Data Glacier

Your Deep Learning Partner

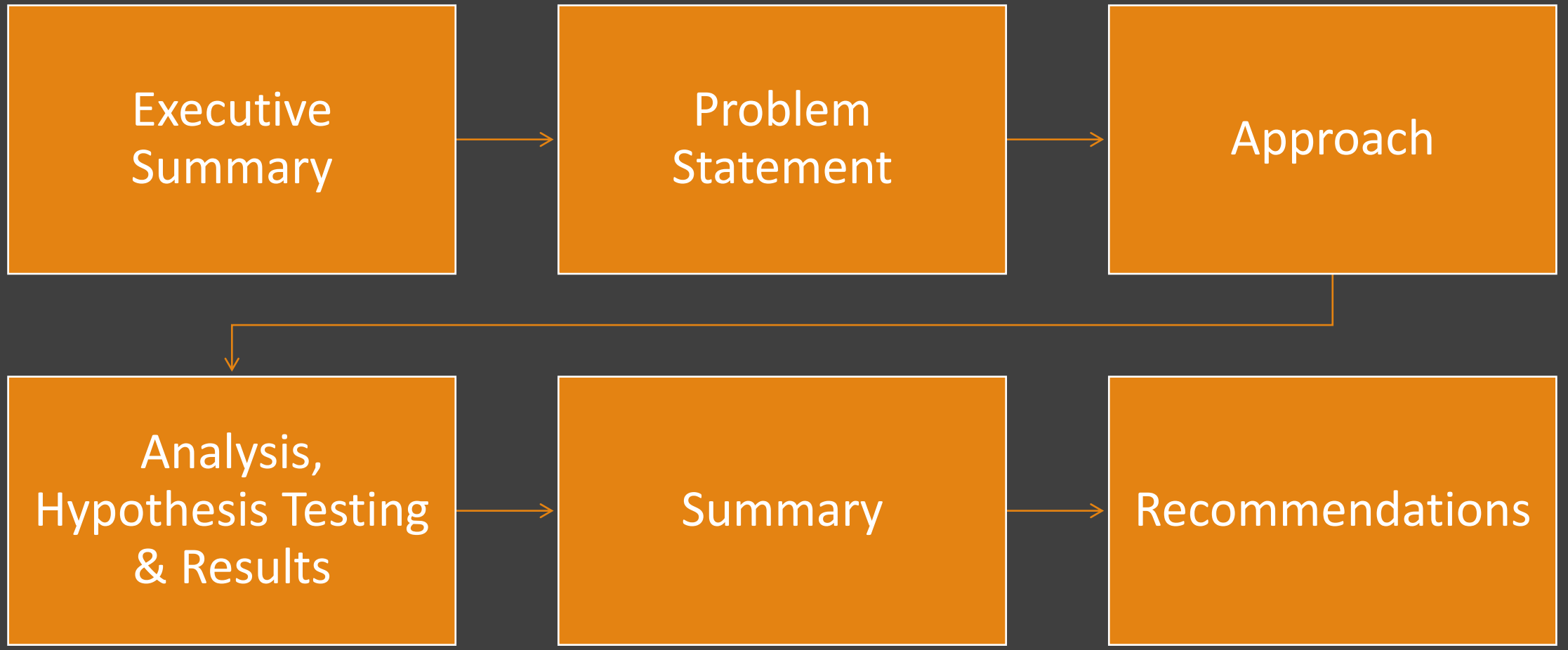
G2M: Cab Investment Analysis

Virtual Internship Case Study

18-Feb-2023

CHUKWUJEKWU JOSEPH EZEMA

AGENDA


Executive Summary
Problem Statement
Approach
Analysis, Hypothesis Testing & Results
Summary
Recommendations

EXECUTIVE SUMMARY

- The objective of this study was to analyze the operations of Pink Cab and Yellow Cab, two ride-hailing services operating in the United States. We performed several hypothesis tests and statistical analyses on a dataset of 359,392 rides to gain insights into the market and help investors make data-driven decisions.

- Our analysis found that Yellow Cab dominates the market, covering 76.43% of trips compared to Pink Cab's 23.57%. We also found that distance traveled was positively correlated with the price charged and profit, suggesting that longer trips generate higher profits. Additionally, we found that cab usage is higher in larger cities and on certain days of the week and months of the year.

- Regarding customer demographics, we found that age group does not have a significant impact on payment preferences, but there is a statistically significant association between age group and gender with respect to the number of cab service transactions.

- Finally, we found that there is a relationship between the number of rides and US holidays, which could potentially be leveraged by cab services to generate more revenue during peak holiday seasons.

- Based on our findings, we recommend that investors looking to invest in the ride-hailing sector consider Yellow Cab as a more profitable option due to its larger market share. Also, we suggest that investors pay close attention to peak demand times, such as holidays, as they could present opportunities for increased revenue.

PROBLEM STATEMENT

The US Cab industry has experienced remarkable growth in the last few years with multiple key players in the market. A private firm, XYZ, is planning to invest in the Cab industry as part of their Go-to-Market (G2M) strategy, and they want to understand the market before making a final decision. The goal of this project is to provide actionable insights using the provided data sets to help XYZ identify the right company to make their investment. The challenge is to analyze the provided datasets, including details of transactions, customer demographics, transaction-to-customer mapping, and the number of cab users in US cities, to gain insights into the US cab industry. The outcome of the project will be a presentation to XYZ's Executive team, which will be evaluated based on the quality of analysis, the value of recommendations and insights, and the quality of the visuals provided.

APPROACH

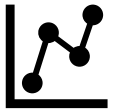
1. **Data Analysis (EDA) with Python**
   - ❖ Data Cleaning and Preparation – pandas, numpy
   - ❖ Data Visualization – seaborn, matplotlib
   - ❖ Comparative Analysis
2. **Hypothesis Testing and Modelling**
   - Test 1 – used T-Test
   - Test 2 to 3 – used Pearson Correlation
   - Test 4 to 5 – applied One-way ANOVA Test
   - Test 6 to 8 – used chi-squared test
3. **Presentation Slide**
   - ➢ Executive Summary
   - ➢ Report

Data Cleaning and Preparation

- Loaded and merged the four datasets into a single DataFrame
- Checked for missing values and duplicates
- Performed necessary data cleaning, including formatting, renaming columns, and converting data types as required
- Created additional columns as needed to support analysis

Data Visualization

Examined the distribution of the key variables, including cab types, cities, payment modes, and time periods.



Performed univariate and bivariate analysis to understand the relationships between variables and how they affect the overall business.

Used appropriate visualization techniques to communicate insights and trends, including scatter plots, line charts, bar charts, and histograms.

Comparative Analysis

- Checked relationships of different variables
- Used the different Cab companies to measure trends
- Made glaring comparisons with so many metrics in respect to the Cab companies

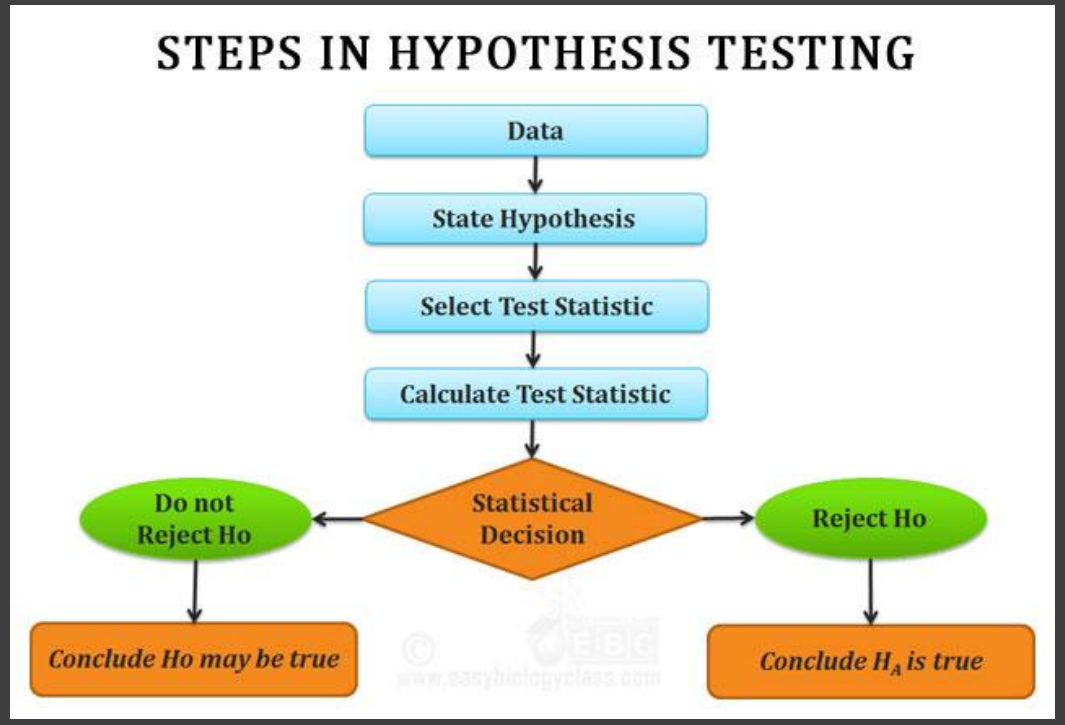
Hypothesis Testing and Modelling

Developed clear hypotheses and conducted statistical tests to validate or reject the hypotheses.

Used appropriate tests such as chi-squared tests, t-tests, and ANOVA, depending on the nature of the hypotheses and the data.

Reported the results of the tests in a clear and concise manner.

