Data Mining Project

Master in Data Science and Advanced Analytics

NOVA Information Management School

Universidade Nova de Lisboa

ABCDEatsInc. - Data-driven Customer Segmentation for Business Growth

Group 46

Burcu Yesilyurt, 20230763

Arda Kaykusuz, 20230766

Emir Kaan Kamiloglu, 20240945

Fall/Spring Semester 2024-2025

1. Introduction **3**
2. Data Preprocessing **3**
2.1 Handling Missing Values 4
2.2 Dropping Unnecessary Columns 4
2.3 Handling Outliers 4
2.3.1 Handling Outliers With Transformation 4
2.3.2 Without Skewness 5
3. Clustering **5**
3.1 Hierarchical Dendogram 6
3.2 K-Means Clustering 6
3.3 Birch Clustering: 6
3.4 Self-Organizing Map (SOM): 6
3.5 DBSCAN Clustering: 7
3.6 Mean Shift Clustering: 7
4. Compare Results **7**
5. Conclusion **9**
5.1 Deep-Dive Analysis for Marketing Strategy 10
5.1.1. Cuisine Preference 10
5.1.2. Day & Hour 10
5.1.3. Regions 11
5.1.4. Payment & Promotion Behaviour 11
5.2 Our Strategy and Goals 11
5.2.1. Cluster 1: High-Value, Potentially Recoverable Customers 11
5.2.2. Cluster 2: Frequent, Consistent Spenders 12
5.2.3. Cluster 3: Fresh, High-Spending Chain Enthusiasts 12
Appendix **13**
Figure 1. Correlation Matrix 14
Table 1. Comparing DBSCAN eps and minPts for outliers 15
Figure 2. Transformation 15
Figure 3. Outliers with skewness 15
Figure 4. DBSCAN - PCA 16
Figure 5. Outliers without skewness 16
Figure 6. Hierarchical Dendogram with Skewness 17
Table 2. Dendogram Comparision 18
Figure 7. Optimal k 18
Table 3. Comparison BIRCH vs K-Means 18
Figure 8. SOM Clustering 19
Figure 9. DBSCAN Clustering 19
Figure 10. KNN Distance Graph 20
Table 4. KNN Distance Graph 20
Figure 11. Comparison for Clusters 21

Table 5. Conclusion.....21
Table 6. Cuisine Preference..... 22
Table 7. Day Preference.....22
Table 8. Hour Preference.....22
Table 9. Region..... 22
Table 10. Payment Method Preference..... 23
Table 11. Promotion Usage.....23
Annexes..... 23

1. Introduction

Customer segmentation is important in helping businesses understand their customer base and improve marketing strategies. This report analyzes customer data using unsupervised machine learning techniques, specifically comparing the performance of six popular clustering algorithms: Hierarchical Clustering, K-Means, Birch, Self-Organizing Maps (SOM), DBSCAN, and Mean Shift.

The analysis starts with data preprocessing, which includes handling missing values through mean/median imputation and encoding categorical variables. A significant part of the preprocessing involved identifying and dealing with outliers, which can influence clustering results. Various methods, such as IQR, Z-Score, DBSCAN, and visual analysis, were tested to find the most appropriate approach. Additionally, we explored how skewness in the data impacted outlier detection and affected the final clustering results.

After preprocessing the data, the six clustering algorithms were applied. The methodology for each algorithm was explained.

2. Data Preprocessing

In preparation for cluster analysis, additional columns were integrated into the dataset to enhance segmentation insights. During this process, we also noticed that some of our customers were not making purchases and we fixed/removed the issue before we started. This anomaly was addressed by cleaning the dataset to exclude or appropriately handle these cases, ensuring data consistency and reliability.

The primary objective of the preprocessing stage, several key steps were undertaken to prepare the data for analysis and clustering. The primary goal was to clean, transform, and scale the data to ensure it was suitable for modeling and to generate meaningful insights. The initial dataset had undergone some preprocessing in an earlier phase, and the resulting cleaned data was saved as a CSV file. We began by loading this dataset and inspecting it for missing values. The columns with missing data included:

- **customer_age**: 727 missing values
- **first_order**: 106 missing values
- **total_spend:** 138 missing values
- **average_order_revenue**: 138 missing values
- **age_group**: 727 missing values
- **first_order_rearranged**: 106 missing values
- **weeks_after_first_order**: 106 missing values
- **Is_chain_percentage:** 63 missing values

After checking missing values we created ***cui_df*** to keep cuisine preference for each customer. In the Exploratory Data Analysis step, we already created **cui_x_percentage** for each cuisine, where *x* denotes a specific cuisine, to represent the percentage of each customer's preference for that cuisine. These variables were derived during the Exploratory Data Analysis (EDA) phase and further validated for consistency.

Additionally, we dropped the ***'weeks_after_first_order', 'weeks_after_last_order', 'first_order', 'last_order',*** and ***'is_chain'*** columns and kept customers who have at least one order.

2.1 Handling Missing Values

We started by handling missing values. As we mentioned *customer_age, first_order, average_order_revenue, age_group, first_order_rearrenged, and weeks_after_first_order* columns have missing values. To address the missing values in the dataset, we applied the following strategies:

- **customer_age**: Missing values were filled with the mean of the available ages. This approach was chosen because age is a continuous variable, and the mean minimizes the risk of distorting the overall distribution.
- **first_order_rearranged**: These temporal columns had missing values replaced with the ***median*** to ensure robustness against outliers and better representation of central tendencies.
- **average_order_revenue**: Missing values were replaced with the mean to maintain the numerical integrity of the data.

2.2 Dropping Unnecessary Columns

We removed columns that were irrelevant to the customer segmentation process or had no predictive value: We handled this step by changing the data types of the columns in the previous steps. Additionally, we wanted to see the linear correlations between numerical features using a correlation matrix. (Figure 1)

2.3 Handling Outliers

Handling outliers is a critical step in data preprocessing, as extreme values can distort results and negatively impact model performance. By detecting and addressing outliers, we can enhance the robustness of clustering models and ensure more reliable insights. For this project, we explored four different approaches to identify outliers: ***IQR (Interquartile Range), Z-Score, visual analysis,*** and ***DBSCAN***.
We applied these approaches using three different strategies:

- Applying skewness transformations first, followed by outlier detection.
- Detecting outliers without addressing skewness.
- Detecting and handling outliers first, then addressing skewness.

