DATA MINING PROJECT

Master's in data science and advanced Analytics

NOVA Information Management School

Universidade Nova de Lisboa

ABCDEats Inc.

Exploratory Data Analysis

Group 02

Sarah Leuthner, 20240581

Catarina Silva, 20240558

Bruna Duarte, 20210669

Afonso Gião, 20240495

Fall/Spring Semester 2024-2025

TABLE OF CONTENTS

INTRODUCTION

Businesses today face new challenges as consumers become more selective about their spending habits. Therefore, the food delivery service, *ABCDEats Inc*., will make use of customer segmentation techniques to meet the preferences and needs of each customer. To develop sustainable and economic strategies, the company must first gain a comprehensive understanding of its customer base. The main goal of this report is to find descriptive patterns among customers, as well as try to understand customer behavior to offer the best personalized strategies for each customer type, starting with the data exploration.

1. DATA EXPLORATION

1. Key statistics

To initiate the data exploration, the primary statistics are examined. The first key statistics table (Table 1) focuses only on the numeric data, with the exception of the values in CUI, DOW, HR. The *Customers Age* ranges between 15 and 80 years with an average of 27 years. Half of their customers are aged between 23 and 31, which indicates a significant presence of young adults. The food delivery service should focus on this segment as core customers, nonetheless, expanding the business to other age groups.

On average, a typical customer orders from 3 different unique *Vendors*; a maximum of 41 different vendors served at least one customer. Given that 50% of our buyer's orders are from 1-4 vendors, the maximum presents itself as an outlier. As it seems likely that most customers use the same vendors regularly, a possible marketing strategy is using loyalty cards (10 meals – one free). The minimum value of 0 on this metric implies that the company has registered customers that have not ordered yet, which creates a possibility for a "first-order" discount. This occurrence can also be seen when the *Product Count* is 0. The average of this metric is 5.67. While 50% of our customers have ordered 2-7 products, the maximum products ordered from a customer is 269 products (which could be an outlier, for example, a customer using the services for catering purposes, or an extremely active customer with a lot of past orders). The low product count indicates that *ABCDEats Inc.* has only a few loyal customers and the broad use of its services is not used regularly or on a long-term basis. Therefore, more incentives in general for all customers are necessary.

According to the *Is Chain* feature, on average a customer places 2-3 orders that are from chain restaurants. This average is highly influenced by the maximum value of 83. For instance, 50% of the customer base places less than 2 chain orders, which is less than the average. The dataset includes data from the first 90 days (3 months) of the company's services, as evidenced by the maximum value in the *Last Order* field being 90. According to *First Order*, on average the customers started placing orders from the 28th day of services. Three quarters of the customer base joined in the first half of the 90-day period pointing to a trend of slower customer growth. This means a strategy to consistently gain new customers for the company is needed. An important feature is also the *Last Order,* with an average of 64 days (about 2 months). Most of the last orders are from 49 to 90 days from the start of the dataset, which means that most customers ordered quite recently. Only a few last orders are from the beginning of the dataset reflecting the company has steady orders. The distributions of the orders per hour of the day, or day of the week and money spent on each cuisine are shown in the histograms in the Appendix.

Regarding the different cuisines offered (Figure 1), by far the most monetary unit spent by an average customer is for Asian Food followed by American Foods and Street Food / Snacks. Cuisines

that do not earn even one monetary unit per average customer include: Café, Chicken, Dessert, Healthy, Noodle and Thai. *ABCDEats Inc*. can either promote neglected cuisines with discounts or increase incentives for the popular cuisines to further increase their revenue. The orders tend to increase starting on Sunday and have a bigger volume from Thursday to Saturday (Figure 2). Most orders are placed on Thursday as well as Saturday and from 9-13 am and from 3-7 pm, which are the two order time peaks (lunch and dinner). There is a small order peak around 3 am, which could be considered a late-night snack (Figure 3).

The distribution for the categorical features is shown in the pie charts in the appendix (Figure 4). Almost 90% of customers are from three of all registered *Customer Regions*, which could be bigger cities and should be the focus for strategies. The remaining 10% are from 4 different regions which could be smaller towns around the cities. Within this 10% there are also customers with unspecified locations, which could include customers who have not ordered yet and therefore did not put their address in the system or missing data. More than half of the customers did not use *Promotion Codes* in their last purchase, which should be taken into account in further marketing campaigns. The other half used promotions for delivery, discounts and freebies and are almost evenly distributed, with delivery promotions taking a bigger slice. Regarding *Payment*, the most used method in last purchases with almost two-thirds of the customers is card payment followed by digital and cash.

1.1.1. **Cuisine Analysis**

For further analysis on metric variables, another key statistics table was created (Table 2) in order to study cuisine types. Most of the cuisine's data is very sparse, meaning there are a lot of null values. The only exception to this is the American and Asian cuisines. On average, a client spends 9,96 monetary units on the Asian cuisine, which is a high value compared to the other cuisines. This cuisine also has a high variance, possibly because it comprehends a huge range of different cuisines in the Asian continent. The American cuisine has more orders as a quarter of its values are between 5,66 and 280,21 monetary units.

When comparing the Japanese, Indian, Chinese and Thai cuisines (Figure 5), the Japanese cuisine has the highest amount of money spent with almost double the money spent on each of the others possibly given the popularity and higher average cost. The Thai cuisine has the least amount of money spent among the 4. Further, when comparing the labeled Asian cuisine data with the other Asian cuisines previously observed, a hypothetical relationship can be seen (Figure 6). If *CUI_Asian* contains the previously observed Asian countries cuisines, then around 2/3 of the money spent on Asian cuisines comes from only Japanese, Indian, Chinese and Thai cuisines. Figure 7 showcases this scenario illustrating the weight of these 4 cuisines in the total Asian cuisine spending, with the Japanese cuisine having a significant impact. Furthermore, Figure 8 compares general cuisines with the rest of the cuisines, displaying the stacked totals of American, Asian, Italian and other cuisines, paired with the rest of the cuisine columns. There is a good chance that these 4 cuisines contain the other types of cuisines in the dataset because the total spending in these 4 cuisines is greater than the total spending in other cuisines. Comparing these 4 cuisines (Figure 9), we see that Asian cuisine has by far the higher amount of money spent, followed by the American and Italian cuisines. Among these types of food, the one that has the biggest spending is street food and snacks (Figure 10).

2. Data Analysis

1.1.2. New features

For deeper analysis, new features were created in addition to the original features. A detailed list of every feature can be found in the appendix (Table 3). To simplify the many features, new features like the sums of specific *DOW* or *HR* as well as total sums of *Orders* and *Monetary Unit Spend* were calculated. The new feature *Customer Duration* shows the number of days a customer has actively used the food delivery services by subtracting the *Last Order* from the *First Order*. The feature *Peak Order Day* identifies the day of the week with the highest order volume by finding the column with the maximum value across fields *DOW_0* to *DOW_6* (representing daily order counts for each day of the week). The *Peak Order Hour* was similarly created. The *Order Frequency* metric calculates the average order rate for each customer by dividing the total order count by the number of days between their first and last orders. This feature provides insight into the customer's ordering frequency over a period of time. The *Churn Risk* (Figure 11) feature is a binary indicator that flags potential churn risk. It is set to 1 if the number of days since the customer's last order exceeds a predefined threshold (set to 30 days, within a maximum limit of 90 days); otherwise, it is 0. This feature allows us to identify customers who may be at risk of leaving. The *Repeat Customer* (Figure 12) feature identifies customers who have placed more than one order, allowing segmentation of repeat customers from one-time buyers. More features about *average spendings* and *cuisine information* were created that will help proceeding with the segmentation about customers.

