Data Science

Machine Learning II

Customer Segmentation

Authors:

Mariana Ferreira 20211637@novaims.unl.pt
Mariana Neto 20211637@novaims.unl.pt
Rui Lourenço 20211639@novaims.unl.pt

2022/2023 – 2nd Semester

Contents

1. Executive summary

Customer segmentation is a strategic approach to classify customers based on shared characteristics into distinct segments. Within the field of machine learning, the use of unsupervised learning algorithms offers powerful techniques for the analysis and clustering of customers, thereby facilitating the identification of hidden patterns.

For the sake of this project, implementing those techniques on the customer data provided, such as purchasing data, provides insights on customers' purchase patterns which allows us to identify targeted marketing strategies tailored to address the specific needs and preferences of the different customer segments.

2. Exploratory Data Analysis

Exploratory Data Analysis (EDA) is the critical initial step in data analysis, laying the foundation to further exploration and modeling. EDA encompasses a series of techniques and methods to explore the data, including descriptive statistics and visualization. In this project, EDA consisted in data collection, data cleaning, feature engineering, univariate and bivariate analysis through the use of visual plots.

2.1. Data Collection

The data assigned consisted in two datasets stored as csv files:

- customer_info: contains customer personal information, demographics and spending patterns, covering a two year period, which facilitate the characterization of the customer;
- customer_basket: contains information about customers' transactions, related to the last 6 months of data. This is essential for the definition of association rules.

2.2. Data Exploration

This section was conducted solely on the 'customer_info' dat set, stored as 'info' During the analysis of this data set, it was observed the existence of infinite values on the variables 'typical_hour' and 'lifetime_spend_videogames'. To ensure the integrity of the data and mitigate any distortions caused by these values, it was decided to treat them as missing values. This approach simplified the process of understanding and gathering information from the data.

Furthermore, it is important to note that missing values were observed in the variables 'loyalty_card_number', 'typical_hour' and 'lifetime_spend_videogames'. To address this issue, the 'loyalty_card_number' column was used to create a new variable that stored information on the ownership of loyalty cards, and was then removed from the data. Regarding the 'lifetime_spend_videogames' variable, the missing values were imputed by assigning a value of zero, assuming that these clients did not spend on video games. Lastly, in respect to the 'typical_hour' variable, it was carried in the Customer Segmentation step.

By means of descriptive statistics, data incoherences and inappropriate feature formats were noted. These situations were further considered in the next section.

It is also relevant to take into account that in clustering algorithms, highly correlated variables can artificially inflate the importance of certain features, leading to biased cluster assignments. Removing those variables can help to alleviate this issue and ensure a more accurate clustering. Therefore, a correlation matrix was plotted and analyzed on all numeric and non-binary features.

From the correlation matrix [Appendice 6.1], it was possible to identify that 'lifetime_spend_groceries', 'lifetime_spend_meat' have a correlation coefficient of 0.93 with ‘lifetime_total_distinct_products' features. This correlation was the one that was subject to more attention, indicating that 'lifetime_spend_groceries' can potentially account for the behavior of these variables. Specifically, it suggests that the expenses allocated to meat ('lifetime_spend_meat') might be encompassed within the category of grocery expenditures ('lifetime_spend_groceries'), or at least its behavior may be integrated with the second variable. Similarly, concerning the 'lifetime_total_distinct_products', the same can be said about its behavior being explained by grocery expenses.

Through experimentation, it was ultimately determined that excluding any of those highly correlated variables would yield inferior results in terms of clustering techniques, leading to less precise outcomes.

2.3. Feature Engineering

The comprehensive analysis of the data, including correlation patterns and visual insights, has enabled the formulation of new variables, contributing to a more refined understanding of the underlying relationships within the dataset.

1. Transforming year-related variables

To improve the analysis of customer demographics, two variables were created using the current year (2023) as a reference. The first variable, 'years_as_customer,' provides

valuable insights into customer loyalty and identifying long-term customers, understanding their behavior patterns will help to make personalized marketing strategies.

