
# Final Project

Machine Learning II

Dinis Gaspar nº 20221869

Dinis Fernandes nº 20221848

Luís Mendes nº 20221949

**Table of contents**

# Executive Summary

The main purpose of this project was to perform customer segmentation based on two datasets, one which included data on customer's transactions and another one which included data regarding customer's personal information, such as their gender, name, etc…

Initially, data about customer information was imported and subjected to exploratory data analysis. During this phase, certain variables were adjusted for clarity, while new variables were introduced to improve interpretation. The process also involved verifying missing values, outliers and data types. Additionally, the dataset's variables were examined for distribution, and various graphical representations were generated, including those related to client gender, education level, number of children, and customer tenure, among others. Inconsistencies were identified and rectified during this process. Finally, the pre-processed data was exported to a new CSV file.

The following step consisted of scaling of the data and implementing clustering techniques, such as K-Means, Hierarchical Clustering. Later the clusters were visualized in a multidimensional space, using a UMAP dimensionality reduction, and a final clustering solution was reached.

On the final clustering solution, we ended up with nine different groups of customers, which by their characteristics were given the following names: Fishermen, Gamers, Pet Lovers, Young Party People, Vegetarians, Young and lots of Electronics, Loyal Customers, Promotions and Parents.

After careful analysis of the clusters, targeted promotions were created for each of them.

By using machine learning techniques to segment our customers, we gained valuable insights that will improve company's marketing strategies, personalize customer experience, and increase business growth. Therefore, we recommend implementing these strategies to take advantage of the unique traits of each customer group and provide targeted solutions that best match their behaviours.

## Exploratory Data Analysis

The first step in any problem is to get a good understanding of the data we're working with; the Exploratory Analysis section aims to obtain the most information about our data to better judge what techniques will be required regarding the possibility of missing values and outliers as well as what transformations will be made to the data to improve its usability.

From the analysis, we concluded that there are missing values in our data, as well as outliers looking at the quartiles, minimum and maximum values. From the .describe() we can also conclude that a missing value in the number_complaints variable doesn't mean that a customer has no registered complaints, but instead means that the data on that for that customer is just missing, we can conclude this because the minimum number of complaints is 0. We can also verify that the only variable with problematic values is percentage_of_products_bought_promotion because it has negative values and values higher than 1.

To get an even better idea of what our variables and their distributions look like we visualized them.

