Machine Learning II Project

Data Science Bachelor's Degree

Spring Semester '23

Project Developed by:

Afonso Cadete | 20211519

Joana Rosa | 20211516

Rita Centeno | 20211579

Table of Contents

Executive Summary

The main purpose of this project was to be able to perform customer segmentation - which consists of the division of a large customer base into smaller groups based on their similar characteristics - on data from a supermarket chain through the implementation of unsupervised machine learning techniques. For this, customer behaviour was analysed in order to create better and tailored marketing strategies.

On an initial stage, the data regarding customer's information was imported and an exploratory data analysis process was carried out. In this phase, the names of some of the variables were modified while other were created for interpretation purposes; missing values and data types were also verified. In addition, a map of the Lisbon metropolitan area was generated to allow a better visualization of the distribution of the supermarkets and clients locations. Furthermore, the distribution of the variables of this dataset was checked, as well as some graphics were plotted regarding gender of clients, their level of education, number of children, customer antiquity, among many others. Alongside this, inconsistencies were checked and corrected. Finally, the now pre-processed data was exported into a new csv file.

The following step consisted on the scaling of the data, conducting a principal components analysis (PCA), and implementing clustering techniques; such as K-Means, Hierarchical Clustering and Mean Shift. Later the clusters were visualized in a multidimensional space, using UMAP, and a final clustering solution was reached.

The final clustering solution consisted of 10 different groups to which the following names were given: Veggies, Parents, Gamers, Young Adults, Spenders, Alcoholics, Low Standards, Promo Hunters and Olds.

The clusters were carefully analysed, and several marketing campaigns were created for each one of them (Targeted Promotions). Some examples worth noting are: BOGO discounts (Buy one, Get one free), wine tasting exclusive pack for the alcoholic cluster, vegetables basket for veggies cluster campaign, weekly discounts for low-standards whose results shall be analysed after a defined period of time and even basket checkout discounts (for instance, on a 50€ purchase, it is possible to acquire a 5€ discount for the next transaction).

By successfully segmenting our customer base through machine learning techniques, it was possible to gain valuable insights which will enhance the company's marketing efforts, personalize customer experiences, and drive business growth. Hence, the implementation of the aforementioned strategies is recommended in order to capitalize on the unique characteristics of each customer segment; as well as, deliver targeted solutions that align with their preferences and behaviours.

Exploratory Data Analysis

Our exploratory data analysis was divided into four main phases: Data Pre-processing, Data Visualization, Correlation Analysis and Inconsistencies Correction.

Data Pre-Processing

As an initial step for the data pre-processing phase, the names of certain variables were changed. Ten of the variables contained the word "lifetime", and since the data provided only concerned the last two years, this word was dropped from the variables' name.

Next, data types and missing values were checked. The only variable containing missing values was the *loyalty_card_number*, which had 5825 registered values from the original 30000. In what regards data types, only *customer_name*, *customer_gender* and *customer_birthdate* were objects, while the remaining ones were considered floats, therefore some changes were made in the data types so that they make more sense given the data. The *customer_gender* variable was changed into a binary variable, the *customer_birthdate* variable was replaced with *customer_age*, *year_first_transaction* was replaced with *customer_antiquity* and, finally, *loyalty_card_number* was replaced with the binary variable *customer_loyalty*, which indicates if the customer has a customer loyalty card or not.

Apart from the variables already mentioned above, other nine variables were created. For each spending category, a percentage spent on that category was calculated in order to better understand the weight that each category has in the consumption of each customer.

Looking at the *customer_name* it was easily noticeable that that there was a distinct group of customers which represented supermarkets. Therefore, the original dataset was divided into a dataset called supermarkets and a dataset called people. These two types of customers were also plotted on a map (Figure 1), allowing to visualize the geographic distinction in their location. Still using the same variable, it was clear that the degree of education could be extracted for some of the customers. This information was stored in a new variable called customer_education.

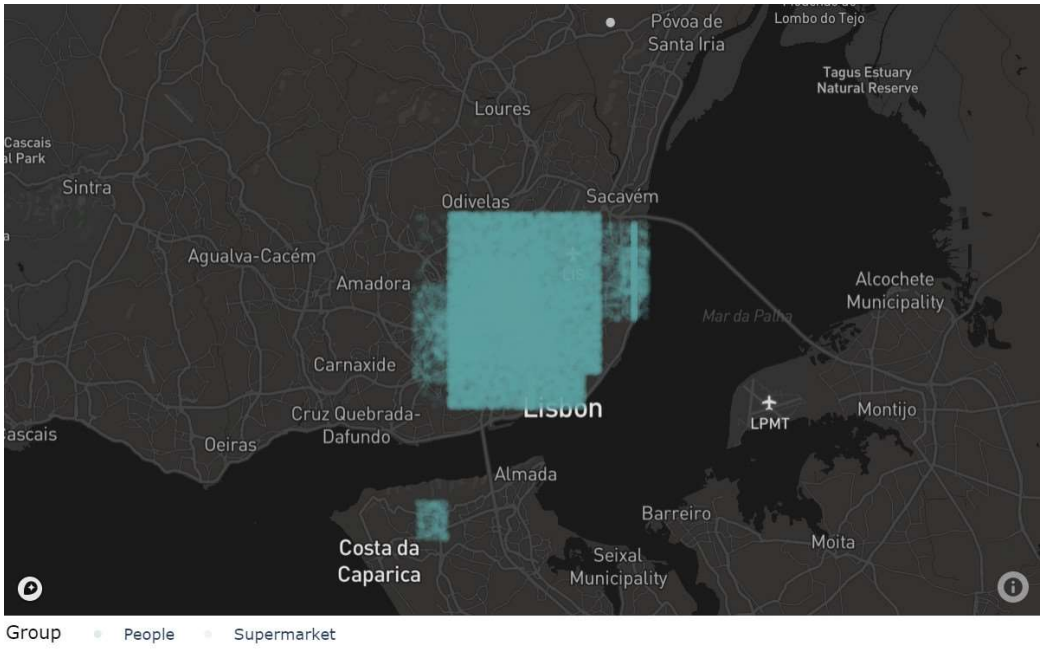

Figure 1: Customers locations by group

Starting by looking at the people dataset, the columns were reindexed so that they are better organized, by category. In addition to that, and following the previous analysis of the data types, a study was made to determine which variables were actually floats and which could, and would later, be converted into integers.

Data Visualization

With all of the final changes to the dataset done, some conclusions were made based on visualizations.

