Exploratory Data Analysis

Project name: G2M Insight for Cab Investment Firm
Data Analyst Virtual Internship
Date: 21-Aug-2022

Agenda

Executive Summary

Problem Statement

Approach

EDA

EDA Summary

Recommendations

EXECUTIVE SUMMARY

- Private equity business XYZ is based in the US. It is preparing to make an investment in the Cab industry as a result of the sector's impressive recent expansion and the presence of numerous important players.

- PROBLEM STATEMENT:

In this project, we'll give our customers insightful information on the market for the Cab sector using an exploratory data analysis (EDA) methodology in order to assist them in making a decision before investing. To determine which cab company offers the best investment opportunity, our analysis compares the two (Pink cab company and Yellow cab company).

- DATA ANALYSIS:

The analysis has been divided into 5 sections:

❑ Data Exploration

❑Exploratory Data Analysis

❑Identifying the most profitable Cab Company

❑Hypothesis Testing

❑Recommendations for investment

DATA EXPLORATION

- Description of Datasets:

Four datasets are provided for this analysis:

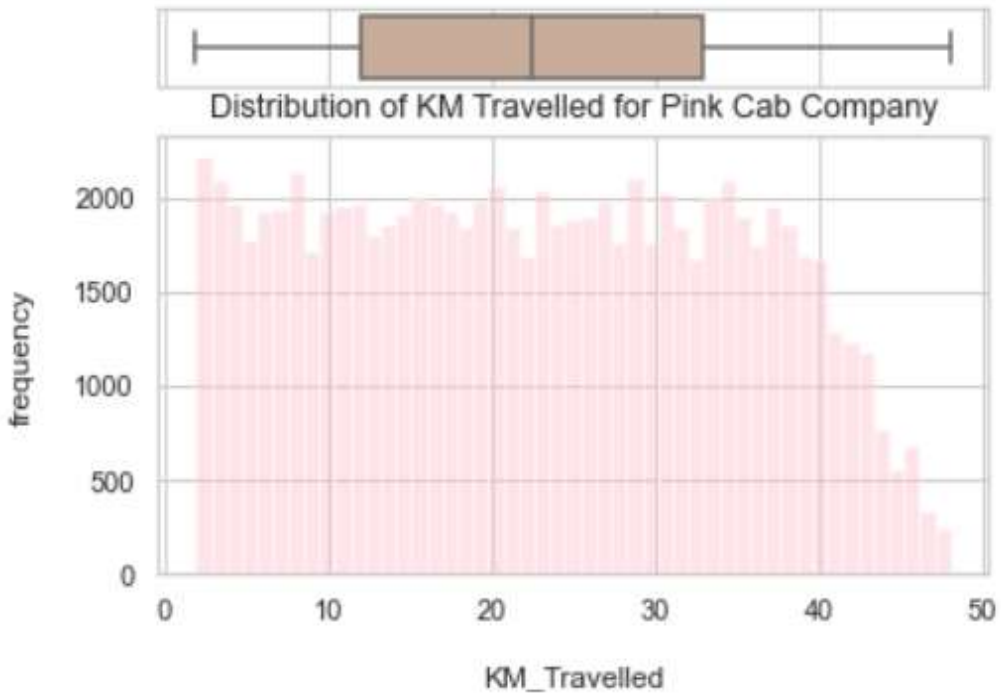
1. Cab_Data.csv: This file lists the details of Transactions, including the companies involved, the distance travelled, the cost, etc.
2. Customer_ID: this file contains the distinct customer ids, along with the customers' ages and incomes.
3. Transaction_ID: this file contains Transaction Ids along with the payment mode.
4. City.csv: this file lists different cities together with their populations and user counts.

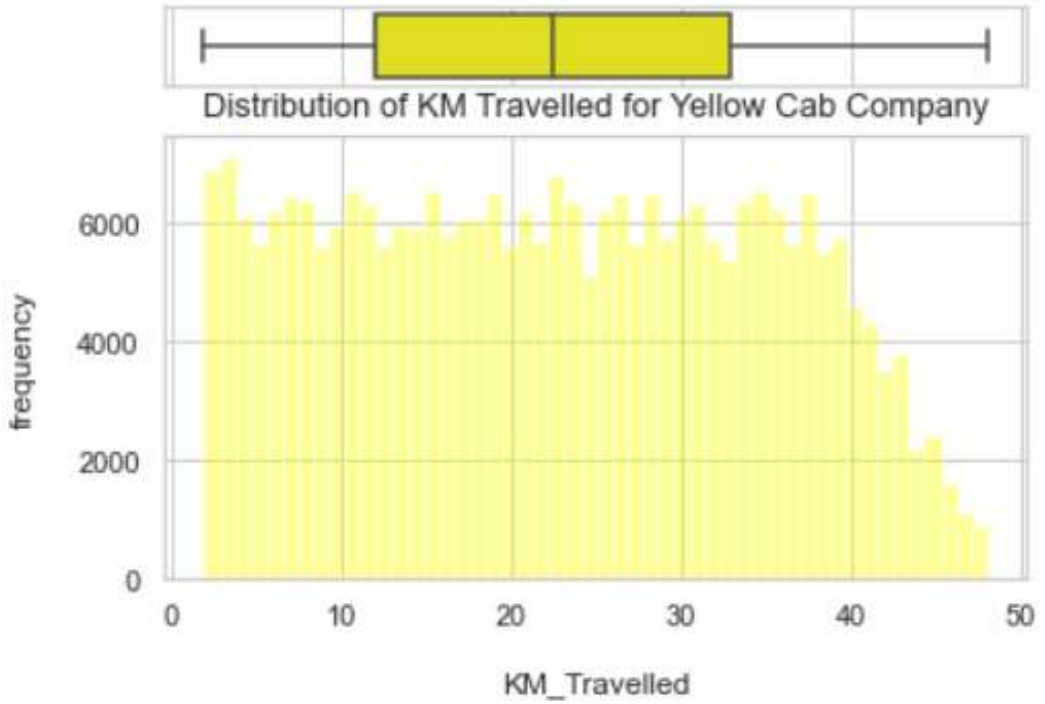
- Assumptions:
- The four datasets listed above are combined to generate the master dataset.
- Timeframe of the data: 2016-01-31 to 2018-12-31.
- Outliers are present in the ‘Price_Charged’ feature. Due to the lack of additional information, we are not treating this as an outlier.
- In the datasets, there are neither duplicated rows nor missing values.
- The "Profit" feature is calculated using the following formula:

  Profit = Price Charged – Cost of Trip

EXPLORATORY DATA ANALYSIS

Distribution of the ‘KM Travelled’ feature for both Companies





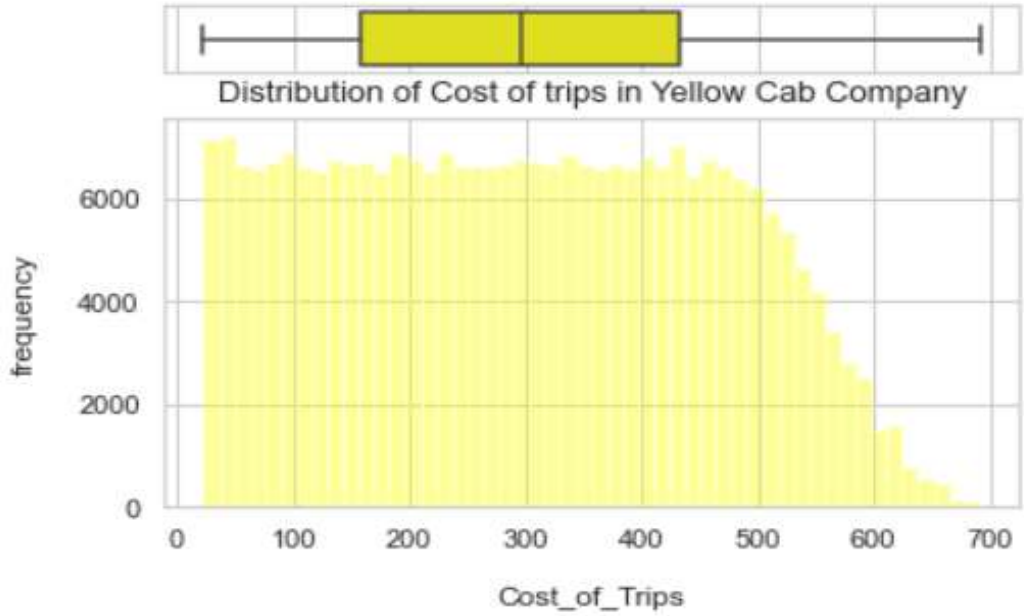
- Most of the rides for both companies vary from 2 to 48 KM.
- Yellow cab company has more frequent rides than the Pink cab company.

Distribution of the ‘Price Charged’ feature for both Companies

- The price charged for Pink cab company ranges between: 15.6 $ and 1623.48 $
- The price charged for Yellow cab company ranges between: 20.73 $ and 2048.03 $

➢ The price charged range for Yellow cab company is higher than the Pink cab company.

