CUSTOMER SEGMENTATION

Executive Report

Ana Farinha 20211514
Ana Reis 20211631
Beatriz Barreto 20211547

Machine Learning II | June 2023

Table of Contents

Executive Summary *3*

Exploratory Data Analysis *3*

Customer Segmentation *10*

Algorithms 11

DBSCAN 11

Mean Shift 11

K-Means 11

Bisecting K-Means 12

Hierarchical Clustering 12

Comparing clustering solutions 12

Cluster Description 14

Targeted Promotion *19*

Cluster 0 - ‘Thrifty’ 19

Cluster 1 - ‘Young and Unattached Customers’ 20

Cluster 2 – ‘Parents’ 20

Cluster 3 – ‘Meat and Alcohol Connoisseurs’ 21

Cluster 4 – ‘TechSip Gamers’ 22

Cluster 5 – ‘Fish Enthusiasts’ 22

Cluster 6 – ‘Minimalist Green Shoppers’ 23

Conclusion *24*

References *25*

Annexes *25*

Annexe 1 – Plots from notebook ‘’1_EDA” 25

Annexe 2 – Plots and tables from notebook ‘’2_Model” 28

Annexe 3 – UMAPs from notebook ‘’3_Clusters” 34

Annexe 4 – Box plots to characterize each cluster from notebook ‘’3_Clusters” 35

Executive Summary

This project aimed to help a business gain valuable insights about their customers and tailor targeted marketing strategies that maximize customer engagement and loyalty. Those objectives are obtained through many processes in which stand out two steps. The first is the identification of relevant customer segments that represent a specific customer profile and will enable the development of targeted marketing strategies. The other is the analysis of customer behaviour by cluster that allows the characterization of segments to facilitate the creation of marketing strategies through association rules.

To organize our project we used Jupyter Notebook and GitHub. The methodology involved many little steps that we separated into four main ones. The first is Exploratory Data Analysis where we made data-related customer information go through pre-processing that involved removing duplicates, handling missing values, identifying anomalies and outliers, feature engineering, data visualization and calculating a correlation matrix. Next, many clustering algorithms were applied and exported in a new dataset for further analysis. Then, to choose the solution that better segmented our data we identified the qualities of each cluster and made a customer profile for each one. Finally, we performed association rules using the apriori algorithm and used plots to visualize our rules based on 2 quality measurements.

In the end, we discovered that seven was the ideal number of clusters and that some of them were very well-defined. While analysing each one we discovered singularities in each one that aided us in naming each cluster. The most well-defined cluster ended up being formed by the customers that spend a lot of money on fish.

We ended up recommending five different promotions for each cluster based on their descriptions and the association rules. The goal we had in mind was for customers to view this business as something they can rely on for every one of their shopping needs.

Exploratory Data Analysis

Exploratory Data Analysis, also known as EDA, is a crucial step in data analysis that involves examining and understanding the data, identifying patterns and trends, detecting outliers and anomalies and exploring relationships between variables.

In this part of our project, we used the dataset called ‘customer_info’ which contained information about each customer and their behaviour for the past two years. The information mainly consisted of demographic data and money spent in different categories.

After importing the dataset, we checked that each column had a unique ‘customer_id’ and used that as an index for our data frame. Our next step was to see the data types of our variables and apply changes to those that needed it. Those changes involved converting 2 variables (an object and a float) to datetime objects and transforming 6 variables that were identified as floats into integers as they were discrete variables. Following, we separated categorical from numerical variables.

After, we identified the variables that had missing values. The only variable that had missing values was ‘loyalty_card_numbers’ which meant that the customer did not have a loyalty card. Next, we saw some descriptive statistics of our variables and some details stood out. The first one was the existence of infinite and missing values in ‘typical_hour’ and 'lifetime_spend_videogames'. The second one was the fact that 'kids_home' had a maximum of 10 which meant that at least a customer was raising 10 kids. Thirdly, the minimum spent on vegetables was 1 and the minimum percentage of products bought in promotion is only about 2%. Fourthly, the maximum spent on a category was 36243 on fish.

Following, we identified and treated some anomalies. The first one was related with the minimal age of each customer when they did their first transaction. In Portugal, there are no restrictions related to age when we want loyalty cards, however most companies only make them available to customers above 16. With that in mind, we decided to implement that as a rule. So, when the customer did not have the required age, we made their year of birth 16 years before the year of their first transaction. The next anomaly that we treated had already been identified. It was the infinite values in 'typical_hours'. As there were only 2 cases, we decided to drop those rows. With that anomaly treated, we proceeded to transform the variable into a datetime object. Following, we treated the other identified anomaly related to infinite values in 'lifetime_spend_videogames'. As there were 24 customers with this anomaly, we decided to replace it with 0.

After identifying and treating anomalies, we decided to perform some feature engineering. Firstly, we created a much-needed feature of 'age'. Our second one was the extraction of the education level from 'customer_name'. Thirdly, we transformed 'customer_gender' into a binary variable called 'female' in which 1 meant female and 0 meant male. Fourthly, we created a variable called 'cust_per_card' that has the total number of customers that use the same loyalty card. This means that if the value is 0, the customer does not have a loyalty card. If the value is 1, the customer is the only that uses that loyalty card. As for 2, it means that 2 customers use the same loyalty card. Following this, we created the variable 'total_lifetime_spend' which is the sum of the money that the customer spent in all categories. Lastly, we created a variable that had the number of dependents that each customer had which was the sum of the number of kids and the number of teenagers. Below, we have a simple list with our new variables:

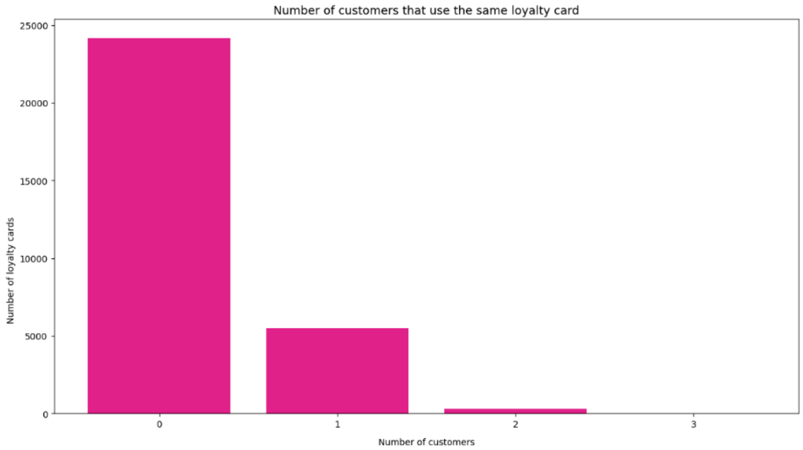
- 'age': Age of the customer.
- 'education': Level of education.
- 'female': Gender where 1 is female and 0 is male.
- 'cust_per_card': number of customers that use the same card as the customer (included).
- 'total_lifetime_spend': Total value spent by the customer on all categories.
- 'number_dependents': Total number of kids and teens at home.

Our next step in the EDA process was to identify outliers. We used the interquartile range method but there were too many outliers. So, we decided to see only the extreme outliers. Unfortunately, we barely reduced the number of outliers. Because of that, we decided to use plots to identify them.

Data visualization uses charts, graphs, and maps to communicate data patterns, relationships, and trends in a more instinctive way to comprehend them. It's extremely helpful in making the data more understandable and easier to explore and analyse. It can also help to identify details that might have been missed or overlooked in previous explorations as well as identify outliers. It allows more informed decision-making.

The first thing we did was plot a bar chart with the number of customers who use the same loyalty card using the variable 'cust_per_card'.

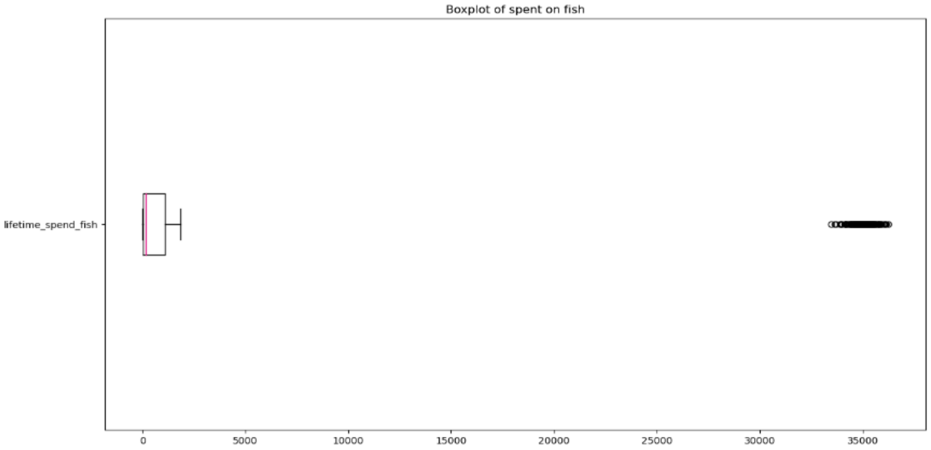
Figure 1 – Number of customers with loyalty cared (shared or not)



Above, we can see that the majority of customers shop without a loyalty card. For the customers who have a loyalty card, it is usually only used by themselves.

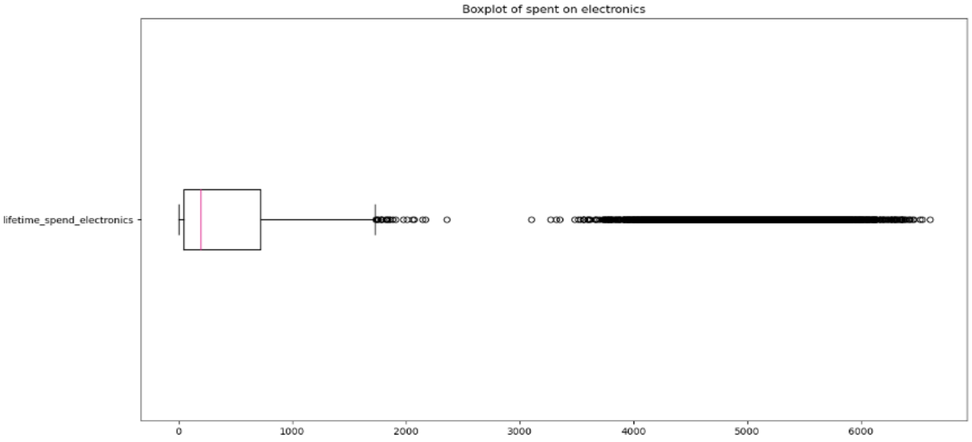
After that, we decided to plot some boxplots for all the variables that had 'lifetime_spend' in the name so that we could identify outliers. The ones that stood out were 'lifetime_spend_fish', 'lifetime_spend_electronics' and 'lifetime_spend_groceries'.

Figure 2 – Boxplot of lifetime spent on fish



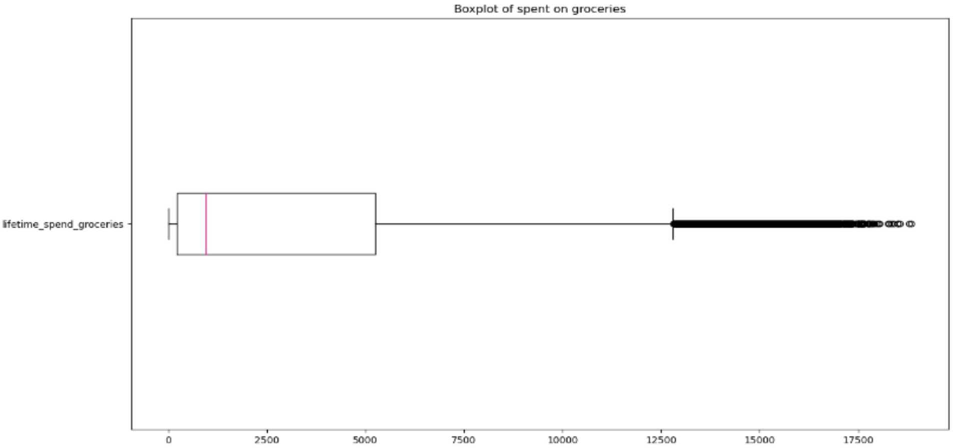
In this first boxplot, we can see that most customers don't spend much money on fish. However, there is a clear group of them that expended a lot of money on fish. This group could be a possible cluster in our customer segmentation.

Figure 3 – Boxplot of the lifetime amount spent on electronics



In the boxplot above, we can see the same phenomenon that happened in the other one. Most customers don't spend much money and a clear group spends a lot which could be a cluster.

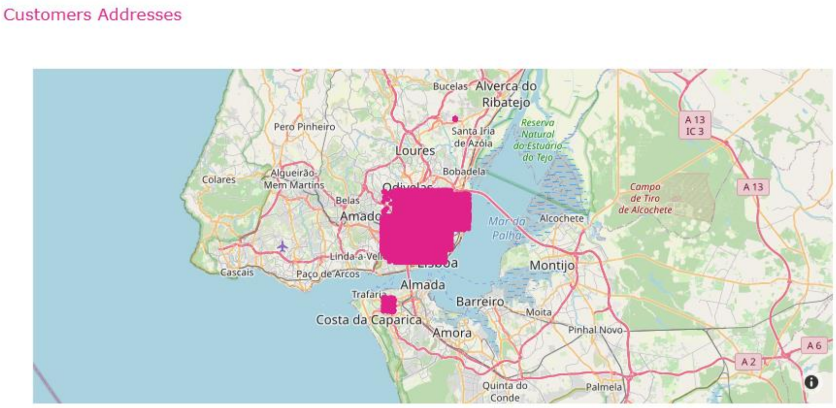
Figure 4 – Boxplot of lifetime spent on groceries



As for this boxplot, we can see that there are some outliers. However, they are really near the other ones so it probably is not a cluster but it is something that we should pay attention to in the future.

