DATA MINING PROJECT

Master in Data Science and Advanced Analytics

NOVA Information Management School

Universidade Nova de Lisboa

Final Report:

ABCDEats Inc. Costumer Base Segmentation

Group 36

Eduardo Mendes, 20240850

Joana Esteves, 20240746

João Afonso Freire, 20240528

Tomás Figueiredo, 20240941

Fall/Spring Semester 2024-2025

TABLE OF CONTENTS

Index of Figures

Index of Tables

INTRODUCTION

This project focuses on the segmentation of ABCDEats Inc.'s customer base, and makes use of a dataset collected over a three-month period. Our primary objective was to ensure data quality and enhance the available information to improve the cluster results. We aimed to experiment with various clustering algorithms to determine which approach best suited our data and explore potential benefits of combining algorithms to optimize results. Following the clustering process, we intended to properly assess cluster quality in order to identify any limitations in our analysis, and evaluate the relevance of segmentation features in the clusters' formation. Ultimately, our goal was to effectively profile the identified customer segments from multiple perspectives, and develop targeted marketing strategies tailored to each distinct group. This report outlines the clustering process undertaken in this project, explaining key decisions and providing the results of our approaches.

This document begins by explaining the preprocessing steps taken to enhance information for the clustering process. We started by addressing missing values through methods such as K-nearest neighbors. Additionally, we explained our rationales for the approaches taken to corrected data inconsistencies identified during the EDA (exploratory data analysis). Since the feature engineering results differ greatly from the EDA phase, we introduce the new features created. Following this, we outline how we conducted the outlier analysis, identifying both global and univariate outliers using methods such as the interquartile range and z-score. Additionally, we explain how we managed features with different units to ensure they equally influence the clustering results. Before proceeding with the models, we analyzed the available features to determine which numeric variables would contribute most effectively to segmentation while reducing redundancies.

The report then moves into the clustering process, starting with the experimentation phase, where several algorithms were explored. Following this, we present the final clustering solution, which incorporates a hybrid approach combining hierarchical clustering and K-means. For the first stage we explain how we implemented hierarchical clustering using an agglomerative algorithm, and determined the best linkage methods and the optimal number of clusters. In this report we also include the intermediate results from this stage to identify critical areas that required improvement. In the next step, we used the K-means algorithm to further optimize cluster formation and present the results of this approach. Furthermore, we assessed the improvements using the silhouette coefficient plot and mean, and present visualizations of the clusters. Lastly, we analyze the number of observations for each cluster to detect possible imbalance.

Following this, we introduce an additional cluster for the extreme values initially removed and explained the rationale behind this decision. In order to understand if the feature selection resulted in relevant features and distinct clusters, we examined the contribution of each feature to cluster formation using various methods, including feature range analysis across centroids, permutation tests, statistical testing, and boxplot analysis.

The last chapter focuses on cluster profiling, exploring economic, demographic, and preference-based perspectives based on available features. Each cluster is differentiated by segmentation features and additional profiling variables, allowing for further characterization. The final clusters are compared to highlight key differences in the final segments of ABCDEats' customer base. A marketing strategy is then developed for each group, tailored to their distinct characteristics.

1. DATA PREPROCESSING

The preprocessing steps outlined in Table 1 summarize the transformations applied following the EDA conclusions. Below, the key decisions are described and justified in more detail.

1.1. Missing data Imputation

The **customer_age** feature had 2.28% of missing data. Given this percentage, we decided to retain it to preserve as much information as possible for our analysis. To fill in missing values, we used K-Nearest Neighbours (KNN) with five neighbours, which preserves local structure and relationships within the dataset. This method is particularly effective for numerical features, and can provide more accurate estimates but it can also introduce some bias into the clustering process. As we could see in Ramu and Shahram' s document.

For the **first_order** feature, as we have already mentioned in our EDA Report, we will replace the missing data (0.33% of the rows) with the value 0, because it happens when variable **last_order** is also 0, indicating that both correspond to the first day of our dataset.

The **HR_0** feature had 3.65% missing data. Since we don't have missing values for the number of orders in the other timelines (HR_1 to HR_23), and the number of orders by days of the week have no missing values either, we can calculate the true values as follows: HR_0 = ∑ (DOW_0 to DOW_6) - ∑ (HR_1 to HR_23).

1.2. Incoherencies in the data

The **customer_region** for 1.39% of the customers was given by an unknown region, labeled as "-". Given that, we replaced the symbol for missing values and once again employed the KNN with five neighbours for imputation.

For **last_promo**, as we have already mentioned in the previous report, we believe that the category "-" represents customer who did not use a promotion code, indicating a common trait among these customers. Since this is a categorical variable, it makes sense to label this category as "No Promo Code";

For the **vendor_count** and **product_count,** in the EDA we noticed high frequencies for zero while all customers have made at least one order. So, in order to remove these inconsistencies, we decided to drop these 138 observations, where both features were 0. For the 18 observations of region "2660" where only product_count was 0, we consider there was a mistake in this region and there needs to be at least one product ordered per vendor, so we replaced the zeros by the vendor_count value.

1.3. Feature Engineering

In the feature engineering section, we created eighteen new features, in addition to the ones created in the EDA phase, present in Table 2 which include transformations like ratios, proportions, averages, and category preferences, aimed at capturing relevant patterns.

1.4. Outliers' Analysis

In order to identify global outliers, we devised a function called obvious_outliers() to detect outliers across all features simultaneously. However, there was no observation in the dataset within these circumstances. Additionally, we implemented the Interquartile Range (IQR) method to detect univariate outliers for each feature. We noticed that if we would only rely on this method to remove observations, we would have to delete a large proportion of our dataset (96.17% of the observation) so we did not use it.

To conclude, we made a Z-score study to identify extreme values in our metric features. This approach calculates the number of standard deviations is each observation is from the mean, allowing for a standardized comparison across different features. We set a conservative threshold of 7 to determine observations with extreme values for at least one feature. The percentage of rows removed using this method was 4.81%.

1.5. Data Scaling

The algorithms we intended to use for clustering rely on distance metrics and our dataset contains features with different units, the magnitude of these units proportionally impacts the distance values. To deal with this problem we applied standardization with standard scaler to ensure that no single feature disproportionately influences the clustering results. One disadvantage of this method is that features may scale differently in the presence of outliers, as we only removed observations with very extreme values, leaving some outliers identified by the interquartile range method. However, when analyzing the maximum and minimum values for each feature after standardization we concluded there were no major differences. Therefore, the impact of outliers on scaling may not be as critical and the standardization seems to have been effective in ensuring that no single feature dominates due to scale differences.

During the experimentation process we also tested other scaling options which included Min-Max Scaling to rescale features to a fixed range of [0, 1] and Robust Scaling which is a method robust to outliers. However, the clustering algorithms performed better with the Standard scaler implementation.

1.6. Features' analysis

Regarding this section, we checked the Spearman's Correlations between metric variables and identified redundant features. We avoided using them simultaneously when testing sets of features for clustering.

Additionally, we computed the mean absolute difference for each numeric variable, a measure of spread that can be an indicative of the discriminative power of the features, as higher values originate more pronounced differences between groups. In general, the final selected features had high values for this method: **vendor_loyalty:** 0.85; **order_frequency:** 0.84; **avg_spent_per_order:** 0.72 and **total_orders:** 0.69.

2. CLUSTERING APPROACH

3.1 Experimentation process

One of our approaches was the SOM (Self-Organizing Maps), in which the training process was made by applying a two-phase strategy in a 50x50 grid: Unfolding Phase: We started with a higher learning rate and sigma (neighborhood radius) to allow the map to adapt to the input data distribution over 150,000 iterations. Gradually, in order to improve precision, the sigma was reduced by 10% after every 20,000 iterations. Then a 2nd phase was followed: Fine-Tuning Phase where was reduced the learning rate and the sigma slightly for 550,000 additional iterations, enabling finer-grained adjustments in order to reduce the quantization error. After Fine Tuning Phase we got a QE=0.0615 and a TE=0.2083. After training the SOM, we applied both K-Means and Hierarchical clustering on top of the resulting SOM units to group similar observations. For 5 clusters, the silhouette scores were 0.33 for K-Means and 0.361 for Hierarchical clustering.

