NOVA IMS Information Management School

Data Mining Clustering Project Proposal

Master's in Data Science and Advanced Analytics

Done By **Group 13:**

Ana Pedro Martins Caleiro	**20240696**
Érica Yeranosyan Parracho	**20240583**
Oumaima Ben Hfaiedh	**20240699**
Rute D'Alva Teixeira	**20240667**

Fall/Spring Semester 2024-2025

TABLE OF CONTENTS

1. INTRODUCTION

Following a thorough preliminary analysis of the features and data distribution, we, as ABCDEats data analysts, are focusing on customer segmentation in this project, using customer data collected over three months. The goal is to develop marketing strategies that improve customer retention and foster stronger engagement. This involves categorizing the customer base into segments based on various characteristics and behaviors to uncover actionable insights.

In the earlier stages, we conducted data cleaning and clustering to prepare the dataset for effective segmentation. This included resolving issues like missing values, duplicates, outliers, and inconsistencies. After feature selection, the dataset was split into two groups: one centered on behaviors and the other on preferences. To perform clustering, we applied models such as K-means, Hierarchical Clustering, Self-Organizing Maps (SOMs), and DBSCAN, Mean-shift and Gaussian Mixture Model. Through merging clusters from both perspectives we were able to build a data-driven approach that enhances the company's understanding of its customers, enabling more personalized services and better-targeted marketing strategies.

2. PREPROCESSING

2.1. Duplicates

We identified 120 rows with duplicate values and removed them from our dataset. Additionally, we set **customer_id** as the index, as it uniquely represents each customer.

2.2. Handling Incoherences

As mentioned in our first handout, we encountered several anomalies in the dataset. One of the significant issues was the presence of 138 rows where ***vendor_count* = 0,** indicating that these customers had not placed any orders with any vendors in region 8670. We considered these instances as order cancellations and, given the small number of rows, decided to remove them from the dataset. Similarly, we applied the same approach to handle 156 rows where ***product_count* = 0.** Given the nature of the attribute ***is_chain*,** which indicates whether the customer purchased from a food chain, we transformed this variable into a binary one, instead of discarding it. We assumed that entries exceeding 1 represent the number of chain restaurants a customer bought from. Therefore, we replaced any value greater than 1 with 1. Although we will not use this variable for clustering, it could still be valuable for profiling.

2.3. Missing Values

For handling missing values, we chose to apply different strategies based on each scenario. We notice that when the ***last_order*** is 0, the ***first_order*** is Nan. For the ***first_order*** column, we decided to fill its missing values with 0 if the corresponding value in ***last_order*** is 0. The rationale behind it is that if ***last_order*** is 0, it is reasonable to assume that ***first_order*** is also 0. However, if ***last_order*** is

not 0, we leave the missing value in ***first_order*** unchanged. For the 1165 missing values in ***HR_0*****,** we filled them by calculating the difference between the total orders per day and the total orders per hour, as the number of orders per day should reflect the hourly data. Finally, for the missing values in ***customer_age***, we filled them using KNN imputation, assigning the values based on the closest neighbours. However, this step was performed only after normalizing the data, since KNN imputation relies on distance metrics - in our case, the Euclidean distance.

2.4. Outliers

For outliers treatment, it's important to distinguish two types of outliers: the ones displayed by univariate features, detected through boxplots, and the outliers introduced by multivariate relationships, from the interaction of multiple features while combining perspectives. For the univariate treatment, we conducted a quantile-based removal, defining as threshold 0.0001% of extreme values on the right end, since our data is heavily left skewed. We aimed to retain as much data as possible, preserving 99.81% of the dataset. This approach is performed after missing values imputation. For multivariate treatment, we made use of DBSCAN's outlier detection and removal properties. This clustering-based approach was performed after combining perspectives and scaling the data, ensuring all features had equal importance, for optimal performance. This operation filtered 0.43% of the existing data at that stage.

2.5. Feature Engineering

We generated 15 new features or variations of existing ones. First, for time and frequency-based features we aimed to reduce dimensionality while still capturing meaningful patterns in customers' purchasing behavior. This resulted in two sets: ***Weekdays Transactions*** and ***Weekend Transactions***, capturing frequency, discretized intuitively; and ***Morning Transactions***, ***Afternoon Transactions***, ***Evening Transactions***, and ***Night Transactions***, a set of features with equal width discretization, capturing purchases' timeline throughout the day. Combining these features into bins provided a better distribution than when they were used individually (see figures 2, 3, 4, 5, 6, 7). Second, we introduced features that may be useful in analyzing clusters which are:

- ***Recency****:* Calculated by subtracting the last_order value from 90 which represents the last day of the study period.
- ***Engagement Span****:* Calculated by computing the difference between last_order and first_order.
- ***Monetary Spending***: Total monetary value for each customer.
- ***Frequency***: Total number of orders for each customer.
- ***Buyer type***: Categorical feature indicating whether a customer is a one-time or repeated buyer. It was determined by computing the transactions where the buyer only issued his order once in all hours. (see figure 1)
- ***Spending Budget*****:** Categorical feature that profiles our customer based on how much they are willing to spend on the business. We conducted a quartile analysis to identify highest and lowest spending customers, defining for each quartile a threshold criteria based to

determine whether a customer spends up to the average amount for their respective category within the sector. (see figure 13)

- ***Customer Segment:*** Created to profile whether transactions were made for an individual purchase or a group/family. This classification defines as criteria whether a customer buys 3 or less products per transaction and spends in this purchase the same or less than what 75% of our customers would spend for a purchase of such volume. If these criteria are simultaneously met, the transaction is classified as individual, otherwise, it is a group/family transaction.(see figure 14)
- ***Cities*****:** Based on our EDA, we observed that regions sharing similar prefixes tend to exhibit the same preferences. As a result, we grouped regions into three cities: City 2 for regions with the 2xxx prefix, City 4 for regions with the 4xxx prefix, and City 8 for regions with the 8xxx prefix. For the unidentified region, values were allocated to region 8550, as they displayed similar patterns. (see figure 12)
- ***Aggregated Cuisines***: We created the cuisine category ***CUI_Asian_Fusion***, which included all Asian-related cuisines except Indian food. Beverages, desserts, and cafe items were grouped under ***CUI_Sweets_and_Beverages***. Finally, we combined ***CUI_Street_Food_Snacks*** with ***CUI_Chicken_Dishes*** into a single category.

Given that these new features introduced new outliers, we treated them using the multivariate outlier removal DBSCAN. (see figures 8 to 12, 15)

2.6. Data Normalization/Scaling

Our dataset displays values represented in different units and ranges, making data scaling crucial before clustering. To address this, we used the MinMaxScaler from library scikit-learn. However, several features, particularly those related to cuisines and time periods, exhibited a highly left-skewed distribution. For this reason, we applied a logarithmic transformation to help data become more symmetrical and closer to a normal distribution prior to scaling, ensuring a consistent range of values. Although alternative approaches like StandardScaler were tested, MinMaxScaler produced superior clustering results, leading us to adopt it as our primary scaling method.

2.7. Categorical Features Encoding

To incorporate categorical features into our profiling, we opted to encode them for a more effective analysis. After evaluating various encoding methods, we selected one-hot encoding due to the low cardinality of our features. This approach minimizes the risk of creating an excessive number of columns and ensures a clear representation of each category. Additionally, one-hot encoding provides better compatibility with profiling by maintaining the interpretability of categorical data.

3. Feature Selection

For feature selection, we applied two filter methods: Spearman's correlation, suited for non-Gaussian data, and a variance test to assess data point deviation from the mean. The variance test showed no features to discard. In Spearman’s correlation, we focused on identifying highly

correlated pairs (|0.8| threshold), as redundancy may occur when two features provide similar information to the model. We also considered how many other features each was significantly correlated with, below the redundancy threshold. For relevance, we excluded non-essential features, ensuring the selected ones contributed meaningful value to the project. In our case, ***product_count*** was highly correlated with ***weekdays transactions***, ***frequency***, and ***vender_count***. , so we chose to remove it. Additionally, we removed **frequency** because it was highly correlated with both ***vender_count*** and ***engagement span***. Through the heatmap, we also identified ***customer_age*** with zero correlation to the other features, indicating its irrelevancy. As a result, we decided to remove it.

4. Contingency Table

To prepare our data for this analysis, we begin by binning each feature into equal-depth categories based on quartile values. From the resulting tables, we observe that:

Monetary Spending **vs** ***Frequency*****:** From this perspective, distribution of frequency shows low variance between quartiles, but this is not the case for monetary spending, with 44% of our customers' highest expenditure deriving from one single transaction. Of our highest spending customers, 15% of purchases total at least 45 monetary units, spread over 5 or more transactions, our second most expressive quartile.Of the customers spending up to 24 monetary units - our business' median for transaction, more customers spend this much in one single transaction than on 2,3, 4 or over 5, combined.

Marketing Strategy: We observe potential for growth, as the highest spending occurs during the first transaction for a segment of 6,239 customers. Retention strategies should focus on this group, as they are the least recent customers and spend the lowest amounts. Strategies like loyalty programs, follow-up discounts, incentives, and personalized email marketing could boost spending and engagement. This segment is primarily from City 2, purchases meals from morning to evening, favors chain restaurants, and 62% use discounts, validating their responsiveness to such strategies. In contrast, the top 25% of customers, representing over 1,200 individuals, are the highest spenders and most frequent buyers. This group primarily resides in City 8, with 71% preferring chain restaurants and often purchasing as groups or families. Notably, 57% of this segment buys without a promotion, suggesting opportunities for further engagement without heavy discounts.

Monetary Spending **vs** ***Recency*****:** From this perspective, data is more evenly distributed across quartiles. However, we can point at almost half of our highest spending sector purchasing very recently, in 7 or less days. However, in the customers spending 44 or less monetary units, there are more customers who haven't engaged in over 41 days than within any shorter span.

Marketing strategy: This analysis shows potential in our low recency, high spending segment, representing 4783 of our customers base, validating previous finds on frequency perspectives. Similar marketing perspectives can take place, focusing more on activation campaigns that reignite interest in our business. Most of these customers buy from chain restaurants, half responds to promotions, 44% from city 2, 33% from city 4 and the residual from city 8.

5. RFM Analysis

After identifying in our customer's database information on recency, frequency and monetary value for all transactions, it was intuitive to conduct an RFM analysis and understand how this segmentation model can provide further insights about our clients' behaviour. Because the features were already binned in a meaningful way from the previous analysis, the next step was simply to rename such categories more comprehensively, using the initials of each feature 'R,F,M' applicably, and rank each feature from 1 to 4, where 1 is the least recent, frequent or monetary value and 4 the most. Lastly, we performed an aggregation, segmenting customers by their RFM score, each a possible combination of rankings for ***recency***, ***frequency*** and ***monetary value,*** totaling 4x4x4=64 scores.

R4F4M4: The most dense segment of our business comprises customers who order frequently, engage recently, and spend high amounts, making up nearly 3000 of our core clientele. This entire segment of repeated buyers can be classified as 'Champions', purchasing as a group/family, mostly buying without any promotion and living in city 2.

R1F1M1: Following**,** with very similar values, however, are the customers who spend the lowest amounts, order sporadically and who haven't engaged in a while. This segment can be called 'Hibernating', buying for individual consumption and 72% can be classified as a one time buyer. Majority is buying without any promo and lives in city 4.

The gap between both segments likely confirms the business' need for more tailored marketing strategies that retain the most valuable clients while stimulating re-engagement. In parallel, such values could hint at the lifecycle of our customers, with new customers spending a lot but after a certain period of time disengaging from our business.

6. Clustering

6.1 Segmentation

Our project segmentation was based on two distinct approaches. The first focused on customer behavior, using features such as ***monetary spending***, ***recency***, ***weekends transactions***, ***weekdays transactions***, and the remaining time-related features to capture patterns in how customers use the ABCDEats application. The second approach is centered on ***cuisine*** types in the dataset, providing insights into customer preferences.

6.2 Behavior Features Clustering

6.2.1 Hierarchical and K-Means clustering

In our analysis, we compared *K-means* clustering with hierarchical linkage methods (single, complete, average, and ward) by visualizing the R^2 scores for each cluster. We selected the clustering method that provided the highest R^2, which was Kmeans for this perspective (see Figure 18). We can observe

in this figure that the elbow point is clearly when k=3. We further confirmed this through plotting inertia and silhouette (see figure 19), thus validating our observation. Therefore, we carried our analysis with the choice of *K-means* with 3 clusters as a more suitable fit than hierarchical clustering.

6.2.2 SOM clustering

Self-Organizing Maps (SOM) are neural networks that map high-dimensional data onto a lower-dimensional grid for visualization and clustering. With this clustering method, we examined the component planes of the features related to the behavior of customers, that is, a visual representation of the values of each feature across the map's grid of neurons (see figure 22). For a better understanding of the visualizations, we generated a Hits Map, providing information of how frequently different units of the map are activated or hit by data points during the clustering process (see figure 23), and a U-Matrix which displays a spatial representation of how clusters relate to each other in terms of proximity and helps identify patterns in the data.(see figure 24). After, we applied K-Means (see figure 25A) and Hierarchical Clustering (see figure 25B) on top of SOM, for a defined number of 3 clusters. Each hexagon represents a cluster, with the colors of the hexagons representing the cluster labels following the same behavior. We concluded that SOMs on top of hierarchical clustering would be a better solution.

