Exploratory Data Analysis

Project: G2M Insight For Cab Investment Firm

Date : 12.7.2023

Submitted By: Saanjanna Yuvaraj

Agenda

Problem Statement
Dataset Information
EDA
Hypothesis
Summary

Problem Statement

XYZ is a private firm in US. Due to remarkable growth in the Cab Industry in last few years it is planning for an investment in Cab industry and as per their Go-to-Market(G2M) strategy•

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

The analysis has been divided into four parts:

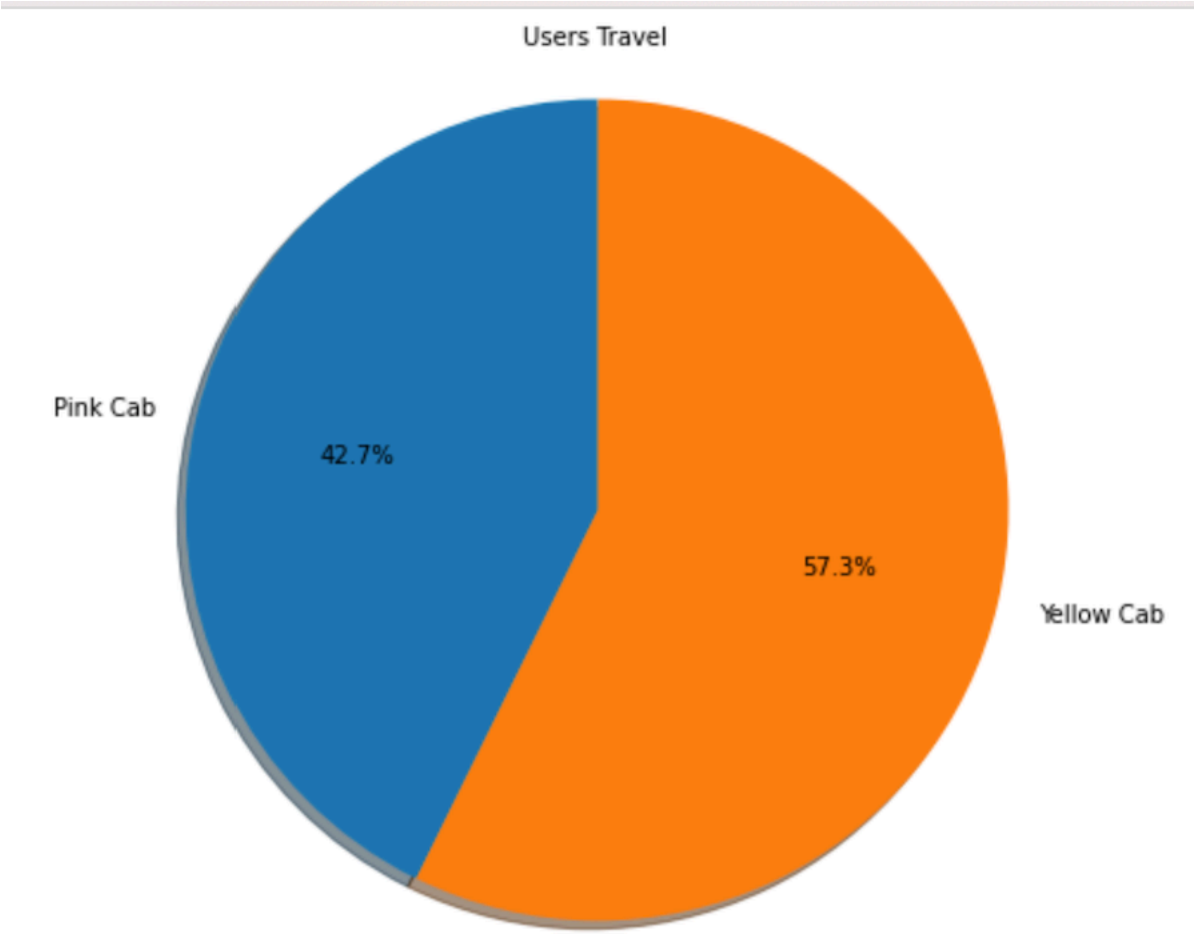
- Data Understanding and Visualisation
- Finding the most users Cab company
- Finding the cheapest Cab company for users
- Finding the most profitable Cab company

Data Information

1. Cab data.csv – this file includes details of transactions for 2 cab companies
2. City.csv – this file contains a list of US cities, no of cab users and their population
3. Transaction_id.csv – this is a mapping table that contains payment mode and transaction to customer mapping
4. Customer_ID.csv – this is a mapping table that contains a unique identifier that links the customer's demographic details