The second variable, 'customer_age,' is derived from the customer's date of birth and offers a straightforward measure of their age, enabling targeted marketing campaigns for different age groups.

2. Transforming 'customer_name' into 'education' and 'customer_name'

The feature 'column_name' includes both the customer's name and education level. Therefore, it is split into two different features: 'customer_name', which will solely contain the customer's name and the 'customer_education', which will carry their education level. In cases where the customer does not have an explicit education level, the value 'basic' will be assigned. This helps in targeting strategies and products based on educational needs.

3. Creating dummy variables

By performing the transformation of numerical variables into binary, it becomes easier to examine the impact of those variables on some variables of interest, such as expenses related to the different sections available. This will be useful to examine the clusters in the Segmentation step. Two dummies were created, ‘has_child’ and ‘has_loyalty_card’, previously mentioned.

These transformations allow for straightforward identification of customers with children and customers with loyalty cards, providing valuable information for further analysis and decision-making.

4. Creating variables through linear combination

Linear combination may be very useful to extract relevant variables from the ones included in the initial data set. In this project, the linear combinations performed are based on the addition of different variables. Firstly, the feature 'num_kids' is created through the sum of the features 'kids_home' and 'teens_home' for each customer. This is useful, since the parents of kids and teens share several characteristics, such as their prioritizing needs and possibly budgeting.

The second variable extracted from the initial data set is named 'total_spend'. This new variable is obtained through the sum of the values of each column whose name starts

with ‘lifetime_spend', meaning it includes the total amount of money each customer spent in the considered timeline. To create this variable, we consider all the variables that start with 'lifetime_spend' as not being included in one another.

The creation of the 'total_spend' variable provides a comprehensive measure of customer spending behavior and serves as a useful metric for analyzing overall customer expenditure within the specified timeframe.

2.4. Feature Selection

To ensure the final 'info' data set contains only relevant variables, some of the initial variables will be excluded based on the insights gained during data exploration. The specific variables that will be discarded are those that were not in the most desirable format, and were, eventually, used to create new features in the 'Features Extraction' section, that were more useful. These variables are: 'customer_birthdate', 'loyalty_card_number', 'kids_home', 'teens_home' and 'year_first_transaction'.

2.6 Data Visualization

Data visualization is an essential aspect of data analysis that involves representing information and insights visually. Data visualization is valuable for presenting findings, such as patterns and trends that may influence decision-making processes.

A notable finding from the application of visual representations, specifically the Scatter Mapbox, is that all customer residences are concentrated within the metropolitan area of Lisbon. This finding provides valuable insight into the geographical distribution of customers and highlights the predominant location of their residences.

1. univariate analysis [Appendice 6.2]

In the analysis, the majority of customers have a high school education or lower (61.89%). The education levels of Bachelor's, Master's, and PhD are, almost, uniformly distributed (around 14% each). The customer age distribution challenges the expectation

of age group concentration. Most customers have up to three children, with a notable proportion having multiple children.

2. Customer Segmentation

Customer segmentation can be effectively achieved through clustering. The method gathers customer data and identifies distinct clusters that can be derived from. By analyzing the distinguishing characteristics of these clusters, valuable insights regarding customer segments can be derived.

The data set used consisted of all variables from the EDA stage, except 'customer_birthdate', 'loyalty_card_number', 'kids_home', 'teens_home', 'year_first_transaction' and ‘typical_hour’
variables.

One of the requirements for proper functioning of the algorithms is to have standardized variables. For that to occur it was considered to divide the data set, in order that only some specific variables were used, excluding dummies, qualitative variables and variables regarding geographical information. This will ensure that the customer segmentation process focuses on relevant and quantifiable variables, the others will be used a posteriori.

We start by using the standard scaler to standardized the variables in all algorithms, however given the fact that K-means and Hierarchical DBSCAN were giving better results than DBSCAN and Fuzzy C-Means and in order to check for potential better results in the representation of segments, MinMax scaler were, also, implemented.

