Data Glacier

Your Deep Learning Partner

Presentation On: G2M Insight For Cab Investment Firm

Submitted By: Shweta Singh

Date: 21-Oct-2022

Background

XYZ is a private equity 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.

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

The analysis has been divided into four parts:

- Data Understanding
- Forecasting profit and number of rides for each cab type
- Finding the most profitable Cab company
- Recommendations for investment

Data Information

Cab_Data
Total no.of observations : 359392
Total no.of features : 7
Base format of the file : csv
Size of data : 20.663 MB

Transaction_ID
Total no.of observations : 440098
Total no.of features : 3
Base format of the file : csv
Size of data :8.788 MB

Customer_ID
Total no.of observations : 49171
Total no.of features : 4
Base format of the file : csv
Size of data : 1.027 MB

City
Total no.of observations : 20
Total no.of features : 3
Base format of the file : csv
Size of data : 0.001 MB

Proposed Approach: There is no missing values in all four datasets that mentioned above.

Master Data

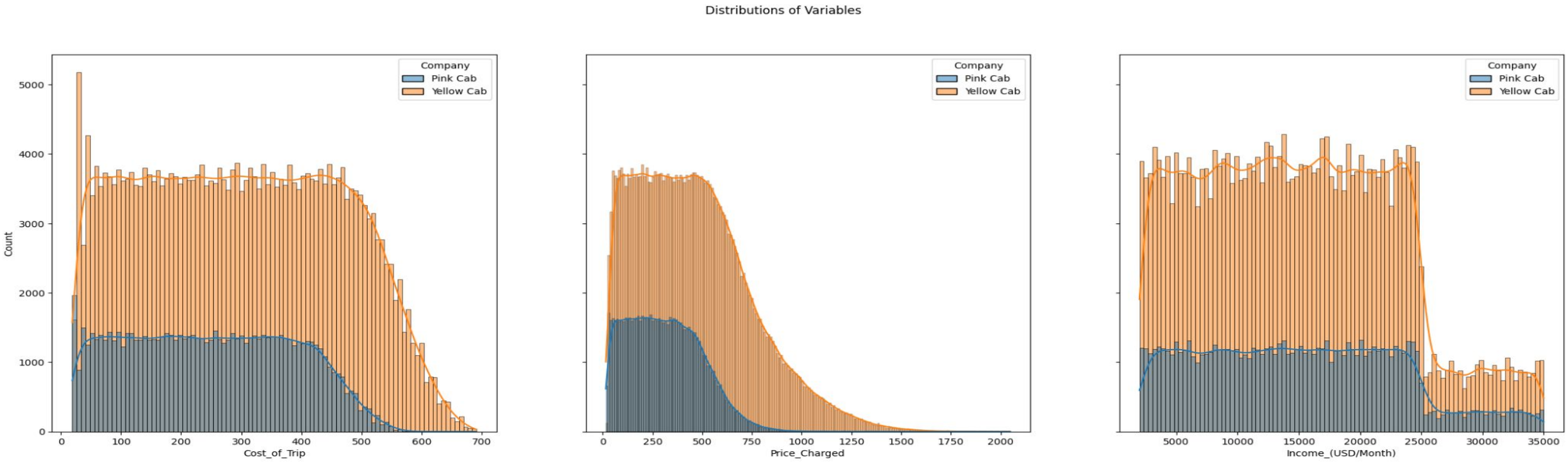
	Transaction ID	Date of Travel	Company	City	KM Travelled	Price Charged	Cost of Trip	Customer ID	Payment_Mode	Gender	Age	Income (USD/Month)	Population	Users
0	10000011	01/08/2016	Pink Cab	ATLANTA GA	30.45	370.95	313.6350	29290	Card	Male	28	10813	814,885	24,701
1	10351127	07/21/2018	Yellow Cab	ATLANTA GA	26.19	598.70	317.4228	29290	Cash	Male	28	10813	814,885	24,701
2	10412921	11/23/2018	Yellow Cab	ATLANTA GA	42.55	792.05	597.4020	29290	Card	Male	28	10813	814,885	24,701
3	10000012	01/06/2016	Pink Cab	ATLANTA GA	28.62	358.52	334.8540	27703	Card	Male	27	9237	814,885	24,701
4	10320494	04/21/2018	Yellow Cab	ATLANTA GA	36.38	721.10	467.1192	27703	Card	Male	27	9237	814,885	24,701
5	10324737	05/04/2018	Yellow Cab	ATLANTA GA	6.18	138.40	87.5088	27703	Cash	Male	27	9237	814,885	24,701
6	10395626	10/27/2018	Pink Cab	ATLANTA GA	13.39	167.03	141.9340	27703	Card	Male	27	9237	814,885	24,701
7	10000013	01/02/2016	Pink Cab	ATLANTA GA	9.04	125.20	97.6320	28712	Cash	Male	53	11242	814,885	24,701
8	10079404	09/21/2016	Yellow Cab	ATLANTA GA	39.60	704.30	494.2080	28712	Card	Male	53	11242	814,885	24,701
9	10186994	06/23/2017	Yellow Cab	ATLANTA GA	18.19	365.63	246.6564	28712	Card	Male	53	11242	814,885	24,701

MasterData dataset has 359392 entries ,

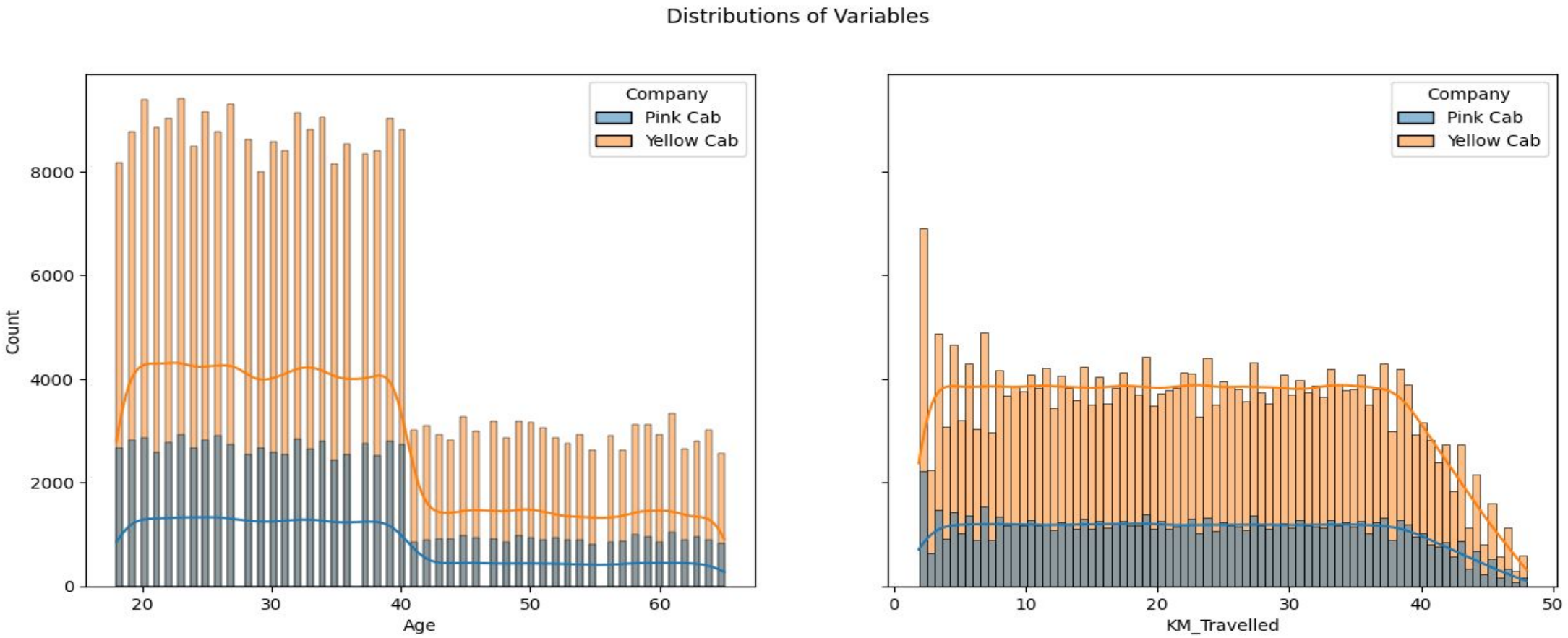
14 features and 0 missing values

Timeframe of the data: 2016-01-31 to 2018-12-31

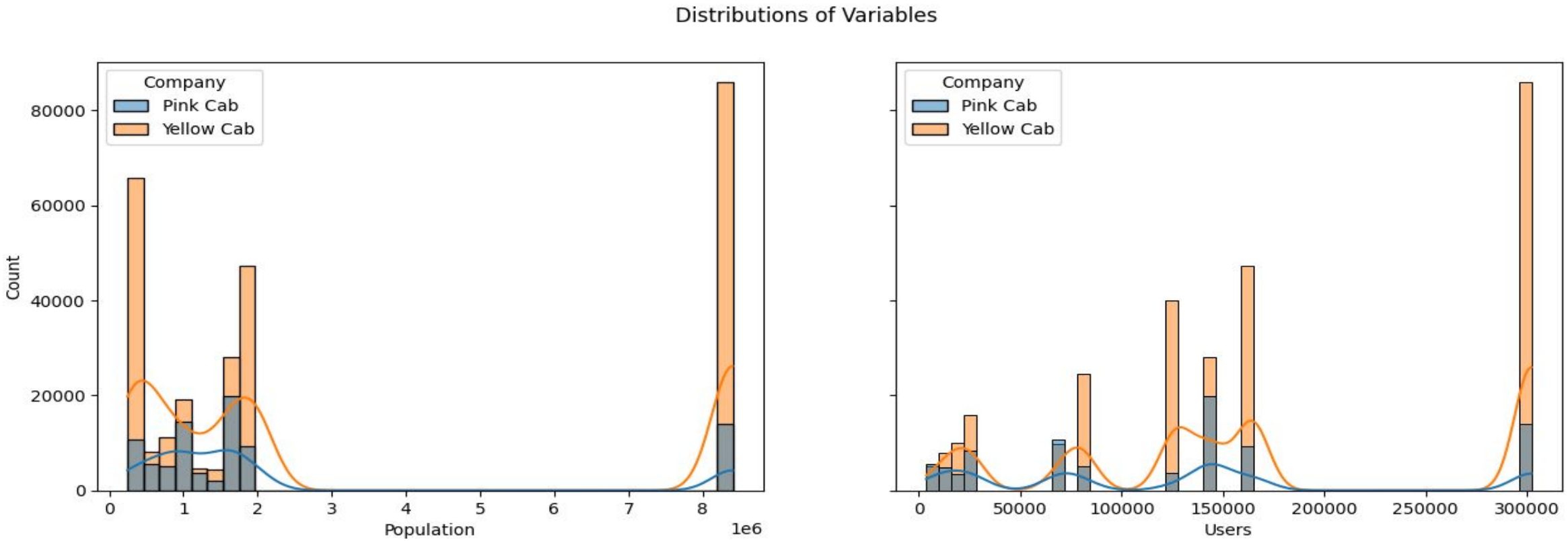
The KDE curves and distribution plots of Cost_of_Trip,Price_Charged,Income_(USD/Month) with respect to Cab Firms drawn below.



