DATA MINING PROJECT

Master in Data Science and Advanced Analytics

NOVA Information Management School

Universidade Nova de Lisboa

Customer segmentation of ABDEats

Group 24

Carolina Pinto, 20240494

Iris Moreira, 20240659

Francisco Pontes, 20211583

Maria Margarida Cardoso, 20240493

Fall Semester 2024-2025

TABLE OF CONTENTS

1. INTRODUCTION

The first part of the report will be feature-based. For each feature we will: summarise key statistics for the data, discuss possible implications, identify trends, patterns and anomalies and discuss methodology for missing values and outlier treatments. We will then explore relationships between features and create new features. All visualizations are included in the appendix for easy reference.

2. IN-DEPTH EXPLORATION OF THE DATASET

2.1.1. Categorical Variables

2.1.1.1. Customer_id:

Based on the total count and the number of unique customer_id's observations we identified 13 duplicates with identical values across other columns, indicating true duplicates. We should drop them to avoid misleading information. This feature should become the index, as it serves as a unique identifier for each customer and statistical operations with its values are meaningless.

2.1.1.2. Customer_region:

According to Figure 1, there are a total of 9 regions, with one region recorded as '-' (1,39% of records). It seems reasonable to assume that customers are distributed across 3 broader regions: 2, 4 and 8, as all the regions that start with the same number are very similarly distributed when compared with other features. Instead of treating '-' as a missing value we identified a similar behavior to the ones starting with the number 8, so we will assign it to this broader region (Figures 21, 33, 36).

2.1.1.3. Last_promo:

There are 4 types of promotions, with one labeled as '-', which appears in over half of the observations (52.5%), likely indicating non-use of promotions (Figure 1). Our hypothesis is that promotions are either badly advertised or are available briefly, leading to low usage. Among the labeled promotions Delivery is the most popular, though usage differences among the other promotions are relatively small. This suggests that last promotion may not strongly differentiate customer behaviors.

2.1.1.4. Payment_method:

There are 3 types of payment methods. The most common is CARD, that represents 63,3% of the observations. Digital payments make up 19,1%, and together with CARD, 82.3% of customers use electronic payment methods. This aligns with the fact that the average customer age is 27.5 years, an age group more likely to favor electronic payment methods due to their widespread adoption, although older users also make a similar use of digital payments (Figures 1, 2). So, independently of age, our customers prefer electronic payment methods, which proves an opportunity for digital engagement strategies.

2.1.2. Numerical Variables

2.1.2.1. Customer_age

The average customer's age is 27, suggesting that most of the customer base is relatively young. Age quantiles show that: 50% of the customers are 26 years old or younger and 75% of the customers are 31 years old or younger, with 63,5% being in their twenties (Figure 3). The maximum age of 80 years old appears to be an outlier.

There are missing values in 2,28% of the age column, which could be filled in with the median due to the narrow age range of the majority. In fact, according to Figure 4, people over 40 years old (3% of the dataset) are outliers. But looking at the pairplots (Figure 22), older customers don't seem to be severe multivariate outliers. We could remove them and proceed only with the younger customers, keep them and try to develop a marketing strategy attractive to older customers or define a threshold to reduce the impact of the outliers.

2.1.2.2. Vendor_count and product_count

The vendor_count (Figure 5) median value is 2, and 75% of the customers ordered from 4 or less different vendors. The maximum value 41 is clearly an outlier, which may be skewing the mean to approximately 3. This indicates that customers develop loyalty towards a specific vendor or a small group of vendors.

The average product count is 5.67, with the median being 3. The high variability of this feature, confirmed with Figure 6, suggests that while the majority only bought a few products, there is a group of customers that purchased a significantly larger quantity.

In 18 rows, the product_count is shown to be less than the total number of orders, which is impossible since each order must include at least a product. In these rows, product_count and is_chain are zero, and vendor_count is 1, while other variables hold meaningful values, indicating likely incorrect storing of values. Dropping these rows is the easiest solution. Furthermore, there are 138 rows where product_count, vendor_count and the sum across DOW are zero, with first_order being the same as last_order, possibly representing customers who didn't complete an order or an error in the storage of the database. We will likely drop the rows.

The outliers of vendor_count (Figure 5) seem to be relevant until the threshold of about 35 vendors, when they start to become sparser. Looking at the pairplots (Figure 22), we see that in the row of vendor_count, there are always 2 customers that appear farther apart and are clearly multivariate outliers, so we will apply a threshold. There is also an evident outlier for product_count, that appears in the pairplots as well. After the value 100, the outliers start to become sparser, so we will define that as the threshold and drop those.

2.1.2.3. Is chain:

According to the metadata, is_chain appears to be a binary variable, but the statistics tell a different story. The mean is 2.81 and the median is 2, what could indicate that the variable reflects the number of times the customer has ordered from a chain restaurant. Considering this interpretation, 75% of customers ordered from 3 or less chain restaurants, with an outlier that reaches 83 orders from a chain. This means that most customers order little times from chain restaurants, but there is a small group of customers that do it often. Looking at Figure 7, the outliers start to become sparser after the value 45. The pairplots (Figure 22) show some multivariate outliers, so we will likely apply a threshold of about 45.

When comparing the total orders to the is_chain values, we find 75 rows where is_chain exceeds the total number of orders. In those rows vendor_count and product count are zero, and the sum across DOW is also zero. If we drop the incoherences with product_count and vendor_count we are simultaneously treating is chain incoherences as well, allowing chain segmentation.

2.1.2.4. First order and last order:

Both first_order and last_order (Figures 8, 9) have a maximum value of 90, which is consistent with the three-month data collection period. The last_order median and mean exceed those of first_order, meaning that most customers placed multiple orders. With a median first_order of 22, half of the customers placed their first order in the first 22 days of the dataset, what shows strong early engagement. The 75th percentile for first_order being 45, suggests that 25% of customers placed first orders in the last month and and half of data collection. Last order equaling first order implies a single day activity.

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.

In the boxplot, there are no outliers for these two variables. Moreover, the boxplots are symmetric, meaning that there are a significant number of customers that used the app more than once (Figures 8, 9).

2.1.2.5. CUI (15):

None of the CUI features have missing values. Most of them, except for American and Asian, have all their quantiles equal to zero, with only the 75th percentile being non-zero for American and Asian cuisines (the ones with most expenditure) (Figure 14). This suggests a high level of customer segmentation based on cuisine preferences. The high maximum values for each cuisine further support this observation, indicating that some customers consistently order from one cuisine type. Therefore, the extreme values in the cuisine features seem to be relevant and we will keep them (Figures 25, 26).

The cuisines with the lowest expenditure are Cafes, Chicken Dishes and Desserts, which makes sense since they are usually cheaper items (Figure 14). For geographic coherence, the sum of CUI Thai, Chinese and Japanese should be at least less or equal to CUI Asian, but that is not the case.

2.1.2.6. DOW:

The days of the week have all their quantiles equal to zero, except for the 75% quantile that is consistently equal to one. The maximum values for these features are very close to each other, indicating that even though some customers order more often, they are relatively consistent across the week. We have less orders on Sunday and Monday and more orders on Thursday and Saturday, indicating that orders tend to increase as the week goes by, peaking at the end of the week (Figure 15).

Outliers are present in each DOW, likely due to customer segmentation based on ordering times, making these extreme values relevant for analysis. Looking at the pairplots (Figure 24), to compare the DOW with the other numerical features, we understand that there are some severe outliers. Adjusting the threshold of the outliers to when they start to become sparser in the boxplot may help improve segmentation and identify high-activity customers more accurately.

2.1.2.7. HR (24):

All hours have their quantiles equal to zero, indicating customer distinguishing based on the specific times they place orders. The maximum values of order counts per hour are higher around 11 and 17 o'clock (Figure 13). HR_0 is the only hour that has missing values, and all non-missing values are zero. To address this, we will compare the sum of orders across DOW (days of the week) with the sum of orders across HR (hours) to impute the missing values.

The outliers in these features may occur because we have small slots of time (hours) to describe the parts of the day. When grouping them we may significantly reduce the number of outliers and target peak times effectively.

2.2. Multivariate Analysis

Regarding the relationships between features, neither last_promo, payment_method nor customer region offer good insights since when plotted 1:1 with customer_age, last_order, first_order, vendor_count, product_count and is_chain, the distribution patterns across different categories are very similar within each group (Figures 16, 17, 18, 19, 20). Different payment methods and promos also show no evident pattern in customer segmentation by cuisine. When plotted against each other, categorical variables are not helpful in pattern detection (Figures not presented in the appendix are in sections 3.4.1. and 3.4.2. of the Jupyter Notebook and not included to avoid overloading with information not as insightful).

On the other hand, region is well differentiated with cuisine expenditure and time of day of ordering. Customer regions that belong to the same broader region have similar cuisine expenditure and order placing hours patterns. It is important to notice that region '-' has a similar bar chart distribution to the other bar charts that represent the broader region 8 (Figures 21, 33, 34, 35, 36, 37).

From the pairplots of numerical features and correlation matrices (Figure 22, 27), we see that vendor_count, product_count and is_chain have a high linear correlation, implying that customers that order from multiple vendors consequently order more products and customers that order more frequently from chain restaurants also buy more products and vary vendors more. In fact, an increased number of unique vendors doesn't increase proportionally the count of products bought. So, a client who buys from a lot of vendors is not buying a lot from each but is rather a diversified customer. While we understand that information from vendor_count, product_count and is_chain is redundant, we will only drop features after feature engineering, to not diminish the range of potential new variables.