Within the new feature *Order Intensity*, most of the data points are concentrated on the lower end (close to zero), indicating that most customers place only a few orders. Therefore, the frequency of customers with low order intensity is significantly high, (Figure 13). The long tail suggests the presence of outliers — customers who order more than 2 or 3 times per active day are not common. These outliers might warrant further investigation to see if they represent a specific customer segment or behavior that differs from the norm. This distribution means that most customers do not make frequent or large orders daily. The few customers who do have higher order intensity might represent more valuable or high-frequency shoppers. Understanding this skewed distribution could help in segmenting customers based on their order intensity. Businesses could target high-intensity customers with loyalty programs or promotions, while encouraging low-intensity customers to increase their order frequency through personalized offers.

1.1.3. Trends and Patterns

After exploring the data and creating new variables to help enhance analysis, looking at each feature individually was the main priority, to identify their distributions, specific characteristics and possible outliers or impossible values. For the metric variables, both histograms and box plots were implemented for each feature, while count plots were implemented for the categorical features. Regarding the first set of features, the metric ones, a first analysis of the **histograms** (Figure 14) emphasized a clear discrepancy between the features. *Customer Age* shows a young customer base, represented by a clear skewness towards lower values as already stated in the key statistics. Both *Vendor Count* and *Product Count* are right-skewed, meaning most of our customer base buys from fewer than 5 vendors and fewer than 50 products. Furthermore, it is interesting to notice that *First Order* and *Last Order* have a remarkably similar graphic representation, almost symmetrical. This supports the previous conclusions of slower customer growth and good customer retention. Similarly,

Spend per Product demonstrates that the customers make small purchases. Moving on to the multiple cuisine types, the histograms suggest that our customers do not order from a wide variety of cuisines and *CUI American* and *CUI Asian* show a favoritism over the others (Figure 1), conclusions that are supported by the *Cuisine Diversity* and *Favorite Cuisine Concentration* graphs. The sharp slope in *Product Frequency* and *Customer Duration* depicts that our customer base has a short duration of activity, and a low frequency of products ordered. Lastly, according to the *DOW* and *HR* histograms, most periods have (not surprisingly) low values.

The box plots (Figure 17) sustain all the conclusions reached in the analysis of the histograms and amplify even more the presence of outliers in our features, especially *Customer Age*, *Vendor Count*, *Product Count*, *Total Spend*, and *Average Spend Per Vendo*r. It's also worth mentioning the tight interquartile range of certain variables, like *Product Count*, *Is Chain*, *Total Spend*, *HR* and *DOW*. These insights ensure that the data used in analysis is both high-quality and meaningful, enabling more accurate and actionable results.

1.1.4. Correlations

The scatter plot is a useful method for providing a first look at bivariate data to see clusters of points and outliers, or to explore correlations. With the intent of studying the pairwise relationship of Numerical Variables (Figure 18), we analyzed the scatter plots between a selected subset of variables. After a closer look at each graph, it is possible to distinguish a positive correlation between *Total Spend* and *Average Spend per Vendor*, which portraits a tendency for customers with higher expenditures buying from more vendors. *Product Frequency* and *Customer Duration* also correlate positively, indicating that long-term customers consume products frequently. Moreover, there seems to be a high relation between *Is Chain* and *Weekdays*, suggesting there is a higher demand for chain restaurants orders (and, therefore, food and services in which the customers have more trust in or are more familiar with) during business days. A similar conclusion is reached by analyzing the scatter plot between *Lunch Hour* and *Is Chain*, and *Lunch Hour* and *Weekdays*. There also seems to be a relation between *Average Spend per Order* and *Spend Per Product,* and also *Average Spend Per Vendor*, meaning the more a customer spends in one order, the more they spend per product bought and, consequently, the more they spend on the vendor. In terms of negative correlations, a relation between *Customer Duration* and *Weekdays*, which means that the longer someone has been a customer, the lower the number of orders they place during business days.

1.1.5. Heatmap

Within the heatmap (Figure 19), we can find **high positive correlations**, such as *Weekdays* and *Weekend days* (0.59), which shows that customers who are active on weekdays also tend to be active on weekends. *Order Frequency* and *Product Intensity* (0.95) are very strongly correlated, indicating that frequent orders are associated with higher product intensity. *Favorite Cuisine Concentration* and *Spend Per Product* (0.81) also show a strong positive correlation, suggesting that higher concentration on a favorite cuisine aligns with higher spending per product. On strong **negative correlations**: *Favorite Cuisine Concentration* and *Cuisine Diversity* (-0.81) indicate that as the concentration on the favorite cuisine increases, the diversity in cuisines decreases significantly. *Churn Risk* and *Favorite Cuisine Concentration* (-0.52), a higher focus on a particular cuisine (lower diversity) is associated with lower churn risk, possibly indicating loyalty. *Favorite Cuisine Concentratio*n and *Chain Preference* (-0.57), those with a strong preference for certain cuisines may not be as inclined toward chain restaurant orders, showing that cuisine preference diverges from chain loyalty. Furthermore, we can find some

interesting feature interactions, *Churn Risk* shows some notable correlations, such as *Favorite Cuisine Concentration* (-0.52), indicating a lower churn risk is associated with high concentration on a favorite cuisine. *Order Frequency* (-0.34), indicating that customers who order frequently are less likely to churn. *Customer Duration* has moderate to strong positive correlations with multiple variables, for instance *Cuisine Diversity* (0.62) and *Weekdays* (0.54), implying that longer-term customers may engage more in diverse cuisines and order on weekdays.

Some features have relatively **weak or minimal correlations**, like *Chain Preference* with most other features, suggesting that chain preference does not strongly impact other variables in this dataset. *Peak Order Day* has low correlations with most features, indicating that the specific day with the highest orders does not significantly align with other behavioral or demographic characteristics. The negative correlation of *Churn Risk* with *Favorite Cuisine Concentration* and *Order Frequency* suggests that customer loyalty might be tied to frequent orders and a preference for a particular cuisine type. High correlations between *Spend Per Product* and *Favorite Cuisine Concentration* may indicate that customers with strong cuisine preferences tend to spend more per product, which could be used to tailor marketing strategies. Strong correlations within time-based features (e.g., between *Morning_HR*, *Lunch_HR*, *Afternoon_HR*) show that customers have consistent order patterns across time, which can help optimize promotion timing.

1.1.6. Anomalies

Firstly, in the **metadata**, the definition of the *Is Chain* feature needs to be discussed. The description of the metadata states that the feature "indicates whether the customer's order was from a chain restaurant" which would indicate a Boolean feature (1/True - the order is from a chain; 0/False - the order is not from a chain). However, the values range from 0 to 80. Except for the possibility of incorrect data, the feature *Is Chain* most likely is the number of orders from chain-vendors per customer.

Furthermore, the dataset contains **missing values** for the features *Customer Age* (2,3%), *First Order* (0,3%), and *HR_0* (3,7%), which need to be treated. For instance, the mean or median *Customer Age* and *First Order* can be used in replacement, so the missing values will not impact any statistics. The missing data in HR_0 (midnight hour) can be recovered. As this is the only hour with missing data, the sums of all orders per hour and all orders per week can be compared, as they should add up to the same value. A difference should be found, when the value of HR_0 is missing, which then can be replaced with the found remainder. Any **duplicate values** in the data set will be dropped to not misguide the analysis. Moreover, **"nonsense" values** in the dataset are found. This means customers with no products orders, therefore also 0 vendors, no money spent on any cuisine and a lower number of products than orders. These could be customers who registered and did not use the food delivery services yet. With this data ABCDEats Inc. could build a segment "customers to be" as they have a chance to reach the customers. When analyzing these customers, we notice a similar right-skewed age distribution as the whole dataset, but two specific customer regions, where one customer region makes up 90% (Figure 20). This region should be the focus for the strategy.