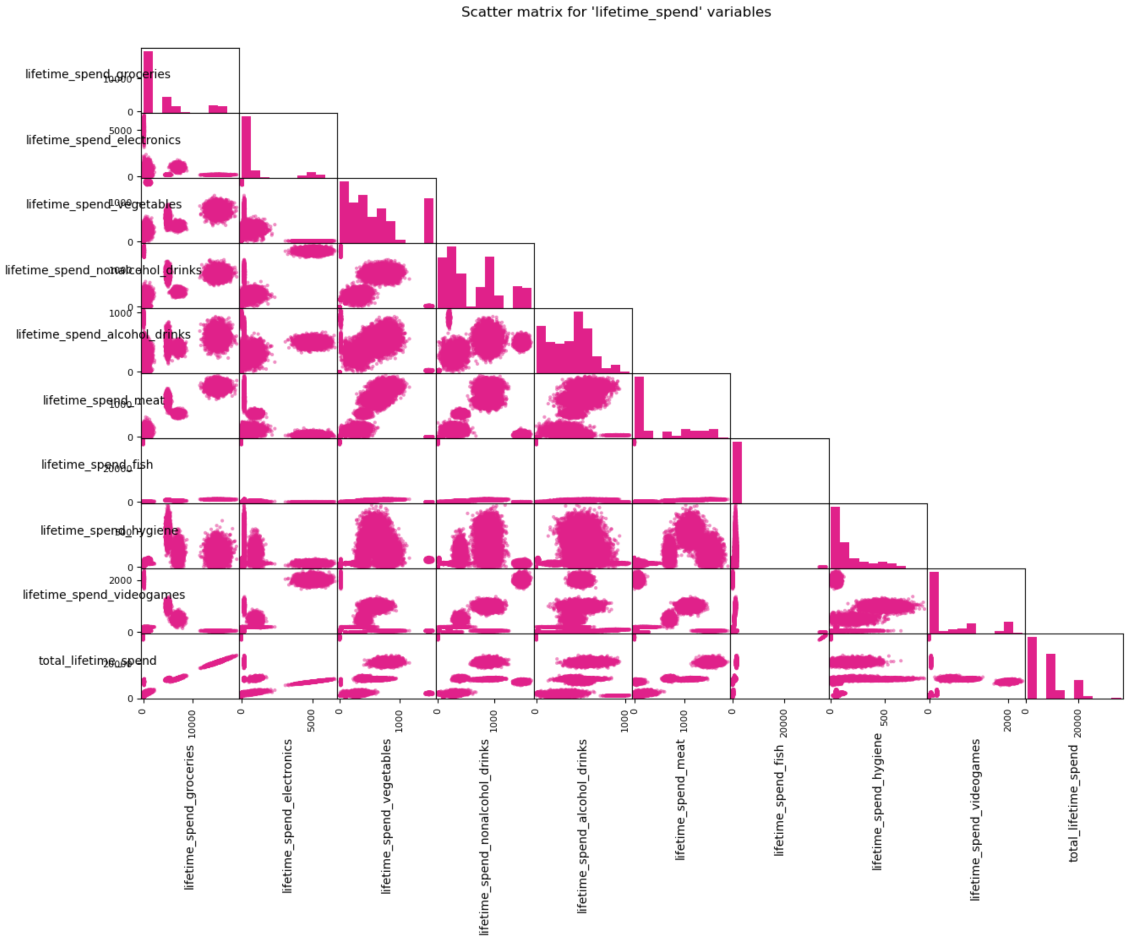
Then we decided to map out the customers' residences. Firstly, we did a scatterplot so we could see the range of coordinates. Afterwards, we plotted the map and saw that all the customers reside in the Lisbon Metropolitan Area. By interacting with the map, we discovered that there were a group of customers from Mercado Abastecedor da Região de Lisboa (MARL) and a group that resides around Nova FCT.

Figure 5 – Map of Lisbon's Metropolitan Area



Our next step was to use the same variables we used in the boxplots to create a scatter matrix with the distribution of the variable on the diagonal. A scatter matrix is useful to visualize relationships between variables. It can help find potential correlations, patterns, subgroups with the same behaviour in the data and outliers. There we can see the distribution of each variable and analyse the relationship between two different variables.

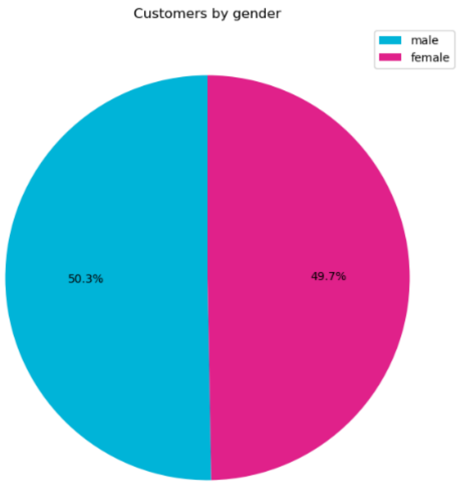
Figure 6 – Scatter matrix



By analysing these plots, we can see that there is a group of customers that spend a lot on vegetables and one that buys significantly more fish than the rest. By examining the relationship between ‘lifetime_spend_electronics’ and ‘total_lifetime_spend’ we can see that there are different defined groups but that overall electronics is not a big contributor to the total spent. Other things we can see are the fact that ‘total_lifetime_spend’ and ‘lifetime_spend_groceries’ almost have a linear relation, as one increases so does the other. The same can be said for ‘lifetime_spend_meat’ and ‘lifetime_spend_groceries’. By looking at ‘lifetime_spend_electronics’ and ‘lifetime_spend_videogames’ we see that customers either don’t spend a lot of money in either or spends in a lot in both.

Next, we used a pie chart to see if there were any discrepancies in our customers’ gender.

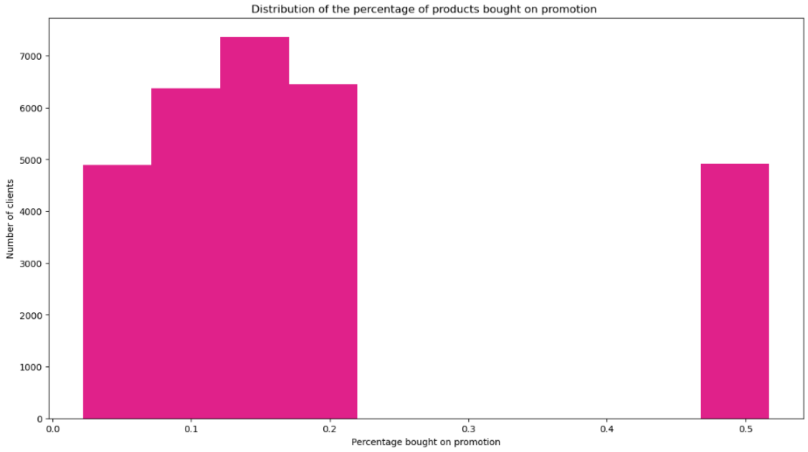
Figure 7 – Gender pie chart



Above, we can see that there is no significant difference in our customers' gender. So, we can conclude that our dataset is balanced in terms of gender.

After, we decided to plot some bar charts using the variables 'year_first_transaction', 'number_dependents', 'typical_hour', 'number_complaints' and 'distinct_stores_visited'. Then, we plotted a histogram and a boxplot for 'percentage_of_products_bought_promotion' and observed that the majority of customers bought less than 20% of their products on promotion. However, there was a group that bought about 50% on promotion.

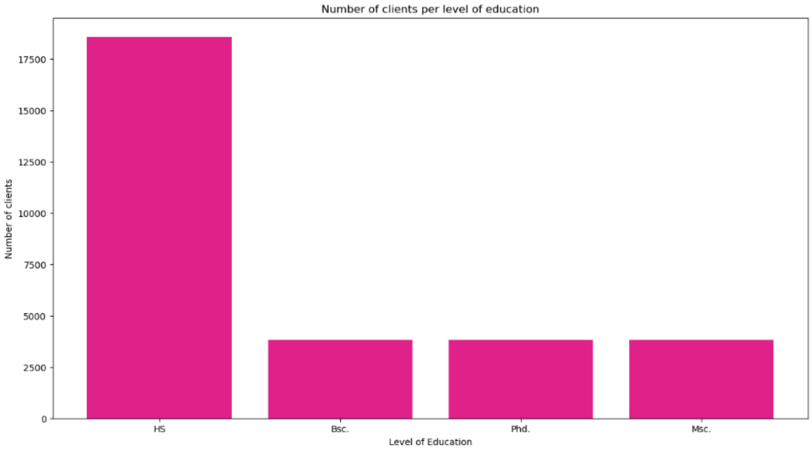
Figure 8 – Histogram of percentage of products bought on promotion



The same graphs were plotted for 'lifetime_total_distinct_products' and we noted that in general customers bought less than 1000 different products.

After this, we plotted a bar chart for the variable education where we discovered that most customers don't have higher education.

Figure 9 – Level of education

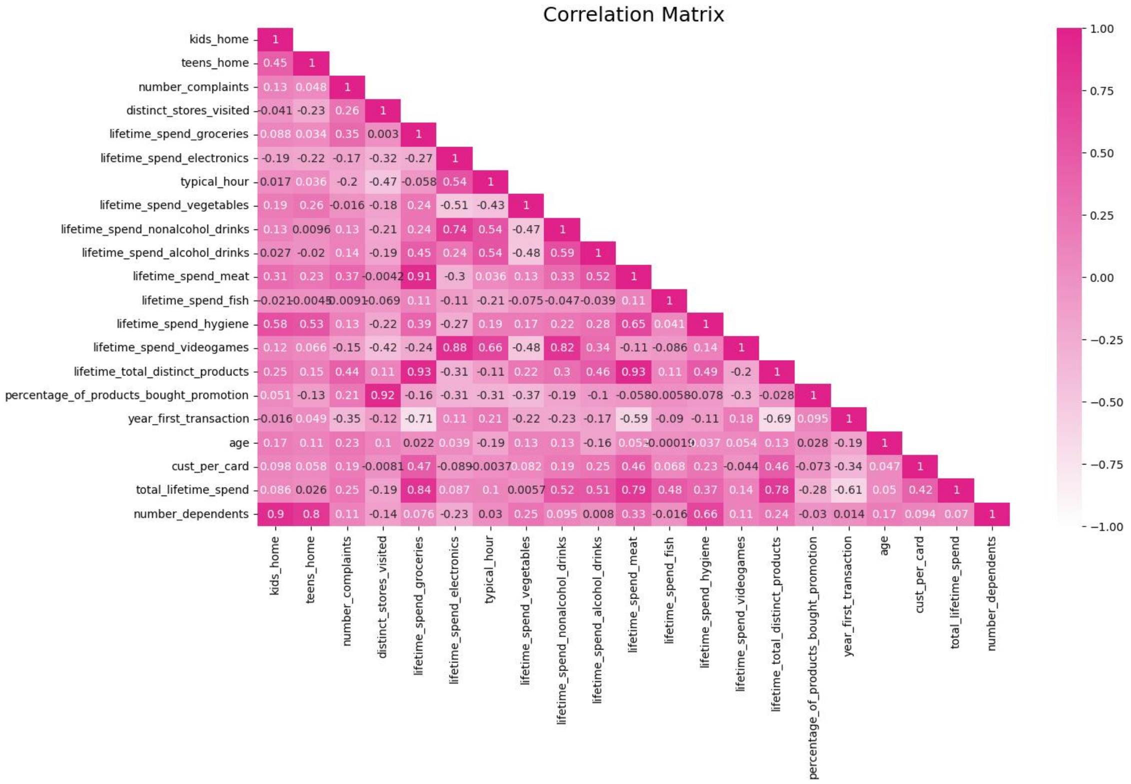


Our next step was to create a Correlation matrix. It allows a better understanding of relationships between variables. By plotting the correlation matrix, we saw that the following pairs of variables had high correlation values (bigger than 0.7 or smaller than -0.7):

- nonalcohol drinks and electronics
- meat and groceries
- videogames and electronics
- videogames and nonalcohol drinks
- distinct products and groceries
- distinct products and meat
- percentage bought on promotion and stores visited
- year first visited and groceries
- total spent and groceries
- total spent and meat
- total spent and distinct products
- number of dependents and teens
- number of dependents and kids

These last two were expected since the number of dependents is made from summing teens and kids. The same can be seen for the variable that refers to the total spent and the ones used to make it.

Figure 10 – Correlation Matrix



Almost at the end of this notebook, we decided to explore the variable ‘customer_name’. To do that, we saw if for the same last name and we found a pattern. The most noticeable was the one from Supermarket in which the mean and the maximum were the same for many variables.

Lastly, we exported our dataset with the treated data.

Customer Segmentation

Customer Segmentation is the name given to the process of dividing a company’s customers into different groups, also called segments. The customers in each segment will have similarities with the members of their groups. This process aims to give a better understanding of different types of customers, in order to target them with products, services, and tailored marketing campaigns for each one of these groups.

To carry out the customer segmentation, we started by importing into our notebook the data treated in the EDA. After, we created a new data frame without the variables “loyalty_card_number”, “latitude”, and “longitude” since they were irrelevant to our analysis. The first one was dropped because we had a lot of missing values (N/A) for the “loyalty_card_number”, considering that it is not obligatory to have a loyalty card to make purchases, and was already accounted for in ‘cust_per_card’. We also dropped the variables that gave the coordinates, because we thought that these could lead to discriminatory segmentations (for example, socioeconomic profiling). Next, we proceeded to check the variance of the variables since we learned that a variable whose variance was equal to zero would not provide useful information for our modelling. Not having found any, we proceeded to scale the data, using the standard scaler, and testing the algorithms.

Algorithms

Clustering Algorithms are very powerful tools used for data analysis. These algorithms group similar data points by identifying structures and patterns hidden in the data. Next, we will describe the algorithms and its implementation.

DBSCAN

The first algorithm we tested was DBSCAN which is a density-based clustering algorithm. It is commonly used to handle outliers as well as create segments. When implementing it, one needs to define the maximum distance to consider a neighbour point and the number of points to consider a point a 'core point' of a cluster solution.

The way it works is by randomly selecting an unvisited data point. Next, if the number of points within the distance given is greater or equal to the minimum number of neighbours needed to be a 'core point', then it is marked as a 'core point'. If the point is called by a 'core point' but does not have a specified number of neighbours, it is considered a 'satellite point'. If the point is never called by another, it is an outlier. All the data points within reach of the 'core point' are assigned to the same cluster. The process is repeated until all data points are visited.

We started with this algorithm because we wanted to see if we could identify outliers with it. However, after a few tests, we found that the results given were not satisfactory, independently of the changes we made to the parameters. We either got cluster solutions with almost no outliers, but very few clusters where one of them had almost 20000 observations, or we, reducing the distance and the number of observations needed to be considered a 'core point', got a solution with more than 15 clusters and more than 4000 outliers.

As we found the results, unsatisfactory we proceeded to test another algorithm.

Mean Shift

The Mean Shift is also a density-based clustering algorithm typically used for image processing and data analysis. This algorithm works by first defining the kernel function, initializing the data points, and defining the bandwidth, which is the size of the neighbourhood around each data point. We start by selecting random points, which are our centroids, and shift them towards the mean of the points in the higher-density region in their window. This process is repeated until there is no region of higher density for each centroid. Then the points are assigned according to their distance to each centroid.

We tried to implement this algorithm but found our results to be unsatisfactory, just like in the DBSCAN. After that, we concluded that density-based clustering algorithms where not the best solution for our data.

K-Means

The K-means is an unsupervised learning algorithm used for clustering that aims to minimize intra-cluster distance and maximize inter-cluster distance. It works by dividing a dataset into a pre-defined number of clusters using distances. It does so by assigning each data point to the nearest cluster and then re-computes the cluster centroids using the mean of all observations in a cluster. This process is repeated until one of the following cases happens, either the centroids stop shifting or the maximum number of iterations is reached.

The first thing we did to implement the K-means was to find the ideal number of clusters for our dataset. To do so we plotted the 'elbow curve' graph and concluded that we should have seven clusters. Next, we implemented 2 K-means, one with all our variables and another without the binary variable 'female', and after we compared our solutions. We obtained very similar results in both as most of our clusters had the same number of customers with similar means for each variable.

Bisecting K-Means

The Bisecting K-Means is a variant of the K-Means that is explained above. The algorithm halves clusters into sub-clusters until the selected number of clusters is achieved. It starts with all the observations in one cluster, then splits into 2 equal-sized clusters and calculates the sum of square errors (SSE). After, the algorithm bisects the cluster with the higher SSE, and the SSE is calculated again for all clusters. This continues until the number of clusters is reached.

