Exploratory Data Analysis

<Cab Exploratory Data Analysis>

<October 21st 2022>

Agenda

Executive Summary

Problem Statement

Approach

Data Exploration

EDA

- Trip Count
- Profit Analysis
- Loss Analysis

EDA Summary

Recommendations

Executive Summary

- This project's goal is to perform exploratory data analysis to decide which Cab company, Pink Cab or Yellow Cab, is a better investment.

Problem Statement

- XYZ is a private firm in US. Due to remarkable growth in the Cab Industry in the last few years and multiple key players in the market, it is planning for an investment in the Cab Industry .
- There are two companies: Pink Cab and Yellow Cab. With multiple data sets provided, each file represents different aspects of the customer profile, XYZ is interested in using insights to help them identify the right company to invest.

Approach

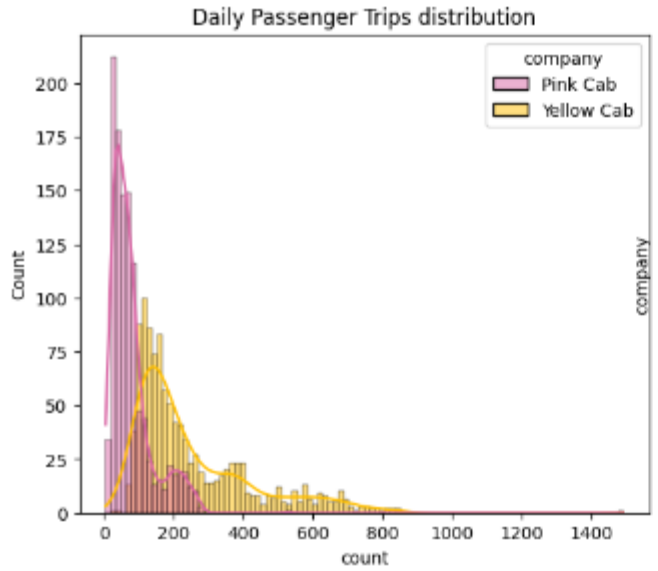
- With the four datasets provided
  - Cab_Data.csv (details of transactions)
  - Customer_ID.csv (customer's demography)
  - Transaction_ID.csv (transaction, payment)
  - City.csv (US cities, population, cab users)
- After investigating and formatting the data, we will be:
  - Joining all four datasets to create a master dataset while filtering null values and duplicates.
  - Performing exploratory data analysis to discover trends and useful insights within the data.

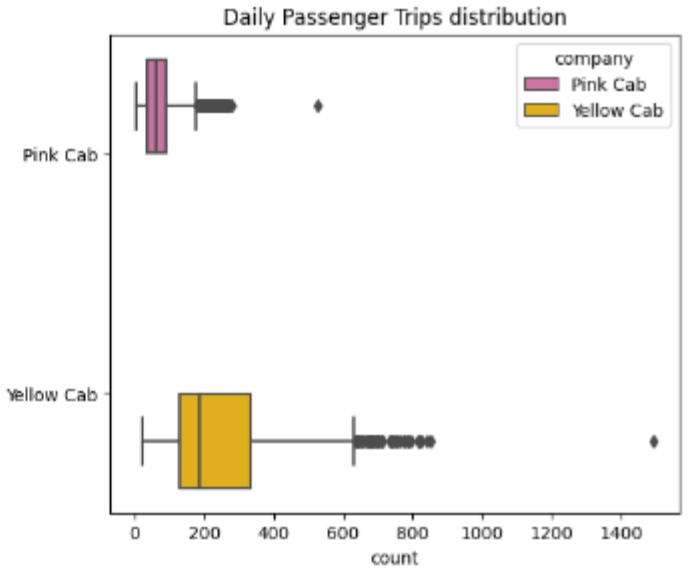
Data Exploration

- Master Set Information:
  - 19 Features (Including 4 derived features)
  - Timeframe of the data: **31/01/2016** to **31/12/2018**
  - Shape of dataset is (359392, 14)

- Assumptions:
  - There are outliers present in Price_Charged feature but due to unavailability of trip duration details, we are not treating this as an outlier
  - Profit of rides are calculated by keeping other factors constant and only Price_Charged and Cost of Trip features used to calculate profit.
  - Expenses for the trip only involves fuel charges.
  - There isn't sufficient data on the internet for the base fares per year for each city in the US. This will be left out of the analysis
  - Users feature of city dataset is treated as number of cab users in the city.

