, we concluded that the best solution to start was K-Means with 8 clusters. However, the result obtained still had some flaws, as three clusters were poorly defined. Thus, we repeated the whole process only for the data points allocated in those three segments. Finally, we reached a solution that generated better-defined clusters by turning those three segments into four. In the next section of this report, we will give a more in-depth interpretation of each cluster, describing its details.

6. Cluster Specific Analysis

Before we begin, it is worth mentioning that in this section, we will look at each cluster, following a descending order regarding the number of clients they include. In addition, we will refer to each segment by its color. To help you keep track, see the figure below.

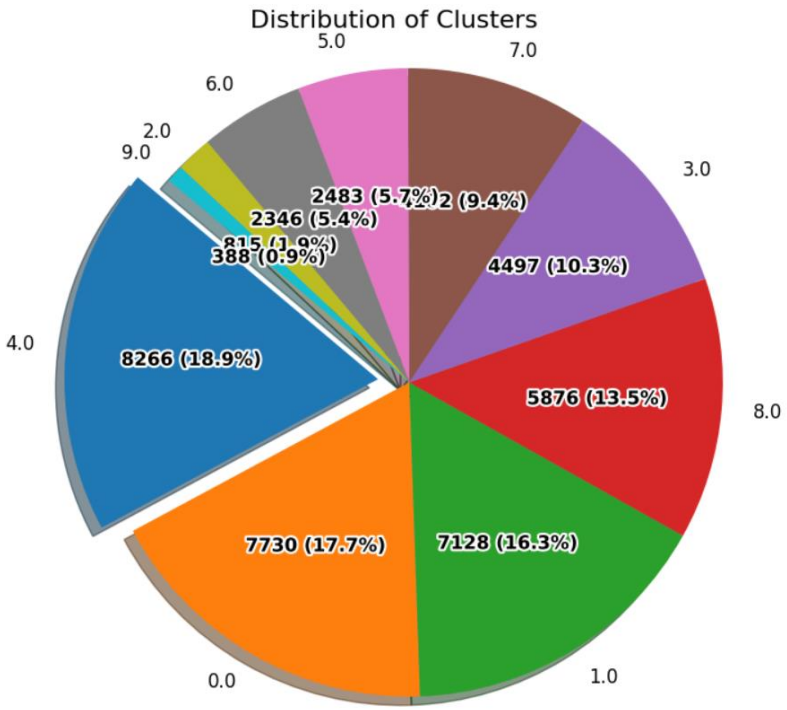


Fig 15: percentual representation of each cluster

Dark blue cluster – 18.9%

- Second lowest average age (28.85) by a wide margin.
- The first transaction is, on average, the second most recent, also by a significant difference (about seven years).
- One of the three clusters in which the percentage of products bought in promotion is higher than the mean (however, not by a significant value).
- The number of kids/teens at home is relatively smaller than the average.
- It does not stand out for the purchase of any specific type of product and only exceeds the average in alcoholic drinks.

Briefly, and considering its age, it has the expected description. Younger clients usually are more recent and do not yet have kids. Also, they probably have a smaller income, which explains the products bought in promotion. Regarding the type of products, they likely have a more active social life, which explains the spending on alcoholic drinks.

Orange cluster – 17.7%

- A high percentage of clients have a loyalty card.
- The highest number of distinct stores visited, almost double the average.
- A relatively large number of complaints.
- Includes customers who buy the highest percentage of products on promotion, with a value over 2x the mean.

In summary, this segment seems to have the type of customers who are constantly looking for promotions, which is likely the reason for the high loyalty card ratio. This behavior could be leveraged for targeted marketing, especially in relation to the greater number of different stores visited, as they are actively seeking these discounts. Their frequent visits to various stores, each with a unique apearence, could also be an opportunity for improvement, as they may be more critical of them, noticing the defects in each one more easily.

Green cluster – 16.3%

- It doesn't stand out in any of the features unrelated to the products themselves.
- Regarding consumption habits, it's worth noting that spending on vegetables is, by a very wide margin, the highest. In contrast, expenditure on meat and fish is close to zero. It is also considerably below average in the other categories.

