DATA MINING PROJECT

Master in Data Science and Advanced Analytics

NOVA Information Management School

Universidade Nova de Lisboa

Customer segmentation of ABCDEats

Group 24

Carolina Pinto, 20240494

Iris Moreira, 20240659

Francisco Pontes, 20211583

Maria Margarida Cardoso, 20240493

Fall Semester 2024-2025

TABLE OF CONTENTS

1. INTRODUCTION

Understanding customer behavior, their spending habits and preferences is essential for any business aiming to effectively tailor its strategies. In this report, we analyze the dataset of ABCDEats, a food delivery service, to uncover insights that will inform clustering algorithms and support data-driven customer segmentation decisions. These insights will then be leveraged to develop a personalized marketing strategy for each customer group, maximizing both business profitability, customer loyalty and satisfaction as well as optimizing the operational efficiency of the company.

This report provides an overview of the previously conducted Exploratory Data Analysis to establish context and rationale, followed by the Pre-Processing steps employed to clean the dataset in a consistent and methodical manner. It then examines the various clustering perspectives and techniques explored and concludes with the final clustering solution, alongside business recommendations tailored to each specific customer group, that allows for an effective allocation of the company resources.

2. EDA OVERVIEW

Before proceeding with any data modifications, we conducted a comprehensive review of our previously developed on the Exploratory Data Analysis.

Key insights included the adjustments for unclear regions in *customer_region*, and the dominant use of electronic payments amongst our predominantly young-adult user base. For numerical variables, we identified trends such as customer loyalty to specific vendors, product preferences, low spending patterns, and app usage trends. Time-based variables (*DOW*, *HR*) revealed patterns of peak ordering behaviors both daily and weekly, while cuisine expenditure indicated strong customer segmentation by preference.

Our preliminary findings also highlighted a majority of users who exhibit low frequency and spending behaviors, alongside with a minority of high-value customers identified as our strategic focus. We also recognized that *broader_region* showed to be the strongest bet for segmenting with categorical variables.

As the analysis progressed, we uncovered additional duplicate rows and developed new features essential for shaping our clustering strategy, as outlined in the preprocessing section of this report. All anomalies stated in the original EDA were corrected, and the EDA of the new features can be found in table 1, along with relevant visualization [Figures 1, 2 and 3].

3. PRE-PROCESSING

3.1. Removing duplicates

The original data set has 13 duplicates. When *customer_id* is set as index, we can see that this amount goes to 60 observations, which are immediately dropped, since it is very unlikely that different people have the exact same consumption pattern, especially when it comes to monetary expenditure that takes on a very specific and continuous variable.

3.2. Missing values treatment

With the information gathered with the development of our EDA, we identified several variables with different types of missing values:

Variable	% Missing Values	Treatment
customer_region	1,39%	Identified a similar behavior of '-' with broader region 8, so we assign 8 value to it.
last_promo	52.5%	Replaced with: 'NO PROMOTION'
customer_age	2,28%	Filled in with the median due to the narrow age range of the majority.
first_order	0.33%	Zero imputation*
HR_0	3,65%	Replace missings with the difference between the sum of orders across DOW and the sum of orders across HR.

*All rows with missing values for *first_order* have *last_order* equal to zero. If *last_order* is zero, that means that the last order for that customer was placed in the day the dataset was created, and no other order was placed. We believe that in this case, the value for *first_order* was not updated and the zero imputation is coherent, since *last_order* equal to zero and *first_order* being a missing value happens in the exact same rows.

3.3. Anomalies

Firstly, we drop the rows (18) where *product_count* is less than the total number of orders, since an order needs to have at least one product. By dropping this inconsistency, we also drop the rows where *product_count* is equal to 0.

We then identify rows where *product_count*, *vendor_count* and *total_orders* have 0 values, but there's a positive value for *first_order* and *last_order,* possibly representing customers who didn't complete an order or a data-entry error. These are also dropped since no useful information can be withdrawn from them. By dropping these rows, we no longer have the anomaly associated with *is_chain* (the number of orders from a chain restaurant being higher than the total number of orders). The metadata of *is_chain* is corrected to: "Number of orders the customer has placed from a chain restaurant".

3.4. New Features

Besides the features created during the initial Exploratory Data Analysis, we developed the following additional features as their necessity became evident during the clustering development process:

- **hr_diversity_score:** (if HR_>0, add 1): number of unique hours where the customer placed an order, to understand purchase preferences.
- **dow_diversity_score:** (if DOW_>0, add 1): number of unique days of week where the customer placed an order, to understand weekly purchase preferences.
- **prefered_cuisine:** (Column of Max CUI_): cuisine where the customer spent the most.

3.5. Outliers' removal

To begin the process of outlier treatment, we focused exclusively on numerical variables. Our process combines univariate outlier detection with boxplots and multivariate outlier detection with DBSCAN. Initially, we examine boxplots for each numerical variable individually, using them to identify and iteratively refine a manual filter for extreme outliers. This approach accounts for the fact that some outliers are so extreme that they obscure other potentially problematic values. By iteratively excluding

these extreme points (sparser outliers in the boxplot) and re-evaluating the boxplots, we uncover and address additional outliers that might initially go unnoticed.

We focus our outlier's removal filter on key variables, including *vendor_count*, *product_count*, is_chain, time grouping features, DOW features and newly created features such as total_spending, *total_orders*, *avg_product_cost*, *avg_order_cost, productos_per_vendor* and *cui_diversity_score*. We decided to not include the CUI features in this process since we are not going to use them for clustering, but rather to refine our profiling.

After not finding a visual difference on the boxplots for different scaling methods for all numerical features [Figures 5, 6, 7, 8, 9 and 10], we decided to go for the one less sensitive to the presence of outliers and then performed Principal Component Analysis (PCA). This reduces the complexity of the dataset by reducing the feature space while preserving the most relevant information, since a high number of existing features could worsen the performance of DBSCAN. We chose 8 PCAs, that explain 78% of the variance in our data (above 75%) [Figure 11]. We only do this transformation for the features that we consider relevant for outlier treatment stated above.

For tuning DBSCAN we first made a graph [Figure 12] with the difference of each point to its 20 nearest neighbors to choose the value for "eps", that we state to be between 3 and 5, where there is an "elbow" in the graph. This "elbow" likely represents the transition from points in dense clusters to outliers or sparsely populated areas. We set 12 as the number of "min_samples", because we have a large dataset, and this number, iterating over different values for the "eps" (3, 3.5, 4), leads to different approaches for the outlier removal that look meaningful. "eps"=3 and "min_samples"=12 led to the removal of 447 outliers. After checking the scatter plots of the different PCA components with the outliers highlighted, the boxplots and the descriptive statistics for the key features, we thought this to be the best parameter setting. Extreme values are clearly removed, and even if there are some boarder points apparently incorrectly identified in the scatter plots [Figures 13, 14 and 15], that is probably because of the different dimensions that we cannot represent at the same time.

In the beginning, our aim was to combine the univariate identification of outliers with the DSBSCAN multivariate method, making sure that we were only removing outliers that were highlighted by the 2 methods. However, this approach led to the removal of very few outliers, keeping extreme values that could lead to bias in clustering. As a consequence, we decided to first identify outliers with DBSCAN and then check the boxplots of the relevant features to refine a manual filter to remove outliers considered extreme that were not removed by DBSCAN [Figures 16, 17, 18, 19, 20 and 21]. This final step ensured a clean dataset for further analysis keeping 96% of observations.

After the outlier removal we scale the data again to prevent skewing the mean and inflating standard deviation. The data to pass through the clusters must be scaled without the influence of outliers to prevent distorted cluster.

3.6. Feature Selection

For the clustering stage, we adopted a perspective-based feature selection approach, requiring us to define specific perspectives.

The first perspective combined features related to value and loyalty, informed by domain knowledge and the EDA. Initially, two separate perspectives were planned: one for value (*total_spending*,

avg_product_cost, and *avg_order_cost*) and another for loyalty (*order_span*, *order_frequency*, *first_order, last_order*, *total_orders*, and *product_count*). However, multicollinearity prompted adjustments. In the value perspective, *avg_product_cost* was dropped due to high correlation with *avg_order_cost*. In the loyalty perspective, *order_span* replaced *first_order* and *last_order*, and *product_count* was excluded due to high correlation with *total_orders*. This refinement left two value variables and three loyalty variables, which were merged into a single combined perspective to provide richer clustering features while maintaining coherence. [Figures 22, 23]

Our initial thoughts for the preference perspective were to segment our clients based on how much they have spent on each cuisine and so use the *CUI_* variables. Based on our Exploratory Data Analysis, we decided not to advance with this strategy due to the fact that there are no relevant correlations between these variables. In order to still obtain relevant information related to the preference perspective, we decided to utilize another set of variables (*hr_diversity_score*, *cui_diversity_score*, *dow_diversity_score*, *is_chain*, *vendor_count*, and *products_per_vendor*). However, correlation analysis revealed that *vendor_count* exceeded the 0.8 multicollinearity threshold with three variables and was dropped. *Products_per_vendor* was also removed due to low correlation with other features. Between *dow_diversity_score* and *hr_diversity_score,* we retained *dow_diversity_score* for its lower granularity and alignment with objectives. [Figures 24, 25 and 26]

Lastly, the time perspective included variables such as *late_night*, *early_morning*, *morning*, *midday*, *afternoon_evening*, *night*, *weekday*, and *weekend*. No strong correlations were found, but EDA insights showed that aggregating hourly and day-of-week variables present similar results. To reduce redundancy, weekday and weekend were removed, as their information was reflected in other features. Excluded features will be revisited during cluster profiling. [Figures 27, 28] [Figure 4]

4. GROUPING METHODS

4.1. Cell-Based Segments

A Cell-Based Segment method was carried out with the aim to compare two possible insightful variables, segmenting each into quartiles. The amount of *total_spending* of each customer is usually the most valuable information for businesses, so we used it as a main variable to segment along with others.

4.1.1. Total Spending VS Average product cost

Average_product_cost can be useful to identify customers with the potential to spend more per product. In our business, we can state that customers who spend more per product are likely to have higher total spending overall. Most customers fall into a pattern of low total spending and low average product cost, suggesting that a significant portion of the customer base is contributing less to overall revenue. However, there is also a notable representation of customers in the higher spending segment, both in total spending and per product on average.

Based on our analysis, customers who spend less in total but more per product seem to be newer customers. This is likely why their total spending is not yet high. These customers are important to retain as they tend to buy more expensive products on average, and stimulating this segment can encourage them to become high-value, long-term customers. [Figure 29]

4.1.2. Total Spending VS Order Span

Although there are customers with a shorter span of days in our business who manage to spend more than the "expected" amount, there are also customers that have been purchasing with our business for a longer period but exhibit lower total expenditure. Despite not representing the majority, these customers should be encouraged to engage and spend more. [Figure 30]

4.2. RFM Analysis

For this analysis, the variables *last_order, order_frequency,* and *avg_order_cost* were used, aligning with the main principles of RFM (Recency, Frequency, Monetary) analysis. Each variable was divided into quintiles, assigning scores to customers for each dimension. To categorize the results, thresholds were defined to classify customers from the best category (1st) to the worst category (5th). The scores were assigned using intervals: [5, 4.5[, [4.5, 3.5[, [3.5, 2.5[, [2.5, 1.5[and [1.5,1], representing the highest to the lowest-performing category. [Figure 31]

From the Radar Chart, we observed that the primary distinction between the 1st (best) and 2nd categories is the Monetary score. In terms of Frequency and Recency average scores, these two categories exhibit similar behavior. [Figure 32]
Regarding the other categories, customers in the 3rd category appear to have the potential to become future best customers. Although their Frequency score is significantly lower than that of the 2nd category, their Monetary score is very close to the 2nd. Moreover, the Monetary score of the 2nd category is quite similar to the lower-performing categories (3rd, 4th and 5th), which is not an expected behavior taking their frequency into consideration. Customers in the 4th category have a markedly lower recency score, suggesting that these customers may have been lost or could simply be sporadic users.

All in all, our best customers appear to be those in the 1st and 3rd categories, rather than in the 1st and 2nd as initially expected when defining the thresholds. As a strategy, we could focus on investing in the 3rd category customers to encourage continued high spending and increase loyalty and frequency as they transition into becoming the best customers.

5. CLUSTERING

For Cluster Analysis, various clustering methods were explored for each perspective defined. To build a diverse approach and to not rule out any potential optimal method we analyzed hierarchical, agglomerative (K-Means) and density-based clustering (Gaussian Mixture Model) algorithms for each perspective, as well as a neural network (Self Organizing Map) and combined methods. Having the final solution of the merged perspectives, we then applied a semi-supervised algorithm (Decision Tree) to improve our results and to predict to which cluster the previously removed outliers belong.

5.1. Value and Loyalty Perspective

The first perspective tested was the **value and loyalty perspective**. The expectations were for us to be able to differentiate customers based on whether they consume frequently or not, on how much they spend for each order and product and for their purchasing frequency.
After trying all the clustering methods mentioned above, the one that provided us with the best results for the value perspective was an integrated solution of a Self-Organizing Map with K-means clustering.

To reach satisfactory results, we tested different parameters of the algorithm, as well as different numbers of clusters. The criteria used to decide on these two components were both the lowest scores of the topographic and quantization errors, and the highest R-squared. To decide on the number of clusters of the K-mean to apply to the SOM solution the metrics used were the highest R-squared, silhouette score, and overall visualization, distribution and interpretability of the clustering solutions. We settled for a final number of 4 clusters, which provided us with a good balance of granularity and interpretability, which led to an R-squared score of 0.640 and silhouette score of 0.363. [Figure 33]

Cluster 0 is a small cluster that represents high-value customers, characterized by the highest total spending, total orders, and a long order span. These customers are frequent and consistent users who significantly contribute to overall revenue overtime and not necessarily with highly valuable orders.

Cluster 1 represents customers with the highest average order cost but lower total orders and a shorter order span. These customers tend to make fewer, higher-value purchases, contributing moderately to total spending. This group is probably the one previously identified on the RFM analysis, so these are probably new customers with a high spending potential.

Cluster 2 is the largest cluster, and it represents customers with low *total_spending*, *total_orders,* and a short *order_span,* probably indicating the customer with single day activity behavior, previously mentioned. The low total orders mainly indicate they are low-engagement users, who also contribute the least to revenue, due to their low *average_order_cost*. This type of pattern is consistent with previous findings both in the EDA and RFM analysis.

Cluster 3 represents customers with low *total_spending* and *average_order_cost*. The difference between these customers and cluster 2 customers is that cluster 3 customers have a higher *order_span* and *total_orders*, which indicates that while they have been active on our app for some time they do not contribute significantly to overall revenue, placing very low-valued orders.

5.2. Preference Perspective

The *cui_diversity_score*, *dow_diversity_score* and *is_chain* features aim to analyze customer behavior, focusing on whether they are spontaneous buyers, or if they follow a consistent consumption pattern by favoring specific cuisines, vendors, or days of the week.
Among the various algorithms explored, the best clustering solution for this perspective was achieved using the K-means algorithm applied to the results from the Self-Organizing Map (SOM). The same evaluation criteria as previously mentioned were used, leading to the selection of 3 clusters. This configuration provides a R-squared value of 0.657 and a silhouette score of 0.474. [Figure 34]

Cluster 0 is the largest cluster, including more than 20,000 customers. These customers have the lowest scores for cuisine diversity, chain preference, and day-of-week diversity. This suggests that most of our customers have repetitive habits, preferring a narrow range of cuisines and ordering primarily from non-chain vendors on specific days.

Cluster 1 represents a moderate-sized group of customers with average scores across all metrics. These customers show some variety in cuisine choices and ordering days, and do not have a high preference for chain vendors. They exhibit stable and predictable ordering behaviors, what makes them a good target group for experimenting with marketing strategies.

Cluster 2 is the smallest cluster, and it contains clients with the highest scores across all metrics: cuisine diversity, chain preference, and day-of-week diversity. These customers have wide and varied ordering habits, with a strong preference in chain restaurants, and they order on diverse days of the week. This group represents flexible and adventurous consumers, who value the daily availability of our service, and the diverse catalog of cuisines offered.

5.3. Time Perspective

After testing all the different clustering methods for the time perspective and even trying numerous different combinations of variables and aggregations for the perspective, we are not completely satisfied with any solution outputted by the clustering methods.
Therefore, we will not include this perspective in our final clustering solution and move on with the perspectives that were already assessed.**Error! Reference source not found.**

6. FINAL CLUSTERING SOLUTION

When merging the two clustering perspectives previously mentioned, we opted for using the hierarchical clustering, in order to fuse clusters based on the distances that separate them. [Figure 35]
We used the dendrogram to determine the optimal number of clusters, which was set to 4. We then tried to improve our results by applying the Decision Tree Algorithm to this final solution, training it with 20% of observations, and when applied the algorithm to the entire dataset our results improved, raising the silhouette score from 0.116 to 0,29 and maintaining the R-squared value of 0,488.
According to the Decision Tree, the most importance features for cluster prediction are *total_orders* (0.42), *avg_order_cost* (0.40), *is_chain*(0.09), *total_spending*(0.09), which is valuable outcome since it gives meaning to key insights, that businesses rely on to succeed, including loyalty, monetary, and preference features.
We then proceeded to use the Decision Tree to predict in which clusters the outliers previously removed during the pre-processing stage would fall and assigned them to those clusters. Most outliers were assigned to cluster 1 (1016 customers). The remaining, that represent a small share, were assigned to cluster 2 (220 customers) and to cluster 3 (14 customers). This indicates that most outliers represent valuable customers, that we want to retain.

