Data Glacier

Your Deep Learning Partner

Exploratory Data Analysis

- G2M insight for Cab Investment firm

Nov. 2023

Agenda

Executive Summary

Problem Statement

Approach

EDA

EDA Summary

Recommendations

Executive Summary

- The purpose of this project is to provide actionable insights to XYZ, a private firm in the US, to support their investment decision in the cab industry.
- The project will involve the analysis of four datasets, including Cab_Data.csv, Customer_ID.csv, Transaction_ID.csv, and City.csv, covering the time period from 31/01/2016 to 31/12/2018.
- The analysis will focus on identifying the right company for investment based on market trends, customer behavior, and company performance.

Problem Statement

- XYZ is planning to invest in the cab industry, and they need to identify the company that presents the best investment opportunity.
- The key challenge is to analyze the provided datasets and external data sources to understand the market dynamics, customer segments, and company performance.
- The project aims to answer the following questions: Which company has the maximum cab users at a particular time-period? Does the margin proportionally increase with an increase in the number of customers? What are the attributes of these customer segments?

Approach

The project will follow a structured approach to provide XYZ with the necessary information to make an informed investment decision in the cab industry.

The approach includes the following steps:

1. **Data Understanding**: Review the field names and data types in the provided datasets. Identify relationships across the files and perform field/feature transformations.
2. **Hypothesis Generation**: Create 5 initial assumptions to investigate, such as seasonality in the number of customers using the cab service or the relationship between margin and the number of customers.
3. **Data Preparation**: Clean the data, handle missing values, and detect outliers. Create a master data set by joining or appending relevant files.
4. **Analysis**: Use various analytical techniques, such as data visualization, to explore the data and test the hypotheses. Incorporate external data, such as US holiday data or weather data, to enrich the analysis.
5. **Presentation**: Prepare a presentation that summarizes the analysis, recommendations, and insights. Identify which company is performing better and represents a better investment opportunity for XYZ.

EDA

For the four CSV files provided (Cab_Data.csv, Customer_ID.csv, Transaction_ID.csv, and City.csv), the EDA process can be structured as follows:

1. **Data Understanding**: Begin by loading the data from the CSV files into a suitable data structure. Understand the structure of the data, including the variables, data types, and any missing values.
2. **Missing Values and Outliers**: Identify and handle any missing values in the datasets. Additionally, detect and address any outliers that may impact the analysis.
3. **Data Visualization**: Create visualizations, such as histograms, box plots, and scatter plots, to explore the distribution of the data and the relationships between variables. This can help identify patterns, trends, and potential outliers in the data.

EDA Summary

1. **Data Understanding:**
    a. Importing libraries for EDA
    b. Loading datasets
    c. Checking types: date -> integer, population and users -> object
    d. Rename columns: Change long name to short, or replace space to '_'
    e. Change formats: date -> yyyy/mm/dd, population and users -> integer

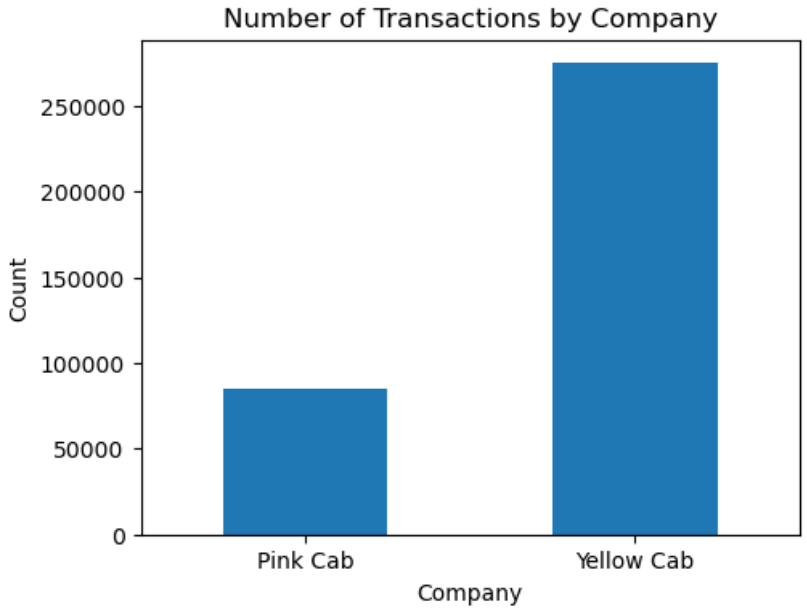
2. **Missing Values and Outliers**
    a. Check duplicate rows: None
    b. Check missing or null values: None

