Exploratory Data Analysis

<G2M Insight for Cab Investment Firm>

Student: Khanh Bui

<22-June_2022>

Agenda

- Executive Summary
- Problem Statement
- Dataset
- EDA
- EDA Summary
- Recommendations

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 that required to understand the market before taking final decision.

Data Glacier
Your Deep Learning Partner

Problem Statement

Goal: Decide which cab company is better by providing statistical information, which helps in choosing which company to invest in.

Route of Analyzing:

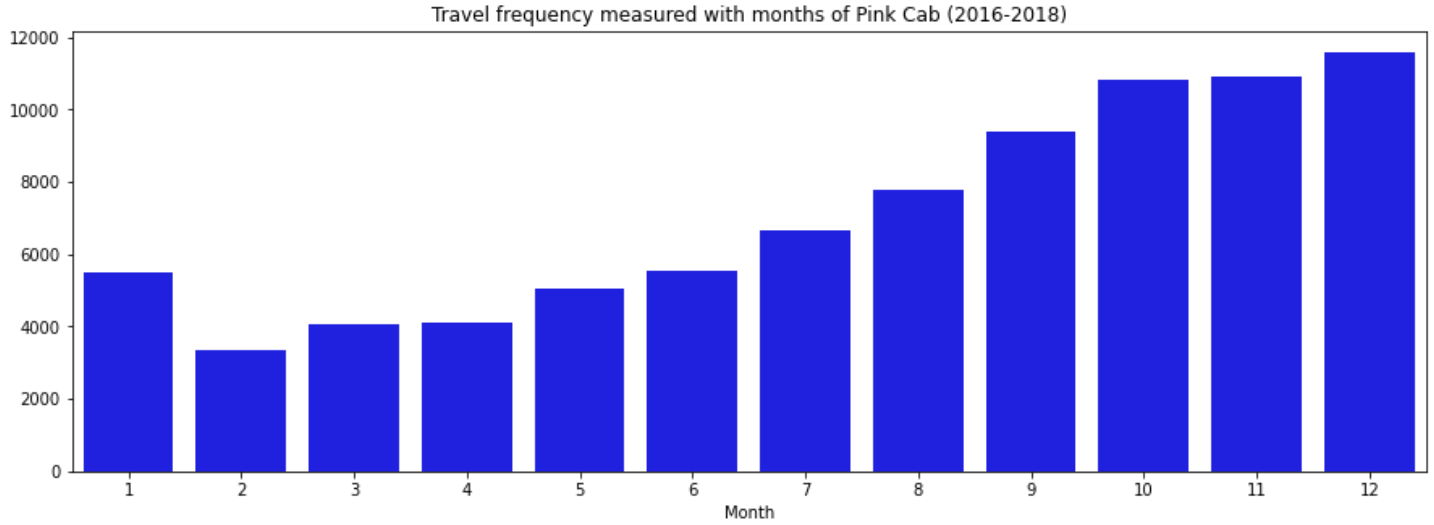
- Understanding the data
- Find the most preferred Cab depends on the customer's usage
- Find the most profitable Cab Company
- Find the rank of the age group that uses the Cab
- Find the cities that operate best
- Find the spread of income group uses Cab
- State and Prove Hypothesis

