DATA MINING PROJECT

Master in Data Science and Advanced Analytics

**NOVA Information Management School**

Universidade Nova de Lisboa

# Comprehensive Customer Segmentation Strategy for ABCDEats, Inc.

## Group 04

Beatriz Monteiro, 20240591

Luís Semedo, 20240852

Pedro Santos, 20240295

Rodrigo Miranda, 20240490

Fall/Spring Semester 2024-2025

# Abstract

This project aimed to build an unsupervised learning framework to conduct robust clustering analysis on a dataset of customers from "ABCDEats, Inc." as part of a consulting project. The goal was to develop a marketing-oriented strategy based on clustering insights derived from customer patterns and behaviors. To do so, the dataset provided contained 56 features and 31,888 rows.

The project followed a structured methodology, starting with extensive exploratory data analysis (EDA) and visualization to identify key patterns and insights. This was followed by a rigorous data preprocessing pipeline to address inconsistencies, missing values, outliers, and variable transformations. Feature engineering played a pivotal role in the project, enabling the creation of derived, aggregated, and proportional variables that formed the foundation of the analysis.

Various clustering techniques and visualizations were explored, leveraging insights from our EDA and self-organizing maps (SOM) to define meaningful customer perspectives. Outliers were extensively addressed using DBSCAN, and clustering solutions were derived through KMeans and Hierarchical clustering methods. The final pipeline culminated in customer profiling and segmentation, providing actionable insights to properly segment customer groups.

**Keywords**: *Unsupervised Learning, Data Mining, Exploratory Data Analysis, Self-Organizing Maps, Outlier Treatment, KMeans, Hierarchical Clustering, DBSCAN, Data Preprocessing, Clustering*

# TABLE OF CONTENTS

# 1. Introduction

Understanding human behavior has long been a cornerstone of study, providing insights into individual and societal evolution. In the modern, technology-driven era, data-driven strategies play a crucial role in analyzing and predicting consumer behavior, enabling companies to enhance profitability and improve customer experiences with personalized solutions. Customer segmentation, a vital practice in consumer-focused industries, further refines this process by distinguishing distinct customer groups. This approach delivers actionable insights that boost revenue, improve efficiency, and provide tailored offerings aligned with consumer preferences.

As consultants for ABCDEats Inc., our goal is to analyze customer behavior and develop a data-driven strategy to foster growth. By segmenting customers based on their purchasing behavior, we aim to help the company optimize its marketing efforts and target specific customer groups more effectively. Using a dataset encompassing demographic features and customer preferences, we explore the question: "How can we group customers based on their purchasing behavior to design targeted marketing strategies?"

Our approach begins with thorough Exploratory Data Analysis (EDA) to comprehensively understand the dataset. EDA serves as the foundation for identifying key patterns, inconsistencies, and relationships, enabling effective data management and providing insights essential for strategic decision-making. This process emphasizes the importance of cleaning and transforming raw data into meaningful features.

Feature engineering plays a vital role in this analysis by transforming raw data into structured and informative variables for modeling purposes. Using methods such as aggregation, transformation, extraction, and binning, we aim to enhance the dataset's predictive power.

To address the complexity of the problem, a robust analytical pipeline was developed. Techniques such as Self-Organizing Map (SOM) visualizations, KMeans clustering, and Hierarchical Clustering were employed alongside careful outlier removal to ensure accuracy. These methods, supported by prior insights, formed the backbone of our analysis.

Finally, the preliminary cluster solutions were refined and enriched through profiling, using our categorical features and advanced visualizations, providing actionable insights to drive targeted marketing strategies.

# 2. Exploratory Data Analysis

The previous EDA (Annex 1) work completed in the first delivery included univariate analysis of categorical and numerical features, bivariate analysis between all features, and feature engineering. Therefore, no major changes were made in this second phase, except for a slight enhancement of the feature engineering section, which will be addressed shortly.

# 3. Preprocessing

Preprocessing proved to be a crucial step in achieving satisfying results in this unsupervised learning project. It is important to note that many of the techniques and processes applied were the result of extensive trials and experimentation, ultimately leading to the final solution presented below.

## 3.1. Addressing Inconsistencies

Through our meticulous Exploratory Data Analysis (EDA) and visualizations, we identified several inconsistencies that needed to be addressed as a priority.

- **Duplicate Rows:** Duplicate rows were identified where some “*customer_id”’s* and their corresponding feature values were identical. This issue, likely deriving from data extraction or system anomalies, was resolved by eliminating the 13 duplicate entries from the dataset.
- **Logical Corrections:** For customers with a “*last_order”* value of 0 (indicating their last order occurred on the first day of the dataset), the “*first_order”* feature was found to be always NaN, and all cases where “*first_order”* was NaN corresponded to a value of 0 in “*last_order”*. To ensure consistency, “*first_order”* was filled with 0 in these cases. Additionally, we identified 156 instances where “*product_count”* was 0 for specific “*customer_id”*‘s. Since it is illogical for customers with no recorded orders to appear in the dataset, these anomalies were flagged and removed.

### 3.1.1. Addressing Incoherences - Categorical Features

- ***last_promo:*** In prior analysis, we determined that a “-” in the “*last_promo”* feature represented the absence of a promotion. This logic was maintained to ensure consistency.
- ***is_chain:*** We observed inconsistencies in the “*is_chain”* feature metadata. Based on our insights, we assumed this feature represents the number of orders placed from chain restaurants. This assumption was carried forward in our analysis, hence, a simple renaming of this feature was conducted.
- ***customer_region:*** The “-” values in the “*customer_region”* feature were previously hypothesized to belong to the “8000’s” city region. This assumption, supported by our initial findings, was also maintained.

## 3.2 Missing Values

Fortunately, missing values were not a significant issue in our dataset. However, there were a few cases that needed to be fixed.

- ***HR_0:*** The feature “*HR_0*” presented a unique challenge as it contained no valid values. To address this, we followed a logical imputation strategy based on the relationship between the total number of orders and their distribution across hours and days of the week. The sum of orders across all hours and days should be equal. This logic held for most cases, except where “*HR_0”* had missing values. For these instances, we imputed “*HR_0”* as the difference between the total sum of orders from the other hourly and daily features in cases where this difference did not equal zero.

- ***Residual Missing Numerical Values:*** For the remaining values, we utilized the KNN Imputer as a fallback technique. Since this method requires scaled data, we carefully standardized the dataset before applying the imputation. The scaling process will be discussed in more detail in a subsequent section.

## 3.3 Feature Engineering

Arguably one of the most important steps in our work, this section was pivotal in shaping our final solution. The results were achieved after numerous experiments, including the creation and subsequent refinement of features. Building on insights from the EDA, we retained key features such as “*promo_used*”, “*total_money_spent*”, “*peak_hours*”, “*off_peak_hours*”, “*avg_spent_per_product*”, and “*different_CUI*”. Below, we provide a detailed overview of additional features we developed, each of which significantly impacted the progress and outcomes of the work that followed:

***Weekdays* and *Weekends*:** Instead of analyzing individual days of the week, we aggregated them into two broader categories: weekdays and weekends. Useful for pattern identification across periods.

**Cuisines:** The original dataset contained an excessive number of features, many of which exhibited high variability within similar categories, making clustering a hard task (Figure A1: Appendix A). To address this, we aggregated related cuisine features into broader categories based on the total money spent per cuisine (Figure A2: Appendix A). We conducted the following aggregations:

- ***CUI_Popular_Asian*** represents *“CUI_Japanese”, “CUI_Chinese”, “CUI_Thai”, “CUI_Indian”, “CUI_Noodle Dishes”.*
- ***CUI_Western*** combined American (“*CUI_American*”) and Italian (“*CUI_Italian*”) cuisines.
- ***CUI_Other_Cuisines*** combines a selection of cuisines like *“CUI_OTHER”, “CUI_Cafe”, “CUI_Beverages”, “CUI_Desserts”, “CUI_Healthy”, and “CUI_Chicken Dishes”,* making it the second most predominant feature just behind “*CUI_Asian*” (Figure A2: Appendix A), which in turn, will aid us in the clustering process.

This resulted in five primary cuisine types (“*CUI_Popular_Asian”, “CUI_Western”, “CUI_Other_Cuisines”, “CUI_Asian”, and “CUI_Street Food / Snacks”*).

**Cuisines proportions:** We calculated proportions for each remaining cuisine, where the expenditure for each cuisine is divided by the total money spent across all cuisines. This method highlights the relative share of each cuisine in the overall spending. Thus, we have 5 more features named *“proportion_CUI_popular_asian****”****, “proportion_CUI_Western****”****, “proportion_CUI_asian****”****, “proportion_CUI_other_cuisines****”*** *and “proportion_CUI_Street Food / Snacks****”.***

**Orders:** Several derived features were created from order-type related features, including **“***total_orders***”, “***avg_spent_per_order***”**, and **“***avg_products_per_order***”**. These represent, respectively, the total sum of all orders placed on a given day, the average amount spent per order (calculated as “*total_money_spent”* divided by “*total_orders”*), and the average number of products per order (calculated as “*product_count”* divided by “*total_orders”*).

**Customer Activity:** “*customer_time*” being the difference between “*last_order*” and “*first_order*” allowing us to measure the time span of customer activity.

### 3.4. Scaling

Scaling was a fundamental step in our analysis, during which several hypotheses were tested. Different versions of our final pipeline were explored using both the StandardScaler and MinMaxScaler (with a range of [0, 1]) (Figure A3: Appendix A). The following insights were drawn from this experimentation:

**MinMaxScaler:** MinMaxScaler was particularly effective for generating visualizations. When using the same clustering solution, we achieved visually appealing cluster representations with dimensionality reduction techniques such as T-SNE and UMAP. However, while the visualizations were satisfying, the clustering analysis itself was less effective. The resulting clusters showed limited variability, and there were no signs of apparent patterns and distinctions across cluster solutions.

**StandardScaler:** StandardScaler, on the other hand, provided more insightful clustering results. This choice was especially suitable due to its ability to handle outliers proportionately, ensuring that they were not excessively compressed (an issue we found with MinMaxScaler). By maintaining the relative scale of features, StandardScaler allowed us to preserve the interpretability of the dataset, which proved critical for meaningful cluster analysis.

### 3.5. Encoding - Profiling

We used One-Hot Encoding for categorical features, transforming each category into binary (0 and 1) vectors. This step was crucial for cluster profiling, enabling us to identify and interpret categorical patterns within each cluster.

## 4. Clustering

### 4.1. SOM - Insights

Before proceeding with clustering, we conducted a Self-Organizing Map (SOM) analysis of all numeric features to identify correlations and detect anomalies in our variables, aiding us in selecting perspectives. By analyzing each feature's heatmap, we observed its distribution and behavior. The darker red areas, representing higher feature values, allowed us to identify potential clusters of nodes exhibiting similar feature behavior.

Additionally, we created a U-Matrix (Unified Distance Matrix) visualization (Figure B1: Appendix B), which revealed that the distances between neighboring nodes in the trained SOM were mostly uniform, except for the bottom-middle red area, where the distances were notably higher. This provided insights into the topological structure of the data as it was mapped onto the SOM grid.

Lastly, we generated a HITS Matrix visualization (Figure B2: Appendix B) to examine the distribution of data points across the SOM nodes. From the hits map, we observe that the data points are unevenly distributed across the grid, with clusters of high density located in specific regions. For instance, the bottom-left, center, and far-right edges of the grid exhibit a concentration of nodes with a large number of data points. These dense clusters align with regions identified as potential groups during clustering.

## 4.2. Perspectives

We decided to focus our analysis on two distinct perspectives: value features and preference features. The value features consist of two variables: "*total_money_spent*" and "*avg_spent_per_product*". Features such as "*customer_time*", "*first_order*", and "*last_order*" were excluded due to their high variability and lack of clear clustering patterns in their SOM visualizations.

We included all cuisine proportions for the preference perspective, as each aggregation successfully captured distinct clusters. Additionally, we retained "*orders_from_chain*" in this perspective while excluding features such as "*different_CUI*", "*vendor_count*", "*total_orders*", and "*product_count*", as they were highly correlated with "*orders_from_chain*".

## 4.3. Outliers

We addressed the outliers in our perspective features using three methods: DBSCAN, DBSCAN + Manual, and the 99% Quantile approach.

1. **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**: This clustering algorithm groups data points based on density, identifying points in low-density regions as outliers. The eps parameter was determined using the K-Distance Graph (Figure B3: Appendix B), where the "elbow" point in the curve was identified as the optimal value for eps. Using this method, we removed 0.13% of the rows.
2. **DBSCAN + Manual Adjustment**: After applying DBSCAN, we manually removed outliers by analyzing the boxplots of each feature. While this manual approach introduces subjectivity, it enabled us to remove 2.77% of the rows.
3. **99% Quantile Method**: This method retains only the rows where all feature values fall below the 99th percentile threshold. Using this approach, we removed 2.69% of the rows.

