



Data Glacier

Your Deep Learning Partner

G2M insight for Cab Investment firm

LISUM15

2022.11.19

Introduction

Problem Statement:

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.

Analysis Outline:

Understand data sets

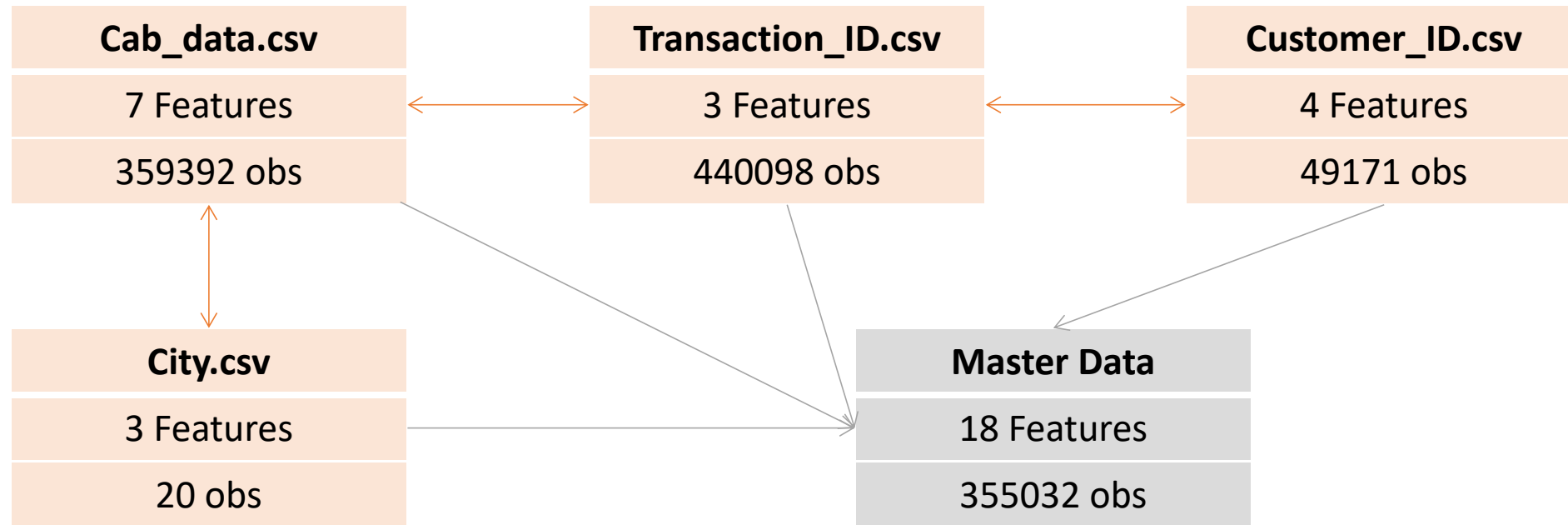
Descriptive analysis of user groups

Explore potential factors for profit

Forecast the profitability of each cab

Recommendations

Data Source



- The time period of this analysis is from 01/31/2016 to 12/31/2018.
- When creating master data, four new features are generated:
Profit, Unit Price, Unit Cost and Number of active users

Data Manipulation

Data Cleaning

- Missingness - Use `pandas.merge()` and drop the records with missing values
- Duplication - Use `df.drop_duplicates()` to drop duplicate records
- Outlier - Use boxplot to detect outliers

New Features:

$$\text{profit} = \text{price_charged} - \text{cost_of_trip}$$

$$\text{unit price} = \frac{\text{price_charged}}{\text{KM_Travelled}}$$

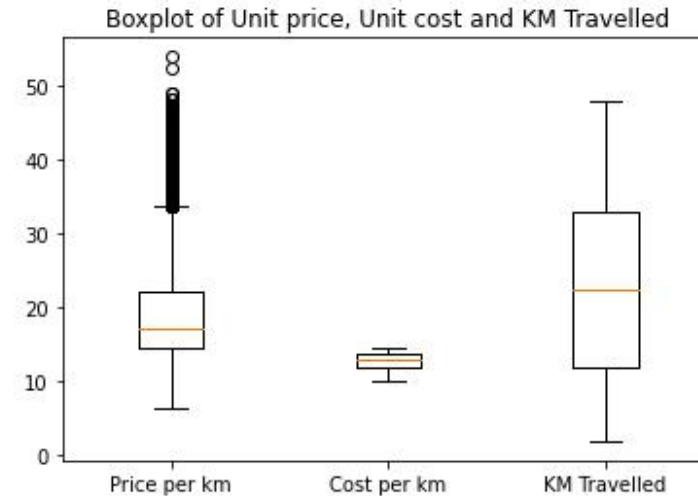
$$\text{unit cost} = \frac{\text{cost_of_trip}}{\text{KM_Travelled}}$$

