Data Mining Project
Master in Data Science and Advanced Analytics

Nova Information Management School
Universidade Nova de Lisboa

Customer Exploration Report for ABCDEats Inc.

Group 54
Chloé Deschanel, 20240693
Diogo Carvalho, 20240694
Ingrid Lopez, 20240692
Ruben Marques, 20240352

Fall Semester 2024/ 2025

1. Introduction

The objective of this report is to communicate the findings from the exploratory data analysis of ABCDEats Inc., a food delivery service, and identify key customer segments. The aim of the analysis is to help the company create highly effective marketing strategies and improve its service offerings. The dataset includes multiple attributes, such as age, geographic region, customer habits, payment preferences and spending behaviours. During this exploratory phase, the report will include:

- Key trends, patterns, and anomalies within the customer data
- New features to deepen the analysis and provide additional insights
- Visualisations to effectively communicate relationships across features

Overall, these insights will be invaluable in generating inputs to segment customers, develop marketing strategies, and tailor promotional campaigns to reach customer groups effectively.

2. Exploratory Data Analysis

2.1. Dataset Overview

ABCDEats dataset has 31,888 entry rows and 56 columns including customer demographics, behaviours, and spending patterns. There are 4 categorical columns, which are customer ID, region, promotion type, and payment method. The rest of the columns are numeric and pertain to customer age, vendor count, product count, different cuisines spent, and day and time of day for different orders.

To facilitate the analysis, the index is set to customer_id to identify each unique customer. However, 13 duplicate entries were located, where the same customer appears twice in the dataset. These duplicate entries will need to be removed in the next part of the project. Additionally, some features contain missing values, notably in customer_age, first_order, and HR_0, representing orders between midnight and 1 am. The exploratory analysis has been conducted with these missing values in mind, and recommendations for handling these have been included in the report. Accordingly, the next part of the project will apply imputation or filtering techniques as needed to ensure reliability of the insights derived from this dataset.

2.2. Feature Analysis

i. Customer Demographic

The age distribution of customers shows a concentration around the median age of 26, with the majority of the customers being between 18 and 35 years old. This indicates that the primary users of the platform are young adults (20-29), the second largest being adults (30-39) followed by teenagers (15-19). The histogram, with the density curve further reinforces the skew towards ages with very few customers over the age of 50. Thus, missing values could be treated by inputting the median to maintain this skew. (Fig.1)

Regarding customer regions, even though customers are distributed throughout nine areas, the three main hotspots of activity, 8670, 4660, and 2360, represent around 88% of order volume, likely accounting for major cities. It can also be assumed that codes starting with the same digit can be different areas of the same city. For instance, codes 8370 and 8550 are likely to be part of the a bigger city or densely populated area represented by code 8670.

ii. Ordering Behaviours

The mean vendor count of 3.1 and median of 2 suggest that customers tend to order from only a few vendors. The vendor count ranges from a minimum of 0 to a maximum of 41. There are 138 entries with a vendor count of 0, which appears incorrect. Further analysis reveals that customers with a vendor count of 0 also have a product count of 0. This could suggest that these entries might be default records where no orders were placed.

For product count, the mean is 5.67 and the median is 3, indicating that customers generally tend to purchase a modest number of products. Product count ranges from a minimum of 0 to a maximum of 269. Similarly, there are 156 rows with 0 product count. Only 138 of these rows have a vendor count of 0, which is consistent with entries where no orders were made. The remaining 18 entries have a product count of 0, but a vendor count above 0, indicating potential error. It can be assumed that these rows should have a product count of 1, as according to variables DOW_ and HR_, only one order appears to have been placed. Additionally, having a maximum value of 269 could be an outlier, as it significantly deviates from the typical range. While four customers have ordered over 100 products, these values could still be acceptable under certain situations (e.g. big events).

Regarding chain restaurants, this feature reflects the number of orders each customer placed at chain restaurants. The huge majority of customers have very low values of chain restaurant orders (between 0 and 3). As the number of orders from chain restaurants increases, so does the count of customers, with the exception that very few customers place more than 10 orders from chain restaurants.

An average of 2 chain orders per customer shows that most of the customers have 2 or fewer chain orders. The range from 0 to 83 indicates a few outliers with a strong preference for chains, but this is not representative of the majority population. However, the chain order per se is insufficient without knowing each customer's total orders. For example, 3 chain orders out of 3 total shows high preference, while 3 out of 20 indicates low preference.

iii. Spending Patterns

The dataset includes 13 different types of cuisines, each representing the amount spent by customers on a given cuisine type. In terms of total spending (i.e. the cumulative amount all customers spent on each cuisine), Asian cuisine leads with 317,618.9 monetary units, followed by American cuisine with 155,627.4 monetary units. This could suggest that Asian cuisine may either be the most popular or the most expensive. (*Fig.2*)

When examining the average spending per order, Street Food & Snacks leads with an average of 29.37 monetary units, compared to 26.68 for Asian cuisine. This could imply that, although ordered less frequently, Street Food & Snacks has a higher spending per order, suggesting that it may be viewed as an occasional option, while Asian cuisine is a more regular choice for customers.