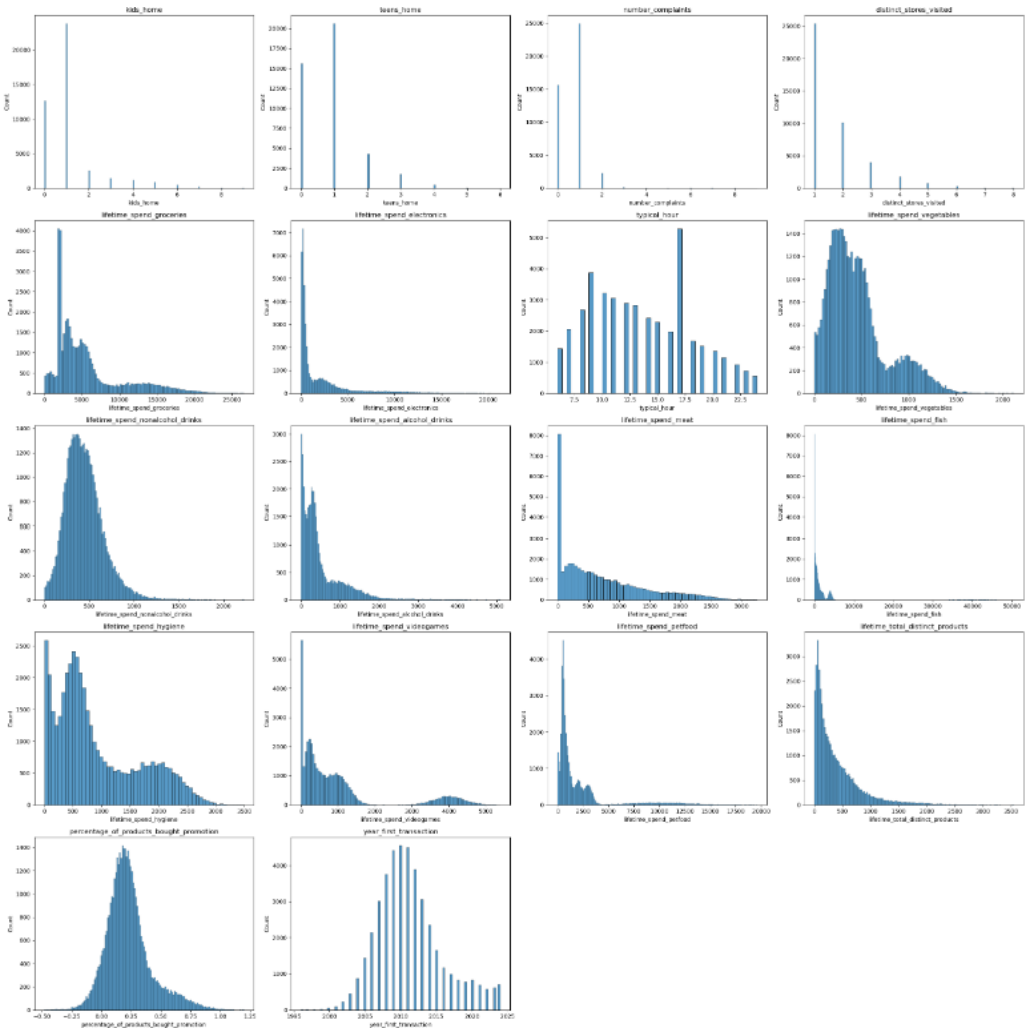

*Figure 1 Variables Distribution*

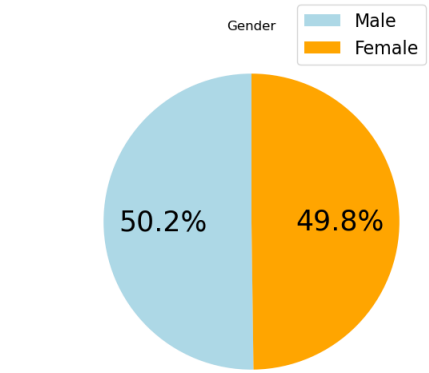


*Figure 2 Gender Distribution*

From the plots above we concluded that our dataset has a nearly identical distribution of male female customers. We can also confirm that there are outliers in the data, especially in the lifetime_spend_ variables.

*Duplicate Evaluation*

Before proceeding with further analysis, it was crucial to ensure that there were no duplicated individuals in our data, after analysis we concluded that all individuals in the dataset are unique.

*Feature Transformation*

In this section, we performed transformations to our data with the aim of obtaining more valuable and interpretable insight from our data.

- Birthdate of customers is replaced by their age as it's easier to work with as well as more interpretable.
- Gender of customers is transformed into a binary variable where 1's are males and 0's are females.
- Customer education (attainable from the prefix of the customers' name) transformed into dummy variables. This purely for analysis and not for modelling which means it makes sense not to include them in the imputation of missing values, for this reason this transformation is actually performed at the end of the pre-processing stage.

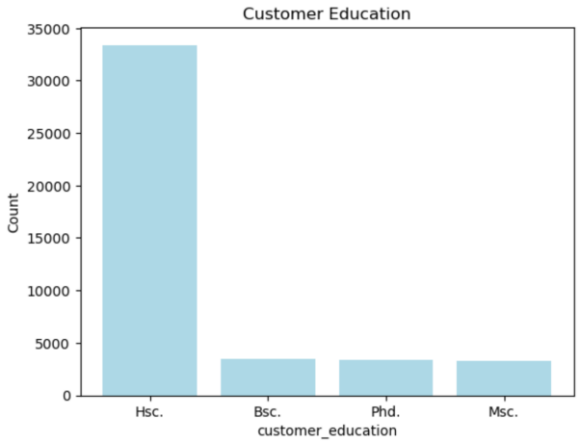


*Figure 3 Customer education distribution*

- As we can see from the bar chart, the distribution of our education variable has a very high disparity, this further confirms our decision to exclude it from modelling.

- Number of loyalty card adjusted into a binary variable, where 1 means the costumer has a loyalty card and 0 means the customer doesn't, as the number itself isn't very useful information.

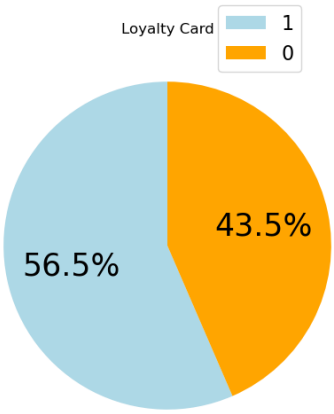


*Figure 4 Loyalty Card Distribution*

- From this transformation, we can see that 56.5% of our customers have a loyalty card.

- To have more relevant information on customers spending habits, instead of using the actual value spent on a given category we used the percentage of the total spent that was spent on that category, we also stored the overall total spent for each customer. These changes helped us in our task of grouping customers later in this project.
- Year of a customer's first transaction is often not very interpretable and useful for the purpose of this project we will replace it by the tenure of customer by simply subtracting the current year to the year of the first registered transaction.
- Percentage of products bought through promotion variable, as it had values lower than 0 and higher than 1 which are obviously not possible, as such lower values than 0 will be replaced by 0 and higher values than 1 will be replaced by 1. We will also transform the variable into an actual percentage by multiplying by 100 as it is more easily interpretable this way and doesn't negatively impact our data in any way.

*Outliers*

In this section, we re-evaluated the outliers in our data and decided what to do with them.

