School Year 2022/2023 - Spring Semester

Machine Learning II
FINAL PROJECT REPORT

Group:

Daniel Kruk

Marcelo Junior

Index

Figure Index

Executive Summary

In today's highly competitive and overpopulated market, it is key for any company to understand who their customers are and their preferences so that the experience of any person can be tailored in order to meet one's specific needs.

The goal of this project was to execute customer segmentation analysis by identifying unique groups of customers based on their purchasing habits. This segmentation process holds immense value for companies, as it provides valuable insights into customers and facilitates the development of targeted marketing strategies. By maximising customer engagement and loyalty, businesses can gain a competitive edge.

After correcting a series of irregularities and assessing every variable, the customer segmentation process took place, where various clustering algorithms were tested and their results were compared. The one which yielded the best results was an HDBScan and eight different clusters were identified. These were the **Golden Oldies**, the **Promo Seekers**, the **Green Beans**, the **Family Friendly**, the **Bang-Average Consumers**, the **Young&Rich**, the **Geeks**, the **Young Alcoholics** and finally the **Loyals**. To these previously mentioned clusters, another one was added which was found during the Exploratory Data Analysis process - a cluster including nothing but **Supermarkets**.

By understanding the consumer demographics and purchasing patterns behind every cluster that was formed, a series of promotions were created in order to try and bolster the interest of these customers in shopping with the company. These explored the most usual and trustworthy combinations of products that were acquired together - by making use of association rules, which were created using a series of historic transactions from the withdrawn customer sample. It was attempted to create at least two promotions for every cluster that was created but in some cases more were developed. Some of the created promotions had a more general goal, of attracting more people regardless of their spending habits, some were tailored using as a basis their purchasing habits and some were created using the association rules mentioned before. To check the complete list one can move to the section Targeted Promotions of this same report, as they are properly organised and structured to promote an easy reading.

Exploratory Data Analysis

This process (also usually referred to as EDA, for simplicity's sake) was divided into two different steps. Firstly, a global view was taken of the initial dataset and some transformations took place during what we called a **Pre-Processing phase.** Then, during a more thorough examination, possible outliers were analysed, and using visual elements were used to check for the variables' distribution. This second phase was called the **Exploratory Data Analysis phase,** and the main goal here was to detect possible patterns and characteristics of the company's customers**.**

Note: It is worth noting that in this report, only the most important aspects will be mentioned regarding both phases mentioned above. To follow the analysis that was executed with more depth and detail, one can check the provided Jupyter Notebooks, which contain a more thorough and exhaustive analysis of the datasets.

Pre-Processing Phase

The first step consisted of analysing the dataset and looking out for existent irregularities within the data, while simultaneously refactoring some features. This refactoring could be required in order to enable some of the variables to be worked with in the future.

Inconsistencies Correction

During this stage, initially, a brief look at the descriptive statistics of every feature was taken and it was noted that some of the dataset's features had *-inf* values (more specifically *typical_hour,* and *lifetime_spend_videogames)*. It was found that every customer with those values were Supermarkets and these values were set to 0 as a placeholder not to affect the rest of the dataset. Because these observations will influence the clustering a lot - as they cannot be considered as humans - they will be later on separated from the dataset and put back only at the very end of the project.

Only one feature displayed missing values in the dataset - loyalty_card_number. Because not every customer must have a loyalty card number associated with their purchases the variable was left as it was.

Feature Engineering

This stage started with us changing the customer names, which were written featuring the highest education level (Bsc., Msc., Phd.). To allow for future analysis, these titles were split from the customer names and put into a new column, named *highest_education*. For those customers without a title, a *Missing* value was given. A basic encoding was then done to turn the string into numbers to make it more viable for clustering.

The next step was to change their *customer_birthdate* into their ages. The actual date was assumed to be 08/03/2023, with it being represented as MM/DD/YYYY in the code in accordance with the values in the column. By making the difference between their birthdate and the actual data, it was possible to know their *Age*.

The Gender feature was also turned into a binary feature, with 1 being assigned to males and 0 to females.

Three new variables were created as a precaution. The first one was *total_minors*, which combines *kids_home* with *teens_home*. The second was *total_spending*, which sums up every *lifetime_spend* variable. Finally, *lifetime_spend_tech* was done to combine *lifetime_spend_videogames* and *lifetime_spend_electronics*. This variable was created to fuse these two variables which could be combined into a single category but it was not deemed the final choice, but rather an option if the needs arise.

A new binary variable was created in regards to the *loyalty_card_number*, filled with 1 if the customers have a value in the previously mentioned column and 0 if not.

A correlation matrix was created to check if the original variable of the dataset and the new ones could pose some redundancy. Some variables showed to be highly correlated, which could indicate patterns to bear in mind for the clustering phase. In particular, *percentage_of_products_bought_promotion* and *distinct_stores_visited*, which can indicate that there are customers who go to the stores holding a promotion and are not particularly loyal to a single one. Another interesting relation is that of *lifetime_total_distinct_products* and *lifetime_spend_meat* and *lifetime_spend_groceries*, with they themselves also being highly positively correlated. The last non-self-explanatory relation is that *lifetime_spend_videgames* is highly correlated with *lifetime_spend_electronics* and *lifetime_spend_nonalcohol_drinks*, which might indicate another possible cluster.

Regarding the Product Mapping dataset, what was found is that there were 2 duplicate products: AirPods and asparagus, with the latter even having a different category. These products were removed and the AirPods kept were those pertaining to the class of electronics.

Finally, to conclude this Preprocessing step, the Supermarkets were separated from the dataset, as implied before, and the final datasets were exported - ***Customer Info PPC***, ***Supermarket Info*** and ***Product Mapping PPC***.

Exploratory Data Analysis phase

The main goal during this phase was to retrieve some information about our datasets in a more intuitive manner in order to realise what could be the "regular" patterns among our customers. To do so, a graphical exploration of every variable was made and also, relationships between different variables were explored.

Various aspects were noted - but as mentioned before, to check the complete analysis of the dataset, one can check the provided Notebooks that contain the complete overview of every dataset.

Customer_info dataset

Regarding the overall characteristics of the selected customer pole, it can be stated that the **gender distribution** is **evenly balanced** with **50.34%** of our customers **being males** and the **remaining 49.66% being females.** Overall the **age distribution** among our customers is also **evenly balanced** with the possible ages **ranging from the early-20s** all the way down **to the mid-80s**, with every inserted age value having about 500 observations. Generally speaking, the **number of complaints** made by the customers selected **tends to be low** - with the **majority of the cases being between 0** (meaning no complaints) **and 1 complaint -** yet it could be worth noting that there are still some customers who make a lot of complaints with the remaining observations ranging from 2 complaints all the way up to 8 complaints per customer.

