Exploratory Data Analysis

G2M insight for Cab Investment firm

Date: 21-April-2023

Agenda

Problem Statement

Data Exploration

EDA

Hypothesis Testing

EDA Summary

Recommendations

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.

Objective

Analyze the provided data sets, identify key trends and insights, and provide actionable recommendations to XYZ's executive team to help them identify the right cab company to make their investment.

Data Exploration

The Data is taken from GitHub. It contains four individual data sets.

Below is the list of datasets that are provided for the analysis:

Cab_Data.csv – This file includes details of transactions for two cab companies (Yellow and Pink Cab).

Customer_ID.csv – This is a mapping table that contains a unique identifier that links the customer's demographic details.

Transaction_ID.csv – This is a mapping table that contains transaction to customer mapping and payment mode.

City.csv – This file contains a list of US cities, their population and the number of cab users.

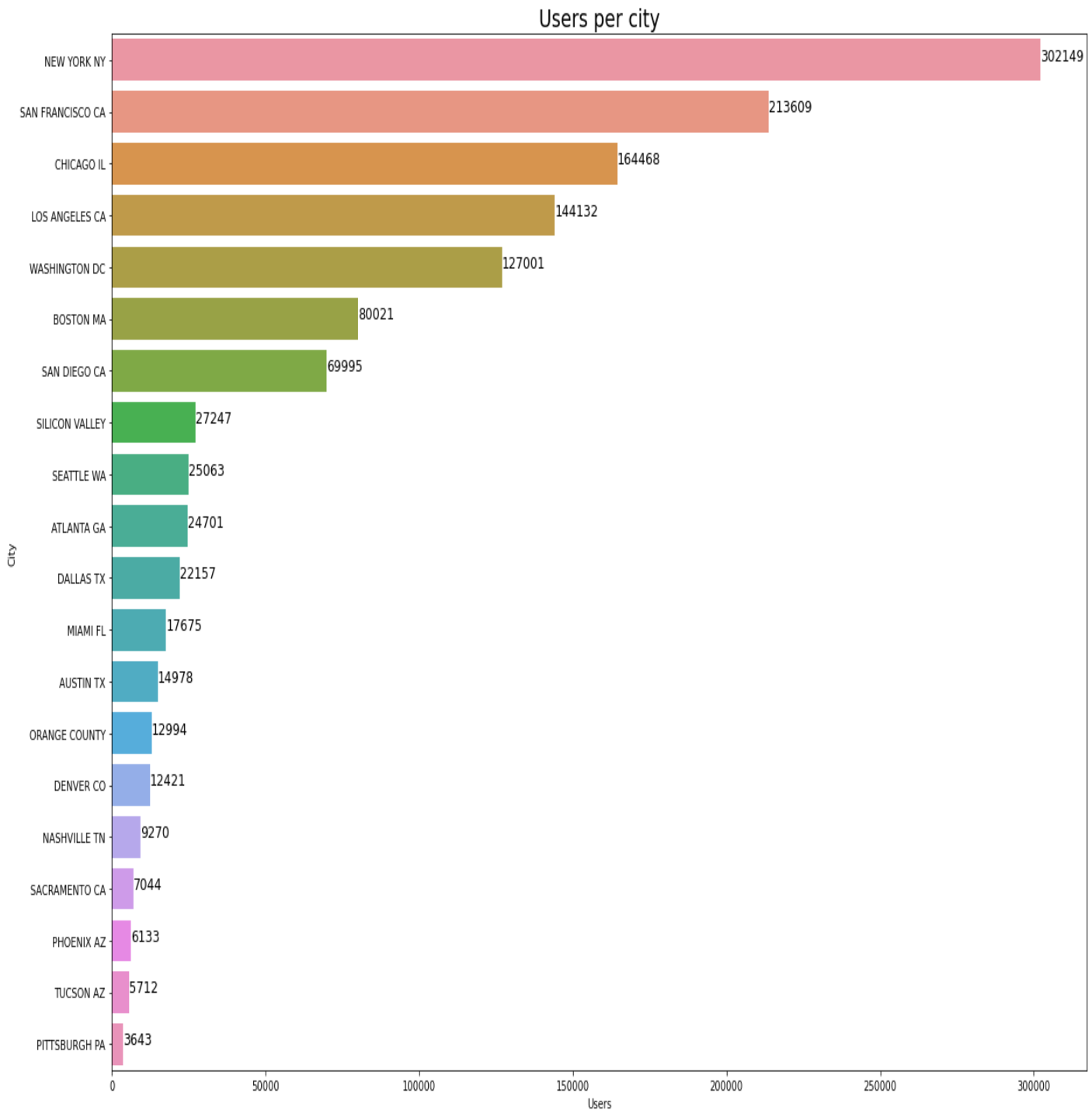
Data Exploration

- The time period of data is from 31-01-2016 to 31-12-2018
- There is No Null and Duplicate values present in this dataset.
- To analyze this dataset we have to preprocess this dataset by converting the Date of Travel column into Date format. Also we are adding the Profit, Weekday, Year, Month columns to our dataset.
- To analysis the dataset we have to merge all the dataset into one table.

