DATA MINING PROJECT

Master in Data Science and Advanced Analytics

NOVA Information Management School

Universidade Nova de Lisboa

Exploratory Data Analysis and Clustering for Customer Segmentation in ABCDEats Inc.

Group 44

Bruna Simões, 20240491

Inês Major, 20240486

Rafael Bernardo, 20240510

Fall Semester 2024-2025

TABLE OF CONTENTS

1. INTRODUCTION

In today's data-driven business landscape, companies must skilfully interpret customer information to remain competitive. With markets saturated by options, generic marketing strategies are no longer sufficient. The primary challenge lies in the lack of personalized engagement strategies that truly resonate with diverse customer segments.

Previous research has shown that segmenting customers based on purchasing behaviour and demographic factors can significantly enhance targeted marketing efforts (Smith, 2020)

Building upon this foundation, our academic project employs data from **ABCDEats Inc.**, preprocessed as described in **"DM_Report_EDA_Group44"**, and applies data mining methodologies, including clustering and pattern recognition, to uncover meaningful customer segments.

2. INITIAL ADJUSTMENTS

To obtain a more nuanced understanding of customer spending behavior, we created the new features (*spent_top_cuisine, spent_per_order, products_per_order, spend_main_food* and *spend_beverages*_addons) described in Table 7.

3. PREPROCESSING

3.1. Missing Values

This section details the methods used to address missing values, ensuring data integrity and minimizing bias in the analysis. Upon analysis, we identified the following:

- **HR_0:** This feature had 3.7% missing values and was imputed by ensuring consistency between total daily orders and the sum of hourly orders. Specifically, when HR_0 was missing, the sum of the remaining hourly columns did not match the total daily orders. By subtracting the sum of the known hourly orders from the daily total, we determined the missing value for HR_0, thereby preserving the alignment between the daily and hourly order counts.
- **Night Orders:** This feature aggregates orders placed during night hours. For this reason, it needed to be recalculated, as it was contingent on accurate values for *HR_0*. Once *HR_0* was imputed, *night_orders* was updated by summing the (now complete) hourly orders for the relevant night hours.
- **Customer Age:** To address the 2.3% missing values in this column, we employed KNN and the results were rounded to integers to preserve consistency. Notably, *age_group* is derived from *customer_age* so it exhibited an identical pattern of missingness. Consequently, after *customer_age* was imputed, *age_group* was updated accordingly.
- **Customer Region:** This categorical feature had 1.4% missing values and imputation was based on the most frequent region within each customer's preferred cuisine group because of the relationship between *customer_region* and *top_cuisine* seen in the EDA.
- **First Order:** This feature exhibited 0.3% missing values (equivalent to 106 clients). Upon closer examination, these clients' total number of orders ranged from one to two, indicating that

they had indeed ordered from the business. Furthermore, their *last_order* value was 0, suggesting that their most recent purchase was made at the earliest recorded date in the dataset. Consequently, *first_order* was set to 0 for these observations.

- **Purchase Frequency:** The missing values of this feature were addressed following the treatment of missing values in *first_order*, as its computation depends directly on its values. Once the imputation process for *first_order* was completed, the updated values were utilized to calculate and impute missing values for *purchase_frequency.*

3.2. Outliers

A systematic approach was adopted to handle outliers, with careful consideration of each variable's underlying characteristics and the relationships among them.

First, outliers in ***customer_age*** were left untreated because they appear to represent the genuine range of customer ages rather than anomalies. Nevertheless, we acknowledge the potential need to define a threshold for customers aged 65 and above, whereby grouping these observations could yield more nuanced insights in future analyses.

For **features related to orders** - including *vendor_count*, *product_count*, *is_chain*, *total_orders*, and the features organized by day of the week, hour of the day, and meal period - an interconnected pattern of outliers was observed. These were primarily driven by extreme values in ***total_orders***, which we treated as the representative feature. Rather than relying on the interquartile range (IQR), which would have excluded 9.50% of the data, we opted to remove only those observations above the 99th percentile (number of orders greater than 25), resulting in a 0.93% reduction of the dataset. Although this approach successfully eliminated the most problematic outliers, it didn't eliminate all the sparsity. To address this issue, we applied Winsorization at the 99.5th percentile to *product_count* (capped at 30), *vendor_count* (capped at 14), and *is_chain* (capped at 18). As *purchase_frequency* and *products_per_order* depend on *product_count*, its values were recalculated. We chose not to apply winsorization to *total_orders* or the day/hour/meal-period features because they jointly capture the same total information and, for this reason, adjusting one subset would compromise the consistency among these interconnected features.

A similar strategy was employed for the **features related to spending**, namely *total_spent*, *spent_top_cuisine*, *spend_per_order* and the columns regarding the cuisine types. Since high values in *total_spent* appeared to drive outliers across these variables, it served as the primary focus for outlier treatment. Since these group of features seemed to have more outliers, we addressed them by removing observations above the 97.5th percentile, corresponding to spending exceeding 148 units. This approach resulted in the exclusion of 2.48% of the data set. This threshold was found more appropriate than the IQR threshold, which would have removed 7.86% of the dataset.

Boxplots showing the distribution of these features prior to outlier treatment and after are provided in the annex [Figure 1, Figure 2, Figure 3], illustrating the rationale behind the chosen thresholds.

3.3. Scaling

In our dataset features vary in magnitude and those with larger ranges (e.g. *total_orders*) can dominate the clustering process, overshadowing features that exhibit smaller value ranges (e.g., *product_count*).

This imbalance can bias clustering results, assigning greater weight to variables with larger values. To address this and ensure that all features contribute comparably, we standardized the data.

In determining the most suitable scaling method, we analysed the distribution of each feature. We opted for **Standard Scaling**, which subtracts the mean and divides by the standard deviation. This transformation retains the relative variance of each feature, which is particularly advantageous for distance-based clustering algorithms used in this study, that rely on numerical distances.

By contrast, Min Max Scaling, while it rescales features to a fixed range (commonly [0, 1]), does not explicitly account for the underlying distribution of the data. In the presence of substantial variability or outliers, it can compress significant differences among data points, undermining the ability of clustering algorithms to capture meaningful patterns.

3.4. Encoding

A separate encoding step was carried out for the categorical features to facilitate more detailed cluster profiling, even though these variables were not directly involved in our clustering process since we are using algorithms that rely on Euclidean distance. For this reason, we employed **One Hot Encoding**, which creates binary features (0 or 1) for each possible category in a categorical feature. By retaining the encoded features in the dataset, we can subsequently analyze and profile each cluster with respect to these categorical attributes, yielding more nuanced insights into the composition of each cluster.

3.5. Feature Engineering after Missing Values and Outliers Treatment

The dataset contains an extensive array of features related to customer ordering behaviour and spending patterns, such as HR_0 to HR_24 and DOW_0 to DOW_6. Given the high dimensionality of these variables, it was essential to perform feature reduction to reduce the input space and streamline the analysis. For this reason, we implemented **K-Means Clustering** to group the features into clusters, reducing complexity in the subsequent stages of the project.

3.5.1. Clustering Hourly Data

In an initial attempt to reduce dimensionality, features representing meal periods were created based on expert judgment, by aggregating hourly features into predefined categories. While this approach provided a preliminary structure, we sought to refine these groupings through clustering analysis.

We evaluated several configurations to identify the optimal number of clusters for grouping the hourly order variables. We concluded that four clusters provided the most meaningful segmentation of the data. The results of the clustering analysis are summarized in the table below:

Cluster	Feature	Hours
Cluster 1	*breakfast_orders*	HR_5 to HR_10
Cluster 2	*lunch_snack_orders*	HR_11 to HR_13 HR_15 to HR_19
Cluster 3	*night_orders*	HR_0 to HR_4
Cluster 4	*dinner_orders*	HR_14 HR_20 to HR_23

Table 1 – Aggregations of hourly features considering K-Means

The clustering results aligned well with the anticipated ordering behaviour patterns, confirming the method's effectiveness. However, HR_14, typically associated with lunch hours, was grouped with the dinner cluster. Despite this, we chose to retain HR_14 within *lunch_snack_orders*, as domain knowledge indicates that it more accurately belongs to this period.

By aggregating the hourly order data, we replaced the original 24-hourly features with four features, each representing a meal period and summarizing the proportion of orders occurring during that time.

As illustrated in Figure 4, the highest proportion of orders is observed during lunch and snack (64% of orders), followed by breakfast (20%). This suggests that lunches and snacks are the preferred meals of the customer. However, it is worth mentioning that a higher proportion of orders during lunch and snacks is expected, as this period spans a significantly longer time frame compared to other periods.

3.5.2. Clustering Weekday Data

Following a similar approach, we aimed to group the weekday features into broader categories to reduce the dimensionality of the dataset. By applying K-Means, we identified two distinct groups, *startweek_orders*, that consider orders from Sunday to Thursday, and *endweek_orders*, that capture orders made on Friday and Saturday.

This clustering goes accordingly with the general perspective of probability of ordering food throughout the week. As we can see in Figure 5, the start week period exhibits a higher proportion of orders (68%), which was expected due to the greater number of days included in this group (five days compared to two in the end week). However, this imbalance is acceptable as the objective is to distinguish customers based on their ordering patterns during the start or end of the week, rather than focusing on an even distribution.

As a result of this step, we reduced the original seven day of week features (DOW_0 to DOW_6) to two features representing the proportions of orders in each period, effectively simplifying the dataset.

4. CLUSTER ANALYSIS

Customers engage in varied ways, so analyzing their actions from different perspectives helps uncover more detailed segments. For this reason, this section aims to show the results of the clusters using different perspectives identified as the most relevant for understanding customer behavior.

It is important to note that, for each perspective, we used different clustering techniques to ensure robust results. Techniques like **Hierarchical Clustering** using Ward and Euclidean Distance, **K-Means** using Euclidean Distance, **Self-Organizing Maps** (SOM), and **Mean Shift** were applied to each perspective. This approach not only allows for a comparison of clustering outcomes but also ensures that the chosen clusters are representative and actionable. The performance of the different clustering techniques was evaluated using metrics like Silhouette Score, Adjusted R-Squared, Calinski-Harabasz Index **(2)** and Davies-Bouldin Index **(3)**.

4.1. Demographic and General Behaviour

The first perspective was designed to analyse the general behavioural patterns in the dataset, using *customer_age*, *purchase_frequency, total_spend*, *diversity_cuisine_types* and *different_weekdays*.

As mentioned before, it's crucial to use different clustering techniques to evaluate the dataset from multiple angles. The decision process to decide the optimal number of clusters for each clustering technique is detailed in Table 8 and Table 9 and the final results are presented in Figure 6.

A segmentation based on age and engagement emerges from the analysis. One cluster group of older individuals characterized by low engagement. The others consist of younger individuals. For these, we were able to differentiate people with higher levels of engagement and spending, frequent but low-value customers and people with moderate levels of spending and ordering behavior. Further details on these clusters are provided in Figure 6 and Table 10**Error! Reference source not found.**.

Recognizing the robustness of this clustering pipeline, we applied it to other clustering combinations. Additionally, considering the results of this initial clustering approach, we decided from now on to use *customer_age* solely for profiling rather than as a primary variable in the clustering algorithms, since our goal is to highlight spending and ordering behavioral differences.

Additionally, during the analysis of the hexagons heatmap after training SOM [Figure 7], a red dot in *purchase_frequency* was observed, representing customers who engaged significantly on one occasion but did not return. In our understanding, it is crucial to retain these customers in the dataset to inform marketing strategies aimed at increasing engagement, however their presence caused the SOM to focus excessively on this group, obscuring other patterns. We decided to keep these customers acknowledging this issue but allowing our clustering techniques to identify these clients.

4.2. Spending Profile

In this perspective, we aimed to evaluate which features capture customers' financial contributions and their culinary diversity. To achieve this, we tested several combinations of features, by applying K-means and measuring the resulting silhouette scores. The choice of K-Means was informed by previous analyses, which showed superior performance of this algorithm. We initially settled on four clusters as a reasonable balance between detailed segmentation and interpretability.

Through this process, we identified the most promising features, ultimately selecting *total_spend*, *spent_top_cuisine*, *diversity_cuisine_types*, *spent_main_food*, and *spent_beverages_addons*.