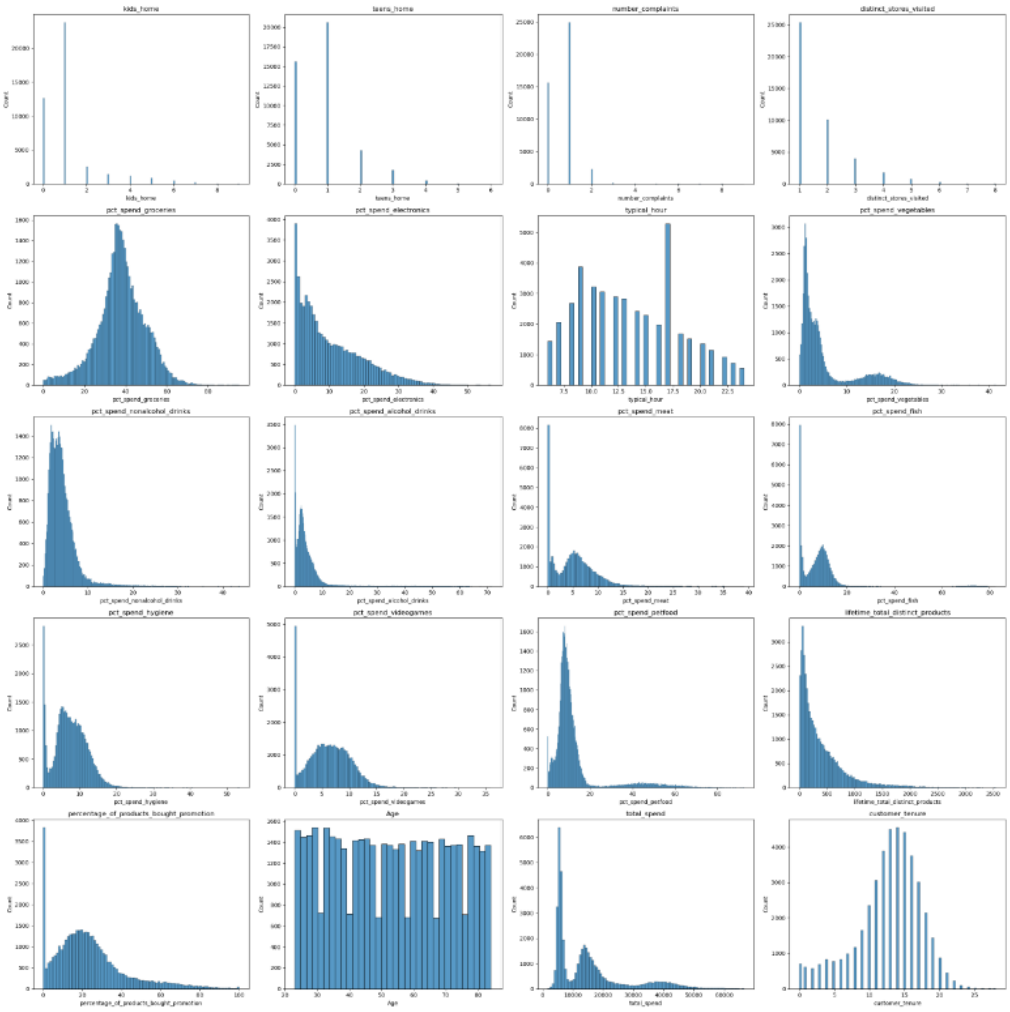

*Figure 5 Variable distributions after transformations*

In the treatment of Outliers, we created a function to evaluate the outliers of a dataset (outlier_check_IQR) which prints the percentage of the dataset that is retained after excluding outliers using the boxplot method and can optionally return a filtered DataFrame without outliers.

After the transformation which reduced the extremity of some outliers it was clear that there were still outliers but removing them using a method such as the Boxplot method (IQR method) would require a very large loss of information (38% of observations using normal and 60% using extreme outliers). In most cases, these outliers are groups and not just single points, which means that clustering with these outliers may not be as problematic. For these reasons, we will not exclude any outliers. Given the fact that no outliers are removed, going forward whenever there is a need to scale data, Robust Scaler will be used.

*Missing Values*

In this section we identified and treated missing values.

We have some missing values in 7 variables: kids_home, teens_home, number_complaints, distinct_stores_visited, typical_hour, pct_spend_vegetables, pct_spend_fish.

To deal with these missing values we will impute them using KNNImputer with 5 neighbours (default value) with a scaled version of our dataset scaled using Robust Scaler. From this, we then obtain a dataframe of imputed non scaled data using inverse_transform method of the scaler to obtain the imputed data in the original scale with the goal of removing float values in 4 of our variables, as float values aren't admissible in them: kids_home, teens_home, number_complaints, distinct_stores_visited, this was done simply by rounding those values to the nearest integer number. After this we scaled back our data without refitting our scaler to obtain a scaled and imputed DataFrame of our data, which had no missing values.

*Correlation*

In this section, we checked for highly correlated variables in our data.