In short, this cluster includes people who are most likely vegetarians. As it represents a considerable part of the customer base, it suggests that the offer of vegetarian products in this store is good.

Red cluster – 13.5%

- By far, the larger number of kids/teens at home.

- Typically, they go shopping later in the day.
- Tend to buy more diverse products than the average customer.
- Includes the buyers that complain a lot.

In a nutshell, this "red" group comprises large households, which explains the later time at which they shop. Despite the average consumption trend, they are responsible for buying a wide variety of products, which explains their greater exigency and thus the number of complaints.

Purple cluster – 10.3%

- Customers with the longest tenure (more loyal).
- Biggest diversity in products chosen, high expenditure on beverages and hygiene products.
- Visit the store very early in the day.
- People who complain the most and visit more establishments than the average.

In short, this cluster includes the most loyal customers, which is in accordance with the fact that they visit more stores and buy more diverse products. This longevity and diversity may explain their high standards, which is why they complain a lot. Despite this, a regular pattern is noticeable in their purchases, with the exception of their preference for drinks and hygiene products.

Brown cluster – 9.4%

- Almost no one in this cluster has kids/teens at home.
- Always go to the same store.
- A large part does not have a loyalty card.
- Absurd spending on pet food. Contrarily, there is very low expenditure on most of the remaining types of products.

This segment is mainly characterized by two factors: no kids/teens at home and the enormous expends in pet food. It is relevant to mention that this group of clients always go to the same retailer and mainly don't have a loyalty card, potentially missing on its benefits.

Pink cluster – 5.7%

- Oldest customers.
- Do not tend to look for discounts.
- Overall, they spend a lot in most categories, especially electronics, video games, and meat.

Remarkably, this cluster comprises our high-value customers, who spend across various product categories and do not benefit from promotions. What's particularly unexpected is that these customers, who are predominantly the oldest, show a significant expenditure in videogames and electronics, warranting further investigation.

Grey cluster – 5.4%

- Very similar characteristics when compared to the previous cluster described, with an even higher spending on electronics and groceries.
- Younger customers than usual.

In conclusion, premium group of clients that is very important for the company, since they buy almost every type of product in a large scale, while not worrying about saving money.

Yellow cluster – 1.9%

- Just like the previous two segments, high-quality consumers who are willing to buy the best products in different categories.
- The group with the youngest people and most recent clients.

Briefly, they are very valuable customers for the company, considering their buying habits and age.

Note: These last 3 clusters only differ in age. However, that is an essential aspect for a company, given that younger people have a longer customer lifespan.

Light blue cluster – 0.9%

- Usually, go to the store very early.
- There is an incredible amount of spending on fish. Contrarily, there is zero expenditure on video games and pet food. Apart from that, regular consumption patterns.
- Only go to one establishment.
- Live in an area far away from other customers (and close to the sea).

To sum up, if we consider their location near the coast, the time they visit the store, and the fact that they always go to the same establishment, these customers are probably fishmongers or restaurant owners. The latter may help explain the significant spending on other types of products.

7. Targeted Promotion

The first promotion we suggest implementing is a Euro's related promotion where we would give a years of purchases in our stores for free (for up to 300 euros spent per month) to everyone who guesses correctly the results of everyone of the fifty one games (1X2), for every 7 years of loyalty card a client has they would earn an extra life, meaning that even if they fail to guess one result they would still win the final prize. As an extra prize everyone that guesses 20 results will be eligible to a contest where 200 randomly chosen people will win a BMW X7. To be eligible to participate in these challenges costumers must spend at least 100 euros in June and 100 euros in July in our stores. This would be a great way to attract new costumers and make the ones we have already motivated to spend more.

Note: All the promotions/campaigns described below are based not only on the general description of each cluster but also on some patterns we have identified in the association rules.

Dark blue cluster

Given the characteristics of this cluster, the promotions approach should be tailored to the specific needs and behaviors of younger, recent customers with fewer responsibilities and likely a tighter budget.

