Exploratory Data Analysis

G2M Insight For Cab Investment Firm

July 20, 2022

By Jeffery Su

Agenda

Executive Summary

Problem Statement

Approach

EDA

EDA Summary

Recommendations

Summary

The Client

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.

OR

The Datasets

There is 4 individual data sets.

Time period of data is from 31/01/2016 to 31/12/2018.

Cab_Data.csv

- Transaction ID : ID of each transaction
- Date of Travel : date of the transaction
- Company : company (Pink cab/Yellow cab)
- City : city the transaction took place
- KM Travelled : distance travelled in KM
- Price Charged : price of the transaction
- Cost of Trip : maintenance cost/fees of the transaction

City.csv

- City : city names (all in USA)
- Population : population of the city
- Users : amount of users in the city

Customer_ID.csv

- Costumer ID : ID of each costumer
- Gender : gender of the costumer
- Age : age of the costumer
- Income : income of the costumer in USD per month

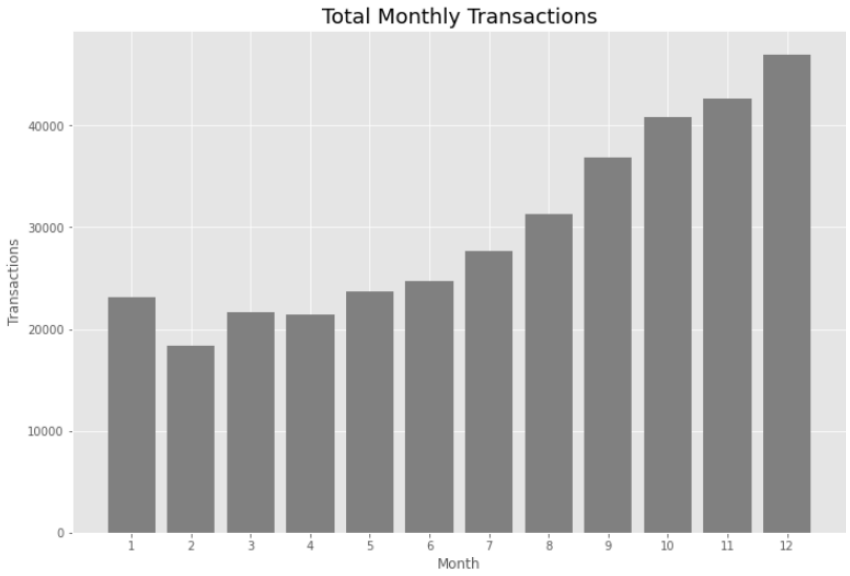
Transaction_ID.csv

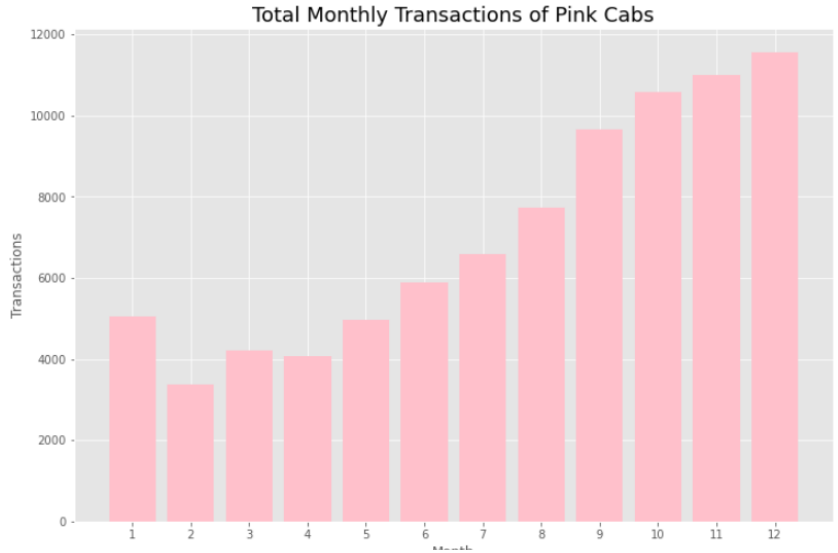
- Transaction_ID : ID of each transaction
- Costumer ID : ID of each costumer
- Payment Mode : payment method (Cash/Card)

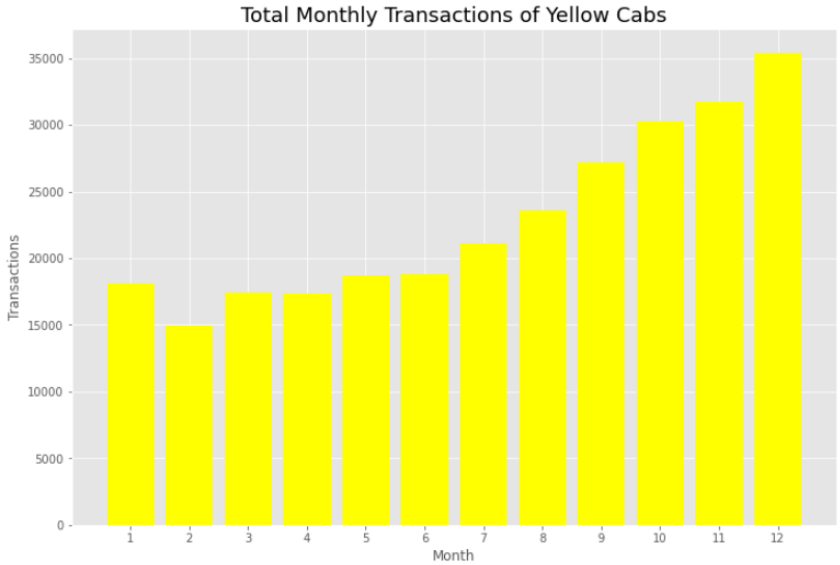
Added a Profit column in Cab_Data.csv to store the profit made for that transaction.

