DATA MINING PROJECT

Master in Data Science and Advanced Analytics

NOVA Information Management School

Universidade Nova de Lisboa

ABCDEats Inc. (2nd Part)

Group 37

André Silvestre, 20240502

Filipa Pereira, 20240509

Umeima Mahomed, 20240543

Fall/Spring Semester 2024-2025

TABLE OF CONTENTS

1. INTRODUCTION

In an increasingly competitive market, businesses have been aiming to better understand their customers, offering products and services that closely align with their needs. By partitioning their customers into groups, they are able to tailor their strategies and improve satisfaction, loyalty, and profits. This practice, which is becoming more essential, is supported by many studies that highlight the importance of customer segmentation as a foundation for lasting client relationships. [1] [2] When executed effectively, segmentation enables a more comprehensive outline of resource allocation, allowing for more impactful investments in areas that yield the greatest return. [3]

While looking at what was done before, numerous studies have investigated customer segmentation, employing various methodologies (**Appendix A - Table A1**). A comparative analysis of unsupervised learning algorithms [4] highlighted the effectiveness of techniques like K-Means, Agglomerative clustering, DBSCAN, and Mini-Batch K-Means for grouping customers. Further research [5] demonstrated the power of combining clustering with the RFM model, while other studies (like [6] and [7]) emphasize the use of K-Means with RFM metrics. Several approaches also conduct a comparative study to determine the best method.

Industry-specific studies, such as [11] in the pizza delivery sector, further illustrate the widespread relevance of customer segmentation. Additionally, research in banking [12] has shown the value of segmentation using transactional activity data. The use of PCA for dimensionality reduction is also addressed in some articles [10] and [13], with some studies confirming PCA can improve clustering performance. This literature indicates that different clustering approaches can provide valuable customer segmentation insights across industries, with some studies confirming that K-Means, GMM, Agglomerative, and DBSCAN often produce good results.

Building upon these insights, this project will apply several clustering algorithms to segment ABCDEats Inc.'s customers. These algorithms include Hierarchical Clustering (HC), K-Means, Self-Organizing Maps (SOM), Mean Shift, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Gaussian Mixture Models (GMM). These methods were selected for their ability to uncover distinct groups based on different underlying data patterns and density characteristics. Furthermore, given the potential for high dimensionality in the dataset, particularly regarding cuisine preferences (*CUI_American*, *CUI_Asian*, etc.), day-of-week (*DOW_0* to *DOW_6*), and hour-of-day order frequency (*HR_0* to *HR_23*), we will also leverage Principal Component Analysis (PCA). This will help us mitigate the curse of dimensionality and to improve the quality of our clustering algorithms by removing noise and focusing on the most relevant components of our data.

The remainder of this report follows a structured approach based on the CRISP-DM methodology (**Annex A.**). [14][15] **Section 2** details the data preprocessing steps, which include cleaning, feature engineering, and dimensionality reduction, to prepare the dataset for modelling. **Section 3** covers the application of diverse clustering algorithms, including a preliminary exploration through different segmentation perspectives, as well as a detailed comparative evaluation of their performance. **Section 4** builds on the clustering by presenting the profiling of the resulting clusters and how these can be translated to business opportunities and strategies. Finally, **Section 5** outlines our conclusions, summarizing the main findings, limitations, and avenues for future research. A visual representation of the overall project workflow is included in **Appendix B**.

2. PREPROCESSING

This section details the data cleaning, preprocessing, and feature engineering steps undertaken to prepare the dataset for customer segmentation analysis. [14]

2.1. Handling Duplicate Entries

Our dataset currently consists of 31,737 rows and 66 columns, following the preprocessing steps conducted in *Part 1* (**Appendix C.**). To enhance the EDA, based on feedback received, we searched for duplicate entries (excluding *customer_id*) and identified 47 cases. Given that this number is small and nearly half of these cases involved customers who placed only one or two orders at restaurants that may offer fixed-price menus, we considered that these could represent distinct customers with similar preferences, which would be valuable for segmentation. Therefore, we chose to preserve these cases.

2.2. Handling Data Inconsistencies

Inconsistencies were identified in Part 1 (**Appendix C.**), with most addressed during that phase. However, we retained 18 cases where *vendor_count* exceeds *product_count* and *order_count* exceeds *product_count*, which are illogical. After analysis, we found no clear pattern to explain these inconsistencies. Recognizing that we might lose minimal information by potentially excluding customers with valid values, but given that these cases occur only 18 times, we chose to discard these customers to minimize noise and prioritize the overall data quality of our solution.

2.3. Addressing Missing Values

Missing values were found in *customer_age* (2.29%), *first_order* (0.33%), and *HR_0* (3.67%). Consequently, derived variables like *customer_age_group*, *days_between_orders*, and *days_between_orders_per_order* also contain missing data, which will be resolved when the original variables are imputed. Since all missing values are well below 5%, we proceeded with imputation. Imputation strategies, along with their pros and cons, are summarized in **Table D1**.

2.4. Feature Engineering

In addition to the 10 new features derived in Part 1, we explored creating 15 new variables representing the proportion of spending per customer on each specific type of cuisine. This approach aimed to create a set of metrics that were better aligned with the behavioural perspective that we are focusing on. However, after visualizing the distributions of these new variables, we noticed that most of them were heavily skewed towards zero, with their means and medians also being near zero, and variances concentrated near zero (**Figure D2**), which suggested that these new variables were not capturing any new relevant information, as many customers had zero spending on a specific type of cuisine. Therefore, we decided to only use the *CUI* original variables.

2.5. Outlier Detection and Handling

Initially, we applied the Interquartile Range (IQR) method for outlier detection but led to the removal of all records. This was due to the presence of many zero values, due to the nature of the dataset, particularly within the *CUI*, *HR*, and *DOW* features, signalling any values different from zero as outliers. To address this, we adopted a mixed strategy that combined:

1. **Modified IQR:** The IQR method was applied to variables not dominated by zero.
2. **Manual Outlier Removal:** Boxplots and domain knowledge were used to identify and remove only the most extreme outliers. (**Table D2**)

This mixed method retained 98.61% of the original data, adhering to the rules of thumb of not removing more than 5% of data. Post-removal boxplots and histograms (**Figure D3**) revealed significantly fewer extreme values, ensuring the data remained representative and suitable for analysis.

2.6. Variable Selection: Redundancy vs. Relevance

To enhance the robustness of variable selection, Pearson correlation was replaced with Spearman correlation, which better handles outliers and captures non-linear relationships. Using a redundancy threshold of **0.8**, several variables with multicollinearity were identified (**Figure D4**):

- ***product_count* vs. *order_count* (0.95):** Prioritized *order_count*, as it aligns with ABCDEats' operational perspective and has fewer outliers;
- ***vendor_count* vs. *CUI_Total_Food_Types* (0.89):** Retained *CUI_Total_Food_Types* due to fewer outliers and lower correlations with other variables. Removing *vendor_count* also resolved its high collinearity with *product_count*;
- ***order_count* vs. *days_between_orders* (0.83):** Selected *order_count* as it is more relevant to the analysis; *days_between_orders* had missing values and was primarily used as an auxiliary variable to calculate *days_between_orders_per_order*;
- ***customer_age_group* vs. *customer_age* (0.83):** We removed both of these variables because they had little correlation with other variables and low relevance to the overall analysis, as observed in Part 1. (**Figure D3**, **Table D3**)

As a result of this process, the following variables were not selected: *vendor_count*, *product_count*, *days_between_orders*, *customer_age* and *customer_age_group* due to redundancy or low relevance. By eliminating these variables, we may improve the clustering algorithms and focus on the most impactful variables.

In order to visualize the correlation between categorical variables using *Cramer's V* (**Figure D5**), we encoded them (using *One Hot Encoder*) however, this type of variable is not relevant to the next steps, due to only using Euclidean Distance based algorithms for clustering, although still relevant for the cluster profiling.

2.7. Feature Scaling

Since the algorithms we will use rely on distance to create clusters, as mentioned before, it's crucial that our features are measured on a consistent scale to avoid any bias that might result from giving more weight to certain variables. We selected the ***Standard Scaler*** because it is more robust to outliers compared to *Min-Max Normalization* — an important factor given our conservative approach to outlier removal. Additionally, the *Standard Scaler* is well-suited for the subsequent PCA, which calculates principal components to maximize variance in the data. Without proper scaling, variables with larger ranges could dominate the principal components, skewing the analysis.

While the *Standard Scaler* has its limitations—most notably making scaled values less intuitive — it remains the most appropriate choice given our objectives and dataset characteristics.

2.8. Principal Component Analysis (PCA)

Given the considerable number of numerical variables remaining after preprocessing, we employed Principal Component Analysis (PCA) to reduce dimensionality, remove noise and redundancy, although losing information. By reducing our dataset to a lower dimensional feature space, we aimed to improve

the performance of our clustering algorithms. We applied PCA separately to the Cuisine (CUI), Day of the Week (*DOW*), and Hour of Day (*HR*) feature groups, in order to simplify the interpretation of the component vectors and their impact in customer behaviour and also because different types of variables might have different levels of variance explained.

For the *CUI* variables, based on our analysis of the scree plots and retaining most of the data's variance, we opted for the first 7 principal components (PCs), capturing over 80% of the variability (**Table D4**, **Figure D6**). Similarly, for *HR* variables, we retained 4 PCs, based on the elbow method and variance explained criteria (**Table D6**, **Figure D8**). Regarding the *DOW* variable, after observing that the first component captured considerable variance across all days, thus lacking the desired interpretability, we decided to drop the corresponding PCs and use the original *DOW* variables instead (**Table D5**, **Figure D7**). After the application of the PCA to the *CUI* and *HR* variables, we then assigned interpretable names to these components based on the high or low loading that each variable had in the resulting vectors, reflecting the data's underlying structure and aided in later analysis (**Table D7**, **Table D8**).

After applying PCA, we finalized our preprocessing by performing a final exploratory data analysis. This step involved a visual assessment of the new distributions using histograms and boxplots (**Figure D9**) which confirmed the effective removal of outliers, while also evaluating the final correlations between metric features using a heatmap (**Figure D10**). As indicated by the heatmap, we were able to reduce the initial multicollinearity of several variables, which are now represented by their corresponding PCs (for the *CUI* and *HR*). Furthermore, the distributions of numeric variables, combined with boxplots, now show fewer extreme values, reflecting the effectiveness of our preprocessing.

3. CLUSTERING

This section details the application of various clustering techniques to segment the *ABCDEats* customer base, transitioning from the data preparation phase to the modelling stage of the CRISP-DM methodology. [14] In this phase, we will perform clustering with the aim of identifying distinct customer groups based on behaviour and value characteristics, thus enabling *ABCDEats* to tailor marketing and service strategies effectively. Ultimately, we will evaluate and compare clustering models to ensure robust and actionable results.

3.1. Analysing by perspectives

To effectively segment customers, we first applied clustering to all previously selected metric variables, which we term the *"overall analysis"*. In addition to this, we then explored both value-based and behaviour-based perspectives, each offering a unique lens for understanding customer heterogeneity. We focus on these two types of segmentation since they better portray the kind of analysis we want to do. Other perspectives, such as demographic segmentation, could also be valuable, but are not included in our current analysis due to limitations in the data, which doesn't provide sufficient metric variables for this kind of approach.

1. **Value-Based Segmentation**: This approach aims to group customers based on their economic contribution to the company. Key variables include *CUI_Total_Amount_Spent, CUI_Total_Food_Types, CUI_Avg_Amount_Spent, order_count, days_between_ orders_per_order,* and *chain_count.* These variables allow identification of high-value customers who contribute significantly to revenue and enable strategies for personalized marketing and retention, crucial for profitability. [1]

2. **Behavior-Based Segmentation:** This method focuses on grouping customers according to their purchasing habits and preferences. Variables include *first_order*, *last_order*, preference in food categories (*CUI_NOTAsian_Italian_OTHER_NOTSnack_PC*, *CUI_American_Cafe_Japanese_PC*, etc.), order timings (*HR_Lunch_Dinner_PC*, *HR_LateNight_Breakfast_PC*, etc.), and day of the week ordered (*Sunday*, ..., *Saturday*). This enables targeted promotions and services based on how and when customers order, aligning with their usage patterns. [3]

3.2. Hierarchical Clustering (HC)

HC is a general-purpose clustering method that produces a hierarchy of clusters either via an agglomerative (bottom-up) process, in which one point at a time is incorporated into a cluster, or via a divisive (top-down) process, in which larger clusters are split into smaller and smaller clusters. One of the most significant features of HC is its capacity to generate a dendrogram and choose the optimal number of clusters, depending on the data. HC is especially well suited to work with datasets in which non-linearity is present. It can, however, have difficulty with noisy data and outliers, as well as the high computational complexity, that makes it impractical when working with massive datasets. [17][18]

We applied agglomerative HC to both the *overall metric features* and each segmentation perspective, using various linkage methods (*Ward*, *Complete*, *Average*, and *Single*) and exploring a range of cluster numbers (1 to 10). We used the R^2 metric to compare different linkage methods, noting that the ward linkage consistently achieved the best R^2 values (**Figure E1**). Following this, to determine the optimal number of clusters, we examined dendrograms produced using the ward linkage (**Figure E2**), and with 2 different thresholds, as it was possible to see clear jumps in the distance when forming 2 or 3/4 clusters, depending on the specific perspective. For the **overall analysis**, we visualized the dendrogram with *Euclidean* distance (ED) thresholds of 250 and 400; for the **value-based** perspective, thresholds of 200 and 300 were used; and for the **behaviour-based** perspective, we used thresholds of 250 and 350. We used this visual inspection, the dendrogram, to guide our selection of the most suitable number of clusters for each case.

Consequently, for the overall, we opted for 4 clusters ($ED = 250$) and for value and behaviour-based analysis we opted for 3 clusters ($ED = 200$ and $ED = 250$ resp.) based on the first jump we were able to observe in the dendrograms and the relative frequency (%) of each cluster (**Figure E3**).

3.3. K-Means

K-Means [19] is one of the most popular clustering algorithms that is simple and efficient. It divides the data into k clusters to minimize the within-cluster variance. This algorithm is computationally efficient, and it is scalable to large datasets, thus it is best suited for customer segmentation. Instead, it postulates that clusters are of fixed size and spherical shape, which would not hold in the real-world situations. Moreover, the algorithm requires the user to predefine the number of clusters, which can be challenging without prior knowledge of the data structure. [17][18]

We tested the K-Means algorithm with different values of k (ranging from 1 to 10 clusters), on the *overall metric features*, as well as for each segmentation perspective. To determine the optimal number of clusters, we used a combination of approaches. First, we analysed the inertia score, seeking the "elbow" in a plot of the sum of squared distances within (SSw) each cluster (**Figure E4**), and observed that $k = 3$ or $k = 4$ was an optimal range of values. However, since relying only on this method can lead to less meaningful results, this was complemented with visual inspection of the data with silhouette analysis plots, calculating the silhouette coefficient for each data point, that measures

how similar a data point is to its own cluster (cohesion) compared to other clusters (separation) (**Figure E5**), and evaluating the cluster distribution. Using this combination of inertia and silhouette analysis, we then selected a k value that produced a more reasonable cluster distribution, while also aiming for a manageable number of clusters. The final chosen cluster solutions were $k = 3$ for all approaches.

3.4. Self-Organizing Maps (SOM)

SOM are an artificial neural network for dimensionality reduction and unsupervised learning. [20] They project high-dimensional data to be represented on a low-dimensional grid and at the same time conserve the topological structure useful for visualization. [17][18] In contrast to other clustering algorithms, SOMs yield interpretable visual representation of the data, which can aid understanding of relationships between the data features, the cluster, and the outliers. One of the major strengths of SOMs is its capability to model nonlinear relationships and translate them into a structured grid, which enables SOMs to help in exploratory data analysis and pattern recognition applications.

Applying it to the project, we first applied SOMs using a 10x10 hexagonal grid with the *overall metric features* and our *perspectives*, having very similar performance, with a learning rate of 0.7, a Gaussian neighbourhood function, and *Euclidean* distance for activation calculations. Following training over 20,000 iterations, both the Quantization Error (QE) and Topographic Error (TE) decreased, indicating the SOM's effective learning and representation of the data (**Figure E6**), showcasing initial distributions of feature weights across the grid. Visualization techniques, such as component planes, the U-Matrix (**Figure E7**) – exhibiting higher distances at certain specific hexagons, suggesting possible outliers – and hit maps (**Figure E8**) – which highlights an imbalance in data point distribution across the map, with some nodes capturing more data – were then utilized to analyse feature relationships and discern the initial structure in the data.

To further refine the analysis and interpretability and address the data imbalance, we expanded the SOM grid to a 50x50 matrix for the *overall analysis* and both the *behaviour-based* and *value-based* perspectives. We retrained the SOM over 100,000 iterations (**Figure E9**, **Figure E10** and **Figure E11**). Despite this expansion, the issue of imbalanced data distribution persisted for the *overall analysis* with some units having a much higher number of data points. As a result, we decided to combine the SOM output with K-Means and HC (Ward linkage). For the *overall case* and *value perspective*, we applied HC using inertia to select $k = 4$. For the *behaviour perspective*, we also applied HC using inertia to determine $k = 3$ (**Figure E12**). We additionally applied K-Means to the SOM weights, and the k considered is the same as previously, except, for the **value** and **behaviour perspectives**, where an inertia and visual inspection indicated that $k = 3\ and\ k = 4$ resp. was the most suitable choice. (**Figure E13**)

The final solution combined the K-Means clustering results with a Best Matching Unit (BMU) mapping approach, by associating each data point with the cluster label of its corresponding BMU on the 50x50 SOM grid. The resulting R^2 value indicated a moderate level of clustering success, suggesting that the clustering solution captures some, but not all, of the variability present in the dataset. This approach allowed us to leverage the SOM's dimensionality reduction and organization capabilities while using K-Means to obtain well-defined clusters.

3.5. Density based Clustering

Density-based clustering algorithms identify clusters based on the density of data points, which is advantageous for detecting non-spherical clusters. We explored three different methods.

3.5.1. Mean Shift

Mean Shift identifies clusters by iteratively shifting data points toward high-density regions [21]. It does not require a predefined number of clusters but is highly dependent on the bandwidth parameter. For the overall analysis, we manually tested multiple bandwidth values and, after using *Silverman's* and *Scott's* rules of thumb, determined that a bandwidth of 9 resulted in a more interpretable four-cluster solution, discarding solutions with lower values since they did not provide enough information. For the *value* and *behaviour-based perspectives* the algorithm performed very poorly following the same trend as the *overall analysis*, with a very low R^2.

3.5.2. Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

DBSCAN algorithm groups tightly packed data points while treating sparsely populated areas as noise. G[22] We explored different eps values for the *overall analysis* using the k-distance, opting for $\varepsilon = 8$, since the dendrogram show a jump on this specific value, with $MinPts = 10$, and observed the identification of only two clusters (1 with 31020 obs. and the other with 259). For the *value* and *behaviour perspective* the performance was also very poor for the different eps and $MinPts$ values.

3.5.3. Gaussian Mixture Models (GMM)

GMM uses Gaussian distributions to model the dataset, assigning probabilities to each point for belonging to each cluster. [17] Based on AIC and BIC, a three-cluster solution was identified as optimal for the *overall analysis*. For the *value-based perspective*, a five-cluster solution from the density-based algorithms performed best, achieving an R^2 close to 0.45. However, for the *behaviour-based perspective*, a three-cluster solution was adopted. This density-based algorithm performed better than the previous two for all approaches. (**Figure E14**)

3.6. Final Clustering Solution

Our final clustering approach involved selecting the best solutions from different perspectives and algorithms based on a trade-off between R^2 values and the interpretability and number of clusters generated (**Table E1**). Specifically, for the *overall analysis* utilizing all metric features, we chose **SOM + K-Means** given its highest R^2 value among all methods (0.283). For the *value-based* perspective, we selected the **K-Means** approach with $k = 3$, yielding the R^2 (0.466). For the *behaviour-based* perspective, we selected **SOM + K-Means** again, which yielded an R^2 of 0.249.

Finally, to produce a more robust solution for segmentation, we **manually** merged the value-based K-Means output ($\boldsymbol{k = 3}$) with the behaviour-based SOM + K-Means results ($\boldsymbol{k = 4}$) by using their combined cluster centroids. In this manual intervention, clusters with less than 1000 observations were merged into closer clusters, producing a final solution with $\mathbf{5}$ **clusters** that are relatively more balanced. (**Figure E15**) We also tried merging with HC, however, it resulted in extremely imbalanced clusters. (**Figure E16**) These choices were also supported by previous research in customer segmentation, which suggests that a reduced set of 3 to 5 clusters for practical application in marketing strategies is a good option. (**Appendix A.**).

While both the overall and merged clustering approaches capture similar behavioural patterns, the merged approach resulted in a more balanced distribution of data points across the clusters and a better segregation between the different segmentts, therefore, providing a more insightful solution, despite having only one additional cluster.

4. PROFILING & BUSINESS APPLICATIONS

In order to assess the clustering outcome, we characterized each cluster by using barplots (**Figure F1**) that depict cluster sizes and heatmaps to visualize average values of selected characteristics (**Figure F2**). **Table 4.1** summarizes the 5 clusters, including their key features and suggested marketing strategies.

Table 4.1 – Segment Profiles & Recommended Marketing Approaches.

	Segment Profile & Key Characteristics	Recommended Marketing Approach
0 | The Mainstream Base 13057 (41.74%)	- Largest group, with spending and behaviour like the overall dataset, which shows moderate to low engagement levels. - Spends more on *Asian* and *American* cuisines. - Balanced across regions; primarily uses card payments.	- Offer tiered loyalty, with discounts and exclusive perks for higher spending and order frequency. - Target promotions for *American* and *Asian* cuisines and combo deals.
1 | The Promo Pursuers 11885 (38.00%)	- Has the lowest orders from all other groups, indicating a low engagement. - Has low total spending, but a significant average spends per order. - Utilizes delivery promotions, probably the motivation behind their orders. - Shows a slight preference for evening orders compared to other consumers. - Displays slight lower-than-average interest in *Noodles*, *Chinese* and *Chicken* dishes.	- Offer free delivery for orders exceeding a slightly higher value. - Implement a rewards program where frequent orders earn points that can be redeemed for discounts or free delivery to motivate more frequent orders.
2 | The Convenience Seekers 2679 (8.56%)	- High concentration in *Region 2*. - Shows a greater preference for *Chicken*, *Chinese*, *Noodles*, and *Other* cuisines compared to the average and less for *Asian*, *Street Food*, and *Snacks* cuisines compared to the average. - Highest order frequency across all days of the week especially during lunch and dinner. While they place the most orders, they are not the highest spenders, though they still spend significantly.	- Focus on promoting a premium dining experience over discounts, especially in *Region 2*, by providing personalized services. - Exclusive menu sneak peeks, early-access to new cuisine options. - Introduce a loyalty program that rewards not only order frequency but also spend per order, with bonus points for premium items or larger combos.
3 | The Balanced Spenders 2115 (6.76%)	- Similar behaviour to *Cluster 2*, but lower on order numbers and spending amounts. - Mostly located in *Region 2* and *Region 4*. - Also places orders mostly during lunch and dinner. - Prefers *Italian* and *other* cuisines compared to other groups; less keen on *Street Food*, *Snacks*, or *Asian* cuisines.	- Highlight *Italian* and other preferred cuisines in promotions, offering exclusive deals and limited-time specials. - Target promotions for lunch and dinner orders. - Offer discount combos for greater amount spent.

