Data Glacier

Your Deep Learning Partner

G2M insight for Cab Investment firm

LISUM15

2022.11.19

Introduction

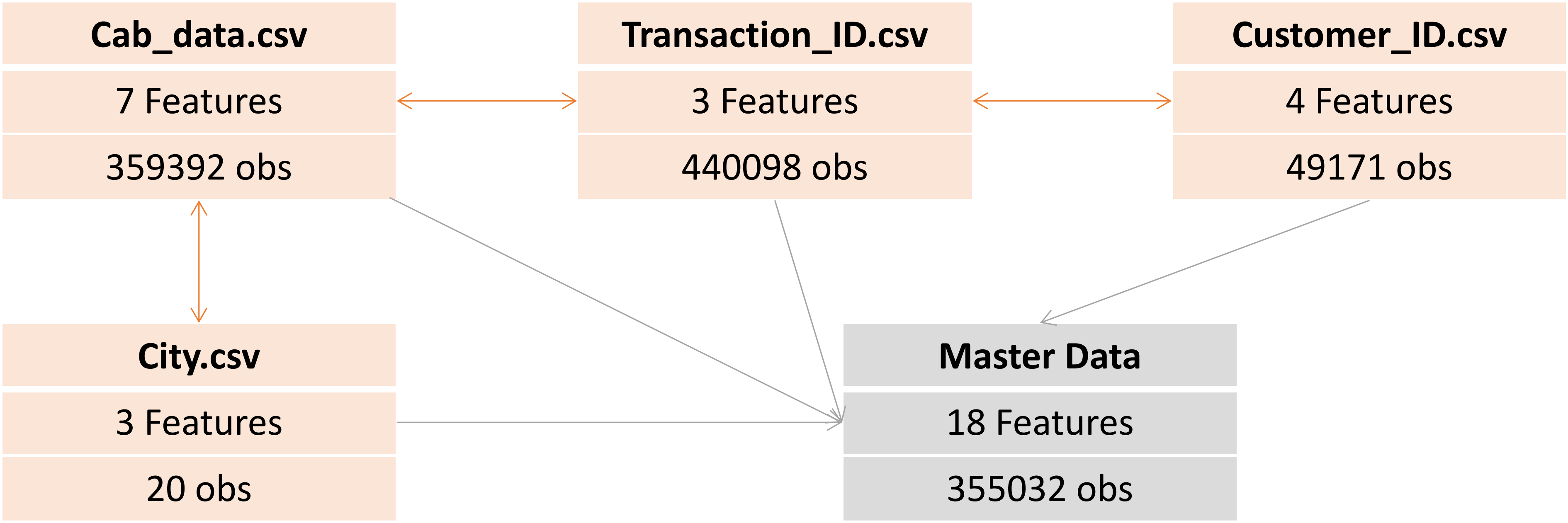
Problem Statement:

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



Cab_data.csv

7 Features

359392 obs

Transaction_ID.csv

3 Features

440098 obs

Customer_ID.csv

4 Features

49171 obs

City.csv

3 Features

20 obs

Master Data

18 Features

355032 obs

- 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:

$$\text{profit} = \text{price\_charged} - \text{cost\_of\_trip}$$

$$\text{unit price} = \frac{\text{price\_charged}}{KM\_Travelled}$$

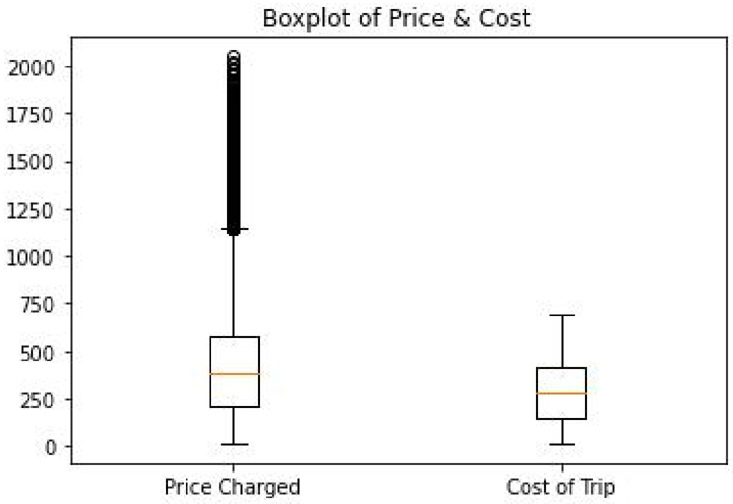
$$\text{unit cost} = \frac{\text{cost}\_of\_trip}{KM\_Travelled}$$

