DATA MINING PROJECT

Master's in Data Science and Advanced Analytics

NOVA Information Management School

Universidade Nova de Lisboa

ABCDEats Inc.

Final Report

Group 02

Sarah Leuthner, 20240581

Catarina Silva, 20240558

Bruna Duarte, 20210669

Afonso Gião, 20240495

CONTENTS

Introduction 1
1. Data Exploration 1
    1.1. Data Review 1
2. Data Preprocessing 2
    2.1. Anomalies, Duplicates and Strange Values 2
    2.2. Missing Values 2
    2.3. Outlier Treatment 2
    2.4. Feature Engineering 2
    2.5. Feature Selection 2
    2.6. PCA Check 3
    2.7. Scaling 3
3. Cell-based segmentation – RFM analysis 3
4. Perspectives 4
5. Clustering 4
    5.1. K-Means with perspectives and Hierarchical Clustering 4
    5.2. Self-Organizing Maps (SOM) 4
    5.3. Density Clustering 5
    5.3.1. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) 5
    5.3.2. Mean-Shift Algorithm 5
    5.3.3. Gaussian Mixture Model (GMM) 5
    5.4. Comparison of Clustering Methods 5
6. Clusters Analysis & Profiling 6
    6.1. Profiling with RFM 8
7. Marketing Strategies 9
8. Conclusion 10
Bibliographical References 11
Appendix Figures 12
Appendix Tables 19
Annex 23

INTRODUCTION

This project focuses on customer segmentation for *ABCDEats Inc.*, a food delivery company, to enhance their marketing strategies and improve customer retention. The dataset reunites three months of customer data. Our analysis aims to extract meaningful insights into *ABCDEats Inc.'s* customer base, addressing challenges like low retention rates and underutilized promotional offers, while identifying opportunities for tailored marketing strategies.

Firstly, a profound exploratory data analysis to identify patterns, trends, and anomalies in the data was done. This analysis revealed a young customer base (age 23–31), which predominantly orders from a limited selection of vendors and prefers Asian, American, and street food cuisines. Many customers exhibit low order frequency, and almost half do not use promotional codes. Further preprocessing steps were addressed such as treating missing values, outliers, and redundancy to ensure high-quality inputs for the subsequent clustering.

Feature engineering and scaling prepared the dataset for advanced clustering methods, such as k-means, hierarchical clustering and density-based algorithms. The clustering analysis focused mainly on two perspectives: customer behavior and demographics. The optimal clustering solution reached combined both these perspectives in four distinct clusters, that highlight important differences in spending patterns, loyalty, and cuisine preferences. Techniques like Principal Component Analysis (PCA) and Self-Organizing Maps (SOM) supplemented the work but were ultimately excluded due to interpretability concerns.

The final segment of this project provides an impactful array of marketing strategies. Key recommendations include loyalty programs for high-frequency users, targeted promotions for underperforming cuisines, and engagement strategies for low-retention customers. These strategies are expected to boost customer satisfaction, drive growth, and establish *ABCDEats Inc.* as a leader in the competitive food delivery market.

1. DATA EXPLORATION

1.1. Data Review

The customer base of the dataset is primarily composed of young adults, aged between 23 and 31, who tend to place orders from a limited number of vendors (3 on average), suggesting potential for marketing strategies based on their loyalty (e.g.: loyalty cards - after 10 meals, the customer gets 1 free). Most customers order infrequently with a high concentration of orders from Asian, American, and street food cuisines, indicating a need for tailored promotions. *ABCDEats Inc.* also faces challenges with customer retention, as many customers place only a few orders (average of 6, but median of 3). According to *First Order*, recently the growth of new customers has slowed significantly. On average, the customers started placing orders from the 28th day of services and three quarters of the customer base joined in the first half of the 90-day period.

Opportunities for targeted marketing are clear, especially in major customer regions, as almost 90% of customers are from three of all registered customer regions. It's crucial to apply strategies that increase engagement, such as offering incentives for underperforming cuisines or customers at risk of churning (not forgetting, that more than half of the customers are not using *Promotion Codes*). It might be important to study the causes of this more deeply, to get marketing strategies that are tailored to

our customer base. Furthermore, the report highlights anomalies, such as missing data and "nonsense" values (e.g. customers who are registered on the system but have not ordered yet). These issues were swiftly addressed to refine customer segmentation and optimize future business operations.

2. DATA PREPROCESSING

2.1. Anomalies, Duplicates and Strange Values

As discussed in the previous work, some inconsistencies were present in the dataset, such as duplicate rows and customers with 0 orders. These observations were dropped as they constituted an insignificant part of the dataset. Additionally, the strange value '-' was identified in both *Customer Region* and *Last Promo* columns. In the *Customer Region* column, it was replaced with NaN and addressed in the following section as a missing value. In the *Last Promo* column, where this value was very frequent and there was no category to express the absence of a discount, it was assumed that '-' indicated a full price last purchase. This category was renamed 'FULL PRICE' to enhance interpretability.

2.2. Missing Values

The missing values in the dataset were treated or dropped from the Dataframe. Some values were treated since deleting the variables or the instances would lead to loss of relevant information. By treating each variable by itself, the potential for introducing bias or the risk of misinterpretation is minimized, and data quality is improved. Other missing values were dropped as the number of rows was low and more difficult to replace with reasonable values given dependencies with another variable. Table 3 in the Appendix displays the treatment strategy for each feature.

2.3. Outlier Treatment

To prevent a misleading impact on the interpretation of the results, the outliers in the data set were treated. By manually defining the threshold based on boxplots, instead of using the Interquartile Range (IQR), subjectivity and lack of generalization were introduced, but it was necessary due to the high skewness of some features. Consequently, the outliers were defined context-specific based on knowledge of each feature. A table with each threshold can be found in the Appendix (Table 4). For some *DOW* and all *HR* features no threshold was defined, as their outliers were giving the most or all the relevant information.

2.4. Feature Engineering

Referring to Table 5 in the Appendix for a list of all the new features created, the old features were transformed into new ones that better capture customer behavior and variance in the data.

2.5. Feature Selection

Most of the original features were dropped to reduce redundancy and enhance the dataset's quality (*Vendor Count, Product Count, Is Chain and First Order*). We checked the correlation between numerical variables using a **threshold of 0.75**, and decided to remove some of the new features that could possibly give us some repeated information (*Product Intensity*, *Avg_Spend_per_Vendor*, *Spend per Product*, *Churn Risk*) – refer to Figure 6 in the Appendix for the correlation heatmap with all the features. As the analysis was further designed to use specific perspectives, only a subset of these features was retained, as detailed in Chapter 4 of this report.

2.6. PCA Check

To reduce the dimensionality of our dataset while still retaining key variability, we tried applying Principal Component Analysis (PCA). After closely analyzing the Scree plot and the variance explained by each principal component (Figure 7), four components were selected that accounted for a major amount of variability. Despite their significant explanatory power, they were not included in the final analysis. Incorporating these components made the results harder to interpret and the loss of information was too critical for the clustering analysis later (e.g.: PC0 combined *Week_days_mean*, *Weekend_days_mean*, *Lunch_HR_mean*, *Customer_Duration* and *Total_Orders*, which are all important features for the demographic perspective).

2.7. Scaling

The numerical data in the dataset was scaled to prepare for clustering. The two approaches, Minmax Scaler and Standard Scaler, were compared to finding the scaler which best fits the dataset. After comparing features using boxplots, before and after scaling, the **Standard Scaler** was chosen due to its robustness and resistance to outliers.

3. CELL-BASED SEGMENTATION – RFM ANALYSIS

The Recency Frequency Monetary (RFM) analysis is a technique used for customer segmentation to identify valuable customer niches with significant economic importance. Each segment is ranked according to 3 customer importance criteria: recency, frequency and monetary. The best customers are the ones that score high in these 3 metrics. This means that valuable clients tend to be the ones that made a purchase recently, order products very frequently and spend a lot of money.

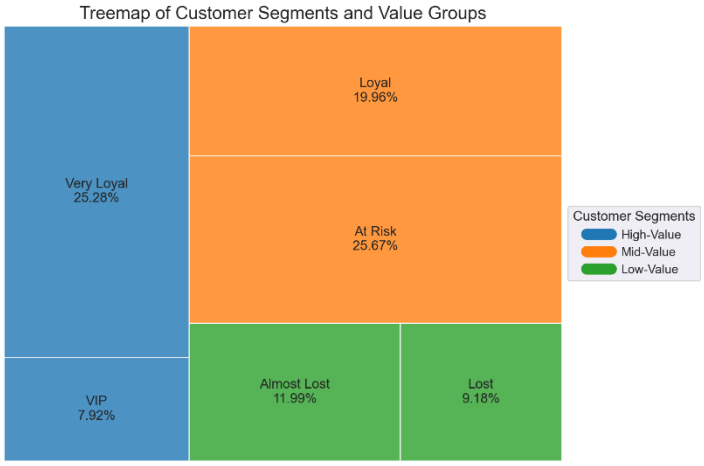

Figure 1 – Tree Map of Customer Segments and Value Groups

To evaluate customers relying on these 3 concepts, *Last Order* and *Total Orders* columns were used, combined with a new column made from the sum of all cuisine columns representing the total money the customer has spent. Each customer was assigned a score of 1 to 4 for each criterion based on their attributes relative position. The final score was the sum of all individual scores.

