NOVA IMS Information Management School

PROJECT REPORT

MACHINE LEARNING

Customer Segmentation: A Key to Unlocking Business Growth and Success

Project By

ANTÓNIO OLIVEIRA 20211595

DAVID MARTINS 20201628

INÊS GRAÇA 20211598

Table of Contents

Executive Summary

The purpose of this report is to analyse customer data, by finding similarities between customers and their purchases, with the aim of providing better services and an increased number of sales and revenue. To start exploring our data, our group started by understanding each variable of the *Customer Information Dataset*, as this was the dataset that allowed us to implement the clustering algorithms previously learnt. Following this, we imported the data, having then performed EDA and implemented clustering algorithms that helped us to segment the customers.

After careful consideration, we decided that the segmentation that better described our clients was the combination of the UMAP and a DB Scan, with 10 clusters, which were the following: Students, Gamers, Promotion Lovers, Big Families, Vegetarians, Loyal Clients, Millennials, Average Clients, Elderly and Supermarkets.

The behaviour and demographics of each of these clusters will be deeply explored later in the report. The next step was to perform Association Rules, which allowed us to see which products were commonly purchased together, giving us an idea on promotions that could be lucrative for the company.

Exploratory Data Analysis

To start exploring the Customer Information Dataset, we began by applying the Profile Report to our data. This is an exploratory method that gives us an overview of all the variables we have, as well as plotting their respective distribution. This was especially useful for us to carefully observe the situations it alerted us to. After this we applied .describe() and .info() to our dataset, to further explore it and look for inconsistencies or incoherent values.

Our first main conclusions were that we had 30000 observations with 24 variables being 21 numeric and 3 categorical. There were 3 variables with missing values (*loyalty_card_number*, *lifetime_spend_videogames* and *typical_hour*), a problem we will later address.

Data Cleaning

In order to ensure our data was ready for the next steps, we must first look for incoherencies, as well as eliminate missing values and outliers. While looking for incoherencies, we found some unexpected values in the *typical_hour*, as we found values like 2.119509, when we were only expecting to find integers between 0 and 24. For now, our decision was not to do anything, having noted that the majority of these observations contained Supermarket in their name.

The next variable we decided to observe was *lifetime_spend_fish*, as we had previously noted a very big maximum value, when comparing with a similar value like *lifetime_spend_meat*. From this we again concluded that these observations also contained Supermarket in their names. Both these conclusions were very useful, as it looked like there was a distinct group of customers to look out for in the future.

Continuing looking for incoherent values, we decided to look for customers with a high percentage of products bought in promotion, which allowed us to conclude that there are some costumers that purchased around 50% of their products while they had discounts. Finally, as previously showed by the *describe* method, we looked for customers whose first transaction was before 2005. A possible conclusion to these results is considering this group as the loyal customers.

Then we approached the missing values present in the previously identified variables. Having said that, our first decision was to transform the *loyalty_card_number* variable into a new one, *loyalty_card*, which represented the presence (1) or absence (0) of a loyalty card being owned by the customer. The next step was to identify the missing values in the *typical_hour* variable. In this case, there were only two clients whose values were missing. In this case, we decided to replace them by 0, as they were both Supermarkets, which we previously checked had some inconsistent values in this variable. Finally, we took care of the missing values present in *lifetime_spend_videogames* by filling the missing values with zeros, as we assumed these were missing when the customers had never bought videogames.

In the end we just confirmed that all variables had the same amount of non-null observations. The only one that was different was the *loyalty_card_number*, as it had not been dropped yet, after being replaced by a new variable.

The final step in the Data Cleaning section is to look for outliers. To do so, we created a function that plotted boxplots and histograms for all the variables. Both these methods showed some outliers in a few variables, but as we know these are real customers, our decision was not to do anything about them.

Feature Engineering

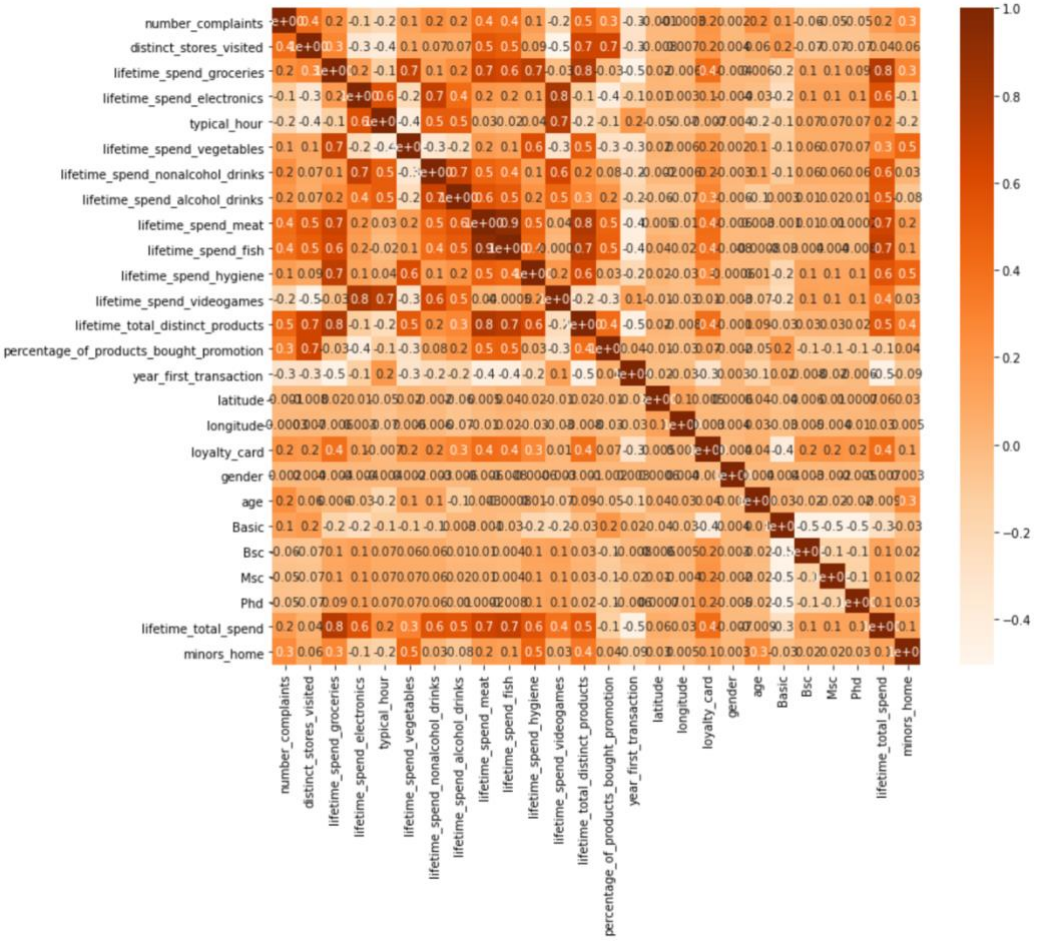
In this section, we proceeded with the creation of new variables and with the transformation of some of the already existing ones, in an attempt of giving them more value for future usage. As so, we transformed the *customer_gender* variable into the *gender* one, which was a binary representation of the previous (male – 1, female – 0). Then we decided to also transform *customer_birthdate* into *age*, to make its interpretation easier. Our final transformation was transforming part of *customer_name* into *education*, that told us the level of education achieved by each individual: PhD, MSc, BSc, Basic (in the case none of the previous cases were present) and the special case of Supermarkets. This was then replaced by a dummy variable for each type of education, where the supermarkets are identified by a zero in all the previously referred types of education.

For variable creation, we decided it was helpful to sum the lifetime expenses in all products into a new variable, *lifetime_total_spend*. We also created *minors_home*, which was a combination of *kids_home* and *teens_home*, that gave us similar information with half the variables.

After finishing this process, we decided to drop columns that were now unnecessary and then proceed to the Correlation Analysis.

To understand if some of our variables had a high correlation with each other, we decided to plot a heatmap with the Spearman method. Following are the displayed results.

