Exploratory Data Analysis

G2M insight for Cab Investment firm

21-August-2022

Agenda

Executive Summary

Problem Statement

Approach

EDA

EDA Summary

Recommendations

Description:

- XYZ is a private equity 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.

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

- **Cab Companies:**
- Yellow Cab
- Pink Cab

- **The Analysis include :**
- Data Understanding,
- Data Visualization,
- Creating multiple hypothesis,
- Building models and finding the best fit model based on Accuracy.

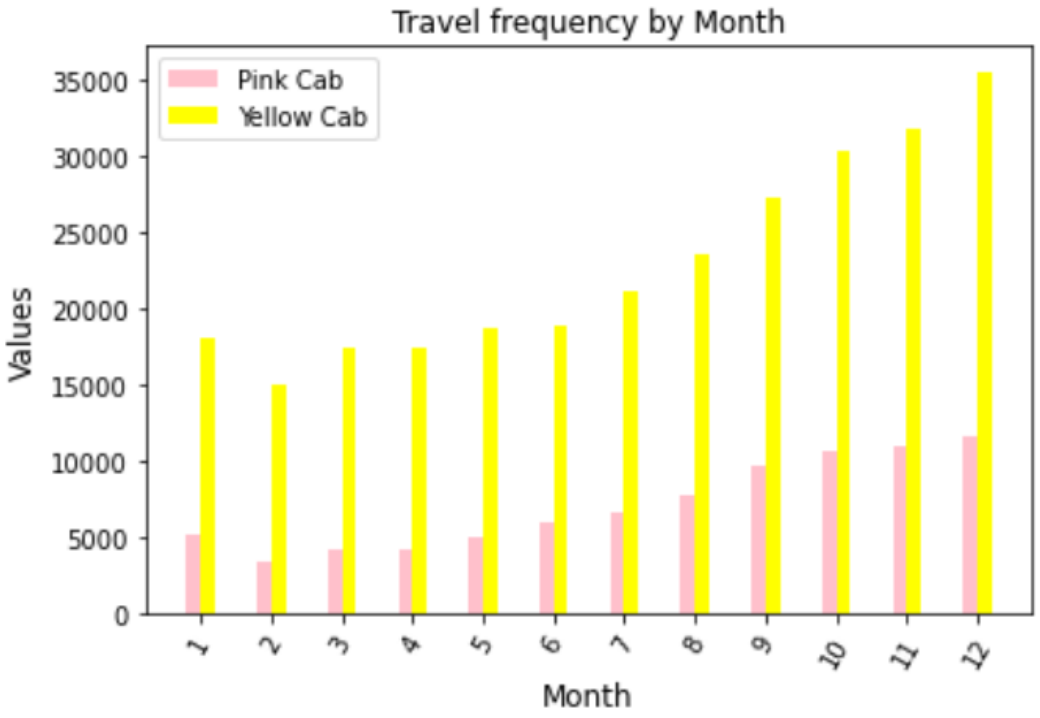
Data Preparation:

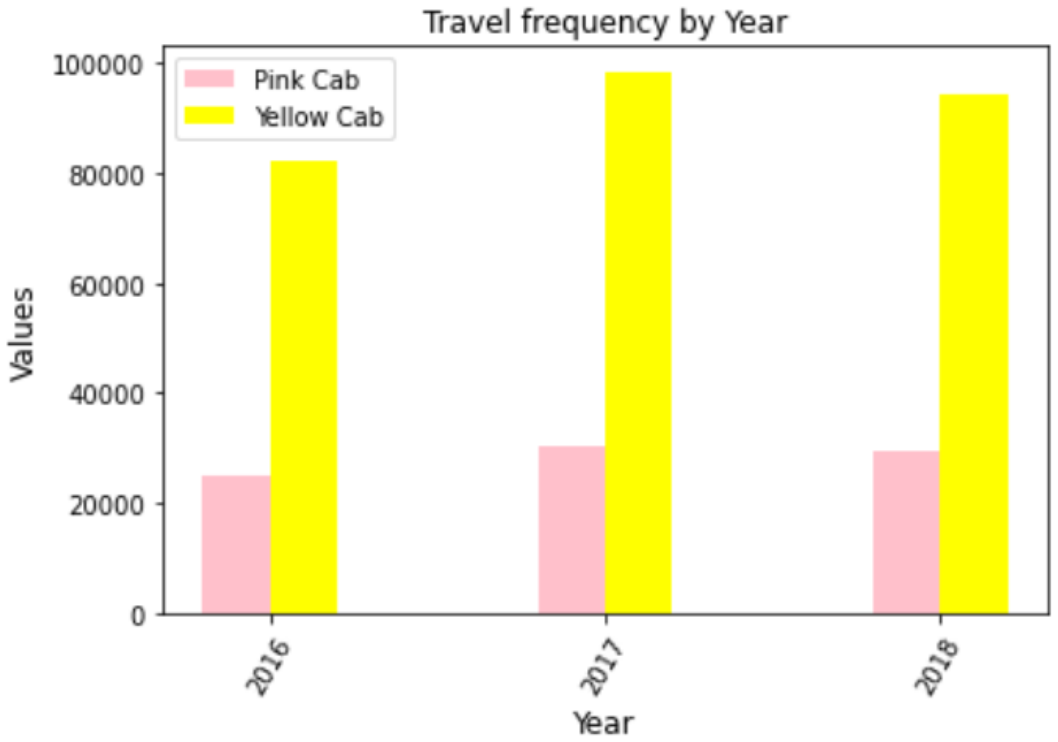
- **There are 4 datasets:**

- 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.