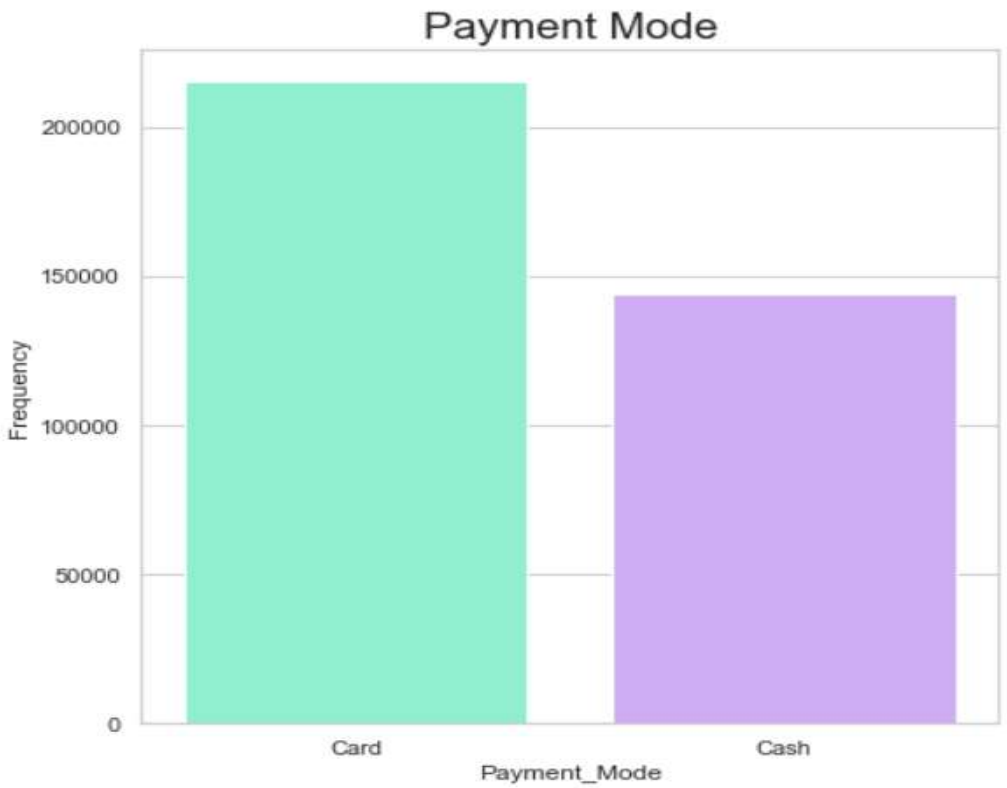
Distribution of ‘Cost of Trip’ feature for both Companies

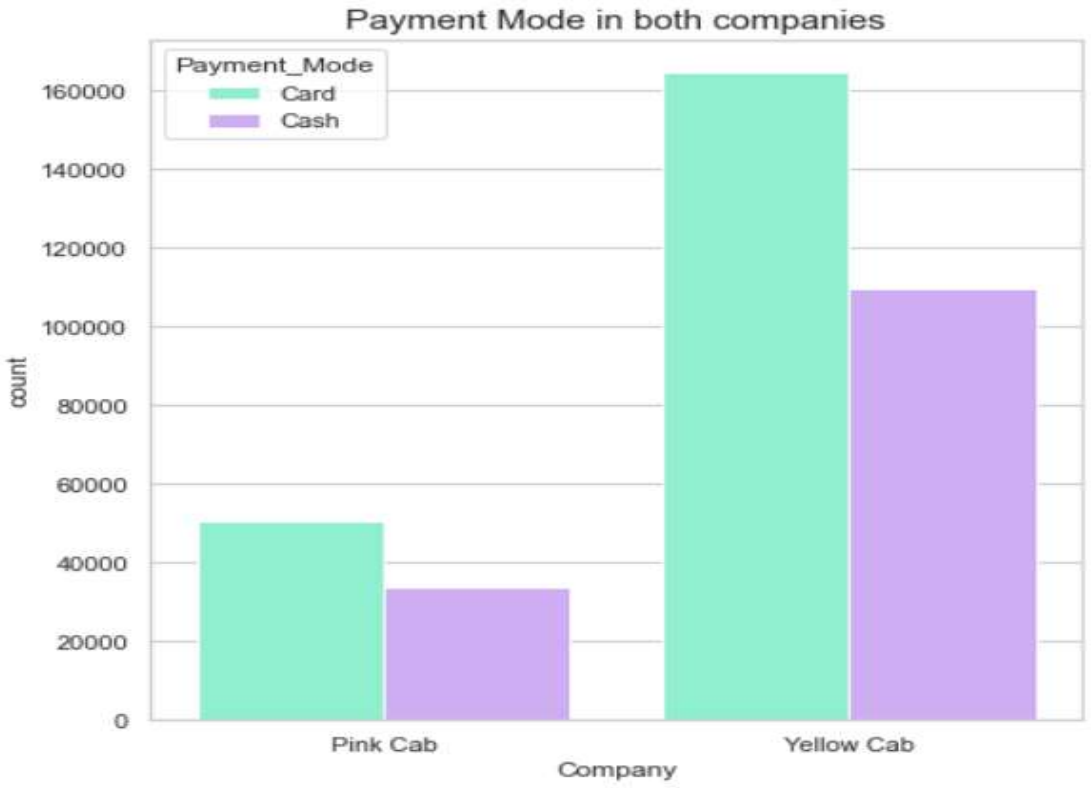


- The cost for Pink cab company ranges between: 19.0 $ and 576.0 $
- The cost for Yellow cab company ranges between: 22.8 $ and 691.2 $

➢ The Cost of Trip range for Yellow cab company is higher than the Pink cab company (expected since the price charged is also higher)

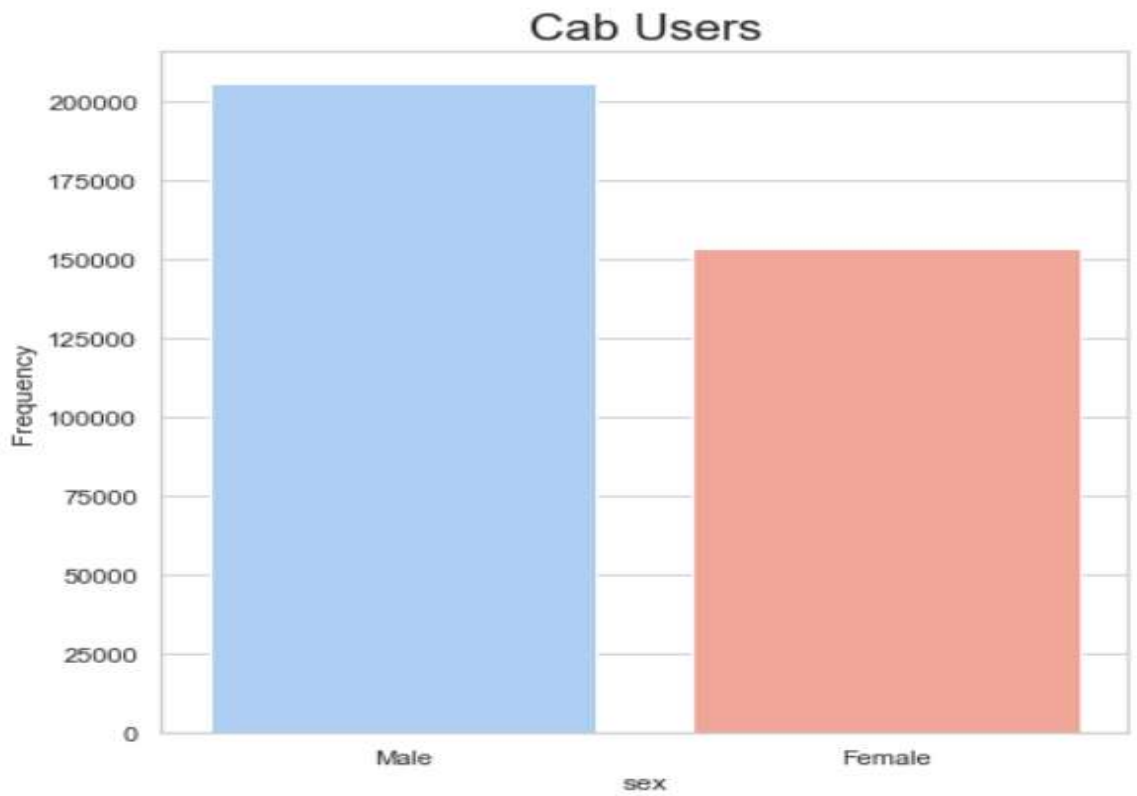
Distribution of ‘Payment Mode’ feature for both Companies

