Investment Report of Cab Company

Name: Xiaoyan Zhang

Report Date:5/13/2024

Internship Batch

Version:1.0

Data Intake Reviewer:

Data Storage location:

Code

https://github.com/xiaoyanhouston/cab_investment/blob/main/Cab_investment.ipynb

Tabular data details:

Total number of observations	359392
Total number of files	4
Total number of features	14
Base format of the file	CSV
Size of the data	31.1MB

Proposed Approach:

Data Visualization and summary statistic

Statistic Test

The dataset for this project comprises four main components, detailing transaction records of Pink company and Yellow company spanning from January 31, 2016 to December 31, 2018.

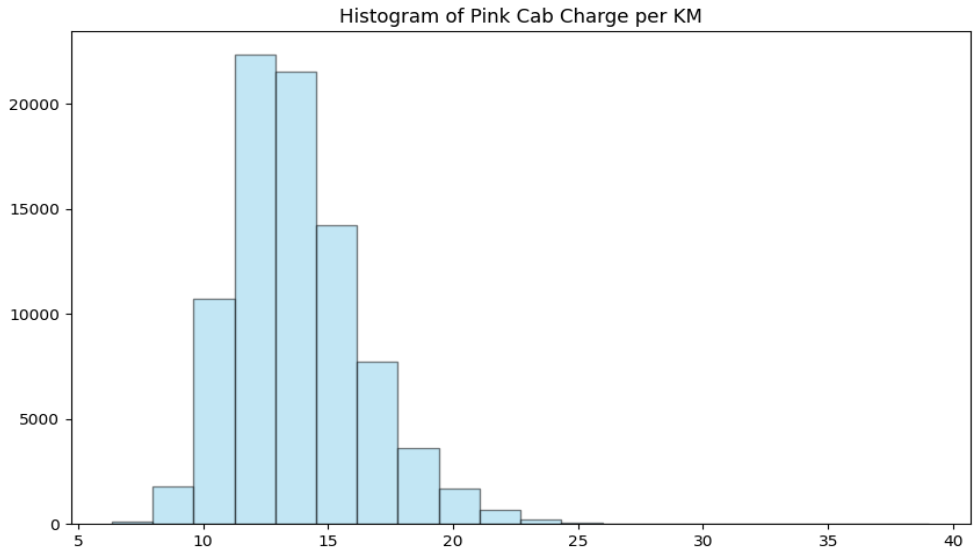
These records are categorized at the levels of company, city, customer, and individual transactions.