EXPLORATORY DATA ANALYSIS

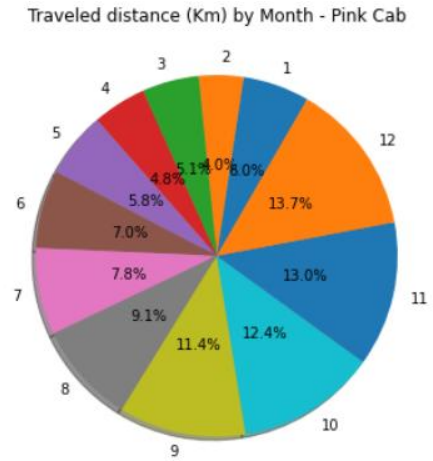
Travel frequency by Month/Year

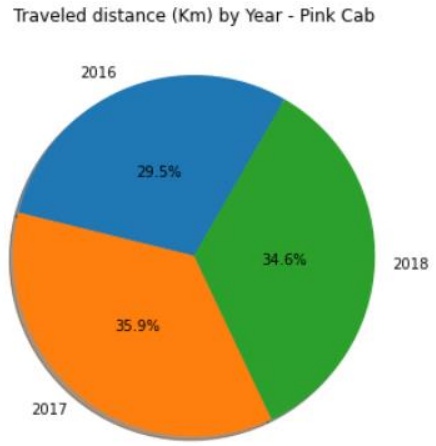


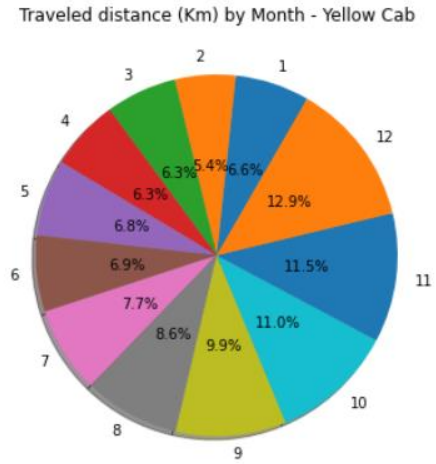


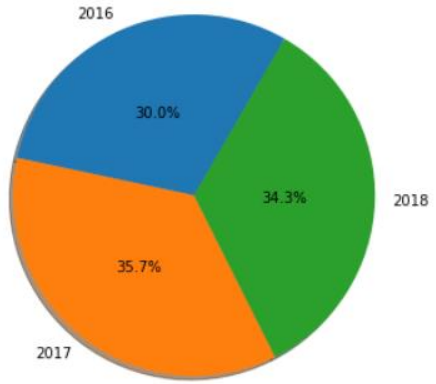
- **Analysis: It is observed that most of the travels occur in December for both companies. It is also observed that the year 2017 has more travels for both cabs.**

Traveled Distance Percentage for Each Month/Year.



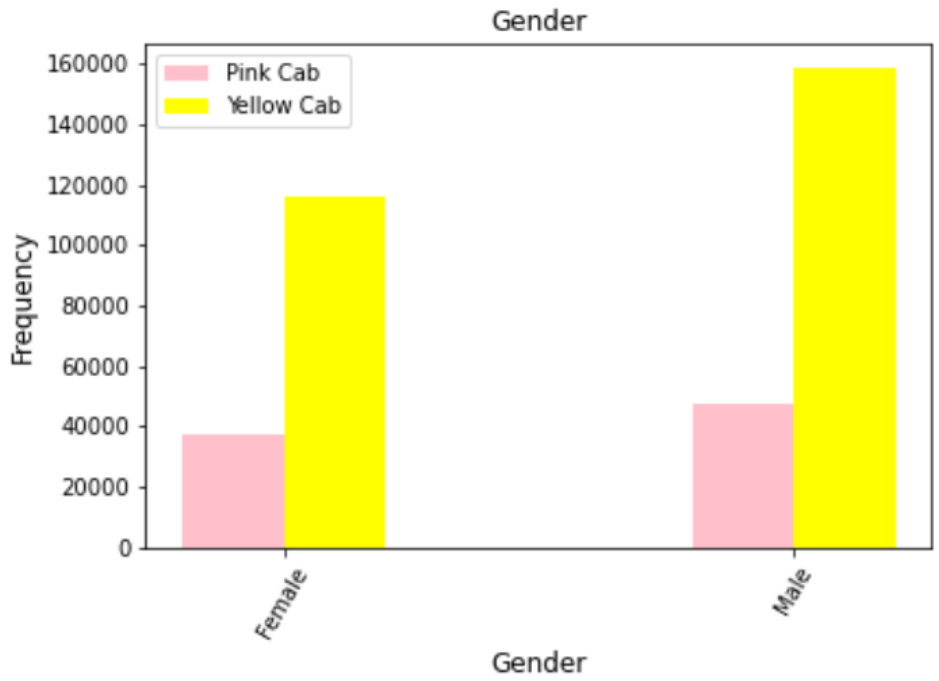






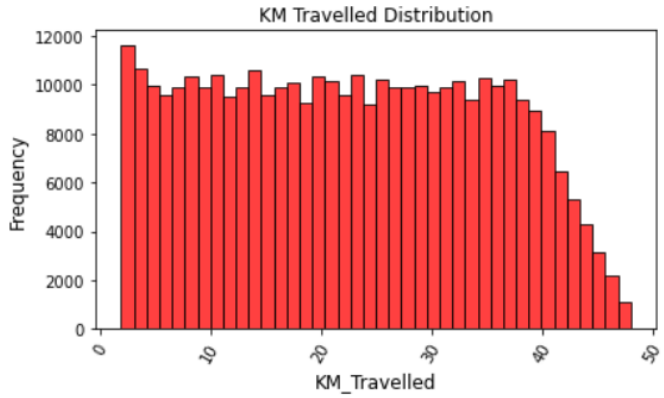
- **It is observed that most of the travelled were occurred on December.**

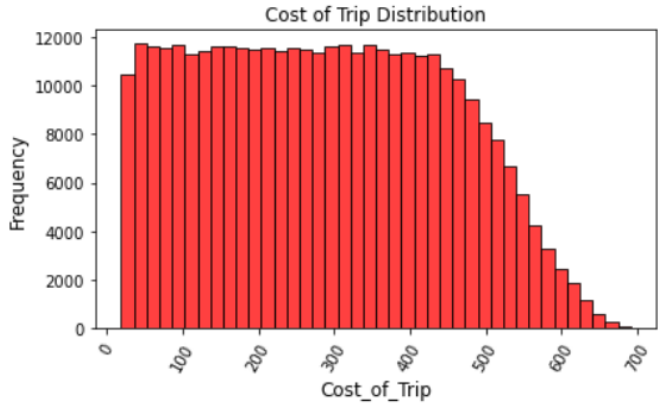
Payment Mode and Gender

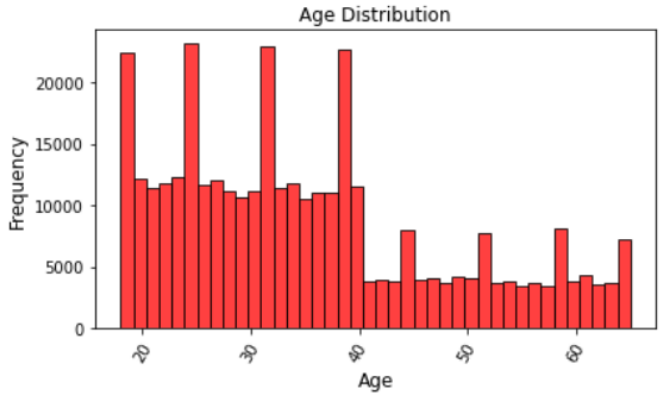


- **Analysis: it is observed that customers of both companies prefer to pay by card. It is also observed that most customers are Males.**