Exploratory Data Analysis

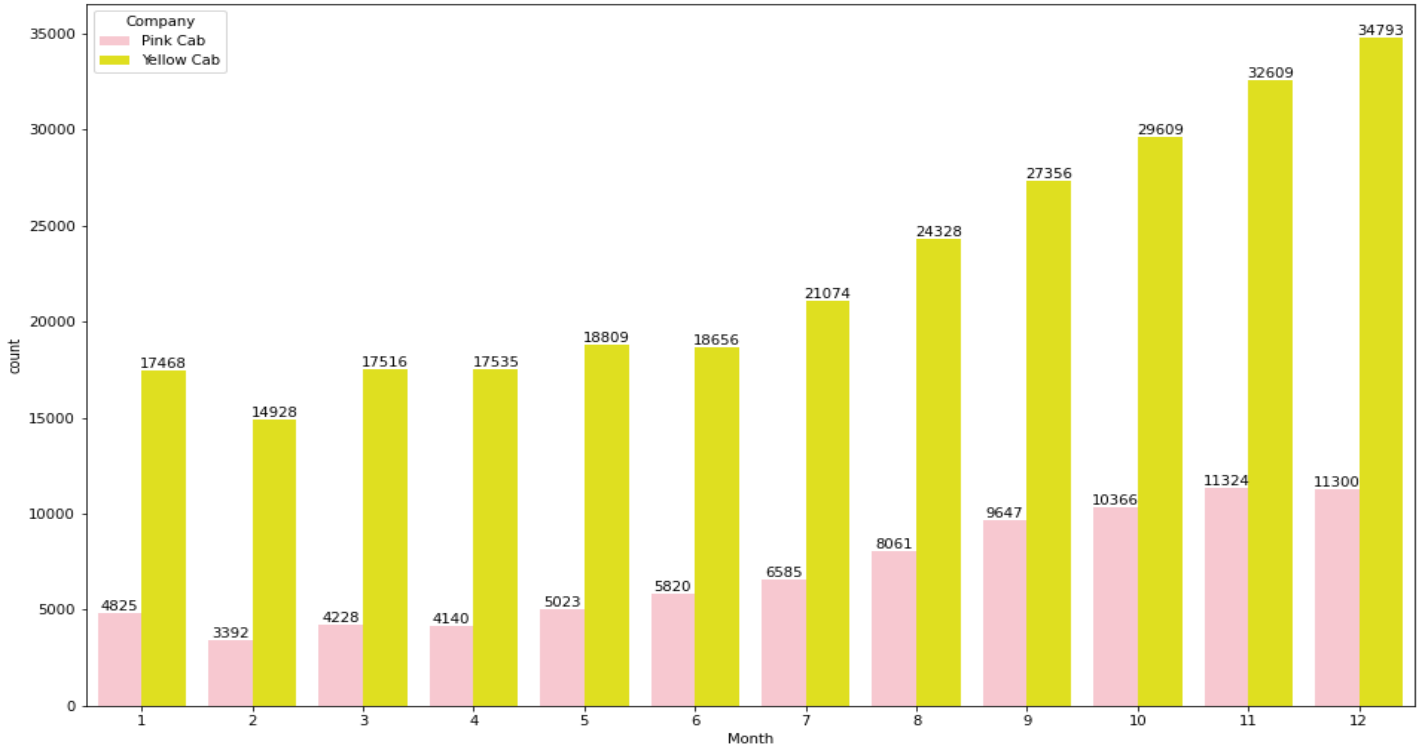
Users Per City

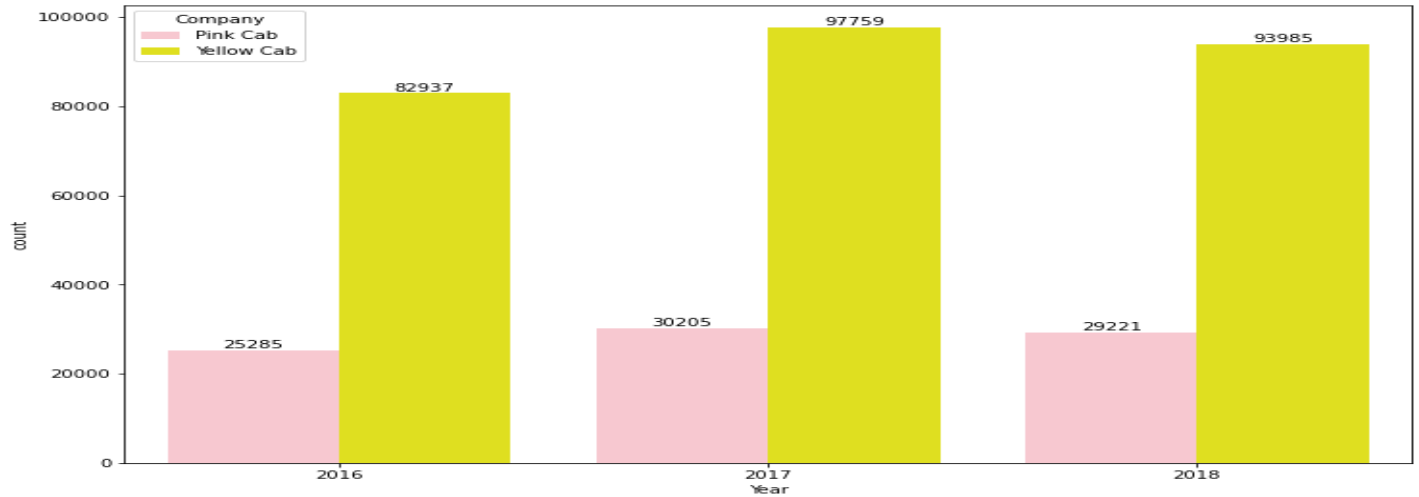
- From the plot we can say that ‘New York’ has the highest number of cab users and ‘Pittsburgh’ has the lowest number of cab users.