Active users: Users who made transactions during this time period

User_ratio: The proportion of cab users in the population of a city

$$\text{User_ratio} = \frac{\text{Number_of_Users}}{\text{Total_population}}$$

Data Manipulation



Number of Outliers in Price Charged

Yellow Cab

Pink Cab

5861

18

Because outliers only exist in Price Charged, which may arise due to some abnormal operation and will interfere the correlation analysis between mileage and charge, so we drop these outliers, and the remaining master data is:

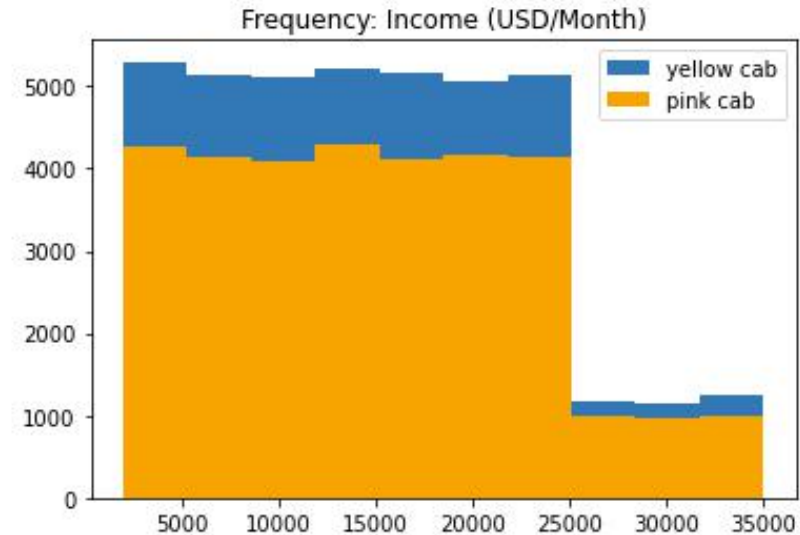
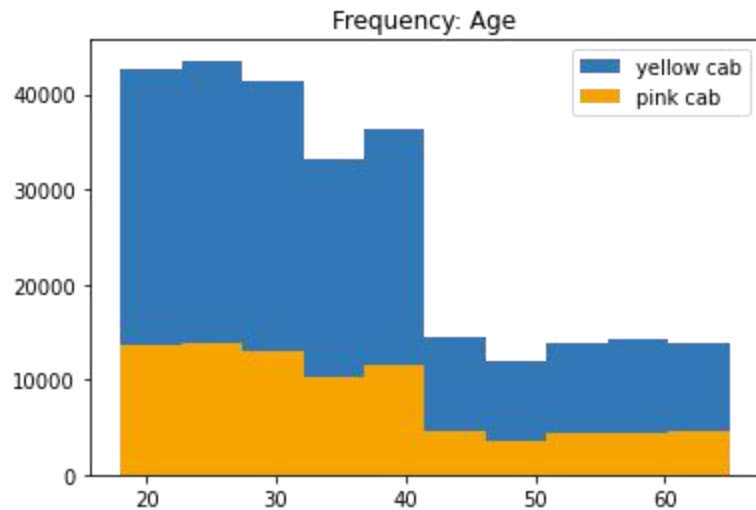
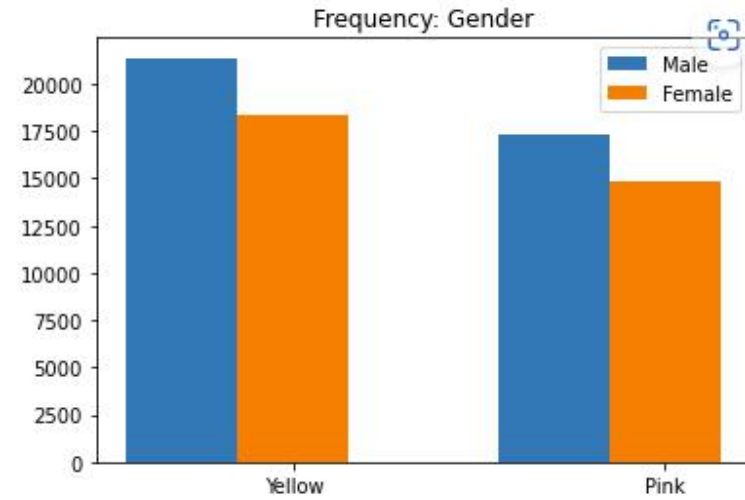
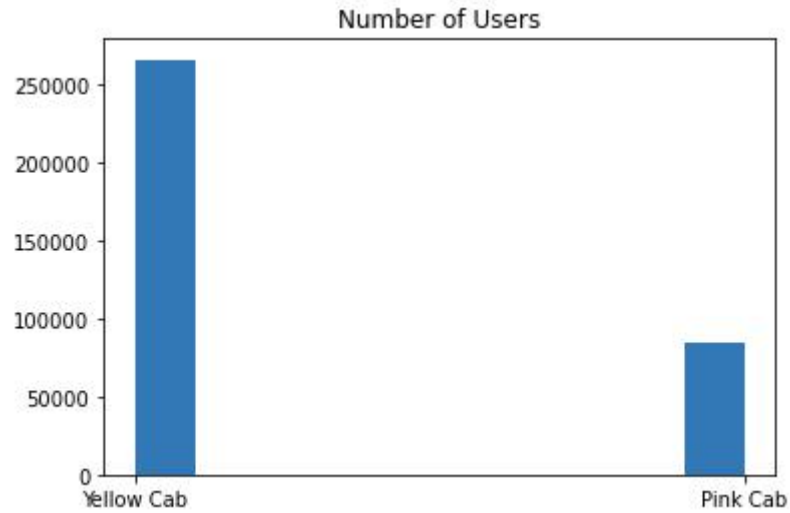
Master Data