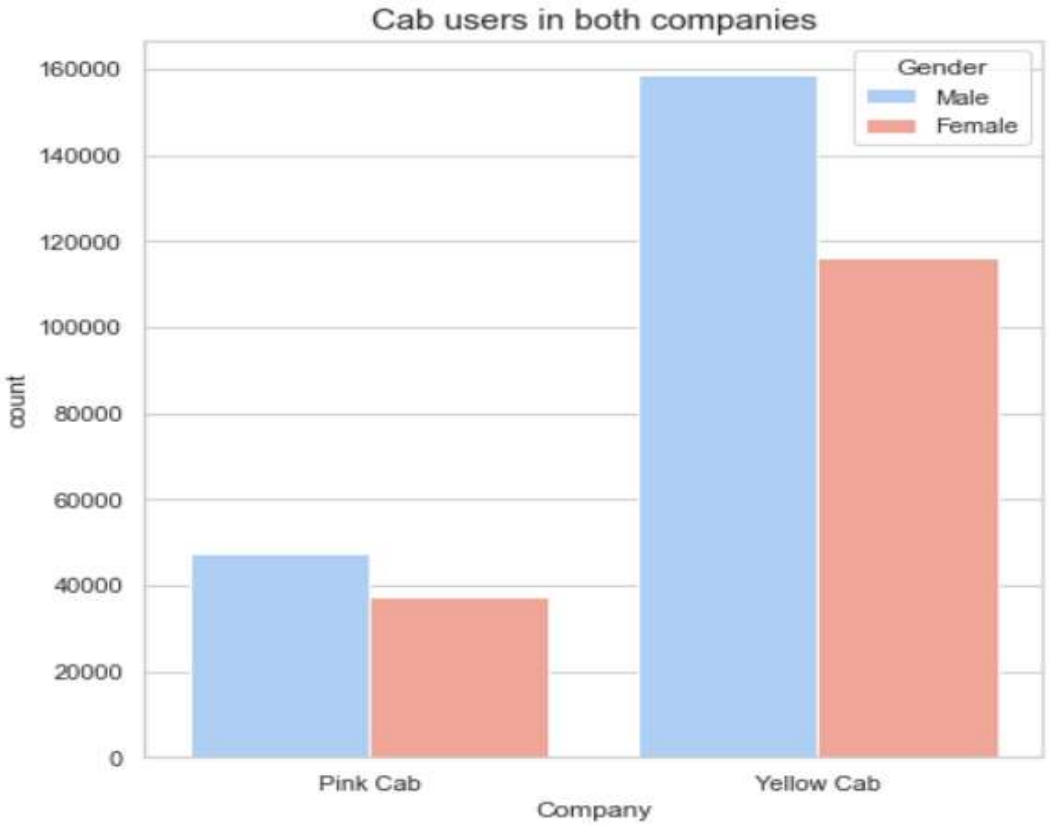




- Cab users prefer to pay with card for their rides than cash.

Distribution of ‘Gender’ feature for both Companies



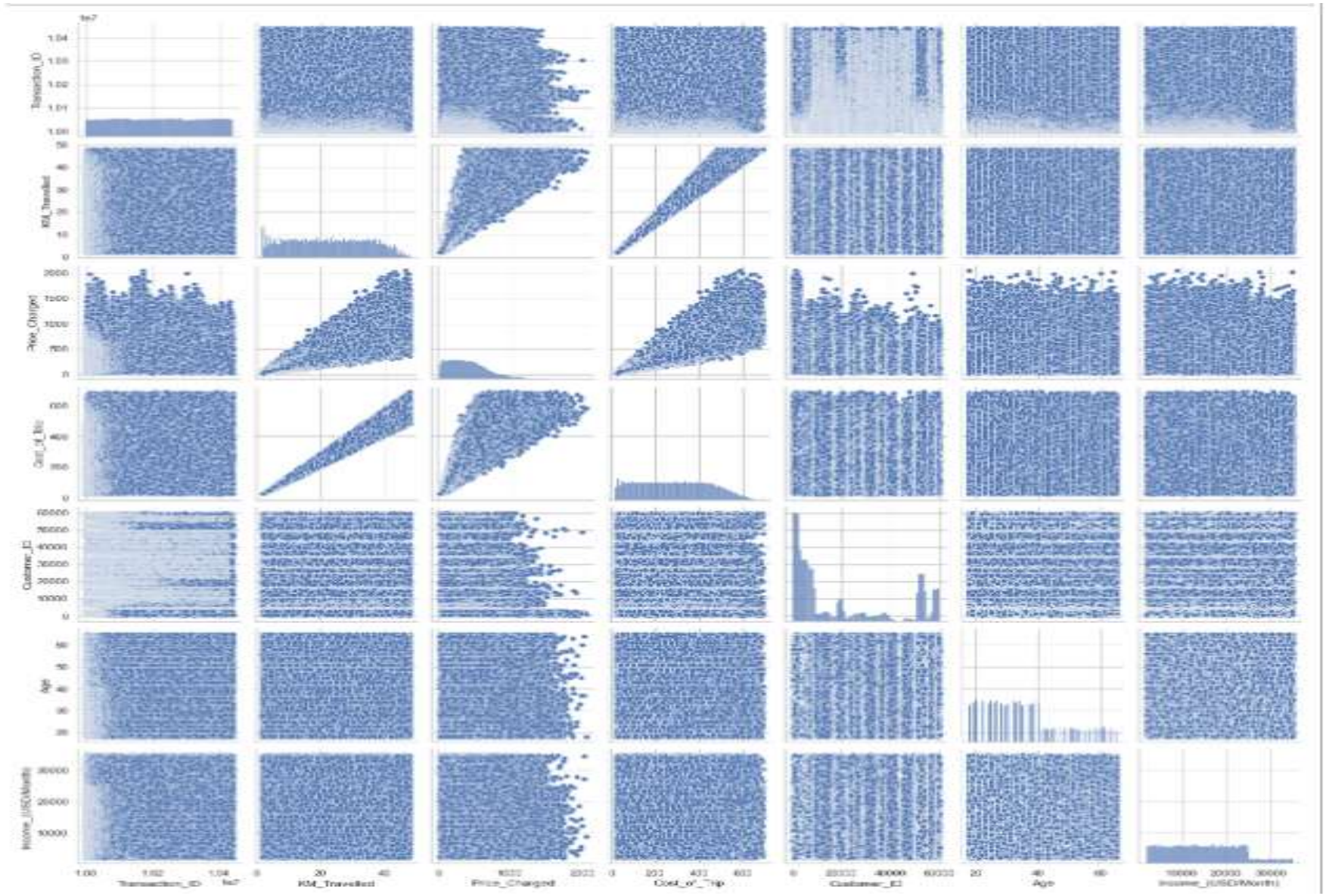


- Most of the female cab users prefer taking the Yellow cab.

Correlation

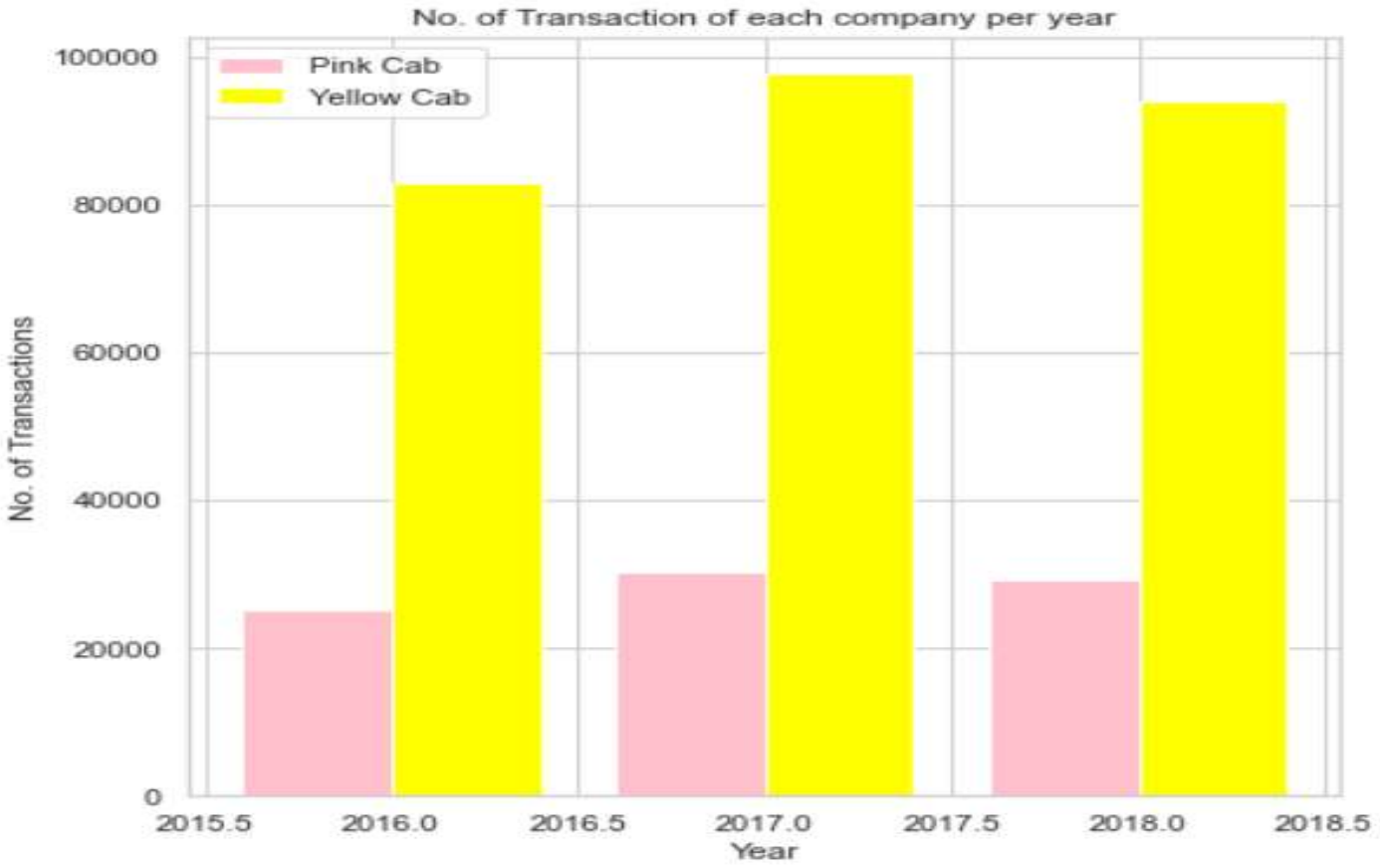
- There is a positive correlation between ‘Price Charged’, ‘KM Travelled’, and ‘Cost of Trip’.

