NOVA IMS
Information Management School

CUSTOMER SEGMENTATION

A KEY TO UNLOCKING BUSINESS GROWTH AND SUCCESS

MACHINE LEARNING II PROJECT

Carolina Almeida | 20221855
Marta Monteiro | 20221954
Teresa Simão | 20221873

Table of Contents

1. Executive Summary

In today's fiercely competitive business landscape, success hinges on understanding customers and tailoring marketing strategies to meet their unique needs and preferences. Customer segmentation, the practice of categorizing clients into specific sub-groups based on shared traits like demographics and behaviors, offers invaluable insights for businesses. This approach allows companies to develop targeted marketing campaigns, enhance customer engagement, and foster lasting customer loyalty. Recognizing that a one-size-fits-all marketing strategy is often inefficient, customer segmentation proves essential in effectively meeting diverse client expectations while optimizing available resources.

The main objective of this project is to conduct customer segmentation by identifying distinct customer groups based on common characteristics. The focus is on addressing the following business questions:

- Identifying relevant customer segments: Utilizing statistical and machine learning methods, the aim is to pinpoint significant segments within the customer base that exhibit similar traits.
- Analyzing customer behavior: Once customer segments are identified, the next step involves delving into their behavior to gain insights into their motivations, preferences, and needs.
- Developing targeted marketing strategies: The goal is to create tailored marketing strategies that resonate with the unique needs and preferences of each customer segment. This could entail devising personalized promotions, launching focused advertising campaigns, or adjusting product offerings to align with specific customer requirements.

In addressing the specified questions, our methodology follows a structured approach:
During the data preprocessing phase, the dataset is meticulously organized to align with our requirements. This involves rectifying missing values and inconsistencies, as well as conducting feature engineering. By leveraging various data visualization techniques, initial insights are uncovered. Moving into the data modelling phase, the data is standardized and subjected to diverse clustering techniques aimed at effectively defining customer segments. Subsequently, an in-depth cluster exploration is conducted to analyze the characteristics and behaviors of the identified clusters, while also pinpointing any outliers. Taking a step further, association rules are generated to unveil the preferences unique to each cluster. This critical insight plays a pivotal role in devising tailored campaigns and strategies that cater specifically to the distinct needs of each customer segment.

The customer segmentation exercise revealed eight distinctive clusters among the customer base. The "Big Spenders" maintain an average spending level while displaying a tendency to purchase a high number of products; The "Pet Lover," prioritize spending on pet food while showing lower expenditures on alcohol and video games; "Savings Squad" excel in making purchases during promotional events; The "Veggie Society" demonstrate a preference for vegetables over meat and fish in their spending habits; "Drunkards" stand out for their high purchases of alcohol and consist mainly of younger customers; The "Gamer Community" focus on acquiring video games in their purchases, The "Big Families" are characterized by a higher number of children and complaints within the customer base; Lastly, the "Fishy Pals" comprise customers inclined towards purchasing fish products. These clusters were derived using data scaled with Min-Max Scaler and the Ward hierarchical clustering technique, providing valuable insights into customer segmentation and behavior.

To make the most of this segmentation analysis, we created tailored promotions for different customer segments. They focus is on enhancing customer service and product availability to retain current customers and attract new ones.

2. Exploratory Data Analysis and Pre-Processing

Exploratory data analysis (EDA) is a crucial initial step in any research study. The main purpose of EDA is to inspect data for patterns, outliers, and irregularities that can guide the subsequent hypothesis testing. It also offers tools to generate hypotheses by visually representing and understanding the data [1]. For this project, the objective is to analyze customer demographics and behavior using the dataset 'customer_info', which includes client-centric data and expenditure history. Additionally, two other datasets named 'customer_basket' (detailing random transactions) and 'product_mapping' (associating products with categories) were provided but not employed at this stage.

2.1. Preliminary Data Analysis and Arrangements

Preliminary Data Analysis plays a crucial role in ensuring data cleanliness, structuring, and preparation before diving into clustering and deeper customer insights. The process began with the importation of the *customer_info* dataset, followed by the creation of a copy to preserve the original data integrity. Setting the *customer_*id column as the index enhanced dataset readability. Subsequently, columns starting with *lifetime_spend* were streamlined by removing the redundant prefix. The *Unnamed:0* column was removed for the same reason. Sorting the index in ascending order was carried out to enhance organization. To optimize memory usage, some variables were converted to more appropriate data types, aligning them with machine learning model expectations:

Data type conversions are fundamental for memory efficiency and model compatibility, enabling streamlined analysis and modelling processes. Consistent data types also ensure smooth execution of machine learning algorithms, promoting more precise decision-making based on results. While some conversions weren't feasible due to missing values, the closest suitable replacements were applied. This optimization led to a reduction in memory footprint from 10MB to 10.8MB.

On further analysis, it was revealed that there were no duplicate rows in the dataset. However, missing values were present in 8 columns, with percentages mostly at or below 4%, except for *loyalty_card_number*, which had a higher percentage at 43.5%. To handle this, the *loyalty_card_number* column was proposed for removal, to be replaced by a new binary column *has_loyalty_card* denoting card ownership status.

A detailed examination of the dataset's descriptive statistics provided initial insights influencing subsequent decisions. The distributions of kids and teens in households were slightly positively skewed, with most households having 1 or fewer kids. Average complaints per customer were around 0.71, and distinct stores visited averaged at 1.67. Notably, a high variability between minimum and maximum values across all categories suggested possible skewness or outliers. Customer purchases of distinct products averaged about 9.38, with a notable standard deviation indicating significant variability. Also, it was observed that transactions spanned from 1996 to 2024, with a loyalty card possession by over 50% of customers.

Inconsistencies in the *percentage_of_products_bought_promotion* variable were noted, with negative values and percentages exceeding 100%, affecting 3352 rows. These were replaced with NaN values, treated as missing data, and handled using K-Nearest Neighbours (KNN) imputation. To determine the optimal number of neighbours, mean square error metrics were used for each column with missing data.

2.2. Feature Engineering

Feature Engineering is a critical process that harnesses domain knowledge to extract, transform, and select features from raw data, enhancing machine learning model performance. By crafting new features and modifying existing ones, feature engineering captures essential data insights to optimize model accuracy and efficiency.

In addition to the previously created variable, several new variables were identified as valuable for analysis:

Variable	Variable name	Description
Education	*Education*	Customers' level of education, categorized into three levels: bachelor, master, and PhD. Those without a specified education level are labelled 'No Degree'.
Children	*Children*	The merging of 'kids_home' and 'teens_home' was done to streamline analysis and modelling processes. This consolidation offers a clearer representation of overall household size.
Age	Age	Age of a customer, derived from their birthdate.
Total Amount of Money Spent	*Total Amount Spent*	Total amount of money spent by customer, across all categories.
Percentages	*percentage_spend_*	Percentage each "spend_" column contributes to the total spend for each row.

***Table 1** – Feature Engineering*

To further delve into these insights, a correlation matrix was constructed using Spearman's correlation, encompassing the newly created variables (*Annex Figure 1*). This analysis unveiled intriguing relationships among different variables within the dataset. Notably, a strong positive correlation was observed between the *Total Amount Spent* and expenditures on groceries and pet food. This correlation implies that households allocating more towards general spending tend to allocate proportionately more towards groceries and pet supplies. Consequently, the former variable will be excluded during the modelling phase.

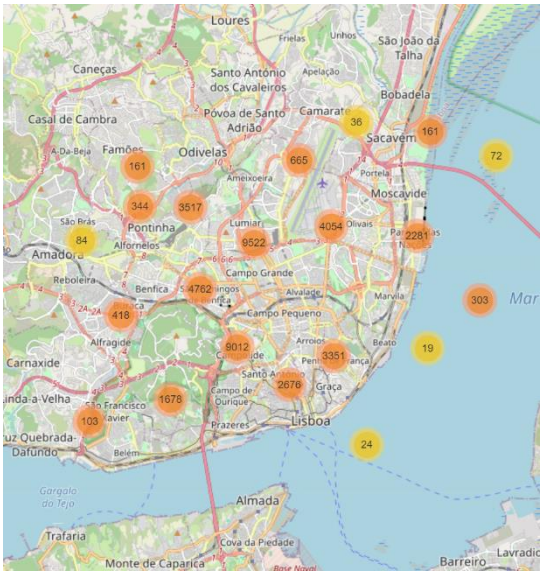
Moreover, a significant correlation was detected between expenditures on meat and fish, highlighting potential interdependencies in consumer preferences within these categories. The expenditure on videogames displayed strong correlations with purchases of electronics, meat, fish, and hygiene products, indicating potential associations between these product categories. These correlations inspired further exploration through visualization, paving the way for a deeper understanding of the relationships between these variables.

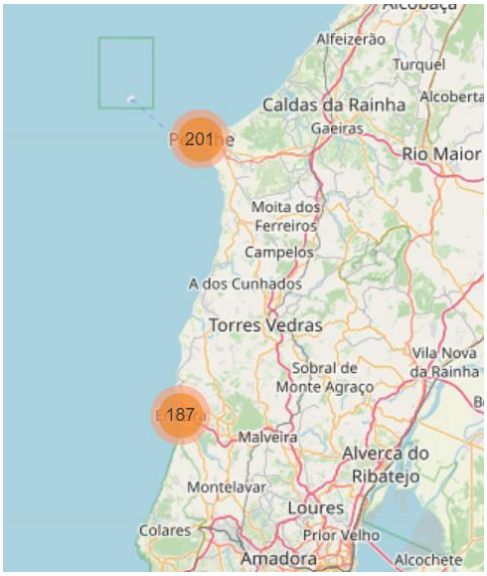
Additionally, as anticipated, a strong correlation was observed between the number of children and the combined count of kids and teens. Furthermore, a high correlation was evident between the proportion of funds allocated to a specific category and the actual amount spent in that particular category.

2.3. Data Visualization

Data Visualization is a powerful tool that enables the identification of trends, outliers, and patterns within datasets through visual representations such as charts, graphs, and maps. This approach simplifies complex data interpretation, aiding decision-making processes and actionable insights.

In light of the frequent occurrence of customers with 'fishy' in their names, a segregation into two distinct groups was implemented: *customers* and those identified as *fishy_customers*. Exploring their respective locations revealed distinct distribution patterns, underscoring the separability of the dataset





Figures 1 and 2 *– customers' location on the left; fishy_customer's location on the right*

Regular customers predominantly inhabit the Lisbon district, with noticeable concentrations in areas like Campolide, Campo Grande, and Benfica. Interestingly, an unusual spatial distribution is observed where some customers appear to be located in the sea – a location that raises questions about data accuracy.

Conversely, customers designated as 'fishy' are dispersed between Ericeira and Peniche, indicating a different geographical distribution compared to regular customers.

2.3.1. Single Variable Analysis

Within the regular customer segment, there is a balanced gender distribution, with 50.2% male customers and 49.8% female customers (*Annex Figure 2*). Education-wise, the majority of both males and females either have no specified degree or lack educational information in the dataset. Other education levels are evenly distributed across genders (*Annex Figure 3*). Most customers have between 0 and 9 children, with the most prevalent number being 1 child. Notably, there are instances of fractional child counts, such as 1.1 children, which have been rounded to the nearest whole number for consistency (*Annex Figures 4 - 5*).

Regarding loyalty card ownership among regular customers, 56.3% do not possess a loyalty card compared to 43.7% with a card, indicating an imbalance that may not be advantageous for the business (*Annex Figure 6*). Transaction trends show peak years for initial transactions in 2010 for females and 2011 for males, aligning with a normal distribution pattern (*Annex Figure 7*). The majority of customers, irrespective of gender, have visited only one store, with a few cases showing fractional store visits that were rounded to the nearest whole number for accuracy (*Annex Figure 8*).

The majority of customers typically have a single complaint; however, instances where customers have 0.7 complaints are not realistic (*Annex Figure 9*). To rectify this inconsistency, these fractional values will be rounded to the nearest integer. Transaction peak times are synchronized for both males and females, with 5 PM being the most favored transaction hour. Conversely, the least popular transaction hours generally fall between 1 PM and 2 PM. All transactions occur within the timeframe of 6 AM to 12 AM (*Annex Figure 10*).

They predominantly allocate a significant portion of their expenses towards groceries (*Annex Figure 11*). A total of 3896 customers, comprising 1922 females and 1974 males, took advantage of a promotional offer on items priced between 0.2525 and 0.2575. This price range delineates a distinct bracket where customers made purchases under the promotion (*Annex Figure 12*).

Among fishy customers, there is a balanced gender distribution, with 51% females and 49% males (*Annex Figure 13*). Similarly, the majority of customers within this segment either have an unspecified degree or have not provided educational background information. Interestingly, males exclusively hold Ph.D. and master's degrees, while only females possess bachelor's degrees (*Annex Figure 14*).

The fishy customers typically have either 0 or 1 child, with a higher prevalence of 0 children. Notably, there are instances of fractional child counts like 1.1 children that have been adjusted to the nearest whole number for accuracy (*Annex Figures 15 - 16*). Additionally, a significant proportion of fishy customers do not possess a loyalty card, with 79.9% falling within this category (*Annex Figure 17*).

Examining the histogram of clients per year of first transaction dataset, a generally standard distribution is observed, barring a singular anomaly where a customer made their initial purchase in 2000. For male customers, the peak initial transaction year is 2008, whereas for female customers, it is 2010 (*Annex Figure 18*).

The majority of customers, both male and female, predominantly visit only one store. However, there are instances of fractional store visits, such as 1.7 stores, which have been rounded to the nearest whole number for practicality (*Annex Figure 19*). Complaint counts display a trend of most customers having either no complaints or a single complaint, with fractional values rounded to the nearest integer for consistency (*Annex Figure 20)*.

Gender distribution across hours exhibits similarities for the most part, with minor variations at specific hours. Notably, the peak transaction hour remains at 5 PM, with transactions observed between 6 AM and 5 PM in the dataset (*Annex Figure 21)*. Spending habits of fishy customers predominantly focus on fish, followed by groceries as the second most significant category. Other product categories, such as videogames and pet food, show minimal to no expenditure (*Annex Figure 22)*. Lastly, males exhibit a preference for promotions in the 0.275 - 0.325 range, while females lean towards promotions in the 0.375 - 0.425 range. This suggests that these customers are inclined to make more purchases during promotional periods in comparison to regular clients (*Annex Figure 23)*.

This analysis starts revealing distinct customer groups based on characteristics like the number of children or buying behaviour during promotions. It provides valuable insights for decision-making.

2.3.2. Multiple Variable Analysis

In terms of buying preferences, the majority of regular customers who purchase fish also tend to opt for meat, showing a comparable number of customers choosing to buy both fish and meat, only fish, only meat, or neither (*Annex Figure 24)*. The prevailing dietary preference among our client base leans towards a carnivorous diet. On the other hand, most fishy customers typically buy both fish and meat, with only a small subset opting for just fish (*Annex Figure 25)*. Interestingly, there are no vegetarians included in this dataset.

For regular customers, across all categories, a positive correlation is noticeable between the number of children and the amount spent per category, except for spending on alcoholic beverages (*Annex Figure 26)*. This suggests that customers with more children tend to consume fewer alcoholic drinks. In the case of fishy customers, a contrasting observation emerges as these customers tend to have minimal or no children. An intriguing finding in the video games category shows that regardless of the number of children, fishy customers consistently spend zero dollars on videogames (*Annex Figure 27)*.

When analyzing regular customers, a scatterplot reveals a negative correlation between age and expenditure on alcohol, indicating that younger individuals allocate more money towards purchasing alcoholic drinks (*Annex Figure 28)*. Conversely, for fishy customers, there seems to be no apparent correlation between age and spending on alcoholic beverages (*Annex Figure 29)*. Here, the data suggests that regardless of age, fishy customers' alcohol consumption remains relatively constant or unaffected by age-related patterns.

As we attempted to explore the correlation between variables identified in the correlation matrix, no meaningful insights were derived due to the absence of data scaling (*Annex Figures 30 – 33)*.

3. Customer Segmentation

Customer Segmentation is a vital process that involves grouping a customer base based on shared characteristics such as demographic details, purchasing habits, geographic location, and more. By categorizing customers into distinct segments, businesses can tailor their marketing strategies to target each group effectively. This approach aims to enhance marketing efforts by reaching out to specific customer segments that are likely to respond positively to tailored strategies.

3.1. Methodology

The process of segmentation began with featuring selection. We opted to drop all categorical variables, remove *kids_home* and *teens_home* since *Children* sums both, and discard *Total Amount Spend* due to its high correlation with *spend_groceries*.

Following this, we decided to form four distinct datasets, one being the original in its untouched unscaled form, while the remaining three employ diverse scaling techniques [2]:

- Standard Scaler – standardizes features by removing the mean and scaling to unit variance. It centers the feature values around 0 and scales them to have a standard deviation of 1, ensuring all features are on a similar scale.
- Min-Max Scaler – scales features to a given range, usually between 0 and 1 (or any other specified range). It rescales each feature by subtracting the minimum value and then dividing it by the range, giving all features a common minimum and maximum.
- Robust Scaler – scales features using percentiles and median instead of mean and variance. It removes the median and scales the data based on the IQR (Interquantile Range), making it robust to outliers by using more robust estimates of central tendency.

Subsequently, we proceeded to evaluate various clustering methods, each of which will be elaborated upon in detail below.

K Means

The k-means algorithm partitions data into k clusters, each represented by a centroid. It aims to minimize intra-group variance by iteratively assigning data points to the nearest centroid and updating the centroids until convergence. The quality of clustering is based on the proximity of points within the same cluster. Adjusting the number of clusters (k) can impact the clustering's effectiveness in reducing the sum of squared errors. [3]

Upon analyzing the dataset without scaling, we observed an 'elbow' in the inertia plot around 5 clusters and a peak in the silhouette plot between 2 to 4 clusters (*Annex Figure 34)*. Given these insights, we decided to prioritize the silhouette plot for our evaluation. Subsequently, we tested the K-means algorithm with 2, 3, and 4 clusters for further assessment. Utilizing the silhouette score and a dedicated plot, we evaluated these cluster configurations (*Annex Figure 35-37)*. Unfortunately, none of the results yielded satisfactory outcomes.

After repeating the process with the other datasets, the following conclusions were drawn:

For standard scaling (*Annex Figures 38 – 41)*: The inertia plot showed an 'elbow' at approximately 6 clusters, with the silhouette plot peaking between 6 to 8 clusters. Consequently, the decision was made to prioritize the silhouette plot. Subsequently, testing the K-means algorithm with 6, 7, and 8 clusters did not yield satisfactory results, leading to the removal of the columns containing the K-means values for each cluster configuration.

For Max scaling (*Annex Figures 42 – 45)*: With the inertia plot indicating an 'elbow' at around 5 clusters and the silhouette plot peaking between 7 to 8 clusters, the focus was shifted to the silhouette plot. Further evaluation with 7, 8, and 9 clusters via the K-means algorithm did not produce desirable outcomes, resulting in the removal of the relevant K-means value columns.

For robust scaling (*Annex Figures 46 – 49)*: The inertia plot exhibited an 'elbow' at roughly 6 clusters, while the silhouette plot peaked between 3 to 5 clusters. Hence, emphasis was placed on the silhouette plot. Testing the K-means algorithm with 3, 4, and 5 clusters did not lead to satisfactory scores, prompting the elimination of the columns containing the K-means values for each cluster count.

Every cluster outcome from the K-means clustering analysis was disregarded since none of the clusters displayed silhouette scores approaching 1. This criterion was established to identify clearly defined and distinct segments within the data.

Hierarchical Clustering

Hierarchical clustering begins by constructing a distance matrix, which computes values by applying a distance function to every pair of objects in the dataset. The Euclidean distance function is a typical choice for this calculation. [4]

Single – Linkage [5]

In single hierarchical clustering, the algorithm initiates with each data point forming its own cluster. Progressively, the two clusters with the smallest minimum pairwise distance are successively merged together. This iterative merging process continues until all data points are encompassed within a single cluster. Single hierarchical clustering method emphasizes linking clusters based on the closest data points, potentially forming elongated clusters and being susceptible to outliers.

In this instance, all dendrograms generated for the four datasets did not offer valuable insights and were not pursued for further exploration (*Annex Figures 50 – 53)*.

Average – Linkage [5]

This methodology calculates the distance between clusters as the average distance among all pairs of data points within the clusters being compared. Known for generating more balanced and cohesive clusters in contrast to alternative linkage methods, average linkage aids in creating meaningful relationships between data points.

The dendrograms generated for standard scaling and min-max scaling did not yield valuable insights and were consequently not further investigated (*Annex Figures 54 – 55)*. The dataset without scaling suggested an optimal number of 3 clusters, while the robust scaling pointed towards 4 clusters. In each case, hypothetical clusters were established and further assessed using a table presenting average values for each cluster and variables, along with UMAP plots. However, both of these hypotheses were ultimately dismissed (*Annex Figures 56 – 57)*.

Complete – Linkage [5]

In this method, cluster proximity is established by considering the most dissimilar members within each cluster. Through the iterative merging of clusters with maximum distances, a hierarchical tree-like structure is shaped, offering insights into the diverse relationships between data points at various resolutions.

Among the dendrograms generated, the one associated with standard scaling was the only one that did not offer insightful information (*Annex Figure 58)*. The optimal number of clusters were determined to be 6 for the dataset without scaling, 5 for min-max scaling, and 6 for robust scaling. However, as before, all these configurations were ultimately discarded (*Annex Figures 59 – 61)*.

Ward – Linkage [5]

Ward's method calculates the increase in variance that occurs when merging clusters at each step and selects the merge that minimizes this increase. Known for generating compact, spherical clusters, Ward's method is less susceptible to outliers compared to other clustering techniques.

The dendrograms created for the four datasets indicated that the ideal number of clusters were 3, 8, 7, and 7, respectively (*Annex Figures 62 – 65)*.

3.2. Results

After a thorough analysis, it was determined that the most effective method for segmenting customers was the Ward hierarchical clustering technique with robust scaling, resulting in the creation of 7 clusters.

The plot reveals that three clusters exhibit clear separation, while the remaining clusters show less distinct boundaries, possibly indicating the presence of outliers.

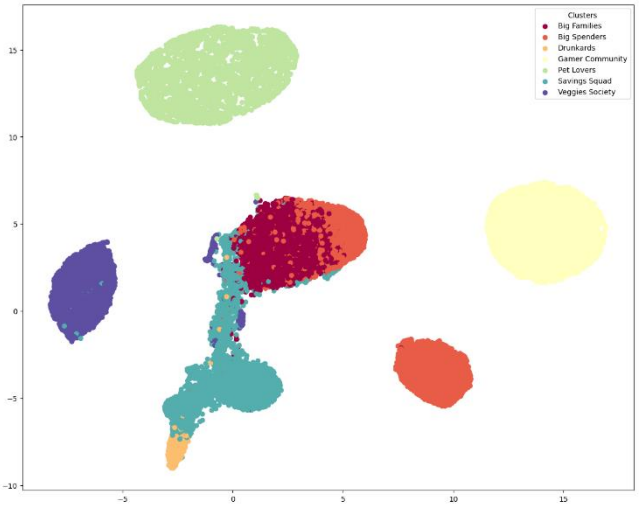


Figure 3 *– UMAP Plot*

Subsequent analysis will delve deeper into outlier detection as part of the upcoming phase of the study.

It should be noted that the *fishy_customers*, handled during the data preprocessing stage, are also accounted for as a distinct cluster. Consequently, the total number of clusters amounts to 8, each defined below:

- **Big Families:** This cluster stands out for having the highest count of children and complaints compared to the other clusters. Typically, these customers shop later. The cluster composition includes slightly more females than males.
- **Big Spenders:** This cluster is notable for its higher total of distinct products purchased. On average, customers in this cluster visit more than one store, likely in search of a variety of products. They tend to shop earlier, possibly to avoid long queues or crowd congestion.
- **Pet Lovers:** This cluster exhibits the highest expenditure on pet food while showing zero expenditure on videogames and minimal purchases of alcohol. The gender composition leans slightly towards females, with a higher percentage of individuals holding a master's degree, possibly in a field related to biology.
- **Savings Squad:** This cluster is characterized by the highest number of discounted product purchases. These customers typically visit more than two stores on average, likely seeking optimal promotions. The gender distribution slightly favors females. Intriguingly, this group also registers a significantly high number of complaints, prompting a deeper exploration that resulted in the creation of a sub-cluster. The **Karens** sub-group comprises individuals with 4 or more complaints, all situated within the *Savings Squad* cluster. In total, there are 126 customers classified as Karens.
- **Veggies Society:** This cluster demonstrates minimal expenditure on meat and fish, although this does not necessarily contradict a vegetarian or vegan diet, since customers may purchase these products for their families. They notably allocate a considerable budget towards vegetables. The cluster shows an equal distribution of customers possessing and lacking loyalty cards, positioning them as one of the less loyal clusters within the segmentation.
- **Drunkards:** These customers are the youngest, averaging 23 years of age, and allocate a significant portion of their budget to alcohol. Being among the most recent customers, likely owing to their age, they also rank as one of the least loyal customer segments. This aligns with the expectation that younger individuals may be less inclined to acquire a loyalty card.
- **Gamer Community:** This group shows a high expenditure on videogames and electronics, as well as on non-alcoholic drinks. Similarly, a significant portion of their customers do not possess a loyalty card.
- **Fishy Pals:** This cluster, previously identified as the fishy_customers, has distinct coordinates compared to other clusters. With their spending habits, they predominantly allocate their budget towards fish purchases. Remarkably, most individuals within this cluster hold a Ph.D.

Please note that all plots showcasing this information will be presented below in the annexes section.

Attached below is the plot illustrating the sizes of each cluster. It is evident from the plot that the *Savings Squad* comprises the largest number of customers, while the *Fishy Pals* cluster has the fewest members.

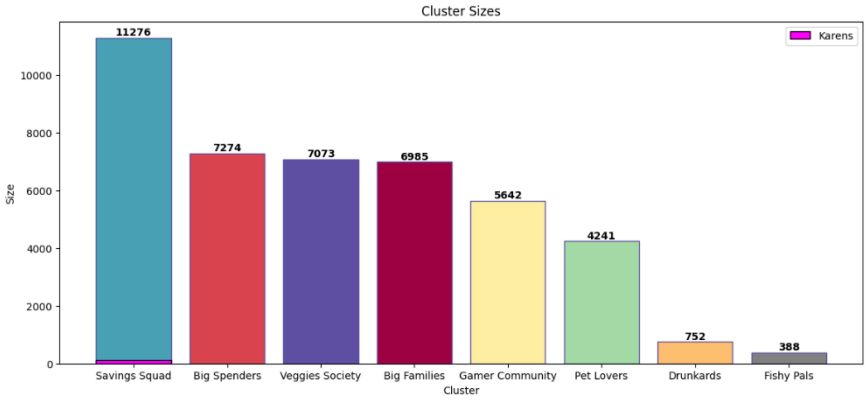


Figure 4 *– Cluster Sizes*

Furthermore, we conducted an examination of the customers' locations. It is essential to mention that this assessment is founded on the average coordinates of each cluster, implying that there might be slight variations in the locations of individual customers within a cluster.

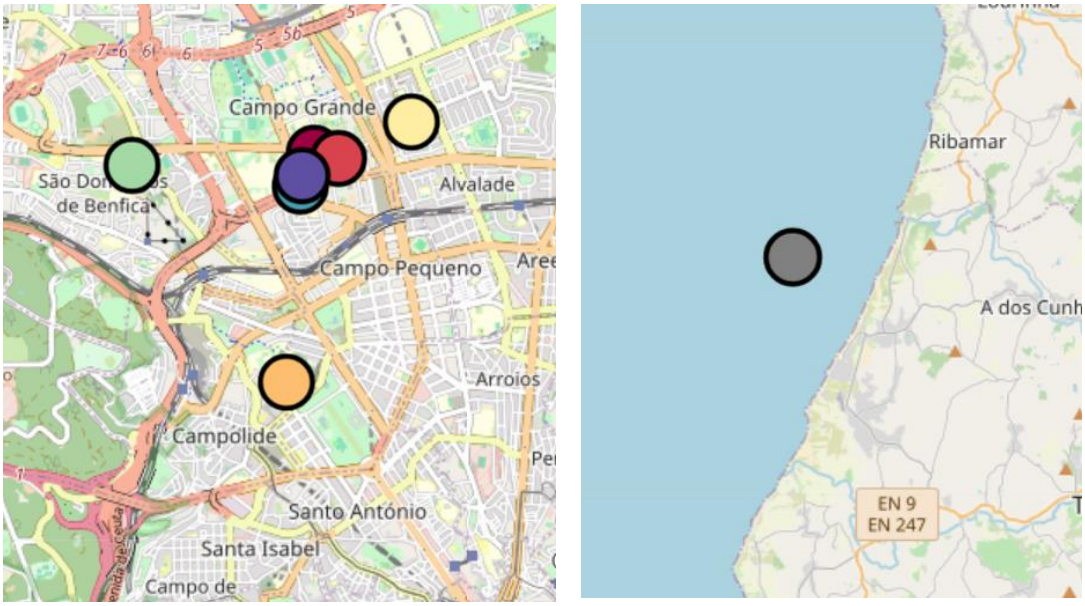


Figure 4 *– Clusters' location*

The *Fishy Pals* cluster is situated between Ericeira and Peniche. On the other hand, the *Pet Lovers* cluster, represented in green, is located in Jardim das Laranjeiras, in proximity to a park and a veterinary office. The *Savings Squad* cluster, marked in blue, resides in Avenidas das Forças Armadas. Moving on, the *Veggie Society* cluster, designated in purple, is positioned in Campo Grande near Cidade Universitária. Both the *Big Families* cluster, depicted in bordeaux, and the *Big Spenders* cluster, in red, are also located in Campo Grande. The *Drunkards* cluster, distinguished in orange, resides near NOVA IMS, potentially linked to their study location. Lastly, the *Gamer Community* cluster, shaded in yellow, lives in close proximity to Alvalade.

In concluding this chapter, we opted to conduct an analysis of outliers, *Annex Figure 87*, identified during our cluster analysis. It was noted that certain clusters lack clear definition, potentially due to the presence of outliers.

For example, the *Savings Squad* cluster displays outliers in alcohol spending, which could obscure its distinction from the *Drunkards* cluster. Similarly, outliers in the number of children within the *Savings Squad* cluster may impact its differentiation from the *Big Families* cluster. Furthermore, the plot depicting the treated percentage of products bought on promotion highlights outliers in both the *Big Families* and *Drunkards* clusters, creating challenges in distinguishing them from the *Savings Squad*.

Despite the presence of outliers, a decision was made not to address them, as we are confident that our unsupervised learning algorithm has provided a satisfactory solution.

---

4. Targeted Promotions

In essence, association rules involve identifying patterns that highlight frequent co-occurrences of specific product sets. It's crucial to note that these rules do not imply causation but rather indicate the likelihood of certain items being purchased together. The evaluation of these rules is typically based on three key metrics: confidence, lift, and support.

An interesting observation was noted with products like cooking and other oil variants emerging as top-selling items, resulting in their frequent appearance in many association rules. As a result of their commonness, these items were often overlooked during the rule analysis process.

Lastly, it's crucial to highlight that no association rules exist for the *Pet Lovers* cluster due to the absence of recorded transactions. Therefore, general promotions were devised for these customers.

Furthermore, for each of the clusters, we will focus on the rules with the highest lift.

Big Families

Their purchases primarily consist of cake and baby food.

Big Spenders

Their purchases primarily consist of cake and baby food.

Given the similarity in purchase history among these clusters, they will be targeted with the same promotions.

Family Feast	Enjoy a special discount on family-sized meals and receive a complimentary cake for every family member!
Bake and Serve	Receive a 15% discount when buying cake mix, cooking oil, napkins, and oil together.
Family Favorites Bundle	*Enjoy 10% off when purchasing a package of cake mix and baby food and receive a surprise gift for every child!*

Pet Lovers

While transaction details are unavailable, it's known that this cluster demonstrates significant expenditure on pet food.

Paw Party	Buy 2 pet toys, get 1 free
Furry Feasts	Enjoy 20% off on all pet food and treats

Savings Squad

Their purchases are generally straightforward, comprising oil and candy bars, but they notably stand out for their high frequency of buying items on promotion. Additionally, they are known to visit multiple stores for their shopping needs.

Multi-Store Explorer Deal	Shop at more than one store in one day and receive a €10 voucher for your next purchase.
Karen's Kompensation	Receive a 15% discount on your next purchase in exchange for feedback
Promotion Hunting	Unlock exclusive discounts through a loyalty program by scanning the app at checkout, earning points for every item purchase!

The implementation of the last two promotions underscores our commitment to enhancing the customer experience at the company's sites. We recognize that improving customer satisfaction is paramount for the growth and success of the business**.**

Veggies Society

Their transactions primarily involve mashed potatoes, tomatoes, asparagus, and carrots.

Garden Fresh Quartet	This month's veggie box: Enjoy a selection of carrots, asparagus, and tomatoes!
Veggie Mania	Stick with the loyalty card, score double points on all veggie buys, and snag surprise veggie treats if you're our top earner of the month!
Veggie Collaborations	Collaborate with influencers in the vegetarian/ vegan community to promote the store, in exchange receive a personalized basket with vegetables of choice.

The second promotion aims to encourage customers from this less loyal cluster to stick with the loyalty card.

Drunkards

Their purchases consistently feature white wine, cider, beer, and dessert wine. They're youthful and haven't yet opted for loyalty cards.

Back to School	Buy any beer and get 15% off on white wine during the first week of classes!
Exam Week Special	Stock up for study sessions: Buy two bottles of wine get a free beer!
Sip & Sample Sessions	Adhere to the loyalty card and receive invitations to exclusive tastings and workshops.
Santos Specials	During the month of June, buy one pack of beer and get a litre bottle for free!

Gamer Community

Their purchases primarily consist of laptops, samsung galaxy 10 devices, bluetooth headphones, and a significant amount of champagne. Additionally, they belong to one of the least loyal customer clusters.

Pop & Shop	Buy champagne and get a discount on your next tech purchase. The more you buy, the bigger the discount!
Tech VIP Nights	Enjoy exclusive champagne tastings paired with tech demos by adhering to the app!

Fishy Pals

Their purchase history includes salmon, fresh tuna, and shrimp. They are located between Ericeira and Peniche.

Tuna Tuesdays	Enjoy special discounts on all tuna purchases every Tuesday!
Seafood Trio	Choose any combination of tuna, shrimp and salmon and get a special discount on your purchase.
Local Celebration	Celebrate the flavors of Ericeira and Peniche with special events featuring local seafood delicacies. Loyalty card holders receive priority access!

