DATA MINING PROJECT

Master in Data Science and Advanced Analytics

NOVA Information Management School

Universidade Nova de Lisboa

ABCDEats Inc. - Customer Segmentation

EDA

Group 10

Alexandra Pinto, 20211599

Marco Galão, r20201545

Sven Goerdes, 20240503

Tim Straub, 20240505

Fall/Spring Semester 2024-2025

TABLE OF CONTENTS

1. Introduction

This report details the exploratory data analysis (EDA) phase for ABCDEats Inc., a fictional food delivery platform offering a variety of restaurant options across regions. By analyzing customer data, we aim to identify patterns and anomalies to support a customer segmentation strategy. This stage involves data inspection and appropriate treatments, alongside feature engineering, all backed by visualizations, to enable accurate clustering and enhance our understanding of customer preferences.

2. Description of the Dataset

2.1. Data Overview

The original dataset from ABCDEats Inc. contains 31,888 rows and 56 columns. Each row represents a unique customer, identified by *customer_id,* which was set as an index, with attributes detailing customer behavior over three months. This section outlines the dataset's initial structure, categorizing the columns into six main groups (see Table 1).

Table 1 - Initial Dataset Overview

Category	Columns	Description	Segmentation purpose	# Columns
Demographics	*customer_region* *customer_age*	Geographic location and age of customers	Aid in segmenting customers by age and region, enhancing targeted marketing	2
Ordering Patterns	*DOW_0 to DOW_6* *HR_0 to HR_23*	Frequency of orders placed by day of the week and hour of the day	Reveal peak ordering times, informing time-sensitive promotions	31
Cuisine Spending	*CUI_American* *CUI_Asian* *…*	Monetary amounts spent on various cuisine types	Highlights popular cuisines, guiding menu, and campaign optimizations	15
Purchase Behavior	*vendor_count* *product_count* *is_chain*[1]	Counts of unique vendors and products ordered	Indicates vendor and product diversity, facilitating personalized recommendations	3
Order Timing	*first_order* *last_order*	Timing of the customer's first and last orders	Provides insights into customer loyalty and recency for re-engagement efforts	2
Promotional Activity	*payment_method* *last_promo*	Method of payment and promotion applied to the last order	Refine discount strategies based on payment methods and promotional engagement	2
∑ columns (excluding *customer_id*)				55

1: *is_chain* should indicate chain orders but needs review due to values not equal to 0 or 1.

2.2. Multivariate Relationships

The pairplot in Figure 1 reveals multiple scatterplots for the initial dataset, highlighting strong positive correlations among purchase behavior variables. This suggests that customers who place more orders[2] tend to explore more vendors. Additionally, higher spending[2] is associated with greater product variety and vendor engagement. Most Pearson correlations among purchase behavior variables exceed 0.75 (see Figure 2), reinforcing these observations. The plots also show a significant number of multivariate outliers, evident from the isolated data points across the charts.

Figure 3 illustrates that most customers are under 55 years old. Similar distributions are observed in *vendor_count*, *product_count*, *is_chain, total_orders*[2], and *total_cui_spending*[2] across age groups. The near-zero correlations in Figure 2 confirm that age has a lower impact on these purchasing behaviors. Additionally, Figure 4 reveals consistent spending on each cuisine across all age groups.

3. Exploring the Dataset

3.1. Constant Features, Duplicates, Missing Values, and Data Types

This section focuses on identifying and addressing constant features, duplicates, and missing values in the dataset, along with the rationale behind each decision (see Table 2).

Table 2 - Constant Features, Duplicates, Missing Values, and Data Types Summary

Category	Feature	Issue Type	Missing Type [Bib. 1]	Decision
Constant Features	*HR_0*	Zero variance	-	Removed *HR_0* variable [1]
Duplicate Records	*customer_id (index)*	Duplicate records	-	Removed duplicates based on *customer_id* [2]
	All features	Duplicate records	-	Removed duplicate records across all features [3]
Missing Values	*last_promo*	"-" values	MNAR	Labeled as "NOPROMO" [4]
	customer_region	"-" values	MAR	Reassigned to region "8670" [5]
	customer_age	NaN values	MCAR	Imputed using median [6]
	first_order	NaN values	MAR	Imputed as 0 for most cases; removed 2 ambiguous cases [7]

[1] The *HR_0* feature provided no variance, thus no valuable information.

[2] 13 duplicate customer IDs were removed to ensure a unique customer identifier.

[3] 47 fully duplicated records across all features were removed, totaling 60 rows (0.19%) alongside *customer_id* duplicates, to maintain data integrity.

2: The *total_cui_spending* and *total_orders* features aggregate the *CUI_* and *DOW_* variables, respectively.

[4] Distribution analysis (Figures 5, 6, and 7) suggested that "-" values (52.54%) likely indicate customers who did not use promotions, as their patterns align with other labeled categories.

[5] The "-" values (1.39%) were analyzed against other region codes (Figures 8, 9, and 10). Distribution patterns suggest that codes with similar prefixes indicate geographically close areas. Based on this, we inferred that these "-" values align most closely with the region "8670" over "8370" and "8550" (Figure 11). In the future, a KNN approach will be considered for imputing these missing values.

[6] Due to the right-skewed age distribution and outliers, missing values (2.28%) were imputed using the median rather than the mean for greater robustness (Figure 12). In the future, an alternative approach would be to drop NaN values, given that they are MCAR and unlikely to bias the results: "Accepting missing data is best for MCAR because they are unlikely to bias your results." [Bib 1.]

[7] Records with missing *first_order* values (0.33%) all had a *last_order* of 0, with most cases also indicating only one order placed in total (Figure 13). Thus, it is reasonable to assume these customers placed a single order on the dataset's first day, making an imputed value of 0 appropriate. Two exceptions (0.01%) showed two orders (Figure 14), creating ambiguity in order spacing, so these rows were removed to maintain dataset integrity.

These steps removed 62 records (0.19%), and 833 values (representing 0.05% of all data points) were imputed, resulting in a more complete and reliable dataset for analysis and segmentation. Subsequently, we adjusted the data types for *customer_age* and *first_order* to integers to accurately reflect their nature.

3.2. Coherence Checking

The coherence checks were conducted to identify and correct inconsistencies, thereby improving the dataset's accuracy and usability (see Table 3).

Table 3 - Coherence Checking Summary

Check	Purpose	Findings	Decision
Minor Customers	Identify customers under 18	365 minors (1.15%)	Removed minor customers to comply with targeting regulations
Order Consistency	Ensure *last_order* ≥ *first_order*	0 rows inconsistent	No action needed
CUI_Asian vs Specific Cuisines	Check for overlaps/redundancy in cuisine labeling	No evidence of multi-label classification (Figure 15); 3 customers with identical spends	Remove the 3 customers with identical spending to maintain the integrity
Sum *DOW_* vs Sum *HR_*	Verify consistency between daily and hourly order sums	1,153 inconsistencies (3.66%) in order counts (Figure 16)	Adjusted minor discrepancies (1-2) and removed larger ones (0.1%). Moving forward, total orders will be calculated based on the sum of *DOW_* rather than *HR_*, as it provides a more robust measure

Total Orders Check	Identify customers with 0 orders placed in the last 3 months	136 entries (0.43%) with no orders	Removed these entries as they do not contribute to segmentation
is_chain Variable	Validate the meaning of the *is_chain* variable	Analysis (Figure 17) suggests it tracks chain restaurant orders	Rename the variable to *chain_orders* for clarity
Total Orders and Vendor Count Consistency	Ensure total orders and *vendor_count* do not exceed *product_count*	18 rows (0.06%) with inconsistencies in orders or vendors	Remove the rows with inconsistencies to maintain data accuracy

The coherence checks resulted in the removal of 590 rows (1.85%) from the dataset. With these inconsistencies resolved, the next step is to examine the outliers.

3.3. Outliers

This section describes the boxplots and evaluates the outliers determined for all numerical features. Outliers were determined using the lower boundary criterion of Q1 – 1.5×IQR and the upper boundary criterion of Q3 + 1.5×IQR (see Table 4).

Table 4 - Outliers Summary

Feature Category	**Boxplot Description**	**Outliers Evaluation**
Cuisine Spending	High concentration of values near zero and a substantial spread of outliers extending far to the extremes (Figure 18, 19 and 20)	One extreme outlier in *total_cui_spending*. No anomalies in terms of spending per order across outliers of all cuisines.
Ordering Patterns	Median of features near zero (Figures 21 and 22), resulting in 9.41% outliers in cumulative feature (Figure 23)	The Outlier range is similar across all *DOW_* and *HR_*. No extreme values.
Order Timing	First_order median at 22 and Last_order median at 70 (Figure 24 and 25)	No outliers and all values are in a 90-day timeframe. No error-based values.
Purchase behavior	Median values close to zero across all features, with one extreme outlier across all (Figure 26 and 27)	One extreme multivariate outlier across all purchase behaviors. Matching outlier also with *total_cui_spending*.
Demographics	For *customer_age* feature, Q1-Q3 range: 23-31. Outliers above 43 (Figure 28)	Outliers result from the overrepresentation of Gen Z and Millenials in the data set. No abnormal outliers.

For the next phase, we plan to take two key actions: (1) eliminate outliers that are considered to be data errors and (2) adjust outliers that are deemed to be extreme values, but not errors, by capping them at the upper or lower boundary limits.

4. Feature Engineering

This section introduces new features aimed at enhancing the dataset's depth and quality for clustering and segmentation analysis (see Table 5).

Table 5 - Feature Engineering Summary

Feature Name	Description	Purpose	Figure Appendix
generation	Customer age is categorized into generations[3]	Segment spending patterns for targeted marketing	Figure 29
total_cui_spending	Sum of spending across cuisines	Estimate the customer's lifetime value	Figure 30
health_index	Proportion of spending on Cui_Health	Identify health-conscious customers	Figure 31
total_orders	Total number of orders	Assess customer engagement and reduce dimensionality	Figure 32
weekend_orders weekday_orders	Total orders on weekends and weekdays	Analyze ordering trends and reduce dimensionality	Figure 33 Figure 34
orders_dawn orders_morning orders_afternoon orders_evening	Distribution of orders by time of day	Understand preferences and consolidate hourly data	Figure 35 Figure 36 Figure 37 Figure 38
order_recency	Normalized recency of last order	Assess customer engagement levels	Figure 39
avg_daily_orders	Average number of daily orders	Distinguish frequent from occasional customers	Figure 40
avg_order_value	Average spending per order	Identify high-value customers	Figure 41
promo_used	Indicator for promotion use	Evaluate promotion effectiveness on spending	Figure 42
chain_orders_prop	Proportion of orders from chain restaurants	Assess brand loyalty trends	Figure 43
city	Geographic categorization of customers	Analyze regional distribution and behavior impact	Figure 44

5. Conclusion

This first part, exploratory data analysis has laid a strong foundation for understanding custom behavior at ABCDEats Inc. Through careful data cleaning, we addressed missing values while improving data integrity. The development of new features such as spending metrics and order patterns, has enriched our dataset, offering deeper insights into customer preferences and potential segmentation strategies. While some redundancy remains, we will address this in the next phase to ensure each feature contributes disting and valuable information.

3: Assuming data collection in 2024. The age distribution aligns with industry norms based on our knowledge of the food delivery sector.

Bibliographical References

[Bib.1] Frost, J. (2024) Missing Data Overview: Types, Implications & Handling .https://statisticsbyjim.com/basics/missing-data/

Appendix A

AI Statement: We use AI tools such as Grammarly to refine and correct English sentences for clarity and accuracy, and ChatGPT to enhance and improve the presentation of our plots.