2.3.1 Handling Outliers With Transformation

We analyzed the skewness for each numeric column and applied three transformations: original skewness, log, and square root to reduce the skewness (Figure 2). For each column, we created a function to select the best transformation based on skewness values and visualized the effects of each transformation. After this process, we checked outliers and visualized them by boxplot (Figure 3) Detected outliers using 4 methods:

- **IQR (Interquartile Range):** Outliers were identified by computing the lower and upper bounds based on the IQR method. Values outside these bounds were flagged as outliers. This method identified ***5.97% of the data*** as outliers.
- **Z-score:** Were calculated, and any values with Z-scores exceeding ±3±3 (more than 3 standard deviations from the mean) were flagged as outliers. This method identified ***2.01% of the data*** as outliers.
- **DBSCAN:** We set epsilon to 0.2 and min_samples to 3. Points labeled as -1 were considered outliers. We selected these values after trying different values. This method flagged only ***0.03% of the data*** as outliers, making it the most conservative approach.
    - **Principal Component Analysis (PCA)** was used to reduce the dataset's dimensions to four components, retaining the most significant variance. We tested a grid of eps values [0.2,0.3,0.4,0.5] and min_samples values [3,5,10], evaluating each combination using the number of clusters, the number of noise points, and visual inspection of the clustering results. We created a table to compare the results. (Table 1) We have not used PCA since it is customer segmentation, we have tried multiple combinations, and we tried to fine-tune, unfortunately, we could not reach the desired level. (Figure 4)
- **Visual Analysis:** Outliers were manually identified using bar plots and custom-defined ranges. Based on visual inspection, ***10.02% of the data*** were marked as outliers.

After comparing these methods, we selected a Z-score with the transformation of some variables since we identified the fewest outliers, minimizing information loss while still addressing extreme values effectively [1].

2.3.2 Without Transformations

In this approach, we skipped skewness transformations and directly applied outlier detection methods. Outliers were visualized using boxplots (Figure 5). Detected outliers using 4 methods:

- **IQR (Interquartile Range):** ***19.9% of the data*** were outliers according to the IQR method.
- **Z-score:** ***25.1% of the data*** were outliers according to the Z-score method.
- **DBSCAN:** epsilon is set to 0.3, and min_samples is set to 5. DBSCAN labels outliers as -1. ***0.05% of the data*** points are considered outliers according to DBSCAN.
- **Visual Analysis:** ***99.96% of the data*** were marked as outliers based on a visual analysis.

Given the high proportion of outliers detected by IQR and Z-Score, as well as the unreliability of visual analysis, we again opted for DBSCAN, as it identified only a minimal percentage of outliers, thereby preserving most of the data.

3. Clustering

For clustering, we created a function to scale our datasets. We applied scale to our three datasets: the skewness dataset, the without skewness dataset, and the skewness after removing the outliers dataset. Then we checked the hierarchical dendrogram, optimal k, and silhouette score for three of them to pick the best one.

We implemented six clustering techniques and then compared each algorithm. The following techniques are used: Hierarchical Clustering, K-Means, BIRCH, Self-Organising Maps, DBSCAN, and Mean-Shif. Feature scaling is essential for algorithms sensitive to feature scales, such as K-Means and SOM. The scaled data and feature names are stored in ***scaled_data*** and ***feature_names***, respectively.

3.1 Hierarchical Clustering

For a skewed dataset optimal ***k = 3*** and ***score = 243*** (Figure 6) which is the best one. We implemented 3 linkage methods: ***ward, complete,*** and ***average***. To determine the best hierarchical clustering method and the optimal number of clusters (kk), we evaluated clustering performance using three widely recognized metrics: ***Silhouette Score, Calinski-Harabasz Score,*** and ***Davies-Bouldin Score***. Best Method: ***Ward***, Best ***k: 3***, Best Silhouette Score: ***0.243*** (Table 2)

3.2 K-Means Clustering

To find the optimal number of clusters we used various evaluation metrics and methods. Several clustering quality measures were computed, such as ***Inertia, Silhouette Score, Davies-Bouldin Index,*** and the ***Gap Statistic***, to identify the most meaningful clustering configuration. The optimal number of clusters determined ***k = 3*** based on the combined results of the Elbow Method, Silhouette Scores, Davies-Bouldin Index, and Gap Statistic. (Figure 7)

3.3 Birch Clustering

Birch (Balanced Iterative Reducing and Clustering using Hierarchies) is a clustering algorithm particularly suitable for large datasets. It builds a Clustering Feature Tree (CFT) to summarize the data and then performs clustering on the leaf nodes of the CFT. Birch clustering was applied with ***n_clusters = 3,*** and cluster labels were assigned. We tried but we didn't use it because it's useful for huge datasets and we set the cluster number as we found optimal k in k-means [3]. You can find a comparison table in the appendix. (Table 3)

3.4 Self-Organizing Map (SOM)

The SOM was trained on the scaled data using a 7x7 grid, random initialization, a Gaussian neighborhood function, batch training, and a hexagonal lattice. Feature names were provided to enhance the interpretability of the SOM. The SOM achieved a ***Topographic Error*** of ***0.0644***, indicating strong topological preservation, with similar data points mapped to neighboring neurons. The Best Matching Unit (BMU) for each data point was identified, and a clustering scheme was implemented based on the BMUs. The dataset was clustered into ***4 distinct groups*** using a modulo operation to assign cluster IDs, ensuring that all data points were consistently assigned to one of the ***4 clusters***. (Figure8)

3.5 DBSCAN Clustering

DBSCAN was applied with ***epsilon=0.25*** and ***min_samples=5***. Epsilon defines the radius around a data point to search for neighbors, and min_samples specifies the minimum number of points

required to form a dense region. Cluster labels were assigned, with -1 indicating noise points (outliers). With these values, we got ***1 cluster.*** (Figure 9)

K-Nearest Neighbors (KNN): While not used for direct clustering in this code, KNN was used to calculate the distances to the 5 nearest neighbors for each data point. This information can be valuable for further outlier analysis by examining points with unusually large distances from their neighbors. (Figure 10)

3.6 Mean Shift Clustering

A crucial parameter in Mean Shift is the bandwidth, which determines the size of the neighborhood used for shifting. We created estimate_bandwidth to automatically estimate a suitable bandwidth based on the data's density. The quantile parameter controls the fraction of data points used to estimate the bandwidth. A lower quantile results in a smaller bandwidth.

Mean Shift clustering was performed using the estimated bandwidth (***bandwidth = 0.6947***), and cluster labels were assigned to each data point. The bin_seeding=True parameter accelerates the convergence of the algorithm. Additional experiments with smaller bandwidth values (***0.1, 0.05, and 0.02***) showed that the number of clusters increases significantly as the bandwidth decreases, resulting in ***14630, 23931,*** and ***28640 clusters,*** respectively. This highlights the sensitivity of the Mean Shift algorithm to the bandwidth parameter.

4. Compare Results

The ***hierarchical clustering*** with ***Ward*** linkage was the most appropriate for itself, selecting ***k = 3*** as an optimum number of clusters with a ***Silhouette Score*** of ***0.243***. When we checked ***K-Means***' clusters we obtained similar results, though a bit higher in its ***Silhouette Score*** of ***0.273***. ***SOM*** maintained the topological relationships and grouped the data into ***4 clusters***, providing different insights and less agreement with the other techniques. On the other hand, ***DBSCAN*** and ***Mean Shift*** did a ***poor*** job, having a uniform density of data points and very sensitive parameters yielded no meaningful cluster generation (Figure 11). When we compared the results of the cluster results, we decided to use SOM's output since it showed more meaningful clusters and we could interpret our clusters rather than the other techniques we applied.

After that, we analyzed the mean values of each cluster (Table 4) to understand customer behaviors and identify actionable patterns in purchasing habits. By interpreting these averages, we will try to define meaningful customer segments such as high-value loyal customers, low-engagement users, and re-engaged or mid-tier customers, and tailor some marketing strategies. In the conclusion part, the clusters will be examined in detail.