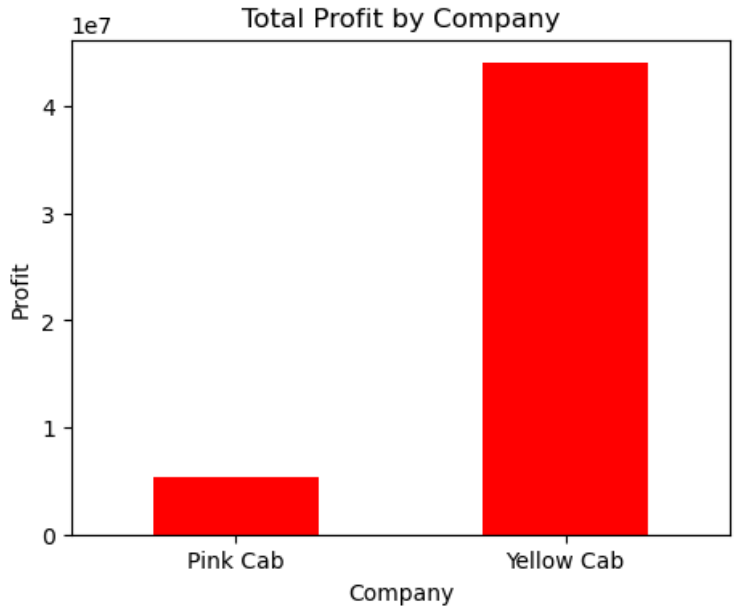
EDA Summary (cont.)

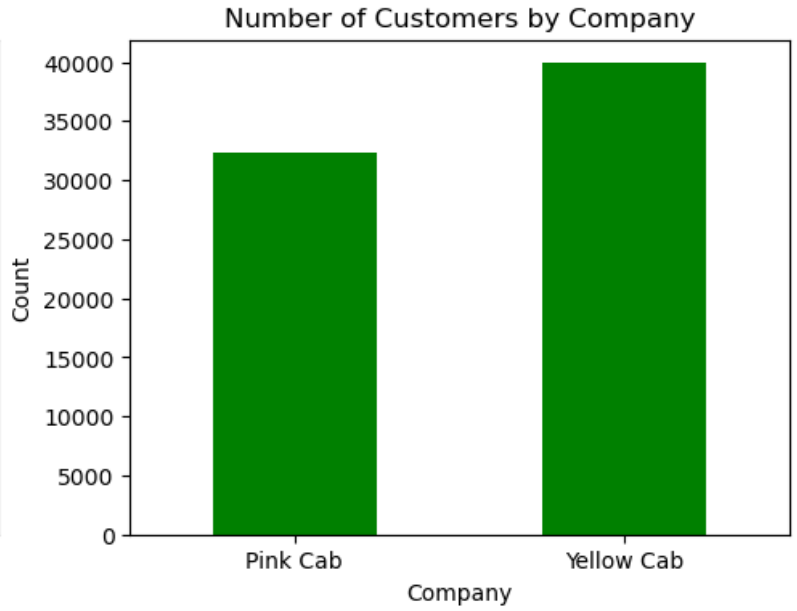
3. Data Visualization :

a. Market share

In terms of the number of transactions and revenue, Yellow Cab holds a higher market share. Given that the number of unique customers over the three years is not significantly different, we can assume that Yellow Cab's operational capabilities are more efficient.





