Exploratory Data Analysis

G2M Case study

16-Feb-2022

Problem statement

XYZ is a private firm in the US planning for an investment in Cab industry.

The **objective** was to generate actionable insights to identify the right Cab company for XYZ to make investments.

The data described fares and customers' data of two different cab companies.

The time period of the data was from 31/01/2016 to 31/12/2018.

The data consisted of 4 different datasets:

- **Cab_Data.csv** – Details of transaction for 2 cab companies.
- **City.csv** – Contains list of US cities, their population and number of cab users.
- **Customer_ID.csv** – Mapping table that contains a unique identifier which links the customer's demographic details.
- **Transaction_ID.csv** – Mapping table that contains transaction to customer mapping and payment mode.

Approach

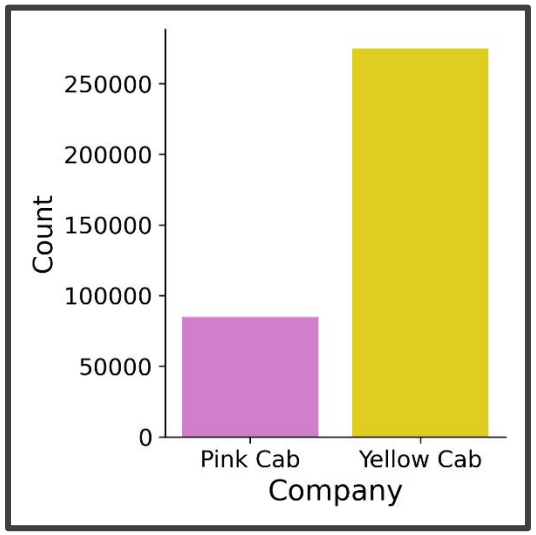
- The Cab_Data.csv dataset was joined with the Transaction_ID.csv dataset, and then with the Customer_ID.csv dataset, both with **full joins**. Finally, the resulting data frame was joined to the City.csv dataset using a **left join** to exclude the data of the cities that had no observations (San Francisco, CA).
- The resulting Master dataset showed null values. They represented almost the 20%, so imputing techniques were not recommended. The null values did not seem to have a relationship with other variables of the dataset, so it was assumed a Missing Completely At Random (MCAR) mechanism. Moreover, without the observations with null values, the number of rows was considered still high (350,000). In consequence, **the observations with null values were dropped from the dataset**.
- The Master dataset did not show duplicates.
- The Master dataset presented outliers only on the Price Charged variable.
- The analysis was carried on examining the categorical variables first, then the quantitative variables, and finally, the relationships between the quantitative variables. For the most part, **comparisons between the two cab companies were made** to figure out the best one to make investments.

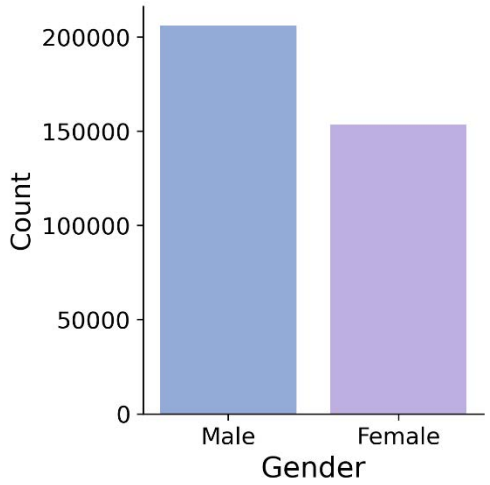
EDA

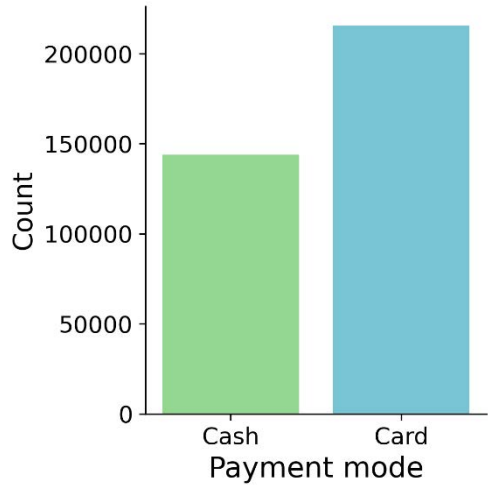
Categorical variables

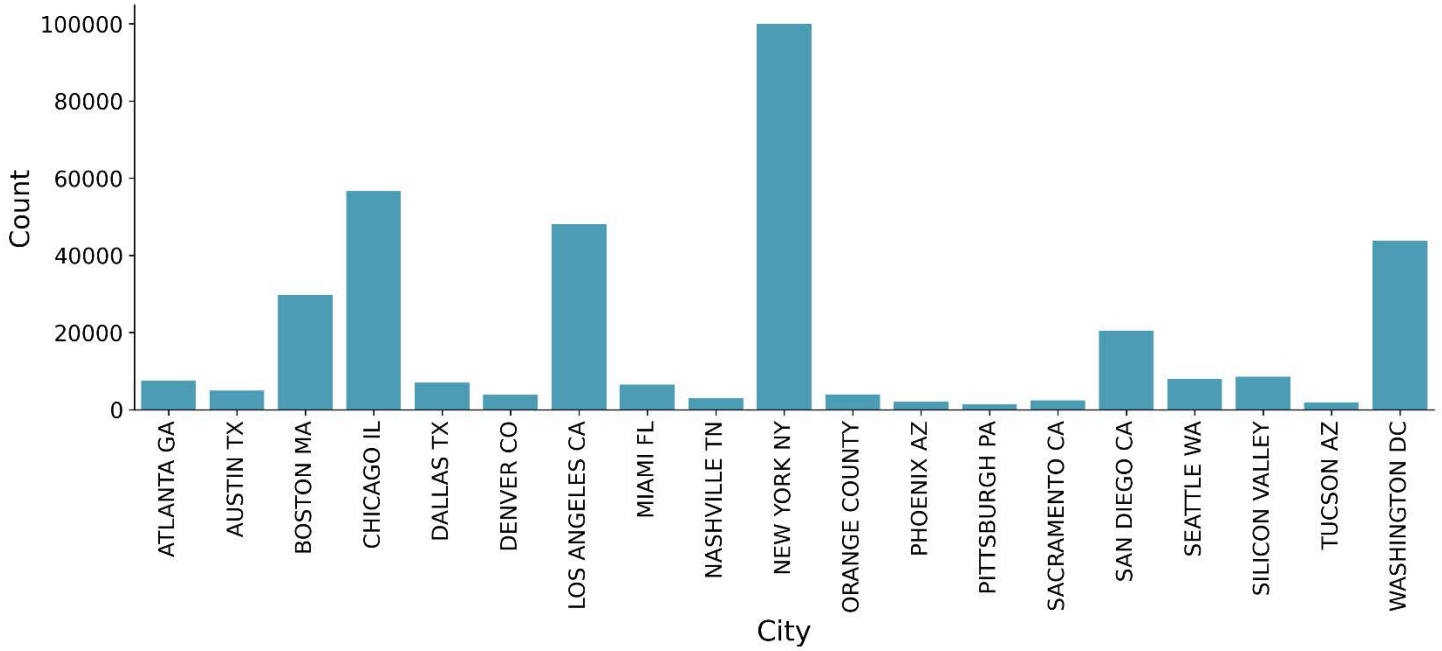
There was a notorious difference on the number of rides between both cab companies. There were a lot more rides for Yellow Cab.