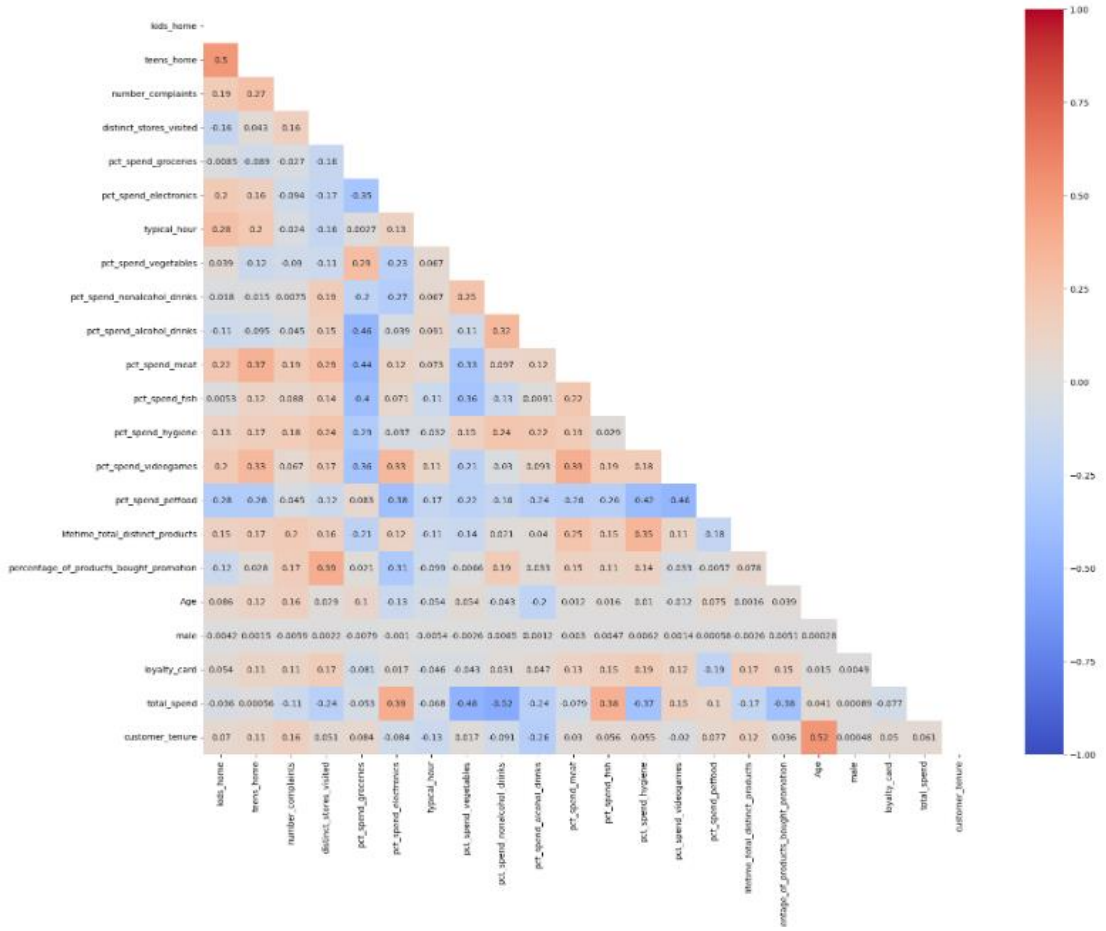


*Figure 6 Correlation Heatmap*

From the heatmap above we can conclude that in our data there aren't any pairs of highly correlated variables, so we won´t use PCA to reduce dimensionality because it won't give us valuable results.

# Customer Segmentation

*Geo Analysis*

Before using more advanced clustering techniques, we will perform a geography-based analysis of our customer base.

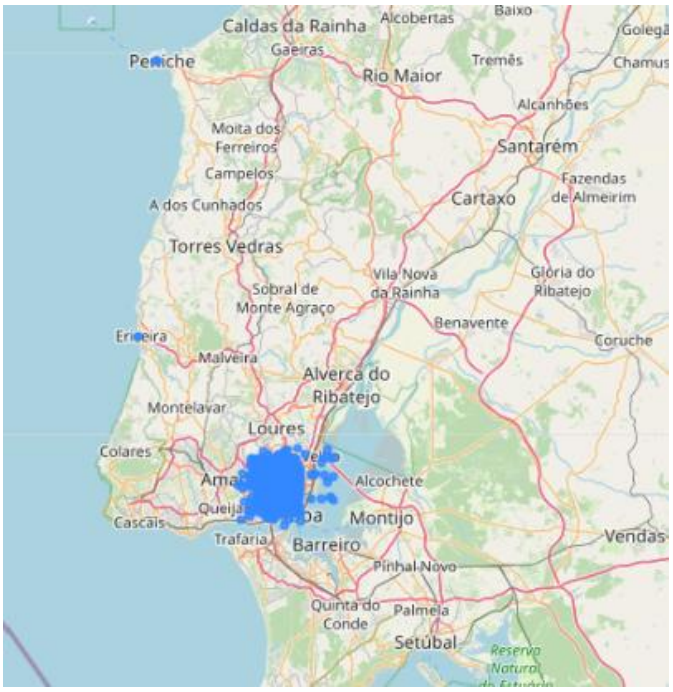


*Figure 7 Geographic Distribution*

We can clearly see that most of our customers are in Lisbon, but we can also see two small groups in Ericeira and Peniche. We split our customers in two groups, the ones from Lisbon and the ones outside of Lisbon, to do this we found a random customer in Lisbon and establish a buffer around them.

Looking at the customer_name it was easily noticeable that that there was a distinct group of customers which had “Fishy” in their name. This indicates that we could potentially have a particular segment of customers outside of the Lisbon area. Furthermore, we can see that all our customers that have "Fishy" in their name are from outside of Lisbon. With all this in mind, a comparative analysis will be performed between the two groups, using the non-geographic dataset.

After the analysis we can see that there is a clear difference in the way the group outside of Lisbon buys products, it's clear (by also looking at their name) that these customers buy a lot of fish. Potentially they are fishermen buying bait, as all of their expenses in other areas are extremely low ruling out the option of these being restaurants for example.

*Important Notes*

Now that we have identified the first segment, we split our data into classified individuals (individuals who have been assigned to a cluster, so far only the Fishermen)

and unclassified individuals (customers who hadn't been assigned to a final cluster), this dataset was used to evaluate different clustering solutions as well as to perform cluster comparisons, finally, data for modelling which had the data scaled used for fitting models in cases where we didn't want to perform additional transformations to the data.