The decision process to decide the optimal number of clusters for each clustering technique is detailed in Table 11. The table below shows the results obtained using the different clustering techniques:

Clustering Techniques	Number of Clusters	Silhoutte Score	Adjusted R squared	Calinski-Harabasz	Davies-Bouldin
Hierarchical Clustering	K = 5	0.3497	0.6432	13875.15	1.1778
K-Means	K = 7	0.4300	0.7652	16722.78	0.9859
SOM Hierarchical	K = 7	0.3730	0.7203	13212.93	1.2374
SOM K-Means	K = 5	0.3710	0.6619	15126.75	1.0822
Mean Shift	K = 7	-	0.6917	-	-

Table 2-Results of Clustering Process for Perspective 2

Among the methods tested, **Mean Shift Clustering** emerged as the least effective approach. Despite the flexibility of its quantile parameter and its relatively high Adj. R-Squared, this method produced highly imbalanced clusters, as demonstrated in [Figure 8], failing to provide meaningful and interpretable clusters for practical analysis. On the other hand, **K-Means** clustering stands out as the most robust method in this analysis. It achieves the highest Silhouette Score, indicating well-separated clusters. Furthermore, it boasts the highest Calinski-Harabasz Index and Adj. R-Squared (76.5%), and the lowest Davies-Bouldin Index (<1). For this reason, it's evident the superior quality and practical interpretability of K-Means clustering, making it the preferred method for this perspective. Although **Hierarchical Clustering** and **SOM techniques** exhibit strong overall performance - including an especially high Adj. R-squared value of 72% for SOM Hierarchical - they do not surpass the results obtained with K-Means.

Taken together, we consider that both K-Means and SOM K-Means offer the most promising clustering solutions, since we achieved similar metrics with less clusters. K-Means [Figure 9 and Table 12] and SOM K-Means [Figure 10 and Table 13] both highlight a distinction between individuals who prioritize spending on diverse cuisines and those who focus more on beverages. However, K-Means, with its additional clusters, further differentiates moderate spenders by grouping them into four clusters instead of two, which varies in the amount spent on main food, with Cluster 1 allocating nearly 100% of their spending. Additionally, K-Means distinguishes a group of lower spenders. In contrast, SOM K-Means lacks this level of granularity, failing to make clear distinctions among these customer groups.

4.3. Temporal Ordering Profile

This perspective was designed to analyze temporal patterns in customer interactions with the platform, focusing on their ordering behavior across different meal periods and days of the week. For this reason, we used *breakfast_orders*, *lunch_snack_orders*, *dinner_orders*, *night_orders*, *startweek_orders*.

The decision process to decide the optimal number of clusters for each clustering technique is detailed in Table 14. The table below shows the results obtained using the different clustering techniques:

Clustering Techniques	Number of Clusters	Silhoutte Score	Adjusted R squared	Calinski-Harabasz	Davies-Bouldin
Hierarchical Clustering	K = 5	0.3829	0.6583	14829.26	0.9994
K-Means	K = 5	0.4354	0.7206	19852.94	0.8410
SOM Hierarchical	K = 5	0.3725	0.6713	15715.70	0.9450
SOM K-Means	K = 6	0.4540	0.7191	18453.75	0.8402
Mean Shift	K = 9	-	0.7434	-	-

Table 3- Results of Clustering Process for Perspective 3

Similarly to what happened in the previous perspective, the **Mean Shift** showed a good Adj. R-Squared but the nine clusters formed exhibit substantial imbalance. In contrast, **K-Means** and **SOM K-Means** demonstrate consistently strong performance across most evaluation criteria. Notably, SOM achieves the highest Silhouette Score, but this comes with a slightly lower Adj. R-squared compared to K-Means. Consequently, while SOM may provide more refined cluster separation, it offers a marginally less comprehensive explanation of variance.

Taken together, we consider that K-Means and SOM K-Means offer the most promising results. K-Means [Figure 11] demonstrates more evenly distributed cluster sizes, while SOM [Figure 12] displays more imbalance, suggesting that may overfit specific behavioral segments and underrepresent others.

In terms of behavioral trends, illustrated in Table 15 and Table 16, both methods appear to excel in differentiating customers based on their preferences, identifying clusters where customers display high behaviors during lunch/snack period and at dinner. Additionally, both approaches capture a cluster that exhibits a tendency to place orders predominantly toward the end of the week. However, SOM creates more granular clusters, splitting behaviors such as *night_orders* into multiple clusters. Notably, in K-Means, the segment with elevated *lunch_snack_orders* also shows high ordering activity toward the end of the week, while in SOM, these patterns are subdivided into multiple clusters that differ in whether their high-order activity occurs early or late in the week. This distinction highlights K-Means as a more suitable option due to its balanced clusters, while SOM, despite offering a more descriptive approach, tends to produce less evenly distributed segments.

4.4. Ordering Behaviour Dynamics

This perspective was developed to explore the dynamics of customer ordering behavior, focusing on identifying patterns related to order volume, product diversity, and engagement across different days of the week. Using the same reasoning for feature importance described in section 4.2, we selected *vendor_count*, *products_per_order*, *total_orders*, *different_weekdays*, and *purchase_frequency*.

The decision process to decide the optimal number of clusters for each clustering technique is detailed in Table 17. The table below shows the results obtained using the different clustering techniques:

Clustering Techniques	Number of Clusters	Silhoutte Score	Adjusted R squared	Calinski-Harabasz	Davies-Bouldin
Hierarchical Clustering	K = 6	0.414	0.762	19720.56	0.9243
K-Means	K = 6	0.4434	0.781	21985.242	0.8470
SOM Hierarchical	K = 5	0.4051	0.681	16489.51	0.8336
SOM K-Means	K = 5	0.3368	0.6650	15294.49	0.9547
Mean Shift	K = 19	-	0.5768	-	-

Table 4-Results of Clustering Process for Perspective 4

As seen previously, **Mean Shift** formed exhibit substantial imbalance, creating 19 clusters. **K-Means** achieves the highest silhouette score and Adj. R-squared values (78%), indicating that this technique provides a relatively well-defined cluster structure with strong explanatory power regarding the observed variance. **Hierarchical Clustering** displays a slightly lower silhouette score but still maintains a robust Adj. R-squared of 76.2%. In contrast, **SOM** approaches return lower scores on most metrics, suggesting they offer less robust segment separation under these parameter settings.

Overall, the results show K-Means as the most effective clustering method, largely owing to its capacity to segment customers according to the volume of orders and the diversity of ordering days [Table 18 and Figure 13]. Specifically, Cluster 5 includes customers who place frequent orders across multiple days, reflecting high engagement levels. By contrast, Cluster 4 comprises customers who order only once but purchase the largest number of products per order. Meanwhile, Clusters 1 and 0 both exhibit low ordering habits, with Cluster 0 showing especially infrequent purchases, thereby underscoring potential opportunities for outreach to stimulate higher engagement.

4.5. Multiple Perspectives

For a more comprehensive representation of customer behavior, we combined the previously examined Perspectives 2, 3, and 4 into a single perspective, enabling a multifaceted view of the data. Building on prior findings indicating that K-Means and Hierarchical Clustering have a good balance between performance and simplicity, we employed both methods into the merged perspective.

In the case of **Hierarchical Clustering**, both the R-squared plot [Figure 14] and the dendrogram [Figure 15] indicate that five clusters may be optimal, as evidenced by a clear elbow in the R-squared curve and a well-defined branching structure. However, upon testing six clusters, we observed better results on nearly every metric. Despite these gains, K-Means still produced stronger overall outcomes [Table 19], including more balanced segments. Focusing on **K-Means** and referencing the Inertia plot [Figure 16], we noted a substantial drop in inertia up to k=5, after which the decrease became more gradual - suggesting an elbow at five clusters. Nonetheless, when we explored **seven clusters**, the resulting segmentation appeared both more balanced and more representative of the underlying customer population, as we considered them our final results, represented in Figure 17 and Figure 18 where we can see a good separation of clusters, showed in the T-SNE in Figure 19.

The details of the resulting clusters are summarized below:

Cluster	**Global Characteristics**
0	Composed of individuals who exhibit high purchase frequency within a short timeframe - often placing a single, high-volume order—but demonstrate low overall engagement (as noted in previous analyses). On average, these customers allocate 30% of their spending to beverages and add-ons, aligning with a preference for lunch/snack periods.
1	Customers that have a high engagement, reflected in both frequent ordering and elevated spending. These customers demonstrate notable diversity, placing approximately six orders, spending 50 units from five different vendors.
2	Customers that order during night (75% of orders), placing relatively few orders (about three), yet with high spending behavior considering the orders placed (38 units). This suggests a higher spending per order.
3	Customers that order at dinner time (63% of orders) that demonstrates an average of number of orders similar to Cluster 2 but at a significantly lower average spending level (20 units).
4	"Typical" customers, who show moderate ordering and spending behaviors. Their preferred meal period is lunch/snack, closely matching the overall distribution observed in the dataset.
5	These customers shares similarities with Cluster 1 but with a more extreme behaviour, being the ones who order and spend the most. On average, they place 14 orders and spend about 80 units from eight different vendors and five distinct cuisines, without any strong day-of-week preference.
6	Customers that demonstrate a marked inclination toward breakfast orders (76% of orders), allocating around 28% of their spending to beverages/add-ons. This behavior may reflect differing time zones or cultural norms, as these customers rarely order during lunch/snack periods.

Table 5-Description of Final Clusters

The results reveal a clear segmentation of customers based on their ordering and spending behaviors, as well as their diversity in cuisine choices, preferred days of the week, and favored meal periods. For more details about the obtained clusters, refer to Table 20.

5. PROFILING

Having established the clusters based on customers' ordering and spending patterns, as well as temporal factors, we next carried out a profiling analysis to gain deeper insight into each segment. Specifically, we examined how members of each cluster engage with promotions (*last_promo*), their preferred payment methods (*payment_methods*), dominant cuisine choices (*top_cuisine*), and demographic features (*customer_age* and *customer_region*).

For these features, we noted the following insights:

- **Last promotion** (*last_promo*): As shown in Figure 20, Cluster 0, characterized by having a high purchase frequency, since includes customers that placed only a single but high-volume order (as seen in previous analysis), exhibits a 40% of customers who benefited from free delivery in their last order. This suggests that free delivery may have been a decisive factor for these customers. Therefore, **increasing delivery promotions** for them would be a good marketing strategy to increase engagement. By contrast, the clusters associated with higher overall engagement (Clusters 1 and 5) displayed no promotions in the last order for nearly 60% of individuals. This pattern highlights their loyalty despite promotional incentives.

- **Payment Method** (*payment_method*): By examining Figure 21, it becomes evident that the most highly engaged customers predominantly use card payments, likely reflecting the convenience of having payment details saved on the platform. In contrast, customers who place few orders but of higher volume, relies considerably more on cash (26%) and digi (33%) than other segments.

- **Region** (*customer_region*): The distribution of regions across each cluster, as presented in Figure 22, reveals that more than half of the customers in Cluster 3 (frequent dinner orders) and Cluster 5 originate from region 2360. Meanwhile, 43% of Cluster 4 - representing more "typical" customers with moderate behavior - are from region 4660. In contrast, Cluster 2 (night order customers) is heavily concentrated in region 8670 (85%), possibly reflecting local dining customs that extend later into the evening. Consequently, **targeting notifications to these customers at later hours may be beneficial**. Additionally, 60% of the customers in Cluster 6 (primarily breakfast orders) also reside in region 8670, further emphasizing the different meal period in this region.

- **Top Cuisine** (*top_cuisine*): As illustrated in Figure 23, clusters with a high concentration of customers from region 8670 exhibit a preference for Asian food, reflecting the local inclination of that region. In contrast, the "typical" customer group displays a comparatively stronger affinity for Italian food (16% of customers). Notably, individuals who order less frequently but purchase a high number of items per order prefer Street Foods/Snacks, suggesting that their large-volume orders may consist primarily of smaller snack items rather than full meals.

6. MARKETING APPROACH

In the concluding phase of this analysis, we leveraged the insights gained from the clustering exercise to propose targeted marketing strategies. Each cluster exhibits unique characteristics, indicating that a one-size-fits-all approach would be suboptimal. Consequently, the strategies outlined below are tailored to address the specific needs and behaviors of each segment, with the overarching goal of enhancing both consumption levels and overall customer spending.