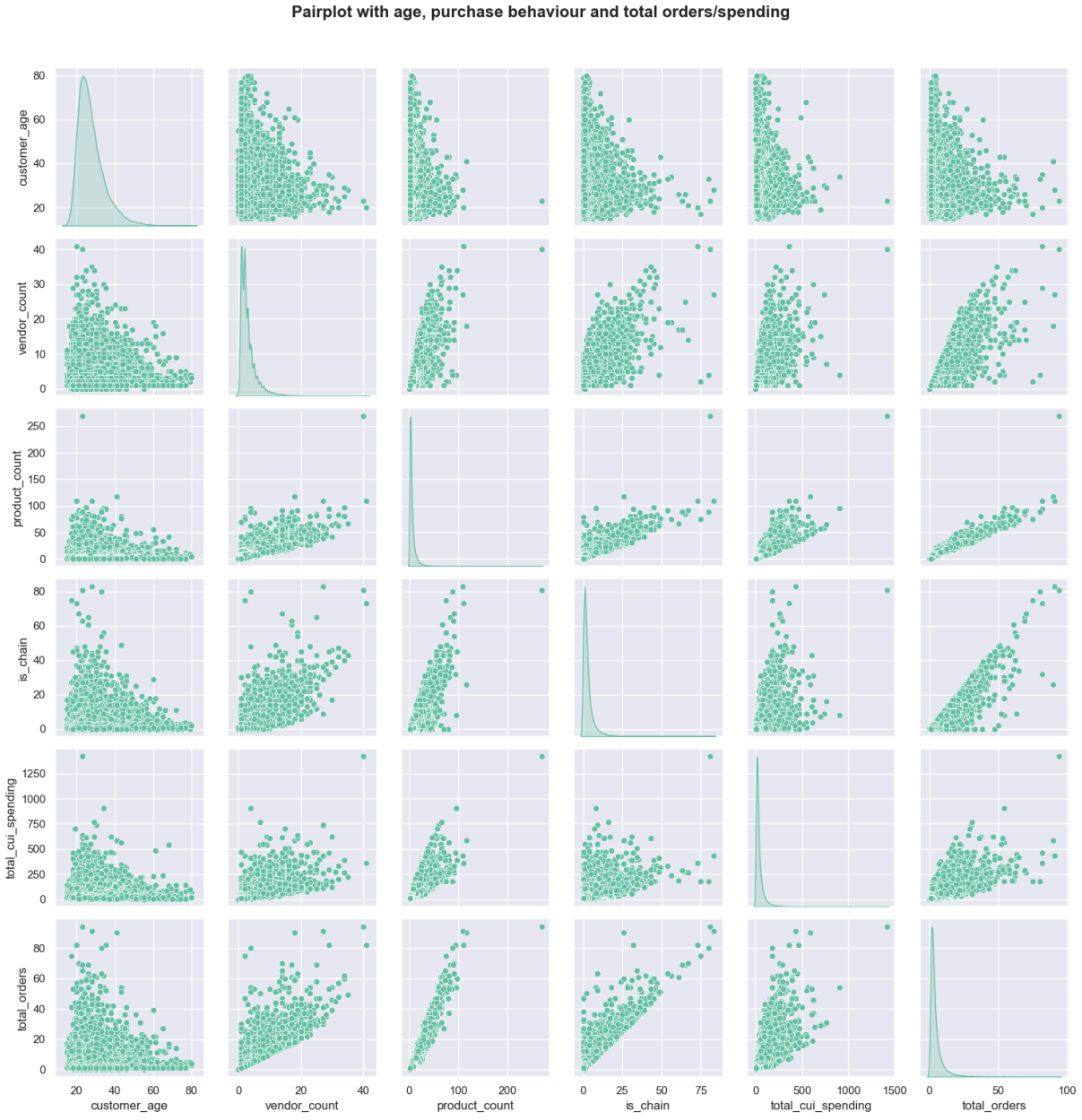

Figure 1 - Pairplot for initial dataset without cuisine, order patterns and order timing*

* The *total_cui_spending* and *total_orders* features aggregate the *CUI_* and *DOW_* variables, respectively.

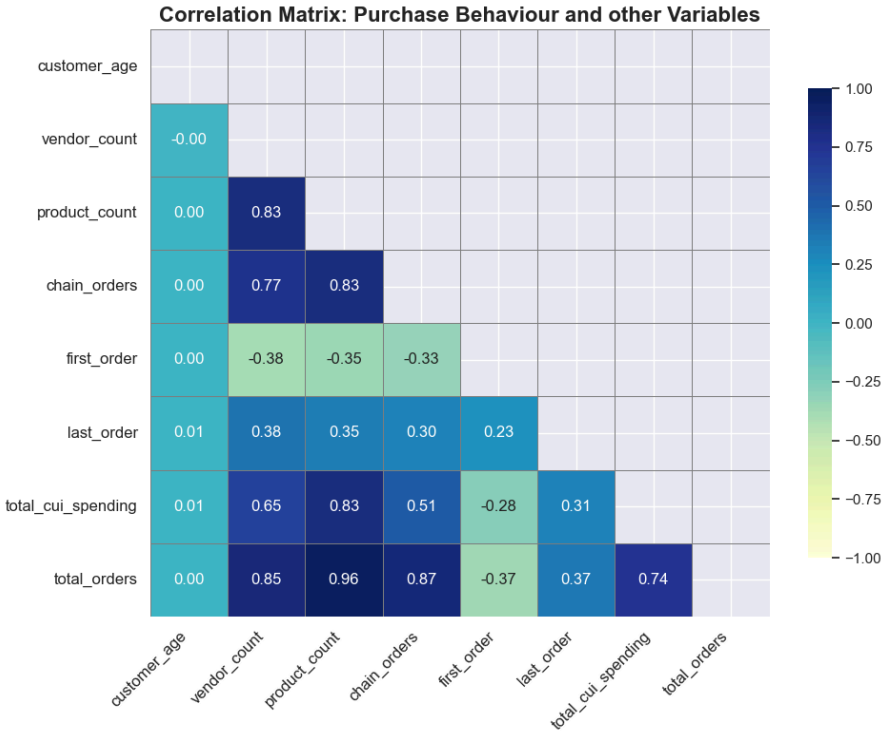

Figure 2 - Correlation Matrix for Purchase Behaviour & Other Variables*

* The *total_cui_spending* and *total_orders* features aggregate the *CUI_* and *DOW_* variables, respectively.