First of all, the team looked at the customers' personal information. The number of male and female customers is almost perfectly balanced. The number of customers without any degree of education is significantly higher than the number of the ones that do, being the distribution between Bachelors, Masters and Phd degrees fairly equal. In what concerns the number of kids and teens at home, it is possible to assume that the majority of our clients does not have a lot of children, the numbers revolving around 0 or 1 child.

Next, the team studied the customer's behaviour. Regarding the customer's antiquity, it's visible that it ranges from 3 years to 34 years, having the most common values in the 10 to 16 years interval. These numbers are correspondent to 19,5% of the customers present in the whole dataset, which means that 80,5% of the customers do not possess a loyalty card. The typical hour in which the clients visit the store also has some interesting patterns, being able to identify some peaks in the intervals between 9am and 11am, and between 6pm and 10pm. The distributions of the number of distinct stores visited and the number of distinct products bought also give some interesting insights. The first one's normal values range between visiting 2 or 3 different stores, having some peaks around extreme values, such as 8 and 20 stores visited. The second one has a higher concentration towards lower values (below 1000) but also has some occurrences around 2000 and 4000 distinct products bought. In regards the percentage of products bought in promotion, the values mainly range from 5% to 20%; nevertheless, there is a group of people that has a 50% ratio of products bought. The number of complaints has its majority of records in the values 0 and 1, having an exponential decrease until reaching 9 complaints.

Afterwards, the histograms were plotted for the total of spends in each category in absolute values (Figure 2) and in percentages, alongside a tree map (Figure 3) for the sum of totals spent in each category.

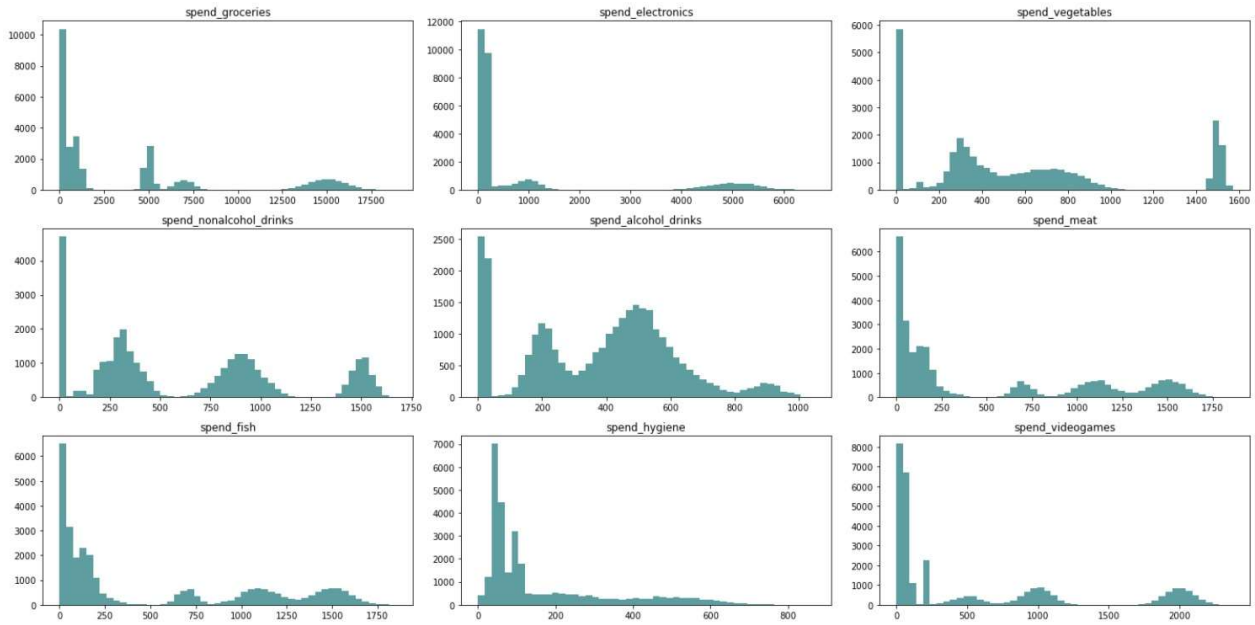

Figure 2: Spends distributions

From the first mentioned graph, presented above, it was possible to notice the presence of extreme values in many of the variables. Other than that, the second plot, presented below, made it easy to understand that there are some of the spending per category variables which have a much wider range than the rest of them. One good example of a category in which this happens is groceries, which has more than 4 times the amount of the second largest category.

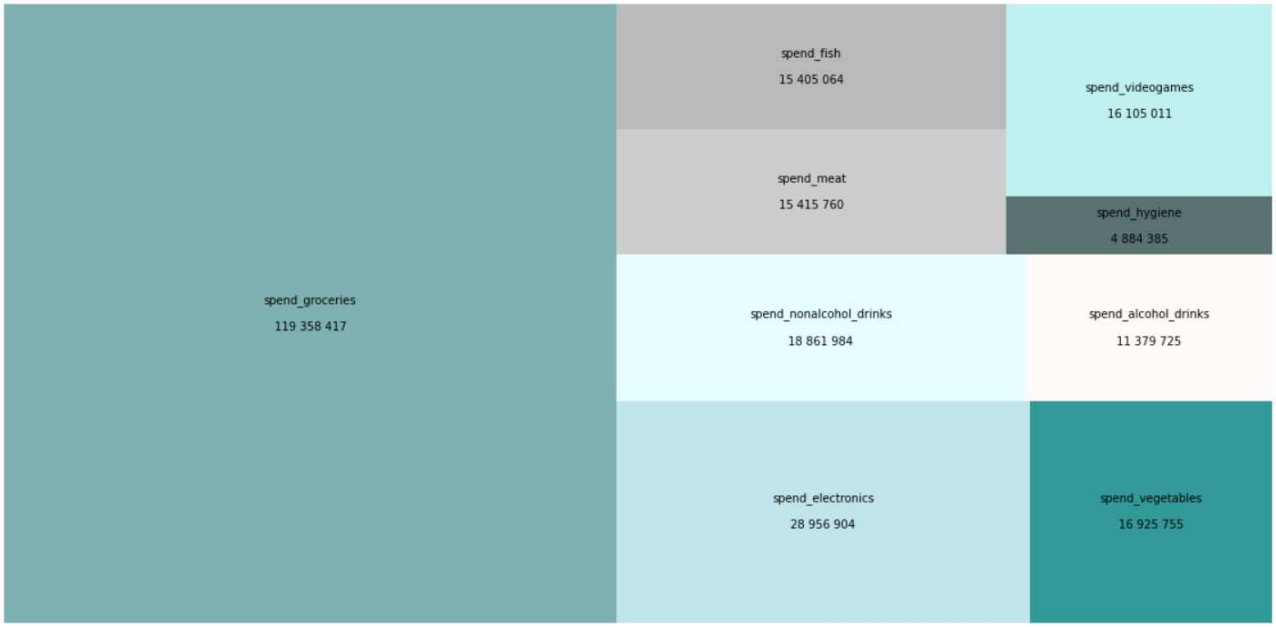

Figure 3: Spends tree map

All in all, it is important to state that the extreme values found in many of our study variables will not be discarded as they will probably be an important factor in deciding the clusters to which those observations will belong.

