DATA MINING PROJECT

Master in Data Science and Advanced Analytics

NOVA Information Management School

Universidade Nova de Lisboa

Comprehensive Customer Segmentation Strategy for ABCD Eats Inc.

Group 04

Beatriz Monteiro, 20240591

Luís Semedo, 20240852

Pedro Santos, 20240295

Rodrigo Miranda, 20240490

Fall/Spring Semester 2024-2025

TABLE OF CONTENTS

1. INTRODUCTION

Customer segmentation is a vital practice across consumer-driven industries, helping businesses understand consumer behavior elaborately. This understanding directly impacts revenue generation and overall growth. Beyond business benefits, segmentation provides tailored options that cater to individual interests and preferences, enhancing convenience for consumers.

As consultants for ABCDEats Inc., our objective is to analyze customer behavior and develop a data-driven strategy to support company growth. We aim to optimize marketing efforts by segmenting customers and effectively target different customer groups. Our analysis uses a dataset from ABCDEats Inc., covering demographics and customer-oriented preferences.

Before developing strategic recommendations, it's crucial to understand the dataset thoroughly. This initial analysis focuses on Exploratory Data Analysis (EDA), forming the basis for deeper insights and a comprehensive strategy. Effective data management, including cleaning and identifying variable relationships, is essential. EDA aims to uncover dataset characteristics, identify inconsistencies, anticipate future problems, and extract meaningful insights, supported by our main pillar: data.

We initialize our analysis with Feature Engineering, which is a vital step in any Machine Learning or Data Mining process. It involves manipulating and transforming raw data into features that may prove relevant for modeling purposes. According to research and industry consensus, Feature Engineering is crucial for capturing valuable information that might not be directly evident in the raw dataset. This process includes methods such as aggregation, transformation, extraction, and binning, all of which enhance our ability to build effective predictive models.

Following up, we divided this analysis into several subtopics, beginning with an initial preview of the dataset and followed by an in-depth examination of numerical and categorical variables. Data quality assessment is addressed as part of this process. We then proceed with bivariate analysis, focusing on relationships between categorical and numerical variables, supported by visualizations to substantiate our assumptions.

Overall, EDA and Feature Engineering are essential for uncovering trends and behaviors that form the foundation for further analysis. The insights gained here will be crucial for customer segmentation, predictive modeling, and supporting data-driven decisions for ABCDEats Inc.

2. EXPLORATORY DATA ANALYSIS

2.1. FEATURE ENGINEERING

In our work, we found that creating some derivations of our initial features might be interesting to capture relevant behavior or trends. Please note that although it may not be a typical procedure, we relocated this section to the start of our notebook, to further capture visualizations and analysis along the way, enhancing the relationships of our newly created features with the existing ones.

Binary Variable - "*Last_Promo***"**: An assumption that will be discussed in point 2.2.2 is that "-" (which represents the majority of cases) could be interpreted as "No Promotion Used". Therefore, for

modeling purposes, we created a binary variable to capture this behavior: 0 indicates no promotion was used, and 1 indicates a promotion was used. This allows us to globally examine the relationship between orders and this binary variable, which is potentially valuable for our model.

Binning – "*customer_age***"**: To cluster different behaviors effectively, we decided to group individuals into age categories. Different age groups and generations are likely to exhibit distinct behaviors. Our visualizations provided valuable insights, clearly demonstrating differences in patterns across various age groups.

Aggregation – "*Hours***"** and **"***Total Money Spent***"**: To capture trends more effectively, we aggregated Hours into two variables: Peak_Hours, representing high-order hours, and Off_Peak_Hours for the rest. Calculating the total volume was essential to define these variables *(Figure A1: Appendix A)*. An alternative grouping by *Meal_Times* was also considered for potential insights, with one method to be chosen later to avoid the curse of dimensionality. For *Total Money Spent*, we combined spending across cuisines into a single variable, enabling analysis of spending patterns across customer segments like age groups and regions. This will support our modeling phase, offering insights to tailor marketing strategies for specific customer segments.

Transformation - "*Average Money Spent per Product"* and **"***Number of Different Cuisines"***:** Simple transformation between the *Total Money Spent* and *Product Count* features. Captures how much money per product is being spent. Interesting patterns are unveiled when combining this with other main features. As for our *Number of Different Cuisines*, it's crucial to capture the relationships between each type of food for clustering. We also opted to analyze these relationships numerically. Instead of only considering the types of cuisines people spend money on, we're interested in the variety of cuisines they order. This becomes intriguing when combined with age groups, demographic regions, etc …

2.2. UNIVARIATE EDA

2.2.1. NUMERICAL ANALYSIS

Our main issue with the *customer_age* variable is its missing values. Once resolved, we'll convert it to an integer type to accurately represent customer ages. No other inconsistencies were found initially. However, we encountered our first issue when examining the relationship between *vendor and product counts*. Some customers had null *product count* values, suggesting they never purchased any products despite placing orders and spending money on different cuisines.

A second issue arose with the supposed binary feature *is_chain*, intended to indicate whether an order was from a chain restaurant (1) or not (0). However, we found different values present, indicating potential data errors. We suspect this feature represents the number of orders from chain restaurants, supported by its strong relationship with product count *(Figure C19: Appendix C)*, with a Pearson's correlation of 0.82. Notably, 75 missing values were found here, corresponding to cases where product_count = 0, suggesting errors in recording customers who haven't purchased anything yet have spent money. Another assumption we could make would be that *is_chain* represents the number of different chains a customer frequents, but as we can see in *Figure C18: Appendix C* the values below the y=x line would be impossible since *vendor_count* would always have to be bigger or equal to *is_chain.*

Regarding *first and last orders*, there are 106 missing values for the *first order* feature, coinciding with 106 customers who haven't placed orders since the dataset began, despite recorded spending on cuisines. All missing *first_order* values align with null *last_order* values.

Examining the distributions of these 6 first features *(Figure B1: Appendix B)* and potential outliers *(Figure B2: Appendix B)* is crucial for understanding these variables. Histograms indicate right-skewed distributions for *vendor_count* and *product_count*. The *first order* distribution declines, suggesting recent first orders, while the *last order* distribution increases, indicating recent purchases by many customers. Boxplots reveal numerous high-value outliers for *vendor_count* and *product_count*, possibly indicating a subset of highly active customers. Outliers for *customer_age* skew older, whereas *first_order* and *last_order* show more evenly distributed ranges. Managing outliers, especially for *vendor_count* and *product_count*, may contribute to a more efficient analysis.

As for our *Cuisines (Figure B3: Appendix B)* and *Days of the Week (Figure B5: Appendix B)* distributions, *Cuisines* show a heavily right-sided skew, with a large concentration near the zero value, which makes sense, since most customers are not spending on every single type of cuisine, instead, they make choices. This same pattern can be recorded when observing *Days of the Week*. Additionally, our boxplots *(Figure B4: Appendix B) and (Figure B6: Appendix B),* indicate a predominant presence of outliers, which will need to be handled.

An important issue is the *Hours* feature, particularly *HR_0* where no values are present. By analyzing order volumes per hour, especially at 1 AM and 11 PM, we can see that midnight (HR_0) should follow a similar pattern *(Figure A1: Appendix A)*. Boxplots and histograms exhibit similar patterns to previous variables, showing significant outliers and skewness. As mentioned earlier, we computed *peak* and *off-peak hours* to identify trends and patterns. Although *peak hours* span from 10 AM to 12 PM and 3 PM to 6 PM, it still manages to capture the majority of order volume *(Figure B7: Appendix B)*! Please note that histograms and boxplots for newly created features were elaborated in *Figure B8: Appendix B and Figure B9: Appendix B*. Since some of these derive from existing features, skewness and outliers remain, except for *different_CUI* and *Average Money spent per Product*.

2.2.2. Categorical Analysis

Regarding the categorical variables, we can make two distinct assumptions about the *customer_region* variable due to the presence of the value "-". According to the project description, the data refers to three separate cities. Initially, we identified a logical pattern suggesting that each city consists of three unique regions, as indicated by the notation used for each. Specifically, City 2 includes regions "*2360*", "*2440*", and "*2490*", City 4 includes regions "-", "*4140*", and "*4660*", and City 8 includes regions "*8370*", "*8550*", and "*8670*". This approach reveals that each city has one region that accounts for a significant proportion of the total number of customers. For instance, region "*2360*" in City 2 contains 27.7% of the customer base, region "*4660*" in City 4 accounts for 30%, and region "*8670*" in City 8 also holds a substantial share of customers. Alternatively, we can assume that region "-" belongs to City 8, as it displays the following: the *average spent by product*, the proportion of *money spent* on the top 7 cuisines, and the distribution of orders by the time of day are very similar to those of the other regions within city 8, as we can see in the *Figure C5: Appendix C, Figure C6: Appendix C and Figure C17: Appendix C*.

We can identify a relationship between the variables *last_promo* and *promo_used*, where 52% of customers possess the value "-". This variable includes valid values such as "*DELIVERY*", "*DISCOUNT*", and "*FREEBIE*", which we consider legitimate promotional codes. On the other hand, the value "-" likely represents customers who have never used a promotion, given its significant presence in the variable and the nature of the business at hand *Figure B10: Appendix B*.

There were no inherent problems with our *payment_method* feature. A large proportion of our values, totalling more than 20,000 entries, represent "*CARD*" payments *Figure B11: Appendix B***.** Subsequently, our *age_group* variable provides a much clearer understanding. By plotting the distribution of customers across 10-year age groups, we can see that the majority of customers fall between the ages of 15 and 35 *Figure B12: Appendix B*.

2.3. Bivariate EDA

2.3.1. Numeric - Numeric

This section includes two main plots displaying correlations among our numeric features. To simplify, we visualized only the first eight unaggregated features *(Figure C1: Appendix C1)*. Each scatterplot, excluding diagonal histograms of feature distributions, helps identify stronger or weaker linear trends between features. For instance, p*roduct count* shows a positive trend with *total money spent*, indicating higher spending as *product count* increases. Feature-specific observations can also be made. There appears to be an inverse relationship between *customer age* and *product count,* suggesting that older customers tend to buy fewer products compared to younger ones. Additionally, we may observe cluster patterns within the data, represented by distinct groups of points, which could imply the existence of different subgroups within our dataset. Moreover, there is a strong relationship between the *first_order* and *last_order* variables. The triangular-shaped plot indicates that as the value of *first_order* increases, the range of possible *last_order* values shrinks. This suggests a natural dependency, a *last order* can't occur before the *first order* has taken place, hence forming the triangular pattern we observe.

The second, larger plot (*refer to subchapter 2.3.1 Numeric-Numeric Correlations in our notebook for more details*) displays all numeric feature correlations using Spearman's method, due to non-normal distributions *(Figure B1*; *Figure B3*; *Figure B5*: *Appendix B)*. This plot helps us observe multicollinearity, which we may want to address in modeling. Strong correlations are shown in darker blue (close to 1), while lighter colors indicate weaker correlations *(Figure C2: Appendix C)*.

2.3.2. Numeric - Categorical

Analyzing the *total money spent* by *customer region* in *(Figure C3: Appendix C)*, we can see that the regions "*8670*", "*4660*", and "*2360*" are the ones with bigger values, respectively, with a much bigger difference in comparison to the other regions, and consequently, a much higher number of orders too *(Figure C4: Appendix C)*, which can be mainly caused by the larger population size *(Figure B13: Appendix B)*. Curiously, the region with the most orders is "*2360*", followed by "*8670*" and then "*4660*", which means that "*2360*" will have a much smaller amount of *money spent per product* on average *(Figure C5: Appendix C)*. This could either indicate that regions with more orders but less money spent, are spending on products with a lower price, or they're making orders in smaller quantities, or even that they use more promotions. To try to analyze this, we can refer to *(Figure C6: Appendix C)*, which indicates the proportion of money each region has spent by cuisine type. Curiously, the regions with

higher *average money spent per product* all have a larger proportion of *money spent* on Asian cuisine, while the ones with smaller *average money spent per product* have a larger proportion of money spent on "*beverages"* cuisine type. Looking at the *proportion of promotion usage on the last order by customer region (Figure C7: Appendix C)*, we observe that the *promotion usage* on the *last order* is fairly balanced, except in region "*8550*", however, we won't draw conclusions about promotion usage, as we lack data on its use in orders other than the last one.

