DATA MINING PROJECT

Master in Data Science and Advanced Analytics

NOVA Information Management School
Universidade Nova de Lisboa

Customer Segmentation Project
ABDCEats Inc.

Course: Data Mining

Professores : Fernando Bação | Farina Pontejos

Group 8

Alexandre Gonçalves, 20240738

Bráulio Damba, 20240007

Victoria Goon, 20240550

Fall/Spring Semester 2024-2025

TABLE OF CONTENTS

Github Code: https://github.com/VGoon/DM2425_Group8

1. INTRODUCTION

This project is a continuation of our earlier EDA report. We will build upon what was done in that report, namely, as the title may already suggest, performing customer segmentation on the data collected from ABCDEats Inc. For this, we will preprocess the data to address any anomalies and concerns found during the EDA analysis to ensure that we get the most meaningful clusters to interpret. From there on out, we shall seek to ensure the validity of the clusters using techniques discussed in class, such as R2 score and T-SNE. An individual analysis of each cluster shall then be conducted and only afterwards will we draw business conclusions. Since we're building upon the EDA, some steps that were already taken there shall not be discussed in depth or at all and in these situations, we will direct the reader to the EDA or use excerpts from it. We will also not discuss the structure of the dataset nor the associated metadata, as this was also already done in the EDA.

2. PREPROCESSING

2.1. Index Definition and Missing Values

Our approach first started off with setting the column "customer_id" as the index of our data frame, "df". This flagged several duplicate IDs which we removed from the dataset to ensure no redundancies.

The columns "first_order", "customer_age" and "HR_0" were the only columns of the 52 that had missing values that had to be dealt with. The "first_order" column had missing values that represented only 0.33% of the data, so dropping these values would have a negligible effect on the data set. We found that these specific rows represent customers that didn't order within the timeframe the data was collected so it was determined it would be best to drop these customers from the analysis as they wouldn't provide useful customer purchasing insights. We imputed our missing customers age with the median, since this approach is more robust to outliers. Values in our "HR_0" column that were missing were imputed with 0. The team chose to use "0" to fill these values as the column was filled with only "0"s for all other rows. The lack of variance suggests that the ABCDE app could potentially be down for an hour from midnight to 1:00 am due to a variety of reasons, platform maintenance being the most feasible since it's normal that the traffic on the platform would naturally be down when most people are sleeping. Hence, imputing the missing values with 0 and not dropping them is the most logical approach. The columns already mentioned in this paragraph and the columns "product_count", "vendor_count", "is_chain" and "last_order" were all subsequently changed to "int" to better match the information they represent.

2.2. Feature Creation

We have created seven new variables to better represent our data:

- Total Spent - sums up the total amount of money spent in the app across all cuisines
- Spender Segment - dividing customers into spending segments [Low, Mid-Low, Mid-High, High]

- Order Range - calculates the number of days between the first time ordered and last time ordered
- Age Group - dividing customers into age groups [Teen, Young Adult, Adult, Middle-Aged Adult, Older Adult, Senior]
- Cuisine Variety - counts the total number of different cuisines a customer bought from
- Total Orders - adding the total number of orders made in the app by a customer
- Cuisine Spender Segment - categorizes the amount spent per cuisine for each customer into a spending segment [None, Low, Mid-Low, Mid-High, High]

For the feature “Total Orders”, as stated in the EDA, was composed of the sum total of orders from either the Days or Hours columns. It was expected that the Days and Hours counts would match but some discrepancies were found. There were no indications on which values were more accurate than the other so it was determined to drop these rows as their validity was in question.

With potential business applications in mind for our clustering approach, we also altered the hour (HR_X) features that represented all 24 hours of the day to reflect human eating habits more accurately by defining them into six new variables: "Early_Morning", "Breakfast", "Lunch", "Snack", "Dinner" and "Late_Night". As people generally follow a conventional eating schedule based around the previously mentioned meal times, it would make interpreting the clusters easier using these variables as opposed to the provided hour columns.

The cuisine (CUI_X) features were transformed using 2 different methods as well. Cuisine Variety helps determine the number of unique restaurants a customer orders from whereas the “Cuisine Spender Segment” determines whether they spent a large amount of money compared to the rest of the data population. Their values were re-represented as the following: “None”, “Low”, “Mid-Low”, “Mid-High”, “High”.

2.3. Outlier Treatment

As already discussed in the EDA, the initial data set a large number of outliers. We determined though that most of the customers that were considered as outliers were all simply because they ordered more than others which skewed the metrics for everything else and most of the other features are collected based on the order (adding a tally to CUI_American when ordering American cuisines). From this logic, we decided to turn select variables into ratios based on the total number of times a customer ordered from the application: ('products_per_order' ("product_count" **/** "total_orders”), 'vendor_per_order' ("vendor_count" **/** "total_orders"), 'chain_per_order' ("is_chain" **/** "total_orders"), "cuisine_per_orders" ("cuisine_variety" **/** "total_orders") and "spent_per_orders" ("total_spent" **/** "total_orders")). There were still plenty of outliers when examining these ratios, with all of them tending to higher spenders/ higher orderers. So from there, we found the interquartile range (IQR) for each feature and filtered the entire dataset to exclude any that exceed the IQR +/- the upper/lower range limits respectively. This filtered out

approximately 6000 records which were saved in their own dataset where they were scaled to their own group and set aside for clustering on their own.

2.4. Feature Selection and Transformation

A twofold approach was used for feature selection. Through a Spearman correlation graph (figure 3) with the threshold set to 0.85 (it is usually recommended that this value be equal to or higher than 0.7), we identified “vendor_count” and “product_count” as being features that showed up frequently when finding pairs of highly correlated features. To play it safe, we then conducted a variance check and found out that “vendor_count” had a low variance. Based on all this information, we decided to drop these two variables.

With our feature engineering approach, we were able to drop the original variables for hours (HR_X), cuisines (CUI_XXX), and days (DOW_X) which all had very low variance. The new cuisine features created were categorical and created for cluster interpretation more than to be used for the creation of the clusters. The time of day features however were integrated into the clustering algorithms.

When scaling our numerical data, we used the three main scalers: standard, minmax and robust. Initially we decided to go with the robust scaler, because of the abundance of outliers in our data. However, since our initial approach to outlier treatment changed, we decided to instead use minmax scaler. Not only because of our change in approach for outlier treatment, but also because this type of scaling improved the power of our clustering techniques (these techniques will be discussed further along in the project).

3. CLUSTERING

3.1. The Approach

When all is said and done, we ultimately came to the conclusion that doing clustering without outliers produced more defined clusters. Not only that, the R^2 scores for different cluster methods were on average much better without outliers (compare figures 4,5,6 and 7). Therefore, for marketing purposes, we chose to primarily focus on the interpretation of the clusters based on a dataset without outliers. This, however, comes with the cost of losing some interpretability grounded in reality: it’s unrealistic to, for example, not consider people that spend way above the average. As this is the case, we created a second clustering notebook for the high spenders. Given that the main objective is to determine a marketing strategy to increase user engagement, in the remainder of the report we will focus on the interpretation and business applications of the main body of data which ultimately are users who weren’t too actively engaged during the 3 month data collection period.

With this clarification, we went on to determine the best cluster methods to apply based on an economical perspective and customer preference perspective using subsets of our features based on their context and on R2 scores. We only used clustering methods discussed in class, namely K-Means,

with different distance computing metrics: Complete, Average, Single and Ward. As the perspectives may imply, the “economical” perspective is focused on the spending behaviours of customers and aggregates the 'Products_per_orders', 'chain_per_orders', 'vendor_per_orders','cuisine_per_orders' and 'spent_per_orders' features. On the other hand, the “preference” perspective relates to the choices made by our customers i.e. what and when they order. It aggregates the "cuisine_variety", "order_range" ,"is_chain", 'Early_Morning','Breakfast','Lunch','Snack','Dinner' and 'Late_Night' features.

Analysis of figures 6 and 7, show that the best clustering approach based on the explained variance is K-Means. Using the elbow method, we’re able to further conclude that the optimal number of clusters for “economic_features” is 4 and for “preference_features” it’s 3. Given that we applied K-Means on these features, the number of clusters for our scaled data set ended up at 7 (3+4). To validate this claim, we produced a T-SNE graph as shown in figure 8.

3.2. Clustering Profile Approach

When creating marketing strategies aimed at boosting engagement and driving business profits, it’s essential to not only optimize but also carefully identify the target audience. The initial clustering solution, particularly one created without proper profiling, may not always be the best option. This is because clustering algorithms lack human intuition and do not factor in the costs associated with implementing specific marketing techniques.