5. Conclusion

Personalization is a key factor in customer retention, as it allows companies to customize interactions and transactions to match individual customer profiles, thereby increasing satisfaction and fostering loyalty. Effective personalization strategies often involve targeted advertising, tailored promotions, and customized communication based on detailed customer data[2]. This approach enables us to

optimize marketing efforts, build stronger relationships, and enhance retention rates. In this section, we will dive into cluster details and look into their purchase habits. We start with cluster definitions.

Cluster 0: Inactive, Low-Value Chain Lovers
Inactive customers who previously preferred ordering from chain restaurants but had low average order values. They rank 3rd in order frequency, with an average of 17 days between orders. Due to their inactivity and historically low spending, this cluster might not generate a substantial Return on Investment (ROI) through targeted marketing efforts.

Cluster 1: High-Value, Potentially Recoverable Customers
High-value customers who tried the application briefly but stopped using it. They have the highest average order value, indicating significant potential to impact revenue if re-engaged. Since they do not show a strong preference between chain and non-chain restaurants, this cluster offers flexibility in targeting. A well-executed remarketing strategy could yield a strong ROI by regaining their activity.

Cluster 2: Frequent, Consistent Spenders
This cluster comprises customers with the 2nd most recent last order dates, making them the "second freshest" group. They exhibit the shortest time between orders and frequently place orders, which compensates for their moderately lower average order value. They are the 2nd highest spending cluster overall. To retain these customers, businesses should implement loyalty programs and incentives. Additionally, offering discounts or coupons with minimum order value requirements could encourage them to increase their average order value.

Cluster 3: Fresh, High-Spending Chain Enthusiasts
This group prefers chain restaurants and has a moderately high average order value but achieves the highest total revenue per customer. They have the most recent last order dates, making them the "freshest" group. Maintaining engagement with this cluster is crucial. Focus on retaining them through loyalty programs, exclusive promotions, and personalized offers to keep them actively ordering.

Based on our analysis of the cluster groups, we identified and described distinct customer segments. After careful consideration, we decided to exclude Cluster 0 from our marketing operations. The primary reason is that customers in Cluster 0 can be classified as churned customers. While reactivating churned customers is sometimes more cost-effective than acquiring new ones, this cluster has a very low average revenue per order. Additionally, the low revenue is not compensated by frequent orders, making reactivation efforts less viable from a return on investment point of view.

To refine our marketing strategy for the targeted clusters (***Cluster 0, Cluster 1, Cluster 2, Cluster 3***), we plan to conduct a deeper analysis of these groups' behaviors (Table 5). This analysis will focus on their cuisine preferences, the days and hours they prefer to place orders, their preferred payment methods, and their usage of promo codes. By understanding these behavioral patterns, we aim to tailor our marketing efforts to better align with the unique characteristics of each cluster.

5.1 Deep-Dive Analysis for Marketing Strategy

5.1.1. Cuisine Preference

The data in this table provides a breakdown of which cuisine generates the highest revenue for each cluster (Table 6). By separating the clusters and analyzing their spending patterns on cuisines, we found the following insights:

- **Cluster 1:** Asian cuisine ranks first (26.62%), followed by Italian (13.35%) and American (11.69%).
- **Cluster 2:** Asian cuisine also takes the lead (22.94%), with Italian (10.74%) in second place and Street Food & Snacks (10.14%) in third.
- **Cluster 3:** Asian cuisine is again in first place (23.43%), followed by American (14.71%) and Italian (9.97%).

This analysis highlights that **Asian cuisine** consistently generates the highest revenue across all three clusters**,** making it a **main product for our marketing campaigns.** This result aligns with our expectations**,** as Asian cuisine had the highest overall revenue based on our EDA. Following Asian cuisine, **American and Italian cuisines** occupy the second and third spots, respectively, with **Street Food & Snacks** appearing once in third place for Cluster 2.

Based on these findings, our marketing campaign will focus on these four cuisines, with Asian cuisine as the flagship, ensuring our efforts are aligned with customer preferences and revenue potential.

5.1.2. Day & Hour

We then move on to analyze the Day & Hour ordering patterns of customers (Table 7).

- **Cluster 1**: 71.8% of orders are placed on weekdays, and 28.2% on weekends.
- **Cluster 2**: 70.3% of orders are on weekdays, and 29.7% on weekends.
- **Cluster 3**: 66.49% of orders are on weekdays, and 33.51% on weekends.

Most orders across all clusters occur on weekdays. However, this pattern aligns with the expected outcome since there are fewer days on the weekend. Therefore, no specific focus on weekdays or weekends is necessary for the marketing campaign at this stage (Table 8).

- **Cluster 0** (for reference): Most orders occur in the ***Afternoon*** (25.95%), followed by ***Lunch*** (24.92%), and ***Breakfast*** (21.13%).
- **Cluster 1**: The ***Afternoon*** sees the highest activity (28.37%), followed by ***Late Night*** (23.21%) and **Breakfast** (22.83%).
- **Cluster 2**: Similar to Cluster 1, the ***Afternoon*** dominates (27.27%), with ***Late Night*** (17.51%) and ***Lunch*** (21.48%) following.
- **Cluster 3**: The ***Afternoon*** also leads (25.89%), followed by ***Breakfast*** (24.99%) and ***Late Night*** (20.9%).

Timing: Campaigns should target customers during ***Breakfast, Afternoon,*** and ***Late Night,*** as these are peak ordering times for most clusters.

Days: While no specific day focus is needed initially, ongoing monitoring of weekday vs. weekend order trends post-campaign will help identify any shifts in behavior and optimize future strategies.

5.1.3. Regions

- ***Region 8670*** consistently appears among the top three regions for all clusters and is the highest volume region for ***Clusters 1*** and ***2***.
- ***Region 4660*** is the highest volume region for ***Cluster 3*** and a significant contributor for ***Clusters 1*** and ***2***.
- ***Region 2360*** holds the third spot across all clusters, indicating consistent but slightly lower activity compared to the top two regions (Table 9).

5.1.4. Payment & Promotion Behaviour

The card payment method is consistent among all clusters we want to target on our marketing campaigns, we can use this information to offer card-specific discounts or cashback deals that could enhance engagement (Table 10).

We can see that a high percentage of customers do not use promo codes, especially customers in ***cluster 3*** with ***%62.11*** of these customers haven't used promo codes (Table 11).

Additionally "***DELIVERY***" is the most used promo between ***cluster 1*** and ***cluster 2***, however, we can see more diversified promo usage in ***cluster 3*** with the most promo used being "***FREEBIE***".

5.2 Our Strategy and Goals

5.2.1. Cluster 1: High-Value, Potentially Recoverable Customers

Goal: Reactivate and retain high-value customers who have the potential to significantly impact revenue.