Now, looking at the different *age groups (Figure C8: Appendix C)*, we can observe that those who spend the most are by far those between 15 and 35 years, which was expected when we looked at our age distribution *(Figure B2: Appendix B)* and concluded that most of our clients are in these *age groups*. Analyzing the average spending per person across *different cuisine types* by *age groups (Figure C9: Appendix C)*, we can see that the highest average spending is on Asian cuisine, which makes sense if we refer to the total money spent on each cuisine *(Figure C10: Appendix C)*, since Asian cuisine has by far the highest amount of money spent and, most likely, the highest number of orders, as it encompasses many cuisine types.

We can also observe different behaviors by customer regions and age groups when analyzing the hours. For example, we see that regions in City 8 and the unsigned region have a higher proportion of orders in the morning and at night, while regions in City 4 show more orders in the afternoon, and regions in City 2 have a well-distributed proportion of orders between morning and dinner time *(Figure C11: Appendix C)*. Finally, we see that most age groups behave similarly during the day, except for the 66-75 group, which does not order much during dinner or at night *(Figure C12: Appendix C).*

2.3.3. CATEGORICAL - CATEGORICAL

Our preliminary analysis in this section focused on calculating the statistical significance between our categorical features. Please note that missing values and outliers have not been removed, which could affect the results. We used the Chi-square test with a significance level of 5%. If the p-value is below 0.05, it suggests that the two variables are not independent, meaning there is a statistical association between them. For more details on the specific variables, please refer to our notebook.

We focused on analyzing *customer regions* by plotting them with other features to gain insights. A clear pattern emerged across all regions, showing a strong preference for card payments *(Figure C13: Appendix C)*. This makes sense, as most orders are likely made through an app where credit card information is already saved, making card payments the most convenient option. Additionally, this rationale is supported by our *age groups* also showing a strong preference for paying by card, although some older customers occasionally prefer paying with cash *(Figure C14: Appendix C)*.

When looking at *promotion usage*, we found an interesting outlier. Region "*8550"* stands out with a significantly higher proportion of *last orders* made using promotions, while other regions show a more balanced distribution *(Figure C15: Appendix C)*. This difference could be attributed to the demographics of the region. Region "*8550*" has a large number of younger customers, primarily between the ages of 15 to 35 *(Figure C16: Appendix C)***.** These younger customers are likely more on par with new promotions and eager to take advantage of them. Please note that these visualizations are based on proportions, and values that equal zero represent a very low number rather than a true null value. Overall, these observations suggest that age and convenience play a role in payment and *promotion usage* trends across regions.

APPENDIX A (FEATURE ENGINEERING)

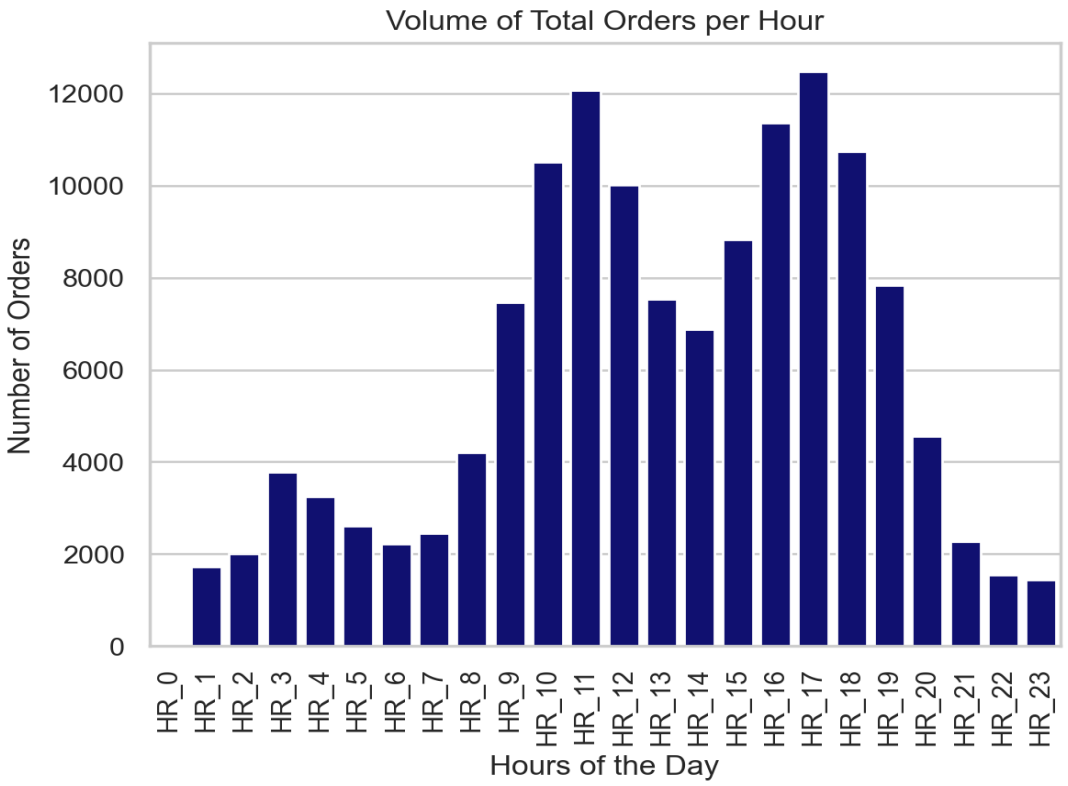

Figure A1: Comprehension of Total Volume of Orders per Hours

APPENDIX B (UNIVARIATE EDA)

