Data Mining Project

Master in Data Science and Advanced Analytics

NOVA Information Management School

Universidade Nova de Lisboa

Exploratory Data Analysis

ABCDEats Inc.

Group 36

Joana Esteves, 20240746

Eduardo Mendes, 20240850

João Afonso Freire, 20240528

Tomás Figueiredo, 20240941

Fall Semester 2024-2025

TABLE OF CONTENTS

LIST OF FIGURES

Introduction

For this project, we analysed three months of customer data from “ABCDEats Inc.”, a food delivery service, to uncover patterns and trends in customer behaviour. Our ultimate aim is to segment the customer base, allowing us to design effective strategies for each group.

We performed an exploratory data analysis (EDA) to examine key statistics, visualize relationships, and identify any anomalies. Our notebook includes a detailed exploration of every statistic and plot, and in this report, we focus on the most important insights.

First, we addressed data quality issues like data type inconsistencies, missing values, unusual values, and duplicates, and we explore how these affect our analysis. Additionally, we performed a univariate analysis to explain distributions, identify outliers and highlight the main characteristics of the customer base. The variables that are related, namely, the days of the week, hours of the day, and types of cuisine, were analysed as groups while the other variables were analysed individually.

Through bivariate and multivariate analysis we explored customer behaviour and identified trends between variables to understand the features that best differentiate customer behaviours. These insights lead to developing new features to further explore these relationships, as a result we were able to identify relationships between the new variables using statistical tests and will also consider them for the future clustering process.

While data preprocessing was done in the notebook, we won’t discuss it here as our focus is on the exploratory analysis. This report summarizes the key insights that emerged from our exploratory analysis, laying a foundation for the future customer segmentation.

1. Exploratory Data Analysis (EDA)

The dataset provided for this project includes customer data from ABCDEats Inc. and consists of multiple columns capturing various customer attributes and aggregating their behaviour over a three-month period. Each row of the dataset represents a customer, adding up to 31 888 customers across 56 features.

1.1. Data Quality

Data Type Issues: ‘customer_age’, **‘first_order’**, and **‘HR_0’** are float features, but they should be integers because they refer to the age of the customers, the number of days, and the number of orders, respectively.

Missing values:

- **‘customer_age’**: 2.28% of the entries are missing, this could be because of errors in collecting the data or the customers' preference to not share their age, as it is personal information.
- **‘first_order’**: 0.33% of the rows have missing values and we noticed that in this cases ‘last_order’ has the value 0 (indicating the first day of the database, meaning 0 days have passed). Since ‘first_order’ cannot be a negative value it must also correspond

to 0, indicating that both first_order and last_order correspond to the same day. We can see that for other cases where the customer only made one order for the last 3 months, the value for first_order and last_order correspond. This reinforces that when last_order is 0, first_order should also be 0.
- **'HR_0'**: The dataset has a missing value percentage of 3.65% for this variable, there may have been an error collecting the data.

Strange values:

- **'vendor_count'** and **'product_count'**: Given that the variables 'payment_method' and 'last_order' have no missing values we can assume all customers have made at least one order. The variables 'vendor_count' and 'product_count' both exhibits very high frequencies for zero, and since we know all customers have ordered before, this seems to be an inconsistency. The dataset has 156 instances where either 'vendor_count' or 'product_count' has a value of 0.
- **'is_chain'**: This variable has values that range from 0 to 83, which doesn't add up with the metadata description. This variable might hold the values for the number of times the customer has ordered from a chain restaurant, because each instance represents several purchases from each customer.
- **'customer_region'**: One of the categories is represented by the symbol "-", this may be used when the region of location was not shared by the customer, resulting in having no value to fill this column. This happens for 1.39% of instances.
- **'last_promo'**: The symbol "-" is used to represent one of the categories. This symbol could be used when the customer has never used a promotion code, because in those cases there would be no code to associate with this variable, resulting in the use of the symbol "-" to represent the absence of code.

Duplicates: There are 13 duplicated customers with the same information for all columns, this seems to be an error when retrieving the data. The duplicated values should be removed for an accurate analysis.

1.2. Characteristics of ABCDEats' Customer Base

Age of customers: According to Figure 2 (1st plot), the majority of ABCDEats customers are young individuals, between the ages of 20 and 30, with the data revealing a drop in older customers especially from 40 and up.

- **Outliers**: Customers over around 45, as we can see in Figure 4 (1st plot) who likely represent another segment of the customer base.

Ordering Behaviour - Vendors, Products, and Chains: We can analyse from Figure 2 most customers order from a small set of vendors, with only a few choosing a wider number of different vendors. This trend is similar for the number of products ordered, while many customers order only a few items, there is a minority of "heavy buyers". For orders from chain restaurants, higher order counts are not common at all, which might point to a preference for local or diverse options. Even though a minority of buyers seem to favour chain restaurants heavily.

- **Outliers:** Customers that are heavy buyers, have a very diverse choice of vendors or a high number of orders in chain restaurants are outliers (Figure 3 and Figure 4).

Recency: ABCDEats seems to constantly attract new customers, this is because many have placed their first order recently. We can visualize this in Figure 1. However, there are still a few customers that have been inactive for a long period.

Cuisine Preferences: Asian cuisine stands out as a choice among customers because the total expenditure on this cuisine surpasses other types of cuisine by far, as Figure 5 shows. This could either reflect a bigger number of purchases or pricey items. American cuisine comes next, but with significantly lower spending, indicating it's less of a highly priced or highly ordered choice. Other cuisines like noodles, chicken, and desserts receive relatively low spending, suggesting there is not much preference for these. Given that all cuisines have a high number of customers who have not spent anything on them (Figure 6) and high values have a low frequency, it indicates customers seem to have a group of preferred cuisine types, rather than spreading their spending evenly across all options.

- **Outliers:** For all types of cuisine, we can see in Figure 7 it's unusual to spend high amounts on a specific type of cuisine, and customers who have this behaviour are classified as outliers. Extremely low spending on Japanese and cafe items by those who have made any purchases in these categories is also classified as an outlier.

Ordering Time: The most popular ordering times seem to align with mealtimes. We can observe in Figure 8 orders peak before lunch and dinner, while late-night orders are much less common. However, a small peak appears around 3 am, this could possibly be related to night shift workers or social events. For the days of the week, the order activity doesn't fluctuate much, as we visualize in Figure 11 but it is noticeable it rises on Thursdays, Fridays, and Saturdays, with Saturday being the day with most orders. It is not common for customers to focus on a specific hour or day, but rather have a range of times where they usually order.

- **Outliers:** Very high values for the most popular hours are considered outliers, as well as any relatively high value during other hours (Figure 10). Customers that have a high number of orders on any of the days of the week are also classified as outliers (Figure 13).

Regional: Geographically, the customer distribution is unbalanced according to Figure 14. Three regions, namely "8670", "4660", and "2360" account for the majority of the customers, while other regions have much less activity. This may suggest ABCDEats has a stronger market presence in certain regions.

Promotions: Most customers do not use any codes (represented by "-"), though "Delivery" discounts are slightly more popular among the available promotion codes, after analysing Figure 15.

Payment Methods: We can see in Figure 16_there's a clear preference for card payments, surpassing the other methods like digital payment or cash.

1.3. Relationships Between Variables: Insights on ABCDEats Customer Behaviour

Ordering Habits by age: Customers between 20 and 40 years old are the bigger spenders across all types of cuisines according to Figure 18, with younger customers (20-30) placing the highest number of product orders. We can visualize in Figure 21 that older customers tend to order a smaller number of products and have less variety in vendor choices. Younger customers, particularly in their 20s, show higher vendor diversity and product count, with chain restaurants being part of ordering habits, as we can visualize on Figure 29 (1st and 2nd plot). Figure 19 shows that customers in the age group 60+ show a unique preference for Snacks and Italian cuisine, while Indian cuisine is generally unpopular among this group. The other age groups exhibit similar habits, favouring Asian and American cuisines.

Ordering Frequency: After analysing the Figure 20 we can affirm there is a strong trend among frequent customers who initially placed orders long ago but have made at least one recent order. On the other hand, a smaller group of customers, whose last orders were a while ago represent inactive users. We can also identify customers who made their first order recently, meaning new customers are joining the customer base.

Vendor and Product Count: According to Figure 29 (3rd plot), customers who order from a wide variety of vendors tend to buy a larger number of products overall. These customers also show a preference to buy in a larger number of chain restaurants.

Payment Methods and Promotions: Figure 23 shows that across all regions, the majority of customers prefer using cards over digital payment or cash, this trend also happens regardless of day (Figure 27), time (Figure 28), or type of promotion code used (Figure 30). Figure 30 also shows that customers who pay with cash or digital methods tend to spend less, especially when using “Delivery” and “Discount” codes. In contrast, customers who don’t use promotions at all tend to spend more overall.

Regional Differences: The regions ‘8550’, ‘8670’, and ‘8370’ concentrate their spending on just a few cuisines such as American, Asian, and street food. While regions like ‘4660’ and ‘4140’ display a more diverse but less intense spending across different cuisines, as can be seen on Figure 24. According to Figure 22, most of the customers don’t use any promotion at all, but certain regions like ‘8550’ favor the “Freebie” option.

