



Data Glacier

Your Deep Learning Partner

G2M Insight for Cab Investment Firm

Navodith Shankar

02/21/2023

Agenda

Executive Summary

Problem Statement

Approach

EDA

EDA Summary

Recommendations

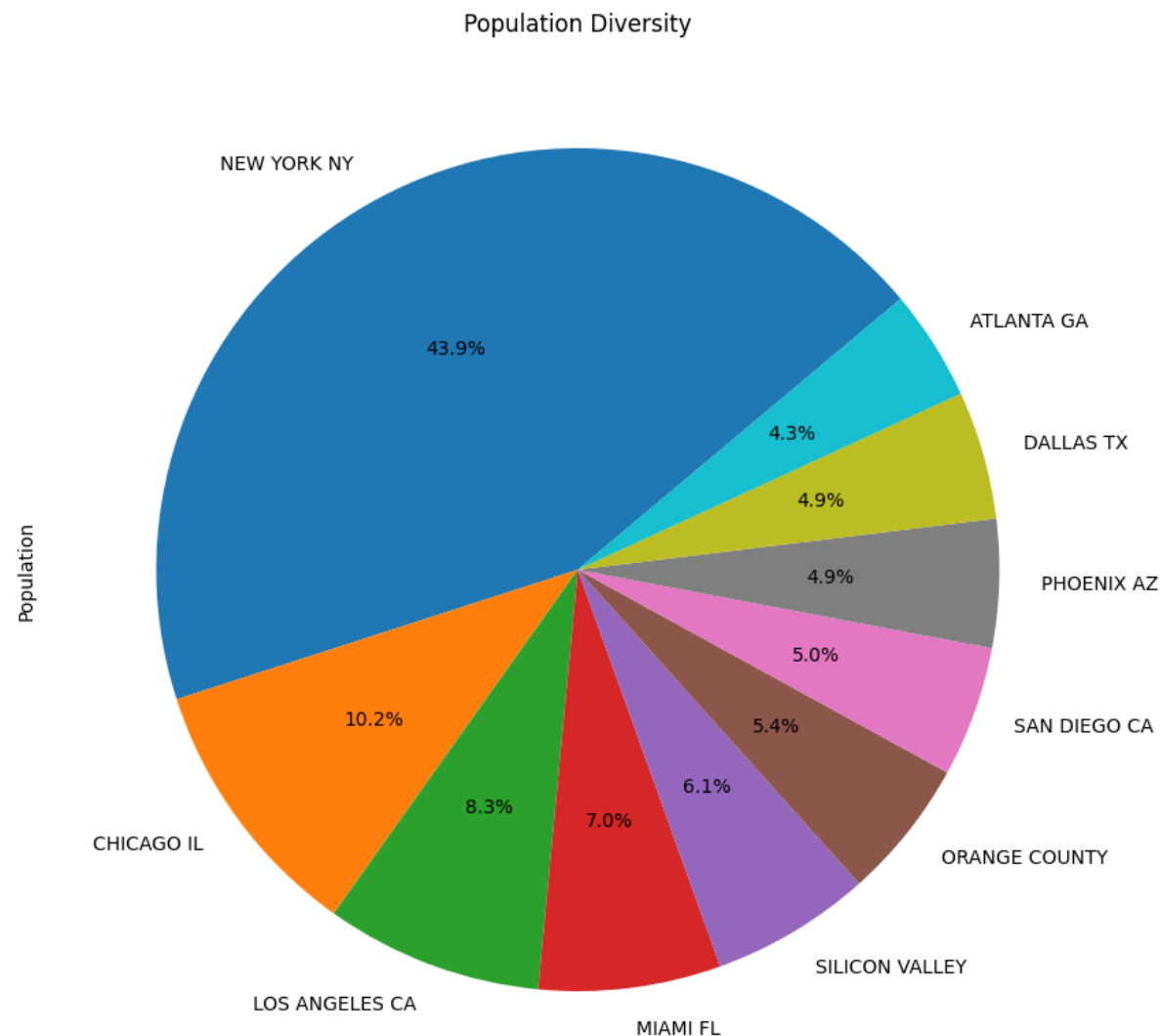
Problem Statement

- 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.
- XYZ is interested in using your actionable insights to help them identify the right company to make their investment.
- Analysis:
 - Find the most popular cab company
 - Find the best cab company
 - Find the best company for XYZ to invest in
 - Test multiple hypothesis statements

Data Sets

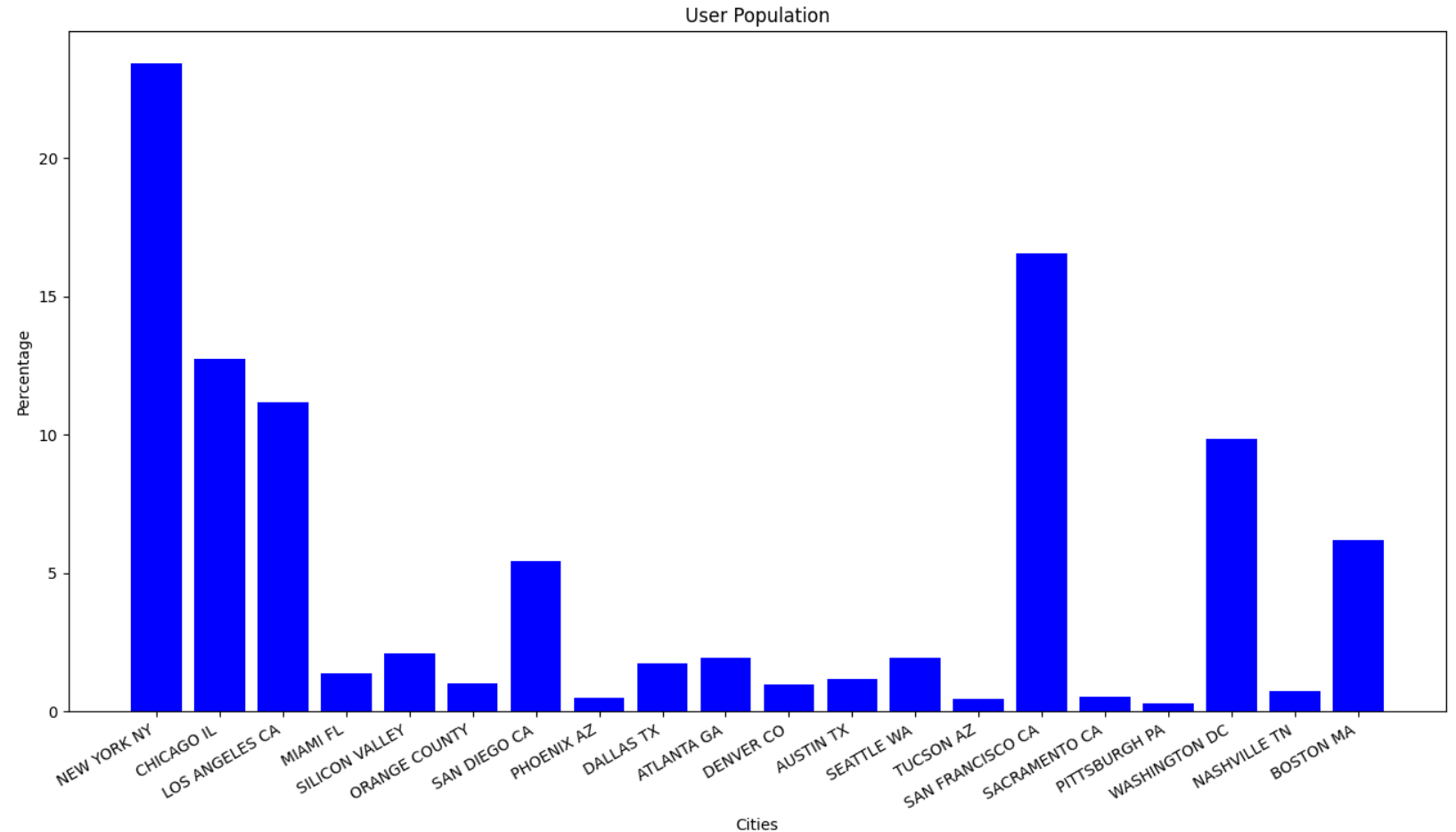
- Cab_Data – This Data Set includes data about Pink and Yellow Cabs
- City – This data set includes data about different US cities and users of each cab in each city
- Customer_ID – This data set includes data about the users of the cabs like their income and gender
- Transaction_ID – This data set includes data about the details of transactions made by users of the cabs