4 |The Late-Night Enthusiasts 1543 (4.93%)	- The highest spenders both in absolute and average terms. - Predominantly located in *Region 8*. - Strong preference for *Asian*, *Snack* and *Street Food*. - Less preference for *Italian* and *other* cuisines, contrary to cluster 3. - Orders primarily during late night and breakfast hours.	- Highlight breakfast and late-night specific food items and offerings. - Introduce city-specific promotions targeting customers in *Region* 8, offering exclusive discounts or early access to new menu items - Offer special discounts or VIP access for the highest spenders, incentivizing continued high-value orders.

Cluster 0, the '*Mainstream Base*' (41.74% of customers), exhibits average spending and ordering behaviour with balanced choices as compared to the rest of the clusters, as they are similar to the whole dataset's EDA. **Cluster 1**, '*The Promo Pursuers*' (38%), shows a propensity for delivery promotions and tends to order during the evening, while presenting a lower total spending in the period but higher average order spending. **Cluster 2**, the '*Convenience Seekers*' (8.56%), is highly active and has an above average order frequency during lunch and dinner but has less preference for *Asian* and *Street Food*. **Cluster 3**, named the '*Balanced Spenders*' (6.76%), has similar activity periods as the previous cluster but less frequent, however with a slight preference for *Italian* food. Finally, **Cluster 4**, identified as '*The Late-Night Enthusiasts*' (4.93%), presents unique characteristics such as the highest spending, and a propensity for ordering during late night and early breakfast hours, and are most concentrated in *Region 8*.

To further understand what drove these segment differences, we assessed feature importance using both R^2 analysis, highlighting *order_count*, *HR_Lunch_Dinner_PC* and *CUI_Total_Food_Types* as key separation factors, and decision tree analysis, which highlighted *days_between_orders_per_order*, *order_count*, *HR_Lunch_Dinner_PC*, and *CUI_Total_Food_Types* as the primary differentiating features (**Figure F4**). The high importance of *order_count* and the *HR_Lunch_Dinner_PC*, highlighted how the frequency of orders and specific ordering periods have a high effect on customer behaviour, meaning that they should be a key factor in future targeting strategies.

Finally, we used dimensionality reduction techniques were used to represent and test the cluster separation through t-SNE (**Figure F5**) and UMAP plots (**Figure F6**). They both showed a relative distinction between clusters, but also a continuous behaviour with some overlapping of the data points, showing that perfect separations are difficult to obtain for this kind of multidimensional analysis. This implies that our chosen clustering solution provides a strong framework for customer segmentation, identifying relevant groups that are supported through all the analysis.

Beyond the clustering algorithms, we examined customer behaviour through a cell-based approach, specifically focusing on the interplay between order frequency and total spending. This approach involved using quartiles to categorize customers based on their order count and total spending (**Table F1**). This revealed that a large proportion of *ABCDEats* customers fall within the first or second quartiles of spending, with many also in the first quartile for order count. Notably, we also identified a significant segment (15.17%) who are high spenders and place a high number of orders. This shows an imbalanced behaviour from a large part of the user base, while a small group of users concentrates the biggest value of our database.

5. CONCLUSION

In this project we were able to segment the *ABCDEats* customer base, using an effective, multifaceted technique, that involves several clustering methods, which finally resulted in practical outcomes for the company. We first performed an "*overall*" analysis of all metric variables chosen and then narrowed our comprehension by examining value and behavioural perspectives. These several different strategies, and several clustering algorithms, made a plausible and more in-depth analysis. Key findings include the identification of five distinct and interpretable customer segments, achieved through a merged clustering solution derived from K-Means and SOM + K-Means.

Our methodology involved comparison between different algorithms, including HC, K-Means, SOM, Mean Shift, DBSCAN and GMM. We used consistent evaluation metrics, i.e., R^2 and silhouette scores, to choose the most fitting clustering method for each view and considered the trade-off between explained variance and cluster complexity. Notably, the summary of the resulting clusters (**Table 4.1**) provides practical intelligence on client preferences, ordering, and promotion responsiveness, and they both play an important role in the plan for developing effective targeted marketing plans.

Main findings, based on our analysis, demonstrated the *"Late-Night Enthusiasts"* to be high-value customers, suggesting that *ABCDEats* may offer specific VIP discounts, while categories such as *"Balanced Spenders"* display contrasting preferences in terms of ordering time, region, cuisine and spending amount, suggesting, for example, the introduction of special combos for *Italian* and "*Other Cuisines"* to boost their engagement. Incorporating these insights meets up with the aims of the project by highlighting key customer activities (e.g., order number and choice of dishes) that are very important for the purpose of personalized (target) marketing.

Despite these achievements, we acknowledge that the clustering methods used — including K-Means, GMM, DBSCAN and SOM—are subject to inherent assumptions related to cluster shape, data distribution and parameters selection, thus affecting their performance. This resulted in a necessary reliance on techniques such as PCA and SOMs, which while enhancing the clustering performance also did result in some loss of feature interpretability. These clustering algorithms also relied in Euclidean distance, restricting the use of demographic data (represented mainly categorically), which constrains our ability to fully comprehend the customer segments' needs and behaviours, and prevented further refinement based on demographic attributes.

In the future we suggest further research include other types of algorithms (that support categorical data) to generate a more granular segmentation, and to advance longitudinal studies that can track customer behaviour over time, thereby increasing the ability to respond to changing business environments. Enhancing the reliability and real-world relevance of the results by using increasingly complex algorithms, plurality data has a complementary effect.

In conclusion, this project provides *ABCDEats Inc.* with a robust and insightful framework for customer segmentation. Because analysis of this nature can be built upon and acted on appropriately, *ABCDEats* is able to continuously improve its approaches, identify its market and improve business.[1]

[1] Throughout this project, we utilized AI tools (e.g. *ChatGPT* and *Github Copilot*) to assist in code development and to help summarize repetitive conclusions derived from the data analysis. However, all AI-generated information was carefully reviewed and validated by the team to ensure accuracy and relevance.

BIBLIOGRAPHICAL REFERENCES

[1] Kotler, P., & Armstrong, G. (2018). *Principles of marketing* (17th ed.). Pearson Education South Asia Pte Ltd.

[2] Mcdonald, M., & Dunbar, I. (2012). *Market Segmentation: How to Do It and How to Profit from It* (4th ed.). John Wiley & Sons.

[3] Wedel, M., & Kamakura, W. A. (2000). Market Segmentation: Conceptual and Methodological Foundations. In *International Series in Quantitative Marketing*. Springer US. https://doi.org/10.1007/978-1-4615-4651-1

[4] Gupta, R., Sanjeev Subedi, Singh, A., & Shivam Kumar Singh. (2024). Comparative Study of Unsupervised Learning Algorithms for Customer Segmentation. *IEEE*, *2024 11th International Conference on Computing for Sustainable Global Development (INDIACom)*. https://doi.org/10.23919/indiacom61295.2024.10498443

[5] Abednego, L., Cecilia Esti Nugraheni, & Salsabina, A. (2023). Customer Segmentation: Transformation from Data to Marketing Strategy. *IAIC International Conference Series*, *4*(1), 139–152. https://doi.org/10.34306/conferenceseries.v4i1.645

[6] Liu, Y. (2023). Customer Segmentation in User Behavior Analysis: A Comparative Study of Clustering Algorithms. *Highlights in Business, Economics and Management*, *21*, 758–764. https://doi.org/10.54097/hbem.v21i.14758

[7] Wani, A., M Priyanka, & R Prasath. (2023). Unleashing Customer Insights: Segmentation Through Machine Learning. *IEEE*, *2023 World Conference on Communication & Computing (WCONF)*. https://doi.org/10.1109/wconf58270.2023.10235136

[8] Kulkarni, S. (2022). Advance Customer Segmentation. *Academia.edu*. https://doi.org/107255556/s200_swanand

[9] Bartels, C. (2022). Cluster Analysis for Customer Segmentation with Open Banking Data. *2022 3rd Asia Service Sciences and Software Engineering Conference*. https://doi.org/10.1145/3523181.3523194

[10] Abdulhafedh, A. (2021). Incorporating K-means, Hierarchical Clustering and PCA in Customer Segmentation. *Journal of City and Development*, *3*(1), 12–30. https://doi.org/10.12691/jcd-3-1-3

[11] Koca, O. (2021, June 30). *Determining customer segmentation and behaviour models with database marketing and machine learning*. Pressacademia. https://www.academia.edu/86952482/Determining_customer_segmentation_and_behaviour_models_with_database_marketing_and_machine_learning

[12] Afonso, P., & Ferreira, B. (2019). *Business clients´segmentation based on activity: a banking approach* [MSc Thesis]. https://run.unl.pt/bitstream/10362/93269/1/TGI0262.pdf

[13] Afrin, F., Al-Amin, Md., & Tabassum, M. (2015). *Comparative Performance of Using PCA With K-Means And Fuzzy C Means Clustering For Customer Segmentation*. https://www.ijstr.org/final-print/oct2015/Comparative-Performance-Of-Using-Pca-With-K-means-And-Fuzzy-C-Means-Clustering-For-Customer-Segmentation.pdf

[14] Provost, F., & Fawcett, T. (2013). *Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking*. O'Reilly.

[15] Waqar, M. (2023, October 26). *Unlocking CRISP-DM: Your Path to Data Science Success*. Medium. https://medium.com/@mwaqarbatalvi/mastering-the-crisp-dm-framework-your-path-to-successful-data-science-projects-56f15d6f4c54

[16] Maćkiewicz, A., & Ratajczak, W. (1993). Principal Components Analysis (PCA). *Computers & Geosciences*, *19*(3), 303–342. https://doi.org/10.1016/0098-3004(93)90090-r

[17] Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys*, *31*(3), 264–323. https://doi.org/10.1145/331499.331504

[18] Berry, M. J. A., & Linoff, G. (2004). *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*. Wiley Pub.

[19] Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, *28*(2), 129–137. https://doi.org/10.1109/tit.1982.1056489

[20] Kohonen, T. (1982). Self-Organized Formation of Topologically Correct Feature Maps. *Biological Cybernetics*, *43*(1), 59–69. https://doi.org/10.1007/bf00337288

[21] Yizong Cheng. (1995). Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *17*(8), 790–799. https://doi.org/10.1109/34.400568

[22] Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*. https://file.biolab.si/papers/1996-DBSCAN-KDD.pdf

[23] A Com, L., & Hinton, G. (2008). Visualizing Data using t-SNE Laurens van der Maaten. *Journal of Machine Learning Research*, *9*, 2579–2605. https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf

[24] A McInnes, L., & Healy, J. J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv (Cornell University)*. https://arxiv.org/abs/1802.03426

APPENDIX A. LITERATURE REVIEW

Figure A1. – Literature Review

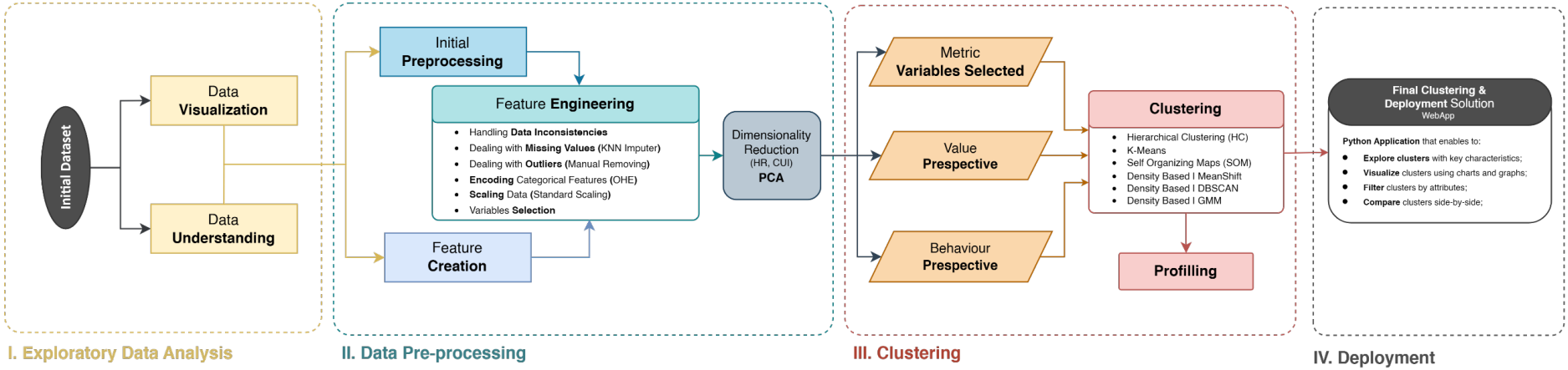
(Chronologically ordered from the most recent article to the oldest)

Paper Title	Reference	Abstract Summary	Methodology	Main Findings	PCA Usage	Algorithms Used	Number of Clusters
Comparative Study of Unsupervised Learning Algorithms for Customer Segmentation	[4]	This paper presents a comparative analysis of various unsupervised learning algorithms for customer segmentation from online platforms, aiming to identify the most effective method for grouping customers based on their characteristics.	- Data preprocessing (standardization) on the customer dataset. - Using PCA. - Applying multiple unsupervised learning algorithms. - Evaluating and comparing the performance of each algorithm using relevant metrics.	- K-means algorithm showed better performance for customer segmentation.	Yes	K-Means, Hierarchical Clustering, MeanShift	5 (K-Means)
Customer Segmentation: Transformation from Data to Marketing Strategy	[5]	Presents a customer segmentation approach using clustering algorithms and the RFM model.	- Collecting a customer dataset - Cleaning the dataset - Implementing the proposed methodology on a real customer dataset of P.T. Jamkrindo - Performing RFM model - Comparing k-Means and DBSCAN algorithms, with Silhouette scores and Davies Bouldin Indices	- DBSCAN outperformed k-Means. - The choice between k-Means and DBSCAN depends on the data's characteristics and goals of analysis. - Results can be used to develop tailored strategies, improving customer engagement, loyalty, and revenue.	No	DBSCAN, K-Means	5 (DBSCAN)
Customer Segmentation in User Behavior Analysis: A Comparative Study of Clustering Algorithms	[6]	Compares the performance of different clustering algorithms using demographic and behavioural data.	- Utilized three clustering algorithms: K-means, hierarchical clustering, and DBSCAN - Analysed age and spending score- Performed K-means, hierarchical clustering and DBSCAN, and visualized the results.	- DBSCAN outperformed K-means and hierarchical clustering. - DBSCAN can automatically determine number and shape of clusters. - DBSCAN's performance may be constrained for high-dimensional data.	No	K-means, Hierarchical Clustering, DBSCAN	7 (DBSCAN)

Paper Title	Reference	Abstract Summary	Methodology	Main Findings	PCA Usage	Algorithms Used	Number of Clusters
Unleashing Customer Insights: Segmentation Through Machine Learning	[7]	This paper explores the application of machine learning techniques, specifically K-means clustering, for customer segmentation then using a decision tree based on the clusters for future data.	- Applying K-means clustering, Agglomerative, Spectral, DBSCAN and analysing the silhouette score. - Training a decision tree with the best model according to the score for future data inserted.	- K Means Clustering effectively segments customers into distinct. - the results can be used to develop targeted Marketing strategies. - K-Means outperformed the other three clustering algorithms by a lot.	No	K-means, Agglomerative, Spectral, DBSCAN, Decision Trees	4 (K-Means)
Advance Customer Segmentation	[8]	Analyses three clustering algorithms to identify target customer segments based on demographic, psychographic, and behavioural factors.	- Used three clustering algorithms: K-means, Agglomerative, and Mean shift - Segmented customers based on characteristics such as age, income level, gender, as well as demographic, psychographic, and behavioural factors.	- Three prominent algorithms were analysed (K-means, Agglomerative and Mean shift) to identify target audience segments for customer segmentation. - The goal of the segmentation was to identify potential customers who are likely to buy the product.	No	K-means, Agglomerative, Mean shift	3
Cluster Analysis for Customer Segmentation with Open Banking Data	[9]	Demonstrates K-Means and DBSCAN clustering to segment customers based on RFM data.	- Used an RFM model to characterize customer behaviour - Employed two clustering techniques: K-Means and DBSCAN - Analysed data from Open Banking datasets.	- K-Means clustering identified three distinct segments based on RFM characteristics. - The study was able to identify both valuable and vulnerable customer segments. - K-Means clustering outperformed DBSCAN.	No	K-means, DBSCAN	3 (K-Means)
Incorporating K-means, Hierarchical Clustering and PCA in Customer Segmentation	[10]	Demonstrates the use of K-means, hierarchical clustering, and PCA for customer segmentation.	- K-means clustering-Hierarchical clustering without PCA - Principal component analysis (PCA) for dimensionality reduction - Using PCA as a validation tool for the K-means and hierarchical clustering results.	- PCA can be effectively used as a validation tool. - The study applied K-Means, Hierarchical clustering, and PCA to segment customers based on their credit card transaction history and it proved more suitable.	Yes	K-means, Hierarchical Clustering, PCA	4 (K-Means)

Paper Title	Reference	Abstract Summary	Methodology	Main Findings	PCA Usage	Algorithms Used	Number of Clusters
Determining customer segmentation and behaviour models with database marketing and machine learning	[11]	Demonstrates customer segmentation using machine learning on a large pizza delivery dataset.	- Used a dataset of 24M order records - Performed customer segmentation using clustering algorithms - Conducted clustering and multiple regression analyses.	- DBSCAN was not suitable for the dataset. - K-Means and Gaussian Mixture performed better. - Multiple regression was used to identify clusters with similar behaviour and discover valuable sub-clusters.	Yes	K-means, Gaussian Mixture, DBSCAN	4 (K-Means)
Business Clients' Segmentation Based on Activity - a Banking Approach	[12]	This study focuses on segmenting business clients of a bank based on their transactional activity, using clustering algorithms to identify distinct groups and provide insights for targeted banking strategies.	- Data preprocessing and feature engineering using business client transaction data. - Implementing K-means clustering algorithms. - Create *personas* with identifiable names from the cluster analysis	- K-means initialized with random seeds was used to segment the data due to its better performance. - The results offer valuable insights for the bank to tailor its services and strategies.	Yes	K-means, PAM	4
Comparative Performance of Using PCA With K-Means And Fuzzy C Means Clustering For Customer Segmentation	[13]	This study compares the performance of K-means and Fuzzy C-means clustering algorithms for customer segmentation, with and without the application of Principal Component Analysis (PCA) for dimensionality reduction.	- Applying PCA for dimensionality reduction on the customer dataset. - Implementing K-means and Fuzzy C-means clustering on both original and PCA-transformed data. - Comparing the performance of the clustering algorithms using appropriate metrics.	- PCA improves the performance of both K-means and Fuzzy C-means clustering algorithms. - Using PCA can lead to more efficient and accurate customer segmentation.	Yes	K-means, Fuzzy C-Means, PCA	3

APPENDIX B. PROJECT FLOWCHART



APPENDIX C. SUMMARY OF PART 1 (EDA)

This section summarizes the Exploratory Data Analysis (EDA) conducted on **Part 1**.

Data Cleaning and Preprocessing:

- **Handling Missing Data:** Missing values were identified in the *customer_age*, *first_order*, and *HR_0* columns.
- **Duplicate Removal:** Thirteen (13) duplicate rows were detected and removed.
- **Data Integrity Enhancements:**
  - Missing region values were consolidated to *Unknown*.
  - '-' was reinterpreted as *NO PROMO* within the *last_promo* variable.
  - One hundred thirty-eight (138) rows exhibiting a zero *order_count* were removed.
  - Inconsistencies between *vendor_count* and *product_count* as well as between *is_chain* and *product_count* were flagged.

Feature Engineering | New Attributes

Several new attributes were engineered to better understand customer behaviour and patterns within the data. Key derived features include:

Table C1 - Feature Engineering - New features created.

	Column Name	Description
+1	*order_count*	The total number of orders placed by a customer.
+2	*customer_region_buckets*	Simplified version of customer regions.
+3	*customer_age_group*	Customer ages grouped into meaningful segments (15-28, 29-41, 42-54, 55-67, 68-80)
+4	*days_between_orders*	The time difference between a customer's first and last order.
+5	*days_between_orders_per_order*	The average time between consecutive orders made by a customer.
+6	*last_promo_bin*	A binary flag indicating if the last order included a promotion.
+7	*CUI_Total_Amount_Spent*	The total amount spent on all food cuisines by the customer.
+8	*CUI_Most_Spent_Cuisine*	Identifies the food cuisine a customer spent the most amount of money.
+9	*CUI_Total_Food_Types*	The number of different food cuisines a customer ordered.
+10	*CUI_Avg_Amount_Spent*	The average amount spent per order by a customer.

Exploratory Data Analysis (EDA):

1. **Data Distribution**: The distributions of various numerical variables were examined using histograms, Kernel Density Estimations (KDE), and box plots. These visualizations provided insights into data symmetry, modality, and outlier presence.
2. **Outlier Detection:** Outliers were identified in the cuisine expenditure variables and removed from boxplot visualizations to improve readability and focus on central tendencies.
3. **Temporal Patterns:** The distribution of orders by the day of the week and hour of the day was analysed to reveal peak ordering times and days for each customer.

Correlation Analysis:

1. **Numerical Correlation:** The Pearson correlation was utilized to evaluate linear relationships among numeric variables and the numerical variables in relation to the target variables.
2. **Categorical Correlation:** Cramer's V correlation was calculated to evaluate associations between categorical variables.
3. **Categorical-Numerical Correlation:** The ETA squared was calculated to evaluate associations between categorical variables and numerical variables.

Key Findings:

- Customer activity is more frequent on weekends than on weekdays.
- The 20s and 30s demographic is the most frequent customer base.
- Most customers tend to order from a limited number of vendors and cuisines.
- The number of *is_chain* restaurants tends to be higher than the number of vendors a client uses
- A high proportion of customers (52.6%) do not use promotions.

APPENDIX D. PREPROCESSING

Table D1 - Filling the missing values.

	Method Used	Thought Process	Advantages	Disadvantages
first_order	Deterministic	When analysing the cases with missing *first_order*, it was observed that *last_order* was always zero. Since *first_order* $\leq$ *last_order*, it follows that *first_order* must also be zero.	This approach is highly reliable when applicable, preventing noise introduction and ensuring data consistency.	The only disadvantage arises if the *last_order* variable was erroneously obtained. However, the data is trusted unless proven incorrect.
$\boldsymbol{HR_0}$	Deterministic	In **Part 1**, we established that the total number of orders should equal the sum of DOW, which, in turn, should match the sum of HR. However, this was not the case, as the sum of DOW exceeded the sum of HR. Additionally, there was no case where the sum of a consumer's HR exceeded the sum of DOW, reinforcing the validity of the DOW values. HR_0 was imputed using the following equation: $$\mathrm{HR}_0 = \sum_{i=0}^{6} \mathrm{DOW}_i - \sum_{i=1}^{23} \mathrm{HR}_i$$ Where: • HR_i is the number of orders made by the customer at hour i. (0 = midnight, 23 = 11 PM) • DOW_i is the number of orders made by the customer at day of the week i (0 = Sunday, 6 = Saturday)	This approach is highly reliable when applicable, preventing noise introduction and ensuring data consistency.	The disadvantage arises if the DOW variables contain errors, which would propagate into the imputed HR_0 values. However, assessment of inconsistencies found no issues, therefore, the data is trusted unless proven otherwise.
customer_age	*KNNImputer*	Data was standardized to ensure variables are equally weighted when calculating distances. KNN was then used with 5 nearest neighbours and uniform weights to impute missing values using only metric features, followed by adjusting the imputed value to the nearest integer for the specific variable and reverting the scaling to handle outliers.	Preserves the data distribution (**Figure D1**) and avoids biases commonly introduced by imputing with the mean or median.	Imputation may yield different results if performed after removing outliers, which is common when outliers are considered as errors, however, we don't believe this is required given our analysis.