6.2.3 Density-Based Clustering

6.2.3.1 DBSCAN

This clustering solution is a great candidate for clustering high-dimensional data with nested clusters that are hard to identify on a first sight. Besides, it doesn't require defining the number of clusters *a priori.* This model is initialized after defining a radius perimeter for how close the data points should be to each other (*epsilon)* and a minimum number of samples (*min_samples*) nearest to each point within that perimeter- these are the main parameters of the model. Such parameters are then used as criteria to classify each datapoint as a core point or a non-core point. The model randomly selects a core point, expanding that selection up to all points that meet the criteria, forming a cluster. Lastly, the nearest non core points can join that cluster. Following, the model picks a new core point, repeating the process all over again.

To define *epsilon,* we plot a *k-distance* graph that describes the distance of each datapoint to its nearest neighbor. *Min_samples* is defined from the rule of thumb. Then, we use the elbow method to choose the optimal number of *eps*. Over trial and error, we obtain an *eps*=0.2175 and a *min_samples*=17. This results in 4 clusters, excluding outliers. However, the smallest cluster was assigned to as little as 9 clients, making up too niche segmentations.

6.2.3.2. Mean-Shift Clustering

We tested mean-shift clustering, a non-parametric method where data points are iteratively shifted towards the nearest mean, gravitating to the most dense region defined as the attraction basin. This shift continues until data points converge into a stable location around the mode, forming a cluster.

The parameter for this model is the *bandwidth*, which determines the reach towards a dense region around each data point. In this case, *bandwidth* was estimated with *scikit-learn's* ingrained function, *estimate_bandwith*. To optimize the clustering solutions, we tested with different quantiles until we obtained a small enough value that would provide enough clusters. In our attempts, this clustering method resulted in low quality solutions, with only 2 clusters assigned, despite the time-costly parameter tuning performed. The optimal parameters were set at *quantile*=0.2, resulting in a *bandwidth=0.54.*

6.2.3.3. Gaussian Mixture Model

The number of clusters to fit the data, is controlled with the *n_components* parameter. To define it, we introduce Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC), plotting both to understand how much information is lost by this model, for different values on clusters. (see figure 26). Then, using the elbow method, we define the optimal *n_components=4.* Other important parameters are how the model is initialized and what type of covariance matrix is being considered. In our case study, we use class material's settings, defining covariance_type=full and init_parameters=kmeans. From fitting this model, we obtain an estimate for the **weights,** proportion of data in each cluster, **covariance,** describing the modelling of each cluster, and **means,** the center of each cluster.

6.2.4 Model Assessment

To select the best model, we performed a comparison analysis, validating the highest score. Our evaluation metric was R-squared. We also considered the cluster's size, but prioritized the R-Squared results. While GMM had the highest R-squared score, it shows unbalanced clusters after merging labels, including zero instances. Such poor clustering solution led us to choose SOMs, the second-highest score with better quality clusters (see Figure 27).

6.3 Preference Features Clustering

6.3.1 Hierarchical and K Means clustering

Following the same evaluation criteria, K-means was the selected model from a preference perspective (see Figure 20). However, in this figure the elbow was not very clear so we relied on the results of inertia and silhouette plots (see figure 21) suggesting 6 as the optimal number of clusters.

6.3.2 SOM clustering

Similarly to the Behavior perspective, we analyzed the Preference Features by plotting their component planes (see figure 28). Next, we examined the Hits Map (see figure 29) to observe the data density and cluster membership across the SOM grid. Additionally, we analyzed the U-Matrix (see figure 30) to evaluate the distances between neighboring nodes. Finally, after applying K-Means (see figure 31A) and Hierarchical Clustering (see figure 31B) on top of SOMs, and upon a careful analysis we defined a number of 6 clusters, concluding that Hierarchical Clustering is the best option.

6.3.3 Density-Based Clustering

6.3.3.1 DBSCAN

In line with the workflow previously described, this model was tuned into the optimal combination of *eps*=0.19 and *min_samples*=30. This results in 9 clusters, excluding outliers, fit to describe nine different cuisine preferences. However, this solution resulted in micro segments, ranging from 23 to 165 customers, excluding the ruling class. For that reason, we consider this clustering of lower quality.

6.3.3.2 Mean-Shift Clustering

From this perspective, like on the previous one, results 2 cluster solutions, with a tuned *quantile*=0.4, obtaining a *bandwidth*=0.70. These cluster solutions can be rated lower quality, since overall there is no strong association in any particular label. However, most recent customers are likely to belong to cluster 2 instead of 1.

6.3.3.3 Gaussian Mixture Model

Following the process outlined on the preference perspective, we initialize this model with the optimal *n_components*=5 (Figure 32), while the remaining parameters are set based on the case study provided in the class materials. This approach is consistent across both perspectives.

6.3.4 Model Assessment and Final Cluster Discussion

Selecting the model for the preference perspective was more challenging than for the behavior perspective, as K-means and SOMs showed similar R-squared values and cluster counts. To decide between the two, we tested both models and analyzed their performance separately for preference clusters and in the final merged clusters. For that reason, in this section we will also discuss the final clustering solution, as it serves as an important factor for assessment. (Figure 33).

First, we began with preference-based clusters, noting a strong separation between certain cuisines in both solutions, supported by R-squared values. We tested K-means (3 and 5 clusters) and SOMs (4 clusters) for preferences, using SOMs consistently for behavior. Our evaluation prioritized separation, interpretability, and balance, avoiding dominance by any single cluster.

Second, we carried our analysis by merging both perspectives. The number of merged clusters varied based on adjustments to the Hierarchical Clustering Ward's Diagram. Using SOMs as the chosen method in preference perspective, we obtained 4 merged clusters with balanced sizes which provided clear separation in features like ***Engagement_Span***, ***Monetary_Spending***, and ***Cuisines,*** along with a versatile and interpretable analysis. The downside, however, lies in the increased number of clusters, which makes interpretation more complex and, from a business standpoint, requires more resources. On the other hand, with *K-means* as the chosen method in preference perspective we obtained 3 merged clusters and observed balanced feature representation, but

potential oversimplification of the insights. We tried increasing to 5 clusters but that only added granularity and disrupted balance highlighting smaller cluster dominance. We validated our analysis using the t-SNE visualization tool, which confirmed our initial findings. (Figure 34 to 36)

To conclude based on this analysis we chose to use 4 merged clusters with the SOMs model for both the behavior perspective and preference perspective. This approach provided the best trade-off across evaluation criteria, delivering detailed and actionable marketing insights. But if the company wants a simpler and more broad strategy, K-means would be also a good choice for preferences with 3 merged clusters. (see table 1)

7. Cluster Analysis and Marketing Approaches

Based on the previous decisions we ended up with 4 distinct groups of individuals with varying characteristics. (Figure 37 to 43)

Cluster 0, the **Balanced Explorers**, represents 26.6% of app clients. This group includes customers like Dora and Diego, who are highly active during weekdays, particularly in the afternoon (lunch time) and frequently during dinner hours. With a preference for ***CUI_OTHER*** and diverse cuisines, avoiding healthy options. Geographically, these customers are primarily based in City 2, followed by City 4, and they form the second-largest cluster. They also rank as the second-highest group for repeat buyers, indicating that Dora and Diego are likely residents in these areas. At this stage in their lives, they are likely starting their careers and rely on delivery due to busy schedules, maintaining a passion for food exploration. With steady spending and engagement span, high vendor diversity, and openness to trendy cuisines, they are a strong segment for loyalty and revenue growth. Possible marketing approaches could include offering time-based discounts specifically for lunch orders and implementing a loyalty-based subscription. After reaching a certain number of deliveries, customers could receive a mystery box featuring trendy food items of the moment, keeping them engaged.

Cluster 1, primarily based in City 4 and City 2, represents our most populated cluster (44.3%) and also composed of our most valuable and loyal customers, **Top-Tier Customers**. This group, including individuals like Michael Scott and Dwight, orders consistently throughout the day, with a slight preference for weekdays and early afternoons. They prefer Asian Fusion and American cuisines, often opting for delivery instead of bringing food to work. As the highest spenders and most active users, they also make extensive use of promo codes and are the top repeated buyers, indifferent to whether a business is part of a chain. To retain them, we could offer premium loyalty rewards, such as exclusive discounts and a points-based system where customers earn more for frequent orders. Additionally, creating package deals (e.g., a €10 menu with a drink) for lunchtime would be appealing, especially if we focus on their preferred cuisines (Asian Fusion and American).

Cluster 2, our smallest group (10.4% of clients), is primarily located in City 2 with a slightly distributed presence across the other cities. Named **Guilty Pleasure Customers**, they show low engagement span, moderate spending.. These customers, like Mónica and Rachel, prefer dining out or cooking at home but occasionally use delivery for specific cuisines like sweets, snacks, and beverages. They are infrequent app users, often one-time buyers, and mostly use delivery promo codes. Their activity

lacks a consistent time based pattern but slightly leans towards weekdays and afternoons, similar to other clusters . Our goal is to move these customers to the **Top-Tier** segment by sending notifications that highlight hassle-free ordering and encourage repeat purchases, such as: *"Busy day? Treat yourself with a delicious dessert!"* or *"We miss you! Here's a 15% discount to try our new cocktails"* We can also offer discounts tailored to their food preferences (Snacks and Sweets and Beverages).

Lastly, Cluster 3 (18.7%), which has the highest number of one-time buyers, exhibits behavior characteristics very similar to our **Guilty Pleasure Customers**. The main distinction lies in their food preferences, which is why we've named this cluster **Chill Asian Vibes**. This group includes customers like Uncle Roger or Ted Mosby. Although they don't use the app consistently or frequently, when they do, they already know what they want (low ***vendor_count***), and it's typically Asian food. This cluster, larger than its lookalike, is mainly located in City 4 and City 8. They equally use delivery and discounts, prefer cash payments, followed by digital, and rarely pay by card. They also avoid chain restaurants, favoring authentic Asian dining experiences. To build loyalty, we could strengthen partnerships with authentic Asian restaurants in City 4 and City 8, highlight cash payment options via notifications, and celebrate Asian events like Lunar New Year with special menus and discounts.

This analysis aligns with the RFM segmentation with **Balanced Explorers** and **Top-Tier Customers** in **R4F4M4** and the others in **R1F1M1**. We also identified similar behaviors and marketing strategies through the contingency table analysis. The main goal is to shift more clients to the higher-performing group.

7.1. Assessing Feature Importance and Reclassifying the outliers

Outliers excluded earlier were reassigned to existing clusters using the Decision Tree algorithm, as their minimal number makes them irrelevant to the analysis. Feature importance (we used R-Squared values and Decision Tree algorithm) confirmed previous findings, highlighting the significance of ***CUI_OTHER*** and ***Asian Fusion*** cuisines, along with ***afternoon*** and ***weekday transactions***. This alignment supports our marketing strategy development.

8. Conclusion

To improve the company's marketing strategy, initially we focused on Data Exploration and Preprocessing targeting all inconsistencies. Then we created important new features to combine and extract more information from existing ones. For clustering, we divided features into two segments: **behavior** and **preference**. We tested the clustering of these perspectives using various models, including k-means, hierarchical clustering, Gaussian mixture, SOMs, DBSCAN, and mean-shift, and selected the best model for each segment based on R^2 which were SOMs in both perspectives. Finally, we merged the segments using hierarchical clustering, identifying four distinct customer groups that we further analysed to capture meaningful marketing insights.