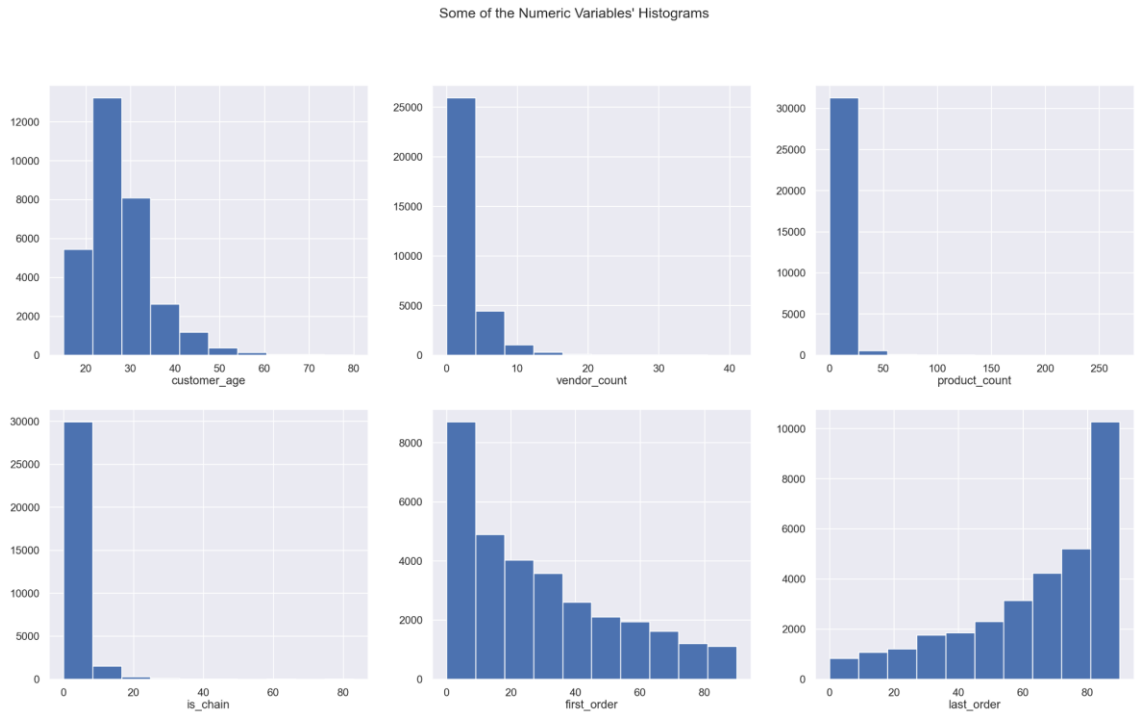

Figure B1: Some of the Numeric Variables' Histograms

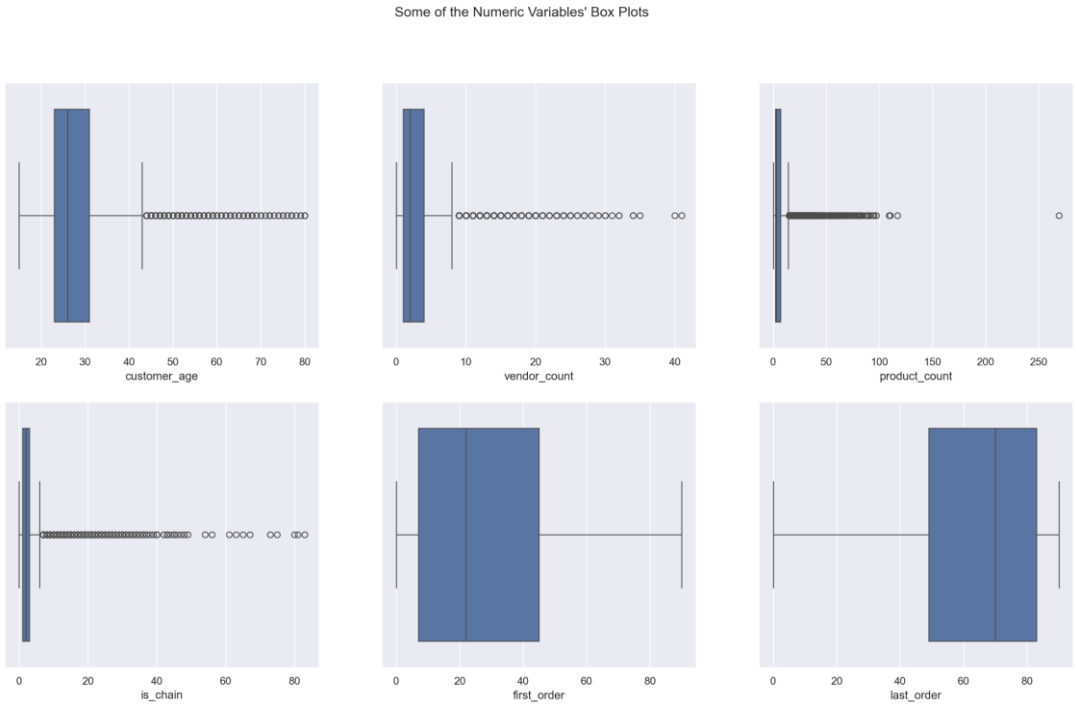

Figure B2: Some of the Numeric Variables' Boxplot

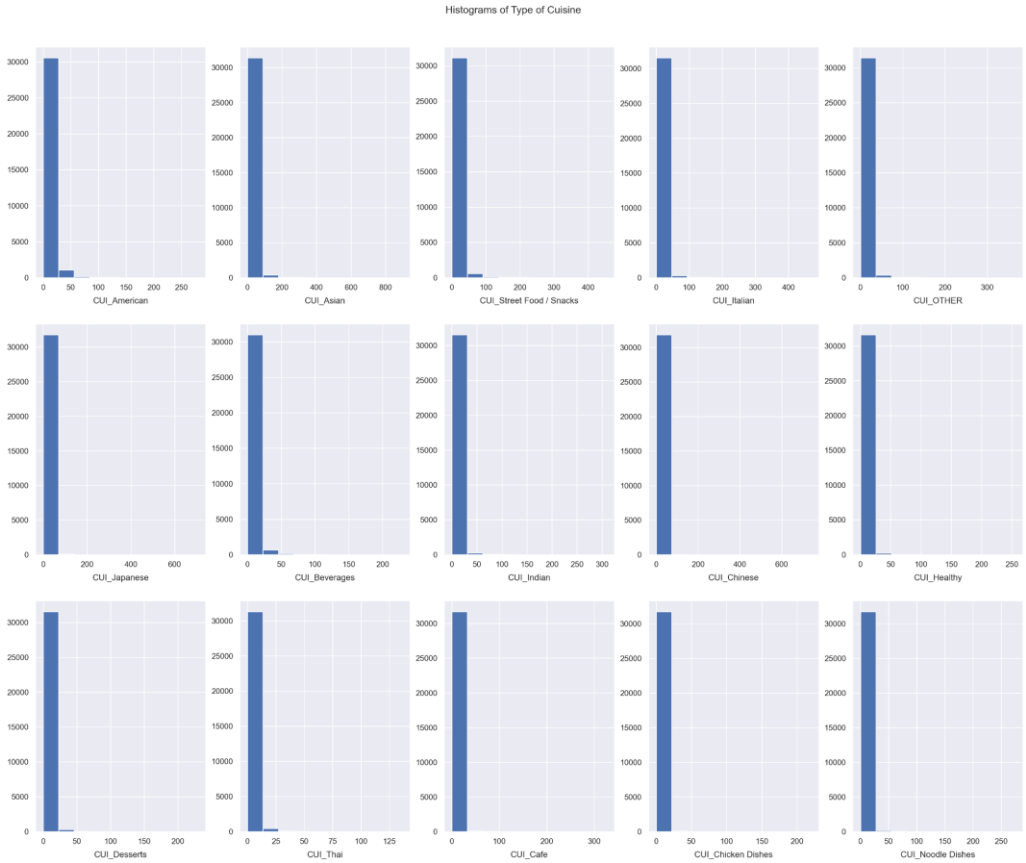

Figure B3: Cuisines' Histogram

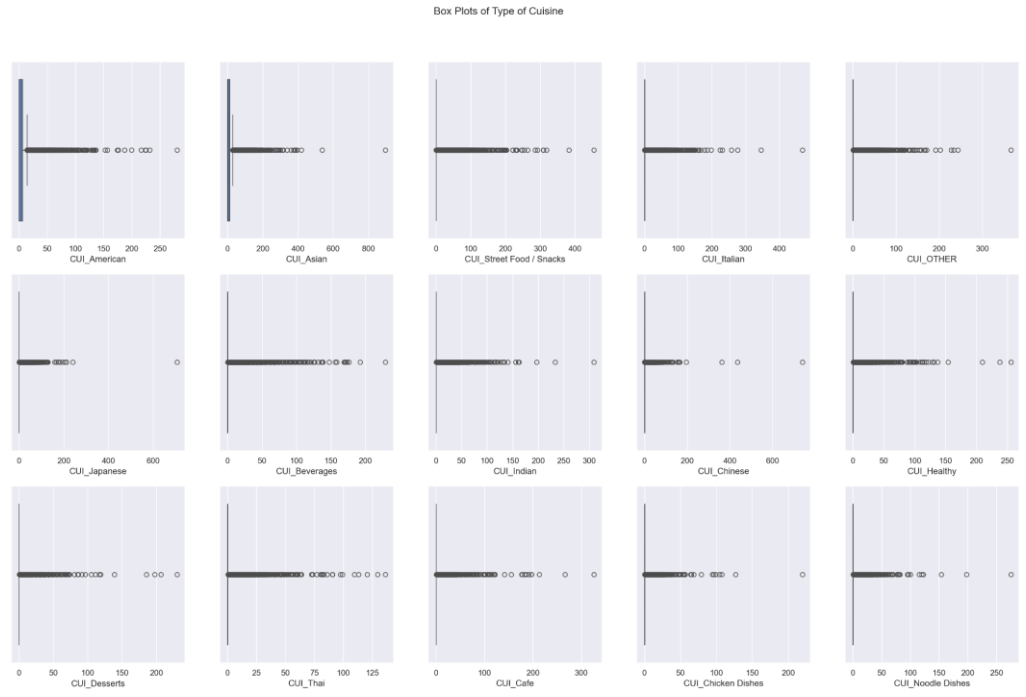

Figure B4: Cuisines' Boxplot

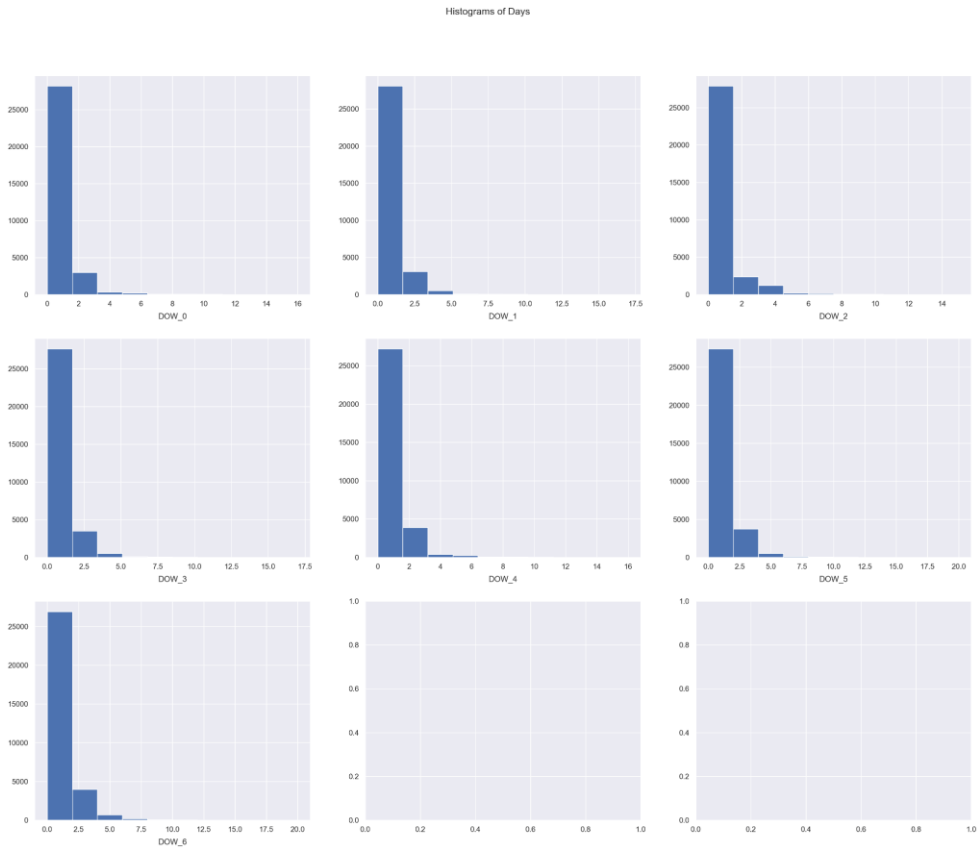

Figure B5: Days of the Week Histogram

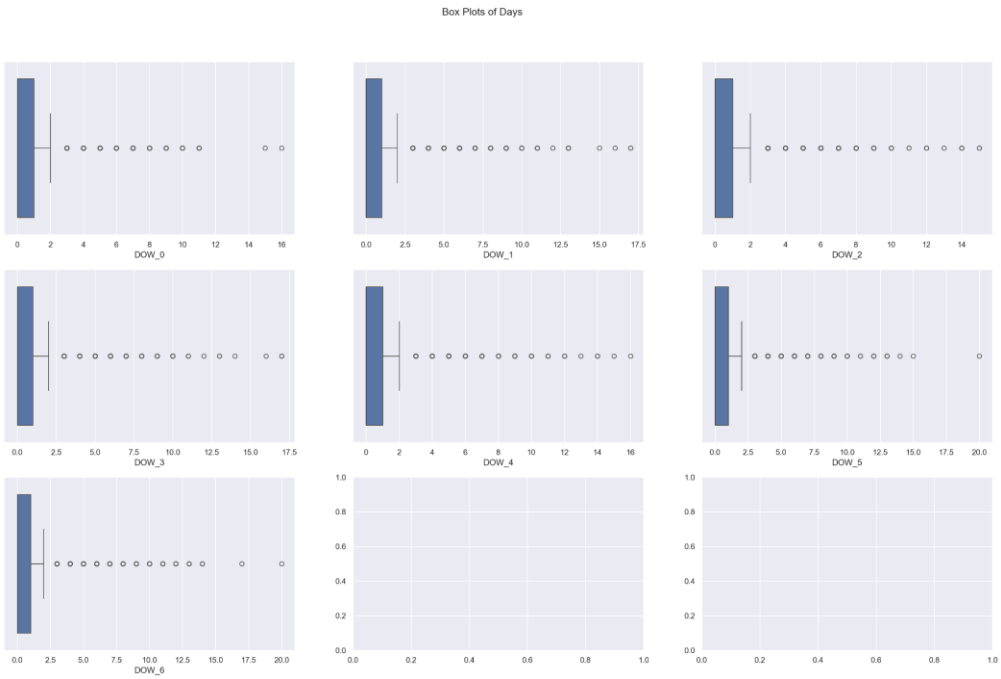

