NOVA IMS Information Management School

MACHINE LEARNING II

CUSTOMER SEGMENTATION:

A KEY TO UNLOCKING BUSINESS GROWTH AND SUCCESS

ALEXANDRA PINTO
20211599

ILONA NACU
20211602

RAFAEL PROENÇA
20211681

5 june 2023

Index

1. Executive Summary **2**
2. Exploratory Data Analysis **2**
2.1. Data Inconsistencies 2
2.2. Features' Transformation 2
2.3. Plotting the customers based on their location 3
2.4. Outliers 3
2.5. Spearman Correlation Matrix 4
2.6. Final Features 4
3. Customer Segmentation **5**
3.1. Previous Attempts 5
3.1.1. RFM 6
3.1.2. Self-Organizing Maps (SOM) 7
3.1.3. DBSCAN 7
3.1.4. Mean Shift 7
3.2. Before Final Segmentation 8
3.2.1. Hierarchical Clustering: Dendogram 8
3.2.2. K-means: Deciding k 9
3.3. Segmentation 9
3.3.1. T-SNE and UMAP Analysis 10
3.4. Choosing Final Model 11
3.5. Segment Description 11
3.5.1. General Overview 11
3.5.2. In-depth exploration of each segment 12
3.5.3. Geographical Distribution of our Clusters 14
4. Targeted Promotion **15**
4.1. Targeted Promotions 15
3. Vegetarian Family 16
6. Promotion Hunters 16
8. Young Educated Moderate 17
9. Wholesale Meat and Seafood Distributors 17
5. Conclusion **17**
6. References **18**
7. Annex **19**

1. Executive Summary

People instinctually group things into categories to ease understanding and comprehension. Companies seek to do the same, with a fairly similar goal: understand customers and how to keep them. In this information age, this becomes vital for companies to stay afloat as the offer of products keeps increasing and the competitiveness does as well.

Thus, it is necessary to adopt marketing strategies that target specific customer groups, while being broad enough to not alienate others. Dividing customers based on their preferences and needs is a way to ensure retention and that companies are focused on the right aspects of their clients, which maximizes profit and loyalty.

From this context, this project arises customer segmentation to derive useful business insights. Furthermore, we also created campaigns for the clusters that were identified, each targeting a specific group, exemplifying how useful clustering can be and one of the things companies may do with this division.

This report details the steps executed and the choices made, as well as the reasoning behind them.

In terms of methodology, first, we looked at our dataset and explored what columns we had to understand what we were working with. We moved on to check for inconsistencies, such as duplicates, constant variance, and missing values. Next, we performed some transformations on our features. and, with the help of a plot, we identified our first cluster and decided it would be better to separate it from the rest of our data. Having completed all these steps, we moved to scaling our variables and applying our models.

We tried to use many models, but we picked K-Means as our final solution, nine customer segments. We explored our clusters and came up with appropriate marketing campaigns for each, aiming to target each of their specific tastes and needs.

2. Exploratory Data Analysis

The dataset provided to us has information on a company's customers, their spending habits, and some personal information as well. We start by exploring it, looking at the available features, and trying to glean as much information as possible before applying our models.

2.1. Data Inconsistencies

We started with the identification and treatment of possible data inconsistencies, such as duplicates, missing values, and constant features. There were no duplicates in our dataset, but we did find them in the basket, so we proceeded to eliminate those. There were, also, missing values in loyalty_card_number so we treat that issue in the next section, features' transformation.

The only inconsistency was some infinite values in two of the columns, "lifetime_spend_videogames" and "typical_hour", but that issue disappeared once we removed the supermarket clients to a separate cluster.

2.2. Features' Transformation

Next, we performed some transformations. Mainly, we noticed a few categorical features that would work better as dummies, when applying clustering methods, which tend to deal with distances and need scaling. We created the variable Education, from the information present before each customer's name. Initially, we also created 3 dummies from it (Bsc. , Msc. and Phd.) but they ended up not being used in our final solution, so we kept only Education.

Other dummies created were for the Loyalty card feature, as it is more useful to know if a customer has a loyalty card or not, and we state that if it has missing values on loyalty_card_number than is 0, the customer does not have a loyalty card, and 1, otherwise. A dummy for gender as well was created. Furthermore, we created a variable Age, from the 'customer_birthdate'.

2.3. Plotting the customers based on their location

In the dataset, there is a longitude and latitude column, so we plotted them on a map, to check if any interesting patterns appeared. Figures 2 and 3 are approximations of specific parts from Figure 1. We noticed a group of clients clumped together in a region a bit to the north of most of the other customers. Upon further analysis, we saw that these clients were actually supermarkets, located next to the “Mercado Abastecedor de Lisboa”, as seen in Figure 3. We also noticed that these customers had a different spending behavior than the rest of the clients, so we decided to separate them, forming our first cluster so that their behavior could not affect the models used for clustering. We also noticed a group farther south, in Figure 2., and they also had something in common: none of them had any education information before their names, so they only finished high school, probably.

After creating the supermarket cluster we had to treat some strange values, by setting features ‘lifetime_spend_videogames’ and ‘typical_hour’ to zero.

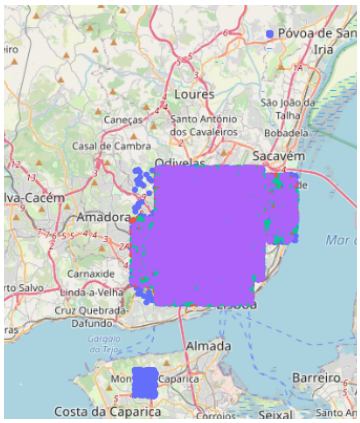

Figure 1.

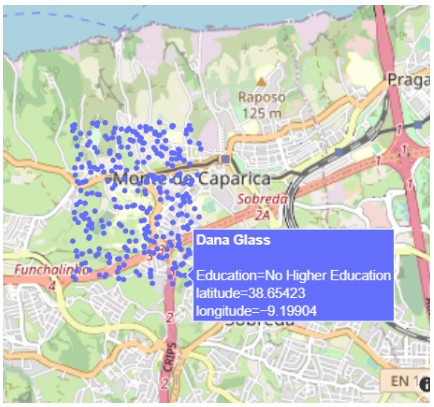

Figure 2.

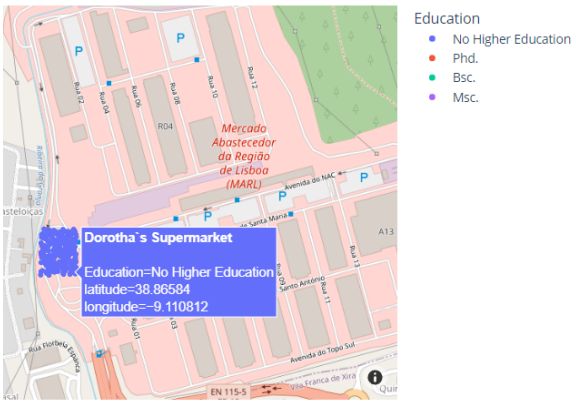

Figure 3.

2.4. Outliers

After separating this cluster from the main dataset, we checked for outliers, because this cluster had very extreme values for some columns, like lifetime_spend_fish_and_meat, as they are buying in bulk, which is not the case for the usual customer. Initially, we had checked for outliers before identifying this cluster, but it became clear that there were a lot, which disappeared after the removal of the cluster.

For this step, we looked at the histogram for each variable, except for binary and non-numerical ones. There are some more extreme values, such as the ones in Figure 4., and we believe that they may belong to specific clusters, as there are very specific and seem significant. The rest of the histograms can be found in 7. Annex, Fig. 5.

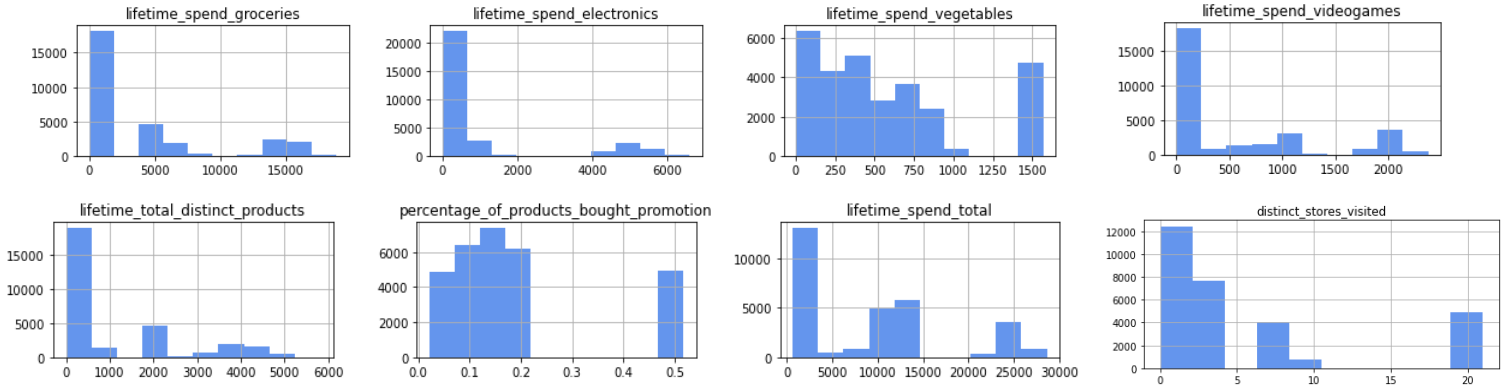

Figure 4.

2.5. Spearman Correlation Matrix

Still in the preprocessing phase, we needed to check if there were any highly correlated variables because if so, it could impact model performance. Also, it would mean that we only need one of the two variables for the clustering stage, as they likely have similar effects. For that, a Spearman Correlation Matrix, which is a statistical tool that is used to evaluate the correlation between two variables, either continuous or discrete, so we proceeded to remove temporarily the other ones. We picked this matrix over the Pearson Correlation Matrix, another commonly used method because Pearson has some assumptions that are too restrictive, such as needing the variables to be normally distributed and have linear relationships with the other variables. The Spearman can deal with non-linear relationships and does not need the variables to follow a normal distribution, thus it is the more robust method.

Highly correlated variables (we considered a value of 0.8 or higher for the variables to be in this list):

- lifetime_total_distinct_products and lifetime_spend_groceries;
- lifetime_total_distinct_products and lifetime_spend_meat;
- lifetime_total_distinct_products and lifetime_spend_fish;
- lifetime_spend_videogames and lifetime_spend_electronics;
- lifetime_spend_fish and lifetime_spend_meat;
- kids_home and children_home;

We decided to combine lifetime_total_distinct_fish and lifetime_spend_meat into one feature, as there was a chance that we would be eliminating important information by dropping one of the features. We applied the same reasoning for lifetime_spend_videogames and lifetime_spend_electronics.

The feature 'lifetime_total_distinct_products' was dropped, since having highly correlated features introduces redundancy in our models. It is possible that there is some loss in information, but we tested the final model with the eliminated feature and without and there was not a big difference between them, meaning that the variable did not add necessary information. One particular case of note if for the variable 'lifetime_spend_groceries', as we initially tried to remove it but, after looking at our clusters with it, we realized that it is useful in our analysis as it was kept in the dataset.

We also dropped 'teens_home' and 'kids_home', because we created a variable that combines both of them, ‘children_home’. We tested the model with them, and then only with the combined variable, and having the two variables did not provide much information, so, in order to have fewer columns, we preferred to keep only ‘children_home’.

2.6. Final Features

These are our final features, after all the transformations, eliminations, and combination of features:

Group	Features	Data	Description
Customer characteristics	customer_name	Categorical	Name of the customer
	latitude	Continuous	Approximate location of the customer's home (Latitude)
	longitude	Continuous	Approximate location of the customer's home (Longitude)
	Male	Binary	Gender of the customer. If 1 its a male customer, if 0 otherwise
	Education	Categorical	Degree of Education. Options are: no higher education, Bsc. , Msc. and Phd.
	children_home	Discrete	Number of children

	age	Discrete	Age of the customer
Spending Habits	lifetime_spend_groceries	Continuous	Total value spent by the customer on groceries
	lifetime_spend_electronics_video	Continuous	Total value spent by the customer on electronics including videogames
	lifetime_spend_fish_and_meat	Continuous	Total value spent by the customer on fish and meat
	lifetime_spend_vegetables	Continuous	Total value spent by the customer on vegetables
	lifetime_spend_nonalcohol_drinks	Continuous	Total value spent by the customer on non-alcoholic drinks
	lifetime_spend_alcohol_drinks	Continuous	Total value spent by the customer on alcoholic drinks
	lifetime_spend_hygiene	Continuous	Total value spent by the customer on hygiene
	percentage_of_products_bought_promotion	Continuous	Percentage of products that were bought with some promotion
	lifetime_spend_total	Continuous	Total value spent by the customer
Store related	loyalty_card	Binary	1 if customer has loyalty card, 0 otherwise
	number_complaints	Discrete	Number of complaints formally done by the customer
	distinct_stores_visited	Discrete	Number of distinct stores visited by the customer
	typical_hour	Discrete	Typical hour when the customer visits the store
	year_first_transaction	Discrete	Year of the first transaction made by the customer

3. Customer Segmentation

This section is dedicated to explaining the models we used for customer segmentation, what led us to our final solution, and an explanation of each of our clusters' meanings.

3.1. Previous Attempts

During Machine Learning II classes, many clustering models were presented and since we wanted to be as thorough as possible, we tested most of them, including RFM (this was was mostly to try to get to know our dataset better, not as an actual final option), K-means, Hierarchical Clustering, SOM, DBSCAN, and Mean Shift. To pick a final model, we based ourselves on how good were the final suggested clusters and on the T-SNE and UMAP graphs as well. We used the same variables for each model, except, of course, the RFM. The model that gave us the best results was K-means, which we will explore in more detail later. All of these were left out of the final code, as they are not useful for our analysis.

We will go more in-depth on our procedure for doing the RFM as we had to make some alterations in the variables and the process itself had a little more to it. We will also explore the results from the Hierarchical Algorithm, which were similar to our final solutions, and we will go more in-depth on why we chose one over the other, but for the other methods, we will be more concise, focusing on the results and why we did not believe them to be satisfactory.

For each model, we also looked at the T-SNE and UMAP of the suggested clusters, in order to access if the solutions were good, except for RFM. T-SNE, T-Distributed Stochastic Neighbor Embedding, is a dimensionality reduction tool useful for visualizing highly dimensional data, which is the case of our data, in a 2D or 3D plot, meaning it's useful for displaying clusters and locating links in large datasets. Uniform

Manifold Approximation Projection, UMAP, is also a visualization method, allowing for the preservation of the local structure of the data while reducing dimensionality. UMAP has some advantages over t-SNE, such as not needing pairwise similarities, running faster, and having better parametrization options, but we used both, in order to be able to have more methods to access the quality of our results.