Another algorithm tested was DBSCAN (Density-Based Spatial Clustering of Applications with Noise), that involved tuning key parameters: eps and min_samples. To determine eps, we used a k-distance graph, selecting 0.8 based on the point where the distance increased sharply. Min_samples was set to twice the number of features, following a rule of thumb. DBSCAN identified six clusters, including one cluster for noise. After clustering, we evaluated the clustering quality using the Silhouette Score, which yielded a silhouette coefficient of 0.3576.

Additionally, we applied the GMM (Gaussian Mixture Model) to determine the optimal number of clusters, we tested values from 2 to 8 and calculated the Silhouette Score for each. The best coeffient of 0.4 was achieved with 2 clusters, while 7 clusters yielded a score of 0.25. Despite the higher score with 2 clusters the results did not provide meaningful segmentation.

2.1. Final approach: Hierarchical-K-means

Our final segmentation was generated using a hybrid Hierarchical-K-means clustering approach. On the first stage Hierarchical clustering was used to determine the number of cluster and initial centroid assignments for K-means. The second step involved applying K-means with to further refine the cluster assignments and improve compactness and quality.

Several feature sets were tested according to the results of the feature analysis, leading to the following selection based on the best performance. These features provide valuable insights from multiple perspectives:

- **”total_orders”** and **“order_frequency”** capture the number and frequency of customer orders, reflecting their overall ordering behavior.
- **”avg_spent_per_order”** represents the value-based aspect of customer behavior which can capture spending patterns.
- **”vendor_loyalty”** is associated with the customer preferences, either towards specific vendors or broader selection choices.

We avoided including a large number of features due to the limitations of the Euclidean distance metric in high-dimensional spaces. Additionally, our experimentation process consistently demonstrated that

better results were achieved with a smaller set of features, reinforcing the importance of reducing dimensionality to preserve meaningful distances.

2.1.1. Hierarchical clustering

For the Hierarchical clustering we used an agglomerative algorithm with the distance metric Euclidean, for which we analyzed the silhouette coefficient with different linkage methods and a k range from 2 to 10, which can be seen in Figure 1. The linkage methods "single" and "average" displayed the best silhouette coefficient for k values up to 5, but when further analyzing the number of observations in each cluster, both methods had assigned close to the total number of observations to the same cluster. As such, the linkage method ward was selected, because despite having a lower silhouette score the observations were being assigned to different clusters, which indicated it was best suited for our dataset.

For the ward method the number of clusters that led to the best coefficient was 5 with a silhouette coefficient of 0.394. In order to further analyze the optimal number of clusters we plotted a dendrogram that can be visualized on Figure 2. While the most obvious choice was 3 clusters for a threshold distance of 150, our object was to further segment the customer base. To achieve this, we lowered the threshold distance for 120 which indicated 5 clusters, making our final choice at this mark.

2.1.1.1. Visualization of intermediate clustering results

For the visualization of the results of the hierarchical clustering we made use of a dimensionality reduction technique, PCA (principal components analysis) to reduce the segmentation variables to 2 components, which allowed us to plot a 2D visualization of the clusters. The visualization of Figure 3 showed overlapping points between some clusters with no clear boundaries, which paired with the silhouette coefficient indicated room for improvement. The principal components can be interpreted as follows: PC1 is a more general component, mostly represents "total_orders", "vendor_loyalty" and "order_frequency" and PC2 is mostly related to "avg_spent_per_order".

2.1.2. K-means optimization

We conducted an optimization of the results using the K-means algorithm with the objective of improving cluster results, given the overlapping points and not very district borders. This algorithm was also tested in a separate study for different values of k, using K-means++ method to select the initial centroids, however the results were inferior, so we decided not to choose it.

For the optimization, K-means was computed making use of the hierarchical cluster centroids and defined number of clusters (k=5). The silhouette coefficient indicated improvements, rising to 0.427. To further analyze the results, we checked the number of observations that were assigned to each cluster, which indicated imbalanced groups as can be seen on Figure 4. However, in the context of client segmentation for a delivery service the imbalance might reflect real differences in customer behavior, so we consider a normal event that a big portion of clients has more common habits, while some smaller clusters can represent niche segments with more unique behaviors.

2.1.2.1. Cluster quality assessment

To assess cluster quality, we made use of the silhouette plot of Figure 6 to understand the silhouette coefficients of each point and made the following analysis:

- **Clusters 0 and 4** are the most well-formed cluster, with have a significant majority of points with silhouette scores above 0.5, indicating that most points are well assigned. However, a small portion of points in these clusters exhibit lower scores, suggesting they may not be as strongly cohesive.
- **Clusters 2 and 3** show approximately half of their points with silhouette scores above 0.4, while the remainder fall below the average silhouette score. Some points have values close to zero, indicating they are near the border with neighboring clusters. Additionally, there is a small number of points with negative silhouette scores, particularly in Cluster 2, which exhibits more pronounced negative values. As such, clusters 3 and 2 exhibit signs of potential misassignments, with Cluster 2 being the least cohesive.
- **Cluster 1** has a considerable number of points with scores above the average silhouette of 0.4, but the majority remain below this threshold. A minority of points have scores close to zero, indicating they are near the border between clusters, while a very minor number exhibits low negative values.

The final silhouette coefficient of 0.427 suggests the most points are well assigned to clusters, while there may be some overlapping which lowers the coefficient and can results in not so clear borders, and it's possible that there is a minor number of misassigned points. This indicates room for improvement but given that the majority of points have are correctly assigned the results might still provide useful insights on customer segmentation, with can be enhanced with additional profiling features.

2.1.2.2. Visualization of final clustering results

The visualization of the final clustering also involved using PCA, where we reduced the dimensionality of the segmentation variables to 2 and 3 components, allowing for 2D and 3D visualizations, respectively. Despite the optimization, the overlapping clusters persisted. The 3D visualization facilitated a better understanding of the areas where overlap occurred and helped to more clearly identify boundaries that were not fully distinct. The results of the visualization align with the silhouette plot and mean coefficient analysis, confirming that the clusters demonstrate overall satisfactory cohesion, though some areas of overlap and potential misassignments remain. The results can be seen on Figure 5 and the notebook includes an interactive plot, the principal components breakdown is as follows: PC1 is a more general component, mostly related to “total_orders” and “vendor_loyalty”; PC2 is mostly related to “avg_spent_per_order” and PC3 is mostly related to “order_frequency”.

2.1.3. Outliers

The points with extreme values removed in the preprocessing phase, were later reassigned to an “outlier cluster”. These observations showed a wide range of values across all segmentation features, with more extreme values than any of the observations in the main clusters, meaning that including them in the clustering process could negatively affected cluster cohesion. Given their wide range and abnormal values we concluded that these observations could benefit from a generalized marketing strategy. Additionally, during the experimentation phase, before removing the outliers, we observed that their presence negatively impacted cluster formation, leading to a lower silhouette coefficient.

2.1.4. Contribution of segmentation features for clustering

To understand the contribution of each segmentation feature in the clustering process we started by analyzing each feature range across centroids, the results can be consulted in Table 3. This method provided insights into the extent to which the features differentiate the clusters. “total_orders” exhibited the most significant variation between clusters, which indicates it is the most critical feature in distinguishing them. “avg_spent_per_order” and “order_frequency” range indicated these also play an important role in cluster differentiation. Lastly, “vendor_loyalty” shows the least variation among clusters, indicating it has a smaller impact than the remaining features, however the contribution is still notable.