To evaluate the effectiveness of these methods, we calculated the $R^2$ scores and created visualizations comparing the $R^2$ values for each clustering solution, focusing on both the value (Figure B4: Appendix B) and preference (Figure B5: Appendix B) variables. Our final decision was to retain the DBSCAN method, as it provided the best trade-off between $R^2$ and the amount of data removed. We considered that while removing more data tends to increase $R^2$, this method achieved a balance that preserved as much data as possible without compromising performance.

We then conducted the SOM analysis again, this time focusing exclusively on the value and preference variables separately, with the outliers removed. The individual feature heatmaps (Figure B6: Appendix B, Figure B7: Appendix B) revealed that the clusters became more distinct and well-defined following the removal of outliers.

## 4.4. Feature Selection

To analyze relationships between variables, Spearman correlation matrices were calculated for value and preference features, which are suitable for our non-parametric data. Heatmaps of these matrices revealed moderate correlations among value features (e.g., 0.41 between "*total_money_spent*" and "*avg_spent_per_product*"), indicating mostly associations without redundancy. In parallel, correlations among preference features were weaker, with no pairs exceeding the 0.75 threshold for high correlation.

### 4.5. K-means and Hierarchical Clustering

We began by evaluating various clustering methods to identify the optimal strategy for clustering both the value and preference features. The R² metric was calculated for different methods, including K-Means (using K-Means ++ initialization) and Hierarchical Clustering, using various linkage criteria (complete, average, single, and Ward's method). The resulting plots (Figures B8, B9: Appendix B) revealed that K-Means and Ward's method consistently achieved significantly higher R² values compared to the other approaches. By analyzing the 'elbow' in the plots, we determined that K-Means with 4 clusters is the most effective method for clustering the value features, while K-Means with 5 clusters is optimal for clustering the preference features.

Following that, we merged the two perspectives using Hierarchical clustering with Ward's linkage method. We generated a dendrogram (Figure B10: Appendix B) with two threshold lines to determine the optimal number of clusters.  When opting for 7 clusters instead of 6, the cluster with 2383 individuals (cluster 0 in Figure B12: Appendix B) is split into two groups: 2052 and 331 individuals (clusters 3 and 6 in Figure B11: Appendix B). Analyzing the 7-cluster plot (Figure B11: Appendix B), the main difference between clusters 6 and 3 lies in the average "*total_money_spent*" and average "*orders_from_chain*". While there are small changes in preferences, such as in Asian cuisine and in Street Food / Snacks, the split primarily isolates higher-spending customers. In the 6-cluster plot (Figure B12: Appendix B), we observe that Cluster 0 exhibits a higher average "*total_money_spent*" than Cluster 3 (in 7-cluster plot) but lower than Cluster 6 (in 7-cluster plot). This is expected, as merging Cluster 6 with Cluster 3 slightly increases the average "*total_money_spent*" of the resulting Cluster 0. Given the similarities in preferences and the small size of the new cluster—which contributes the least revenue in absolute terms—we decided to retain the 6-cluster configuration

We use the **KMeans + Hierarchical clustering approach** to combine the strengths of both methods. KMeans provides an efficient initial partitioning of the data into clusters, while Hierarchical clustering refines these clusters by analyzing relationships between their centroids, offering a detailed view of their structure through a dendrogram. This hybrid approach ensures computational efficiency, better alignment with the data's natural structure, and more meaningful segmentation.

## 5. Profiling

We will proceed with the profiling for each identified cluster. Refer to Figure B17: Appendix B for details on "*customer_region*", Figure B12: Appendix B or Figure B19: Appendix B for an overview of cluster profiling, Figure B15: Appendix B for the mean of unused metric features across clusters, and Figure B16: Appendix B for their sum. However, we recommend referring to *sections 4.* and *4.1.* of our notebook for better visualization and a clear understanding of the metric features distribution across our clusters.

**Cluster 0 - Selective Spenders** (2383 customers): This group consists of customers who predominantly spend on Street Food / Snacks, which accounts for 70% of their total money spent. They stand out for having the highest average spent per product, despite not achieving the highest average total money spent. This is primarily because their average number of orders is significantly lower compared to the group with the highest total spend (Cluster 2 - Revenue Powerhouses). These customers are particularly valuable as, while their purchase frequency is low, their orders tend to be large, as reflected by the average number of products per order, and they favor higher-priced items, as evidenced by their high average spent per product. Although they currently generate the least revenue for the business, they represent a highly promising segment due to their substantial spending per order. Regionally, most of these customers are concentrated in region 8670, offering a strategic focus area for targeted marketing efforts.

**Cluster 1 - The Average Joe** (8165 customers): This cluster has the largest number of customers, stands out for its preference for the aggregated category "Other Cuisines" (70% of the total money spent), and has a slightly higher number of orders from chains than other clusters besides Cluster 2 - Revenue Powerhouses. Within this category, most of their spending is concentrated in "Other" (25%) and "Beverages" (19%). Despite having the highest absolute number of orders, they are only the second-highest revenue-generating cluster for the company. This is due to their lower average spent per product and per order (the lowest across all clusters), which significantly reduces their average total money spent, even though their large size drives a high volume of orders. Analysing customer_region we see that almost half of the customers of this cluster are located in region 2360 and represent almost half of the people of this region.

**Cluster 2 - Revenue Powerhouses** (1945 customers): These customers are the highest spenders on average and stand out for their preference for chain restaurants (high *“orders_from_chain”*). Regarding cuisine types, their spending is fairly balanced across the different groups, with a slight preference for Asian cuisine (32% of their spending). Although they do not have the highest average spending per product, they are the group that generates the most revenue overall due to their high number of orders (*“total_orders”*) and relatively large order sizes (*“avg_products_per_order”*).
Additionally, this cluster has the smallest number of customers but brings in the largest amount of revenue, with the highest total_money_spent across all clusters. For that reason, we consider these our best customers. Interestingly enough, it appears that this group has the least promotion usage on their last order across groups. Most of the customers are located in regions 8670 and 4660.

**Cluster 3 - Western Preference Shoppers** (6389 customers): This relatively large cluster stands out for its preference for the "Western" aggregation (80% of their total spending), which includes American and Italian cuisines. A closer look reveals a slight preference for American cuisine (43.91%) compared to Italian (36.75%). These customers have the lowest average total money spent, which is justified by their reduced number of average orders combined with the lowest average spent per order. While only Cluster 4 - Occasional High Spenders on Asian Cuisine has fewer average orders, it still achieves a higher average total money spent due to a greater average spent per order. This explains why this cluster is the second-lowest revenue-generating group for the business, despite its relatively large size. The distribution of customers across regions stands out in region 4660.

**Cluster 4 - Occasional High Spenders on Asian Cuisine** (5999 customers): Among our clusters with 6000 to 7000 customers, this group stands out as the highest revenue generator, despite having the fewest customers. However, they are also the cluster with the lowest average number of orders. The revenue they generate is justified by their high average_spent_per_product, which indicates that they are valuable and promising customers for our business. If we can encourage them to increase their order frequency, their contribution to revenue could grow even further. This group also stands out by the preference for Asian cuisine, dominating their spending pattern (77.58%). The customers of this region are mostly located in region 8670.

**Cluster 5 - Popular Asian Cuisines Enthusiasts** (6751 customers): This cluster stands out for its preference for popular Asian cuisines (73% of their total money spent) aggregation, including Japanese, Chinese, Thai, Indian, and Noodle Dishes. Among these, Japanese cuisine (26% of their spending), Indian cuisine (17%), and Chinese cuisine (13%) are the most prominent. While this cluster has the second-lowest average total money spent, it generates more revenue than Cluster 0 - Selective Spenders due to its significantly larger size. The distribution of customers across the different regions stands out mainly in regions 2360 and 4660.

## 5.1. Visualization

This section was crucial in confirming the robustness of the clustering and profiling process. The primary visualization tool used was t-SNE, a statistical method for visualizing high-dimensional data. As shown in *Figure B13: Appendix B*, six distinct clusters are evident, primarily separated and concentrated in specific regions of the plot, with minor overlaps potentially indicating similar patterns or behaviors.

Additionally, we applied UMAP, another dimensionality reduction technique that often preserves global relationships better than t-SNE *(Figure B14: Appendix B)*. However, in this case, t-SNE provided a clearer and more accurate visualization of the clusters.

# 6. Assessing Feature Importance and Reclassifying Outliers

### Feature Importance

To better understand cluster separation and refine our analysis, we employed two techniques: $R^2$ evaluation and a decision tree classifier. We calculated the $R^2$ for each variable to measure the proportion of total variance explained by the clusters. Cuisine-related features, such as "*proportion_CUI_Street_Food / Snacks*" (0.86) and "*proportion_CUI_popular_asian*" (0.78) showed higher $R^2$ values, indicating their significant role in cluster separation.

A Decision Tree Classifier, trained using perspective features, achieved 82.74% accuracy on the test set, demonstrating its effectiveness in classifying customers into their clusters. Feature importance analysis revealed that the most impactful features for cluster prediction were the cuisine-related ones, especially the aggregations "*Other*", "*Western*" and "*Popular Asian*". Using the Gini Index as the criterion for splits, the decision tree visualization (Figure B18: Appendix B) highlights these features at the top levels of the tree, emphasizing their role in cluster separation.

### Outlier Reclassification

The decision tree mentioned above was used to predict the cluster labels for the identified outliers, ensuring these customers were analyzed. These customers exhibited extremely high values in variables such as “*total_money_spent”* and “*orders_from_chain”* (Figure B20: Appendix B, FigureB21: Appendix B), contributing 385.75 monetary units. While relevant, this represents a minimal fraction of the overall impact in a dataset containing more than 31000 customers, further reinforcing the exceptional nature of these cases.

Approximately 40 outliers (representing only 0.13%) were removed to maintain the consistency and representativeness of the main clusters. We acknowledge, however, that these outliers may represent a special segment of consumers, such as high-value or VIP customers. This decision was based on the need to prioritize creating representative and useful clusters for the majority of customers without compromising the robustness of the analysis or the relevance of key variables such as “*total_money_spent”*. Despite the removal of outliers, the post-removal analysis, as we saw before, demonstrated that we were still able to identify groups representing high-value customers, such as Cluster 2- Revenue Powerhouses.

# 7. Business Strategies

Based on the cluster profiles, we propose targeted strategies to improve customer engagement and profitability. While these strategies are tailored to each cluster, they remain flexible and can be adapted as needed. Most clusters exhibit clear preferences for specific cuisines and are concentrated in a few key regions, which means marketing efforts should prioritize these areas and cuisines to maximize impact.

**Cluster-Specific Strategies**

**Selective Spenders (Cluster 0):** Although this cluster generates the least revenue, its high average spending per product makes it a promising segment and the focus should be on capitalizing their high spending per order. Campaigns should emphasize Street Food and Snacks to encourage more frequent orders through, for example, free delivery for orders exceeding a certain value, or encouraging larger order sizes by offering discounts on bulk purchases. Targeting region 8670, where most of these customers are located, could further enhance their impact.

**The Average Joe (Cluster 1):** Focus on boosting average spending and increasing engagement**.** As the largest cluster, these customers represent a significant share of the customer base. While their spending per order is relatively low, their volume of orders is critical for sustaining revenue. Regional campaigns targeting 2360 and promotions for "Other" and "Beverages" cuisines should be prioritized. Additionally, low-cost incentives to increase order size could improve profitability**.**

**Revenue Powerhouses (Cluster 2)**: These are the highest-value customers, generating the most revenue despite being the smallest cluster. Retention should be the primary goal. Idea would be to launch premium loyalty programs (e.g., "ABCDEats Elite") offering benefits such as discounts for high monthly spending, free delivery, or priority customer service. Any experimental engagement strategies should be meticulously tested to ensure they enhance customer satisfaction and loyalty without risking attrition.

**Western Preference Shoppers** (Cluster 3): Focus on increasing order frequency and average spending. Diversifying the Western cuisine offerings or introducing promotions for complementary items could boost engagement. Regional campaigns in 4660, where this cluster is mainly concentrated, should accompany these efforts.

**Occasional High Spenders on Asian Cuisine** (Cluster 4): Focus should be to convert occasional spenders into frequent customers. This group has the lowest order frequency but the highest average spending per product. Strategies should aim to slightly increase order frequency by emphasizing Asian cuisine (not the popular ones aggregation) and offering targeted promotions or discounts. Regional campaigns focused on 8670 could further enhance engagement.