Then the distributions of each variable were analysed for the Supermarkets group and some insights were easily extracted. First of all, and as expected, all the observations registered the number of kids and teens at home as zero. The number of different stores visited was also constant for every observation, being always one. There were many strange values in variables such as spend_videogames and typical_hour, having some of the statistical measures equal to -inf. This anomaly caused the percentage variables created to go to zero, as it induced a division by infinite. Another noticeable occurrence was that this group did almost all of its spending in fish. Having all of this said, and reaching the conclusion that the supermarket's values are very different from the people's values, a decision was made to consider the supermarkets as one of our final clusters, not being included in the clustering algorithms used.

Correlation Analysis

A next step consisted in analysing the correlation between variables. The team started by looking at the correlation matrix between each variable and found several highly correlated variables.

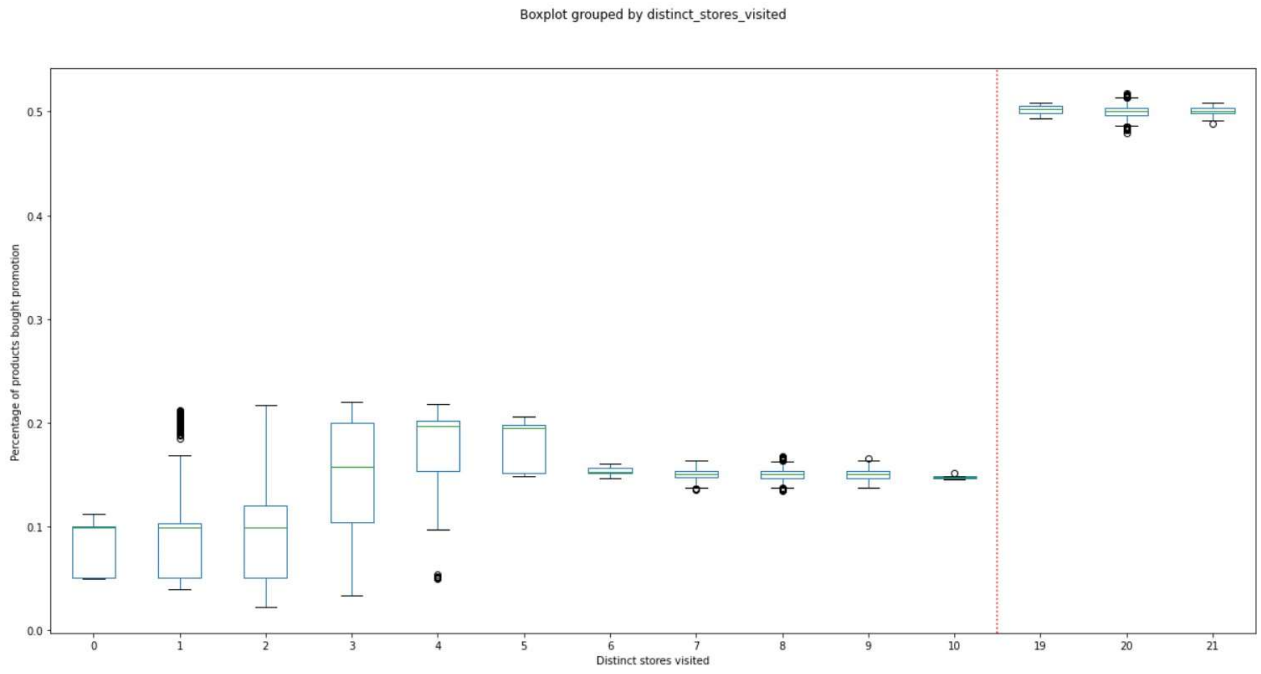

Figure 4: Percentage of products bought promotion per distinct stores visited

On pair of variables that were shown as having a strong correlation (0.92) was *distinct_stores_visited* and *percentage_of_products_bought_promotion* which was then plotted using the graph above. From the plot it was possible to better understand this correlation. It is clear that, when the number of stores visited is very high (more than 10 stores), the percentage of products bought on promotion was also very high, approaching values near the 50%, while for the other values for number of stores visited the

percentages were overall smaller. It was also interesting to see that for a lower number of stores visited, the variance in the percentages was also higher.

Furthermore, scatterplots between the remaining highly correlated variables were plotted. As is seen from Figure 5, presented below, which plots the values from *spend_eletronics* against *spend_videogames*, *spend_nonalcohol_drinks* and *percentage_spend_videogames*. Beyond the high linar correlation, it is interesting how the observations shape to the cluster picture. This also happens in other cases such as *spend_meat & spend_fish*, which are the two most correlated variables, almost with a perfect correlation.