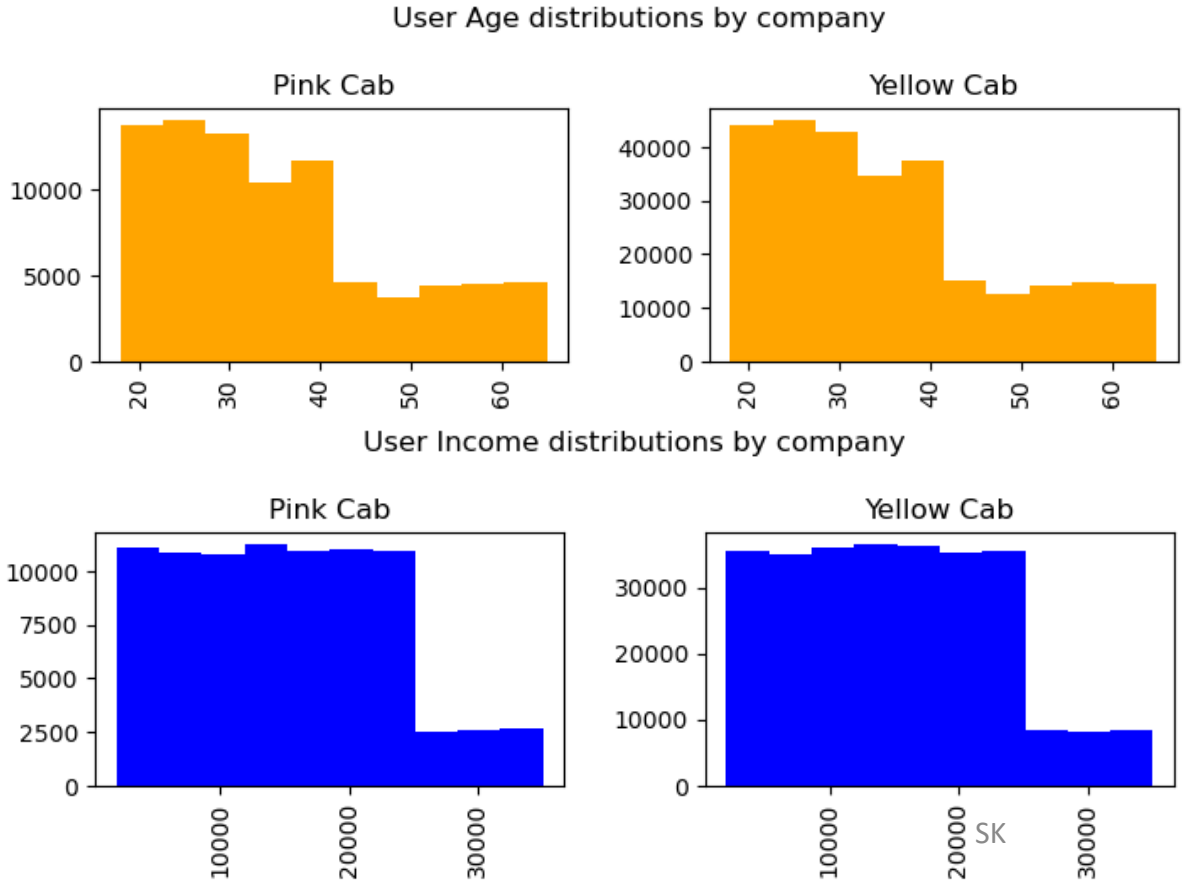


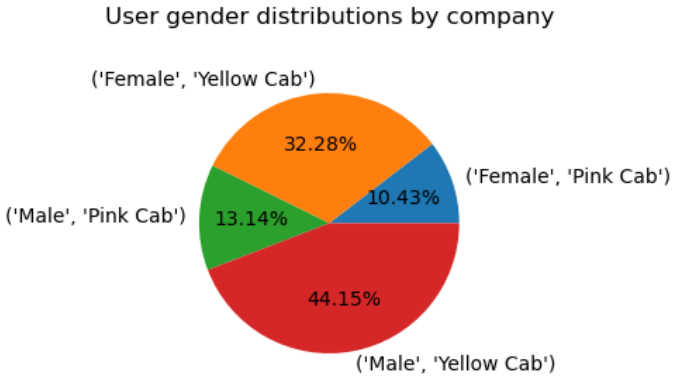
EDA Summary (cont.)

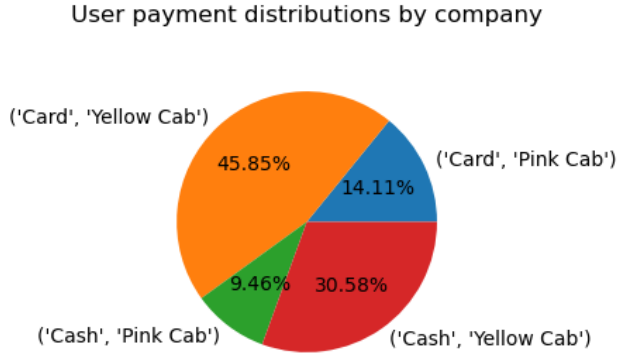
3. Data Visualization :

The histograms of age and income reveal nearly identical patterns for the two companies. Ratios by gender and payment method also demonstrate the same ranking. Therefore, it is reasonable to assume that these specific demographic factors or choices of payment method do not influence customer preferences.