APPENDIX FIGURES

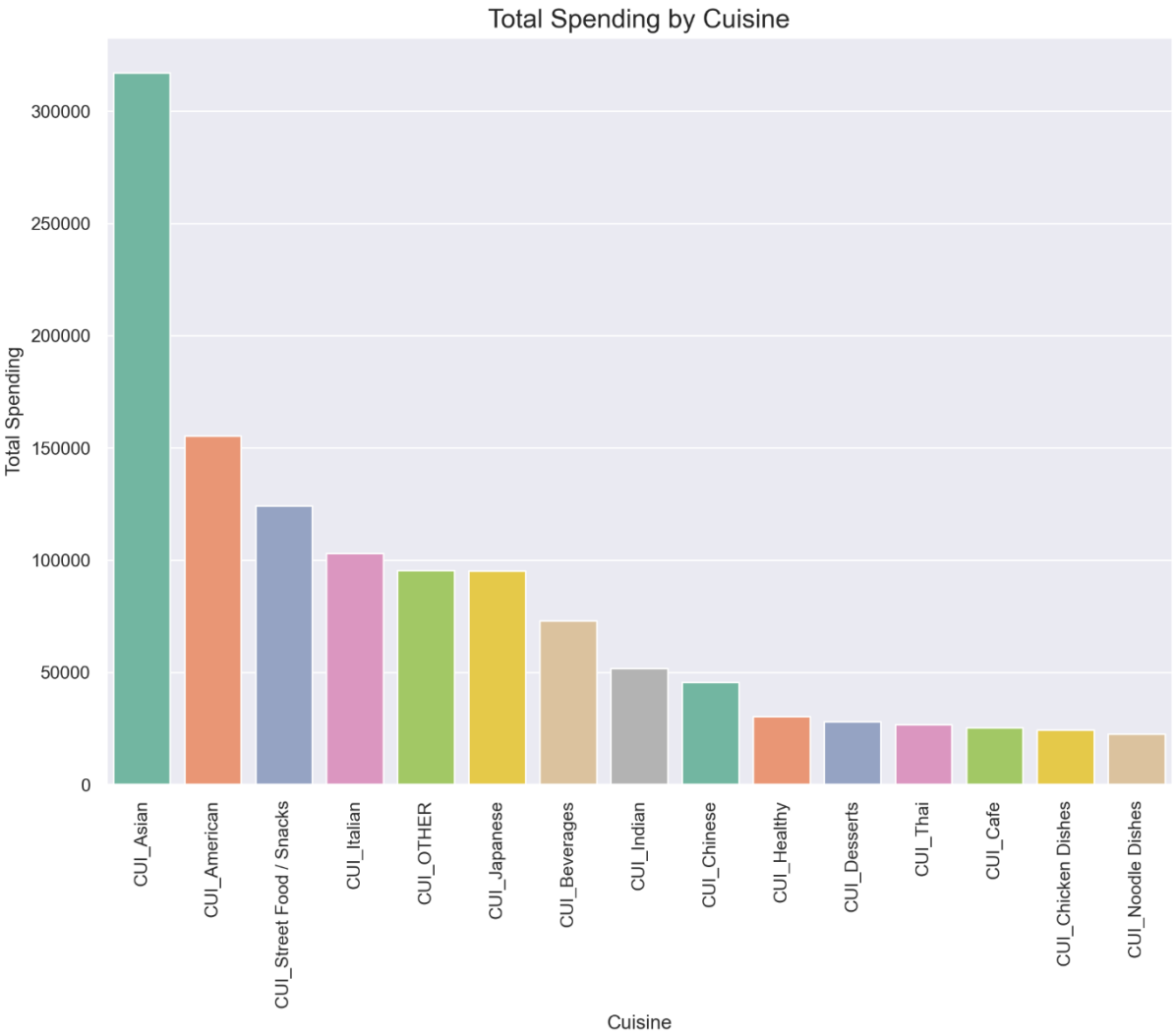

Figure 1: Total Spending by Cuisine

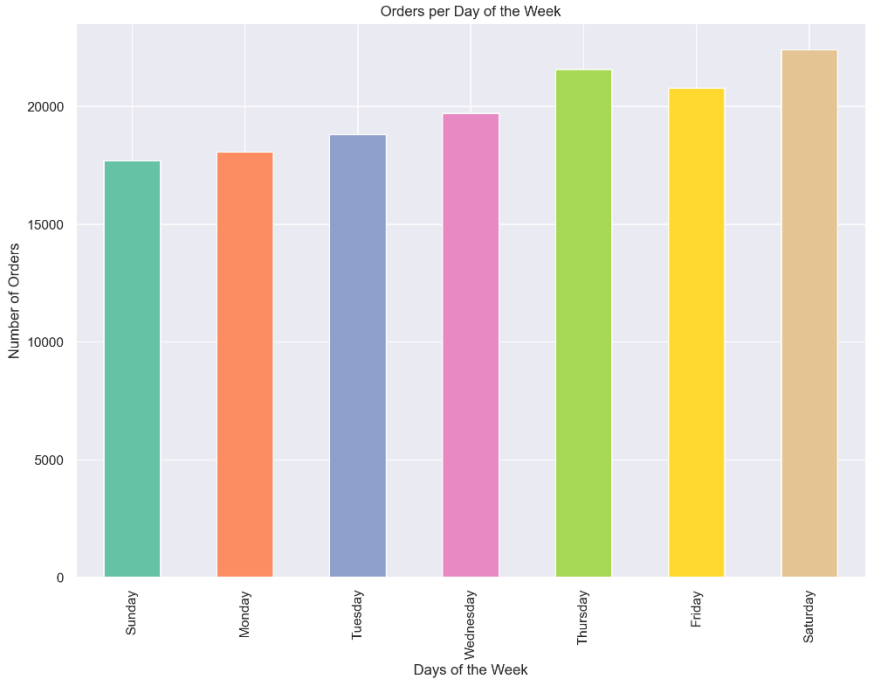

Figure 2: Orders per Day of the Week

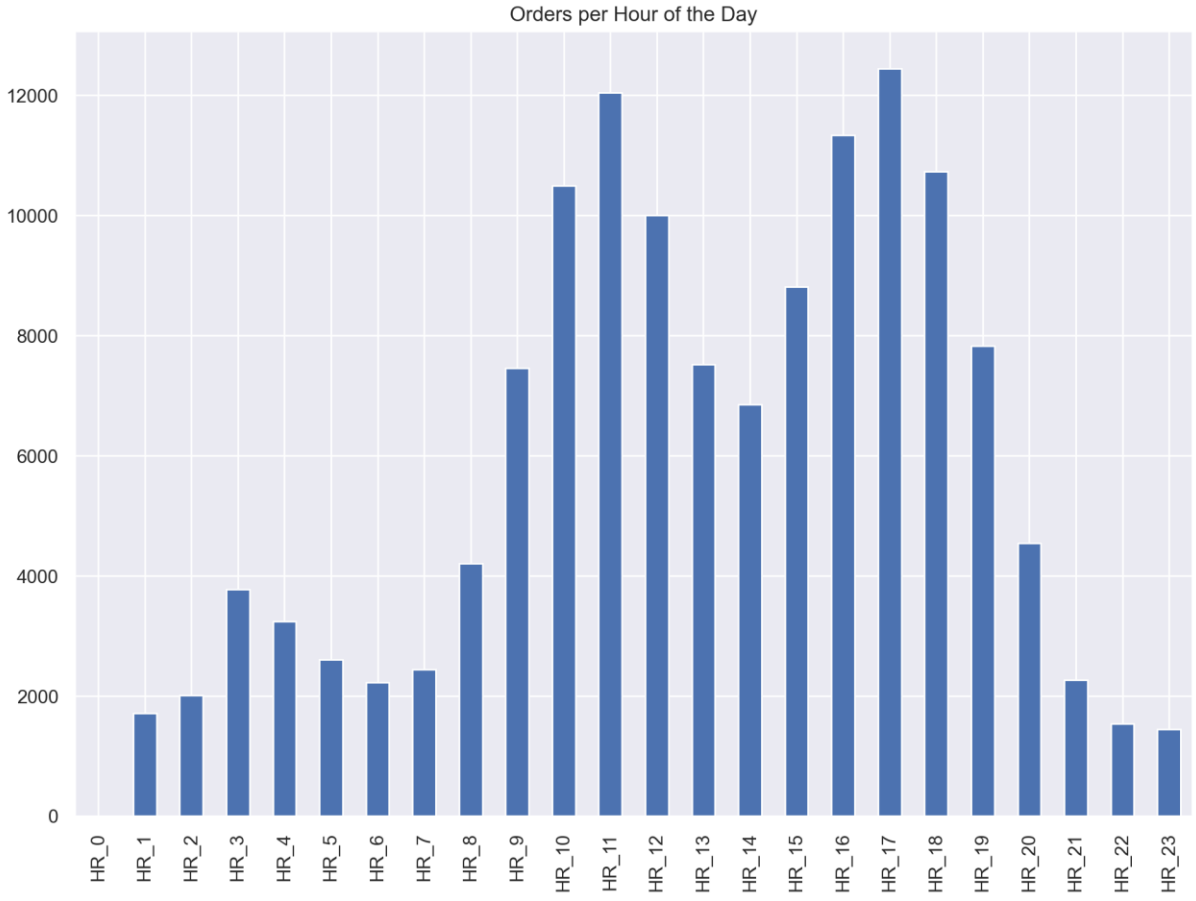

Figure 3: Orders per Hour of the Day

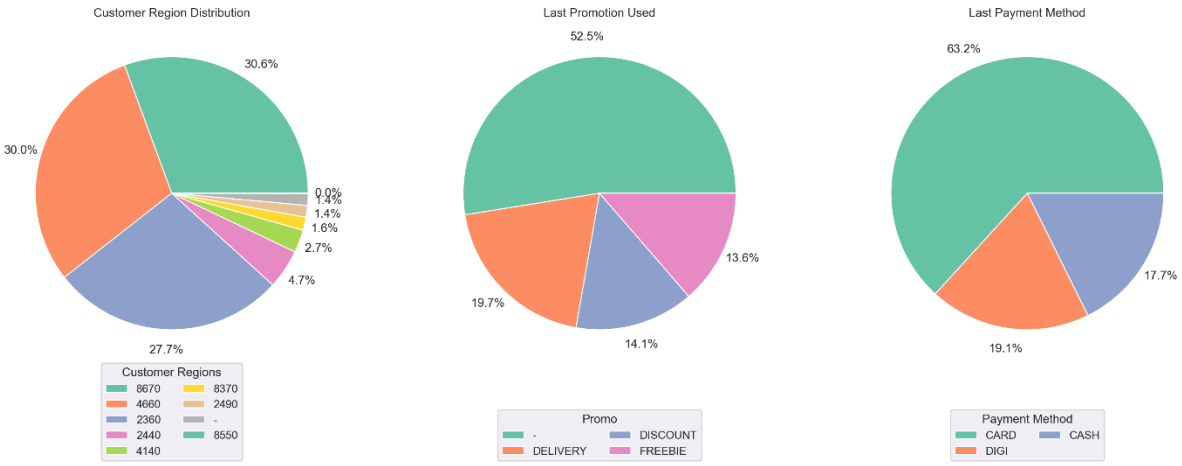

Figure 4: Distribution of categorical data

