Data Glacier

Your Deep Learning Partner

G2M Case Study

Virtual Internship

Submitted by:Han-Fu Lin

14-OCT-2022

Background –G2M(cab industry) case study

- 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.

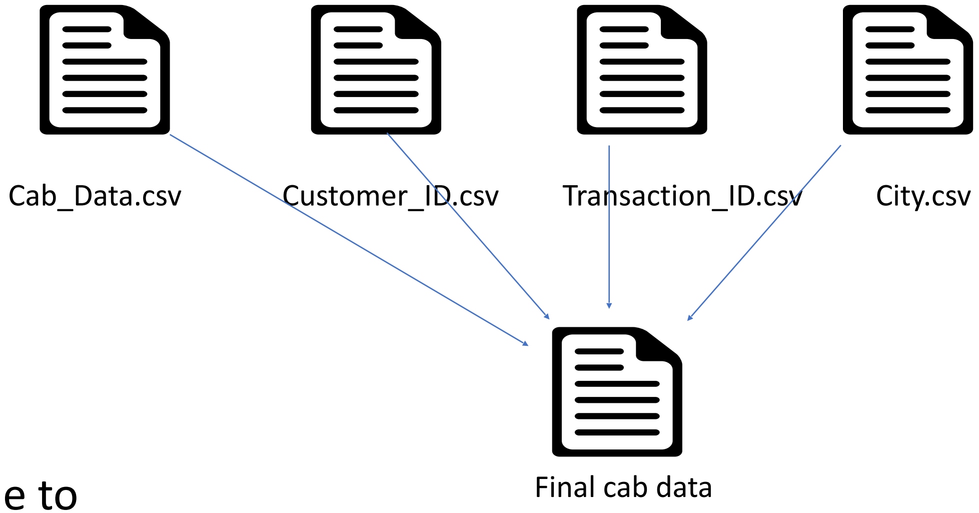
- Objective : Provide actionable suggestion to help XYZ firm in identifying the right company for making investment. Based upon 4 datasets of the two firm

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 Exploration

- Total of 17 feature
- Timeframe of the data: 2016-01-31 to 2018-12-31
- Total data points :355,032



Assumptions:

- Outliers are present in Price_Charged feature but due to unavailability of trip duration details ,we are not treating this as outlier.

- Profit of rides are calculated keeping other factors constant and only Price_Charged and Cost_of_Trip features used to calculate profit.

- Users feature of city dataset is treated as number of cab users in the city. we have assumed that this can be other cab users as well(including Yellow and Pink cab)

DATA INFORMATION

Data Information Explanation

- Customer_ID.csv – this is a mapping table that contains a unique identifier that links the customer’s demographic details
- Transaction_ID.csv – this is a mapping table that contains transaction to customer mapping and payment mode
- City.csv – this file contains a list of US cities, their population, and the number of cab users
- Cab_Data.csv – this file includes details of transactions for 2 cab companies

Problem identification

- Data interpretation and Visualization

- Finding Most popular Cab company

- Compare the price advantages and disadvantages for users

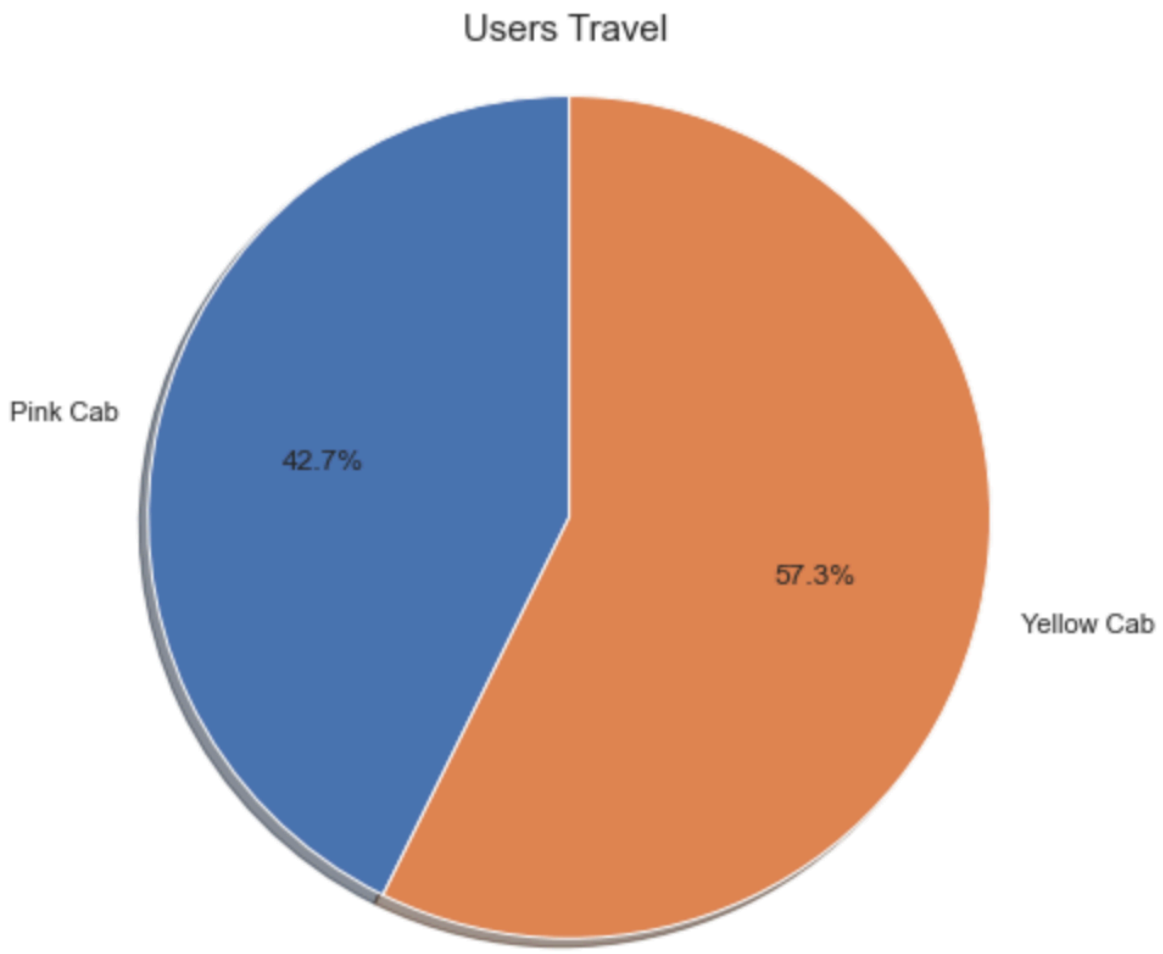
- Understanding the customers and examine profit

