NOVA IMS Information Management School

Machine Learning II

PROJECT REPORT

Customer Segmentation Study

Francisco Gomes - 20221810
Maria Henriques - 20221952
Marta Almendra - 20221878

2023/2024

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Nowadays, identifying client behavior and adapting marketing tactics to match unique demands is critical to corporate success. This project focuses on using customer segmentation to drive business success by breaking down a vast client base into smaller, and easier to handle groups with similar characteristics.

To accomplish so, two datasets, customer_info and customer_basket, were used, which included detailed information on customer demographics, spending habits, purchase behavior, and history transactions.

Our approach began with thorough Exploratory Data Analysis (EDA), which included data cleaning and preprocessing to guarantee that the datasets were suitable for further analysis. This step was also highly useful for conducting a preliminary analysis and visualization of the data to better understand the underlying patterns and relationships.

The next element of our strategy was Segmentation, in which we employed statistical and machine learning methodologies to identify separate customer segments. This was done by using a variety of clustering techniques, including K-means and different types of hierarchical clustering, to diverse kinds of scalled data.

The analysis showed multiple separate client groups, each with their own traits and habits, and our final results consisted of the following 10 segments: 'promo hunters', 'high spenders', 'vegetarians', 'pet lovers','tech enthusiasts', 'demanding foodies', 'large families','college students', 'karens' and 'fishy', which will be described in this report.

Understanding these groups provided us with valuable insights into their motivations, interests, and demands, which, combined with the results of association rules created for each of them, enabled us to develop targeted marketing strategies to address the unique needs of each segment.

Finally, our customer segmentation study provides a strong tool for companies to better understand their customer base and design tailored marketing strategies that generate engagement and loyalty, giving them a competitive advantage and achieving long-term success.

EXPLORATORY DATA ANALYSIS

The Exploratory Data Analysis was divided into five phases: Preliminary Data Analysis, Feature Transformation and Engineering, Data Segmentation and Cleansing, Missing Values Imputation, and Multidimensional Outlier Detection. For all prior actions, the customer_info dataset was used.

PRELIMINARY DATA ANALYSIS

To begin pre-processing, we conducted a basic data analysis. After importing the data, we discovered an odd column called 'Unnamed', which included no relevant information and was therefore removed.

Following this, the data types and missing values were checked. In terms of data types, only three variables were objects: 'customer_name', 'customer_gender', and 'customer_birthdate', with the remaining being all float64.

Regarding the missing values, eight variables had **missing data**. To gain a clear insight, the percentage of the missing values was calculated for each variable, aiming to better comprehend the impact on the dataset. We found small percentages of missing values in the 'kids_home', 'teens_home', 'number_complaints','distinct_stores_visited','typical_hour', 'lifetime_spend_vegetables', 'lifetime_spend_fish', and a great percentage in the 'loyalty_card_number' variable.
Since this is an unsupervised learning project, there's a need to be carefull on how we deal with missing data as it can significantly influence the quality of the clusters formed, so the most suitable approach was to deal with the missing data at the end of the EDA section.

Next, we analyzed both the numerical and categorical data, looking at their statistics and particular variable information. Concerning the numerical data, an **inconsistency** was discovered in the variable regarding the percentage of product bought on promotion, as the values ranged from -0.477986 to 1.196858, indicating that the dataset had values that were outside of an expected percentage range of 0 to 1.

Specifically, 7.68% of the dataset had this issue. This problem was addressed later in the Feature Transformation and Engineering section.

Regarding the categorical data, we initially saw that an abbreviation of each customer's level of education preceded their name ('Bsc.', 'Msc.', 'Phd.') and assumed consumers with only their name had no degree. Using this information, a variable was later constructed for each customer's education.
It was also discovered that there were a significant number of instances with **'Fishy'** in customer's name. In order to perform a more thorough study, the data was separated into two datasets, one of which was solely focused on the 'Fishy' clients.

FEATURE TRANSFORMATION AND ENGINEERING

Following an initial examination of the data, we implemented several variable modifications to enable for a more thorough investigation and comprehension of it. That being said, four variables were altered and five new ones were created.

1. Customer Age

We used python's datetime module, more specifically the pd.to_datetime function, to convert the customer's birthdate to their age in years.

2. Customer Loyalty

To create a variable that examined customer loyalty, we transformed the loyalty card number variable to a binary indicator that indicated if the consumer had a loyalty card or not. The missing values were assumed has if the client didn't have a card.

3. Customer Gender

A binary transformation was performed for the customer's gender, with a 0 representing females and a 1 representing males.

4. Total Spend

The ‘total_spend’ variable was created by summing the 'lifetime_spend' columns for each customer.

5. Percentage Lifetime Spend

As the variable name implies, we also chose to establish variables that reflect the percentage of a client's lifetime spending across all categories.

6. Percentage Products Bought on Promotion

In our preliminary analysis, we discovered that the values for this variable were not appropriate for percentages, thus we implemented an adjustment to ensure that a percentage was never more than one (100%) or lower than zero.

7. Customer Education

As previously mentioned, a customer’s level of education was present in their name. For this reason, with the help of a function created, we were able to extract an education level from the 'customer_name' field and clean the name by removing education abbreviations. So each education level was represented as the following: no degree - 0; 'Bsc.' - 1; 'Msc.' - 2, 'Phd.’ - 3.

8. Customer Tenure

For this variable, we transformed a client's year of first transaction to customer tenure in years, which represents the number of years since their initial purchase.

9. Dietary Preferences

This variable was constructed to assess the clients' dietary preferences, categorizing them into four purchase options: 'Only Fish', 'Only Meat', 'Fish and Meat', and 'No Fish or Meat'. A small number of clients who did not identify with any of these categories were classified as unknown and excluded from the dataset.

DATA SEGMENTATION AND CLEANSING

In response to previously reported customer name information, specifically the existence of clients with 'Fishy' in their names, we opted to separate the customers into two independent DataFrames, allowing for a more advanced examination. This allowed us to identify and address trends and behaviors unique to each group.

Before dividing the dataset, we investigated their locations and determined that the majority of these 'Fishy' clients lived in more coastal places such as Peniche and Ericeira, whereas the remaining customers were all in Lisbon (Figure 1).

This step was critical to ensure that any insights or strategies developed were relevant and effective for each distinct customer group within the dataset.

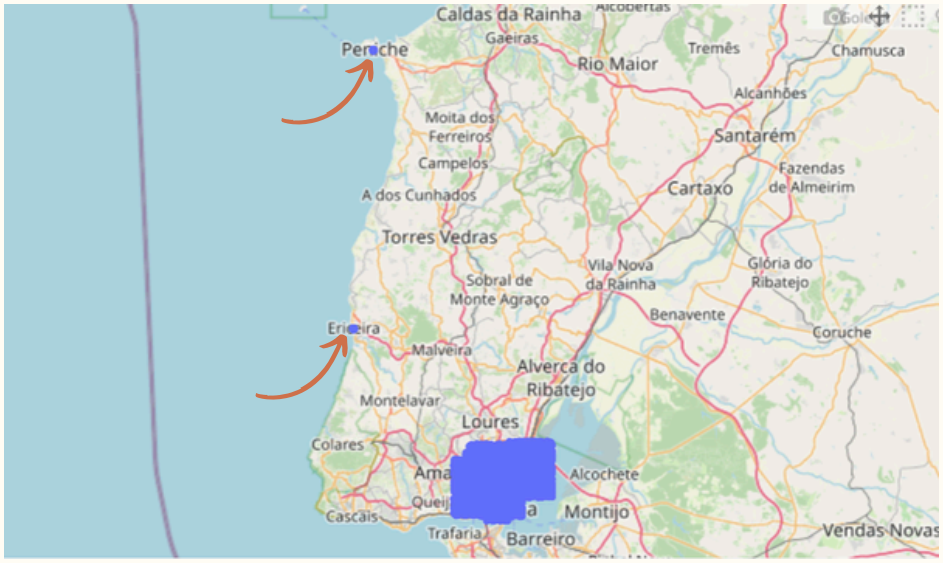

Fig. 1 - Customer Locations

1. ‘ PERSONS’ DATA

To have a better understanding of each variable and identify possible patterns or trends, some conclusions were taken based on visualizations.

1.1 Visualizations

We started by analysing the distribution of the customers’ age and their spendings on all categories. Regarding **customer_age**, the distribution appears relatively uniform across the different age groups, showing slight peaks in the 20's to 30's and in the 70's to 80's.

Moving on to the **lifetime_spend** variables it is possible to see that the spending patterns across the various categories, generally show a skew towards lower amounts, suggesting that most customers tend to spend less, with only a few spending more. Additionally, the data reveals that while customers purchase a wide variety of products, the majority tend to buy a smaller range of items. A significant portion of products are bought on promotion, highlighting customers' sensitivity to discounts and special offers.

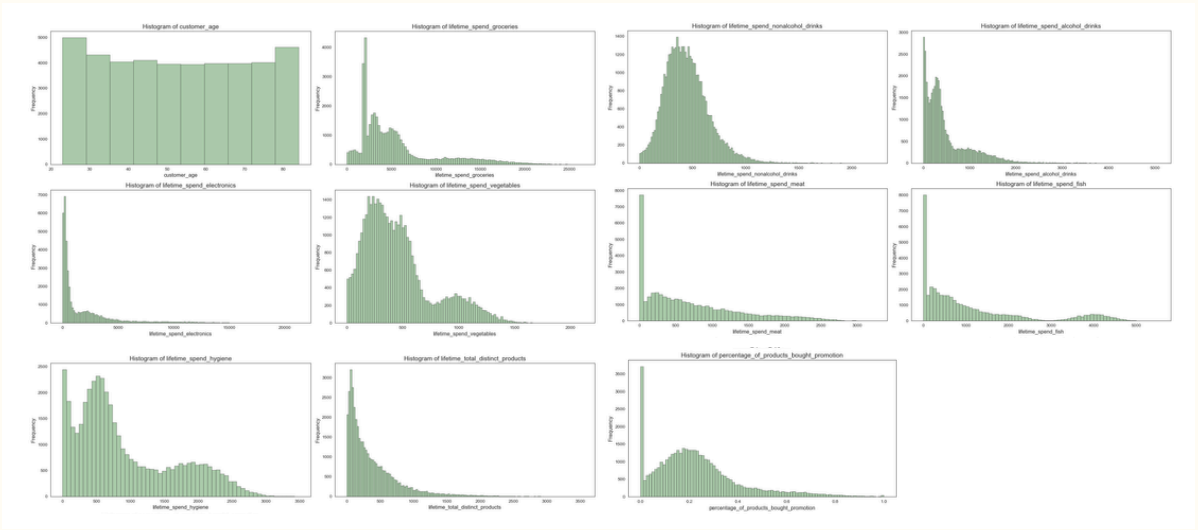

Fig. 2 - Histogram Distributions

To better understand our client's spending patterns, we examined their lifetime spending by category. We determined that groceries were the most common purchase among customers, followed by pet food and electronics, whereas the categories with the least amount spent were alcoholic drinks, vegetables, and nonalcoholic drinks.

