Data Glacier

Your Deep Learning Partner

G2M insight for Cab Investment firm

Company Name : XYZ

Location: US

Team: Data and Analytics

Date: 12-November-2022

Agenda

Executive Summary

Data Exploration

EDA

EDA Summary

Hypothesis Testing

Recommendations

EXECUTIVE SUMMARY

Problem Summary:

In this project, we are going to provide productive insight through an EDA (Exploratory Data Analysis) approach about the Cab industry market for our client in order to help them take a final decision before investment. Our analysis is based on a comparison between two cab companies (Pink cab company / Yellow cab company) to show which one represents the best opportunity to invest in.

Analysis:

The analysis has been divided into 5 parts:

- Data Exploration.
- EDA
- Finding the most profitable Cab company.
- Hypothesis Testing.
- Recommendations for investment

DATA EXPLORATION

DATASET DESCRIPTION

- For this analysis, there are 4 datasets provided:

1) cab_data.csv : this file describes attributes of Transactions like Companies, Km travelled, price charged etc.
2) Customer_ID.csv : this file consists of unique customer ids with their ages and income.
3) Transaction_ID.csv : this file consists of Transaction Ids with the payment mode.
4) City.csv : this file consists of various cities, their populations and number of users.

- Time frame of the data : 2016-01-31 to 2018-12-31.
- The main dataset is created by merging mentioned 4 datasets.

ASSUMPTIONS:

- Outliers are present in 'Price Charged' feature. We are not treating this as outliers because of unavailability of more details. (We Assume these might be due to high end cars or harsh weather conditions.)
- There are no duplicated rows neither missing values in the datasets.
- The 'Profit' feature is calculated as follows:

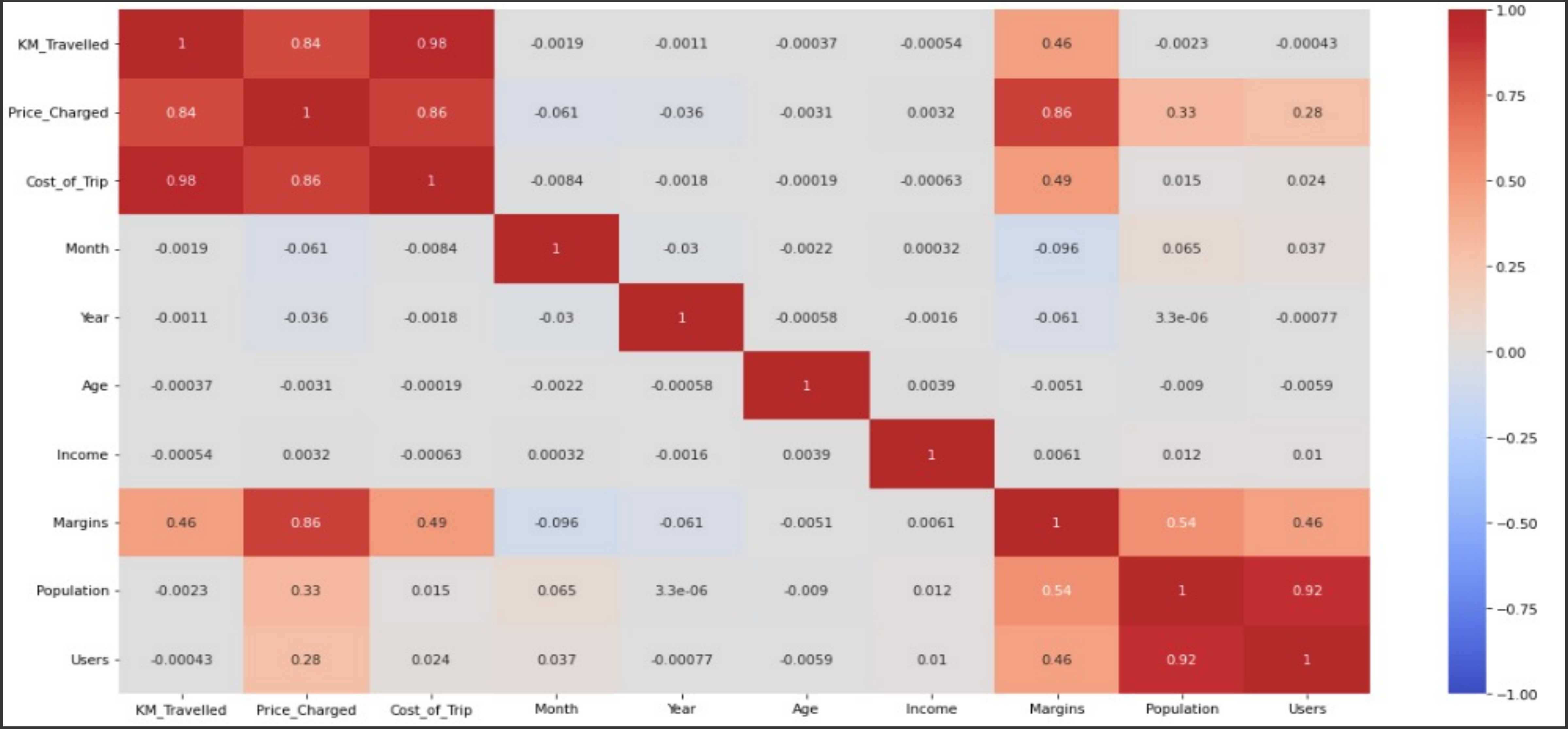
Profit = Price Charged – Cost of Trip

Data Glacier

Your Deep Learning Partner

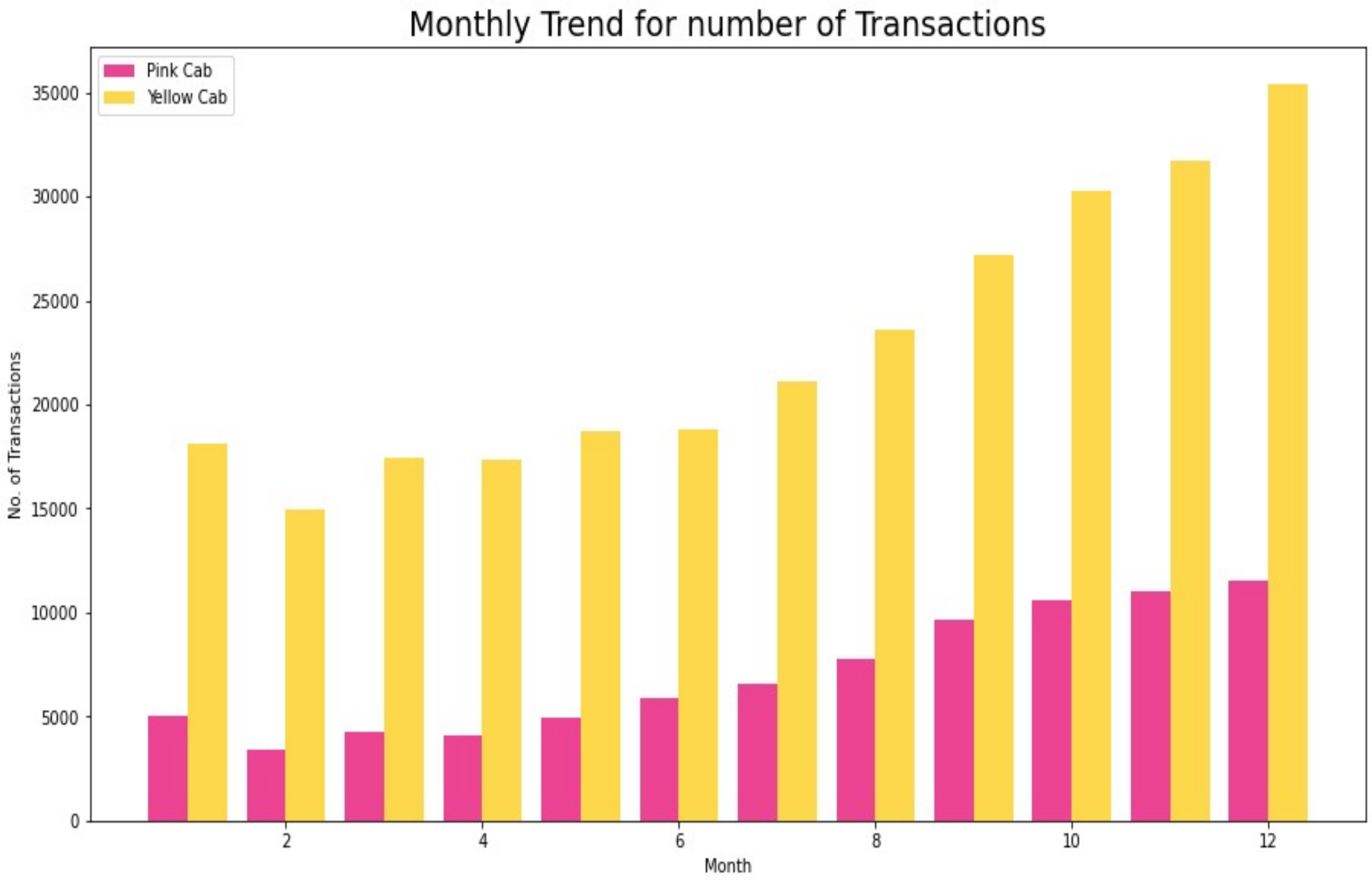
EXPLORATORY DATA ANALYSIS(EDA)

