Exploratory Data Analysis

G2M insight for Cab Investment firm

XYZ

16/09/2022

Agenda

- Executive Summary
- Steps take for the Analysis
- Data Exploration
- Datasets Info
- EDA
- Model Selection

EXECUTIVE SUMMARY

- XYZ is a private firm in US. Due to remarkable growth in the Cab Industry in last few years and multiple key players in the market, it is planning for an investment in Cab industry and as per their Go-to-Market(G2M) strategy they want to understand the market before taking final decision.

- Objective: Provide actionable insights to help XYZ firm in identifying the right company for making investment.

Steps taken for Analysis

- Data Exploration
- Importing Datasets
- Creating Master Data
- Data Wrangling
- Exploratory Data Analysis
  - ❖ Descriptive Statistics
  - ❖ Data Visualization
  - ❖ Hypothesis Testing
- Investment Decision
- Model Selection

Datasets Info

There are 4 datasets:

- **Cab_Data.csv –** this file includes details of transaction for 2 cab companies.
- **Customer_ID.csv** – this is a mapping table that contains a unique identifier which links the customer's demographic details.
- **Transaction_ID.csv –** this is a mapping table that contains transaction to customer mapping and payment mode.
- **City.csv –** this file contains list of US cities, their population and number of cab users.

The above four datasets are merged into **MasterData**

After cleaning the data, MasterData is normalized into **Pink_MasterData** & **Yellow_MasterData** for analysis purpose.

Data Exploration

19 Features(including 7 derived features)
Timeframe of the data: 2016-01-01 to 2018-12-31
Total data points : 359,392

Assumptions:

Outliers are present in Price Charged feature but due to
unavailability of trip duration details ,we are not treating this as outlier.

Profit of rides are calculated keeping other factors constant and only
Price Charged and Cost of Trip features used to calculate profit.

Users feature of city dataset is treated as number of cab users in the city.
we have assumed that this can be other cab users as well(including Yellow and Pink cab)