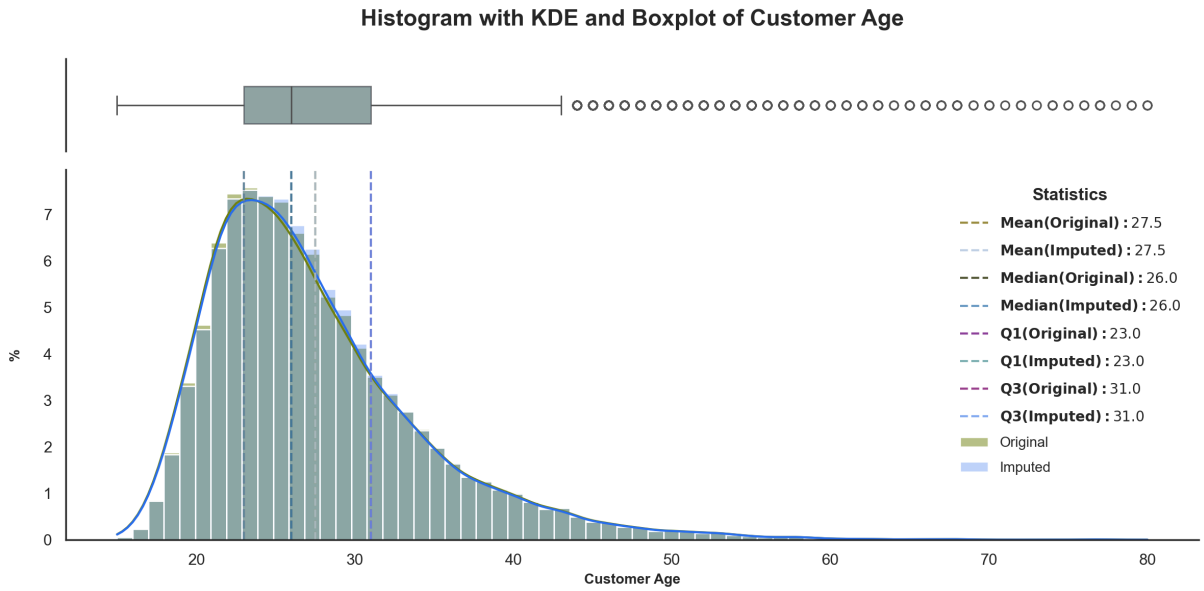


Figure D1 - Histogram with KDE and boxplot for Costumer Age.

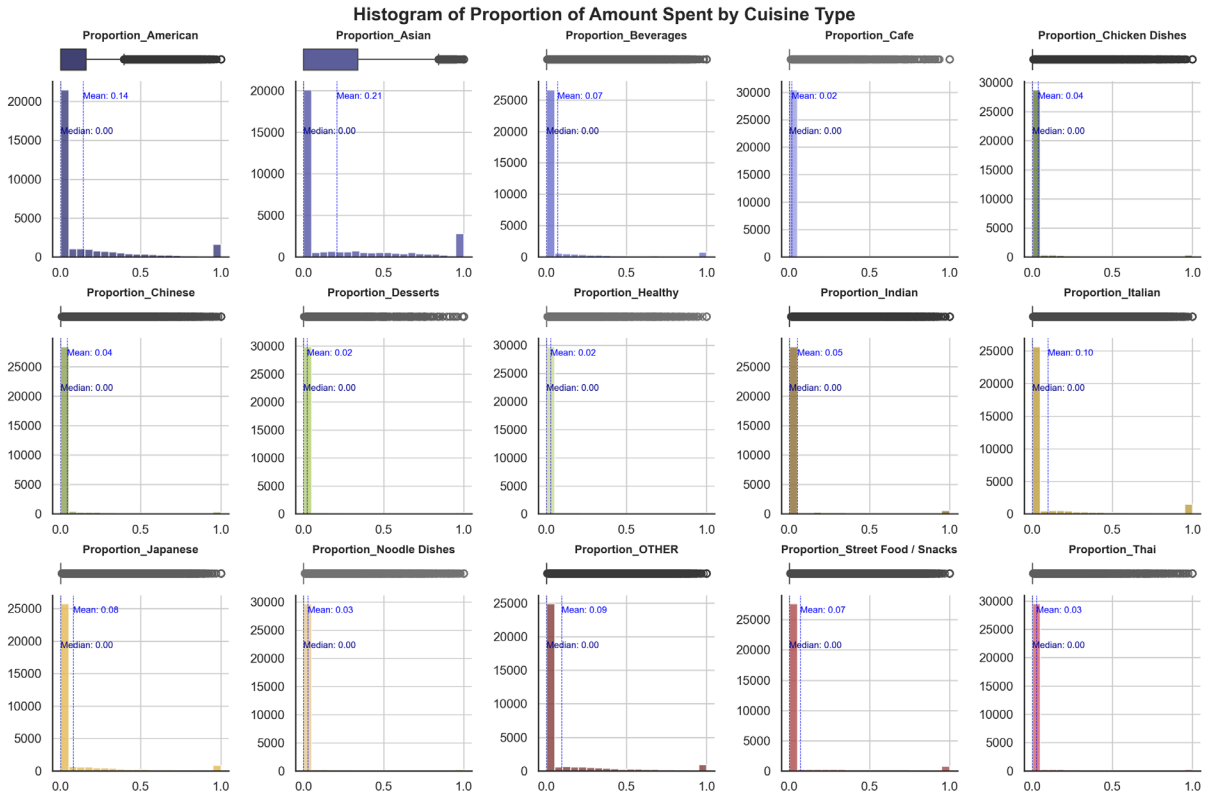


Figure D2 – Histogram of Proportion of Amount Spent by Cuisine Type.

Table D2 - Thresholds for Manual Outlier Removal.

Variable	Threshold
customer_age	≤ 65
product_count	≤ 100
vendor_count	≤ 30
chain_count	≤ 50
order_count	≤ 60
CUI_American	≤ 100
CUI_Asian	≤ 250
CUI_Beverages	≤ 100
CUI_Cafe	≤ 100
CUI_Chicken Dishes	≤ 50
CUI_Chinese	≤ 175
CUI_Desserts	≤ 75
CUI_Healthy	≤ 75
CUI_Indian	≤ 75
CUI_Italian	≤ 175
CUI_Japanese	≤ 150
CUI_Noodle Dishes	≤ 75
CUI_OTHER	≤ 125
CUI_Street Food / Snacks	≤ 175
CUI_Thai	≤ 60
DOW_0	≤ 10
DOW_1	≤ 10
DOW_2	≤ 10
DOW_3	≤ 15
DOW_4	≤ 15
DOW_5	≤ 15
DOW_6	≤ 15

Variable	Threshold
HR_0	≤ 10
HR_1	≤ 10
HR_2	≤ 10
HR_3	≤ 10
HR_4	≤ 10
HR_5	≤ 6
HR_6	≤ 10
HR_7	≤ 10
HR_8	≤ 17
HR_9	≤ 10
HR_10	≤ 15
HR_11	≤ 15
HR_12	≤ 15
HR_13	≤ 10
HR_14	≤ 10
HR_15	≤ 10
HR_16	≤ 10
HR_17	≤ 15
HR_18	≤ 15
HR_19	≤ 15
HR_20	≤ 10
HR_21	≤ 7
HR_22	≤ 10
HR_23	≤ 7

Table D3 – IQR Analysis after outliers' removal.

Feature	1st Quartile	3rd Quartile	IQR	Lower Bound	Upper Bound	Number of Outliers	Percentage of Outliers (%)	Min	Max
customer_age	23	31	8	11	43	1020	3,26	44	65
vendor_count	1	4	3	-3,5	8,5	1344	4,3	9	30
product_count	2	6	4	-4	12	2731	8,73	13	61
chain_count	1	3	2	-2	6	2891	9,24	7	43
first_order	7	45	38	-50	102	0	0	-	-
last_order	49	83	34	-2	134	0	0	-	-
CUI_American	0	5,54	5,54	-8,31	13,85	3417	10,92	13,86	99,33
CUI_Asian	0	11,79	11,79	-17,685	29,475	3349	10,71	29,49	244,59
CUI_Beverages	0	0	0	0	0	5314	16,99	0,32	98,95
CUI_Cafe	0	0	0	0	0	1285	4,11	0,66	99,52
CUI_Chicken Dishes	0	0	0	0	0	3183	10,18	0,34	79,07
CUI_Chinese	0	0	0	0	0	3442	11	0,44	160,97
CUI_Desserts	0	0	0	0	0	1961	6,27	0,51	85,42
CUI_Healthy	0	0	0	0	0	2079	6,65	0,43	73,7
CUI_Indian	0	0	0	0	0	3328	10,64	0,46	87,82
CUI_Italian	0	0	0	0	0	6290	20,11	0,34	167,12
CUI_Japanese	0	0	0	0	0	6153	19,67	0,43	128,24
CUI_Noodle Dishes	0	0	0	0	0	2174	6,95	0,37	82,07
CUI_OTHER	0	0	0	0	0	6861	21,93	0,36	121,43
CUI_Street Food / Snacks	0	0	0	0	0	4132	13,21	0,44	172,71
CUI_Thai	0	0	0	0	0	2293	7,33	0,5	73,39

DOW_0	0	1	1	-1,5	2,5	1207	3,86	3	10
DOW_1	0	1	1	-1,5	2,5	1272	4,07	3	10
DOW_2	0	1	1	-1,5	2,5	1361	4,35	3	10
DOW_3	0	1	1	-1,5	2,5	1417	4,53	3	11
DOW_4	0	1	1	-1,5	2,5	1549	4,95	3	10
DOW_5	0	1	1	-1,5	2,5	1530	4,89	3	14
DOW_6	0	1	1	-1,5	2,5	1772	5,67	3	13
HR_0	0	0	0	0	0	1118	3,57	1	9
HR_1	0	0	0	0	0	1234	3,95	1	8
HR_2	0	0	0	0	0	1443	4,61	1	10
HR_3	0	0	0	0	0	2450	7,83	1	10
HR_4	0	0	0	0	0	2300	7,35	1	8
HR_5	0	0	0	0	0	1955	6,25	1	5
HR_6	0	0	0	0	0	1676	5,36	1	8
HR_7	0	0	0	0	0	1804	5,77	1	7
HR_8	0	0	0	0	0	2796	8,94	1	13
HR_9	0	0	0	0	0	4711	15,06	1	10
HR_10	0	0	0	0	0	6345	20,29	1	15
HR_11	0	0	0	0	0	7283	23,28	1	14
HR_12	0	0	0	0	0	6296	20,13	1	14
HR_13	0	0	0	0	0	5262	16,82	1	10
HR_14	0	0	0	0	0	4910	15,7	1	10
HR_15	0	0	0	0	0	5822	18,61	1	9
HR_16	0	0	0	0	0	6910	22,09	1	10
HR_17	0	0	0	0	0	7392	23,63	1	15

HR_18	0	0	0	0	0	6515	20,83	1	15
HR_19	0	0	0	0	0	4758	15,21	1	13
HR_20	0	0	0	0	0	2980	9,53	1	10
HR_21	0	0	0	0	0	1697	5,43	1	6
HR_22	0	0	0	0	0	1132	3,62	1	7
HR_23	0	0	0	0	0	1079	3,45	1	6
order_count	2	5	3	-2,5	9,5	2692	8,61	10	55
days_between_orders	3	61	58	-84	148	0	0	-	-
days_between_orders_per_order	1,2	12,8	11,6	-16,2	30,2	583	1,86	30,5	44,5
CUI_Total_Amount_Spent	12,9	43,81	30,91	-33,465	90,175	2407	7,7	90,18	471,86
CUI_Total_Food_Types	1	3	2	-2	6	594	1,9	7	12
CUI_Avg_Amount_Spent	5,046833	12,86	7,813167	-6,67292	24,57975	1605	5,13	24,58	104,32


Numeric Variables' Histograms with Boxplots
customer_age
vendor_count
product_count
chain_count
first_order
last_order
CUI_American
CUI_Asian
CUI_Beverages
CUI_Cafe
CUI_Chicken Dishes
CUI_Chinese
CUI_Desserts
CUI_Healthy
CUI_Indian
CUI_Italian
CUI_Japanese
CUI_Noodle Dishes
CUI_OTHER
CUI_Street Food / Snacks
CUI_Thai
DOW_0
DOW_1
DOW_2
DOW_3
DOW_4
DOW_5
DOW_6
HR_0
HR_1
HR_2
HR_3
HR_4
HR_5
HR_6
HR_7
HR_8
HR_9
HR_10
HR_11
HR_12
HR_13
HR_14
HR_15
HR_16
HR_17
HR_18
HR_19
HR_20
HR_21
HR_22
HR_23
order_count
days_between_orders
days_between_orders_per_order
CUI_Total_Amount_Spent
CUI_Total_Food_Types
CUI_Avg_Amount_Spent

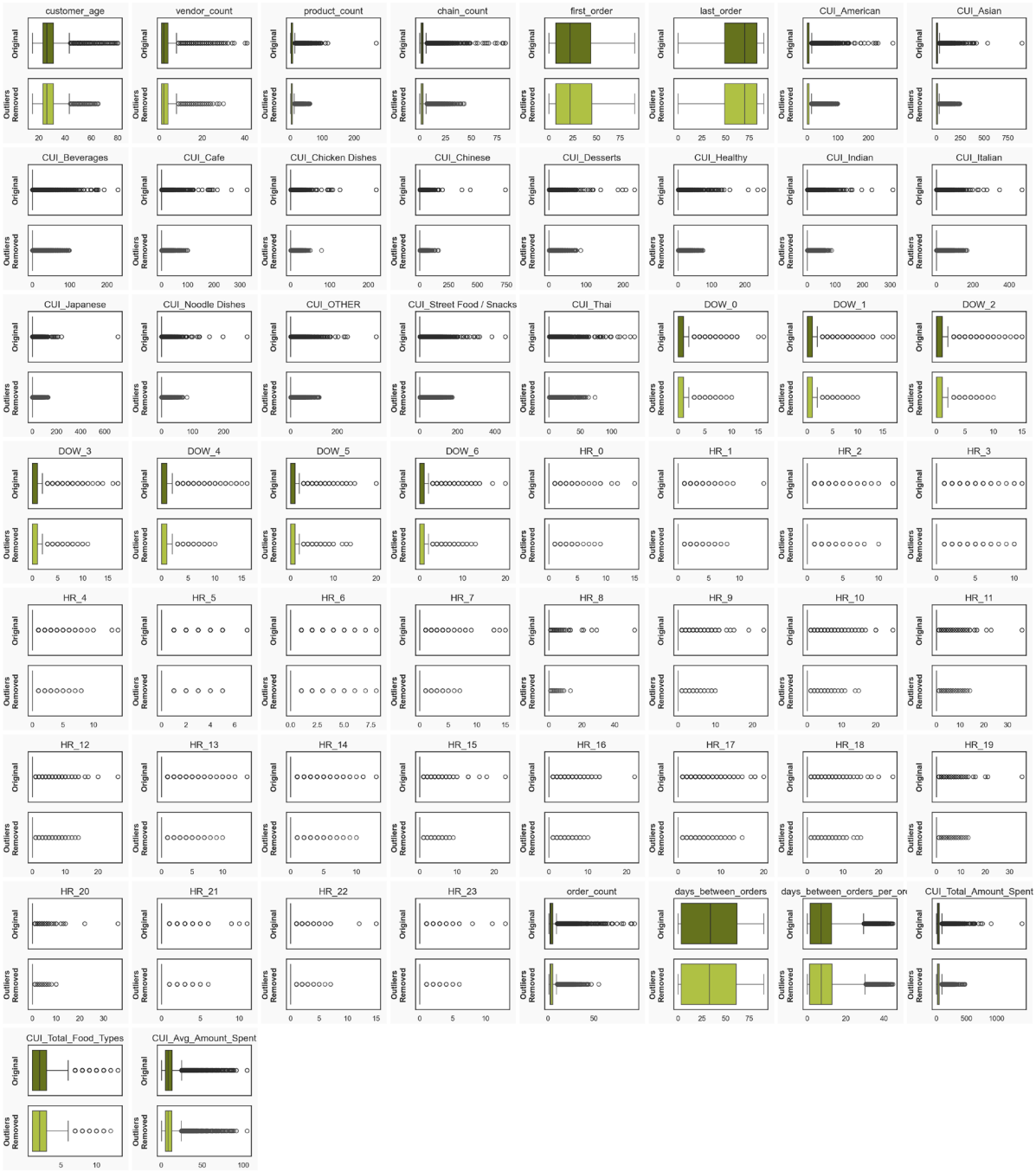


Figure D3 – Comparative Histograms & Boxplots before and after outlier removal.

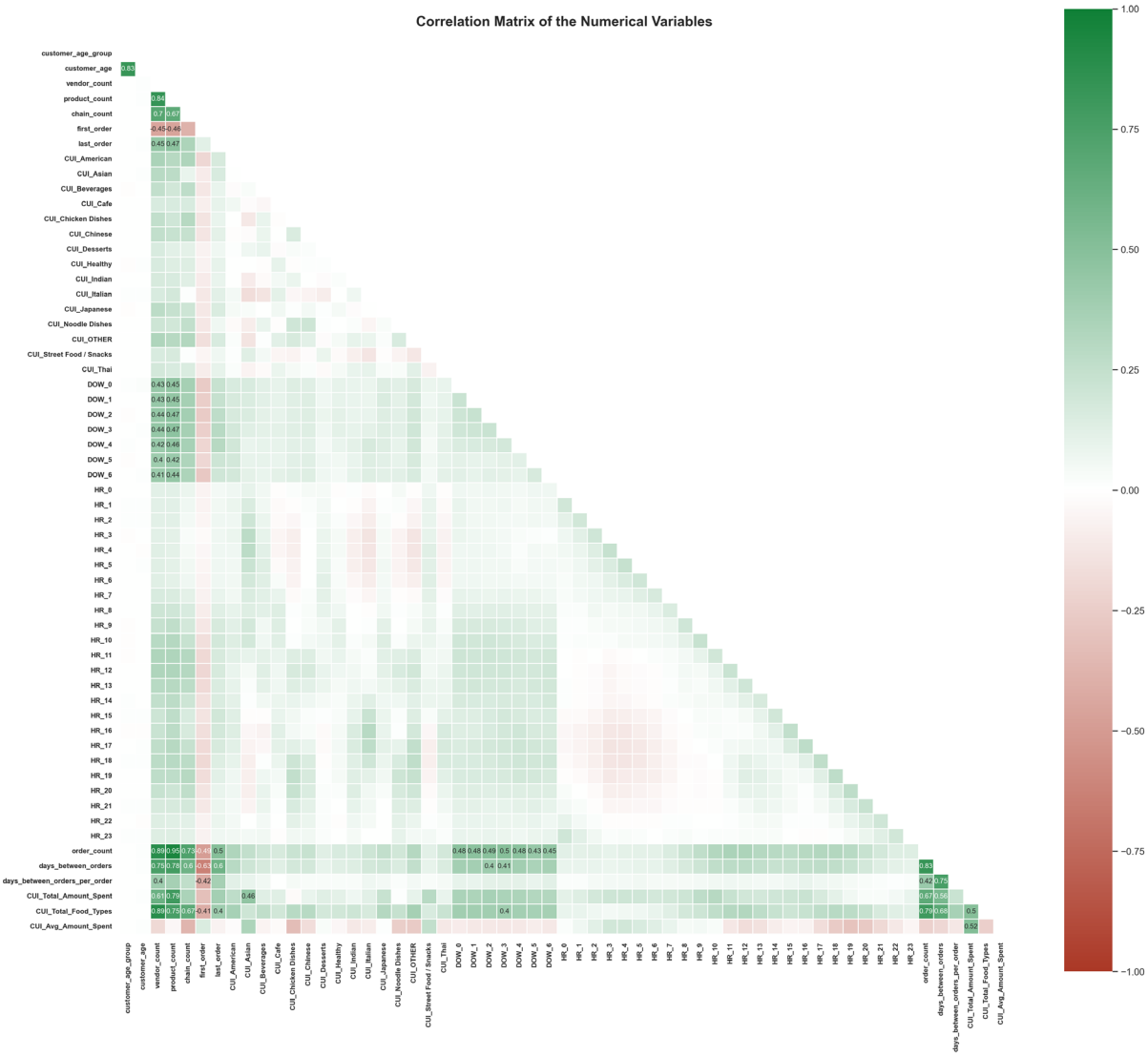


Figure D4 – Spearman's Correlation Matrix after preprocessing (Numerical Variables).

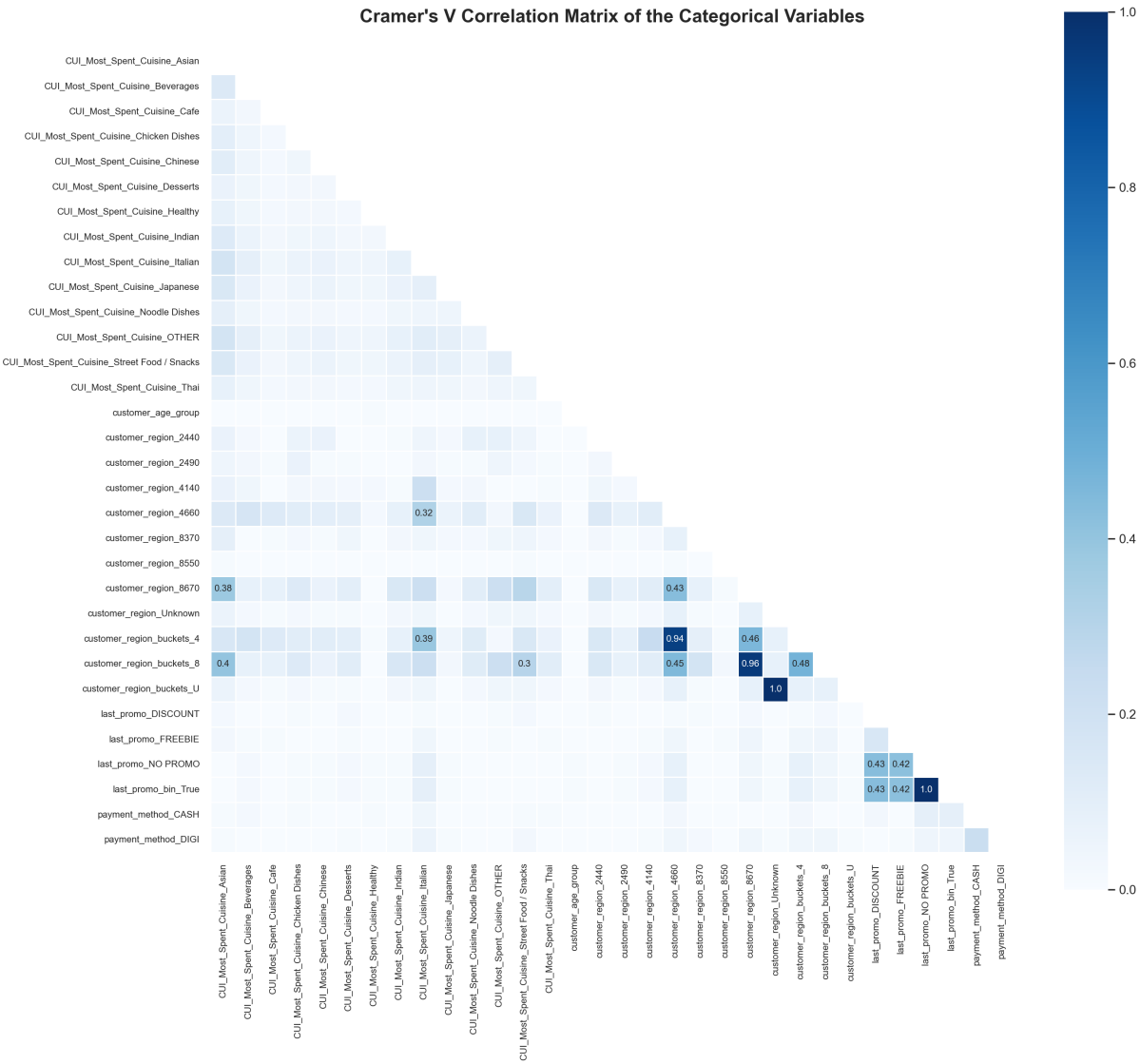


Figure D5 – Cramer's V Correlation Matrix after preprocessing (Categorical Variables).

