Exploratory Data Analysis

G2M Case Study

20/10/2022

Agenda

- Data Summary
- Problem Statement
- Approach
- EDA
- EDA Summary
- Recommendations

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.

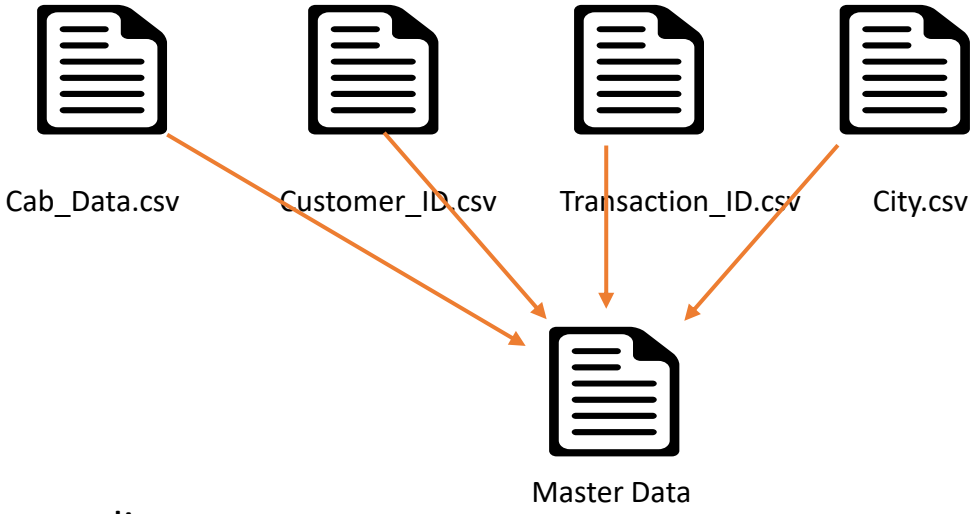
- **Objective** : Provide actionable insights to help XYZ firm in identifying the right company for making investment.

The analysis has been divided into following parts:

- Data Understanding
- Data Visualization
- EDA
- Finding the most profitable Cab company
- Recommendations for investment

Data Summary

- 22 Features(including 8 derived features)
- 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.

Problem Statement

- To find out from the data of Pink Cab Company and Yellow cab Company, which is better to do an investment, so the investor XYZ can make a choice to invest.

Approach

- Dataset provided is cab data, city data, customer data and transaction data in csv format. So data were merged into Master Data.
- Performed analysis on KM Travelled, Cost of trip and Price Changed for Pink cab as well as Yellow cab.

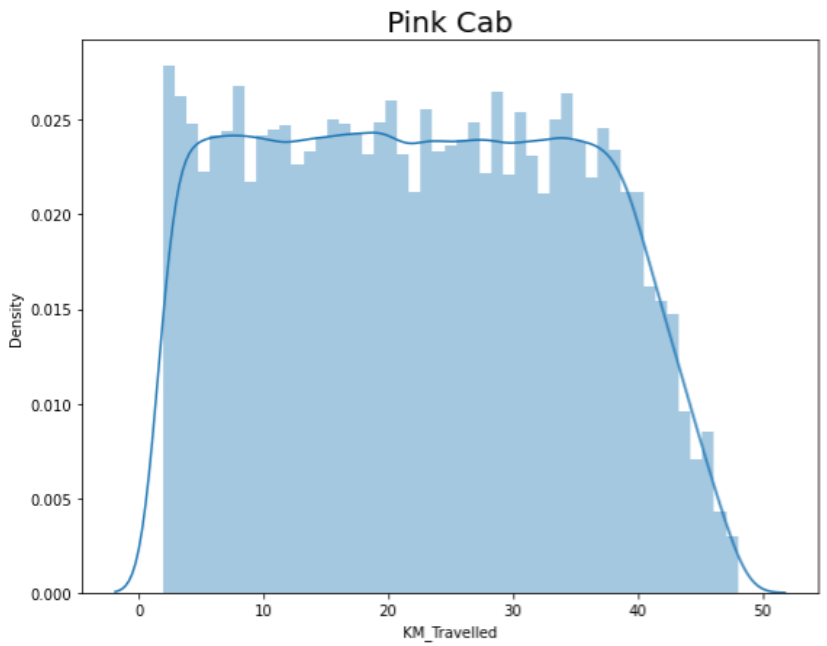
Following steps are performed on dataset to gain some insights