The **majority** of human customers **do not have a loyalty card (80.51%)**, with the **percentage of clients with loyalty cards** being of **corresponding to just 5802** of the 29774

clients withdrawn in this dataset **(19.49%)** and it is also important to mention that out of the sample of customers analysed in this project, a **vast majority made their first purchase** with the company **in the years surrounding 2010** (before and afterwards), with the **first dated transaction** coming from the **year of 1990** and the **latest dating from 2020**.

The sampled customers usually visit the stores with a **higher influx during the hours** which come **prior to the start of a day of work** (surrounding 9.00 a.m) **or** in the hours **following the end of a normal day of work** (surrounding 6.00 p.m). There is **also a high incoming of customers** in the hours coming **after dinner** - probably with some of our visitors being customers that are arriving late home and that on their way back home are passing by the stores to acquire some products. The **distribution of the customer inflow** is represented in the line plot **below**.

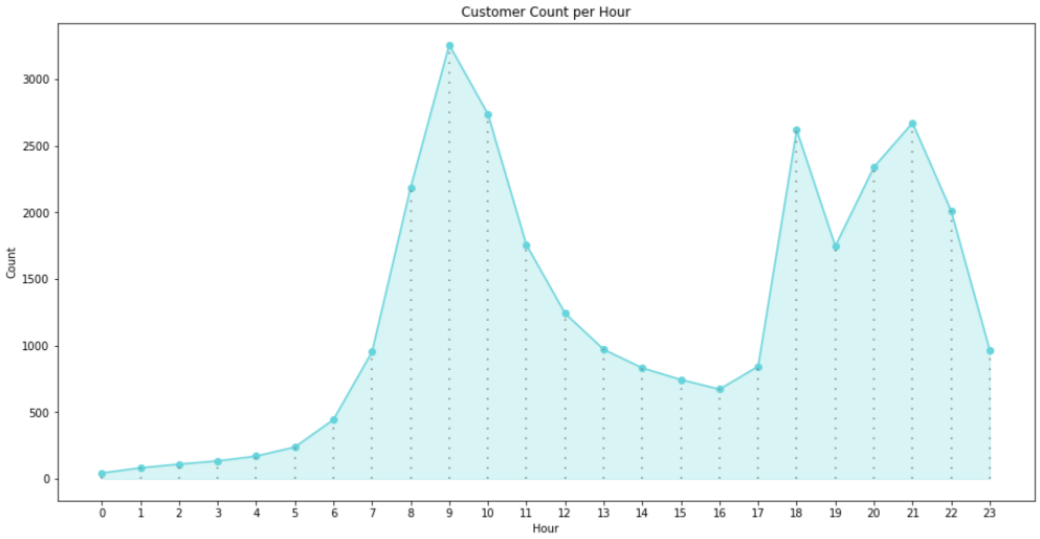


Fig.1 - Line Plot displaying hourly customer inflow

Another curious aspect regarding the demographic characteristics of our customers is that the **majority of the families** represented in the dataset **have 0 to 2 minors per household** with **some ranging** all the way **up to 14**. An analysis was conducted on these cases which displayed a bigger and less common number of minors per household - to verify if the observations were not outliers - and it was concluded that these observations appeared to be indeed real and actual customers (as they did not possess any other irregular values for any other feature. This analysis can be followed using the Notebooks already mentioned previously.

In terms of the money that is spent, various analyses took place. Regarding the money that is spent per category, it can be stated that per category the splits every customer spends do not follow any kind of distribution, which is relatively normal. The **treemap** displayed **below shows** what is the **comparison of the total money** that is **spent between every category** and it is clear that there are some categories which seem to be the most common for the customers to spend their money on.

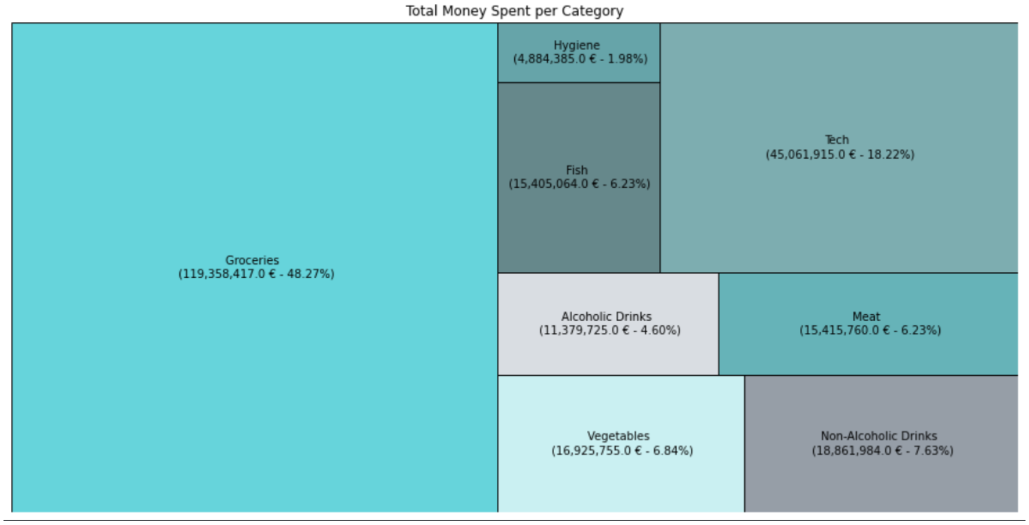


Fig.2 - Treemap of Total Money Spent per Category

An analysis regarding the money that was spent per education level was also conducted, but the graphical exploration showed that the dispersion of money spent per education level was equal all around.

An interesting curiosity regarding the observations is that there are **several customers** who **were found to be visitors to multiple stores**. The **majority visited only 1 or up to 3 stores**, but **there are customers who were found to visit around 20 stores**, as is displayed in the bar chart below.

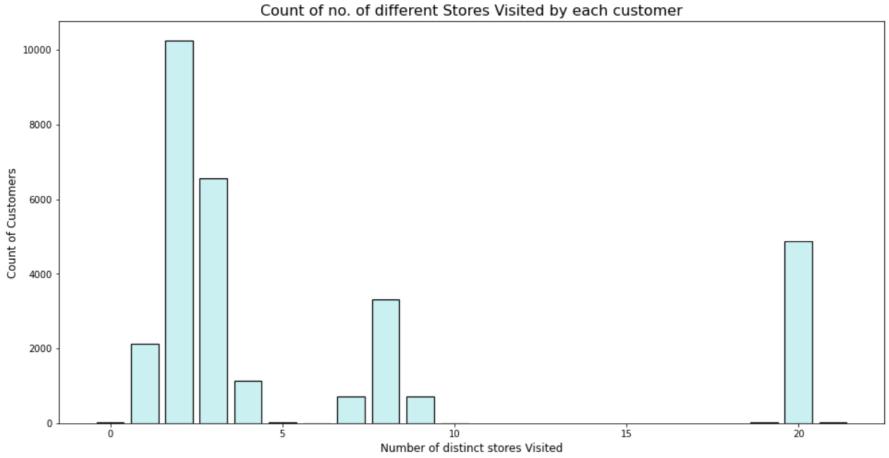


Fig. 3 - Distribution of stores visited

Since in the dataset, there was a variable which regarded the percentage of products bought in promotions (*percentage_of_products_bought_promotion*), the instant reasoning was that **these customers could be promotion seekers**. This **relation was checked** and indeed **confirmed** using the boxplot which is displayed below.