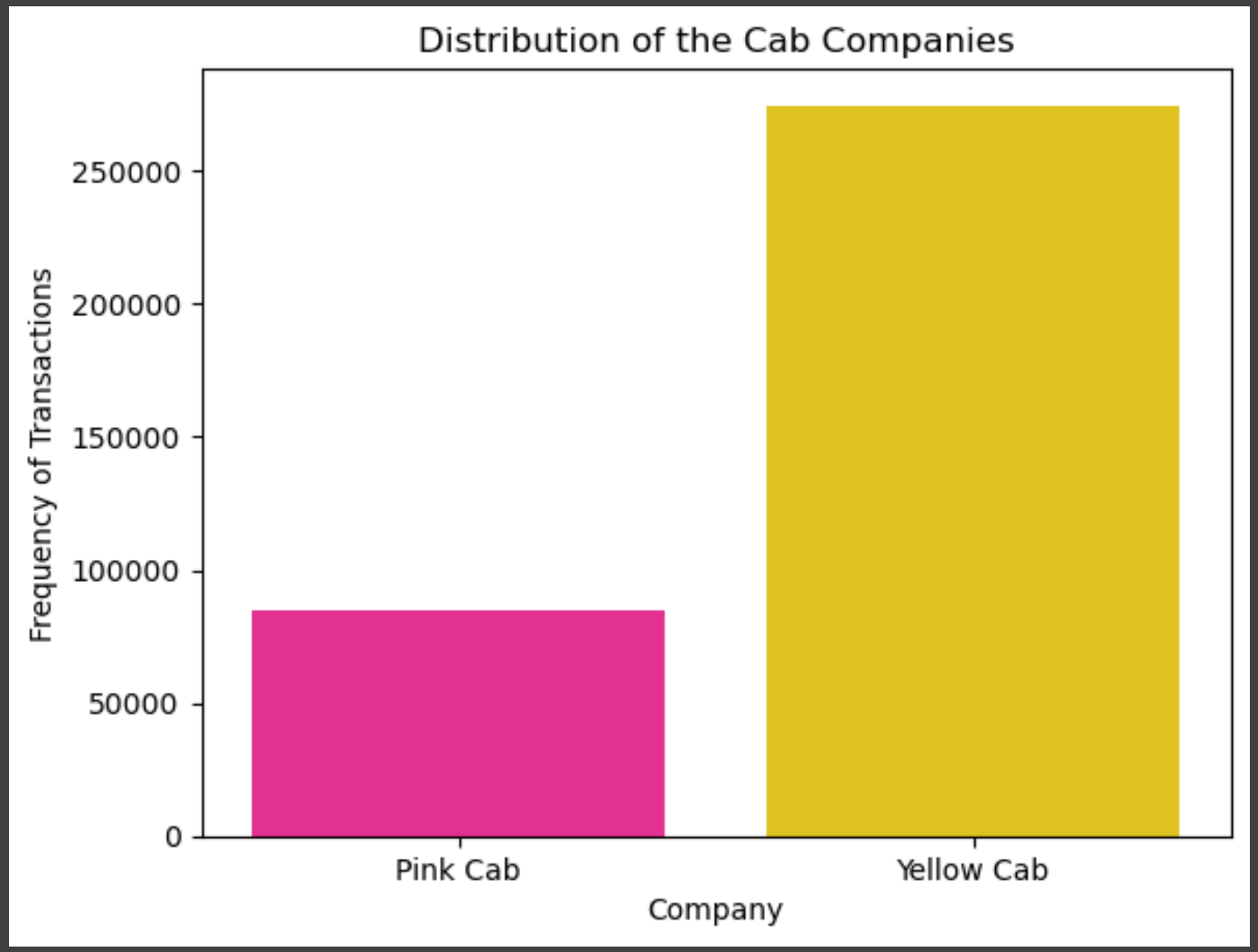
ANALYSIS, HYPOTHESIS TESTING, AND RESULTS

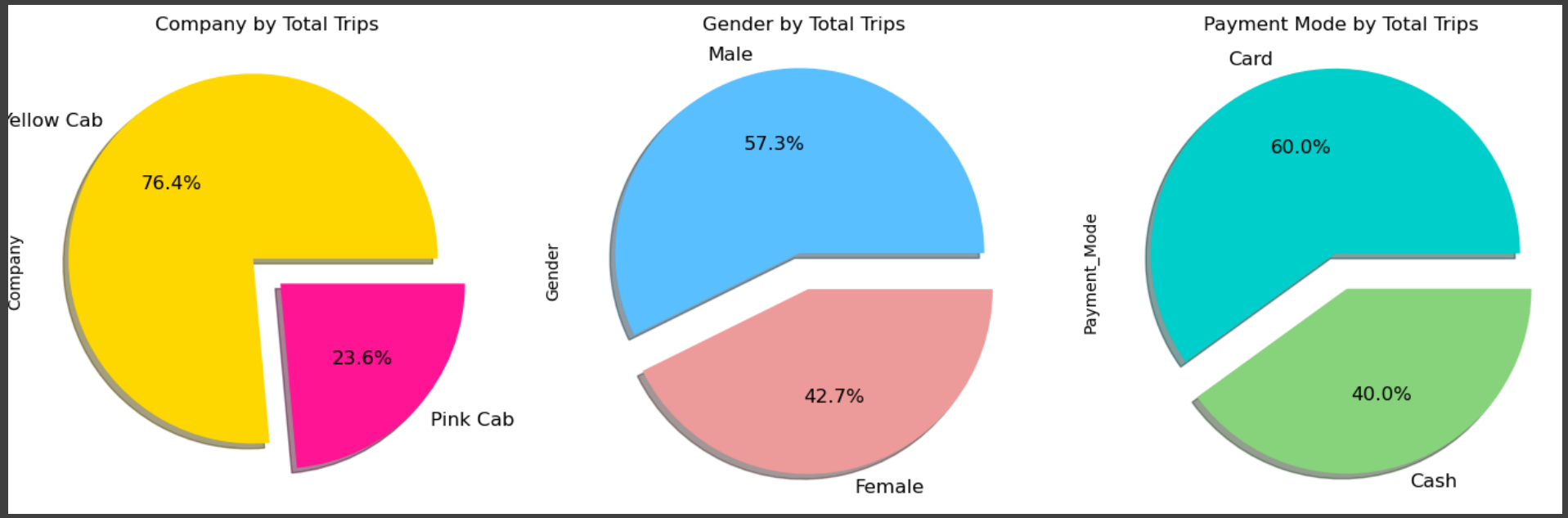


We will look inward into the dynamics of this analysis by investigating customer behaviour and preferences in the ride-sharing service, and identifying potential drivers for increased service demand.

Market Spread of Yellow and Pink Cabs by Trips

- When placed side by side from 359,392 trips, Yellow Cab has more trips covered
- Yellow Cab has 274,681 trips with 76.43% market advantage over Pink Cab which covered 84,711 at 23.57%