2. New features

To further analyse relationships in the dataset, improve customer segmentation and data interpretability, we created new features by transforming existing variables.

We grouped **orders by weekdays vs. weekends** and **mealtimes** (breakfast, lunch, snack, dinner and late night), reducing the dataset’s dimensionality and highlighting behavioural patterns, while simplifying the hourly data into key periods improves clarity in clustering. This helps us understand what times of the day customers tend to be most interested in ordering, thinking from a business perspective, for example, issuing notifications close to those times. It also serves to analyse what times there is the most demand and from the perspective of those who manage businesses, see what times of the day they need more people working, etc.

By categorizing **age into groups** (young, adult, senior), we simplify age analysis, helping identify trends across demographics. Also, we believe these different groups of people have different purchasing power, so we need to differentiate the ages to be able to maximize customer engagement. Adding a **promotion usage** binary variable will allow us to pinpoint segments responsive to promotions. This is a good way to see if the promotions are effective or not, for example if a high percentage of customers used promotions, it suggests the promotion was appealing. Total **orders** and **total spending** were also calculated to measure customer activity, adding in differentiating between high and low engagement and spend levels.

Table 1 - New Features

New Feature Name	Feature Pseudo code	What is it?
weekend_orders	DOW_0 + DOW_6	Captures the total orders placed from Monday to Friday.
weekday_orders	DOW_1 + DOW_2 + DOW_3 + DOW_4 + DOW_5	Captures the total orders placed on Saturday and Sunday.
total_spent	cui_columns.sum	Indicates the total spending amount for each customer across all orders.
total_orders	DOW_0 + DOW_1 + DOW_2 + DOW_3+ DOW_4 + DOW_5 + DOW_6	Represents the cumulative number of orders placed by each customer.
breakfast,lunch,snack, dinner,late_night	Sum the order counts from HR_x to HR_y	Result of grouping number of orders per hour into categories according to mealtimes.
use_promo	last_promo != '-' 1 if the customer used a promotion, 0 if they did not	Indicates whether a customer has ever used a promotion code.
age_group	'young' if age < 25, 'adult' if 25 <= age < 60, 'senior' if age >= 60	Categorizes customer age into three groups: Young (0-24), Adult (25-59), and Senior (60+).

2.1. Relationships between new features

The new variables allow us to **further explore relationships** in the dataset. We performed **statistical tests** to identify patterns in customer behaviour.

The analysis of ***total_spent*** vs ***customer_region*** indicates a statistically significant difference in spending patterns across different regions, which suggests that the customer region has a role on spending behaviour. The comparison of **total spending** across different **age groups** indicates a statistically significant difference, which means that age can influence the spending habits. For ***weekdays_orders*** vs ***weekend_orders*** the test shows us there is also a statistically significant difference between the number of orders according to the time of the week. Additionally, ***total_spent*** vs ***total_orders*** test indicated that there is a statistically significant difference between the total orders of high-spending and low-spending customers. Lastly, there is no significant relationship between ***use_promo*** and ***age_group***, this suggests that customers' age does not influence whether they use promotional codes.

Bibliographical References

GeeksforGeeks. (2024, May 16). *What is Exploratory Data Analysis?*
https://www.geeksforgeeks.org/what-is-exploratory-data-analysis/