In our implementation, we once again did an 'elbow curve' graph to decide the ideal number of clusters, and just like in the other k-means we decided on 7. Our solution gave us similar clusters to the other k-means solutions with 2 clusters that are the same in each cluster solution (same number of customers and similar means of each variable).

Hierarchical Clustering

Hierarchical clustering is an unsupervised algorithm that builds a hierarchy of clusters by merging or splitting clusters using distances. There are 2 types of hierarchical clustering: Agglomerative Hierarchical Clustering and Divisive Hierarchical Clustering. Agglomerative Hierarchical Clustering begins with all the data points as their own cluster and joins them until they are only a cluster. Divisive Hierarchical Clustering does the opposite, it starts with all data points in a single cluster and divides it repeatedly until all data point becomes their own cluster.

In our pipeline, we decided to use Agglomerative Hierarchical Clustering with 2 different ways to connect the elements: the Single or Minimum Linkage and Ward's method.

The first starts with each data point as an individual cluster, then it computes the distances between all pairs of clusters and joins the 2 closest clusters into a single cluster. Next, it computes the distances between the new and remaining clusters. This happens until all data points belong to only a cluster.

The latter calculates the distance between all pairs of clusters and the sum of squared differences (SSD) of each pair. After, it compares it with the SSD of each cluster and finds the pair where the increase between the two values of SSD is minimal, merging that pair into a new cluster. This happens until all data points belong to only a cluster.

After implementing both solutions and plotting their corresponding dendrograms, we found that the ward's method gave us the best results. The minimum linkage solution for 7 clusters, put around 20000 observations into a cluster and left 3 clusters with only one observation each. The results of the ward's method clusters were more similar to our k-means segments.

Comparing clustering solutions

Following the implementation of the clustering algorithms we had to choose what we considered to be the best solution. To do that we started by remembering some insights we had discovered in the EDA. The most obvious one was that we had to have a cluster of clients that spent a very large sum of money on fish, and this was the only category where they spent a significant amount of money.

Taking that into consideration, we had 4 solutions where this happened: The 2 k-means implementations, the bisecting k-means, and the ward's hierarchical. Our next step was to look at the tables that compared the observations present in each cluster to 2 given solutions. With this, we found out that, for our solutions of k-means, including or not the variable ‘female’ would make minimal changes to our solution, so we decided to exclude the solution that did not take it into consideration and keep the one that took into account more information about our customers.

After, we made the same comparison between the K-means and the Bisecting K-means. We concluded that both solutions had created 2 clusters that were exactly the same in both. This made us believe that these 2 clusters were well-defined since they could be easily found.

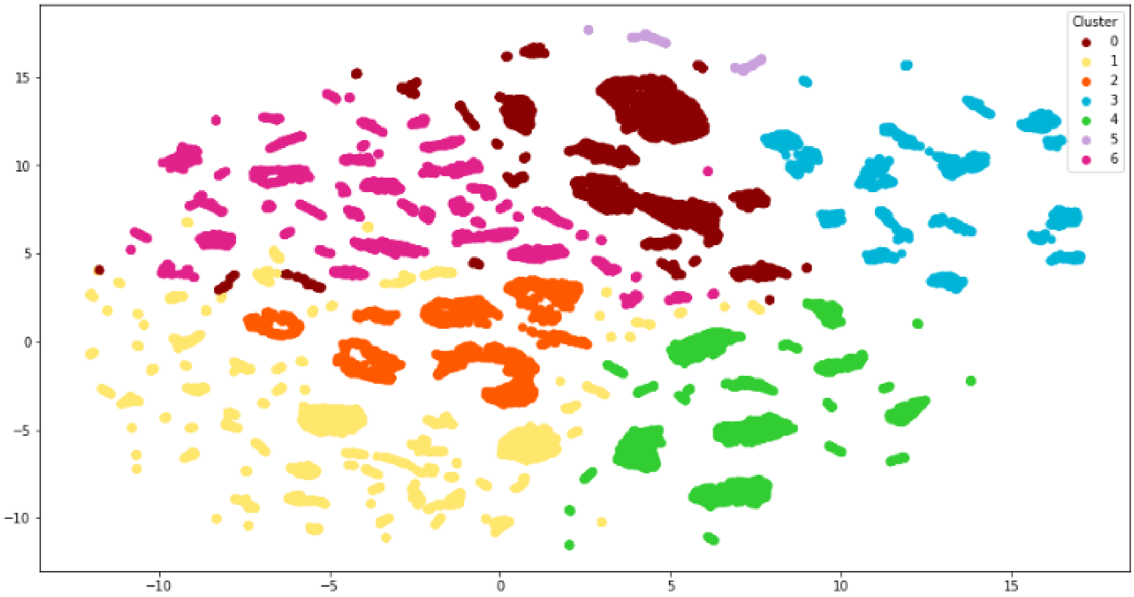
We also made these comparisons between these solutions and the one given by the Ward’s Hierarchical. We found out that the 2 clusters that were equal in the other solutions were also present in this. After these comparisons, we thought that we could affirm that we had at least 2 clusters that were certainly well-defined.

Because these comparisons only told us that our solutions were similar and we wanted to choose one to further analyse, we decided to create a new notebook where we started by using UMAP to compare the remaining 3 different customer segmentations which were defined by the following unsupervised algorithms: k means, bisecting k-means and hierarchical using ward linkage.

Before proceeding with the comparison, it is important to know that UMAP is a dimension reduction technique that can be used to visualize high dimensional data. It has two main parameters: the number of nearest neighbours (controls balance between local and global structure in data) and the minimum distance (defines the minimum distance between points in the low dimension representation).

At the beginning of the comparison between the three visualizations, we noticed that they were quite similar, as we expected from the conclusions in the previous notebook. As such, we decided to use the clusters obtained by the k-means algorithm because there was practically no point that was not near one from the same cluster. The only situation that is important to notice were the existence of small groups of points distant from the others of the same cluster. Below, we can see a figure of the UMAP of the chosen algorithm.

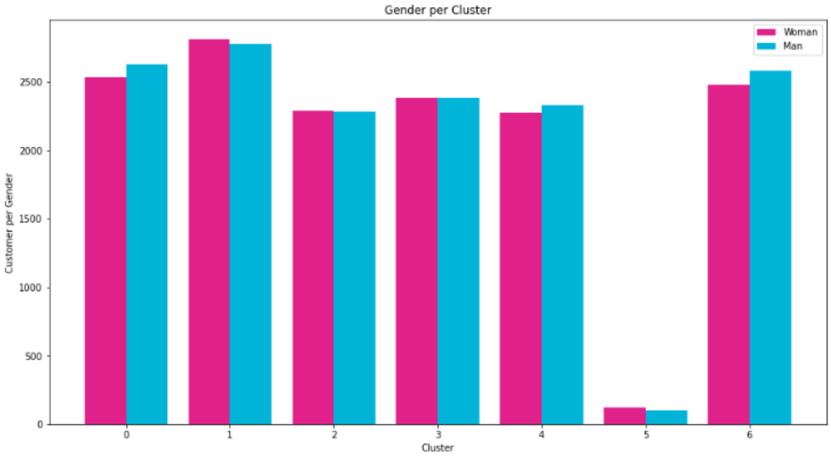
Figure 11 – UMAP of k-means



Cluster Description

Now that we had our clusters defined and chosen, we started analysing the characteristics of each one. Previously, we had concluded that our dataset was balanced in terms of gender. So, we decided if the same succeeded with our clusters. Below, we can see a bar graph that presents the number of women and men per cluster.

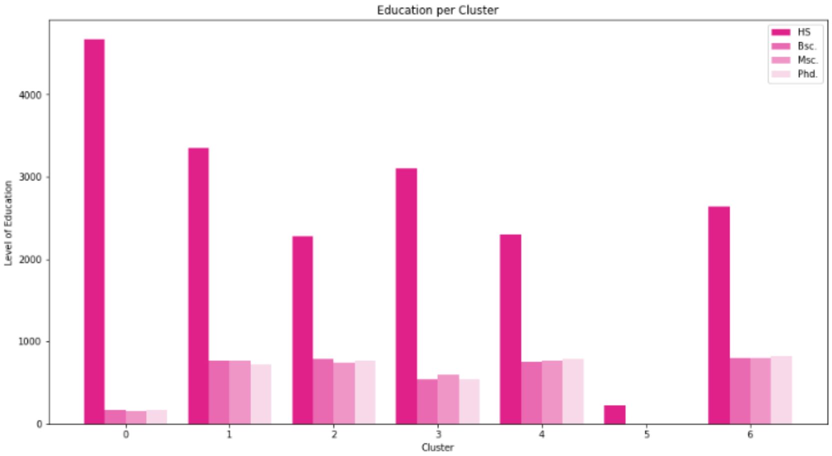
Figure 12 – Gender distribution in each cluster



Here, we can see that our clusters are almost balanced with some having a little more women customers and other men. We can also see that cluster 5 is really small compared to the others, which is something that we should pay attention to for future reference.

Our next step was to see the level of education by cluster keeping in my that we already concluded that most customers only had a high school diploma.

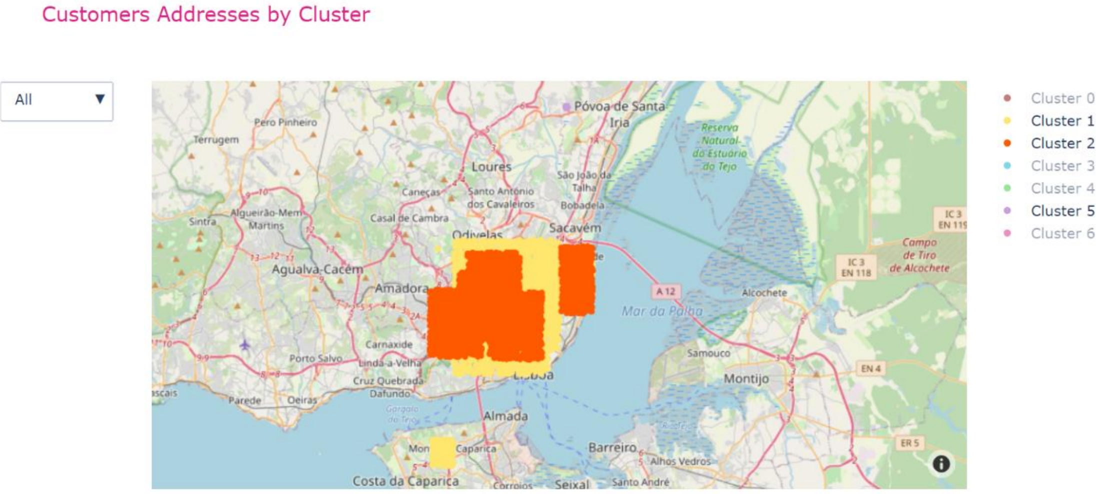
Figure 13 – Level of education per cluster



In the plot above, we can see that cluster 5 is composed only of customers with high school diplomas. As for the remaining clusters, they have most education levels balanced except for the one that refers to high school which is higher for all. The cluster where this difference is the biggest is 0.

Following this, we decided to see our customers' addresses per cluster. On our first try, did not have a clear view to see where the customers from each cluster lived. So, we decided to create an interactive map where we not only can navigate but also can select the clusters we want to see.

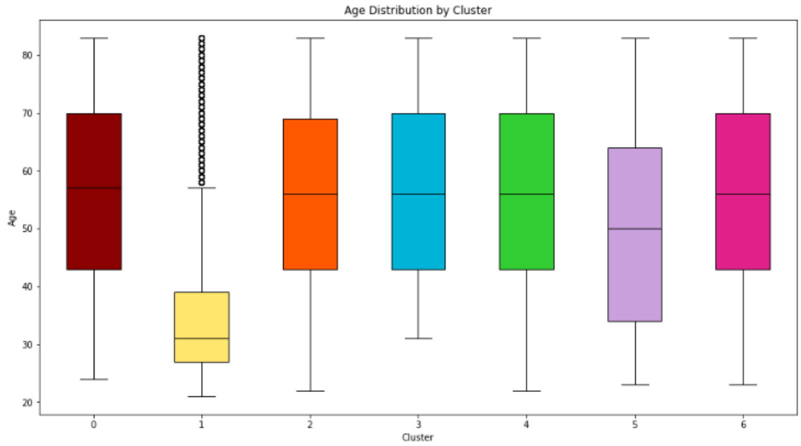
Figure 14 – Metropolitan Area of Lisbon



Using our map, we can see that cluster 5 is well-defined and isolated near Mercado Abastecedor da Região de Lisboa (MARL). The remaining clusters have their customers in mostly well-defined groups, but they are overlapping with each other in some parts, mainly in Lisbon municipality.

Our next step was to see the distribution of our customers' age by cluster. To do that we created a boxplot for each, as we can see below.

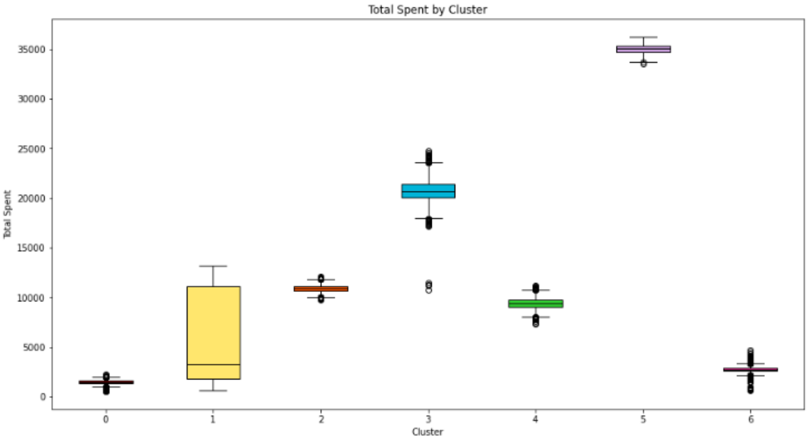
Figure 15 – Age distribution by cluster



Here we can see that most of our clusters have about the same age distribution. However, one is different from the pattern. In cluster 1, most of the customers are younger in comparison to the others but it still has some that are a bit older.

Next, we decided to explore some of the variables that we had some conclusions about in the previous notebooks. The first we explored was the variable 'total_lifetime_spent' through box plots.

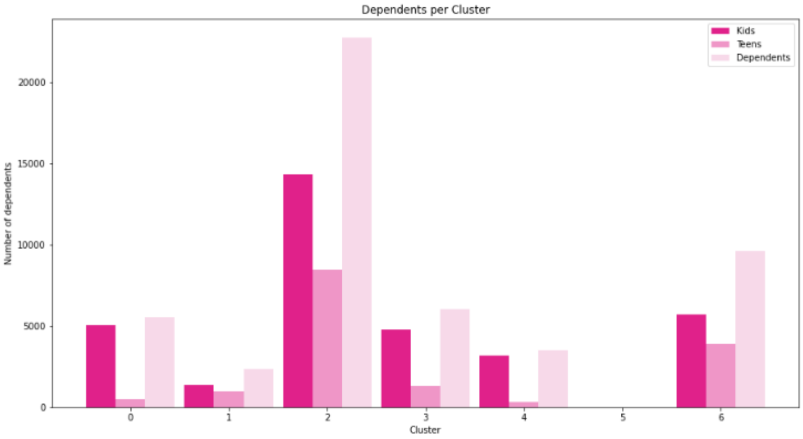
Figure 16 – Distribution of total lifetime spent per cluster



In this variable, we can see several differences between the clusters. The first one is the high amounts of money spent by the customers in cluster 5. Not far away in terms of spending more is cluster 3 but has some that escape that rule. As for the ones that spent less money, they are in clusters 0 and 6. We also noticed that cluster 1 is the least defined in terms of spending, as it has the biggest interquartile range.