**Popular Asian Cuisines Enthusiasts** (Cluster 5): This cluster generates moderate revenue despite having the second-lowest average total spending. Their strong preference for Asian cuisines, particularly Japanese, Indian, and Chinese presents an opportunity for campaigns to highlight variety and premium offerings. Regional strategies targeting 2360 and 4660 will be essential.

# 8. Conclusion

As consultants for ABCDeats Inc., our primary objective was to segment customer groups from a standard dataset and develop targeted business strategies tailored to their behaviors and patterns to enhance profitability.

To achieve this, a robust and technical approach was necessary. A coherent exploratory data analysis (EDA) formed the foundation of our work, enabling us to clean the dataset and engineer new features that improved the explainability of patterns and behaviors. Advanced techniques were employed, including DBSCAN for outlier removal, Self-Organizing Maps for visualization and pattern analysis and K-Means & Hierarchical Clustering for segmentation.

These methods provided a scientific and systematic framework for cluster analysis and profiling. Through iterative experimentation, we successfully profiled customer groups, identifying distinct consumption behaviors in terms of cuisine preferences and purchase frequencies. These insights became the backbone of our business strategy development.

Some strategies were generalized, and applicable across multiple groups, focusing on increasing order frequency and unlocking untapped potential. Other strategies were tailored to specific clusters, requiring a careful, customized approach to complement and enhance consumption behaviors.

While our analysis was comprehensive, we acknowledge the inherent assumptions and ambiguities in this type of work. Alternative feature engineering approaches or additional preprocessing methods, such as scaling and imputation, could potentially yield even more powerful insights. Additionally, exploring other techniques for clustering or visualization might further refine the results and strategies. However, through this extensive and robust pipeline, the results demonstrate potential and significant impact in this sector, ultimately, the project highlights the power of data-driven decision-making in developing strategies that align with business goals, ensuring a continuous harmony between technological growth and business strategies.