Figure 1 - Heatmap



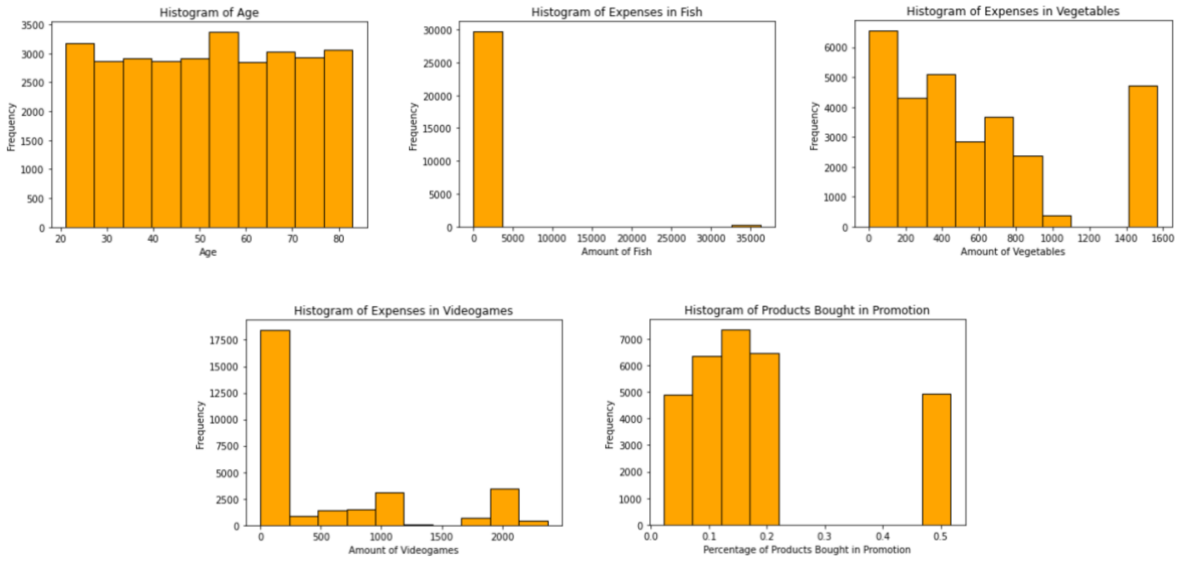
As one can see, there are a few variables with high correlations. For example, *lifetime_spend_fish* and *lifetime_spend_meat*, as well as *lifetime_spend_videogames* and *lifetime_spend_electronics*. However, as we know the true meaning of these variables, we know that the effect of *lifetime_spend_fish* cannot be replaced by *lifetime_spend_meat*, nor vice versa. With this in mind, we decided not to remove any variables from our dataset.

Data Visualization

With our data clean, without missing values, and the inconsistencies and outliers completely understood, we decided to perform some individual plots in certain variables, that we thought would give us some insights for the next steps, along with the previously acquired knowledge. Our next decision was what type of visualisations would give us the results we were looking for. We decided to plot histograms, pie charts and a map.

Firstly, we plotted histograms for the age, lifetime_spend_fish, lifetime_spend_vegetables, lifetime_spend_videogames and percentage_of_products_bought_promotion. The results were the following.

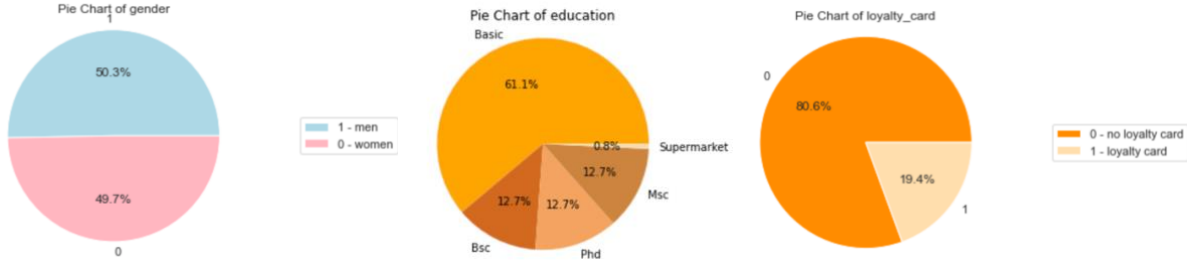
Figure 2 - Histograms



As it is shown in these boxplots, the distribution of the above-mentioned variables is very important for us to know what to expect in the next section of this project. We decided to highlight specifically these variables as we believed they were the most important ones until now.

Our next visualisation are the pie charts, where we just wanted to see if the amount of males and females was more or less the same. Additionally, we also looked at the proportion of the observations in each education type and at the number of customers that had a loyalty card. Here are the results we got.

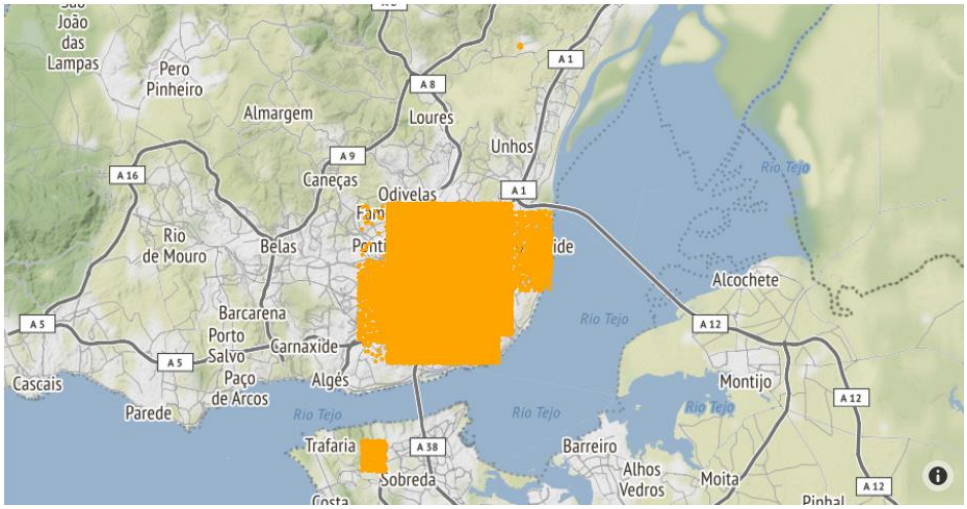
Figure 3 – Pie Charts



With this we concluded that there were practically as many men as women, being the Basic Education the most common and more than ¾ of our costumers did not own a loyalty card.

To end this section, we just plotted a map to be able to see where our customers were located. This allowed us to see that our customers (represented in orange) were mainly located in the Lisbon Metropolitan Area.

Figure 4 - Map



With the end of this section, we decided to export the final results. We created a second notebook. We decided to do so for better organizational purposes, as well as to distinguish the EDA section from the Customer Segmentation and Targeted Promotion ones.

Customer Segmentation

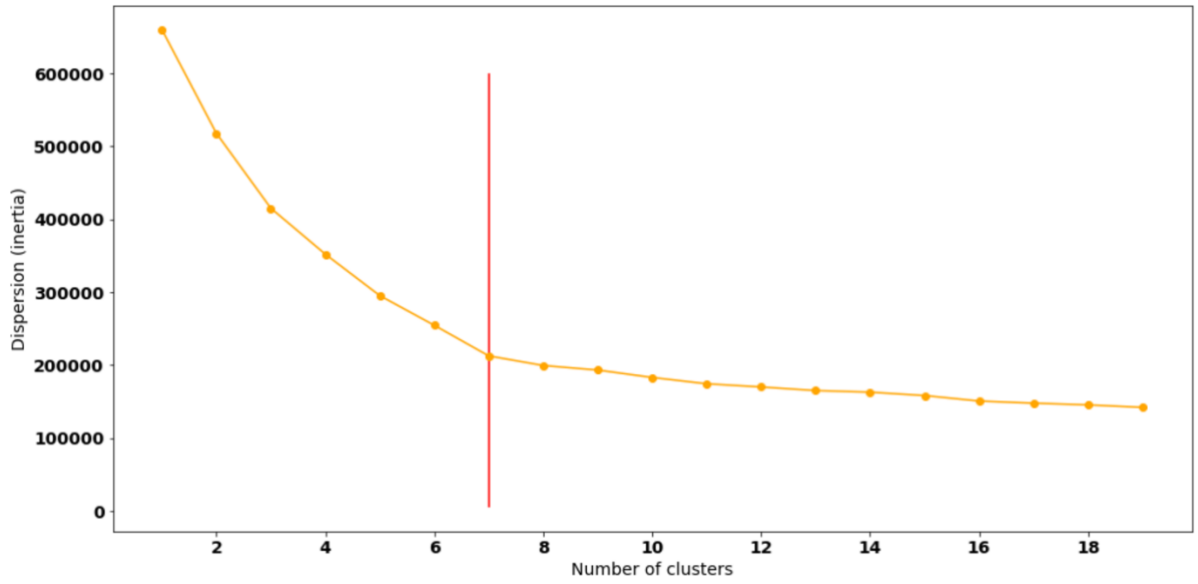
In this part of the project, we decided to start by creating a variable with only numeric columns, where we did not include the dummies for education, as they can create difficulties for algorithms that base their approach on computing distances between data points. Following this, we standardized our data and performed all clustering algorithms learnt in class and understood their results. Having said that, when doing so, we concluded that some of the methods did not produce good and/or interpretable results, having then decided to keep the

best performing clustering algorithms in the notebook. A description of our findings in each one of them follows.