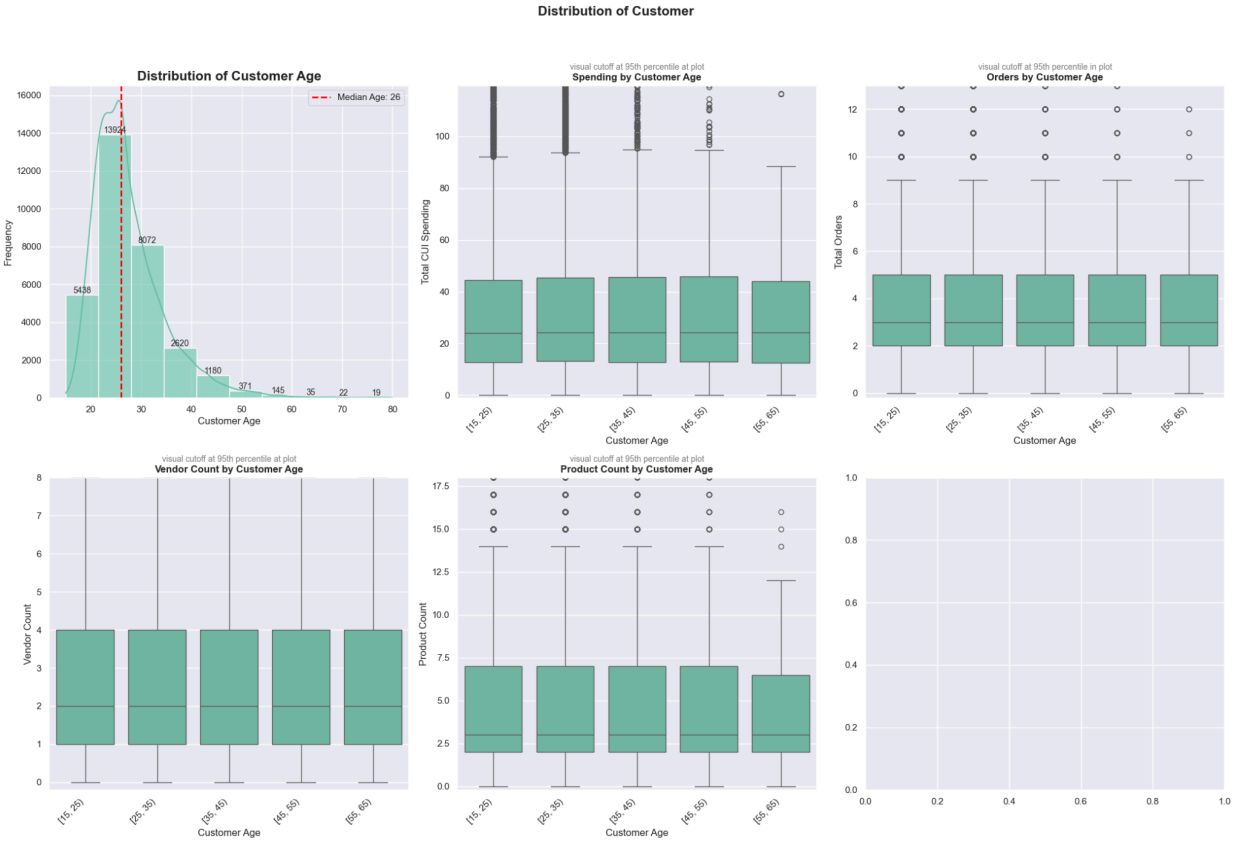

Figure 3 - Customer Age Distribution & Spending Behavior

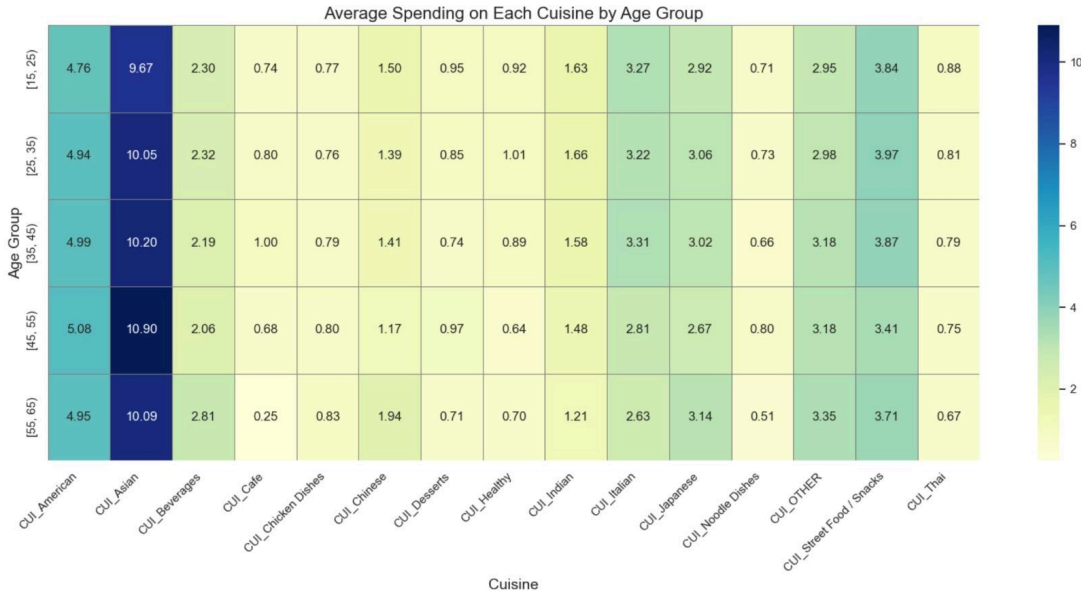

Figure 4 - Average Spending on Cuisines Types by Age Group

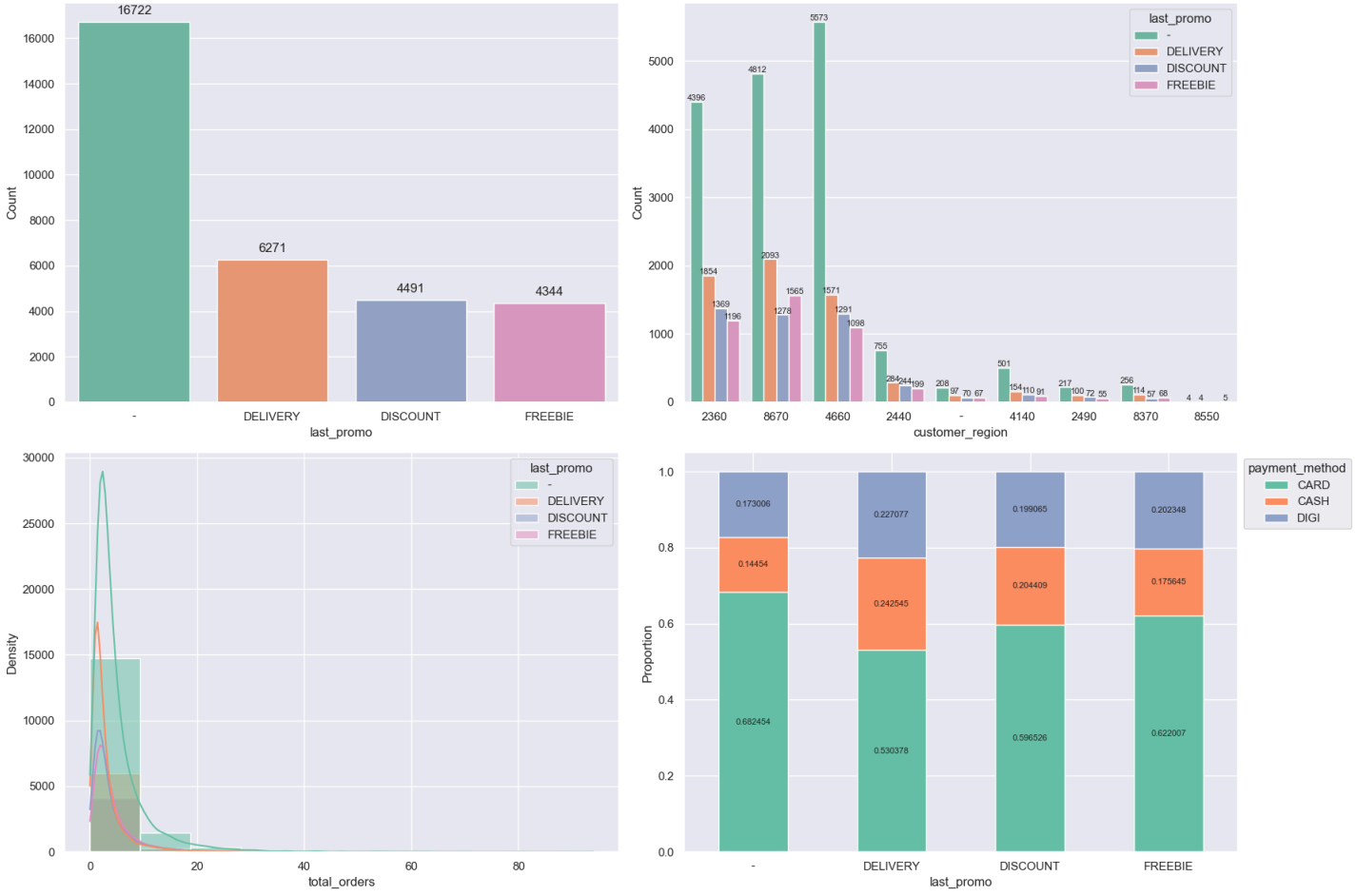

Figure 5 - Distribution of Variables by Last Promotion Category

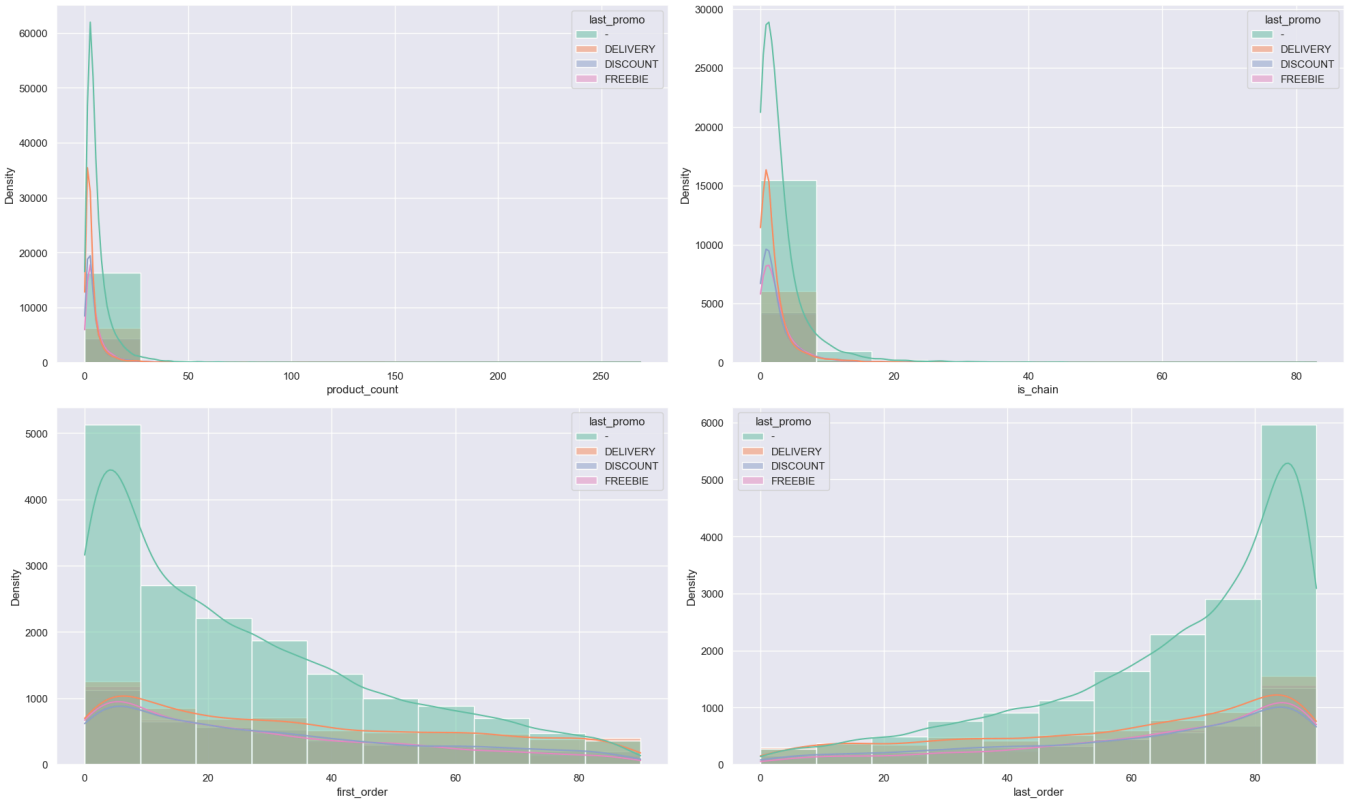

Figure 6 - Distribution of Variables by Last Promotion Category (Continuation)