Exploratory Data Analysis

The analysis is carried through

Descriptive Statistics

Data Visualizations

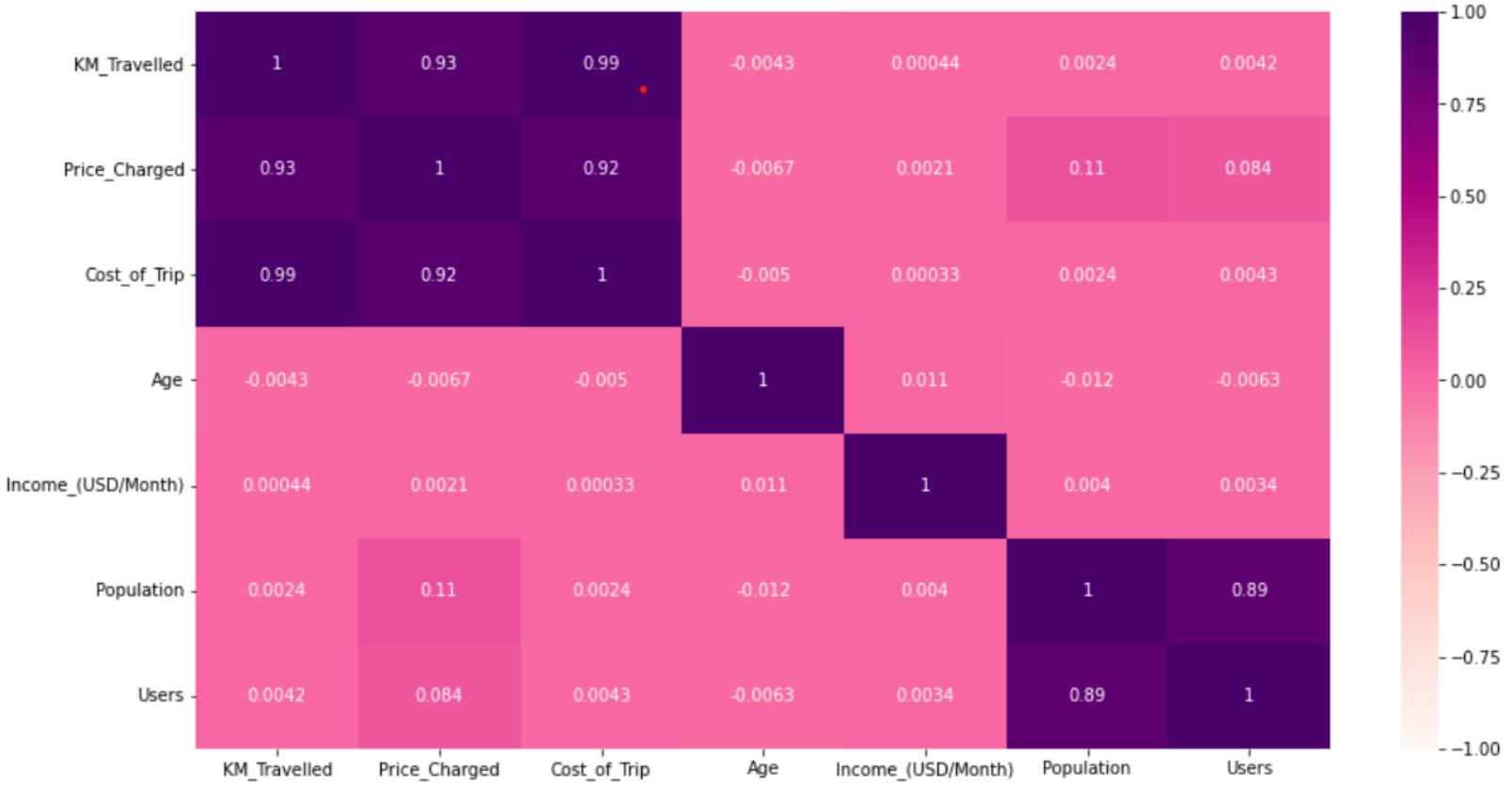
Hypothesis testing

Following are the key aspects of Analysis

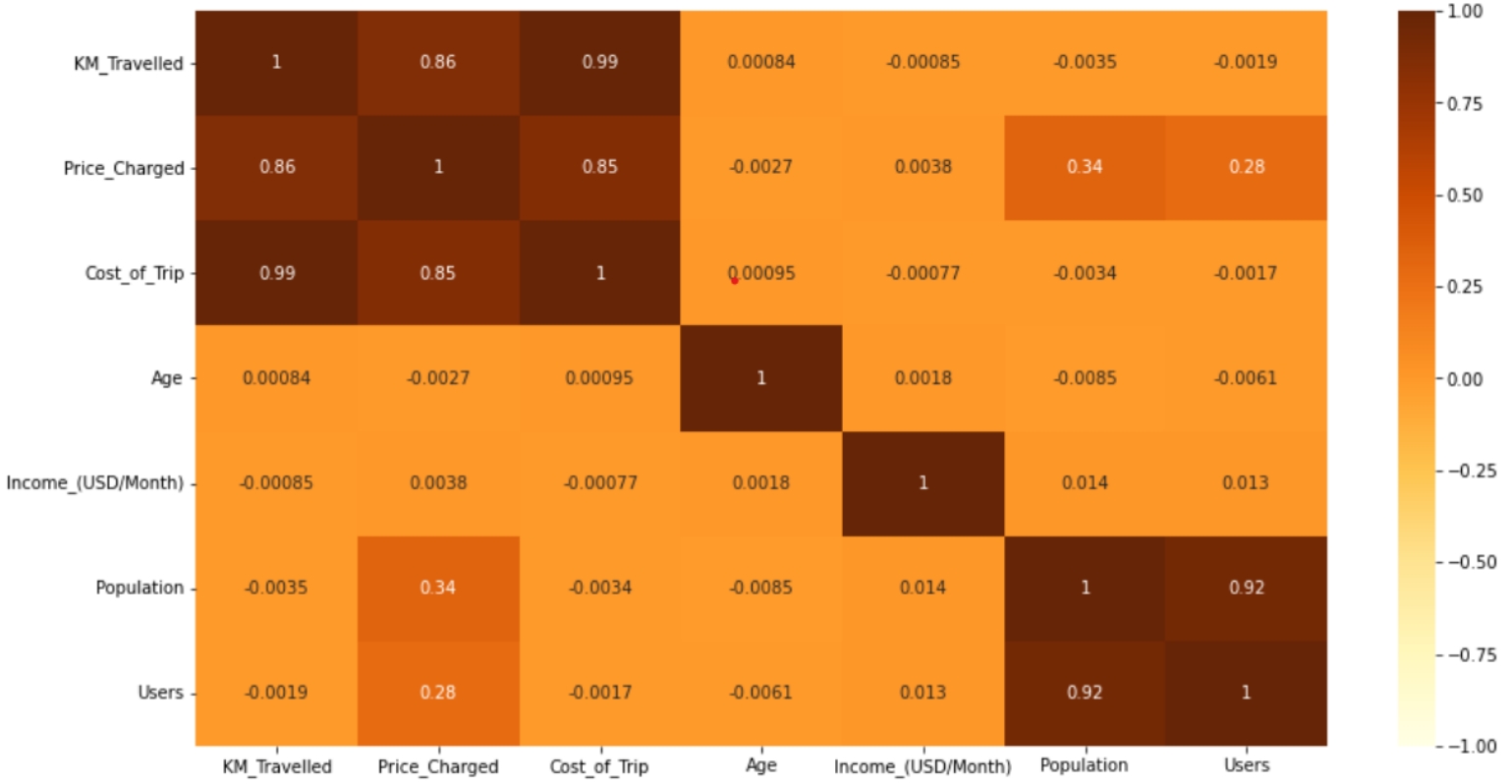
- **What are the attributes of these customer segments?**
- **What's the profit rate of Cab market? Which company has higher profit rate?¶**
- **Does Yellow Cab have bigger market share than Pink Cab?**
- **Do Customers prefer Yellow Cab to Pink Cab?**
- **Who serves the higher income?**
- **Which company prevail in which city?**
- **Does margin proportionally increase with increase in any variable?**
- **Is there any difference in Margins for Customers based on Genders, Age & Mode of Payment?**
- **Which company has maximum cab users at a particular time period?**