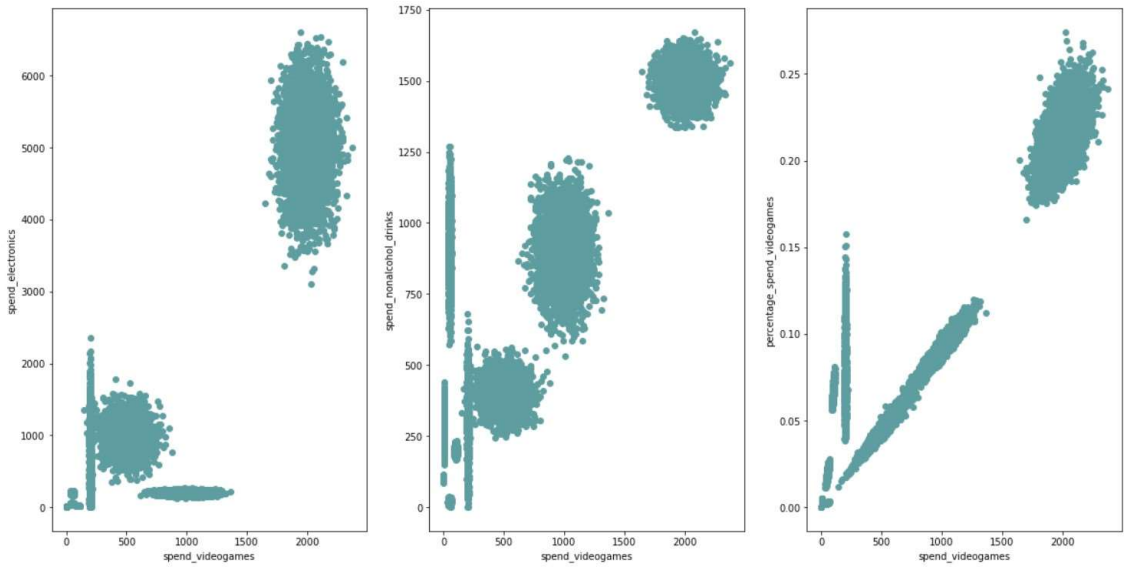

Figure 5: 'spend_videogames' correlations with 'spend_eletronics', 'spend_nonalcohol_drinks' and 'percentage_spend_videogames'

Lastly, a density plot (Figure 6) was done for the following variables: *spend_meat, spend_fish, spend_groceries* and *total_distinct_products*. From this plot, we can see that the distribution between these variables follow overall a very similar pattern.

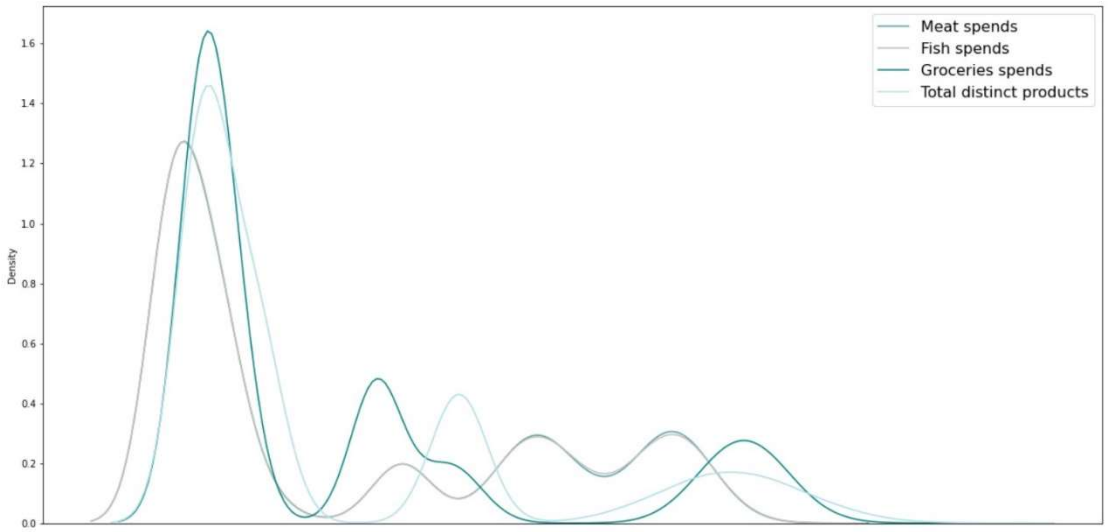

Figure 6: Scaled density comparision between the variables 'spend_meat', 'spend_fish', 'spend_groceries' and 'number_distinct_products'

Inconsistencies Correction

In what regards checking for anomalies in the dataset, the first thing done was to check if there were any duplicated values, which there were not.

After making sure that the dataset didn't have redundancies, the team found out that there were observations in which customers were clients even before being born. Having this in mind, the observations in which customers became clients before being 16 years old were considered anomalies. These invalid values were present in close to 10% of our dataset and were set as NaN, in order to be later imputed with an acceptable value by a KNN imputer. The optimal number of k was determined (5) and the KNN imputer was run. However, this solution failed as not even half of the previous anomalies were solved. In the end, the variable *customer_antiquity* was dropped from the people dataset.

As a last step of the Exploratory Data Analysis, the datasets *people* and *supermarkets* were exported as a csv for later usage.

Customer Segmentation

The clustering methods used were K-means, Hierarchical Clustering with different linkage methods and Mean Shift. Each clustering method was run at least once for each of the following data inputs: no scaling, PCA output, standard scaling, minmax scaling and robust scaling.

Considering that good results were achieved using K-Means and Hierarchical Clustering and as SOM also works with linear patterns, the team opted to leave SOM out of the study. Nevertheless, some tests were performed, but the calibration of the hyperparameters was hard and the interpretability was lower than with the previously mentioned clustering methods. In this same line of thought, DBSCAN was also discarded as Mean Shift was already used, the hyperparameter calibration was time-consuming and overall, the clusters from the dataset describe a circular shape, not having any need for additional testing with complex methods.

Segmentation Approach

To do the customer segmentation, the approach taken consisted of an exhaustive stepwise analysis of several options. Different combinations of clustering methods were joined with different scaling methods as well as with a PCA output. The parameters used for each algorithm were not kept fixed throughout the whole analysis, but instead they were constantly being changed according to new insights acquired throughout the whole testing phase.

As a small note into what concerned the usage of PCA in this approach is that it never directly influenced the analysis of the metrics of the clusters. It was only used in the clustering process, being the clusters always analyzed based on the original variables' means. This was included in the analysis to study if having only 80% of the variance from our original variables would bring any good outcome to the clustering solutions.

The K-means algorithm was tested for different values of k, suggested by the inertia/dispersion and silhouette score plots. The number of clusters to be considered in the hierarchical clustering was defined by looking at the dendrograms, having several values been tested for each linkage method too. In addition to these two algorithms, Mean Shift was also used. It will be able to identify more complex and non-linear relationships.

As mentioned previously, several tests were conducted, and the results constantly compared with each other. During testing, it was clear that some of the clusters that were being formed were very consistent, appearing in several of the clustering solutions. From these occurrences, some of the final clusters started to be noticeable.

In an initial phase, the clustering method chosen as the best one was a hierarchical clustering calculated with Ward distances and 8 clusters. This solution was the same for the principal components output, the standard scaled data, and the robust scaled data. The clusters formed were *alcoholics*, *gamers*, *promo hunters*, *low standards*, *parents*, *spenders*, *veggies* and *young adults*. All of these clusters are represented in Table 2 and will be later explained in detail.