Appendix

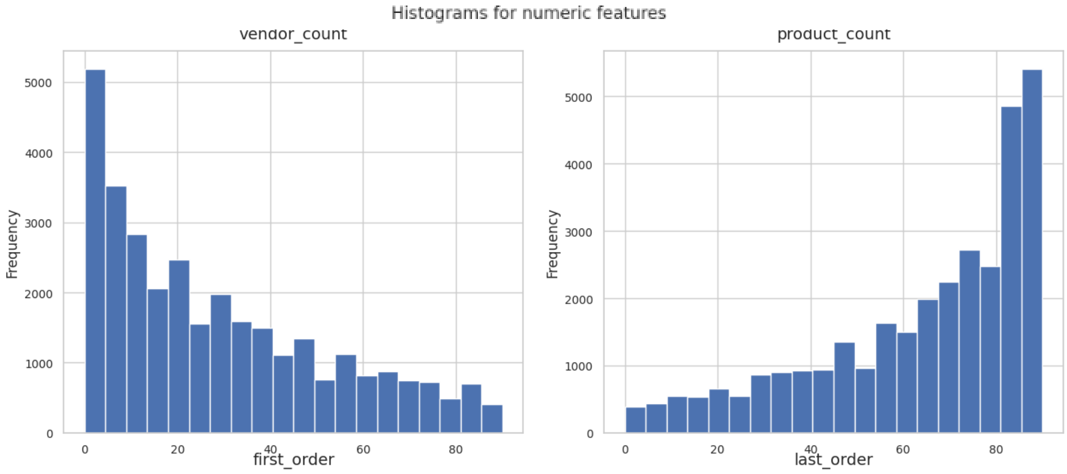

Figure 1 - Histogram for metric features

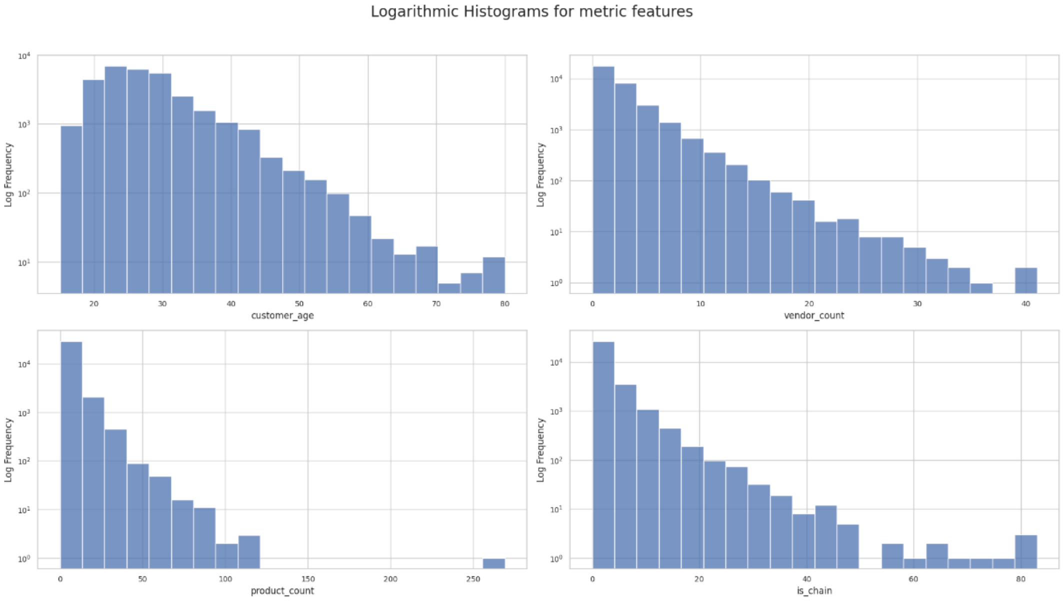

Figure 2 - Logarithmic Histograms for Metric Features

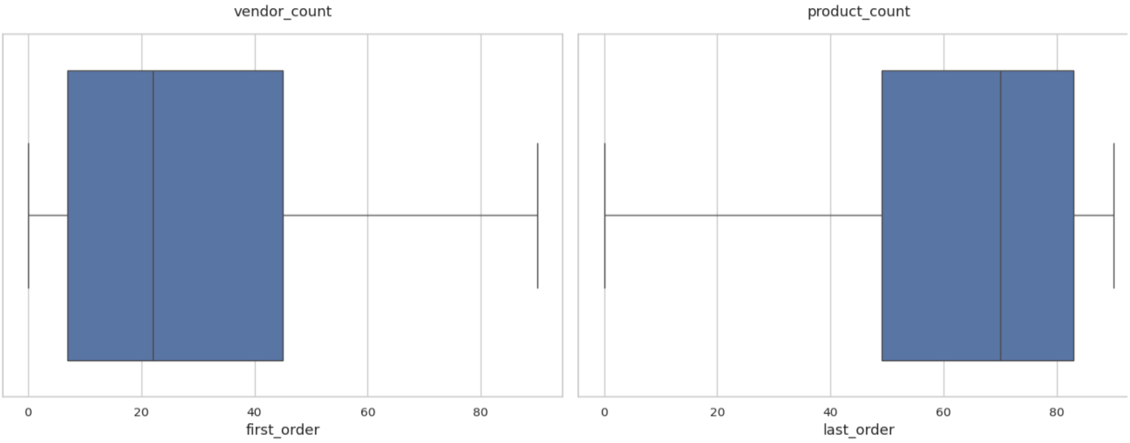

Figure 3 - Boxplots for metric features

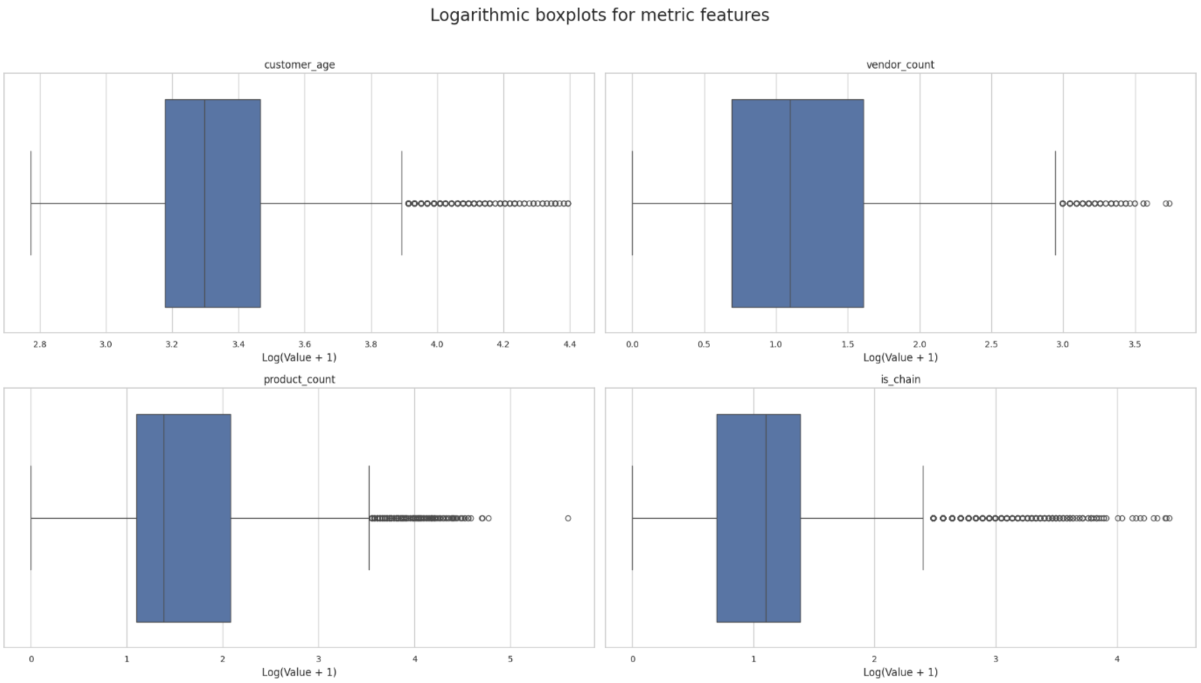

Figure 4 - Logarithmic Boxplots for metric features

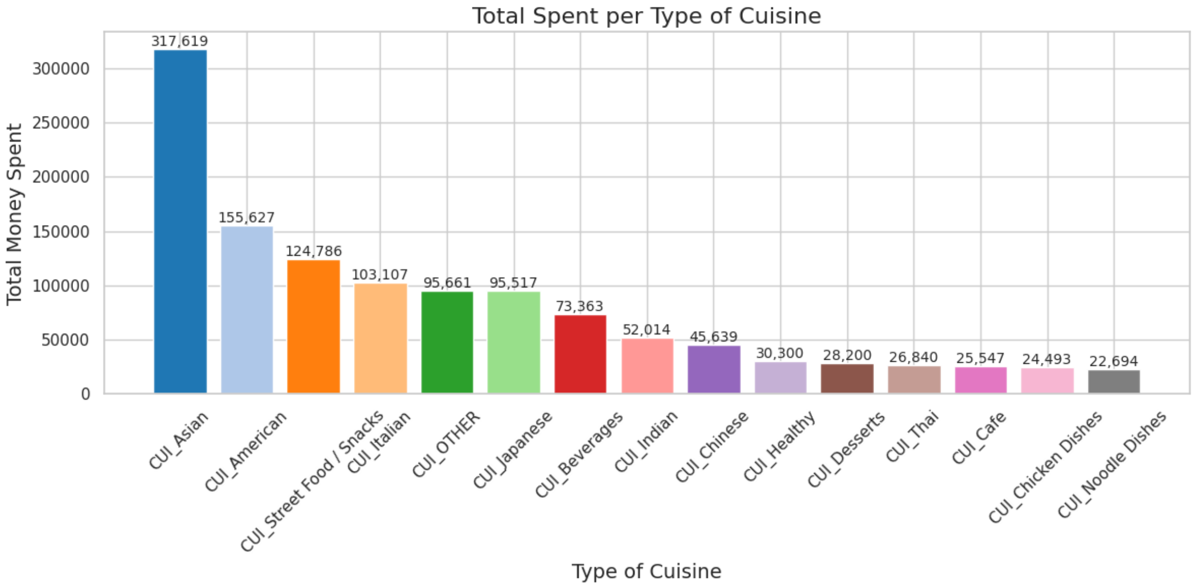

Figure 5 – Bar chart of total spent per type of cuisine

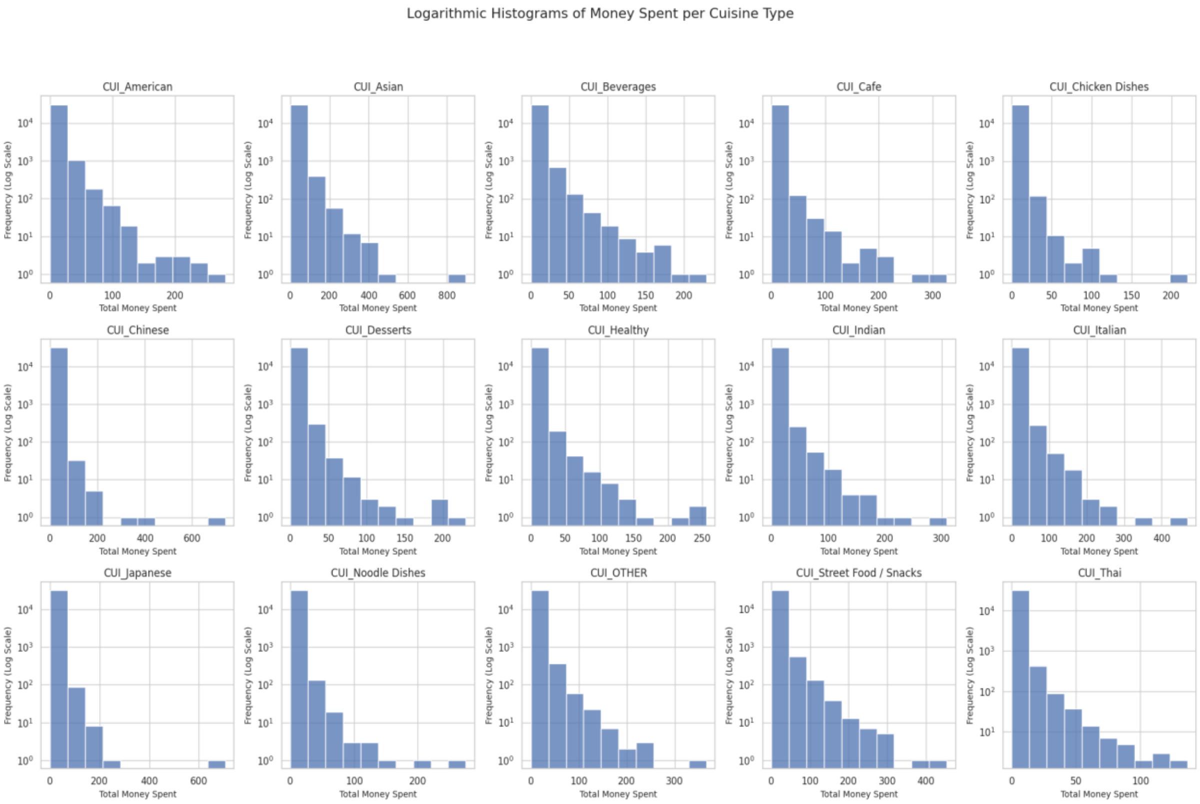

Figure 6 - Logarithmic Histograms of Money Spent by each customer per Cuisine Type