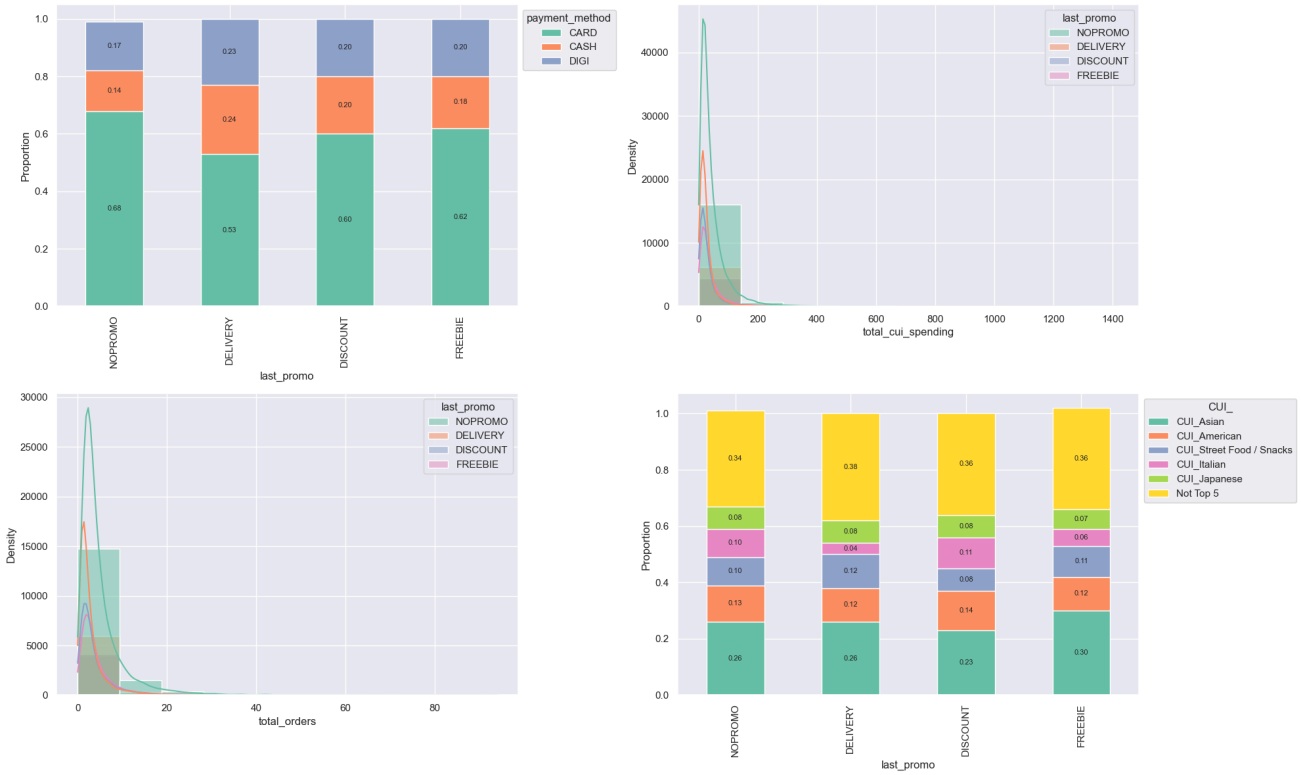

Figure 7 - Distribution of Variables by Last Promotion Category

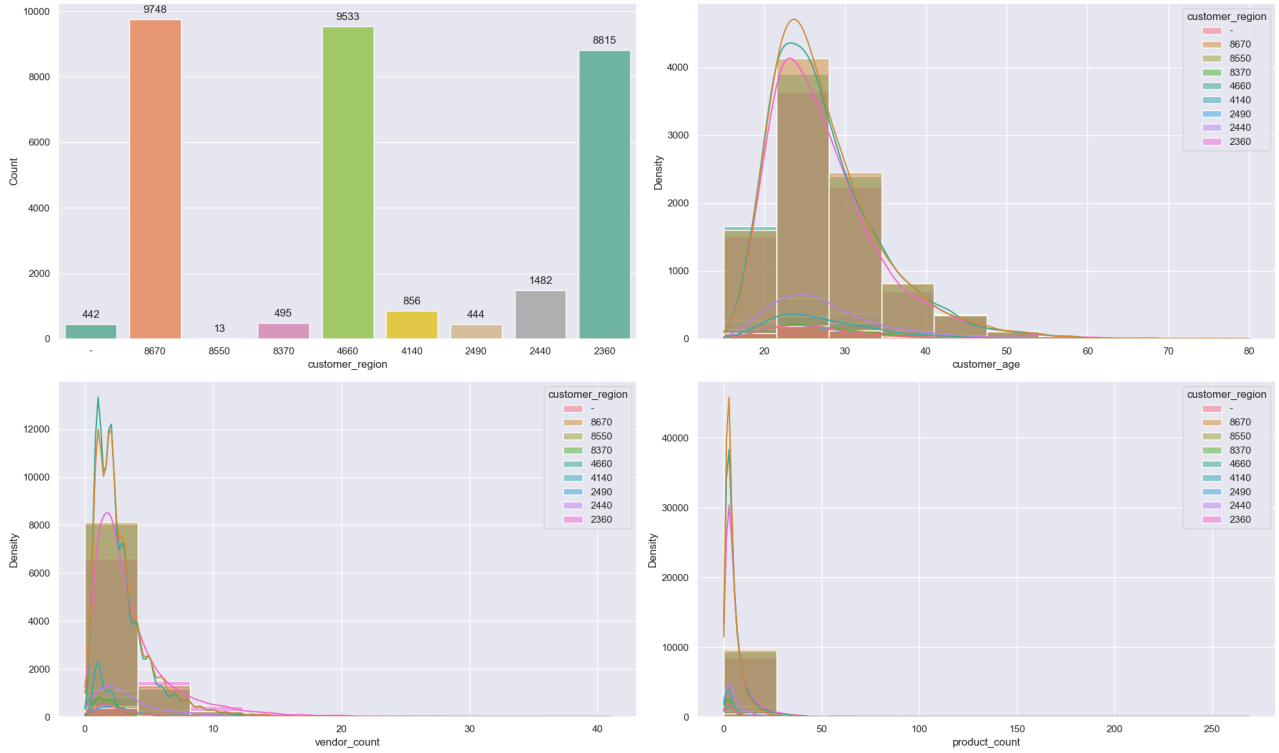

Figure 8 - Distribution of Variables by Customer Region

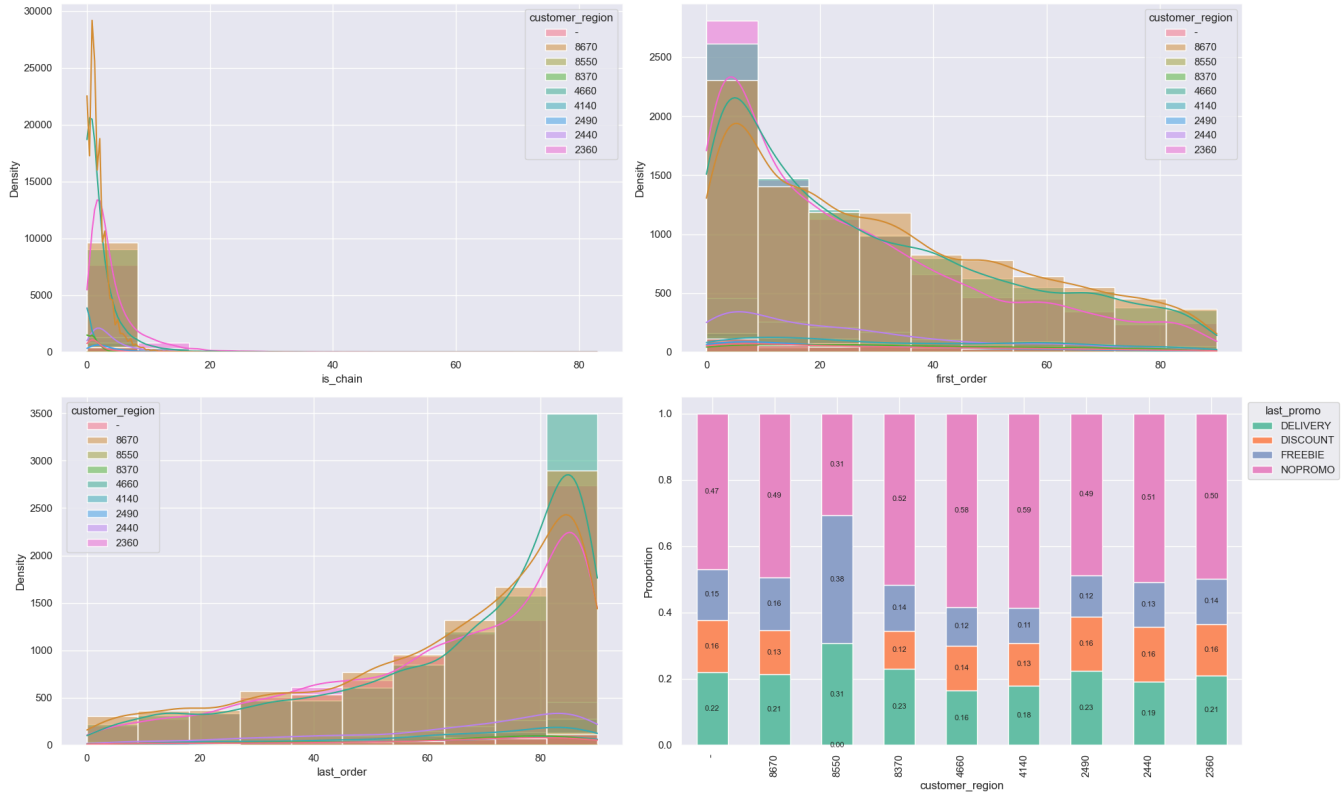

Figure 9 - Distribution of Variables by Customer Region (Continuation)

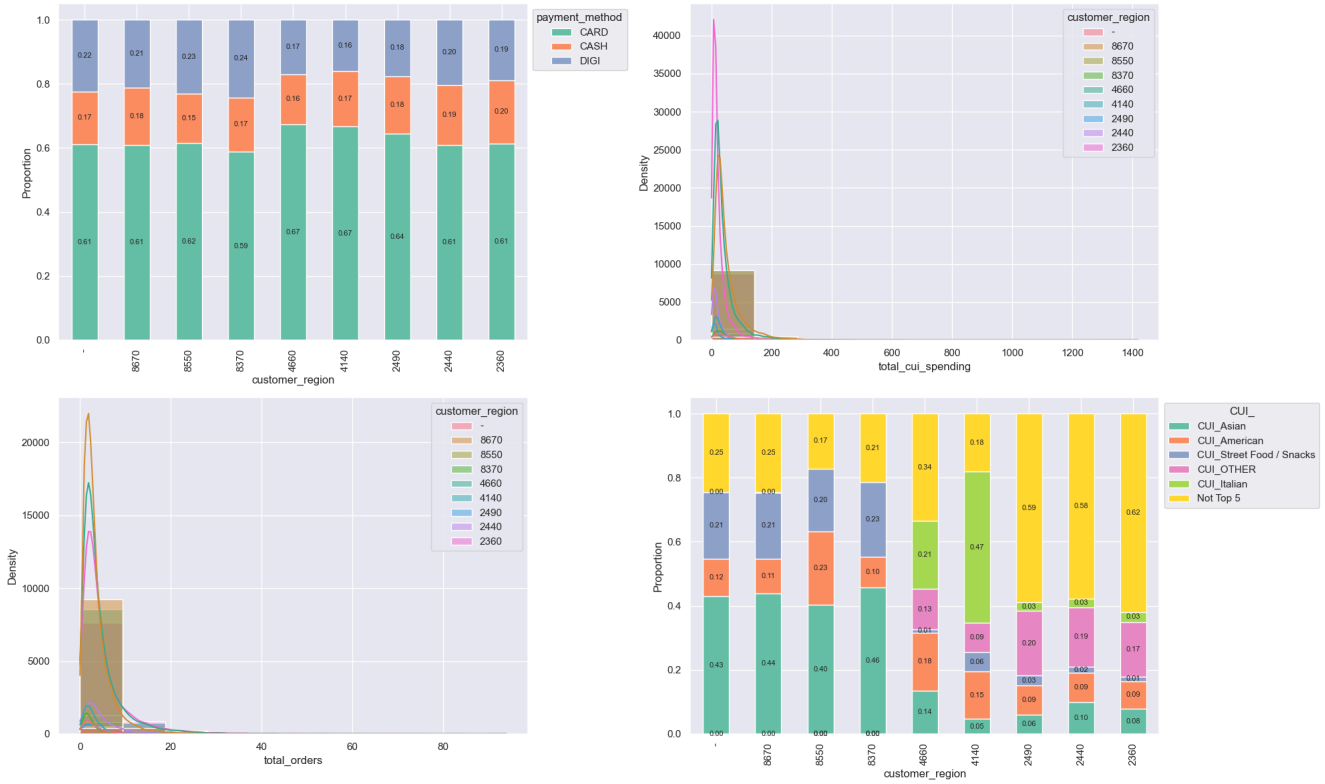

Figure 10 - Distribution of Variables by Last Promotion Category (Continuation) *

* The *total_cui_spending* and *total_orders* features aggregate the *CUI_* and *DOW_* variables, respectively.

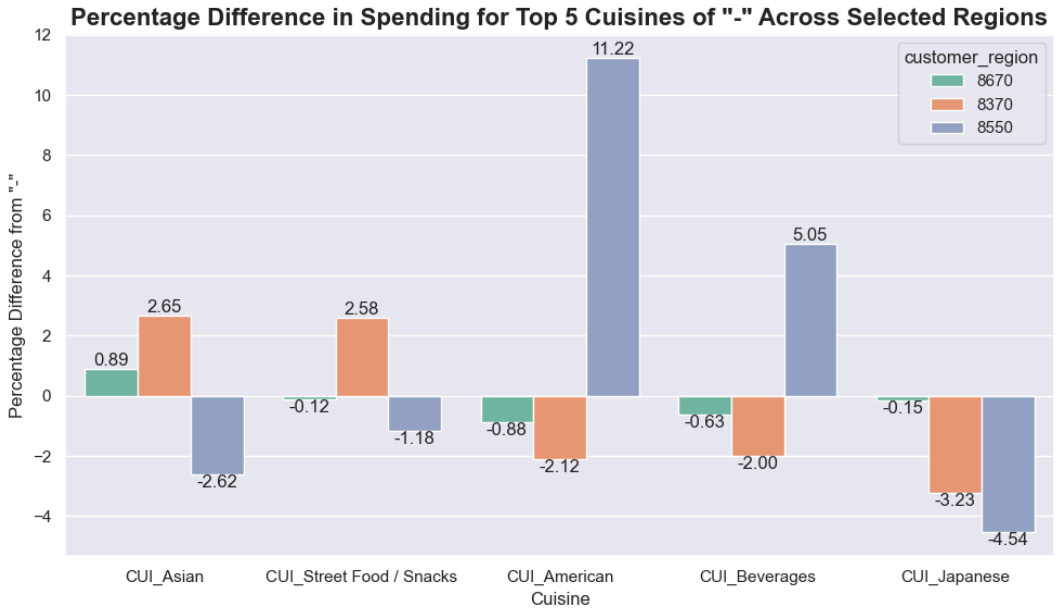

Figure 11 - Comparison of Spending Percentages for top 5 of '-' across selected regions

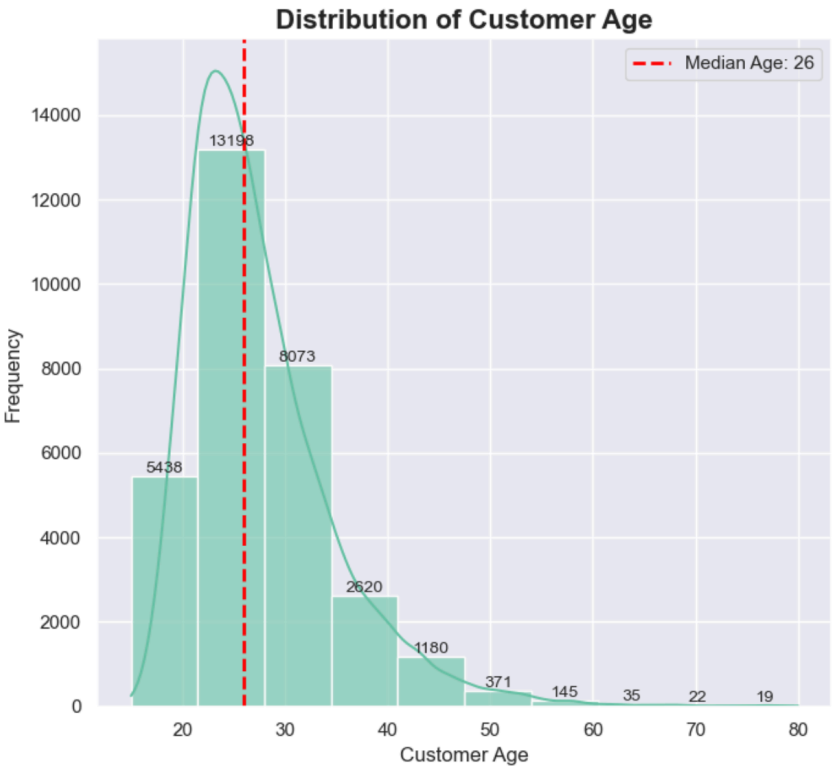

Figure 12 - Distribution of Customer Age

	count	mean	std	min	25%	50%	75%	max
customer_age	106.0	28.75	7.42	18.0	23.00	27.50	32.00	60.00
vendor_count	106.0	1.02	0.14	1.0	1.00	1.00	1.00	2.00
product_count	106.0	1.31	0.72	1.0	1.00	1.00	1.75	7.00
is_chain	106.0	0.57	0.52	0.0	0.00	1.00	1.00	2.00
first_order	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
last_order	106.0	0.00	0.00	0.0	0.00	0.00	0.00	0.00
total_cui_spending	106.0	14.23	12.41	1.0	6.14	12.08	18.84	105.55
total_orders	106.0	1.02	0.14	1.0	1.00	1.00	1.00	2.00