Below, we will describe our final clustering solution, it is important to note that other methods such as Meanshift, DBSCAN and SOM were tested but do not feature in the final solution and therefore won't be mentioned, more details on their usage as well as the final reason for their exclusion can be found in the Project_clustering notebook. It is also crucial to mention that the data user for clustering in our final solution is a slightly transformed version of the original modelling data, where kids_home and teens_home are transformed into binary variables, this was tested at an early stage of the clustering process before any observations (excluding the fishermen) were finalized and yielded better results.

*Hierarchical Clustering - Ward Linkage*

The first algorithm to feature in our final solution is an Agglomerative Hierarchical clustering method using 'ward' linkage method fitted on the transformed data specified above.

To define the number of adequate clusters we plotted a Dendrogram.

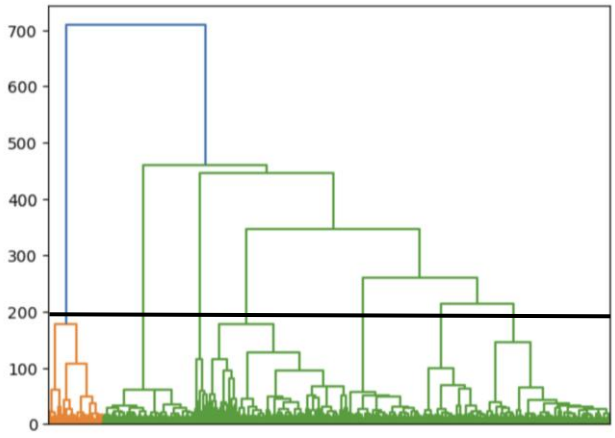


*Figure 8 Hierarchical Clustering Dendrogram (ward)*

After analysing the dendrogram we defined 7 as the number of adequate clusters.

The following UMAP dimensionality reduction allow us to get a better idea of the performance of the model.

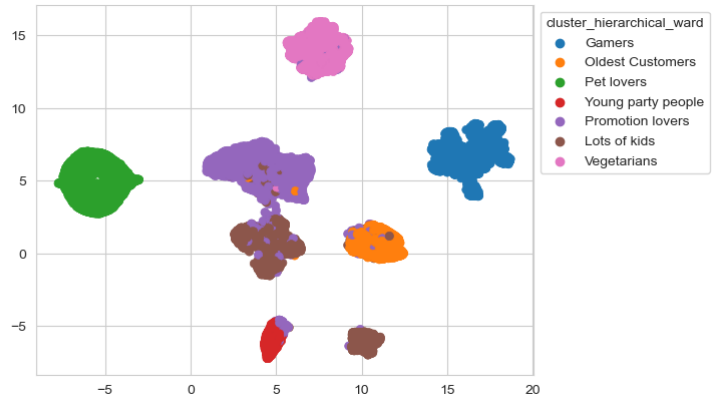


*Figure 9 UMAP of first Hierarchical Clustering (ward)*

Looking at the dimensionality reduction of the ward hierarchical model, we defined the final cluster as the ward hierarchical cluster for the following clusters: 'Pet lovers', 'Gamers', 'Vegetarians', 'Young party people', as these clusters can be easily identified. Subsequently, we added the newly classified customers to our classified_data and removed them from unclassified_data. We also removed them from both our modelling datasets. As for the remaining customers that haven't been classified a deeper analysis will be performed to find an effective method of splitting them into groups.

*K-Means*

We have established that simpler techniques which are good at finding spherical clusters are required to adequately cluster the remaining customers in our data, through testing the methods mentioned above, we will start by using KMeans once again. We use the elbow plot to find an adequate k.

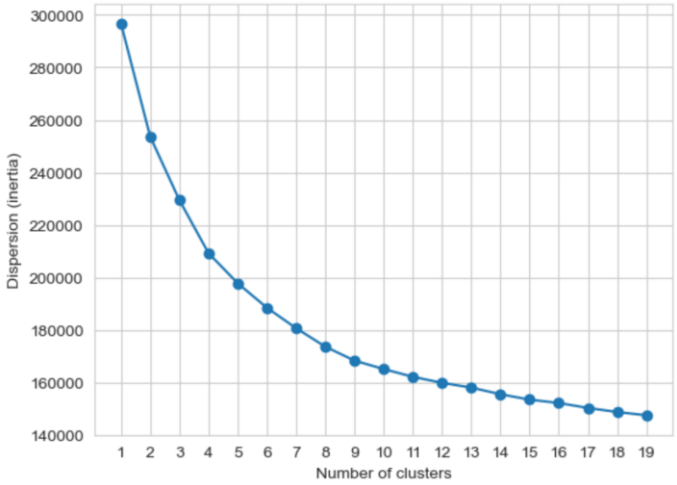


*Figure 10 elbow plot of K-means*

We choose k=5 because it allows us to reduce complexity and because in the UMAP dimensionality reduction (available in some of the above models) there are only 4 clear groups of points in the remaining data and thus it makes no sense to go much higher than 4.