Exploratory Data Analysis

Transactions Per Month



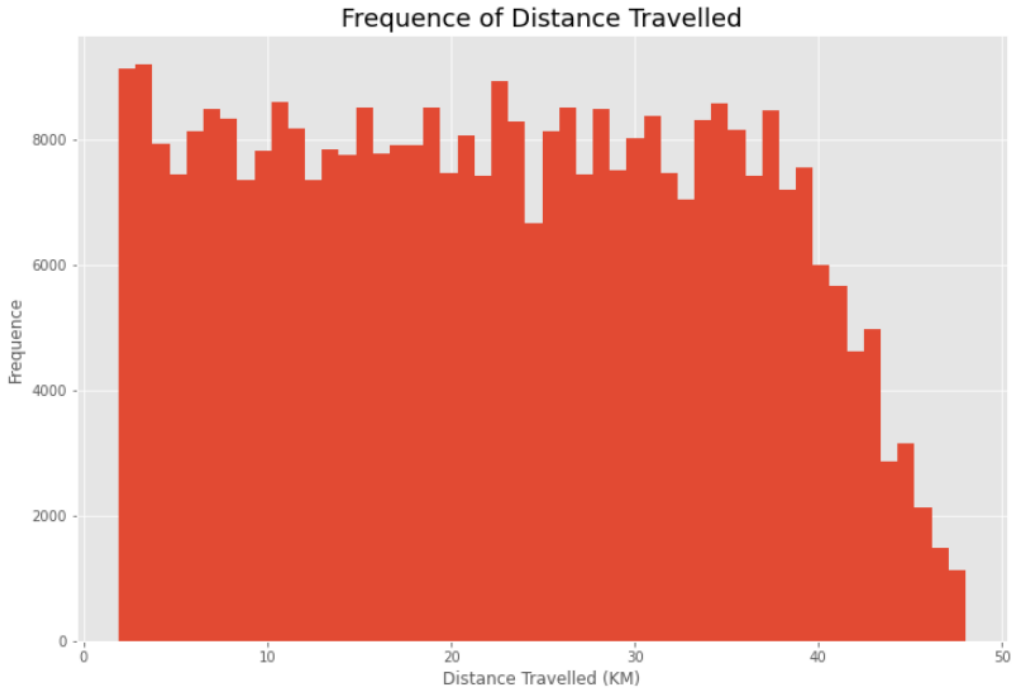


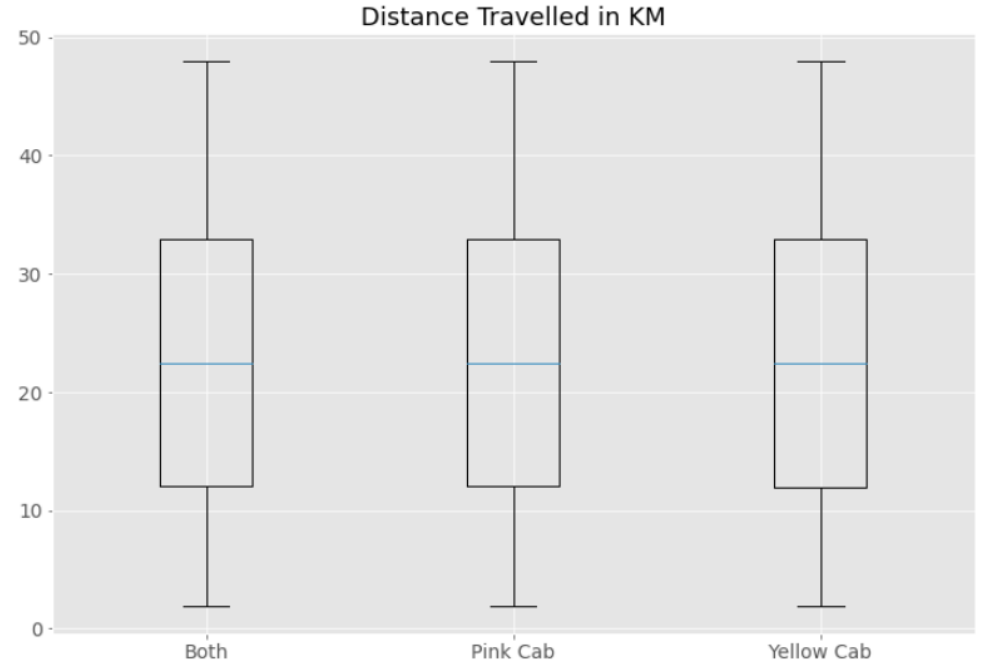


- February has the least transactions while December has the highest transactions.
- There is a trend of increasing transactions from February and peaks in December
- January has a lot less transactions.
- Yellow Cab has noticeably many more transaction in all the months