What are the attributes of these customer segments?

Pink Cab

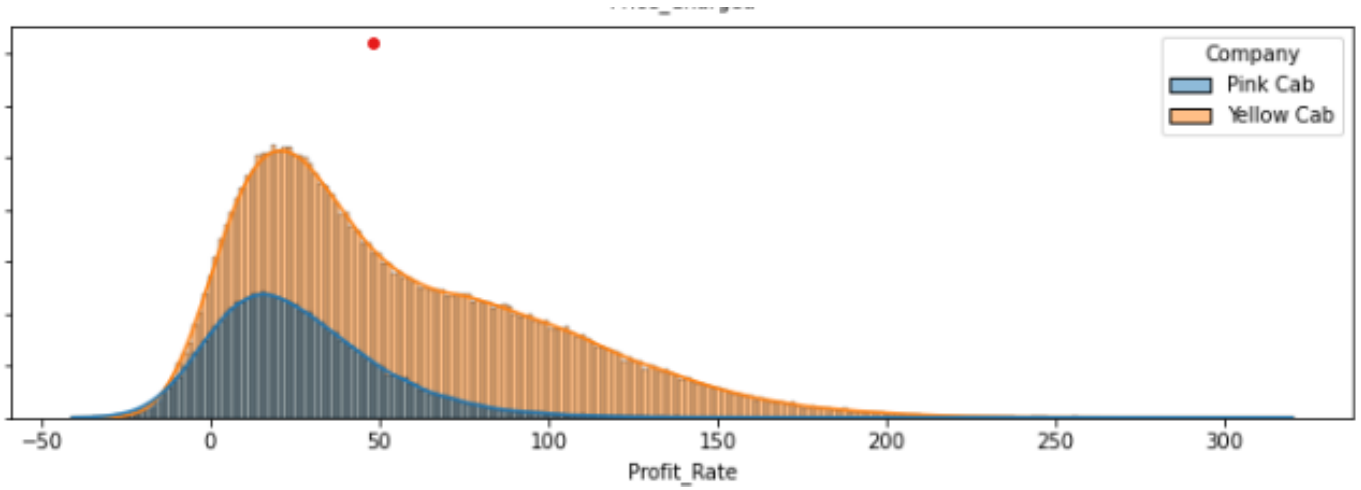


Yellow Cab

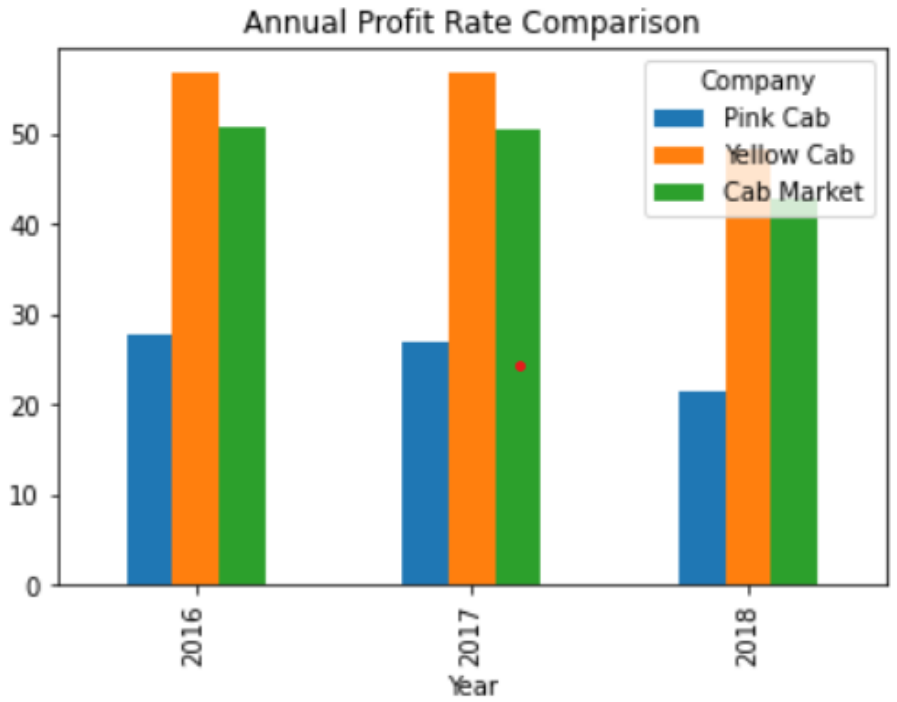


Here we have analyzed the correlation between various attributes(variables) from various segments. Key attributes analyzed are Cost of Trip, Price Charged, Income, Age, KM Travelled, Population, Numbers of Users. KM Travelled, Cost of Trip, Price Charged are highly correlated with Population & Users The price charge range for Yellow cab is more than the Pink cab. There is positive correlation between Price and KM Travelled for both Pink & Yellow Cab .

