



Data Glacier

Your Deep Learning Partner

Exploratory Data Analysis

G2M insight for Cab Investment firm

13th March 2023

By Hira Fahim

Background –G2M(cab industry) case study

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.

Approach

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 Understanding

CAB DATA

```
Numeric Features:  
Index(['Transaction_ID', 'Date_of_Travel', 'KM_Travelled', 'Price_Charged',  
      'Cost_of_Trip'],  
      dtype='object')  
=====
```

Categorical Features:
Index(['Company', 'City'], dtype='object')

CUSTOMER ID

```
Numeric Features:  
Index(['Customer_ID', 'Age', 'Income_(USD/Month)'], dtype='object')  
=====
```

Categorical Features:
Index(['Gender'], dtype='object')

CITY

```
Numeric Features:  
Index([], dtype='object')  
=====
```

Categorical Features:
Index(['City', 'Population', 'Users'], dtype='object')

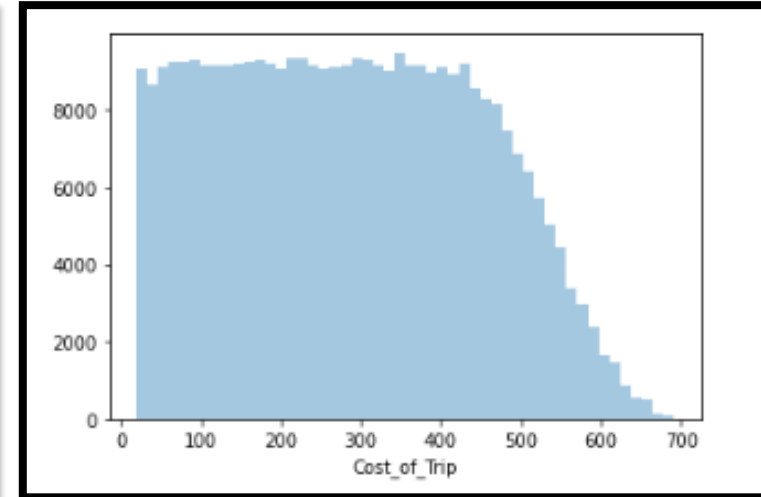
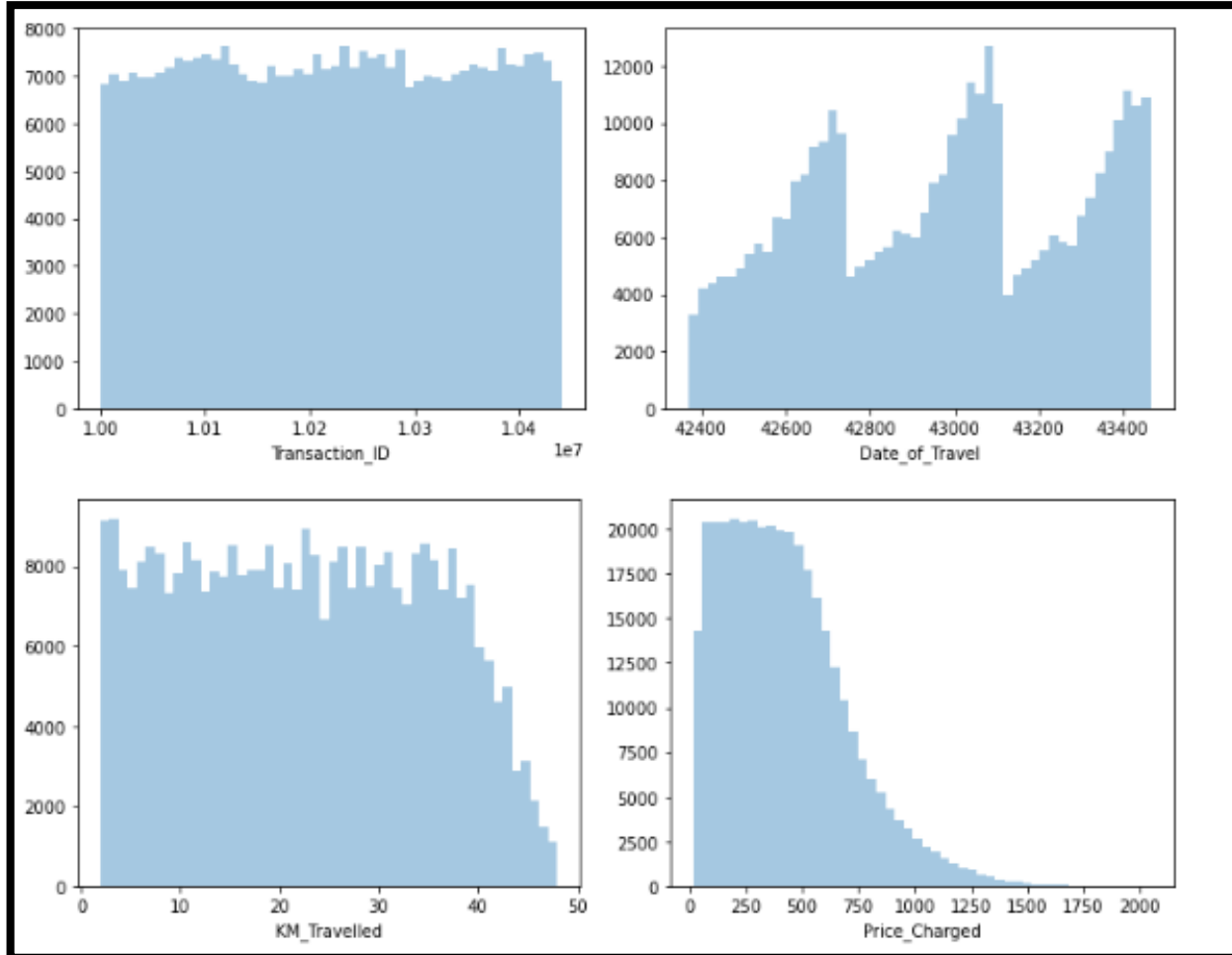
TRANSACTIONAL ID

```
Numeric Features:  
Index(['Transaction_ID', 'Customer_ID'], dtype='object')  
=====
```

Categorical Features:
Index(['Payment_Mode'], dtype='object')

Univariate Analysis: Cab Data

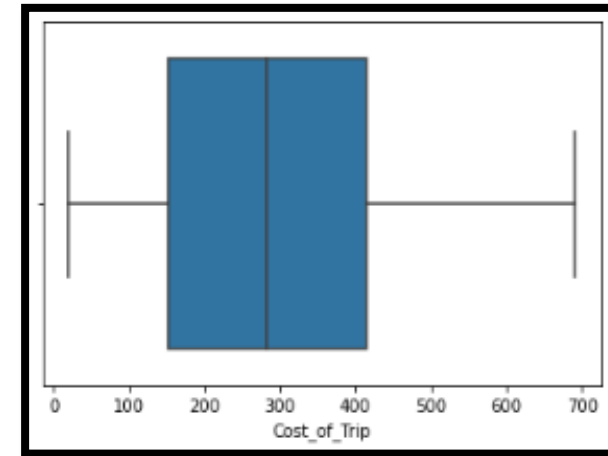
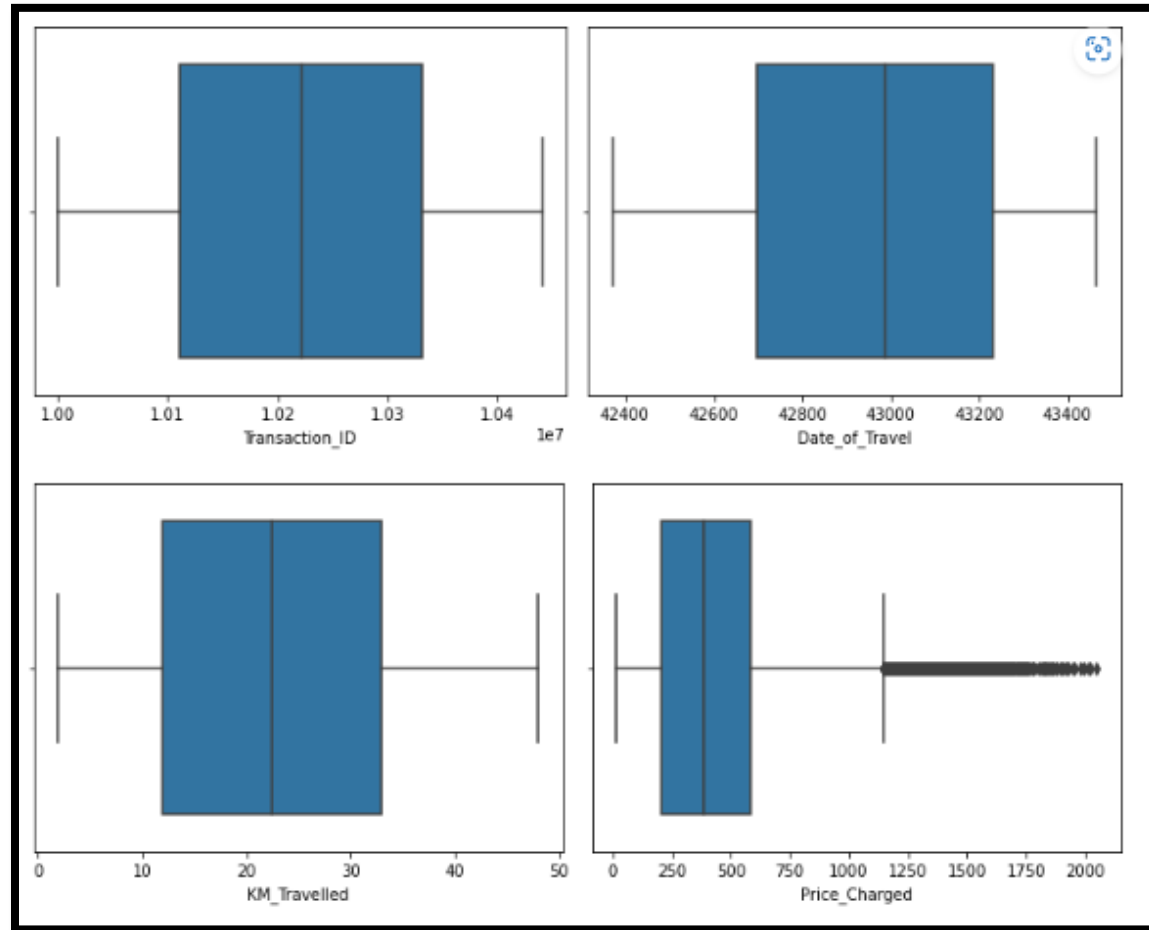
CAB DATA



Here we see the graphical representation of different variables of Cab Data .