18 Features

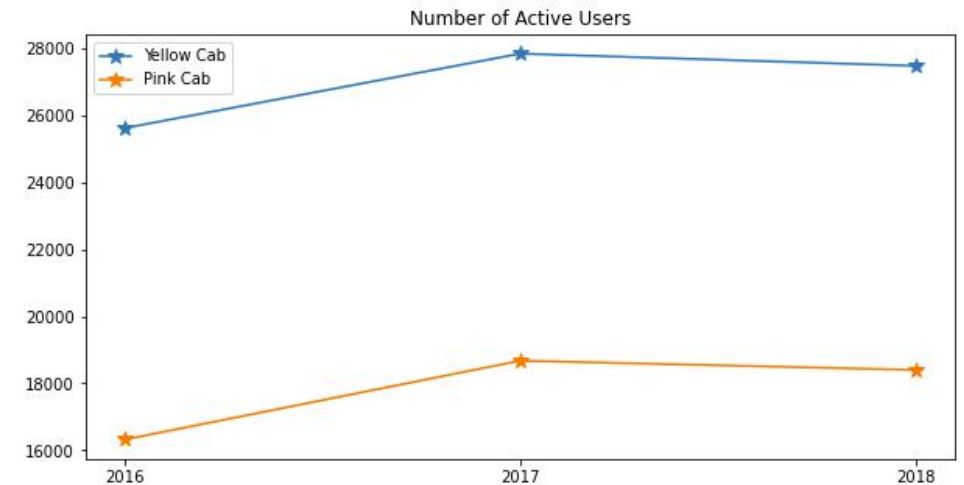
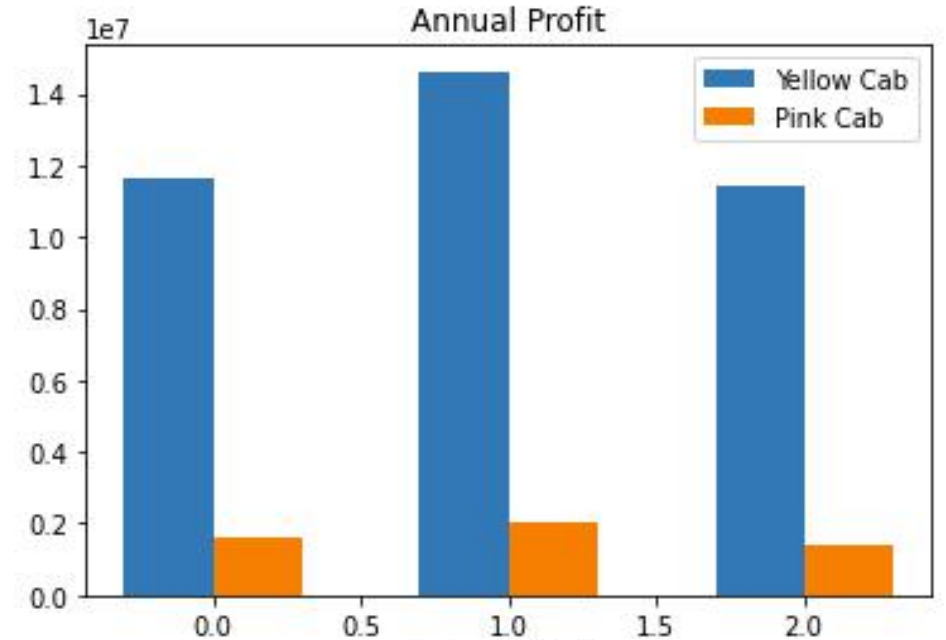