Correlation



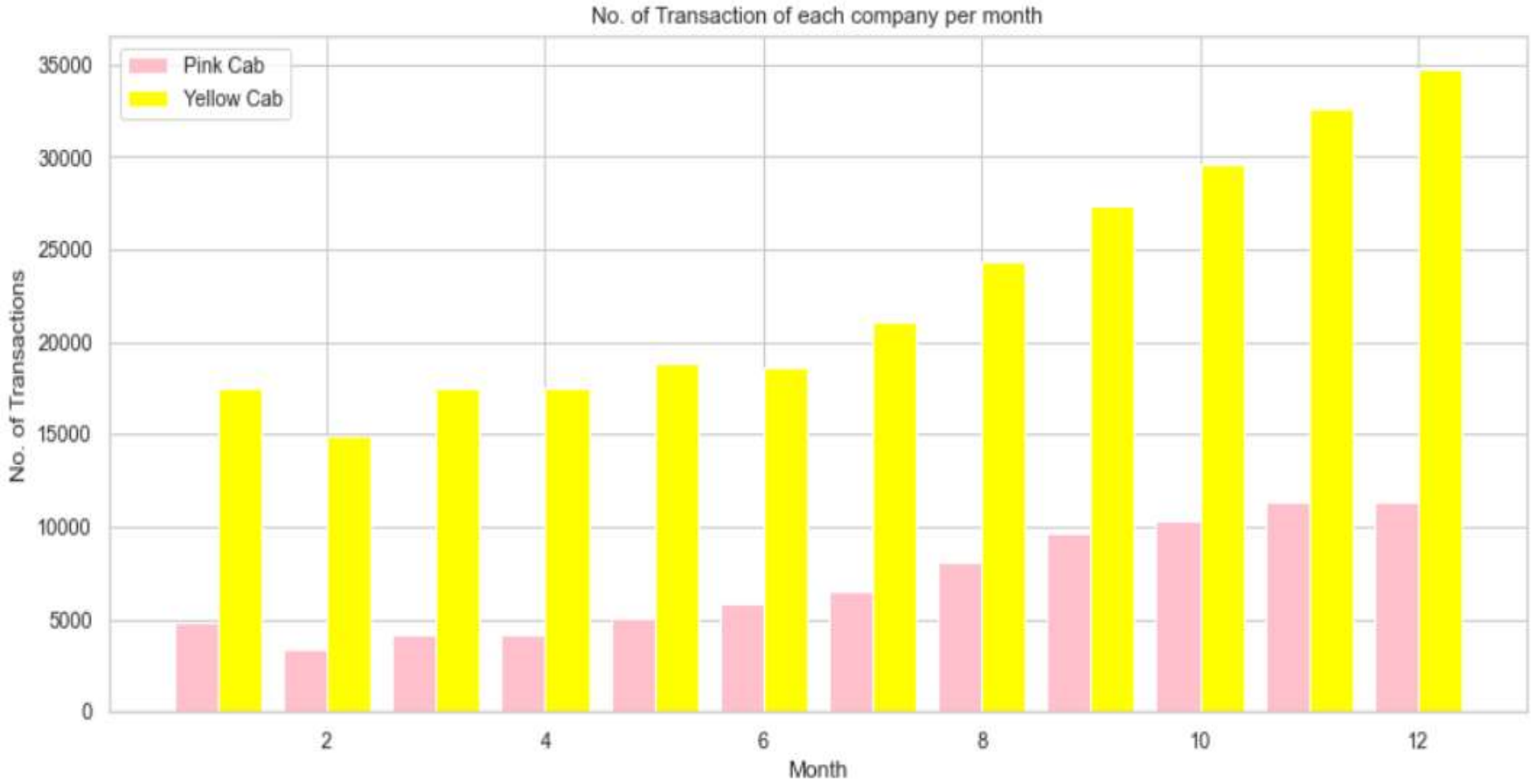
- This pairs plot allows us to see both distributions of single variables and relationships between two variables.

Transactions per year for both Companies



- The Yellow Cab company looks more active than the Pink Cab company on an annual basis.

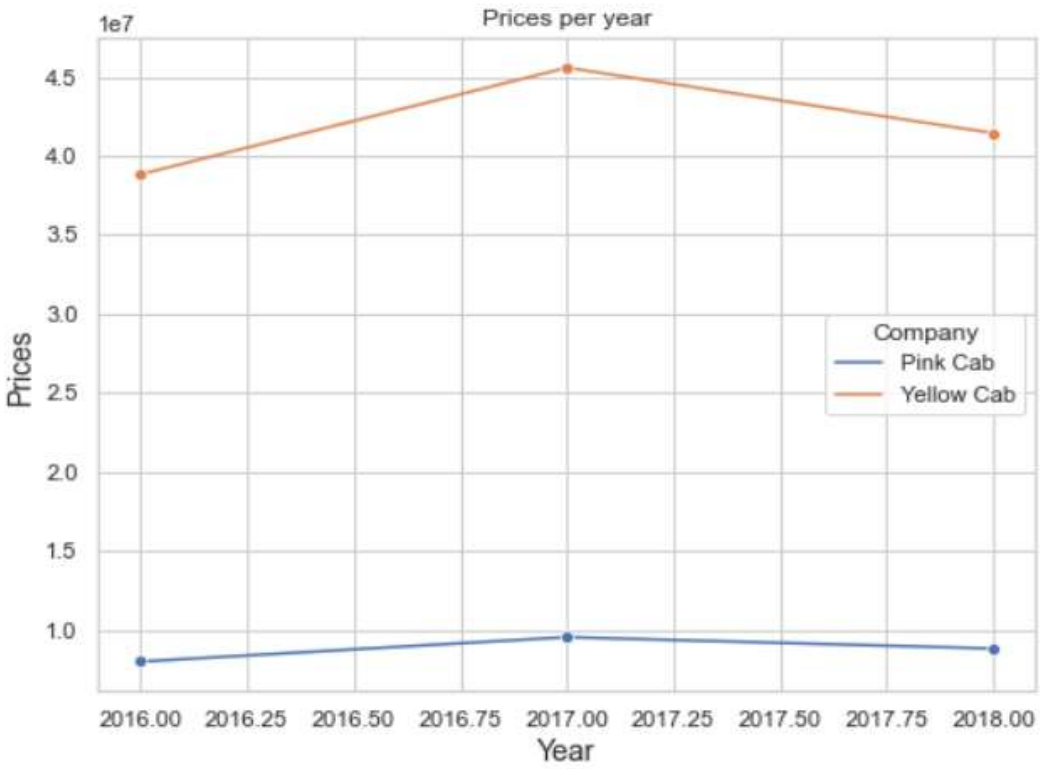
Transactions per month for both Companies