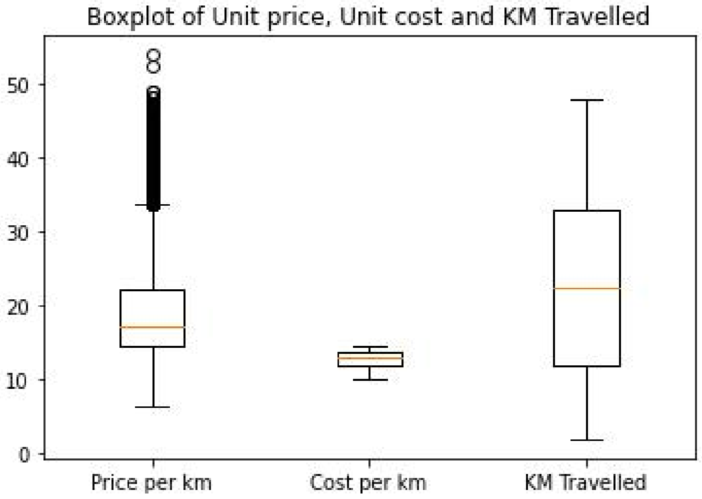
Active users: Users who made transactions during this time period

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

$$User\_ratio = \frac{Number\_of\_Users}{Total\_population}$$

Data Manipulation





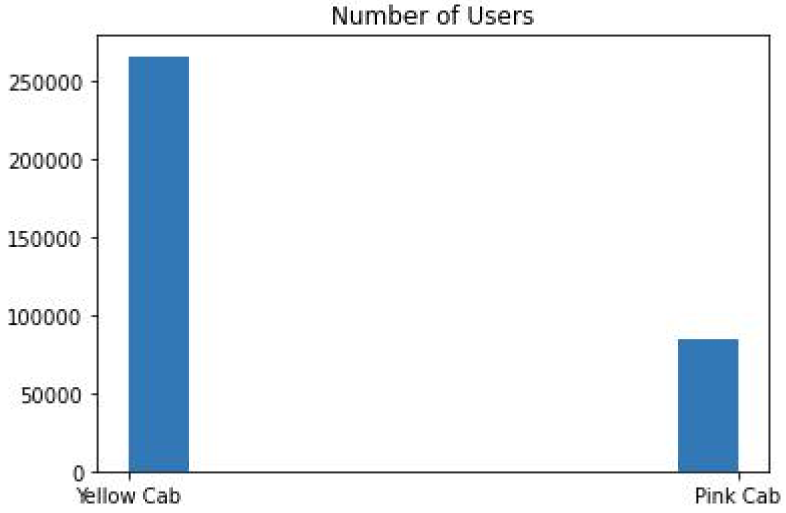
Number of Outliers in Price Charged	
Yellow Cab	Pink Cab
5861	18