The next step was creating the value segments and the value groups based on the final score. The first segment 'High-Value' contains VIP and very loyal customers. The 'Mid-Value' segment contains loyal customers as well as users that make good contributions but are at risk and need to be supervised. The last segment, 'Low-Value', encompasses customers with very low activity that may need some re-

engagement strategies or targeted incentives to increase their activity and loyalty. These customers were divided into 2 ordered categories: almost lost and lost. The tree map below provides a clear visualization of how customers are ranked (Figure 1). The high-value and mid-value segments hold a great portion of *ABCDEats Inc.*'s customers. Almost half of the customers are considered mid-value and only around 20% are low value. VIP customers make up just around 8%, yet they contribute the most value individually.

4. PERSPECTIVES

Table 1 – Perspectives Created

Customer Behavior		**Demographic Information**
- CUI_American	- Order Frequency	- Weekdays (mean)
- CUI_Italian	- Average Spend per Order	- Weekend Days (mean)
- CUI_Asian	- Product per Order	- HR peak
- CUI_OTHER	- Total Orders	- DOW peak
- Chain Preference	- Customer Duration	- Customer Age

For a more detailed and tailored clustering analysis, two perspectives were used: customer behavior and demographic information (Table 1). The perspective "Customer Behavior" gives relevant insights into the purchasing habits and engagement levels of the customers, which are essential for understanding their priorities like food variety, spending patterns, order frequency, etcetera.

On the other hand, the demographic perspective includes average activity and customer age. This perspective was built with the intention of segmenting customers based on their lifestyle, daily routines, and basic demographic traits to study how these factors influence ordering tendencies. By combining these two perspectives, we created a more detailed segmentation, allowing *ABCDEats Inc*. to develop more personalized marketing strategies that attend to the specific needs of each customer segment.

5. CLUSTERING

5.1. K-Means with perspectives and Hierarchical Clustering

To determine the optimal clustering method for demographic and behavioral data, the quality measure score R^2 was calculated for each cluster solution on the variables. Using both k-means and hierarchical clustering (with various linkages, namely complete, average, single and ward), k-means performed the best, with the optimal number of clusters being between 4 and 5 for both perspectives. Validating this choice with inertia plots, revealed an elbow somewhere at 3 or 4 clusters. Therefore, 4 clusters for each perspective were kept, ensuring a robust solution supported by both R^2 and inertia analysis. Once the number of clusters was chosen, the perspectives were then merged. The results obtained with manual and hierarchical merging are presented in the Appendix, Table 6 and Table 7.

5.2. Self-Organizing Maps (SOM)

Another clustering method and multidimensional data visualization technique used was Self-Organizing Maps, where clusters and outliers can be detected. Each perspective was brought into a 2-dimensional space using component planes, U-matrices and Hits-plots for visualizing potential clusters and interpretation. The graphs for both perspectives can be seen in the Appendix (Figure 8 to Figure 13). Despite an unfolding phase with a lot of iterations, the SOM with k-means and hierarchical

clustering did not show clear structured clusters and present a low R^2-value. However, some interpretation could be found in the component planes regarding outliers.

5.3. Density Clustering

Three density clustering methods were tested. The methods applied were DBSCAN, Mean-Shift Algorithm and Gaussian Mixture Model, with the latter performing the best.

5.3.1. DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

Retaining the number of clusters of 5 for both perspectives, the variables of epsilon and minimum samples were tuned. A K-distance graph to find out the right epsilon value was used. This resulted in a proportion of the variance in the data, R^2, of 11.03% for the demographic perspective, and of 17.92% on the behavioral perspective.

5.3.2. Mean-Shift Algorithm

Testing the Mean-Shift algorithm, to contain the number of clusters into 5 in both perspectives, the value of quantile was raised, to merge nearby clusters into fewer groups. This resulted in an unsatisfying proportion of the variance in the data, R^2 value, (Demographic: 9.71% and Behavioral: 7.91%) for both perspectives.

5.3.3. Gaussian Mixture Model (GMM)

Referring to the graphs on Figure 14 and Figure 15 in the Appendix, the number of components for this density clustering model was selected based on AIC and BIC. The quality measure R^2 was of 39.29% on the demographic perspective, and 33.36% for the behavioral perspective. These were the best results of R^2 achieved with density clustering methods.

The visual representation of these clustering solutions was created through PCA dimensionality reduction, and can be found on Figure 16 and Figure 17, for demographic and behavioral perspectives, respectively. In both figures, we can detect the presence of 6 clusters, that don't seem to be clearly segregated, which is also not very satisfying as a result.

5.4. Comparison of Clustering Methods

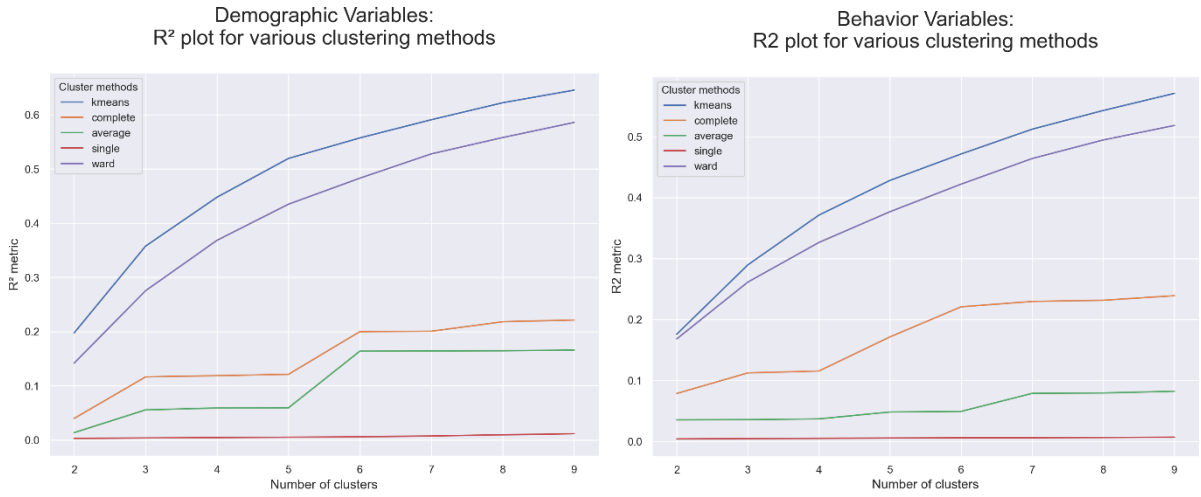

Figure 2 – R^2 plots for multiple Clustering Methods

To choose the overall best clustering method for this dataset, the quality measure score R^2 was compared across all applied techniques. Figure 2 shows, for both perspectives, a line graph with the R^2-score for several clusters of k-means and different forms of hierarchical clustering. Table 2 shows the same for a specific number of clusters, for all other tested clustering methods, and for both perspectives. After analyzing and comparing all the clustering methods used, **K-Means with Hierarchical** was selected, as it presented the highest R^2-score in both perspectives.

Table 2 – Clustering Methods' Quality Analysis

Methods	Clusters	R^2 - Demographic	R^2 - Behavioral
K-Means + Hierarchical	**4**	**0.4477**	**0.3715**
SOM	4	0.4414	0.3135
DBSCAN	5	0.1103	0.1792
Mean Shift Clustering	5	0.0971	0.0791
Gaussian Mixture Model	6	0.3929	0.3336

6. CLUSTERS ANALYSIS & PROFILING

For the final clusters, a manual merging strategy for K-Means results was applied. The results depict 4 clusters.

From a **behavioral perspective**, two very imbalanced clusters were obtained, with cluster 0 having almost five times more observations than cluster 1. Overall, cluster 0 shows a very standard behavior, only distinguishing itself by its low frequency. Contrarily so, individuals in cluster 1 show more irregular tendencies. They tend to order on very uncommon days, both during the week and weekend, but especially at the weekend. Nonetheless, they spend a bit less per order than individuals from cluster 0. Cluster 1 also shows a profoundly low loyalty to the service, presenting drastically small values for *Customer Duration*. Nevertheless, they order very frequently, even if they don't order many products per order. Cluster 0 seems to portrait an average customer, while cluster 1 shows more unpredictable and "opportunistic" users. Cluster 1 individuals might only use the app sporadically or when specific promotions or circumstances are in favor of their needs. Despite their frequent ordering, the small basket sizes and lower overall spending indicate a focus on potentially impulsive purchases rather than planned shopping experiences. This cluster represents a distinct segment that might require tailored strategies to encourage greater loyalty or to better deal with their unique behavior patterns. It's also important to notice that none of the clusters order any specific type of Cuisine, but cluster 1 presents a slightly higher tendency for chain restaurants. Cluster 1 also displays higher averages for weekend activity and since cluster 0 is significantly larger than cluster 1, this indicates that regular weekday customers form most of our customer base.