What's the profit rate of Cab market? Which company has higher profit rate?¶

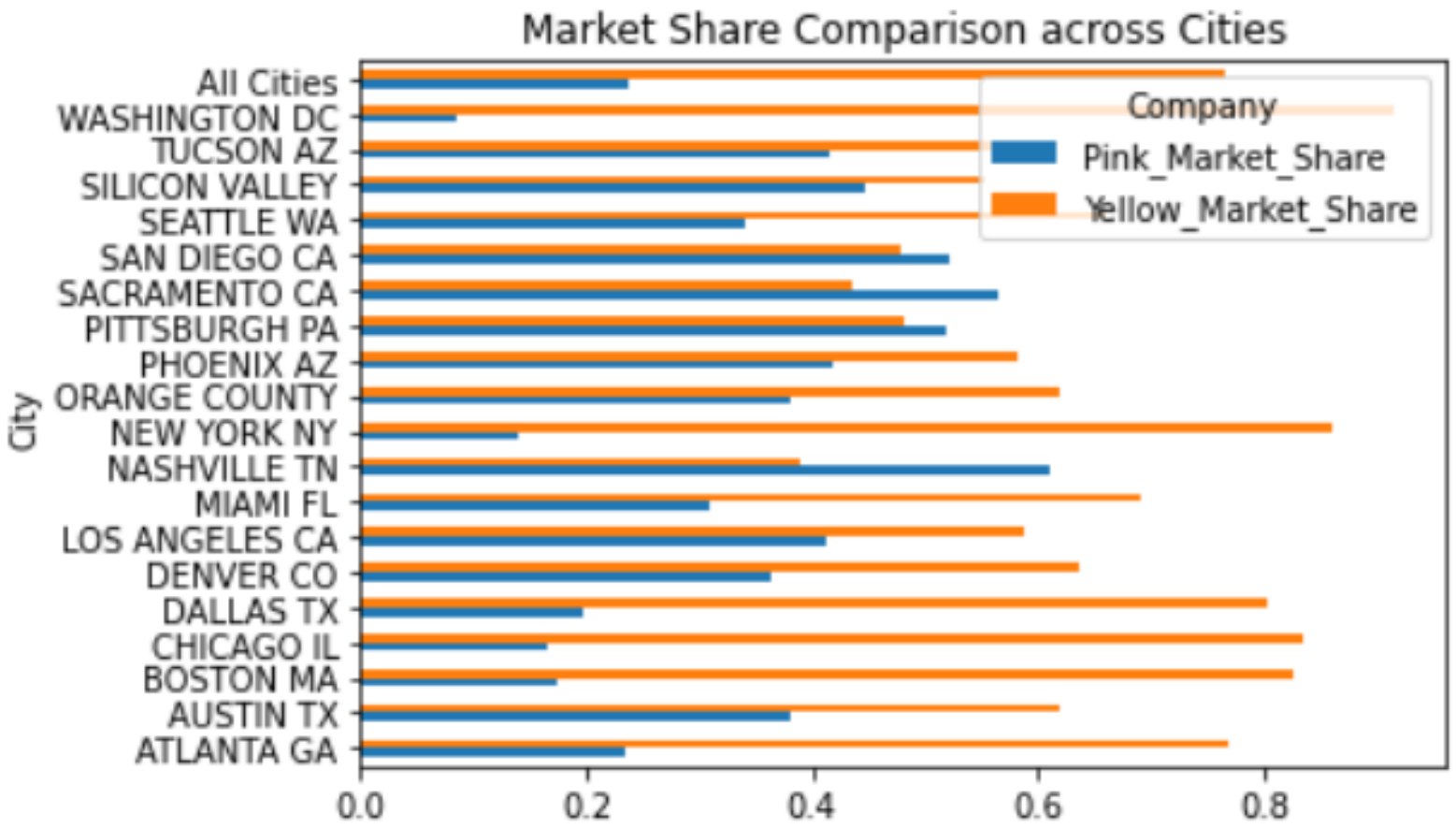


Comparatively Yellow Cabs are skewed towards right implying higher total profit rate



