DATA MINING PROJECT

Master in Data Science and Advanced Analytics

NOVA Information Management School

Universidade Nova de Lisboa

ABCDEats Inc.

Group 26

Report made as part of the Curricular Unit of Data Mining

Diogo Miguel Calisto Rodrigues, 20240512

Daniel Rodrigues Rainho, 20240607

Duarte Queiróz Miguel, 20240608

Fall Semester 2024-2025

TABLE OF CONTENTS

1. INTRODUCTION

This project aims to act as consultants for a fictional food delivery service called ABCDEats Inc. In this case, our goal is to analyze all the customers data collected over three months from three different cities and assist the service in developing a data-driven strategy for various customer segments.

We are free to try and analyze various approaches and perspectives in this project with the intention of giving the company a final segmentation to help them develop a marketing strategy.

2. DATA DESCRIPTION

The sample we received contains **31888** observations and **56** variables that we will need to manage for an easier understanding of the problem. In the following table there's a description of them.

Tabel 1 - Variable Description

VARIABLE	TYPE	DESCRIPTION
customer_id	object	Customer ID
customer_region	object	Geographic region where the customer is located
customer_age	float64	Age of the Customer
vendor_count	int64	Number of unique vendors the customer has ordered from
product_count	int64	Total number of products the customer has ordered
is_chain	int64	Number of times the costumer ordered from a chain restaurant (*)
first_order	float64	Number of days from the start of the dataset when the customer first placed an order.
last_order	int64	Number of days from the start of the dataset when the customer most recently placed an order.
last_promo	object	The category of the promotion or discount most recently used by the customer.
payment_method	object	Method most recently used by the customer to pay for their orders
CUI_American, CUI_Asian, CUI_Chinese, CUI_Italian,etc.	float64	The amount in monetary units spent by the customer from the indicated type of cuisine.
DOW_0 to Dow_6	int64	Number of orders placed on each day of the week (0 = Sunday, 6 = Saturday).
HR_0	float64	Number of orders placed during each hour of the day (0 = midnight, 23 = 11 PM).
HR_1 to HR_23	int64	

(*) Originally the metadata implied that the variable **is_chain** should be boolean. The description of the variable given by the problem did not correspond well to the data and in our interpretation, this description better suits the type of the problem, and the date received. We'll justify this decision in the Analysis of Variables chapter.

To obtain trustworthy results, we must check our data and clean it. The first thing we notice is that our sample has **13 duplicate observations** that we will delete from our database. Secondly, we had to check the missing values. We have **727 missing values in customer_age (around 2.28%)**, 106 missing values in **first_order (around 0.33%)** and 1165 missing values in **HR_0 (around 3.65%)**. Further forward, these missing values will be processed.

3. ANALYSIS OF VARIABLES

Let's analyze the **customer_id** first. Our idea is to set this one as the index of our DataFrame, it is beneficial because identifies each customer, making it easier to locate and manage individual data efficiently. There are no duplicated id's, so we set this variable as the index.

Next, we have the **customer_age**. In our analysis of the data. As we can see in *Figure 1* by examining the histogram and boxplot, we can see that the age distribution has a positive skew. A significant number of outliers skew toward older age groups, those are not errors but some older users using the application. Most of the users belong to a younger age group.

For the **customer_region** we wanted to see the main target regions. The three main clients are from the regions **8670**, **4660** and **2360** by observing *Figure 2*. The category ' – ' means unknown and we should decide if we keep it or change it.

Looking now for the **vendor_count** and the **product_count** variables' histogram (*Figure 3* and *Figure 4)*, we can see a positive skew revealing a concentration at lower counts. Both variables have their outliers, as seen in the bloxplots are unlikely to be data errors, representing clients who made more orders/products. It is remarkable that product_count variable has an outstanding outlier that could negatively impact our analysis and visualization, that was omitted.

Now, let's look at two variables at the same time, **first_order** and **last_order**. Looking at both histograms (*Figure 5*) we could see contrasting shapes between them. In the first_order histogram, most customers made their first purchase early in the dataset timeline, with fewer joining over time. In the last order histogram, many customers remained active or re-engaged toward the end, suggesting strong retention or successful reactivation efforts.

Looking at the **last_promo** variable, we interpreted that the rows with value **' – ',** indicates customers who did not use any promotions. Checking *Figure 9*, most of our users didn't use any promotions (52.5%), followed by Delivery Promotion with 19.7% of the clients.

At the **payment_method**, the majority use card as their payment method (63.2%) seen in *Figure 10*.

Looking at all **CUI_** variables, *Figure 6*, the cuisine with the most money spent is Asian being followed by the American and Street food / Snacks.

Next, we analyzed the number of orders placed on each day of the week, the **DOW_** variables. The data indicates an increase in orders throughout the week as we can see in *Figure 11*. Peaking on Sunday. This could reflect a behavior where people are more likely to shop during the weekend.

As for the **HR_** variable, we checked that in the hour 0, all rows take the value 0, indicating either that there were no orders placed at midnight or that was an error in collecting data for this variable. Orders peak at hours 17 and 11 (as we can see in *Figure 8*) are likely aligning with lunch and early dinner times or after work hours. In contrast, order volumes are lowest during early morning when people typically don't order food.

Originally the **is_chain** variable was described as 'Indicates whether the customer's order was from a chain restaurant.', however it takes values between 0 and 83 (not Boolean) so the data given to us does not correspond to the original description. We decided to check its distribution, seen in *Figure 7*, and we noticed that it's very similar to the vendor_count's distribution. Our conclusion is that the variable itself gives us the number of times the costumer ordered from a chain restaurant. Also, it has the same problems as the vendor_count.

4. CREATION OF NEW VARIABLES

In our work, we decided to create new variables to help approach the problem and the relationship between the original variables.

