DATA MINING PROJECT

Master in Data Science and Advanced Analytics

NOVA Information Management School

Universidade Nova de Lisboa

ABCDEats Inc.
Exploratory Data Analysis

Group 20

Afonso Gamito, 20240752
Gonçalo Pacheco, 20240695
Hassan Bhatti, 20241023
Moeko Mitani, 20240670

Fall/Spring Semester 2024-2025

TABLE OF CONTENTS

1. INTRODUCTION

This project focuses on segmenting ABCDEats Inc. (ABCDE) customers using data from three cities over three months to identify groups based on demographics, behavior, and value. By analyzing customers through these perspectives, we uncover distinct segments within the customer base. The unique characteristics of each segment will guide specific marketing strategies, enabling ABCDE to deliver personalized service, enhance customer satisfaction, and drive revenue growth.

This project involves two main processes. In this report, we focus on conducting an Exploratory Data Analysis of the dataset and creating new features.

2. IN-DEPTH EXPLORATION OF THE DATASET

2.1. Anomalies within The Dataset

Data processing is a critical step in enhancing our model's performance and results. During this stage, we identify anomalies within the dataset for data processing in the next process.

2.1.1. Missing Values

The dataset consists of 31888 rows and 56 features. The five features were identified having missing values in the dataset as follows: 727 NaNs in *customer_age*, 106 NaNs in *first_order*, and 1165 NaNs in *HR_0*. In addition, there are 442 missing values "-" in *customer_region* and 16748 missing values "-" in *last_promo* (see Appendix A). The feature *last_promo* has 16748 missing values, which is more than 50% of the rows. Therefore, it is necessary to consider how to handle this later.

2.1.2. Wrong dtypes

The five features were identified as having incorrect dtypes in the dataset as follows: *customer_region* should be int instead of obj*, customer_age*, *first_order,* and *HR_0* should be int instead of float*, vendor_count* should be bool instead of int.

2.1.3. Outliers

The box plots were generated to identify outliers in the features (see Appendix B). As can be seen, each feature has outliers that should be considered in later processing.

2.1.4. Duplicates

There are 13 rows that can be duplicated in *customer_id*. It should be considered in a later process.

2.2. Key Statistics Summary

The average age of customers is 27.5 years old, which is a young generation. Moreover, 25% of the customers are 23 or younger, and 75% are 32 or younger (see Appendix C).

The mean of the number of vendors the customers have ordered from is 3.1. 75% have ordered from four vendors. Thus, most of the customers have ordered from less than four vendors in three months.

The mean of the total number of products the customers have ordered is 5.67. 25% of the customers have ordered two products, 50% have ordered three products, and 75% have ordered seven products in three months. There is a high standard deviation of 6.96, reflecting significant variability in order volume.

It was also observed that most customers paid by card rather than by cash or digital.

Eight different regions can be seen in *customer_region*. There should be three cities instead, thus it seems like they are postal codes. Then, we can categorize them into three groups (cities): city A (2360, 2440 and 2490), city B (4660 and 4140), and city C (8670, 8370 and 8550) in the next process.

The region with the highest number of orders was 8670, followed by 4660 and 2360, far ahead of the other regions (see Appendix D). As expected, the region 8670 had the highest number of unique customers who placed orders, followed by 4660 and 2360 (see Appendix E). It can be observed that they show similar distributions.

2.3. Trends among The Customers

The dataset reveals that the majority of food delivery customers fall within the younger demographic range, predominantly between the ages of 18 and 32. Notably, 25% of customers are 23 years old or younger, while 75% are 32 years old or younger, indicating a strong skew towards younger age groups in the user base (see Appendix C and Appendix F).

In alignment with the younger demographic, analysis of payment methods indicates a preference for digital options, underscoring the trend toward cashless transactions.

Overall, Asian cuisine is the most popular among the customers, followed by American cuisine and Street food and snacks (see Appendix G). Further examination of customer behavior demonstrates a marked preference for specific types of cuisine. Asian cuisine is the most popular in region 8670, while Italian cuisine leads in region 4660. This trend is consistent across other city regions, though with significant variation in total order volumes. In region 4140, however, the popularity of these cuisines is less pronounced compared to the two largest regions. City A shows greater diversity in cuisine preferences than the other two cities. (Appendix H).

Temporal analysis of order patterns shows peak food ordering on Thursdays (DOW_4) and Saturdays (DOW_6). High demand occurs from 10:00 AM to 12:00 PM and 4:00 PM to 6:00 PM, representing the busiest times for food delivery activity (see Appendix I for the peak hours and days for the orders).

Most of the customers stay for two to three days, this suggests that customers may use the service sporadically, this aligns with the platform's popularity among younger customers, who may use food delivery services to fulfill immediate, occasional needs (see Appendix J).

2.4. New Feature Creation

Developing new features is essential for more effective customer segmentation, enabling tailored marketing strategies for each segment group. In our view, the following features could be highly valuable for our future strategy.

*2.4.1. **Regional Demographic** [city_A; city_B; city_C]*

By consolidating the eight customer regions (based on postal codes) into three new categorical features - City A (2360, 2440, 2490), City B (4660, 4140), and City C (8670, 8370, 8550) - we enhance segmentation for more insightful analysis. This grouping clarifies geographic trends in customer preferences, particularly cuisine choices, allowing for targeted marketing strategies tailored to regional tastes.

*2.4.2. **Cuisine Popularity Rate** [low_popularity; medium_popularity; high_popularity]*

We categorize cuisine types into three popularity levels - Low, Medium, and High - based on order volume thresholds. This segmentation supports targeted marketing: high-popularity cuisines can be broadly promoted, while low-popularity cuisines benefit from niche marketing to specific customer segments.

*2.4.3. **Age Segmentation** [15-23; 24-26; 27-31; 32-80]*

Based on statistical analysis, our age segmentation reveals valuable insights into customer demographics, spanning ages 15 to 80 and highlighting diverse preferences across age groups. This age spread emphasizes the need for tailored marketing strategies: younger consumers may respond to trends and social media, while older ones may value quality and brand reputation.

*2.4.4. **Peak Hours** [morning_peak; afternoon_peak]*

By analyzing daily activity patterns, we observe that the majority of orders occur between 9 AM and 7 PM, identifying these as peak hours. This timeframe can be further divided into two segments: the morning peak (9 AM to 1 PM) and the afternoon peak (2 PM to 7 PM). Understanding customer preferences and behaviors during these distinct periods allows us to tailor our logistical and marketing strategies to better meet the demands of different cities throughout the day. This proactive approach enables us to optimize resource allocation, enhance customer satisfaction, and adapt our offerings to align with customer tendencies during peak ordering times.

*2.4.5. **Average Revenue per Order** - Total Revenue (Cuisine) / Total Orders (Cuisine)*

Understanding how much revenue is generated per order can shed light on customer preferences and spending habits, allowing for targeted promotions and upselling opportunities. Also, it

helps assess the profitability of each cuisine. A higher average revenue per order indicates that customers are willing to spend more on certain cuisines.

*2.4.6. **Customers Ordering Level** [frequent_customers; moderate_customers, infrequent_customers]*

Setting thresholds for order frequency enables effective segmentation of customer behavior into Low, Medium, and High ordering levels. This approach identifies frequent customers, allowing us to implement targeted marketing strategies to drive retention and loyalty among high-value segments.

2.5. Visualization

Creating effective data visualizations is crucial for uncovering insights and enhancing decision-making processes. In our view, the following visualizations could significantly contribute to our analysis by highlighting trends, patterns, and relationships within the data that inform our strategic direction.