Furthermore, when last order is high, i.e., when the last order the customer placed was done in the last days of the data set, the product count and the vendor count have higher values as well (Figure 22, 27). This might indicate that customers with a high last order are frequent customers and order from a wider range of vendors. The symmetry between last order and first order distributions mentioned priorly is also evident. According to Figure 23, we can distinguish two types of customers: the customers represented in the superior left corner of the graph are the ones that place a first order in the start of data collection and the last in last few days of data collection, this means they are active for a long period of time. On the other hand, we have customers in the diagonal that are single-day users. There's a negative relationship between customer_age and vendor_count, suggesting that younger customers tend to order from a higher variety of vendors (Figure 22).

2.3. New Features

To deepen our analysis, we will create new features. Firstly, we plan to group variables to reduce dimensionality. Based on Figure 10, we see that the total number of orders peaks at 11 AM and 5PM. Instead of aggregating the **hour variables** by standard eating times, we will consider the busiest hours and the least busy, while associating them to a more personalized period (Figure 13). We will group features by the following distribution: 2:00 to 7:00, 8:00 to 13:00, 14:00 to 19:00, and 20:00 to 1:00. The correlation matrix also shows stronger correlations within consecutive hours, especially in the described peaking slots (Figure 30).

We also plan to aggregate the 7 **weekly** variables into Weekday (Monday to Thursday), and Weekend (Friday to Sunday). We include Friday in the weekend category to have a more balanced distribution. This decision is because the correlation matrix (Figure 29) between days of the week doesn't show a specific relationship between the days of the week other than sequential correlation. Based on the Last_Promo we plan to create used_promo. It will behave as a binary, assuming 0 if Last_Promo is '-', meaning that the client never used a promotion, and 1 if otherwise.

Based on customer **region** we will create broader_region, that aggregates regions that start with 2, 4 and 8, as discussed above in 2.1.1.2. Region '-' is aggregated within broader region 8 (Figures 34, 35, 37).

Apart from grouping these variables, we also plan to create other features, such as:

- **Total_spending** (Sum of all CUI): total expense
- **Total_orders** (Sum of all DOW): total number of orders placed, to understand frequency
- **Avg_product_cost** (total_spending/product_count): average expenditure per product, to understand price-based preferences
- **Avg_order_cost** (total_spending/total_orders): average expenditure per order
- **Order_span** (last_order – first_order): for how long the customer has been placing orders, to understand customer loyalty patterns
- **Order_frequency** ((last_order – first_order)/total_orders): average number of days between orders, to understand engagement regularity
- **Products_per_vendor** (product_count/vendor_count): average products ordered from each vendor
- **CUI_diversity_score** (if CUI_...>0, add 1): number of unique cuisines ordered by customer, to understand variety preference.

Finally, we thought about grouping the features of cuisines, however, after checking the correlation matrix (Figure 28), and looking at the pairplots (Figures 25, 26) we don't find any linear correlations between these features. In fact, when a customer spends a lot of monetary units in a type of cuisine, it doesn't do so in other types of cuisine.

To avoid any multicolinearity issues, we will remove redundant variables (the ones used to aggregate, and the ones used to create new variables), using Pearsons's correlation matrix to evaluate.

Further analysis on multivariate relationships with new features and other univariate trials are done in the Jupyter Notebook and was not included here to respect the page limit since it was not the main goal of the EDA.

3. CONCLUSION

We conclude with the EDA that we have a young customer base, highly segmented by cuisine expenditure and with spending patterns influenced by time of day and customer region. Our customers are either frequent or single users, with the majority having a low average of placed orders and products. However, a smaller group of high-frequency, high-spending customers emerges as outliers and are in fact a valuable segment.

4. 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

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

Stack Overflow (This Forum was used for some code construction and debugging.)

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

5. APPENDIX

5.1. Univariate visualizations

Disclaimer: Key statistics information was taken with the output of the describe function, in sections 3.1.1 and 3.1.2 of the Jupyter Notebook. We opted to not include the table since this information is described above. We only included in the report visualizations that we perceive as relevant.