5. Conclusion

The study aimed to identify various customer groups with distinct characteristics, establish association rules among them, and subsequently develop tailored marketing strategies for each group. This goal was effectively achieved by applying machine learning techniques such as exploratory data analysis (EDA) involving feature engineering and data visualization, as well as scaling and utilizing unsupervised learning algorithms like k-means and hierarchical clustering.

Among these methods, ward hierarchical clustering with robust scaling yielded the best results. The analysis revealed eight distinct customer groups: **Big Families** which comprise customers with a large number of children, requiring products in bulk; **Big Spender**, a group in which customers purchase a wide variety of products in significant quantities, indicating a higher level of expenditure; **Pet Lovers** that allocate a notable portion of their budget to pet food; **Saving Squad**, a group in which the customers make multiple purchases during promotions or special deals; This clusters contains a sub-group, **Karens**, which are customers with more than three complaints; **Veggies Society**, that prioritize vegetables in their spending, with minimal expenditure on meat and fish; **Drunkards**, contains the youngest customers who purchase a substantial amount of alcohol; The **Gamer Community** comprises enthusiasts of electronics and videogames; and **Fishy Pals** focus their spending on fish products, and are particularly situated between Ericeira and Peniche.

Once the clusters were established, we analysed the customer basket to identify purchasing patterns within each group. These insights were essential to strategically plan marketing campaigns to attract new customers and retain the existing ones.

This study's success underscores the potential for more targeted and personalized retail marketing strategies. Additionally, it highlights the importance of Unsupervised Learning techniques in optimizing sales approaches.

6. References

[1] Komorowski, Matthieu; Marshall, Dominic; Salciccioli, Justin; Crutain, Yves. “Exploratory Data Analysis” Secondary Analysis of Electronic Health Records, September 2016, https://www.researchgate.net/publication/308007227_Exploratory_Data_Analysis

[2] Sharma, Vinod. “A Study on Data Scaling Methods for Machine Learning” International Journal for Global Academic & Scientific Research, February 2022,

https://www.researchgate.net/publication/360109009_A_Study_on_Data_Scaling_Methods_for_Machine_Learning

[3] https://elearning.novaims.unl.pt/mod/resource/view.php?id=74695

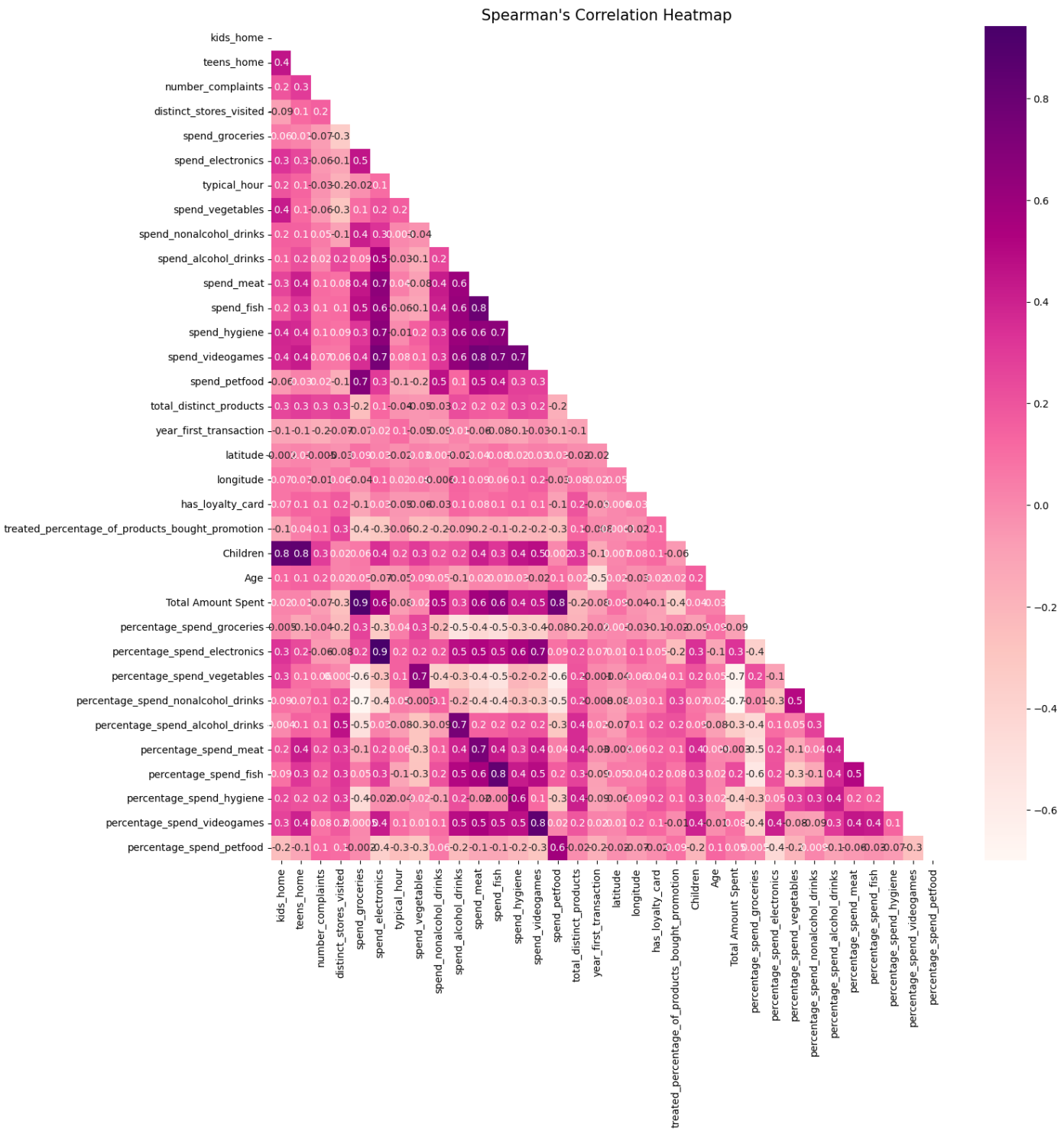
[4] https://elearning.novaims.unl.pt/mod/resource/view.php?id=74891d

[5] Shetty, Pranav; Singh, Suraj. “Hierarchical Clustering: A Survey, April 2021,

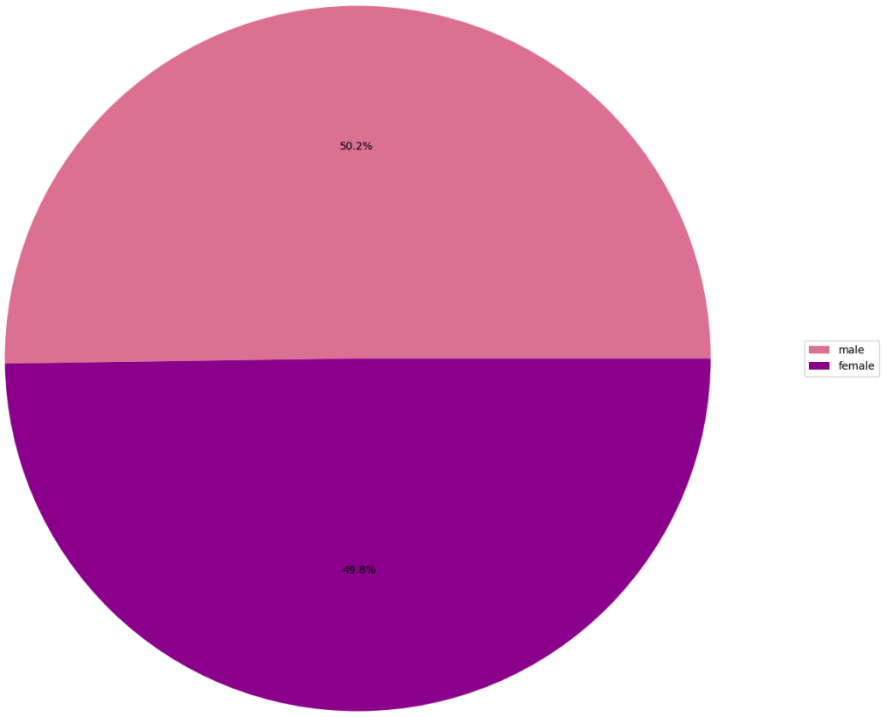
https://www.researchgate.net/publication/351076785_Hierarchical_Clustering_A_Survey

7. Annexes

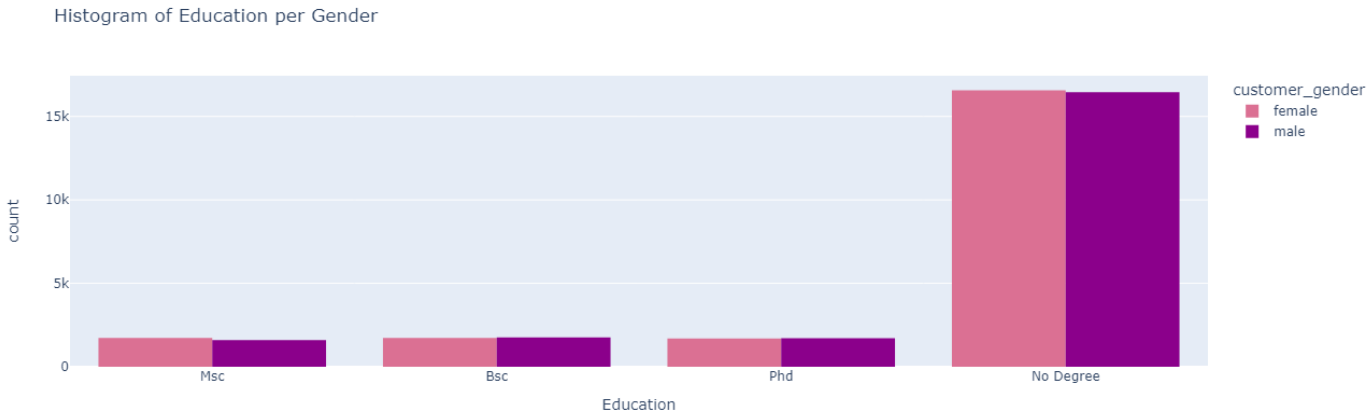
[1] SPEARMAN'S CORRELATION HEATMAP



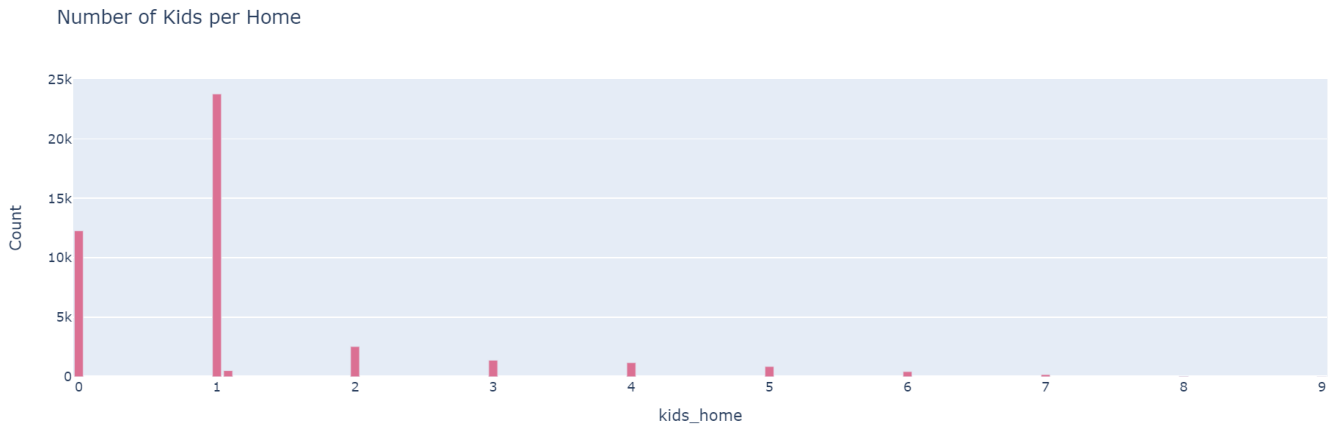
[2] GENDER DISTRIBUTION IN THE REGULAR CUSTOMER SEGMENT



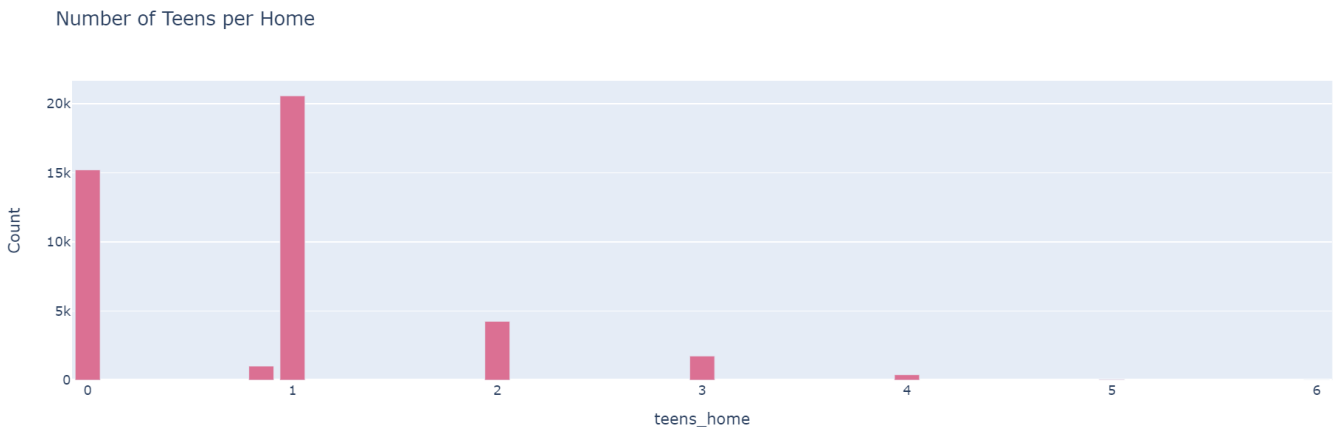
[3] HISTOGRAM OF EDUCATION PER GENDER IN THE REGULAR CUSTOMER SEGMENT



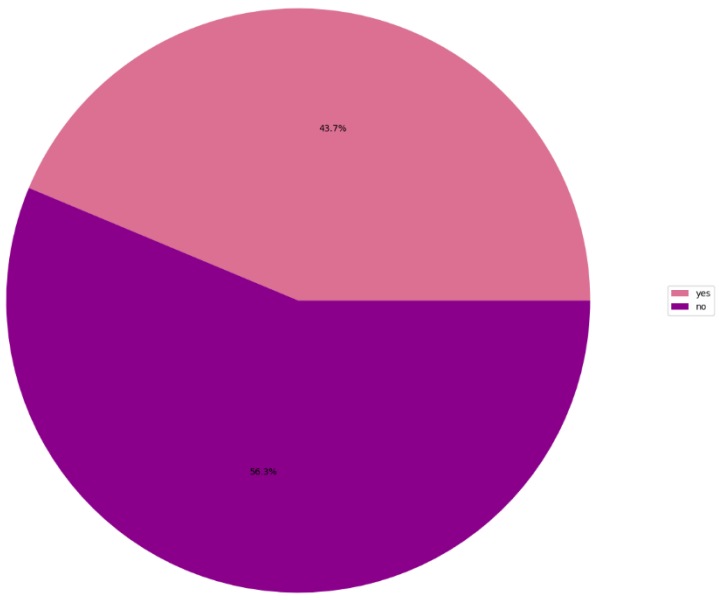
[4] DISTRIBUTION OF KIDS IN HOUSEHOLDS IN THE REGULAR CUSTOMER SEGMENT



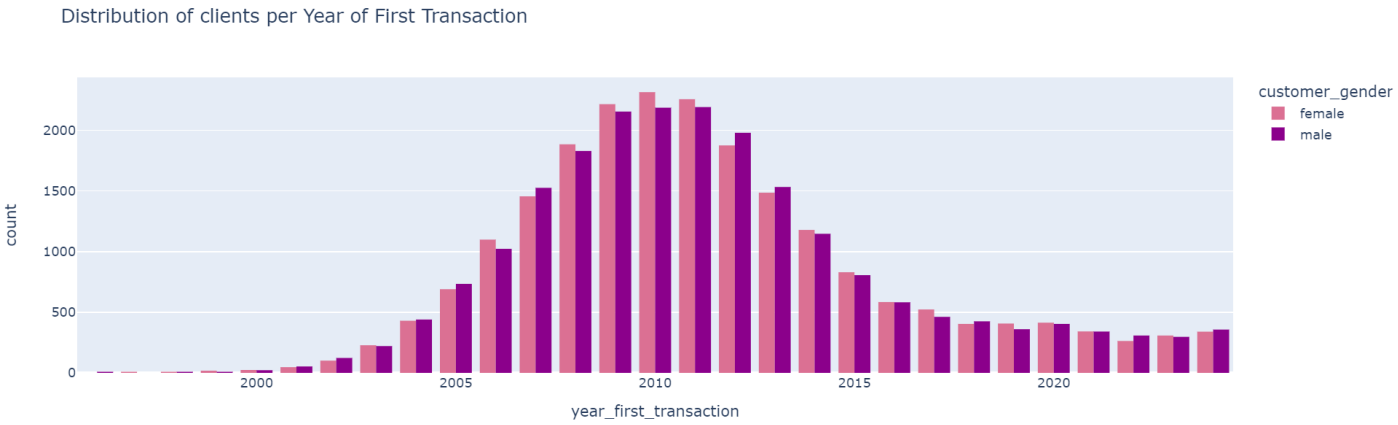
[5] DISTRIBUTION OF TEENS IN HOUSEHOLDS IN THE REGULAR CUSTOMER SEGMENT



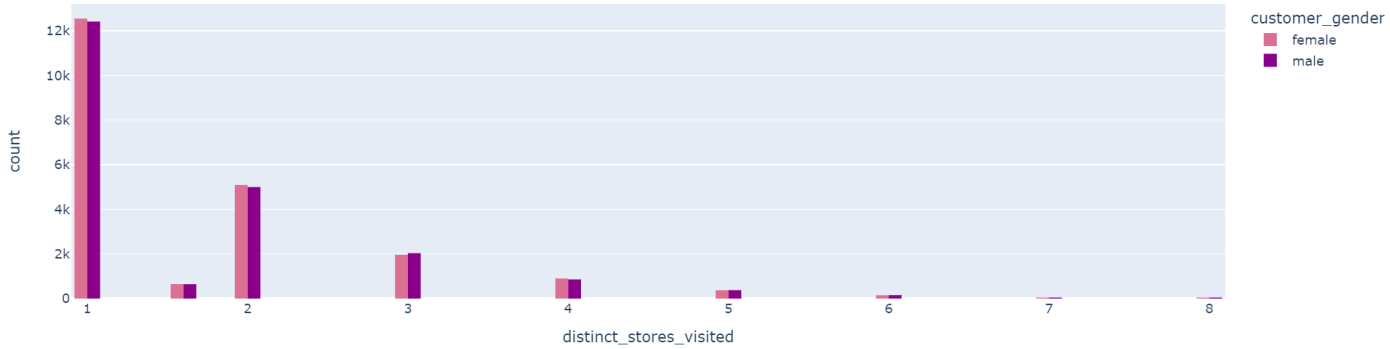
[6] LOYALTY CARD DISTRIBUTION AMONG THE REGULAR CUSTOMER SEGMENT



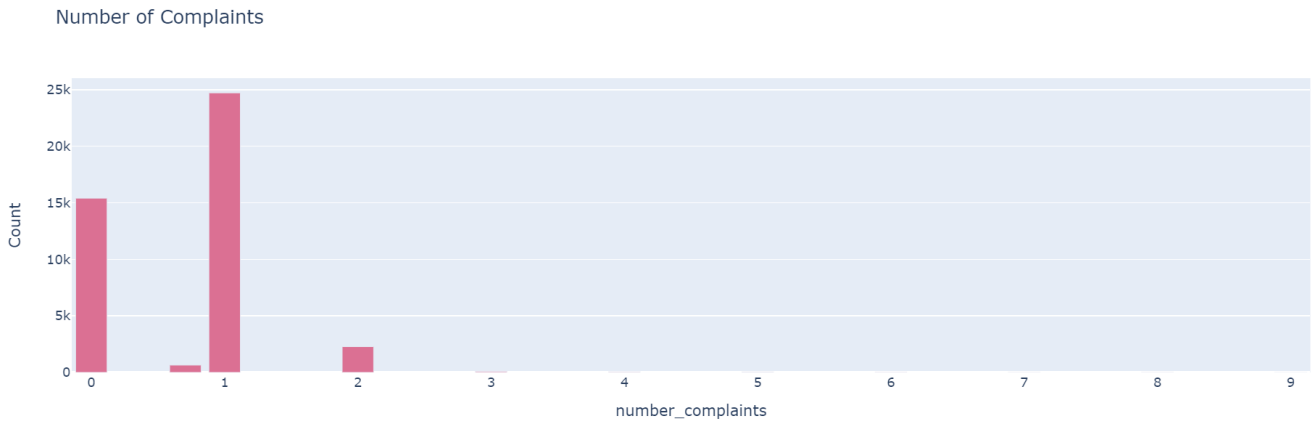
[7] CLIENT ACQUISITION BY YEAR IN THE REGULAR CUSTOMER SEGMENT



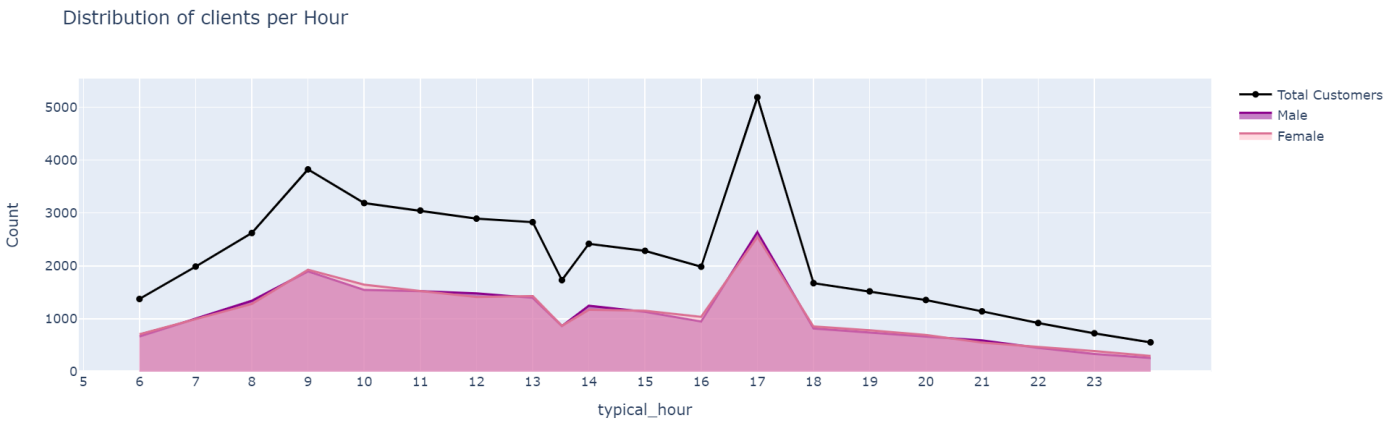
[8] DISTRIBUTION OF CLIENTS PER NUMBER OF DISTINCT STORES VISITED IN THE REGULAR CUSTOMER SEGMENT



[9] DISTRIBUTION OF CLIENTS PER NUMBER OF COMPLAINTS IN THE REGULAR CUSTOMER SEGMENT

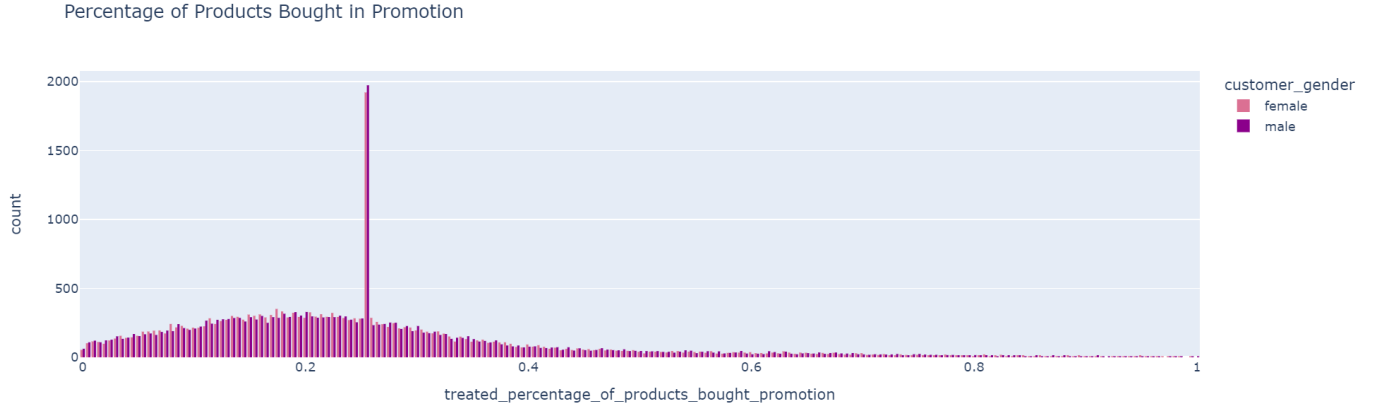


[10] DISTRIBUTION OF CLIENT ATTENDANCE BY HOUR IN THE REGULAR CUSTOMER SEGMENT

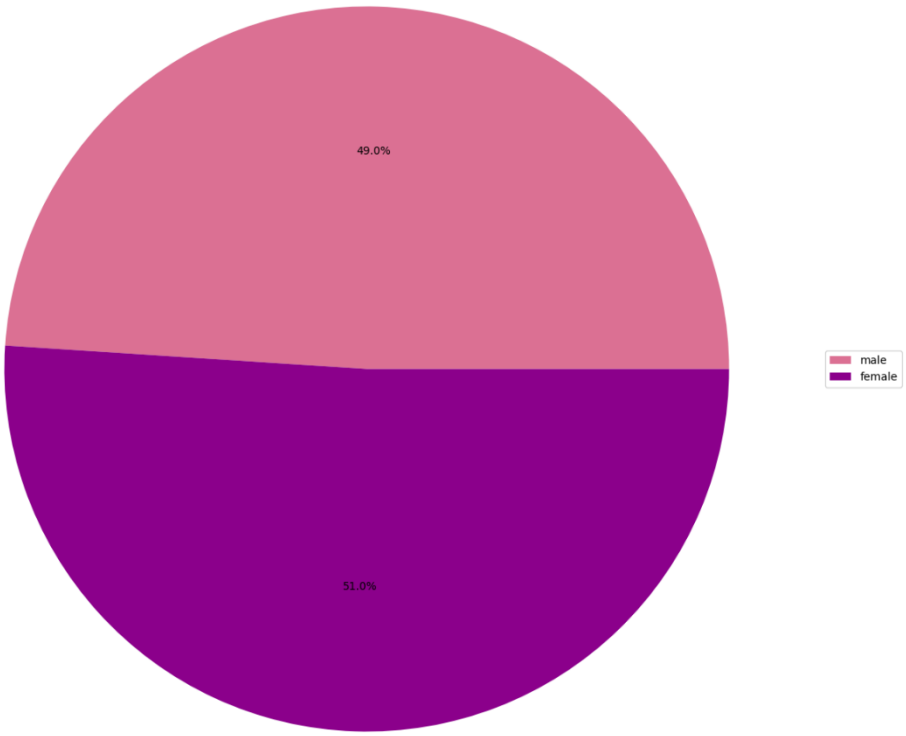


[11] TOTAL MONEY SPENT BY CATEGORY IN THE REGULAR CUSTOMER SEGMENT

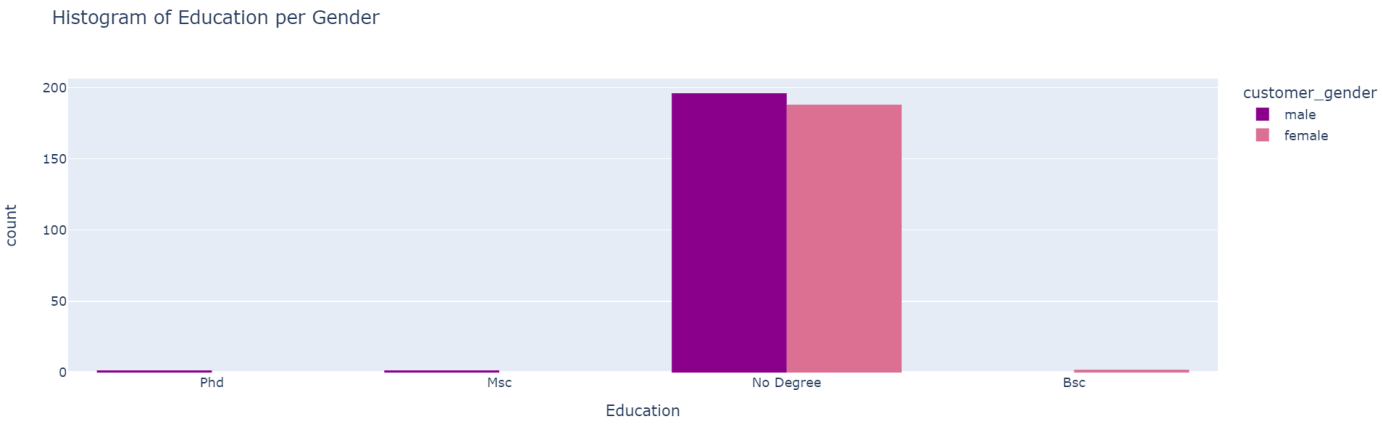
[12] PERCENTAGE OF PRODUCTS BOUGHT IN PROMOTION IN THE REGULAR CUSTOMER SEGMENT



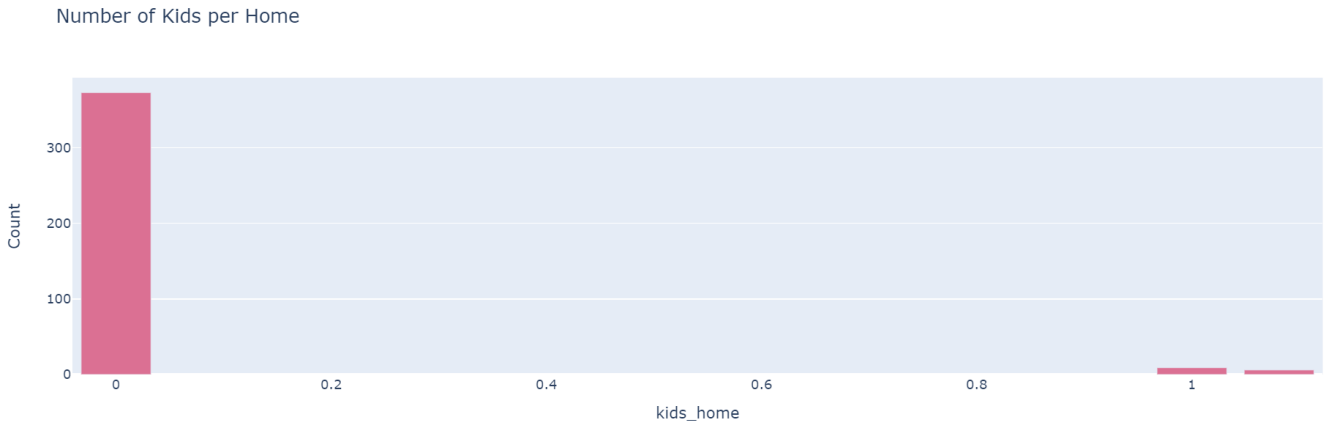
[13] GENDER DISTRIBUTION IN THE REGULAR FISHY SEGMENT



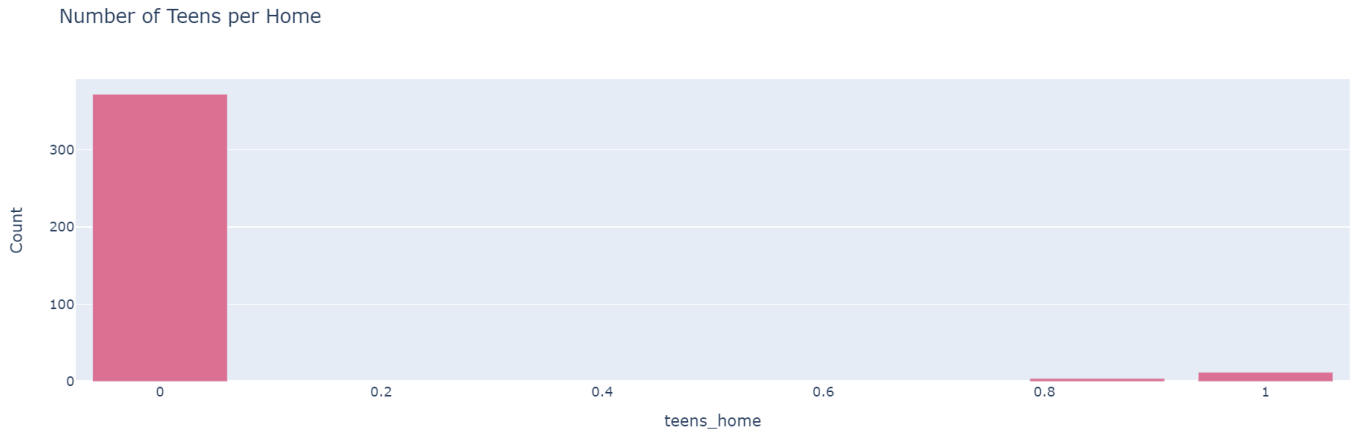
[14] HISTOGRAM OF EDUCATION PER GENDER IN THE FISHY SEGMENT



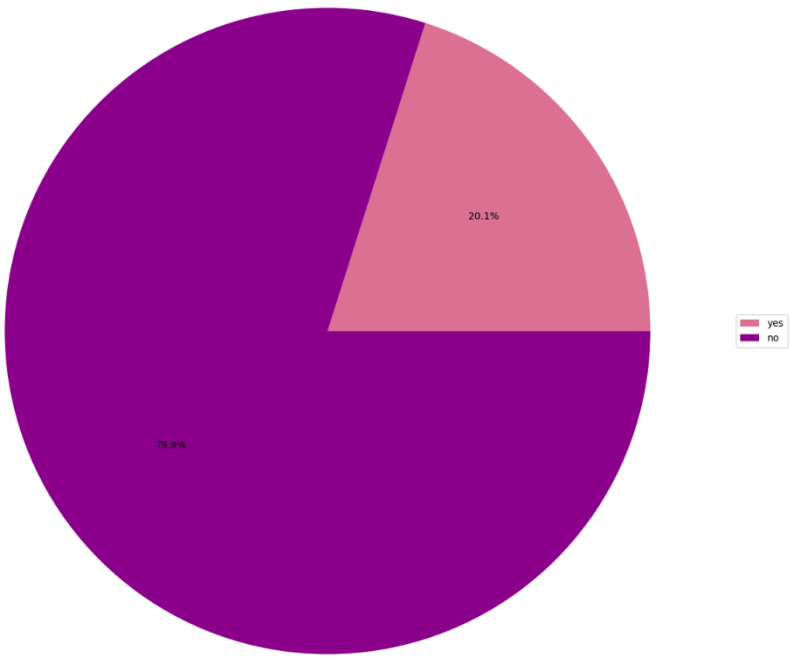
[15] DISTRIBUTION OF KIDS IN HOUSEHOLDS IN THE FISHY SEGMENT



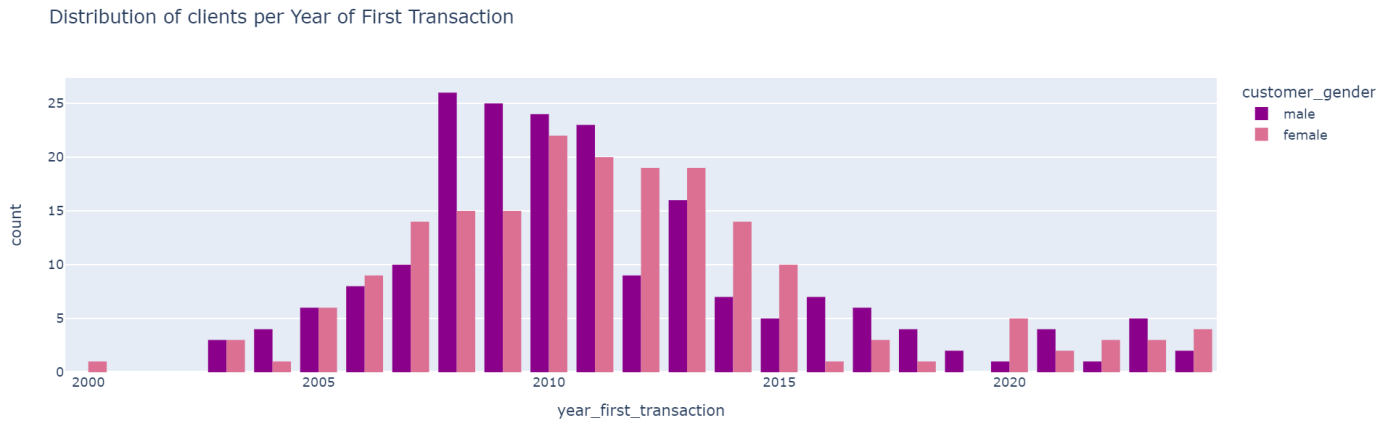
[16] DISTRIBUTION OF TEENS IN HOUSEHOLDS IN THE FISHY SEGMENT



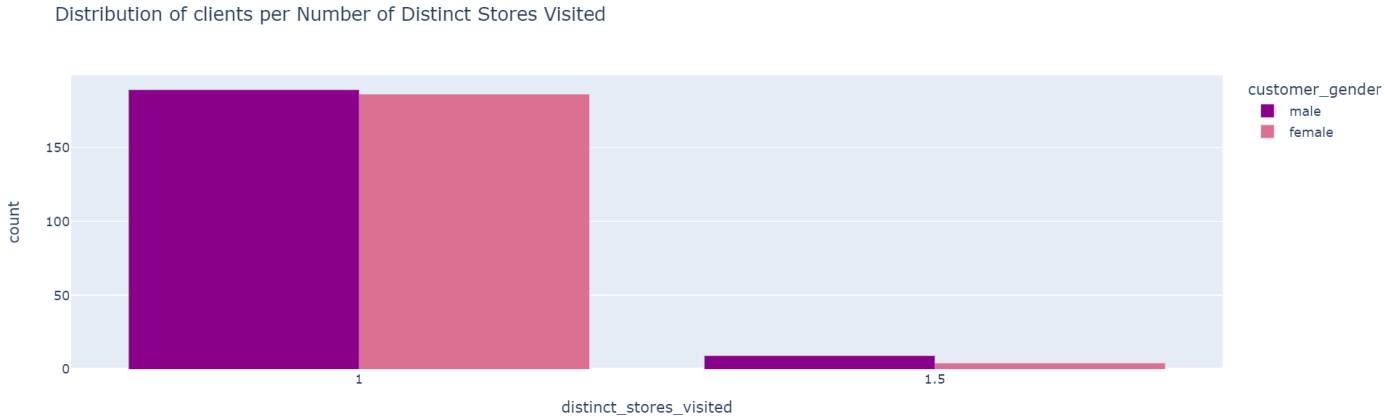
[17] LOYALTY CARD DISTRIBUTION AMONG THE FISHY SEGMENT