CORRELATION MATRIX



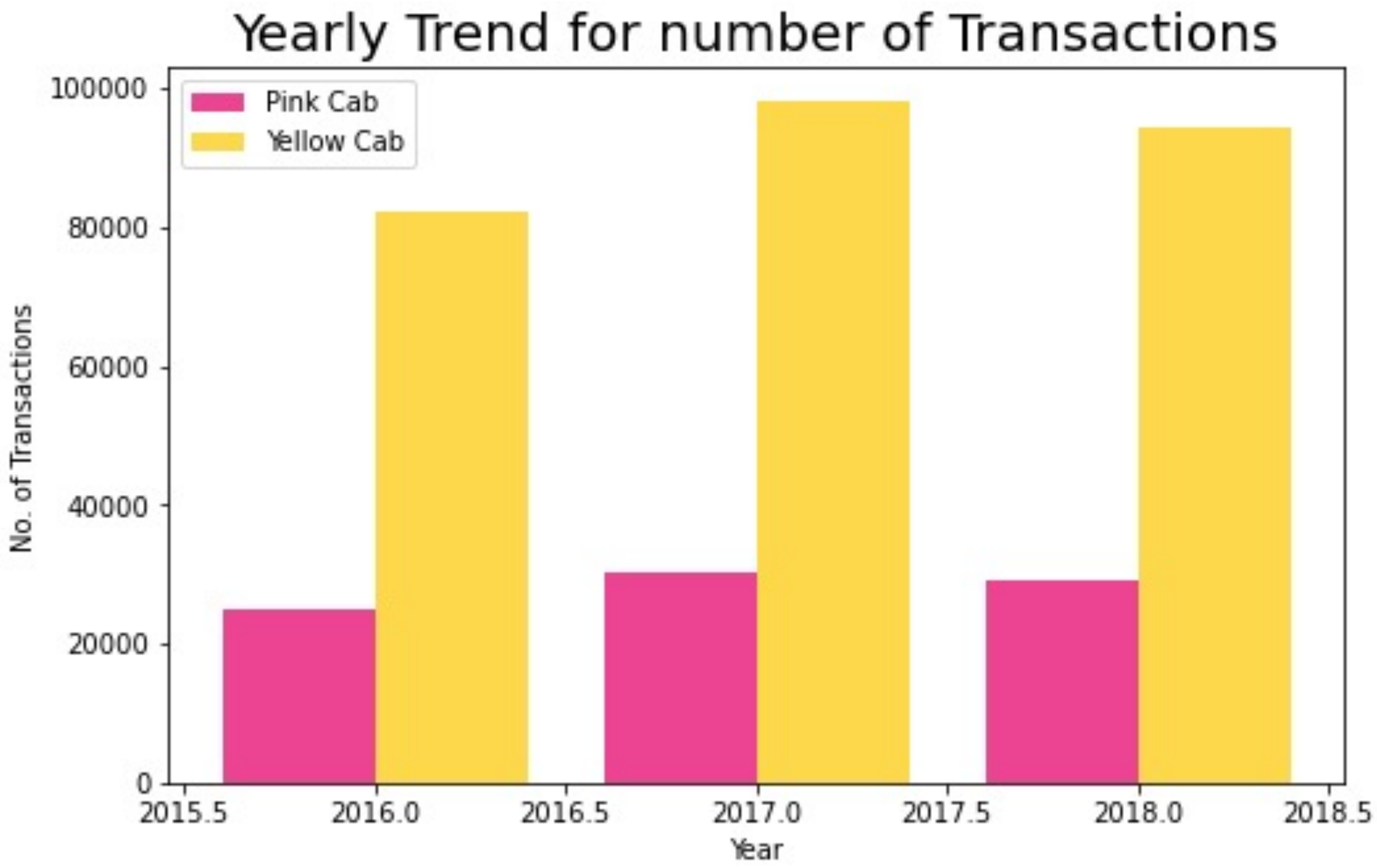
	KM_Travelled	Price_Charged	Cost_of_Trip	Month	Year	Age	Income	Margins	Population	Users
KM_Travelled	1	0.84	0.98	-0.0019	-0.0011	-0.00037	-0.00054	0.46	-0.0023	-0.00043
Price_Charged	0.84	1	0.86	-0.061	-0.036	-0.0031	0.0032	0.86	0.33	0.28
Cost_of_Trip	0.98	0.86	1	-0.0084	-0.0018	-0.00019	-0.00063	0.49	0.015	0.024
Month	-0.0019	-0.061	-0.0084	1	-0.03	-0.0022	0.00032	-0.096	0.065	0.037
Year	-0.0011	-0.036	-0.0018	-0.03	1	-0.00058	-0.0016	-0.061	3.3e-06	-0.00077
Age	-0.00037	-0.0031	-0.00019	-0.0022	-0.00058	1	0.0039	-0.0051	-0.009	-0.0059
Income	-0.00054	0.0032	-0.00063	0.00032	-0.0016	0.0039	1	0.0061	0.012	0.01
Margins	0.46	0.86	0.49	-0.096	-0.061	-0.0051	0.0061	1	0.54	0.46
Population	-0.0023	0.33	0.015	0.065	3.3e-06	-0.009	0.012	0.54	1	0.92
Users	-0.00043	0.28	0.024	0.037	-0.00077	-0.0059	0.01	0.46	0.92	1

Monthly trend of no. of transactions



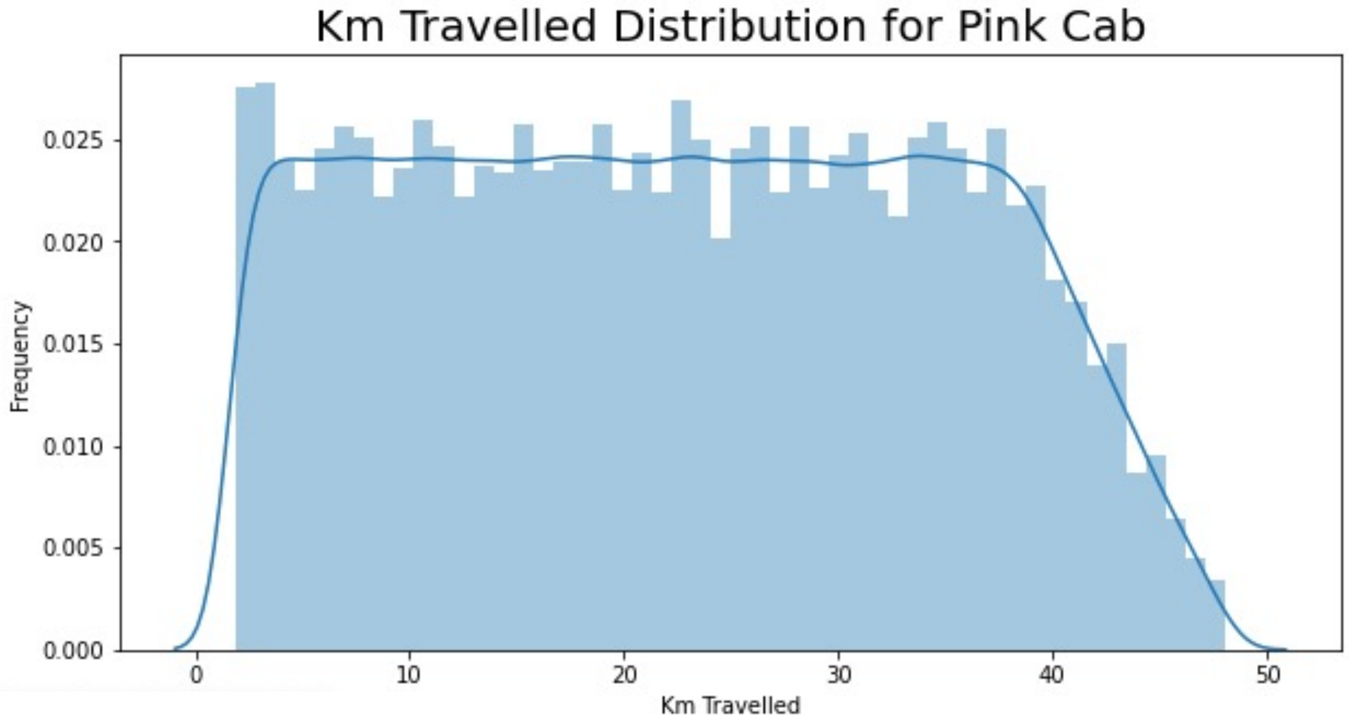
- ❖ Trend of number of transaction month wise is almost same for both the companies.
- ❖ No. of transactions for yellow cabs is nearly three times higher than pink cabs.

Yearly trend of no. of transactions

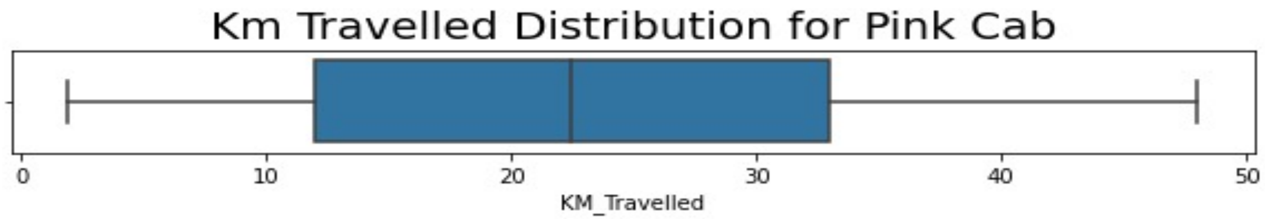


- Yearly transactions also show same trend for both with highest transactions in year 2017.
- Yearly transactions for yellow cabs are three times higher than pink cab.