b. Customer demographics and Payment methods





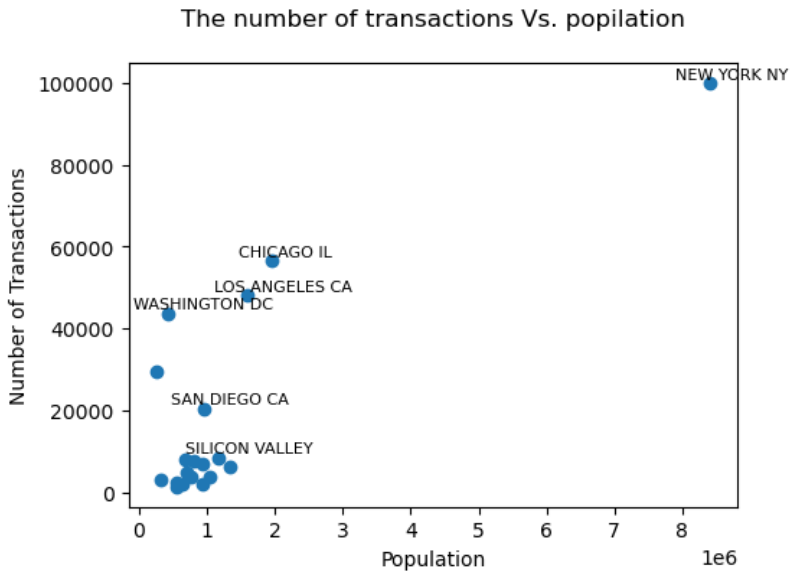


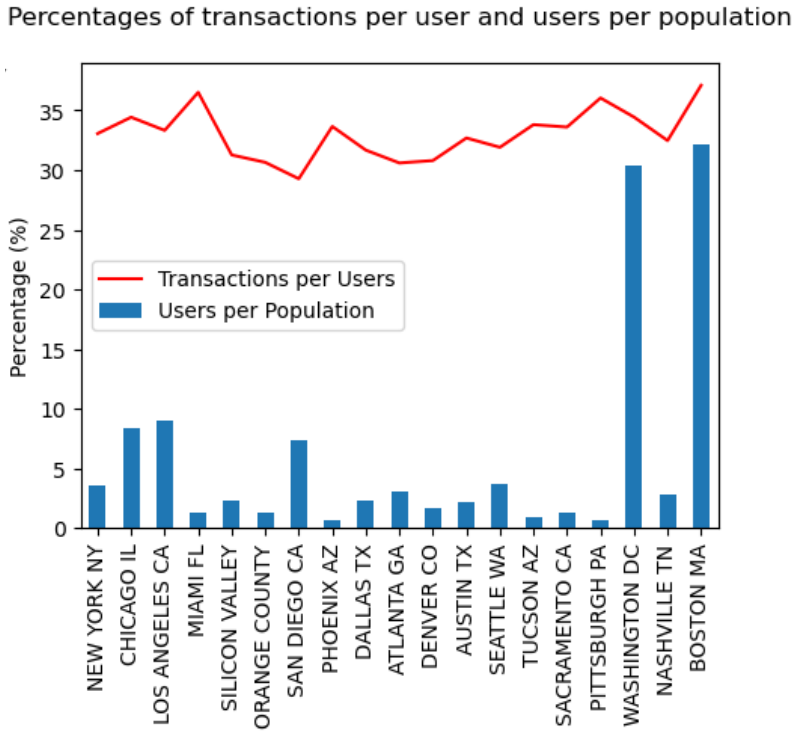
EDA Summary (cont.)

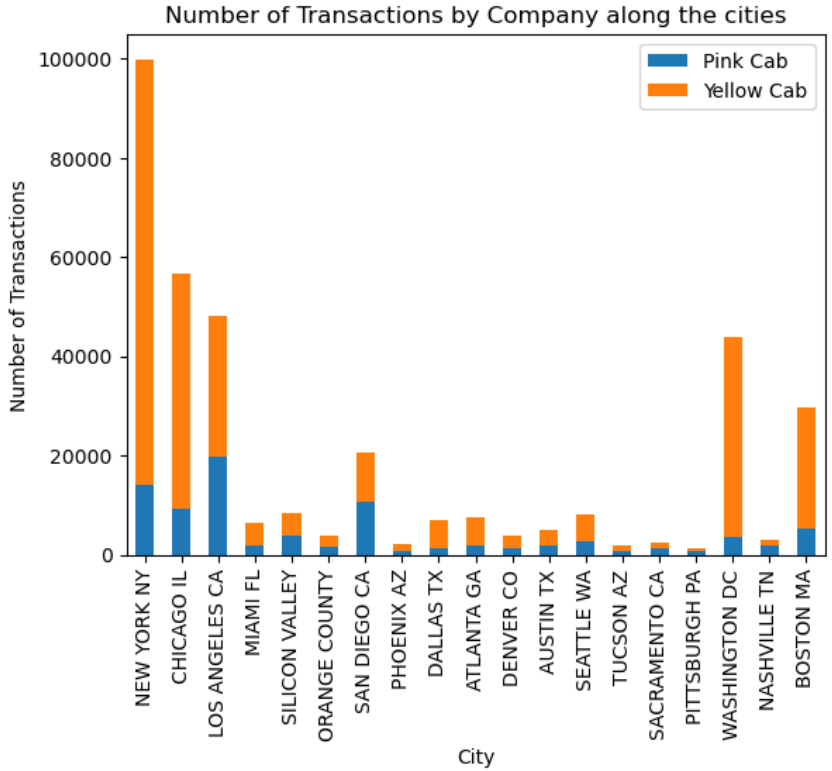
3. Data Visualization :

c. City population

The hypothesis that cities with large populations will exhibit a higher demand for taxi services is notably supported by two cities with small populations: Washington DC and Boston MA. These cities have more than three times the number of users compared to other cities, and Yellow Cab's operating pattern aligns well with these statistics. In other words, in the top five cities by transaction number, the number of Yellow Cabs significantly exceeds that of other companies.