Figure B6: Days of the Week Boxplots

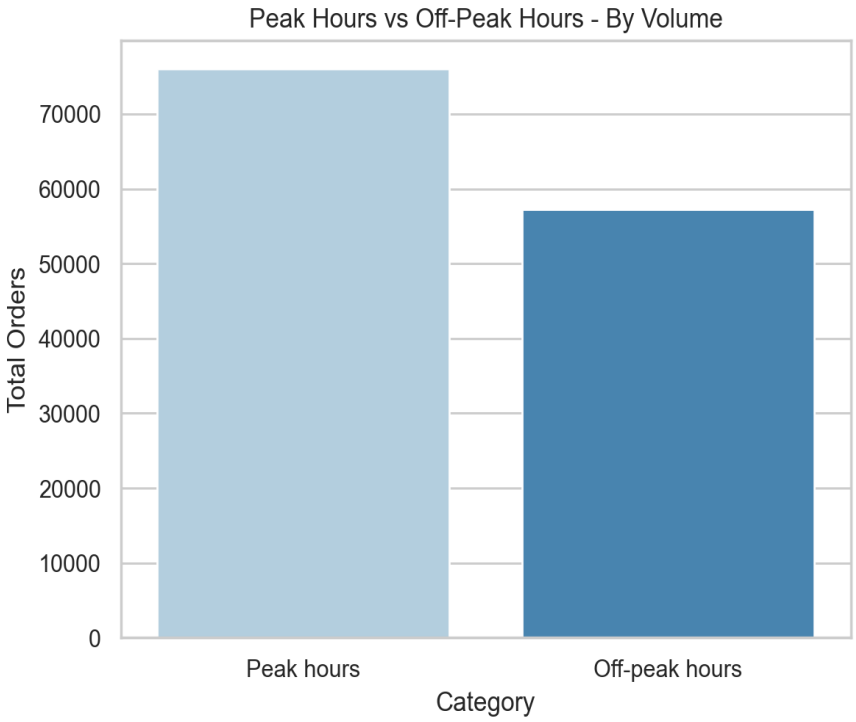

Figure B7: Peak Hours and Off-Peak Hours

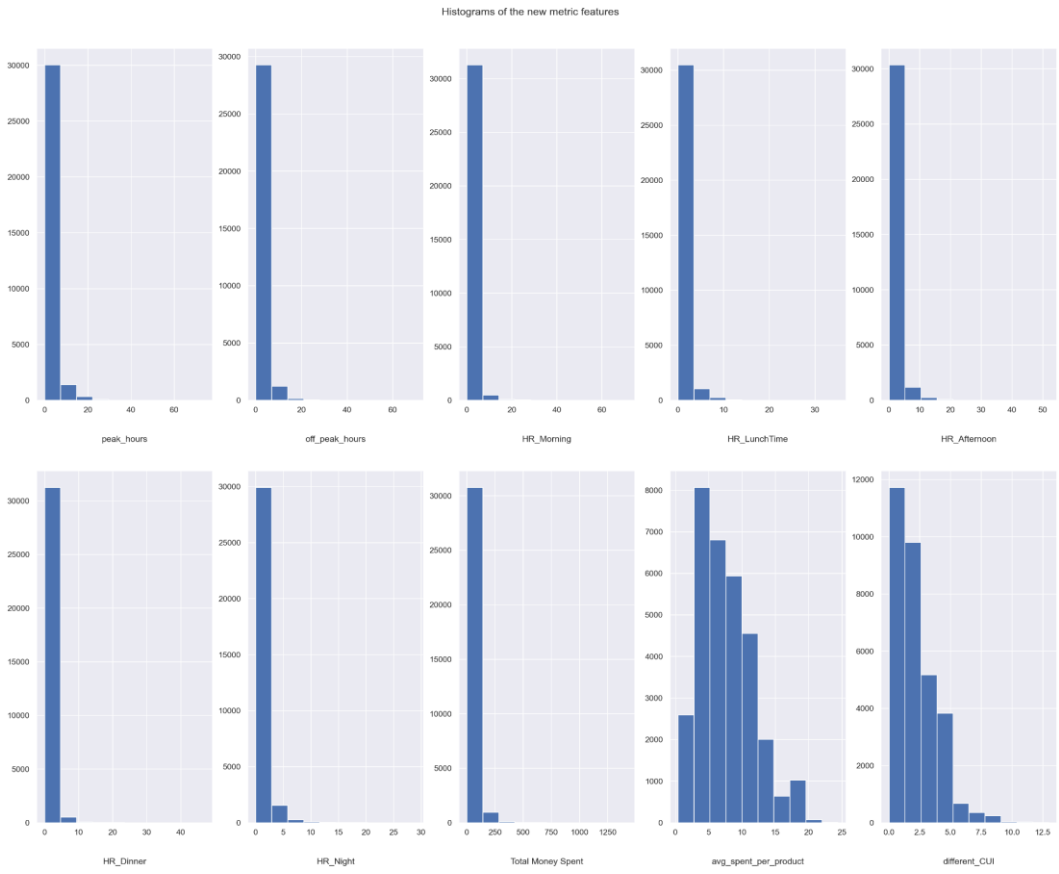

Figure B8: New Numeric Features' Histogram

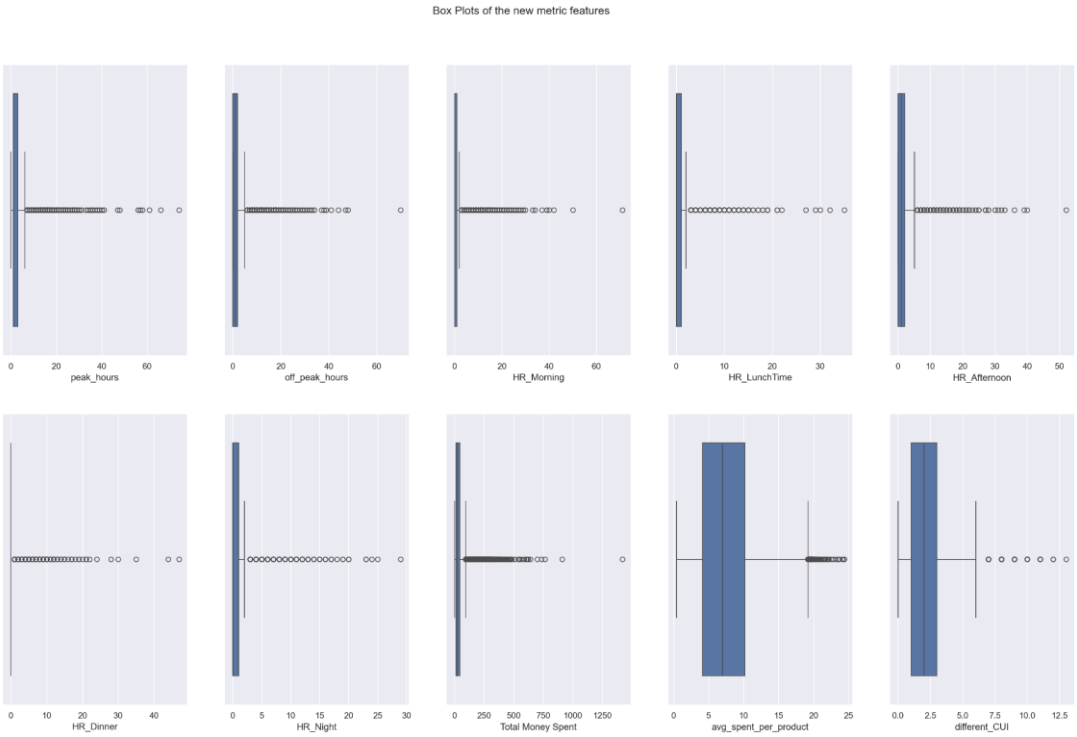

Figure B9: New Numeric Features' Boxplots

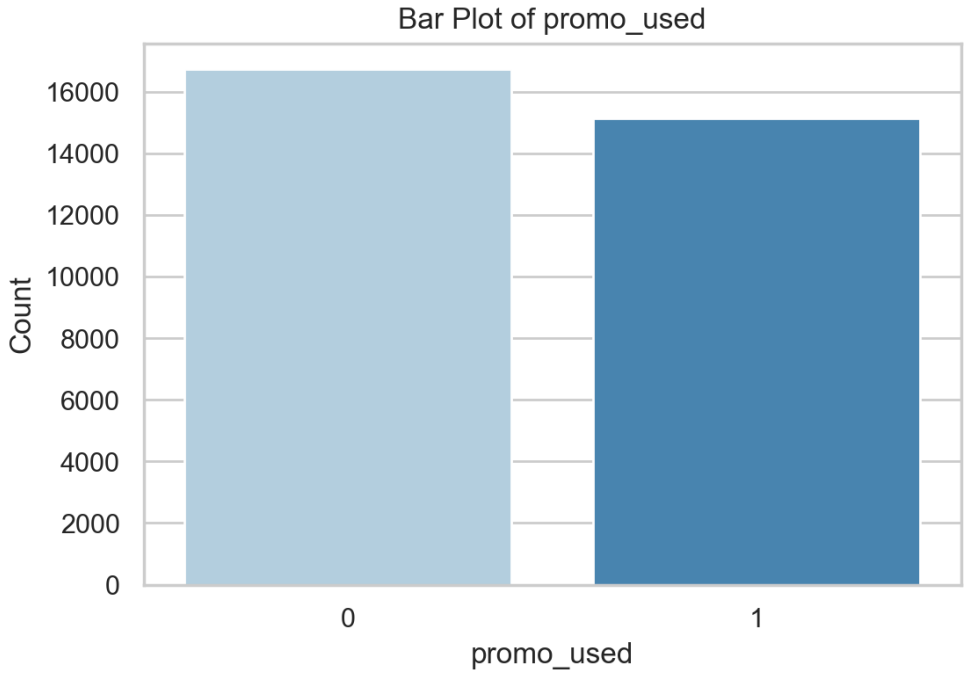

Figure B10: Frequency of Promotional Method Used

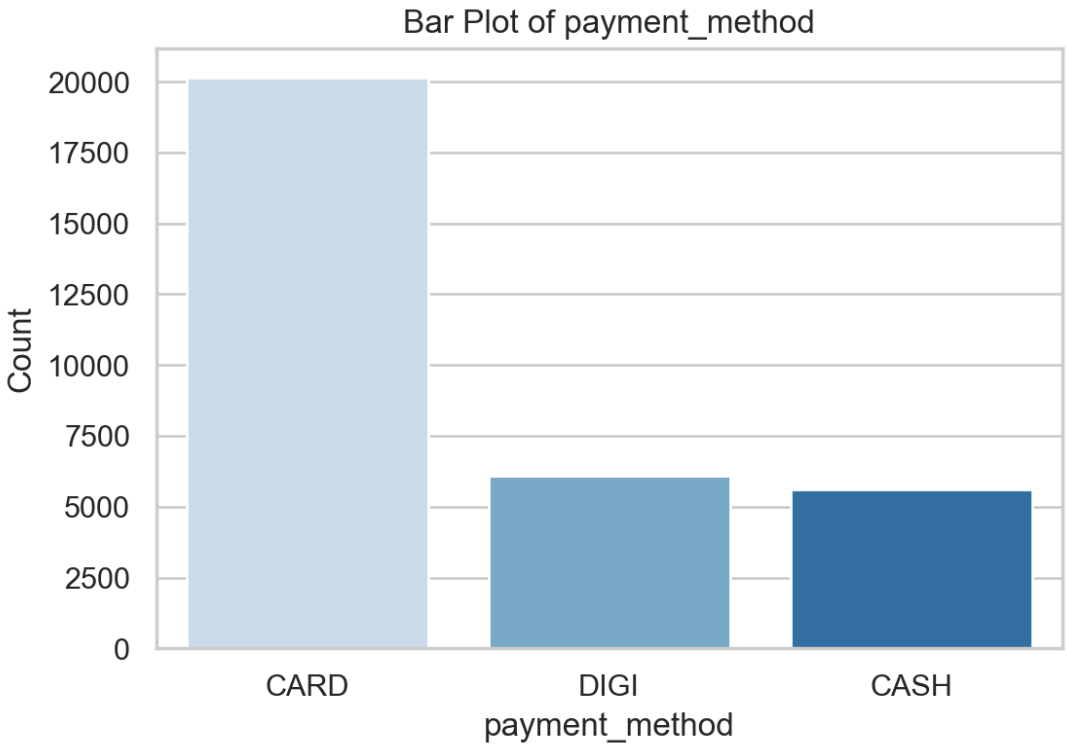

Figure B11: Frequency of Each Payment Method