Figure 13 - Summary Statistics of *first_order* Missing

customer_id	1f9cd0d268	c3690a6fa1
customer_age	30.00	27.00
vendor_count	2.00	2.00
product_count	2.00	7.00
is_chain	2.00	1.00
first_order	NaN	NaN
last_order	0.00	0.00
total_cui_spending	4.08	105.55
total_orders	2.00	2.00
DOW_0	0.00	0.00
DOW_1	0.00	0.00
DOW_2	0.00	0.00
DOW_3	0.00	0.00
DOW_4	0.00	0.00
DOW_5	0.00	0.00
DOW_6	2.00	2.00

Figure 14 - Customers with *first_order* Missing and *total_orders* = 2

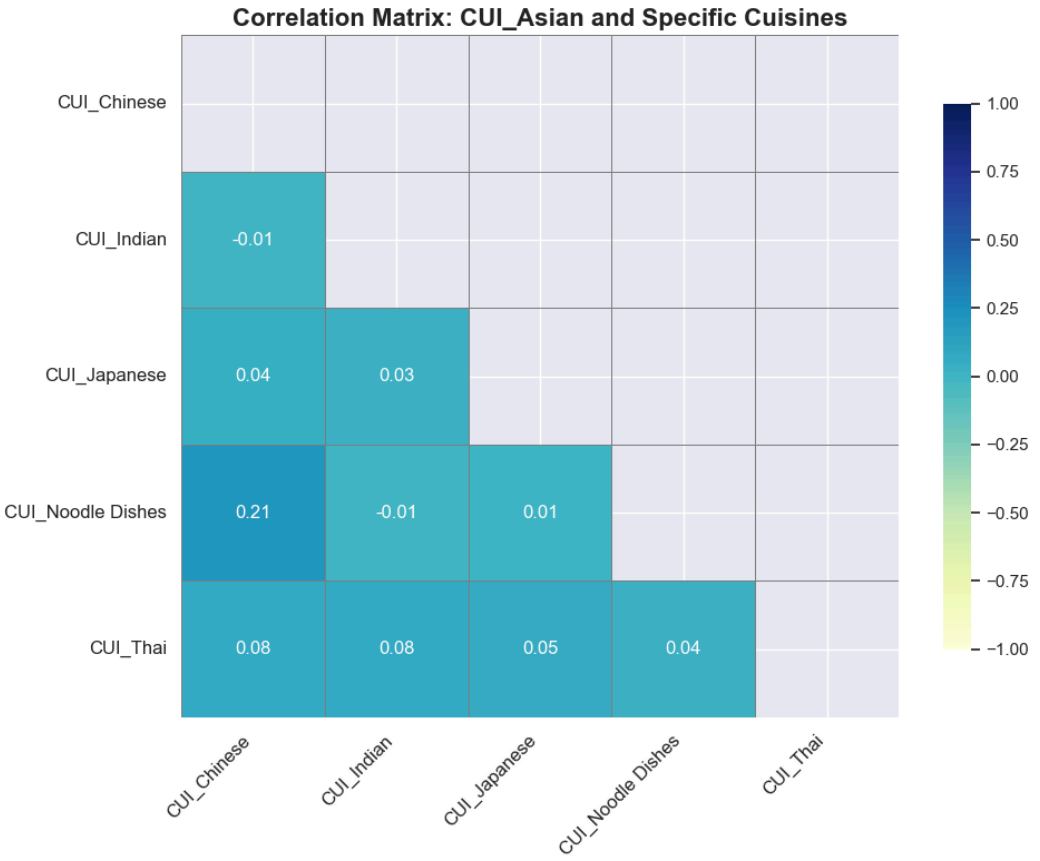

Figure 15 - Correlation Matrix: *CUI_Asian* and Specific Cuisines

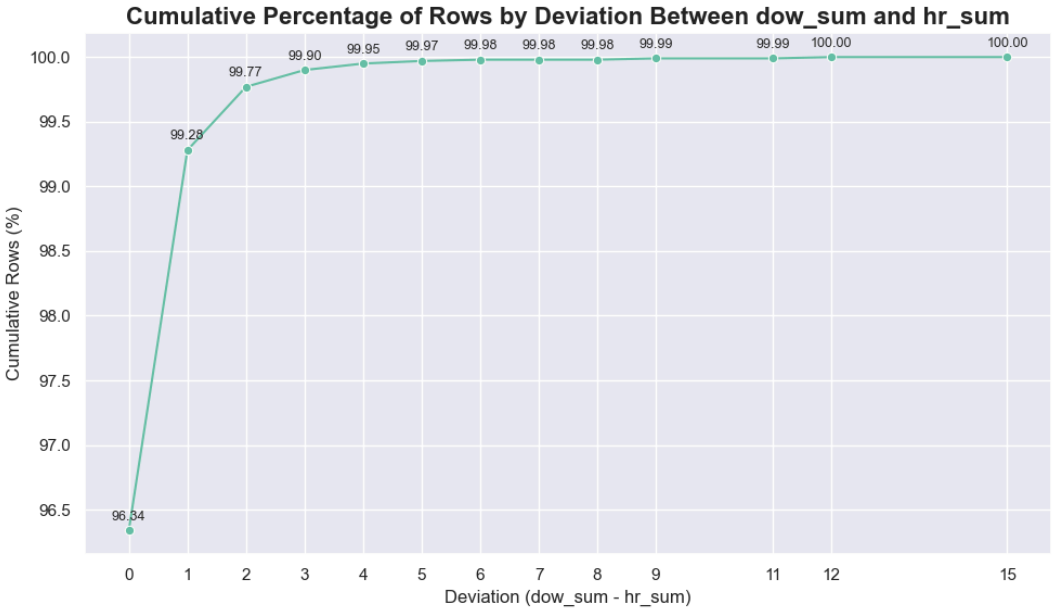

Figure 16 - Cumulative Percentage of Rows by Deviation Between *dow_sum* and *hr_sum*

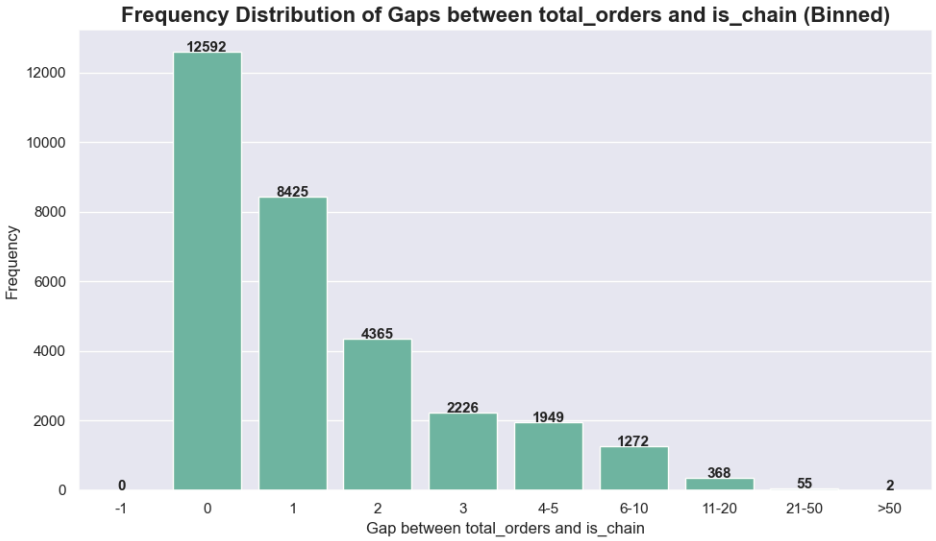

Figure 17 - Frequency Distribution of Gaps between *total_orders* and *is_chain* (Binned)*

* The *total_orders* feature aggregates the *DOW_* variables.