Demographic clustering highlights 3 clusters, with cluster 1 standing out as a dominant group. The smallest one, cluster 2, represents a very aged group, which spends highly on every cuisine type (but especially Asian and "Other") and has an above average tendency to order from chain restaurants. They order mostly during the middle/end of the week after lunchtime. Even though they don't order frequently and don't spend much per order, they make a lot of orders and have been using the app for a long time, which might suggest they are quite loyal, although they don't spend much money per order. Cluster 0 and 1, on the other hand, represent younger individuals but while cluster 0 shows a clear tendency of ordering during weekends and at after-lunch hours, cluster 1 shows an average

behavior of placing orders during weekdays at lunchtime. Cluster 1 has higher expenditure habits and order frequency than cluster 0 but they are not loyal customers. On the other hand, individuals from cluster 2 don't order regularly nor extravagantly, but present very good customer durations.

Finally, in the merged clustering analysis, there is a clear differentiation across the final 4 clusters once behavioral and demographic variables are combined. A visual summary of these clusters can be found in Figure 3.

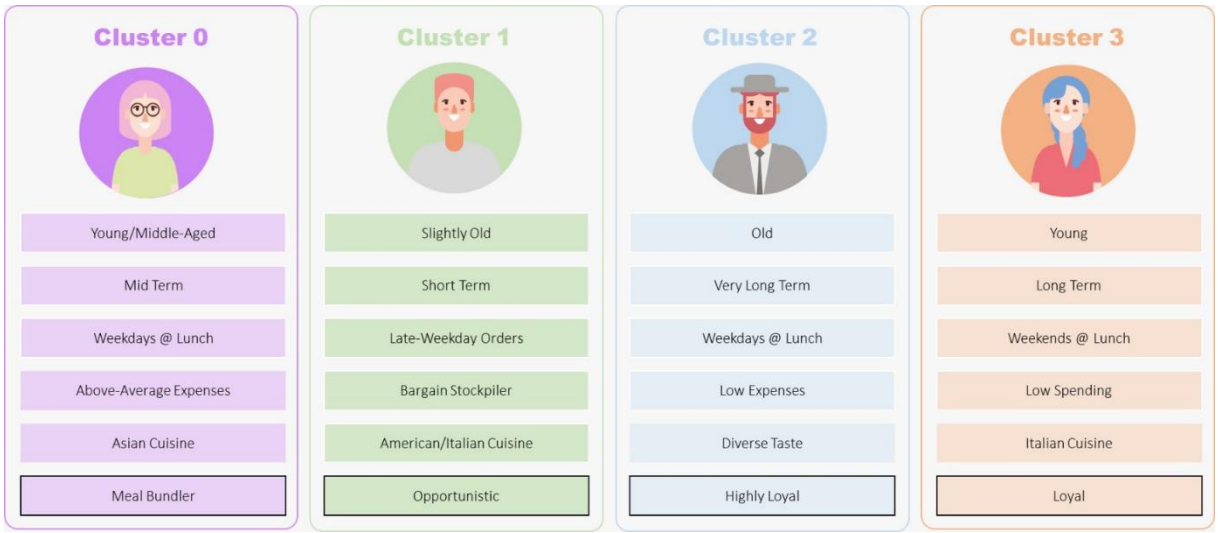

Figure 3 – Customer Segments

For example, **Cluster 0,** the largest group, represents the core customer base, young/middle-aged and orders during the week, at lunchtime, showing a preference for both Asian and "Other" Cuisines. They order less from chain restaurants than other clusters and make above-average expenses. Their low total orders are balanced by the high number of products per order. They represent the average and most frequent consumers, although not the most loyal. They're quality-conscious and not extravagant but are willing to spend more on better meals. Health consciousness might play a certain role, considering their higher spending habits. It's noticeable they appreciate quality and diversity. These customers demonstrate strong purchasing habits and could easily become the most valuable customers if provided with the right loyalty programs.

Cluster 1 has slightly older individuals with very high activity but low economic spending habits, ordering predominantly in the middle/end of the weekdays after lunchtime, mostly from chain restaurants. They tend, though, to not stay for a long time on the app and are highly unloyal, aligning with the "opportunistic" customers identified in the previous behavioral analysis. Cluster 1 represents cost-sensitive customers who engage a lot in short periods of time, probably seeking new promotions constantly, lacking loyalty and depth of interaction. They prioritize cheap options, speed, and practicality, which dominates over health or food quality. This group is most likely to respond to discounts and deals instead of long-term relationship-building strategies.

Cluster 2, our smallest cluster, demonstrates a lower-than-average *Order Frequency* and *Average Spend per Order*, representing aged customers who engage a little bit less frequently in the app and spend less money per order. Although they are not as frequent as the others, they have spent a lot of money overall and they are the most loyal by far. They provide consistent economic value for the business overtime and have a very high retention rate, which highlights their importance since they'll

spend a lot of money overall due to their long *Customer Duration*. They tend to order during the final weekdays at lunchtime, especially from chain restaurants and with a tendency to order from every cuisine type (their tastes are quite diverse, but there's a slight preference for Asian).

Lastly, **Cluster 3** has the lowest order frequency but has one of the longest durations using the service. They tend to order on weekends, after lunchtime, and mostly Italian Cuisines, preferring the familiarity and speed of chains. They don't spend much, which might be an indicator of fast-food lovers and a representation of low-frequency but high-commitment customers, whose behaviors center around leisurely weekend dining, generic cuisine preferences, and convenience-driven choices. Their significant (if average) number of orders makes them a valuable segment despite their less frequent engagement.

6.1. Profiling with RFM

In this section, each cluster is going to be analyzed using the segments and groups created in the RFM analysis. The bar charts below illustrate the composition of each cluster in terms of value segments and economic groups.

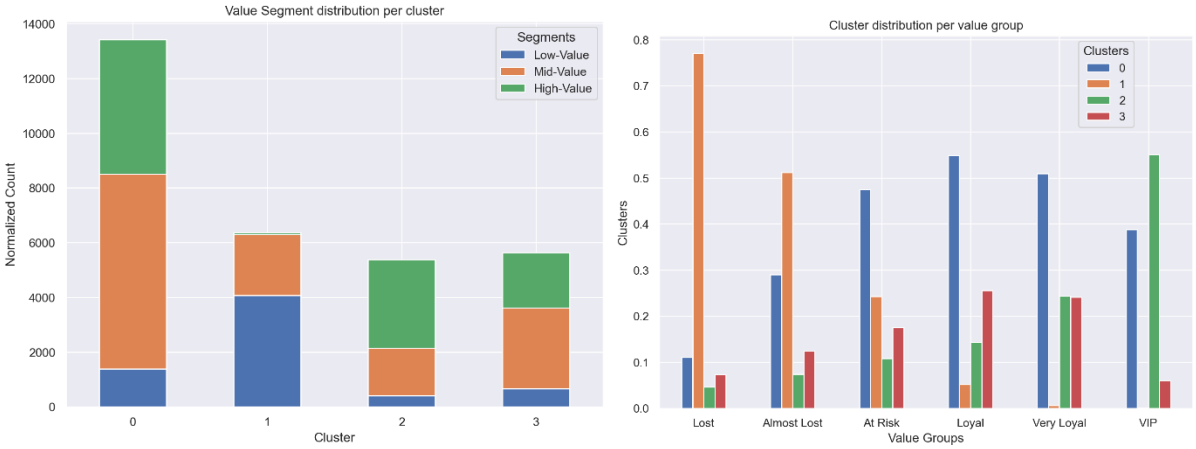

Figure 4 – Value Segment Distribution per Cluster and Figure 5 – Cluster Distribution per Value Group

Cluster 0 has a lot of high and mid-value customers compared to other clusters, indicating that this is a core customer segment. The proportion of mid-value customers is slightly greater than high value and there is a big percentage of "At Risk" customers, which is a little worrying because of the high number of customers that are in this cluster. To address this issue, Cluster 0 could be further engaged through loyalty programs or curated offers emphasizing quality and diversity, appealing to their appreciation for higher-end meals.

Cluster 1 is primarily represented in the low value segment, aligning with the description of these customers as cost-sensitive and opportunistic. The absence of high value customers in this cluster emphasizes their behavior of prioritizing low-cost options and promotions. They lack representation in the loyal categories, with a significant portion falling into "Lost" and "Almost Lost", indicating that they are not tied to the company and are likely to switch based on promotions. This cluster may include customers that use the platform for catering or group orders, given their high order frequency combined with low overall economic spending and engaging duration. To appeal to these customers, the company could offer discounts for bulk orders or highlight catering options.