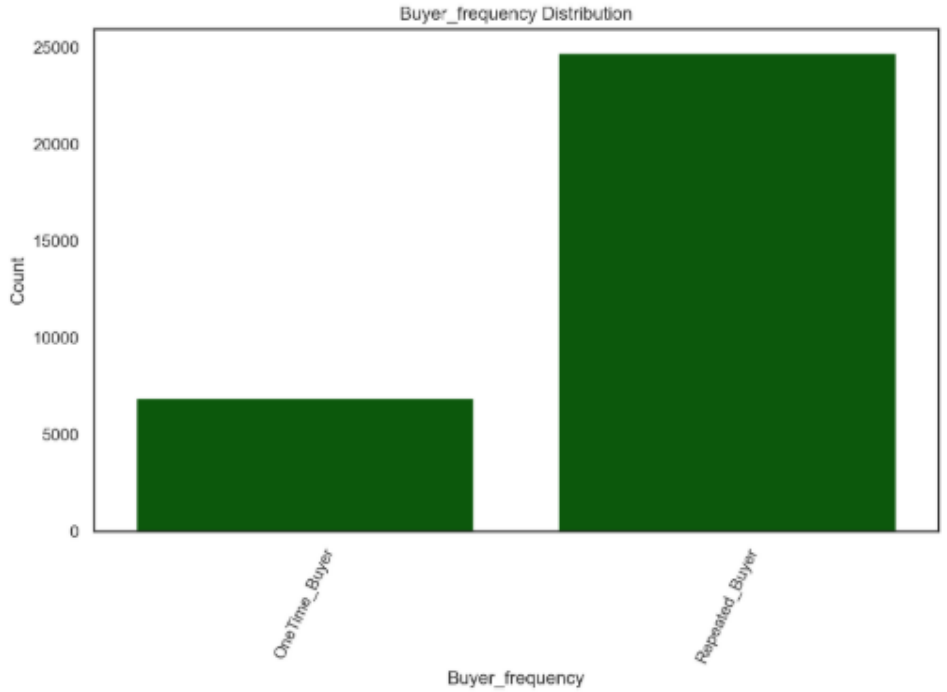
9. References

- Market Segmenation Book “Market Segmentation Analysis” by Sara Dolnicar Bettina Grün Friedrich Leisch: Link
- Mean-Shift Clustering: Link
- Gaussian Mixture-Model: Link
- DBSCAN: Link
- UMAP : Link
- RFM : Link

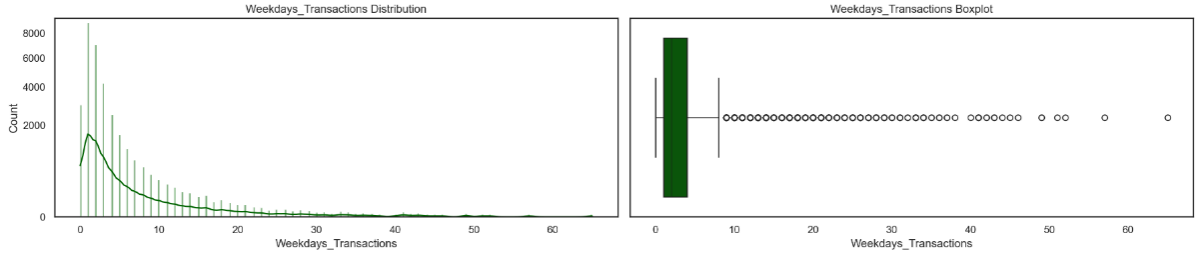
10. Appendix

10.1. Preprocessing Appendix

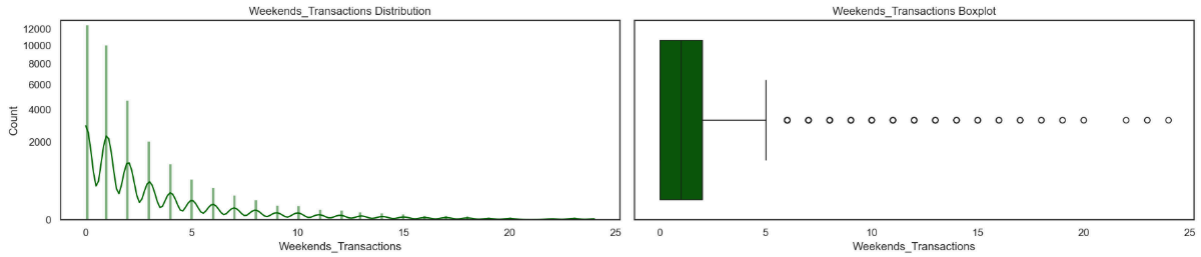
10.1.1 Feature Engineering : Buyer Type, Figure 1



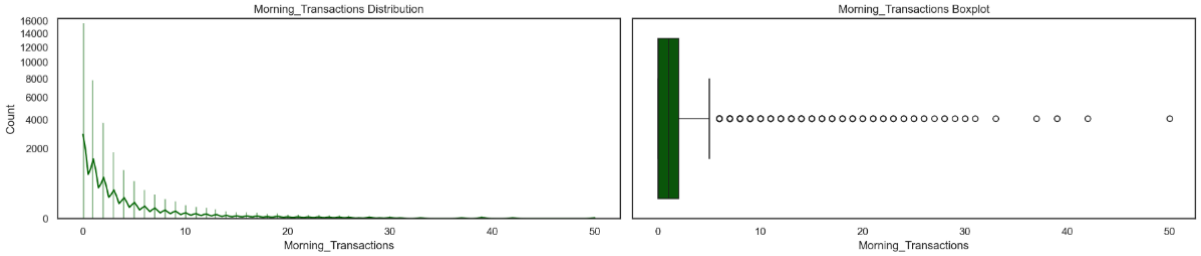
10.1.2 Feature Engineering : Weekdays Transactions, Figure 2



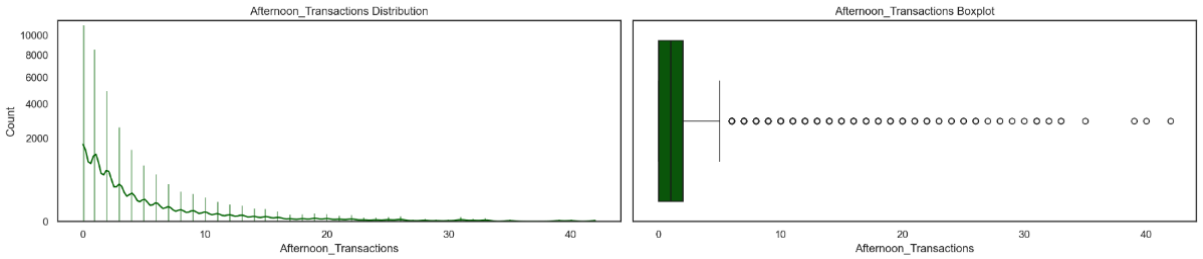
10.1.3 Feature Engineering : Weekends Transactions, Figure 3



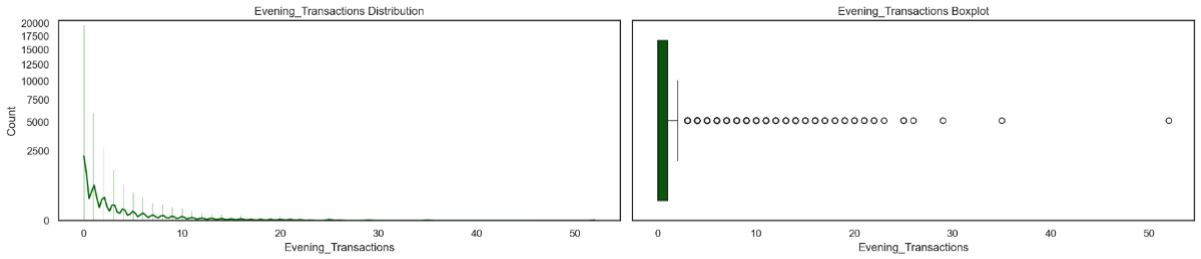
10.1.4 Feature Engineering : Morning Transactions, Figure 4



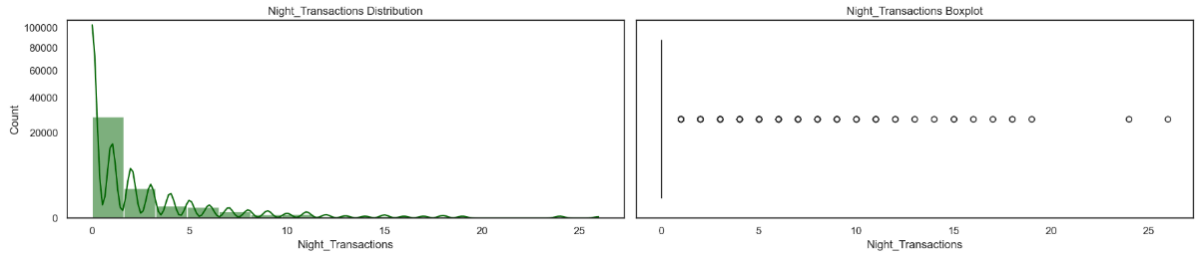
10.1.5 Feature Engineering : Afternoon Transactions, Figure 5



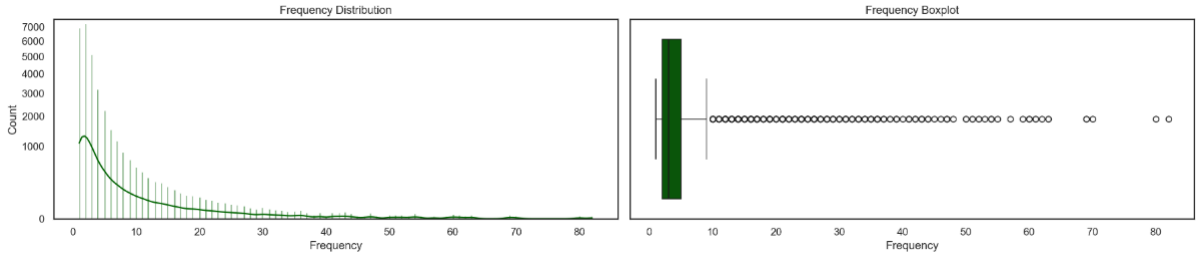
10.1.6 Feature Engineering : Evening Transactions, Figure 6



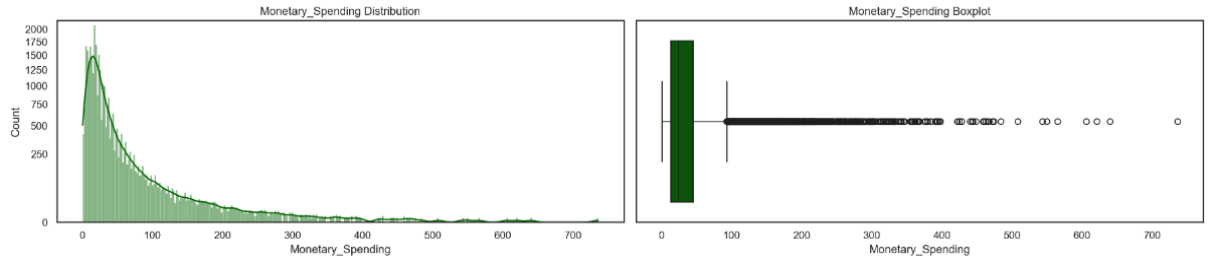
10.1.7 Feature Engineering : Night Transactions, Figure 7



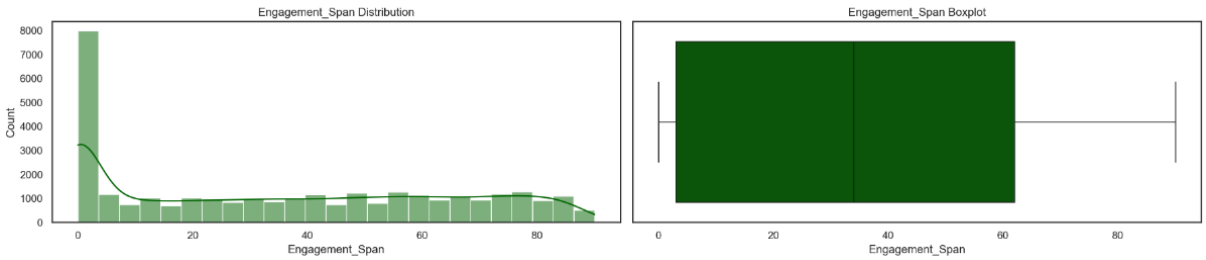
10.1.8 Feature Engineering : Frequency, Figure 8



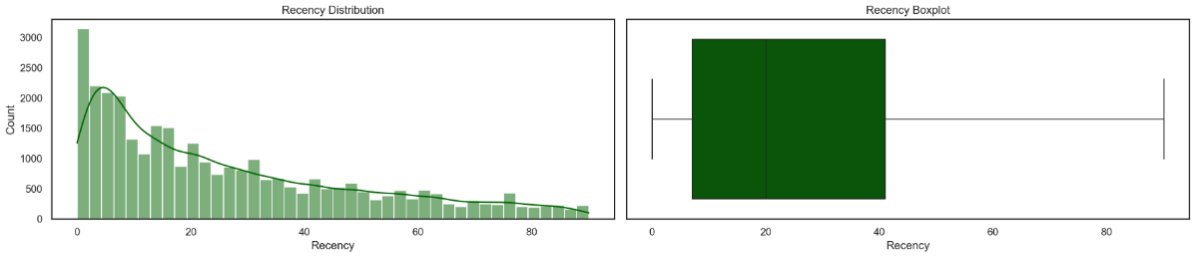
10.1.9 Feature Engineering : Monetary Spending, Figure 9