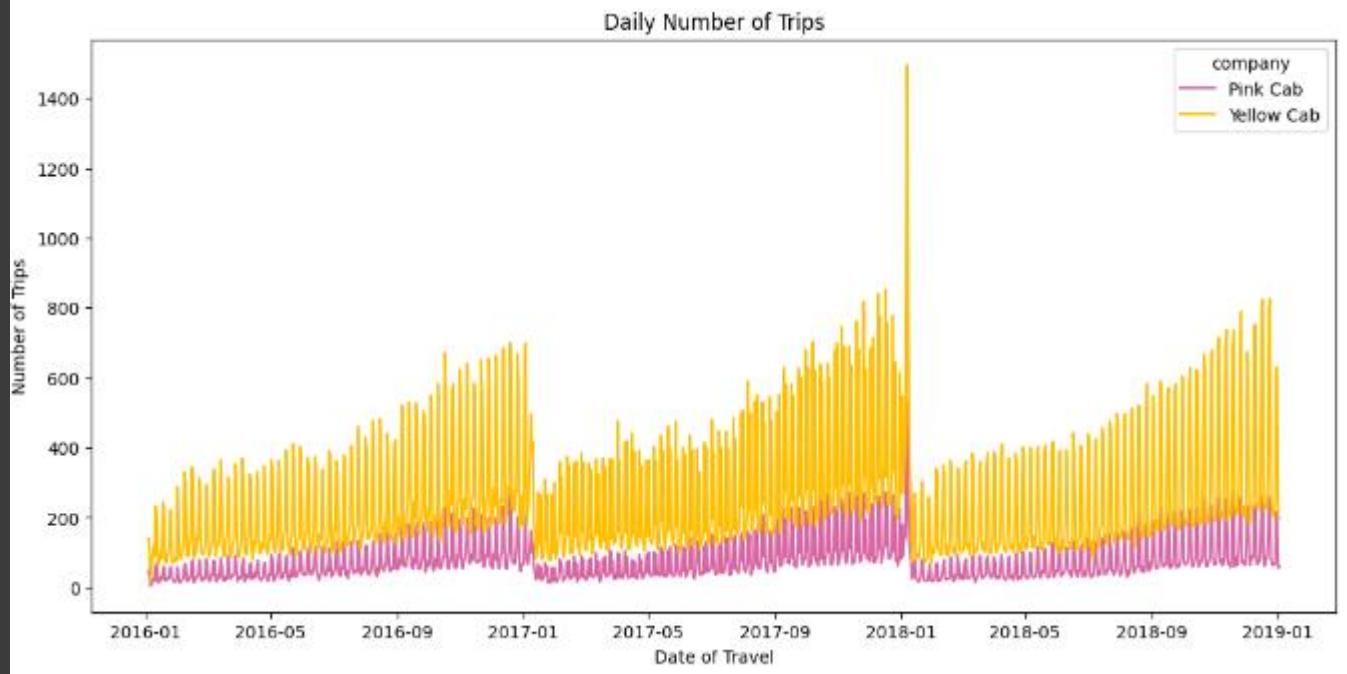
Exploratory Data Analysis (Trip Count)





The above plots depicts the distribution of daily trips by both Cab companies. **Yellow Cab** has higher median trips compared to **Pink Cab**. Both distributions rightly skewed, signifying that it is rarer to have a greater number of trips on some days.

Trip Count Cont.