Customer_time is a variable that represents the duration of each customer's time with the delivery service. We get this variable by subtracting last_order by first_order. Because we have missing values in the first_order variable, customer_time will have the same number of missing values. Looking at the histogram (*Figure 12*), we can see that most users have a **customer_time** of zero, indicating that they only used the application one day. Excluding the 0 value, the distribution is almost uniform.

Order_count represents the total number of orders for each customer. We get it by doing the sum of the values of HR_0 to HR_23 or DOW_0 to DOW_6. Order_count is very similar and has the same problems as the vendor_count, seen in *Figure 14*.

Intensity_of_activity quantifies how active a customer is by calculating the average time interval between two orders. This variable is obtained by dividing customer_time by order_count. The variable has 0.33% of missing values derived from first_order. This variable has a positive skew (*Figure 15*).

Total_Spent represents the total amount of money spent by each customer by doing the sum of the values from the CUI_ columns. It's noticeable the huge number of outliers as we can see in its boxplot (*Figure 16*), unlikely to be errors and just users who spent a large amount of money.

Diversity_cuisine is a variable that measures how many different types of cuisines a customer has ordered. With that we conclude that most of the users only order from 2 different types of cuisines.

Customer_loyalty tells us how diverse a customer's ordering behavior is. A low value suggests a preference for some vendors. A high value indicates a willingness to try new vendors. This variable takes values between 0 and 1. The mean of the variable is 0.8 so it suggests that most of the users have no preference in some specific vendors when they order.

Age_category categorizes individuals into distinct age groups. We set 4 type of age groups:

- Young [15, 20 [
- Young-Adult [20, 30 [
- Adult [30, 50 [
- Senior [50, 80]

The main group is **Young-Adult** with 63.1%.

Product_intensity represents the average number of products per order of a customer. The variable shows a positive skew, with a concentration of lower values and a longer tail extending toward higher values. A significant number of outliers skew toward higher counts. Those are not errors but costumers that made a bigger purchase of products in certain orders.

5. INCOHERENCE CHECKING

After the data description, we had to analyze the coherence of our data and whether it made sense in our context. The first case that was analyzed was whether the sum of the number of orders placed on each day of the week is equal to the number of orders placed on each hour of the day. We saw that **30711** observations complied with this rule and **1164** did not.

The second case was that the total number of orders cannot be smaller than the vendor count. All the observations comply with the rule. We have no problem here.

With the data we have, to make sense in the context of this work, we decided to check if there are values in the product_count and in the vendor_count with value 0. For the first one we confirmed **156** rows with value 0 and for the second one **138** rows with value 0. Those 138 rows coincide with the lines whose product_count is 0. We decided to go further about these 138 rows and all of them have value 0 in all the cuisine types (CUI_), is_chain and order_count variables. Right now, we will interpret them as errors.

The last case is that the last order can't come before the first order. None of them reject this rule.

6. RELATIONS BETWEEN VARIABLES

In this chapter we'll explore some multivariable analysis and examine how different variables interact with each other.

First, customers are more likely to pay with a card for expensive purchases due to convenience. To test this statement, we will check the relation between **total_spent and payment_method**. We will group customers into sets of 500 based on ascending order of money spent and then analyze the percentage of each payment_method in each spending group. Our groups were chosen based on their

Total_Spent, for example, the 500 customers that spent the least monetary units were assigned to group 1. We see that the group of costumers that spend the least money, about 35% uses their payment as CASH, and about 45 % uses their payment as CARD. However, the group that spends the most money, about 11% uses their payment as CASH and about 75% uses their payment as CARD. We can conclude that as purchases become more expensive, the use of cash becomes less frequent, and the use of cards becomes more frequent, and it doesn't seem to be a relation between DIGITAL payment and expensive purchases. (*Check Figure 19*).

Second, we will see the relationship between **last_promo** and **customer_time**. We thought that customers that didn't stay in the app had a different behavior than the customers that are loyal to. This would make sense has the unique buyers only bought in the app because they could enjoy a promotion. As we can see in *Figure 20*, customers that only bought in one day, have a higher probability of using a promotion type, however repetitive buyers appear to have no problem not using a promotion.

We found it interesting checking the relationship between **CUI_** and **customer_region**. Using the *Figure* 21, the regions 8370, 8550, 8670 and Unknown have preference in Asian Cuisine. The region 4140 prefers by a fine margin the Italian Cuisine, making it the most differentiated. The rest of the regions don't appear to have a specific preference in cuisine type.

The last relationship we looked at was **last_promo** and **customer_region**, we thought that there might be regions where it would be preferable to use a different type of discount, perhaps because they are geographically more distant than other regions. With the help of its plot (*Figure 22*), we concluded that region 8550 prefers the Freebie or Delivery Promo Code, however none of them used the Discount Promo Code, implying that possibly there is no such promo code in this region. Also, we can see that the regions 4140 and 4660 have no preference in a promotion.

7. FIRST THOUGHTS ABOUT THE DATA

In this analysis of the delivery app dataset, we identified key patterns and insights that reveal important relationships between variables. This enhances our understanding of user behavior and operational performance. By examining each variable individually, we highlighted significant statistics that illustrate the central tendencies and variability within the data. Additionally, we explored the interactions among variables and uncovered potential areas for improvement. These findings offer actionable insights that can help inform strategies for optimizing the delivery app, increasing user satisfaction, and boosting app engagement. Future analyses could further enhance this understanding, for example, by clustering customers.