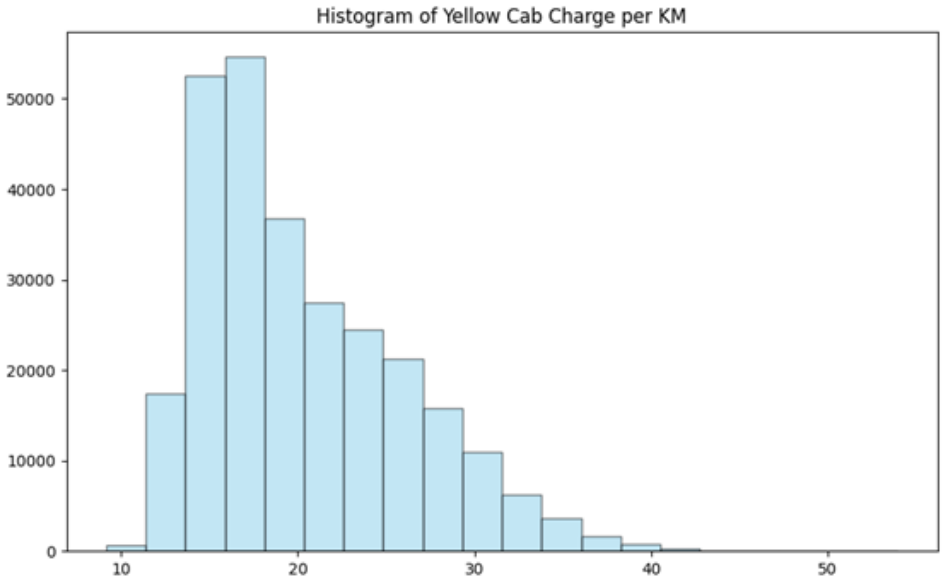
Explore Data

All code has been completed on the Machine Learning Studio in Azure. First, pandas is utilized to read in all four datasets. For the company dataset, some features possess lengthy names. To streamline the analysis process, the rename function is employed to assign shorter names. Such as “Transaction ID” to “trans_id” and “Date of Travel” to “trans_date”. Second, check each dataset whether it has missing value. Third, create tools, such as define a function to check the distribution. From the overall dataset and grouped by company, we calculated the minimum, maximum, median, and average values for each piece of data. In addition some new variables are created, such as ‘profit’, “roi”(which means the return of investment), “profit_km”, “charge_km”, “cost_km”.

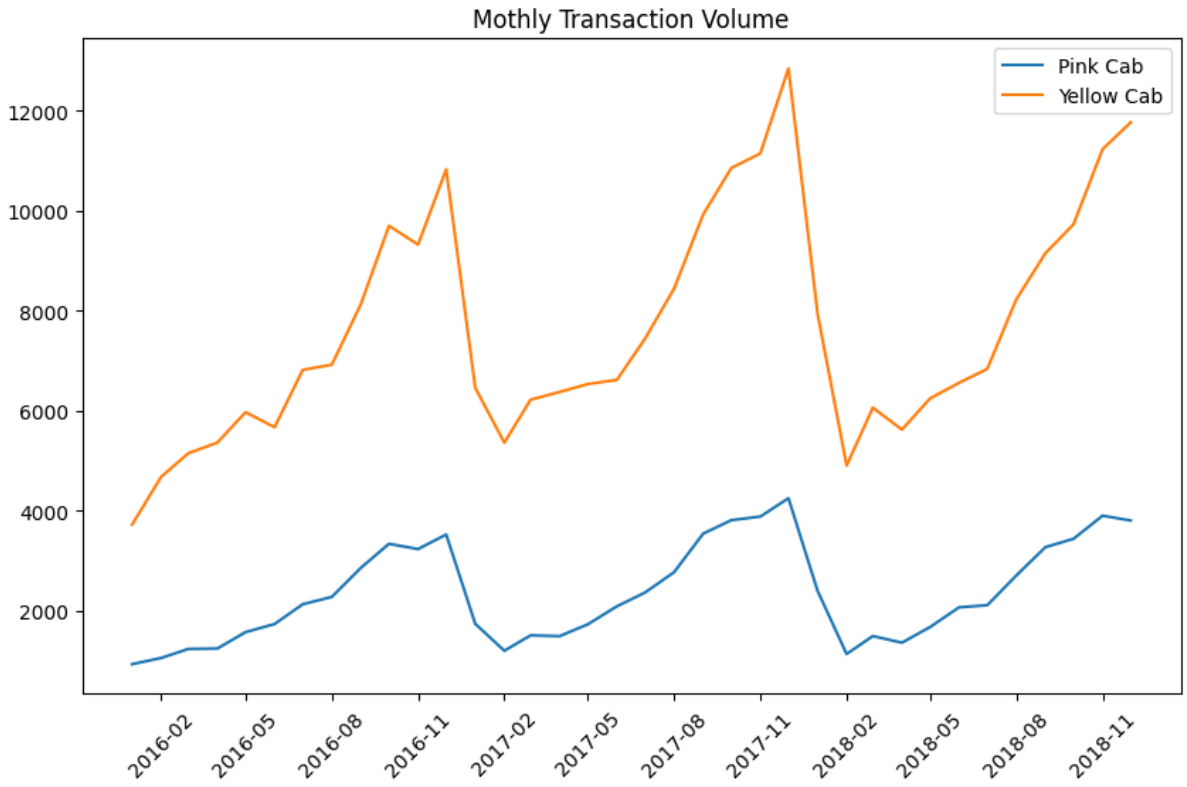
Data Analysis

From the following Plot of Histogram of Pink Cab Charge per KM, it shows the charge of per kilometer in Pink Cab company is mainly between 12 to 17 and it is bell shaped. The plot of Histogram of Yellow Cab company is mainly between 15 to 20 and it is right skewed.