- Set Multiple Hypothesis and Investigate

EDA

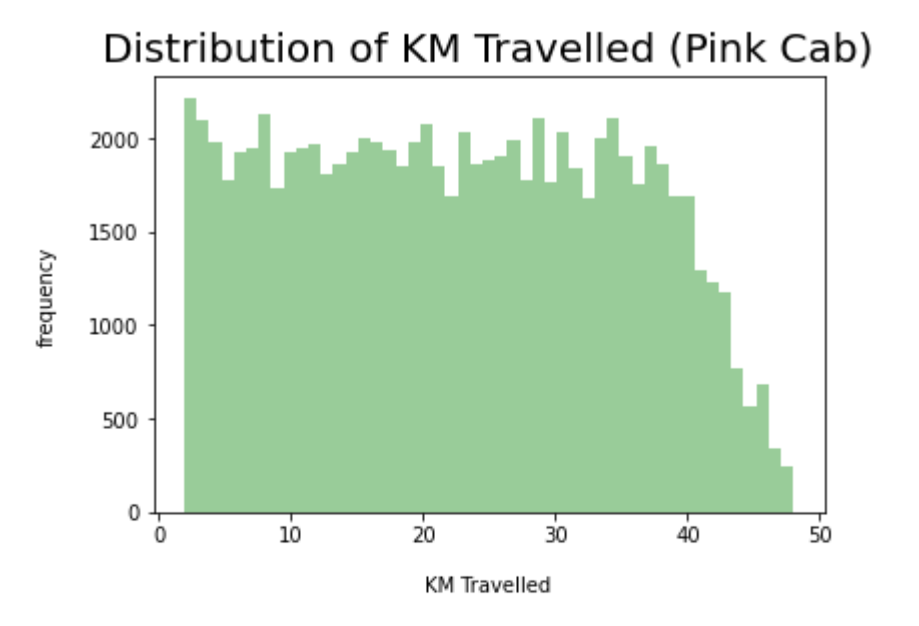
Relationship Exploration

The Choice of people between two cabs

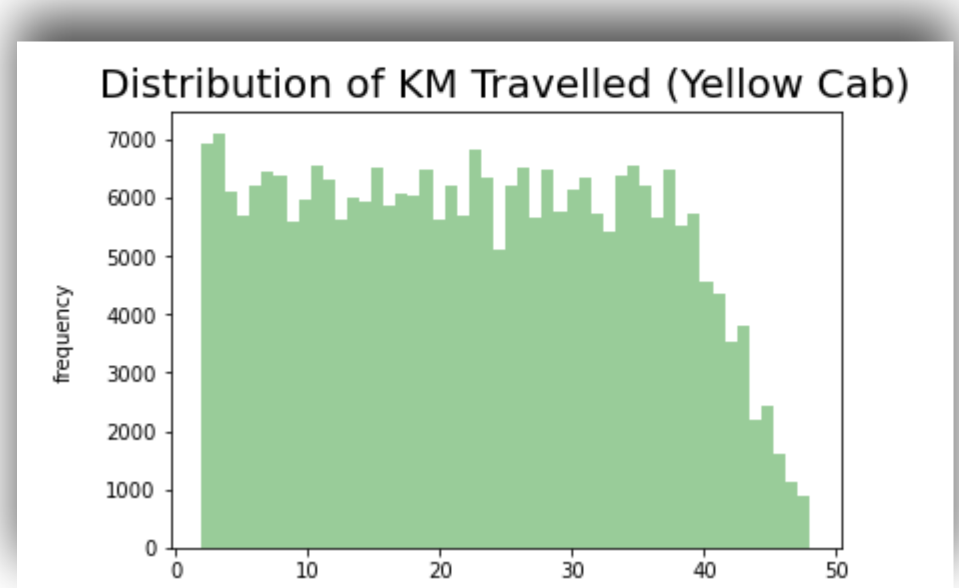


- ❑ **The choice between the two is similar no special preference**

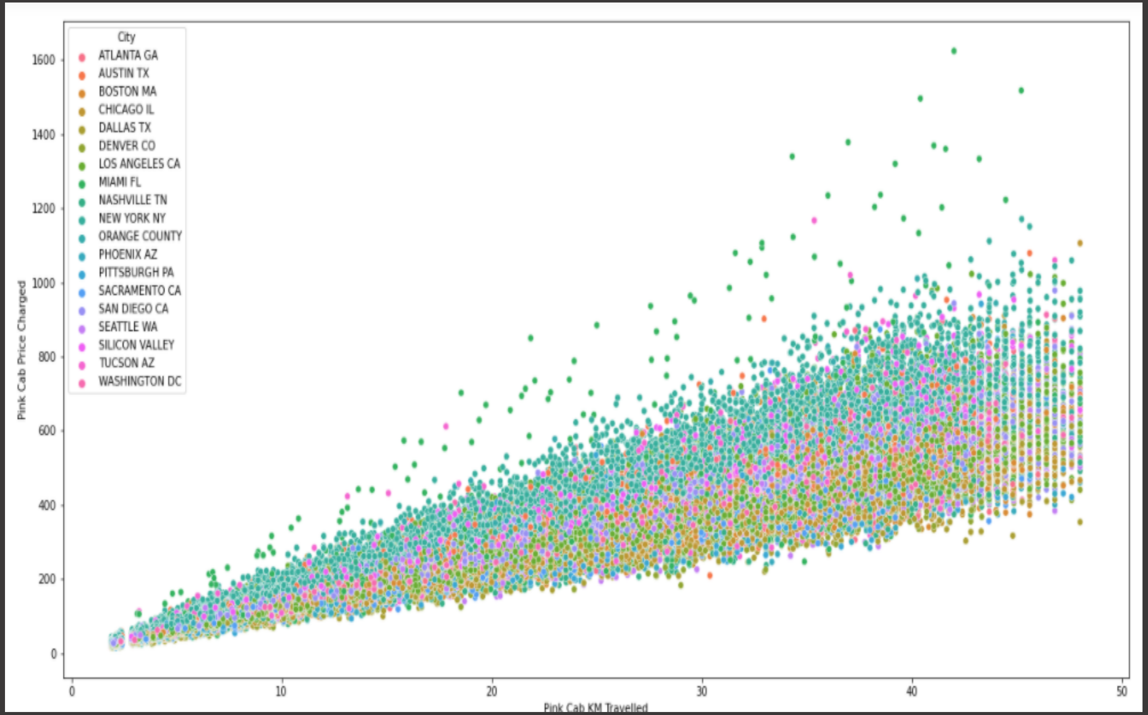
Kilometres Distribution between the two kinds of Cabs

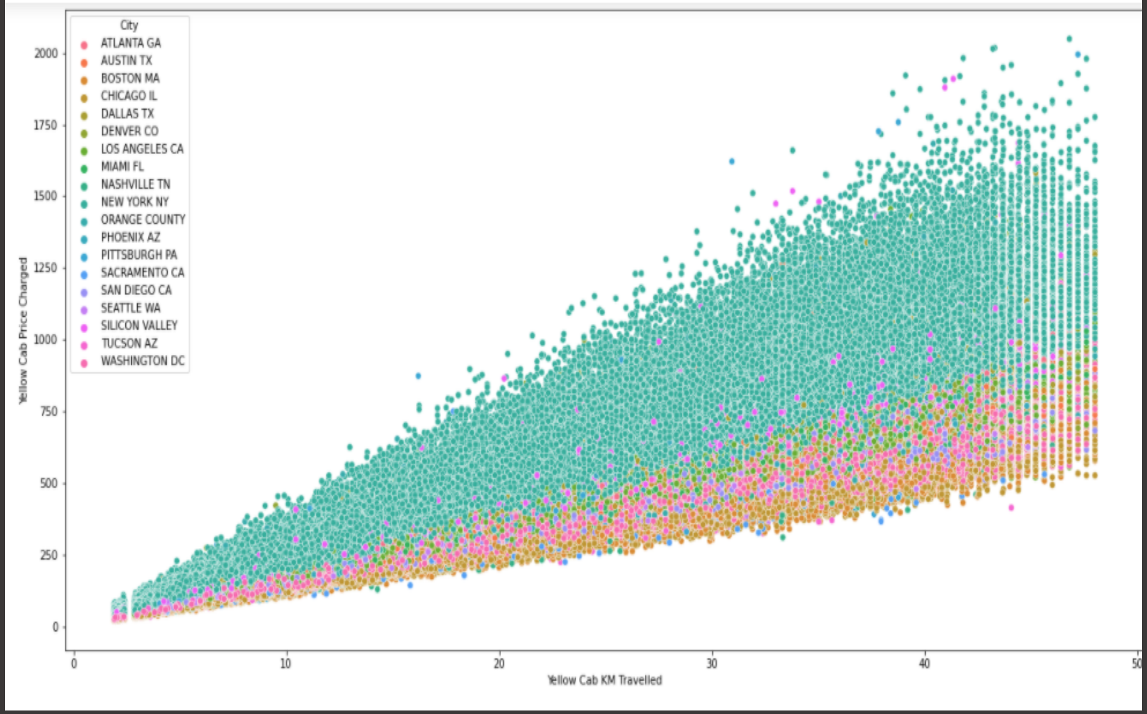


In most cases, the amount of kilometres rode was similar from the range of 2 to 46