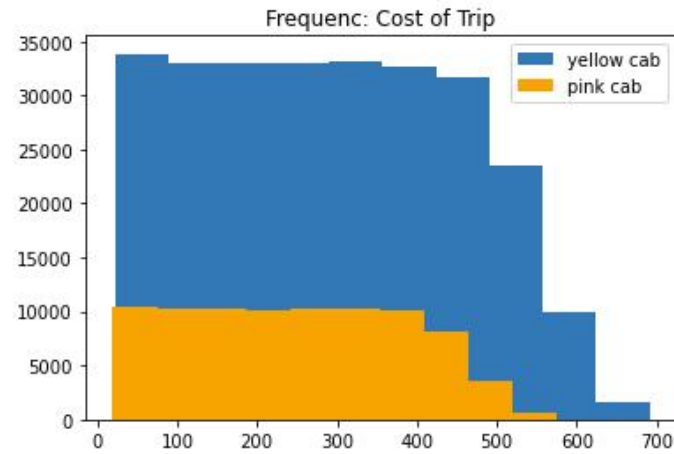
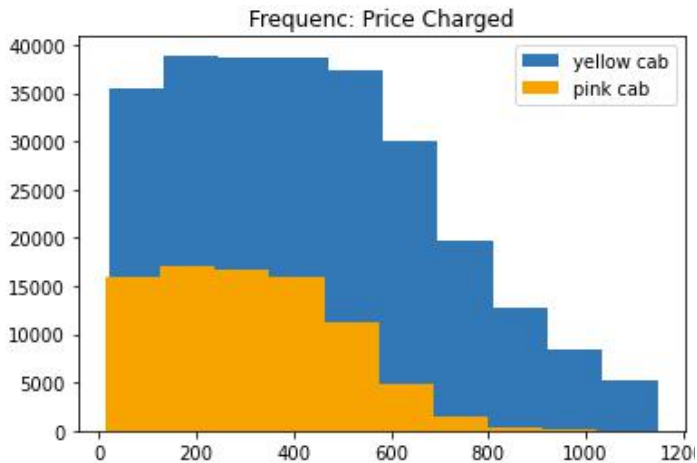
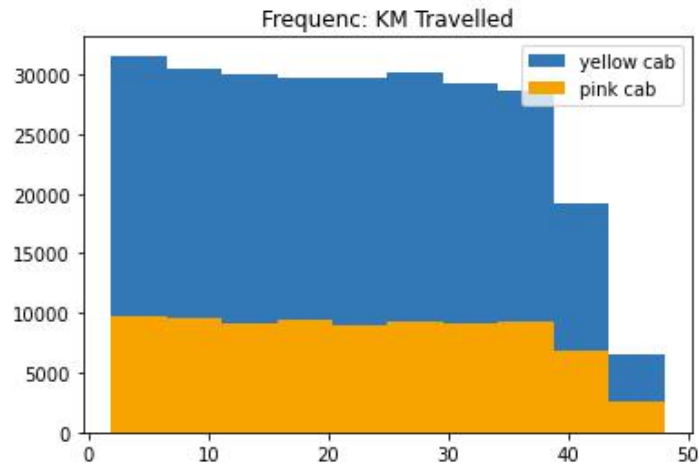
349153 obs

