Exploratory Data Analysis Presentation

Project name: **G2M Insight for Cab Investment Firm**

Team: **Data & Analytics**

Date: **June 25th, 2021**

Executive Summary

Data Exploration

EDA

Hypothesis Testing

Recommendations

Agenda

Executive Summary

- **Problem Statement:**

  In this project, we are going to provide fruitful insight through an EDA (Exploratory Data Analysis) approach about the Cab industry market for our client in order to help them take a final decision before investment. Our analysis is based on a comparison between two cab companies (Pink cab company / Yellow cab company) to show which one represents the best opportunity to invest in.

- **The client:**

  XYZ is a private firm in US.

- **Analysis:**

  The analysis has been divided into **5 parts**:

  - Data Exploration.
  - EDA
  - Finding the most profitable Cab company.
  - Hypothesis Testing.
  - Recommendations for investment.

Data Exploration

- **Datasets Description:**

  For this analysis, there are **4 datasets** provided:

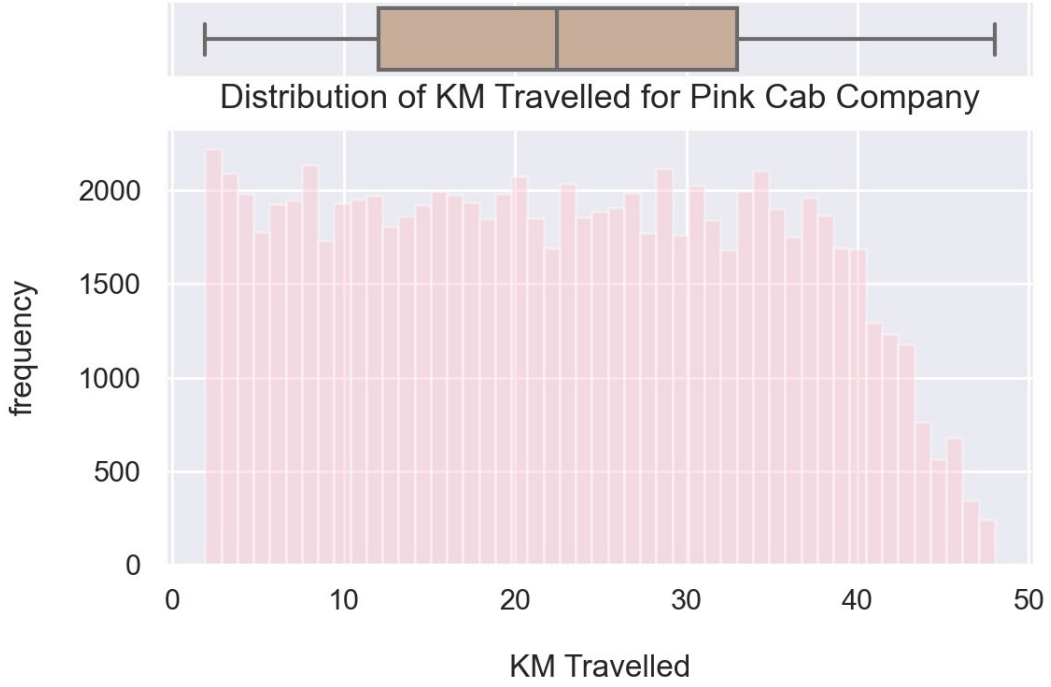
1) **cab_data.csv** : this file describes attributes of Transactions like Companies, Km travelled, price charged etc.
2) **Customer_ID.csv** : this file consists of unique customer ids with their ages and income.
3) **Transaction_ID.csv** : this file consists of Transaction Ids with the payment mode.
4) **City.csv** : this file consists of various cities, their populations and number of users.

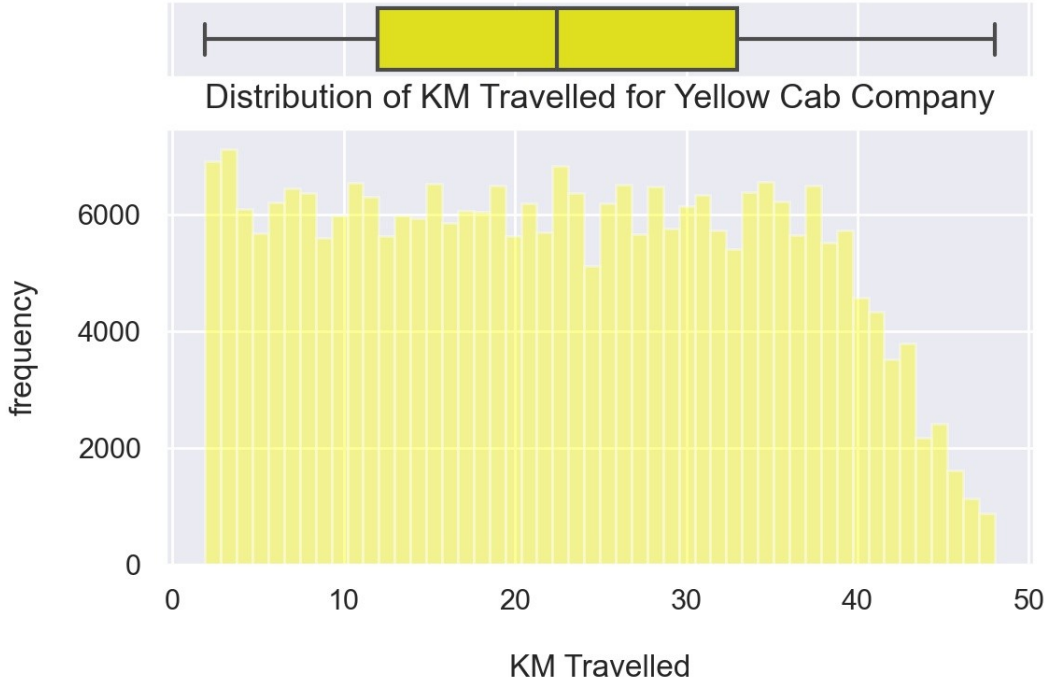
- **Assumptions:**
- Time frame of the data : **2016-01-31** to **2018-12-31**.
- The main dataset is created by **merging** mentioned 4 datasets.
- **Outliers** are present in 'Price Charged' feature. We are not treating this as outliers because of unavailability of more details.
- There are no **duplicated rows** neither **missing values** in the datasets.
- The 'Profit' feature is calculated as follows:

  Profit = Price Charged – Cost of Trip


EDA

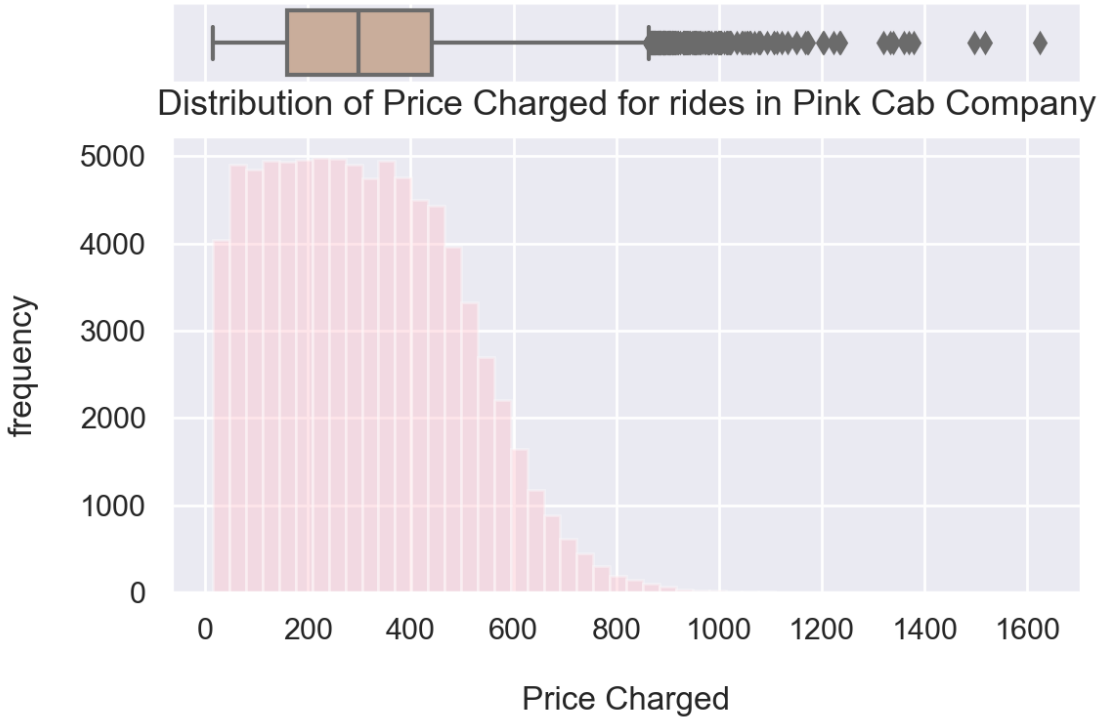
Distribution of ‘KM Traveled’ feature for both Companies