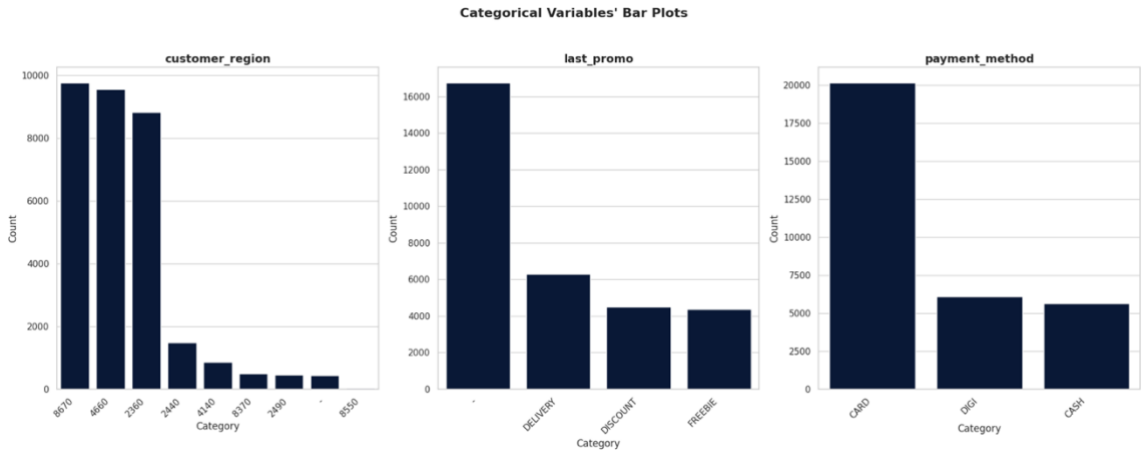

Figure 1 - Bar Chart of categorical variables

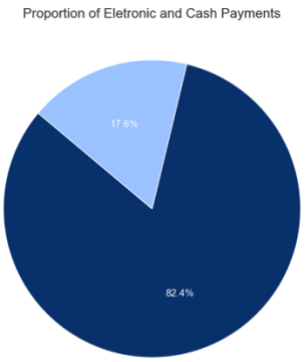

Figure 2 - Pie Chart of Payment Methods

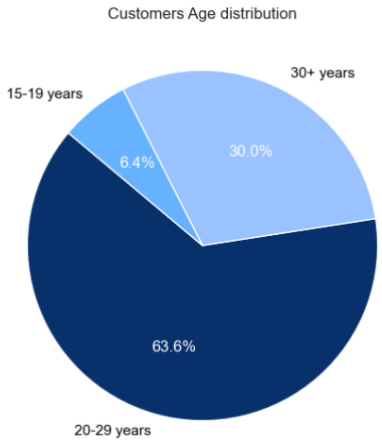

Figure 3 - Pie Chart of customer_age

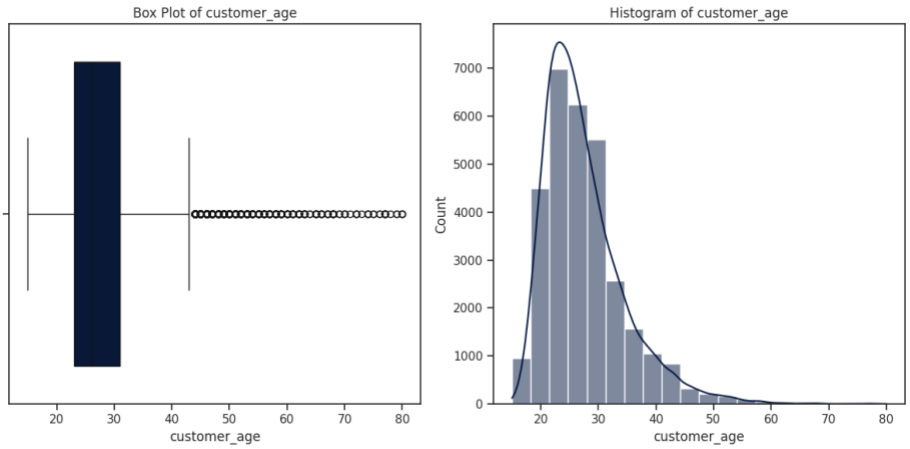

Figure 4: customer_age boxplot and histogram

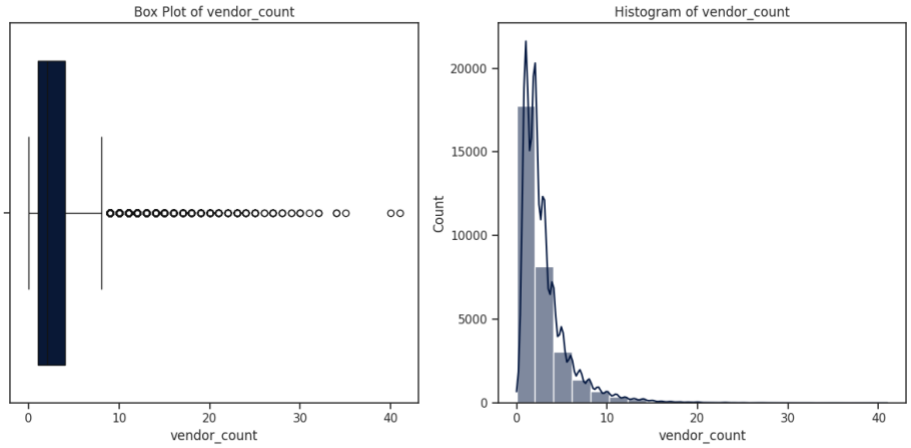

Figure 5 - vendor_count boxplot and histogram

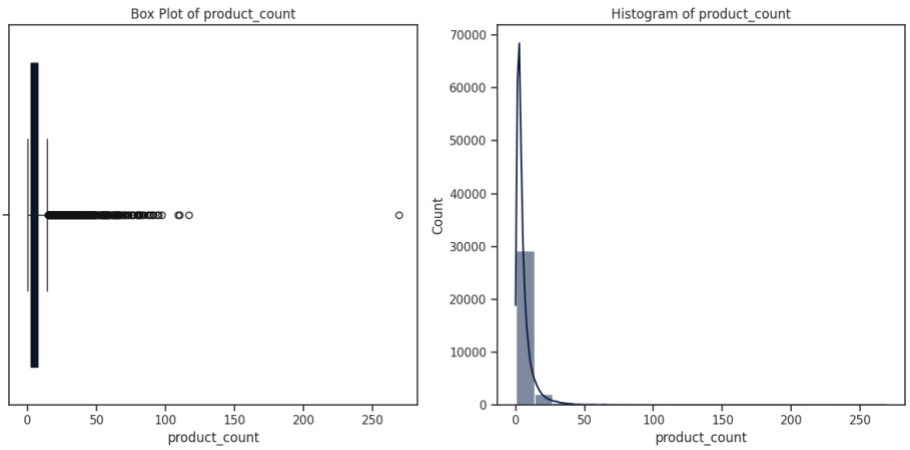

Figure 6 - product_count boxplot and histogram

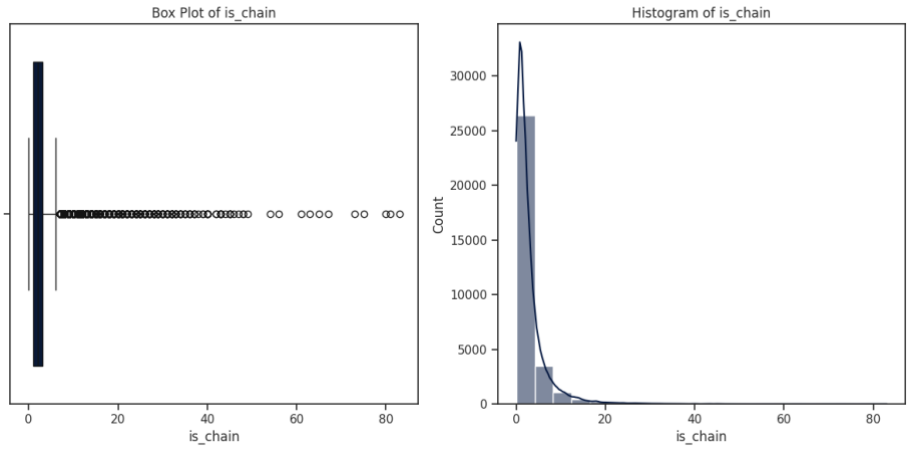

Figure 7 - is_chain boxplot and histogram

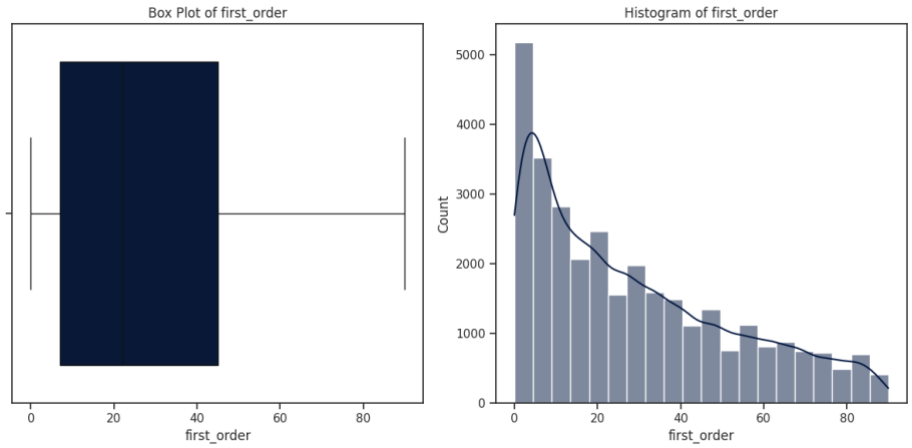

Figure 8 - first_order boxplot and histogram

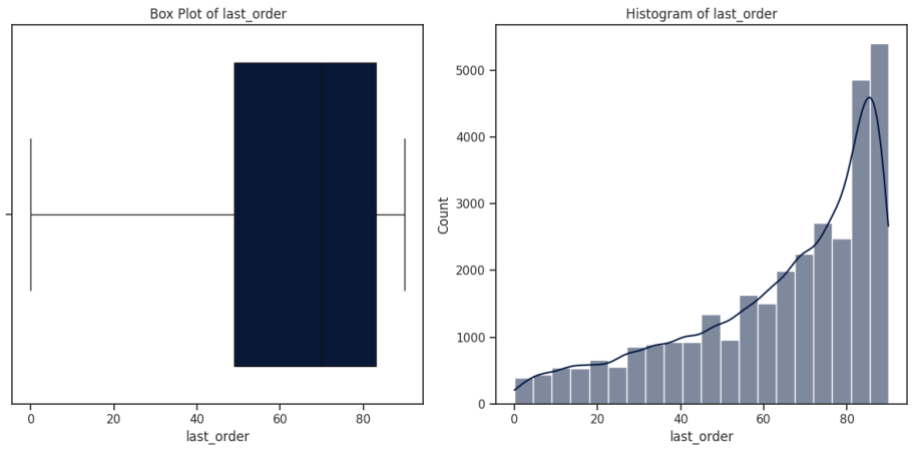

Figure 9 - last_order boxplot and histogram

In order to not overload the appendix, we will only include one example of the boxplot and histogram for day of the week, since the distributions are very similar between them.

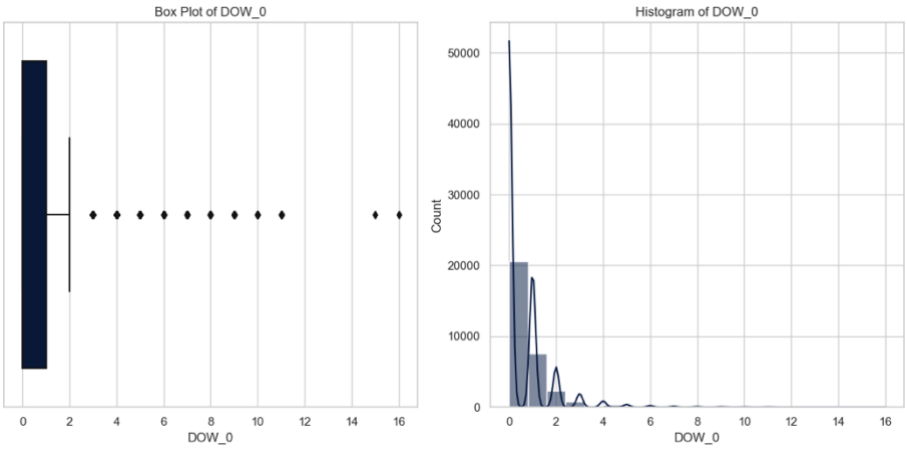

Figure 10 - Boxplot and Histogram of DOW_0

For HR and CUI, the boxplot and the histograms were not insightful, so we decided to include only one example of each.