10.1.10 Feature Engineering : Engagement Span, Figure 10



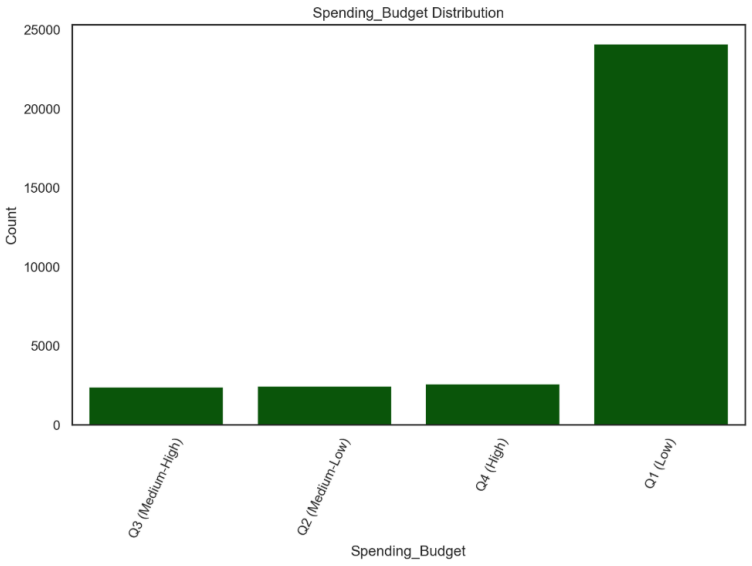
10.1.11 Feature Engineering : Recency, Figure 11



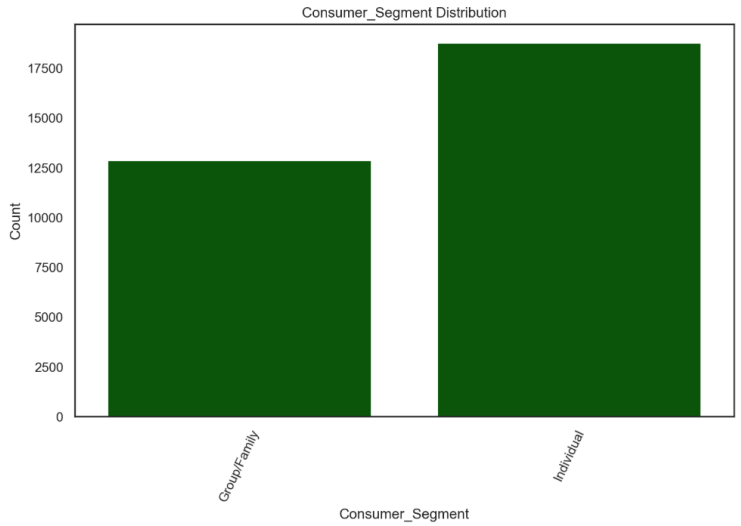
10.1.12 Feature Engineering : Cities, Figure 12

	cities			2		4			8
	customer_region	2360	2440	2490	4140	4660	8370	8550	8670
Monetary_Spending	min	0.370000	0.860000	0.500000	1.35000	1.120000	0.500000	5.750000	0.470000
	max	392.220000	297.150000	122.940000	242.04000	565.270000	283.590000	355.040000	736.150000
	mean	23.142913	21.778346	19.285192	28.89276	40.313405	43.336626	50.443626	52.171552

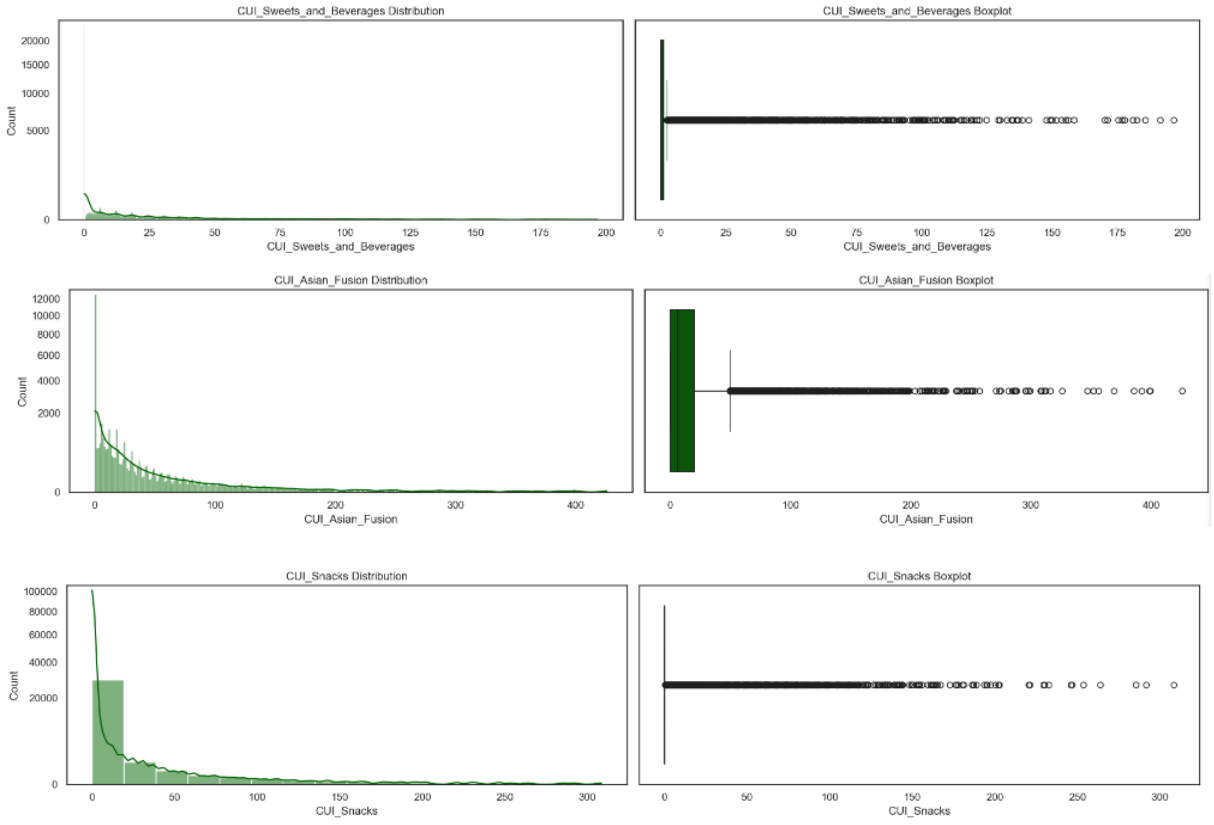
10.1.13 Feature Engineering : Spending Budget, Figure 13



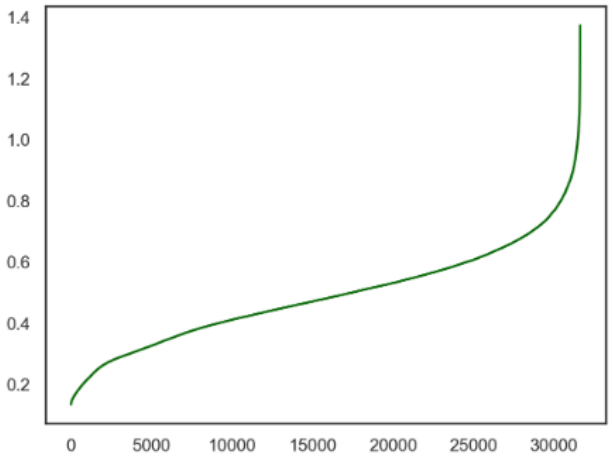
10.1.14 Feature Engineering : Consumer Segment: Individual vs Group transactions, Figure 14



10.1.15 Feature Engineering : Combining Cuisines: Asian_fusion, Sweets_and_Beverages and snack, Figure 15

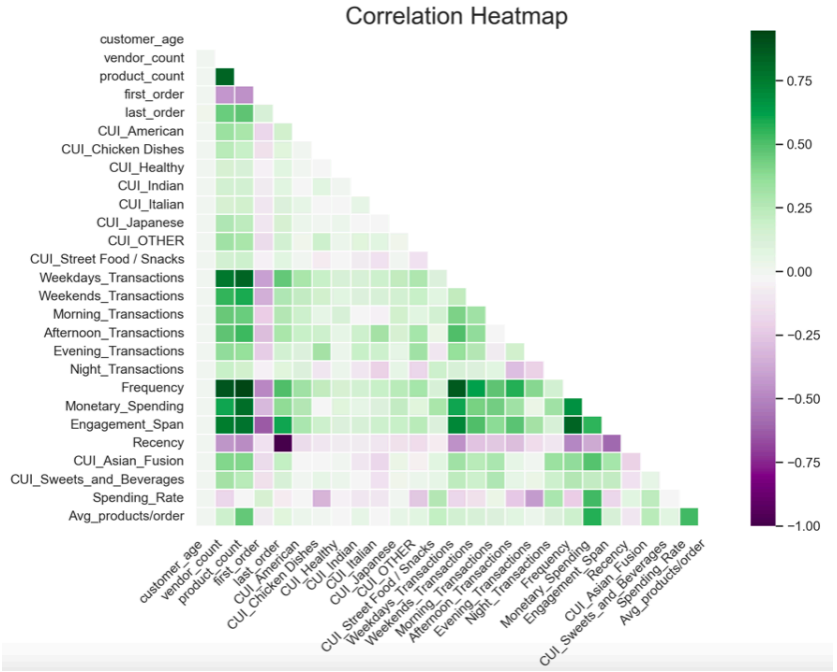


10.1.16 DBSCAN Multivariate Outliers: K-distance graph to find out the right epsilon value, Figure 16



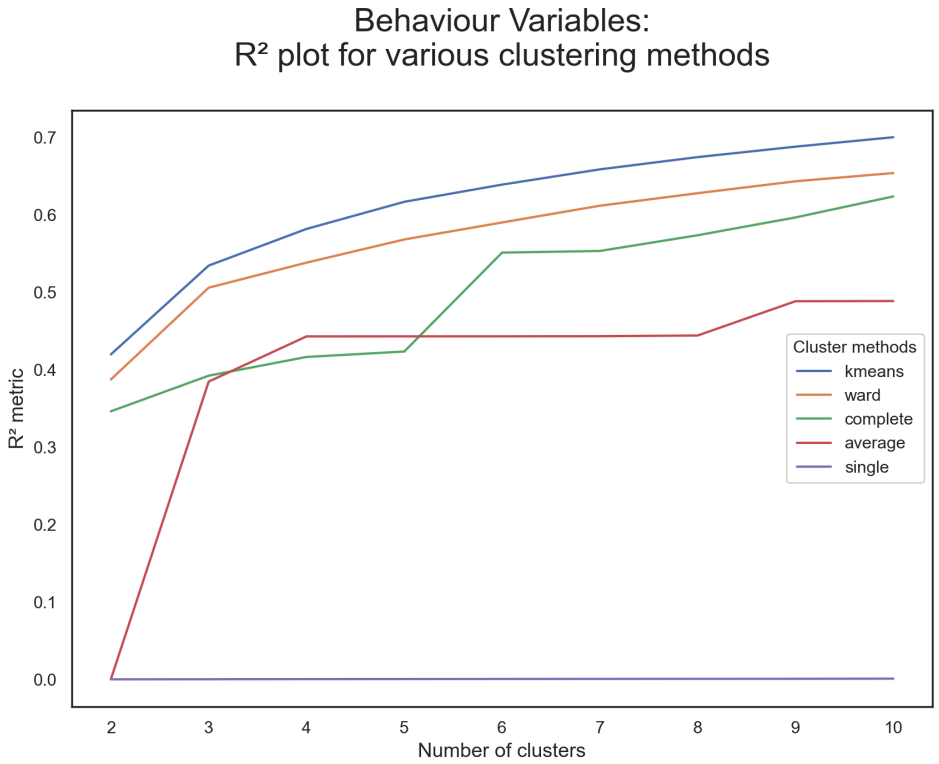
10.2 Feature Selection Appendix:

10.2.1. Correlation Matrix of New Numeric Features, Figure 17

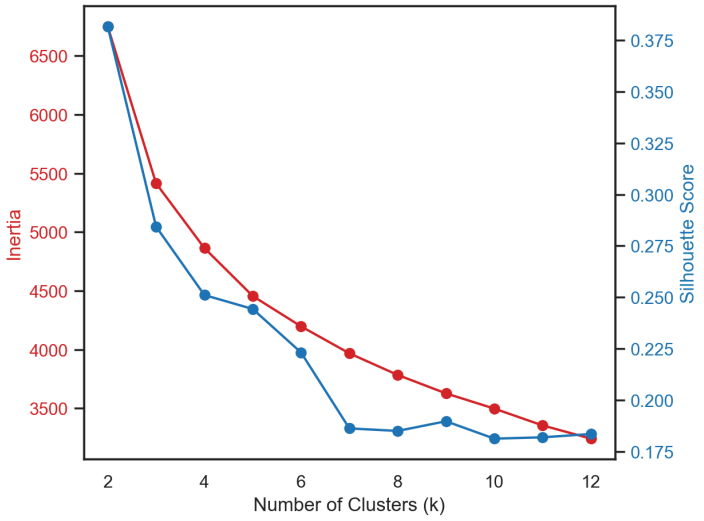


10.3 Clustering Appendix

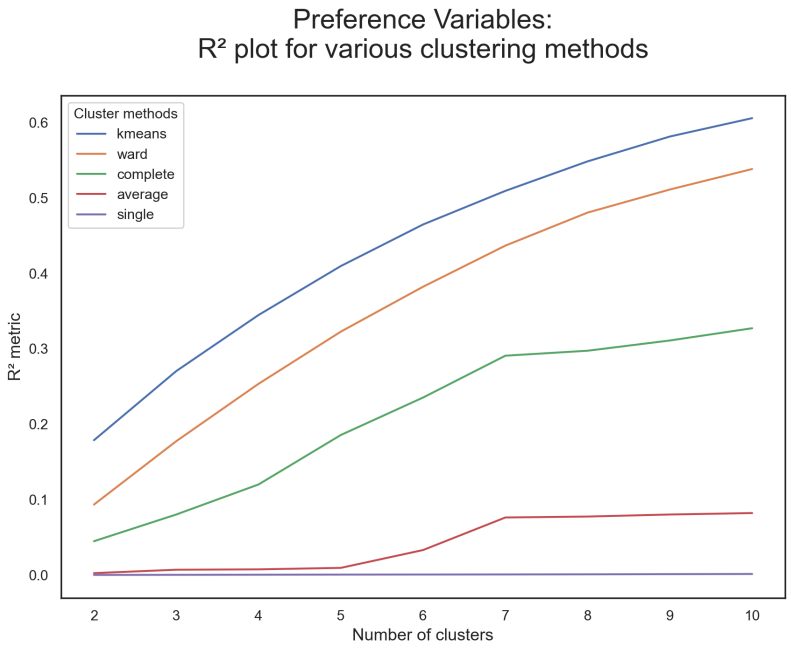
10.3.1 Hierarchical Clustering for Behavior Variables, Figure 18