City Population



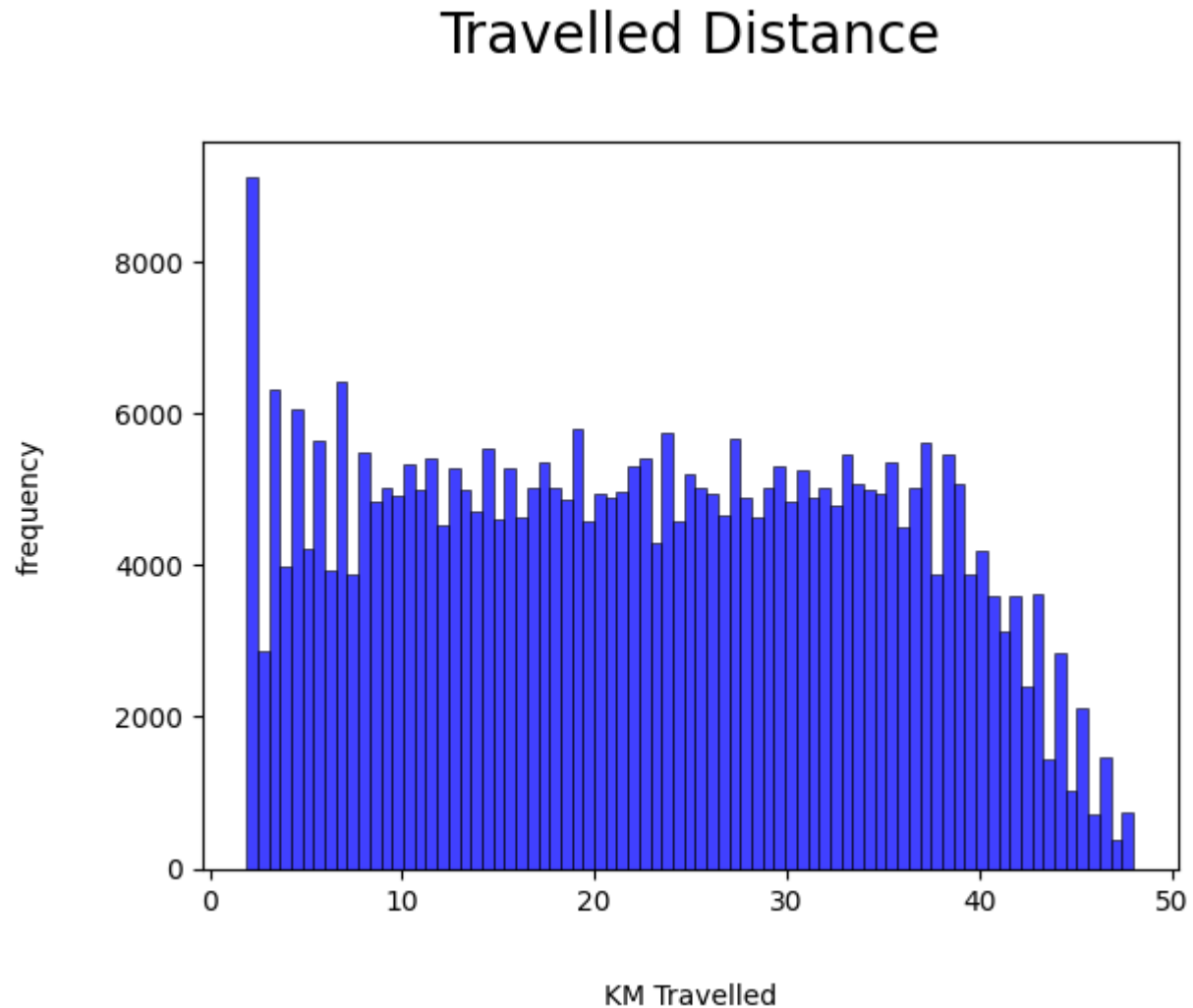
- As we can see, New York has the largest population percentage of 43.9%. We can presume that most users come from New York as it has the highest population. We can also presume that population of a city plays a major role in the number of users.