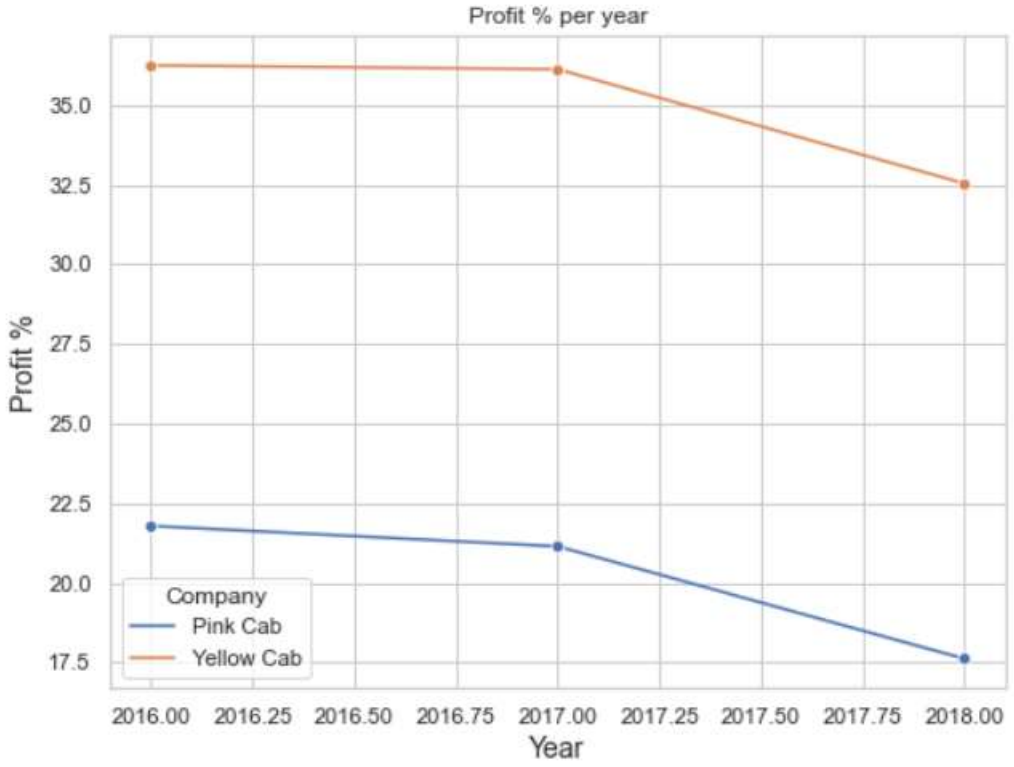


As we can see from this bar plot, on a monthly basis, the Yellow Cab company is in high demand than the pink cab company, especially during the Holiday season.

PROFIT ANALYSIS

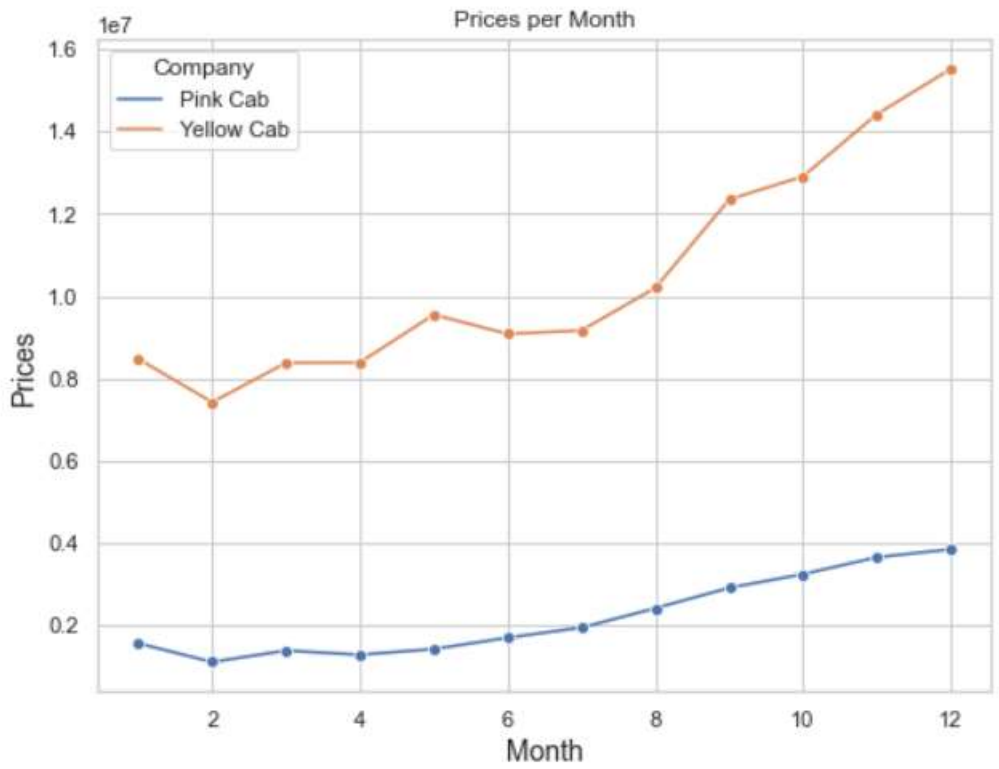
Evolution of Prices and Profit percentage per year

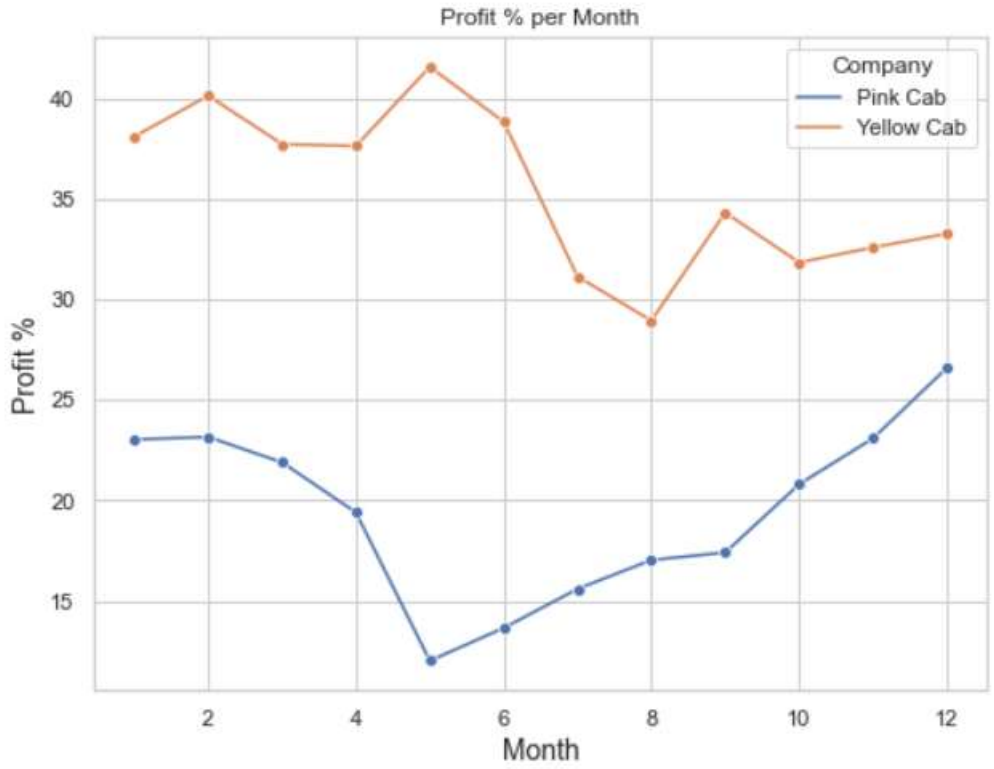




- The percentage of the Profit deviation for the Yellow Cab company is 23.07 %
- The percentage of the Profit deviation for the Pink Cab company is 61.09 %