There is no duplication and missingness existing in this master data.

Descriptive Analysis - User groups

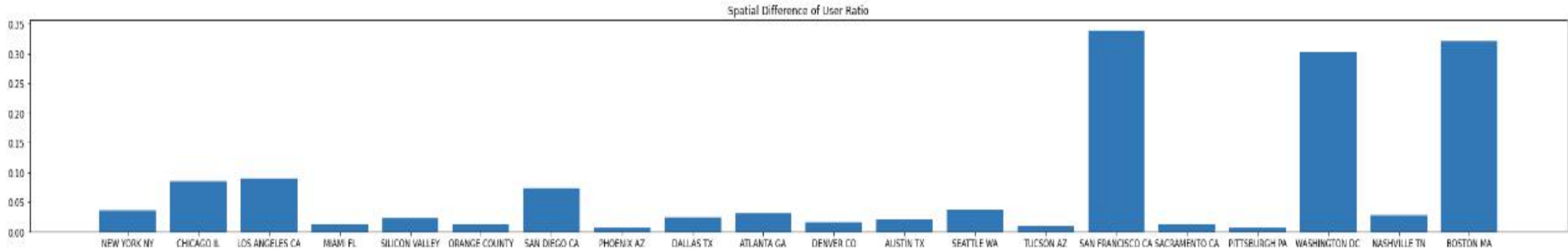


Descriptive analysis - Profitability

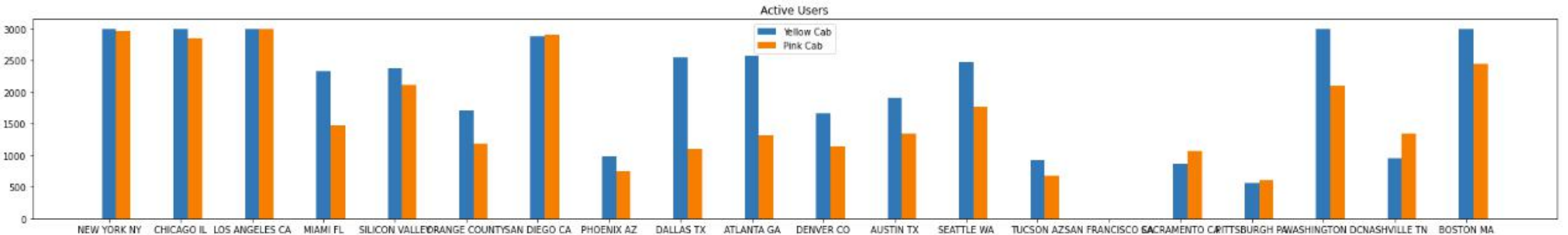


Descriptive analysis - Spatial difference

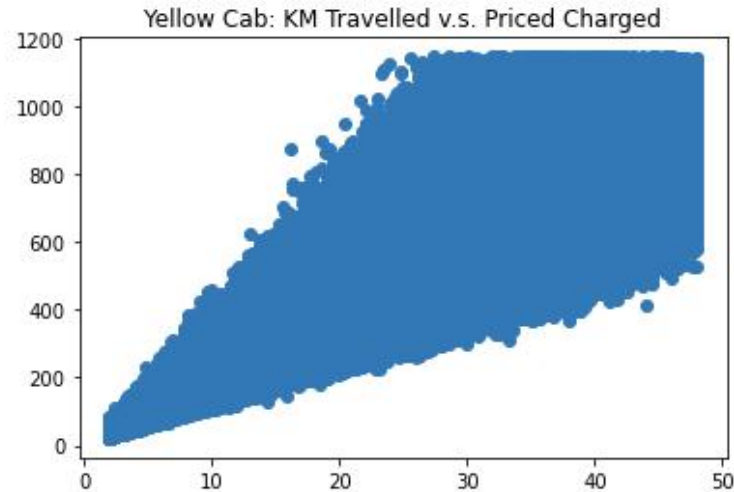
We use User_ratio to represent the cab market scale of each city.