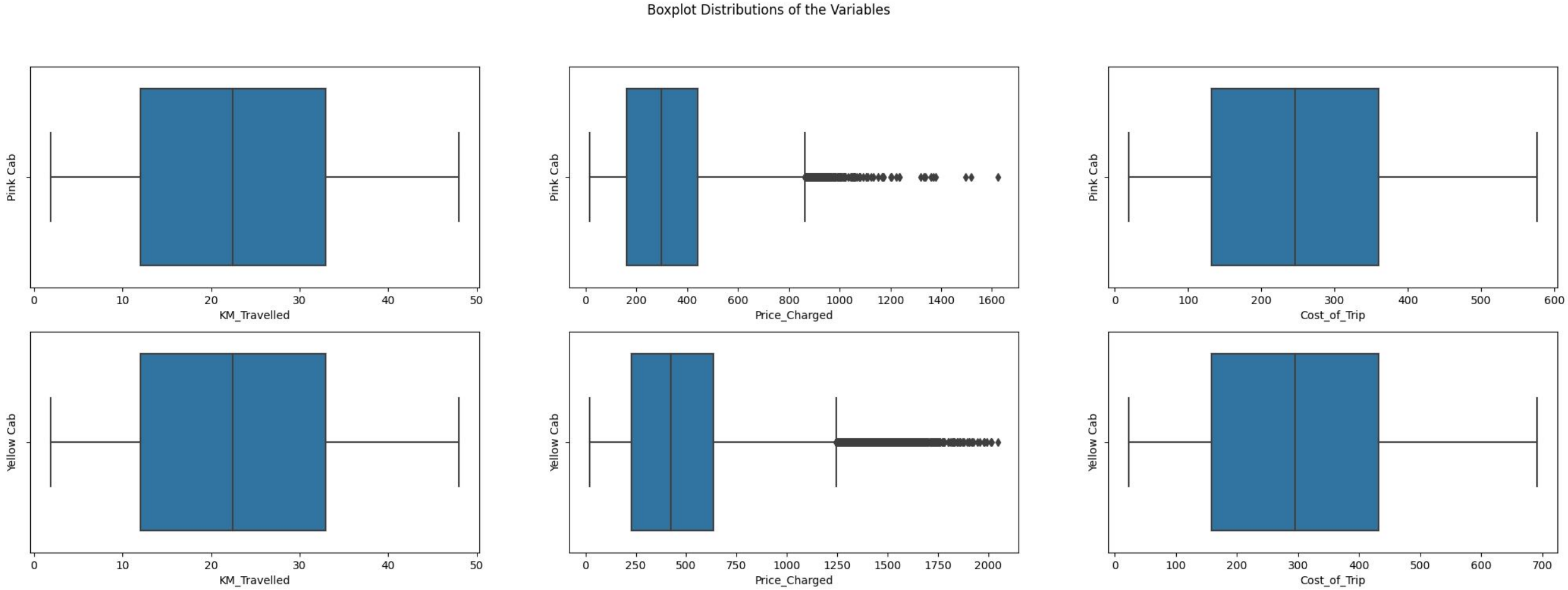
The KDE curves and distribution plots of Age & Km_travelled with respect to Cab Firms drawn below.



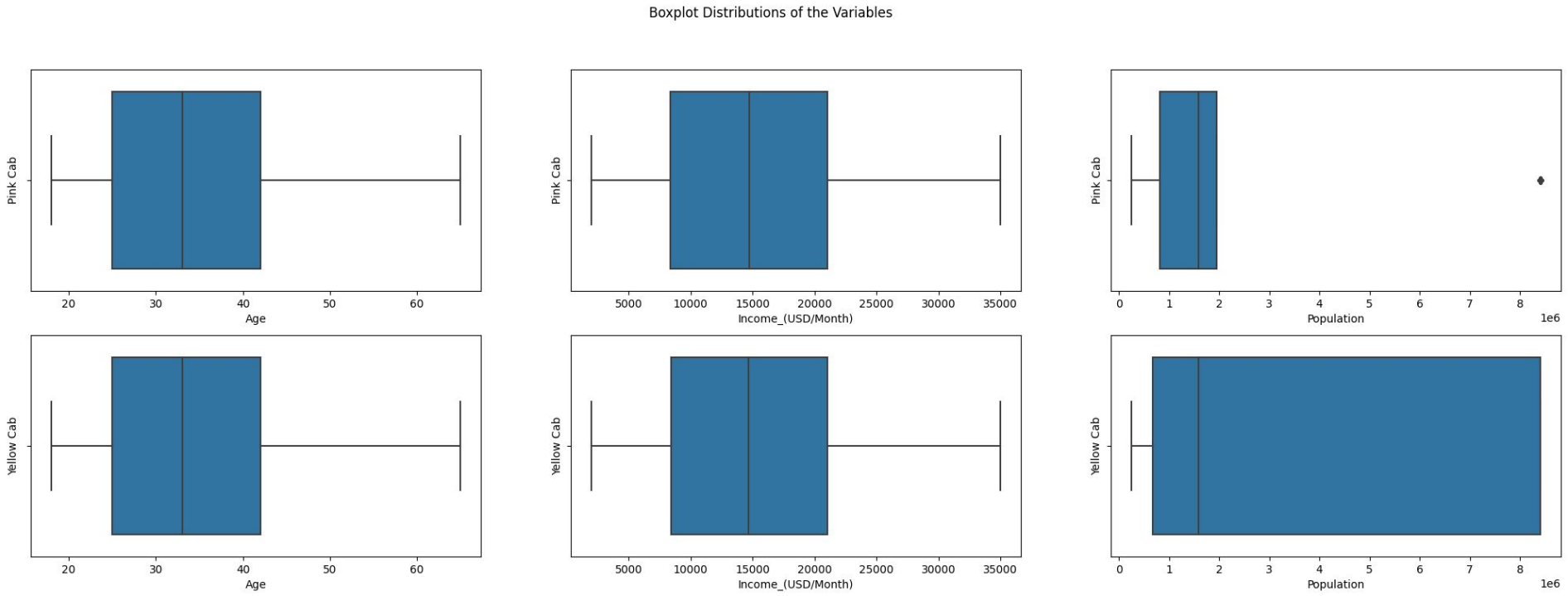
The KDE curves and distribution plots of Population & Users with respect to Cab Firms drawn below.



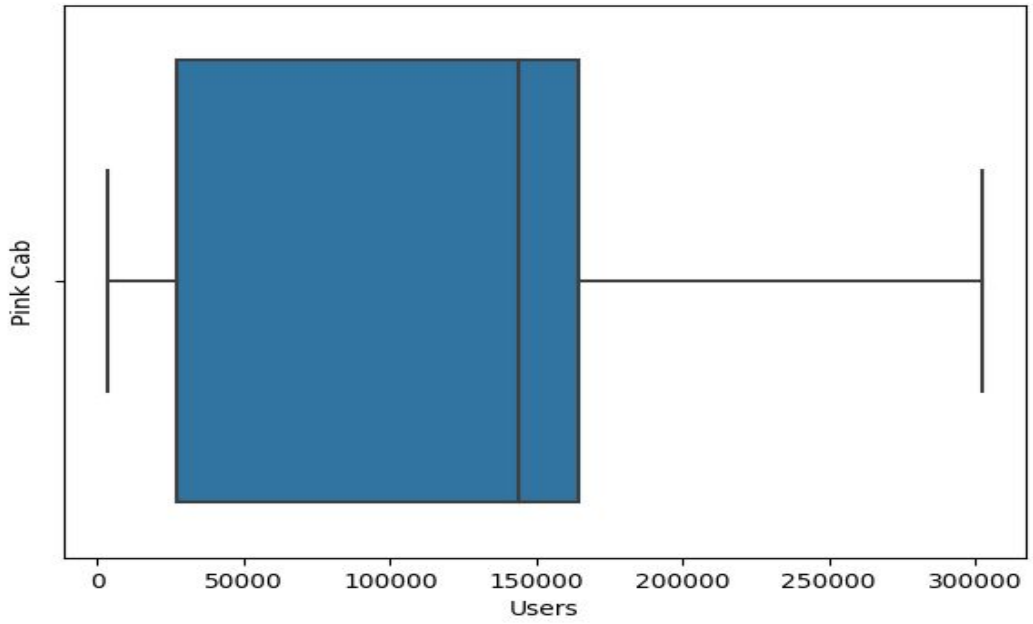
Boxplot distributions of Km_Travelled,Price_Charged,Cost_of_trip with respect to Cab Firms were drawn below

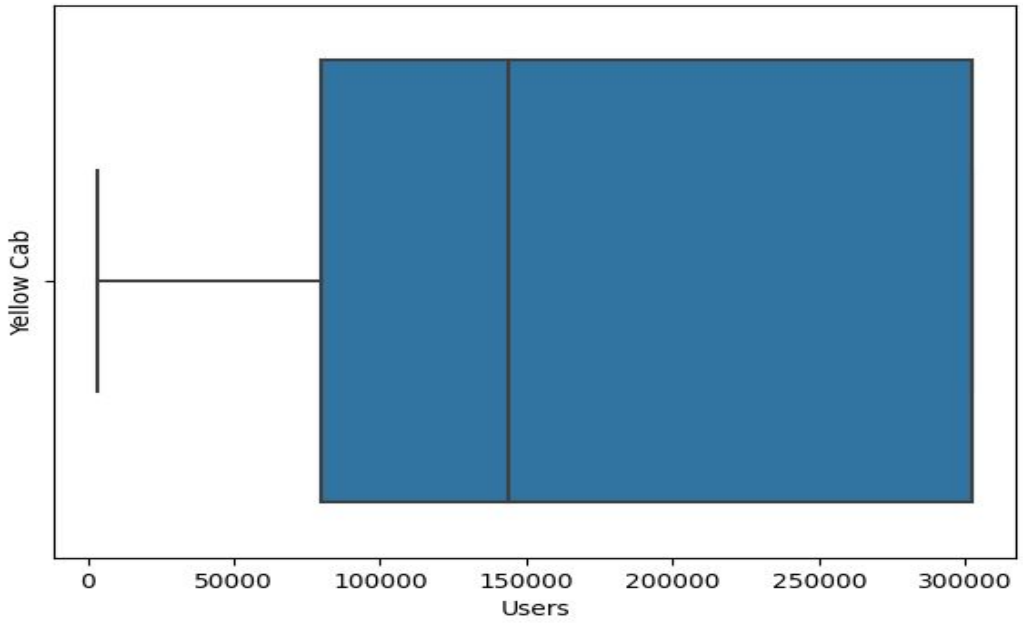


Boxplot distributions of Age,Income_(USD/Month),Population with respect to Cab Firms were drawn below