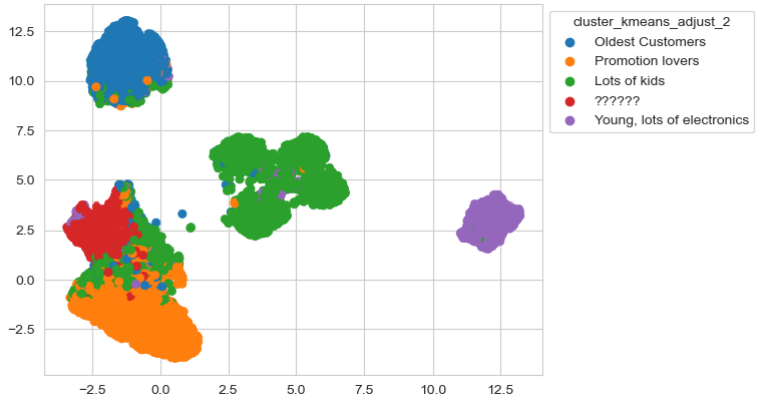


*Figure 11 UMAP of K-means (remaining customers)*

K-Means singles out a group of points (Young, lots of electronics) despite providing, in general, bad results. For this reason, we assigned to those observations their final cluster and repeat the process of moving them from one dataset to another, then re-filtering the modelling datasets.

*Hierarchical Clustering*

We are now going to use Agglomerative Hierarchical clustering using ward linkage again. We will use the dendrogram to find the optimal number of clusters which remain uncovered in our data.

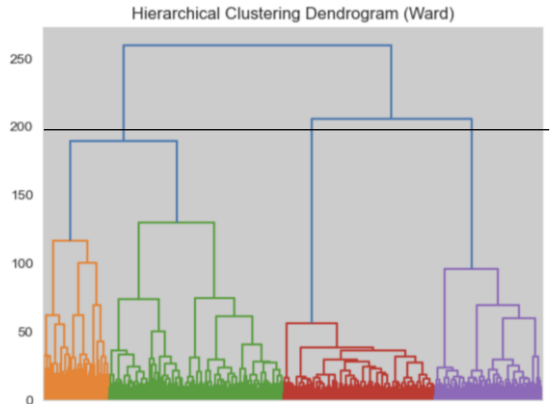


*Figure 12 Dendrogram of Hierarchical Clustering*

Despite the dendrogram showing us that 4 clusters are the right number, after testing both 4 and 3, 3 provides better results and it also coincides with the number of groups left in the UMAP dimensionality reduction.

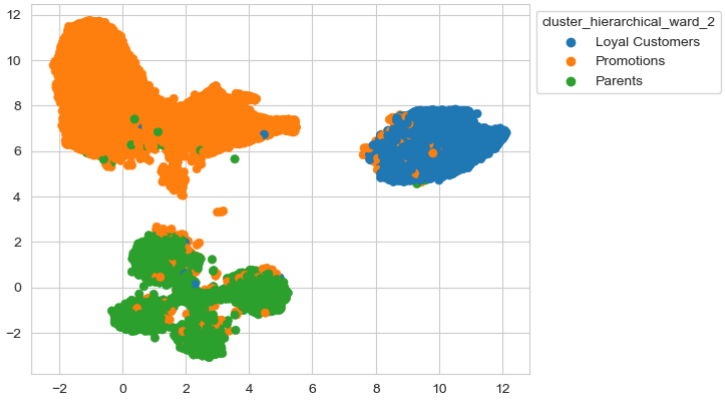


*Figure 13 UMAP of Hierarchical Clustering (remaining observations)*

Despite this still not being a perfect solution, we will finalize these clusters as it is the best solution we have been able to find.

## *Feature Importance*

We used a Random Forest to estimate which of our variables have more impact in the separation of our clusters.

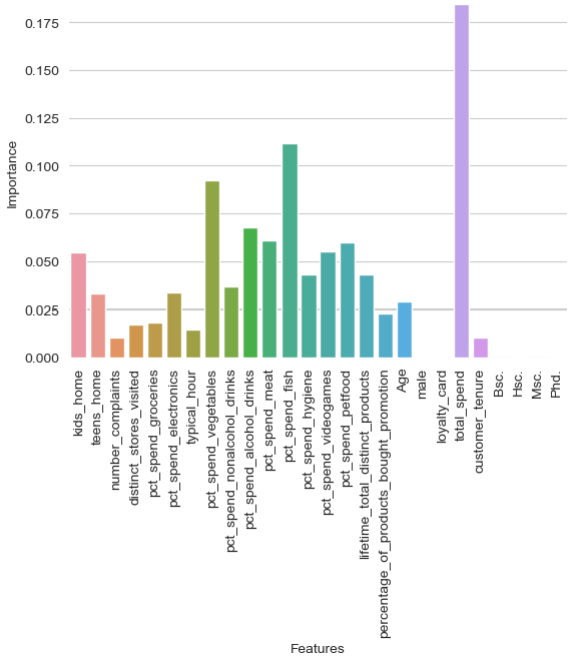


*Figure 14 Feature Importance*

We can see that the importance is relatively balanced, with a couple of spikes, in total_spend and pct_spend_fish. This can be attributed to the 'Fishermen' cluster which has the highest total_spend of all clusters and has an average pct_spend_fish greater than 70% but is the smallest cluster. We can also see that the gender and loyalty card variables have essentially no importance, as well as the education variables, which makes sense as these education variables weren't used for clustering. The purchase behaviour variables show a decent amount of importance, especially vegetables, which also makes sense from the vegetarian's cluster.

