G2M insight for Cab Investment firm

Company Name: XYZ
Location: US
Date: 21-07-2023

Agenda

Data Glacier
Your Deep Learning Partner

Summary:

- 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.

- XYZ is interested in using your actionable insights to help them identify the right company to make their investment.

- Cab companies:

➢Yellow cab

➢Pink Cab

Data Preparation:

- There are 4 datasets:

➢Cab_Data.csv : this file contains transaction details for 2 cab companies.

➢Transaction_ID.csv : this dataset contains transaction data including payment mode.

➢Customer_ID.csv : this file contains unique customer details with identification numbers.

➢City.csv : this file contains list of US cities and population with the number of cab users.

Distribution of Kilometres travelled

From the above graph we can derive that cab rides ranges between distance of 2 to 48 kms. Also count of long-distance rides drastically decreases after a ride distance of 40 kms.
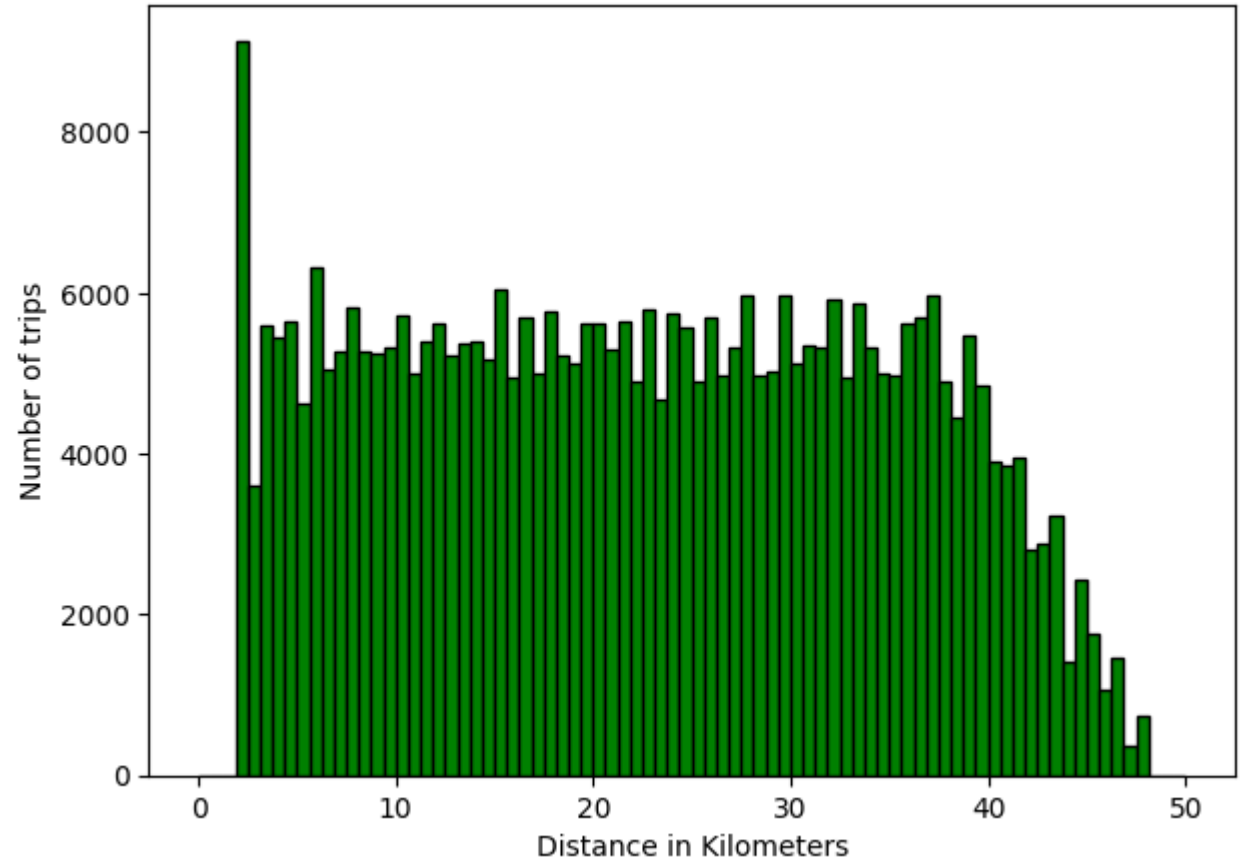
Count of Pink cab rides are about 80000 and Yellow cab rides is over 2.5 lakhs.
Also, the city with the largest number of cab rides is New York.
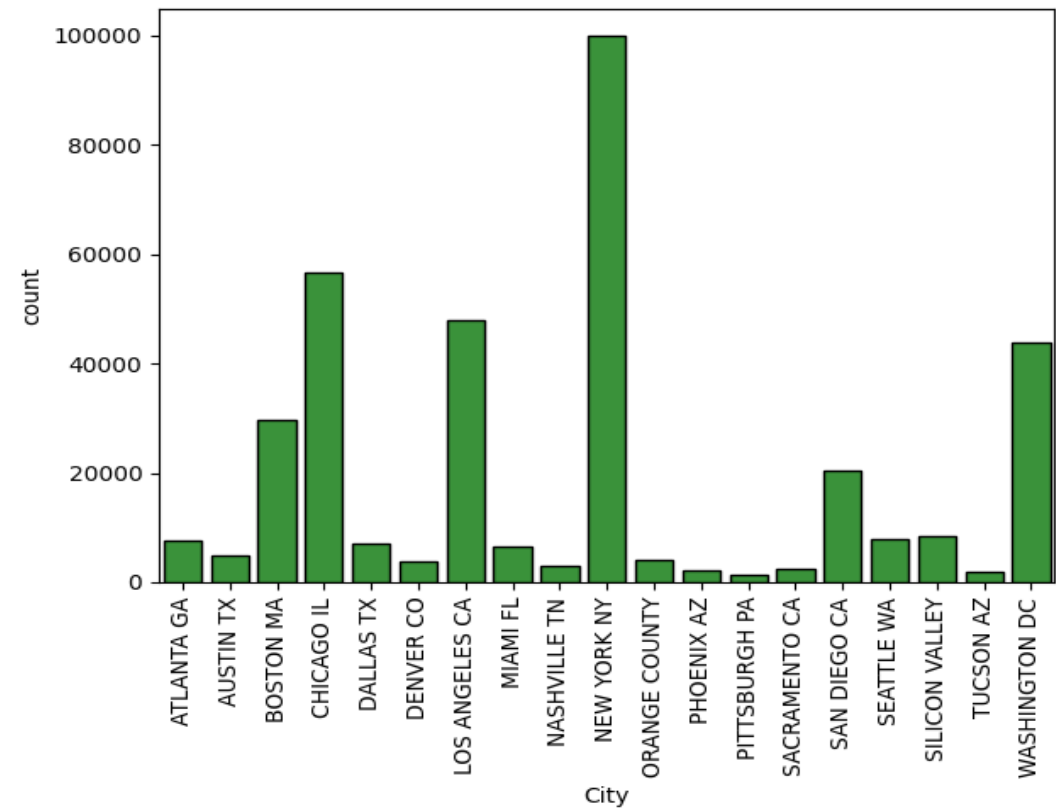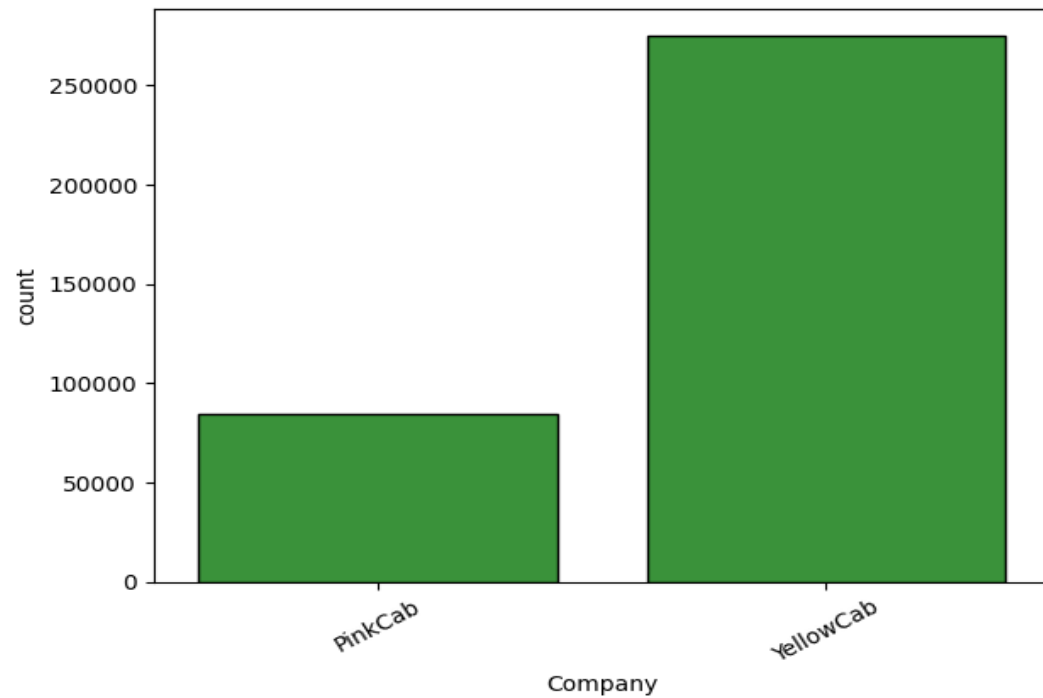
Number of card payments is over 2 lakhs whereas cash transactions are about 1.5 lakhs.
Number of Male passengers is over 2 lakhs, greater than Female passengers by about 50 thousand.
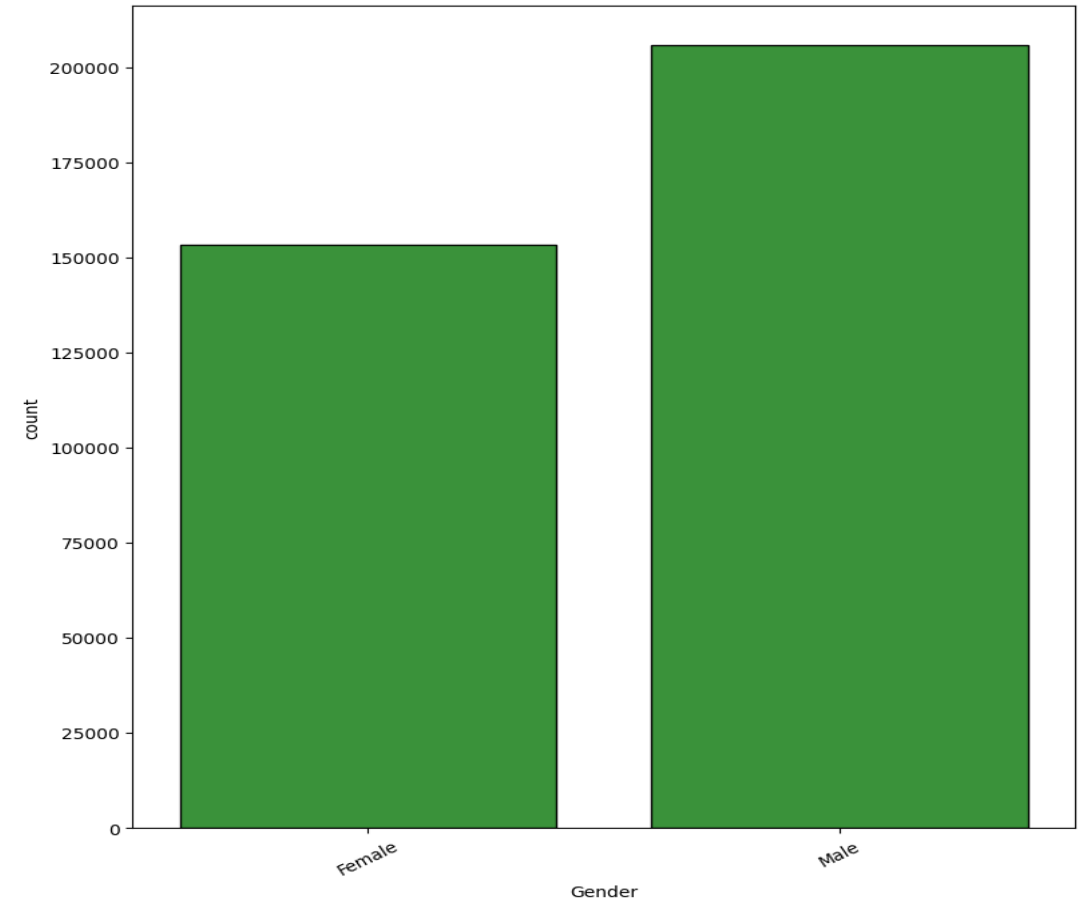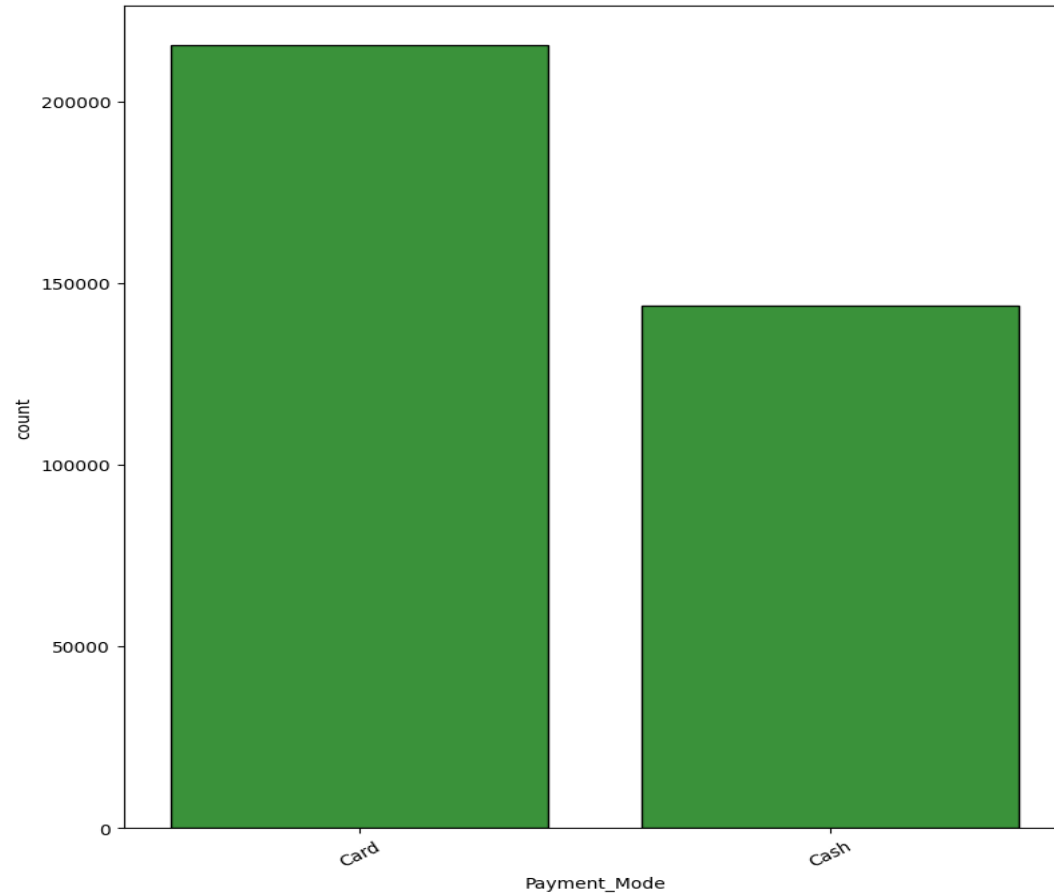
Price charged in Yellow cabs is higher than price charged in Pink cabs. Cost of trips in Yellow cabs is higher than price charged in Pink cabs.