Univariate Analysis: Cab Data

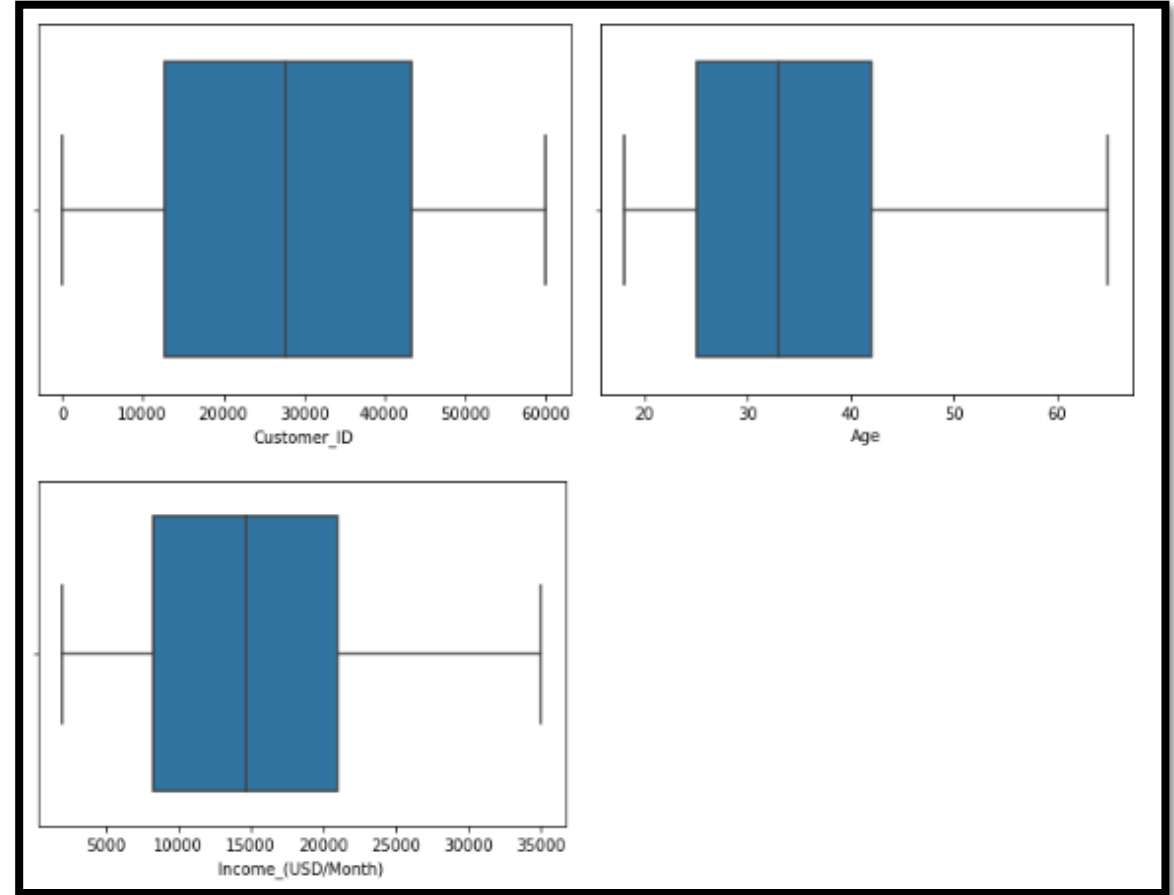
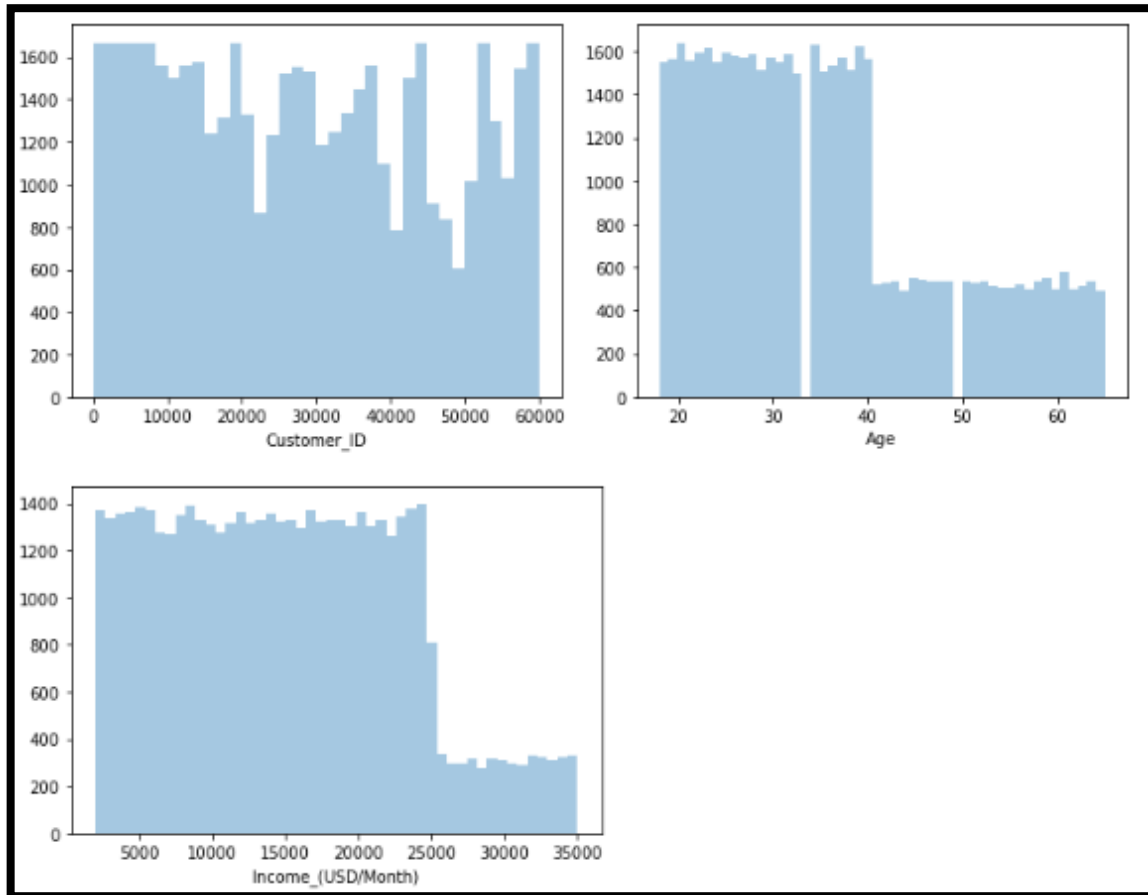
CAB DATA



Here we see the data distribution of all variables of Cab Data. We can also see the outliers in the column of "Price_Charged".