8. APPENDIX

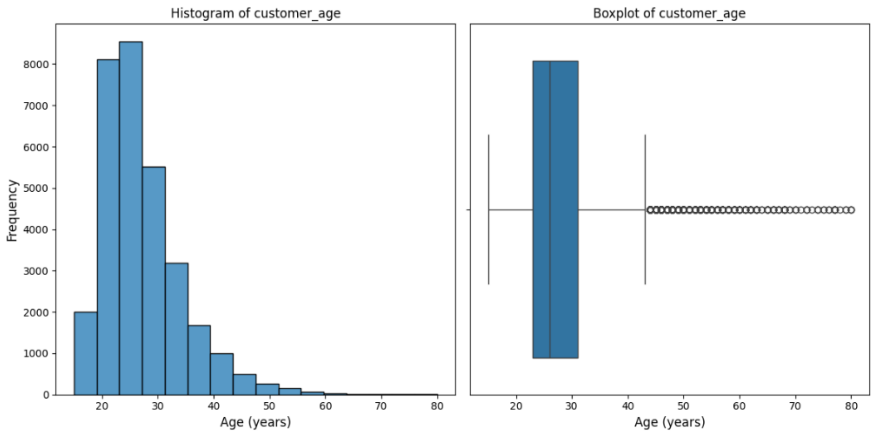

Figure 1 – Plots for customer_age

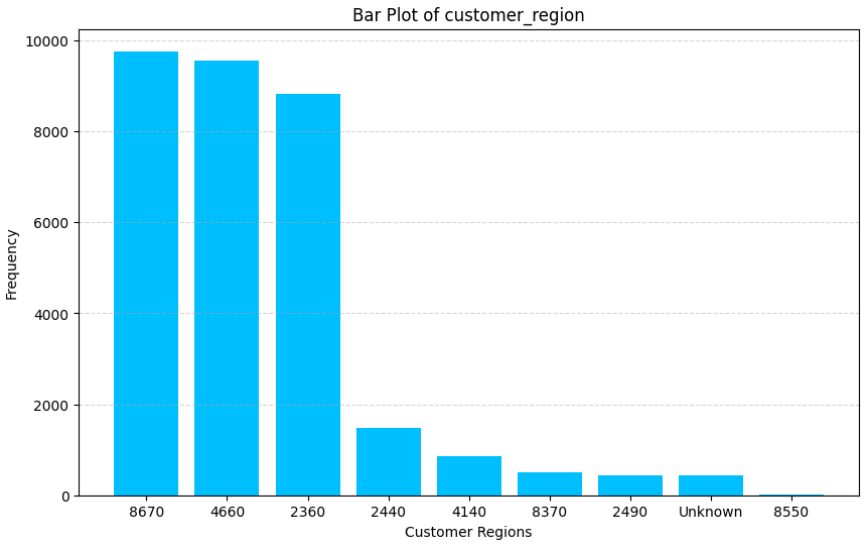

Figure 2 – Bar plot for customer_region

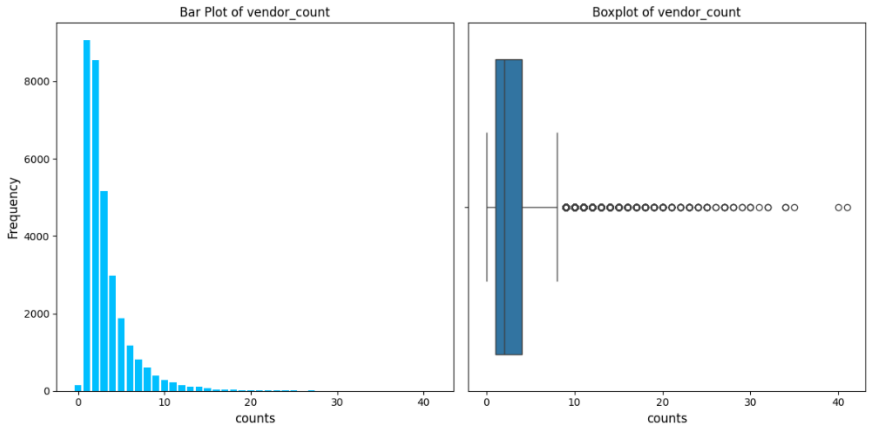

Figure 3 – Plots for vendor_count

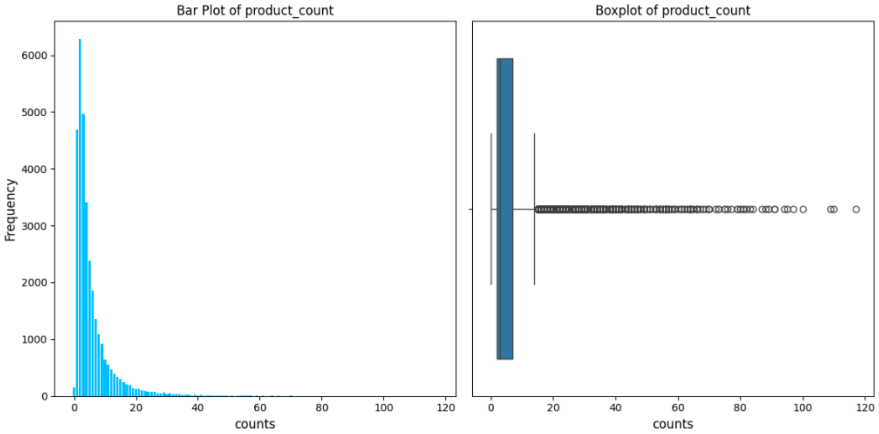

Figure 4 – Plots for product_count