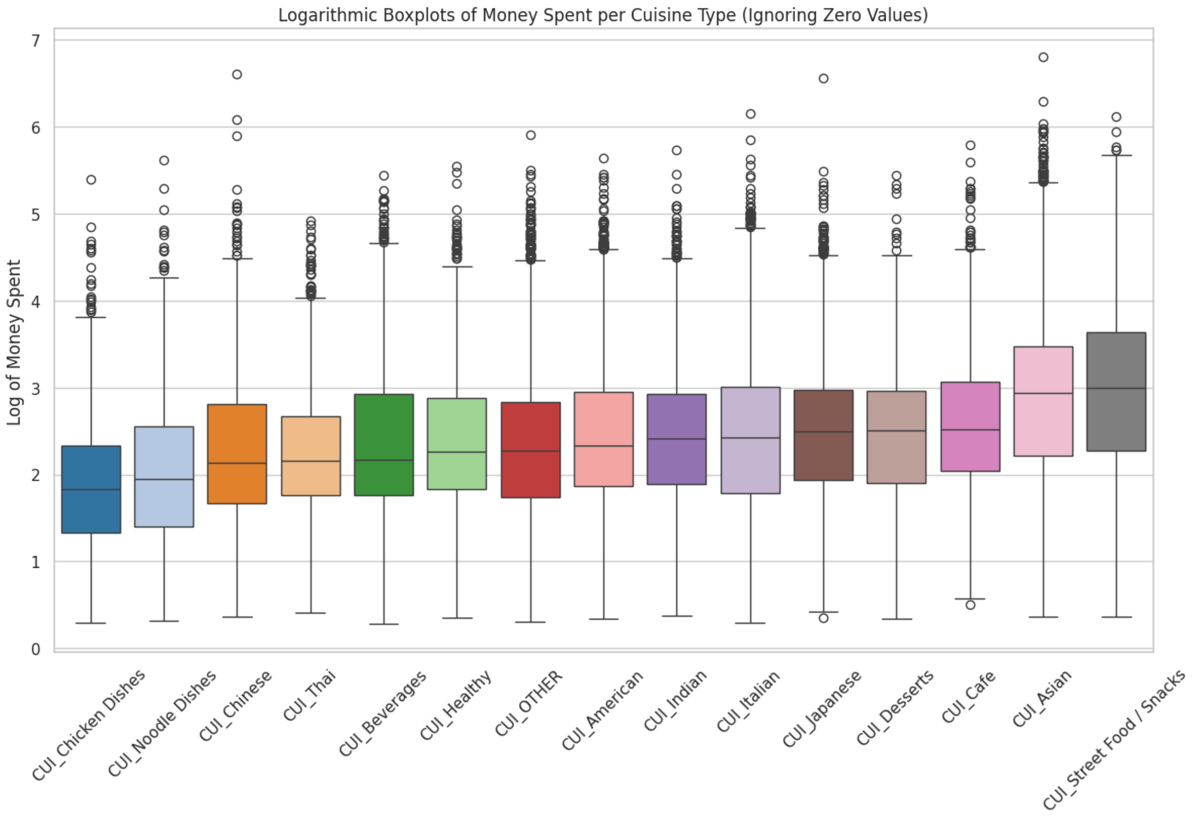

Figure 7 - Logarithmic Boxplots of Money Spent per Cuisine Type (Ignoring Zero Values)

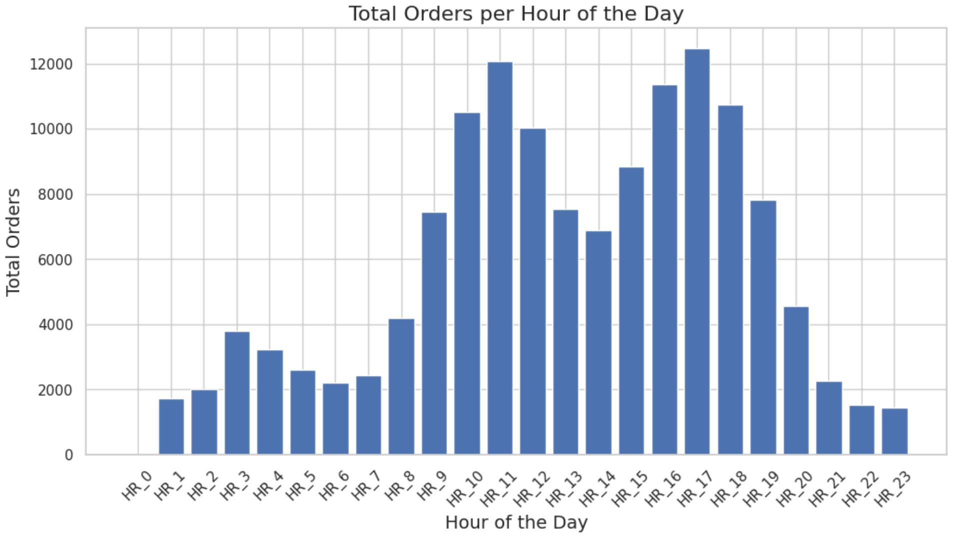

Figure 8 - Histogram for Total orders by hours of the day

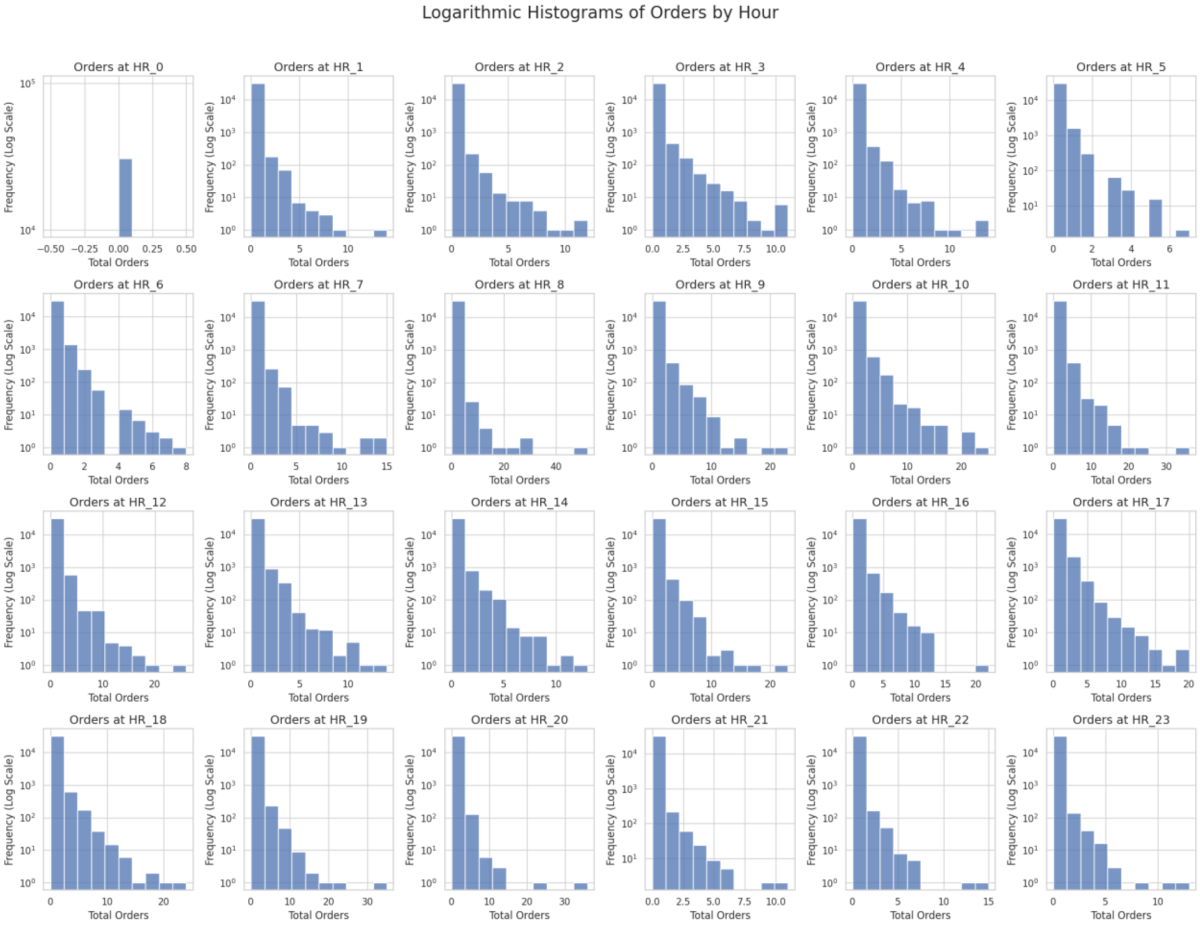

Figure 9 - Logarithmic histograms of orders by hours

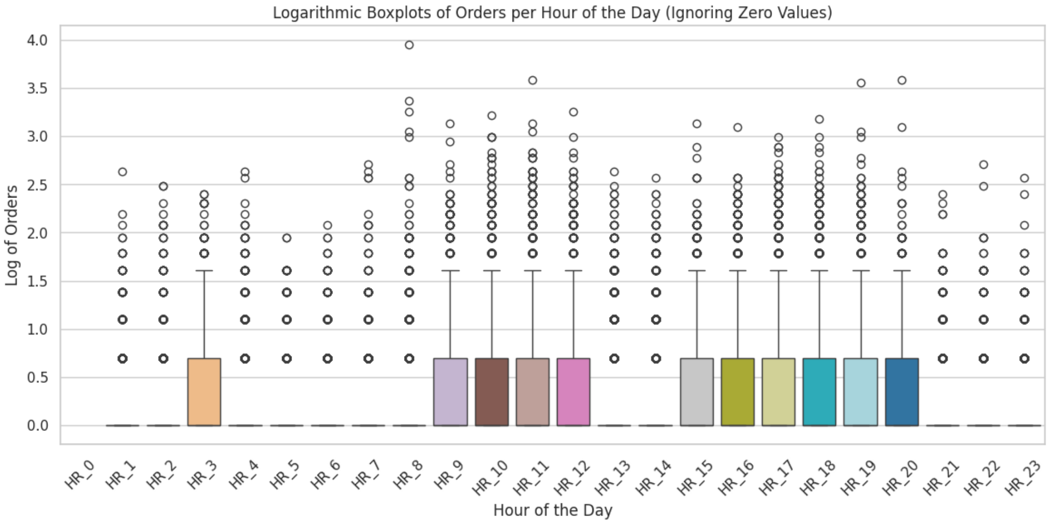

Figure 10 - Logarithmic Boxplots of Orders per Hour of the Day (Ignoring Zero Values)