There was a big difference in rides considering the different cities. New York was the city with the biggest number of rides.





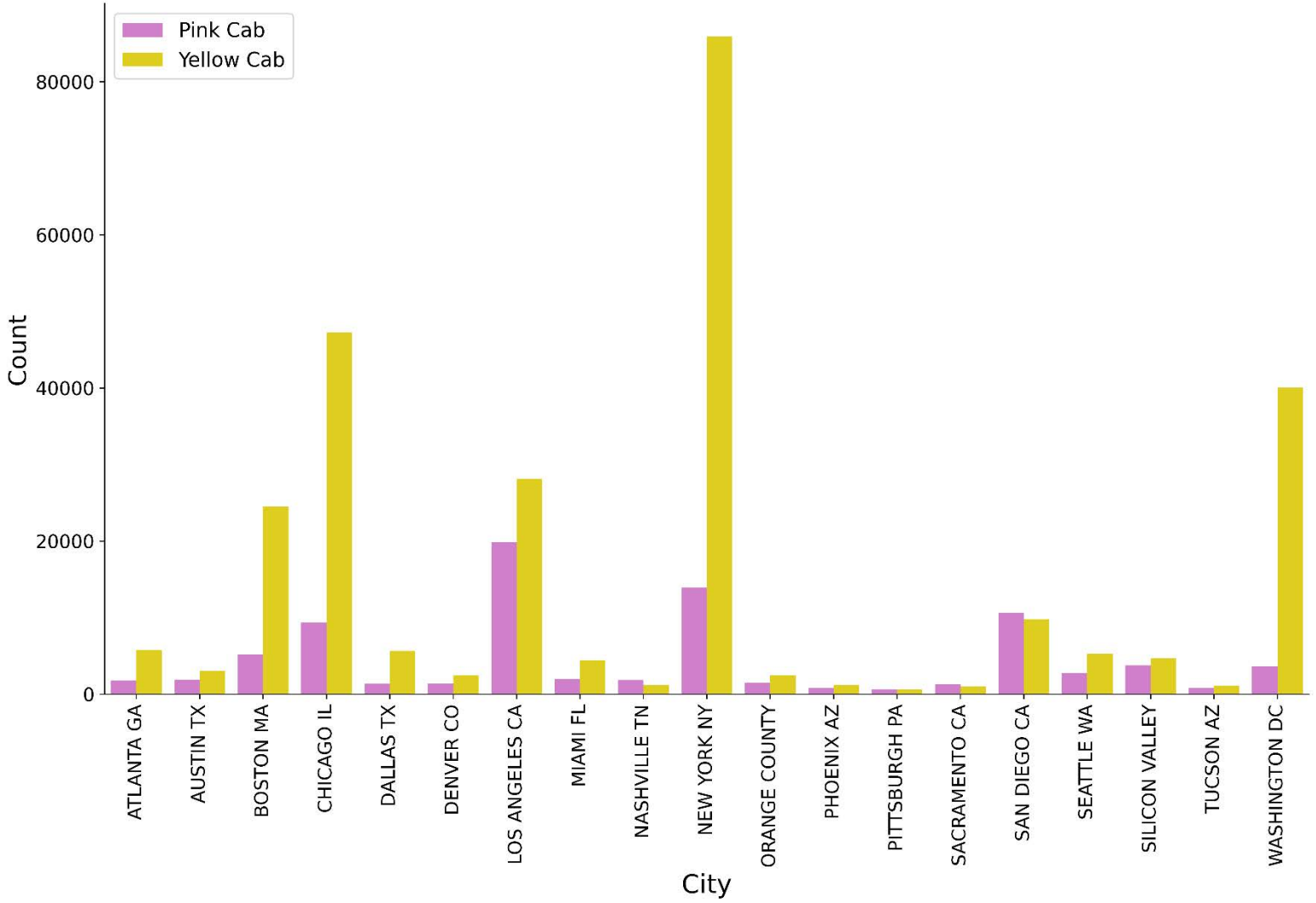




EDA

Trips per City

In examining the number of trips per city, and divided by company, it is notorious that, in general, Yellow Cab had a bigger number of rides. This was observed especially in New York, Chicago, Washington, and Boston.