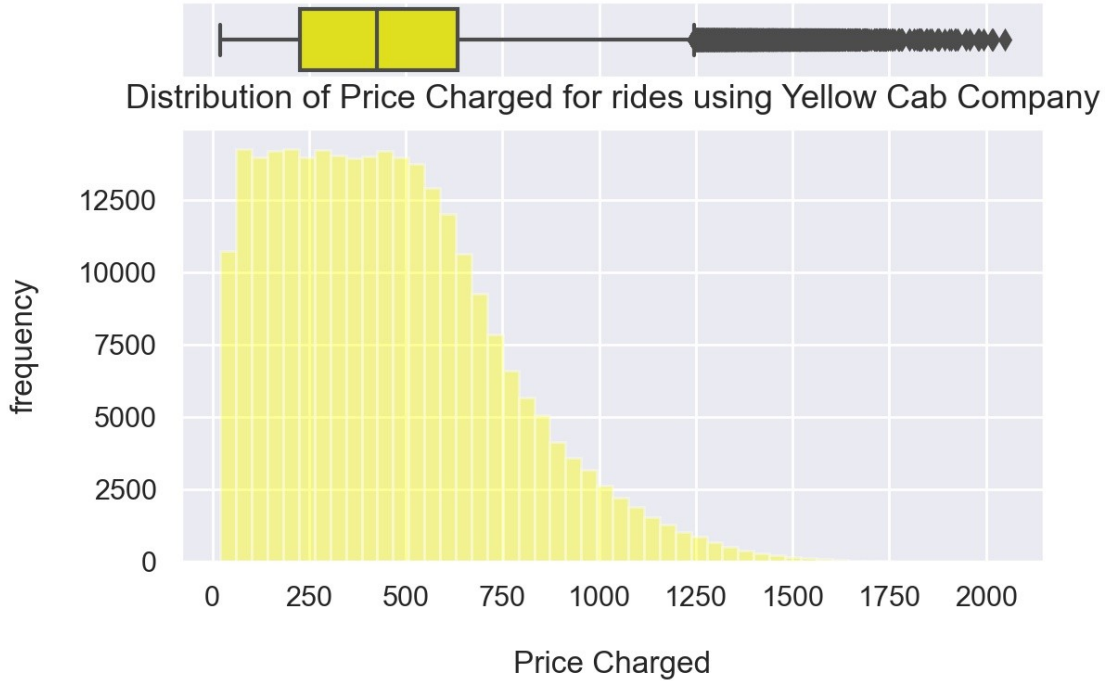




- Most of the rides for both companies varies from 2 to 48 KM.
- Yellow cab company has more frequent rides than the Pink cab company.

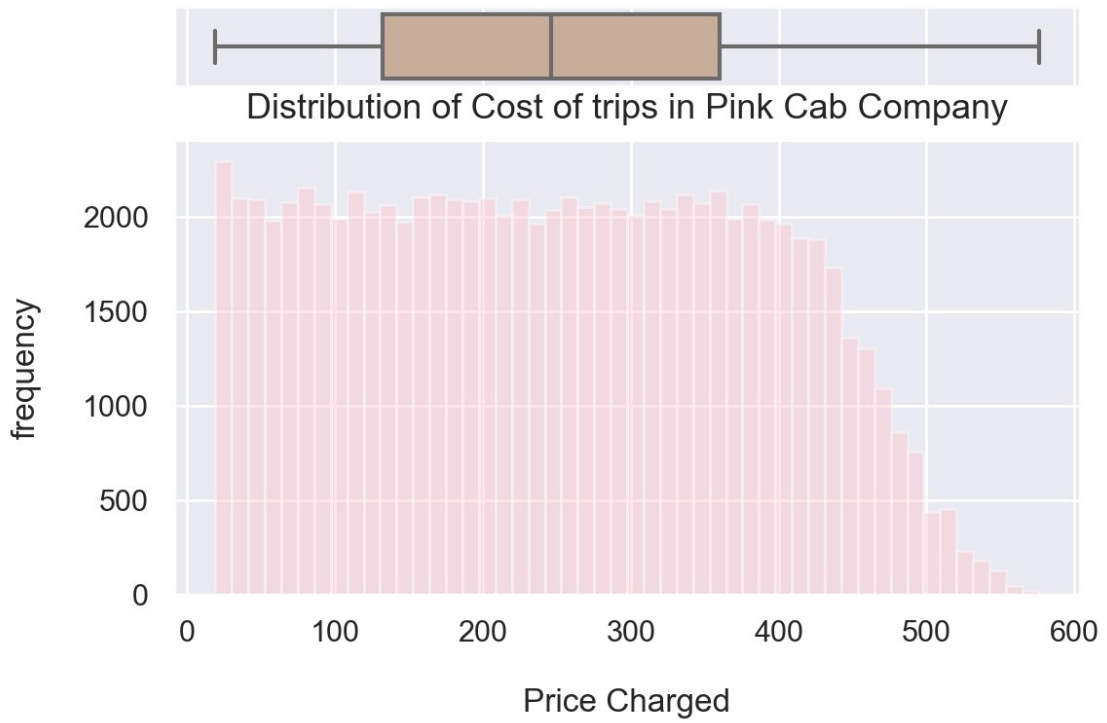
Distribution of ‘Price Charged’ feature for both Companies

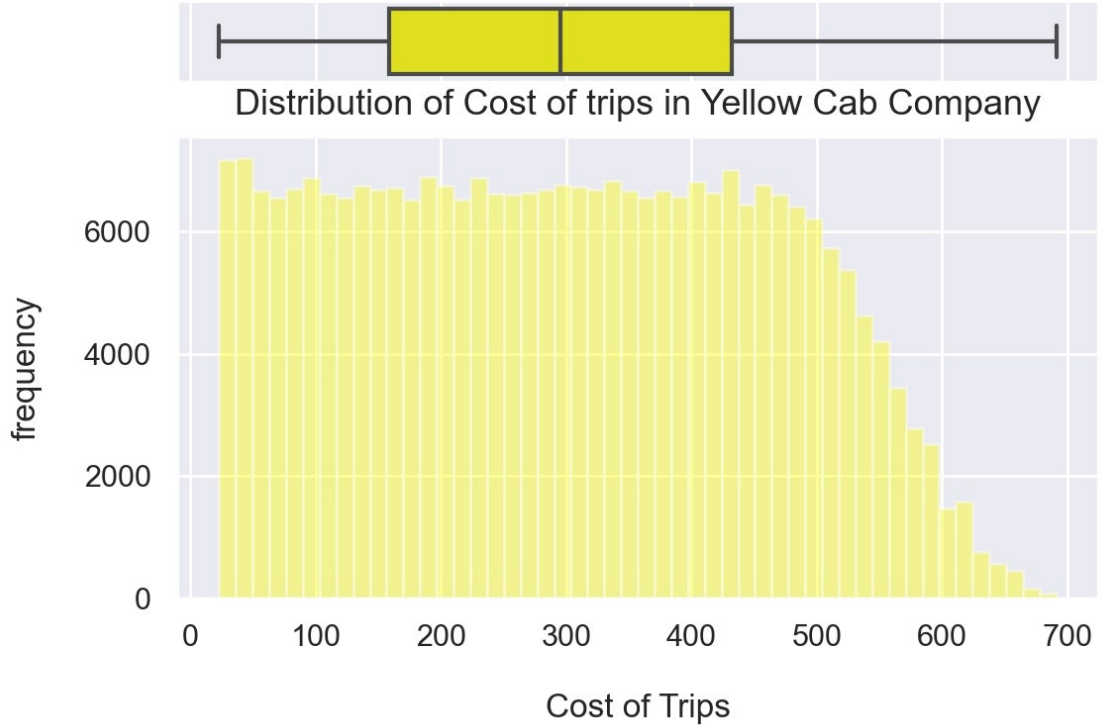




- The price charged for Pink cab company ranges between : 15.6 $ and 1623.48 $
- The price charged for Yellow cab company ranges between : 20.73 $ and 2048.03 $
- ➔ The price charged range for Yellow cab company is higher than the Pink cab company.