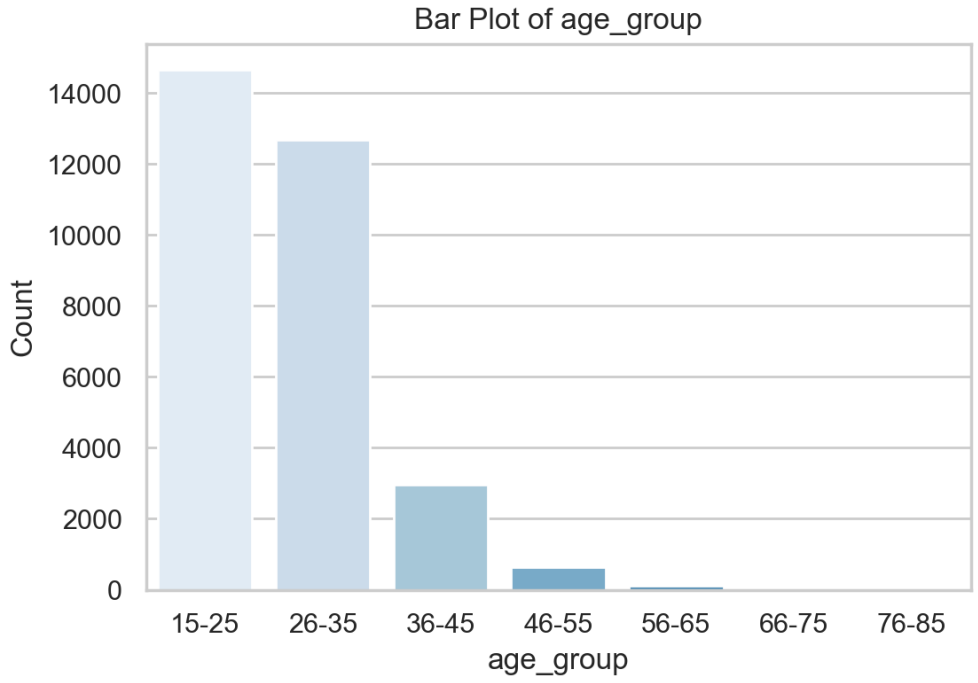

Figure B12: Frequency of Customers by Age Group

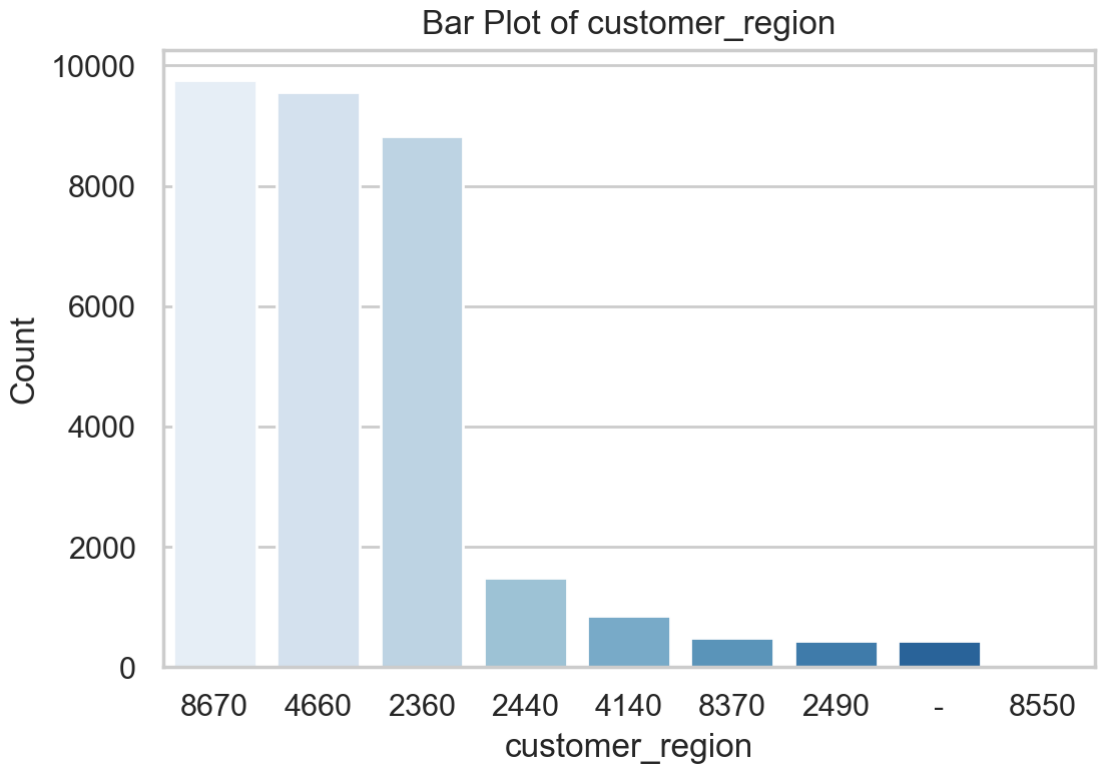

Figure B13: Frequency of Customers by Customer Region

APPENDIX C (BIVARIATE EDA)

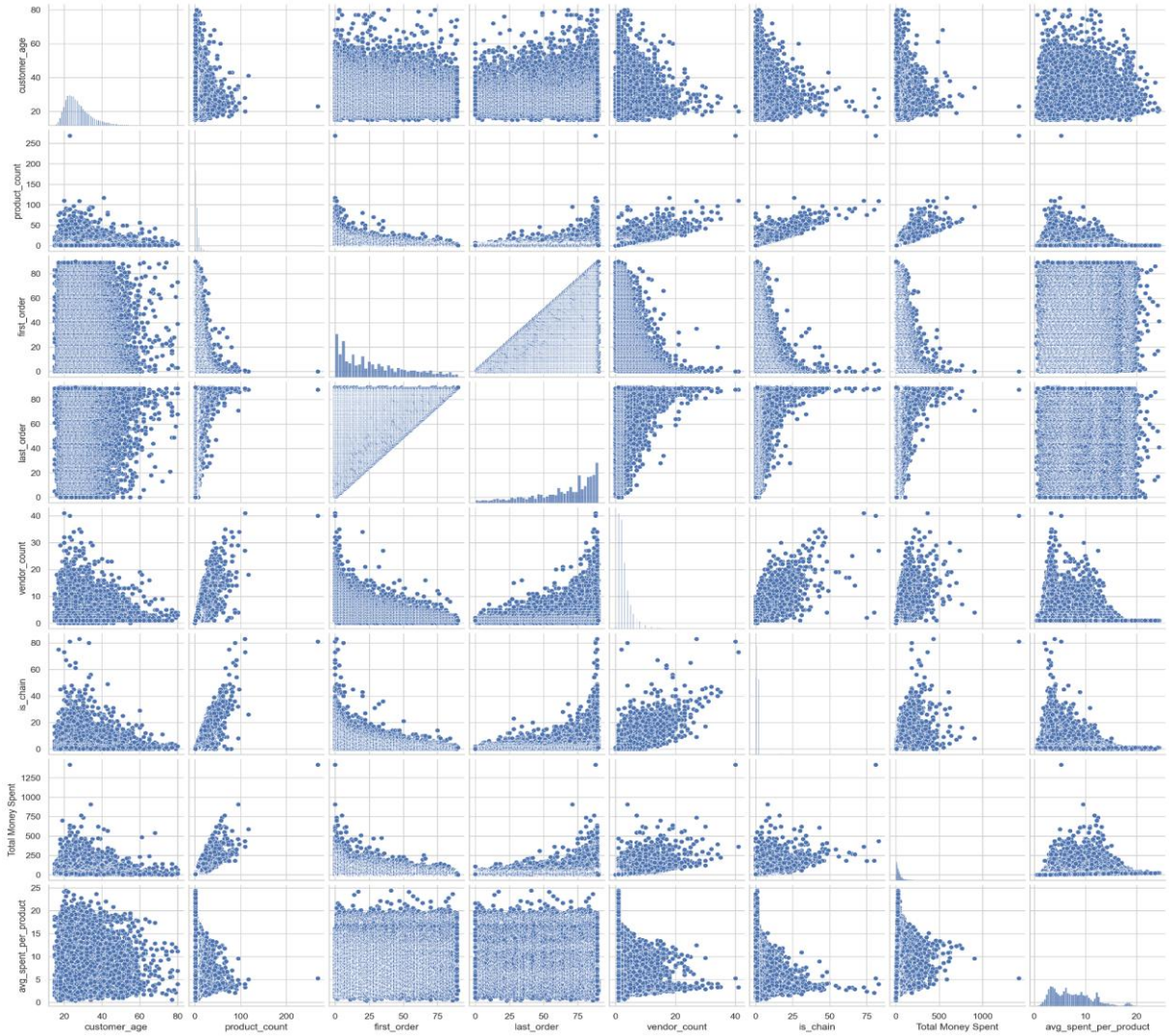

Figure C1: Scatterplot Correlation

(Scatterplot correlation for features: "customer_age", "product_count", "first_order", "last_order", "vendor_count","is_chain","Total Money Spent","avg_spent_per_product")