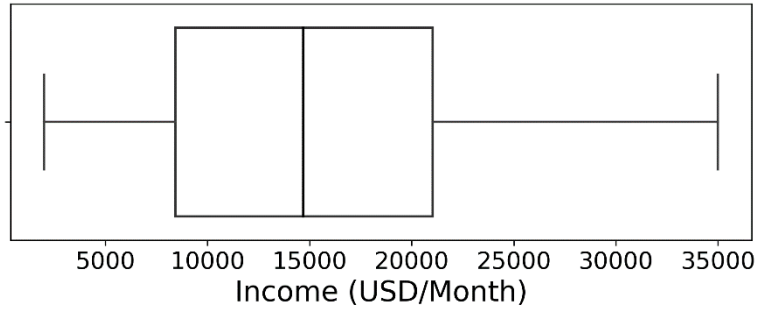


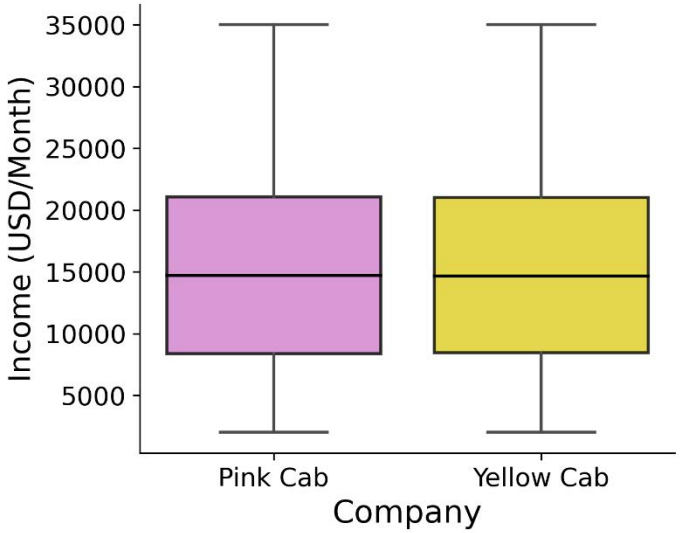
EDA

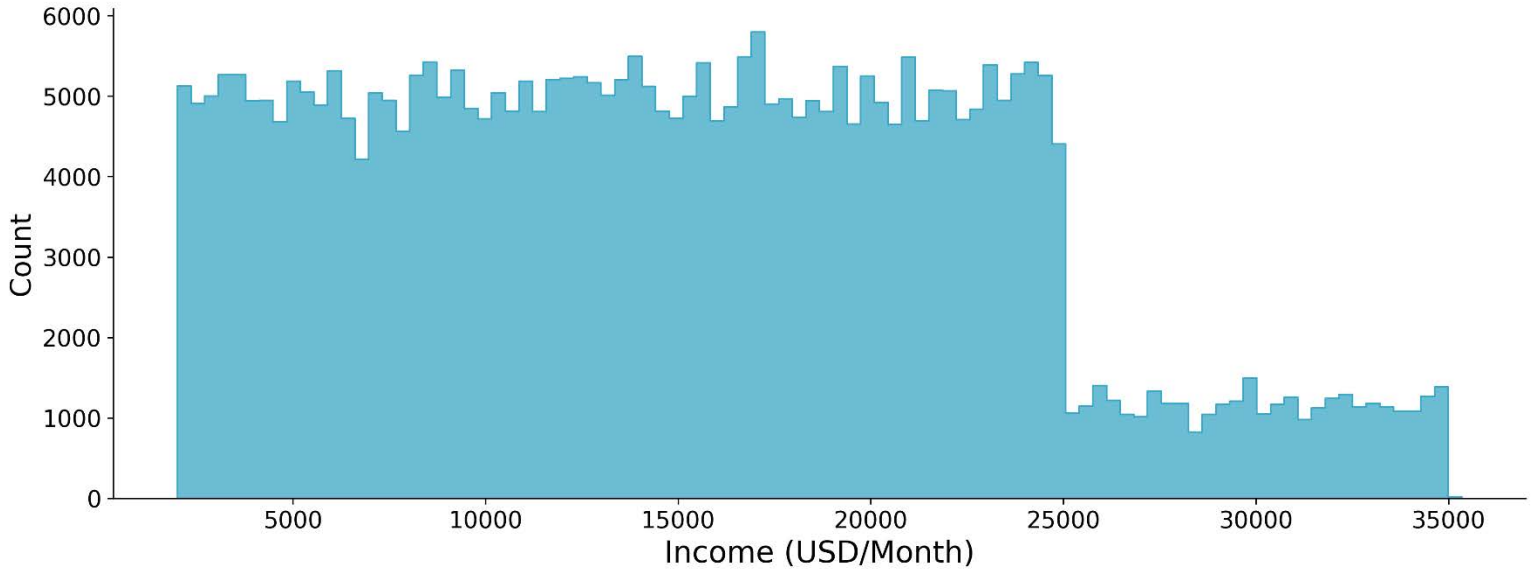
Income of customers

Income of customers seemed to have a right skewed distribution. Also, It appears there were two different groups of people considering their monthly income (<=$25,000 & >$25,000).

When dividing the data by Company there seemed to be no differences between the amount of money earned monthly by the customers of each cab company. In fact, their means didn't show statistically significant differences (p>0.05).