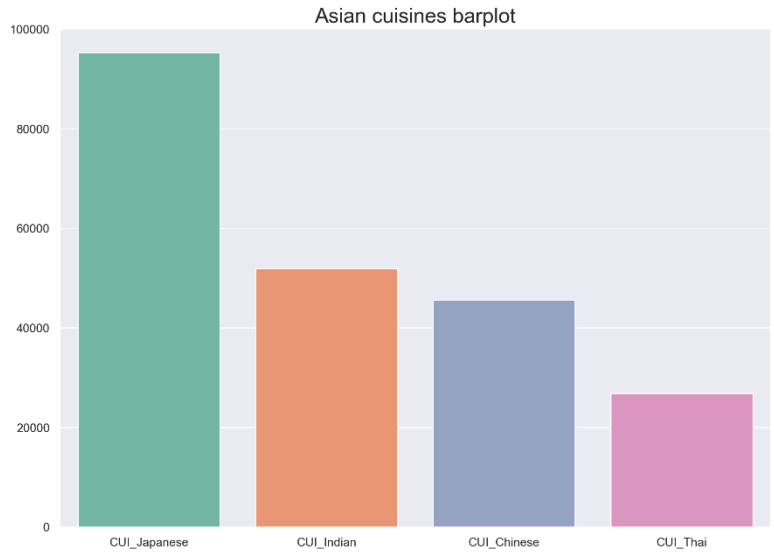

Figure 5: Monetary units spend on Asian cuisines

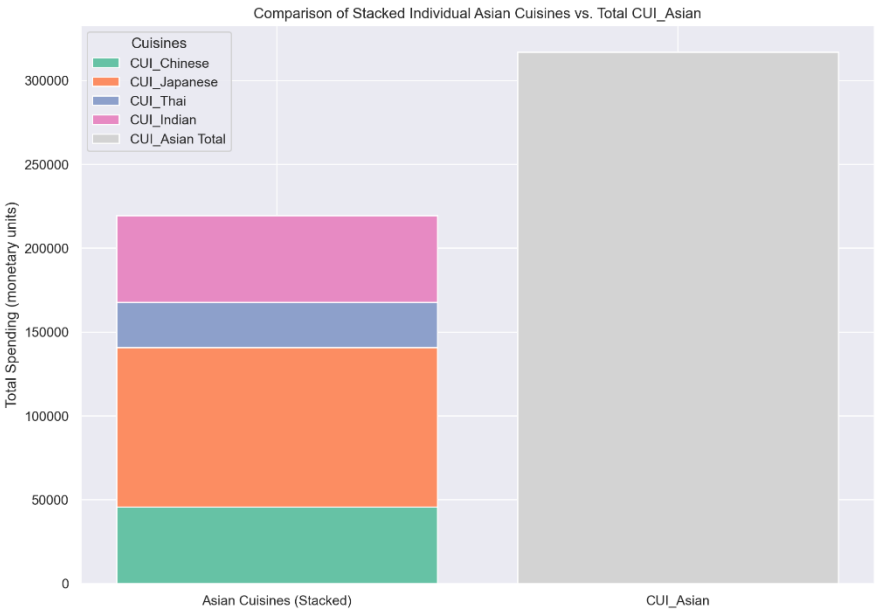

Figure 6: Comparison of individual Asian cuisines vs. total CUI_Asian

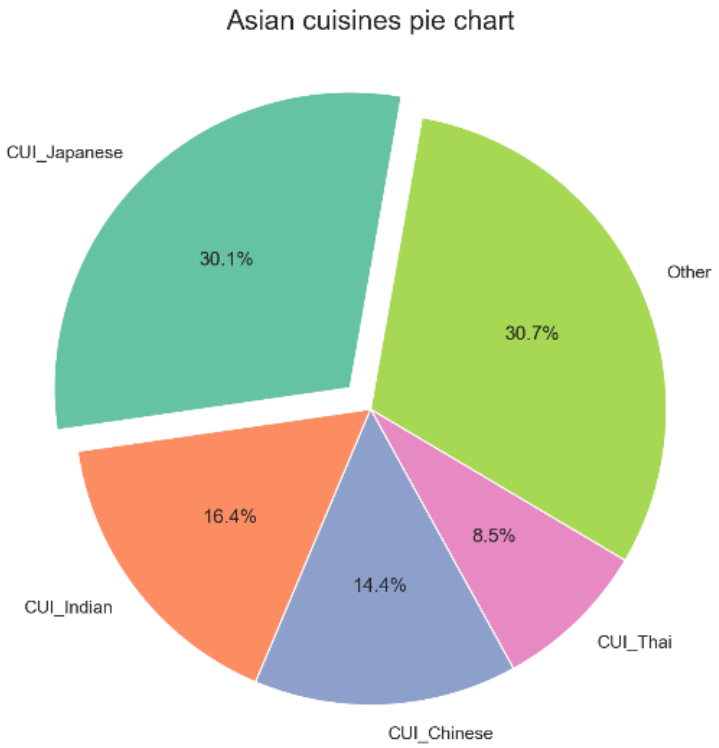

Figure 7: 4 most popular Asian cuisines in the total spending of all Asian cuisines