2.1. Clustering Techniques

Two distinct types of clustering algorithms were implemented to represent the segments within the data: classical clustering and fuzzy clustering. For classical clustering, the DBSCAN (Density-Based Spatial Clustering of Applications with Noise), K-MEANS, and hierarchical DBSCAN algorithms were implemented. These algorithms partitioned the dataset into distinct groups based on similarity and observed patterns in the data. Additionally, fuzzy clustering was performed using the Fuzzy C-Means algorithm, which also aimed to identify meaningful clusters but allowed for soft assignments, in which the same data point can be assigned to more than one number, which contrasts with the classical clustering.

To gain a visual comprehension of the dispersion of the clustered data, the UMAP (Uniform Manifold Approximation and Projection) technique was employed. By reducing dimensionality and visualizing the data in a lower dimensional space, it is possible to better understand the relationships between the clusters, detect any overlaps or separations, and gain insight into the underlying structure of the data.

2.1.1. DBSCAN

DBSCAN is a density-based clustering algorithm that groups densely grouped data points into a single cluster. The algorithm starts by randomly selecting a point, defining an area around that point, based on the epsilon provided. To determine whether the area is dense or not, a minimum number of points needs to be established. After conducting multiple attempts, a minimum value of 50 points and an epsilon value of 2 were considered appropriate for defining density in the algorithm.

This means that if the number of neighboring points within the specified area is equal to or greater than 50, the area is considered dense. The epsilon value of 2 denotes the maximum distance allowed between points to be considered neighbors.

The density of the clusters varied, with some clusters being more densely populated than others, namely cluster 0, with 10748 observations. This suggests that cluster zero represents a dominant group or a common pattern present in the data. In contrast, cluster five appears to be significantly smaller, containing only 224 observations. This indicates

that cluster five represents a distinct subset within the data, potentially characterized by unique properties or behaviors. DBSCAN was also capable of identifying 9 outliers.

2.1.2. Fuzzy C-Means

Fuzzy clustering is a form of clustering in which each data point can belong to more than one cluster. The algorithm used is the Fuzzy C-means clustering (FCM). Similarly to the k-means algorithm, a k number of observations are selected, being k the number of clusters.

The Fuzzy Partition Coefficient (FPC) is the measure used to assess the quality of clustering results from the algorithm. The coefficient is in the range of 0 and 1. When the FPC value is maximized, the clusters are well separated. By plotting the fuzzy partition coefficient, we were able to identify the quality of the clustering results, in which the best number of clusters occurs when k is equal to 6.

The algorithm calculates the strength of association for each point in multiple clusters and iteratively adjusts cluster centers and association values. The primary objective of FCM is to minimize the variance within the clusters. This process continues until the algorithm achieves convergence and stability.

The results of the FCM algorithm shows that the clusters have relatively similar sizes, contrasting with the cluster sizes observed in the DBSCAN algorithm, which exhibit significant differences.

Despite having similar cluster sizes, the UMAP visualization indicates that the clusters generated by the Fuzzy C-Means algorithm are not as distinct as those produced by DBSCAN. There is evidence of cluster overlap, particularly in cluster two, where individuals appear to be closer to individuals from other clusters.

2.1.3. K-Means

The K-means algorithm aims to partition a given dataset into a predetermined number of clusters, where each data point is assigned to the cluster with the closest centroid, using the Euclidean distance. The process continues until the centroids are stabilized, by minimizing the within-cluster sum of squares.

The Elbow Method

Using the elbow method, the optimal number of clusters was determined. The optimal value of the number of clusters occurs when the distortion begins to increase rapidly. When using StandardScaler, the optimal number of clusters were found to be 7, while with MinMaxScaler it was 6.

Silhouette analysis on K-Means clustering¶

Silhouette analysis is a useful technique for evaluating the quality of clustering by measuring the separation distance between clusters. The values obtained from the silhouette analysis have significance in evaluating the quality of clustering.

A silhouette value close to 1 suggests that the samples are well-clustered, while a value close to -1 indicates that the samples may have been assigned to the wrong cluster. Values around 0 suggest overlapping.

