DATA MINING PROJECT

Master in Data Science and Advanced Analytics

NOVA Information Management School
Universidade Nova de Lisboa

Data Mining Project Report

Group 77

Ana Farinha, 20211514

António Oliveira, 20211595

Mariana Neto, 20211527

Salvador Domingues, 20240597

Fall Semester 2024-2025

TABLE OF CONTENTS

1 INTRODUCTION

Consumers today are becoming more selective about where they buy their products and where they spend their money. Consequently, it is essential for companies to better understand their clients and be able to tailor sales and discounts to certain groups of customers.

1.1 The Project

Knowing this, *ABCDEats Inc.* approached *TargetSphere Advisors* about a Customer Segmentation project, whose goal was to segment customers into distinct groups based on shared characteristics and their purchasing behaviours. By identifying these unique segments, the company can create more targeted sales strategies, offer personalized discounts and enhance customer satisfaction and loyalty.

1.2 The Team

This project was presented to *TargetSphere Advisors*, a team of Consultants, Business Analysts and Data Scientists focused on implementing Machine Learning solutions for prediction or clustering purposes. *TargetSphere Advisors* were founded in 2021, and since then have developed several projects in the Machine Learning area of expertise. *TargetSphere Advisors* started with three members, and recently expanded its team with a fourth member, as it is believed the project presented by *ABCDEats Inc.* would require more manpower than initially available.

2 INITIAL EXPLORATION

After importing the data and the necessary libraries, we conducted an initial exploration of the data using a Profile Report generated by the 'ydata_profiling' Python package. This allowed us to see that the dataset comprised 55 variables, 51 numeric and 4 categorical. It contained 31888 observations. After looking at the 'Alerts' section of the Profile Report, it highlighted a few problems within the dataset, such as the variable HR_0 being constant, the existence of missing values and duplicate rows. Additionally, most of the 'hour variables' features (the ones that start with HR_) had a high percentage of zeros.

3 EXPLORATORY DATA ANALYSIS

3.1 Duplicates

Before proceeding with an in-depth analysis, we checked for duplicate rows, having found 120 observations. Our analysis confirmed that the duplicates were linked to the same customer but had different customer IDs. To address this issue, we removed one of the duplicate entries.

3.2 Descriptive Statistics

The key insights from the data were obtained by analysing the descriptive statistics for both numerical and categorical features and were further confirmed throughout the notebook. For instance, we found out that customers' ages range from 15 to 80-year-olds (Fig. 1), with 75% being 31

or younger, indicating that most of our customers are relatively young. The customers spent most in Asian (mean = 9.96) and American (mean = 4.88) restaurants. Thursday and Saturday show slightly higher average orders. There was a noticeable spike in orders during typical lunchtime (**HR_11** to **HR_13**) and snack time (**HR_16** to **HR_18**). Activity during late-night and early-morning hours (**HR_0** to **HR_7**) is significantly lower. We, also, noticed that variable **HR_0** was univariate. In terms of categorical variables, we found that the values of **customer_region** were codes. Additionally, there was a value ('-') that was changed to '0000' to be in the same format as the other. The most frequent value for the last promotion category was the unknown category ('-'). This category was considered as 'NO DISCOUNT'.

3.3 Incoherencies

In this project's next step, we focused on ensuring our data had no inconsistencies with reality. The first variable we checked was **customer_age** which can have two problems: people who are too young to buy anything via an app (depending on the country) and people who are too old to be alive. 365 customers had not reached the age of 18 (the age of majority in most countries). Afterwards, we verified the age of our oldest customer, who is 80, an age that falls within the reality limits.

The next feature that went through incoherency checking was **vendor_count**. It had an interesting minimum of 0 unique vendors for 138 customers. Those rows had no orders placed at any hour and no money spent on any cuisine in the three months our data is referring to. However, they had values for the columns of **is_chain**, **first_order** and **last_order**.

Following this, we turned our attention to the variable **product_count**. To those who did not buy any product (**product_count = 0**), there was some amount spent on specific cuisines and orders in specific hours in the last three months, all of them with a unique vendor. After, we made a comparison between **product_count** and the sum of the orders placed by adding the **HR_** features. Here, we identified 18 cases (the same clients identified above) where the customers did not even order a product for each delivery made. This might have happened due to delivery mistakes, freebies sent, cancelled orders or other errors. Additionally, using the sum of orders placed by adding the **HR_** features we could also conclude that 97 clients bought products, but there is no information on the hour the order was placed.

The last variable that was subject to incoherency checking was **is_chain**. This variable, which was binary according to the metadata, had values between 0 and 83 which makes it a discrete feature (Fig. 2). By comparing it with the unique number of vendors, we could also conclude that it does not indicate the number of unique chains the customer has ordered from. Instead, it refers to the frequency of orders made to chain restaurants. We found out that there were 422 customers that had a total number of orders from chain restaurants that were higher than the total number of orders made.

In this section of the project, we found out that whenever the first order day was missing, the last order was always 0. We assume that the customer's last order was placed on the same day the dataset started. Therefore, these customers only made an order once and never came back. Given this insight, we considered replacing the missing values in **first_order** to 0.

3.4 Missing Data Analysis

Regarding missing values, we identified three variables with missing values in our dataset: **customer_age**, **first_order**, and **HR_0**, each with less than 5% missing data (Fig. 3). To understand the nature of this missingness, we used the '*missingno'* library to analyse the nullity correlation, shown in Figure 4. In this context, a nullity correlation close to 1 indicates a strong positive relationship, meaning if one column has a missing value, the other likely does too. On the other hand, a correlation value near -1 suggests a strong inverse relationship, where one column has missing values while the other does not. From the output of Figure 4, we were able to identify that the correlation between the 3 variables was nearly or exactly 0, indicating that the missing values in one column had no significant relationship with those in another, meaning the occurrence of missing data was likely independent across these variables - Missing Completely at Random (MCAR).

Afterwards, we analysed pairs of columns with missing values and we concluded that there weren't customers with missing values on the three variables at the same time. However, there were 2 customers with missing values in **customer_age** and **first_order**, 1 customer with missing values in **first_order** and **HR_0** and 27 customers with missing values in customer_age and **HR_0** (Fig. 5). Considering this, there was no significant relationship or pattern between the missingness of the variables analysed, as already verified in Fig. 4. The highest rate of missingness occurred in the combination of **customer_age** and **HR_0**. Given that HR_0 is a univariate variable, its relevance may be minimal in the context of our analysis, and it can potentially be discarded. For **customer_age** and **HR_0**, we took the median to fill the missing values, while, for **first_order** we decided to impute the missing values with 0, given the insights we took during the previous section.

3.5 Outliers

First, we use the `IQR_outliers` function, which identifies the potential outliers in the columns by applying the interquartile range method. With this, we detected variables with a huge number of outliers like **HR_17**, **HR_11**, **HR_16** and **CUI_OTHER** with more than 7000 outliers and **first_order**, **last_order** and **HR_0** the lowest with 0 outliers. Then, to better visualize them, we use the `plot_multiple_boxes_with_outliers1` function to generate a matrix of boxplots. The output can be seen below on the Appendix, from figure 6 to 9. For now, we decided not to remove them, and address them in the next Project Deliverable.

3.6 Data Wrangling

This section consists of transforming existing variables, when required, as well as creating new variables from the existing ones, that we believe would add valuable information and could be useful later in the project.

3.6.1 Feature Transformation

In order to maintain data types consistency, we started by converting the features **first_order** and **HR_0** to integers. By doing this we could align with the data types in **last_order** and the hours features, respectively.

3.6.2 Feature Engineering

Upon examining unique values in categorical features, we identified some variables suitable for conversion into dummy variables. Taking as an example **last_promo**, which had values 'DELIVERY', 'DISCOUNT', 'NO DISCOUNT' and 'FREEBIE', we decided to create a column for each of these options. 1 if the value was present in the column **last_promo** and 0 otherwise. Since customers can have only one promotion at a time, we can drop the **promo_NO_DISCOUNT** column to avoid redundancy—its value is implied when **promo_DELIVERY**, **promo_DISCOUNT**, and **promo_FREEBIE** are all 0. This process was repeated for feature **payment_method**. We decided to take this approach, called one-hot encoding, since we are dealing with a Customer Segmentation problem.

Replacing each categorical variable with a number could lead to algorithms interpreting these numeric values as ordinal or continuous, which would be incorrect as categories like promotion types do not have a natural order or magnitude. Having created dummy columns for these two features, we also thought it could be interesting to transform them using `ordinal encoder` and `frequency encoder` respectively. The best approach for each feature will be decided later in the Project
More variables were created based on the existing ones. The features and respective descriptions can be found in Table 1.

3.7 Data Reduction

Having 66 different features to work with, we thought that this could be an excessive number of features which could lead to problems later in the project. Consequently, we added another step to this section, which consists of trying to reduce the dimensionality of the dataset by grouping some variables that would still make sense together.

The first change was combining the **DOW_** variables by weekday and weekend orders. The same idea was applied to the **HR_** variables, as we had 24 columns, one for each hour. By visually analysing the total number of orders per hour (Fig. 10), we observed peaks at 3 AM, 11 AM and 5 PM, which align with the most frequent times for each meal. In a Portuguese context, these times would not make sense. However, we must remind ourselves that this plot may be influenced by the data being collected in a different time zone, or in countries with 'meal schedules' a little different than what it is like in Portugal. Having said this, we decided to group these variables into 4 groups: 1-7h, 8-14h, 15-19h, 20-23h. Our choice was based on the distribution of the Total Number of Orders per Hour, as referred previously.

4 KEY INSIGHTS

While checking the features **first_order** and **last_order**, we found out that 2 customers first placed an order on the very first day of the dataset (day 0) and last placed an order on the last day of the dataset (day 90). These customers were active throughout the entire 90-day period, starting their activity right away and making their most recent purchase on the final day. One customer made their first order on the last day of the dataset. 7179 customers made only one purchase during the dataset period (first order day is the same as their last order day). This means these customers only placed a single order and did not return to make any additional purchases afterward.

From a customer region analysis, we identified that most of the customers in the dataset came from regions '8670', '4660' and '2360' (Fig. 11), while other regions showed fewer customers, with

only 13 from region '8550'. Across all regions, the customer base is relatively young (Fig. 12), with a median age of 25/26 years. In regions '0000', '4140', '2490', and '8550', customers that have ordered products are all under 60, while in others some customers even reach the age of 80.

In terms of average spending (Fig.13), regions '8670' and '4660' exhibit broader spending ranges, with some customers even spending on average more than 20 monetary units per product and others not spending at all. Regions '2360', '2440', and '2490' are more concentrated around lower spending levels. In regions '0000', '8370', and '8550', spending ranges between 5 and 20 units on average, though there are customers who spent less or more than the majority of customers from their region.

Cuisine spending (Fig. 14) indicates Asia as the most spent in regions '8670', '0000', '8370', and '8550', followed by Street Food/Snacks, while Italian and OTHER seem to be the ones that the customers never spend. Regions '4140' and '4660' spend more on Italian cuisine.