Data Glacier
Your Deep Learning Partner

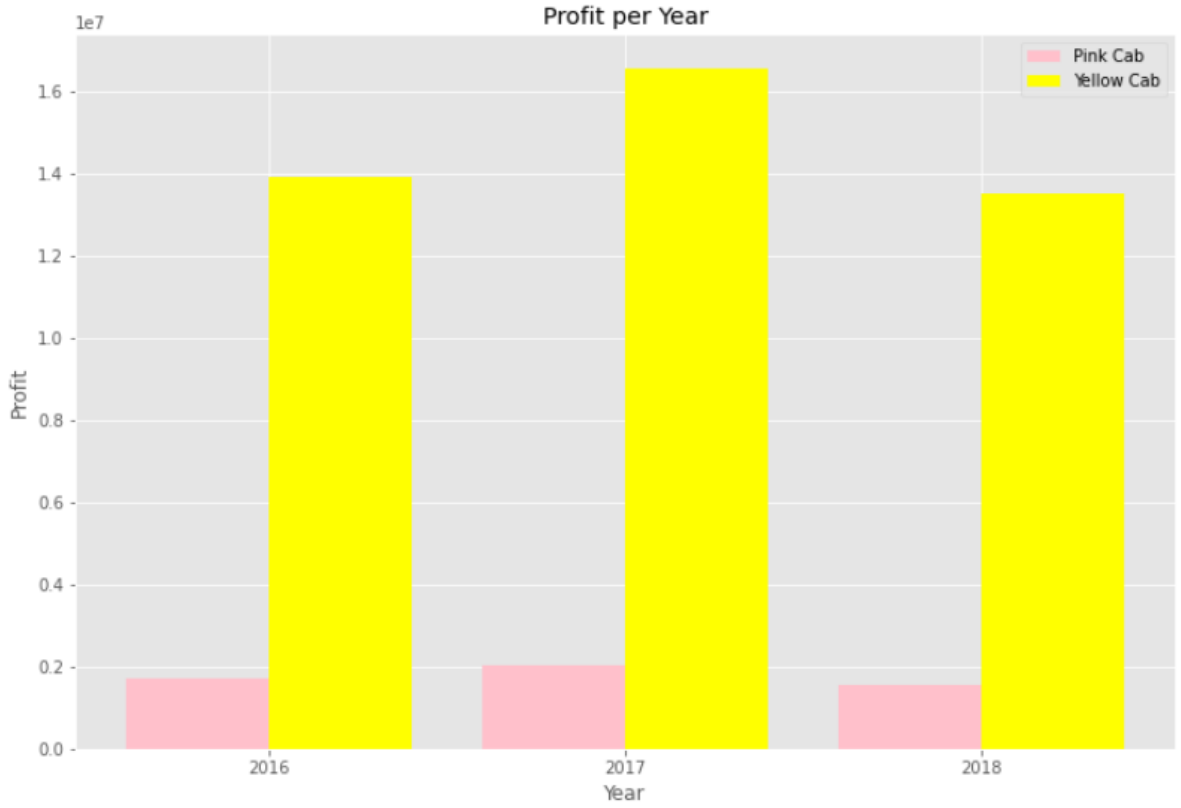
Frequency of Distance Travelled





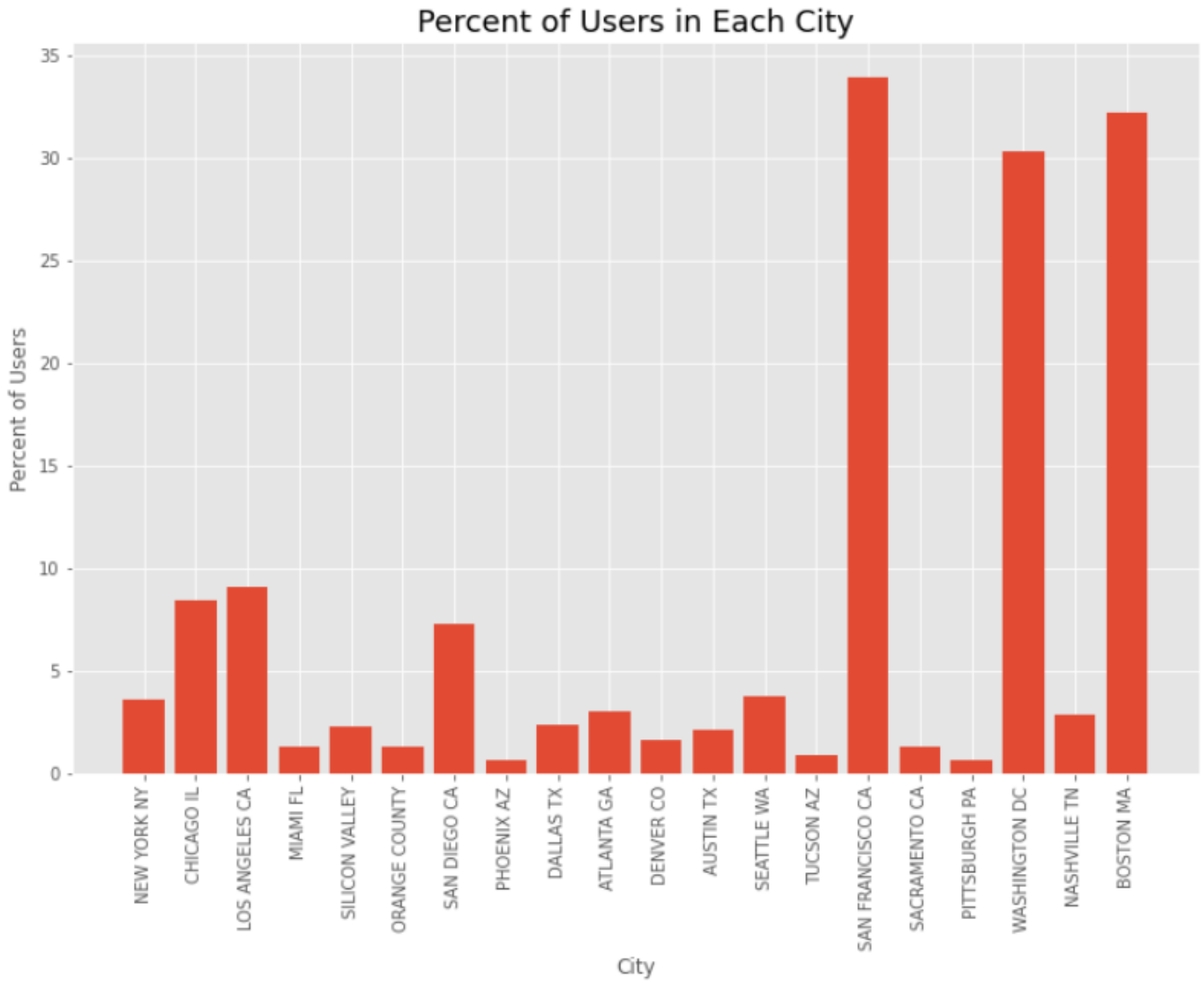
- Travel distance is range from 2 KM to 48 KM
- Majority of the distance are from 12 KM to 33 KM