3.1.1. RFM

Recency, Frequency, and Monetary (RFM), is a marketing technique that analyzes customer behavior based on three factors: Recency, Frequency, and Monetary Value. Recency measures the time since a customer's last purchase, Frequency assesses how often a customer makes purchases, and Monetary Value quantifies the amount of money a customer has spent.

By assigning scores to these factors, customers can be segmented into different groups, allowing businesses to identify valuable customer segments. RFM analysis helps businesses tailor their marketing strategies by targeting specific customer groups based on their scores. Since it's a simple, low-cost method, it is sometimes preferred over more complex ones.

As RFM can only be used for datasets with purchase history, and we only have more general columns, we had to make some alterations to the original method, so as to be able to use it.

As we do not have transaction history, we used the variable 'year_first_transaction' and calculated the years since the first transaction and called it 'year_first', with the idea that older clients, who still shop to this day, are likely to be good to pinpoint.

The frequency part of RFM analysis focuses on quantifying the level of customer activity or engagement based on the number of distinct transactions or purchases. To access the frequency we used the basket data set and grouped by invoice id. In this way, we aim to get the frequency of each customer.

In the monetary part of RFM analysis, we calculate the monetary value for each customer. To do this, we use the variable created 'lifetime_spend_total', obtaining an aggregate measure of the customer's overall spending behavior.

We merge the customer data into a single table, combining Recency, Frequency, and Monetary segments. This comprehensive view allows us to identify patterns and segments within our customer base, enabling targeted strategies for retention, engagement, and personalized marketing.

	monetary_value	customer_recency	customer_frequency
monetary_value	1.000000	0.597770	0.335003
customer_recency	0.597770	1.000000	0.265602
customer_frequency	0.335003	0.265602	1.000000

Figure 5.

The table shows the correlation values between the three dimensions. To simplify the analysis, we can focus on the relationship between frequency and recency, disregarding the monetary dimension, as it had a higher correlation with frequency.

Next, to see the final suggested clusters, we looked at the bubble plot. We did not use 5 for each variable, preferring instead 3, to have fewer clusters, 9 in total, making them easier to interpret.

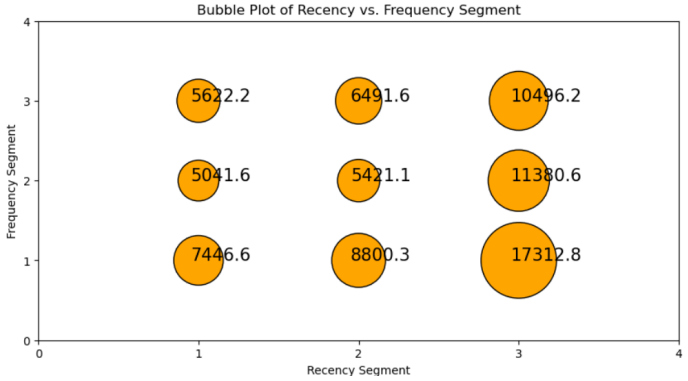

Figure 6.

The size of the bubble indicates how many customers are in each segment, and the values pertain to the monetary value associated with each group. This plot shows that there is not much information that can be gleaned from the formed clusters. They do not seem significant and none of them have interesting values.

Thus concludes our analysis of RFM, with these disappointing results, which makes sense taking into account that we did not even have all the necessary variables to perform the classic RFM. Also, taking into account that this is a very simple method, that only takes those three dimensions into account, it was expected that the resulting clusters would not be very comprehensive, considering all the other useful information is being left out.

3.1.2. Self-Organizing Maps (SOM)

A SOM is a type of unsupervised neural network that organizes data points into clusters based on their similarities. It does this by mapping the data onto a 2D grid of nodes and adjusting the weights of the nodes to match the input patterns. Basically, the nodes compete for the capacity to represent the input data.

Before starting, we performed a quantization errors plot, which helps to choose the number of iterations for our SOM, as it allows us to see how well the neurons are representing the real observations. After looking at Fig.10 Form 7.Annex, we picked 2000.

We performed a distance map and we looked at the influence of each feature in each node, but since the SOM provided so many clusters and the T-SNE and UMAP, Fig. 15 and Fig. 16 in 7. Annex, does not look good, meaning that the obtained clusters are not good, we did not explore this solution further. Our analysis of SOM was very superficial, once we realized that it did not work very well in our case, we did not spend much more time with it, nor did we analyze the results with as much care as with the methods that presented better results.

3.1.3. DBSCAN

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a density-based algorithm, meaning that, unlike k-means for example, it is not based on the distance between points but on density, so it groups points that are packed together, meaning that they have many neighbors close by. The algorithm has two important hyperparameters, epsilon, and minimum points. It starts by selecting an arbitrary data point and checks if there are at least MinPts data points within a distance of ε (epsilon) from it. The process is repeated until either every point is in a cluster or deemed as an outlier. An advantage of DBSCAN is that we do not need to set the number of clusters a priori.

The results given by the DBSCAN, with an eps = 2 and min_samples = 700, were not awful, but it suggested too many clusters, many very similar between each other, such as clusters 0 and 3, as seen in Fig. 17 from 7.Annex. The T-SNE and the UMAP, Fig. 19 and 20 from 7.Annex, also show some overlap between clusters. We tried other combinations of values for the eps and min_samples, but the results were even worse. Either we ended up with a lot more clusters or much less.

Overall, the segmentation was not satisfactory and the results did not provide useful information.

3.1.4. Mean Shift

Mean shift is also an iterative density-based algorithm. It automatically discovers clusters in a dataset without knowing the number of clusters beforehand. It shifts windows towards dense regions, gradually converging to cluster centroids. It can handle clusters of different shapes and sizes and does not require specifying the number of clusters in advance. One important parameter is the bandwidth, which has a similar function as that of the epsilon in DBSCAN, meaning that is the radius of the sliding window.

We used a bandwidth of 5, which is lower than the suggested bandwidth (obtained with the function estimate_bandwith()), as that one only gave one cluster. With a bandwidth of 4, we obtained 6 clusters. Looking at the means of each variable for each cluster, we can identify some clusters in common with our final solution. But since there are only 6, and our final method finds more, in a more conclusive way, we

prefer the solution given by k-means. Looking at the T-SNE (Fig. 23 in 7. Annex), it appears that our model did not capture all clusters, looking at clusters 5 and 1. The UMAP (Fig. 24 in 7. Annex), looks even worse, cluster 5 is separated and appears to overlap with some elements from clusters 6 and 1.

Overall, we think density-based methods in general are not appropriate for our dataset. They managed to capture some patterns, but not all the necessary ones to make them good solutions. The other methods, especially hierarchical clustering, could possibly be made to give better results, but we preferred to focus our attention on K means.

3.2. Before Final Segmentation

Both Hierarchical Clustering and K-means gave us good results and fairly similar final clusters. We explored both models, looking at the advantages and disadvantages of each. Ultimately, we chose K-means, but we will show the results for both, in order to better explain why we made our decision. For the rest of this section, we will present the results of both models, the reasons behind our final decision, and, at last, explain the obtained clusters.

Hierarchical clusters seek to build a hierarchy of clusters. The most common type, and the one we used, is the Agglomerative type, in which each point starts as a cluster, and based on a similarity measure, like distance, the two closest clusters are merged into a single cluster. This process is repeated until all data points are combined into one large cluster or until a specified stopping criterion is met.

K-means is a very popular method, used to cluster data into groups, decided at the beginning, based on similarity between points, using distances. We use the default Euclidean distance. The algorithm iteratively assigns each data point to the nearest cluster center and then re-computes the cluster centers based on the average of all the data points in the cluster. This process continues until the cluster assignments do not change.

3.2.1. Hierarchical Clustering: Dendogram

An important part of this method is the dendrogram, which is a tree-like representation of the algorithm, it's where we can look to have an idea of the appropriate number of clusters.

As previously stated, we decided to use the Agglomerative approach, and we tested it with the single and the ward linkages. We tested all linkage methods, including single, complete, ward, and average linkage, whose dendrograms can be found in Fig. 25 to 27. in 7. Annex. With Ward Linkage, we obtained the best results. Single computes the minimum distances between every observation in one cluster and every observation in the other clusters. Complete is similar, but it uses the maximum distances, and the Average linkage uses the average of the distances. Ward linkage, instead, minimizes the variance of the clusters it tries to merge. [1]

In Figure 7 is the dendrogram with the Agglomerative method, using the Ward Linkage method, from it, we decided to implement 8 clusters.

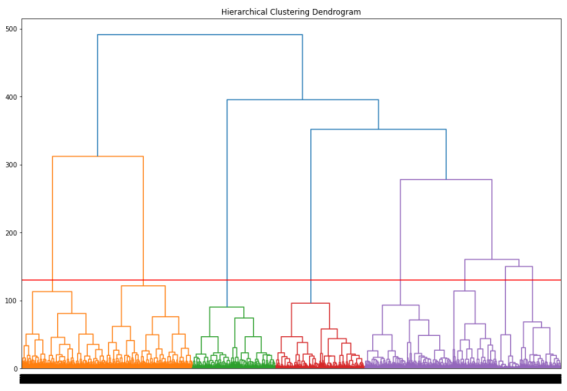

Figure 7.

3.2.2. K-means: Deciding k

When applying K-means, we realized that our results changed drastically when we did not use the binary variables[2] in the algorithm, and our results improved a lot, as the final clusters were much more homogeneous, meaning that we only used discrete and continuous features. Not using binary variables tends to be good practice in the case of K-means. [3]

In order to decide the number of clusters to use in clustering with K-means, we used the Elbow method and the Silhouette Method, to see an appropriate number.

The Elbow Curve, whose name is due to the usual shape of the graph, will show how much value is added, or not, when increasing the number of clusters, this value is shown by the attribute 'inertia', which measures the average distance of each observation to its centroid. We pick k according to where the value falls suddenly and afterward there does not seem to be much of a difference in values.

The Silhouette Method, on the other hand, determines how much a point resembles its own cluster in comparison to other clusters.[4] Higher silhouette coefficient values indicate that the sample is more separated from the surrounding clusters, hence that is what we want. [5]

The elbow method seems to suggest between 6 and 10, as seen in Figure 8. We do not want to have more than 9, as we may lose interpretability. From the Silhouette Method, figure 9., 8 has the highest score, followed closely by 6 and 7.

The silhouette scores for 6, 7, and 8 are:

For 6 clusters, the average silhouette_score is: 0.47149387392009234

For 7 clusters, the average silhouette_score is: 0.4707188153007983

For 8 clusters, the average silhouette_score is: 0.479362030393261

So, the values are quite similar. But for 8 clusters it's slightly higher.

We tried 6, 7, and 8, and with 8 clusters we obtained the best results. With 6 and 7 we lost some information, and some conclusions were not as clear.

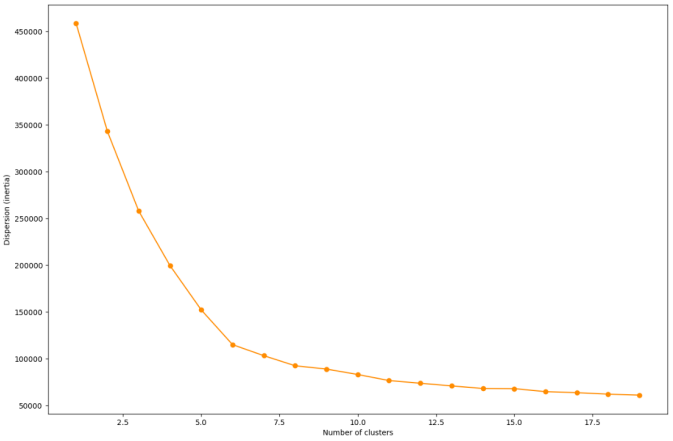

Figure 8. Elbow Method

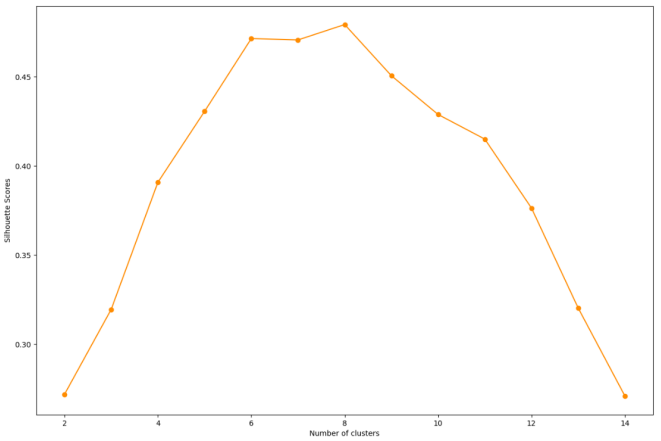

Figure 9. Silhouette Scores Plot

3.3. Segmentation

After realizing that the best number of clusters for both methods is eight, we proceed to look at each suggested segmentation individually, in order to access each solution and see what clusters it suggests and if they are comprehensible and have characteristics that make sense.

Both solutions give very similar clusters, with very similar solutions, as we can see in Fig. 32 in the 7. Annex and Figure 15 in section 4.5.1. General Overview.

We will explain each cluster a bit better after choosing our final solution, but it's very clear that the values are very similar for both methods and that the clusters in both solutions match each other, as can be seen in the table below, where each cluster in a model is matched with the corresponding one in the other model.

K-Means Cluster	Hierarchical Solution Cluster
1	7
2	2
3	4
4	5
5	1
6	3
7	0
8	6

3.3.1. T-SNE and UMAP Analysis

The differences are more pronounced here for the T-SNE and the UMAP. For the Hierarchical Clustering solution, Figures 10 and 11, cluster 4 does not seem well defined, and there is some overlap between 5 and 6. Cluster 7 also seems like it could be a bit better. Whereas for k-means, all clusters seem well defined, with the exception of cluster 8, which we can see in Figure 13, that it overlaps with clusters 5 and 7.