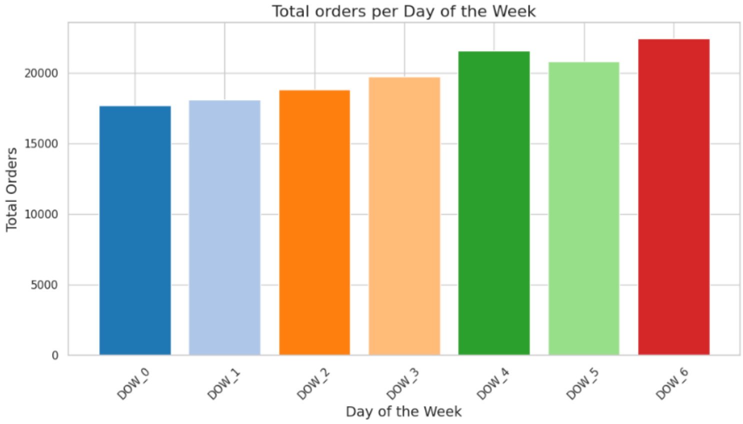

Figure 11 - Bar chart for total orders by day of the week

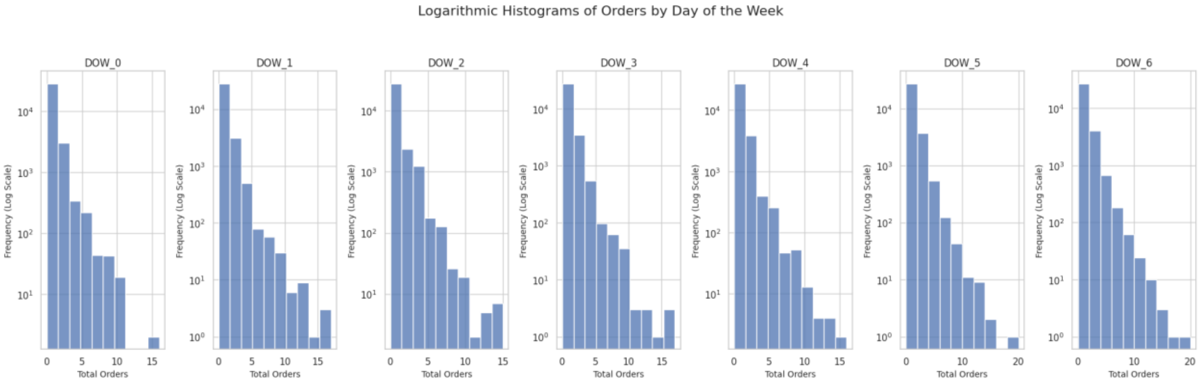

Figure 12 - Logarithmic histograms of orders by day of the week

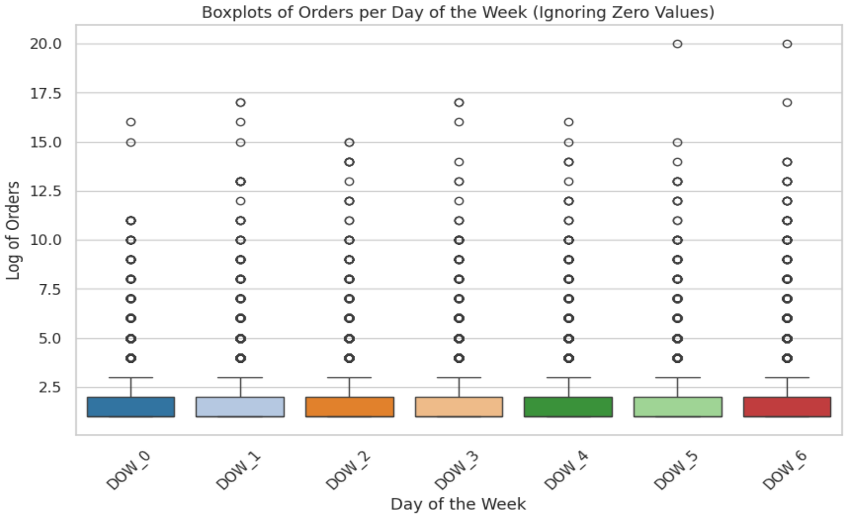

Figure 13 - Boxplots of Orders per Day of the Week (Ignoring Zero Values)