K-Means

The K-means is an algorithm capable of creating clusters based on the similarities of the given observations. In this case of K-Means, we started by finding the optimal number of clusters k, which was not that straightforward, as the Elbow Curve showed that using 7 clusters could be a good choice but using 8 or 9 was not completely out of the question. After experimenting with these 3 different values, we started to believe the best choice is using 7 clusters.

Figure 5 - Elbow Method Visualisation



After this, we plotted a bar chart that told us how many clients were in each cluster. All clusters had approximately more or less the same number of observations, except one. As we had previously identified that there might be a small cluster of observations with 'Supermarket' on their names, this did not worry us. Our next step was to use the average value of each variable for all clusters to try and identify which groups of customers were being selected by our K-Means.

We were able to find 7 distinctive groups: the *vegetarians*, the ones that rarely buy meat and fish, and spend a lot of money on groceries; the *promotion lovers,* who buy half of their products in promotion; the *loyal customers*, who were identified by the year of their first transaction, that was quite earlier (2000) when comparing with other clusters (2009, 2010, 2011); the *gamers,* that on average spent a lot on electronics and videogames, as well as having a very low average spent in groceries and a very high expenditure in non-alcoholic drinks (maybe energy drinks, typical of a gamer); *big families*, which are a group of customers that basically has high values on everything (on average, they also have about 5 minors at home); finally, the *average customer*, that does not have particularly big values of anything, as so being

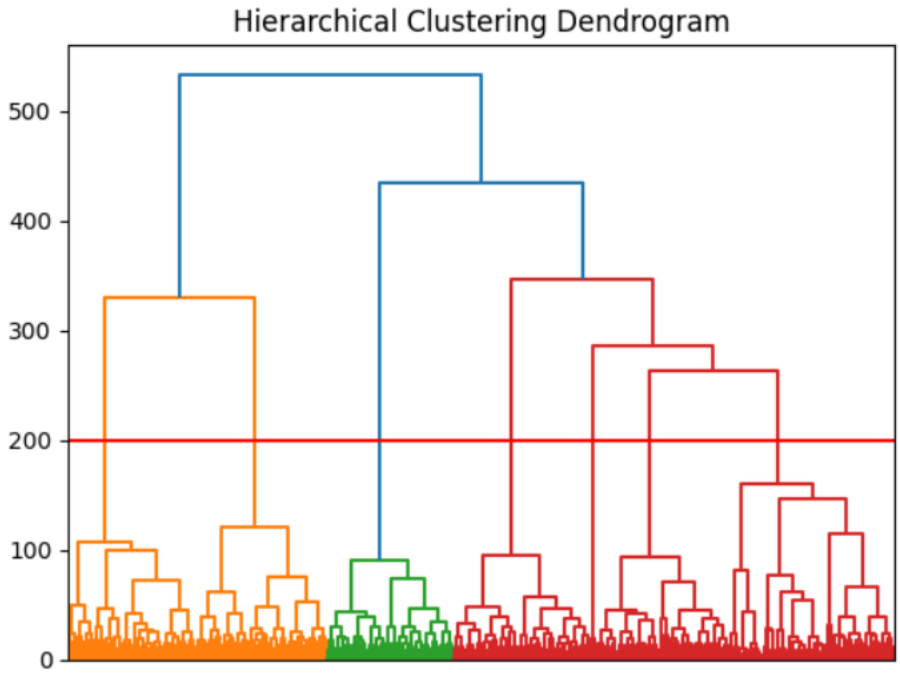
referred to as the average customer: the one that buys a little of everything; as previously referred, we also found a group of Supermarkets, which as expected were the ones with significant expenses in fish.

Hierarchical Clustering

The Hierarchical Clustering method is based on organizing the data hierarchically, with an Agglomerative or Divisive approach. In our case we were interested in the first approach, so we had to decide which linkage criterion should be used in the algorithm. We opted for the Ward method which minimises the variance of the cluster being formed, as all other methods produced extremely poor results.

After this we had to analyse the following dendrogram to decide on the ideal number of clusters. We came to the conclusion that 7 was the optimal number of clusters to use, but 9 also seemed promising.

Figure 6 - Dendrogram



After analyzing the clusters, we saw that they were the same ones that were found by K-Means algorithm with 7 clusters, so we decided to compare the constitution of those clusters with the following table.

Figure 7 - Clusters of K-means vs Hierarchical

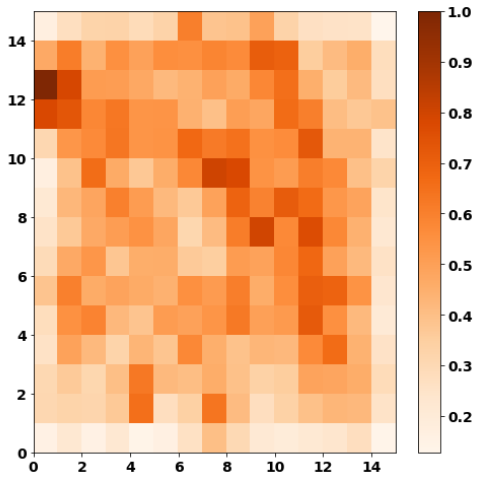
	Ward 0 Cluster	Ward 1 Cluster	Ward 2 Cluster	Ward 3 Cluster	Ward 4 Cluster	Ward 5 Cluster	Ward 6 Cluster
K-means 0 Cluster	5749	0	117	0	0	0	0
K-means 1 Cluster	0	4764	0	0	0	0	0
K-means 2 Cluster	22	0	4722	0	0	0	0
K-means 3 Cluster	0	0	0	0	4667	0	0
K-means 4 Cluster	0	0	0	0	0	0	4610
K-means 5 Cluster	202	0	0	4921	0	0	0
K-means 6 Cluster	0	0	0	0	0	226	0

As we can see these two methods agree on which cluster most observations should belong, these made us feel confident in our clustering solution.

Self-Organizing Map (SOM)

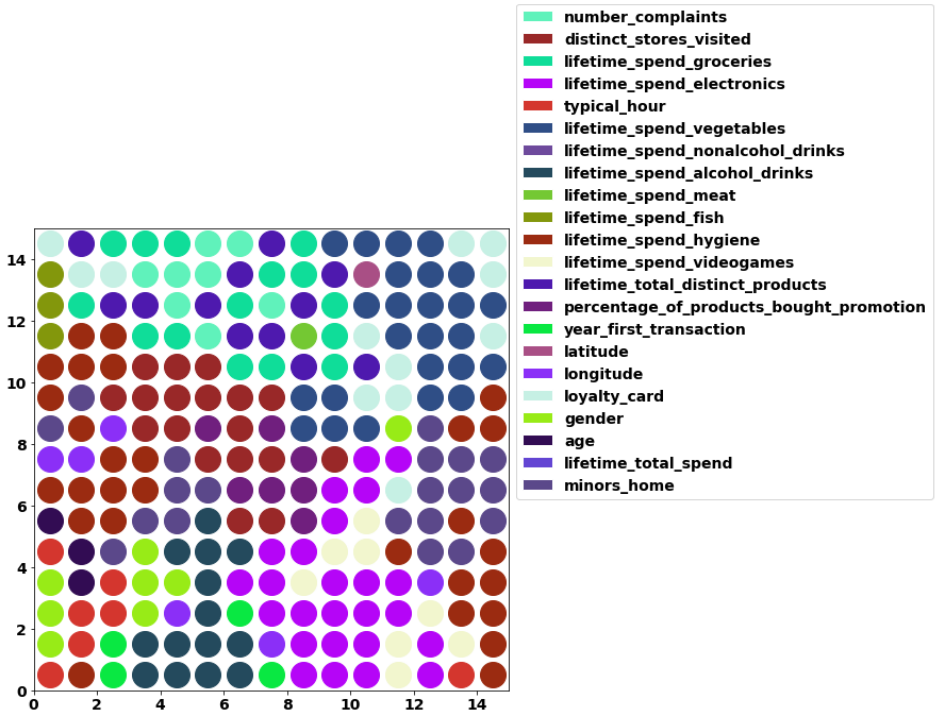
The Self-Organizing MAP (SOM) is a type of unsupervised neural network that is used for dimensionality reduction and visualization of high-dimensional data. We decided to use this algorithm to further explore our data and see if there were any other conclusions, we were able to draw. We started by training our SOM using MiniSom and then visualized its distance map as well as its colour mapping.