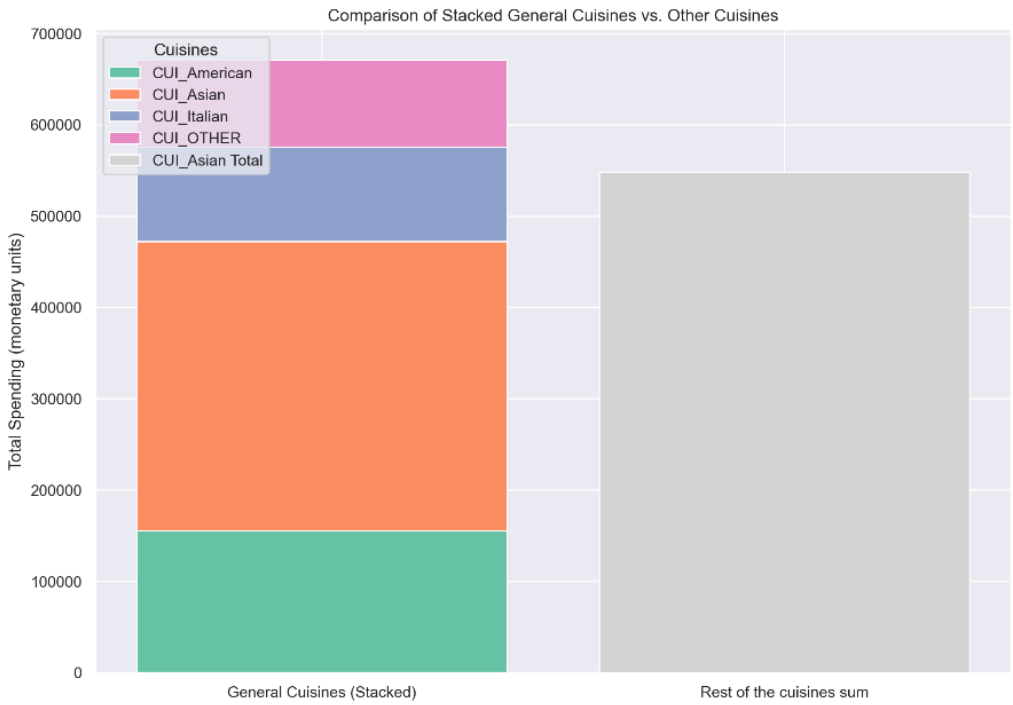

Figure 8: Comparison of Stacked General Cuisines vs. Other Cuisines

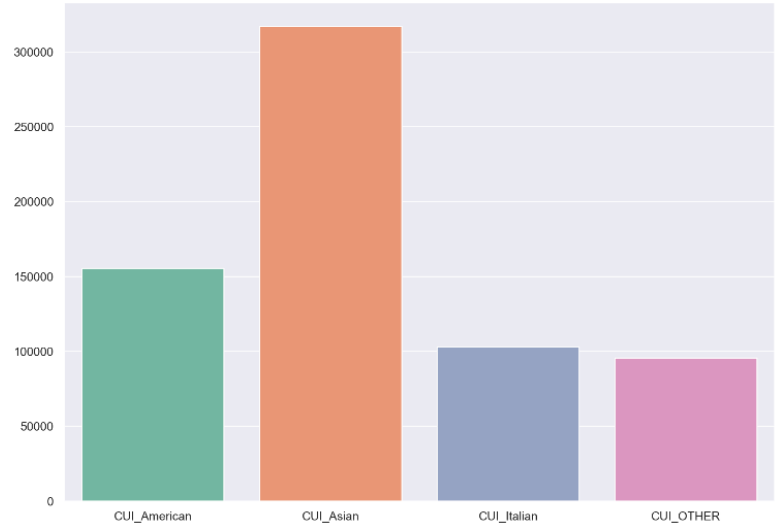

Figure 9: General Cuisines Comparison

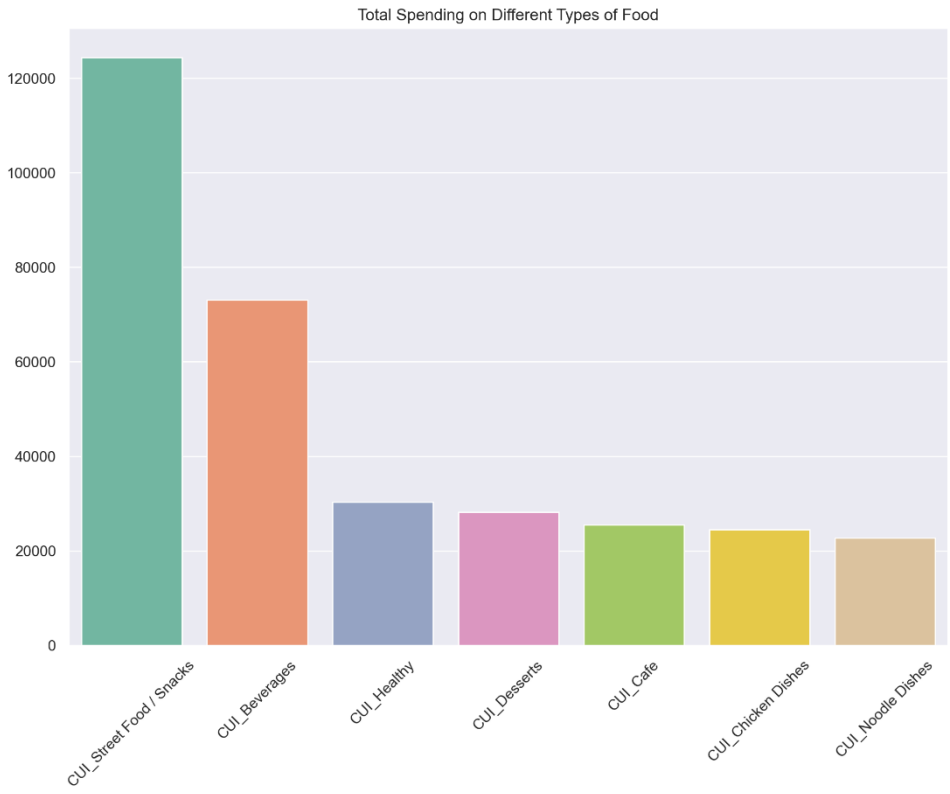

Figure 10: Total Spending on Different Types of Food

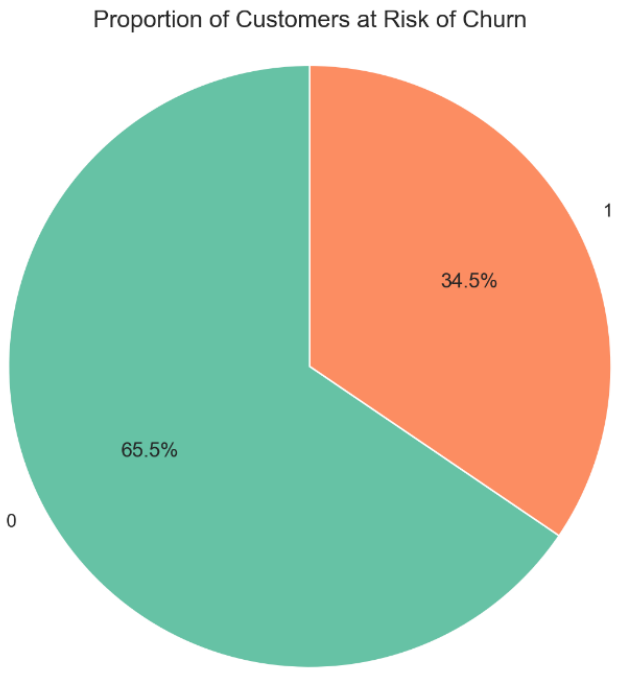

Figure 11: Percentage of Customers at Risk of Churn

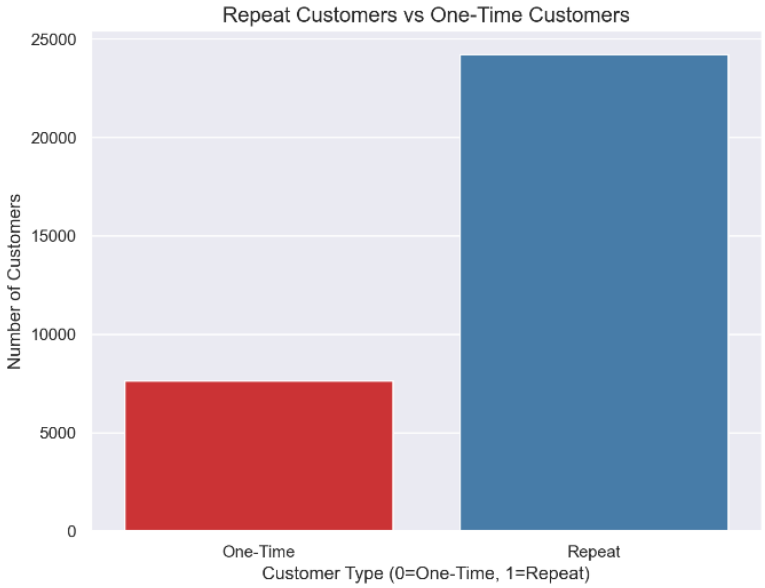

Figure 12: One-time VS Repeat Customers

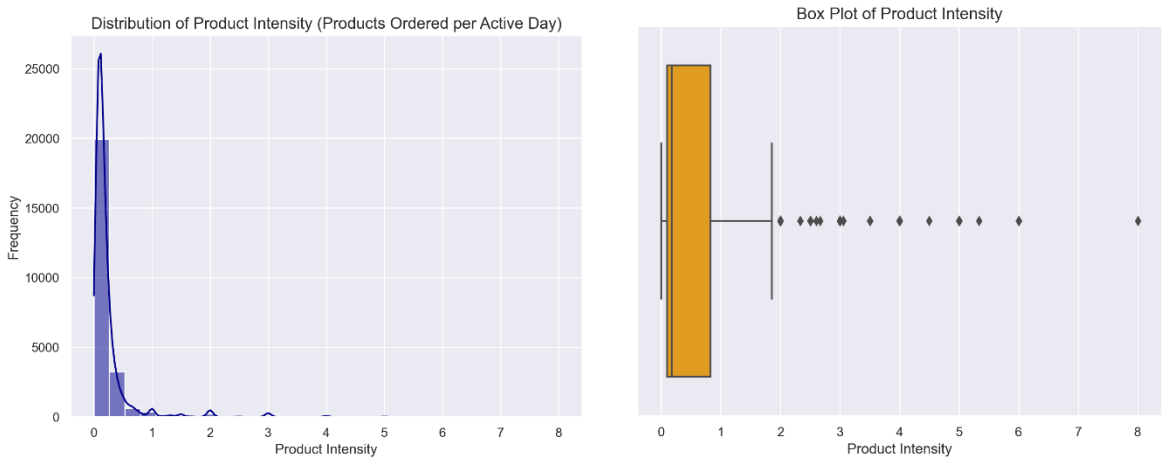

Figure 13: Order Intensity Distribution

Figure 14: Numeric Variables' Histograms

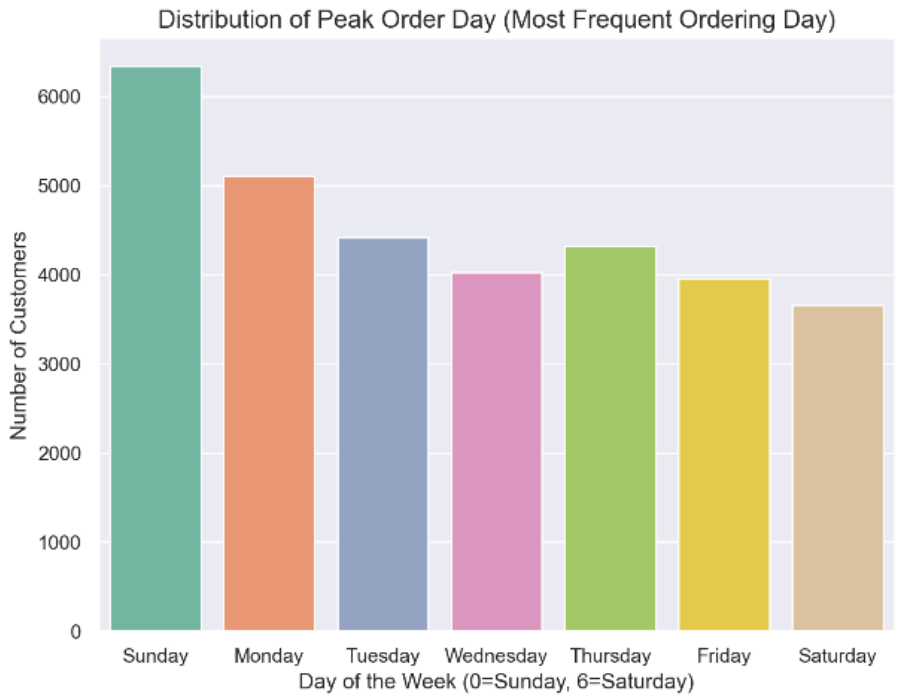

Figure 15: Distribution of Peak Order Day

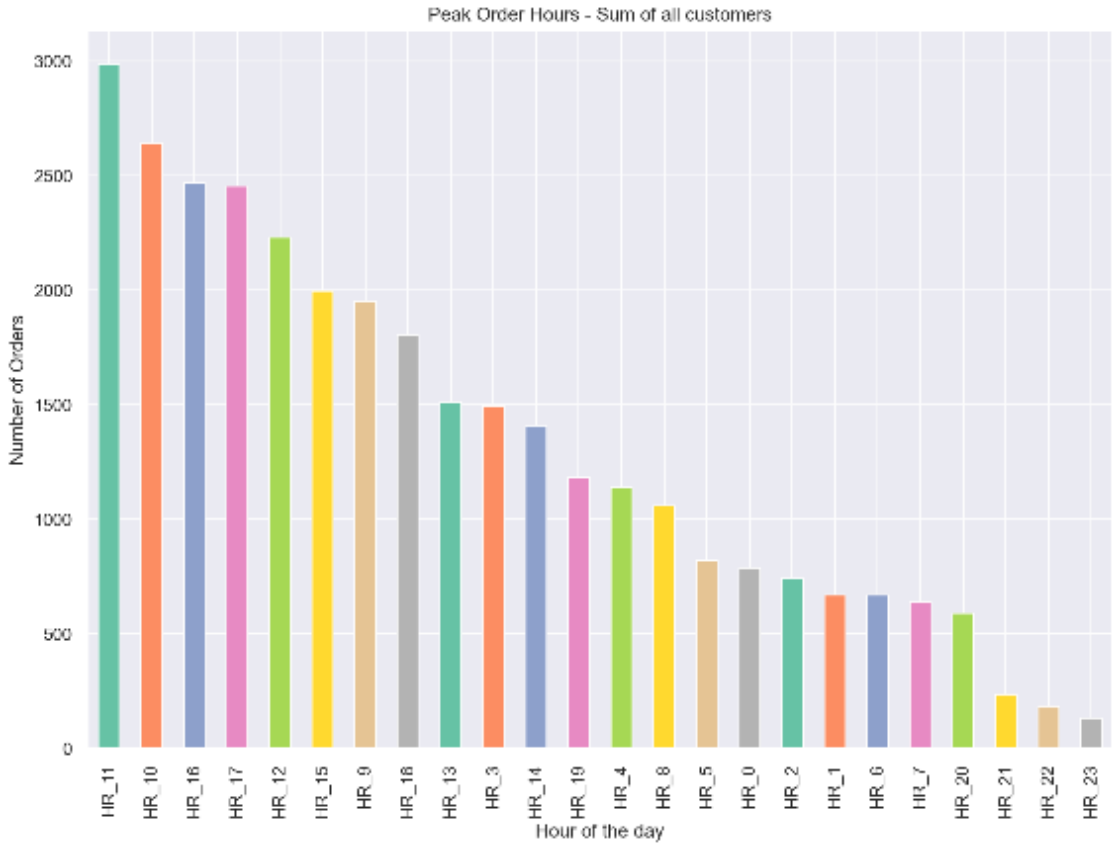

Figure 16: Peak Order Hours

Figure 17: Numeric Variables' Box Plots

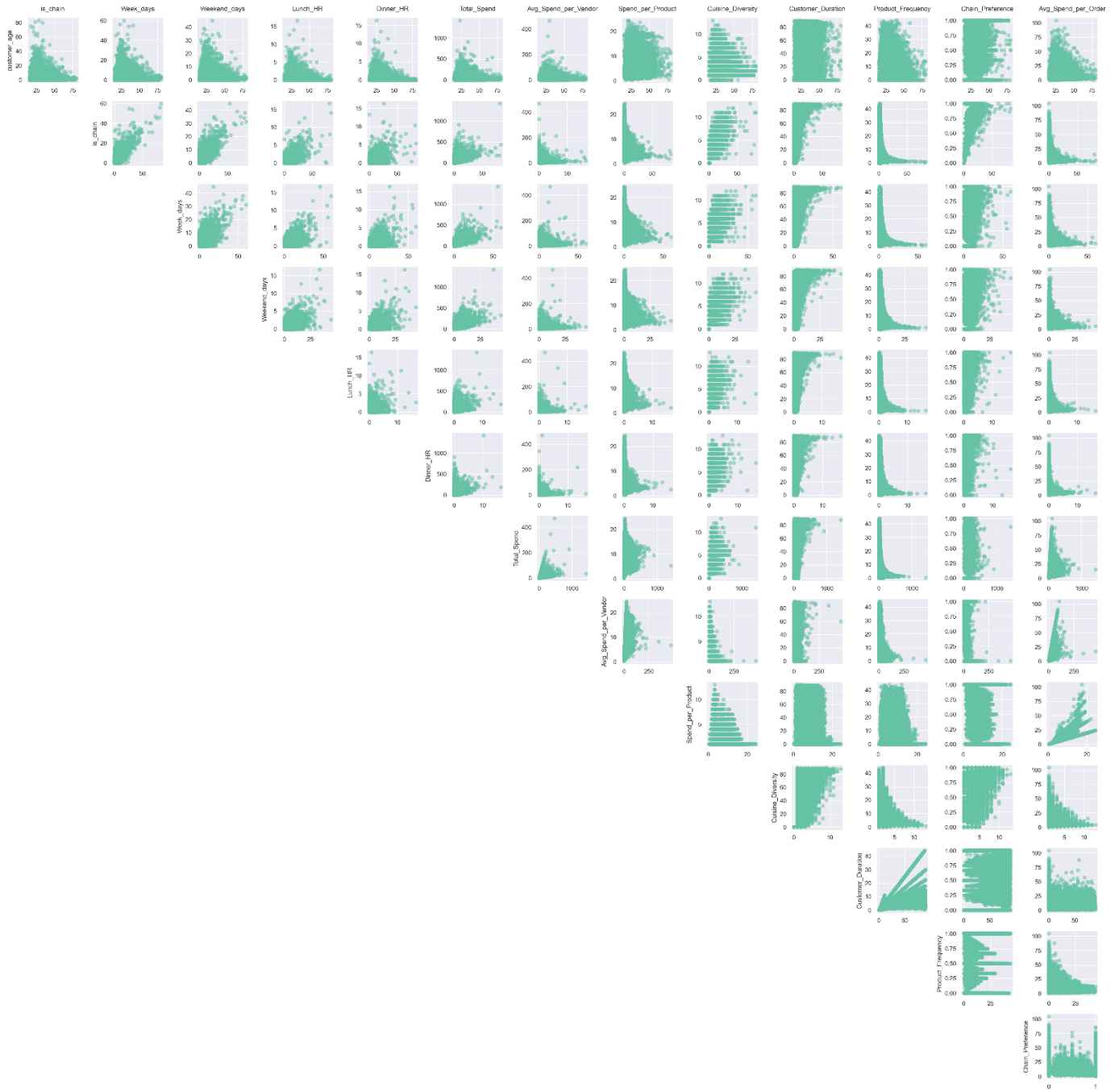

Figure 18: Scatter Plots

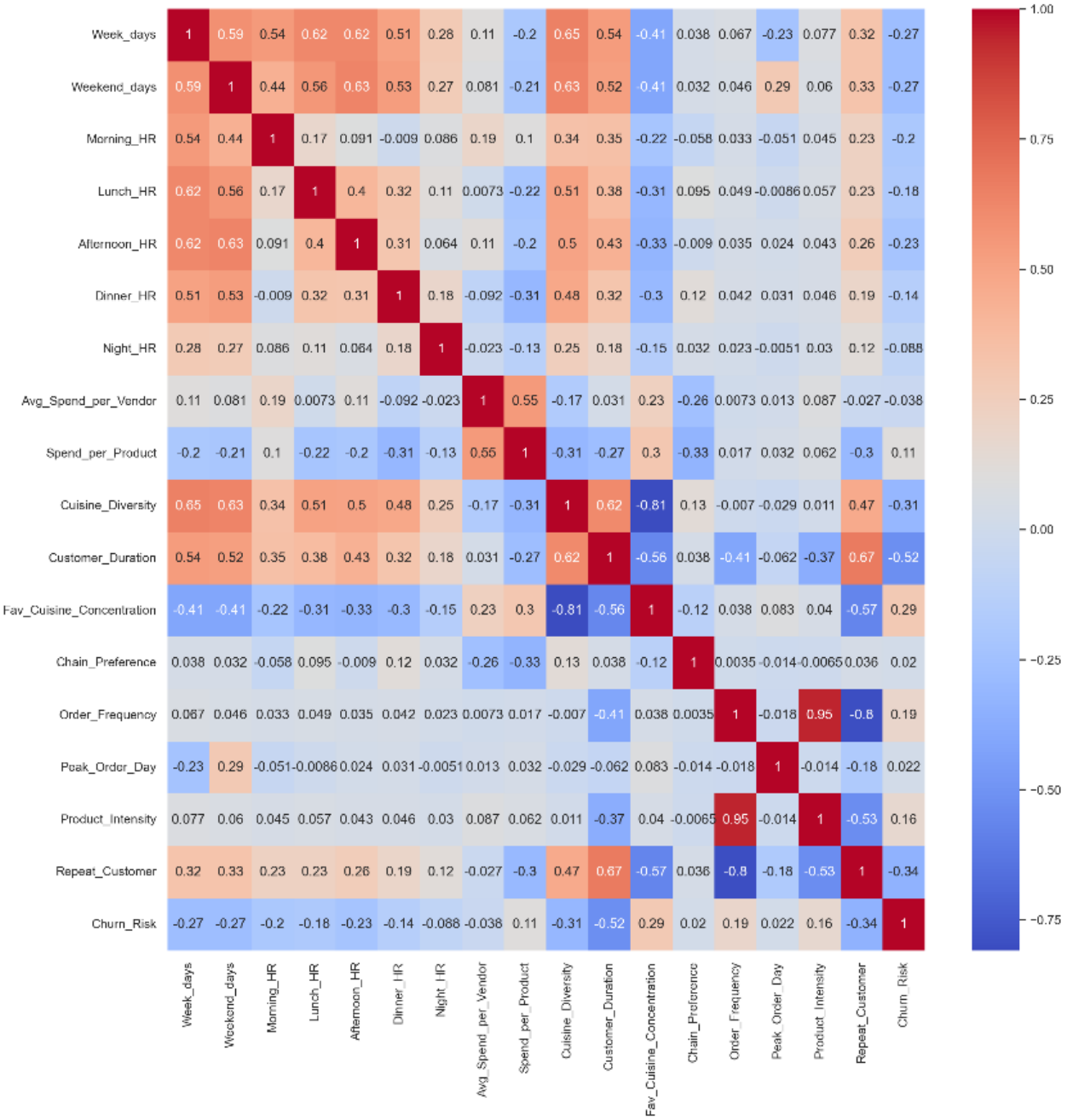

Figure 19: Heatmap

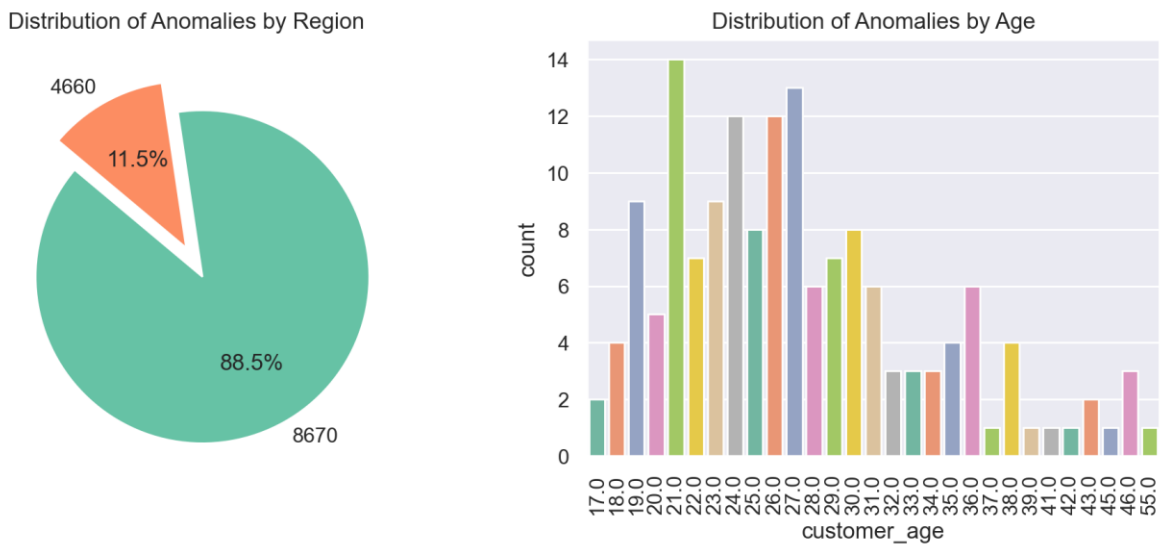

Figure 20: Analysis of Anomalies in the data

APPENDIX TABLES

Table 1: Key statistics numeric data

	customer_age	vendor_count	product_count	is_chain	first_order	last_order
count	31828.00	31828.00	31828.00	31828.00	31828.00	31828.00
mean	27.47	3.10	5.67	2.82	28.44	63.67
std	7.09	2.77	6.96	3.98	24.06	23.23
min	15.00	0.00	0.00	0.00	0.00	0.00
25%	23.00	1.00	2.00	1.00	7.00	49.00