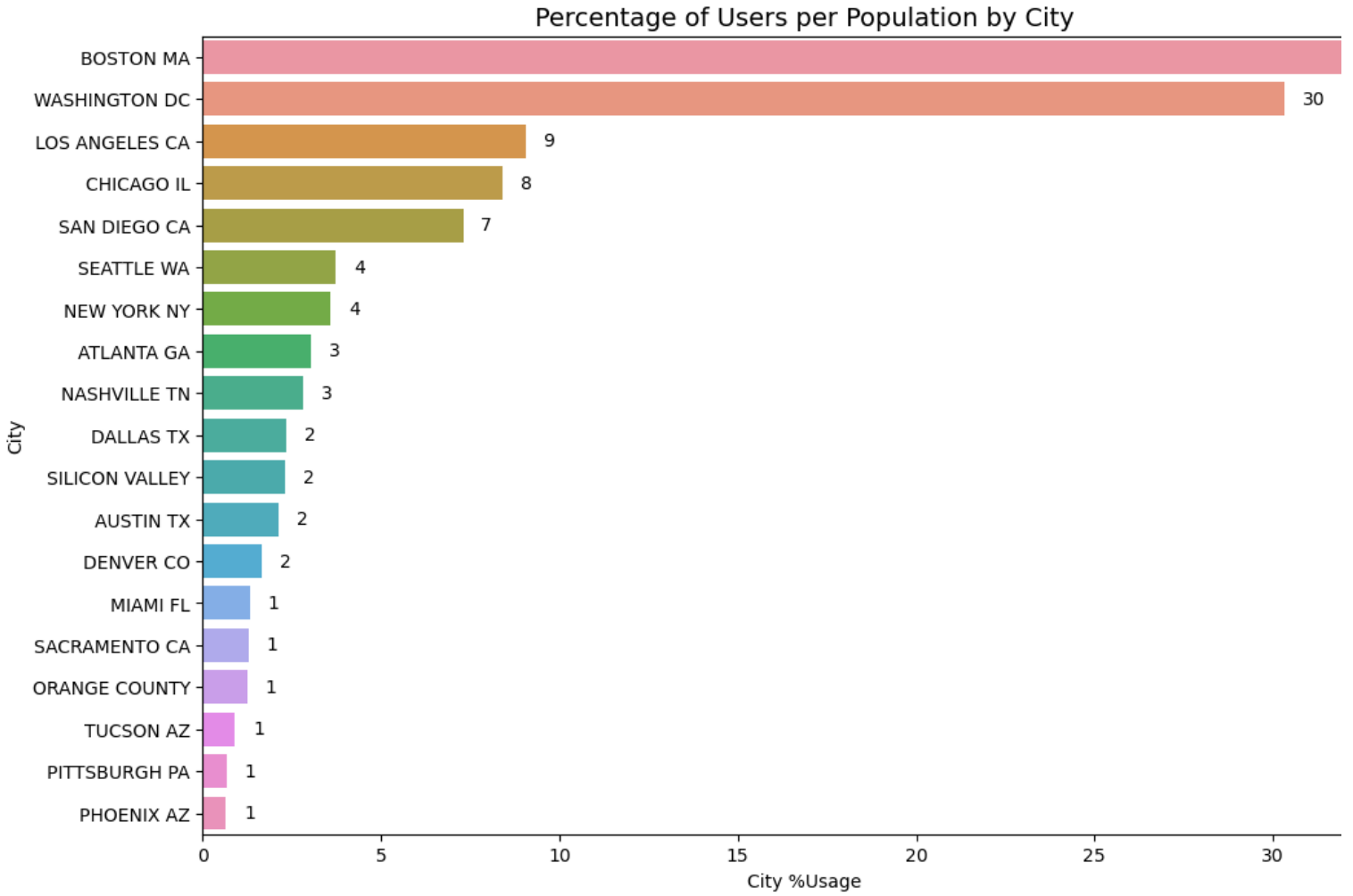


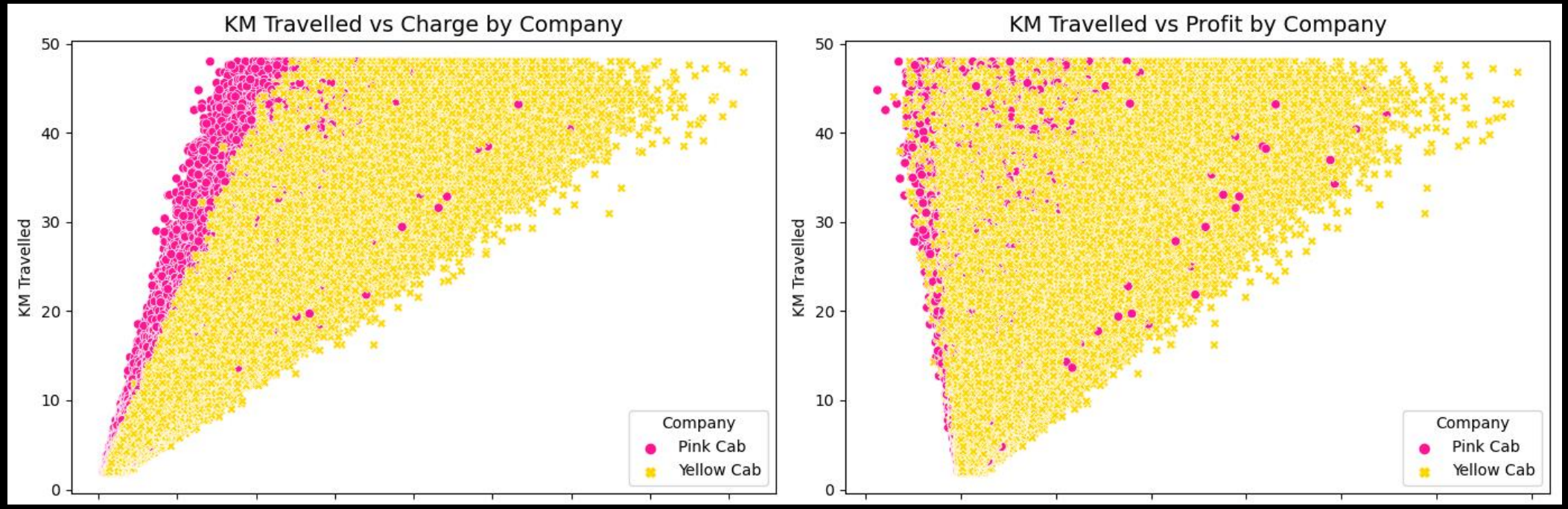


Univariate Analysis by Total Trips

- There is almost equal distribution of gender
- Cards are used more by users.

Percentage of Users per Population





Who controls the Market?

Hypothesis Testing

Is there a difference in average profit between customers who pay with card and those who pay with cash?

Is distance travelled positively correlated with the price charged and profit?

Is cab usage higher in larger cities with more population?

Are there specific days of the week when demand for cab services is higher?

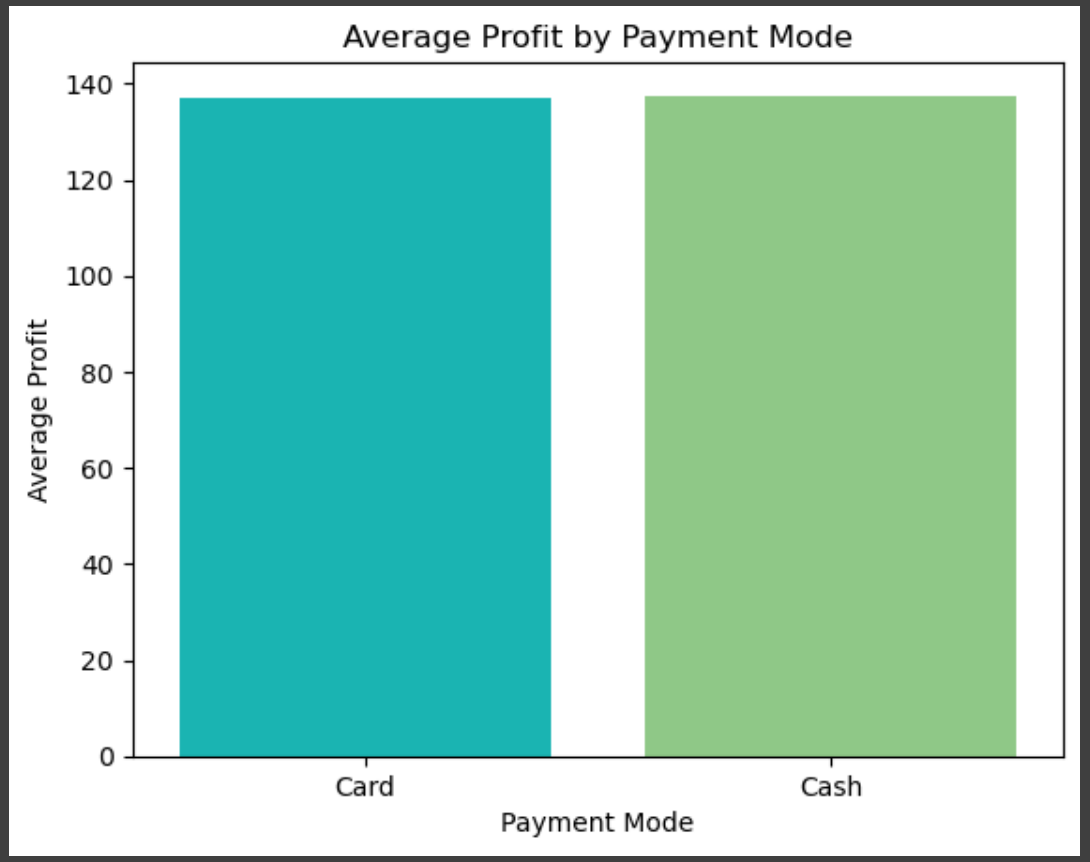
Are there specific Months of the year when demand for cab services is higher?

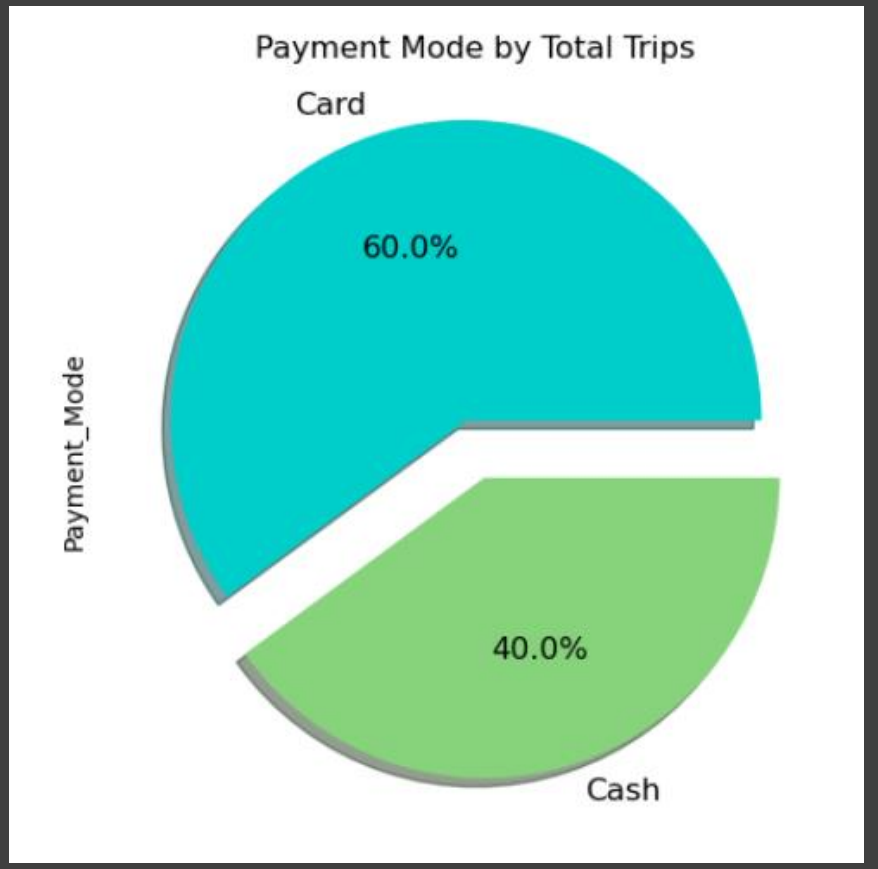
Do customers of different age groups have different payment preferences?

Are there specific customer gender that are more likely to use cab services?

Is there a relationship between the number of rides and the US Holidays?