Suggested promotions and marketing strategies:

- "Party Pack" – Buy two items from each of the following sections (Alcoholic beverages, Mixers, Snacks and Party essentials) and get a 5% discount on a future purchase.
- "Ultimate Party Pack" – Buy five items from each of the following sections (Alcoholic beverages, Mixers, Snacks and Party essentials) and get a 15% discount on a future purchase.
- Install our mobile app and get a 10€ coupon for a future purchase.
- For each 20€ spent on either mashed potatoes, tomatoes, carrots, asparagus or melons, get a credit. When you reach 15 credits, earn a free beverage cooler or a snack dispenser.

Orange cluster

This segment is highly motivated by discounts and promotions, and their critical nature requires a focus on quality and customer feedback. It is composed by critical customers who frequently use loyalty programs and visit multiple stores.

Suggested promotions and marketing strategies:

- Exclusive sales to reward their loyalty.
- Shop at three different stores within a month and get a 5% discount. For each additional store visited, get an additional 5% off the immediate purchase.
- For each suggestion/complaint that directly results in an improvement, get bonus loyalty points that can be redeemed into prizes in the future.
- Place technology items between the drinks and groceries.
- Buy two bottles of any type of wine, get one free.
- Spend a minimum of 50€ in the technology section and get a 5€ coupon for drinks.

Green cluster

Given the focus on vegetables and vegetarianism, the promotions approach should emphasize the store's offerings in this category while also catering to their broader lifestyle preferences.

Suggested promotions and marketing strategies:

- Spend 25€ in vegetables, receive a book of vegetarian recipes.
- 10% discount every 1st Monday of the month in all the categories (except vegetables and fruits).
- Supermarket stamp booklet: earn a stamp for every 25€ spent on vegetables and fruits. Extra stamps for 15€ spent in other categories. When 20 stamps are collected, claim a re-usable bag.

Red cluster

Considering the characteristics of this cluster, the promotional approach should be adapted to larger households that shop later in the day. These customers buy a wide variety of products and tend to complain more. Promotions should focus on family-friendly deals, evening discounts, and better customer service to meet their needs and improve their shopping experience.

Suggested promotions and marketing strategies:

- Family time at the cafeteria: buy four menus and get a free coffee or “pastel de nata”.
- Buy four family packs throughout a certain month and earn a 30% discount on a variety of offers, such as cinema, cooking classes, or craft workshops.
- For each suggestion/complaint that directly results in an improvement, get bonus loyalty points that can be redeemed into prizes in the future.
- Buy any variety of oil and receive a 10% discount on any variety of french fries and dressings.

Purple cluster

These customers buy a wide range of products, focusing on beverages and hygiene items. They tend to have high standards, resulting in more complaints. Promotions should focus on exclusive deals on beverages and hygiene products and enhanced customer service to reward loyalty and improve the shopping experience.

Suggested promotions and marketing strategies:

- Product bingo: complete the product bingo and get 15% off on a future purchase.
- For each suggestion/complaint that directly results in an improvement, get bonus loyalty points that can be redeemed into prizes in the future.
- For each 20€ spent at the store after 3 pm, earn a credit. Use those credits to claim a variety of hygiene products (set of towels, premium hand soaps or aromatic room sprays, for example). Alternatively, can trade those credits for a pack of artisanal beers, wine oxygenator or a set of Asian teas.

Brown cluster

This group of customers is loyal to one store but often doesn't have a loyalty card, missing out on its perks. Promotions should focus on pet food discounts, incentives for joining the loyalty program, and personalized offers to encourage broader spending. This will help enhance their shopping experience and maximize their benefits from being frequent shoppers.

Suggested promotions and marketing strategies:

- For each store that you visit and make a purchase of at least 15€, receive a 5% discount on vet services (maximum four different establishments).
- Sign up for the customer card and receive a 15€ voucher to spend in the pet accessories section.
- Place pet food and related items away from the retailer´s entrance.