From the analysis of the silhouette of both MinMax and StandardScaler the better results were the MinMaxScaler, where clusters are well represented, compared to StandardScaler where clusters have a negative value, indicating that the cluster is not well represented.

Based on the analysis of silhouette scores for both MinMaxScaler and StandardScaler, the MinMaxScaler showed better results. In the MinMaxScaler transformation, the clusters were well represented, as indicated by positive silhouette scores. Alternatively, when using the StandardScaler transformation, the clusters had negative silhouette scores, suggesting that they were not well represented.

The results suggest that using MinMaxScaler yields better results in comparison to StandardScaler.

To seek for better clusters within the K-means results, Hierarchical DBSCAN (HDBSCAN) is used on the K-means clusters.

2.1.4. Hierarchical HDBSCAN

HDBSCAN, which stands for Hierarchical Density-Based Spatial Clustering of Applications with Noise, is a density-based clustering algorithm that extends the traditional DBSCAN algorithm.

The results for both MinMaxScaler and StandardScaler were both equal, and also represent the best results among all algorithms presented. The choice between the two was random, selecting the one with the MinMax scaler.

2.2. Cluster Insights

After the clusters have been defined using HDBSCAN segmentation, it is crucial to understand the underlying meaning and representation of each cluster. This helps in gaining insights into the distinctive characteristics associated with each cluster.

To facilitate the process of understanding the significance of each cluster, it is essential to take the average value of the features used for clustering for each cluster. This technique helps to understand the central tendencies and typical patterns of the data points within each cluster.

Cluster Insights:

1. Cluster 0 - Fish Enthusiasts:

This cluster is characterized by customers who do not complain, have no children, and primarily visit one store at around midnight. Their expenditure in various store sections is minimal, except for a significant focus on fish. They allocate a considerably higher portion of their expenses to fish compared to other customers.

2. Cluster 1 - Old and Cranky:

This cluster is characterized by customers who have registered the highest number of complaints. They tend to visit only one store and generally spend less than the average customer. With an average age of 71 years, they are the oldest segment within the dataset.

3. Cluster 2 - Young Drunk:

This cluster is characterized by customers who do not complain, visit an average of two stores at approximately 10 pm, and typically do not have children. They are the youngest customers, with an average age of 23, and exhibit the highest expenditure on alcoholic drinks.

4. Cluster 3 - Not Even Average:

This cluster is characterized by customers who fall below the average in most features, with some of them close to average. They have, on average, one child and an average age of 52 years.

5. Cluster 4 - Millennials:

This cluster is characterized by customers with almost no complaints, visits an average of three stores at around 6 pm, and have an average age of 29 years. They belong to the second cluster with the highest expenditure on groceries and total expenses.

6. Cluster 5 - Gamers:

This cluster is characterized by customers who do not frequently complain, or do it only once. They visit an average of two stores at approximately 9 pm and typically have one child. This segment demonstrates the highest expenditure on electronics, video games, and non-alcoholic drinks, most likely in energetic drinks.

7. Cluster 6 - Parents:

This cluster is characterized by customers who register an average of one complaint, spend above average across various categories, including groceries, vegetables, alcoholic drinks, and occasionally total expenses. They have an average age of 56 years and four children, which is the highest average number of children within a cluster.

8. Cluster 7 - Vegetarians:

This cluster is characterized by customers who rarely complain, visit an average of two stores at 10 am, and typically have two children. They do not allocate any expenses to meat or fish but spend the most on vegetables.

9. Cluster 8 - The Heart of the Company:

This cluster is characterized by customers who complain one or twice and visit around eight stores at midnight. They are the ones that have the longest average tenure with the company, 22 years and exhibit the highest expenditures in total, spending more in categories, such as groceries and meat.

10. Cluster 9 - Promotion Seekers:

This cluster is characterized by customers who tend to spend below average across all categories. They visit, on average, a single store and are the ones that purchase the largest portion of products in promotion, approximately 50%.