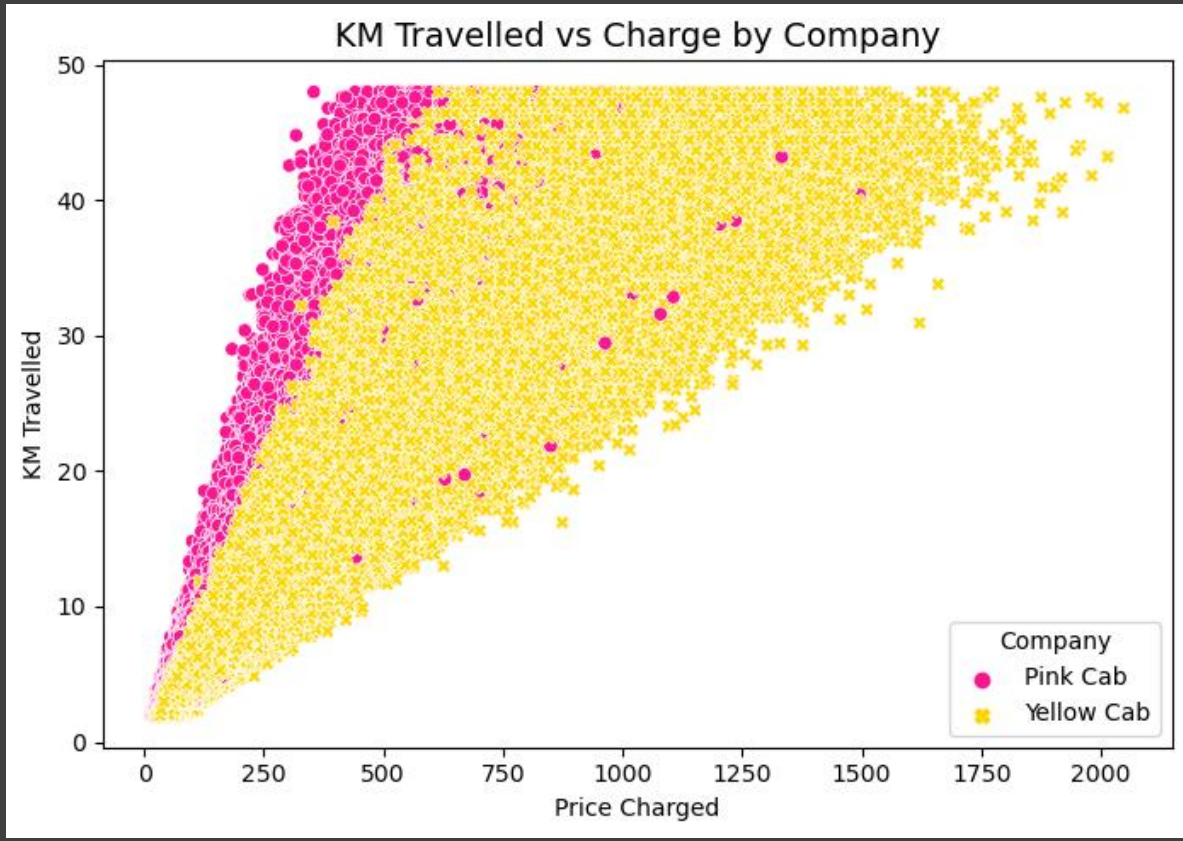
Test 1: Is there a difference in average profit between customers who pay with card and those who pay with cash?

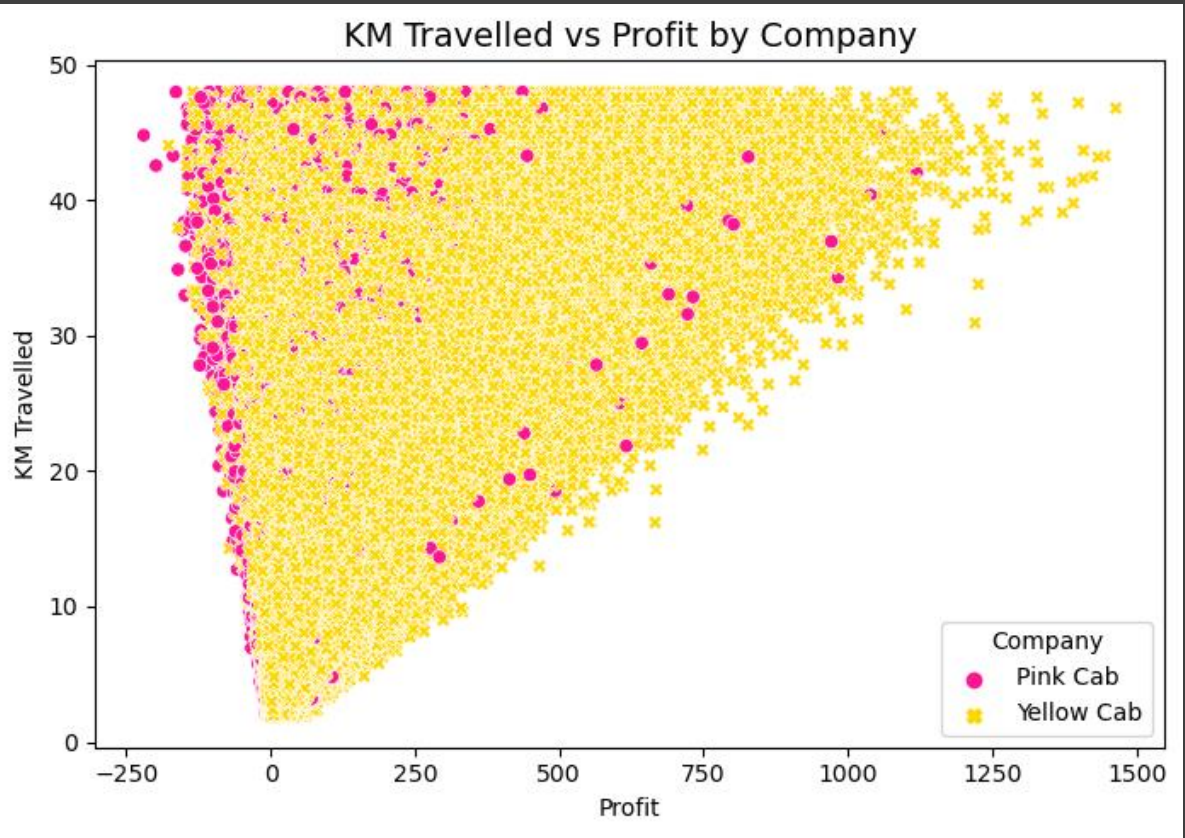




- **Result 1**: We accept the Null Hypothesis (H_0) in Test 1, hence there is no difference in average profit between customers who pay with card and those who pay with cash
- **Reason**: We accept the H_0 because the p-value (0.4460) is NOT less than 0.05 or it's greater than 0.05

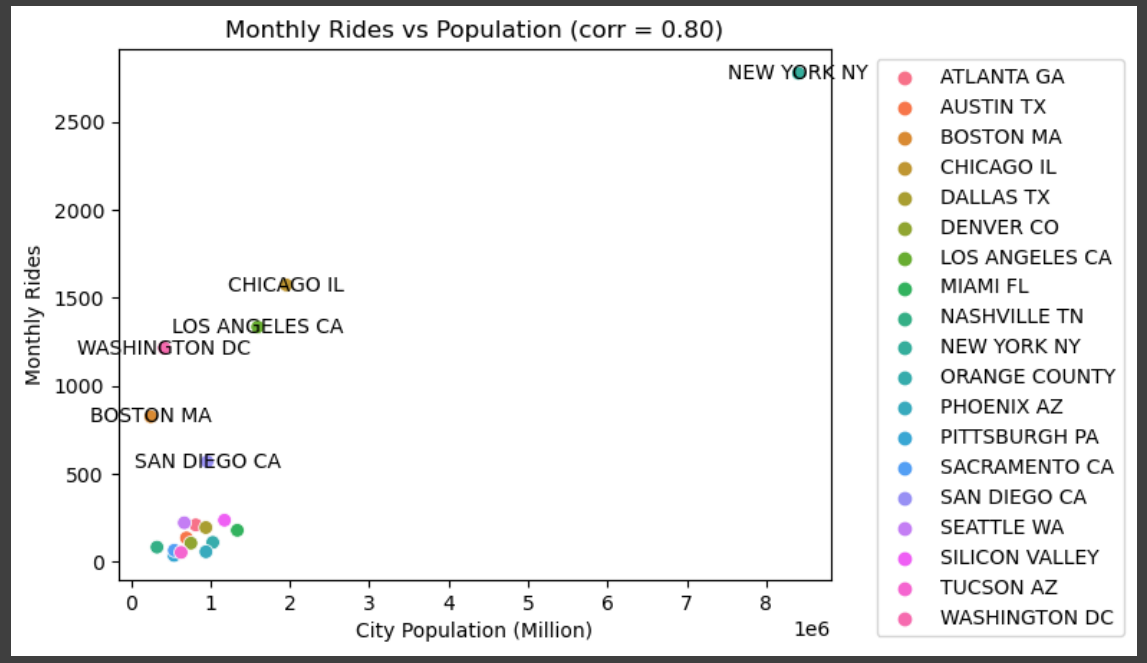
Test 2: Is distance travelled positively correlated with the price charged and profit?