Spending is highest among younger customers, particularly for ages between 20 and 30 (Fig. 15). A standout customer is a 23-year-old who spent a total of 1418.33 monetary units. The same customer has ordered from 40 different vendors (Fig. 16) and it is the one that has one of the largest differences between the first order and last order (Fig. 17), meaning that throughout 3 months he was the one that spent the most.

Most customers, however, order from 0 to 15 vendors, with spending concentrated below 500 units (Fig. 16). Orders within shorter intervals (0-60 days) show lower and more consistent spending, with most customers spending less than 500 (Fig. 17).

In terms of cuisines, in general customers spend most in Asia (Fig. 18), followed by American and OTHER. Spending decreases slightly on Fridays and peaks on Saturdays. Desserts and noodle dishes are the cuisines that customers spend less.

During the 3 months of data, there were 7285 new customers recorded (Fig. 19). And as it can be seen by Figure 20, these customers spent less. Most of the repeat customers spent less than 500 monetary units.

Orders are more frequent on weekdays, peaking on Thursdays and Saturdays (Figs. 21-22). These orders were mostly done between 8h to 14h and between 15h to 19h (Fig. 23). The heatmap (Fig. 24) shows order peaks at 11h and 17h daily, with minimal orders between 21h and 7h and none at midnight.

Looking at the pairwise correlation (Fig. 25), if we assume a threshold for feature relevance of 0.2 and redundancy of 0.8, there are various highly correlated pairs. The feature **product_count** is highly correlated with **vendor_count** (0.83), **weekday_orders** (0.93), **weekend_orders** (0.80), **total_spend** (0.82), **total_orders** (0.97) and **is_chain** (0.83). This suggests that customers with high product variety tend to be frequent purchasers who spend and order more. Then **vendor_count** is with **weekday_orders** (0.81) and **total_orders** (0.84). Customers with a higher **vendor count** tend to place more orders on weekdays. Also, **is_chain** with **weekday_orders** (0.83) and **total_orders** (0.87). This high correlation suggests that customers ordering from chains have higher order frequency and tend to order more on weekdays. The feature **total_orders** is highly correlated with **8-14h** (0.81), **weekday_orders** (0.96) and **weekend_orders** (0.82). This suggests that customers with high order volume tend to place orders throughout the week (more frequent on weekdays), with peak activity during daytime hours (8-14h). Overall, there is no highly negative correlation. High feature redundancy suggests that these features may be capturing similar information, which could lead to multicollinearity issues.

5 APPENDIX A – VISUALIZATIONS

Some visualization codes were partially done with the help of ChatGPT.

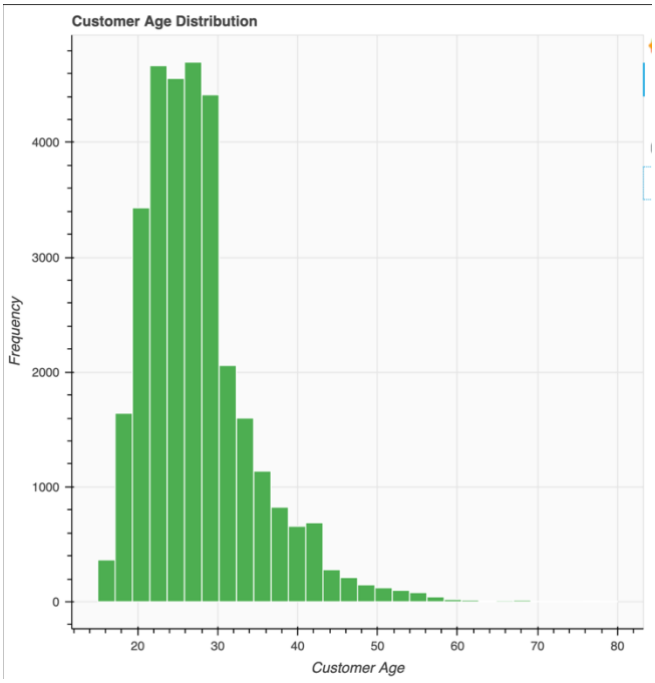


Figure 1

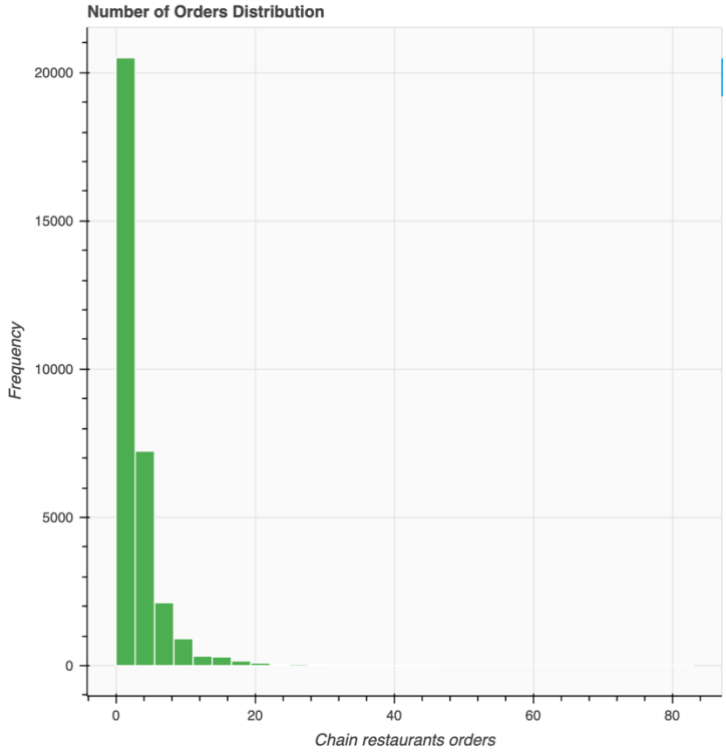


Figure 2


0.0
0.2
0.4
0.6
0.8
1.0
HR_23 31828
HR_22 31828
DOW_1 31828
DOW_0 31828
CUI_Thai 31828
CUI_Street Food / Snacks 31828
CUI_OTHER 31828
CUI_Noodle Dishes 31828
CUI_Japanese 31828
CUI_Italian 31828
CUI_Indian 31828
CUI_Healthy 31828
CUI_Desserts 31828
CUI_Chinese 31828
CUI_Chicken Dishes 31828
CUI_Cafe 31828
CUI_Beverages 31828
CUI_Asian 31828
CUI_American 31828
payment_method 31828
last_promo 31828
last_order 31828
is_chain 31828
product_count 31828
vendor_count 31828
DOW_2 31828
DOW_3 31828
DOW_4 31828
HR_10 31828
HR_21 31828
HR_20 31828
HR_19 31828
HR_18 31828
HR_17 31828
HR_16 31828
HR_15 31828
HR_14 31828
HR_13 31828
HR_12 31828
DOW_5 31828
HR_11 31828
HR_9 31828
HR_8 31828
HR_7 31828
HR_6 31828
HR_5 31828
HR_4 31828
HR_3 31828
HR_2 31828
HR_1 31828
DOW_6 31828
customer_region 31828
first_order 31722
customer_age 31101
HR_0 30664
0
6365
12731
19096
25462
31828