Number of Active Users in each city



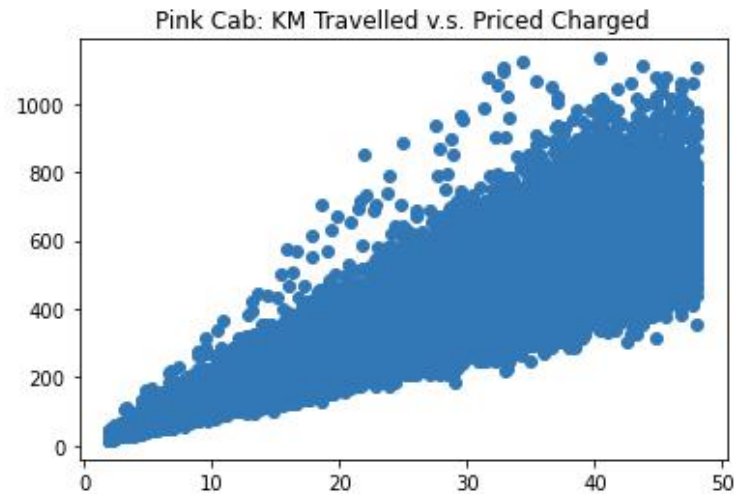
Regression analysis - Price



Yellow Cab:

$$\text{Price_Charged} = 20.23 * \text{KM_Travelled} + 0.88$$

$$R^2 = 0.74$$

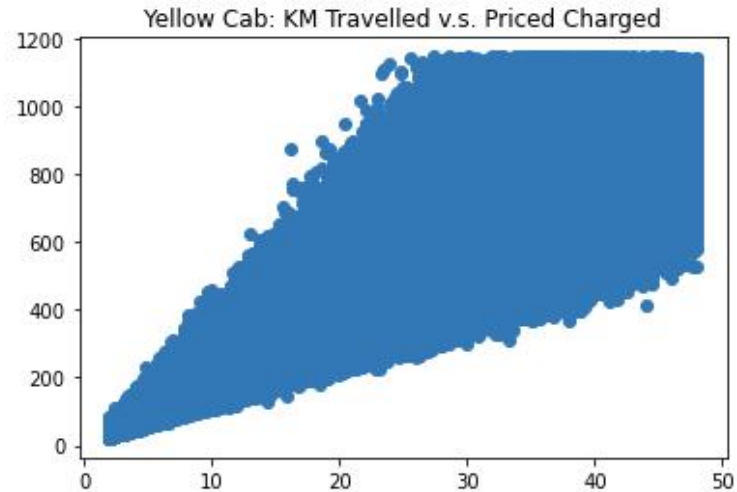


Pink Cab:

$$\text{Price_Charged} = 13.79 * \text{KM_Travelled} - 0.55$$

$$R^2 = 0.86$$

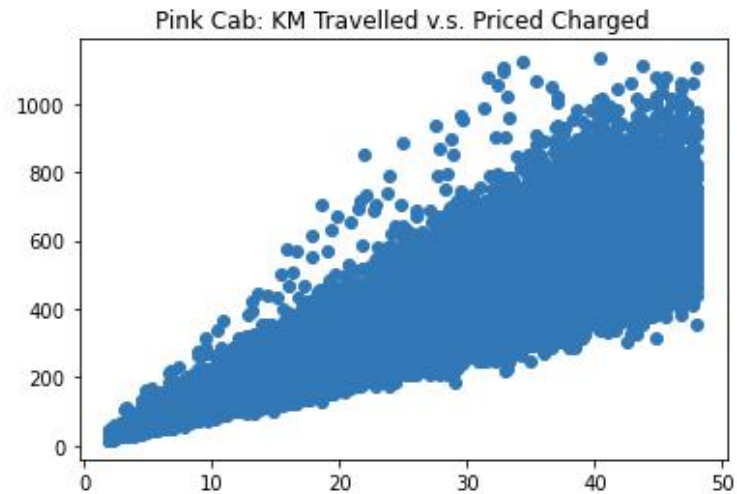
Regression analysis - Cost



Yellow Cab:

$$\text{Cost_of_Trip} = 13.20 * KM_Travelled + 0.03$$

$$R^2 = 0.987$$

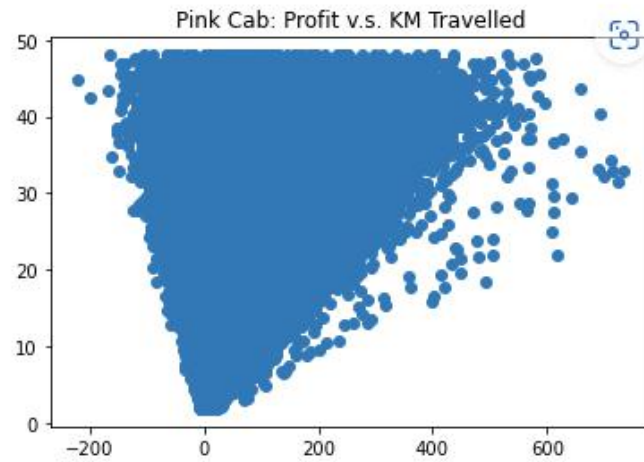
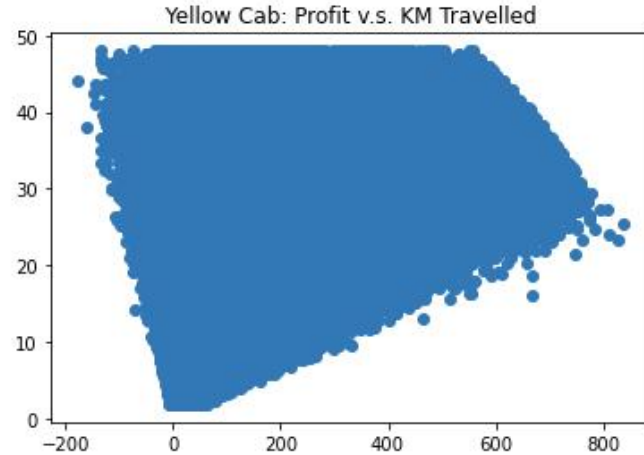


Pink Cab:

$$\text{Cost_of_Trip} = 11.00 * KM_Travelled + 0.05$$

$$R^2 = 0.987$$

Causality Analysis - Profit



Pearson Correlation Coefficient Profit v.s. KM Travelled

Yellow Cab

0.502

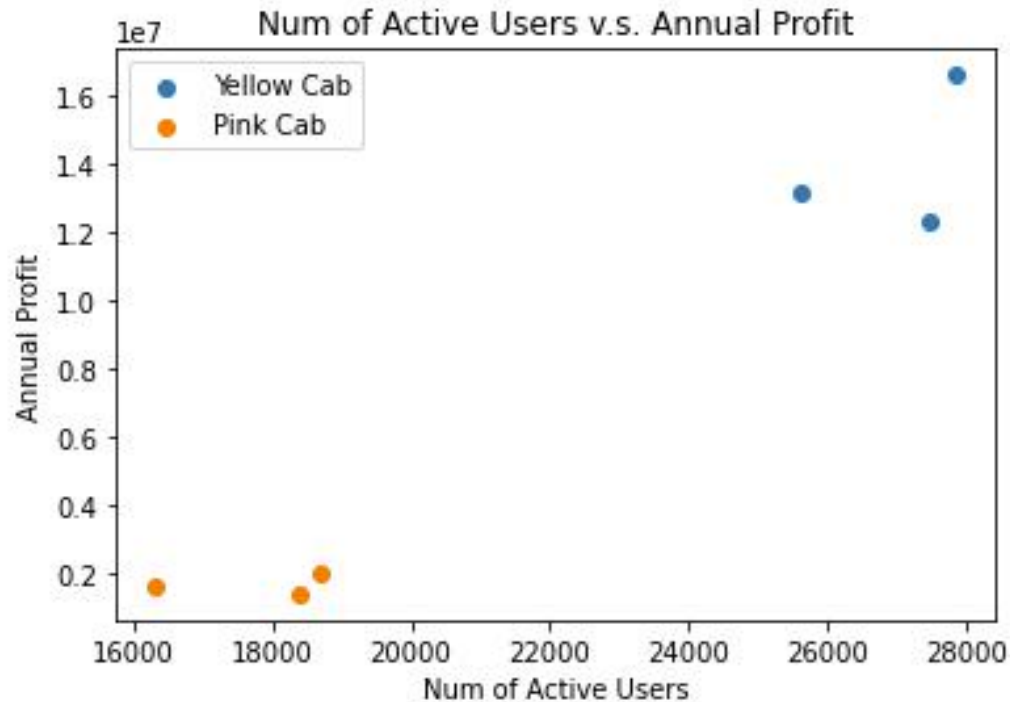
Pink Cab

0.442

Conclusion:

Profit is correlated to KM_Travelled, but not strong, so there are some other factors affecting the profit of each trip.

Causality Analysis - Profit



Pearson Correlation Coefficient Profit v.s. Num of Active Users

Yellow Cab

Pink Cab

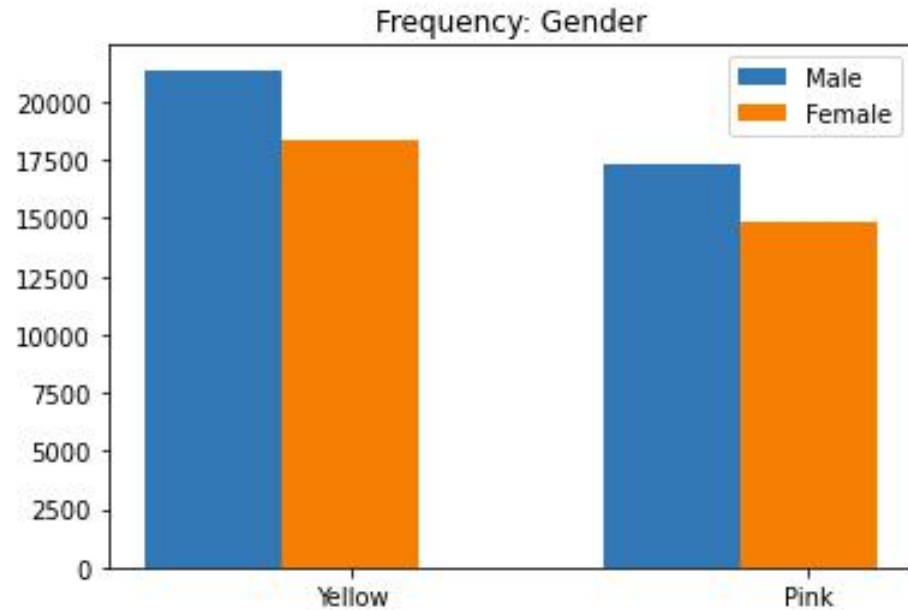
0.46

0.23

Conclusion:

Profit is correlated to Number of Annual Active Users, but not strong.

Hypothesis Test - Gender preference



Contingency Table:

	Male	Female
Yellow	21376	18379
Pink	17363	14811

Chi square test result:

p-value = 1 >> 0.05

Conclusion: there is no gender preference of choosing cabs.

Recommendation

- Main customer group:
 - under 40 years old
 - monthly income is less than 25000 USD
- Size of market:
 - The size of user group in Yellow Cab is three times larger than that in Pink Cab
 - The cab market in San Francisco CA, Washington DC and Boston MA is better developed than other cities, where Yellow Cab has a larger market share than Pink Cab. However, the transaction data of San Francisco is missing.
- Profitability:
 - Yellow Cab creates the annual profit more than six times larger than Pink Cab.
 - Yellow Cab can generate more profit per kilometre of operation than Pink Cab.
 - These two companies have similar trend of annual profit and active user retention
- Existing problems:
 - Yellow Cab has much more outlier of charged price than Pink Cab, which may indicate abnormal charging and abnormal operation.
 - Both companies have many transactions with negative profit, and there is no more detailed information to analyse the reason.

Thank You