User Population



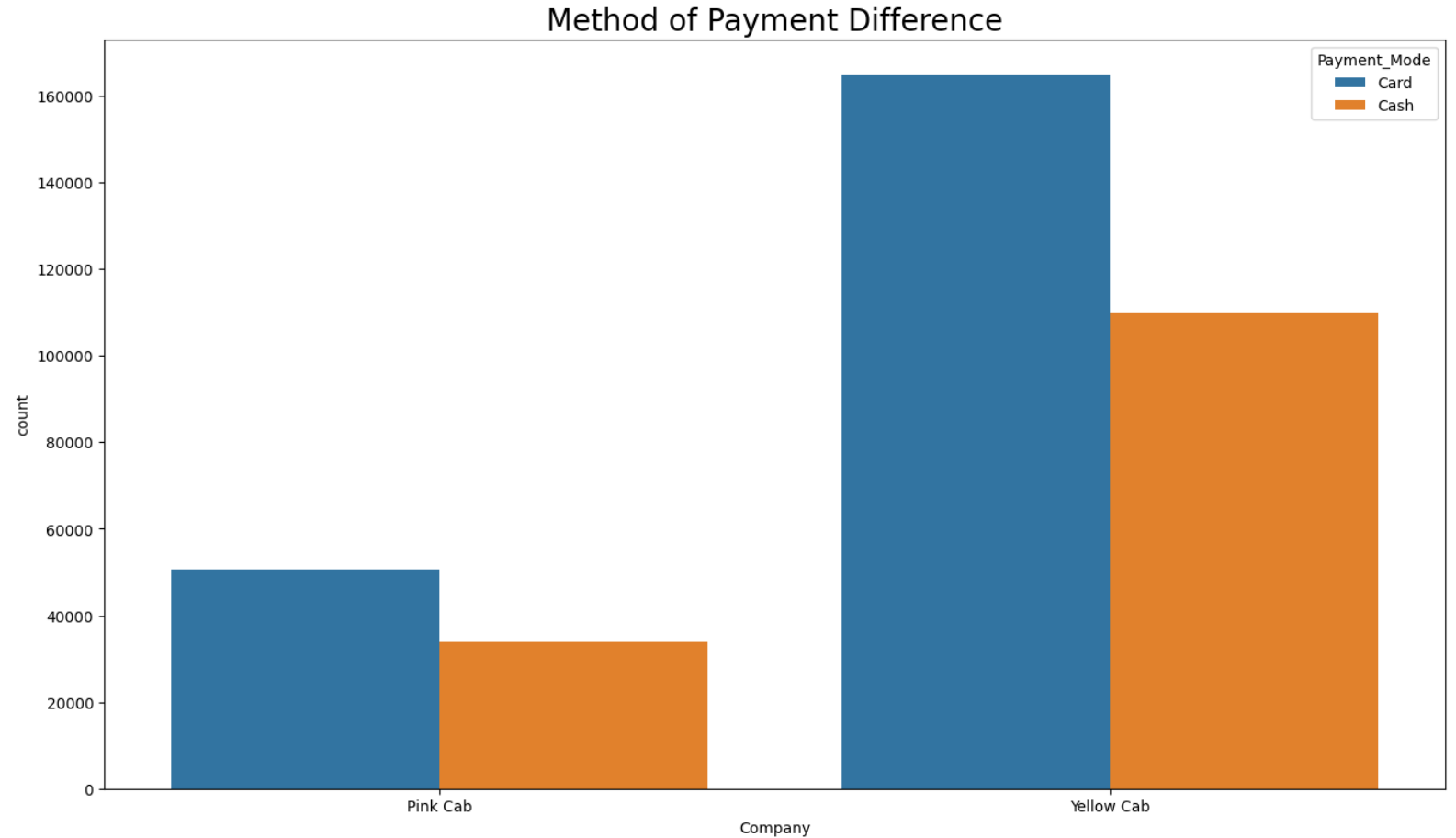
- As we can see, from what we presumed before, majority of users reside in New York with San Francisco being the second highest in user population. From this we can conclude that city population heavily affects user population in each city.

Common Trip Distance



- As we can see, between both cab companies the most travelled distance is 2km. This makes sense as most people in cities travel short distances.