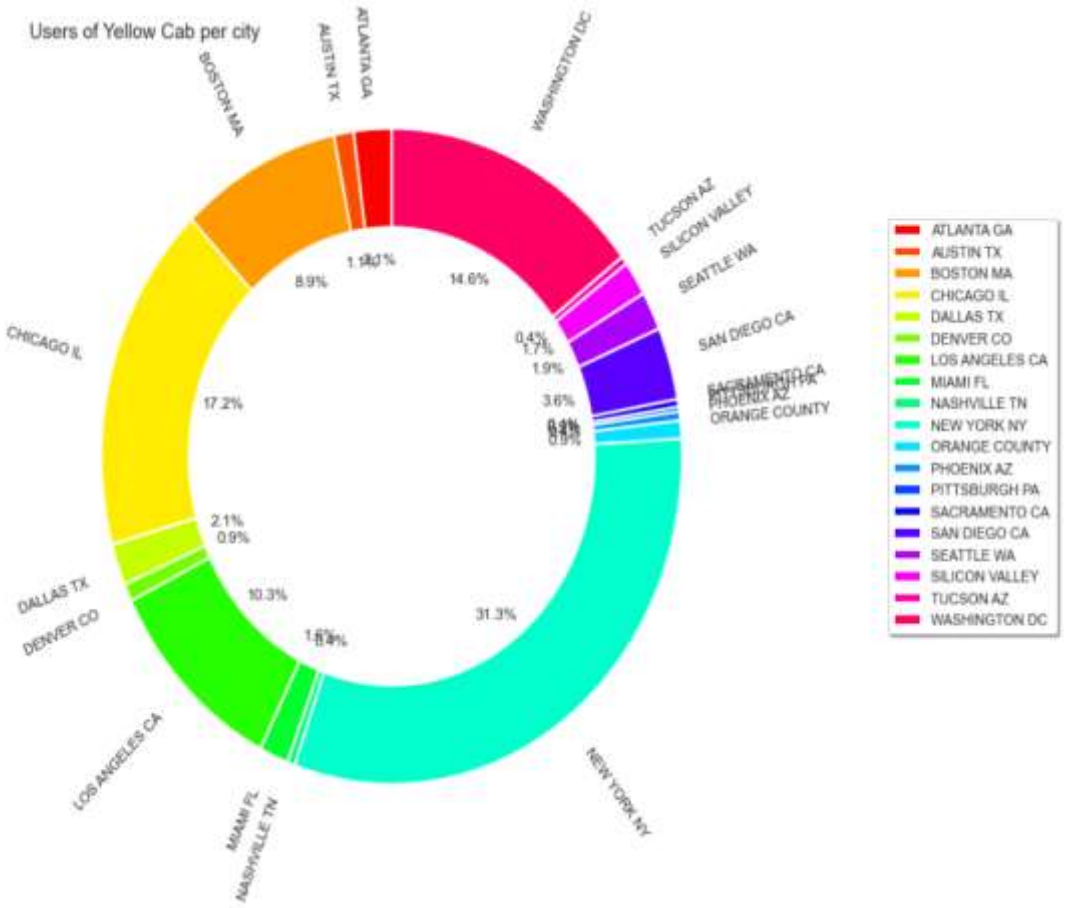
Evolution of Prices and Profit percentage per month

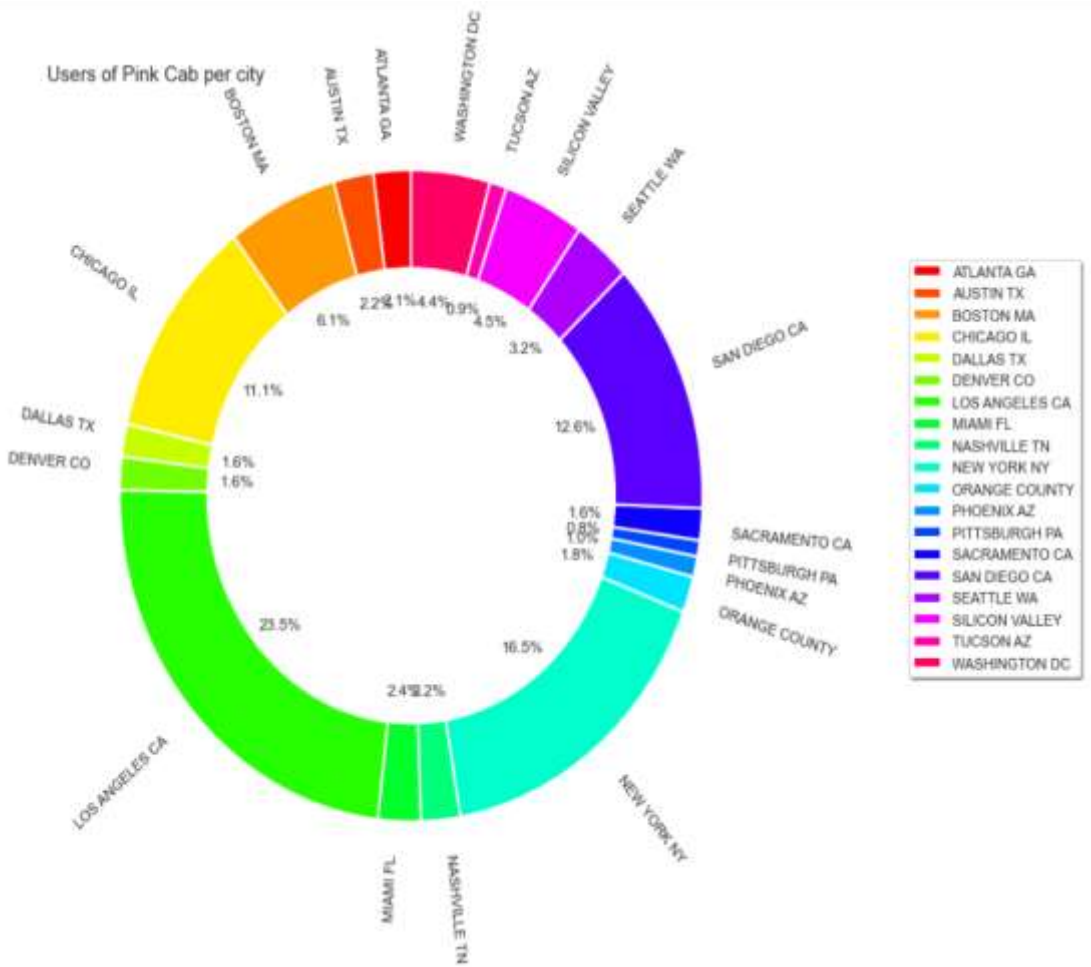




- The earnings of the Pink Cab Company fluctuate by around 61.09%, whereas those of the Yellow Cab Company fluctuate by around 23.07%.