Feature contribution was further assessed through permutation tests by analyzing the changes in silhouette scores after feature values were permuted across clusters. According to the results that can be consulted in Table 4, “vendor_loyalty” and “order_frequency” had the greatest impact on cluster formation, as their permutation caused the largest reduction in silhouette scores. “total_orders” and “avg_spent_per_order” also contributed significantly, though their impact was slightly less pronounced than the top two features. Overall, every feature demonstrated notable relevance, which indicates a collective importance in shaping the clusters.

Additionally, we implemented a Kruskal-Walli’s test, which is non-parametric and adequate for our dataset because it does not assume normal distributions. It allowed us to compare the medians of segmentation features across clusters to understand if the segmentation features provided significant differences. All p-values were very close to zero, which indicates highly significant differences in medians across clusters for all features. The H-statistic, which can be consulted in Table 5, confirms the extent of these differences, which indicated that greatest differences are for “vendor_loyalty”, followed by “total_orders” and “order_frequency”, with “avg_spent_per_order” as the feature that provides the least differences across clusters.

To further investigate the differences, pairwise comparisons between clusters were conducted using Post Hoc Dunn’s test. For all comparisons the p-values was extremely small, demonstrating that every pair of clusters exhibited significantly different distributions across all tested features.
Finally, to visualize the differences we analyzed the box plots for each segmentation feature across clusters that can be visualized in Figure 7. From our analysis, summarized in the Table 6,all clusters displayed variations in at least one feature, with most clusters differing significantly across multiple features.

3. CLUSTER PROFILING

The cluster profiling was based on the visual analysis of multiple plots made for both numeric and categoric features. For the numeric features, the more relevant visualizations can be seen in Figure 8, which shows the boxplots of each feature across clusters. In relation to the categoric features the main plots were stacked bar charts with relative frequencies, which can be consulted in Figure 9. Additionally, Table 7 conveniently summarizes the key characteristics of each segment.

Segment -1.0: Outlier Customers (4.81%)

This segment has a wide range of customers, with extreme preferences across varied characteristics. However, they stand out for having numerous extreme values in areas like last order, first order, vendor count, average spent per order, average products per order and recency. Despite the presence of these extreme values, in most aspects, the median values do not differ much from the other clusters.

Segment 0 - Vendor loyal and moderate spending customers (38.14%):

Economic Perspective: These customers demonstrate slightly above average spending habits, this is true for both the average spending per and the average price per product when compared to other clusters. They tend to purchase few products per order which keeps their expenditure relatively low.

Demographic Perspective: Customers in this cluster are fairly evenly distributed among regions 8670, 2360, and 4660, with the remaining regions seem to have much lower presentations.

Preference-Based Perspective: These customers exhibit a tendency for an intermediate active period of 41 days and have made an order not very long ago, while during this time it's notable the low order frequency. The individuals have a very high vendor loyalty indicating a strong preference for their chosen set of vendors but an above average cuisine diversity, reflecting some willingness to explore food options. Most of the customers order more on weekdays and show an inclination for lunch and breakfast time orders. The majority does not use promotion codes and makes most payments by card. The preferred cuisine type is Asian, American and Italian.

Segment 1 - Moderate customers with varied preferences (22.73%):

Economic Perspective: This cluster is characterized by moderate spending habits indicated by the intermediate average price per product and spent per order. The average number of products per order is low, which is common across clusters.

Demographic Perspective: Nearly half of the customers in this cluster belong to region 4660, with a considerate percentage spread across regions 8670, 2360. The remaining regions are much less popular.

Preference-Based Perspective: Most of the customers in this cluster have been on the platform for a considerable time, with a median of 47 days, and have ordered relatively recently, however they have a low order frequency. Customers in this cluster display the lowest vendor loyalty while the cuisine diversity diverges among the group but also shows greater diversity than other clusters, which means that in general they are open to experiment different vendor and cuisine offers. The most popular cuisine types among the group are Asian, American and Italian. They have a preference for card payments and the majority do not use promotion codes. The majority of orders are placed at lunch or afternoon snack.

Segment 2 - High spenders and short-term customers (9.89%):

Economic Perspective: This group represents high spending customers, they stand out for the highest average spending per order, average price per product, and average products per order. Specifically, the average price per product is greatly larger than in customers from other segments. This means they not only buy more expensive products, but they also buy larger quantities than other clusters.

Demographic Perspective: The vast majority of customers in this cluster come from region 8670, with a smaller yet notable representation from region 4660. There is a specific regional concentration on this cluster that stands out from the remaining.

Preference-Based Perspective: This customer segment is mostly composed of customers that have very short active periods, while there are a minor number of individuals with higher active periods. The recency varies greatly with most customers not being active for quite some time. Given that the majority of individuals in this group have only done one unique order this indicates one-off purchases spread throughout the database's 3-month period. As such, there are no specific conclusions about vendor loyalty, order frequency and cuisine diversity for this group. They also tend to order on weekends more than other clusters, while still having a preference for weekdays. Most of the customers in this group use promotion codes, especially "DELIVERY" and there is a diverse choice of payments, which includes card, cash and digital payments. The preferred mealtimes are breakfast, lunch and late night, while it's very rare to order for dinner. The preferred cuisine type stands out to be Asian, while there is a notable preference for street food and snacks as well.

Segment 3 – Balanced and long-term customers (7.74%):

Economic Perspective: These individuals show the lowest average price per product and average expenditure per order, while the number of products per order is also low but similar to other groups. This indicates that there is a tendency for cheaper alternatives as well as low quantities ordered at each time.

Demographic Perspective: Almost half of the customers in this cluster are from region 2360, with significant representation from region 4660 as well.

Preference-Based Perspective: This segment has long-term customers that have recent activity and have joined the platform a long time ago. They have a slightly higher order frequency than other clusters. The individuals of this group have a diverse behavior for vendor loyalty and cuisine diversity. The majority seem to be more inclined to order in a specific group of vendors, however they stand out for the most inclination to have a very diverse of choice of cuisine types. They order the most at lunch, while they are still likely to order at dinner and afternoon snack. The preferred payment method is by card, and they are unlikely to use promotion codes. This groups ordered heavily during weekdays and have a preference for Asian cuisine, American cuisine and the category "Others".

Segment 4 – One-time buyers and promotion driven customers (16.7%):

Economic Perspective: These individuals are average spenders for both the total cost of the purchase and the individual items. However, it's notable that most customers only order one product.

Demographic Perspective: Customers are almost evenly split between regions 2360 and 4660, these being the most popular regions.

Preference-Based Perspective: The great majority of these customers have only been active on the platform for 1 day or an extremely short time, this takes place during the interval of 3 months captured by the database. Almost all individuals have only ordered once which point to one-off purchases with a single item that then resulted in abandoning the platform, as such there are no specific conclusions about vendor loyalty, order frequency and cuisine diversity. These customers mostly order for lunch, breakfast and dinner, and they show a bigger preference for weekend orders than in any other cluster

while still making most of the orders during the week. The cuisine types they order the most are Asian, American and Italian. Of all segments, this group has the most clients that use promotion codes with a specific preference for “DELIVERY” followed by “DISCOUNT”, which could be the reason behind the one-off purchases. These individuals also show a stronger preference for cash payments than in any other cluster, not much lower than the number of card payments.

3.1. Business Application Suggestions

For **Cluster 0**, these are **vendor loyal customers that show moderate spending activity**. We recommend a collaboration between the platform and vendors to offer exclusive deals and loyalty rewards for preferred vendors. Highlight these partnerships in app notifications and email campaigns. Focus marketing efforts on weekdays, promoting lunch and breakfast specials. Encourage continued usage with personalized meal suggestions and occasional discounts on their favorite cuisines: Asian, American and Italian.