DISTRIBUTION OF KMs TRAVELLED



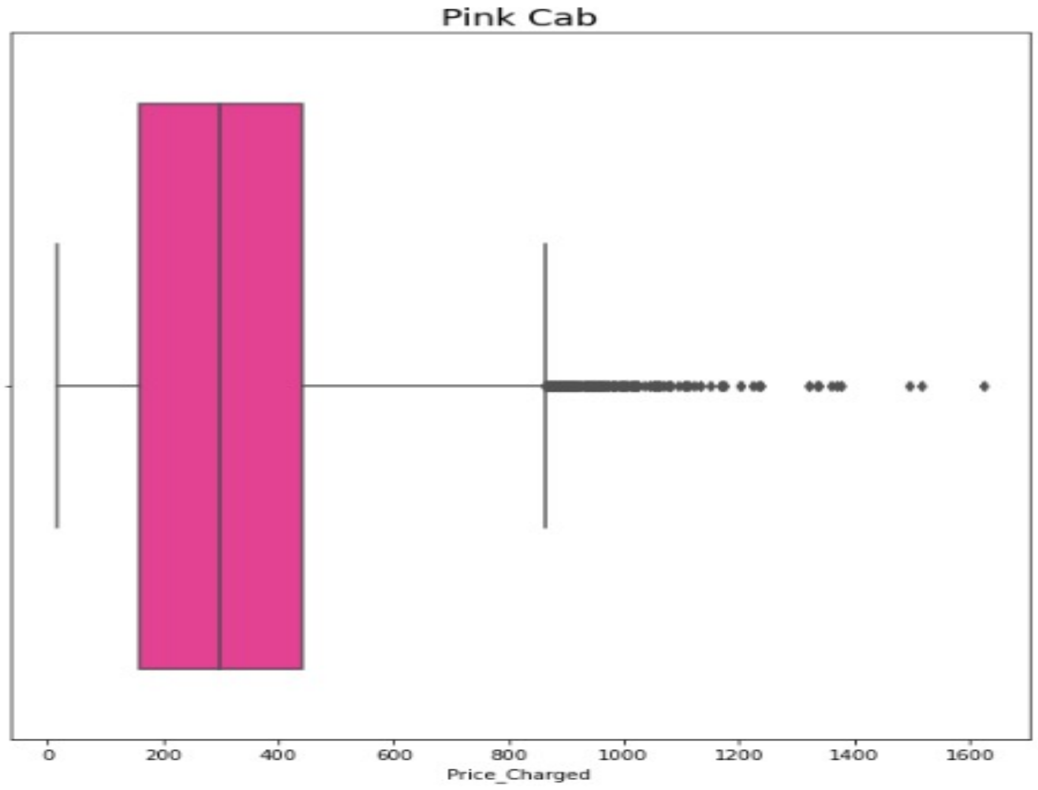
Km Travelled Distribution for Pink Cab

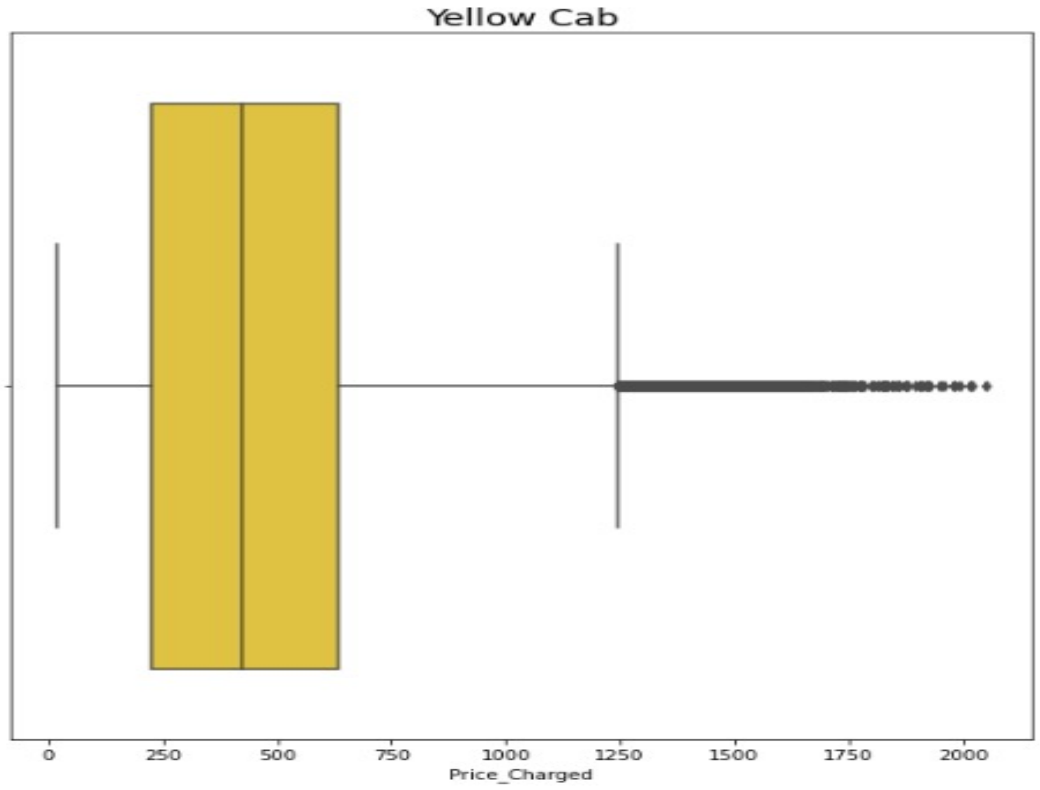


Km Travelled Distribution for Pink Cab

- KM's Travelled for all transactions are between 1.9 KM to 20 KM
- Median of KM's Travelled is 22.44 KMs

DISTRIBUTION OF PRICE CHARGED

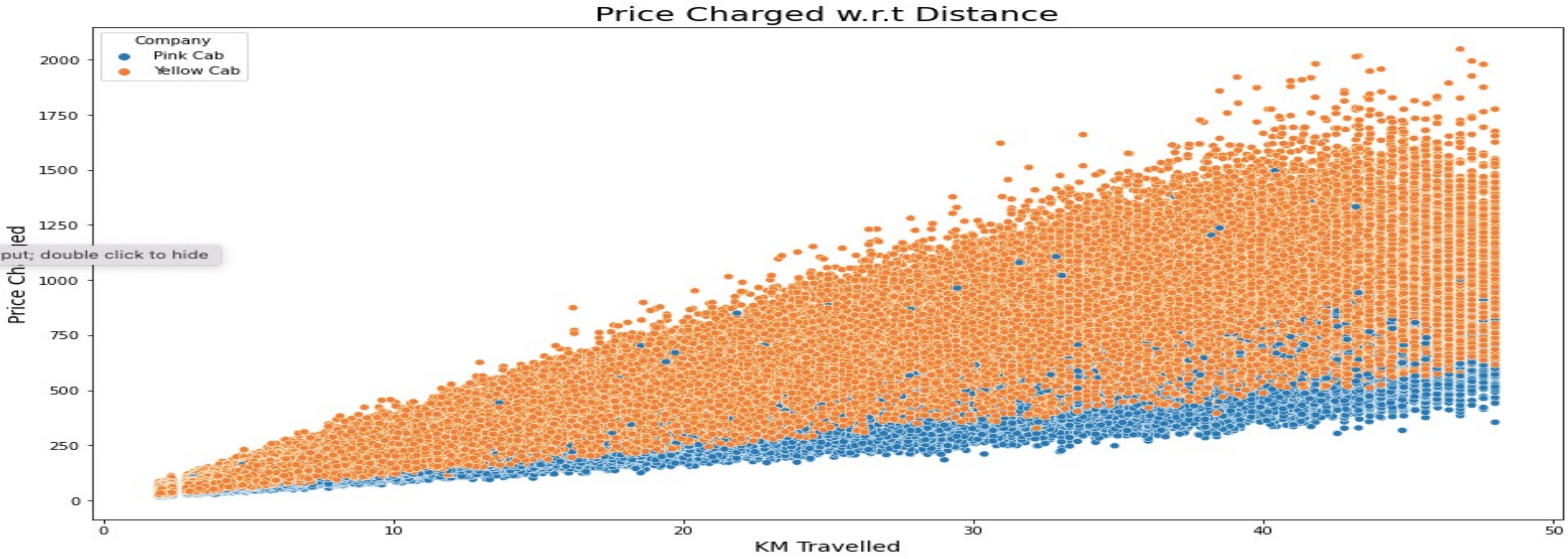




- The prices charged for Yellow Cabs are comparatively more as compared to that of Pink Cab.

- We are predicting that the outliers in the Price_charged are because of Holiday season or the high end services(Cars).

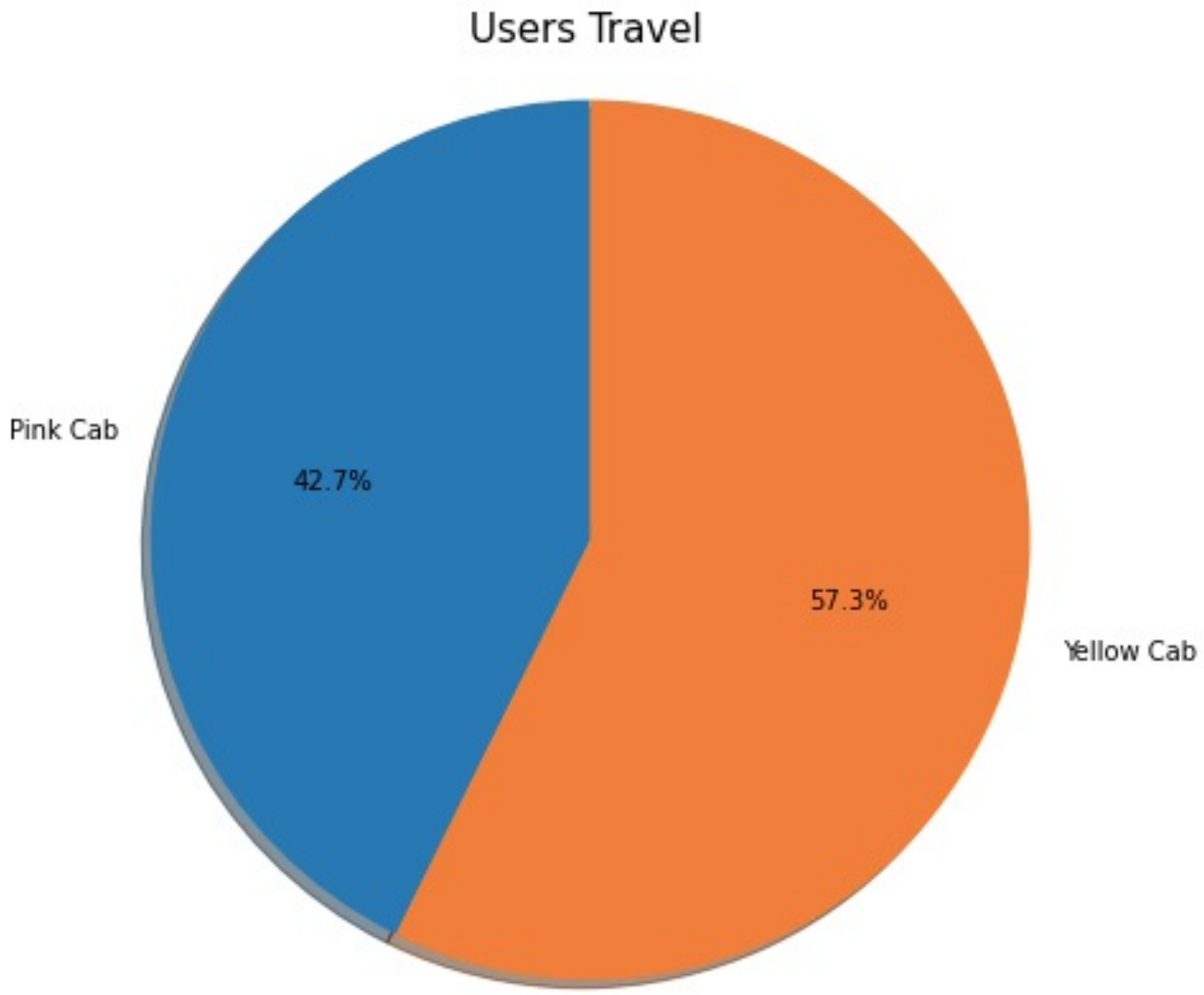
PRICE CHARGED W.R.T KM TRAVELLED



- ❖ There is linear relationship between price charged and KM's Travelled.

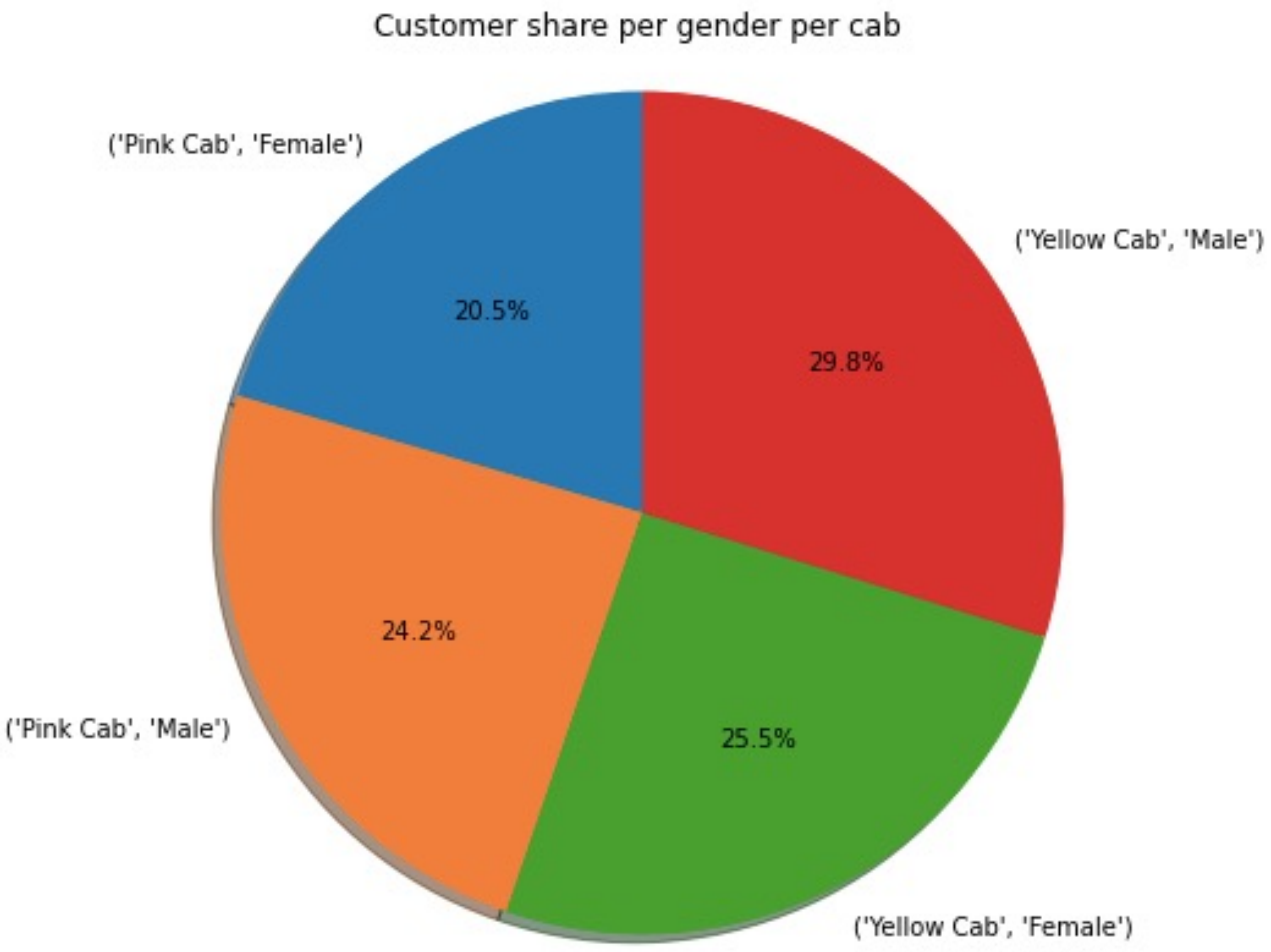
- ❖ Yellow cabs charges more price as compared to Pink cabs for KM travelled.