After, we saw the sum of the quantities of three different variables per cluster. Those variables were: 'kids_home', teens_home' and 'number_dependents'. To facilitate our analyse, we made a visualization which was a multiple bar chart.

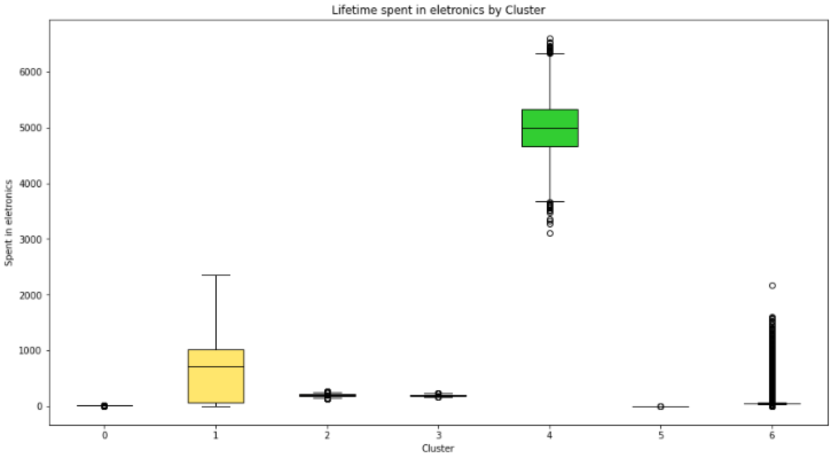
Figure 17 – Kids, teens and dependents per cluster



Above, we can see that our customers that were in cluster 5 neither had kids or teens living at home. The other two clusters that had a small number of dependents were 1 and 4. As for cluster 2, it stands out by having the most kids and teens living at home. We also found it interesting that most clusters had customers with more kids than teens.

Next, we saw another variable that we had taken note of in the previous analysis which has the total spent on electronics. Below, we used box plots to see the distribution of this variable in different clusters.

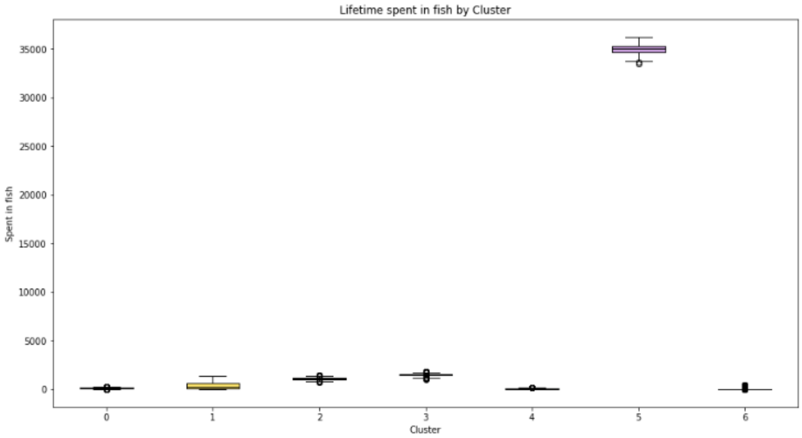
Figure 18 – Lifetime spent on electronics per cluster



In the figure above, we can see that there is a cluster that is characterized by customers that spend a considerable amount on electronics (cluster 4). It is also noteworthy to observe that clusters 0, 2, 3 and 5 are barely distinguishable with this variable.

Following this, we decided to see our clusters using the variable related to the amount spent on fish. In this variable, we had already noticed a distinct group that spent a significant amount of money. As such, we wanted to see if big spenders on fish were considered a cluster. Below we can see several box plots with the distribution of the variable in question by cluster.

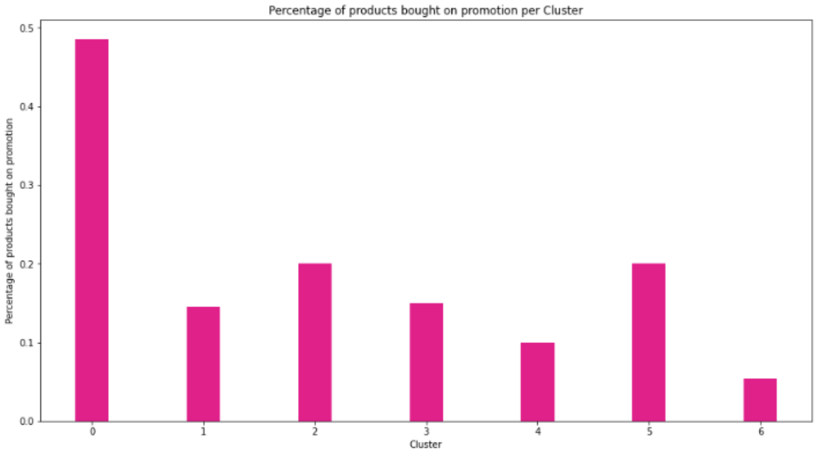
Figure 19 – Lifetime spent on fish by cluster



Above, we can see that our prevision of a cluster of big spenders on fish became true as we can see in cluster 5. As for the remaining clusters, there is nothing noteworthy as they are too alike when comparing them with 'lifetime_spend_fish'.

Our last visualization per cluster was related to the variable 'percentage_of_products_bought_promotion'. In this variable, we already noticed the existence of a possible cluster where the customers bought about half of their products on promotion. We also wanted to confirm its existence so below we can see a bar chart that has the mean of each cluster for the variable in question.

Figure 20 – Mean of the percentage of products bought on promotion per cluster



Here, we can confirm our last hypothesis of the existence of a cluster of clients that bought many products on promotion in cluster 0. We also can see that in cluster 6 the customers buy very few products on promotion.

In our notebook, we saw a few more visualizations that can be seen in annexes. After seeing all the above visualizations and reaching some conclusions. Below, is a summary of the characteristics of each cluster:

- Cluster 0 – customers that buy most products on promotion, don't spend much money, and have an overwhelming number of customers that only the high school diplomas. (Thrifty)
- Cluster 1 – youngest customers on the dataset and almost don't have dependents. (Young and Unattached Customers)

- Cluster 2 – customers that have the biggest number of dependents, are the second biggest spenders on meat and are the biggest spenders on hygiene products. (Parents)
- Cluster 3 – first clients (based on the year of the first transaction), biggest spenders on meat, alcoholic beverages, and groceries. It is also the second biggest spender in general and on vegetables. (Meat and Alcohol Connoisseurs)
- Cluster 4 – these customers are big spenders on electronics, non-alcoholic drinks and video games. (TechSip Gamers)
- Cluster 5 – the smallest cluster that has the biggest spenders, the customers don't buy almost any vegetables, they have no dependents, all of them only have high school diplomas and they spend a lot of money on fish. (Fish Enthusiasts)
- Cluster 6 – barely spends any money with and without promotions and they mostly spend money on vegetables. (Minimalist Green Shoppers)

Targeted Promotion

After creating our clusters and describing them, we started doing target promotions for each one of our clusters. Target promotions offer customers exclusive deals, discounts, and special offers. These promotions are designed to attract shoppers, encourage purchases, and provide value to customers.

To make them, we started by applying association rules to our cluster. Association rules are the extraction of compact patterns to describe subsets of data. It provides information about things that tend to happen together. To evaluate their quality, we used three important measures:

- Support – shows how frequently a combination occurs.
- Confidence – reflects the strength of the association by giving the percentage of the appearance of the consequent when the antecedent has occurred.
- Lift – shows the likelihood of the consequent increasing given an antecedent.

Below, we can see a brief summary of each cluster, some promotions and our reasoning behind those promotions.

Cluster 0- 'Thrifty'

The defining characteristic of this cluster is that they love a good promotion. They normally only buy things from us if they have discount. Because they buy things from every category, our main focus is to make sure we become their number one supermarket. To do so we suggest these promotions:

1. 10% off on a purchase superior to 100€ - These clients spend very little money, to try and force them to spend more, this promotion is only valid on purchases superior to 100€. However, it is valid for every product to make. The idea of not limiting this promotion to one or only a few categories is that this way they might go through the whole store and buy more things.
2. 20% off on cooking oil on the purchase of butter – cooking oil appears as a consequent on a large number of our association rules for this cluster. We decided that this would be a good promotion, since the rule that has butter as an antecedent and cooking oil has a consequent has one of the higher lifts and one of the highest confidence. This means that people in this cluster buy butter when they buy cooking oil. Because they usually only buy products on promotion, we thought it would be a good idea to try and convince

them to buy a product for its full price (butter) while offering a discount on a product that appears in a lot of our rules.

3. 25% 'Party Mix' Bundle – we found out that cake was present in a large amount of our association rules, both as an antecedent and a consequent. It also appears a lot paired with other sweets like candy bars, gum, cookies, and muffins. For this reason, we think that we should offer a bundle consisting of one unit of all the products above and price them at 25% off. So, if they want to buy them with a discount, they have to buy them all.
4. Buy 3 get one free on oil – oil is one of the products that appear in a great number of association rules, not only that but they also appear on the ones with the best confidence. We think it is a very popular product so we should incentivize these thrifty customers to get it with us. This promotion offers a free oil bottle on the purchase of 3 others, making them buy a higher number of products in order to feel like they got a good deal.
5. 20% on groceries – Even if they do buy from every category, the majority of our association rules have products corresponding to the groceries category on them. If we give them a discount on these products, it is more probable that they will spend money on other complementing products from other categories that we have seen present on association rules with them.

Cluster 1- 'Young and Unattached Customers'

This cluster, 'Young and Unattached Customers', is the youngest cluster and barely has any dependents. From observing the association rules we were able to conclude that they buy a lot of alcoholic drinks together. We want them to see us as their prime provider of alcoholic drinks. To achieve that objective, we suggest these promotions.

1. 15% off on beer – beer is one of the items that are usually bought with other drinks. Therefore, by increasing the amount of beer they can buy with the same money we hope to encourage the purchase of other beverages.
2. Buy one champagne bottle get 20% discount on the second one – we observed that champagne appears as an antecedent on a lot of transactions. We want our clients to buy more champagne so, we will offer a discount on the second bottle.
3. Buy 2 get 1 free on oil – by analysing the association rules we noticed that oil appears as a consequent on a lot of them. By encouraging the purchase of oil we hope customers will also want to buy more products.
4. Get 30% off cider if you buy from the groceries category – groceries appear in a lot of antecedents on this cluster. Since these customers tend to buy a lot of cider, by giving 30% on cider if they buy from groceries, we hope to encourage them to start buying groceries from our store.
5. Buy dessert wine get 40% off white wine – dessert wine and white wine appear in a lot of the rules we discovered both alone and as a set with a high lift and high confidence. We thought that giving a discount on one of them would encourage customers in buying both more times.

Cluster 2 – 'Parents'

The 'Parents' cluster behaves exactly how you would expect it to behave. They have the most dependents and are the biggest spenders on hygiene products and the second biggest in the meat category. We want these customers to lean on us and to view us as a store that has everything they could ever need. For that, we propose the promotions listed below.

1. 50% off on baby food if you buy beats headphones – baby food was a consequent of a lot of association rules with a high confidence. Since these customers will buy baby food, we want to promote the purchase of products of a different category they don't usually buy from. We choose beats headphones because while a parent with small kids might not use them if the client has teenagers at home, they might enjoy it.
2. 10% of on cake – cake appears in a lot of the association rules for this cluster and if we think about it makes sense since the customers in this cluster are the ones that have more dependents at home. There is nothing that children love more than sweets and if we can convince these parents to buy them a cake as a dessert then we can maximize the number of aisles they go through during their shopping trip and hopefully put into action some of the association rules by buying other products.
3. Get 10% off on spaghetti if you buy from the meat category – feeding children is a difficult task, sometimes parents don't have time to do elaborate meals, and sometimes the kids are picky eaters. It was thinking about this that we made this promotion. These clients are already the second cluster that spends more on meat and we wanted to incentivize them to spend more on other categories by offering a simple solution for dinner. So, these parents can come and buy some kind of meat and also buy spaghetti. Not only is this a simple and easy way to feed children, but hopefully they will also buy more products while picking these up.
4. 25% 'Sauce bundle' – we noticed these customers tend to buy ketchup, so we will offer ketchup, mayonnaise and barbecue sauce as a bundle. This way, if they want a discount on ketchup, they must buy the other two sauces as well.
5. 25% of candy bars and gum – candy bars and gums appear as antecedents on transactions with high confidence. For this reason, we decided to give 25% off on these products. This promotion is not cumulative with other promotions.

Cluster 3 – 'Meat and Alcohol Connoisseurs'

These clients have been with us the longest, spend the most on meat, alcoholic beverages, and groceries. They are the second biggest spenders in general and on vegetables. Since we have had these customers for so long, we don't want to lose them and need to reassure them we are here for all their needs. To achieve this, we propose the promotions below:

1. 30% off on all vegetables – These clients are our biggest spenders in general, but a category where they aren't is vegetables. So, to try and incentivize them to spend even more money with us, we decided to offer them a discount on the vegetable category. Our idea is that if we are able to make these clients start to buy their vegetables from us then it is one less thing they will feel the need to look for elsewhere and hopefully, it will incentivize them to buy everything from us.
2. Get a bottle of oil for each 40€ spent on the hygiene category – oil, as well as cooking oil, appear often as a consequent on the transactions with higher confidence. For this reason, we will encourage this group of customers to buy more hygiene products.
3. Buy 2 get one free on French fries – French fries are seen a lot as an antecedent on the transactions with the highest lift. If we encourage more purchases of French fries, we hope to encourage the purchase of different products.
4. Buy napkins get 20% off on meat – Using our association rules for this cluster, we saw that napkins appeared on a lot of transactions as the antecedent. So, we thought that we could use it to incentivize even more the spending of this cluster on the meat category. Hopefully, this way these customers buy even more meat, and since the

napkins are so heavily featured on our rules, we also hope it will influence them to buy other products.

5. Spend 50€ on meat and get 10% off on any type of wine – Because these customers are big fans of meat and also spend more than average on alcoholic beverages, we decided to create a promotion that combined two of their favourite things. Ideally, this will make them buy even more meat and guide them toward more expensive alcoholic beverages. As they already spend quite a lot on these categories it will also serve to make them feel valued as customers and help to retain them as clients, since they are our biggest spenders.

Cluster 4 – 'TechSip Gamers'

The 'TechSip Gamers' are a particular bunch, most of the money they spend with us is on videogames and electronics. They also have the tendency to buy drinks, being the biggest spenders on non-alcoholic drinks.