50%	26.00	2.00	3.00	2.00	22.00	70.00
75%	31.00	4.00	7.00	3.00	45.00	83.00
max	80.00	41.00	269.00	83.00	90.00	90.00

Table 2: Key statistics Cuisines

	count	mean	std	min	25%	50%	75%	max
CUI_American	31828.0	4.88	11.65	0.0	0.0	0.0	5.66	280.21
CUI_Asian	31828.0	9.96	23.57	0.0	0.0	0.0	11.83	896.71
CUI_Beverages	31828.0	2.30	8.47	0.0	0.0	0.0	0.00	229.22
CUI_Cafe	31828.0	0.80	6.43	0.0	0.0	0.0	0.00	326.10
CUI_Chicken Dishes	31828.0	0.77	3.66	0.0	0.0	0.0	0.00	219.66
CUI_Chinese	31828.0	1.43	8.20	0.0	0.0	0.0	0.00	739.73
CUI_Desserts	31828.0	0.88	5.26	0.0	0.0	0.0	0.00	230.07
CUI_Healthy	31828.0	0.95	5.84	0.0	0.0	0.0	0.00	255.81
CUI_Indian	31828.0	1.63	7.44	0.0	0.0	0.0	0.00	309.07
CUI_Italian	31828.0	3.23	11.25	0.0	0.0	0.0	0.00	468.33
CUI_Japanese	31828.0	2.99	10.18	0.0	0.0	0.0	0.00	706.14
CUI_Noodle Dishes	31828.0	0.71	4.54	0.0	0.0	0.0	0.00	275.11
CUI_OTHER	31828.0	3.00	9.78	0.0	0.0	0.0	0.00	366.08
CUI_Street Food / Snacks	31828.0	3.91	15.52	0.0	0.0	0.0	0.00	454.45
CUI_Thai	31828.0	0.84	4.44	0.0	0.0	0.0	0.00	136.38

Table 3: List of new features created

Average Spend per Order	*= Total Spend / Total Orders* Customers spending on average per order
Average Spend per Vendor	*= Total Spend / Vendor Count* Customers spending on average per vendor