Additionally, when exploring the cuisine on which each customer spends the most (i.e. their top cuisine), 22.04% of customers spend the most on Asian cuisine, followed by American cuisine with 13.52%. Street Food & Snacks is in 6th position, with 7.19% of customers, consistent with it being a less frequent option, but a higher-value one.

Based on the cuisine type variables, it is possible to calculate the total spent by each customer. On average, customers spend 38.3 monetary units on delivery services, with expenditures ranging from a minimum of 0 to a maximum of 1,418.33. This aligns with the 138 entries where both the vendor and product count are 0 (i.e. default information with no orders placed). The boxplot reveals significant outliers in spending, especially when totals surpass 600 monetary units. (Fig.3)

iv. Time-based Features

The order days (i.e. orders per day of the week) represents the number of orders that occur each day of the week, with a mean of 4.37 and a median of 3. It can be noticed that most customers place a low number of orders during the week, with some customers placing higher than average raising the mean. Furthermore, as the weekend gets closer the number of orders for those days also tend to increase with the busiest days being Thursday, Friday and Saturday.

The orders per hour of the day (i.e. hourly trend) indicate two main spikes during lunch and dinner times. In hour 0, there are missing values and the other values are 0. There is also an inconsistency when comparing the total number of orders from days of the week to hours of the day, which can be due to the missing values in hour 0. For now, the total number of orders based on days of the week will be used, and once duplicates and missing values are treated, it will be possible to have a better view of how to handle this inconsistency. (Fig.4)

v. Promotional Engagement & Payment Methods

The last promotion feature shows the most recent type of promotion or discount used by the customer. In this feature, the majority of the customers (52.52%) have not used any type of promotions since the start of the database. As for the customers that did use a discount or promotion in the most recent order, the most used recently was the Delivery one with 19.71% of the orders, followed by the Discount with 14.10%, and then Freebie with 13.67%.

Regarding the preferred payment method, the analysis shows the distribution of the customers, through card, cash, and digital payments. The graph indicates that 63.22% of customers pay by card and 19.13% use DIGI. Only 17.64% of customers use cash as a more conventional payment method. These findings may translate to future decisions on promotions, incentives, or payment methods.

2.3. Feature Engineering

From the above analysis, the variables below have been created for further exploration of the dataset:

- **Age Group:** Categorize customers into age groups (teenagers, young adults, adults, middle-aged, older adults, seniors).
- **Vendor Diversity:** Categorize vendor choice diversity (low, moderate, high).
- **Active Period:** Calculate active days per customer (difference between last_order and first_order).

- **Total Spending:** Sum of monetary units across cuisine types to get a customer's total spending.
- **Top Cuisine:** Identify the cuisine type with the highest spending per customer.
- **Total Number of Orders:** Calculate the total number of orders per customer based on orders made during weekdays.
- **City:** Classify each region code into one of three cities based on the starting digit of the code.
- **Promo Used:** Indicate if any type of promotion was used in the last order.

The below features also have been added for further analysis:

- **Spending habit:** Categorise customers based on total spending, using percentiles. Categories include: "Low" for customers spending less than 6.72, "Medium" for spending between 6.72 and 24.14, "High" for spending between 24.14 and 84.34, and "Very High" for spending above 84.34.
- **Week Periods:** Group DOW_ into week periods: start of the week (Monday & Tuesday), mid week (Wednesday & Thursday), and end of the week (Friday, Saturday & Sunday).
- **Time of Day:** Group HR_ into time of day: night (from 10 pm to 5 am), morning (from 6 am to 10 am), midday (from 11 am to 2 pm), afternoon (from 3 pm to 5 pm) and evening (from 6 pm to 9 pm). HR_0 is not relevant as it contains missing values or values of 0.
- **Non chain orders:** Calculate the number of non-chain orders by taking the difference between total_number_of_orders and is_chain.
- **Chain preference:** Classify customer preferences based on their preference for chain restaurants or not.