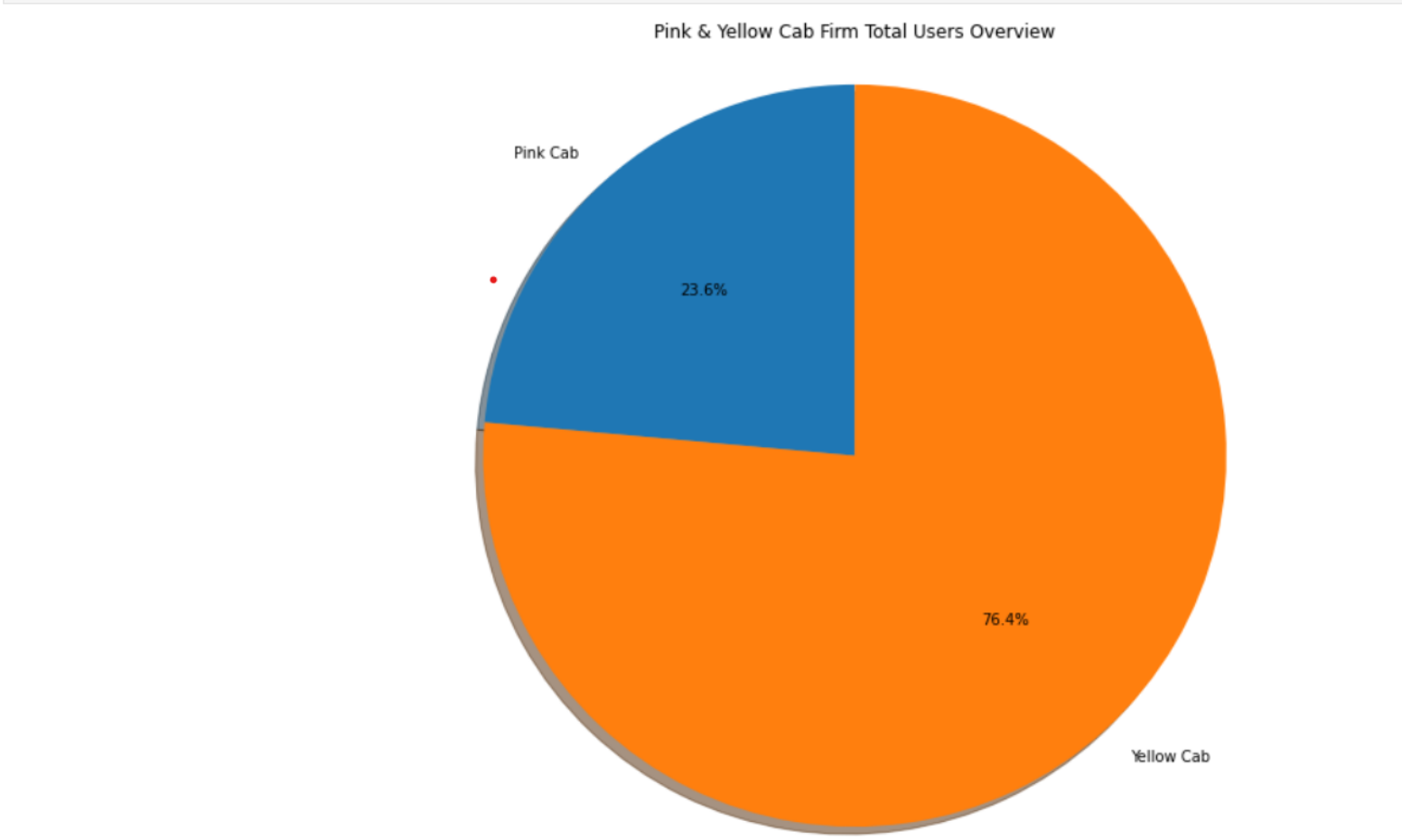
Yellow Cab has higher yearly profit

Does Yellow Cab have bigger market share than Pink Cab?



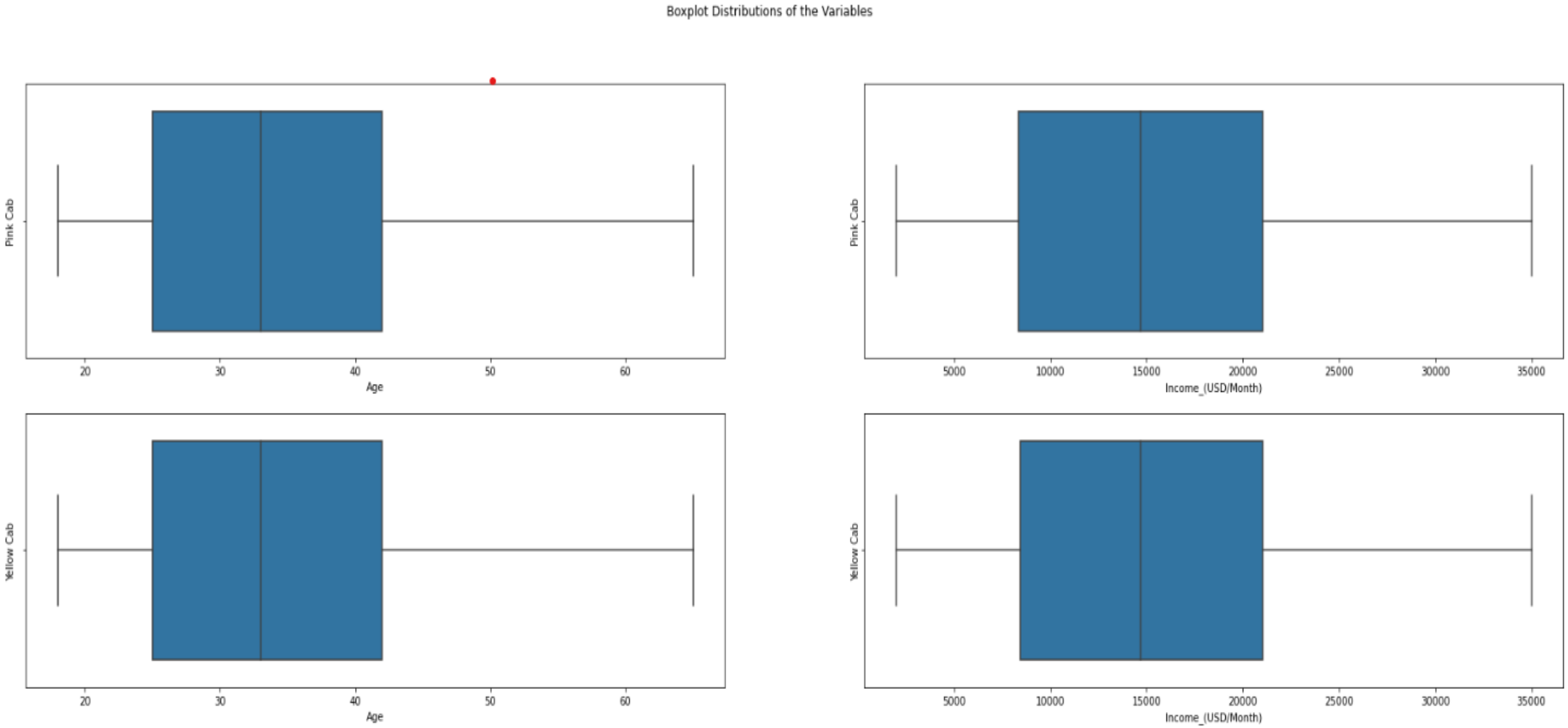
Yellow cab market share is three times greater than Pink Cab