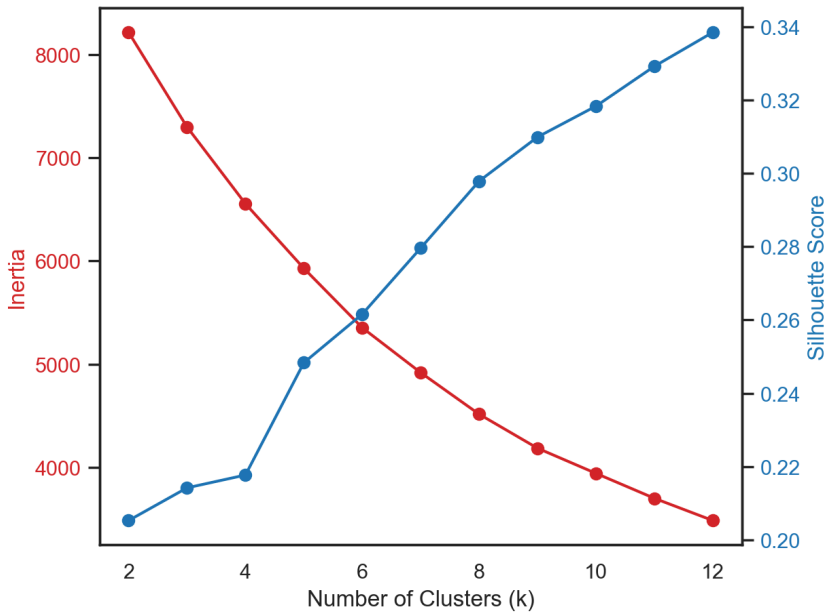
10.3.2 Hierarchical Clustering for Behavior Variables Inertia and Silhouette graphs, Figure 19



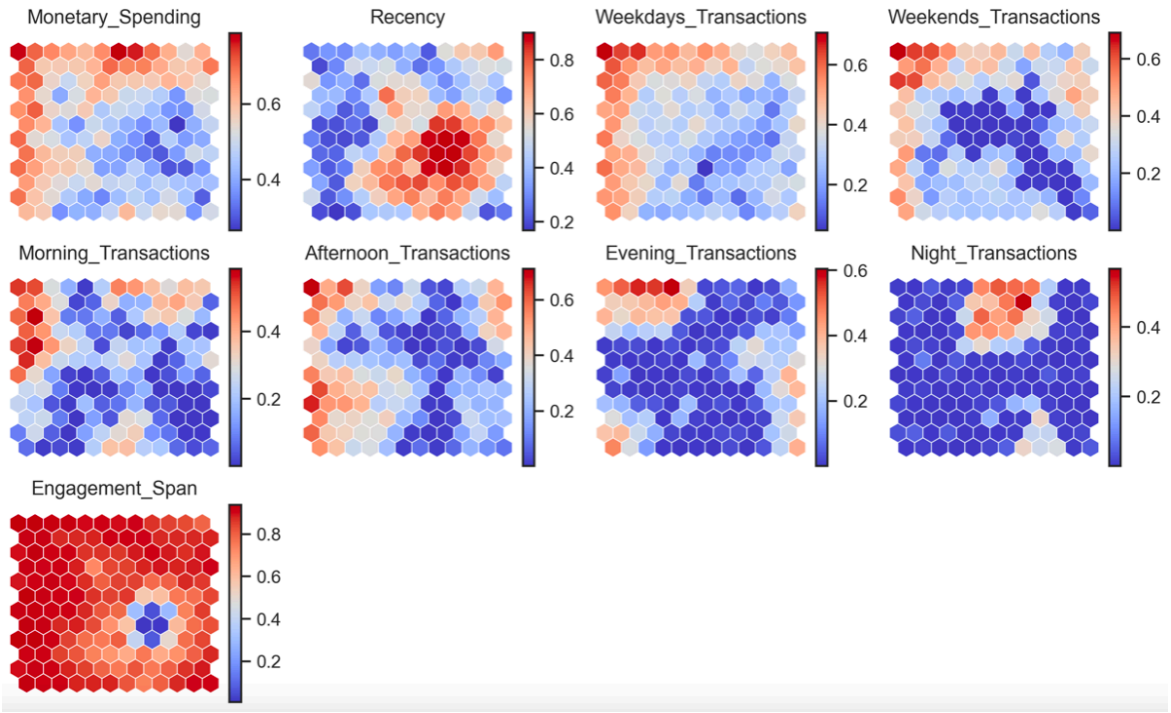
10.3.3 Hierarchical Clustering for Preference Variables, Figure 20



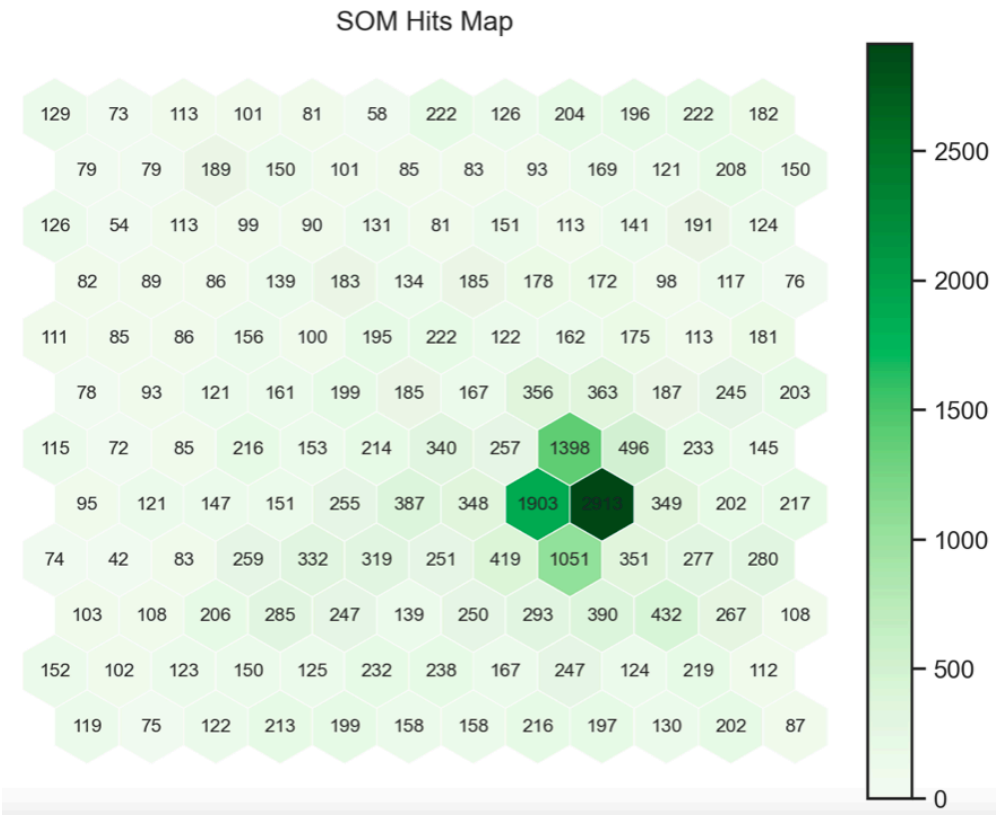
10.3.4 Hierarchical Clustering for Preferencer Variables Inertia and Silhouette graphs, Figure 21



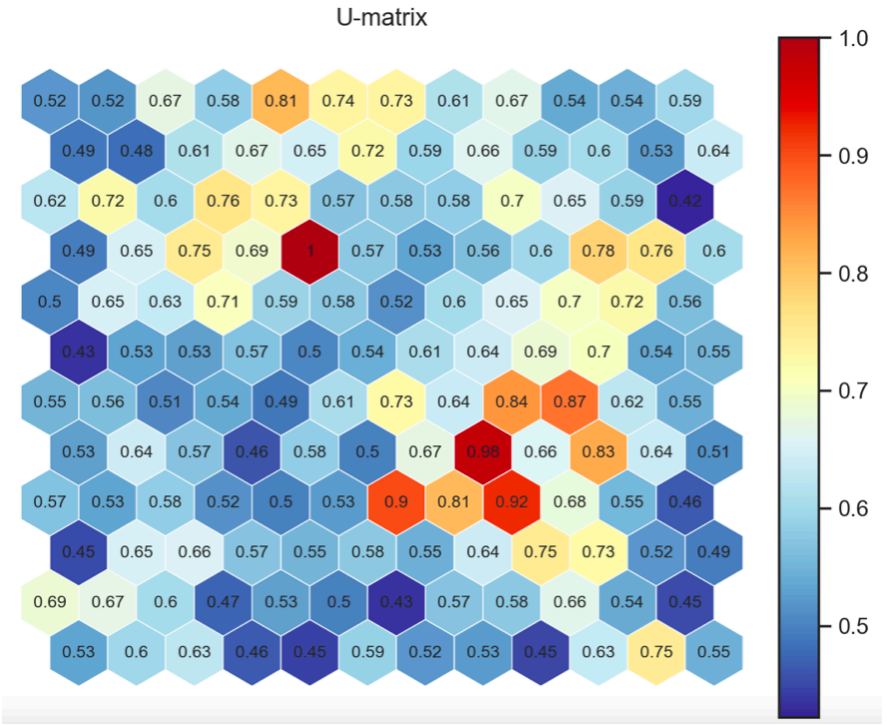
10.3.5. Self Organizing Maps (SOM) for Behavior Variables, Figure 22



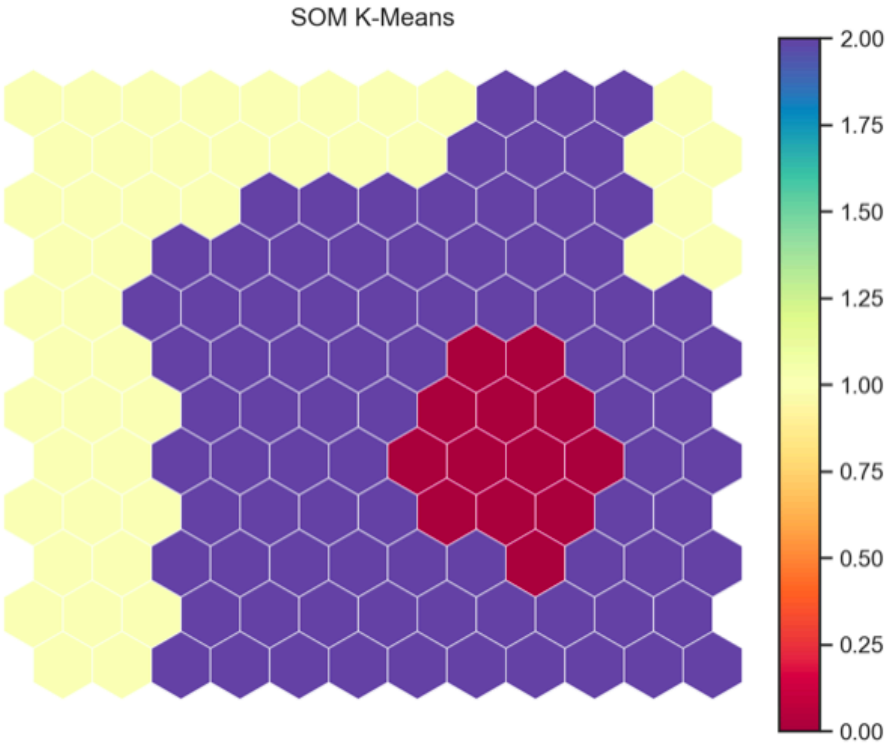
10.3.5. Self Organizing Map: HitsMap for Behavior Variables, Figure 23