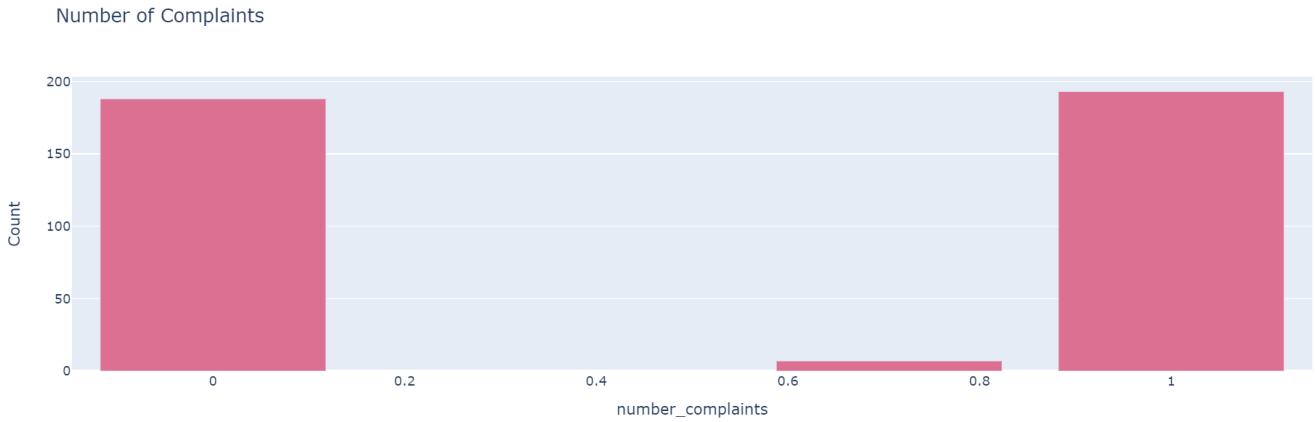
[18] CLIENT ACQUISITION BY YEAR IN THE FISHY SEGMENT



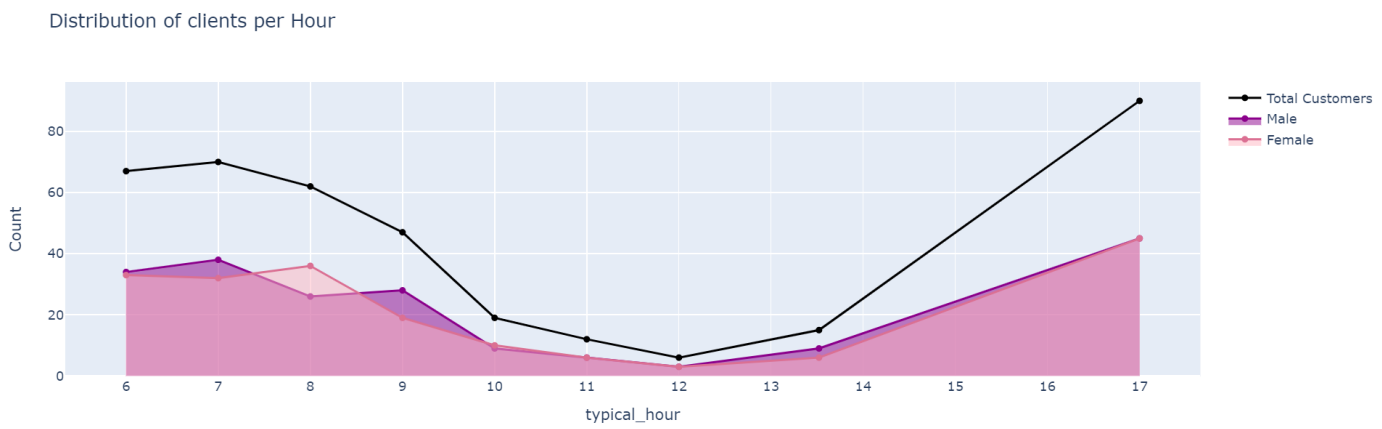
[19] DISTRIBUTION OF CLIENTS PER NUMBER OF DISTINCT STORES VISITED IN THE FISHY SEGMENT



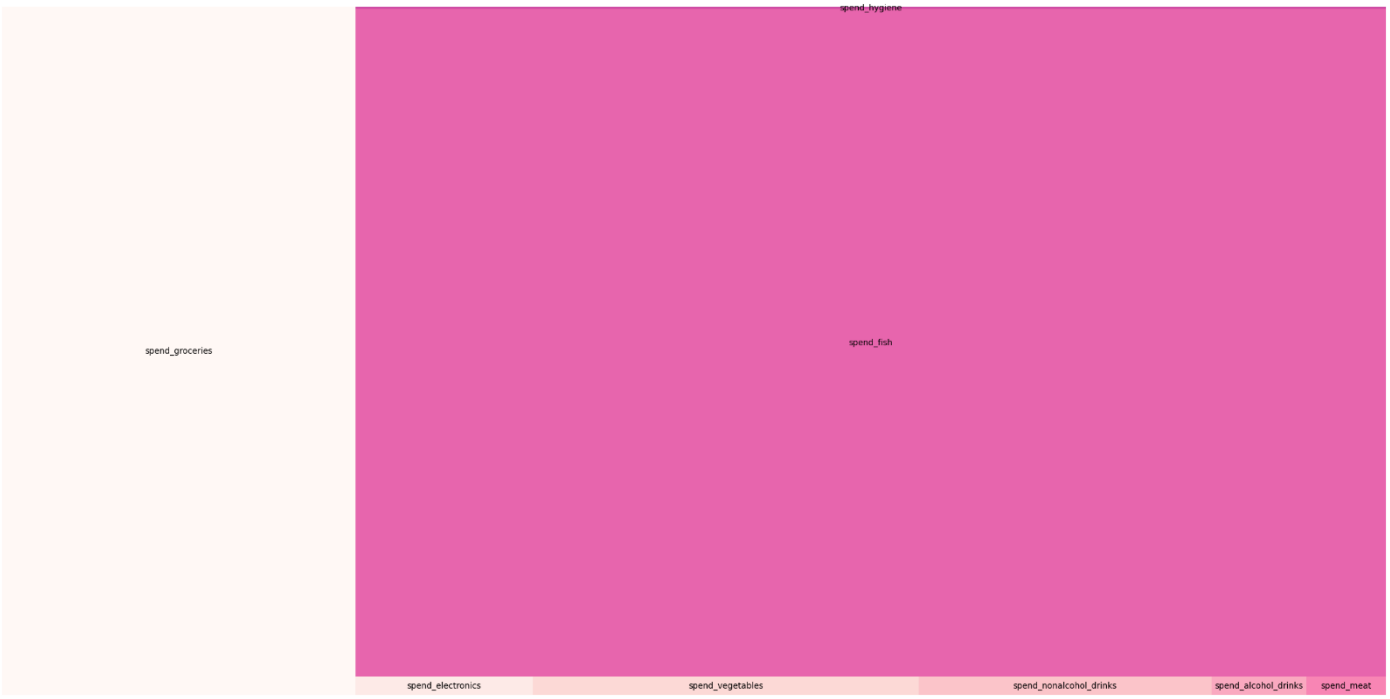
[20] DISTRIBUTION OF CLIENTS PER NUMBER OF COMPLAINTS IN THE FISHY SEGMENT



[21] DISTRIBUTION OF CLIENT ATTENDANCE BY HOUR IN THE FISHY SEGMENT

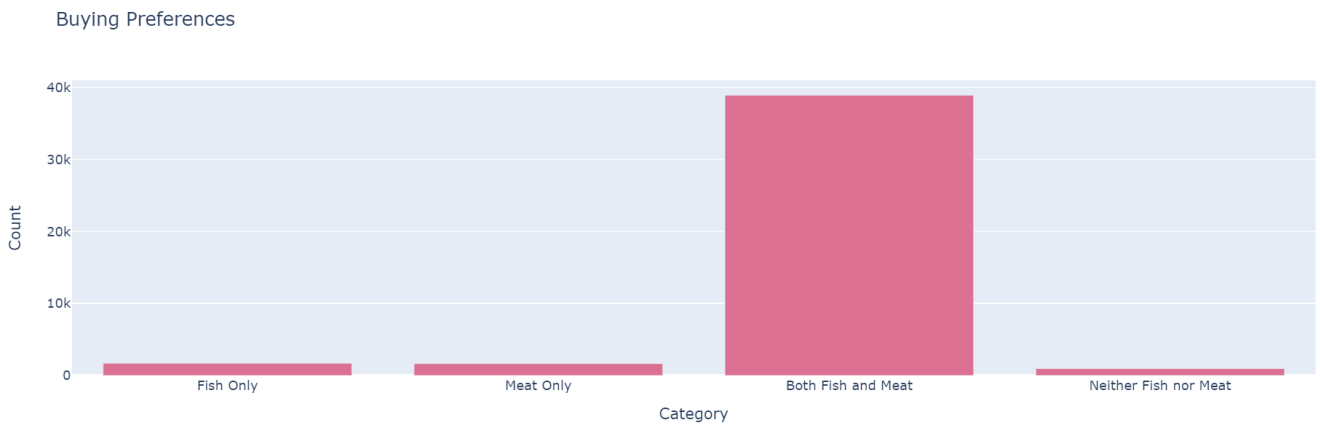


[22] TOTAL MONEY SPENT BY CATEGORY IN THE FISHY SEGMENT

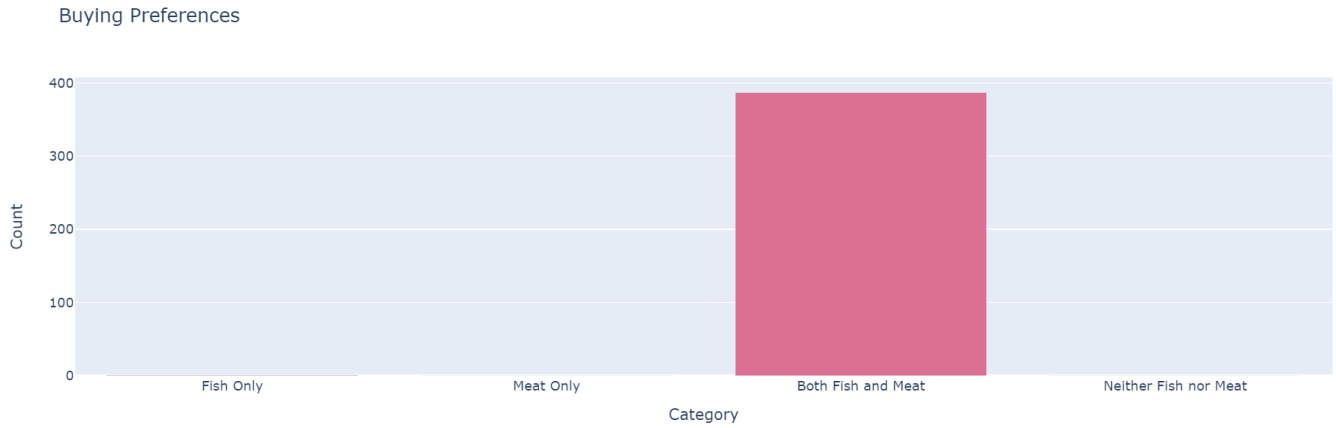


[23] PERCENTAGE OF PRODUCTS BOUGHT IN PROMOTION IN THE FISHY SEGMENT

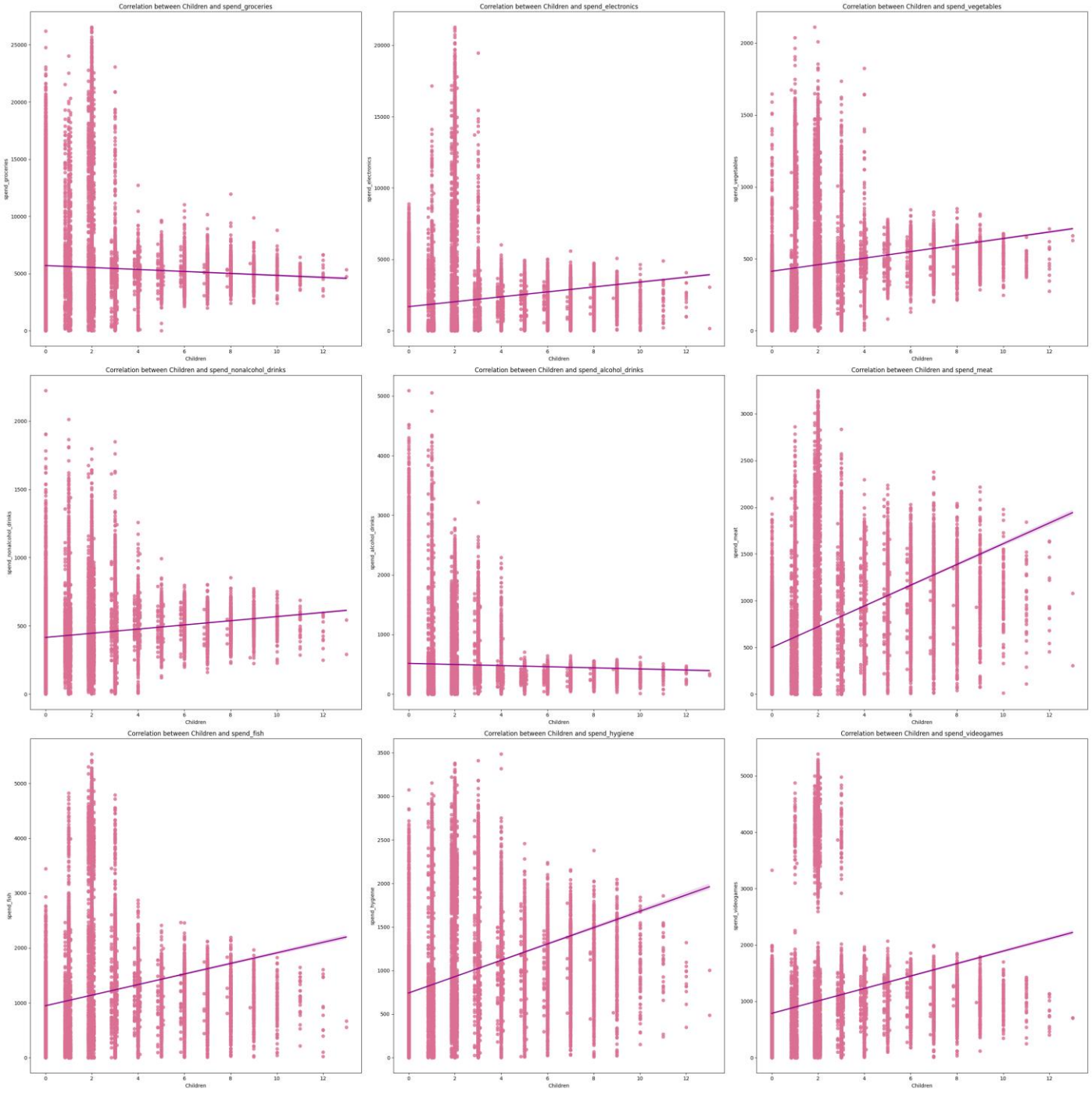
[24] BUYING PREFERENCES IN THE REGULAR CUSTOMER SEGMENT



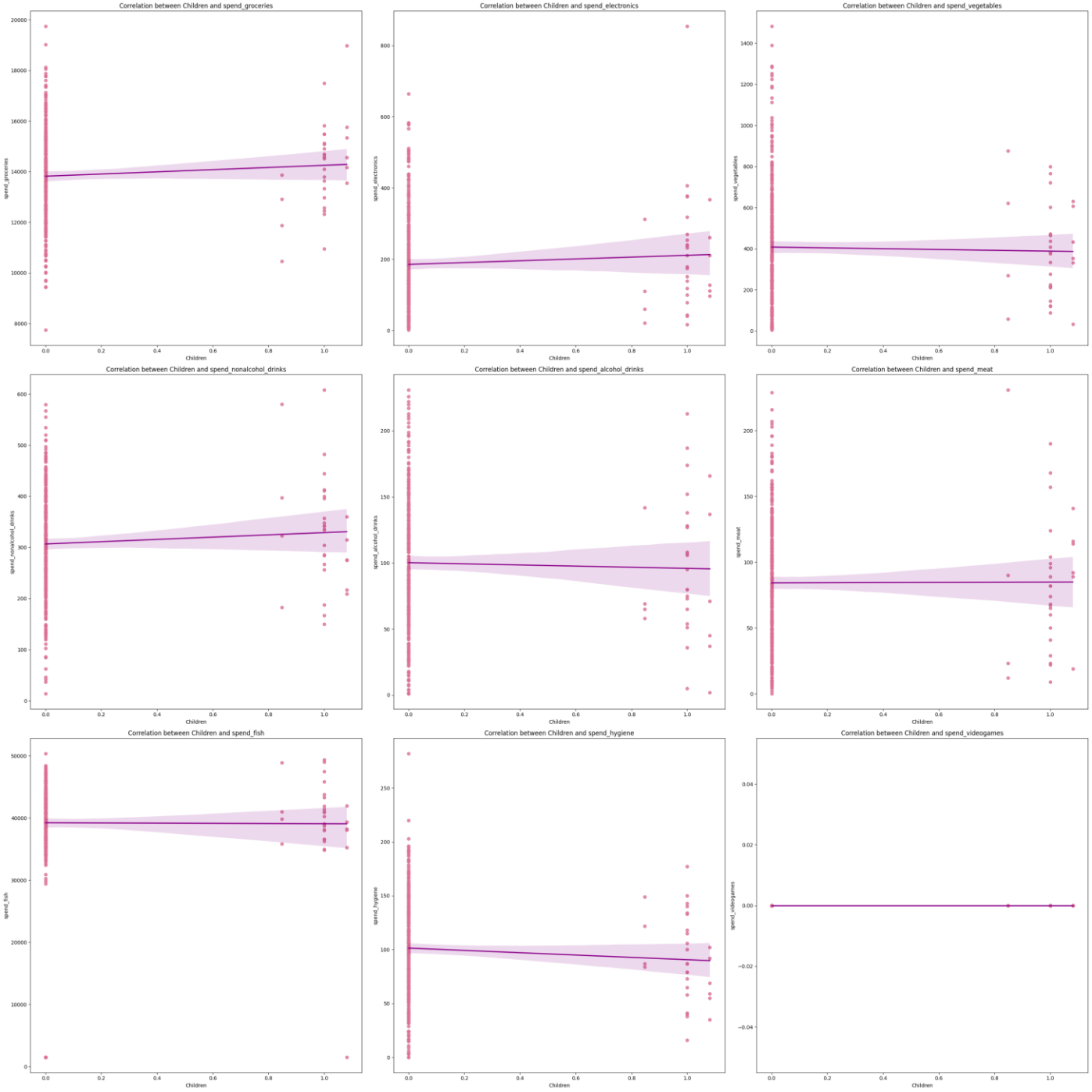
[25] BUYING PREFERENCES IN THE FISHY SEGMENT



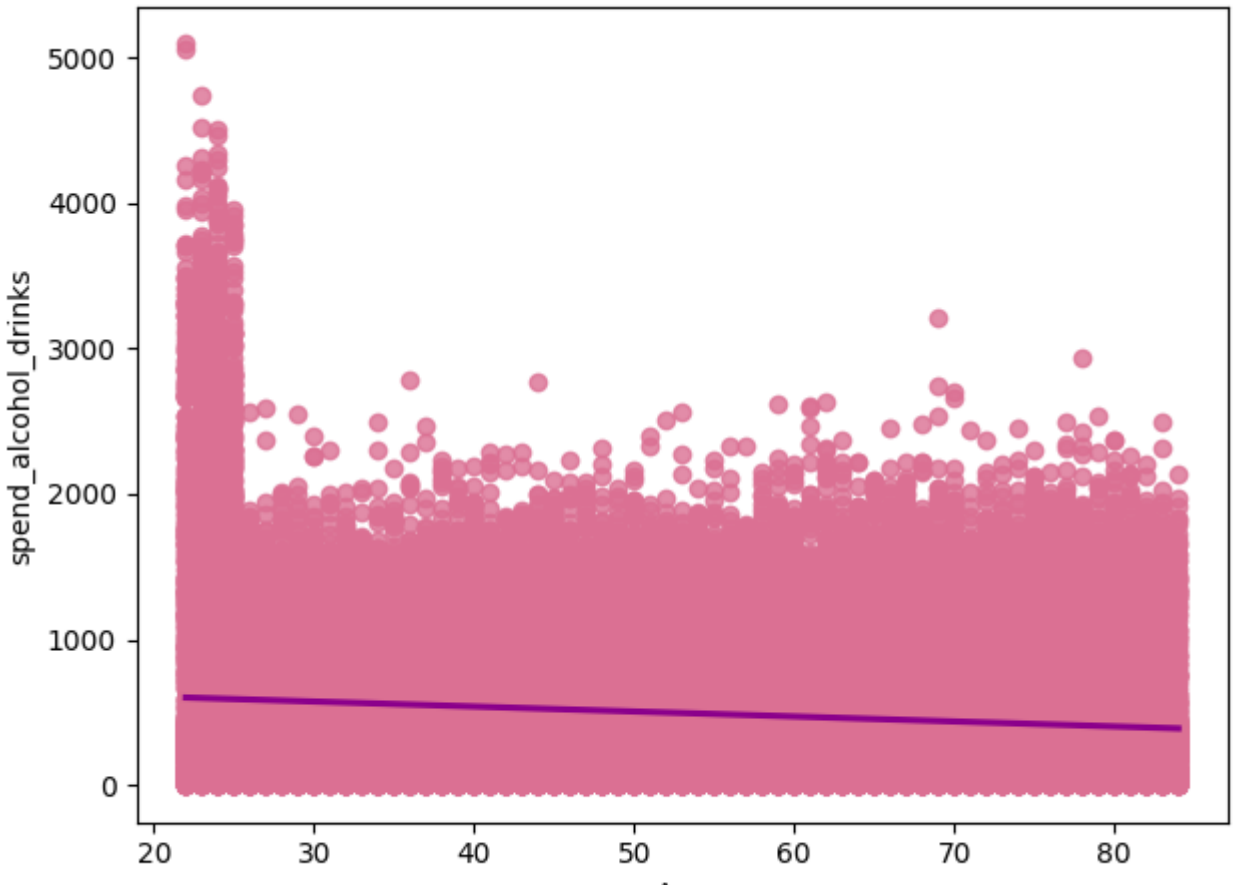
[26] CORRELATION BETWEEN THE NUMBER OF CHILDREN AND THE PRODUCT CATEGORIES IN THE REGULAR CUSTOMER SEGMENT



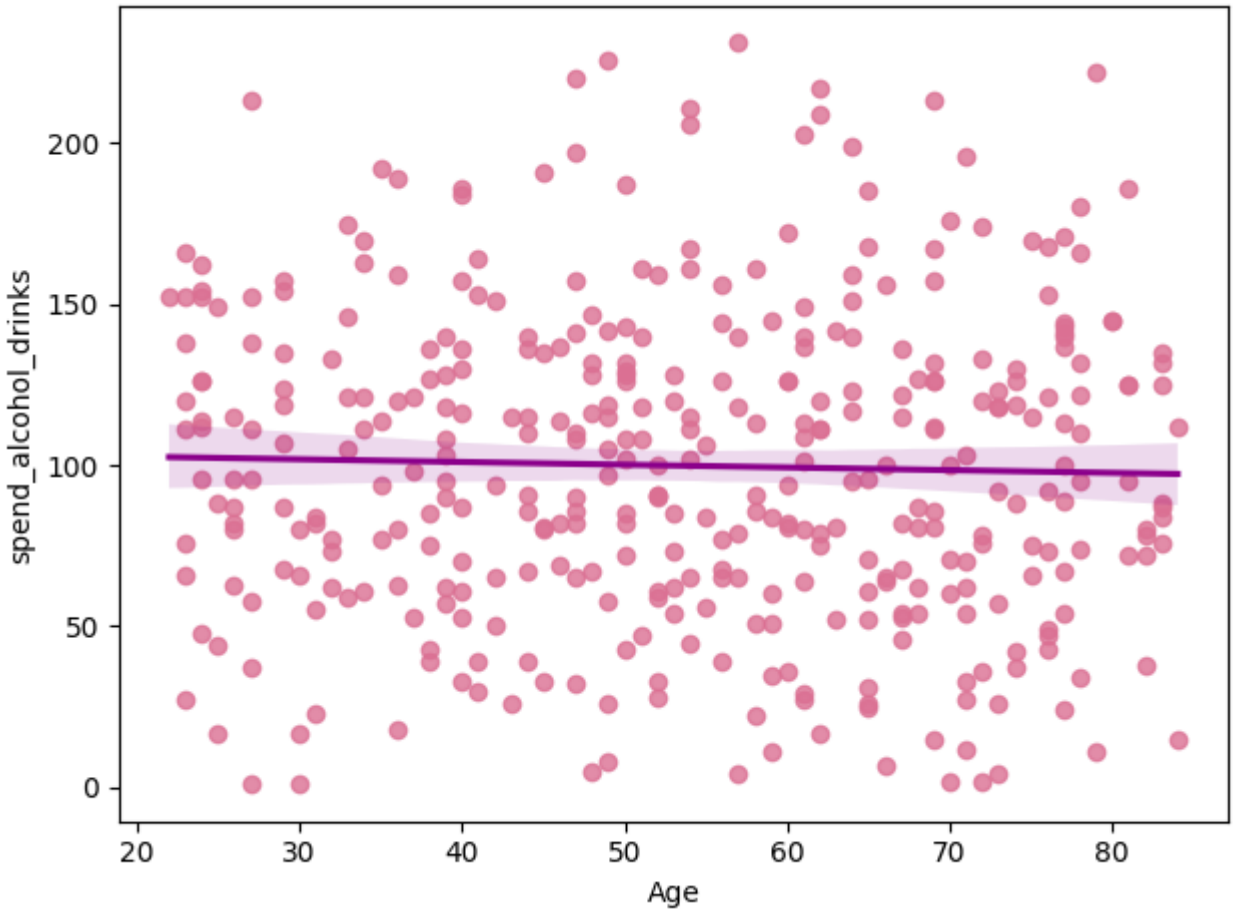
[27] CORRELATION BETWEEN THE NUMBER OF CHILDREN AND THE PRODUCT CATEGORIES IN THE FISHY SEGMENT



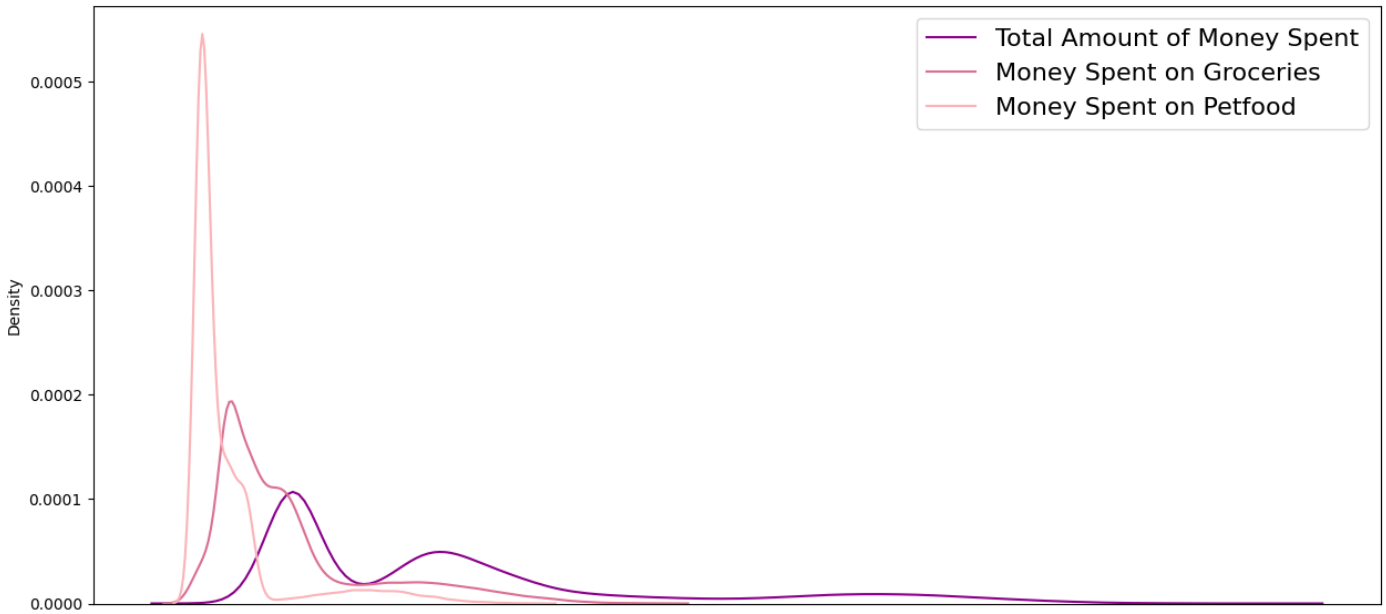
[28] CORRELATION OF AGE WITH THE MONEY SPENT ON ALCOHOLIC DRINKS IN THE REGULAR CUSTOMER SEGMENT