Cluster	**Marketing Approaches**
0	To re-engage these customers, increase free-delivery promotions around lunchtime, bundle lunch with the add/ons. Since many customers order only once, provide post-first-order discounts offering discounts for next purchases. Encourage them to buy other types of cuisines with in-app ads to broaden cuisine choices.
1	Since they are high-frequency customers, introduce a reward system or tiered membership. To leverage their willingness to spend and since they have a high diversity of cuisine, recommend complementary items at check-out or showcase new cuisines.
2	Offer specials or discounts and schedule push notifications/emails during night to match their ordering habits. Highlight Asian cuisines and businesses from region 8670. Provide free add-ons or discounts above a certain spending threshold to encourage larger orders, as these customers show willingness to spend more per order.
3	Offer dinner-special combos or small upgrades since they show preference of this meal period. Send targeted offers or reminders before dinner time to stimulate additional orders and encourage them to buy more and expensive products to consequently promote spending.
4	Provide appealing lunch/snack promotions, sending an email at the morning to lunch orders. Since almost half of the customers are from 4660, highlight local vendors. Offer loyalty incentives to turn moderate customers into frequent. Encourage to explore new cuisines, offering discounts if they buy products for other cuisines.
5	Upgrade loyal, high-spending customers to VIP, to distinguish them from Cluster 1. Additionally, we can encourage referral with bonuses, use personalized communications (emails, small gifts or free-delivery), informing them of their loyalty and encourage to continuing ordering.
6	After buying breakfast give discounts for lunch/dinner to leverage their spending with main food orders.

Table 6-Marketing Approaches for Final Clusters

7. CONCLUSION

This study set out to explore the behavioural patterns of ABCDEats Inc. customers by applying different clustering methods and feature selection strategies. Throughout the process, multiple algorithms were tested to assess their ability to segment the customer base. Although several approaches were examined, K-Means and SOM K-Means emerged as the most promising, consistently delivering higher scores and generating meaningful clusters - albeit with SOM K-Means introducing greater complexity.

Given the substantial number of features (e.g., cuisines, days of the week, meal times), reducing the input space became a critical step. In some instances, aggregating them using K-Means facilitated a more streamlined representation and improved the performance of subsequent clustering steps. Equally important was the feature-importance analysis, which guided the selection of the most informative features for clusters, ultimately improving the clarity and coherence of the final clusters.

To gain a comprehensive view of customer behaviour, the different perspectives - spending, ordering frequency, and temporal factors—were ultimately merged, offering a richer understanding of how these dimensions overlap. Upon obtaining distinct clusters, various categorical variables (such as region and cuisine preferences) were then used to profile each segment and uncover additional insights into their preferences. This holistic approach proved instrumental in crafting targeted marketing strategies tailored to the nuanced characteristics of each group.

Looking ahead, the current outlier treatment may benefit from a more flexible approach, allowing for a broader representation of atypical cases without compromising the reliability of cluster formation. Such an adjustment could further refine the clustering results and reveal new opportunities for personalized marketing efforts. Overall, this research provides a solid foundation for understanding customer engagement at ABCDEats Inc., illuminating paths for both immediate strategic initiatives and future analytical enhancements.

8. BIBLIOGRAPHY

Smith, A. (2020). *Consumer Behaviour and Analytics.* Newgen Publishing UK.

APPENDIX A – FIGURES

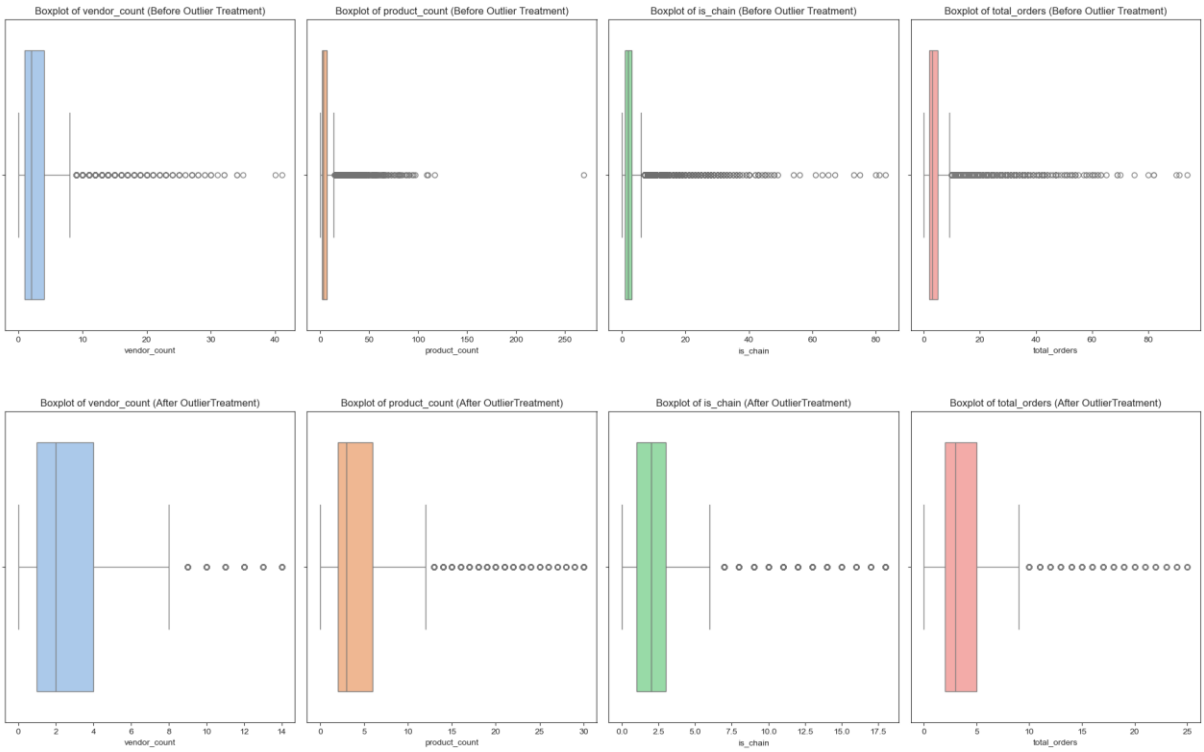

Figure 1 – Boxplots before and after outlier treatment for features related to number of orders

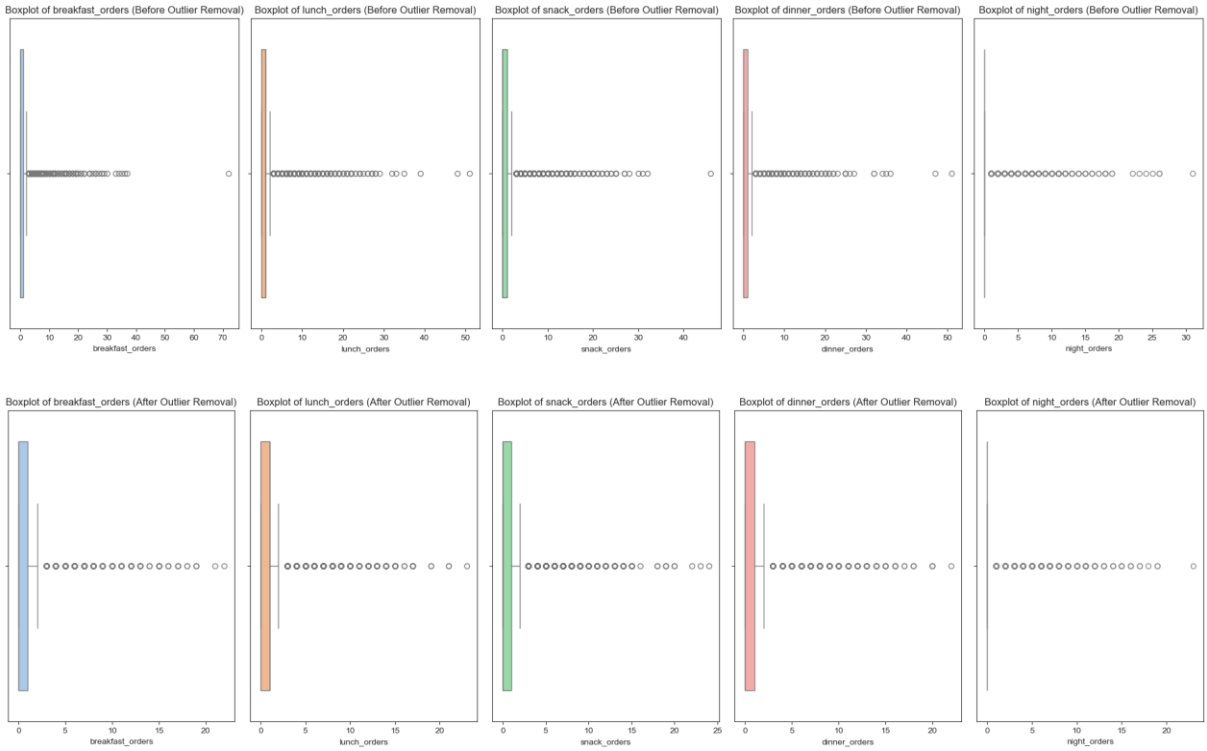

Figure 2 – Boxplots before and after outlier treatment for features related to meal period

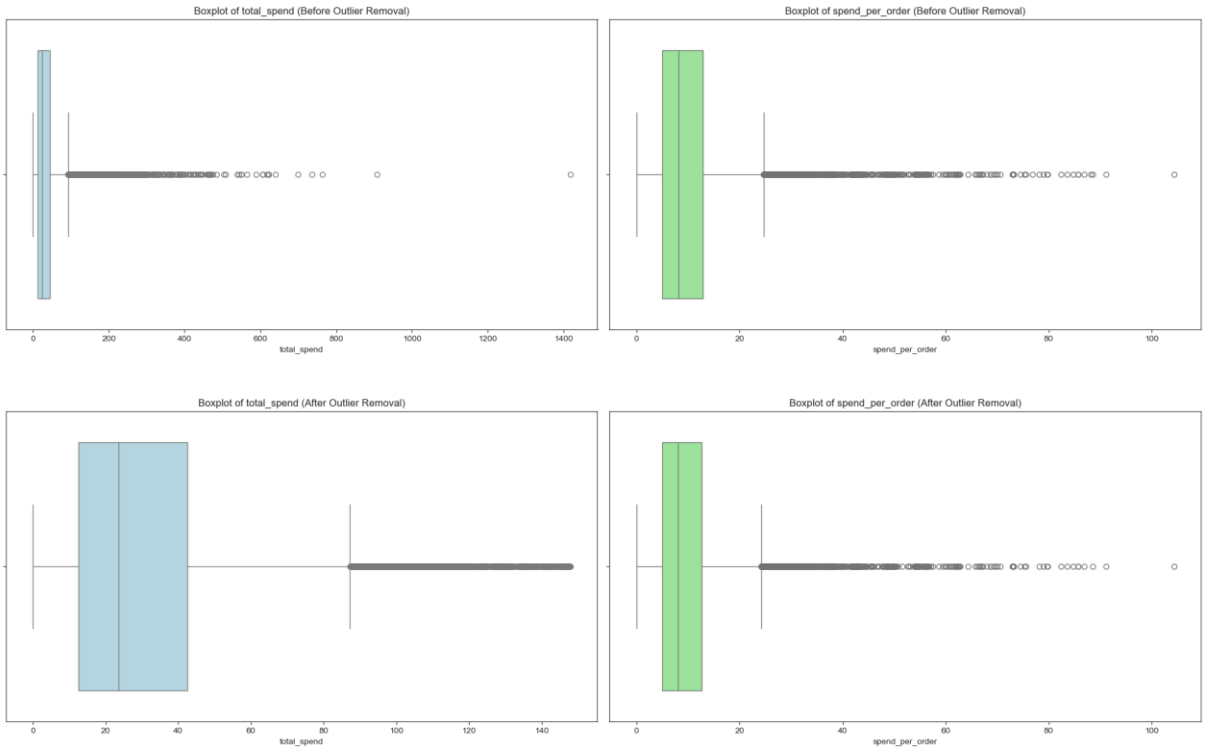

Figure 3 – Boxplots before and after outlier treatment for features related to amount spent