USER DISTRIBUTION IN BOTH THE COMPANIES



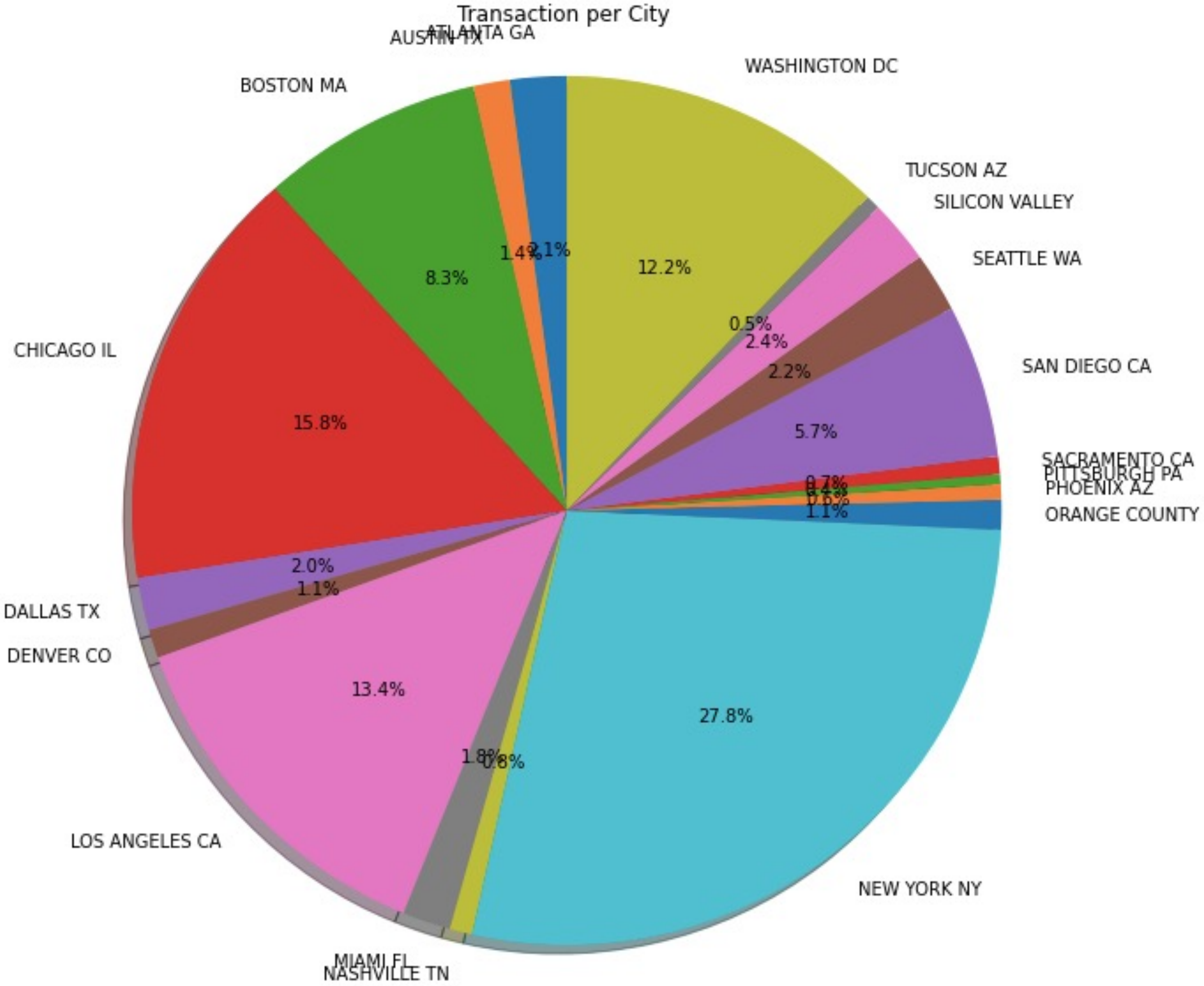
- ❖ Yellow Cabs have more number of users as compared to pink cabs.

GENDER DISTRIBUTION IN BOTH THE COMPANIES



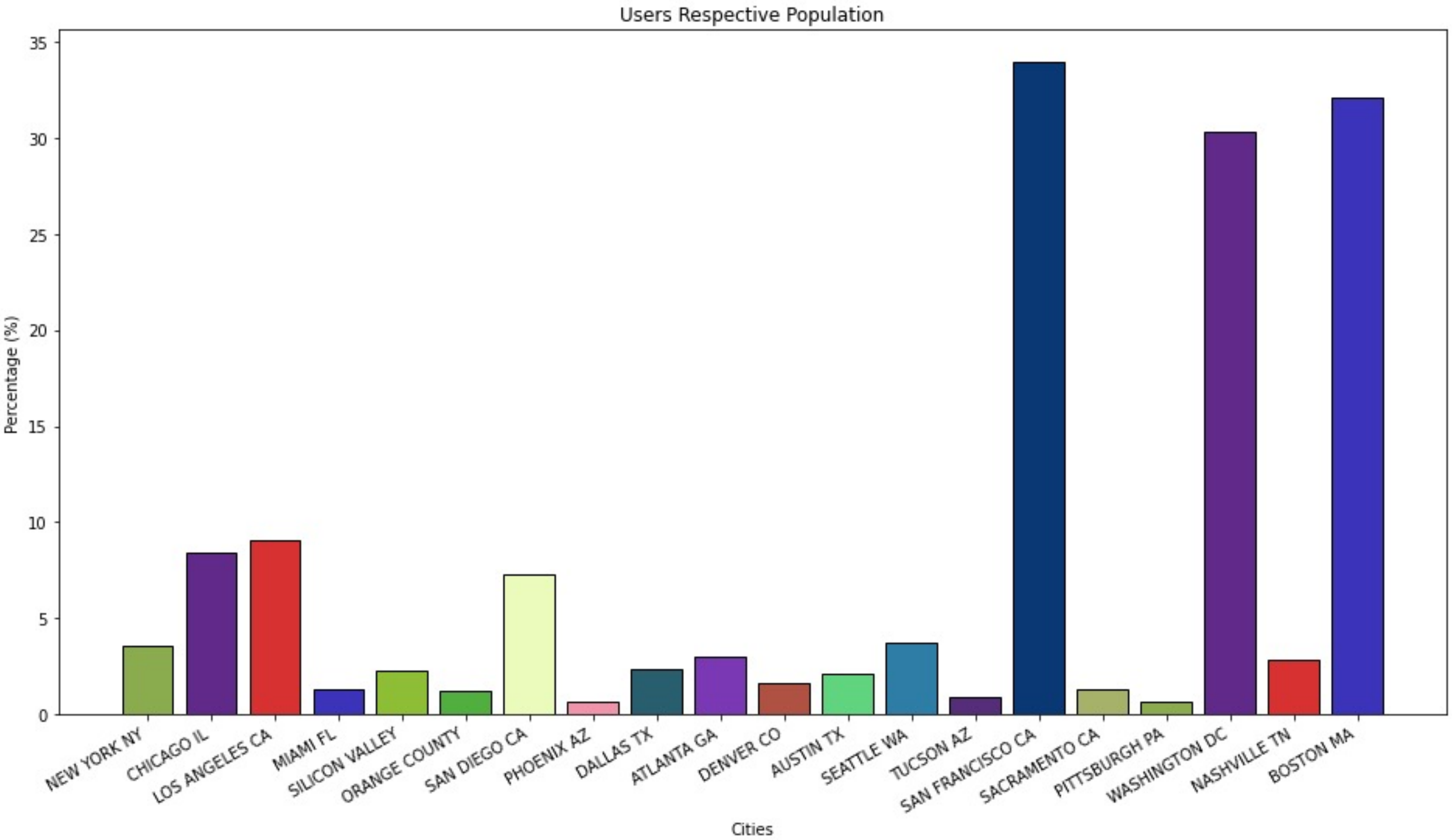
- For both companies male users are more as compared to female users
- Yellow Cab male users are highest in number

USER DISTRIBUTION ACROSS DIFFERENT CITIES



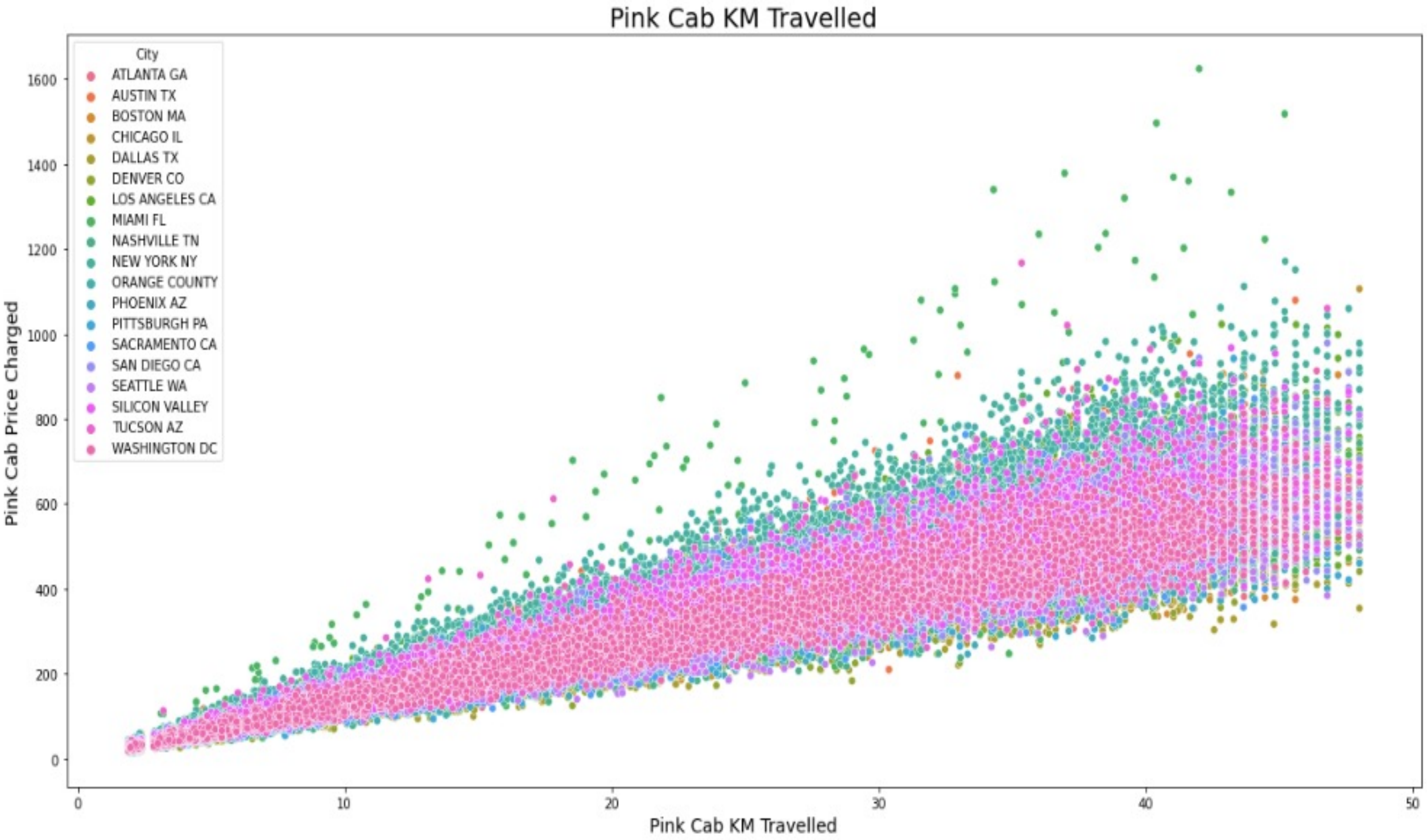
- ❖ New York has most number of users.
- ❖ Los Angeles, Chicago and Washington also has high number of users