2.4. Multivariate Analysis

payment_method & product_count: Customers who use card as a payment method, on average tend to order more products.

city & top_cuisine: city_8's top cuisine is Asian, while city_4 favours Italian, and city_2 favours "Other", followed by Chinese. In city_2, it is harder to distinguish a top_cuisine as there is a bit of every cuisine type. This pattern suggests that each city may have a preferred cuisine type or fewer options available in certain cuisine categories. (Fig.5)

spending_habit & order behaviour (i.e. vendor_count & total_number_of_orders): Higher spending customers tend to place more frequent orders and order from a greater number of vendors. Specifically, customers in the "Very High" spending category have significantly higher average order frequency (13.14 orders) and vendor diversity (7.44 vendors) compared to lower spending categories. This suggests that high spenders are also more engaged with the platform, both in terms of frequency and variety of vendors. (Fig.6)

last_promotion & total_spent: Customers who did not use a promotion ("-") have the highest average total spending, followed closely by those who used "Freebie" promotions. This suggests that non-promotion users and freebie users may tend to be higher spenders, while "Delivery" and "Discount" promotions are associated with slightly lower spending on average.

chain_preference & total_spent: The analysis reveals that 59% of customers prefer chain restaurants, while 31% favour non chain ones, and 10% show a balanced preference. Further analysis examines whether customer preference for chain vs. non-chain restaurants impacts total spending by categorising preferences and comparing average spending across groups. Results show that non-chain preferred customers tend to spend more on average than chain-preferred customers. (Fig.7)

Week period and time of day: The peak ordering time happens at the end of the week (Friday, Saturday, and Sunday) around midday (i.e. lunch time). Generally, midday is the busiest ordering period throughout the week. In contrast, the least number of orders happen at the beginning, on Monday and Tuesday evenings. (Fig.8)

City & spending_habit: city_8 has customers with the highest spending habit, which could indicate it's the wealthiest city. city_4 shows a balance between medium and high spenders, and city_2 primarily consists of medium spenders. (Fig.9)

spending_habit & top_cuisine: Asian cuisine and Street Food & Snacks have the customers with the highest spending habits. (Fig.10)

3. Conclusion

The exploratory data analysis provides valuable information about the dataset and its anomalies, as well as customer demographics, behaviours and patterns. In addition to the anomalies mentioned in the report, others have been identified along the analysis which will also need to be dealt with (i.e. see non_chain_orders in the notebook). This exploratory data analysis forms the basis for the next phase of the project, and is necessary to accurately preprocess the data, identify clusters in final segments, and provide business recommendations for ABCDEats Inc.