Figure 3

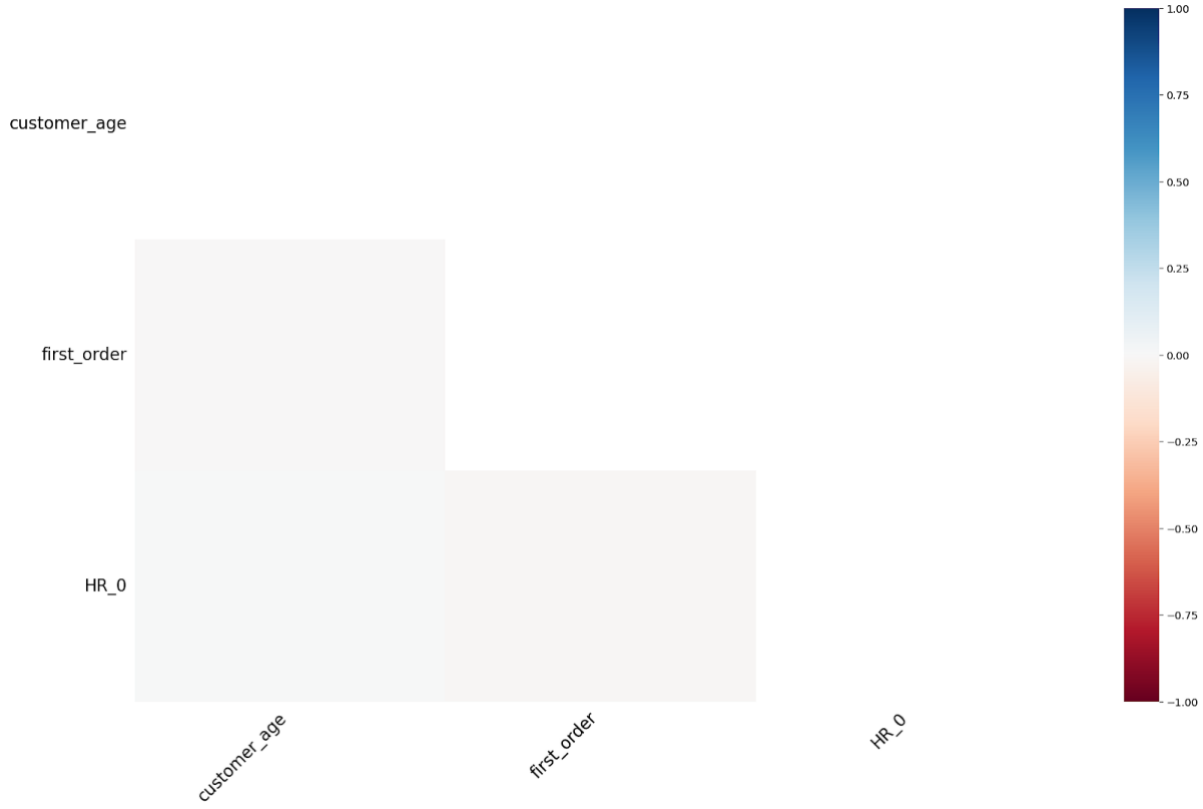


Figure 4

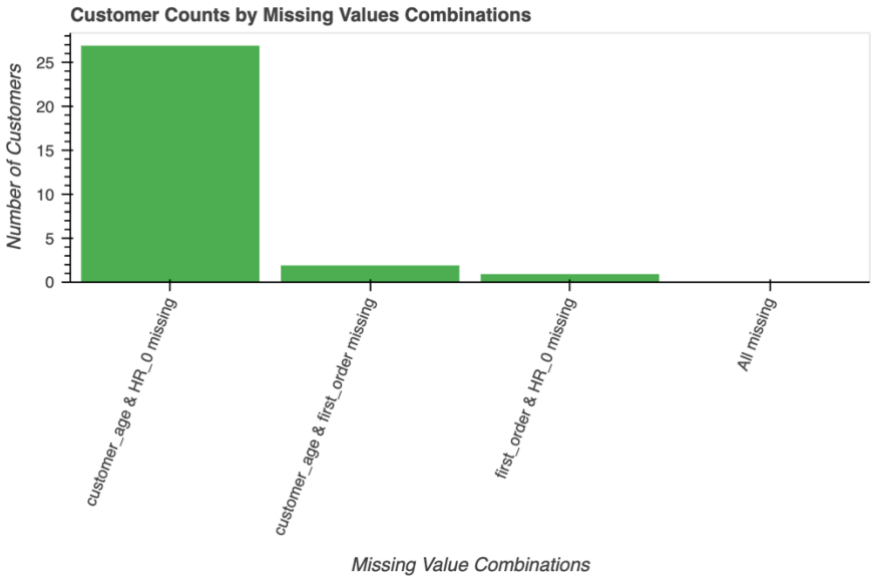


Figure 5

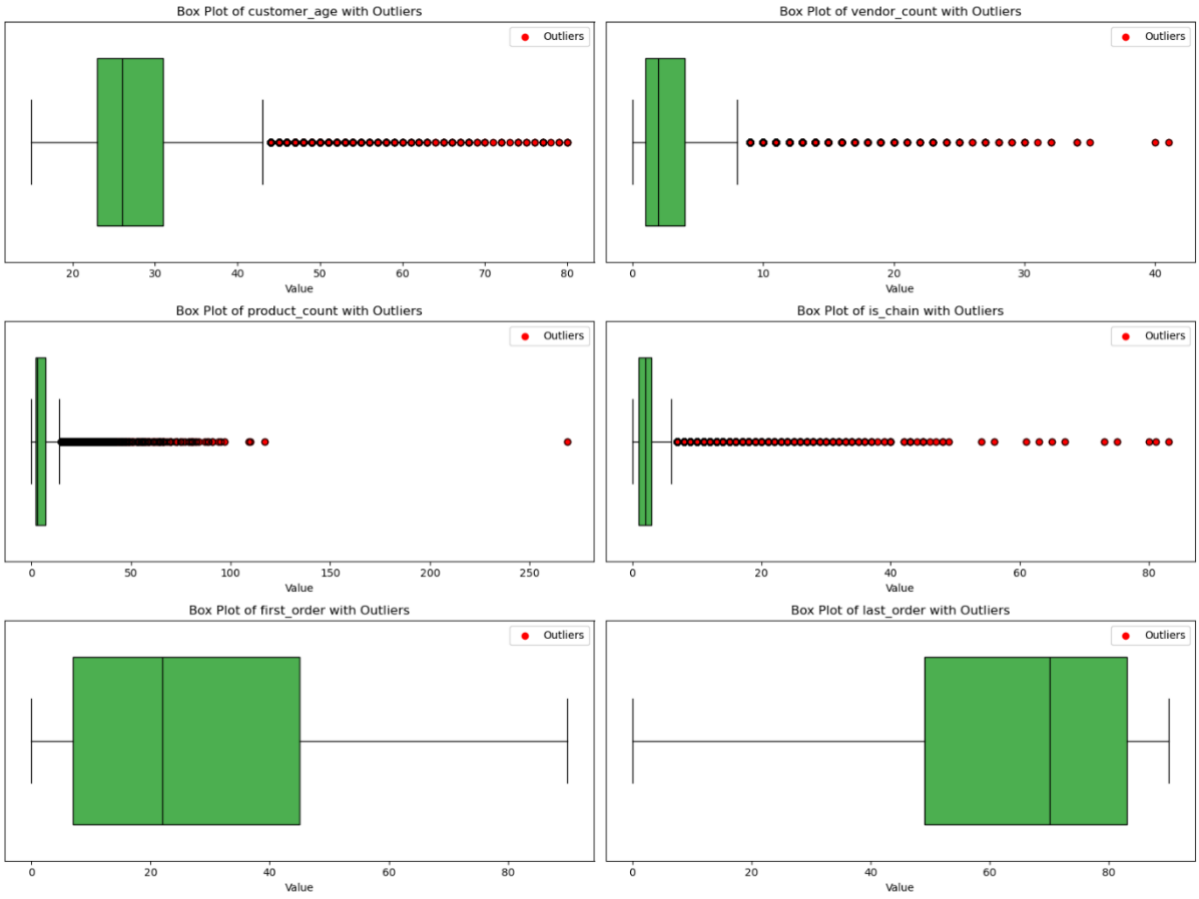
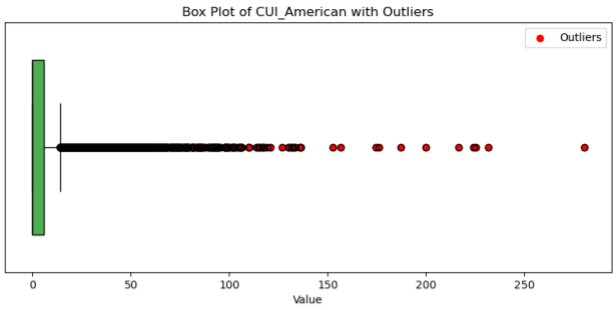
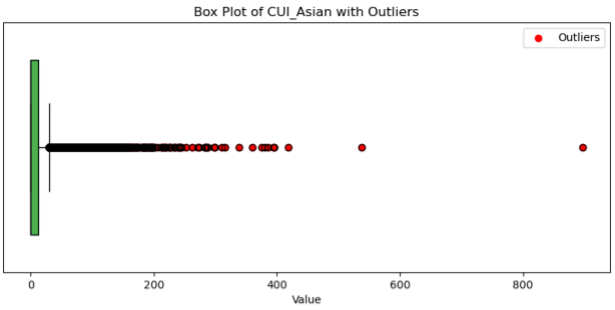
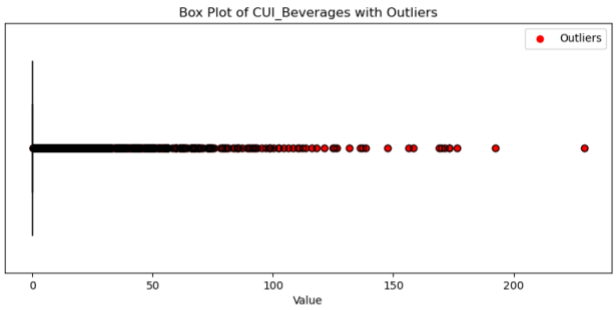
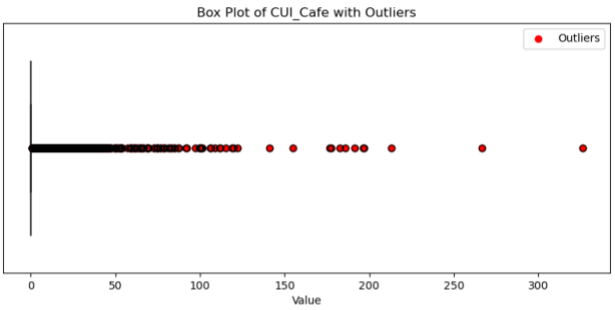

Box Plot of customer_age with Outliers
Outliers
20
30
40
50
60
70
80
Value
Box Plot of vendor_count with Outliers
Outliers
0
10
20
30
40
Value
Box Plot of product_count with Outliers
Outliers
0
50
100
150
200
250
Value
Box Plot of is_chain with Outliers
Outliers
0
20
40
60
80
Value
Box Plot of first_order with Outliers
Outliers
0
20
40
60
80
Value
Box Plot of last_order with Outliers
Outliers
0
20
40
60
80
Value

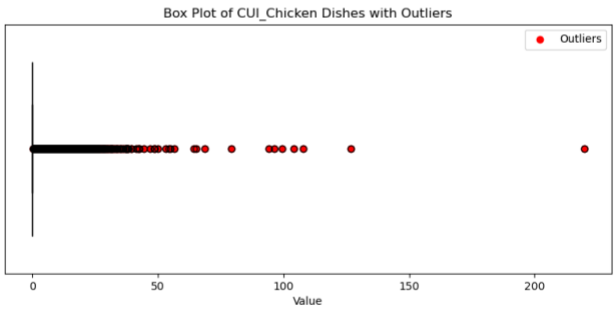
Figure 6

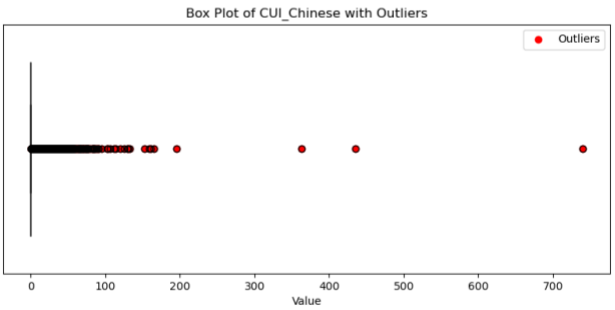

Box Plot of CUI_American with Outliers
Outliers
0
50
100
150
200
250
Value

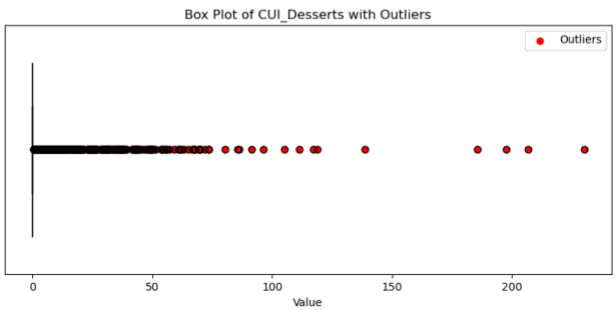

Box Plot of CUI_Asian with Outliers
Outliers
0
200
400
600
800
Value

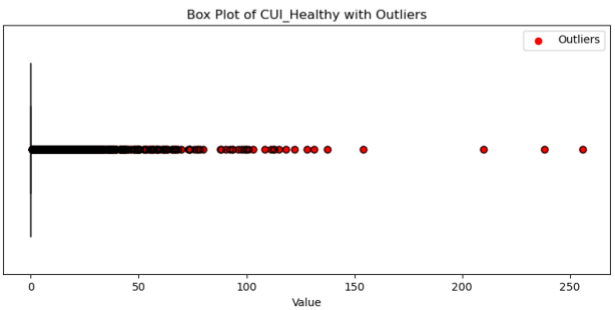

Box Plot of CUI_Beverages with Outliers
Outliers
0
50
100
150
200
Value


Box Plot of CUI_Cafe with Outliers
Outliers
0
50
100
150
200
250
300
Value


Box Plot of CUI_Chicken Dishes with Outliers
Outliers
0
50
100
150
200
Value


Box Plot of CUI_Chinese with Outliers
Outliers
0
100
200
300
400
500
600
700
Value