Distributions of Variables
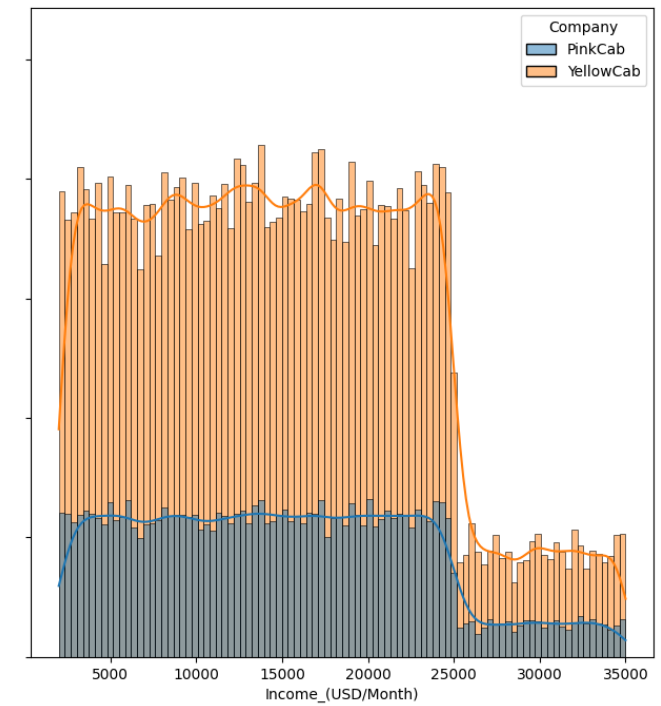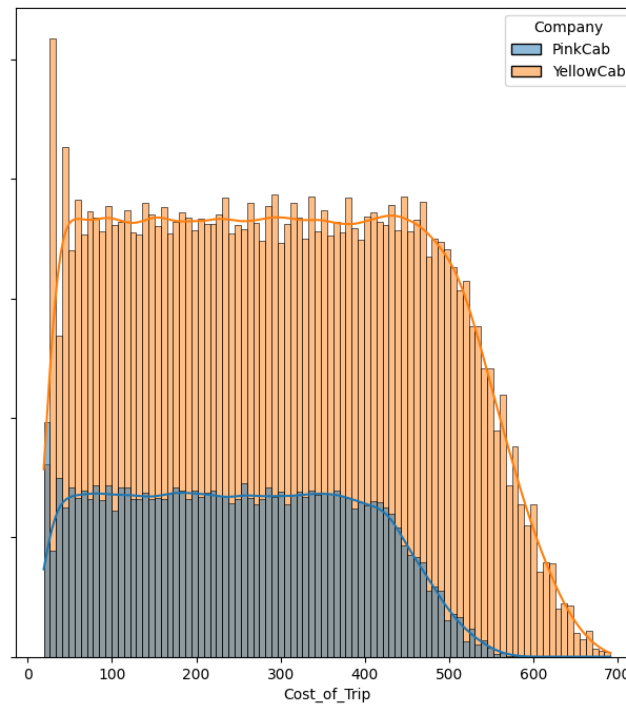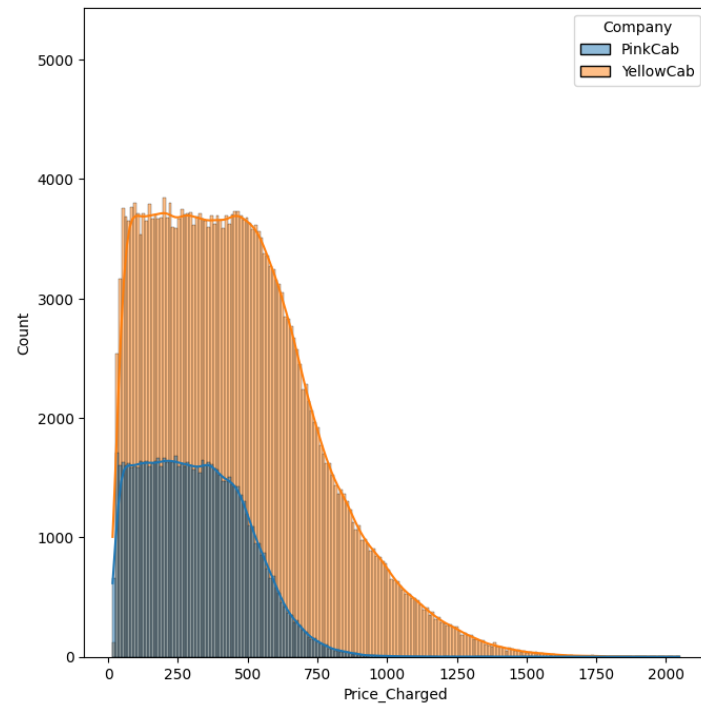
Yellow cabs have a higher profit margin than Pink cabs. Mean cost in Yellow cabs is higher than mean cost in Pink cabs.
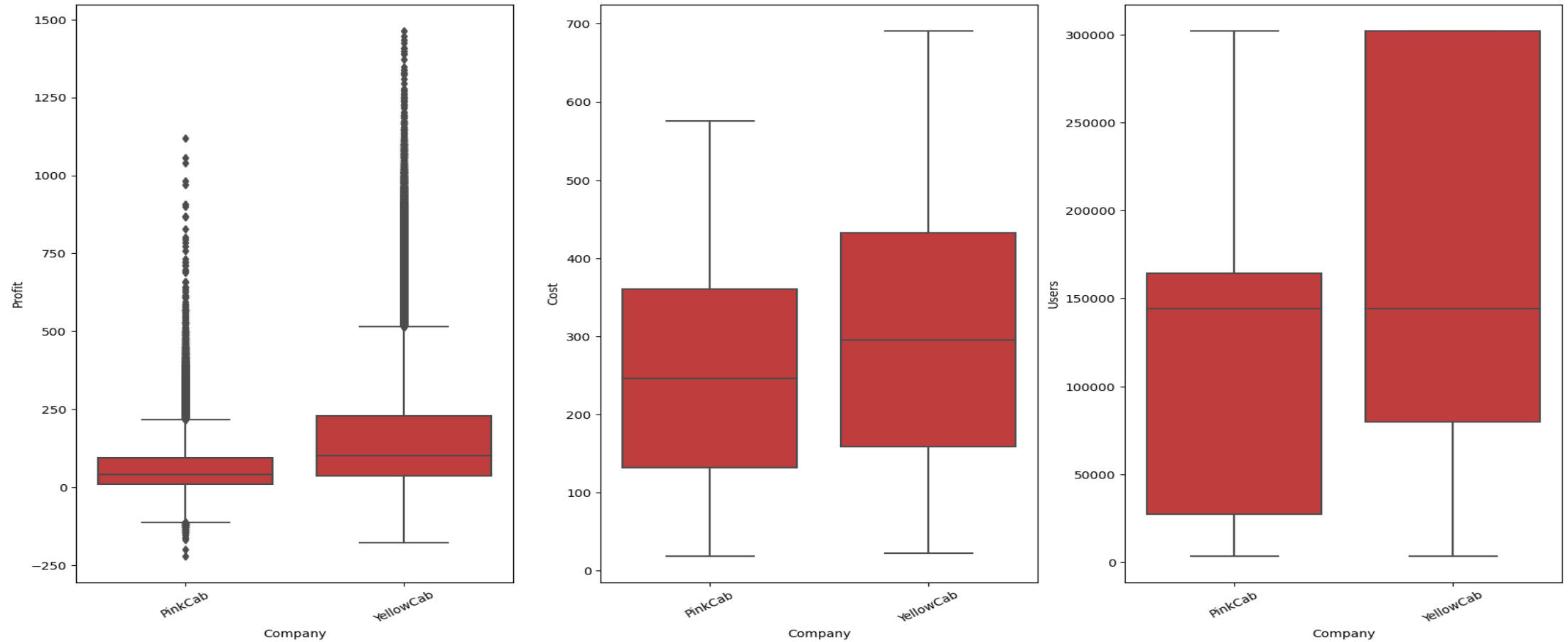
December has the highest number of cab users while in the month of February we can observe least number of cab users.