4. Appendix

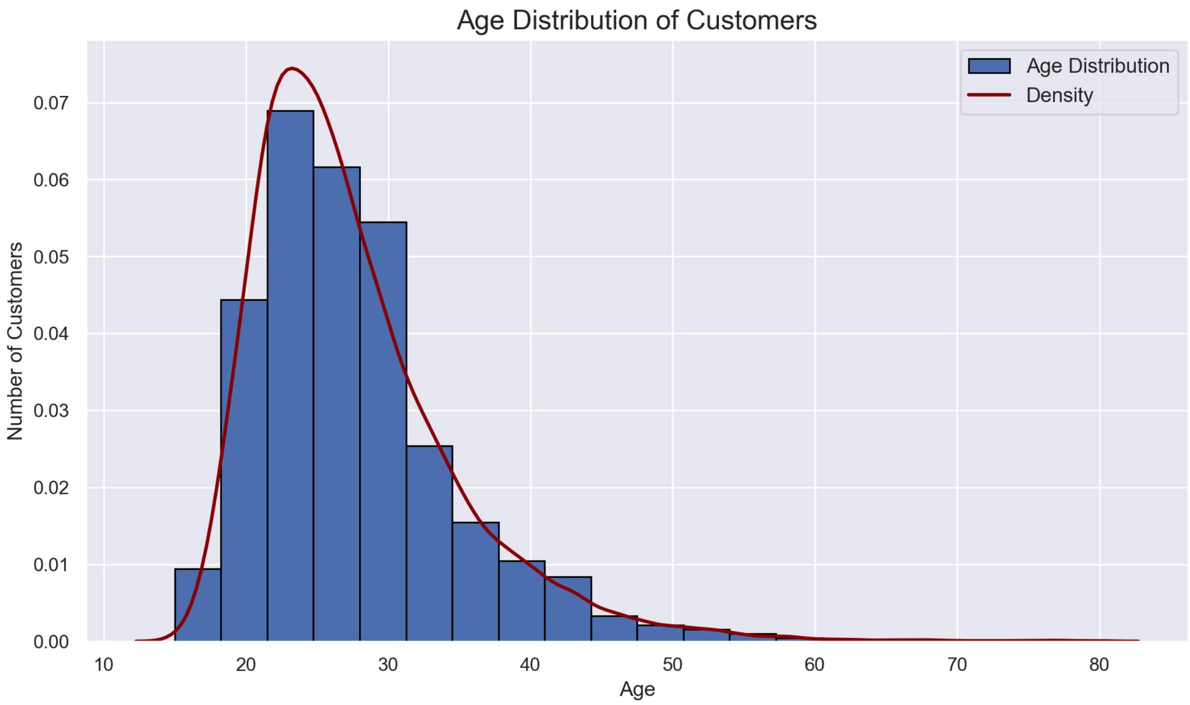


Fig.1 - Age Distribution of customers

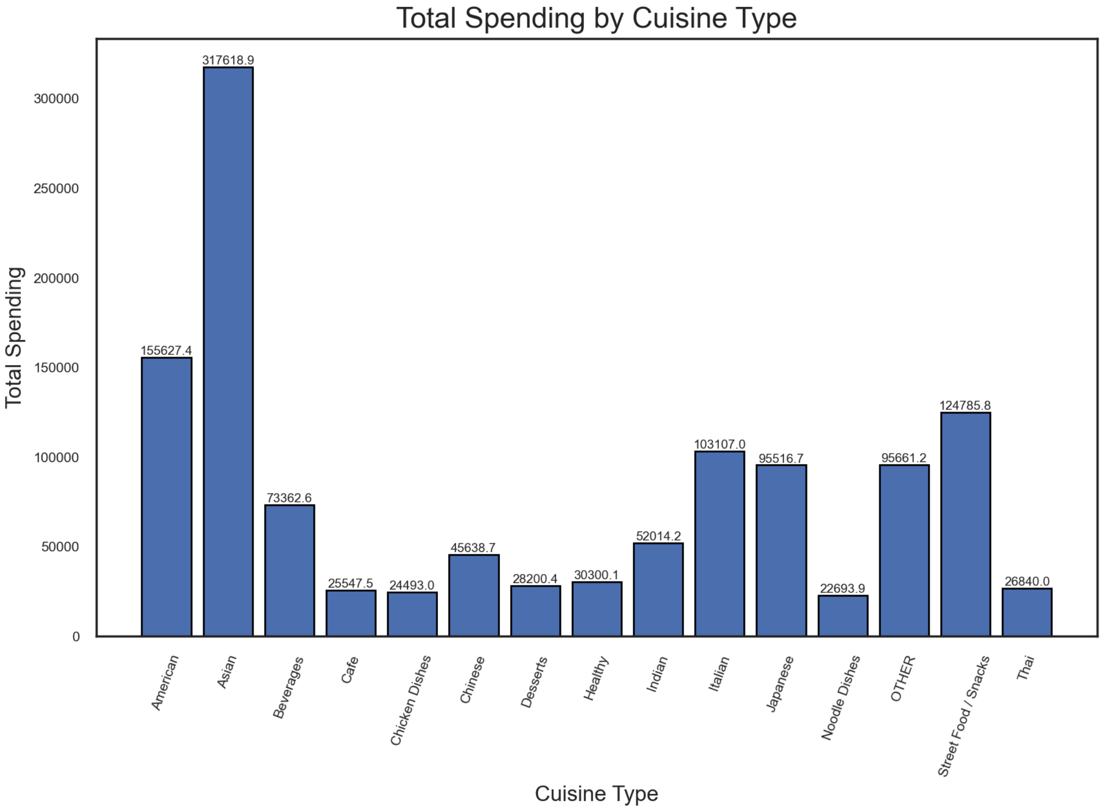
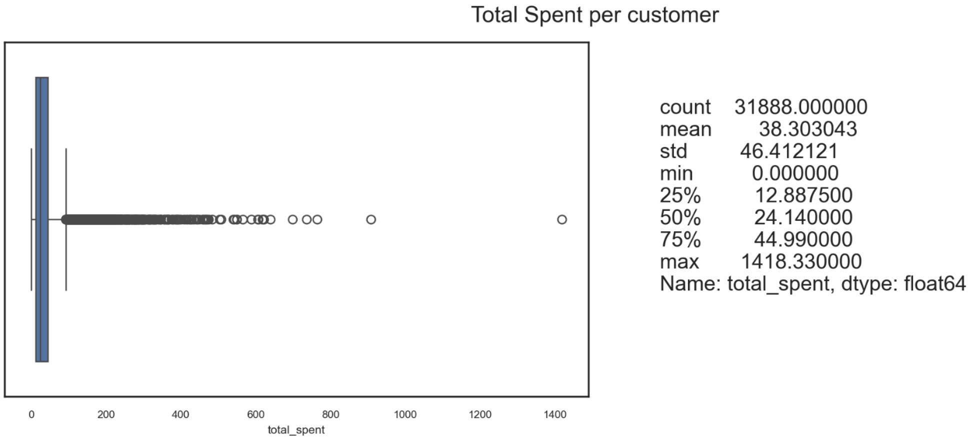