Price Distance distribution

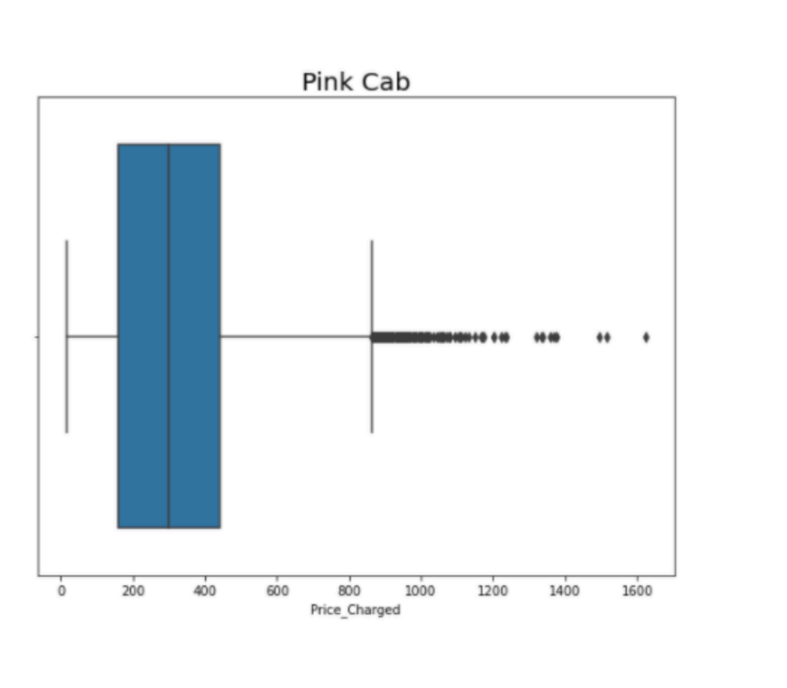


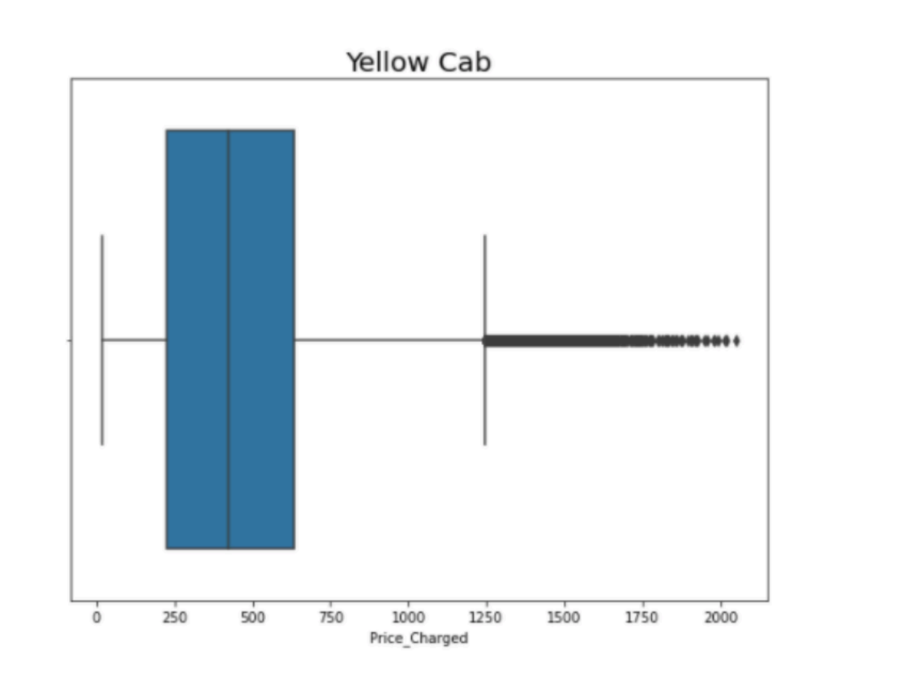


❑ **Pink Cab has Almost same riding price**

❑ **Yellow Cab has higher price in more developed city**

Price Range Distribution

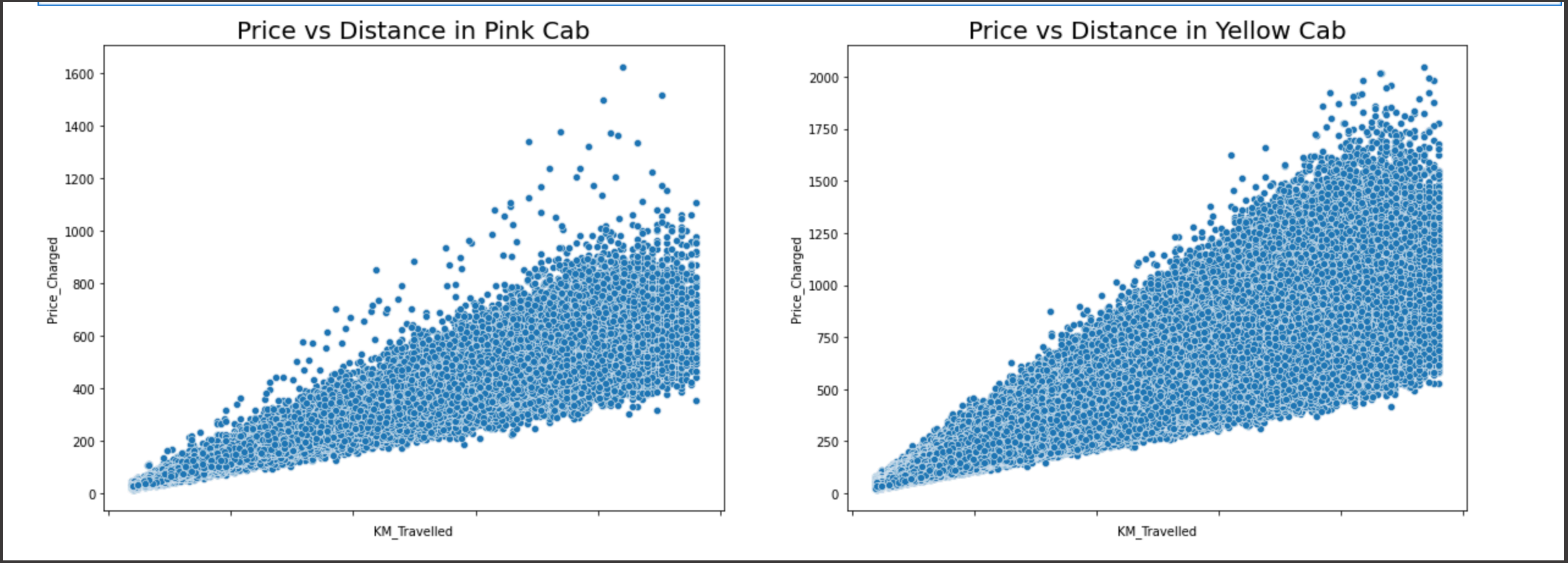




❑ **The Average price charge for yellow cab is higher**

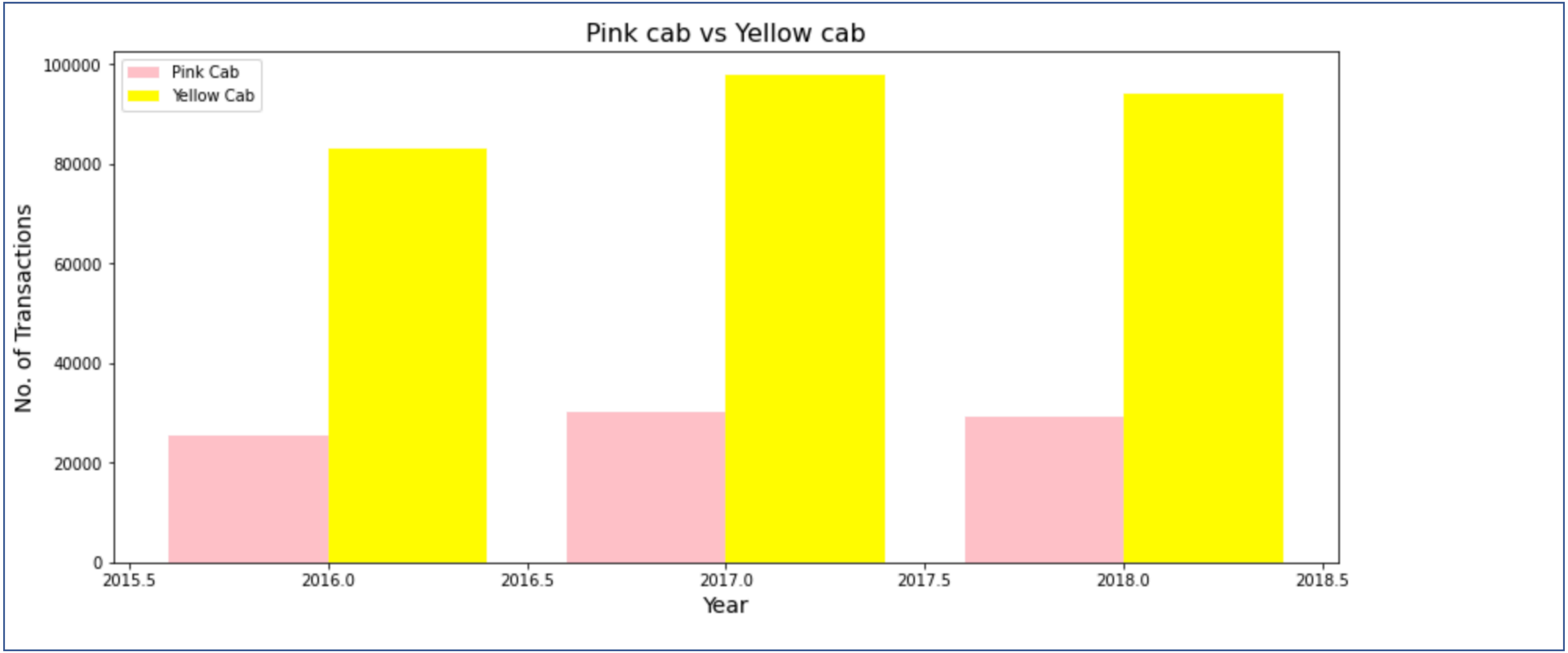
❑ **As Assumption needs to be established that Outliers will not influence the result, we can see it is almost the same.**