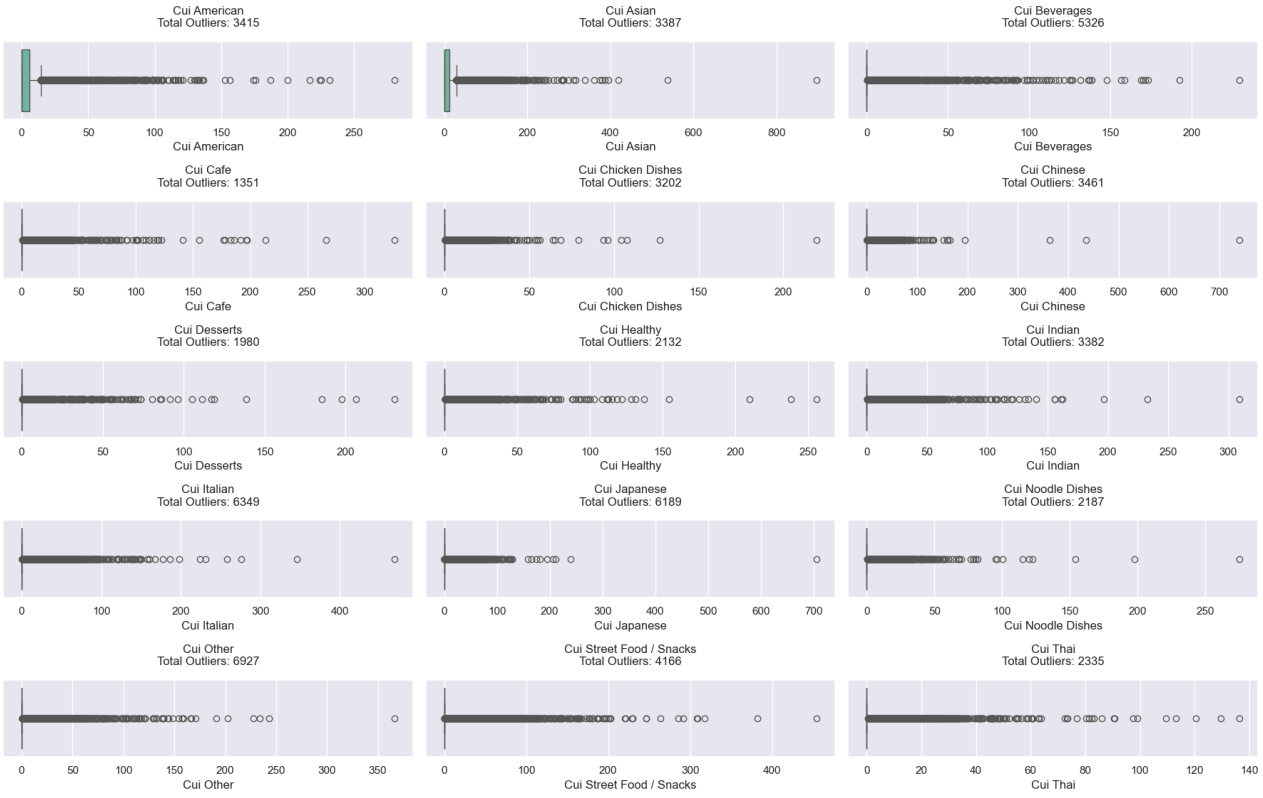

Figure 18 - Boxplots Cuisine Spendings

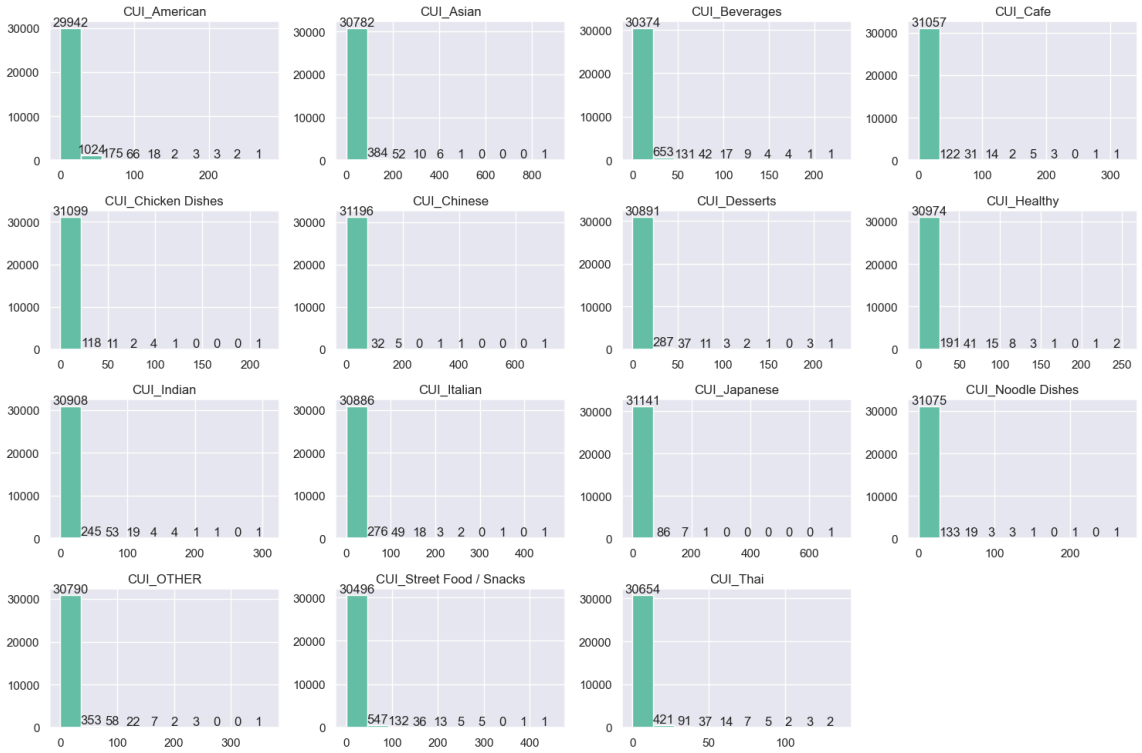

Figure 19 - Histograms Cuisine Spendings

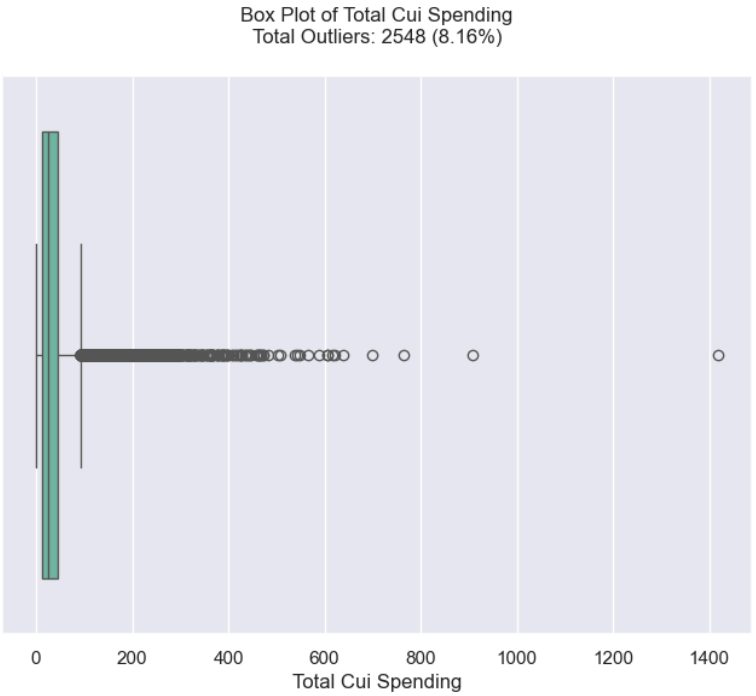

Figure 20 - Total Cuisine Spending Boxplot*

* The *total_cui_spending* feature aggregate the *CUI_* variables.

