DATA MINING PROJECT

Master in Data Science and Advanced Analytics

ABCDEats Inc.

Group 16

Ana Margarida Valente, 20240936

Catarina Carneiro, 20240690

Rui Reis, 20240854

Mara Mesquita, 20241039

Fall/Spring Semester 2024

TABLE OF CONTENTS

TABLE OF APPENDIX

1. INTRODUCTION

In today's competitive market, consumers are increasingly selective about the businesses they choose to support and in order to improve, ABCDEats Inc. is determined to understand their customers, providing personalized services and effective marketing strategies. Customer segmentation allows the company to group customers based on shared characteristics or behaviours, enabling personalized approaches that address specific needs and preferences.

This project analyses three months of customer data from three cities (2, 4 and 8) to identify meaningful segments for ABCDEats company. These segments are based on factors such as a customer's financial contribution to the company, their purchasing habits and preferences, and their demographic characteristics. The goal is to uncover actionable insights that will help ABCDEats Inc. improve customer satisfaction, retention, and overall, the total income. This project focuses on preparing the data, implementing clustering methods, defining customer profiles, and developing strategies that provide better business results.

In the first delivery of our work, we carried out the initial exploration of the data and some minor data preparation of the variables (grouping the variables hours and age into groups, and changing the type of the variables customer_age, first_order, last_order and HR_0 to integer type). We also did the description of the variables, the covariance check, analysed the numerical features taking into account boxplots and histograms, and the categorical features, observing count plots. After data preparation, we continued with feature engineering, where we created new variables, and feature relationships, where we were able to compare the relationship between certain variables.

In this report, we describe the second part of our project, focusing on data treatment, feature selection, clustering, and marketing strategy development for the ABCDEats Inc. company. We began by addressing data issues, such as handling missing values, strange values and outliers, to ensure the quality of our analysis. Following this, we selected relevant features by observing the correlation matrix and using Self-Organizing Maps (SOMs). The clustering process was then handled from two perspectives: value-based and behavioural-based. These perspectives were then merged to form the final clusters. Lastly, we profiled the clusters and proposed targeted marketing strategies adapted to each cluster's characteristics.

2. DATA PREPROCESSING

We first started by renaming the columns: DOW_0 to Sunday, DOW_1 to Monday, DOW_2 to Tuesday, DOW_3 to Wednesday, DOW_4 to Thrusday, DOW_5 to Friday and DOW_6 to Saturday, for better understanding while analysing it. Previously, we reviewed the data and determined that the columns **HR_0**, **first_order**, and **customer_age** would be more appropriate as integers rather than floats.

Missing Values:

HR_0: To fill missing values in the variable HR_0, we compared the sum of orders of the hours with the sum of orders of the weekdays. If there was a difference, that value was assigned to HR_0, if not, HR_0 should be 0.
first_order: When analysing the missing values in the first_order variable, it was observed that in these cases, last_order was also equal to 0. We investigated further to see if, under this condition, only one order was placed. However, there were two exceptions where both orders occurred on the same day (Saturday). This likely happened because that corresponded to the starting day of the dataset. Based on this observation, we decided to replace the missing values in first_order with 0 to ensure consistency, making sure that both first_order and last_order refer to orders placed on the same day (the starting day of the dataset). This approach was also led by the logic that first_order cannot occur after last_order.
customer_age: For the customer_age variable, we observed the histogram, which was left-skewed. To better represent the central tendency, missing values were replaced with the median value of 26.

Strange Values:
customer_region: The value "-" represented 1.4% of the dataset and was replaced with the mode (region 8670) of the variable to ensure consistency.
last_promo: For the last_promo variable, the value "-" was replaced with "No Promo" to clarify the information.

Binarization:
Is_chain: Was converted into a binary format to align with the metadata. Any value greater than 0 was recoded as 1, ensuring the variable represented a clear "yes" or "no" outcome.

3. OUTLIERS

For the outlier's treatment we started by performing an automated outlier detection, using the Interquartile Range (IQR) method, this method also detected "global outliers", meaning observations that are outliers in all features. However, we observed that relying solely on this method would have led to the elimination of the entire dataset. Therefore, we concluded with a manual outlier removal process, removing the extreme outliers using null-eye analysis of the boxplots to define the intervals from which the outliers should be excluded, resulting in 99.288% of the data being retained.

In addition to the previous methods, we used the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm to detect and handle the remaining outliers. We began by determining the optimal epsilon (eps) for the algorithm by plotting a k-distance graph, which helped us identify that it should be set to 1. DBSCAN labelled 677 observations (2.14%) as outliers, marked with

-1. These outliers were temporarily removed from the dataset and stored in a separate dataset to prevent them from influencing the clustering process. Afterward, they were reintroduced and relabelled with the appropriate final merged labels. Ultimately, the model estimated that, on average, 86.75% of customers were correctly predicted. This approach allowed us to effectively detect and isolate outliers that deviated from the general patterns in the dataset, improving the quality of our analysis and following clustering steps.

4. FEATURE ENGINEERING

Time-Periods - To simplify the analysis and reduce dimensionality, the 24-hourly activity columns (HR_0 to HR_23) were aggregated into four time periods: early_morning(0h-5h), morning(6h-11h), afternoon(12h-17h) and night(18h-23h). These new features provide an orders to time period ratio activity patterns.

Age_group - Facilitating the analysis of customer trends by age, we decided that customer_age column should be turned into a categorical variable representing age groups. We defined the aged groups by 'Teenagers' (15-19), ’Young Adults’ (20-29), ‘Adults’ (30-49), ’Middle Aged’ (50-64) and 'Seniors' (65+), providing a better understanding costumer behaviours by age.