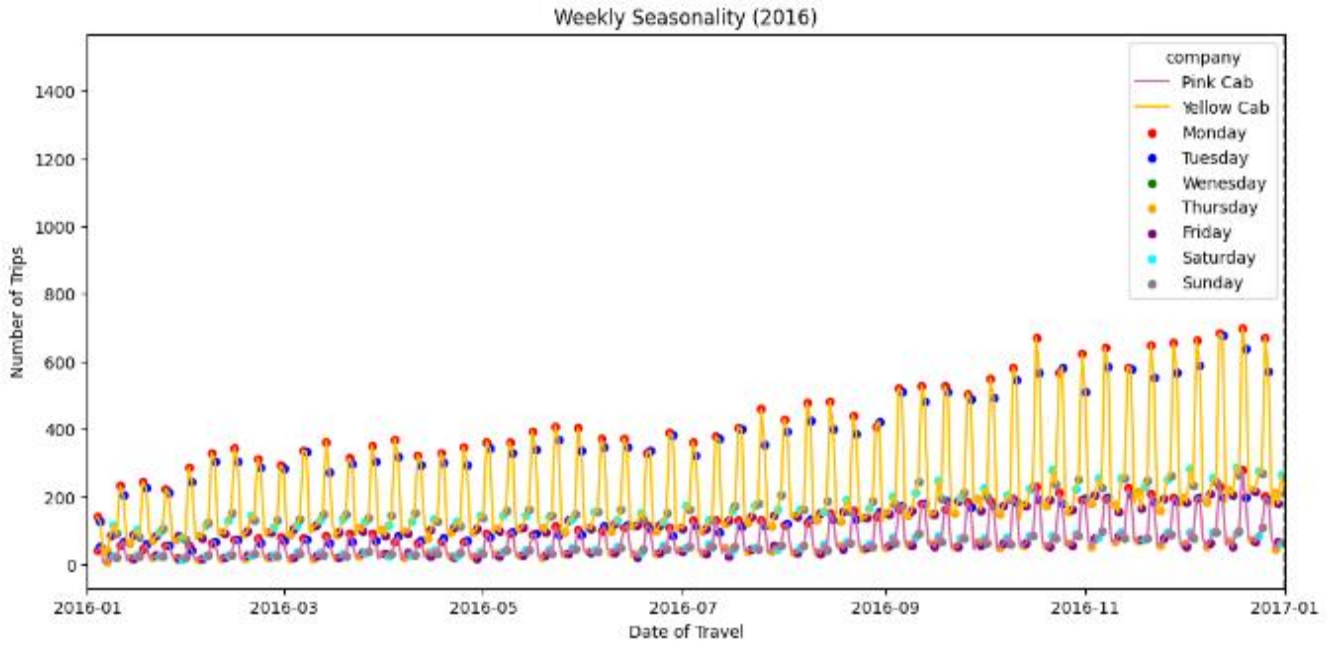
The above plot displays the daily trips made by both Cab companies from the beginning of 2016 till the end of 2018. There is a clear seasonality on a weekly, monthly, as well as yearly count for both Cab companies. Both Cab companies follows generally the same patterns.

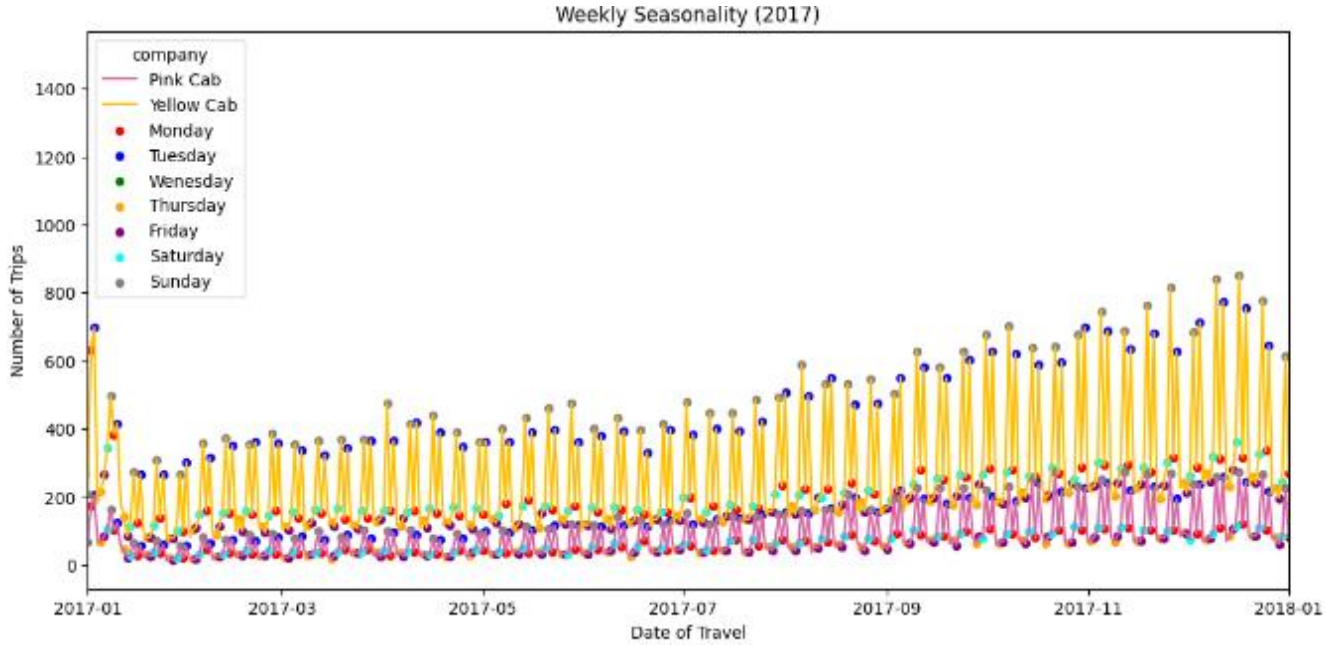
On a monthly level, there is a clear upward trend. It seems that the lowest number of trips occur at the start of the year, and increases gradually as the months go by. On a new year, the number of trips drop back down to the lowest count during that year.