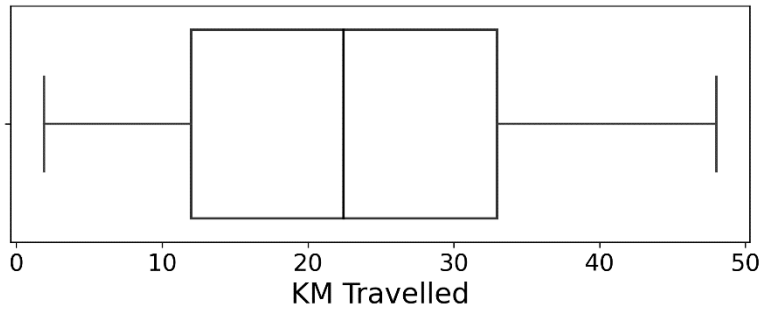


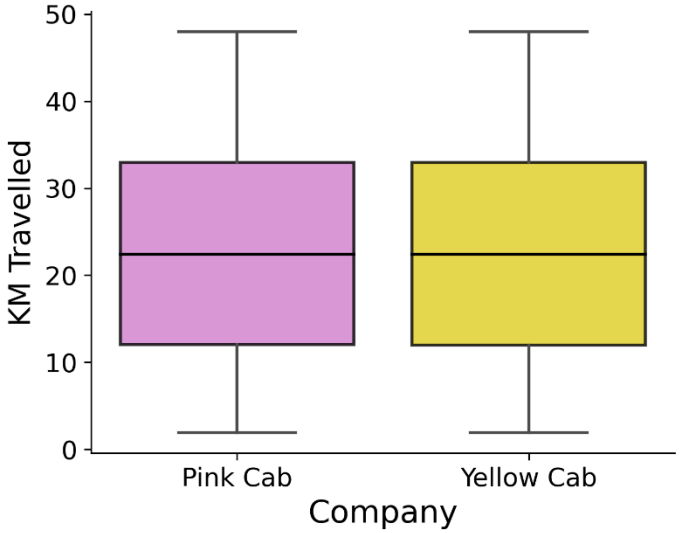
EDA

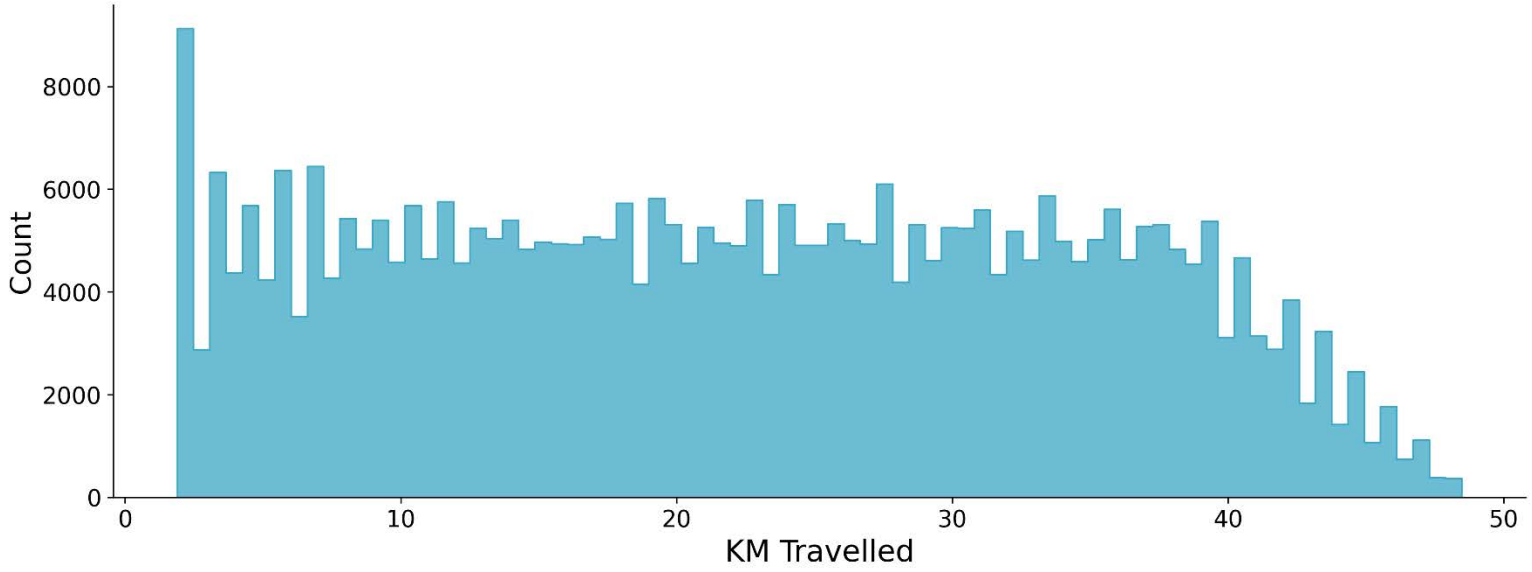
Kilometers traveled

The Km Traveled variable seemed to have a distribution with a slight skew to the right. The least common trips consisted of big distances.

Between the two companies, there seemed to be no big differences on distance traveled. The means were not different with statistical significance (p>0.05). The two cab companies seemed to have both short and long trips.