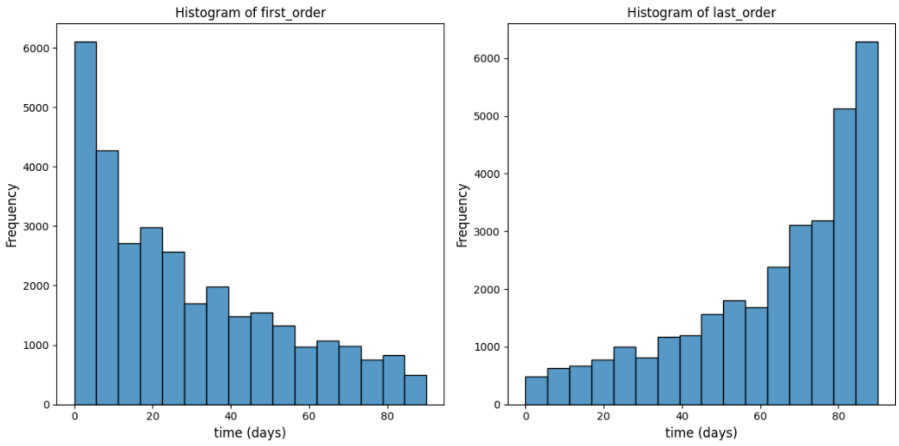

Figure 5 – Histograms of first_order and last_order

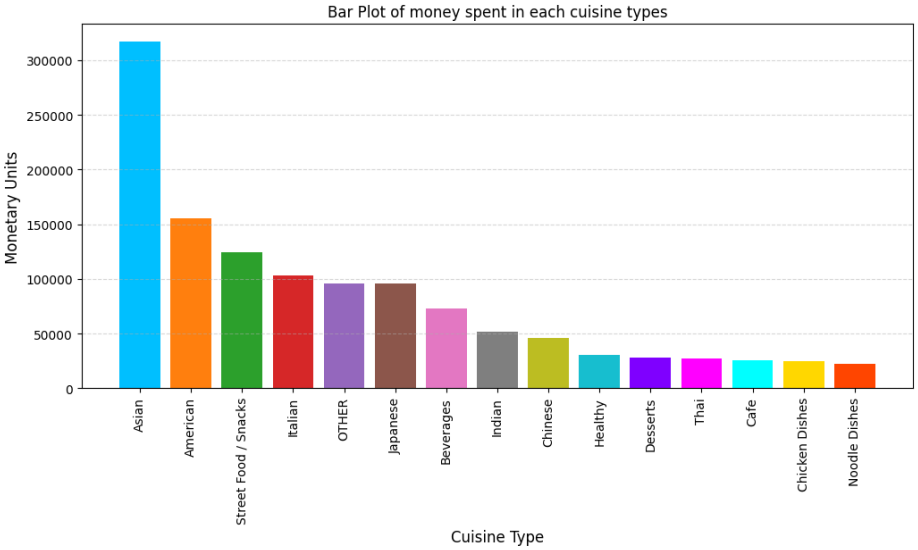

*Figure 6 – Bar plot for CUI_*

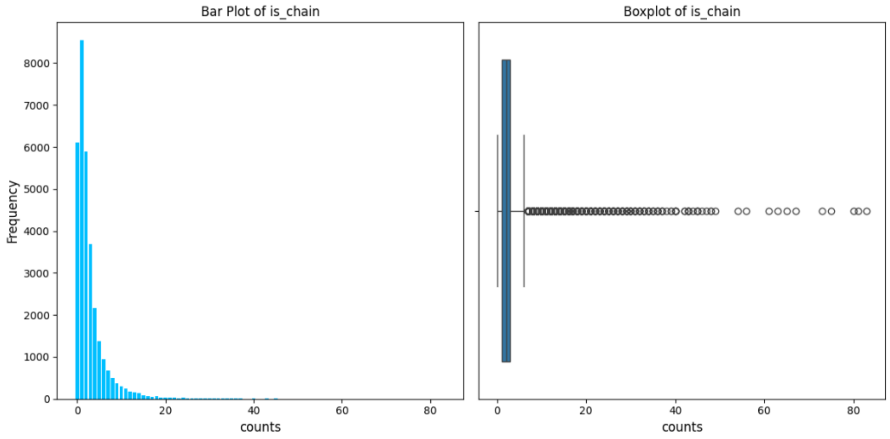

Figure 7 – Plots for is_chain

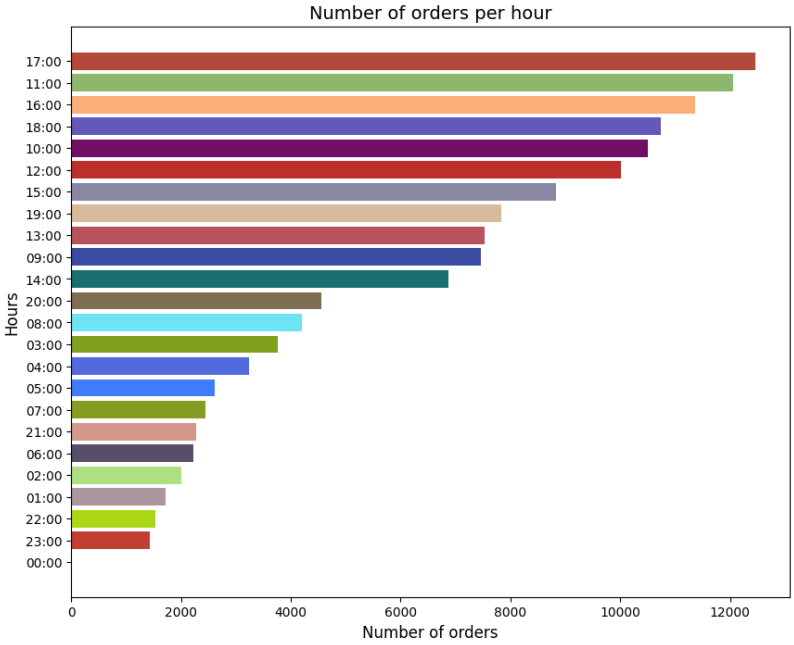