Travel Frequency By Month and Year

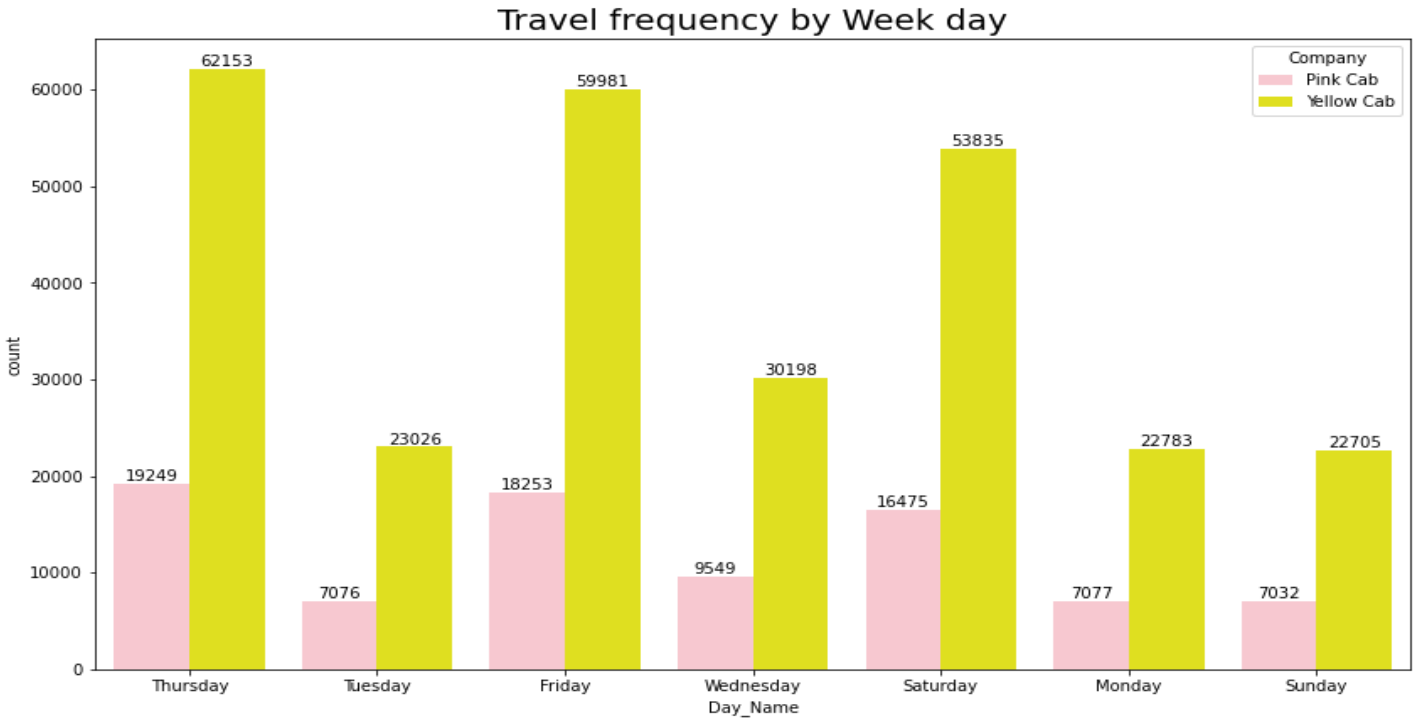
- From the Month Bra plot we can say that In December higher travelers use cab services it is because of the holiday season.
- While comparing we can say that the Yellow cab company has higher customers compared to the Pink cab.
- From the Year Bar plot we can say that in 2017 there are higher travelers using cab services compared to Years 2016 and 2018.