Box Plot of CUI_Desserts with Outliers
Outliers
0
50
100
150
200
Value


Box Plot of CUI_Healthy with Outliers
Outliers
0
50
100
150
200
250
Value

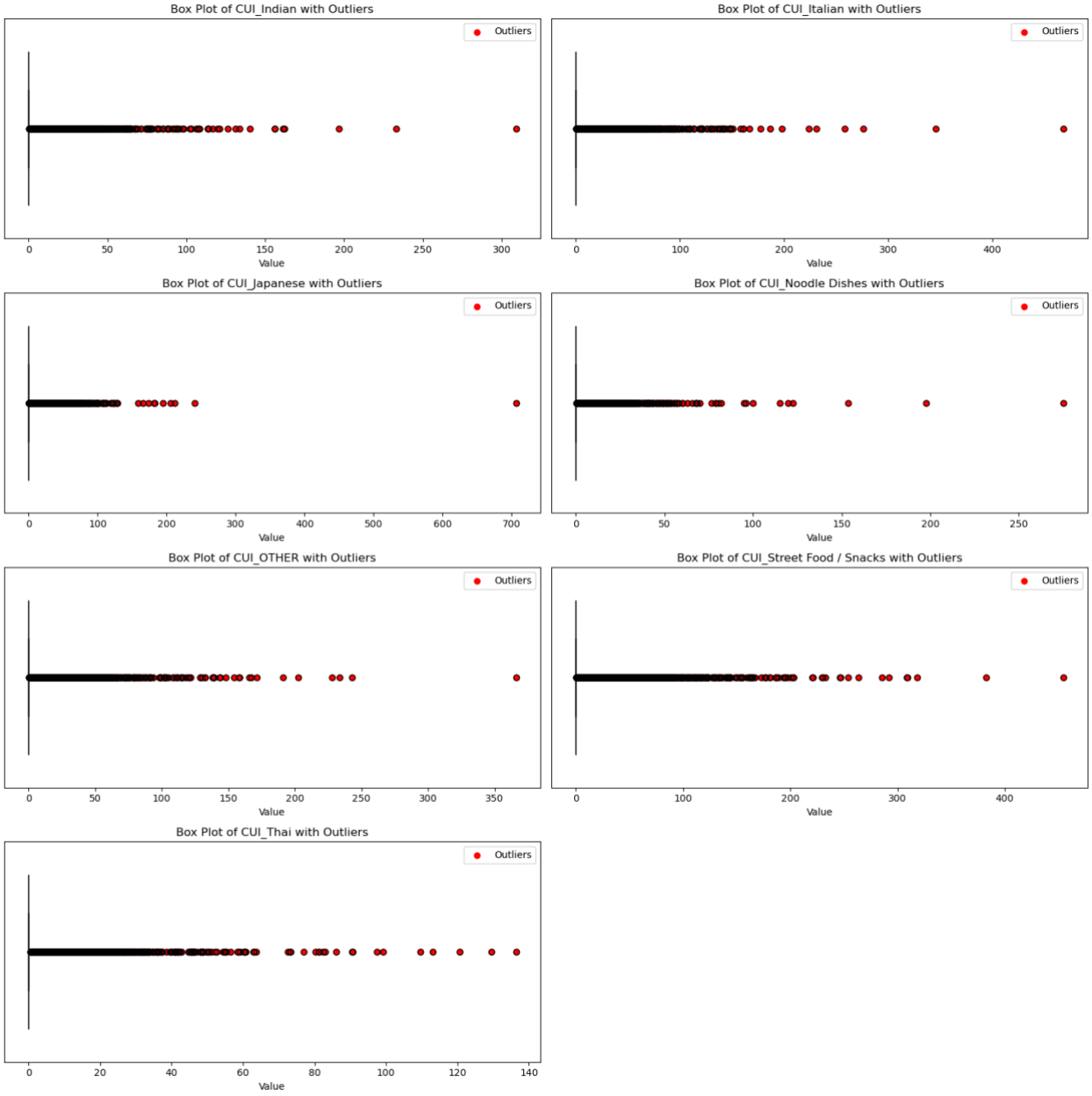

Box Plot of CUI_Indian with Outliers
Outliers
0 50 100 150 200 250 300
Value
Box Plot of CUI_Italian with Outliers
Outliers
0 100 200 300 400
Value
Box Plot of CUI_Japanese with Outliers
Outliers
0 100 200 300 400 500 600 700
Value
Box Plot of CUI_Noodle Dishes with Outliers
Outliers
0 50 100 150 200 250
Value
Box Plot of CUI_OTHER with Outliers
Outliers
0 50 100 150 200 250 300 350
Value
Box Plot of CUI_Street Food / Snacks with Outliers
Outliers
0 100 200 300 400
Value
Box Plot of CUI_Thai with Outliers
Outliers
0 20 40 60 80 100 120 140
Value

Figure 7

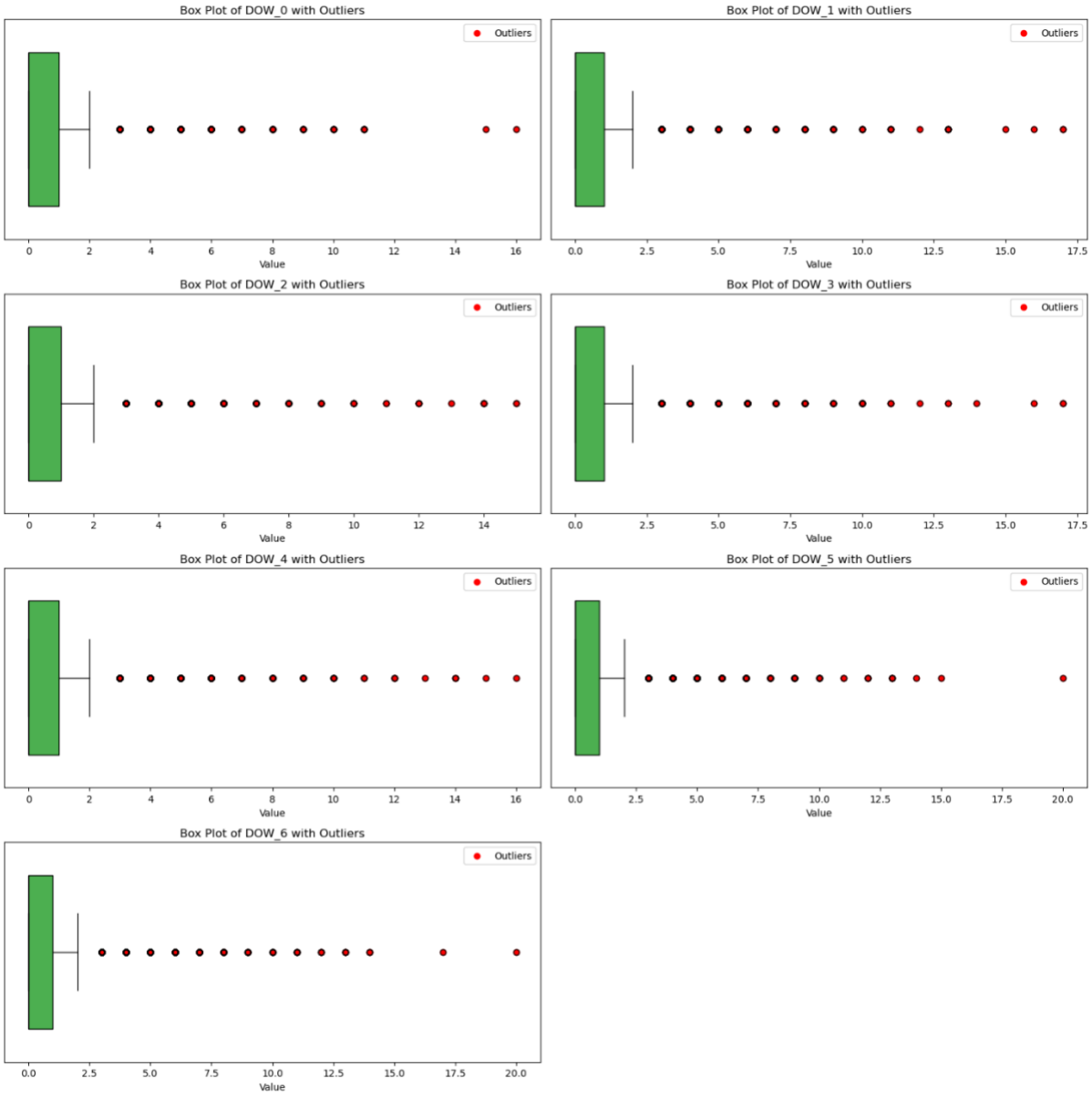
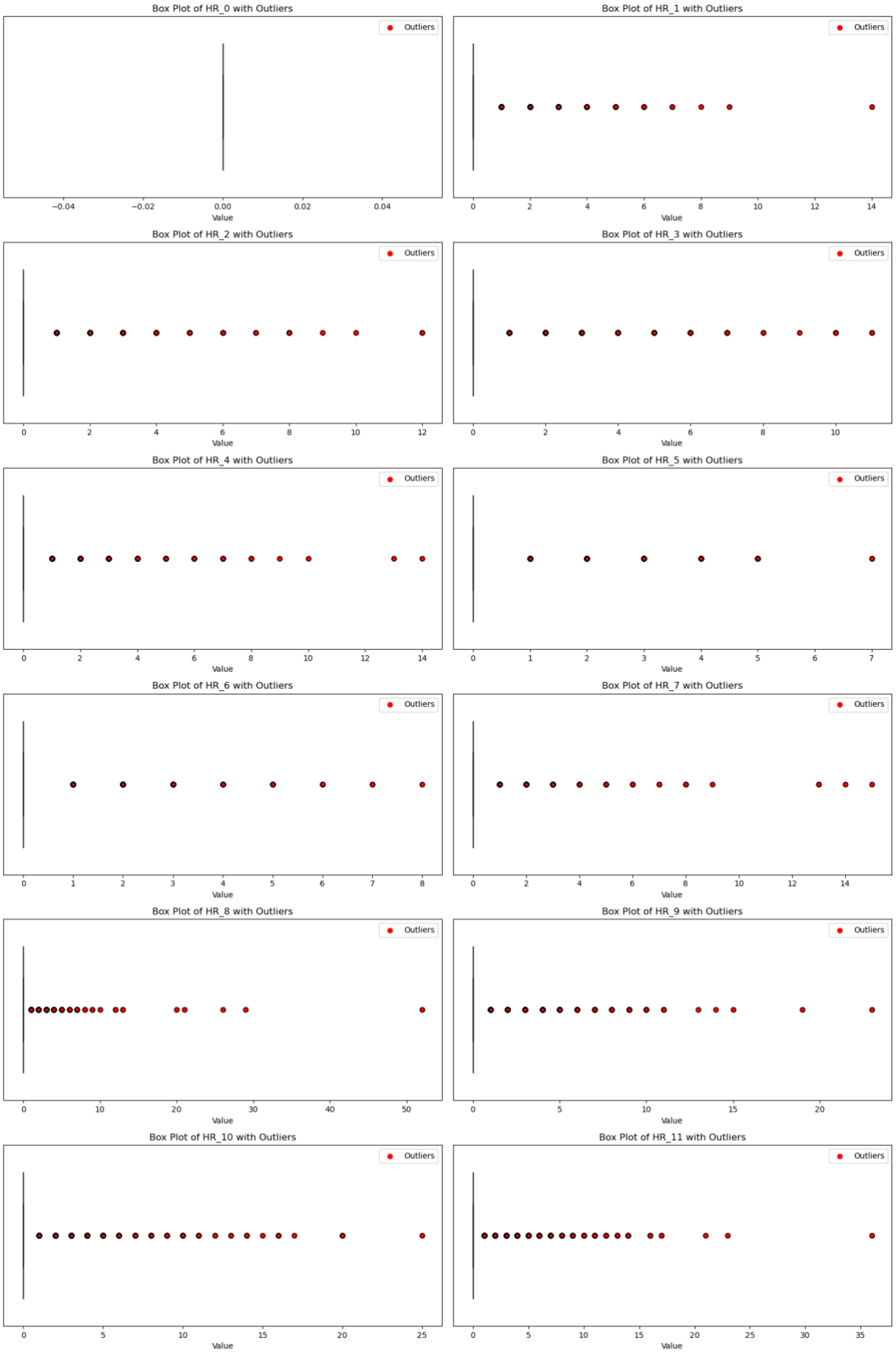