Despite their low order frequency, **Cluster 2** contains a significant proportion of high-value customers, which can be attributed to their exceptional loyalty and consistent economic contributions over time. These older customers have built a long-standing trust relationship with the company, making them the foundational pillars of our customer base. It is important to notice that, although these customers have a low order frequency, they have the longest duration, many total orders, and they spend the most money overall on any cuisine. Most of these customers belong to the "VIP" and "Very Loyal" categories, highlighting their crucial role in the company's success and their significance in ensuring long-term revenue stability. To retain these customers, the company should offer them exclusive offers, tailored rewards programs, and premium experiences that make them feel valued. Building on their loyalty, providing early access to new features or promotions, and offering benefits like priority customer service or occasional surprises could further solidify their relationship with the brand. Additionally, maintaining open communication channels and seeking feedback can help ensure their needs are consistently met, reinforcing their trust and commitment over time.

Lastly, **Cluster 3** consists of mid-value customers and some high-value users who belong to the "Very Loyal" group. Like Cluster 0, there is slight concern about the higher proportion of mid-value customers compared to high-value ones. This cluster has the lowest order frequency, comparable to that observed in Cluster 2. These issues should be addressed with targeted initiatives, such as personalized engagement strategies, loyalty rewards, and weekend-specific promotions to encourage higher spending and more frequent interaction. Additionally, there should be a focus on converting mid-value customers into high-value ones by offering them premium experiences or incentives tailored to their preferences, such as exclusive deals on their favorite cuisines or enhanced service options.

7. MARKETING STRATEGIES

The clustering analysis provides an understanding of *ABCDEats Inc.*'s customer base, highlighting significant differences in customer behavior and demographics. Our final clustering approach suggests four distinct customer segments – refer to Table 8 for detailed cluster descriptions.

These groups showcase a diverse customer base with different needs, habits, and preferences, which highlights the importance of tailored strategies to optimize retention and engagement. Intrinsically, we propose the creation of 4 targeted marketing programs that satisfy each cluster's unique behavior.

For **Cluster 0**, we suggest the creation of a loyalty program, offering discounts for every X number of orders or exclusive rewards for frequent purchases. This aims to extend their duration and frequency in the app in a simple, yet alluring way. Since they appreciate diversity and quality in their meals, introducing healthy and balanced meal options is also important. Additionally, forming partnerships and giving them early access to new, high-quality restaurants for limited-time offerings can create a sense of exclusiveness that drives engagement. Since it's one of our youngest groups, creating digital campaigns on platforms like Instagram or TikTok to connect with them might be a good investment, featuring exclusive weekend promotions and influencer collaborations that emphasize the benefits of healthy and premium meals.

For **Cluster 1**, the key marketing strategy should revolve around frequent promotions, coupons, and time-limited deals to take advantage of their opportunistic behavior. Budget-friendly meal combo deals will capture their attention and, since they're so active, push notifications are a powerful tool, alerting them to new discounts or flash sales. Moreover, incorporating gamification elements into the

app, like rewards for specific actions or for ordering from certain cuisines, can help increase their cuisine diversity and encourage repeat orders.

Moving to **Cluster 2**, *ABCDEats Inc.* should focus on convenience and simplicity. It's important to tailor the marketing strategy for these older customers. Sending them personalized emails or SMS reminders of future promotions or easy-to-order combos could boost their engagement. Simplifying the app experience by offering pre-selected meals or famous options based on their past choices will make ordering more convenient for them. Since they have "adventurous" tastes and are likely to try something new, it's important that these pop-ups emphasize diversity and show a vast range of cuisine options.

Lastly, for **Cluster 3**, weekend promotions can be an effective way to enhance activity. Special weekend deals, new menu items, or exclusive offers will appeal to their leisurely dining habits. Upselling strategies are very important to increase their spending-per-order. Offering premium upgrades, extras or family sizes can further the profit obtained on their orders.

8. CONCLUSION

The clustering analysis conducted for *ABCDEats Inc*. provided significant insights into the company's diverse customer base, identifying four distinct customer segments based on behavioral and demographic attributes. These clusters presented clear differences in spending patterns, cuisine preferences, and loyalty levels, emphasizing the importance of tailored marketing approaches.

The findings indicate that Cluster 0 represents the core customer base, composed of younger to middle-aged, quality-conscious individuals who place moderate, higher-value orders, primarily on the last weekdays after lunchtime. Cluster 1 consists of opportunistic, cost-conscious customers who order frequently, particularly on weekdays after lunch, with a preference for chain restaurants. Cluster 2 represents older, less frequent customers who spend less per order, with a preference for routine meals from chain restaurants, but with a taste for diverse food. These customers have a surprisingly high customer duration and make consistent economic contributions over time. Finally, Cluster 3 includes low-frequency, low-spending customers who make larger orders, especially during weekends, probably representing families.

The proposed strategies aim to address each cluster's unique characteristics. Loyalty programs, personalized promotions, and in-app engagement are some of the recommendations designed to boost retention, engagement, and revenue. Implementing these strategies will help *ABCDEats Inc.* strengthen customer relationships, positioning itself as a leader in the food delivery market.

BIBLIOGRAPHICAL REFERENCES

Han, J., Kamber, M. 2006, Data Mining – Concepts and Techniques, Morgan Kaufmann, Elsevier Inc.

A. K. Jain, M.N. Murthy and P.J. Flynn, 1999 Data Clustering: A Review, ACM Computing Review.

Provost, F., Fawcett, T. (2013) Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking, O'Reilly Media, ISBN-13: 978-1449361327.

Han, J., Pei, J., & Tong, H. (2023). Data, measurements, and data preprocessing. In Elsevier eBooks (pp. 23–84). https://doi.org/10.1016/b978-0-12-811760-6.00012-6