Distributions of Variables
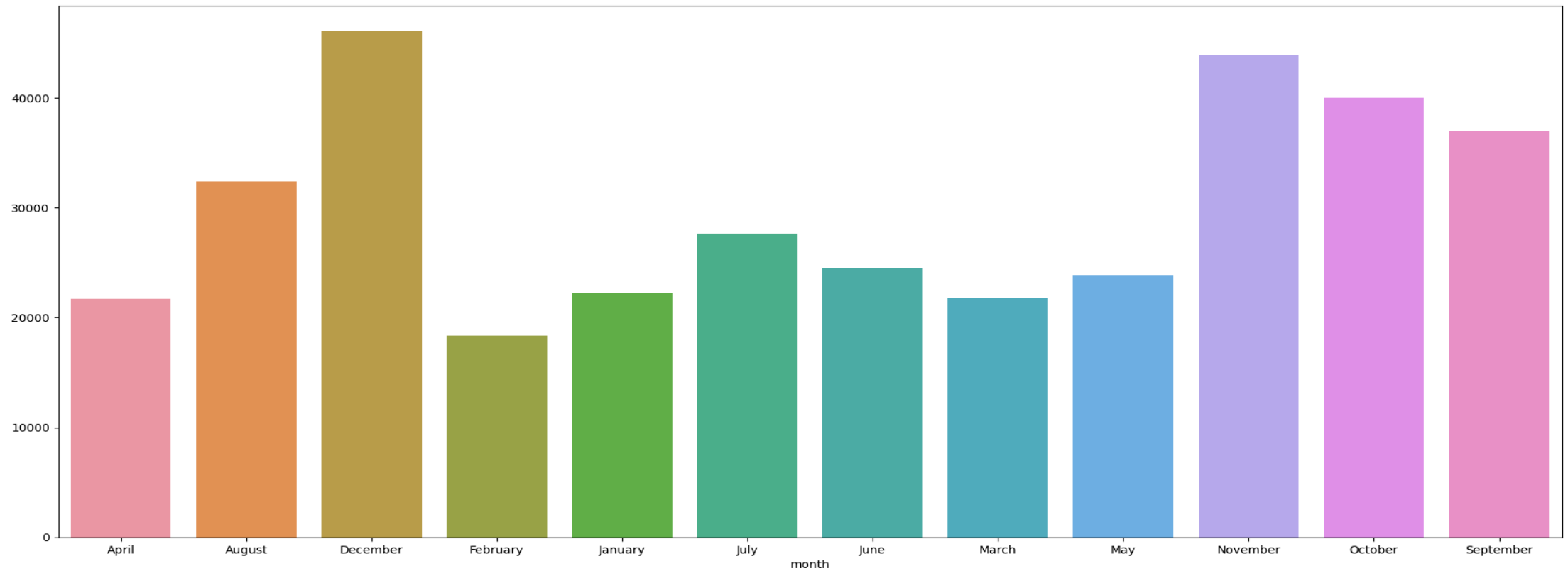
Thursday seems to be the busiest day for cab companies while Tuesday, Sunday and Monday seem to be least busiest days.