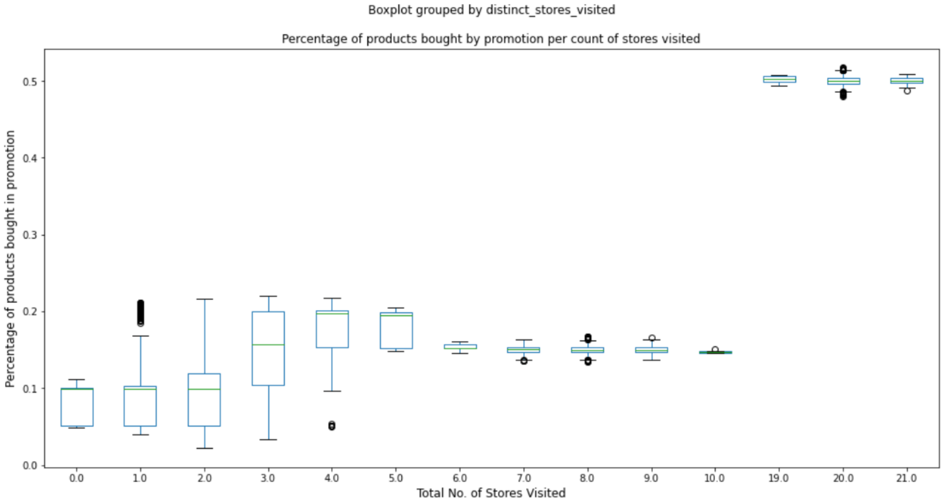


Fig.4 - Distribution of the percentage of products bought in promotion ber stores visited

Finally, and to establish a bridge between the analysis of the customer_info dataset and the supermarket_info dataset, a map was created to verify what the **overall geographical distribution of our customers** looked like.

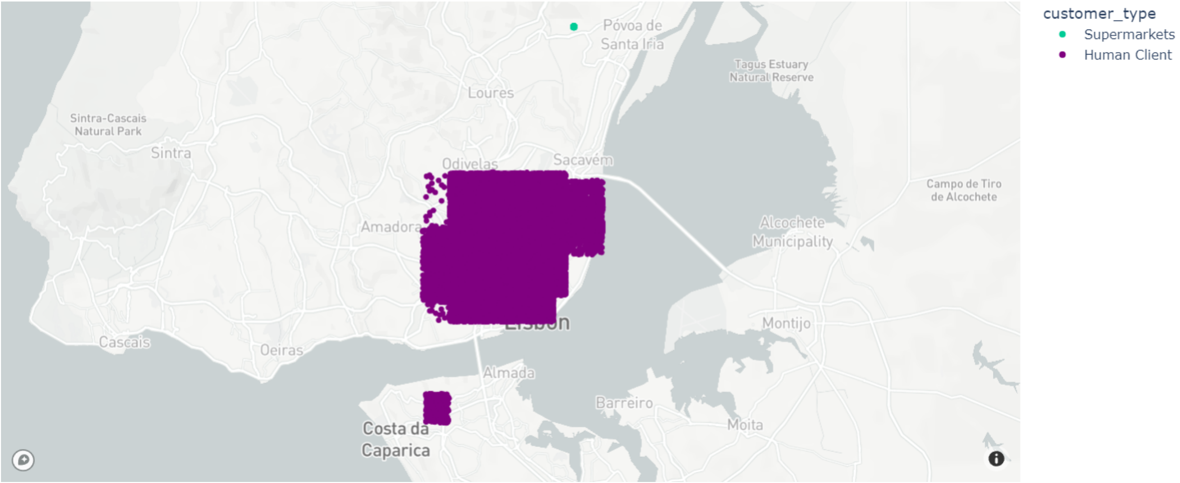


Fig. 5 - Geographical Distribution of all Customers

It is curious to verify that **all supermarkets are located around the same area**, near *Póvoa de Santa Iria*. This could be due to the fact that in that region is located in the Lisbon Region Supply Market, which could mean that all supermarkets refill their stocks from the same source point. Also and interestingly enough we can see that near the area of *Costa de Caparica*, there is a small group of observations which are grouped together. These could be distributed in the shown way as in the region there is a University meaning all observations could correspond to students.

Supermarket_info dataset

Regarding the dataset containing the information pertaining to the supermarkets previously separated, there are not many analyses that can be extracted as the majority of the dataset's columns regard characteristics which are more human-like than anything else. Some variables weren't even considered for analysis, mainly variables like *highest_education*, *customer_gender*, *kids_home*, etc. for the reason previously mentioned but also, some were not analysed with more depth, due to the fact that whilst making an overall verification of this dataset's descriptive statistics it was verified that the features were univariate - with the value being the same for every observation. Cases like such include *number_complaints* and *distintc_stores_visited* where all values were 0 and 1 respectively - meaning that **no supermarket ever made a complaint** and **all supermarkets buy their products in the same origin point** (just as was mentioned in the prior geographical analysis).

Some of the more interesting insights taken regarding the supermarkets include the fact that regarding the **loyalty card percentage**, as it was verified for the human customers the **percentage of customers with loyalty cards is much smaller (10.18%)** in comparison to that of **customers without (89.92%)** and also that **the category in which more products are bought is the Fish Category by a very considerably large margin**. In terms of their **tenure** with our company, it can be said that these values follow a normal distribution, with the **majority of supermarkets making their first transaction with us in the years surrounding 2010**.

Customer_Basket dataset

This dataset was not analysed in detail as it only contained information that would be pertinent for the latter stages of the creation of association rules.

Product_mapping dataset

This dataset only contained a list containing all possible products that could be bought. No further analysis was made.