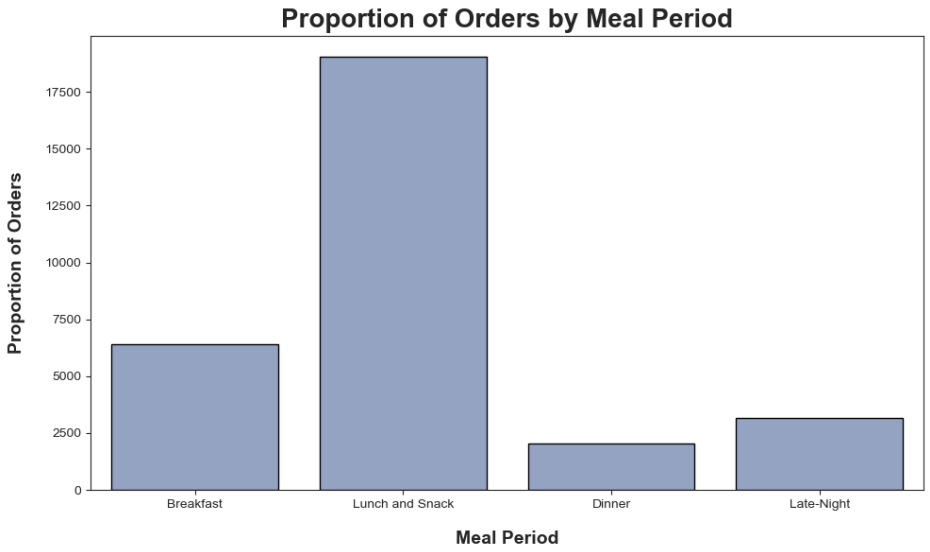

Figure 4- Orders by Meal Period

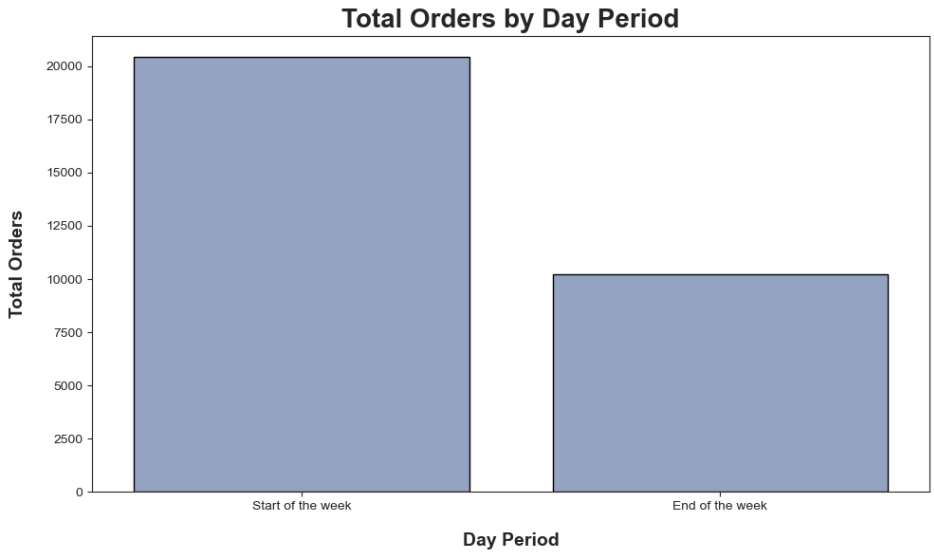

Figure 5- Orders by Day Period

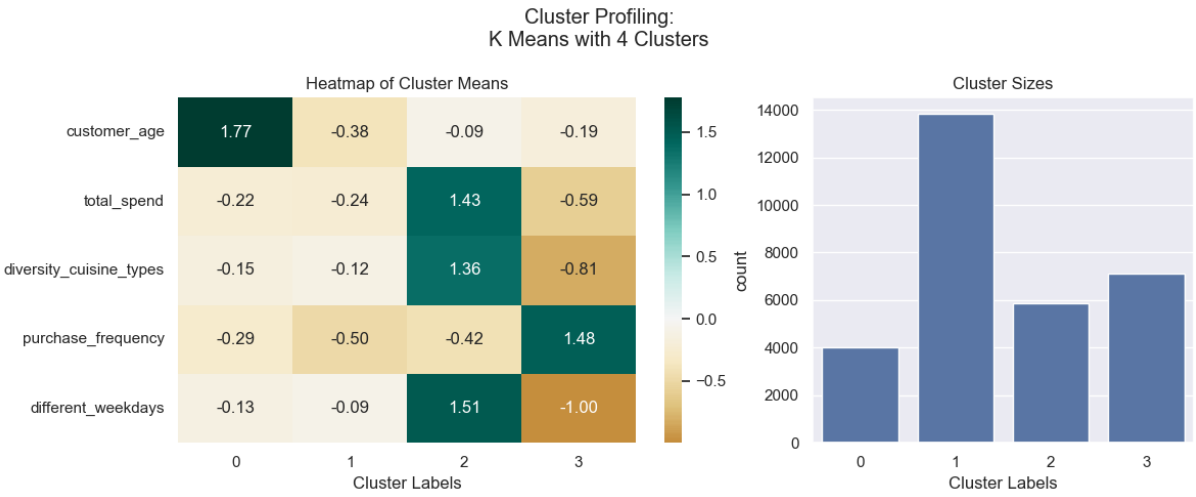

Figure 6- Final Clustering Results for Perspective 1 using K-Means

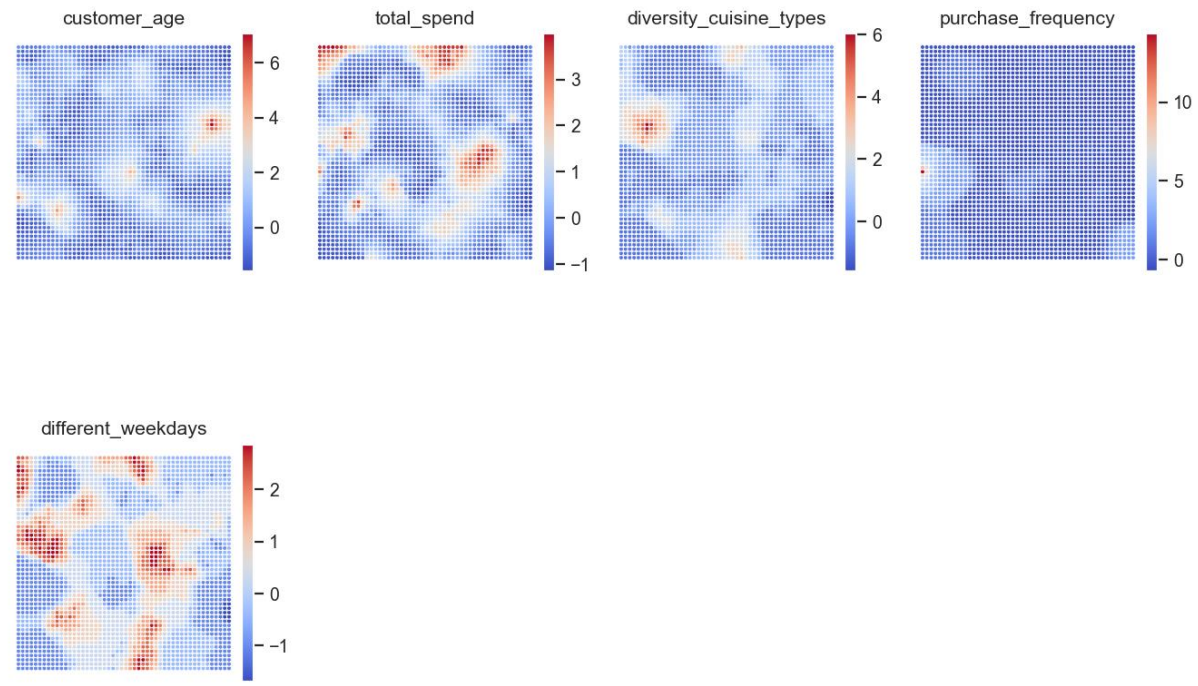

Figure 7 - Hexagons Heatmap after training SOM for Perspective 1

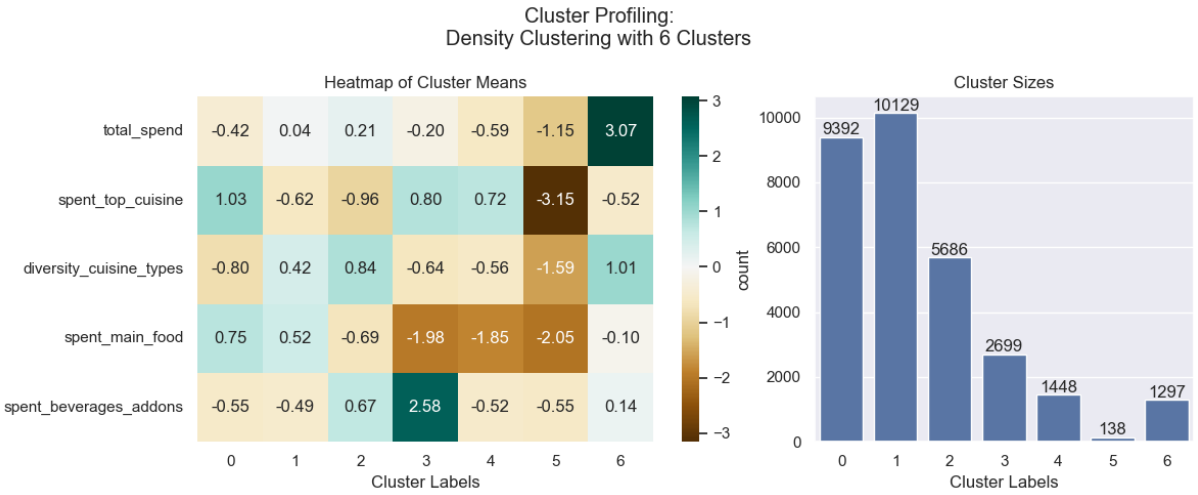

Figure 8 - Clustering Results for Perspective 2 using Mean Shift

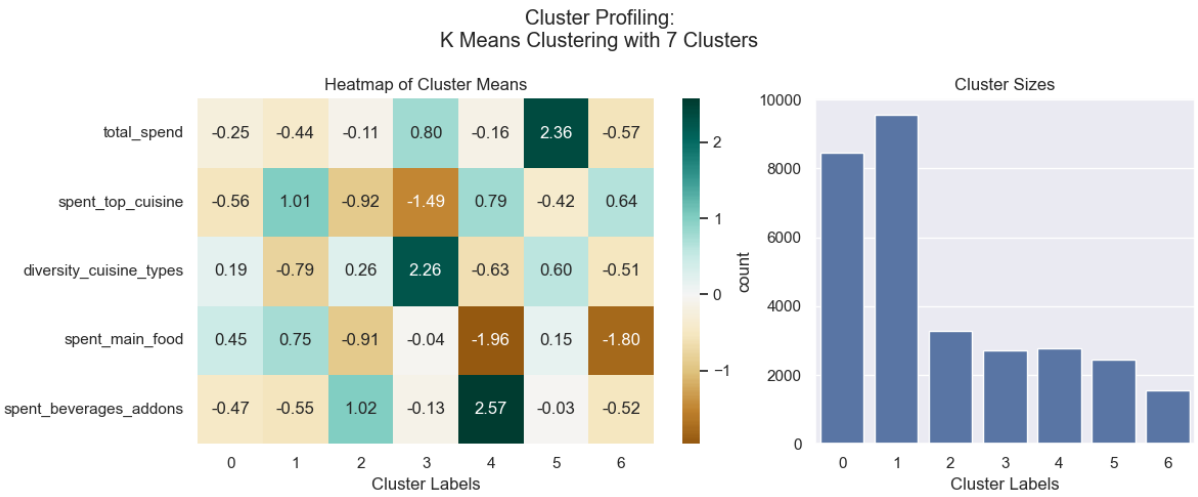

Figure 9 - Clustering Results for Perspective 2 using K-Means