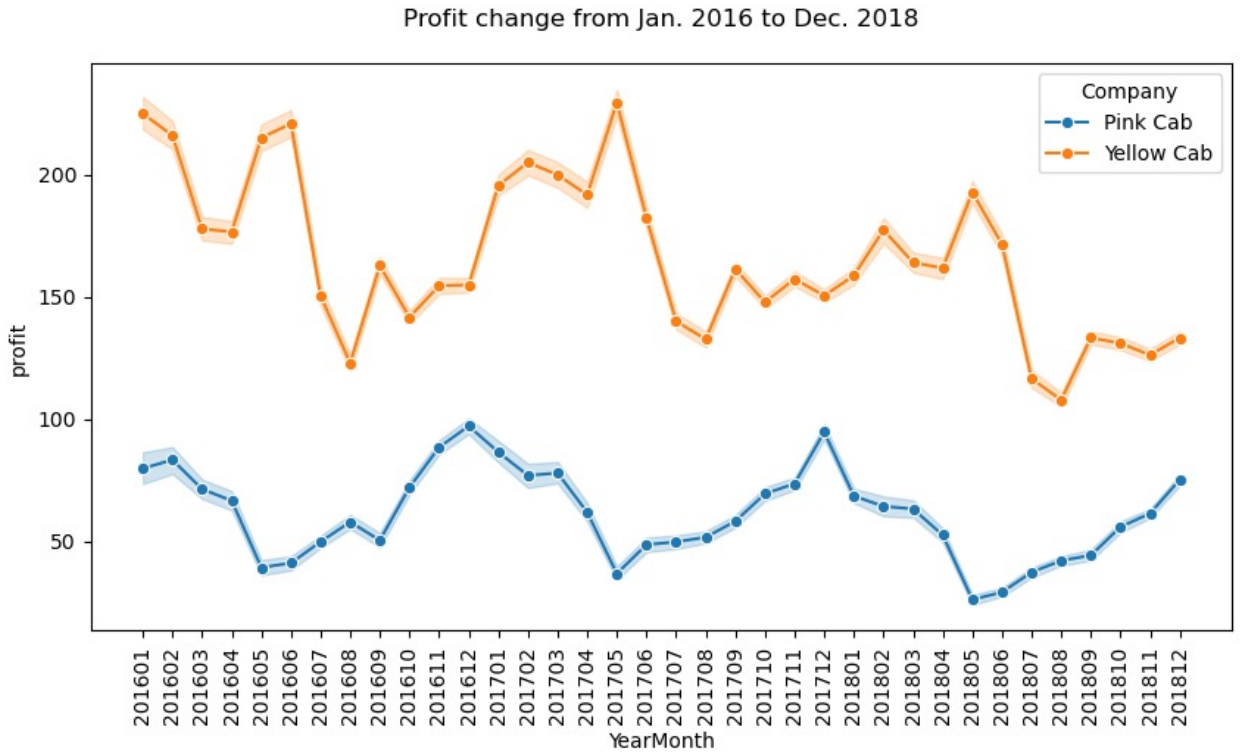


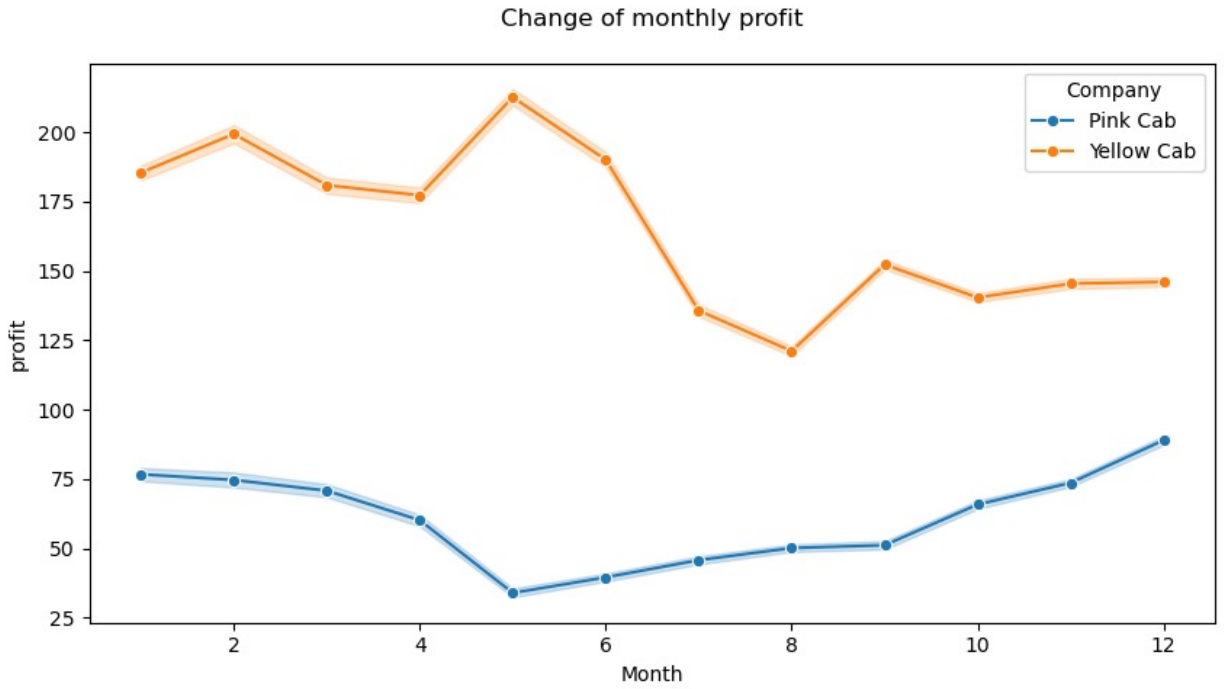
EDA Summary (cont.)

3. Data Visualization :

d. Seasonality

Finally, we estimate that there will be significant seasonal effects on the number of customers using taxi services. Monthly profits over three years exhibit different patterns for the two companies. Yellow Cab records the highest revenue in May and the lowest revenue in August, while Pink Cab shows the lowest revenue in May and the highest revenue in December. Pink Cab's seasonal pattern suggests that many taxis operate in warmer regions.





Conclusion

Based on the provided context and the explored hypotheses, the following recommendations can be made:

- **Market Share and Operational Efficiency**: Yellow Cab holds a higher market share in terms of the number of transactions and revenue. The operational capabilities of Yellow Cab appear to be more efficient, as the number of unique customers over the three years is not significantly different from that of the other company.
- **Customer Preferences**: The histograms of age and income, as well as the ratios by gender and payment method, reveal nearly identical patterns for the two companies. It is reasonable to assume that specific demographic factors or choices of payment method do not influence customer preferences.
- **City Population and Demand**: The hypothesis that cities with large populations will exhibit a higher demand for taxi services is supported by the data. Two cities with small populations, Washington DC and Boston MA, have a significantly higher number of users, and Yellow Cab's operating pattern aligns well with these statistics. In the top five cities by transaction number, the number of Yellow Cabs significantly exceeds that of other companies.
- **Seasonal Effects**: There are significant seasonal effects on the number of customers using taxi services. Monthly profits over three years exhibit different patterns for the two companies. Further research on people's movements in specific months may reveal more efficient ways to operate.

Recommendation

Based on these conclusions, the following recommendations can be made:

- ✓ **Investment Decision**: Considering Yellow Cab's higher market share and operational efficiency, it is recommended that XYZ consider Yellow Cab as a better investment opportunity in the cab industry.
- ✓ **Further Research**: Conduct further research on the seasonal effects and customer preferences to gain a deeper understanding of the market dynamics and customer behavior.

Thank You