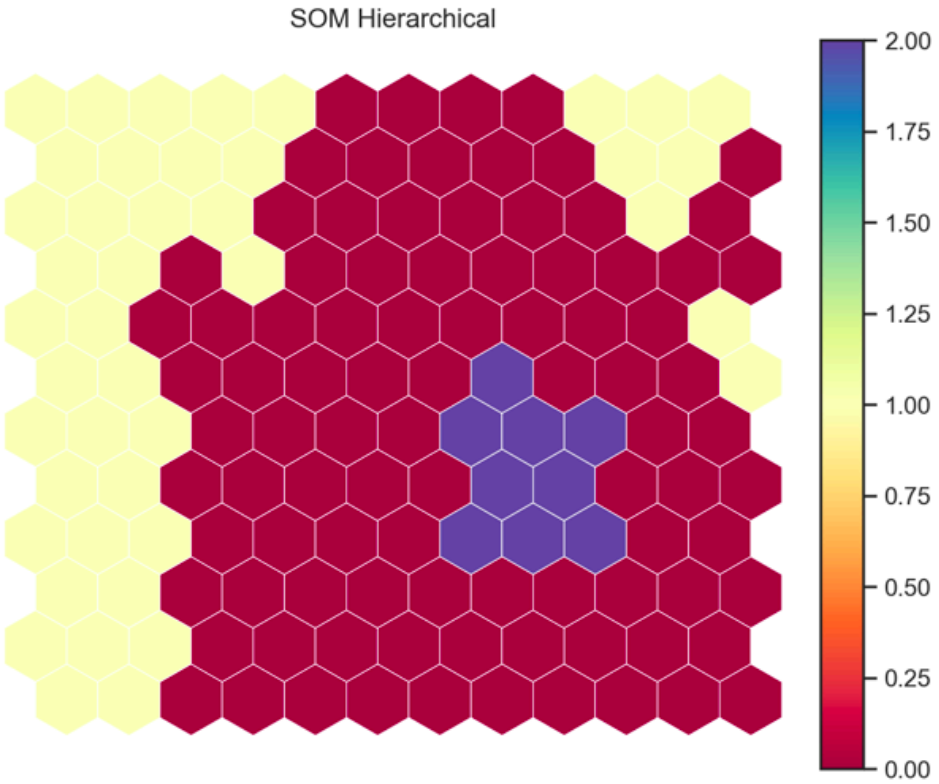
10.3.6. Self Organizing Map: U-Matrix for Behavior Variables, Figure 24



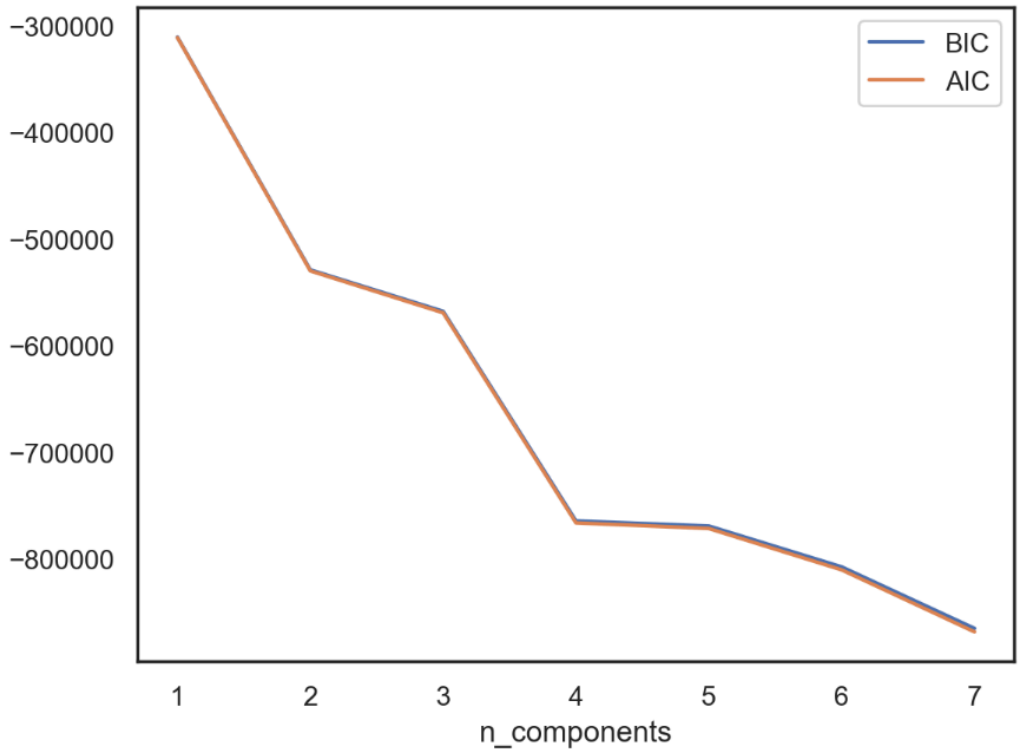
10.3.7. Self Organizing Map with K-Means for Behavior Variables, Figure 25 A



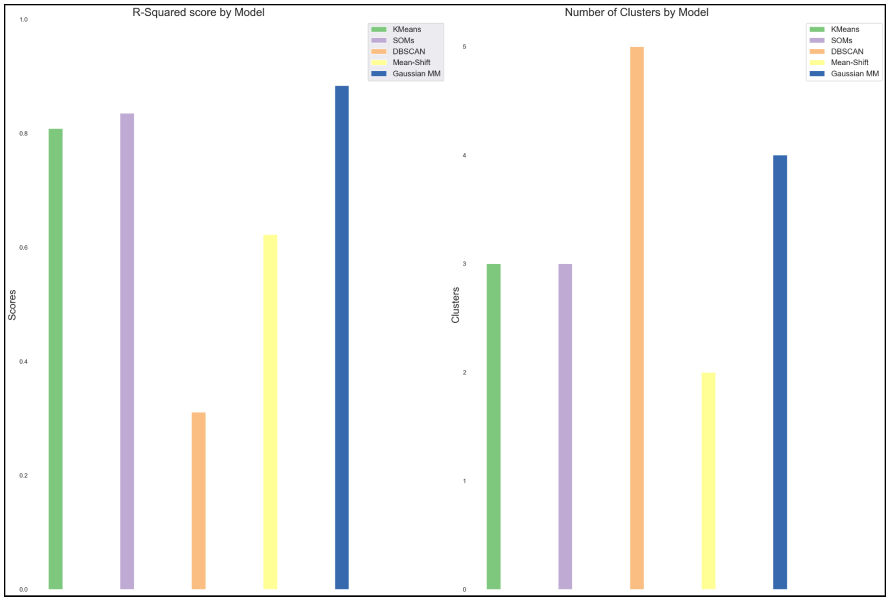
10.3.8. Self Organizing Map with Hierarchical Clustering for Behavior Variables, Figure 25 B



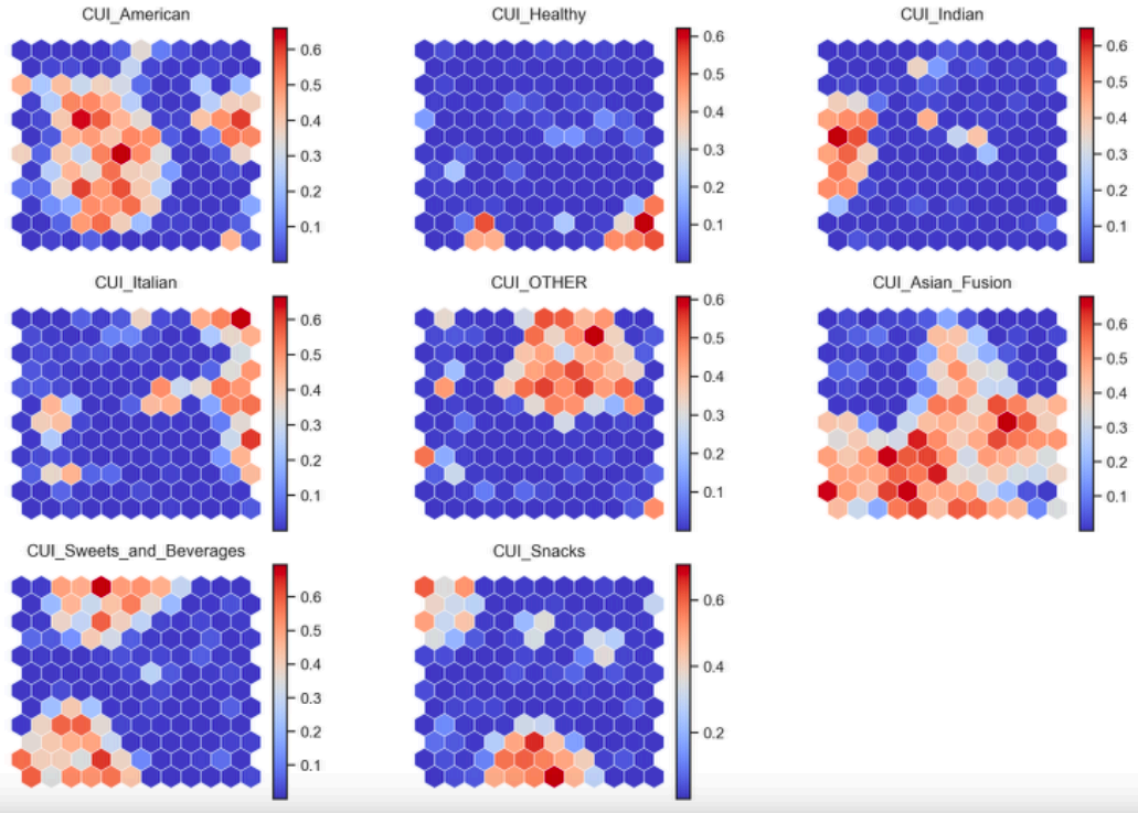
10.3.9. Plot of AIC and BIC evolution for different values of components- Behaviour Perspective , Figure 26



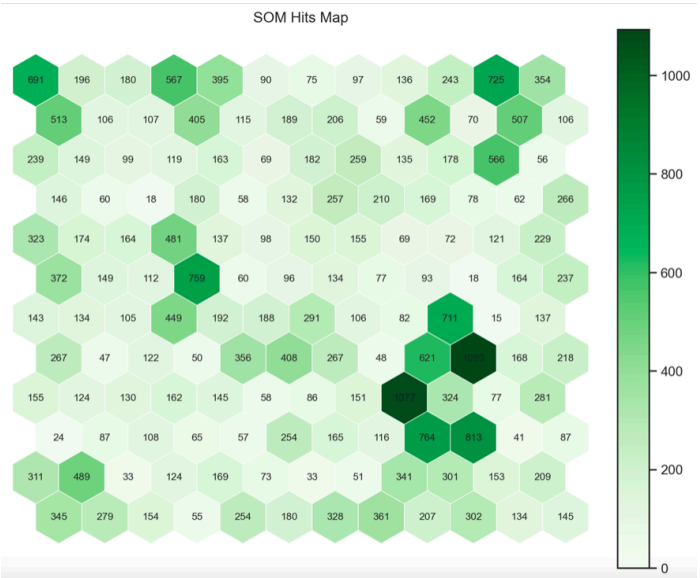
10.3.10. Comparing Models for Behavior Perspective, Figure 27



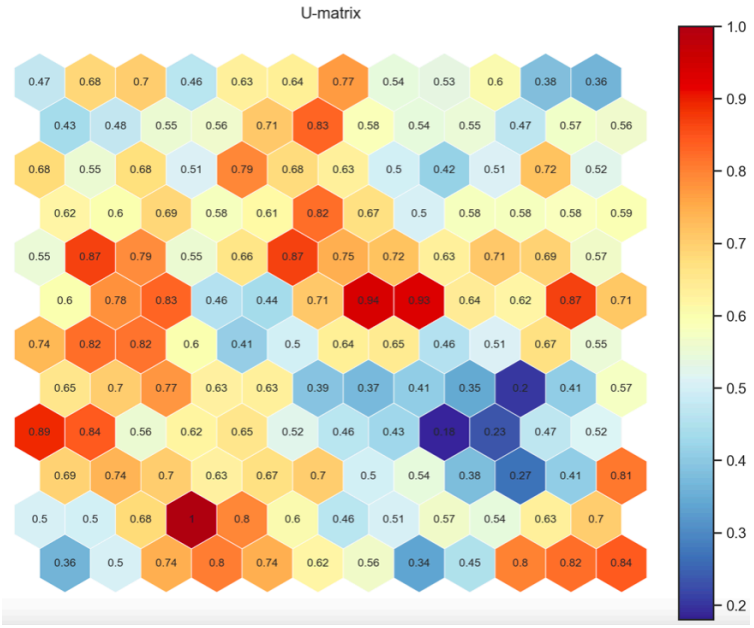
10.3.11. Self Organizing Maps (SOM) for Preference Variable, Figure 28