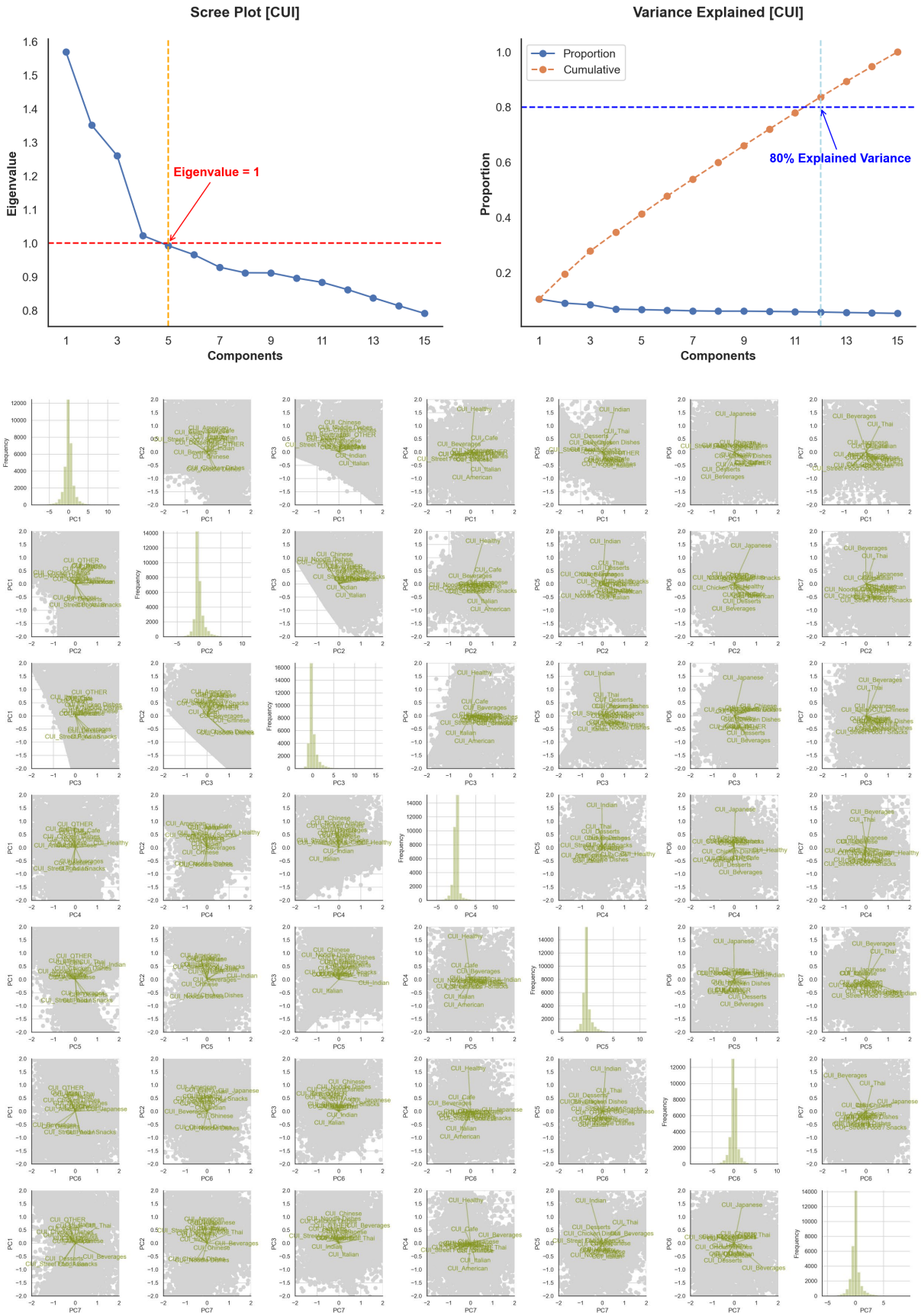


Figure D6 – Principal Component Analysis (PCA) for Cuisine *(CUI)* variables.

Table D4 – Loadings for Cuisine (*CUI*)

	PC0	PC1	PC2	PC3	PC4	PC5	PC6
CUI_American	0.037358	0.486789	0.089294	-0.458737	-0.144423	-0.223790	-0.059434
CUI_Asian	-0.474456	0.396779	0.239199	-0.045912	0.039430	0.108679	-0.046109
CUI_Beverages	-0.309726	-0.012336	0.340090	0.136858	0.161299	-0.435868	0.602341
CUI_Cafe	0.365371	0.432628	0.097943	0.245110	-0.130247	-0.186988	-0.126043
CUI_Chicken Dishes	0.236590	-0.333722	0.425155	-0.081554	0.161871	-0.075017	-0.213252
CUI_Chinese	0.097057	-0.104442	0.577726	-0.142222	-0.158806	0.161726	0.094703
CUI_Desserts	-0.349090	0.188227	0.276264	-0.043839	0.274494	-0.306928	-0.230296
CUI_Healthy	0.096361	0.302301	0.093443	0.760432	-0.118760	-0.050729	-0.096568
CUI_Indian	0.318784	0.067626	-0.077427	-0.012220	0.741159	0.067919	-0.241717
CUI_Italian	0.402454	0.291183	-0.232868	-0.314533	-0.242591	-0.195398	0.095954
CUI_Japanese	0.040962	0.414687	0.211655	0.021351	-0.010138	0.656572	0.158594
CUI_Noodle Dishes	0.170890	-0.357952	0.478451	-0.024907	-0.216956	0.104420	-0.101669
CUI_OTHER	0.508497	0.149793	0.332883	-0.008190	-0.027629	-0.195868	-0.153438
CUI_Street Food / Snacks	-0.470914	0.250065	0.134290	-0.131425	0.034591	0.096844	-0.284412
CUI_Thai	0.376347	0.214898	0.086139	-0.082789	0.359309	0.134159	0.469077

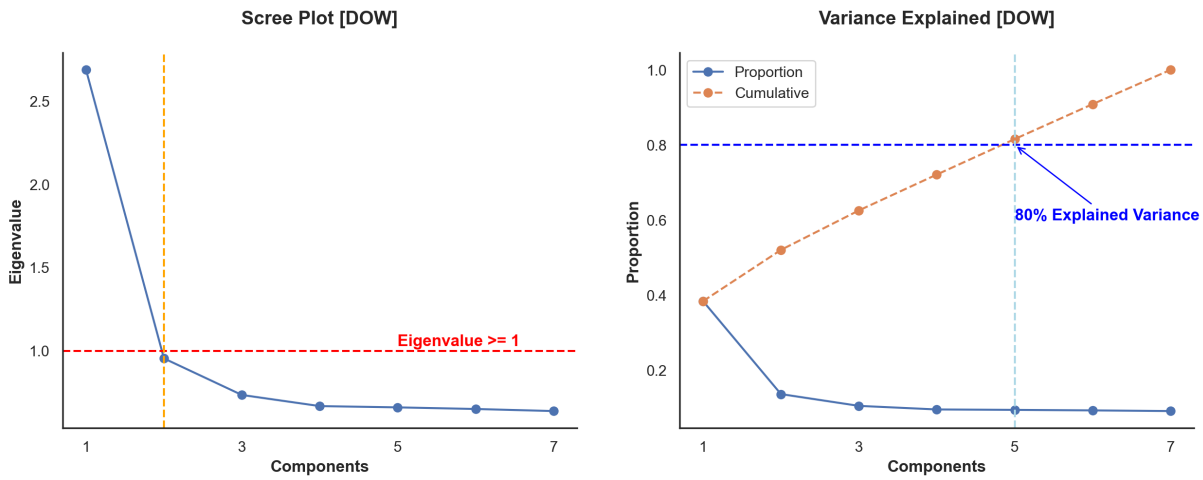


Figure D7 – Principal Component Analysis (PCA) for Day of the Week (*DOW*) variables.

Table D5 – Loadings for Day of the Week (*DOW*)

	PC0	PC1	PC2
Sunday	0.649737	-0.220289	-0.328973
Monday	0.653105	-0.294649	-0.147805
Tuesday	0.661376	-0.265825	-0.066398
Wednesday	0.656979	-0.204820	0.018828
Thursday	0.611662	-0.005684	0.749962
Friday	0.544797	0.603009	-0.018295
Saturday	0.547401	0.586222	-0.195349

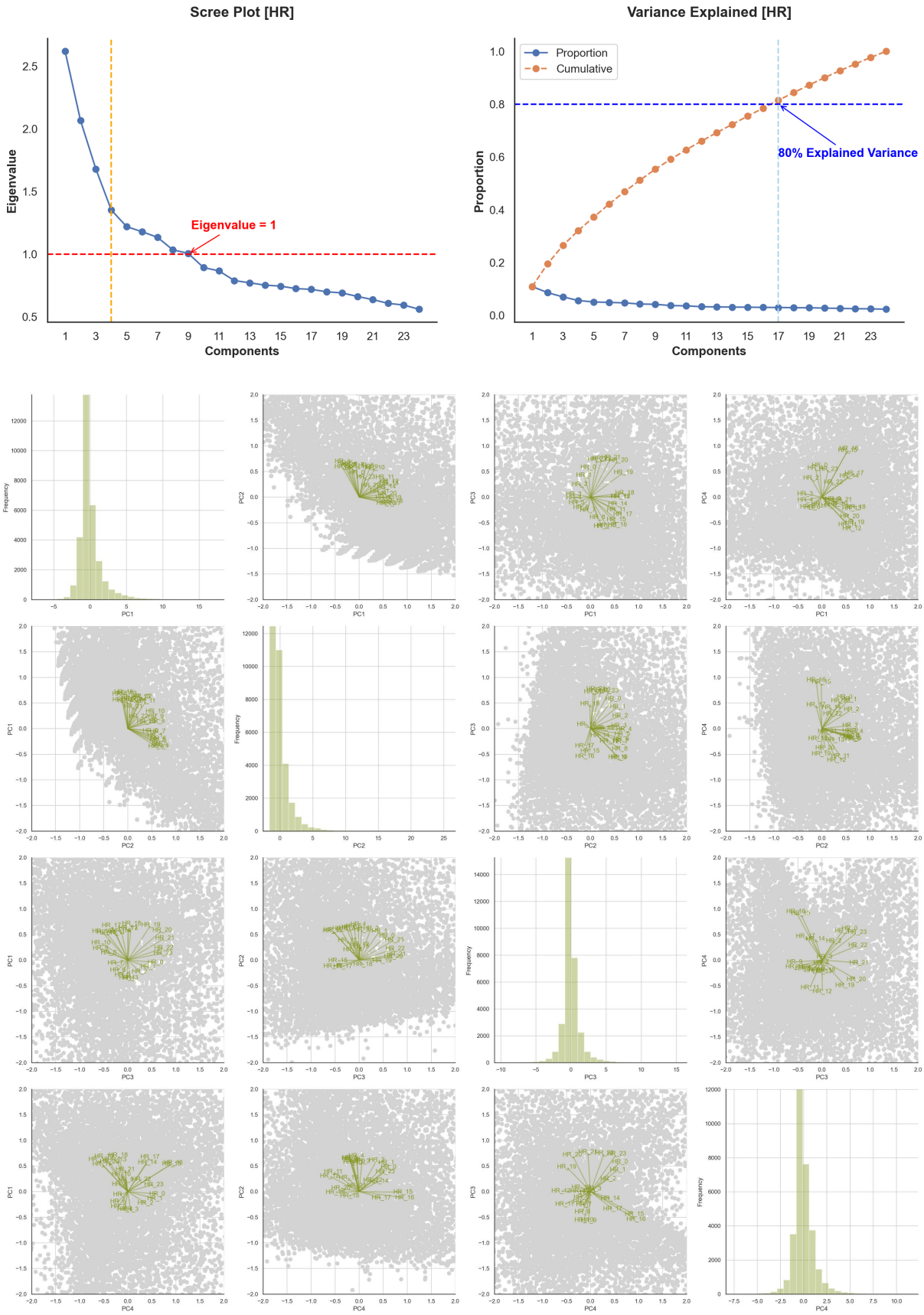


Figure D8 – Principal Component Analysis (PCA) for hours of the day (*HR*) variables.

Table D6 – Loadings for the hours of the day (*HR*)

	PC0	PC1	PC2	PC3
HR_0	-0.035134	0.315851	0.348847	0.332033
HR_1	-0.102967	0.377006	0.256574	0.306890
HR_2	-0.171052	0.411288	0.149463	0.199881
HR_3	-0.263096	0.394599	0.033426	0.032263
HR_4	-0.261347	0.465525	-0.010080	-0.031096
HR_5	-0.220443	0.439306	-0.065809	-0.103306
HR_6	-0.149799	0.427102	-0.123697	-0.109066
HR_7	-0.044864	0.416786	-0.155101	-0.075721
HR_8	0.107003	0.407955	-0.239983	-0.091562
HR_9	0.171035	0.390339	-0.342702	-0.021668
HR_10	0.251413	0.380541	-0.335542	-0.072264
HR_11	0.401666	0.256854	-0.149340	-0.292730
HR_12	0.470850	0.201493	0.005787	-0.333841
HR_13	0.465097	0.178736	0.005793	-0.121522
HR_14	0.426353	0.133211	-0.080977	0.221773
HR_15	0.404190	-0.008873	-0.266827	0.492110
HR_16	0.420828	-0.080132	-0.326443	0.512095
HR_17	0.505175	-0.083893	-0.205105	0.248941
HR_18	0.534609	-0.055812	0.047242	-0.105868
HR_19	0.514186	-0.009450	0.288447	-0.268659
HR_20	0.437118	0.042611	0.431194	-0.204901
HR_21	0.321567	0.075246	0.465560	-0.029143
HR_22	0.177675	0.147896	0.451647	0.154269
HR_23	0.101702	0.260560	0.442258	0.293133

Table D7 – Criterion for PC Retention and Final Decision.

Variable	Kaiser's Rule	Scree Plot	Cumulative % of Total Variance (80%)	Nr of PCs Kept
CUI	4	5	12	7
DOW	1	2	5	-
HR	9	4	17	4

Table D8 – PCA Components kept with descriptions.

PC Name	Description
CUI_NOTAsian_Italian_OTHER_NOTSnack_PC	Principal component representing a combination of non-Asian, non-Italian, non-snack cuisines.
CUI_American_Cafe_Japanese_PC	Principal component representing a combination of American, Cafe, and Japanese cuisines.
CUI_Chicken_Chinese_Noodle_PC	Principal component representing a combination of Chicken, Chinese, and Noodle cuisines.
CUI_Healthy_NOTAmerican_PC	Principal component representing a combination of healthy, non-American cuisines.
CUI_Indian_PC	Principal component representing Indian cuisine.
CUI_Japanese_NOTBeverages_PC	Principal component representing Japanese cuisine excluding beverages.
CUI_Beverages_Thai_PC	Principal component representing a combination of Beverages and Thai cuisines.
HR_Lunch_Dinner_PC	Principal component representing orders placed during lunch and dinner hours.
HR_LateNight_Breakfast_PC	Principal component representing orders placed during late-night and breakfast hours.
HR_Evening_PC	Principal component representing orders placed during evening hours.
HR_AfternoonSnack_PC	Principal component representing orders placed during afternoon snack hours.

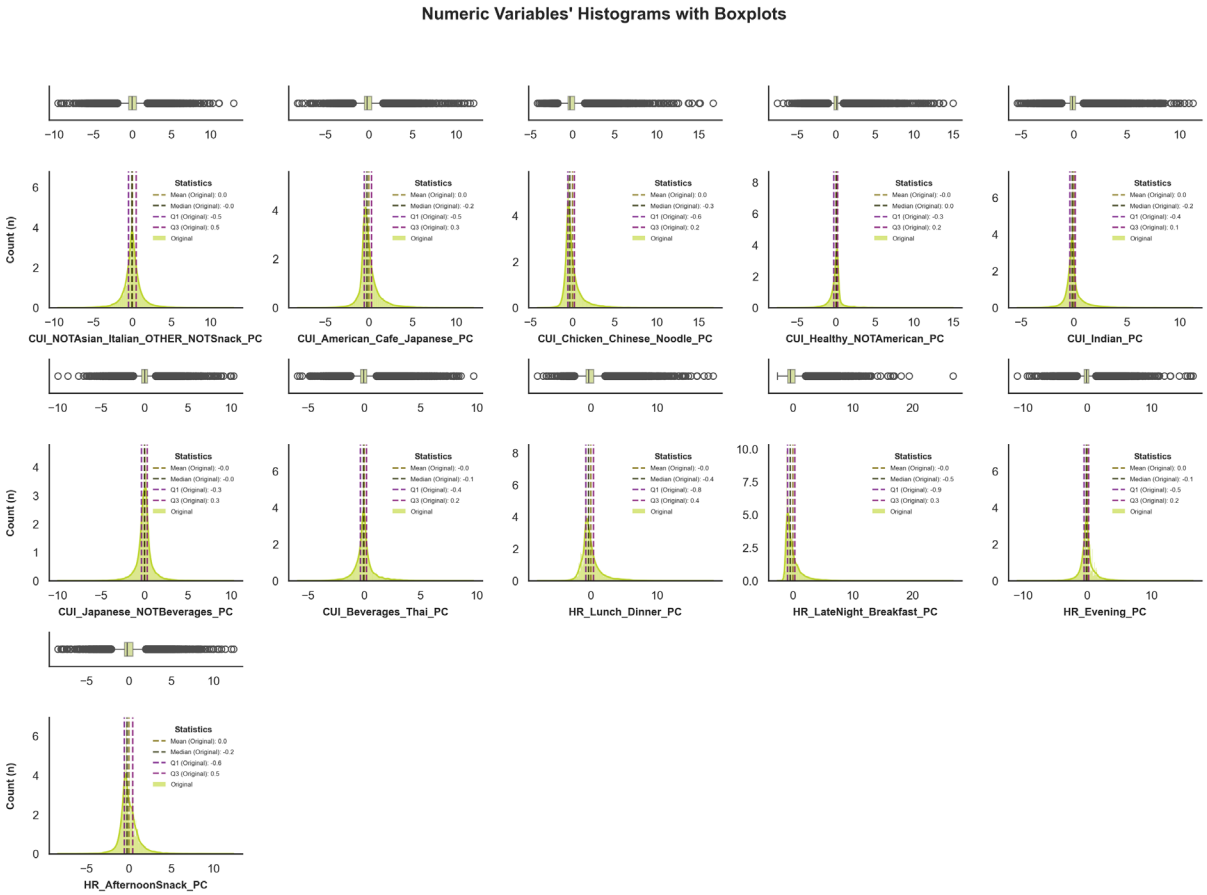


Figure D9 – Histograms & Boxplots of the PC variables created.

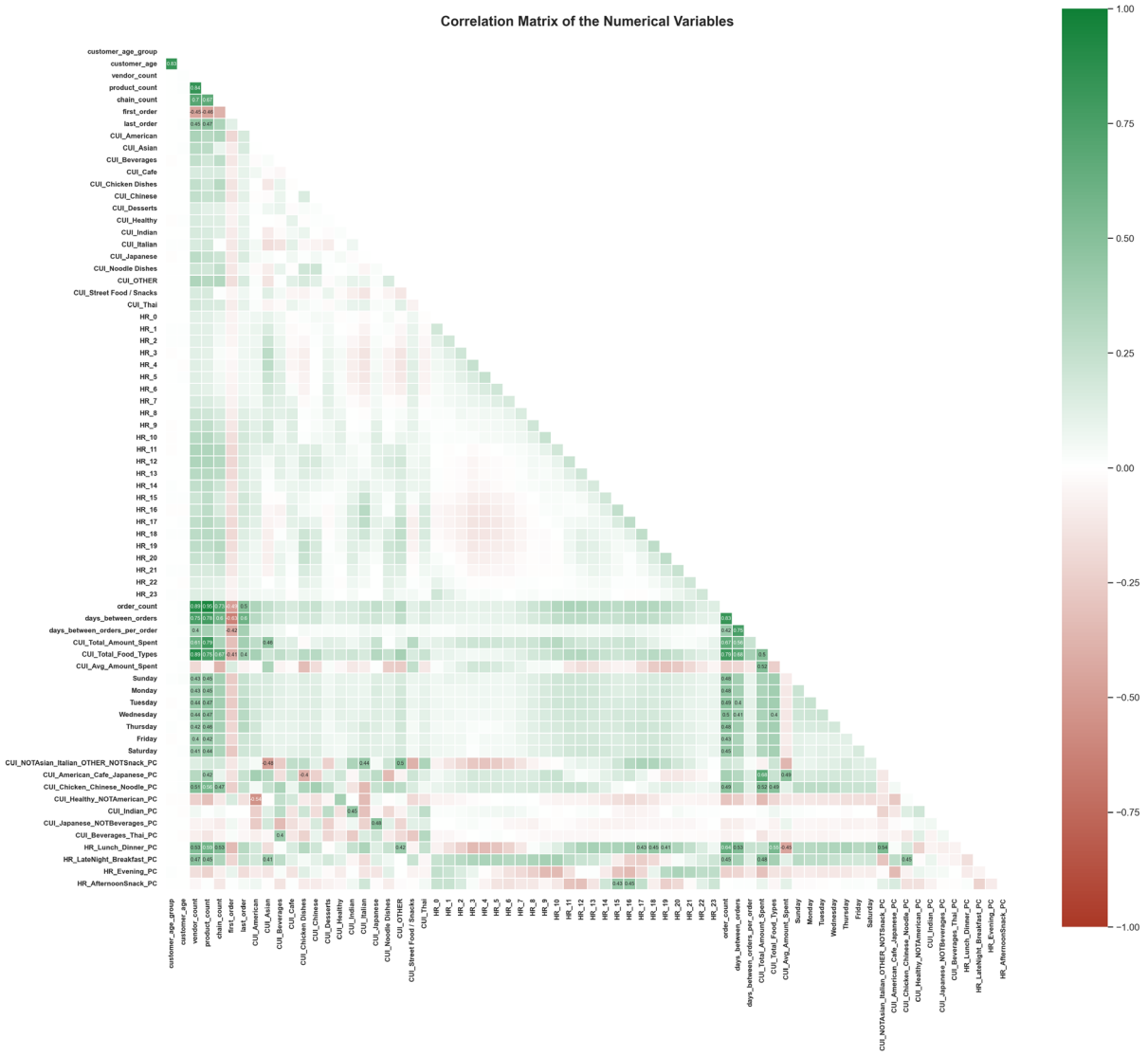


Figure D10 – Correlation Matrix of the principal component's variables created with the numerical variables.