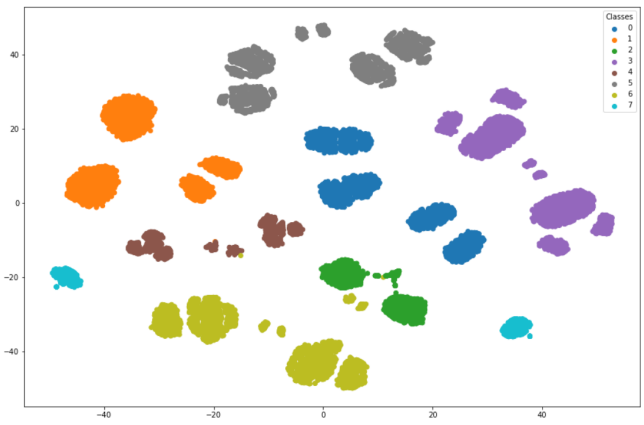

Figure 10. Hierarchical Clustering: T-SNE

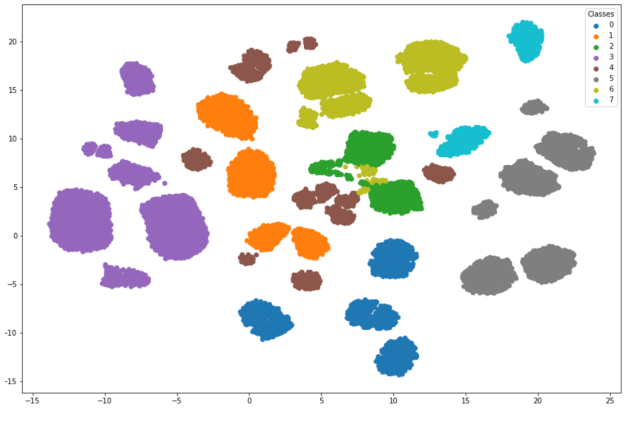

Figure 11. Hierarchical Clustering: UMAP

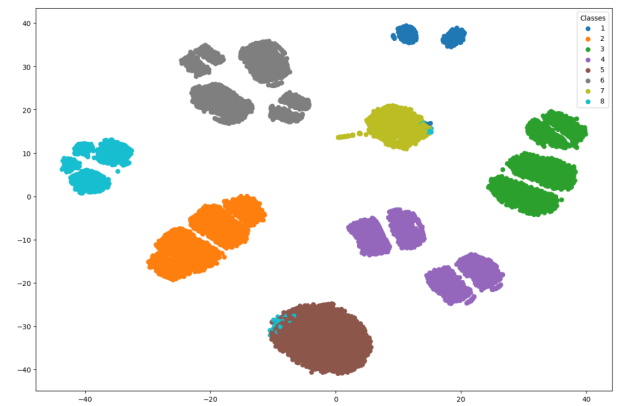

Figure 12. K-Means: T-SNE

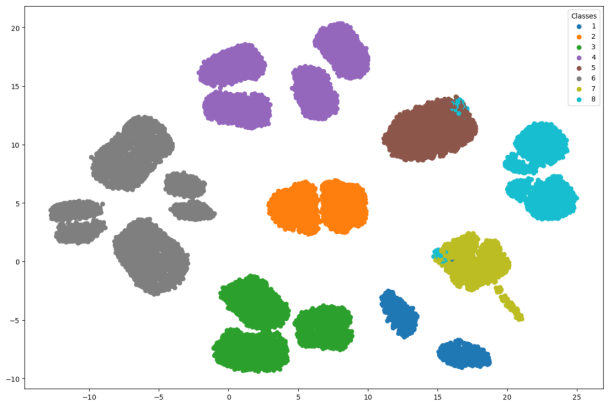

Figure 13. K-Means: UMAP

3.4. Choosing Final Model

Now that we have the results from both our models, it's time to decide on the final solution. Since they are very similar, we will base our decision on technicalities.

The biggest drawbacks of K-means are that it does not deal well with outliers and the number k needs to be decided beforehand. Since our dataset does not have outliers, and there are some methods that can be used to glean what the best k is, these are not issues for our particular case.

As K-means is less computationally expensive than hierarchical clustering, and the UMAP from K-means looks a little better than that of the hierarchical clustering, we will use the solution obtained from K-means as our final one.

3.5. Segment Description

Having a K-means of 8 clusters as our final model, it's time to analyze its results and conclude about exactly what characteristics each segment has and what are the main differences between each of them.

3.5.1. General Overview

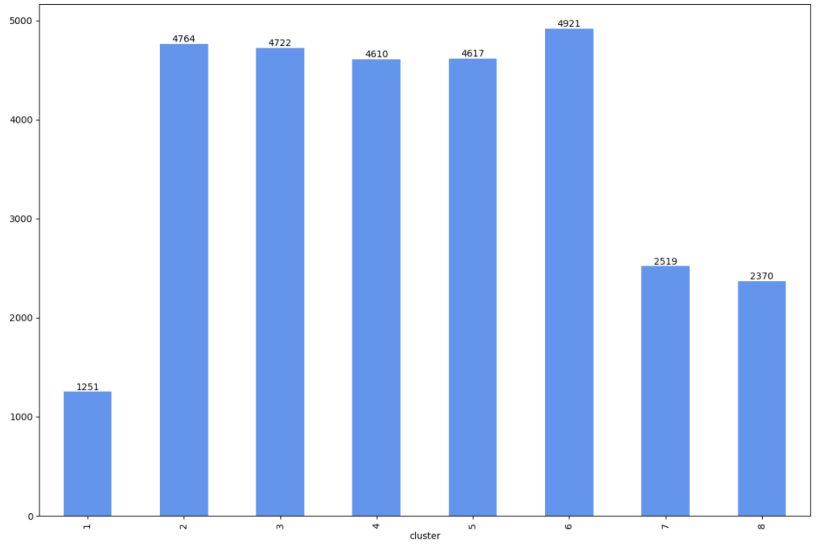

Figure 14. - Cluster Sizes

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Supermarket Cluster	Original
customer_name	James Hernandez	David Jackson	Christopher Smith	Dana Davis	Anne Jones	Mary Smith	Elizabeth Perez	Dorothy Johnson	James`s Supermarket	Robert Smith
Education	No Higher Education	No Higher Education	No Higher Education	No Higher Education	No Higher Education	No Higher Education	No Higher Education	Bsc.	No Higher Education	No Higher Education
loyalty_card	0	1	0	0	0	0	0	0	0	0
Male	1	1	1	1	1	1	1	0	0	1
age	22.0	56.0	55.0	56.0	56.0	55.0	57.0	29.0	50.0	52.0
number_complaints	0.0	2.0	0.0	0.5	1.0	1.0	1.0	0.0	0.0	1.0
distinct_stores_visited	2.0	8.0	2.0	2.0	3.0	20.0	2.0	3.0	1.0	3.0
children_home	0.0	1.0	2.0	1.0	5.0	1.0	1.0	0.0	0.0	1.0
typical_hour	22.0	12.0	10.0	21.0	18.0	9.0	14.0	19.0	0.693147	14.0
year_first_transaction	2017.0	2000.0	2010.0	2010.0	2010.0	2010.0	2010.0	2010.0	2010.0	2009.0
lifetime_spend_electronics_inc_video	120.0	250.0	100.0	7012.0	1199.0	25.0	671.0	1492.0	1.151293	258.0
lifetime_spend_vegetables	20.0	799.5	1499.0	20.0	603.0	300.0	286.0	403.0	2.126912	429.0
lifetime_spend_nonalcohol_drinks	200.0	900.0	20.0	1501.0	902.0	301.0	283.0	401.0	2.303334	431.0
lifetime_spend_alcohol_drinks	900.0	598.0	20.0	500.0	502.0	199.0	298.0	402.0	2.704669	422.0
lifetime_spend_fish_and_meat	100.0	2999.5	0.0	113.0	2203.0	301.0	370.0	1398.0	35013.612777	326.0
lifetime_spend_hygiene	50.0	199.0	100.0	50.0	502.0	50.0	50.0	199.5	1.553652	90.0
lifetime_spend_groceries	100.0	14983.0	998.0	200.0	5000.0	301.0	472.0	6976.0	2.70714	961.0
lifetime_spend_total	1568.0	24756.0	2940.0	9490.0	12908.0	1975.0	2673.0	11821.0	35027.082672	9230.0
percentage_of_products_bought_promotion	0.199785	0.149969	0.049996	0.099991	0.200023	0.499987	0.103815	0.149832	0.200166	0.149596

Figure 15. - K-means: Median and Mode per Cluster

From Figure 14, it's clear that cluster 1 is the smallest, with only 1251 customers, followed by clusters 8 and 7, with 2370 and 2519 customers each. The other five clusters have fairly similar sizes around 4700 each. This seems like a good enough distribution.

Looking at some descriptive statistics, figure 15, the median for numerical features, and mode for non-numerical ones, we can see some patterns for each cluster. We used the median as it is more robust to extreme values than the mean. [6]

- Cluster 1 most interesting values are the median age of their customers, which is 22, being the youngest cluster we have, and we can see that thy only began their transactions in 2017, which makes sense due to their young age. Their spending habits revolve mostly around alcoholic drinks.
- For Cluster 2, loyalty card mode being 1 is what jumps to our attention right away, that and the fact that they, in general, began their transactions in 2000, being our oldest customers. Most of their spending is on groceries.
- As for Cluster 3, their spending is mostly around vegetables, spending around 0 on meat and fish.
- Cluster 4's customers' spending habits are mostly on electronics and video games and non-alcoholic beverages.
- In cluster 5, we see that the median number of children is 5, the highest of all clusters, and their spending habits are focused on electronics and video games, non-alcoholic drinks, and fish and meat.
- The customers in cluster 6 are the ones who go to the stores the earliest, around 9 am, and are the ones that have visited more distinct stores, around 20. Another aspect to note is that they also have the highest percentage of products bought on promotion.
- Regarding cluster 7, we see that they have mostly below-average spending, except for electronics and video games, and even then the difference is not significant.
- For cluster 8, the final cluster, we see that they are customers whose mode is having completed their bachelor, and their age is around 29, being the second youngest customers. Most of their purchases are fish and meat and electronics and video games.

This was just a very brief description of the most noteworthy features for each cluster, we will explain to what conclusions we came from them.

3.5.2. In-depth exploration of each segment

So, based on the description provided so far, here are all the cluster we found, including the Supermarket cluster that we found in section 2, with comprehensive and useful insights, that will then allows us to move on to provide an effective targeting campaign for each.

Cluster 1 - Students:

This cluster represents young male customers with no higher education. They have no complaints and visit a few distinct stores. They have no children at home and prefer shopping at night. These customers started their transactions in 2017 and do not hold a loyalty cards. Their spending is relatively lower compared to other clusters, with minimal values on electronics, groceries, fish and meat, hygiene products, and vegetables. They show a moderate preference for buying products on promotion. An important characteristic is that they spend a lot of their budget on alcoholic drinks.

Cluster 2 - Loyal Grocery Shoppers:

Cluster 2 consists of customers who have shown a strong commitment to their shopping habits, as they hold a loyalty card and have been making purchases since 2000 and they are mostly male. Notably, these customers demonstrate a significant focus on fish, meat, and groceries, indicating a preference for essential food items. Their substantial spending in these categories suggests a possible emphasis on quality, freshness, or adherence to a particular dietary lifestyle. With their long-standing loyalty and dedicated approach to grocery shopping.

Cluster 3 - Vegetarian Family:

In Cluster 3, customers show a complete avoidance of meat and fish. They prioritize a vegetarian or plant-based diet, focusing primarily on vegetables. They also exhibit minimal spending on drinks and alcohol. These individuals demonstrate a strong inclination towards a healthy lifestyle, emphasizing plant-based food choices and reducing their consumption of animal products. This name reflects the cluster's focus on vegetable consumption, minimal spending on meat and fish, and a commitment to a healthy and sustainable lifestyle, including the fact that it has a median of two children.

Cluster 4 - Tech Enthusiasts:

Composed mostly of middle age male customers with no higher education and one child. It comprises customers with distinct spending patterns. They demonstrate a strong affinity for electronics and video-related products, indicating a tech-savvy nature. Additionally, they allocate a significant portion of their spending towards various beverages, including both alcoholic and non-alcoholic drinks. However, their expenditure on meat, vegetables, hygiene products, and groceries is relatively lower compared to other customer clusters.

Cluster 5 - Family Shoppers:

Mature Family Shoppers cluster consists of customers who have a higher number of children at home, specifically five children. These customers prioritize meeting the needs of their household and exhibit a higher lifetime spend on groceries, hygiene products, and other household-related items. They understand the importance of providing for their family's needs and maintaining a well-stocked home. This cluster represents families who engage in responsible and strategic shopping, focusing on essential items for their larger household.

Cluster 6 - Promotion Hunters:

This cluster represents mostly middle age male customers with no higher education and one child. It's characterized by a strong inclination towards promotional offers, as nearly half of the products purchased by customers in this cluster are on promotion. These customers actively seek out discounts and special deals, indicating a cost-conscious approach to shopping. Additionally, they tend to visit a significant number of distinct stores, the median is 20, suggesting a willingness to explore various shopping destinations in search of the best bargains. Important characteristic is that these customers spend very little on video games and electronics.

Cluster 7 - Casual Shoppers:

These shoppers consist of moderate spenders. They have no higher education and have made about 1 complaint. These customers visit around 2 distinct stores and typically shop at 2 PM. They have been engaged with the supermarket since 2010. Their highest values are for electronics and video games and for fish and meat, but even then they are not very high. The rest of their spending fall below the general average, which may indicate that they are not avid customers, maybe just visiting when it's convenient, and not actually their preferred stores. This means that they are just casual shoppers, being the ones with the highest risk of being lost to the competitor stores as they do not seem loyal in the first place.

Cluster 8 - Young Educated Moderate:

This cluster represents young, female-educated customers who have a moderate spending pattern. They visit a few distinct stores and do not have any complaints. They have no children at home and prefer shopping in the evening. These customers started their transactions in 2010 and do not hold a loyalty cards. They predominantly spend on electronics, groceries, and hygiene products, while their spending on vegetables, non-alcoholic drinks, and alcoholic drinks is moderate. These customers have a high lifetime spending in total, considering that they do not have kids, being the third cluster with the highest value for

that feature, which could mean that their daily needs for products are met with our stores, making them important clients to keep.

Supermarket Cluster - Wholesale Meat and Seafood Distributors:

The Supermarket Cluster represents a segment of customers, predominantly females, without higher education, who primarily function as resellers, specializing in the distribution of fish and meat products. They have no children at home and show limited spending on other product categories other than meat and fish. These customers exhibit a significantly higher lifetime spending indicating their role as resellers.

3.5.3. Geographical Distribution of our Clusters

We also decided to visualize our clusters in a map, in order to see if any more interesting patterns arise.

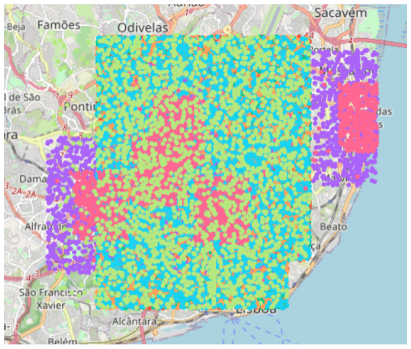

Figure 16.

Figure 17.