Since videogames and electronics are expensive, we wanted to make sure these customers saw us as their first choice to buy these. For that we suggest the following promotions:

1. 50% off on 'Pokemon Shield' if bought with 'Pokemon Violet' and 'Pokemon Scarlet' – looking at our association rules, we saw that the ones with the biggest lift and confidence where the ones that included 'Pokemon Violet' and ' Pokemon Scarlet' on the antecedent, so we know that the customers that buy these two games have the tendency to then buy 'Pokemon Shield', but to secure that they do we think it would be a good idea to give them a discount.
2. Get a pack of soda with 'Pokemon Sword' – to ensure our gamers continue to buy drinks from us, we thought it would be a good idea to offer them a pack of soda with the highly popular 'Pokemon Sword'. This way they have to go to the beverage aisle and might be influenced to buy other drinks.
3. 25% off on the 'ratchet & clank' trilogy bundle – for limited time, we should incentivize these customers to buy a highly popular and highly rated trilogy. By selling a bundle at 25% off the original price we are increasing the number of games sold, since the majority of people don't usually buy 3 games at once, and trying to make these customers associate the company with good deals on video games.
4. Buy one, get 50% off on the second pair of Bluetooth headphones – A week's promotion on the second pair of Bluetooth headphones is a good way to incentivize these clients to come to our store. The limited time and the big discount might be a good way to make them not want to lose the deal. Furthermore, we can make this promotion closer to a holiday so these customers see it as an easy way to get someone a gift.
5. Buy an iPad, get a 30% discount on AirPods – this promotion is good for those who seek seamless integration and compatibility. With this promotion, we wanted to make our customers buy different expensive products from the same brand and same category. This could incentivize them to buy the rest of the products of the same brand with our stores.

Cluster 5 – 'Fish Enthusiasts'

Our cluster 5, also known as Fish Enthusiasts, is the smallest of them all. However, they are still the biggest spenders, mostly on fish. Our first step was to see if they spend some money in any other category, which they barely did. However, most of our customers in this group bought

20% of their products on promotion. This indicates that if we increase the number of products on promotion, they may increase the money spent in the stores.

To do promotions for this specific cluster, our line of thought was that they already spend a lot on fish, so we could do promotions on other categories so that these big spenders can enjoy other products of our stores. An important detail is that those discounts should not be very high as they are already our biggest spenders. Here are some promotions for Fish Enthusiasts and our reasoning behind them:

1. Get a 15% discount on Fish if you buy groceries – for this promotion, we saw that most association rules with high lift, high confidence and low support had products from the groceries category. So, we incentive the purchase of groceries by giving them discounts on fish.
2. Buy 3 avocados and get a 10% discount on trout. – there is an association rule where the antecedent is avocado and the consequents are oil and cooking oil. As it has lift above 1.5, with confidence of 1 and low support, we decided to increase the times the bundle is likely to appear by giving a discount on a specific fish with many cooking oil recipes.
3. Get a 10% discount on fish, if you buy pet food. – there are many rules with pet food as an antecedent (alone and with others). So, we decided to increase the probability of selling them and at least one more product by, giving a small discount on every type of fish.
4. Get a 35% discount on seabass if you buy vegetables. – we thought that this promotion could be a good incentive to buy vegetables that go along with cheaper fish. As these customers already spend a lot, giving them a higher discount on cheaper fish won’t affect our sales much.
5. Buy a package of mashed potatoes, get canned tuna. – this promotion also pretends to incentivize the purchase of a product in the vegetable category. We decided to make mashed potatoes and canned tuna as a bundle because they are the main products to make a quick meal known as tuna pie.

Cluster 6 – ‘Minimalist Green Shoppers’

The last cluster, known as Minimalist Green Shoppers, barely spends any money; when they do, it’s mostly on vegetables. However, in this cluster, many customers are far from the mean spent in each category. So, we decided to do promotions that relate the vegetable category with the others that had some outliers. Below, are our promotions for this customer segment:

1. Buy 1 white wine, get 1 free cider. – this promotion is based on two associations that have a lift of approximately 9. We decided to offer the cider because it’s usually cheaper.
2. Buy carrots and hot dogs, get a 50% discount on tomatoes. – an association rule with hot dogs and carrots as antecedents and tomatoes as consequent has a lift of 1.11, a confidence of almost 0.9 and a support of 0.048. We decided to increase the frequency of this bundle by offering a discount on the tomatoes. This rule was also chosen because the majority of the products are from the category vegetables which is the one where these customers spend the most money.
3. Buy a package of frozen vegetables, get 1 kg of tomatoes for free. – as these customers barely spend any money, we thought that an increase in the amount spent on vegetables would increase the probability of taking up other promotions.

4. Get a 20% discount on vegetables if you buy electronics. -this promotion was related with the fact that these customers barely spend money on our stores. We thought that giving discounts in the category in which they spend the most money by making them buy from another in which some spend a little more, would increase the interaction between the store and these customers.
5. Get a 50% discount on a basket in the entire store just for you! – this promotion also has as its main objective increasing our sales with these customers. They buy less than 10% of their products on promotion (maybe because they only buy vegetables), so this could increase the probability of buying from another category as they are not going to pay the full price.

Conclusion

Being able to tailor and accommodate the needs of their customers is vital to the survival of a business in today's market. To do this, machine learning has become a very powerful tool for retail businesses. With the help of data scientists, big retailers are able to handle their data to gain insights about their clients and give them a better shopping experience.

In this project, the objective was to use the provided datasets about customer demographics, spending habits, purchasing behaviour, and historical transitions to create customer segmentations that made it easier to manage the different types of clients.

To do so, we started by doing an exploratory data analysis to better understand what we were working with. After, we tested different clustering algorithms to try and find one that gave us a solution that we could easily understand and work with. Next, we further analysed our clusters in order to better understand the customers present in our segments. Finally, we used those insights and association rules to aid us in the more creative part of this project, creating promotions for each of our clusters.

As for our clusters, using different unsupervised algorithms, we got similar customer segments. As such, we can see that they were well-defined. We were also able to create a customer profile for each cluster. In cluster 0 we had people who mostly bought products on promotion. The cluster 1 had the youngest customers with no dependents. In the complete opposite spectrum, we had cluster 2 where we recorded the biggest number of dependents and, as such, they were the biggest spenders on hygiene products and the second ones on meat. Next, we had cluster 3 which had our first clients which were the biggest spenders in several categories. Following, we had cluster 4 with customers that were the top spenders on videogames, electronics, and non-alcoholic drinks. Cluster 5 was composed by very few customers, but they were our biggest clients, especially on fish. Lastly, in cluster 6 we had customers that only spent money on vegetables.

Regarding the overall success of the project, we think that the objective was successful since we were able to further explore and consolidate what we learned during the semester. Furthermore, even if the data was a lot better than what we will find while working, we believe that this project was important to see one of the possible uses of machine learning in the real world.

References

Li, Y., & Chung, S. M. (2007). Parallel bisecting k-means with prediction clustering algorithm. The Journal of Supercomputing, 39, 19-37. Available on: https://doi.org/10.1007/s11227-006-0002-7

Bação, Fernando. Introduction to Clustering PowerPoint. 2023. Pdf. Available on: Nova IMS moodle: 202223 - Aprendizagem Máquina II - S2

Bação, Fernando. Hierarchical Clustering Algorithm. 2023. Pdf. Available on: Nova IMS moodle: 202223 - Aprendizagem Máquina II - S2

Bação, Fernando. Partition Algorithms. 2023. Pdf. Available on: Nova IMS moodle: 202223 - Aprendizagem Máquina II - S2

Bação, Fernando. Density-based clustering. 2023. Pdf. Available on: Nova IMS moodle: 202223 - Aprendizagem Máquina II - S2

Bação, Fernando. Mean shift clustering. 2023. Pdf. Available on: Nova IMS moodle: 202223 - Aprendizagem Máquina II - S2

Bação, Fernando. Association Rules. 2023. Pdf. Available on: Nova IMS moodle: 202223 - Aprendizagem Máquina II - S2

Bação, Fernando. Multidimensionality Visualization. 2023. Pdf. Available on: Nova IMS moodle: 202223 - Aprendizagem Máquina II - S2

Annexes

Annexe 1 – Plots from notebook "1_EDA"

Figure 21 – Boxplots of lifetime spent categories

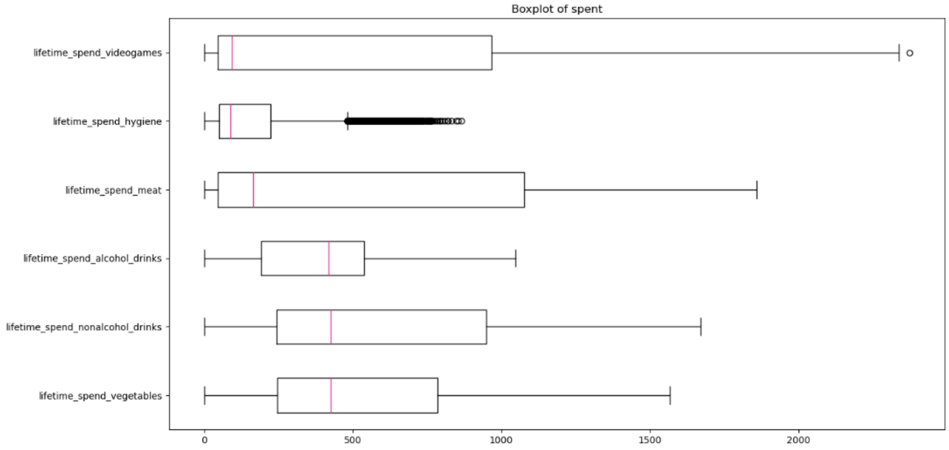


Figure 22 – Boxplot of total lifetime spent

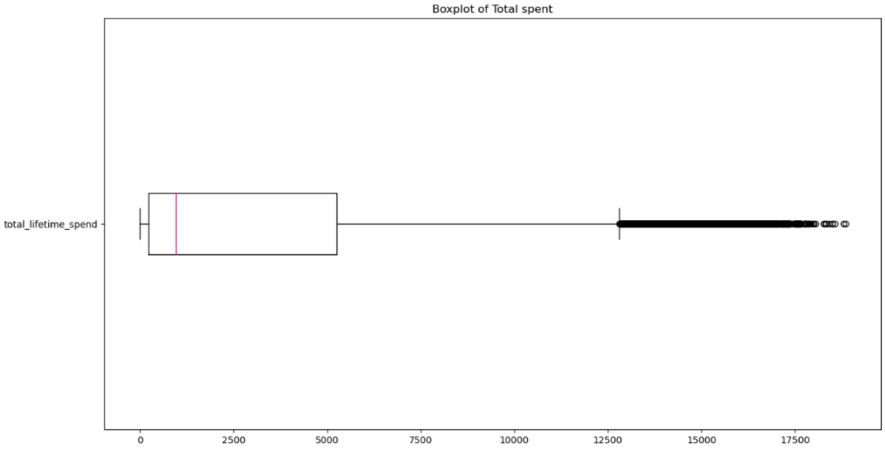


Figure 23 – Boxplot of total distinct products customers have ever bought

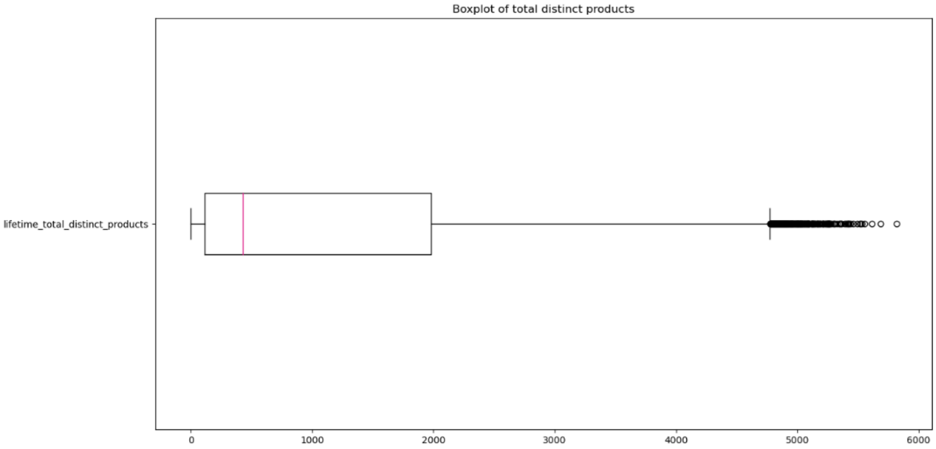


Figure 24 – Boxplot of the percentages clients buy on promotion

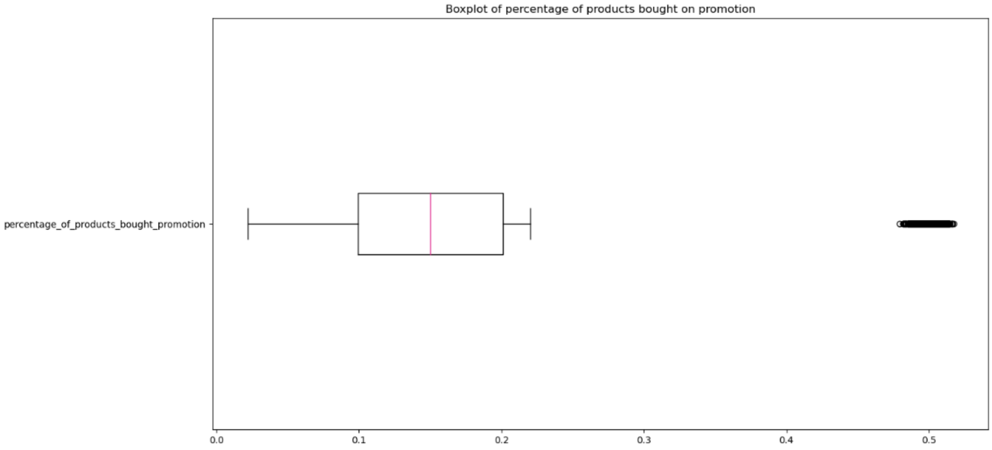


Figure 25 – Bar chart with the number of customers that share the same amount of dependents