APPENDIX FIGURES

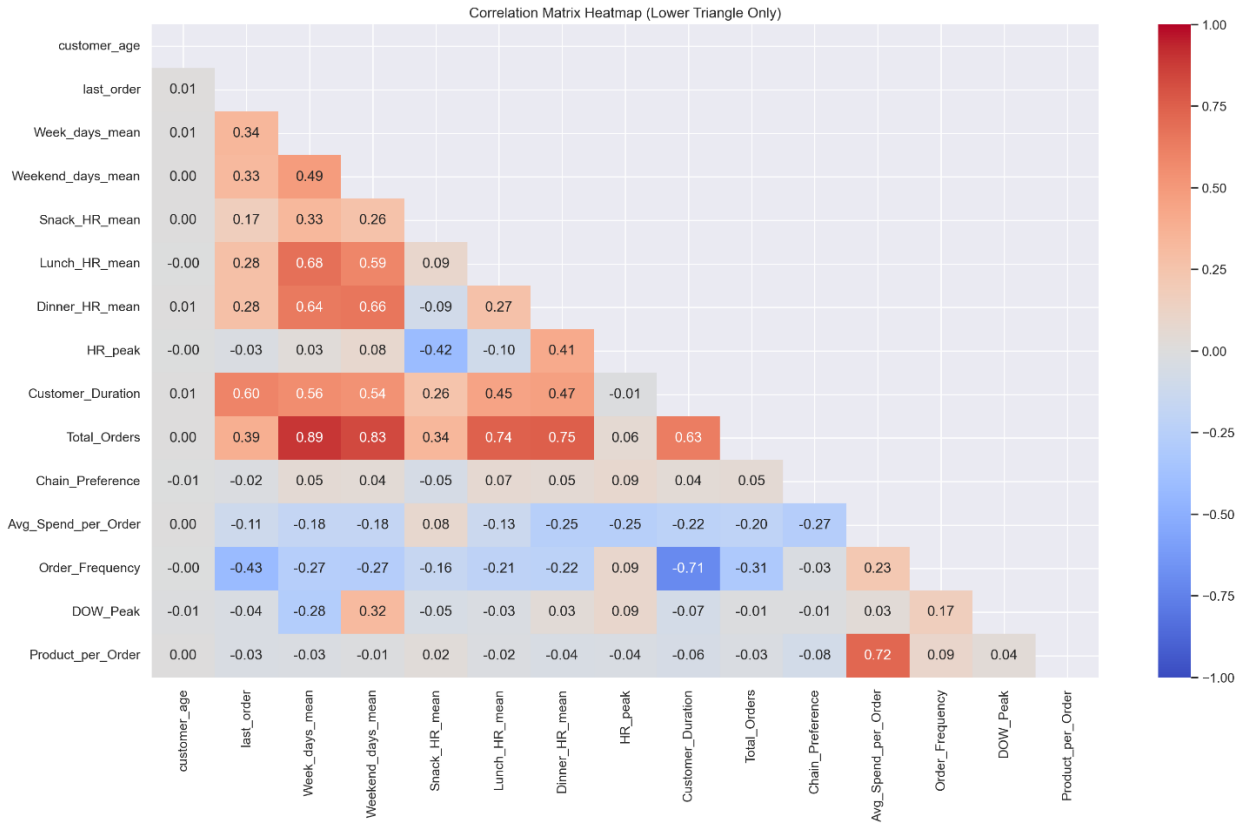

Figure 6 – Correlation Heat Map with All Features

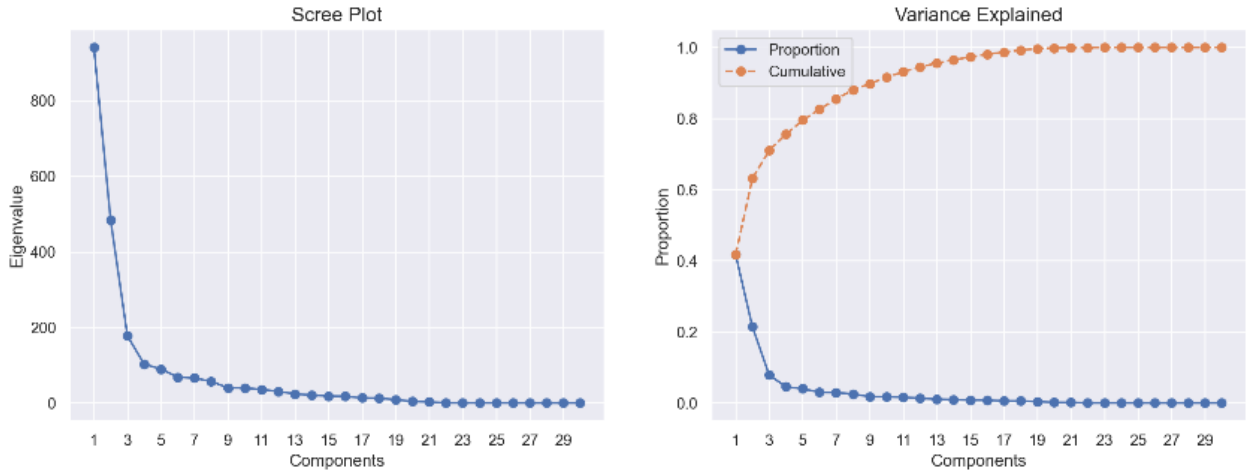

Figure 7 – PCA Analysis

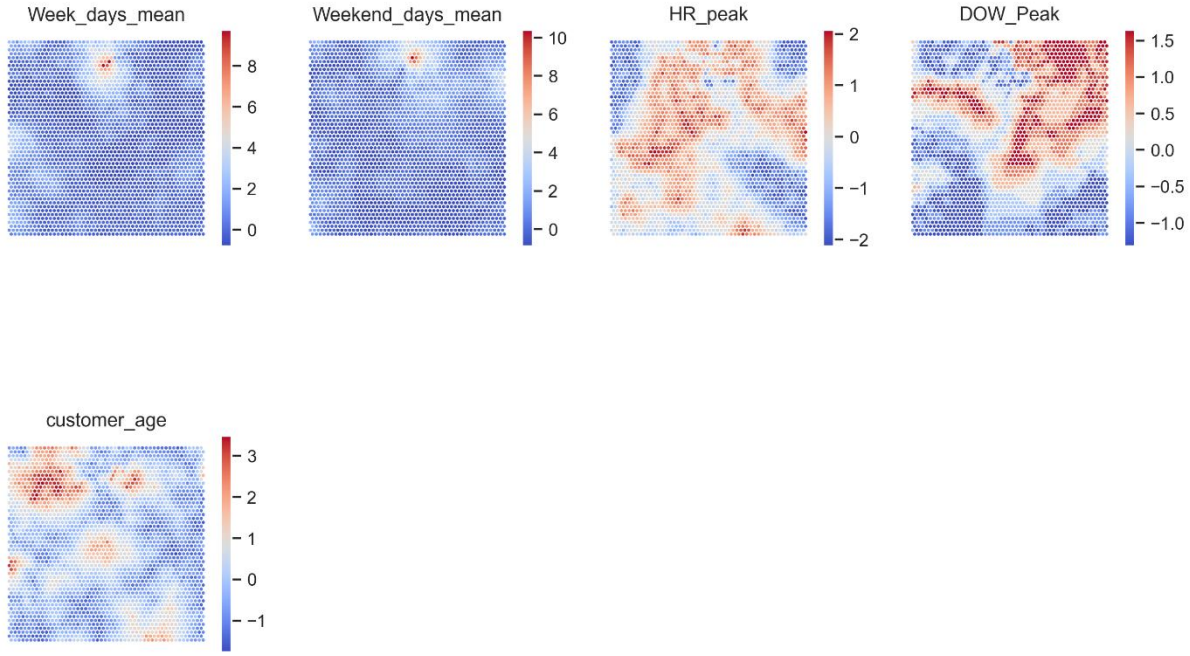

Figure 8: SOM Component Planes (Demographic Perspective)

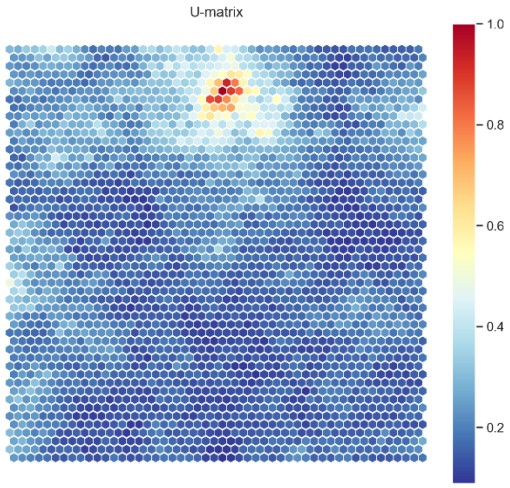

Figure 9: SOM U-Matrix (Demographic Perspective)

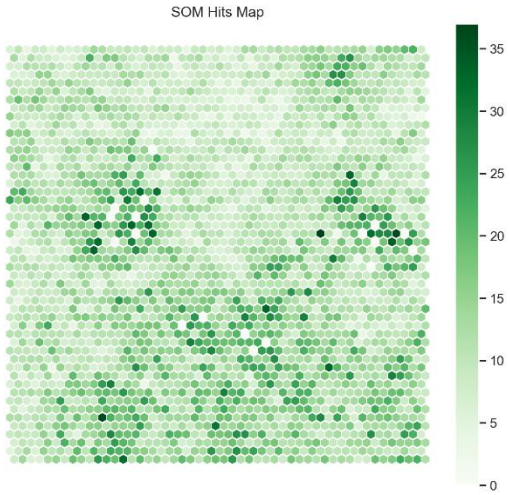

Figure 10: SOM Hits Map (Demographic Perspective)

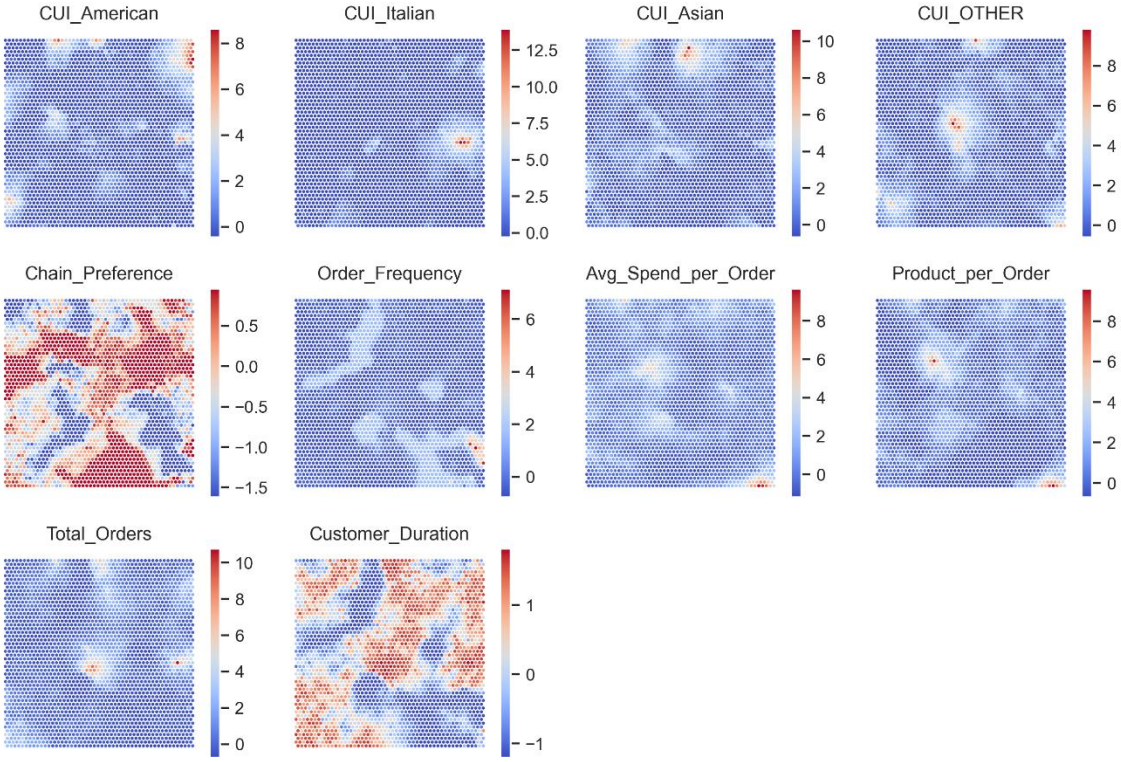

Figure 11: SOM Component Planes (Behavioral Perspective)