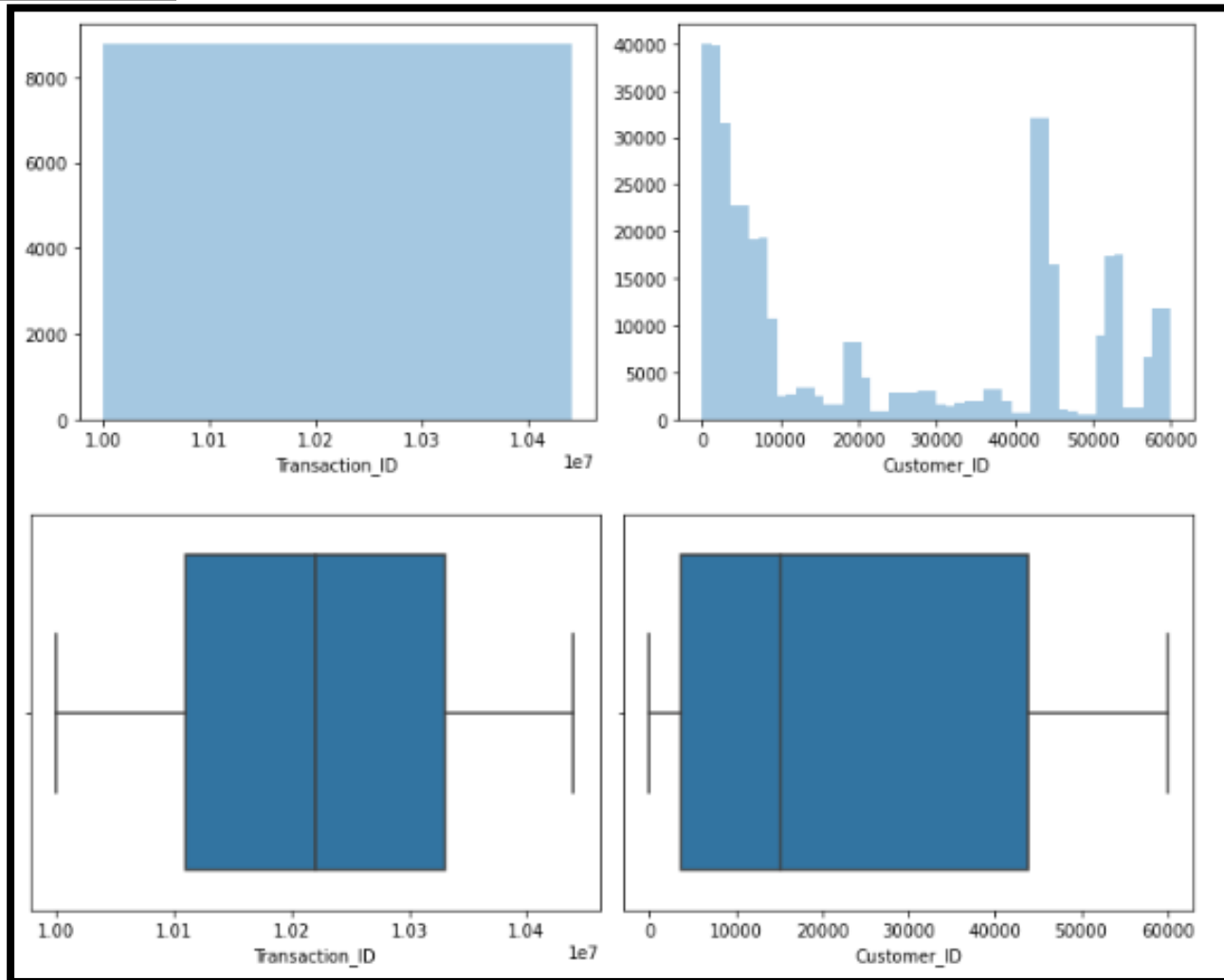
Univariate Analysis: Customer ID

CUSTOMER ID



Univariate Analysis: Transactional ID

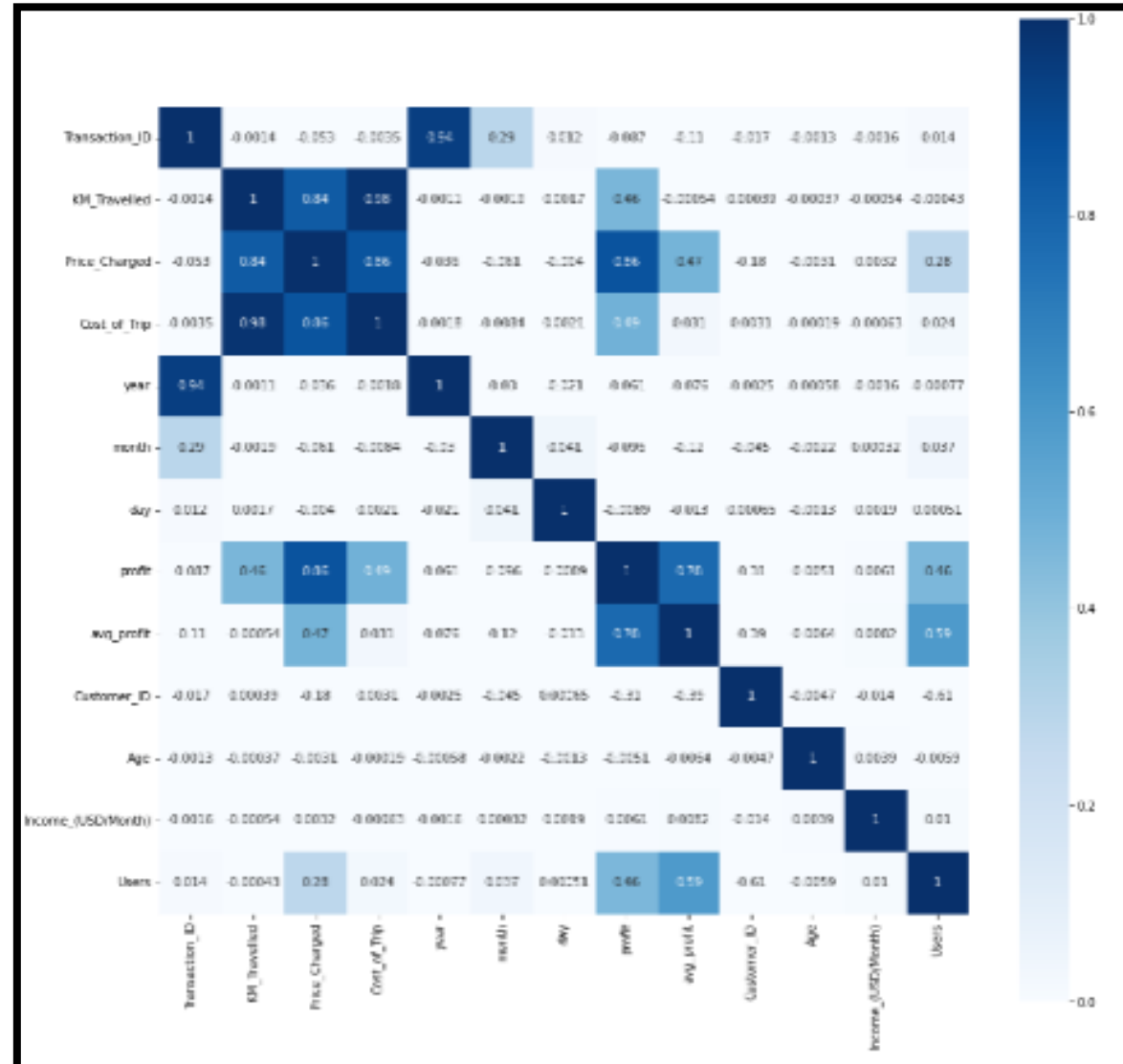
TRANSACTIONAL ID



Here we see the graphical representation and data distribution of different variables of Transactional_ID dataset.

Correlation of all Numerical features with Heatmap

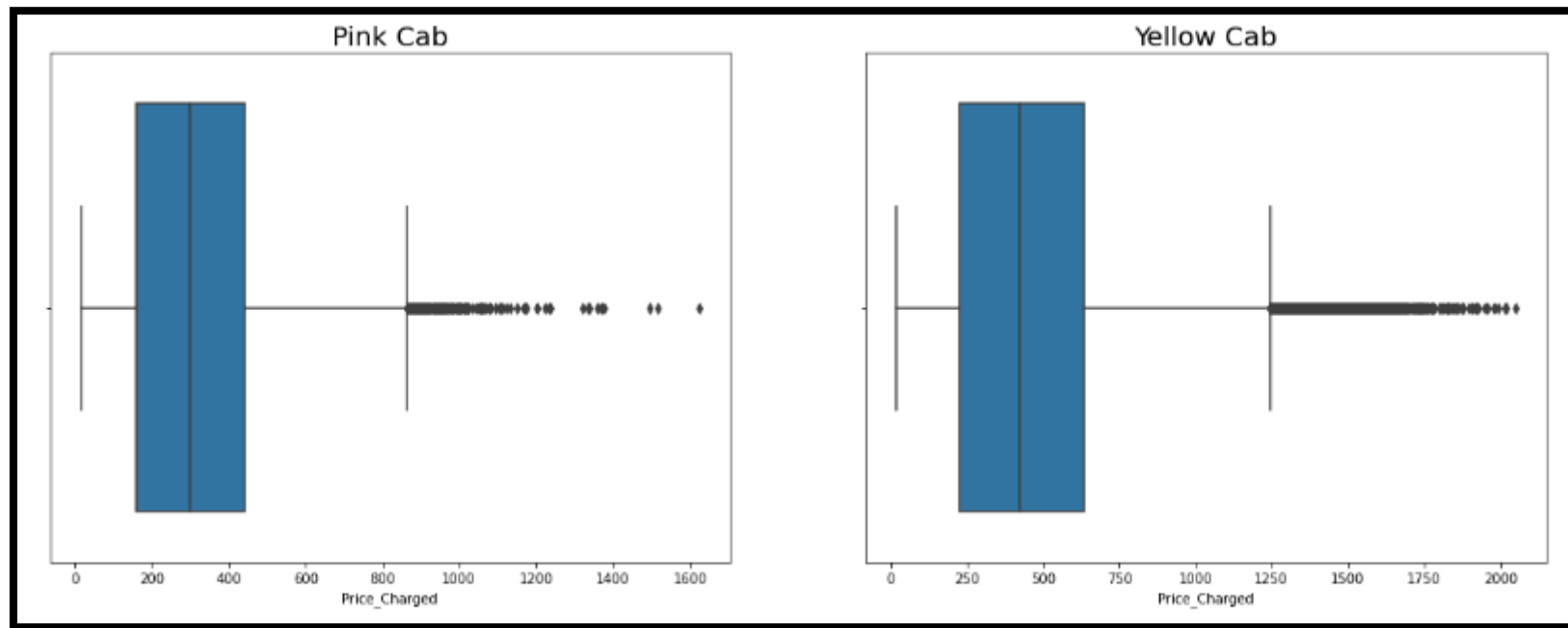
This Heatmap shows the correlation of all variables of Master Dataset for Cab industry.