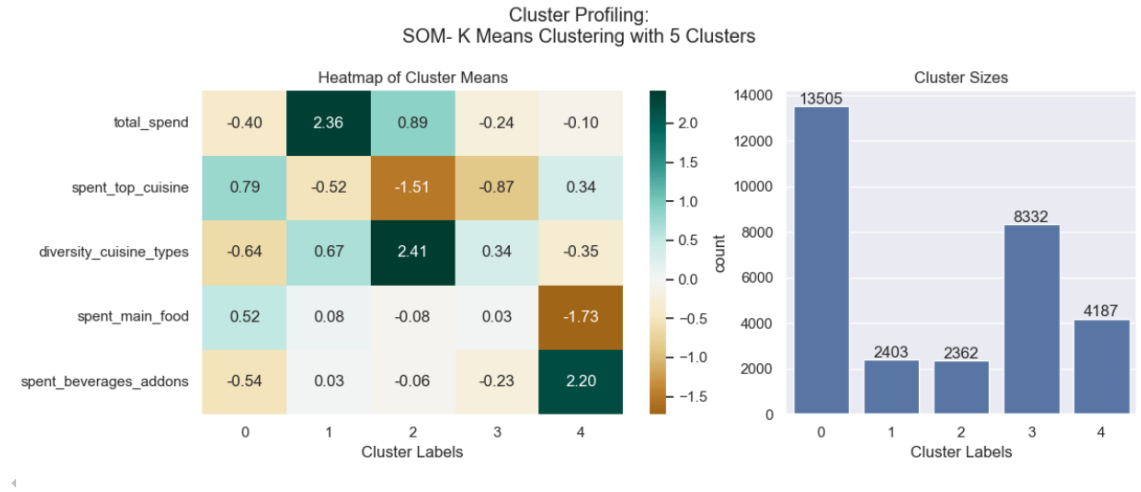

Figure 10 - Clustering Results for Perspective 2 using SOM K-Means

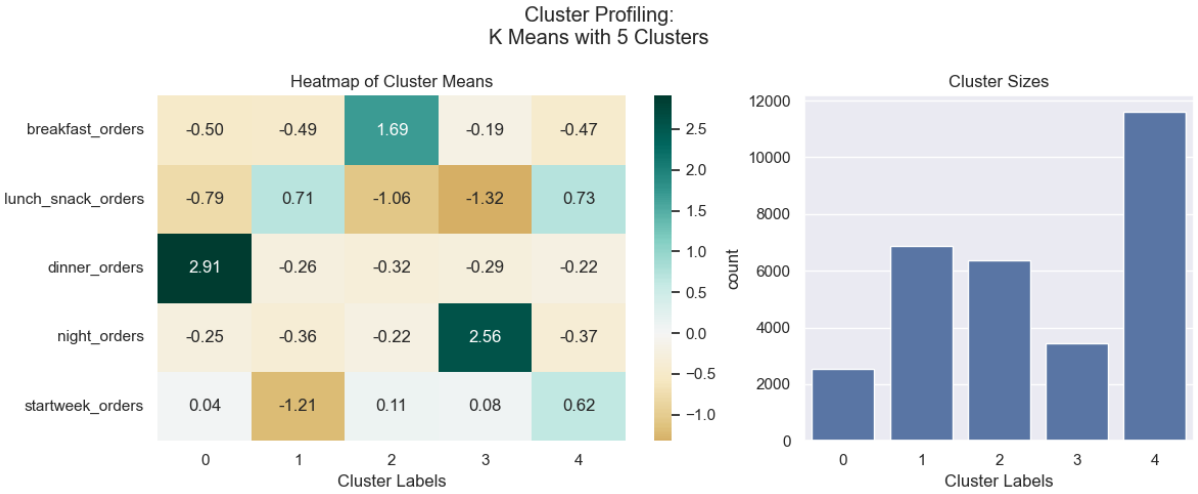

Figure 11 - Final Clustering Results for Perspective 3 using K-Means

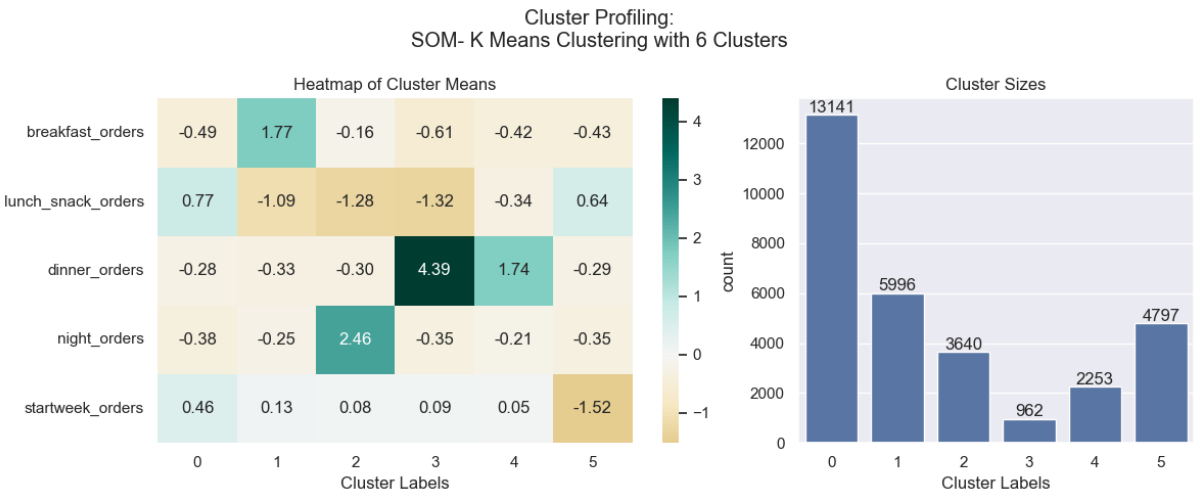

Figure 12 - Final Clustering Results for Perspective 3 using SOM K-Means

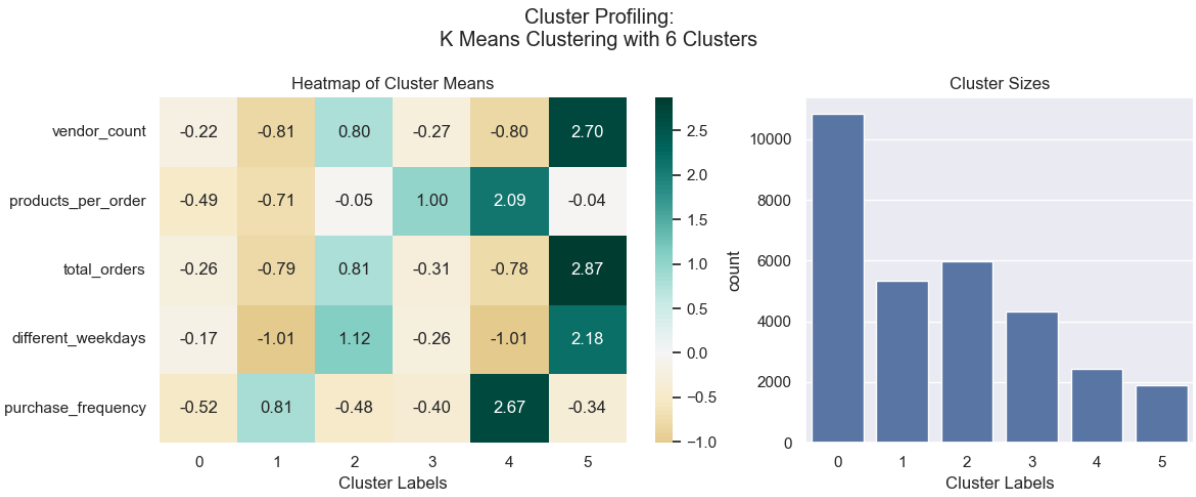

Figure 13 - Final Clustering Results for Perspective 4 using K-Means

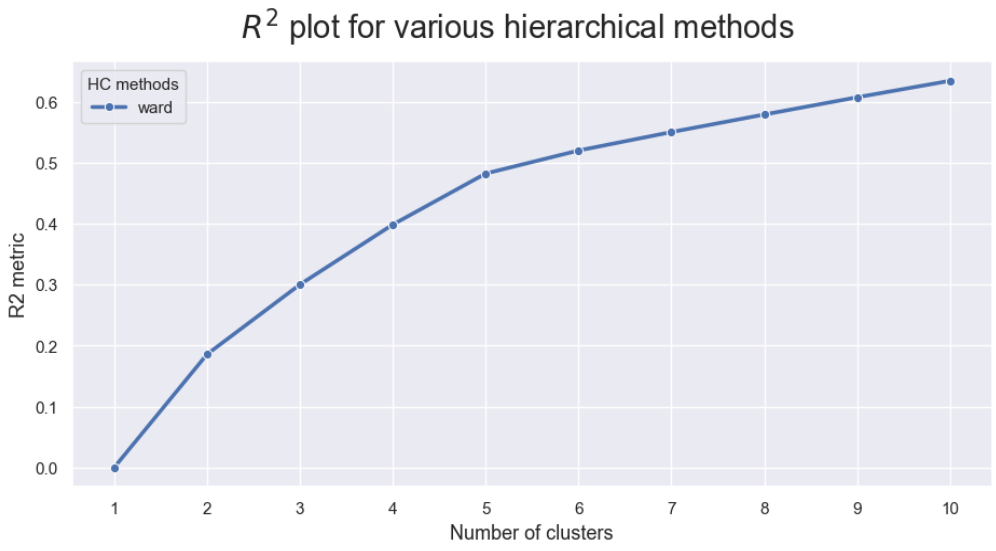

Figure 14 – R-Squared Plot for Hierarchical Clustering using merged perspectives