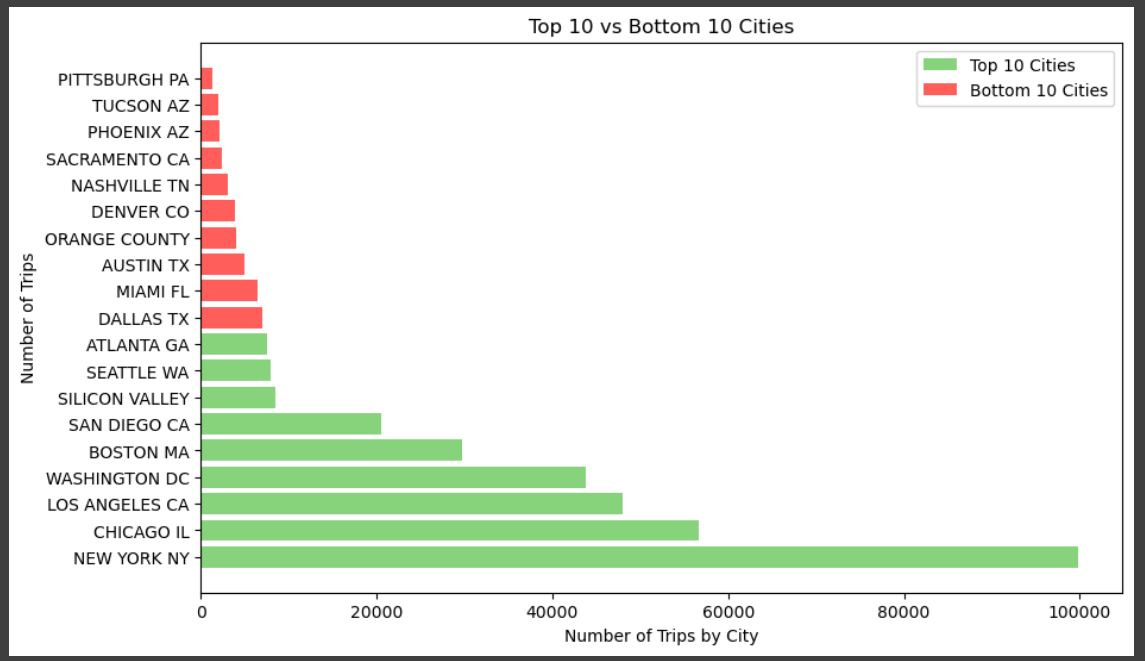




- **Result 2:** We reject the Null Hypothesis (H_0) - which means the distance travelled is positively correlated with the price charged and profit
- **Reason:** The correlation (0.83) is approximately 1 and p-value (0.00) is less than 0.05, hence we reject the null hypothesis and accept the alternative hypothesis. Admittedly, the correlation for the profit is not as strongly correlated with the distance as that of the price, but this doesn't alter our result because the p-value is still less than 0.05.

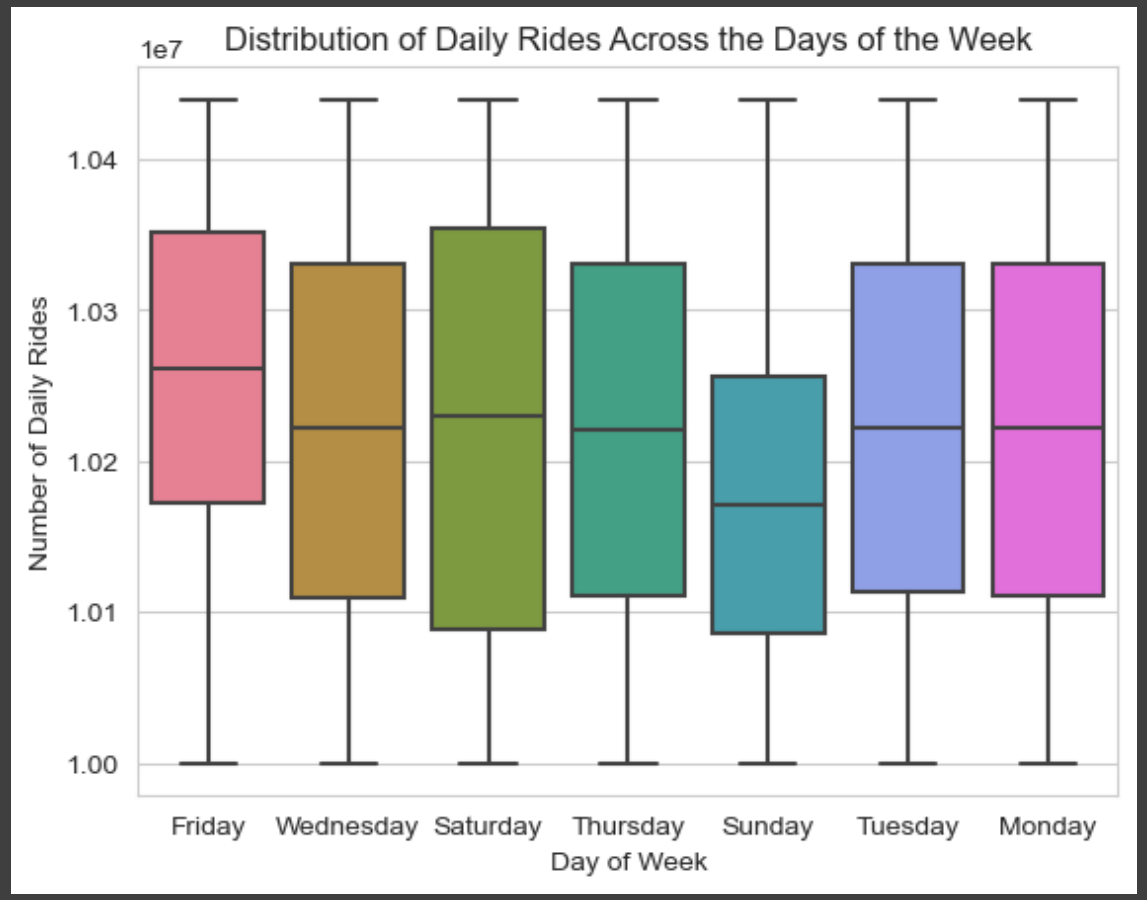
Test 3: Is cab usage higher in larger cities with more population?

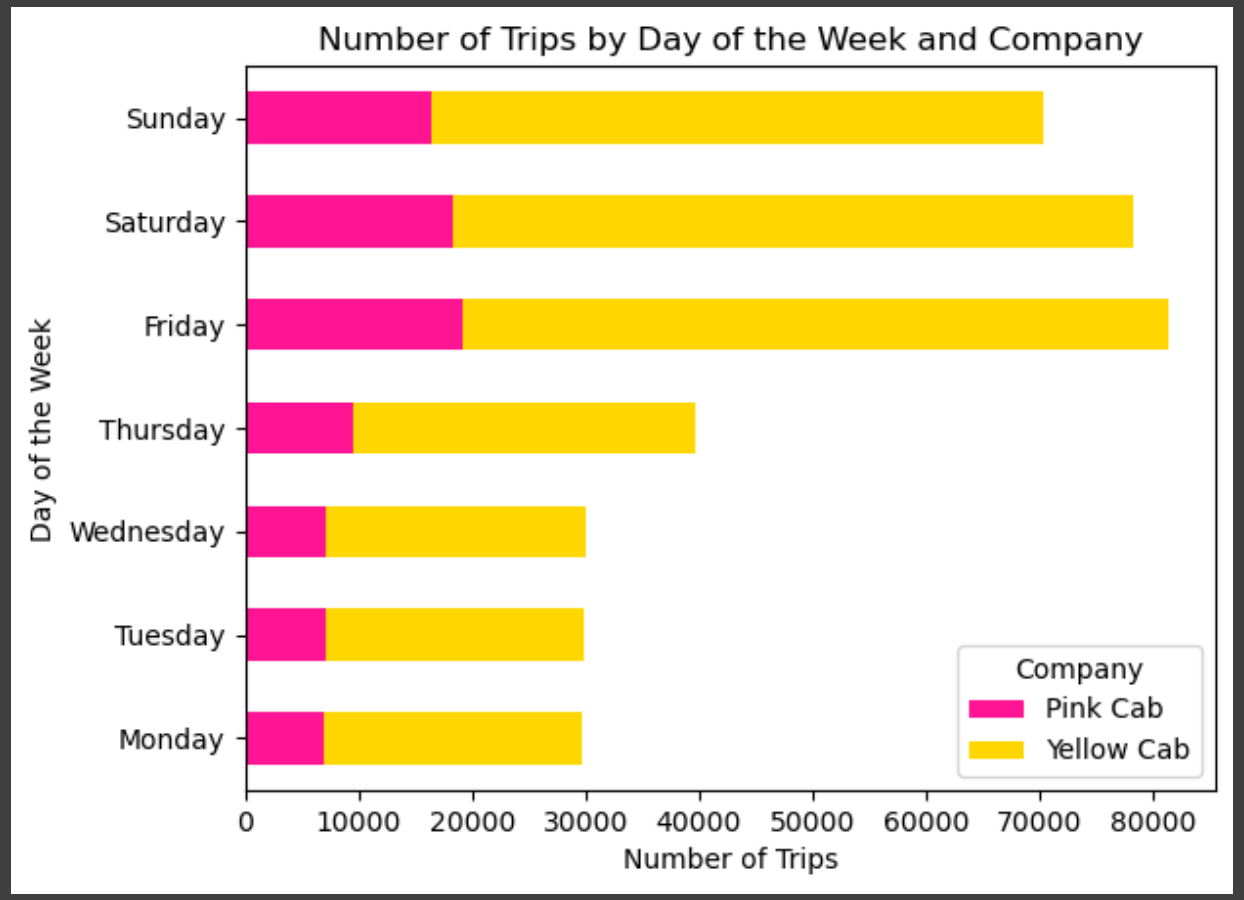




- **Result 3:** We can reject the null hypothesis and conclude that there is a significant correlation between the population of a city and the number of monthly rides.
- **Reason:** Since the test statistic is greater than the critical t-value, and which means it is significant enough, we reject the null hypothesis.