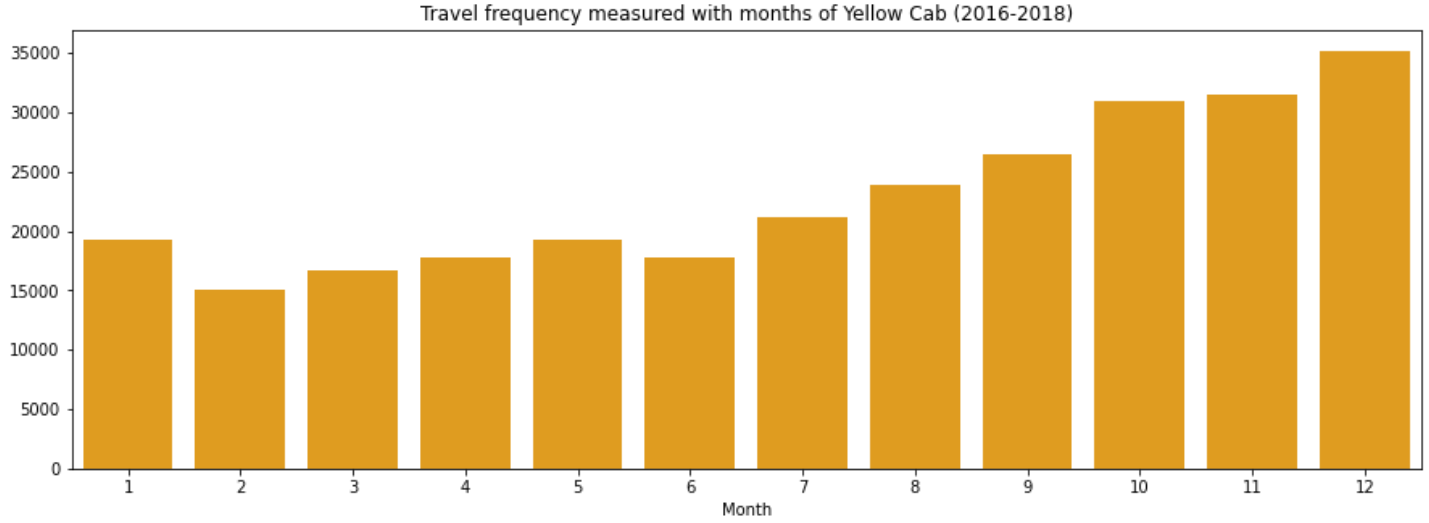
Dataset

Main data are taken from 4 CSV files:

- Cab_Data.csv: Contain Transaction ID, Date of Travel, Company, City, KM Travelled, Price Charged, Cost of Trip
- Customer_ID.csv: Contain Customer ID, Gender, Age, Income (USD/Month)
- Transaction_ID.csv: Provide Transaction ID, Customer ID, Payment_Mode
- City.csv: Provide City, Population, Users

EDA
Travel frequency by month from two cab companies

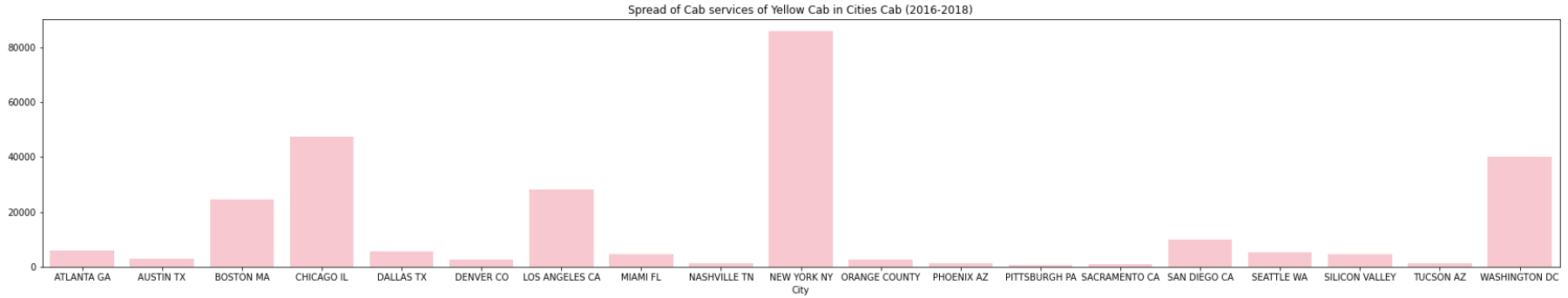


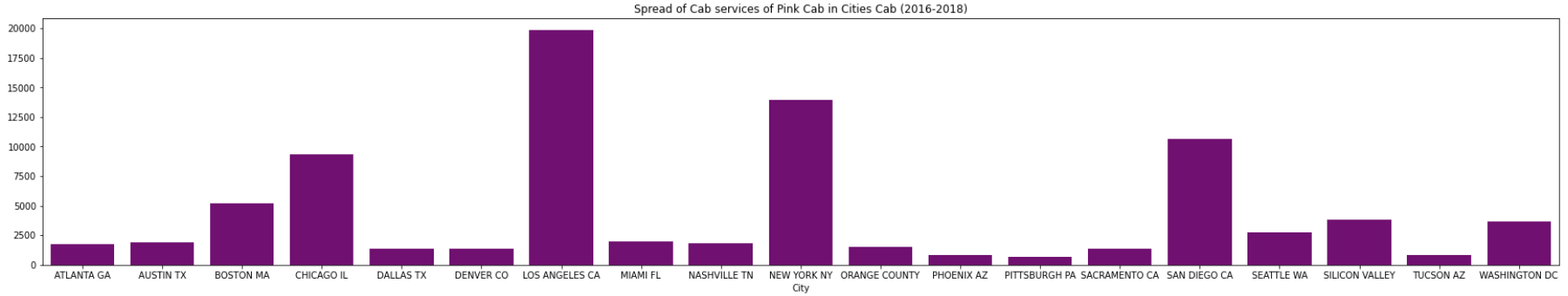


Following the graph beside:

- Yellow Cab has more users compared to Pink Cab in the year 2016-2018
- The highest amount of cab usage is in December for both Company
- The usage also seasonally depended, since most of the high data groups are around August to December and the lowest are around February to June.