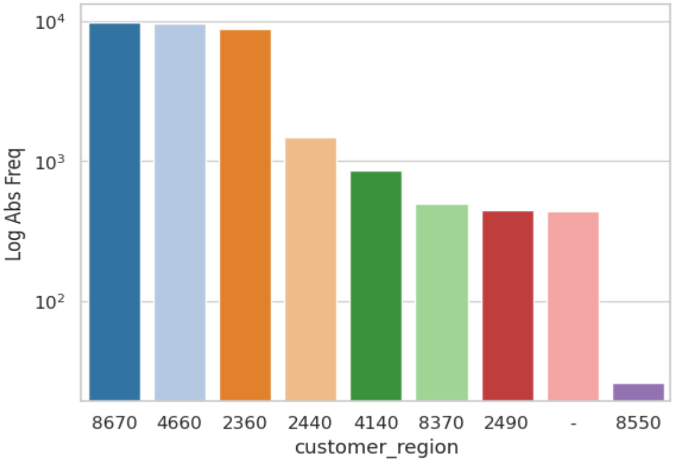

Figure 14 - Logarithmic bar chart for absolute frequency per customer region

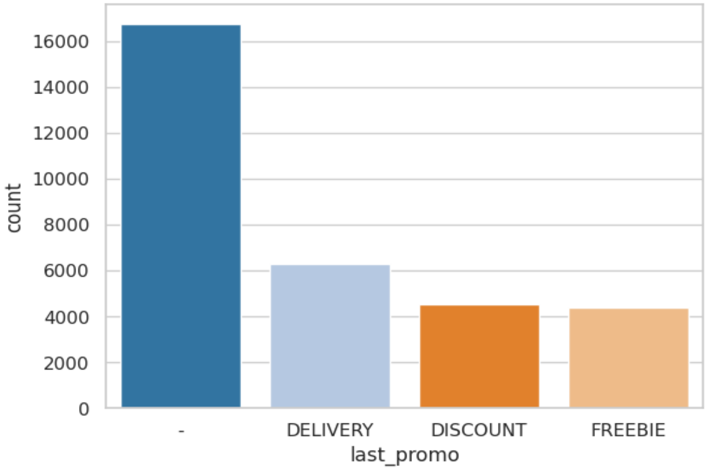

Figure 15 - Bar chart for absolute frequency per last promo

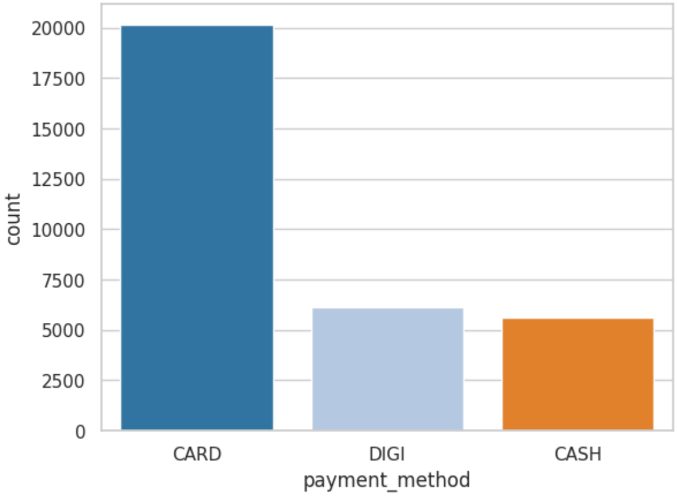

Figure 16 - Bar chart for absolute frequency per payment method

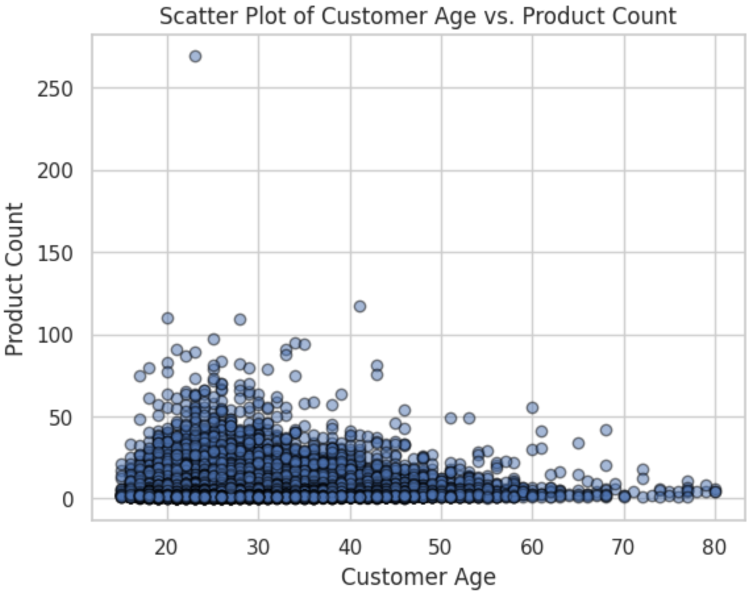

Figure 17 - Scatter Plot of Customer Age vs. Product Count

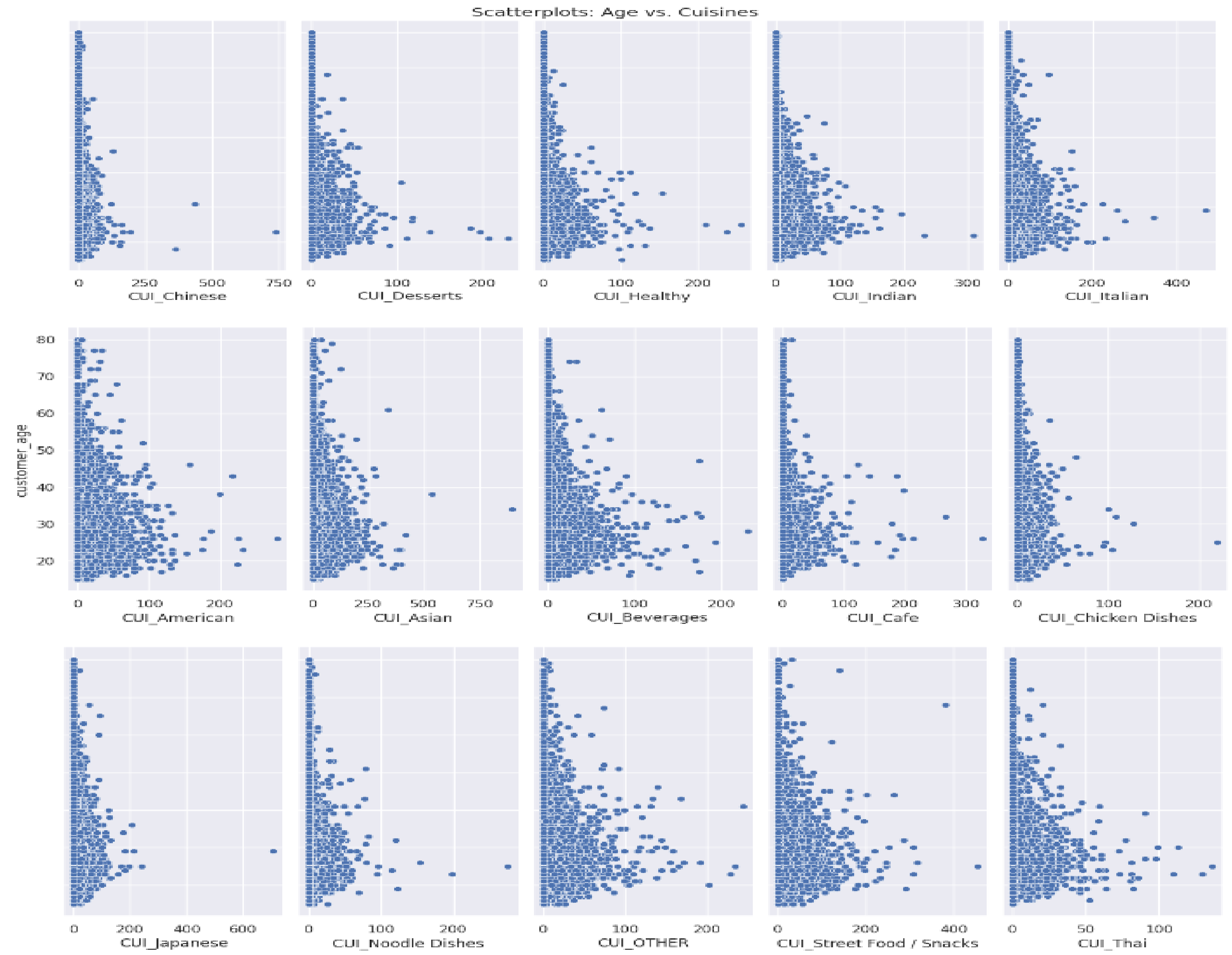

Figure 18 - Scatterplots for Age vs. Cuisines

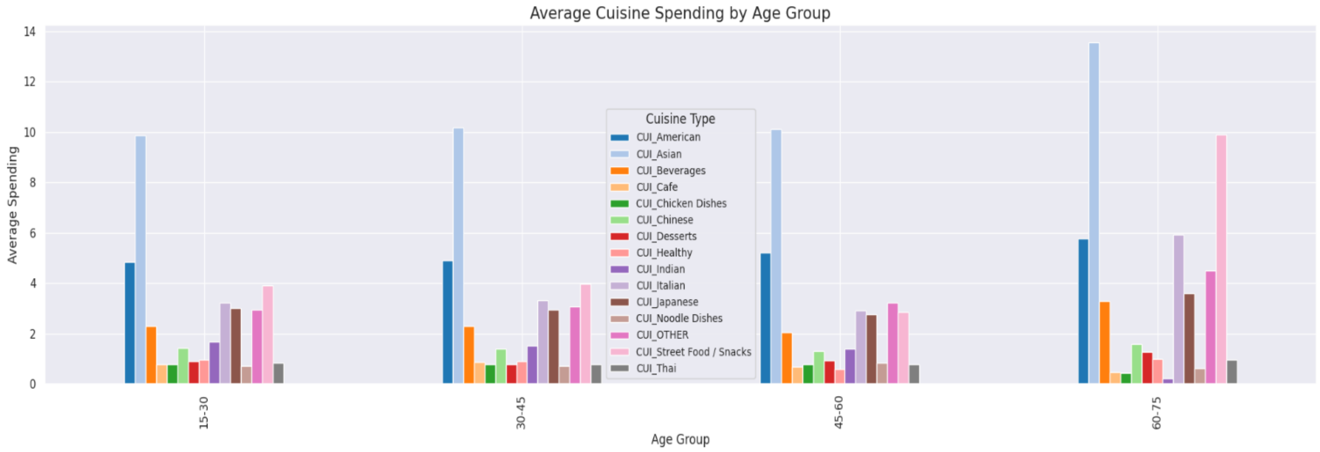

Figure 19 - Bar plot for Average Cuisine Spending by Age Group

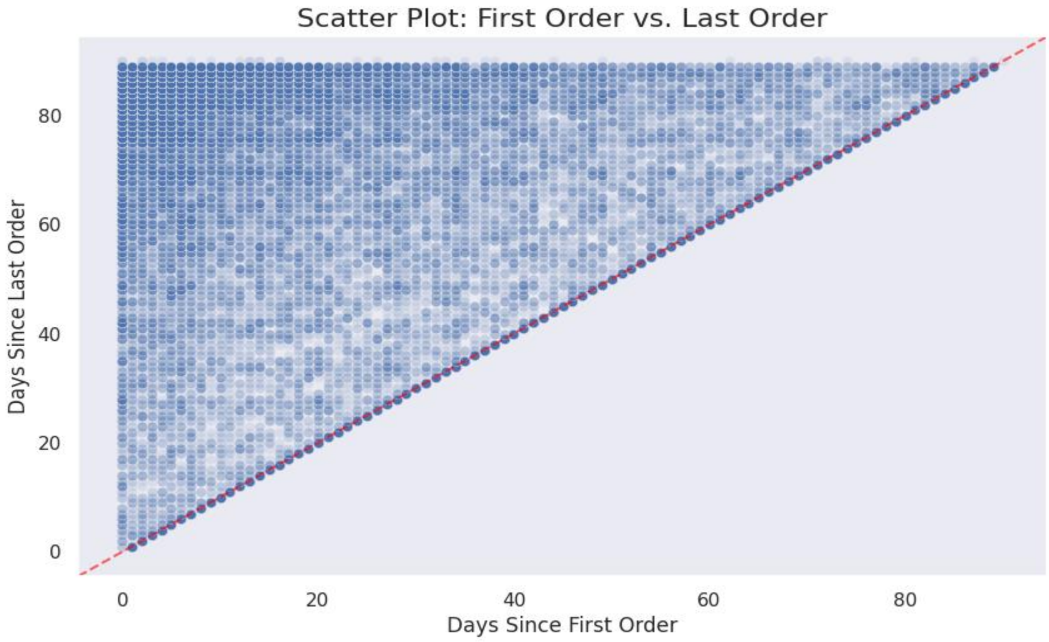

Figure 20 - Scatter Plot for First Order vs. Last Order

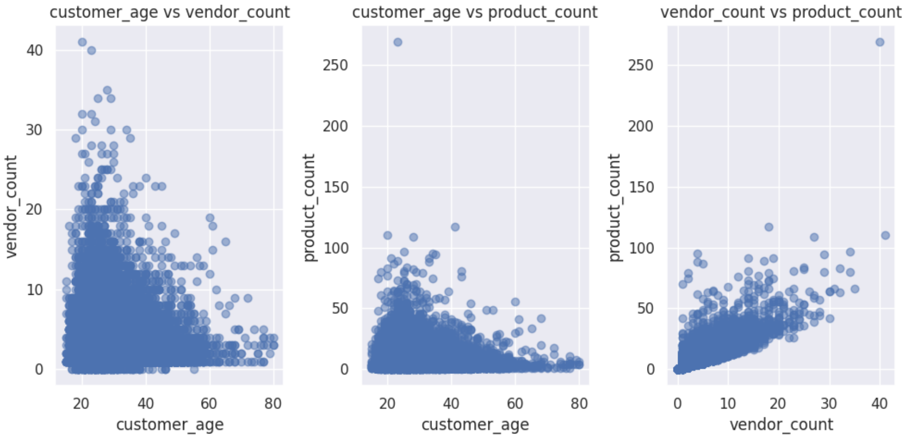

Figure 21 - Pairwise scatter plots for customer age, vendor count and product count