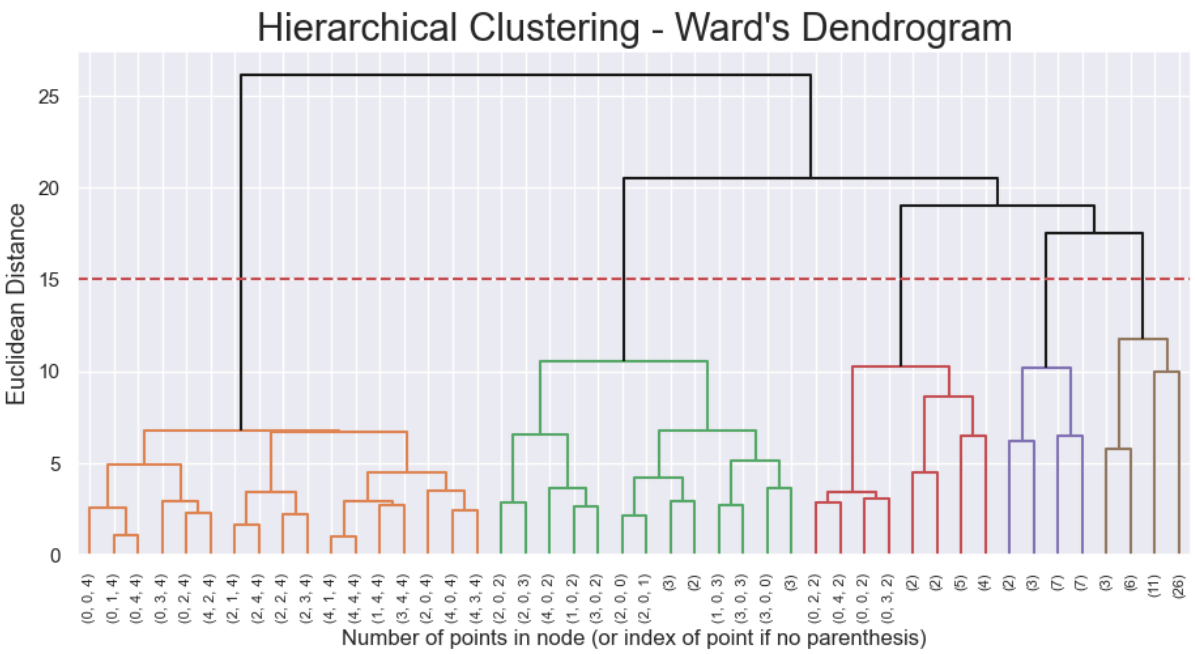

Figure 15 – Ward's Dendogram using merged perspectives

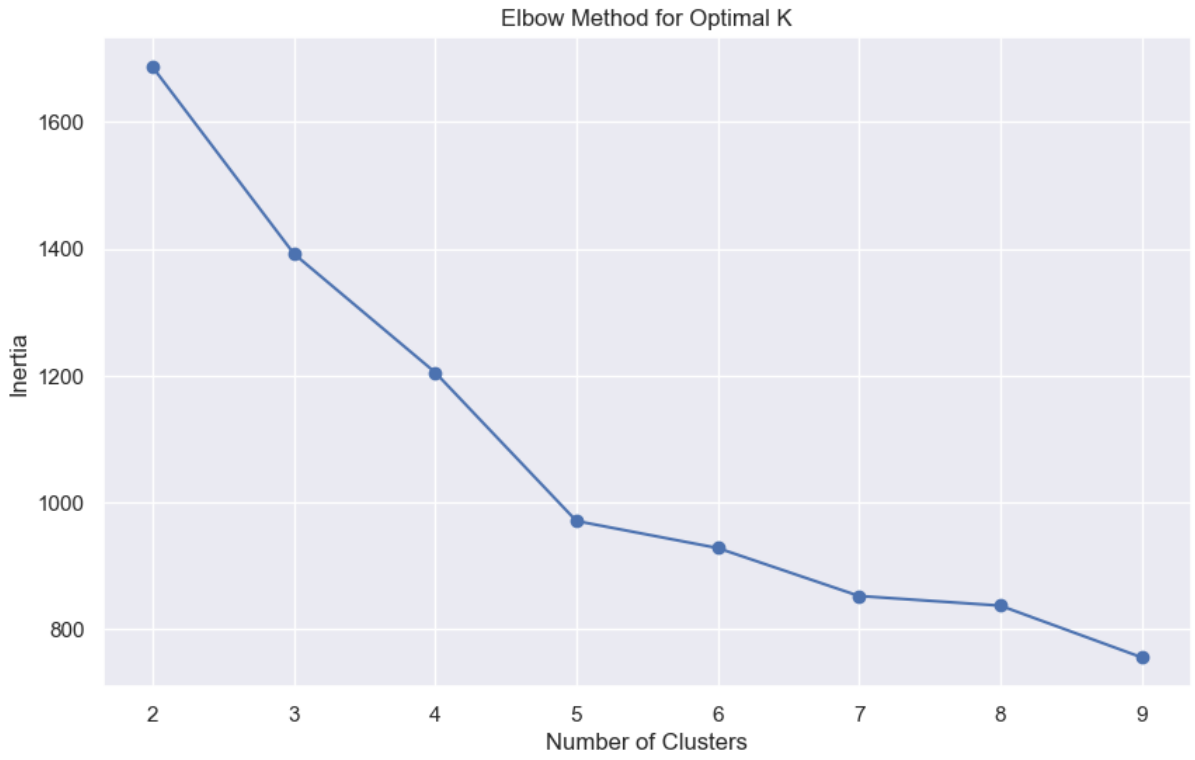

Figure 16 – Inertia Plot for K-Means using merged perspectives

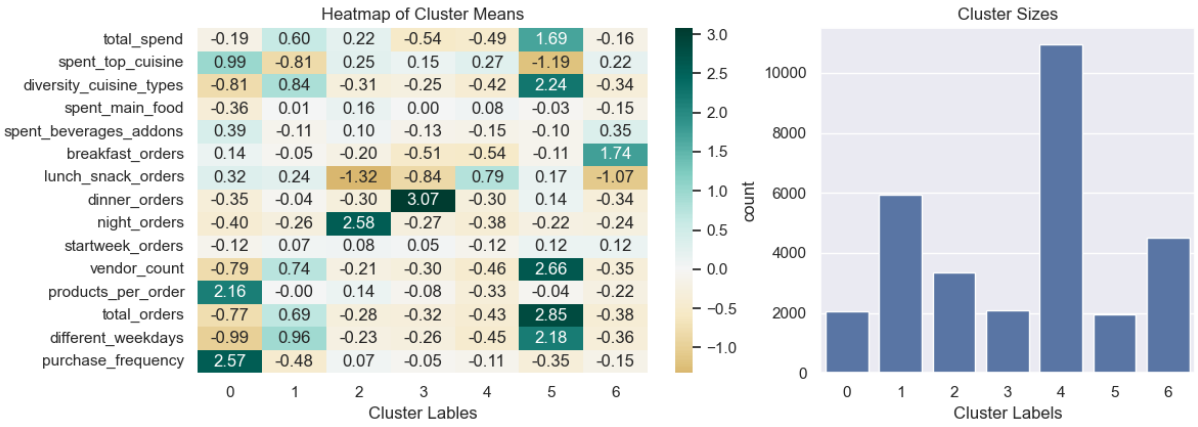

Figure 17 – Final Clustering Results for merged perspectives using K-Means

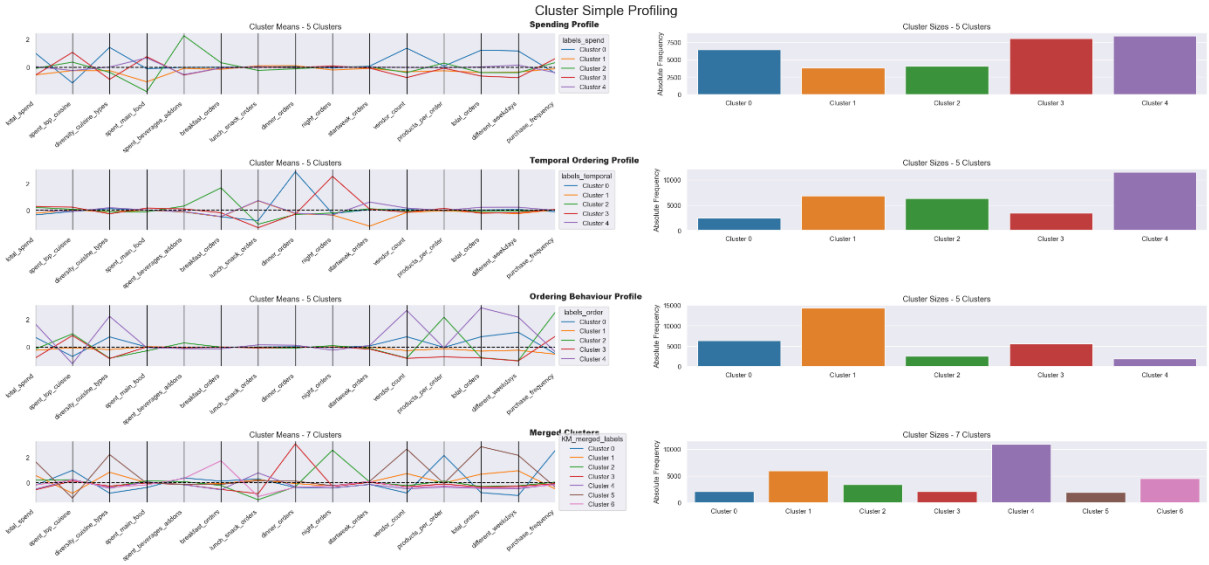

Figure 18 – Profile of final Clustering Results for merged perspectives

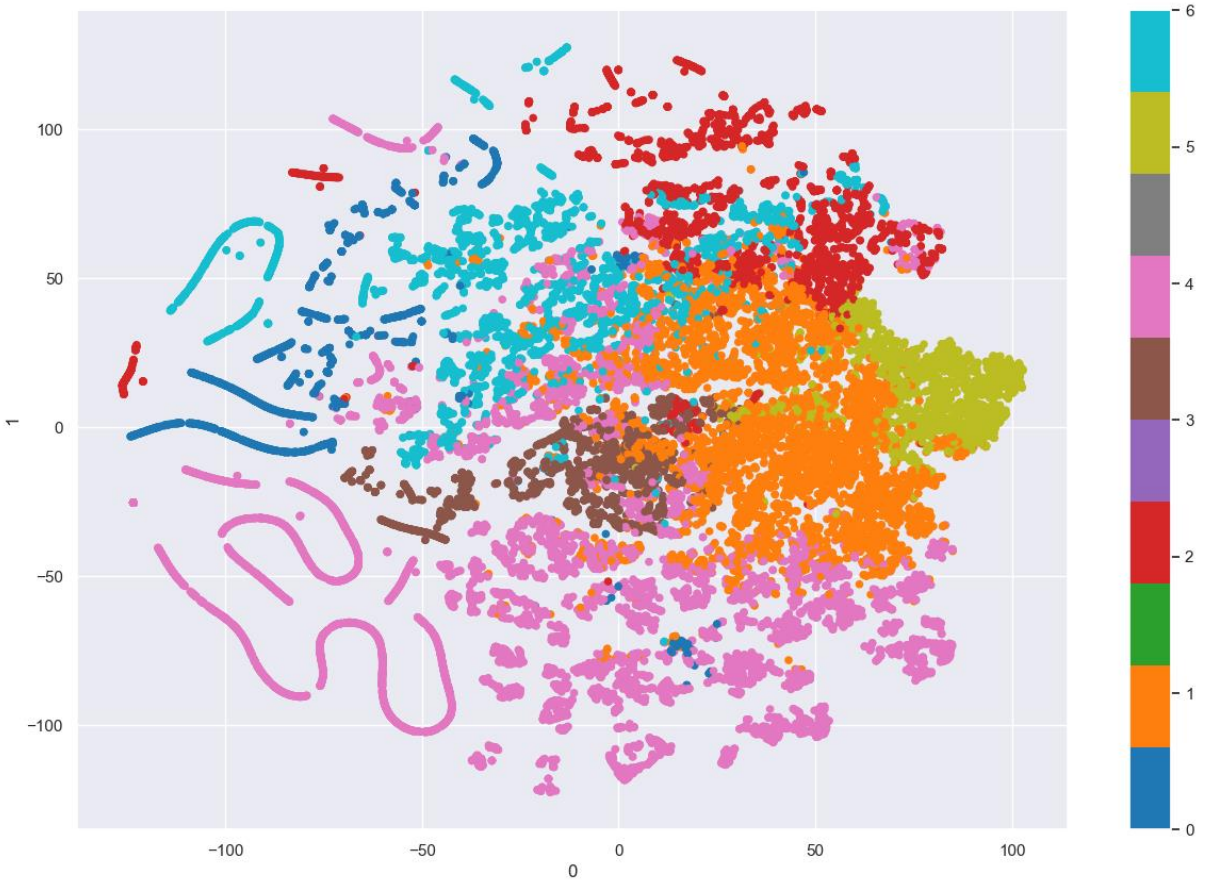

Figure 19 – T-SNE Visualization of final Clustering Results for merged perspectives