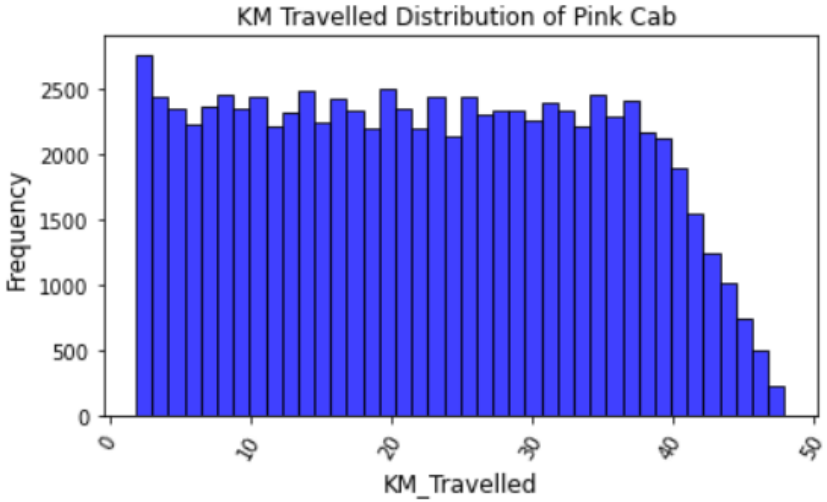
Distribution of KM travelled, price, cost of tirp and age:

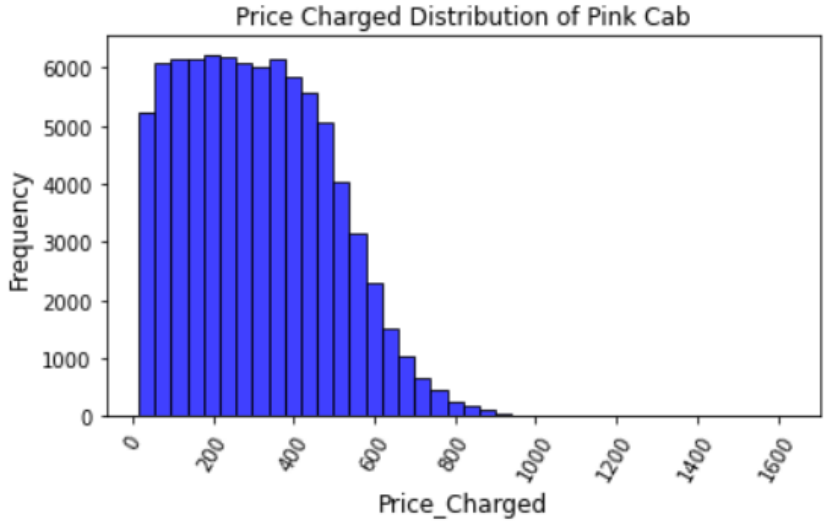


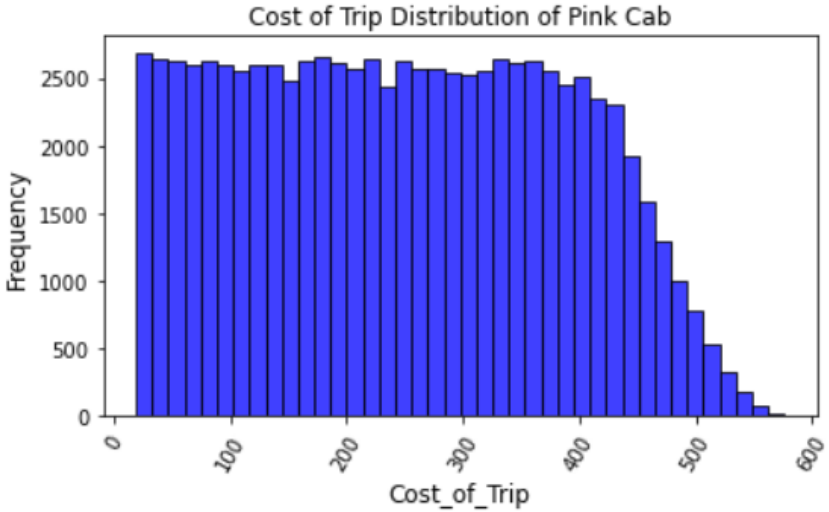


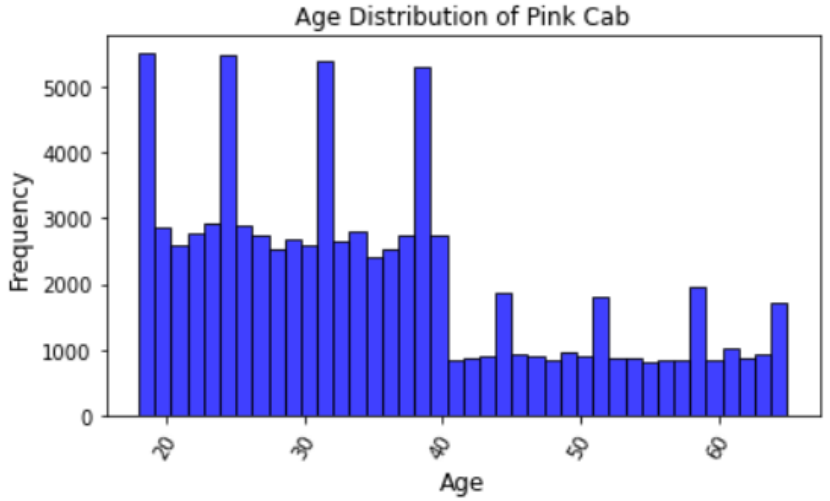


- **Analysis: It is observed that the most of the rides varies from 2 to 48 KM. Most of the charged prices are between 15.6 and 600 Dollar, and then the number starts to decrease. Most of the costs are varied from 19.0 to 691.2 Dollar where the maximum cost is around 690. Regarding the age of the customers, up to 80k are customers within the age of 18, 24, 31 and 39 years; except these ages, there are up to 10k customers for each age between 19 and 40, and les than 7k customers for each age >40 years.**