The City Spread of 2 Cab Companies in Cities (real number)

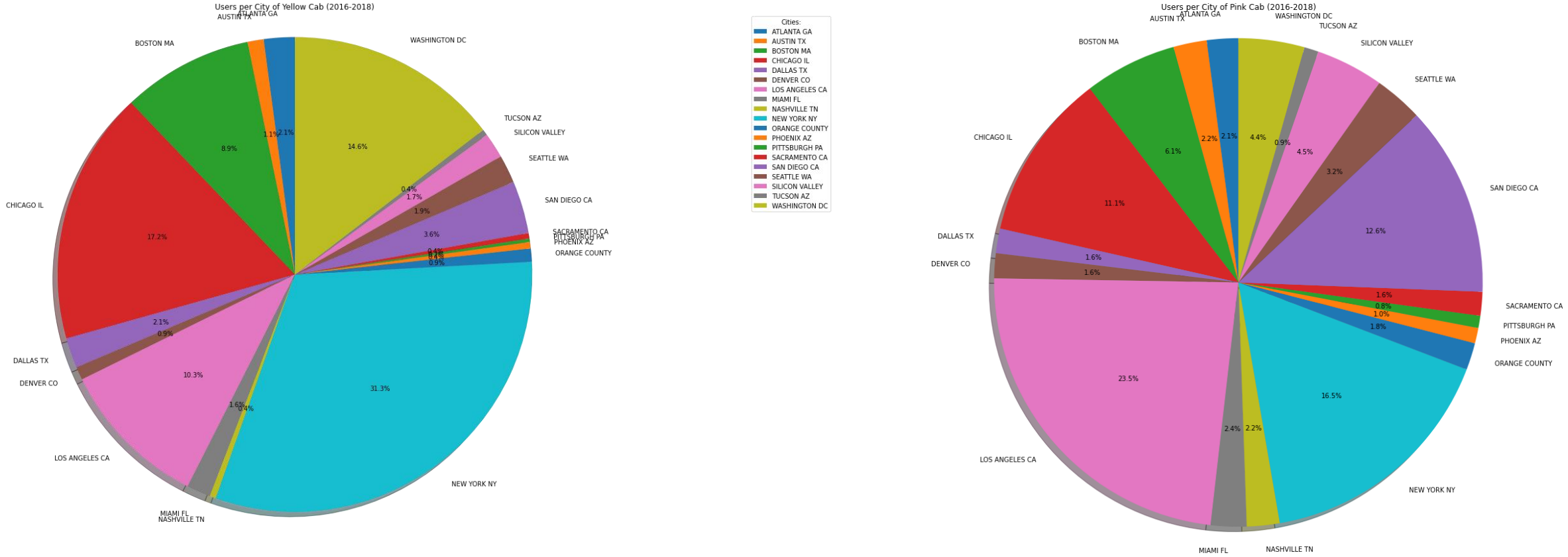




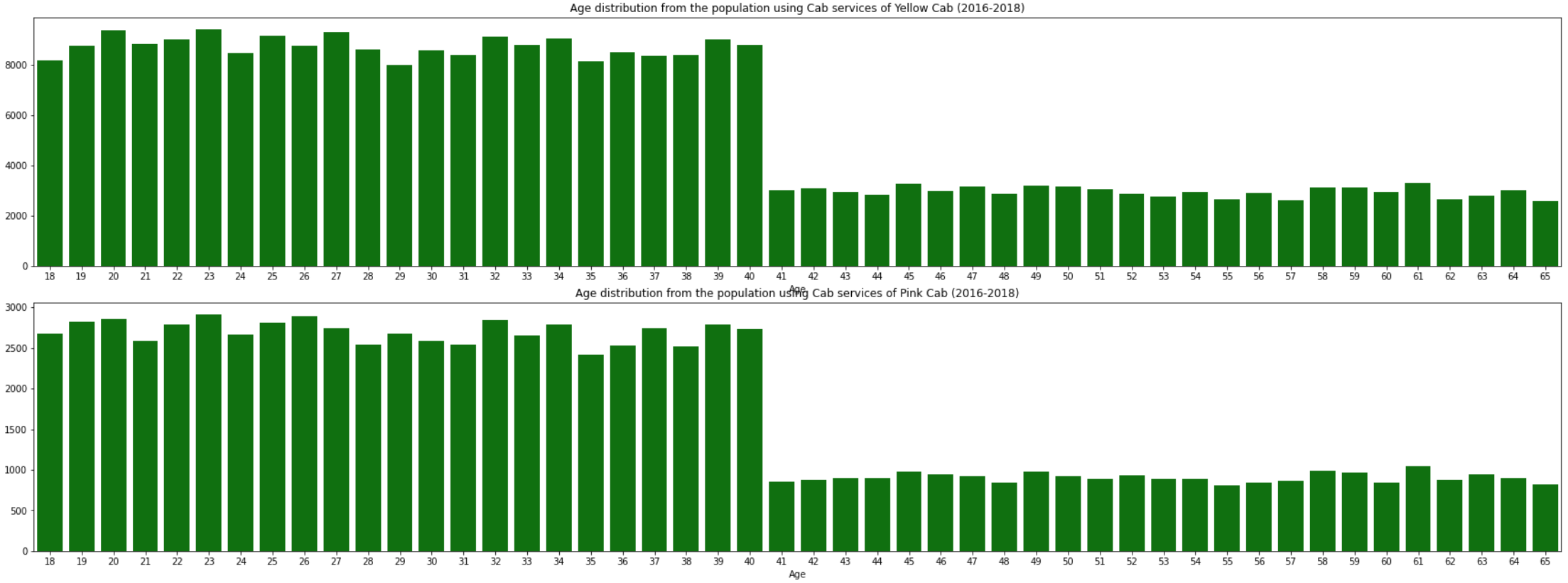
The graph illustrates the spread of Cab services in cities from 2 Cab Company

- For Yellow Cab, the most populated Cab is in New York City, Washington DC, Chicago, and Los Angeles.