Cluster 1 is composed of **moderate customers with varied preferences.** We could appeal to their diversified taste by offering a variety of cuisine options through themed weeks. Use push notifications to introduce new vendors and menu items to encourage them to explore the platform. Additionally, offer points for trying different cuisines and leverage their preference for lunch and afternoon snack orders with occasional discounts around these meal times.

Cluster 2 represents **premium customers with briefly behavior**. We could highlight premium offerings to match their high spending habits and offer discounts for high number of products. Additionally, offer exclusive weekend promotions, as weekend proportion is high in this cluster. The majority do one-off purchases and preferred discount code is “DELIVERY”, so we could offer free delivers in the first month to incentive be active in the platform. And because the majority of individuals in this cluster are from a specific region (8670), some campaigns could be created specifically targeting customers in this region, for example outdoor advertising.

For **Cluster 3**, referring to **balanced long-term customers,** that don’t usually spend a lot in their orders. We could reward their loyalty with tiered membership discounts and exclusive early access to new products, since they are unlikely to use promotion codes. Focus on highlining diverse meal options in their preferred mealtime, lunch and dinner, to keep their interest, since they have a high cuisine diversity.

Finally, **Cluster 4** comprises **one-time and promotion-driven customers.** As they are motivated by deals and promotions, personalized welcome campaigns could introduce them to the platform and provide exclusive offers for new customers. A good option would be encouraging referrals with bonuses, that they would be able to spend in later purchases in order to retain them for more time. Another possibility would be sending a message to reactivate inactive customers by offering personalized discounts, leveraging their preference for promotions like "DELIVERY" and "DISCOUNT".

4. CONCLUSION

A crucial part of this project was the preprocessing phase to improve the quality of the dataset. In this phase, we also identified and removed extreme observations that negatively impacted the segmentation quality and scaled the data to prevent any single feature from disproportionately influencing the clustering results. The improvements in feature engineering also significantly contributed to the outcomes, as some of these features were selected as segmentation variables, while the rest provided valuable insights during the profiling phase.

The clustering experimentation process identified the Hierarchical-K-means method as the best choice for our dataset. This approach resulted in 5 clusters with a silhouette coefficient of 0.427, representing the most cohesive and meaningful clusters among all algorithms tested. However, our analysis revealed some issues with the final segmentation, including overlapping areas and a considerable number of points near the borders between clusters, suggesting that the overall cohesion of the groups could be improved. All segmentation features significantly contributed to the segmentation, providing meaningful differences across all clusters, as confirmed by the various methods employed.

We believed that our results could still yield meaningful segments despite the issues we identified. This proved true, and the profiling of each group resulted in several interesting insights. The results led to two major groups, together accounting for more than half of the customers, these groups shared common traits of being moderate spenders with medium active periods; however, one group exhibited highly varied preferences, while the other showed high vendor loyalty and less diverse choices. Additionally, there were two groups with extremely low active periods, seemingly associated with customers making one-time purchases; one group comprised high spenders, while the other was drawn to promotions and opted for cheaper products. The segmentation also identified a minor group of very loyal platform users exhibiting more balanced behaviors. Lastly, the outlier cluster included a wide range of customers associated with extreme values across various characteristics.

In conclusion, the initial goal of segmenting customers into meaningful groups and identifying behavioral trends was achieved. The clustering process provided valuable insights about the customer base and allowed us to develop actionable marketing strategies for each segment.

As future work, this project could be enhanced in several areas. The clustering process might benefit from further feature engineering to enhance available information. Additionally, exploring different outlier handling methods could offer new ways to manage extreme values in a more robust way. Further experimentation with different algorithms could also help identify a better fit for our dataset. Overall, these strategies could lead to more cohesive clusters and improve the quality of the segmentation.

BIBLIOGRAPHICAL REFERENCES

1. Aggarwal, C. C. (2015). Data mining: The textbook. Springer.
2. Analytics Vidhya. (2020, October). Feature selection techniques in machine learning. Retrieved from https://www.analyticsvidhya.com/blog/2020/10/feature-selection-techniques-in-machine-learning/
3. Chen, B., Tai, P. C., Harrison, R., & Pan, Y. (2005). Novel hybrid hierarchical-K-means clustering method (H-K-means) for microarray analysis. In Proceedings of the IEEE Computational Systems Bioinformatics Conference (CSB'05) (pp. 505–506). IEEE. Retrieved from https://ieeexplore.ieee.org/abstract/document/1540560
4. Gautam, R. & Latifi, F. (2023). Comparison of Simple Missing Data Imputation Techniques for Numerical and Categorical Datasets.
5. Han, J., Kamber, M., & Pei, J. (2011). Data mining: Concepts and techniques (3rd ed.). Morgan Kaufmann.
6. Scikit-learn. (n.d.). AgglomerativeClustering. Retrieved from https://scikit-learn.org/dev/modules/generated/sklearn.cluster.AgglomerativeClustering.html
7. Scikit-learn. (n.d.). Clustering. Retrieved from https://scikit-learn.org/1.5/modules/clustering.html
8. Scikit-learn. (n.d.). Compare the effect of different scalers on data with outliers. Retrieved from https://scikit-learn.org/1.5/auto_examples/preprocessing/plot_all_scaling.html
9. Scikit-learn. (n.d.). DBSCAN. Retrieved from https://scikit-learn.org/1.5/modules/generated/sklearn.cluster.DBSCAN.html
10. Scikit-learn. (n.d.). Gaussian mixture models. Retrieved from https://scikit-learn.org/1.5/modules/mixture.html
11. Scikit-learn. (n.d.). KMeans. Retrieved from https://scikit-learn.org/1.5/modules/generated/sklearn.cluster.KMeans.html
12. Scikit-learn. (n.d.). Permutation feature importance. Retrieved from https://scikit-learn.org/1.5/modules/permutation_importance.html
13. Scikit-learn. (n.d.). Selecting the number of clusters with silhouette analysis with K-mean clustering. Retrieved from https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html#sphx-glr-auto-examples-cluster-plot-kmeans-silhouette-analysis-py
14. Scikit-learn. (n.d.). silhouette_score. Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html
15. Scikit-learn. (n.d.). StandardScaler. Retrieved from https://scikit-learn.org/1.5/modules/generated/sklearn.preprocessing.StandardScaler.html
16. Scikit-posthocs. (n.d.). scikit_posthocs.posthoc_dunn. Retrieved from https://scikit-posthocs.readthedocs.io/en/latest/generated/scikit_posthocs.posthoc_dunn.html
17. Scipy. (n.d.). scipy.cluster.hierarchy.dendogram. Retrieved from https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.dendrogram.html
18. SciPy. (2014). scipy.stats.spearmanr. Retrieved from https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.spearmanr.html
19. SciPy. (n.d.). scipy.stats.kruskal. Retrieved from https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.kruskal.html
20. SciPy. (n.d.). scipy.stats.zscore. Retrieved from https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.zscore.html

APPENDIX A

Summary of preprocessing steps				
Data types	customer_age, first_order, HR_0: changed to Int64			
Duplicates	Duplicated clients removed			
Incoherencies	customer_region (1.39%): replace "-" with nan	last_promo: replace "-" with "No Promo Code"	vendor_count = 0 & product_count = 0 (138obs.): drop rows	vendor_count != 0 & product_count = 0 (18 obs.): input vendor_count value in product_count
Missing values	customer_age (2.28%): KNN imputer	first_order (0.33%): input 0	HR_0 (3.65 %): calculation of true values	customer_region (1.39%): KNN inputer
Outliers	For a Z-score > 7 in any feature: observation removed			
Scaling	Standardization of data using Standard Scaler			