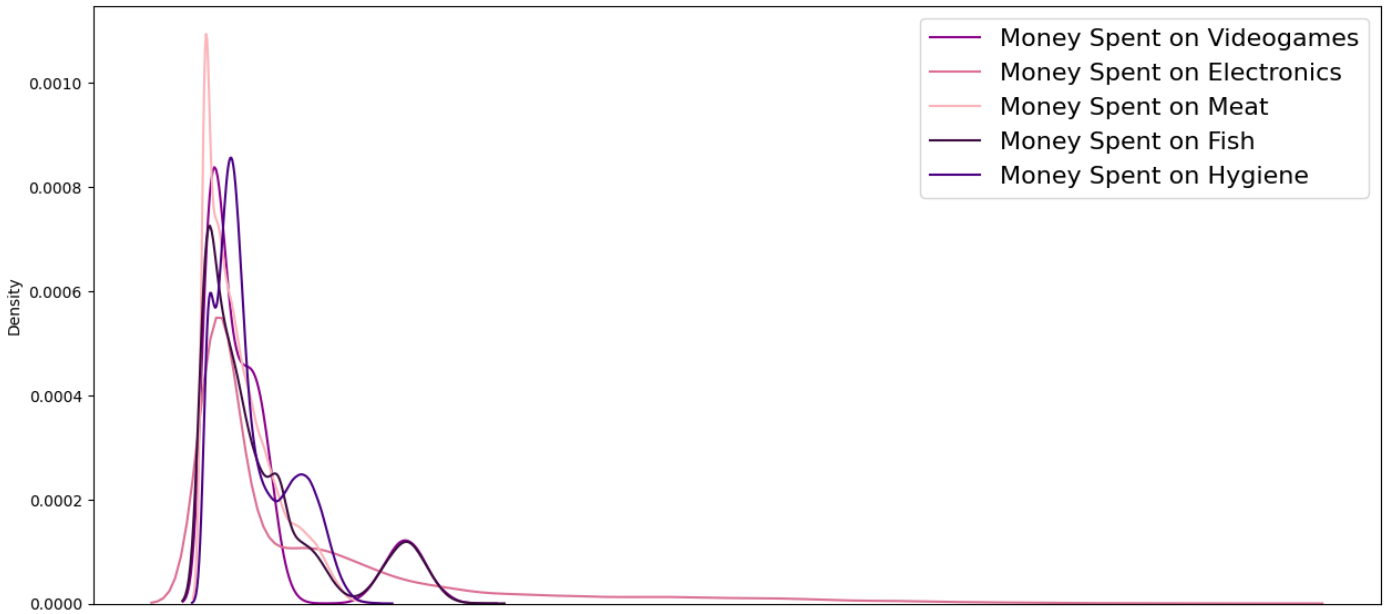
[29] CORRELATION OF AGE WITH THE MONEY SPENT ON ALCOHOLIC DRINKS IN THE REGULAR CUSTOMER SEGMENT



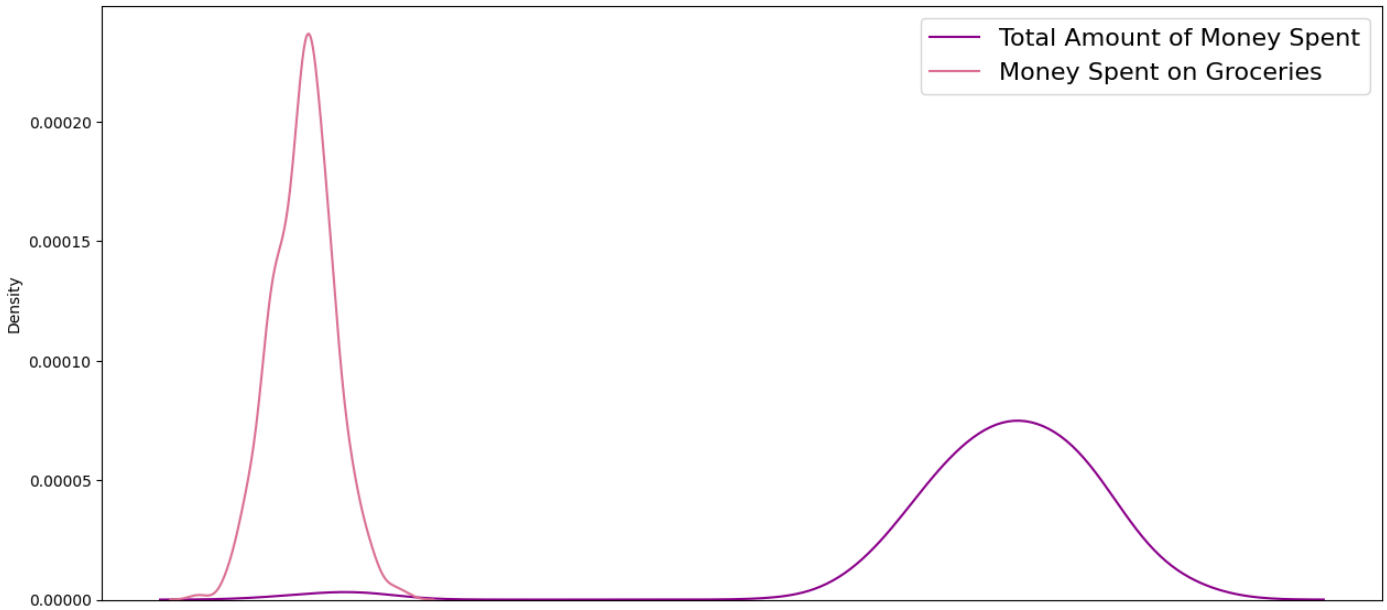
[30] TOTAL AMOUNT SPENT & GROCERIES & PET FOOD IN THE REGULAR CUSTOMER SEGMENT



[31] VIDEOGAMES & ELECTRONICS & MEAT & FISH & HYGIENE IN THE REGULAR CUSTOMER SEGMENT

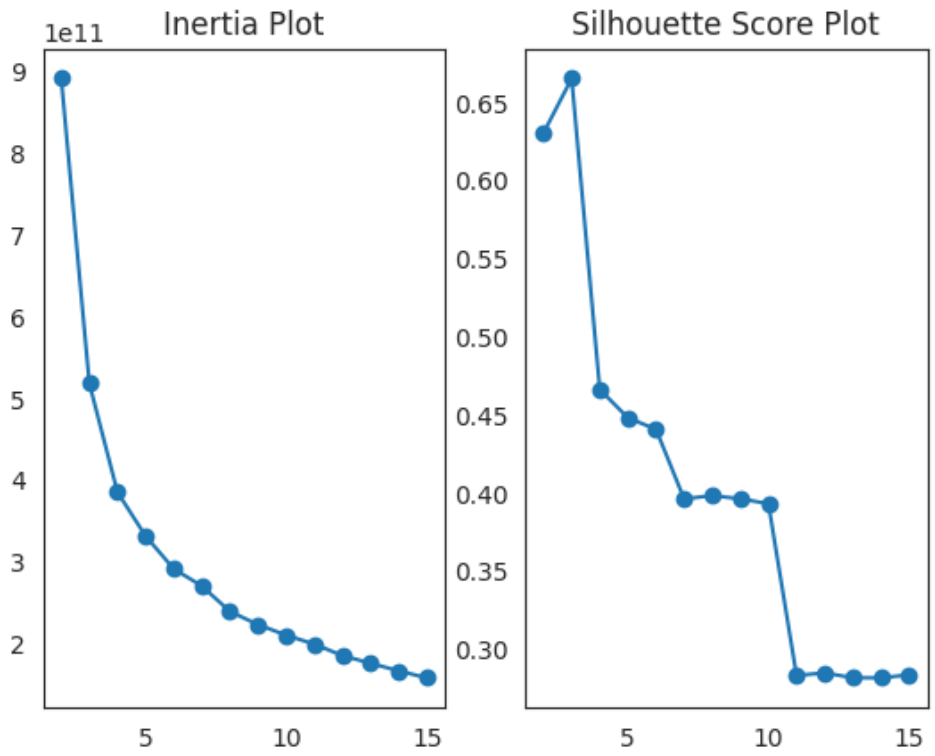


[32] TOTAL AMOUNT SPENT & GROCERIES IN THE FISHY SEGMENT



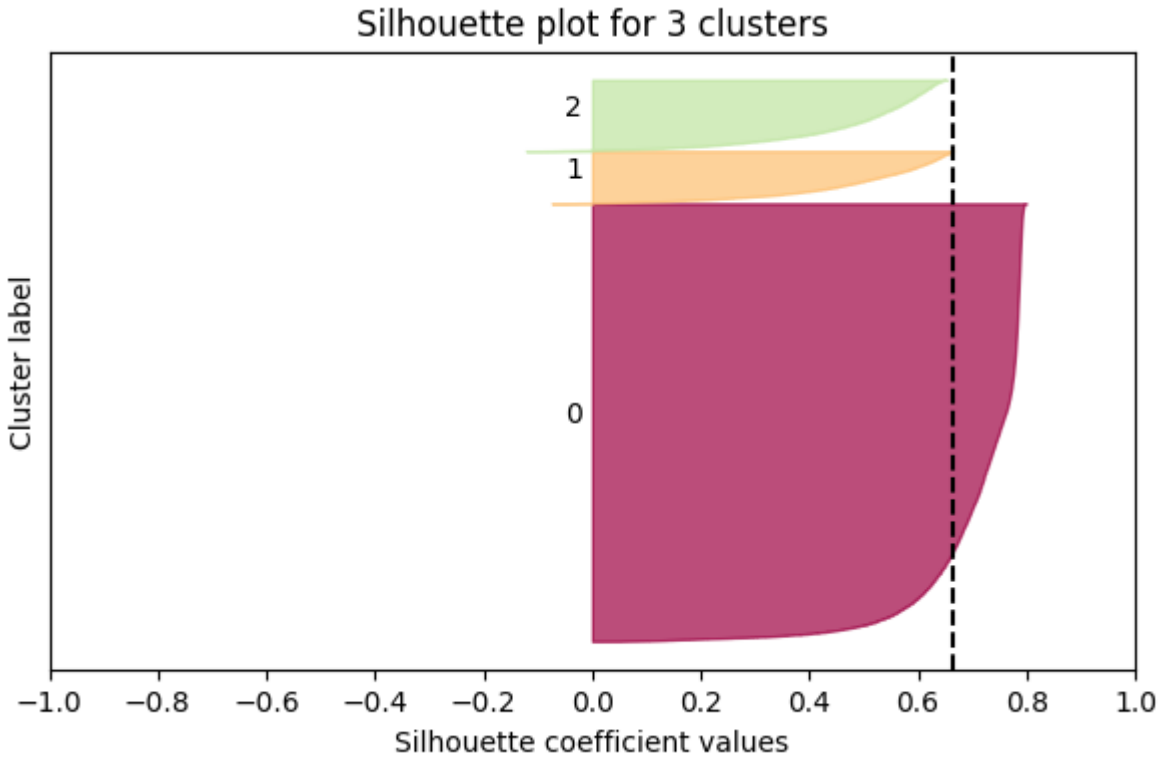
[33] VIDEOGAMES & ELECTRONICS & MEAT & FISH IN THE FISHY SEGMENT

[34] INERTIA AND SILHOUETTE PLOT FOR K-MEANS CLUSTERING WITHOUT SCALING THE DATA

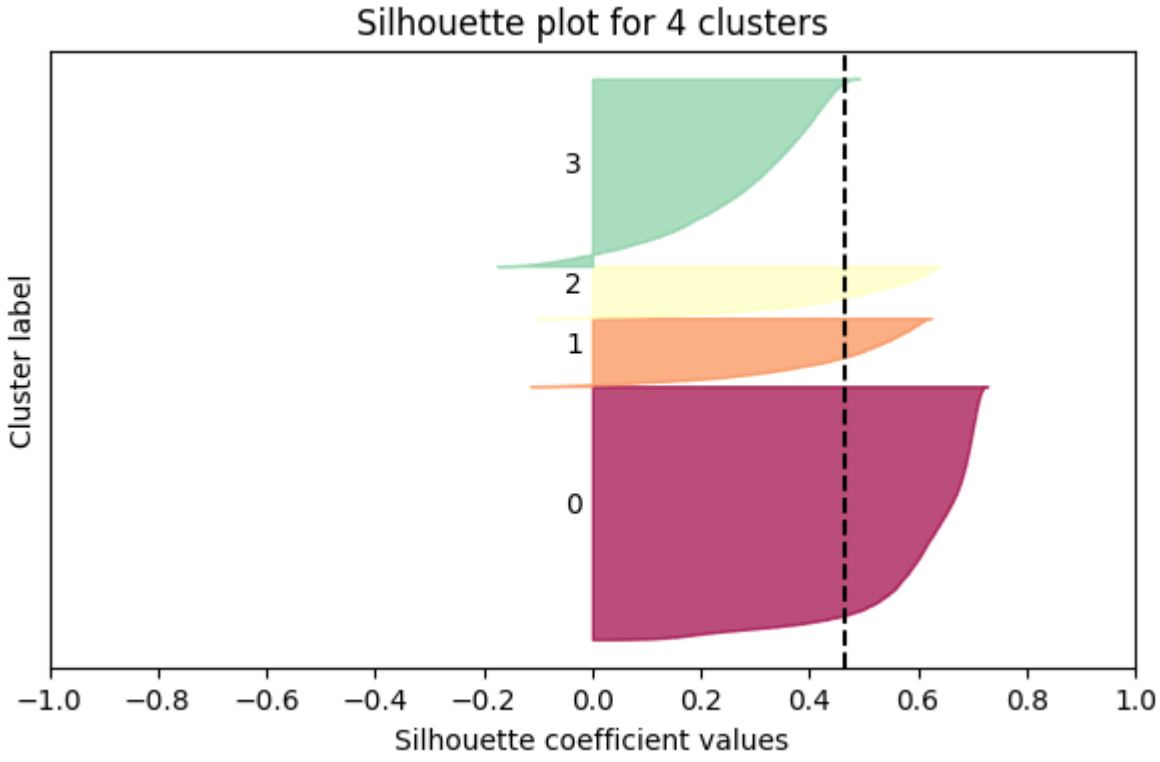


[35] SILHOUETTE PLOT FOR K-MEANS CLUSTERING WITHOUT SCALING THE DATA WITH 2 CLUSTERS

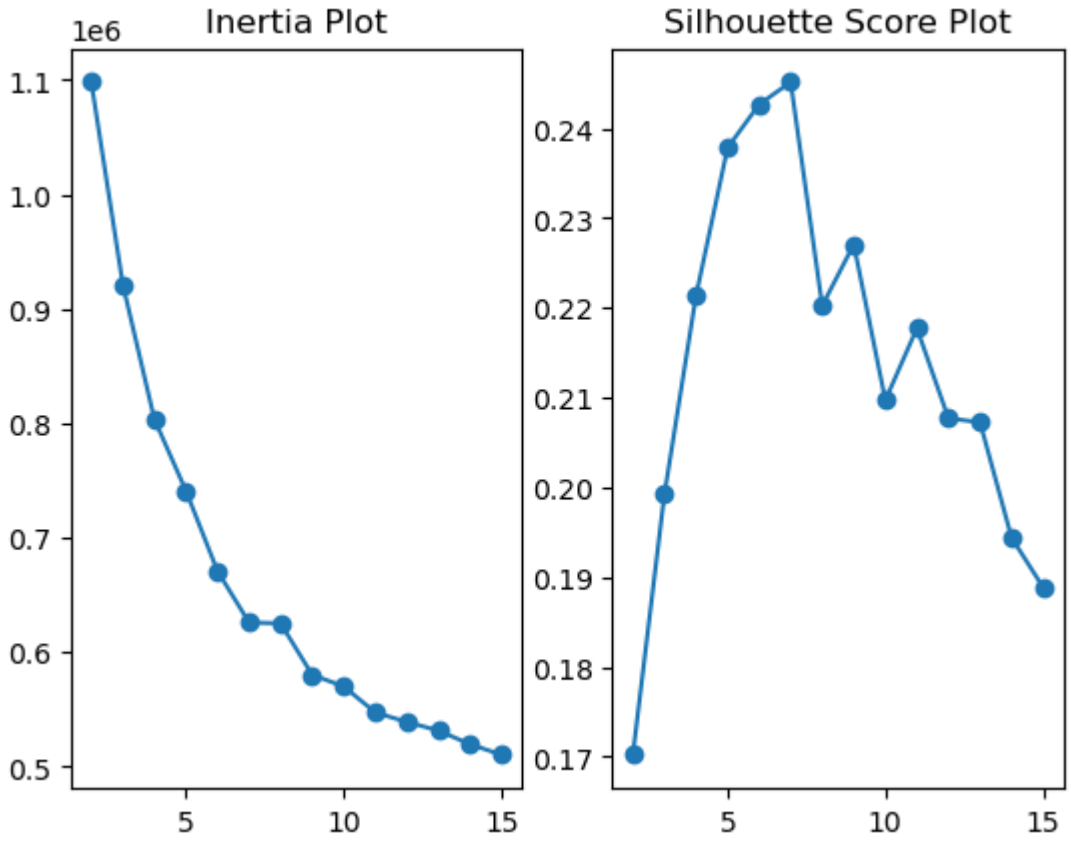
[36] SILHOUETTE PLOT FOR K-MEANS CLUSTERING WITHOUT SCALING THE DATA WITH 3 CLUSTERS



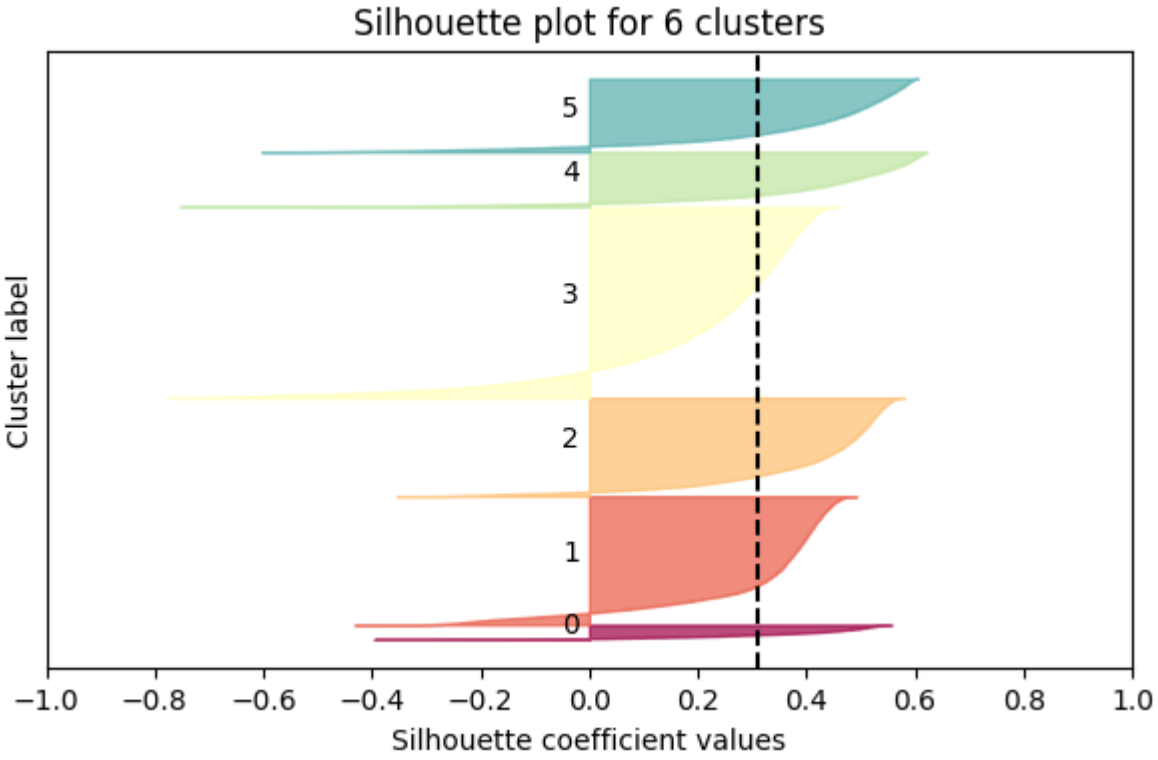
[37] SILHOUETTE PLOT FOR K-MEANS CLUSTERING WITHOUT SCALING THE DATA WITH 4 CLUSTERS



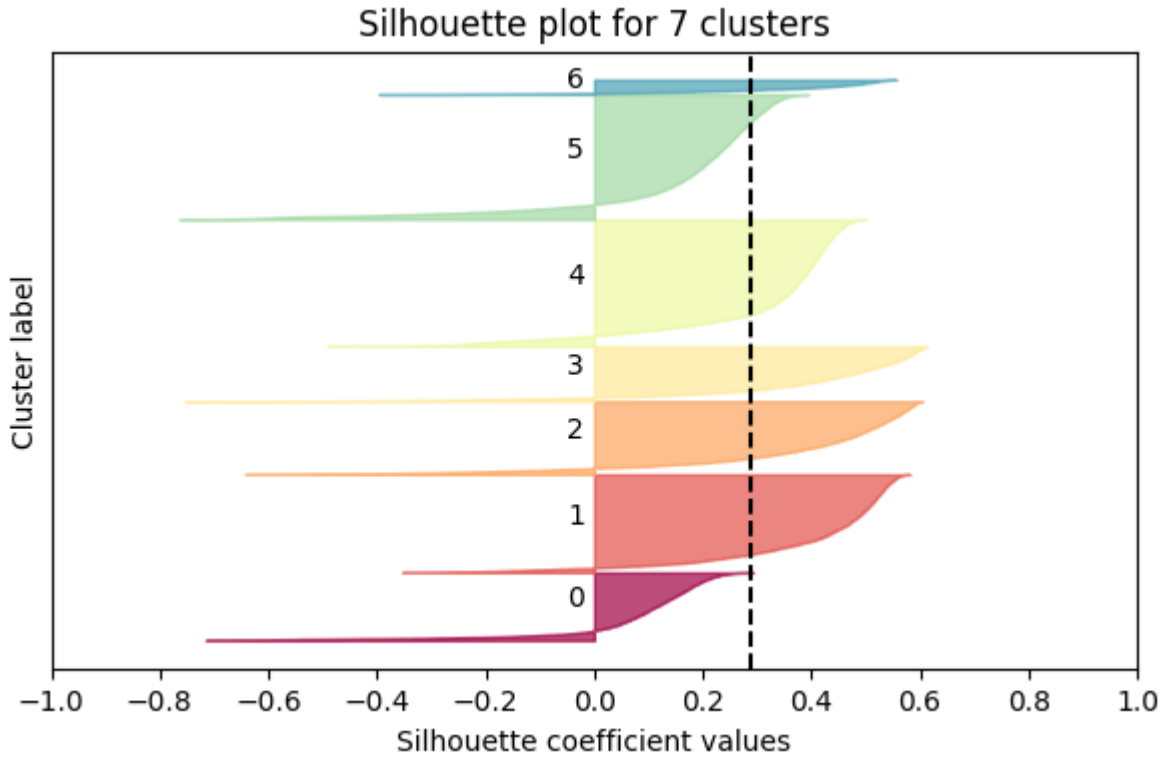
[38] INERTIA AND SILHOUETTE PLOT FOR K-MEANS CLUSTERING WITH STANDARD SCALING



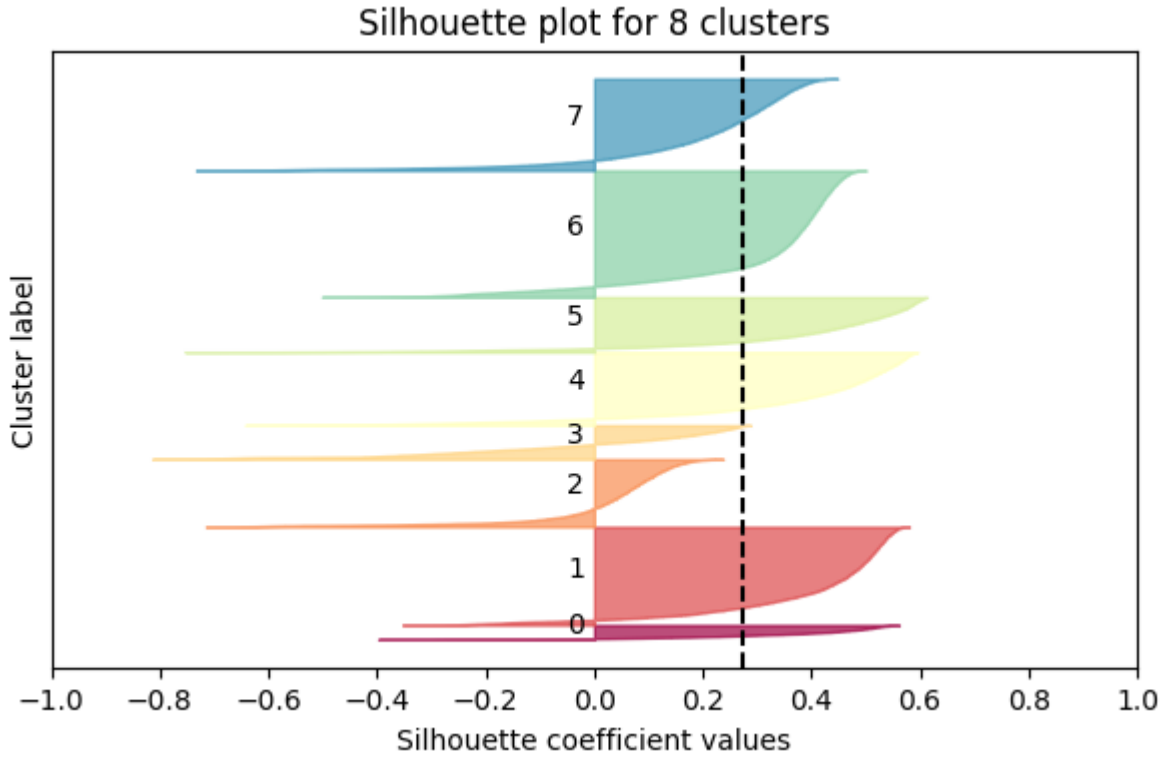
[39] SILHOUETTE PLOT FOR K-MEANS CLUSTERING WITH STANDARD SCALING AND 6 CLUSTERS



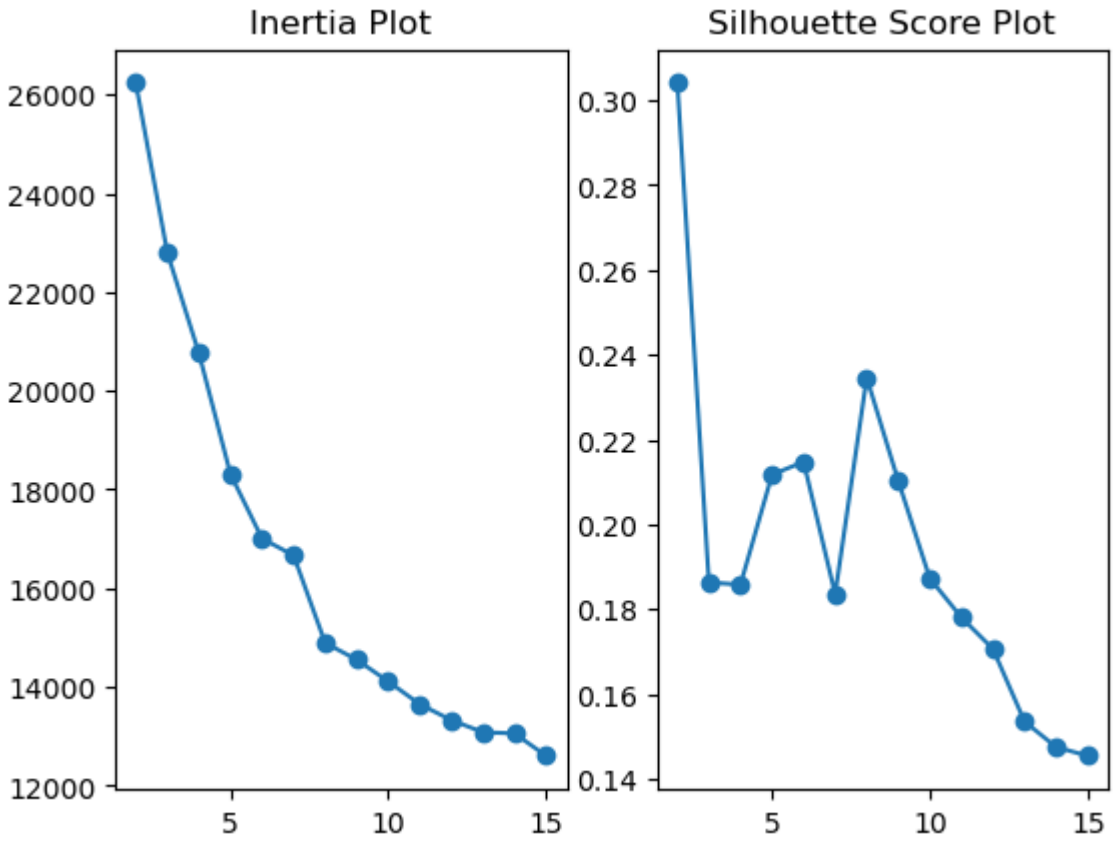
[40] SILHOUETTE PLOT FOR K-MEANS CLUSTERING WITH STANDARD SCALING AND 7 CLUSTERS



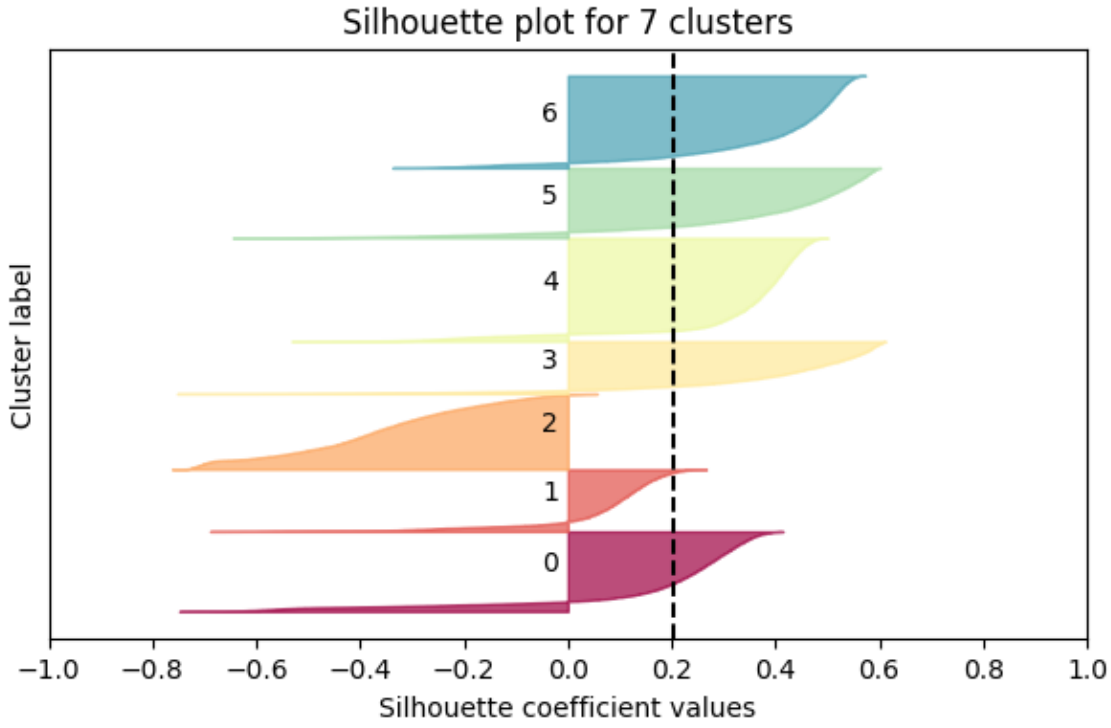
[41] SILHOUETTE PLOT FOR K-MEANS CLUSTERING WITH STANDARD SCALING AND 8 CLUSTERS