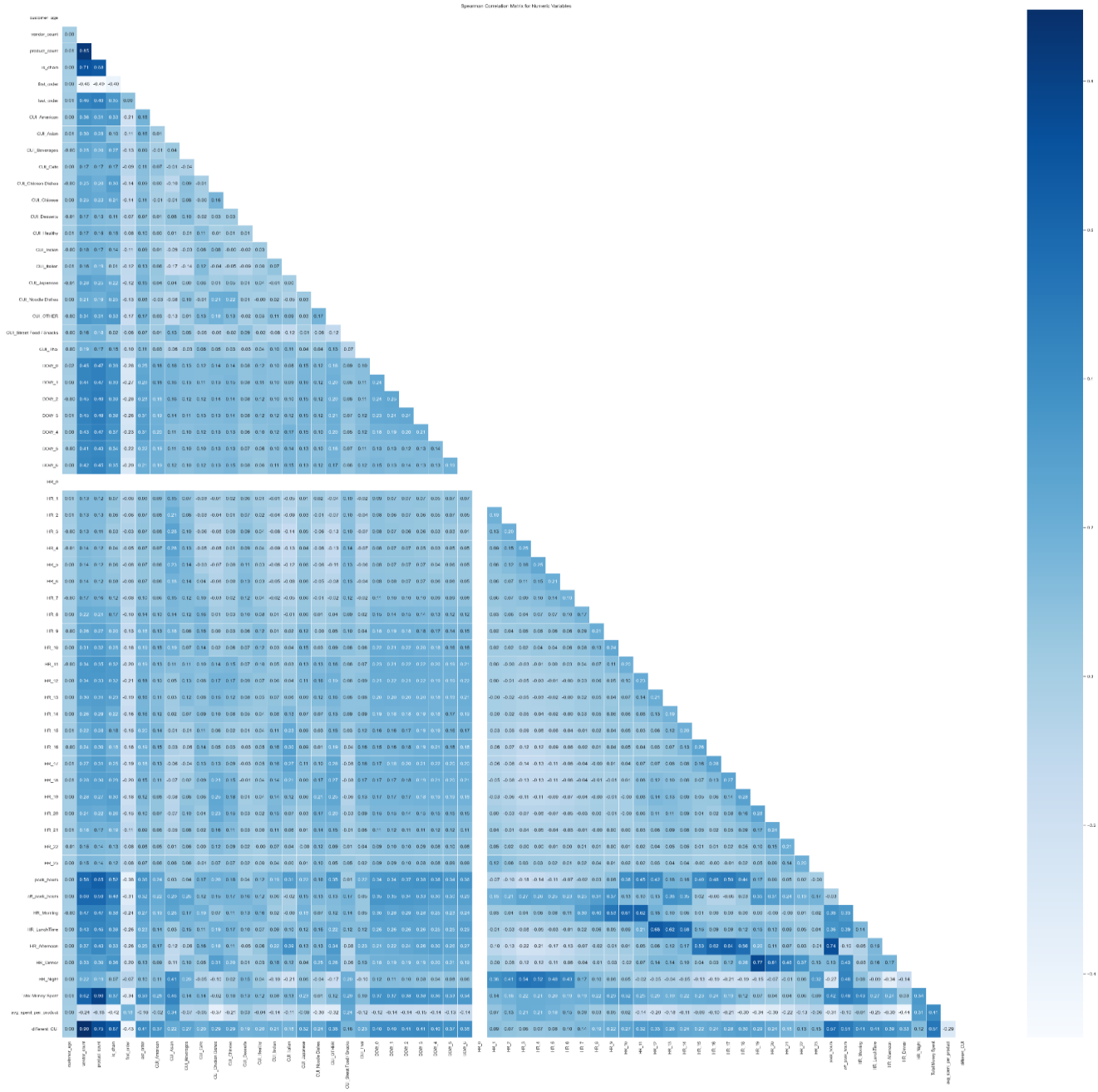

Figure C2: Spearman's Correlation Matrix

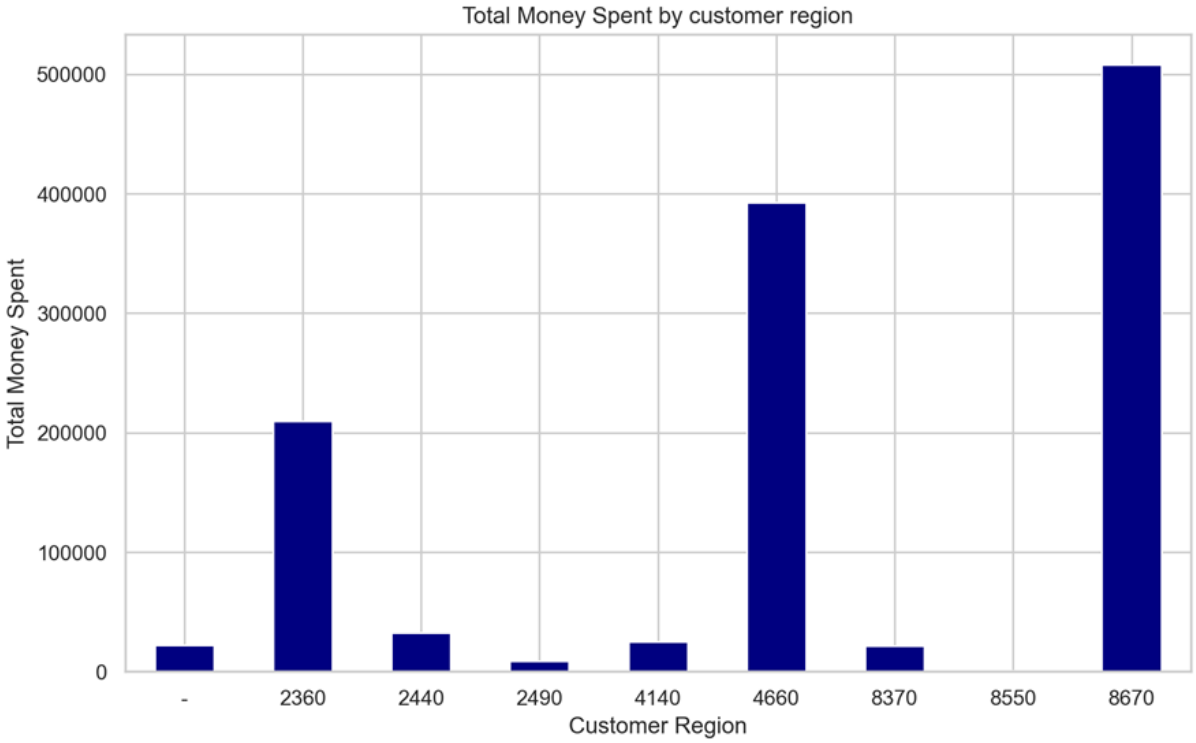

Figure C3: Total Money Spent by Customer Region

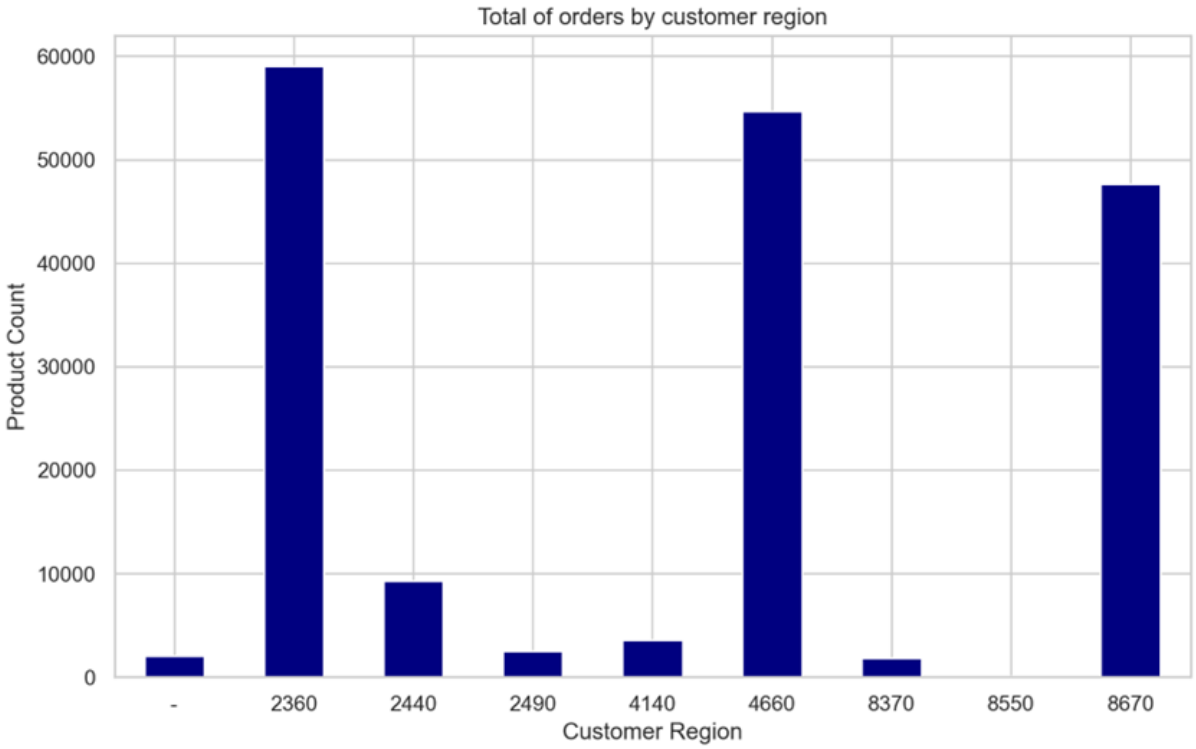

Figure C4: Total Number of Orders by Customer Region

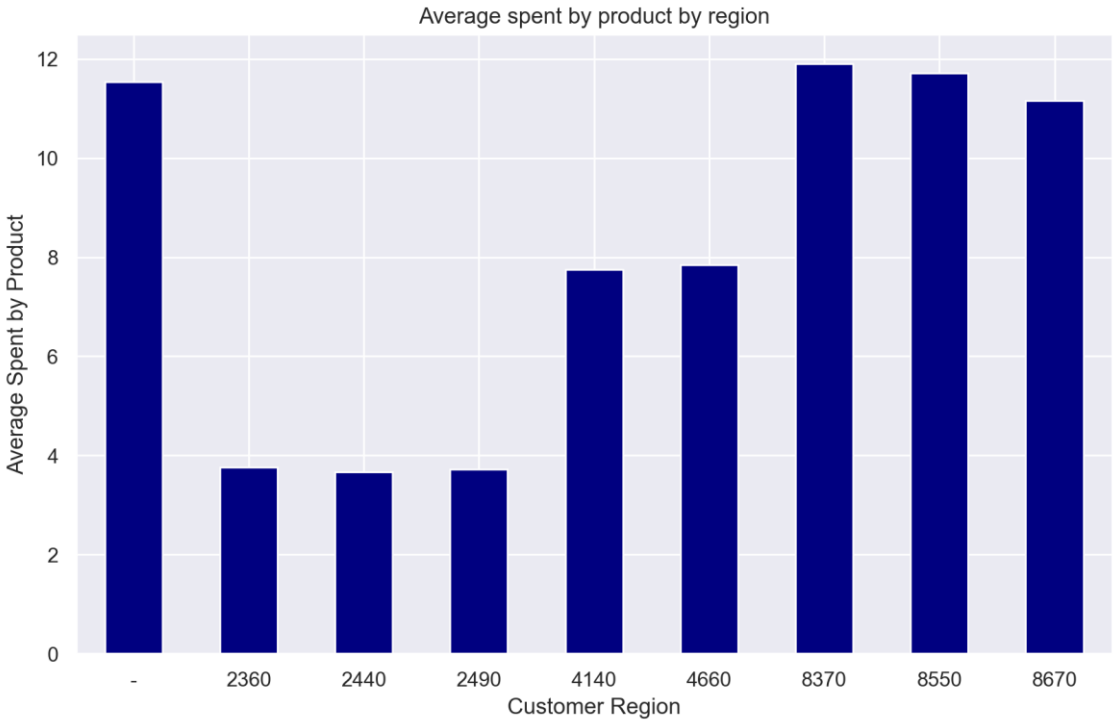

Figure C5: Average Spending per Product by Customer Region

Figure C6: Proportion of Money Spent by Demographic Region on the 7 Cuisines where Most Money is Spent