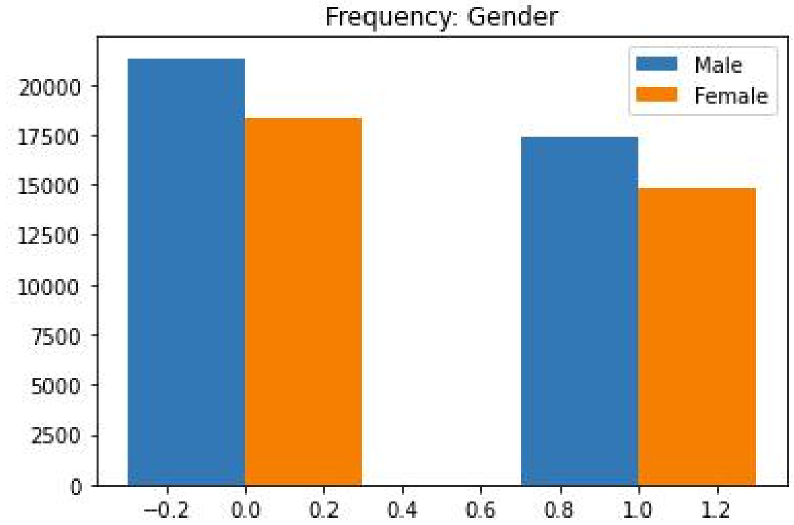
Because outliers only exist in Price Charged, which may arise due to some abnormal operation and will interfere the correlation analysis between mileage and charge, so we drop these outliers, and the remaining master data is:

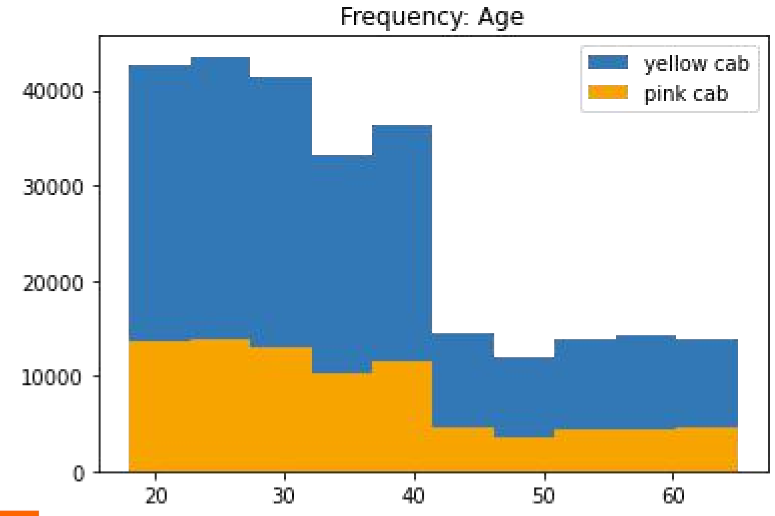
Master Data
18 Features
349153 obs

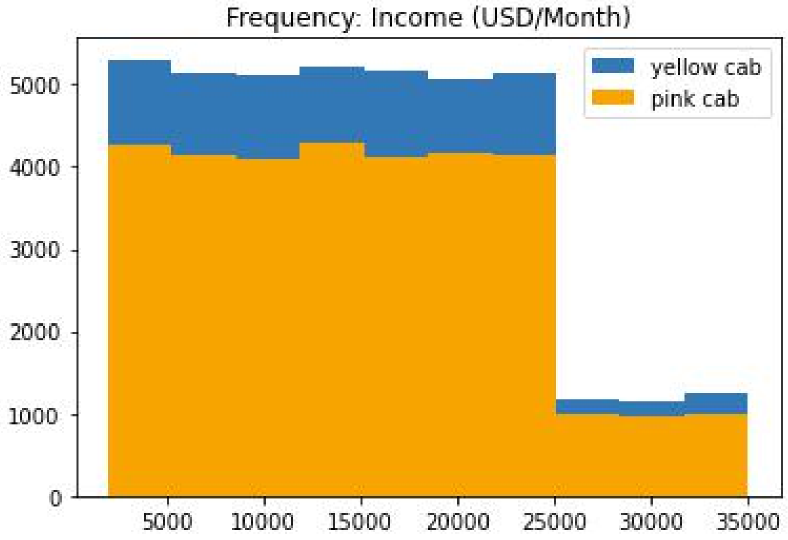
There is no duplication and missingness existing in this master data.