Distribution of ‘Cost of Trip’ feature for both Companies

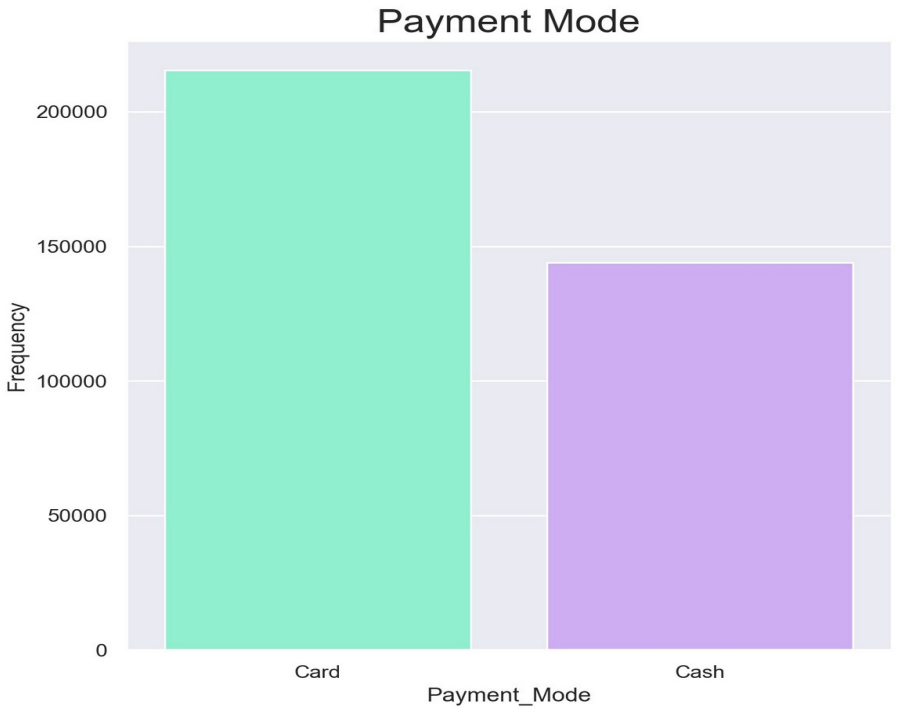


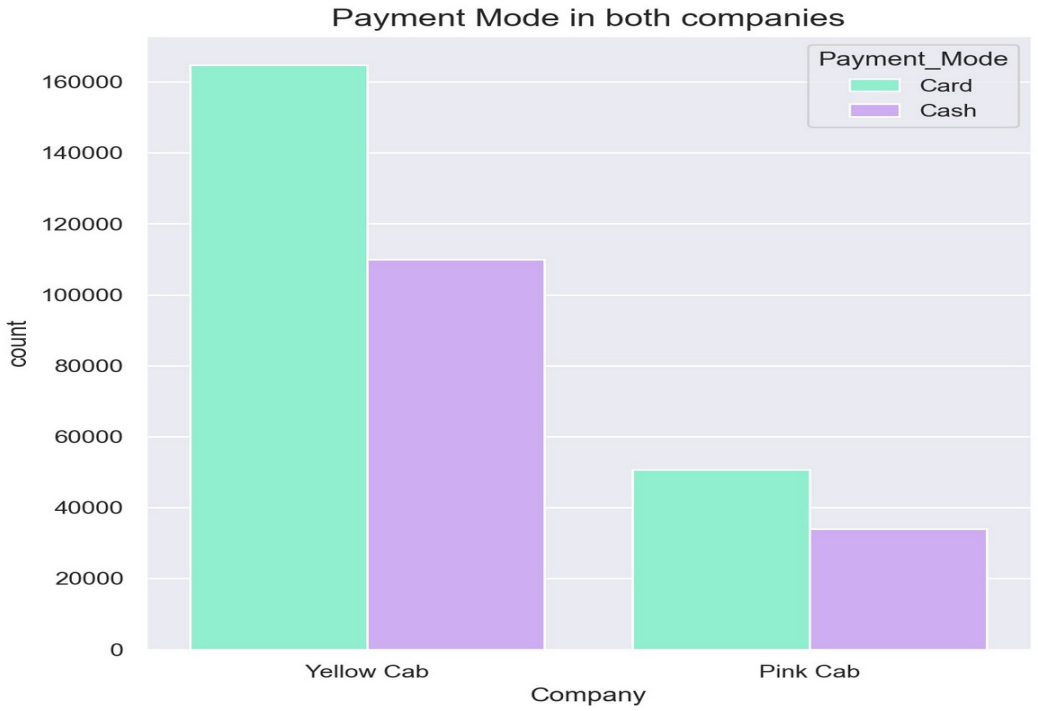


- The cost for Pink cab company ranges between : 19.0 $ and 576.0 $
- The cost for Yellow cab company ranges between : 22.8 $ and 691.2 $

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

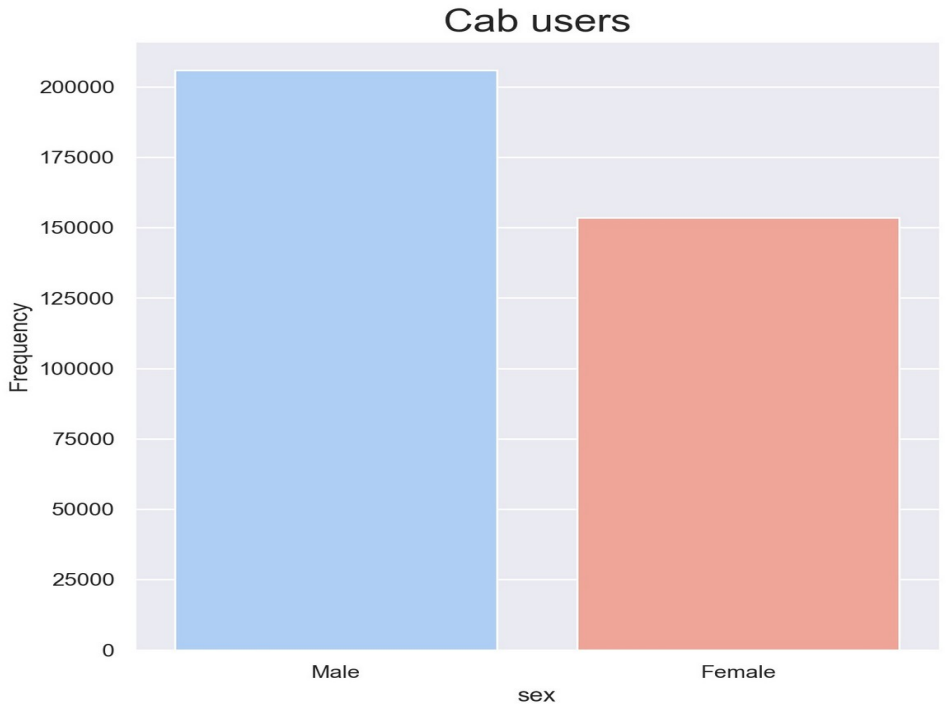
Distribution of ‘Payment Mode’ feature for both Companies

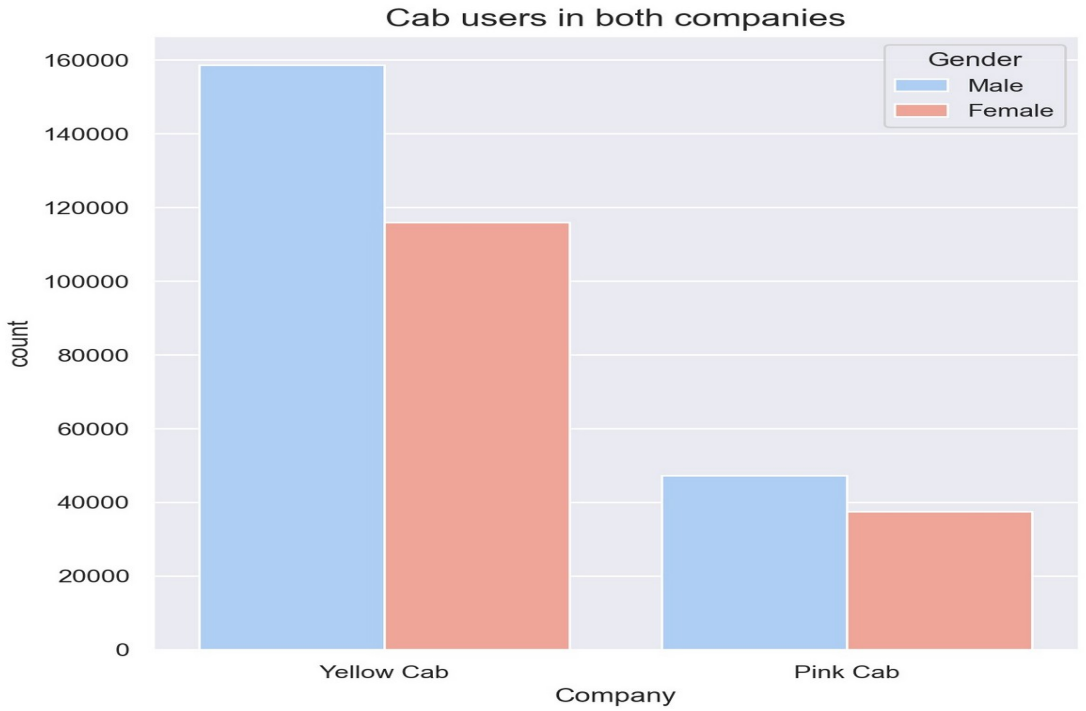




- Cab users prefer to pay with card for their rides.

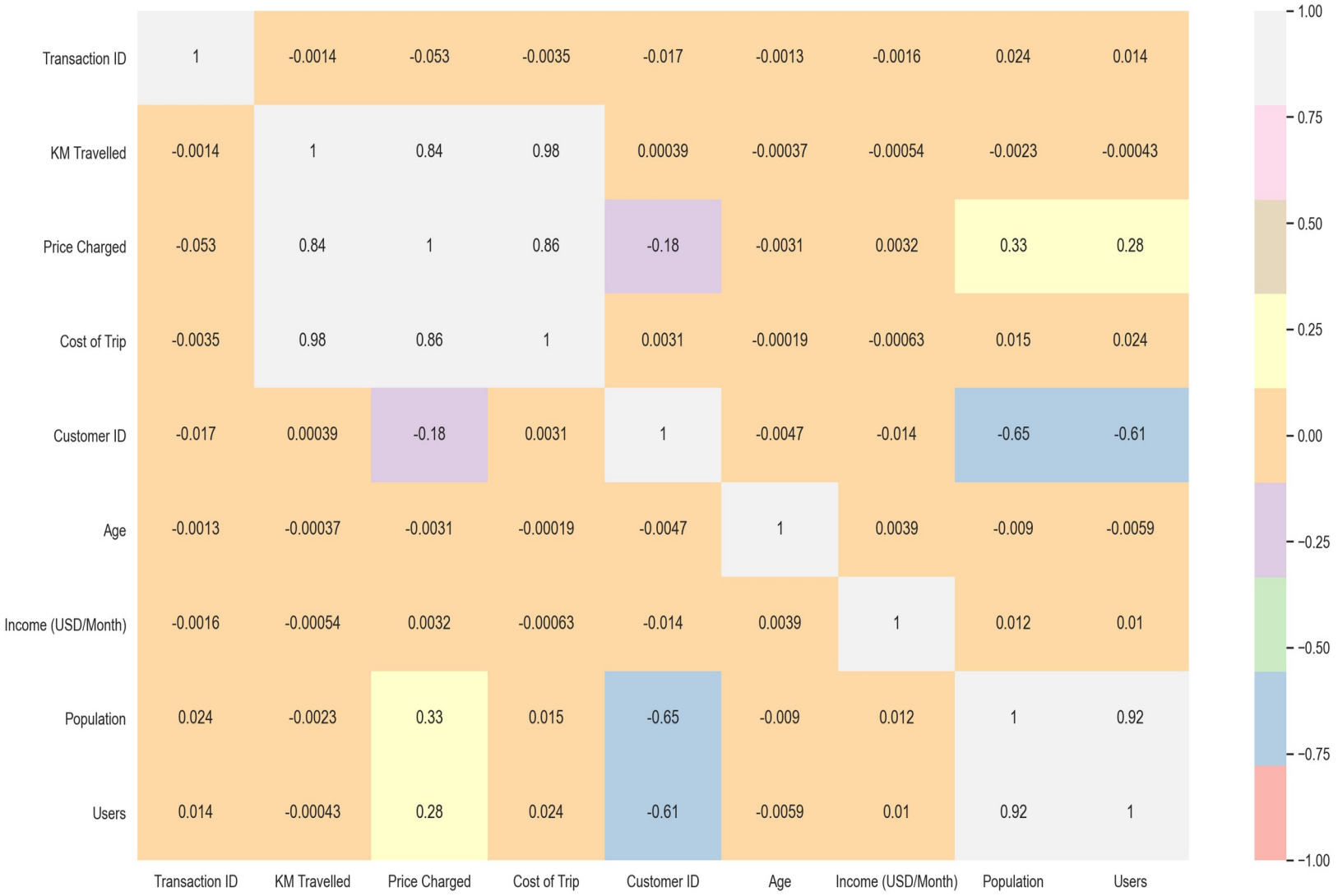
Distribution of ‘Gender’ feature for both Companies