2.5.1. Total Number of Orders per Hour

This chart (see Appendix K) visualizes the total number of orders for each hour, providing insight into hourly ordering patterns that are essential for scheduling and optimizing operations. A bar plot was chosen for its effectiveness in highlighting differences across discrete time categories (hours), making it easy to identify high- and low-ordering periods. Findings from this analysis reveal peak order hours, likely during evening or lunch times, which assist in identifying demand spikes and informing resource allocation decisions.

2.5.2. Total Order per Day of Week (DOW)

This chart (see Appendix L) visualizes the distribution of orders across the days of the week, providing insights into which days experience higher customer activity. A bar chart was selected for its ability to clearly represent total orders for each day, making it easy to observe daily trends and identify peak ordering days. The analysis indicates that certain days, such as weekends, may show higher order volumes, which can guide staffing and inventory planning to meet demand fluctuations effectively.

2.5.3. Cuisine Popularity based on Total Orders

This bar chart (see Appendix M) visualizes the popularity of various cuisines based on total orders. A bar chart was chosen for its simplicity in allowing easy comparison across different cuisines. The color scheme indicates popularity levels, with light grey for low, medium grey for moderate, and dark grey for high popularity. This visualization highlights which cuisines

are performing well and reveals those that may benefit from increased marketing efforts or menu adjustments.

2.5.4. Key Purchasing Trends across Age Demographics

These scatter plots (see Appendices N through R) explore the relationship between customer age and various factors. Appendix N highlights age-related patterns in product count, while Appendix O shows how vendor interactions vary by age. Appendix P and Q focus on preferences for Asian and American cuisines among different age groups, and Appendix R examines age-based interest in street food and snacks. Together, these visualizations reveal key purchasing trends across age demographics.

2.5.5. Total Orders of All Cuisines by Customer Age

This bar chart (see Appendix S) illustrates total orders across all cuisines, segmented by customer age. This visualization helps identify age demographics that contribute most to overall orders, revealing key customer preferences.

2.5.6. Distribution of Client Duration (Days as Clients)

This histogram (see Appendix J) displays the frequency distribution of client duration in days. The x-axis represents the number of active days, and the y-axis shows the count of clients within each range. This visualization provides insights into typical client lifespans and patterns in retention.

2.5.7. Total Orders by Region and Cuisine Type

This bar chart (see Appendix H) illustrates the distribution of total orders across various customer regions, segmented by cuisine type. The x-axis represents different regions, while the y-axis indicates the total order count for each cuisine. This visualization helps to understand customer preferences and the popularity of different cuisines in each region.

3. CONCLUSION

In this report, we identified several anomalies within the dataset, including missing values, incorrect data types, outliers, and duplicates, along with a key statistical summary. We analyzed customer trends related to our service and explored potential new features to strengthen our models and results in the next process. Finally, we developed visualizations to enrich our exploratory data analysis and provide clearer insights.

The next process involves clustering customers based on the findings from this process, enabling a targeted marketing approach for each segment to fulfill their specific preferences and needs and eventually boost their retention and our revenue.

APPENDIX A

TABLE OF MISSING VALUES IN THE DATASET

Features	Missing value types	Total missing values
customer_age	NaN	727
first_order	NaN	106
HR_0	NaN	1165
customer_region	-	442
last_promo	-	16748

APPENDIX B

BOX PLOTS TO IDENTIFY OUTLIERS IN THE FEATURES

APPENDIX C

KEY STATISTICS SUMMARY OF THE DATASET

	customer_age	vendor_count	product_count	is_chain	first_order	last_order	CUI_American	CUI_Asian	CUI_Beverages	CUI_Cafe
count	31161.000000	31888.000000	31888.000000	31888.000000	31782.000000	31888.000000	31888.000000	31888.000000	31888.000000	31888.000000
mean	27.506499	3.102609	5.668245	2.818866	28.478604	63.675521	4.880438	9.960451	2.300633	0.801163
std	7.160898	2.771587	6.957287	3.977529	24.109086	23.226123	11.654018	23.564351	8.479734	6.427132
min	15.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	23.000000	1.000000	2.000000	1.000000	7.000000	49.000000	0.000000	0.000000	0.000000	0.000000
50%	26.000000	2.000000	3.000000	2.000000	22.000000	70.000000	0.000000	0.000000	0.000000	0.000000
75%	31.000000	4.000000	7.000000	3.000000	45.000000	83.000000	5.660000	11.830000	0.000000	0.000000
max	80.000000	41.000000	269.000000	83.000000	90.000000	90.000000	280.210000	896.710000	229.220000	326.100000

CUI_Chicken Dishes	CUI_Chinese	CUI_Desserts	CUI_Healthy	CUI_Indian	CUI_Italian	CUI_Japanese	CUI_Noodle Dishes	CUI_OTHER	CUI_Street Food / Snacks
31888.000000	31888.000000	31888.000000	31888.000000	31888.000000	31888.000000	31888.000000	31888.000000	31888.000000	31888.000000
0.768096	1.431218	0.884359	0.950203	1.631153	3.233411	2.995379	0.711676	2.999913	3.913253
3.657273	8.191755	5.259868	5.830590	7.443234	11.247990	10.180851	4.536457	9.768300	15.548507
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
219.660000	739.730000	230.070000	255.810000	309.070000	468.330000	706.140000	275.110000	366.080000	454.450000

CUI_Thai	DOW_0	DOW_1	DOW_2	DOW_3	DOW_4	DOW_5	DOW_6	HR_0	HR_1	HR_2
31888.000000	31888.000000	31888.000000	31888.000000	31888.000000	31888.000000	31888.000000	31888.000000	30723.0	31888.000000	31888.000000
0.841697	0.555914	0.567486	0.591006	0.619449	0.677747	0.652973	0.704246	0.0	0.053845	0.063190
4.433047	1.013601	1.044090	1.045907	1.069672	1.088122	1.069947	1.167446	0.0	0.317013	0.351498
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000
0.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	0.0	0.000000	0.000000
136.380000	16.000000	17.000000	15.000000	17.000000	16.000000	20.000000	20.000000	0.0	14.000000	12.000000

HR_3	HR_4	HR_5	HR_6	HR_7	HR_8	HR_9	HR_10	HR_11	HR_12	HR_13
31888.000000	31888.000000	31888.000000	31888.000000	31888.0000	31888.000000	31888.000000	31888.000000	31888.000000	31888.000000	31888.000000
0.118759	0.101700	0.081943	0.069681	0.0768	0.131899	0.233912	0.329560	0.378167	0.314162	0.236453
0.500862	0.437493	0.358705	0.329461	0.3777	0.635582	0.724906	0.891161	0.959961	0.842484	0.637502
0.000000	0.000000	0.000000	0.000000	0.0000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	0.000000	0.0000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	0.000000	0.0000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	0.000000	0.0000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
11.000000	14.000000	7.000000	8.000000	15.0000	52.000000	23.000000	25.000000	36.000000	26.000000	14.000000

HR_14	HR_15	HR_16	HR_17	HR_18	HR_19	HR_20	HR_21	HR_22	HR_23
31888.000000	31888.000000	31888.000000	31888.000000	31888.000000	31888.000000	31888.000000	31888.000000	31888.000000	31888.000000
0.215630	0.277032	0.356435	0.390962	0.336961	0.245610	0.142812	0.071155	0.048263	0.045189
0.599006	0.738162	0.874449	0.943721	0.893949	0.795296	0.586529	0.348536	0.298265	0.282006
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
13.000000	23.000000	22.000000	20.000000	24.000000	35.000000	36.000000	11.000000	15.000000	13.000000