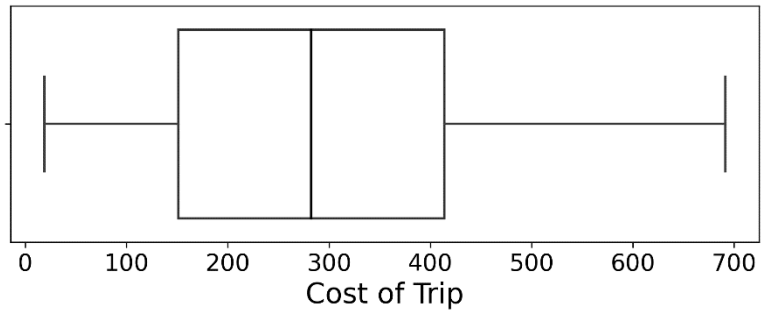


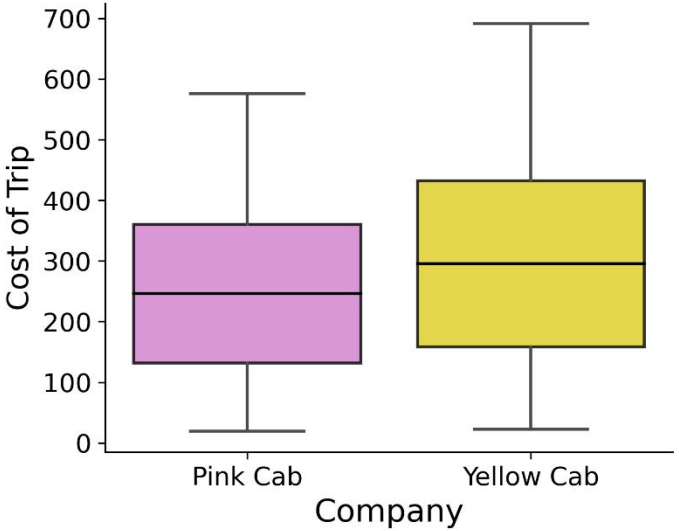
EDA

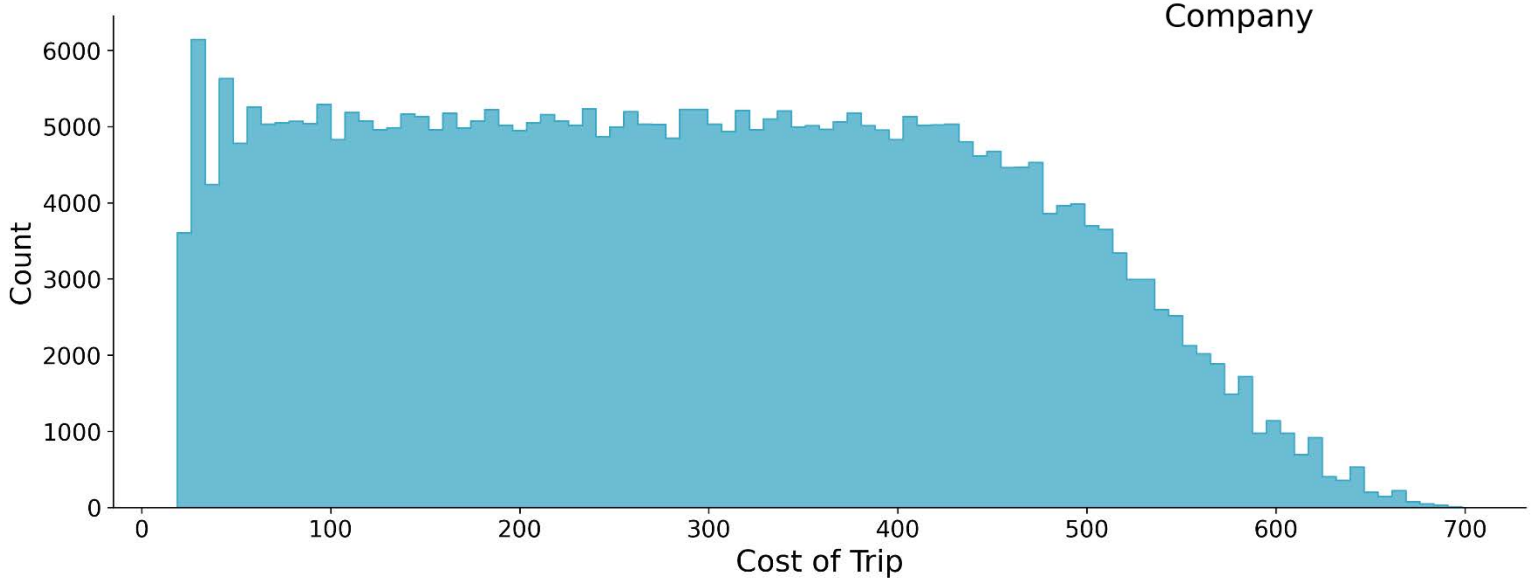
Cost of trip

The distribution of the Cost of Trip variable was right skewed, with no outliers.

When dividing by company, Yellow Cab showed higher values. Also, the differences in means were statistically significant ($p<0.05$).





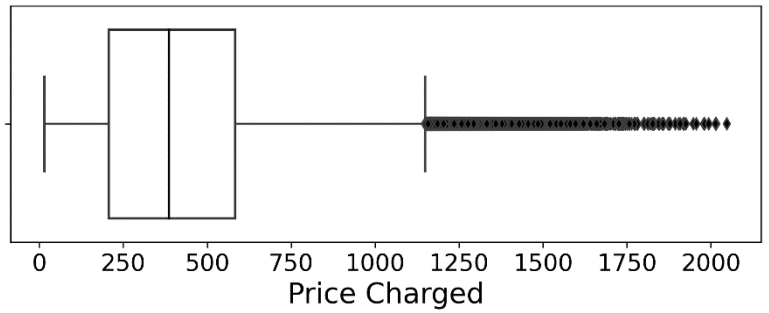


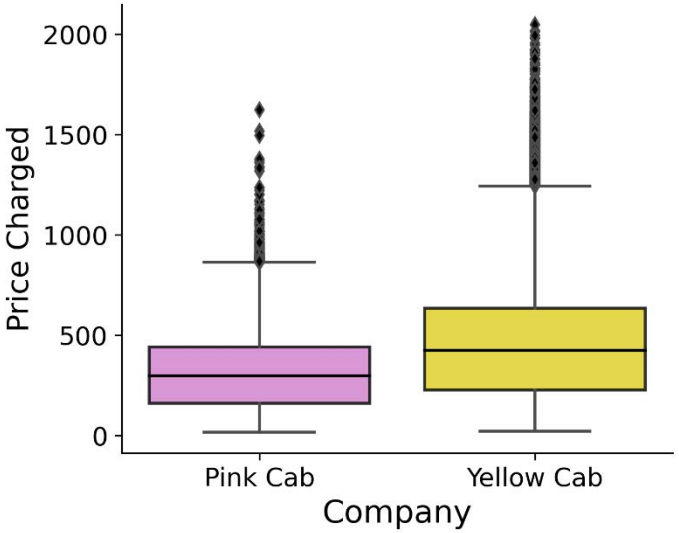
EDA

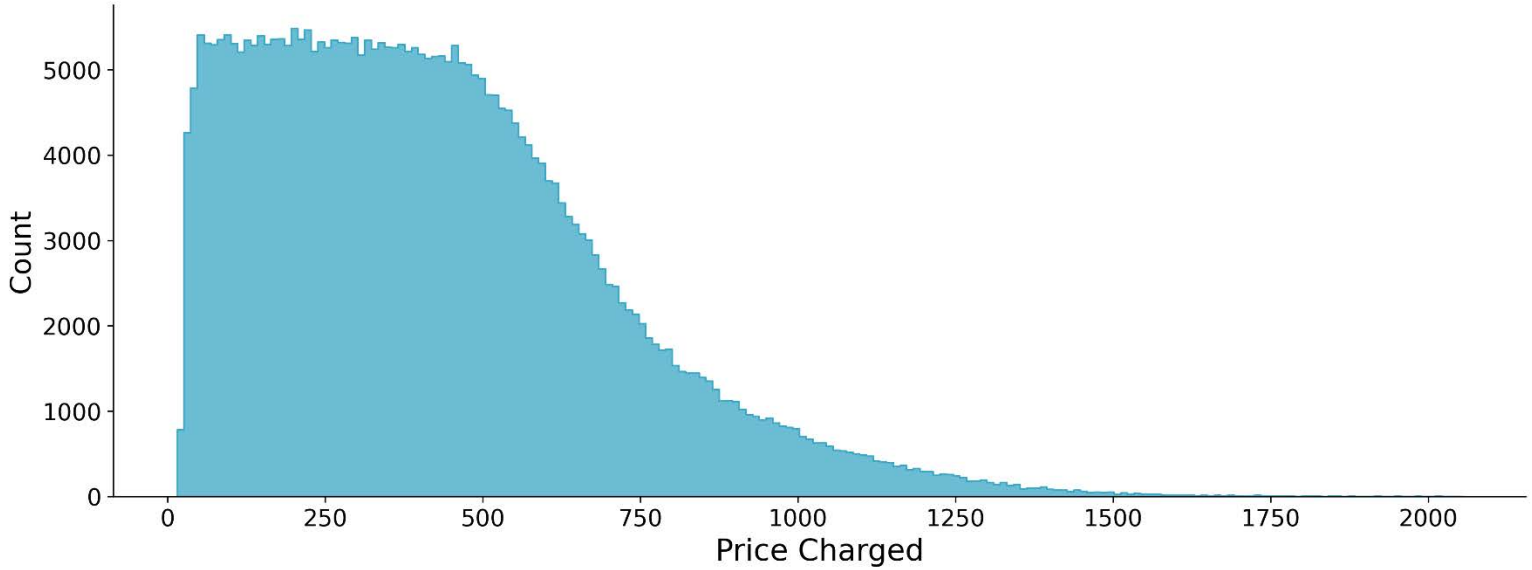
Price charged

The Price Charged variable showed a right skewed distribution with 5,958 outliers (1.66%).