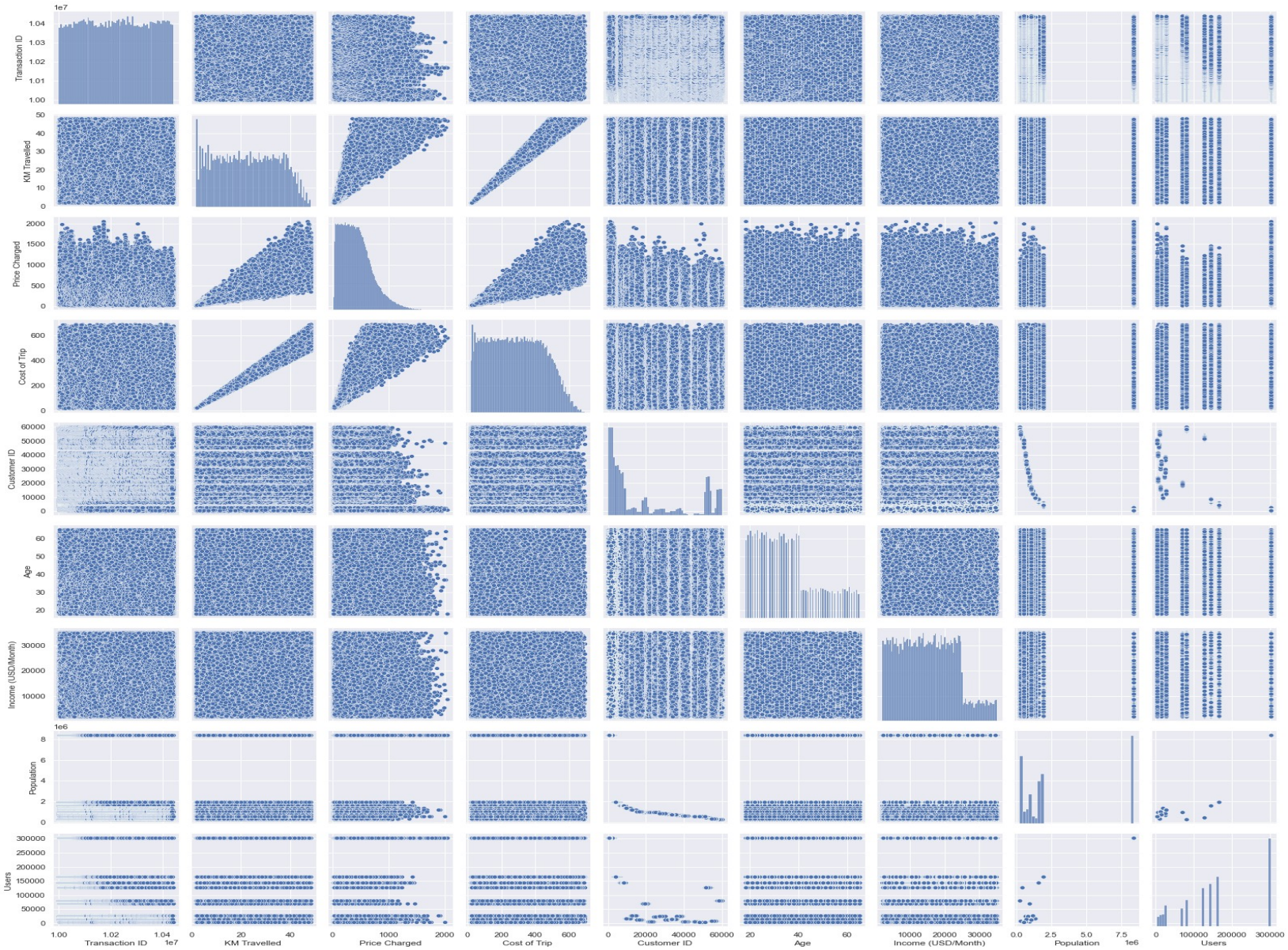
- Looks like the majority of female cab users prefer taking the Yellow Cab.

Correlation



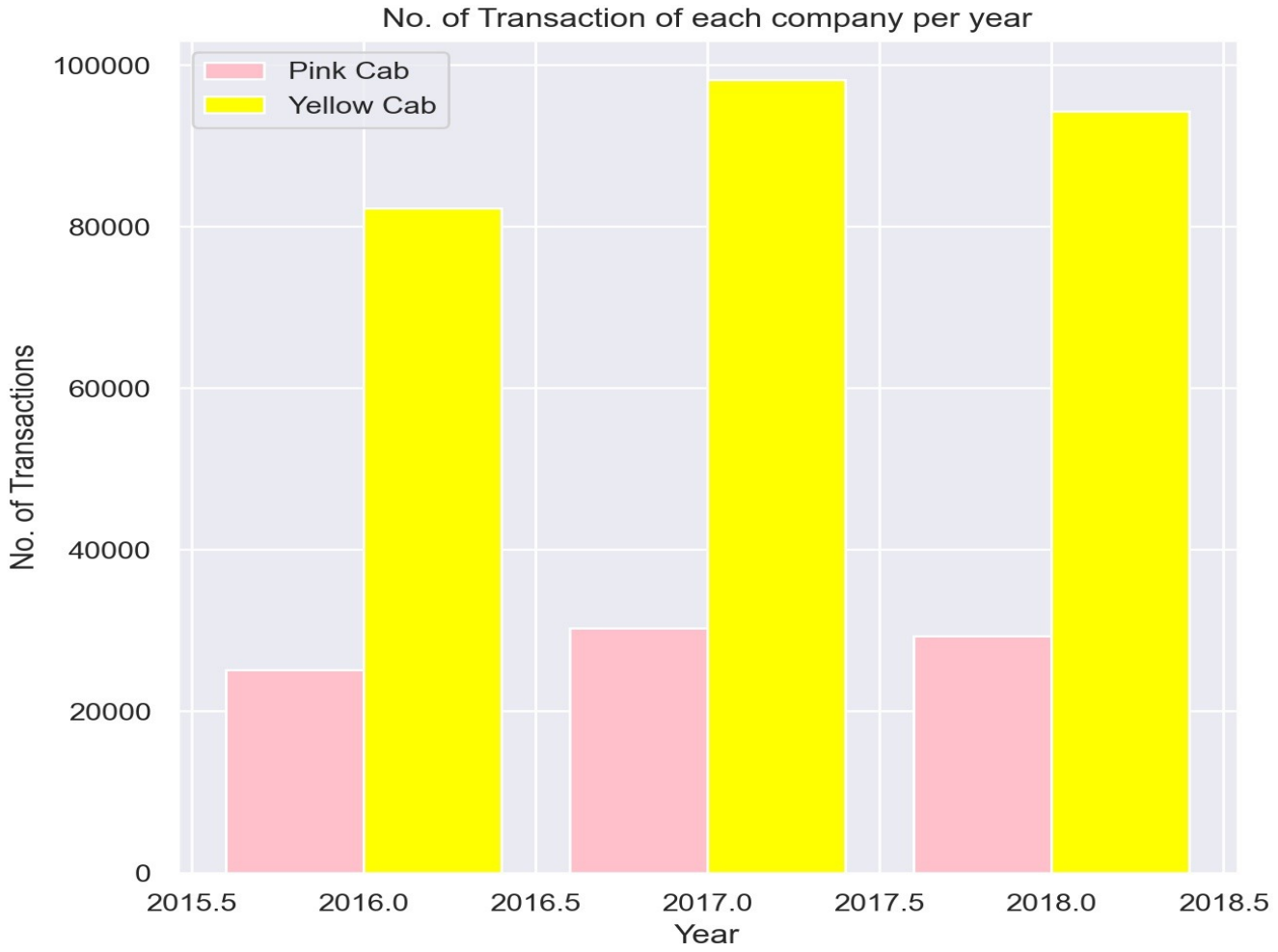
- There is a positive correlation between ‘Price Charged’ and ‘KM Traveled’ and ‘Cost of Trip’.