	KMeans		Single		Complete		Ward		Average	
No scaling	5	G, P, S, Y	6	G, P, S, V, Y	5	G, S*	6	G, P, S, V, Y	4	G, S
PC	8	A, G, V	7	A*, G, S, V	7	A, G, P, S, V	8	!	7	A, G, S, V
Standard	7	G, P, S, V	7	A, G, H, V	7	A, G, P*, S	8	!	8	A, G, H, P, S, V
MinMax	8	G, H, S*, V	3	G, H	9	A, G, H, P, S, V, Y	6	G, H, S, V	7	G, H, S , V
Robust	8	G, H, S, V	4	G, H, V	5	G	8	!	4	G

Table 1 – Clustering solutions comparison

After assigning this solution as the base one, the results of the remaining solutions were compared to the results of this chosen one. These results were documented in Table 1. For each combination of data input and clustering algorithms, only the one considered to have the best clusters was chosen to be inserted in this table.

Before the table analysis proceeds it is important to make a small aside regarding the Mean Shift solutions, they were not considered for the table because they have a lot of defects. First, it was not possible to run the algorithm with unscaled data. Second, the standard scaler did not generate any cluster corresponding to those already mentioned. In general, the outputs were unsatisfactory, with only the *gamers* and *veggies* clusters being universal and the *spenders* cluster being captured with the main components data.

In Table 1, for each of the clustering techniques, there are two columns. The one on the left represents the number of clusters the solution has. The one on the right represents the list of clusters it has in common with the base solution previously chosen (which is represented with an exclamation point ("!") in the table). The meaning behind these letters can be better understood by looking at the second and third columns of Table 2. It is important to note that the correspondence between clusters was only considered if they were a complete match. Variations of even one observation counted as non-matching clusters. The only exception to this rule was represented as an asterisk ("*"), meaning that the cluster from the best solution was represented in the clustering solution in case, however divided into two or more clusters.

Num of obs	Abreviation	Cluster Name	Count
1248	A	alcoholics	8
4610	G	gamers	22
4921	H	promo hunters	9
2570	L	low standards	0
4667	P	parents	8
4764	S	spenders	16
4722	V	veggies	15
2272	Y	young adults	4

Table 2 – Base solution's clusters frequency

In Table 2 the first column corresponds to the number of customers associated with each specific cluster, represented in the third column, having an abbreviation included in the second column, so that

it was easier to match in Table 1. The last column corresponds to the number of times the cluster appeared, as said previously, without any changes, in other clustering solutions. From this table, we can see that clusters such as *spenders* or *veggies* appear in the vast majority of the clustering solutions, *gamers* being the most consistent cluster of all, having appeared in every one of the solutions. The clusters which appeared fewer times were *young adults* and *low standards*, with a respective number of appearances of 4 and 0. Throughout the study, these two clusters tended to get mixed together, having overall changing representations. Having this in mind, these two clusters were merged into a new dataset and Hierarchical Clustering was performed with all the different linkage methods to get a better understanding of these two clusters. Hierarchical Clustering was the method used as it was the one that provided better results in the previous analysis, and because, by leveraging the dendrograms, a more precise analysis would be possible of the potential clusters to be considered. From the observation of the dendrograms generated, the team decided that it was important to keep them separated and that, most likely, there were 308 observations that could form a new cluster.

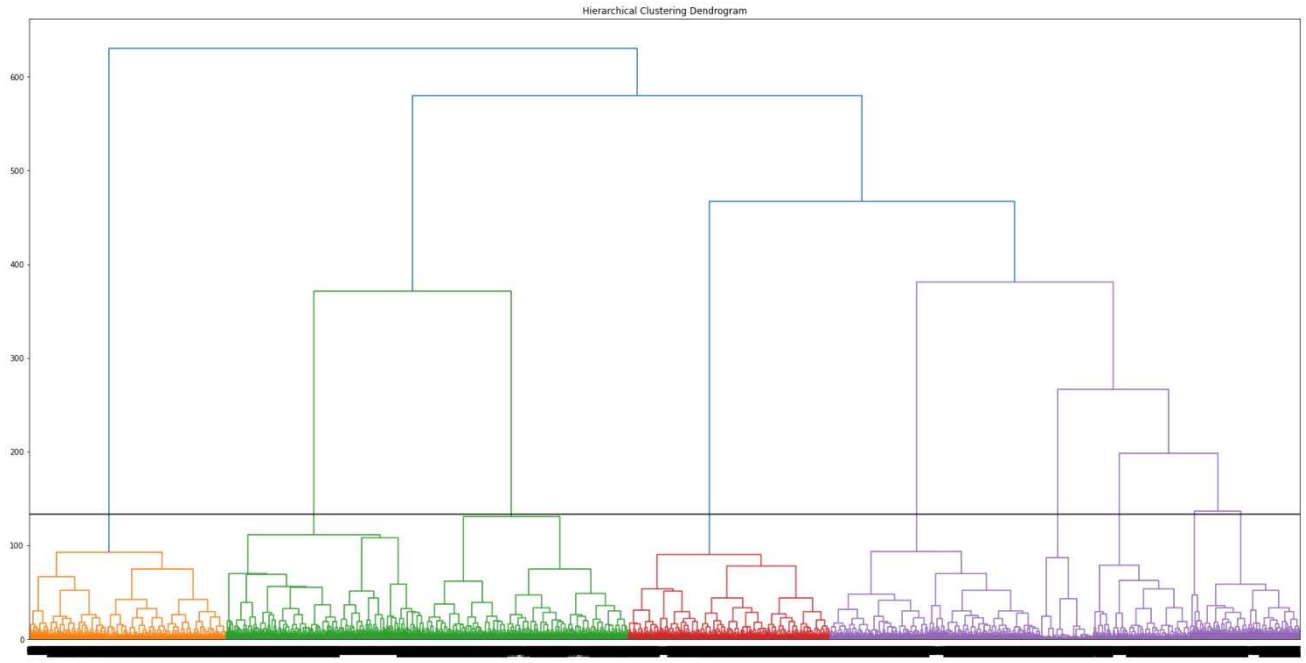

Figure 7: Hierarchical clustering Ward's method dendrogram with standard scaled data

Having the previously base solution and analysis of results in mind, the next step was to visualize the clusters in a multidimensional space. For this, the team opted to use UMAP instead of TSNE, based on the increase in running times for this last-mentioned technique. The UMAP, as seen in Figure 8, ended up showing 9 different clusters, the ones mentioned in the previously mentioned base solution, however, separating those 308 observations as a separate group. Having this said, the team decided to consider those observations in a different cluster called *olds*.