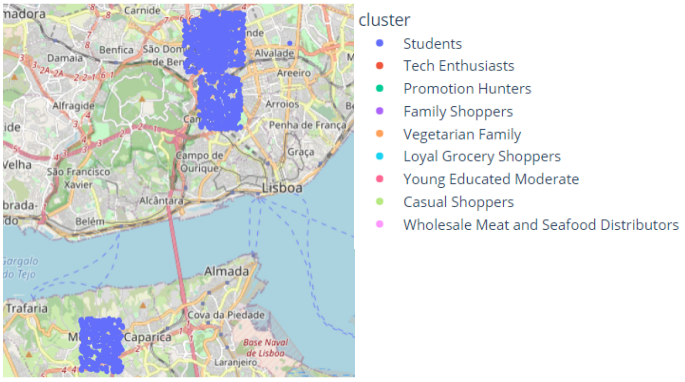

Figure 18 - Only Student Shoppers selected

In addition to the geographical location of the Wholesale Meat and Seafood Distributors cluster, that we had already seen, figure 17, there are other interesting insights that can be gleaned from these maps:

1. The Students cluster, as seen in Figure 18, is situated in areas such as Margem Sul, Campolide, and Campo Grande, which are known for hosting various faculties. This suggests a focus on catering to university students.

2. The Young Educated Moderate, Figure 16, customers predominantly reside in affluent areas like Parque das Nações, Telheiras, and Benfica. This indicates that these customers are financially comfortable, as reflected in their higher lifetime spending.

3. The Family Shoppers cluster, Figure 16, includes customers located in Lisbon, but they tend to reside farther away from the city center.

4. Clusters such as Tech Enthusiasts, Promotion Hunters, Vegetarian Families, and Loyal Grocery Shoppers, Figure 16, but can be seen more clearly in Fig. 45 and 46 in 7. Annex, are concentrated in a specific area between Odivelas and Alcantara. The Casual Shoppers cluster also shares this general region, although some of them are located in the direction of Sintra, which is considered a more affordable area.

Overall, the geographical distribution of these clusters confirms our earlier description of some of them, showing that our insights are valuable and make sense considering the comprehensive results obtained, meaning that we now understand the preferences and characteristics of different customer segments.

4. Targeted Promotion

Now that we have classified and named our clusters, it's important to create association rules for each. That will allow us to suggest customized promotions based on the distinct characteristics and preferences of each group.

These rules enable us to unveil patterns and correlations within the clusters, facilitating personalized recommendations and targeted promotions. Leveraging these insights allows us to optimize our marketing strategies and enhance customer satisfaction by delivering relevant products that align with their individual purchasing behaviors.

In order to create targeted promotions for our segments, we based some of them on the mode and median of each cluster and the rest on association rules.

The Apriori algorithm, the one we used, is a technique used to find frequent item sets and discover association rules in large datasets based on support, confidence, and lift measures. Before diving into the analysis of the association rules for each cluster, it's important to understand a few key concepts:

- Support: This indicates the frequency of occurrence of a particular rule, showing how often the items in the rule appear together in the dataset.
- Confidence: It represents the conditional probability of the consequent item given the antecedent item. In other words, it shows how likely the consequent item is to be purchased when the antecedent item is already in the customer's basket.
- Lift: This metric measures the strength of association between two items in the context of association rules analysis. It quantifies the extent to which the presence of one item influences the presence of another item. A lift value greater than 1 suggests a positive association, indicating that the presence of one item increases the likelihood of the other item being present. In our analysis, all clusters displayed lift values greater than 1, signifying a significant association between items within each cluster.

These insights can be highly valuable for targeted marketing efforts, optimizing product placements, and enhancing product recommendations to enhance customer satisfaction and boost sales.

An important note is that we had to look for a lift higher than 1 in each cluster. When we tried to be more selective, imposing some conditions about confidence and support the results were null.

The results of the association rules for each cluster can be found in 7.Annex, Fig. 37 to 44.

4.1. Targeted Promotions

In this section we will propose some promotions for each cluster that will be catered to the need of each group of customers.

1. Students

a) 20% off on every alcoholic drink, money that will go to the card, and can be later spent on other products, on the first purchase. This exclusive offer is applicable on your first purchase of the month;
b) Spend 20€ on alcoholic drinks and receive 5€ to use in groceries in your next purchase;
c) Buy one bottle of wine and receive a complimentary dessert wine from the same brand;
d) 10% off on every ketchup and mayonnaise.

2. Loyal Grocery Shoppers

a) Enjoy a monthly discount of 20% on all fish and meat products, money that will go to the card, and can be later spent on other products. This exclusive offer is applicable on your first purchase of the month;
b) Spend 40€ on fish and meat and receive 5€ to use in groceries and 3€ to use in hygiene products in your next purchase;
c) Spend 60€ on store and receive a basket with a product from groceries, vegetables, hygiene, electronics, meat and fish, and drinks;
d) 20% Off on All Oils and eggs.

3. **Vegetarian Family**

a) 20% off on every vegetable, money that will go to the card, and can be later spent on other products. This exclusive offer is applicable on your first purchase of the month;
b) Spend 30€ on vegetables and receive 5€ to use in hygiene products in your next purchase;
c) For Every purchase of asparagus, enjoy a fantastic discount of 50% off on all tomatoes;
d) 10% on toothpaste.

4. **Tech Enthusiasts**

a) 20% off on every electronic and videogames product, money that will go to the card, and can be later spent on other products. This exclusive offer is applicable on your first purchase of the month;
b) Buy 3 get 4 products about *Pokemon* Collection - only available for electronic and videogame products - a free item of equal or lesser value;
c) Get all *Ratchet & Clank* videogames: buy 3 *Ratchet & Clank* video games and gain 15% at card to spend on non-alcoholic drinks;
d) 10% of *Apple* products.

5. **Family Shoppers**

a) 20% off on every groceries product, money that will go to the card, and can be later spent on other products. This exclusive offer is applicable on your first purchase of the month;
b) Spend 30€ on groceries and receive 5€ to use in hygiene products in your next purchase;
c) Buy 3 get 4 products about babies' food - free item of equal or lesser value;
d) 15% off on all gums and candy bars.

6. **Promotion Hunters**

a) 10% off on store, money that will go to the card, and can be later spent on other products. This exclusive offer is applicable on your first purchase of the month;
b) Spend 60€ store and receive a basket with a product from groceries, vegetables, hygiene, electronics, meat and fish, and drinks - worth 15€;
c) In the purchase of a cake, receive 5% to spend on fish and meat;
d) 15% off on all pet-food.

7. **Casual Shoppers**

a) 20% off on every electronic and videogames product, money that will go to the card, and can be later spent on other products. This exclusive offer is applicable on your first purchase of the month;
b) Spend 15€ on electronic and videogames and receive and receive 5€ to use on the store;
c) On the purchase of 3 bottles of champagne, we offer a wine of up to 5€;
d) Purchase 10€ on electronic and videogames and gain 15% at the card to spend on non-alcoholic drinks.

8. **Young Educated Moderate**

a) 20% off on every groceries product, money that will go to the card, and can be later spent on other products. This exclusive offer is applicable on your first purchase of the month;
b) On the purchase of 3 packs of French fries, we offer a sauce - ketchup or mayonnaise.
c) On the purchase of two cakes, we offer cooking oil or oil.
d) Buy 2 get 3 fromage blanc.

9. **Wholesale Meat and Seafood Distributors**

a) IVA off on store, money that will go to the card, and can be later spent on other products. This exclusive offer is applicable on your first purchase of the month;
b) 5% off on all fish and meat products;
c) For every purchase totaling over 1000€, we offer a complimentary family basket filled with essential items, including spaghetti, milk, vegetables, and oil.
d) Purchase cookies and yogurt cake and receive a 20% discount on milk.

Promotion a), on every cluster, will encourage the customer to create a loyalty card and return to the supermarket to spend the “money” that is on the card.

5. Conclusion

To conclude, we successfully managed to segment 30000 customers into comprehensive groups, providing valuable descriptions of each of them. Our approach involved a thorough exploratory data analysis, with inconsistencies checking, feature transformation, and looking for outliers, where we found our first cluster as well.

After attempting many models, we finally chose K-means as our final solution. From that, we obtained 8 more clusters. The distinctness of the clusters was further confirmed by the T-SNE and UMAP studies. Each market category displayed distinct traits in terms of preferences and buying habits. We learned more about each segment's requirements and motives through an in-depth investigation of the most common values for each cluster, including their spending habits and demographic characteristics, to help us better understand them.

Afterward, we developed targeted promotions for each group, tailored to each purchasing behavior. By doing so, we aimed to dissuade customers from seeking alternative offers elsewhere and encourage continued shopping with our stores.

The project was successful overall. We were able to find separate and homogeneous groups, enabling specialized marketing tactics. This project, with the help of K-Means, resulted in 9 clusters. We produced interesting findings on the tastes and behavior of the target market. We expect that the deployment of targeted promotions will raise customer satisfaction and customer retention rates, which will benefit the company's overall growth.

6. References

1. scikit-learn. (2023). sklearn.cluster.AgglomerativeClustering — scikit-learn 1.2.2 documentation. Retrieved June 5, 2023, from: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html

2. IBM. Clustering Binary Data: K-Means Should Be Avoided. Retrieved June 5, 2023, from https://www.ibm.com/support/pages/clustering-binary-data-k-means-should-be-avoided

3. Patidar, S. Clustering Algorithm for Data with Mixed Categorical and Numerical Features. Towards Data Science. Retrieved June 5, 2023, from https://towardsdatascience.com/clustering-algorithm-for-data-with-mixed-categorical-and-numerical-features-d4e3a48066a0

4. Bhandari, A. Silhouette Method: Better Than Elbow Method to Find Optimal Clusters. Towards Data Science. Retrieved June 5, 2023, from https://towardsdatascience.com/silhouette-method-better-than-elbow-method-to-find-optimal-clusters-378d62ff6891

5. scikit-learn. Silhouette Analysis of K-Means Clustering. Retrieved June 5, 2023, from https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

6. Central Statistics Office Ireland. (n.d.). Mean vs Median Information Note. Retrieved June 5, 2023, from https://www.cso.ie/en/releasesandpublications/in/rrppi/meanvsmedianinformationnote/

7. Annex

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 30000 entries, 4239 to 2570
Data columns (total 23 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   customer_name                         30000 non-null  object
 1   customer_gender                       30000 non-null  object
 2   customer_birthdate                    30000 non-null  object
 3   kids_home                             30000 non-null  float64
 4   teens_home                            30000 non-null  float64
 5   number_complaints                     30000 non-null  float64
 6   distinct_stores_visited               30000 non-null  float64
 7   lifetime_spend_groceries              30000 non-null  float64
 8   lifetime_spend_electronics            30000 non-null  float64
 9   typical_hour                          29998 non-null  float64
 10  lifetime_spend_vegetables             30000 non-null  float64
 11  lifetime_spend_nonalcohol_drinks      30000 non-null  float64
 12  lifetime_spend_alcohol_drinks         30000 non-null  float64
 13  lifetime_spend_meat                   30000 non-null  float64
 14  lifetime_spend_fish                   30000 non-null  float64
 15  lifetime_spend_hygiene                30000 non-null  float64
 16  lifetime_spend_videogames             29774 non-null  float64
 17  lifetime_total_distinct_products      30000 non-null  float64
 18  percentage_of_products_bought_promotion  30000 non-null  float64
 19  year_first_transaction                30000 non-null  float64
 20  loyalty_card_number                   5825 non-null   float64
 21  latitude                              30000 non-null  float64
 22  longitude                             30000 non-null  float64
dtypes: float64(20), object(3)
memory usage: 5.5+ MB
```

Fig. 1 - Dataset Information

	count	mean	std	min	25%	50%	75%	max
kids_home	30000.0	1.146700	1.234111	0.000000	0.000000	1.000000	1.000000	10.000000
teens_home	30000.0	0.512933	0.912502	0.000000	0.000000	0.000000	1.000000	9.000000
number_complaints	30000.0	0.764367	0.836341	0.000000	0.000000	1.000000	1.000000	9.000000
distinct_stores_visited	30000.0	6.120767	6.511410	0.000000	2.000000	3.000000	8.000000	21.000000
lifetime_spend_groceries	30000.0	3978.634251	5280.641466	0.000000	220.000000	950.000000	5253.000000	18852.000000
lifetime_spend_electronics	30000.0	965.238793	1755.323397	1.058527	45.000000	194.000000	721.000000	6603.000000
typical_hour	29998.0	14.357796	NaN	0.000000	9.000000	14.000000	20.000000	23.000000
lifetime_spend_vegetables	30000.0	564.207835	481.819086	1.000000	247.000000	425.000000	785.000000	1568.000000
lifetime_spend_nonalcohol_drinks	30000.0	628.750142	496.346400	0.000000	244.000000	425.000000	949.000000	1671.000000
lifetime_spend_alcohol_drinks	30000.0	379.344543	236.206400	0.000000	193.000000	419.000000	537.000000	1048.000000
lifetime_spend_meat	30000.0	513.878374	575.448347	0.000000	46.000000	164.000000	1078.000000	1860.000000
lifetime_spend_fish	30000.0	777.151833	3036.405862	0.000000	48.000000	166.000000	1091.000000	36243.000000
lifetime_spend_hygiene	30000.0	162.824538	168.547217	0.000000	50.000000	89.000000	223.000000	867.000000
lifetime_spend_videogames	29774.0	540.908544	NaN	0.000000	46.000000	95.000000	971.000000	2375.000000
lifetime_total_distinct_products	30000.0	1123.887300	1423.956421	0.000000	116.000000	427.000000	1980.000000	5818.000000
percentage_of_products_bought_promotion	30000.0	0.190981	0.145402	0.021920	0.099313	0.149777	0.201240	0.517190
year_first_transaction	30000.0	2008.684067	4.992296	1989.000000	2006.000000	2009.000000	2012.000000	2020.000000
loyalty_card_number	5825.0	949911.497854	28920.116430	900039.000000	924547.000000	950215.000000	974937.000000	999997.000000
latitude	30000.0	38.748224	0.024788	38.653348	38.731150	38.748200	38.765098	38.866765
longitude	30000.0	-9.157740	0.025749	-9.215240	-9.177770	-9.159280	-9.139674	-9.091217

Fig. 2 - Descriptive Statistics