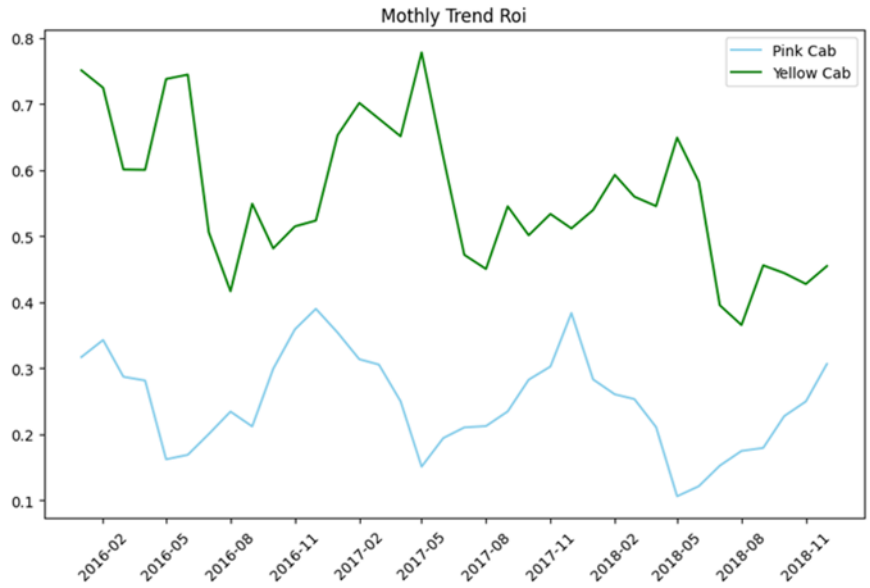




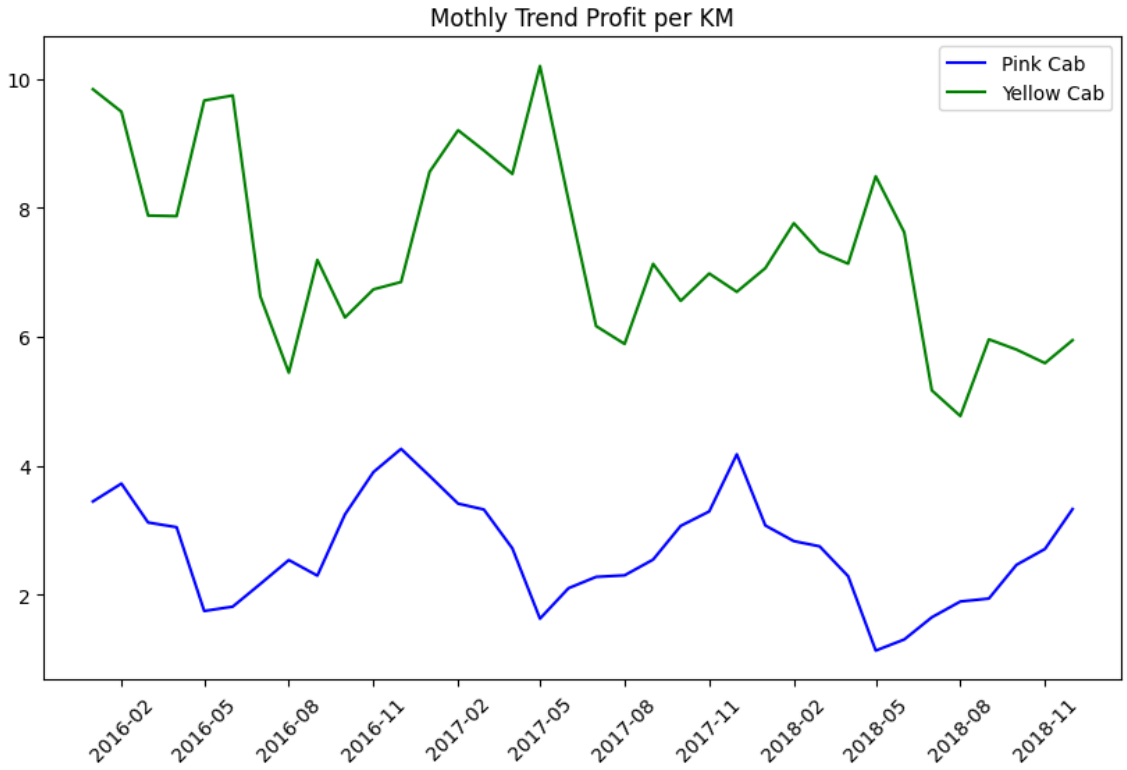
The Monthly Transaction Volume Plot tell us the volume from Yellow company is much larger than Pink company.