- Pink Cab, is different compared to Yellow Cab, the most populated one is in Los Angeles, New York, San Diego, and Chicago

The City Spread of 2 Cab Companies in Cities (percentage)



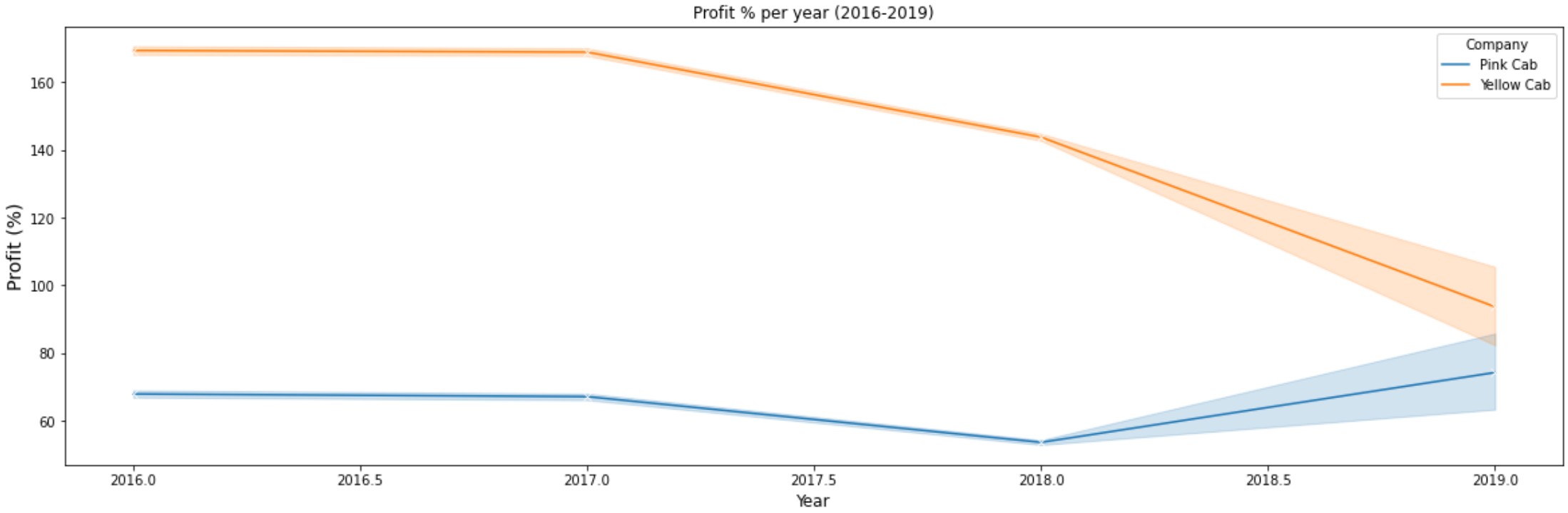
Age distribution from the population using Cab services



Which age has the most Cab Usage in 2 Cab Company?