Price Distance distribution



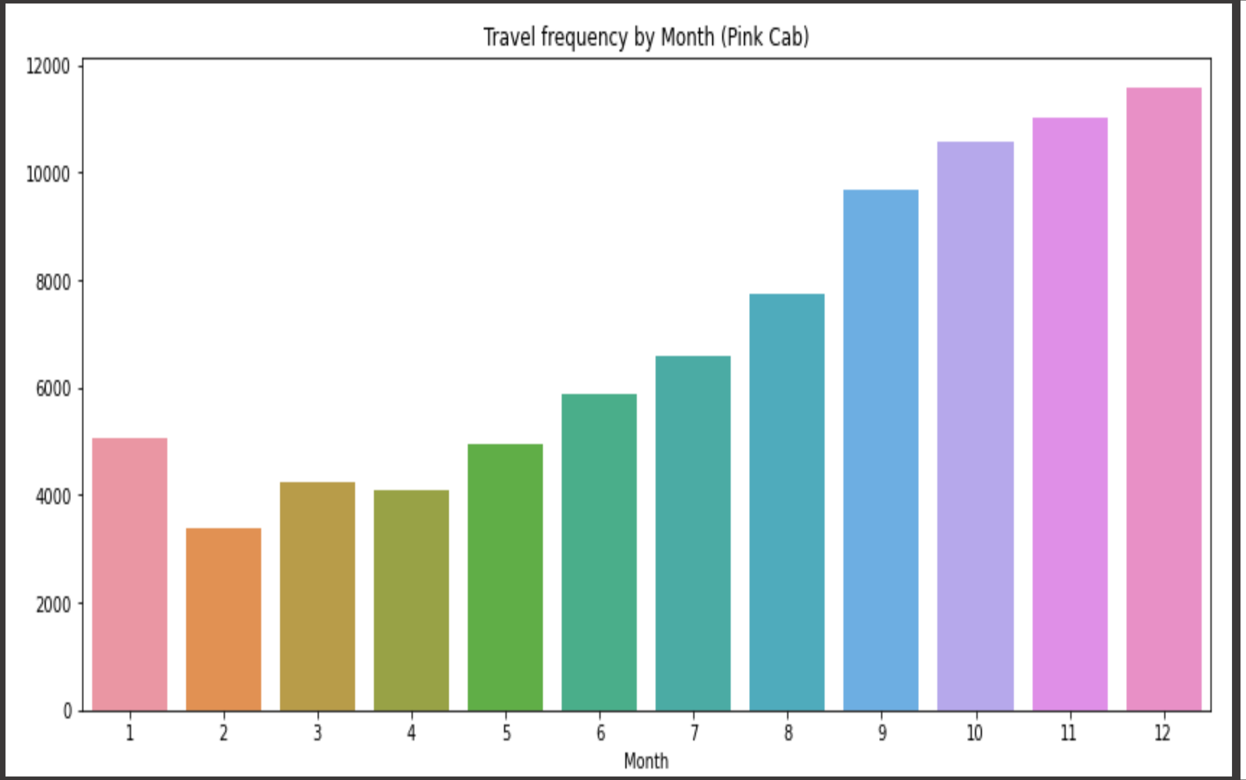
- ❑ **The Lower KM price is nearly the same between two cabs**
- ❑ **The Higher KM price is lower in pink cab(on AVG)**

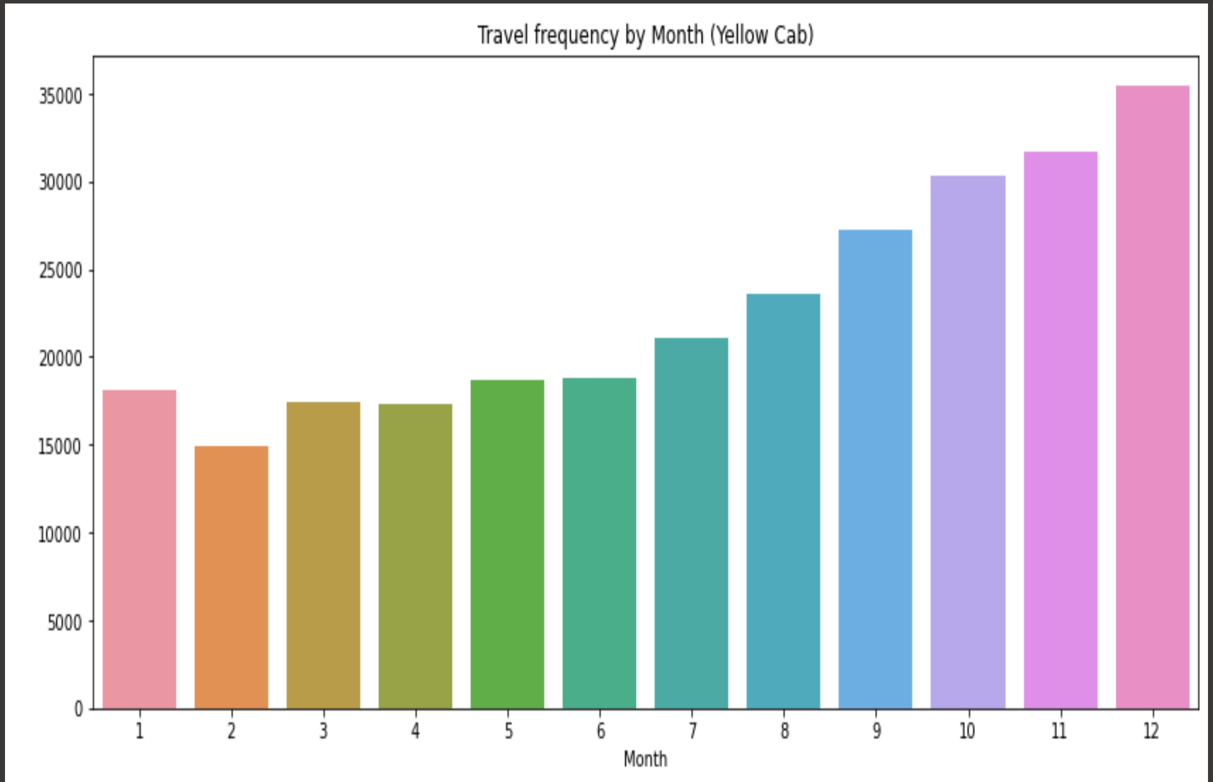
Choice of Cab



- ❑ **Yellow has higher No. Of transaction throughout three years**
- ❑ **Indicating being a higher “percent” of market holding**