Strategy:

- **Launch remarketing campaigns** targeting these customers.
- **Offer flexible promotions** (applicable to both chain and non-chain restaurants) to accommodate their diverse preferences.
- **Introduce exclusive deals on Asian, Italian, and American cuisines** to align with their top preferences.
- **Focus promotions on Breakfast, Afternoon, and Late Night time slots**, which align with their peak ordering behavior.
- **Leverage CARD-specific discounts** and cashback offers to incentivize re-engagement.
- **Target Region 8670 and 4660 with location-specific campaigns** to maximize impact, as this is their highest volume region.

5.2.2. Cluster 2: Frequent, Consistent Spenders

Goal: Strengthen loyalty and increase average order value.
Strategy:

- **Implement a loyalty rewards program offering points** for frequent orders, redeemable for discounts or freebies.
- **Provide discounts with minimum order value** requirements to encourage higher spending per order.
- **Focus promotions on Breakfast, Lunch, and Afternoon time slots**, which align with their peak ordering behavior.
- **Introduce exclusive deals on Asian, Italian, and Street Food & Snacks** in targeted marketing efforts.
- **Leverage CARD-specific discounts** and cashback offers to incentivize re-engagement.
- **Target Region 8670 and 2360 with location-specific campaigns** to maximize impact, as this is their highest volume region.

5.2.3. Cluster 3: Fresh, High-Spending Chain Enthusiasts

Goal: Retain and maximize revenue from the highest spending cluster.
Strategy:

- **Offer personalized promotions for chain restaurants**, aligning with their strong preference.
- **Roll out exclusive promotions like limited-time "FREEBIE" deals**, as this is the most popular promo type in this cluster.
- **Maintain engagement with loyalty programs and personalized recommendations** to keep them actively ordering.
- **Focus promotions on Breakfast, Afternoon, and Late Night time slots**, which align with their peak ordering behavior.
- **Leverage CARD-specific discounts** and cashback offers to incentivize re-engagement.
- **Target Region 4660 and 8670 with location-specific campaigns** to maximize impact, as this is their highest volume region.
- **Introduce exclusive deals on Asian, Italian, and American cuisines** to align with their top preferences.

5. References

[1] Jagabathula, S., Subramanian, L., & Venkataraman, A. (2017). A Model-based Projection Technique for Segmenting Customers. http://arxiv.org/abs/1701.07483

[2] Ballı, A. (2024). The Effect of Product Personalization on Consumer Purchasing Intention, Customer Satisfaction, Brand Loyalty and Artificial Intelligence Applications with Machine Learning. Fiscaoeconomia, 8(3), 1240-1263. https://doi.org/10.25295/fsecon.1449755

[3] Peng, K., Zheng, L., Xu, X., Lin, T., Leung, V.C.M. (2018). Balanced Iterative Reducing and Clustering Using Hierarchies with Principal Component Analysis (PBirch) for Intrusion Detection over Big Data in Mobile Cloud Environment. http://doi.org/10.1007/978-3-030-05345-1_14

Appendix

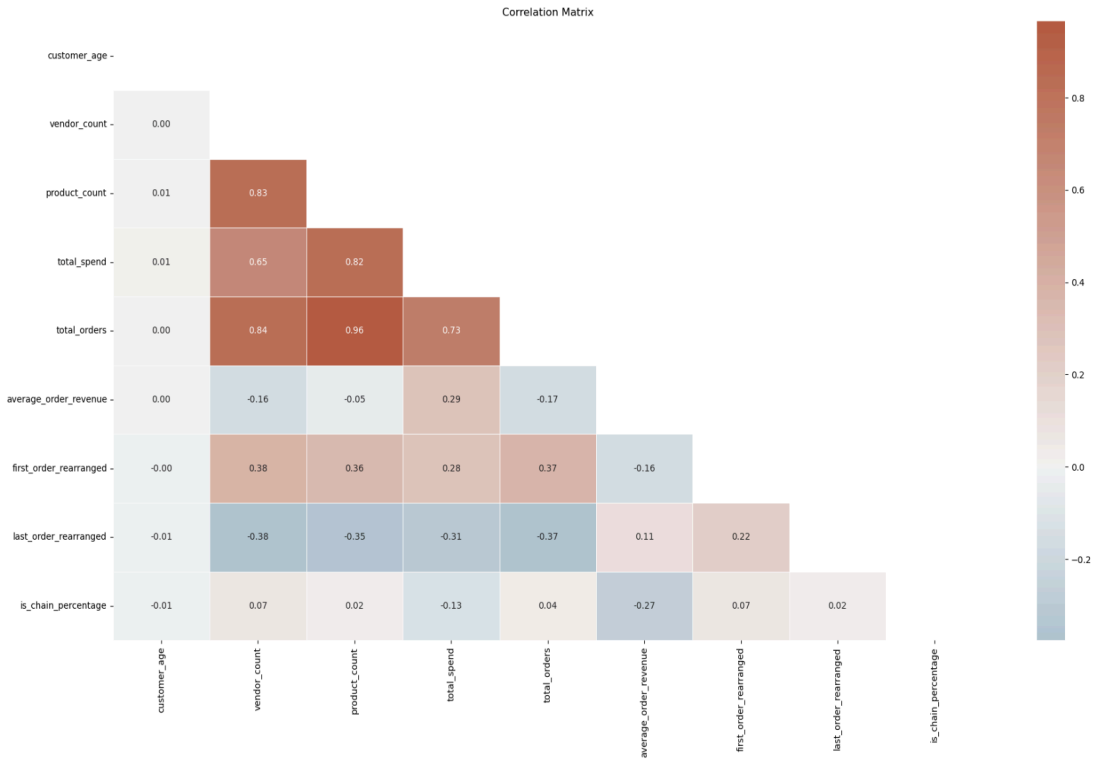

Figure 1. Correlation Matrix

	eps = 0.2	**eps = 0.3**	**eps = 0.4**	**eps = 0.5**
minPts = 3	***Number of clusters:*** 2 ***Number of noise points:*** 294	***Number of clusters:*** 1 ***Number of noise points:*** 10	***Number of clusters:*** 1 ***Number of noise points:*** 0	***Number of clusters:*** 1 ***Number of noise points:*** 0
minPts = 5	***Number of clusters:*** 1 ***Number of noise points:*** 368	***Number of clusters:*** 1 ***Number of noise points:*** 11	***Number of clusters:*** 1 ***Number of noise points:*** 0	***Number of clusters:*** 1 ***Number of noise points:*** 0
minPts = 10	***Number of clusters:*** 1 ***Number of noise points:*** 574	***Number of clusters:*** 1 ***Number of noise points:*** 18	***Number of clusters:*** 1 ***Number of noise points:*** 0	***Number of clusters:*** 1 ***Number of noise points:*** 0