# Appendix A – Preprocessing

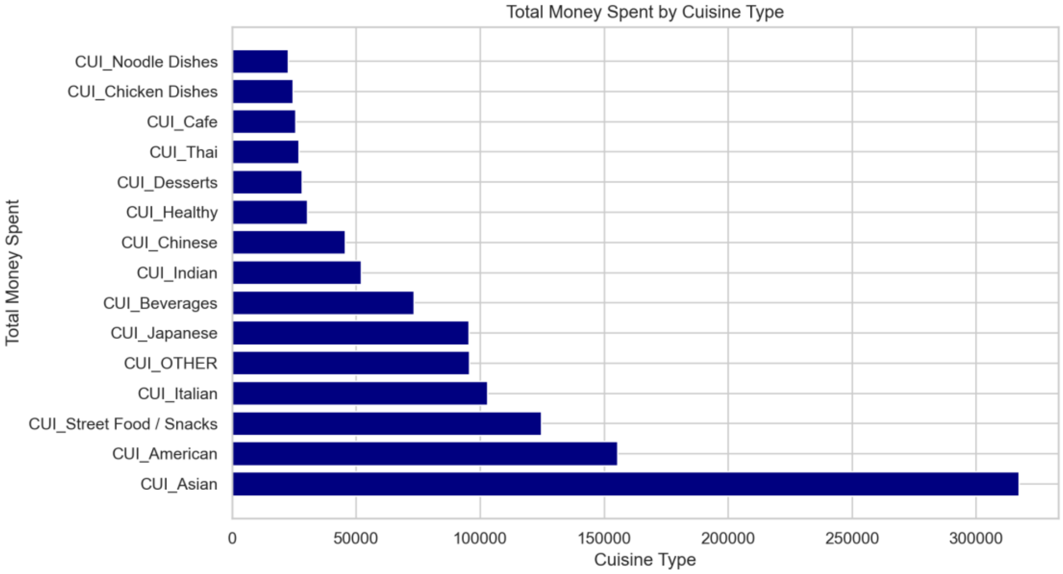


Figure A1: Total money spent per cuisine

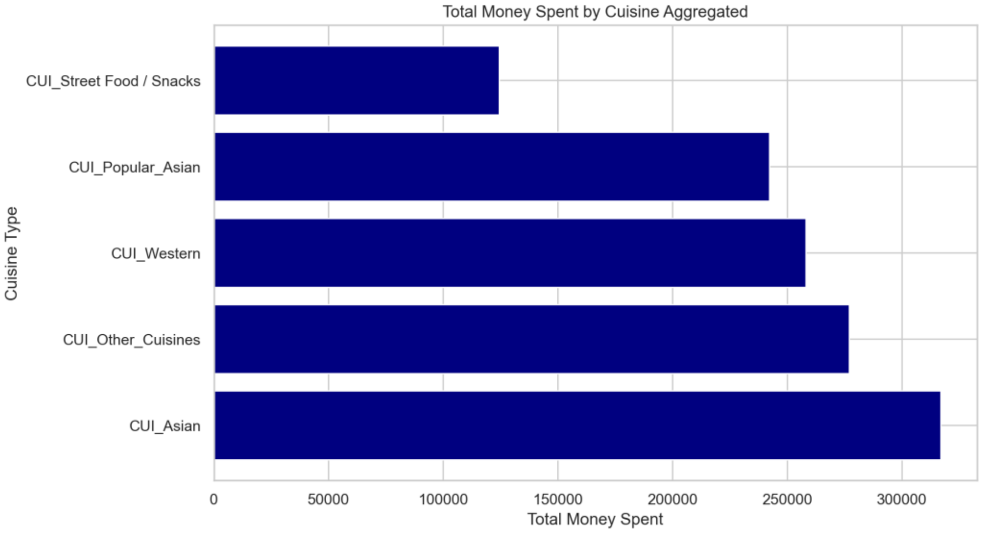


Figure A2: Total money spent per cuisine after aggregation

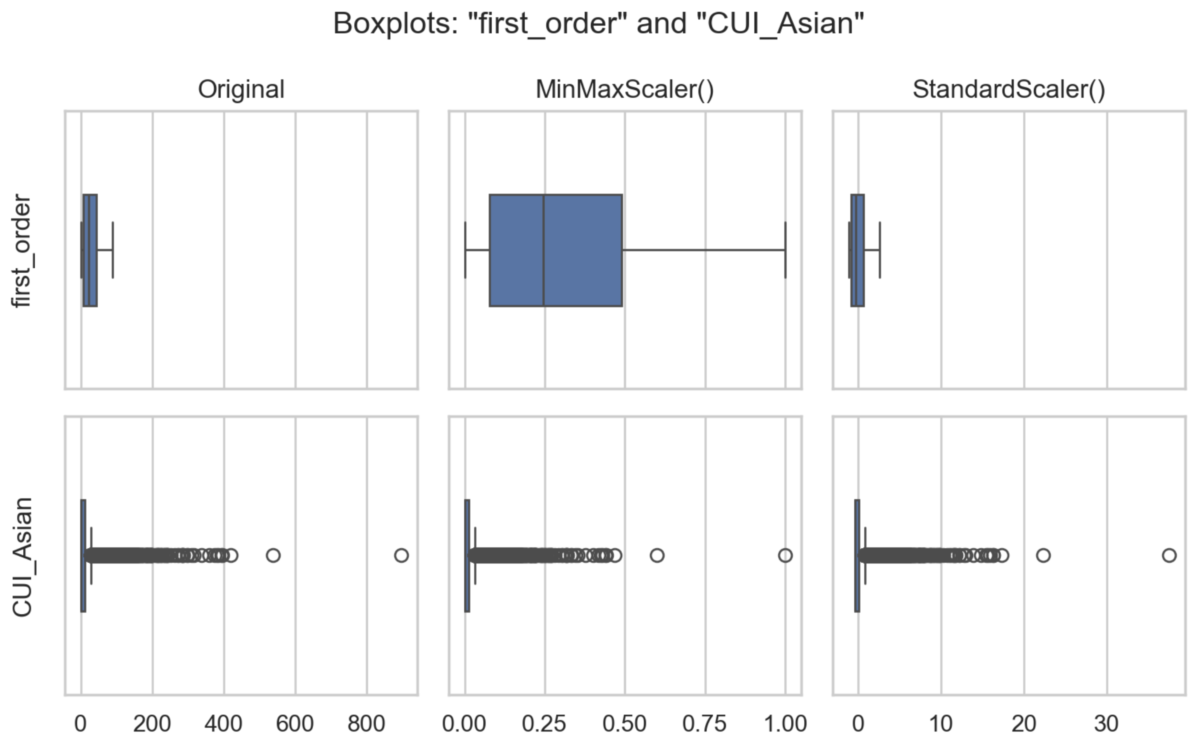


Figure A3: Boxplots ‘first_order’ and ‘CUI_Asian’ to compare scaling methods

# Appendix B - Clustering and Profiling

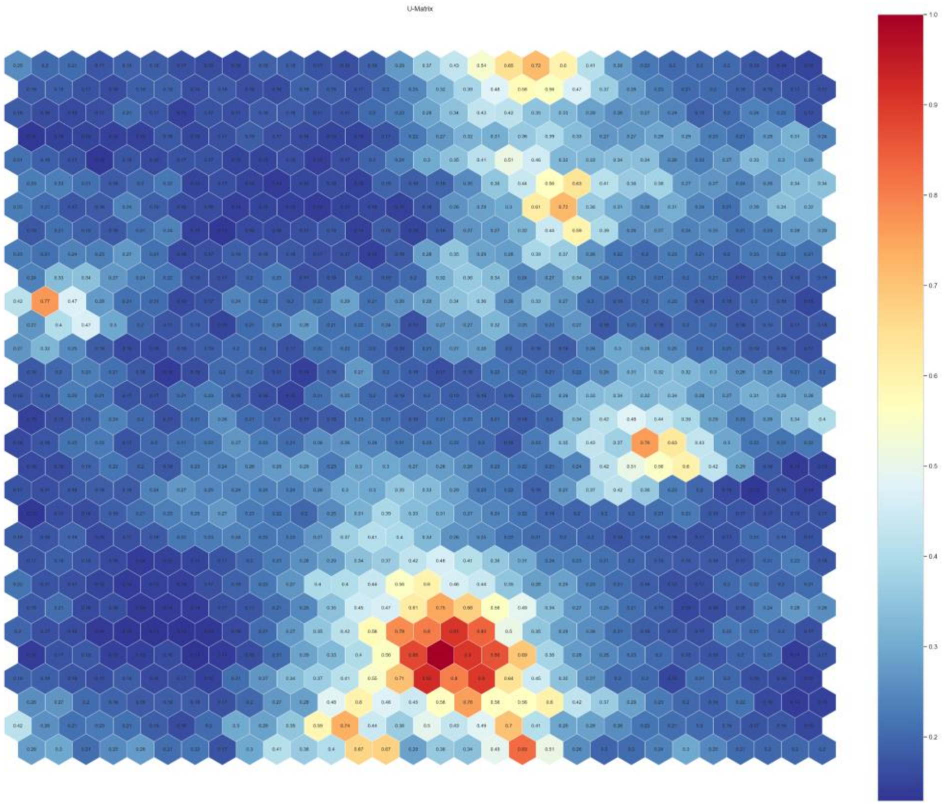


Figure B1: SOM U-Matrix

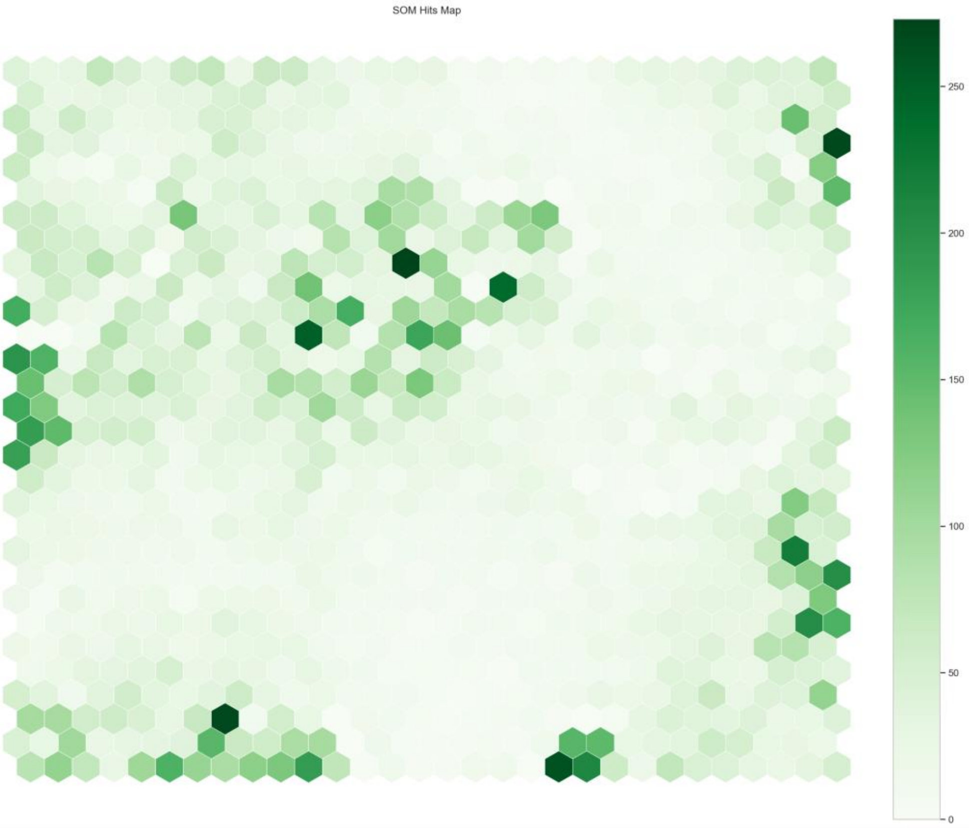


Figure B2: SOM Hits Map

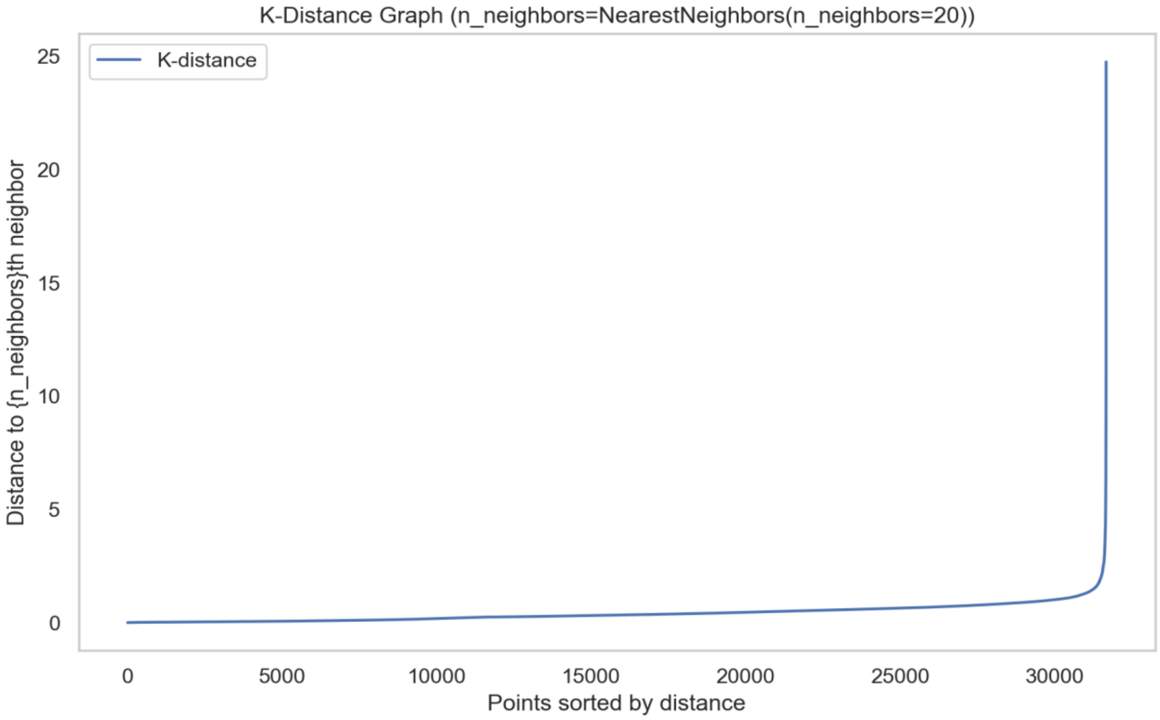