Box Plot of DOW_0 with Outliers
Box Plot of DOW_1 with Outliers
Box Plot of DOW_2 with Outliers
Box Plot of DOW_3 with Outliers
Box Plot of DOW_4 with Outliers
Box Plot of DOW_5 with Outliers
Box Plot of DOW_6 with Outliers
Outliers
Value

Figure 8


Box Plot of HR_0 with Outliers
Box Plot of HR_1 with Outliers
Box Plot of HR_2 with Outliers
Box Plot of HR_3 with Outliers
Box Plot of HR_4 with Outliers
Box Plot of HR_5 with Outliers
Box Plot of HR_6 with Outliers
Box Plot of HR_7 with Outliers
Box Plot of HR_8 with Outliers
Box Plot of HR_9 with Outliers
Box Plot of HR_10 with Outliers
Box Plot of HR_11 with Outliers
Outliers
Value

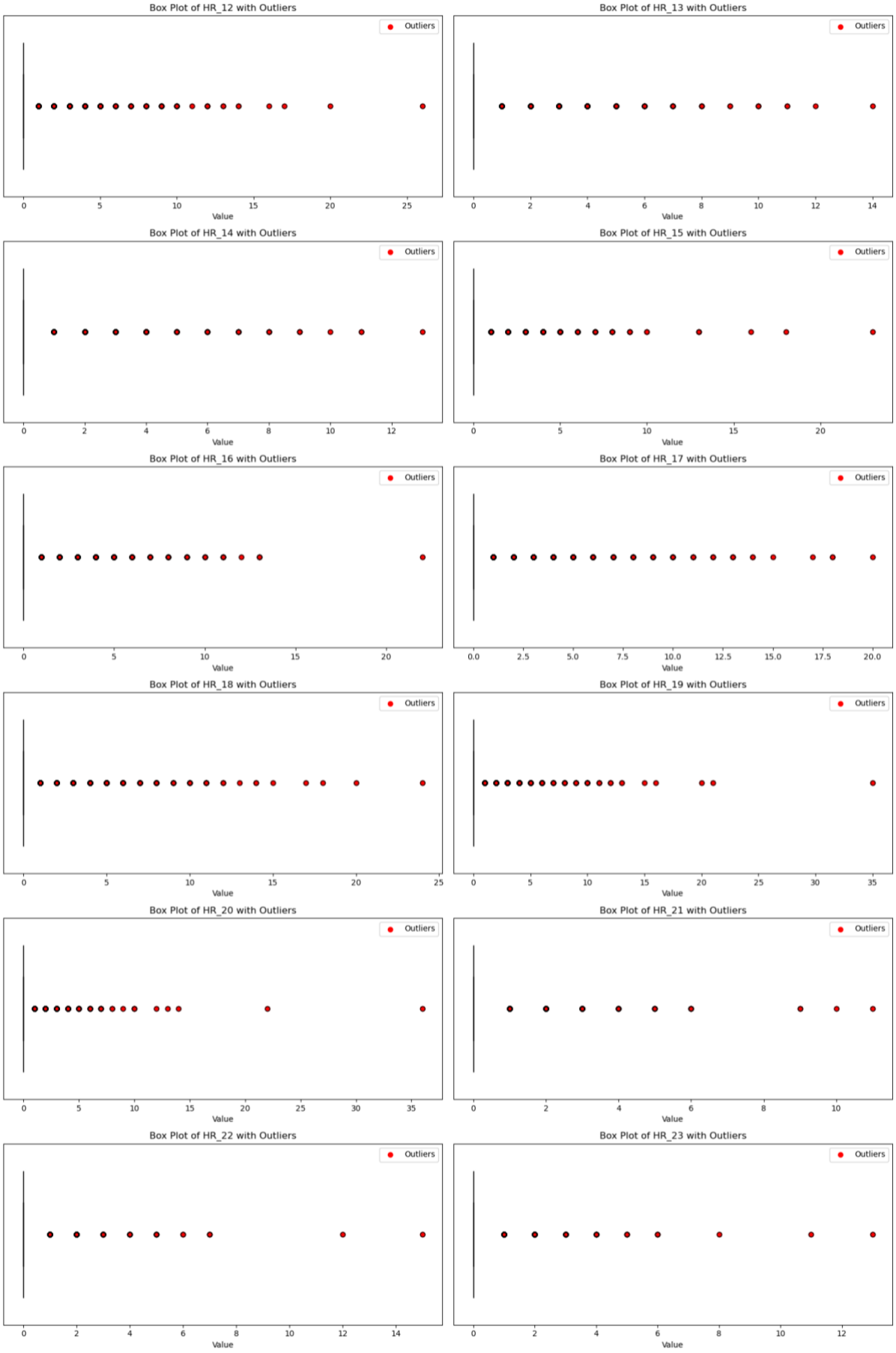

Box Plot of HR_12 with Outliers
Box Plot of HR_13 with Outliers
Box Plot of HR_14 with Outliers
Box Plot of HR_15 with Outliers
Box Plot of HR_16 with Outliers
Box Plot of HR_17 with Outliers
Box Plot of HR_18 with Outliers
Box Plot of HR_19 with Outliers
Box Plot of HR_20 with Outliers
Box Plot of HR_21 with Outliers
Box Plot of HR_22 with Outliers
Box Plot of HR_23 with Outliers
Outliers
Value