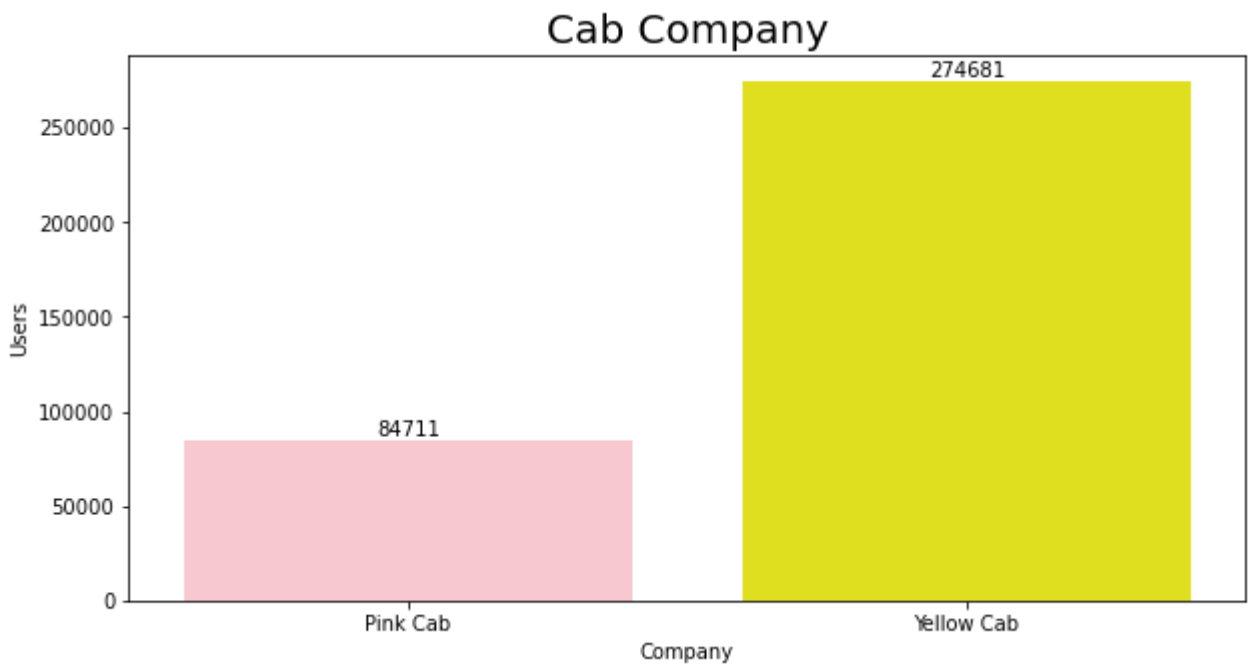
Travel Frequency By Week Days

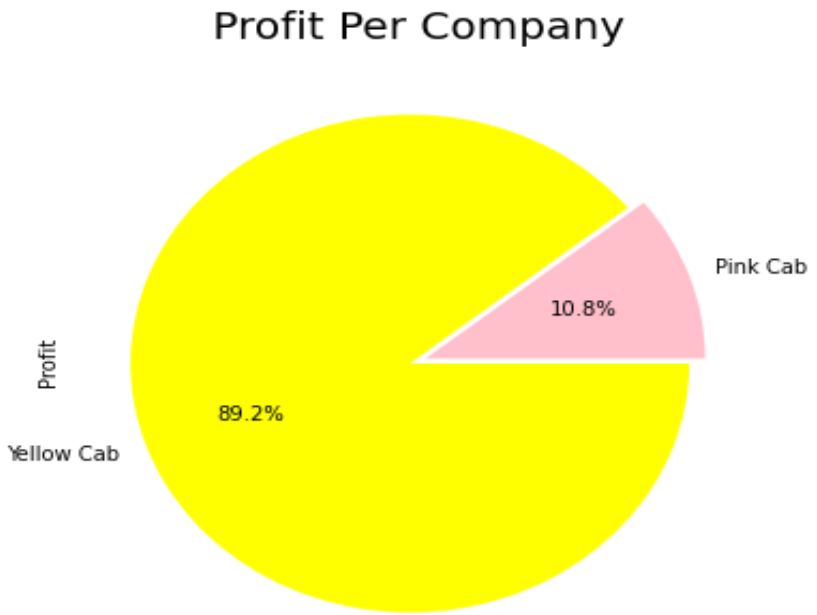
- From the Bar plot we can say that on Thursday there are higher travelers using cab services.
- On Weekends there are fewer travelers who use cab services.



Compare Users and Profit

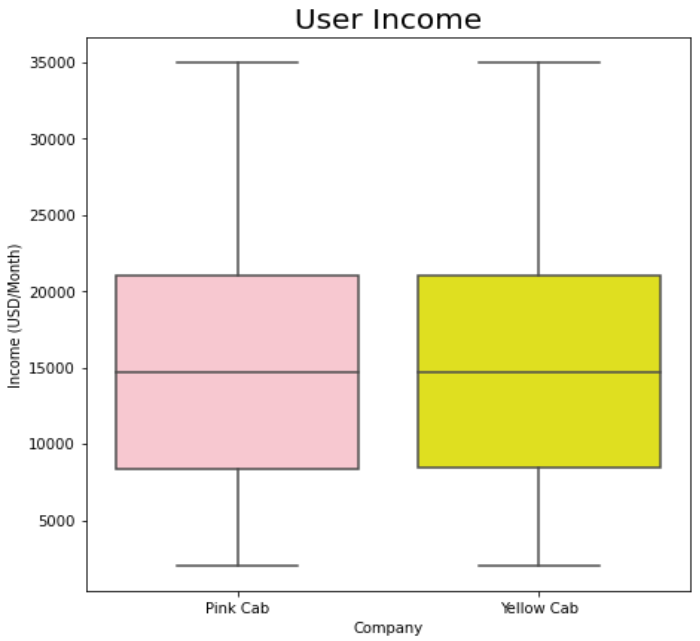
- From the Bar Plot we can say users like to ride in Yellow cabs compared to Pink cab
- Also, From the Pie chart we can say that Yellow Cab has high profit compared to Pink Cab. It is because it has more number of users.