Figure B3: K-Distance Graph to find the optimal eps

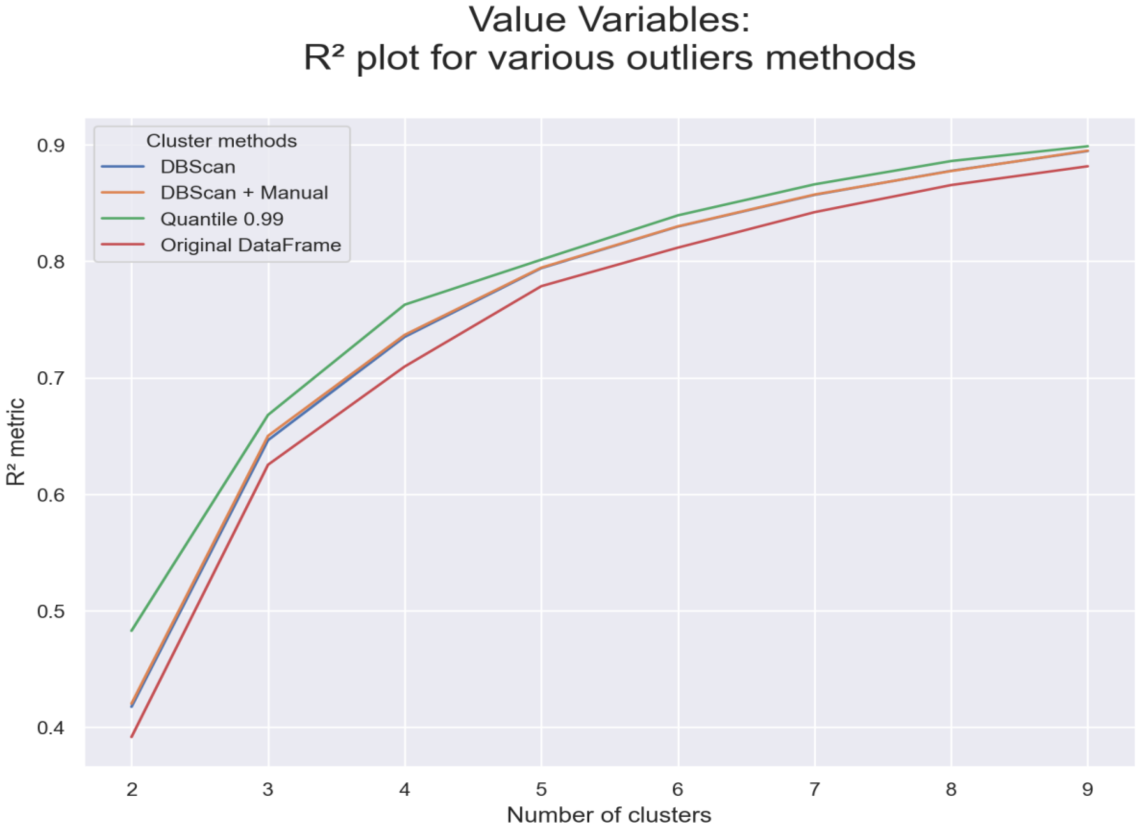


Figure B4: R$^2$ Scores for value features

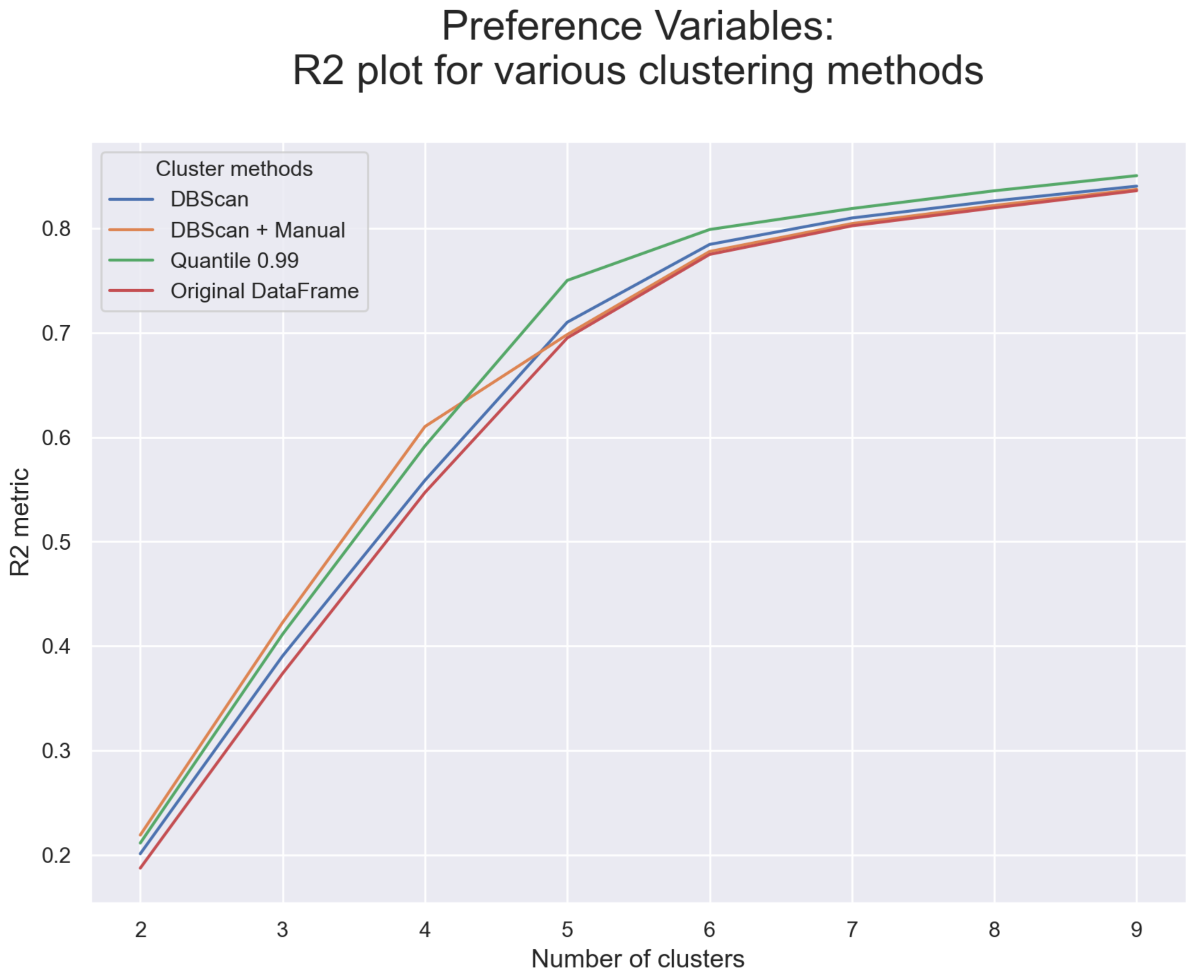


Figure B5: R² Scores for preference features

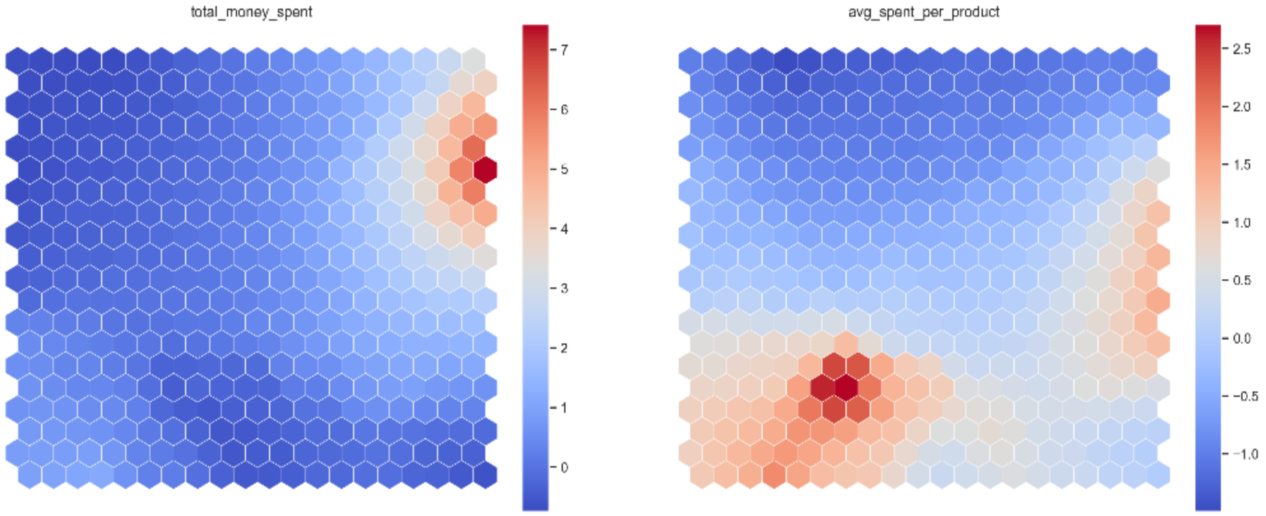


Figure B6: Heatmaps for Value Features

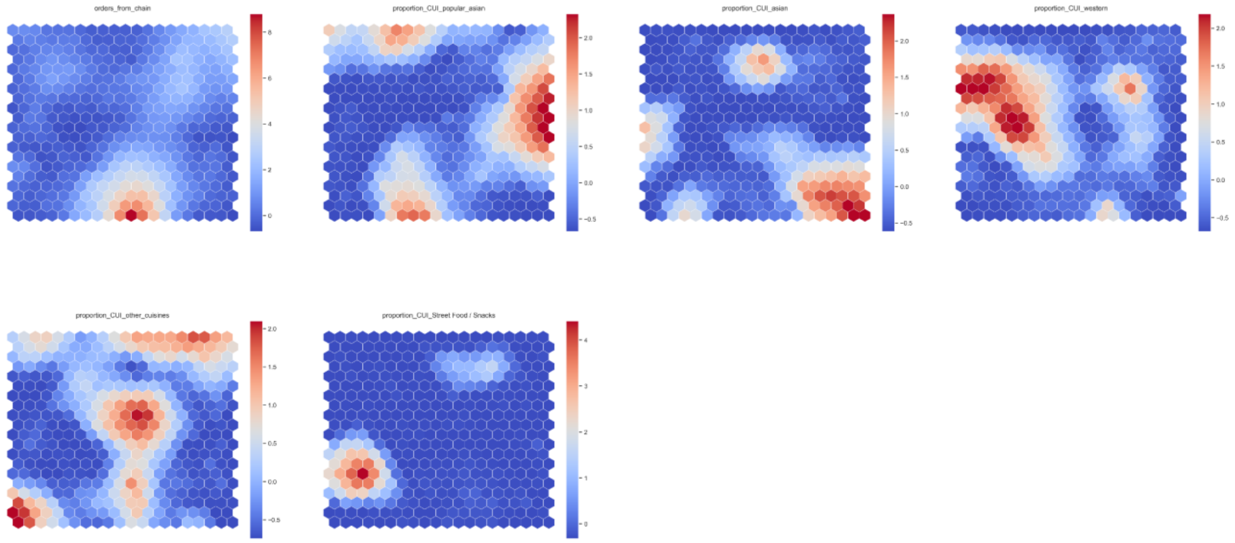


Figure B7: Heatmaps for Preference Features

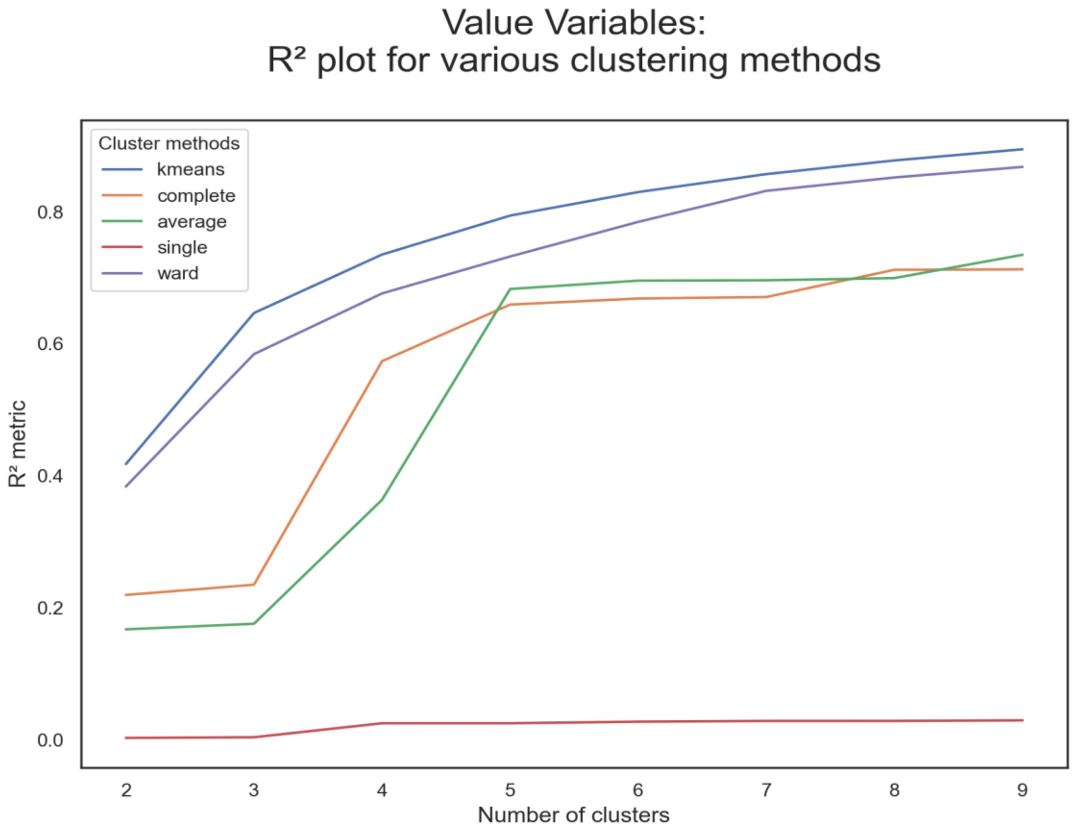


Figure B8: Cluster methods and number of clusters for value features

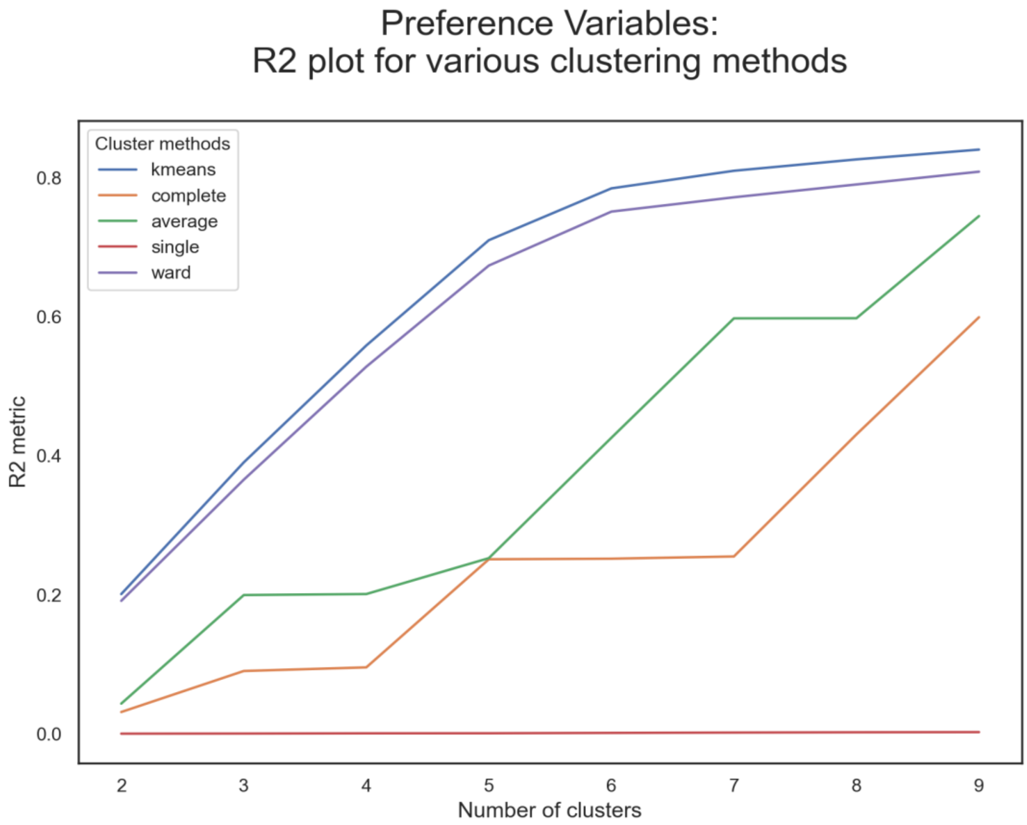