Distributions of Variables
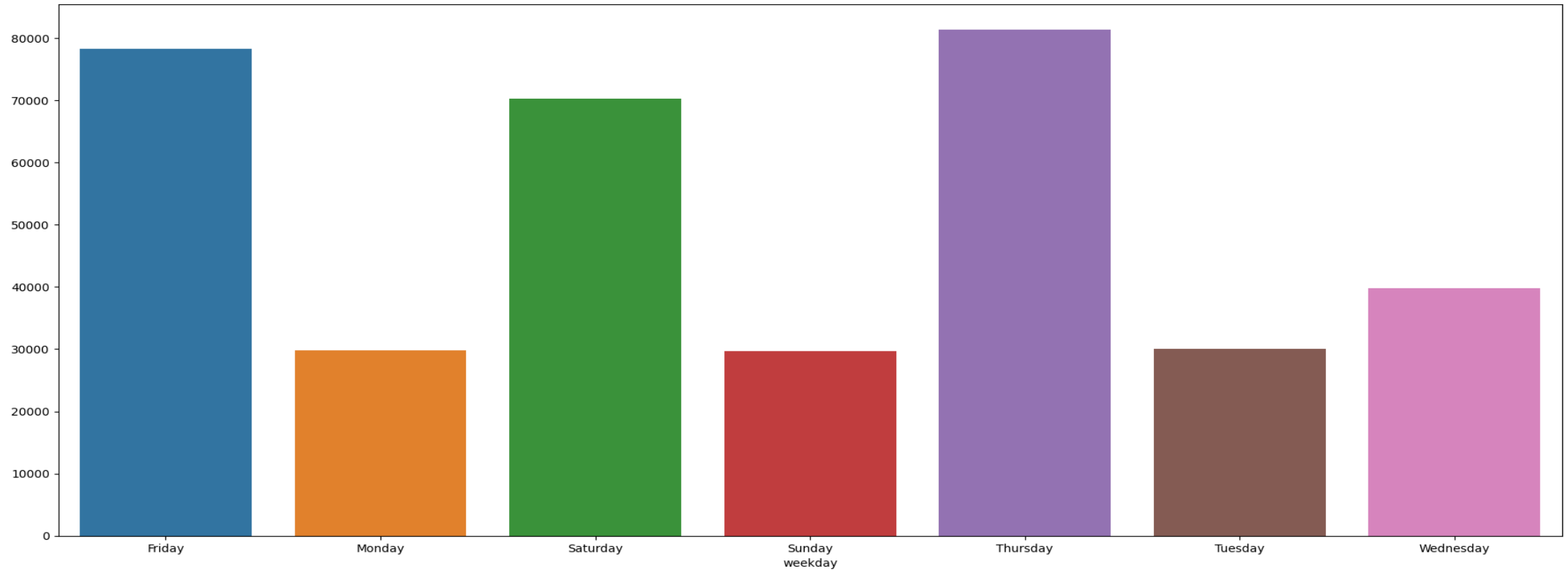
There seems to be positive correlation between Profit and Price Charged
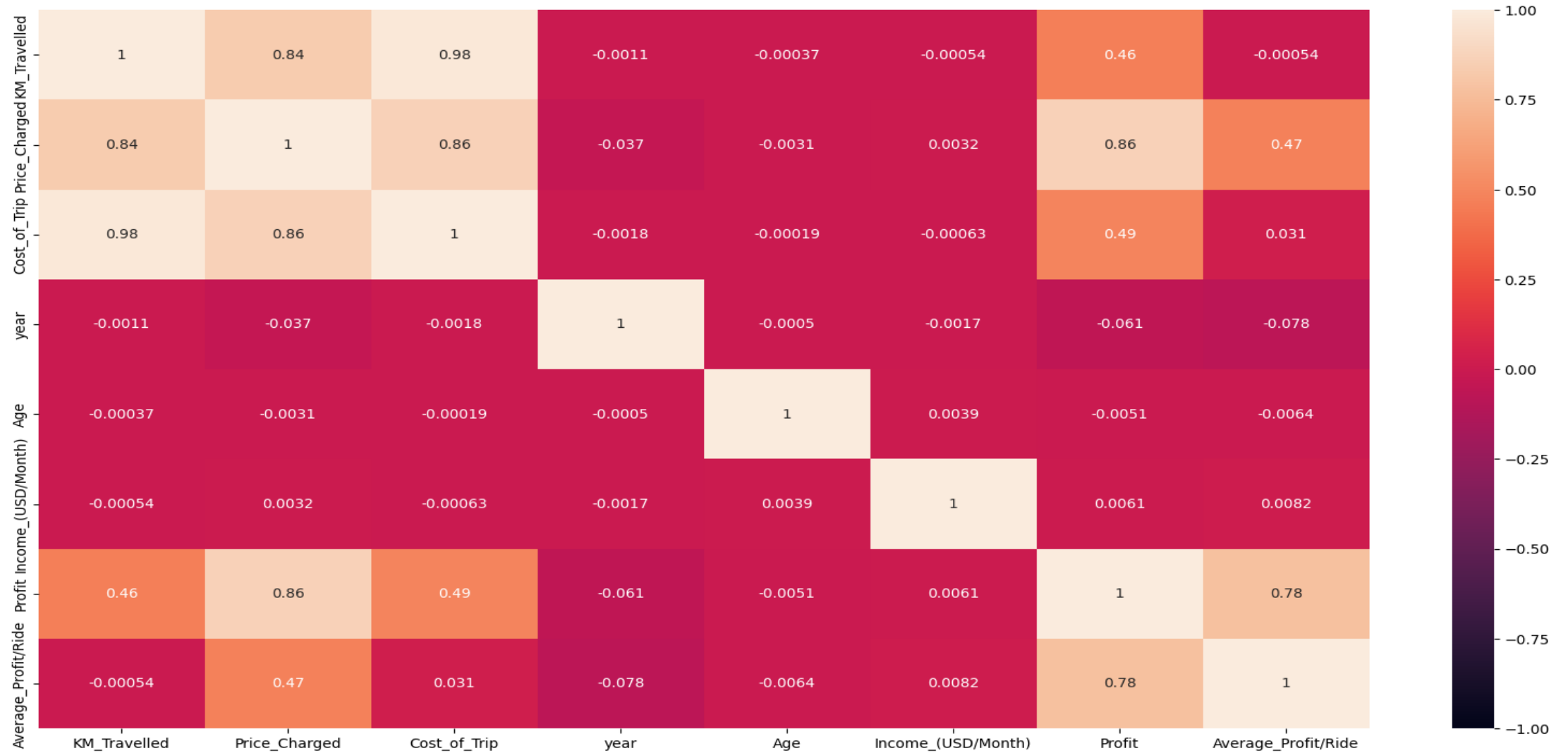
Transaction per Year for both Cabs:
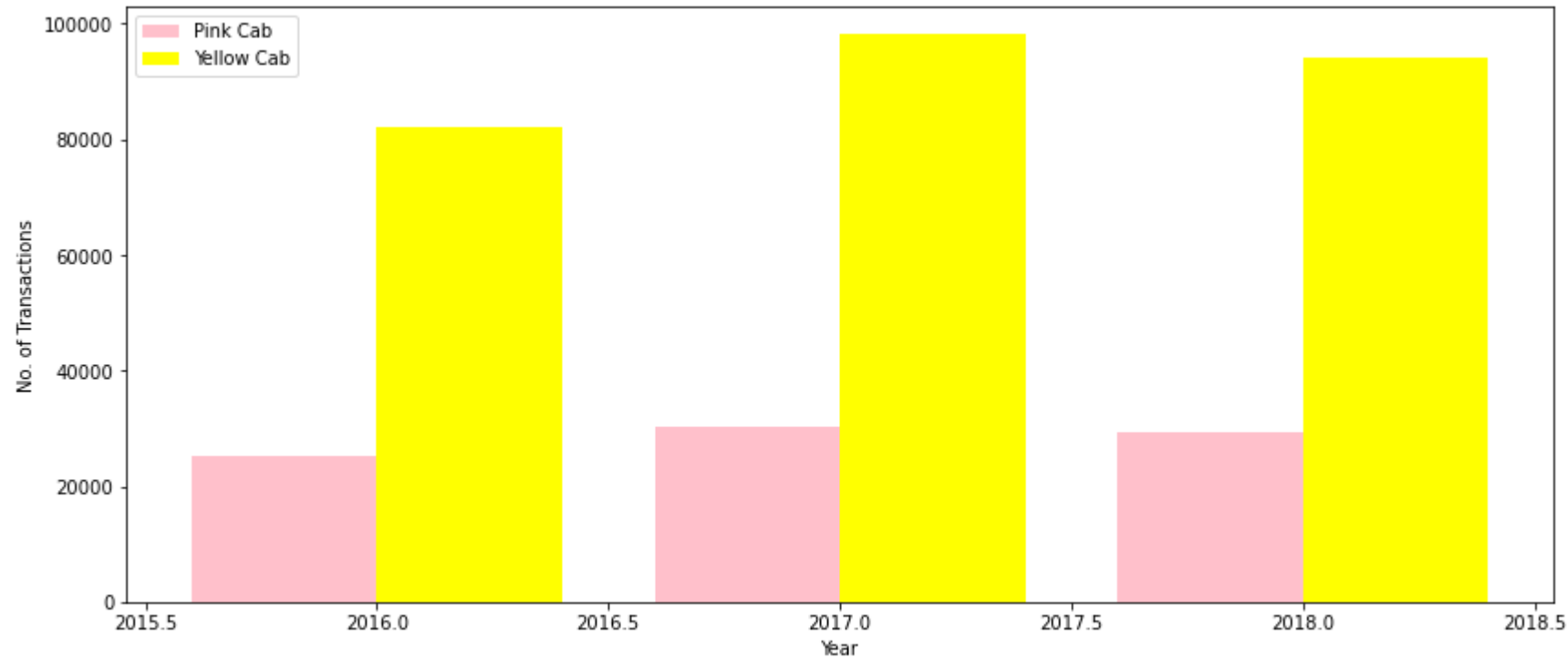
- From the graph it shows that on yearly basis no. of transactions for Yellow cab is higher than Pink cab.