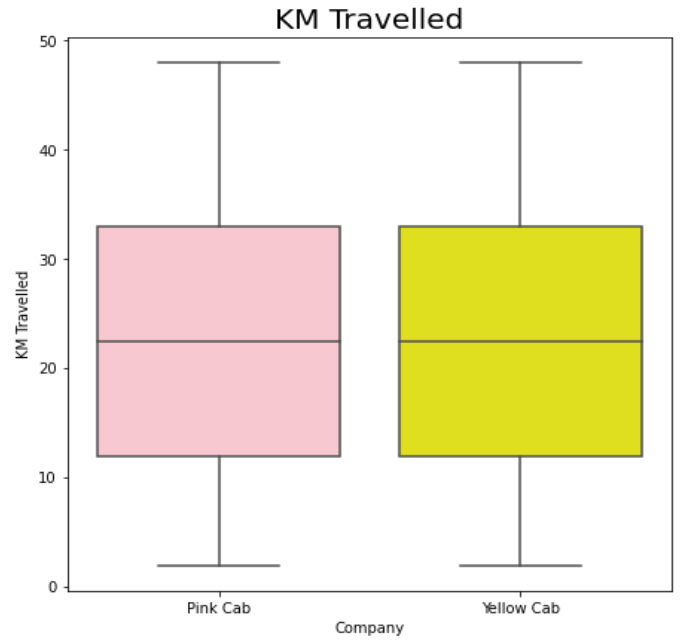


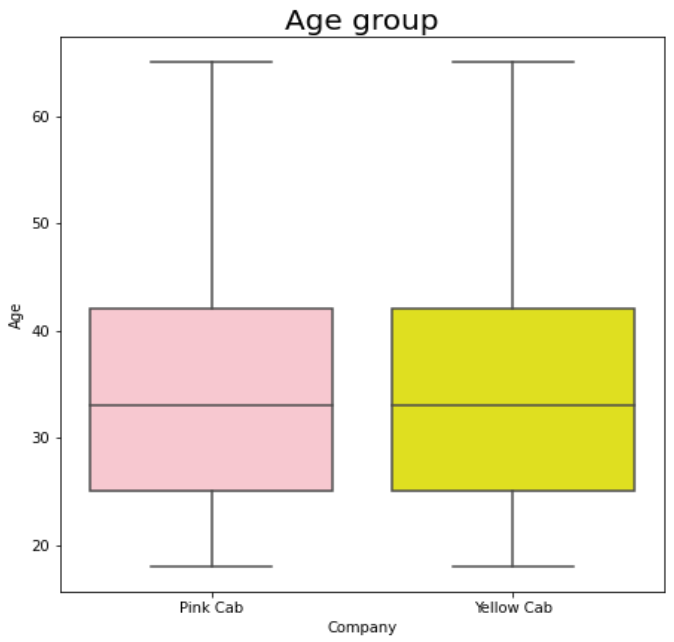


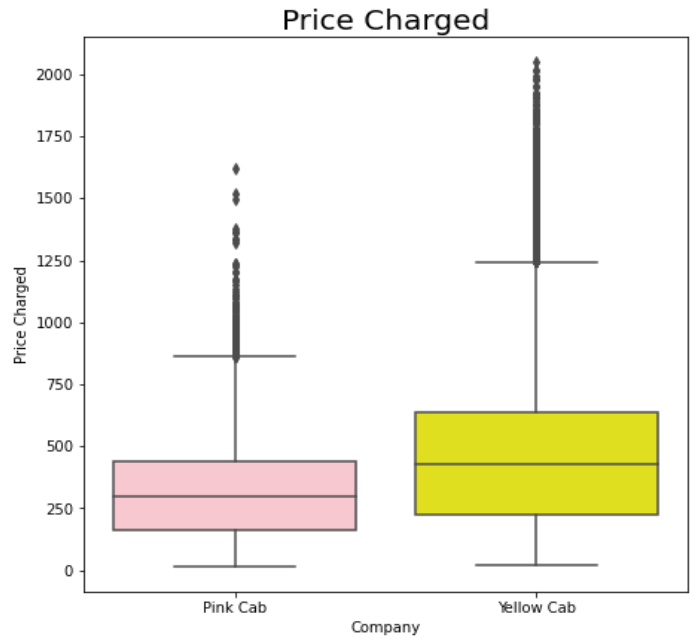
Box Plots

- Most of the users traveled in the range of 2 to 48 km for both cabs.
- Also, we can say that most of the people who travel in a cab are from the age group 20 to 45 years.
- We can say that Yellow Cab charged more money than Pink Cab.