```
age                                   3.205280e+02
kids_home                             1.524534e+00
teens_home                            8.369687e-01
number_complaints                     7.003078e-01
distinct_stores_visited               4.251974e+01
lifetime_spend_groceries              2.797594e+07
lifetime_spend_electronics            3.097440e+06
typical_hour                          3.035954e+01
lifetime_spend_vegetables             2.314954e+05
lifetime_spend_nonalcohol_drinks      2.452283e+05
lifetime_spend_alcohol_drinks         5.513200e+04
lifetime_spend_meat                   3.316552e+05
lifetime_spend_fish                   3.312487e+05
lifetime_spend_hygiene                2.842488e+04
lifetime_spend_videogames             5.116004e+05
lifetime_total_distinct_products      2.033399e+06
percentage_of_products_bought_promotion  2.130154e-02
year_first_transaction                2.503442e+01
latitude                              5.125046e-04
longitude                             6.515160e-04
loyalty_card                          1.568997e-01
Male                                  2.499967e-01
children_home                         3.376962e+00
lifetime_spend_total                  6.274600e+07
dtype: float64
```

Fig. 3 - Variance of each feature

```
cust_info1['Education'].value_counts()
```

```
No Higher Education    18568
Bsc.                    3815
Phd.                    3810
Msc.                    3807
Name: Education, dtype: int64
```

Fig. 4 - Creation of 'Education' feature

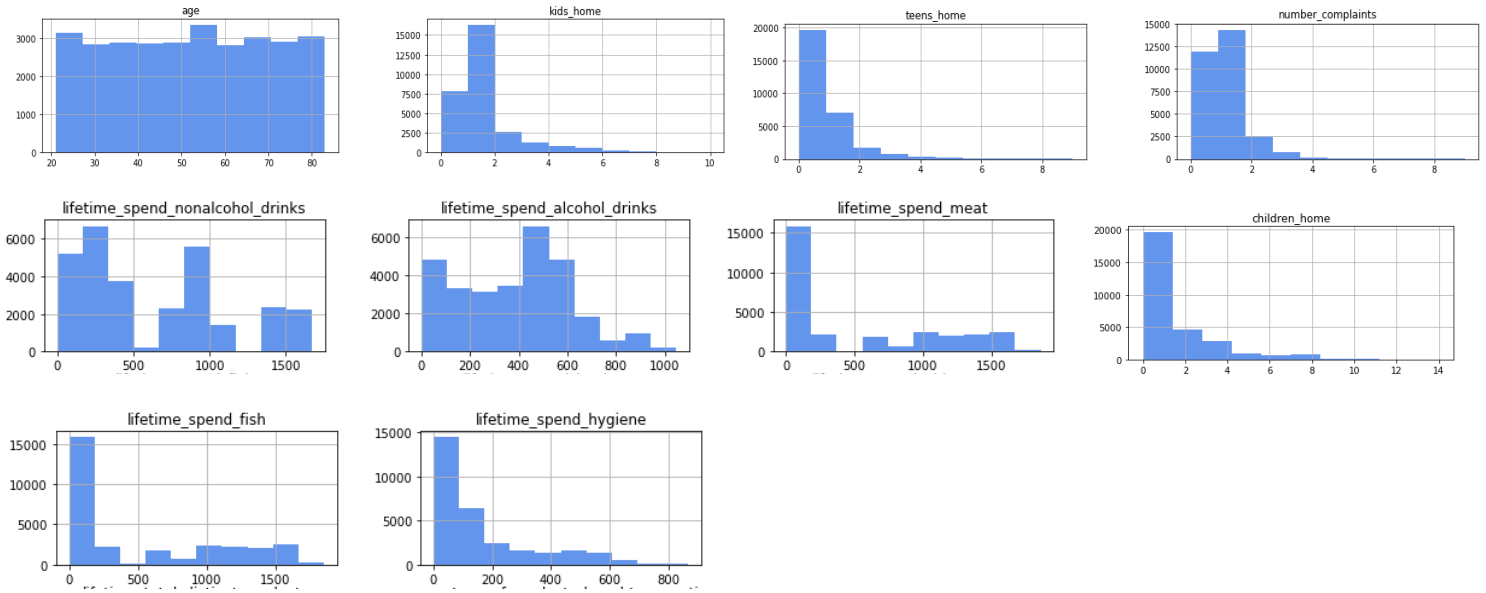

Fig. 5 - Histograms

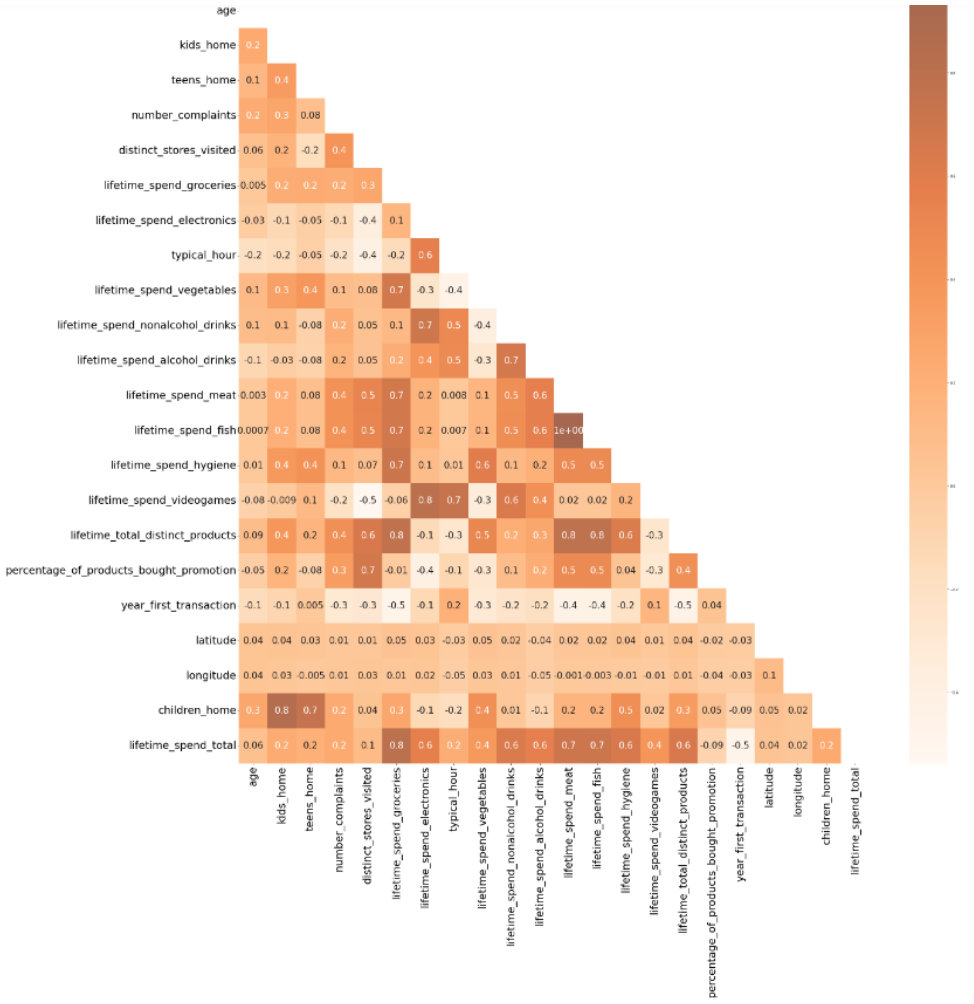

Fig. 6 - Spearman's Correlation Matrix

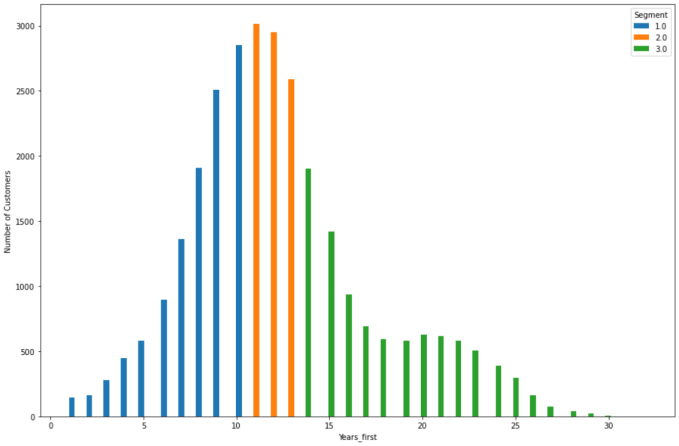

Fig. 7 - Number of customers per year since first transaction

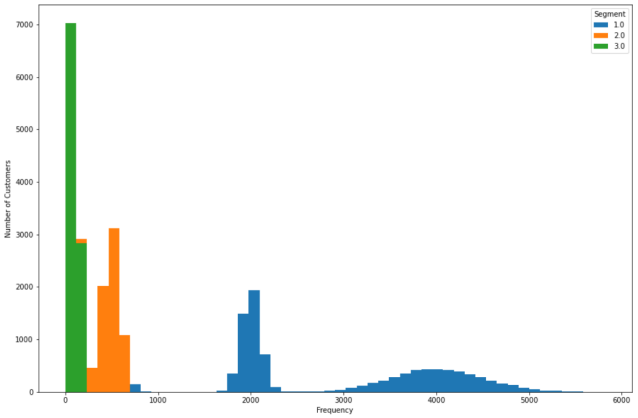

Fig. 8 - Number of customers per 'lifetime_total_distinct_products'

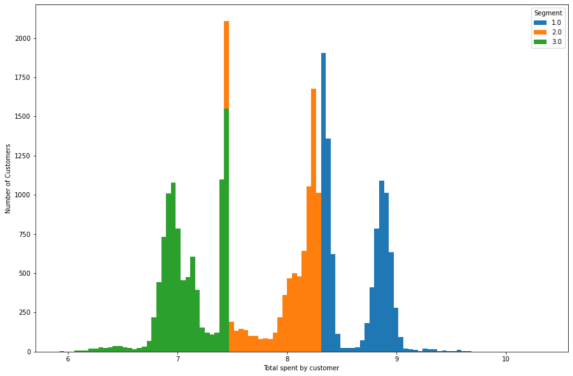

Fig. 9 - Number of customers per total amount spent

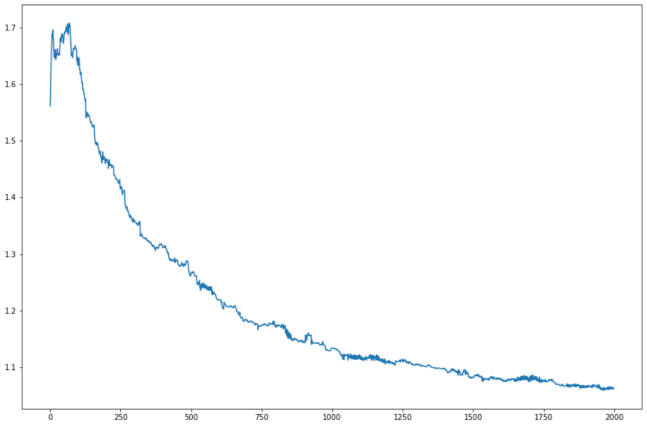

Fig. 10 - q-errors plot

winner_node	
(5, 0)	4921
(14, 4)	4610
(13, 13)	2288
(5, 6)	2198
(4, 7)	1518
(7, 3)	1335
(9, 2)	1248
(8, 3)	1066
(3, 8)	915
(12, 1)	904
(12, 8)	898
(11, 4)	843

Fig. 11 - Most populated nodes in SOM

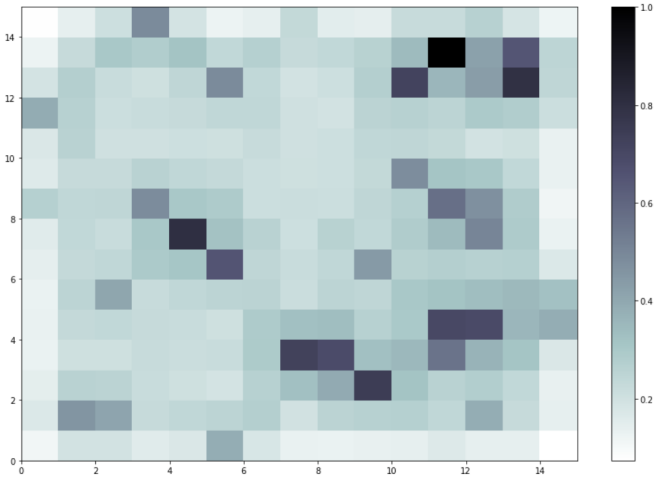

Fig. 12 - Distance Map

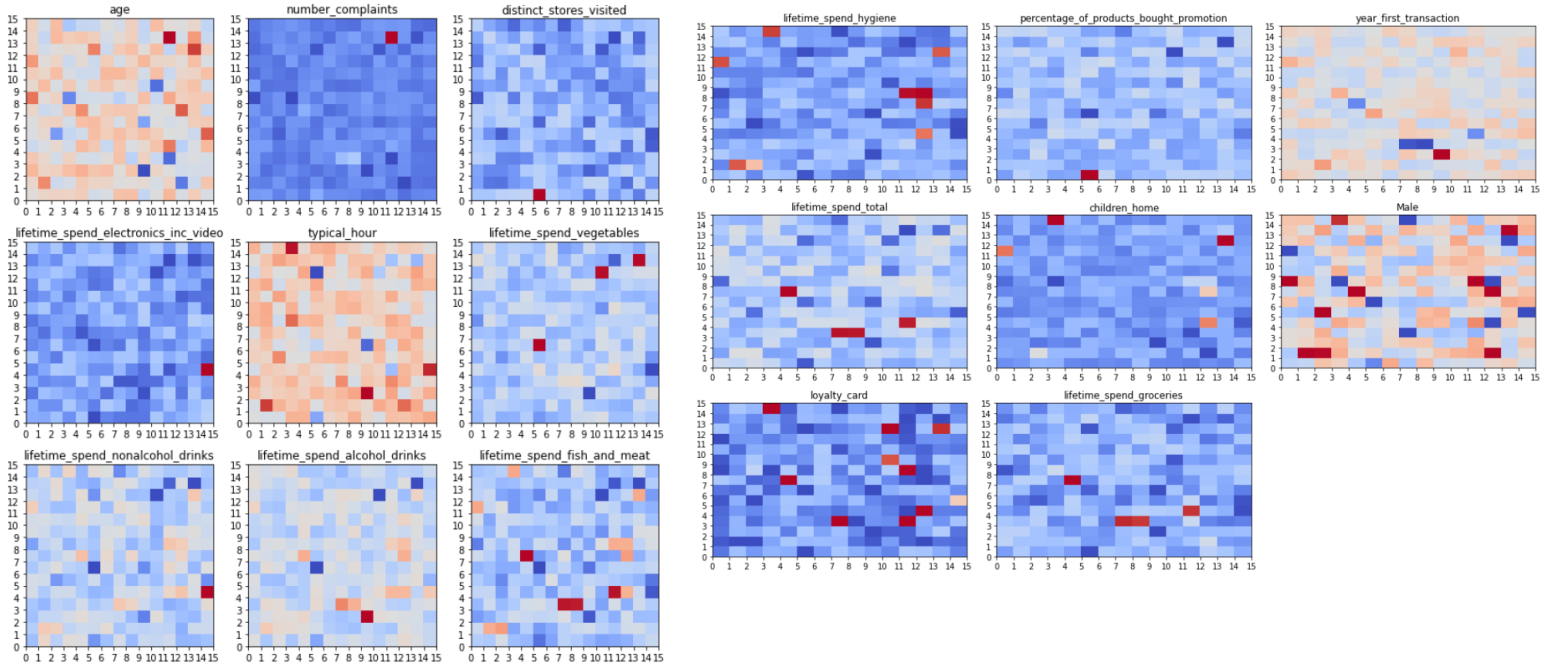

Fig. 13 - Feature Influence

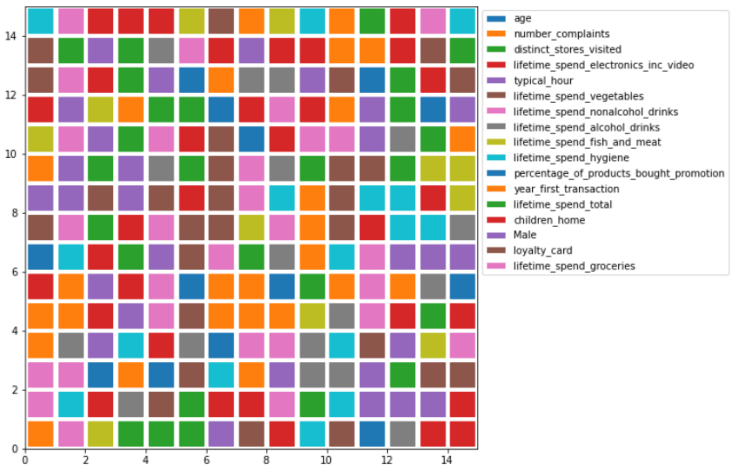

Fig. 14 - Most important variable per node

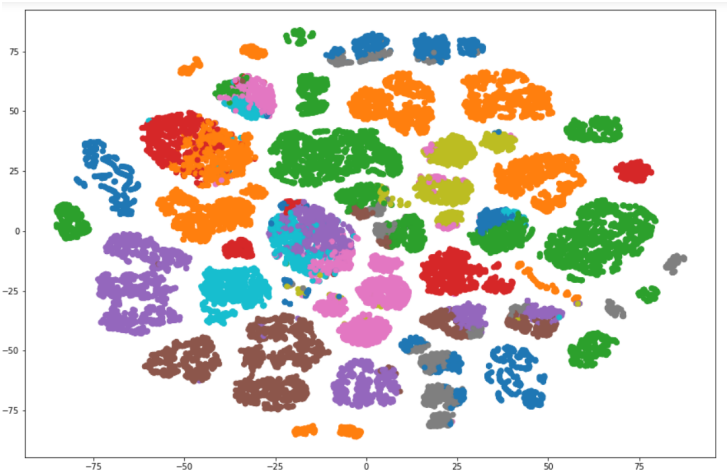

Fig. 15 - SOM's t-SNE plot

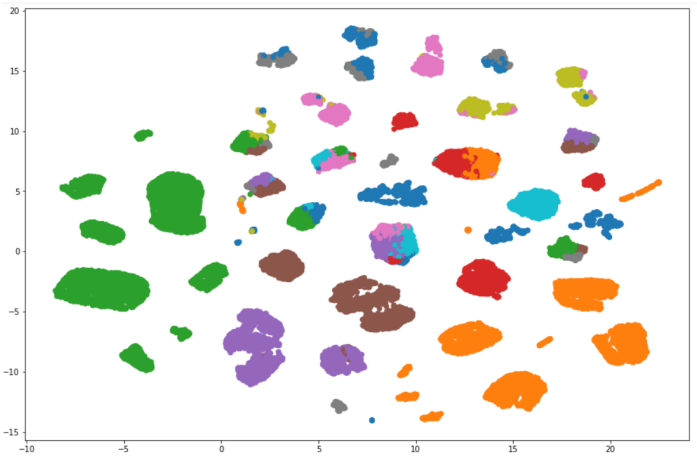

Fig. 16 - SOM's UMAP plot