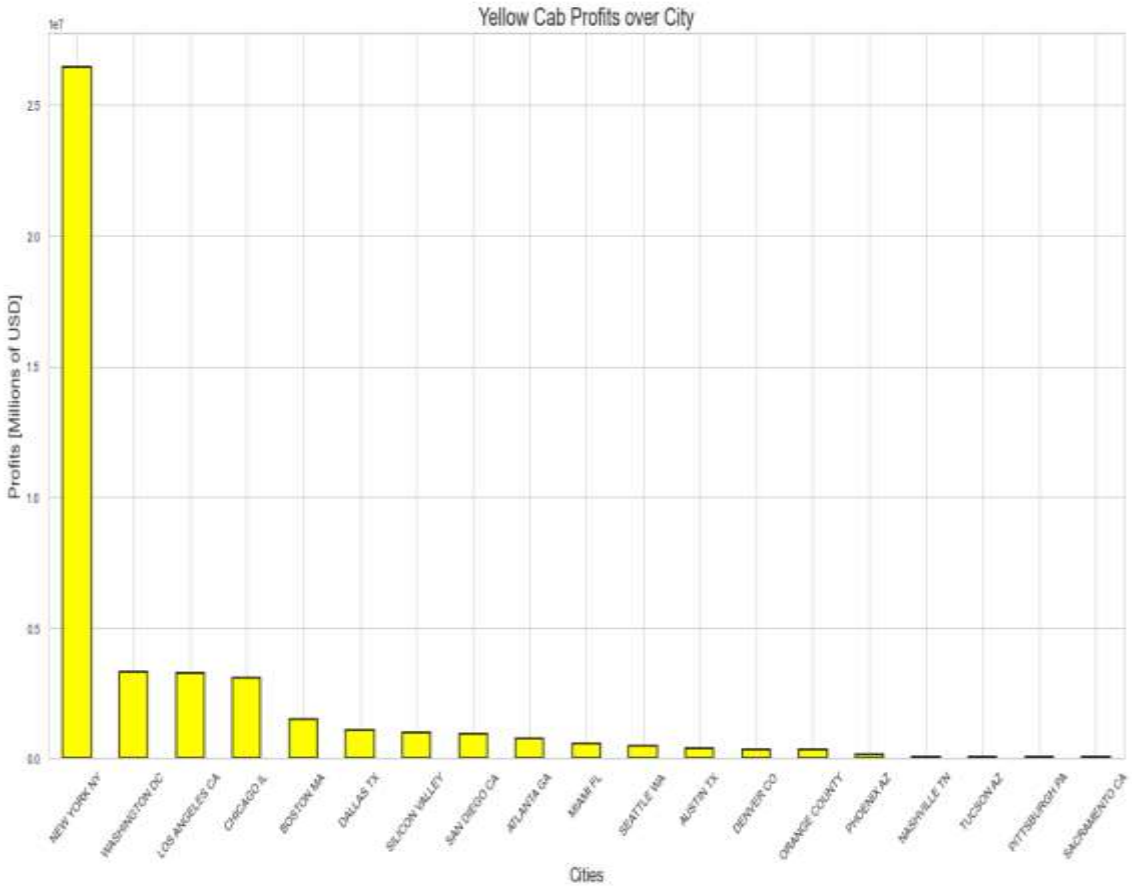
Cab Users per City

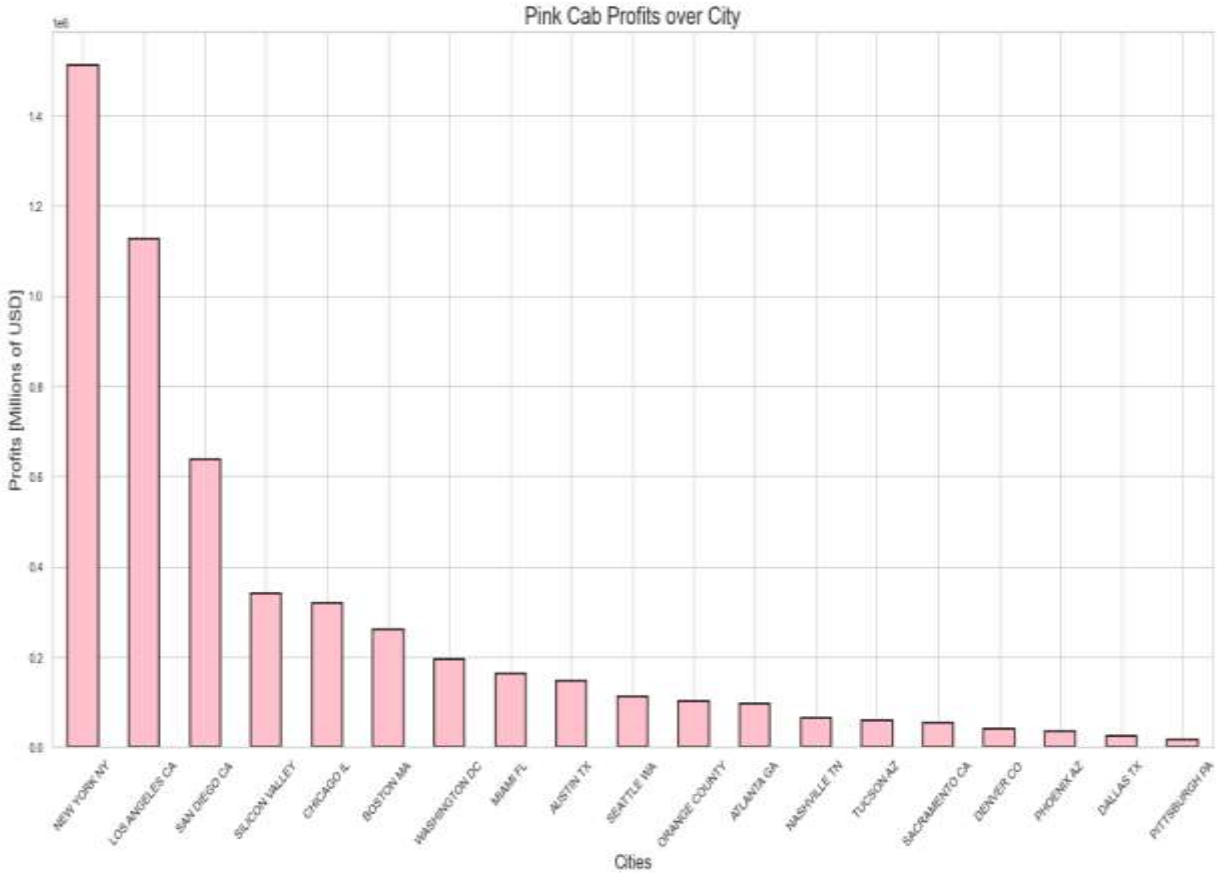




- Transactions for Yellow Cab are highest in New York City which has the highest Cab Users of 47% in total.
- Transaction for Pink Cab is highest in Los Angeles CA City with 34% of users in total.

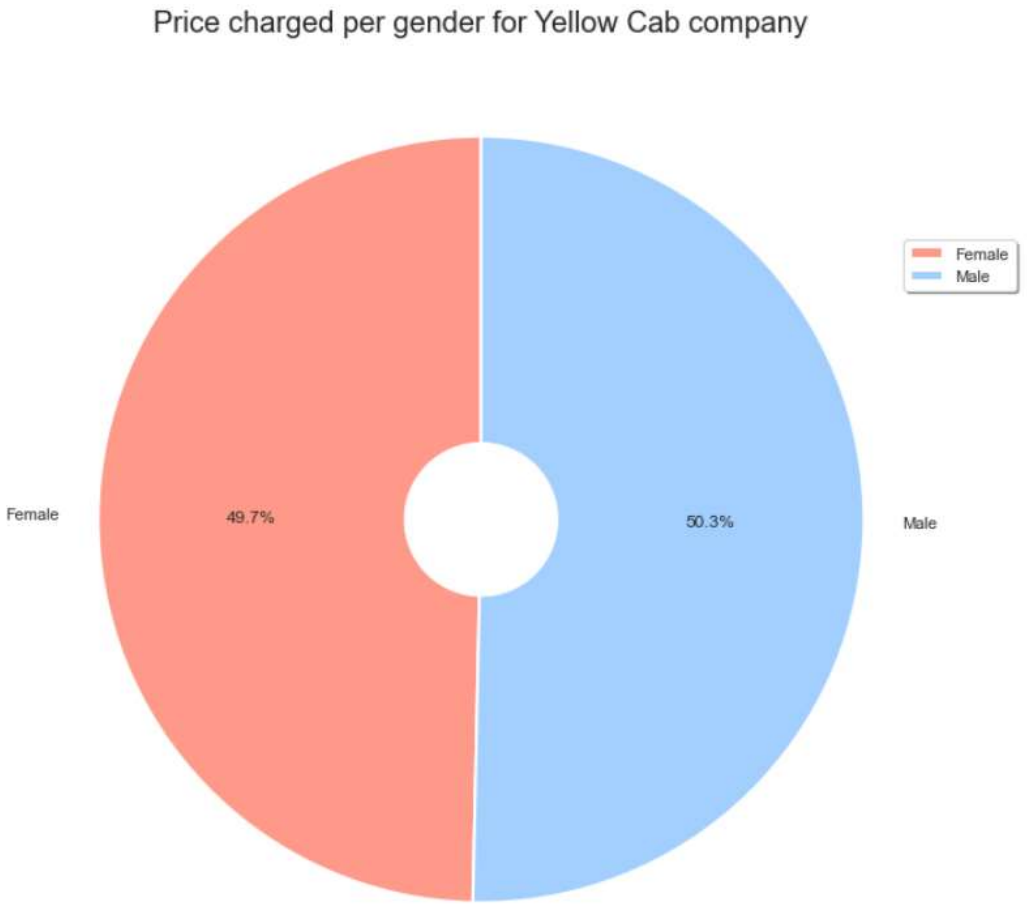
Profit over cities for both companies

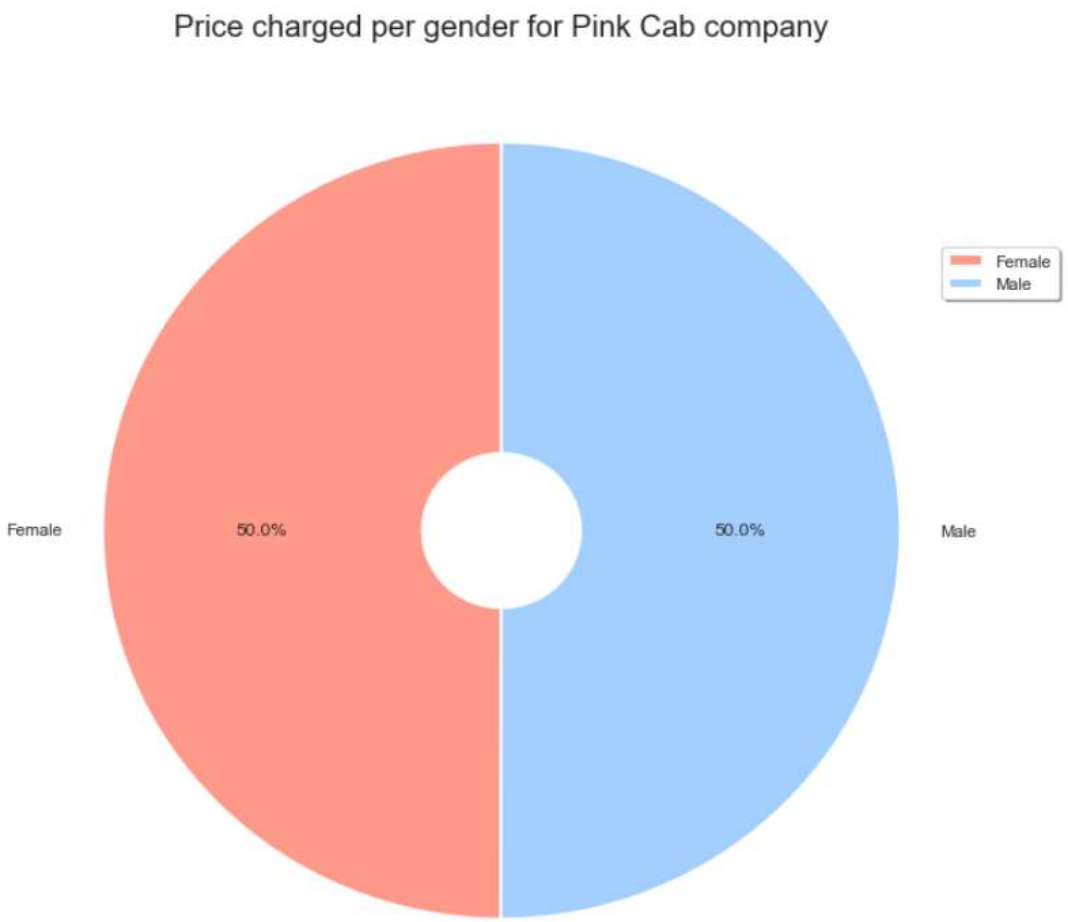




- Since both businesses conduct the majority of their business in New York Metropolis, the Yellow Cab Company's profit in this city exceeds that of the Pink Cab Company.