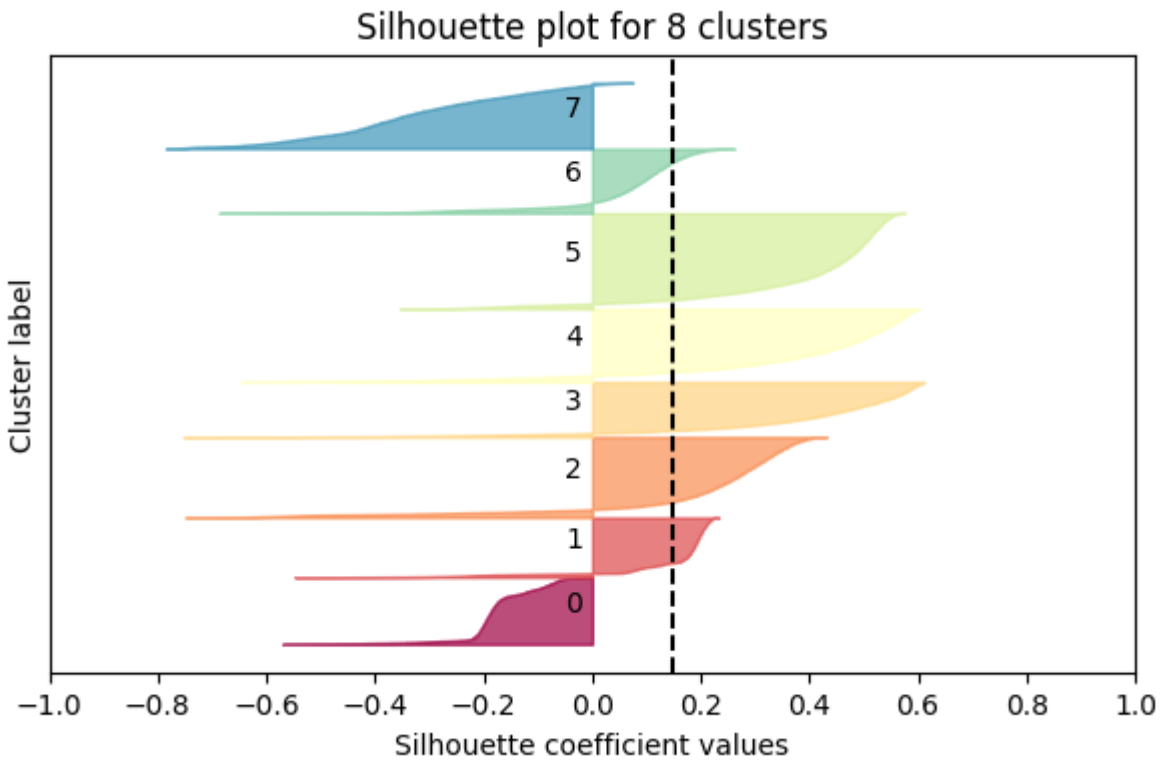
[42] INERTIA AND SILHOUETTE PLOT FOR K-MEANS CLUSTERING WITH MIN MAX SCALING



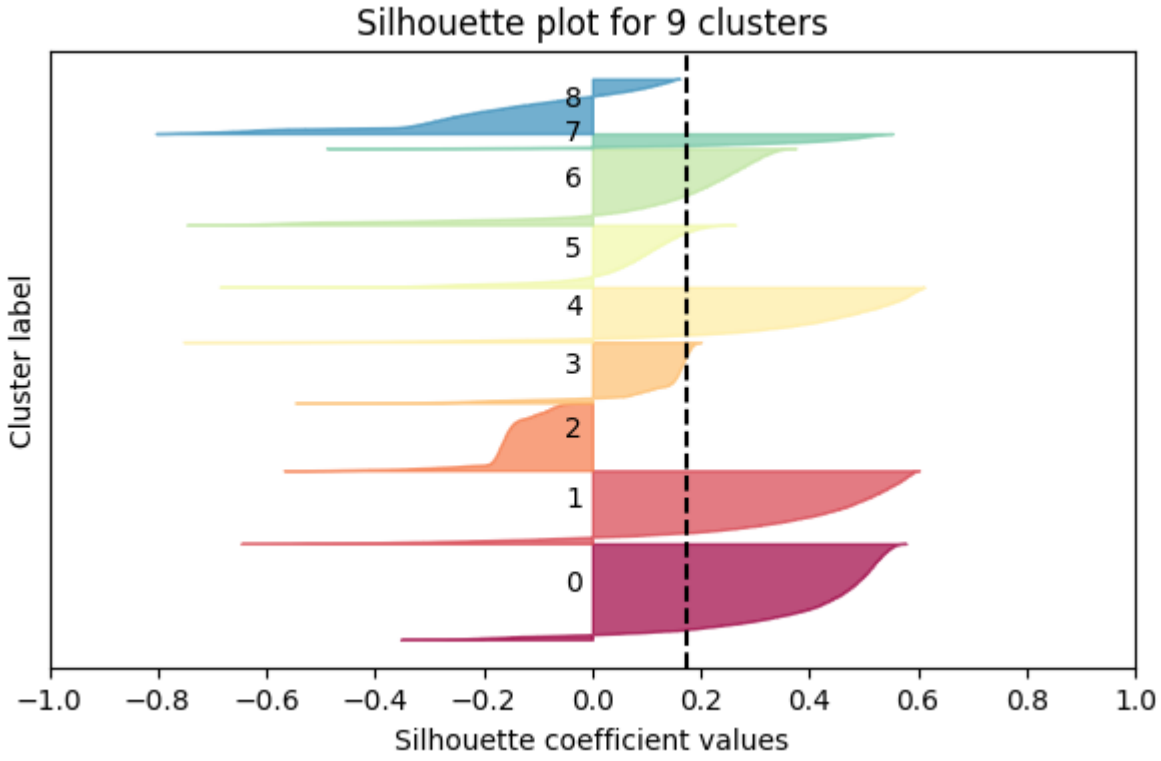
[43] SILHOUETTE PLOT FOR K-MEANS CLUSTERING WITH MIN MAX SCALING AND 7 CLUSTERS



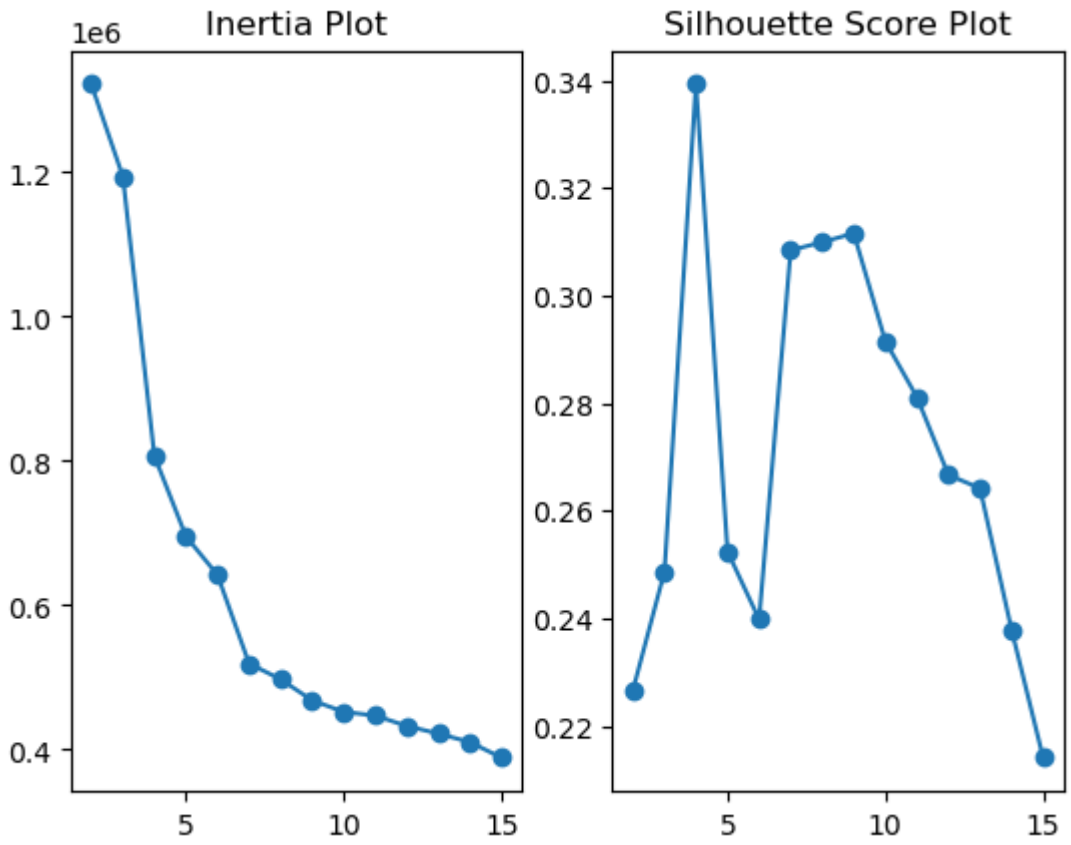
[44] SILHOUETTE PLOT FOR K-MEANS CLUSTERING WITH MIN MAX SCALING AND 8 CLUSTERS



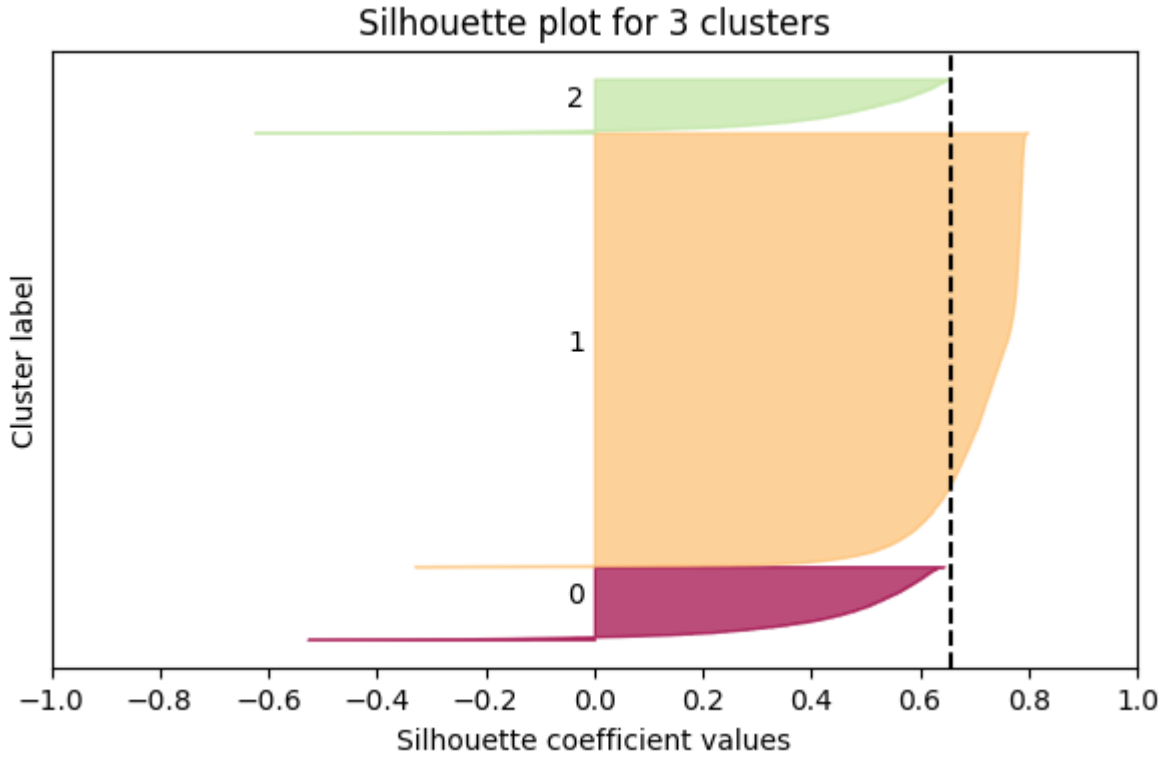
[45] SILHOUETTE PLOT FOR K-MEANS CLUSTERING WITH MIN MAX SCALING AND 9 CLUSTERS



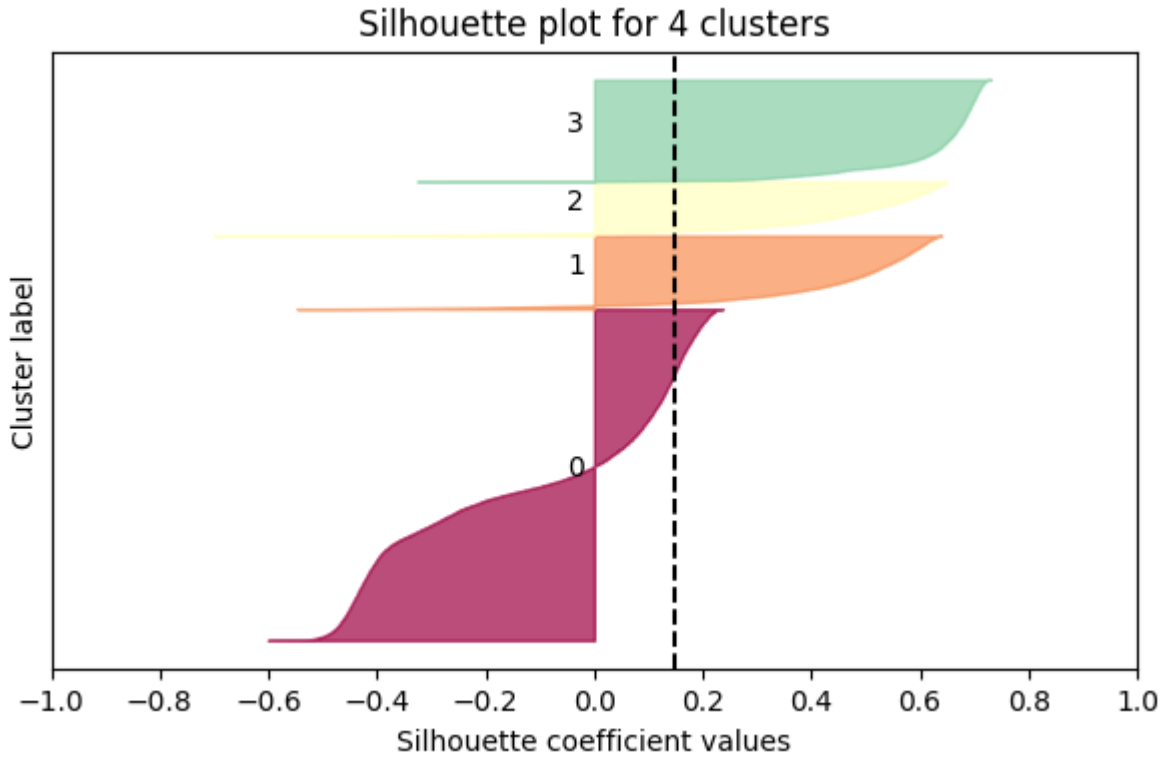
[46] INERTIA AND SILHOUETTE PLOT FOR K-MEANS CLUSTERING WITH ROBUST SCALING



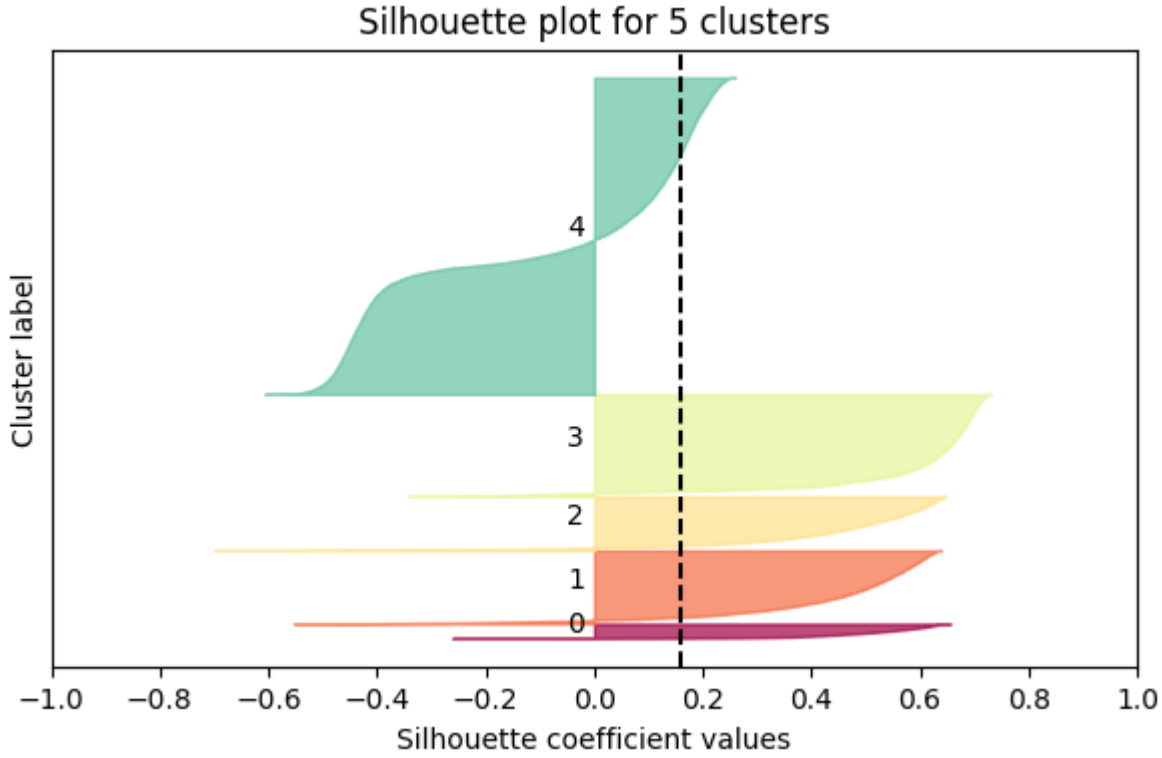
[47] SILHOUETTE PLOT FOR K-MEANS CLUSTERING WITH ROBUST SCALING AND 3 CLUSTERS



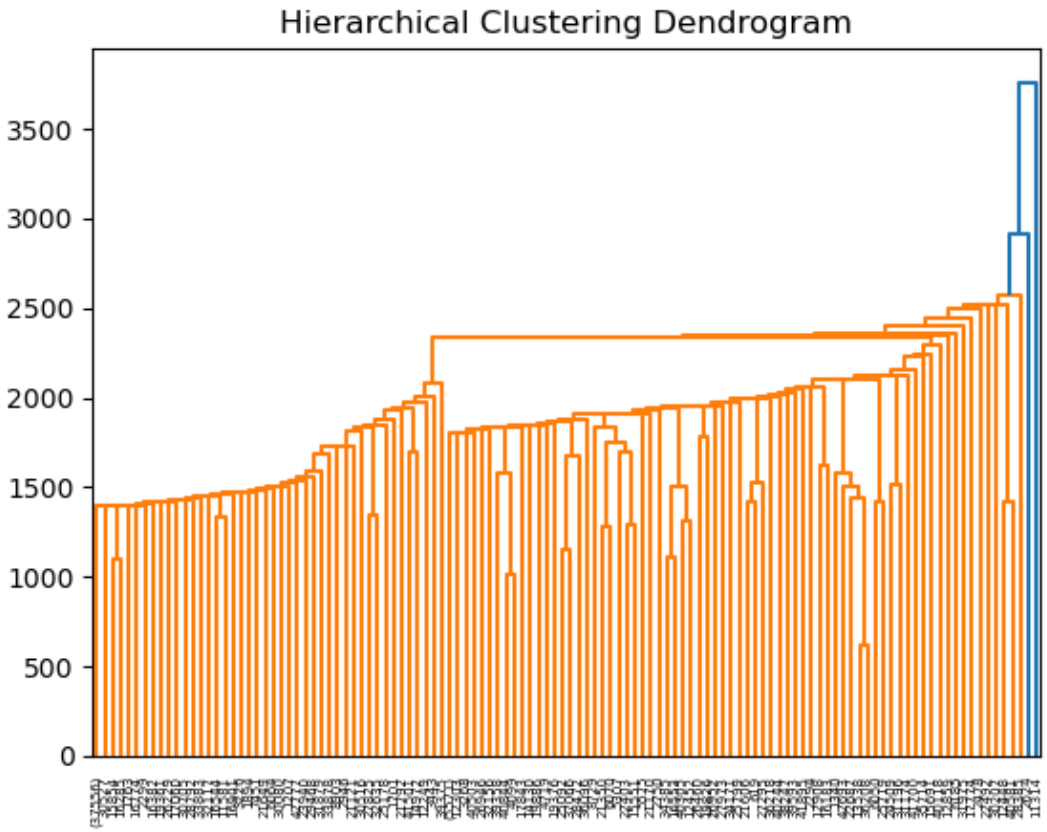
[48] SILHOUETTE PLOT FOR K-MEANS CLUSTERING WITH ROBUST SCALING AND 4 CLUSTERS



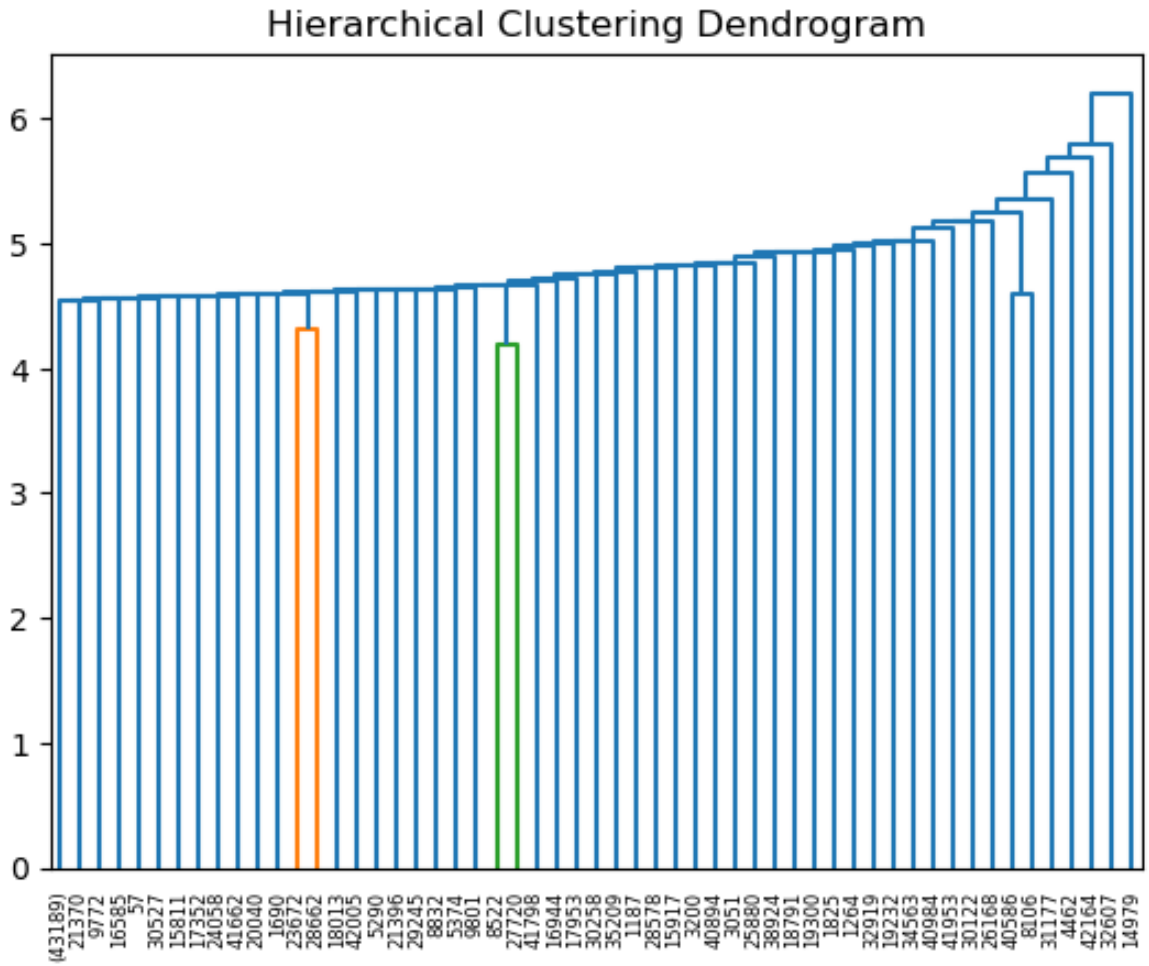
[49] SILHOUETTE PLOT FOR K-MEANS CLUSTERING WITH ROBUST SCALING AND 5 CLUSTERS



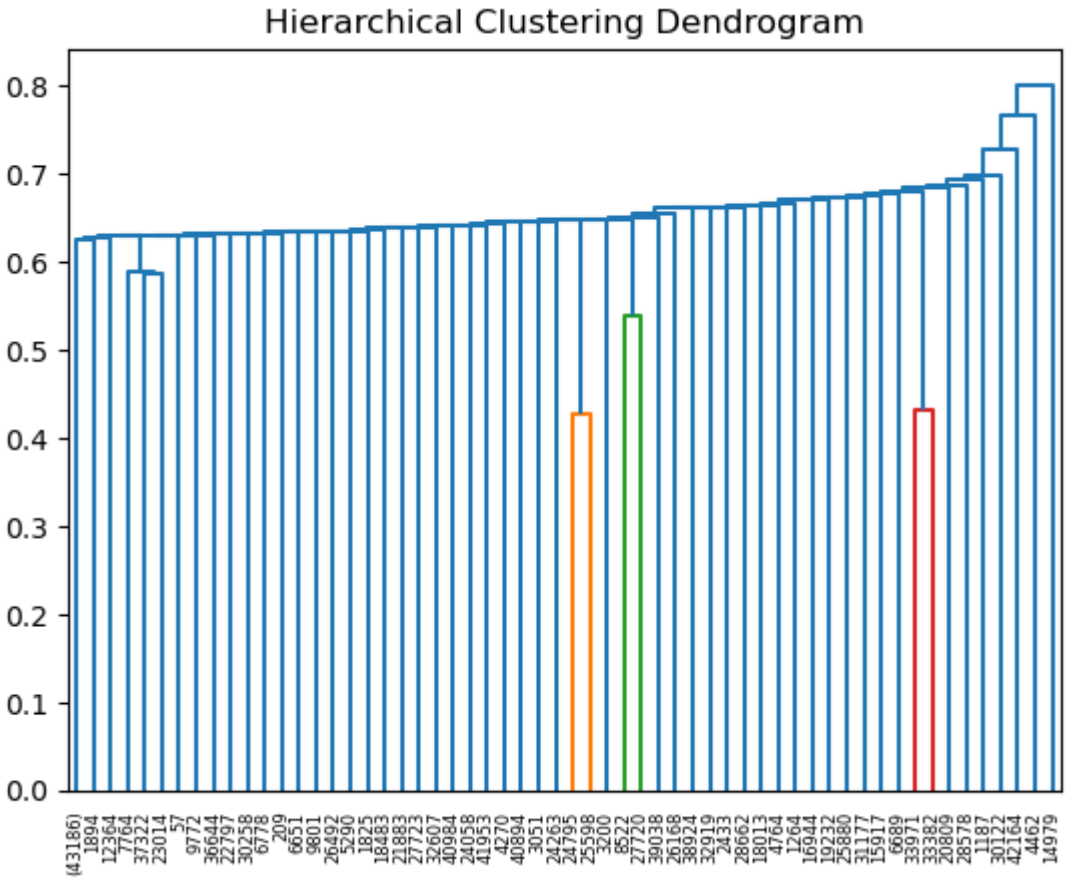
[50] HIERARCHICAL CLUSTERING DENDROGRAM: SINGLE-LINKAGE WITHOUT SCALING THE DATA