```
cust_info1['dbscan_clustering']  = DBSCAN(eps=2, min_samples=700).fit_predict(scaled_cust)
```

```
cust_info1['dbscan_clustering'].value_counts()
```

```
-1     9491
 7     2389
 5     2289
 1     2286
 2     2197
 0     2112
 3     2037
 8     1541
 9     1473
 11    1116
 6     1020
 4     1009
 10     814
Name: dbscan_clustering, dtype: int64
```

Fig. 17 - DBSCAN Algorithm and resulting clusters

	Cluster -1	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10	Cluster 11	Supermarket Cluster	Original
age	50.0	56.0	55.0	56.0	56.0	29.0	55.0	29.0	55.0	56.0	56.0	54.0	58.0	52.5	52.0
number_complaints	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
distinct_stores_visited	3.0	2.0	20.0	2.0	2.0	3.0	2.0	3.0	20.0	3.0	3.0	8.0	8.0	3.0	3.0
children_home	1.0	1.0	1.0	2.0	1.0	0.0	2.0	0.0	1.0	5.0	5.0	1.0	1.0	1.0	1.0
typical_hour	16.0	21.0	9.0	10.0	21.0	19.0	10.0	19.0	9.0	18.0	18.0	13.0	12.0	15.0	14.0
year_first_transaction	2009.0	2010.0	2010.0	2010.0	2010.0	2010.0	2010.0	2010.0	2010.0	2010.0	2010.0	2000.0	2000.0	2010.0	2009.0
customer_name	James Grant	Donald Brown	Mary Smith	Gloria Campbell	Mary Williams	Dorothy Johnson	Christopher Smith	David Howard	Brian Anderson	Anne Jones	Peter Jackson	Noemi Martin	David Jackson	Albert Hurt	Robert Smith
Education	No Higher Education	No Higher Education	No Higher Education	No Higher Education	No Higher Education	No Higher Education	No Higher Education	No Higher Education	No Higher Education	No Higher Education	No Higher Education	No Higher Education	No Higher Education	No Higher Education	No Higher Education
Bsc.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Msc.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Phd.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
loyalty_card	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0
Male	0	1	0	0	0	0	1	1	1	0	1	0	1	1	1
lifetime_spend_electronics_inc_video	884.701928	7001.517045	25.0	99.837961	7004.701031	1462.297324	100.121887	1442.809804	25.0	1197.124594	1201.45757	250.22973	249.734767	1545.607143	1513.465272
lifetime_spend_vegetables	477.663155	20.060133	299.934383	1499.632681	20.008837	399.919722	1499.970293	397.762745	300.956886	605.131733	602.803123	806.886978	797.390681	526.459821	568.47434
lifetime_spend_nonalcohol_drinks	617.672321	1501.721117	301.380577	20.11015	1500.32597	404.744301	19.934906	400.210784	300.295521	899.792343	906.107943	899.701474	900.937276	659.022321	633.505206
lifetime_spend_alcohol_drinks	506.839005	498.204072	198.956255	20.109695	501.481591	400.374628	20.09917	396.491176	200.545835	499.537313	500.038018	604.153563	595.844086	411.053571	382.203433
lifetime_spend_fish_and_meat	1453.896639	116.589962	301.106737	0.0	116.494845	1341.921705	0.0	1318.606863	300.331101	2200.538611	2207.801765	2996.552826	3002.901434	1058.834821	1035.158998
lifetime_spend_hygiene	180.038984	50.221117	50.095801	100.198452	49.774669	191.888008	99.918305	191.253922	50.036835	501.094095	506.534963	195.291155	196.503584	160.473214	164.048667
percentage_of_products_bought_promotion	0.159933	0.100016	0.499986	0.049979	0.099972	0.147247	0.049979	0.14587	0.499962	0.200162	0.199937	0.150026	0.149981	0.194808	0.190912
lifetime_spend_total	11584.173638	9488.78125	1975.311899	2938.109695	9492.655376	11195.733399	2937.304063	10997.959804	1983.26036	12892.317975	12926.921928	24852.087224	24733.554659	9534.424107	9438.080305
lifetime_spend_groceries	5852.720367	200.362216	300.771654	998.502959	200.626411	6491.700694	997.479249	6371.326471	305.486396	4990.699546	4998.979633	15070.85258	14976.144265	3992.919643	4008.813629

Fig. 18 - DBSCAN: Mean (or median for categorical or mode for non numerical) of every feature per cluster

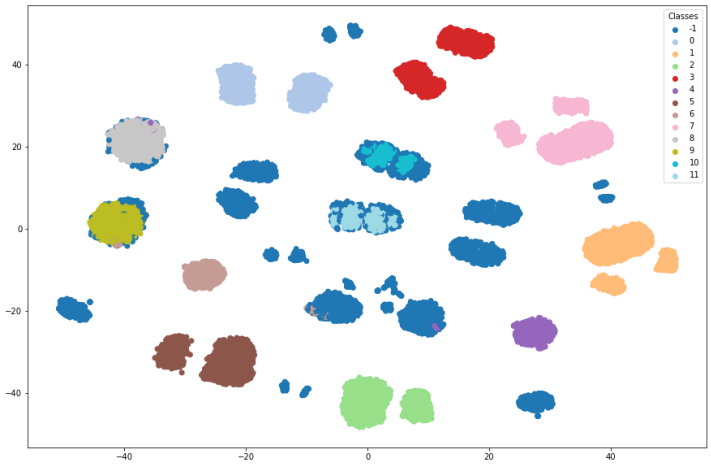

Fig. 19 - DBSCAN: t-SNE

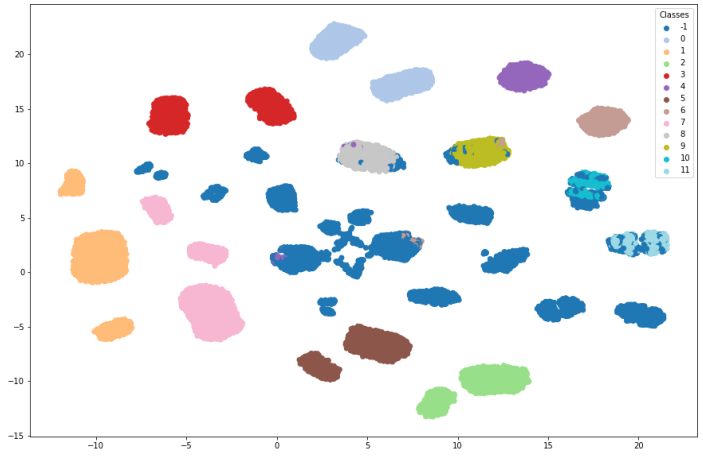

Fig. 20 - DBSCAN: UMAP