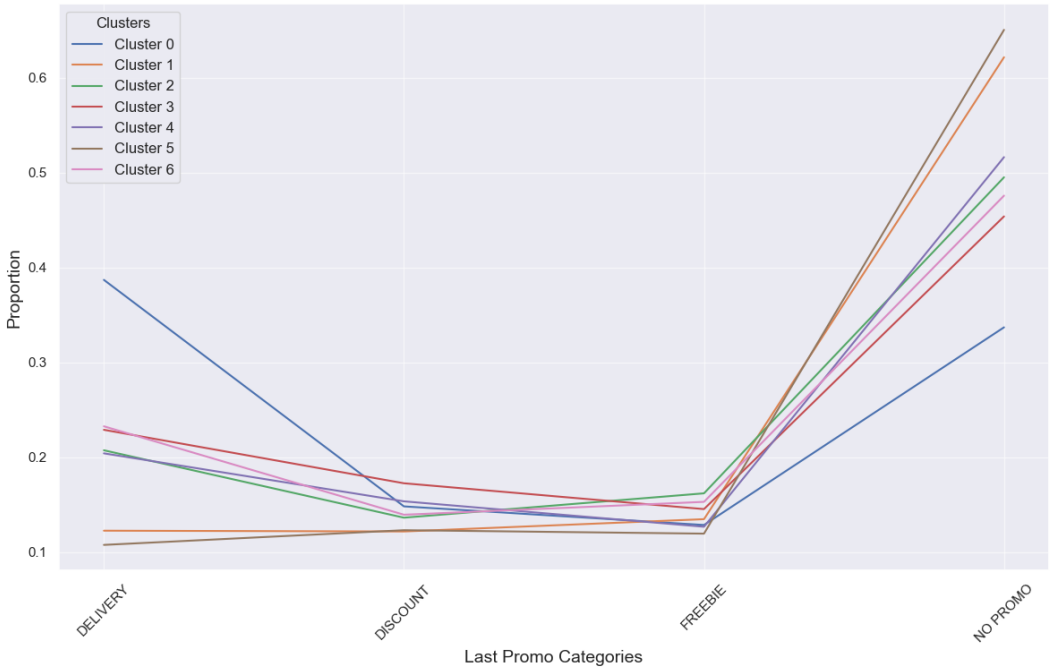

Figure 20 – Proportions of Last Promotions across final clusters

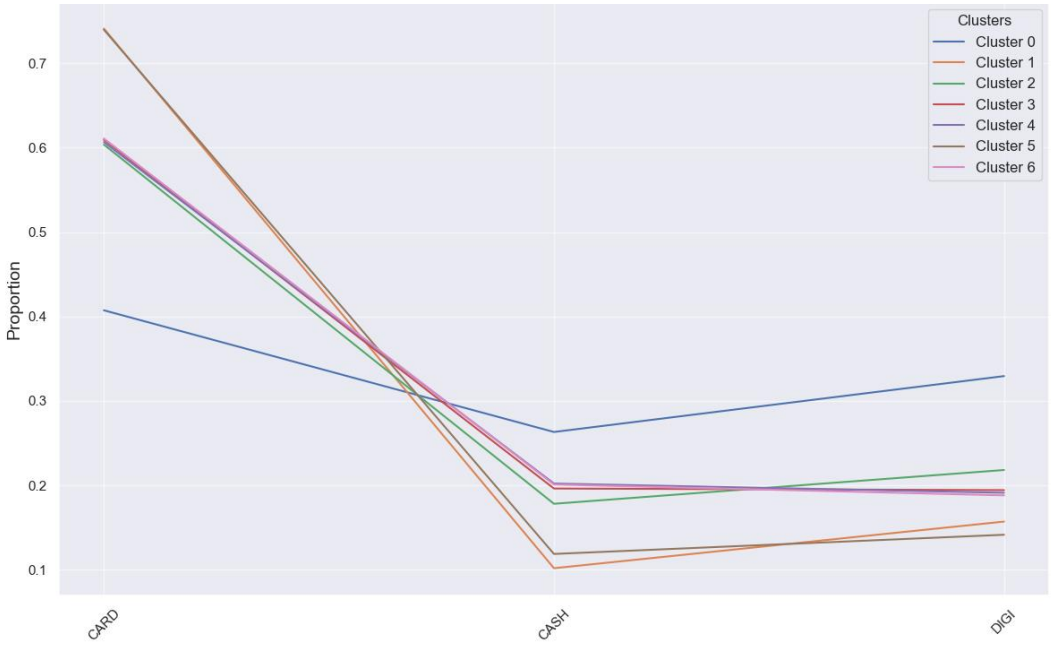

Figure 21 – Proportions of Payment Methods across final clusters

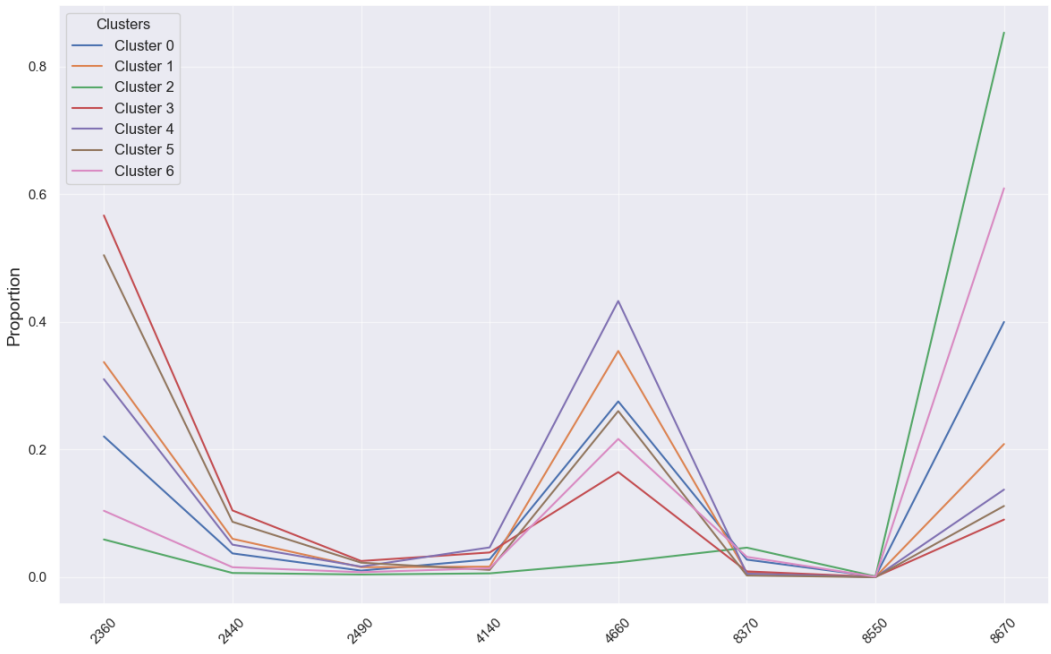

Figure 22 – Proportions of Regions across final clusters

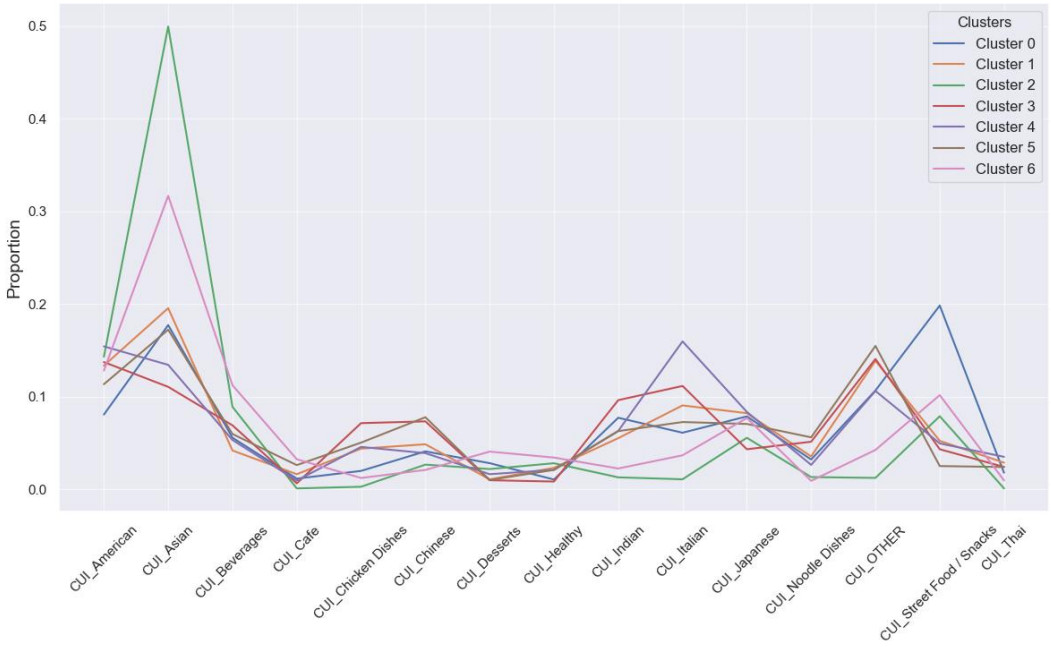

Figure 23 – Proportions of preferred cuisines across final clusters

APPENDIX B – TABLES

Feature	Description
spent_top_cuisine	Represents the proportion of a customer's total spending allocated to their most frequently chosen cuisine, providing insight into how preferences influence spendings.
spent_per_order	Since *total_spent* often parallels *total_orders* (the more orders placed, the higher the total spending), this feature calculates the average amount spent per order. By focusing on spending per order, we can see the variations in individual purchasing habits that might be obscured by aggregated spending measures, offering a more detailed perspective on spending tendencies.
products_per_order	Following a similar approach, this feature measures the average number of products per order. This allows for a detailed examination of ordering habits, highlighting variations in product preferences and order composition across customers.
spend_main_food and ***spend_beverages_addons***	Considering the high dimensionality in the dataset regarding cuisine types, these features were created to summarize spending into two categories: main food items and beverages/add-ons.

Table 7-Description of the new features created in Feature Engineering

Clustering Technique	Decision Process
Hierarchical Clustering	The R Squared metric revealed an optimal range of 3 to 5 clusters. To further validate the best number of clusters, a dendrogram was performed to help visualize the distances, and the silhouette score was calculated for a range from 3 to 6 clusters, with the results pointing to 4 or 5 clusters as the optimal solutions. After conducting the analysis and testing different cluster configurations, it was decided to proceed with **five clusters**.
K Means	The inertia plot was used to evaluate cluster validity, leading to a solution of 4 or 5 clusters. However, after conducting a silhouette analysis and checking the used metrics, we concluded that **four clusters** represent the dataset the better.
SOM Hierarchical	**1.** The SOM was trained using a 50x50 grid, which exceeds the size recommended by the rule of thumb **Rule of Thumb SOM Grid** This larger grid size yielded significantly better results in terms of quality and topology preservation. After training, the **Quantization**

	Error (QE) decreased by approximately **23%**, reaching a final value of **0.163**, while the **Topographic Error (TE)** showed an improvement of about **71%**, dropping to **0.2851**. These results suggest that the SOM successfully learned the underlying structure of the data, accurately representing both the input data points and their topological relationships. To determine the optimal number of clusters (K), hierarchical clustering was applied to the trained SOM weights using Ward linkage and Euclidean distance. The analysis of silhouette scores ranging from approximately 0.21 to 0.31 across different clusters, with good results and separation using five clusters. Additionally, the dendrogram provided a clear distinction among the clusters, supporting the selection of K = 5.
SOM K Means	Using the same training batch, K-Means clustering was applied to the SOM's weight vectors, identifying distinct groups within the mapped data. To determine the optimal number of clusters, we applied KMeans with different values of **K** and evaluated the performance using the **Silhouette Score**, **Calinski-Harabasz Index**, and **Davies-Bouldin Index** The hexagon plot showed a good definition between clusters, confirming the use of this five, the result of R^2 was 0.63 which means **37% of the variance** remains unexplained (SSW), indicating potential noise or unaccounted complexity in the data.
Mean Shift	For this analysis, the Mean Shift algorithm was used to identify clusters based on the density of data points. The bandwidth parameter, which defines the size of the neighbourhood considered for clustering, was estimated using a 0.15 quantile. This value determines how far the algorithm searches for neighbouring points to form clusters. A smaller bandwidth results in more granular clusters, while a larger bandwidth produces fewer, broader clusters. The estimated bandwidth was approximately **1.72**, suggesting a moderate neighbourhood size for density estimation. With the calculated bandwidth, Mean Shift clustering was performed, resulting in **11 estimated clusters**. These clusters were determined based on density peaks, which represent areas with high data point concentrations. Unlike previous clustering methods, Mean Shift does not require predefining the number of clusters, making it suitable for exploratory analysis in datasets with unknown cluster structures.

Table 8-Decision Process for determining the optimal number of clusters in Perspective 1

Clustering Techniques	Number of Clusters	Average Silhoutte Score	Adjusted R squared	Calinski-Harabasz Index	Davies-Bouldin Index
Hierarchical Clustering	K = 5	0.2623	0.5717	10273.80	1.4104
K-Means	K = 4	0.3179	0.5660	13384.98	1.1046
SOM Hierarchical	K = 5	0.2616	0.5920	11162.64	1.4108
SOM K-Means	K = 5	0.2955	0.6309	11942.20	1.2163
Mean Shift	K = 11	-	0.332	-	-

Table 9 - Results of Clustering Process for Perspective 1

labels	customer_age	total_spend	diversity_cuisine_types	purchase_frequency	different_weekdays
0	40.119492	25.980826	1.975106	0.285162	2.395818
1	24.794035	25.493893	2.028524	0.148038	2.460074
2	26.871375	71.959901	4.065677	0.201461	4.921187
3	26.175849	15.500876	1.065662	1.453759	1.064253

Table 10 - Results of Clustering Process for Perspective 1 using K-Means

Clustering Technique	Decision Process
Hierarchical Clustering	The R Squared metric revealed an optimal range of **3 to 6 clusters**. To further validate the best number of clusters, a dendrogram was performed to help visualize the distances, and the silhouette score was calculated for a range from **5 to 6 clusters**, with the results pointing to **5 clusters** as the optimal solutions. After conducting the analysis and testing different cluster configurations, it was decided to proceed with **five clusters**.
K Means	The inertia plot was used to evaluate cluster validity, leading to a solution of 4 or 5 clusters. However, after conducting a silhouette analysis and checking the used metrics, we came to the conclusion that **seven clusters** represent the dataset the better.
SOM Hierarchical	The SOM was trained using a 50x50 grid, which exceeds the size recommended by the rule of thumb[1]. This larger grid size yielded significantly better results in terms of quality and topology preservation. The **Quantization Error (QE)** decreased by approximately 34% after training, to 0.0806, while the **Topographic Error (TE)** improved about 70%, reducing to 0.2965. These results indicate that the SOM effectively captured the topological relationships within the input data.