Figure B9: Cluster methods and number of clusters for preference features

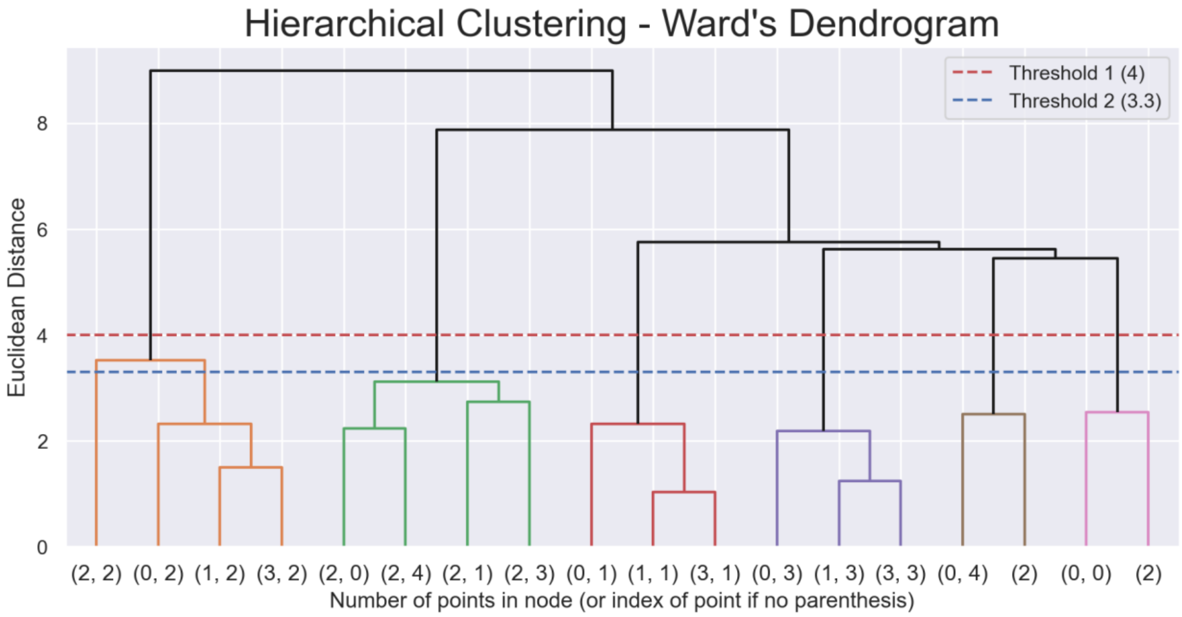


Figure B10: Dendrogram of Hierarchical Clustering

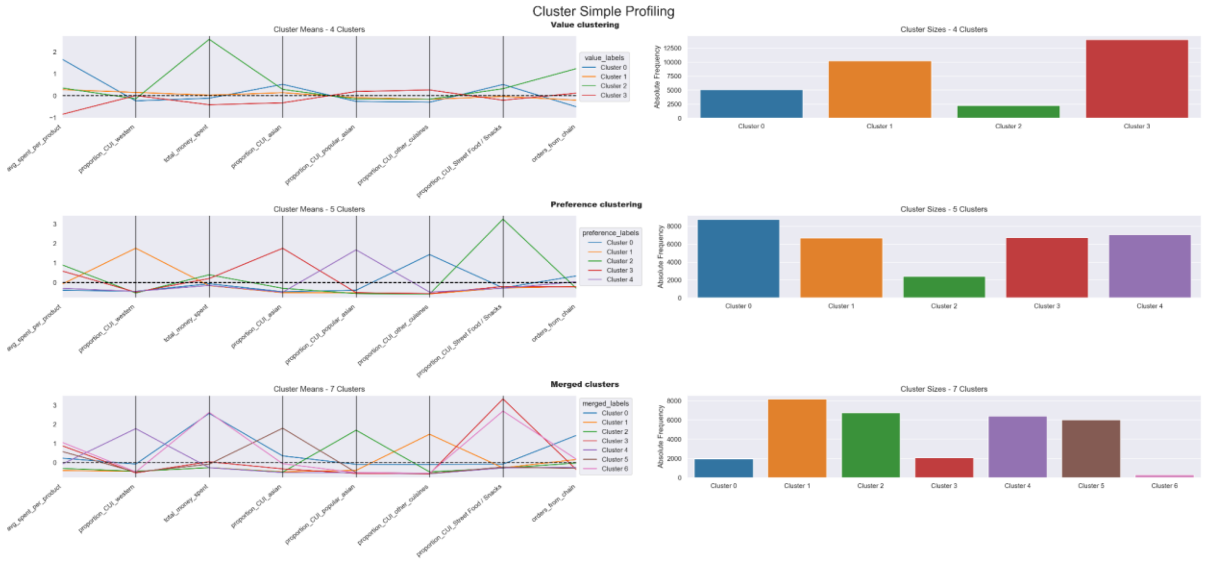


Figure B11: Cluster Profiling considering 7 clusters

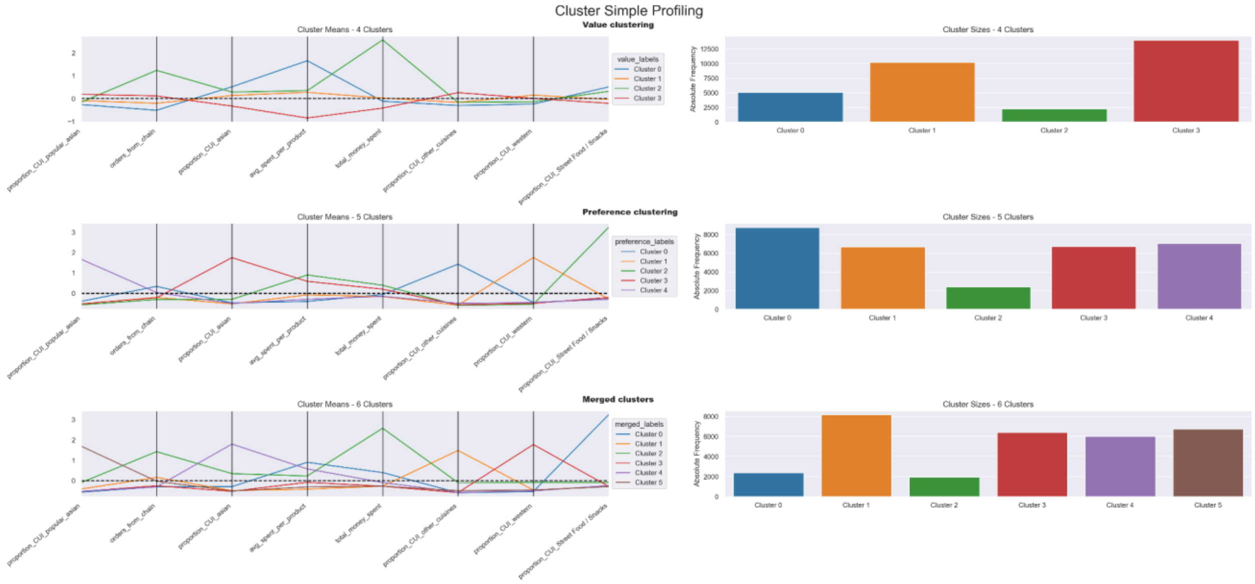


Figure B12: Cluster Profiling considering 6 clusters (final choice)

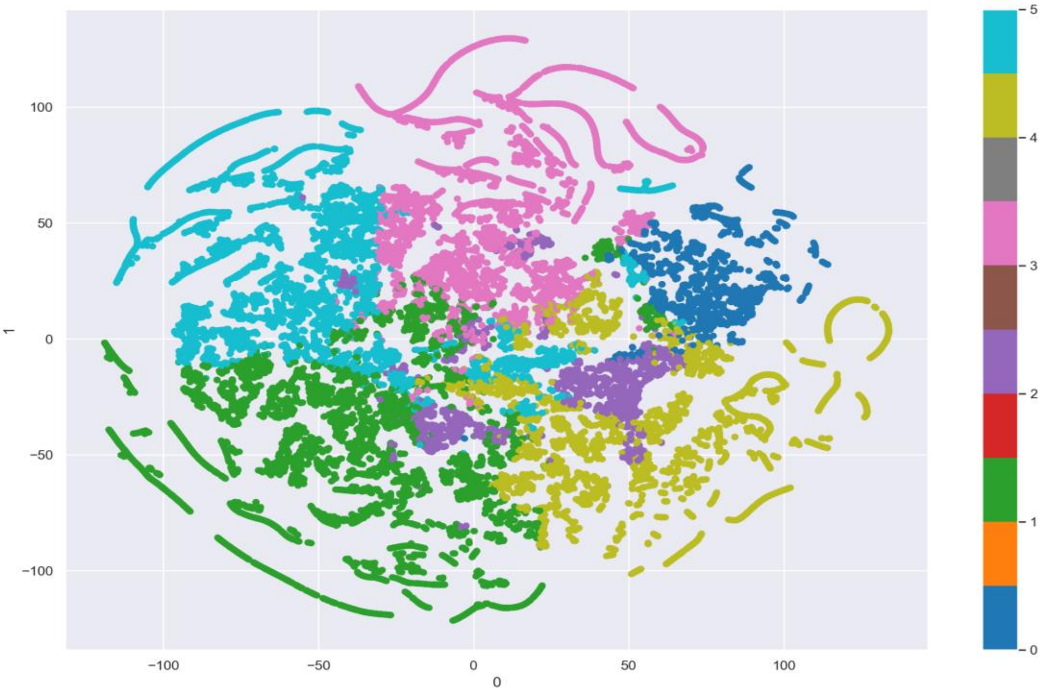


Figure B13: t-SNE Cluster Visualization

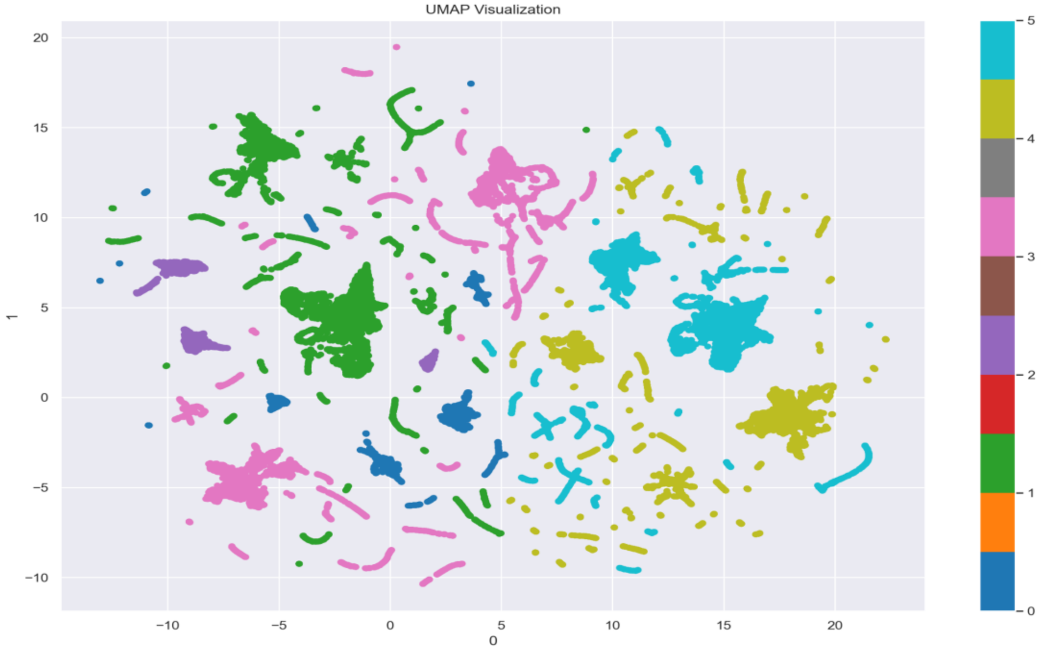
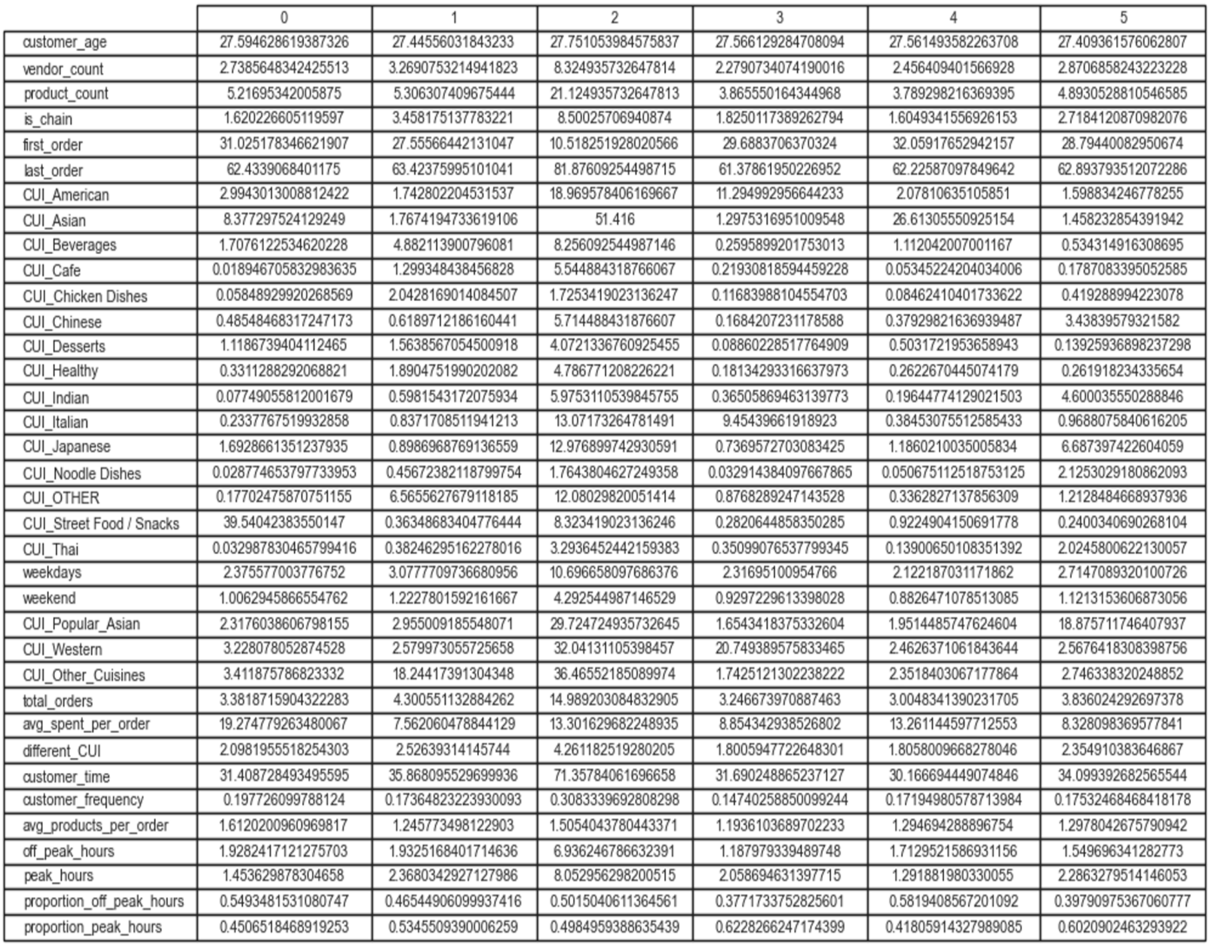
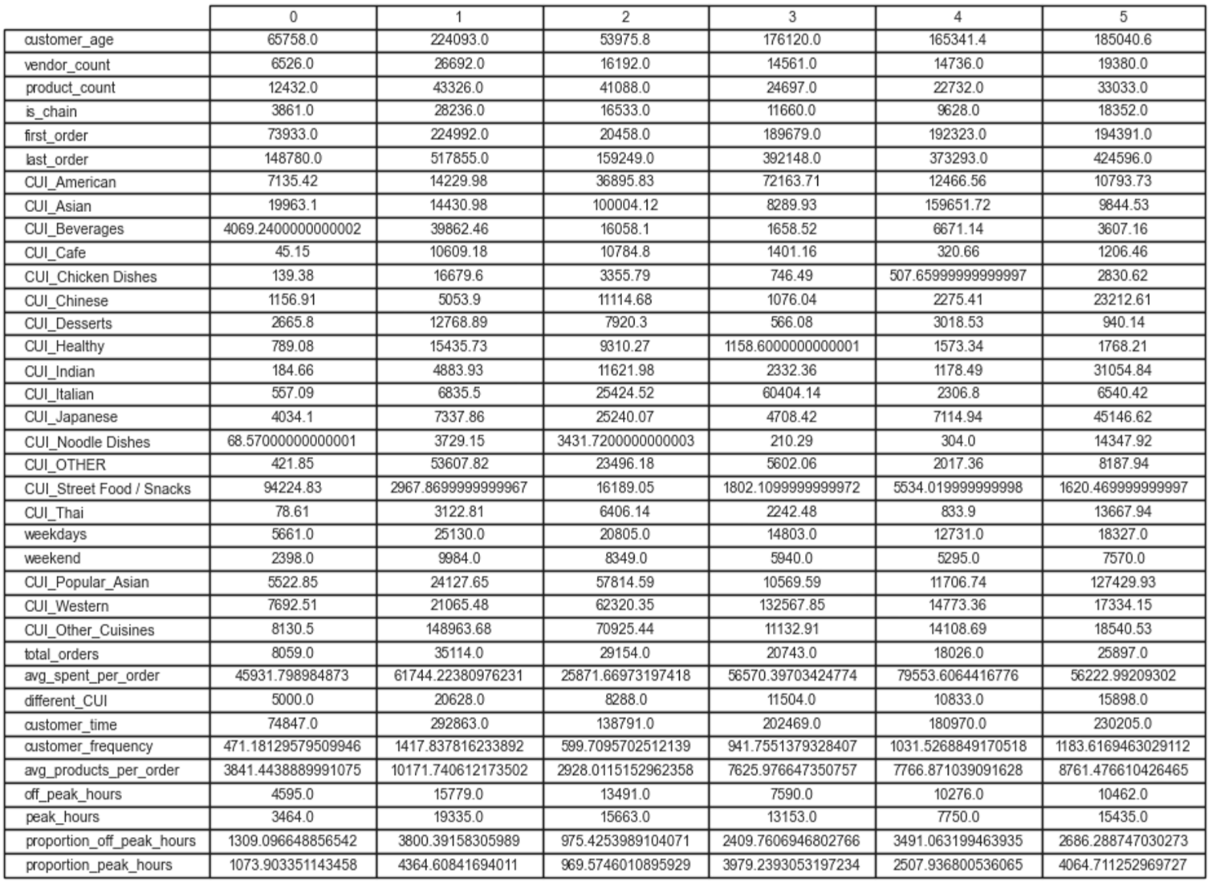