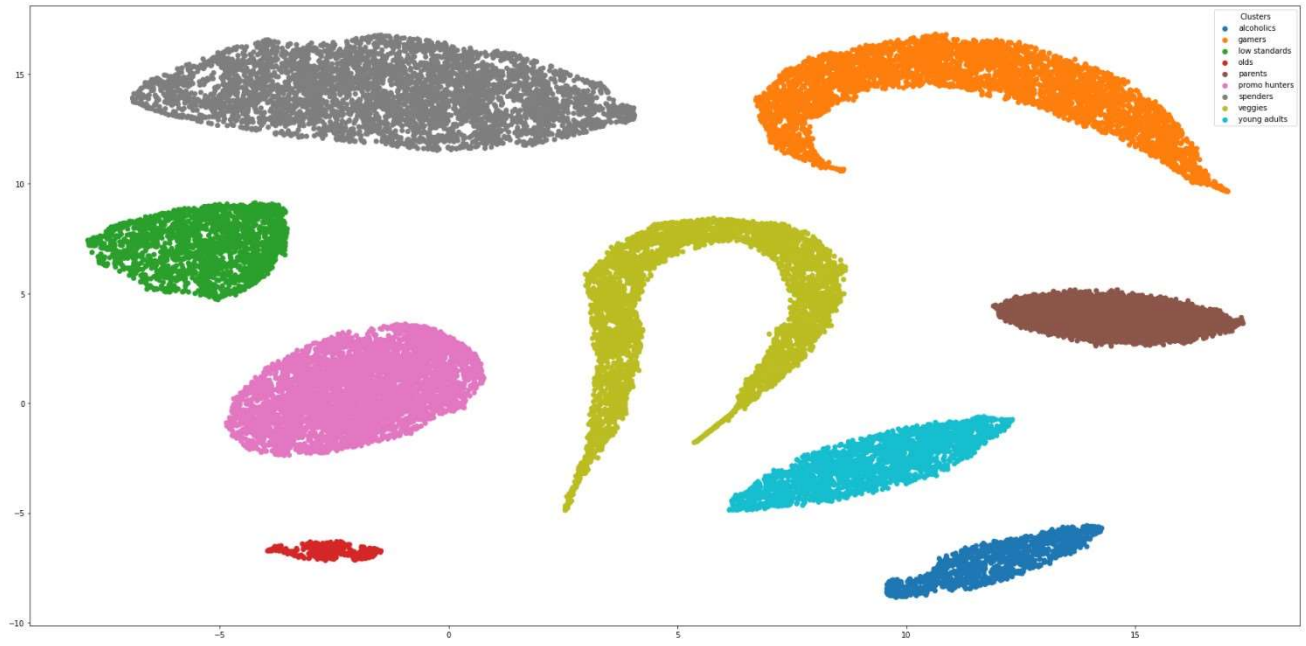

Figure 8: Cluster solution visualization with UMAP implementation

The names attributed to each cluster were chosen based on the mean values of the clusters' observations for each of the independent variables. For a better understanding of the characteristics of each group, Table 3, presented below, shows the relevance each variable has in each cluster's distinction.

nomenclature	high	slightly high	slightly low	low
veggies	(p\)vegetables	teens, kids, education, p\hygiene	hour, (p\)videogames, complaints, products, loyalty	(p\)meat, (p\)fish, promos, (p\)nonalcohol, (p\)alcohol, stores
parents	teens, kids, (\p)hygiene, p\meat, p\fish	meat, fish, (p\)videogames, nonalcohol, products, complaints, (p\)groceries, vegetables, loyalty, promos, education	stores	-
gamers	(p\)eletronics, (p\)videoganes, (p\)nonalcohol	hour, education	groceries, (p\)meat, (p\)fish, teens, promos	(p\)vegetables, stores, p\groceries, (p\)hygiene, products
young adults	education	(p\)groceries, eletronics, hygiene	age, kids, teens, (p\)nonalcohol, stores	complaints
spenders	(p\)groceries, products, loyalty, complaints, meat, fish	stores, vegetable, alcohol, nonalcohol, hygiene, p\meat, p\fish	videogames, (p\hygiene), (p\alcohol), (p\)nonalcohol	(p\)eletronics, (p\videogames)
alcoholics	(p\)alcohol, hour	promos	products, meat, fish, (p\)nonalcohol, (p\)eletronics	age, education, kids, teens, complaints, latitude, longitude, (p\)groceries, eletronics, vegetables, loyalty, stores,
low standards	-	(p\)electronics, p\alcohol, p\meat, p\vegetables	education, vegetables, complaints, alcohol, meat, fish, videogames	products, hygiene
promo hunters	promos, stores, p\nonalcohol, p\meat, p\fish	complaints, teens, p\vegetables, p\alcohol	education, alcohol, (p\)groceries, p\eletronics, vegetables, nonalcohol, meat, fish, loyalty	hour, (p\)videogames, eletronics, hygiene
olds	age, complaints, p\hygiene, p\alcohol	promos, p\vegetables	hour, groceries, alcohol, meat, fish, p\eletronics,	education, longitude, stores, products, eletronics, vegetables, nonalcohol, hygiene, (p\)videogames

Table 3 – Variables' relevance for each cluster

p\ represents the percentage values and
(\p) represents both values: absolute and in percentage

Note: Variables not present in the table, were not positively or negatively important for the cluster

From the analysis of the table above these are the main characteristics of each of our clusters:

Veggies: People who buy a lot of vegetables and do not buy any meat or fish. Probably customers with vegetarian or vegan diets.

Parents: People who have a higher number of kids and teens at home, who essentially have the typical buying habits of parents. They tend to have somewhat elevated expenditures in a lot of different product categories, especially food and hygiene products. They also tend to have loyalty cards and search buy products on promotion.

Gamers: People who spend a lot on electronics, videogames and non-alcoholic beverages. They don't tend to visit stores but when they do it's normally late at night. The type of customer who likes to play videogames and likes to have the latest technologies to improve gaming.

Young Adults: Group of people characterized for being younger and for having a higher degree of education.

Spenders: People who essentially spend a lot of money mainly in the categories with a higher impact (groceries, meat and fish). They also tend to have loyalty cards and have a somewhat high number of complaints when compared with customers from other clusters.

Alcoholics: People who tend to buy many alcoholic products. On average, they also visit stores later in the day. This cluster also has a very low level of education and contains younger people, with the average customers' age being around 23 years old.

Low Standards: This group does not have any strong characteristics that distinguish it from the rest of the clusters. The cluster joined customers that tend to have an average behaviour, not having a lot of noticeable characteristics. People from this cluster tend to have their characteristics aligned with the average of the whole group of clients.