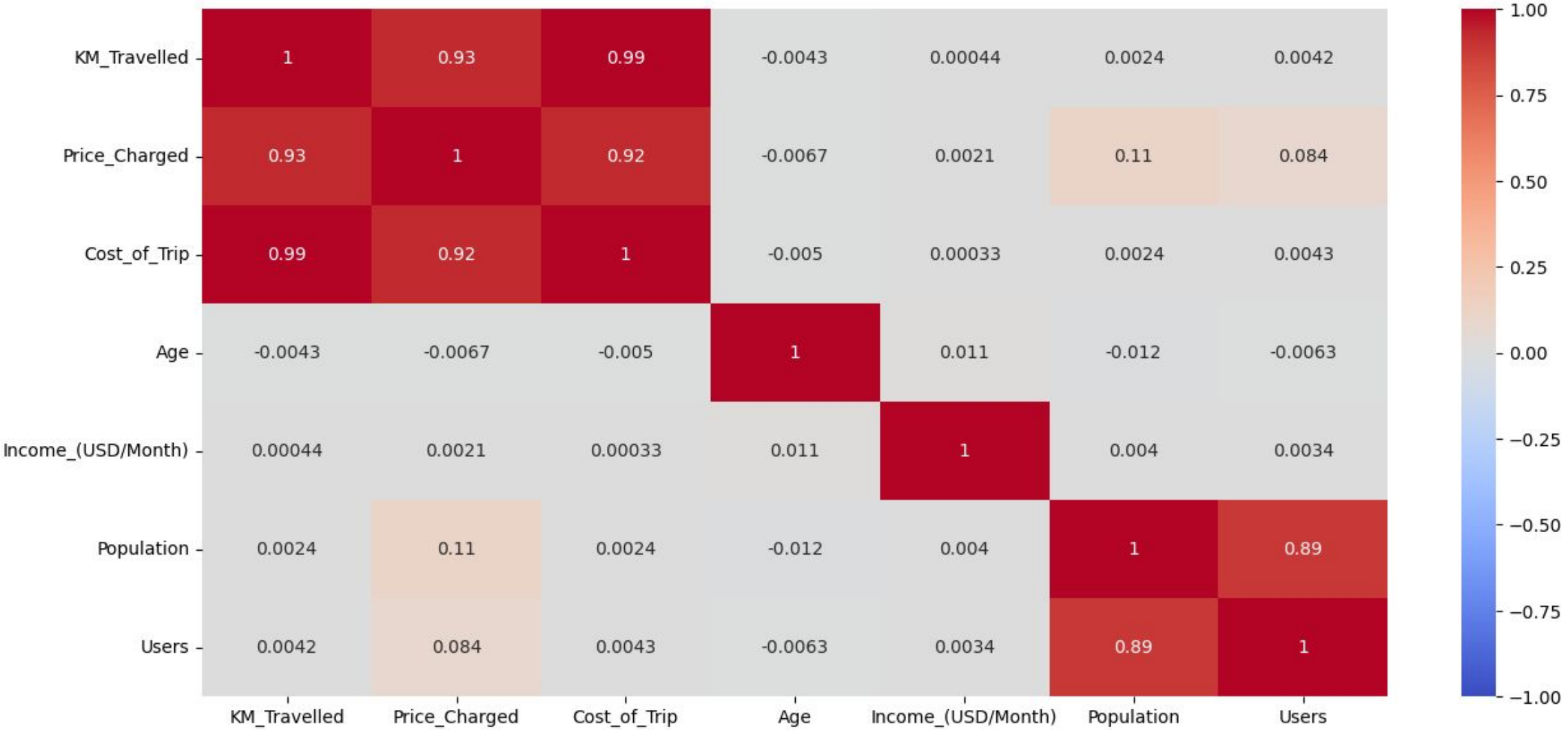
Boxplot distribution of Users with respect to both the firm



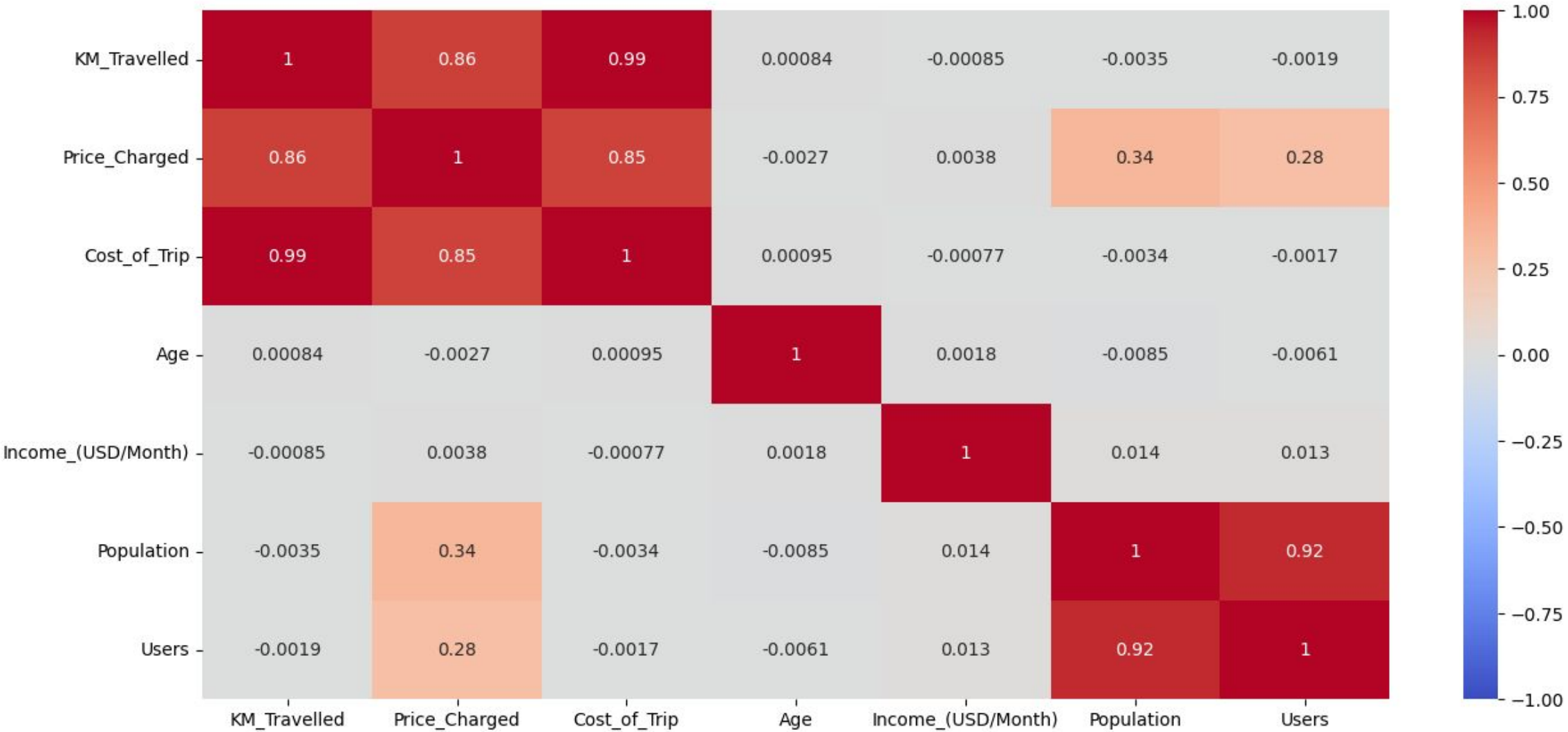


- **Users for Yellow Cab is higher than the Pink Cab**

Correlation of MasterData features by heatmap for Pink Cab Firm.