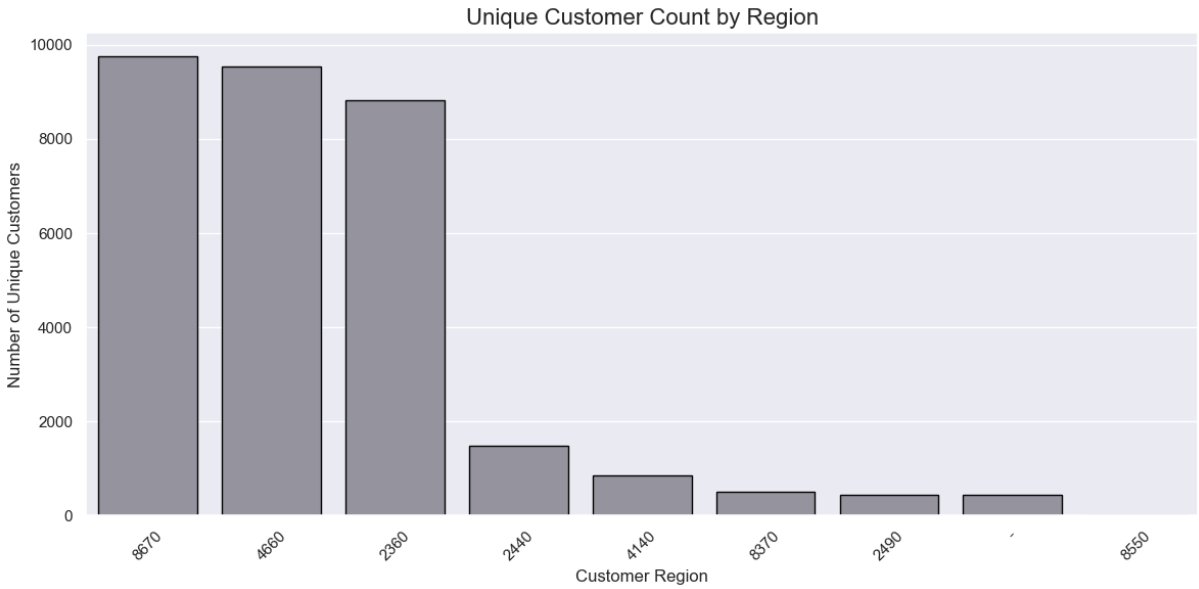
	customer_id	customer_region	last_promo	payment_method
count	31888	31888	31888	31888
unique	31875	9	4	3
top	742ca068fc	8670	-	CARD
freq	2	9761	16748	20161