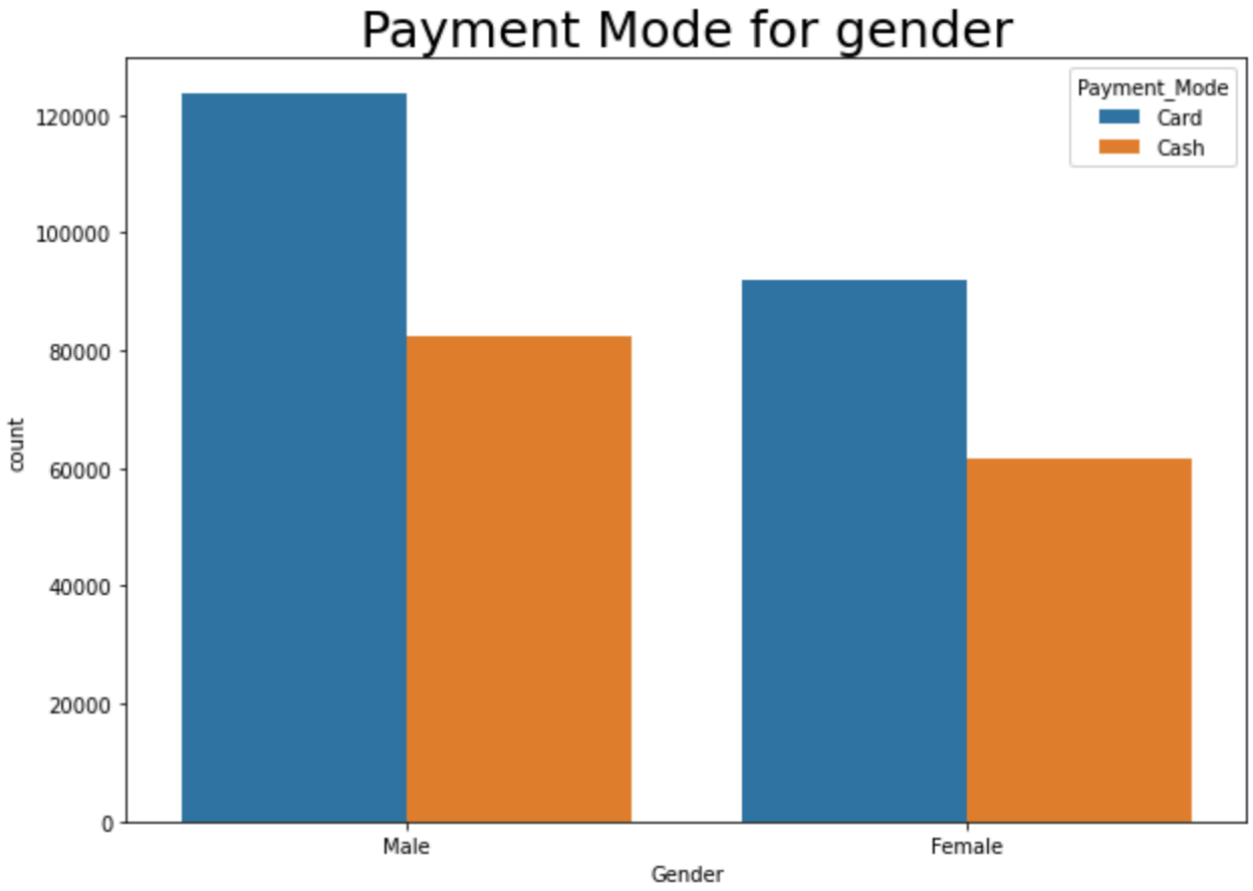
Investigate the data to see which cab the user uses most

As we can see from the pie chart user like to travel in yellow cab compared to pink cab