Sum_of_Orders - was created by aggregating the number of orders placed across the days of the week (Monday to Sunday). This feature captures the total weekly activity for each customer and simplifies the analysis by reducing dimensionality while retaining key information.

To help us make a customer segmentation we created these **RFM** features:

Recency(R): Was created to measure how recently each customer placed their last order. The recency feature enables a clear understanding of customer engagement levels, where lower values indicate more recent activity.

Active_period and Frequency(F): The active_period measures the total number of days between a customer's first and last orders, providing insight into their activity duration. The frequency feature calculates the average number of orders per day during this period, indicating the frequency of customer engagement. These features help differentiate long-term, high-frequency customers, from short-term, or low-frequency users.

total_spend(M): Captures the overall spending behaviour of customers.

By combining these metrics, as it is done in the RFM method, we can classify customers into groups such as loyal, at risk, and new customers, supporting personalized strategies and improved retention.

Cuisine_diversity: We decided to analyse customer preferences for food variety, which led us to a create a new feature, cuisine_diversity. This feature provides insight into customer preferences, distinguishing those with diverse tastes from those who prefer a limited range of cuisines.

Customer_city: Was based on the customer_region first digit (2,4 or 8) to group regions and to simplify and standardize customer location information.

Weekdays and Weekends: In order to interpret customer ordering patterns more effectively, two new features, Weekdays and Weekends, were created. These features aggregate orders to weekdays across the weekdays (Monday to Friday) and weekends (Saturday and Sunday). This is done making a ratio of orders to weekdays or weekends.

CUI Types (Types of Cuisine): Were grouped into 5 categories such as 'Main Courses,' 'Snacks and Street Food', 'Desserts and Beverages', 'Health and Special Diets' and 'Other'. This step simplifies variables and captures overall spending patterns across different cuisine types.

- Main Courses: This group captures the primary meal offerings, which typically include more large dishes and are often associated with specific cuisines. ('CUI_American', 'CUI_Chicken Dishes', 'CUI_Indian', 'CUI_Italian', 'CUI_Thai', 'CUI_Chinese', 'CUI_Asian', 'CUI_Japanese').
- Snacks and Street Food: This group includes lighter, quick-to-eat items, often found in street food settings or cafes. ('CUI_Street Food / Snacks', 'CUI_Noodle Dishes', 'CUI_Cafe').
- Desserts and Beverages: This group appropriately captures sweet items and drinks, which are often consumed after the main meal or as standalone treats. ('CUI_Desserts', 'CUI_Beverages').
- Healthy and Special Diets: This group isolates options focused on health-conscious or special dietary needs. ('CUI_Healthy').
- Other: This is a general category for any items that don't clearly fall into the other categories. ('CUI_OTHER')

5. ENCODING

We decided to use One Hot Encoding approach for the following nominal categorical features:

- 'payment_method'
- 'is_chain'
- 'customer_city'

Regarding ordinal categorical variables, for **age_group**, we used an Ordinal Encoding approach. For the variable **last_promo**, we applied a mapping technique, where ***'NO PROMO'*** is mapped to 0, and all other values are mapped to 1. This approach was chosen because we are primarily interested in understanding whether a promotion was used or not, irrespective of the type of promotion.

6. SCALING

Regarding scaling, two approaches were used as a test to see which was best suited to our project. We began by defining which numerical variables we were going to use, and then proceeded to try the MinMax Scaling. On the other hand, in Standard Scaling since the extreme outliers were already removed it proved to be the best one for KMeans.
We therefore opted for Standard Scaling because it centralises the data more and is more robust to outliers.

7. FEATURE SELECTION

In the feature selection process, we first organized the features into three categories: metric features, non-metric features (encoded), and unused features (original categorical features). We then removed features that were explicitly represented by others through grouping (HR, DOW, and CUI). To better understand the relationships between variables, we visualized the data using correlation analysis and Self-Organizing Maps (SOMs).

Next, we identified and eliminated redundant features with high correlation (*see Appendix A – figure 1*). For instance, **product_count** had a correlation of 0.95 with sum_of_orders, **vendor_count** showed a 0.9 correlation with cuisine_diversity, and **last_order** had a -1 correlation with recency, making them redundant for clustering. We also excluded features unsuitable for clustering, such as time periods, which lacked clear differentiation in the final clusters. The SOMs (*see Appendix A – figure 2*) helped reveal features that did not exhibit meaningful variation or distinct patterns across the data, making them less useful for the clustering analysis, based on these insights, types of meal were removed. Additionally, customer_age, which had no significant correlation with other features, was also excluded (*see Appendix A – figure 3*).

The **final set of features** selected for the cluster analysis included: 'total_spend', 'frequency', 'recency', 'active_period', 'first_order', 'cuisine_diversity', 'Weekdays', 'Weekends' and 'Sum_of_Orders'.

8. CLUSTERING

Starting of the clustering process, the features were divided into two distinct perspectives to better capture the differences in the customer's behaviours.

A **Value-Based Segmentation**, that focuses on features that reflect the value a customer brings to the business. The selected features: 'total_spend', 'frequency', 'recency', and 'active_period' provide insights into how much customers spend, how often they engage, how recently they interacted with the service, and the duration of their active engagement.

We also did a **Behaviour-Based Segmentation**, this perspective aims to understand customer behaviour and preferences. The chosen features: 'first_order', 'cuisine_diversity', 'Weekdays', 'Weekends', and 'Sum_of_Orders' capture patterns such as when customers placed their first order, the variety of cuisines they prefer, and their ordering behaviour across weekdays and weekends.