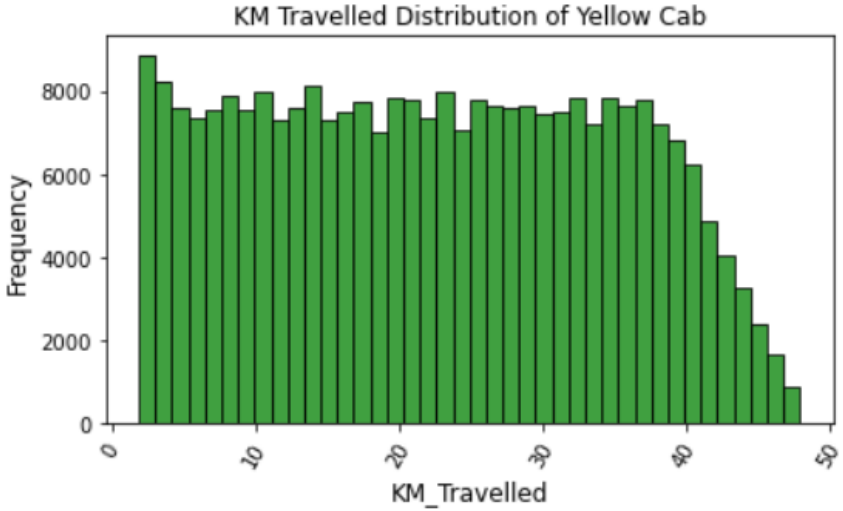


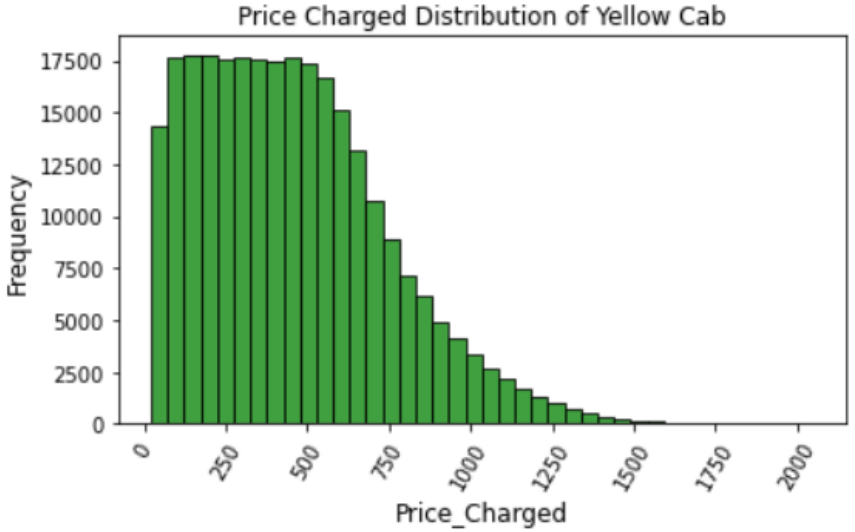


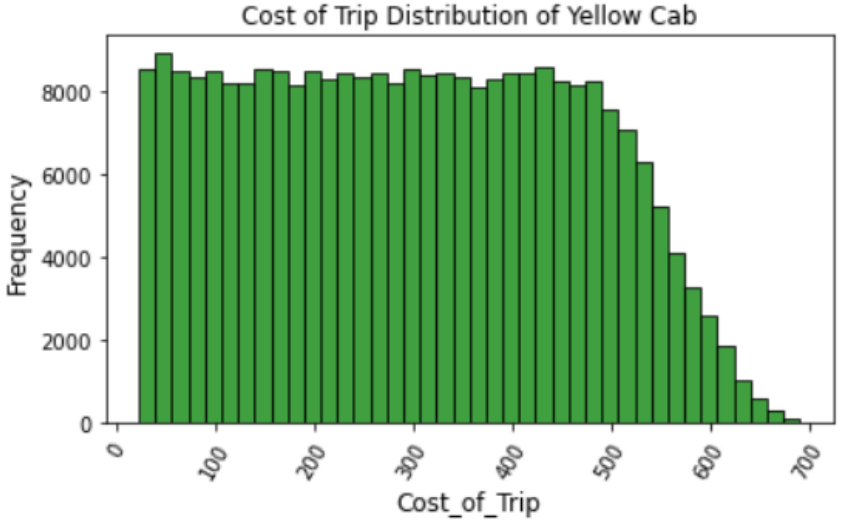


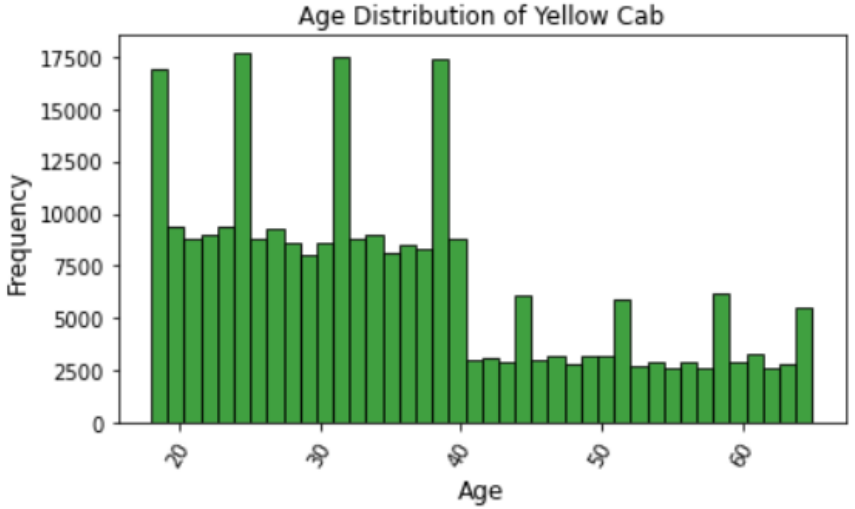


- **Analysis: it is observed that most of the costs of the Pink Cab are varied between 19 and 500 Dollar, where the rest are varying between 500 and 576 Dollar.**



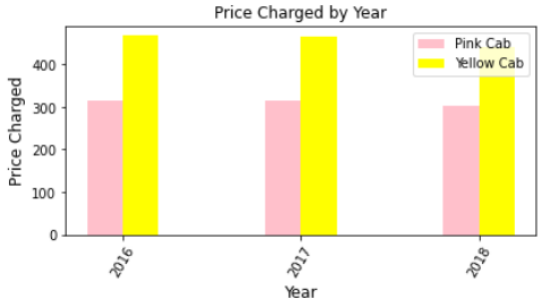


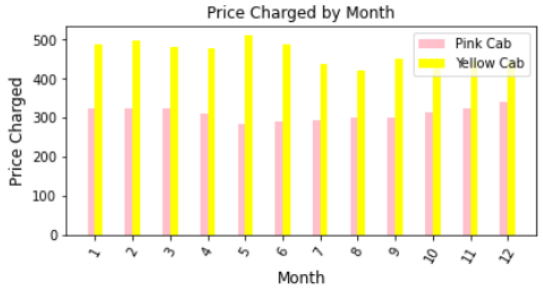


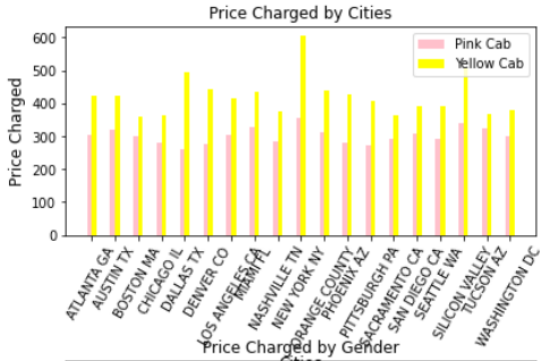


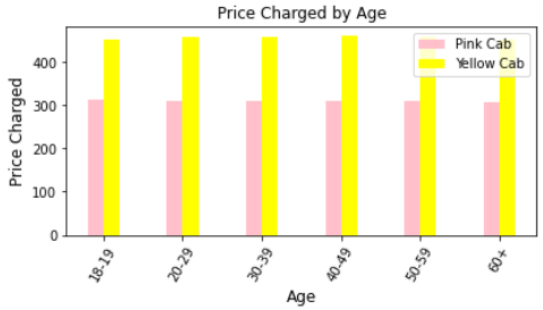
- **Analysis: it is observed that most of the costs of the Yellow Cab are varied between 22 and 600 Dollar, where the rest are varying between 600 and 691 Dollar**