Descriptive Analysis - User groups

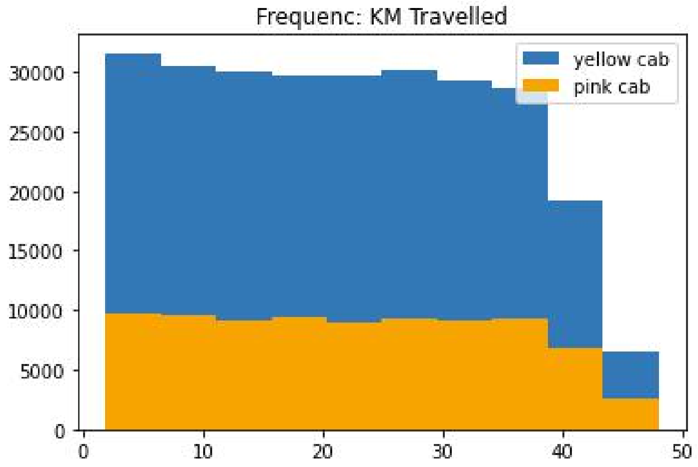

Number of Users

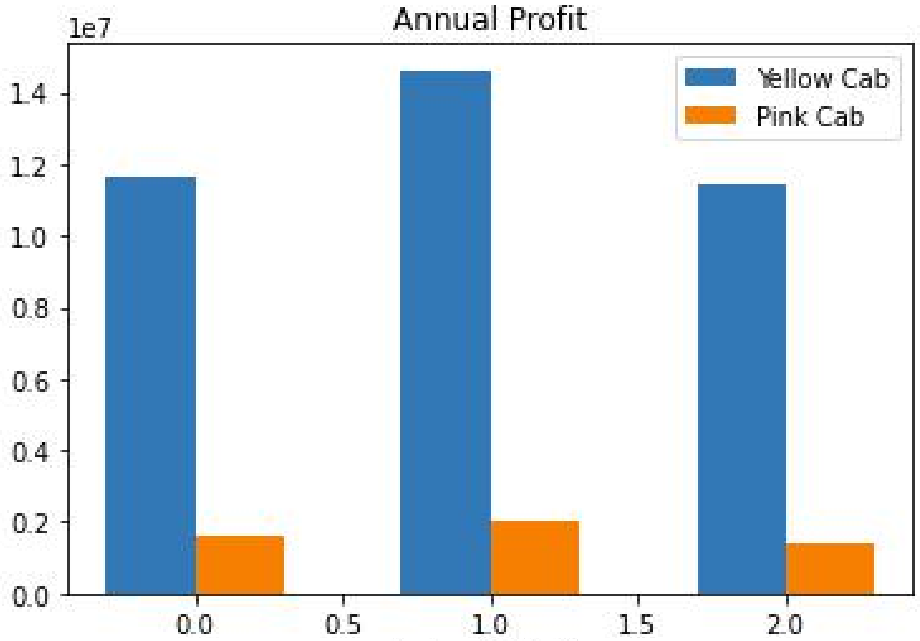

Frequency: Gender

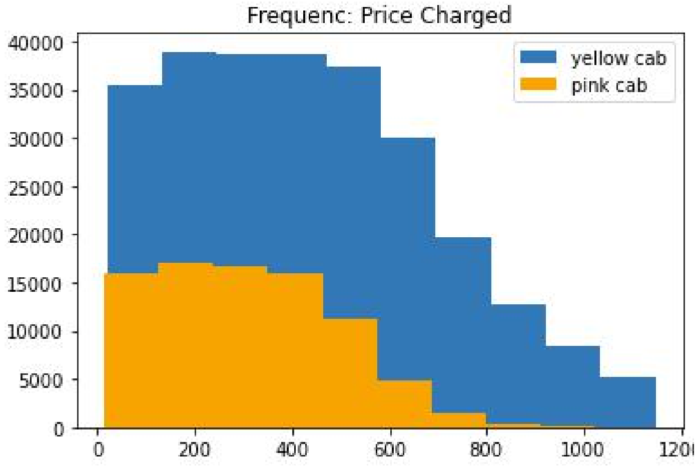

Frequency: Age

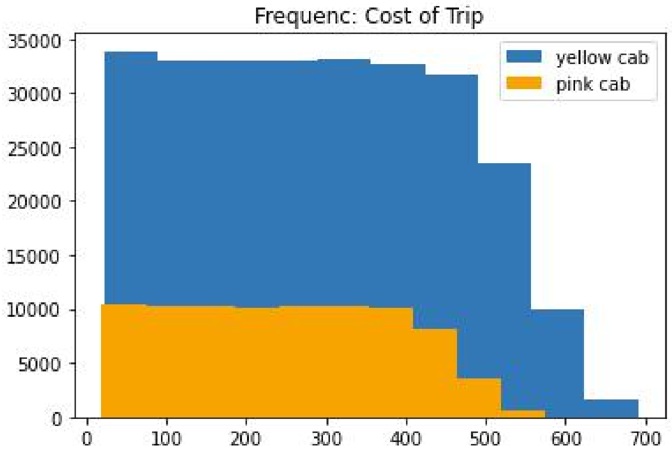

Frequency: Income (USD/Month)