Furthermore, we discovered that the majority of clients lacked a degree, whereas the number of customers with a Bachelor's, Masters, or Ph.D. degree was quite comparable. Most consumers have only gone to one store having rarely visited more than six different ones, most have filled one complaint and the majority made their first purchase 14 years ago. Also, it is more typical for a client to have one kid or adolescent at home, and to purchase food for an omnivorous diet.

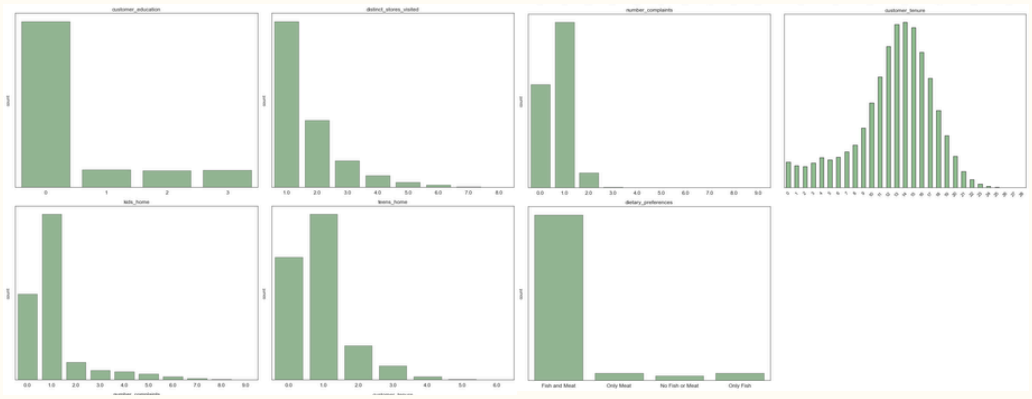

Fig. 3 - Bar Chart Analysis

Next, we discovered that the majority of consumers (56.3%) had a loyalty card, implying that loyalty programs may be quite effective with this client group. The gender distribution was found to be reasonably balanced, with 50.2% of clients being female and 49.8% being male. This indicated that gender may not be a key distinction in this client group, and gender-neutral marketing methods might be beneficial.

It was also found that the most frequent typical hour to visit a store is around 5:00 pm, which might be explained by customers who have just finished their day at work, followed by a large number of clients in the early morning, particularly around 9:00 am.

1.2 Outliers

The presence of probable outliers was also assessed and reported on using the prior visualizations shown in Figures 2 and 3. We were able to identify the number of apparent outliers in each variable by doing a visual assessment and creating a function to count the number of customers who surpassed a specified threshold (chosen based on the distribution of each variable).

We chose this technique above others such as IQR and *z*-score because, in the first trial, they classified points as outliers when they were not. As a result, we removed these tactics from our study.

Our conclusions were that only a small percentage of consumers visited more than seven different stores, had more than eight children or five teenagers at home, filed more than six complaints, purchased more than 90% of items on promotion, and made their first purchase more than twenty-five years ago. The findings in relation to lifetime spending categories are as follows:

- Groceries over 26000 euros,
- Eletronics over 20000 euros,
- Vegetables over 1900 euros,
- Non-alcoholic drinks over 1800 euros
- Alcoholic drinks over 4300 euros,
- Meat over 3200 euros,
- Fish over 5300 euros,
- Hygiene products over 3300 euros,
- Total distinct products over 3200.

Since they represented only a small portion of the data, they were were kept for further analysis.

1.3 Correlations

Furthermore, the correlation between variables was analysed. We started by analyzing the correlation matrix and found some highly correlated variables. To have a better understanding of this correlations, some visualizations were plotted. Some of the variables that presented a high correlation were **lifetime_spent_fish**, **lifetime_spent_videogames** and **lifetime_spent_groceries** with the **total_spent**. The correlation between this spending categories and the total_spent sugget that this variables influence the most the total spent of each customer, we can also see that the distribution of this variables follow a very similar pattern.

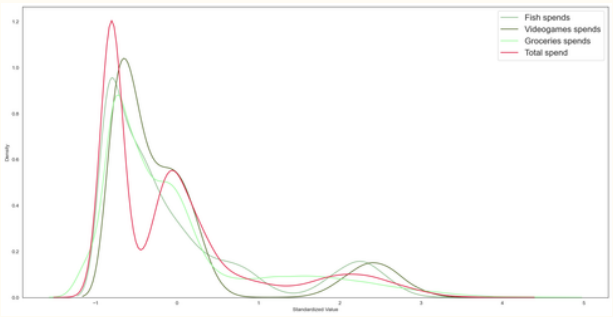

Fig. 4 - Density plot between the variables lifetime_spend fish, groceries, videogames and total_spent

For the remaining highly correlated variables scatterplots were examined. The first pair analysed was **lifetime_spend_meat** and **lifetime_spend_fish**. The plot (Figure 5) presents a positive correlation between spending on meat and fish, indicating that customers who spend more on meat also tend to spend more on fish. It is also possible to see the presence of two distinct groups, one with moderate spending and another with higher spending.

Regarding **lifetime_spend_meat/fish** with **lifetime_spend_eletronics**, the plot (Figure 6) suggests that customers who spend more on electronics also tend to spend more on fish. There is a dense cluster of data points along the lower spend regions for both categories, but the spread along the electronics axis implies that while many customers spend lower amounts, there are several high spenders on electronics among them.

The relationship in the plot concerning eletronics and meat seems weaker compared to the electronics and fish. The data points are more spread out, indicating a broader range of spending behaviors.

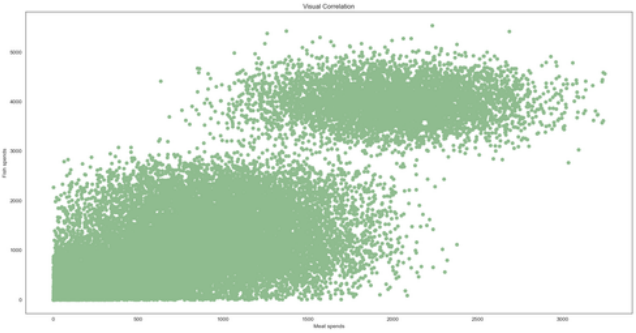

Fig. 5 - Correlation between lifetime_spend fish and meat

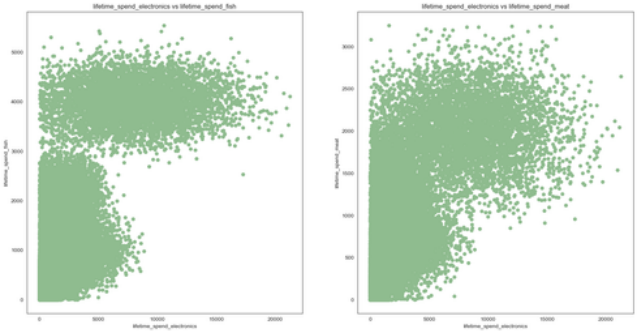

Fig. 6 - Correlation between lifetime_spend fish/meat with eletronics

The last variables analysed were **lifetime_spend_fish/meat/eletronics** with **lifetime_spend_videogames** (Figure 7). Concerning videogames and fish, it is possible to see that as videogame spending increases, fish spending also increases. Regarding videogames and meat, the plot seems to have a weaker correlation compared to the previous, higher spenders on video games do not necessarily spend significantly more on meat. Examining the videogames and eletronics scatterplot, it shows a more pronounced upward trend, suggesting a stronger correlation between spending on video games and electronics. This makes intuitive sense as both are related to technology.

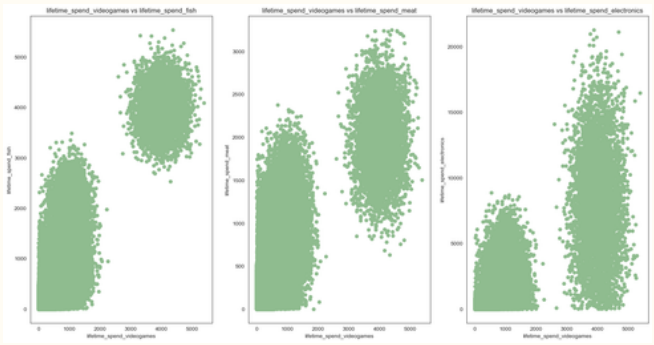

Fig. 7 - Correlation between lifetime_spend fish/meat/eletronics with videogames

2. 'FISHY' DATA

Moving on to the **'fishy'** data, since it had a very small number of consumers with quite different behaviours from those mentioned above, a less thorough examination is necessary, as they will eventually be handled as their own customer group. Nonetheless, to determine that they reflected different behaviours, certain analysis and visualizations were performed.

2.1 Visualizations

Equivalent to what was done in the 'Persons' data, we began to analyse the distributions of the customers' age and their lifetime spends on all the categories. Concerning the customer_age variable, 'Fishy' customers tend to be slightly older on average compared to 'persons' customers, whose age distribution is more uniform.

About the **lifetime_spend** variables, it was noted that their distribution was slightly different from the 'Persons' data. The spending pattern across the different categories tends to cluster around mid-range amounts, indicating that 'Fishy' customers have moderate spending habits, although it was noticed that regarding **lifetime_spend_fish**, the values were way higher than the other categories suggesting that these customers tend to spend more on fish.

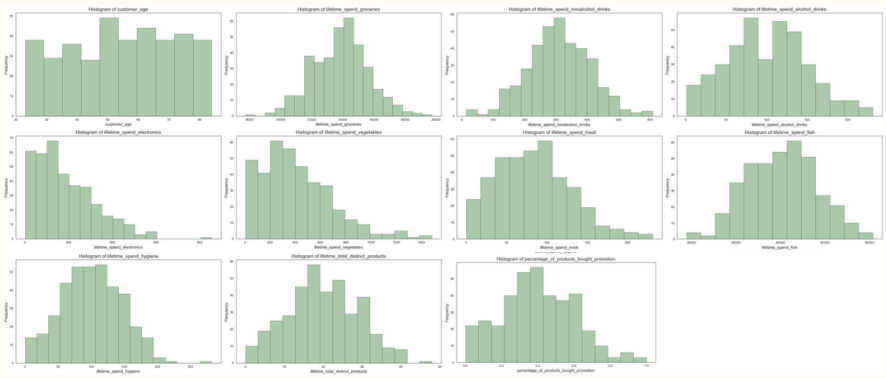

Fig. 8 - Histogram Distributions

Moreover, we discovered that a large number of clients lacked a degree, whereas the number of customers with a Bachelor's, Masters, or Ph.D. degree was small. All consumers have only gone to one store and the number of complaints is quite balanced between zero and one, with one being slightly higher. Regarding the kids and teens, most 'Fishy' customers have no children. Similar to the 'Persons', the distribution of the customers' tenure tends to be left skewed, suggesting that most clients have a great antiquity. About the customer's dietary preferences, the majority tend for fish and meat.