Cab Users per City:
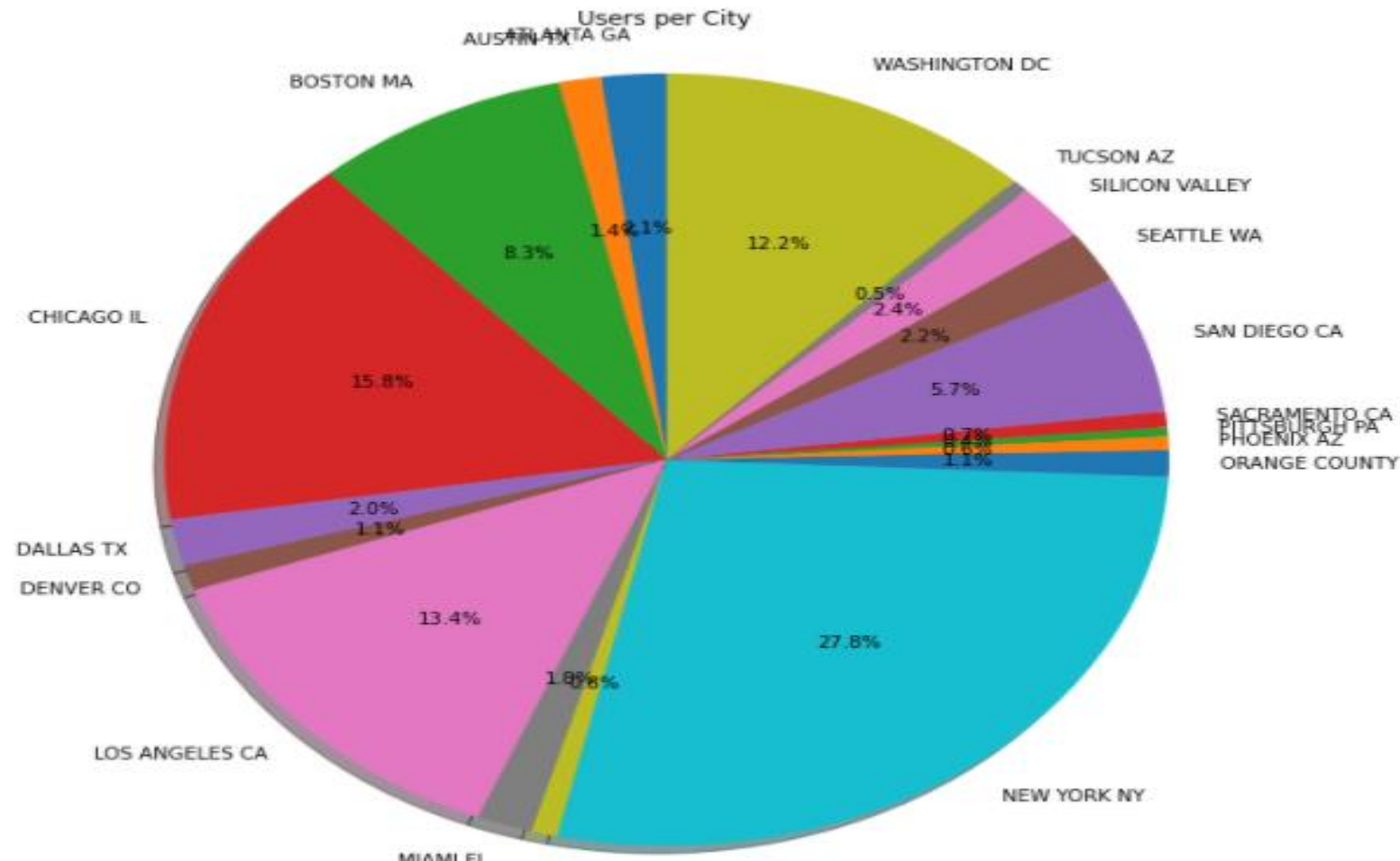
- New York City has the highest Cab users with 28% followed by Chicago with 16% and Los Angeles with 13%

Price Charged per Gender for both Cabs:
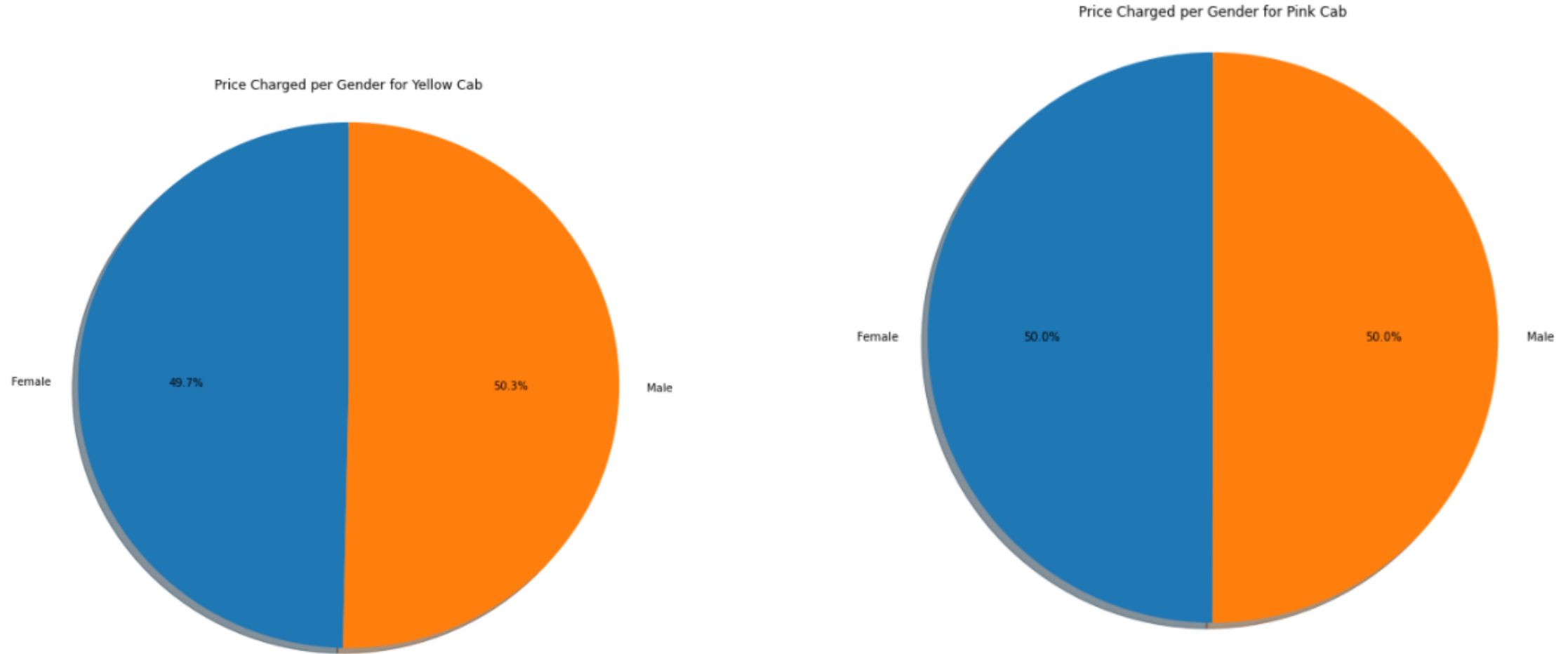
- Yellow Cab charge less from Female Customers whereas Pink Cab charges same for both Male and Female Customers.

Customer Share per Gender for both Cabs:

Customer share per gender per cab

('Pink Cab', 'Female') 20.5%

('Yellow Cab', 'Male') 29.8%

('Pink Cab', 'Male') 24.2%

('Yellow Cab', 'Female') 25.5%
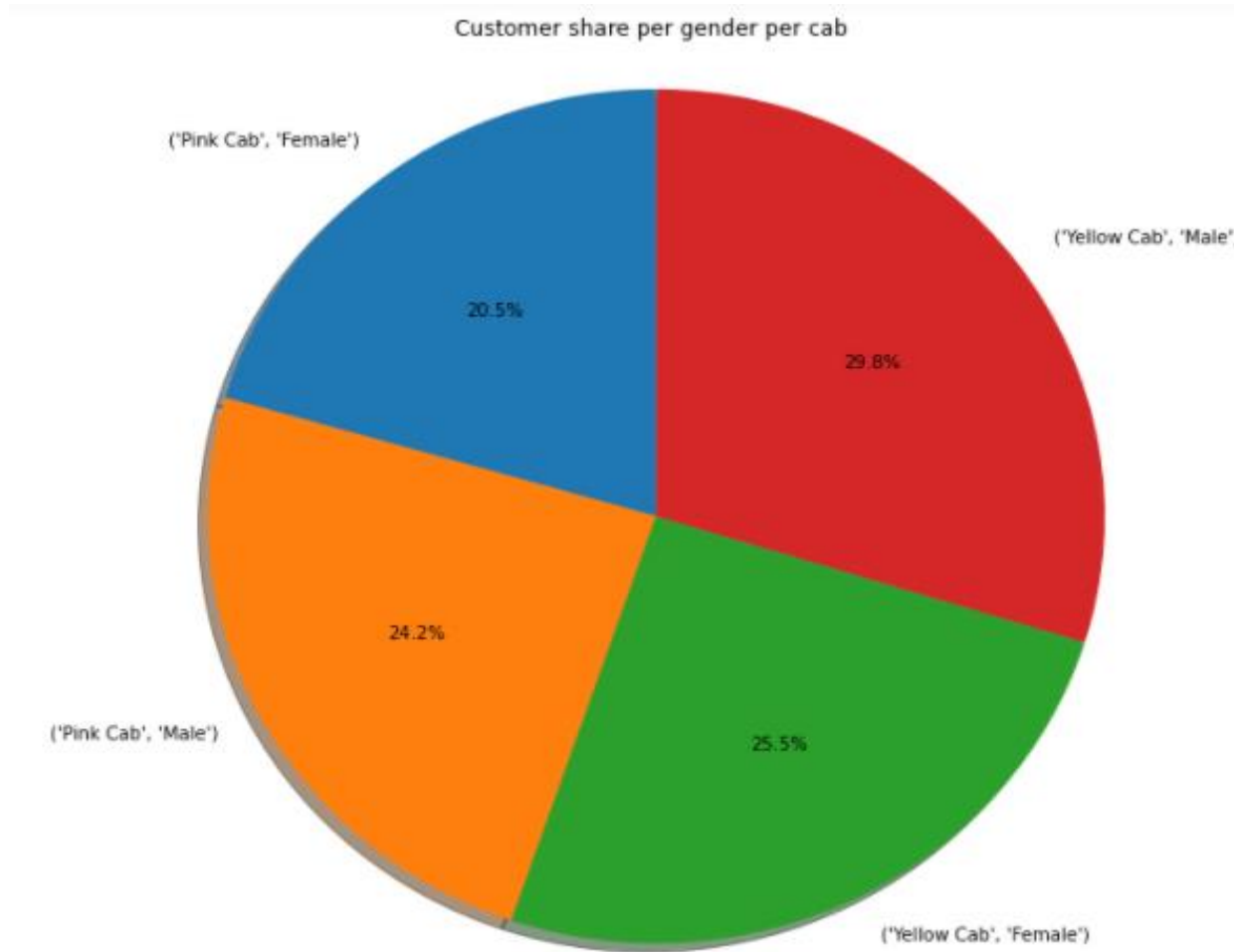
- Female Customers in Yellow Cab(25.5%) is higher compared to Pink Cab (20.5%)

Profit Margin per year for both Cabs:
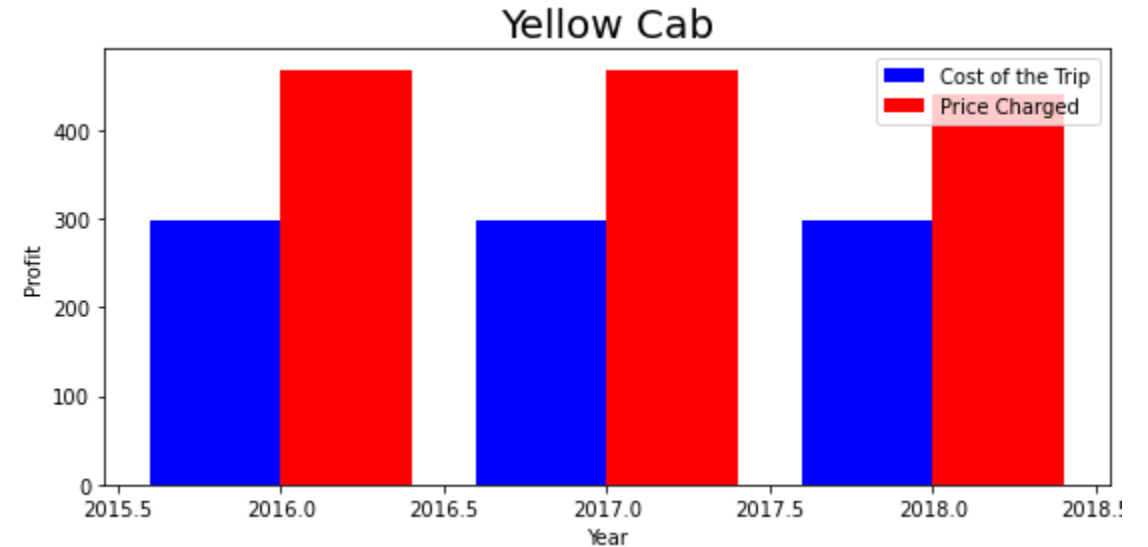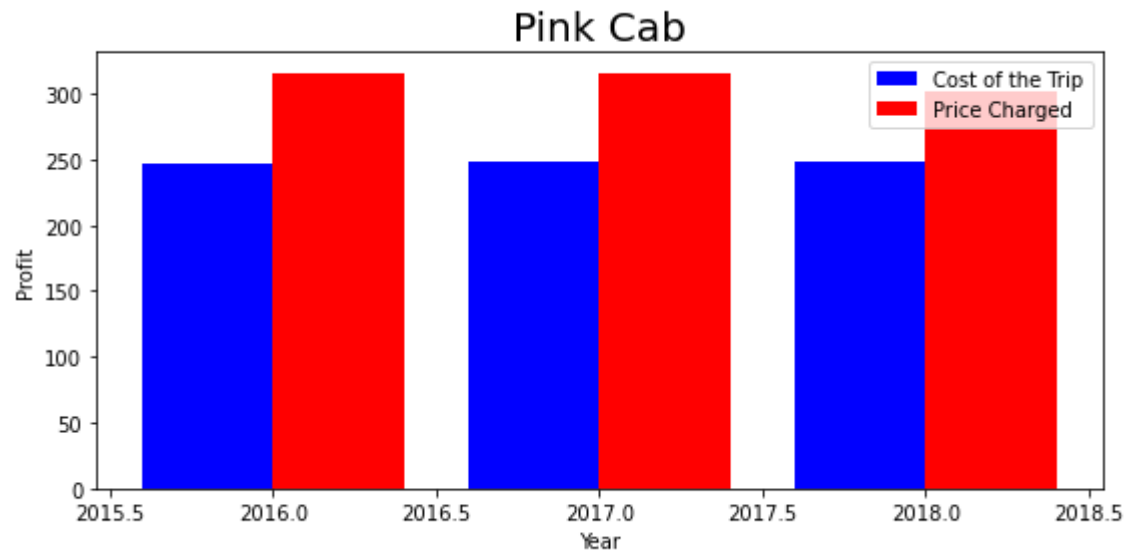
- From the Graphs, it shows that the Yellow cab has a higher Profit Margin (Price Charged - Cost of Trip) compared to Pink cab.

Pink Cab

- Rides are in the range of approximately 2 to 48 KM.

- Price Charge range from 150 to 450 dollars.

- In December which is the holiday season, no. of travels was around 11000.

- Transaction per year:
  - 2016: 20000 – 40000
  - 2017: 20000 – 40000
  - 2018: 20000 – 40000

- All the cities have the same increase in price charge with increase in distance.

Yellow Cab

- Rides are in the range of approximately 2 to 48 KM.

- Price Charge range from 250 to 600 dollars.

- In December which is the holiday season, no. of travels was around 35000.

- Transaction per year:
  - 2016: 80000 – 100000
  - 2017: 80000 – 100000
  - 2018: 80000 – 100000

- In New York City the Price charged for Yellow Cab is more in comparison to the other cities.

Pink Cab

- Pink Cab charges same for both Male and Female Customers.

- Female customers are around 20.5% out of the total Customers.

- Profit Margin is low each year (2016-2018) compared to Yellow Cab.

- Pink Cabs increase margins with increase in number of Transactions.

Yellow Cab

- Yellow Cab charge less from Female Customers.

- Female customers are around 25.5% out of the total Customers.

- Profit Margin is high each year (2016-2018) compared to Pink Cab.

- Yellow Cab decrease Margins with the increase in Transaction.

Observations of One-Way ANOVA Analysis:

- There is difference in Profit Margin for Male and Female users
- There is difference in Profit Margin for Pink Cab company and Yellow Cab company.
- There is no evidence to say that there is any difference in profit margin when different payment modes are concerned.