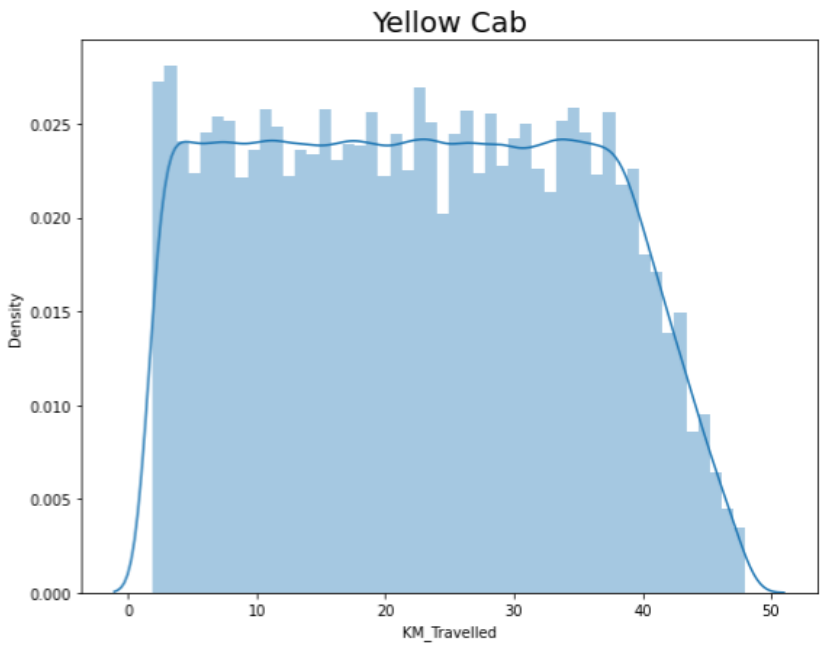
- Merge all csv files into Master Data.
- Perform a quick review of dataset with seaborn.
- Looked for correlation coefficient to find relation between variables.

EDA

- Dataset used for EDA is Master data
- Looked for missing data
- Cleaned the dataset
- Identified numerical and categorical features