For instance, while a clustering solution might appear optimized from a purely algorithmic perspective, when viewed through a business lens—considering key variables like economic value or preferential/behavioral patterns—the resulting clusters may end up being very similar. Treating these as distinct groups could lead to unnecessary costs that are not justified by the limited incremental benefits they provide. To avoid this, it is crucial to balance the technical optimization of clusters with practical business considerations to ensure the solution is both effective and cost-efficient.

Taking this into consideration, our solution takes into account two perspectives : economic value segmentation, and behaviour based segmentation. We chose to exclude the demographic perspective, firstly because the age distribution was almost identical across all clusters, with young adults representing more than 50% of the population in each, as shown in *Figure 11*. Moreover, we decided not to use customers' “region” because we do not have the corresponding zip codes for restaurants. Without this information, we cannot determine if a customer’s location falls within the service range of any restaurant, making it impossible to build strategies to target the customers’ location.

3.3. Economic Value Profiling

Our team began the cluster profiling process by focusing on economic value features, as we believe this is the most defining characteristic of a customer, given its direct correlation with revenue generation.

As shown in *Figure 9* , we initially identified seven clusters. However, from a marketing perspective, this number felt excessive. Some clusters were only slightly distinct from each other, and while merging them may sacrifice a bit of precision, we believe the trade-off is justified. Combining the clusters simplifies the strategy and avoids the need to create additional marketing techniques for marginal gains, such as correctly targeting, for instance, 5% more customers. Our merging methods to reduce the total number of clusters is detailed below with the overview of the final clusters.

3.3.1. Spender Segment

Analyzing the clusters based on the spender segment feature, we observe significant overlap in their characteristics. For instance, from *Figure 9* we can notice that clusters 1 and 6 have nearly identical distributions of spender segments and a similar number of customers. Similarly, clusters 0 and 5 show overlap, though cluster 0 has a slightly larger customer base and a higher proportion of top spenders. A comparable pattern is seen with clusters 2 and 4, where cluster 4 has fewer customers and a smaller percentage of higher spenders.

3.3.2. Average Spending Metrics

Looking at the average spending metrics (total spent, spent per order, and order range) in *Figure 10* , the similarities between clusters become even more evident. Clusters 1 and 6, for example, lack any order range and have comparable total and per-order spending. Before merging we constructed an auxiliary table (*Table 1*) where we did the profiling based on the economical features of the clusters before merging, and after that we merged them based on the similarities.

To align with the previous clustering logic and balance precision with simplicity, we opted to merge clusters 0 and 5, despite cluster 0 showing a slightly broader order range and higher total spending. The same rationale was applied to the merging of clusters 2 and 4.

After merging the clusters, and basing our analysis in the economic value segmentation we created a table to summarize and assign labels to each of the four resulting clusters.

Table 2 : Clusters, labels and economical features

Cluster	Label	Spender segment	Total spent	Spent per order	Order range	Nr. of customers	Initial clusters
1	Inactive customers	Low	Low	Low	0	5060	1 and 6
2	Occasional customers	Low, Mid-Low	Medium	Low	Low	6105	3

3	Regular customers	Mid-Low, Mid-High	Medium	Low	High	9442	0,4 and 5
4	High value customers	High, Mid-High	High	Low	High	3446	2

From *Table 2*, we can characterize the clusters as follows:

- Cluster 1 (Inactive Spenders): Customers with minimal engagement who placed only one order during the 3-month period and spent minimally. These customers have low total spending and low order frequency, making them a low-priority segment for marketing efforts.

- Cluster 2 (Occasional Spenders): Customers with moderate spending but a lower frequency of orders. They show sporadic engagement, with a mix of low and mid-low spending segments. This group has potential for targeted promotions to increase order frequency.

- Cluster 3 (Regular Spenders): Customers with consistent engagement, characterized by moderate total spending and a broader order range. They represent a mid-value segment that can be further targeted to increase loyalty and spending.

- Cluster 4 (High-Value Customers): Premium customers who exhibit high total spending and frequent orders. They consist of high and mid-high spenders and represent the most valuable segment for marketing strategies to retain loyalty.

3.4. Behaviour Based Profiling

After analyzing our customer base through economical features, which helped us understand their spending patterns and overall value, we decided to next characterize the obtained clusters based on behavioural features. We started by analyzing the last promotion used and payment method , and afterwards ordering patterns such as which cuisine each cluster ordered the most. Moreover, as we can see by *Figure 17* and *Figure 18* , the day of the week distribution and hourly distribution of orders across clusters is homogenous, so we decided not to use these two features when doing the profiling, as they do not bring any value to the clustering profiling.

3.4.1. Last Promotion Used

Promotions play a varying role in influencing customer behavior across clusters, with some clusters showing a balanced use while others rely less on offers, as *Figure 14* demonstrates.

In Cluster 1, delivery promotions (1,755) and no promotions (1,704) dominate, with discounts (973) and freebies (628) less common. Cluster 2 also favors no promotions (2,761), though delivery (1,585) has some appeal, while discounts (899) and freebies (860) lag. Cluster 3 is heavily skewed toward no promotions (5,441), but delivery (1,412), discounts (1,236), and freebies (1,353) engage smaller groups equally. Cluster 4 follows a similar pattern, with no promotions (2,110) leading and minimal interest in delivery (446), discounts (424), and freebies (466).

3.4.2. Payment Method

Payment methods show a consistent preference for cards across all clusters, as seen in *Figure 15*. In Cluster 1, card payments are the most popular, with a balanced usage of cash and digital payments. Cluster 2 shows a similar trend but has over 1,000 more card users compared to Cluster 1, with a slight preference for digital payments over cash. In Cluster 3, cards dominate overwhelmingly, with more than 6,000 users relying on this method. Cluster 4 also favors card payments, with digital payments surpassing cash usage, indicating a growing inclination toward modern payment options.

3.4.3. Cuisine Preferences

After segmenting our customers by the last promotion used and by the preferred payment method, which we consider a secondary characterization from a behaviour based point of view, our team decided to look at the key behaviour feature which is the cuisine preferences by cluster.

To help us analyze these preferences, our group did a bar plot as shown in *Figure 16* that shows the % of customers that each cuisine has, in relation to the total number of customers of that particular cluster and also the high + mid-high segment, which is the spender segment that brings more revenue to the company.

Starting with Cluster 1, customers show a clear preference for American cuisine, which accounts for 21% of the total customers and holds the highest share of high and mid-high spenders among all cuisines (7.5%). Asian cuisine follows as the second most popular option, although its representation among high spenders is slightly smaller. Japanese cuisine, beverages, and other options have some presence, but their contribution to high spenders remains modest. Meanwhile, snack options like Café and Desserts, along with healthier choices such as Healthy, have minimal representation in both total customers and high spenders.

Moving to Cluster 2, there is a strong preference for Asian cuisine, which makes up 35% of total customers and 12.2% of high and mid-high spenders. American and Italian cuisines also play substantial preferences, representing 25% and 18.7% of total customers, with high spender percentages of 8.8% and 10.4%, respectively. Beverages attract 14.5% of total customers, but the share of high spenders for this category remains below 10%. Snack options like Café and healthier alternatives such as Healthy continue to have a smaller presence.

Cluster 3 also represents the alignment between total customers and high spenders, particularly for Asian cuisine, which represents 36.9% of total customers and 20.2% of high and mid-high spenders. Western cuisines like American and Italian follow closely, accounting for 27.2% and 22.8% of total customers, with high spender percentages of 14.2% and 14.1%, respectively. The "Other" category captures 13.9% of total customers. Beverages and snack cuisines, while present, have high spender percentages below 10%, suggesting a potential not so valuable targeting on these cuisines.

Finally, Cluster 4 demonstrates a strong preference for American cuisine, which dominates with 46% of total customers and 24.8% of high and mid-high spenders, making it a prime focus for targeting high-value customers. Asian cuisine, particularly Japanese, also is a significant preference, representing 39.5% of total customers and 18.4% of high spenders. Italian cuisine shows a balanced contribution, accounting for 22.8% of total customers and 10.4% of high spenders. The "Other" category is also noteworthy, with 29.2% of total customers and 14.6% of high spenders. Additionally, beverages are relatively more popular in this cluster, capturing 18.9% of total customers, though their representation among high spenders remains smaller.