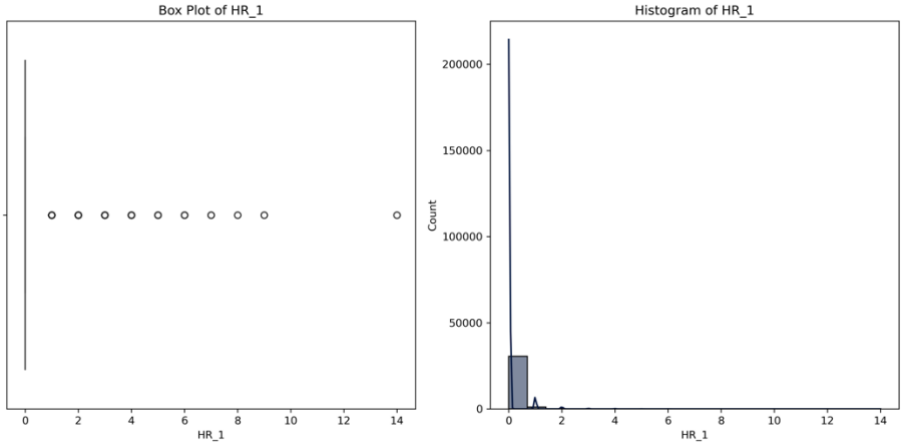

Figure 11 - HR_1 boxplot and histogram

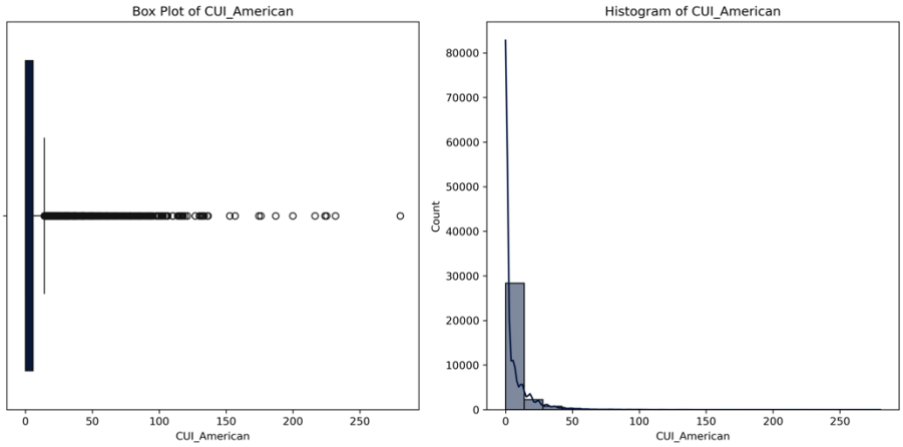

Figure 12 - CUI_American boxplot and histogram

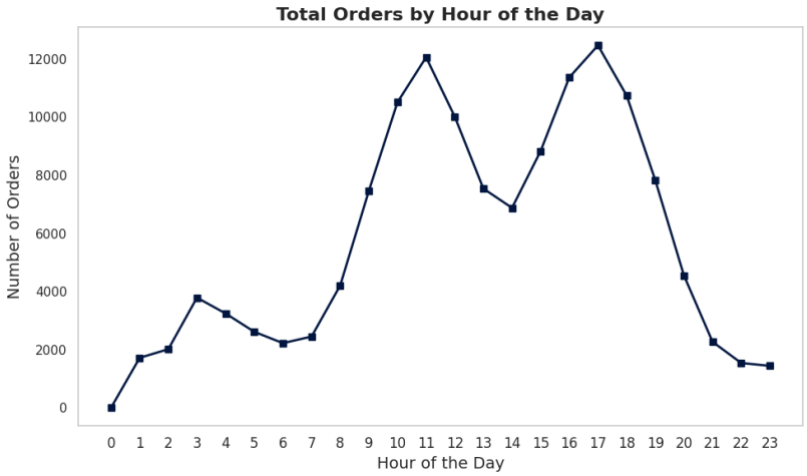

Figure 13 - Plot Of Number of Orders By Hours of the Day

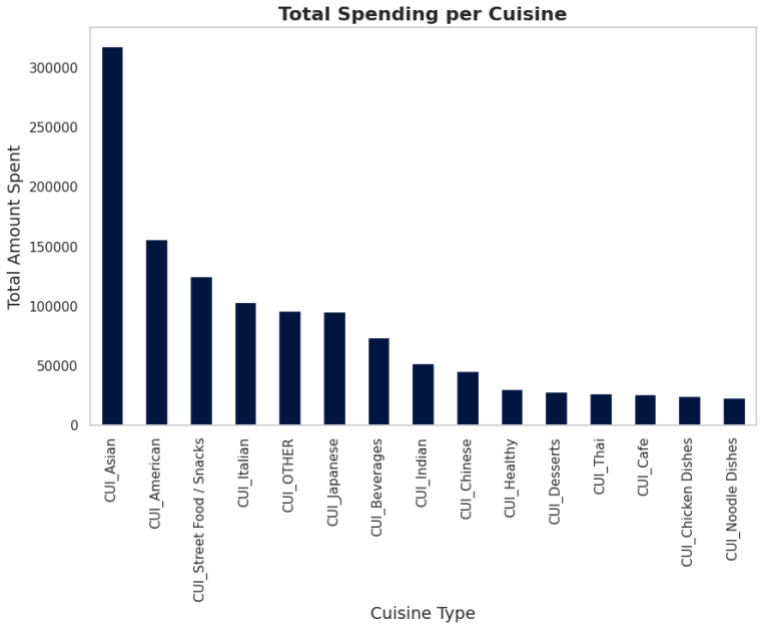

Figure 14 - Bar chart of Total Spending per Cuisine

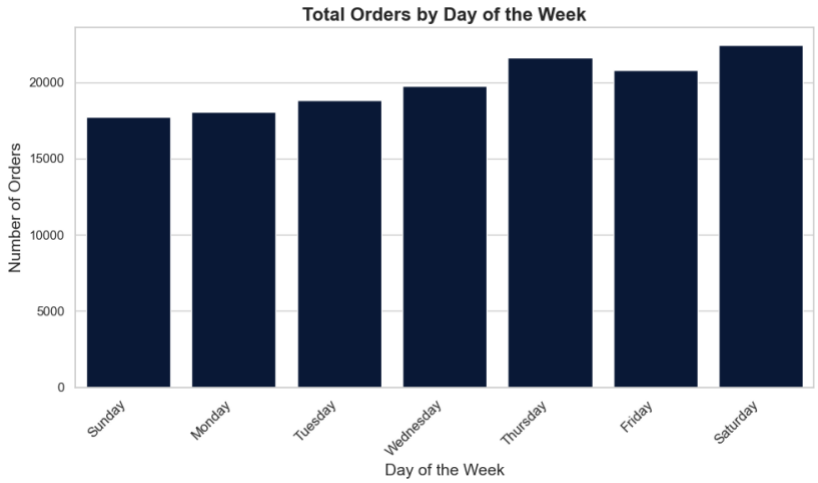

Figure 15- Bar Chart of Total of Orders by Day of the Week

5.2. Multivariate Visualizations

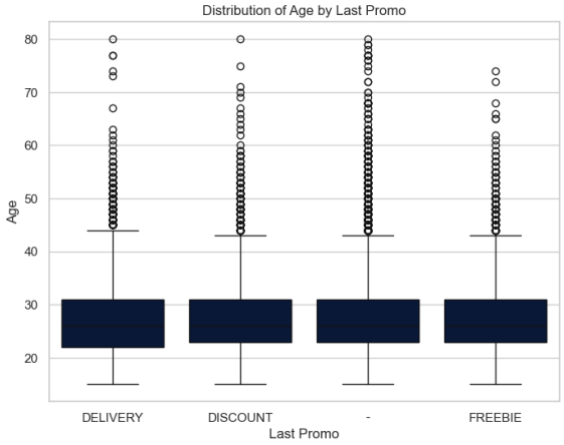

Figure 16 - Boxplot of Distribution of Age by Last Promo

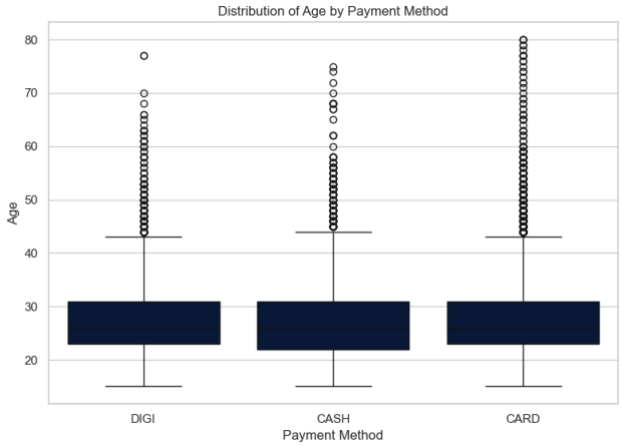

Figure 17 - Boxplot of Distribution of Age by Payment Method

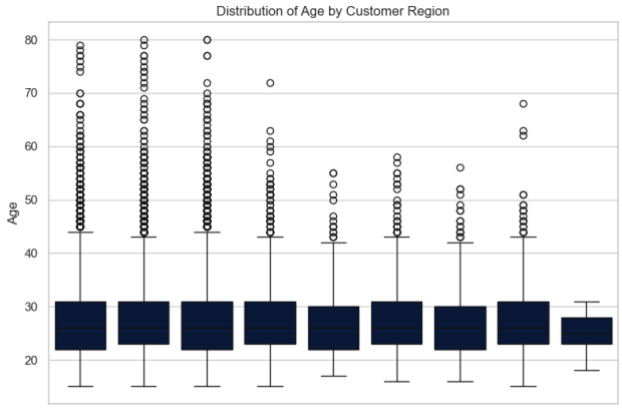

Figure 18 - Boxplot of Distribution of Age by Customer Region

Figure 19 - Boxplot of First order by Customer Region

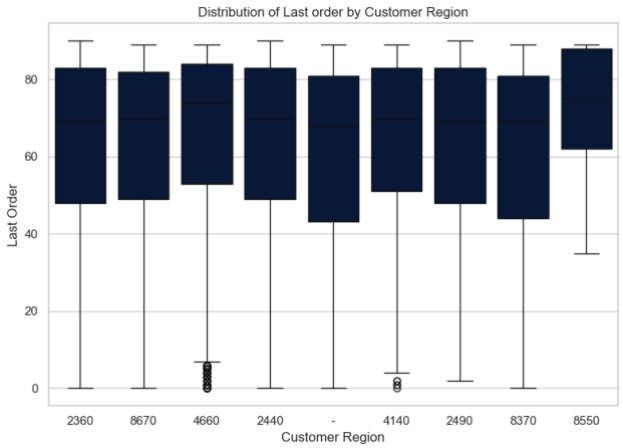

Figure 20 - Boxplot of Last order by Customer Region

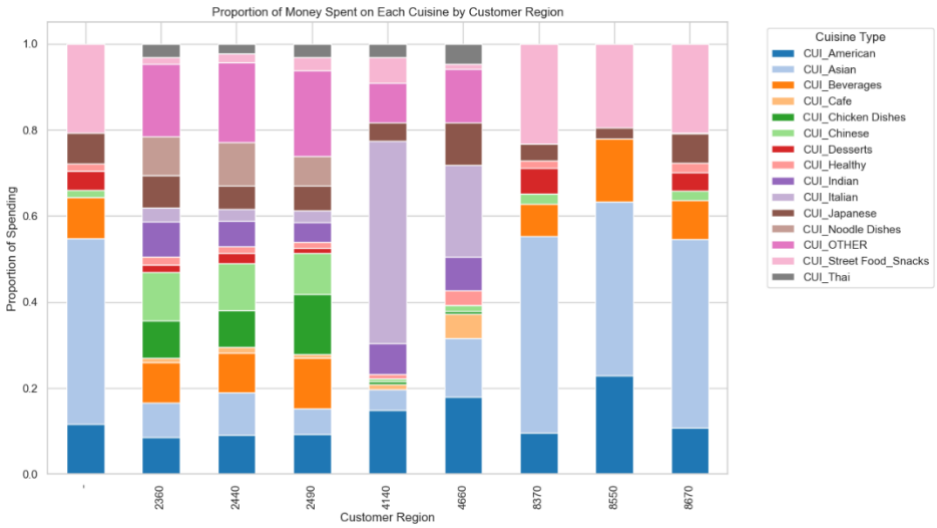

Figure 21 - Bar Chart of Proportion of Money Spent on Each Cuisine by Customer Region