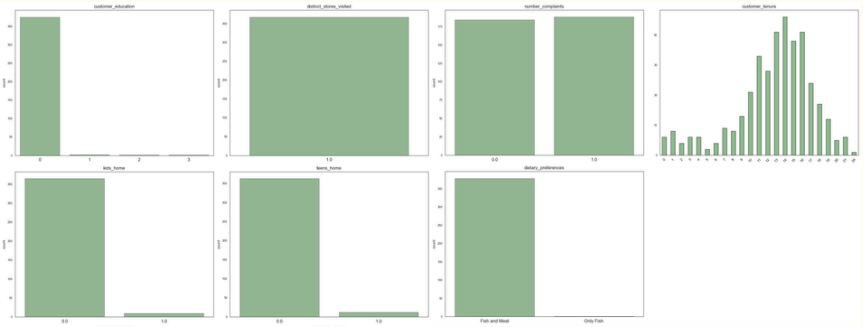

Fig. 9 - Bar Chart Analysis

We also discovered that the majority of the clients possessed a loyalty card (79.7%), which is much greater than in the 'persons' data, and that, similarly to before, the customer gender distribution is pretty comparable, with a little higher proportion of female clients (51.5%) than men (48.5%).

In terms of the most typical time for a store visit, unlike the 'persons' clients, these often arrive at 5 p.m. or incredibly early in the morning. It is worth noticing that there are no values after 5 p.m., which might indicate that these clients have a different work schedule than those previously evaluated.

Finally, to further assess the customers spending patterns, their lifetime total spending by category was plotted. As previously stated, their leading category is fish expenses, followed by groceries. In this graphic, available in the annexes, it is simpler to see the low quantity of remaining category expenditures when compared to these two. This demonstrates, yet again, that these clients' habits differ greatly from those of the 'persons' data.

2.2 Correlations

Possible correlations were then examined, with a focus on the variables that were highly correlated. It was revealed that the only variables with a significant correlation were lifetime_spend_fish and total_spend, showing an almost identical distribution. This finding validates the dataset's name once again, and, after a final analysis, we determine that these clients are most likely fishermen or just pescetarians, with distinct living habits presumably owing to their location.

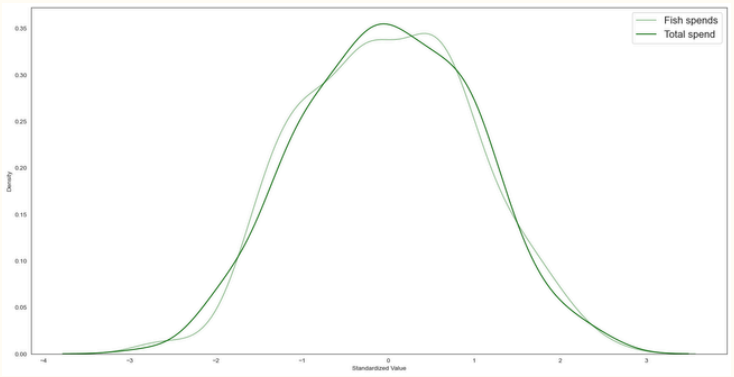

Fig. 10 - 'Fishy' Density plot between the variables lifetime_spend fish and total_spent

MISSING VALUES IMPUTATION & MODELING PREPARATION

This section dealt with preparing the data for customer segmentation, as well as the missing values reported in the Preliminary Data Analysis. For the first point, a final feature transformation was made since, based on the visualization analysis, we believed that kid_home and teen_home were very similar and it would be advantageous to construct a variable that reflected both. As a result, the variable '**nr_child**' was used instead. Lastly, the following parameters were used to fill in the missing values in the 'persons' data using the **KNN Imputer:** 4 nearest neighbours and distance weighting.
This was done because a k number of neighbours equal to 4 is a commonly used and recommended value for analyzing the local neighbourhood structure in data, providing a good balance of sensitivity to local variations and robustness to noise, and distance weighting gives more weight to data points closer to the target, improving the precision of spatial predictions and models.

MULTIDIMENSIONAL OUTLIERS

In order to have some initial insights about potential clusters, and since we had decided to treat the 'Fishy' as its own segment, possible multidimensional outliers were investigated for the 'persons' data. We opted to exclude the variables 'customer_name', 'customer_gender', 'customer_loyalty', 'total_spend', 'customer_education', and 'dietary_preferences' for improved comprehension and analysis. To do so, DBScan which is an algorithm that identifies clusters with varied forms and densities, efficiently distinguishing between dense clusters and sparse noise points was utilized, with the same logic of 4 nearest neighbors.

We found potential anomalies regarding a low number of vegetables and meat purchases, which might represent customers with specific dietary preferences, financial constraints, or other behaviours that differ significantly from the majority of the population.

Moreover, a low number of complaints and promotion purchases was found, which could indicate a group of customers who do not complain but have unusual purchasing behaviour regarding promotions. Regarding these variables, some extreme values were also found, showing customers with highly atypical buying patterns, such as buying a high proportion of promotional items without registering complaints.

We also discovered a small group of clients with atypical drinking behaviours, this is, that have very low spending on both drink categories, possibly due to lifestyle choices or dietary restrictions.

Finally, we discovered a set of clients with low spending on both pet food and video games, as well as some extreme values who give a large proportion of their spending to pet food.

The plots depicting these conclusions can be found in the annexes.

EXTRACTING THE DATA

The last step of our Exploratory Data Analysis consisted in the extraction of the datasets to be used in further steps of the project.

CUSTOMER SEGMENTATION

This section focuses on customer segmentation, a crucial method for identifying and categorizing consumers. Using complex clustering algorithms, we identified diverse groups within our customer data, allowing for more targeted and profitable campaigns. This segmentation not only enabled us to tailor our services to each group's specific needs, but also increased our capacity to anticipate consumer behaviour, raise customer satisfaction, and optimize resource allocation.

METHODOLOGY

The phase began with experiments involving various forms of clustering. Methods such as single, complete, average linkage and Dbscan, had poor results, therefore we did not include them in our analysis and will not cover them in this report. The chosen ones were as follows:

1. KMeans
2. Ward (Hierarchical)

In our first efforts, we tested all available features before deciding to remove some independent variables that impacted the clusters. This occurred because binary variables are given more weight than others in segmentation, and certain types of scaling, such as the MinMax scaler, are sensitive to them. Furthermore, several variables representing categories, such as level of education and dietary preferences, were omitted for the same reason, and the variable 'total_spend', which was strongly correlated with the majority of the variables mentioned in the EDA section and thus unnecessary to help define segments, was also excluded from this step.

The data used for clustering was scaled using several methods, including no scaling at all, Standard Scaler, Min-Max Scaler, and Robust Scaler, and all possible combinations were tested. This thorough approach ensured that we found the most efficient scaling solution for our distance-based clustering approaches.

The initial step for each scaled dataset was to determine the appropriate number of clusters to analyze. For K-means, a plot displaying an inertia and silhouette score was created, and for the other methods, a dendrogram was displayed.

The r-squared (R^2) value for each clustering technique was also calculated based on the number of possible clusters. This allowed us to determine which strategies were most effective and had the highest scores. Given that the Ward method consistently outperformed the other methods in terms of the R^2 metric across all cluster numbers, it was justifiable to focus on analyzing it for hierarchical clustering in this context. In other words, the method provided a superior explanation of the variance within the data, making it the most reliable choice for this analysis, as seen in Figure 11.

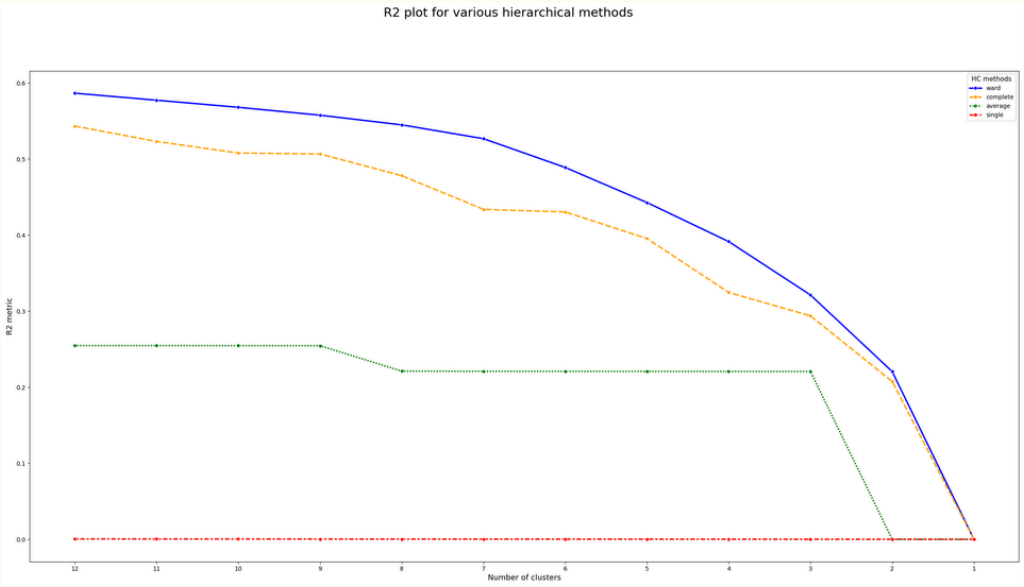

Fig. 11 - R^2 for MinMax Scalled Data Given Different Clustering Methods

This method with no scaling received the highest score, followed by robust scaling and min-max scaling. Even though the r-squared with a min-max scaler was only the third best, the UMAP's interpretability was significantly superior. Given this, it was the preferred method as will be explained later.

The mean values for each variable were then shown per segment, and an analysis of each segment was performed. This persisted in attempting to uncover patterns or characteristics in a variable that defined the segment. In each case, we identified the best solution based on our perspective, particularly in situations where solutions with different numbers of clusters were compared. To help in this process, a comparison of the clusters was offered as a reference, allowing for an understanding of the differences between them.

We concluded after a lengthy period of trial and error in determining the best combination of the proper selection of variables to utilize, the kind of scaling, the clustering method, and the number of segments that should be used. A uniform manifold approximation and projection (UMAP) graph was produced to better visualize the various segments and conclude it was the best possible solution.

Our findings revealed that the most logical and distinguishable segments were created by combining the MinMax scaler with a hierarchical clustering approach (Ward's method) with nine segments. Initially, our "best" solution comprised of only 8 segments with this method, but we had difficulties finding a specific cluster of clients owing to their similarity to the other segment. We first attempted to probe further into the cluster that contained the two groups of clients by using hierarchical clustering, k-means, and DBScan inside it, but were unsuccessful. Finally, we determined that adding one more segment to the final solution was enough to completely separate these two groups.