Model1: Linear Regression

❑ Linear Regression is a method for predicting target value and attempts to model the linear relationship between target and one or more predictors.

❑ In our dataset, Price Charge is the target value, and all the other variables are predictors.

Splitting the data into a training set (75%), and test set (25%).

Yellow Cab

Pink Cab

```
X_train.info()

<class 'pandas.core.frame.DataFrame'>
MultiIndex: 192276 entries, (10033146, 559, 'NEW YORK NY') to (10194745, 58301, 'BOSTON MA')
Data columns (total 6 columns):
 #   Column             Non-Null Count   Dtype
---  ------             --------------   -----
 0   KM_Travelled       192276 non-null  float64
 1   Cost_of_Trip       192276 non-null  float64
 2   Month              192276 non-null  int64
 3   Year               192276 non-null  int64
 4   Age                192276 non-null  int64
 5   Income_(USD/Month) 192276 non-null  int64
dtypes: float64(2), int64(4)
memory usage: 24.8+ MB
```

```
X_test.info()

<class 'pandas.core.frame.DataFrame'>
MultiIndex: 82405 entries, (10263934, 7987, 'LOS ANGELES CA') to (10226996, 3195, 'CHICAGO IL')
Data columns (total 6 columns):
 #   Column             Non-Null Count   Dtype
---  ------             --------------   -----
 0   KM_Travelled       82405 non-null  float64
 1   Cost_of_Trip       82405 non-null  float64
 2   Month              82405 non-null  int64
 3   Year               82405 non-null  int64
 4   Age                82405 non-null  int64
 5   Income_(USD/Month) 82405 non-null  int64
dtypes: float64(2), int64(4)
```

```
X_train.info()

<class 'pandas.core.frame.DataFrame'>
MultiIndex: 59297 entries, (10400049, 1517, 'NEW YORK NY') to (10085754, 13379, 'SILICON VALLEY')
Data columns (total 6 columns):
 #   Column             Non-Null Count   Dtype
---  ------             --------------   -----
 0   KM_Travelled       59297 non-null  float64
 1   Cost_of_Trip       59297 non-null  float64
 2   Month              59297 non-null  int64
 3   Year               59297 non-null  int64
 4   Age                59297 non-null  int64
 5   Income_(USD/Month) 59297 non-null  int64
dtypes: float64(2), int64(4)
memory usage: 17.6+ MB
```

```
X_test.info()

<class 'pandas.core.frame.DataFrame'>
MultiIndex: 25414 entries, (10184224, 46628, 'SACRAMENTO CA') to (10158114, 8037, 'LOS ANGELES CA')
Data columns (total 6 columns):
 #   Column             Non-Null Count   Dtype
---  ------             --------------   -----
 0   KM_Travelled       25414 non-null  float64
 1   Cost_of_Trip       25414 non-null  float64
 2   Month              25414 non-null  int64
 3   Year               25414 non-null  int64
 4   Age                25414 non-null  int64
 5   Income_(USD/Month) 25414 non-null  int64
```

Model2: Decision Tree

❑ **Decision tree** builds regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed.

❑ The final result is a tree with decision nodes and leaf nodes.

❑ The topmost decision node in a tree which corresponds to the best predictor for the target value (Price Charged).

Model3: Random Forest

❑ A **Random Forest** operates by constructing several **Decision trees.**

❑ A prediction from the **Random Forest** is an average of the predictions produced by the **Decision trees** in the forest.

Base Model:

Yellow Cab

Dep. Variable:	Price_Charged	R-squared:	0.745
Model:	OLS	Adj. R-squared:	0.745
Method:	Least Squares	F-statistic:	1.336e+05
Date:	Sun, 14 Mar 2021	Prob (F-statistic):	0.00
Time:	09:57:32	Log-Likelihood:	-1.7581e+06
No. Observations:	274681	AIC:	3.516e+06
Df Residuals:	274674	BIC:	3.516e+06
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.774e+04	700.786	39.591	0.000	2.64e+04	2.91e+04
KM_Travelled	20.3282	0.198	102.704	0.000	19.940	20.716
Cost_of_Trip	-0.0052	0.015	-0.346	0.729	-0.034	0.024
Month	-5.5016	0.080	-68.649	0.000	-5.659	-5.345
Year	-13.7343	0.347	-39.532	0.000	-14.415	-13.053
Age	-0.0835	0.022	-3.780	0.000	-0.127	-0.040
Income_(USD/Month)	0.0002	3.49e-05	4.746	0.000	9.73e-05	0.000

Omnibus:	51903.377	Durbin-Watson:	0.652
Prob(Omnibus):	0.000	Jarque-Bera (JB):	122747.976

Pink Cab

Dep. Variable:	Price_Charged	R-squared:	0.863
Model:	OLS	Adj. R-squared:	0.863
Method:	Least Squares	F-statistic:	8.871e+04
Date:	Sun, 14 Mar 2021	Prob (F-statistic):	0.00
Time:	09:57:34	Log-Likelihood:	-4.7693e+05
No. Observations:	84711	AIC:	9.539e+05
Df Residuals:	84704	BIC:	9.539e+05
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.515e+04	584.885	25.903	0.000	1.4e+04	1.63e+04
KM_Travelled	13.4824	0.165	81.834	0.000	13.160	13.805
Cost_of_Trip	0.0295	0.015	1.985	0.047	0.000	0.059
Month	1.5216	0.069	21.950	0.000	1.386	1.657
Year	-7.5169	0.290	-25.924	0.000	-8.085	-6.949
Age	-0.0400	0.018	-2.185	0.029	-0.076	-0.004
Income_(USD/Month)	3.423e-05	2.9e-05	1.181	0.238	-2.26e-05	9.11e-05

Omnibus:	28936.298	Durbin-Watson:	0.887
Prob(Omnibus):	0.000	Jarque-Bera (JB):	273333.925

As per Base Model:

➢ Cost of Trip, Month, Year, Age, Income are significant variable for **Yellow Cab** which are the best predictors for Price Charged.

➢ Cost_of_Trip, Year, Age, Income are significant variable for **Pink Cab** which are the best predictors for Price Charged. Month is not considered significant.

Best Fit Model: RMSE Value & Accuracy

❑ **RMSE or root mean square error** measures the error which is Prediction values – Actual values.
❑ **Lower the RMSE value the better is the Model.**

RMSE values & Accuracy for Yellow Cab

	Train	Test
Linear Regression	145.4599	146.1994
Decision Tree	107.3967	109.4580
Random Forest	77.2731	78.4734

	Accuracy
Linear Regression	74.43906127028283%
Decision Tree	86.11582117196697%
Random Forest	92.85776861169764%

RMSE values & Accuracy for Pink Cab

	Train	Test
Linear Regression	67.2351	67.9136
Decision Tree	80.7492	84.4882
Random Forest	57.4761	59.7556

	Accuracy
Linear Regression	86.06270464033021%
Decision Tree	79.66683587364297%
Random Forest	89.781966675241622%

As per the above RMSE data and Accuracy, Random Forest Model is the best fit model for further deployment.

Interpreting Random Forest Model: Cost of Trip, Month, Year, Age, Income are the best predictors for Price Charged.

Recommendation

❑ **Transaction per year**: For Yellow Cab Transaction per year from 2016 to 2018 is almost double than Pink Cab.

❑ **Margin per Gender**: For Yellow Cab there is difference in Margin between Male and Female Customers due to which Female Customer percentage is higher in Yellow Cab in comparison to Pink Cab.

❑ **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 people older than 50 yrs, whereas in Pink Cab there is no difference in Margin of all age group.

❑ Yellow Cab **decreases Margins with the increase in Transaction**, hence for Yellow Cab the travel frequency during the Month of December which is the holiday season is 3 times more than Pink Cab.

❑ Customers for Yellow Cab is highest in New York City which has the highest Cab Users of 28%.

On the basis of the above points, Yellow Cab is recommended for investment.

Thank You