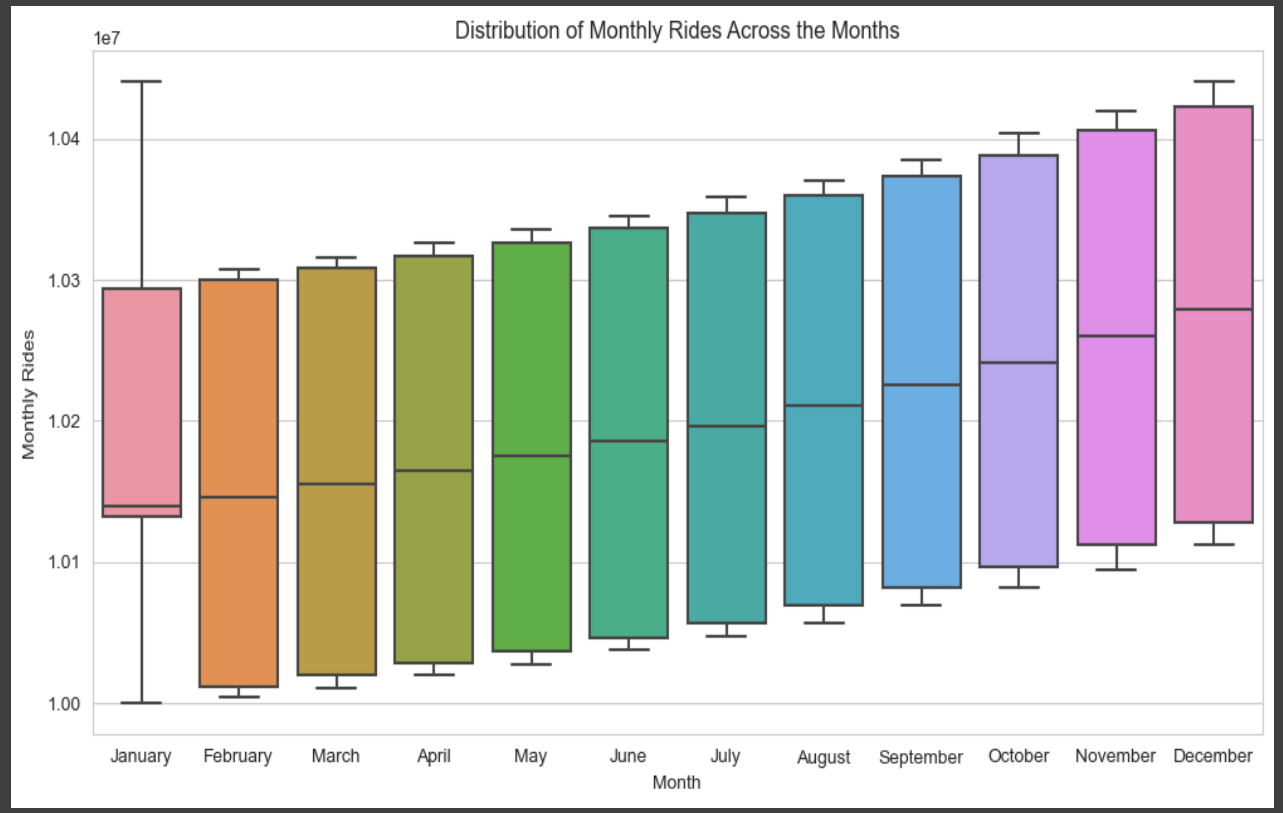
Test 4: Are there specific days of the week when demand for cab services is higher?

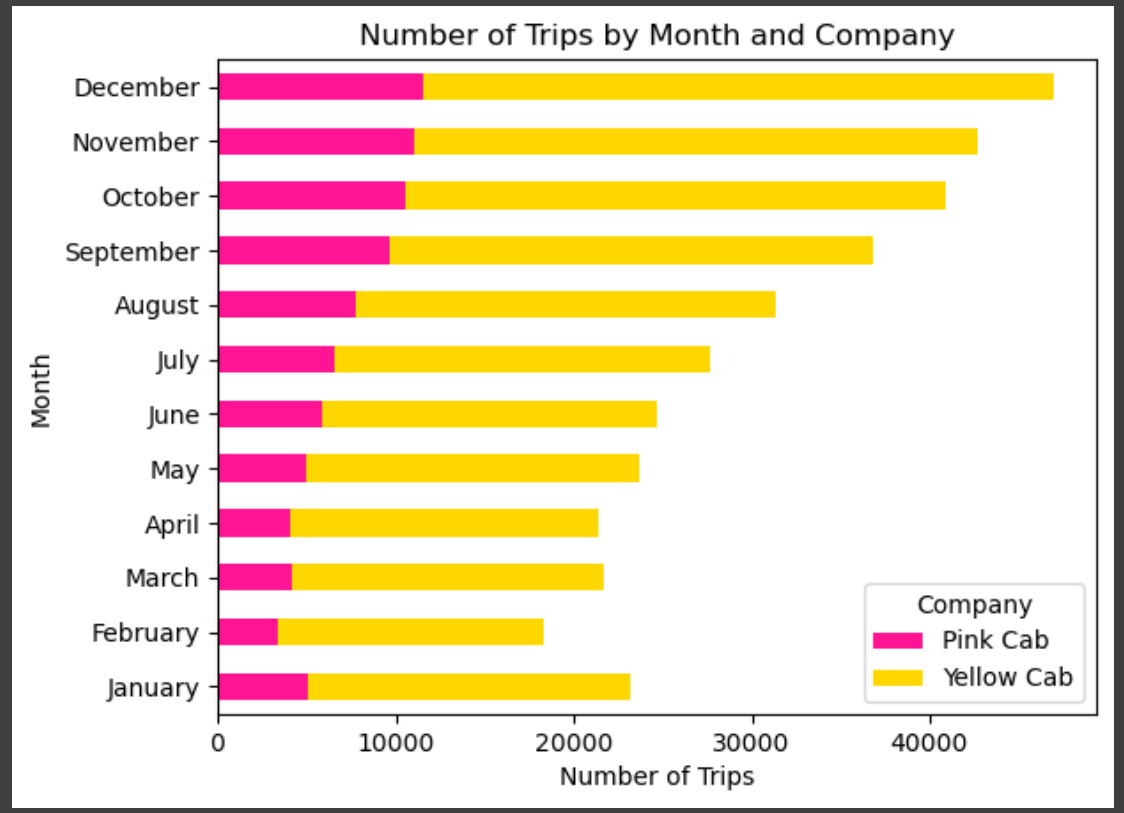




- **Result 4:** We can reject the null hypothesis and conclude that there is a statistically significant difference in the average daily rides across different days of the week
- **Reason:** Since the p-value (PR(>F)) is very small (0.0), this means that we can reject the null hypothesis and say that there specific days of the week when demand for cab services is higher.

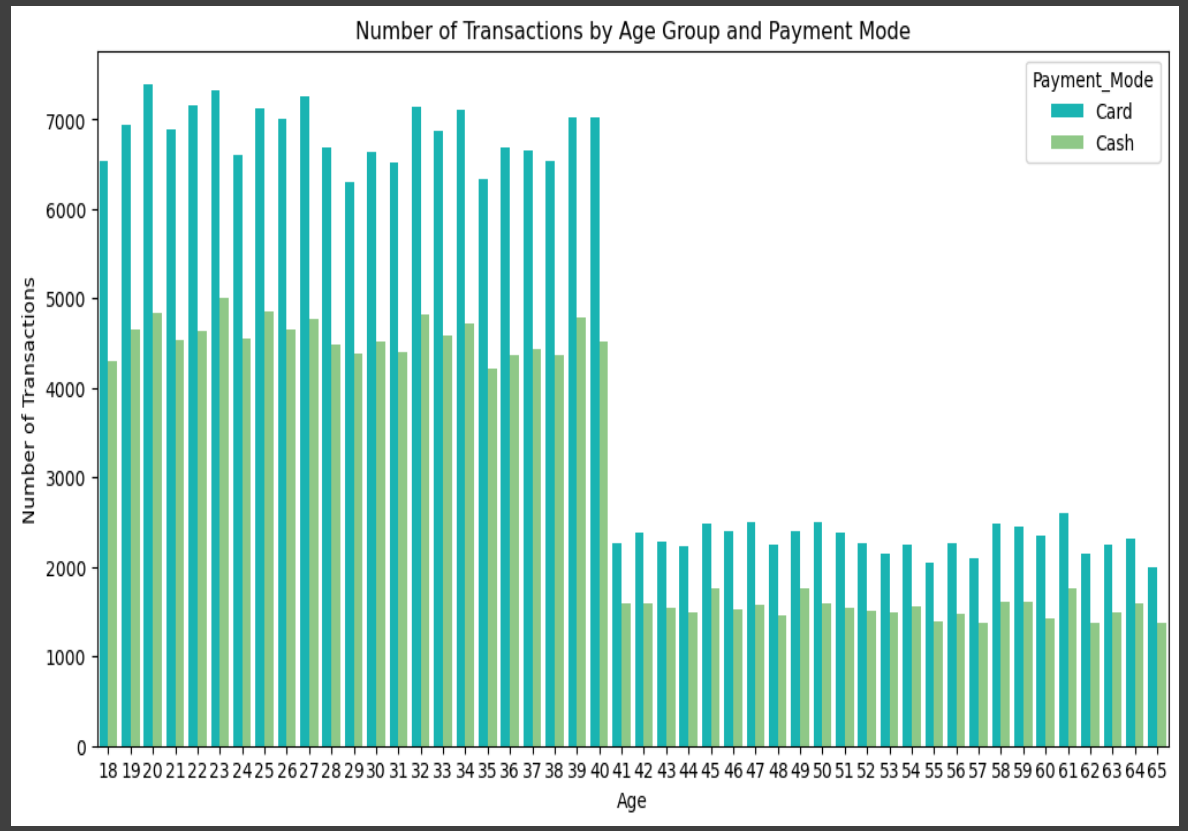
Test 5: Are there specific Months of the year when demand for cab services is higher?