Profit Per Year



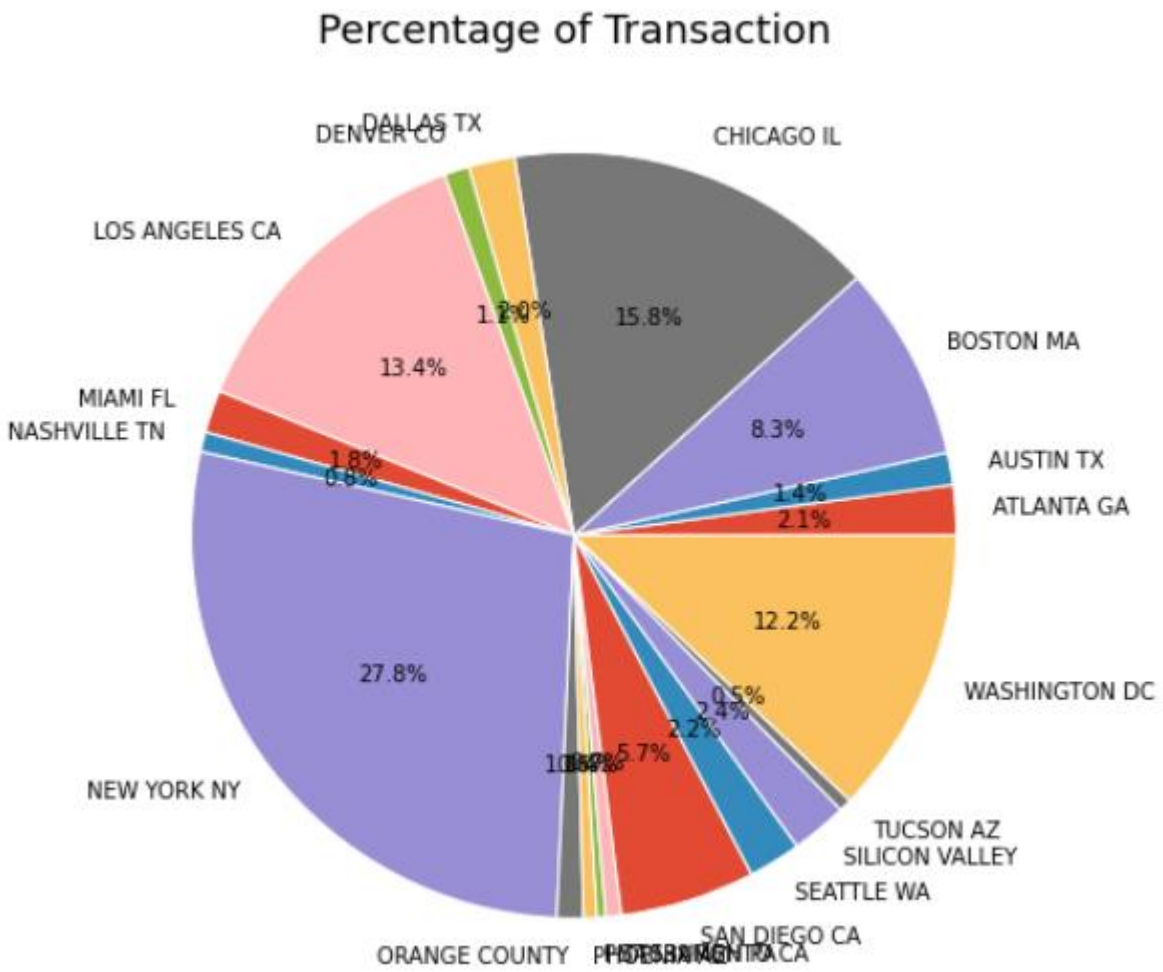
The graph above, shows the profit of both Cab company and the Yellow Cab is more profitable. The profit is the highest in 2017 followed by 2016 and 2018.

Percentage of Population in City



The percentage of population that uses the cab services are highest in San Francisco, Washington and Boston is the highest.

Percentage of Total Transaction

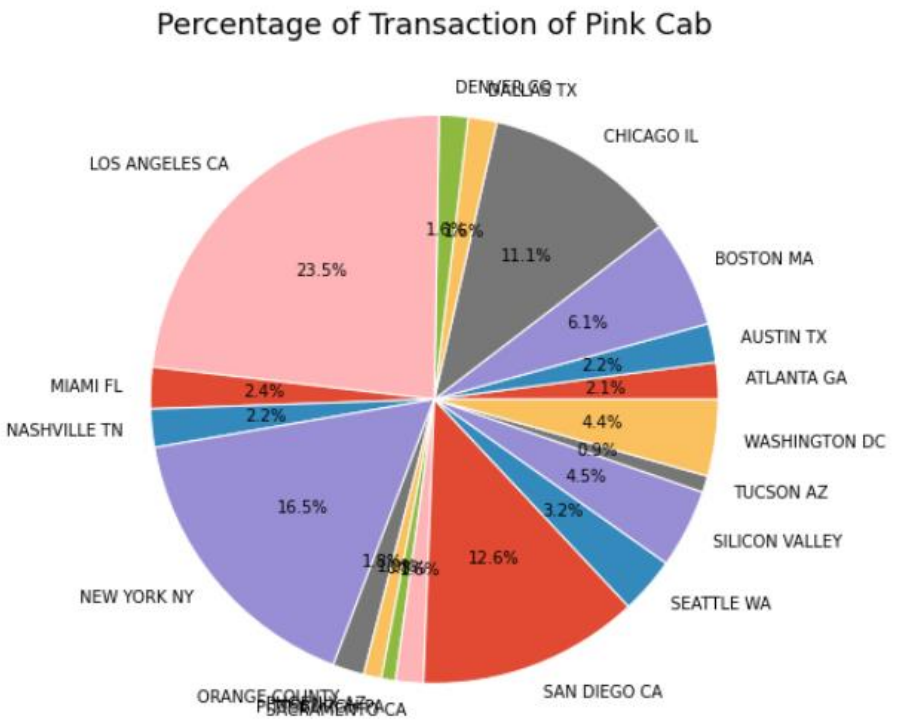


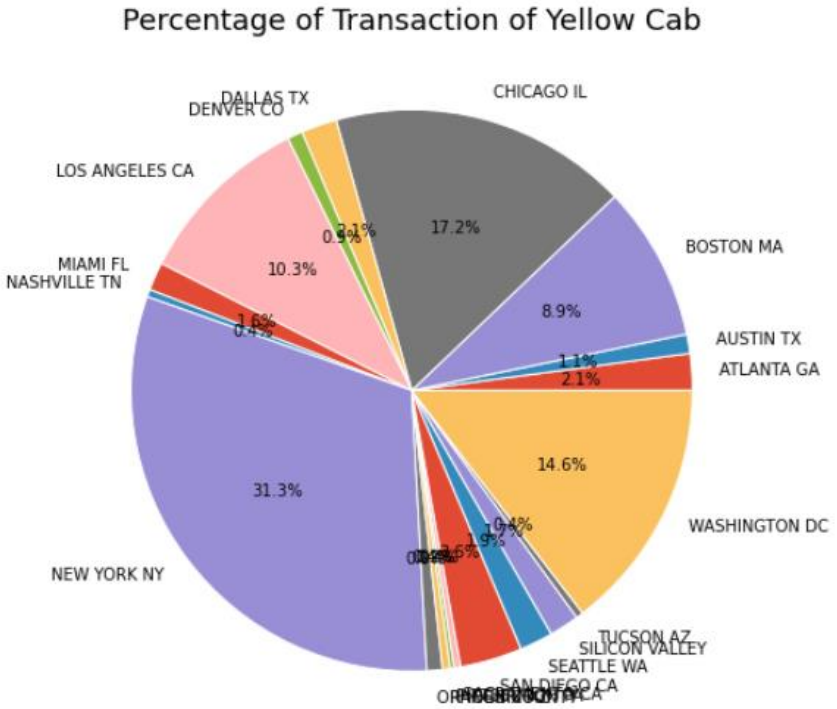
New York has the highest over all percentage of transaction at 27.8%

Followed by Chicago (15.8%), Los Angeles (13.4%) and Washington (12.2%).