	The U-Matrix and the Hit-map revealed significant patterns corresponding to possible clusters. The hexagonal plot also displayed a clean separation between almost every cluster.
SOM K Means	Using the same training batch, K-means was applied and the inertia plot revealed the best range of clusters were **3 to 5** but after the metrics analysis **k=5** was the best one. The hexagon plot showed a clear definition between clusters, confirming the use of this five, the result of R^2 was 0.66 which means **66% of the variance** is explained by this model .
Mean Shift	For the bandwidth, various quantile values were tested so the number of clusters were not too big nor too small for our dataset. The number of clusters given with 0.1 quantile was 7. The **R^2** result was 0.69.

Table 11 - Decision Process for determining the optimal number of clusters in Perspective 2

labels	total_spend	spent_top_cuisine	diversity_cuisine_types	spent_main_food	spent_beverages_addons
0	25.174598	0.613790	2.451521	0.890466	0.026155
1	19.916141	0.987173	1.101739	0.997364	0.001429
2	28.966604	0.529687	2.540226	0.406919	0.485069
3	54.337248	0.393350	5.295371	0.717080	0.131378
4	27.711247	0.933683	1.320273	0.030851	0.964231
5	97.991380	0.648000	3.017915	0.783566	0.161808
6	16.101957	0.899596	1.481198	0.090060	0.010344

Table 12-Results of Clustering Process for Perspective 2 using K-Means

label	total_spend	spent_top_cuisine	diversity_cuisine_types	spent_main_food	spent_beverages_addons
0	21.050512	0.934063	1.301444	0.914751	0.005763
1	98.070524	0.623697	3.113192	0.759473	0.180345
2	56.903251	0.390019	5.508044	0.703121	0.152239
3	25.376074	0.541098	2.659986	0.738987	0.100669
4	29.271849	0.827812	1.700502	0.114055	0.851310

Table 13-Results of Clustering Process for Perspective 2 using SOM K-Means

Clustering Technique	Decision Process
Hierarchical Clustering	The R Squared metric revealed an optimal range of **3 to 6 clusters**. To further validate the best number of clusters, a dendrogram was performed to help visualize the distances, and the silhouette score was calculated for a range from **4 to 7 clusters**, with the results pointing to **5 clusters** as the optimal solutions. After conducting the analysis and testing different cluster configurations, it was decided to proceed with **five clusters**.
K Means	The inertia plot was used to evaluate cluster validity, leading to a solution of **3 to 5 clusters**. However, after conducting a silhouette analysis and checking the used metrics, we came to the conclusion that **five clusters** represent the dataset the better. The **R^2** also showed a positive result of 0.72.
SOM Hierarchical	The SOM was trained using a 30x30 grid, which exceeds the size recommended by the rule of thumb[1]. This larger grid size yielded significantly better results in terms of quality and topology preservation. The **Quantization Error (QE)** decreased by approximately 65% after training, to 0.0281, while the **Topographic Error (TE)** improved about 86%, reducing to 0.1372. These results indicate that the SOM effectively captured the topological relationships within the input data. The U-Matrix and the Hit-map revealed significant patterns corresponding to possible clusters. The hexagonal plot also displayed a clean separation between almost every cluster.
SOM K Means	Using the same training batch, K-means was applied and the inertia plot revealed the best range of clusters were **3 to 6** but after the metrics analysis **k=6** was the best one. The hexagon plot showed a clear definition between clusters, confirming the use of this five, the result of R^2 was 0.75 which means **75% of the variance** is explained by this model .
Mean Shift	For the bandwidth, various quantile values were tested so the number of clusters were not too big nor too small for our dataset. The number of clusters given with 0.15 quantile was 9. The **R^2** result was 0.74.

Table 14 - Decision Process for determining the optimal number of clusters in Perspective 3

labels	breakfast_orders	lunch_snack_orders	dinner_orders	night_orders	startweek_orders
0	0.049134	0.304108	0.608093	0.038665	0.678650
1	0.051032	0.899986	0.018037	0.010814	0.245426
2	0.747922	0.196057	0.007398	0.048623	0.701612
3	0.146555	0.091562	0.011619	0.750263	0.691751
4	0.056721	0.908200	0.025181	0.009899	0.877050

Table 15-Final Clustering Results for Perspective 3 using K-Means

label	breakfast_orders	lunch_snack_orders	dinner_orders	night_orders	startweek_orders
0	0.051899	0.926345	0.014191	0.007565	0.822820
1	0.771941	0.183088	0.005825	0.039146	0.706495
2	0.157894	0.106371	0.010067	0.725669	0.689573
3	0.012111	0.090283	0.883668	0.013939	0.694168
4	0.074932	0.484340	0.390646	0.050081	0.678894
5	0.071242	0.874040	0.011617	0.014333	0.137450

Table 16-Final Clustering Results for Perspective 3 using SOM K-Means

Clustering Technique	Decision Process

Clustering Technique	Decision Process
Hierarchical Clustering	The R Squared metric revealed an optimal range of **3 to 5 clusters**. To further validate the best number of clusters, a dendrogram was performed to help visualize the distances, and the silhouette score was calculated for a range from **3 to 6 clusters**, with the results pointing to **6 clusters** as the optimal solutions. After conducting the analysis and testing different cluster configurations, it was decided to proceed with **six clusters**.
K Means	The inertia plot was used to evaluate cluster validity, leading to a solution of **3 to 5 clusters**. However, after conducting a silhouette analysis and checking the used metrics, we came to the conclusion that **six clusters** represent the dataset the better. The **R^2** also showed a positive result of 0.78.
SOM Hierarchical	The SOM was trained using a 30x30 grid, which exceeds the size recommended by the rule of thumb[1]. This larger grid size yielded significantly better results in terms of quality and topology preservation. The **Quantization Error (QE)** decreased by approximately

	40% after training, to 0.1487, while the **Topographic Error (TE)** improved about 88%, reducing to 0.1222. These results indicate that the SOM effectively captured the topological relationships within the input data. After analysing the metrics used , 5 clusters were the best choice for this method, with an **R^2** of 0.68.
SOM K Means	Using the same training batch, K-means was applied and the inertia plot revealed the best range of clusters were **3 to 5** but after the metrics analysis **k=5** was the best one. The hexagon plot showed a clear definition between clusters, confirming the use of this five, the result of R^2 was 0.67 which means **67% of the variance** is explained by this model .
Mean Shift	For the bandwidth, various quantile values were tested so the number of clusters were not too big nor too small for our dataset. The number of clusters given with 0.2 quantile was 19. The **R^2** result was 0.57.

Table 17 - Decision Process for determining the optimal number of clusters in Perspective 4

labels	vendor_count	products_per_order	total_orders	different_weekdays	purchase_frequency
0	2.358562	1.077164	2.933001	2.339155	0.133977
1	1.060413	0.983890	1.088743	1.047092	1.013946
2	4.604823	1.262230	6.603985	4.321165	0.160503
3	2.257604	1.703123	2.732028	2.198157	0.214324
4	1.086101	2.166134	1.121361	1.042640	2.244177
5	8.760996	1.267198	13.735029	5.948066	0.250294

Table 18-Final Clustering Results for Perspective 4 using K-Means

Number of Clusters	Metric	Hierarchical Clustering	K-Means
K=5	Average Silhoutte Score	0.1931	**0.2009**
	Calinski-Harabasz Index	3774.64	**4349.66**
	Davies-Bouldin Index	**1.5703**	1.6026
	Adj. R squared	0.3291	**0.3611**
K=6	Average Silhoutte Score	**0.1954**	0.1807
	Calinski-Harabasz Index	**3862.77**	3642.63
	Davies-Bouldin Index	**1.5296**	1.6004
	Adj. R squared	**0.3854**	0.3717
K=7	Average Silhoutte Score	**0.1878**	0.1767
	Calinski-Harabasz Index	3301.72	**4516.28**
	Davies-Bouldin Index	1.5262	**1.5688**
	Adj. R squared	0.3914	**0.4682**

Table 19-Comparison between Hierarchical Clustering and K-Means through different number of clusters for Multiple Perspective

merged_labels	0	1	2	3	4	5	6
total_spend	26.912961	48.864010	38.359872	16.878309	18.425222	79.325857	27.666915
spent_top_cuisine	0.981016	0.555213	0.806626	0.782387	0.812186	0.464773	0.798324
diversity_cuisine_types	1.077941	3.344897	1.762687	1.845750	1.607865	5.265496	1.722198
spent_main_food	0.600372	0.733010	0.786721	0.730954	0.758346	0.718925	0.677588
spent_beverages_addons	0.292770	0.137080	0.202222	0.130490	0.126532	0.139674	0.279813
breakfast_orders	0.252451	0.191347	0.144939	0.045208	0.037118	0.172873	0.764129
lunch_snack_orders	0.745531	0.714135	0.089756	0.281106	0.932132	0.685336	0.190950
dinner_orders	0.001038	0.058361	0.010803	0.638925	0.011399	0.093209	0.003589
night_orders	0.000980	0.036157	0.754502	0.034761	0.006730	0.048582	0.041333
startweek_orders	0.623219	0.685864	0.690224	0.678782	0.622165	0.706267	0.704447
vendor_count	1.106863	4.459751	2.386866	2.192932	1.844353	8.678202	2.080961
products_per_order	2.193513	1.282476	1.341736	1.248908	1.143827	1.265625	1.189054
total_orders	1.160294	6.200573	2.839403	2.693410	2.334065	13.663223	2.485988
different_weekdays	1.076961	4.070057	2.248955	2.192455	1.901509	5.946281	2.048710
purchase_frequency	2.176882	0.159708	0.521291	0.445132	0.406685	0.249074	0.376270

Table 20-Final Clustering Results for merged perspectives

APPENDIX C – DEFINITIONS

2. Calinski-Harabasz Index

Also known as the Variance Ratio Criterion, Calinski-Harabasz Index (CH) is a metric used to evaluate the quality of clustering solutions by examining the relationship between the dispersion of data points within clusters and between clusters. A higher value of the index indicates better-defined and more distinct clusters.

The CH is computed as:

$$CH = \frac{Tr\,(B_k)}{Tr\,(W_k)} \cdot \frac{n-k}{k-1}$$

Where n is the total number of data points, k the number of clusters, $Tr\,(B_k)$ the trace of the between-cluster dispersion matrix and $Tr\,(W_k)$ the trace of the within-cluster dispersion matrix.

The Between-Cluster Dispersion Matrix (B_k) quantifies the variance of cluster centroids from the overall mean. The Within-Cluster Dispersion Matrix (W_k) measures the variance of points within each cluster around their respective centroids.

The CH index evaluates how compact the clusters are (minimizing $Tr\,(W_k)$) while maximizing the separation between clusters (maximizing $Tr\,(B_k)$). The adjustment factor $\frac{n-k}{k-1}$ accounts for the number of clusters, penalizing solutions with a larger number of clusters to avoid overfitting.

3. Davies-Bouldin Index

The Davies-Bouldin Index is a clustering evaluation metric that assesses the quality of a clustering solution by measuring the average similarity ratio between each cluster and its most similar cluster. A lower value of the index indicates better clustering, as it means compact and well separated clusters.

The DB is computed as:

$$DB = \frac{1}{k}\sum_{i=1}^{k} \max\,(\frac{s_i + s_j}{d_{ij}})$$

Where k is the number of cluster, s_i the average intra-cluster distance for cluster i and d_{ij} the distance between the centroids of clusters i and j, to calculate the maximum similarity ratio between cluster I and any other cluster j.

The intra-cluster distance (s_i) measures the average dissimilarity between all points within a cluster and its centroid.

The inter-cluster distance (d_{ij}) quantifies the separation between clusters i and j, typically computed as the Euclidean distance between their centroids.

The DB index calculates the worst-case similarity ratio for each cluster and then averages these worst-case values across all clusters. A lower DB score reflects compact clusters (s_i is small) that are well-separated (d_{ij} is large).

4. Rule of Thumb SOM Grid

A widely referenced heuristic for determining the size of a Self-Organizing Map grid is to set the total number of neurons (i.e., the product of the grid's width and height) to approximately $5 \times \sqrt{N}$ where N is the number of training samples. This guideline seeks to balance the representational capacity of the map with its interpretability and computational overhead. In practice, once this approximate total is established, the grid dimensions (e.g., $m \times n$) are chosen so that $m \times n \approx 5 \times \sqrt{N}$.