PROPOTION OF POPULATION TO USERS FOR CITIES



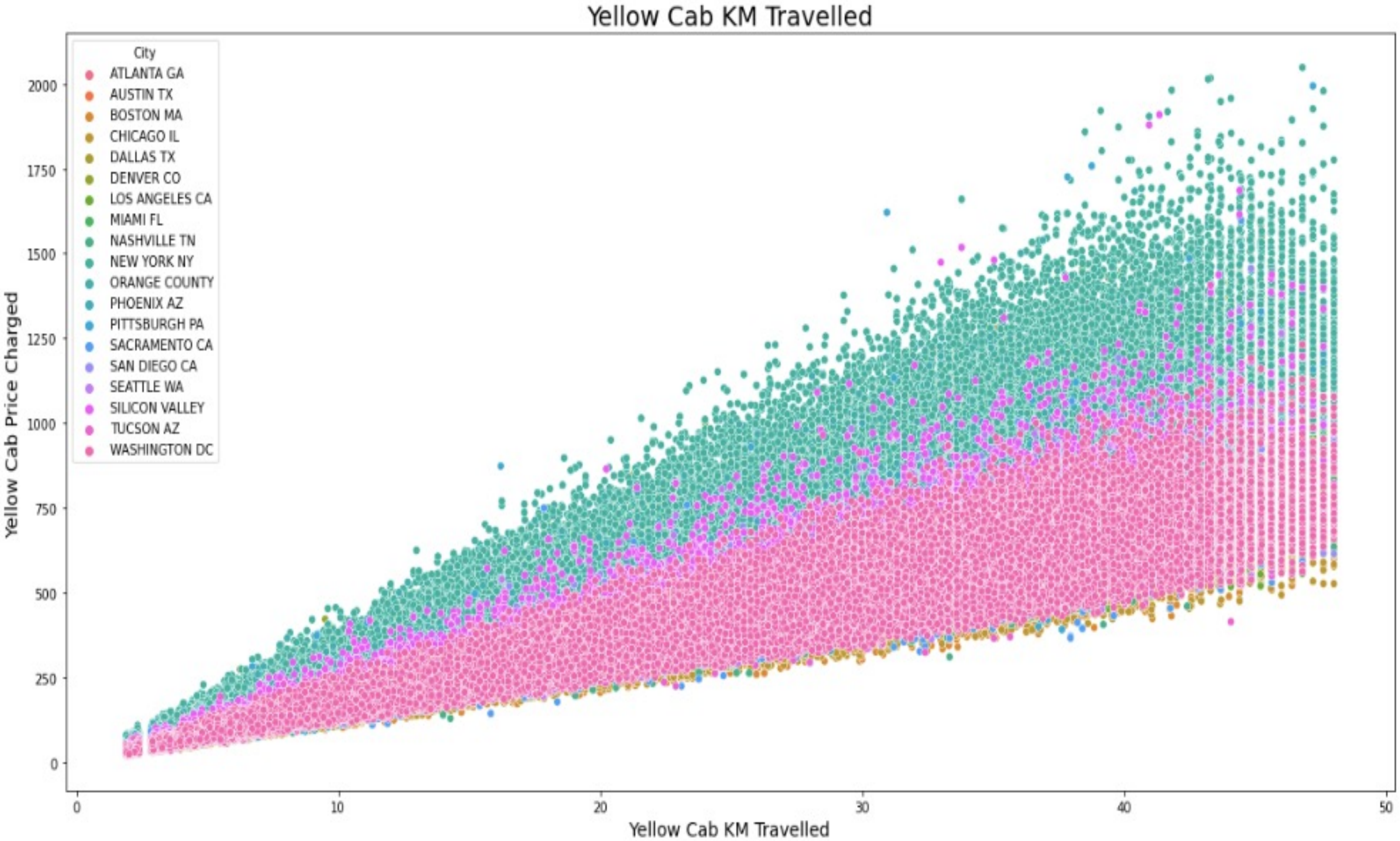
- San Francisco, Washington DC and Boston MA have most user/population proportion

PRICE CHARGED W.R.T. DISTANCE TRAVELLED



- ❖ Pink cabs charges almost same across different cities.

PRICE CHARGED W.R.T. DISTANCE TRAVELLED



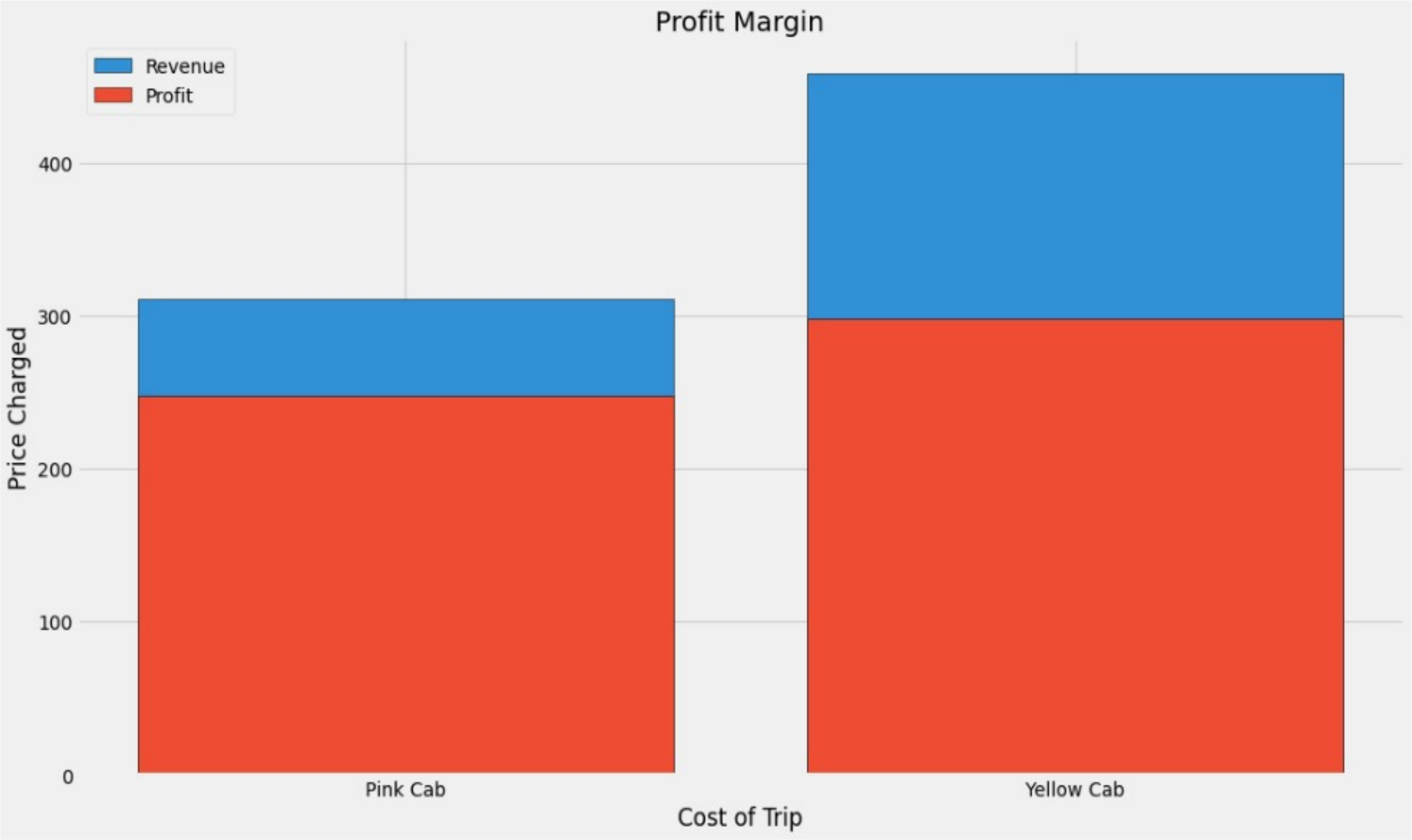
- ❖ Yellow cabs charges more in New York NY as compared to different cities.