Customer Segmentation

To start working towards the most important and central part of this project, we began with the scaling of the data, due to the reasons mentioned previously during the preprocessing phase (that the scale of certain variables was very large in comparison to others), and applying a PCA to it. This decision was also due to the existence of over 20 features in the dataset, with some of them being highly correlated. The PCAs can be interpreted in the following way:

- In PC0, *lifetime_spend_meat* and *lifetime_spend_fish* have the highest positive influence, whereas *year_first_transaction* has the most negative influence. We could call it Meat Eaters being represented.
- For PC1, *lifetime_spend_videogames* has the highest positive influence and *lifetime_spend_vegetables* the most negative one. These would be the Gamers represented.
- Now for PC2, we have that *total_minors* has the biggest positive coefficient and *distinct_stores_visited*, with *percentage_of_products_bought_promotion* also close, the most negative one, even if not much. These are the Loyal Families being represented.
- Regarding PC3, *percentage_of_products_bought_promotion* has the highest coefficient, followed closely by *distinct_stores_visited*, with *lifetime_spend_vegetables* and *highest_education* having the lowest negative values. This would be the Uneducated Store Hoppers, or Promotion Seekers, being represented.
- Finally, PC4 has *lifetime_spend_alcohol_drinks* as its positive highest coefficient and *Age* the lowest, thus they are Young Alcoholics represented.

This transformed dataset was then used on every clustering algorithm used, but the clusters were analysed based on the mean of the original features and not of the PCA, thus avoiding dimensionality problems while not losing interpretability on the clusters.

The way this part was thought of was, to begin with a K-Means and compare the number of clusters to the Hierarchical solution. Where their cluster numbers are the same, that solution would immediately go into consideration

The first algorithm used was K-Means. To choose the best number of clusters, the Elbow method was used, which showed 6 clusters to be the optimal solution, while after 9 clusters there was basically no change. As such, both numbers were considered. Analysing the means, 6 clusters were shown to be too little, while 9 clusters were not as distinct from one another as one would hope. Another problem of the 9 clusters was that many clusters did not have many customers. Nonetheless, these solutions shall be compared to those of the next algorithm, so as to see whether they are to be considered optimal.

Secondly, Hierarchical Clustering was tried, using the Silhouette score and Dendrograms for references. The first metric showed Ward's method was the overall best one to choose, while Average came just behind. While the other linkages were used, they fared terribly on this metric and, as such, were disregarded. In the latter, 9 clusters were shown to be a good reference, with 6 clusters combining too many distinct clusters. Suffering from the same problems as the K-Means, a middle ground was found at 8 clusters, which also seemed to be a reasonable cut-point.

In regards to the MeanShift, it was found, through a grid search, that the bandwidth value of 2.2 yielded the best results, as the default value only produced a single cluster. With this optimised parameter, the MeanShift algorithm was subsequently employed to obtain a solution consisting of 7 clusters. Although the solution was not flawless, it managed to capture a significant amount of patterns.

For the DBSCAN algorithm, another grid search was conducted to determine the optimal values for the epsilon and minimum sample parameters. It was determined that an epsilon value of 1 and a minimum sample value of 300 provided the best results. Using these parameters, the DBSCAN algorithm produced a solution consisting of 7 clusters, with 459 points classified as noise and assigned a label of -1.

Upon further analysis, it became apparent that some of the identified clusters exhibited a high degree of similarity and lacked significant differentiation. This observation made the solution generated by the DBSCAN algorithm less reliable and somewhat questionable, basically discarding it from consideration.

After all of these algorithms were done, they were compared to one another on the basis of a UMAP representation of the dataset. This representation thus exposed many of the problems of the solutions, such as distant clusters being put together under the K-Means with 6 clusters, and with a shape on the UMAP having 2 clusters within it, such was the case of the

K-Means with 9 clusters. With none of the options seeming too reliable at finding the patterns in the data, and in search of a better solution, a new algorithm was tested - HDBScan.

This algorithm extends the aforementioned DBScan algorithm by combining it with a Hierarchical Clustering algorithm (also like the one previously mentioned). HDBScan will work by extending the original DBScan into a Hierarchical Clustering algorithm and then using a technique to extract a flat clustering based on the stability of the obtained clusters. The data's hierarchical structure is captured using what is called a minimal spanning tree. It builds a cluster hierarchy and extracts the most stable clusters by looking at core distances and mutual reachability. Another very positive aspect of this algorithm is that it recognizes outliers as noise.

The application of this algorithm was made by embedding the non-scaled dataset into a UMAP. This would not be used for the final representation, but to train the model. After some trial and error, a minimum sample value of 10 and minimum cluster size of 1000 was found to yield quite reliable results. The solution yielded 8 clusters with some points being assigned the label -1. On that initialisation of the UMAP, those points were together and far from any other clusters, not having sufficient individuals to generate a new one.

Due to the fact that HDBScan managed to capture every trend the other algorithms tested got, while also identifying a group of outliers, which mostly consisted of older customers who did not spend much, this model was chosen to be the final. Its representation in a UMAP cemented the decision, due to how precise it was.

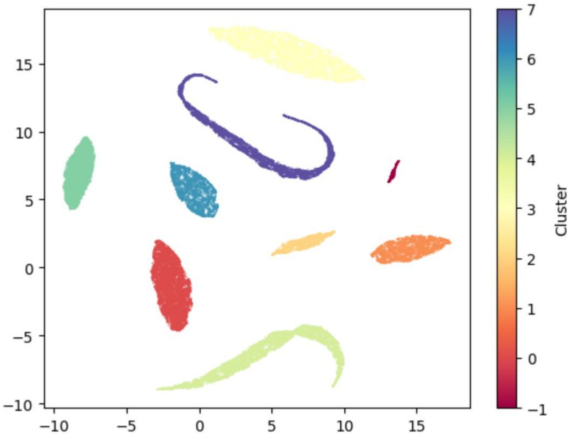


Fig. 6 - Final UMAP Cluster Representation

With the final solution chosen, the clusters and outliers were to be given names, these go as follows:

- **Golden Oldies (-1)**: These customers cannot be really considered a cluster as they were separated from the remaining observations for not fitting any specific case while at the same time creating their own pattern. These clients have a mean age of around 70 years old and do not spend an extraordinary amount in any specific category. Like any human being of a higher age, they tend to make more complaints (approximately 4 per customer);
- **Promo Seekers (Cluster 0):** They visit around 20 stores seeking promotions in general, which can also be verified as they have the highest percentage of products bought in promotion (with approximately 50% of their purchases being made in promotion). As one could expect, their total spending is also one of the lowest values among all clusters;
- **Young&Rich (Cluster 1):** This cluster includes young adults who spend a lot of money in comparison to the majority of the defined clusters. People around 28 years of age, some already having kids which spend way more money than the previously defined cluster of average clients (hence being separated from the bang-average clients);
- **Young Alcoholics (Cluster 2):** This segment of clients is the one which was already envisioned in the creation of the map. A group of University students (whose age revolves around 22 years) that spends the majority of their money buying alcoholic drinks as all university students do. They are also very recent customers and most do not own a loyalty card;
- **Loyals (Cluster 3):** This cluster includes the majority of the people who own a loyalty card and the people who have spent the most money in our company (almost totally buying groceries);
- **Geeks (Cluster 4):** This group of clients is known to spend the majority if not almost the totality of its money buying tech products (whether it is electronics or videogames). Other than that they possess no other specific characteristic;
- **Family Friendly (Cluster 5):** This cluster englobes those clients that make their purchases essentially in the representation of an entire big family. In this case, we are talking about parents of many minors (approximately 5 per household) who spend a lot of money essentially providing groceries for their family;