*Figure 8 – Bar plot for HR_*

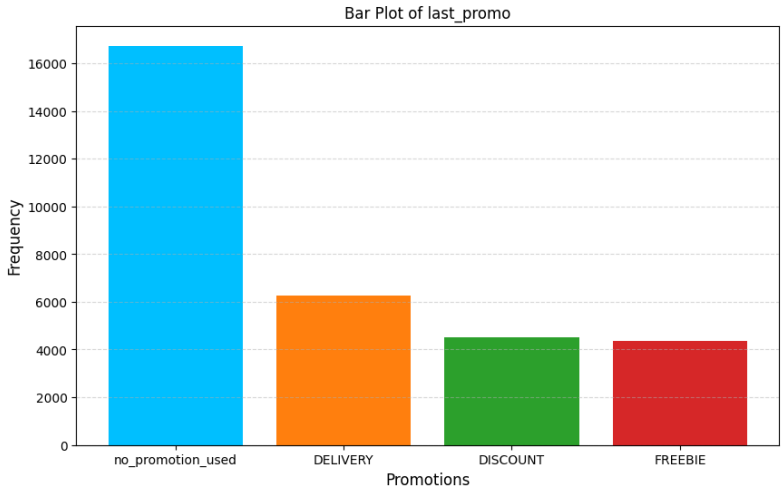

Figure 9 – Bar plot for last_promo

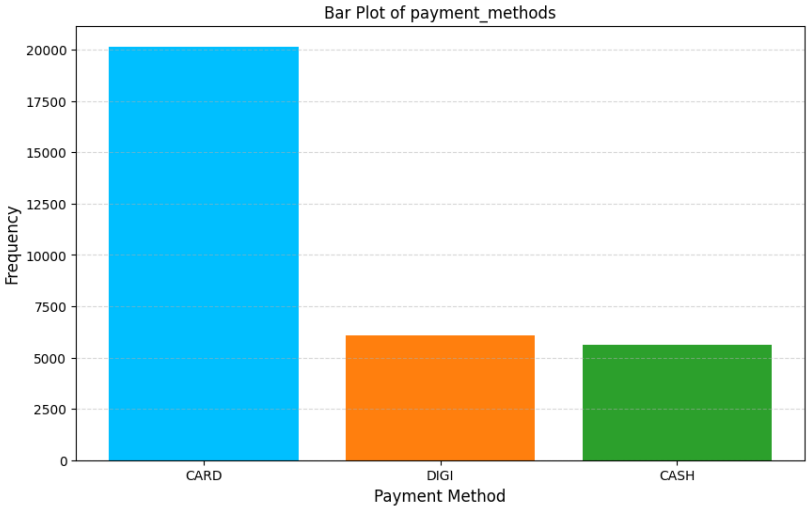

Figure 10 – Bar plot for payment_method

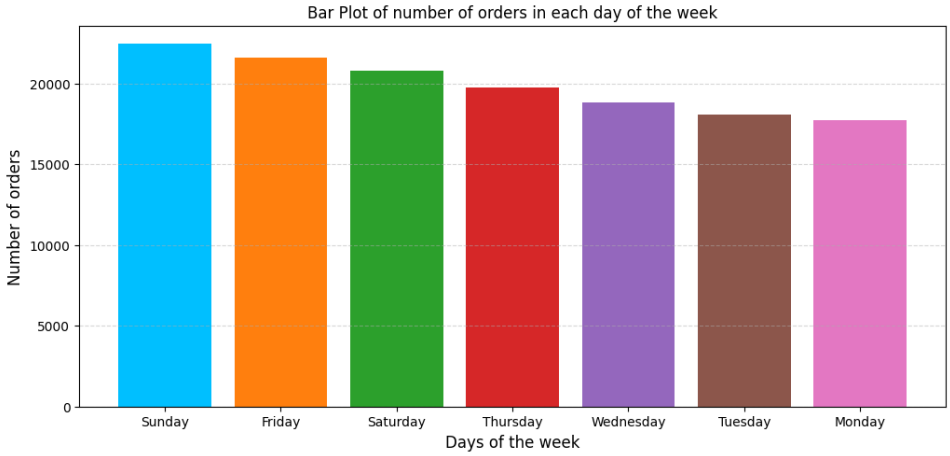

*Figure 11 – Bar plot for DOW_*

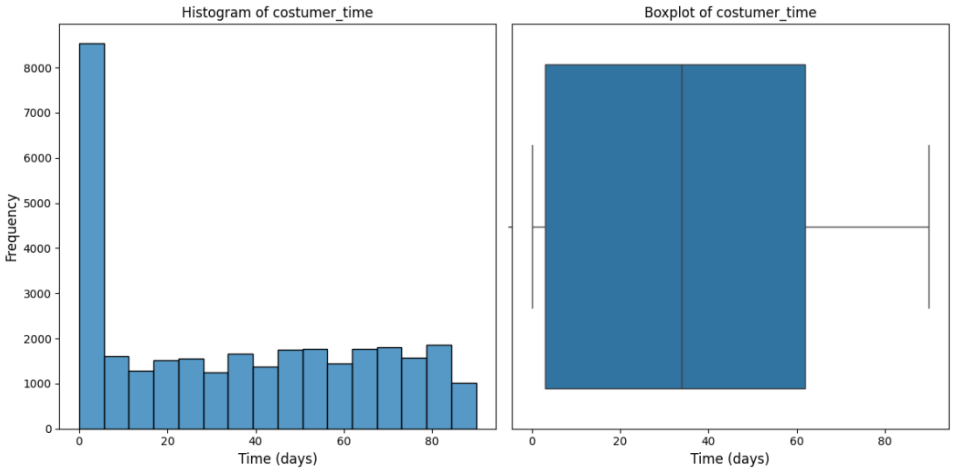