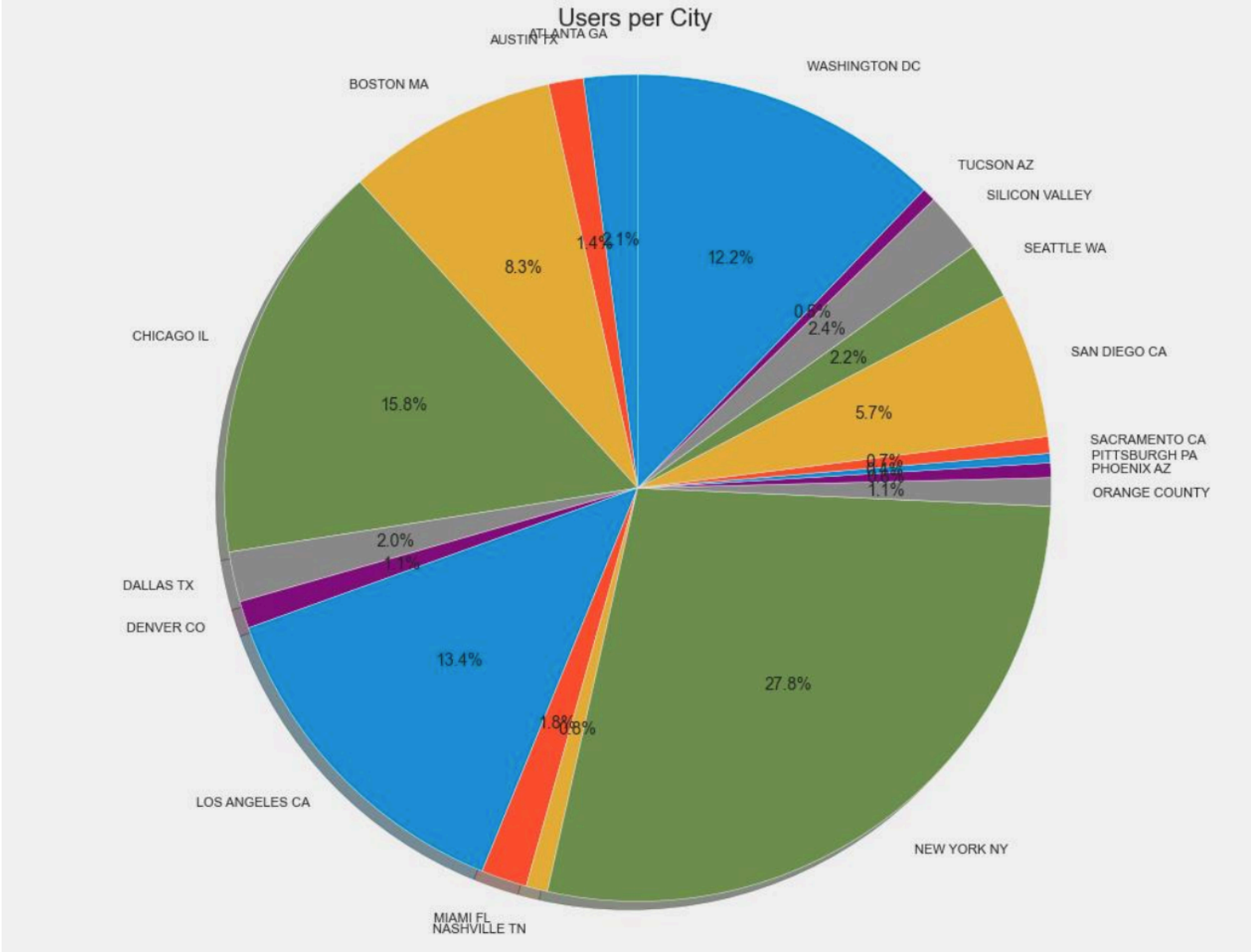
Investigate the data to see which payment the user prefers

As we can see from figure user prefer card compared to cash.