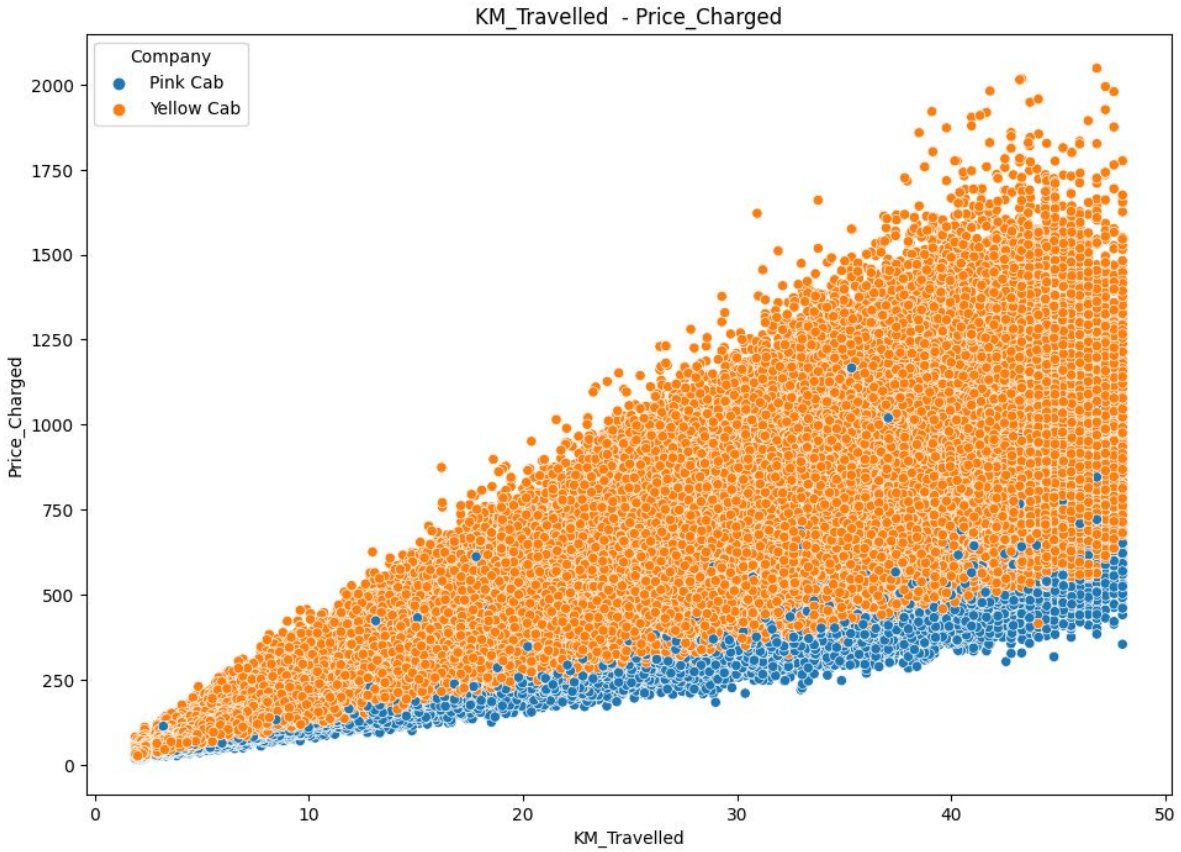


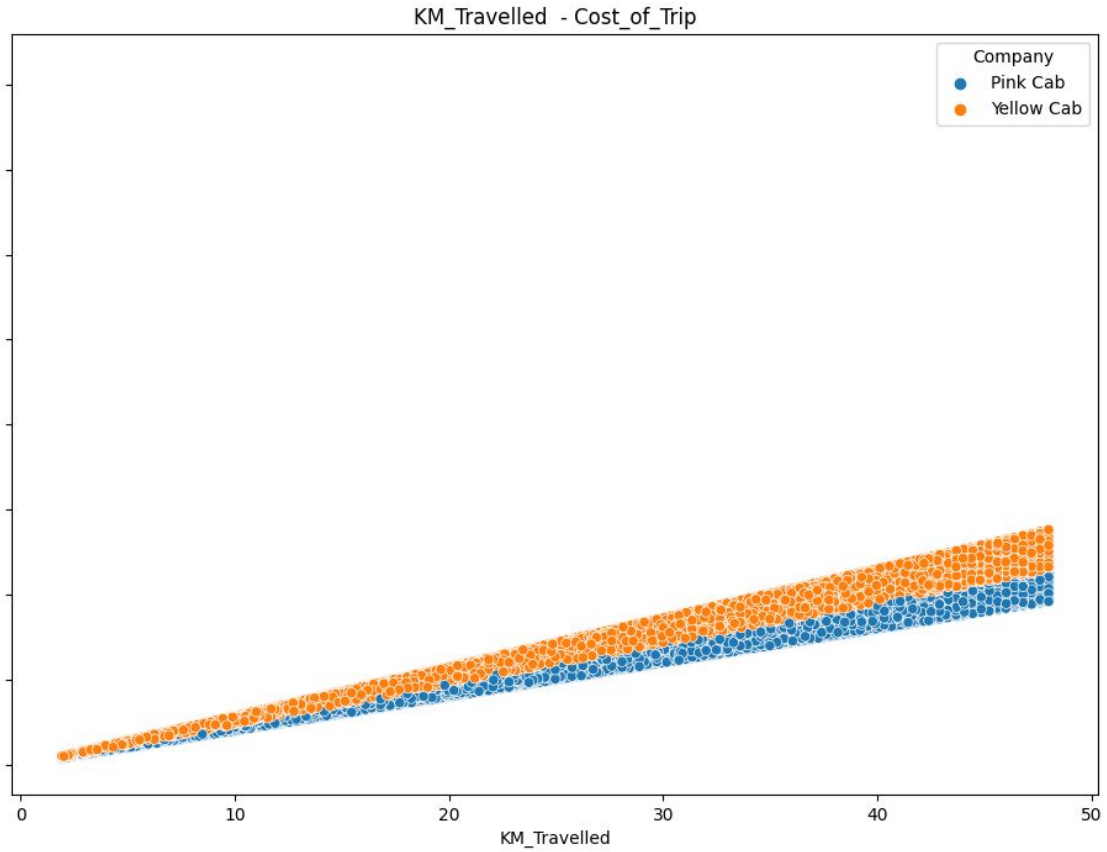
Correlation of MasterData features by heatmap for Yellow Cab Firm.



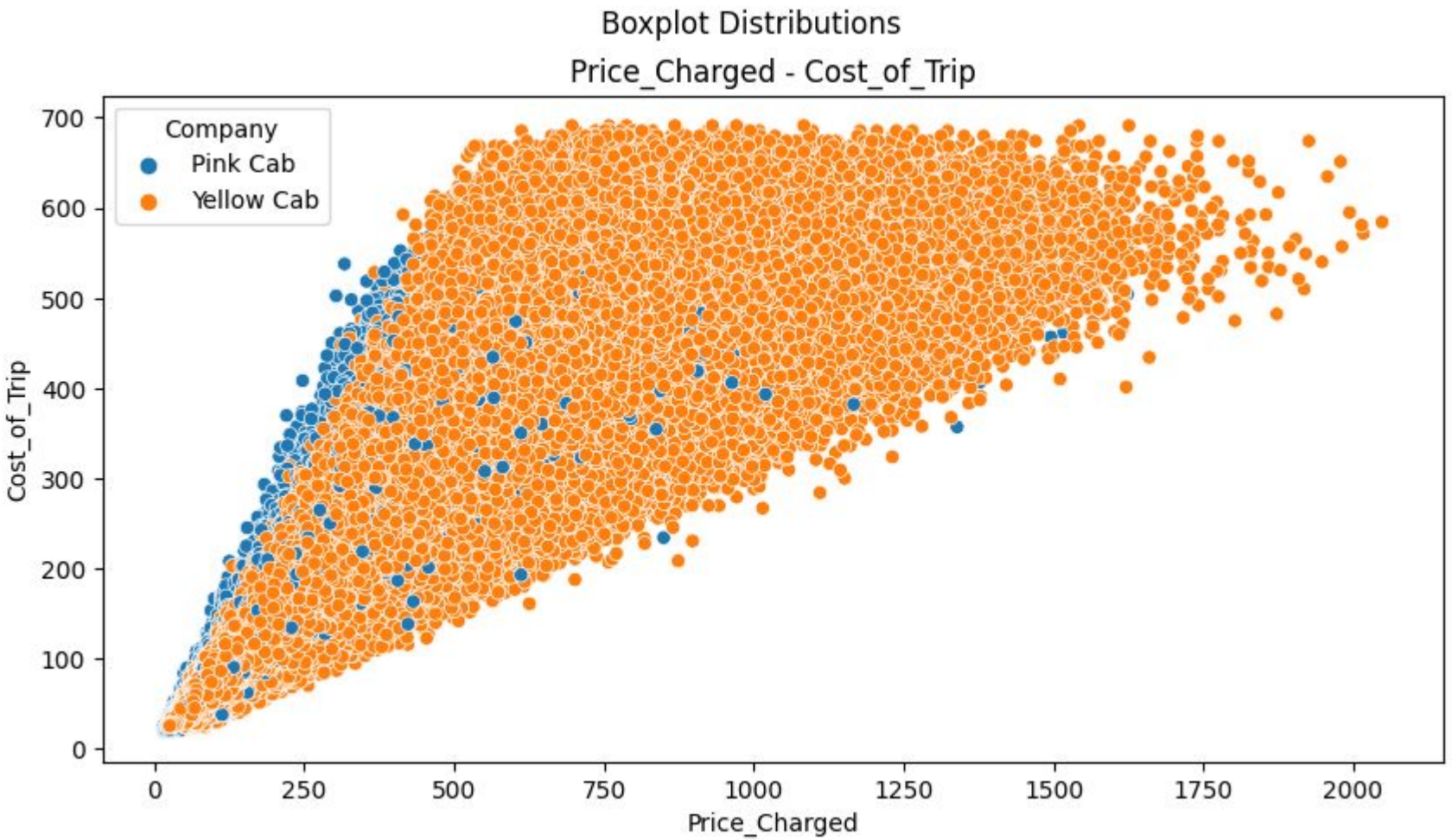
Scatter plots of MasterData features to see correlations between KM_travelled - Cost_of_Trip & KM_travelled - Price_Charged

Boxplot Distributions

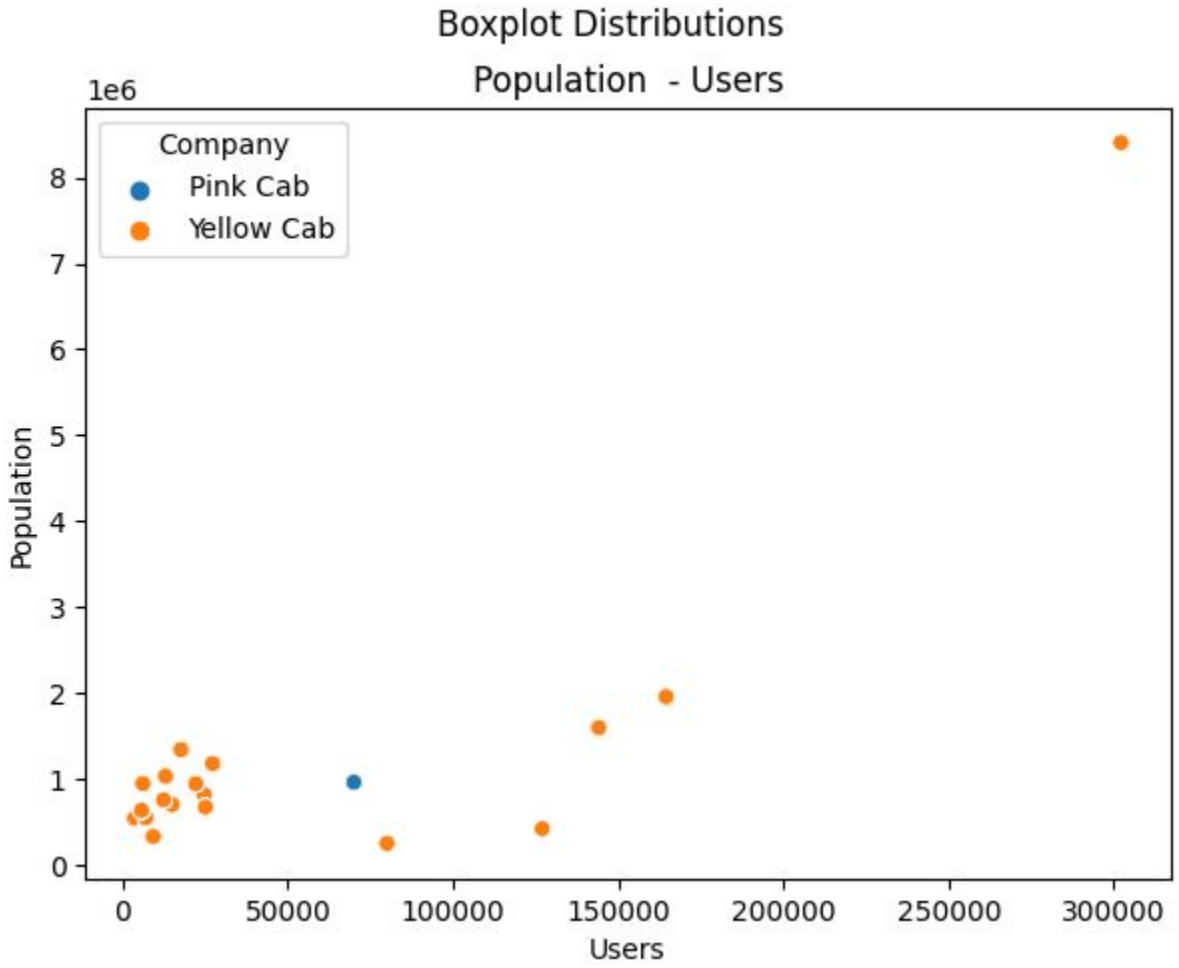




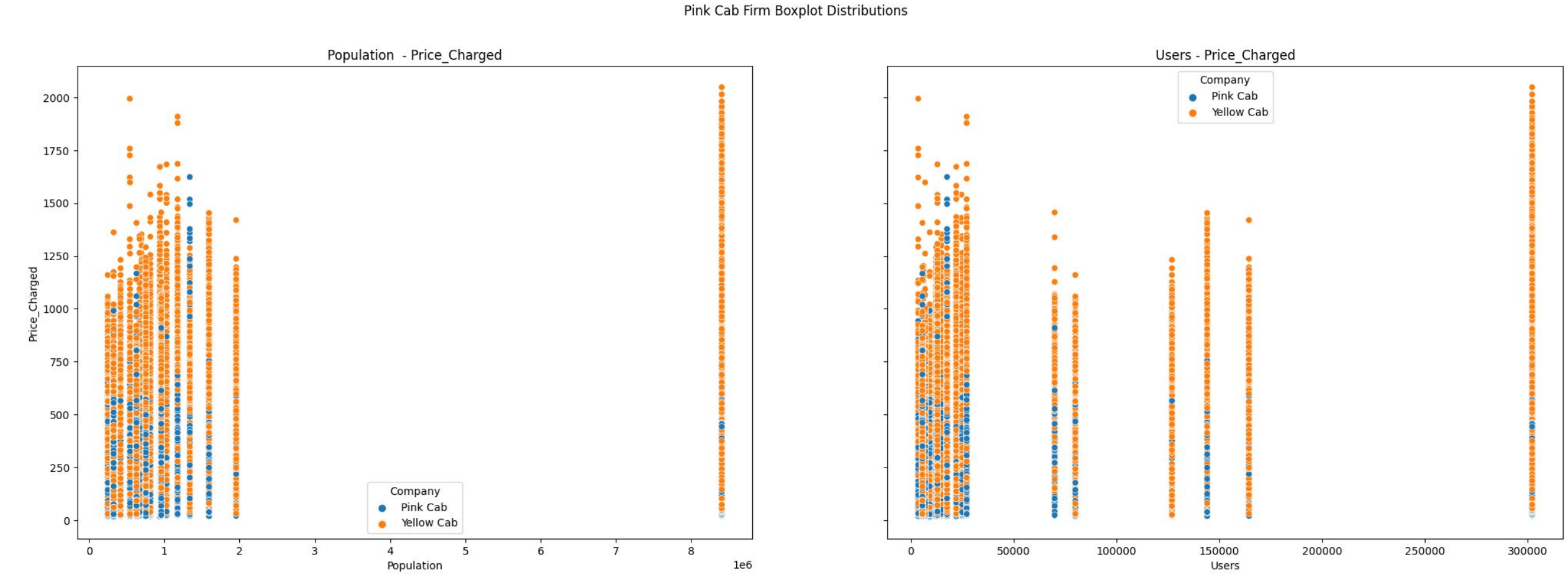
Here visualized scatter plots of MasterData features to see correlations between Price_Charged-Cost_of_Trip if exists