Data Glacier

Your Deep Learning Partner

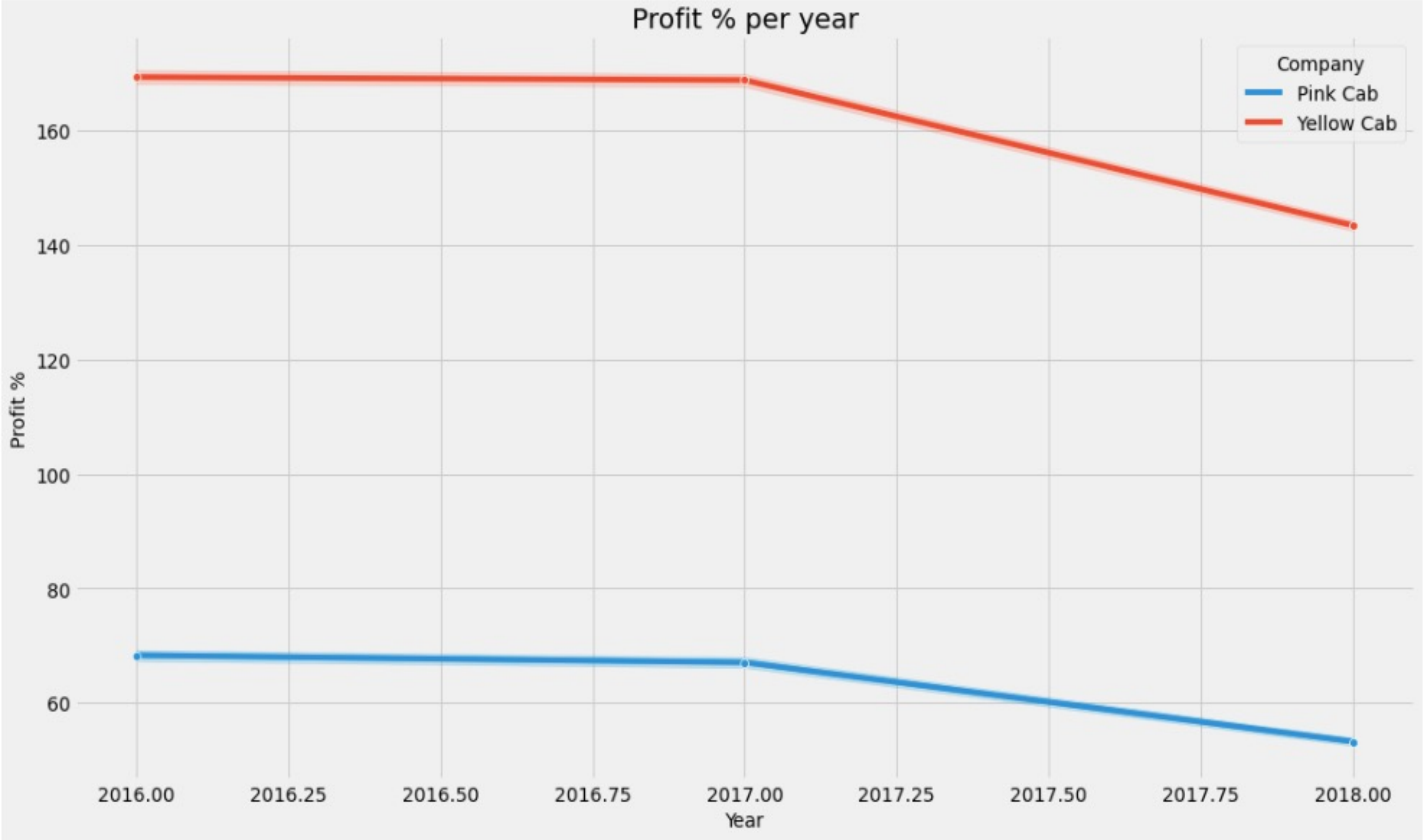
WHICH COMPANY MAKES MORE PROFIT?

PROFIT/MARGIN



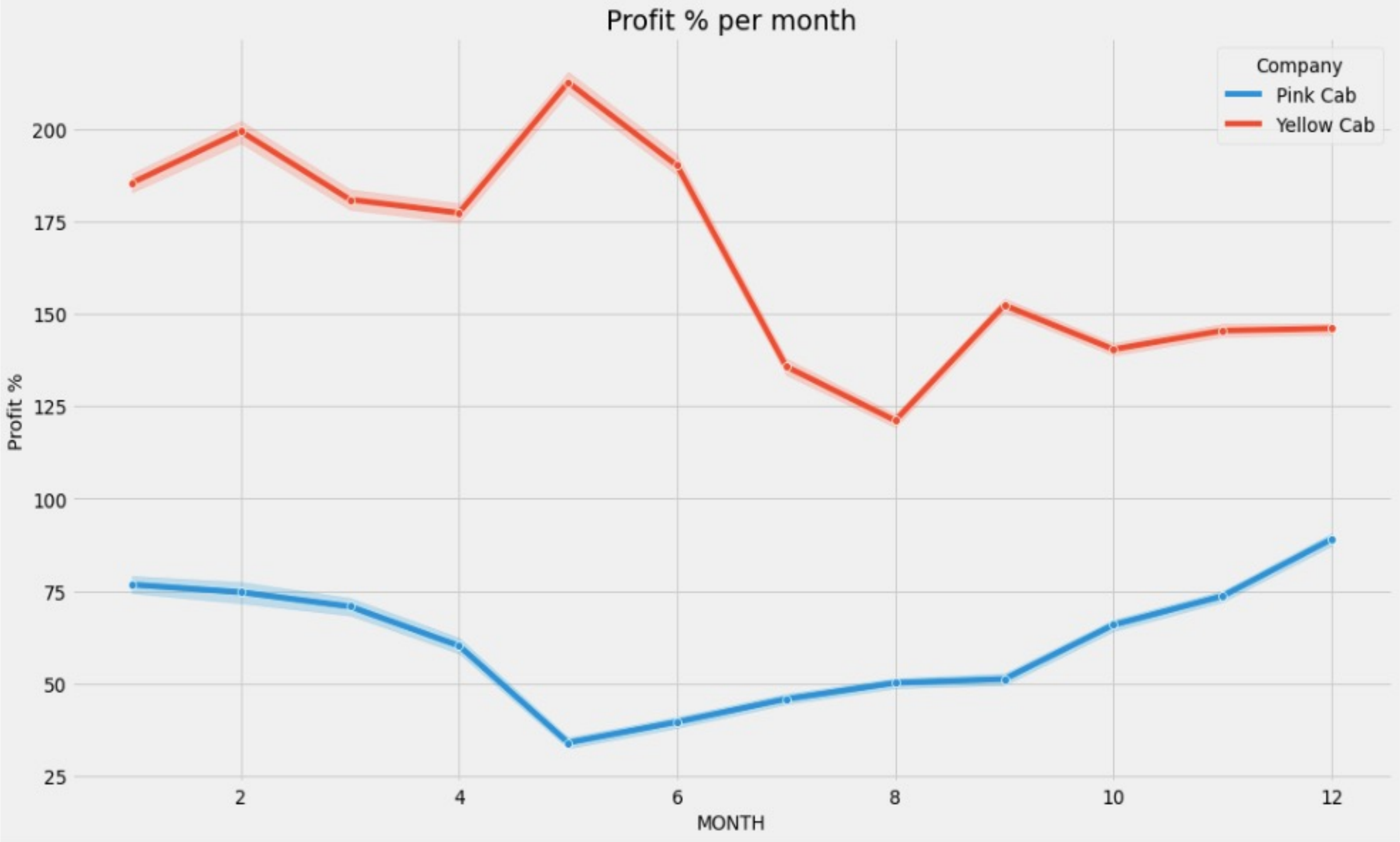
- ❖ Yellow cabs makes more profit as compared to Pink Cabs.

YEARLY TREND OF PROFITS



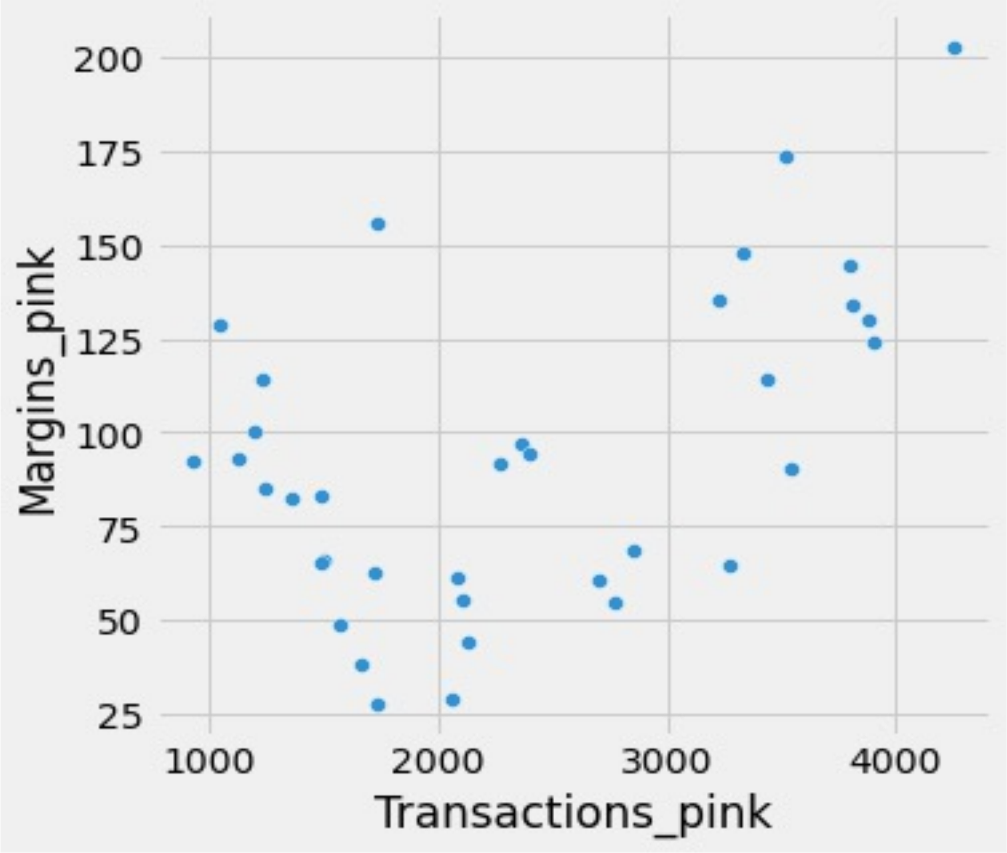
- Yellow cabs makes more profit as compared to Pink Cabs.
- Profits declined for both companies from 2017 to 2018.

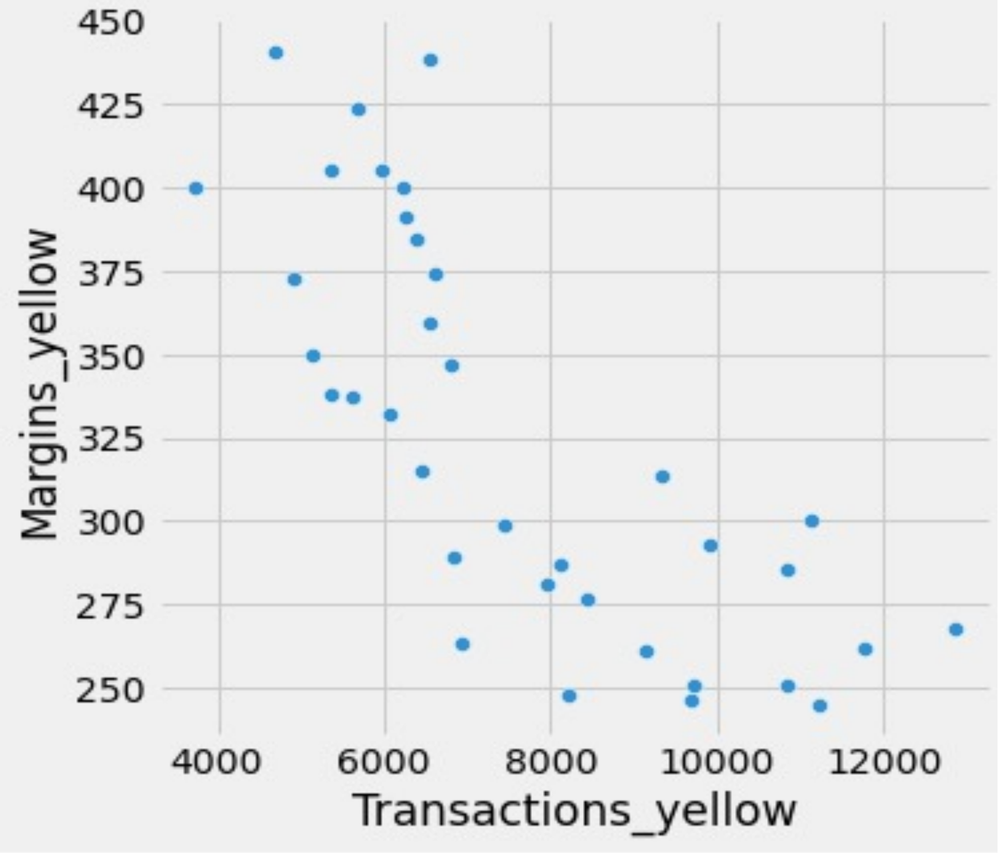
MONTHLY TREND OF PROFITS



- ❖ Profit of Pink cabs have increased after may month but reduced for yellow cabs.
- ❖ Overall profit % for yellow cab is higher then pink cab

PROFIT W.R.T. NUMBER OF TRANSACTIONS





- ❖ Pink cab increases the profit with increase in number of transactions but yellow cab reduces the profit with increase in transactions

Data Glacier

Your Deep Learning Partner

EDA SUMMARY

EDA SUMMARY

YELLOW CABS

- Monthly number of transactions increases from may to December and suddenly reduces in January.
- There is not huge deviation in number of transactions each year but 2017 has most number of transactions.
- Yellow cabs charges users between 0 to 1275.(Outliers present but ignoring it.)
- Yellow cabs have more male users as compared to females.
- Yellow cabs charges more in New York City as compared to other cities.
- Yellow cabs makes more profit (140-160%)
- Profit reduces from may to December.
- Yellow cabs decreases profit % as number of transactions increases.