By creating separate datasets for these perspectives, the analysis can be fitted to explore and cluster customers based on either their value contribution or their behaviour.

The next step was deciding the optimal number of clusters using R^2 plots generated for both perspectives (*see Appendix A – figure 4 and figure 5*). After evaluating multiple clustering methods, including K-Means and Hierarchical Clustering with different linkage criteria (complete, average, single, and ward), for Value-Based Segmentation, the R^2 plot revealed that **two** clusters effectively captured the variance in the value-based features. For the Behaviour-Based Segmentation, the analysis showed that **three** clusters worked best, capturing the key differences in customer behaviour. In both perspectives, K-Means showed to be the better clustering method. For the Value-Based perspective we got an R^2 of 0.47, as far as it goes for the Behaviour-Based Perspective, the R^2 was 0.54.

8.1. MERGING PERSPECTIVES

To draw interesting conclusions, these two perspectives must be merged and we tried two strategies, **Hierarchical Clustering Merging** and **Manual Merging**.

8.1.1. Hierarchical Clustering Merging

The goal is to perform hierarchical clustering on a dataset and create a dendrogram, which shows how data points (or clusters) are progressively merged into larger clusters and decide which how much Euclidian distance we are willing to accept between each smaller cluster and one that aggregates that with another.

Using this approach, we got three merged clusters that performed poorly on our clustering merging, generating uneven cluster distributions and huge disparity of highly frequent clusters from others.

8.1.2 MANUAL MERGING

The goal is to perform a merging that where is clusters where customers belong to two clusters perspectives and the intersection being of low frequency (selected manually with a threshold of less than 5000 occurrences) merged into the nearest cluster.

This is done by finding the centroids, getting all the distances between centroids and then merged these intersections of clusters into the closest pair of clusters centroids. Then finally assign new cluster label to the pair of two clusters to make sure the merging is from two distinct clusters to one.

Using this approach, with four clusters we got a far even distribution of merged clusters that was our main goal (*see Appendix B – table 1*).

8.2 CUSTOMER SEGMENTATION AND CLUSTER PROFILLING ANALYSIS

The clustering process in this project identified four distinct customer segments, each with unique behaviour, spending habits, and preferences. This analysis provides insights into how customers interact with the platform, which can be used to improve customer satisfaction and improve business growth (*see Appendix A – figure 6 and Appendix B – table 2*).

Cluster 0: Weekend-Focused Occasional Customers

Cluster 0, with 8204 customers, includes customers who engage with the platform only occasionally, mostly on weekends. These customers have a low order volume, and their total spending is also below average. Their behaviour shows that they have low engagement, as they rarely place orders and have not made recent purchases. These customers tend to stick with meals that they are most used to, showing a limited diversity in their cuisine choices. Activity during weekdays is very low, with a noticeable spike on weekends.

In summary, Cluster 0 consists of customers who use the platform for convenience on weekends, are more price-sensitive, and are less open to trying new cuisines.

Cluster 1: Regular Weekday-Focused Users

Cluster 1 contains 7509 customers who consistently place orders during weekdays, but their spending remains low and the order volume is slightly below average. While these customers have recently engaged with the platform, their overall frequency of orders is still low. They have a longer engagement period, slightly above average, meaning they have been using the platform for some time but with a low intensity of usage. Their cuisine preferences are somewhat diverse, as they are open to trying different meals, but they still show a preference for options that they are used to. Most of their activity occurs during weekdays, maybe for work related meals, while their weekend usage is limited.

In conclusion, Cluster 1 represents customers who use the platform reliably for weekday meals but do not spend much or engage frequently.

Cluster 2: High-Frequency, Low-Spending Customers, Weekday-Focused

Cluster 2 consists of a total of 6934 customers, who place orders often but spend little. These customers have not ordered recently, which suggests a potential drop in engagement. Their active period is also short, meaning they may not have been using the service for a long time. However, these customers used to order frequently, meaning that when they do engage, they place orders more often within a shorter time span. They stick to known meals, as shown by their low cuisine diversity. Their spending is low, which suggests that despite frequent orders, they tend to choose cheaper or familiar meal options, and they show a clear preference for weekday use.

In summary, Cluster 2 represents customers who engage often but spend little. Their usage is mostly on weekdays, and they prefer sticking to known meals rather than exploring new options.

Cluster 3: Adventurous, High-Spending Customers

Cluster 3, with 8324 customers, includes highly engaged customers who are adventurous and spend more, making them one of the most valuable segments. These customers place a high number of orders, with their sum of orders being significantly above average. Their recency score shows that they have been the most active recently, and their active period shows their consistency on the platform, suggesting strong loyalty. Their frequency is relatively low, meaning they do not place orders as often, but when they do, they tend to spend significantly more. These customers are also highly diverse in their cuisine choices, showing a strong willingness to try different meal options. Their activity on weekdays and weekends is quite balanced, meaning they engage with the platform throughout the week.

In conclusion, Cluster 3 represents loyal, high spending customers who are not only order more frequently but also enjoy exploring a variety of cuisines, contributing greatly to the company's revenue.

8.3 CLUSTER PROFILING ANALYSIS

For the profiling of the clusters, we used the categorical features as well as some numerical features such as time periods and type of meals. However, after the analysis, the most relevant insights came from only two categorical variables, the customer city and the last promotion used.