APPENDIX D

BAR CHART OF TOTAL ORDERS BY CUSTOMER REGION

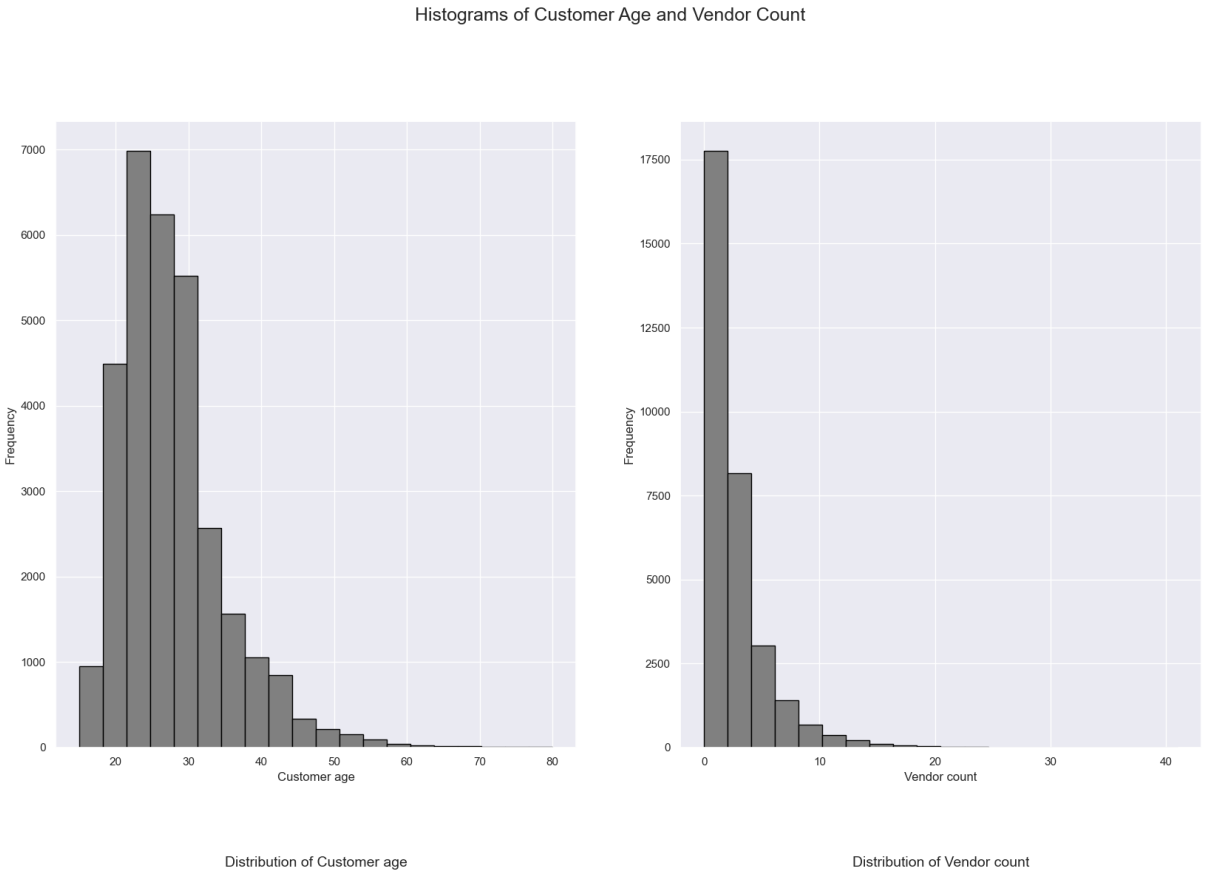
APPENDIX E

BAR CHART OF UNIQUE CUSTOMER COUNT BY CUSTOMER REGION



APPENDIX F

HISTOGRAMS OF CUSTOMER AGES AND VENDOR COUNT

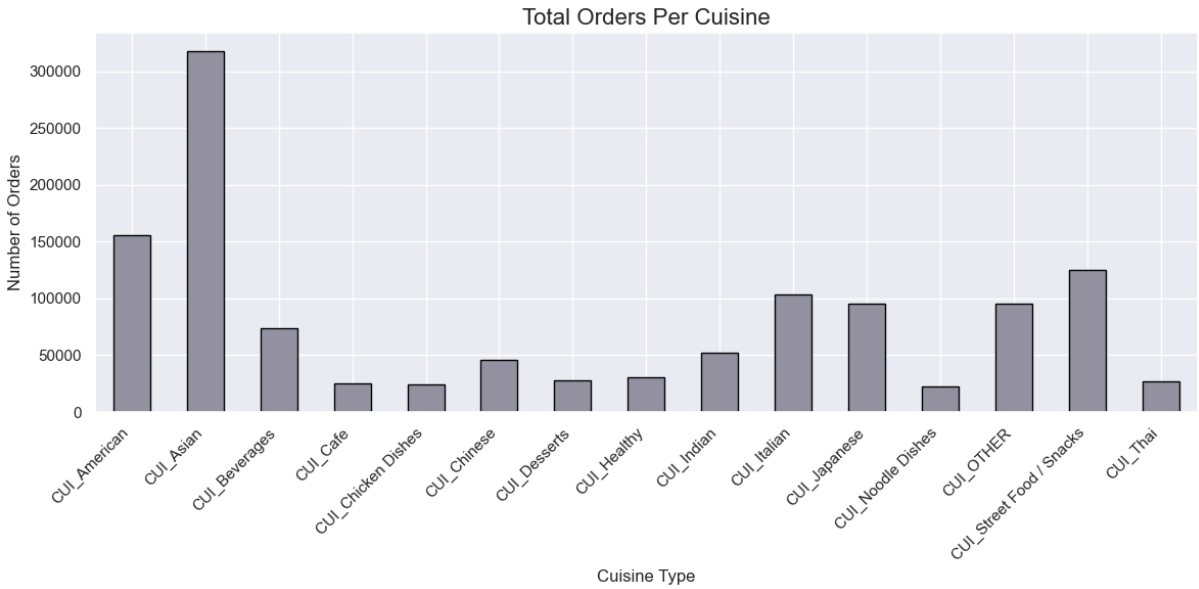


Distribution of Customer age

Distribution of Vendor count

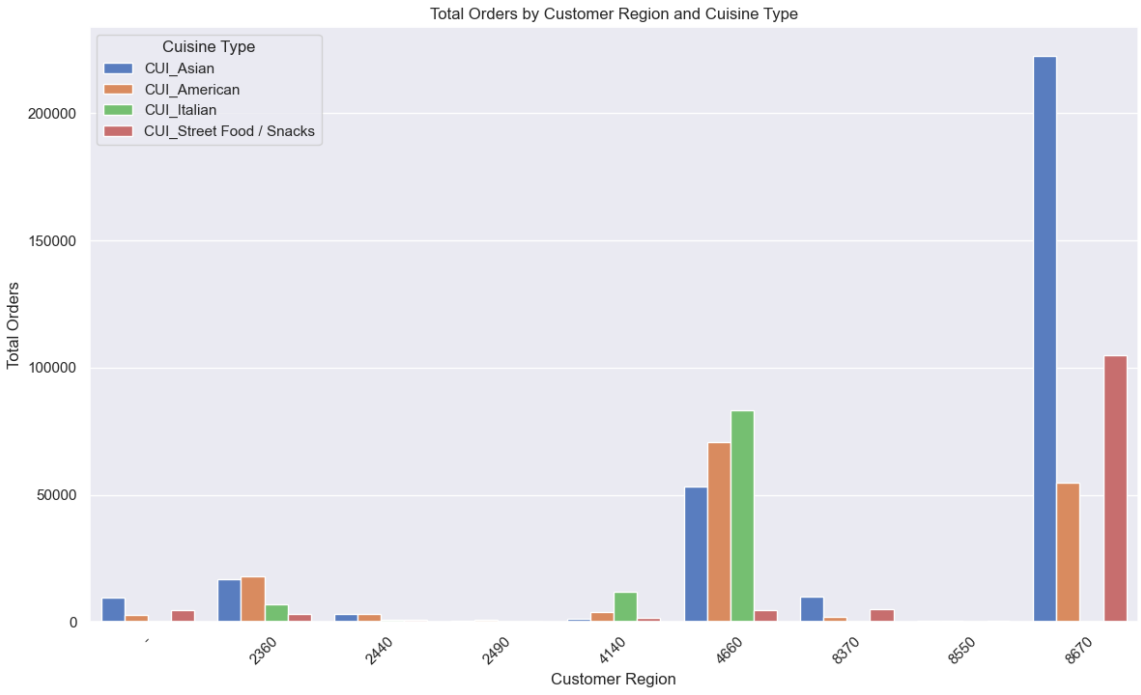
APPENDIX G

BAR CHART OF POPULAR CUISINE

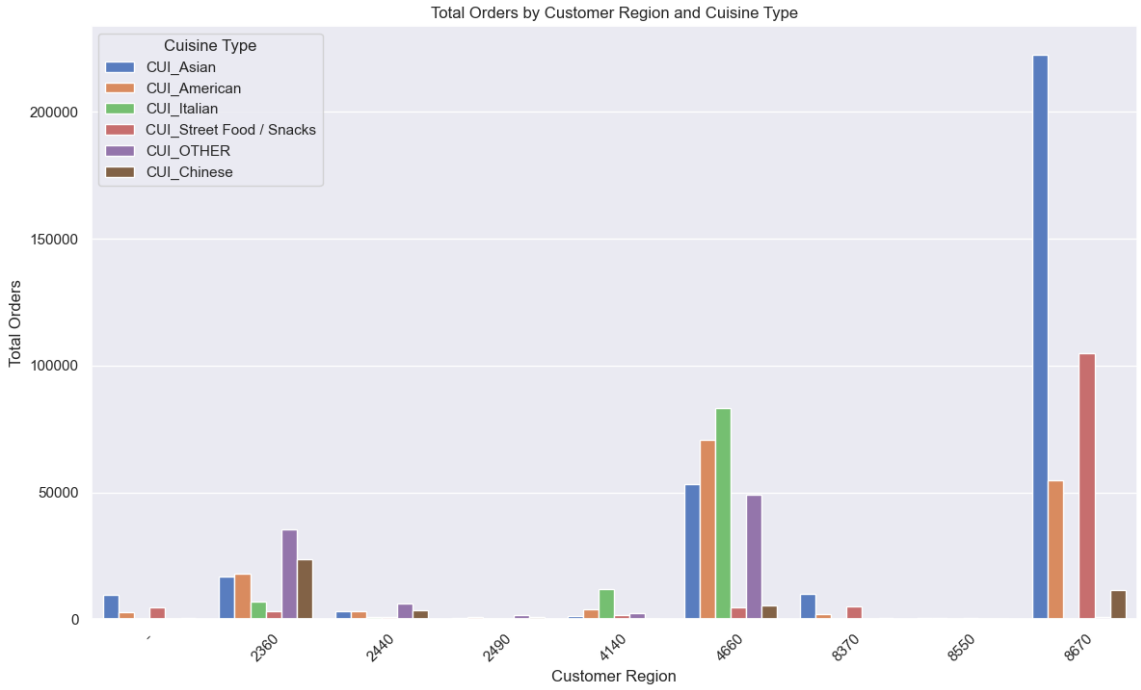