Figure 8 - Distance Map



When looking at the distance map we instantly spot a node that is darker than all the others (0,12). We agreed this was likely to be a node that was grouping supermarkets, as we've realized they are a very distinct group. We can also observe darker nodes around that one, as well as a few ones in the middle. However, this distribution isn't very clear and it's not easy to conclude much else from it.

Figure 9 – Colour Mapping



When analyzing the color mapping, we were able to notice the creation of some clusters by colors. There are a few easy to identify. For example, the ones that are connected by *lifetime_spend_vegeatables*, or by *lifetime_spend_eletronics*. Despite these interpretations it's still hard to determine how many clusters we should use to group our data.

Lastly, we assigned a winner node to each customer. We did this to be able to group our customers by their winner node and then sort these groups by size. The goal of the sort was to find which node had more data points associated to them. We then view the data from the four nodes with more data points. Since the conclusions weren't any different from what we already knew, we decided to not keep this last step in the notebook.

To summarize, the SOM algorithm didn't help us decide how many clusters to use. However, it was useful to reaffirm the results we had already gathered from the other algorithms.

DB Scan and Mean Shift

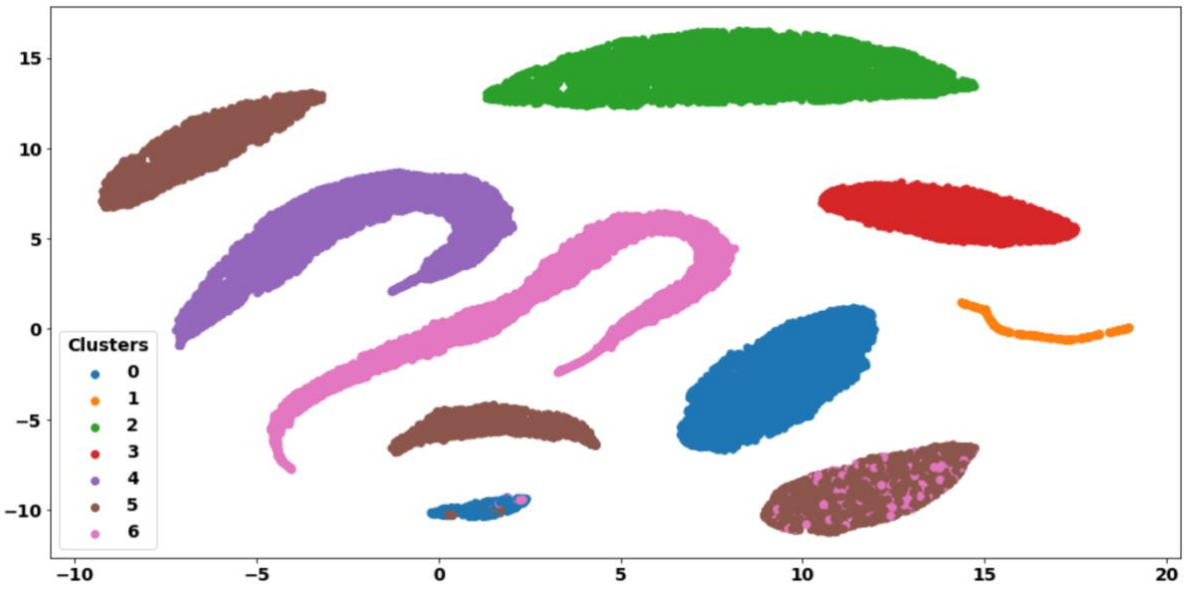
DB Scan and Mean Shift are two density-based clustering algorithms, able to create groups with observations close to each other. With our data, these algorithms did not perform very well, as the vast majority of the observations was put in the outliers' group when using DB Scan, being the rest of the clusters hard to interpret. With Mean Shift, our observations were more evenly spread by each cluster, but when using it with a lower bandwidth (=2), it created too many clusters, some of them very similar between themselves. In addition to that, when using a higher bandwidth (≥5) it just did not create enough clusters.

UMAP

UMAP is a dimensionality reduction technique that is able to preserve the structures of high-dimensionality data. It allows the construction of a visualization on how our observations are spread, in a 2D graph, attributing a different colour to each group of observations. We were able to use this visualisation in all the clustering algorithms previously explained, by taking advantage of a function used in class, *visualize_dimensionality_reduction*.

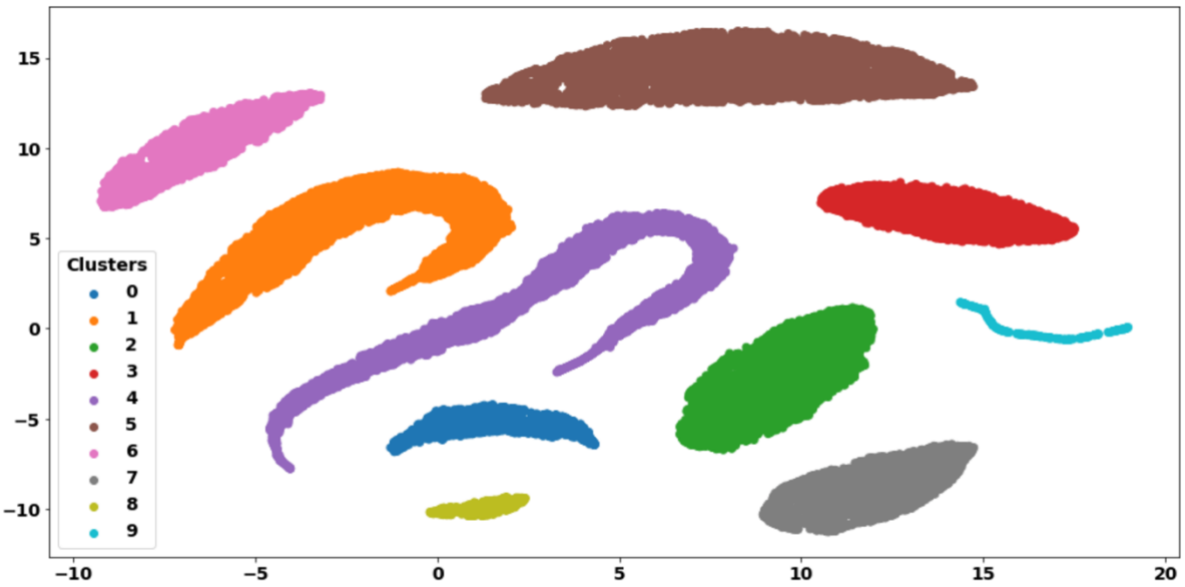
This allowed us to have a better perspective when evaluating which clustering algorithm performed better in our data, as when points of two different colours are close could mean that the algorithm used may not have been able to correctly capture those clusters. An explicative image follows.

Figure 10 - K-Means Clusters



As one can see, this is the representation of the clusters formed by a K-Means with 7 clusters. This shows that there is a mixture between clusters 5 and 6, as well as 0, 5 and 6. Cluster 5 is also split into 3 different agglomerations, and cluster 0 into 2 different ones. All this information could tell us that maybe we should experiment with a K-Means with a couple more clusters, evaluating its results after. The visualisation of the combination of UMAP and DB Scan is next.

Figure 11 – UMAP + DB Scan



This representation seems to be a better solution, as each group of observations has its distinct colour. This was achieved by combining a UMAP with a DB Scan. This process is explained in more details in the next section.

The rest of the visualisations produced with a UMAP will be in the Annexes Section.

UMAP as input of DB Scan

Given that our data has a large number of dimensions, density-based clustering algorithms like DBSCAN and Mean Shift have trouble in finding clusters, as seen previously. Therefore, we decided to reduce dimensionality by using UMAP before running these algorithms.