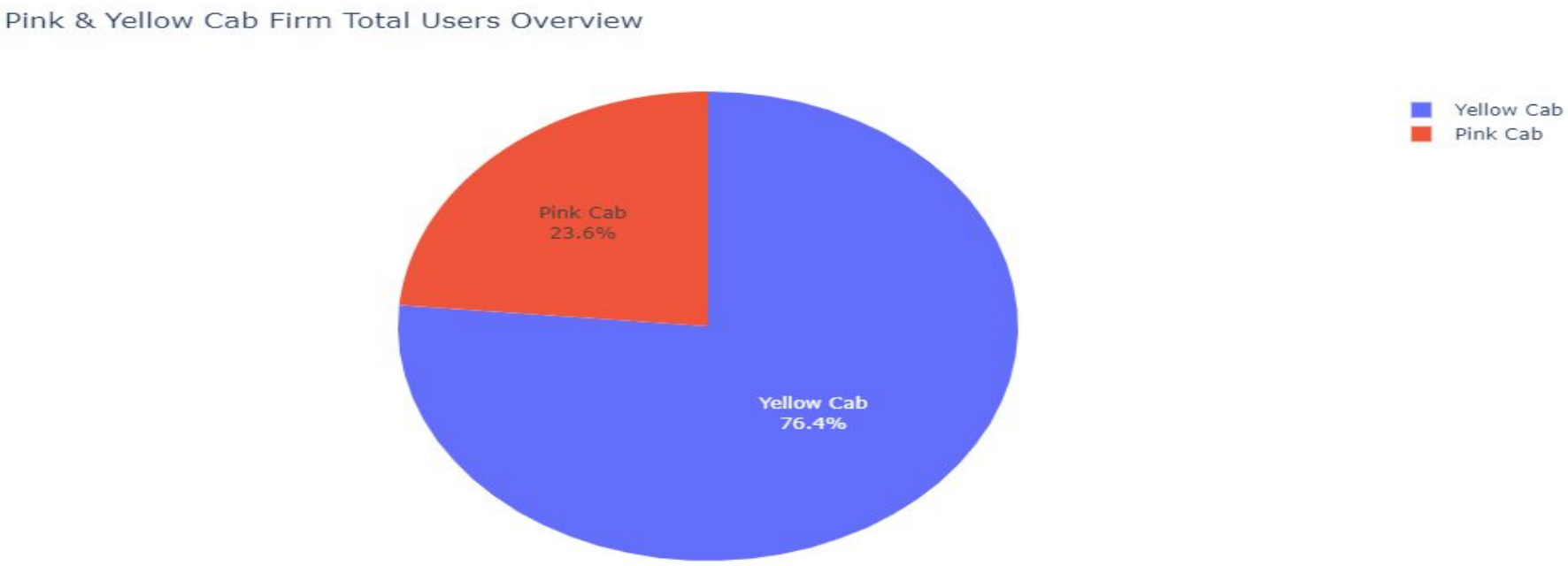
correlations between Population-User if exists



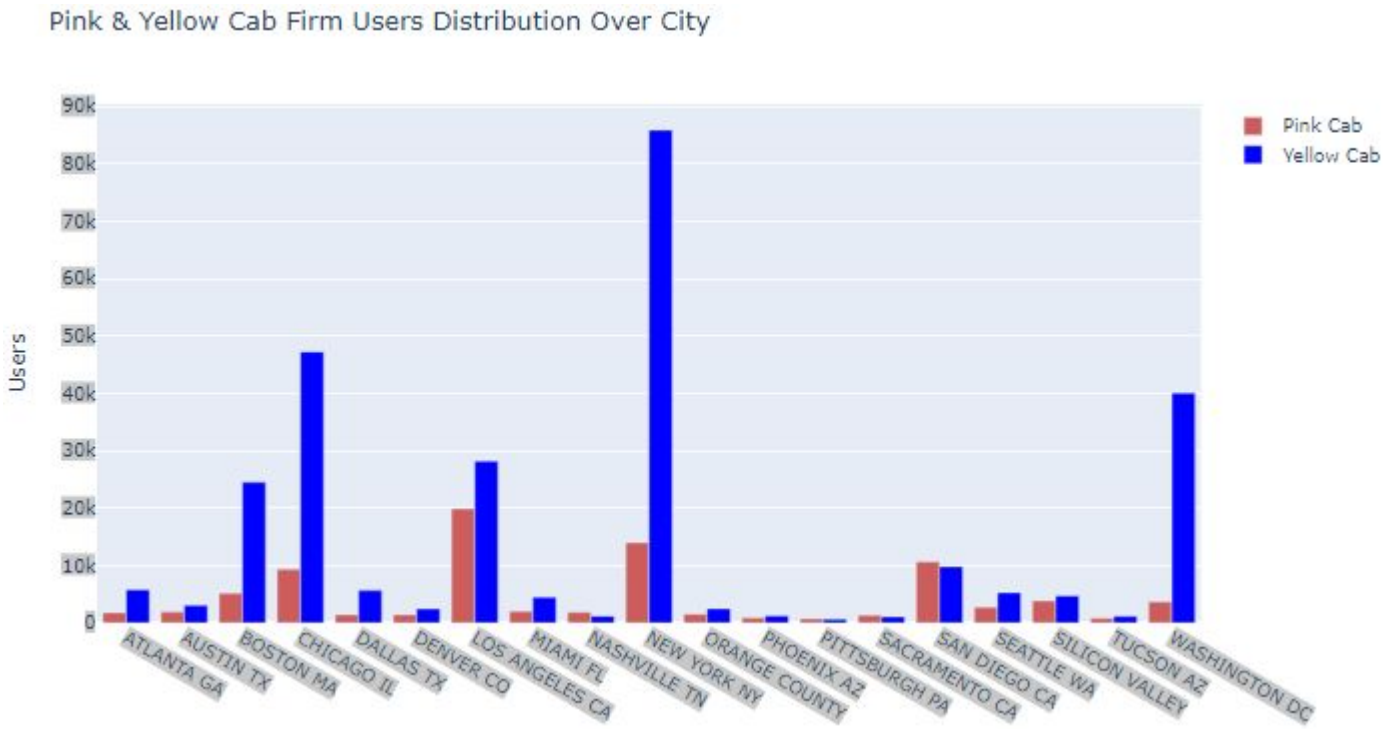
Here visualized scatter plots of MasterData features to see correlations between Population - Price_Charged & Users - Price_Charged if exists



Inferential Data Analysis

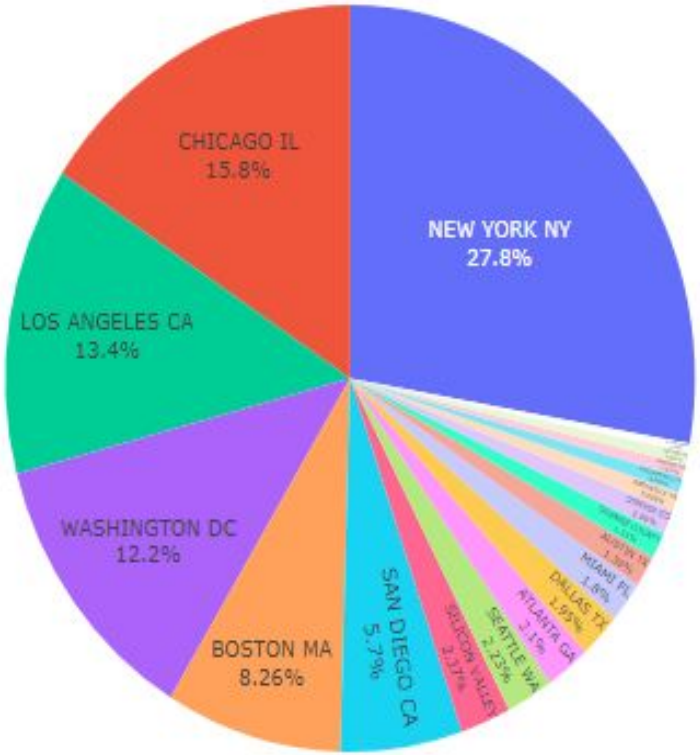


As seen from this Pie Chart; The total number of users of Yellow Cab is approximately 3 times that of Pink Cab.



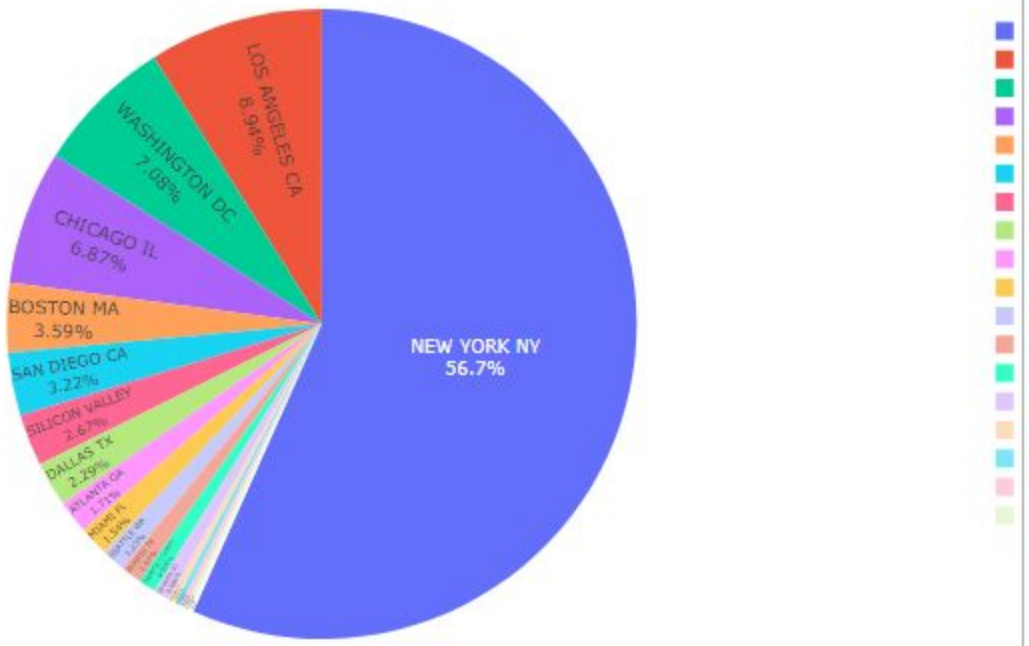
As seen from this Bar Chart; For the Yellow Cab Company, the highest number of users on a city basis are in New York, Washington and Chicago, while for the Pink Cab Company, the most are in Los Angeles, New York and San Diego.