During the Exploratory Data Analysis stage, missing values were observed in the variable 'typical_hour'. These values were not set to zero in order to avoid potential misinterpretations that could arise. To address this issue, the two individuals with missing values were assigned to the segment that aligns best with their behaviors. This determination was based on similarities such as sharing the same last name, having no complaints or children, and exhibiting a significant expenditure on fish while spending minimal amounts on other products. These individuals were identified as fish enthusiasts.

10.1. Association Rules

Association rules analysis allows to uncover patterns and identify frequent associations, enabling the optimization of strategies and personalized recommendations. Measures like lift, confidence, and support provide insights into the strength and significance of these associations.

- Lift indicates the increased likelihood of one item being purchased when another is present. 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.

- Confidence calculates the conditional probability of finding the consequent item in a transaction given the presence of the antecedent item. Confidence values range from 0 to 1, with a higher value indicating a stronger association between the items.
- Support measures the frequency of a particular item or association rule in a dataset.

Association rules were implemented for each customer segment, for which fifteen associations with lift higher than 1.05 were presented. Also, the significance of the dummy variables in each customer is compared, such as the number of women and men, the level of education and loyalty card ownership.

Cluster 0 - Fish Enthusiasts: This cluster consists of customers from the MARL region in Lisbon who have not pursued education beyond high school. The gender distribution is equal, and most of them own a loyalty card.

Cluster 1 - Old and Cranky: Customers in this cluster reside throughout Lisbon and have not pursued education beyond high school. The gender distribution is relatively equal, and a significant portion of customers own a loyalty card.

Cluster 2 - Young and Drunk: Customers in this cluster live in areas of Lisbon near universities and have not pursued education beyond high school. The gender distribution is almost equal, and a majority of customers have a loyalty card.

Cluster 3 - Not even average: This cluster is concentrated in a specific area of Lisbon and has a higher proportion of customers with higher education. The gender distribution is fairly equal, and a significant majority of customers own a loyalty card.

Cluster 4 - Millennials: Customers in this cluster reside in various areas of Lisbon and have higher education. The gender distribution is almost equal, but the majority of customers do not own a loyalty card.

Cluster 5 - Gamers: Customers in this cluster live in a large area of Lisbon and have higher education. The gender distribution is fairly equal, and the majority of customers own a loyalty card.

Cluster 6 - Parents: Customers in this cluster reside in two significant areas of Lisbon and have higher education. The gender distribution is fairly equal, but the majority of customers do not own a loyalty card.

Cluster 7 - Vegetarians: Customers in this cluster are dispersed across a significant area of Lisbon and have higher education. The gender distribution is almost equal, and the majority of customers own a loyalty card.

Cluster 8 - The Heart of The Company: Customers in this cluster reside throughout a large area of Lisbon and have higher education. The gender distribution is almost equal.

Cluster 9 - Promotion Seekers: Customers in this cluster are spread throughout a large area of Lisbon and have higher education. While the majority have at most a high school education, there are also some with higher education. The gender distribution is almost equal, and the majority of customers own a loyalty card.

11. Targeted Promotion

The promotions for the different segments of customer were developed by taking into account the different association rules and cluster characteristics.

For Cluster 0 - Fish Enthusiasts we created the following promotions:

- Became a member today and get five candy bars for free
- If your a member for every purchase above one hundred euros get five euros cash back in cash and a free candy bar
- If you are buy above one hundred euros in fish get 20% discount in groceries and vegetables
- If you buy potatoes, eggs and olives get a 50% discount on cod so you can make Bacalhau a bras
- If you are a member older than five years you pay half the IVA

For Cluster 1 - Old and Cranky we have designed these promotions:

- If you became a member today and you are over seventy, get a lifetime discount on hygiene products.
- Get a treat for yourself, get one for your grandchild, if you buy a bottle of wine get a 25% discount on a cake or muffin.
- Getting ready for a big dinner, if you spend over one hundred euros get a 30% discount on meat or fish related products.
- If you are a member older than five year get a lifetime discount on groceries

For Cluster 2 - Young and Drunk we created the following promotions:

- Be a champion if you spend over 100 euros in video games get a 25% discount in champagne
- If you are a student and our member get a 25% discount in groceries if you get over fifty euros in alcohol drinks, so you can party without worries
- Get our loyalty card today and get a bottle opener and five cash back in the next five purchases of alcohol over fifteen euros.