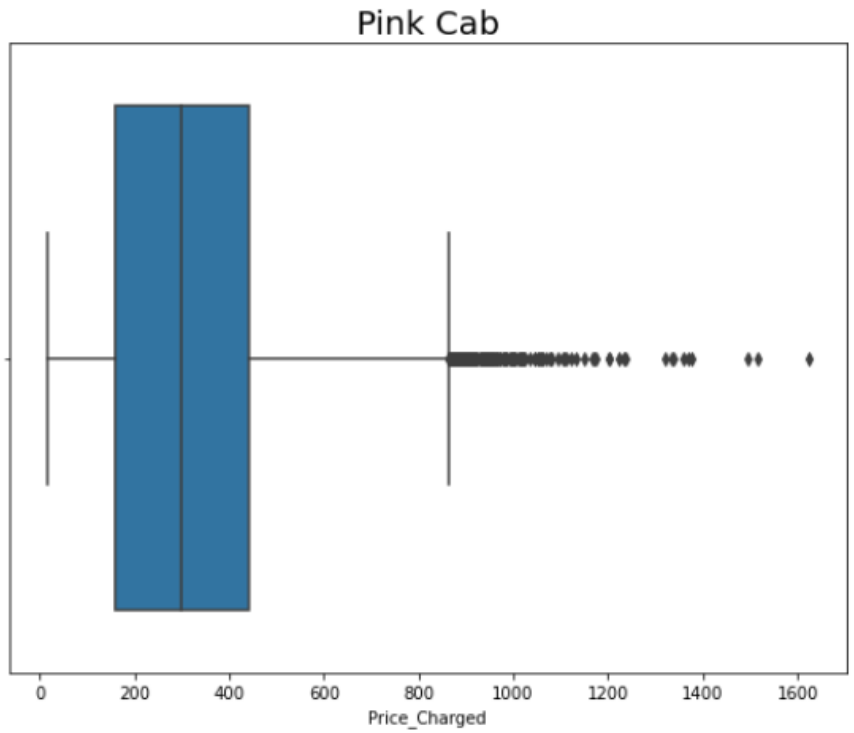
EDA

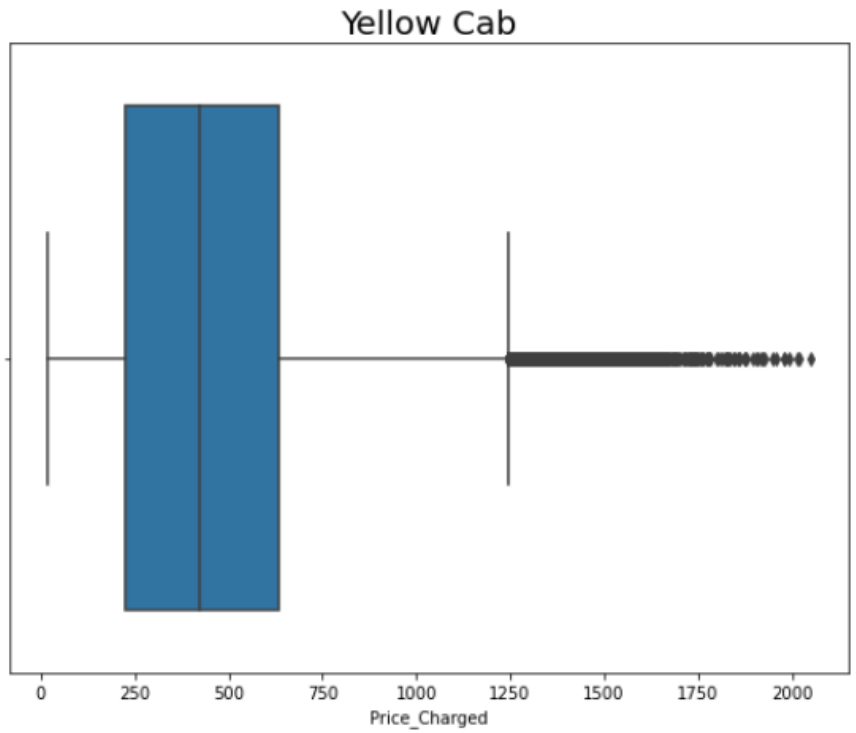




- Above visualization shows that total distance travelled in KM for yellow cab as well as Pink cab is same, which is range from 2- 48 KM.