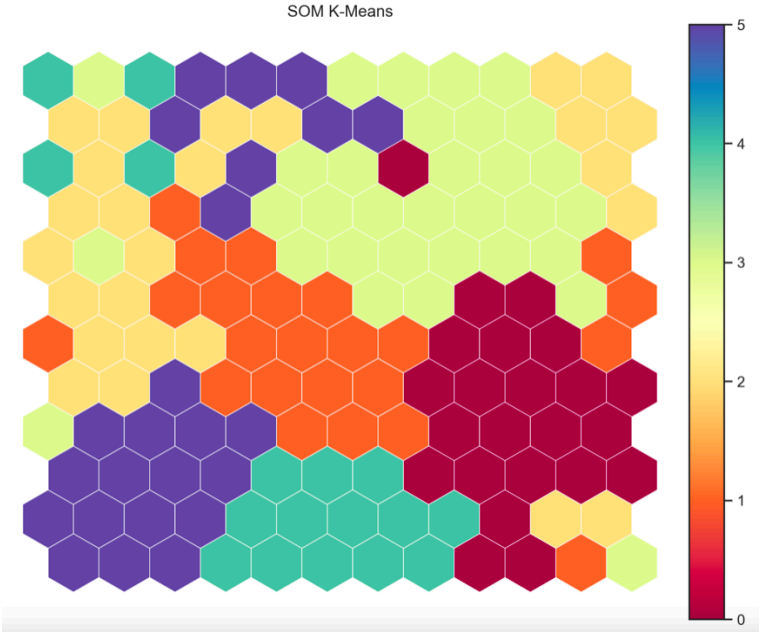
10.3.12. Self Organizing Map: HitsMap for Preference Variables, Figure 29



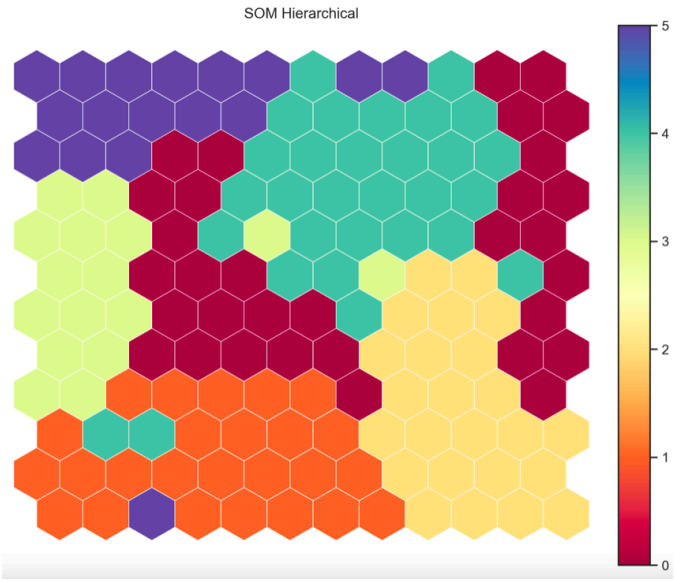
10.3.13. Self Organizing Map: U-Matrix for Preference Variables, Figure 30



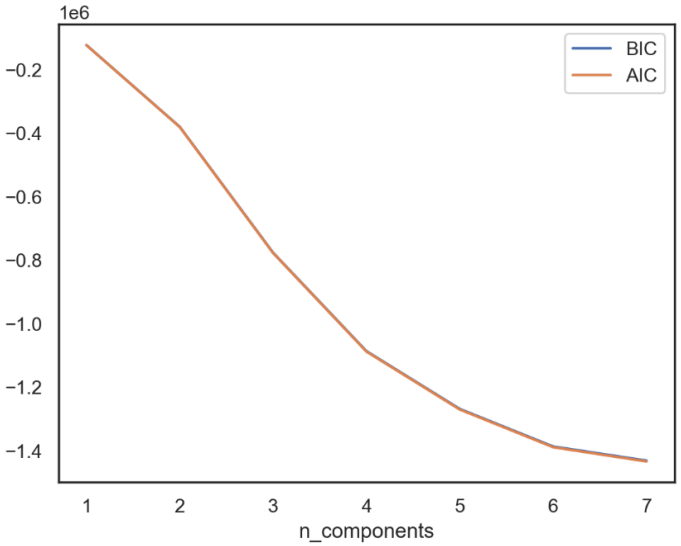
10.3.14. Self Organizing Map with K-Means for Preference Variables, Figure 31 A



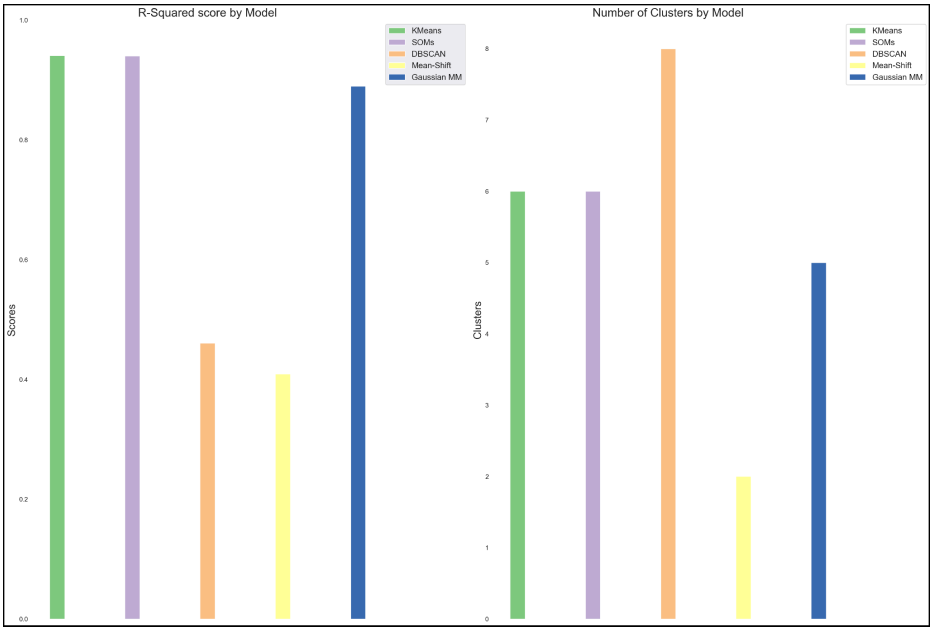
10.3.15. Self Organizing Map with Hierarchical Clustering for Preference Variables, Figure 31 B



10.3.16. Plot of AIC and BIC evolution for different values of components for Preference Perspective, Figure 32

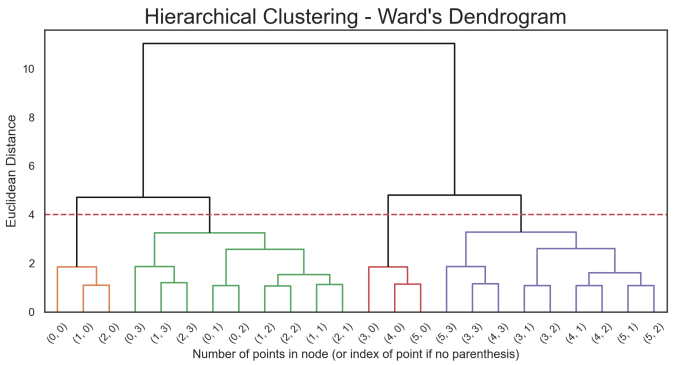


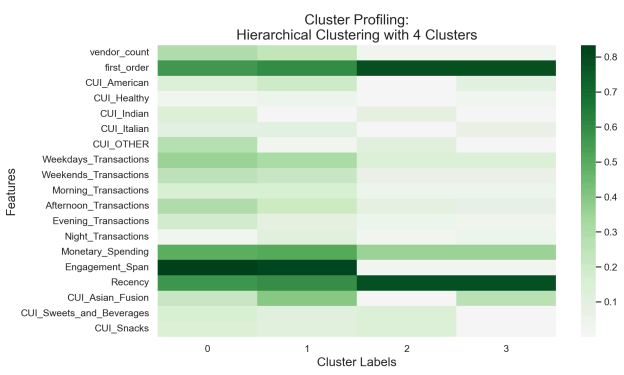
10.3.17. Comparing Models for Preference Perspective, Figure 33

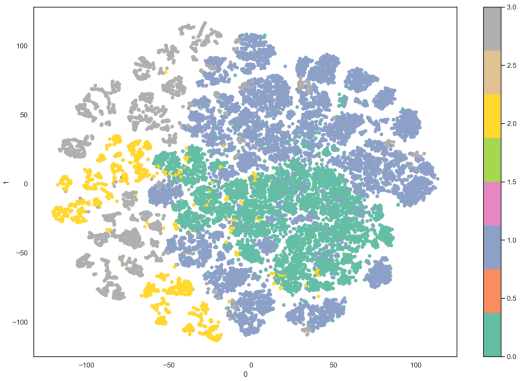


10.4 Final Clustering Appendix

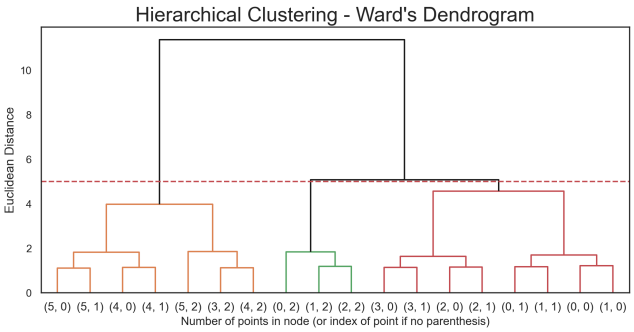
10.4.2. SOMs model for preference perspective and 4 merged clusters, Figure 34

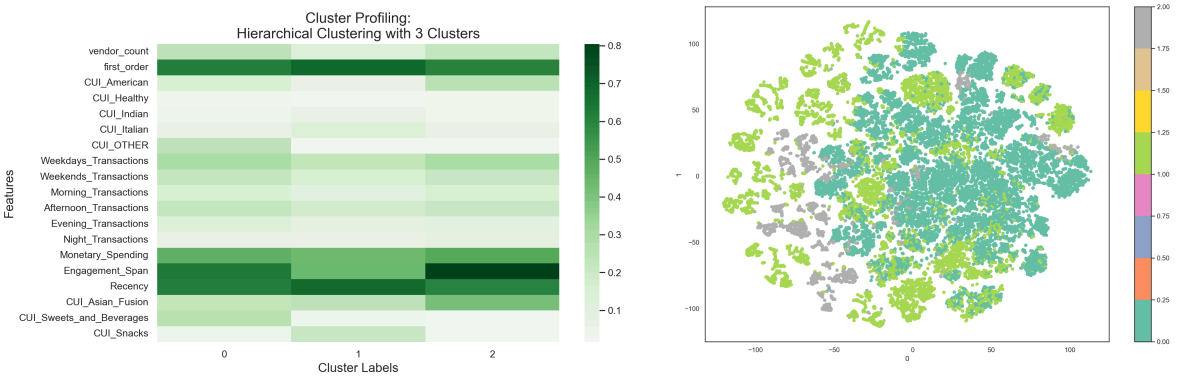




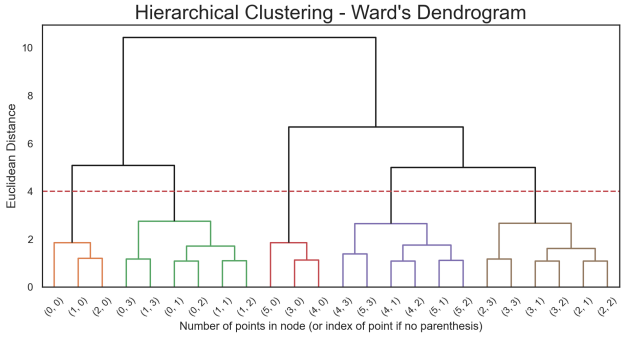


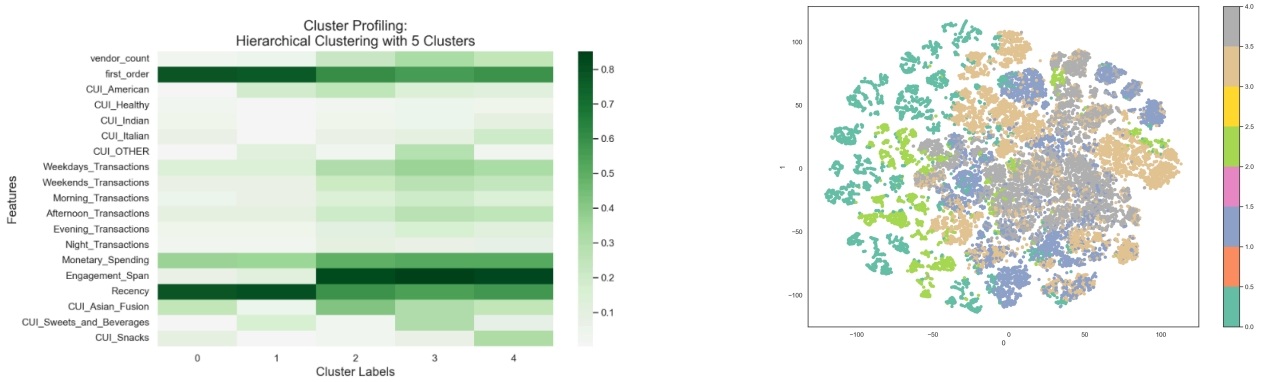
10.4.3. K-means model for preference perspective and 3 merged clusters , Figure 35