APPENDIX E. CLUSTERING

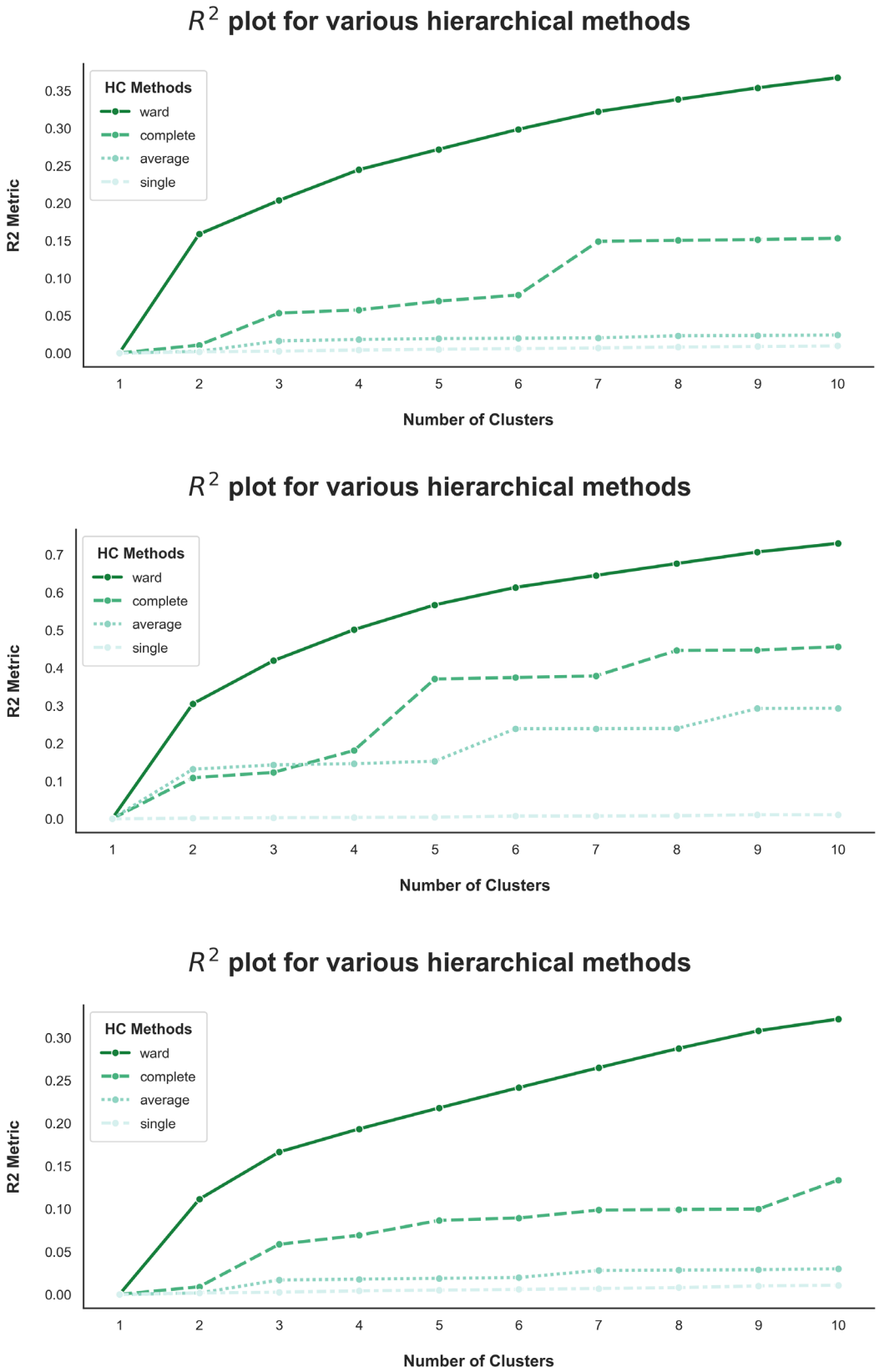


Figure E1 - R^2 plots for various hierarchical clustering methods.

(**Top:** Overall, **Middle:** Value perspective, **Bottom:** Behaviour perspective)

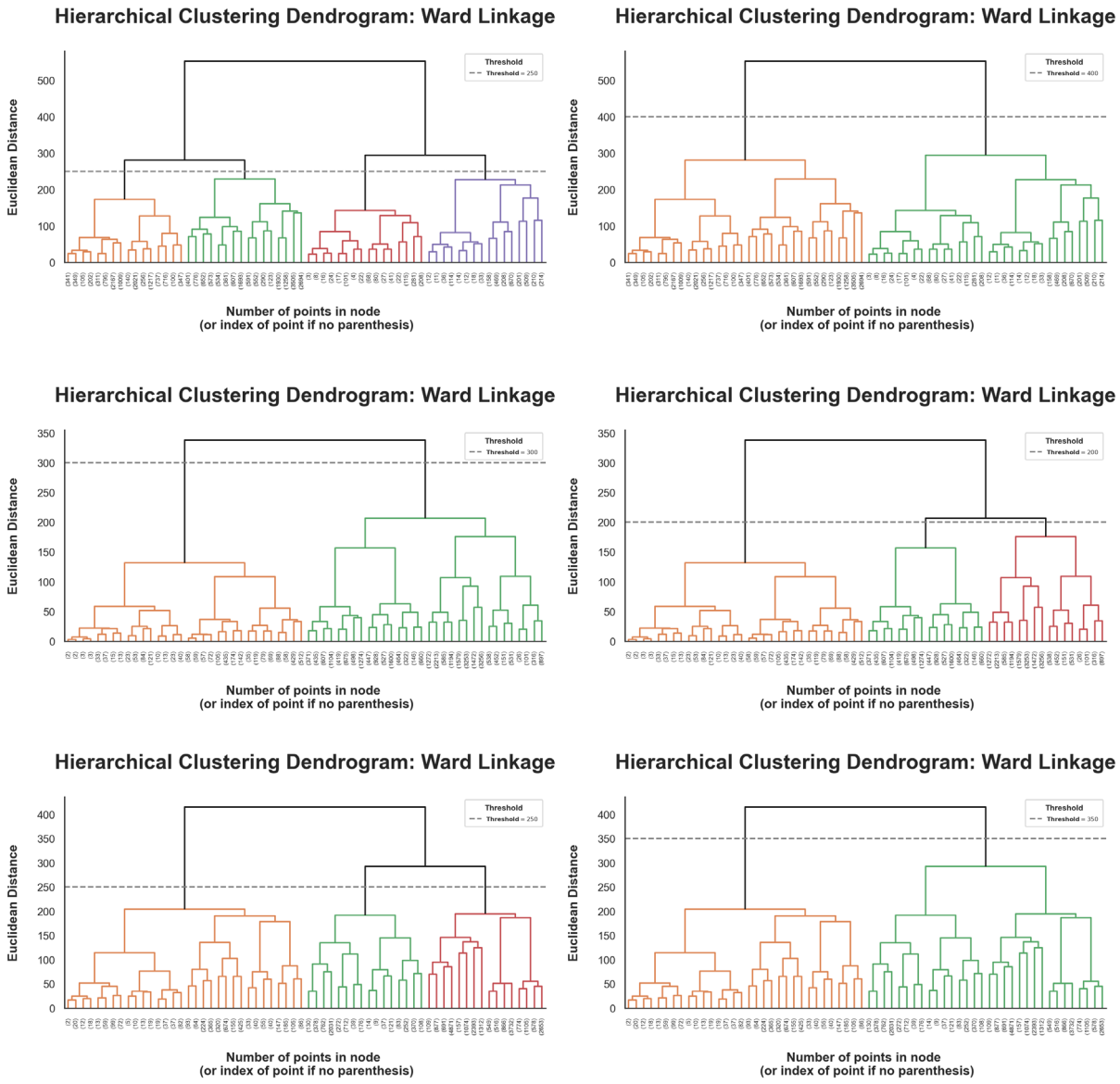


Figure E2 - Hierarchical clustering dendrograms using Ward linkage.

(**Top:** Overall, **Middle:** Value perspective, **Bottom:** Behaviour perspective)

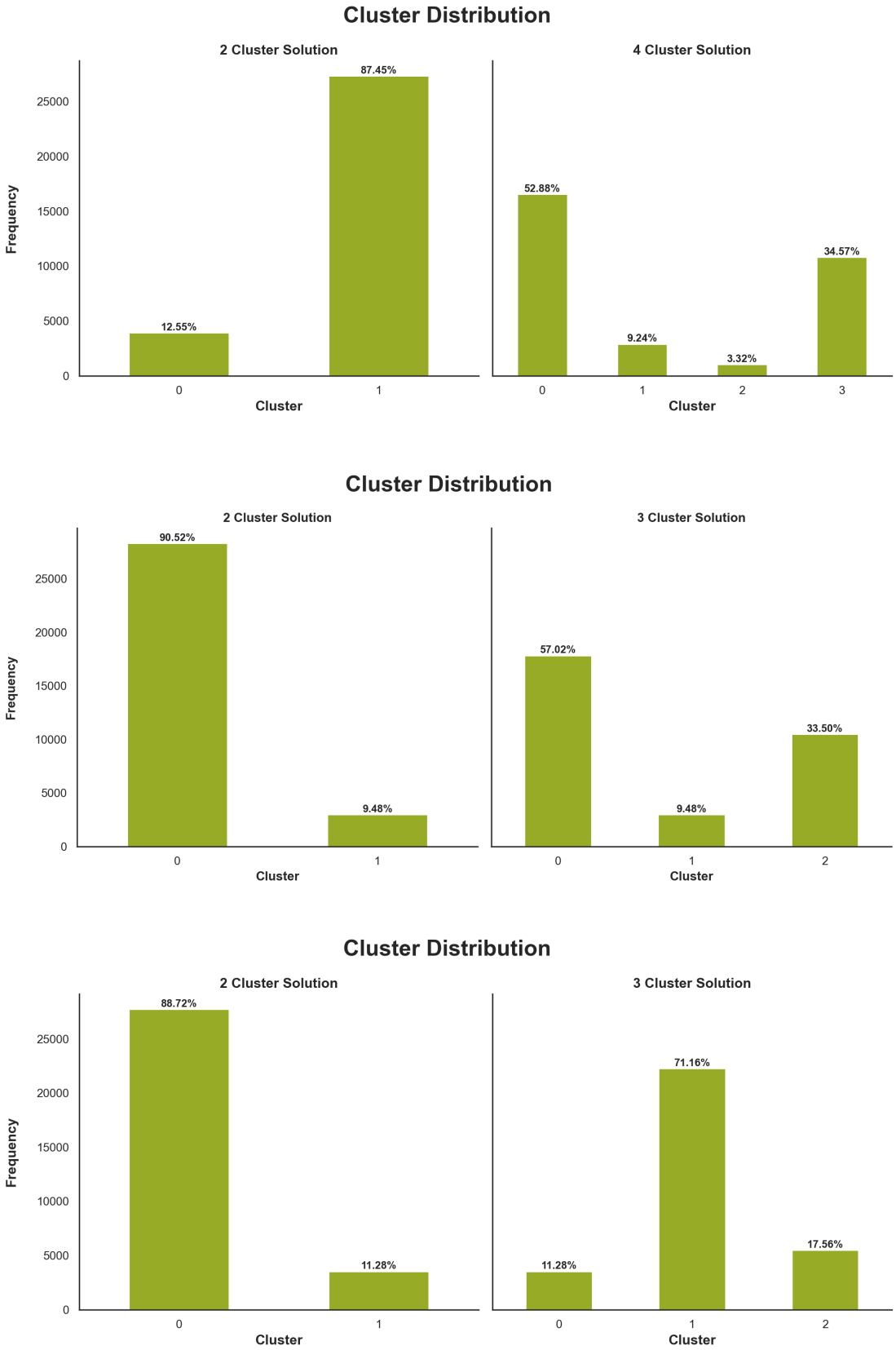


Figure E3 – Barplot with cluster distribution for each *n_cluster*.

(**Top:** Overall, **Middle:** Value perspective, **Bottom:** Behaviour perspective)

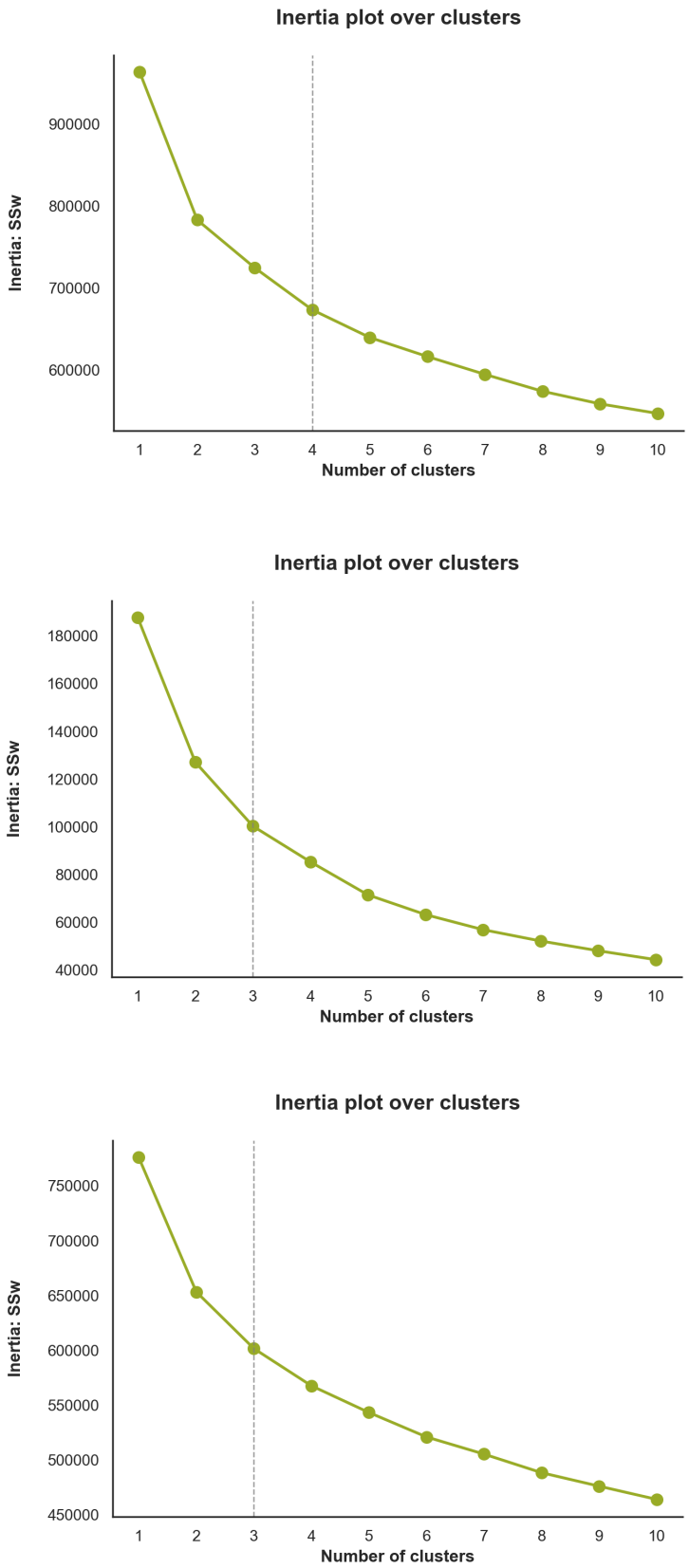
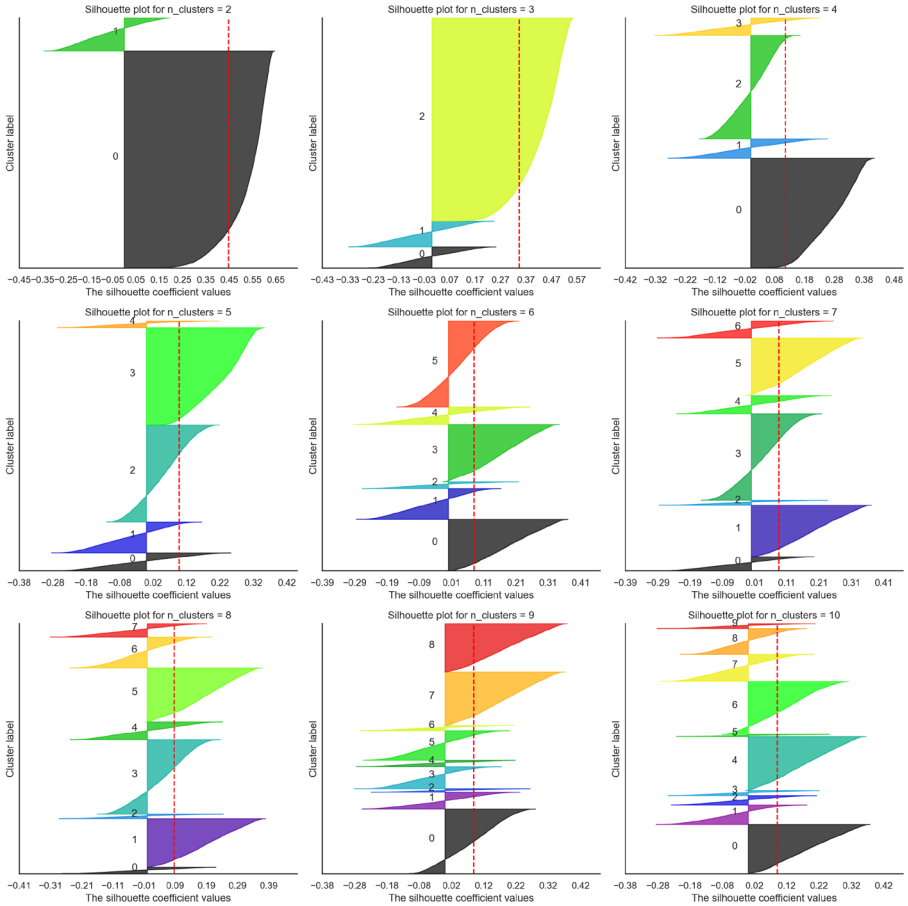
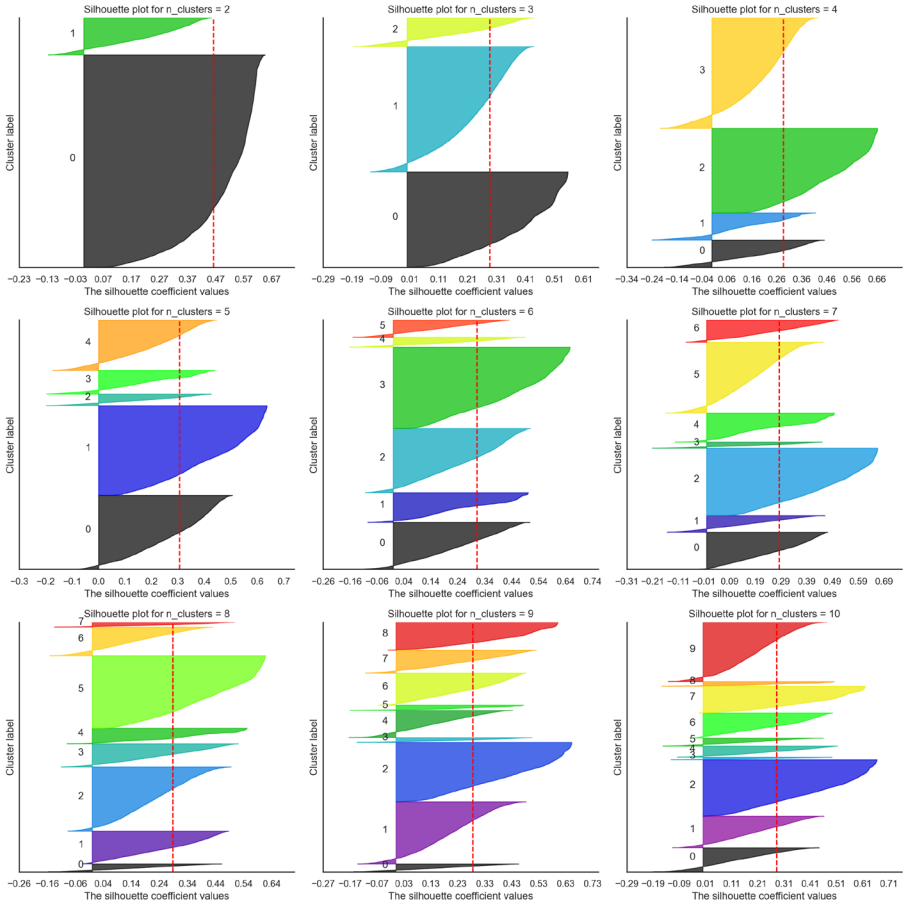


Figure E4 – Inertia Plot.