## Cluster Analysis

From the analysis of the clusters these are the main characteristics for each:

*Fishermen*: Spent 70% or more on fish, only one visited store, basically no children, not a lot of registered complaints, large majority has loyalty card, buy with a promotion 40% of the time.

*Gamers:* High percentage spent on videogames (10%) and electronics (20%), have some children, not a lot of registered complaints, generally only visit one store, similar average age to other clusters (55 years old).

*Pet Lovers*: High percentage spent on pet food (47%), very low number of different products (only 52), don't have children, not a lot of registered complaints, generally only visit one store, similar average age to other clusters.

*Young Party People*: High percentage spent on drinks (10% in non-alcohol and 39% in alcohol ones), don't have children in general, not a lot of registered complaints, only one store visited generally, lowest average age of all clusters (24 years), relatively large percentage spent on hygiene products (9%).

*Vegetarians*: High percentage spent on vegetables (16%) and groceries (50%), very low percentage spent on meat and fish (less than 1%), have young children and in some cases teens, not a lot of registered complaints, one visited store in general, similar average age to other clusters.

*Young, lots of electronics:* High percentage spent on electronics (26%), rare cases where they have kids and teens, essentially no complaints, visit one store in general, lowest average customer_tenure (only 7 years), second lowest average age (30 years), highest level of education.

*Loyal Customers:* Highest average tenure (15 years), highest percentage of customers with a loyalty card (80% of the group), similar average age to other clusters, highest number of different products bought (1071 different ones), have children and teens in general, visit more different stores in general than the average, high number of registered complaints (Karens…).

*Promotion lovers:* Buy with promotions 40% of the time, high percentage of customers with loyalty card (67% of the group), similar average age to other clusters, balanced expenses across categories, high number of registered complaints, visit multiple stores in general, have teens and kids in some cases.

*Parents:* Have a lot of kids and teens, high number of complaints (Karens…), one store visited in general, high expense on electronics (14%), reasonably balanced expenses, visit stores after work(5/6pm), highest average age.

# Targeted Promotions

After completing the initial stages of exploratory data analysis and customer segmentation, the next step involved crafting association rules. This process facilitated the formulation of tailored campaigns aimed at distinct client segments.

In summary, Association rules are patterns found in data that reveal relationships between items. They consist of "if-then" statements that indicate the likelihood of one item or event occurring given the occurrence of another. These rules are derived from datasets, often transactional data, and are measured using metrics such as support, confidence, and lift. Association rules help businesses understand customer behaviour, identify product associations, and optimize marketing strategies for targeted campaigns and personalized recommendations.

*Fishermen*

We had 905 transactions made by the fishermen, we considered 2 possible sets of rules, one with products that appear at least 12% of the time and 5% of the time, appearing at least 50% of the time together. After looking at the results we see that:

- They mostly bought fish products.
- Shrimp is bought with almost everything, meaning that they go to the place which has the best price on shrimp.
- Tuna and Salmon appear together often in different transactions.

With this information the following three targeted campaigns were suggested:

1. **Fish in the morning? Why not?:** 10% discount on fish products in purchases before 1pm.
2. **One Love Tuna and Salmon:** 10% discount off on fresh tuna and salmon if they are bought together.
3. **Shrimp Lovers:** 5% discount if you bought shrimp.

Promotions are delivered using the loyalty card.

*Gamers*

We had 12873 transactions made by the gamers group, we considered 2 possible sets of rules, one with products that appear at least 15% of the time and 8% of the time, appearing at least 50% of the time together. After looking at the results we see that:

- They mostly bought electronic products.
- These people bought laptops, Samsung galaxy or iPhone.
- Champagne is often bought with electronics.
- Headphones are often bought with a phone.

With this information the following three targeted campaigns were suggested:

1. **Catalogs** (digital or physical) about new games, phones or laptops sent to these customers.
2. **Lovers Game Night:** 50% discount on champagne with a purchase of a laptop, Samsung galaxy or iPhone.
3. **No one has to listen your music:** 30% discount on Bluetooth headphones if you bought a Samsung galaxy (the same with airpods and an iPhone).

*Pet Lovers*

We don't have any registered transactions made by this type of customers, but we know that they don't have children, spend the majority on pets' products, so we recommend the following promotions and strategies to encourage purchases in this group:

1. **Catalogs** (digital or physical) about toys, homes and food for pets (being generic about the type of pet) sent to these customers.
2. **One Love Human's Best Friends:** 10% discount on pet products on purchases higher than 50€.
3. **Spoil Human's Best Friend:** Buy 2 toys for your pet and get 1 free.

Promotions are delivered in ad campaigns and tv commercials.

*Young Party People*

We had 2255 transactions made by the young party people group, we considered 2 possible sets of rules, one with products that appear at least 10% of the time and 5% of the time, appearing at least 50% of the time together. After looking at the results we see that:

- They mostly bought alcohol products.
- These people bought mostly cider, white and dessert wine, and beer.

With this information the following three targeted campaigns were suggested:

1. **For the Thirsty People:** 10% discount on alcohol and non-alcohol products on purchases higher than 50€.
2. **"We don't have money" Alcoholics**: 15% discount on beer on the purchase of cider.
3. **For that special night**: 20% discount on dessert wine and white wine if bought together.

Promotions are delivered in ad campaigns and loyalty card.

*Vegetarians*

We had 16229 transactions made by the vegetarian's group, we considered 2 possible sets of rules, one with products that appear at least 15% of the time and 5% of the time, appearing at least 50% of the time together. After looking at the results we see that:

- They mostly bought vegetable products.