3.5. Final Profiling

In order to merge the economical perspective and the behaviour perspective, we decided to build a table, with the main findings, so that we can present the final results to the marketing team, that later on will build the strategies to target the different clusters.

Table 3 : Clusters, labels and relevant economical and behaviour based features

Cluster	Label	Total Spent	Top 3 cuisines	Top 3 High spending cuisines	Last promotion	Preferred payment method
1	Inactive customers	Low	American (21%) Asian (18%) Other(13,4%)	American (7,9%) Asian (6,9%) Beverages(5,7%)	Delivery	Card
2	Occasional customers	Meidum	Asian (35%) American(25%) Other (18,8%)	Asian (12,3%) American (8,8%) Beverages(6,3%)	-	Card
3	Regular customers	Medium	Asian (36,9%) American(27,2%) Italian (22,8%)	Asian (20,3%) American(14,3%) Beverages(14,1%)	-	Card
4	High value customers	High	American (46%) Asian(39,5%) Other (29,2%)	American(24,8%) Asian(18,4%) Other(14,6%)	-	Card

4. BUSINESS APPLICATIONS

Based on our final profiling, ABCDEats should focus its targeting efforts mostly on cluster 4 and 3, as these segments offer the highest potential for revenue growth. When evaluating the potential of the other two clusters, it becomes clear that focusing on them would be less effective for ABCDEats. Cluster 1 is made up of inactive customers who have shown little interest in placing orders, making it costly and time-consuming to try and re-engage them with no guarantee of success. Similarly, Cluster 2, though slightly more active, consists of occasional customers whose sporadic engagement and low spending make them less impactful. Targeting these groups would require significant marketing efforts, which may not deliver enough return to justify the investment.

A potential marketing strategy would best focus on targeting the number of times a customer orders in these groups if company resources aren't a concern. Potentially having promotions / deals on specific high ordering times/days could help capture the audience in clusters 1 and 2.

Cluster 4, composed of high-value customers, leads in both total spending and engagement, particularly with American and Asian cuisines. As these two cuisines have a high percentage of high spenders, ABCDEats should introduce early personalized discounts or cashback rewards, particularly for these cuisines. Additionally, introducing a tiered rewards system could further incentivize repeat orders, where higher spending unlocks perks like free delivery, premium packaging, or access to exclusive dining experiences.

Similarly for Cluster 3, which represents a large base of regular customers, strategies should focus on increasing their spending frequency and basket size. ABCDEats could introduce family bundle deals or combo promotions that cater to their preferences for American and Asian cuisines. To drive engagement, targeted campaigns like "Order 3 Times This Week and Get Free Delivery" or discounts during off-peak hours could be effective.

5. CONCLUSION

To conclude, our preprocessing efforts that expanded past our original EDA report included finding a way to reduce the total number of outliers and features with sparse variance as they were resulting in erratic clusters. This was solved through replacing certain features with ratios based on the number of orders a customer placed as well as transforming old features into new categorical features that placed value in having not placed orders in certain categories as much as placing the order. This resulted in clusters that were evenly partitioned and more well defined when utilizing kMeans. From the resulting clusters, we further analyzed and merged the clusters once more based on 2 different customer perspectives: behavioral and economical. Using this, we found 3 of the original 7 clusters to be similar and found it would be more resourceful and economical to merge the similar clusters as this would help ABCDEats target the majority while conserving resources. With the 4 final clusters, we determined that

each had different spending and ordering habits (low spending and low ordering versus high spending and low ordering, etc…) which could help the ABCDEats marketing team determine which group they want to target if all four aren't feasible depending on company resources. Further, we found the top cuisines that each group prefers so targeting promotions based on the cuisines that the majority of each group spends on (limited to the top 3 cuisines) could help improve customer engagement, especially ones based around their preferred payment methods. Business applications suggested from these findings included targeting primarily cluster 3 and 4 as 1 and 2 had extremely low ordering habits which suggests general uninterest in using the application or that they only used the application for very specific circumstances. Clusters 3 and 4 spent more and ordered more suggesting that they find ABCDEats useful and sending them promotions that place discounts on loyalty or package deals based on their preferred cuisines would help persuade them to use ABCDEats more. These findings provide clear strategies for ABCDEats to target clusters effectively and optimize resources. Tailored promotions based on cluster preferences can boost customer engagement and drive growth.