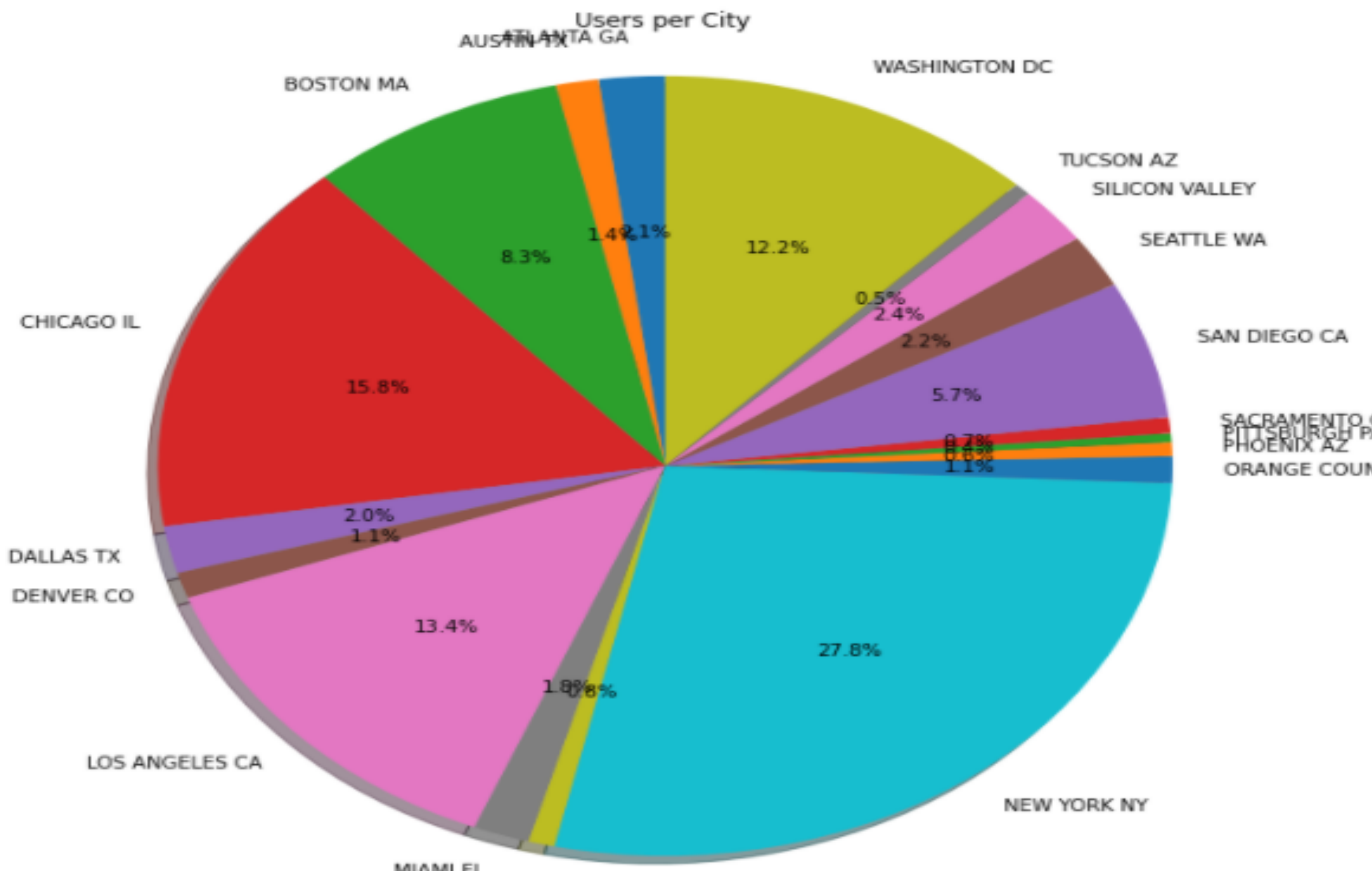
Seasonality Analysis





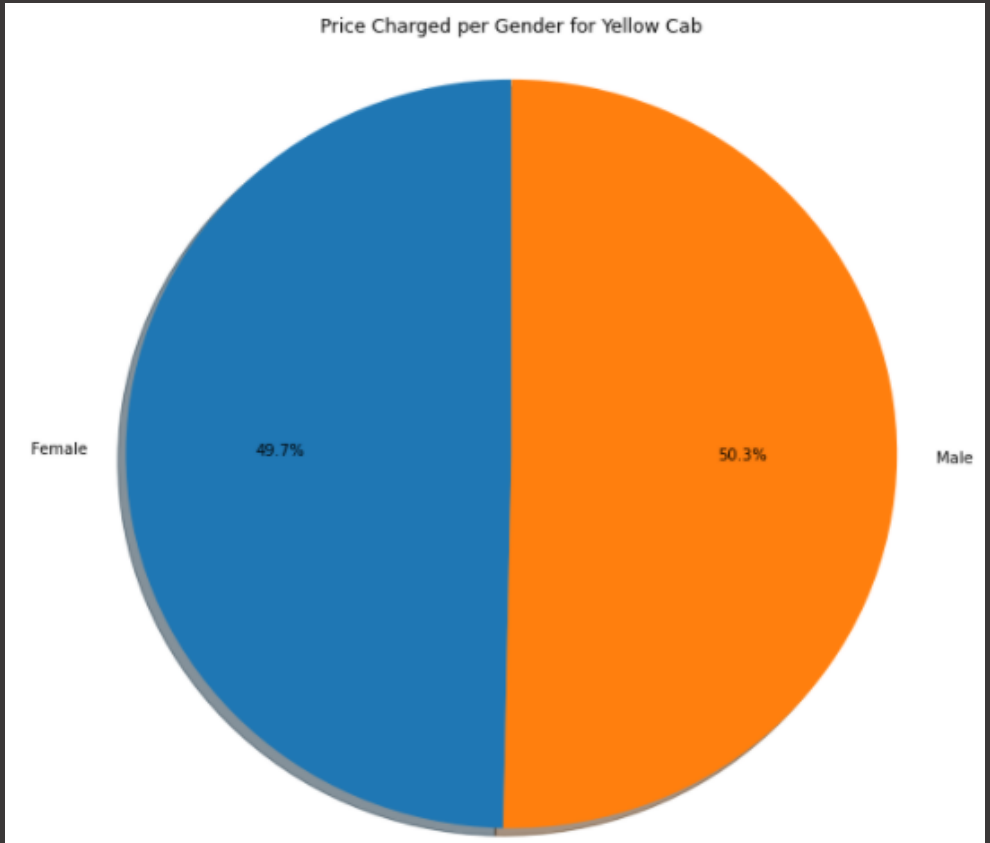
- ❑ **It shows that seasonality does matter the usage of cabs, especially it shows highest with Winter have higher usage of cabs compare to Spring.**

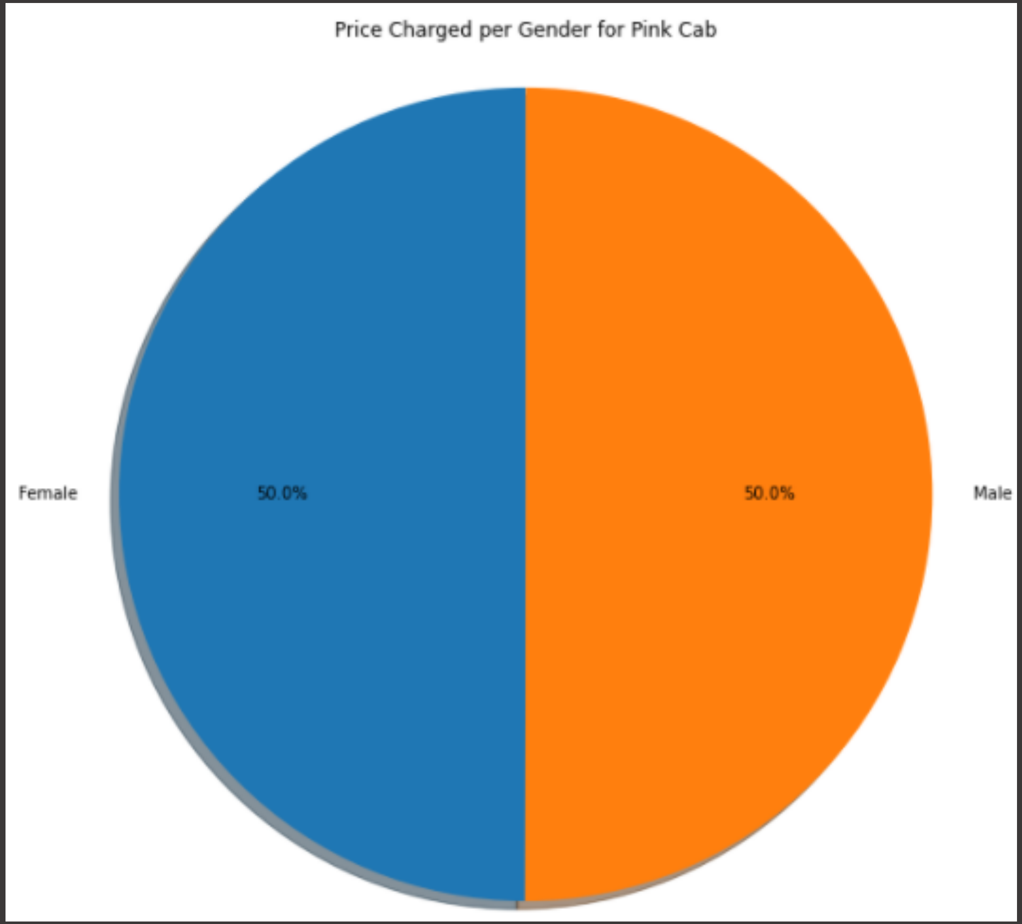
The Choice of cities in cabs



- ❑ **NYC has highest cab use of 27.8% while Chicago and LA also highly relies on cabs**