- These people bought asparagus, tomatoes, carrots and mashed potatoes with everything.
- Melons are bought normally with the other products.

With this information the following three targeted campaigns were suggested:

1. **One love Vegetables**: 10% discount on vegetables products on purchases higher than 30€.
2. **Our Special Basket for veggies**: 15% discount on melons and mashed potatoes on purchase of 1 Kg (combined) of tomatoes and asparagus.
3. **If U love veggies we are here for U**: 10% discount on asparagus, tomatoes, carrots and mashed potatoes.

Promotions are delivered using the loyalty card and in ad campaigns.

*Young Electronics People*

We had 5725 transactions made by the young electronics group, we considered 2 possible sets of rules, one with products that appear at least 10% of the time and 3% of the time, appearing at least 50% of the time together. After looking at the results we see that:

- They buy candy in their basic baskets.

- These people bought mostly basic products like oil, napkins.

We suggest the following promotions and strategies to encourage purchases in this group:

- Separate the oil and cooking oil from the napkins as much as possible, to **make them to walk and see more products.**
- To encourage the purchase of other candies, **put stands with candy bars and gummies near the checkout and the self-checkout.**
- **Cakes and Oil**: 20% discount on cake if they bought oil (or cooking oil).
- **One Love Electronics**: 10% discount in electronic products on purchases higher than 20€.

Promotions are delivered in loyalty card and in ad campaigns.

*Loyal Customers*

We had 11209 transactions made by the loyal customers group, we considered 2 possible sets of rules, one with products that appear at least 10% of the time and 3% of the time, appearing at least 50% of the time together. After looking at the results we see that:

- They buy candy and soup in their basic baskets.
- These people bought mostly basic products like oil, napkins.

We suggest the following promotions and strategies to encourage purchases in this group:

- Separate the oil and cooking oil from the napkins as much as possible, to **make them to walk and see more products.**
- To incentivize the purchase of other candies, put stands with candy **bars and gummies near the checkout and the self-checkout.**
- **Soups and Oil**: 20% discount on soup if they bought oil (or cooking oil).
- **To the Sweet tooth**: 20% discount on candy bars if they bought cake.

Promotions are delivered using the loyalty card and tv commercials.

*Promotion lovers*

We had 24536 transactions made by the promotion lovers customers group, we considered 2 possible sets of rules, one with products that appear at least 12% of the time and 6% of the time, appearing at least 50% of the time together. After looking at the results we see that:

- These people are highly sensitive to promotions.

- There are people that buy only basic products as napkins and oil, and people that buy alcohol (cider, beer and wine).

We suggest the following promotions and strategies to encourage purchases in this group:

- To encourage higher purchases: 20% of discounts on purchases higher than 100€.
- **If U love it, we are here for U**: Specific 10% discount on a certain product category (pets, vegetables, meats, fish, electronics, video games, alcohol and non-alcohol and hygiene).
- **Surprise the wife**: 20% discount on dessert wine and white wine if bought them together.
- **Surprise the girlfriend**: 15% discount on beer on the purchase of cider (apply the same discount to white wine and dessert wine).
- To incentivize the purchase of other candies, put **stands with candy bars and gummies near the checkout and the self-checkout.**

Promotions are delivered using the loyalty card and tv commercials.

*Parents*
We had 905 transactions made by the loyal customers group, we considered 2 possible sets of rules, one with products that appear at least 15% of the time and 8% of the time, appearing at least 50% of the time together. After looking at the results we see that:

- These people buy basic products like oil and candies and baby food.
- These people have many children.

We suggest the following promotions and strategies to encourage purchases in this group:

- **One toy is not enough**: buy 1 get 1 at 50%.
- Separate the oil and cooking oil from the baby food as much as possible, to **incentivize them to walk and see more products.**
- To incentivize the purchase of other candies, **put stands with candy bars and gummies near the checkout and the self-checkout.**

Promotions are delivered using the loyalty card and tv commercials.

# Conclusion

The objective of this project was to effectively cluster customers based on similarities. To achieve this objective, we started with an initial analysis aimed at gaining deeper insights into the dataset at hand.

Based on the analysis of the means of each independent variable within the clusters derived from the final clustering solution, the customer segments were assigned the following labels: "Fishermen", "Gamers", "Pet Lovers", "Promotions lovers", "Young Party People", "Parents", "Vegetarians", "Young lots of electronics", and "Loyal Customers".

Having the clusters defined, an analysis on customer's basket was done to perform Association Rules.

To conclude, if we were to choose which customers should be targeted first and which strategies to be considered more seriously, the order will be the following:

1. **Fishermen:** because they are easy to contact, and they spent a lot in our store.
2. **Gamers:** because they are the second most profitable clients.
3. **Pet lovers:** because they are easy to satisfy, and they are the next in the list of expending more.
4. **Promotions (and young parties):** because they have rules from the young party's group and are the most sensitive to promotions.
5. **Parents:** because the strategies don't imply ads (only 1 but it could be omitted).
6. **Vegetarians:** All promotions to compete for that market group.
7. **Young lots of electronics:** only the strategic decisions and the promotion on electronics.
8. **Loyal customers:** Nothing, they won't leave because changing habits is hard.

# References


- Theoretical class presentations
- Practical class notebooks