Figure 12 – Plots for customer_time

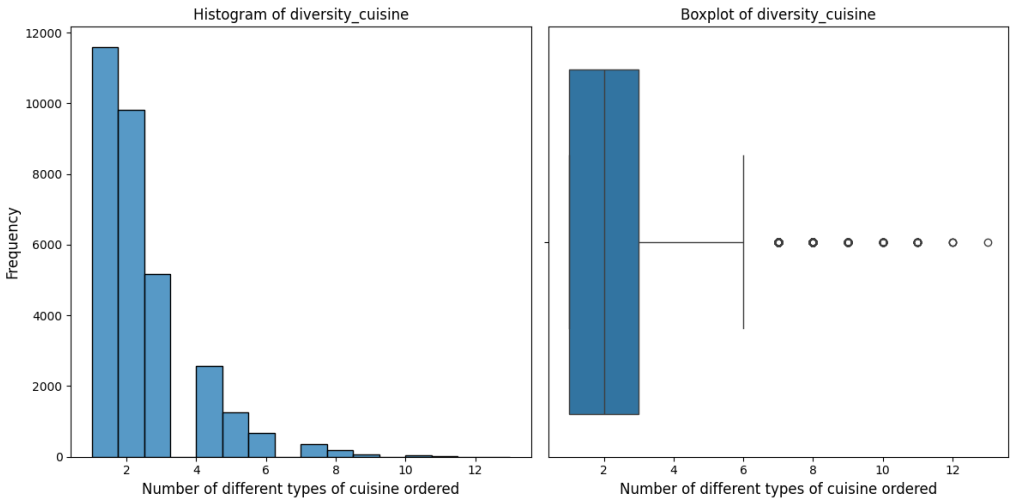

Figure 13 – Plot for diversity_cuisine

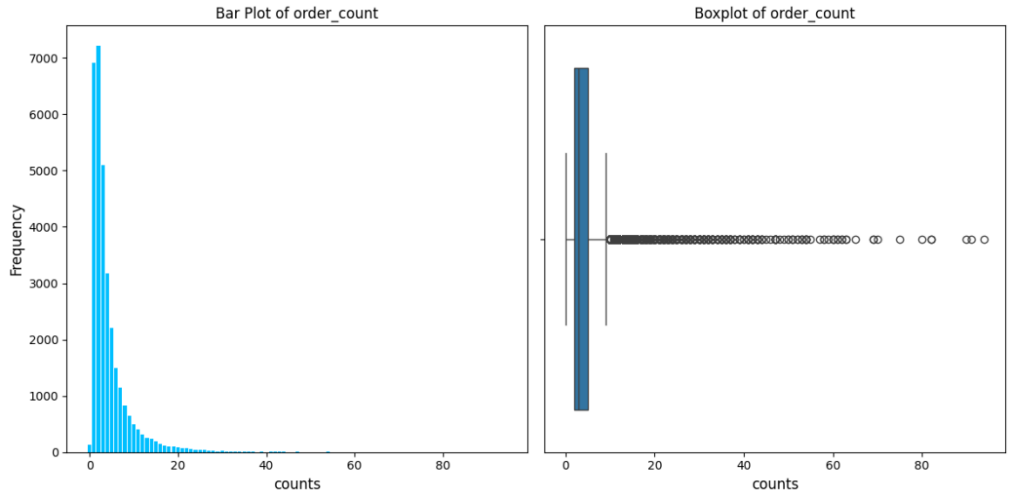

Figure 14 – Plots for order_count

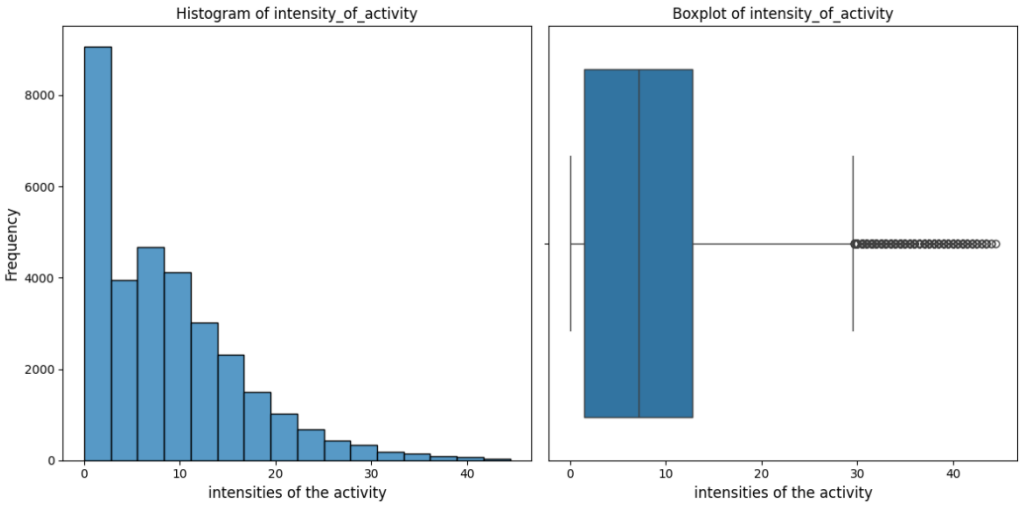

Figure 15 – Plots for instensity_of_activity

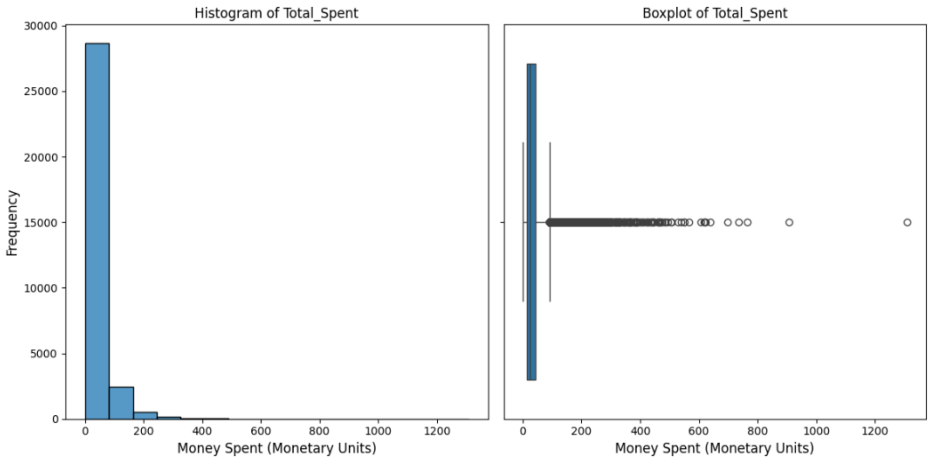