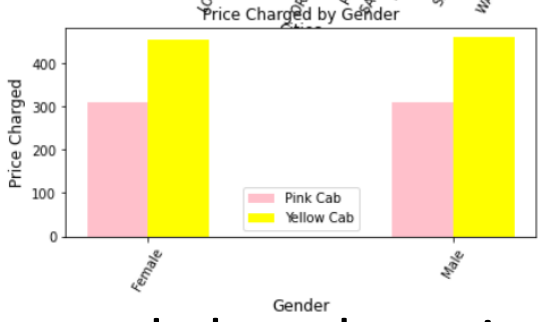
Price Analysis

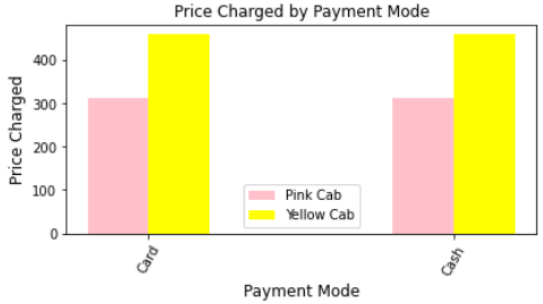






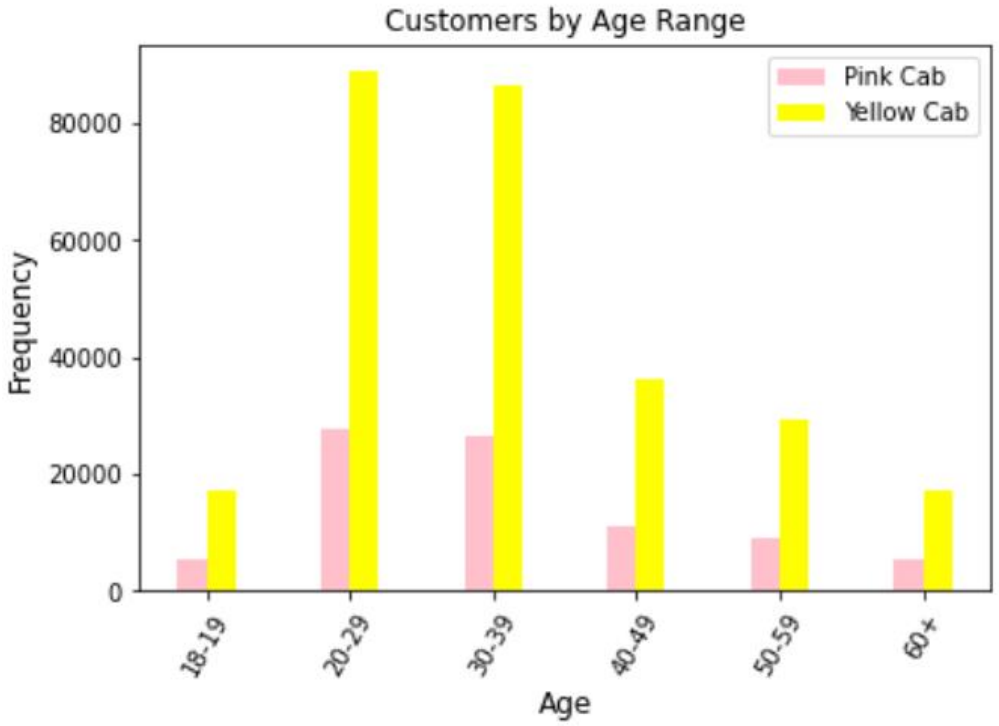


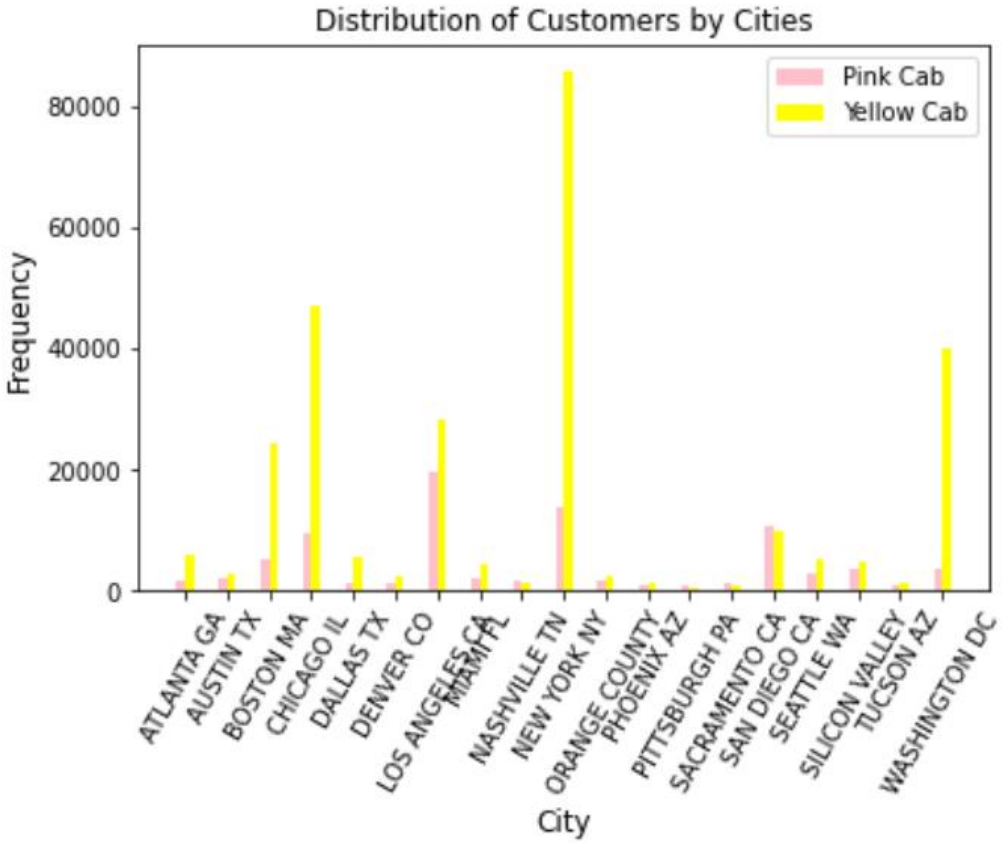




- Analysis: it is observed that the price charged of Yellow Cab is higher for all the categories (seasons, cities, age, gender and payment mode). Besides, it is clear that the price is highly vitiated by month and city.