The results were a lot better this time, as DB Scan was able to find 10 clusters, this is more than both K-Means and Hierarchical methods suggested. The new clusters suggested by DB Scan were *students*, who have an average age of 22 and are characterised by spending a lot in alcoholic drinks, *elderly*, that have an average age of 70 and an extremely high number of complaints and *Millennials,* which have a high percentage of PhD's and MSc's.

However, these results should be taken with a grain of salt, as no dimensionality reduction technique is perfect and, in the process, important information can be lost, UMAP can create false tears in clusters, resulting in a larger number of clusters than is necessarily present in the data. So, it is important to analyse if these new clusters are different enough from previous ones to warrant being called a cluster.

Final Decision

After careful consideration and comparison between our several clustering solutions, we ended up choosing the UMAP + DBSCAN clustering solution which identified the following clusters. The average values for each variable of all clusters are displayed in the Annex section.

Students

When plotting these individuals on the map it is possible to see 3 main clusters of points (Cidade Universitária, NOVA IMS and NOVA FCT in Caparica). This cluster has the lowest average age within all our clusters at 22, also none of them have completed their bachelors yet. This cluster tends to come to our stores at night and buy primarily alcoholic drinks, having a low value in all other categories. This cluster contains 1248 individuals.

Gamers

This group of customers also opt by making their purchases at night, buying primarily videogames, electronics, and non-alcoholic drinks (maybe energy drinks). This cluster contains 4610 individuals.

Promotion Lovers

These clients seem to love going hunting for promotions. Half of their purchases have been made when the item is on sale. They also are willing to travel to different stores in order to save every penny, as they have by far the largest average number of distinct stores visited at 20 (which is more than double of the next cluster). Besides this they also are our earliest shoppers, as if they spend the whole night camping on the sidewalk whenever there is a promotion. This cluster contains 4921 individuals.

Big Families

This cluster is obviously defined by the number of minors at home (almost 5 on average). They usually arrive at our stores right after work. They buy a lot of different distinct products, and their money is spent across several categories. They are also our highest spenders in the hygiene category. This cluster contains 4667 individuals.

Vegetarians

These are a group of customers highlighted by the amount of vegetables they purchase, as well as having never bought meat nor fish. They do not really care about promotions. However, they prefer to shop in the morning (on average, 9 o'clock), maybe to be able to get the freshest vegetables for their meals. This cluster contains 4722 individuals.

Loyal clients

This group of customers was interpreted by us as being the loyal clients, as their first purchase was in 2000. They have also bought many distinct products over their lifetime, which

helps to conclude that they must have been costumers for a while. Additionally, they have the highest total lifetime spent on all products. They also have the highest number of clients with loyalty card at 60%. This cluster contains 4764 individuals.

Millennials

This cluster is emphasized by how educated they are, being on average quite young (28 years old). They have the highest percentage of people with Bachelors, Masters and PhDs. This cluster contains 2272 individuals.

Average Clients

These are just average customers, without high expenses on anything. They are not particularly old nor young, they are really just average. This cluster contains 2262 individuals.

Elderly

We classified this group of individuals as the elderly customers. They have an average age of 70 and have an extremely high number of complaints when compared to the rest of the dataset. They are very loyal to their store of choice, as they only ever visit one store. This cluster contains 308 individuals.

Supermarkets

This group was the easiest one to identify, as we already expected to find it after in the EDA section. This is a group that have Supermarket in their customers names and have very big expenses in fish, not buying basically anything of other products. When plotting these individuals on the map, we can see that they are all nearby MARL (Mercado Abastecedor da Região de Lisboa). This map will be in the Annex section. This cluster contains 226 individuals.

Targeted Promotion

In this section of the project, we aim to understand which products are bought together by each of the clusters previously found. In the section above we only used the Customer Information Dataset, whereas for this one the Customer Basket Dataset and the Product Mapping ones will be the most useful. To be able to associate a purchase in the Customer Basket Dataset to a client, we decided to merge it with the Customer Information Dataset on *customer_id*.

To be able to find relations between products and/or product categories, we must use Association Rules. Having implemented them, we decided to make the following promotions, based on metrics like the lift, confidence, and support.

Cluster 0– Students

This segment is characterised by spending most of their money in alcoholic drinks at night. We suggest that during Friday night's alcohol can be purchased with a discount, making it so that partying students choose our stores to buy their beverages before the party. In addition, association rules were able to see that products such as ketchup and cookies further increase sales of beer, cider, white wine, and champagne.

Cluster 1– Gamers

Looking at the top 10 most bought products, it is possible to see that *Pokémon* games sit comfortably at the top, putting these games on sale might attract new gamers to our stores. Association rules show that clients that buy the videogame *half-life: alyx* are more likely to buy AirPods, so if we put it under promotion our sales of AirPods will also increase.

Cluster 2– Promotion Lovers

This group will almost only come to our stores whenever there is a promotion, to the point of nit-picking which store has active promotions. We suggest having promotions in our lowest grossing stores, and wait for them to flood in.

Cluster 3– Big Families

Big families buy primarily groceries, vegetables, and hygiene products. A good idea would be to sell baskets that include products from these categories that the whole family can make use of. These can be advertised in TV commercials.

Cluster 4- Vegetarians

As vegetarians, these customers seem to purchase a lot of tomatoes, asparagus, as well as many other frozen vegetables. They also look like they care a lot about their hygiene, as they usually buy shower gel or deodorant in the same purchase as they buy vegetables.

A good promotion to these customers could be a discount on hygiene products, for example 20%, but only if vegetables are bought in the same purchase. This may lead to an increase in sales of vegetables. Additionally, this could attract more vegetarians, as they are already interested in vegetables, they may take advantage of the promotion in the hygiene products, which they also love.

Cluster 5 – Loyal Clients

These customers buy many different products, as it was concluded in the customer segmentation section. However, taking a closer look at their purchases, we can see that they really enjoy buying ingredients which they can use to cook, like eggs, oil, tomato sauce, etc. They also enjoy some pre-made food, like cakes, muffins, and cereals.

As they usually buy these together, we can tempt them to buy bigger quantities of both buy creating a discount when more than 5 grocery products are bought at the same time, like buy 5 get 1 free (being the free the cheapest one they buy). This will hopefully lead to more sales in this category of products, whilst rewarding the customers that have purchased here for a long time.

This campaign could be advertised on the store itself, as these customers tend to go there regularly.

Cluster 6 – Millennials

This group of customers were not highlighted by having particularly high values of expenses in a certain type of products. As so, looking at the Association rules, we can see that they buy groceries in almost all their purchases, and more than 50% of the times they shop, they buy groceries and vegetables, like fresh bread, asparagus, and olive oil. To take advantage of this, a possible promotion could be a discount in vegetables in case a client also purchases groceries.

We also know that these customers buy videogames or electronics more than 25% of the times. We could try and take advantage of this, as younger generations tend to enjoy more videogames and new technologies. To do so, we could launch flash sales on selected products of these categories, like on some AirPods and on Minecraft. This would hopefully bring more customers of this generation to our stores, leading them to also purchase groceries and vegetables.

In this case, the best way to advertise this would be through social media and email newsletters.

Cluster 7– Average Clients

These customers were previously described as average, as they did not have high expenses in any type of products. However, we found out that their most purchased products are alcoholic drinks, like wine, beer and champaign. They also tend to buy some groceries here. With this in mind, we would suggest that a promotion is created on groceries, especially snacks, as everyone knows that drinking with your stomach empty is not a good idea. Some steaks could also be included in this promotion, as a barbecue is usually accompanied by alcohol consumption.

Having said that, a good sale would be the creation of Fiesta combo, which includes some alcoholic drinks, snacks, and steaks. As this pack has a low quantity of each, our customers will need more than one, or maybe even buy separately more quantity of a certain product, if needed.

This type of sale should be advertised near the wines and beers section, as these customers tend to already go there, trying to catch their attention with this new Fiesta combo.

Cluster 8– Elderly

This group of customers focuses their purchases on groceries and vegetables, mainly on cooking oil and cake. In addition, as we've seen previously, they complain more often than the others, on average, and don't visit more than one store.

For them, we thought about creating a special type of loyalty card, for example, a loyalty store card. With this card, they would get exclusive promotions on their favourite stores. Some of the promotions would be on their favourite products and others on the ones they don't usually buy. This could keep them happy while simultaneously introducing them to different options available.

With the exclusiveness and benefits of this new card, we are hoping for an increase in sales as well as a decrease in complains – and a chance this group will like us more.

Cluster 9– Supermarkets

As we've previously seen, this cluster really stands out when it comes to fish purchases. In order to increase sales on other categories we could create special promotions for these clients.