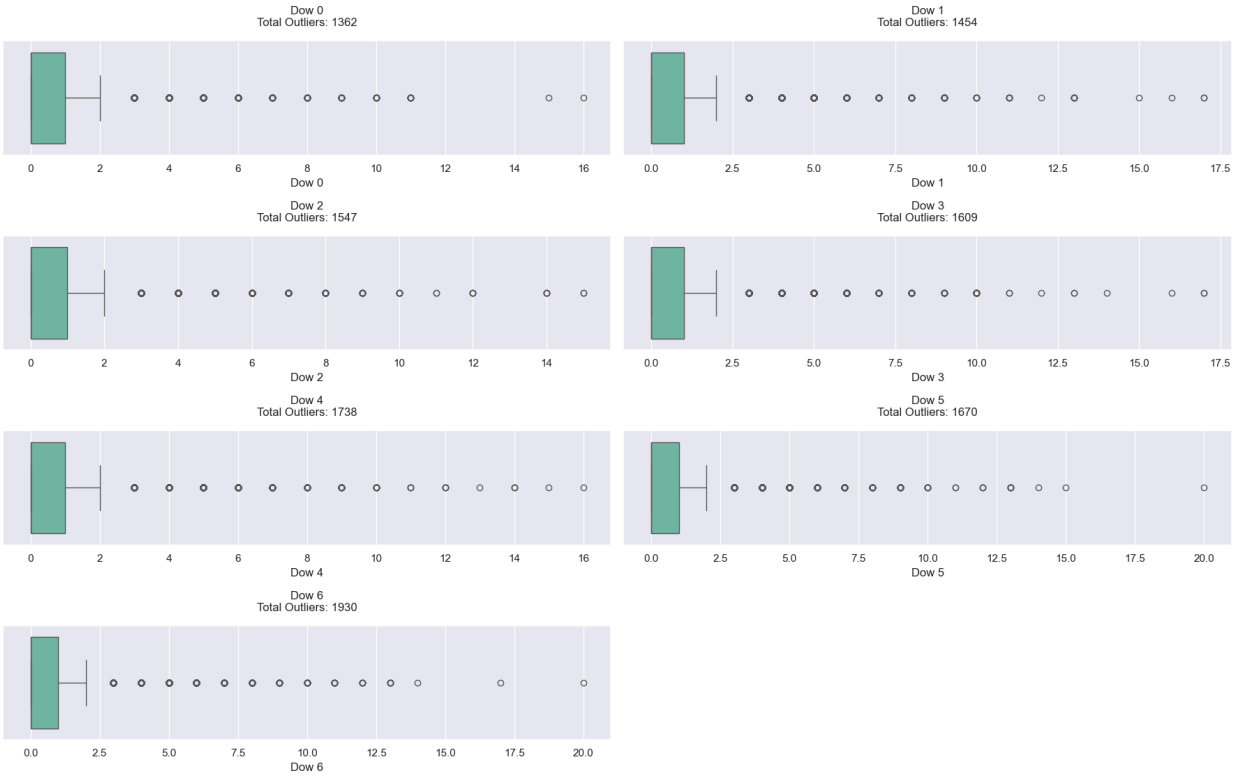

Figure 21 - Boxplots of DOW features

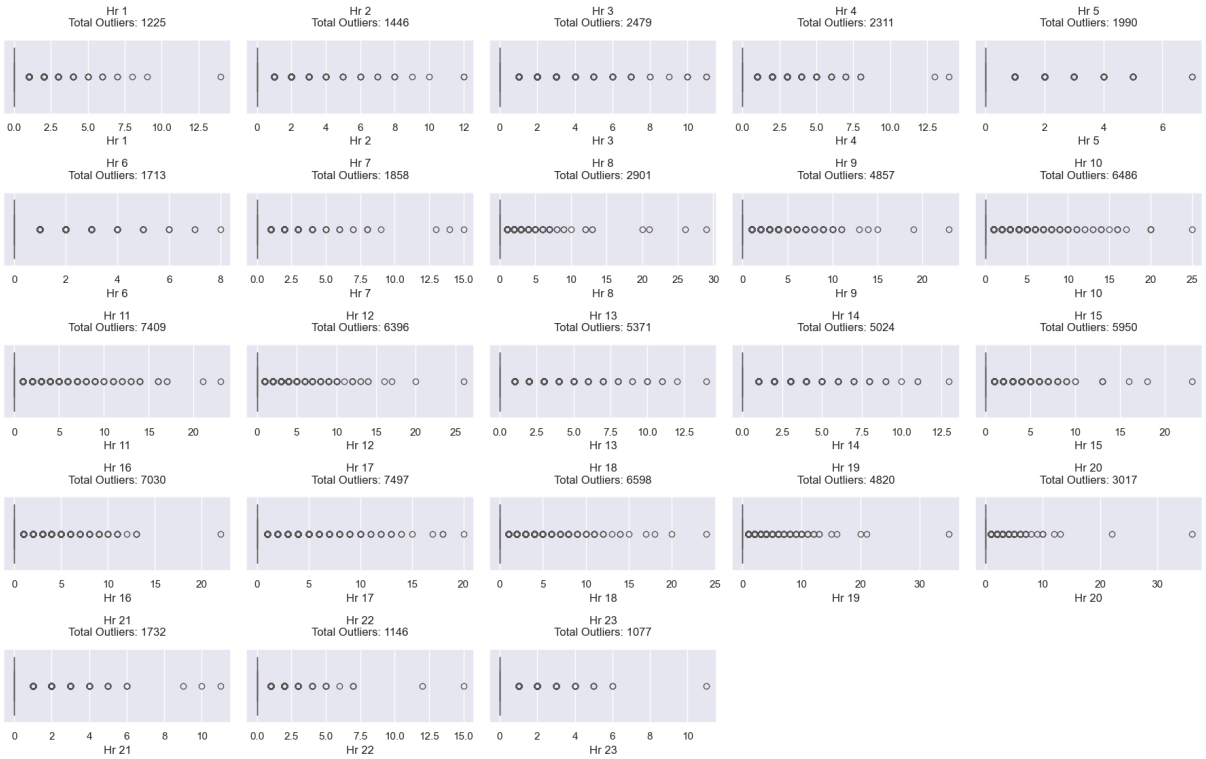

Figure 22 - Boxplots of HR features

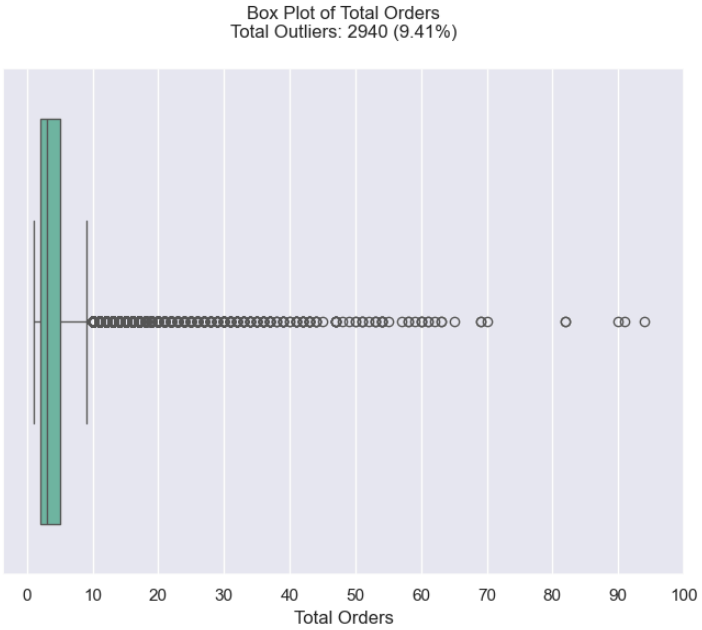

Figure 23 - Boxplot of Total Order

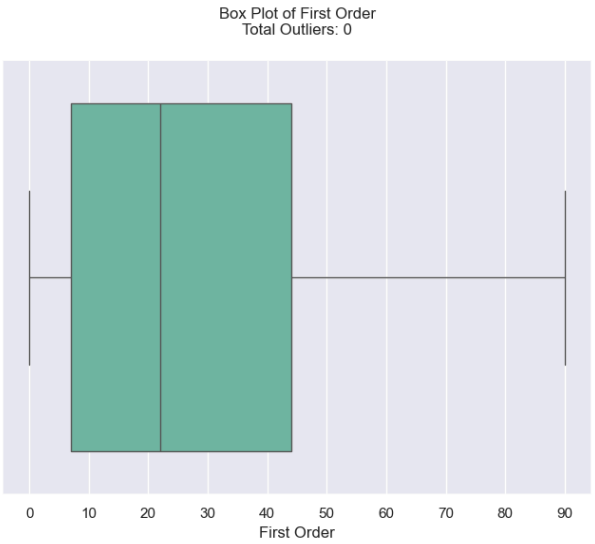

Figure 24 - First Order Boxplot

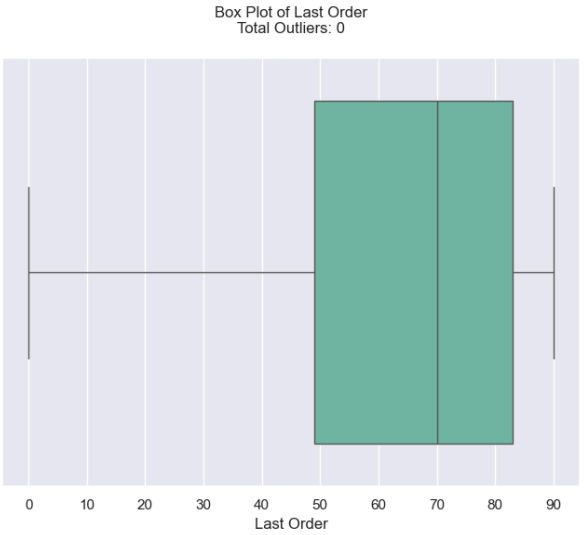

Figure 25 - Last Order Boxplot

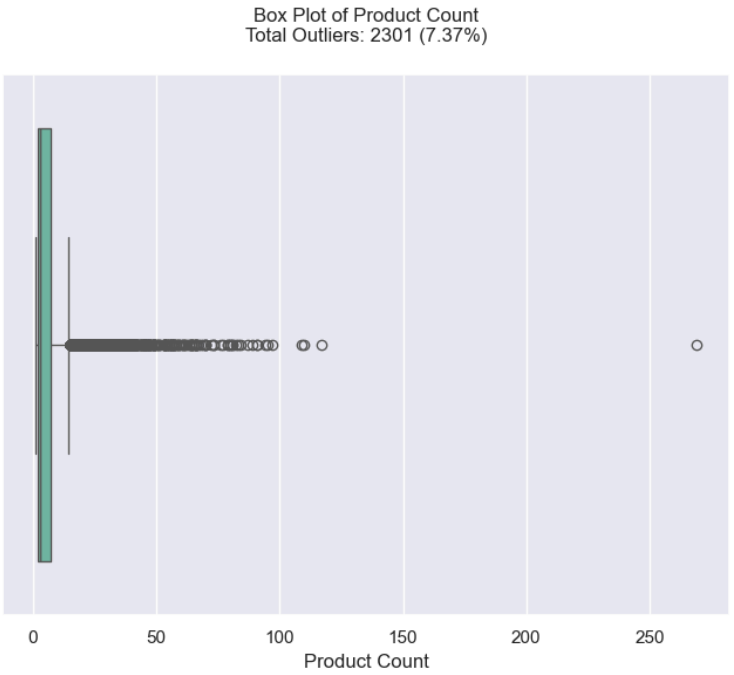

Figure 26 - Product Count Boxplot

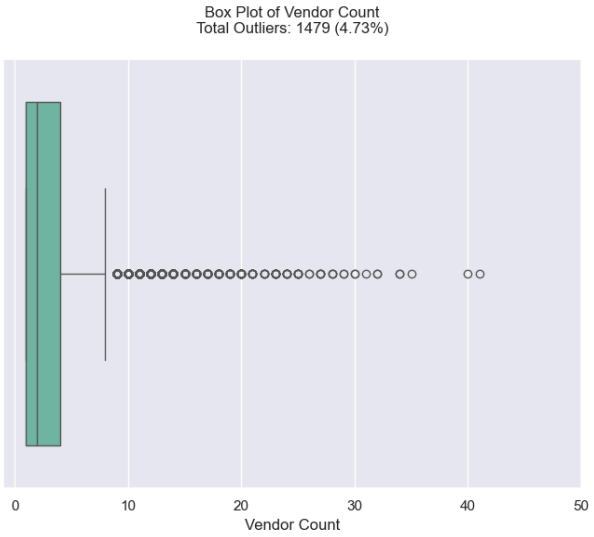

Figure 27 - Vendor Count Boxplot

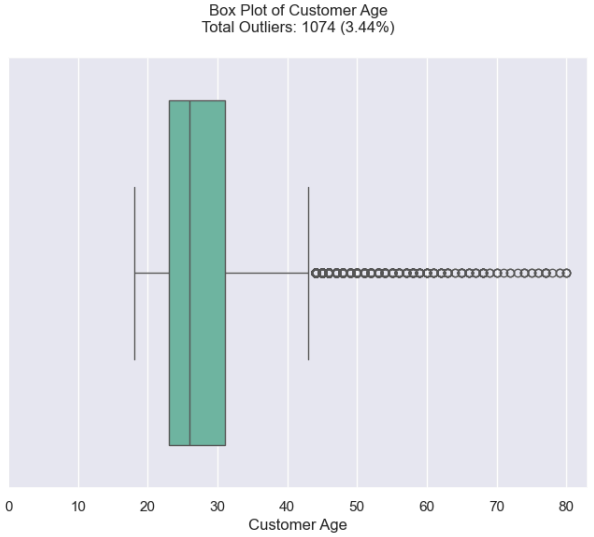

Figure 28 - Boxplot Customer Age

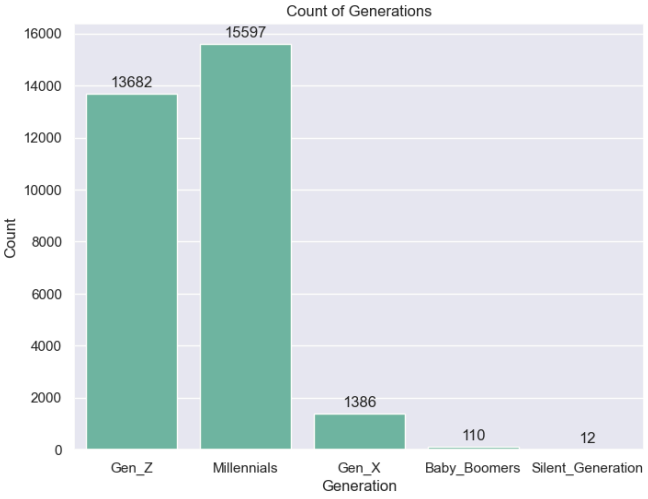

Figure 29 - Generation Frequency Distribution

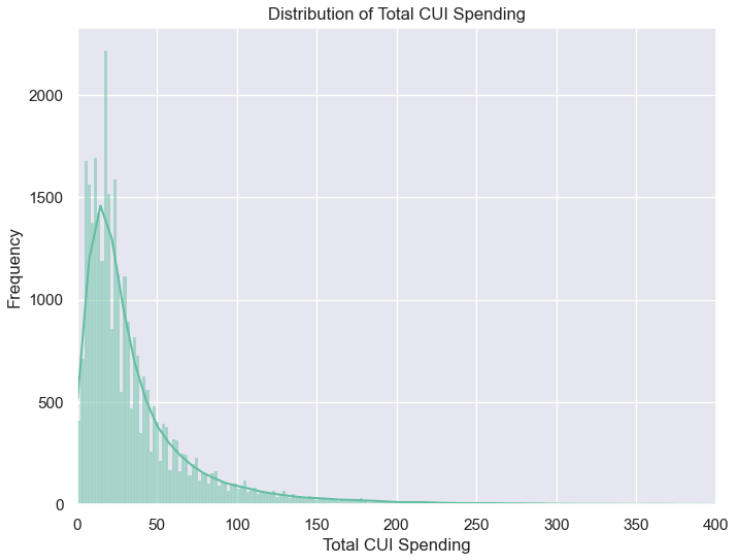

Figure 30 - Total Cuisine Spending Histogram

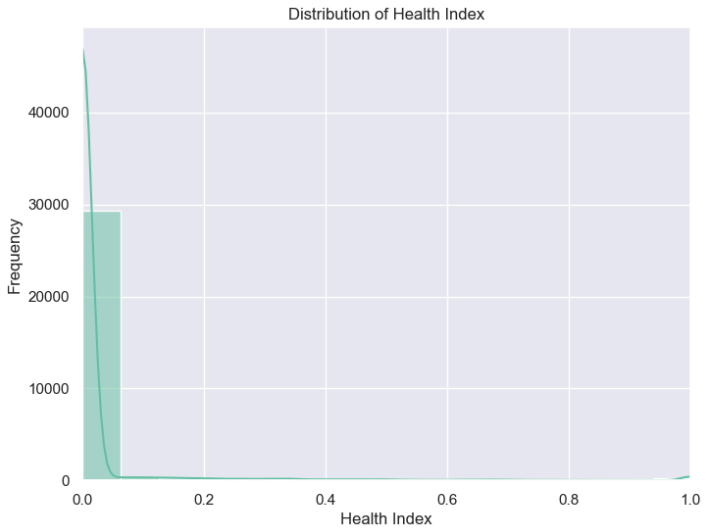

Figure 31- Health Index Boxplot

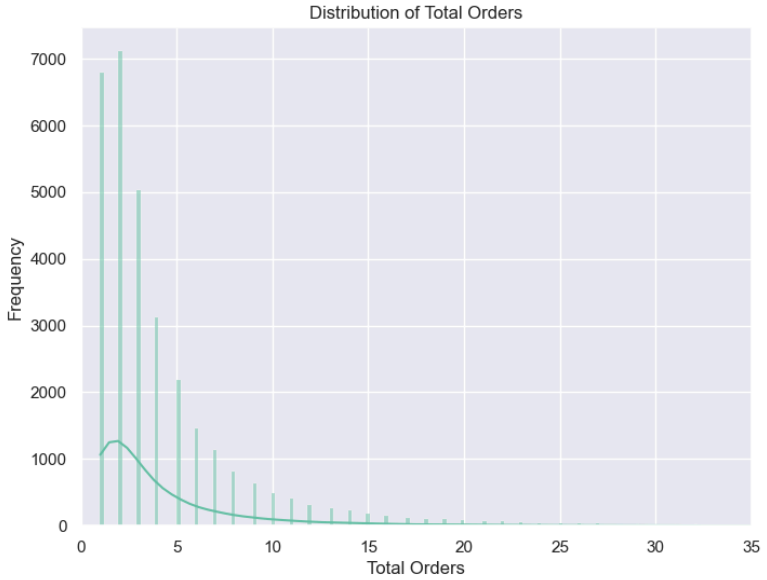