PINK CABS

- Monthly number of transactions increases from may to December and get suddenly reduces in January.
- There is not huge deviation in number of transactions each year but 2017 has most number of transactions.
- Pink cabs charges between 0 to 825. (Outliers present but ignoring it.)
- Pink cabs have more male users as compared to females.
- Pink cabs charges same across different cities.
- Pink cabs makes less profit as compared to yellow cabs.(50-60%)
- Profit reduces from may to December.
- Pink cabs increases profit % as number of transactions increases

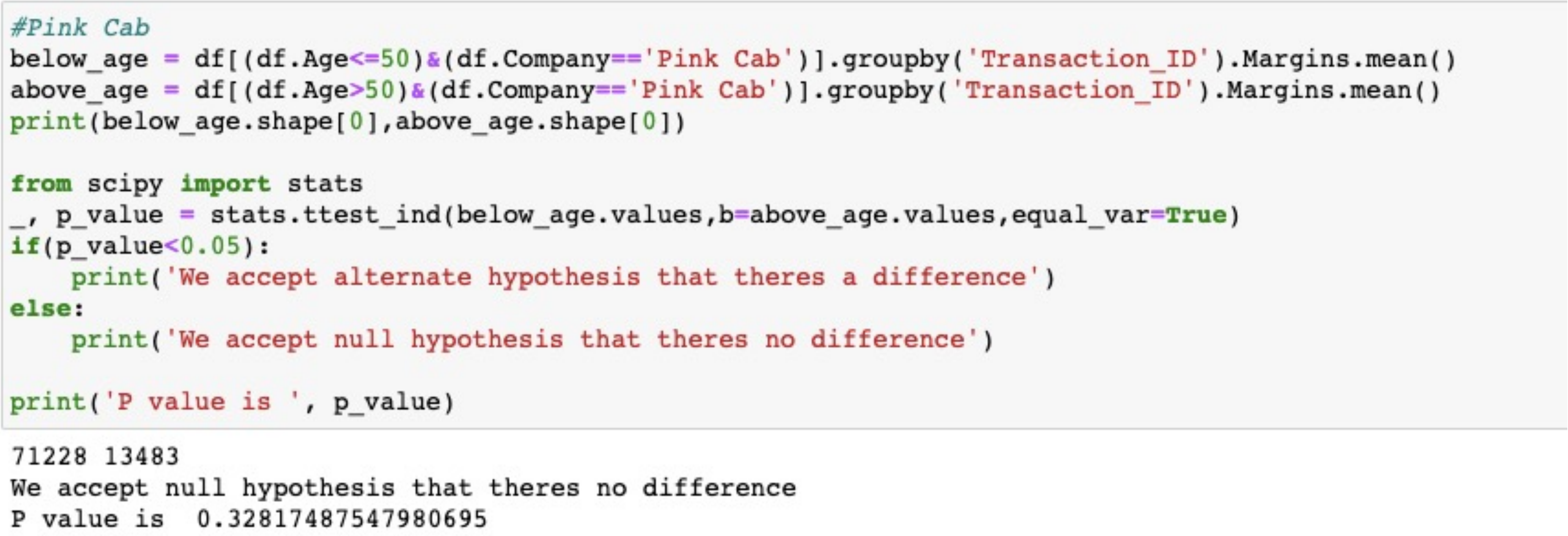
Data Glacier

Your Deep Learning Partner

HYPOTHESIS TESTING

HYPOTHESIS 1

❖ **HYPOTHESIS 1: 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
below_age = df[(df.Age<=50)&(df.Company=='Pink Cab')].groupby('Transaction_ID').Margins.mean()
above_age = df[(df.Age>50)&(df.Company=='Pink Cab')].groupby('Transaction_ID').Margins.mean()
print(below_age.shape[0],above_age.shape[0])

from scipy import stats
_, p_value = stats.ttest_ind(below_age.values,b=above_age.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)
```

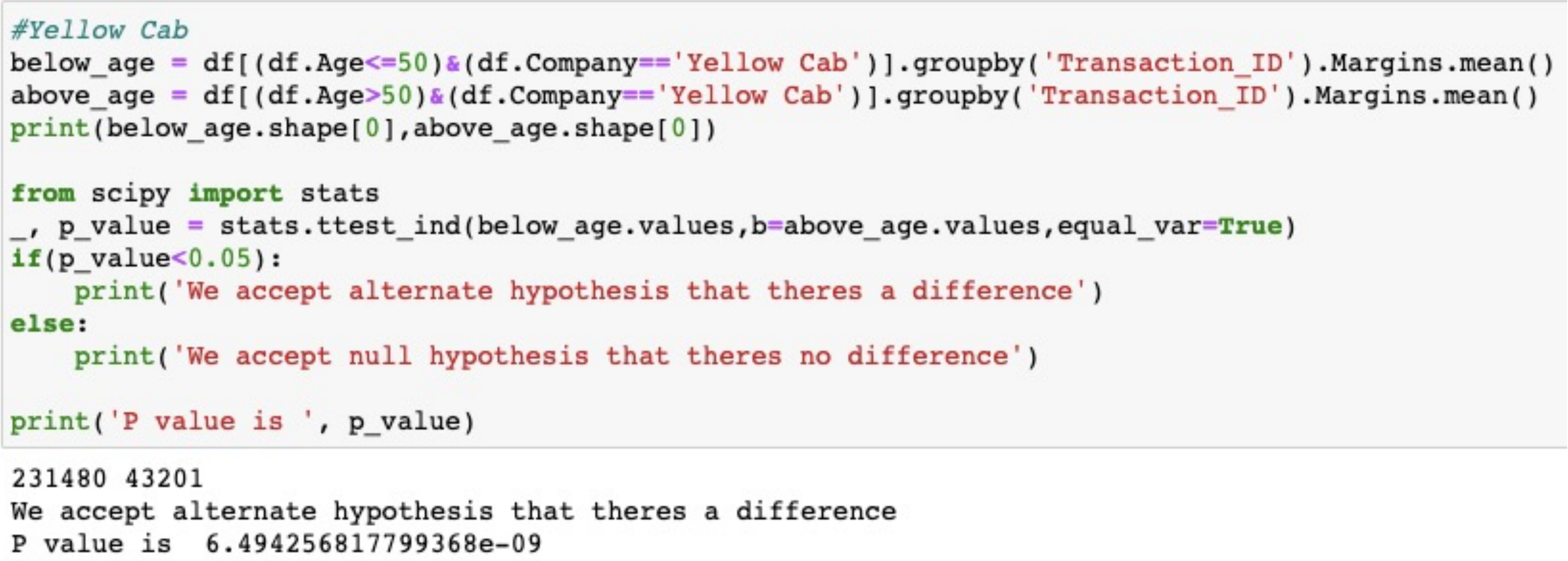
```
71228 13483
We accept null hypothesis that theres no difference
P value is  0.32817487547980695
```

HYPOTHESIS 1

- **HYPOTHESIS 1: 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.

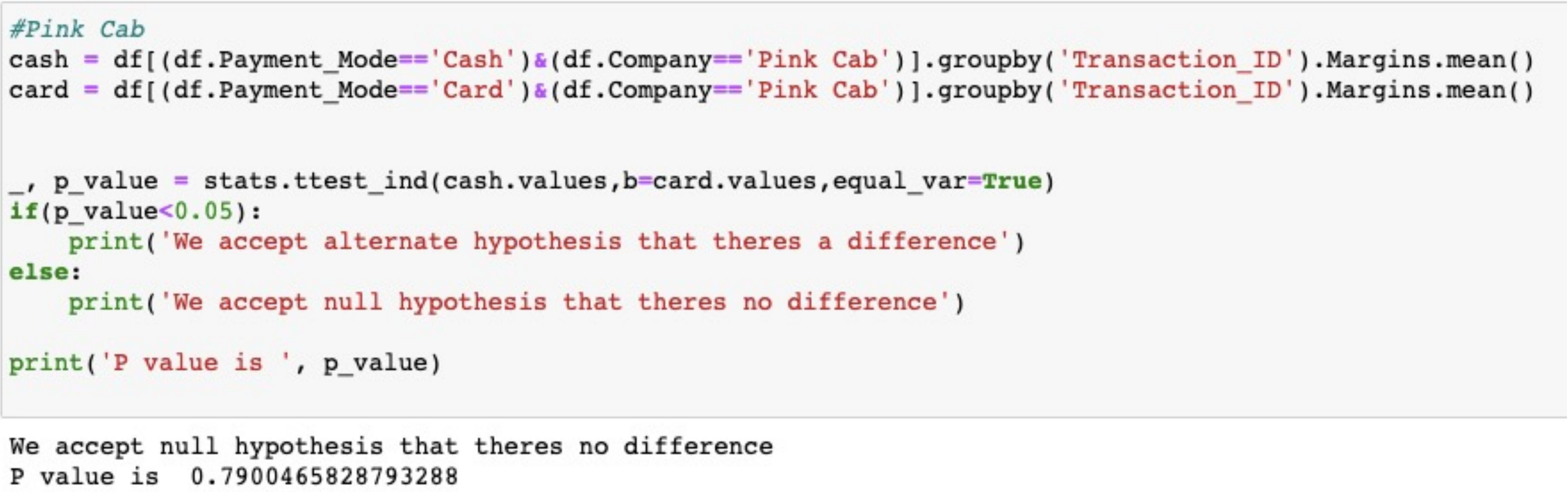


```
#Yellow Cab
below_age = df[(df.Age<=50)&(df.Company=='Yellow Cab')].groupby('Transaction_ID').Margins.mean()
above_age = df[(df.Age>50)&(df.Company=='Yellow Cab')].groupby('Transaction_ID').Margins.mean()
print(below_age.shape[0],above_age.shape[0])

from scipy import stats
_, p_value = stats.ttest_ind(below_age.values,b=above_age.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)
```