- From two Cab Services, the age distribution is most likely the same, the age dist. with high data n,umber varies around 18-40 years old.
- It can be concluded that the most common user's age is in adult and middle-age age group

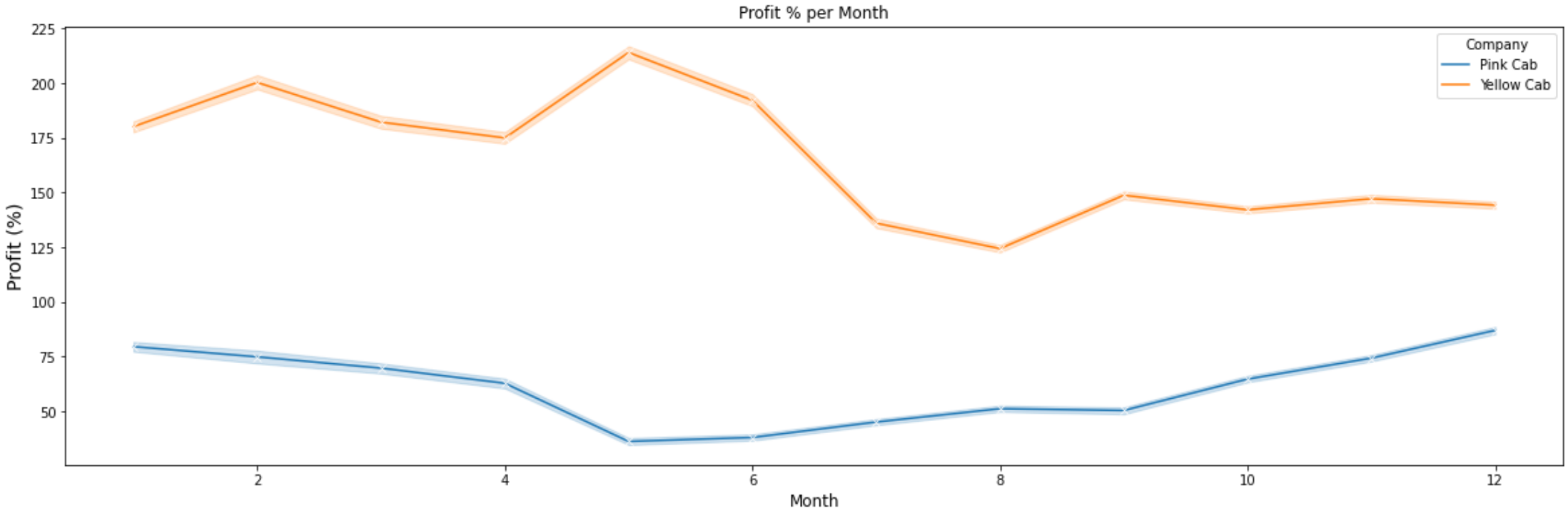
Profit (%) per year of 2 Cab Companies



What are the profit trends?

- The Yellow Cab's profit is at a higher peak than the Pink Cab
- However, the increase of Yellow Cab start to fall off since 2017
- The Pink Cab increase its profit % per year since 2018
- At this rate, the profit % of the 2 Cab is predicted to collide in half of 2019