Data Glacier
Your Deep Learning Partner

Percent of Total Transaction



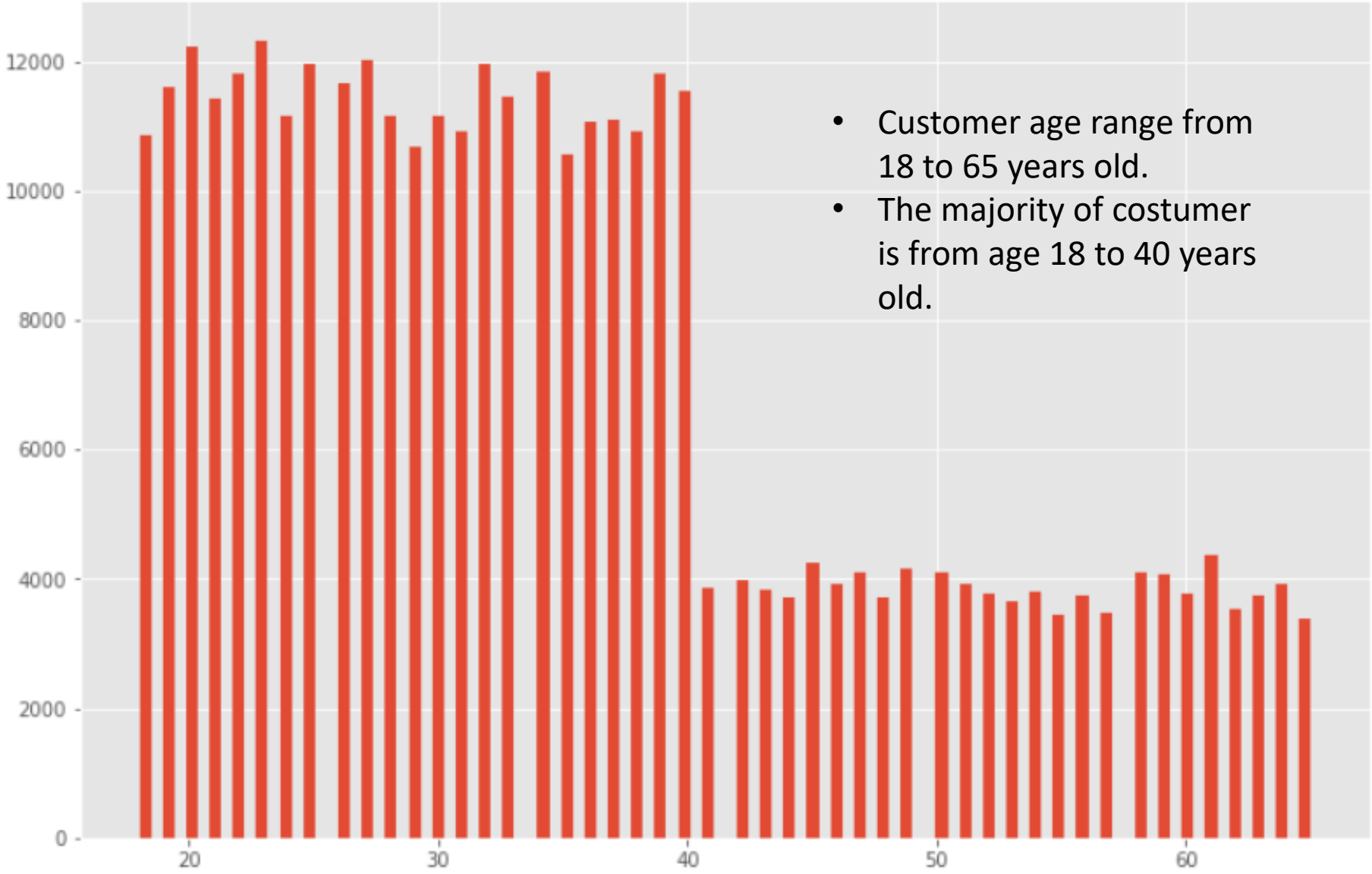


Pink Cab have more transactions in Los Angeles and San Diego

Yellow Cab have more transactions in New York, Chicago and Washington.