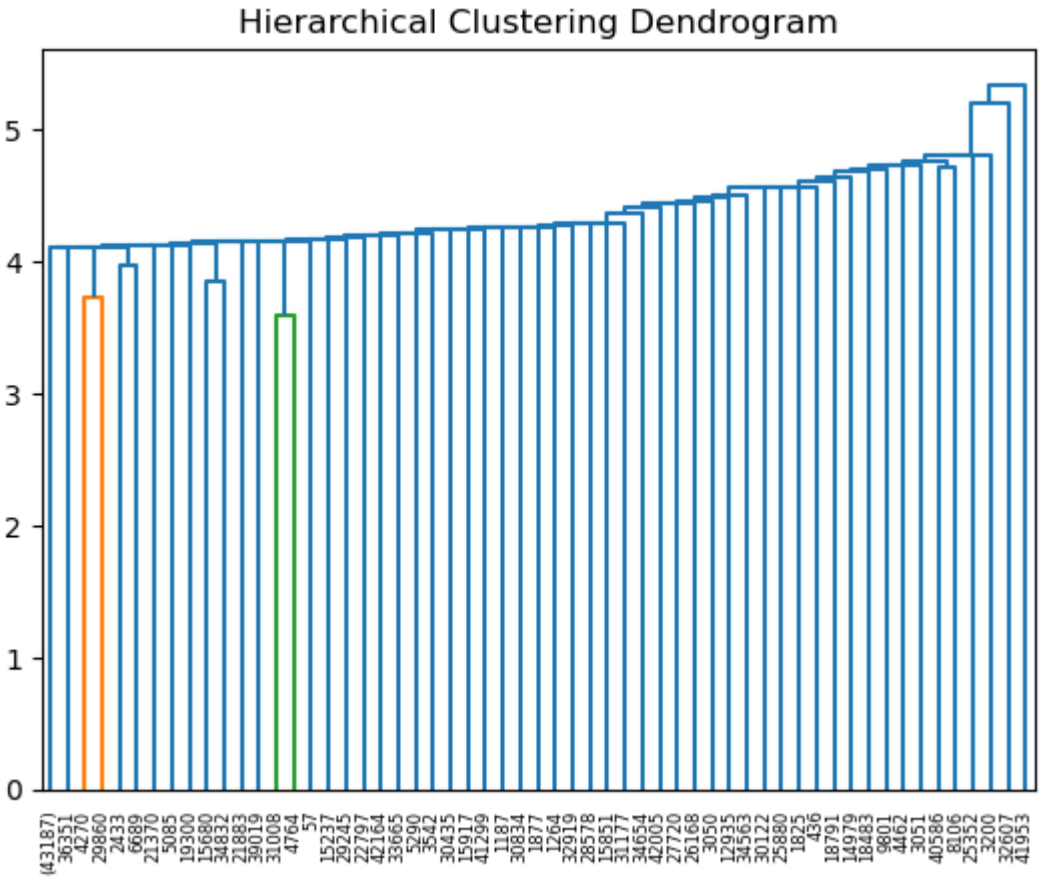
[51] HIERARCHICAL CLUSTERING DENDROGRAM: SINGLE-LINKAGE WITH STANDARD SCALING



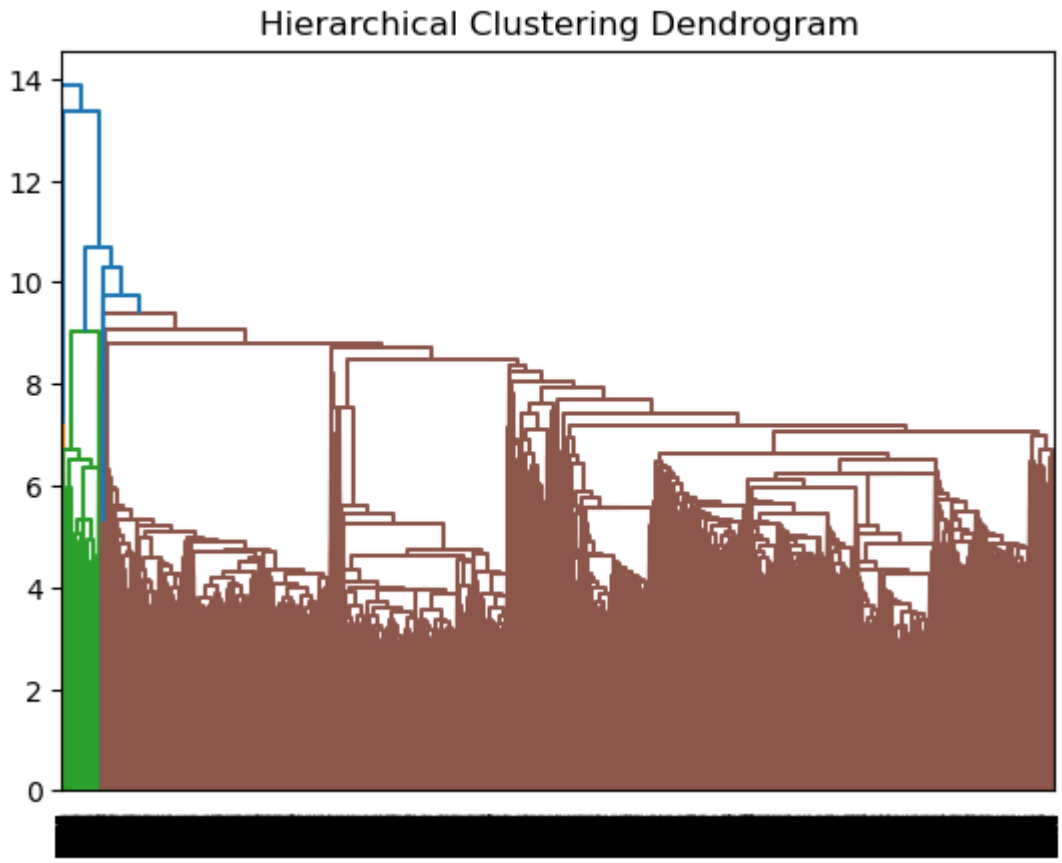
[52] HIERARCHICAL CLUSTERING DENDROGRAM: SINGLE-LINKAGE WITH MIN MAX SCALING



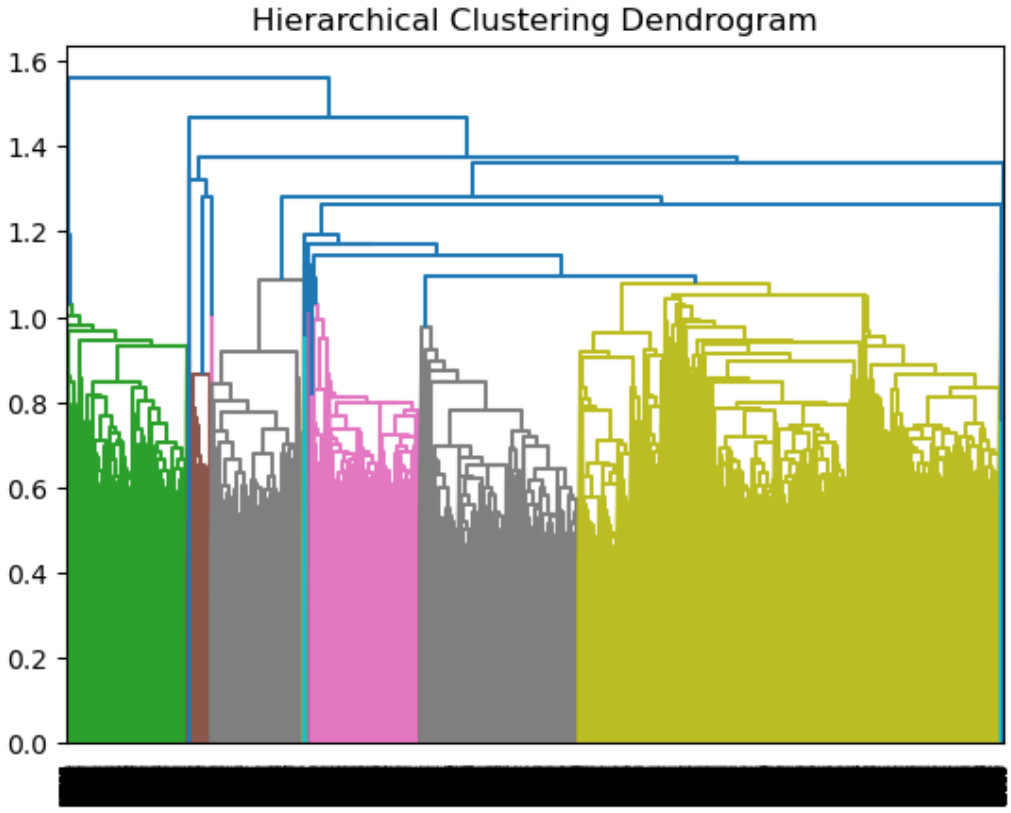
[53] HIERARCHICAL CLUSTERING DENDROGRAM: SINGLE-LINKAGE WITH ROBUST SCALING



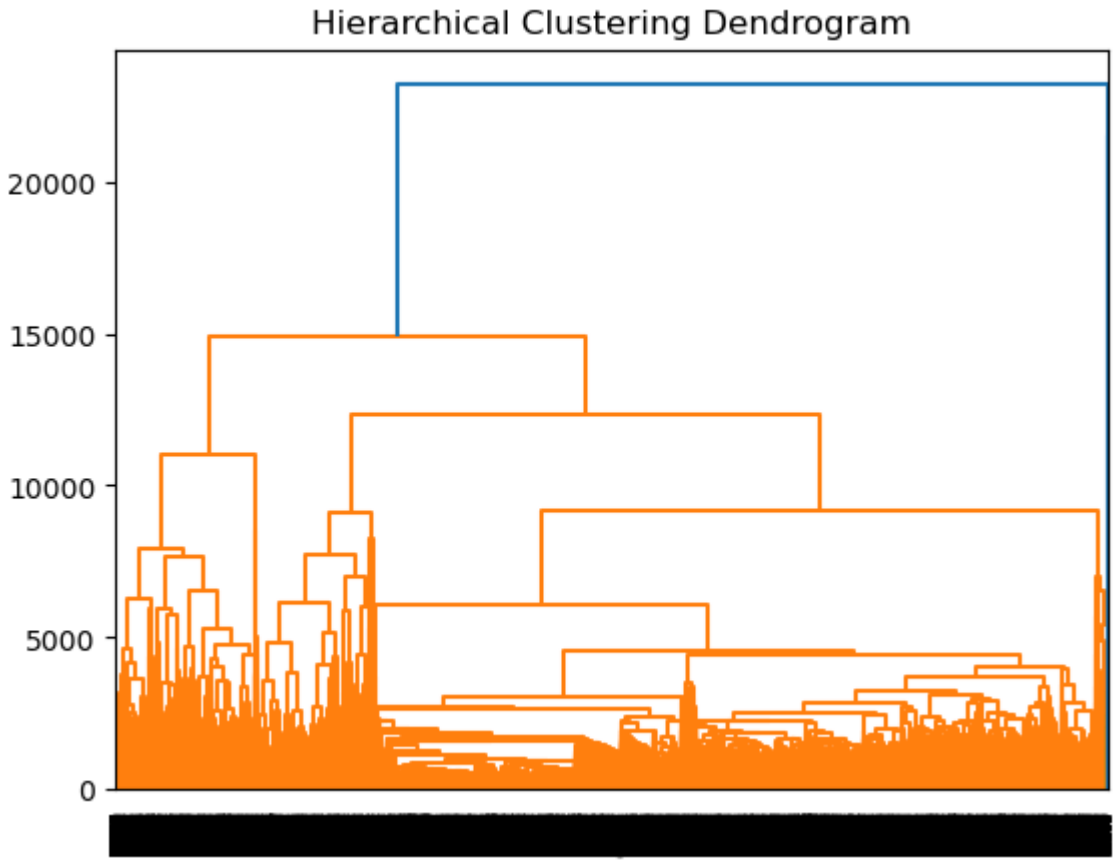
[54] HIERARCHICAL CLUSTERING DENDROGRAM: AVERAGE-LINKAGE WITH STANDARD SCALING



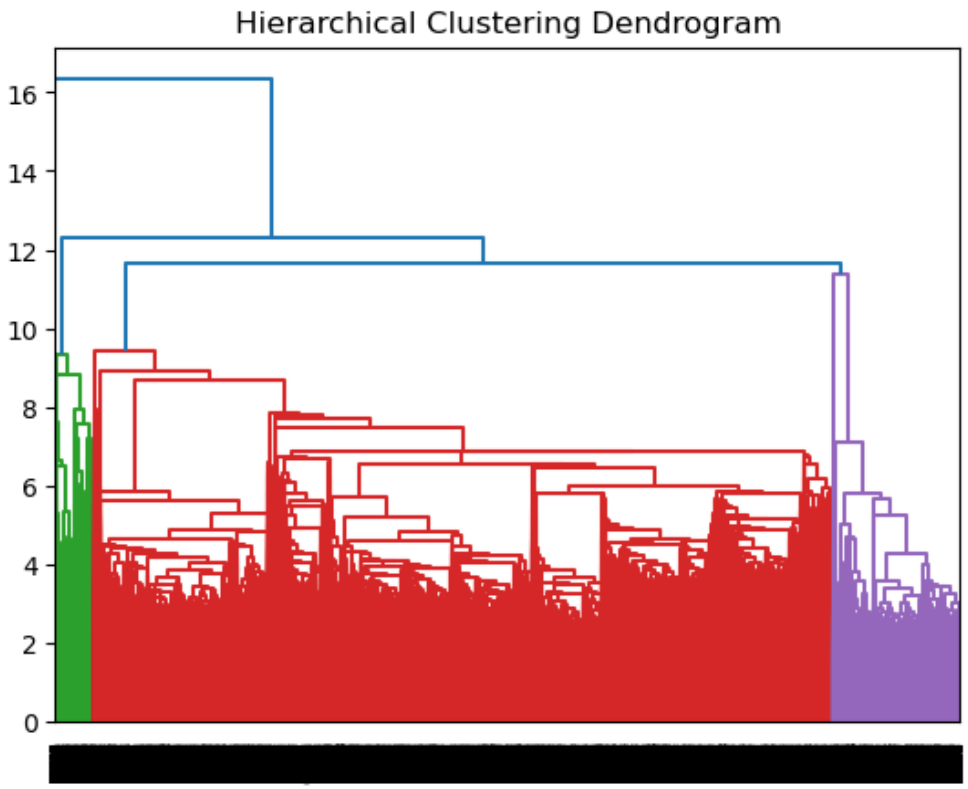
[55] HIERARCHICAL CLUSTERING DENDROGRAM: AVERAGE-LINKAGE WITH MIN MAX SCALING



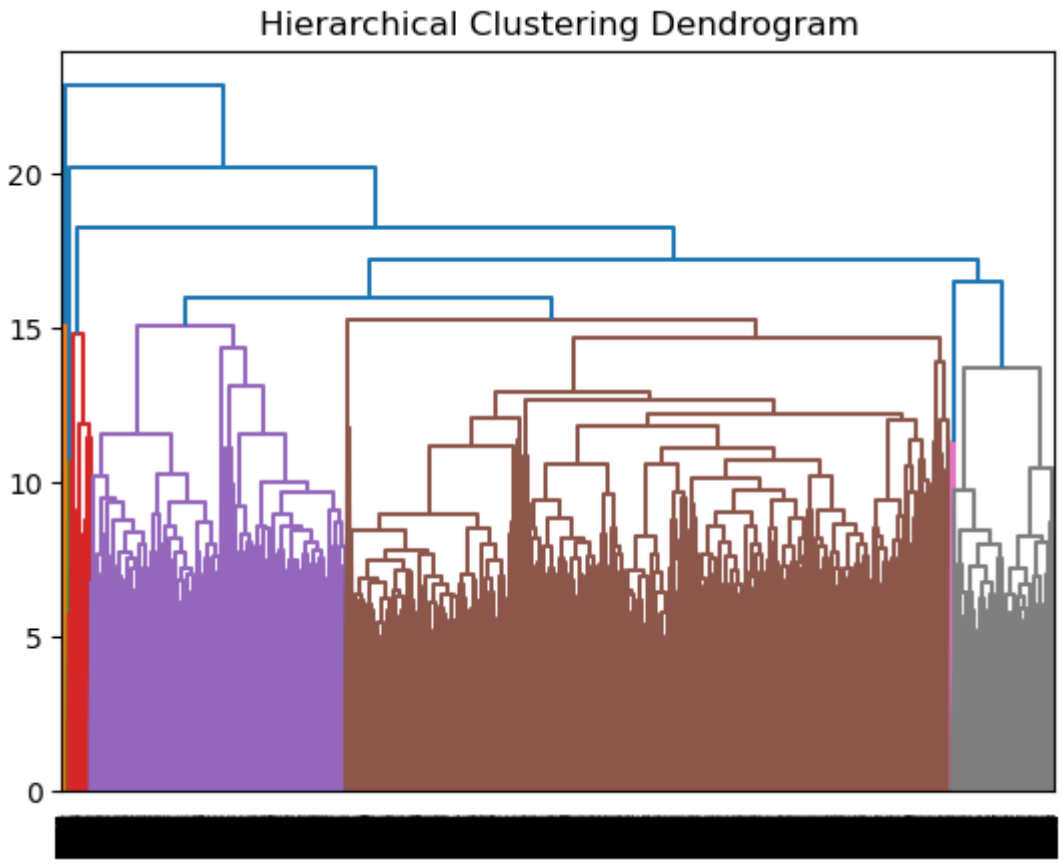
[56] HIERARCHICAL CLUSTERING DENDROGRAM: AVERAGE-LINKAGE WITHOUT SCALING THE DATA



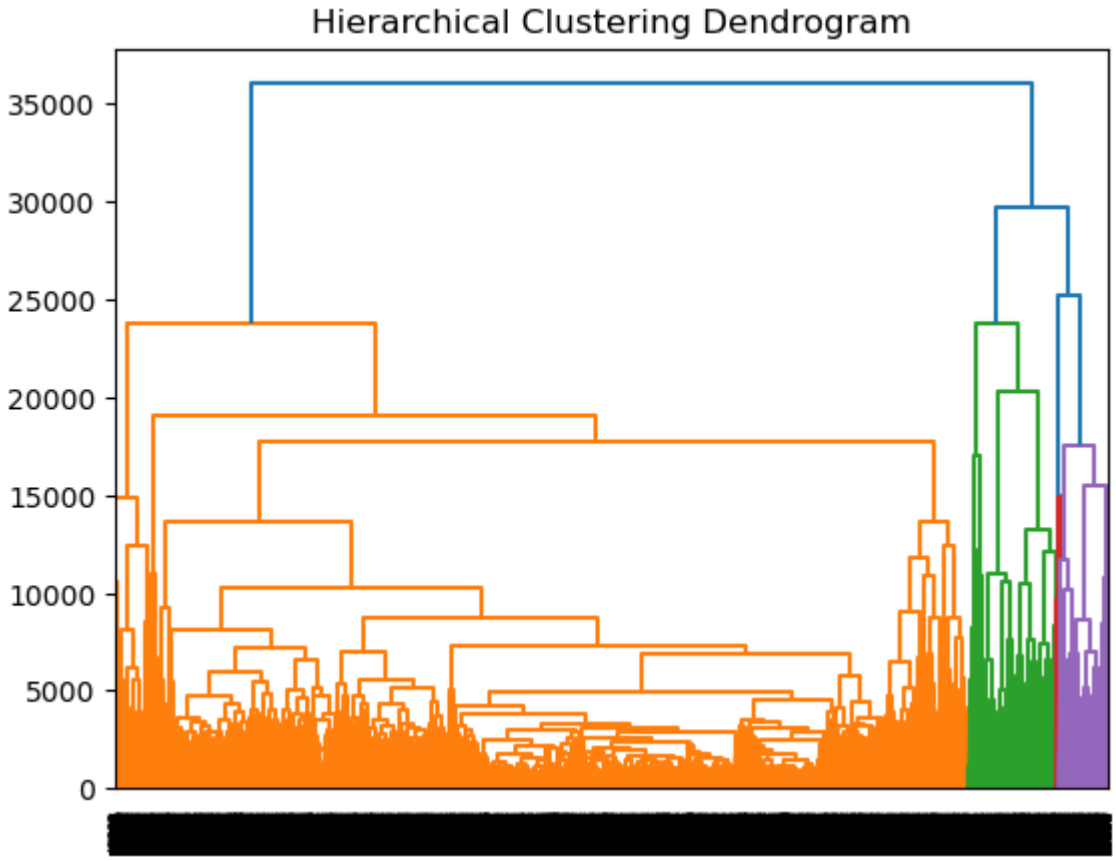
[57] HIERARCHICAL CLUSTERING DENDROGRAM: AVERAGE-LINKAGE WITH ROBUST SCALING



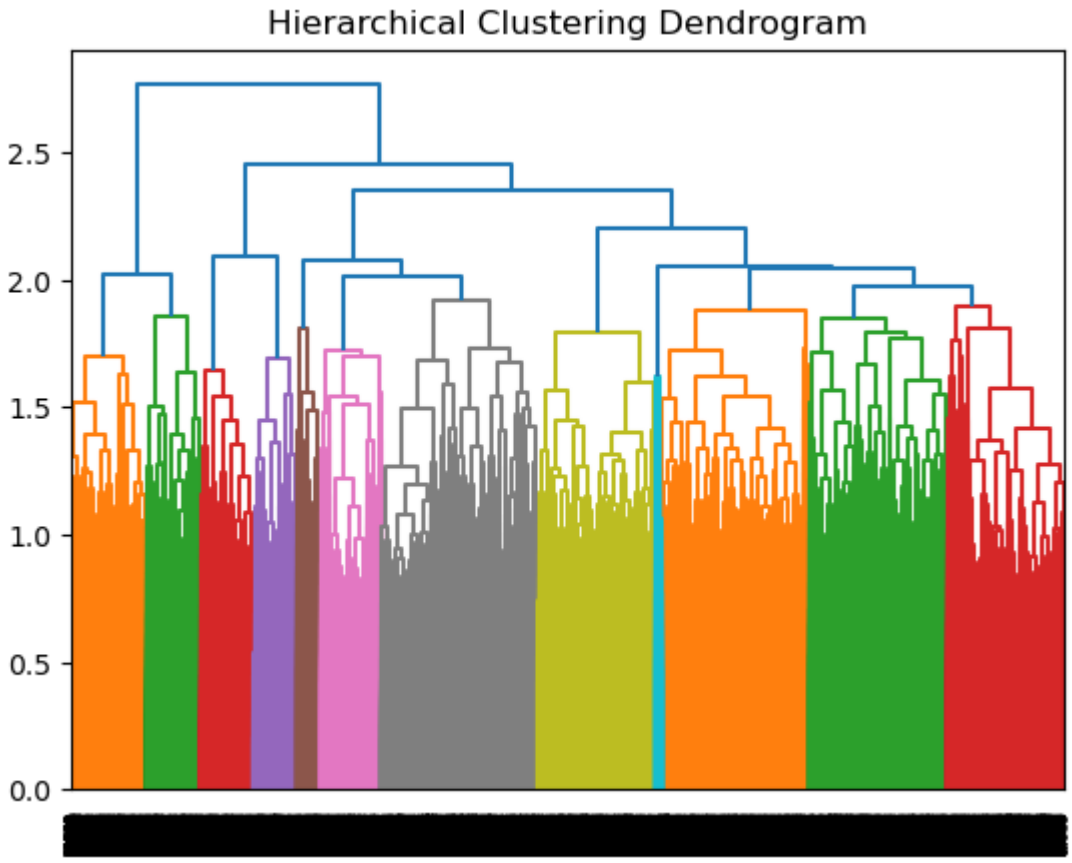
[58] HIERARCHICAL CLUSTERING DENDROGRAM: COMPLETE-LINKAGE WITH STANDARD SCALING



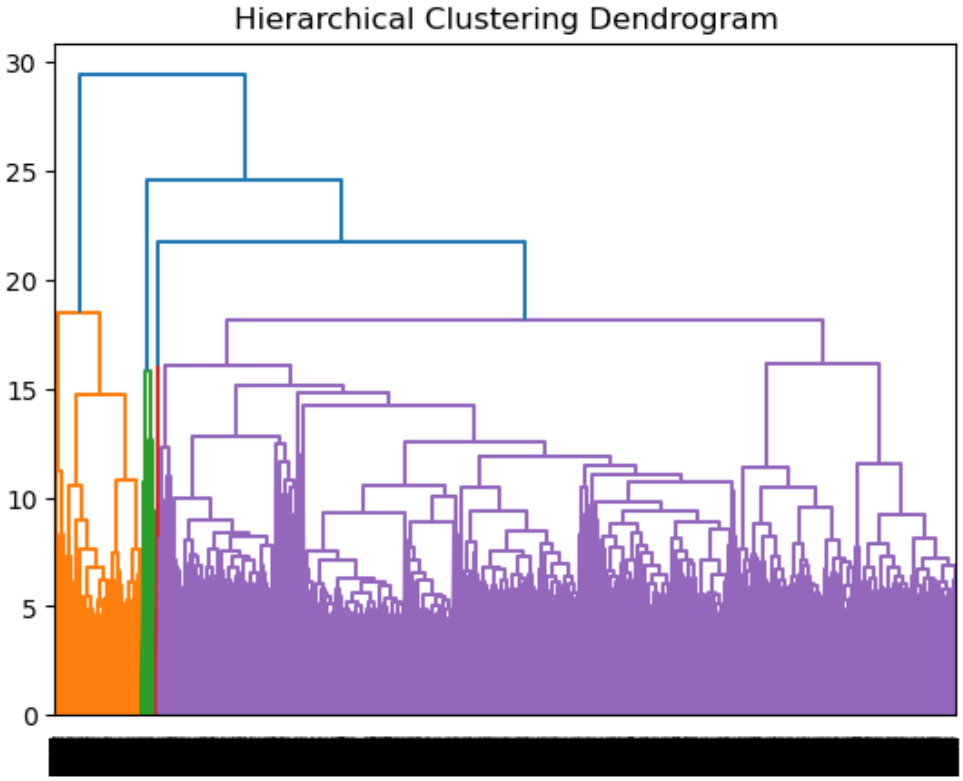
[59] HIERARCHICAL CLUSTERING DENDROGRAM: COMPLETE-LINKAGE WITHOUT SCALING THE DATA