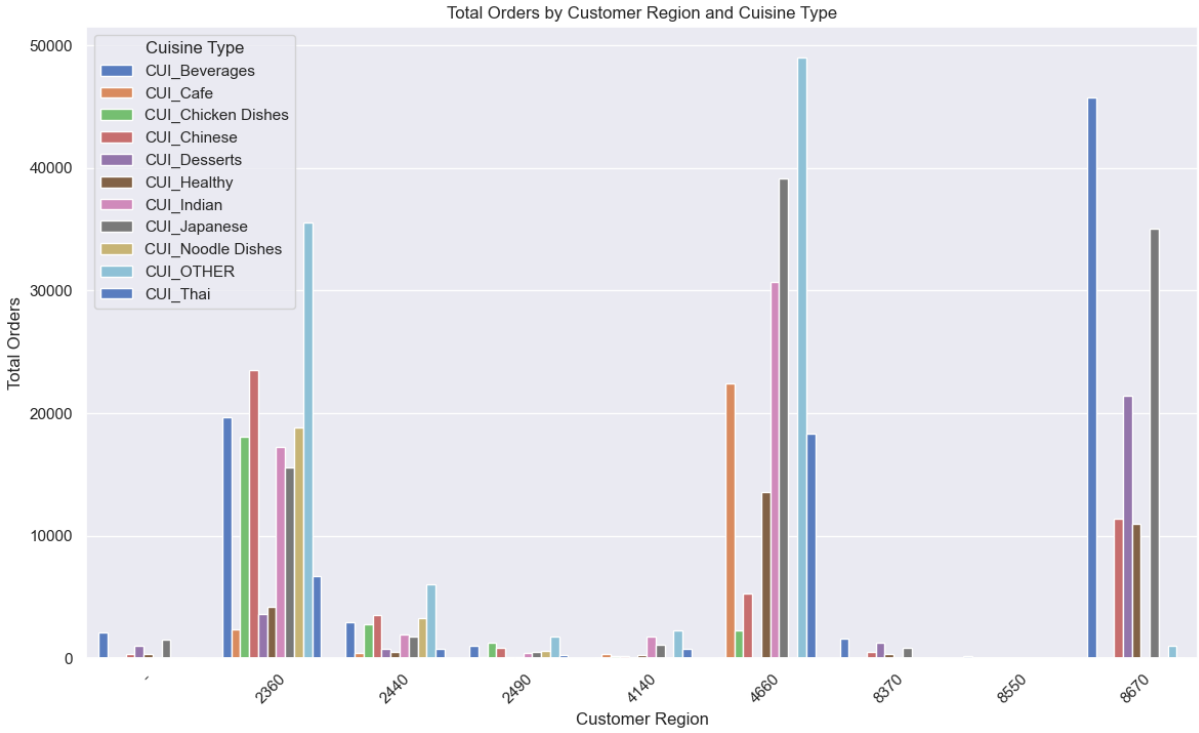
APPENDIX H

BAR CHARTS OF TOTAL ORDERS BY CUSTOME REGION AND CUISINE TYPE



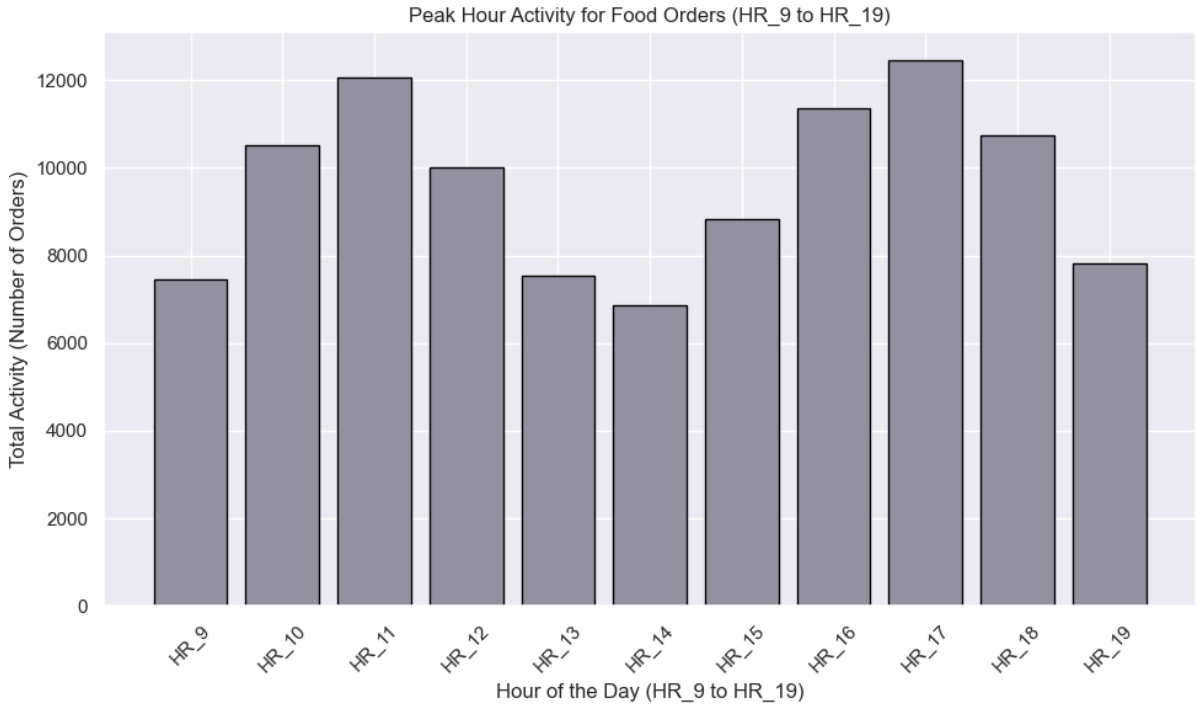
R




Total Orders by Customer Region and Cuisine Type
Cuisine Type
CUI_Beverages
CUI_Cafe
CUI_Chicken Dishes
CUI_Chinese
CUI_Desserts
CUI_Healthy
CUI_Indian
CUI_Japanese
CUI_Noodle Dishes
CUI_OTHER
CUI_Thai
Total Orders
50000
40000
30000
20000
10000
0
2360
2440
2490
4140
4660
8370
8550
8670
Customer Region