Fig.2 - Total Spending by Cuisine Type



```
count    31888.000000
mean        38.303043
std         46.412121
min          0.000000
25%         12.887500
50%         24.140000
75%         44.990000
max       1418.330000
Name: total_spent, dtype: float64
```

Fig.3 - Total Spent per customer

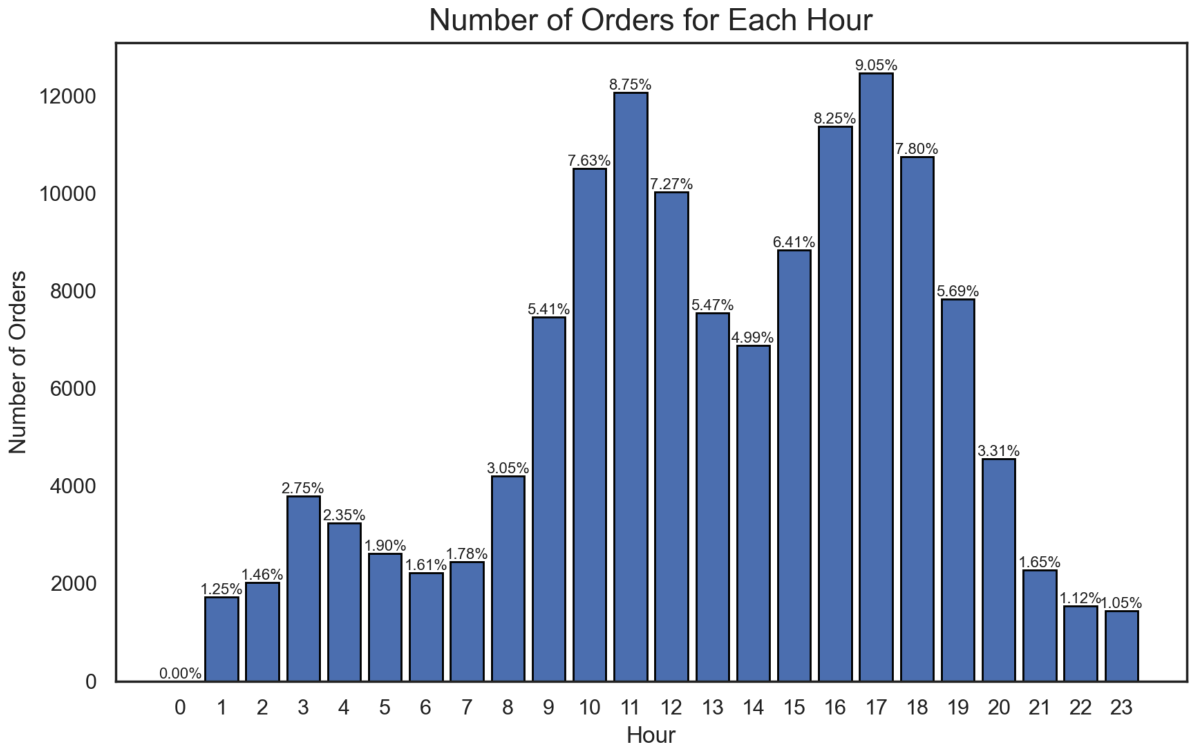


Fig.4 - Number of Orders for Each Hour

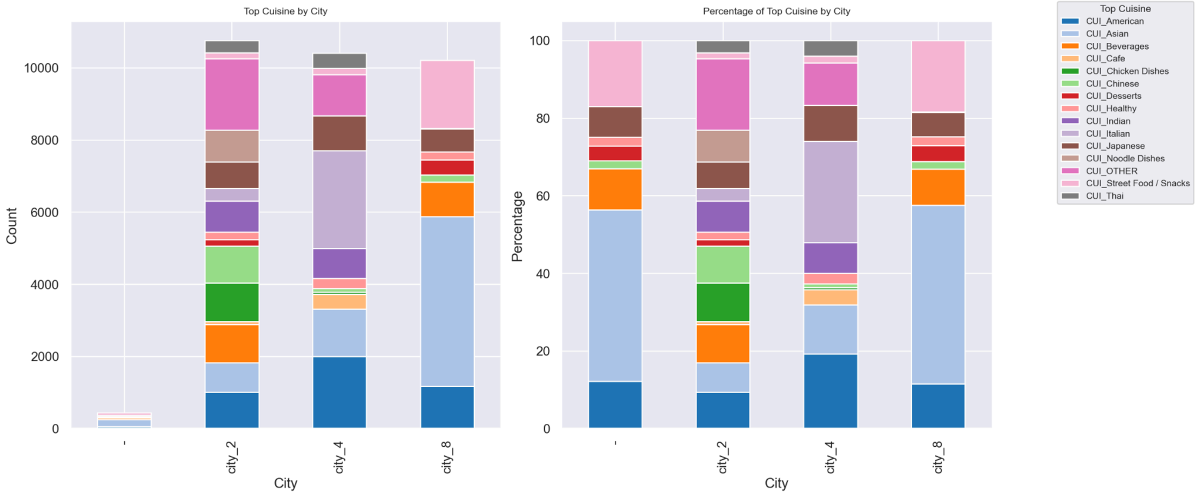


Fig.5 - Top cuisine by city count and percentage

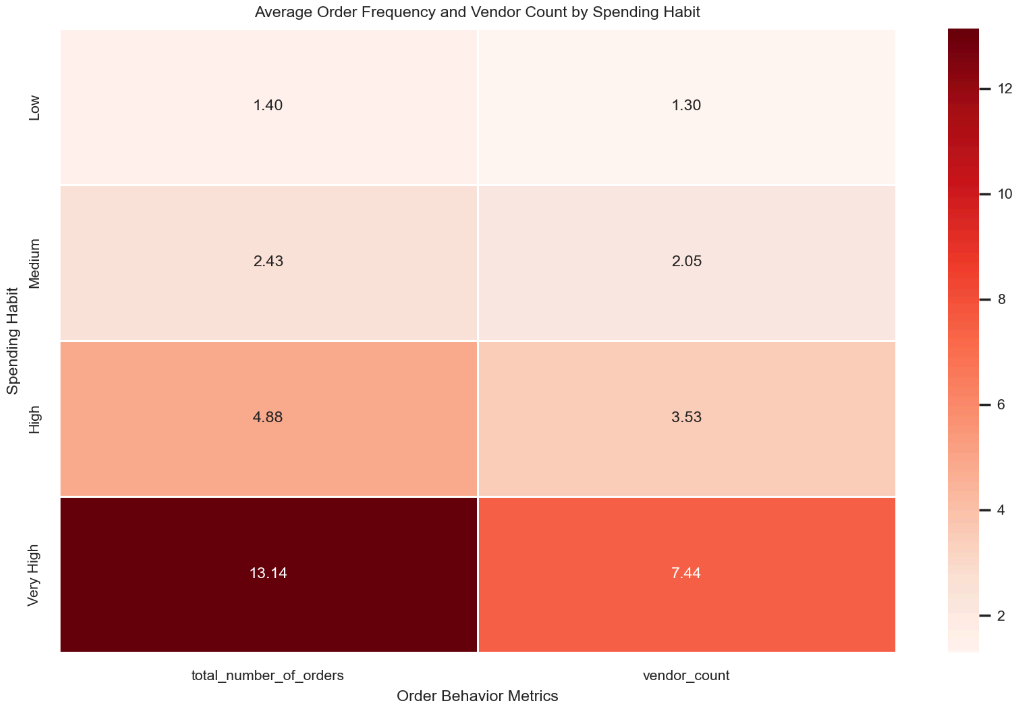