Below is the obtained UMAP of the retrieved segments (except 'Fishy'):

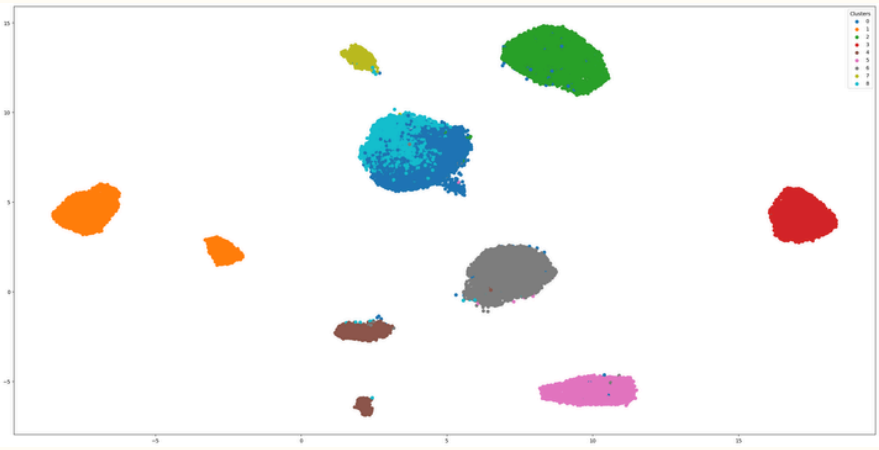

Fig. 12 - UMAP of final segments

OBTAINED SEGMENTS

As previously stated, a final segment depicting the 'Fishy' customers was explored. The analysis identified 10 separate client segments, demonstrating a diverse customer base for the business in question, with their dimensions visible in Figure 13.

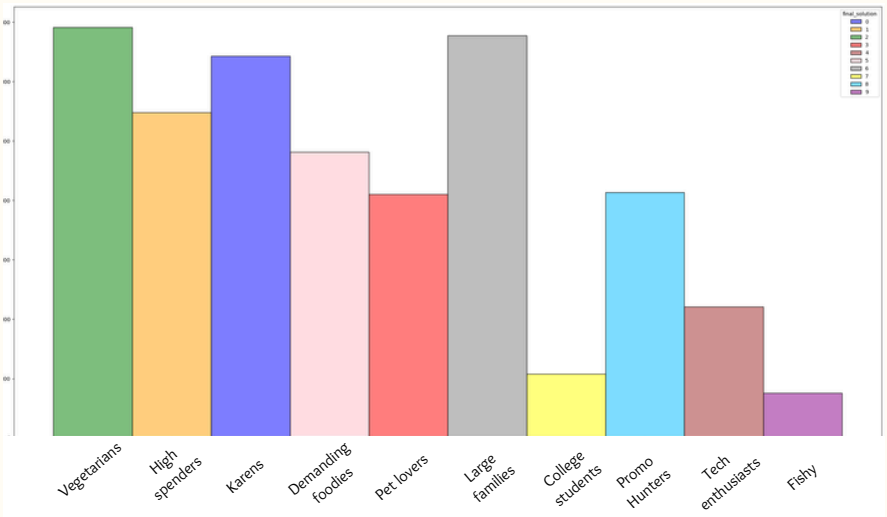

Fig. 13 - Dimension of each Segment

Their respective designations and descriptions are as follows:

1. Karens

The "Karens" section includes older clients, with an average age of 63.7 years. They have a large amount of complaints, indicating that they are more vocal about service difficulties, hence their segment name. These clients visit around two to three stores, demonstrating a willingness to try different stores, and possibly looking for products on promotion. They spend moderately on groceries (32.5% of their lifetime spending) and on alcohol (7.2%), but lesser in other categories. We also noticed that the Karens cluster has the highest count of customer_education at zero, meaning many in this group do not have formal education. Overall, they are older, more diverse consumers who value groceries and are prone to express worries about service quality.
This segment was particularly difficult to find, since these customers share some similarities with the so-called "Promo Hunters", given that both have a relatively high amount of products bought on promotion. To help discover them, a function named check_for_karens was used, which checked in each segment if customers typically had more than four complaints.

2. High Spenders

The "High Spenders" are middle-aged customers, with an average age of 55.2 years, who exhibit high spending behaviour across many categories. They have a low complaint rate and frequent 1.4 stores, indicating a preference for certain stores. Their average spending levels are significant, particularly in groceries and electronics, which make up a significant portion of their total spending. They also have about two children, which may explain part of their spending habits. Regarding their location, these customers tend to live in developed areas of the city (Figure 14), which are expensive in Lisbon, reinforcing once again their high purchasing abilities.
In conclusion, this category is marked by its high spending power, preference for food and technology, and the presence of children in the household.

3. Vegetarians

Customers in the "Vegetarians" sector have an average age of 55.1 years and are defined by their dietary habits. They spend a lot on vegetables but relatively little on fish and meat. They have fewer complaints and have around one to two children. As mentioned, this group spends a considerable portion of their spending on groceries (50.1%) and vegetables (16.3%) but also a significant portion on pet food (8.3%) indicating these clients are likely pet owners. Another insight found was that they represent the highest number of customers with no loyalty card. Interestingly, these customers also frequently visit only one store. This raises an intriguing question: if they consistently visit the same store, why do they choose not to have a loyalty card? Possible reasons could be that they may not see the value or benefits offered by it, might perceive it as an intrusion on their privacy, or even be unaware of the loyalty program or its benefits.
In essence, this large sector comprises of middle-aged clients that value a plant-based diet, have few complaints but no loyalty card, and buy less frequently.

4. Pet Lovers

The "Pet Lovers" have a typical age of 55.3 years and spend an average of 10,009€ on pet food, accounting for 47.1% of their total spending. They make very few complaints, visit only one store, and don't have any video-game expenses. These clients typically don't have any children.

They lay less focus on other categories but show a strong tendency for pet-related items. This segment is marked by a significant investment in pet care, a limited focus on other retail areas, and a low number of children.

5. Tech Enthusiasts

The "Tech Enthusiasts" sector is one of the youngest, with an average age of 30.5 years. They have extremely few complaints and generally visit on average 1.4 shops. Their purchasing habits indicate a strong affinity for electronics, and a significant portion of their overall spending (27.9%) goes to this sector. Their grocery spending is moderate, as are their videogame and pet food expenses. These customers have few children and similar to 'High Spenders' they tend to live in more developed and expensive areas of Lisbon (Figure 14).
Therefore, this segment is distinguished by its young age, emphasis on technology purchases, and low number of complaints.

6. Demanding Foodies

Customers in the "Demanding Foodies" sector are typically 54.8 years old, have one to two children, have a high rate of complaints, and visit about two stores. They have a broad spending pattern, purchasing a high amount of distinct products, with significant expenditures on groceries, alcoholic beverages and other areas. A sizable percentage of their spending is also allocated to meat, fish, and hygiene items. Lastly, these clients tend to visit a store relatively earlier than all other segments, with their average typical hour for a store visit at 9 a.m.
This sector is distinguished by their variety of products bought, large expenditure on food and drinks, and greater likelihood to express their requests or concerns. Additionally, they also have the biggest numbers in customer tenure.

7. Large Families

"Large Families" are consumers with an average age of 55.2 years, known for their high complaint rates and fewer different store visits, usually always visiting the same store. As the name suggests, they have a large family (an average of five children at home) and spend a significant portion of their earnings on groceries and other products such as hygiene and electronics, as is typical for a family.

This cluster is also characterized by having the highest amount of loyalty cards, likely due to the benefits and savings offered by them.
In conclusion, they are characterized by their high family size, high grocery and electronics expenditures, and frequent complaints. These families find value in the loyalty programs and appreciate the convenience and cost savings associated with shopping at a single, familiar store.

8. College Students

The "College Students" sector consists of the youngest clients, with an average age of 24. These are likely university students who have almost no complaints and generally visit around 1.4 stores. Their overall spending is minimal, with a great emphasis on alcohol, which accounts for a substantial portion (36.3%) of their total expenditures. These customers have extremely few children (0.35 on average) and the majority are distributed around the centre of Lisbon (Figure 15).
This group stands out by its young customers, low total spending, a strong emphasis on alcohol purchases, and a minimal amount of complaints.

9. Promo Hunters

"Promo Hunters" are clients with an average age of 42, have around one to two children, most have filed a complaint, and visit many stores (2.8 on average), showing a great desire to seek out discounts. Their spending is reasonable throughout many categories, with a strong emphasis on promotions, having purchased almost 50% of their total products on promotion. They spend modestly on groceries and other remaining categories.
This group is defined by its high retail visiting values, emphasis on promotional offers, and elevated complaint frequency.

10. Fishy

The final segment relates to the previously examined "Fishy" consumers.
In summary, this section is recognized for its high fish expenditures, followed by groceries, and almost none to zero spending in the other categories. These clients have an average age of 54.5 years, exclusively visit a single store, typically buy products on promotion, and live in coastal locations.

Additional Insights

Based on the conclusions of the segment research, we found that some had comparable tendencies in certain respects, such as the location of the "High Spenders" and "Tech Enthusiasts" as shown in Figure 14, and of ‘College Students’ in Figure 15.

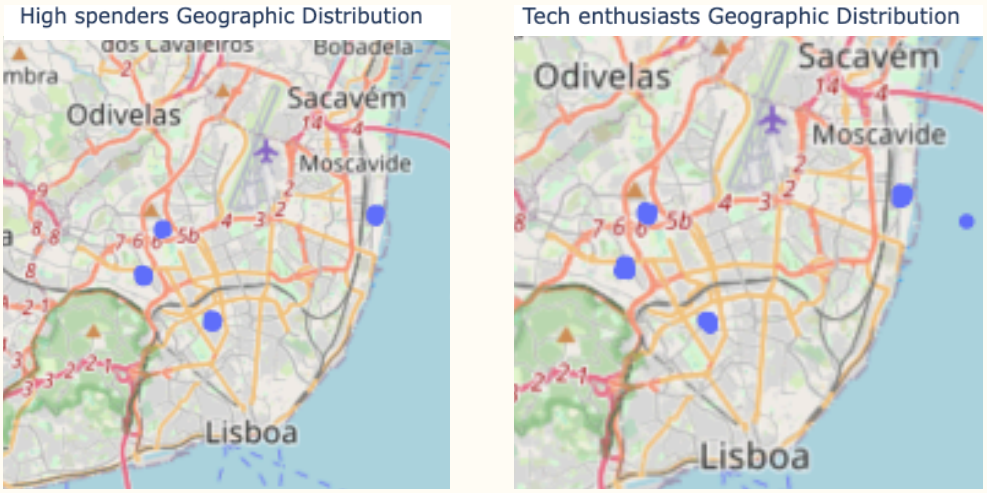

Fig. 14 - Segment Location Comparison

This might be explained by the fact that both segments tend to buy expensive products or products in a greater quantity, therefore it stands to reason that they live in expensive and similarly developed regions of the city.

Regarding ‘College Students’, as said previously, it is possible to see that these customers are spread by the centre region of Lisbon and many are near various universities across the city.