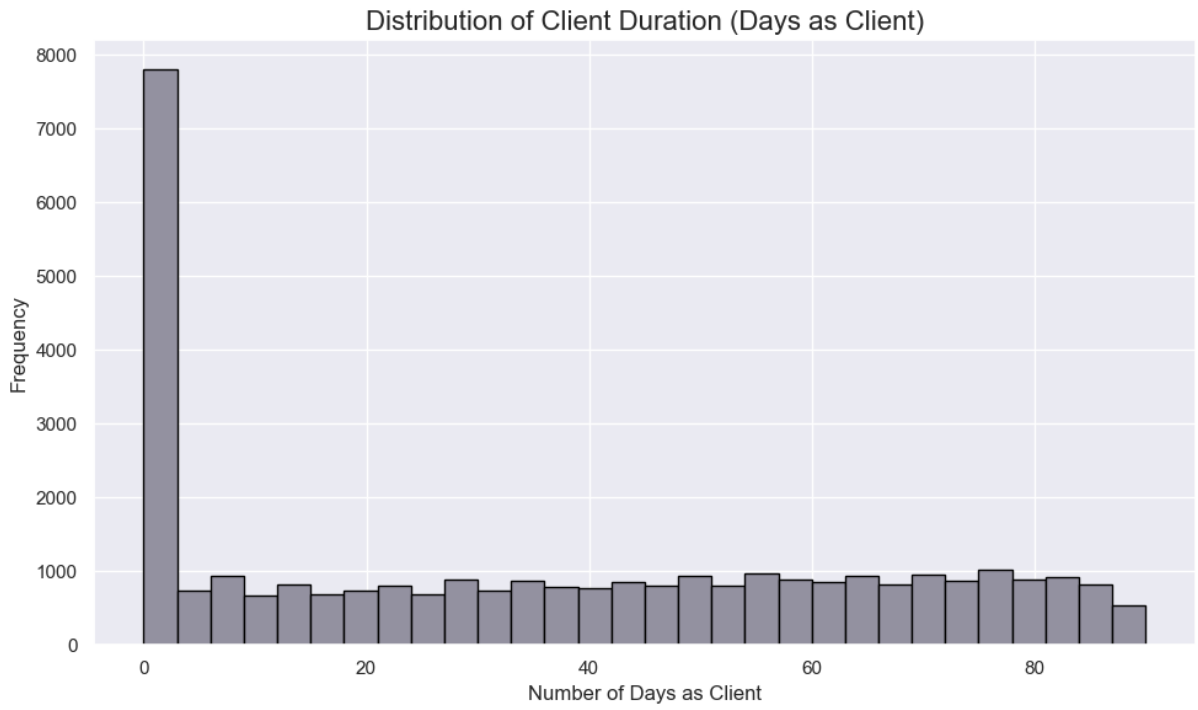
APPENDIX I

BAR CHART OF PEAK HOURS AND DAYS FOR THE ORDERS



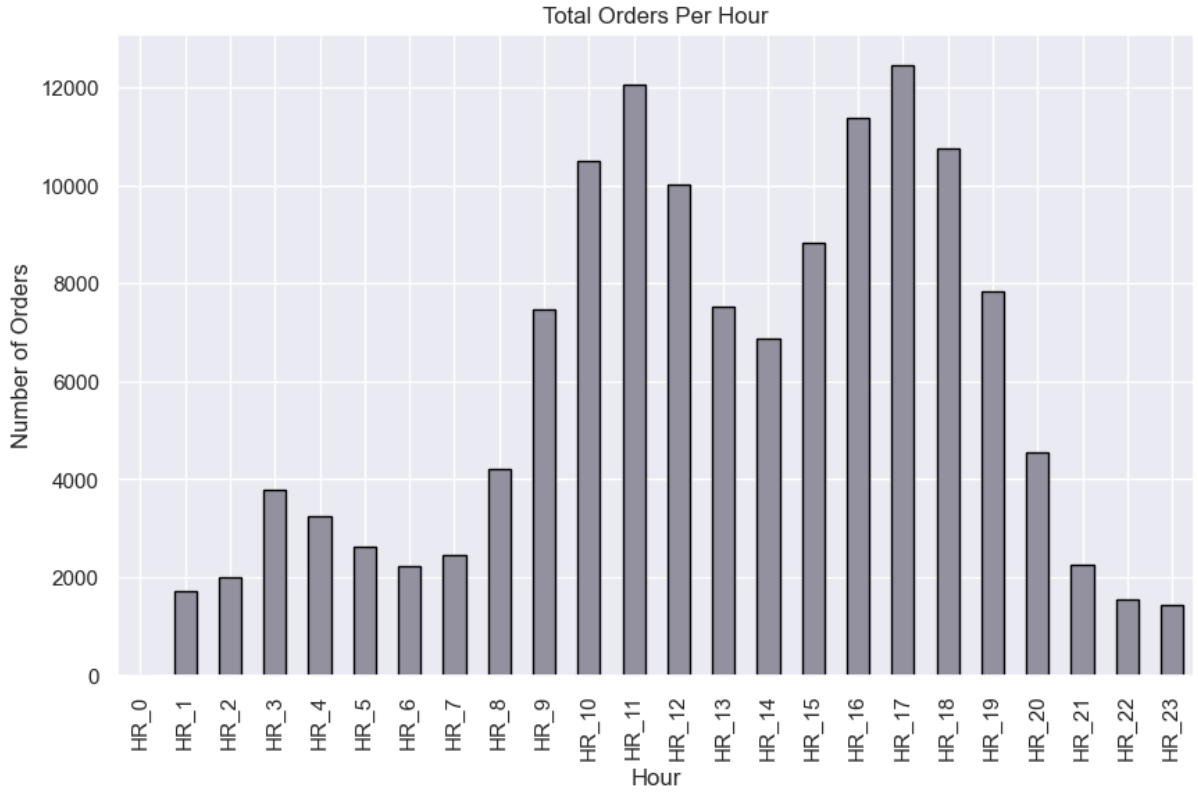
APPENDIX J

BAR CHART OF CLIENT SPREAD OVER 3 MONTHS



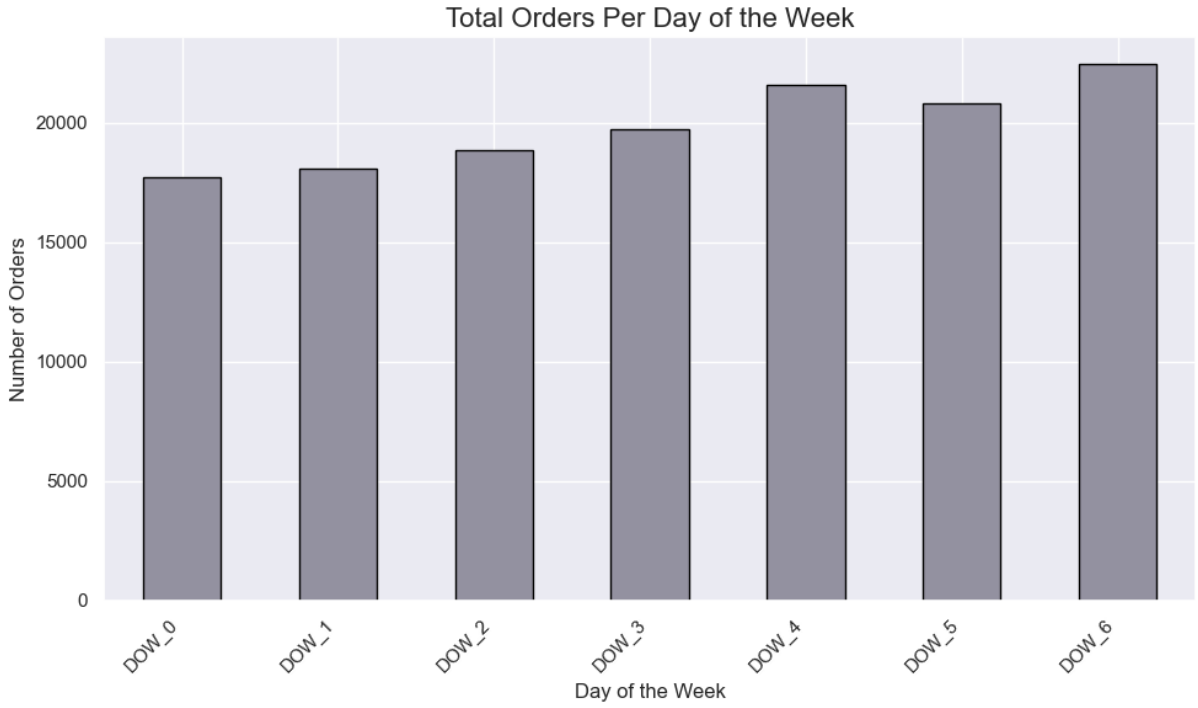
APPENDIX K

BAR CHART OF TOTAL ORDERS PER HOUR



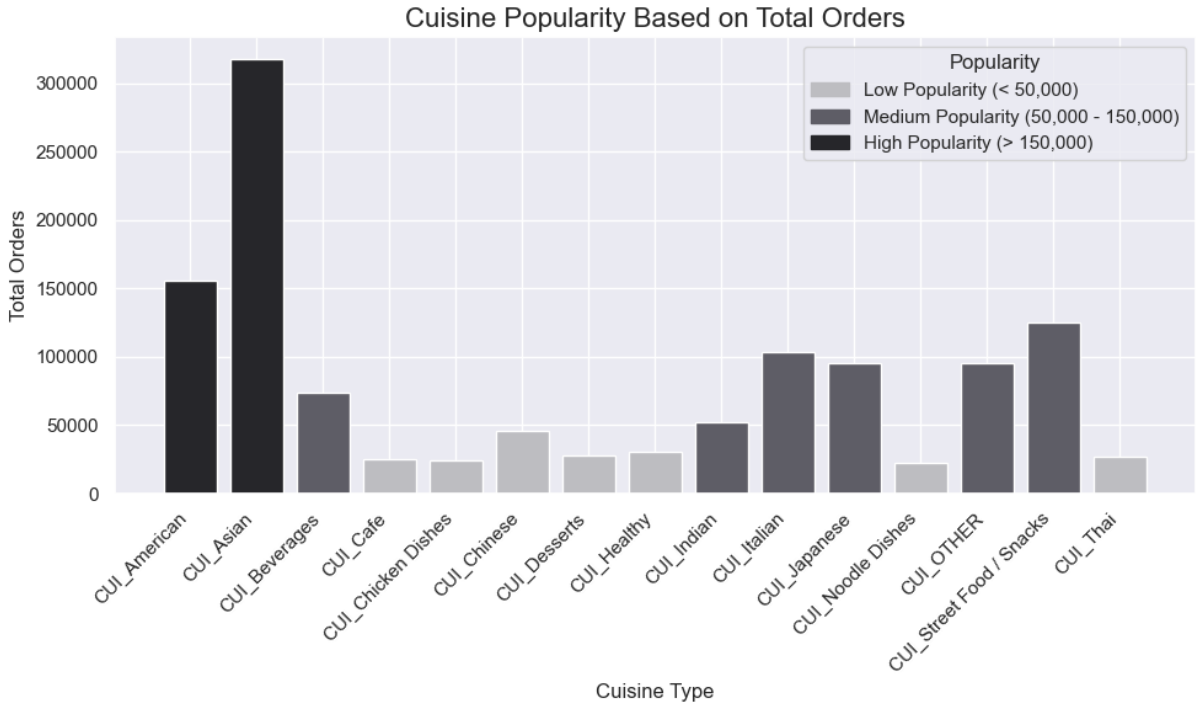
APPENDIX L