Correlation



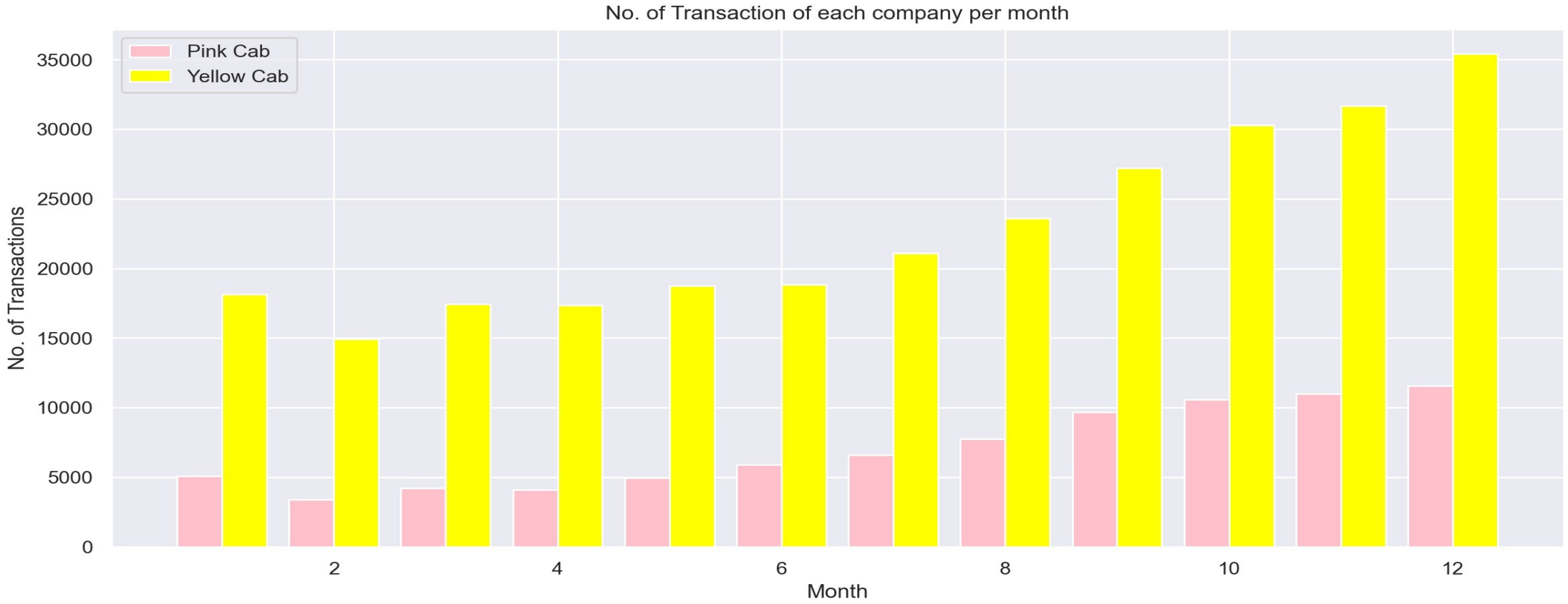
- This pairs plot allows us to see both distribution of single variables and relationships between two variables.

Transactions per year for both companies



- The Yellow Cab company looks more active than the Pink Cab company on a yearly basis.

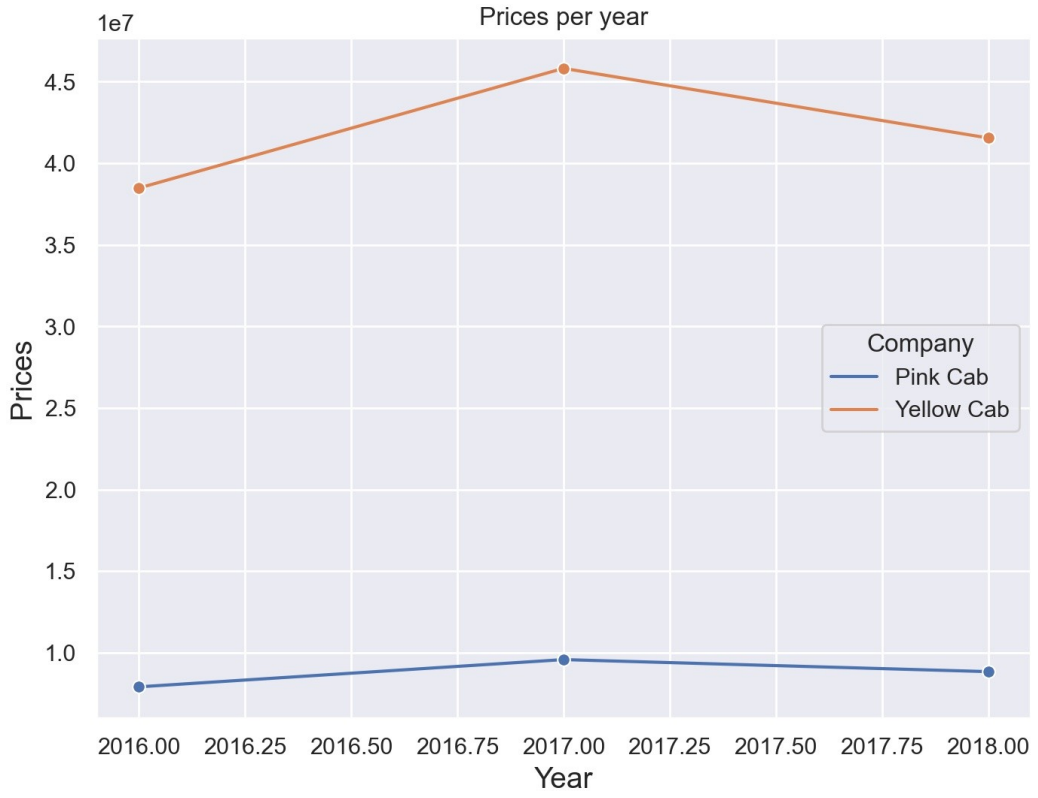
Transactions per month for both companies

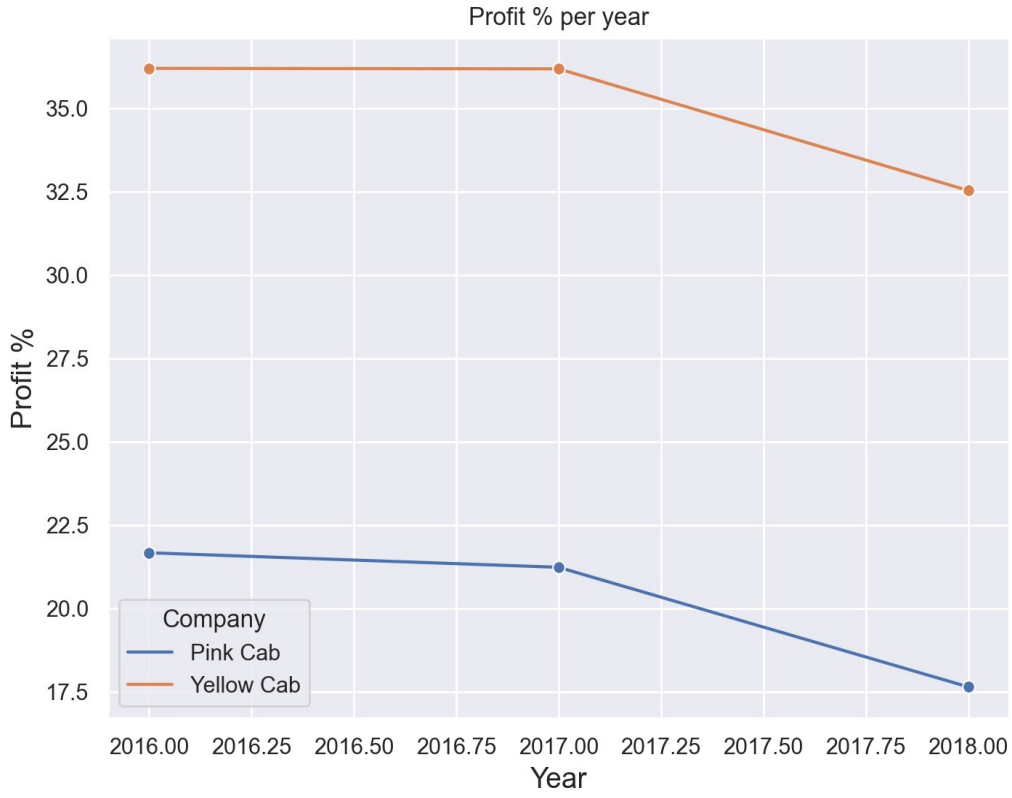


- As we can see from this bar plot, on a monthly basis, Yellow Cab company is in high demand than the pink cab company especially during Holiday season.

Which company is more profitable?