Table 1 - Summary of preprocessing steps

New Feature Name	Feature Pseudo code	What is it?
Weekend_orders	DOW_0 +DOW_6	Captures the total orders placed from Monday to Friday.
weekday_orders	DOW_1 + DOW_2 + DOW_3 + DOW_4 + DOW_5	Captures the total orders placed on Saturday and Sunday.
total_spent	cui_columns.sum	Indicates the total spending amount for each customer across all orders.
total_orders	DOW_0 + DOW_1 + DOW_2 + DOW_3+ DOW_4 + DOW_5 + DOW_6	Represents the cumulative number of orders placed by each customer.
breakfast,lunch,snack, dinner,late_night	Sum the order counts from HR_x to HR_y	Result of grouping number of orders per hour into categories according to mealtimes.
use_promo	last_promo != '-' 1 if the customer used a promotion, 0 if they did not	Indicates whether a customer has ever used a promotion code.
age_group	'young' if age < 25, 'adult' if 25 <= age < 60, 'senior' if age >= 60	Categorizes customer age into three groups: Young (0-24), Adult (25-59), and Senior (60+).
Proportion_weekdays, proportion_weekend	weekend_orders /total_orders weekdays_orders / total_orders	The proportion of orders the client has made for each week time in the total of orders made

Proportion_breakfast, proportion_lunch, proportion_snack, proportion_dinner, proportion_late_night	breakfast/total_orders lunch /total_orders snack /total_orders dinner /total_orders late_night /total_orders	The proportion of orders the client has made for each mealtime in the total of orders made
Avg_spent_per_order	Total_spent /total_orders	Indicates the average value a customer spends per order
Avg_products_per_order	Product_count/total_orders	Indicates the average number of products bought per order
Avg_price_per_product	Avg_spent_per_order/ avg_products_per_order	Indicates on average how much the client spends per product
Recency	90 – last_order	Indicates how recently the customer made their last order relative to the dataset's timeframe
Active_period	Last_order – first_order + 1	Indicates how much time the client was active for the 3-month period
Order_frequency	Total_orders / active_period	Indicates the frequency of orders for the time between first and last order
Cuisine_diversity	Count of nonzero CUI_columns	Represents the number of different cuisines the customer has ordered from
Vendor_loyalty	Vendor_count / total_orders	Measures customer loyalty to the vendors
Weekend_preference	1 if weekend_orders > weedays_orders, 0 otherwise	Indicates the customers' orders more on the weekends or on weekdays
Meal_time_preference	Columns name with max value among meal periods	Indicates the meal time where the customer has ordered the most
Cuisine_preference	Columns name with max value among CUI_columns	Indicates the cuisine type that the customer has ordered the most