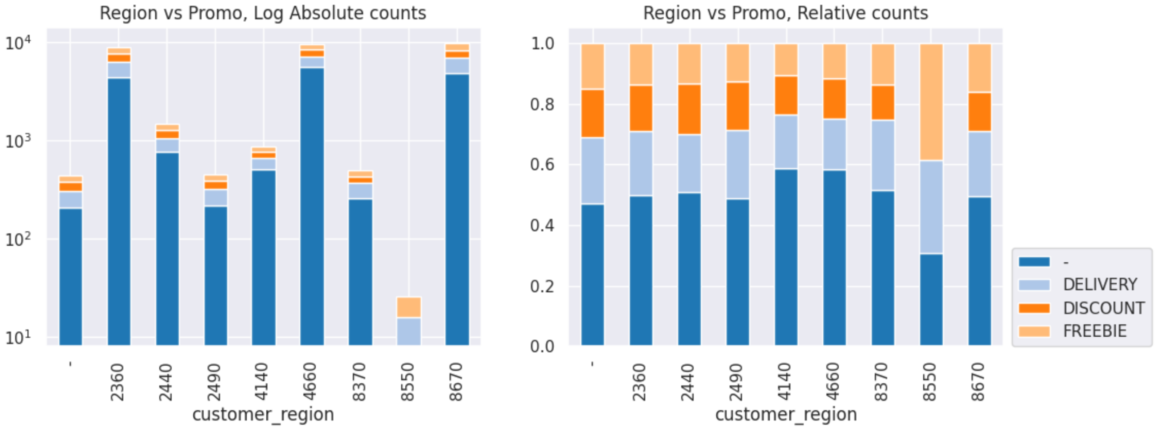

Figure 22 - Logarithmic bar charts for costumer region vs last promo

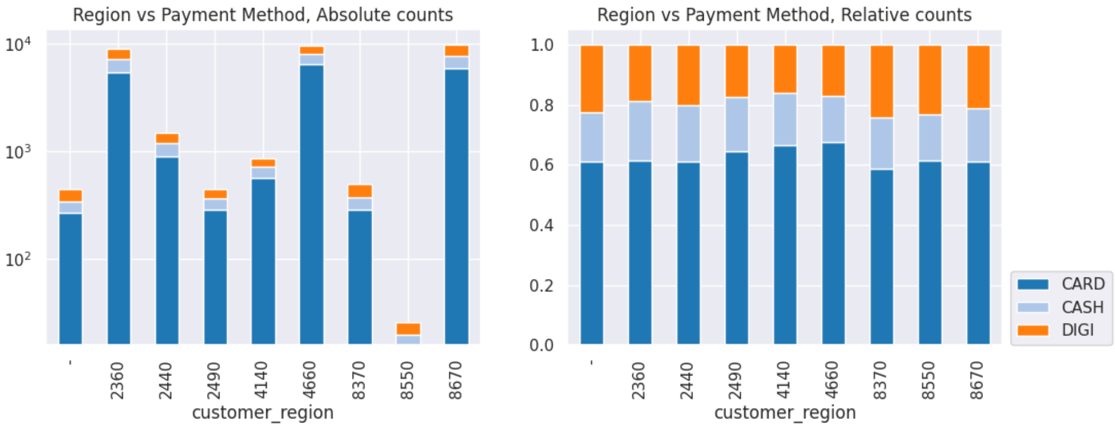

Figure 23 - Logarithmic bar charts for costumer region vs payment method

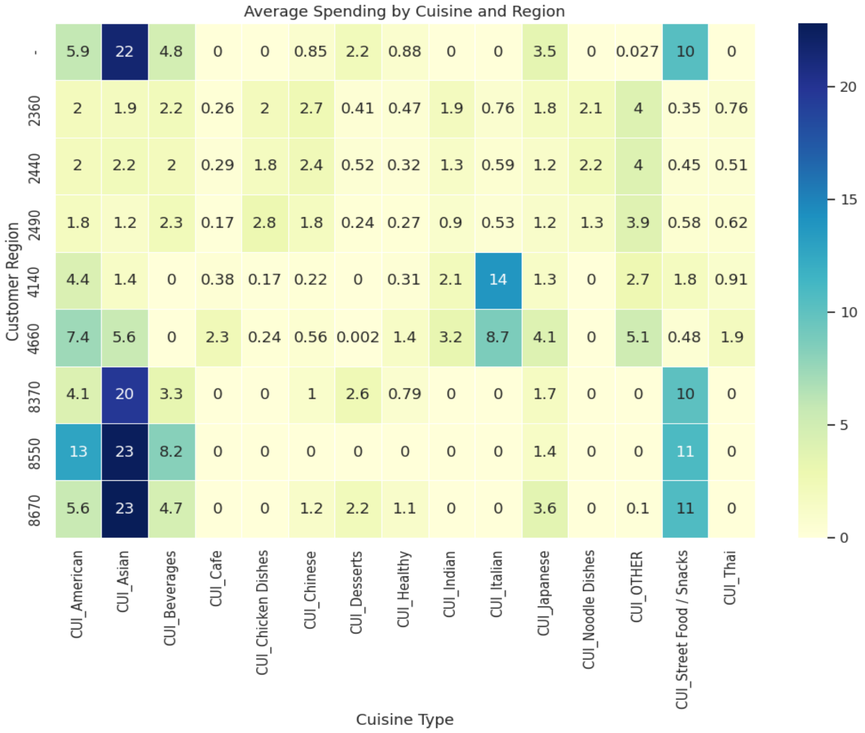

Figure 24 - Heatmap for average spending by cuisine and customer region

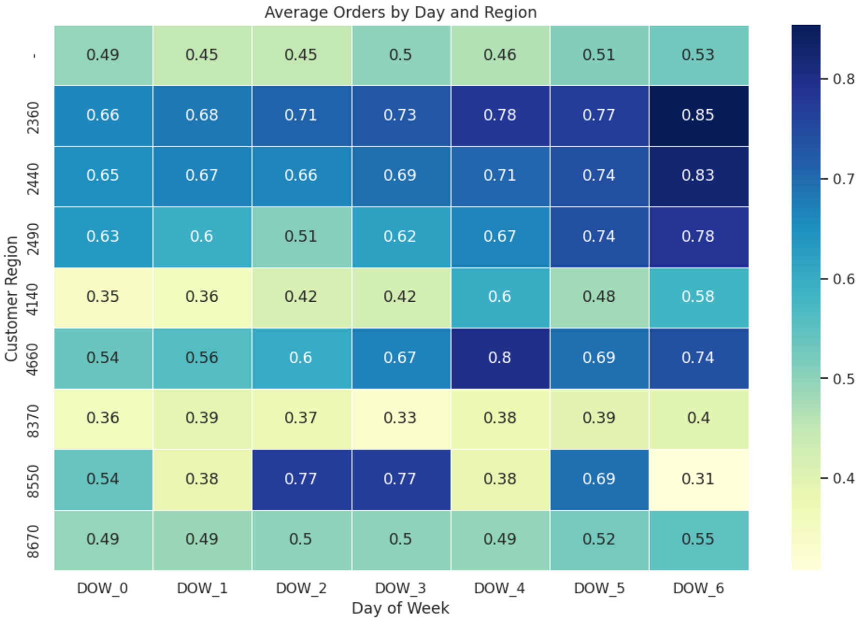

Figure 25 -Heatmap for average orders by day and customer region

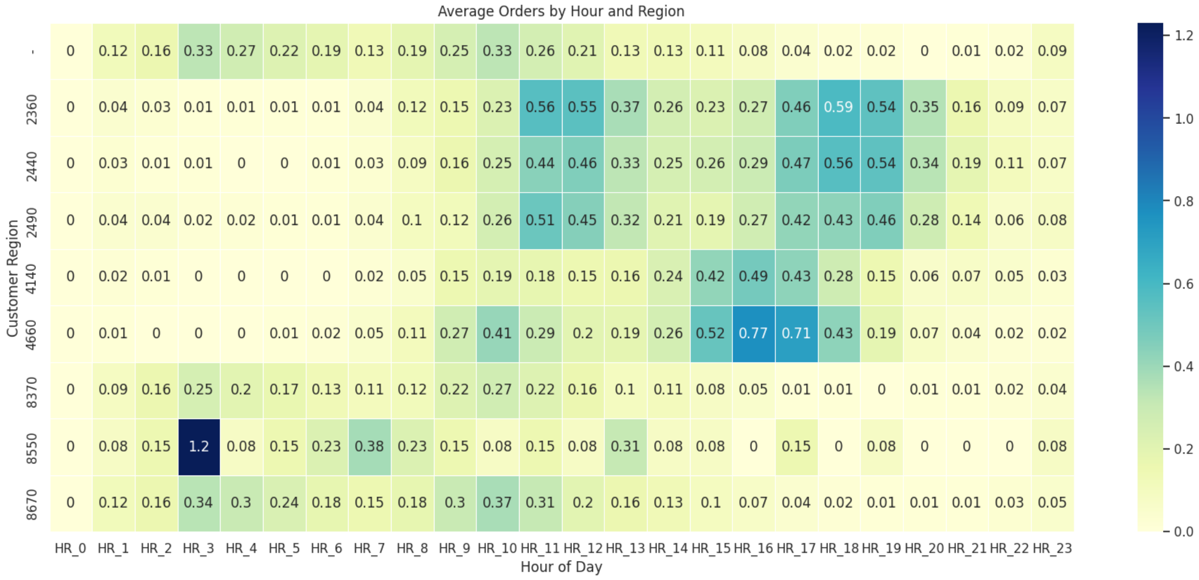