Total Users Overview by Cities



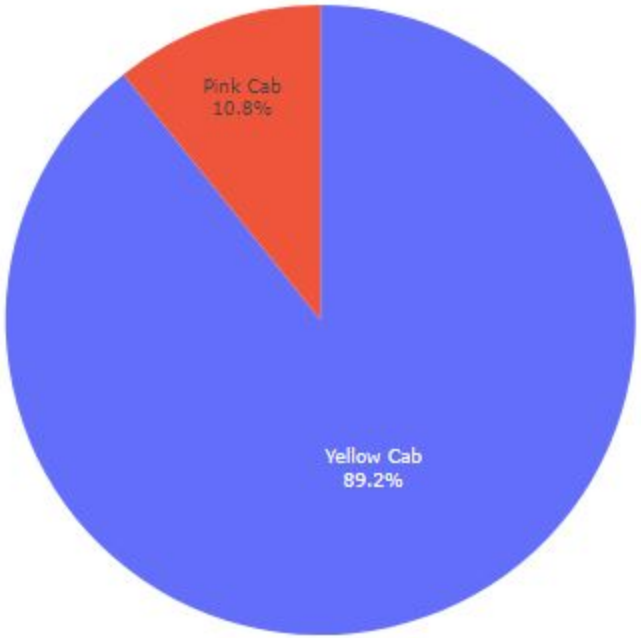
As seen from this Pie Chart; On the basis of cities, the highest number of total users are in New York, Chicago, Los Angeles, Washington and Boston.

Total Market Profit Share by Cities



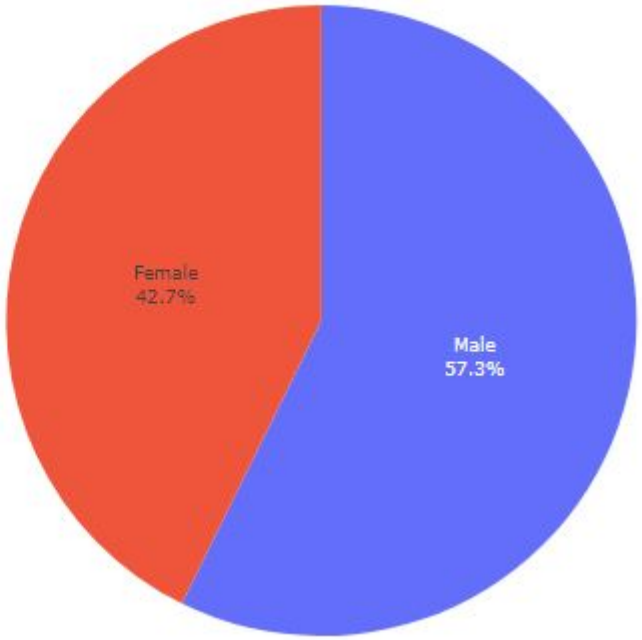
As seen from this Pie Chart; More than half of the total market profit share on the basis of cities belongs to New York.

Total Market Profit Share by Cab Firms

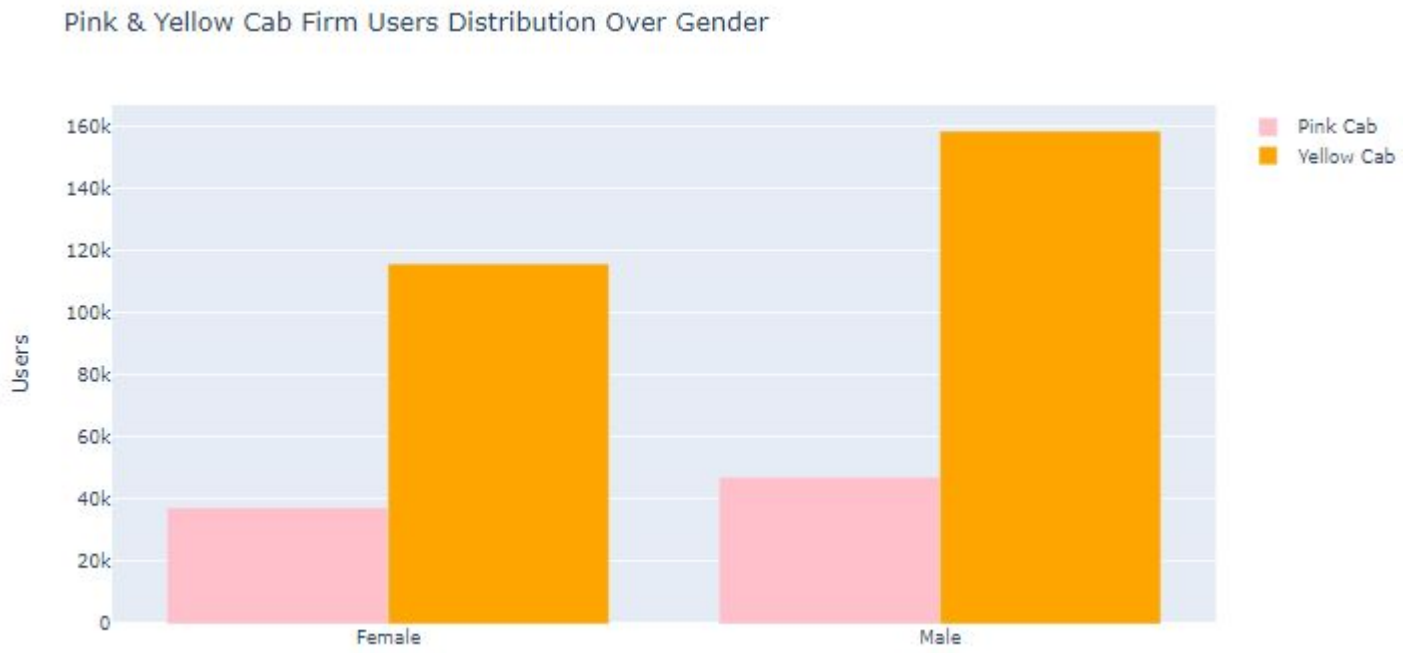


As seen from this Pie Chart; The total market profit share of Yellow Cab is approximately 9 times that of Pink Cab.

Total Users Overview by Gender

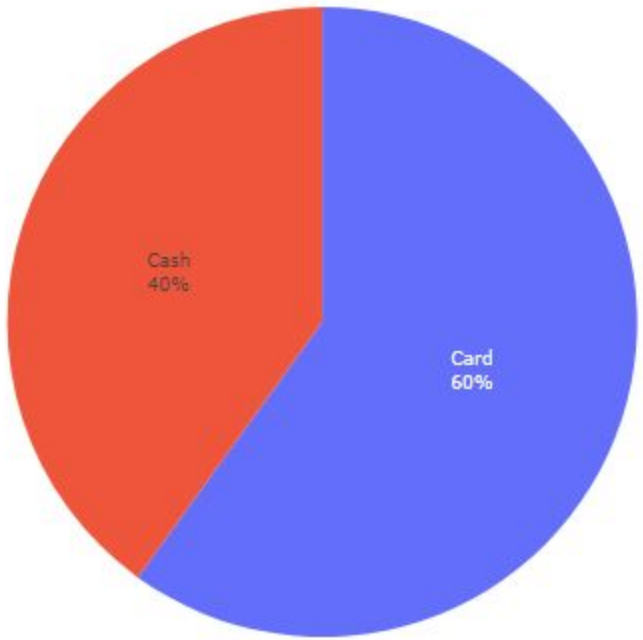


As seen from this Pie Chart; In the distribution of users by gender, there is an approximate 3 to 2 ratio for men and women



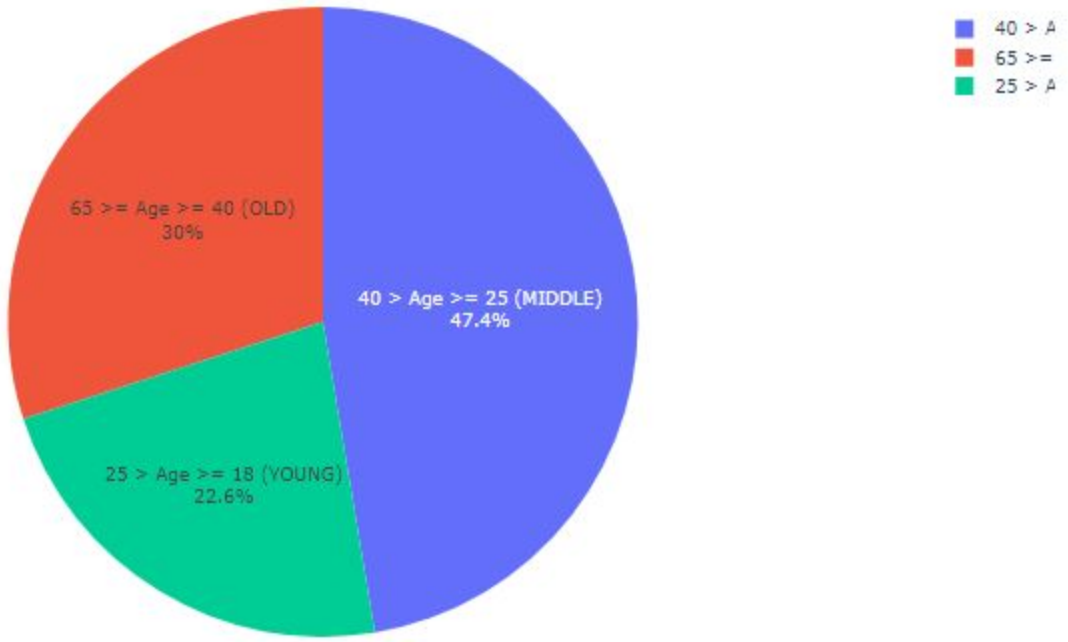
As seen from this Bar Chart; When the distribution of users by gender is analyzed on a company basis, while the male-female ratio is 57.6% - 42.4% in Yellow Cab Company, the male-female ratio is 55.9% - 44.1% in Pink Cab Company.