Table 1. Comparing DBSCAN eps and minPts for outliers

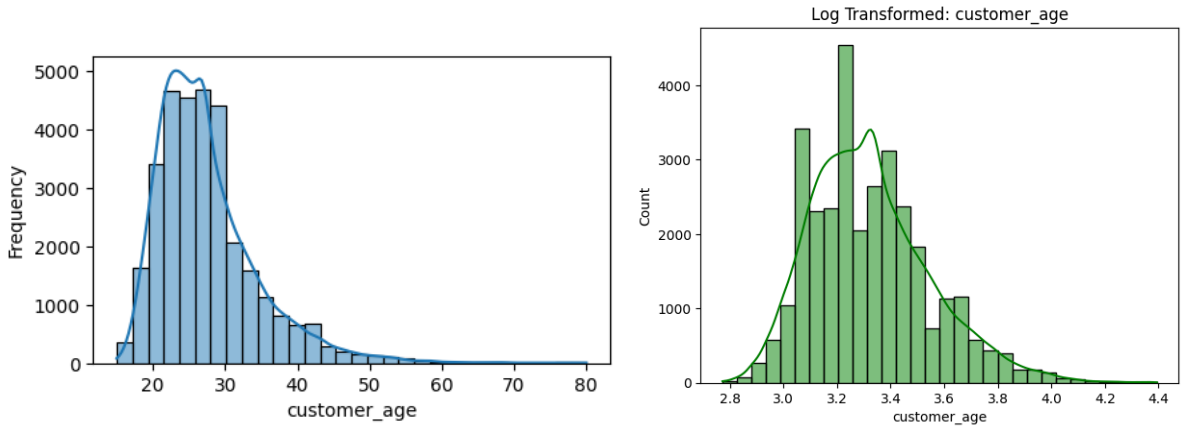

Figure 2. Transformation

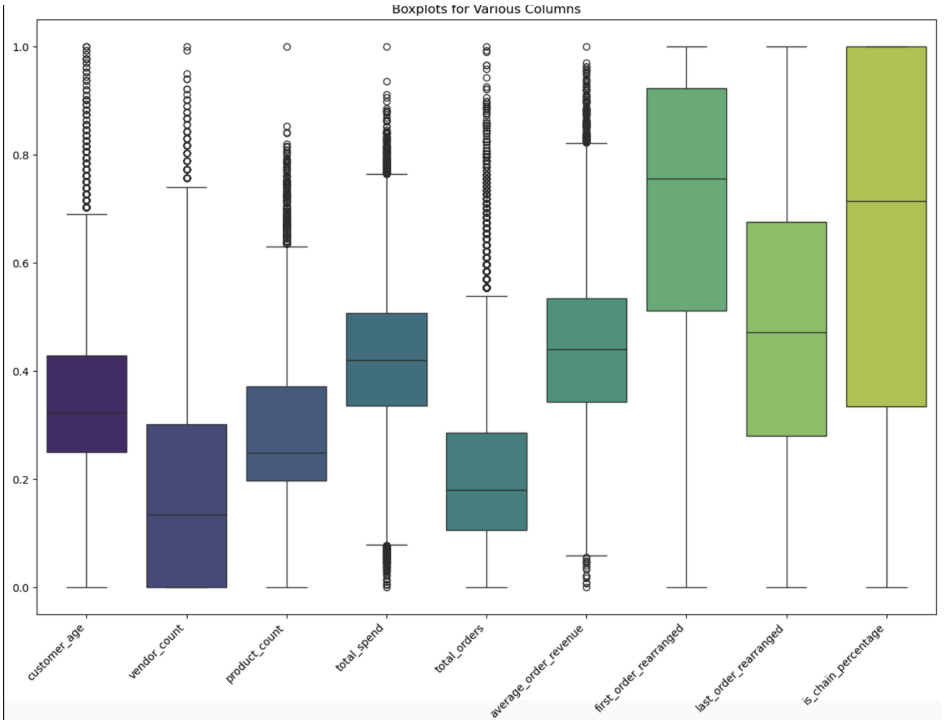

Figure 3. Outliers with skewness

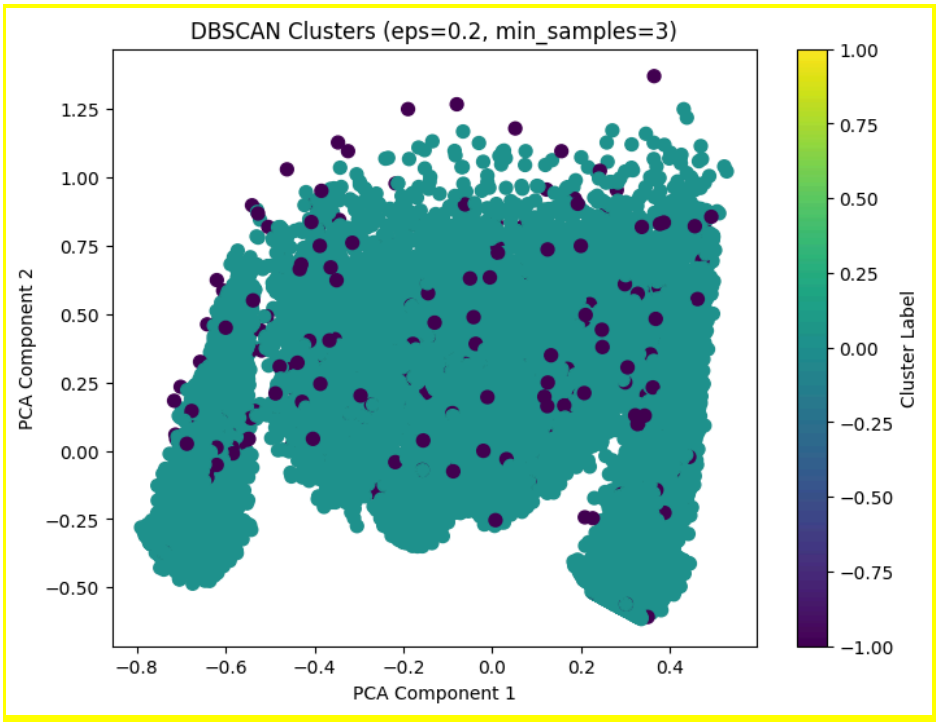

Figure 4. DBSCAN - PCA

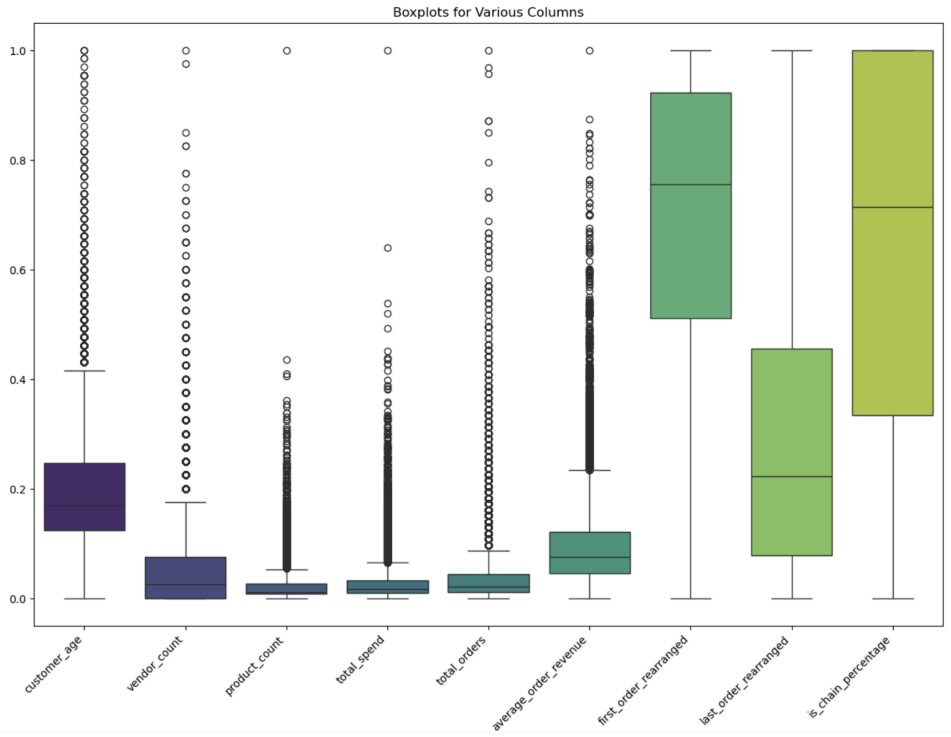

Figure 5. Outliers without skewness

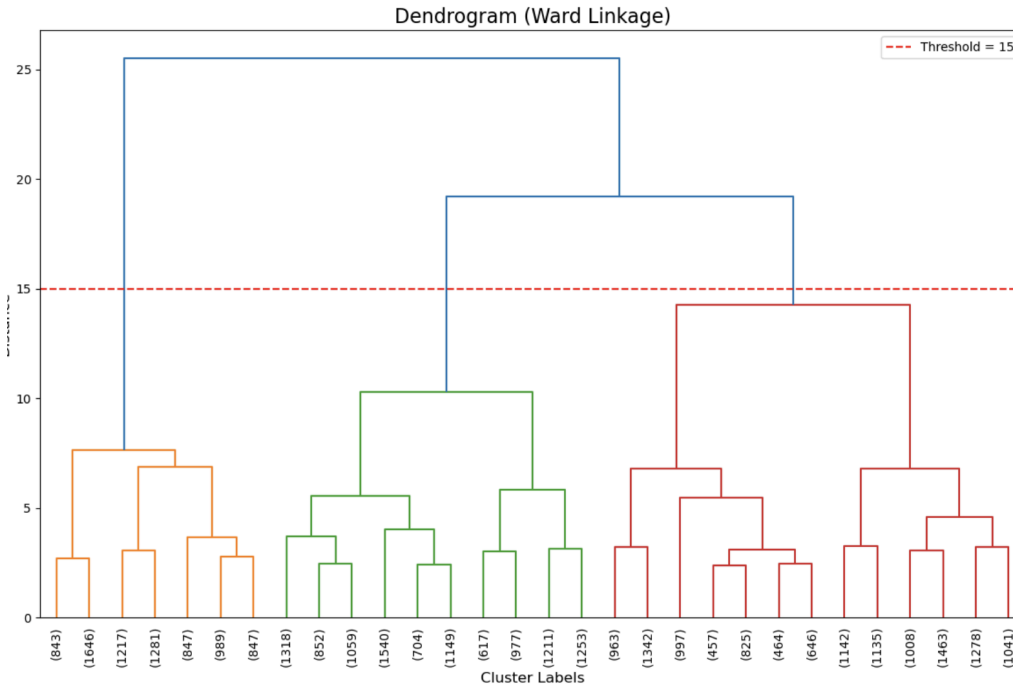

Figure 6. Hierarchical Dendogram with Skewness

Method	k	Silhouette	CH	DB
ward	2	0.204	8425.933	1.502
ward	3	0.243	11002.783	1.399
ward	4	0.198	9939.316	1.641
ward	5	0.173	8500.149	1.779
ward	6	0.144	7555.819	1.780
complete	2	10.174	7635.601	1.951
complete	3	0.106	5538.463	1.915