(**Top:** Overall, **Middle:** Value perspective, **Bottom:** Behaviour perspective)


Silhouette plot for n_clusters = 2
Silhouette plot for n_clusters = 3
Silhouette plot for n_clusters = 4
Silhouette plot for n_clusters = 5
Silhouette plot for n_clusters = 6
Silhouette plot for n_clusters = 7
Silhouette plot for n_clusters = 8
Silhouette plot for n_clusters = 9
Silhouette plot for n_clusters = 10
Cluster label
The silhouette coefficient values


Silhouette plot for n_clusters = 2
Silhouette plot for n_clusters = 3
Silhouette plot for n_clusters = 4
Silhouette plot for n_clusters = 5
Silhouette plot for n_clusters = 6
Silhouette plot for n_clusters = 7
Silhouette plot for n_clusters = 8
Silhouette plot for n_clusters = 9
Silhouette plot for n_clusters = 10
Cluster label
The silhouette coefficient values

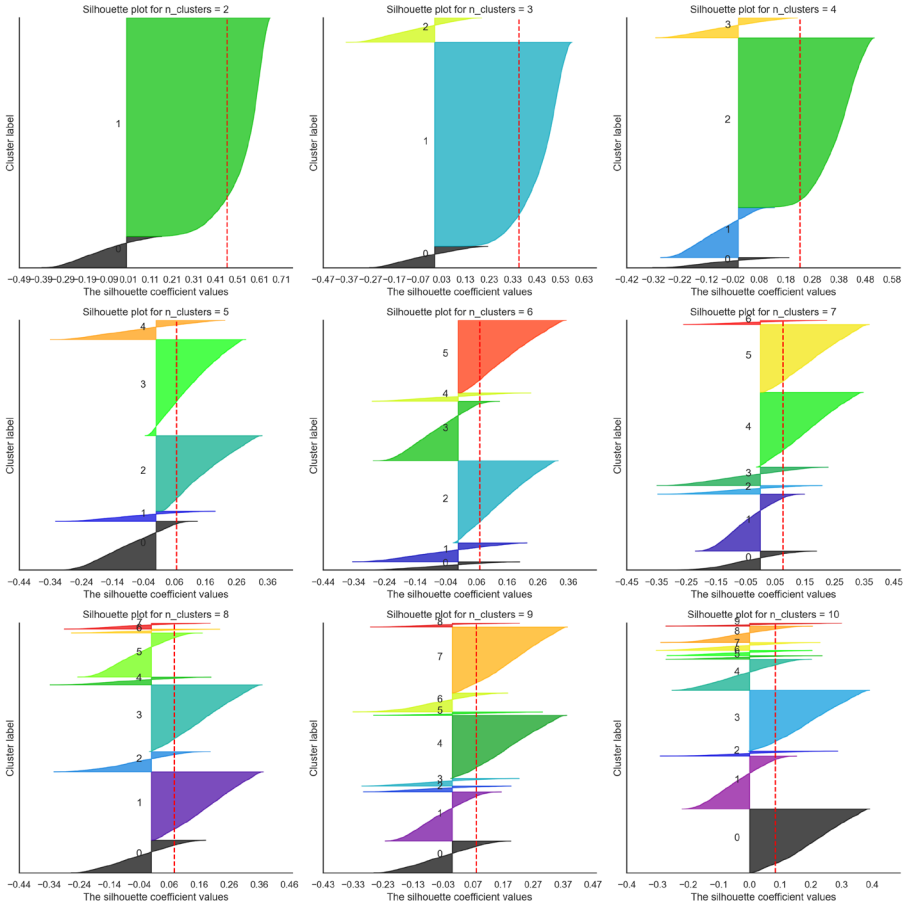
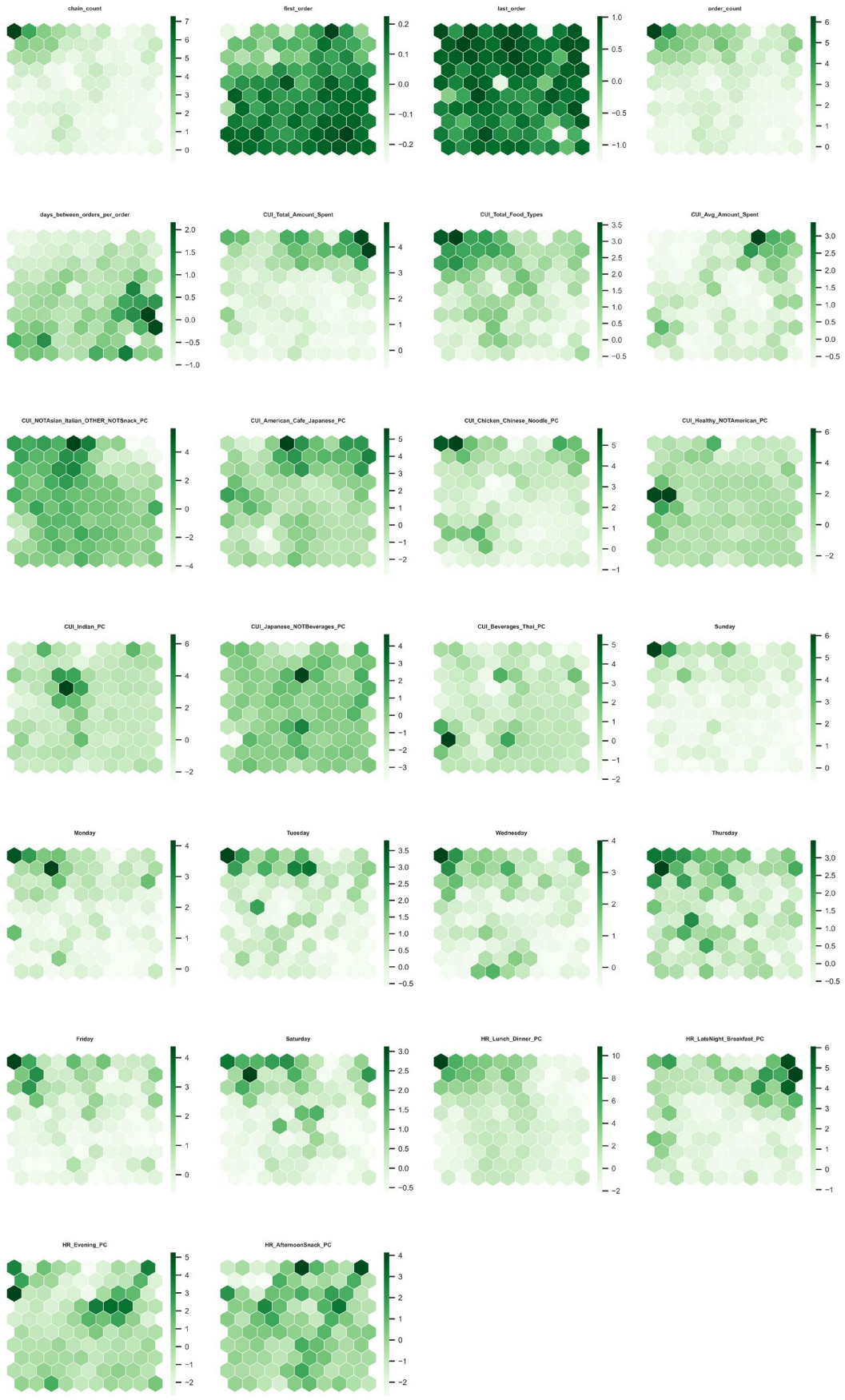


Figure E5 – K-Means Silhouette Plot.

(**Top:** Overall, **Middle:** Value perspective, **Bottom:** Behaviour perspective)


chain_count
first_order
last_order
order_count
days_between_orders_per_order
CUI_Total_Amount_Spent
CUI_Total_Food_Types
CUI_Avg_Amount_Spent
CUI_NOTAsian_Italian_OTHER_NOTSnack_PC
CUI_American_Cafe_Japanese_PC
CUI_Chicken_Chinese_Noodle_PC
CUI_Healthy_NOTAmerican_PC
CUI_Indian_PC
CUI_Japanese_NOTBeverages_PC
CUI_Beverages_Thai_PC
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
HR_Lunch_Dinner_PC
HR_LateNight_Breakfast_PC
HR_Evening_PC
HR_AfternoonSnack_PC

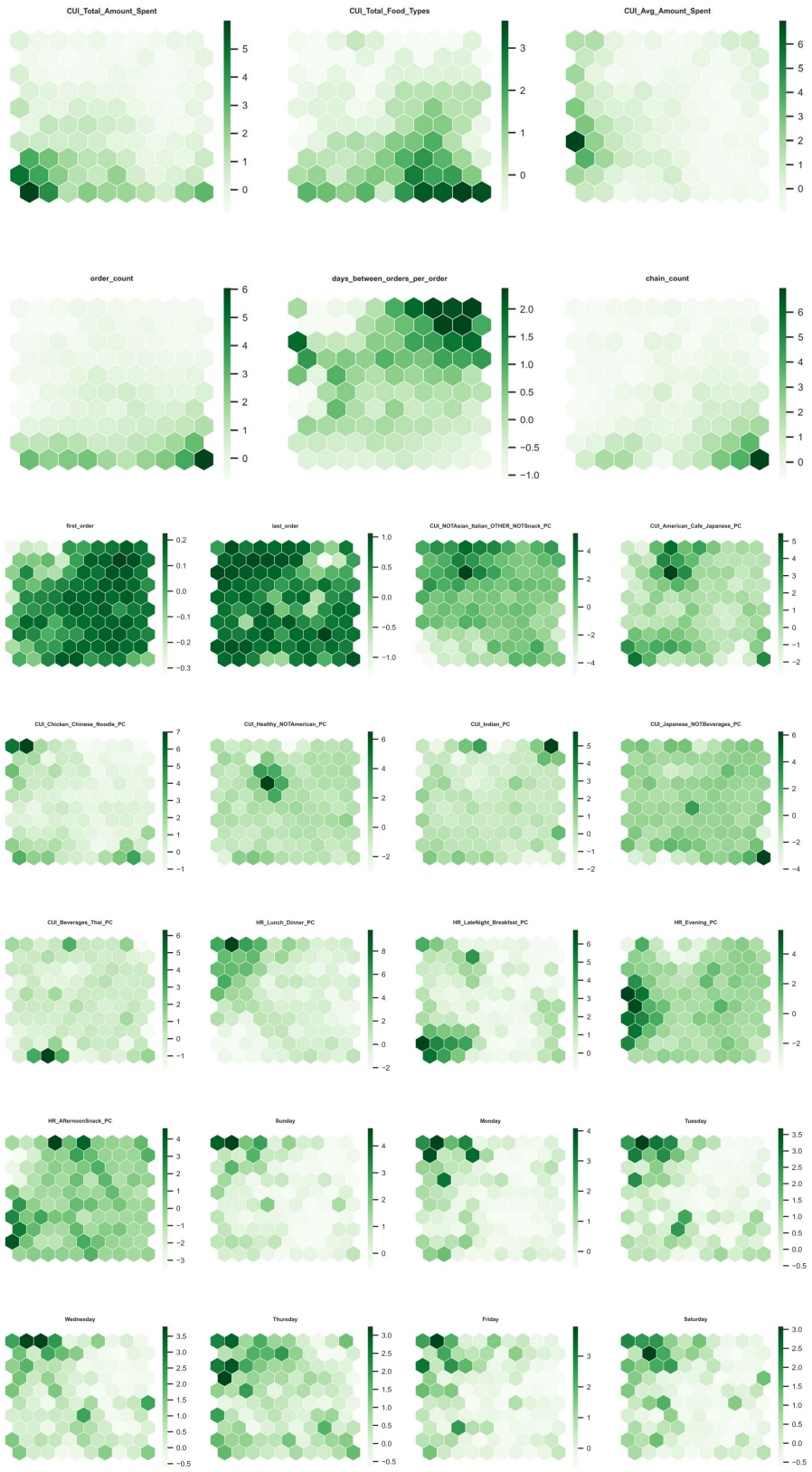


Figure E6 – SOM Component Planes (10x10).
(**Top:** Overall, **Middle:** Value perspective, **Bottom:** Behaviour perspective)

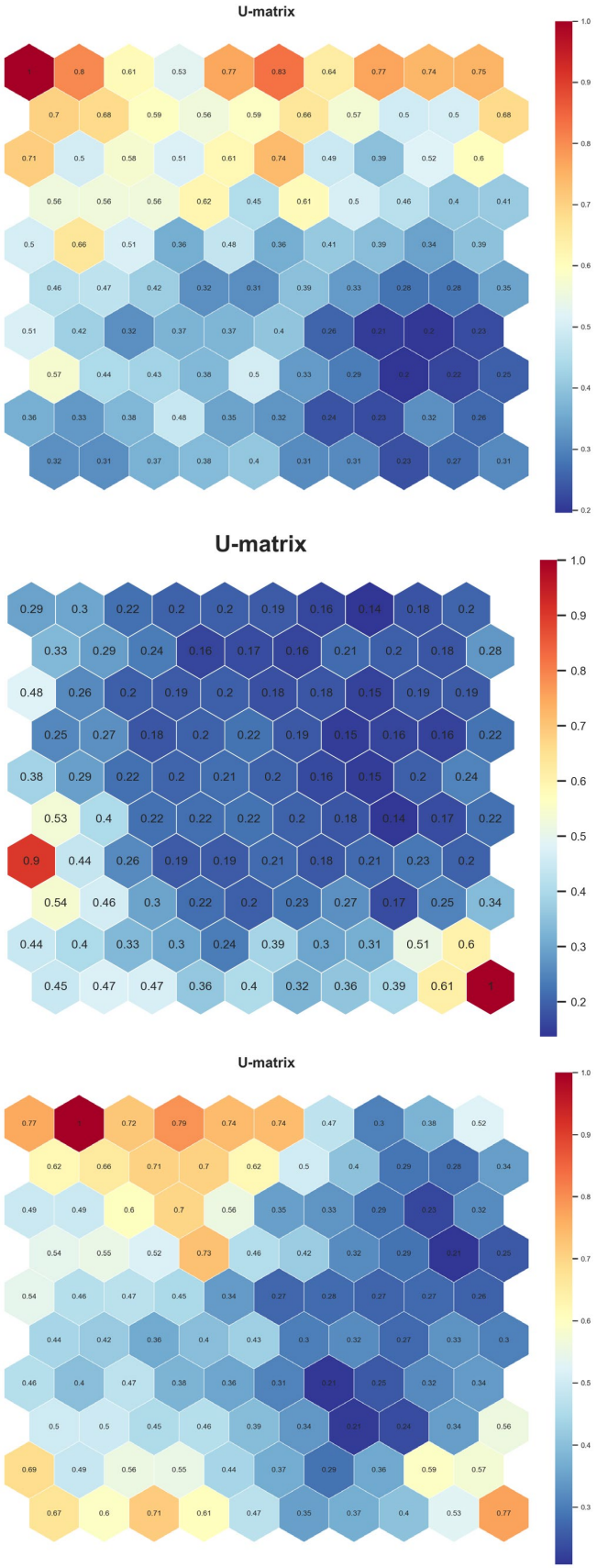


Figure E7 – U-Matrix (10x10).
(**Top:** Overall, **Middle:** Value perspective, **Bottom:** Behaviour perspective)

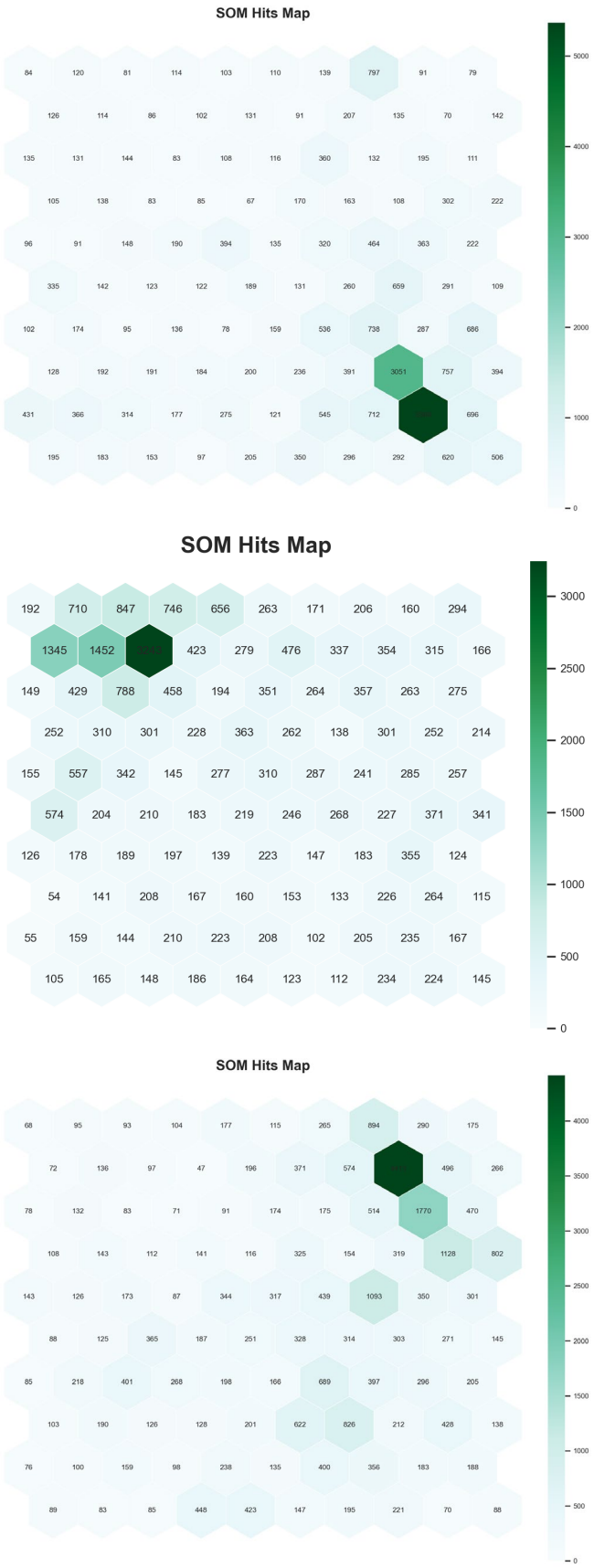
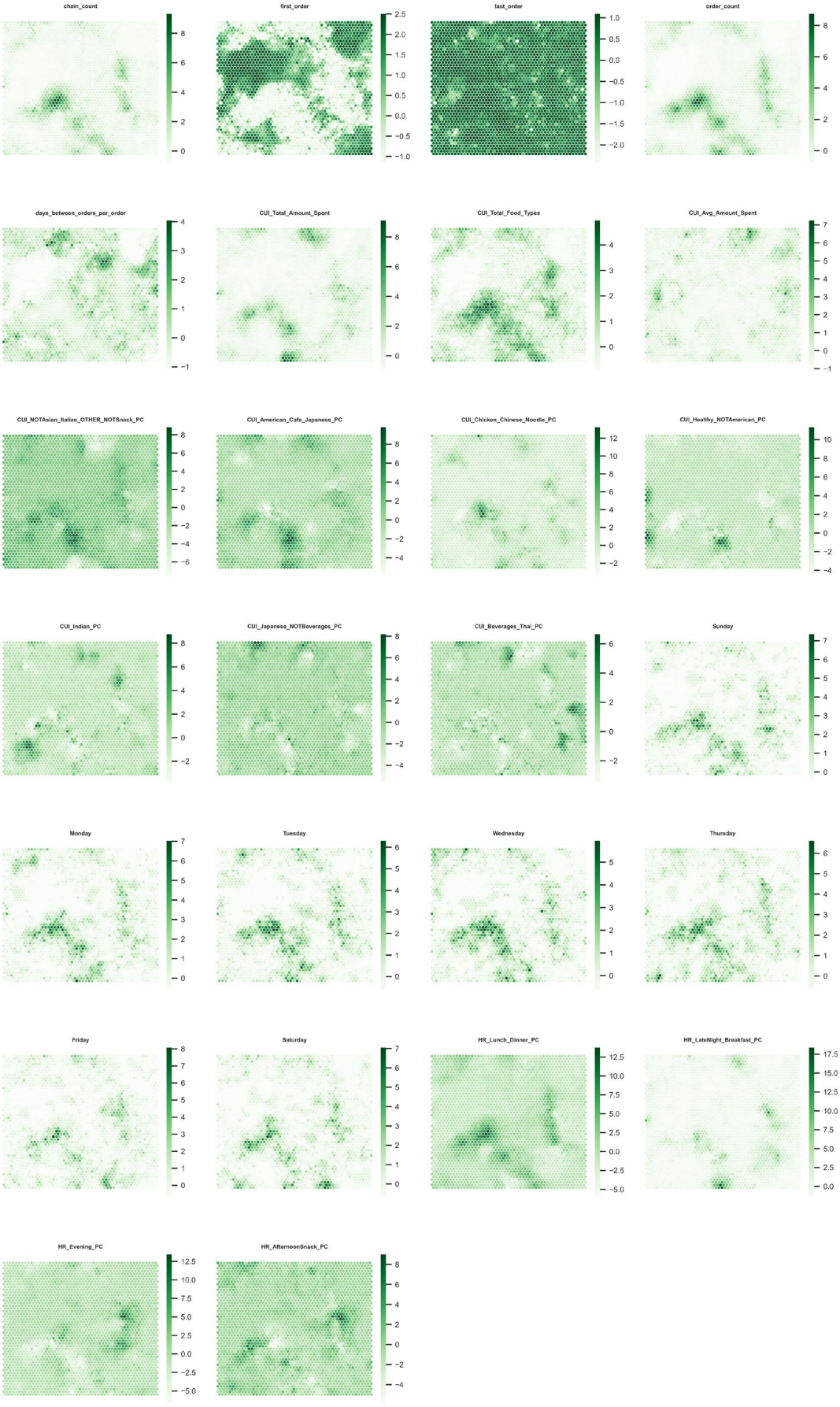


Figure E8 – SOM Hits Map (10x10).

(**Top:** Overall, **Middle:** Value perspective, **Bottom:** Behaviour perspective)


chain_count
first_order
last_order
order_count
days_between_orders_per_order
CUI_Total_Amount_Spent
CUI_Total_Food_Types
CUI_Avg_Amount_Spent
CUI_NOTAsian_Italian_OTHER_NOTSnack_PC
CUI_American_Cafe_Japanese_PC
CUI_Chicken_Chinese_Noodle_PC
CUI_Healthy_NOTAmerican_PC
CUI_Indian_PC
CUI_Japanese_NOTBeverages_PC
CUI_Beverages_Thai_PC
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
HR_Lunch_Dinner_PC
HR_LateNight_Breakfast_PC
HR_Evening_PC
HR_AfternoonSnack_PC

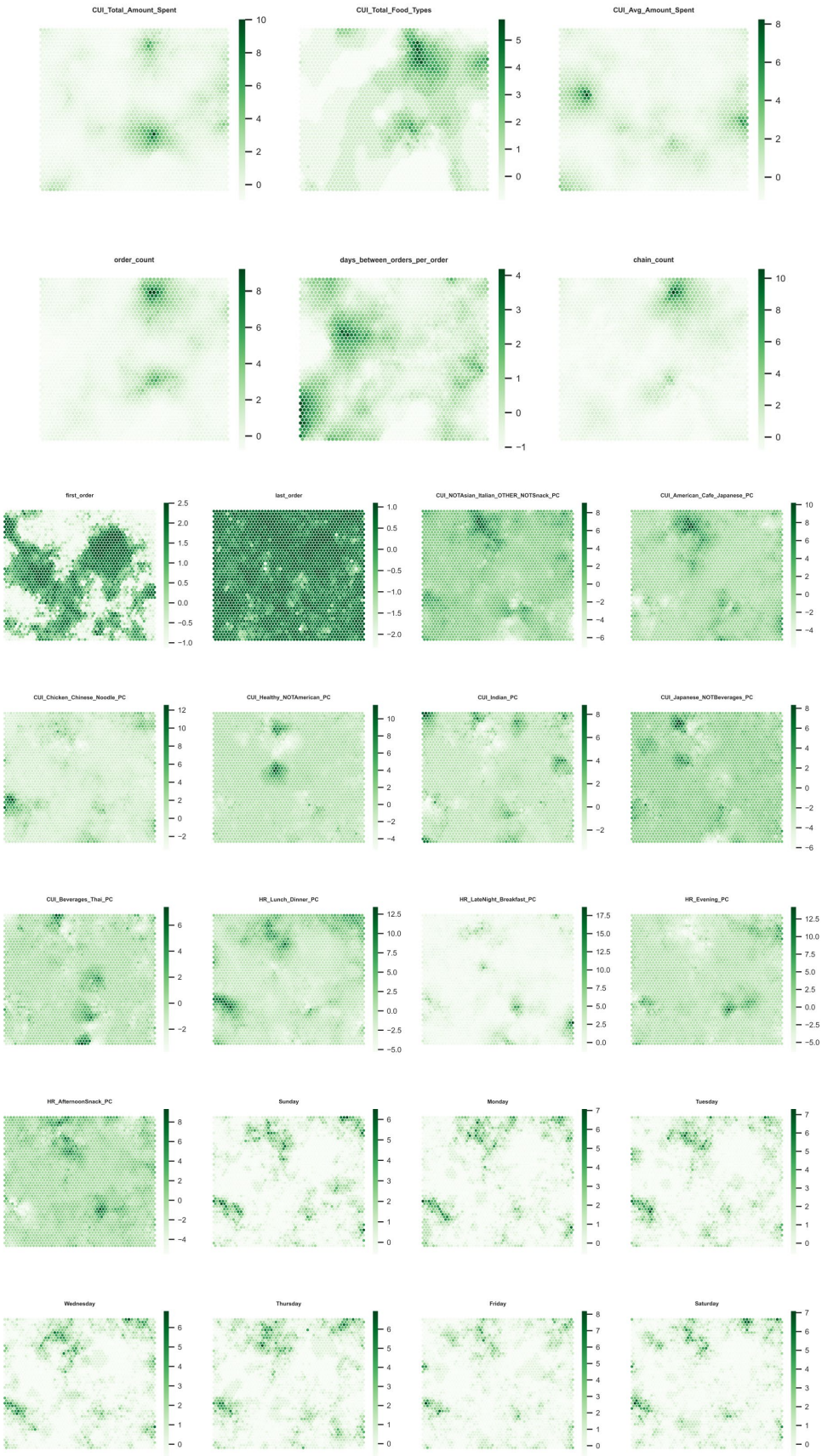


Figure E9 – SOM Component Planes (50x50).