```
ms = MeanShift(bandwidth = 3, bin_seeding=True)
ms.fit(data_temp_k)
```

```
MeanShift(bandwidth=3, bin_seeding=True)
```

```
cust_info1['meanshift_clustering'] = ms.predict(data_temp_k)
```

```
cust_info1.meanshift_clustering.value_counts()
```

```
5    5625
0    4921
1    4921
3    4764
4    4656
2    4610
6     277
Name: meanshift_clustering, dtype: int64
```

Fig. 21 - Mean Shift Algorithm and resulting clusters

	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Supermarket Cluster	Original
age	55.0	56.0	56.0	56.0	56.0	29.0	71.0	50.0	52.0
number_complaints	1.0	0.0	0.5	2.0	1.0	0.0	4.0	0.0	1.0
distinct_stores_visited	20.0	2.0	2.0	8.0	3.0	2.0	1.0	1.0	3.0
children_home	1.0	2.0	1.0	1.0	5.0	0.0	1.0	0.0	1.0
typical_hour	9.0	10.0	21.0	12.0	18.0	19.0	10.0	0.693147	14.0
year_first_transaction	2010.0	2010.0	2010.0	2000.0	2010.0	2011.0	2010.0	2010.0	2009.0
customer_name	Mary Smith	Christopher Smith	Dana Davis	David Jackson	Anne Jones	Dorothy Johnson	Alberta Davis	James`s Supermarket	Robert Smith
Education	No Higher Education	No Higher Education	No Higher Education	No Higher Education	No Higher Education	No Higher Education	No Higher Education	No Higher Education	No Higher Education
Bsc.	0	0	0	0	0	0	0	0	0
Msc.	0	0	0	0	0	0	0	0	0
Phd.	0	0	0	0	0	0	0	0	0
loyalty_card	0	0	0	1	0	0	0	0	0
Male	1	1	1	1	0	0	1	0	1
lifetime_spend_electronics_inc_video	25.0	125.387116	7001.754881	249.983207	1197.640249	937.010844	21.483755	1.149498	1513.465272
lifetime_spend_vegetables	300.231863	1452.689494	20.033623	799.560034	602.228522	278.393244	101.920578	2.12416	568.47434
lifetime_spend_nonalcohol_drinks	300.431823	30.53546	1500.946855	900.786104	900.842569	320.031289	101.389892	2.302082	633.505206
lifetime_spend_alcohol_drinks	199.776671	29.22028	499.866161	599.145886	500.949098	476.366578	296.519856	2.704784	382.203433
lifetime_spend_fish_and_meat	300.573664	15.559033	116.686334	2999.797859	2202.147122	738.572267	102.916968	35000.363755	1035.158998
lifetime_spend_hygiene	50.032514	98.013209	49.912148	199.985306	500.98067	111.405511	49.837545	1.553669	164.048667
rcentage_of_products_bought_promotion	0.499966	0.052218	0.100002	0.149955	0.200116	0.143058	0.1988	0.200151	0.190912
lifetime_spend_total	1978.581792	2927.123146	9489.39154	24754.028967	12903.376074	6194.784	976.032491	35013.877321	9438.080305
lifetime_spend_groceries	302.959764	981.638895	200.406291	14993.307935	4997.198024	3086.584356	201.891697	2.701444	4008.813629

Fig. 22 - Mean Shift: Mean (or median for categorical features, or mode for non-numerical ones) for each cluster

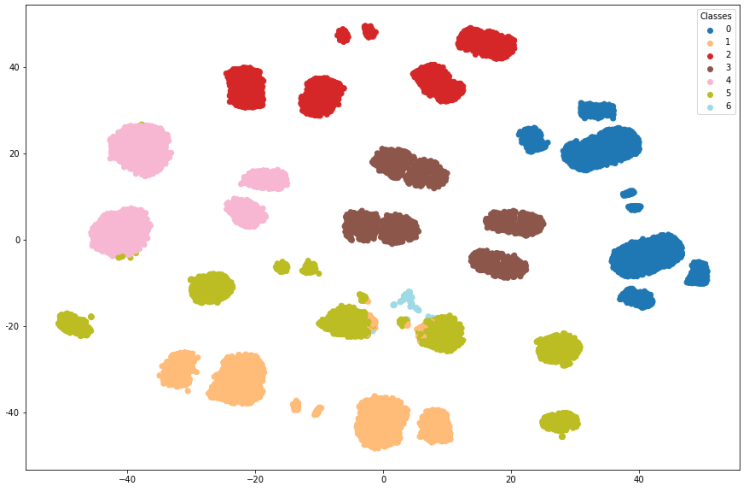

Fig. 23 - Mean Shift: t-SNE

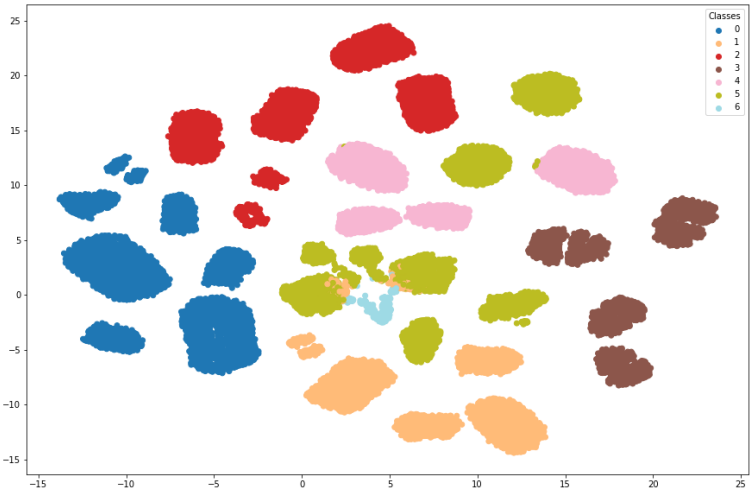

Fig.24 - Mean Shift: UMAP

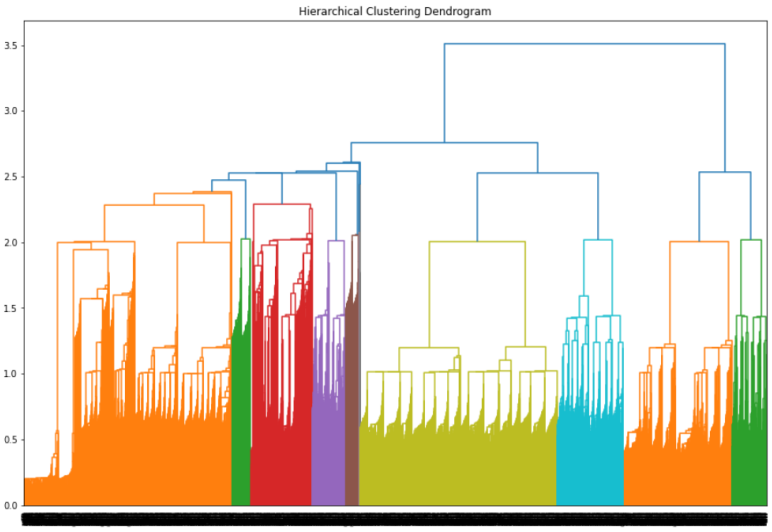

Fig. 25 - Hierarchical Clustering Dendogram: Single Linkage

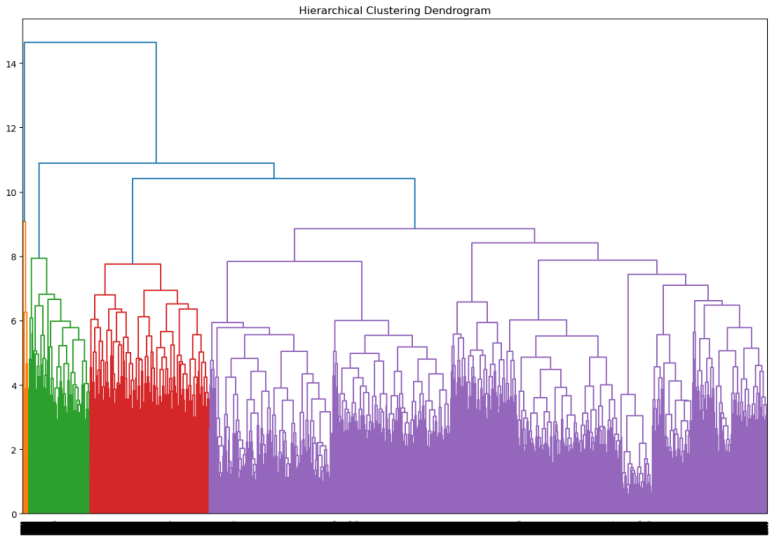

Fig. 26 - Hierarchical Clustering Dendogram: Complete Linkage

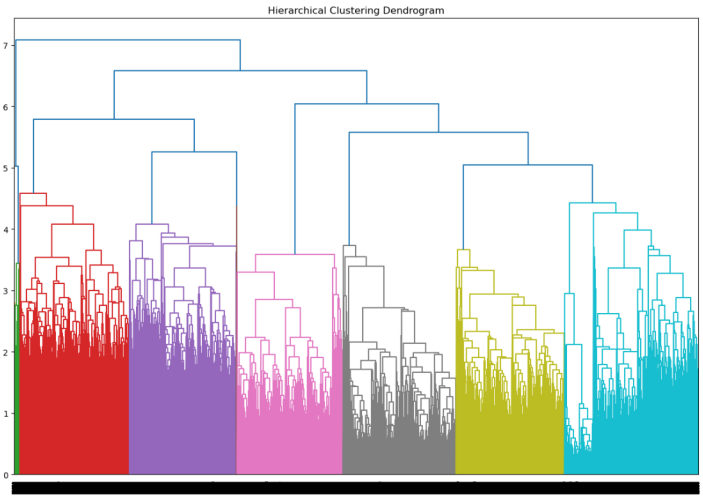

Fig. 27 - Hierarchical Clustering Dendogram: Average Linkage

```
cust_info1['cluster_single'] = AgglomerativeClustering(
    linkage='single', n_clusters=7
    ).fit_predict(data_temp_k)
```

```
cust_info1['cluster_single'].value_counts()
```

```
2    17380
6     4675
0     4610
3     2858
5      246
1        3
4        2
Name: cluster_single, dtype: int64
```

Fig.28 - Hierarchical Single Linkage: Cluster Size

```
cust_info1['cluster_complete'] = AgglomerativeClustering(
    linkage='complete', n_clusters=5
    ).fit_predict(data_temp_k)
```

```
cust_info1['cluster_complete'].value_counts()
```

```
0    23671
1     4764
2     1086
3      212
4       41
Name: cluster_complete, dtype: int64
```

Fig.29 - Hierarchical Single Linkage: Cluster Size

```
cust_info1['cluster_ward'] = AgglomerativeClustering(
    linkage='ward', n_clusters=8
    ).fit_predict(data_temp_k)
```

```
cust_info1['cluster_ward'].value_counts()
```

```
3    4921
2    4764
4    4722
1    4667
5    4610
0    2570
6    2272
7    1248
Name: cluster_ward, dtype: int64
```

Fig.30 - Hierarchical Ward Linkage: Cluster Size

```
cust_info1['cluster_average'] = AgglomerativeClustering(
    linkage='average', n_clusters=7
    ).fit_predict(data_temp_k)
```

```
cust_info1['cluster_average'].value_counts()
```

```
0    10538
5     4921
4     4764
6     4645
3     4610
1      275
2       21
Name: cluster_average, dtype: int64
```

Fig.31 - Hierarchical Single Linkage: Cluster Size

	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Supermarket Cluster	Original
customer_name	Elizabeth Perez	Anne Jones	David Jackson	Mary Smith	Christopher Smith	Dana Davis	Dorothy Johnson	James Hernandez	James`s Supermarket	Robert Smith
Education	No Higher Education	No Higher Education	No Higher Education	No Higher Education	No Higher Education	No Higher Education	Bsc.	No Higher Education	No Higher Education	No Higher Education
loyalty_card	0	0	1	0	0	0	0	0	0	0
Male	1	0	1	1	1	1	0	1	0	1
age	56.0	56.0	56.0	55.0	55.0	56.0	29.0	22.0	50.0	52.0
number_complaints	1.0	1.0	2.0	1.0	0.0	0.5	0.0	0.0	0.0	1.0
distinct_stores_visited	2.0	3.0	8.0	20.0	2.0	2.0	3.0	2.0	1.0	3.0
children_home	1.0	5.0	1.0	1.0	2.0	1.0	0.0	0.0	0.0	1.0
typical_hour	14.0	18.0	12.0	9.0	10.0	21.0	19.0	22.0	0.693147	14.0
year_first_transaction	2010.0	2010.0	2000.0	2010.0	2010.0	2010.0	2010.0	2017.0	2010.0	2009.0
lifetime_spend_electronics_inc_video	675.0	1199.0	250.0	25.0	100.0	7012.0	1504.5	120.0	1.151293	258.0
lifetime_spend_vegetables	287.0	603.0	799.5	300.0	1499.0	20.0	403.0	20.0	2.126912	429.0
lifetime_spend_nonalcohol_drinks	284.0	902.0	900.0	301.0	20.0	1501.0	401.0	200.0	2.303334	431.0
lifetime_spend_alcohol_drinks	298.0	502.0	598.0	199.0	20.0	500.0	402.0	901.0	2.704669	422.0
lifetime_spend_fish_and_meat	371.0	2203.0	2999.5	301.0	0.0	113.0	1398.0	100.0	35013.612777	326.0
lifetime_spend_hygiene	50.0	501.0	199.0	50.0	100.0	50.0	199.0	50.0	1.553652	90.0
lifetime_spend_groceries	481.0	5000.0	14983.0	301.0	998.0	200.0	7007.0	100.0	2.70714	961.0
percentage_of_products_bought_promotion	0.103674	0.200023	0.149969	0.499987	0.049996	0.099991	0.149822	0.199792	0.200166	0.149596
lifetime_spend_total	2679.0	12907.0	24756.0	1975.0	2940.0	9490.0	11821.0	1567.0	35027.082672	9230.0

Fig. 32 - Hierarchical WardLinkage: Most common values for each feature, per cluster, compared to the orignial

	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Supermarket Cluster	Original
customer_name	Elizabeth Perez	Anne Jones	David Jackson	Mary Smith	Christopher Smith	Dana Davis	Dorothy Johnson	James Hernandez	James`s Supermarket	Robert Smith
Education	No Higher Education	No Higher Education	No Higher Education	No Higher Education	No Higher Education	No Higher Education	Bsc.	No Higher Education	No Higher Education	No Higher Education
loyalty_card	0	0	1	0	0	0	0	0	0	0
Male	1	0	1	1	1	1	0	1	0	1
age	56.0	56.0	56.0	55.0	55.0	56.0	29.0	22.0	50.0	52.0
number_complaints	1.0	1.0	2.0	1.0	0.0	0.5	0.0	0.0	0.0	1.0
distinct_stores_visited	2.0	3.0	8.0	20.0	2.0	2.0	3.0	2.0	1.0	3.0
children_home	1.0	5.0	1.0	1.0	2.0	1.0	0.0	0.0	0.0	1.0
typical_hour	14.0	18.0	12.0	9.0	10.0	21.0	19.0	22.0	0.693147	14.0
year_first_transaction	2010.0	2010.0	2000.0	2010.0	2010.0	2010.0	2010.0	2017.0	2010.0	2009.0
lifetime_spend_electronics_inc_video	675.0	1199.0	250.0	25.0	100.0	7012.0	1504.5	120.0	1.151293	258.0
lifetime_spend_vegetables	287.0	603.0	799.5	300.0	1499.0	20.0	403.0	20.0	2.126912	429.0
lifetime_spend_nonalcohol_drinks	284.0	902.0	900.0	301.0	20.0	1501.0	401.0	200.0	2.303334	431.0
lifetime_spend_alcohol_drinks	298.0	502.0	598.0	199.0	20.0	500.0	402.0	901.0	2.704669	422.0
lifetime_spend_fish_and_meat	371.0	2203.0	2999.5	301.0	0.0	113.0	1398.0	100.0	35013.612777	326.0
lifetime_spend_hygiene	50.0	501.0	199.0	50.0	100.0	50.0	199.0	50.0	1.553652	90.0
lifetime_spend_groceries	481.0	5000.0	14983.0	301.0	998.0	200.0	7007.0	100.0	2.70714	961.0
percentage_of_products_bought_promotion	0.103674	0.200023	0.149969	0.499987	0.049996	0.099991	0.149822	0.199792	0.200166	0.149596
lifetime_spend_total	2679.0	12907.0	24756.0	1975.0	2940.0	9490.0	11821.0	1567.0	35027.082672	9230.0

Fig. 33 - Hierarchical Single Linkage: Most common/central values for each feature, per cluster, compared to the orignial

	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Supermarket Cluster	Original
customer_name	Dorothy Johnson	Alberta Davis	Casey Gregg	Dana Davis	David Jackson	Mary Smith	Anne Jones	James`s Supermarket	Robert Smith
Education	No Higher Education	No Higher Education	Msc.	No Higher Education	No Higher Education	No Higher Education	No Higher Education	No Higher Education	No Higher Education
loyalty_card	0	0	1	0	1	0	0	0	0
Male	1	1	0	1	1	1	1	0	1
age	40.0	71.0	41.0	56.0	56.0	55.0	56.0	50.0	52.0
number_complaints	0.0	4.0	1.0	0.5	2.0	1.0	1.0	0.0	1.0
distinct_stores_visited	2.0	1.0	3.0	2.0	8.0	20.0	3.0	1.0	3.0
children_home	1.0	1.0	12.0	1.0	1.0	1.0	5.0	0.0	1.0
typical_hour	13.0	10.0	17.0	21.0	12.0	9.0	18.0	0.693147	14.0
year_first_transaction	2010.0	2010.0	2010.0	2010.0	2000.0	2010.0	2010.0	2010.0	2009.0
lifetime_spend_electronics_inc_video	557.652591	17.669091	1193.380952	7001.754881	249.983207	25.0	1197.721636	1.149498	1513.465272
lifetime_spend_vegetables	826.471057	102.254545	603.666667	20.033623	799.560034	300.231863	602.089559	2.12416	568.47434
lifetime_spend_nonalcohol_drinks	184.273771	101.378182	885.142857	1500.946855	900.786104	300.431823	900.862217	2.302082	633.505206
lifetime_spend_alcohol_drinks	267.497343	297.189091	552.428571	499.866161	599.145886	199.776671	500.687836	2.704784	382.203433
lifetime_spend_fish_and_meat	399.46375	101.723636	2250.47619	116.686334	2999.797859	300.573664	2201.92831	35000.363755	1035.158998
lifetime_spend_hygiene	104.872556	49.807273	503.714286	49.912148	199.985306	50.032514	500.737998	1.553669	164.048667
percentage_of_products_bought_promotion	0.100579	0.199317	0.199636	0.100002	0.149955	0.499966	0.200121	0.200151	0.190912
lifetime_spend_total	4661.871323	970.770909	13001.142857	9489.39154	24754.028967	1978.581792	12902.417008	35013.877321	9438.080305
lifetime_spend_groceries	2101.349402	200.345455	5005.47619	200.406291	14993.307935	302.959764	4997.057051	2.701444	4008.813629

Fig. 34 - Hierarchical Average Linkage: Most common/central values for each feature, per cluster, compared to the orignial

	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Supermarket Cluster	Original
customer_name	Robert Smith	David Jackson	Michael White	Angela Veno	Alberta Davis	James`s Supermarket	Robert Smith
Education	No Higher Education	No Higher Education	Msc.	No Higher Education	No Higher Education	No Higher Education	No Higher Education
loyalty_card	0	1	1	0	0	0	0
Male	1	1	1	0	1	0	1
age	50.0	56.0	58.0	70.0	71.0	50.0	52.0
number_complaints	1.0	2.0	1.0	4.0	8.0	0.0	1.0
distinct_stores_visited	2.0	8.0	3.0	1.0	1.0	1.0	3.0
children_home	1.0	1.0	7.0	1.0	1.0	0.0	1.0
typical_hour	16.0	12.0	18.0	10.0	10.0	0.693147	14.0
year_first_transaction	2010.0	2000.0	2010.0	2010.0	2009.0	2010.0	2009.0
lifetime_spend_electronics_inc_video	1798.396899	249.983207	1195.791897	9.891509	10.121951	1.149498	1513.465272
lifetime_spend_vegetables	525.500951	799.560034	600.544199	100.09434	100.170732	2.12416	568.47434
lifetime_spend_nonalcohol_drinks	573.122302	900.786104	901.487109	99.783019	99.804878	2.302082	633.505206
lifetime_spend_alcohol_drinks	333.962781	599.145886	501.418048	299.160377	297.536585	2.704784	382.203433
lifetime_spend_fish_and_meat	596.213088	2999.797859	2202.06814	99.966981	102.146341	35000.363755	1035.158998
lifetime_spend_hygiene	142.457902	199.985306	503.660221	49.726415	49.195122	1.553669	164.048667
percentage_of_products_bought_promotion	0.198627	0.149955	0.200301	0.199991	0.199825	0.200151	0.190912
lifetime_spend_total	6287.122555	24754.028967	12906.122468	961.240566	948.756098	35013.877321	9438.080305
lifetime_spend_groceries	1793.342867	14993.307935	4999.489871	202.099057	189.95122	2.701444	4008.813629

Fig. 35 - Hierarchical Complete Linkage: Most common values for each feature, per cluster, compared to the orignial