Our final clusters are characterized by the following behaviours: [Figure 36]
Cluster 0 corresponds to our **low-spending, less enganged and sporadic customers**. This is the largest cluster, containing 22755 customers, which is a large amount. They have been users for some time, as indicated by the value in *order_span*, but they present low spending and low engagement patterns. When placing an order, they don't spend much neither per order nor per product. These customers are equally spread across the 3 different broader regions and half of them used a promotion in the three-month period of data collection, which is a good indicator of how to interact with them. Their diversity scores are also very little, which is likely a direct consequence of their little engagement.

We decided to divide cluster 0 into three subclusters based on broader regions, both due to the high segmentation verified and to solve the problem of the high imbalance that the actual cluster solution provides. As identified during the EDA and confirmed by the visualizations in Figures 48, 49, 50 and 51, these regions exhibit distinct cuisine preference patterns. Given that customers in cluster 0 have very low cuisine diversity scores, their preferred cuisine feature strongly reflects their overall preferences.

The other clusters also revealed a segmentation by broader region; however, due to their smaller number of customers and more specific spending patterns, this division would not provide as much insight or relevance for ABCDEats. In fact, segmentation by broader region is very similar in all clusters, this is showed in an example [Figure 52]. Further visualizations supporting this statement are in the Merge_Profile notebook, in the profile section.

This new segmentation into three subclusters will enable ABCDEats to develop targeted marketing strategies, tailored to the specific preferences of these customers, hopefully increasing their engagement with the app.

Customers in **subcluster 0.2** (26.15% of the dataset), have a lot of diversity when it comes to their preferred cuisine type. Most of them have as their preferred cuisine *CUI_OTHER* (18.2% of customers), followed by *CUI_Chicken_Dishes* (10.4% of customers) and *CUI_America* (10% of customers). All cuisine categories have customers who prefer them, highlighting the diversity within this subcluster. Most customers in this subcluster order in the morning, afternoon and evening, not doing it in afterhours [Figure 53].
Customers in **subcluster 0.4** (25.0% of the dataset), are less diverse compared to the previously described subcluster regarding their preferred cuisine type. Certain cuisine types, such as *CUI_Beverages, CUI_Desserts*, and *CUI_Noodle_Dishes*, are not ordered by customers in this subcluster, so marketing strategies should be aligned with that. 27.3% of the customers prefer *CUI_Italian*, 19.5% *CUI_American* and 12.2% *CUI_Asian*. Most customers in this subcluster order in the afternoon and in the evening [Figure 53].
S**ubcluster 0.8** (20.69% of the dataset), is the least diverse in terms of cuisine preference. Customers order mostly from *CUI_Asian*, with 45.2% identifying it as their preferred cuisine, followed by *CUI_American* and *CUI_Street Food / Snacks. CUI_Cafe, CUI_Chicken Dishes*, *CUI_Indian, CUI_Italian, CUI_Noodle Dishes* and *CUI_Thai* do not have any customers ordering from them. Most customers in this subcluster place orders during afterhours and in the morning [Figure 53].
The engagement and spending patterns in the subclusters are the same as the ones described in cluster 0, as expected.

Cluster 1 corresponds to our **loyal and engaged customers**, that have been users for a long period of time (11.46% of the dataset). They order frequently and are the most diverse customers when it comes to the type of cuisines ordered, to the days of the week and the hours of the day in which they order. They do not spend much per order, so they are not our highest spending customers, but are the ones with the most engagement in the app. Their value is built through loyalty and consistency over time and not so much through large purchases. These customers are mostly based on broader region 2, while some of them belong to broader region 4. They tend to prefer American, Asian and Other cuisine types, while also presenting a preference for chain restaurants, what makes sense since this type of cuisine is typically cheaper, and which also explains the low value of *average_order_cost*. These customers present a preference for ordering in the morning and afternoon/evening, being loyal to specific ordering hours. They also have a consistent behavior throughout the week, without any major peaks in specific days.

Cluster 2 represents our **high-spending** customers (8.86% of the dataset). They place fewer but more expensive orders when compared with cluster 1. They are engaged customers, that are active in the app for a significant time period although their diversity isn't influenced by this, since they are less diverse when it comes to the cuisine type than customers from cluster 1. Nevertheless, they are still customers who value diversity in the offered cuisines and in the daily availability of our services.

These customers are mostly based on region 8 and 4, and they tend to prefer American and Asian cuisine type (50% of the customers in this cluster ordered at least once Asian and American cuisine types) which aligns with the general trend of the entire dataset. These customers also present a preference for ordering in the morning and afternoon/evening, being loyal to specific ordering hours.

Cluster 3 represents our **low-span** but **potencially great** customers with **very high *average_order_cost*** (7.82% of the dataset). These customers are not active in the app, as they have a very low order span, but they represent an important segment since that even though they were not using our business for a long period, they still managed to have a high total spending due to their elevated expenditure per order and per product. This means that with the right strategy they can be retained and transformed into our best segment.
Most customers from this cluster are based on broader region 8 and have a significant percentage ordering at dawn (36%), which means that while they have not made a lot of orders, it is still markedly accentuated a pattern of specific preferences: they make use of the promotions offered and prefer Asian and Street Food / Snacks cuisine type, not ordering much from chain vendors and with the majority of them using promotions. [Figures 38, 39 , 40 , 41 , 42 , 43 , 44 , 45 , 46, 47]

This leaves us with a 6-cluster solution, much more balanced than the previous one.

7. BUSINESS APPLICATIONS

Cluster 0: As a general application for all the observations of the cluster, we suggest making a promotion-based approach, due to the fact that the clients have sporadic and budget-friendly purchases. Since around half of the cluster has used a promotion on the few orders they have made, we believe that this segment can be more incentivized to engage with ABCDEats' services when having more promotional benefits.

For cluster **0.2**, based on the fact that the total spending per cuisine is not significantly high on any specific one and most customers are very loyal to their respective preferred cuisine, we believe that whenever a promotion is available for a certain cuisine, the clients that have it has their favorite one are notified. Since this is our most segmented cluster, we believe that this can be a great group for experimenting with marketing campaigns, to see how they would react.

In regards of cluster **0.4**, since they thrive on American and Italian cuisines, our suggestion is to emphasize on these interests, and so, we suggest the implementation of more partnerships with American and Italian restaurants, to be more responsive to the high demand on these cuisines for region 4. Having higher supply for these departments and combining the overall cluster 0 strategy to this plan will likely increase engagement.

For **cluster 0.8**, and similarly to cluster **0.4**, the preference is visible for a certain type of cuisine. In this case, the Asian one. The specified location-based strategy is similar, as we believe that by adding more Asian restaurants to the options to order of the region 8 customers, and combining that with promotions for this cuisine, the overall engagement of these observations with ABCDEats can improve largely.

The 3 subclusters derived from the initial cluster 0 should have push notifications and email campaigns scheduled to coincide with the times when customers within each subcluster are most likely to order, as outlined in the final cluster solution description.

Cluster 1: The goal for this segment is to encourage continued use of our app and enhance their loyalty. To achieve this, ABCDEats should develop exclusive loyalty rewards for frequent orders, such as offering a free meal after a certain number of orders. Another potential strategy is to promote their favorite cuisines in region 2, as previously identified, by offering time-limited promotions and chain-specific deals, given their preference for this type of restaurant. To align with their engagement patterns, our strategy should focus on their peak activity hours (typically morning, afternoon, and evening), rather than specific days of the week. Sending push notifications during these time periods can reinforce their habitual usage.

Cluster 2: The difference between this cluster and Cluster 1 lies in the propensity to diversify the types of cuisines, which is much smaller in Cluster 2. While the *dow_diversity_score* is not significantly lower than Cluster 1, there is a slight difference. Therefore, our goal is to encourage these loyal and high-spending customers to try new cuisines in hopes of increasing their order frequency. One possible approach is to establish a gamification system. This system could allocate bonus points for orders in different cuisines and restaurants, which, upon reaching a certain threshold, could be redeemed for discount vouchers on their preferred cuisines or free add-ons or beverages with their orders.
For both Cluster 1 and Cluster 2, a premium membership plan could be developed, as high-spending customers typically become less sensitive to price changes as their spending increases. The benefits of this plan could include free deliveries for orders exceeding a specific monetary amount, priority deliveries and exclusive promotions and discounts, in hopes that increasing promotion usage will motivate them to continue their overall engagement.

Cluster 3: These customers are potential high-spenders, and our goal is to retain them and transfer them to our best segments (both cluster 1 and 2). We can leverage their promotion usage by designing personalized promotions such as first or second order bonus or rewarding them for a specific number of orders within a timeframe. The rewards can be set up in an escalating way, for example for the first order they get a free drink, for the second a dessert and for the third an extra main dish. To turn them into loyal customers we should introduce incentives such as bonus points for each order placed. These bonus points could then be redeemed in reduced delivery fees for next orders in their preferred cuisines during time-limited offers in the dawn hours.

8. CONCLUSION

ABCDEats has a young customer base whose spending behaviors are influenced by their duration as customers, the region they are based in, and their propensities for flexibility and diversity.
Throughout the project, the original dataset was adjusted and modified to meet the conditions required for effective clustering segmentation, based on perspectives deemed appropriate for the context. The final solution was achieved by merging the perspectives that provided the most distinctive insights. In profiling and describing the clusters with unused categorical variables, the largest cluster was further divided into three smaller ones. This refinement resulted in a customer segmentation of six clusters, offering our app a comprehensive framework for understanding its audience.

ABCDEats' customers can be segmented into: **loyal customers** (who have been clients for a long time, but are not very high spenders), **high-spending customers** (who bring the most value for the business, even though they are only few observations), **low-span but potentially great customers** (have not been customers for a long period of time, but have brought important value every time they have

ordered), and finally **low-spending and sporadic customers** (cost-effective, and do not bring much value to the business; unfortunately, they are the most frequent case). After segmenting the customers, we defined business strategies to apply to each segment of customers.

Limitations of our work include challenges related to interpreting ambiguous metadata, which required us to make certain assumptions and decisions about the data and its handling. The outlier treatment was an iterative process that could be further refined, for instance, by exploring additional dimensions or experimenting with alternative tuning parameters or algorithms for identification.
We also encountered an imbalance in the final clustering solution, which may pose challenges for ABCDEats due to the reduced granularity in the largest cluster. Additionally, the t-SNE visualization [Figure 54] of the final solution prior to the split using the categorical variables, lacked clearly identifiable segmentations, suggesting that our dimensionality reduction technique may have overlooked significant aspects of the data that cannot be captured in a two-dimensional representation. This highlights an aspect for potential improvement.

BIBLIOGRAPHICAL REFERENCES

Pandas 2.2.3 documentation. Link

NumPy Documentation. Link

Matplotlib 3.9.2 documentation. Link

Seaborn: statistical data visualization. Seaborn 0.13.2 documentation. Link

PennState, Eberly College of Science. STAT 200. Identifying Outliers: IQR Method. Link

Scikit-learn: effects of feature scaling. Scikit-learn documentation. Link

Frost, J.; Missing Data Overview: Types, Implications & Handling. Statistics By Jim. Link

Han J, Kamber M, Pei J. Data Mining: Concepts and Techniques. 3rd ed. Morgan Kaufmann; 2011. Link

Pontejos, F.; GitHub of Farina Pontejos, Data Mining 24/25, lab03 Data Exploration. Link

Pontejos, F.; GitHub of Farina Pontejos, Data Mining 24/25, lab 04 Data Visualization. Link

Pontejos, F.; GitHub of Farina Pontejos, Data Mining 24/25, lab 06 Data Preprocesss. Link

Pontejos, F.; GitHub of Farina Pontejos, Data Mining 24/25, lab07 Data Preprocess2. Link

Pontejos, F.; GitHub of Farina Pontejos, Data Mining 24/25, lab08 Data Preprocess3. Link

Pontejos, F.; GitHub of Farina Pontejos, Data Mining 24/25, lab 09 Hierarchical Clustering. Link

Pontejos, F.; GitHub of Farina Pontejos, Data Mining 24/25, lab 10 KMeans Clustering.. Link

Pontejos, F.; GitHub of Farina Pontejos, Data Mining 24/25, lab 11 Self Organizing Maps. Link

Pontejos, F.; GitHub of Farina Pontejos, Data Mining 24/25, lab 12 Density Clustering. Link

Pontejos, F.; GitHub of Farina Pontejos, Data Mining 23/24, lab 13 Cluster Analysis. Link

ChatGPT (This AI tool was used at times to elaborate some code, this usage is mentioned at the respective code in the Jupyter Notebook.)

APPENDIX

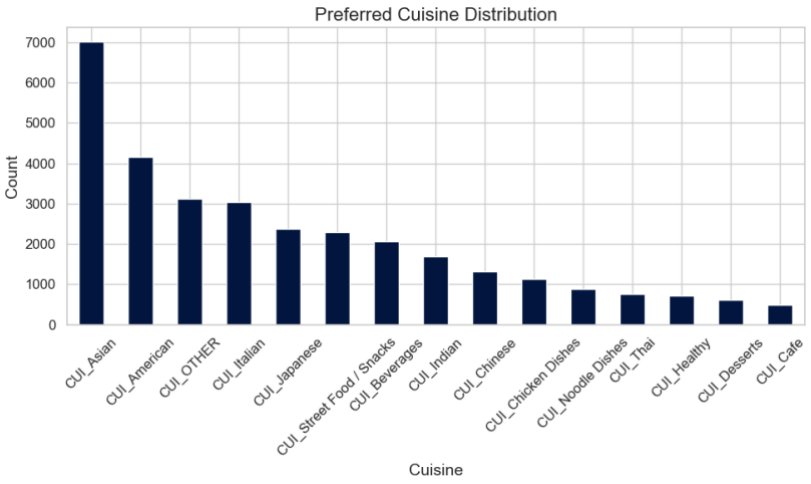

Figure 1 - Distribution of Preferred Cuisine

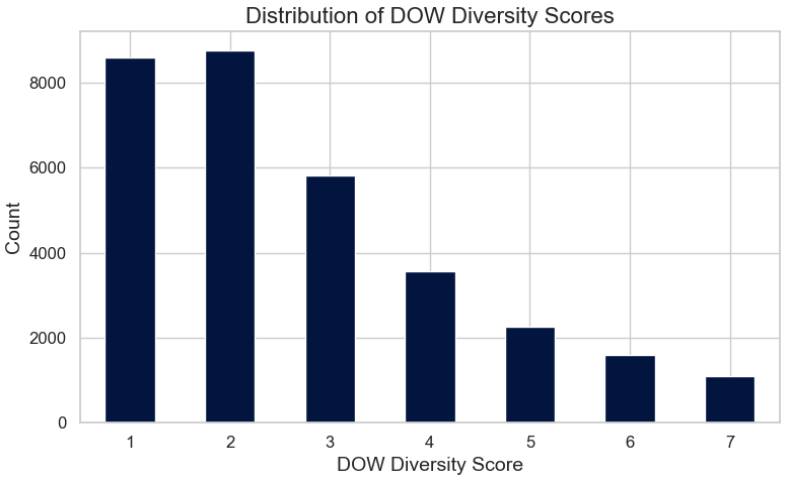

Figure 2 - Distribution of DOW Diversity Score

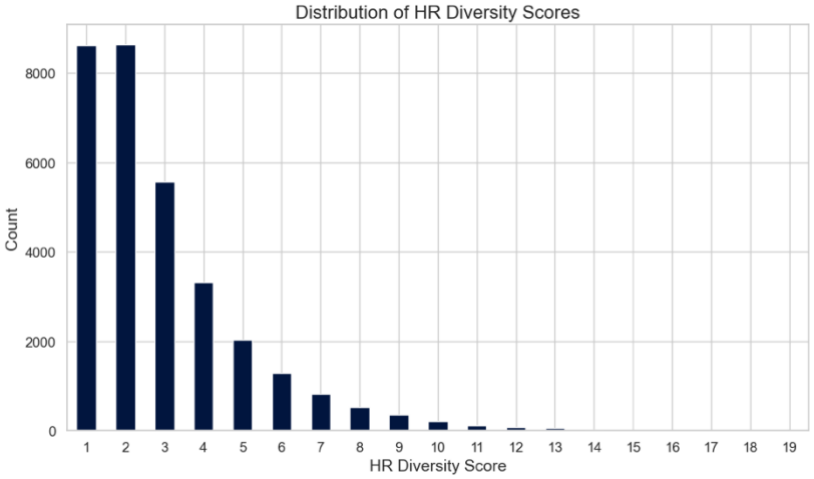

Figure 3 - Distribution of HR Diversity Score

New Variable	Summarized Exploratory Data Analysis
Late_night	Few orders identified in this set of hours (Mean is 0.39, 75% quantile is 0), and the maximum ordered by an individual is 24, which is defenitely an outlier.
early_morning	Less average number of orders than in late_night (0.36), 75% quantile is also 0. Maximum ordered by an individual is 66, which is the most across all time periods.
morning	Fair number of average orders placed (1.26). Median is 1, and the maximum ordered by an individual is 49, which is clearly an outlier.
midday	Less mean orders than morning on this period (0.73), Median is 0, and the maximum ordered by an individual is 37.
afternoon_evening	Peak set of hours (Mean of 1.34), Median is 1, and the maximum ordered by an individual is 49
night	Set of hours where our database orders less on average (0.31). 75% quantile is 0. The maximum ordered by an individual is 43.
weekday	Average of 3.03 orders per customers on weekdays. Median of 2. The maximum ordered by an individual on weekdays is 75.
weekend	Average of 1.36 orders per customers on the weekend. Median of 1. The maximum ordered by an individual on weekends is 29
total_spending	Average of 38.48 monetary units spent. Median of 24.2. The maximum spent by an individual is 1418.33
total_orders	Majority of customers place between 1 and 5 orders, as indicated by the 75th percentile value of 5.
order_frequency	A typical customer places an order roughly every 7 days. However, some customers have much higher frequencies, such as 44.5. For this metric, a lower value represents a customer that, theoretically, places more orders.
order_span	Average of 35.47 days between first and last orders. Maximum is 90, which is how long the data has been retrieved for.
avg_product_cost	The mean expenditure of a product for a customer is around 7.6 monetary units. The maximum average money spent for a product by a single customer is 24.39, which is most likely tied to orders of more expensive restaurants and cuisines.
avg_order_cost	The mean expenditure per order for a customer is 10.31 monetary units. The maximum average amount spent on orders by a single customer is 104.32. This maximum probably represents orders from numerous families.
products_per_vendor	Customers typically purchase 1-2 products per vendor, with occasional extreme cases of up to 70 products, which are clearly outliers.
cui_diversity_score	Most customers have ordered from 2-3 different cuisines, with a maximum of 13 cuisine types. The maximum possible would be 15.