APPENDIX A

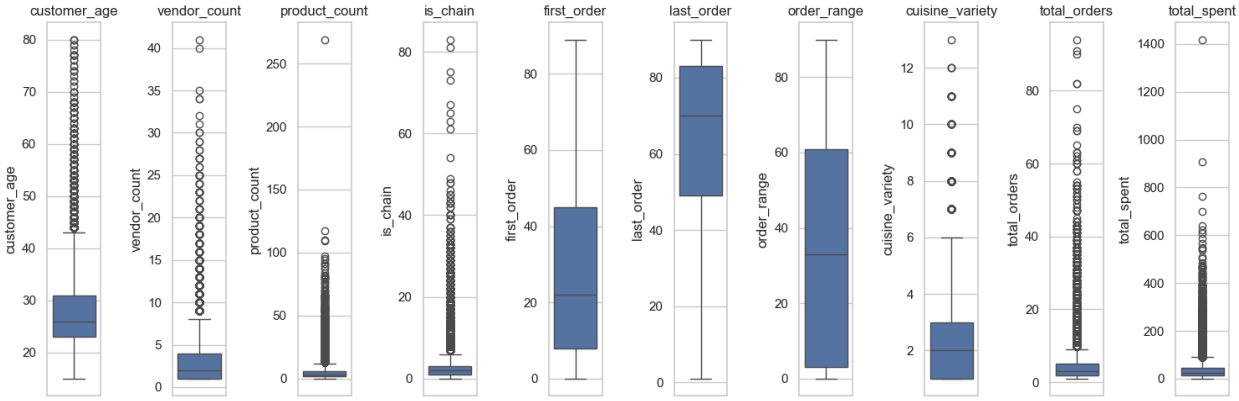

Figure 1 - Boxplots of the data before outlier removal

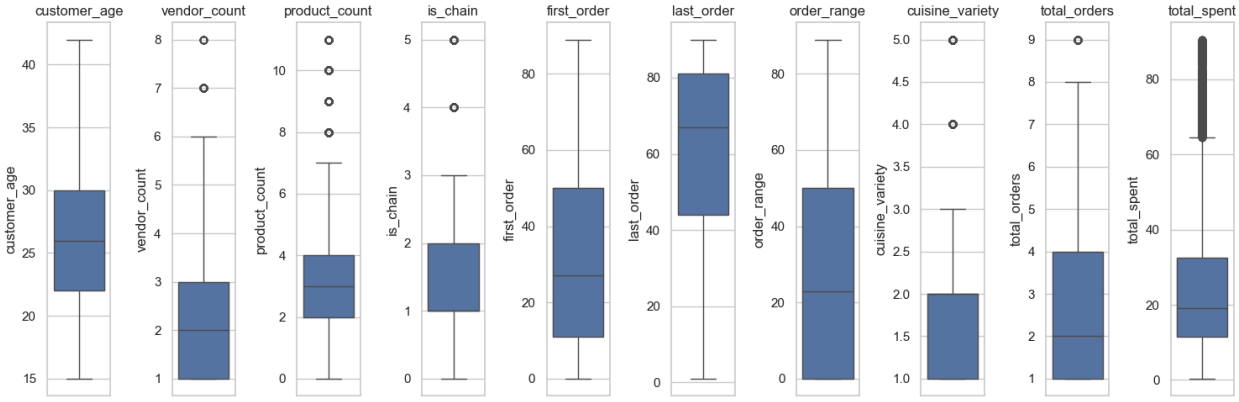

Figure 2 - Boxplots of the data after outlier removal

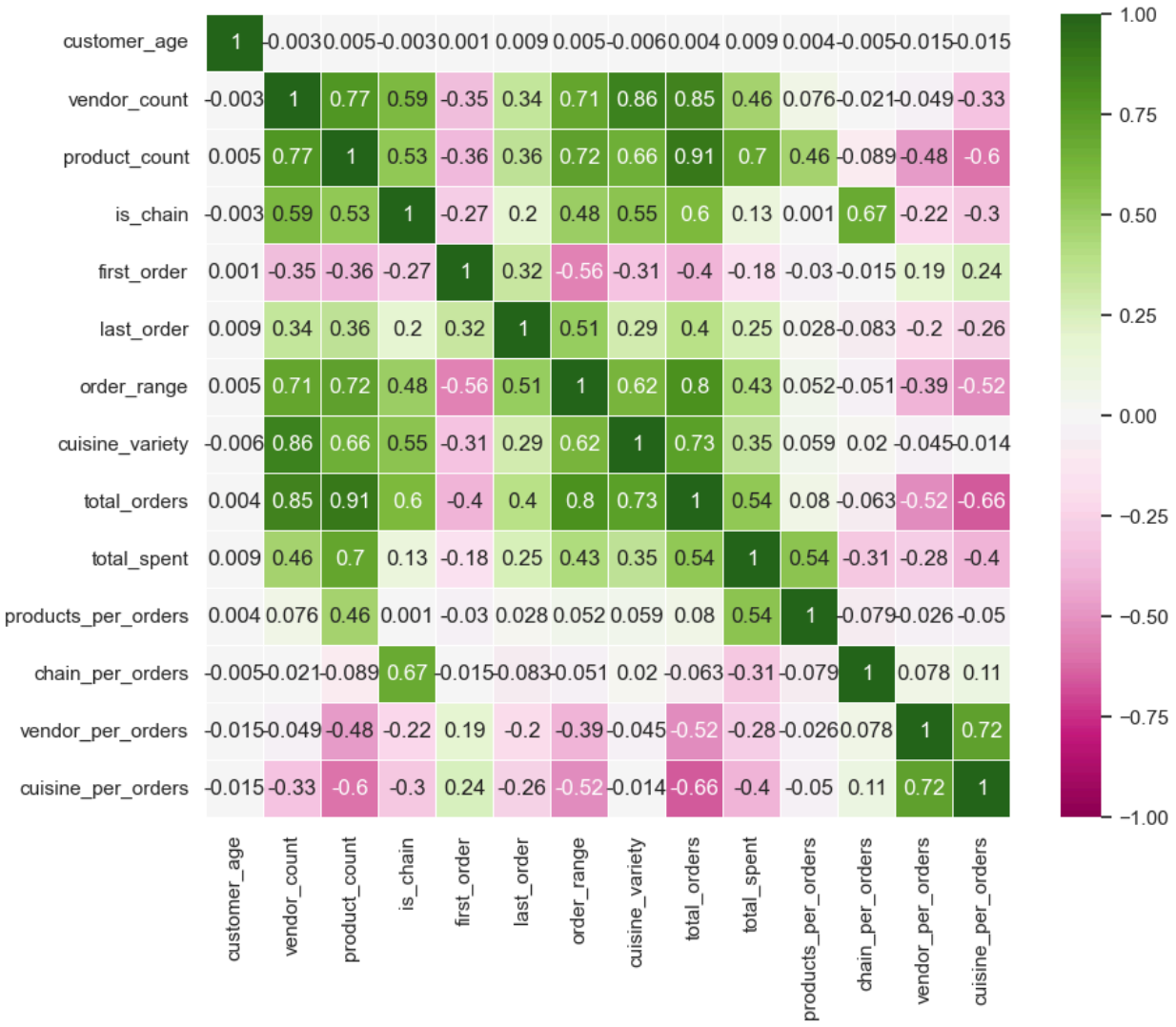

Figure 3 - Spearman Graph

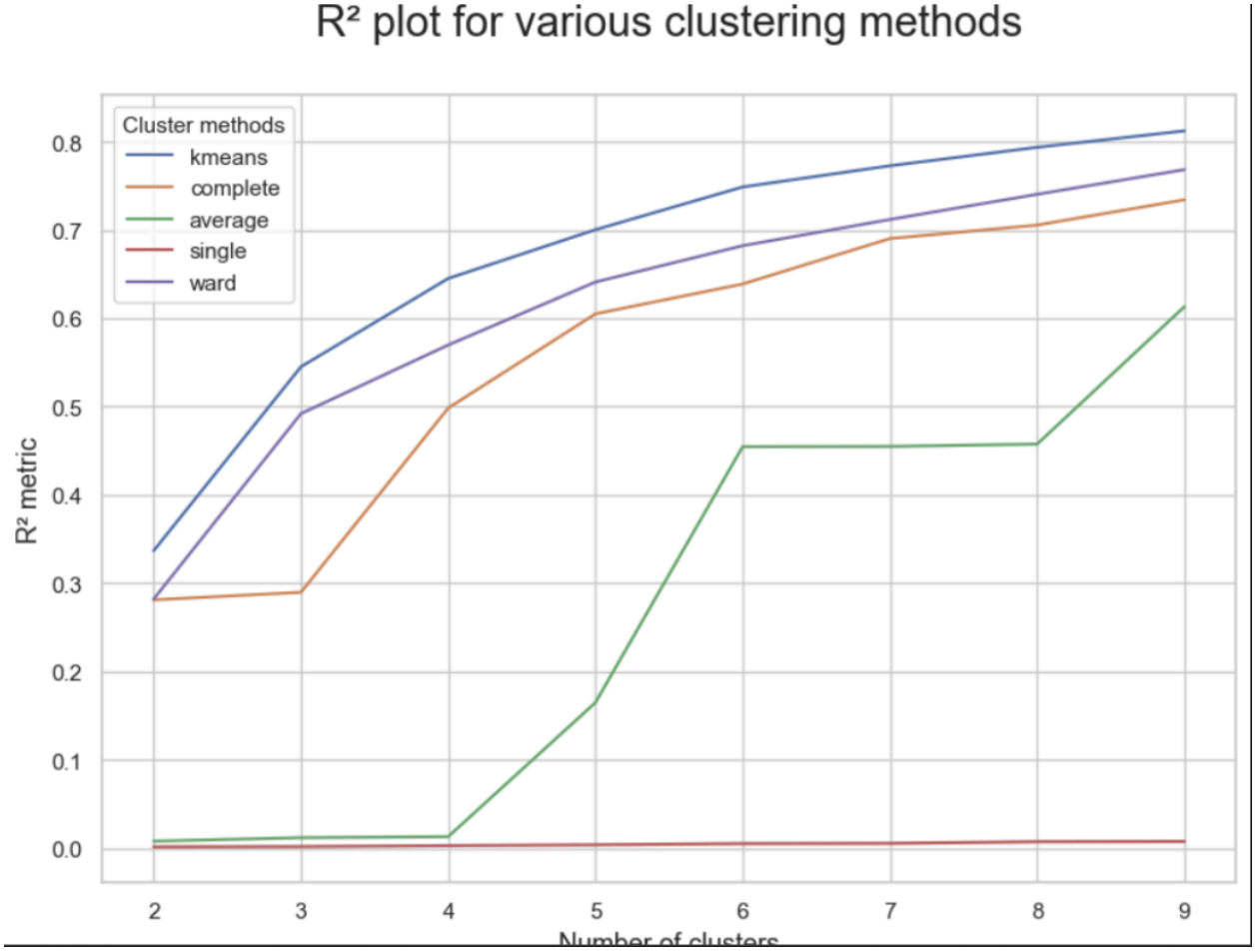

Figure 4 - R^2 plot for "economic variables" on a dataset with outliers

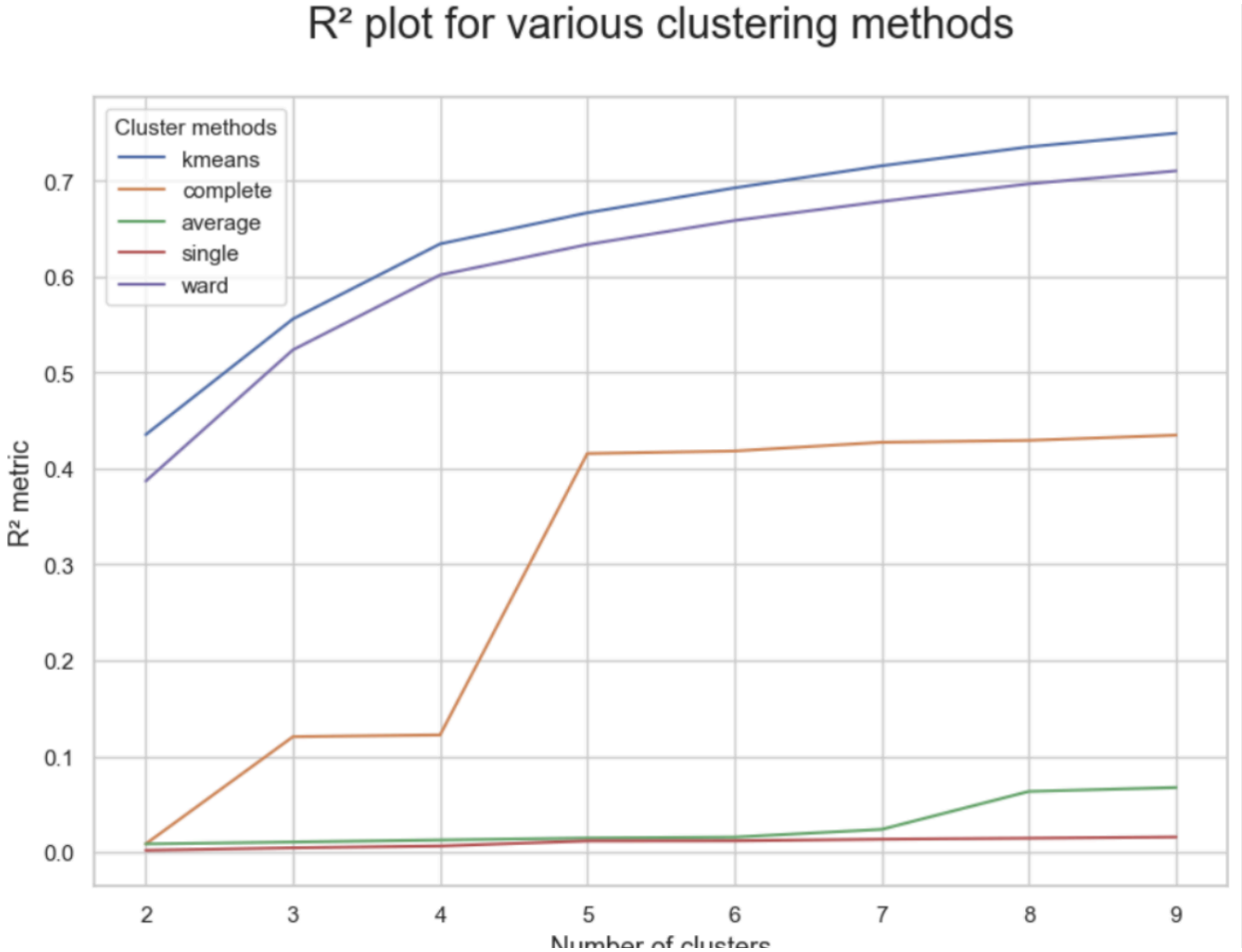