Figure B14: UMAP Cluster Visualization



| | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| customer_age | 27.594628619387326 | 27.44556031843233 | 27.751053984575837 | 27.566129284708094 | 27.561493582263708 | 27.409361576062807 |
| vendor_count | 2.7385648342425513 | 3.2690753214941823 | 8.324935732647814 | 2.2790734074190016 | 2.456409401566928 | 2.8706858243223228 |
| product_count | 5.21695342005875 | 5.306307409675444 | 21.124935732647813 | 3.865550164344968 | 3.789298216369395 | 4.8930528810546585 |
| is_chain | 1.620226605119597 | 3.458175137783221 | 8.50025706940874 | 1.8250117389262794 | 1.6049341556926153 | 2.7184120870982076 |
| first_order | 31.025178346621907 | 27.55566442131047 | 10.518251928020566 | 29.6883706370324 | 32.05917652942157 | 28.79440082950674 |
| last_order | 62.4339068401175 | 63.42375995101041 | 81.87609254498715 | 61.37861950226952 | 62.22587097849642 | 62.893793512072286 |
| CUI_American | 2.9943013008812422 | 1.742802204531537 | 18.969578406169667 | 11.294992956644233 | 2.07810635105851 | 1.598834246778255 |
| CUI_Asian | 8.377297524129249 | 1.7674194733619106 | 51.416 | 1.2975316951009548 | 26.61305550925154 | 1.458232854391942 |
| CUI_Beverages | 1.7076122534620228 | 4.882113900796081 | 8.256092544987146 | 0.2595899201753013 | 1.112042007001167 | 0.534314916308695 |
| CUI_Cafe | 0.018946705832983635 | 1.299348438456828 | 5.544884318766067 | 0.21930818594459228 | 0.05345224204034006 | 0.1787083395052585 |
| CUI_Chicken Dishes | 0.05848929920268569 | 2.0428169014084507 | 1.7253419023136247 | 0.11683988104554703 | 0.08462410401733622 | 0.419288994223078 |
| CUI_Chinese | 0.48548468317247173 | 0.6189712186160441 | 5.714488431876607 | 0.1684207231178588 | 0.37929821636939487 | 3.43839579321582 |
| CUI_Desserts | 1.1186739404112465 | 1.5638567054500918 | 4.0721336760925455 | 0.08860228517764909 | 0.5031721953658943 | 0.13925936898237298 |
| CUI_Healthy | 0.3311288292068821 | 1.8904751990202082 | 4.786771208226221 | 0.18134293316637973 | 0.2622670445074179 | 0.261918234335654 |
| CUI_Indian | 0.07749055812001679 | 0.5981543172075934 | 5.9753110539845755 | 0.36505869463139773 | 0.19644774129021503 | 4.600035550288846 |
| CUI_Italian | 0.2337767519932858 | 0.8371708511941213 | 13.07173264781491 | 9.45439661918923 | 0.38453075512585433 | 0.9688075840616205 |
| CUI_Japanese | 1.6928661351237935 | 0.8986968769136559 | 12.976899742930591 | 0.7369572703083425 | 1.1860210035005834 | 6.687397422604059 |
| CUI_Noodle Dishes | 0.028774653797733953 | 0.45672382118799754 | 1.7643804627249358 | 0.032914384097667865 | 0.050675112518753125 | 2.1253029180862093 |
| CUI_OTHER | 0.17702475870751155 | 6.5655627679118185 | 12.08029820051414 | 0.8768289247143528 | 0.3362827137856309 | 1.2128484668937936 |
| CUI_Street Food / Snacks | 39.54042383550147 | 0.36348683404776444 | 8.323419023136246 | 0.2820644858350285 | 0.9224904150691778 | 0.2400340690268104 |
| CUI_Thai | 0.032987830465799416 | 0.38246295162278016 | 3.2936452442159383 | 0.35099076537799345 | 0.13900650108351392 | 2.0245800622130057 |
| weekdays | 2.375577003776752 | 3.0777709736680956 | 10.696658097686376 | 2.31695100954766 | 2.122187031171862 | 2.7147089320100726 |
| weekend | 1.0062945866554762 | 1.2227801592161667 | 4.292544987146529 | 0.9297229613398028 | 0.8826471078513085 | 1.1213153606873056 |
| CUI_Popular_Asian | 2.3176038606798155 | 2.955009185548071 | 29.724724935732645 | 1.6543418375332604 | 1.9514485747624604 | 18.875711746407937 |
| CUI_Western | 3.228078052874528 | 2.579973055725658 | 32.04131105398457 | 20.749389575833465 | 2.4626371061843644 | 2.5676418308398756 |
| CUI_Other_Cuisines | 3.411875786823332 | 18.24417391304348 | 36.46552185089974 | 1.7425121302238222 | 2.3518403067177864 | 2.746338320248852 |
| total_orders | 3.3818715904322283 | 4.300551132884262 | 14.989203084832905 | 3.246673970887463 | 3.0048341390231705 | 3.836024292697378 |
| avg_spent_per_order | 19.274779263480067 | 7.562060478844129 | 13.301629682248935 | 8.854342938526802 | 13.261144597712553 | 8.328098369577841 |
| different_CUI | 2.0981955518254303 | 2.52639314145744 | 4.261182519280205 | 1.8005947722648301 | 1.8058009668278046 | 2.354910383646867 |
| customer_time | 31.408728493495595 | 35.868095529699936 | 71.35784061696658 | 31.690248865237127 | 30.166694449074846 | 34.099392682565544 |
| customer_frequency | 0.197726099788124 | 0.17364823223930093 | 0.3083339692808298 | 0.14740258850099244 | 0.17194980578713984 | 0.17532468468418178 |
| avg_products_per_order | 1.6120200960969817 | 1.245773498122903 | 1.5054043780443371 | 1.1936103689702233 | 1.294694288896754 | 1.2978042675790942 |
| off_peak_hours | 1.9282417121275703 | 1.9325168401714636 | 6.936246786632391 | 1.187979339489748 | 1.7129521586931156 | 1.549696341282773 |
| peak_hours | 1.453629878304658 | 2.3680342927127986 | 8.052956298200515 | 2.058694631397715 | 1.291881980330055 | 2.2863279514146053 |
| proportion_off_peak_hours | 0.5493481531080747 | 0.46544906099937416 | 0.5015040611364561 | 0.3771733752825601 | 0.5819408567201092 | 0.39790975367060777 |
| proportion_peak_hours | 0.4506518468919253 | 0.5345509390006259 | 0.4984959388635439 | 0.6228266247174399 | 0.41805914327989085 | 0.6020902463293922 |

Figure B15: Mean values of unused metric features across identified clusters



| | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| customer_age | 65758.0 | 224093.0 | 53975.8 | 176120.0 | 165341.4 | 185040.6 |
| vendor_count | 6526.0 | 26692.0 | 16192.0 | 14561.0 | 14736.0 | 19380.0 |
| product_count | 12432.0 | 43326.0 | 41088.0 | 24697.0 | 22732.0 | 33033.0 |
| is_chain | 3861.0 | 28236.0 | 16533.0 | 11660.0 | 9628.0 | 18352.0 |
| first_order | 73933.0 | 224992.0 | 20458.0 | 189679.0 | 192323.0 | 194391.0 |
| last_order | 148780.0 | 517855.0 | 159249.0 | 392148.0 | 373293.0 | 424596.0 |
| CUI_American | 7135.42 | 14229.98 | 36895.83 | 72163.71 | 12466.56 | 10793.73 |
| CUI_Asian | 19963.1 | 14430.98 | 100004.12 | 8289.93 | 159651.72 | 9844.53 |
| CUI_Beverages | 4069.2400000000002 | 39862.46 | 16058.1 | 1658.52 | 6671.14 | 3607.16 |
| CUI_Cafe | 45.15 | 10609.18 | 10784.8 | 1401.16 | 320.66 | 1206.46 |
| CUI_Chicken Dishes | 139.38 | 16679.6 | 3355.79 | 746.49 | 507.65999999999997 | 2830.62 |
| CUI_Chinese | 1156.91 | 5053.9 | 11114.68 | 1076.04 | 2275.41 | 23212.61 |
| CUI_Desserts | 2665.8 | 12768.89 | 7920.3 | 566.08 | 3018.53 | 940.14 |
| CUI_Healthy | 789.08 | 15435.73 | 9310.27 | 1158.6000000000001 | 1573.34 | 1768.21 |
| CUI_Indian | 184.66 | 4883.93 | 11621.98 | 2332.36 | 1178.49 | 31054.84 |
| CUI_Italian | 557.09 | 6835.5 | 25424.52 | 60404.14 | 2306.8 | 6540.42 |
| CUI_Japanese | 4034.1 | 7337.86 | 25240.07 | 4708.42 | 7114.94 | 45146.62 |
| CUI_Noodle Dishes | 68.57000000000001 | 3729.15 | 3431.7200000000003 | 210.29 | 304.0 | 14347.92 |
| CUI_OTHER | 421.85 | 53607.82 | 23496.18 | 5602.06 | 2017.36 | 8187.94 |
| CUI_Street Food / Snacks | 94224.83 | 2967.8699999999967 | 16189.05 | 1802.1099999999972 | 5534.019999999998 | 1620.469999999997 |
| CUI_Thai | 78.61 | 3122.81 | 6406.14 | 2242.48 | 833.9 | 13667.94 |
| weekdays | 5661.0 | 25130.0 | 20805.0 | 14803.0 | 12731.0 | 18327.0 |
| weekend | 2398.0 | 9984.0 | 8349.0 | 5940.0 | 5295.0 | 7570.0 |
| CUI_Popular_Asian | 5522.85 | 24127.65 | 57814.59 | 10569.59 | 11706.74 | 127429.93 |
| CUI_Western | 7692.51 | 21065.48 | 62320.35 | 132567.85 | 14773.36 | 17334.15 |
| CUI_Other_Cuisines | 8130.5 | 148963.68 | 70925.44 | 11132.91 | 14108.69 | 18540.53 |
| total_orders | 8059.0 | 35114.0 | 29154.0 | 20743.0 | 18026.0 | 25897.0 |
| avg_spent_per_order | 45931.798984873 | 61744.22380976231 | 25871.66973197418 | 56570.39703424774 | 79553.6064416776 | 56222.99209302 |
| different_CUI | 5000.0 | 20628.0 | 8288.0 | 11504.0 | 10833.0 | 15898.0 |
| customer_time | 74847.0 | 292863.0 | 138791.0 | 202469.0 | 180970.0 | 230205.0 |
| customer_frequency | 471.18129579509946 | 1417.837816233892 | 599.7095702512139 | 941.7551379328407 | 1031.5268849170518 | 1183.6169463029112 |
| avg_products_per_order | 3841.4438889991075 | 10171.740612173502 | 2928.0115152962358 | 7625.976647350757 | 7766.871039091628 | 8761.476610426465 |
| off_peak_hours | 4595.0 | 15779.0 | 13491.0 | 7590.0 | 10276.0 | 10462.0 |
| peak_hours | 3464.0 | 19335.0 | 15663.0 | 13153.0 | 7750.0 | 15435.0 |
| proportion_off_peak_hours | 1309.096648856542 | 3800.39158305989 | 975.4253989104071 | 2409.7606946802766 | 3491.063199463935 | 2686.288747030273 |
| proportion_peak_hours | 1073.903351143458 | 4364.60841694011 | 969.5746010895929 | 3979.2393053197234 | 2507.936800536065 | 4064.711252969727 |