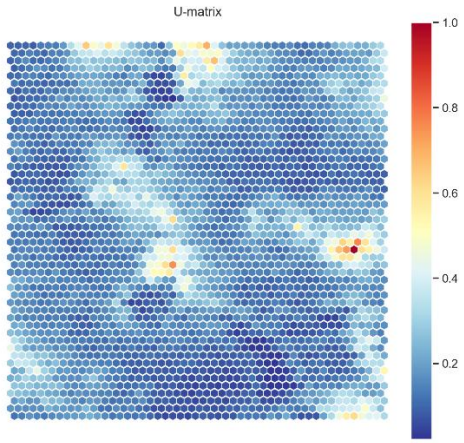

Figure 12: SOM U-Matrix (Behavioral Perspective)

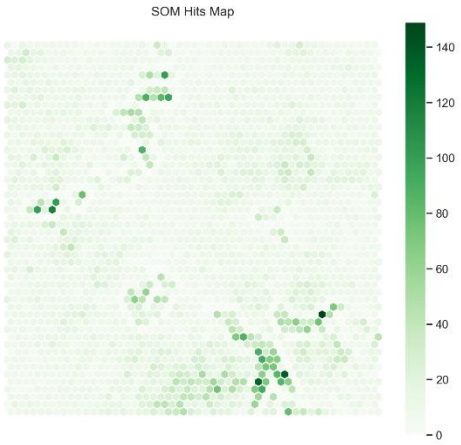

Figure 13: SOM Hits Map (Behavioral Perspective)

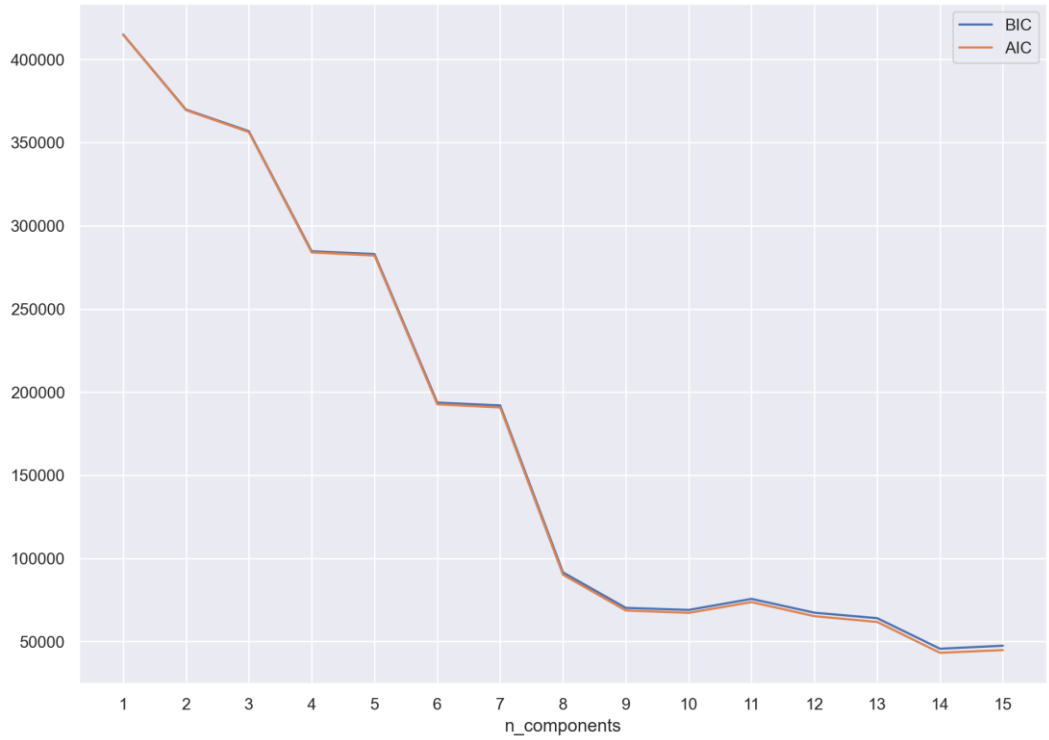

Figure 14 – Gaussian Mixture Model: *Selecting number of components based on AIC and BIC* (Demographic Perspective)

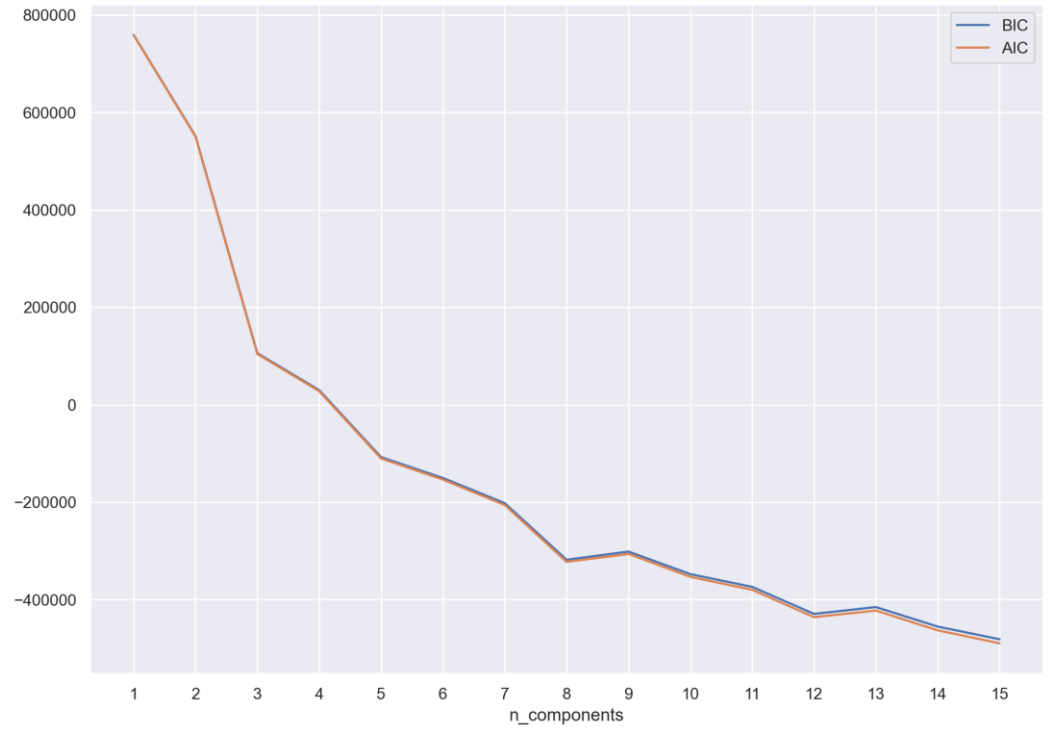

Figure 15 – Gaussian Mixture Model: *Selecting number of components based on AIC and BIC* (Behavioral Perspective)