Figure 16 – Plots for Total_Spent

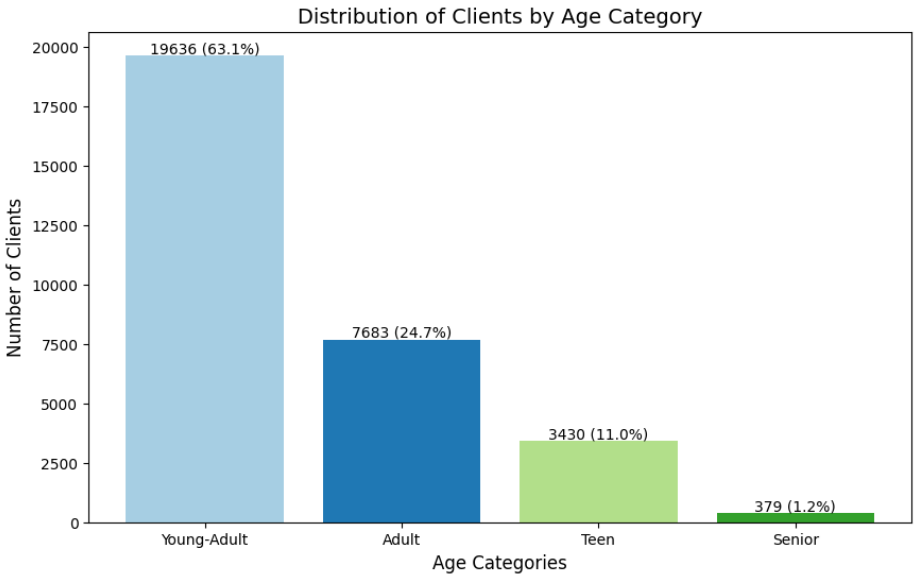

Figure 17 – Bar plot for age_category

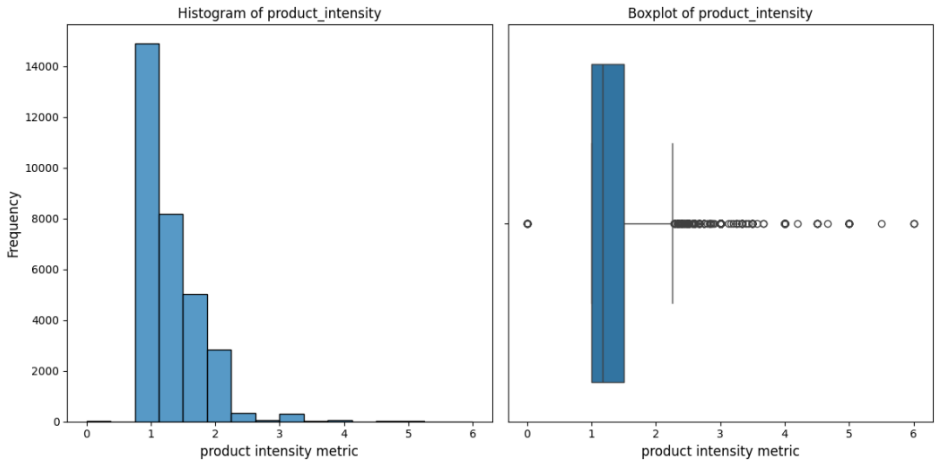

Figure 18 – Plots for product_intensity

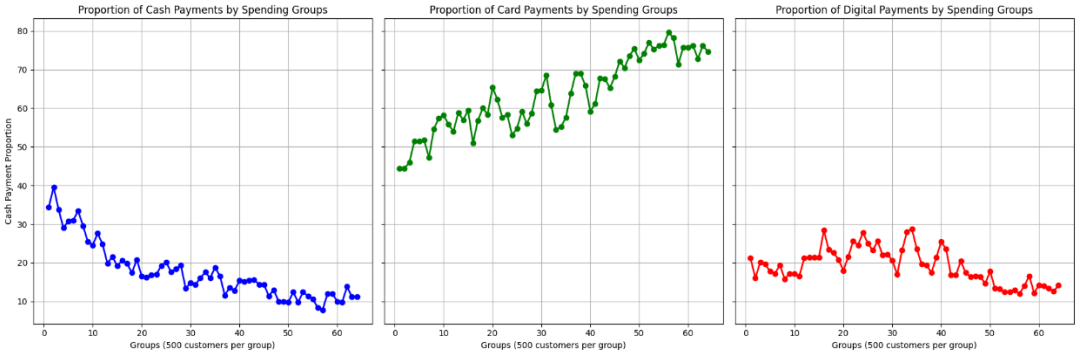

Figure 19 – Relationship between payment_method and Total_Spent