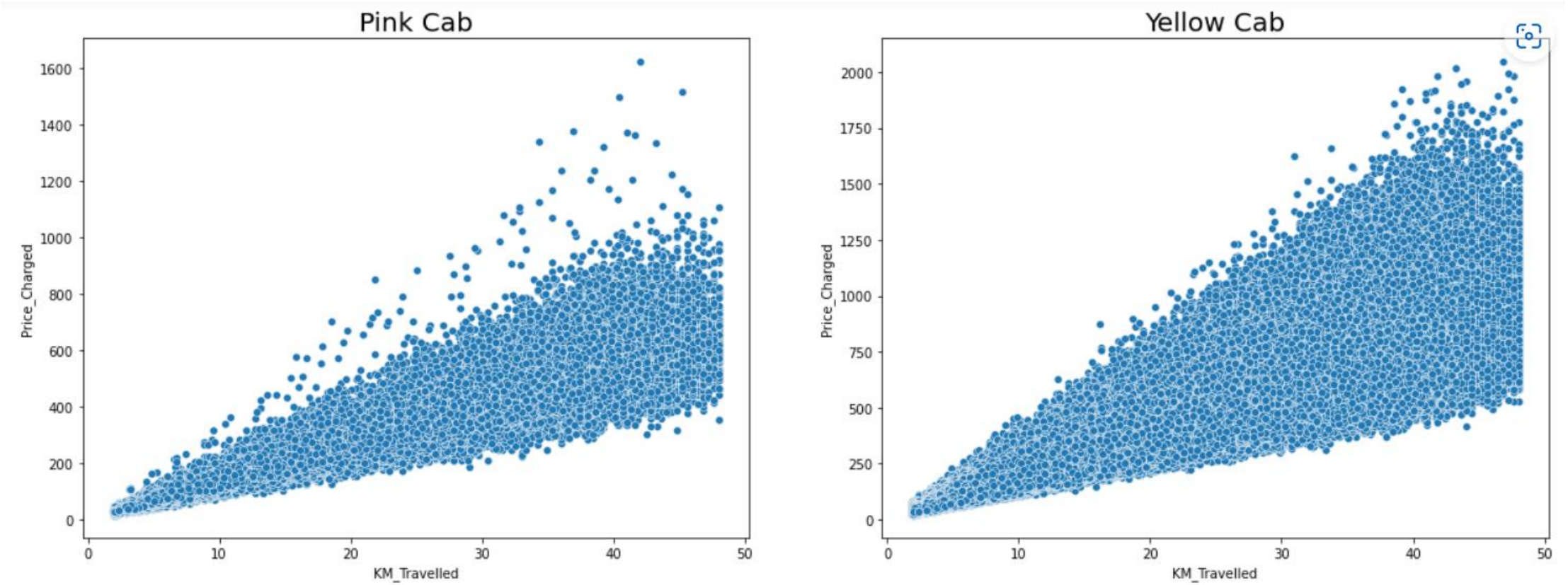
EDA





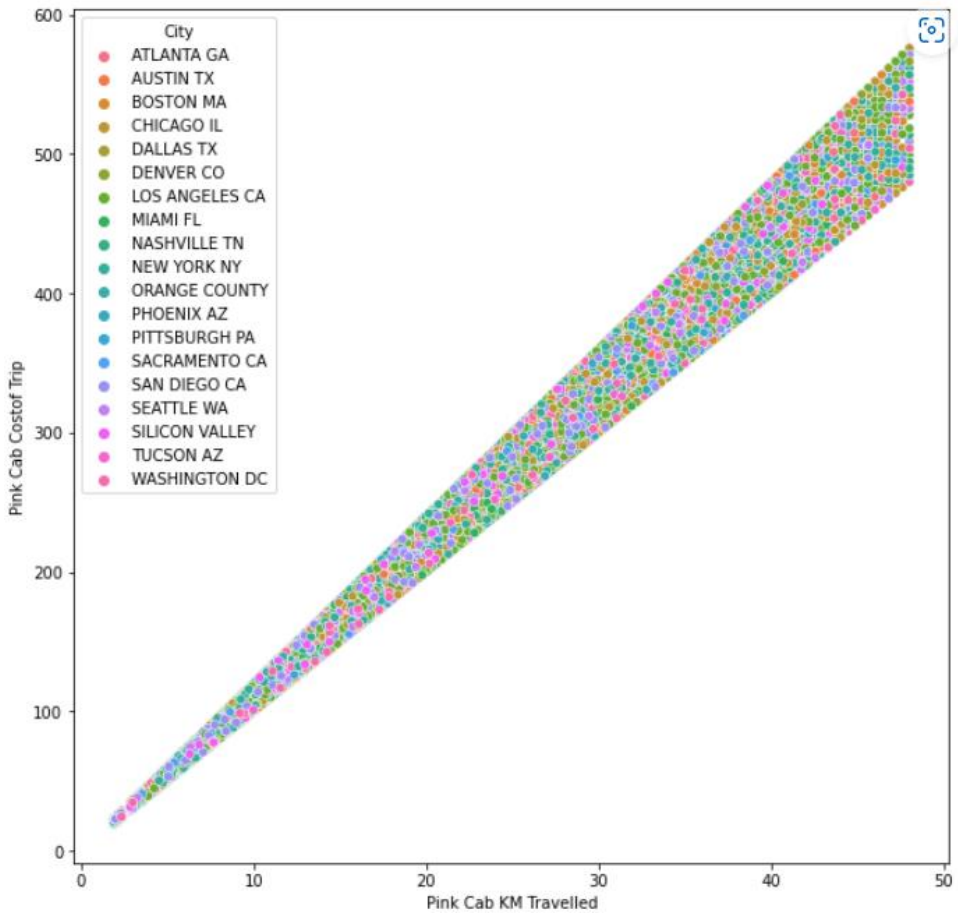
- Above visualization shows that Price charged for yellow cab is greater than price charged for Pink cab.
- Price charged for Box plot shows outliers. But they are not considered as outliers, as unavailability of trip duration.