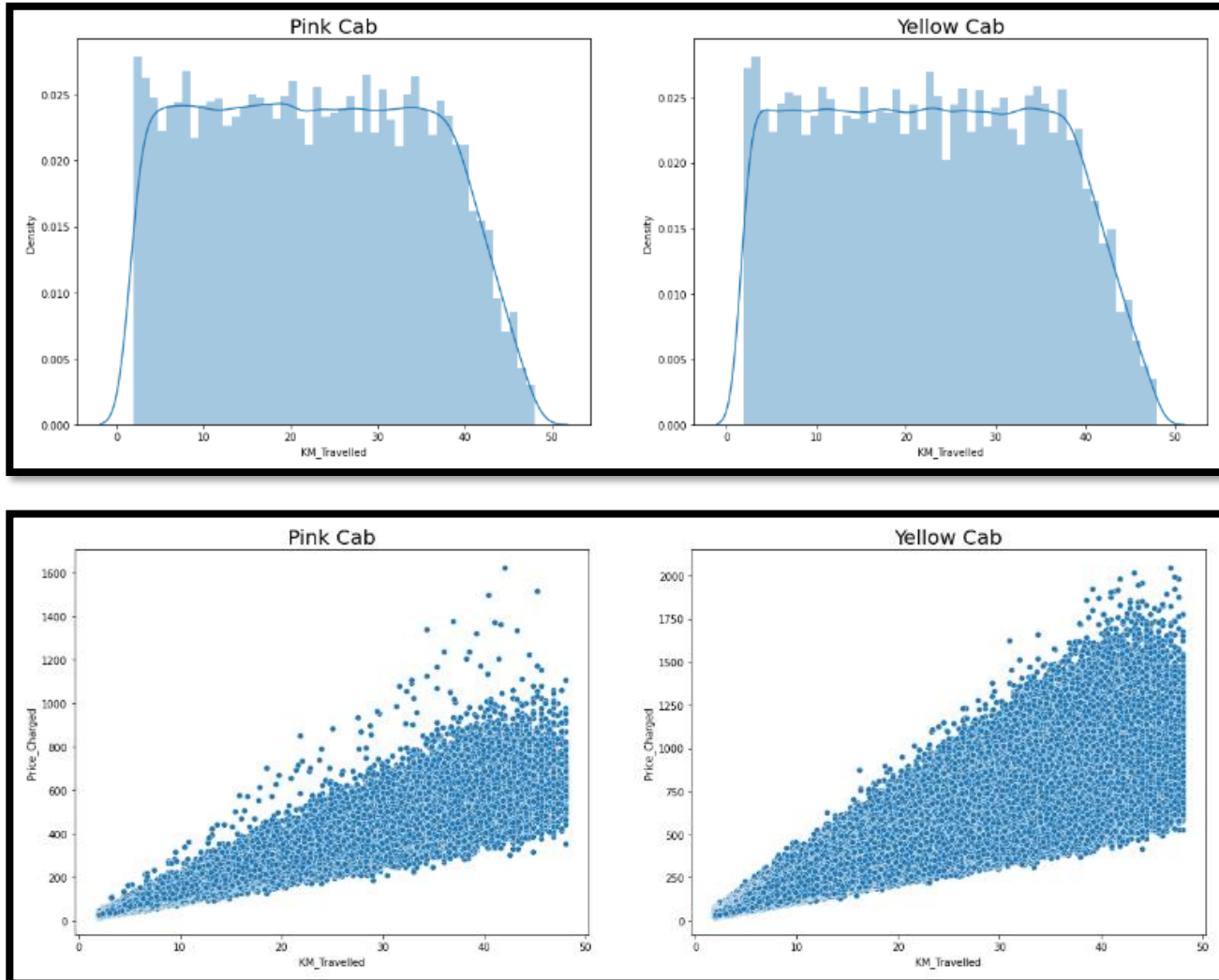
Exploratory Data Analysis

Outliers Detection

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

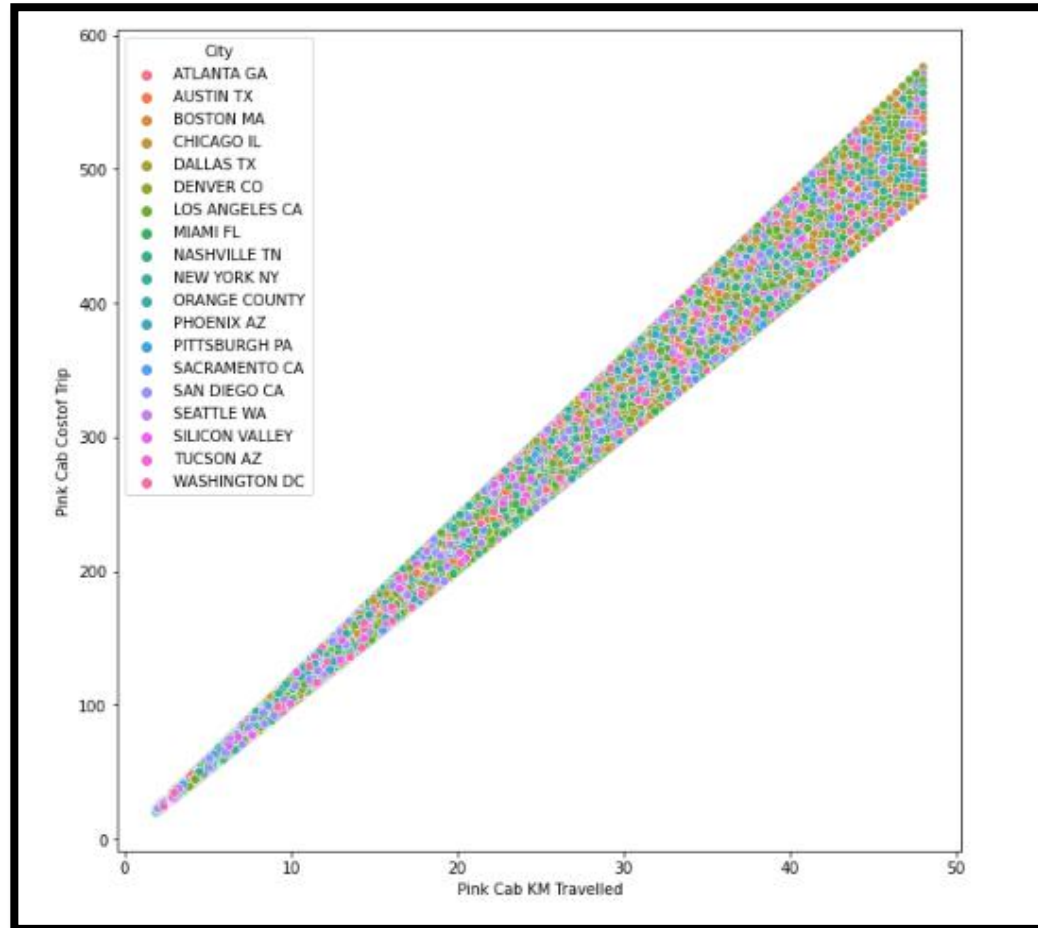


Visualization- Density and Price Charged per KM

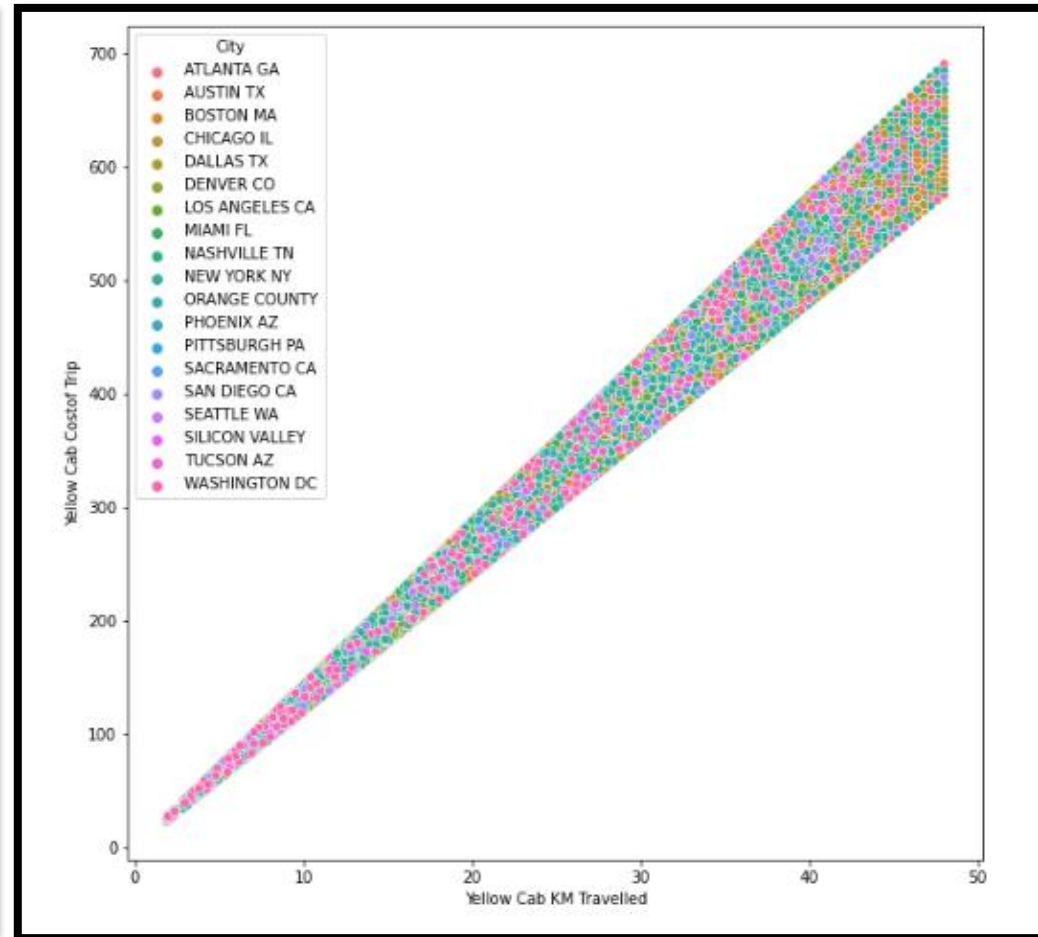


Visualization- Cost of trip per KM Travelled per City

Pink Cab

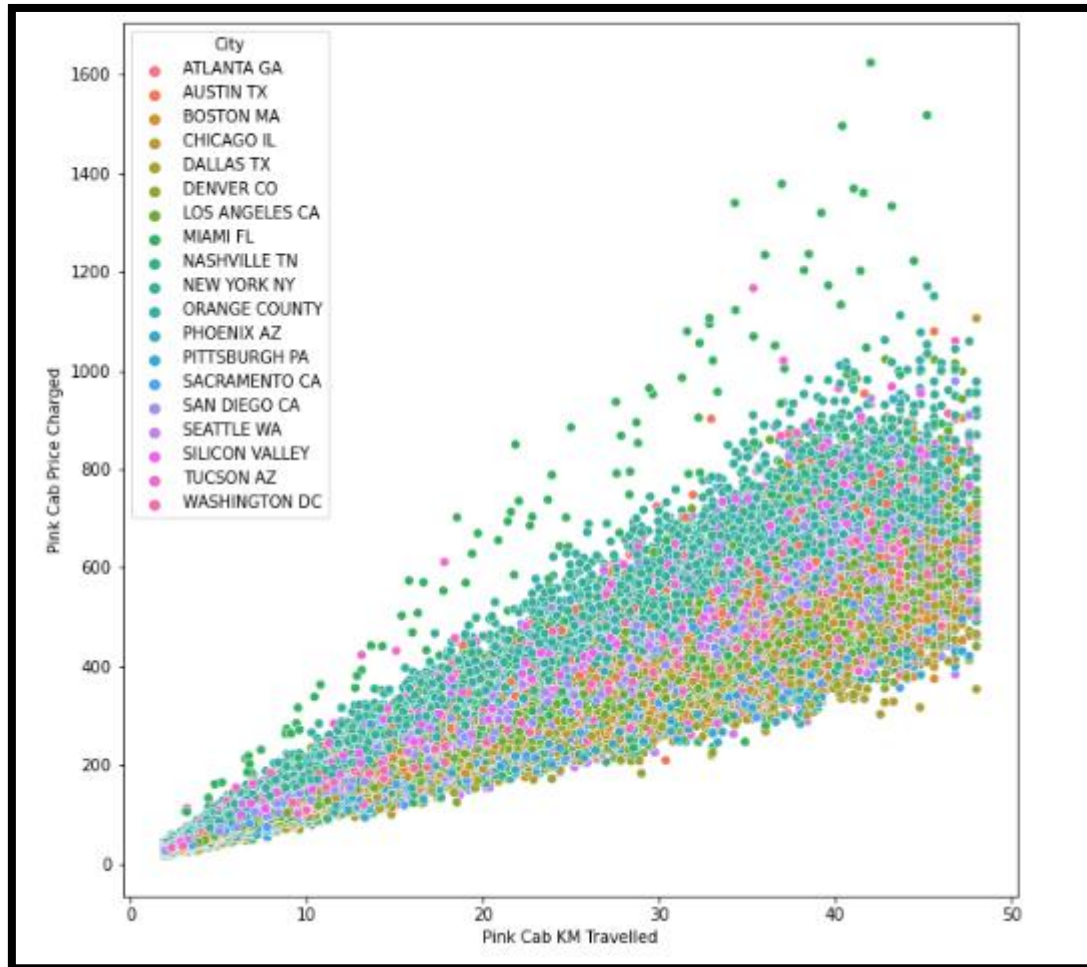