(**Top:** Overall, **Middle:** Value perspective, **Bottom:** Behaviour perspective)

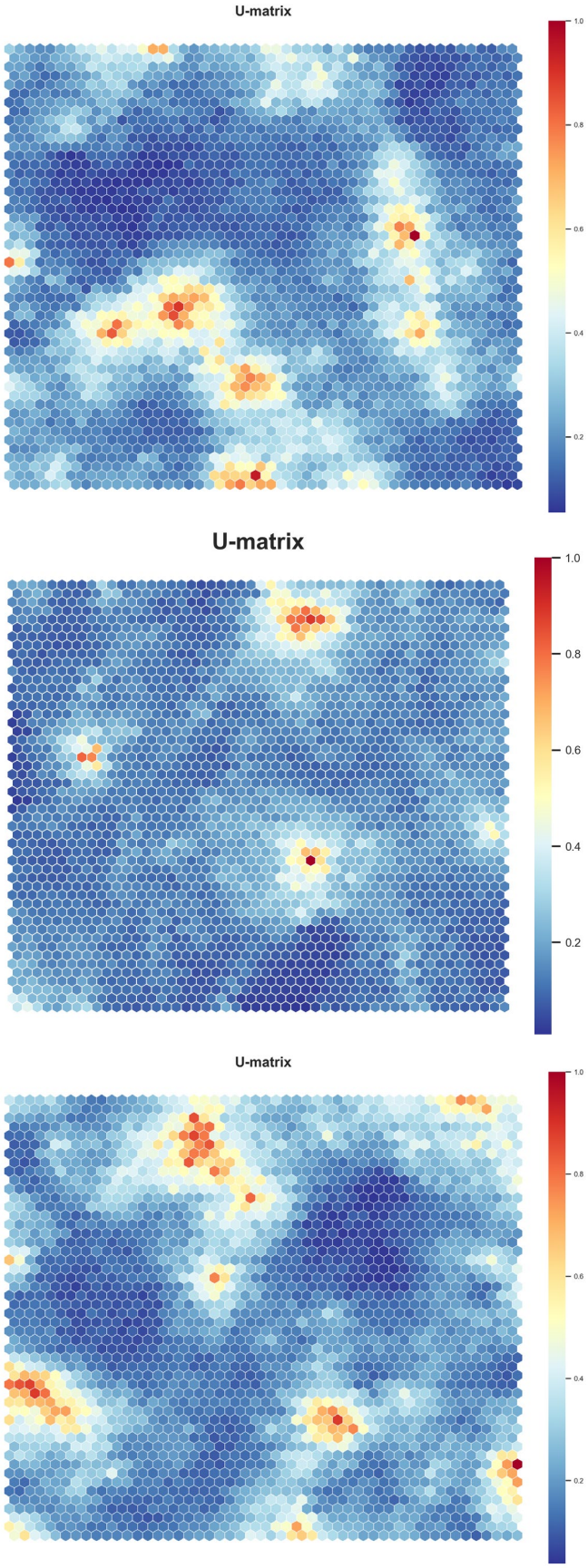


Figure E10 – U-Matrix (50x50).
(**Top:** Overall, **Middle:** Value perspective, **Bottom:** Behaviour perspective)

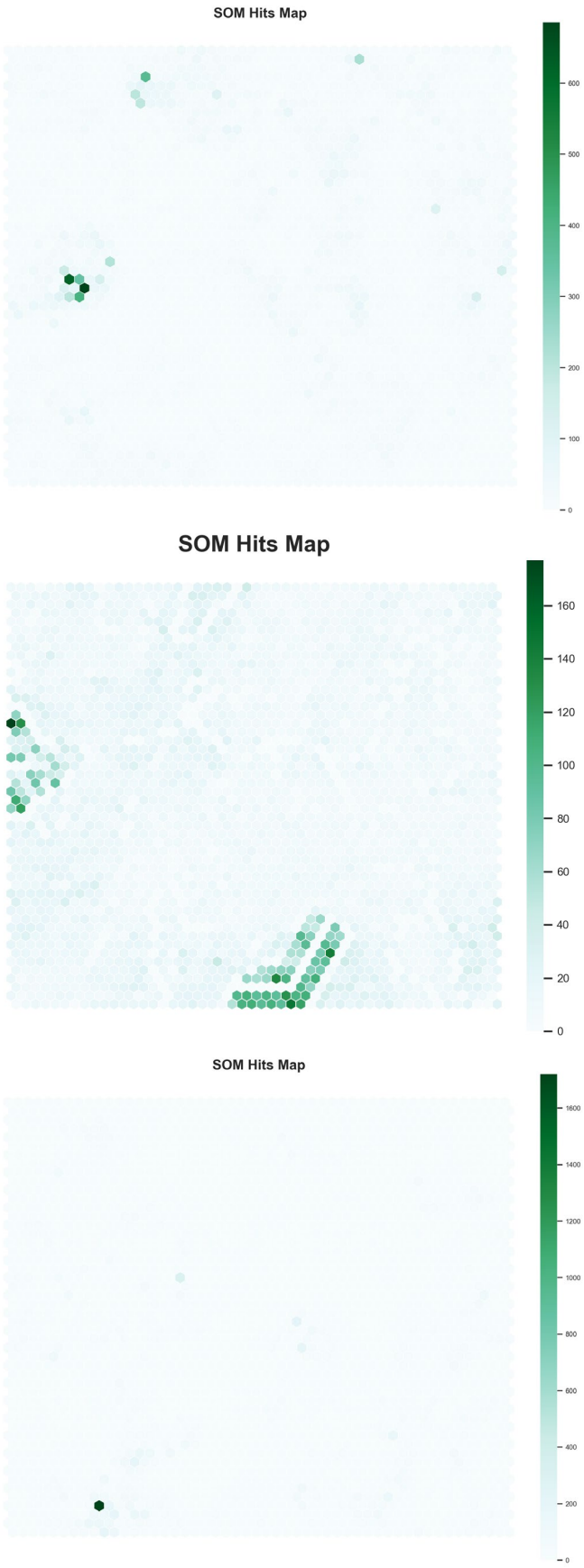


Figure E11 – SOM Hits Map (50x50).
(**Top:** Overall, **Middle:** Value perspective, **Bottom:** Behaviour perspective)

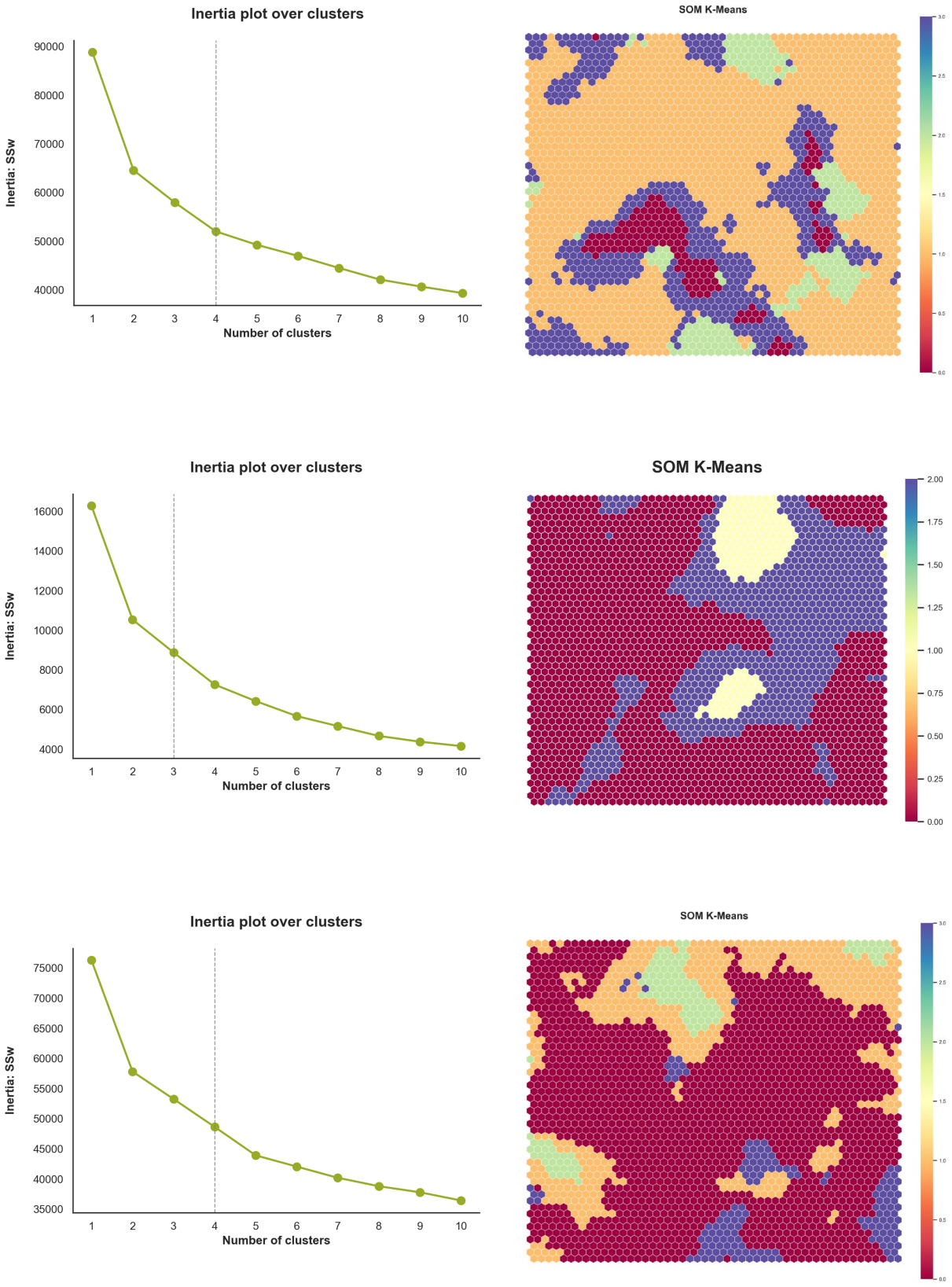


Figure E12 – SOM+K-Means Inertia Plot.
(**Top:** Overall, **Middle:** Value perspective, **Bottom:** Behaviour perspective)

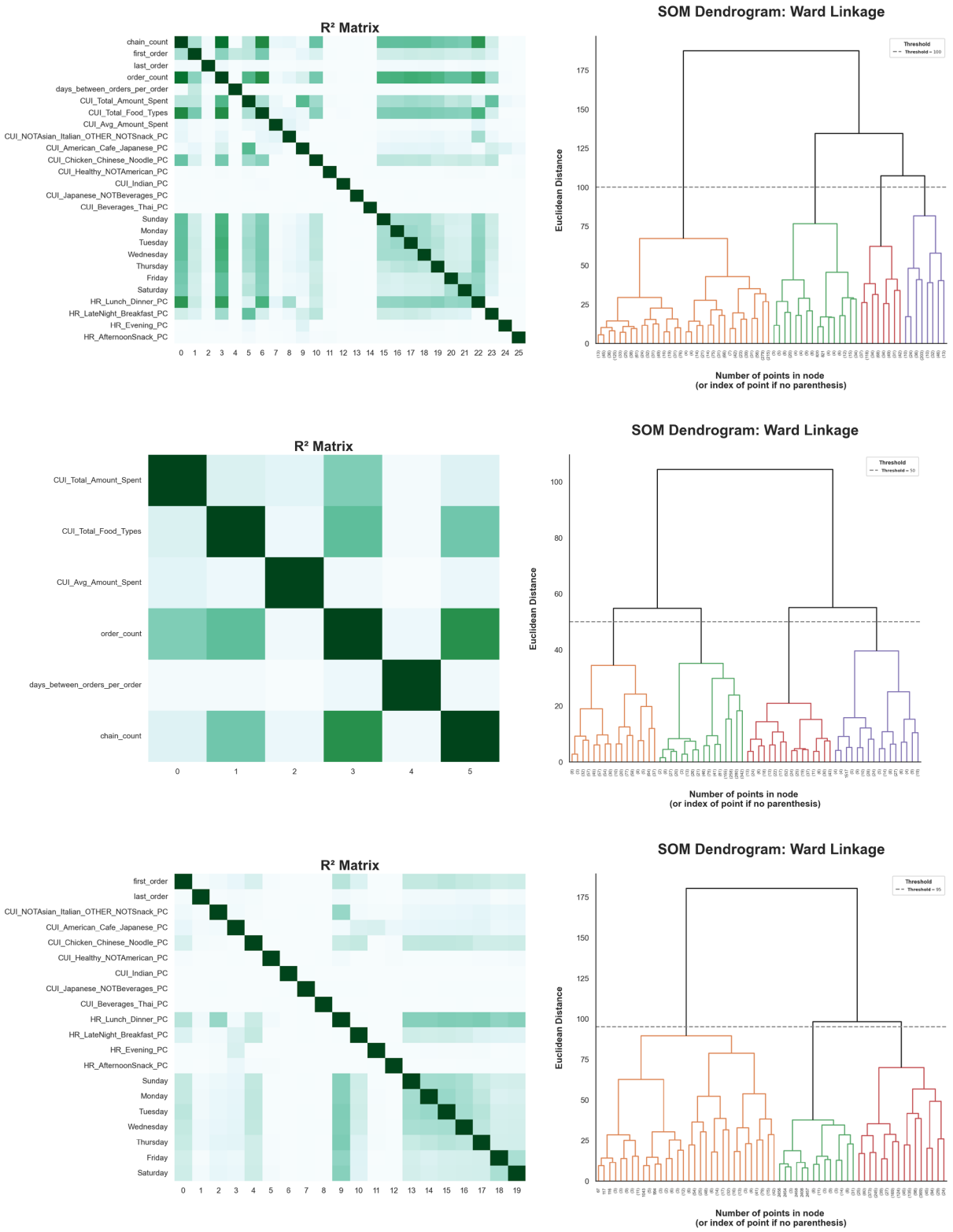


Figure E13 – R^2 Matrix and Dendrogram (SOM+HC).

(**Top:** Overall, **Middle:** Value perspective, **Bottom:** Behaviour perspective)

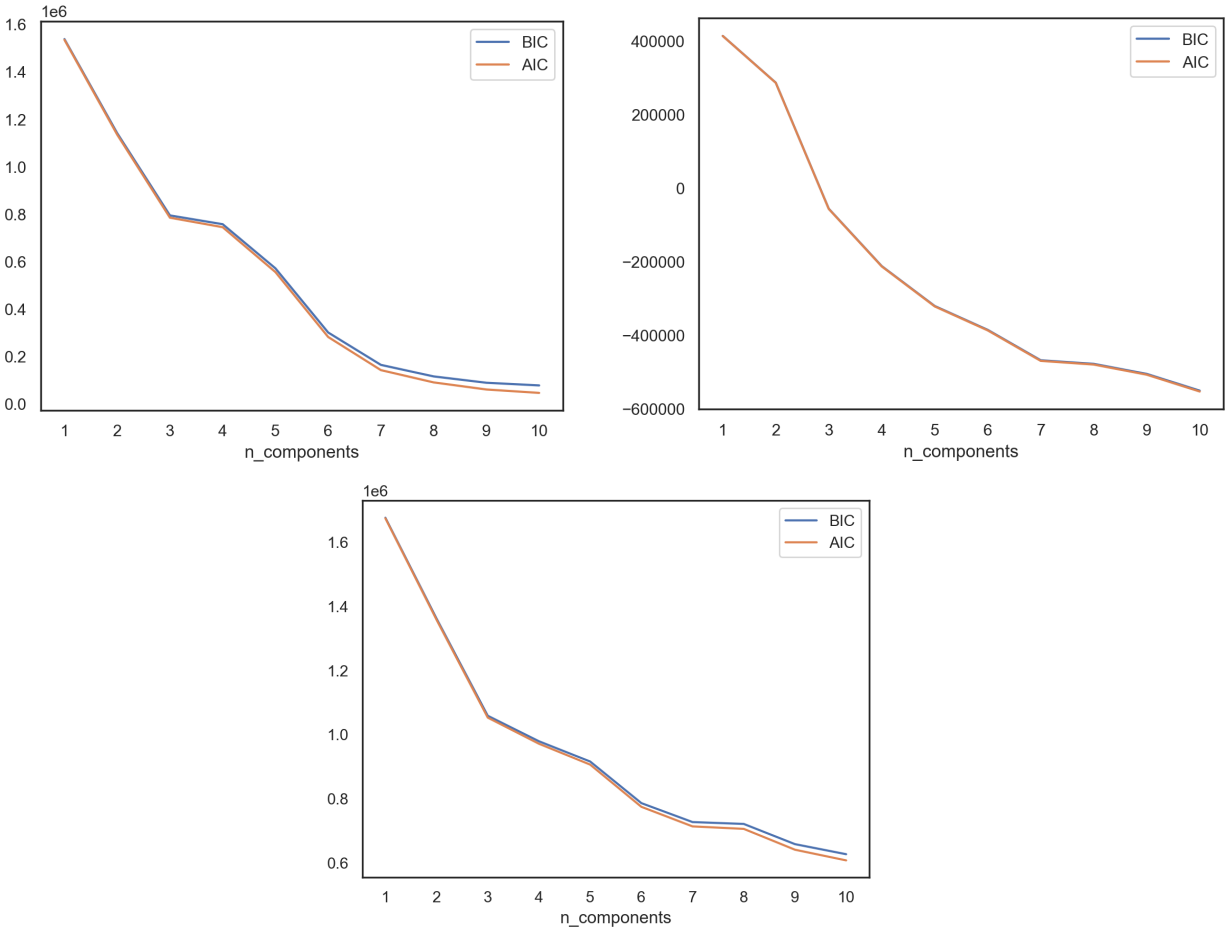


Figure E14 - GMM Algorithm AIC and BIC criteria plots
(**Top Left:** Overall, **Top Right:** Value perspective, **Bottom:** Behaviour perspective)

Table E1 – R^2 values for all clusters studied.
(The value between parentheses indicates the number of clusters that each algorithm generates)

Clustering Method	Overall	Value Perspective	Behaviour Perspective
Hierarchical	0.245 (4)	0.419 (3)	0.166 (3)
K-Means	0.248 (3)	0.466 (3)	0.225 (3)
SOM + HC	0.242 (4)	0.473 (4)	0.181 (3)
SOM + K-Means	0.283 (4)	0.410 (3)	0.249 (4)
Mean Shift	0.072 (4)	0.093 (2)	0.027 (3)
DBSCAN	0.067 (2)	0.065 (10)	0.086 (6)
GMM	0.165 (3)	0.448 (5)	0.123 (3)

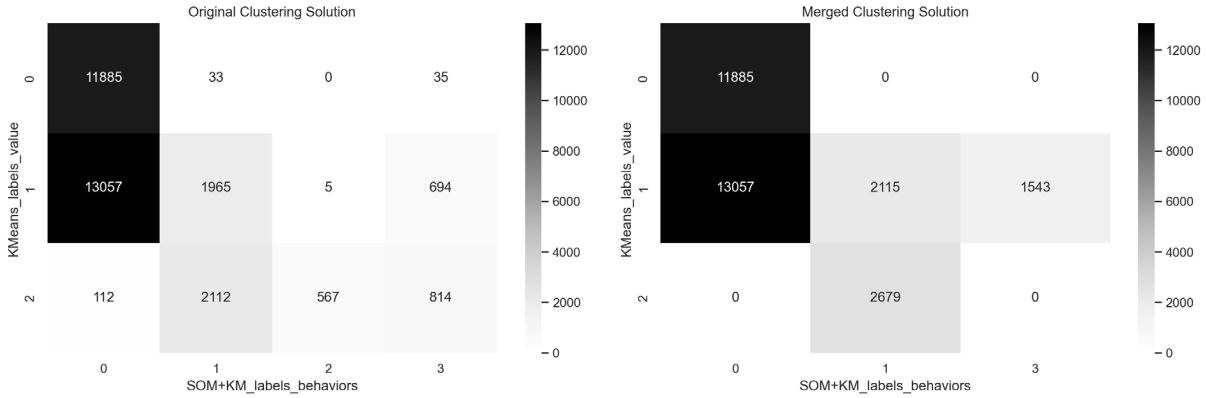


Figure E15 – Manual merging perspectives.

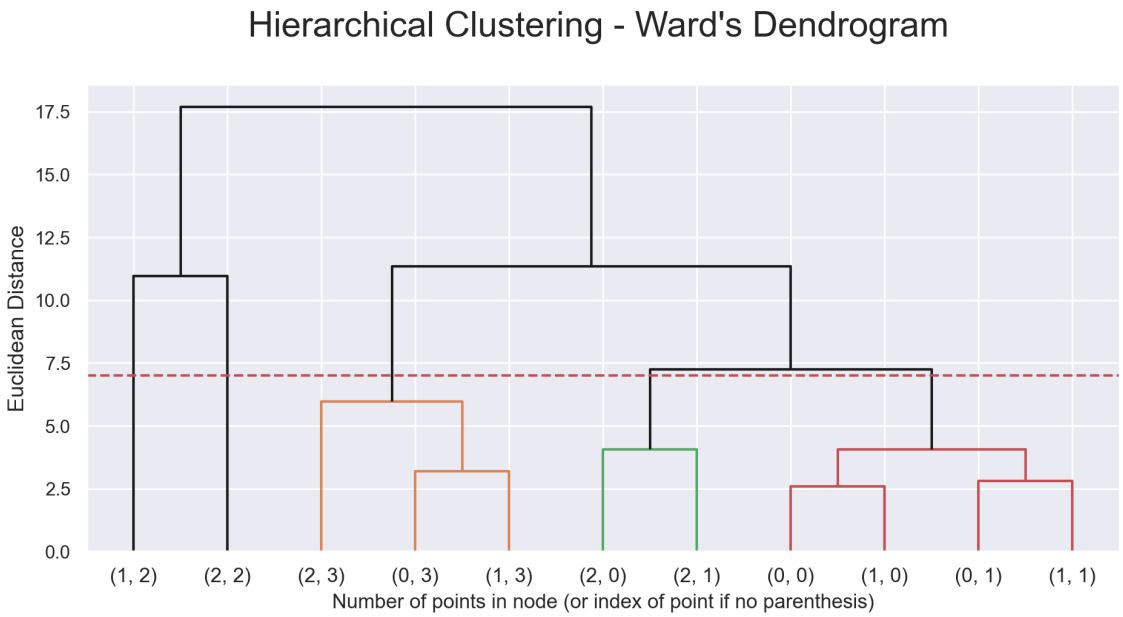


Figure E16 – HC Dendrogram for the merging of both value and behaviour perspectives.