Figure 5 - R^2 plots for "preference variables" on a dataset with outliers

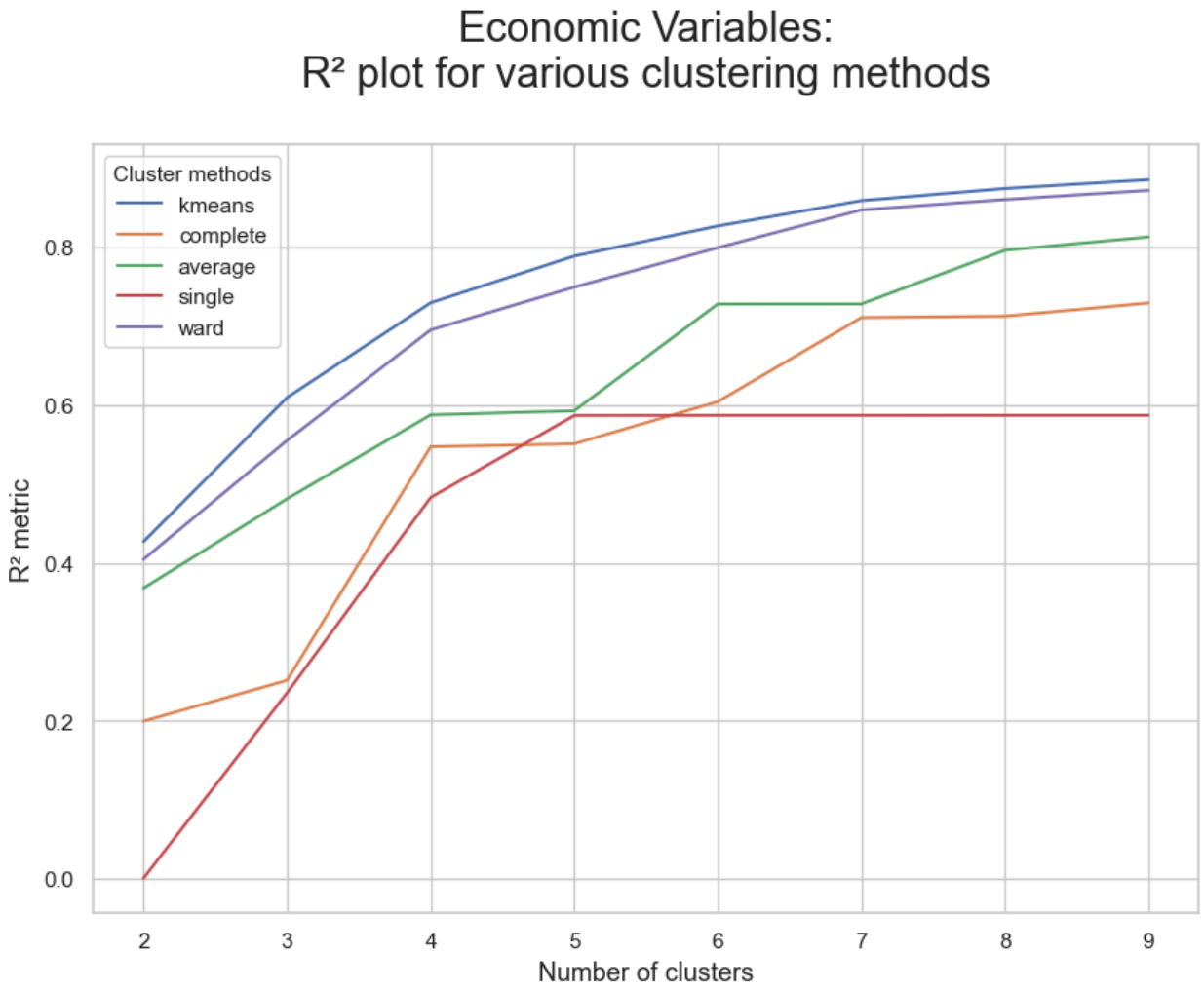

Figure 6 - R^2 plots for "economic variables" on a dataset without outliers

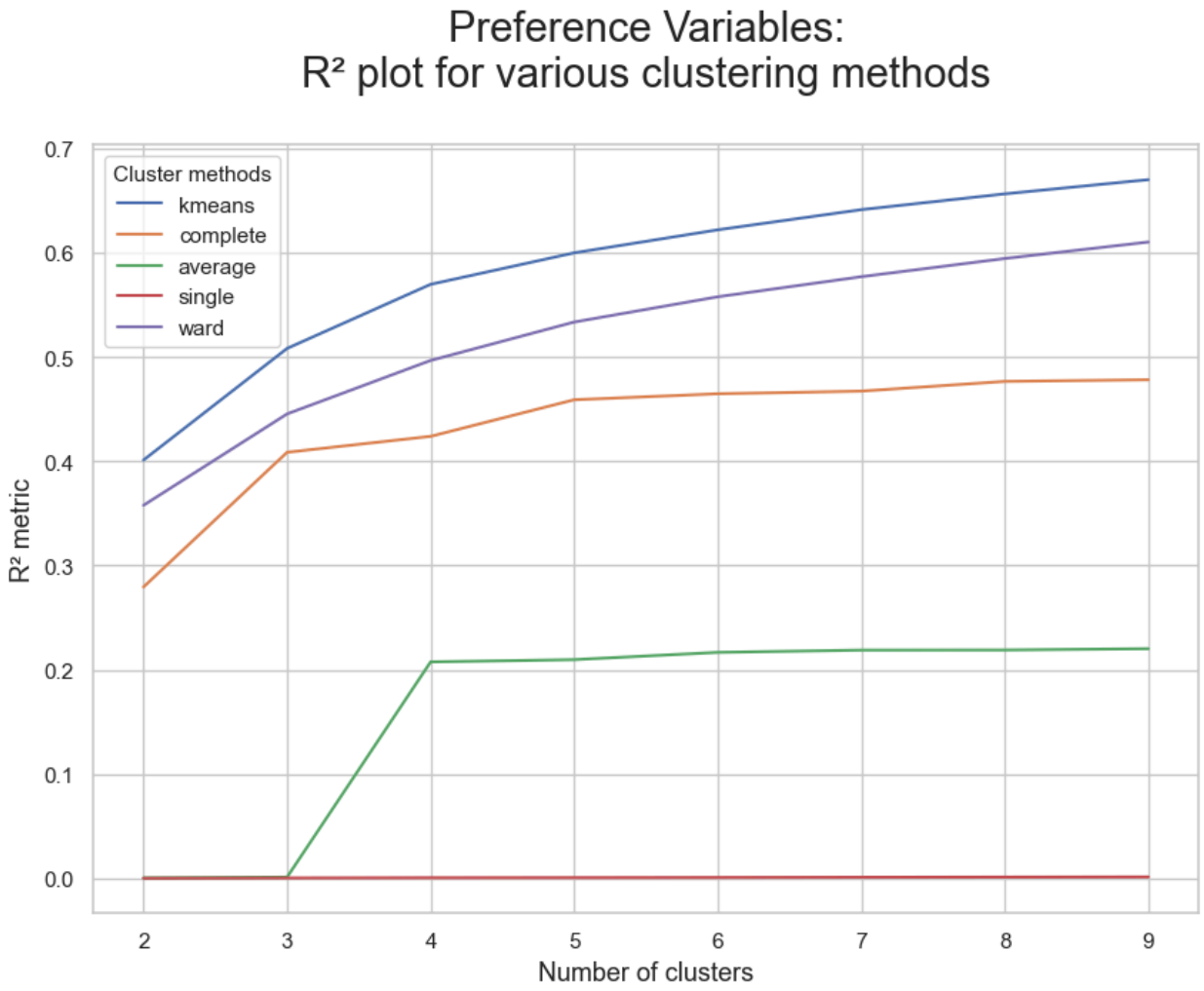

Figure 7 - R^2 plots for "preference variables" on a dataset without outliers

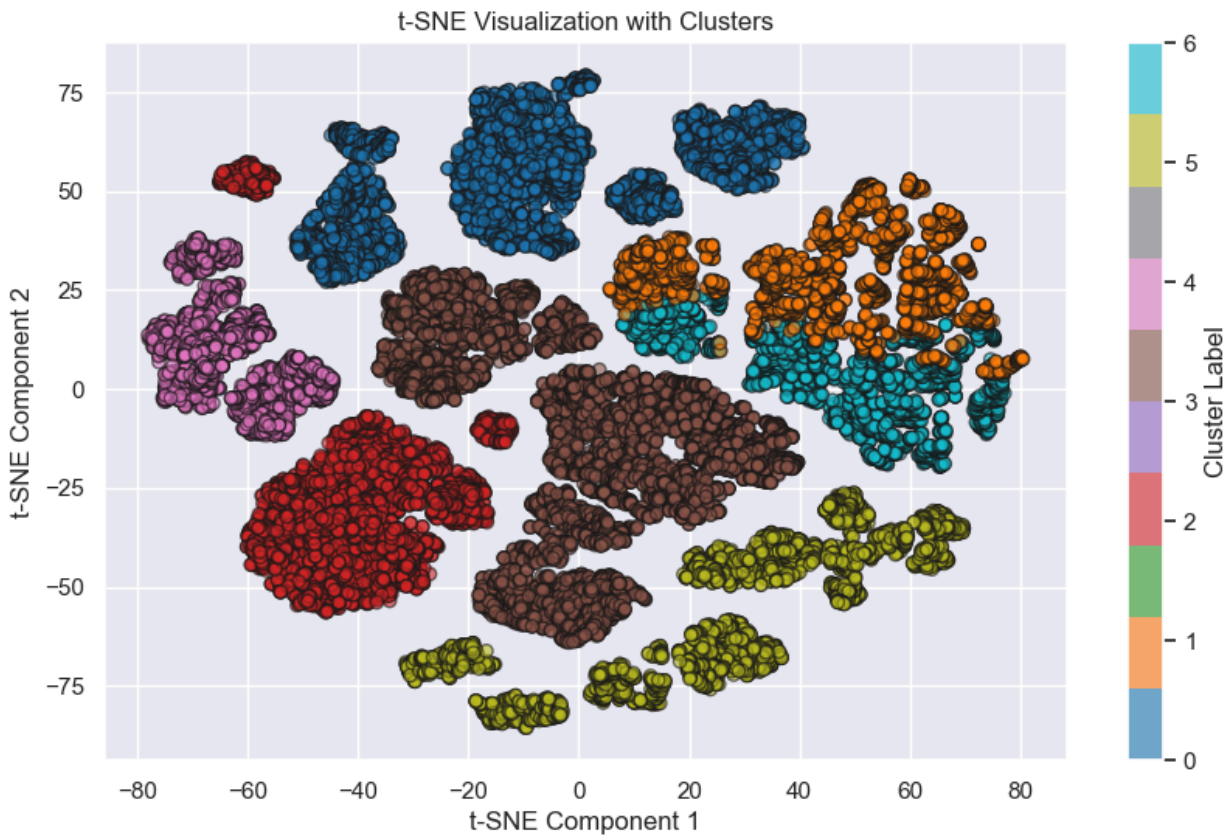

Figure 8 - T-SNE plot with 2 components

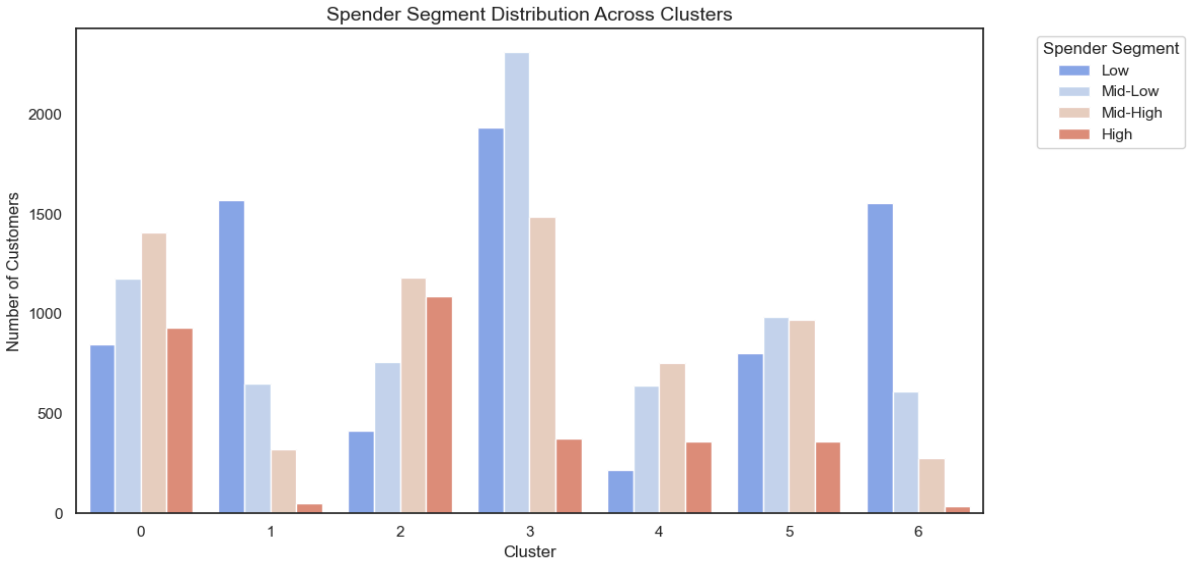

Figure 9 - Initial clustering solution for spender segment feature

Figure 10 - Initial clustering solution for average spending metrics

Cluster	Spender Segment	Total Spent	Spent per order	Order range
0	Medium	Low	Low	Medium
1	Low	Medium	Low	0
2	High, Mid High	High	Low	High
3	Low, Mid Low	Low	Low	Low
4	Medium	Medium	Low	High
5	Medium	Medium	Low	Medium
6	Low	Low	Low	0

Table 1 - Initial clustering solution (Spender segment and average spending metrics)