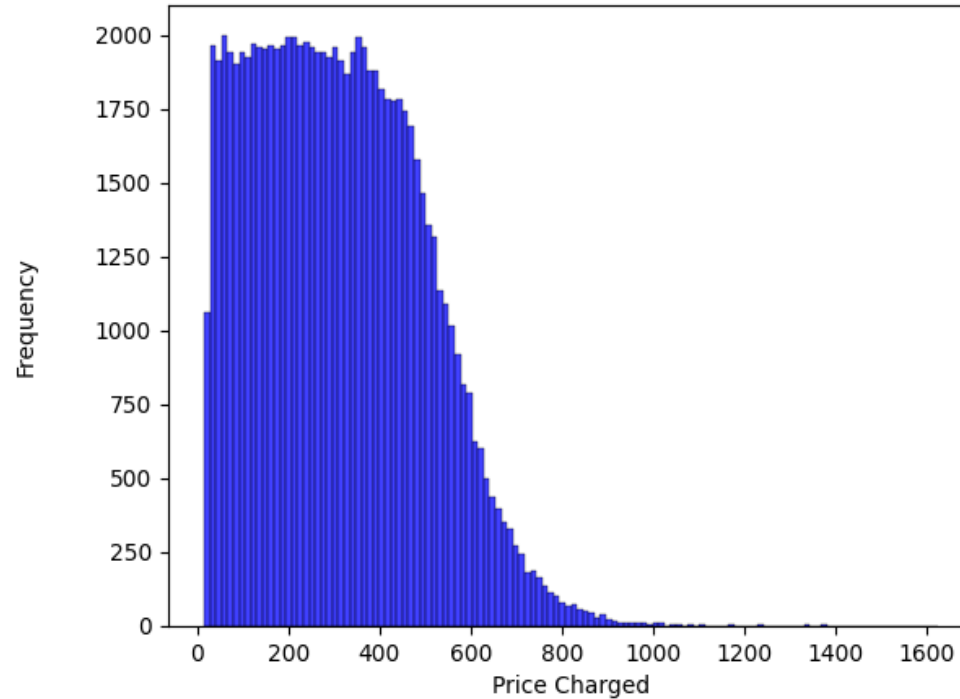
Method of Payment



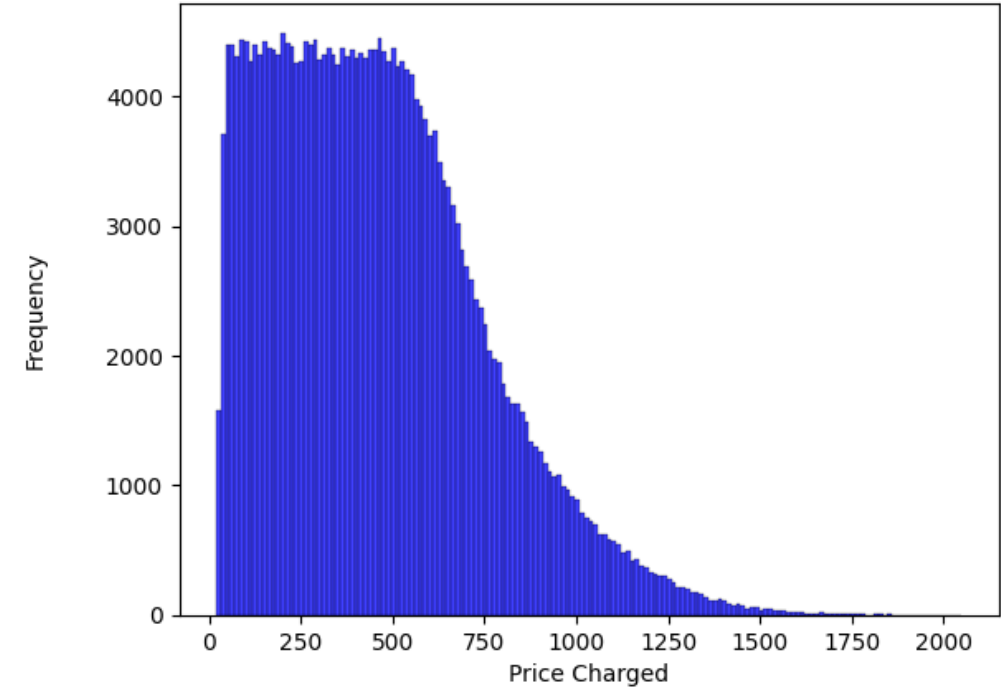
- As we can see, between both cab companies most users pay via card, but we can also see a huge difference in the number of payments made between both cabs. Yellow cab has a larger number of payments, this yields that more users use yellow cab instead of pink cab.

Price Charged

Price Charged (PINK)



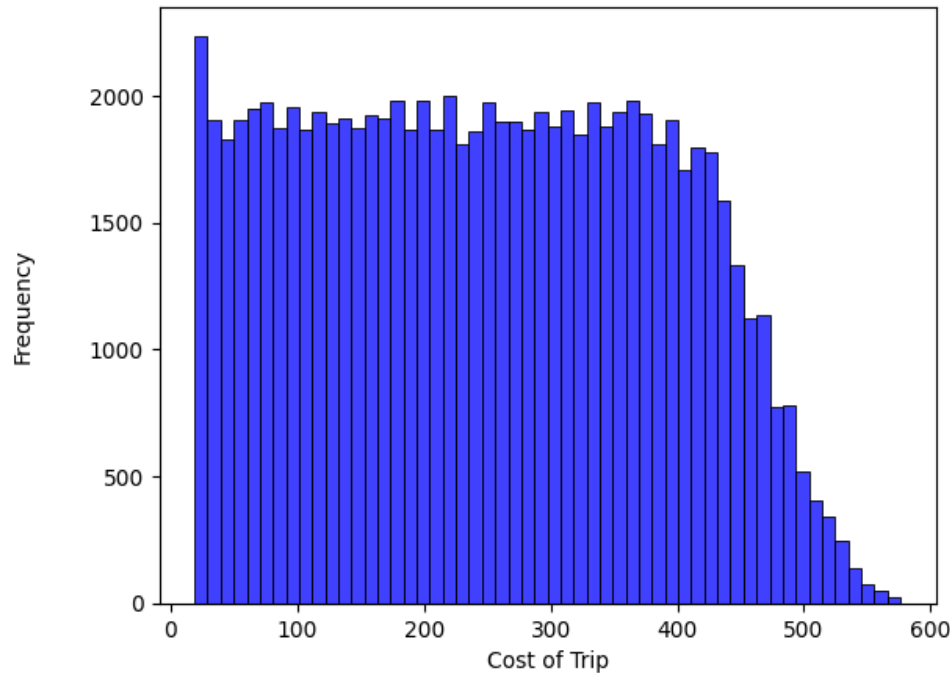
Price Charged (YELLOW)



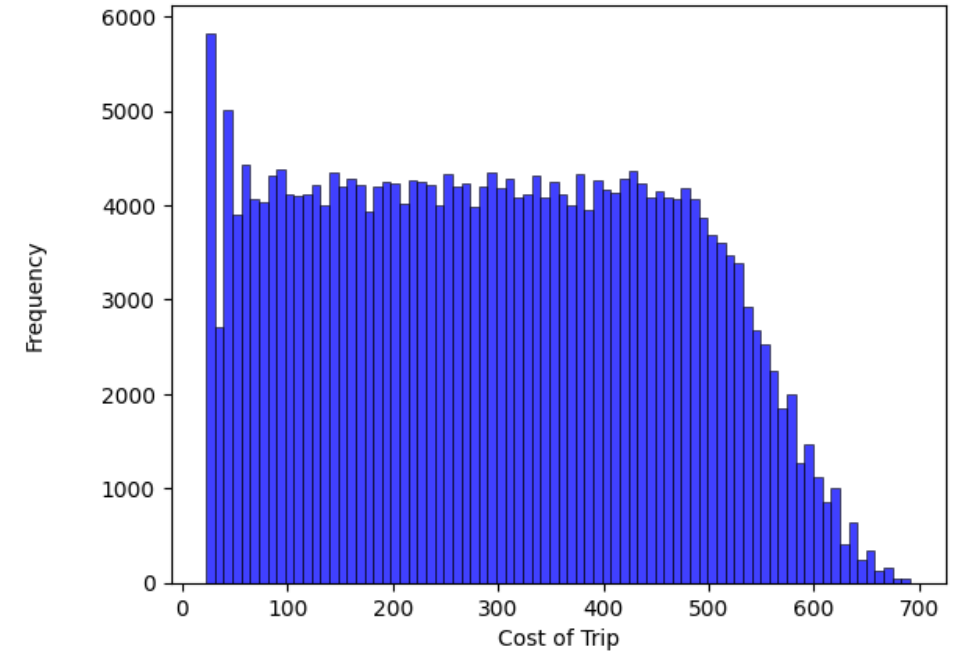
- As we can see, between both cab companies pink cab charges users less compared to yellow cab. Yellow cab has charged close to \$2000 whereas pink cab has never charged more than \$1400.