Total Users Overview by Payment Method

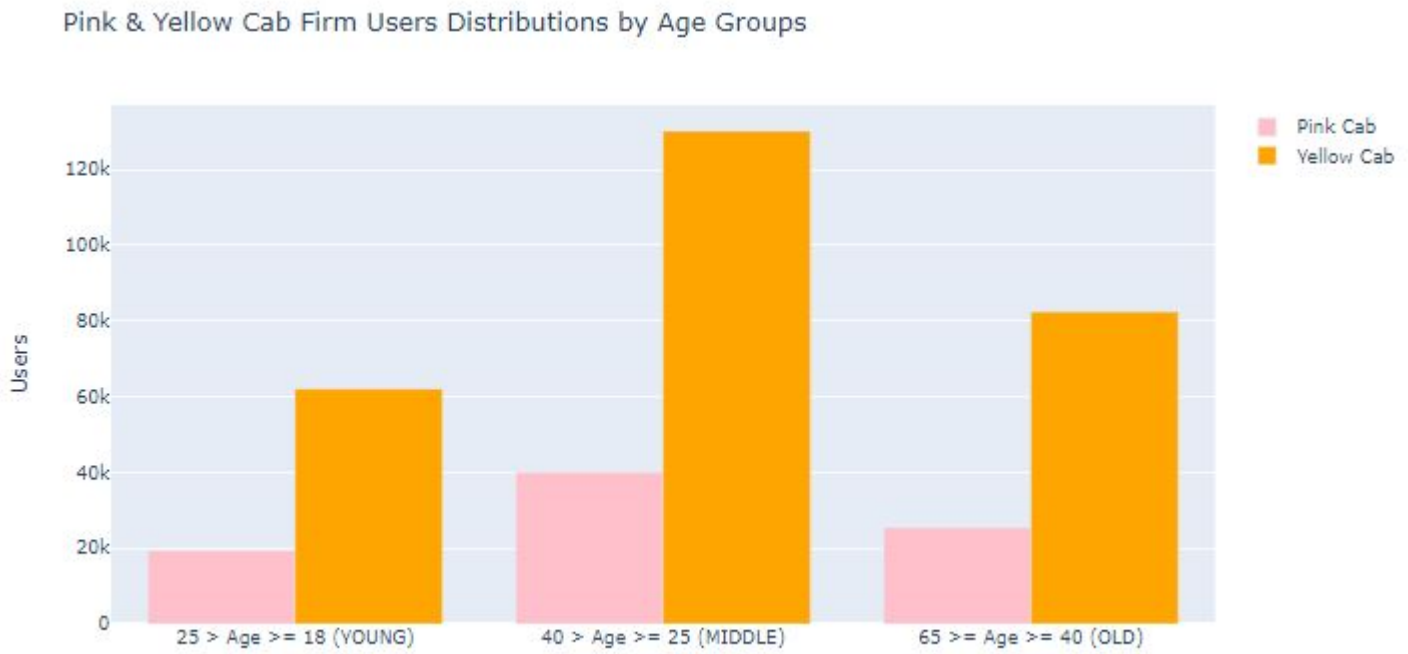


As seen from this Pie Chart; Considering the payment preferences of all users, the credit card- cash payment ratio is 3 to 2.

Total Users Overview by Age Groups

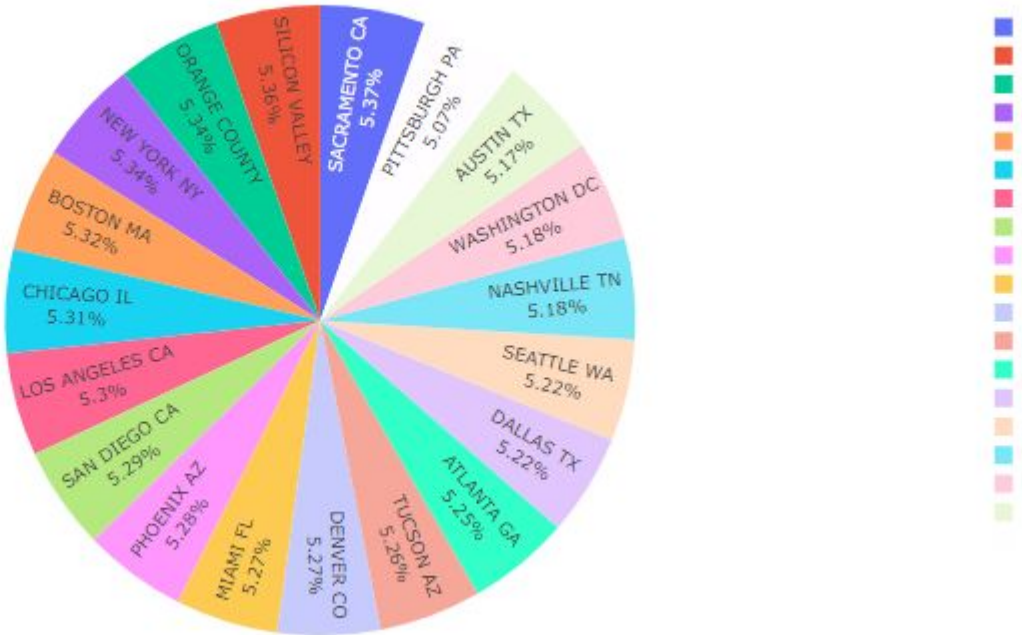


As seen from this Pie Chart; Looking at the age distribution of all users, it is seen that approximately half of them are between the ages of 18-25.

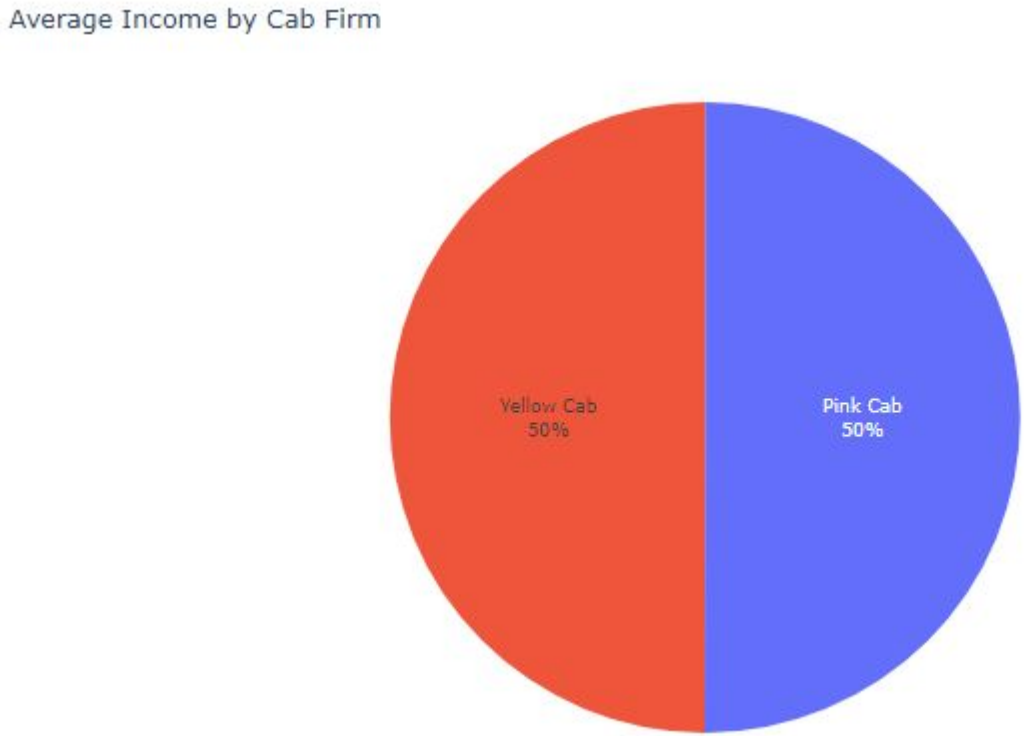


As seen from this Bar Chart; Looking at the age distribution of all users in the basis of companies, it is seen that both have the same percentage distribution for every age group

Average Income by Cities

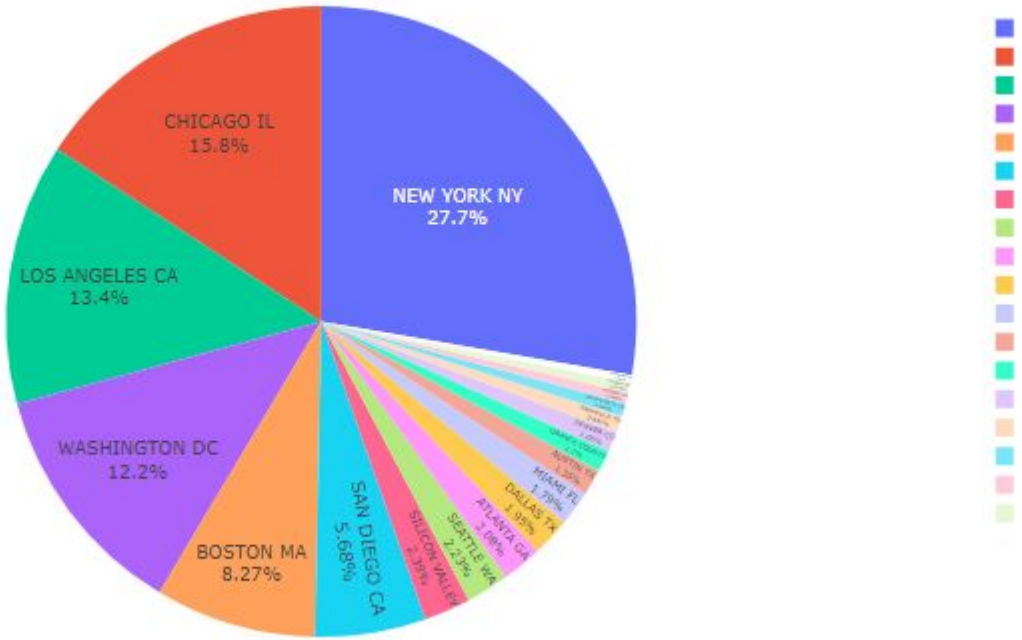


As seen from this Pie Chart; The average income of all users by city is approximately equal.

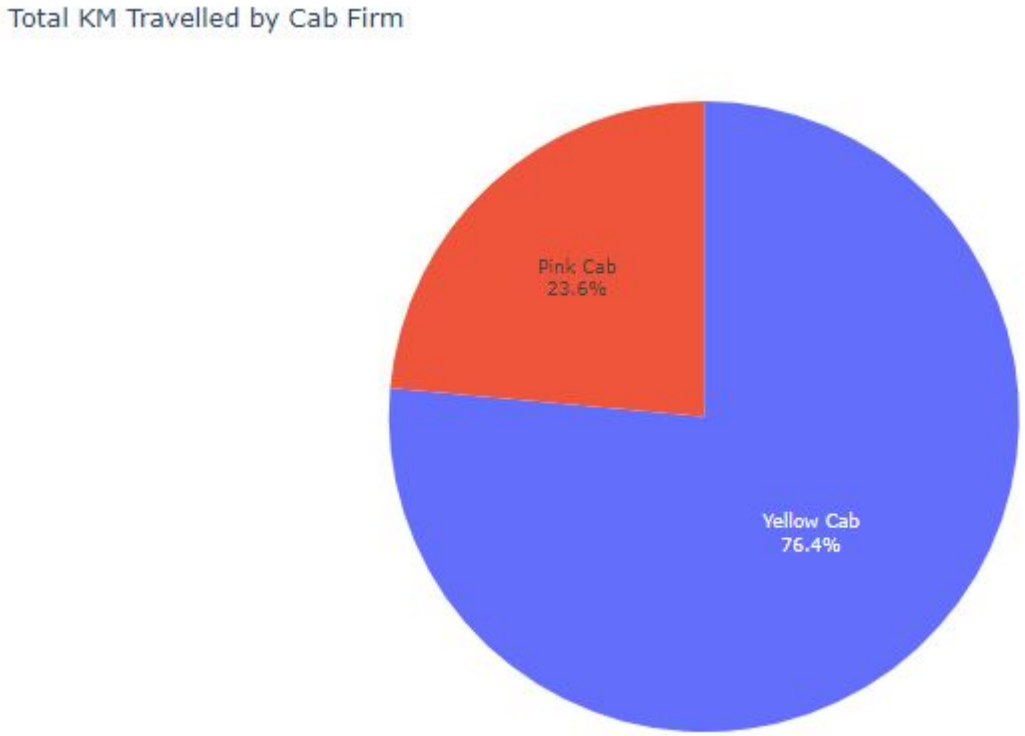


As seen from this Pie Chart; The average income of all users by companies is approximately equal.