Table 2 - New features

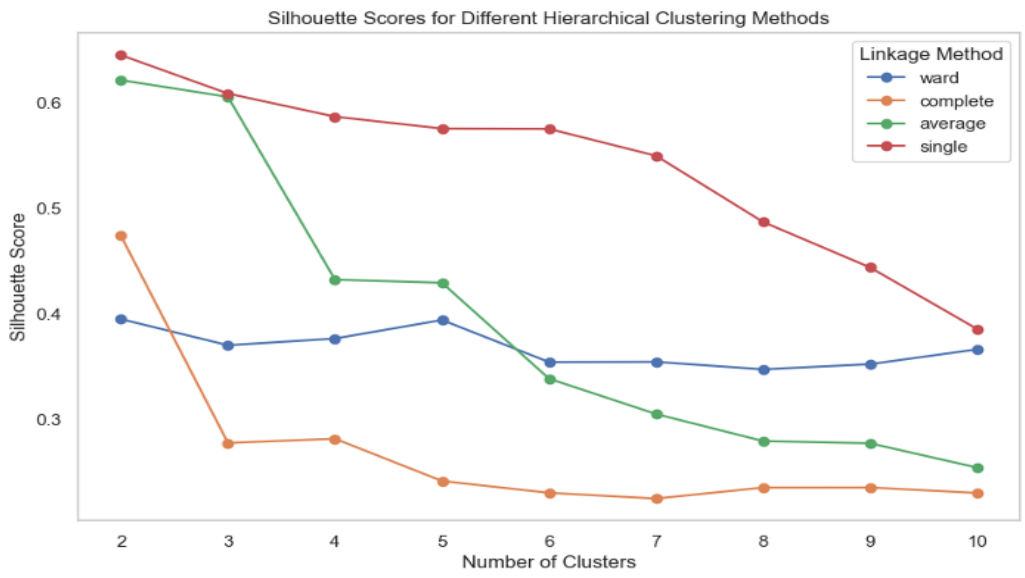


Figure 1 - Silhouette coefficient for different hierarchical clustering methods

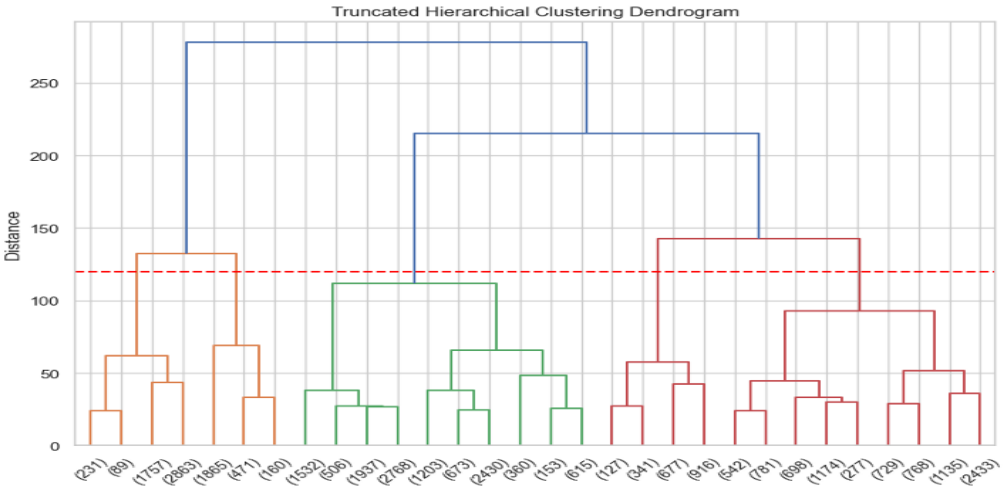


Figure 2 - Truncated Hierarchical Clustering Dendogram

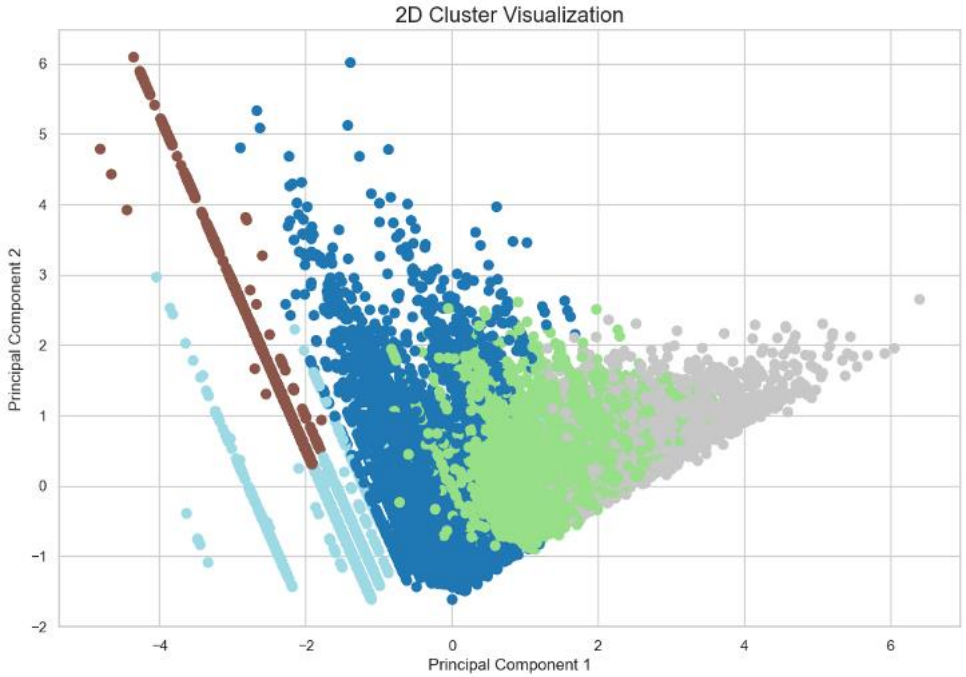


Figure 3 - 2D Scatter plot of Intermediate Cluster Results

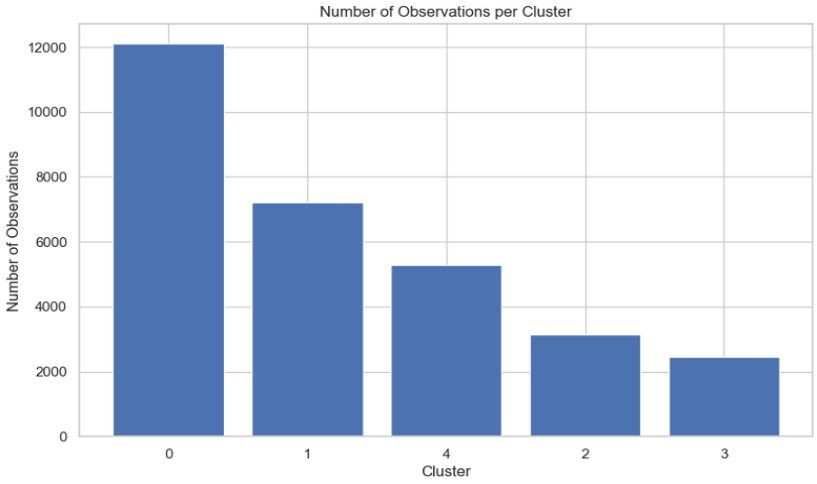


Figure 4 - Bar chart of Number of Observations per Cluster

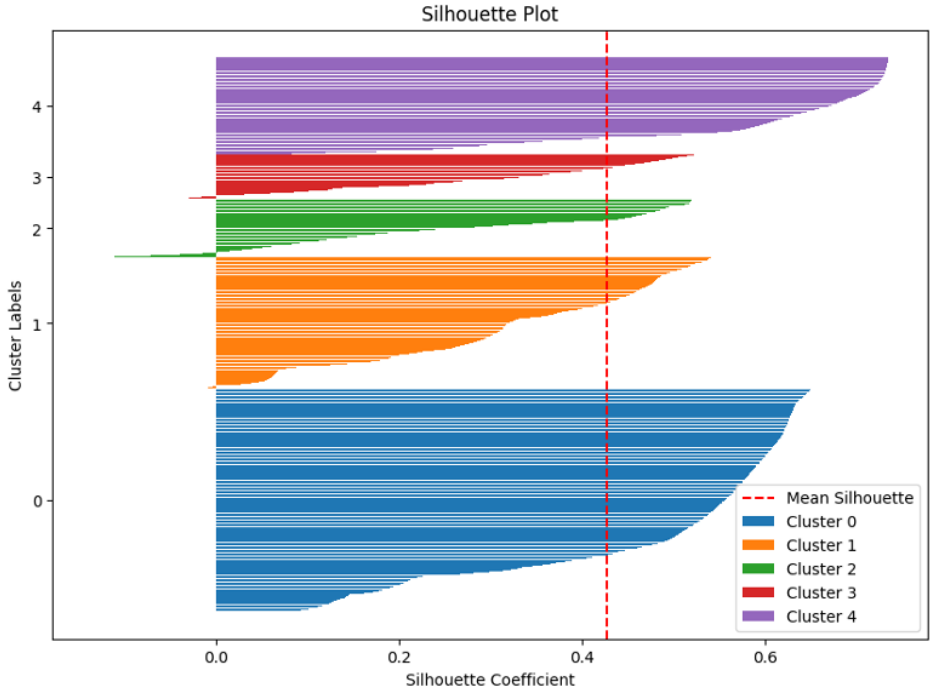


Figure 6 - Silhouette Plot

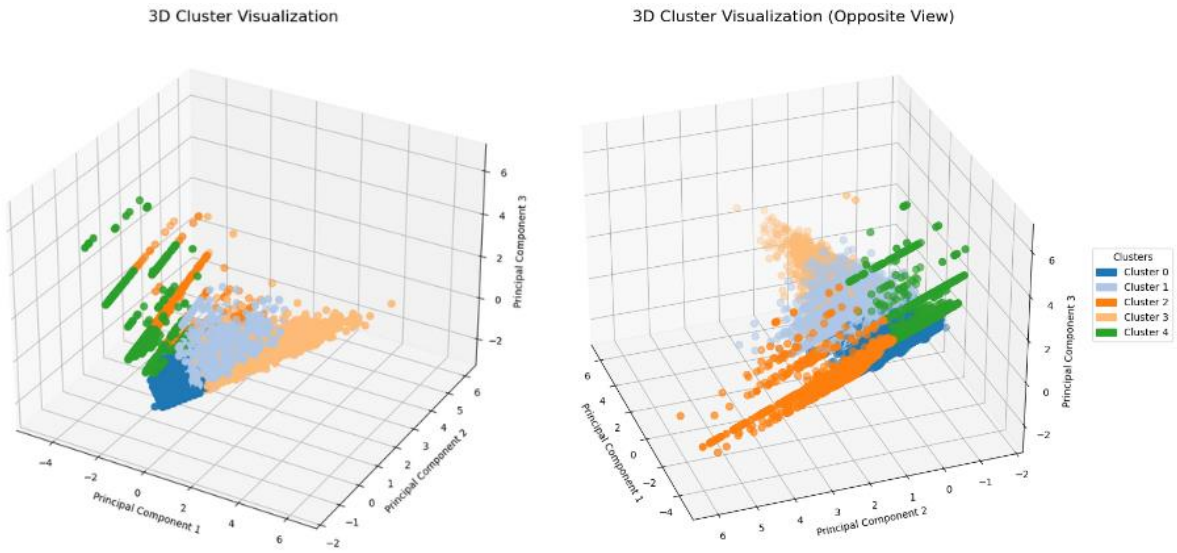


Figure 5 - 3D Scatter plots of Final Cluster Results

"total_orders"	"avg_spent_per_order"	"order_frequency"	"vendor_loyalty"
3.384	2.602	2.215	1.981

Table 3 - Feature range across centroids

"vendor_loyalty"	"order_frequency"	"total_orders"	"avg_spent_per_order"
0.298	0.292	0.185	0.175

Table 4 - Decrease in silhouette score after feature values permutation

“vendor_loyalty”	“total_orders”	“order_frequency”	“avg_spent_per_order”
0.298	0.292	0.185	0.175

Table 5 - H-statistics for Kruskal-Walli’s test

Cluster	“vendor_loyalty”	“total_orders”	“order_frequency”	“avg_spent_per_order”
0	Very high	Medium	Low	Medium
1	Low	Medium	Low	Medium
2	Very high	Low	High	High
3	Medium	High	Medium	Low
4	Very high	Very low	High	Medium
-1	Wide range	Wide range	Wide range	Wide range