Pink cluster

These high-value shoppers spend significantly across various categories, especially electronics, video games, and meat. Promotions should focus on exclusive offers and personalized experiences to cater to their broad purchasing habits and enhance their overall satisfaction. This strategy will reward their loyalty and align with their spending patterns.

Suggested promotions and marketing strategies:

- Exclusive product launch events: provide VIP access to exclusive product unveilings and demonstrations, giving the opportunity to purchase premium products before they are available to the general public.
- Offering samples of high-quality products in the charcuterie and alcoholic beverages sections.
- Buy two out of these four products (champagne, headphones, mobile phones, or laptop) and get 10% off in cheeses.

Grey cluster

Here, the approach is very similar to the previous one. The only difference is that, by analyzing some association rules, the last promotion is more adapted to their preferences.

Suggested promotions and marketing strategies:

- Exclusive product launch events: provide VIP access to exclusive products unveilings and demonstrations, giving the opportunity to purchase premium products before they are available to general public.
- Offering samples of high-quality products in the charcuterie and alcoholic beverages sections.
- Buy 2 out of these 4 products (champagne, headphones, mobile phones or laptop) and get 10% off on meat.

Yellow cluster

Again, the approach is similar to the previous two, as these three clusters have similar characteristics. However, as in this one we had younger and more recent customers, we decided to make some changes that would help to captivate and maintain them.

Suggested promotions and marketing strategies:

- Exclusive product launch events: provide VIP access to exclusive product unveilings and demonstrations, giving the opportunity to purchase premium products before they are available to the general public.
- Offering samples of high-quality products in the charcuterie and alcoholic beverages sections.
- Partner with social media influencers to promote products and events.
- For each purchase above 50€, get 50% discount on entry to a nightclub.
- Buy the bundle of champagne, cottage cheese and turkey and get 5% off a future purchase above 100€.

Light blue cluster

Considering their coastal location and shopping habits, these clients are likely fishmongers or restaurant owners, which explains their significant spending on other products. Promotions should focus on bulk buying discounts and offers suited to their professional needs.

Suggested promotions and marketing strategies:

- Provide personalized assistance to clients that are restaurants owners in order to guarantee their unique requirements and best-quality fish/seafood.
- For each 100€ spent on fish, get a 10 cent/l discount in fuel.
- Essential Pantry Pack: combine 10 different products from the groceries section and get a credit. Then, trade those credit for baby-related products, such as baby health kits, re-usable diapers, soft plush toys or baby books.

8. Conclusion

In this project, we proposed that using statistical and machine learning techniques, we would identify, within the customer base, different clusters representing customers with similar characteristics. After this, the aim was to analyze each segment's characteristics and purchasing trends to develop personalized marketing strategies to meet our client's needs and attain greater satisfaction.

Regarding our approach, after some initial steps to get to know and pre-process given data (distributions, removal of outliers, transformation of variables, normalization), we implemented various clustering techniques, some of them to help with the primary purpose (Hierarchical Clustering, UMAP, KMeans), and others to carry out the segmentation itself (KMeans, SOM, DBScan). Once the final solution was obtained, we analyzed each cluster formed and found the respective association rules. Finally, based on the conclusions from the previous steps, we developed various promotions and campaigns to improve our service and satisfy the customers' preferences and habits.

Concerning our key findings, we concluded that numerous factors can influence customer segmentation. Although some are already known, such as the type of preferred products and customer's age and longevity, it was interesting to realize that there are many other relevant features influencing customer segmentation, such as the number of products bought on promotion, the diversity of purchases, and even the number of complaints. In addition, it was also appropriate to analyze the association rules and identify some patterns that were, in a certain way, unexpected.

Having completed the project, we have obtained a fairly reasonable final result for customer segmentation. We worked hard to improve our results, combining different models, testing combinations of variables, and optimizing the hyperparameters of the various algorithms. We also have carried out a complete analysis of the clusters obtained regarding their more general characteristics and the patterns found in the association rules. Overall, we were satisfied with our work, with the awareness that we could improve. This project allowed us to work on this segmentation process for the first time and with a practical objective, so it was beneficial for us to apply the knowledge we had acquired throughout the semester. Finally, we believe that the final part, which consists of creating campaigns/promotions, makes perfect sense in this type of work, as it gives it a more genuine meaning, ending the connection that exists throughout the various tasks proposed.