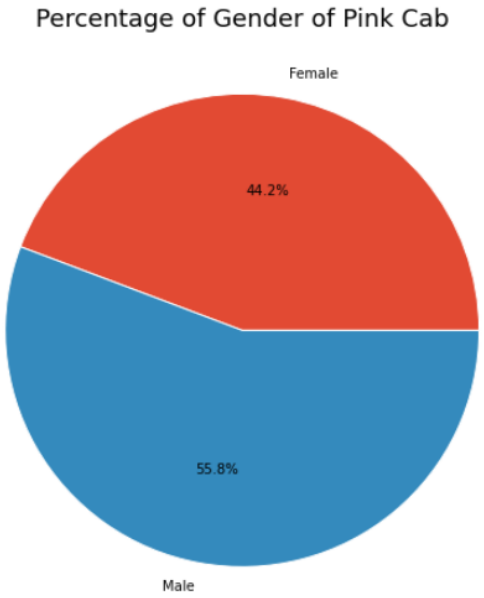
Data Glacier
Your Deep Learning Partner

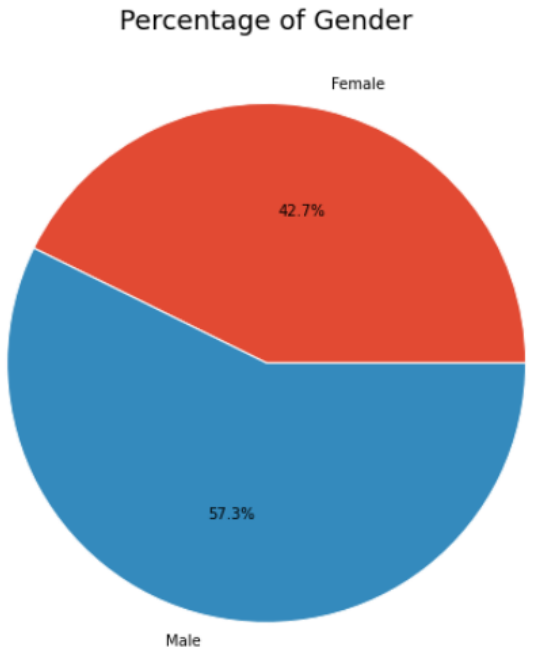
Costumer Age

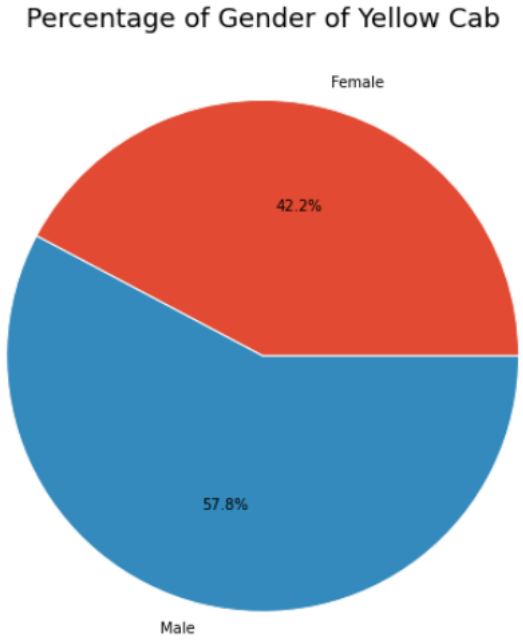


- Customer age range from 18 to 65 years old.
- The majority of costumer is from age 18 to 40 years old.

Gender

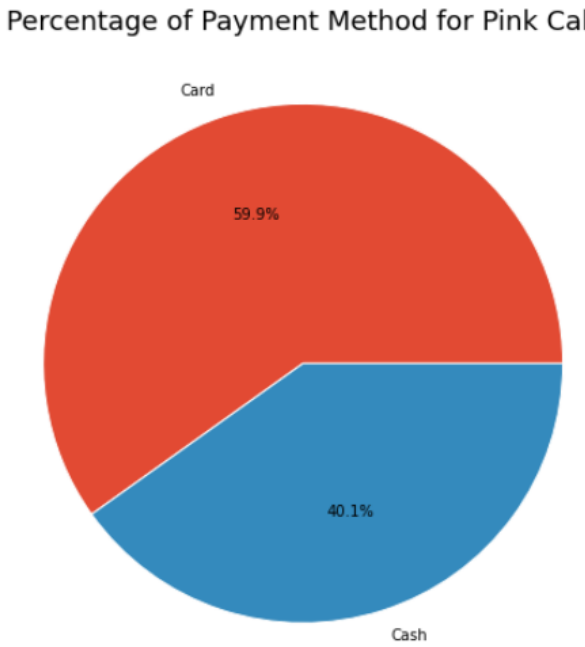


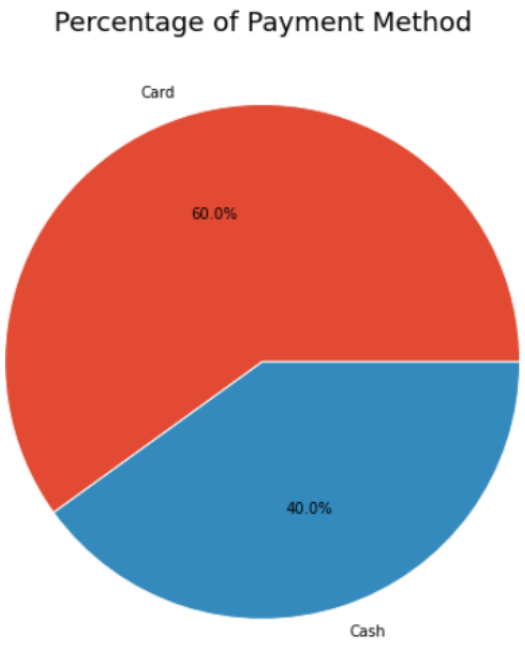


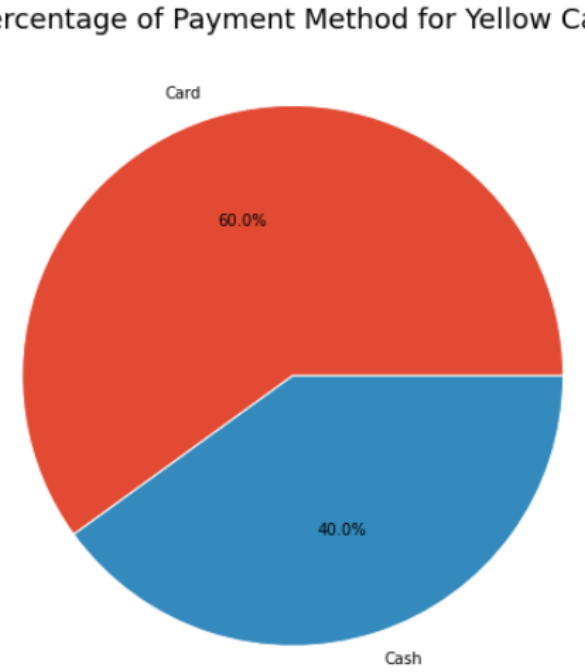


- More males uses the cab then females.