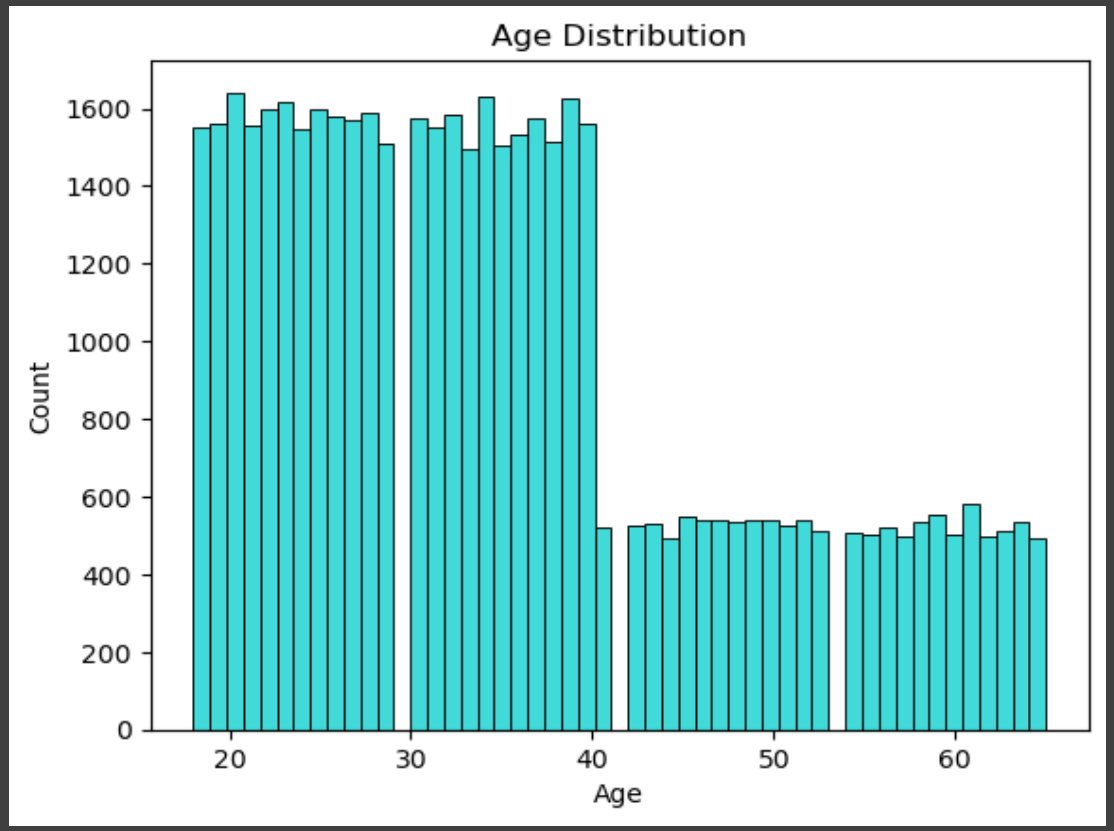




- **Result 5:** We can reject the null hypothesis and conclude that there is a statistically significant difference in the average monthly rides across different Months
- **Reason:** Since the p-value, PR(>F) is very small (0.0), this means that we can reject the null hypothesis and say that there are specific months of the year when demand for cab services is higher.

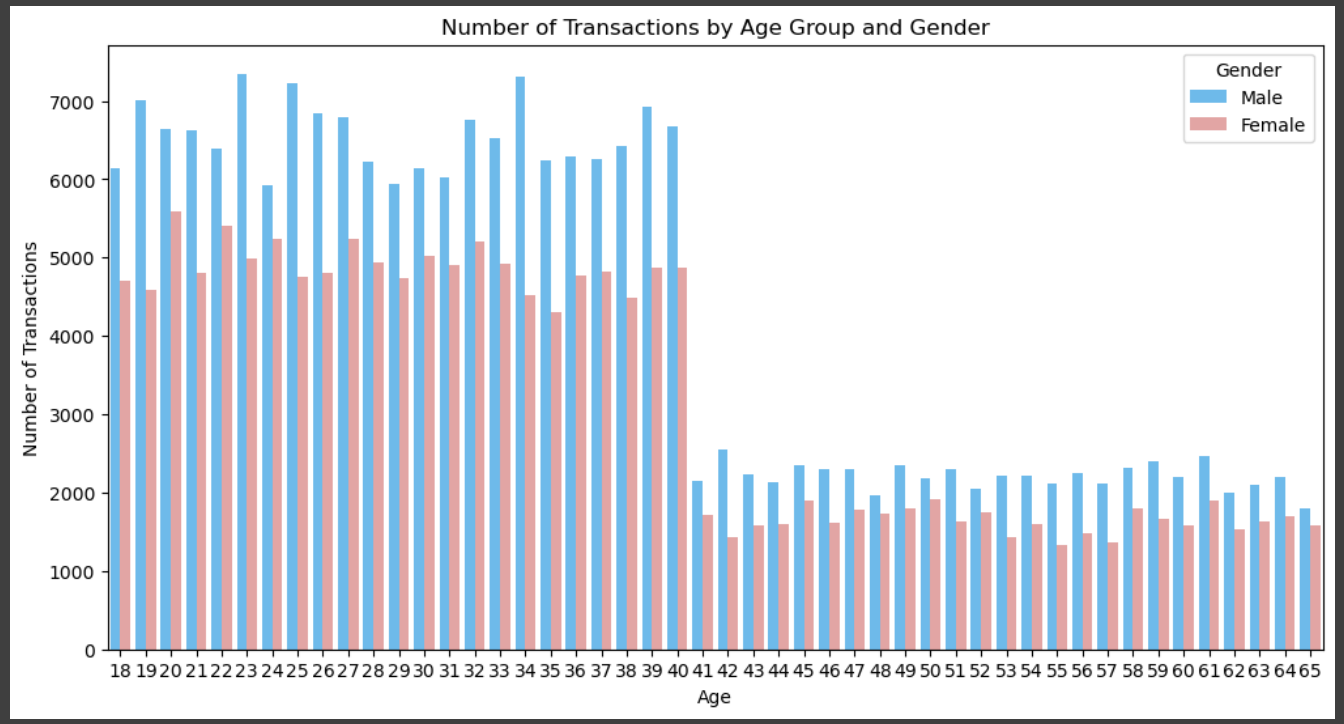
Test 6: Do customers of different age groups have different payment preferences?

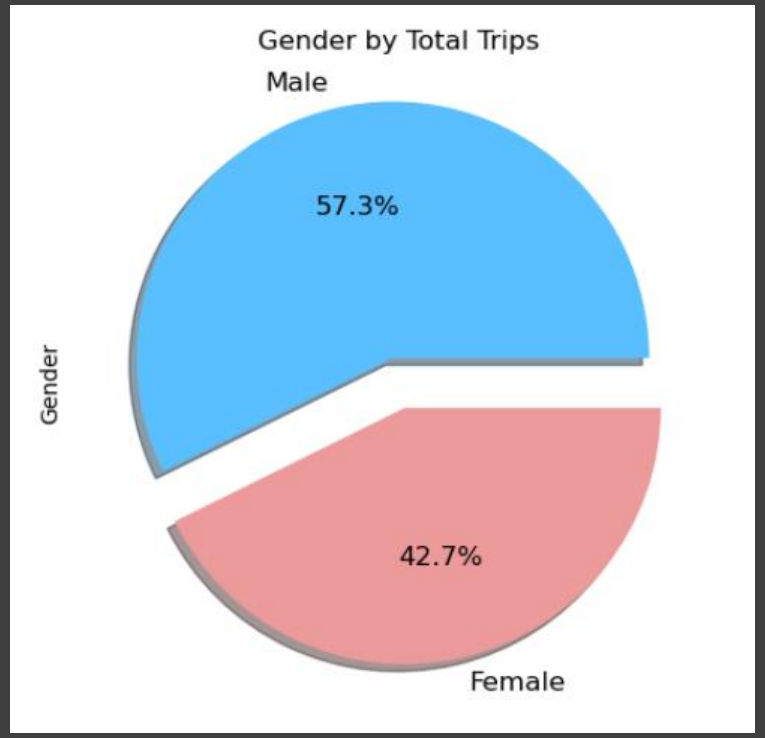




- **Result 6:** The p-value of 0.0561 is slightly above the commonly used significance level of 0.05. This suggests that there is NOT ENOUGH EVIDENCE to reject the null hypothesis that there is no difference in payment preferences between different age groups.