BAR CHART OF TOTAL ORDERS PER DOW



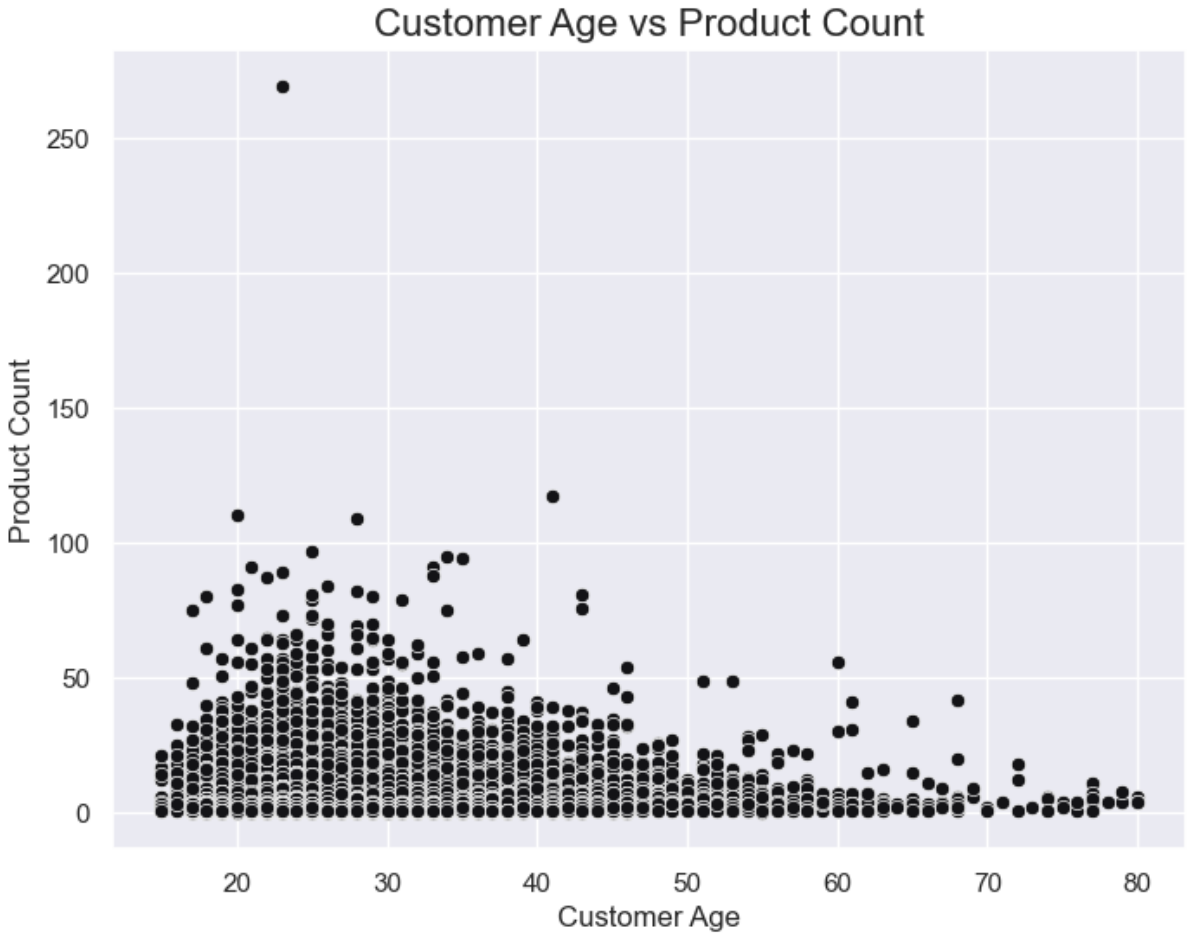
APPENDIX M

BAR CHART OF CUISINE POPULARITY PER TOTAL ORDERS



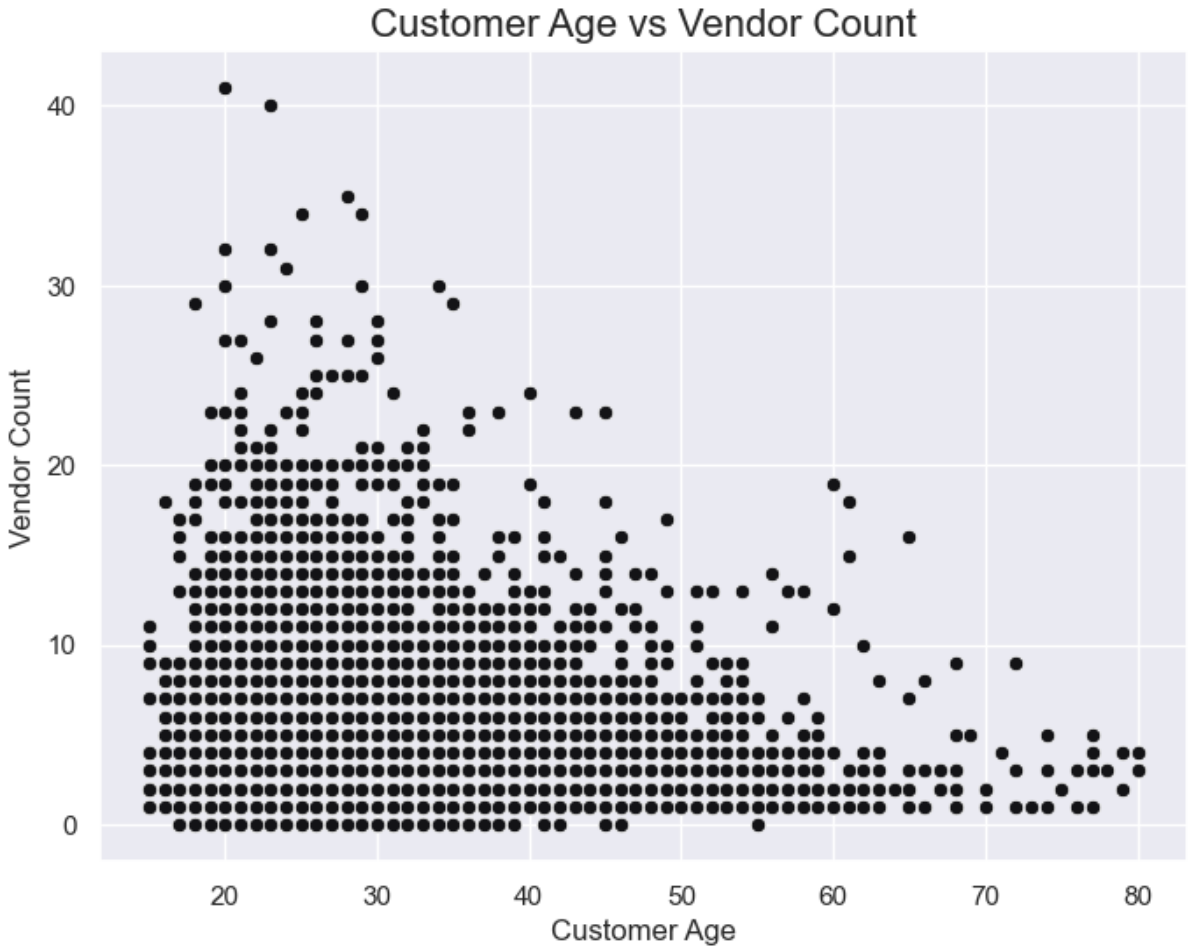
APPENDIX N

SCATTER PLOT OF PRODUCT COUNT PER CUSTOMER AGE



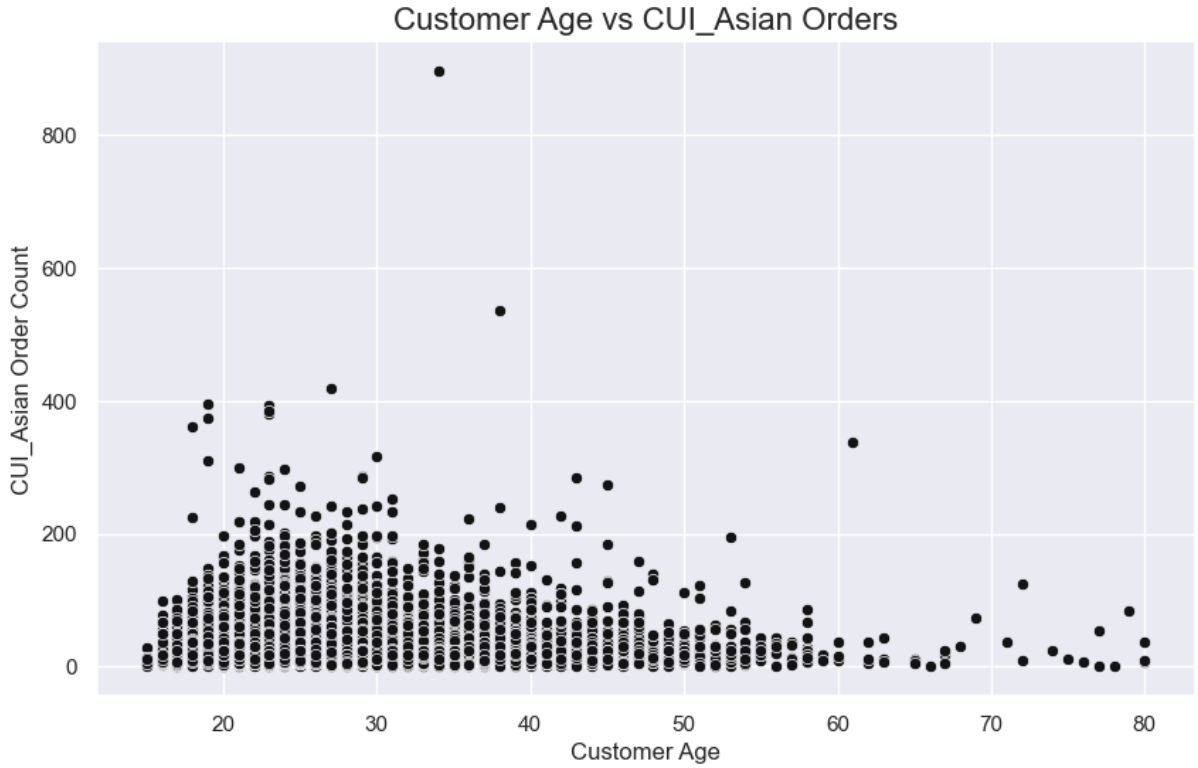
APPENDIX O

SCATTER PLOT OF PRODUCT COUNT PER CUSTOMER AGE