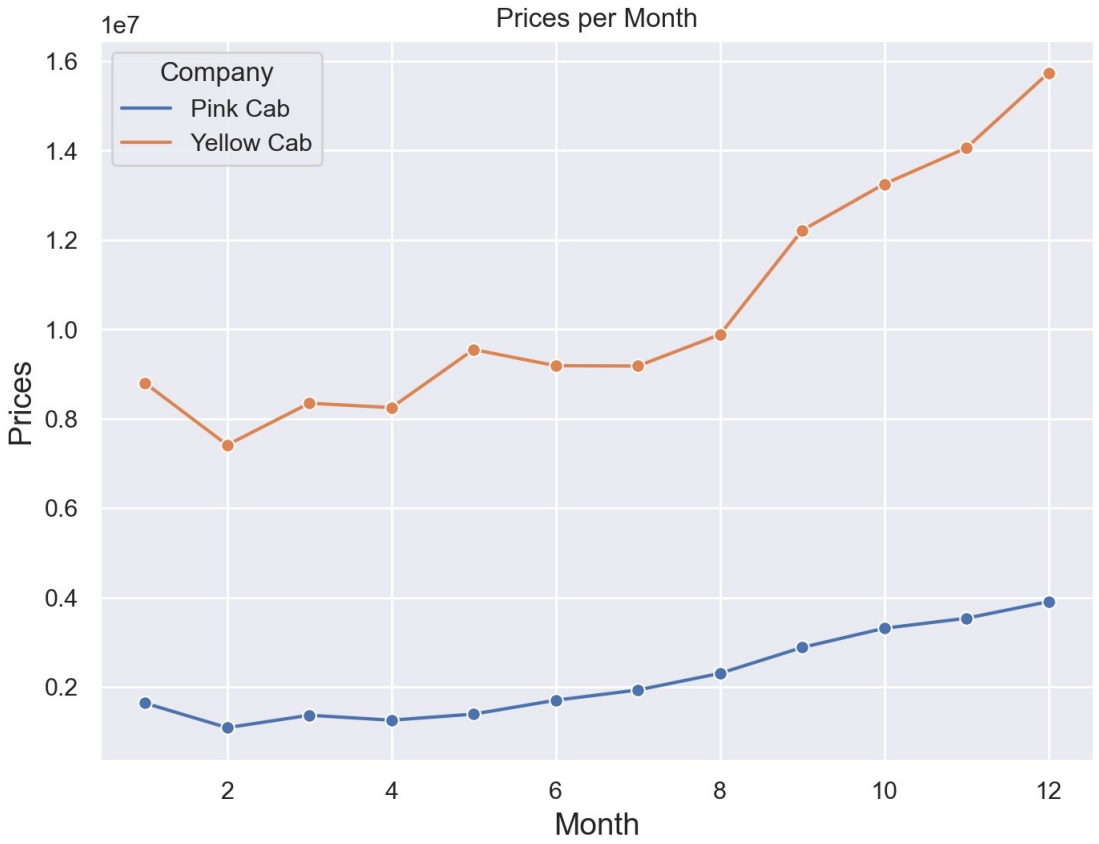
Evolution of Prices and Profit percentage per year

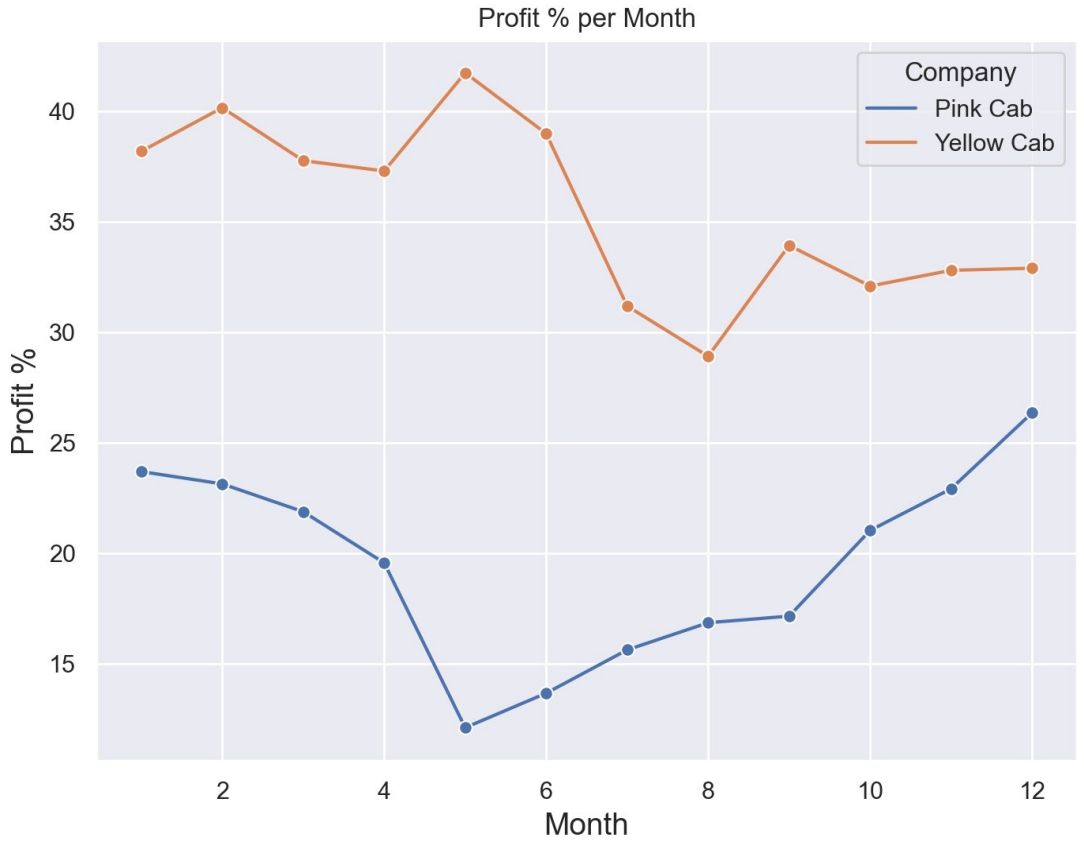




- The percentage of the Profit deviation for the Yellow Cab company is 23.07 %
- The percentage of the Profit deviation for the Pink Cab company is 61.09 %

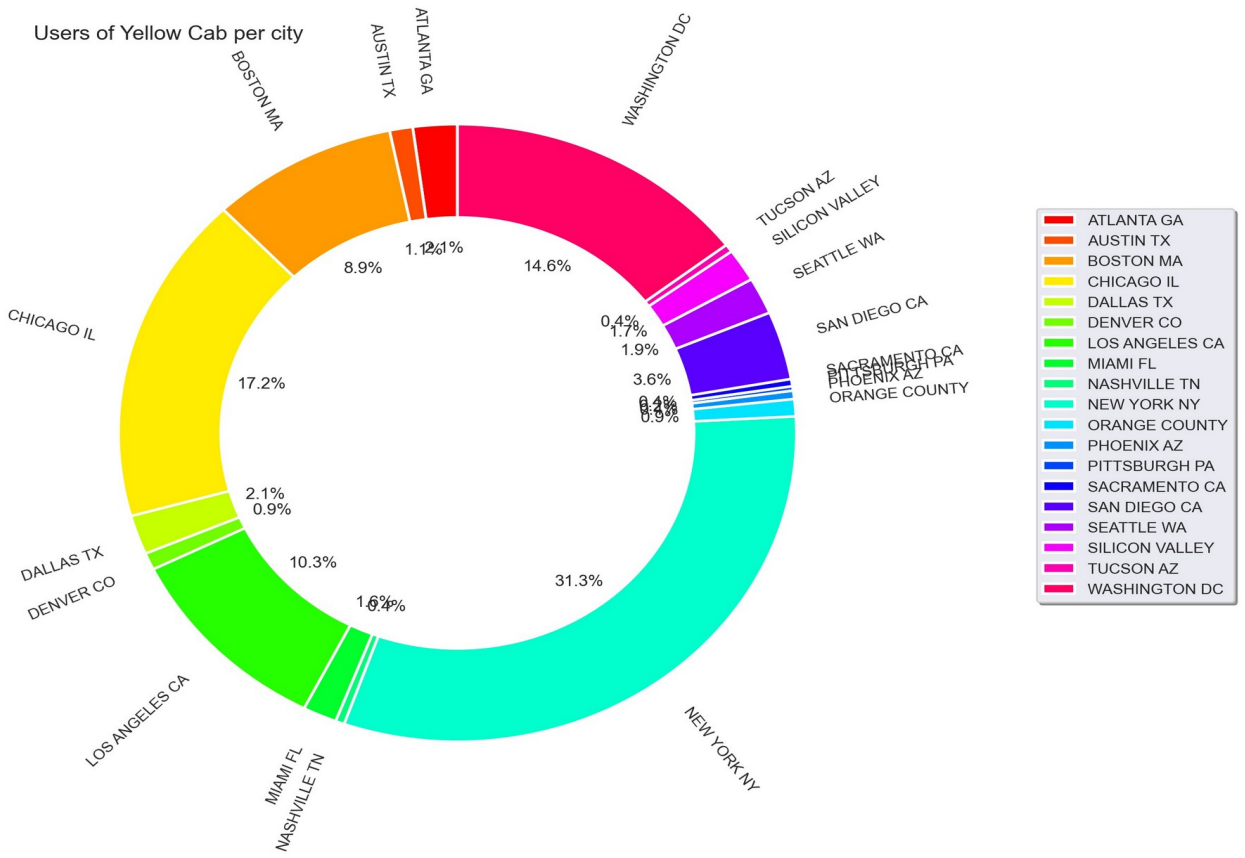
Evolution of Prices and Profit percentage per month