For instance, we could say that an x amount of money spent on something other than fish will equal an y% discount on the fish they buy. We are confident that they'll buy fish anyway, they may want to get it for a cheaper price.

If they start constantly purchasing other categories in order to get fish discounts, they might even, eventually, find another category they would be willing to invest in as much as fish.

Conclusion

In the beginning of this project, we aimed to segment our clusters and after that, creating custom promotions for them, aiming to increase sales and revenue. To do this, we had to clean our data, followed by experimenting with a lot of different clustering methods, reaching the conclusion that the combination of a UMAP with DB Scan gave us the best results.

After this, we moved on to the Targeted Promotion section, where we took advantage of Association Rules to take conclusions about each cluster. We found out that some of our clusters were more into all products that were on sale, whereas others just enjoyed to have a glass, or maybe even a bottle, of an alcoholic beverage.

With the conclusion of this project, we believe to have found an optimal clustering solution, as well as some creative and interesting promotions, targeted to each group of customers.

We would also like to add that this project was quite enjoyable, thus requiring a lot of effort, time, and knowledge application.

Annexes

1. K-Means with 8 clusters

cluster_kmeans8	0	1	2	3	4	5	6	7
customer_id	14496.980769	14995.381192	14903.834273	15070.718655	14358.973451	15129.863875	15071.617017	15010.749530
number_complaints	0.000000	1.551008	1.186929	0.500000	0.000000	0.280579	0.933562	0.313402
distinct_stores_visited	2.000000	7.999580	19.188679	1.992625	1.000000	2.490794	3.007287	1.998955
lifetime_spend_groceries	100.346955	14993.307935	298.519549	200.406291	2.701444	3788.005042	4997.019288	993.002718
lifetime_spend_electronics	19.899038	199.970403	19.566427	4999.557484	1.149498	795.206488	200.005144	56.429856
typical_hour	21.606571	11.914568	9.035207	20.951844	0.620237	16.492986	17.403986	9.900272
lifetime_spend_vegetables	20.133814	799.560034	291.663101	20.033623	2.124160	347.610259	602.087870	1485.383441
lifetime_spend_nonalcohol_drinks	199.929487	900.786104	291.840303	1500.946855	2.302082	347.897852	900.797471	23.244825
lifetime_spend_alcohol_drinks	900.553686	599.145886	204.041432	499.866161	2.704784	351.199036	500.929704	22.712733
lifetime_spend_meat	49.848558	1501.016583	146.725540	57.739913	2.615967	446.011837	1100.672525	2.459753
lifetime_spend_fish	49.891026	1498.781276	145.276211	58.946421	34997.747788	445.417580	1101.494856	2.598369
lifetime_spend_hygiene	49.731571	199.985306	50.000584	49.912148	1.553669	125.031346	500.780969	99.394731
lifetime_spend_videogames	100.166667	50.012804	4.786034	2002.197397	0.000000	347.173608	997.684312	51.851558
lifetime_total_distinct_products	79.710737	4011.462636	482.479090	99.785249	0.977930	280.740026	2001.353622	197.956095
percentage_of_products_bought_promotion	0.199982	0.149955	0.487130	0.100002	0.200151	0.126840	0.200118	0.050699
year_first_transaction	2016.899038	2000.006507	2010.011282	2009.978091	2010.030973	2009.935774	2009.949850	2010.017562
latitude	38.724097	38.747749	38.747763	38.747749	38.866276	38.748681	38.749546	38.748642
longitude	-9.172161	-9.157703	-9.157987	-9.157489	-9.111225	-9.157008	-9.157404	-9.157217
loyalty_card	0.020032	0.599916	0.048823	0.100000	0.101770	0.124726	0.300043	0.049760
gender	0.500801	0.500210	0.509045	0.505857	0.446903	0.494739	0.499786	0.510767
age	22.441506	55.653862	55.984439	55.752711	50.026549	40.534415	55.612516	55.606105
Basic	1.000000	0.650084	0.904299	0.500000	0.000000	0.503507	0.499786	0.504704
Bsc	0.000000	0.112091	0.033067	0.163124	0.000000	0.171197	0.171453	0.162659
Msc	0.000000	0.124055	0.030928	0.164859	0.000000	0.167251	0.161809	0.162868
Phd	0.000000	0.113770	0.031706	0.172017	0.000000	0.158045	0.166952	0.169768
lifetime_total_spend	1440.769231	20542.581024	1402.418596	9339.694143	35011.345723	6868.521701	10400.691170	2637.683253
minors_home	0.016026	1.269521	1.067497	0.758134	0.000000	0.593819	4.891342	1.924733

2. Hierarchical with 7 clusters

cluster_ward7	0	1	2	3	4	5	6
customer_id	15300.045873	14995.381192	14650.983054	14881.347897	15072.095350	14358.973451	15070.718655
number_complaints	0.395948	1.551008	0.318661	1.010567	0.933576	0.000000	0.500000
distinct_stores_visited	2.338021	7.999580	1.998760	20.001829	3.007285	1.000000	1.992625
lifetime_spend_groceries	2915.509124	14993.307935	987.835710	302.959764	4997.018856	2.701444	200.406291
lifetime_spend_electronics	606.648250	199.970403	62.187229	20.000000	200.006642	1.149498	4999.557484
typical_hour	17.245103	11.914568	10.073982	8.991668	17.401971	0.620237	20.951844
lifetime_spend_vegetables	271.326302	799.560034	1470.598884	300.231863	602.098779	2.124160	20.033623
lifetime_spend_nonalcohol_drinks	308.356437	900.786104	26.257285	300.431823	900.787872	2.302082	1500.946855
lifetime_spend_alcohol_drinks	463.373514	599.145886	27.377144	199.776671	500.900364	2.704784	499.866161
lifetime_spend_meat	350.664490	1501.016583	4.923125	151.107702	1100.662738	2.615967	57.739913
lifetime_spend_fish	350.970367	1498.781276	4.253565	149.465962	1101.494750	34997.747788	58.946421
lifetime_spend_hygiene	107.169931	199.985306	98.828890	50.032514	500.748018	1.553669	49.912148
lifetime_spend_videogames	284.166416	50.012804	53.400083	5.000000	997.704307	0.000000	2002.197397
lifetime_total_distinct_products	233.829734	4011.462636	196.383964	499.575493	2001.371759	0.977930	99.785249
percentage_of_products_bought_promotion	0.145053	0.149955	0.051274	0.499966	0.200119	0.200151	0.100002
year_first_transaction	2011.382053	2000.006507	2010.023765	2010.019508	2009.948361	2010.030973	2009.978091
latitude	38.743467	38.747749	38.748622	38.747834	38.749548	38.866276	38.747749
longitude	-9.160891	-9.157703	-9.157162	-9.157208	-9.157408	-9.111225	-9.157489
loyalty_card	0.081031	0.599916	0.072949	0.049990	0.299979	0.101770	0.100000
gender	0.498744	0.500210	0.508369	0.508433	0.499893	0.446903	0.505857
age	37.936883	55.653862	55.361645	55.311725	55.605742	50.026549	55.752711
Basic	0.638875	0.650084	0.488117	0.900020	0.499893	0.000000	0.500000
Bsc	0.123389	0.112091	0.169870	0.034546	0.171416	0.000000	0.163124
Msc	0.122551	0.124055	0.167390	0.032311	0.161774	0.000000	0.164859
Phd	0.115185	0.113770	0.174623	0.033123	0.166917	0.000000	0.172017
lifetime_total_spend	5551.014900	20542.581024	2636.833023	1428.973786	10400.674309	35011.345723	9339.694143
minors_home	0.486523	1.269521	1.911552	1.070108	4.890508	0.000000	0.758134