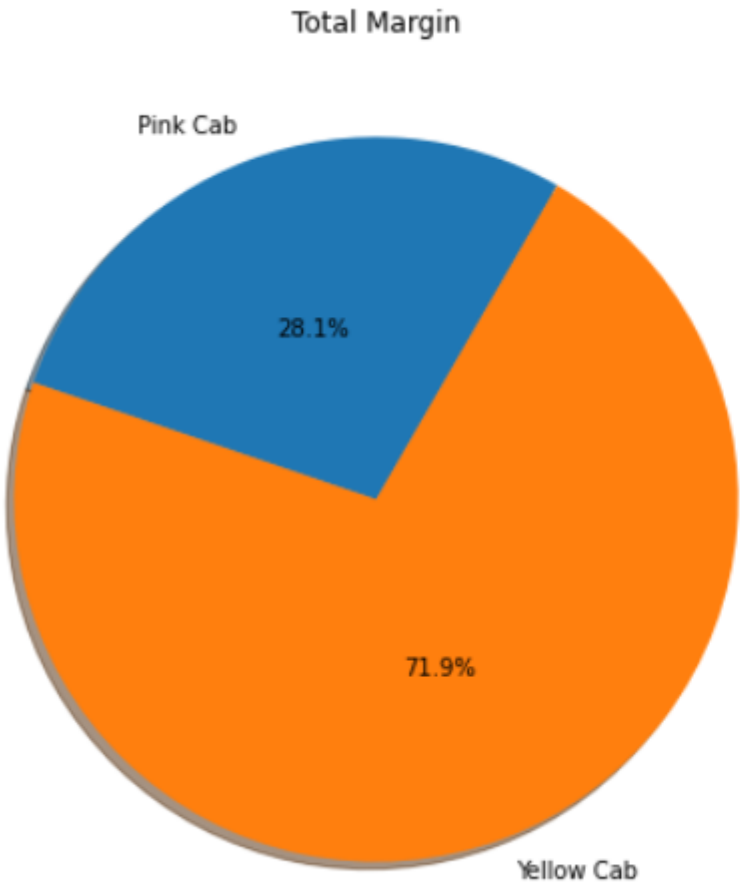
Age and Cities Analysis:

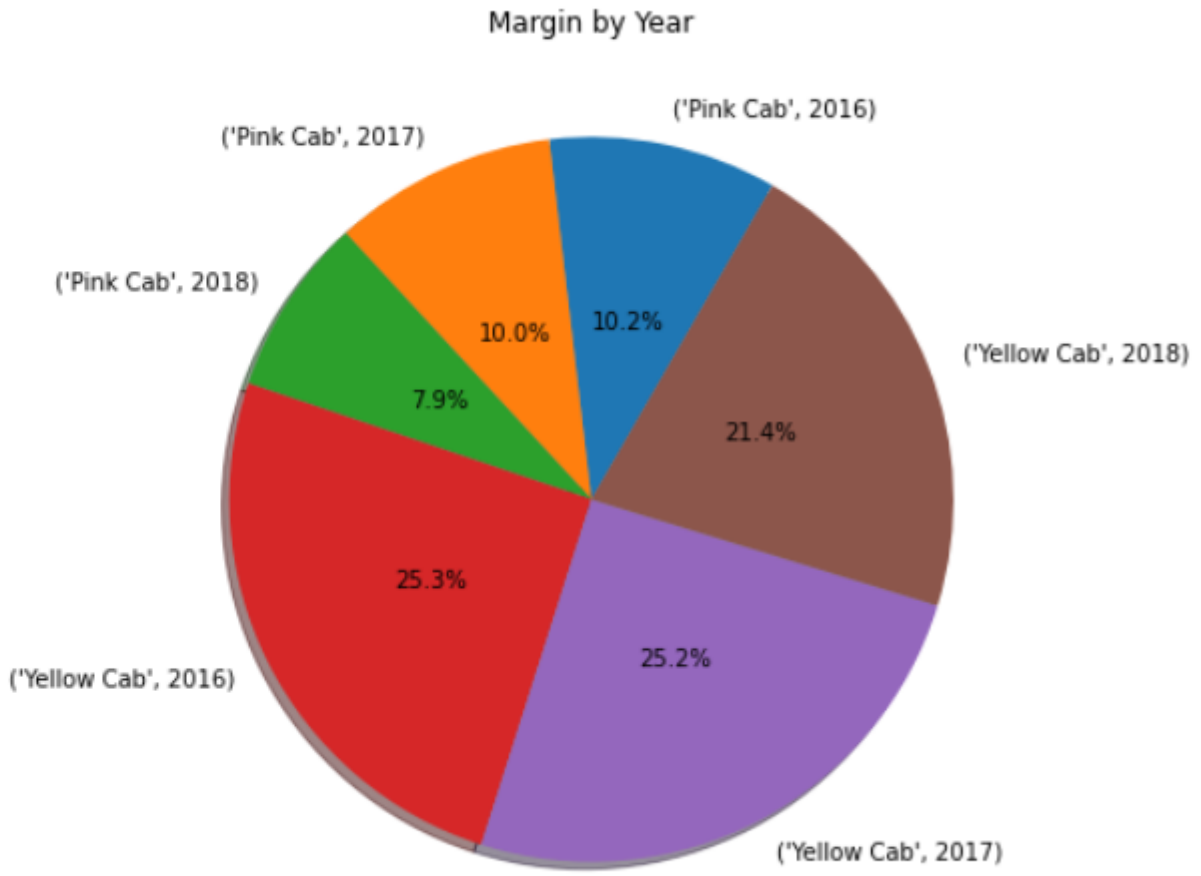




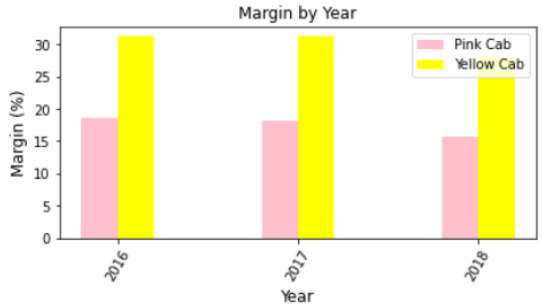
- **Analysis: it is observed that most of the ages are between 20 and 40 years for both companies. It is also observed that most customers of the Pink Cab are from LOS ANGELES CA whereas most of the customers of the Yellow Cab are from NEW YORK.**

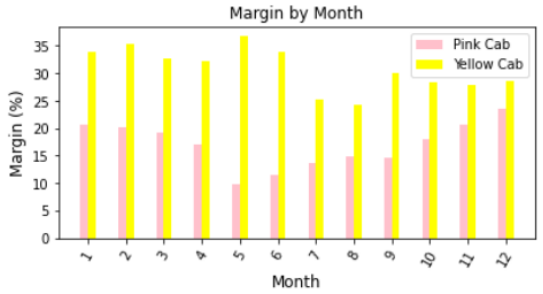
Margin Analysis:

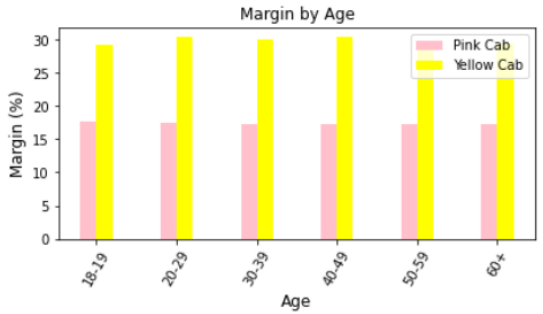


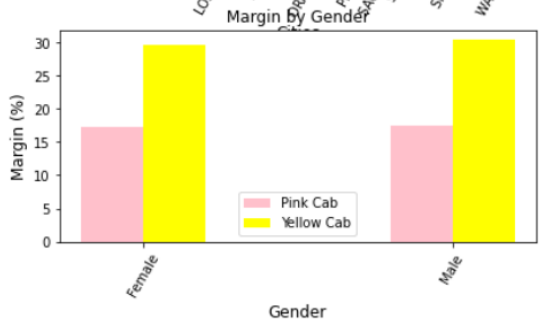


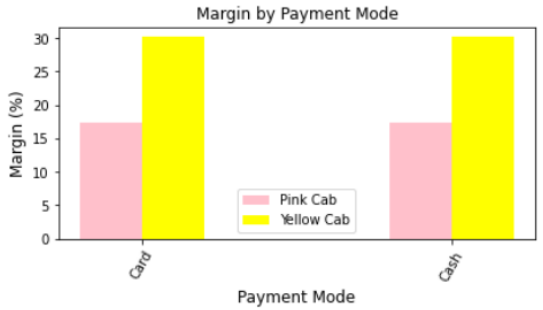
- **Analysis: Yellow cab owns 71.9% of the total margin made by both companies.**





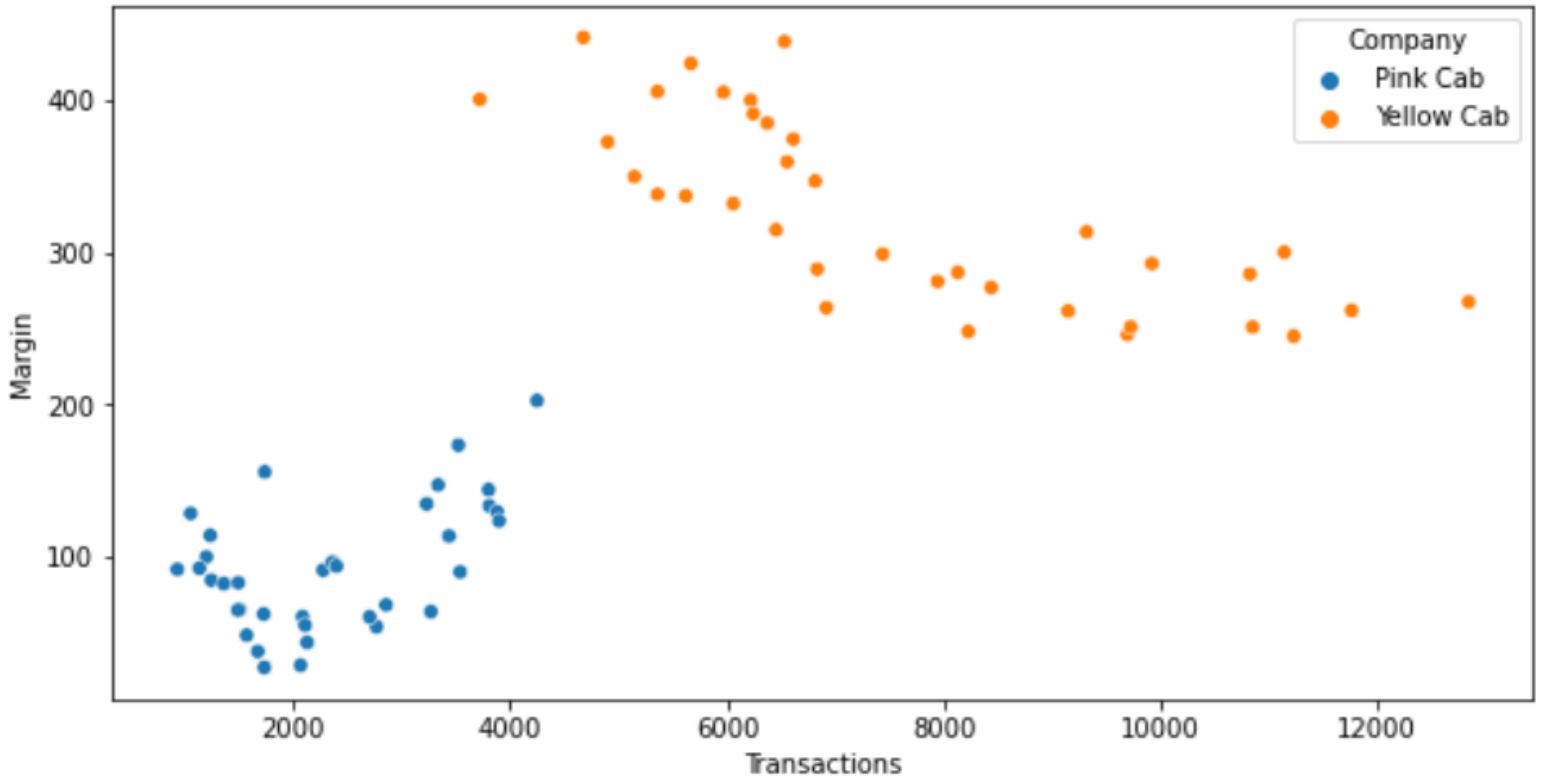






- Analysis: it is observed that in 2016 and 2017 the profit of the two companies kept constant, whereas a slight reduction is occurred in 2018. The highest Margin of the Yellow cab company was in the fifth month whereas its lowest Margin was in the 8th month. Regarding the Pink Cab company it had a high Margin in the first and last months of the year whereas its lowest Margin was in the fifth month. The highest Margin for both companies was from NEW YORK although the Pink company had more customers from LOS ANGELES. It is also observed that the age, Gender and Payment mode looks like that they have no dependency on the Margin since they look constants. However, later, they will be studied in depth.

Margin Vs Number of Customers



Analysis: it is observed that Pink Cab increases its margins with increase in number of Transactions. However, the Yellow Cab decrease its Margin with the increase in Transactions.