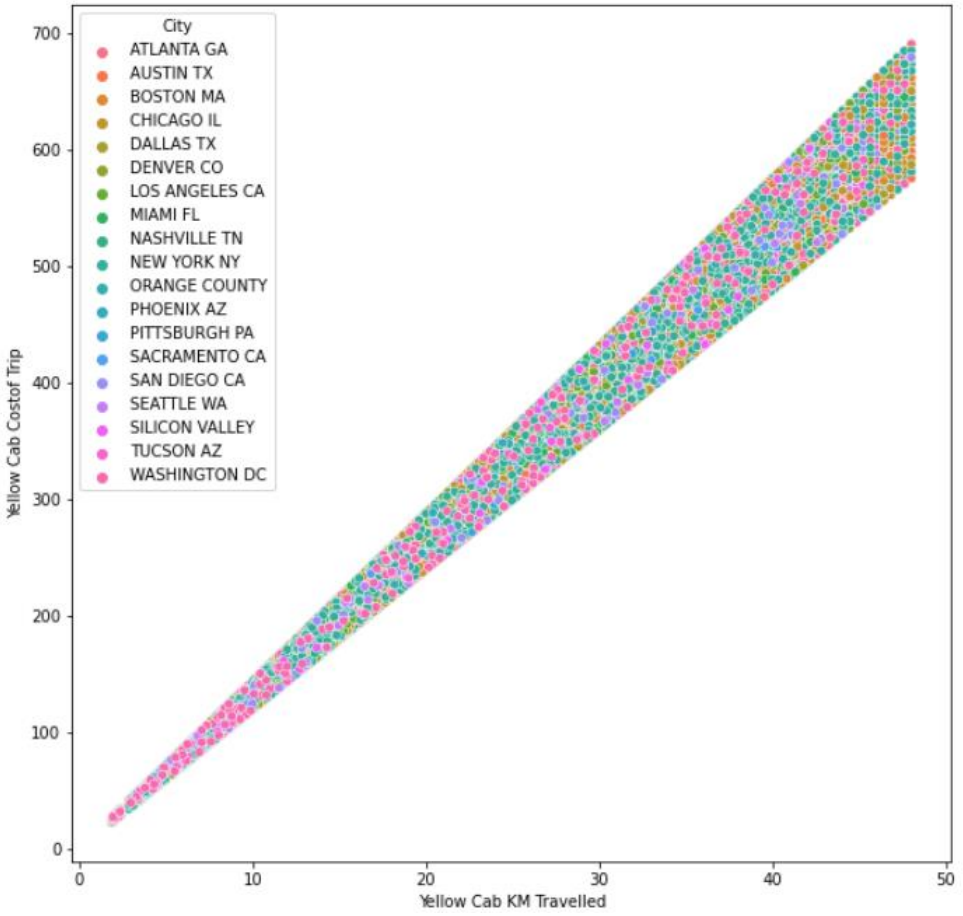
EDA



- Yellow cab charged higher price as compared to Pink cab for same distanced travelled.