For Cluster 3 - Not Even Average we implemented these promotions:

- If you have our loyalty card for purchases over fifty euros in electronics, get a 20% discount on wines and champagne.
- If you buy different types of wine you get a up to 15% discount on them.
- If you spend more than fifty euros on groceries, get a 25% discount on meat and fish related products.

For Cluster 4 - Millennials we selected the following promotions:

- If you are a member shop after 6pm and get a 35% discount on vegetables.
- Become a member of our loyalty card today and get fifteen euros cash back in the next five purchases over fifty euros on meat and fish products.
- Tired after work, I get two cakes for the price of one.
- Nostalgic struck get three retro video games for 75% of the price
- On weekends, get up to 20% discount on alcohol drinks if you spend over one hundred euros in groceries.

For Cluster 5 - Gamers we devised the following promotions:

- If you became a member today or you already have a loyalty card get priority on the queue, after all online videogames can not be paused.
- Nostalgic struck get three retro video games for 75% of the price
- Get a 25% discount in the following bundles:
    - Pokemon Sword + Pokemon Violet
    - Pokemon Shield + Pokemon Scarlet

- Half-Life Alyx + any of the Pokemon Games
- Ratchet & Clank + Half - Life 2

- If you spend more than twenty euros get a 50% discount in hygiene products
- You need to stay hydrated for every twenty euros and get five euros in non-alcoholic drinks.

For Cluster 6 - Parents we implemented these promotions:

- Get our loyalty card today and become exempt from IVA in children related products such as baby foods and hygiene products.
- Tired after taking care of your kids, have a beer. Get a free beer for every five euros spent in baby diapers.
- If you spend more than one hundred euros on groceries, get a 20% discount on baby hygiene products.
- Kids should eat their greens, get a 20% discount on meat and fish products and 20% discount on candy if you spend over fifty euros on vegetables.

For Cluster 7 - Vegetarians we selected the following promotions:

- If you have a loyalty card and you are Vegetarian or Vegan, then receive a 10% of purchase of non vegetables products in vegetables, because not every hero eats meat.
- Get your vegan, animal cruelty free toothpaste and get a 5% discount on your veggies.

For Cluster 8 - The Heart of The Company we created the following promotions:

- If you have a loyalty card and have been a customer for over ten years, then you only pay half the IVA on most essential products and some selected products.
- If you have a loyalty card and have been a customer over fifteen years, then you are exempt from IVA in most essential products and some selected products.

For Cluster 9 - Promotion Seekers we devised this promotion:

- If you have a loyalty card or get one today, you get a 10% cash back on most products that do not have a promotion already.

12.Conclusion

The process of segmentation started with the definition of the variables that should be used. As previously mentioned, the original data set ‘info’ was explored and new variables were extracted from it by means of linear combination and the use of conditions. Some of the variables from which the new ones were extracted, were consequently removed from the dataframe, since they were not in the most desirable format. The data was also subject to visualization using histograms, pie charts and a scatter mapbox, which provided information about the general population that lives in Lisbon, has a fairly uniformly distributed age, whose majority have at most concluded high school and is skewed towards having within three children.

Before starting the procedure of creating the segments, the two observations with missing values in the column ‘typical_hour’ were set aside. Subsequently, the numerical variables selected to be used in segmentation were scaled using the MinMaxScaler and the StandardScaler. The first clustering technique executed was the DBSCAN for the StandardScaler, which was tested multiple times and the parameters considered as optimal were selected. However, this implementation described an inconsistent behavior with slight changes of the parameters, which made it less trustworthy, and therefore it was not selected. Following that, the Fuzzy CMeans was implemented also using the StandardScaler. To select the optimal number of clusters, a plot with the fuzzy partition coefficient was used. The definition of the clusters formed by this algorithm were tested with the usage of a UMAP, which suggested that the clusters were not very well defined. There were clusters that combined observations that appeared not to belong in the same segment, according to the UMAP generated image. Thus, this second implementation was also not selected as the final segmentation.