dow_diversity_score	Customers generally place orders on 2-4 different days of the week, with a maximum of all of at least the days of the week.
hr_diversity_score	Orders are placed during 2-4 distinct hours on average, with some customers placing orders across up to 19 hours.
used_promo	A little over half of the customers (16 667 customers) did not use a promotion, while the rest did.
broader_region	The most frequent broader region of the customers (10 741) is 2.
preferred_cuisine	The most preferred cuisine is CUI_Asian, being the favorite by 7,009 customers out of 15 unique cuisine options.

Table 1 - Summary EDA of New Features

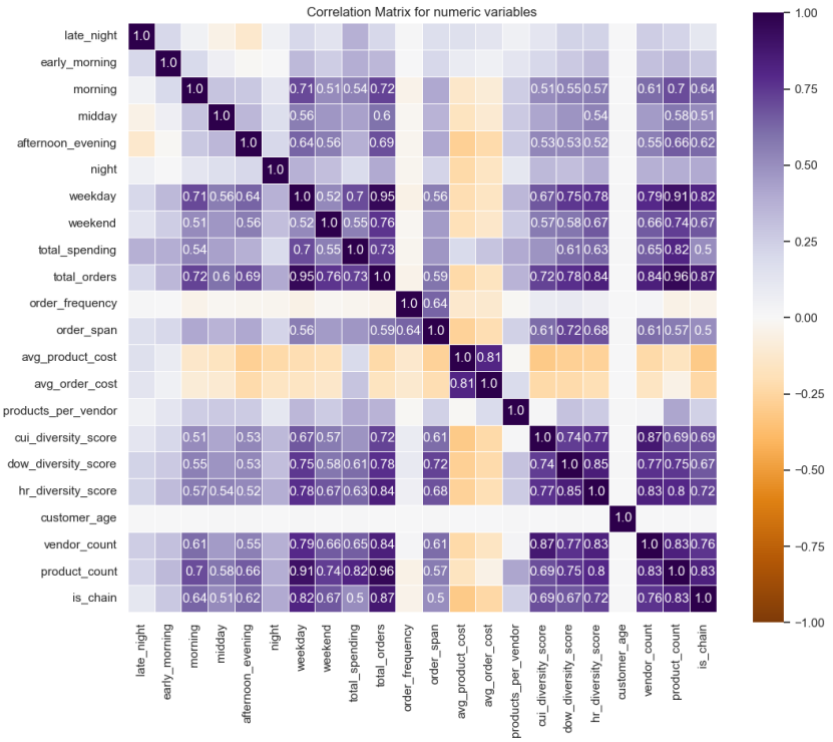

Figure 4 - Correlation Matrix of all Numeric Features

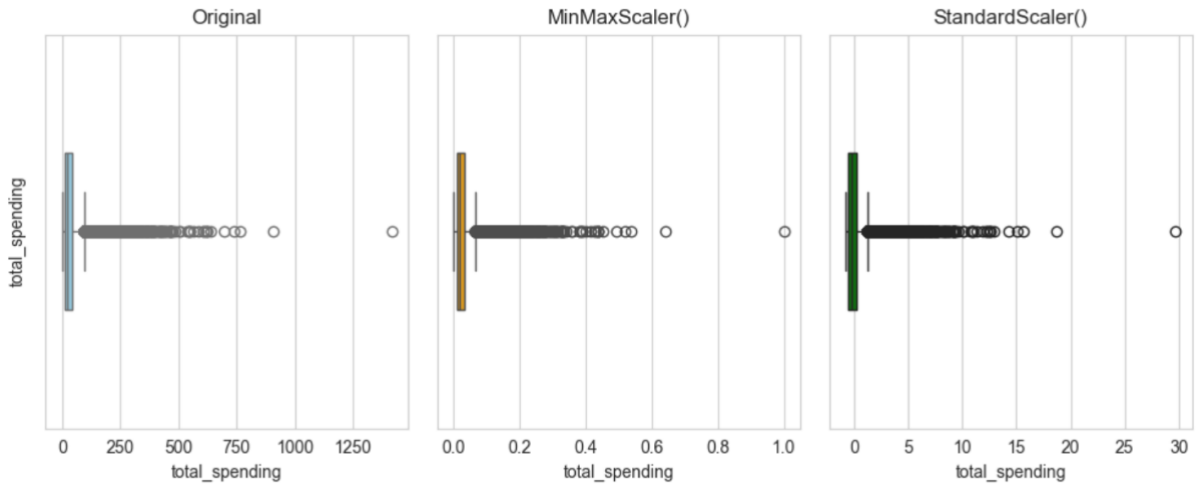

Figure 5 – Impact of different scaling methods on Total Spending

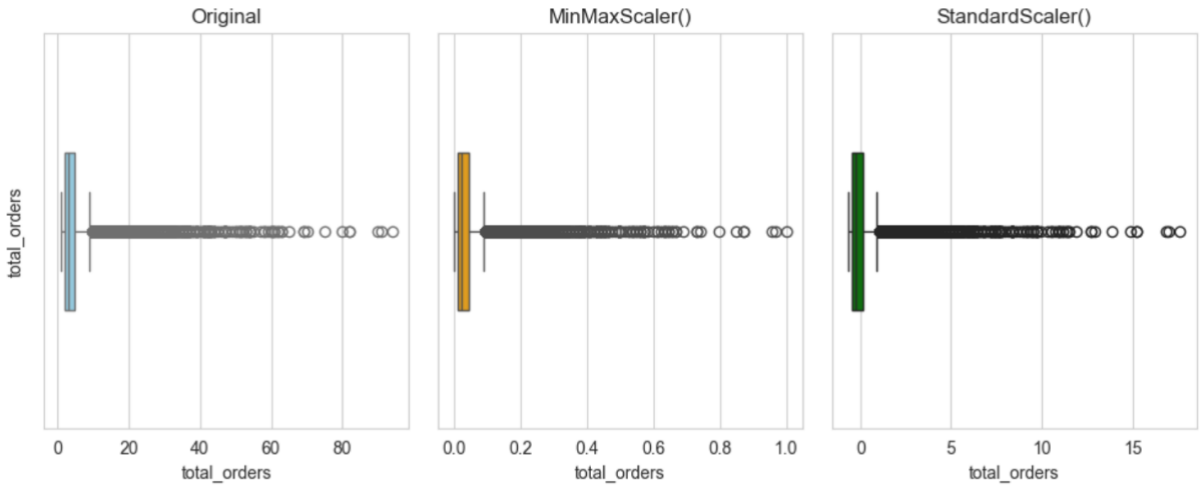

Figure 6 - Impact of different scaling methods on total_orders

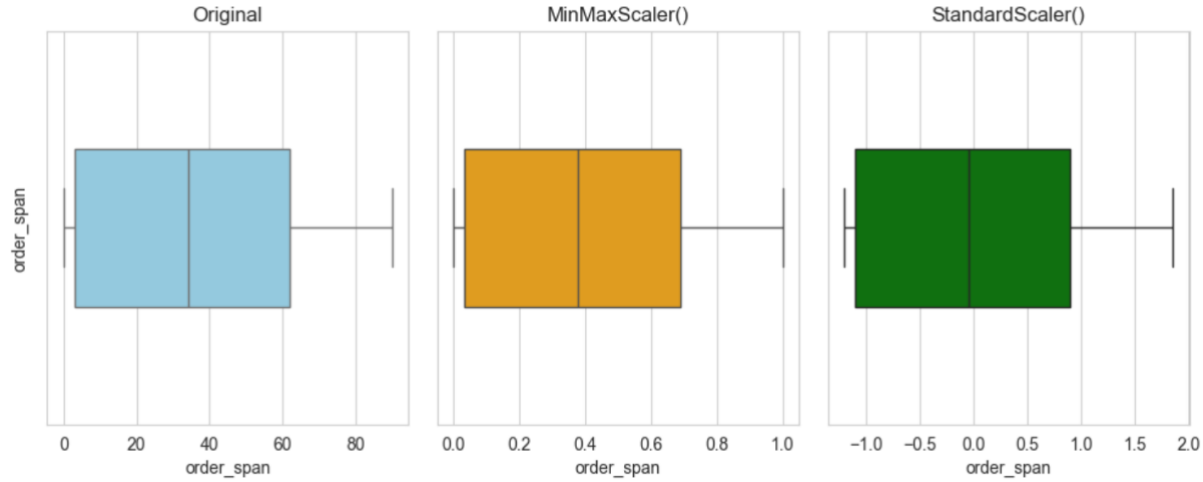

Figure 7 - Impact of different scaling methods on order_span

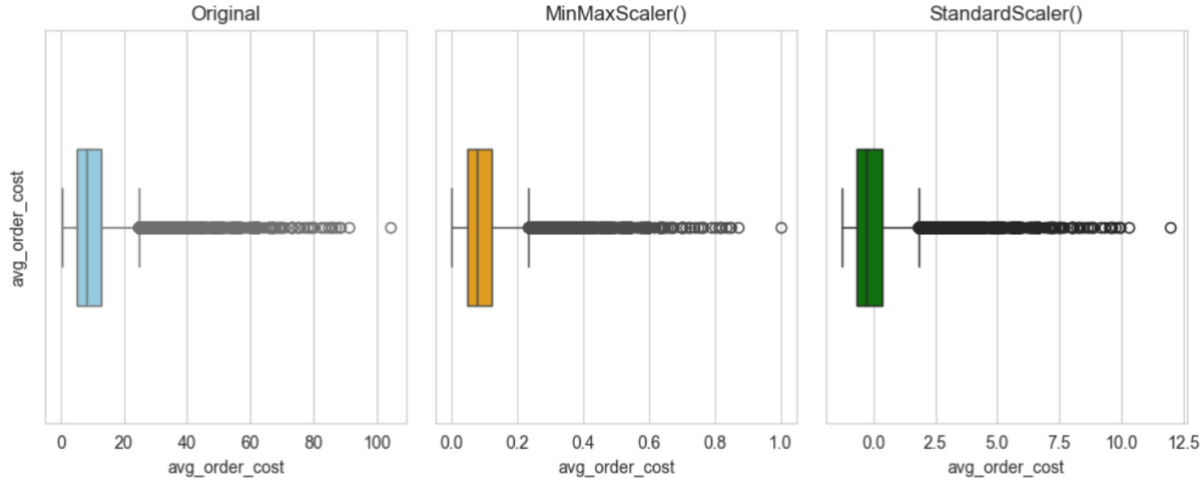