Promo Hunters: This cluster was highly characterized by the number of items bought on promotion and the number of different stores visited. Customers belonging to this cluster tend to always be looking for a good bargain, visiting a vast range of stores in order to find the best offers. Another noticeable characteristic is that customers belonging to this cluster, on average, visit stores early in the morning, somewhere close to 9 am.

Olds: People from this cluster have a tendency to be older, with the average customer's age revolving around 71 years old. These clients also tend to have a higher number of complaints, when compared with the rest of the clusters. These people aren't prone to spend a lot of money; however, a good percentage of their expenditures are on alcoholic drinks.

In the end, the final number of clusters chosen was 10. The nine clusters described above, plus the supermarkets, separated at the beginning of the study. Figure 9, shown below, shows the number of observations per cluster.

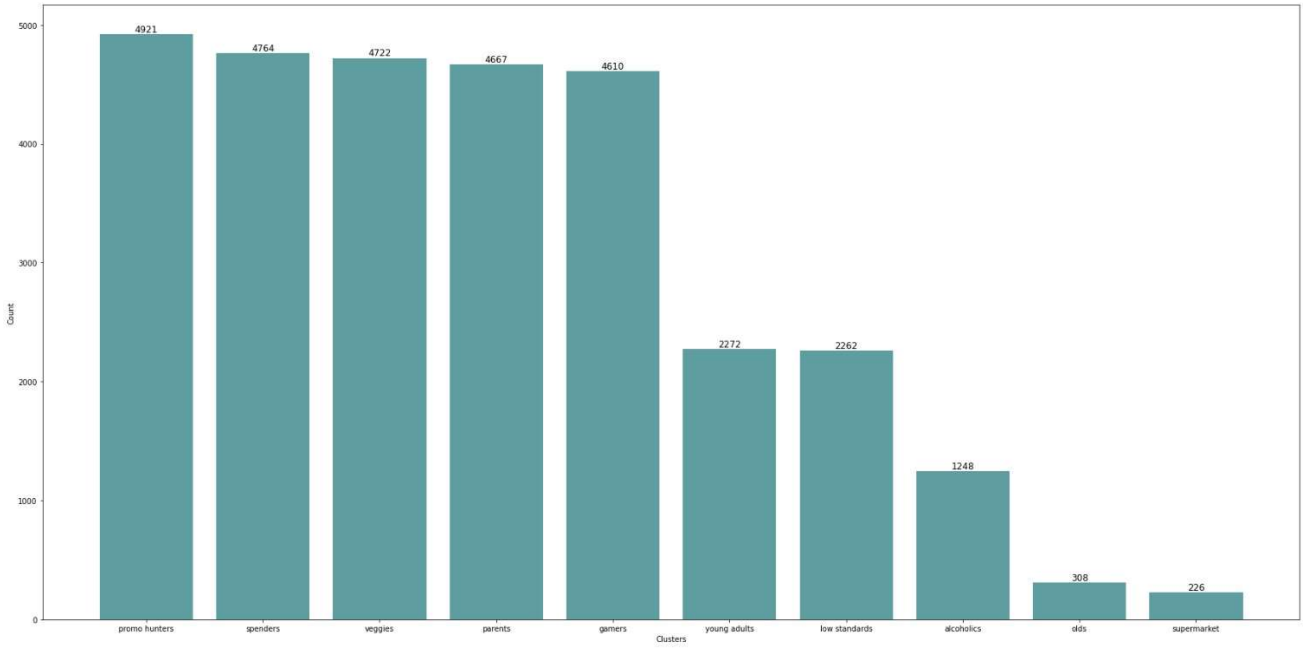

Figure 9: Number of observations in each cluster

Targeted Promotions

Following the previously explained stages – exploratory data analysis and customer segmentation – the next stage consisted of the creation of association rules, and consequently, the development of targeted campaigns for the different segments of clients.

In summary, association rules consist of the discovery of patterns that describe, in this particular case, certain subsets of products which are often bought together. It is also important to mention that these rules do not have any type of causality effect, since they only indicate that when a given item is bought there is an *x* probability that another specific item will also be bought. Additionally, the quality of these can be evaluated based on three main measurements: confidence, lift and support.

Please note that as cooking and other types of oil revealed themselves to be bestselling products, their appearance on most association rules was very common. Therefore, they were mostly disregarded and only taken into consideration to aid in other types of campaigns (i.g.: discounts on oils if *x* item was also bought). Additionally, as there was no information on the prices of the analysed products, the final prices for the suggested campaigns nor their sustainability for the company were topic of discussion on this report.

In the following few paragraphs, the association rules and promotions for each one of the segments will be specified.

Gamers

From the obtained association rules and by only considering the ones with the highest lift (above 1.279) and acceptable support values (around 0.01) it was possible to identify a pattern between the customers from this Gamers segment.

Their most common purchases revolved around Pokémon videogames (pokémon violet, pokémon shield, pokémon scarlet and pokémon sword), bluetooth headphones and ratchet & clank game.

With this information the following three targeted campaigns were suggested:

- **Pokémon gaming bundle**: promotion of get 4, pay 3 of the four most sold Pokémon games (Pokémon shield, sword, violet, scarlet) at the stores under analysis.
- **Pokémon Violet Exclusive:** special offer - get 20% off on the Bluetooth headphones when buying Pokémon Violet.
- **Cross-gaming experience bundle:** package deal of the Ratchet & Clank game with Pokémon sword game – i.g: considering the base value of a game, giving 30% off on the cheapest one (most indicated for customers which usually buy Pokémon, and consequently, Ratchet & Clank games).

Additionally, and specifically for these clients, these promotions could be shared via email, or there could be an extra and exclusive discount if the purchases were done on the company's e-commerce shop.

Spenders

Trying to follow a similar pattern of thought, the minimum lift considered for this segment was 1.226 and support was around 1% of the baskets. In this specific cluster, it is rather common for clients to buy olive or cooking oil, milk, cakes, milk, soup, napkins, sandwiches, and pet food.

The promotions developed for this segment were as follows:

- Milk and sandwiches are often bought together. Hence, a **snacks pack** with milk and a sandwich is suggested. This new product can be sold at a slightly higher price, than the two individual items and promoted as it being more efficient and convenient.
- Knowing that most of the customers who buy pet food, also buy napkins and soup and that the opposite is also verified. Instead of offering discounts on these products which technically have no connection to each other, what is suggested, in hopes of increasing the company's sales, is that pet food is changed, in the supermarket layout, to after the soup and napkins sections.