Figure 9

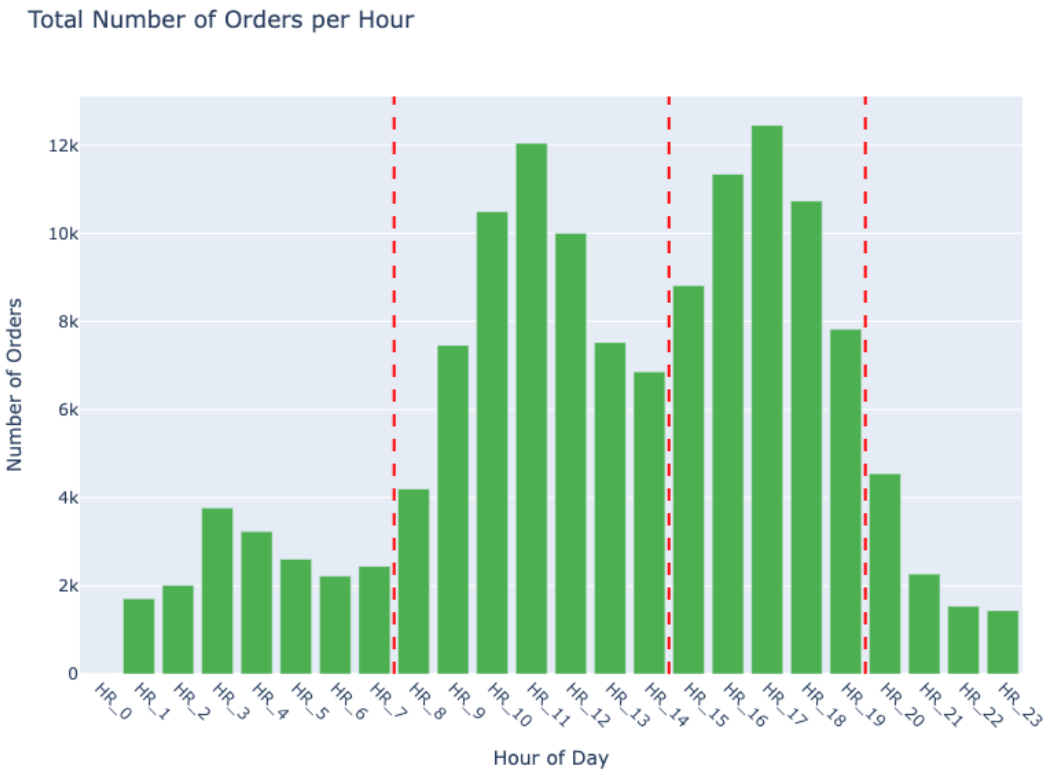


Figure 10

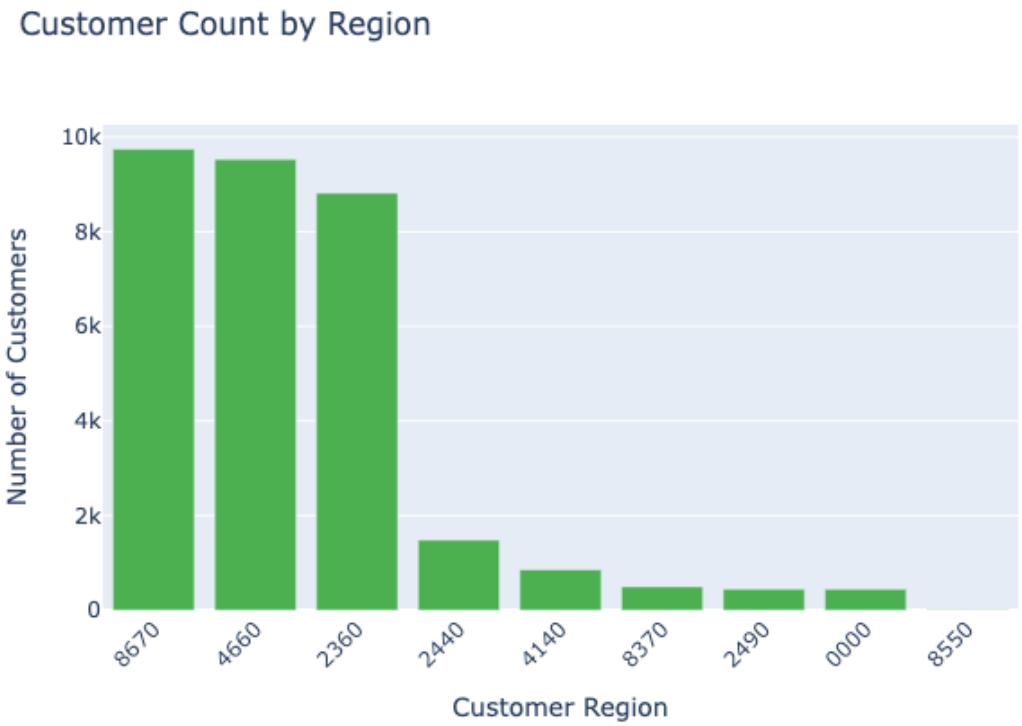


Figure 11

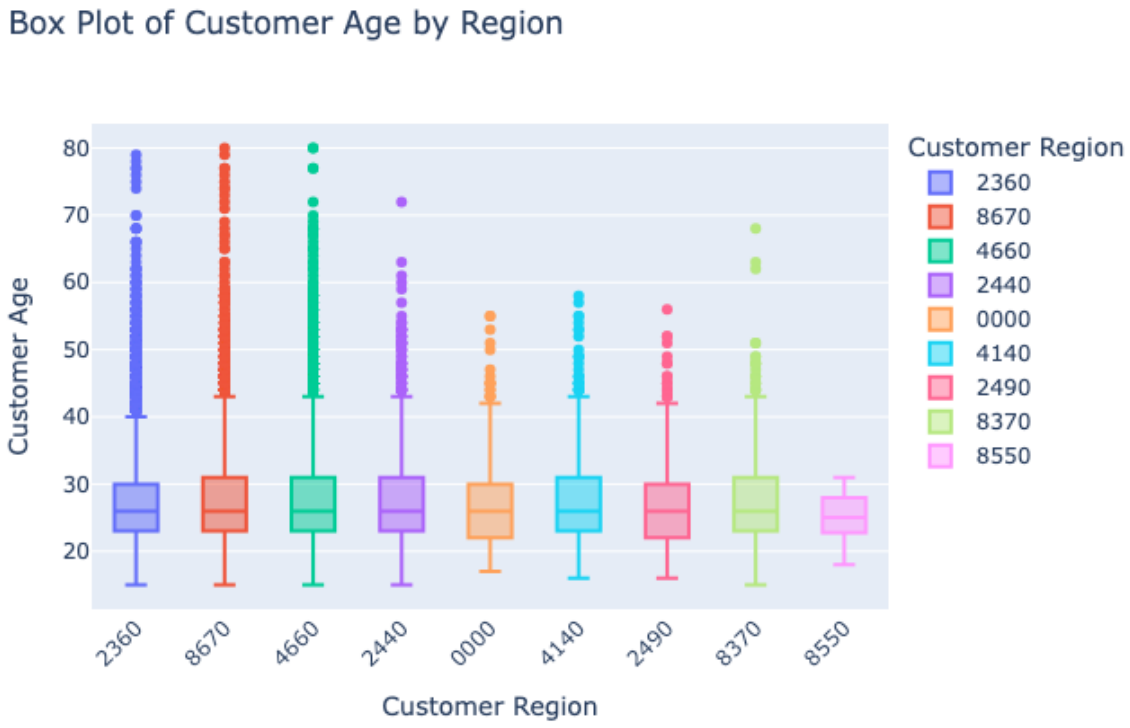


Figure 12

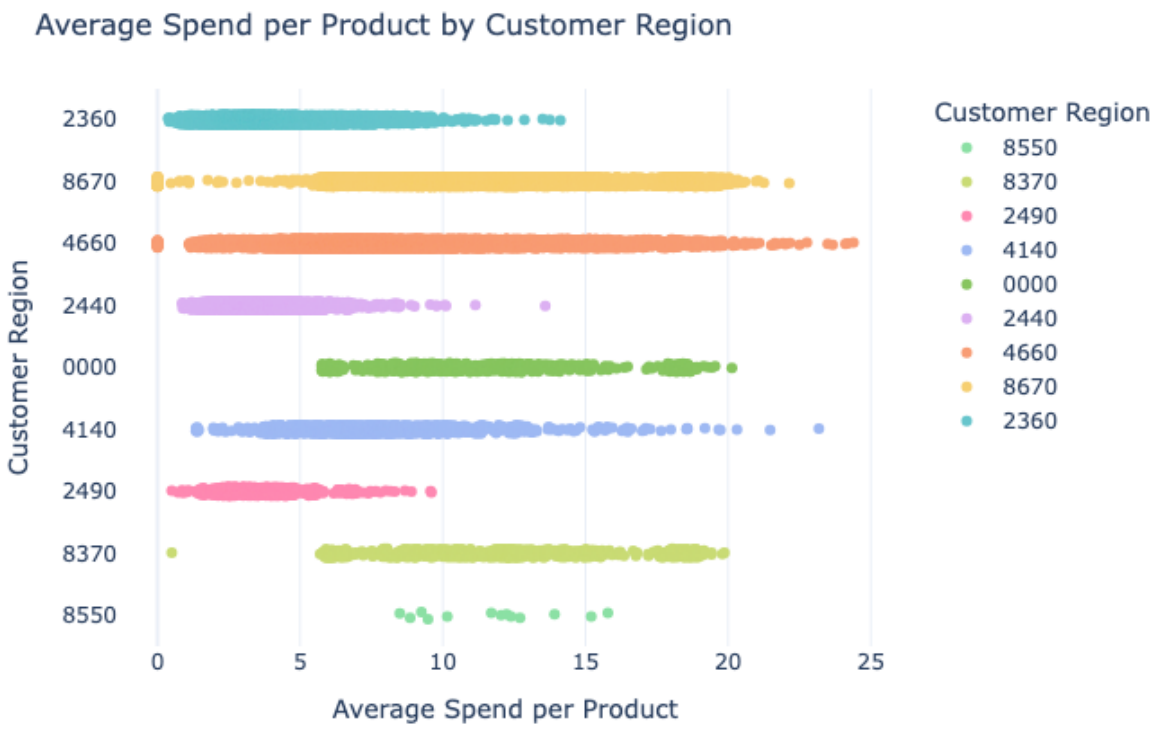

Average Spend per Product by Customer Region
Customer Region
2360
8670
4660
2440
0000
4140
2490
8370
8550
0
5
10
15
20
25
Average Spend per Product
Customer Region
8550
8370
2490
4140
0000
2440
4660
8670
2360