- **Bang-Average Consumers (Cluster 6):** This group of customers has no extraordinary pattern to highlight. A group of people that doesn't spend a lot of money in any specific category with the totals spent per category being very average and low all around;
- **Green Beans (Cluster 7):** These clients spend absolutely no money on meat or fish products and spend almost all of their budget buying vegetables (as expected). They possess no other specific characteristic;
- **Supermarkets (Previously Separated Group):** This cluster represents the supermarkets which are essentially the highest spenders within the withdrawn dataset. They are those who spend the most money and buy almost nothing but fish.

Finally, their size comparison is presented in the following graph:

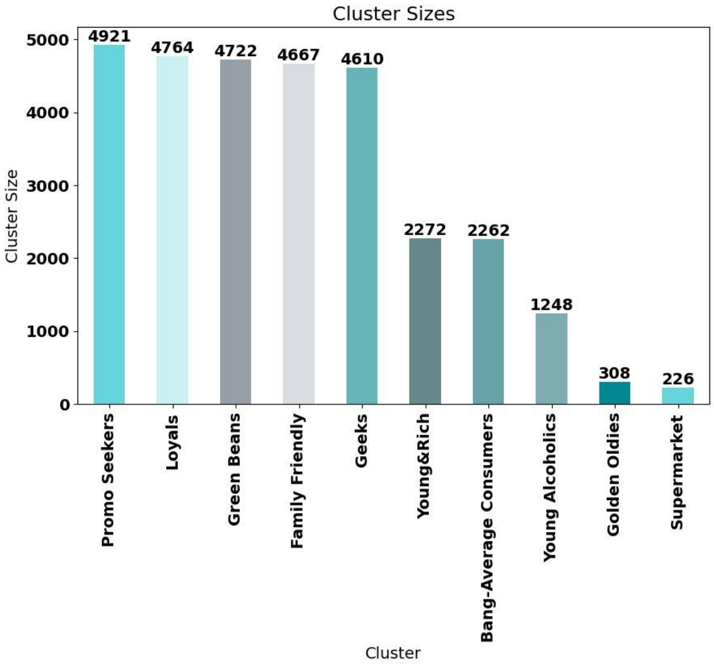


Fig. 7 - Cluster Size Comparison

As one can note the created clusters' sizes vary with the biggest clusters having more than 4000 clients within them and the smallest ones containing just about 300. The cluster which the algorithm deemed as -1 is one whose creation might be a little controversial since the model grouped the observations together due to the fact of them not belonging to any of the other created clusters. Curiously enough, these "outliers" formed a pattern by themselves which indicates that possibly if more observations were considered, an actual cluster would be formed by the algorithm giving more information about this segment in particular.

Targeted Promotions

Having sampled the clusters, association rules were created in order to try and create promotions to attract these customers to spend more money with the company. In this section, we will be stating the created promotions we believe the company should offer/ present to every group of clients in order to attract more interest on their behalf. To create the promotions, strong consideration was given to the lift value and confidence level associated with the association rules that were created. Some promotions will be given names to stand out, due to how creative they seem which could cause for more attractiveness.

Some examples of promotions we believe the company should target are:

- A general offer **for all customers**, in order **to boost the number of loyalty cards** existent among the customers: a promotion giving **a global 10% off discount on the total cost of the next purchase for the creation of a new loyalty card**;

For the Golden Oldies Cluster:

- *To have a barbeque for the grandchildren:* In the case of a **purchase of 1 kilogram of ground beef**, the client is **offered a 15% discount on the purchase of cooking oil;**
- **Overall discount of 20% in the purchase of any products from the groceries category;**

For the Promo Seekers Cluster:

- **Overall 15% discount in the next purchase**;
- In the case of **a purchase amounting to at least 30.00€** a **discount of 20% is given** in the **purchase of cooking oil;**
- (any promotion from the remaining clusters in general can be used, as they seem to enjoy buying products which are in promotion);

For the Young&Rich:

- *Wake n' Bake:* In case of a **purchase containing any kind of pastry (muffins, cake, etc.) is made**, the customer is **offered a 10% discount** in the **next purchase**;

- A **take 3 and pay 2 promotion in the purchase of any type of candy bars;**

For the Young Alcoholics:

- In case of a **purchase with 2 bottles of french wine**, the client is **offered a free bottle of dessert wine of the same brand**;
- *For you to bar like a rapper:* **Overall discount of 10%** in the purchase of **any type of bars** (protein bars, candy bars, etc.)
- *For you and your loyal friends to party:* **Overall 15% discount** in the **purchase of cider or beer** if you create to the loyalty card

For the Loyals:

- *It's Been a long time with you, dear friend:* Due to their long-lasting tenure with the company, **clients belonging to this group will receive a 40% discount on their next purchase**;
- *Rewarding the Loyals:* **For those** who belong in this category **that still don't have a loyalty card**, **if one is created** they **receive an extra 10% discount to the 40%** mentioned in the promotion above;

For the Geeks:

- In case of a **purchase of any Pokémon Game** the client is **offered a 25% discount** in the **purchase of another Pokemon Game**;
- *For a Better Sound Experience*: In case of a purchase is made containing a pair of Bluetooth Headphones and at least 2 games, the customer receives a 10% discount with an additional 5% being added for every additional game that is bought (mounting up to 20%);

For the Family Friendly cluster:

- *Ginger Carrot Soup with Fromage Blanc for the family!:* For the course of a week, customers who have a loyalty card and with at least 4 children get a whopping **20% off fromage blanc and soup** for the family to enjoy!

- *What would we be without our parents!?:* For those numerous families who adhere to the loyalty card program, they get a **15€ discount** on their next purchase to help them in this time of need and inflation.