Users with respect to cities

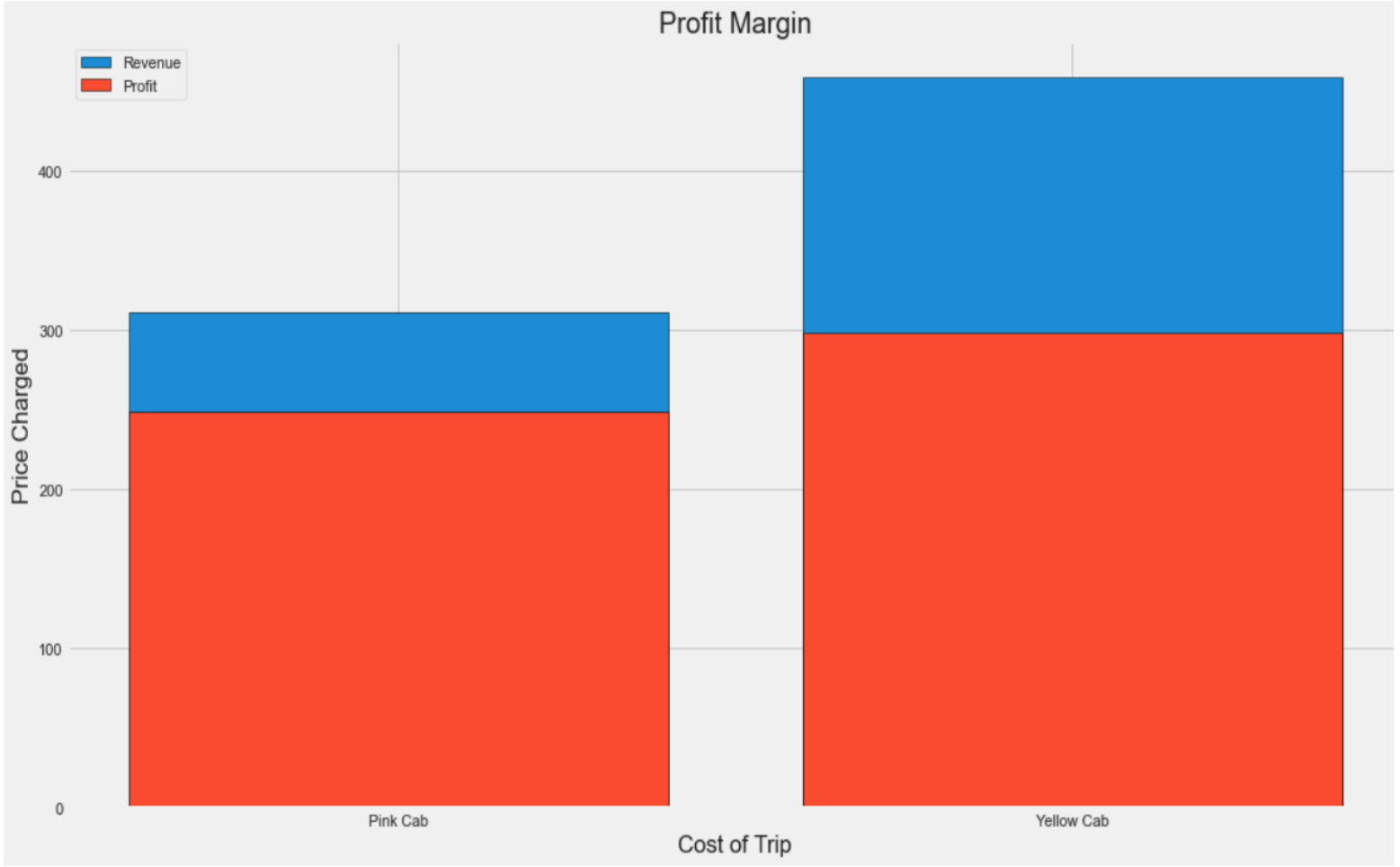
New York City has the highest Cab users compared to other cities.



Data Glacier
Your Deep Learning Partner

cab with respect to profit

Yellow cab has the higher profit compared to pink.



Data Glacier
Your Deep Learning Partner

Hypothesis 1” Is there any difference in payment mode?

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

```
mergedata[(mergedata['Payment_Mode']=='Cash')&(mergedata.Company=='Pink Cab')].groupby('Transaction ID').Profit.mean()
mergedata[(mergedata['Payment_Mode']=='Card')&(mergedata.Company=='Pink Cab')].groupby('Transaction ID').Profit.mean()

p_value = stats.ttest_ind(a.values,
                          b.values,
                          equal_var=True)

nt('P value is ', p_value)

p_value<0.05):
 print('We accept alternative hypothesis (H1) that there is a difference in payment mode for Pink Cab')
e:
 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
```

```
#yellow cab
```

```
a = mergedata[(mergedata['Payment_Mode']=='Cash')&(mergedata.Company=='Yellow Cab')].groupby('Transaction ID').Profit.m
b = mergedata[(mergedata['Payment_Mode']=='Card')&(mergedata.Company=='Yellow Cab')].groupby('Transaction ID').Profit.m

_, p_value = stats.ttest_ind(a.values,
                             b.values,
                             equal_var=True)

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  0.2933060638298729
We accept null hypothesis (H0) that there is no difference in payment mode for Yellow Cab
```

Summary

The above Exploratory Data Analysis (EDA) finds various dynamics of the data of two cab companies: Pink Cab and Yellow Cab.

- We found that there are no null values in the master data frame.
- Yellow Cab is seen having more profit than the Pink Cab.
- Yellow cab is preferred by the users over pink cab. In summary yellow cab would be the preferred one to invest in.

Thank You!