Cost of Trip

Cost of Trip (PINK)

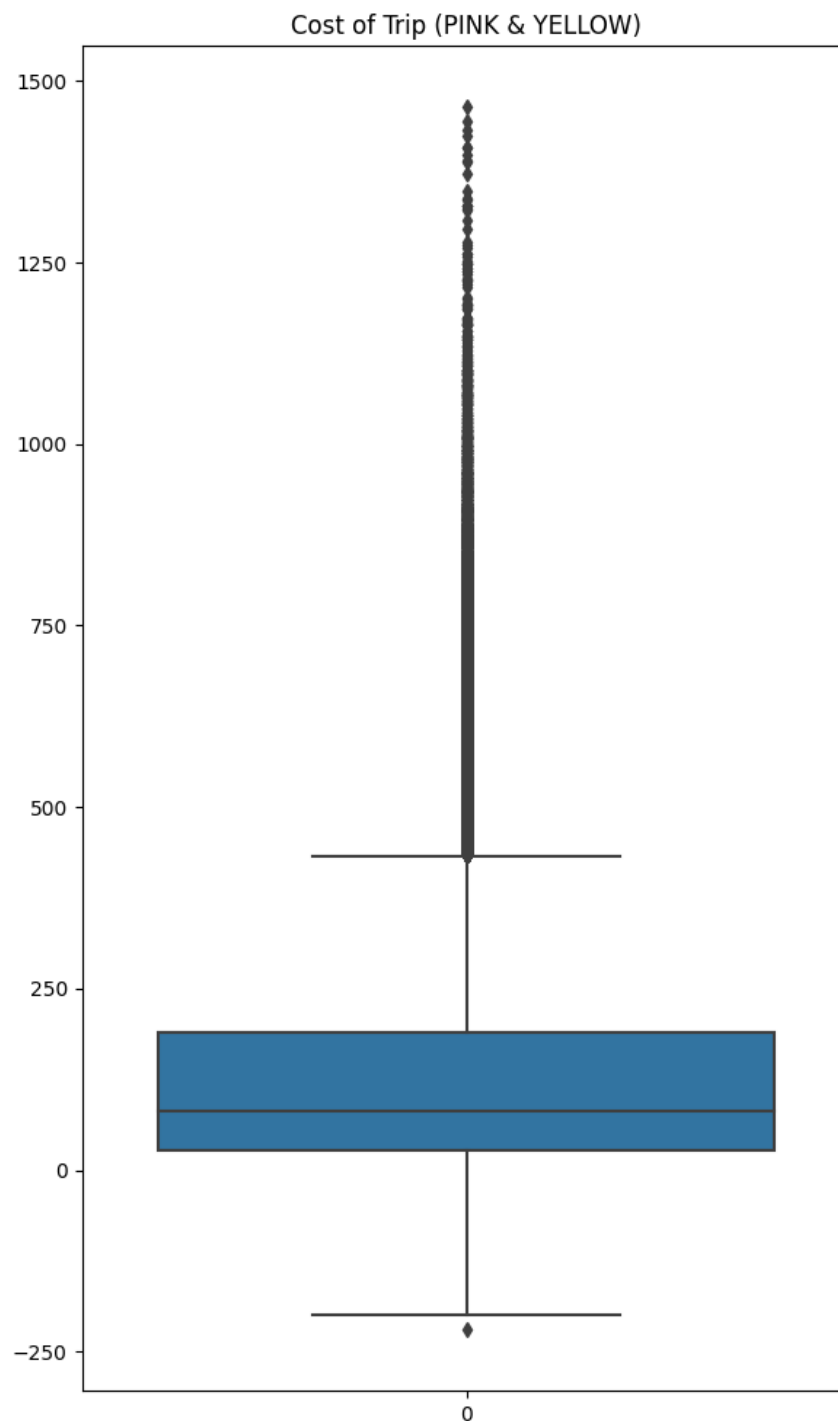


Cost of Trip (YELLOW)



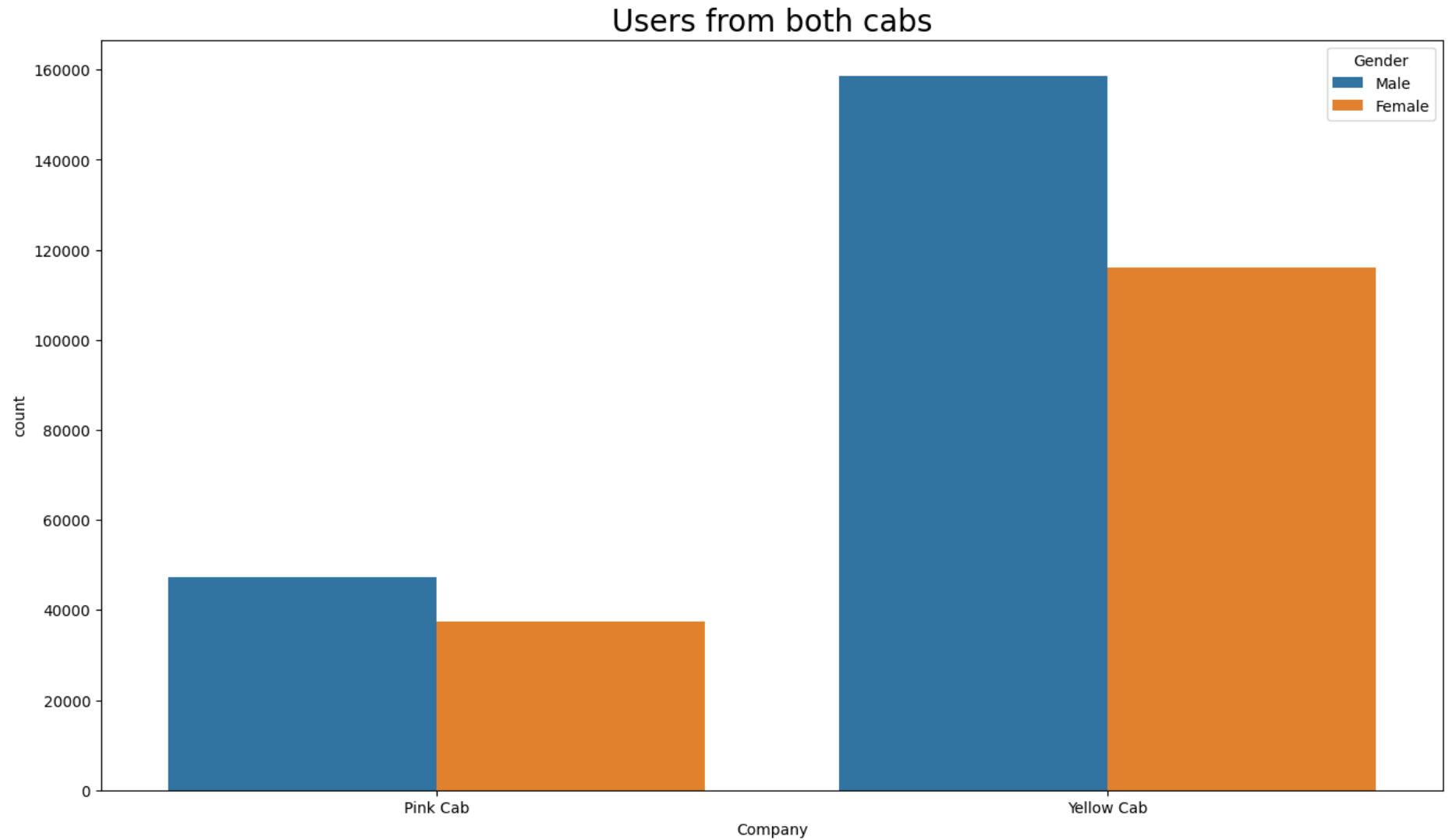
- As mentioned before, yellow cab is more expensive. By looking at the cost of trip we can surely say that yellow cab is expensive. The cause of this could also be because more users use yellow cab which results in different type of users needing to travel different distances.