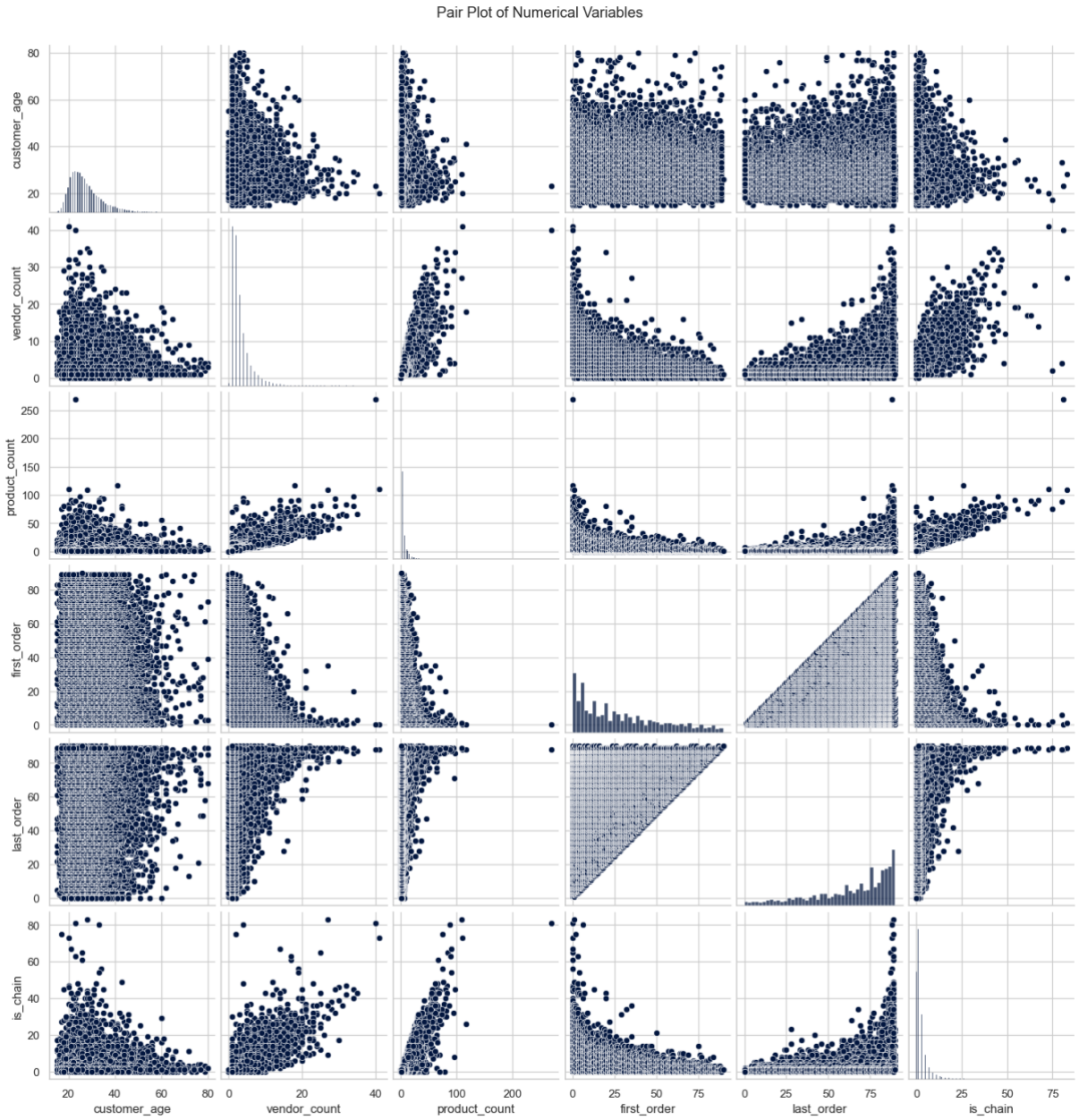

Figure 22 - Pairplots of Numerical Variables

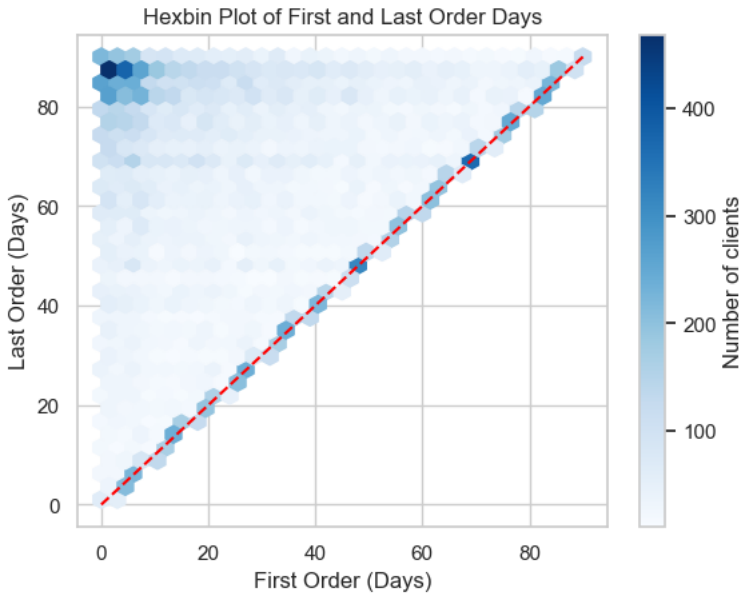

Figure 23 - Hexbin Plot of First order and Last order

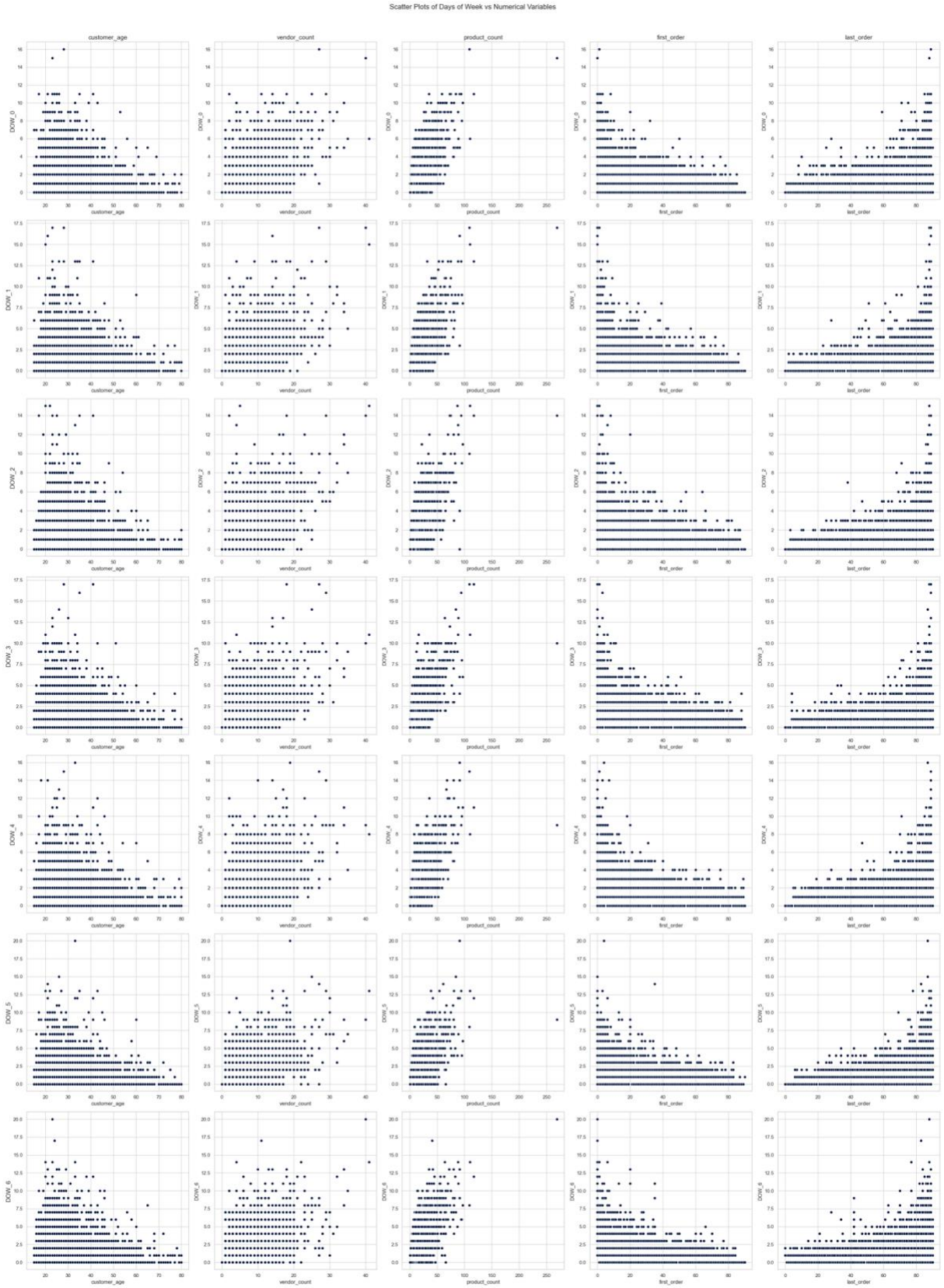

Figure 24 – Scatter plots of days of the week vs numerical variables

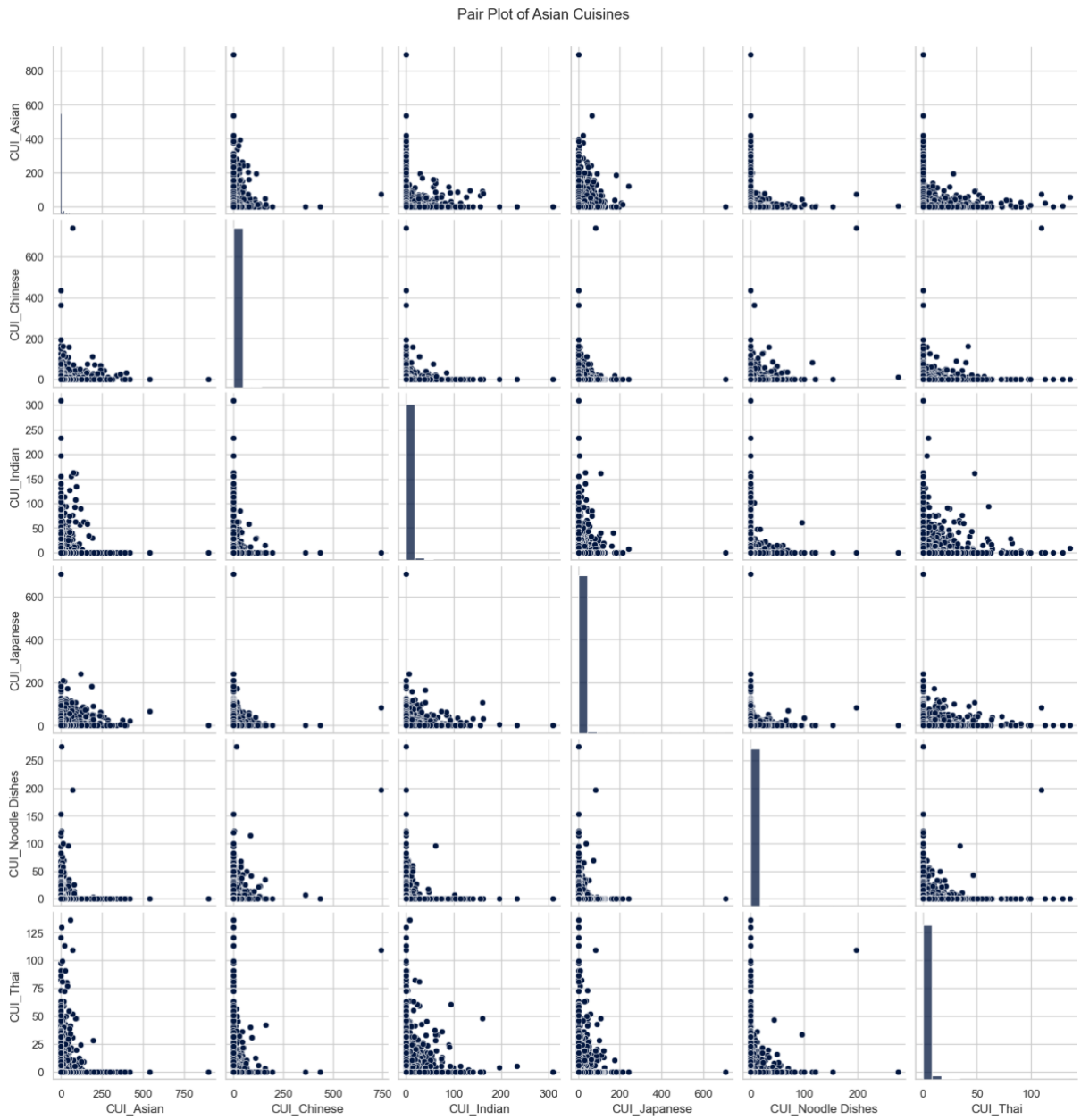

Figure 25 - Pairplots of Cuisines

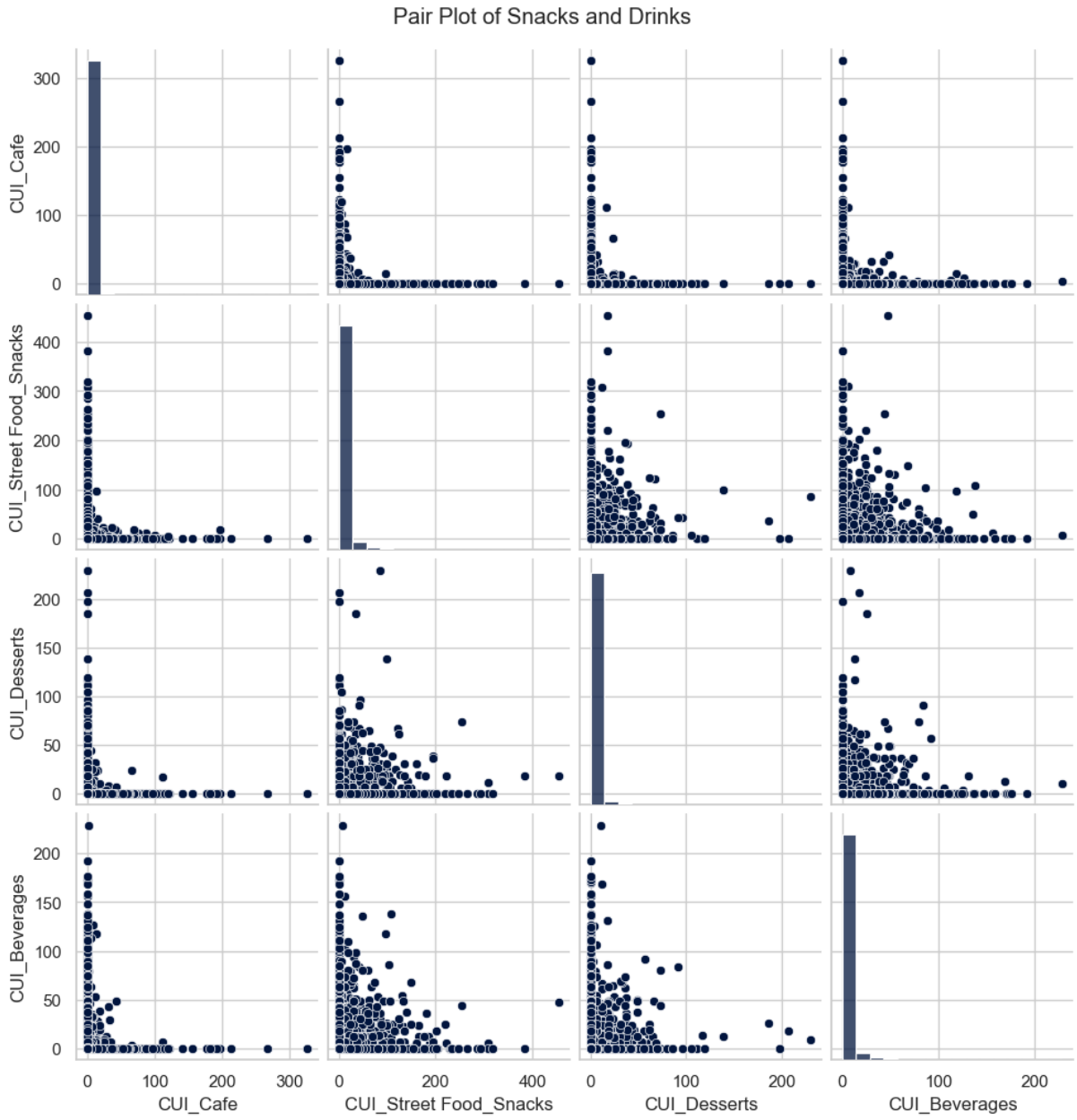

Figure 26 - Pairplots of Cuisines