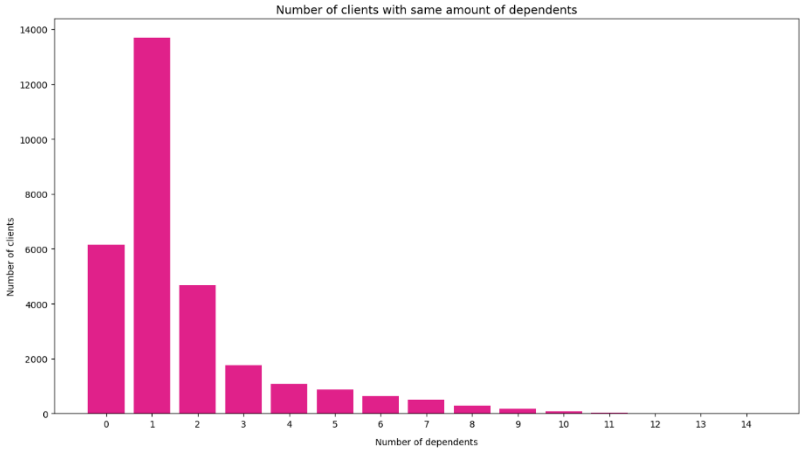


Figure 26 – Bar chart with the year clients started buying from us

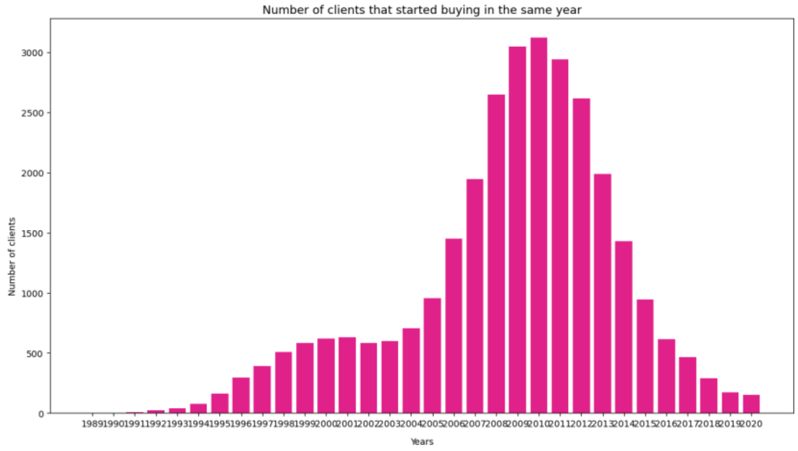


Figure 27 – Bar chart with the typical time of day clients shop

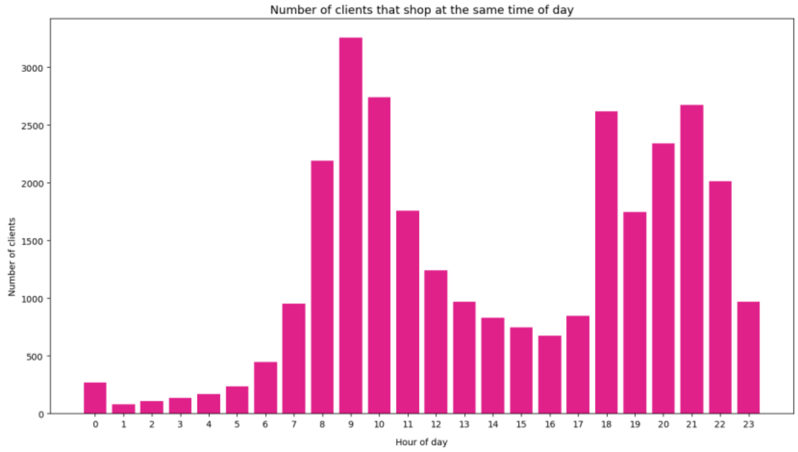


Figure 28 – Bar chart with the number of shops customers visited

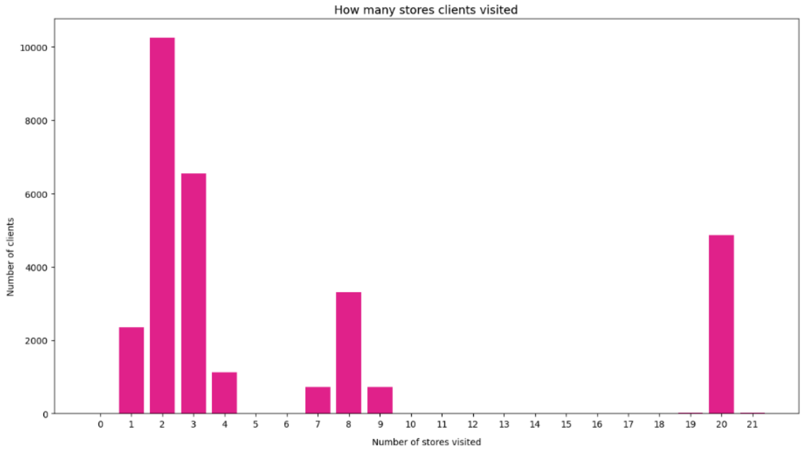


Figure 29 – Histogram of the distribution of products bought

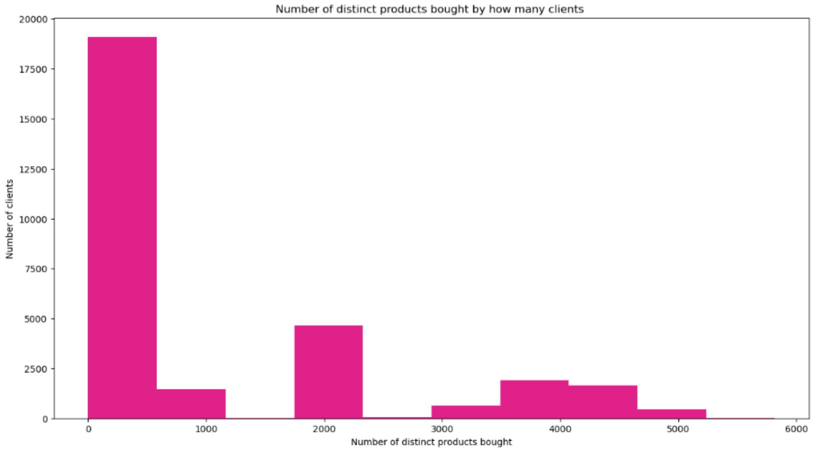
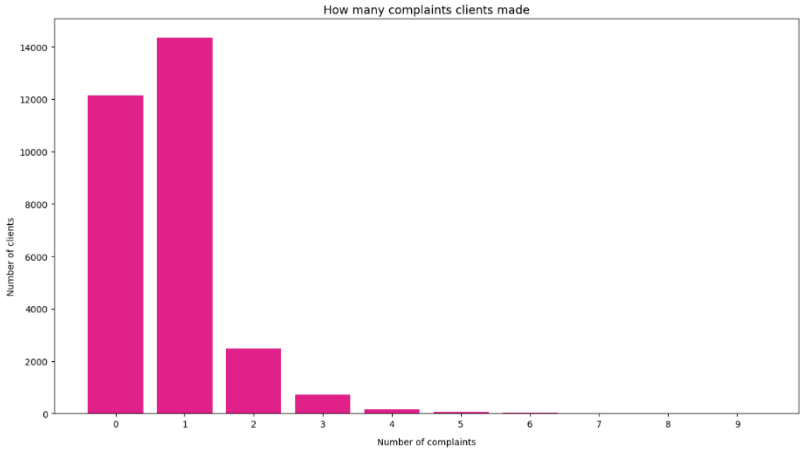


Figure 30 – Bar chart of the number of complaints clients made



Annexe 2 – Plots and tables from notebook ‘’2_Model”

Figure 31 – bar chart of the number of customers in each clusters DBSCAN_1

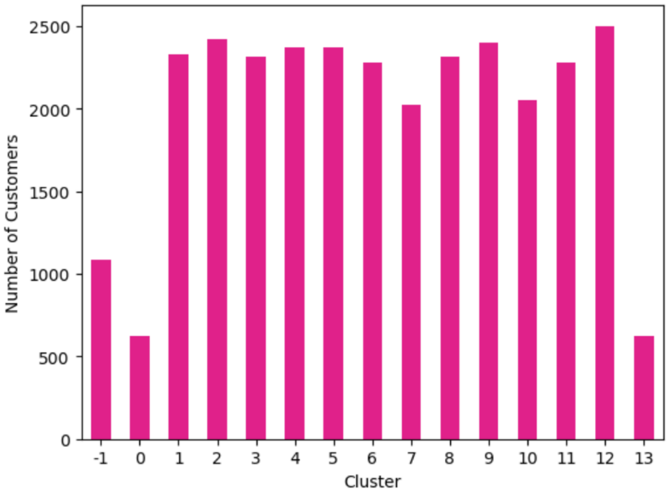


Figure 32 – bar chart of the number of customers in each clusters DBSCAN_2

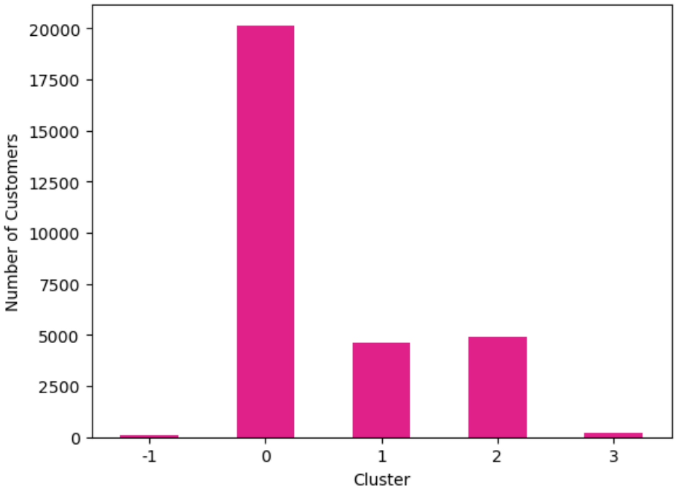


Figure 33 – bar chart of the number of customers in each clusters DBSCAN_nb_1

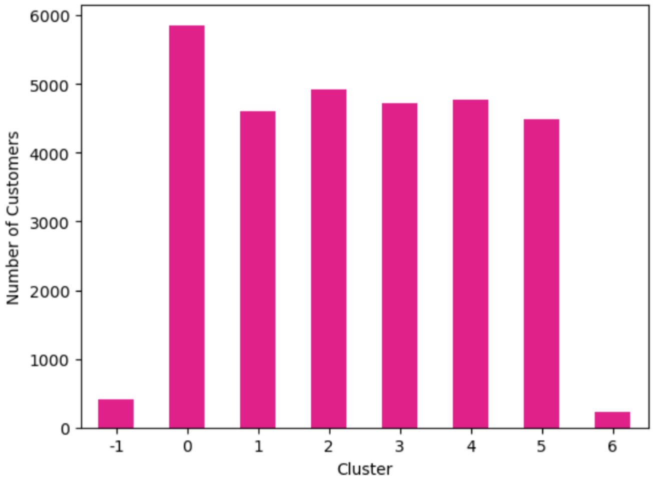


Figure 34 – bar chart of the number of customers in each clusters DBSCAN_nb_2

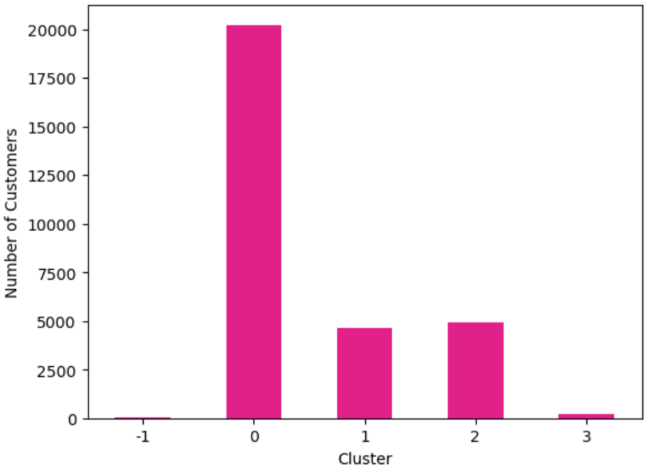


Figure 35 – Dispersion plot K-Means

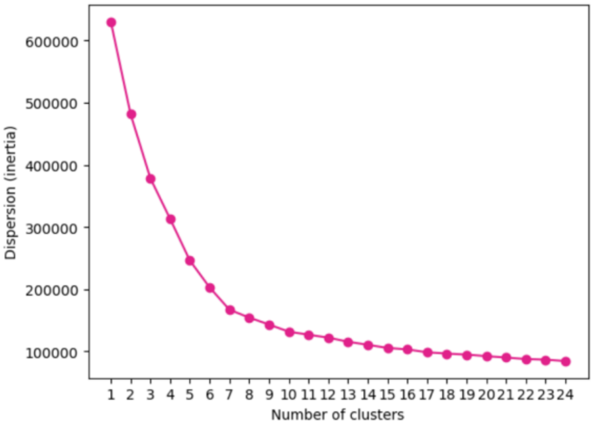


Figure 36 – Dispersion plot K-Means_nb

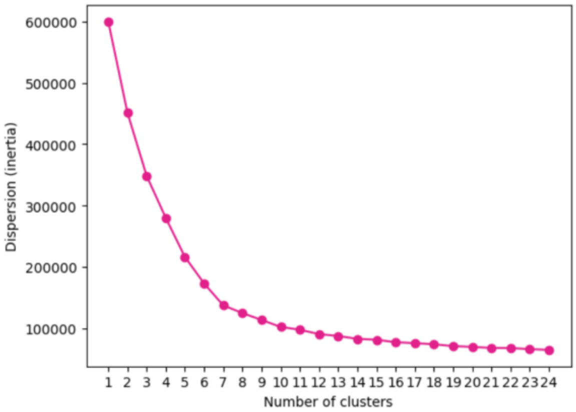


Figure 37 – bar chart of the number of customers in each clusters K-Means

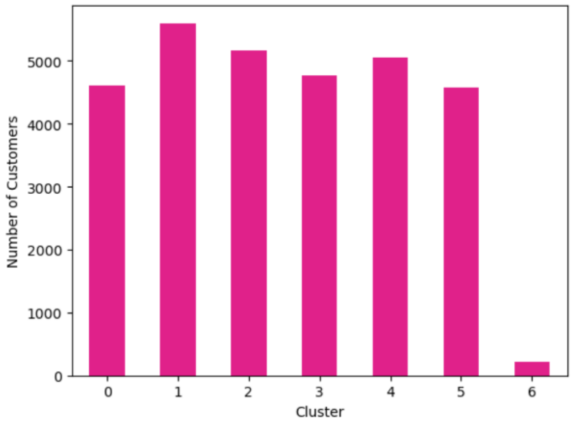


Figure 38 – bar chart of the number of customers in each clusters K-Means_nb

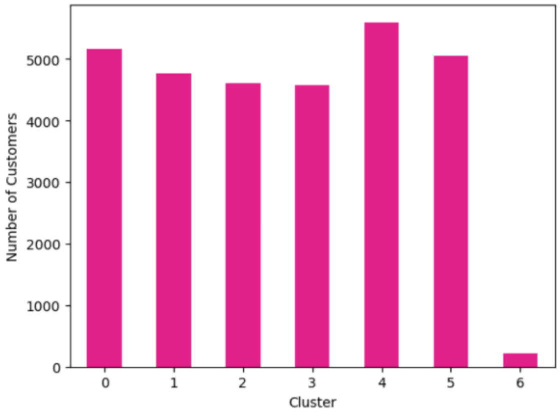


Figure 39 – Dispersion plot Bisecting K-Means

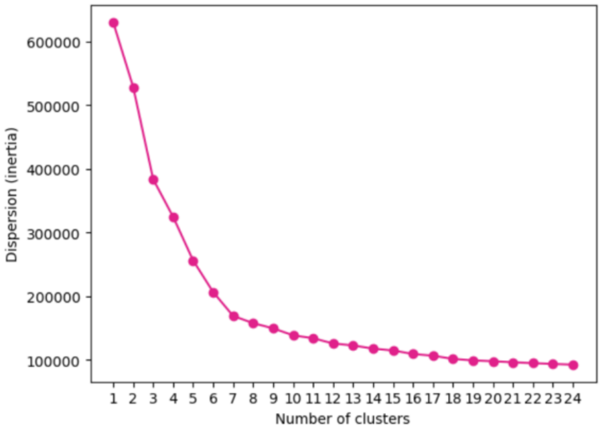


Figure 40 – bar chart of the number of customers in each clusters Bisecting K-Means