Figure 8 - Impact of different scaling methods on avg_order_cost

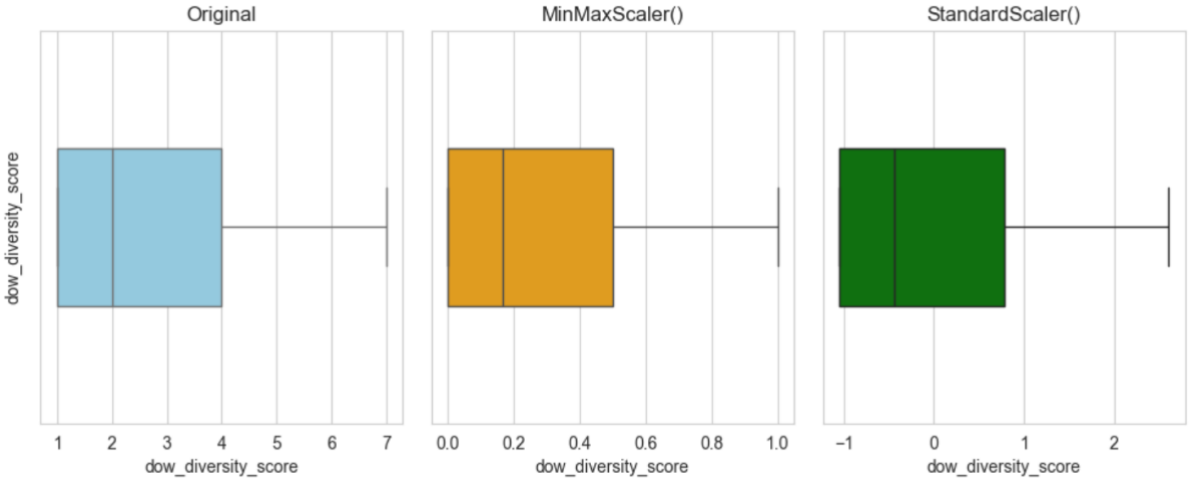

Figure 9 - Impact of different scaling methods on dow_diversity_score

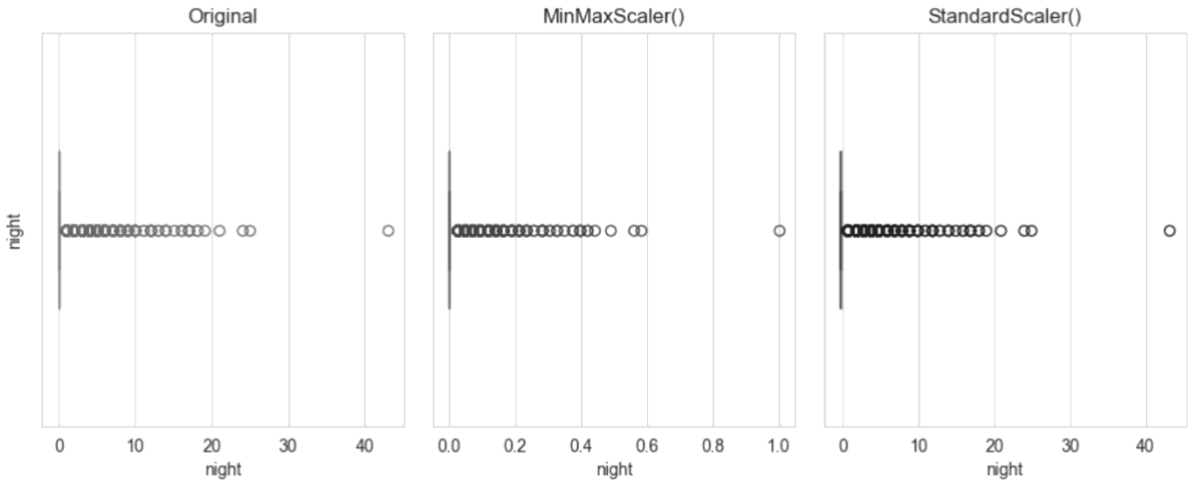

Figure 10 - Impact of different scaling methods on Night

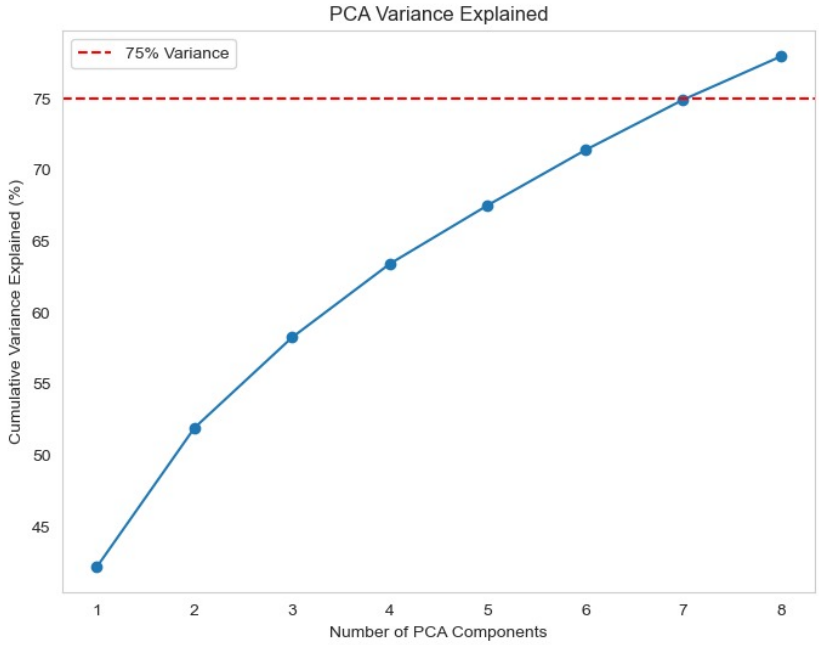

Figure 11 - Variance explained by PCA components

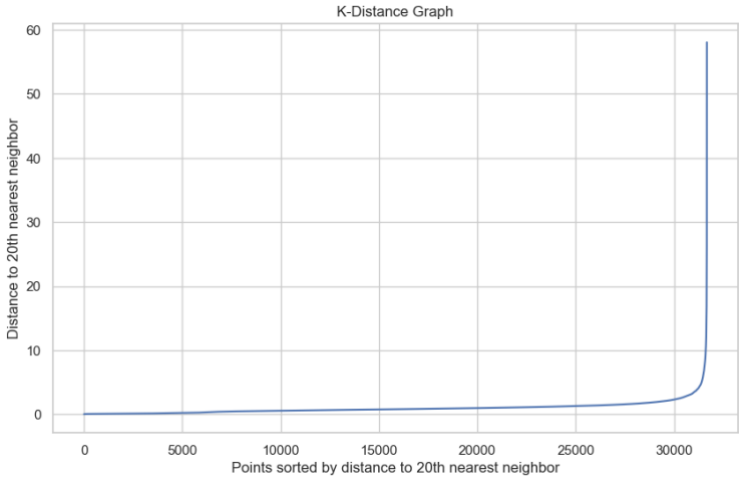

Figure 12 – Points sorted by distance to 20th nearest neighbor

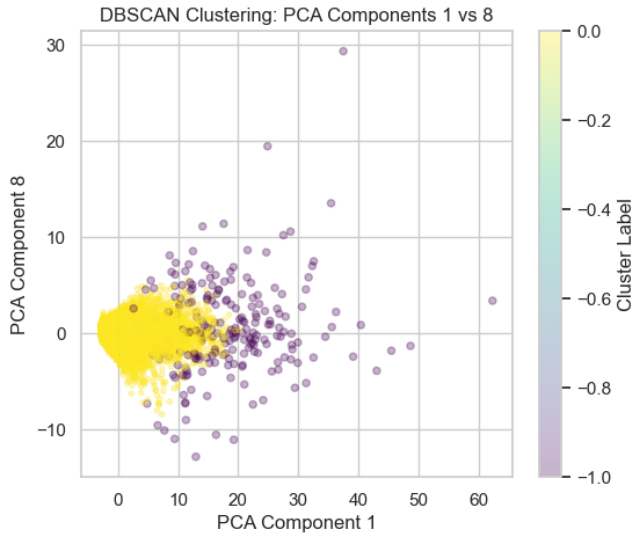

Figure 13 - DBSCAN Clustering – PCA 1 VS. PCA 8 for the first set of parameters

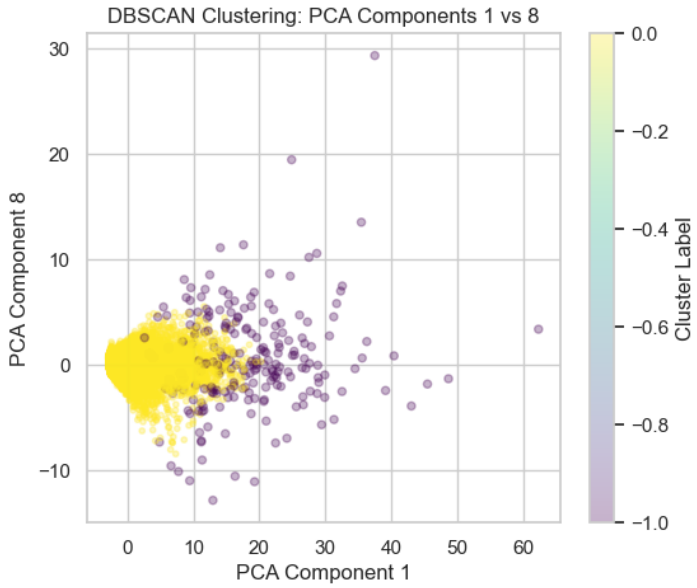

Figure 14 - DBSCAN Clustering – PCA 1 VS. PCA 8 for the second set of parameters

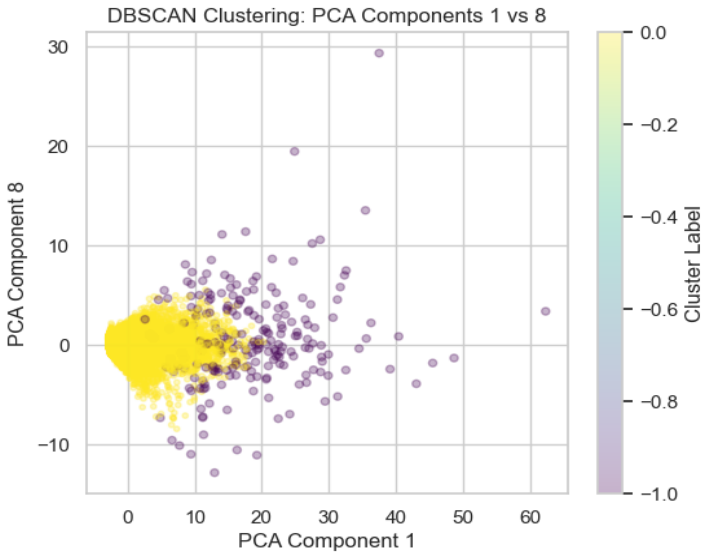

Figure 15 - DBSCAN Clustering – PCA 1 VS. PCA 8 for the third set of parameters

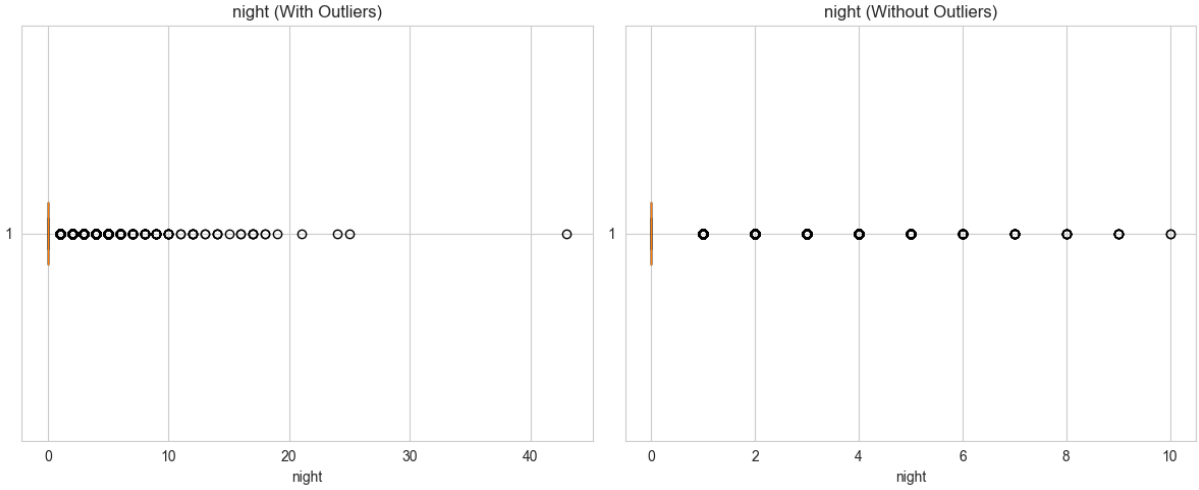

Figure 16 – Impact of final outlier treatment solution on Night

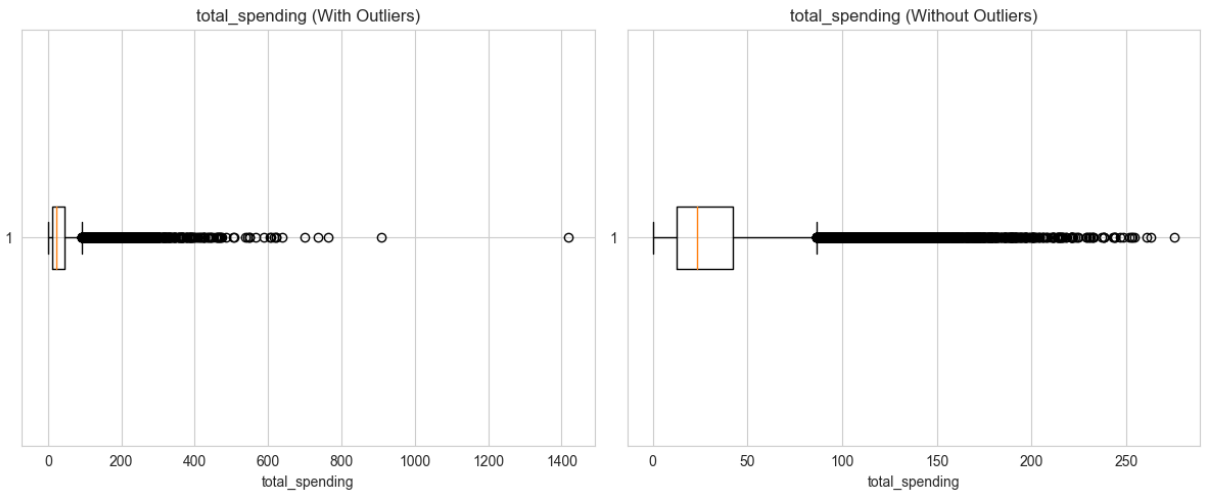

Figure 17 – Impact of final outlier treatment solution on Total Spending

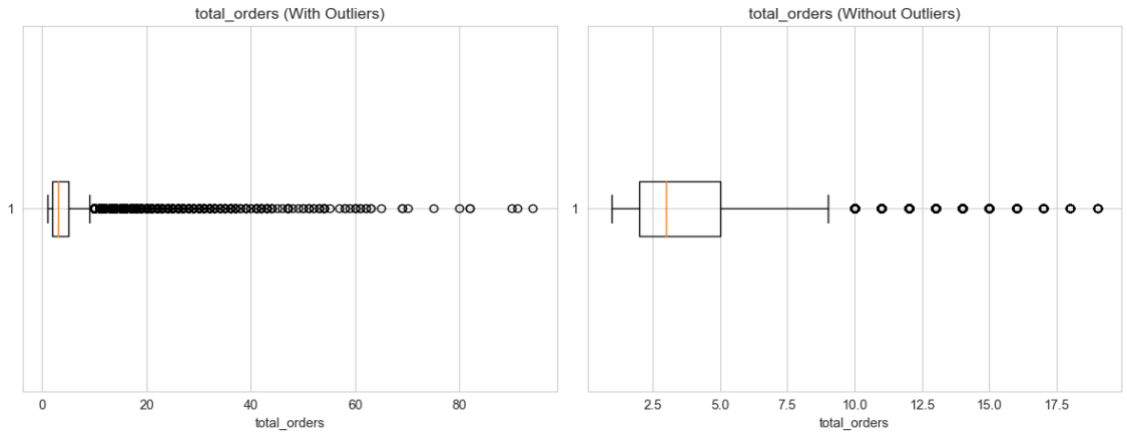