10.4.4. K-means model for preference perspective and 5 merged clusters , Figure 36



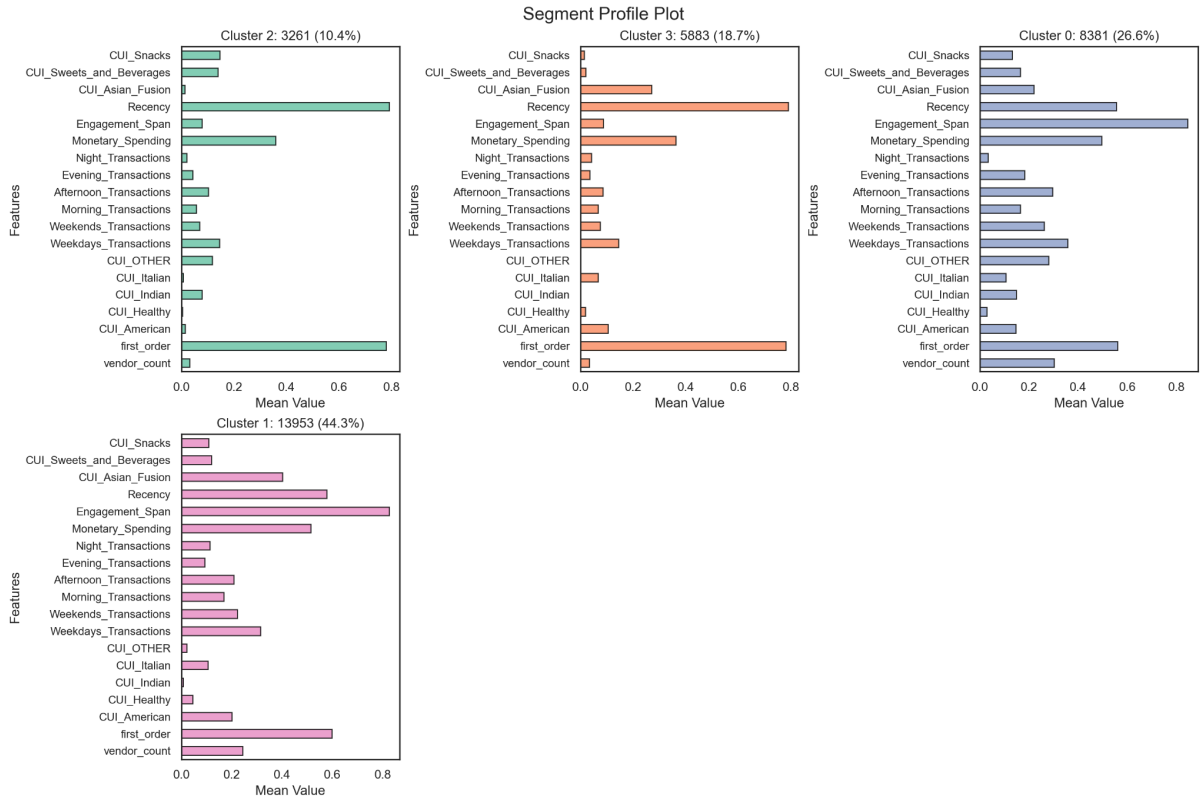


10.4.4. Comparing Models for merged clusters (note: the different models on the table are just being tested on the preference perspective), Table 1

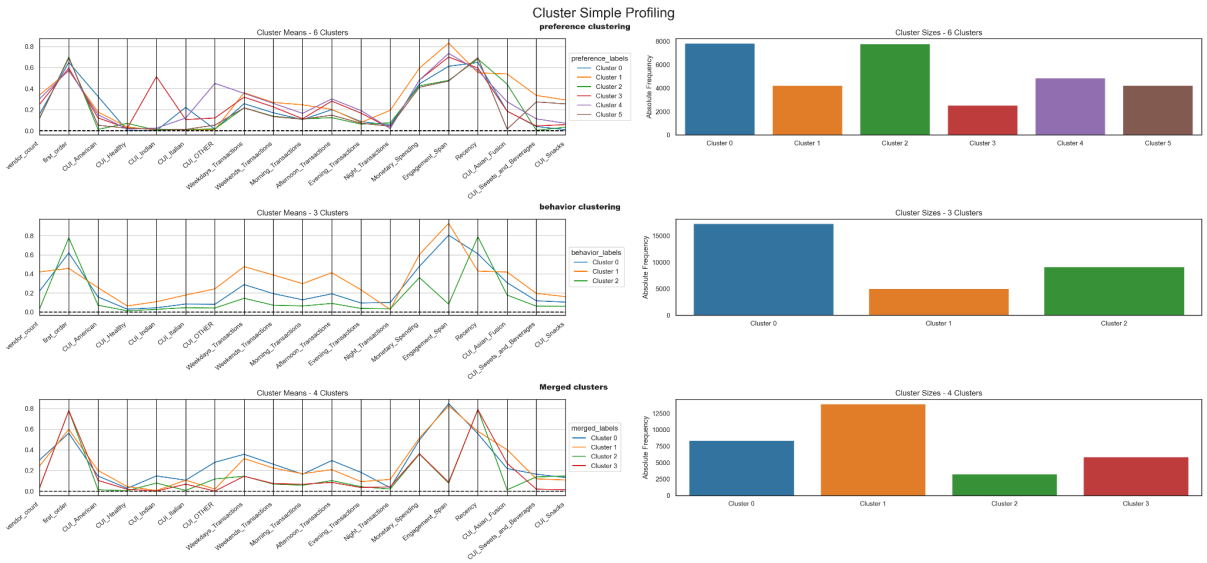
Clustering Method	Cluster Separation	Interpretability	Balance	Observations
K-means (3 Clusters)	Moderate	High	Moderate	Broad segmentation for general strategies.
K-means (5 Clusters)	High	Moderate	Low (smaller clusters may dominate)	More granularity, way less balance
SOMs (4 Clusters)	High	Moderate	High	Best balance of separation and interpretability for detailed marketing.

10.5 Cluster Characterization

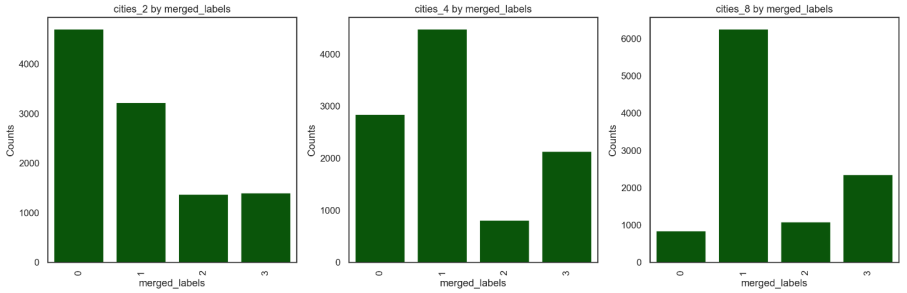
10.5.1. Segment Profile Plot for final merged clusters, Figure 37



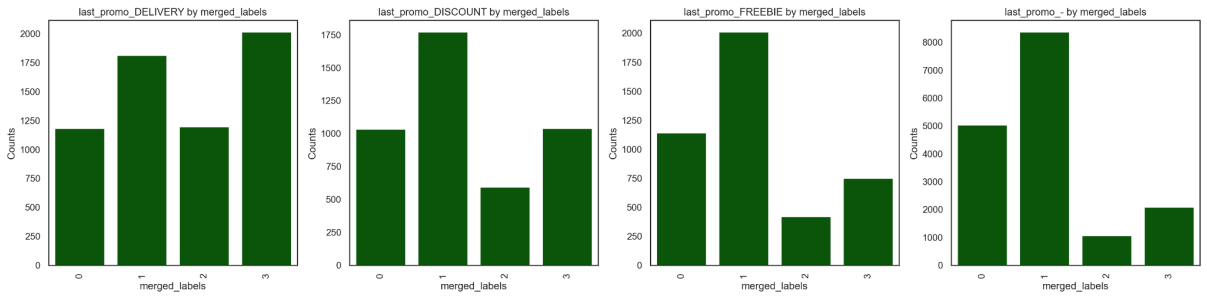
10.5.2. Cluster Simple Profiling, Figure 38



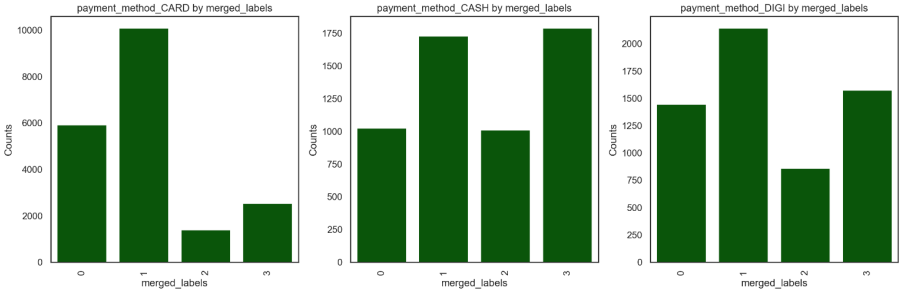
10.5.3. Cities bar plots of merged clusters, Figure 39



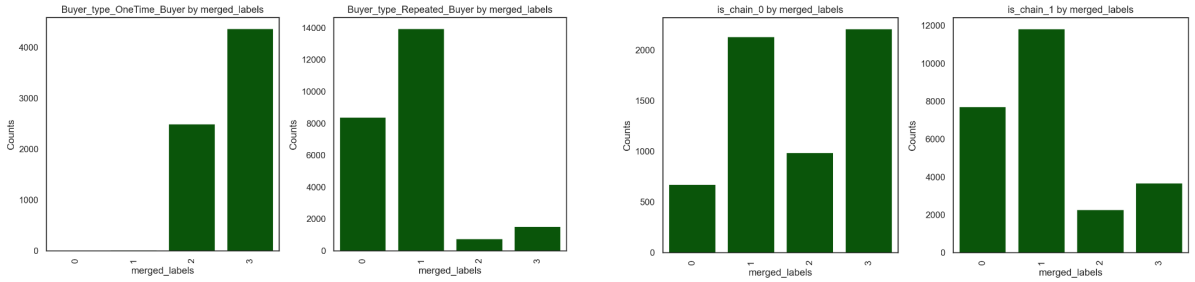
10.5.4. Promo Codes bar plots of merged clusters, Figure 40



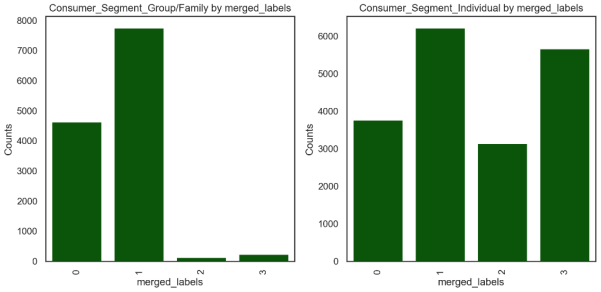
10.5.5. Payment Methods bar plots of merged clusters, Figure 41



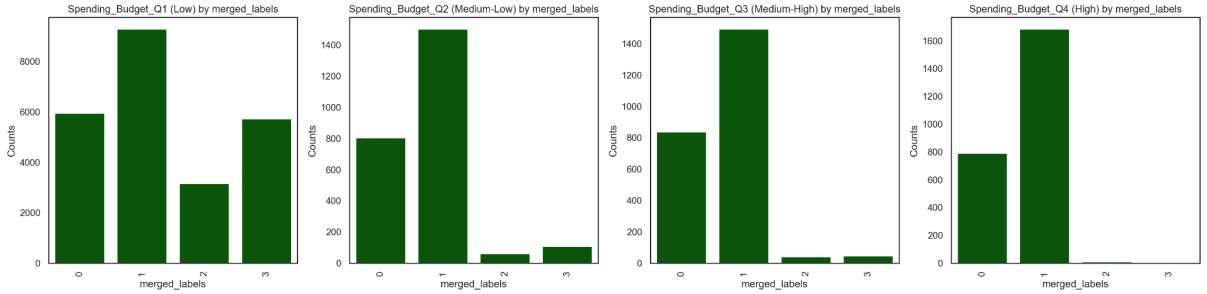
10.5.6. Buyer Type and **is_chain** bar plots of merged clusters, Figure 42



10.5.7. Consumer Segment bar plots of merged clusters, Figure 43



10.5.8. Consumer Segment bar plots of merged clusters, Figure 44



10.6 Additional notes about EDA

We identified more duplicates in this part , since we select the instances the customer_id.

We added more new features, wich are:

- Customer Segment
- Spending Budget
- Buyer type
- Engagement Span

We also analysed the outliers of this new features.