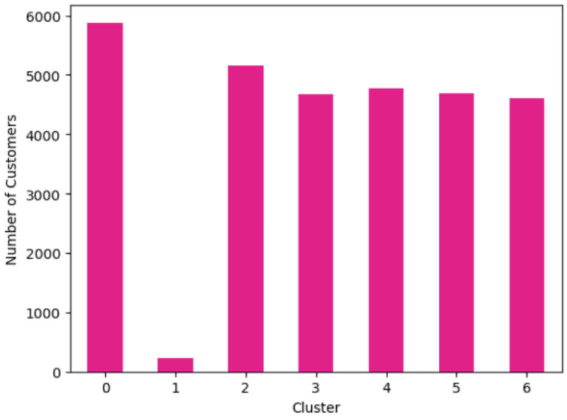


Figure 41 – Minimum Hierarchical Clustering Dendrogram

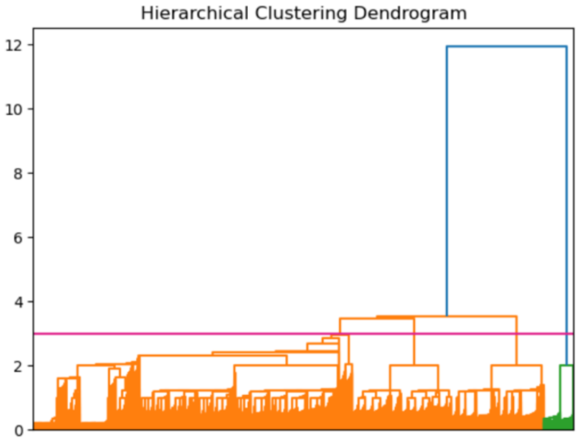


Figure 42 – bar chart of the number of customers in each clusters Minimum Hierarchical Clustering

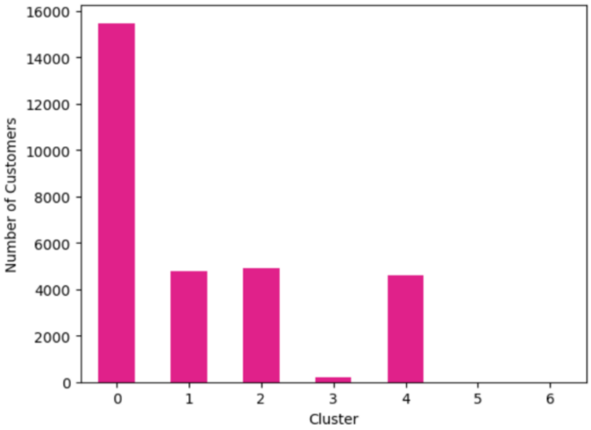


Figure 43 – Ward's Hierarchical Clustering Dendrogram

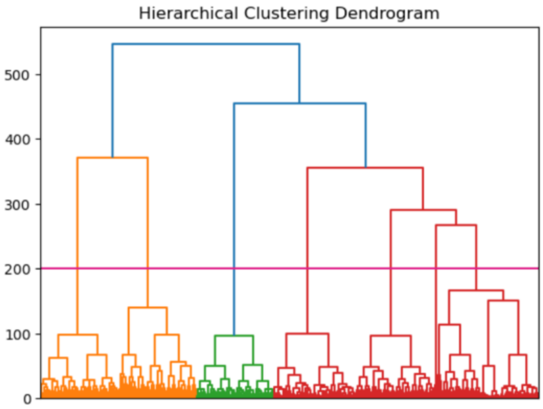


Figure 44 – bar chart of the number of customers in each clusters Ward's Hierarchical Clustering

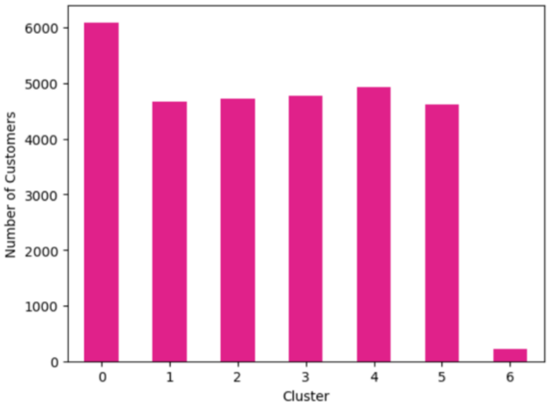
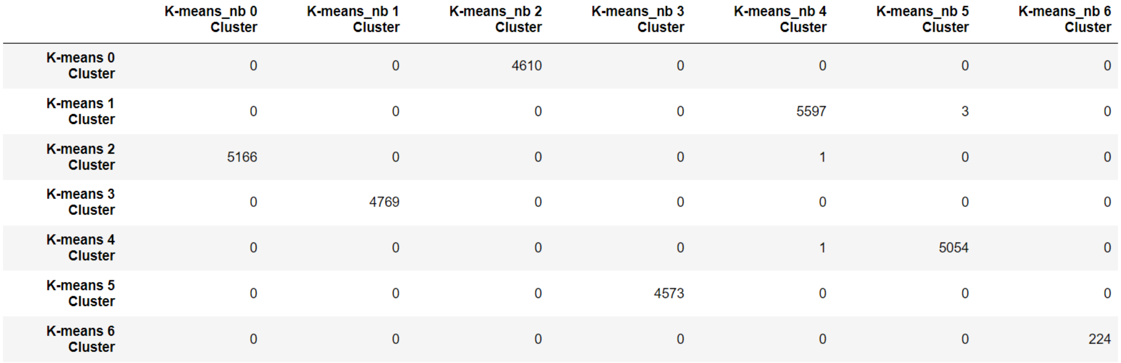
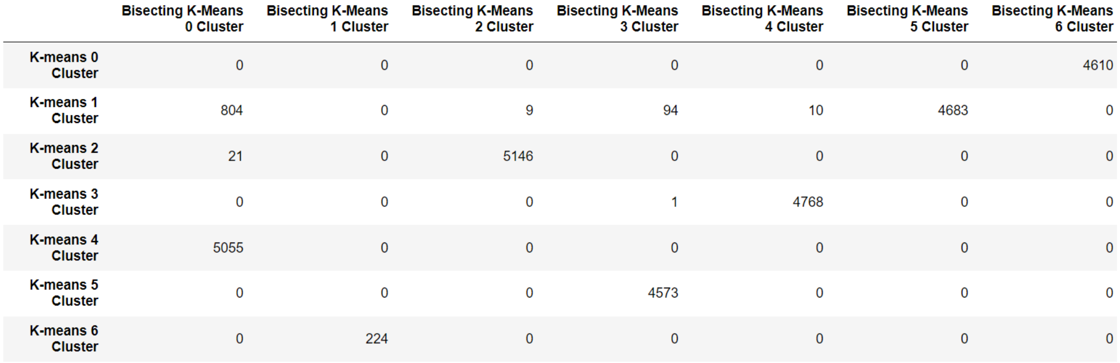


Table 1 – table comparison of the K-means



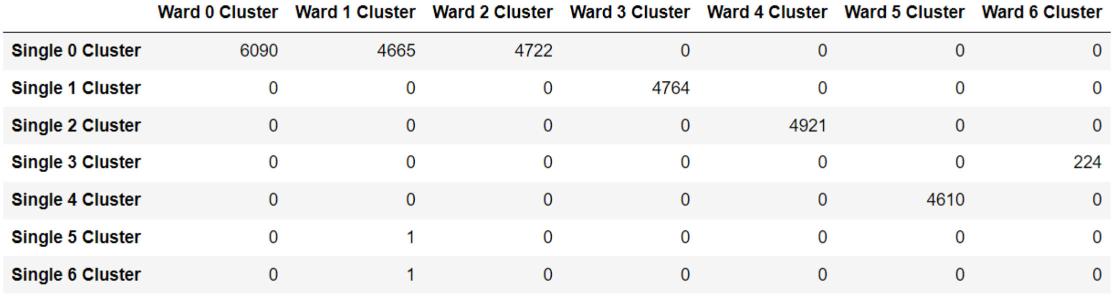
	K-means_nb 0 Cluster	K-means_nb 1 Cluster	K-means_nb 2 Cluster	K-means_nb 3 Cluster	K-means_nb 4 Cluster	K-means_nb 5 Cluster	K-means_nb 6 Cluster
K-means 0 Cluster	0	0	4610	0	0	0	0
K-means 1 Cluster	0	0	0	0	5597	3	0
K-means 2 Cluster	5166	0	0	0	1	0	0
K-means 3 Cluster	0	4769	0	0	0	0	0
K-means 4 Cluster	0	0	0	0	1	5054	0
K-means 5 Cluster	0	0	0	4573	0	0	0
K-means 6 Cluster	0	0	0	0	0	0	224

Table 2 – table comparison of the K-mean with the Bisecting k-means



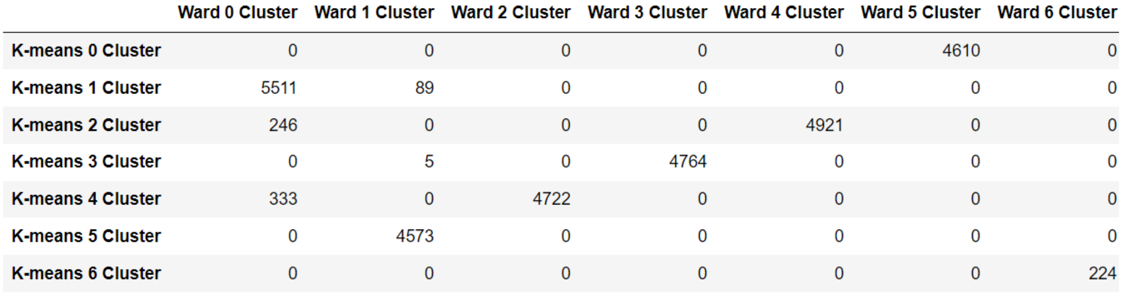
	Bisecting K-Means 0 Cluster	Bisecting K-Means 1 Cluster	Bisecting K-Means 2 Cluster	Bisecting K-Means 3 Cluster	Bisecting K-Means 4 Cluster	Bisecting K-Means 5 Cluster	Bisecting K-Means 6 Cluster
K-means 0 Cluster	0	0	0	0	0	0	4610
K-means 1 Cluster	804	0	9	94	10	4683	0
K-means 2 Cluster	21	0	5146	0	0	0	0
K-means 3 Cluster	0	0	0	1	4768	0	0
K-means 4 Cluster	5055	0	0	0	0	0	0
K-means 5 Cluster	0	0	0	4573	0	0	0
K-means 6 Cluster	0	224	0	0	0	0	0

Table 3 – table comparison of the ward hierarchical with the minimum hierarchical



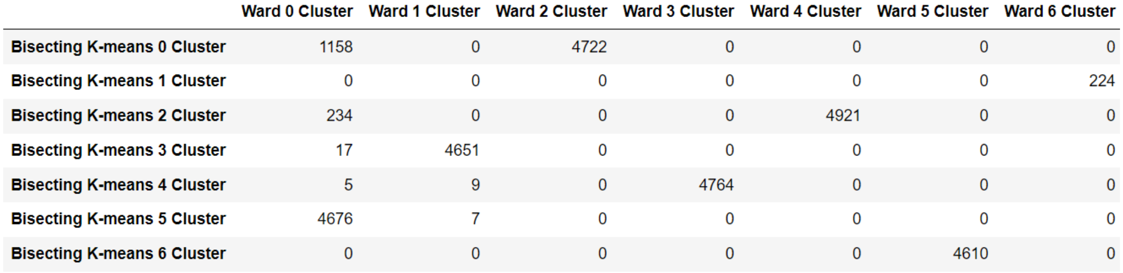
	Ward 0 Cluster	Ward 1 Cluster	Ward 2 Cluster	Ward 3 Cluster	Ward 4 Cluster	Ward 5 Cluster	Ward 6 Cluster
Single 0 Cluster	6090	4665	4722	0	0	0	0
Single 1 Cluster	0	0	0	4764	0	0	0
Single 2 Cluster	0	0	0	0	4921	0	0
Single 3 Cluster	0	0	0	0	0	0	224
Single 4 Cluster	0	0	0	0	0	4610	0
Single 5 Cluster	0	1	0	0	0	0	0
Single 6 Cluster	0	1	0	0	0	0	0

Table 4 – table comparison of the K-mean with the ward hierarchical



	Ward 0 Cluster	Ward 1 Cluster	Ward 2 Cluster	Ward 3 Cluster	Ward 4 Cluster	Ward 5 Cluster	Ward 6 Cluster
K-means 0 Cluster	0	0	0	0	0	4610	0
K-means 1 Cluster	5511	89	0	0	0	0	0
K-means 2 Cluster	246	0	0	0	4921	0	0
K-means 3 Cluster	0	5	0	4764	0	0	0
K-means 4 Cluster	333	0	4722	0	0	0	0
K-means 5 Cluster	0	4573	0	0	0	0	0
K-means 6 Cluster	0	0	0	0	0	0	224

Table 5 – table comparison of the ward hierarchical with the Bisecting k-means



	Ward 0 Cluster	Ward 1 Cluster	Ward 2 Cluster	Ward 3 Cluster	Ward 4 Cluster	Ward 5 Cluster	Ward 6 Cluster
Bisecting K-means 0 Cluster	1158	0	4722	0	0	0	0
Bisecting K-means 1 Cluster	0	0	0	0	0	0	224
Bisecting K-means 2 Cluster	234	0	0	0	4921	0	0
Bisecting K-means 3 Cluster	17	4651	0	0	0	0	0
Bisecting K-means 4 Cluster	5	9	0	4764	0	0	0
Bisecting K-means 5 Cluster	4676	7	0	0	0	0	0
Bisecting K-means 6 Cluster	0	0	0	0	0	4610	0

Annexe 3 – UMAPs from notebook "3_Clusters"

UMAP of the other two algorithms: bisecting k-means and hierarchical clustering with ward linkage.