Yellow Cab

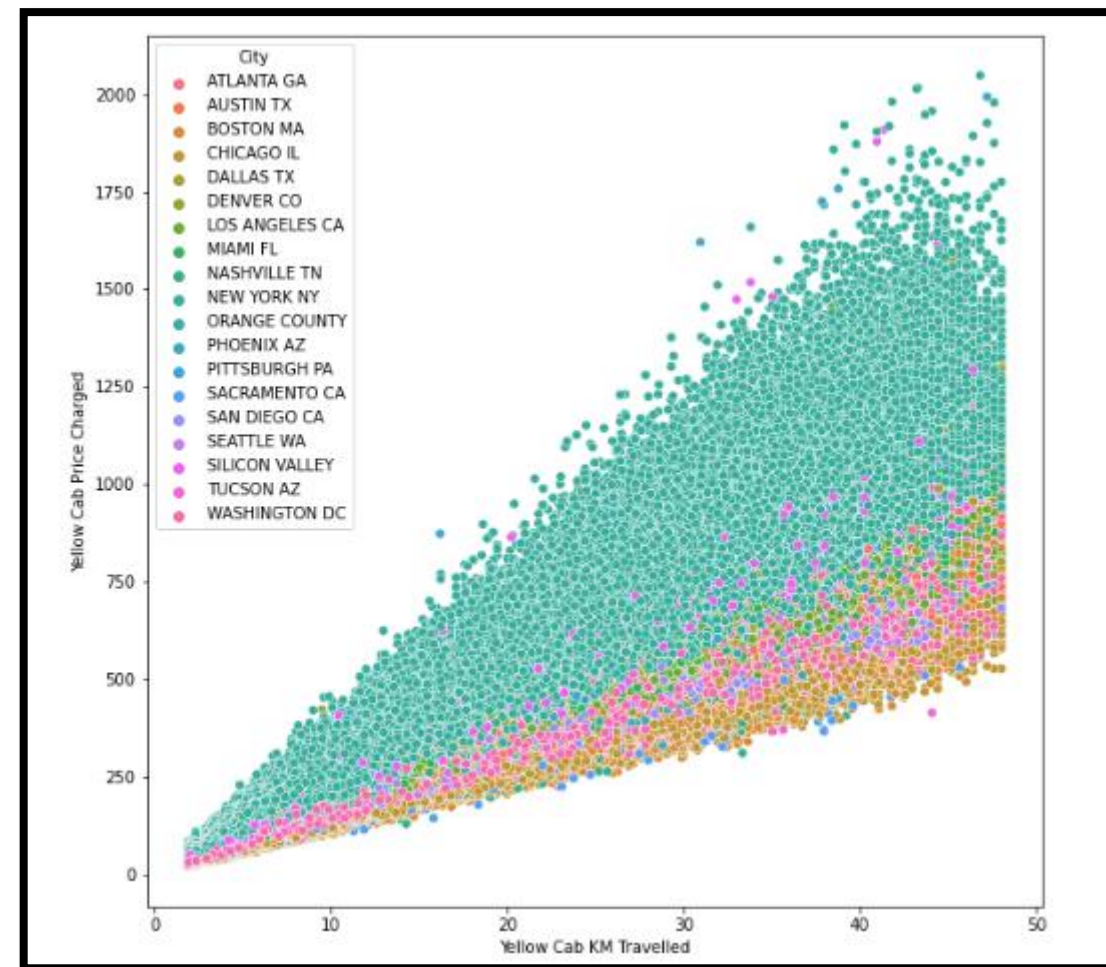


Visualization- Cab priced Charged per KM Travelled

Pink Cab

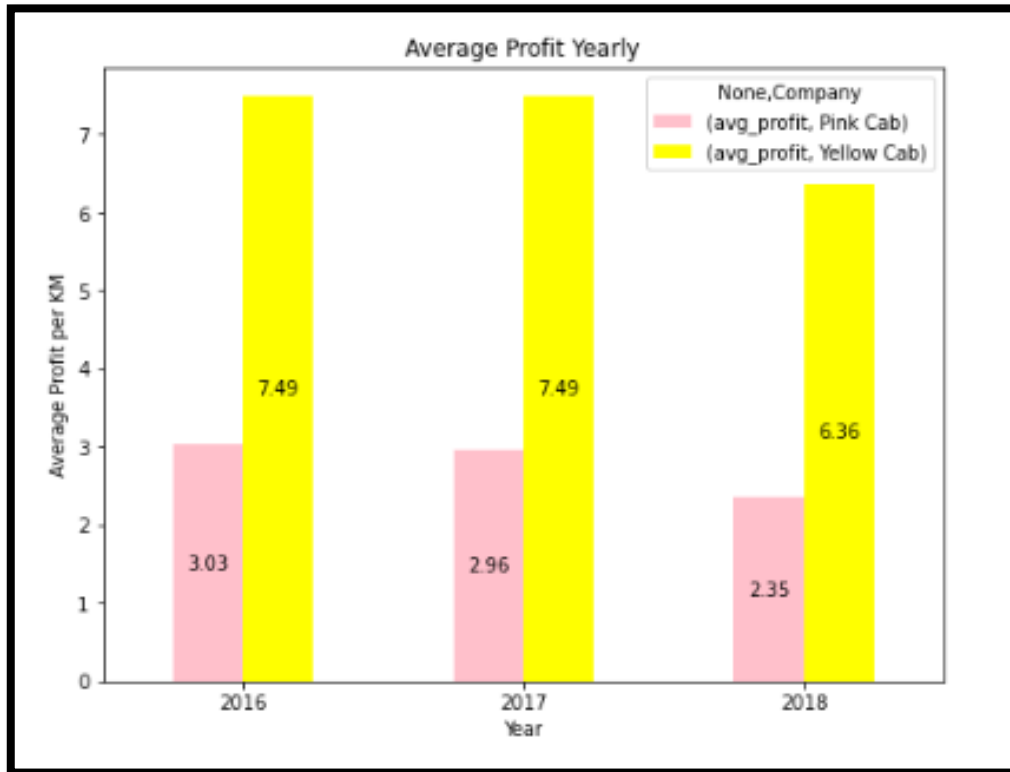


Yellow Cab

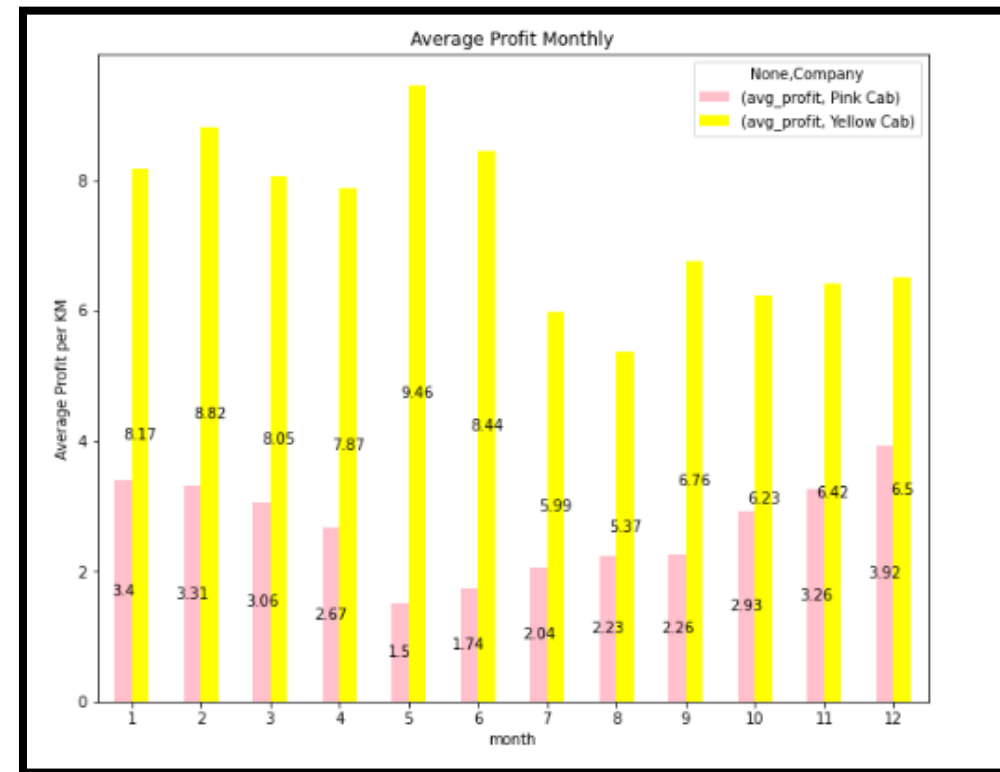


Average Yearly and Monthly Profit Analysis

Average Yearly Profit



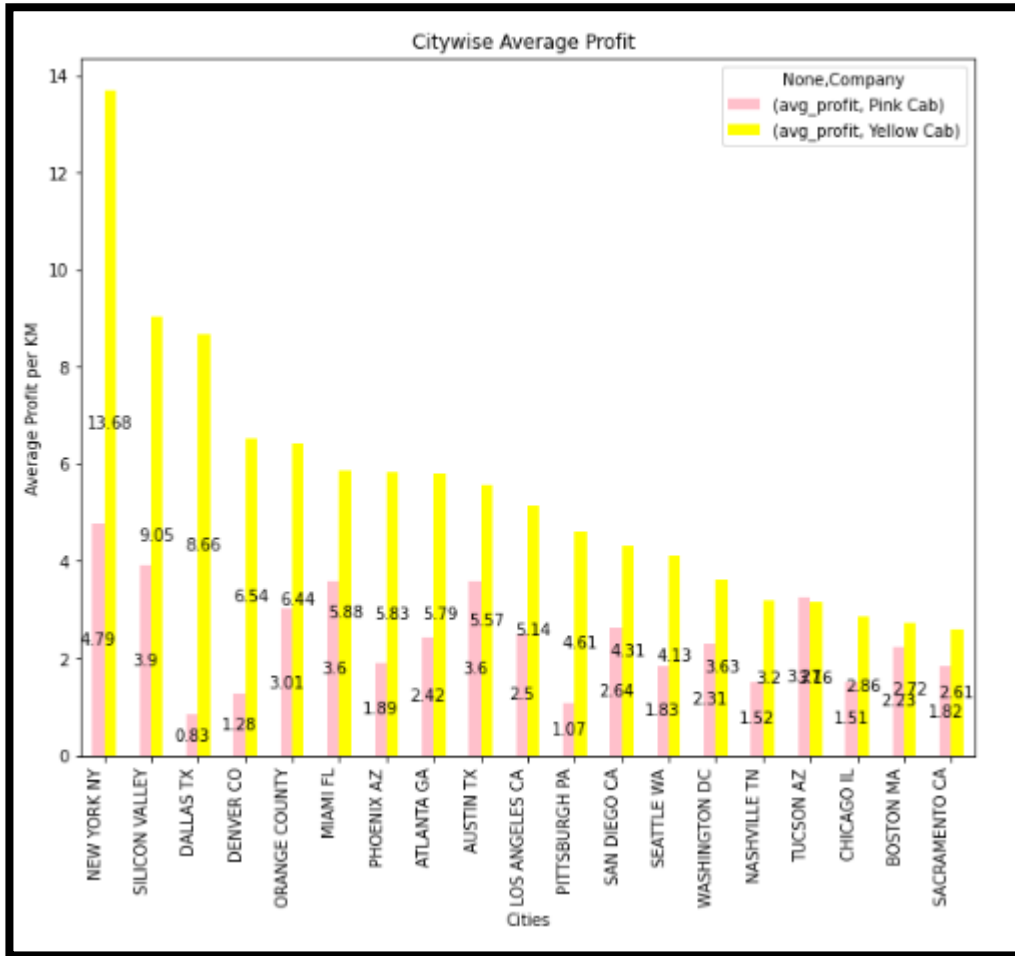
Average Monthly Profit



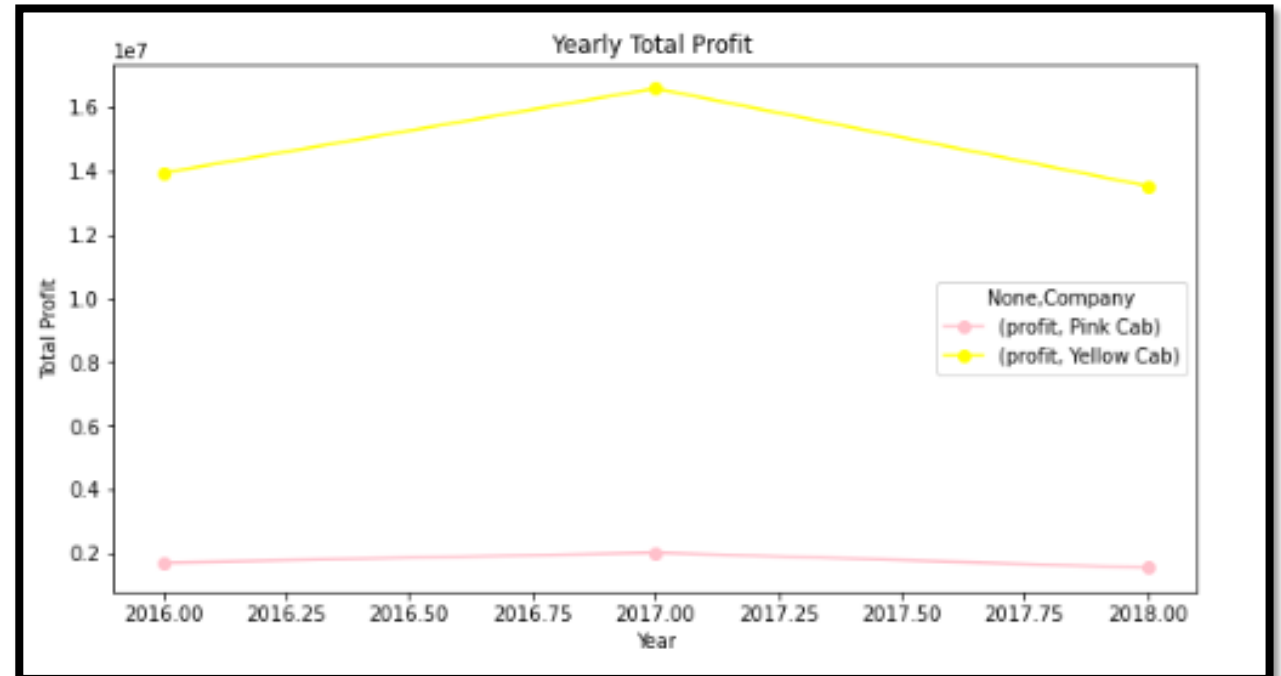