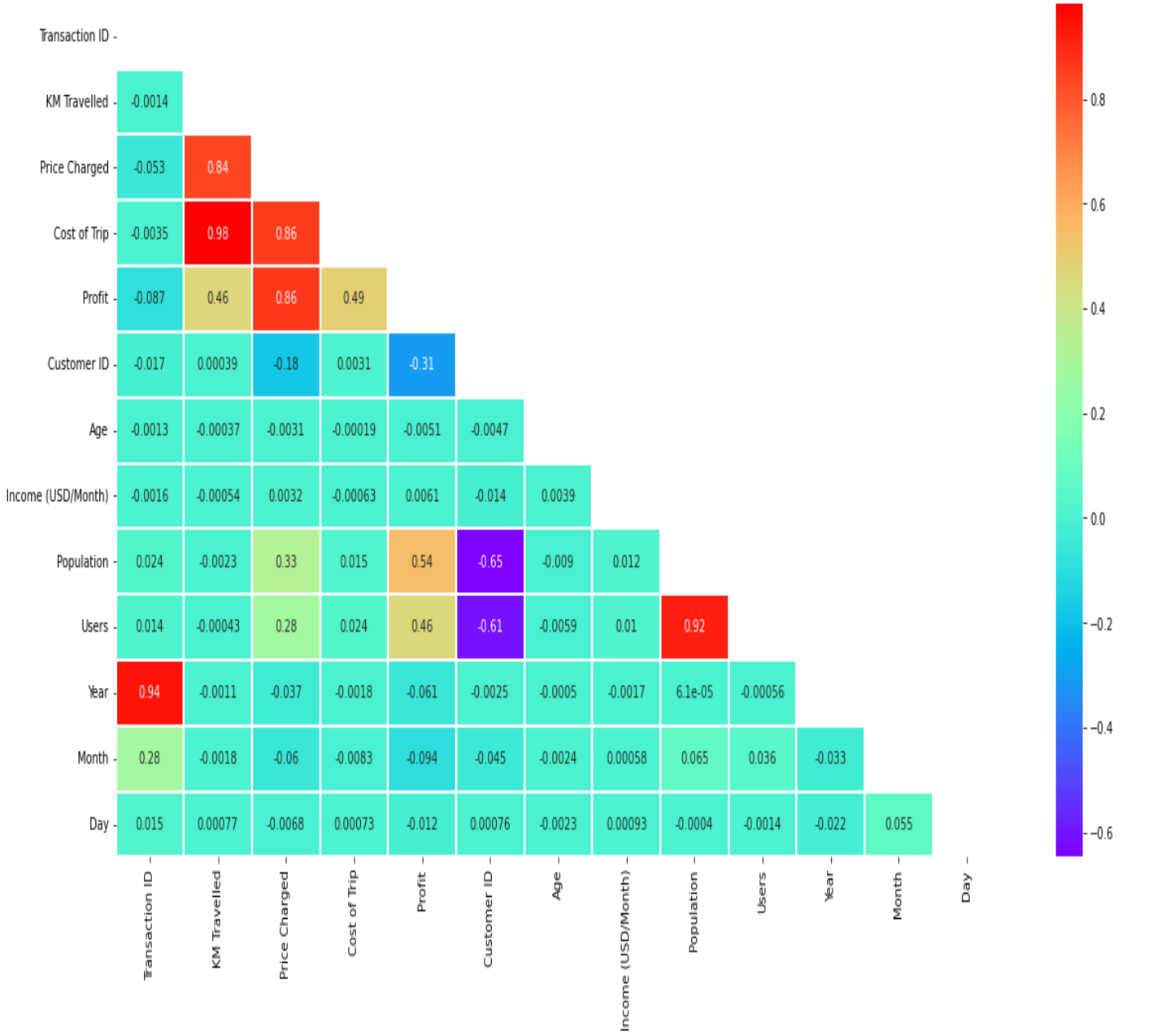




Correlation Matric

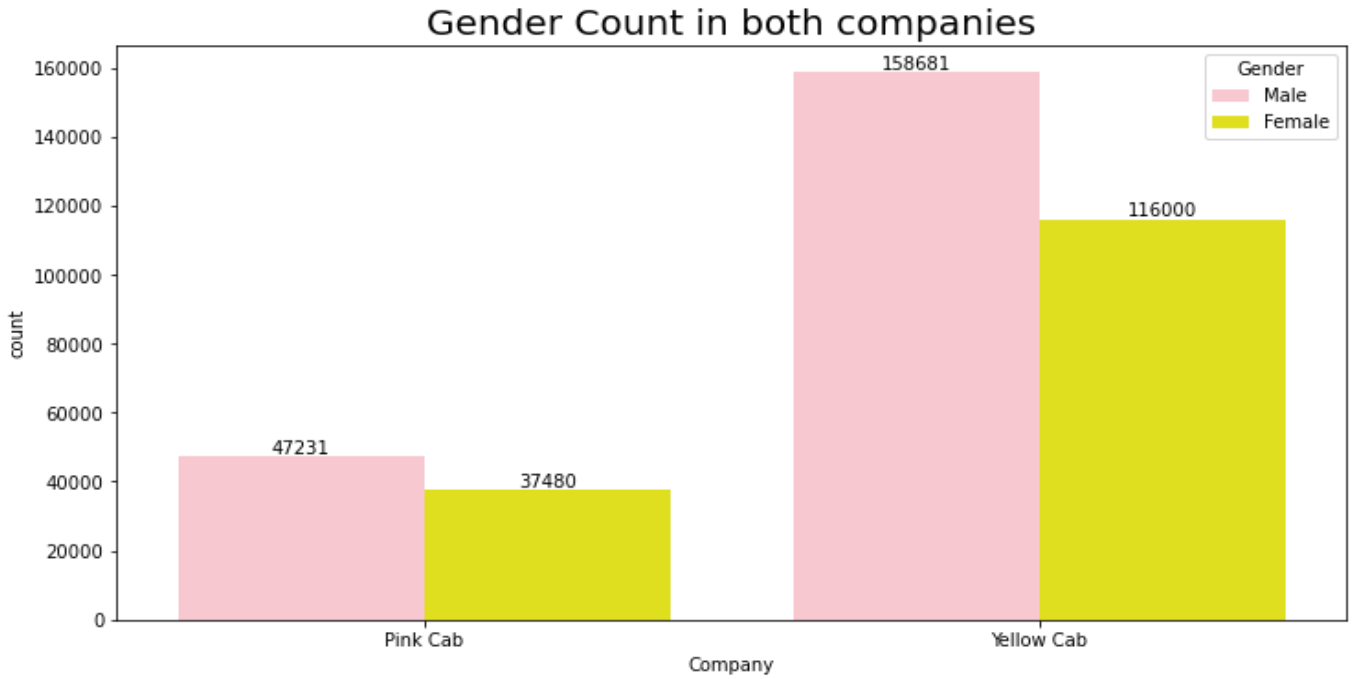
From the above correlation matrix we can say that