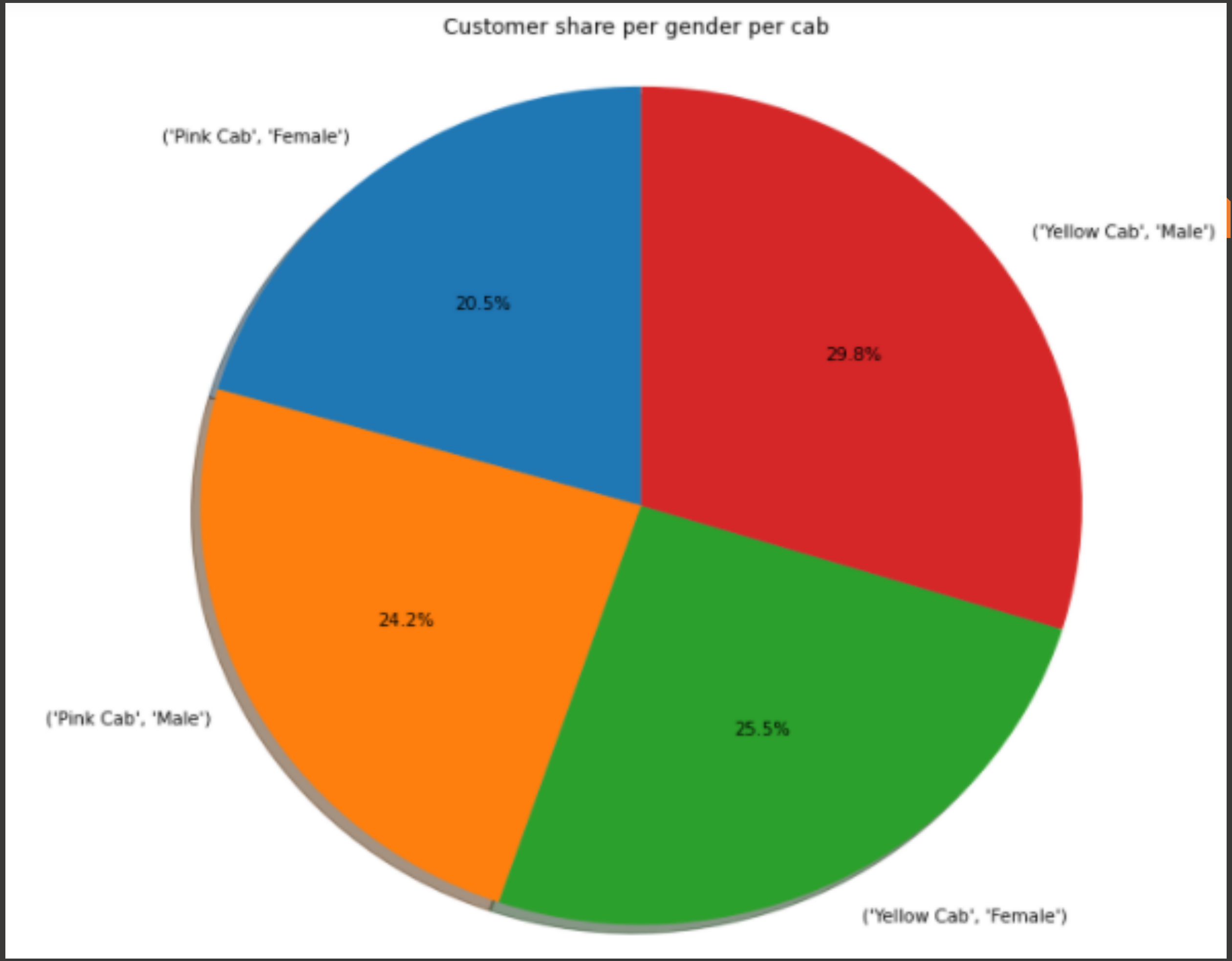
Gender distribution





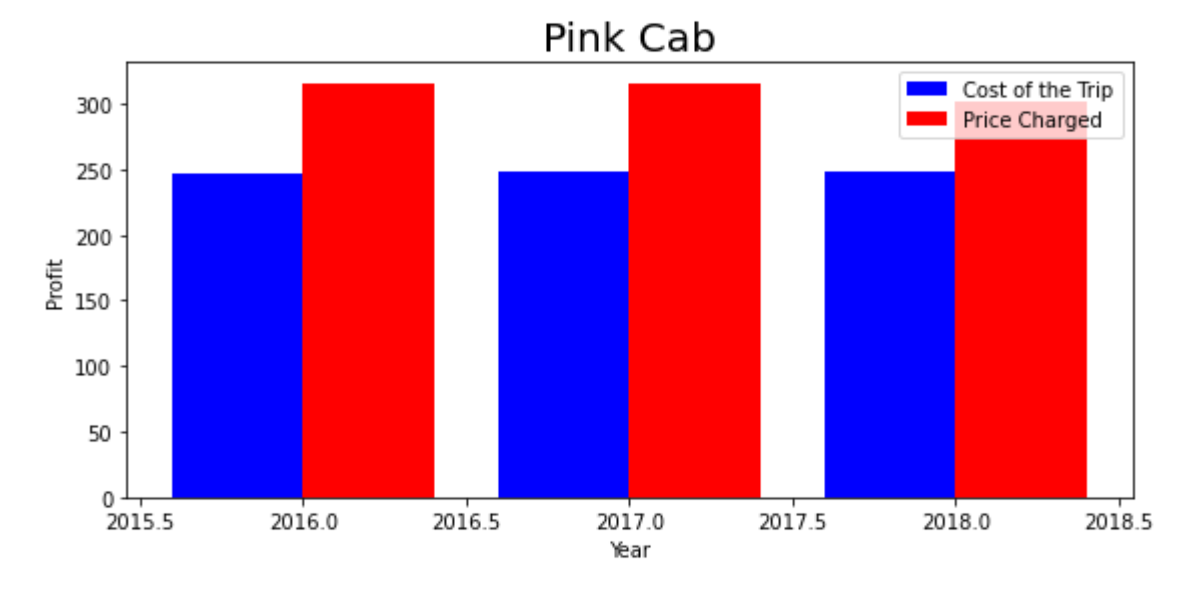
- ❑ **The Gender difference could be very low to none between these two cab**

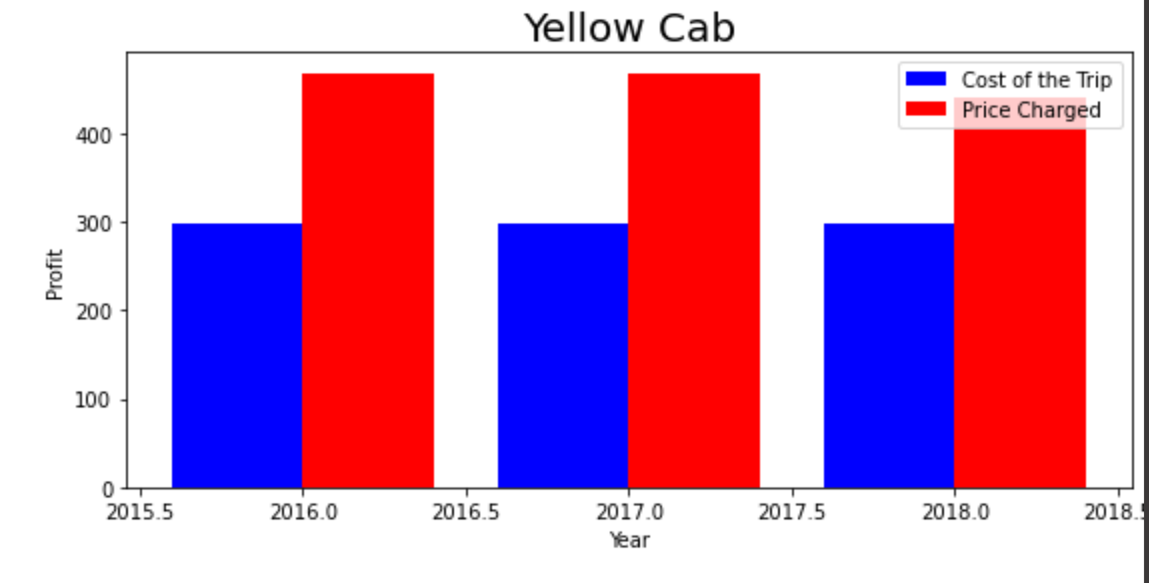
Gender distribution Further



However in this picture we could see that there are more female choosing yellow instead of pink, we yet know whether or not the difference is due to Number of case. Discuss in hypothesis testing