complete	4	0.169	8113.885	1.644
complete	5	0.136	6719.606	1.588
complete	6	0.114	6186.589	1.841
average	2	0.206	7262.315	1.399
average	3	0.120	3633.609	1.198
average	4	0.173	6282.314	1.184
average	5	0.144	4718.892	1.202
average	6	0.125	3776.069	1.177

Table 2. Dendogram Comparision

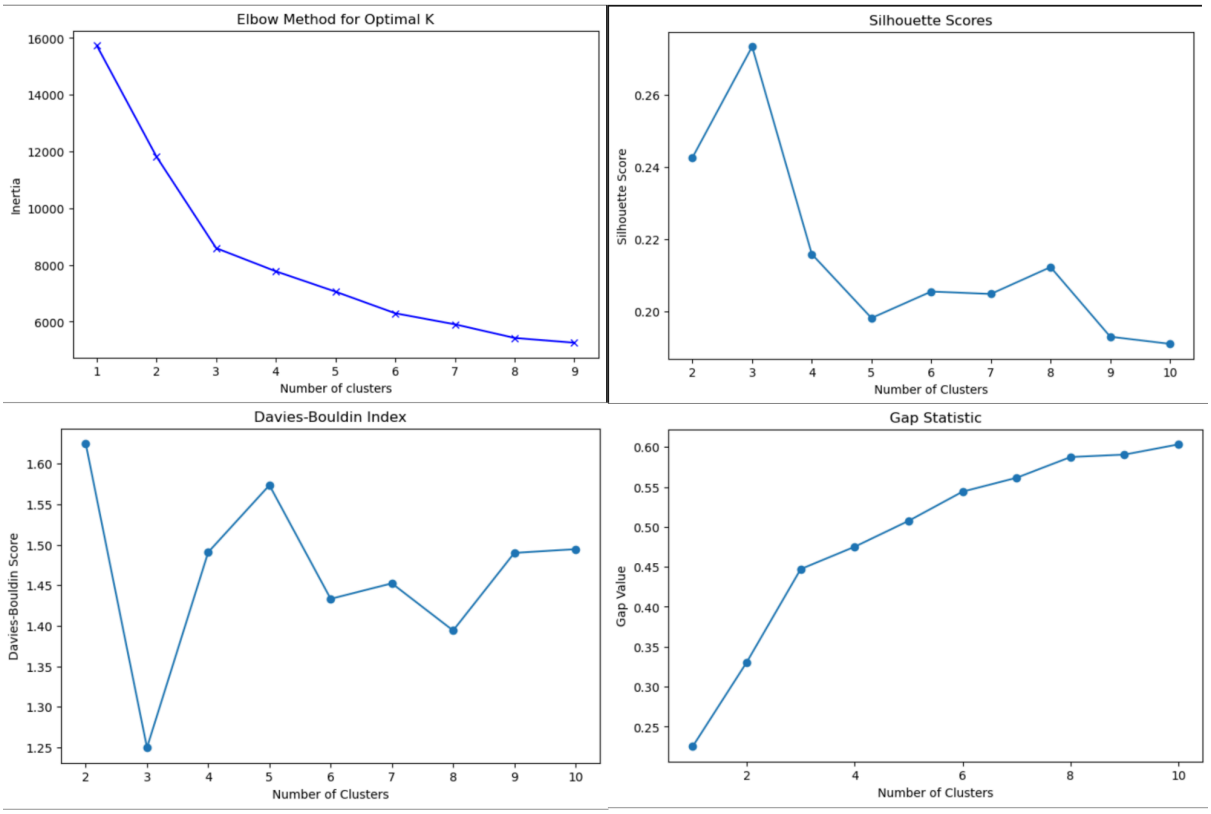

Figure 7. Optimal k

Feature	BIRCH	K-Means
Data Size	Large datasets	Small to medium datasets
Memory Usage	Low	High
Dynamic Data	Handles Streaming data	Not suitable for streaming
Speed	Faster with a larger dataset	Slower for larger dataset
Outliers	Handles outliers better	Sensitive to outliers
Preprocessing	Can be used as a preprocessing step	Final clustering algorithm