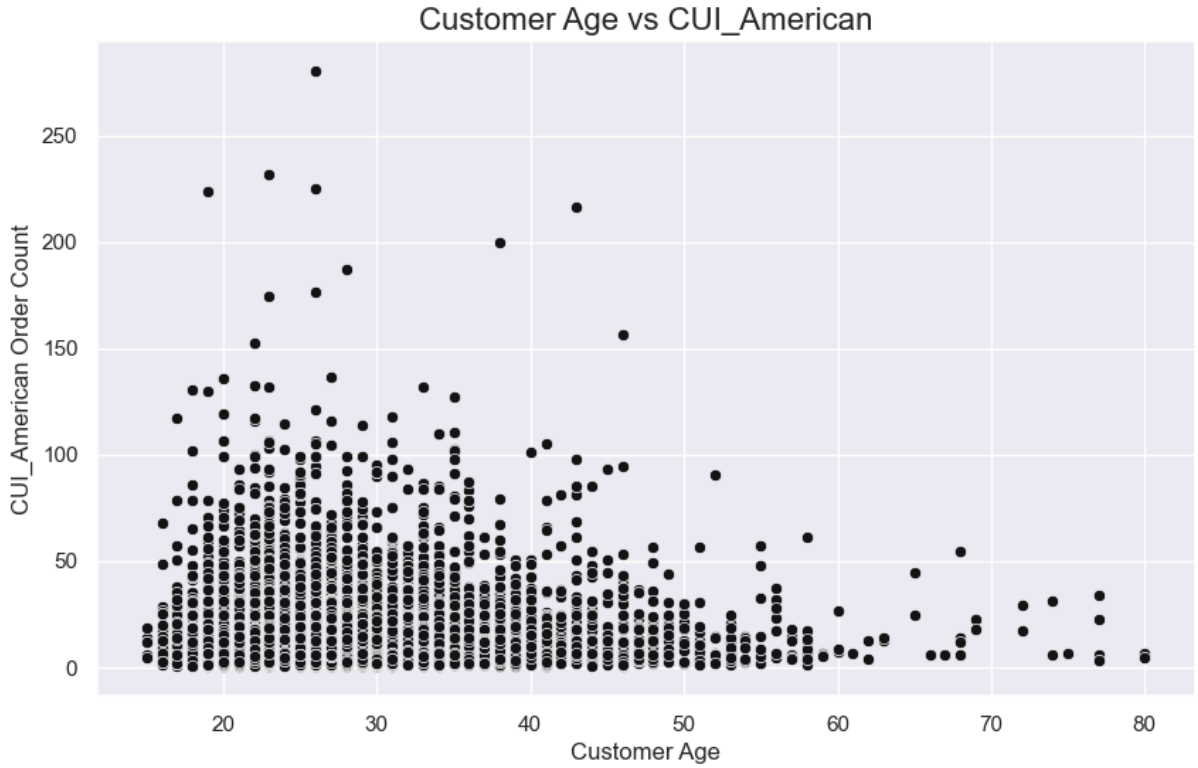
APPENDIX P

SCATTER PLOT OF ASIAN CUISINE ORDERS BY CUSTOMER AGE



APPENDIX Q

SCATTER PLOT OF AMERICAN CUISINE ORDERS BY CUSTOMER AGE



APPENDIX R

SCATTER PLOT OF STREET FOOD / SNACKS ORDERS BY CUSTOMER AGE

APPENDIX S

BAR CHART OF TOTAL ORDERS (FROM ALL CUISINES) BY CUSTOMER AGE