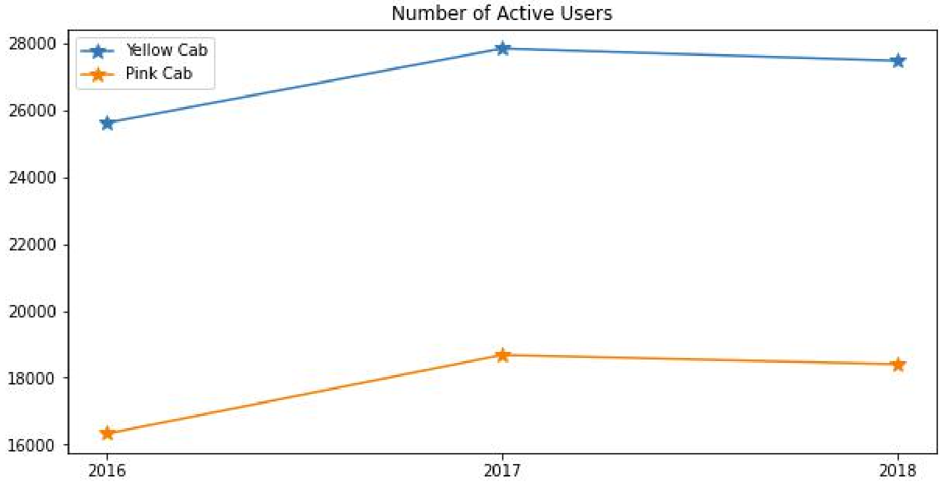
Descriptive analysis - Profitability


Frequenc: KM Travelled


Annual Profit

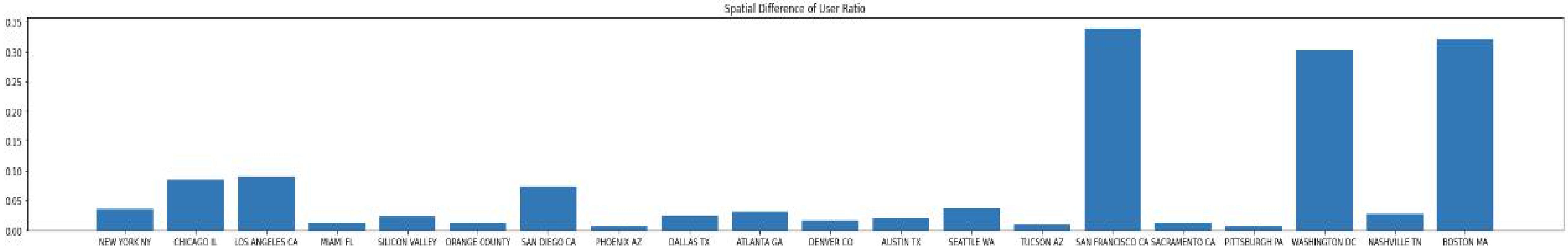

Frequenc: Price Charged


Frequenc: Cost of Trip