Table 6 – Segmentation features Cluster Characterization

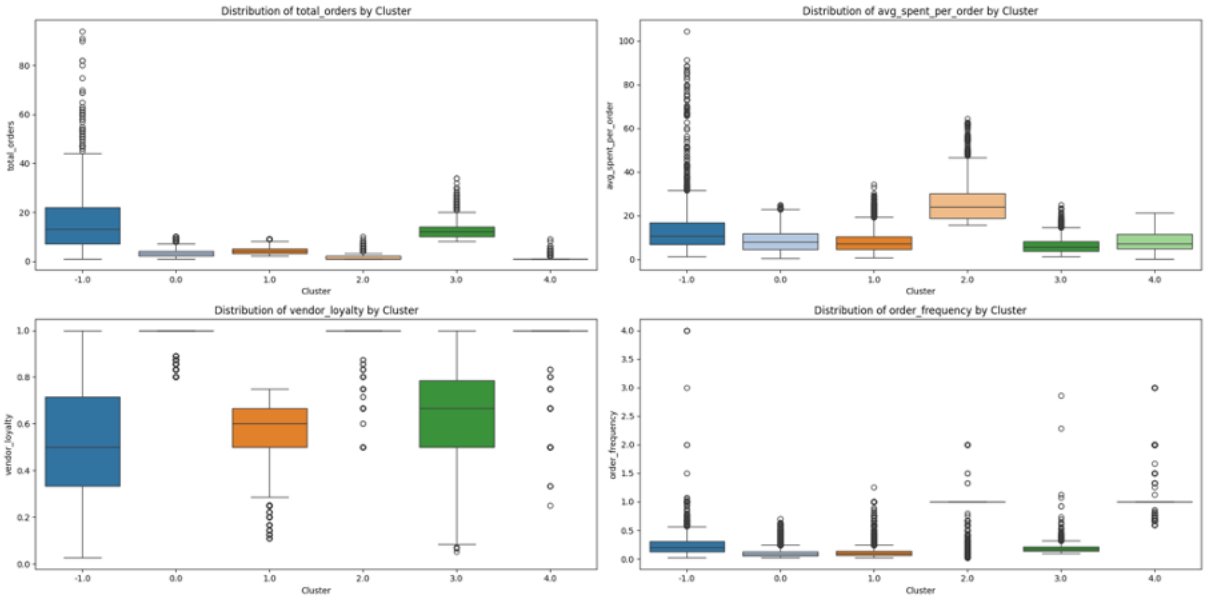
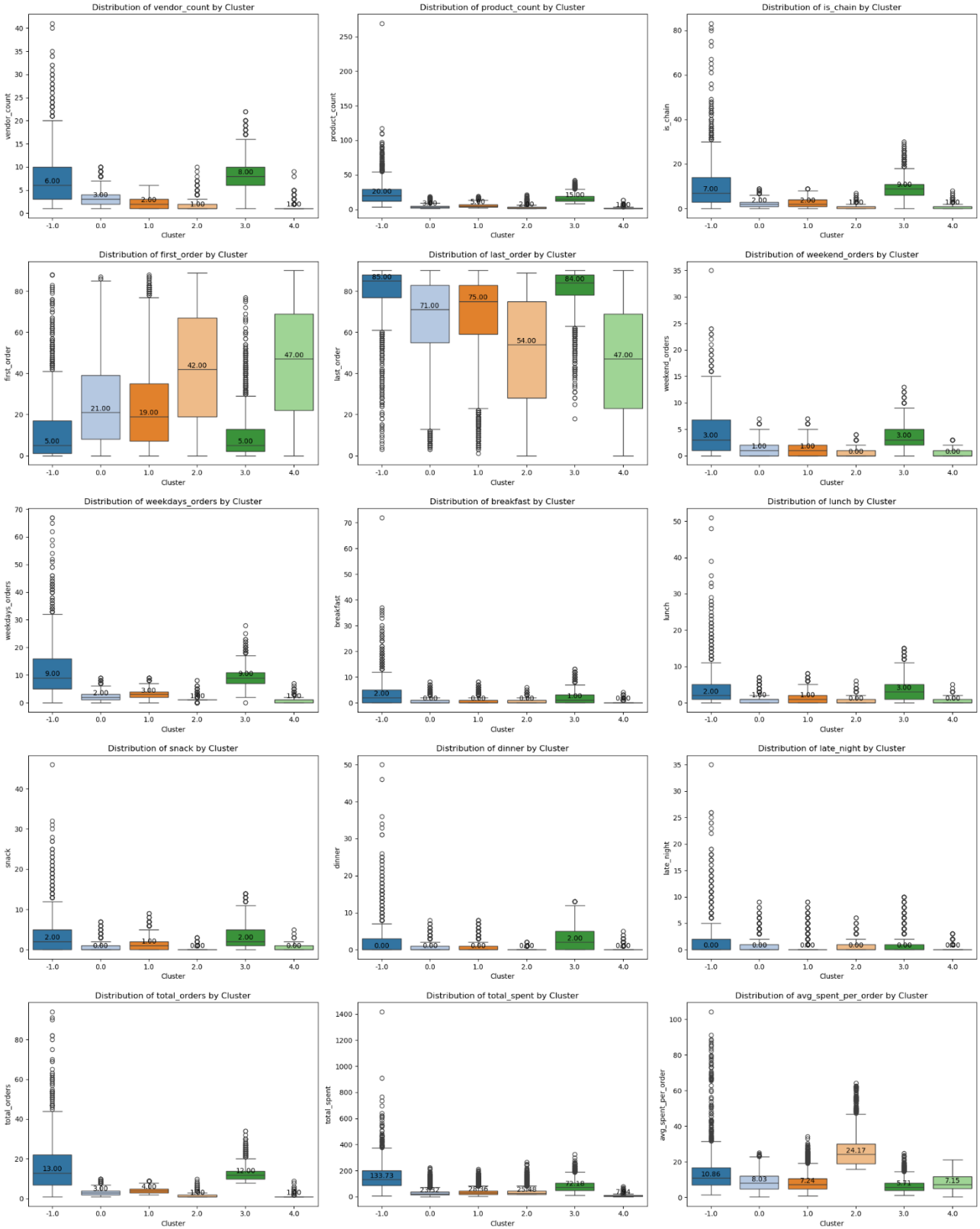


Figure 7 – Box Plots for Each Segmentation Feature Across Clusters


Distribution of vendor_count by Cluster
vendor_count
Distribution of product_count by Cluster
product_count
Distribution of is_chain by Cluster
is_chain
Distribution of first_order by Cluster
first_order
Distribution of last_order by Cluster
last_order
Distribution of weekend_orders by Cluster
weekend_orders
Distribution of weekdays_orders by Cluster
weekdays_orders
Distribution of breakfast by Cluster
breakfast
Distribution of lunch by Cluster
lunch
Distribution of snack by Cluster
snack
Distribution of dinner by Cluster
dinner
Distribution of late_night by Cluster
late_night
Distribution of total_orders by Cluster
total_orders
Distribution of total_spent by Cluster
total_spent
Distribution of avg_spent_per_order by Cluster
avg_spent_per_order
Cluster