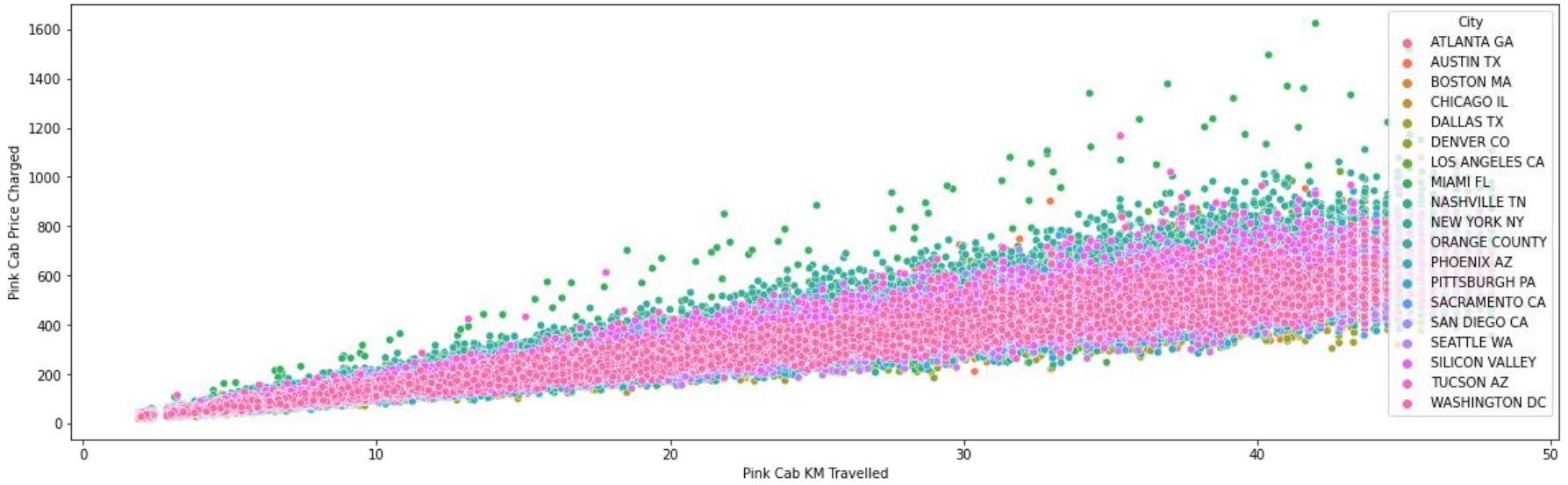
Profit (%) per month of 2 Cab Companies

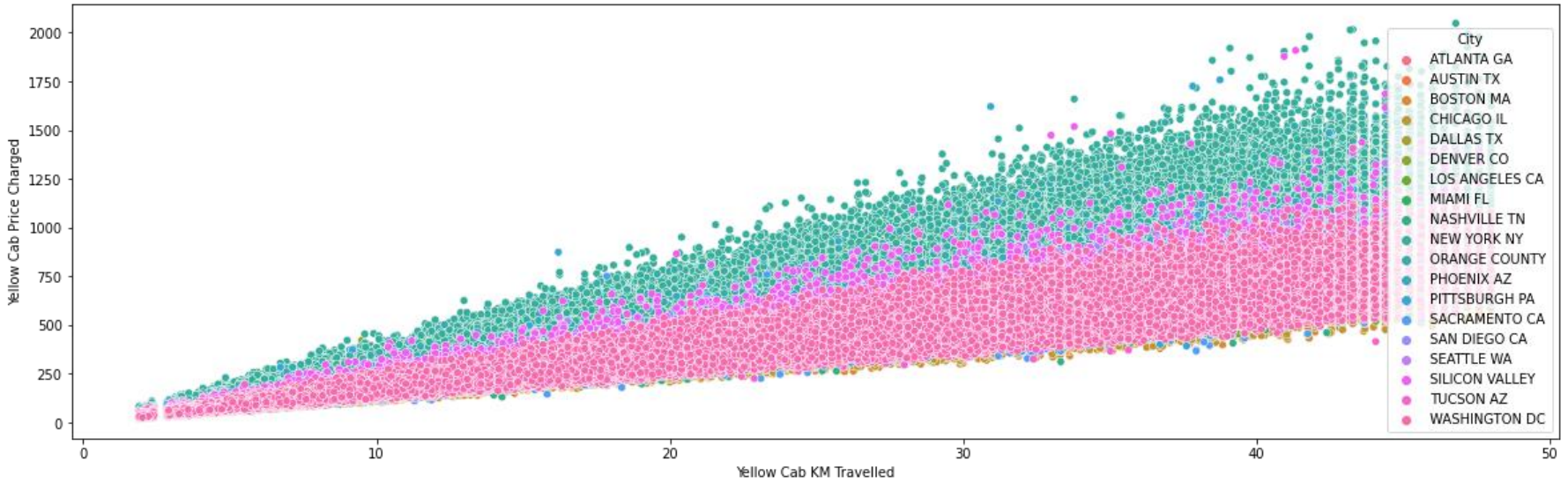


With this graph:

- We can conclude that in Yellow Cab, Although it has a significant increase in May, its fall starts to begin drastically in June till December.
- From Pink Cab, with a steady Fall from Jan to May, the trend starts to increase steadily till December.