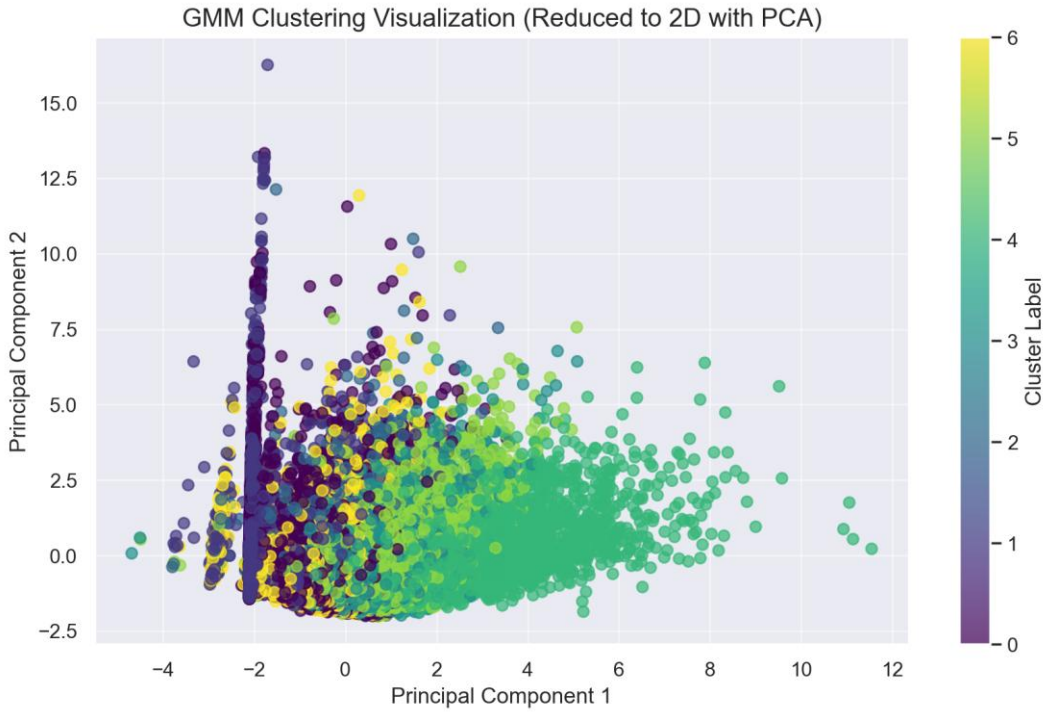

Figure 16 – GMM Clustering Results for the Demographic Perspective

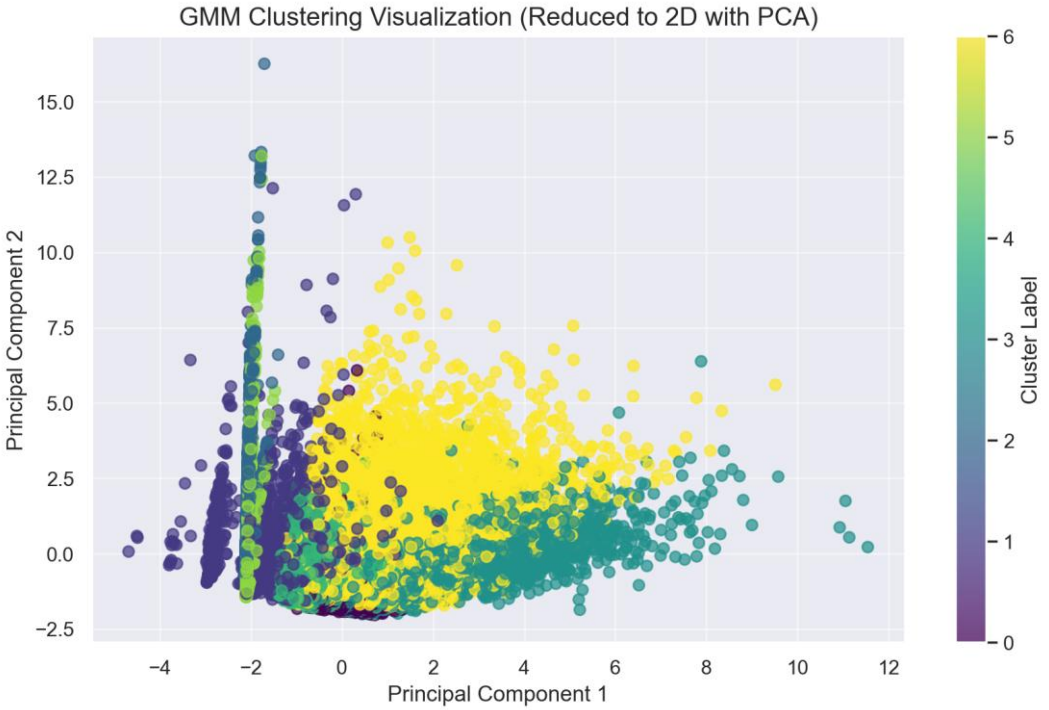

Figure 17 – GMM Clustering Results for the Behavioral Perspective

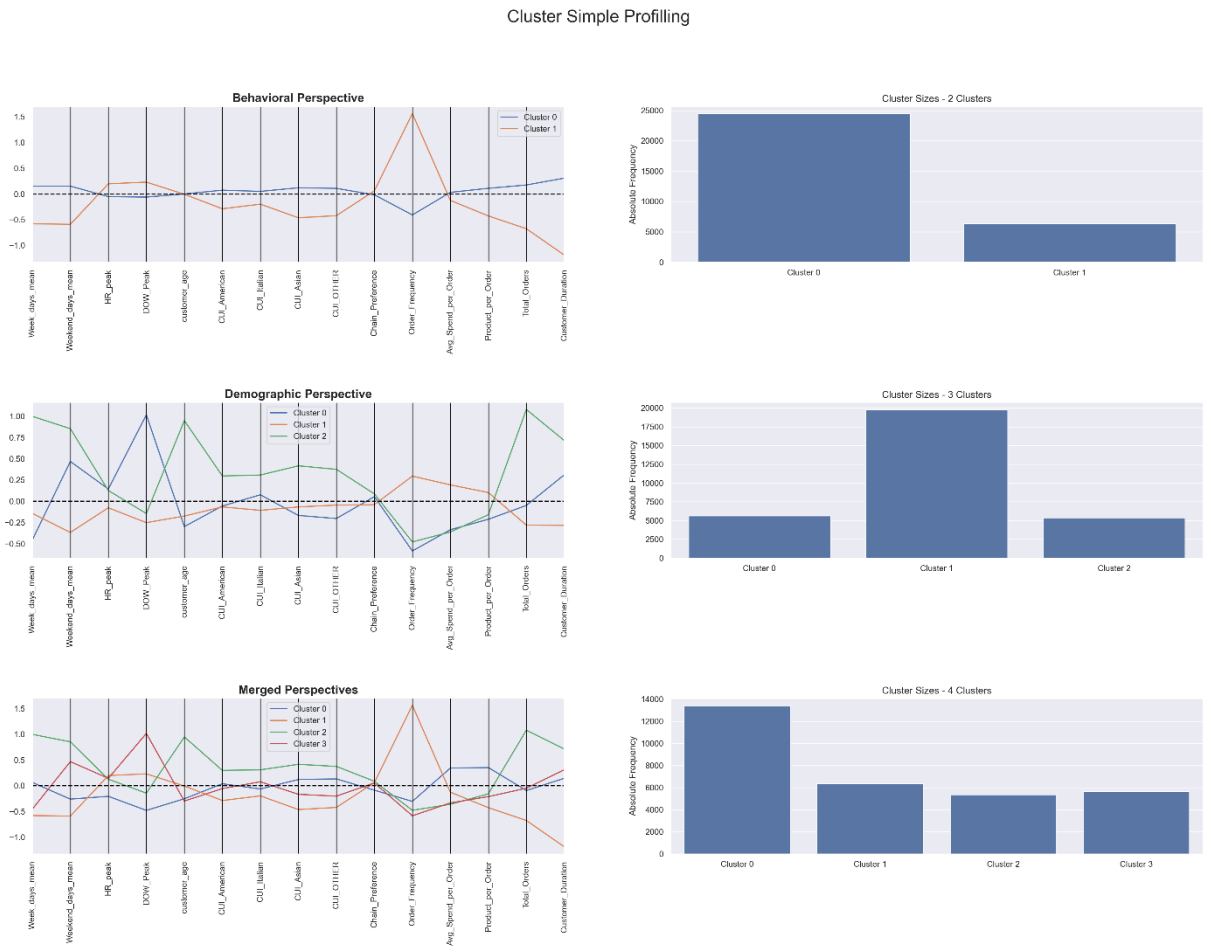

Figure 18 – Cluster Profiling (Manually Merged Solution)