Profit Analysis



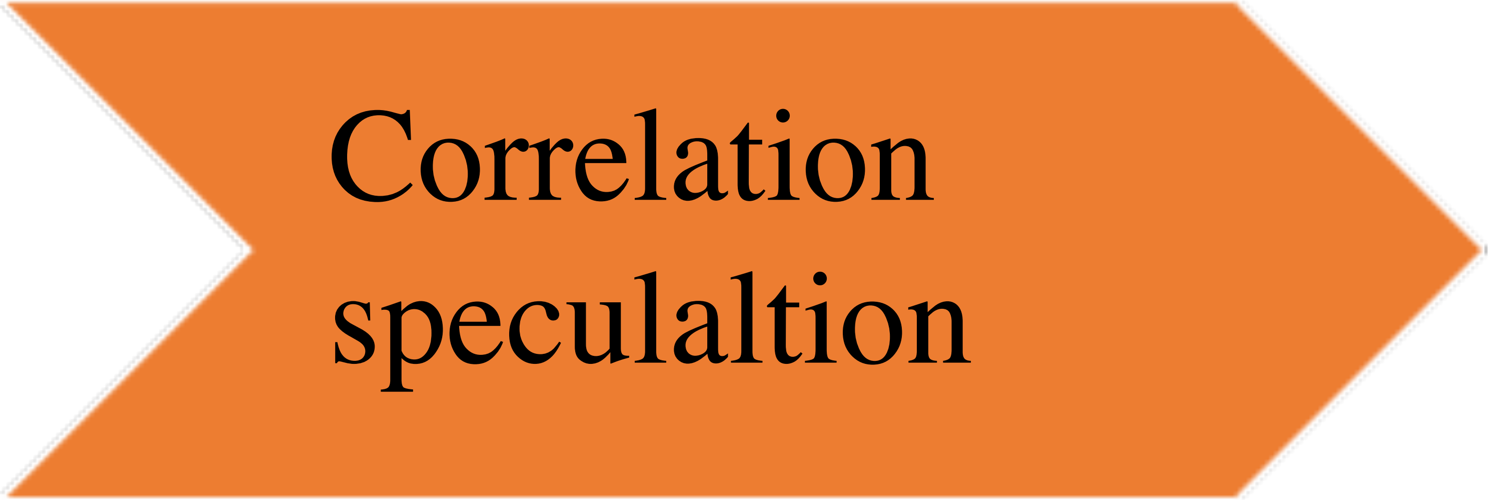


- **On average yellow cab seems to have a higher profit**

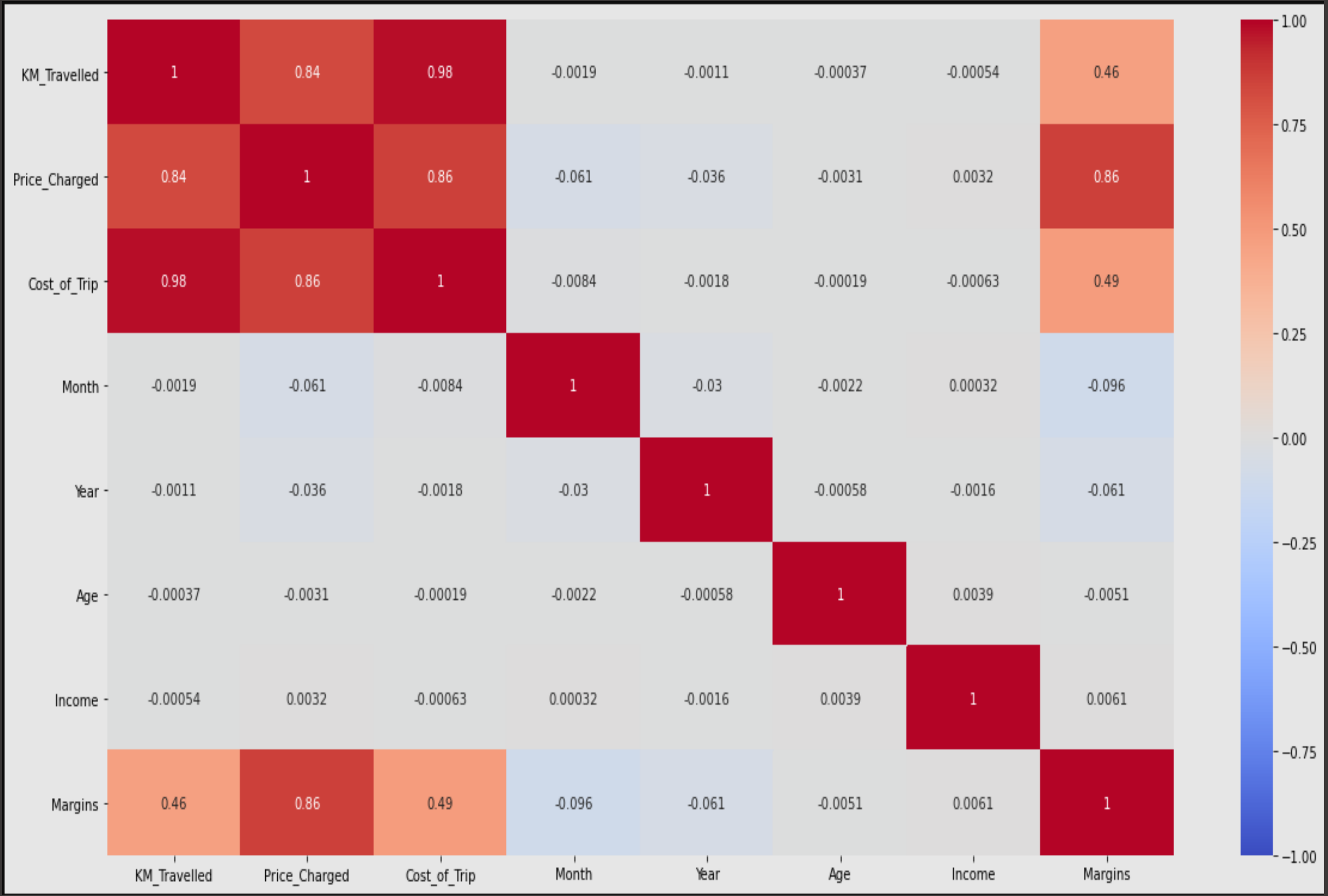
Summarizing EDA analysis

- ❑ **On average yellow cab seems to have a higher profit**
- ❑ **Ride distance is similar**
- ❑ **Yellow car has higher price when distance is longer in developed cities**
- ❑ **Might be due to outliers, Price range of yellow cab might be higher**
- ❑ **No significant “percent” difference in gender between two cabs, but have number difference in gender between two cabs**
- ❑ **Both cabs are influenced by seasonality, but influence is smaller in yellow cab**

Correlation Observation



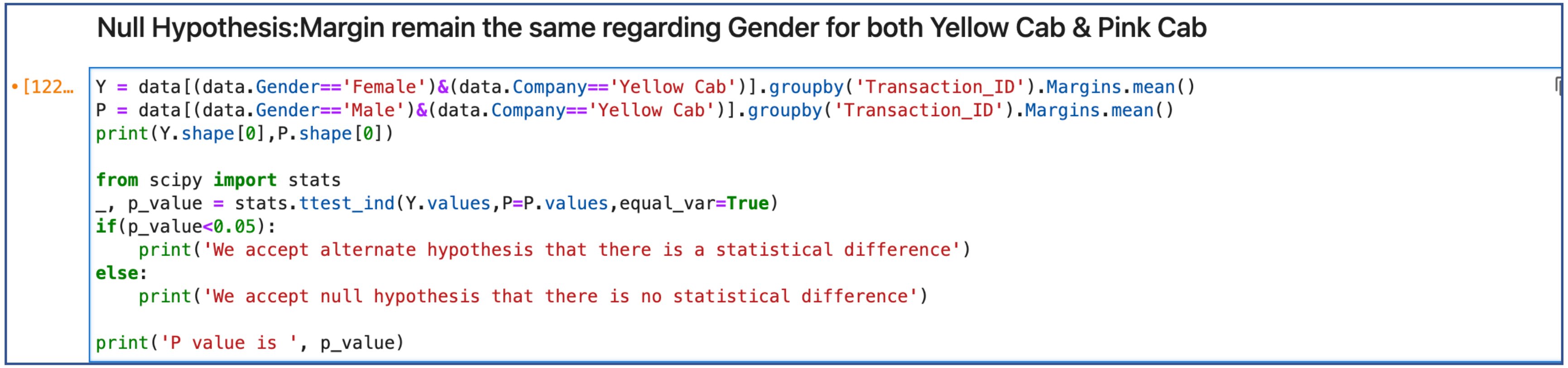
Correlation speculaltion



	KM_Travelled	Price_Charged	Cost_of_Trip	Month	Year	Age	Income	Margins
KM_Travelled	1	0.84	0.98	-0.0019	-0.0011	-0.00037	-0.00054	0.46
Price_Charged	0.84	1	0.86	-0.061	-0.036	-0.0031	0.0032	0.86
Cost_of_Trip	0.98	0.86	1	-0.0084	-0.0018	-0.00019	-0.00063	0.49
Month	-0.0019	-0.061	-0.0084	1	-0.03	-0.0022	0.00032	-0.096
Year	-0.0011	-0.036	-0.0018	-0.03	1	-0.00058	-0.0016	-0.061
Age	-0.00037	-0.0031	-0.00019	-0.0022	-0.00058	1	0.0039	-0.0051
Income	-0.00054	0.0032	-0.00063	0.00032	-0.0016	0.0039	1	0.0061
Margins	0.46	0.86	0.49	-0.096	-0.061	-0.0051	0.0061	1

- ❑ **The Graph shows Margins being correlated to almost everything.**
- ❑ **Definitely Price and KM traveled is correlated**

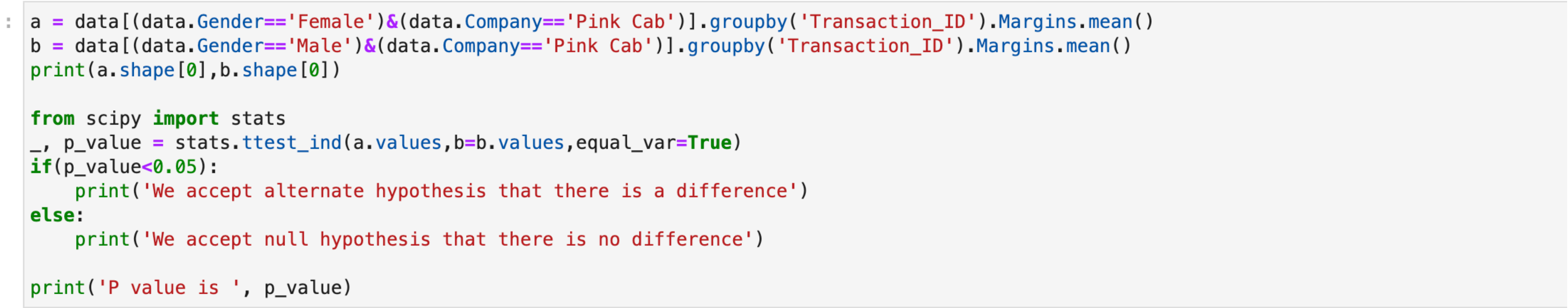
Hypothesis Testing



Null Hypothesis:Margin remain the same regarding Gender for both Yellow Cab & Pink Cab

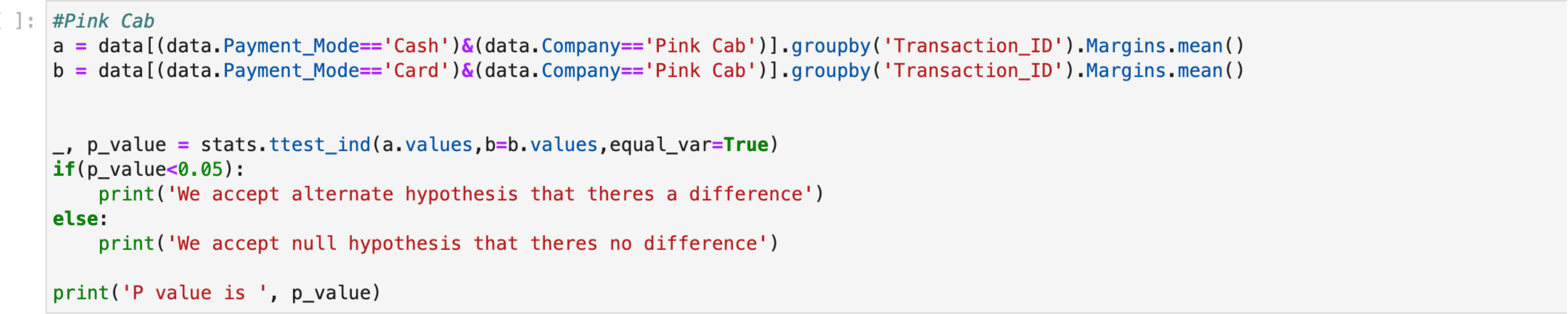
```
Y = data[(data.Gender=='Female')&(data.Company=='Yellow Cab')].groupby('Transaction_ID').Margins.mean()
P = data[(data.Gender=='Male')&(data.Company=='Yellow Cab')].groupby('Transaction_ID').Margins.mean()
print(Y.shape[0],P.shape[0])

from scipy import stats
_, p_value = stats.ttest_ind(Y.values,P=P.values,equal_var=True)
if(p_value<0.05):
    print('We accept alternate hypothesis that there is a statistical difference')
else:
    print('We accept null hypothesis that there is no statistical difference')

print('P value is ', p_value)
```