On a yearly level, the trend seems to be almost uniform.

Yellow Cab makes significantly more trips on any given day compared to Pink Cab. The highest reported trips for both Cab companies was on January 7th, 2018.

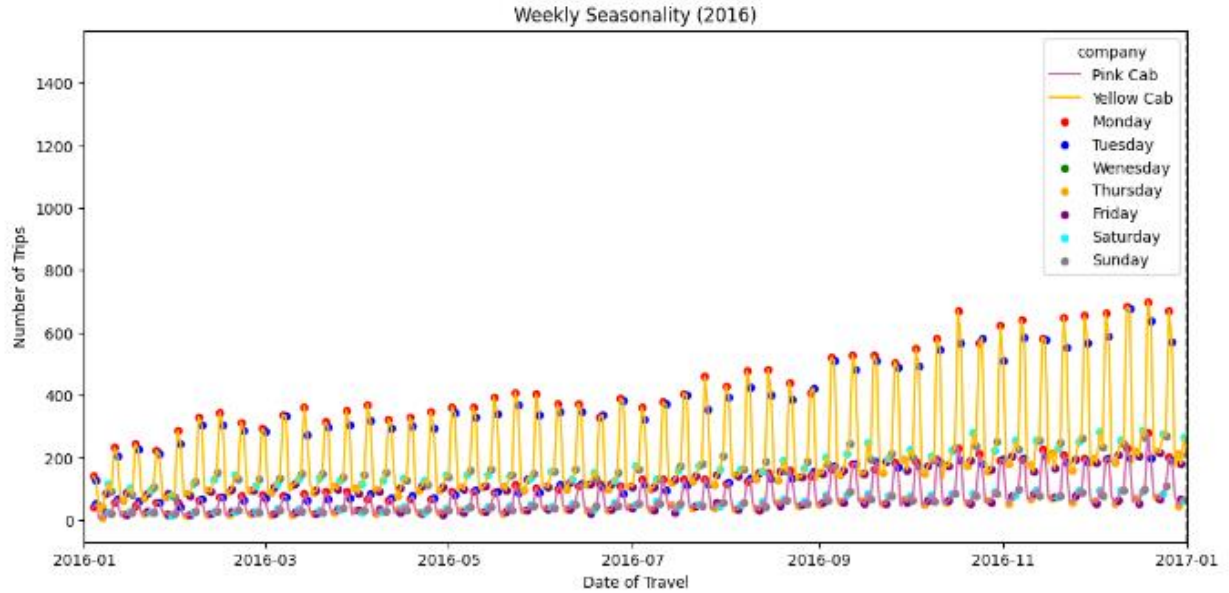
Trip Count Cont.