Figure 18 – Impact of final outlier treatment solution on Total Orders

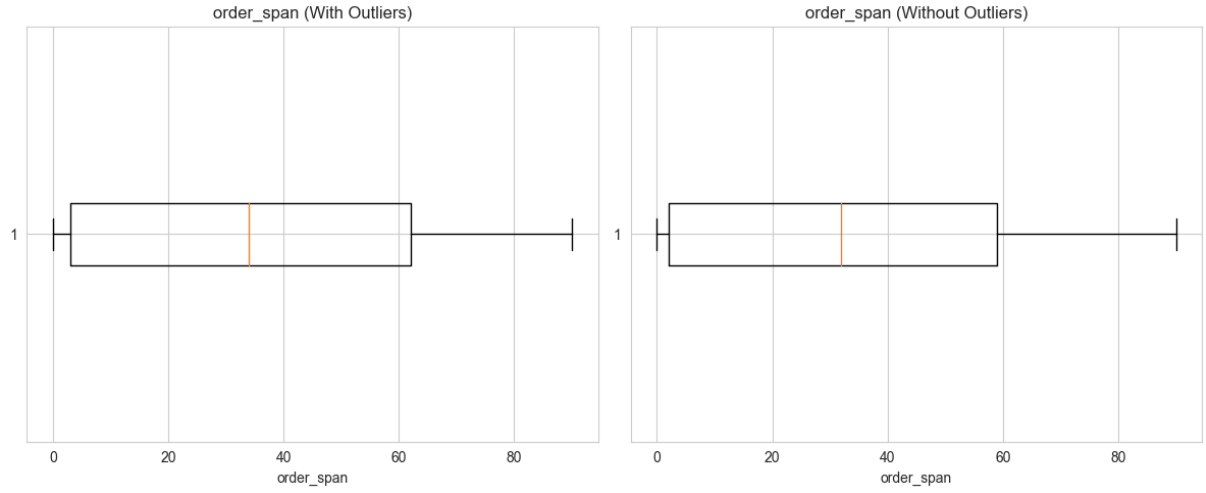

Figure 19 – Impact of final outlier treatment solution on Order Span

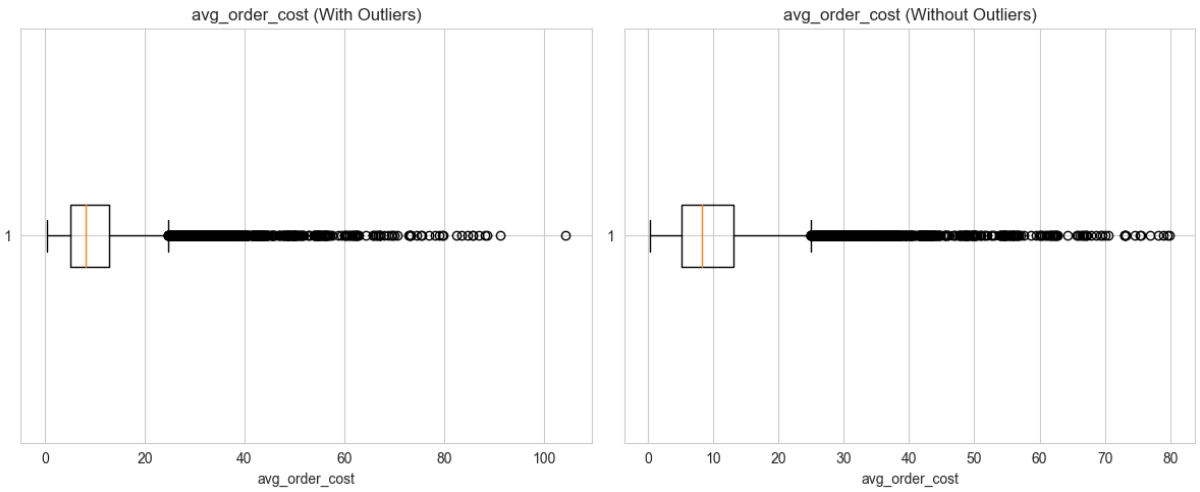

Figure 20 – Impact of final outlier treatment solution on avg_order_cost

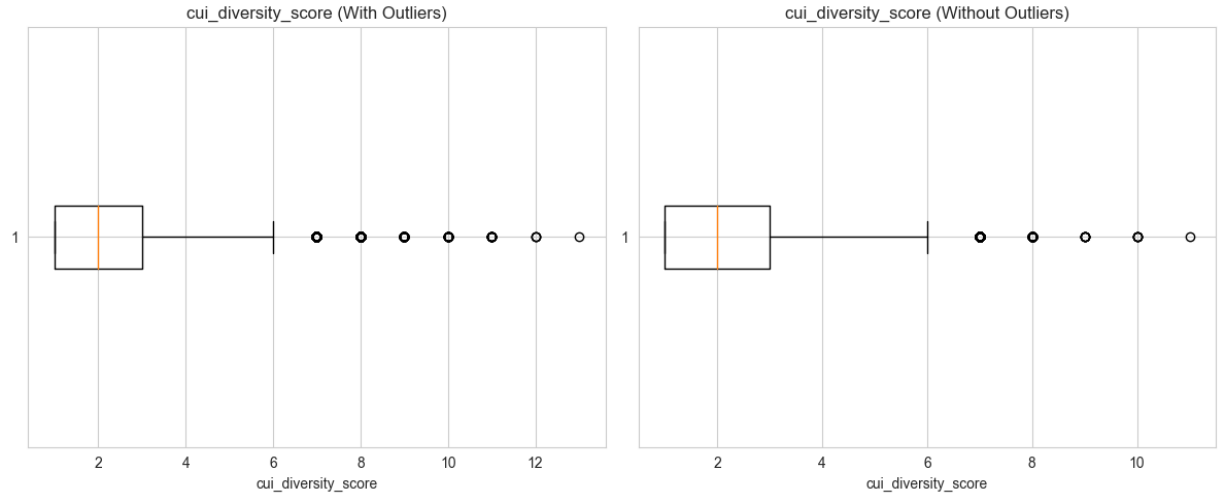

Figure 21 – Impact of final outlier treatment solution on CUI Diversity Score

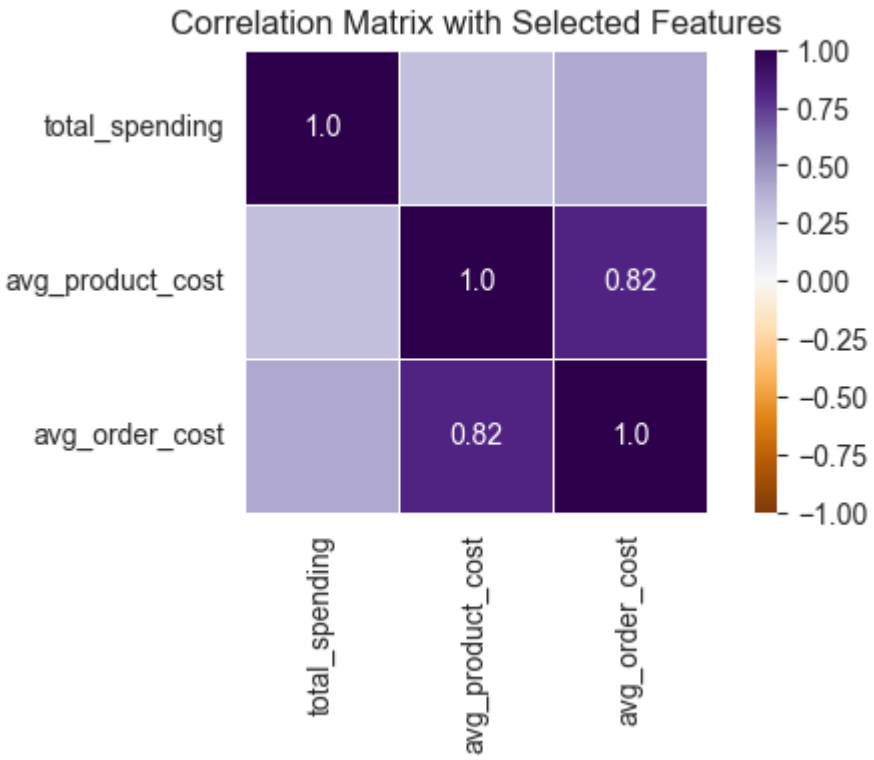

Figure 22 – Correlation Matrix of inital Value Perspective Variables

Figure 23– Correlation Matrix of final Value Perspective Variables

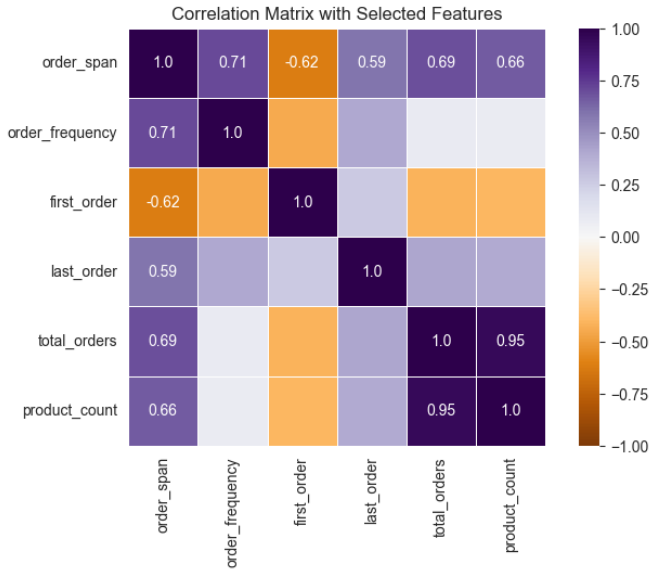

Figure 24 – Correlation Matrix of initial Loyalty Perspective Variables

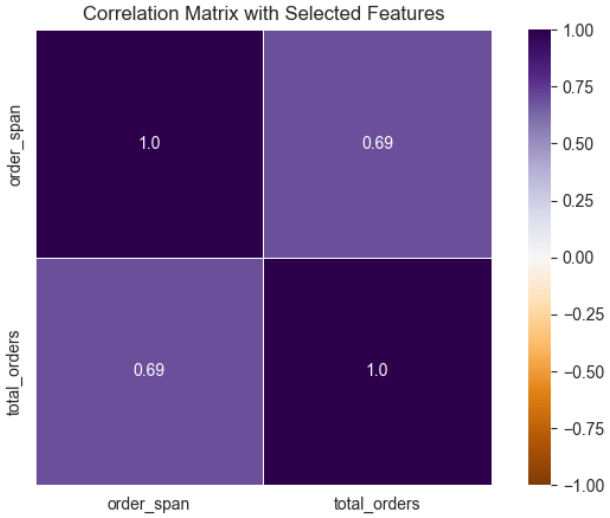

Figure 25 – Correlation Matrix of Final Loyalty Perspective Variables

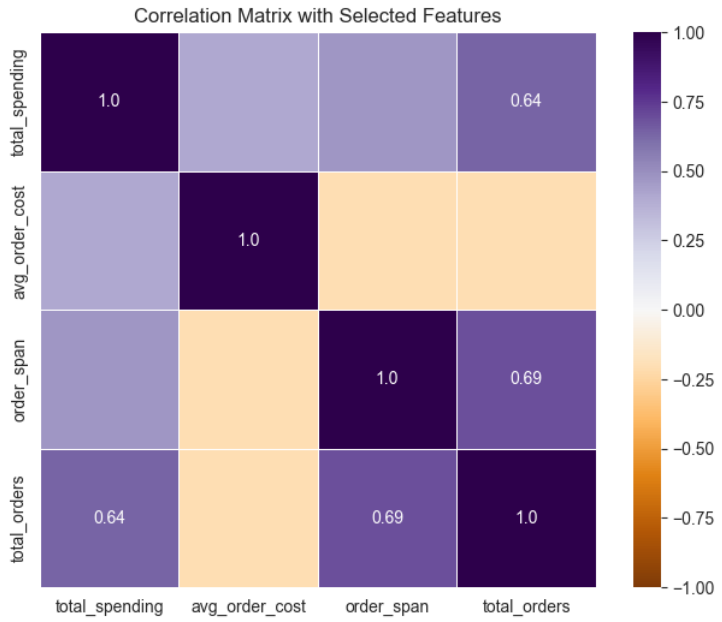

Figure 26 – Correlation Matrix of Merged Value and Loyalty Perspective Variables

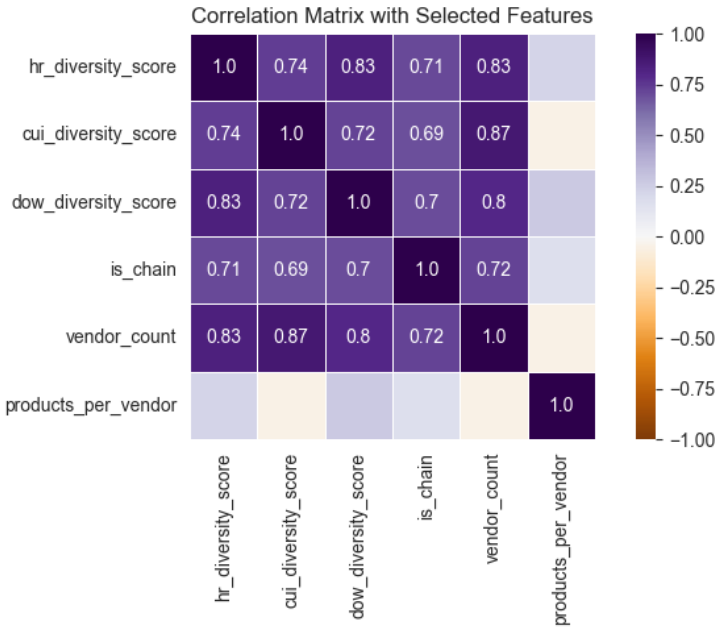

Figure 27 – Correlation Matrix of Initial Preference Perspective Variables

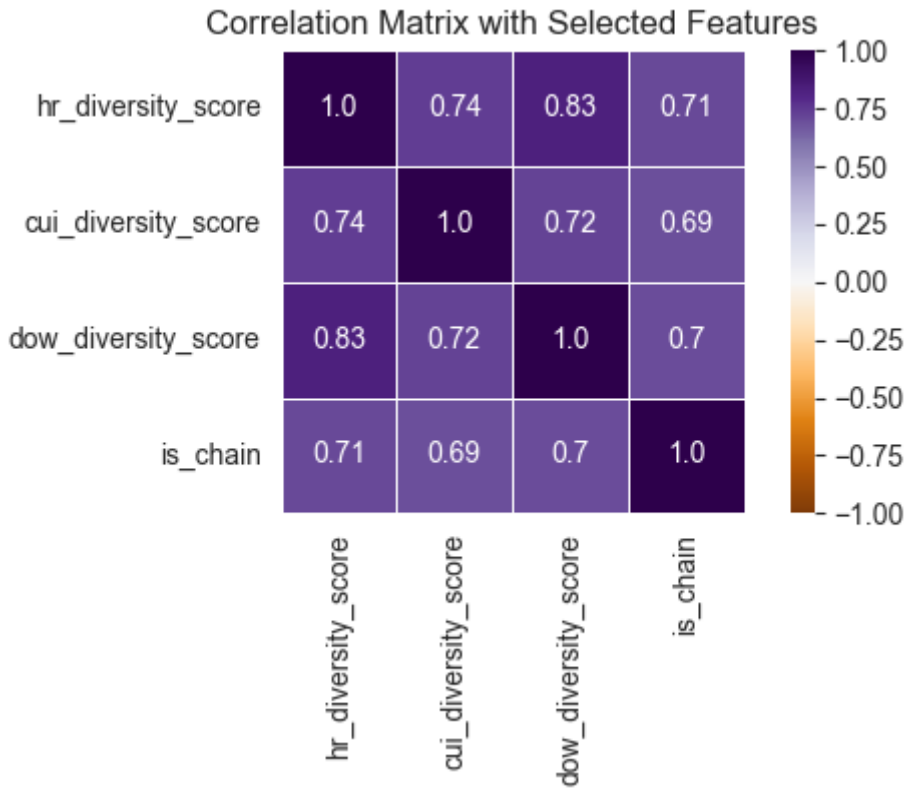