The last segmentation conducted began with the execution of the K-means algorithm for data scaled using the two mentioned scalers. The elbow curve plot was used to choose the optimal number of clusters and their quality was accessed by a silhouette plot and a UMAP. Both these visualization tools displayed that the K-means that used the MinMaxScaler has much better defined clusters than the one with the StandardScaler. Then, a hierarchical DBSCAN was executed using as input the UMAP for both scaled

data. This HDBSCAN exhibited perfectly defined segments, according to its UMAP and it formed the exact same clusters for the two different scaled data.

The last implementation was the only one where two different scaled data were used, since all the clustering techniques were initially conducted with data scaled by the StandardScaler. Once the last segmentation was deemed so successful, the MinMaxScaler was used to check if it could possibly provide a better clustering solution. As a matter of fact, it was not able to provide a better solution, but yes an equally successful one, which was eventually chosen as the final segmentation.

Immediately following the selection of the best segments, they were described based on their means and in comparison to the population mean for each feature. Afterwards, the two observations that were initially set aside, were appended into the data set, with their most suitable cluster. This procedure started by the visualization of the respective observations' characteristics, which were, briefly, spending an unusual amount of money on fish, spending very little on all other sections, having 'Supermarket' as last name, and only purchasing one distinct product. This characteristics made this observations fit perfectly in the cluster zero (Fish Enthusiasts), therefore, their inclusion in the data set was straightforward.

In the association rules section, the only association rules considered for each cluster were those with the largest lifts, a minimum support of 0.05 and a minimum confidence of 0.2. Among the ones who satisfied this conditions, those with the greatest confidence and ultimately, support, were used in conjunction with the description of each cluster to create promotions. For instance, in cluster zero, most of the association rules include candy, and, knowing that their main purchase is fish, their promotions were made to incentivise them to buy other products, such as candy, to eventually have discounts on fish. In cluster one, most associated rules were related to products that we assume to be directed to the customers' grandchildren, since this cluster is composed by seniors. It was also considered that they mostly purchase vegetables, groceries and alcohol. In cluster two, it was considered that they buy a lot alcoholic drinks, which also composed most of their association rules. The remaining clusters were subject to this type of reasoning for the creation of their promotions.

13. References

[1] *How HDBSCAN works*. (2016). The hdbscan Clustering Library. URL: https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html

[2] Nara Ramezani. (2022, May 13). Fuzzy C-means clustering. Medium. URL: https://medium.com/@nafiseramezani1985/fuzzy-c-means-clustering-f9e047e4e458

[3] Avijeet Biswal. (2023, February 17). What is exploratory data analysis? Steps and market analysis. Simplilearn. URL: https://www.simplilearn.com/tutorials/data-analytics-tutorial/exploratory-data-analysis

[4] Ibrahim Abayomi Ogunbiyi. (2022, November 2). *How to perform customer segmentation in Python*. freeCodeCamp. URL: https://www.freecodecamp.org/news/customer-segmentation-python-machine-learning/