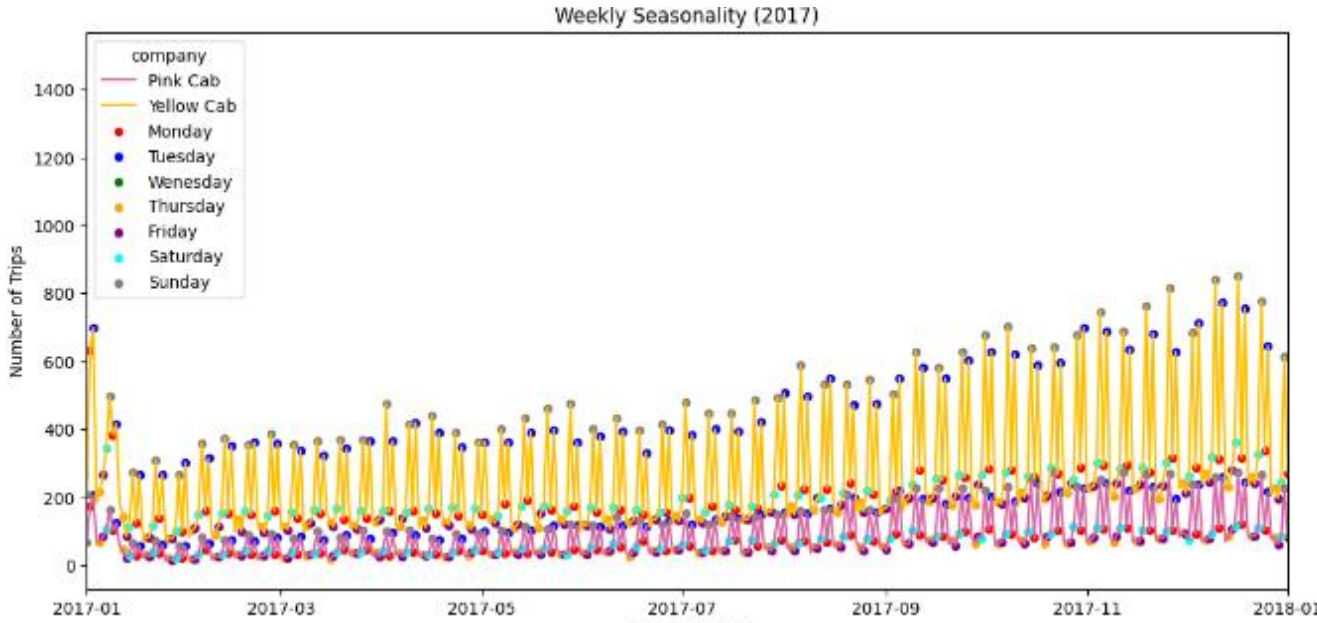


Data Glacier
Your Deep Learning Partner

Trip Count Cont.



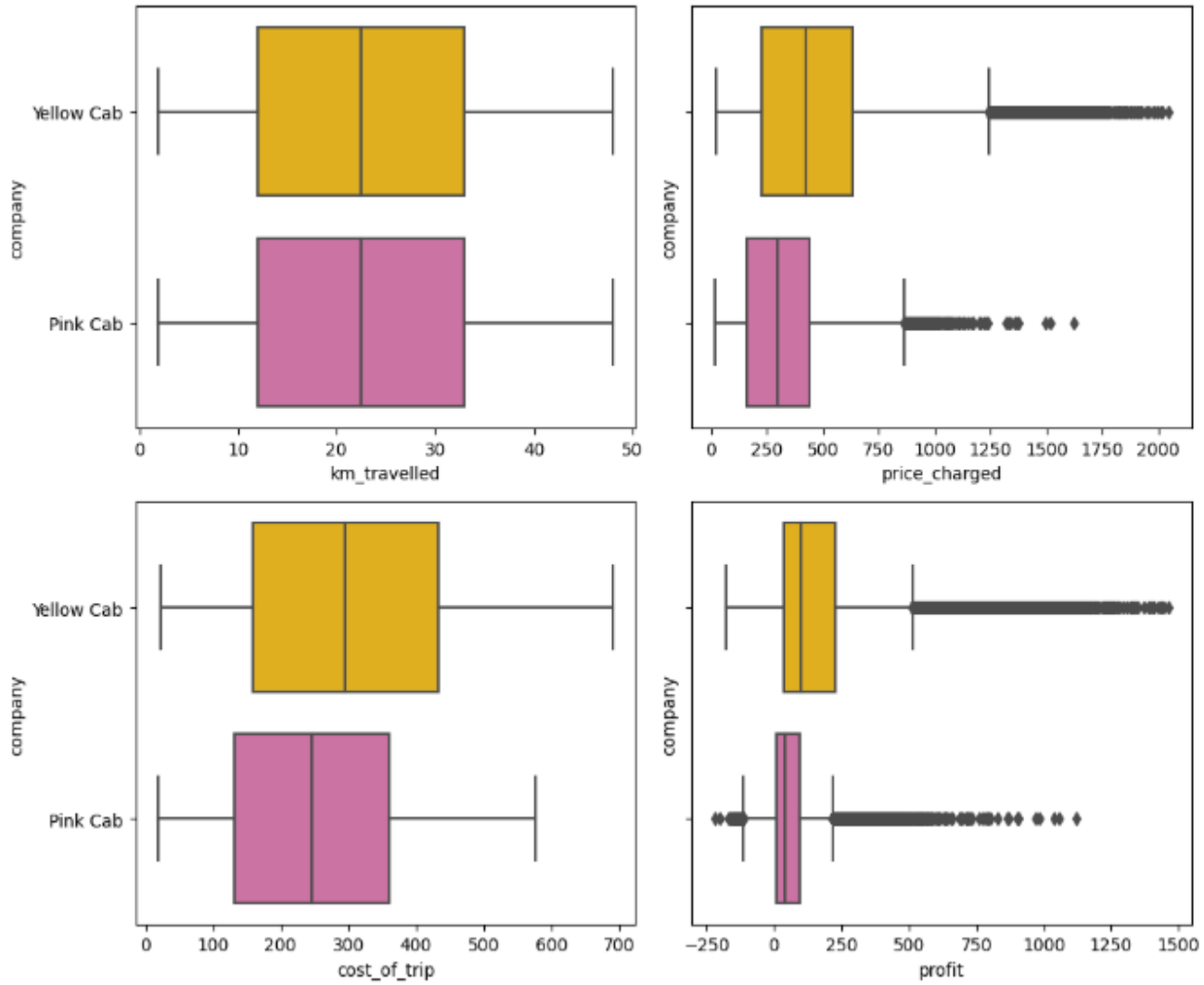
On any given month, there is a weekly seasonality where the number of trips are especially high during Monday and Tuesday for the year 2016.