3. DB Scan with 6 clusters, being the cluster -1 the outliers

cluster_dbscan	-1	0	1	2	3	4
customer_id	14941.476498	15177.457447	14916.213656	13959.238710	18475.755495	17653.762626
number_complaints	0.808597	0.000000	0.000000	0.000000	0.000000	0.000000
distinct_stores_visited	6.339610	2.000000	2.000000	2.000000	2.975275	3.015152
lifetime_spend_groceries	4066.156406	100.695745	100.134361	99.374194	7018.813187	6989.186869
lifetime_spend_electronics	1000.597087	19.948936	19.867841	19.787097	999.563187	989.530303
typical_hour	13.987664	21.614894	21.612335	21.658065	19.005495	19.106061
lifetime_spend_vegetables	588.105824	20.089362	20.352423	20.483871	403.469780	401.121212
lifetime_spend_nonalcohol_drinks	649.580601	199.087234	200.837004	199.561290	397.524725	407.404040
lifetime_spend_alcohol_drinks	359.024799	900.755319	901.266520	899.909677	402.252747	405.636364
lifetime_spend_meat	527.835121	49.859574	49.700441	50.283871	700.587912	700.348485
lifetime_spend_fish	806.329666	50.280851	49.788546	49.612903	700.936813	700.934343
lifetime_spend_hygiene	166.708527	49.568085	50.024229	49.470968	199.766484	154.838384
lifetime_spend_videogames	554.212278	100.465957	100.037445	100.348387	493.791209	505.934343
lifetime_total_distinct_products	1175.864841	80.470213	79.693833	78.212903	501.063187	513.919192
percentage_of_products_bought_promotion	0.191459	0.199982	0.199827	0.200091	0.149783	0.149517
year_first_transaction	2008.341796	2016.827660	2016.995595	2016.954839	2009.914835	2010.580808
latitude	38.748784	38.744617	38.744818	38.661590	38.749591	38.749723
longitude	-9.157090	-9.161305	-9.161473	-9.204891	-9.171490	-9.171649
loyalty_card	0.205402	0.000000	0.000000	0.000000	0.000000	0.000000
gender	0.497796	0.000000	1.000000	1.000000	1.000000	0.000000
age	53.568285	22.470213	22.418502	22.425806	28.950549	29.080808
Basic	0.603653	1.000000	1.000000	1.000000	0.271978	0.227273
Bsc	0.130082	0.000000	0.000000	0.000000	0.225275	0.222222
Msc	0.128954	0.000000	0.000000	0.000000	0.255495	0.287879
Phd	0.129342	0.000000	0.000000	0.000000	0.247253	0.262626
lifetime_total_spend	8551.841783	1441.182979	1441.984581	1439.361290	11116.939560	11100.095960
minors_home	1.752847	0.014894	0.017621	0.019355	0.112637	0.106061

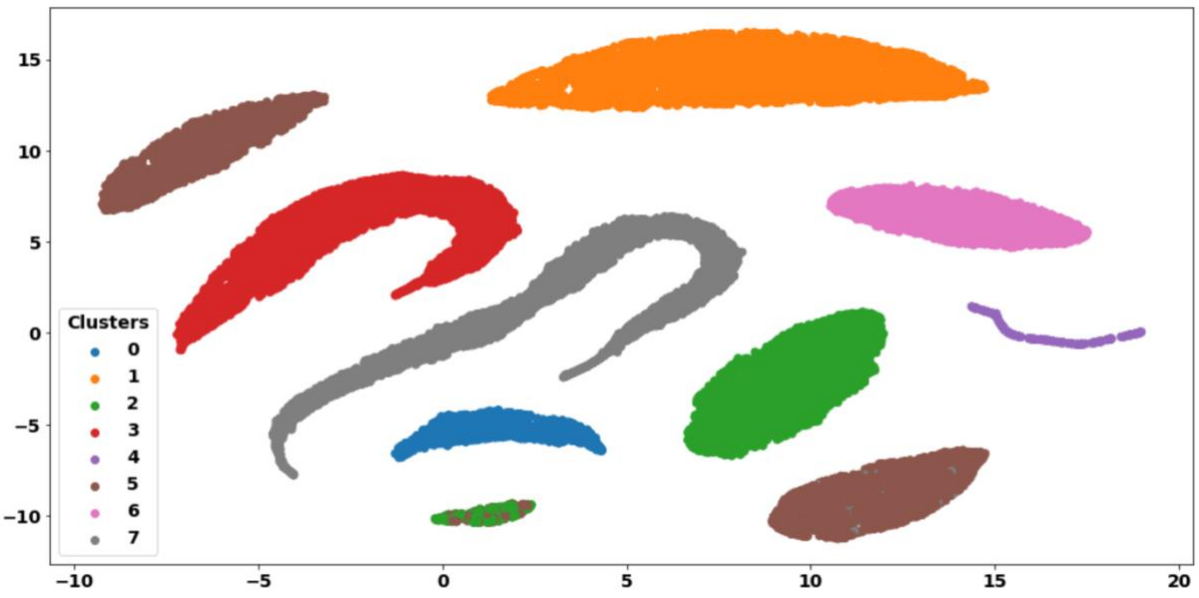
4. Mean Shift with 5 clusters

cluster_meanshift	0	1	2	3	4
customer_id	15036.863067	14883.376832	14995.381192	14978.810330	15070.718655
number_complaints	0.286675	1.142178	1.551008	0.922523	0.500000
distinct_stores_visited	2.170844	19.516436	7.999580	3.008256	1.992625
lifetime_spend_groceries	2005.496455	300.166139	14993.307935	5022.766511	200.406291
lifetime_spend_electronics	356.275598	19.742574	199.970403	209.682896	4999.557484
typical_hour	13.776831	9.019604	11.914568	17.421888	20.951844
lifetime_spend_vegetables	801.852014	295.113861	799.560034	599.651355	20.033623
lifetime_spend_nonalcohol_drinks	178.150320	295.307129	900.786104	894.836156	1500.946855
lifetime_spend_alcohol_drinks	261.435337	202.589703	599.145886	499.883150	499.866161
lifetime_spend_meat	191.035404	148.488713	1501.016583	1095.754234	57.739913
lifetime_spend_fish	919.646056	146.951485	1498.781276	1096.637172	58.946421
lifetime_spend_hygiene	100.461494	49.999208	199.985306	499.408764	49.912148
lifetime_spend_videogames	177.429598	4.872277	50.012804	992.072185	2002.197397
lifetime_total_distinct_products	212.465906	489.358218	4011.462636	1983.290008	99.785249
percentage_of_products_bought_promotion	0.103696	0.492310	0.149955	0.199527	0.100002
year_first_transaction	2010.773867	2010.014455	2000.006507	2009.949407	2009.978091
latitude	38.748261	38.747782	38.747749	38.749556	38.747749
longitude	-9.158062	-9.157641	-9.157703	-9.157391	-9.157489
loyalty_card	0.074825	0.049703	0.599916	0.305461	0.100000
gender	0.501474	0.509901	0.500210	0.499153	0.505857
age	45.607077	55.722178	55.653862	55.286622	55.752711
Basic	0.557040	0.902574	0.650084	0.494708	0.500000
Bsc	0.141725	0.033663	0.112091	0.173793	0.163124
Msc	0.140804	0.031485	0.124055	0.162786	0.164859
Phd	0.139606	0.032277	0.113770	0.168713	0.172017
lifetime_total_spend	4891.320783	1413.231881	20542.581024	10411.283658	9339.694143
minors_home	1.106248	1.068515	1.269521	4.835944	0.758134