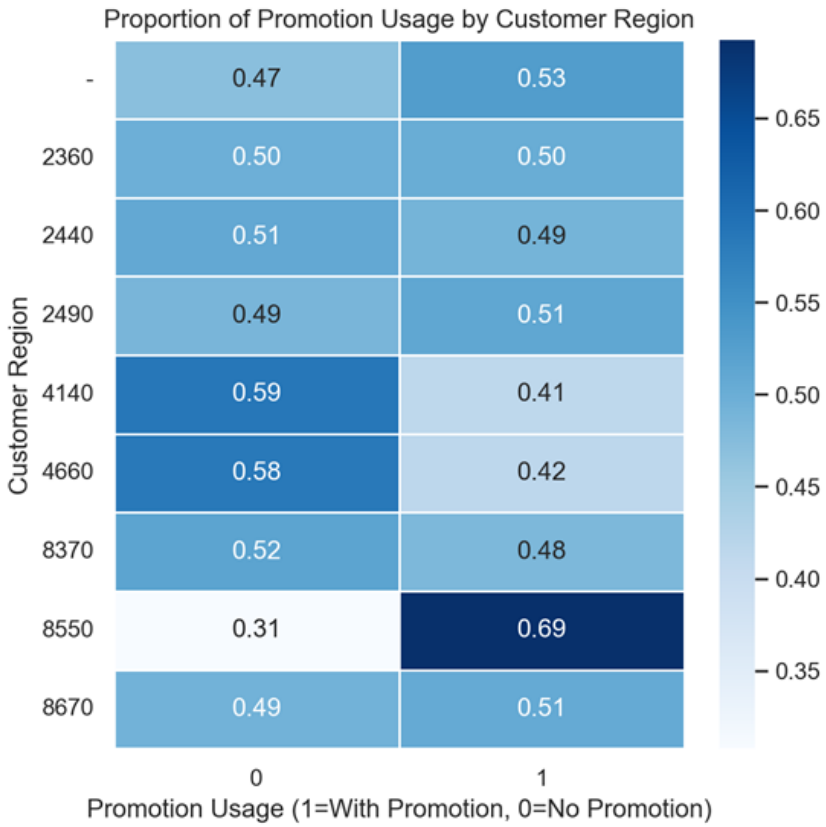

Figure C7: Proportion of Promotion Usage on the Last Order by Customer Region

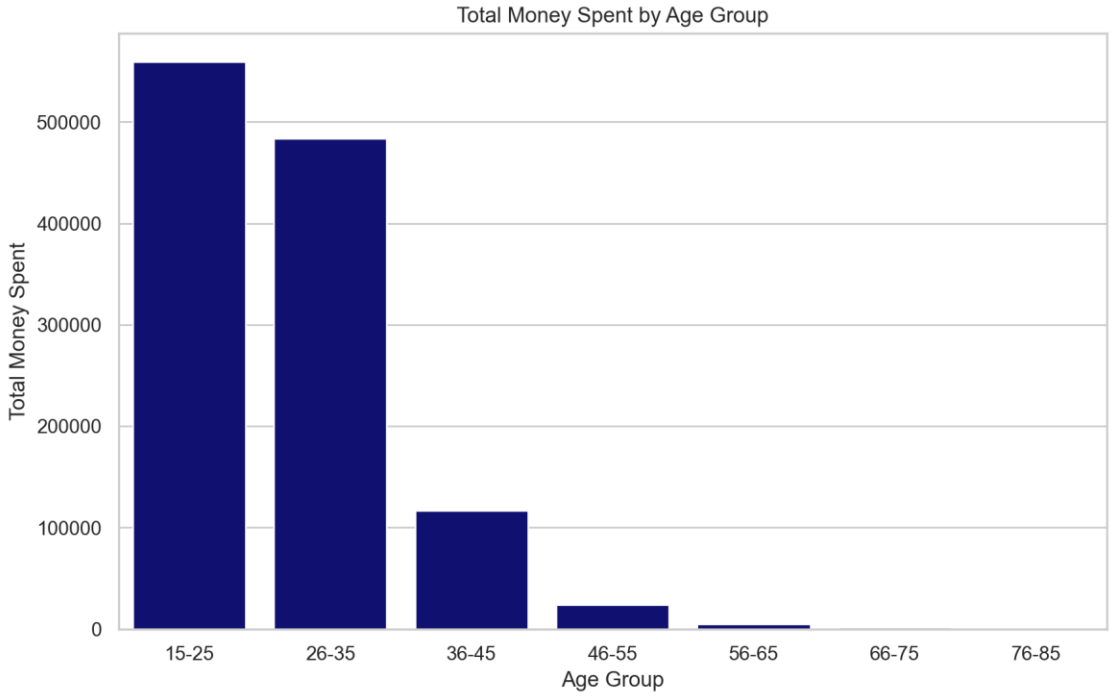

Figure C8: Total Money Spent by Age Group

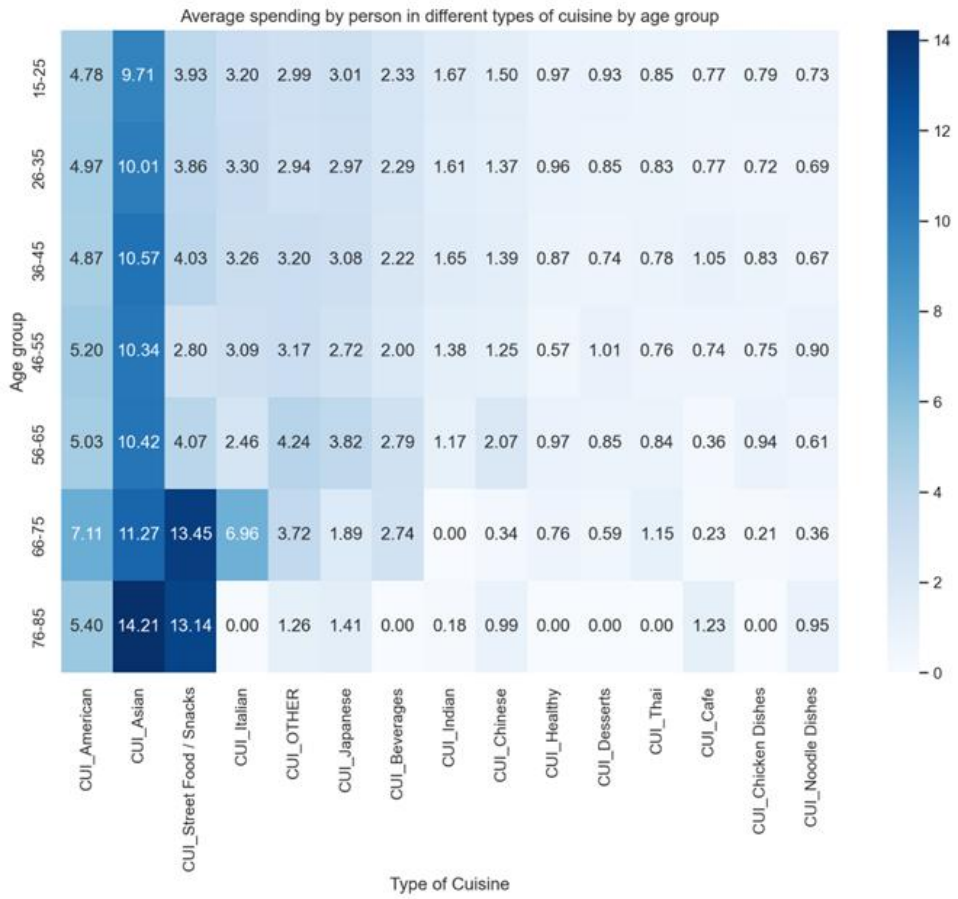

Figure C9: Average Spending per Person per Cuisine Type and Age Group

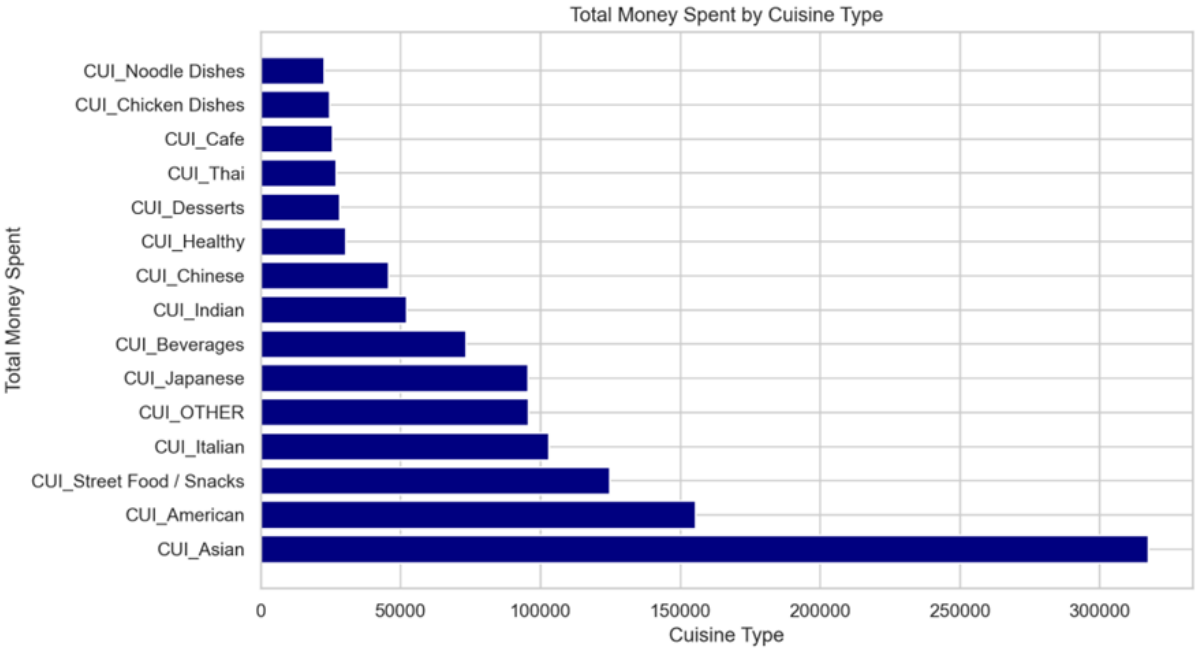

Figure C10: Total Money Spent by Cuisine Type

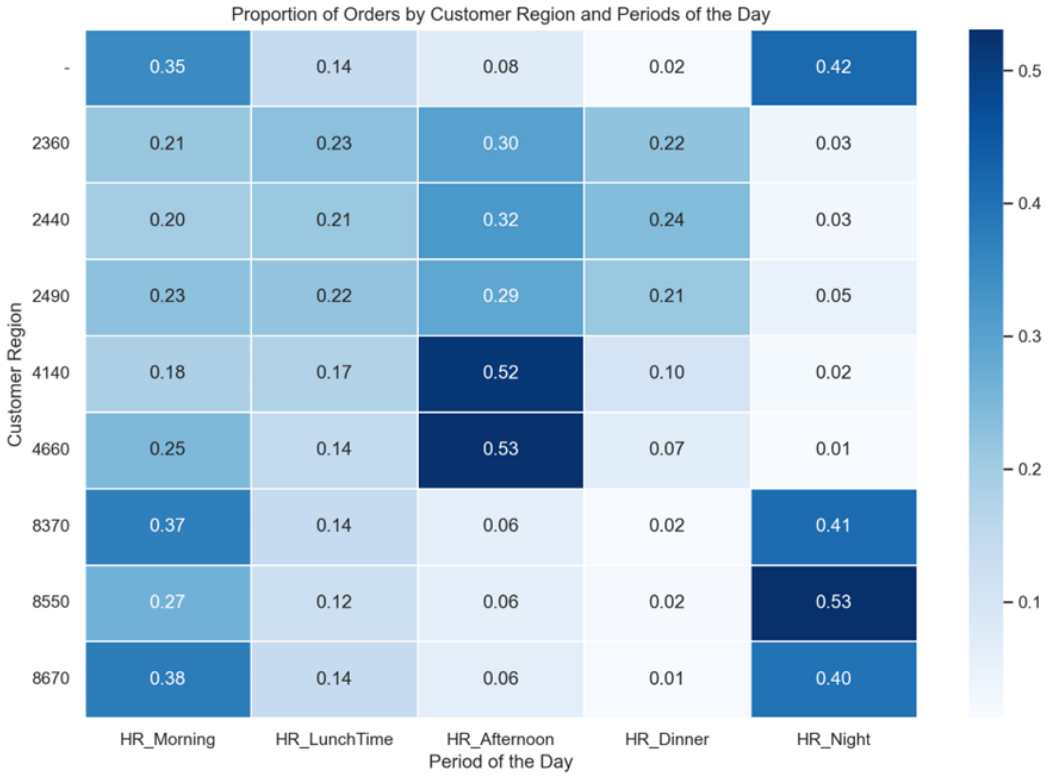

Figure C11: Proportion of Orders by Customer Region and Periods of the Day

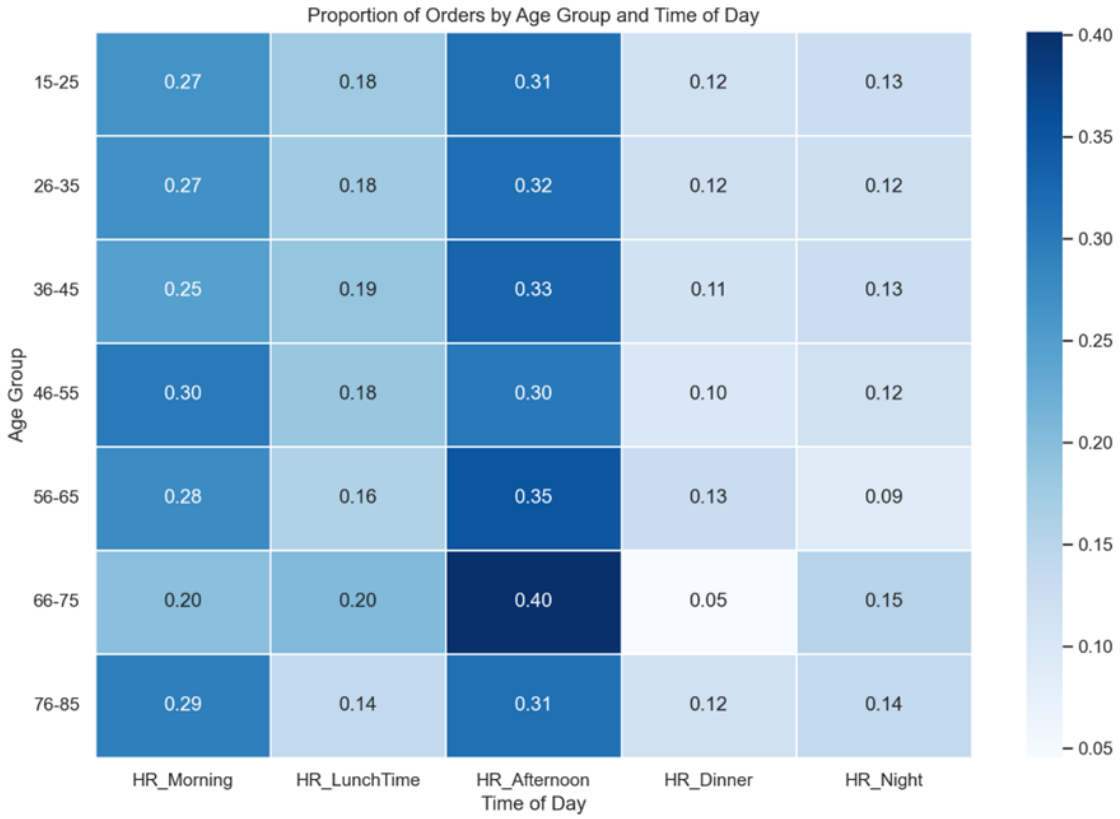

Figure C12: Proportion of Orders by Age Group and Periods of the Day

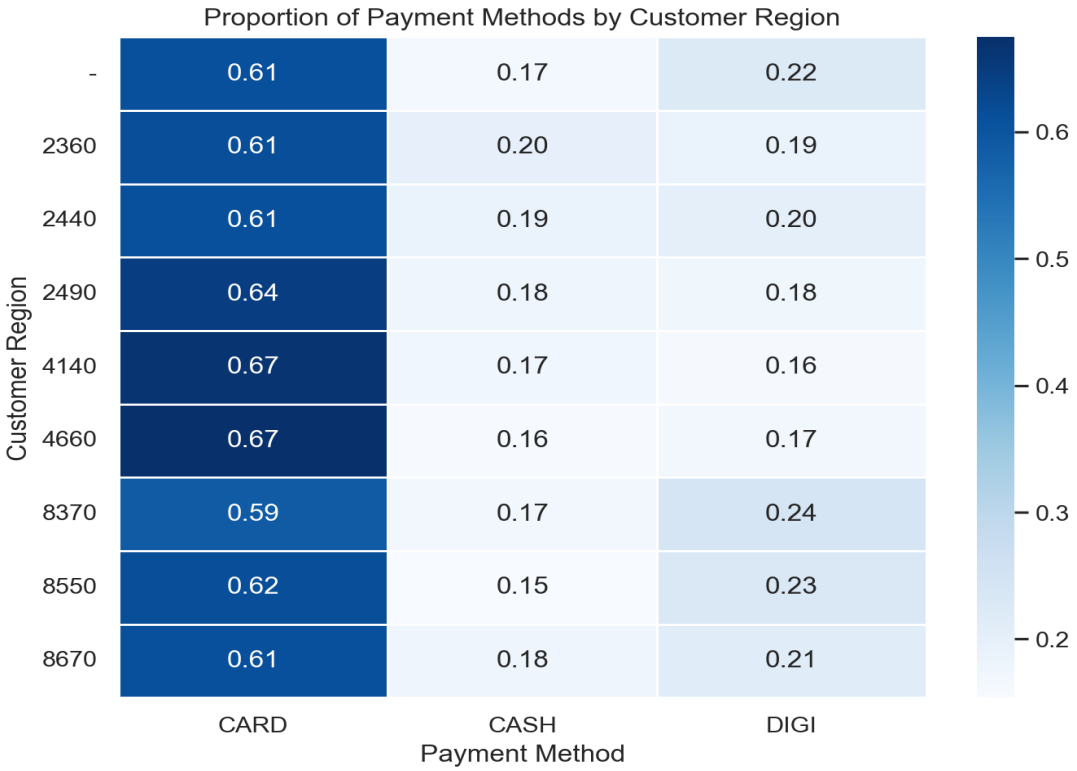

Figure C13: Proportion of Payment Methods by Customer Region

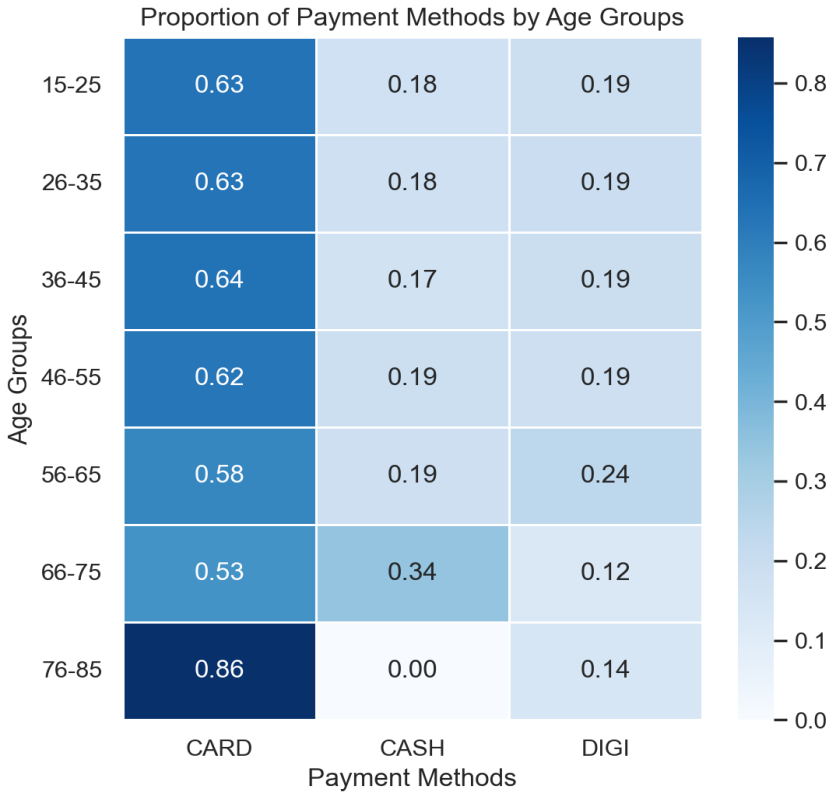

Figure C14: Proportion of Payment Methods by Age Groups

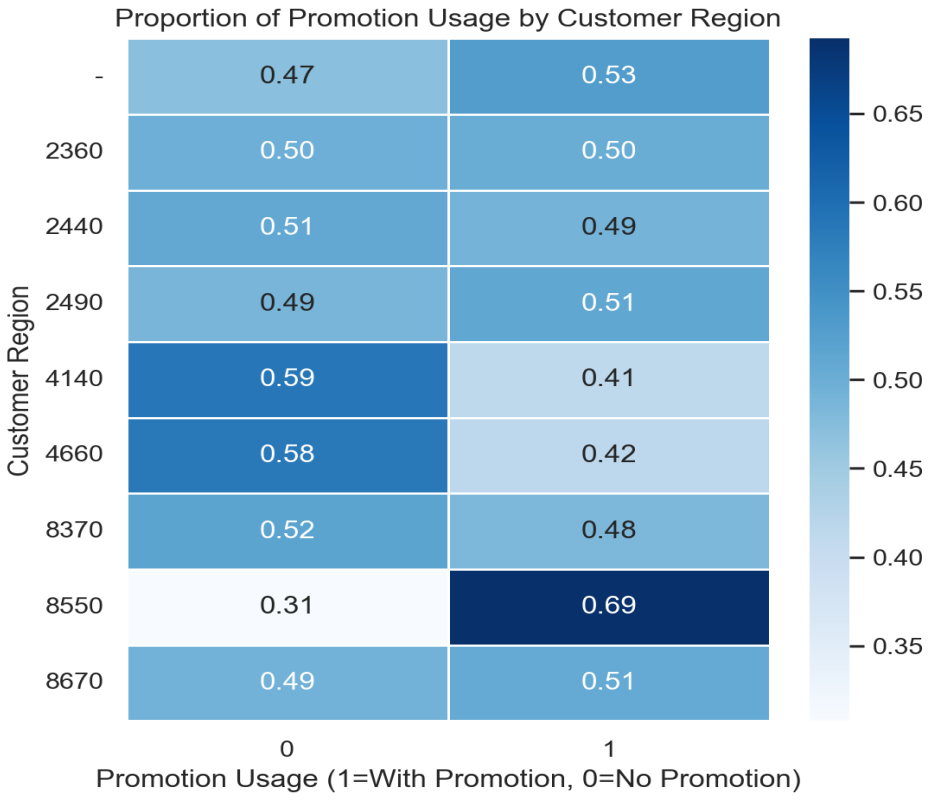