For the Bang-Average Consumers:

- In case of a **purchase of a regular 6-pack of Black Beer**, the customer is **offered a 20% discount** if he/she decides **to buy a 6-pack of Normal Beer;**

- *Grape Success:* In case of a **purchase of any bottle of wine,** the customer **receives a 15% discount** in the purchase of a **bottle of another bottle of wine;**

- *Refresh the tech:* **An overall 15% discount in the purchase of any "tech" products;**

For the Green Beans:

- In case of a **purchase including at least 25.00€, the customer is enrolled in a draw where the winner will win a copy of any pokemon game of choice**;
- *Healthy Living for you and Your Pet:* In the case of a purchase of 3 cans of pet food and any quantity of any vegetable, the customer receives a 15% discount on the purchase of another vegetable;

For the Supermarkets:

- Due to their extremely high levels of money spent on fish products, one promotion we decided to propose is **if they purchase more than 30 kilograms of fish products**, **a 20% discount is given on the purchase of any given products from any other category;**

- Because they are our highest spending group of customers it was thought that **as a compensation prize a global 40% off on their next purchase could be given;**

- **Overall 15% discount on the purchase of any oil (oil, cooking oil, etc.)**

Conclusion

To conclude the project we can say that the final objective proposed was reached - with a very solid representation of the various groups of customers existent within the company being found and a series of promotions being created in order to attract and stimulate more the consuming spirits of our clients.

By utilising the HDBScan clustering algorithm, a grouping of the withdrawn sample of customers was made by joining the observations according to their characteristics and spending patterns. By finding a total of 10 different clusters - 9 of them using the aforementioned clustering and 1 during a simple geographical observation of the data, it was interesting to realise that there exist a series of characteristics which are common to various persons and also, it was interesting to explore ways to attract these groups of customers by exploring their interests and tastes.

However, we believe the project could have had a better ending to it if there was more data regarding the transactions of every client. With the given data regarding the purchases, it was hard to create strong association rules in order to develop decent promotions for some of the clusters that could stimulate the interest of the already existing customers .

Nevertheless, due to the reasons previously mentioned, one can conclude that overall the project succeeded in its goals and it was possible to obtain valuable solutions that can help the company attract more investment and interest from the customers.

References

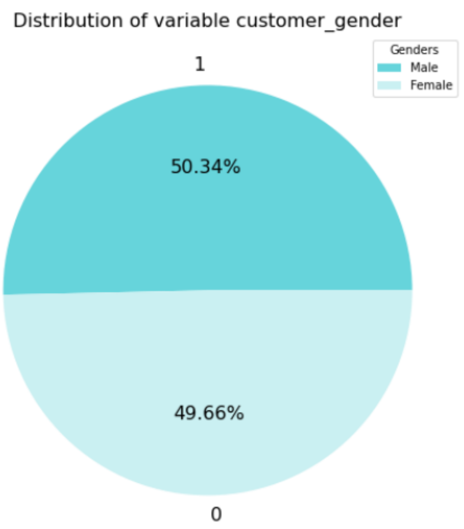
How HDBSCAN Works — hdbscan 0.8.1 documentation.

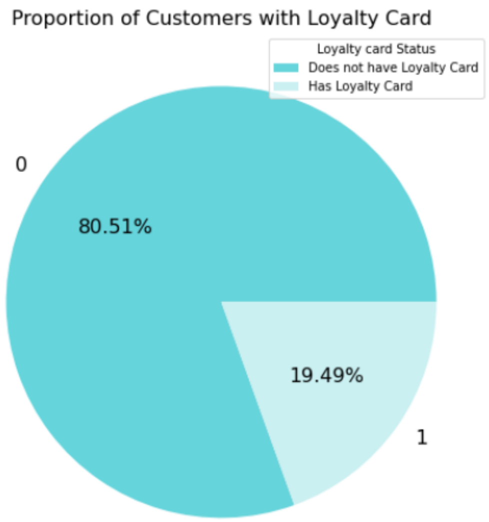
(n.d.).https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html

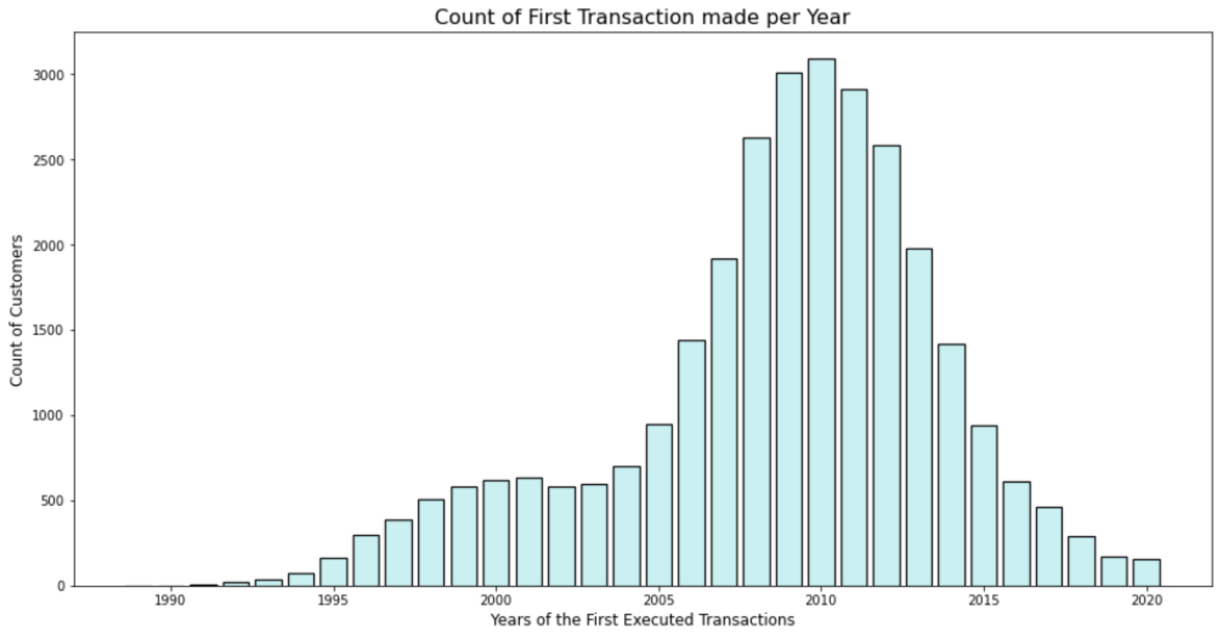
scikit-learn-contrib/hdbscan. (n.d.). GitHub Repository. Retrieved from https://github.com/scikit-learn-contrib/hdbscan