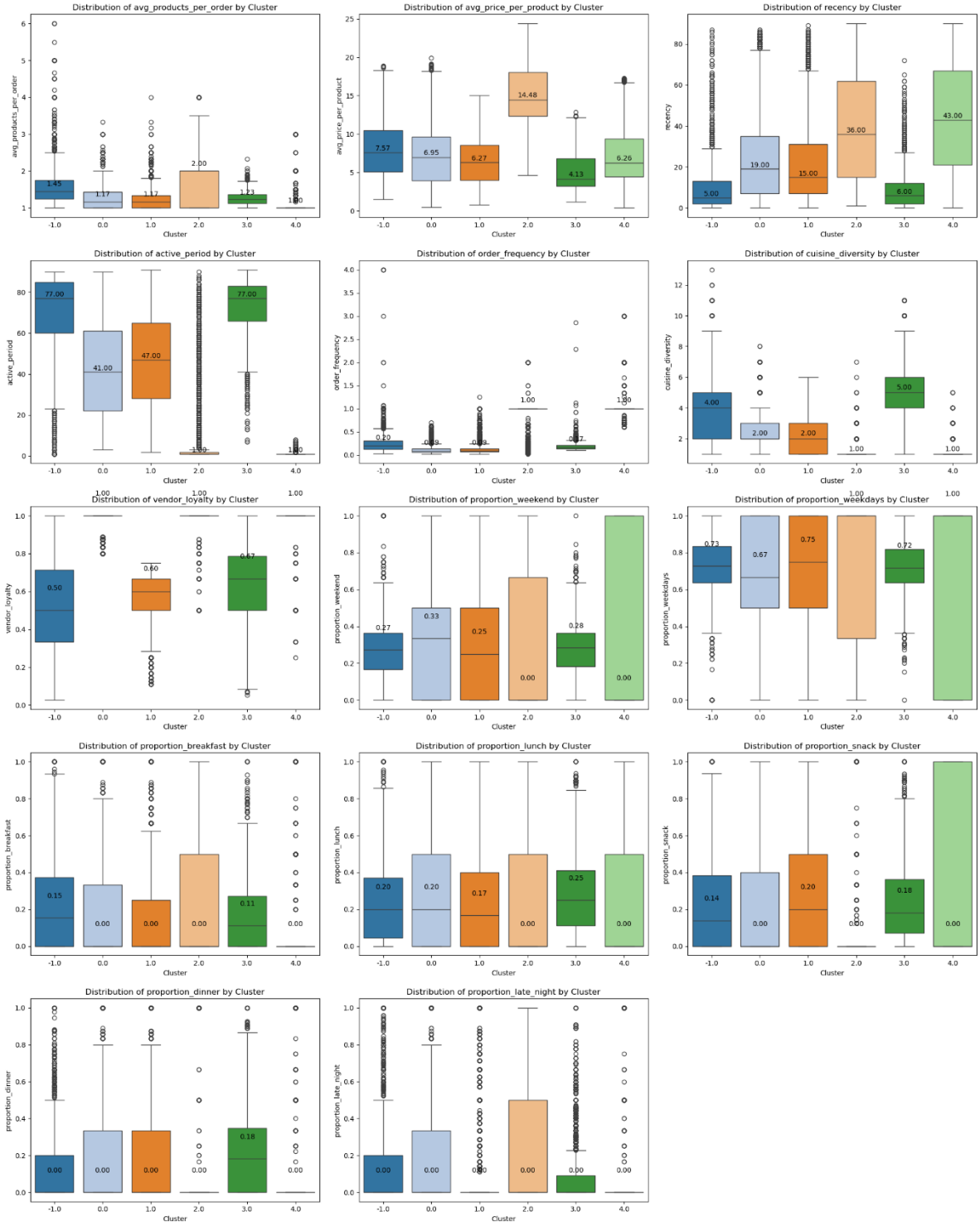


Figure 8 - Boxplots for Each Numerical Feature Across Clusters

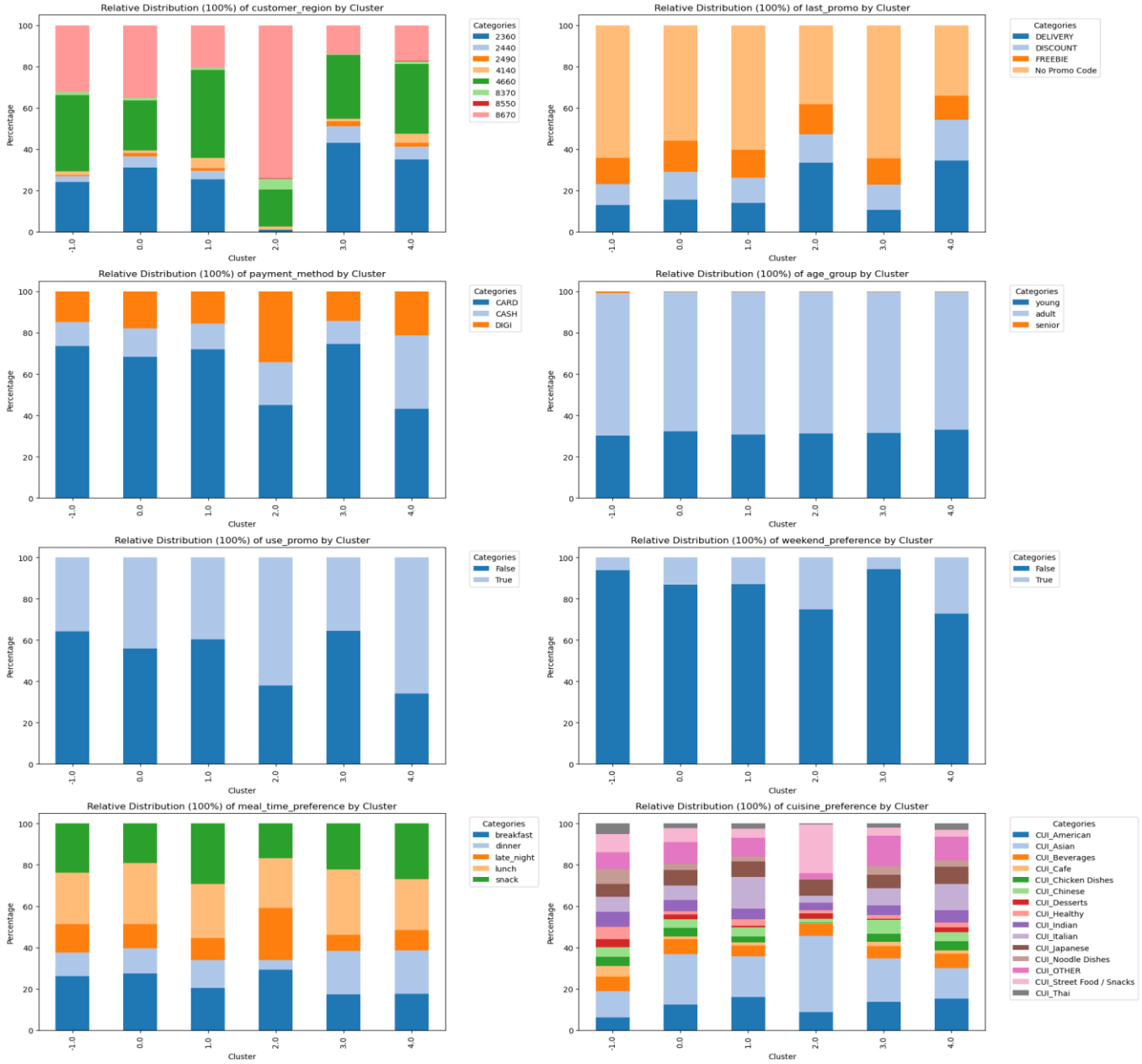


Figure 9 - Stacked Bar Charts with Relative Frequencies of Categoric Features

	Segment 0	Segment 1	Segment 2	Segment 3	Segment 4
Spending	Medium	Medium	High	Low	Medium
Product quantity	Low	Low	High	Low	Very low
Active period	Medium	Medium-High	Short	High	Very short
Recency	Recent	Recent	Varies	Recent	Varies
Order frequency	Low	Low	-	Medium	-
Vendor loyalty	High	Low	-	Medium	-
Cuisine diversity	Medium	High	-	High	-
Mealtime preference	Breakfast Lunch	Lunch Snack	Breakfast Lunch Late night	Lunch Snack Dinner	Breakfast Lunch Dinner
Week time preference	Weekdays	Weekdays	All week	Weekdays	All week
Discount codes	Unlikely	Unlikely	Likely	Unlikely	Very likely
Payment preferences	Card	Card	All	Card	Cash Card
Cuisine preferences	Asian American Italian	Asian American Italian	Asian Street-food	Asian American Others	Asian American Italian
Main regions	8670 2360 4660	8670 2360	8670 4660	2360 4660	2360 4660

Table 7 - Summary of Main Characteristics of each segment