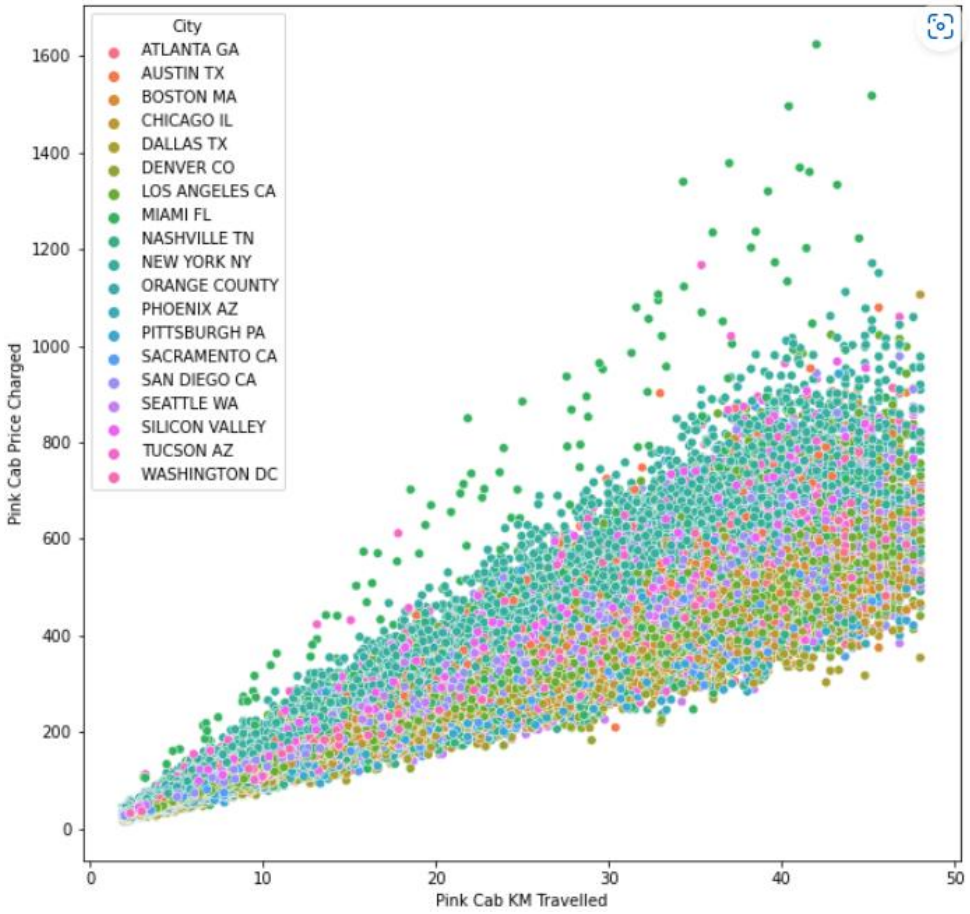
EDA

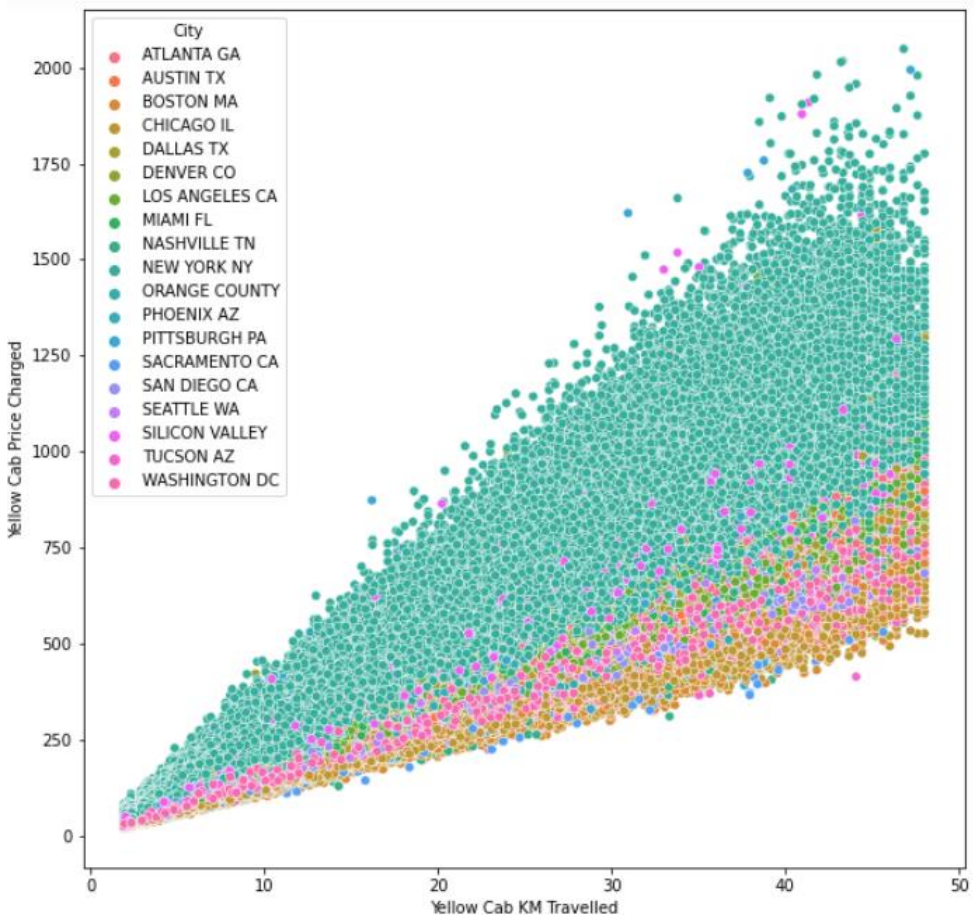




- Yellow cab has higher profit margin as compared to Pink cab.