Do Customers prefer Yellow Cab to Pink Cab?



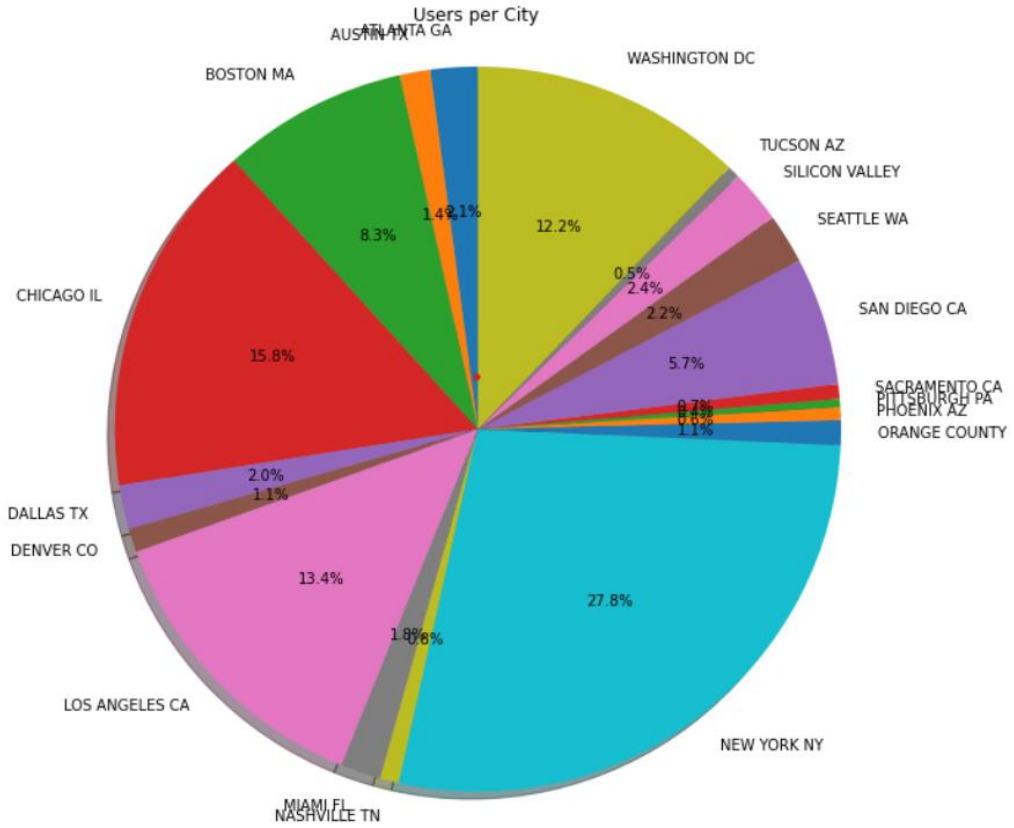
Customers prefers yellow cabs. The number of users of Yellow Cab is 3 times that of Pink Cab approximately.