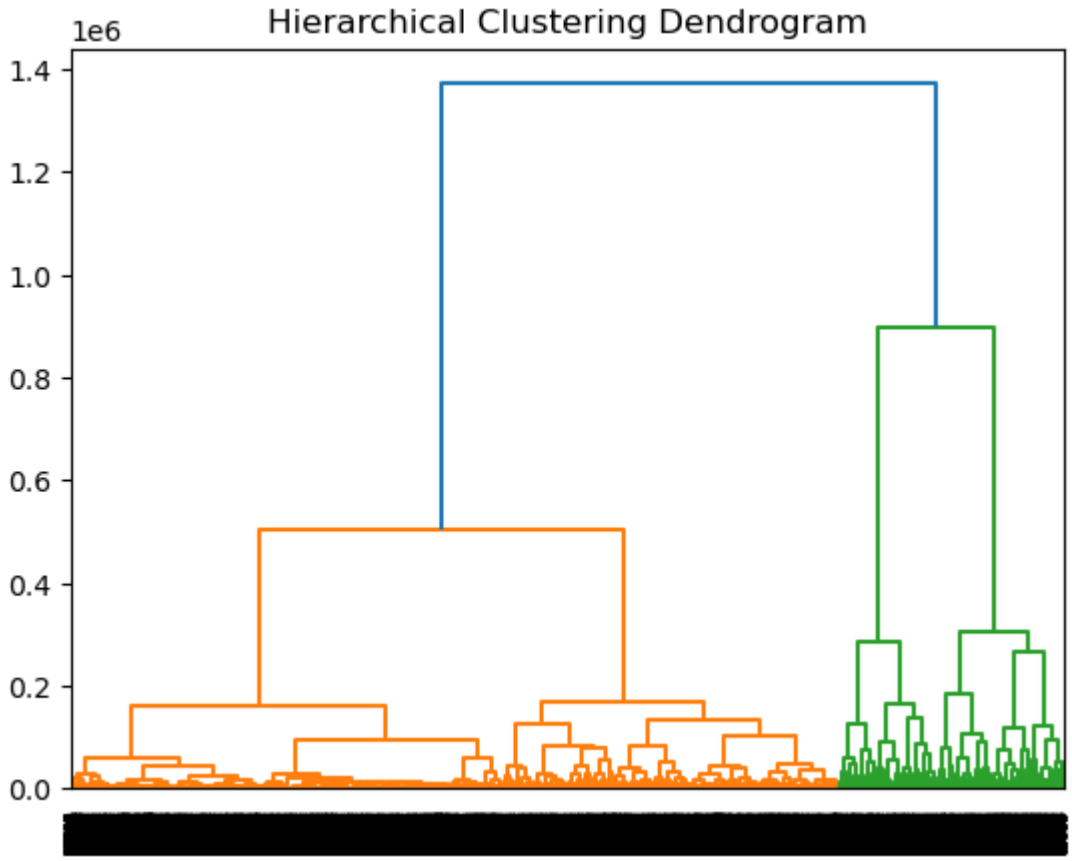
[60] HIERARCHICAL CLUSTERING DENDROGRAM: COMPLETE-LINKAGE WITH MIN MAX SCALING



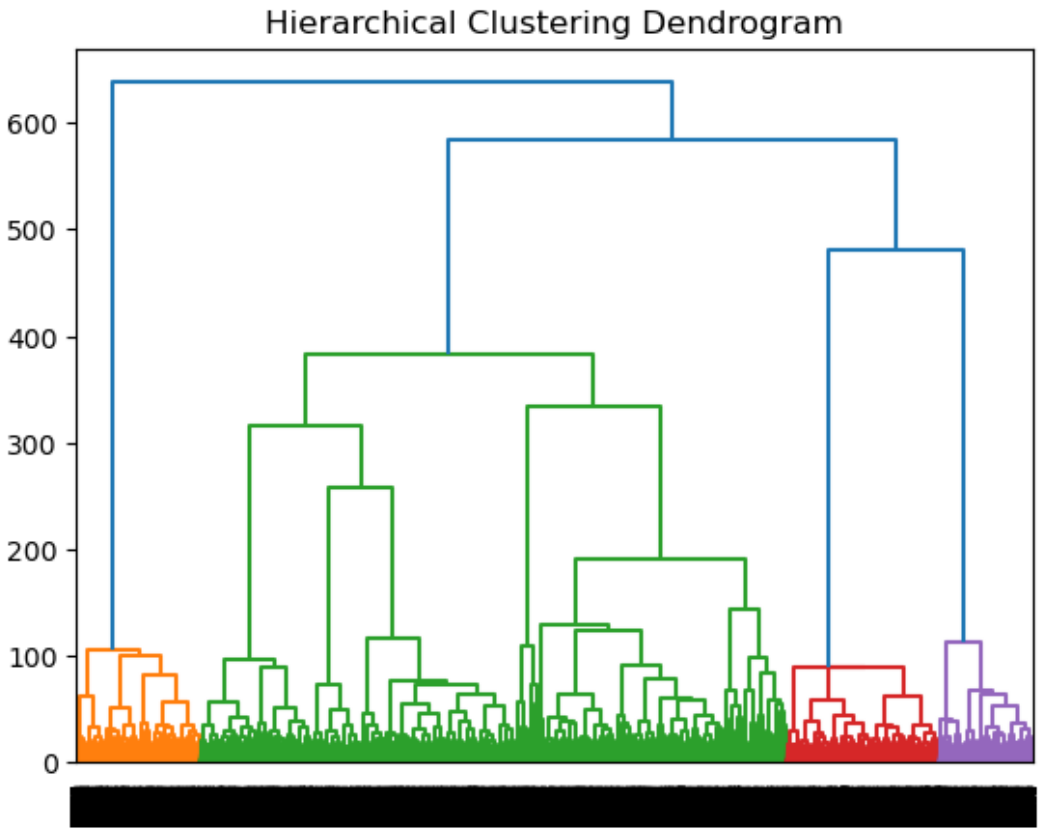
[61] HIERARCHICAL CLUSTERING DENDROGRAM: COMPLETE-LINKAGE WITH ROBUST SCALING



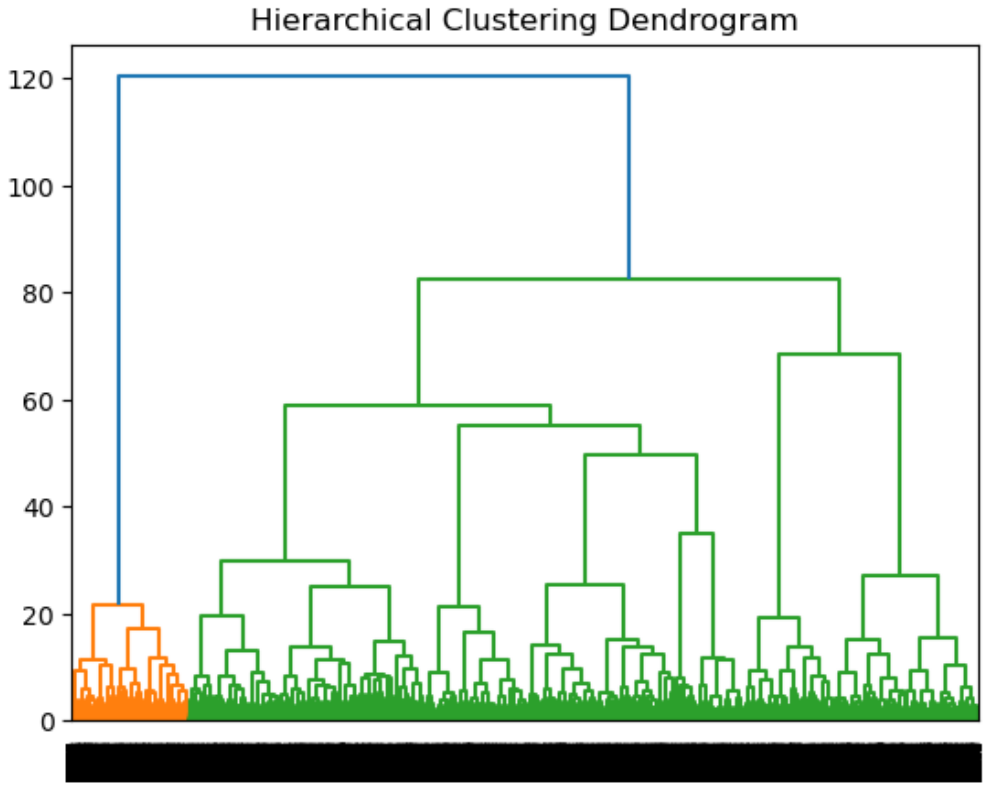
[62] HIERARCHICAL CLUSTERING DENDROGRAM: WARD-LINKAGE WITHOUT SCALING THE DATA



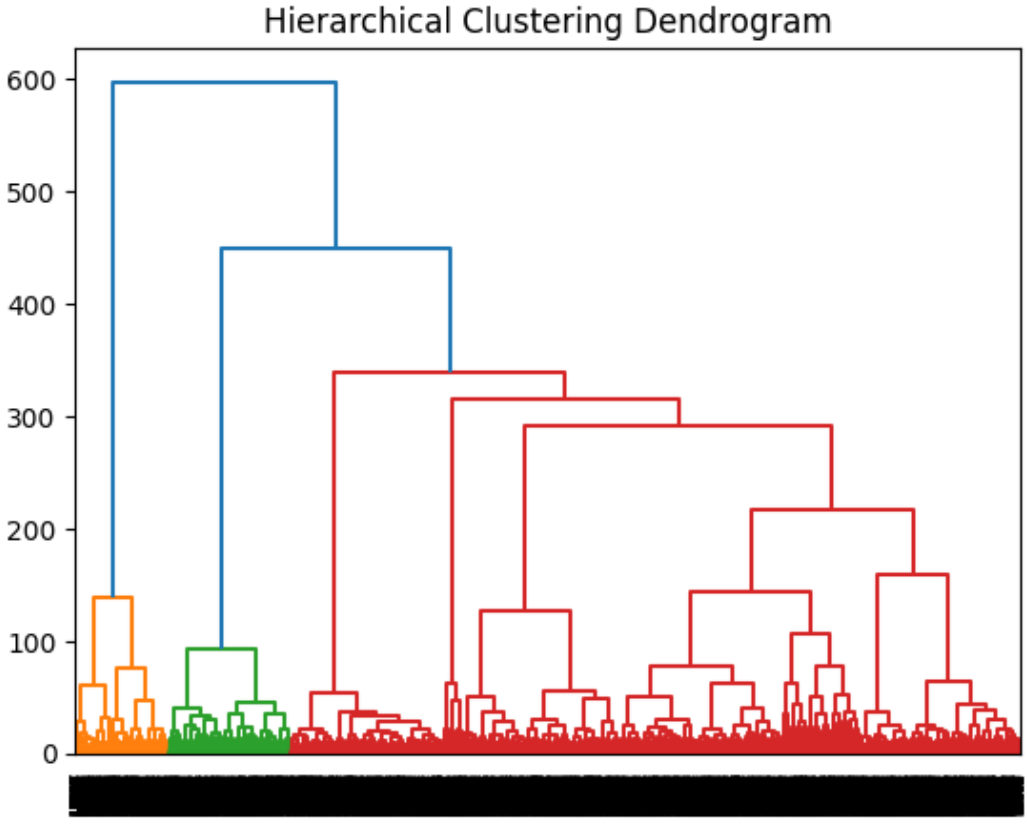
[63] HIERARCHICAL CLUSTERING DENDROGRAM: WARD-LINKAGE WITH STANDARD SCALING



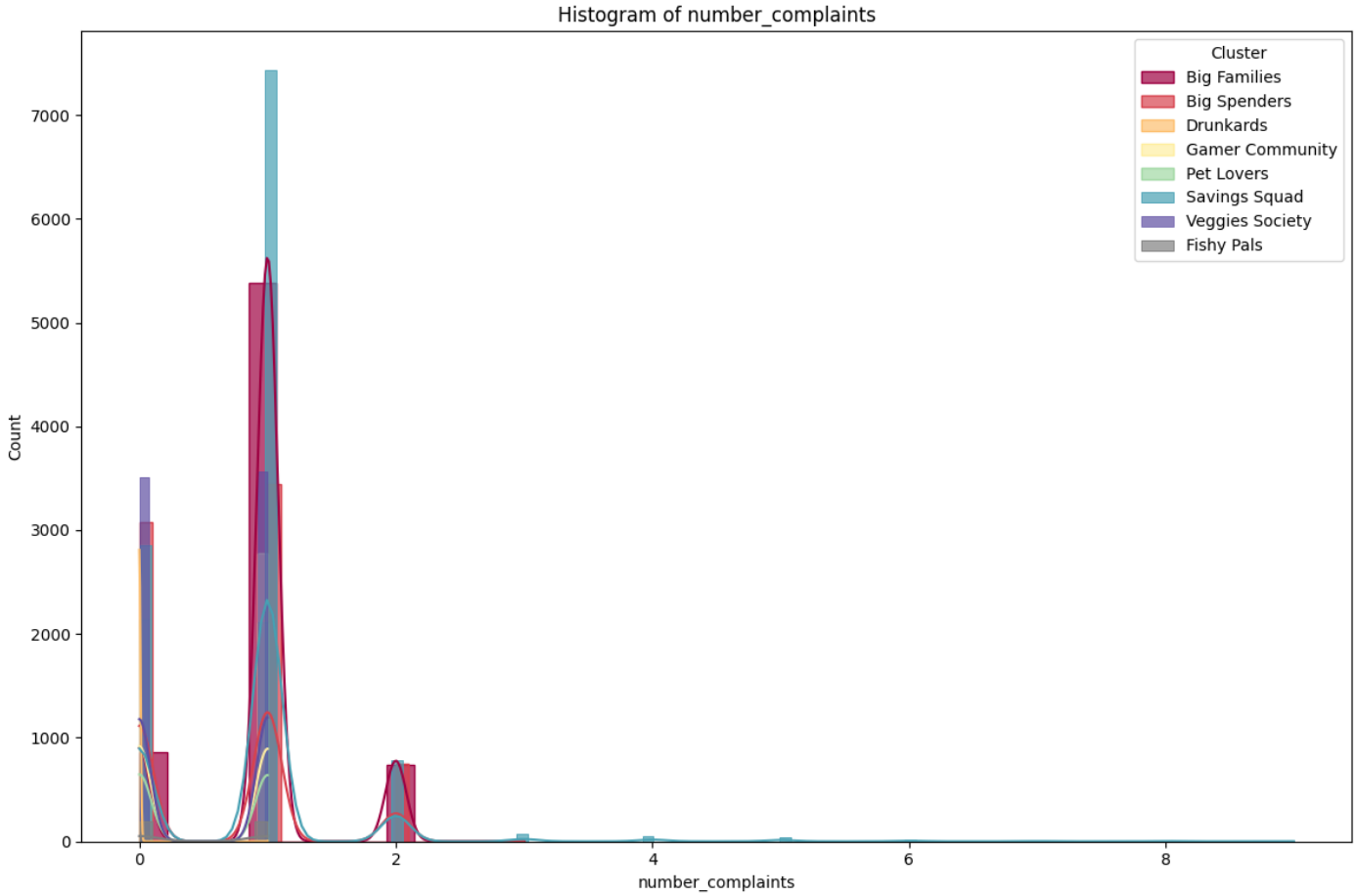
[64] HIERARCHICAL CLUSTERING DENDROGRAM: WARD-LINKAGE WITH MIN MAX SCALING



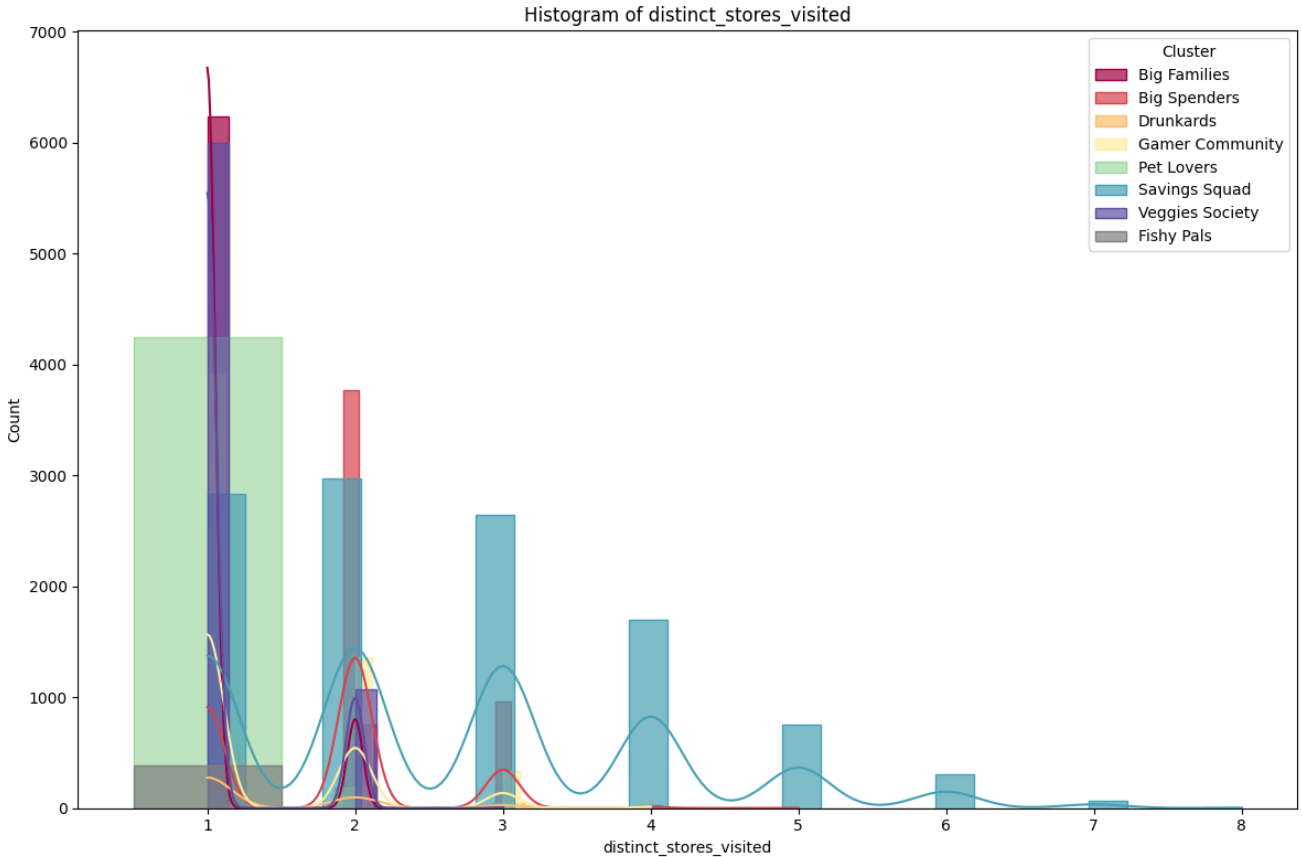
[65] HIERARCHICAL CLUSTERING DENDROGRAM: WARD-LINKAGE WITH ROBUST SCALING



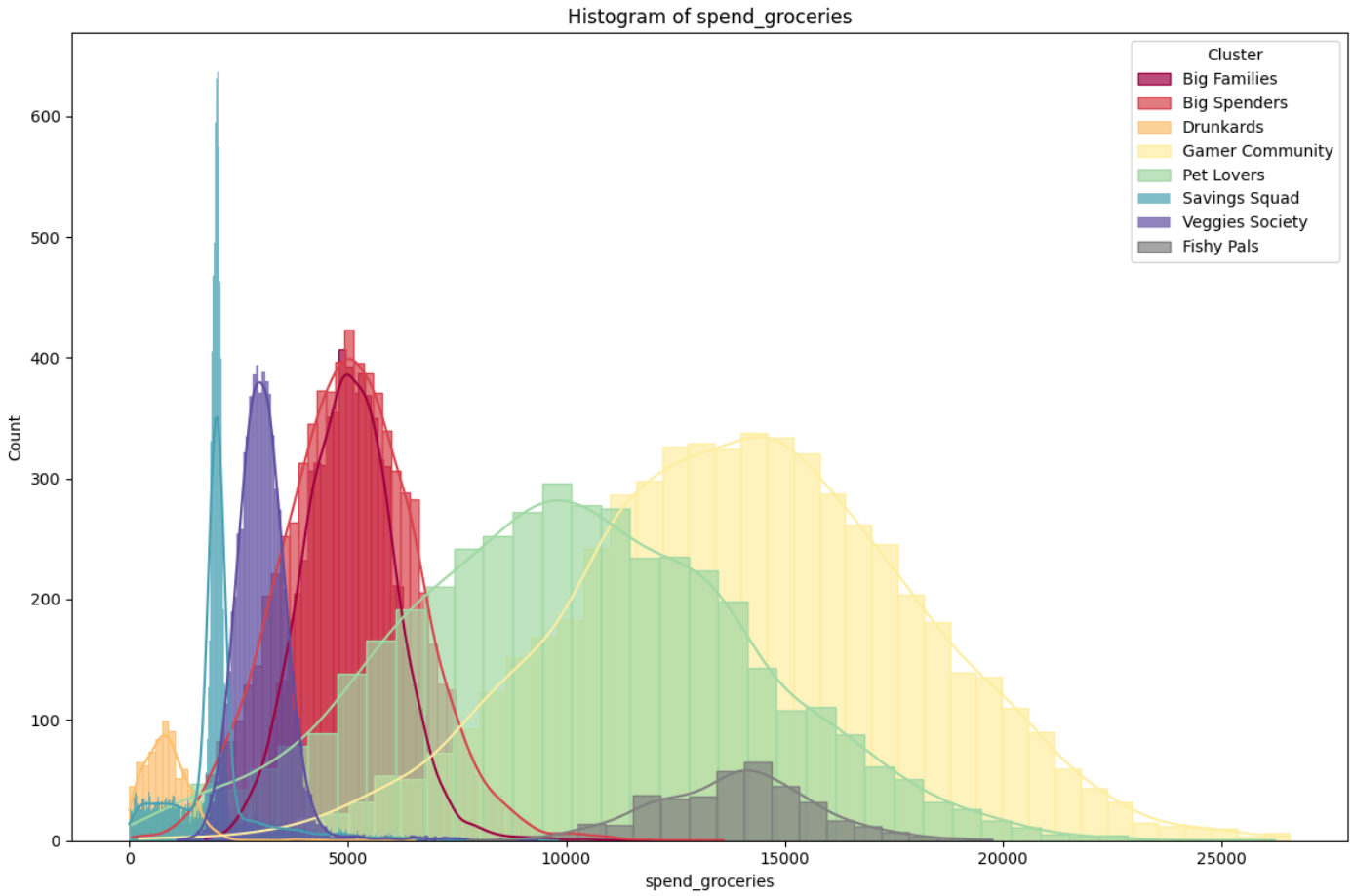
[66] HISTOGRAM OF NUMBER COMPLAINTS FOR THE DIFFERENT CLUSTERS



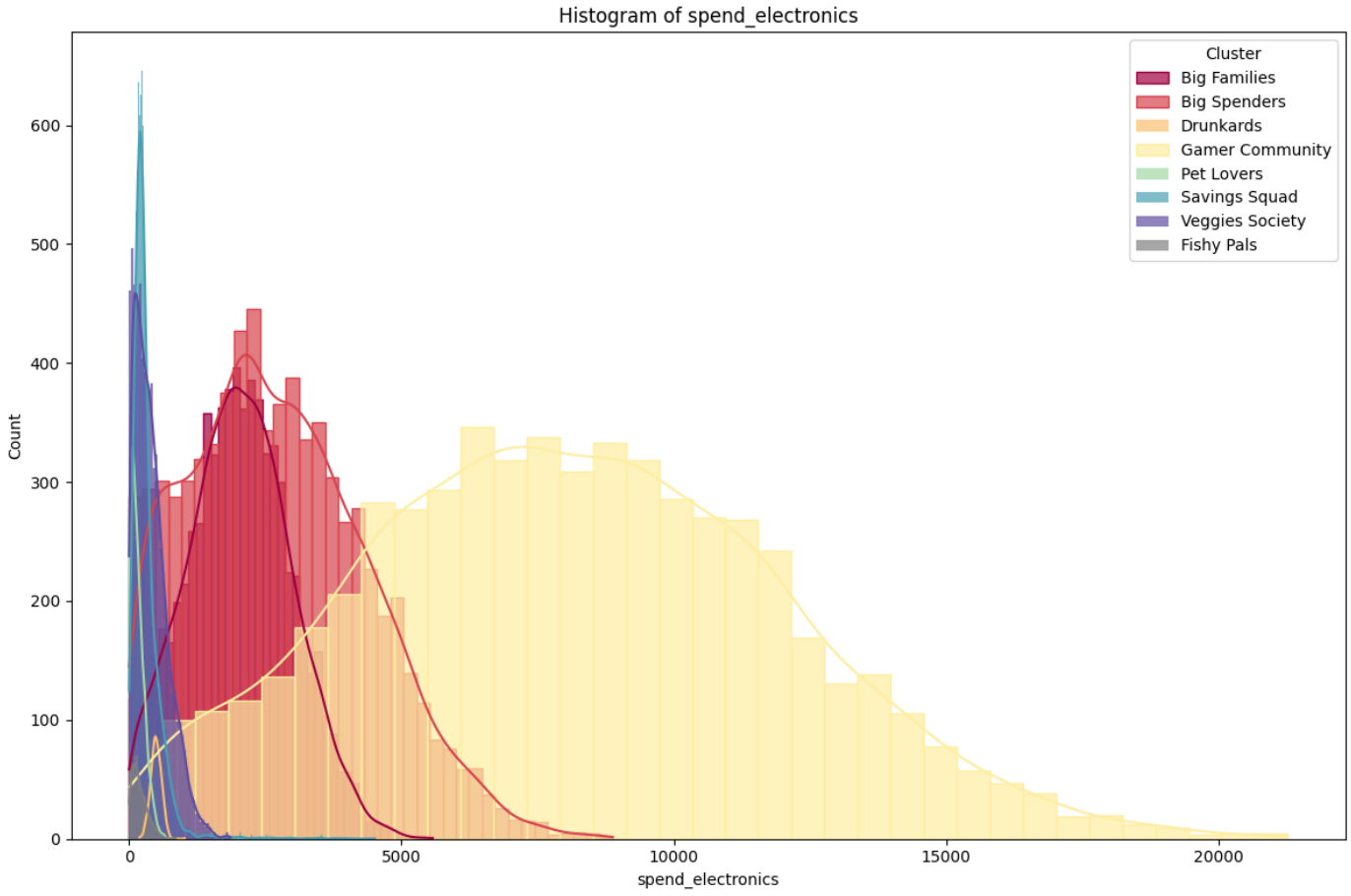
[67] HISTOGRAM OF THE DISTINCT STORES VISITED FOR THE DIFFERENT CLUSTERS



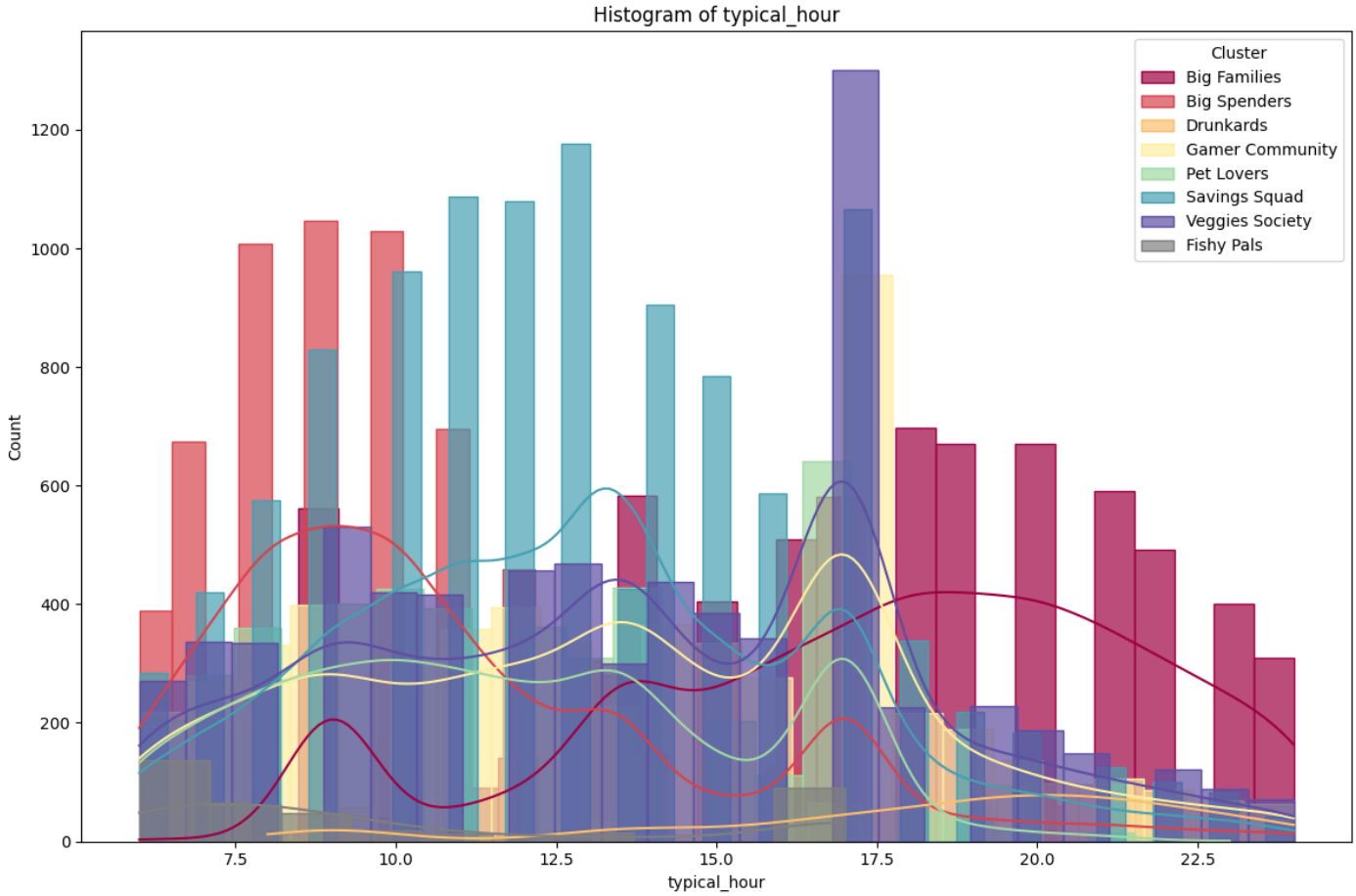
[68] HISTOGRAM OF THE MONEY SPENT ON GROCERIES VISITED FOR THE DIFFERENT CLUSTERS



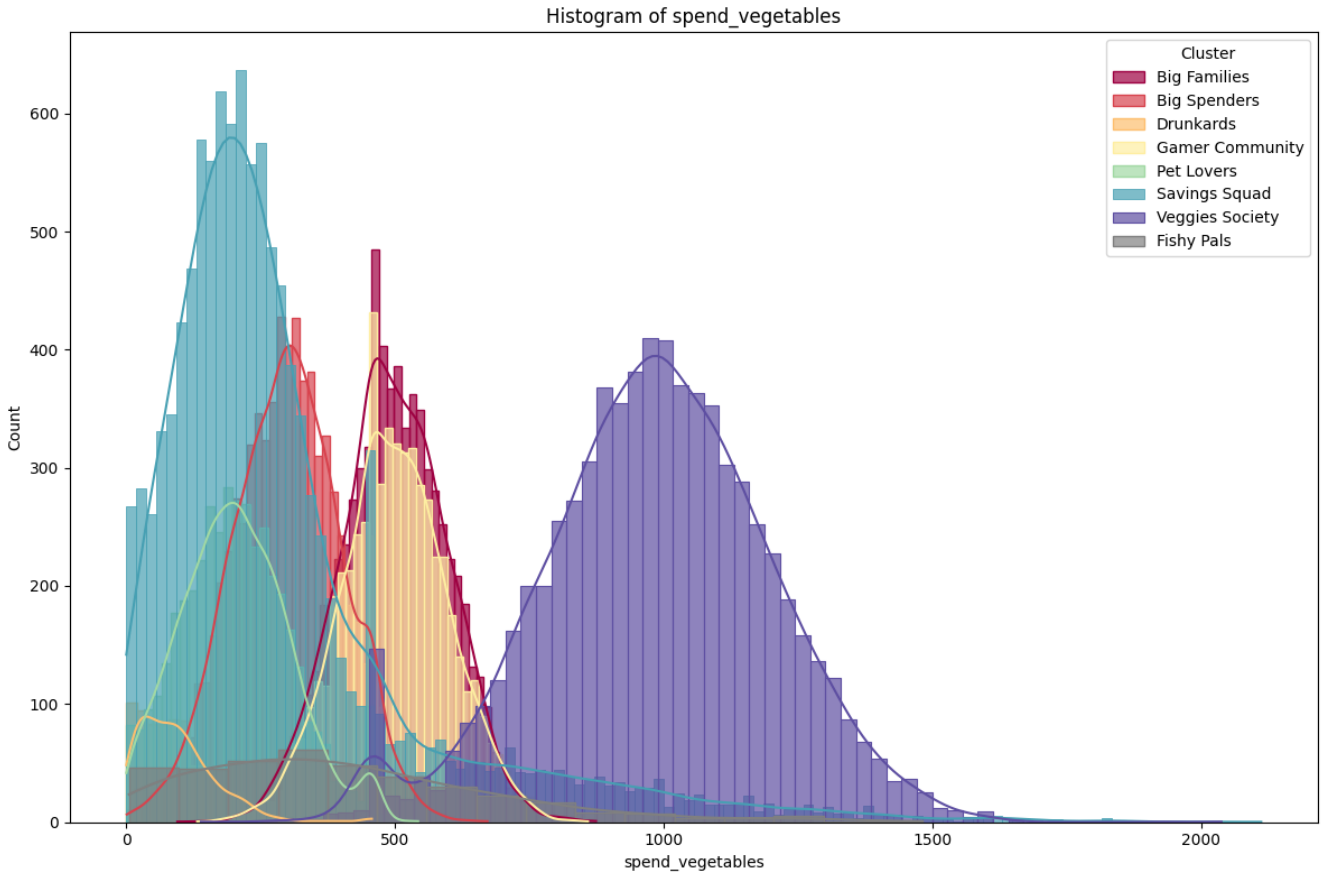
[69] HISTOGRAM OF THE MONEY SPENT ON ELECTRONICS FOR THE DIFFERENT CLUSTERS



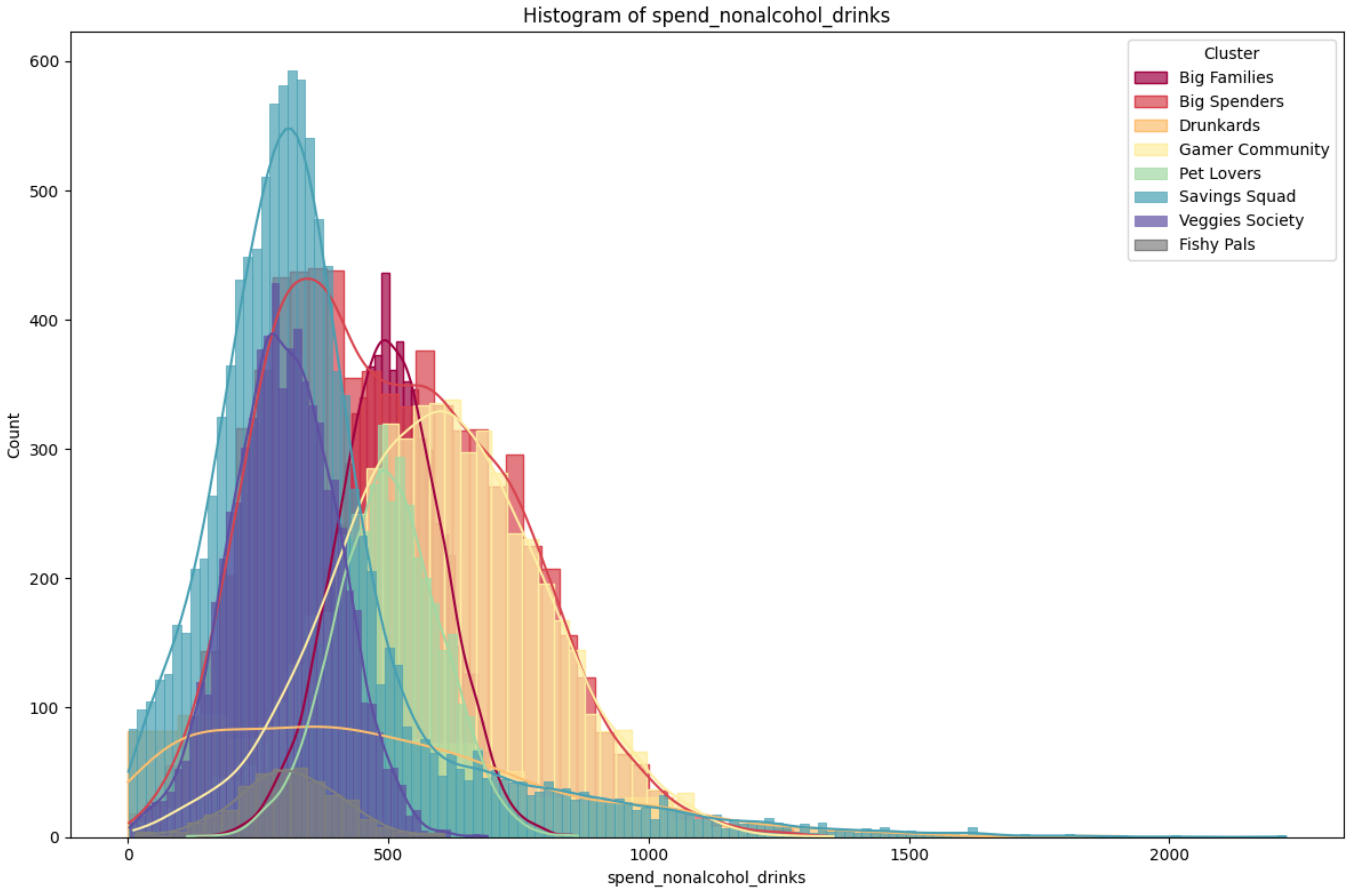
[70] HISTOGRAM OF THE TYPICAL HOUR FOR THE DIFFERENT CLUSTERS



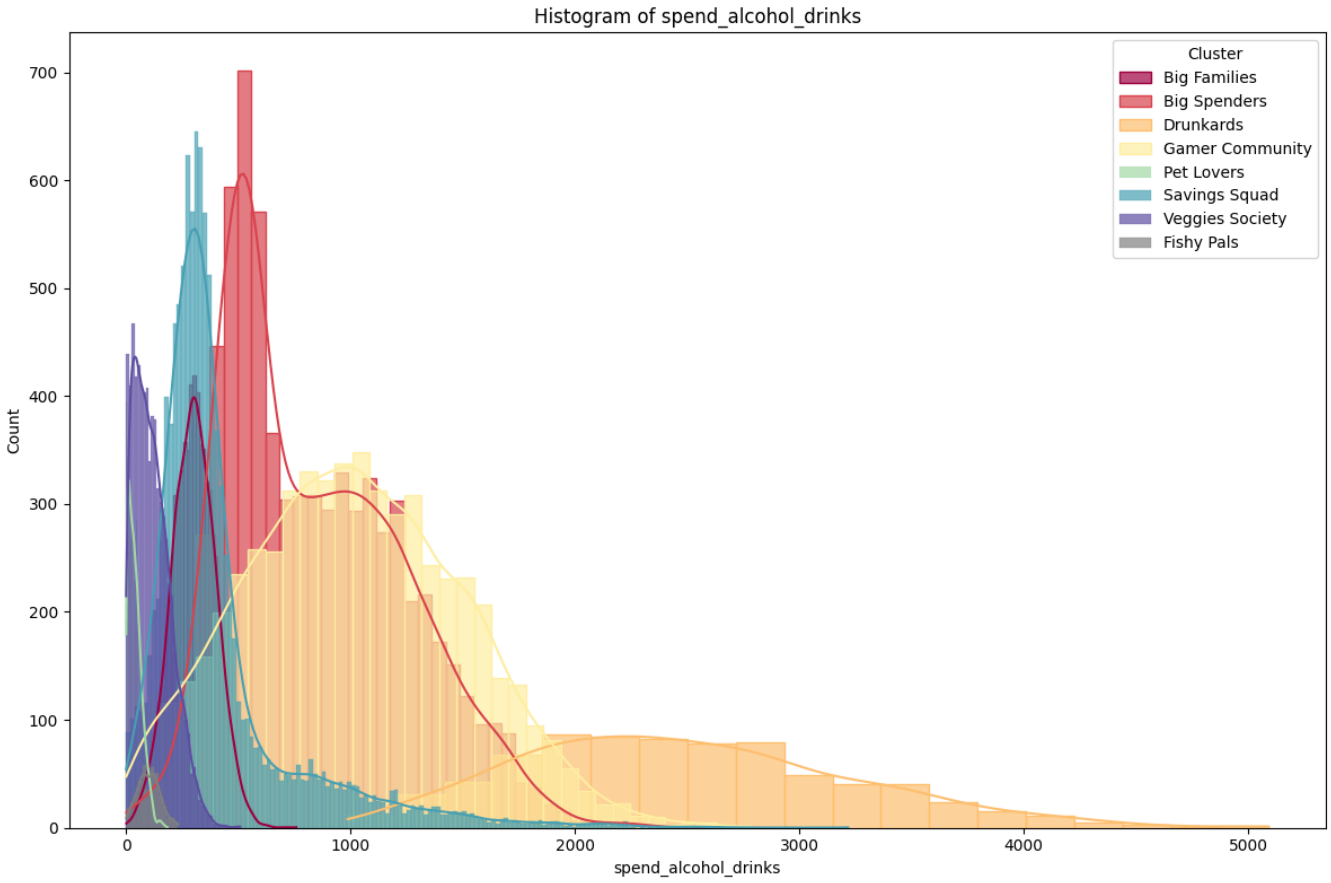
[71] HISTOGRAM OF THE MONEY SPENT ON VEGETABLES FOR THE DIFFERENT CLUSTERS



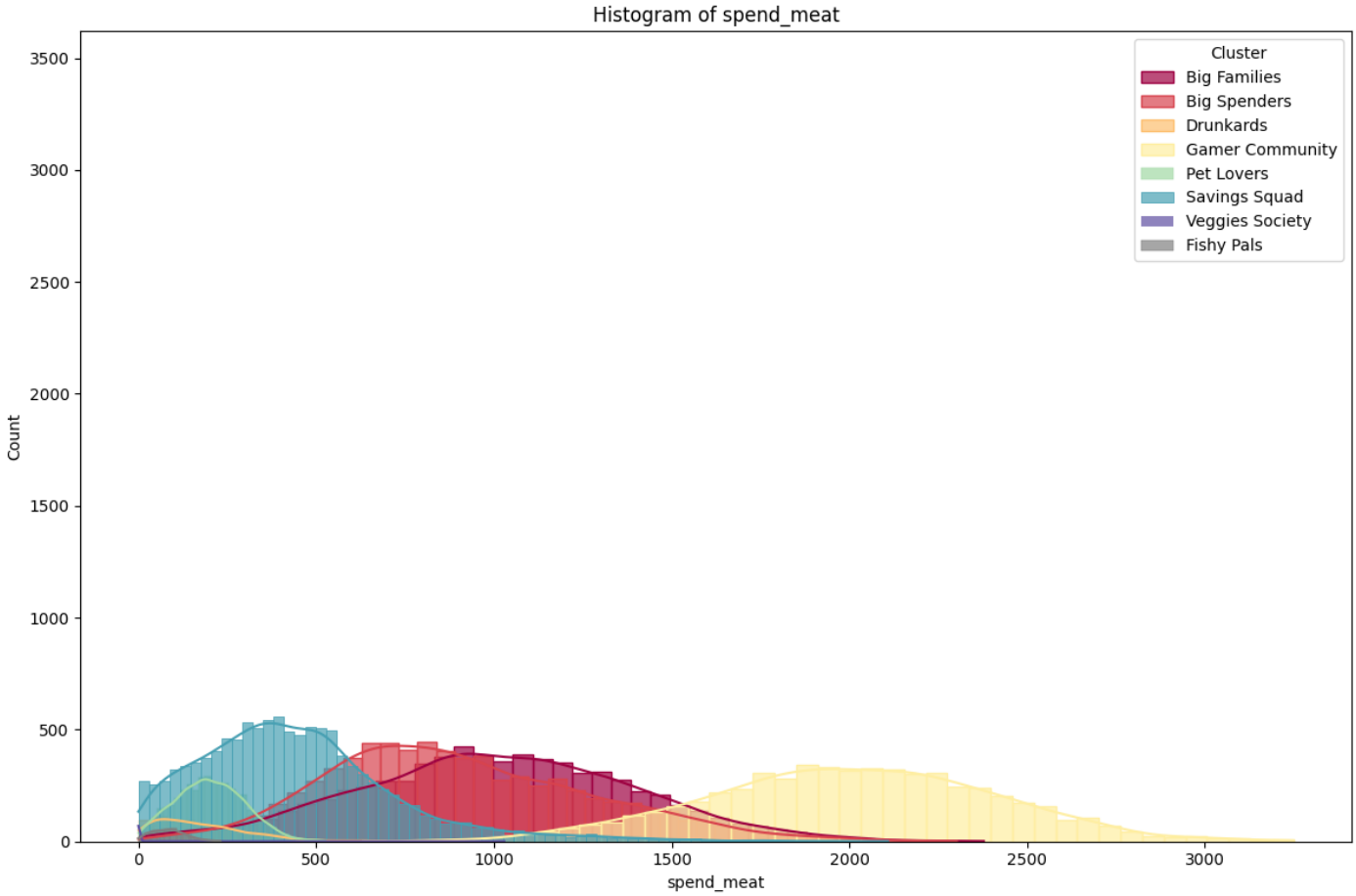
[72] HISTOGRAM OF THE MONEY SPENT ON NON-ALCOHOLIC DRINKS FOR THE DIFFERENT CLUSTERS



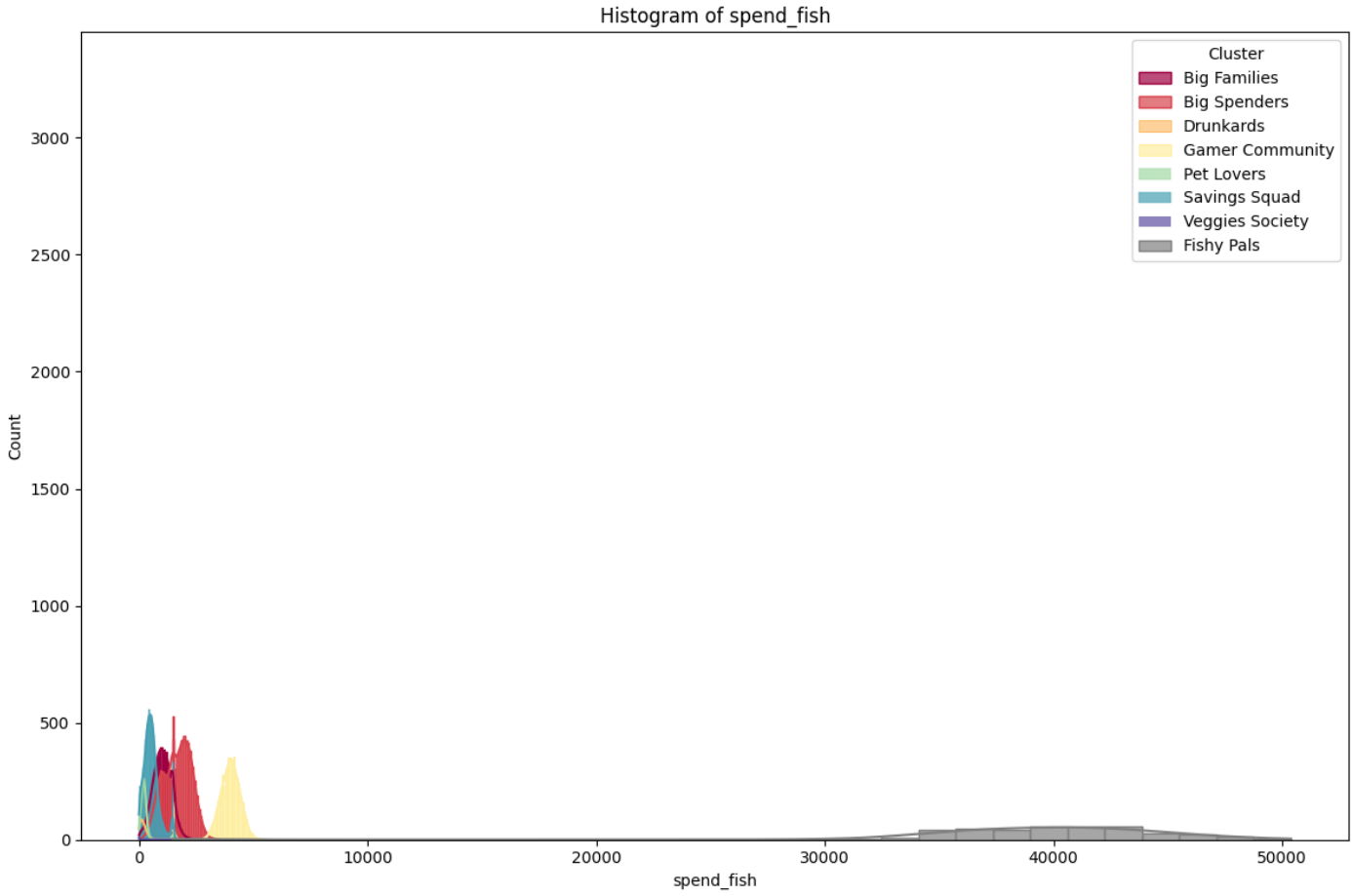
[73] HISTOGRAM OF THE MONEY SPENT ON ALCOHOLIC DRINKS FOR THE DIFFERENT CLUSTERS