Figure 28 – Correlation Matrix of Final Preference Perspective Variables

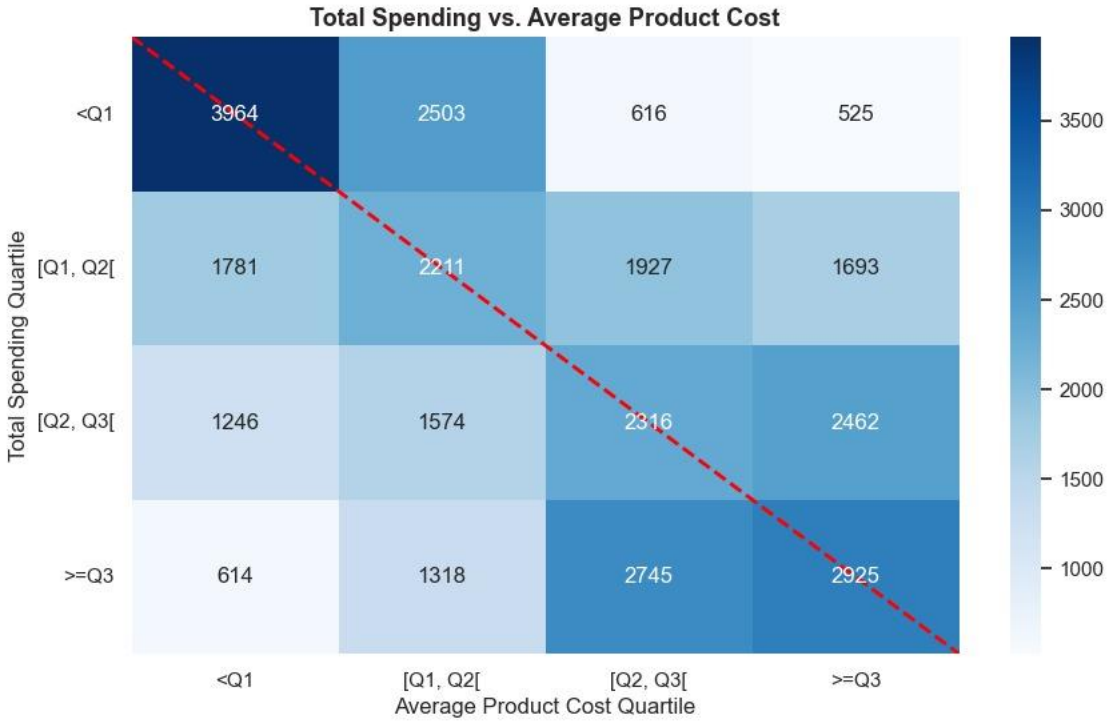

Figure 29 - Quartile Cross Table between Total Spending and Average Product Cost

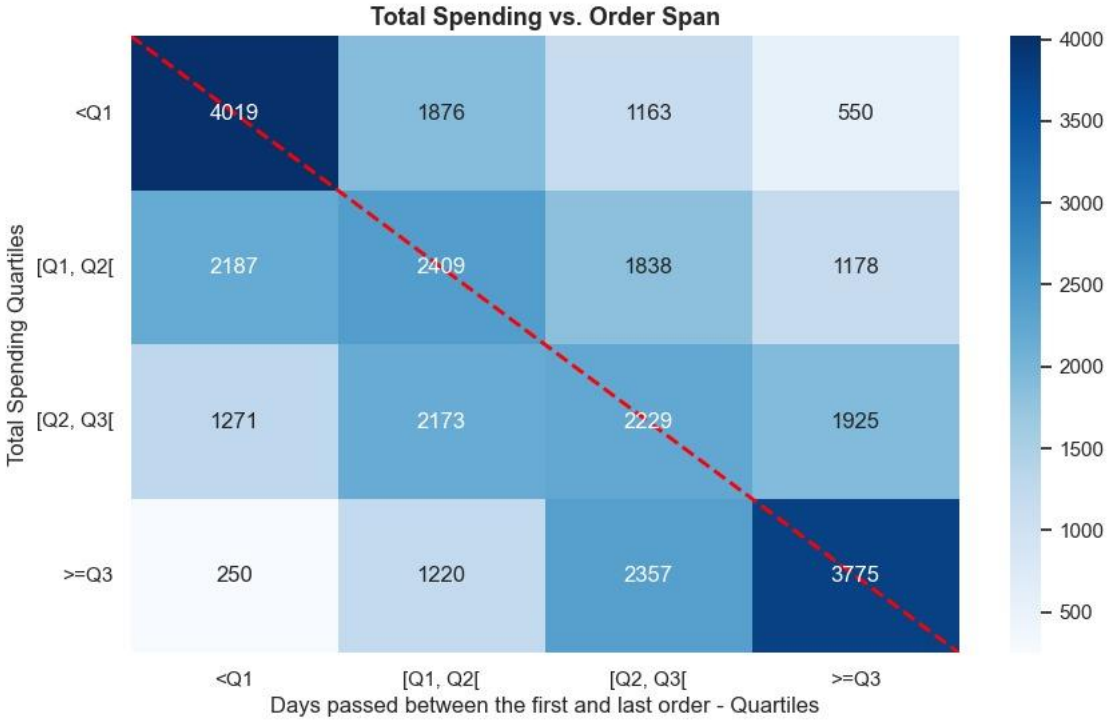

Figure 30 - Quartile Cross Table between Total Spending and Order Span

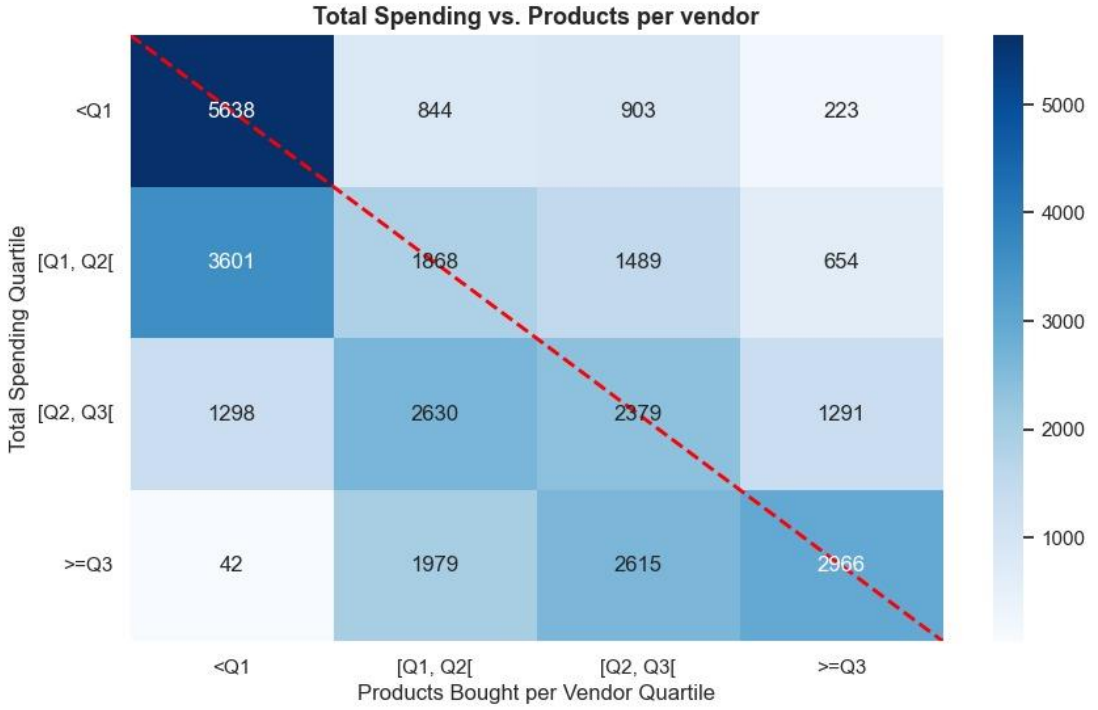

Figure 31 - Quartile Cross Table between Total Spending and Products per Vendor

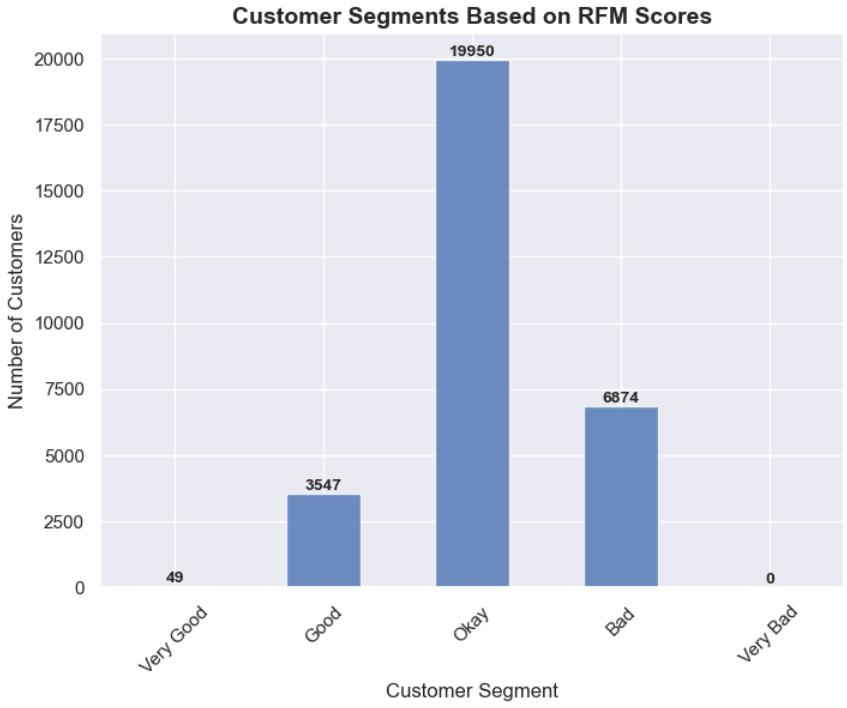

Figure 32 – Bar Chart of Customer Segmentation with RFM analysis

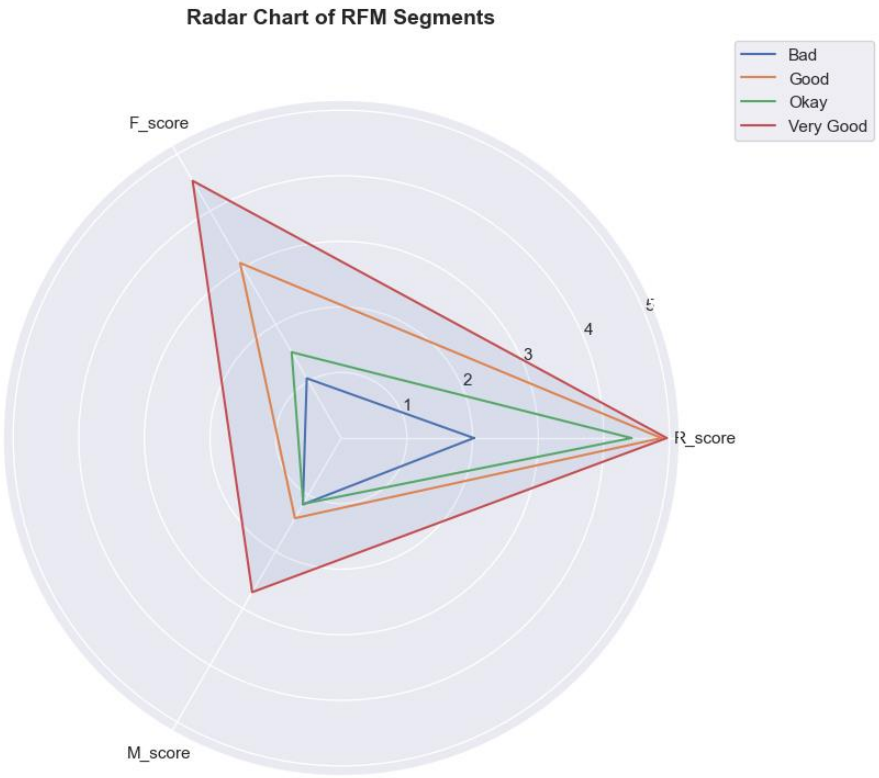

Figure 33 – Radar Chart of RFM Segments

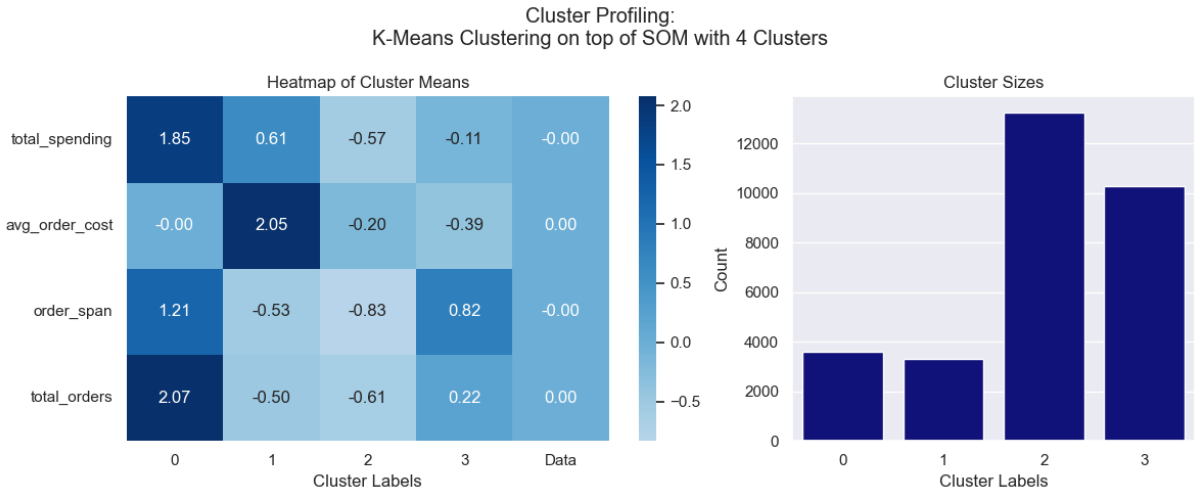

Figure 34 - Final Clustering Result for Value and Loyalty Perspective

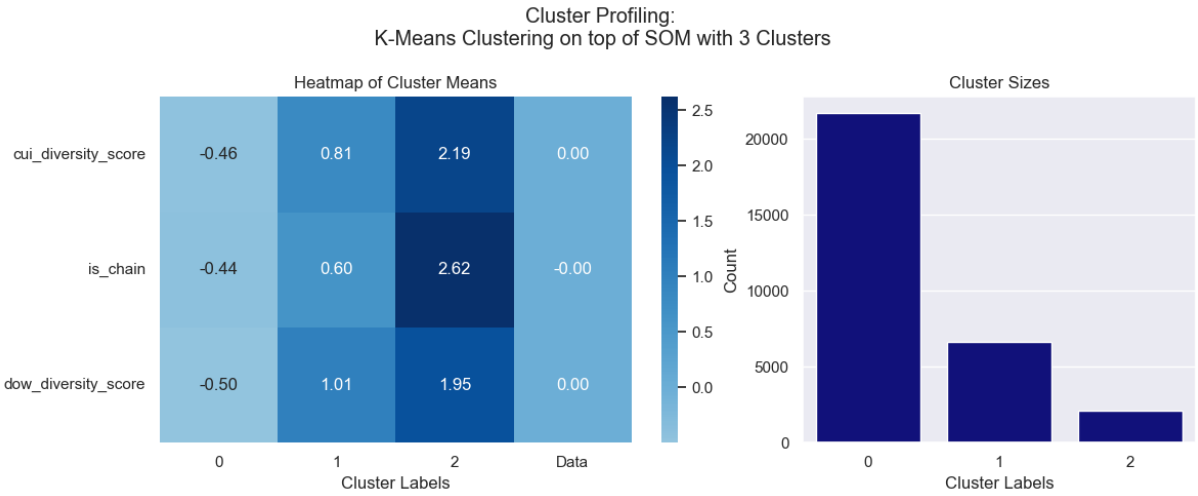

Figure 35 - Final Clustering Result for Preference Perspective

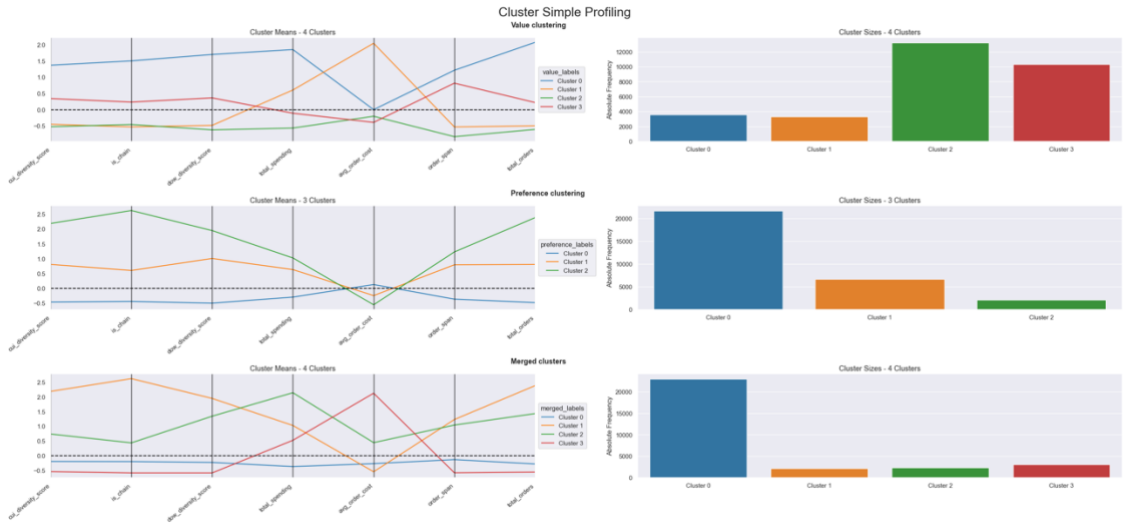

Figure 36 - Final Clustering Result for mergerd perspectives using Hierarchical Clustering

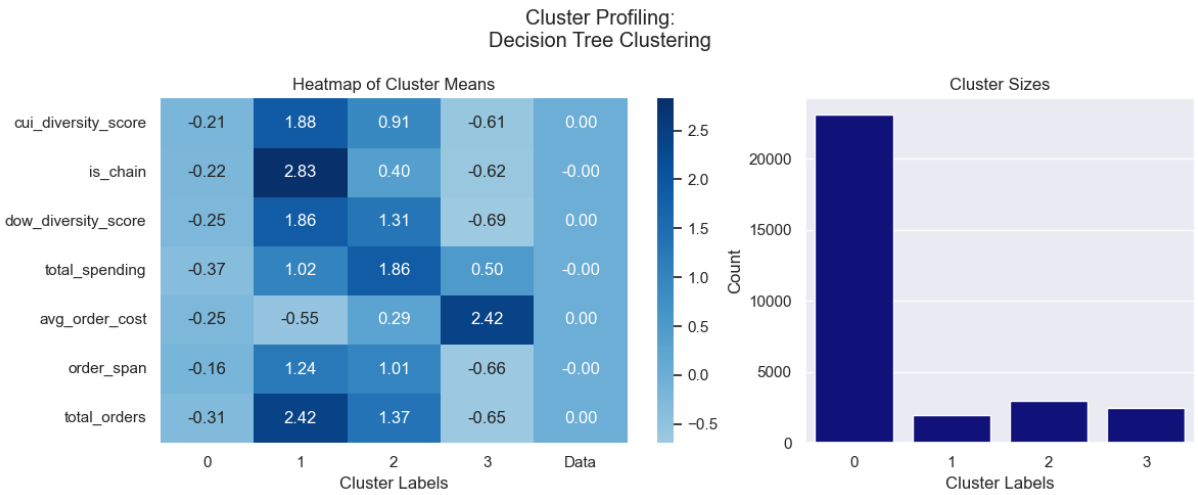

Figure 37 - Final Clustering Result for merged perspectives using the Decision Tree