Cost of Trip



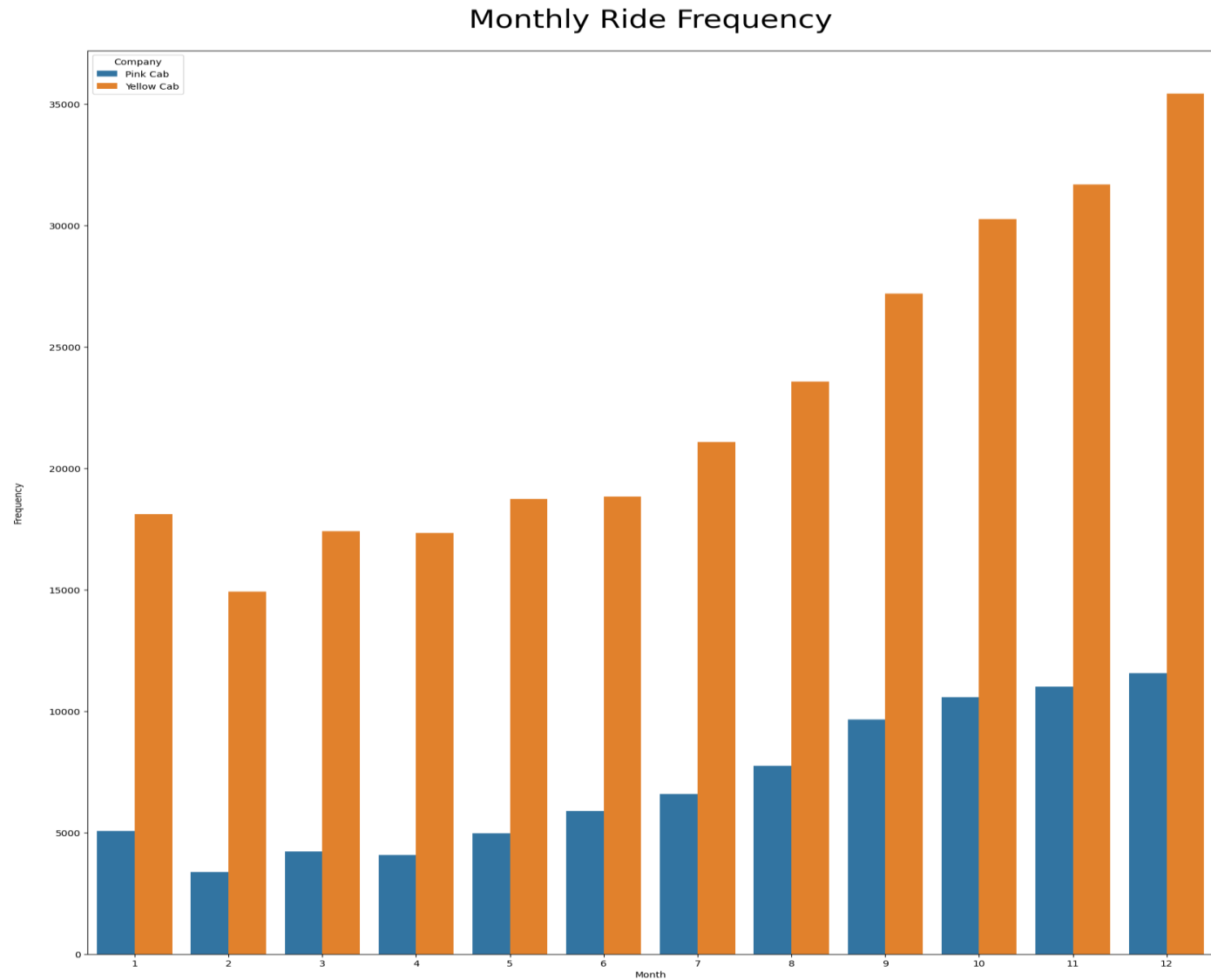
- Here we can see the cost of trip with both companies combined. We can see that most trips do not exceed the cost of \$1200.