When divided by company, the distributions were a little bit different. Yellow Cab seemed to be more expensive. The differences between the means on both companies were statistically significant ($p<0.05$).





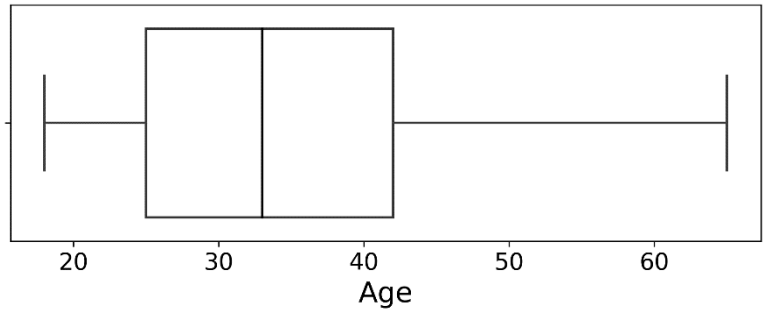


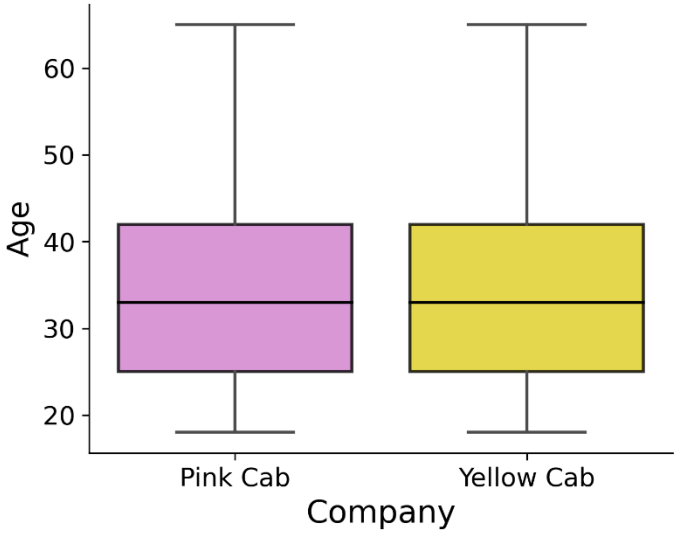
EDA

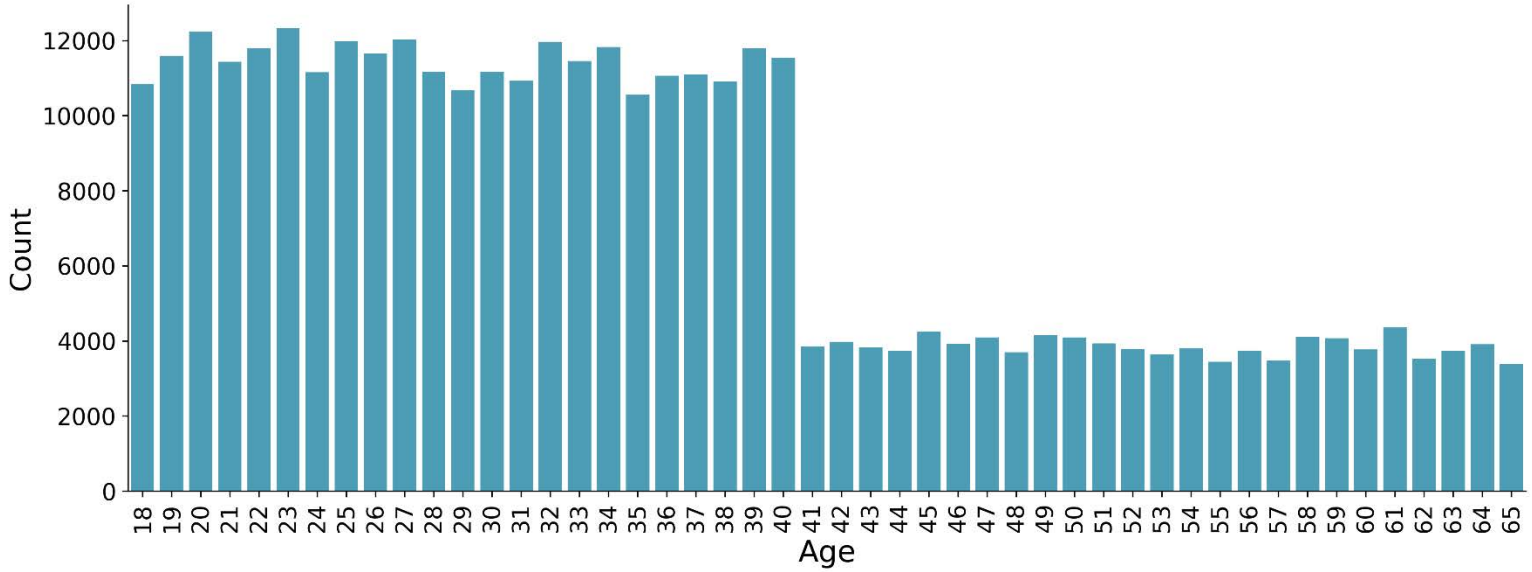
Age of customers

The distribution of Age was right skewed. There seemed to be two big groups divided by ages (18-40 & 41-65).

When divided by company, there seemed to be no big differences in the distributions. Also, the differences between the means were not statistically significant (p>0.05).