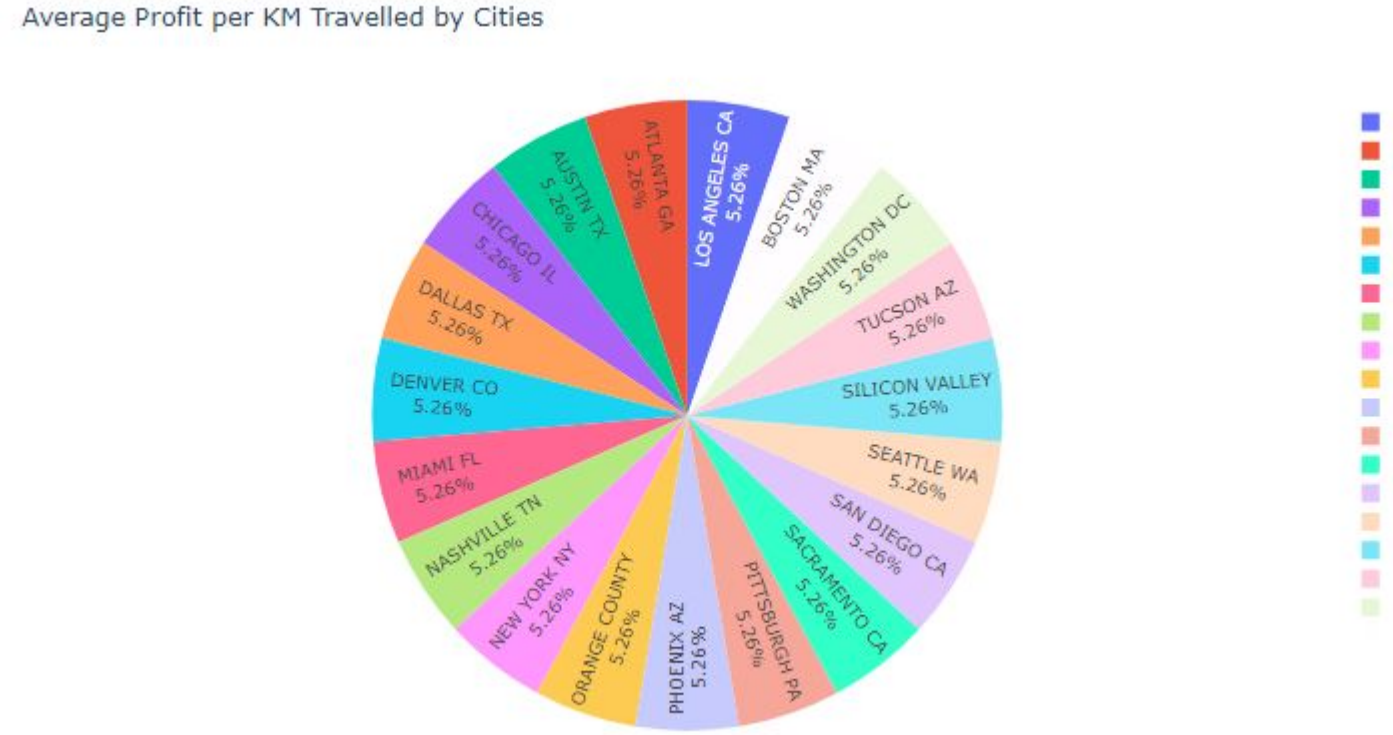
Total KM Travelled by Cities



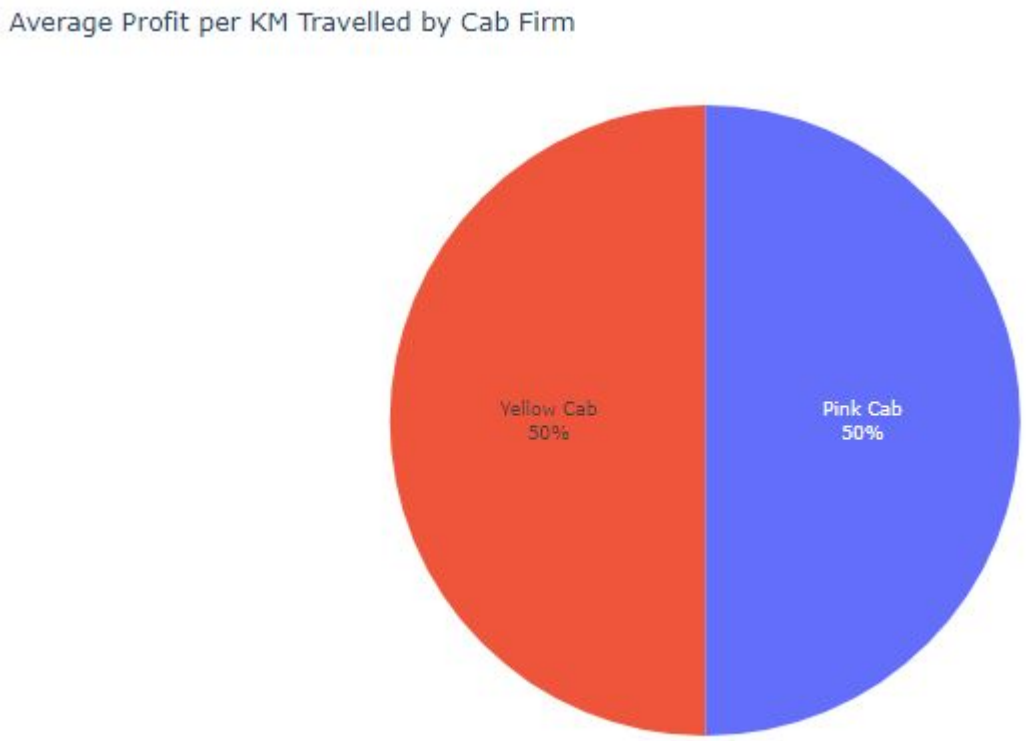
As seen from this Pie Chart; On the basis of cities, the most travelled in KM are New York, Chicago, Los Angeles, Washington and Boston.



As seen from this Pie Chart; The total travelled in KM for Yellow Cab is approximately 3 times that of Pink Cab.



As seen from this Pie Chart; The average profit per travelled in KM by cities is approximately equal.



As seen from this Pie Chart; The average profit per travelled in KM by companies is approximately equal.

As seen from this Pie Chart; In 2016 , The total market profit share of Yellow Cab is approximately 8.15 times that of Pink Cab. In 2017 , The total market profit share of Yellow Cab is approximately 8.16 times that of Pink Cab. In 2018 , The total market profit share of Yellow Cab is approximately 8.66 times that of Pink Cab

Conclusion

when we consider for both Cab Firms in terms of total market profit share , total user share , yearly market profit share , total travelled in KM by Users ; we will recommend Yellow Cab Firm for investment.

Hypothesis 1: Is there any difference in profit regarding Gender

H0 : There is no difference regarding Gender in both cab companies. H1 : There is difference regarding Gender in both cab companies.

Pink Cab

In [64]:
```
a = MasterData[(MasterData.Gender=='Male')&(MasterData.Company=='Pink Cab')].groupby('Transaction_ID').Profit_of_Trip.mean()
b = MasterData[(MasterData.Gender=='Female')&(MasterData.Company=='Pink Cab')].groupby('Transaction_ID').Profit_of_Trip.mean()
print(a.shape[0],b.shape[0])

_, 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 regarding gender for Pink Cab')
else:
    print('We accept null hypothesis (H0) that there is no difference regarding gender for Pink Cab')
```

```
47231 37480
P value is  0.115153059004258
We accept null hypothesis (H0) that there is no difference regarding gender for Pink Cab
```

Yellow Cab

In [65]:
```
a = MasterData[(MasterData.Gender=='Male')&(MasterData.Company=='Yellow Cab')].groupby('Transaction_ID').Profit_of_Trip.mean()
b = MasterData[(MasterData.Gender=='Female')&(MasterData.Company=='Yellow Cab')].groupby('Transaction_ID').Profit_of_Trip.mean()
print(a.shape[0],b.shape[0])

_, 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 regarding gender for Yellow Cab')
else:
    print('We accept null hypothesis (H0) that there is no difference regarding gender for Yellow Cab')
```

```
158681 116000
P value is  6.060473042494056e-25
We accept alternative hypothesis (H1) that there is a difference regarding gender for Yellow Cab
```

There is no difference regarding Gender in both cab companies.

Hypothesis 2: Is there any difference in Profit regarding Age

H0 : There is no difference regarding Age in both cab companies. H1 : There is difference regarding Age in both cab companies.

Pink Cab

In [66]:
```
a =  MasterData[( MasterData.Age <= 60)&(MasterData.Company=='Pink Cab')].groupby('Transaction_ID').Profit_of_Trip.mean()
b =  MasterData[( MasterData.Age >= 60)&( MasterData.Company=='Pink Cab')].groupby('Transaction_ID').Profit_of_Trip.mean()
print(a.shape[0],b.shape[0])

_, 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 regarding age for Pink Cab')
else:
    print('We accept null hypothesis (H0) that there is no difference regarding age for Pink Cab')
```

```
80125 5429
P value is  0.4816748536155634
We accept null hypothesis (H0) that there is no difference regarding age for Pink Cab
```

Yellow Cab

In [67]:
```
a =  MasterData[( MasterData.Age <= 60)&(MasterData.Company=='Yellow Cab')].groupby('Transaction_ID').Profit_of_Trip.mean()
b =  MasterData[( MasterData.Age >= 60)&( MasterData.Company=='Yellow Cab')].groupby('Transaction_ID').Profit_of_Trip.mean()
print(a.shape[0],b.shape[0])

_, 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 regarding age for Yellow Cab')
else:
    print('We accept null hypothesis (H0) that there is no difference regarding age for Yellow Cab')
```

```
260356 17257
P value is  6.328485471267631e-05
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.

Hypothesis 3: Is there any difference in Profit regarding Payment mode

H0 : There is no difference regarding Payment_Mode in both cab companies. H1 : There is difference regarding Payment_Mode in both cab companies..

Pink Cab

In [68]:
```
a =MasterData[(MasterData['Payment_Mode']=='Cash')&(MasterData.Company=='Pink Cab')].groupby('Transaction_ID').Profit_of_Trip.mean(
b = MasterData[(MasterData['Payment_Mode']=='Card')&(MasterData.Company=='Pink Cab')].groupby('Transaction_ID').Profit_of_Trip.mean

_, 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 Pink Cab')
else:
    print('We accept null hypothesis (H0) that there is no difference in payment mode for Pink Cab')
```

```
P value is  0.7900465828793286
We accept null hypothesis (H0) that there is no difference in payment mode for Pink Cab
```

Yellow Cab

In [69]:
```
a =MasterData[(MasterData['Payment_Mode']=='Cash')&(MasterData.Company=='Yellow Cab')].groupby('Transaction_ID').Profit_of_Trip.mea
b = MasterData[(MasterData['Payment_Mode']=='Card')&(MasterData.Company=='Yellow Cab')].groupby('Transaction_ID').Profit_of_Trip.me

_, 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  0.29330606382987284
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.

Thank You