Figure C15: Proportion of Promotion Usage by Customer Region

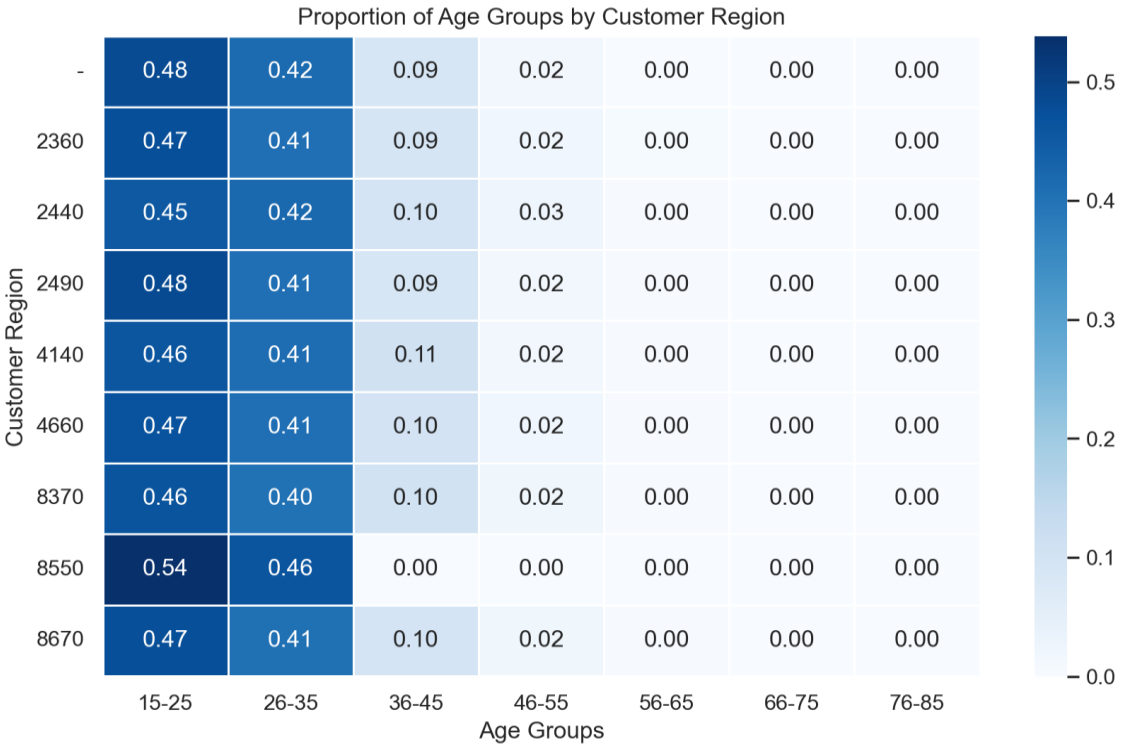

Figure C16: Proportion of Age Groups by Customer Region

Figure C17: Proportion of Orders by Customer Region and Periods of the Day

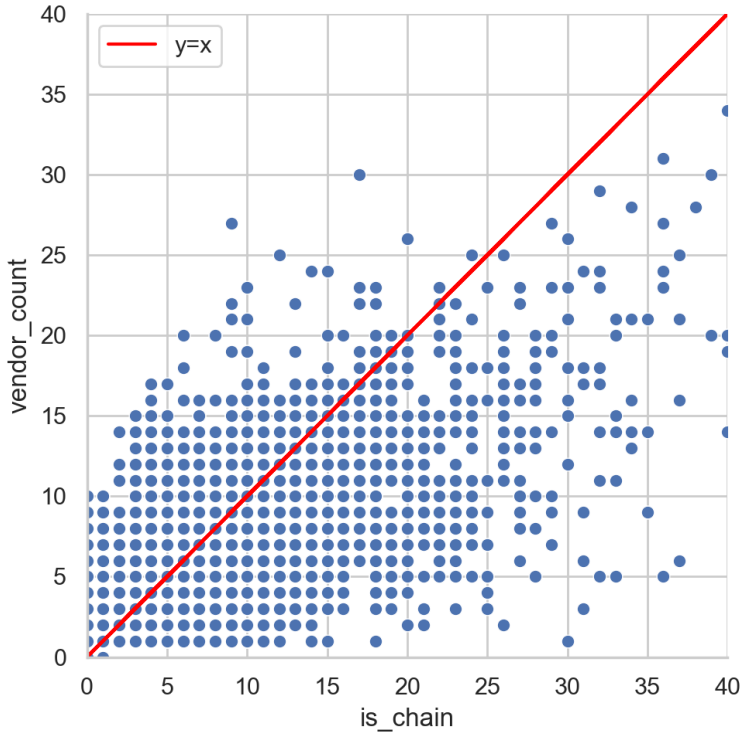

Figure C18: Relation between Is_Chain and Vendor_Count

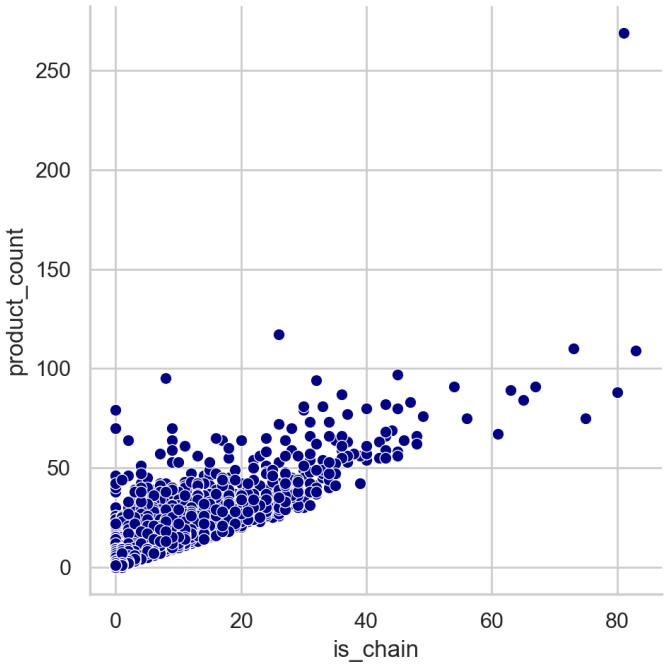

Figure C19: Relation between Is_Chain and Product_Count

ANNEXES

ChatGPT was used in the following cases:

. To generate the y=x line in *Figure C19: Appendix C*

. To aid in the making of the first plot where promotions were used (During our work, future plots were adapted, and knowledge was gained from this first visualization).

. In *Figure C2: Appendix C* so that only the lower triangle is shown in the correlation matrix

. To ensure well-written text and improve textual organization.