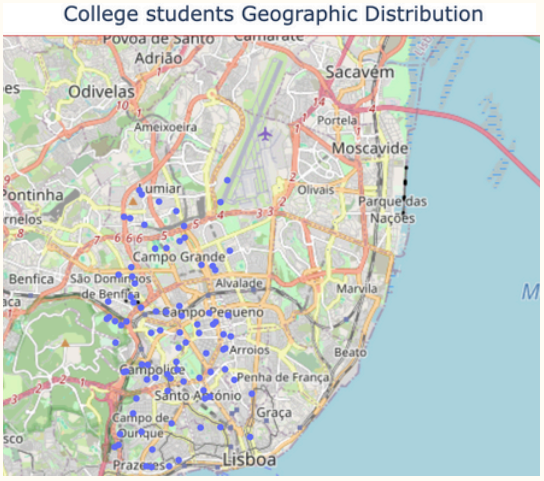

Fig. 15 - ‘College Students’ Location

TARGETED PROMOTION

After identifying different client segments, the next stage was to develop specific promotions and campaigns for each one. For this stage, **association rules** were created using an additional dataset including information on a customer's basket, essentially the products purchased by a specific customer.[1][3]

We have created an approach that incorporates both general and segment-specific marketing for the broad client base and provides templates and design suggestions for the most specific campaigns. For general promotions, we suggest a **birthday discount** campaign, that recognizes our clients by offering them a 10% discount on three in-store purchases made during their birthday month, with coupons sent via email issued a week in advance to remind them of the deal. **Seasonal deals**, such as "Sunshine Specials" and "Frosty Finds," that would offer timely discounts on relevant items, ensuring that our clients have what they need for each season, with a focus on popular products in their respective locations are also recommended. Finally, the "Early Bird Special" which would provide **loyal customers with unique early access** to the latest items, as well as exclusive discounts, making them feel valued and appreciated for their ongoing support, is encouraged.

The following campaigns provide specific promotions for each segment found.

1. Karens

In regards to the "Karens" segment, because we discovered a strong association between cooking oil and oil or cake, and because we know these consumers often have children at home, marketing concentrating on kitchen essentials may appeal to them. Offering a **"Cooking Essentials"** basket with premium cooking oil, baking goods, and other pantry basics required for a family might entice these clients by highlighting quality and efficiency in cooking.
Also, because these customers are more vocal about their concerns and appreciate promotions, as previously determined, providing a special discount on the "Cooking Essentials" package together with a **satisfaction guarantee** could spark their interest and increase loyalty to the company.

2. High Spenders

High spenders have a significant relationship between champagne and laptop purchases, or other eletronic products such as bluetooth headphones, airpods or cellphones. A campaign like **"Tech & Toast"** may appeal to this demographic. Offering a complimentary champagne glass or a discount on a bottle with the purchase of a high-end laptop or tech product, may tempt these clients. This is a particularly clever promotion because it is unexpeted to pair these products, and yet, the idea of celebrating after a major purchase is not. Therefore we are combining those feelings into a promotion that will overall leave the client satisfied.

3. Vegetarians

We suggest offers for vegetarians who, by our findings in the association rules, tend to appreciate asparagus, tomatoes, mashed potatoes, and carrots. To satisfy their tastes, we provide customized meal combinations and discounts. These include, for those looking for quick and easy solutions, **ready-to-cook vegetable kits** containing fresh asparagus, ripe tomatoes, and carrots, perfect for cooking delicious vegetarian dishes at home.
As mentioned before, these customers tend not to have a loyalty card, and to address this, we suggest enhancing the loyalty program's appeal by offering more targeted rewards that align with their interests, such as a **cumulative points** system for purchasing vegetables or other plant-based products.
Moreover, a campaign emphasizing the association of the tomatoes and melons purchases, while providing value with a complementary product, tempting customers to improve their culinary experience, was developed and found below:

"**Duo Delight:** Purchase a bundle of heirloom tomatoes and melons and get a 25% discount on a bottle of balsamic glaze. Elevate your summer salads and fruit buffets with this ideal match!"

4. Pet Lovers

Unfortunately, the invoice information, which contained information on these clients basket and the products they purchased, was not included in the dataset.

Nevertheless, since we examined their typical purchasing behaviors, we suggest a **loyalty program** that offers **discounts on future pet-related purchases** or a subscription **service for regular delivery of pet food** might help to engage this population. This strategy capitalizes on their large pet-care spending and need for high-quality, convenient items.

5. Tech Enthusiasts

The "tech enthusiasts" clients have a strong association between cooking oil and cake and candy bars. A marketing campaign labeled **"Gadgets and Goodies"** might appeal to this audience by joining their need for basic and enjoyable goods and their love for tech products. Bundling electronic products with complimentary candy bars or giving discounts by email on eletronic products may appeal to these consumers.

6. Demanding Foodies

For demanding foodies, the association between cake and cooking oil was quite visible. This segment might benefit from a promotion concentrating on unusual baking product combinations. It may be called **"Unexpected Flavors"** or **"Wordly Tastes"** if it uses ingredients from all around the world. This would be a clever strategy since it pairs the recorded purchasing tendencies for cake and cooking oil with their natural interest in discovering and trying new products. We could send an email with the campaign information and display it in many isles across the store, as customers are likely to move throughout the store to purchase their usual varied items.

7. Large Families

The "Large Families" clients have a tendency to purchase products like cooking oil, cake, and oil. Therefore, and keeping our previous conclusions in mind, specific promotions for this segment may include **discounts on bulk purchases**, as large families require larger quantities, for example, offering a **free cake mix** with the purchase of a certain amount of cooking oil.
Another example would be additional **loyalty points** for purchases of these specific products, rewarding their consistent buying patterns.

8. College Students

There was a clear association noticed between the intake of white wine and beer among college students. To target this population, we suggest giving discounts on both beer and white wine. Options such as offering a **10% discount on wine when purchased with a pack of beer** or designing **combo packages** might effectively persuade these students, who appear to appreciate both drinks.

9. Promo Hunters

Based on the "Promo Hunters" shopping behaviours, which depicted an association between cooking oil/oil and cake, and also interest in gum and candy bars, we recommend the following three types of coupons:

- **Bundle Saver Coupon:** Purchase cooking oil, cake mix, and oil together and receive a discount of 15% on the total price.
- **Oil Essentials Discount**: Provide a voucher for a 10% discount on oil and cooking oil purchases. This addresses their regular purchases of these core products and encourages them to stock up.
- **Sweet Treats Deal:** Coupon that allows people to buy two packets of gum/candy bars and receive one free.

We believe that this promotion will be highly effective given their affinity for discounts in general. Since these promotions align closely with their usual shopping behaviors, we anticipate them to be even more enthusiastic about taking advantage of the offer.

10. Fishy

The “Fishy” clients had a significant association between fresh tuna and salmon or shrimp in their customer baskets. Hence, developing a promotional campaign that offers a **complimentary packet** of any of these fish types **when purchasing three** could be strategic. Introducing a **'4 for 3'** offer enables clients to buy any three packets of either tuna, salmon, or shrimp and receive a fourth, free of charge.

CONCLUSION

The main goal of this project was to utilize customer segmentation to reinforce our understanding of consumer behaviour. Through a meticulous analysis of customer demographics, spending habits, purchase behaviour, and other factors, we were able to identify distinct customer segments, providing interisting insights and tailored business strategies.

Our approach began with exploratory data analysis, followed by segmentation, and concluded with association rules to help in the creation of targeted promotions.

Our results for this client base concluded of the following ten segments: 'Promo Hunters', 'High Spenders', 'Vegetarians', 'Pet Lovers', 'Tech Enthusiasts', 'Demanding Foodies', 'Large Families', 'College Students', 'Karens', and 'Fishy' customers. Each segment exhibited unique traits and spending patterns, providing valuable insights into their motivations and preferences. From these findings, we were able to develop specific targeted promotions and campaigns designed to address the unique needs and preferences of each group, thereby enhancing customer engagement and loyalty.

This project has demonstrated the influence of customer segmentation in driving business success. By gaining a deeper understanding of our customer base, more effective and personalized marketing strategies can be created, ultimately leading to increased customer satisfaction and long-term business growth.

REFERENCES

[1] Fernando Bação, Ivo Bernardo (2024). Machine Learning II. NOVA IMS, Universidade Nova de Lisboa.

[2] Han, J., Kamber, M., & Pei, J. (2011). Cluster Analysis: Basic Concepts and Methods. Data Mining: Concepts and Techniques (3rd ed., pp. 447-493).

[3] Han, J., Kamber, M., & Pei, J. (2011). Mining Frequent Patterns, Associations, and Correlations: Basic Concepts and Methods. Data Mining: Concepts and Techniques (3rd ed., pp. 243-275).