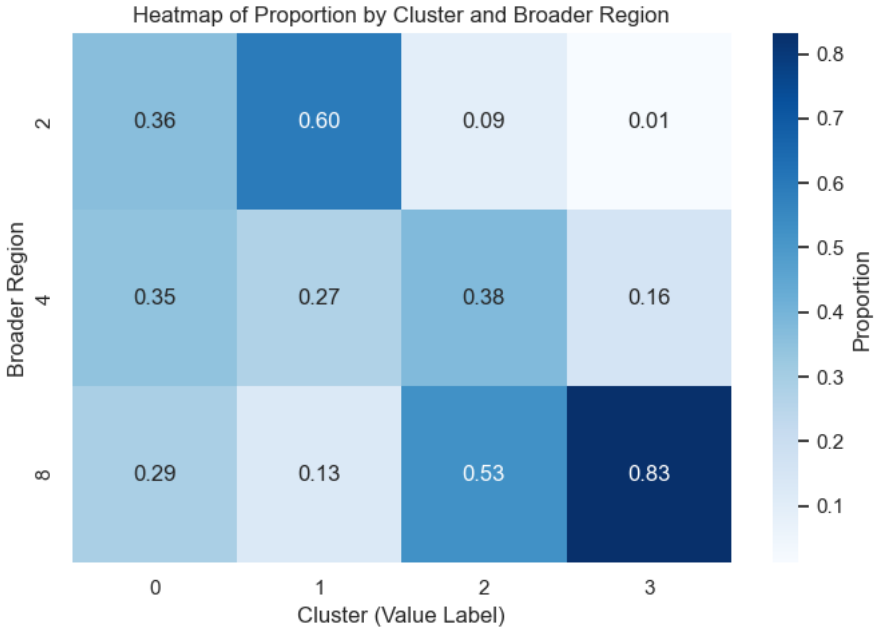

Figure 38 – Heatmap of Proportion of each Broader Region by Cluster

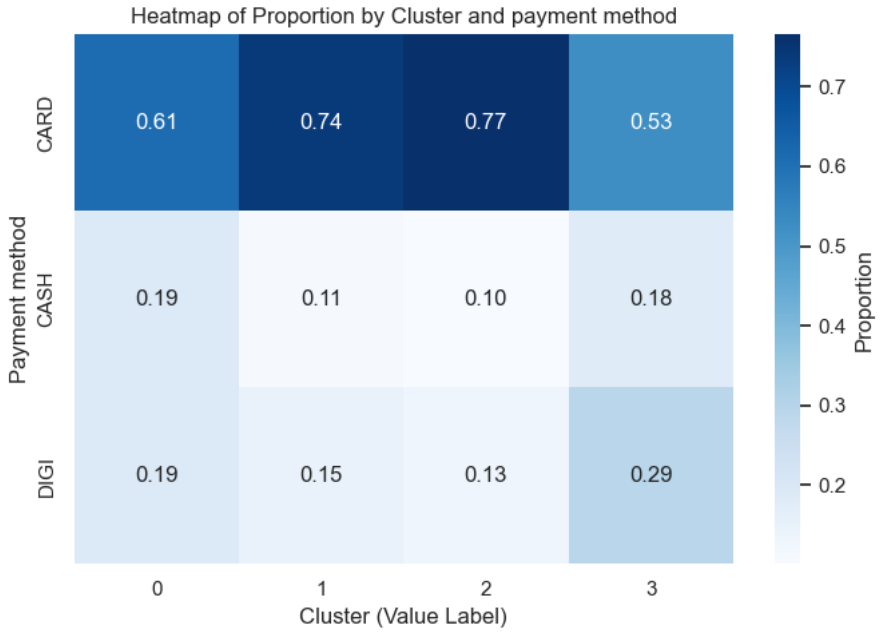

Figure 39 – Heatmap of Proportion of each payment method by Cluster

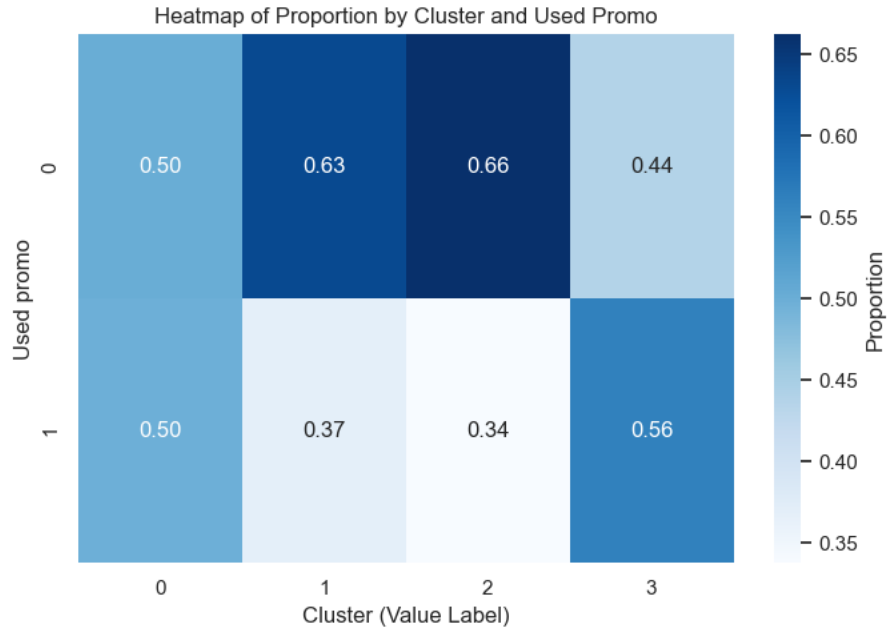

Figure 40 – Heatmap of Proportion of Used Promo by Cluster

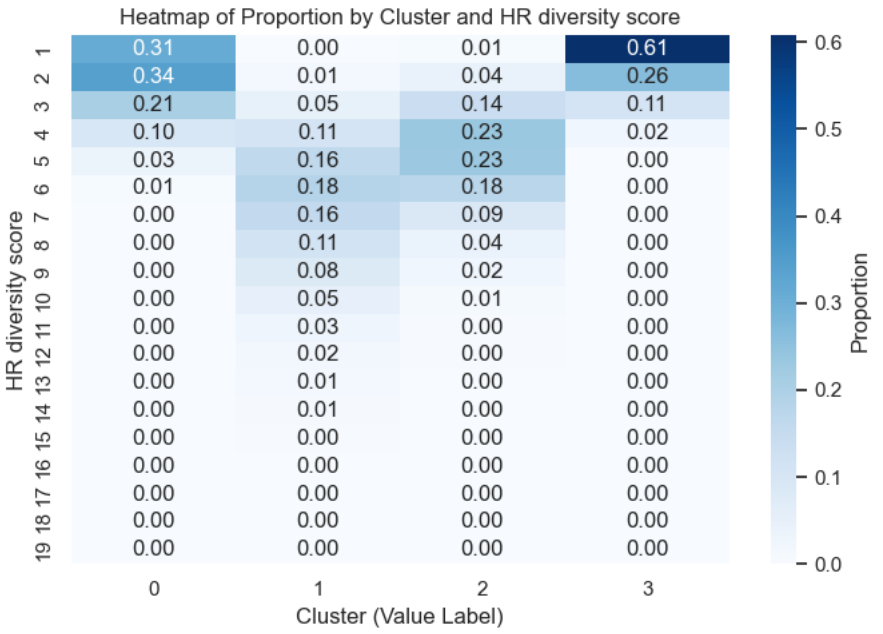

Figure 41 – Heatmap of Proportion of *hr_diversity_score* by cluster

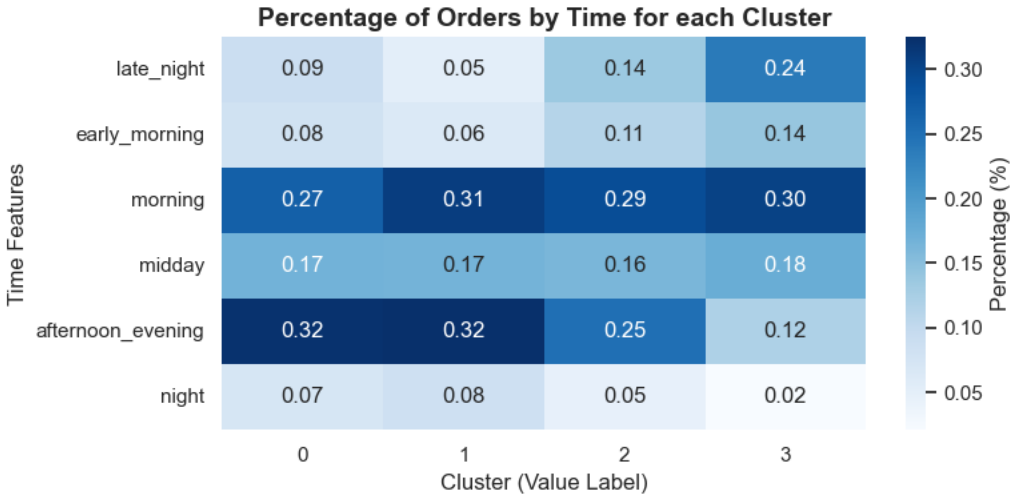

Figure 42 – Heatmap of Proportion of Time Features by cluster

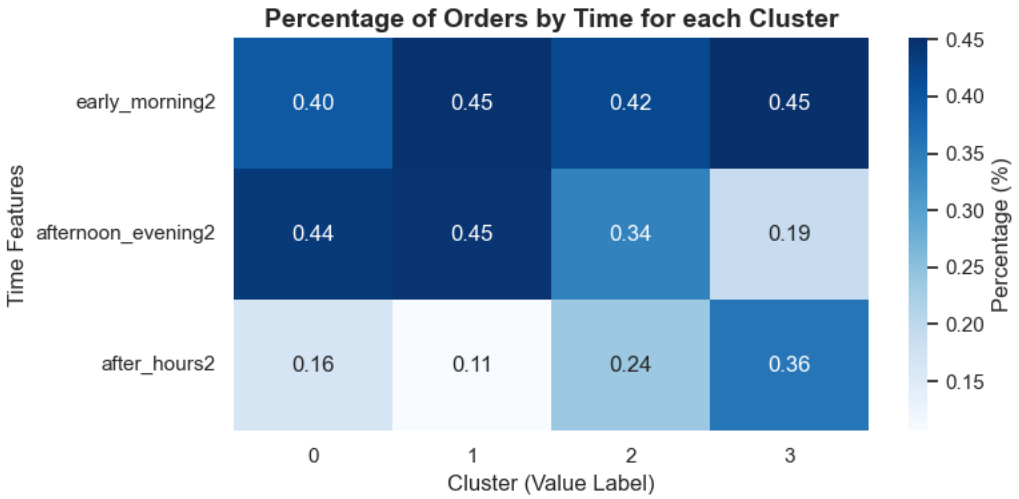

Figure 43 – Heatmap of Proportion of Time Features by cluster (2)

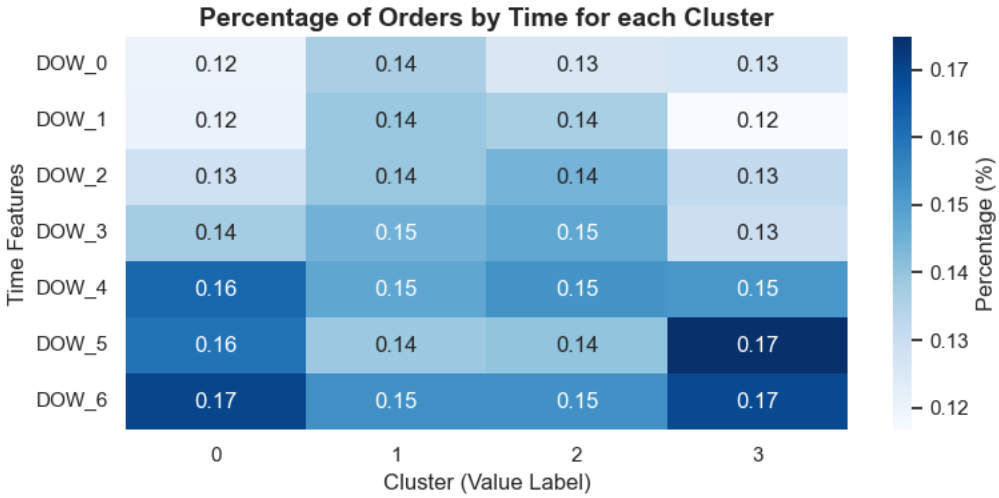

Figure 44 – Heatmap of Proportion of Time Features by cluster (3)

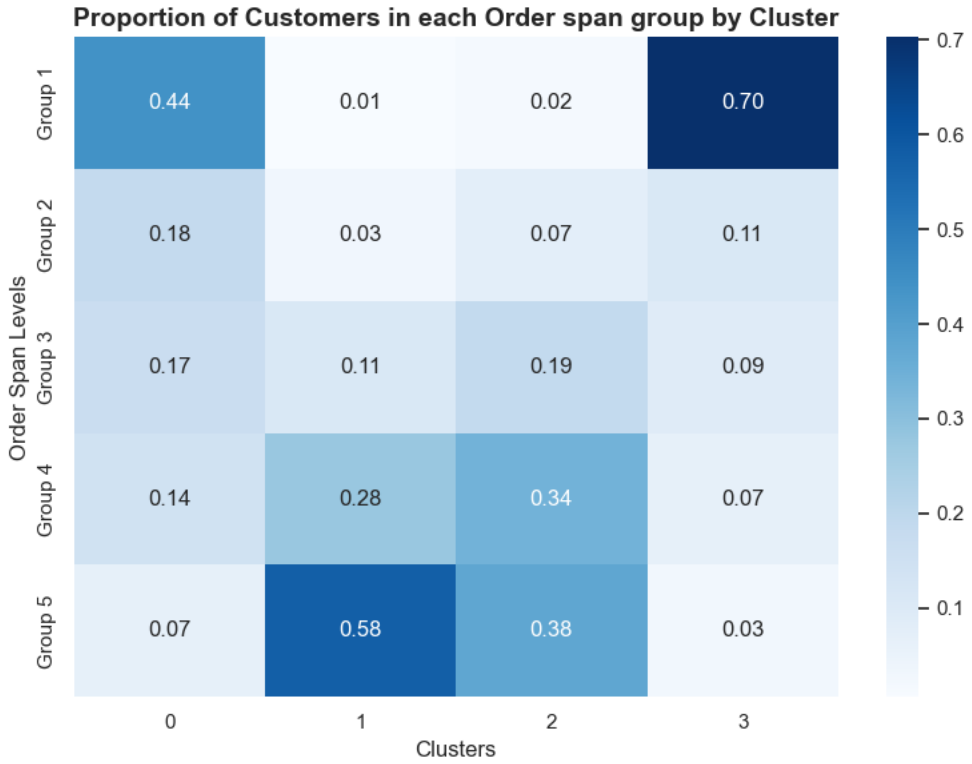

Figure 45 – Proportion of customers in each *order_span* group by cluster (being group 1 the one with lowerst *order_span* and group 5 the one with the highest *order_span*)

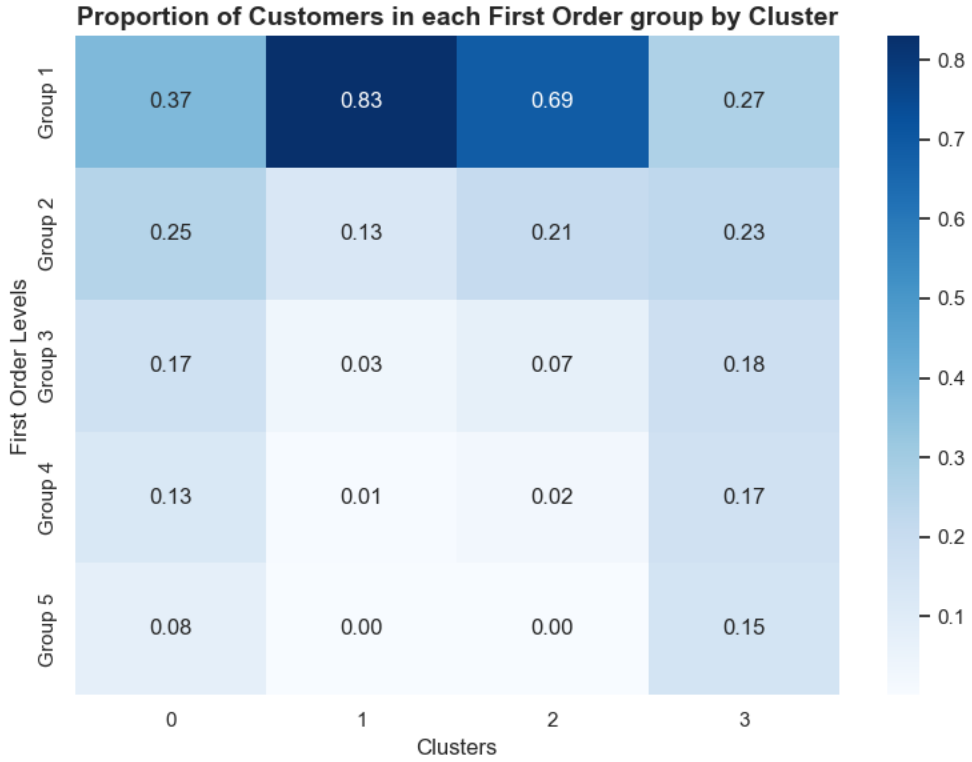

Figure 46 – Proportion of customers in each *first_order* group by cluster (being group 1 the one with lowest *first_order* and group 5 the one with the highest *first_order*)