Table 3. Comparison BIRCH vs K-Means

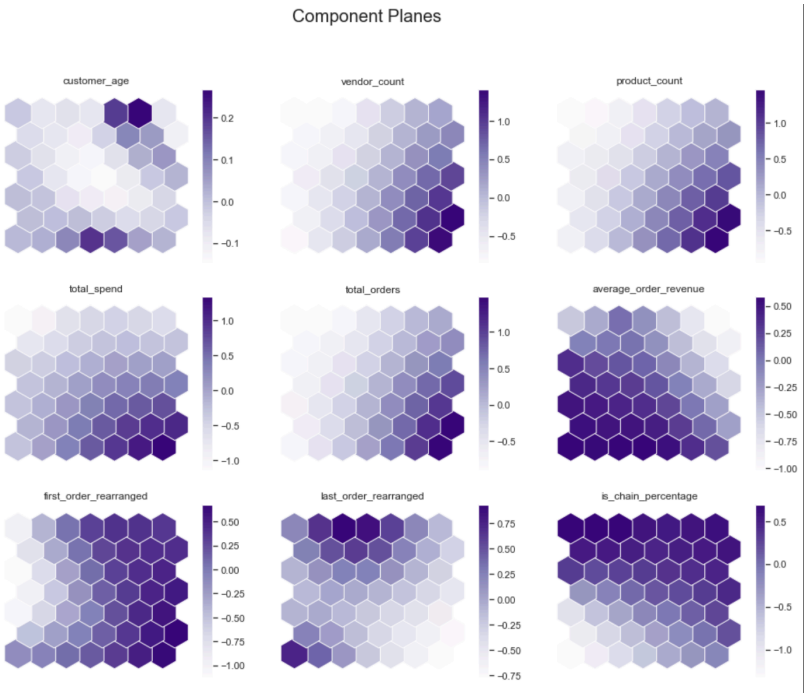

Figure 8. SOM Clustering

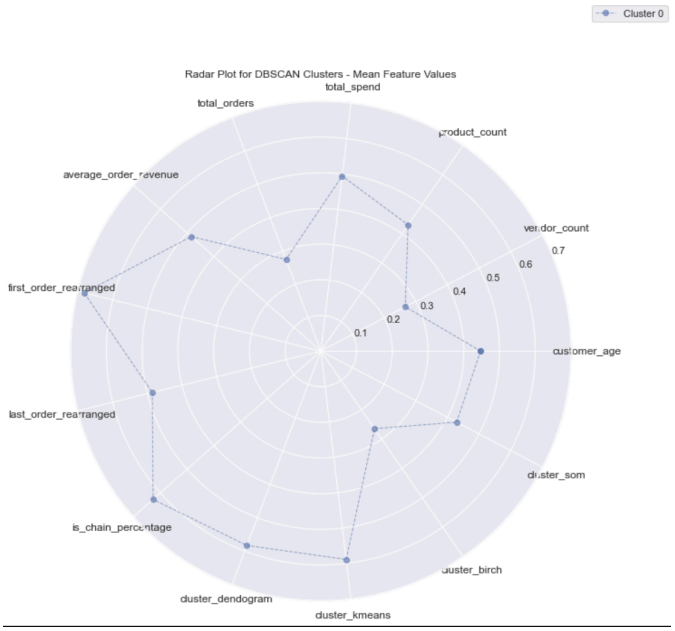

Figure 9. DBSCAN Clustering

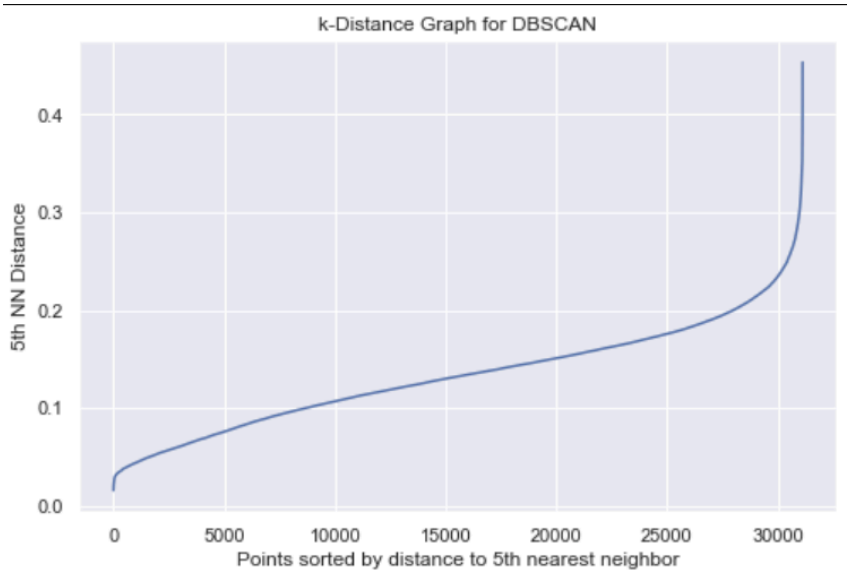

Figure 10. KNN Distance Graph

cluster_kmeans	customer_age	vendor_count	product_count	is_chain	first_order	last_order	total_spend	total_orders	average_order_revenue	first_order_rearranged	last_order_rearranged
0	27.68	7.81	19.61	7.87	10.91	81.06	150.19	13.75	14.51	79.09	8.94
1	27.55	1.81	2.66	1.30	60.76	72.74	20.93	2.12	11.32	29.24	17.26
2	27.35	1.78	2.57	1.39	18.44	30.27	18.67	2.05	10.74	71.56	59.73
3	27.51	3.94	6.86	3.79	14.76	75.92	39.42	5.51	8.36	75.24	14.08