The **city** (*see Appendix A – figure 8*) showed to be an important factor in distinguishing the behaviour patterns, from the analysis, it was observed that Cluster 3, which had the highest total spending, was predominantly made up of customers from City 2. In contrast, Clusters 0 and 2, which had the lowest total spending, were primarily composed of customers from City 8. Meanwhile, Cluster 1 showed a more balanced distribution, with customers almost equally split between City 4 and City 8. In terms of total spending, Cluster 1 fell in the middle range, neither as high as Cluster 3 nor as low as Clusters 0 and 2. This geographical distribution highlights the varying spending patterns across cities and clusters, offering valuable insights for location-based marketing strategies.

When analysing the last **promotion** used (*see Appendix A – figure 9*), it was noted that Cluster 2, which has the lowest total spending, showed almost equal frequencies of applying a promo code and not using one on their last order. In contrast, the other clusters had significantly higher frequencies of orders placed without using a promo code. This distinction suggests that while promo codes play a role in Cluster 2's engagement, the other clusters rely less on such promotions, reflecting different behavioural patterns in their spending habits.

We can see that on Cluster 3 there is a lower usage of promo codes suggesting that is a profitable group of people.

9. MARKETING STATREGIES ADAPTED TO CUSTOMER SEGMENTS FROM CLUSTER ANALYSIS AND PROFILING

Cluster 0: Weekend-Focused Occasional Customers (Prominent in City 8)

For Cluster 0, the focus should be on encouraging more frequent weekend orders.

Limited weekend promotions or discounts can be offered to attract more customers.

Since Cluster 0 is mostly from City 8, we can run geo-targeted ads highlighting exclusive weekend deals in City 8 to make the customer feel that's something "right around the corner".

Also, since this cluster, mostly, did not use a promo code in their last order, we can introduce them with a friend's referral promo codes promotion.

Cluster 1: Regular Weekday-Focused Users (Equally from City 8 and City 4)

To target Cluster 1, weekday-specific campaigns should be used. Offers like "Lunch Break Deals" or "Workday Specials" could motivate them to use the platform more for work-related meals.

As this cluster is spread across both City 8 and City 4, localized ads can be used.

A loyalty program offering points for every order could help reward their long-term engagement.

Knowing their meals are work related targeting them with restaurants near their workplace would be more effective.

Since Cluster 1 did not use promo codes in their last orders, we can introduce them, as well to a co-worker's menu packs discounts.

To further engage them, offering discounts or free trials on new cuisines would motivate them to explore more variety in their meal choices.

Cluster 2: High-Frequency, Low-Spending, Weekday-Focused Customers (Prominent in City 8)

For Cluster 2, these are people that are low investing in the platform that prefer their usual meals that could depend on their wage.

The goal of the strategy for this group is to make them more invested in the app regardless of their wage.

As they use promo codes frequently, offering them exclusive, targeted promo codes like a "Frequent User Discount" can encourage more orders and higher-value transactions, although the latter may not work on lower wage people.

Since Cluster 2 is prominent in City 8, weekday-specific offers, such as free delivery targeted at City 8, could help keep their loyalty.

Encouraging them to try different cuisines can be done by offering points when they buy something so they can later spend them on new dishes that are like their usual preferences, helping them explore new options gradually.

Cluster 3: Adventurous, High-Spending Customers (Prominent in City 2)

For Cluster 3, marketing strategies should focus on their high spending habits and adventurous preferences.

One approach could be offering premium dining experiences, like gourmet meals and at same time appeal to customers seeking a sense of elevated status.

Since this cluster is most active in City 2, campaigns with premium options available in their city can help make the offers feel more exclusive.

Lastly, even though they didn't use promo codes in their last orders, offering a promotion for high-value orders or special promotions might convince them to be bolder on their new experiences.

10. CONCLUSION

In this report, we describe the second part of our project, focusing on data treatment, feature selection, clustering, and marketing strategy development for ABCDEats Inc. We began by addressing data issues, such as handling missing values, strange entries, and outliers, to ensure the quality and reliability of our analysis. We also created new features, such as time periods, age groups, types of meal and RFM metrics (Recency, Frequency, and Monetary value), to better capture customer behaviour and value. Encoding techniques were used for categorical variables, and Standard Scaling was applied to ensure uniformity across numerical data. After this, we selected the relevant features by analysing the correlation matrix and applying Self-Organizing Maps (SOMs) to identify patterns.

The clustering process was approached from two perspectives: **value-based** and **behavioural-based**. These perspectives were later merged to form the 4 final clusters. Finally, we profiled the clusters and proposed targeted marketing strategies adapted to each cluster's characteristics.

Our analysis identified four distinct customer segments, each with unique behaviours and preferences:

Cluster 0 includes occasional weekend-focused customers, mostly from City 8, who are less engaged and prefer already meals that are most used to. Targeting this group with those meals could encourage more activity from this group.

Cluster 1 consists of regular weekday users, equally distributed between City 4 and City 8, who prefer workday meals. Loyalty programs and workplace-focused deals definitely will help strengthen their engagement.

Cluster 2 represents high-frequency but low-spending customers, primarily from City 8. Their high usage of promo codes creates opportunities for tailored promotions to increase spending and encourage them to explore new cuisines.

Finally, **Cluster 3** stands out as adventurous, high-spending customers, predominantly from City 2. Providing premium dining experiences and exclusive campaigns tailored for these high-income individuals will further enhance their loyalty, value and at the same time project higher status (something these types of people always try to be perceived).