Fig.6 - Average Order Frequency and Vendor Count by Spending Habit

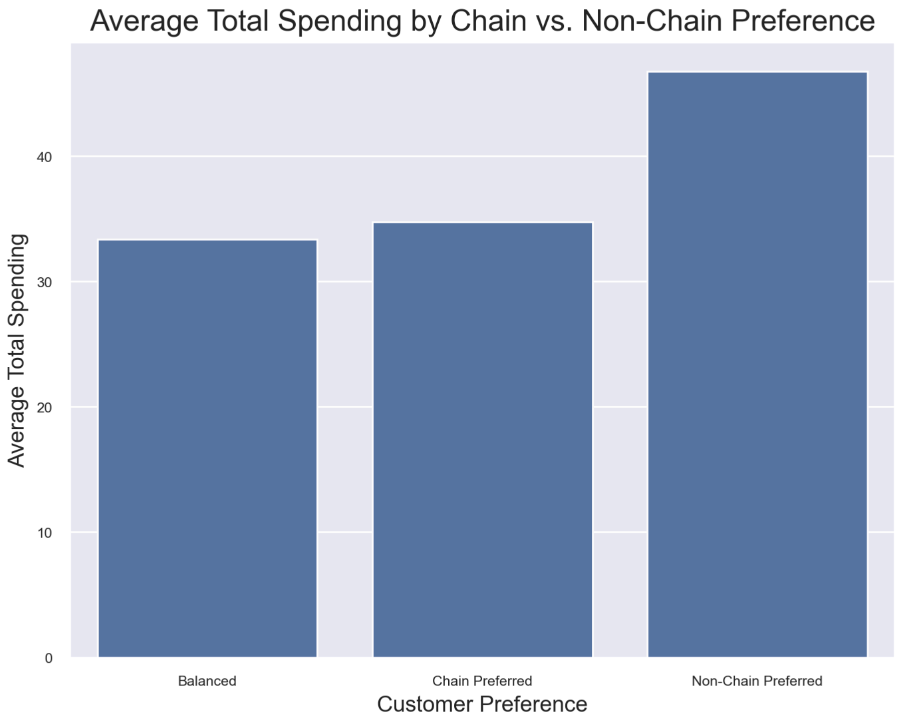


Fig.7 - Average Total Spending by Chain vs. Non-Chain Preference

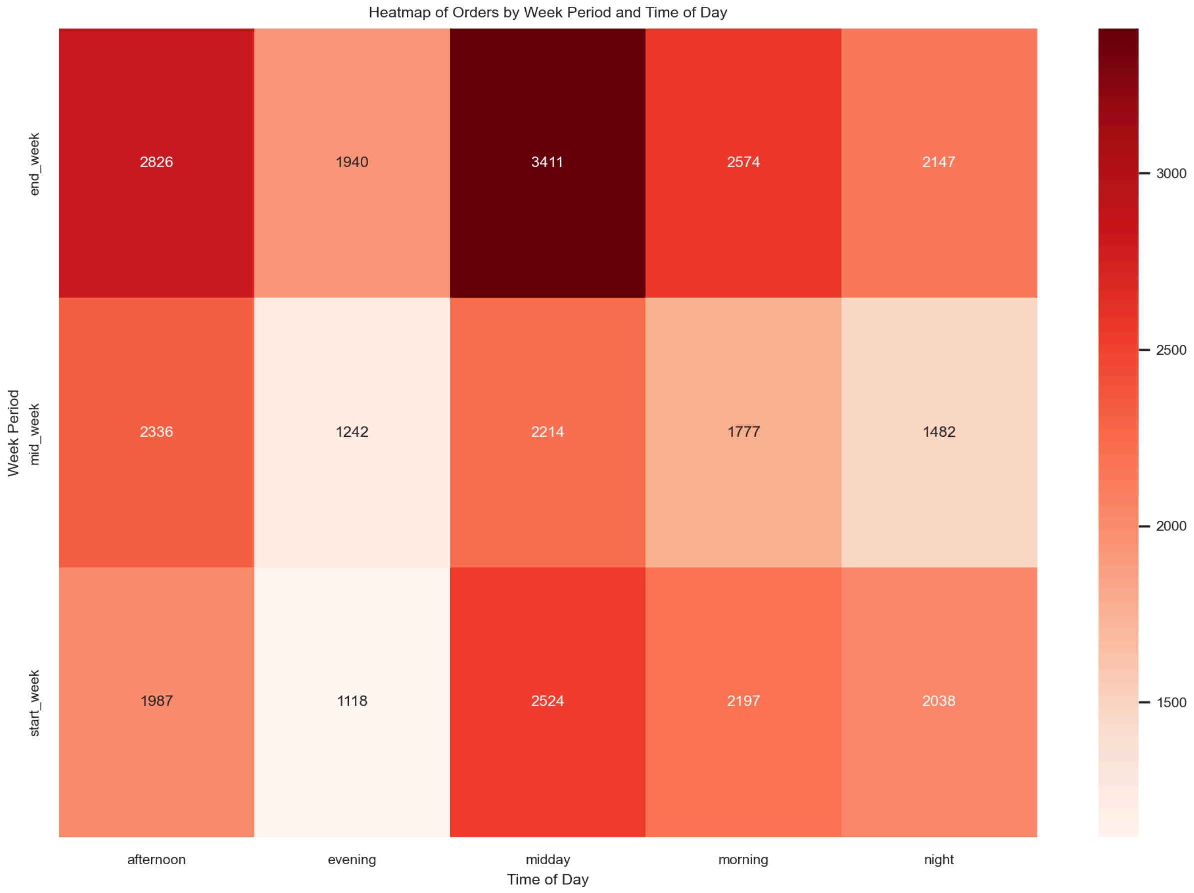