Figure 26 - Heatmap for average orders by hour and customer region

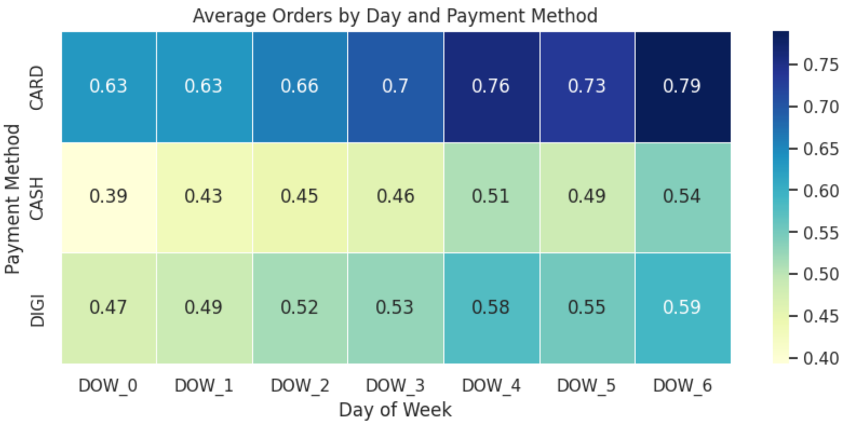

Figure 27 - Heatmap for average orders by day and payment method

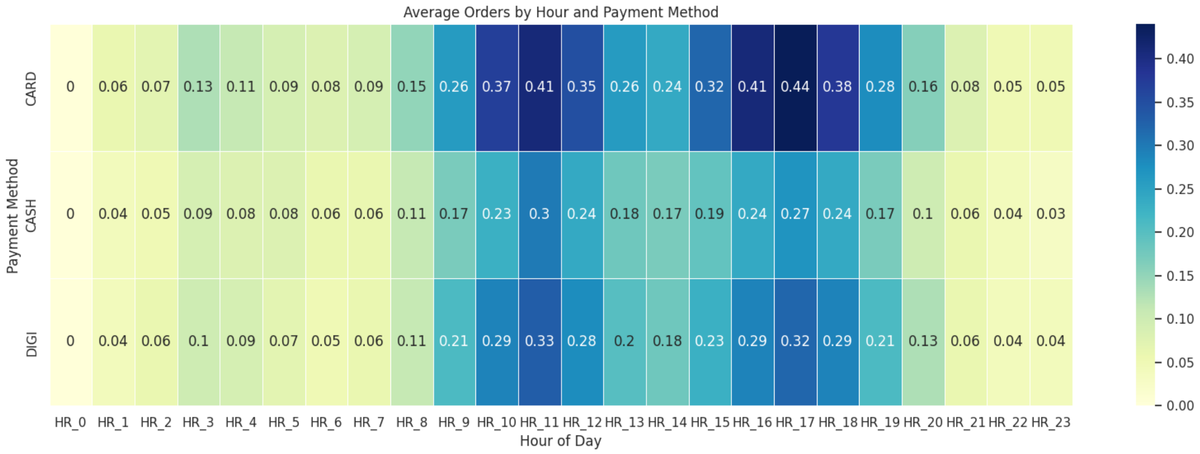

Figure 28 - Heatmap for average orders by hour and payment method

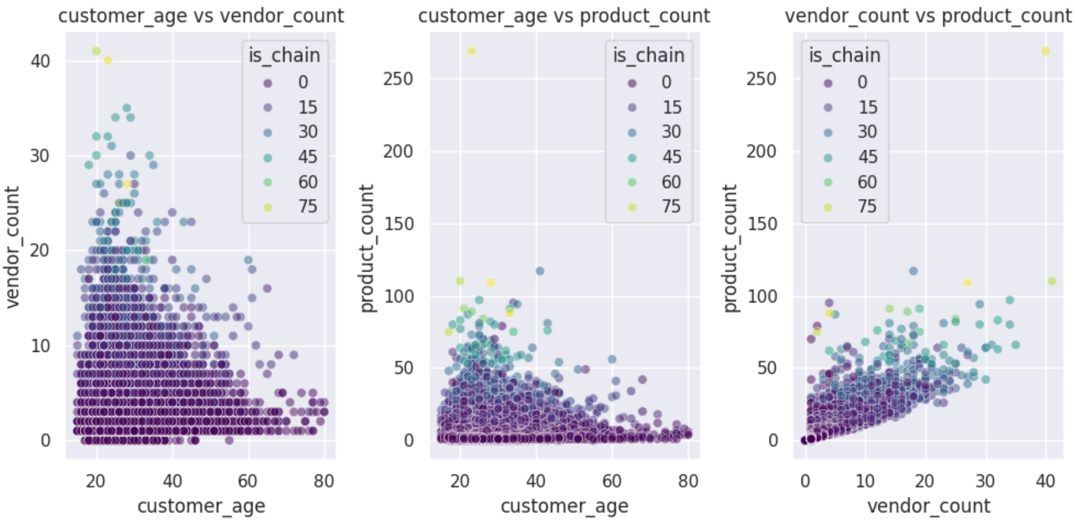

Figure 29 - Pairwise scatter plots of customer age, vendor count and product count by is chain

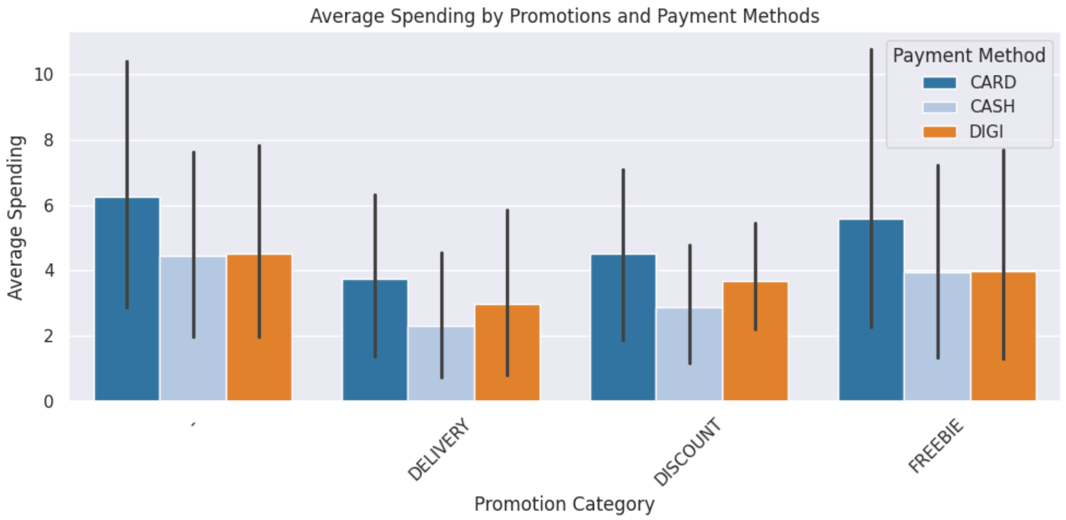

Figure 30 - Bar chart of average spending on cuisines by promotion and payment method