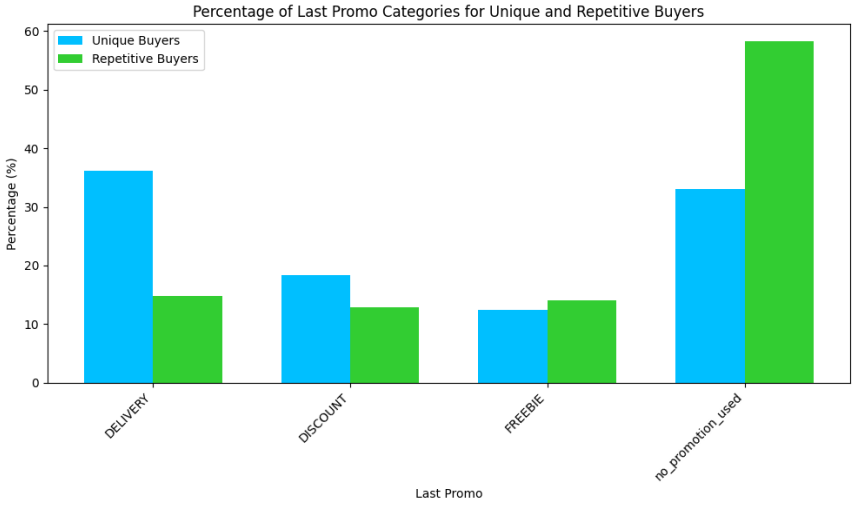

Figure 20 – Relationship between last_promo and customer_time

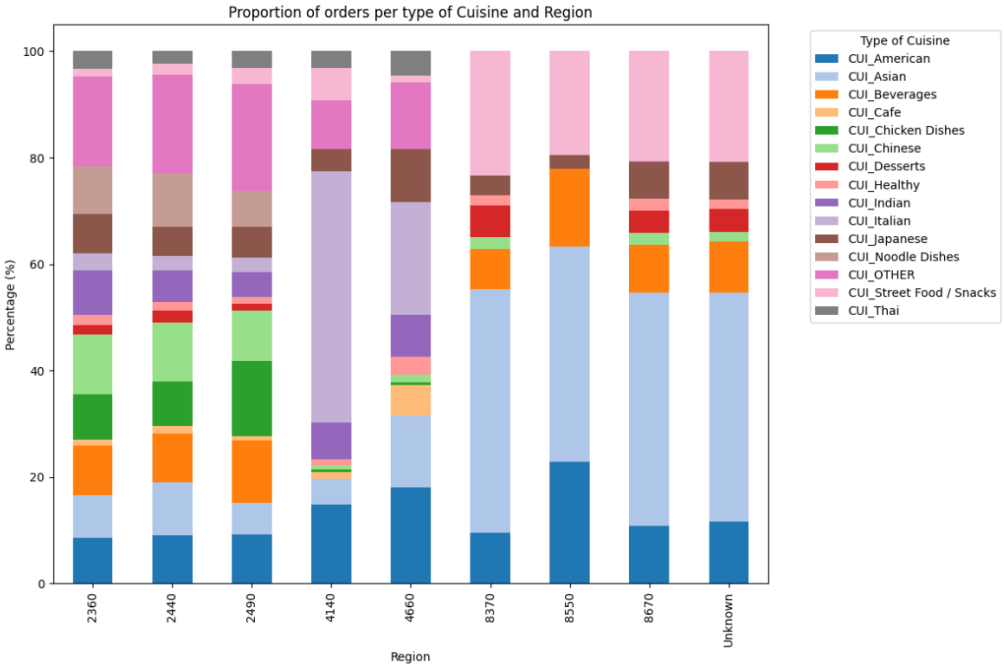

Figure 21 - Relationship between Cui_ and customer_region

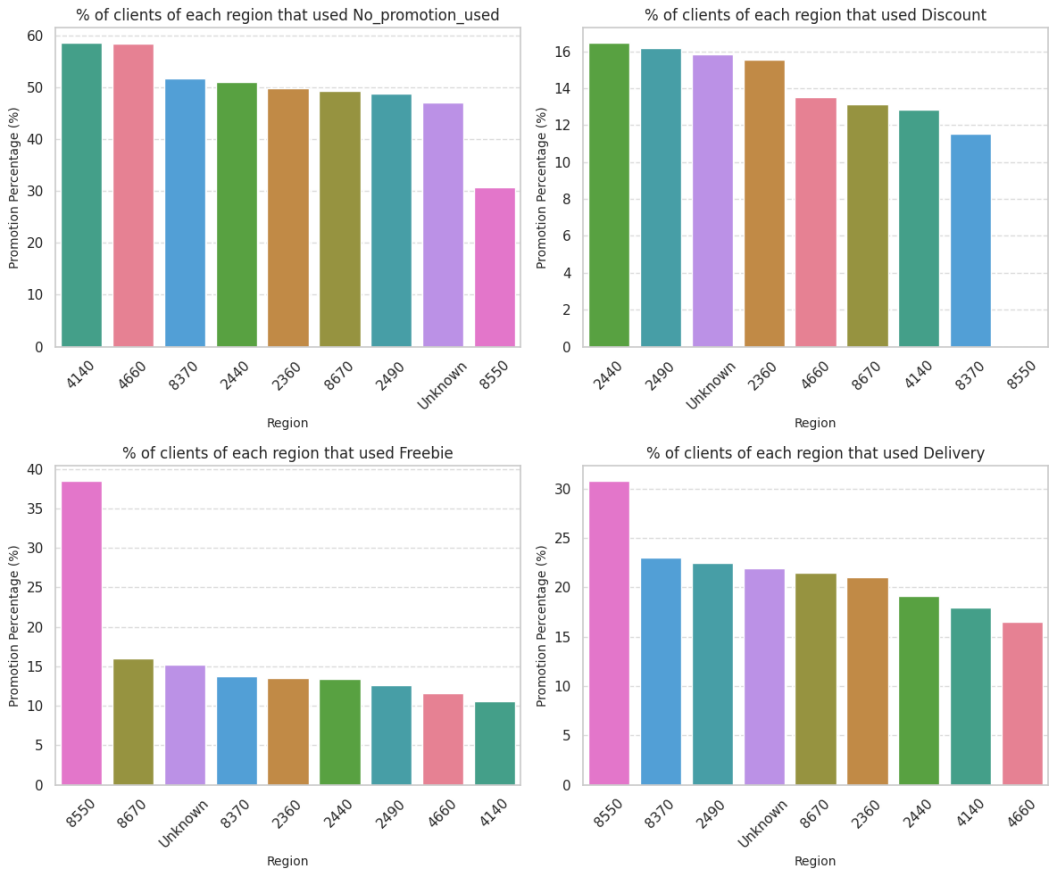

Figure 22 - Relationship between last_promo and customer_region