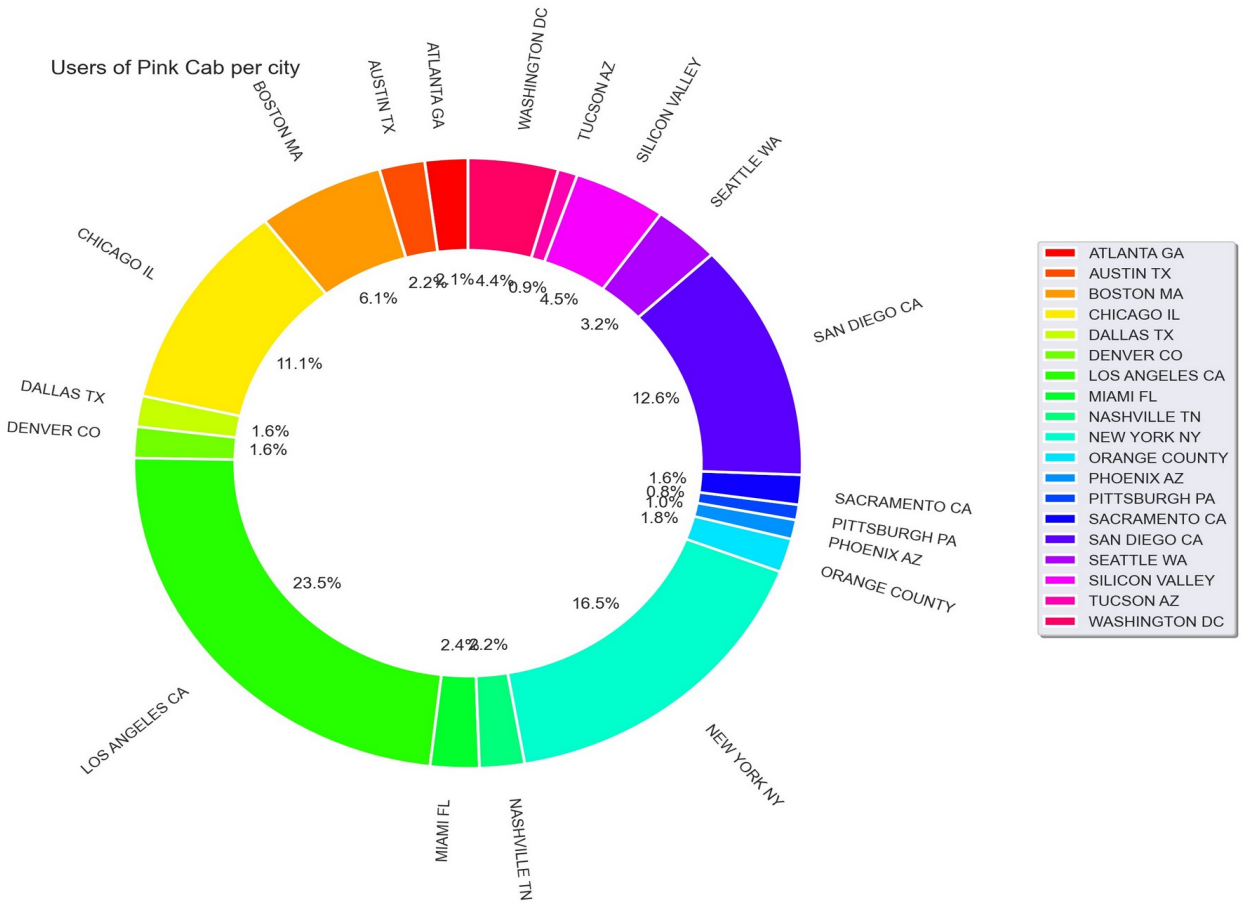




- Yellow Cab Company's earnings are more stable, with fluctuations of 23,07%, while those of the Pink Cab Company's vary in the order of 61,09%

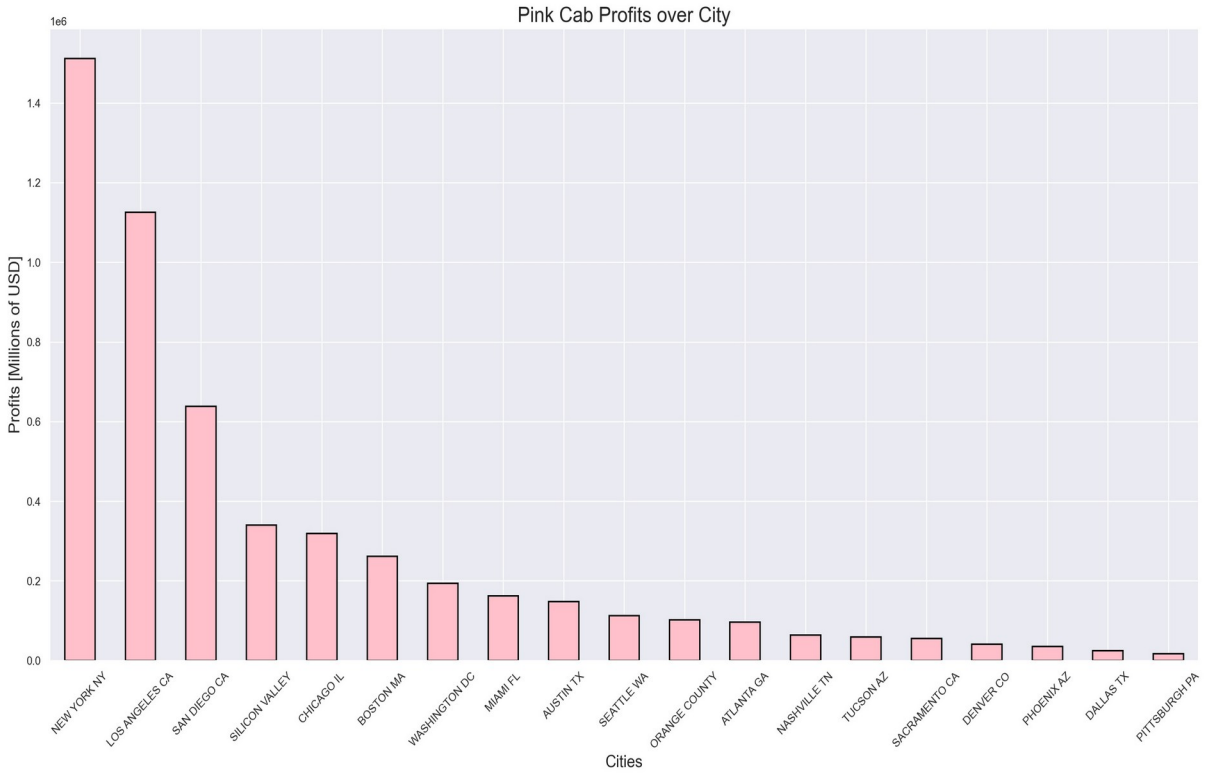
Cab Users per city

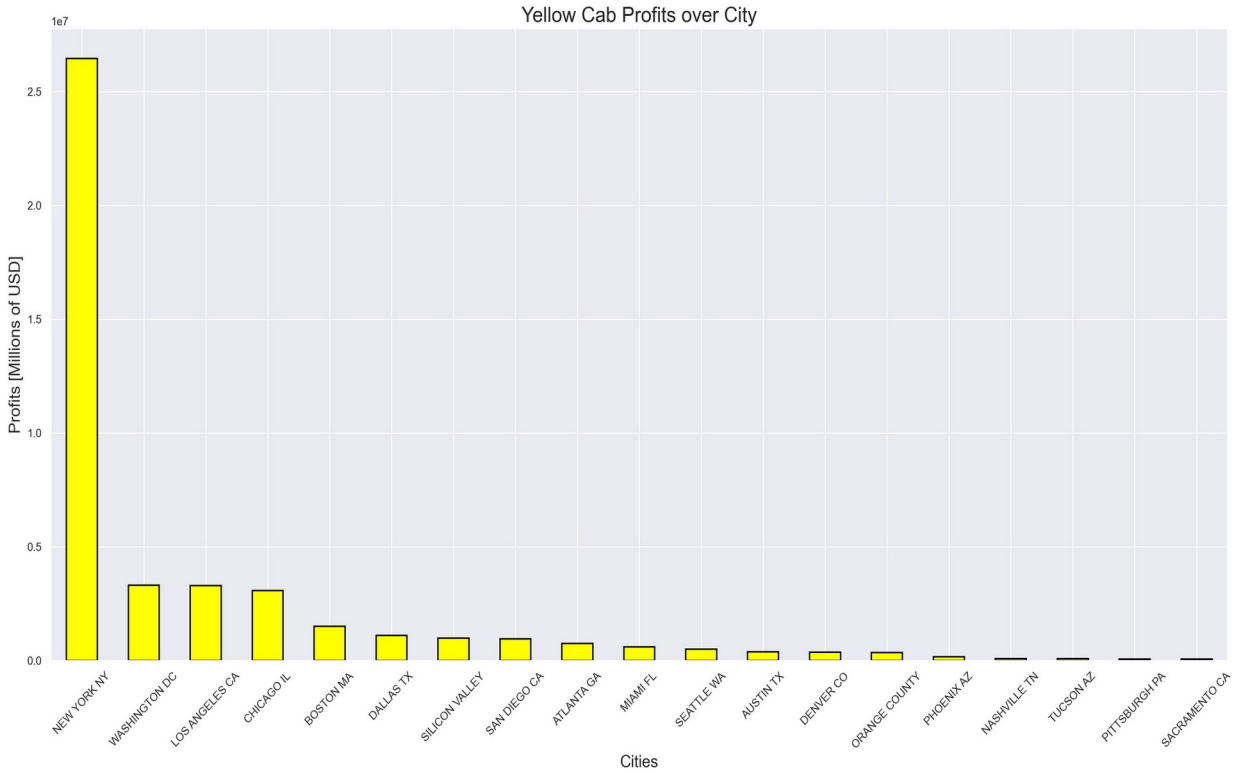




- Transactions for Yellow Cab is highest in **New York City** which has the highest Cab Users of 47% in total.
- Transaction for Pink Cab is highest in **Los Angeles CA City** with 34% of users in total.