For the year 2017, the seasonality starts to change. There is an increase in number of rides during Sunday, then it dips down on Monday, and then increases again during tuesday. This pattern is observed for both cab companies.

Data Glacier
Your Deep Learning Partner

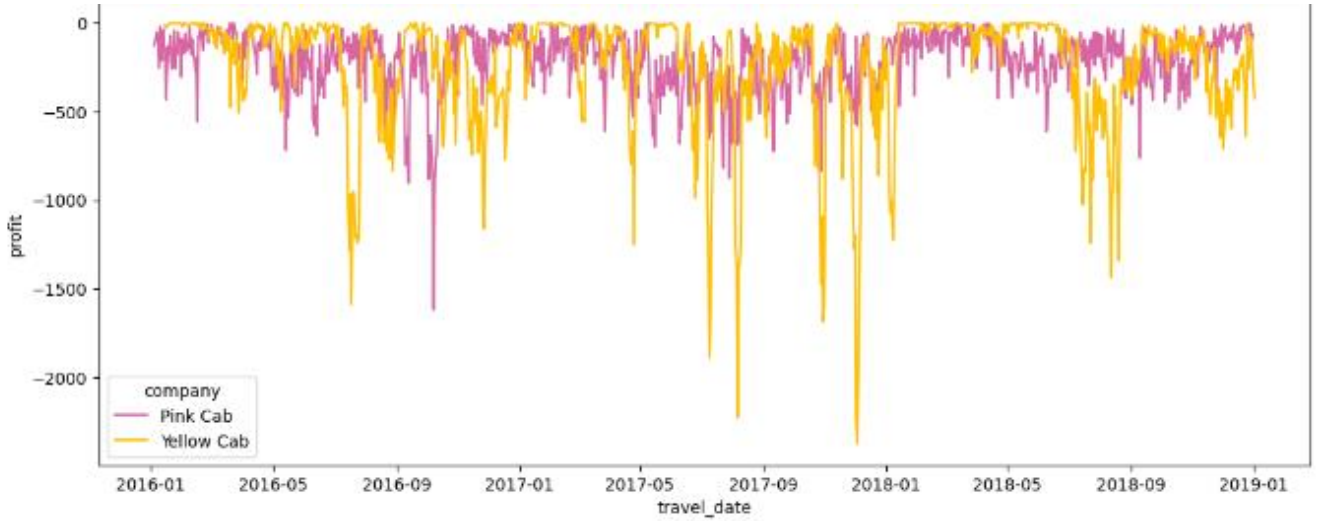
Profit Analysis



The plots above displays a distribution of features related to the trip. The distributions of km_travelled as well as cost_of_trip both follow a uniform distribution. Both Price charged, as well as profit follow a Gaussian distribution that is rightly skewed.