We received a result lower than 0.05 so we reject null hypothesis for yellow cab

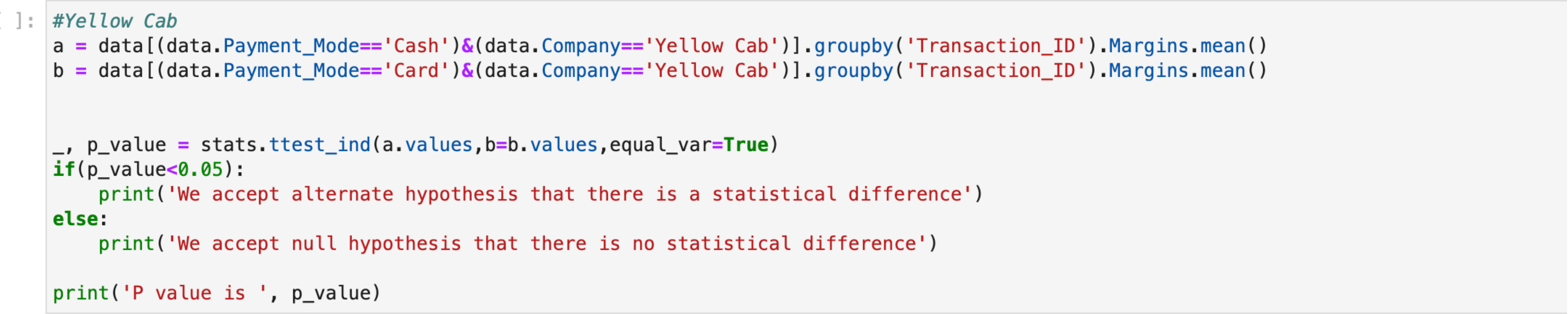


```
a = data[(data.Gender=='Female')&(data.Company=='Pink Cab')].groupby('Transaction_ID').Margins.mean()
b = data[(data.Gender=='Male')&(data.Company=='Pink Cab')].groupby('Transaction_ID').Margins.mean()
print(a.shape[0],b.shape[0])

from scipy import stats
_, p_value = stats.ttest_ind(a.values,b=b.values,equal_var=True)
if(p_value<0.05):
    print('We accept alternate hypothesis that there is a difference')
else:
    print('We accept null hypothesis that there is no difference')

print('P value is ', p_value)
```

Received higher than 0.05 p value, suggesting no difference for pink cab

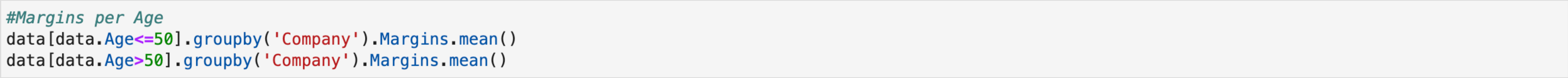


```python
#Pink Cab
a = data[(data.Payment_Mode=='Cash')&(data.Company=='Pink Cab')].groupby('Transaction_ID').Margins.mean()
b = data[(data.Payment_Mode=='Card')&(data.Company=='Pink Cab')].groupby('Transaction_ID').Margins.mean()


_, p_value = stats.ttest_ind(a.values,b=b.values,equal_var=True)
if(p_value<0.05):
    print('We accept alternate hypothesis that theres a difference')
else:
    print('We accept null hypothesis that theres no difference')

print('P value is ', p_value)
```

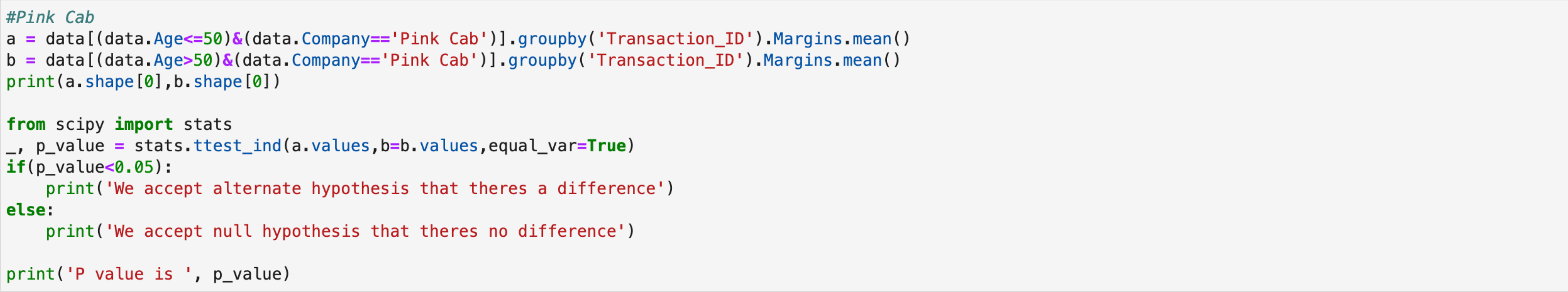


```python
#Yellow Cab
a = data[(data.Payment_Mode=='Cash')&(data.Company=='Yellow Cab')].groupby('Transaction_ID').Margins.mean()
b = data[(data.Payment_Mode=='Card')&(data.Company=='Yellow Cab')].groupby('Transaction_ID').Margins.mean()


_, p_value = stats.ttest_ind(a.values,b=b.values,equal_var=True)
if(p_value<0.05):
    print('We accept alternate hypothesis that there is a statistical difference')
else:
    print('We accept null hypothesis that there is no statistical difference')

print('P value is ', p_value)
```

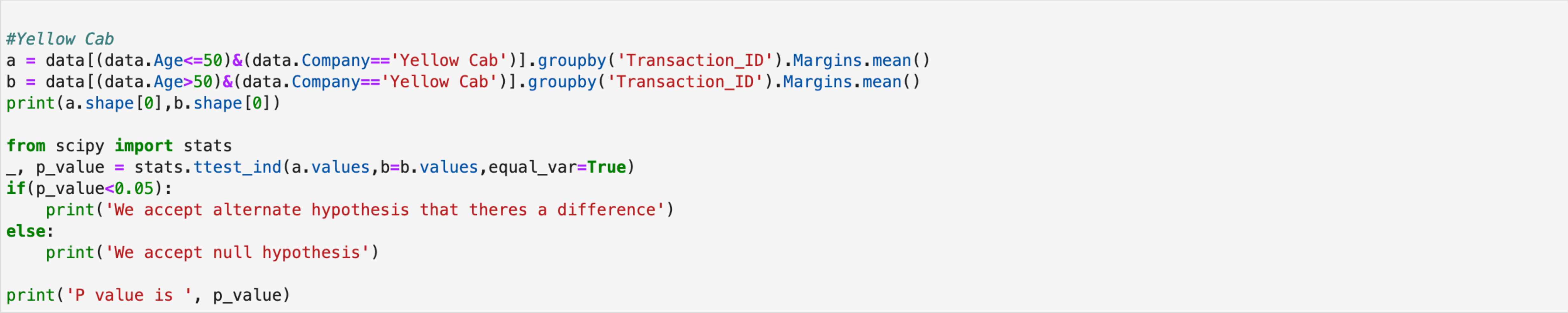
Mode of payment has no difference in this case



```
#Margins per Age
data[data.Age<=50].groupby('Company').Margins.mean()
data[data.Age>50].groupby('Company').Margins.mean()
```



```
#Pink Cab
a = data[(data.Age<=50)&(data.Company=='Pink Cab')].groupby('Transaction_ID').Margins.mean()
b = data[(data.Age>50)&(data.Company=='Pink Cab')].groupby('Transaction_ID').Margins.mean()
print(a.shape[0],b.shape[0])

from scipy import stats
_, p_value = stats.ttest_ind(a.values,b=b.values,equal_var=True)
if(p_value<0.05):
    print('We accept alternate hypothesis that theres a difference')
else:
    print('We accept null hypothesis that theres no difference')

print('P value is ', p_value)
```



```

#Yellow Cab
a = data[(data.Age<=50)&(data.Company=='Yellow Cab')].groupby('Transaction_ID').Margins.mean()
b = data[(data.Age>50)&(data.Company=='Yellow Cab')].groupby('Transaction_ID').Margins.mean()
print(a.shape[0],b.shape[0])

from scipy import stats
_, p_value = stats.ttest_ind(a.values,b=b.values,equal_var=True)
if(p_value<0.05):
    print('We accept alternate hypothesis that theres a difference')
else:
    print('We accept null hypothesis')

print('P value is ', p_value)
```

The Yellow cab provides support for elder group so there is difference in P value, suggest there is difference in age

Recommendation For XYZ firm

❑ **Profit Margin**: For Yellow Cab the Profit Margin is higher per year from 2016 to 2018 in comparison to Pink Cab.

❑ **Margin per Age**: In Yellow Cab there is difference in Margin for elders, Which is a better choice for elders compare to Pink

❑ Yellow Cab **decreases Margins with the increase in Transaction**, Choose December as Yellow cab has the highest difference compare to Pink cab for investment

❑ New York has the highest percent for cab Choose NY for investment

❑ **Transaction per year**:Yellow cab is higher than Pink cab so much

Yellow cab should be better in general