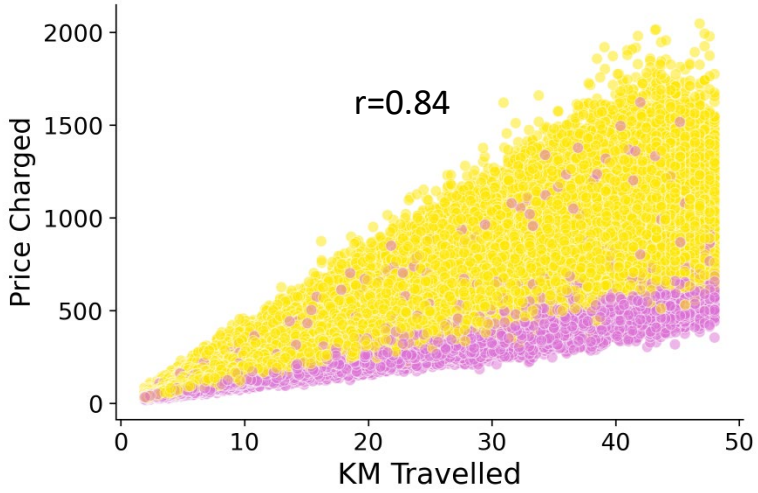


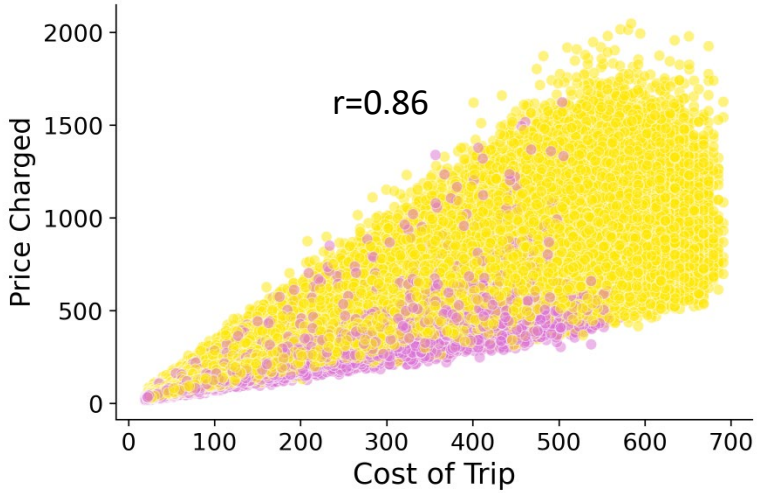
EDA

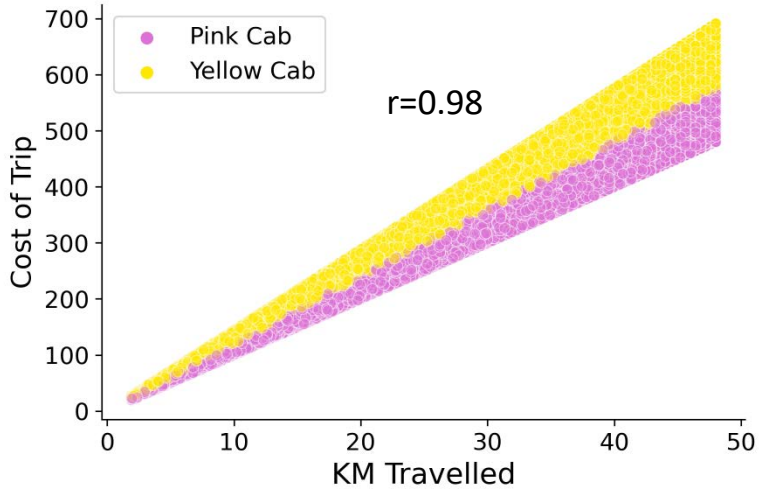
Distance traveled, cost, and price

On examining the relationship between Km Traveled, Price Charge, and Cost of Trip it was observed heteroscedasticity. The variability increased with the increments of the three variables.
As expected, the three variables showed high values of correlation, especially cost and distance traveled.

In general, Yellow cab seemed to be more expensive than Pink cab.





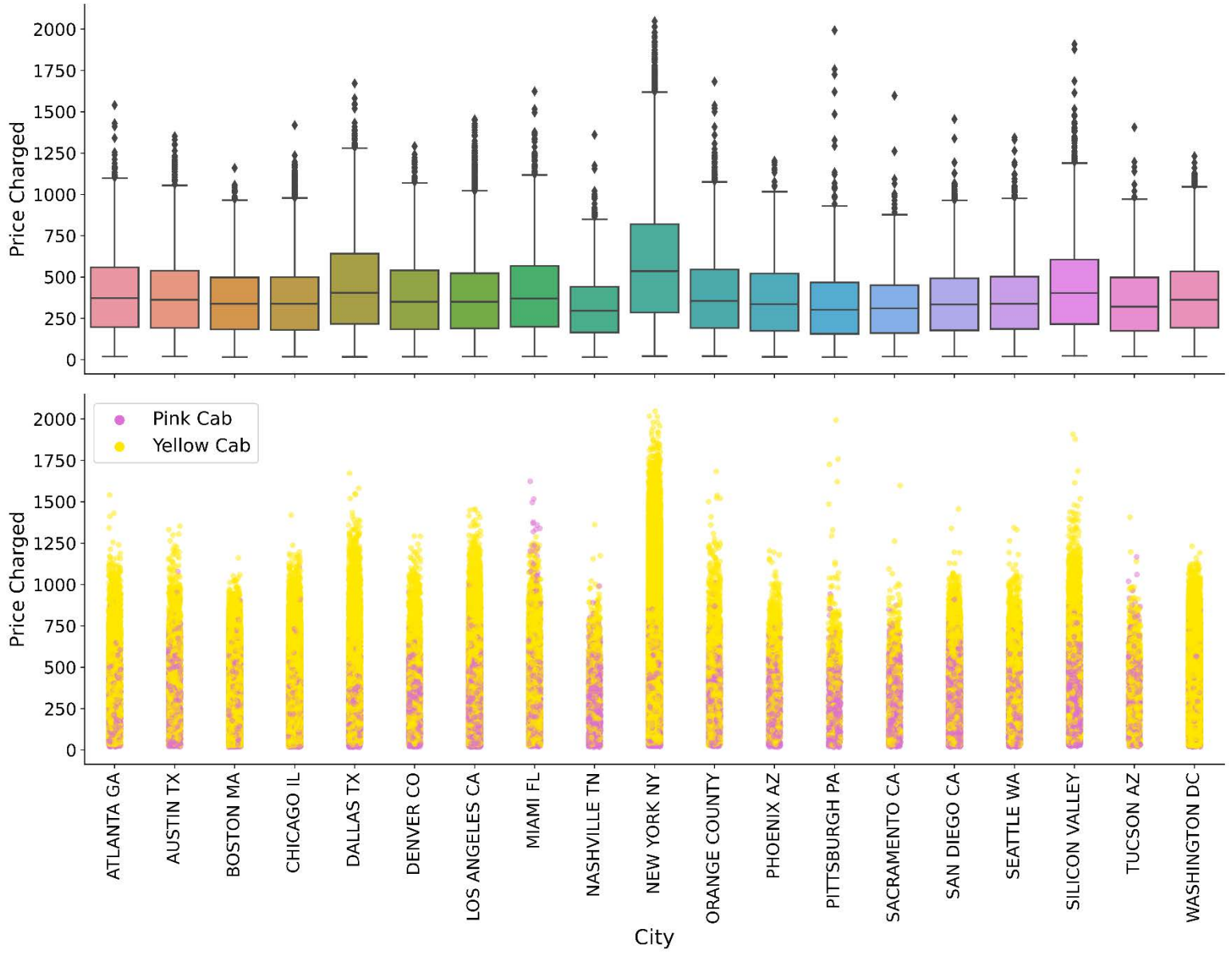


EDA

Price charged per city

There were differences for prices charged between cities. It seems the largest prices were charged in New York.