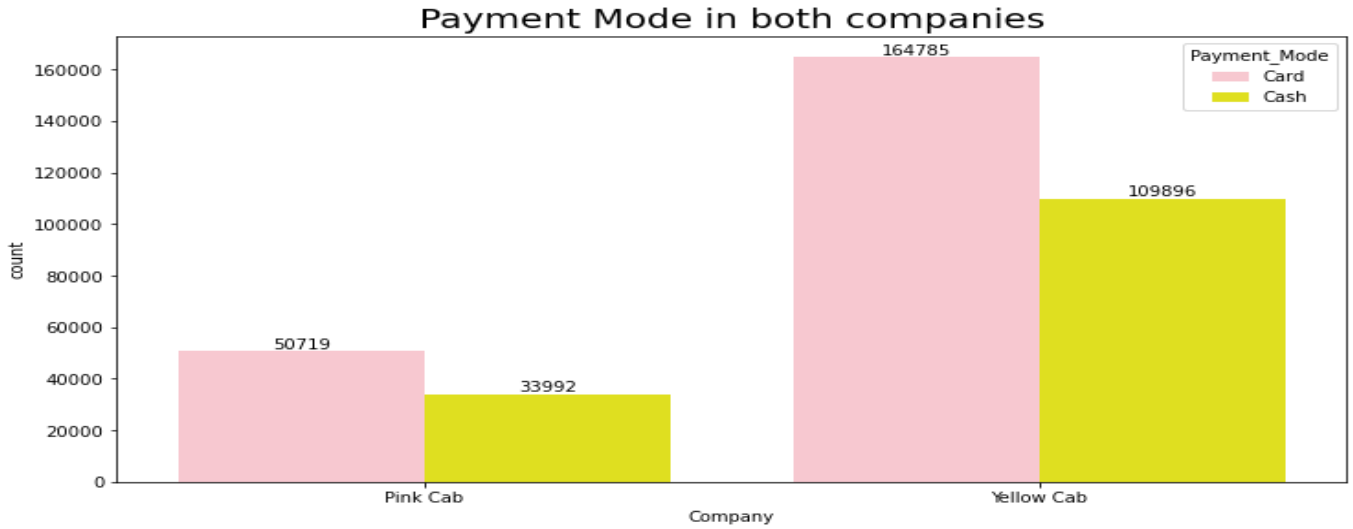
- KM Travelled has a high correlation with the cost of trip and the Price Charged
- Population and Users are also highly correlated



Box Plots

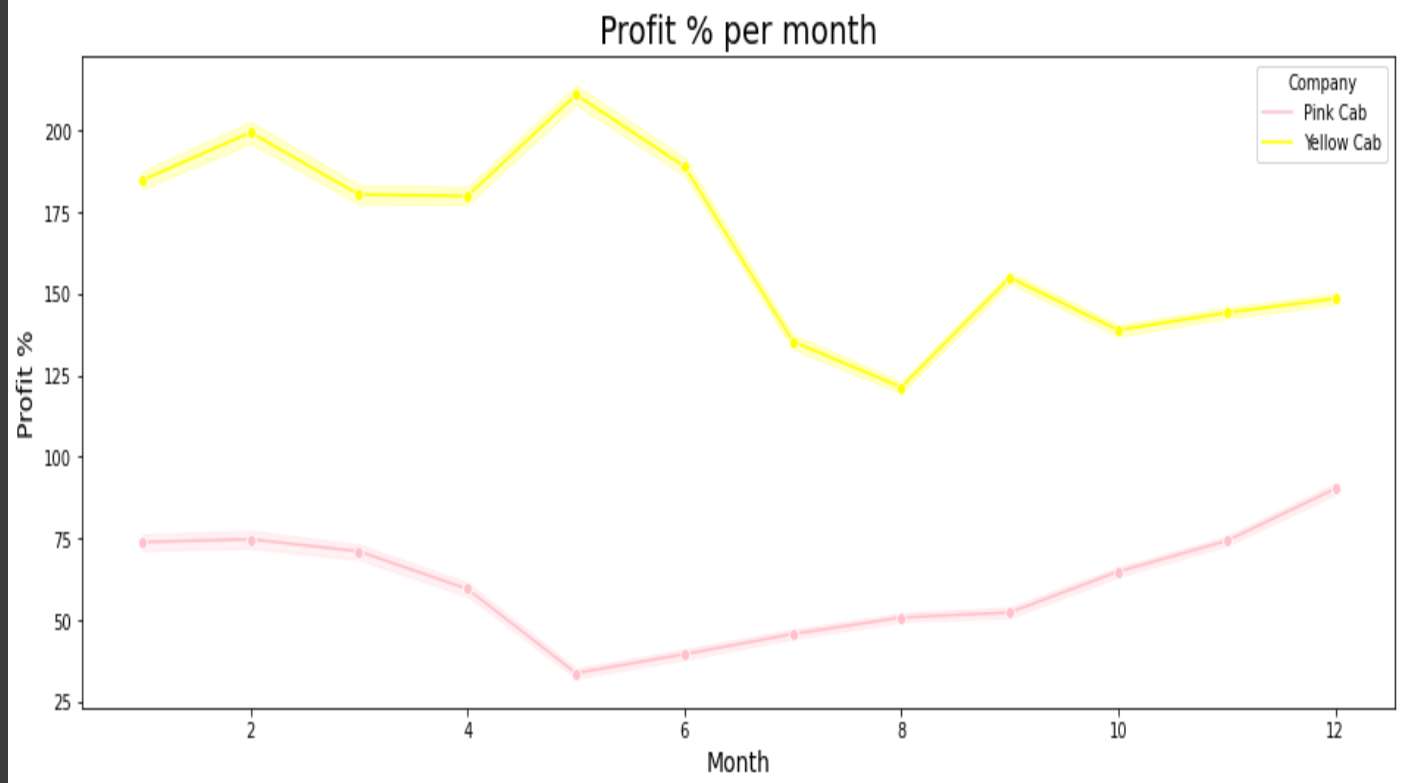
- From the bar plot, we can say that Male users prefer to travel in a cab as compared to Females.
- In both of the cab companies, Users have made payments mostly by card as compared to cash.