Figure B16: Sum of unused metric features across identified clusters

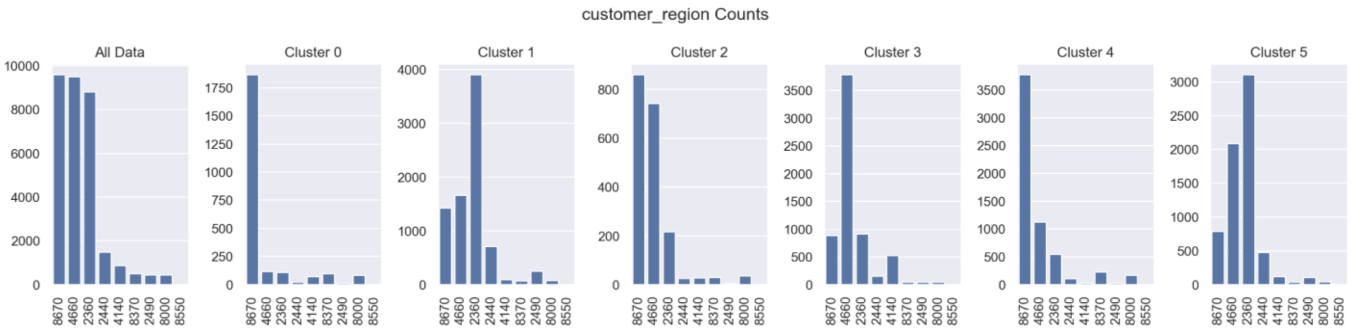


Figure B17: customer_region distribution on all the data and across the clusters

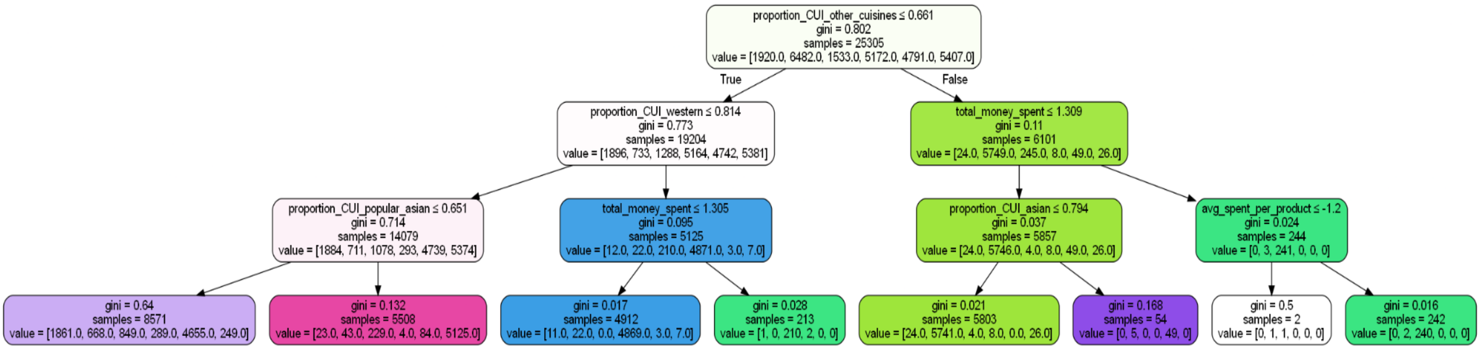

proportion_CUI_other_cuisines ≤ 0.661
gini = 0.802
samples = 25305
value = [1920.0, 6482.0, 1533.0, 5172.0, 4791.0, 5407.0]
True
False
proportion_CUI_western ≤ 0.814
gini = 0.773
samples = 19204
value = [1896, 733, 1288, 5164, 4742, 5381]
total_money_spent ≤ 1.309
gini = 0.11
samples = 6101
value = [24.0, 5749.0, 245.0, 8.0, 49.0, 26.0]
proportion_CUI_popular_asian ≤ 0.651
gini = 0.714
samples = 14079
value = [1884, 711, 1078, 293, 4739, 5374]
total_money_spent ≤ 1.305
gini = 0.095
samples = 5125
value = [12.0, 22.0, 210.0, 4871.0, 3.0, 7.0]
proportion_CUI_asian ≤ 0.794
gini = 0.037
samples = 5857
value = [24.0, 5746.0, 4.0, 8.0, 49.0, 26.0]
avg_spent_per_product ≤ -1.2
gini = 0.024
samples = 244
value = [0, 3, 241, 0, 0, 0]
gini = 0.64
samples = 8571
value = [1861.0, 668.0, 849.0, 289.0, 4655.0, 249.0]
gini = 0.132
samples = 5508
value = [23.0, 43.0, 229.0, 4.0, 84.0, 5125.0]
gini = 0.017
samples = 4912
value = [11.0, 22.0, 0.0, 4869.0, 3.0, 7.0]
gini = 0.028
samples = 213
value = [1, 0, 210, 2, 0, 0]
gini = 0.021
samples = 5803
value = [24.0, 5741.0, 4.0, 8.0, 0.0, 26.0]
gini = 0.168
samples = 54
value = [0, 5, 0, 0, 49, 0]
gini = 0.5
samples = 2
value = [0, 1, 1, 0, 0, 0]
gini = 0.016
samples = 242
value = [0, 2, 240, 0, 0, 0]

Figure B18: Decision tree visualization

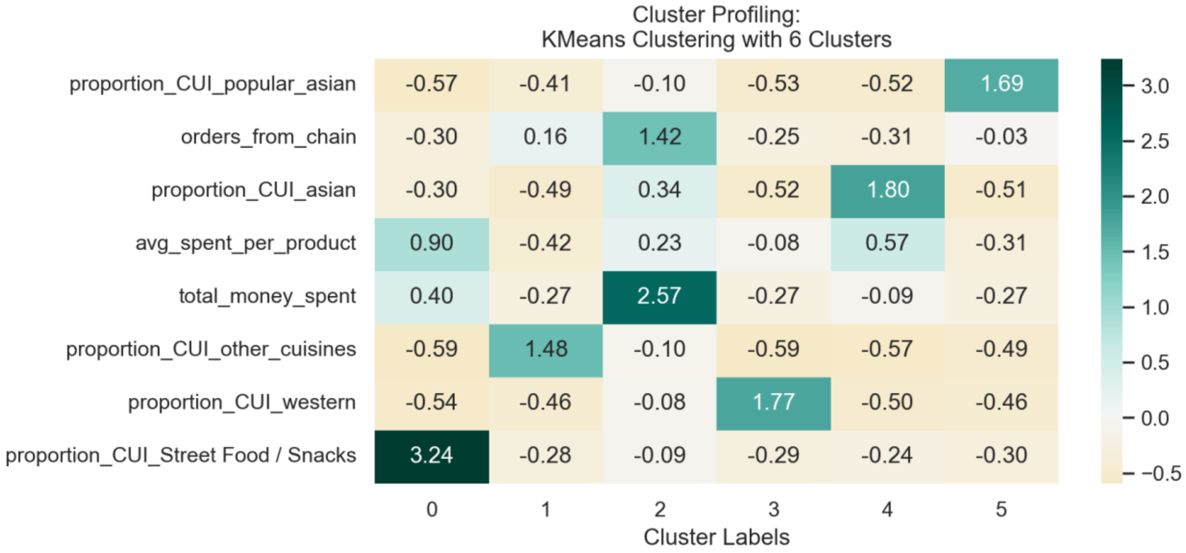
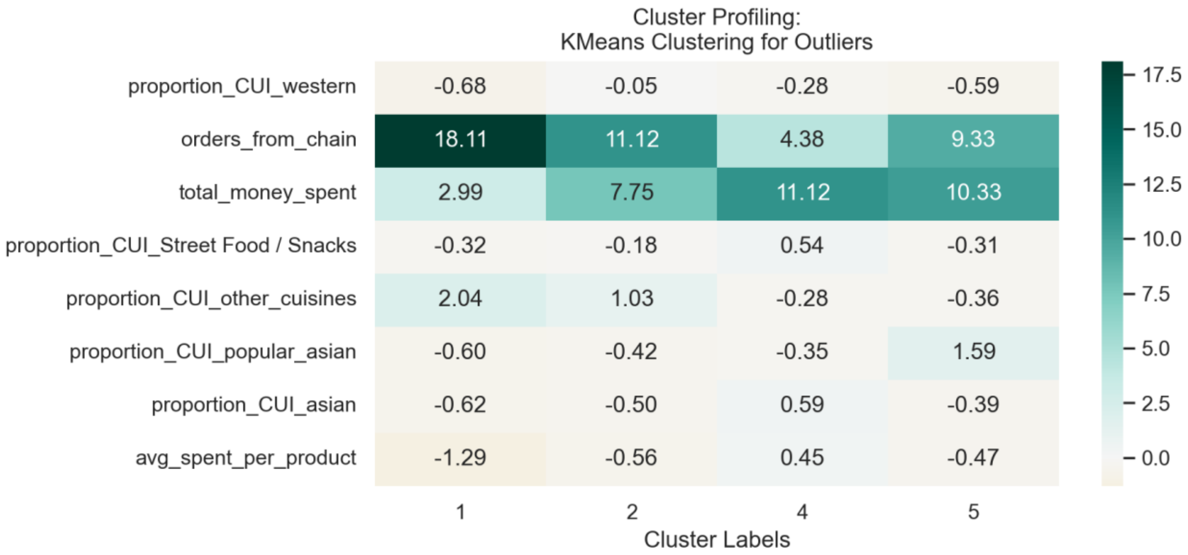


Figure B19: Heatmap of our final solution for the profiling



| | 1 | 2 | 4 | 5 |
|---|---|---|---|---|
| proportion_CUI_western | -0.68 | -0.05 | -0.28 | -0.59 |
| orders_from_chain | 18.11 | 11.12 | 4.38 | 9.33 |
| total_money_spent | 2.99 | 7.75 | 11.12 | 10.33 |
| proportion_CUI_Street Food / Snacks | -0.32 | -0.18 | 0.54 | -0.31 |
| proportion_CUI_other_cuisines | 2.04 | 1.03 | -0.28 | -0.36 |
| proportion_CUI_popular_asian | -0.60 | -0.42 | -0.35 | 1.59 |
| proportion_CUI_asian | -0.62 | -0.50 | 0.59 | -0.39 |
| avg_spent_per_product | -1.29 | -0.56 | 0.45 | -0.47 |

Figure B20: Heatmap to profile the Outliers

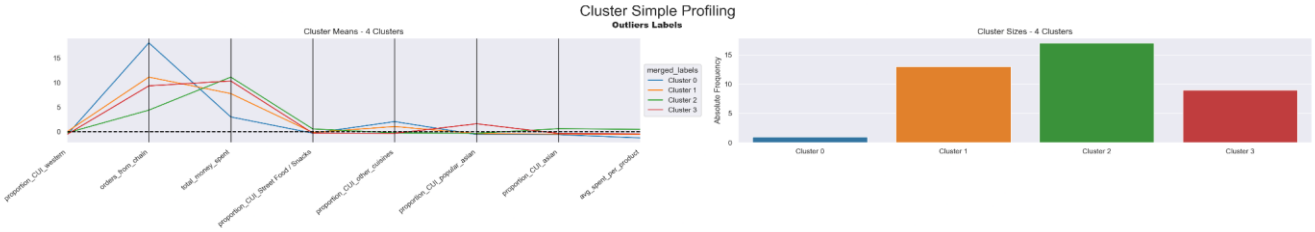


Figure B21: Parallel coordinate plots and bar graph to profile our Outliers

# Annex

1. EDA - In-Depth:
   DM2425_Part1_04.pdf

2. GitHub Repository: https://github.com/luispsDev/Data_Mining_Project_Group_04

3. AI Usage:
   ChatGPT was used in the following cases:
   a. To generate the code that allowed us to make Figures B15 and B16.
   b. To ensure well-written text and improve textual organization.
   c. To make the second threshold on the dendogram on Fiigure B10
   d. In section 4.1 of the notebook, to undo the StandardScaler of the unused metric-features and put them on their original distribution
   e. In section 3.6 of the notebook to provide the correlated pairs of features from the correlation matrix
   f. To help generate UMAP visualization