```
231480 43201
We accept alternate hypothesis that theres a difference
P value is  6.494256817799368e-09
```

HYPOTHESIS 2

❖ **HYPOTHESIS 2 : 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.

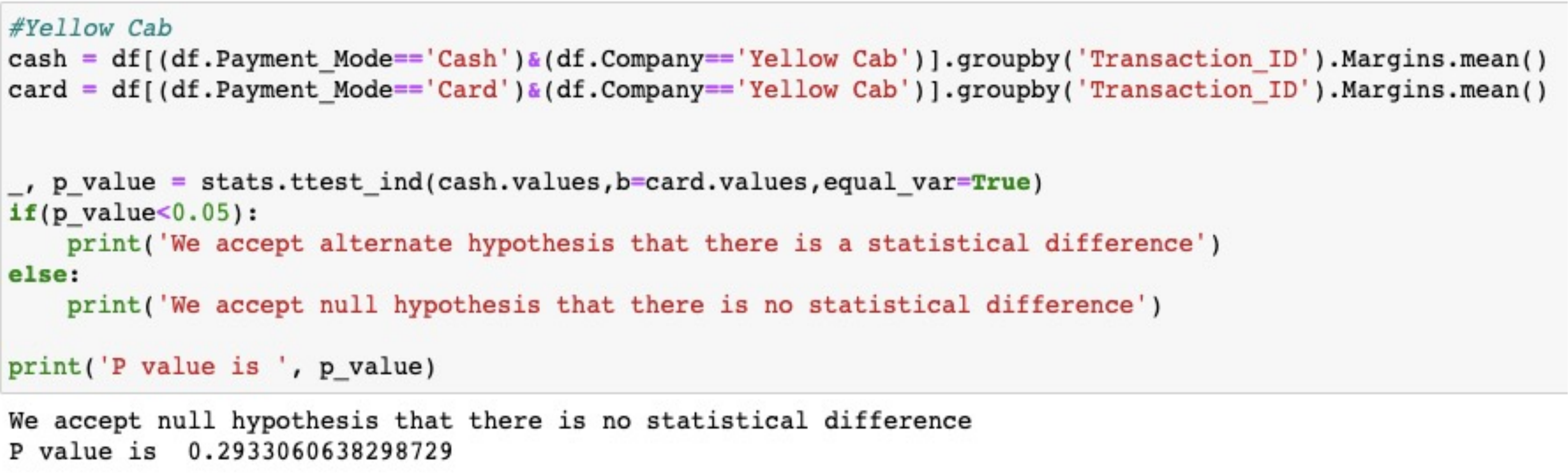


```
#Pink Cab
cash = df[(df.Payment_Mode=='Cash')&(df.Company=='Pink Cab')].groupby('Transaction_ID').Margins.mean()
card = df[(df.Payment_Mode=='Card')&(df.Company=='Pink Cab')].groupby('Transaction_ID').Margins.mean()


_, p_value = stats.ttest_ind(cash.values,b=card.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)
```

```
We accept null hypothesis that theres no difference
P value is  0.7900465828793288
```

HYPOTHESIS 2

❖ **HYPOTHESIS 2 : 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.

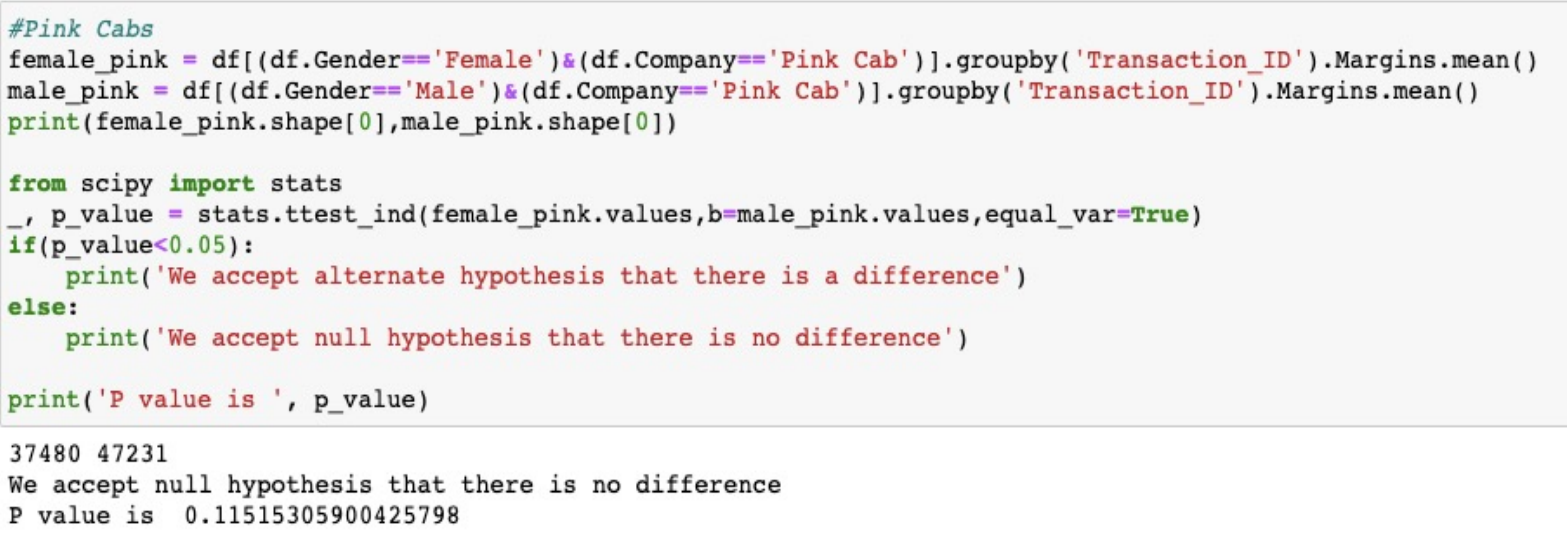


```
#Yellow Cab
cash = df[(df.Payment_Mode=='Cash')&(df.Company=='Yellow Cab')].groupby('Transaction_ID').Margins.mean()
card = df[(df.Payment_Mode=='Card')&(df.Company=='Yellow Cab')].groupby('Transaction_ID').Margins.mean()


_, p_value = stats.ttest_ind(cash.values,b=card.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 accept null hypothesis that there is no statistical difference
P value is  0.2933060638298729
```

HYPOTHESIS 3

- **HYPOTHESIS 3: 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.

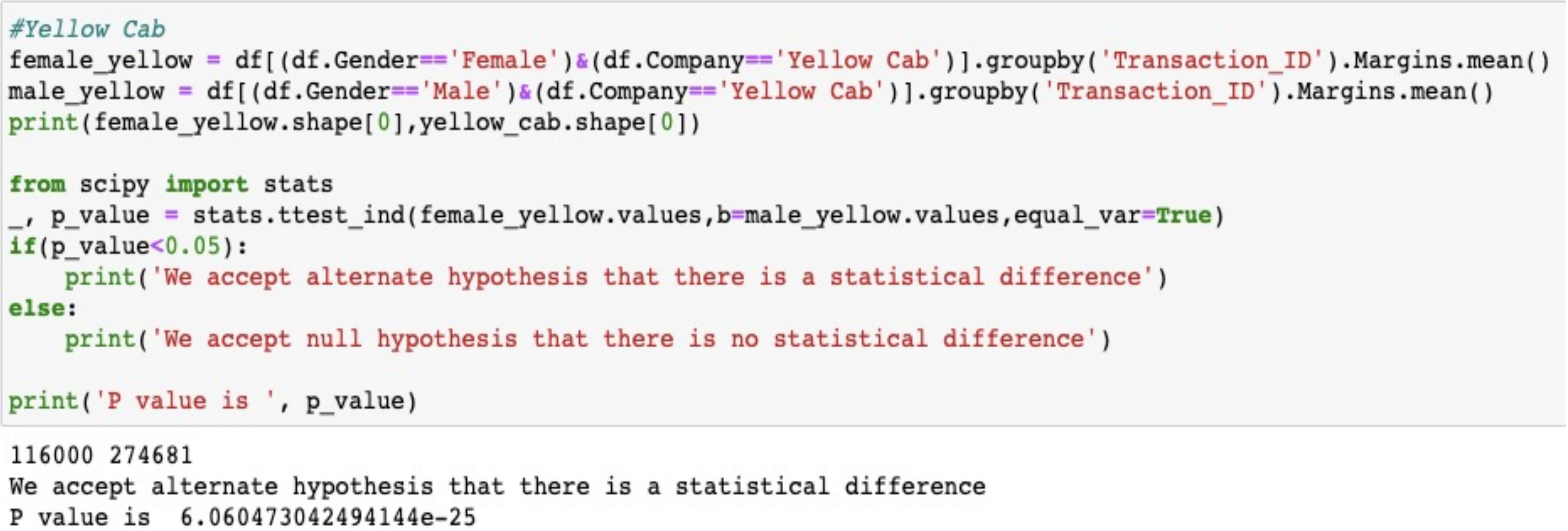


```
#Pink Cabs
female_pink = df[(df.Gender=='Female')&(df.Company=='Pink Cab')].groupby('Transaction_ID').Margins.mean()
male_pink = df[(df.Gender=='Male')&(df.Company=='Pink Cab')].groupby('Transaction_ID').Margins.mean()
print(female_pink.shape[0],male_pink.shape[0])

from scipy import stats
_, p_value = stats.ttest_ind(female_pink.values,b=male_pink.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)
```

```
37480 47231
We accept null hypothesis that there is no difference
P value is  0.11515305900425798
```

HYPOTHESIS 3

❖ **HYPOTHESIS 3: 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.



```
#Yellow Cab
female_yellow = df[(df.Gender=='Female')&(df.Company=='Yellow Cab')].groupby('Transaction_ID').Margins.mean()
male_yellow = df[(df.Gender=='Male')&(df.Company=='Yellow Cab')].groupby('Transaction_ID').Margins.mean()
print(female_yellow.shape[0],yellow_cab.shape[0])

from scipy import stats
_, p_value = stats.ttest_ind(female_yellow.values,b=male_yellow.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)
```

```
116000 274681
We accept alternate hypothesis that there is a statistical difference
P value is  6.060473042494144e-25
```

HYPOTHESIS RESULTS

❖ **HYPOTHESIS 1: Is there any difference in Profit regarding Age**

•Pink cab does not have any difference in profits based on age.
•Yellow cab provides discount for users above 50 years of age.

❖ **HYPOTHESIS 2 : Is there any difference in Profit regarding Payment mode**

•There is no difference in profits based on payment methods for both the companies

❖ **HYPOTHESIS 3: Is there any difference in profit regarding Gender**

•Pink cab does not have any difference in profits charged based on Gender.
•Yellow cab provides discount to Female users as compared to male users.

Data Glacier

Your Deep Learning Partner

RECOMMENDATIONS

RECOMMENDATIONS

Post Analysis it can be concluded that yellow cab company is much better as compared to pink cab company.

❖ **CLIENT ANALYSIS :**

- **AGE :** In Yellow Cab company there is difference in prices for people older than 50 years, whereas in Pink Cab there is no difference for all age groups.
- **GENDER:** Yellow cabs company also provided discount to female customer
- **PAYMENT MODE :** There is no difference in profits based on payment mode for both companies

❖ **PROFIT ANALYSIS :**

- **CITY :** Pink cabs charges more profit from new York city while pink city keeps it same across all cities.
- **TRANSACTIONS :** Yellow cabs have more number of transactions almost 3 times as compared to pink cabs and yellow cabs also reduced profit percentage with increase in transactions.
- **PROFIT:** Yellow cabs reduces the profit percent with number of transactions but still each month it has more profit percent as compared to Pink cabs.

❖ **On the basis of above points , I recommend Yellow cab for investment.**