Correlation:

- Analysis: it is observed that there is a high correlation between Margin and Price_Charged

Hypothesis Testing

Null Hypothesis:

- Margin remain the same regarding Gender for both Yellow Cab & Pink Cab
- Margin remain the same regarding Age for both Yellow Cab & Pink Cab
- Margin remain the same regarding mode of Payment for both Yellow Cab & Pink Cab

Margin per Gender:

- There is no difference in Margin between Male and Female customers for the Pink Cab:
- There is a difference in Margin between Male and Female customers for the Yellow Cab:

Margin per Age:

- Pink Cab does not give any discount to people older than 60
- Pink Cab does not give any discount to young customers (<20)

- Yellow Cab gives discount to people older than 60
- Yellow Cab also gives discount to young customers (<20)

Margin per Mode of Payment:

- There is no difference in Margin regarding mode of Payment for the Pink Cab
- There is no difference in Margin regarding mode of Payment for the Yellow Cab

CONCLUSIONS

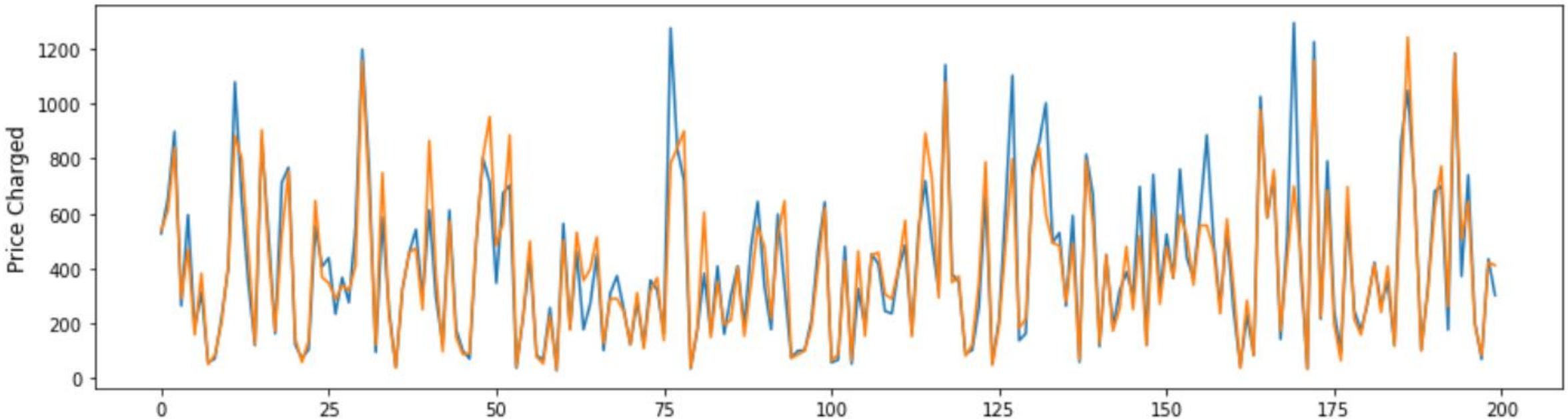
- We have evaluated both cab companies on the following points and found Yellow cab better than Pink cab:
- Most Users prefer travelling with Yellow cab than Pink cab
- Users travelled mostly with Yellow cab on short and long distance trip
- Yellow cab company charged more in populated City which is a good strategy and increase its profit
- Yellow cab charged higher than Pink cab
- Yellow cab owns 71.9% of the total margin made by both companies

We will advise the XYZ company to invest in Yellow Cab company for its glorious benefit.

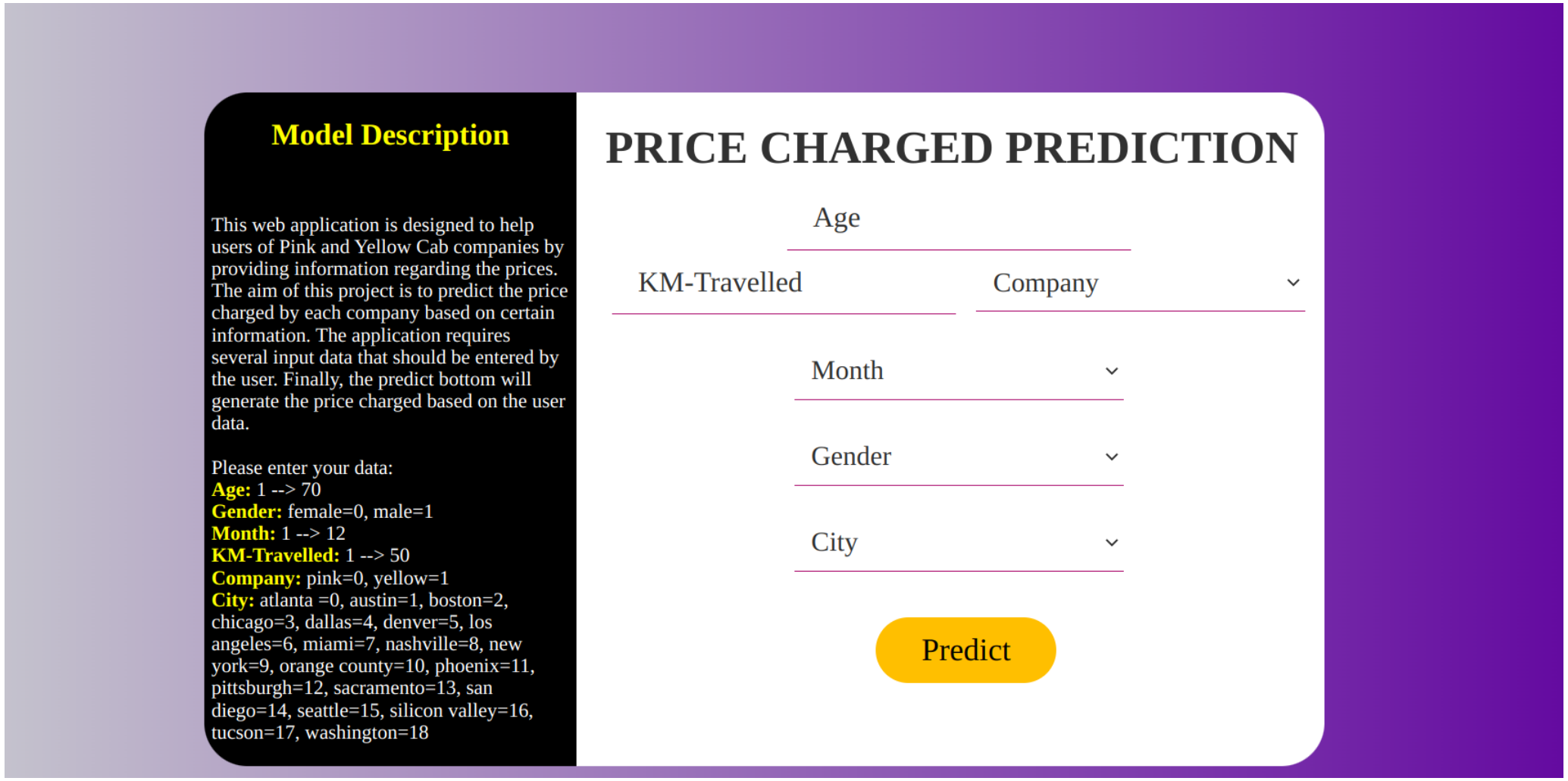
Build the Model:

- In this section I will build a model that predict the price charged for a certain trip based on several information. To do so, first i will remove the features that have no impact on the price in order to simplify the complexity of the model.

- Based on the above analysis it was found that the Gender, Age, Company, City, KM Travelled, Month have considerable effect on the price. It should be noted that number of users (per city), population (per city), and cost of trip also have effect on the price but I will not use them since these information requires prior knowledge and I would like to generate a model that can be used by any user or company.


Price Charged
1200
1000
800
600
400
200
0
0
25
50
75
100
125
150
175
200

Model Deployment:



Thank You