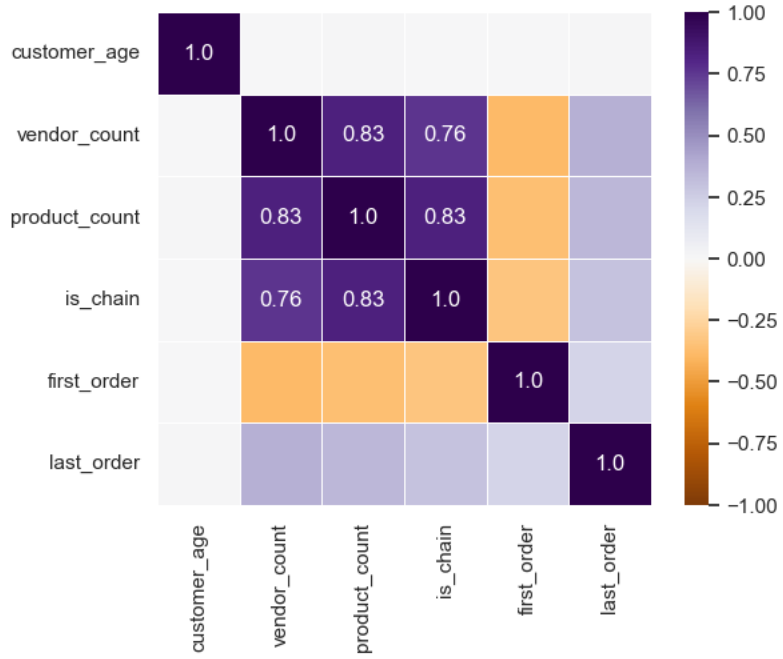

Figure 27 - Correlation Matrix between Numerical Variables

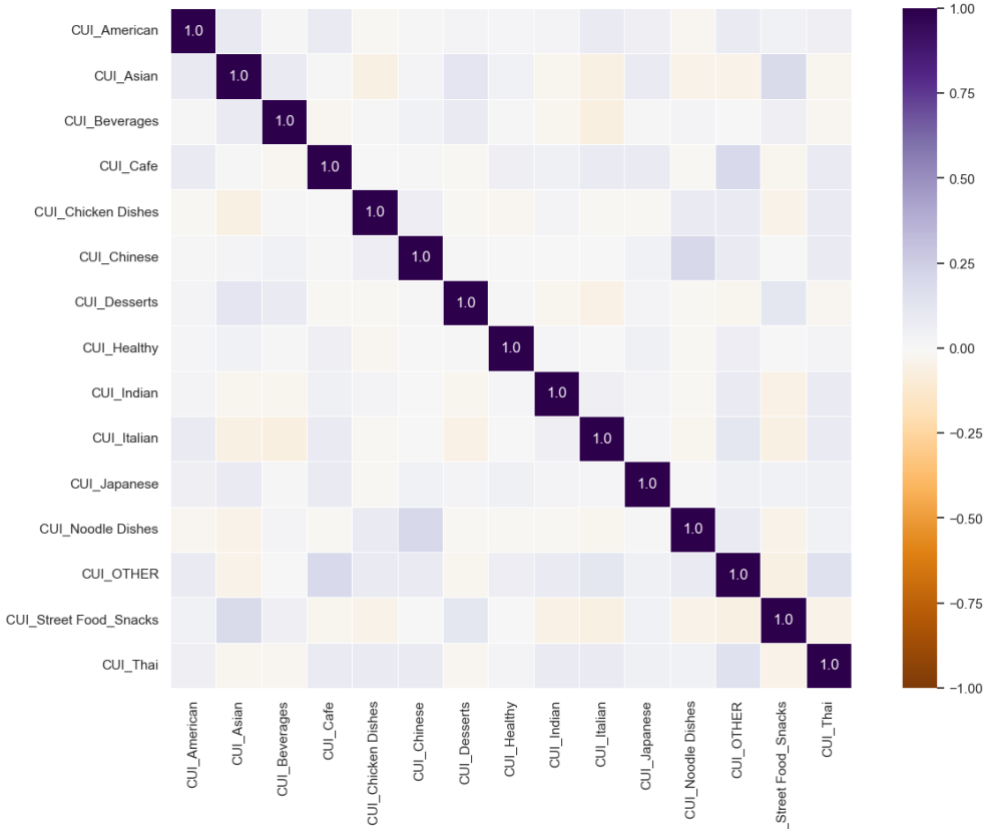

Figure 28 - Correlation Matrix between Cuisine Features

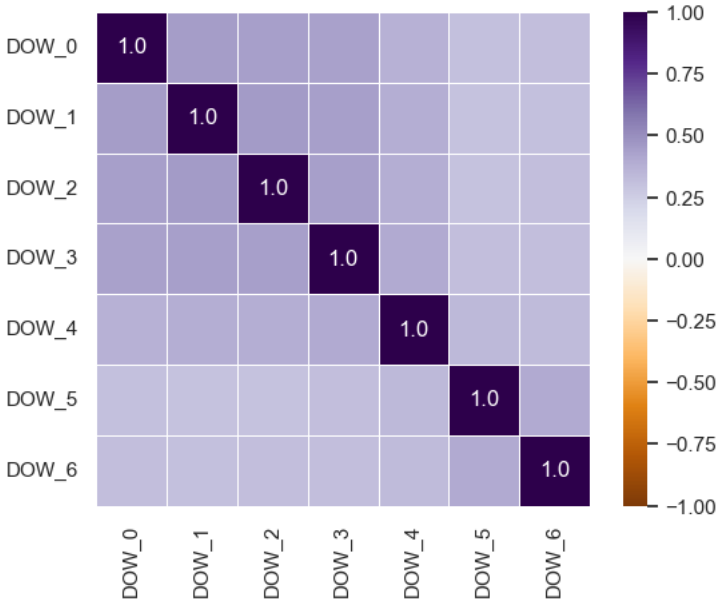

Figure 29 - Correlation Matrix between DOW

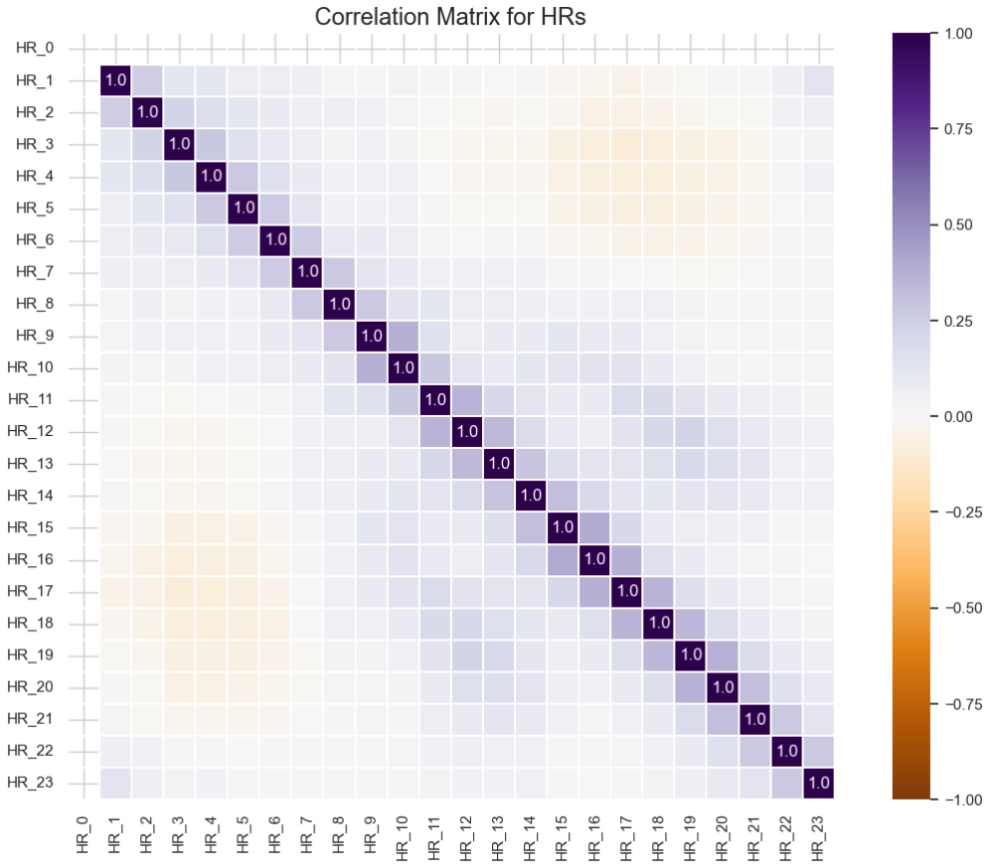

Figure 30 - Correlation Matrix between HRs

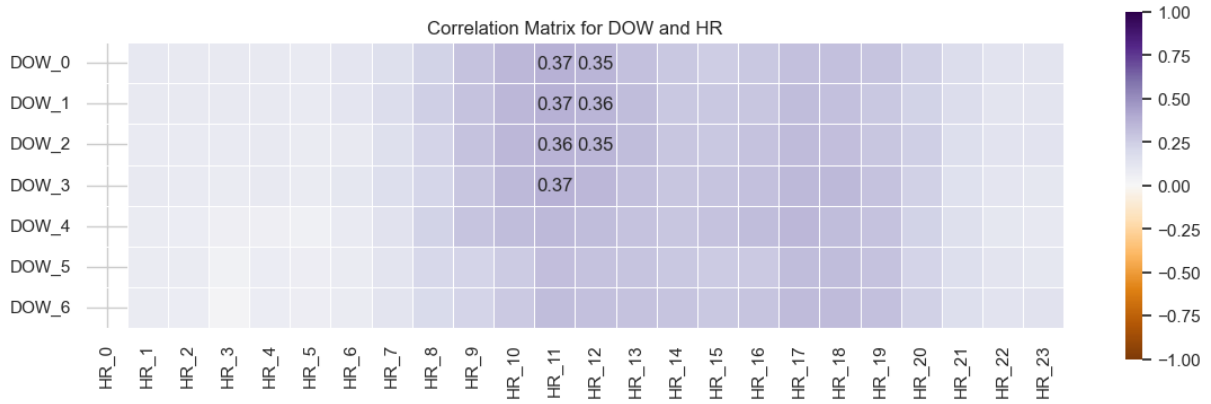