These graphs represent that “Average Yearly and Monthly profit” of Yellow cab is more than Pink cab.

City-wise Average Profit and Total Yearly

City-wise Average Profit



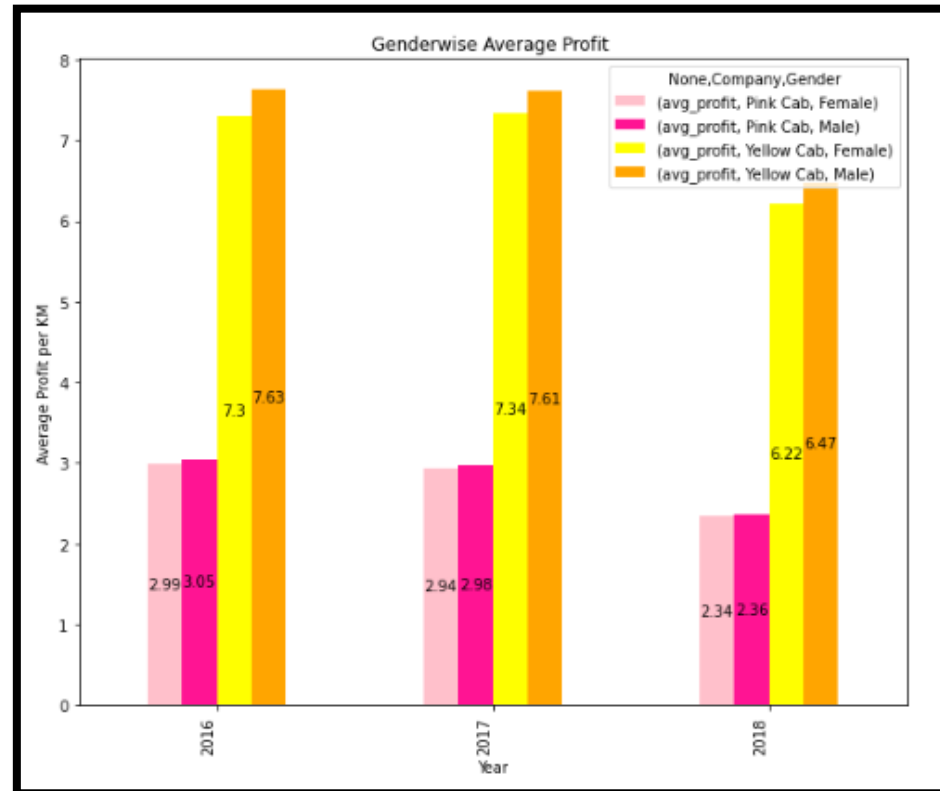
Yearly Total Profit



These graphs represent that “City-wise average profit” and “Total yearly profit” per KM of Yellow cab is better than Pink cab.