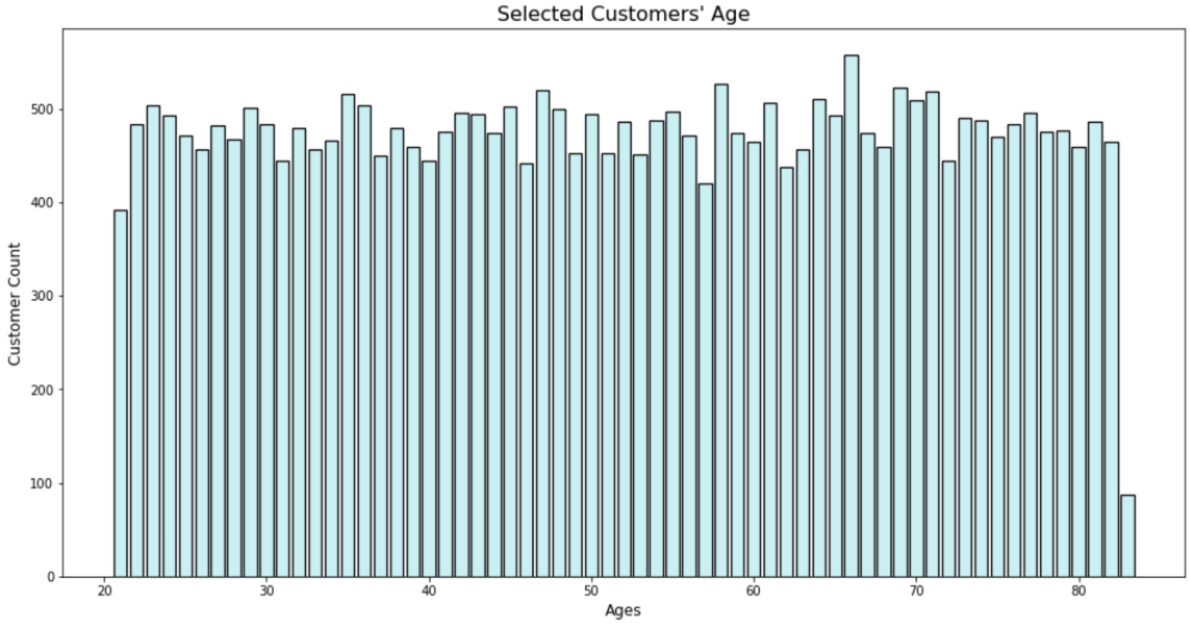
Class codes provided by the professor Ivo Bernardo

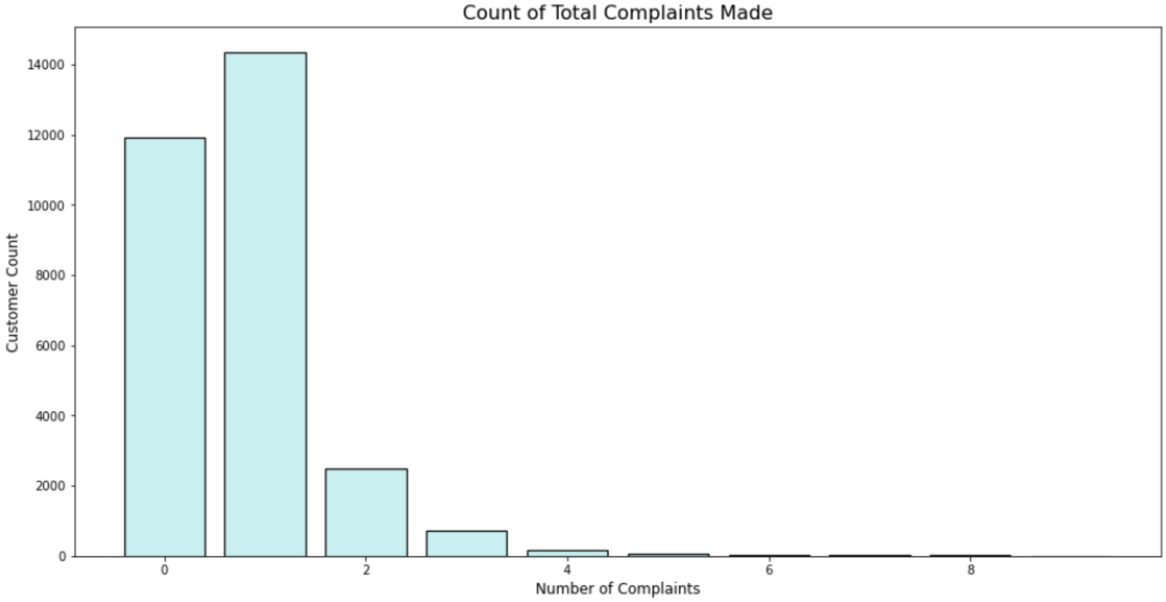
Annexes

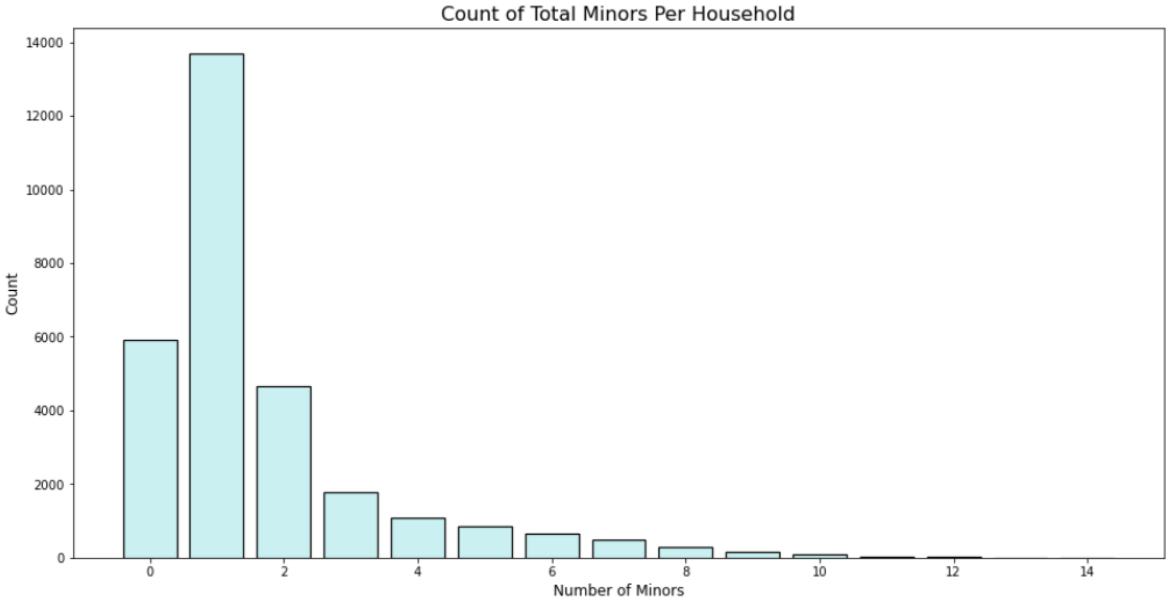

Distribution of variable customer_gender
Genders
Male
Female
1
50.34%
49.66%
0

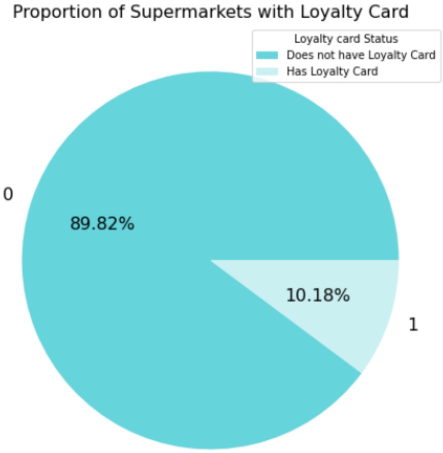

Proportion of Customers with Loyalty Card
Loyalty card Status
Does not have Loyalty Card
Has Loyalty Card
0
80.51%
19.49%
1