Figure 31 - Correlation Matrix between HR and DOW *

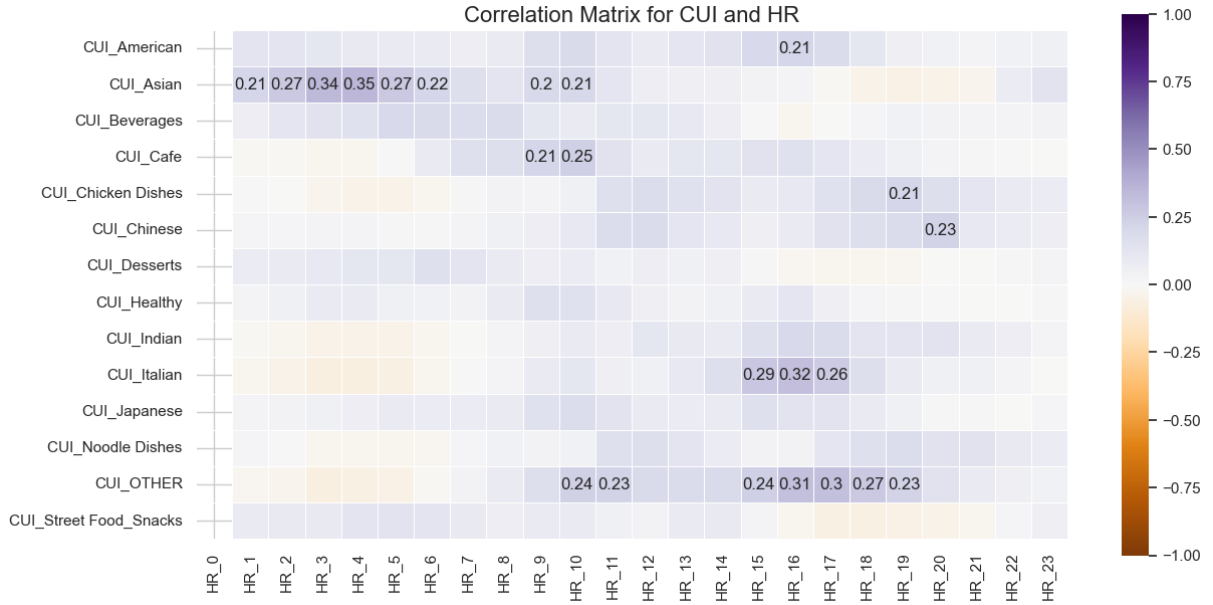

Figure 32 - Correlation Matrix between CUI and HR *

*this figures are not referenced in the report but provide insightful information that is further explored in the notebook

5.3. Insightful visualizations with new features

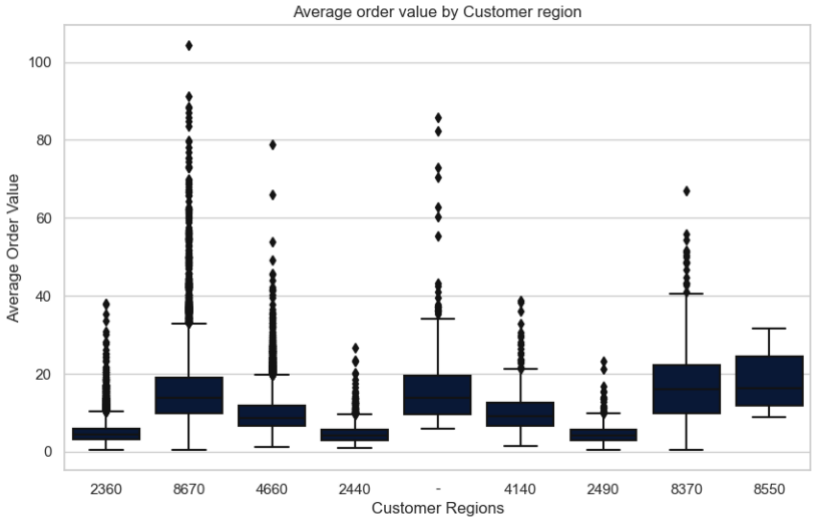

Figure 33 - Box Plot of Average Order Value and Customer

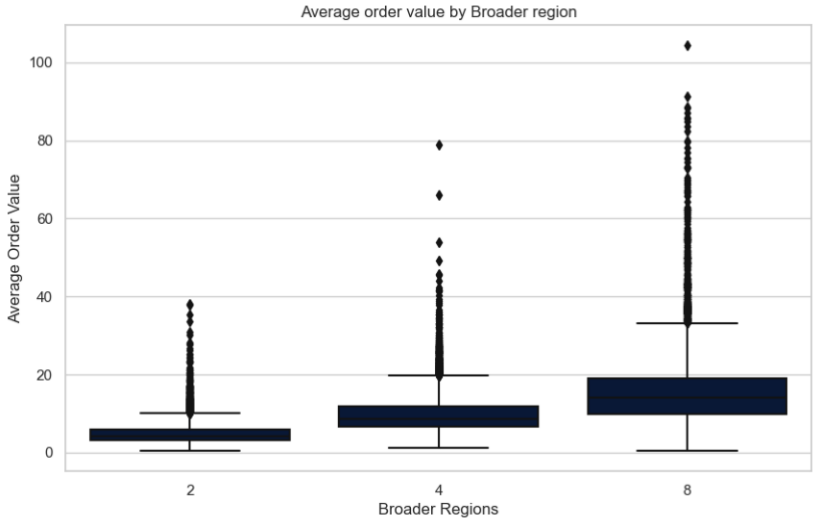