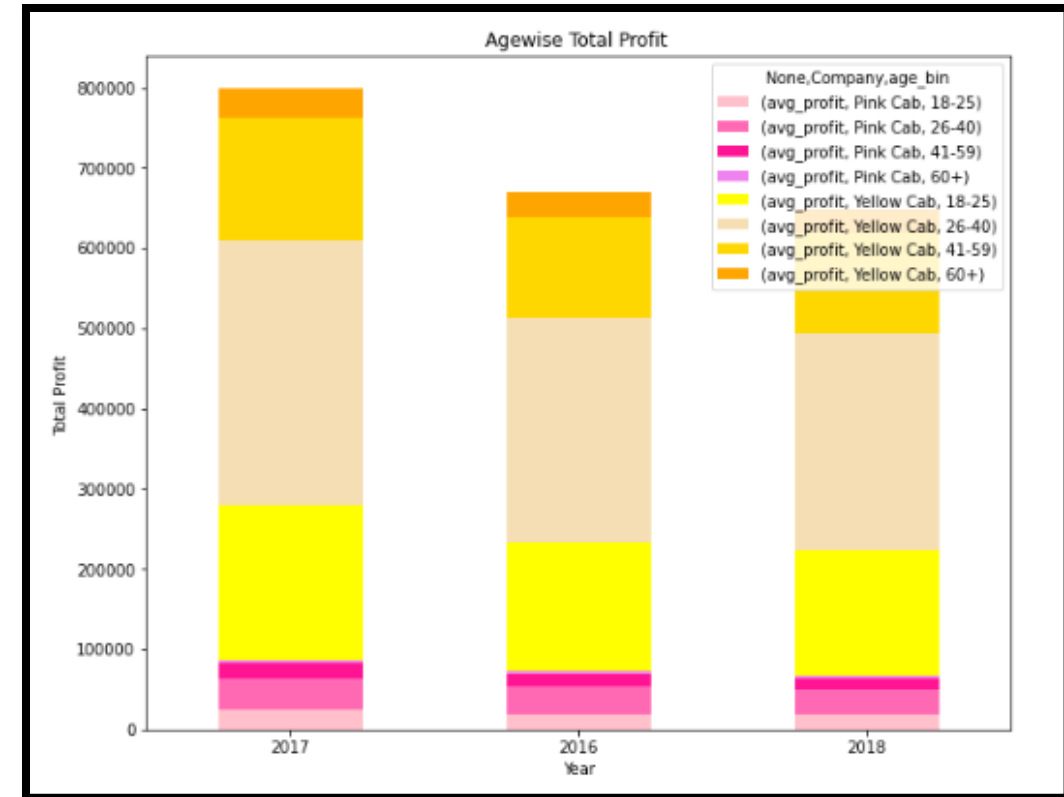
Gender-wise and Age-wise Profit Analysis

Gender-wise Average Yearly Profit



There is almost equal distribution of gender in the Profit and customer base of both the cabs.

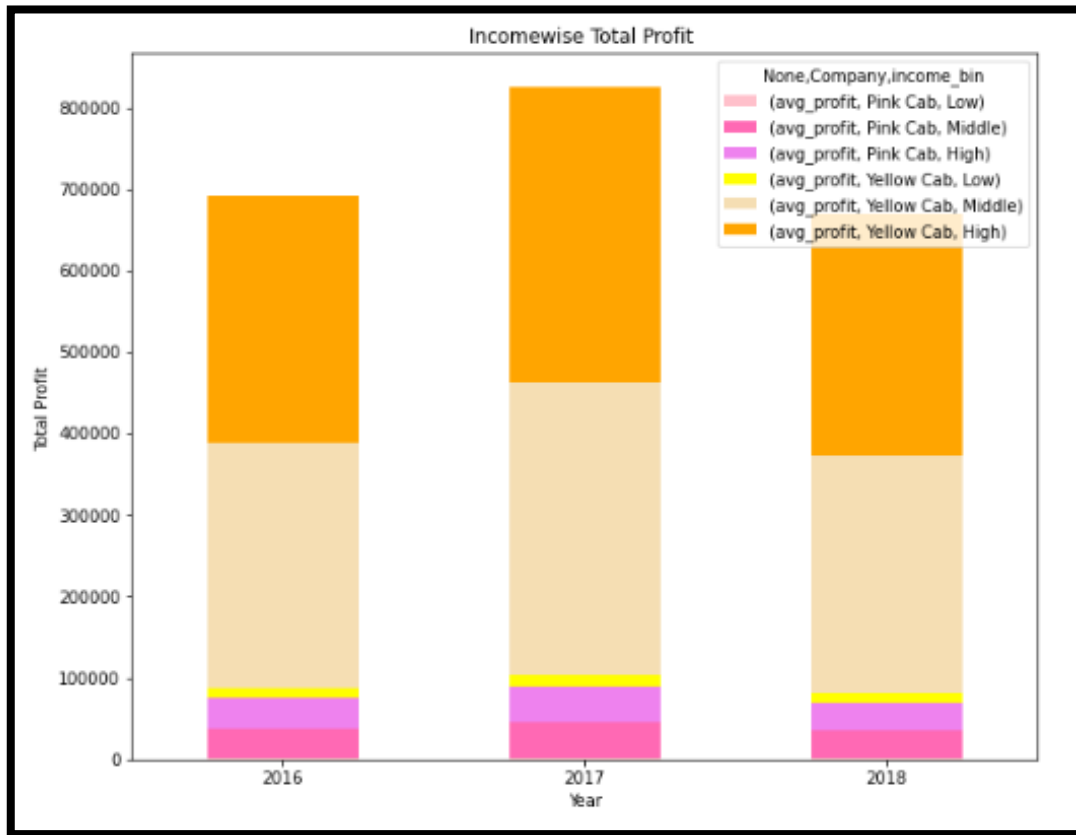
Age-wise Total yearly Profit



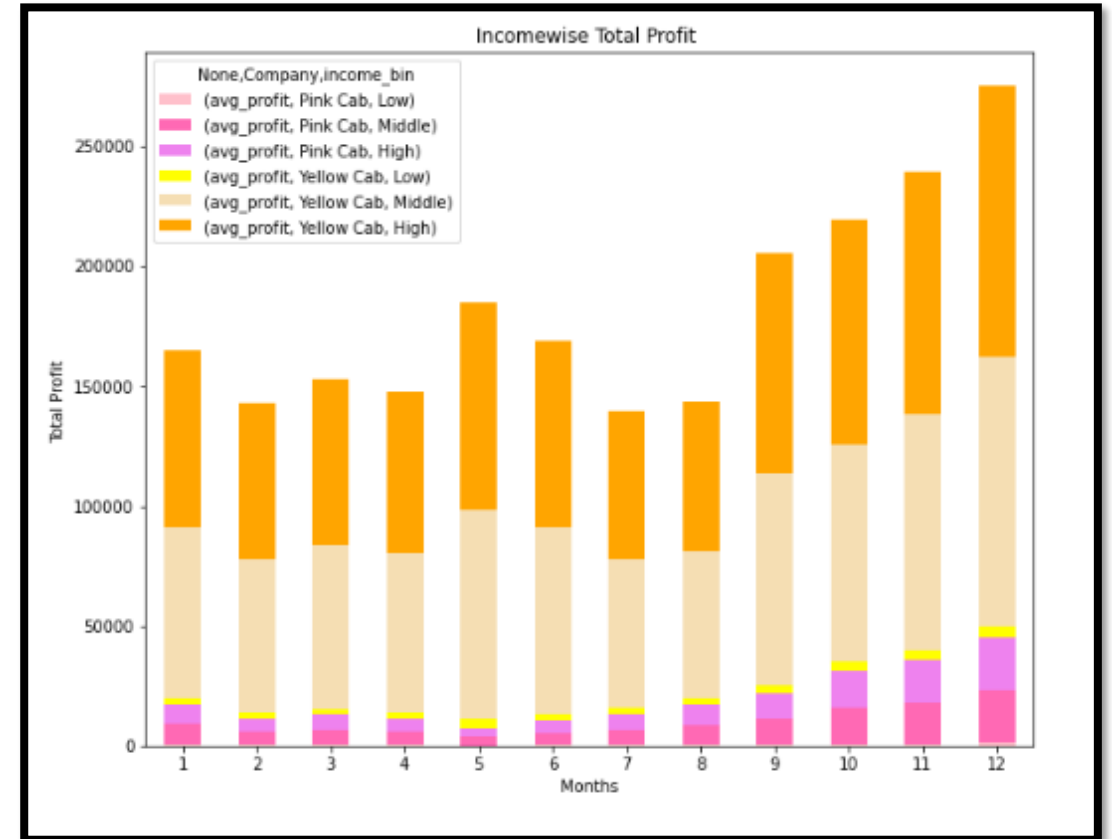
Users of the age group (26 – 40) are main contributors in the Profit of both the cabs.

Income –wise Profit Analysis

Income-wise Total Yearly Profit



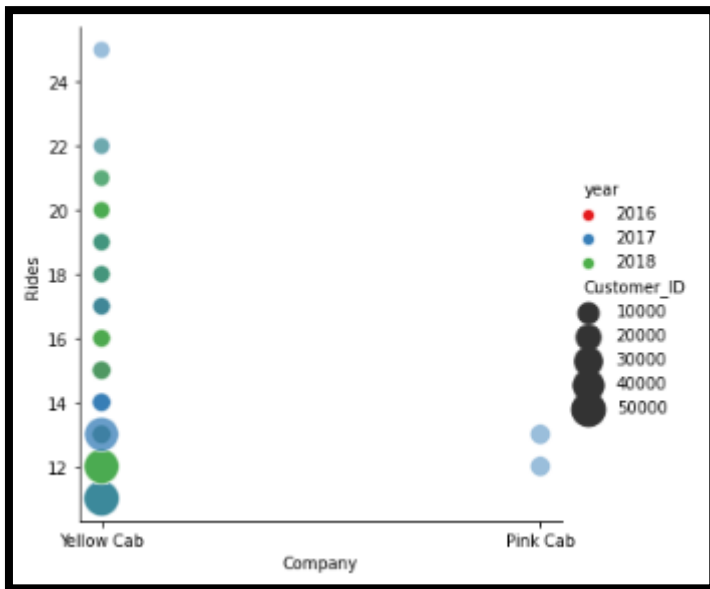
Income-wise Total Monthly Profit



These graphs “Income-wise yearly profit”, and here we can see high- and middle-income users have more contribution in profit. Yellow cab has more profit gain than pink cab.