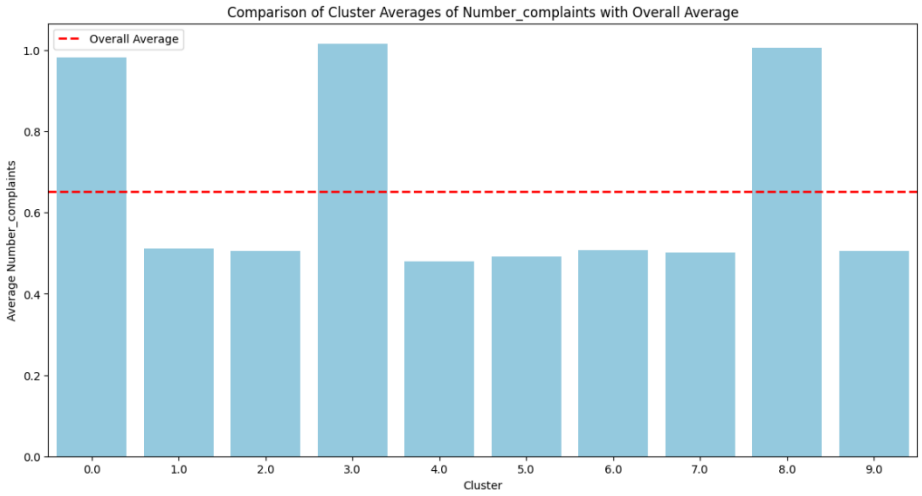
References:

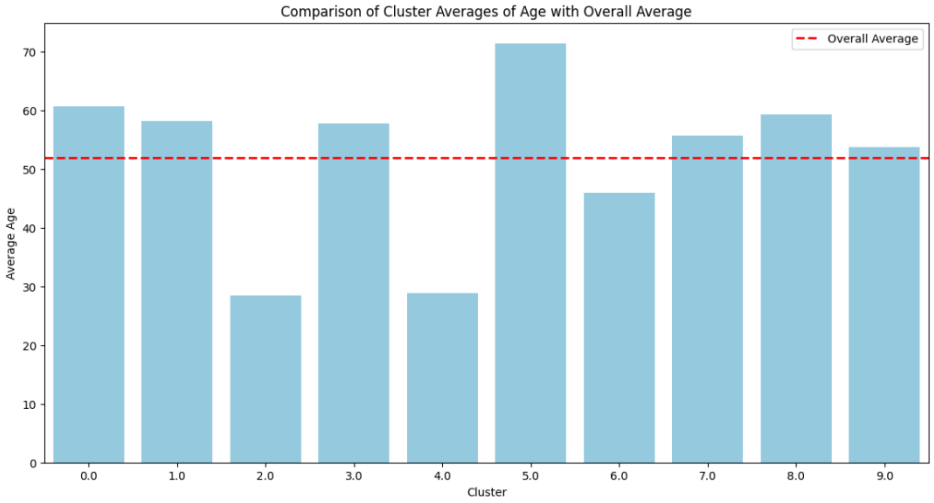
1. https://stackoverflow.com/questions/63799332/how-determine-optimal-epsilon-value-in-meters-for-dbscan-by-plotting-knn-elbow
2. https://umap-learn.readthedocs.io/en/latest/parameters.html
3. https://smorabit.github.io/blog/2020/umap/#:~:text=min_dist%3A%20The%20minimum%20distance%20between%20two%20points%20in,maxiter%3A%20The%20number%20of%20iterations%20for%20UMAP%20optimization.
4. Notes and PowerPoint Slides of the course