APPENDIX F. PROFILING

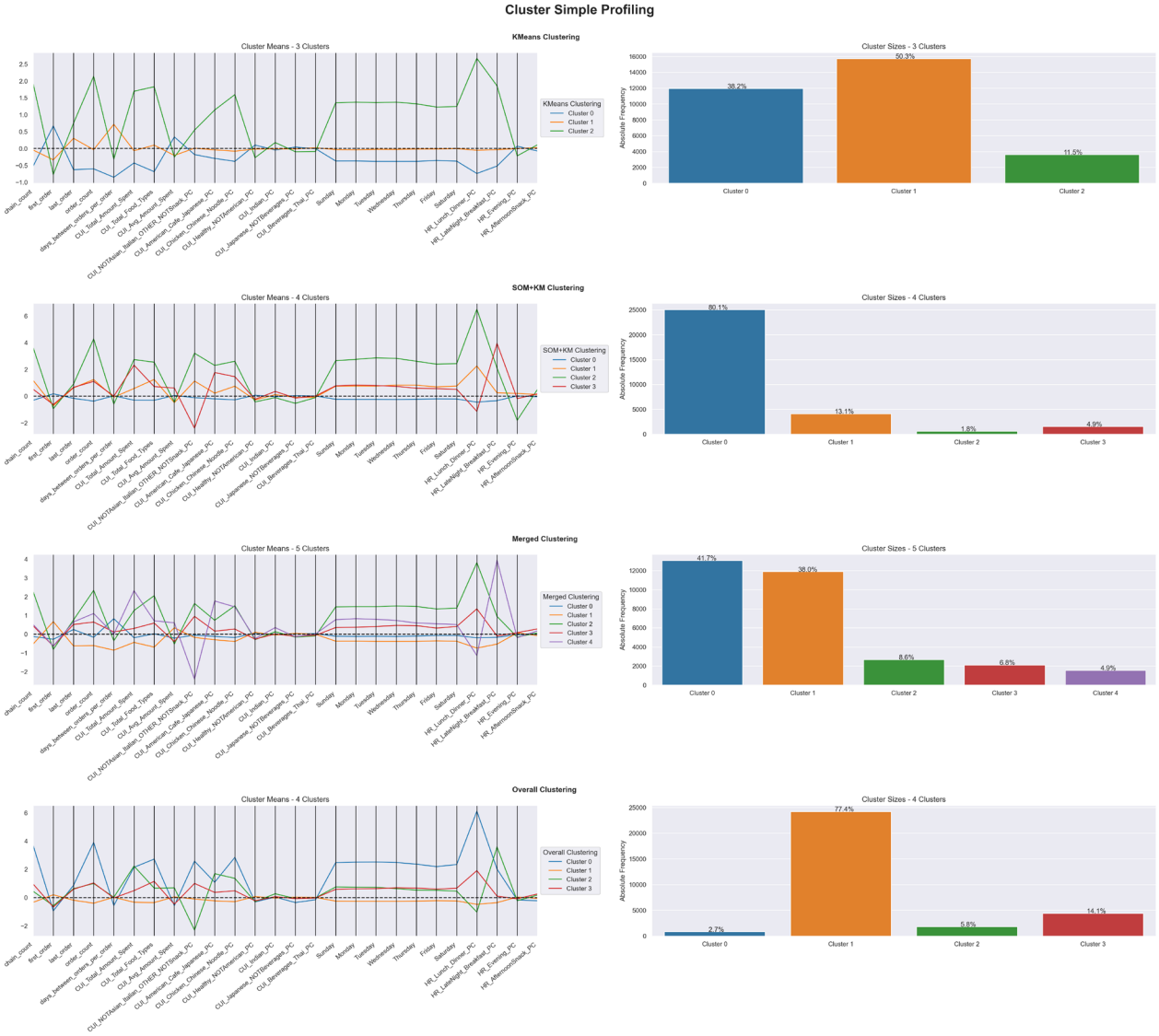
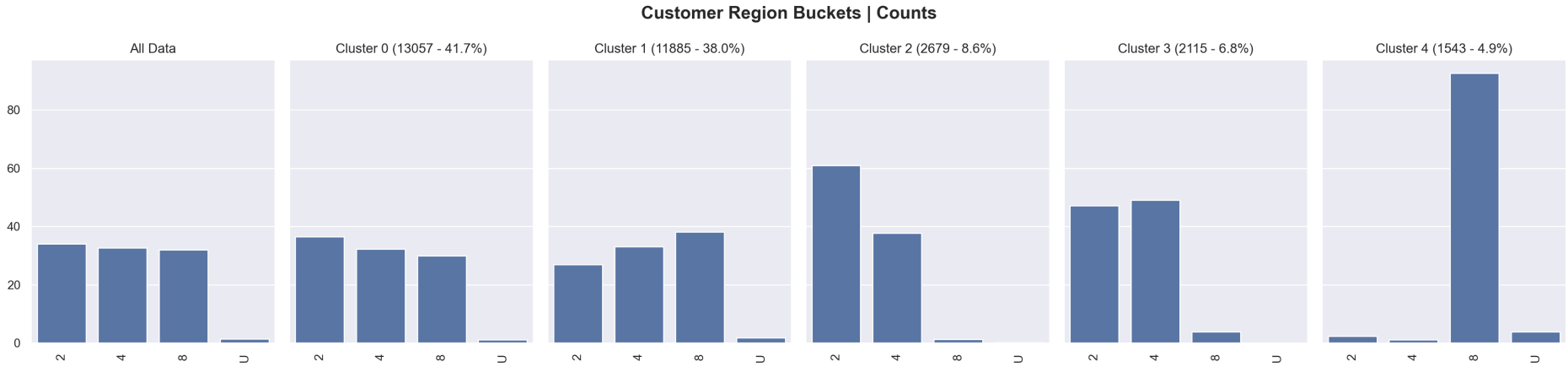
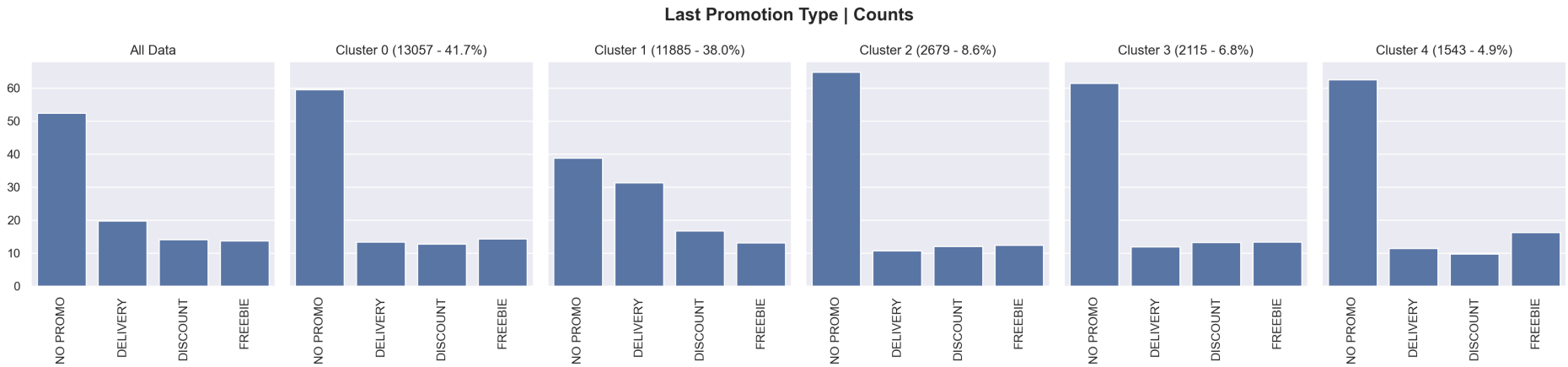


Figure F1 – Cluster simple profiling.

Figure F2 - Heatmap Cluster Profiling.


Customer Region Buckets | Counts
All Data
Cluster 0 (13057 - 41.7%)
Cluster 1 (11885 - 38.0%)
Cluster 2 (2679 - 8.6%)
Cluster 3 (2115 - 6.8%)
Cluster 4 (1543 - 4.9%)
0
20
40
60
80
2
4
8
U


Last Promotion Type | Counts
All Data
Cluster 0 (13057 - 41.7%)
Cluster 1 (11885 - 38.0%)
Cluster 2 (2679 - 8.6%)
Cluster 3 (2115 - 6.8%)
Cluster 4 (1543 - 4.9%)
0
10
20
30
40
50
60
NO PROMO
DELIVERY
DISCOUNT
FREEBIE

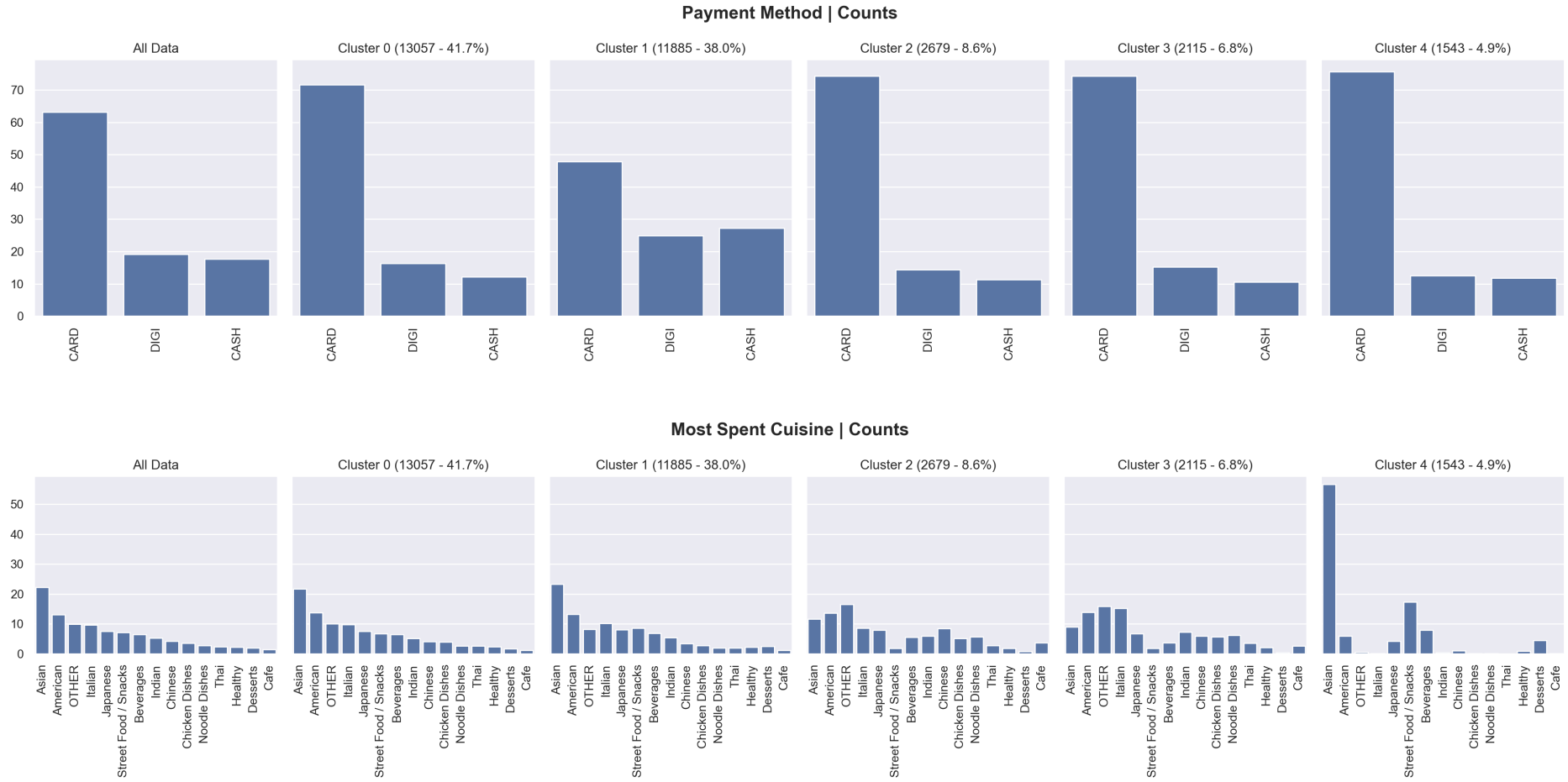


Figure F3 – Barplots with categorical variables and final clustering solution.

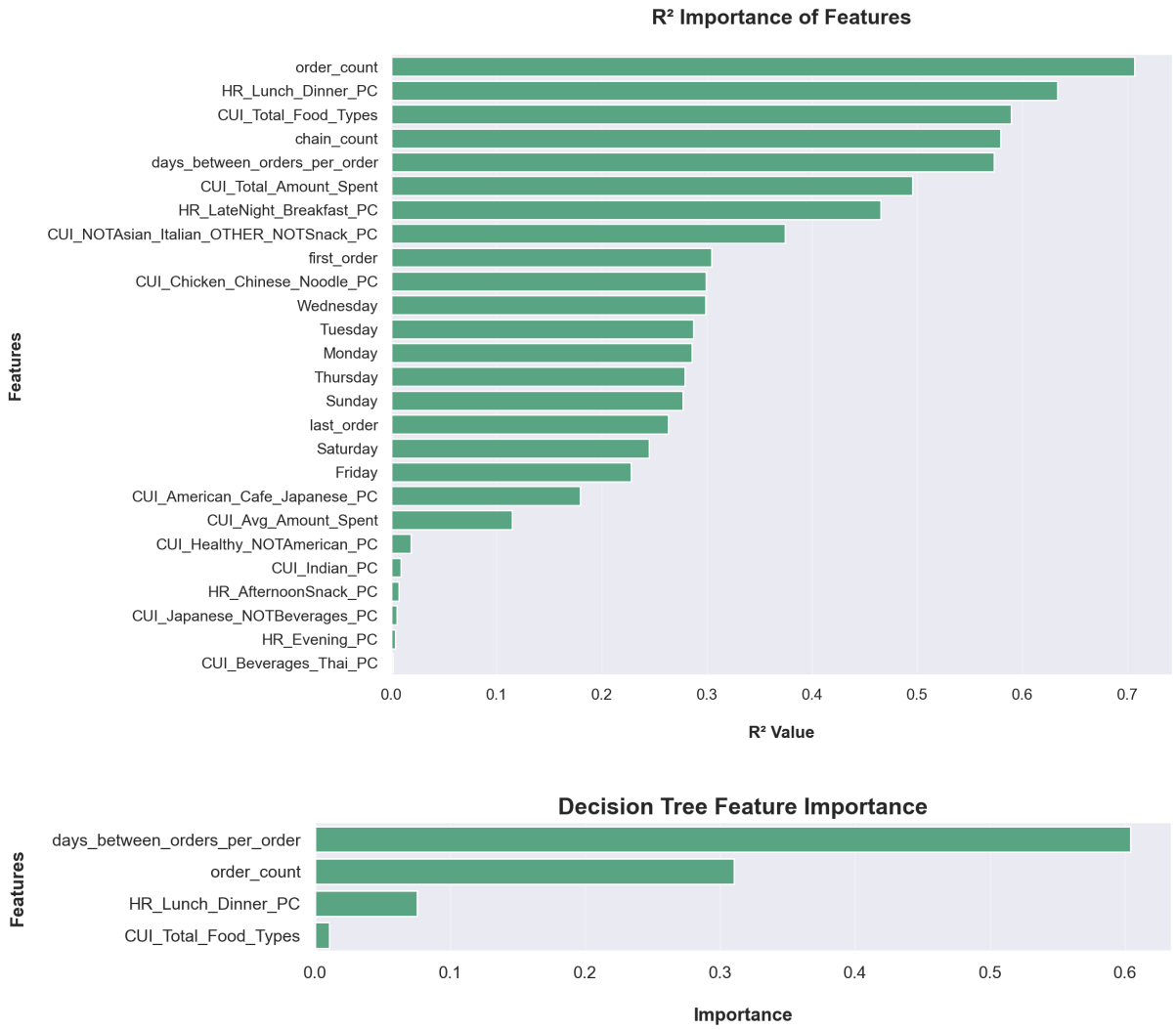


Figure F4 - Feature Importance based on R^2 and Decision Tree Feature Importance

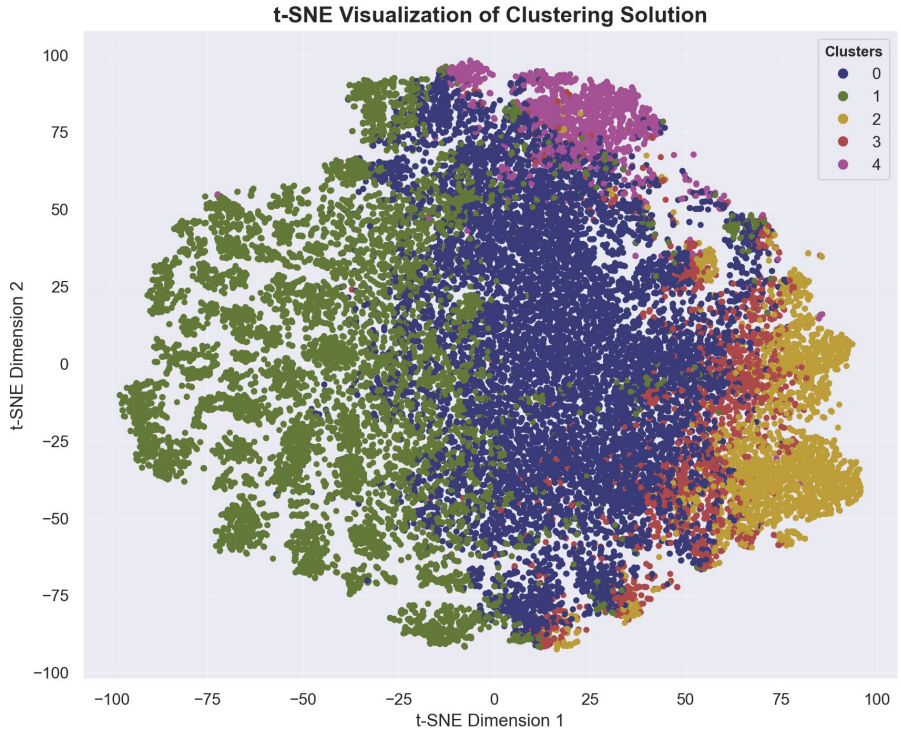


Figure F5 - t-SNE Visualization of Clustering Solution.

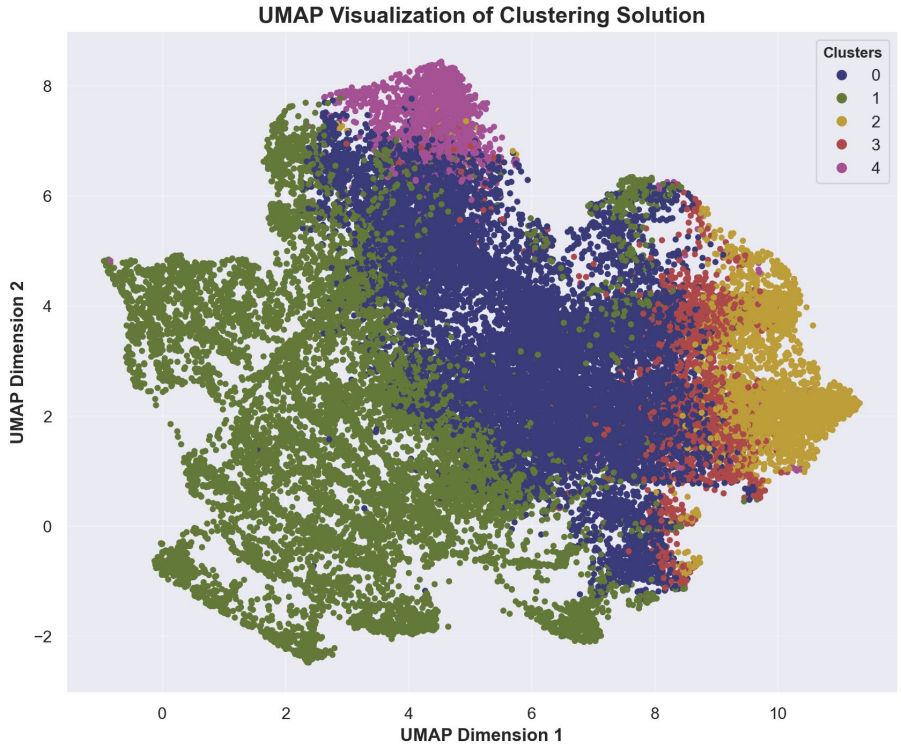


Figure F6 - UMAP Visualization of Clustering Solution.

Table F1 – Contingency Table for *order_count* and *CUI_Total_Amount_Spent* with quartiles.

		***CUI* Total Amount Spent**				
		Q1	**Q2**	**Q3**	**Q4**	**Total**
Order Count	**Q1**	6181 (19.76%)	4511 (14.42%)	2835 (9.06%)	560 (1.79%)	14087 (45.0%)
	Q2	1087 (3.48%)	1594 (5.10%)	1651 (5.28%)	751 (2.40%)	5083 (16.3%)
	Q3	501 (1.60%)	1241 (3.97%)	1849 (5.91%)	1764 (5.64%)	5355 (17.1%)
	Q4	52 (0.17%)	488 (1.56%)	1470 (4.70%)	4744 (15.17%)	6754 (21.6%)
	Total	7821 (25.0%)	7834 (25.0%)	7805 (25.0%)	7819 (25.0%)	**31279 (100%)**

ANNEX A. CRISP-DM

To address the defined objectives, the **CRISP-DM** methodology was employed. This methodology is a dynamic and fluid iterative approach (**Figure A1**) and relies on the exploration and refinement of its different stages to achieve better results, thereby supporting more effective and meaningful decision-making. By definition, this methodology comprises six phases: *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modelling*, *Evaluation*, and *Deployment*, with various tasks within each phase that will be explored in this project. [14]

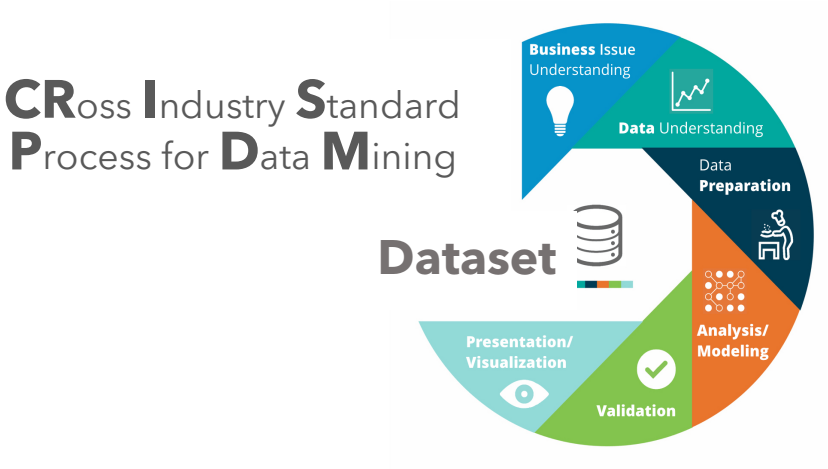


Figure A1 – CRISP-DM Methodology Cycle.
[Adapted from [15]]

Business Understanding

This initial phase, which was started during the initial phase of this project, centred on defining the project's objectives and requirements from the perspective of *ABCDEats Inc.* The aim was to understand the need for effective customer segmentation, recognizing that today's consumers are increasingly selective and demand personalized experiences. Therefore, our goal was to develop a data-driven customer segmentation strategy, enabling to tailor its services and marketing more effectively.

Data Understanding

The second phase, also detailed in our first report, involved collecting and exploring the *ABCDEats* customer dataset, which included data from 3 cities over 3 months, with focus on customer behaviours and purchases. During this phase, we focused in understanding the dataset's quality and suitability for customer segmentation, by calculating descriptive statistics and visualizing various aspects of the data, allowing for a first understanding of the data's characteristics and identifying specific patterns.

Data Preparation

We then moved to data preparation, where we selected, cleaned, and transformed the data. Key tasks included handling inconsistencies, imputing missing values, using deterministic approaches, or imputation, while creating new features. This stage also involved addressing outlier values, to ensure that the data was reliable and well-structured for modelling. We used Principal Component Analysis (PCA) to reduce dimensionality and create new features.

Modelling

During this phase, we applied a range of clustering algorithms, such as Hierarchical Clustering (HC), K-Means, Self-Organizing Maps (SOM), Mean Shift, DBSCAN, and Gaussian Mixture Models (GMM). We also chose to use value-based and behaviour-based perspectives in the clustering process. Each method was assessed based on different metrics such as the inertia, silhouette score, and R^2. In the end, a combination of methods was used to construct the final solution, which was then further profiled and explained.

Evaluation

After the modelling phase, the focus was on evaluating the clustering results, using metrics and visualizations to assess the quality of each model, and by comparing different approaches. The evaluation phase also considered which approach best described the underlying structures of the data.

Deployment

The deployment phase, beyond the typical presentation of results, involved the development of an interactive web application using *Streamlit*. This application, provided as an optional yet valuable extension to the project, empowers *ABCDEats Inc.* to explore and interact with the EDA and customer segmentation analysis in a user-friendly and visually intuitive manner.