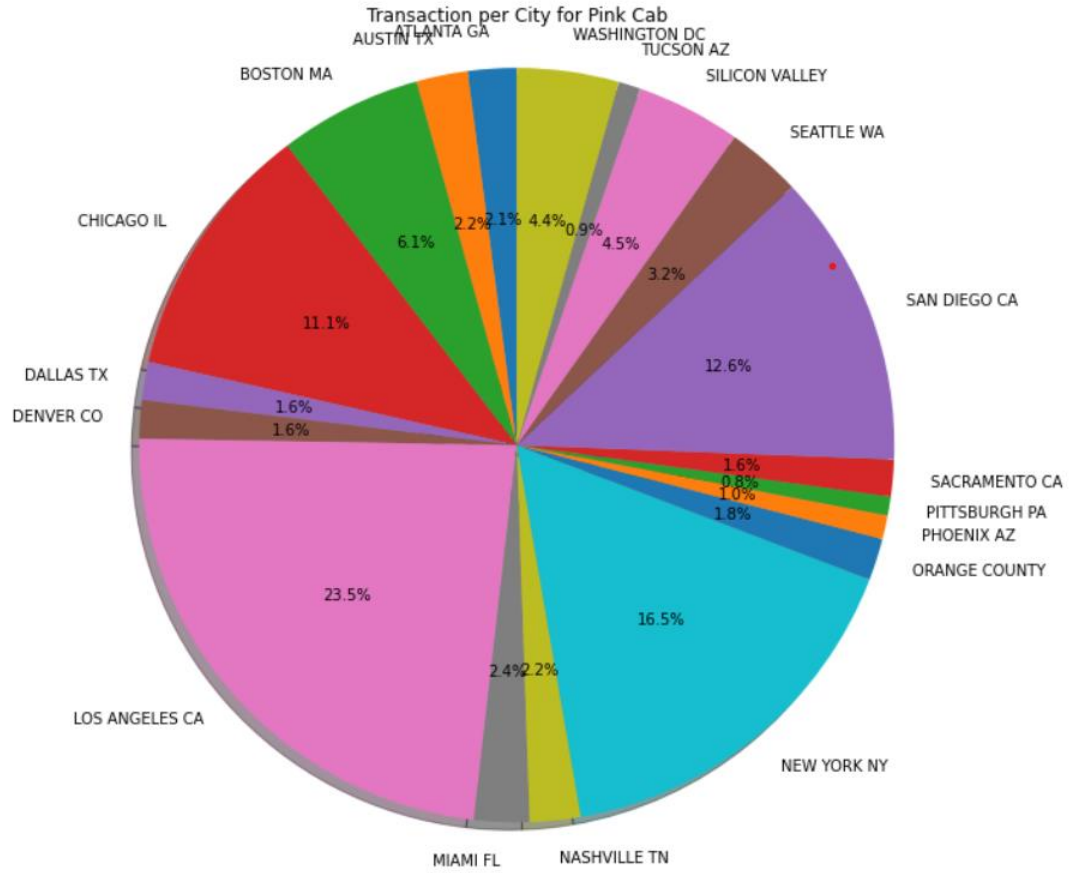
Who serves the higher income?



As per above box plot Pink Cab and Yellow Cab serve similar income.

Which company prevail in which city?





Transaction for Yellow Cab is highest in New York City which has the highest Cab Users of 28%

Transaction for Pink Cab is highest in Los Angeles City

Does margin proportionally increase with increase in any variable?

Hypothesis Testing results

- Does margin proportionally increase with increase in Price?

  There is a correlation between Margin & Price Charged

- Is there any difference in Margins for Customers based on Genders?

  Yellow Cab: There is difference in Margin between Male and Female customers

  Pink Cab: There is no difference in Margin between Male and Female customers

- Is there differences in Margins based on Age Group?

  There are differences in Margin based on Age for Yellow Cab

- Is there difference in margins for Card payer and Cash payers?

  There is no difference in Margin regarding mode of Payment for both Yellow & Pink Cab

Which company has maximum cab users at a particular time period?

Customer Number of Yellow Cab

Year	Customer ID
2016	25937
2017	27789
2018	27470

Customer Number of Pink Cab

Year	Customer ID
2016	16661
2017	18643
2018	18400

Yellow Data has maximum users in 2017

Investment Recommendation

As per the parameters Profits Margins, Income, Customer Availability, Age wise reach, Market Share **YELLOW CAB** is recommended.

Which model is best to predict price from the given dataset?

Multiple Linear Regression (MLR) vs. Polynomial Fit

- The Mean-Square-Error(MSE) for the MLR is smaller than the MSE for the Polynomial Fit.
- R-squared: The R-squared for the MLR is also much larger than for the Polynomial Fit.

The Multiple Linear Regression model is the best model to be able to predict price from our dataset.

Thank You