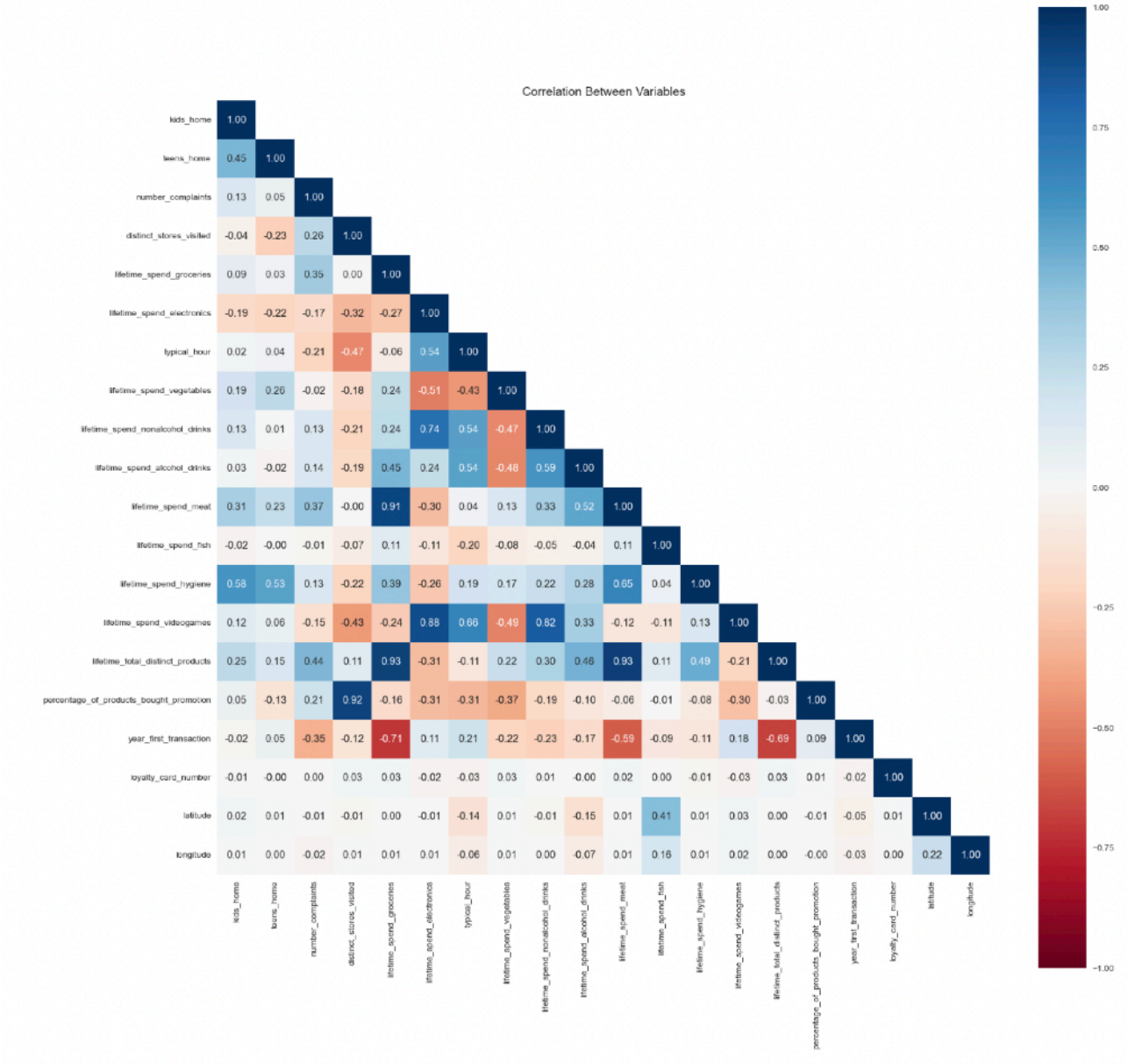
[5] Baruah, I. D. (2020, October 25). Cheat sheet for implementing 7 methods for selecting the optimal number of clusters in Python. Medium. URL: https://towardsdatascience.com/cheat-sheet-to-implementing-7-methods-for-selecting-optimal-number-of-clusters-in-python-898241e1d6ad

[6] Karnika kapoor. (2021, October 8). *Customer segmentation: Clustering*. Kaggle. URL: https://www.kaggle.com/code/karnikakapoor/customer-segmentation-clustering#DATA-PREPROCESSING

[7] Marcelo Marques. (2018, November 27). *Customer segmentation and market basket analysis*. Kaggle. https://www.kaggle.com/code/mgmarques/customer-segmentation-and-market-basket-analysis

14.Appendices

14.1. Correlation Matrix



14.2. Visual Graphs