Fig.8 - Heatmap of Orders by Week Period and Time of Day

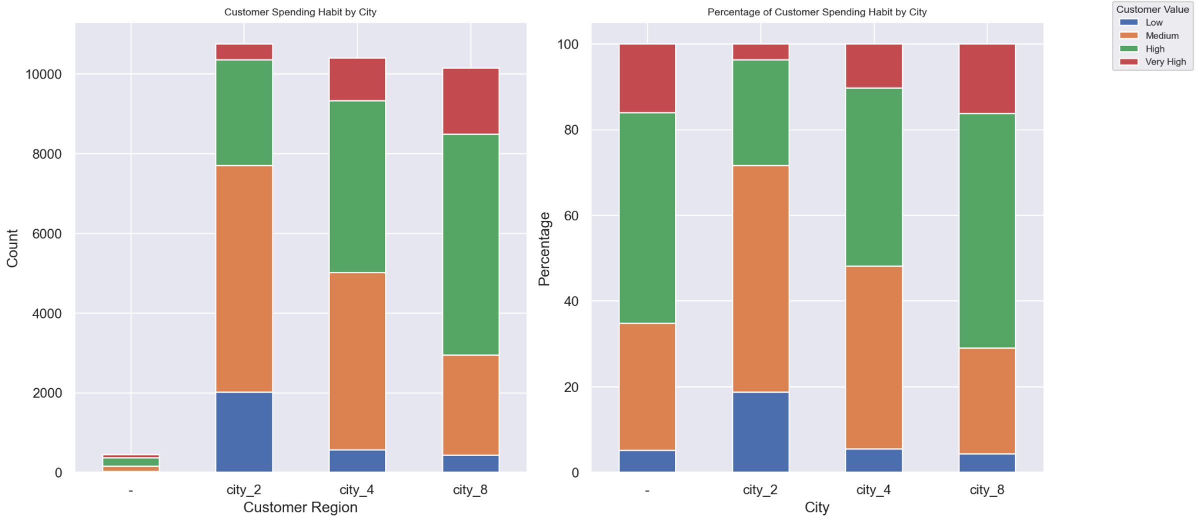


Fig.9 - Customer Spending Habit by City Count and Percentage

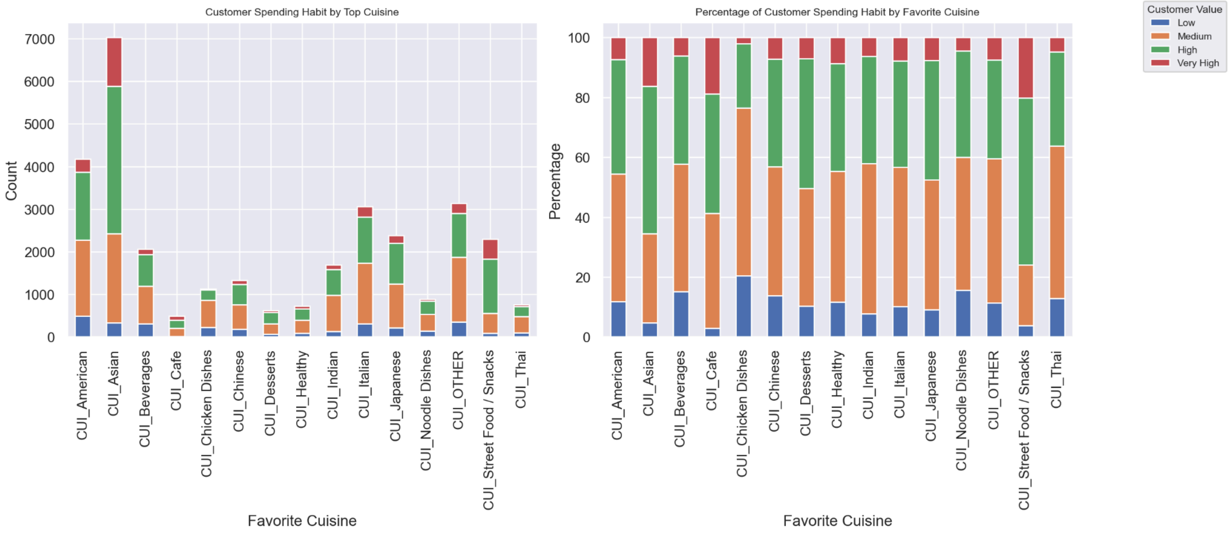


Fig.10 - Customer Spending Habit by Top Cuisine Count and Percentage