```
For n_clusters = 2 The average silhouette_score is: 0.27174043589349856
For n_clusters = 3 The average silhouette_score is: 0.319451313075198
For n_clusters = 4 The average silhouette_score is: 0.39080125644335184
For n_clusters = 5 The average silhouette_score is: 0.4306316061450293
For n_clusters = 6 The average silhouette_score is: 0.4714952823352651
For n_clusters = 7 The average silhouette_score is: 0.4707198677687455
For n_clusters = 8 The average silhouette_score is: 0.4793640624471343
For n_clusters = 9 The average silhouette_score is: 0.45063168522880687
For n_clusters = 10 The average silhouette_score is: 0.4288637955117489
For n_clusters = 11 The average silhouette_score is: 0.4150507286182722
For n_clusters = 12 The average silhouette_score is: 0.3763264991164983
For n_clusters = 13 The average silhouette_score is: 0.32019384092335607
For n_clusters = 14 The average silhouette_score is: 0.2708460447079091
```

Fig. 36 - Silhouette Scores

```
filtered_data1 = rules_cluster1[(rules_cluster1['lift'] > 1.25)]
filtered_data1
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
419	(cider, whole weat flour)	(dessert wine)	0.044164	0.415002	0.023835	0.539683	1.300434	0.005506	1.270858
607	(dessert wine, gums)	(cider, white wine)	0.048370	0.439187	0.026639	0.550725	1.253964	0.005395	1.248262
610	(ketchup, dessert wine)	(cider, white wine)	0.036453	0.439187	0.020329	0.557692	1.269829	0.004320	1.267925

Fig. 37 - Association Rules: Students

```
rules_cluster2 = generate_association_rules(df2)
```

```
filtered_data2 = rules_cluster2[(rules_cluster2['lift'] > 1.075) ]
filtered_data2
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
306	(eggs, cake)	(cooking oil)	0.031305	0.656398	0.022303	0.712460	1.085408	0.001755	1.194970
541	(cotton buds)	(cooking oil, oil)	0.034355	0.576086	0.021303	0.620087	1.076379	0.001512	1.115819
882	(eggs, cooking oil, oil)	(cake)	0.037156	0.508376	0.020353	0.547779	1.077508	0.001464	1.087132
883	(eggs, cake, oil)	(cooking oil)	0.028304	0.656398	0.020353	0.719081	1.095495	0.001774	1.223135
884	(eggs, cake)	(cooking oil, oil)	0.031305	0.576086	0.020353	0.650160	1.128580	0.002319	1.211735

Fig. 38 - Association Rules: Loyal Grocery Shoppers

```
rules_cluster3 = generate_association_rules(df3)
```

```
filtered_data3 = rules_cluster3[(rules_cluster3['lift'] >1.1)]
filtered_data3
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
1310	(frozen vegetables, cologne)	(asparagus, tomatoes)	0.030062	0.656109	0.022010	0.732143	1.115886	0.002286	1.283859
1374	(toothpaste, green beans)	(asparagus, tomatoes)	0.031029	0.656109	0.022761	0.733564	1.118052	0.002403	1.290707
1436	(toothpaste, tomatoes, zucchini)	(asparagus)	0.025123	0.750805	0.020829	0.829060	1.104228	0.001966	1.457789
1437	(toothpaste, zucchini)	(asparagus, tomatoes)	0.028130	0.656109	0.020829	0.740458	1.128559	0.002373	1.324991

Fig. 39 - Association Rules: Vegetarian Family

```
rules_cluster4 = generate_association_rules(df4)
```

```
filtered_data4 = rules_cluster4[(rules_cluster4['lift'] > 1.11)]
filtered_data4
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
778	(ratchet & clank 2, ratchet & clank 3)	(pokemon violet)	0.046848	0.496060	0.026086	0.556818	1.122480	0.002846	1.137094
826	(airpods, pokemon sword, ratchet & clank 2)	(pokemon violet)	0.043441	0.496060	0.023957	0.551471	1.111700	0.002407	1.123537
873	(pokemon sword, ratchet & clank, bluetooth hea...	(pokemon violet)	0.038756	0.496060	0.021614	0.557692	1.124243	0.002389	1.139342
987	(pokemon violet, pokemon sword, iMac)	(pokemon shield)	0.044932	0.653003	0.032581	0.725118	1.110437	0.003240	1.262353
1239	(pokemon sword, ratchet & clank 2, ratchet & c...	(pokemon violet)	0.041844	0.496060	0.023637	0.564885	1.138743	0.002880	1.158177
1240	(ratchet & clank 2, ratchet & clank 3)	(pokemon violet, pokemon sword)	0.046848	0.438458	0.023637	0.504545	1.150726	0.003096	1.133387
1286	(pokemon violet, ratchet & clank, pokemon scar...	(pokemon sword, pokemon shield)	0.064310	0.574532	0.041525	0.645695	1.123864	0.004577	1.200855
1301	(ratchet & clank, white wine, pokemon scarlet)	(pokemon sword, pokemon shield)	0.043122	0.574532	0.028003	0.649383	1.130282	0.003228	1.213484

Fig. 40 - Association Rules: Tech Entusiasts

```
rules_cluster5 = generate_association_rules(df5)
```

```
filtered_data5 = rules_cluster5[(rules_cluster5['lift']>1.08)]
filtered_data5
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
150	(toilet paper)	(cake)	0.052816	0.514131	0.029519	0.558897	1.087072	0.002364	1.101488
190	(mint)	(cooking oil)	0.029916	0.652393	0.021444	0.716814	1.098747	0.001927	1.227489
649	(cookies, candy bars)	(cake)	0.040241	0.514131	0.022569	0.560855	1.090881	0.001880	1.106399
725	(cream, gums)	(cake)	0.036336	0.514131	0.020319	0.559199	1.087658	0.001638	1.102241
790	(french fries, oil)	(cooking oil)	0.030909	0.652393	0.022437	0.725910	1.112689	0.002272	1.268224
1044	(ketchup, candy bars)	(babies food, cooking oil)	0.075386	0.573036	0.047323	0.627744	1.095470	0.004124	1.146962
1120	(ketchup, french fries)	(babies food, cooking oil)	0.036270	0.573036	0.022834	0.629562	1.098643	0.002050	1.152593

Fig. 41 - Association Rules: Family Shoppers

```
rules_cluster6 = generate_association_rules(df6)
```

```
filtered_data6 = rules_cluster6[(rules_cluster6['lift'] >1.11)]
filtered_data6
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
10	(bacon)	(cake)	0.042191	0.503245	0.024239	0.574519	1.141628	0.003007	1.167514
257	(bacon, oil)	(cake)	0.035598	0.503245	0.020487	0.575499	1.143574	0.002572	1.170207
284	(candy bars, whole wheat rice)	(cake)	0.038540	0.503245	0.021602	0.560526	1.113823	0.002208	1.130340
349	(cooking oil, whole wheat rice)	(cake)	0.059838	0.503245	0.033570	0.561017	1.114798	0.003457	1.131603
963	(cooking oil, strong cheese)	(cake, oil)	0.048073	0.445943	0.024037	0.500000	1.121219	0.002599	1.108114
1217	(pet food, napkins)	(cooking oil, oil)	0.037728	0.582454	0.024544	0.650538	1.116890	0.002569	1.194823

Fig. 42 - Association Rules: Promotion Hunters

```
rules_cluster7 = generate_association_rules(df7)
```

```
filtered_data7 = rules_cluster7[(rules_cluster7['lift'] >1.25)]
filtered_data7
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
212	(beats headphones, bramble)	(dessert wine)	0.038368	0.474311	0.022933	0.597701	1.260146	0.004734	1.306713
848	(beer, phone car charger)	(dessert wine, white wine)	0.042778	0.418523	0.022712	0.530928	1.268576	0.004809	1.239633
950	(white wine, bramble, google tablet)	(dessert wine)	0.044101	0.474311	0.027343	0.620000	1.307159	0.006425	1.383392
951	(bramble, google tablet)	(dessert wine, white wine)	0.051819	0.418523	0.027343	0.527660	1.260767	0.005655	1.231056
974	(champagne, half-life: alyx)	(cider, white wine)	0.033076	0.544873	0.022712	0.686667	1.260232	0.004690	1.452532
1120	(beer, champagne, french wine)	(dessert wine, white wine)	0.037266	0.418523	0.020066	0.538462	1.286577	0.004470	1.259868

Fig. 43 - Association Rules: Casual Shoppers

```
rules_cluster8 = generate_association_rules(df8)
```

```
filtered_data8 = rules_cluster8[(rules_cluster8['lift'] >1.12)]
filtered_data8
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
841	(cooking oil, french fries, oil)	(cake)	0.112324	0.500117	0.063263	0.563218	1.126172	0.007088	1.144468
853	(fromage blanc, cake, oil)	(cooking oil)	0.027934	0.645188	0.020305	0.726891	1.126634	0.002282	1.299159
854	(fromage blanc, cake)	(cooking oil, oil)	0.031690	0.558803	0.020305	0.640741	1.146631	0.002597	1.228075
873	(mayonnaise, cooking oil, oil)	(cake)	0.040610	0.500117	0.022770	0.560694	1.121124	0.002460	1.137891
946	(napkins, french fries, oil)	(cake)	0.044601	0.500117	0.025235	0.565789	1.131313	0.002929	1.151245
1068	(ketchup, french fries, oil)	(cooking oil)	0.030282	0.645188	0.021948	0.724806	1.123403	0.002411	1.289318
1069	(ketchup, french fries)	(cooking oil, oil)	0.034624	0.558803	0.021948	0.633898	1.134386	0.002600	1.205122
1200	(cooking oil, gums, french fries, oil)	(cake)	0.037793	0.500117	0.021596	0.571429	1.142589	0.002695	1.166393

Fig. 44 - Association Rules: Young Educated Moderate

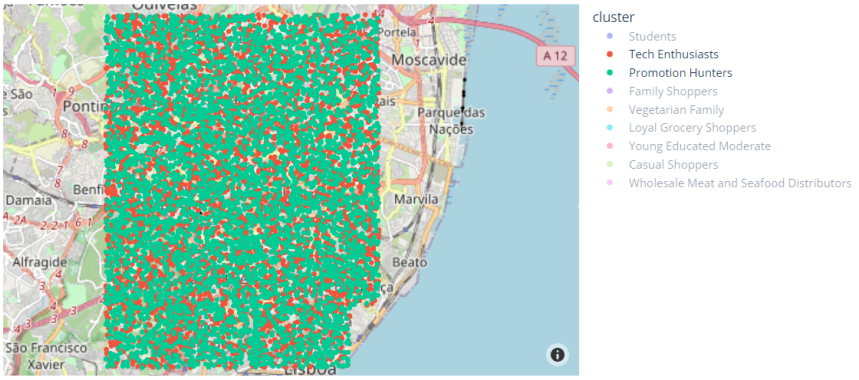

Fig. 45 - Tech Entusiasts and Promotion Hunters location

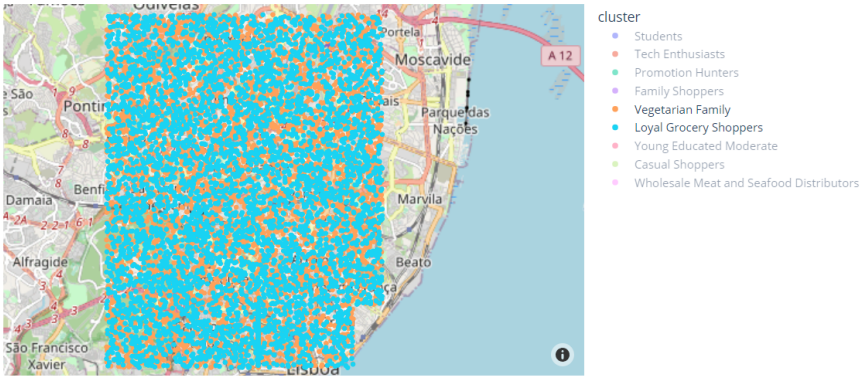

Fig. 46 - Vegetarian Family and Loyal Grocery Shoppers location