By addressing the unique needs and behaviours of each segment, ABCDEats Inc. can implement more targeted marketing strategies, improve customer satisfaction, revenue and consolidate their position in the market.

A significant challenge during the project was the dataset's high number of outliers, which required careful handling to maintain data integrity without losing too much information. Additionally, some features, such as customer age and types of cuisine, did not provide meaningful insights during the analysis. While these variables could have offered valuable information, their lack of differentiation or strong patterns made them less useful for clustering, highlighting the limitations of the dataset.

APPENDIX A - FIGURES

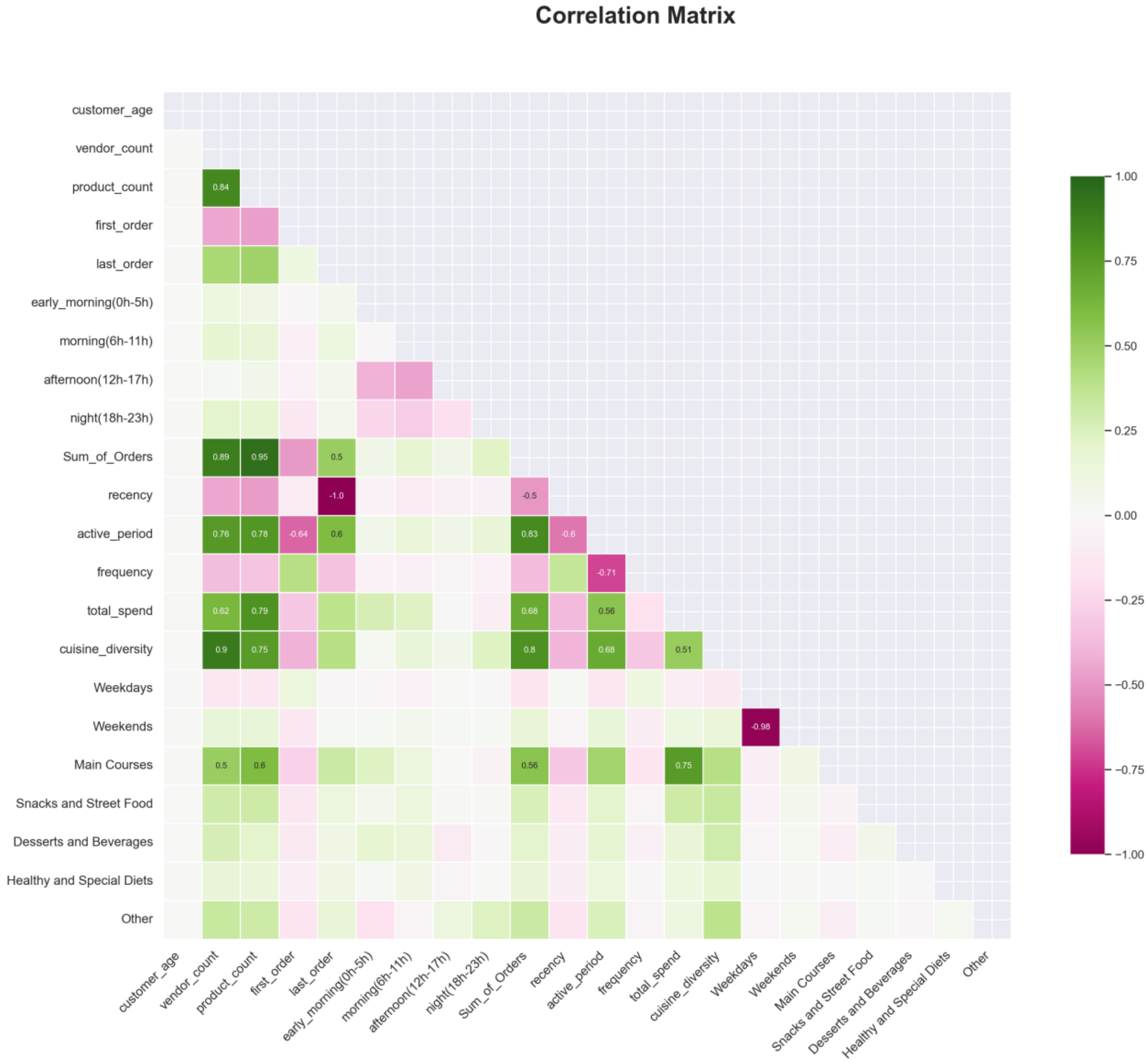

Figure 1: Correlation Matrix Higher 0.5

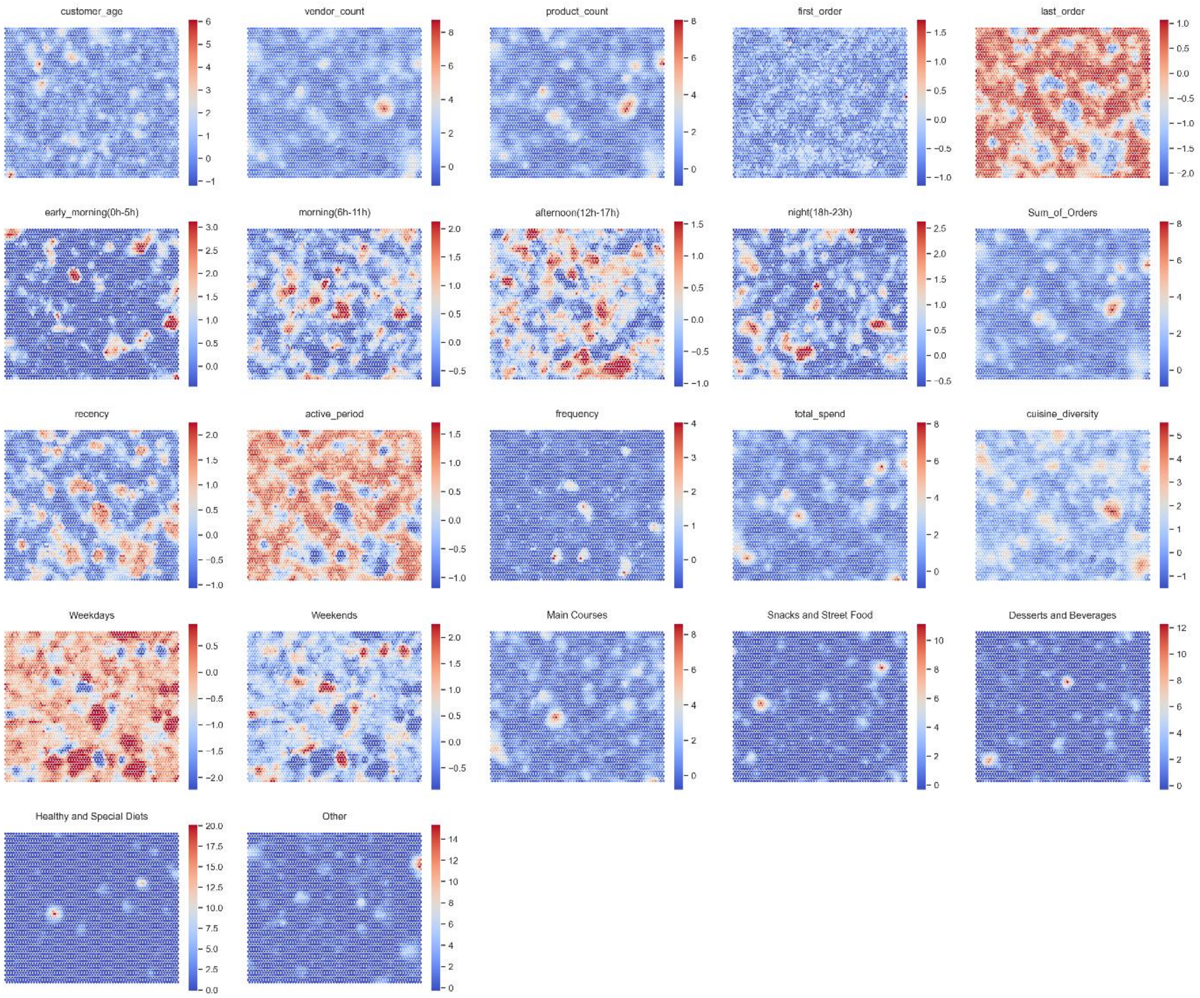

Figure 2: SOMS – Data Visualization

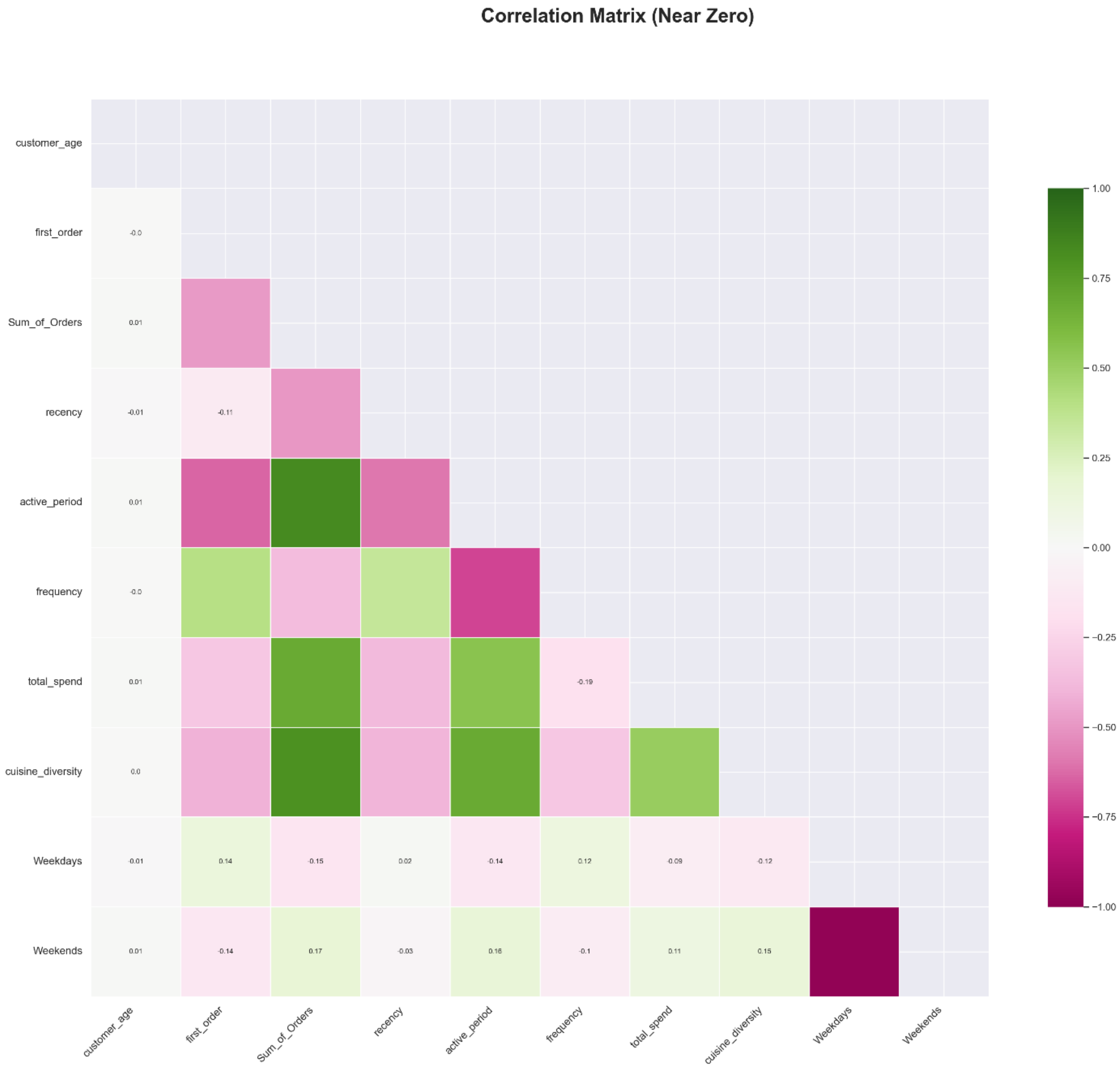

Figure 3: Correlation Matrix Higher 0.2

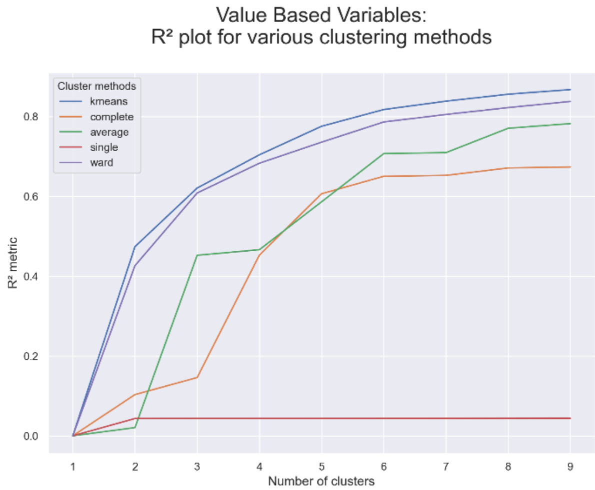

Figure 4: Value Based Variables