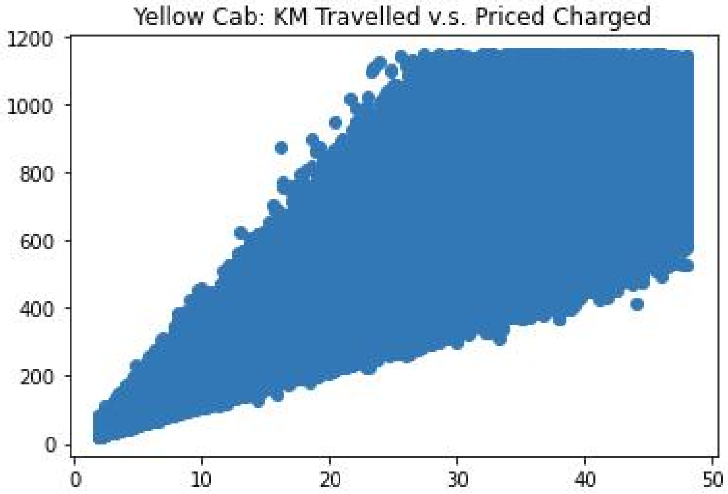

Number of Active Users

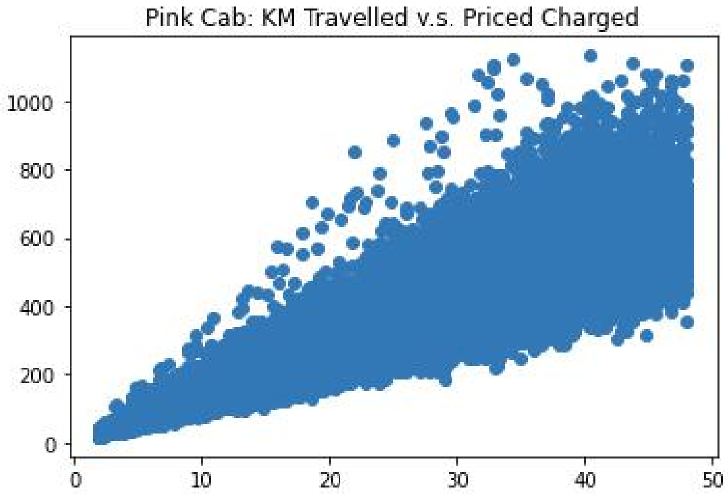
Descriptive analysis - Spatial difference

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



Regression analysis - Price





Yellow Cab:

$$\text{Price\_Charged} = 20.23 * KM\_Travelled + 0.88$$

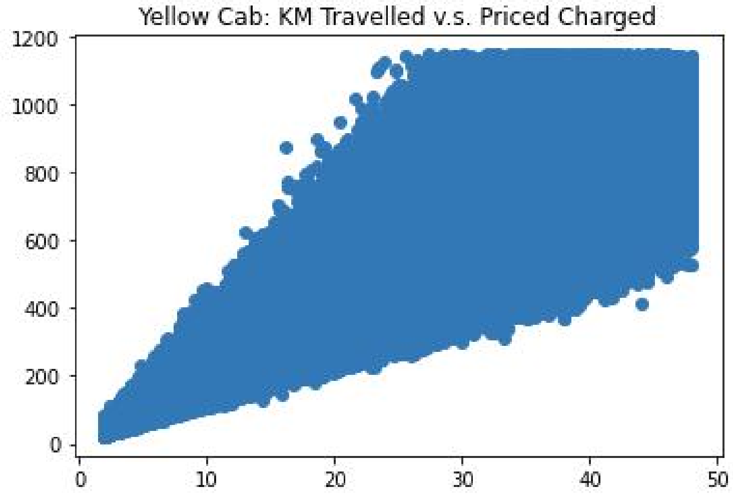
$$R^2 = 0.74$$

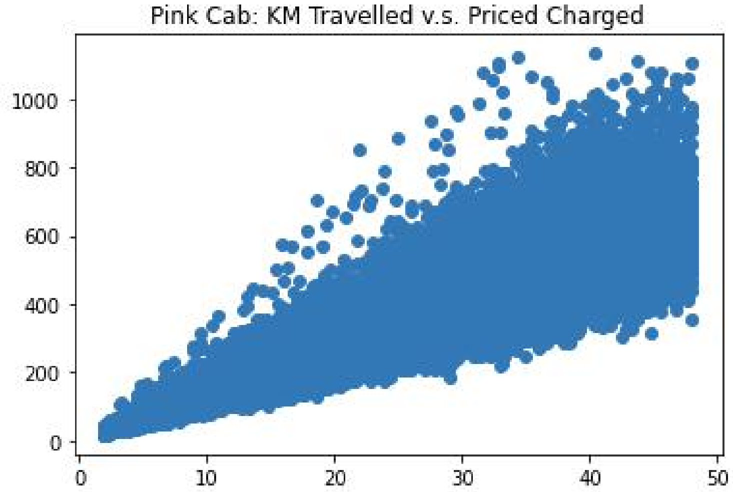
Pink Cab:

$$\text{Price\_Charged} = 13.79 * KM\_Travelled - 0.55$$

$$R^2 = 0.86$$

Regression analysis - Cost


Yellow Cab: KM Travelled v.s. Priced Charged


Pink Cab: KM Travelled v.s. Priced Charged

Yellow Cab:

$$\text{Cost\_of\_Trip} = 13.20 * KM\_Travelled + 0.03$$

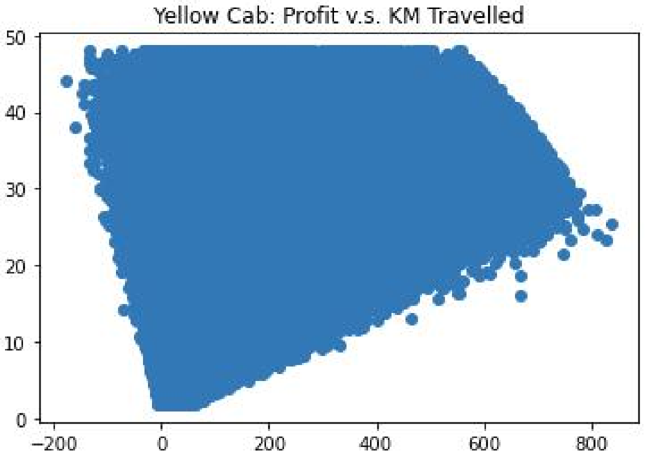
$$R^2 = 0.987$$

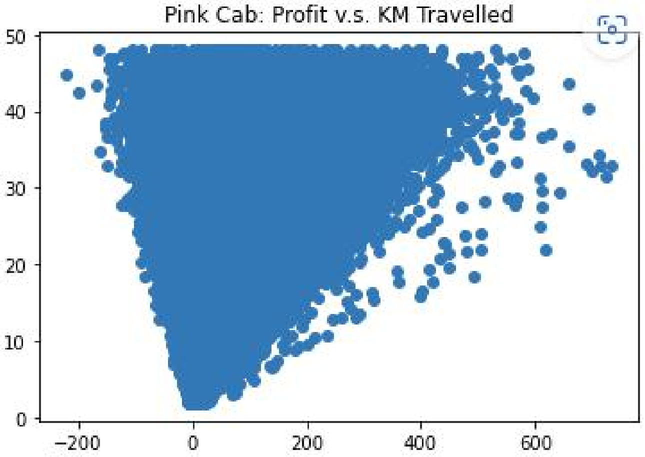
Pink Cab:

$$\text{Cost\_of\_Trip} = 11.00 * KM\_Travelled + 0.05$$

$$R^2 = 0.987$$

Causality Analysis - Profit

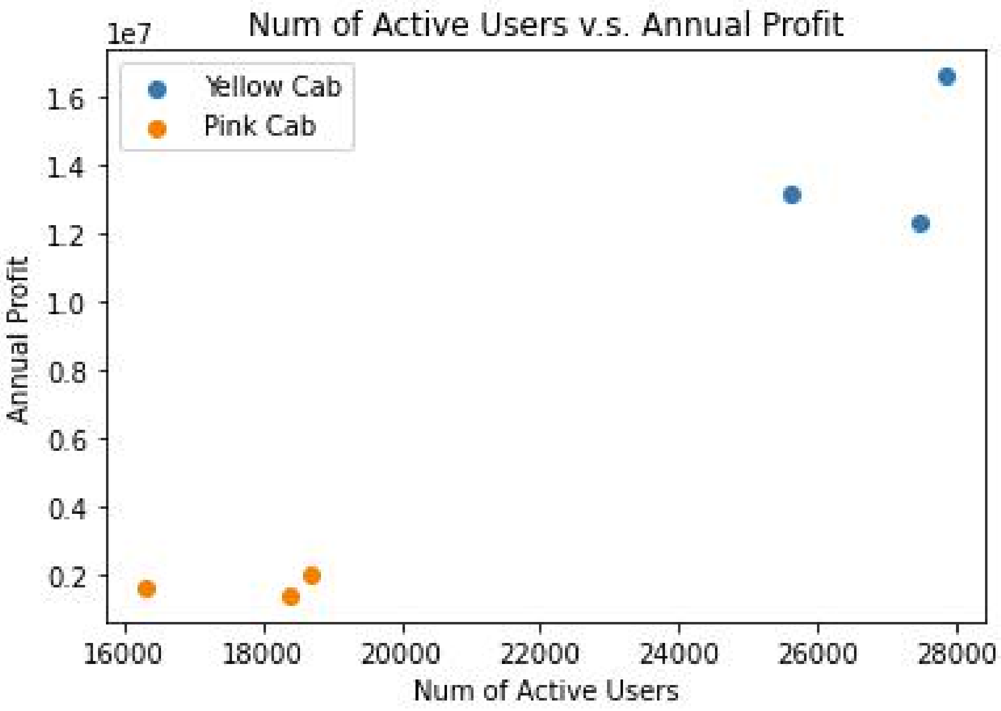




Pearson Correlation Coefficient Profit v.s. KM Travelled	
Yellow Cab	Pink Cab
0.502	0.442

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

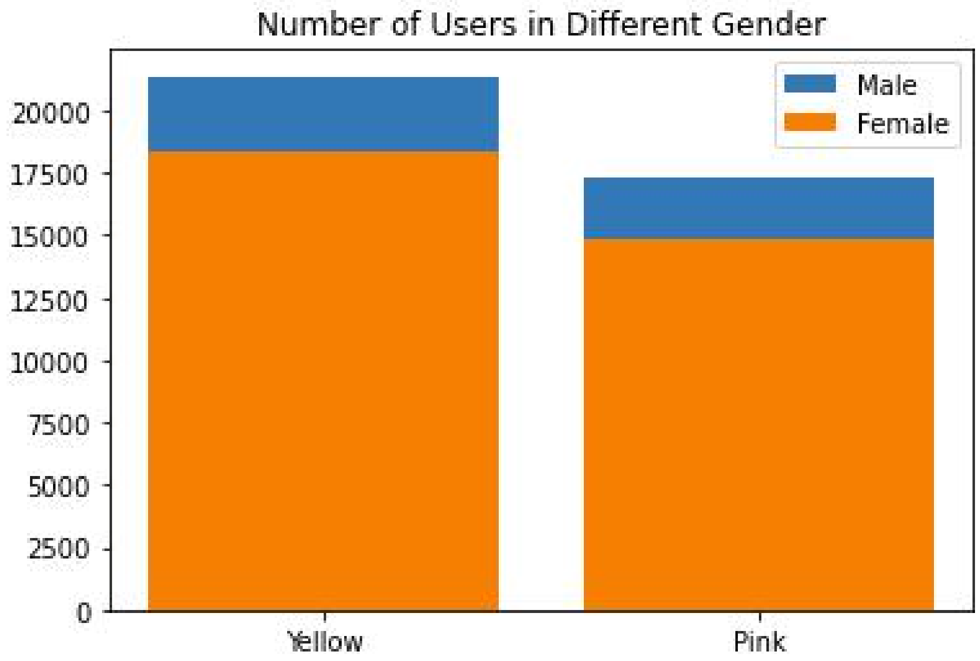


Pearson Correlation Coefficient Profit v.s. Num of Active Users	
Yellow Cab	Pink Cab
0.46	0.23

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

Thank You

Data Glacier

Your Deep Learning Partner