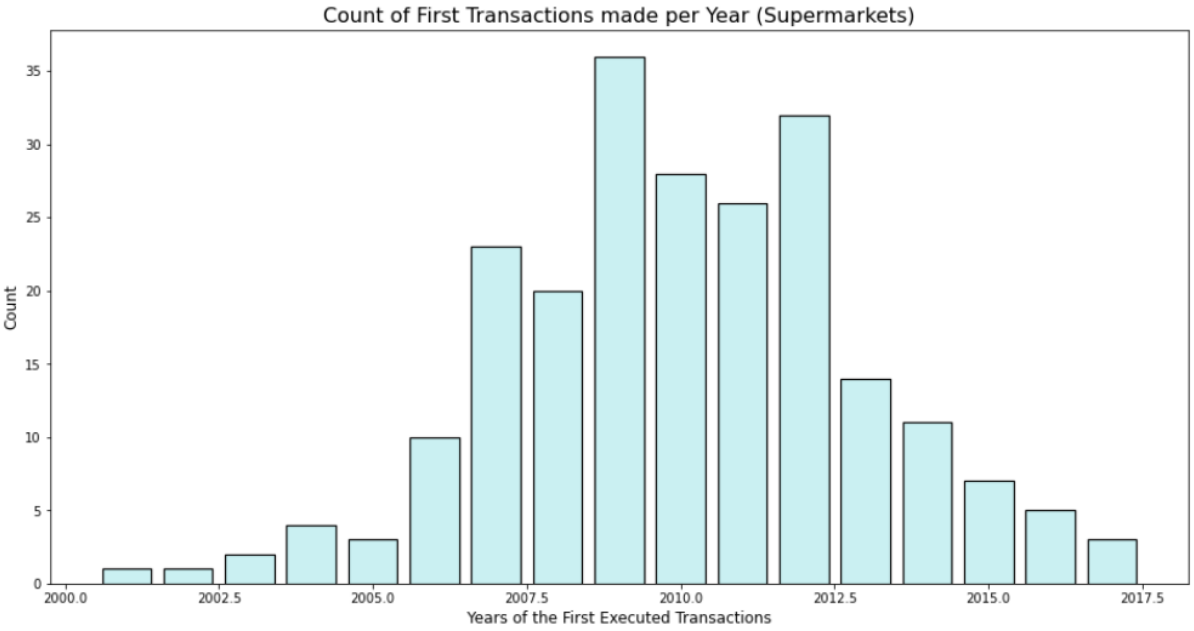

Count of First Transaction made per Year
Count of Customers
0
500
1000
1500
2000
2500
3000
1990
1995
2000
2005
2010
2015
2020
Years of the First Executed Transactions

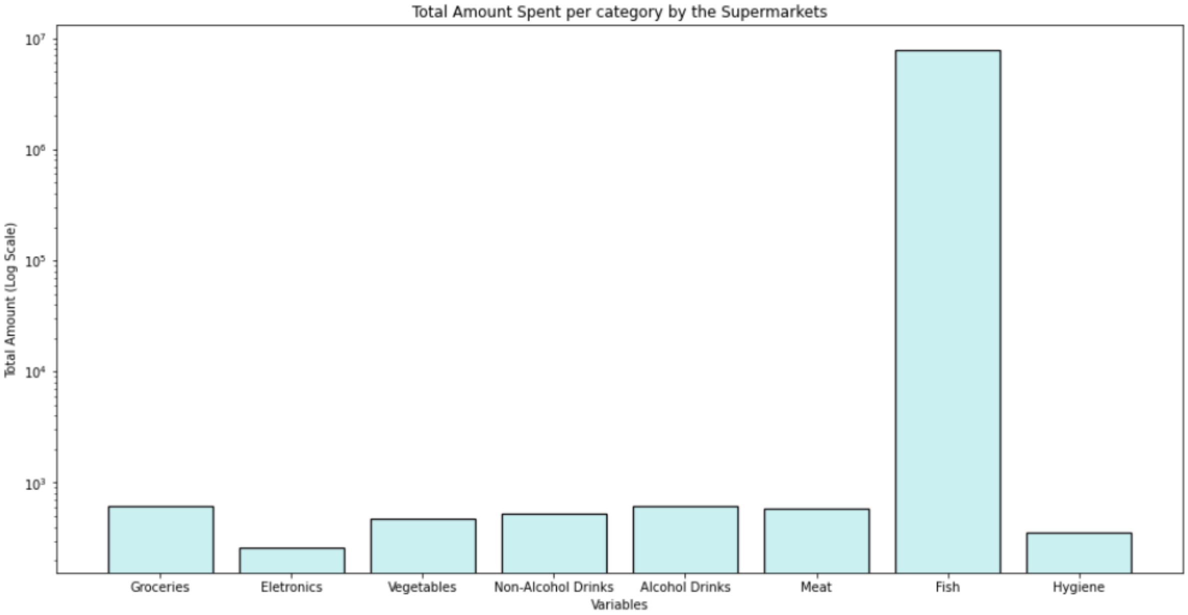

Selected Customers' Age
Customer Count
0
100
200
300
400
500
20
30
40
50
60
70
80
Ages


Count of Total Complaints Made
Customer Count
0
2000
4000
6000
8000
10000
12000
14000
0
2
4
6
8
Number of Complaints


Count of Total Minors Per Household
Count
14000
12000
10000
8000
6000
4000
2000
0
0
2
4
6
8
10
12
14
Number of Minors


Proportion of Supermarkets with Loyalty Card
Loyalty card Status
Does not have Loyalty Card
Has Loyalty Card
0
89.82%
10.18%
1


Count of First Transactions made per Year (Supermarkets)
Count
0
5
10
15
20
25
30
35
2000.0
2002.5
2005.0
2007.5
2010.0
2012.5
2015.0
2017.5
Years of the First Executed Transactions


Total Amount Spent per category by the Supermarkets
Total Amount (Log Scale)
10^3
10^4
10^5
10^6
10^7
Groceries
Eletronics
Vegetables
Non-Alcohol Drinks
Alcohol Drinks
Meat
Fish
Hygiene
Variables