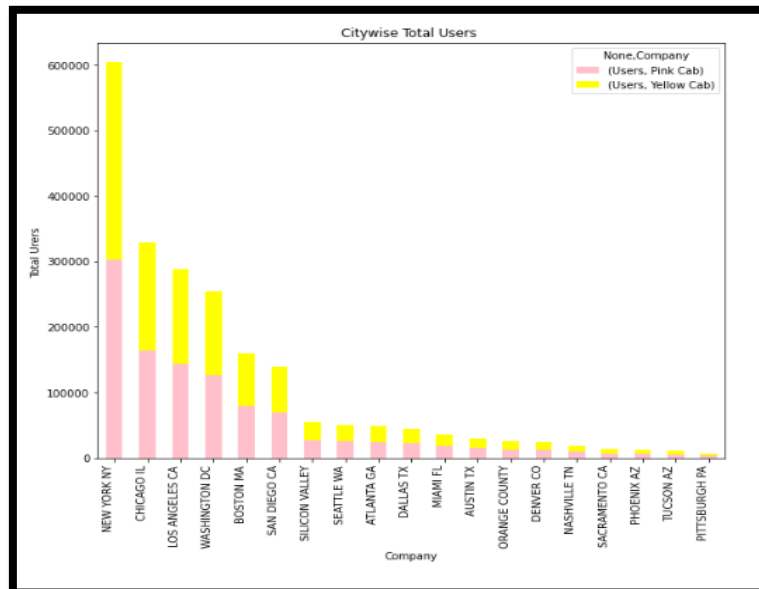
Visualization on Cab Users

Customer Retention



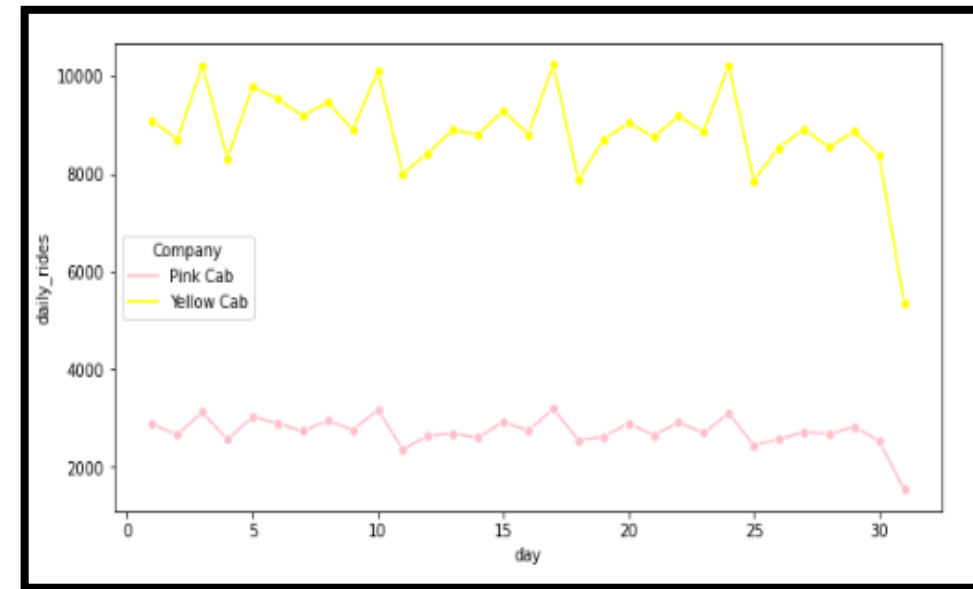
This relplot represents that Customer retention of Yellow cab has increase over the years (2016, 2017 and 2018) as compared to Pink cab.

City-wise Users



This bar chart represents that cab users per city of Yellow cab are more than Pink cab.

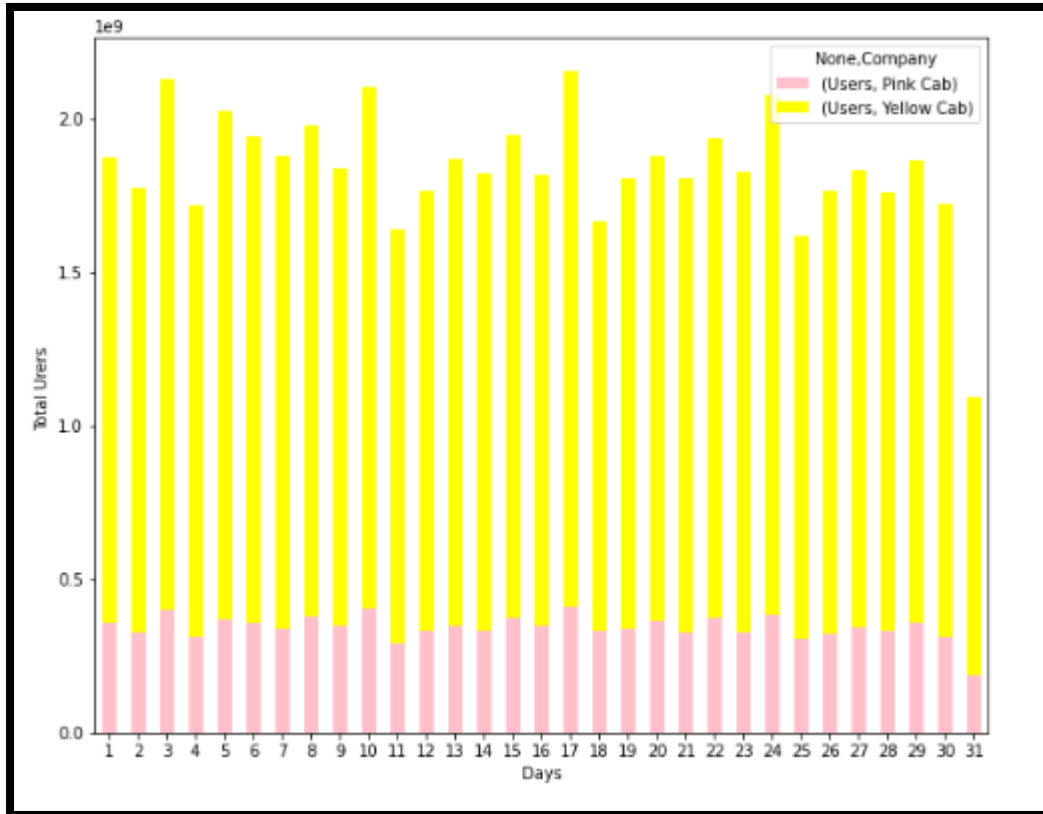
Seasonality



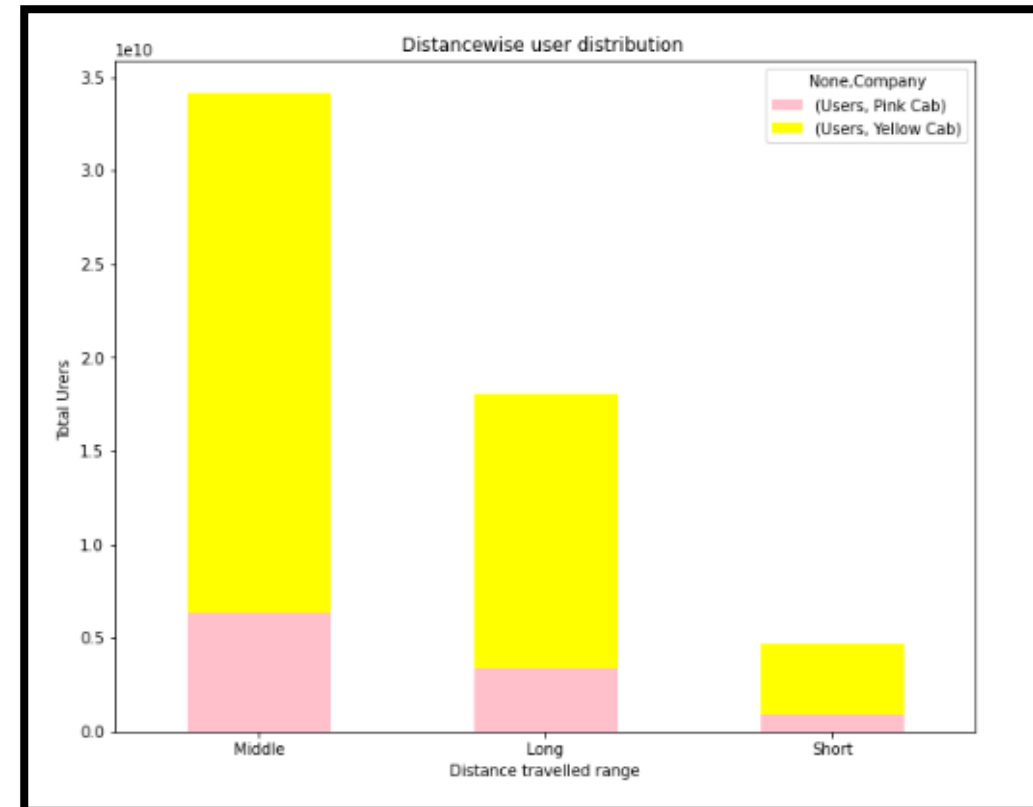
This plot shows that daily rides of Yellow cab are more than Pink cab.

Visualization on Cab Users

Daily Users



Distance Travelled



These graphs represent that “Users per day” and “Total users per distance(middle, long and short) travelled ” of Yellow cab are more than Pink cab.

Hypothesis Testing

Hypothesis : Is there profit remains same for gender for yellow & pink cab?

Null Hypothesis : Profit for Gender remains same for Yellow Cab & Pink Cab ($p - \text{value} < 0.05$)



Yellow Cab:

- $P - \text{value} = 6.06144$, which is greater than 0.05
- We accept alternative Hypothesis that is there is statistical difference.
- There is difference in Profit between Male and Female customer.



Pink Cab:

- $P - \text{value} = 0.115$, which is greater than 0.05
- We accept alternative Hypothesis that is there is statistical difference.
- There is difference in Profit between Male and Female customer.

Hypothesis Testing

Hypothesis : Is there difference in profit for Card payers and Cash payers?

Null Hypothesis : Profit remains same for card and cash users ($p - \text{value} < 0.05$)



Yellow Cab:

- $P - \text{value} = 0.29$, which is greater than 0.05
- We accept alternative Hypothesis that is there is statistical difference.
- There is difference in Profit for card and cash users .



Pink Cab:

- $P - \text{value} = 0.79$, which is greater than 0.05
- We accept alternative Hypothesis that is there is statistical difference.
- There is difference in Profit for card and cash users.

Recommendations for Investment

	Yellow Cab	Pink Cab
Customer Reach	Highest reach in 19 cities	Highest reach in 5 cities
Customer Retention	Number of rides have increased over the years	Number of rides are less compare to Yellow Cab
Age - wise Reach	All age group consumers + profit gain is high	All age group consumers + profit gain is Low comparatively
Income - wise Reach	High and middle income group are main contributor + profit gain is high	High and middle income group are main contributor + profit gain is Low comparatively
Average Profit per KM	4 times profit gained than pink cab	Less profit gained than yellow cab

After complete analysis we recommend yellow is more profitable for Investment



Thank You



Data Glacier

Your Deep Learning Partner

Hira Fahim