Figure 34 - Box Plot of Average Order Value and Broader Regions

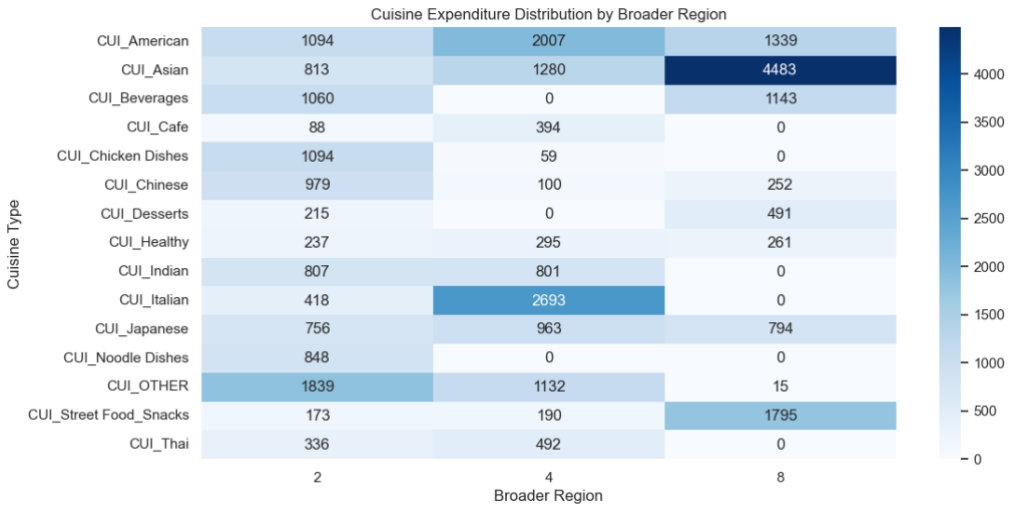

Figure 35 - Heatmap of Cuisine Expenditure Distribution by Broader Region

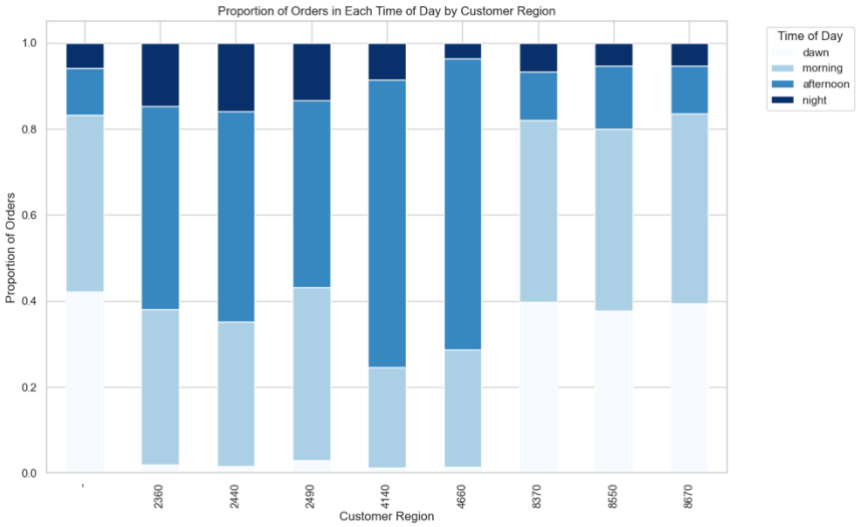

Figure 36 - Bar Chart of Proportion of Orders in Each Time of Day by Customer Region

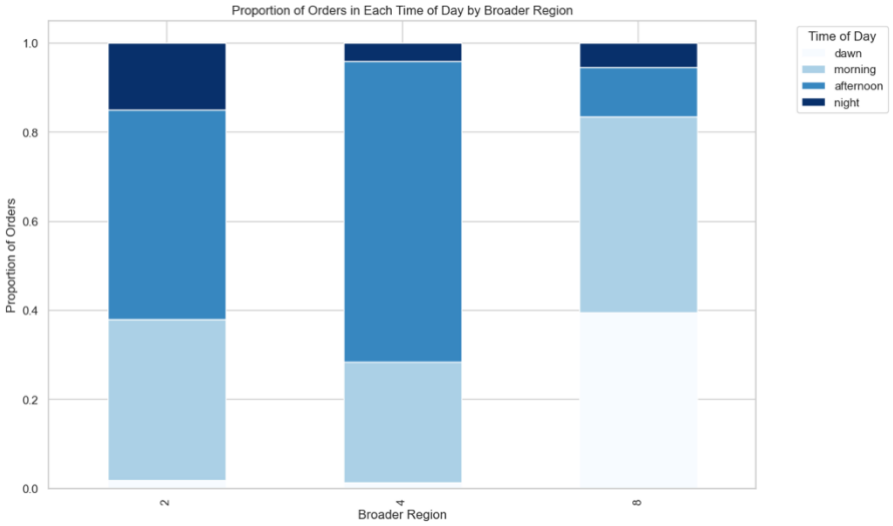

Figure 37 - Bar Chart of Proportion of Orders in Each Time of Day by Broader Region