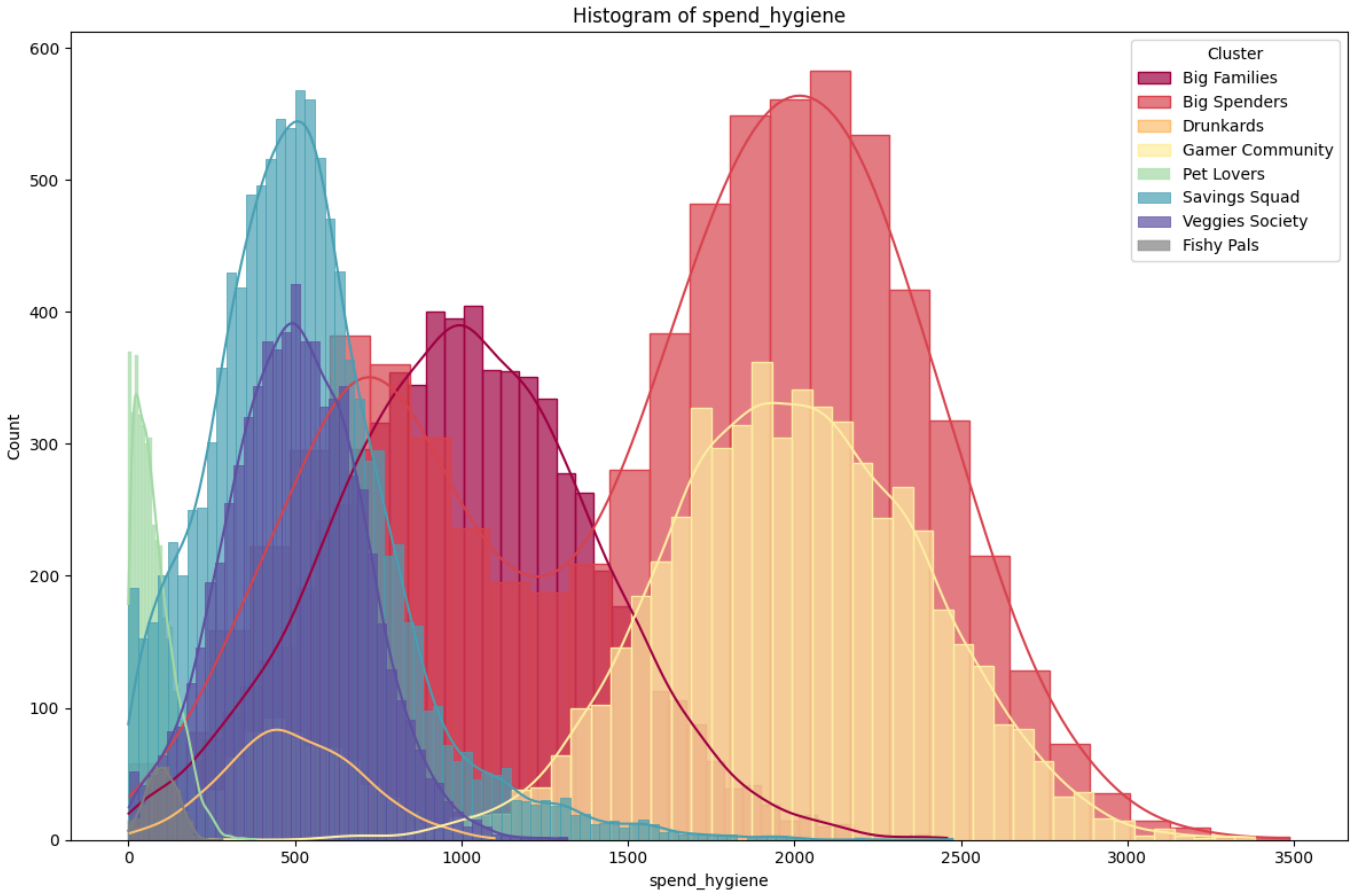
[74] HISTOGRAM OF THE MONEY SPENT ON MEAT FOR THE DIFFERENT CLUSTERS



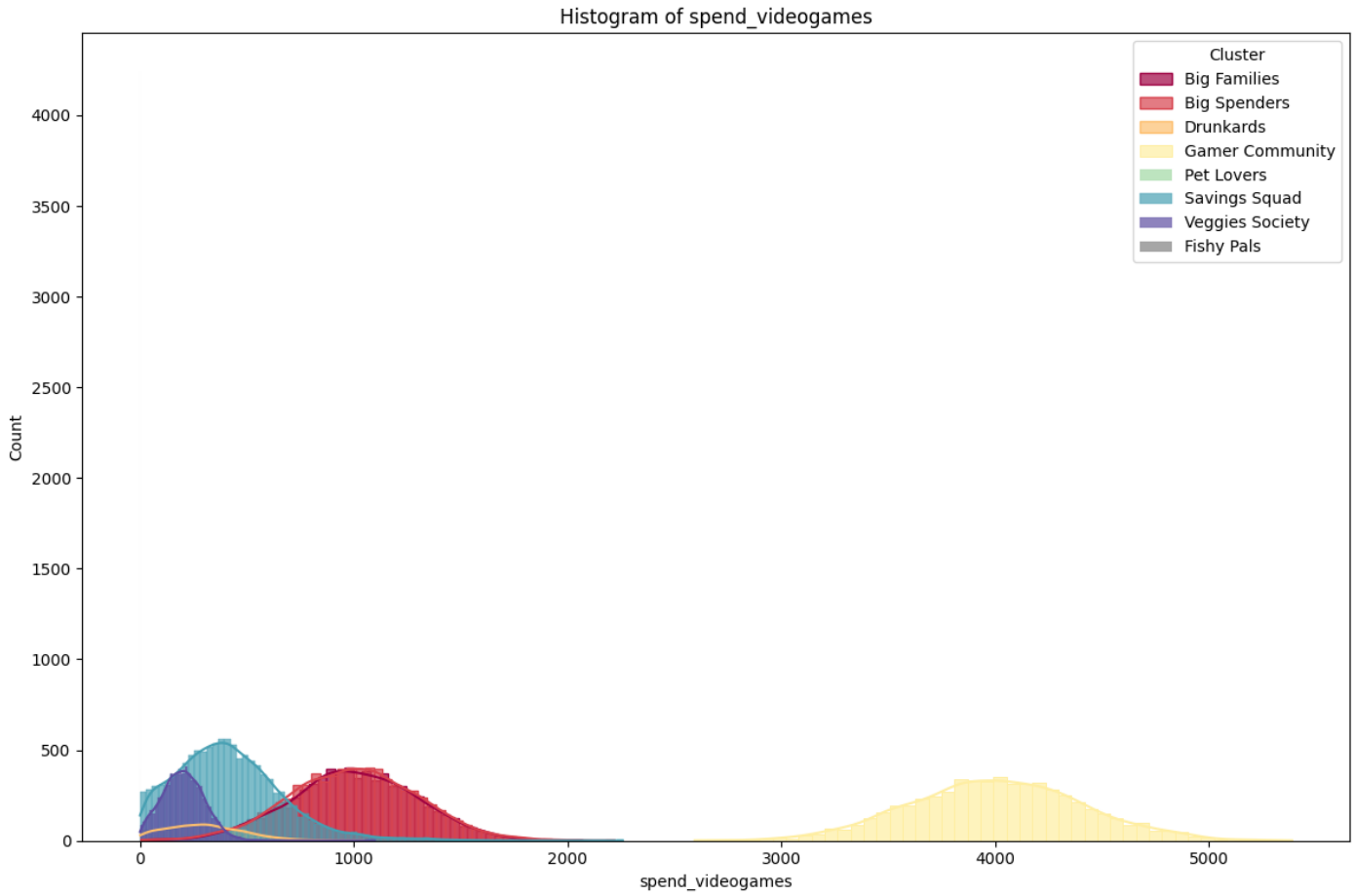
[75] HISTOGRAM OF THE MONEY SPENT ON FISH FOR THE DIFFERENT CLUSTERS



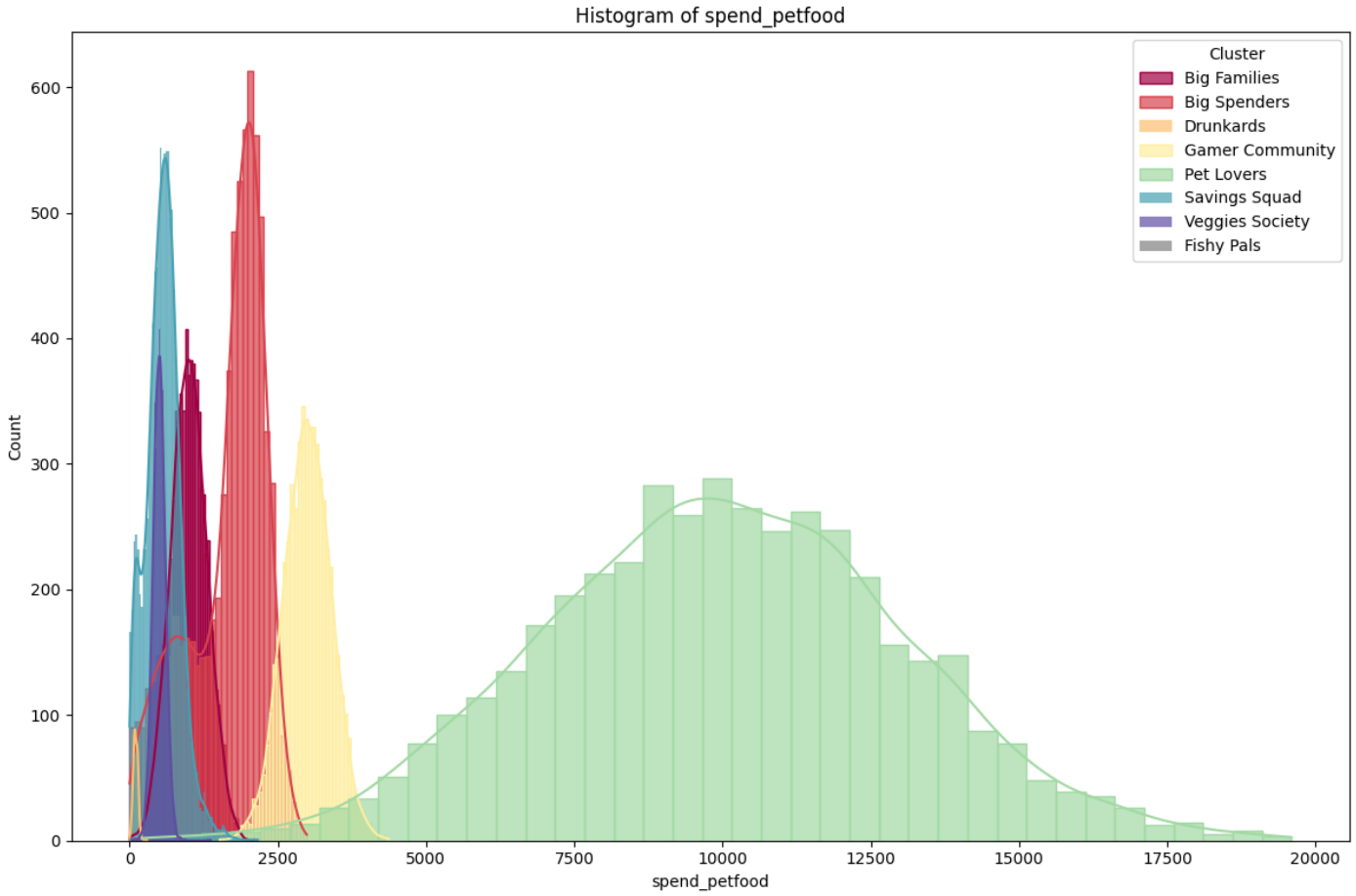
[76] HISTOGRAM OF THE MONEY SPENT ON HYGIENE PRODUCTS FOR THE DIFFERENT CLUSTERS



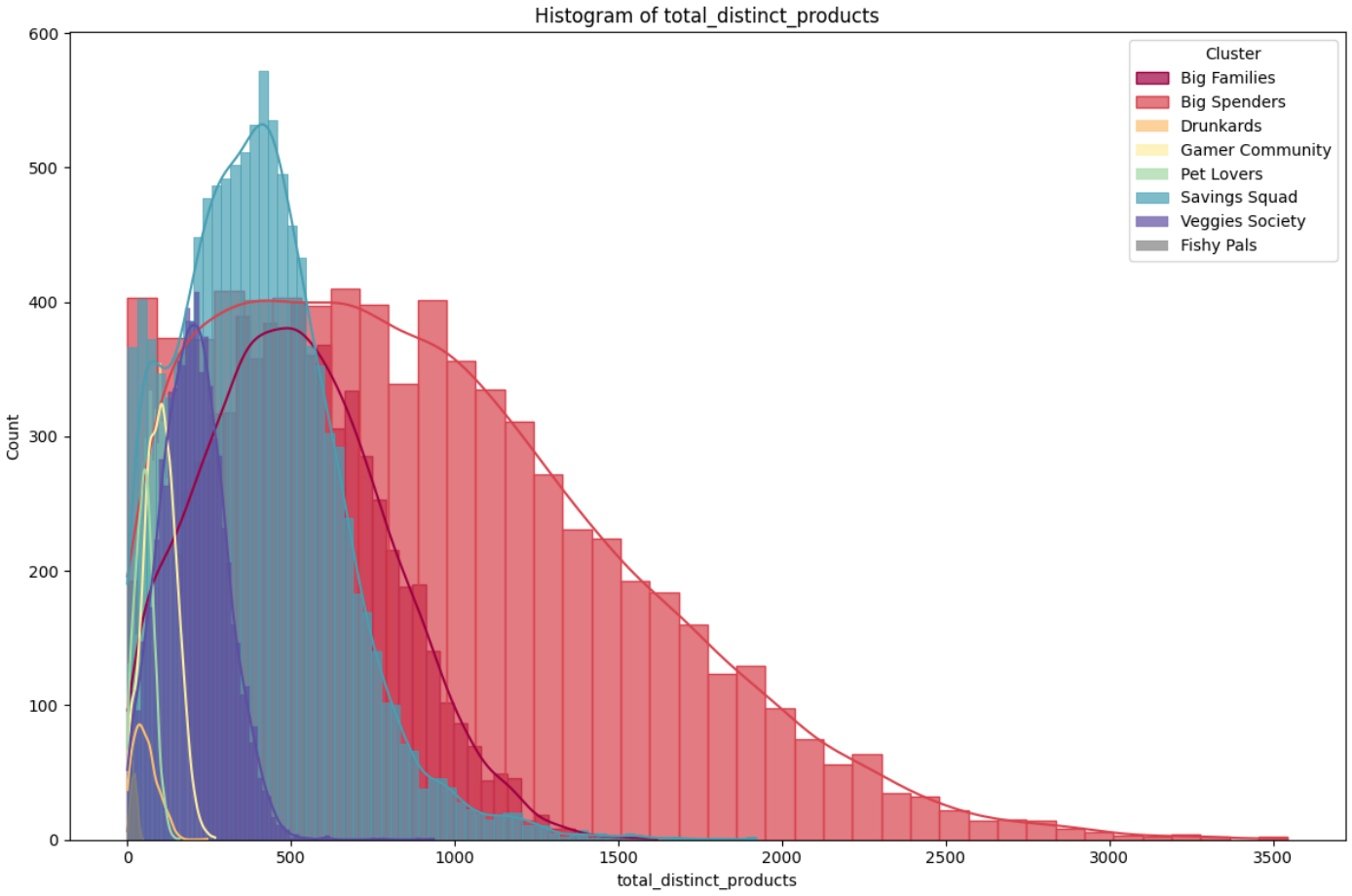
[77] HISTOGRAM OF THE MONEY SPENT ON VIDEOGAMES FOR THE DIFFERENT CLUSTERS



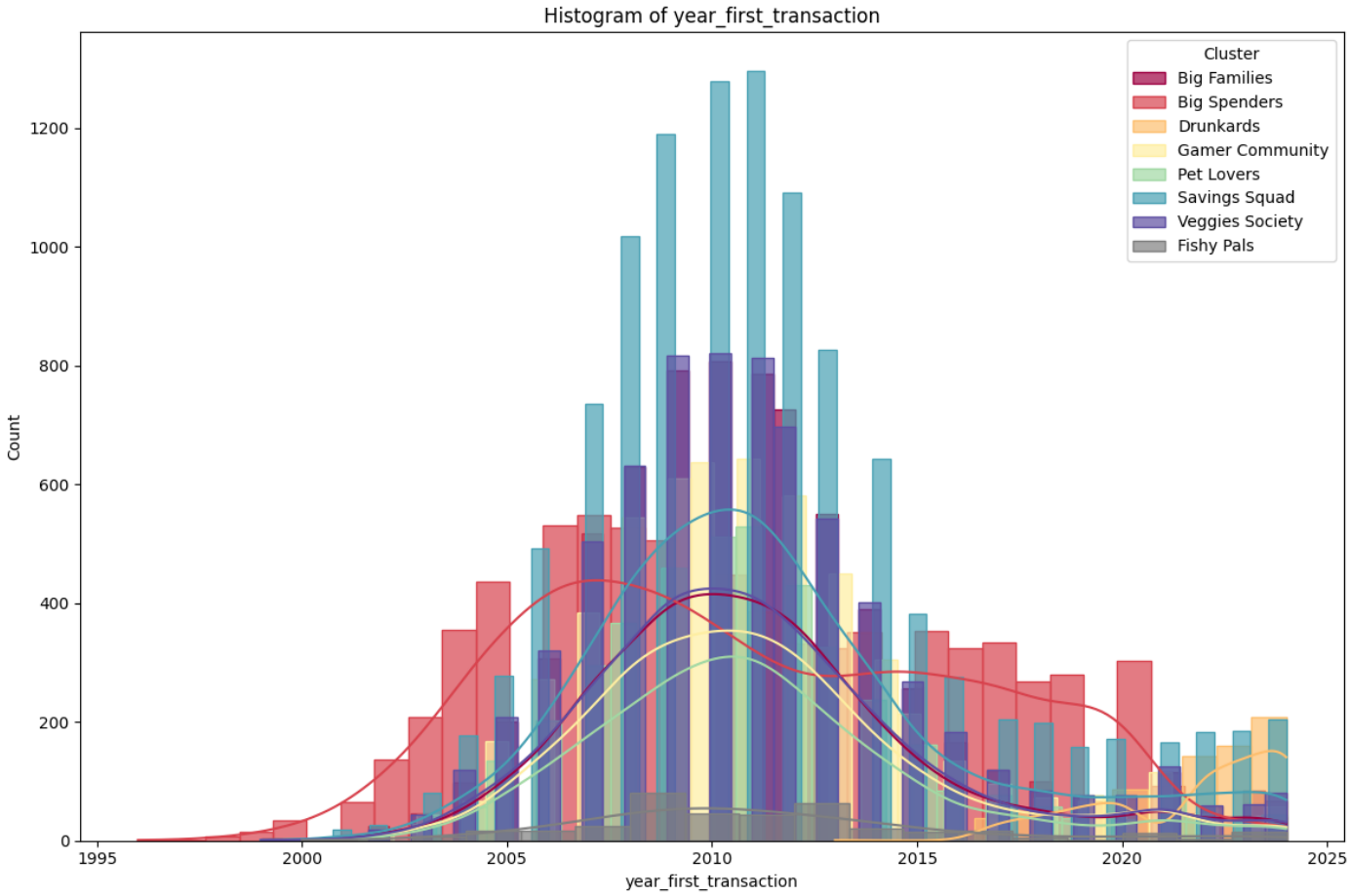
[78] HISTOGRAM OF THE MONEY SPENT ON PET FOOD FOR THE DIFFERENT CLUSTERS



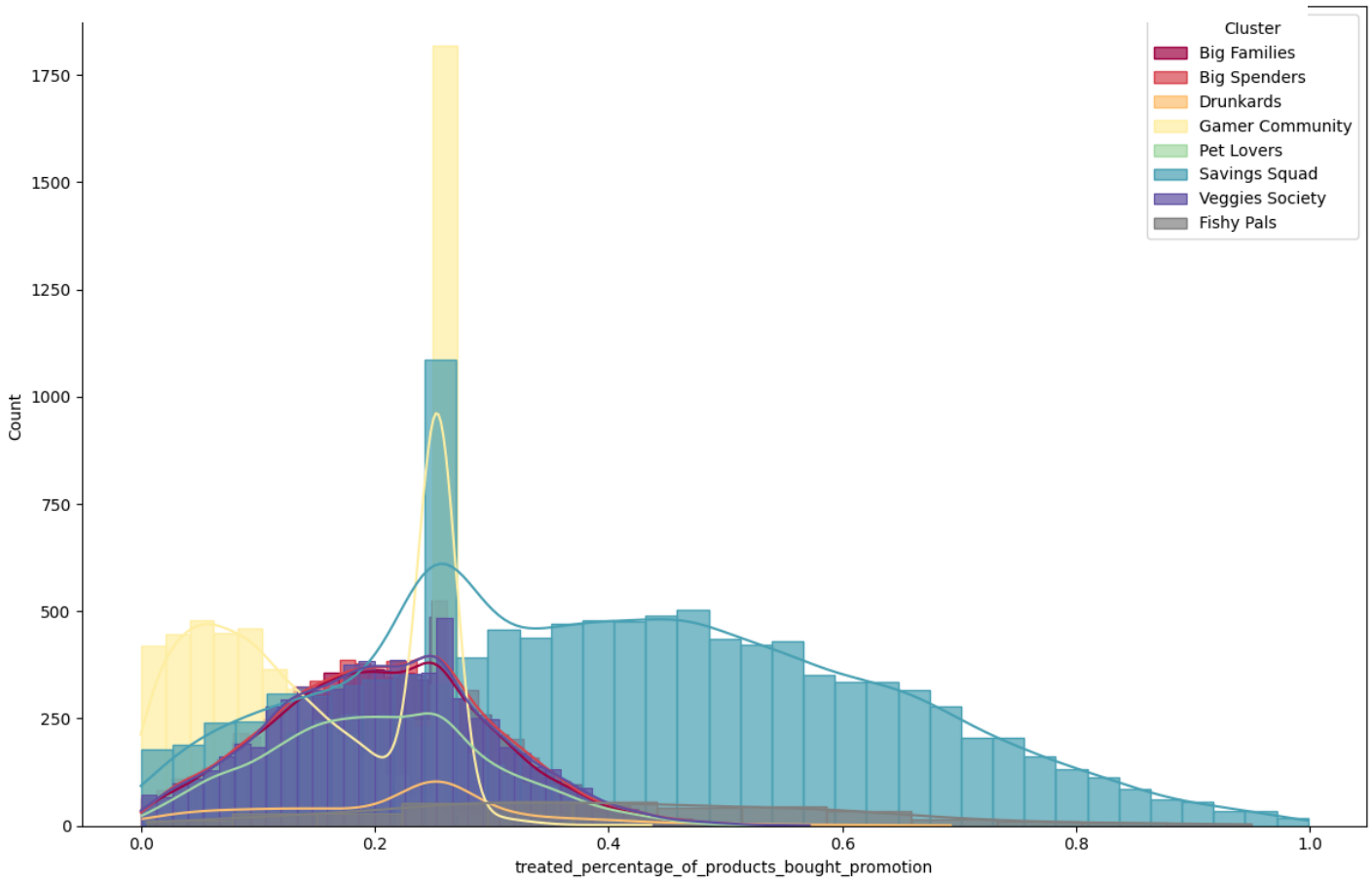
[79] HISTOGRAM OF THE TOTAL DISTINCT PRODUCTS FOR THE DIFFERENT CLUSTERS



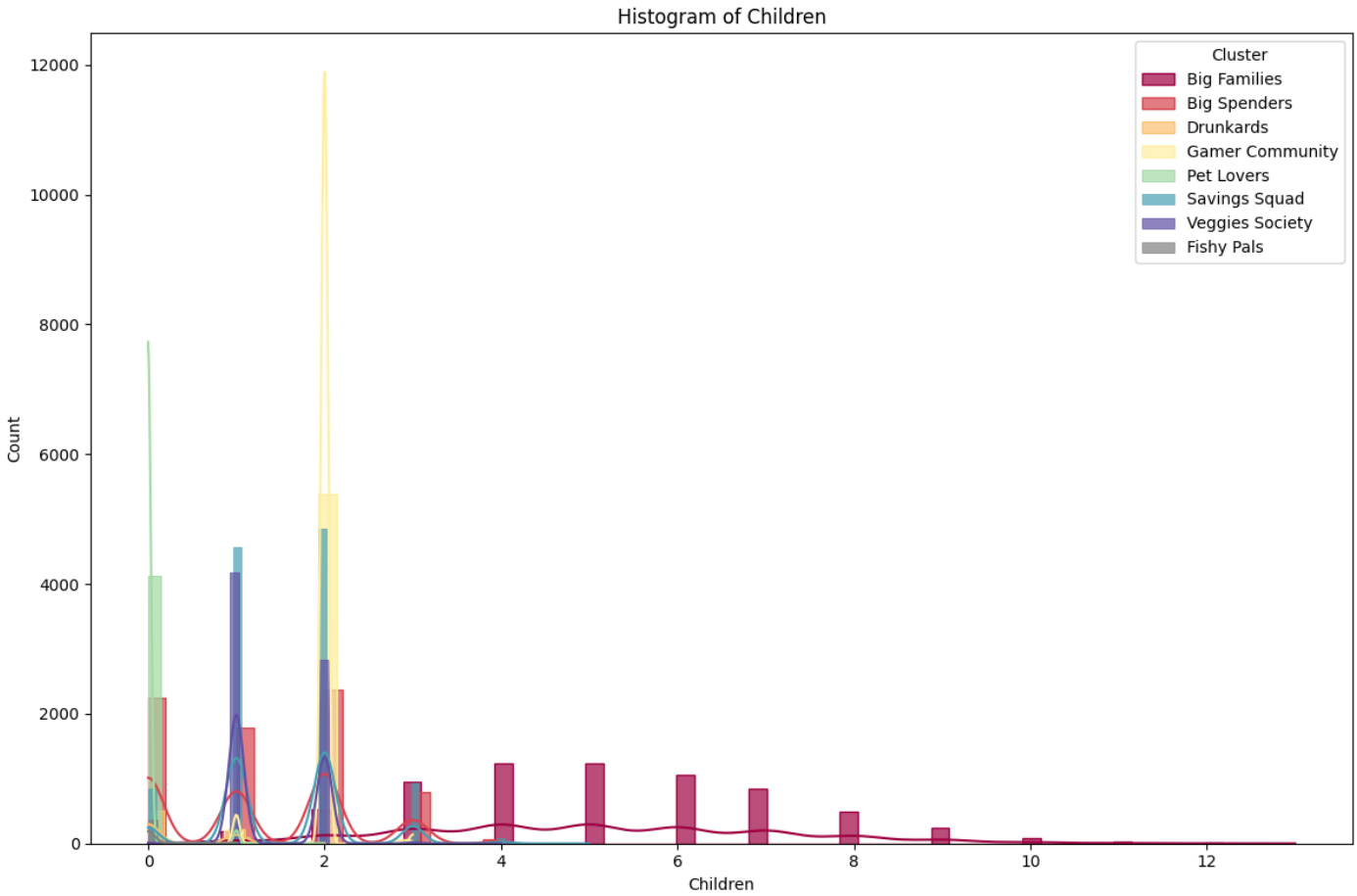
[80] HISTOGRAM OF THE YEAR OF THE FIRST TRANSACTION FOR THE DIFFERENT CLUSTERS



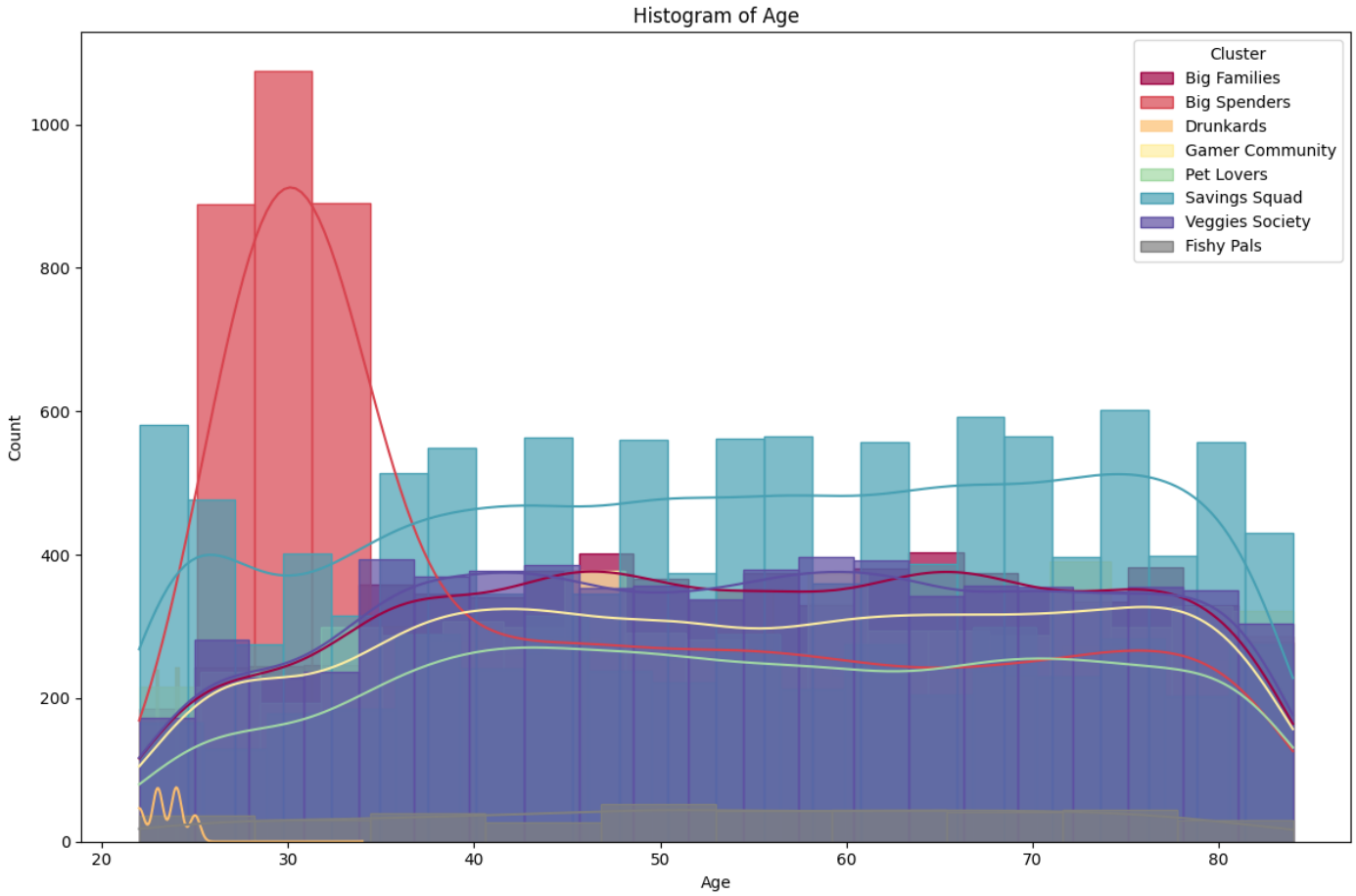
[81] HISTOGRAM OF THE PRODUCTS BOUGHT ON PROMOTION FOR THE DIFFERENT CLUSTERS



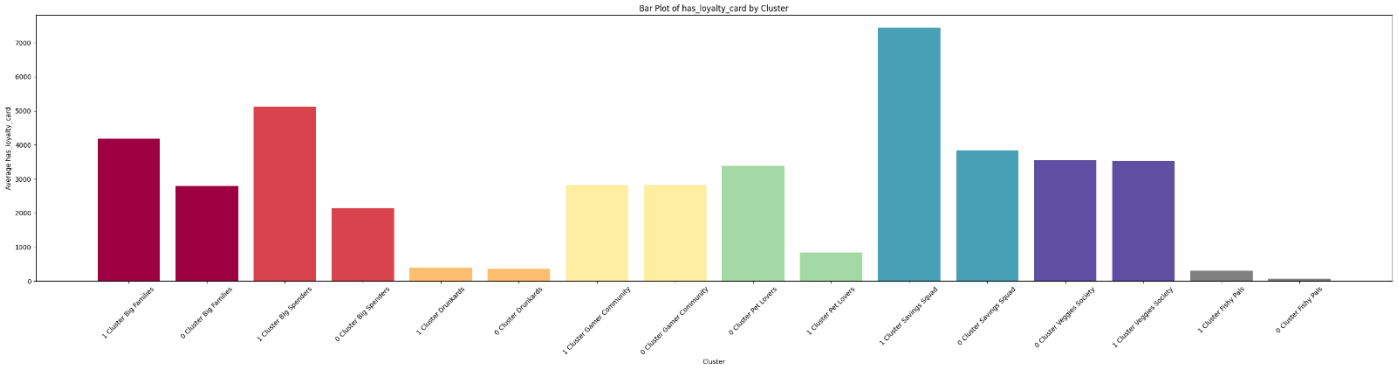
[82] HISTOGRAM OF THE CHILDREN IN HOUSEHOLDS FOR THE DIFFERENT CLUSTERS



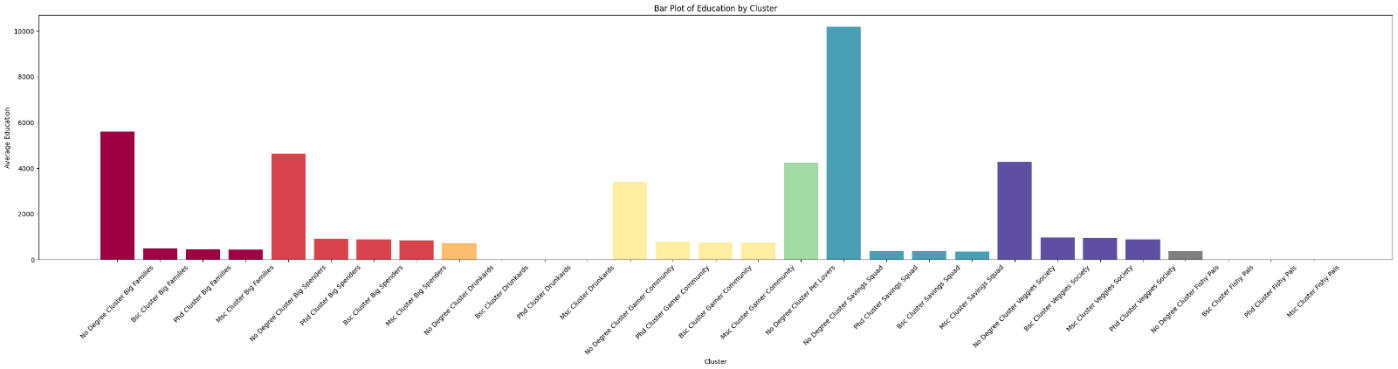
[83] HISTOGRAM OF AGE FOR THE DIFFERENT CLUSTERS



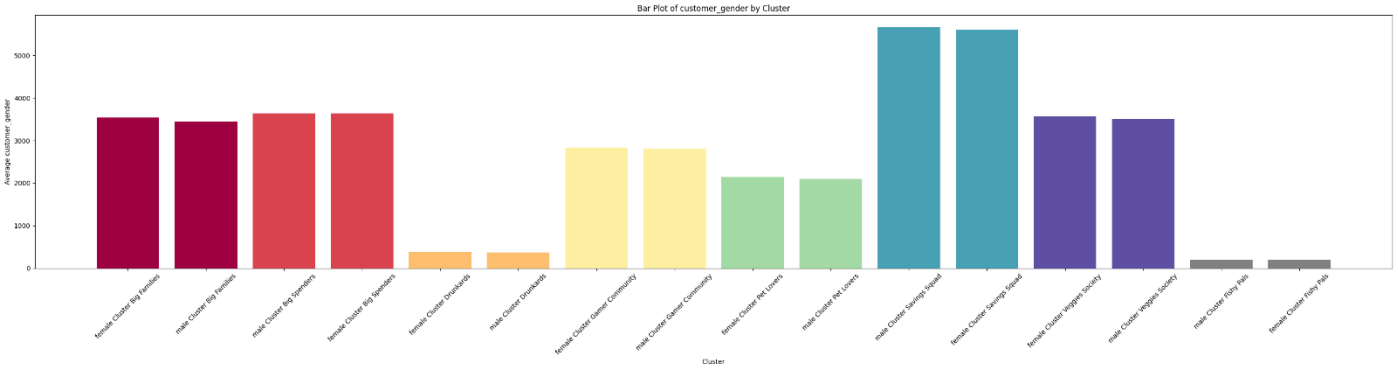
[84] LOYALTY CARD DISTRIBUTION FOR THE DIFFERENT CLUSTERS



[85] EDUCATION DISTRIBUTION FOR THE DIFFERENT CLUSTERS



[86] GENDER DISTRIBUTION FOR THE DIFFERENT CLUSTERS



[87] BOXPLOTS FOR EACH VARIABLE BY CLUSTER