Payment Method

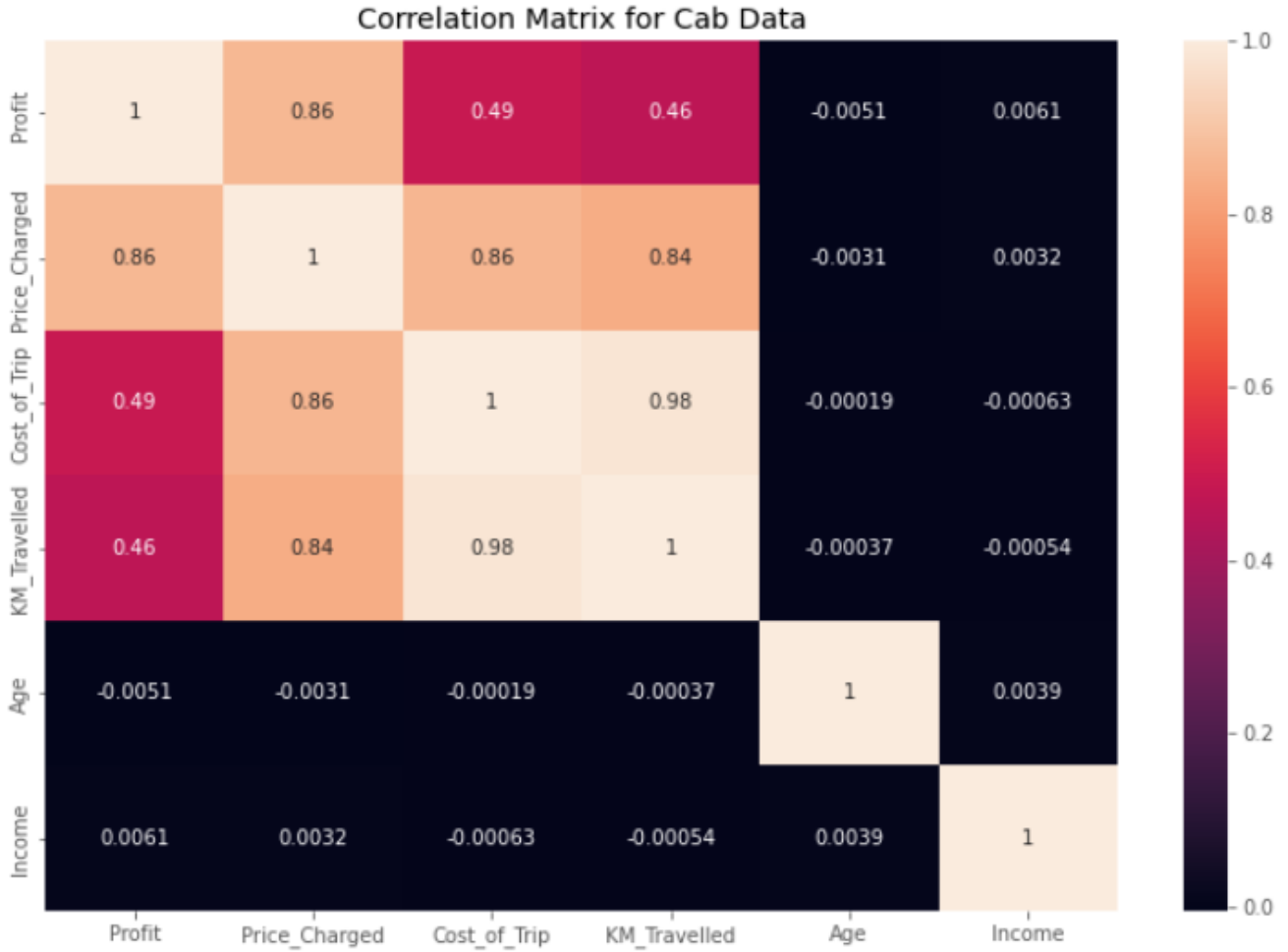






- 60% of the transactions are through card and 40% is with cash.

Correlations

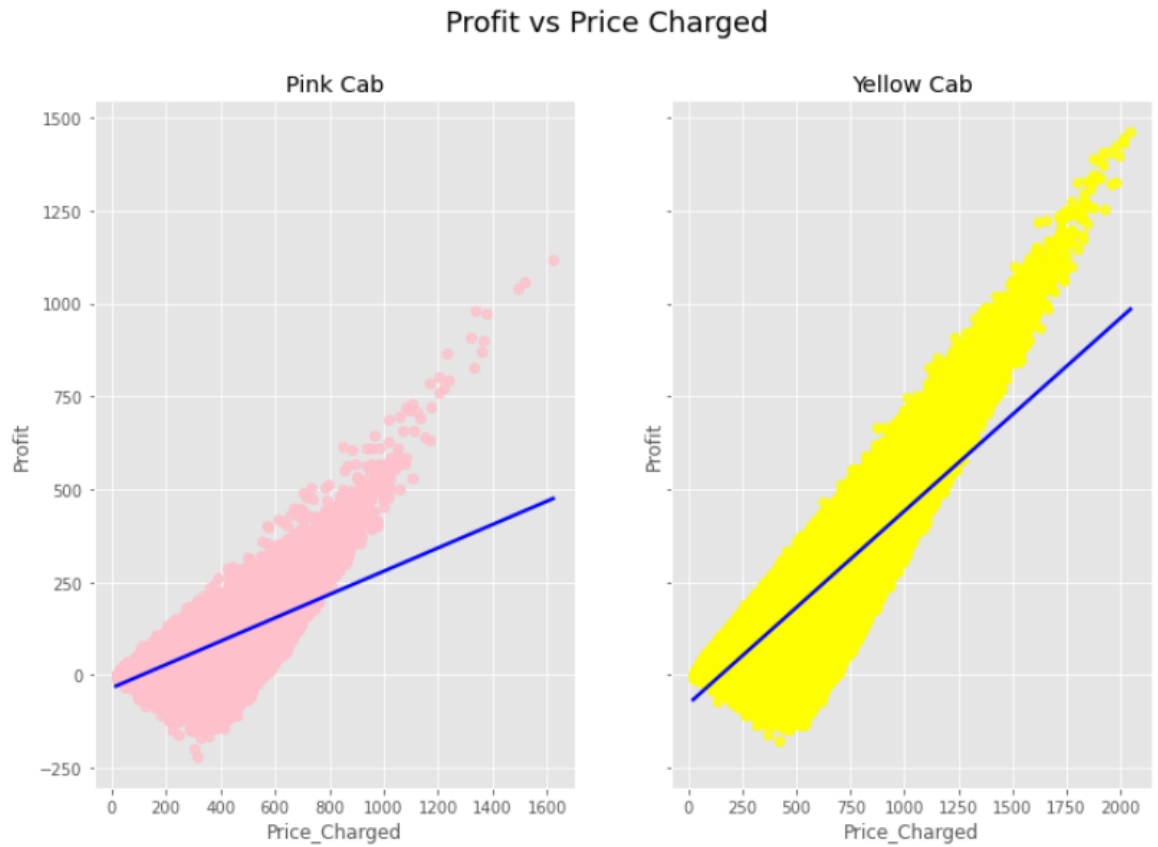


From the above heatmap, there is a big correlation between profit & price charged. There are also good correlation between Profit & Cost of Trip and Profit & KM Travelled

Profit per Distance Travelled

There is increase profit when increase in distance travelled for both company, but Yellow Cab have a higher correlation.