High outliers appear on the right side of both the profit and price change columns. Both cab companies have the same median distance travelled. **Yellow Cab** has higher cab expenses overall and the median price charged of **Pink Cab** is lower than it's rival company, although the profit of **Yellow Cab** is significantly higher.

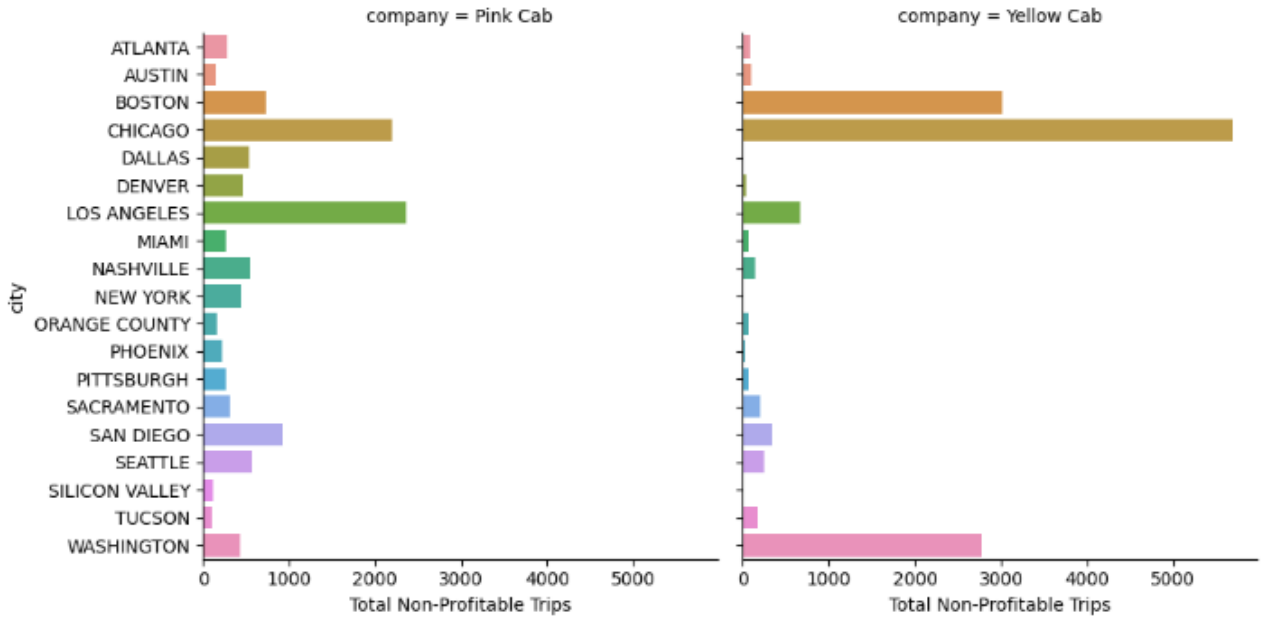
Loss Analysis



The above plot displays the trips that only made losses which is aggregated at a daily level by summing up the losses. that there are few clusters of losses during certain time periods at particular months. Most apparent in August.

- The greatest loss that **Yellow Cab** has experienced during a day was on 12-03-2017, where the company lost around 2373.17 dollars in profit
- The greatest loss that **Pink Cab** has experienced during a day was on 10-08-2016, where the company lost around 1617.50 dollars in profit

Loss Analysis Cont.



According to the data, the most number of non-profit trips made by **Yellow Cab** was Chicaco, Boston, and Washington. For the **Pink Cab**, the highest number of non-profit trips was Los Angeles, Chicago, and Boston.

EDA Summary

- Both Cab companies' financial performance is based on profit. Profit derived from the difference of the price charged and cost of trip for each trip. Both variables are highly correlated with the distance travelled for each trip.
- There is weekly, monthly and quarter seasonality on the number of rides in each period. The number of cab rides are higher during December and lowest during February.
- Yellow Cab seems to perform well almost on all cities and is able to make significantly more profit than its rival.

Recommendations

- In conclusion, we can measure a company's financial performance by looking at the total number of daily trips. Based on the observations, we can determine that Yellow Cab is a better investment than Pink Cab.

Thank You