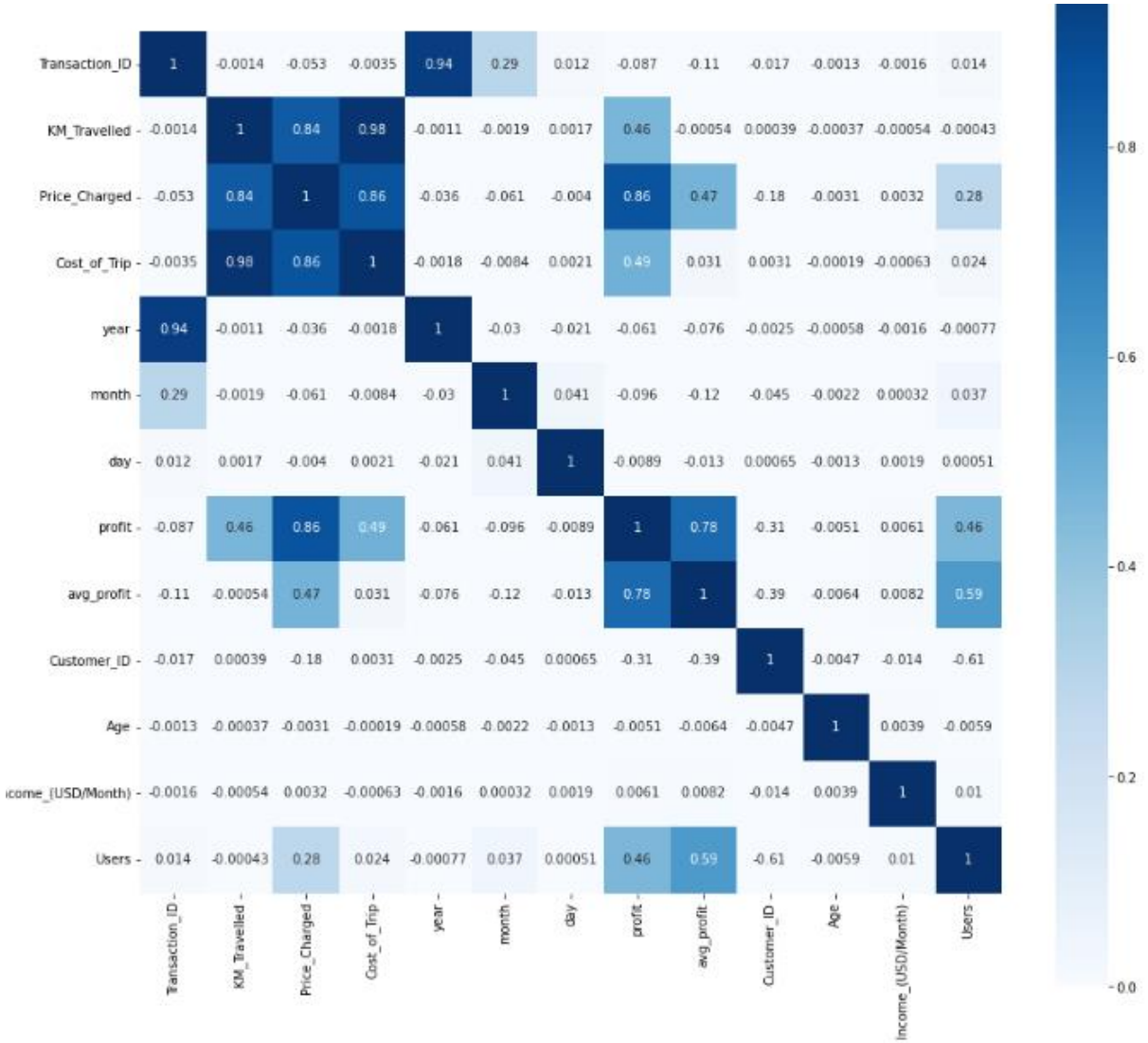
EDA





For New York City the Yellow Cab price charged is more in comparison to the other cities and for Pink cab all the cities have the same increase in prices with increase in distance

EDA



There is positive correlation between price changed and Profit Margin.

EDA Summary

- Performed some operations on dataset and found that there is positive correlation between Margin and price charged.
- Yellow cab charged more than Pink Cab
- For New York City the Yellow Cab price charged is more in comparison to the other cities and for Pink cab, all the cities have the same increase in prices with increase in distance

Recommendation

I would recommend that investor should invest in Yellow cab as

- Even if price charged is more for yellow cab than Pink cab, Margin is more for yellow cab.
- There is price change only in New York city than any other city for yellow cab as New York is Metro Politian city so price is higher.

Thank You