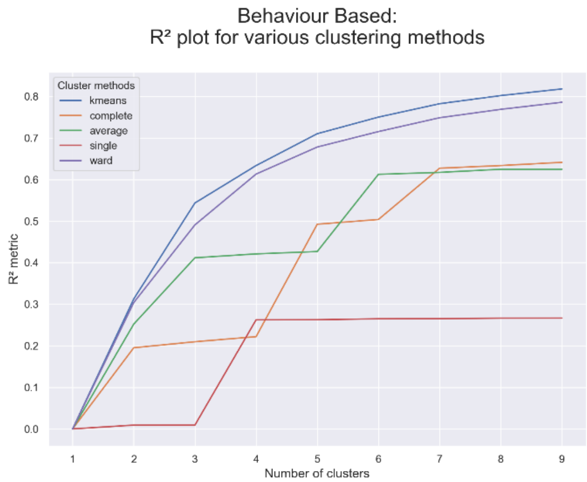

Figure 5: Behaviour Based

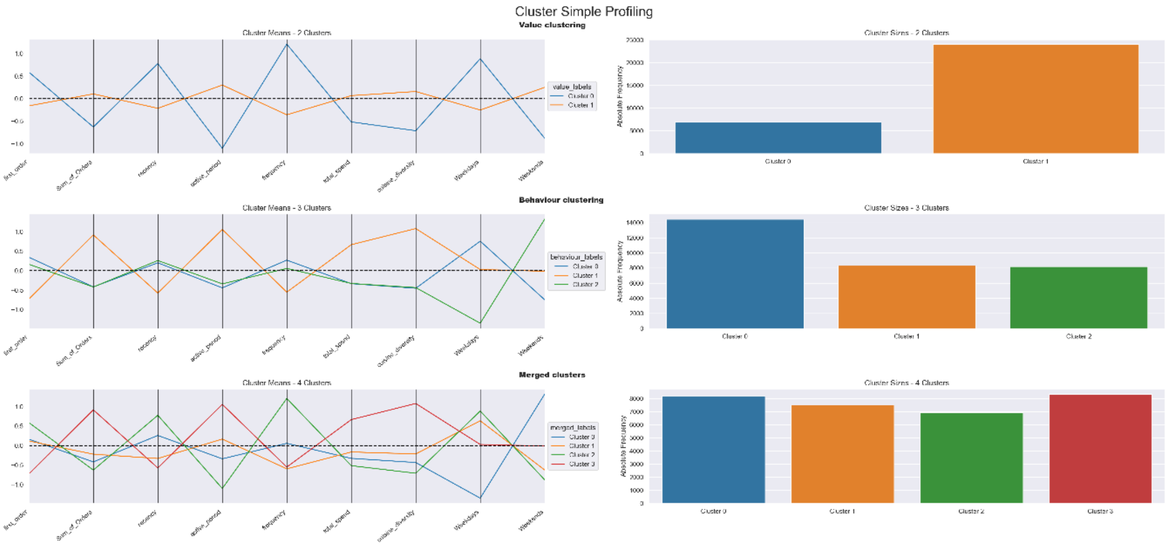

Figure 6: Cluster Profiles and their Frequencies

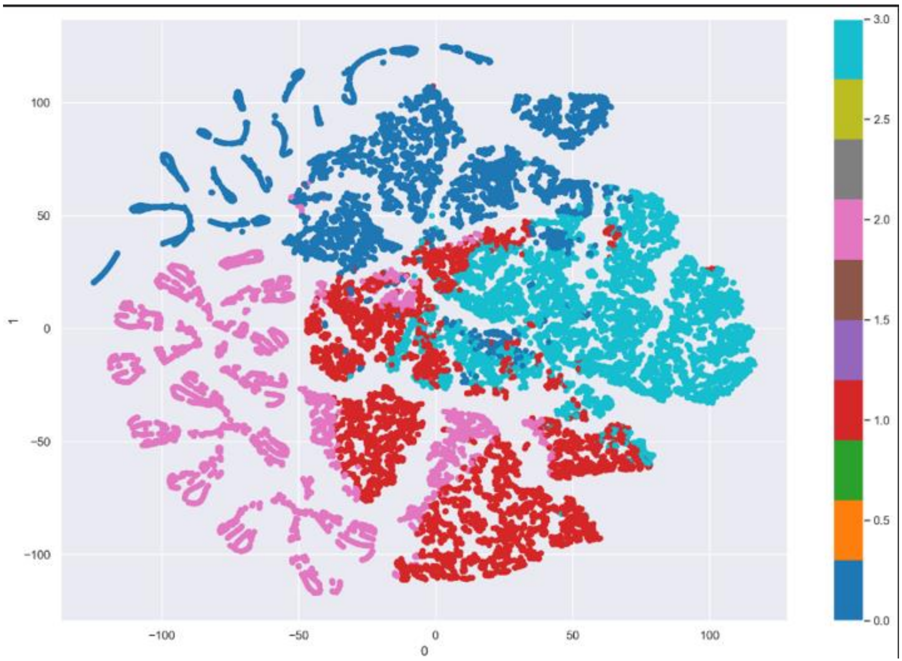

Figure 7: Cluster Visualization using t-SNE

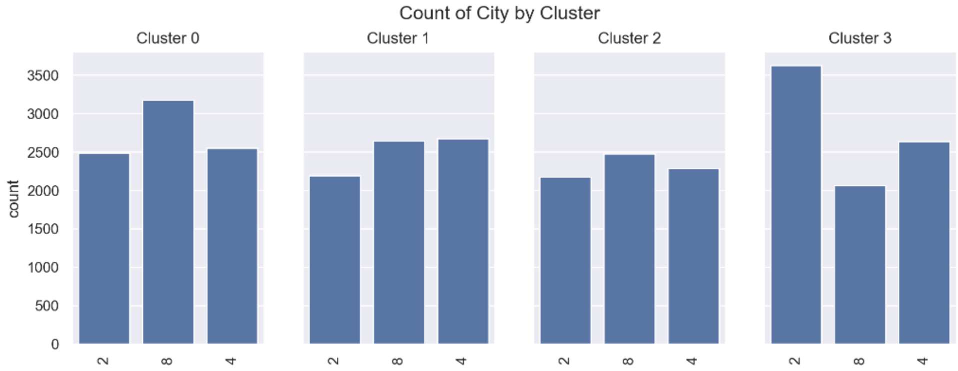

Figure 8: Count of City Cluster

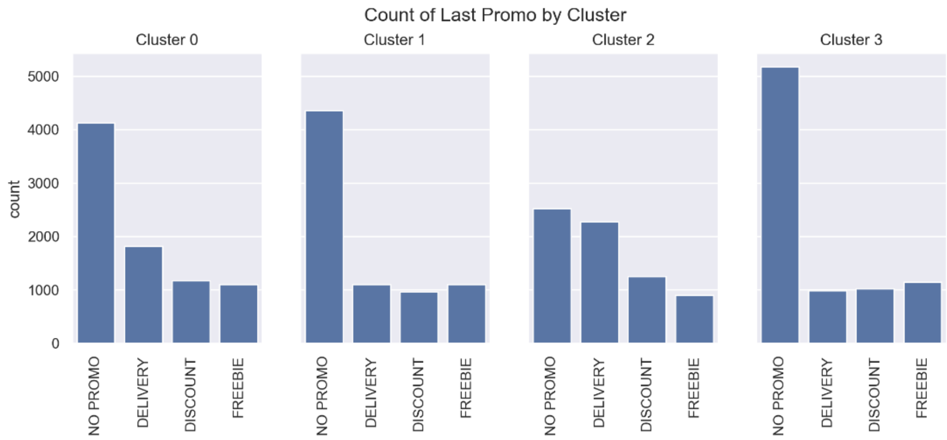

Figure 9: Count of Last Promo by Cluster

APPENDIX B – TABLES

Table 1: Manual merging - frequencies of merged labels

Merged Label	Frequency
0	8204
1	7509
2	6934
3	8324

Table 2: Customer Segmentation

Merged Labels	first_order	Sum_of_Orders	recency	active_period	frequency	total_spend	cuisine_diversity	Weekdays	Weekends
0	0.163100	-0.419784	0.259247	-0.339894	0.057143	-0.328690	-0.434932	-1.349376	1.322059
1	0.117415	-0.221016	-0.331366	0.166154	-0.597401	-0.165654	-0.215285	0.635462	-0.627413
2	0.588529	-0.628882	0.777212	-1.100957	1.207711	-0.516381	-0.711132	0.885753	-0.880018
3	-0.728923	0.915639	-0.572576	1.055821	-0.551751	0.665428	1.079125	0.027733	-0.014064