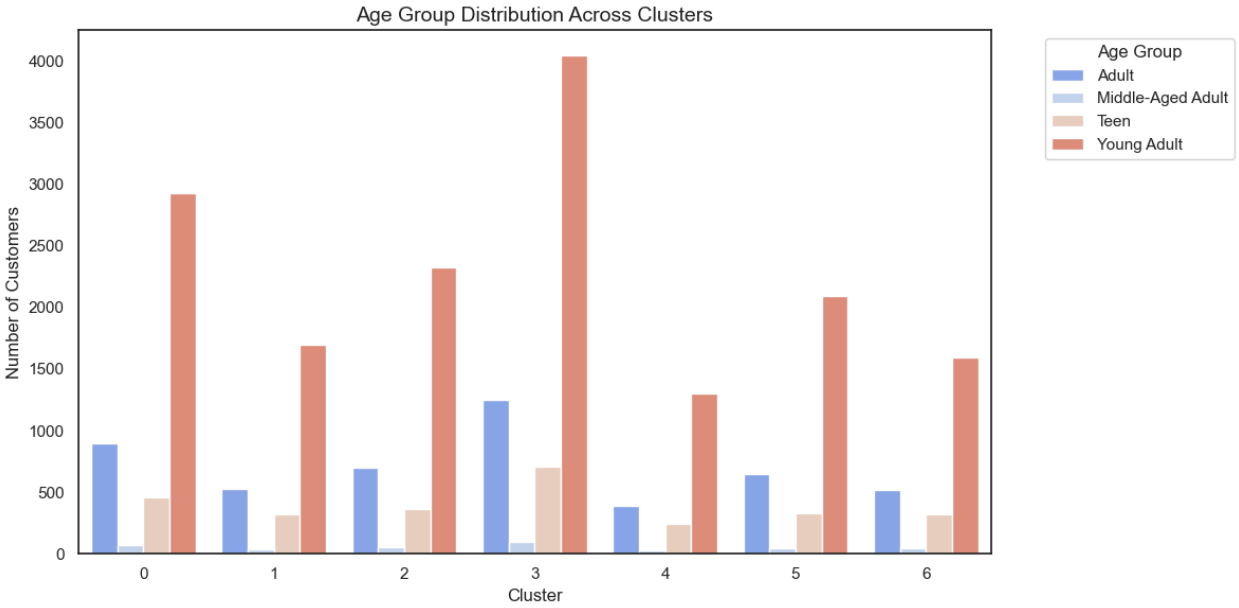

Figure 11 - Initial clustering solution for age distribution

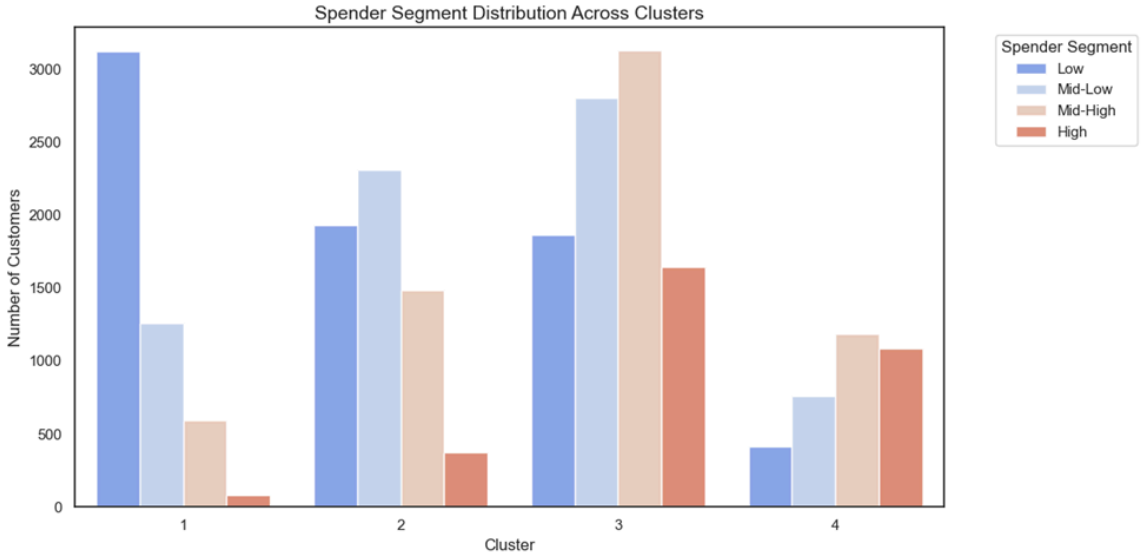

Figure 12 - Final clustering solution for spender segment feature

Figure 13 - Final clustering solution for average spending metrics

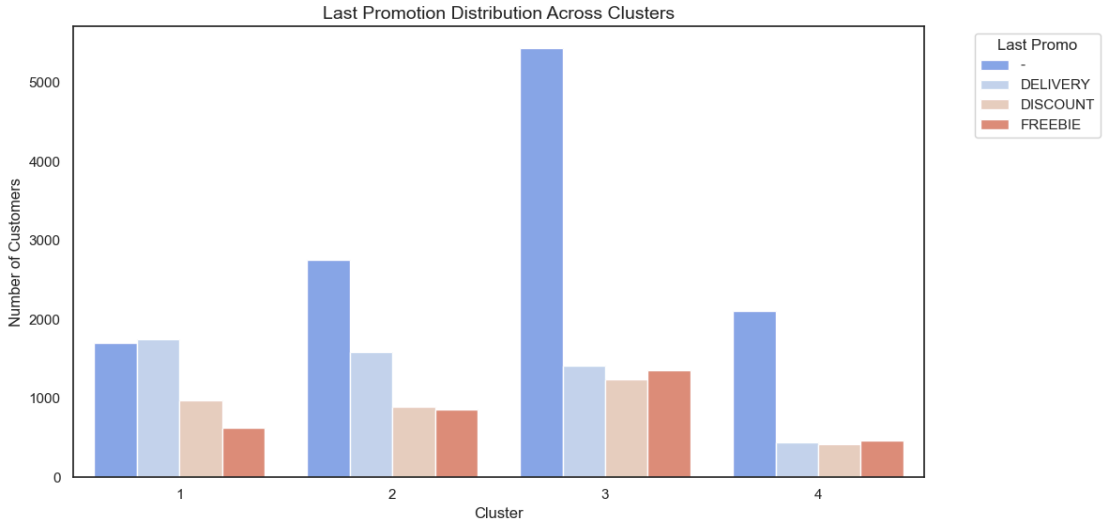

Figure 14 - Final clustering solution for last promotion

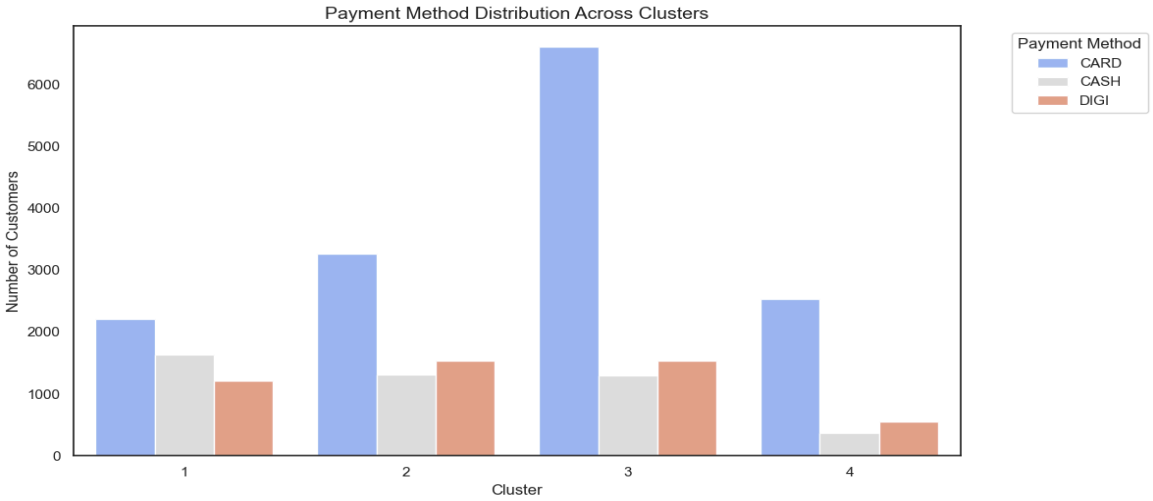