Table 4. KNN Distance Graph

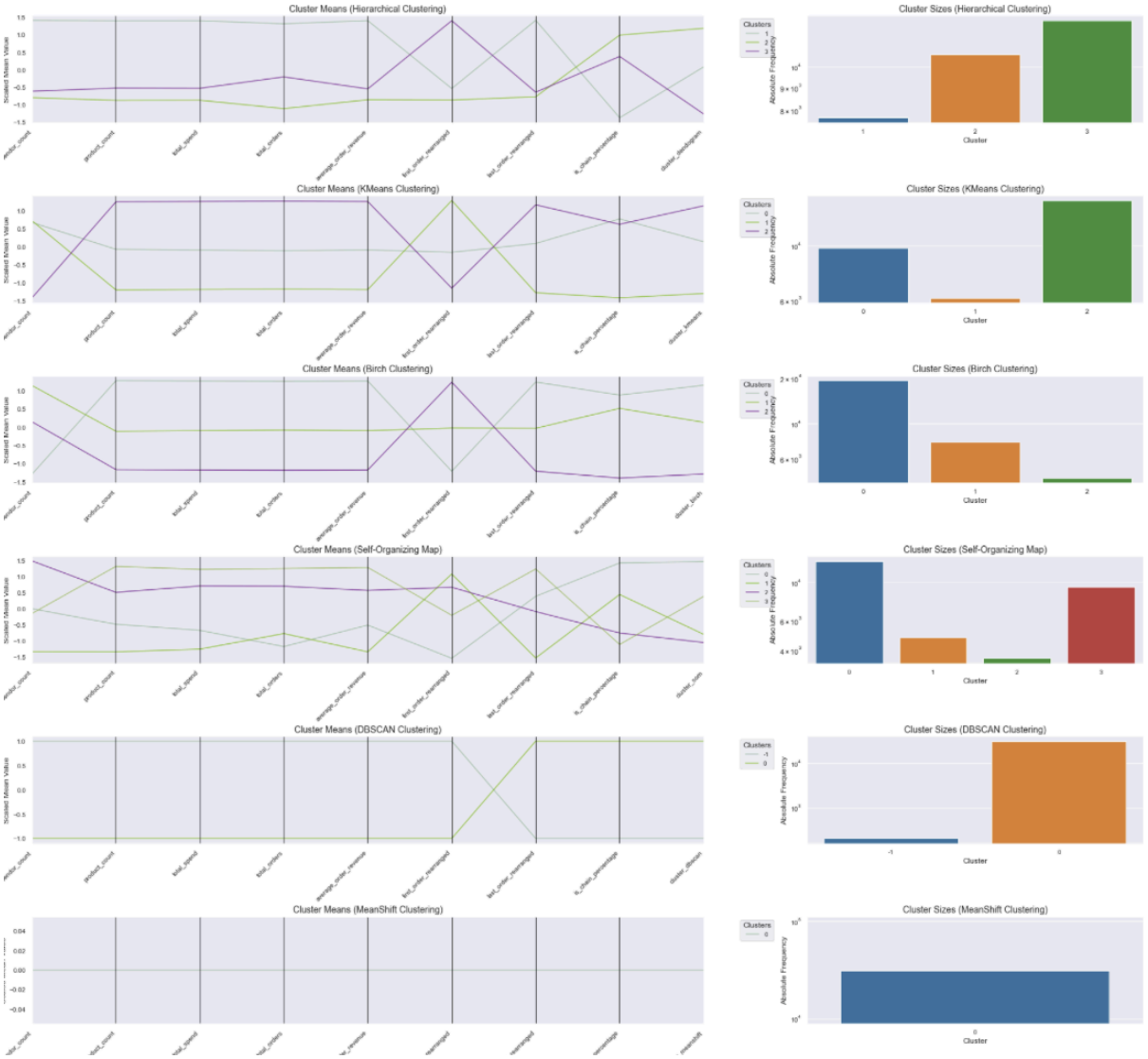

Figure 11. Comparison for Clusters

cluster_som	customer_count	customer_age	first_order_rearranged	last_order_rearranged	total_orders	product_count	total_spend	average_order_revenue	order_frequency	vendor_count	is_chain_percentage	Marketing
0	13375	27,37	62,23	35,15	3,52	4,39	27,89	7,91	17,66	2,60	0,80	No
1	4717	25,61	43,73	28,16	2,08	2,79	27,19	13,07	21,02	1,75	0,40	Yes
2	3576	29,39	57,65	19,84	5,69	7,45	40,87	7,19	10,14	3,91	0,35	Yes
3	9443	27,17	70,42	15,83	5,22	6,88	50,66	9,70	13,48	3,78	0,60	Yes

Table 5. Conclusion

Clusters	1	2	3
American Cuisine Revenue % of Total	11.69	8.82	14.71
Asian Cuisine Revenue % of Total	26.62	22.94	23.43
Beverages Cuisine Revenue % of Total	5.56	5.5	5.81
Cafe Cuisine Revenue % of Total	1.01	0.88	1.85
Chicken Dishes Cuisine Revenue % of Total	1.53	3.49	2.65
Chinese Cuisine Revenue % of Total	3.18	4.83	3.95
Desserts Cuisine Revenue % of Total	2.59	2.47	1.98
Healthy Cuisine Revenue % of Total	2.41	2.16	2.56
Indian Cuisine Revenue % of Total	5.06	6.61	4.6
Italian Cuisine Revenue % of Total	13.35	10.74	9.97
Japanese Cuisine Revenue % of Total	7.74	6.77	8.12
Noodle Dishes Cuisine Revenue % of Total	1.13	2.79	2.28
OTHER Cuisine Revenue % of Total	5.94	9.74	8.36
Street Food / Snacks Cuisine Revenue % of Total	9.83	10.14	7.09
Thai Cuisine Revenue % of Total	2.37	2.12	2.66

Table 6. Cuisine Preference

Clusters	Weekday	Weekend
1	3387 (71.8%)	1330 (28.2%)
2	2514 (70.3%)	1062 (29.7%)
3	6279 (66.49%)	3164 (33.51%)

Table 7. Day Preference

Clusters	Breakfast	Lunch	Afternoon	Dinner	Late Night
1	1077 (22.83%)	756 (16.03%)	1338 (28.37%)	451 (9.56%)	1095 (23.21%)
2	688 (19.24%)	768 (21.48%)	975 (27.27%)	519 (14.51%)	626 (17.51%)
3	2360 (24.99%)	1800 (19.06%)	2445 (25.89%)	864 (9.15%)	1974 (20.9%)

Table 8. Hour Preference

Cluster	Highest Vol. Region	2nd Highest Vol. Region	3rd Highest Vol. Region
1	**8670** 1939 (41.11%)	**4660** 1579 (33.47%)	**2360** 583 (12.36%)
2	**8670** 1127 (31.52%)	**2360** 1070 (29.92%)	**4660** 825 (23.07%)
3	**4660** 3407 (36.08%)	**8670** 3289 (34.83%)	**2360** 1883 (19.94%)

Table 9. Region

Clusters	CARD	CASH	DIGI
1	2662 (56.43%)	1012 (21.45%)	1043 (22.11%)
2	2226 (62.25%)	550 (15.38%)	800 (22.37%)
3	6973 (73.84%)	1091 (11.55%)	1379 (14.6%)

Table 10. Payment Method Preference

Clusters	-	DELIVERY	DISCOUNT	FREEBIE
1	2181 (46.24%)	1203 (25.5%)	709 (15.03%)	624 (13.23%)
2	1855 (51.87%)	767 (21.45%)	468 (13.09%)	486 (13.59%)
3	5865 (62.11%)	1105 (11.7%)	1088 (11.52%)	1385 (14.67%)

Table 11. Promotion Usage

Annexes

Chat GPT was used to refine and correct the grammar throughout the report. At the same time, it was used to adjust charts and display them together in Figure 11 as one output. From time to time, we used it for color adjustments as well.