Chain Preference	*= Chain restaurant orders / total orders* Proportion of orders that came from chain restaurants
Days Since Last Order	*= max (Last Order) - Last Order* *Days that passed since last order to the end of the Dataset*
Churn Risk	*= 1 if days since last order > threshold, else 0*. Binary feature that indicates whether a customer is at risk of churn, based on how long it has been since their last order
Cuisine Concentration	*= Spend on the most ordered cuisine / Total Spend* Customers' spending on a particular type of cuisine. Higher concentration may indicate strong preferences, while lower indicates a more diverse taste
Cuisine Diversity	*= number of non-zero CUI_* columns* Distinct types of cuisine the customer has ordered from, indicating the variety of their taste preferences
Customer Duration	*= last order – first order* Number of days since the customer has been registered to its last order
Favorite Cuisine	= cuisine with highest spending
Favorite Cuisine Concentration	*= spend on the favorite cuisine / total spend* Concentration of the customer's spending on his favorite cuisine
Product Frequency	*= Customer Duration / product count*
HR groups	Dividing all hours in specific daytimes
Money spent on average	*= sum of all the cuisines per customer / total products* Total expenditure
Order Frequency	*= Total Orders / Customer Duration*
Product Intensity	*= product count / Customer Duration* Number of products ordered per active day. A higher number could indicate a high purchase intensity per day.
Other Asian Cuisines	*= CUI_Asian – sum of the money spent in Japanese, Chinese, Thai and Indian cuisines* Money spent on other types of Asian cuisines
Peak Order Day	*= DOW_X with the maximum value*

	Day of the week the customer most frequently orders on. This could identify customers who prefer to order on weekends or weekdays.
Peak Order Hour	*= HR_X with the maximum value*
Repeat Customer	*= 1 if last order – first order > 1, else 0* whether a customer has placed more than one order over a significant period, differentiating between one-time and repeat customers
Spend per Product	*= Total spend / product count* average amount spent per product, which can help identify high-value or budget-conscious customers
Total Orders	*= sum of the DOW_ columns* Total number of orders a customer has placed
Total Spend	*= Sum of all CUI_* columns* Measures the total monetary value a customer has spent over the given period. This helps in customer segmentation based on spending.)
Weekend Orders	*= Sum of DOW_ from 4 to 6*
Week Orders	*= Sum of DOW_ from 0 to 3*
Weekend Days Mean	*= Mean of DOW_ from 4 to 6*
Weekdays Mean	*= Mean of DOW_ from 0 to 3*