5. UMAP + DB Scan with 10 clusters

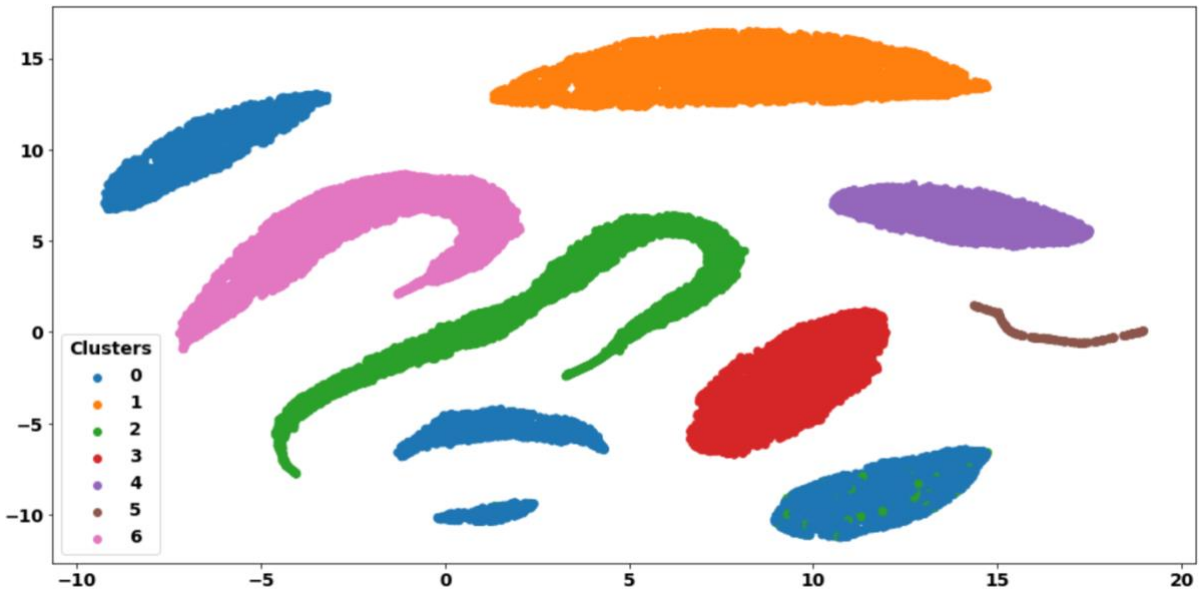
cluster_umap_dbscan	0	1	2	3	4	5	6	7	8	9
customer_id	14496.980769	15070.718655	14881.347897	15072.095350	14998.421855	14995.381192	15105.493398	15222.413351	14986.107143	14358.973451
number_complaints	0.000000	0.500000	1.010567	0.933576	0.310885	1.551008	0.018486	0.496021	4.139610	0.000000
distinct_stores_visited	2.000000	1.992625	20.001829	3.007285	2.000212	7.999580	3.009683	2.011494	1.000000	1.000000
lifetime_spend_groceries	100.346955	200.406291	302.959764	4997.018856	997.551461	14993.307935	6998.608715	617.420866	199.211039	2.701444
lifetime_spend_electronics	19.899038	4999.557484	20.000000	200.006642	50.036637	199.970403	1002.705986	611.014589	9.948052	1.149498
typical_hour	21.606571	20.951844	8.991668	17.401971	9.988564	11.914568	18.956426	13.907162	10.048701	0.620237
lifetime_spend_vegetables	20.133814	20.033623	300.231863	602.098779	1499.739094	799.560034	402.549736	302.609637	100.224026	2.124160
lifetime_spend_nonalcohol_drinks	199.929487	1500.946855	300.431823	900.787872	20.005294	900.786104	400.469190	302.498674	99.925325	2.302082
lifetime_spend_alcohol_drinks	900.553686	499.866161	199.776671	500.900364	20.117535	599.145886	402.982394	298.152962	296.506494	2.704784
lifetime_spend_meat	49.848558	57.739913	151.107702	1100.662738	0.000000	1501.016583	699.633363	199.534925	49.402597	2.615967
lifetime_spend_fish	49.891026	58.946421	149.465962	1101.494750	0.000000	1498.781276	699.194542	199.086207	51.175325	34997.747788
lifetime_spend_hygiene	49.731571	49.912148	50.032514	500.748018	100.033037	199.985306	200.338908	50.158267	49.707792	1.553669
lifetime_spend_videogames	100.166667	2002.197397	5.000000	997.704307	49.929691	50.012804	502.893926	199.990274	0.000000	0.000000
lifetime_total_distinct_products	79.710737	99.785249	499.575493	2001.371759	199.708598	4011.462636	502.506162	58.268789	100.538961	0.977930
percentage_of_products_bought_promotion	0.199982	0.100002	0.499966	0.200119	0.049988	0.149955	0.149835	0.100307	0.199909	0.200151
year_first_transaction	2016.899038	2009.978091	2010.019508	2009.948361	2010.014824	2000.006507	2009.927377	2009.970822	2009.743506	2010.030973
latitude	38.724097	38.747749	38.747834	38.749548	38.748634	38.747749	38.749198	38.748111	38.747348	38.866276
longitude	-9.172161	-9.157489	-9.157208	-9.157408	-9.157192	-9.157703	-9.156626	-9.156902	-9.174110	-9.111225
loyalty_card	0.020032	0.100000	0.049990	0.299979	0.049979	0.599916	0.199824	0.049956	0.029221	0.101770
gender	0.500801	0.505857	0.508433	0.499893	0.509530	0.500210	0.490317	0.500884	0.522727	0.446903
age	22.441506	55.752711	55.311725	55.605742	55.421432	55.653862	28.937500	51.821397	70.840909	50.026549
Basic	1.000000	0.500000	0.900020	0.499893	0.500000	0.650084	0.199824	0.800177	0.990260	0.000000
Bsc	0.000000	0.163124	0.034546	0.171416	0.163702	0.112091	0.274648	0.071176	0.003247	0.000000
Msc	0.000000	0.164859	0.032311	0.161774	0.164972	0.124055	0.268486	0.067197	0.003247	0.000000
Phd	0.000000	0.172017	0.033123	0.166917	0.171326	0.113770	0.257042	0.061450	0.003247	0.000000
lifetime_total_spend	1440.769231	9339.694143	1428.973786	10400.674309	2637.379712	20542.581024	11109.037852	2730.308134	806.392857	35011.345723
minors_home	0.016026	0.758134	1.070108	4.890508	1.932020	1.269521	0.162852	1.029178	1.022727	0.000000

6. UMAP K-Means 8



7. UMAP K-Means 9

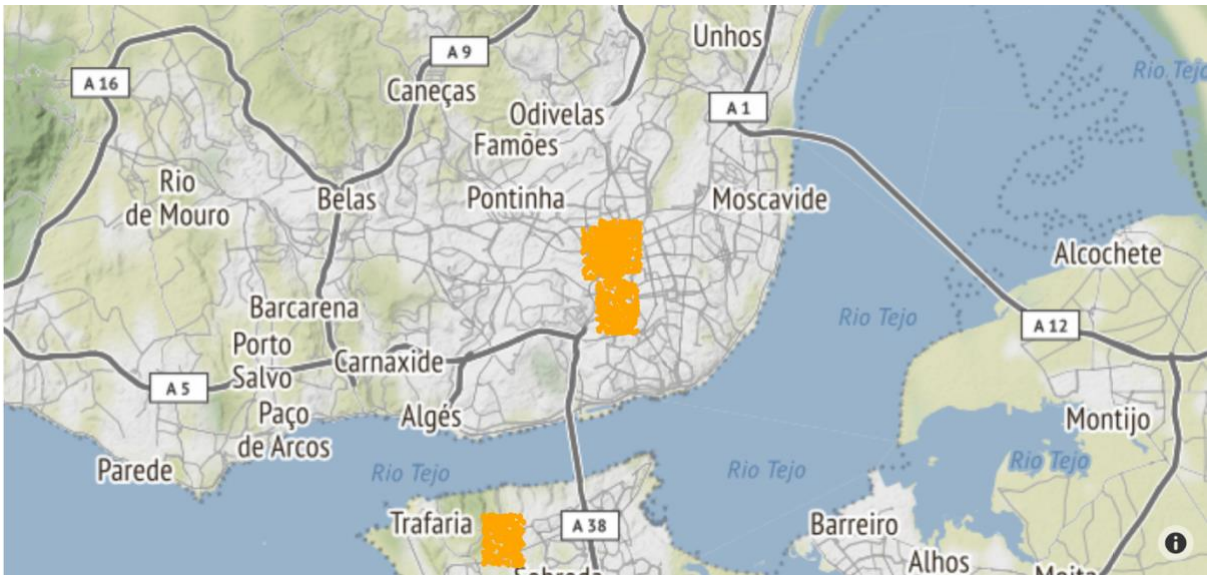
8. UMAP Hierarchical 7



9. Confusion Matrix: UMAP + DB Scan vs K-Means 8

	K-means 0 Cluster	K-means 1 Cluster	K-means 2 Cluster	K-means 3 Cluster	K-means 4 Cluster	K-means 5 Cluster	K-means 6 Cluster	K-means 7 Cluster
DBSCAN 0 Cluster	1248	0	0	0	0	0	0	0
DBSCAN 1 Cluster	0	0	0	4610	0	0	0	0
DBSCAN 2 Cluster	0	0	4921	0	0	0	0	0
DBSCAN 3 Cluster	0	0	0	0	0	1	4666	0
DBSCAN 4 Cluster	0	0	0	0	0	0	0	4722
DBSCAN 5 Cluster	0	4764	0	0	0	0	0	0
DBSCAN 6 Cluster	0	0	0	0	0	2272	0	0
DBSCAN 7 Cluster	0	0	0	0	0	2201	0	61
DBSCAN 8 Cluster	0	0	220	0	0	88	0	0
DBSCAN 9 Cluster	0	0	0	0	226	0	0	0

10. Map of the students



11. Map of the Supermarkets