Gender



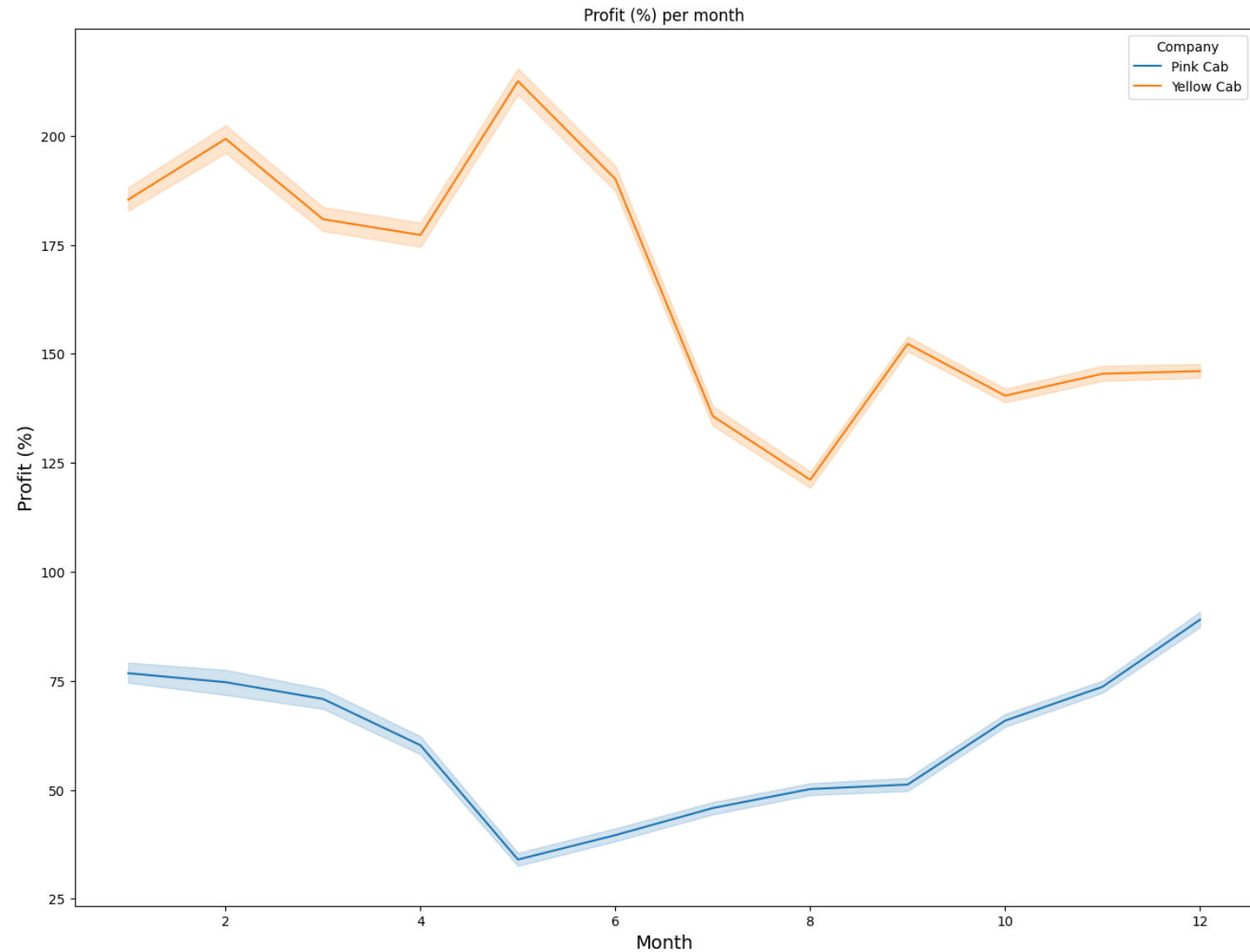
- As we can see, between both cab companies mostly males use pink and yellow cab. Even though mostly males use pink cab, the difference between male and female is minimal where the difference in yellow cab is greater.

Ride Preference



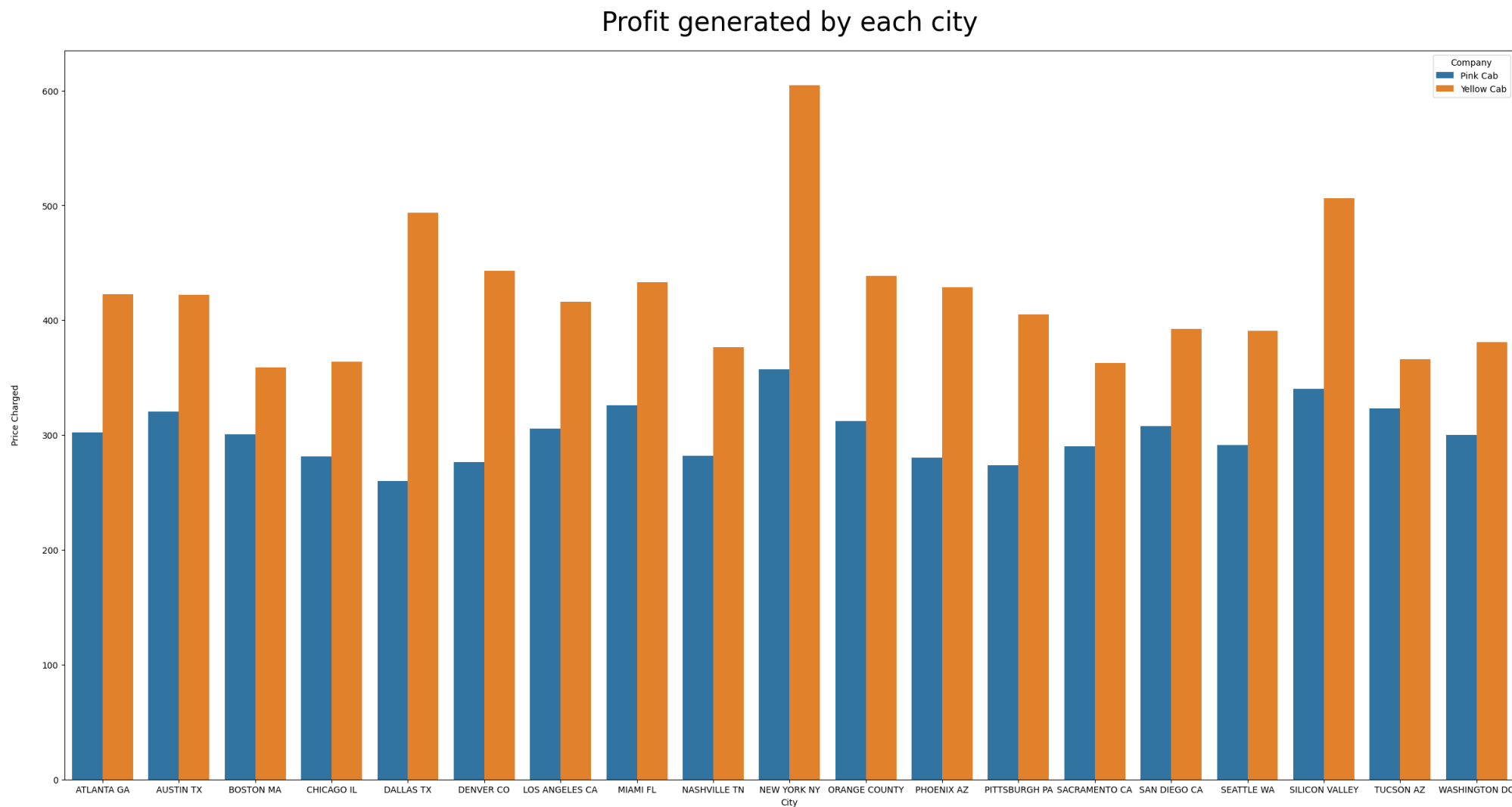
- As we can see, between both cab companies most users tend to travel in yellow cab which ultimately means that yellow cab is the most popular cab company. In both cabs, december is the most popular month but yellow cab has more than double the rides.

Profit



- Previously, it was found that yellow cab was the most popular cab. By looking at the profits between both cabs we can see that the result of more popularity is more profit generated. Yellow cab generates more profit than pink cab.

Profit per city



- By looking at this bar plot, we can see that New York generates the most profit for both cab companies with yellow cab generating the most profit.