Figure 13

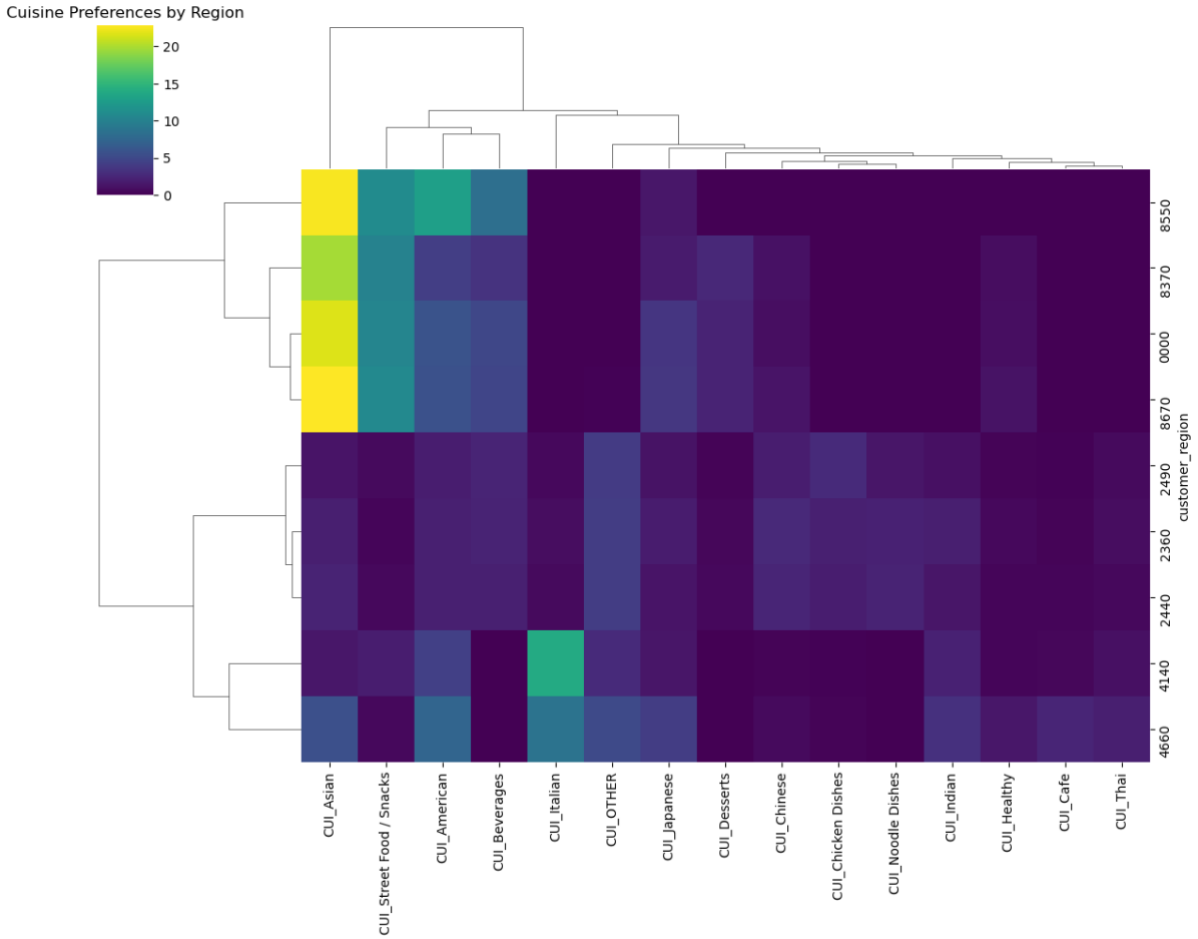


Figure 14

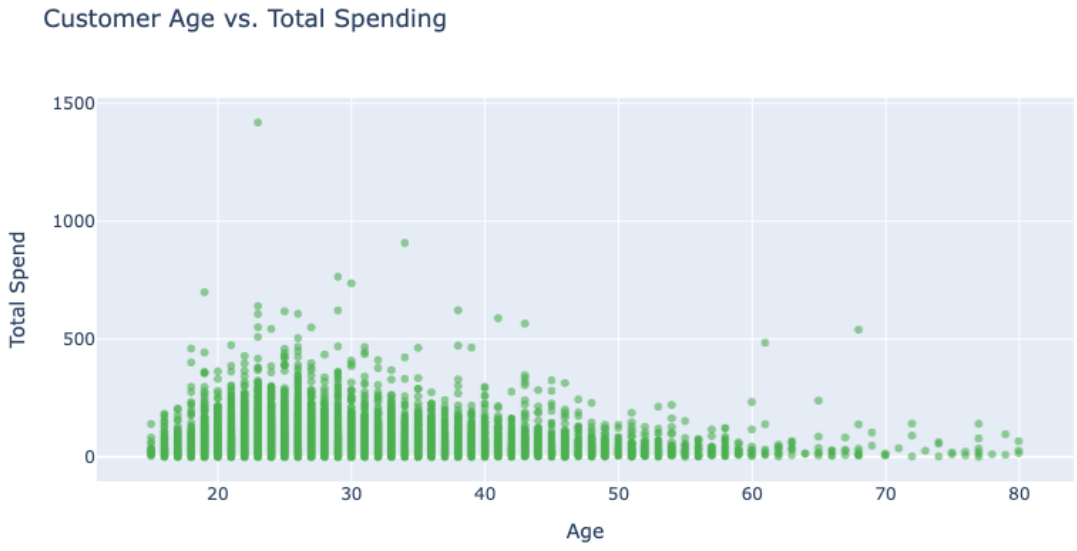


Figure 15

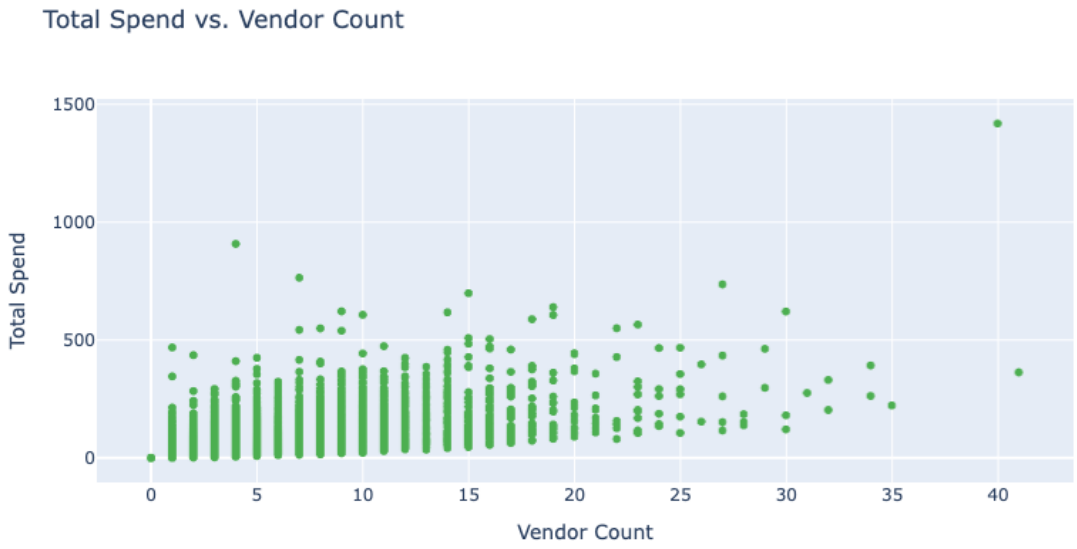


Figure 16

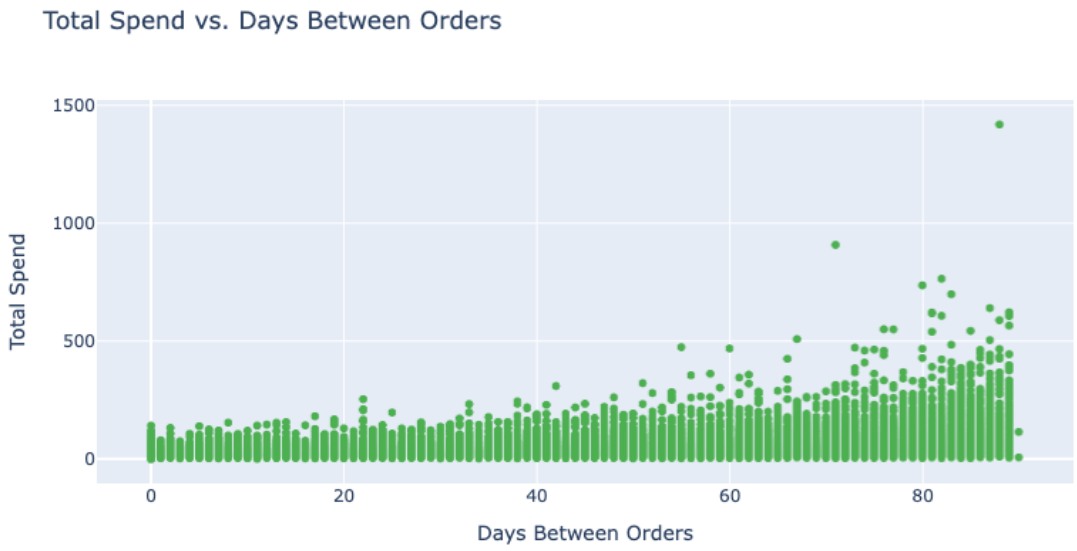


Figure 17

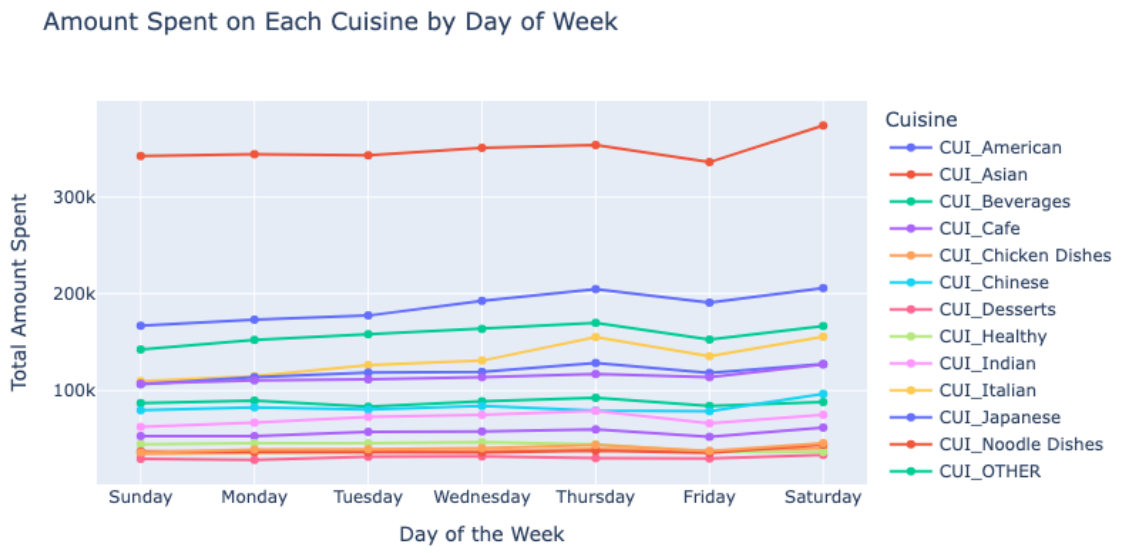


Figure 18

Figure 19

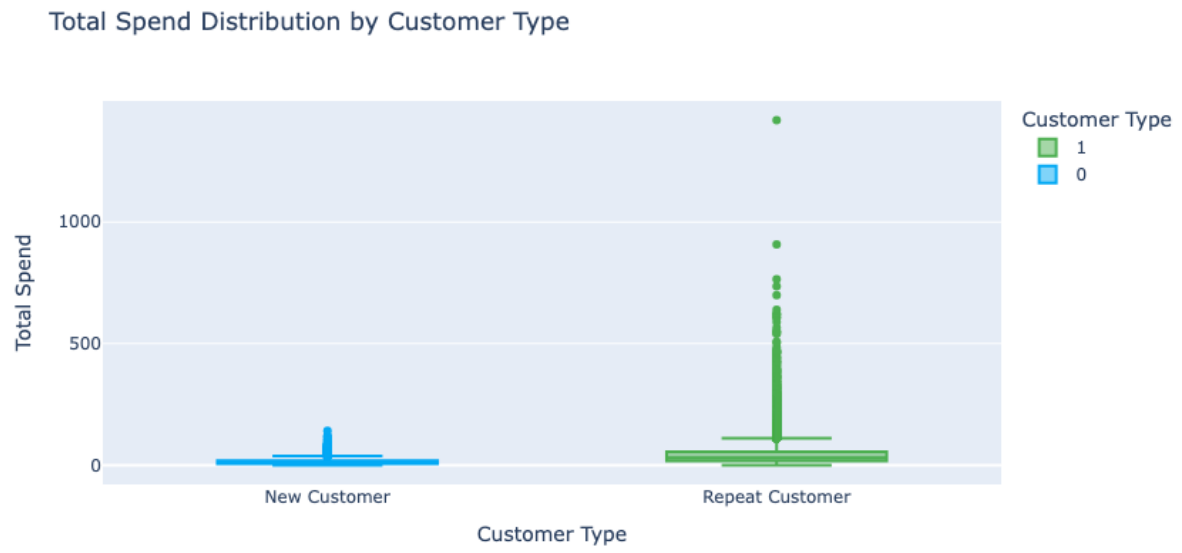


Figure 20

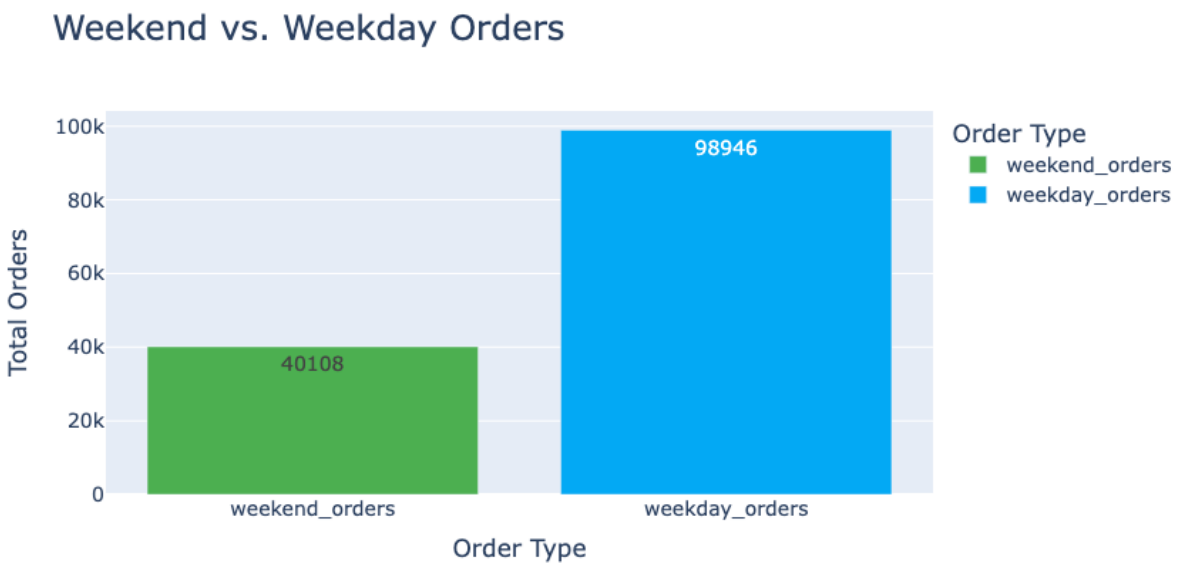


Figure 21

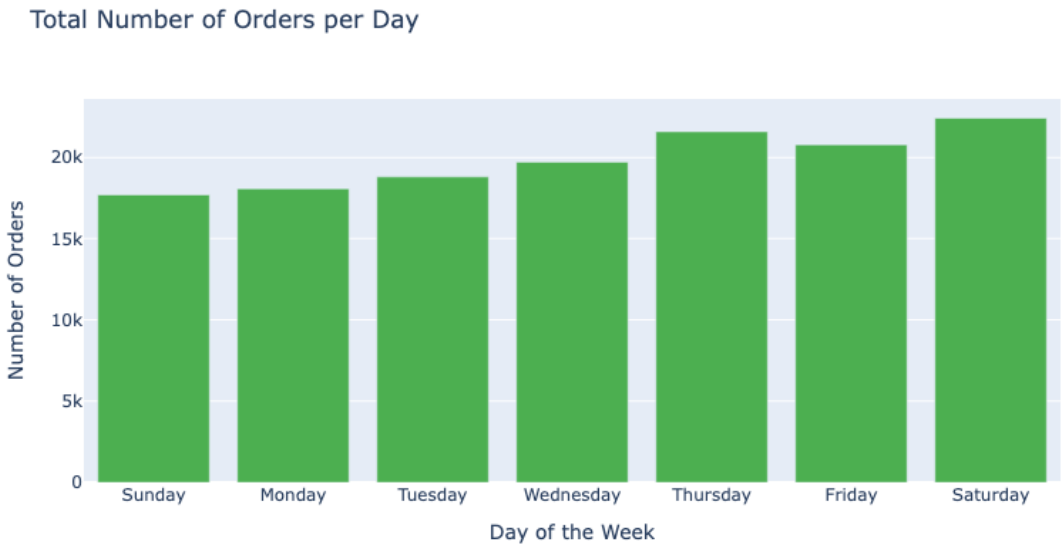


Figure 22

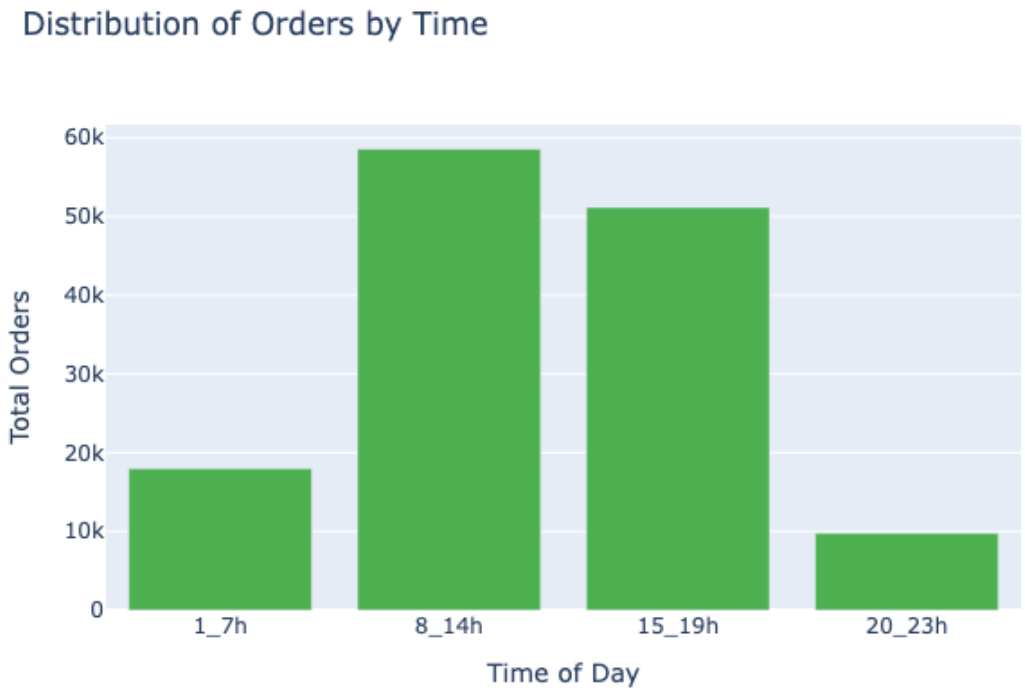


Figure 23

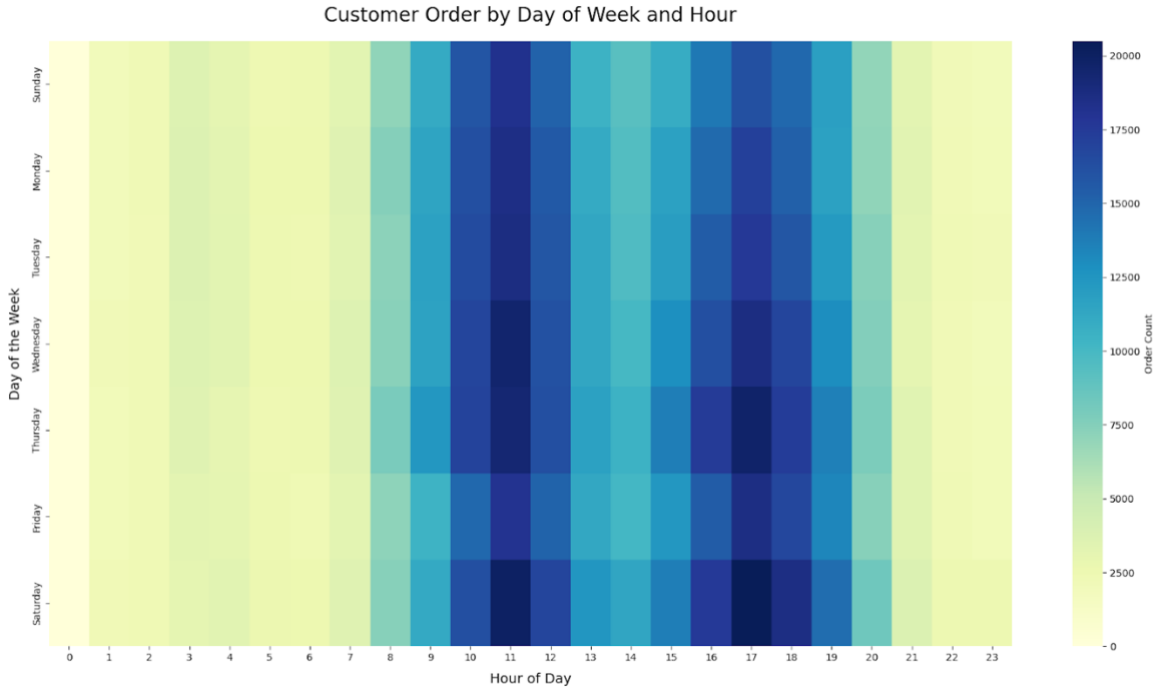

Customer Order by Day of Week and Hour
Day of the Week
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Hour of Day
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Order Count
20000
17500
15000
12500
10000
7500
5000
2500
0

Figure 24

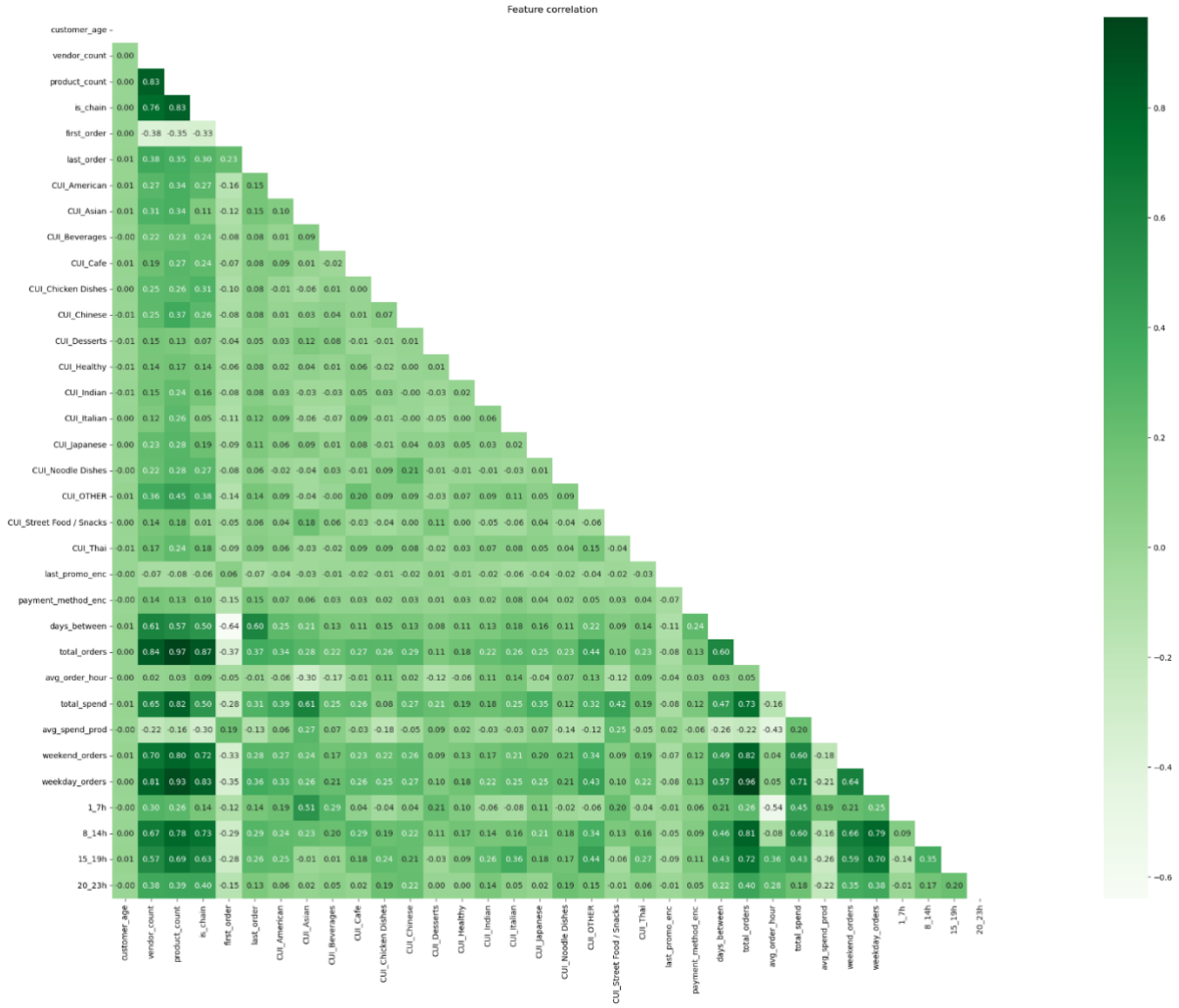

Feature correlation
customer_age
vendor_count
product_count
is_chain
first_order
last_order
CUI_American
CUI_Asian
CUI_Beverages
CUI_Cafe
CUI_Chicken Dishes
CUI_Chinese
CUI_Desserts
CUI_Healthy
CUI_Indian
CUI_Italian
CUI_Japanese
CUI_Noodle Dishes
CUI_OTHER
CUI_Street Food / Snacks
CUI_Thai
last_promo_enc
payment_method_enc
days_between
total_orders
avg_order_hour
total_spend
avg_spend_prod
weekend_orders
weekday_orders
1_7h
8_14h
15_19h
20_23h
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

Figure 25

6 APPENDIX B - TABLES

New Variable	Description	Variable(s) used
days_between	Days between the first and last transaction	last_order first_order
total_orders	Total number of orders placed	all DOW_ variables
avg_order_hour	Weighted average hour at which customers placed an order	all HR_ variables
total_spend	Total amount spent by each customer	all CUI_ variables
avg_spend_prod	Average of money spent per product	total_spend product_count
is_repeat_customer	Customers that made more than one order	days_between

Table 1