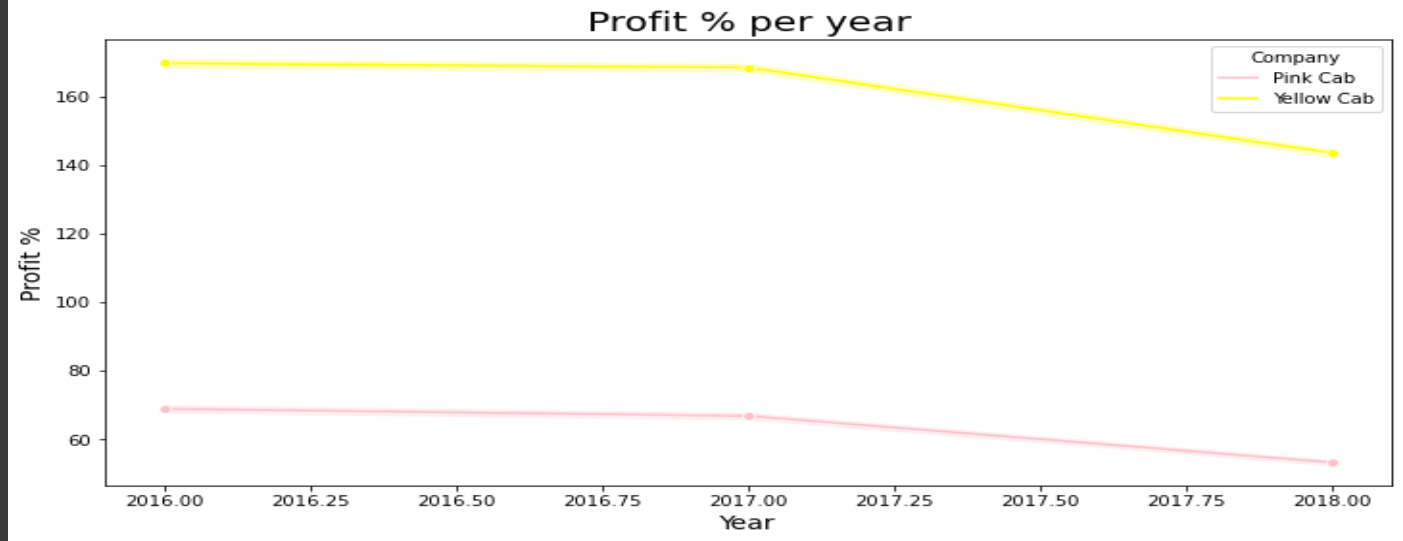




Box Plots

- From the plot, we can say that in the month of May, Yellow Cab has the highest profit margin whereas pink cab has the lowest profit margin.
- Also, In 2018 we can say that both companies decrease their profit margin.





Hypothesis Testing

HYPOTHESIS 1

Is there any difference in Profit regarding Gender of customers in both the cab companies?

H0 (Null hypothesis):

There is no difference in profit regarding gender in both the companies

H1 (Alternate Hypothesis):

There is difference in profit regarding gender in both the companies

Yellow Cab

P value is 6.060473042494144e-25

There is difference in profit regarding gender in Yellow company, therefore alternate hypothesis is selected.

Pink Cab

P value is 0.1151530590042579 8

There is no difference in profit regarding gender in Pink company, therefore null hypothesis is selected.

Conclusion

From the above analysis, we can say that there is no difference in profit regarding gender in both companies.

HYPOTHESIS 2

Is there any difference in Profit regarding age of users in both the cab companies?

H0 (Null hypothesis):

There is no difference in profit regarding age in both the companies

H1 (Alternate Hypothesis):

There is difference in profit regarding age in both the companies

Yellow Cab

P value is: 7.618115793609196e-05

There is difference in profit regarding age in Yellow company, therefore alternate (H1) hypothesis is selected.

Pink Cab

P value is: 0.502996690620347 1

There is no difference in profit regarding age in Pink company, therefore null hypothesis is selected.

Conclusion

From the above analysis, we can say that Yellow Cabs gives a discount to people older than 60.

HYPOTHESIS 3

Is there any difference in Profit regarding mode of payment in both the cab companies?

H0 (Null hypothesis):

There is no difference in profit regarding mode of payment in both the companies

H1 (Alternate Hypothesis):

There is difference in profit regarding mode of payment in both the companies

Yellow Cab

P value is: 0.2933060638298729

There is no difference in profit regarding mode of payment in Yellow company, therefore null hypothesis is selected.

Pink Cab

P value is: 0.7900465828793288

There is no difference in profit regarding mode of payment in Pink company, therefore null hypothesis is selected.

Conclusion

From the above analysis, we can say that there is no difference in profit regarding the mode of payment in both companies.

EDA Summary

After performing EDA on the datasets we can conclude the following points:

- Most Users prefer traveling with a Yellow cab to a Pink cab.
- New York has the highest number of cab users.
- Yellow cab owns 89% of the total profit made by both companies.
- Yellow cab charges higher than Pink cab.
- Male users prefer to travel in a cab as compared to Females.
- Most of the users traveled in the range of 2 to 48 km for both cabs.

Recommendations

After analyzing the dataset I advise XYZ company to invest in Yellow Cab company.

Thank You