Figure 45 – UMAP of bisecting k-means

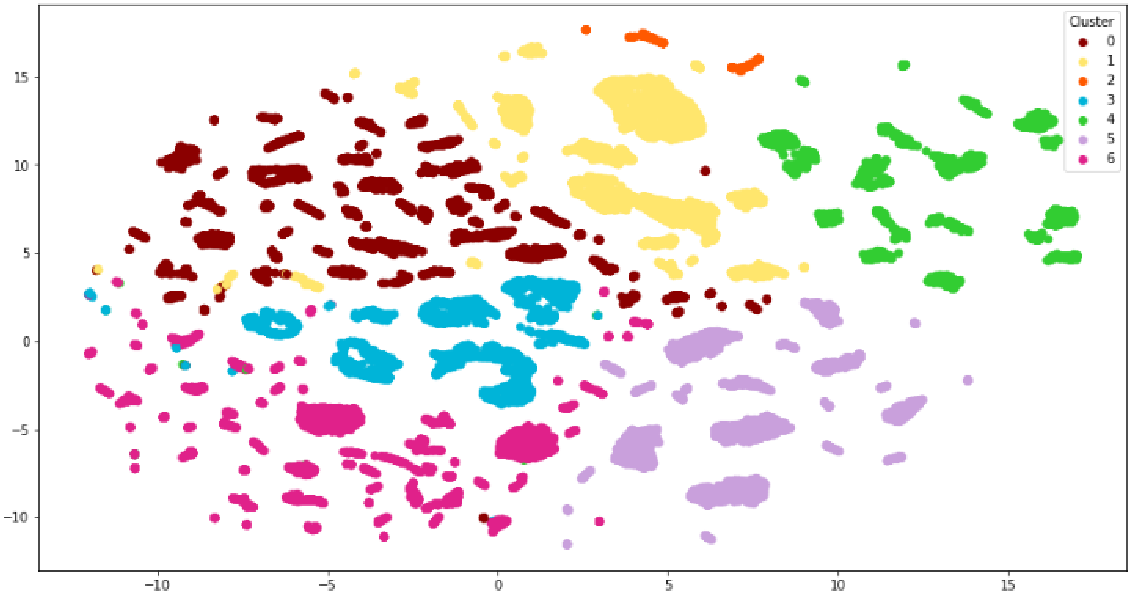
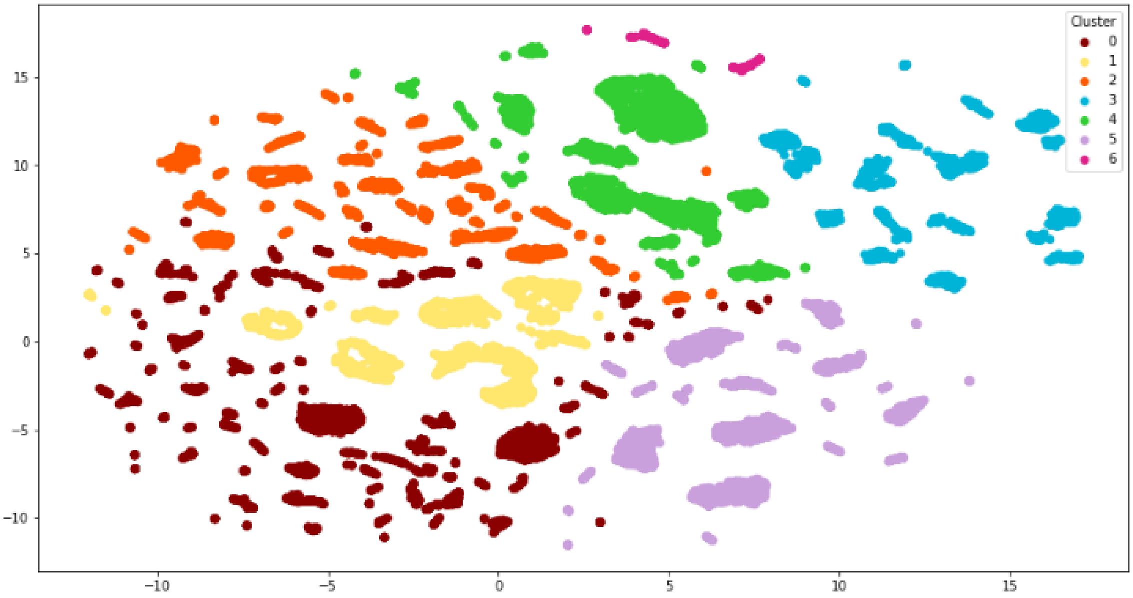


Figure 46 – UMAP of hierarchical with ward linkage



Annexe 4 – Box plots to characterize each cluster from notebook "3_Clusters"

Figure 47 – Lifetime spent in vegetables by cluster

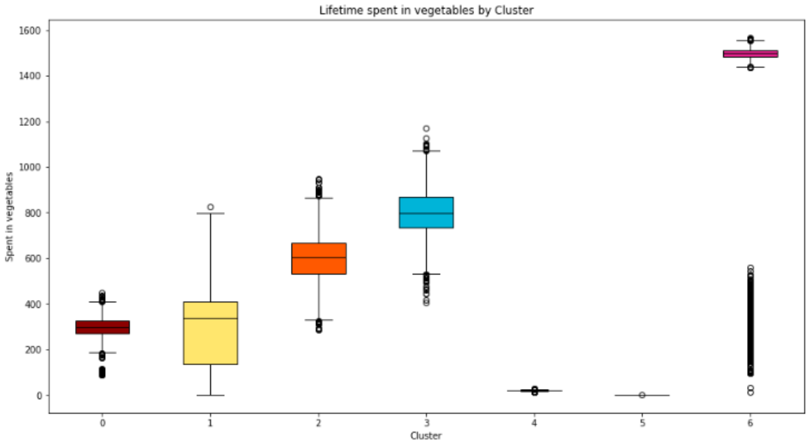


Figure 48 – Lifetime spent on non-alcoholic drinks by cluster

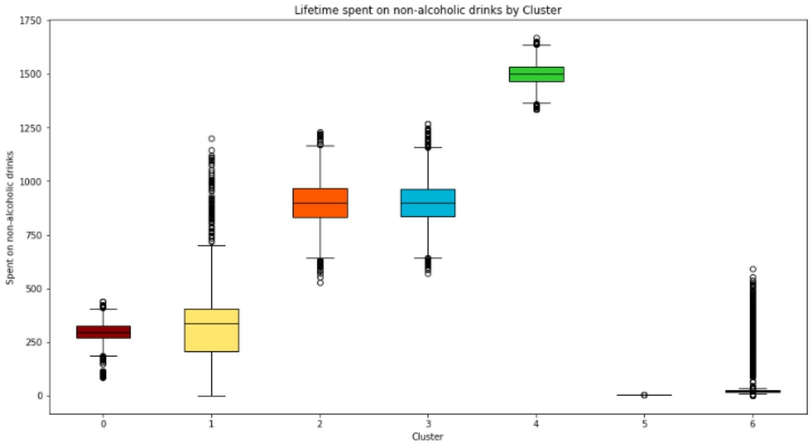


Figure 49 – Lifetime spent on alcoholic drinks by cluster

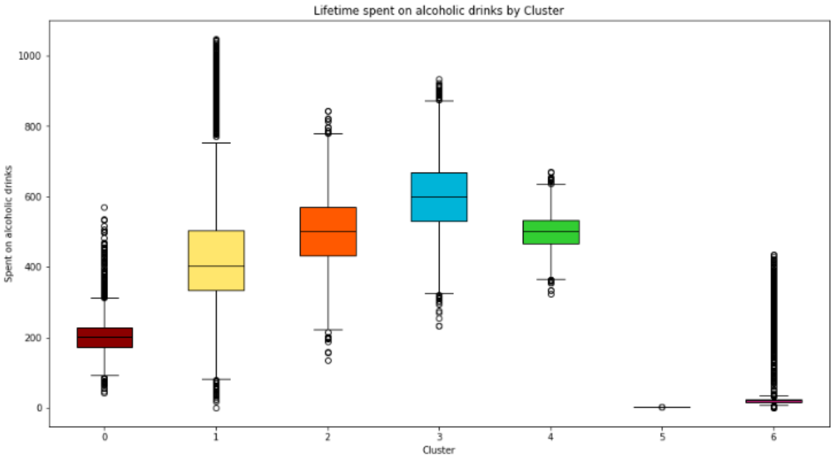


Figure 50 – Lifetime spent on meat by cluster

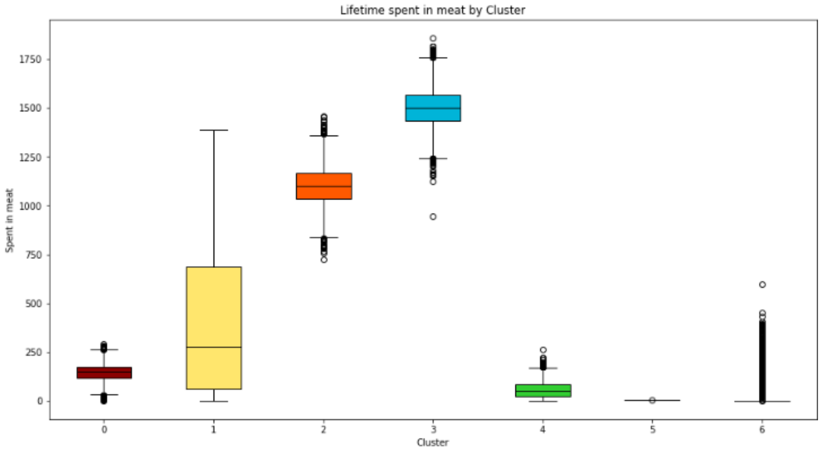


Figure 51 – Lifetime spent on hygiene products

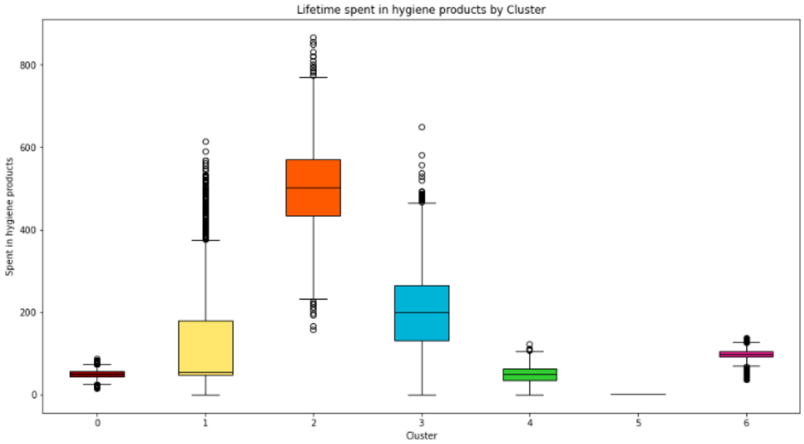


Figure 52 – Lifetime spent on videogames by cluster

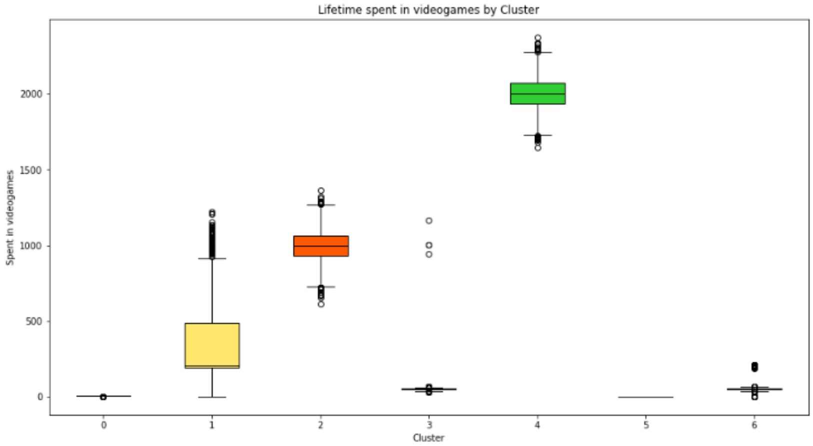


Figure 53 – Lifetime spent on groceries by cluster

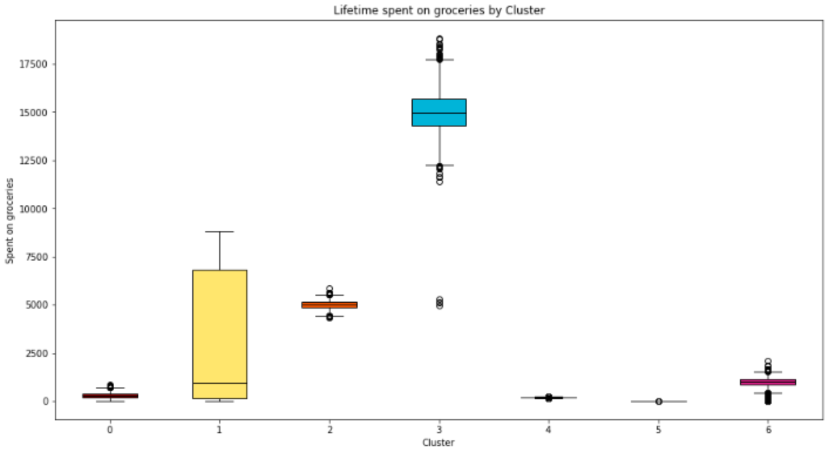


Figure 54 – Total of distinct products bought by cluster

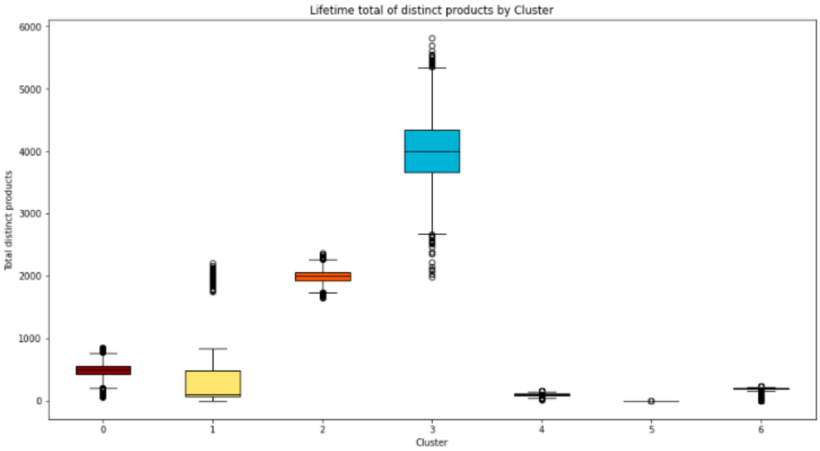
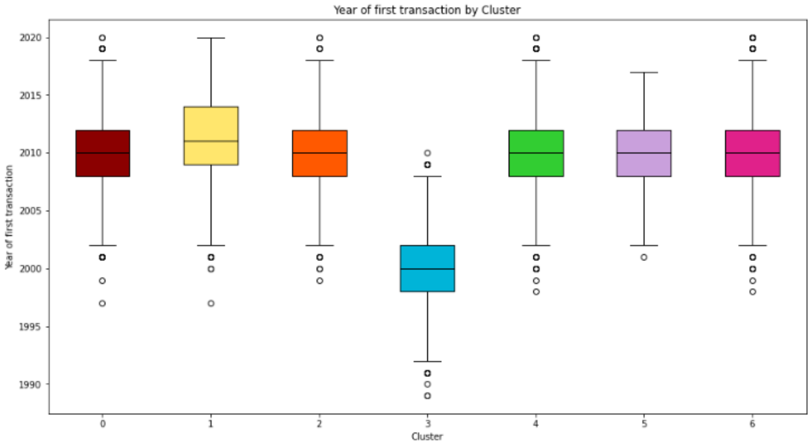


Figure 55 – Distribution of the year of first transaction by cluster


Year of first transaction by Cluster
Year of first transaction
Cluster