ANNEXES

Exploratory Data Analysis

Fig. 1 - Pie Chart Distribution

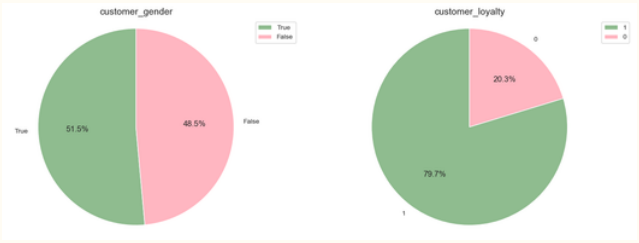

Fig. 2 - 'Fishy' Pie Chart Distribution

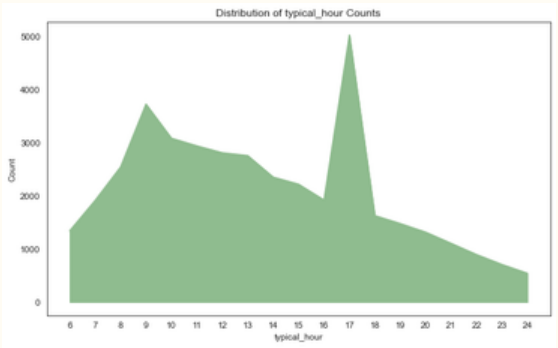

Fig. 3 - Typical hour of Store Visits

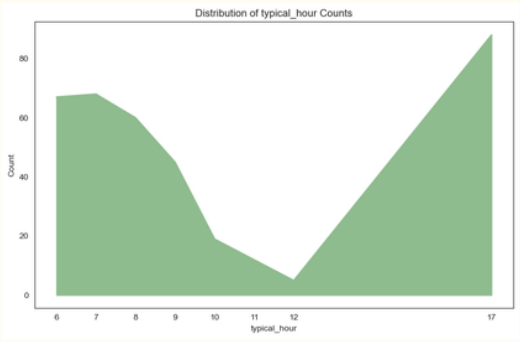

Fig. 4 - 'Fishy' Typical hour of Store Visits

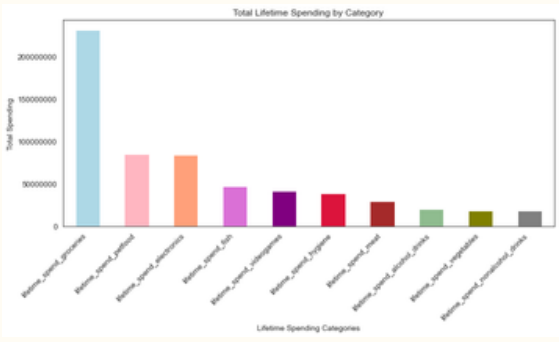

Fig. 5 - Lifetime Spending per Category

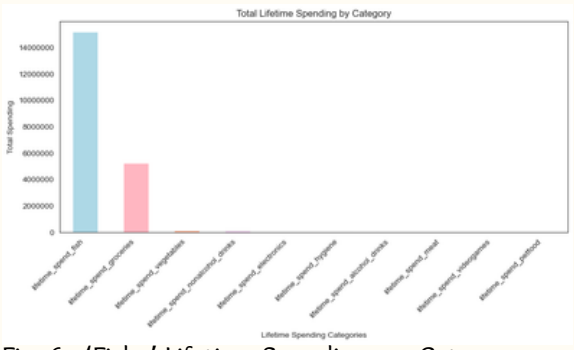

Fig. 6 - 'Fishy' Lifetime Spending per Category

ANNEXES

Exploratory Data Analysis

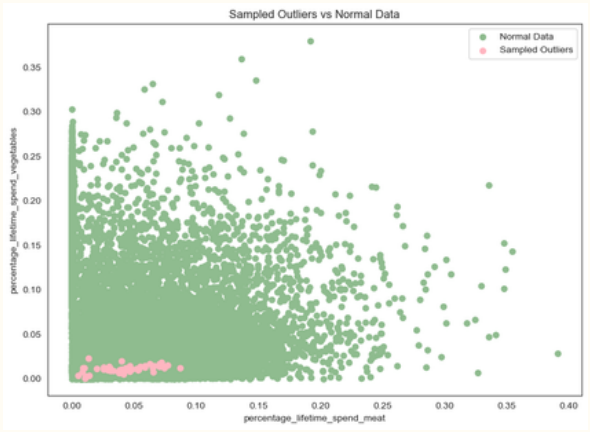

Fig. 7 - Percentage_lifetime_spend_meat and vegetables outliers

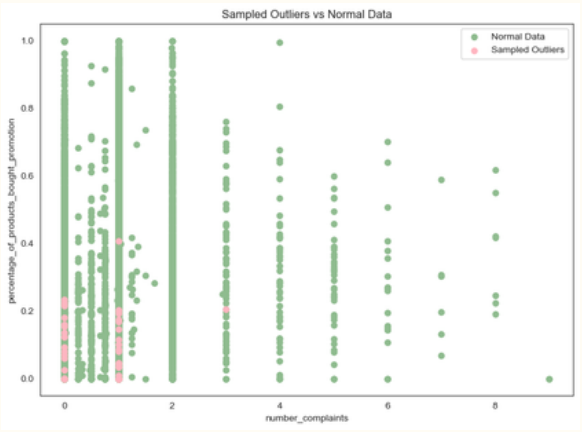

Fig. 8 - number_complaints and percentage_of_products_bought_promotion outliers

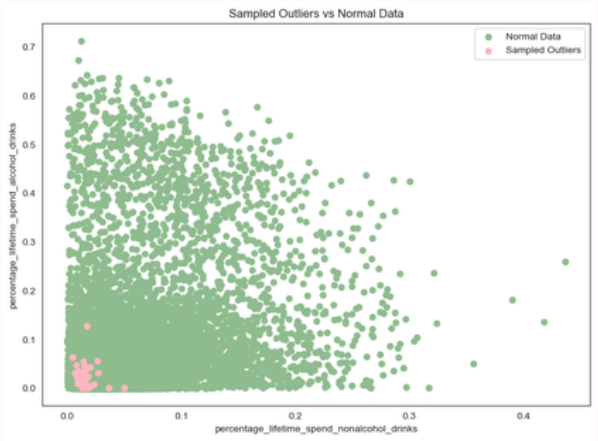

Fig. 9 - percentage_lifetime_spend_nonalcohol and alcohol_drinks outliers

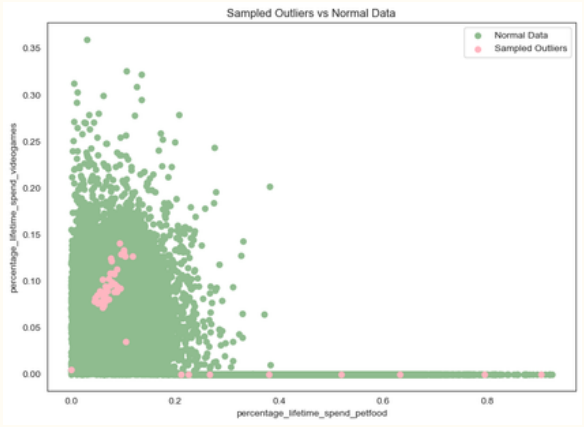

Fig. 10 - percentage_lifetime_spend_petfood and videogames outliers

ANNEXES

Customer Segmentation

Ward with MinMax scaler (chosen method) graphs:

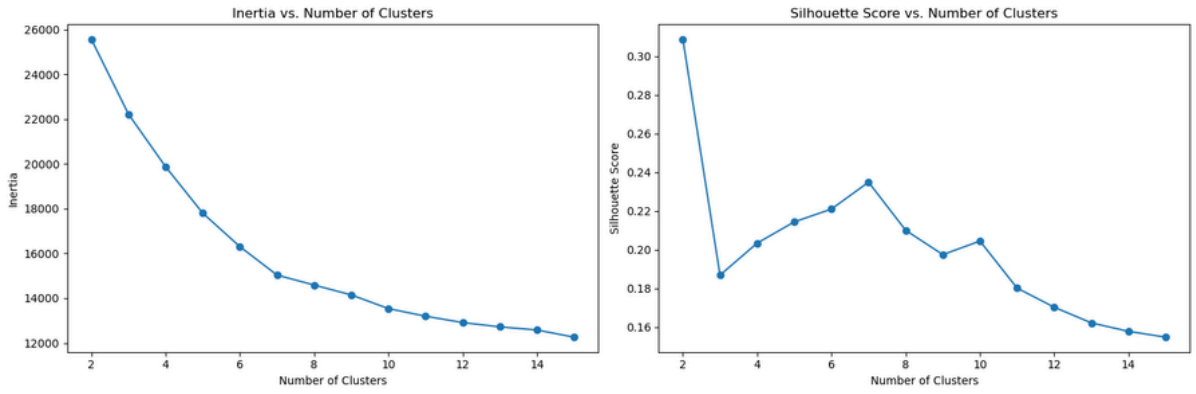

Fig. 1 - Inertia and Silhouette Score Plots

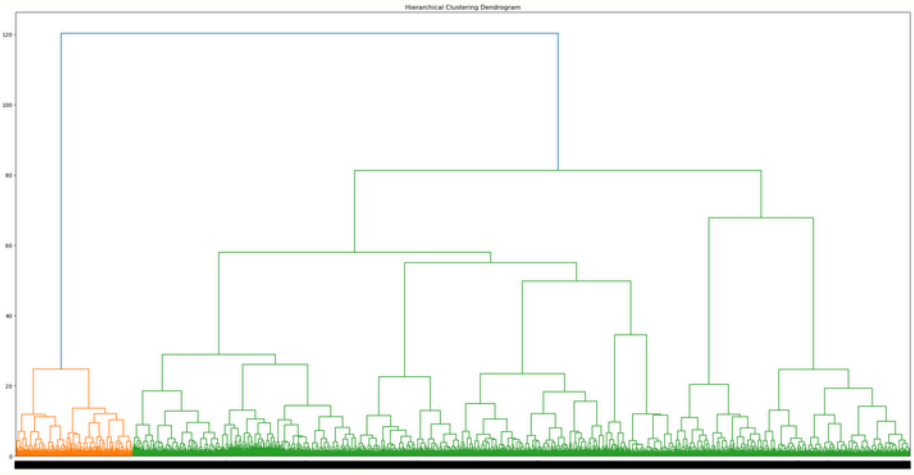

Fig. 2 - Dendrogram for Ward method with MinMax scalled data

UMAP's of different clustering methods

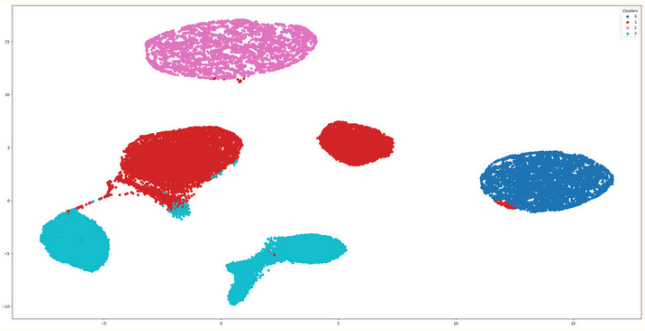

Fig. 3 - Ward with no Scaler

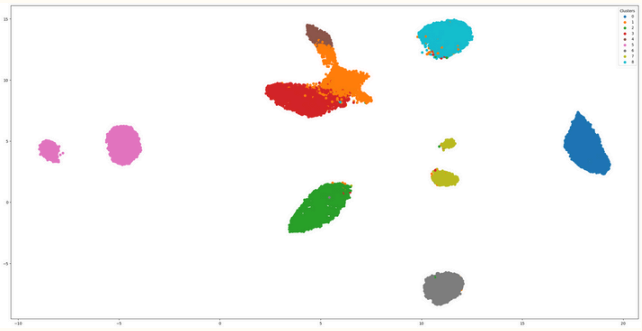

Fig. 4 - Ward with Robust Scaler

ANNEXES

Customer Segmentation

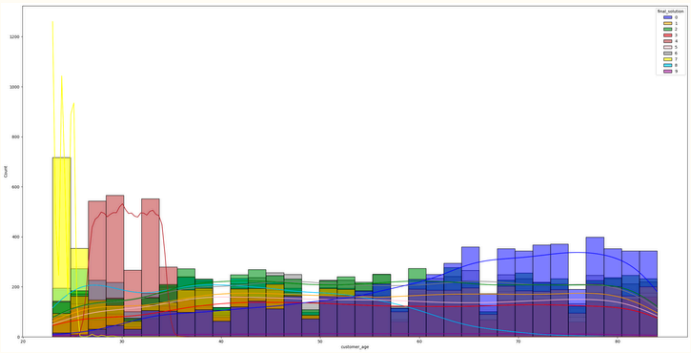

Fig. 5 - Customer Age by the different segments

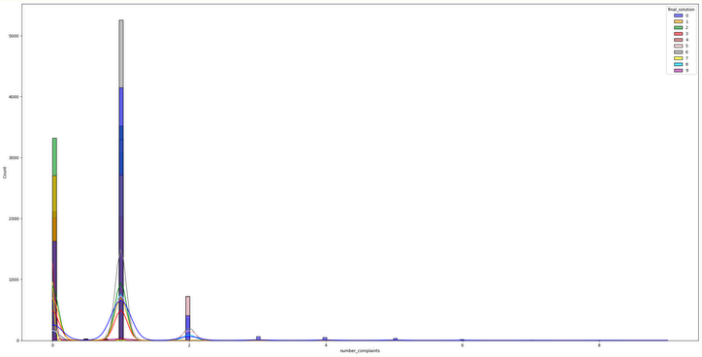

Fig. 6 - Number Complaints by the different segments

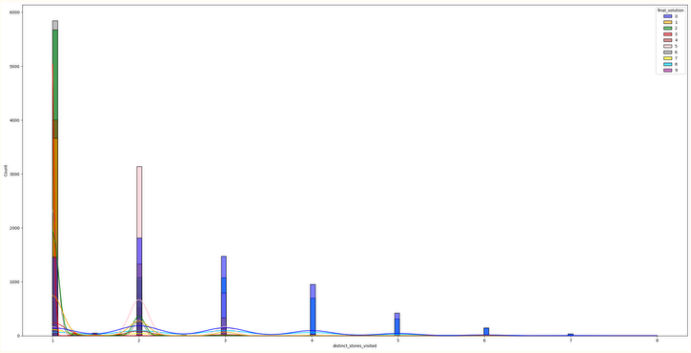

Fig. 7 - Distinct stores visited by the different segments

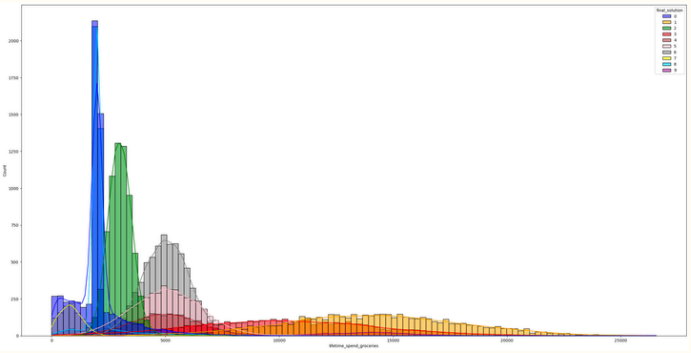

Fig. 8 - Spend Groceries by the different segments

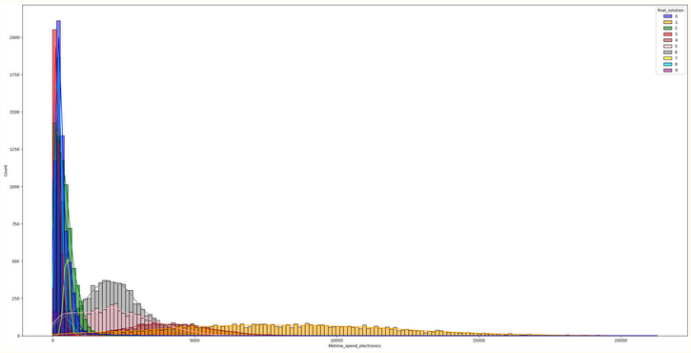

Fig. 9 - Spend Eletronics by the different segments

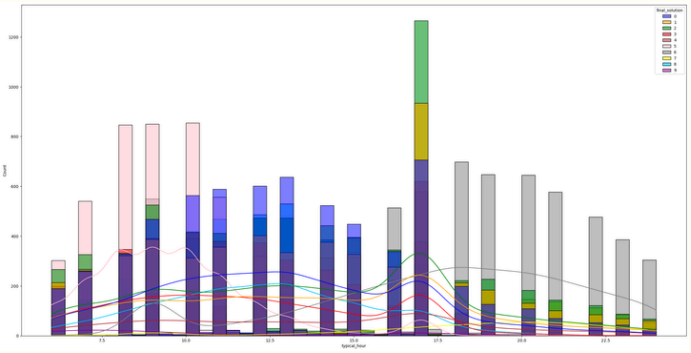

Fig. 10 - Typical Hour by the different segments

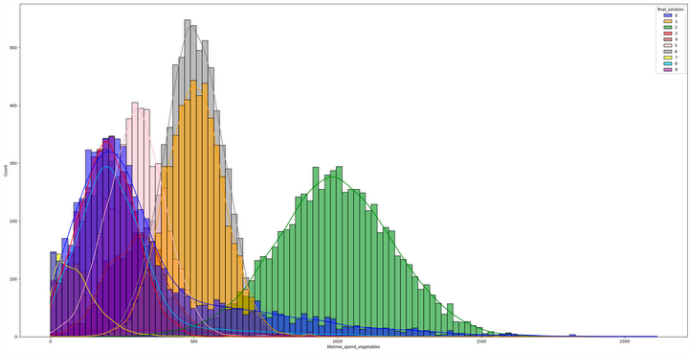

Fig. 11 - Spend Vegetables by the different segments

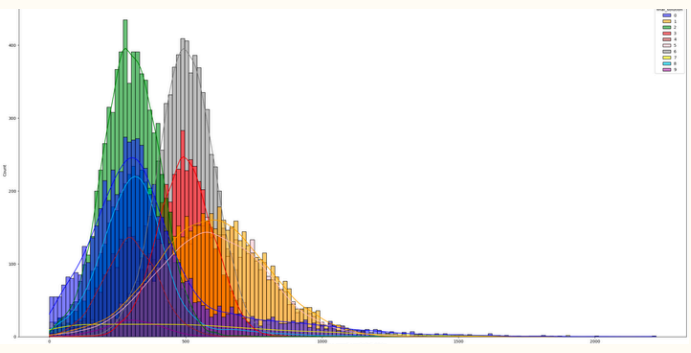

Fig. 12 - Spend NonAlcohol Drinks by the different segments

ANNEXES

Customer Segmentation

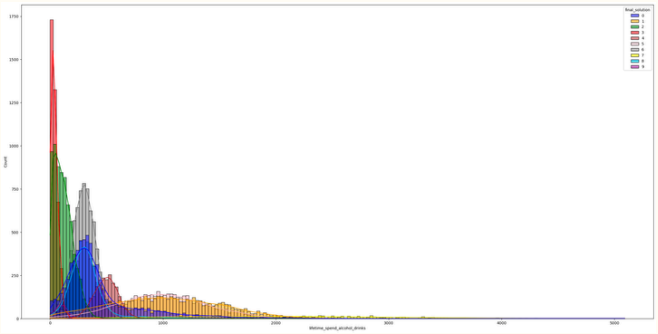

Fig. 13 - Spend Alcohol Drinks by the different segments

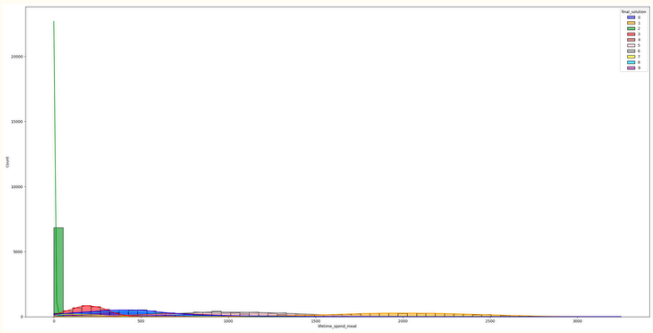

Fig. 14 - Spend Meat by the different segments

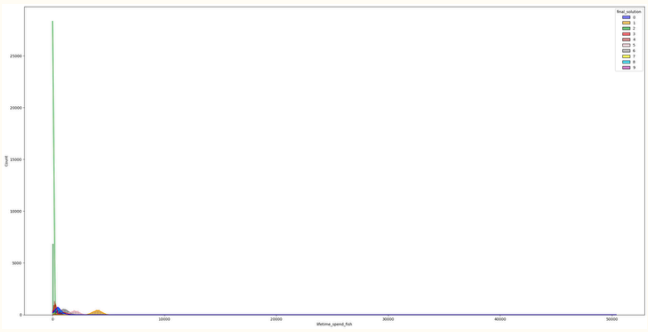

Fig. 15 - Spend Fish by the different segments

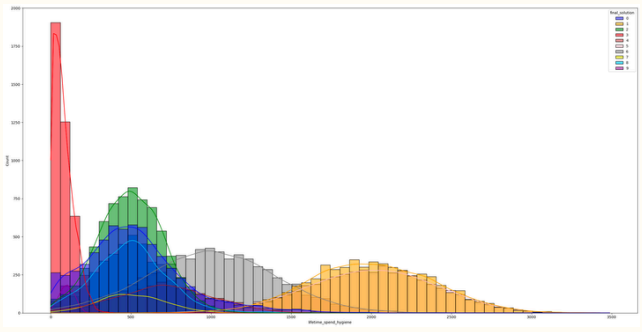

Fig. 16 - Spend Hygiene by the different segments

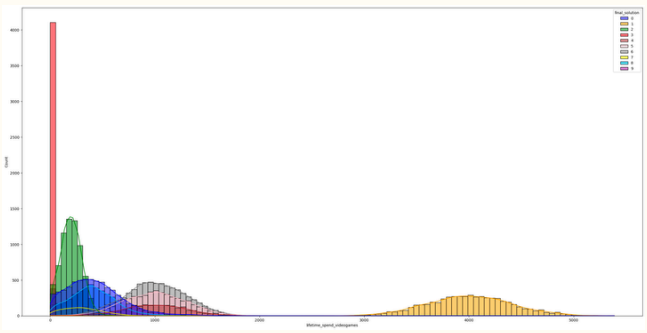

Fig. 17 - Spend Videogames by the different segments

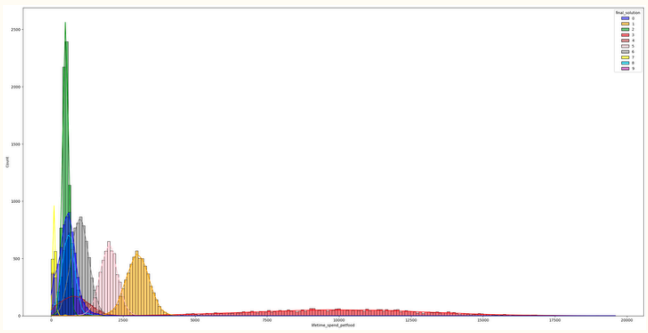