Figure 32 - Total Orders Histogram

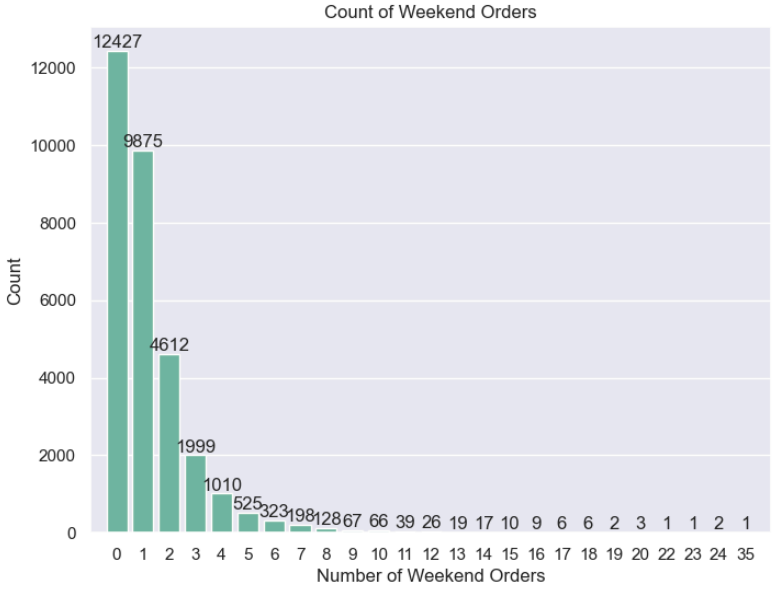

Figure 33 - Weekend Orders Histogram

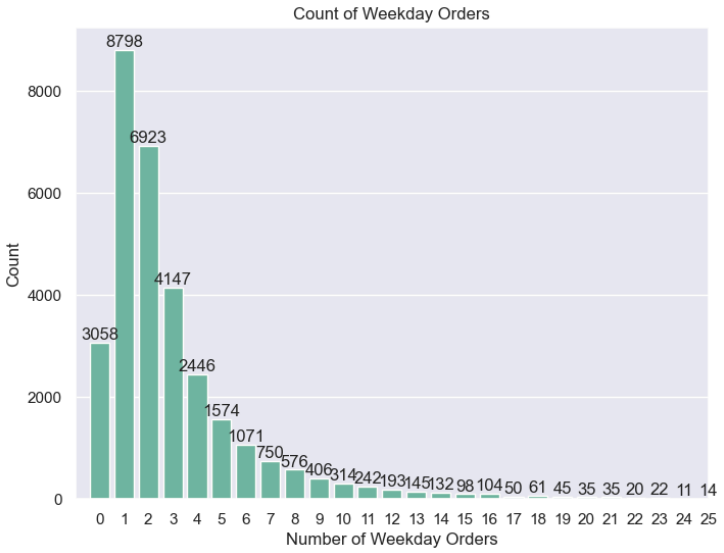

Figure 34 - Weekday Orders Histogram

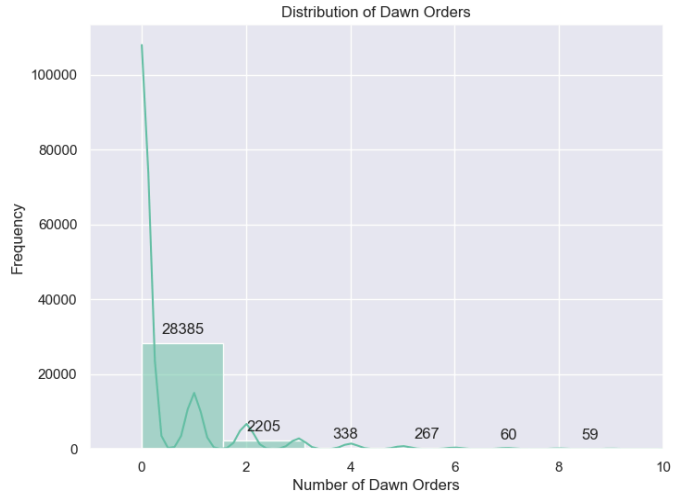

Figure 35 - Number of Dawn Orders

Figure 36 - Number of Morning Orders

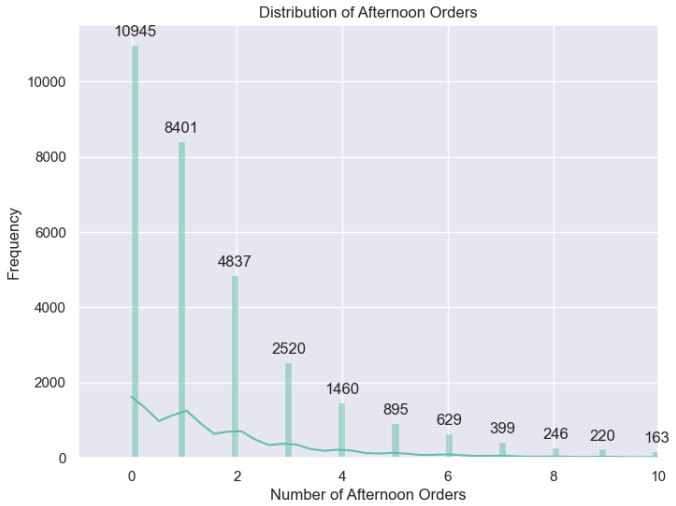

Figure 37 - Number of Afternoon Orders

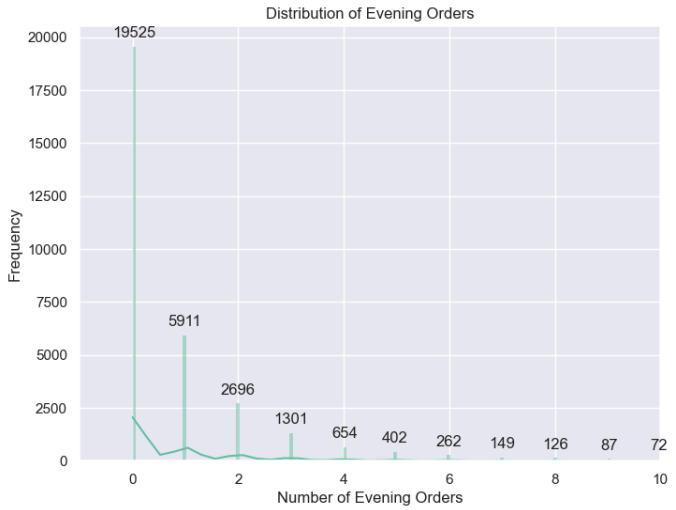

Figure 38 - Number of Evening Orders

Figure 39 - Order Recency Distribution

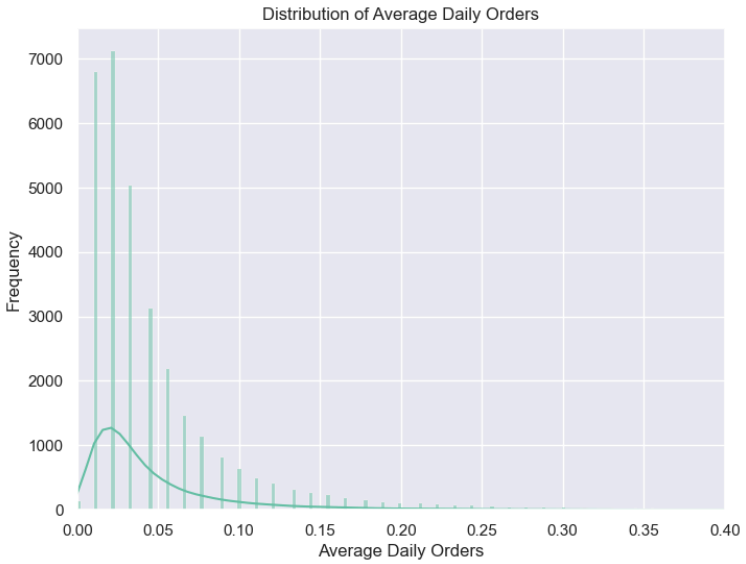

Figure 40 - Average Daily Orders Distribution

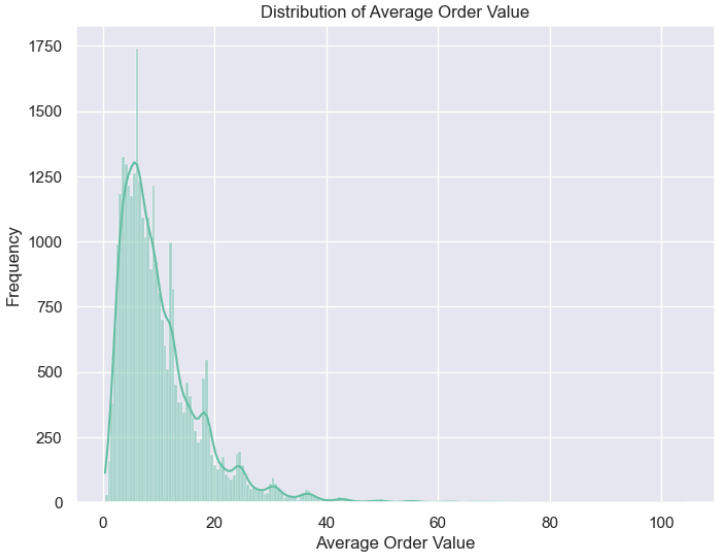

Figure 41 - Average Order Value Distribution

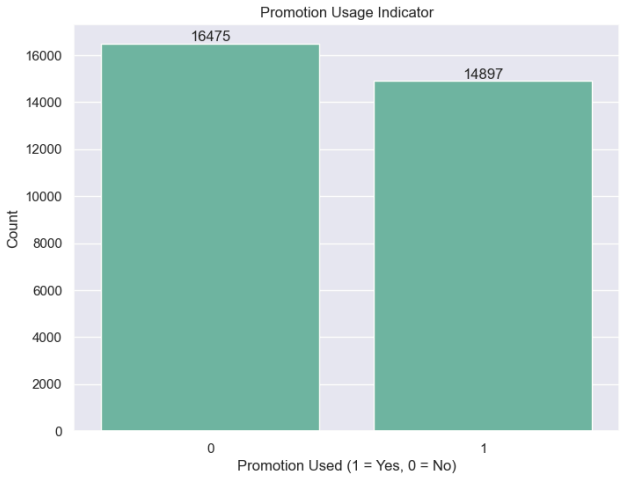

Figure 42 - Promotion Used Frequency Distribution

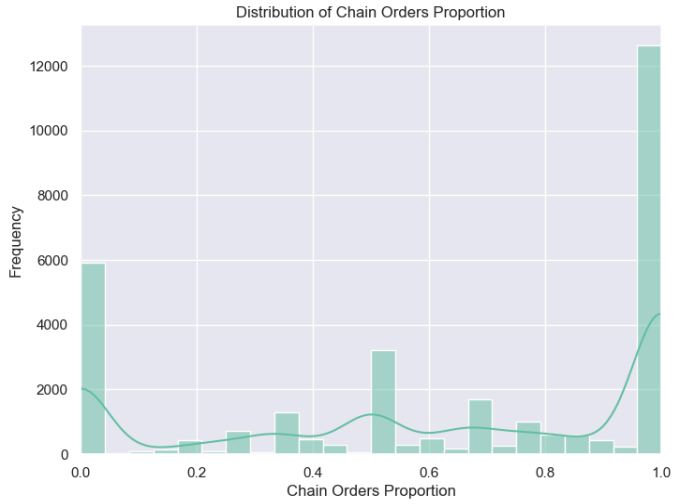

Figure 43 - Chain Orders Distribution

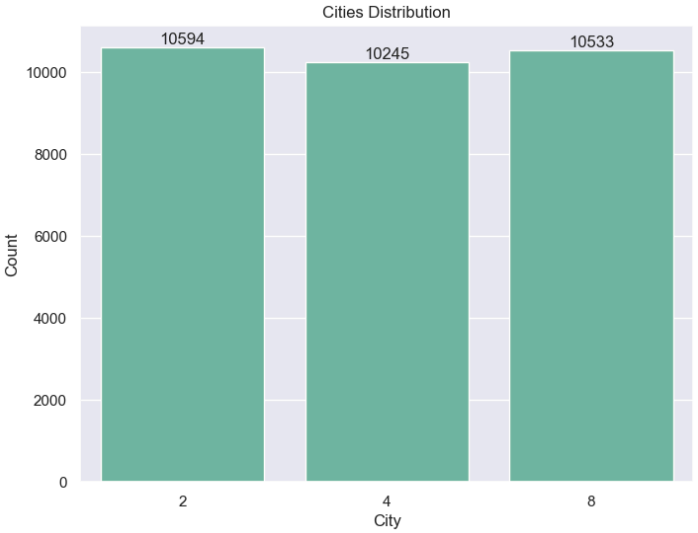

Figure 44 - City Frequency Distribution