From the perspective of monthly return on investment show us the Yellow company is better than Pink company.



From the plot of Monthly Trend Profit per Kilometer we can see the Yellow company's profit per KM is better than the profit per KM of Pink company.



Without statistic test, sometimes it is too subjective to make decision only based on data visualization. So, we need to use statistic test to confirm our recommendations.

Test 1

H0: The profit per kilometer of Pink company and Yellow company are the same

H1: The profit per Kilometer of Pink company and Yellow company are different

Compared with the profit per Kilometer from Pink company and Yellow company, the result is t-statistic: `-160.3715175947806, p-value:0.0, reject the null hypothesis. The profit per kilometer of Pink company and Yellow company are different. According to the result from hypothesis test it is consistent with the visualization plot. The profit per kilometer of two Cab company should be one of the indicators that the investment company considered before investing`

Test 2

H0: The mean of return of investment of Pink company is greater than Yellow company

H1: The mean of return of investment of Pink company is less than Yellow company

With the result from the hypothesis test, `t-statistic: -179.2759885217892, p-value:0.0, Reject the null hypothesis. The mean of roi for Pink Cab is significantly smaller than the mean of roi for Yellow Cab.`

The return of investment of the two company should be the other indicators that the investment company should considered before investing.

Test 3

H0: The average age of customer from Pink company is the same as the customer from Yellow company

H1: The average age of customer from Pink company is different as the customer from Yellow company

The result from t-test is `t-statistic: 0.31633759975120346 p-value: 0.7517471890396208 Fail to reject the null hypothesis. There is no`

significant difference between the average age of Pink company and Yellow company.

The average age of customer from Pink company and Yellow company is the same. This is not the indicators that the investment company should considered before investing.

Test 4

H0: The average income of customer from Pink company is the same as the customer from Yellow company

H1: The average income of customer from Pink company is the different from the customer from Yellow company

The result from Hypothesis test is t-statistic: 0.838939895792237 p-value: 0.40150581507996785 Fail to reject the null hypothesis. There is no significant difference between the average income of Pink company and Yellow company.

It shows us that the average income can not be used as an indicator for investment company to considered.

Test 5

H0: The gender distribution of customer from Pink company is the same as Yellow company

H1: The gender distribution of customer from Pink company is different from Yellow company

The result from the Chi-square test is Chi-square statistic: 0.5063792600762221 P-value: 0.47671046144863083 Degrees of freedom: 1 Expected frequencies: [[14866.89405477 17463.10594523] [18346.10594523 21549.89405477]]

Since the p-value (0.4767) is greater than the chosen significance level (commonly 0.05), we fail to reject the null hypothesis. This means that there is not enough evidence to conclude that there is a significant association between the gender of customers and the company they choose.

Recommendation:

Customer Gender: There is no difference between the customer gender of Pink company and Yellow company.

Customer Age: There is no difference between the average age of customer from Pink company and Yellow company

Customer Income: There is no difference between the average income of customer from Pink company and Yellow company

Volume: The volume from the Yellow company is larger than the Pink company

Return of Investment: The return of investment of Yellow company is higher than the Pink company

Profit of Per KM: The profit of per km from Yellow company is higher than the Pink company

Conclusion: The recommendation is to invest in Yellow Cab company.