Hypothesis #1

H1: Age impacts the amount of Profit/Margin generated by a cab. There is a relation

```
a1 = DataMerge[(DataMerge.Age <= 50)&(DataMerge.Company=='Pink Cab')].groupby('Transaction ID').Profit.mean()
b1 = DataMerge[(DataMerge.Age >= 50)&(DataMerge.Company=='Pink Cab')].groupby('Transaction ID').Profit.mean()
_, p_value = stats.ttest_ind(a1.values, b1.values, equal_var=True)

if(p_value<0.05):
    print('It is concluded there is a relation between Profit and Age in PINK Cab. Hypothesis is true')
else:
    print('It is concluded there is no relation between Profit and Age in PINK Cab. Hypothesis is false')
```

It is concluded there is no relation between Profit and Age in PINK Cab. Hypothesis is false

```
a = DataMerge[(DataMerge.Age <= 50)&(DataMerge.Company=='Yellow Cab')].groupby('Transaction ID').Profit.mean()
b = DataMerge[(DataMerge.Age >= 50)&(DataMerge.Company=='Yellow Cab')].groupby('Transaction ID').Profit.mean()
_, p_value = stats.ttest_ind(a.values, b.values, equal_var=True)

if(p_value<0.05):
    print('It is concluded there is a relation between Profit and Age in YELLOW Cab. Hypothesis is true')
else:
    print('It is concluded there is no relation between Profit and Age in YELLOW Cab. Hypothesis is false')
```

It is concluded there is a relation between Profit and Age in YELLOW Cab. Hypothesis is true

There is a relation between Profit and Age in Yellow Cab and not Pink Cab. If tested with 70 instead, both cabs indicate that there is no relation between Profit and Age.

Hypothesis #2

H2: Gender impacts the amount of Profit/Margin generated by a cab. There is a relation

```
a = DataMerge[(DataMerge.Gender=='Female')&(DataMerge.Company=='Pink Cab')].groupby('Transaction ID').Profit.mean()
b = DataMerge[(DataMerge.Gender=='Male')&(DataMerge.Company=='Pink Cab')].groupby('Transaction ID').Profit.mean()

_, p_value = stats.ttest_ind(a.values,b.values,equal_var=True)
if(p_value<0.05):
    print('It is concluded there is a relation between Profit and Age in PINK Cab. Hypothesis is true')
else:
    print('It is concluded there is no relation between Profit and Age in PINK Cab. Hypothesis is false')
```

It is concluded there is no relation between Profit and Age in PINK Cab. Hypothesis is false

```
a = DataMerge[(DataMerge.Gender=='Female')&(DataMerge.Company=='Yellow Cab')].groupby('Transaction ID').Profit.mean()
b = DataMerge[(DataMerge.Gender=='Male')&(DataMerge.Company=='Yellow Cab')].groupby('Transaction ID').Profit.mean()

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

if(p_value<0.05):
    print('It is concluded there is a relation between Profit and Age in YELLOW Cab. Hypothesis is true')
else:
    print('It is concluded there is no relation between Profit and Age in YELLOW Cab. Hypothesis is false')
```

It is concluded there is a relation between Profit and Age in YELLOW Cab. Hypothesis is true

In Pink cab there is not relation indicated but in Yellow cab there is a relationship indicated between Profit and Gender.



Hypothesis #3

H3: The method of payment does not affect the Profit generated by a cab. There is no relation

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

_, p_value = stats.ttest_ind(a.values,b.values,equal_var=True)
if(p_value<0.05):
    print('It is concluded there is a relation between Profit and Method of Payment in PINK Cab. Hypothesis is true')
else:
    print('It is concluded there is no relation between Profit and Method of Payment in PINK Cab. Hypothesis is false')
```

It is concluded there is no relation between Profit and Method of Payment in PINK Cab. Hypothesis is false

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

_, p_value = stats.ttest_ind(a.values,b.values,equal_var=True)
if(p_value<0.05):
    print('It is concluded there is a relation between Profit and Method of Payment in Yellow Cab. Hypothesis is true')
else:
    print('It is concluded there is no relation between Profit and Method of Payment in Yellow Cab. Hypothesis is false')
```

It is concluded there is no relation between Profit and Method of Payment in Yellow Cab. Hypothesis is false



Hypothesis

#4

H4: Gender impacts the amount of KM travelled which results in the Profit changing. There is a relation

```
a = DataMerge[(DataMerge.Payment_Mode=='Female')&(DataMerge.Company=='Pink Cab')].groupby('KM Travelled').Profit.mean()
b = DataMerge[(DataMerge.Payment_Mode=='Male')&(DataMerge.Company=='Pink Cab')].groupby('KM Travelled').Profit.mean()

_, p_value = stats.ttest_ind(a.values,b=b.values,equal_var=True)
if(p_value<0.05):
    print('It is concluded there is a relation between Gender impacting KM Travelled which in the end changes the Profit in PINK Cab. Hypothesis is True')
else:
    print('It is concluded there is no relation between Gender impacting KM Travelled which in the end does not change the Profit in PINK Cab. Hypothesis is False')
```

It is concluded there is no relation between Gender impacting KM Travelled which in the end does not change the Profit in PINK Cab. Hypothesis is false

```
a = DataMerge[(DataMerge.Payment_Mode=='Female')&(DataMerge.Company=='Yellow Cab')].groupby('KM Travelled').Profit.mean()
b = DataMerge[(DataMerge.Payment_Mode=='Male')&(DataMerge.Company=='Yellow Cab')].groupby('KM Travelled').Profit.mean()

_, p_value = stats.ttest_ind(a.values,b=b.values,equal_var=True)
if(p_value<0.05):
    print('It is concluded there is a relation between Gender impacting KM Travelled which in the end changes the Profit in Yellow Cab. Hypothesis is True')
else:
    print('It is concluded there is no relation between Gender impacting KM Travelled which in the end does not change the Profit in Yellow Cab. Hypothesis is False')
```

It is concluded there is no relation between Gender impacting KM Travelled which in the end does not change the Profit in Yellow Cab. Hypothesis is false

Conclusion

1. Yellow cab is more popular and generates more profit. Users are more likely to use Yellow cab instead of Pink Cab
2. Yellow cab is also more expensive than Pink cab, Pink cab is cheaper to travel in.
3. New York is the most popular city with the most population and the most amount of users
4. Yellow cab also generates the most in New York with Silicon Valley being the second highest
5. Yellow cab also offers discounts for 50 to 60 years of age users

Overall, Yellow cab is more popular and generates more profit than Pink cab. The private firm XYZ, should invest in Yellow cab as the analysis performed illustrates that Yellow cab is the most profitable which will result in a good investment. Pink cab is not worth the investment as they are less popular and do not generate as much profit as Yellow Cab.



G2M Insight for Cab Investment Firm

Navodith Shankar
02/21/2023

Thank You