Alcoholics

This segment of clients was the one which showed the best results in terms of lift; having values from 1.49 to 1.86. Similarly, to what was previously done, from the information that could be taken away from the association rules the following campaign suggestions were produced:

- **Buy two** white wine bottles, **get one** for free (this campaign was indicated, since most of these association rules had white wine in them. Allowing the opportunity to capitalize on this product).
- **Wine tasting pack:** an exclusive pack specially thought for wine lovers; includes a bottle from some of the most popular alcoholic beverages – white wine, bramble, beer, black beer and champagne.

Promotion Hunters

Typically, the clients associated with this cluster do not always buy the same products at the supermarket, since they have revealed a pattern of most commonly buying promotion items. Hence, in the association rules (considering lifts from 1.22 to 1.29 and support between 0.01 and 0.012) it was quite hard to detect some patterns of products that are obviously connected to each other. So, there are two possible approaches to solve this mishap:

- The first solution is based on a stock analysis (to be performed with stock data) reach a conclusion of which products can be in larger quantities and at reduced prices - **flash sale** – as a way to keep the attention of this type of customers.
- The second one and indicated in the case that the first option is not viable for the company at the moment, is offering a **5€ coupon** when a client makes a 50€ purchase that he can use in a future purchase (in the span of 3 months) in any transaction over 25€.

Low Standards

Akin to the previous situation, these were customers which did not have main characteristics of any of the other segments. Hence, their shopping habits are extremely diverse; while some customers tend to spend more on alcoholic beverages, other buy electronic devices or vacuum cleaners.

Take into account that the studied association rules for this client segment had lift values from 1.44 to 1.7 and support between 0.01 and 0.013.

A campaign that can be beneficial in this situation is the implementation of **weekly coupons** for a limited period of time (for example, 3 months) where each week the discount to be applied will be to a different category of products. Some practical examples can be:

- Week 1: 10% discount on any selection of wines;
- Week 2: 20% discount on laptops
- Week 3: 15€ discount on vacuum cleaners

After the defined period of time, the data from these clients can be analysed again to understand the amount of adherence to each campaign and to adapt them to the newly shown patterns of purchase.

Parents

Considering lift values from 1.00 to 1.34 and support around 0.001, the association rules for this cluster showed items such as: soups, baby foods, candy bars, cake and ketchup.

As soup and baby food was the one which appeared more times and it does make sense given that new parents do not have a lot of free time to prepare their meals; an interesting campaign to promote is as follows: for each pack of four baby foods, a 50% discount on soup is given at checkout. Furthermore, an entire basket promotion can be carried out for the parents who usually buy more items on this supermarket chain (5€ available on the customer loyalty card for their next purchase with no minimum limit).

Veggies

This group of clients, and as explained previously, consists of people who mostly buy vegetables and fruits. Thus, considering the association rules of this cluster, with lifts between 1.23 and 1.34 and support of around 0.01, the most common items were: carrots, tomatoes, corn, melons, asparagus, and frozen vegetables.

Therefore, some exclusive campaign suggestions are:

- **Mixed vegetables basket:** create a limited-edition basket with the most bought vegetable items – carrots, tomatoes, corn, melons, asparagus - at a slightly reduced price to keep capturing the interest of these clients.
- **TikTok trend avocado kit:** since avocado is a common item among the association rules, an interesting idea to increase its sales is to create a small section, among the vegetables area, with avocados and other products that are needed for trendy fast recipes, typically popular on social media apps such as TikTok.

Young adults

The young adults cluster is not extremely consistent in terms of bought product; considering lift values between 1.26 and 1.35 and support of around 0.01. However, some patterns can be seen, for instance most purchases have some type of oil.

As so, and to enhance the sales for this segment of customers, with each 15€ of purchases the customer can randomly acquire either a bottle of oil of their preference (cooking, olive, frying oil) or a pack of gums.

The following two clusters were defined based on a smaller number of observations. Hence, the returned values for the lift of the association rules were very high (40 and 85). In order to try to avoid this problem, the minimum support value was then increased to 0.04 and the analysis done below took into consideration these new values.

Olds

Olds is a cluster which clearly indicates transactional patterns from older generations, probably meaning grandparents. This, because most of the association rules, if ordered by lift in descending order, had products such as baby food or non-fat milk.

In such cases, a possible campaign, to be sent via post-mail letters (most of these clients do not check their emails or most likely do not even have one) consists of a set of monthly coupons (10% discount, 5€ on a 20€ purchase) on common products from the association rules – oils, soups, candy bars, muffins, baby food, non-fat milk.

Supermarkets

As for the supermarkets, and as all the above promotions can and shall be applied not only on the customer level but also at the stores, a noticeable pattern from the association rules is that fresh bread, cakes, gums and candy bars are a popular item among client's purchases. Thus, they can be strategically placed around the store to enhance their sales – for example, next to checkout points.

Conclusion

The goal of this project was to successfully create groups of customers based on similar characteristics. To achieve this goal, an initial analysis was performed to better understand the dataset in question.

After an exhaustive testing phase, the clustering method found to provide better solutions was Hierarchical Clustering, when given standard scaled data. Given the results observed in the dendrograms created and a posterior analysis of the UMAP algorithm output, the final number of clusters to be formed was equal to 10.

From the analysis of the means of each independent variable of each cluster obtained with the clustering technique mentioned above, the customer segments were attributed the following names: veggies, parents, gamers, young adults, spenders, alcoholics, low standards, promo hunters, olds and supermarkets.

Having the clusters well defined, an analysis of the customers' baskets provided was done. This analysis was performed separately for each customer segment, in order to extract buying patterns within the clusters previously made and be able to strategically plan marketing campaigns to bring more customers to the business and sustain the ones the business already has.

All in all, the objective was completed with success and good results were obtained. However, to test the efficiency of the marketing strategies proposed, a future analysis will have to be conducted after a suitable period of time.

References

[1] Data mining – Lift in an association rule. (n.d.). Data Mining – Lift in an Association Rule. https://www.ibm.com/docs/en/db2/11.1?topic=SSEPGG_11.1.0/com.ibm.im.model.doc/c_lift_in_an_association_rule.htm

[2] Garg, A. (2019, February 7). Complete guide to Association Rules (1/2). Medium. https://towardsdatascience.com/association-rules-2-aa9a77241654

[3] Garg, A. (2019, February 7). Complete guide to Association Rules (2/2). Medium. https://towardsdatascience.com/complete-guide-to-association-rules-2-2-c92072b56c84