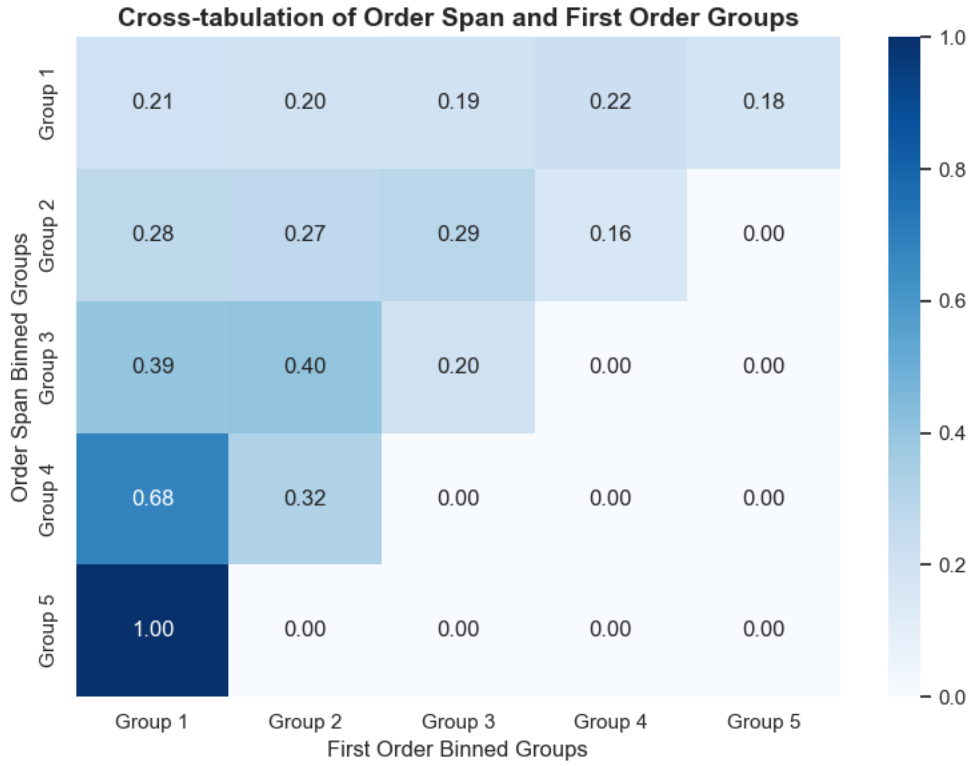

Figure 47 – Cross table of *order_span* and *first_order* groups

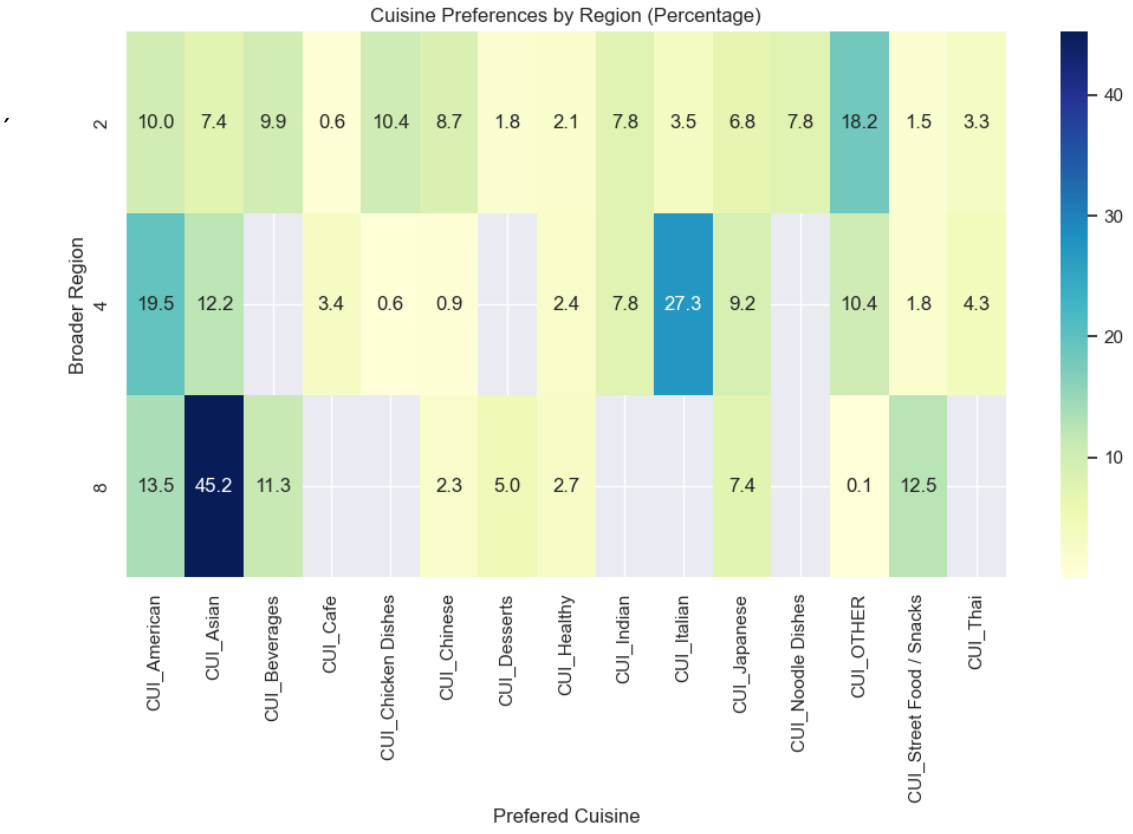

Figure 48 – Heatmap of Cuisine preferences by Broader Region in Cluster 0

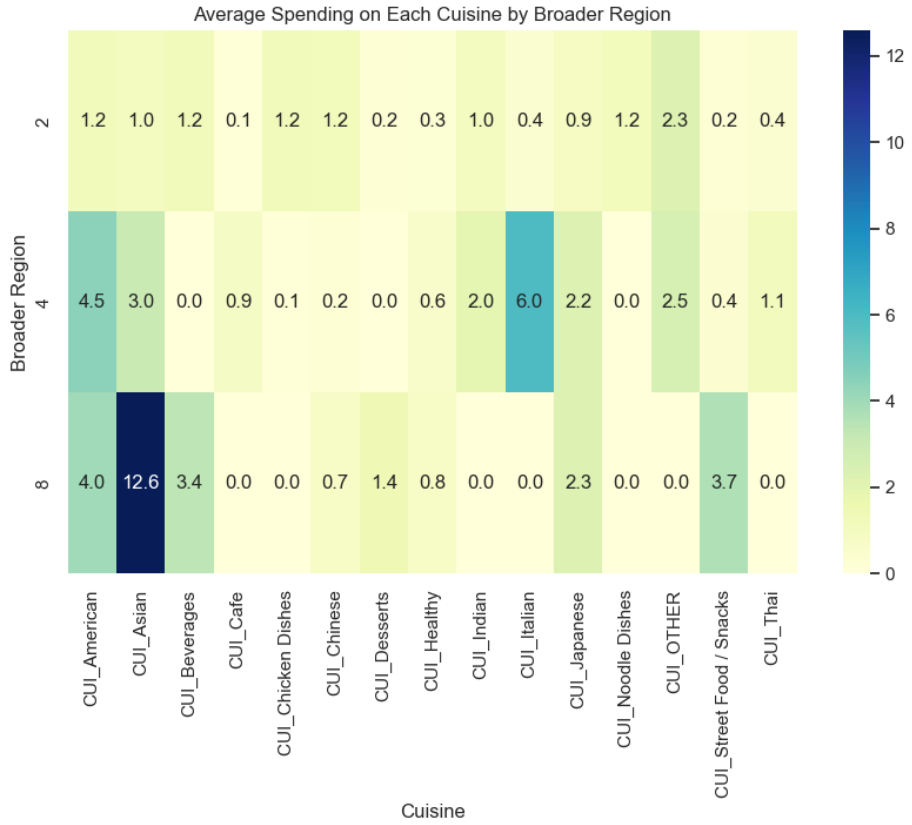

Figure 49 – Heatmap of average spending on each cuisine by broader region in cluster 0

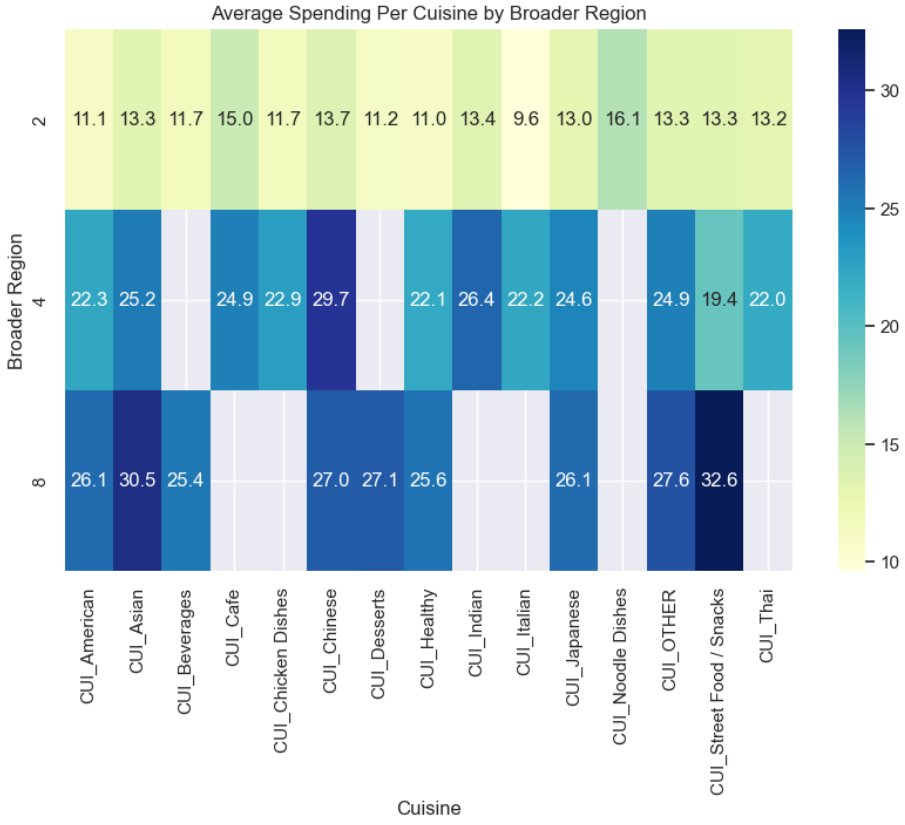

Figure 50 – Heatmap of average spending on each cuisine by broader region in cluster 0

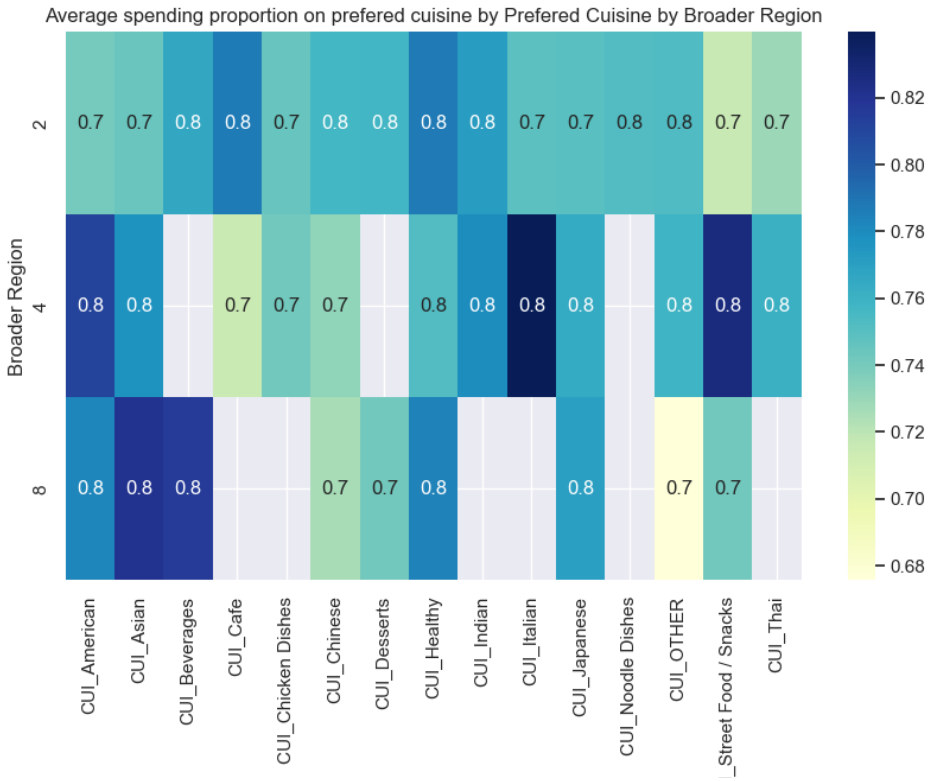

Figure 51 – Heatmap of average spending proportion on preferred cuisine by broader region in cluster 0

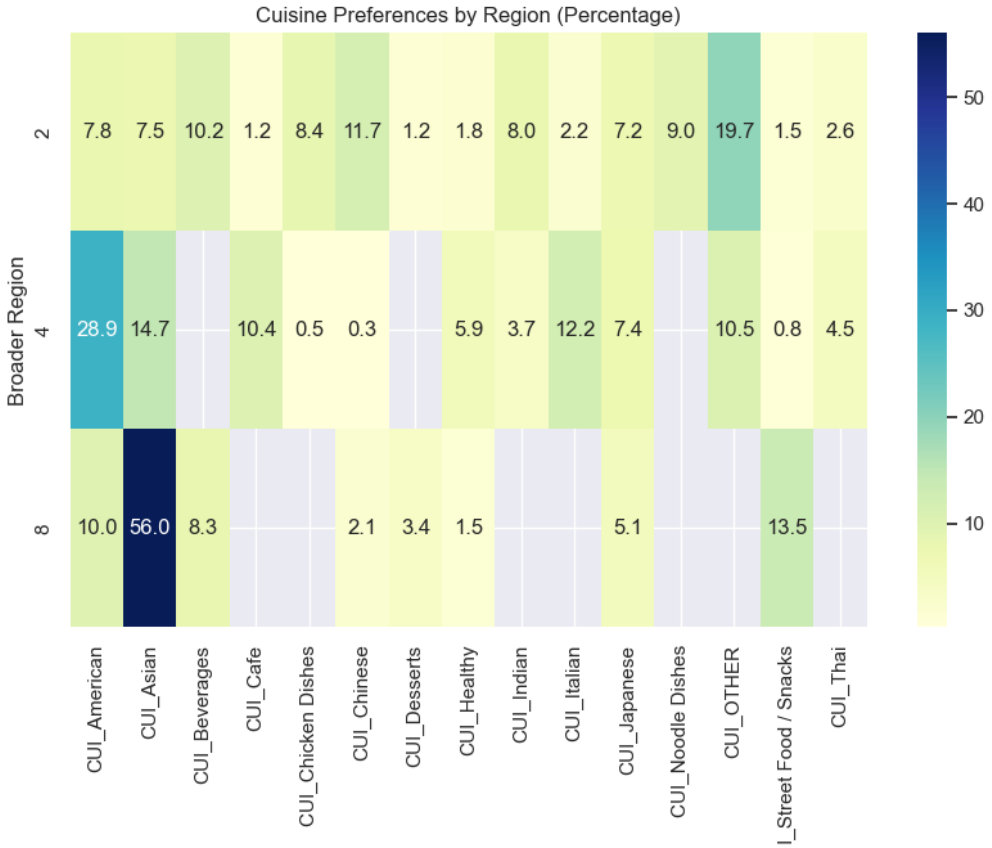

Figure 52 – Heatmap of cuisines preferences by broader region in cluster 0

Figure 53 – Heatmap of time preferences in each cluster

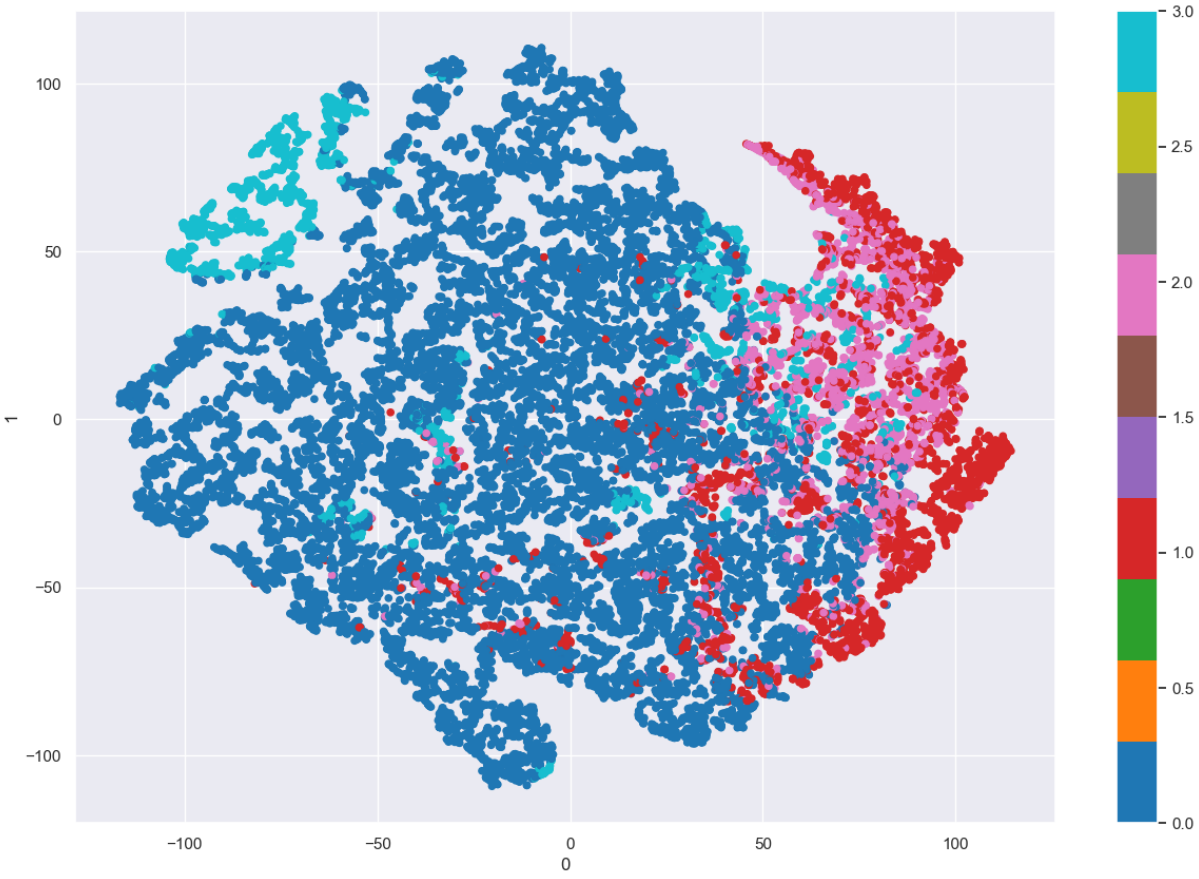

Figure 54 – TSNE of final cluster solution obtained with the decision tree