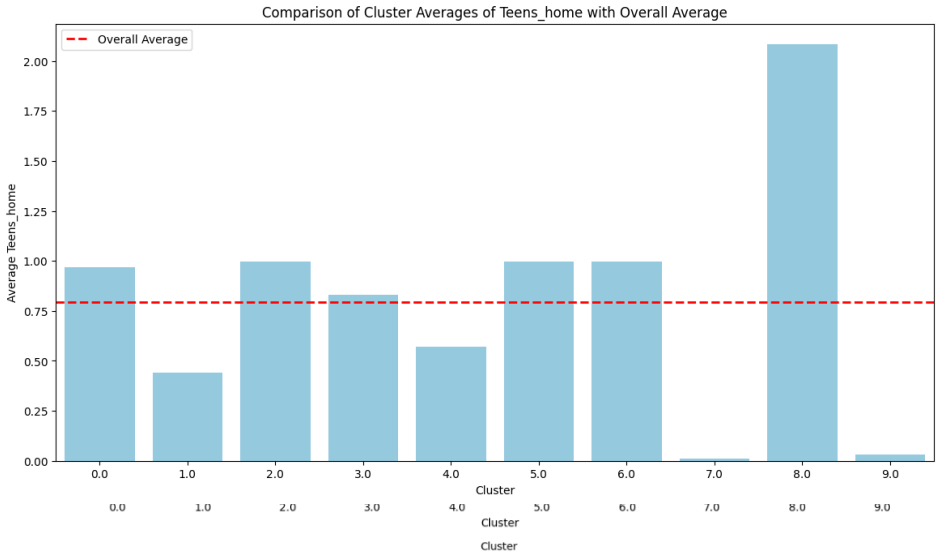
To have access to all the work developed, the following link gives access to our GitHub repository:

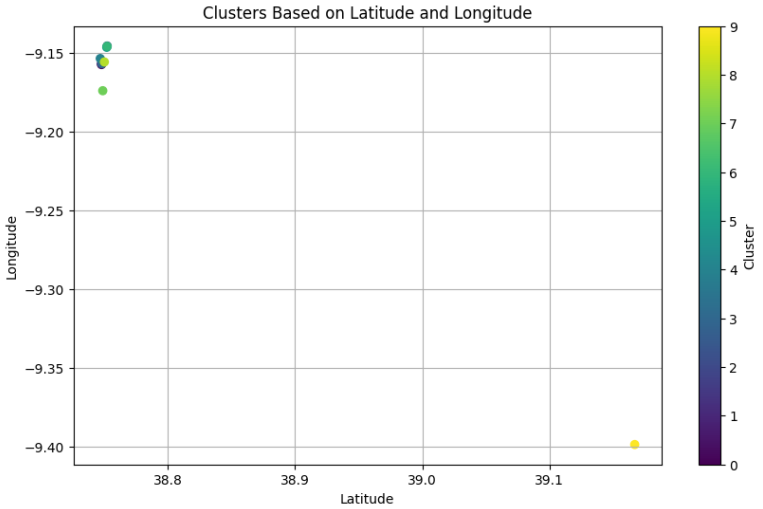
GitHub - JoaoF7/Project-ML

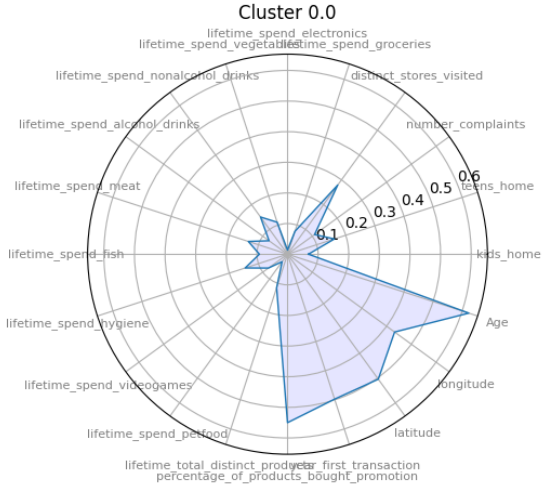
Annex

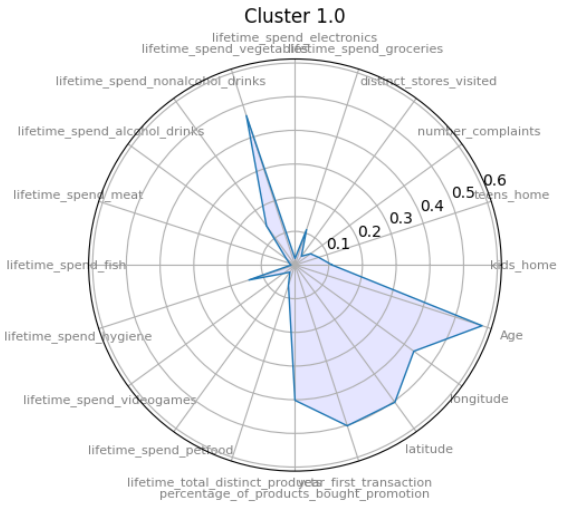

Comparison of Cluster Averages of Number_complaints with Overall Average
Overall Average
Average Number_complaints
Cluster

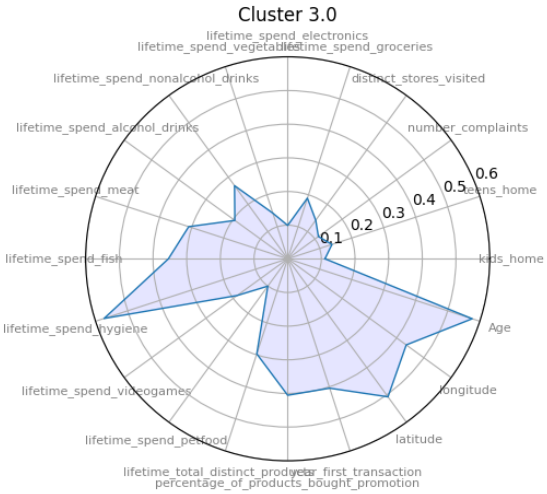

Comparison of Cluster Averages of Age with Overall Average
Overall Average
Average Age
Cluster


Comparison of Cluster Averages of Teens_home with Overall Average
Overall Average
Average Teens_home
Cluster


Clusters Based on Latitude and Longitude
Longitude
Latitude
Cluster
−9.15
−9.20
−9.25
−9.30
−9.35
−9.40
38.8
38.9
39.0
39.1
0
1
2
3
4
5
6
7
8
9


Cluster 0.0
lifetime_spend_electronics
lifetime_spend_vegetables
lifetime_spend_groceries
lifetime_spend_nonalcohol_drinks
distinct_stores_visited
lifetime_spend_alcohol_drinks
number_complaints
lifetime_spend_meat
teens_home
lifetime_spend_fish
kids_home
lifetime_spend_hygiene
Age
lifetime_spend_videogames
longitude
lifetime_spend_petfood
latitude
lifetime_total_distinct_products
year_first_transaction
percentage_of_products_bought_promotion
0.1
0.2
0.3
0.4
0.5
0.6


Cluster 1.0
lifetime_spend_electronics
lifetime_spend_vegetables
lifetime_spend_groceries
lifetime_spend_nonalcohol_drinks
distinct_stores_visited
lifetime_spend_alcohol_drinks
number_complaints
lifetime_spend_meat
teens_home
lifetime_spend_fish
kids_home
lifetime_spend_hygiene
Age
lifetime_spend_videogames
longitude
lifetime_spend_petfood
latitude
lifetime_total_distinct_products
year_first_transaction
percentage_of_products_bought_promotion
0.1
0.2
0.3
0.4
0.5
0.6


Cluster 3.0
lifetime_spend_electronics
lifetime_spend_groceries
distinct_stores_visited
number_complaints
teens_home
kids_home
Age
longitude
latitude
year_first_transaction
percentage_of_products_bought_promotion
lifetime_spend_petfood
lifetime_spend_videogames
lifetime_spend_hygiene
lifetime_spend_fish
lifetime_spend_meat
lifetime_spend_alcohol_drinks
lifetime_spend_nonalcohol_drinks
lifetime_spend_vegetables
0.1
0.2
0.3
0.4
0.5
0.6