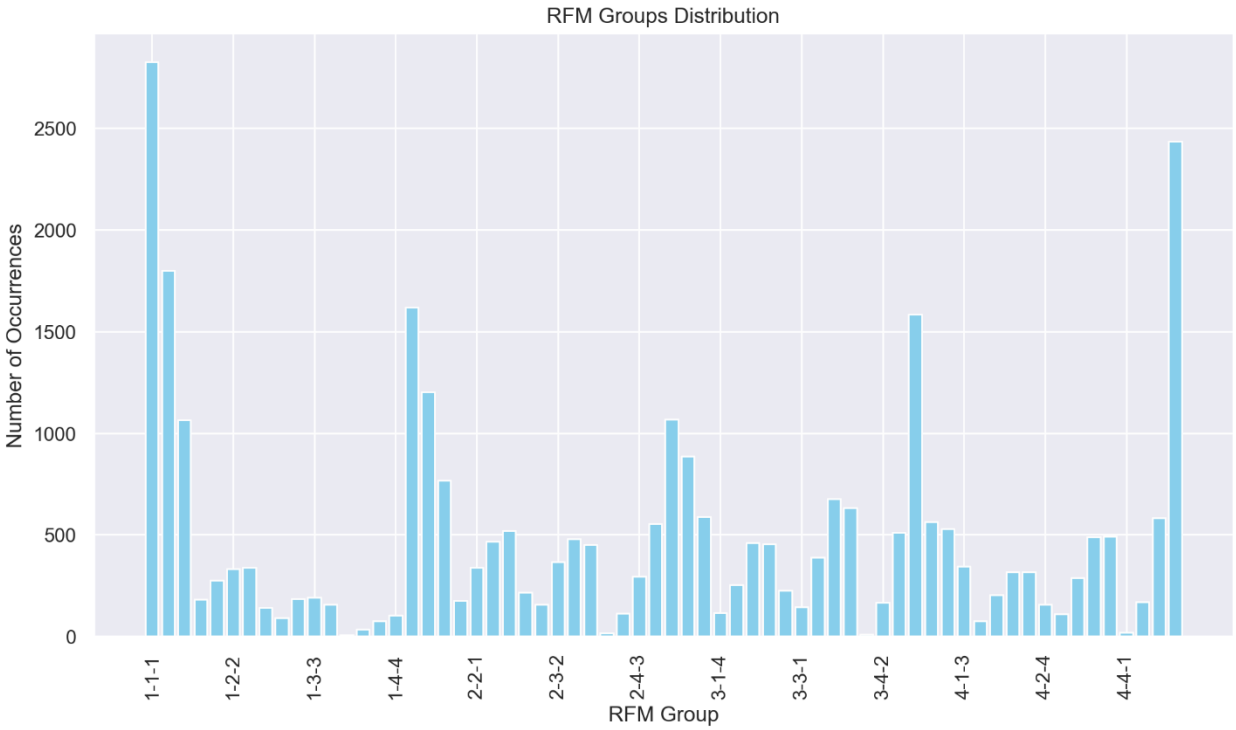

Figure 19 – RFM Groups distribution

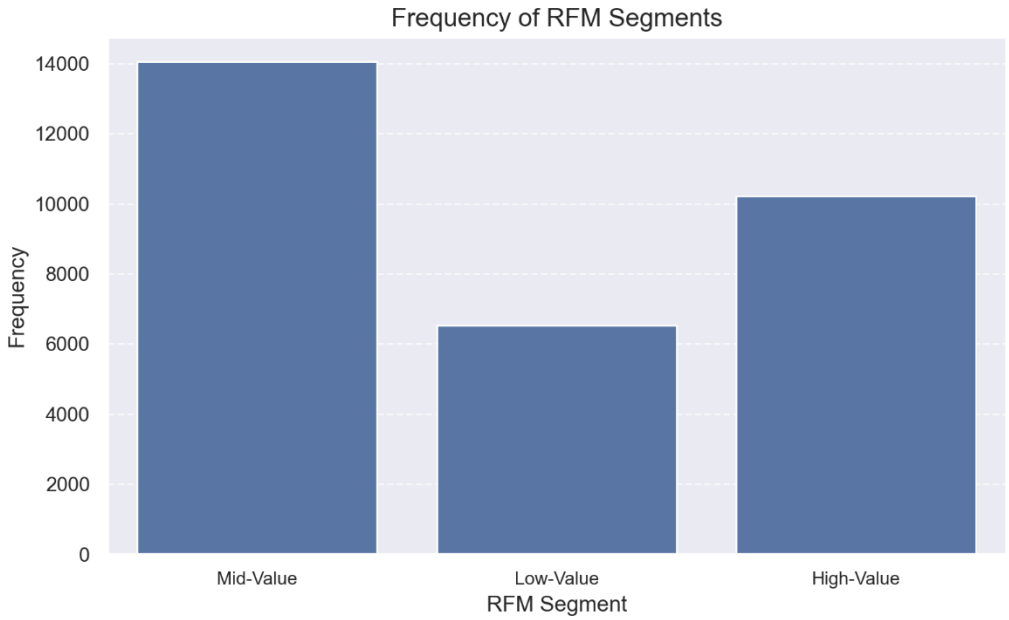

Figure 20 – Customer Value Segments distribution

APPENDIX TABLES

Table 3 - Missing Value Treatment

Feature	Missing Values	Solution
Customer Age	727	Replacing missing values with the median, as it is not as sensitive to outliers than the mean
First Order	106	Dropping values to avoid complicated sensible imputation as it's depended on the Last Order and noise – low impact due to little occurrences
HR_0	1164	Substituting the missing values with calculating the difference in total orders of the sum DOW and HR (as there are no missing values of DOW, the real missing values can be detected in this way)
Customer Region	442	Replacing the values with missing data with the mode.

Table 4 – Threshold Values Defined for Each Feature

Feature	Threshold / Excluded Value	Unit
Customer Age	50	Years
Vendor Count	30	Number of Vendors
Product Count	70	Number of Products
Is Chain	40	Number of Chains
Customer Region	8550 (object, excluded value)	Geographic Region
CUI American	95	Monetary Unit
CUI Asian	200	Monetary Unit
CUI Beverages	120	Monetary Unit
CUI Café	120	Monetary Unit
CUI Chicken Dishes	60	Monetary Unit
CUI Chinese	100	Monetary Unit
CUI Desserts	70	Monetary Unit
CUI Healthy	80	Monetary Unit
CUI Indian	100	Monetary Unit
CUI Italian	150	Monetary Unit
CUI Japanese	105	Monetary Unit
CUI Noodle Dishes	70	Monetary Unit
CUI OTHER	95	Monetary Unit
CUI Street Food / Snacks	160	Monetary Unit
CUI Thai	64	Monetary Unit
DOW 0	11	Number of placed orders
DOW 1	13	Number of placed orders
DOW 3	14	Number of placed orders
DOW 5	15	Number of placed orders
DOW 6	14	Number of placed orders

Table 5 - List of New Features Created

Average Spend per Order	*= Total Spend / Total Orders* Customers spending on average per order
Average Spend per Vendor	*= Total Spend / Vendor Count* Customers spending on average per vendor
Chain Preference	*= Chain restaurant orders / total orders* Proportion of orders that came from chain restaurants
Churn Risk	*= 1 if days since last order > threshold, else 0*. Binary feature that indicates whether a customer is at risk of churn, based on how long it has been since their last order
Cuisine Concentration	*= Spend on the most ordered cuisine / Total Spend* Customers' spending on a particular type of cuisine. Higher concentration may indicate strong preferences, while lower indicates a more diverse taste
Cuisine Diversity	*= number of non-zero CUI_* columns* Distinct types of cuisine the customer has ordered from, indicating the variety of their taste preferences
Customer Duration	*= last order – first order* Number of days since the customer has been registered to its last order
Days Since Last Order	*= max (Last Order) - Last Order* *Days that passed since last order to the end of the Dataset*
DOW Peak	*= DOW_X with the maximum value* Day of the week the customer most frequently orders on. This could identify customers who prefer to order on weekends or weekdays.
Favorite Cuisine	= cuisine with highest spending
Favorite Cuisine Concentration	= *spend on the favorite cuisine / total spend* Concentration of the customer's spending on his favorite cuisine
HR groups	Dividing all hours in specific daytimes
HR Peak	*= HR_X with the maximum value*
Money spent on average	*= sum of all the cuisines per customer / total products* Total expenditure
Order Frequency	*= Total Orders / Customer Duration*
Other Asian Cuisines	*= CUI_Asian – sum of the money spent in Japanese, Chinese, Thai and Indian cuisines* Money spent on other types of Asian cuisines
Product Frequency	*= Customer Duration / product count*
Product Intensity	*= product count / Customer Duration*

	Number of products ordered per active day. A higher number could indicate a high purchase intensity per day.
Repeat Customer	*= 1 if last order – first order > 1, else 0* whether a customer has placed more than one order over a significant period, differentiating between one-time and repeat customers
Spend per Product	*= Total spend / product count* average amount spent per product, which can help identify high-value or budget-conscious customers
Total Orders	*= sum of the DOW_ columns* Total number of orders a customer has placed
Total Spend	*= Sum of all CUI_* columns* Measures the total monetary value a customer has spent over the given period. This helps in customer segmentation based on spending.)
Weekdays Mean	*= Mean of DOW_ from 0 to 3*
Weekend Days Mean	*= Mean of DOW_ from 4 to 6*

Table 6 - Manually Merged Clusters

Demographic Labels / **Behavior Labels**	**0**	**1**	**2**
2	5638	13420	5376
3	0	6352	0

Table 7 - Hierarchically Merged Clusters

Demographic Labels / **Behavior Labels**	**2**	**3**
0	2962.0	0.0
2	0.0	3936.0
3	23886.0	2.0

Table 8 – Cluster Descriptions

Cluster	Description
Cluster 0	Largest group; core customer base. Young/middle-aged, order during the week at lunchtime. Prefers Asian and “Other” Cuisines. Makes above-average expenses with a high number of products per order.
Cluster 1	High activity but low total economic spending. Order on final weekdays/weekends after lunchtime, primarily from chain restaurants. Short app usage duration and highly unloyal, aligns with “opportunistic” customers.
Cluster 2	Smallest cluster; Lower engagement and spending, but highly loyal. Represents older customers who order during weekdays and lunchtime, tends toward chains and Asian, but all in all has a very diverse cuisine taste.
Cluster 3	Low order frequency but good number of total orders and one of the longest app durations. Orders on weekends after lunchtime; prefers Italian and favors chains.

ANNEX

Chat GPT was used throughout this project to shorten text, improve report readability, create cluster visualizations, and improve code.