Price charged to the customer per km traveled

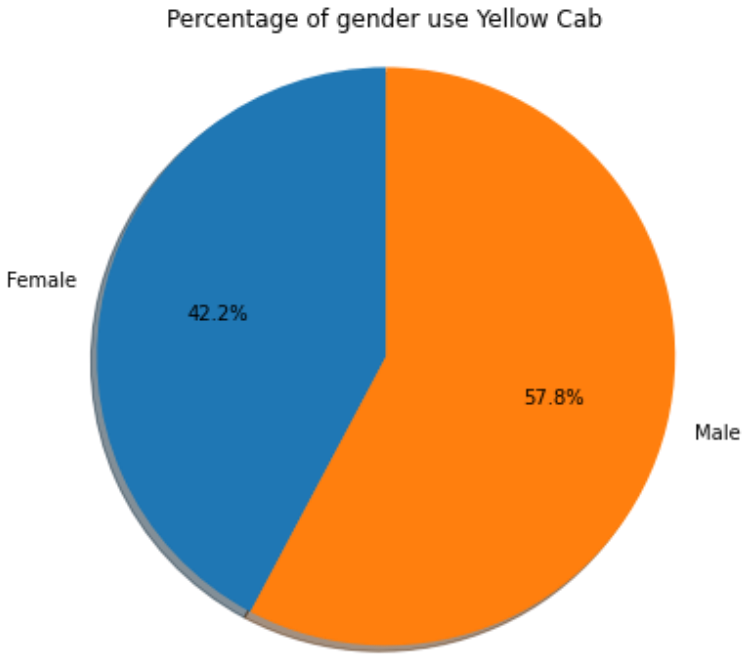


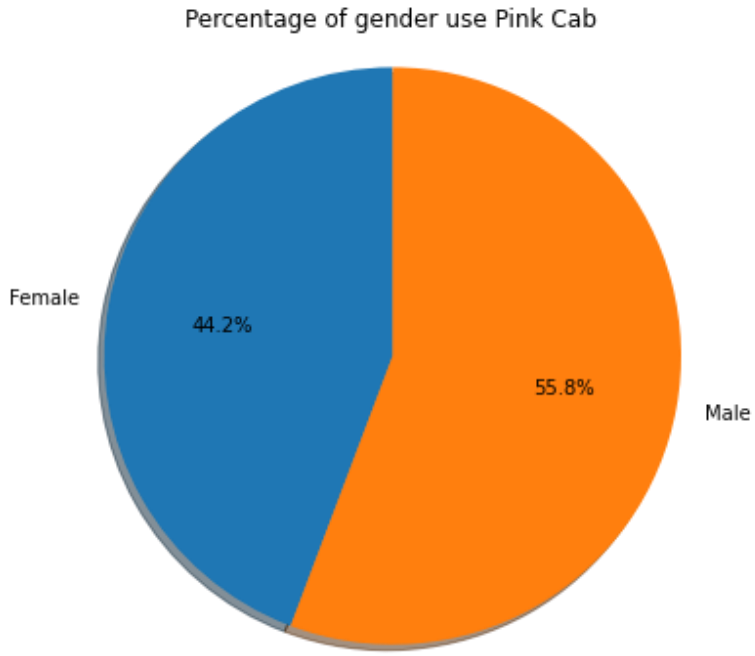


From two graphs:

- The rate of charging the price is linear, as more Km traveled, more prices going to be charged.
- However, the price is different depending on the city that it's operating, Since the green part is the most expensive, it is the most populated city with the most Cab services from both Companies.

Gender percentage that uses the Cab services from 2 given companies





As we observed:

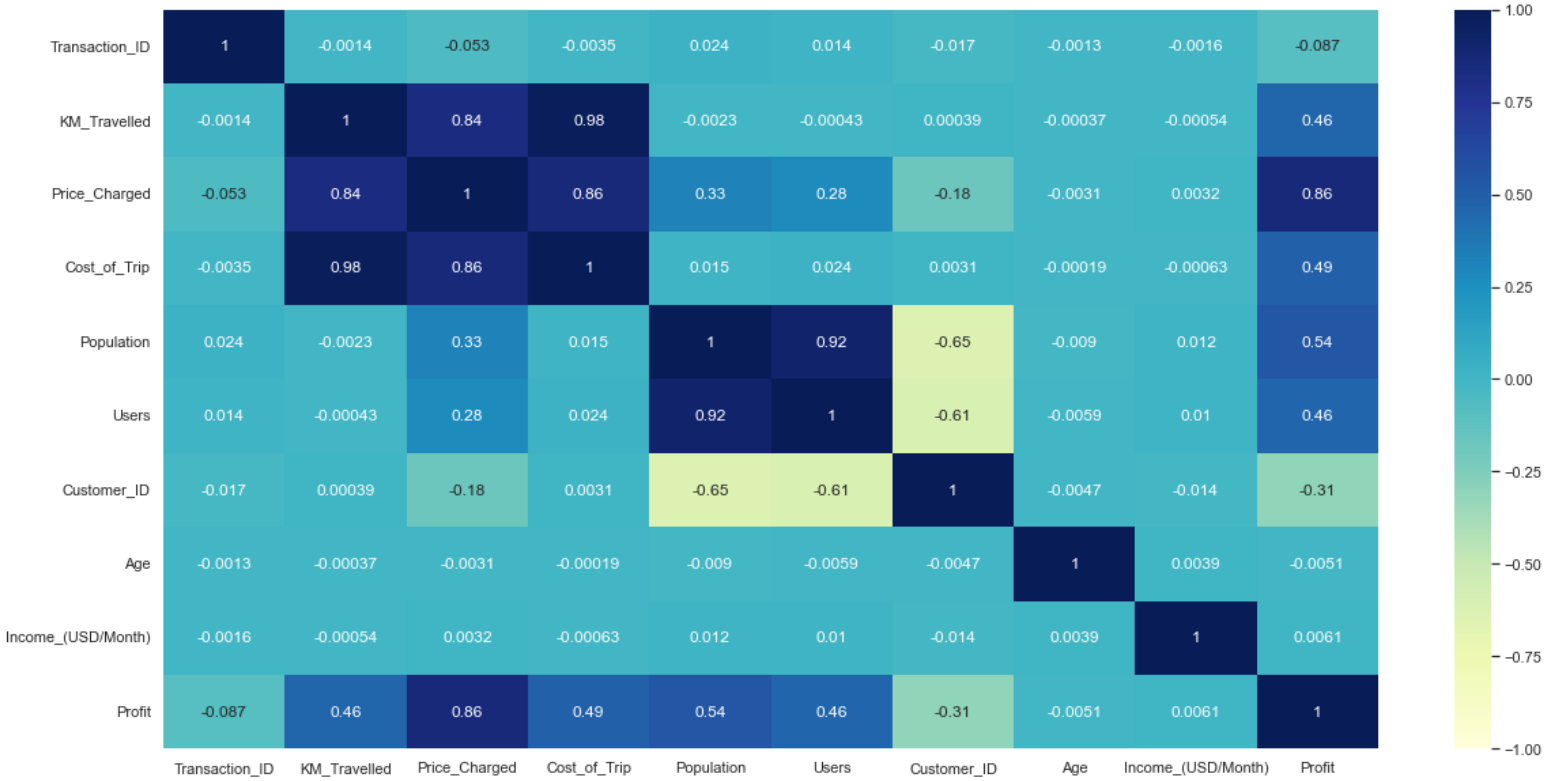
- From two Cab Companies, Male has the to traveled more frequently compared to Female
- For Yellow Cab company, the male user's percentage is bigger than their opponent which is Pink Cab company

Average user income of 2 cab companies

It is illustrated that:

- The income of Cab users from the two Companies is the same.
- The average user's income is about 15000.
- The highest can be 35000.

Correlation of all variables inside the data frame



As we see from the heat map:

- All the data have a strong correlation with each other including:

+ Cost of trips vs lm traveled
+ Price charged vs Cost of trip
+ Population vs the amount of cab user
+ Price charged vs Profit

Hypothesis of the data

Hypothesis 1: Is there any changes to age if profit changes?

H0: There are no changes to consumer's average age if profit changes
H1: There are changes to consumer's average age if profit changes

+ Pink Cab Company

```
We accept alternate hypothesis that theres no difference for Pink Cab
P value is  0.04113385102242378
```

+ Yellow Cab Company

```
We accept alternate hypothesis that theres no difference for Yellow Cab
P value is  0.0036516330682683363
```

- Conclusion:

There is no differences to age if profit change for both Cab Companies

Hypothesis of the data

Hypothesis 2: Is there any changes to age average if Population changes?

H0: There are no changes to consumer's average age if population changes
H1: There are changes to consumer's average age if population changes

+ Pink Cab Company

```
We accept alternate hypothesis that theres no difference for Yellow Cab
P value is  1.1217049899132143e-08
```

+ Yellow Cab Company

```
We accept alternate hypothesis that theres no difference for Pink Cab
P value is  1.4607087470014844e-05
```

- Conclusion:

There are no differences changes in age average if Population changes for both Cab companies

Hypothesis of the data

Hypothesis 3: Is there any changes to age if price charged changes?

H0: There are no changes to consumer's average age if price charged changes
H1: There are changes to consumer's average age if price charged changes

+ Pink Cab Company

```
We accept null hypothesis that theres is differences for Yellow Cab
P value is  0.3413547683791115
```

+ Yellow Cab Company

```
We accept null hypothesis that theres is differences for Pink Cab
P value is  0.23110176404877747
```

- Conclusion:

There will be differences in consumer's average age if the price charged changes
+Consumer's average age is increased if the price increases and vice versa.
+Solution: Charged price lower a bit(5% of the total price charged to consumers) to attract more younger consumers

This is the end of the presentation

Thank You