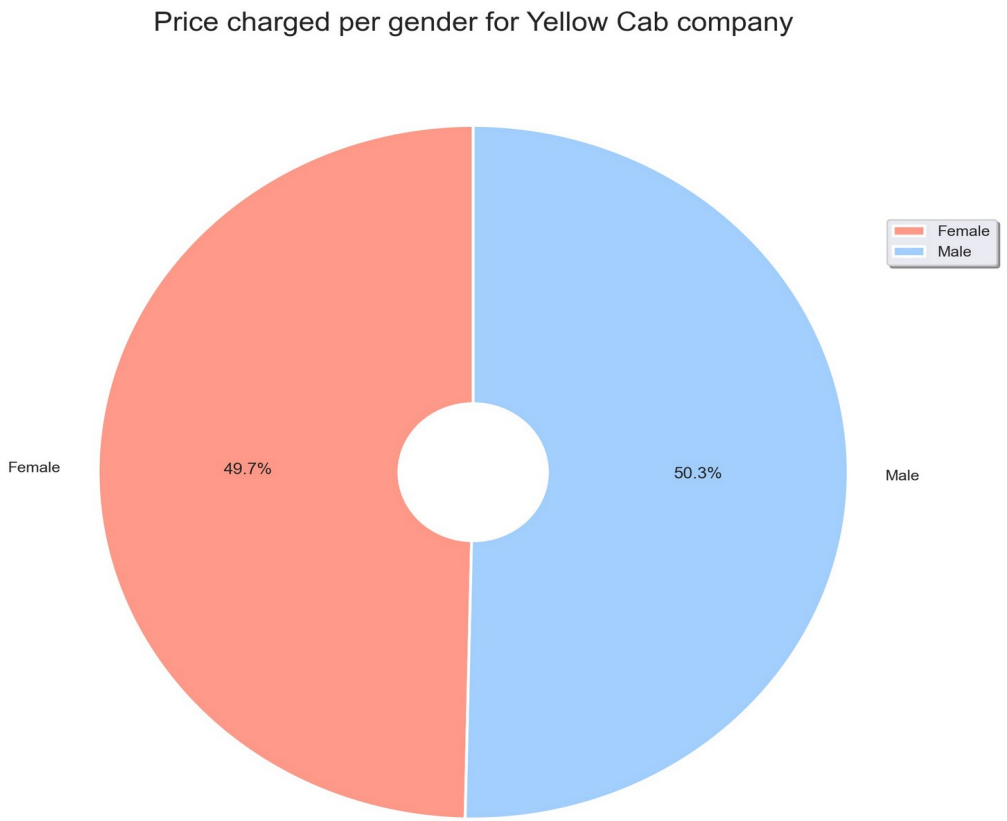
Profit over cities for both companies

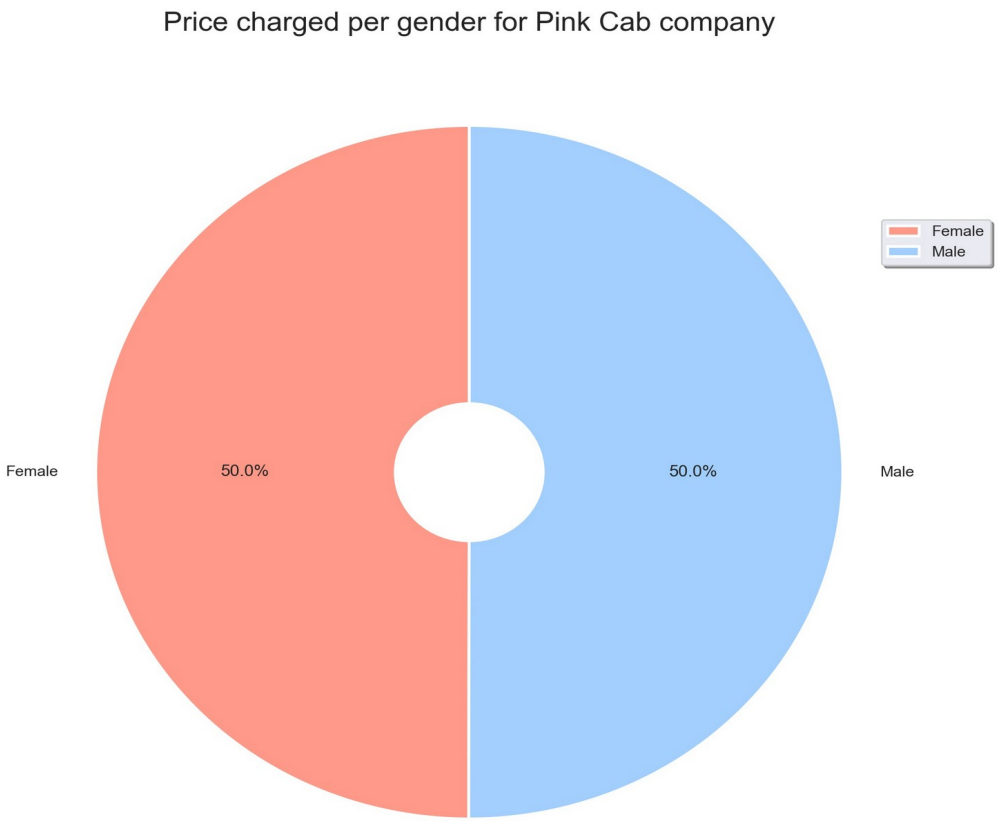




- New York city holds the highest number of transactions for both companies, hence the profit in Yellow cab company is higher than the Pink cab company in this city.

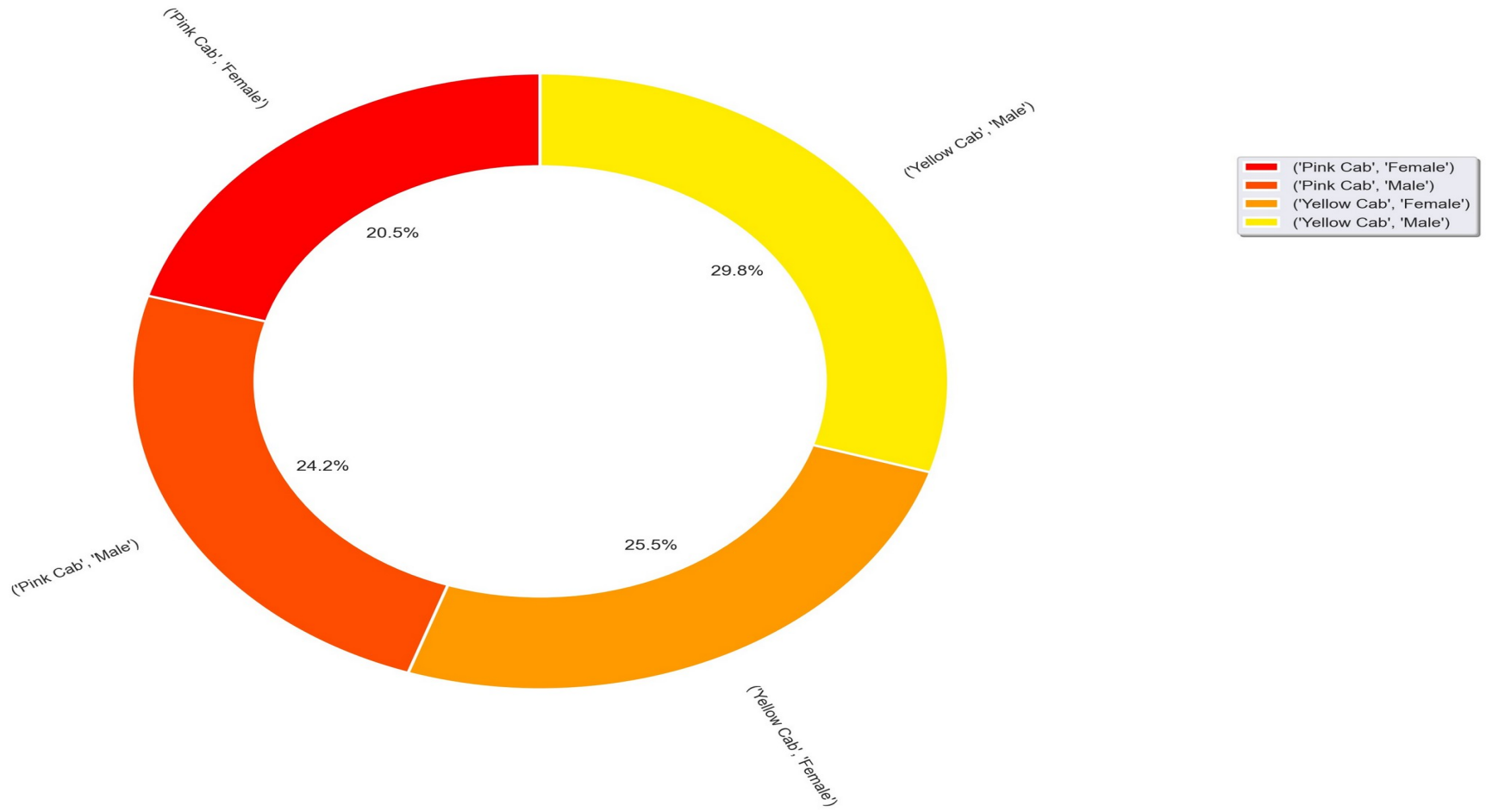
Price charged per gender in both companies





- Yellow Cab charge less from Female Customers whereas Pink Cab charges same for both Male and Female Customers.

Customer share in both companies



- Female Customers in Yellow Cab is higher (25.5%) compared to female customers in Pink cab (20.5%).

Hypothesis Testing

Hypothesis 1

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

H1 : There is 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 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 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
```

→ Looks like Yellow Cab company offers discounts for their customers who are older than 60 years old.

Recommendations

I 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**:

1. Profit Analysis:

- **Profits:** Higher Profits over the time and less fluctuations monthly for the Yellow cab company.
- **Profits City wise:** Yellow Cab company has greater market share in every City.
- **Nb of Transactions:** On a monthly basis, Yellow Cab company is in high demand than the pink cab company especially during Holiday season.

2. Client Analysis:

- **Payment Mode Distributions:** Both companies present the same distribution of Payment Mode over time, city wise and age wise.
- **Gender Age wise:** In Yellow Cab company there is difference in prices for people older than 60 yrs, whereas in Pink Cab there is no difference for all age groups.
- **Gender:** Yellow cab company also charge less for female customers.

→ **On the basis of above points , I recommend Yellow cab for investment.**

Thank You