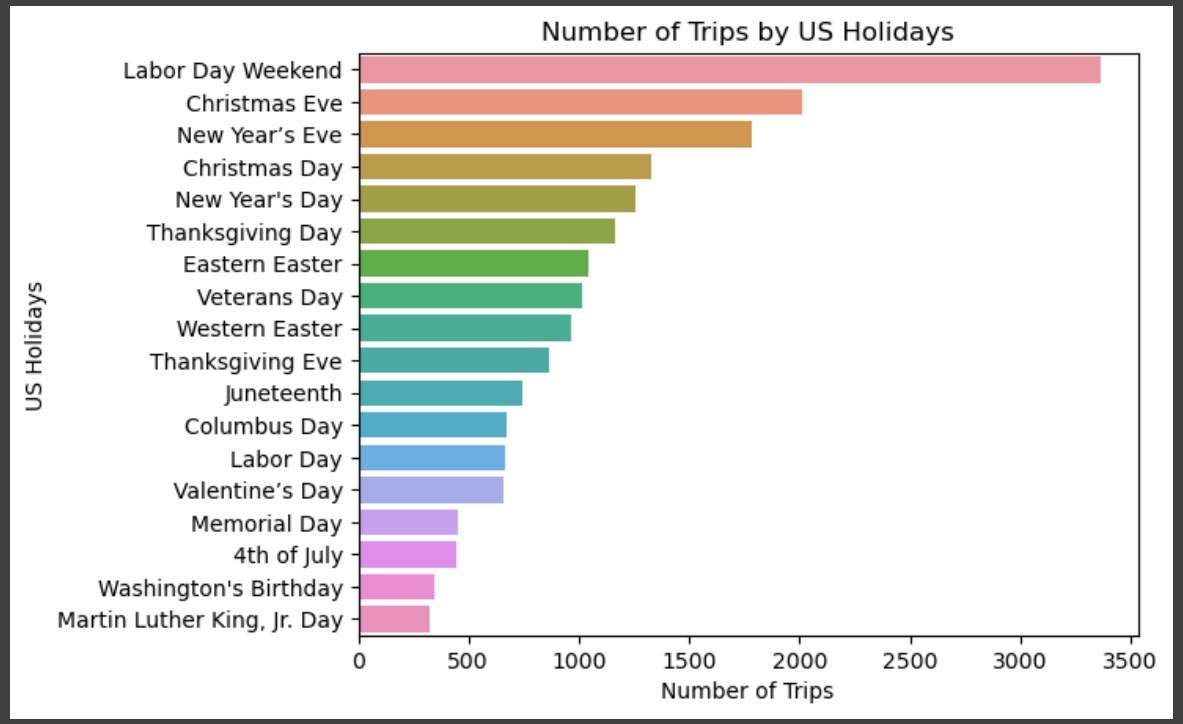
Test 7: Are there specific customer gender that are more likely to use cab services?

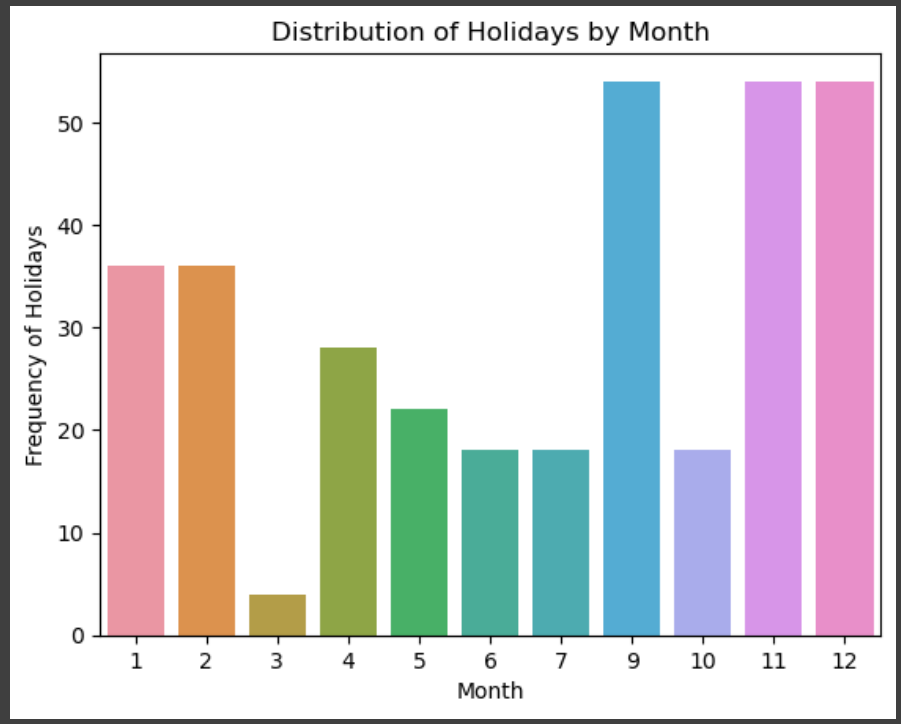




- **Result 7:** The p-value of 4.4576 is highly above the threshold of 0.05. This means that there is a strong evidence to reject the null hypothesis and conclude that there is a statistically significant association between age group and gender with respect to the number of cab service transactions

Test 8: Is there a relationship between the number of rides and the US Holidays?

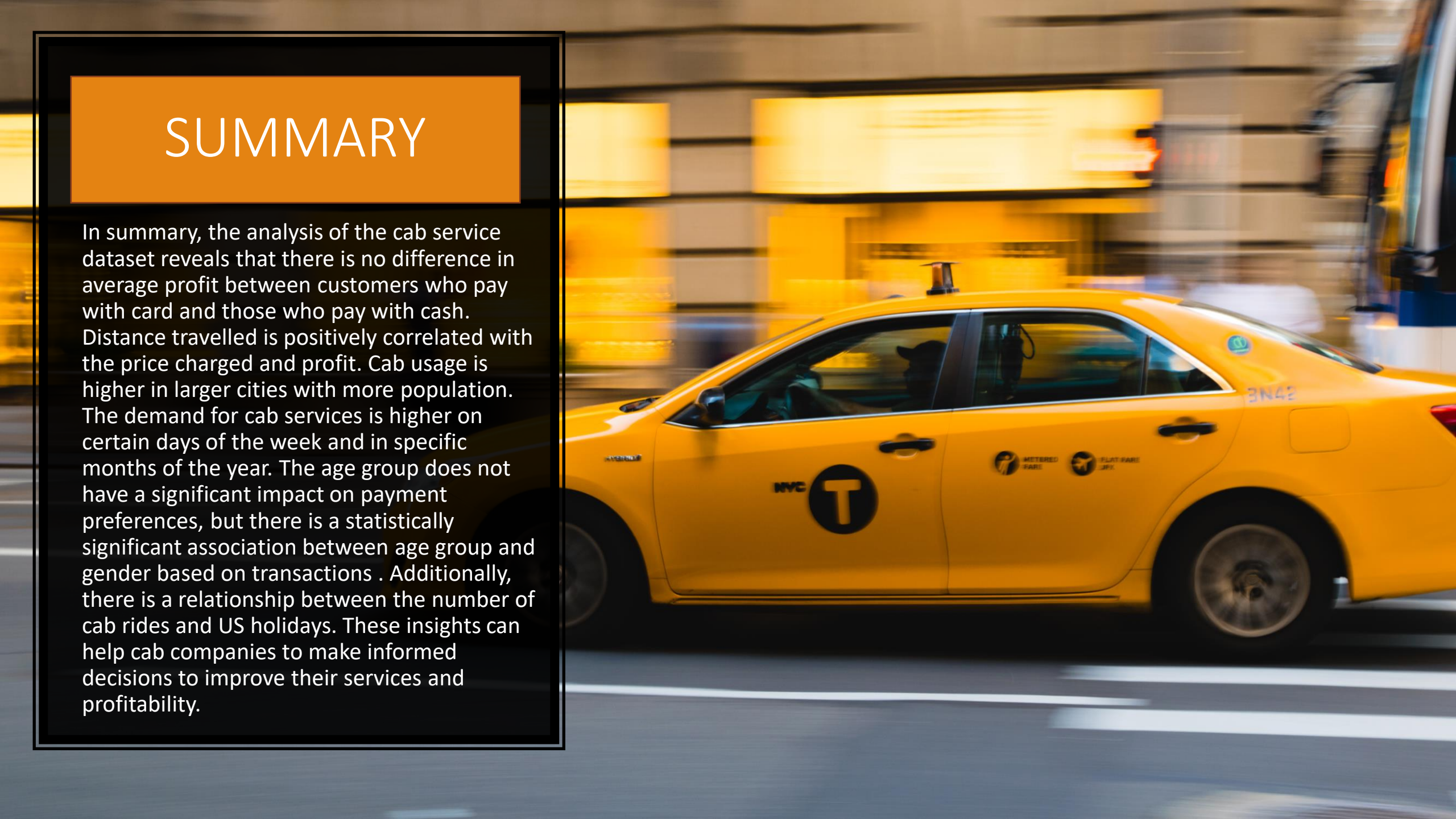




- **Result 8**: The p-value of 1.2341 is well above the determining level of 0.05. This means that there is a strong evidence to reject the null hypothesis and conclude that there is a relationship between the number of rides and US holidays

SUMMARY

In summary, the analysis of the cab service dataset reveals that there is no difference in average profit between customers who pay with card and those who pay with cash. Distance travelled is positively correlated with the price charged and profit. Cab usage is higher in larger cities with more population. The demand for cab services is higher on certain days of the week and in specific months of the year. The age group does not have a significant impact on payment preferences, but there is a statistically significant association between age group and gender based on transactions . Additionally, there is a relationship between the number of cab rides and US holidays. These insights can help cab companies to make informed decisions to improve their services and profitability.



RECOMMENDATIONS

- Based on the analysis conducted, investors looking to invest in the cab industry should consider investing in Yellow Cab as it has a significantly higher market share compared to Pink Cab. Additionally, it was found that cab usage is higher in larger cities with more population, and there are specific days and months when demand for cab services is higher.

- Investors should also consider the fact that distance traveled is positively correlated with the price charged and profit, which means that optimizing routes and increasing the number of trips can lead to higher profits. Furthermore, it was found that there are specific customer demographics that are more likely to use cab services, such as age group and gender. This information can be useful in developing targeted marketing strategies to increase customer acquisition and retention.

- Overall, investing in the cab industry can be a profitable venture, especially when the market share of a particular company is taken into consideration. However, it is important to pay attention to market trends and customer preferences to stay competitive and relevant in the industry.

Thank You!

Data Glacier
Your Deep Learning Partner