Profit vs Price Charged

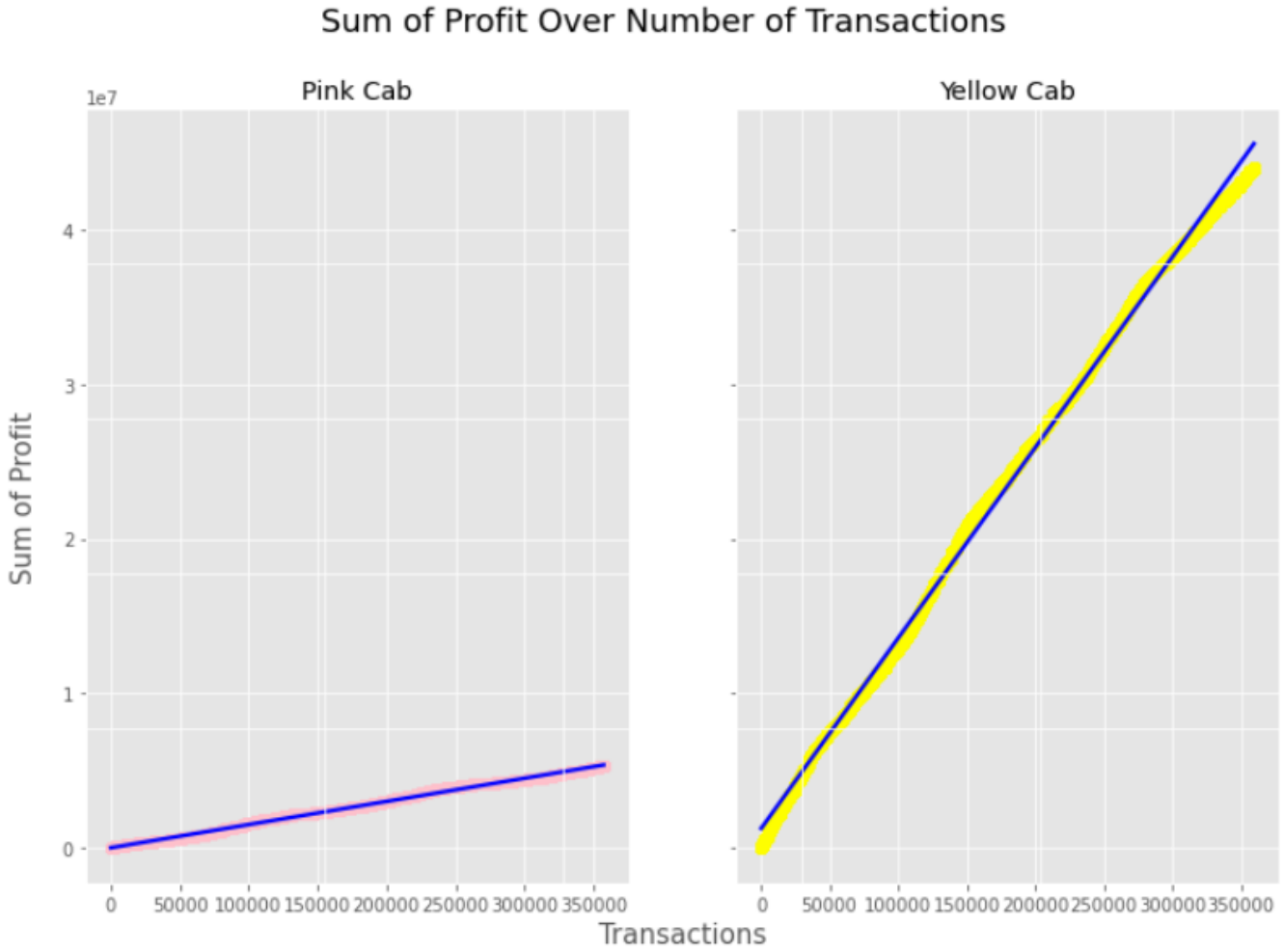


There is increase profit when increase in price charged for both company, but Yellow Cab have a higher correlation.

Price Charged for Distance Travelled

Yellow Cab charges more for distance travelled then Pink Cab

Profit Over Transactions



From the graph above, the profit of the Yellow Cab increase more with increasing transactions then Pink Cab.

Hypothesis Testing

Profit between Cash and Card Payments

H_0: There are no difference in profit between cash and card payments.

H_1: There are difference in profit between cash and card payments.

```
p_value = 0.4454179896745799
Do not reject the null hypothesis
```

H_0: There are no difference in profit between cash and card payments for Pink Cab.

H_1: There are difference in profit between cash and card payments for Pink Cab.

```
p_value = 0.7900535953280892
Do not reject the null hypothesis
```

H_0: There are no difference in profit between cash and card payments for Yellow Cab.

H_1: There are difference in profit between cash and card payments for Yellow Cab.

```
p_value = 0.2933058137696317
Do not reject the null hypothesis
```

Profit between Male and Female

H_0: More profit from female then male.

H_1: More profit from Male then Female.

```
p_value = 5.923495710616242e-37
Reject the null hypothesis
```

Hypothesis Testing

Profit Per Half Year

H_0: More profit is made in the second half of the year.

H_1: Less profit is made in the second half of the year.

```
p_value = 0.0
Reject the null hypothesis
```

Prices of Both Cab Company

H_0: Pink Cab Company charges more then Yellow Cab Company.

H_1: Yellow Cab Company charges more then Pink Cab Company.

```
p_value = 0.0
Reject the null hypothesis
```

EDA Summary

VS

- Monthly transaction range from 3,000 to 12,000.
- Majority of distance travelled are the same for both company (range from 12 KM to 33 KM)
- Yearly profit of around 2 Million
- More transactions in Los Angeles, New York, and San Diego
- Age group range from 18 to 65 year old, majority from 18 to 40
- Around 44.2% female and 55.8% male

- Monthly transaction range from 15,000 to 35,000.
- Majority of distance travelled are the same for both company (range from 12 KM to 33 KM)
- Yearly profit of 14 to 16.5 Million
- More transactions in New York, Chicago and Washington
- Age group range from 18 to 65 year old, majority from 18 to 40
- Around 42.2% female and 57.8% male

EDA Summary

VS

- Around 59.9% uses card and 40.1% uses cash to pay
- Less profit per distance then Yellow Cab
- Less profit per price charged then Yellow Cab
- Charge less per distance then Yellow Cab
- 1/8 times the total profit of Yellow Cab

- Around 60% uses card and 40% uses cash to pay
- More profit per distance then Pink Cab
- More profit per price charged then Pink Cab (almost double)
- Charge more per distance then Pink Cab
- 8 times the total profit of Pink Cab

Conclusion

In conclusion, based on the EDA and hypothesis testing, Yellow Cab company has 7 – 8 times the profit of Pink Cab company per year. The Yellow Cab company also has 3 – 5 times the about of transactions of Pink Cab. Yellow Cab makes more profit per distance, more profit per price charged, and more profit per distance. Therefore, XYZ should invest into Yellow Cab then Pink Cab.

Thank You