When dividing by company, it's notorious the largest prices were charged by Yellow Cab, with the possible exceptions of Miami and Tucson.

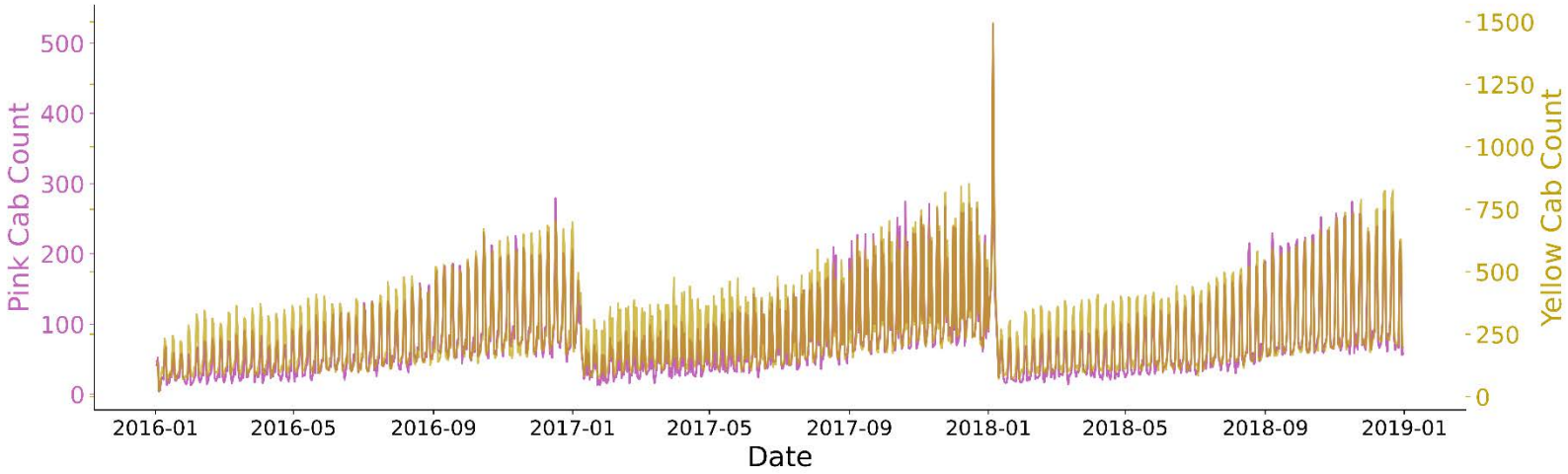


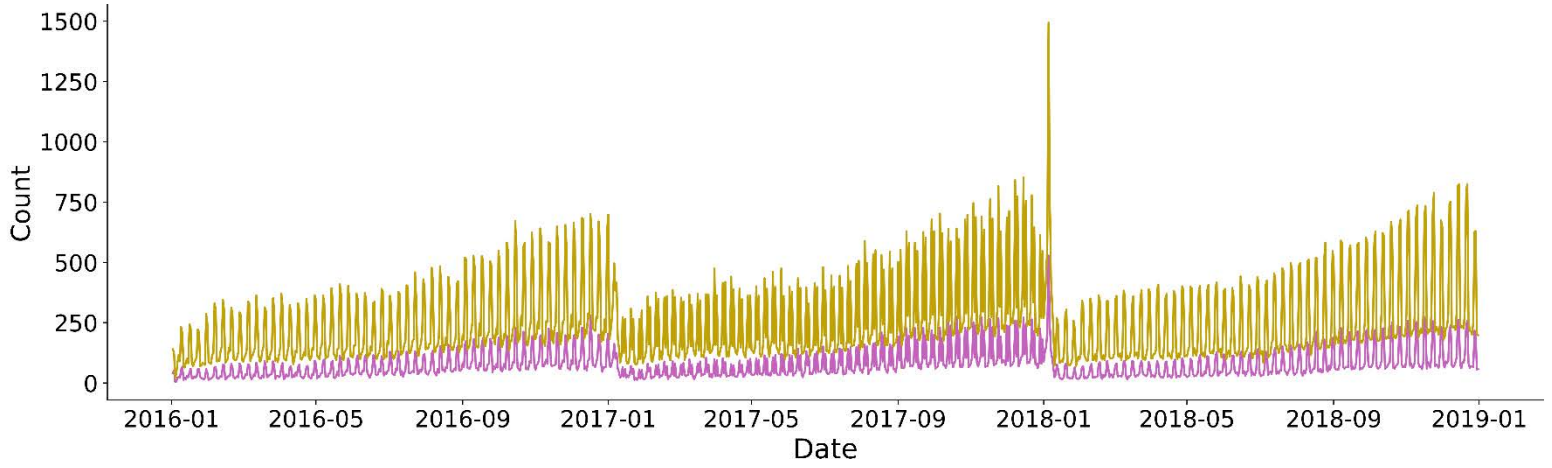
EDA

Trips over time

Besides the absolute number of rides for each date, there seemed to be a similar pattern for both companies. It seemed that there was an annual growing tendency for both. There appeared to be seasonality for both, beginning each year with few rides on the first months and more on the last ones.

Yellow Cab was superior on every date in terms of number of rides.





EDA summary & recommendations

- For both companies, the distributions of income of customers, kilometers traveled, and the age of customers were very similar, and did not show significant differences when comparing their respective means.
- Between companies there were differences on the distributions of the costs of trips and the prices charged, with statistically significant differences on means.
- The prices charged were also different between cities, being New York the one with the highest prices.
- The patterns of number of trips over time were very similar between companies, showing apparent seasonality.
- In a general manner, Yellow Cab was the company with the highest prices and even the highest number of trips.
- Pink Cab didn't stand out in any feature. The preference of one over the other seems to be unrelated to customer income, age nor length of trips.

The recommendation is to **invest on Yellow Cab** as it was superior in prices charged and number of trips. Specifically, it's recommended to invest in Yellow Cab's projects related with New York.

Thank You