Figure 15 - Final clustering solution for payment method

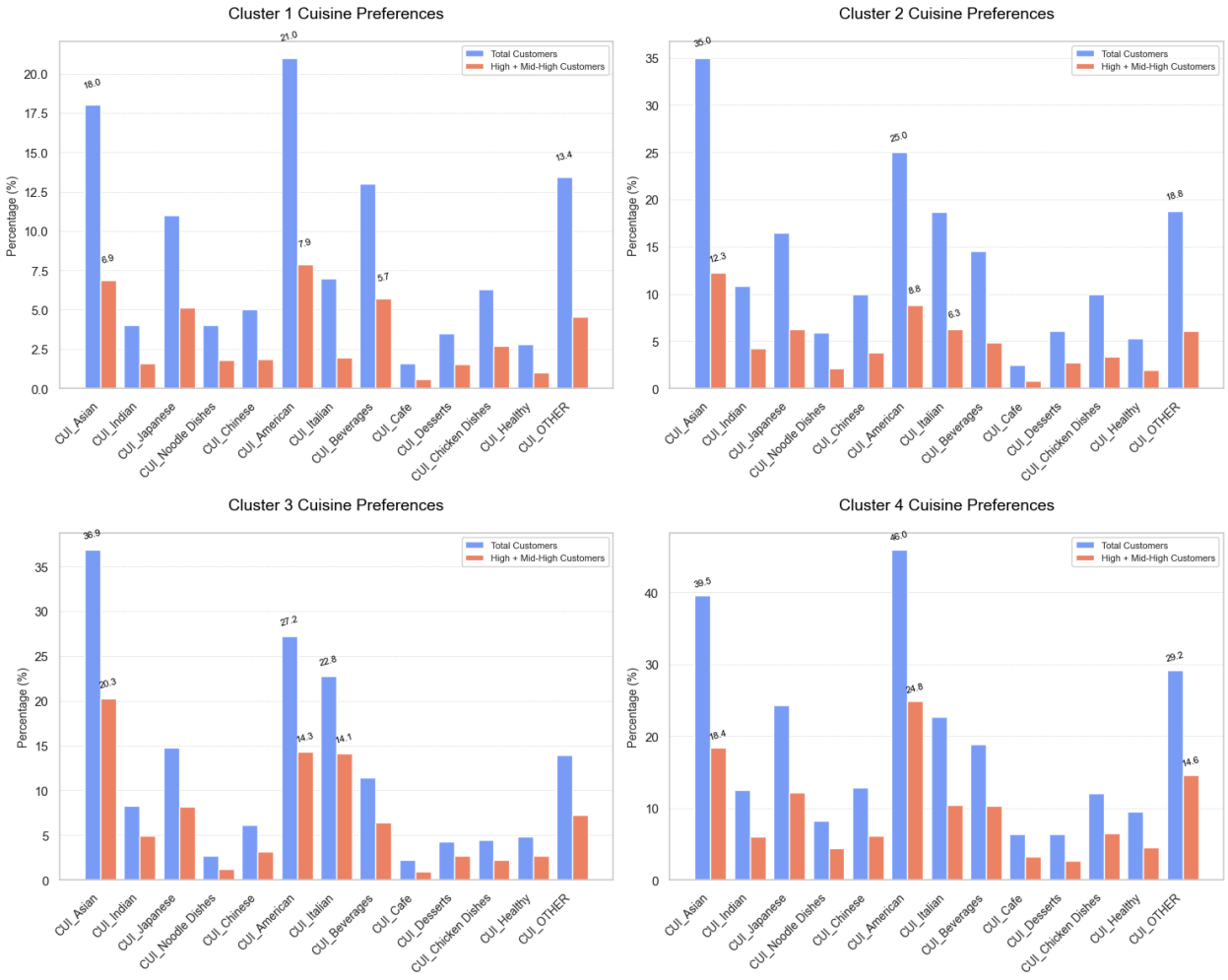

Figure 16 - Cuisine preferences by Cluster (Total customers and High/Mid-High segment in % of the total customers in that cluster)

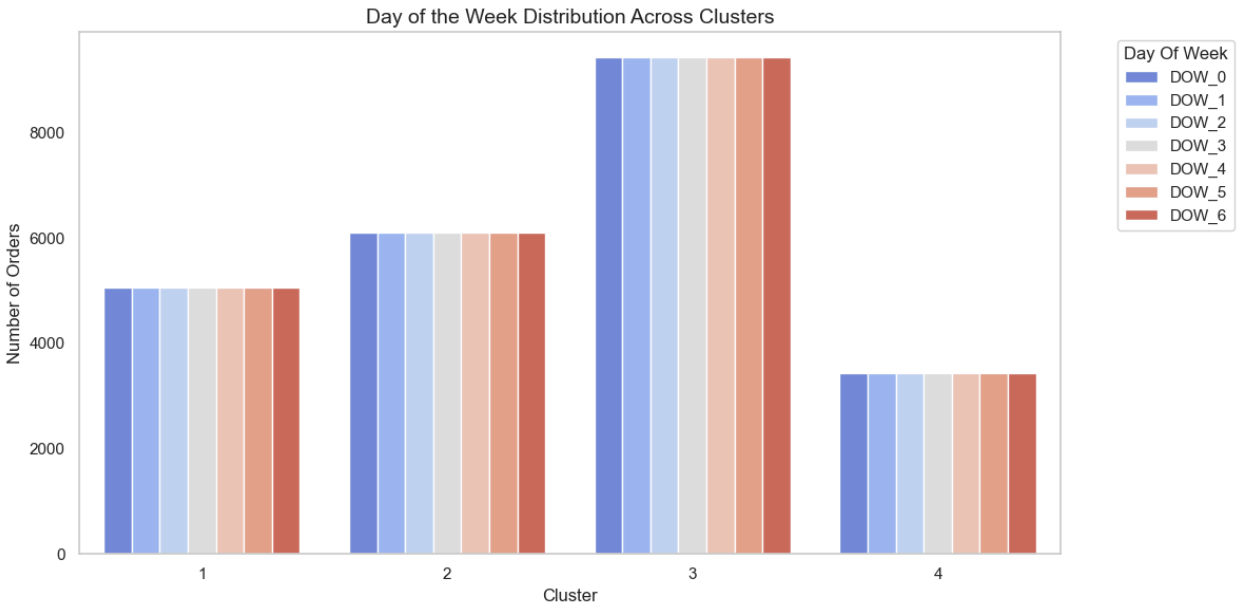

Figure 17 : Day of the week distribution of orders per cluster

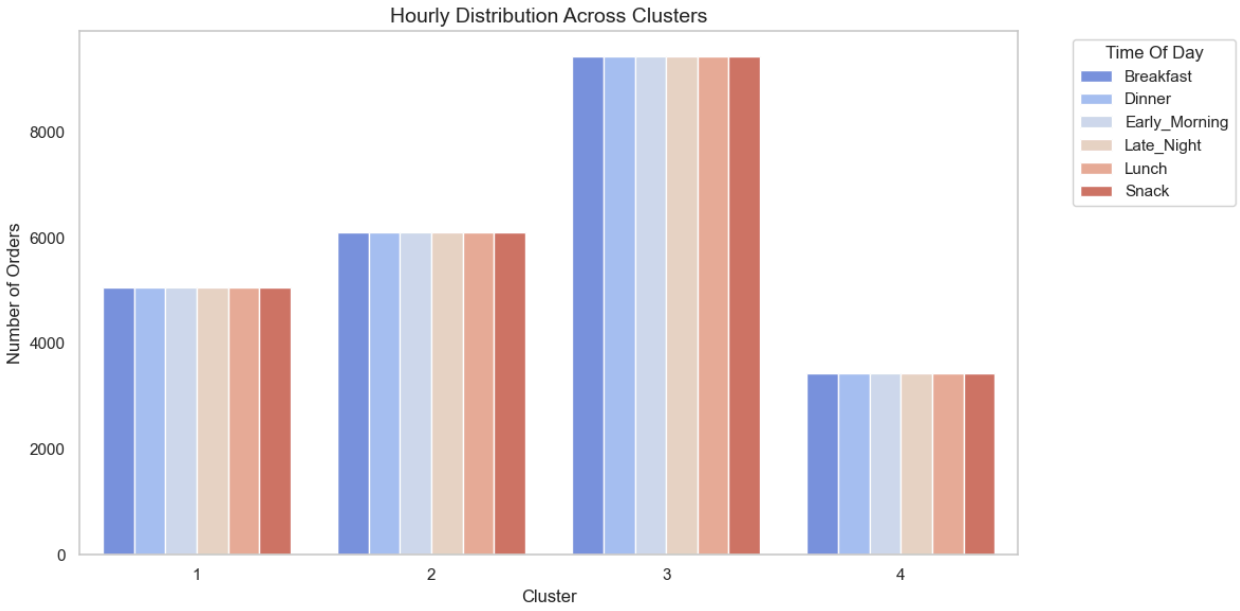

Figure 18 : Hourly distribution of orders per cluster