Fig. 18 - Spend Petfood by the different segments

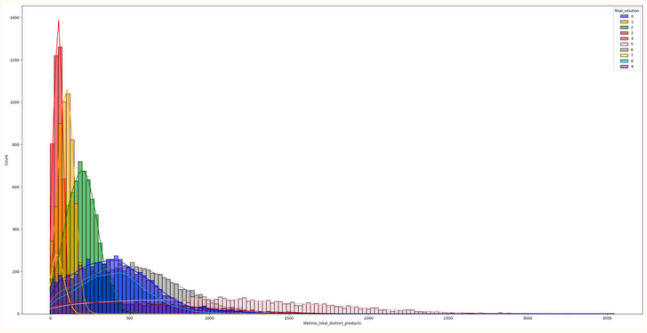

Fig. 19 - Spend Total Distinct Products by the different segments

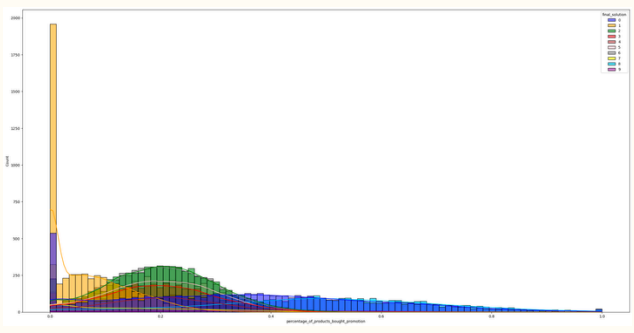

Fig. 20 - Percentage of products bought promotion by the different segments

ANNEXES

Customer Segmentation

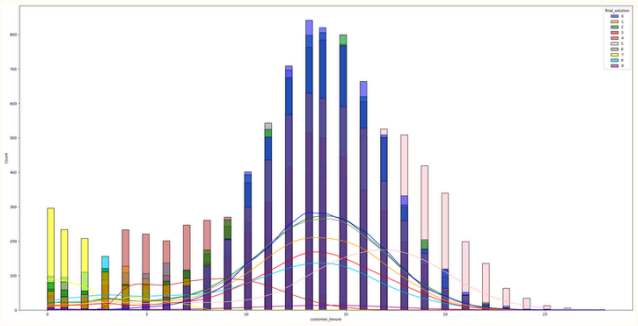

Fig. 21 - Customer Tenure by the different segments

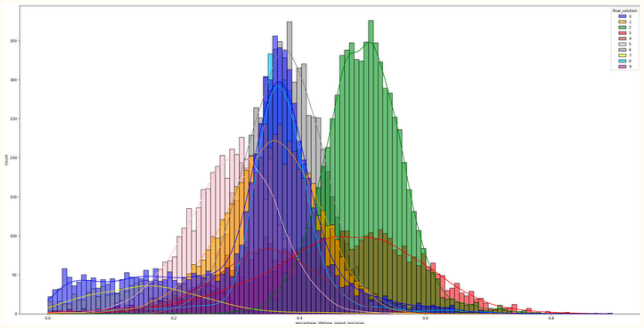

Fig. 22 - Percentage Groceries by the different segments

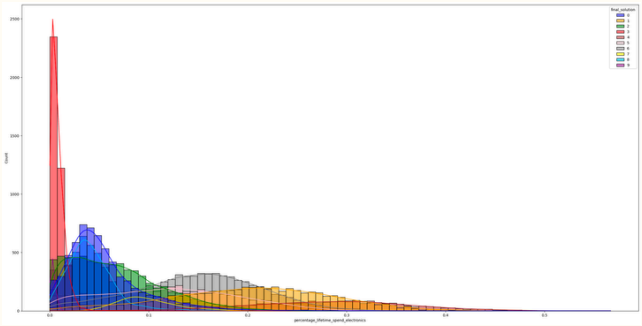

Fig. 23 - Percentage Eletronics by the different segments

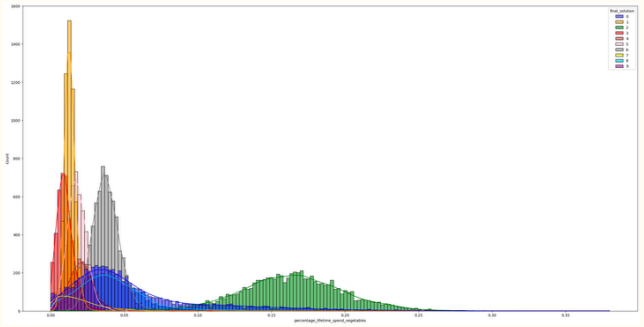

Fig. 24 - Percentage Vegetables by the different segments

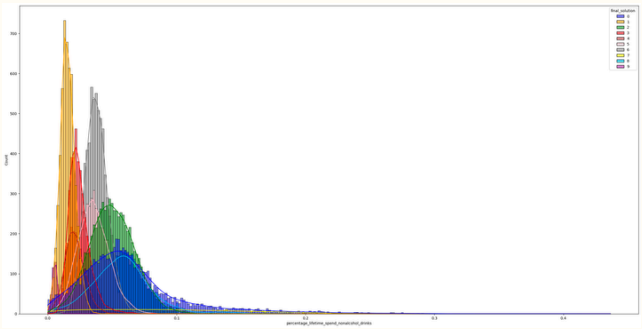

Fig. 25 - Percentage NonAlcohol Drinks by the different segments

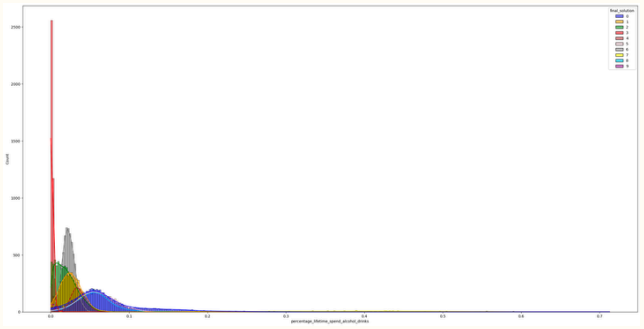

Fig. 26 - Percentage Alcohol Drinks by the different segments

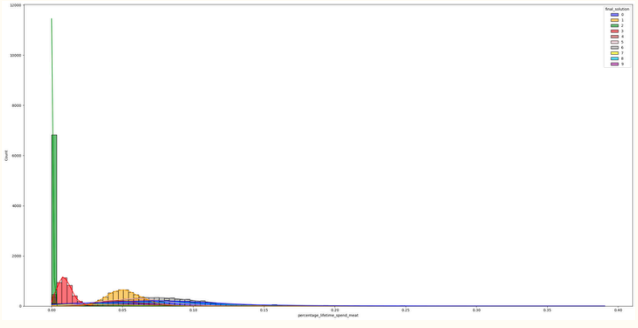

Fig. 27 - Percentage Meat by the different segments

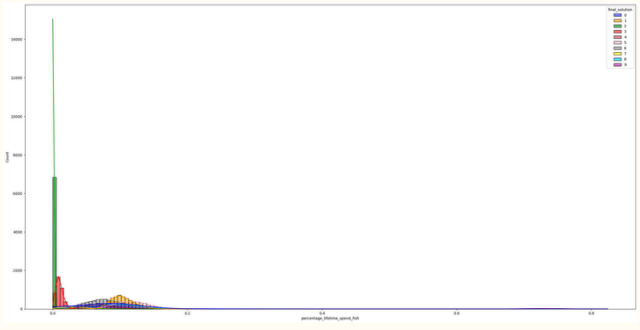

Fig. 28 - Percentage Fish by the different segments

ANNEXES

Customer Segmentation

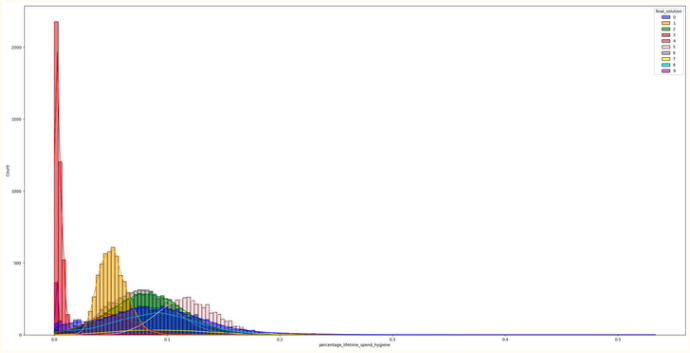

Fig. 29 - Percentage Hygiene by the different segments

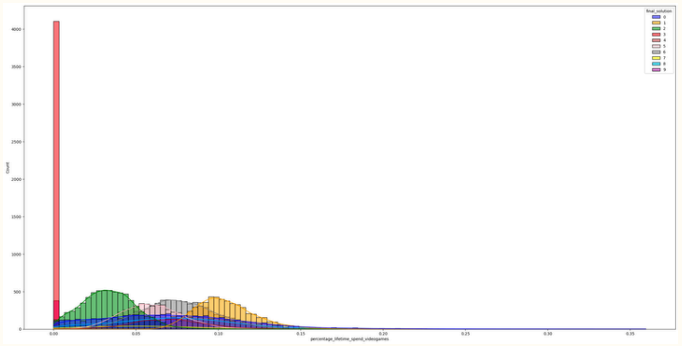

Fig. 30 - Percentage Videogames by the different segments

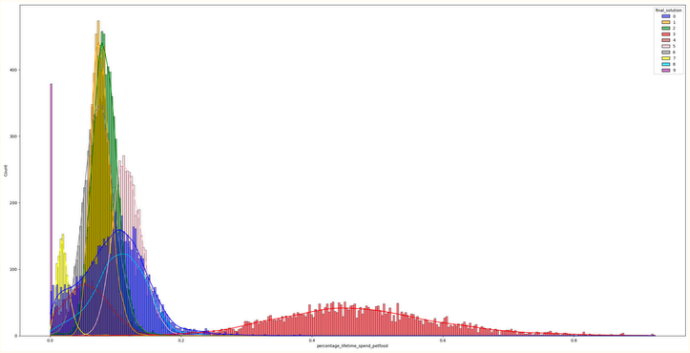

Fig. 31 - Percentage Petfood by the different segments

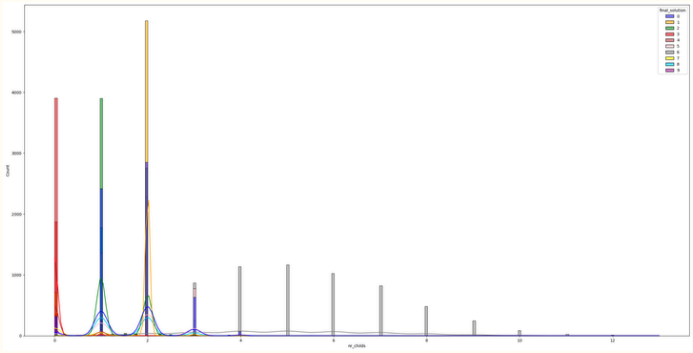

Fig. 32 - Nr Childs by the different segments

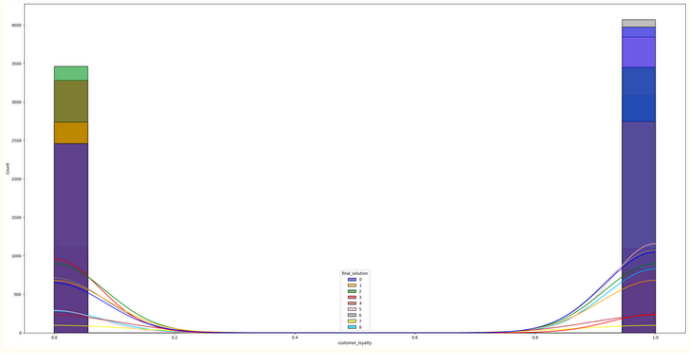

Fig. 33 - Customer Loyalty by the different segments

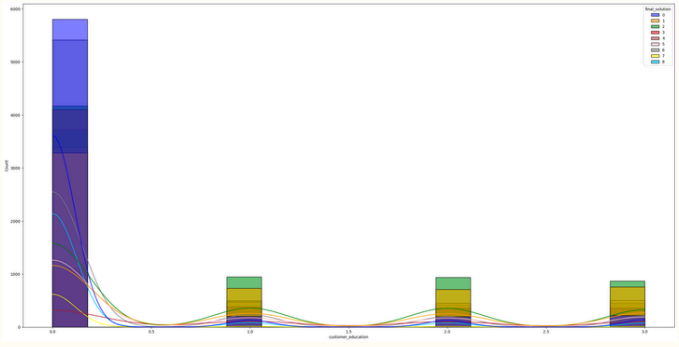

Fig. 34 - Customer Education by the different segments