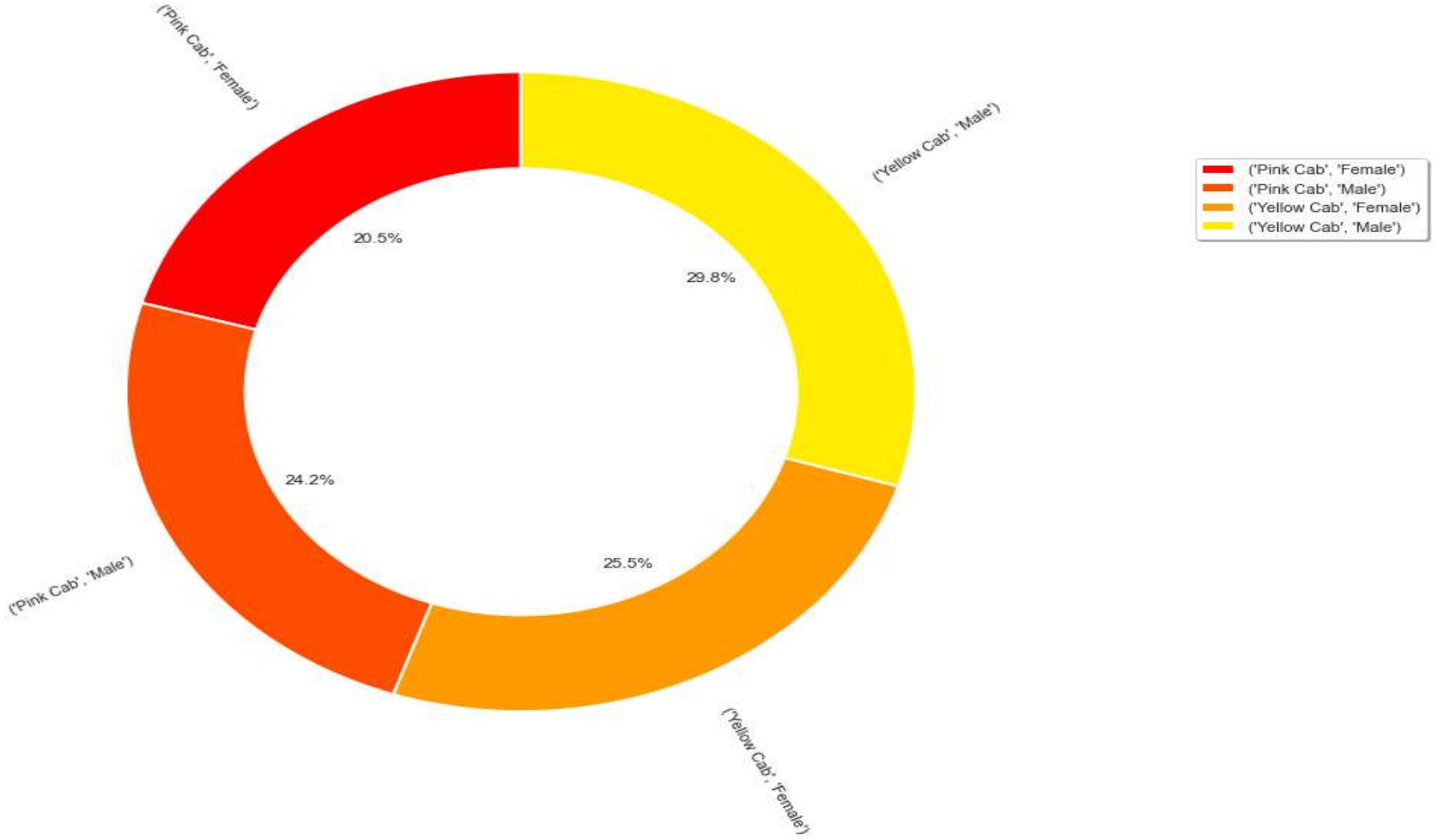
Price charged per gender in both companies





- Pink Cab and Yellow Cab charge the same for both Male and Female Customers, however, Yellow Cab charges less for female customers.

Customer share in both companies



- Female customers in Yellow Cabs are higher (25.5%) than in Pink Cabs (20.5%).

HYPOTHESIS TESTING

Hypothesis 1

- H0: There is no difference regarding Payment Mode in both cab companies.
- H1: There is a difference regarding Payment Mode in both cab companies.

```
print('P value is ', p_value)

if(p_value<0.05):
    print('We accept alternative hypothesis (H1) that there is a difference in payment mode for Pink Cab')
else:
    print('We accept null hypothesis (H0) that there is no difference in payment mode for Pink Cab')
```

```
P value is  0.7900465828793288
We accept null hypothesis (H0) that there is no difference in payment mode for Pink Cab
```

```
print('P value is ', p_value)

if(p_value<0.05):
    print('We accept alternative hypothesis (H1) that there is a difference in payment mode for Yellow Cab')
else:
    print('We accept null hypothesis (H0) that there is no difference in payment mode for Yellow Cab')
```

```
P value is  0.29330606382985325
We accept null hypothesis (H0) that there is no difference in payment mode for Yellow Cab
```

➢ There is no difference in payment mode for both cab companies.

Hypothesis 2

- H0: There is no difference regarding Gender in both cab companies.
- H1: There is a difference regarding Gender in both cab companies.

```
print('P value is ', p_value)

if(p_value<0.05):
    print('We accept alternative hypothesis (H1) that there is a difference for Pink Cab')
else:
    print('We accept null hypothesis (H0) that there is no difference for Pink Cab')
```

```
P value is  0.11515305900425798
We accept null hypothesis (H0) that there is no difference for Pink Cab
```

```
print('P value is ', p_value)

if(p_value<0.05):
    print('We accept alternative hypothesis (H1) that there is a difference for Yellow Cab')
else:
    print('We accept null hypothesis (H0) that there is no difference for Yellow Cab')
```

```
P value is  6.060473042494144e-25
We accept alternative hypothesis (H1) that there is a difference for Yellow Cab
```

➢ There is a difference regarding Gender only for Yellow Cab company.

Hypothesis 3

- H0: There is no difference regarding Age in both cab companies.
- H1: There is a difference regarding Age in both cab companies.

```
print('P value is ', p_value)

if(p_value<0.05):
    print('We accept alternative hypothesis (H1) that there is a difference regarding age for Pink Cab')
else:
    print('We accept null hypothesis (H0) that there is no difference regarding age for Pink Cab')
```

```
P value is  0.18796448671958466
We accept null hypothesis (H0) that there is no difference regarding age for Pink Cab
```

```
print('P value is ', p_value)

if(p_value<0.05):
    print('We accept alternative hypothesis (H1) that there is a difference regarding age for Yellow Cab')
else:
    print('We accept null hypothesis (H0) that there is no difference regarding age for Yellow Cab')
```

```
P value is  2.8426722804525463e-07
We accept alternative hypothesis (H1) that there is a difference regarding age for Yellow Cab
```

➢ It appears that Yellow Cab offers discounts to customers who are above 60 years old.

RECOMMENDATIONS

- We have evaluated both Cab companies based on the following points and found out that the Yellow Cab company is better than the Pink Cab company:

❑ Profit Analysis:

- Profits: Higher long-term profits and fewer monthly variations for the yellow cab company.
- City-wise Profit: In every City, Yellow Cab Company has a larger market share.
- Number of Transactions: Yellow Cab Company is more in demand than Pink Cab Company on a monthly basis, particularly during the holiday season.

❑ Client Analysis:

- Payment Mode Distributions: The distribution of Payment Mode over time, cities, and ages is the same for both companies.
- Gender Age Wise: While Pink Cab does not charge differently for any age group, Yellow Cab charges differently for customers over the age of 60.
- Gender: Additionally, Yellow Cab charges less for female customers.

➢ On the basis of the above point, we will recommend Yellow cab for investment.

Thank You