Exploratory Data Analysis

G2M Cab Data Analysis

20 August 2022

Agenda

Executive Summary

Problem Statement

Approach

EDA

EDA Summary

Recommendations

Executive Summary

Executive Summary for **XYZ**, private firm in US.

Mission:

XYZ wants to invest in growing Cab business. XYZ company follows Go-to-Market(G2M/GTM) strategy.

G2M strategy involves a plan that can connect and convince their customers that their product is better than other companies in the industry.

The analysis covers the following:

- Understanding the cab data and city data.
- Computing the profit of cab companies to see which company holds a large share of profits.
- Conclude with hypothesis to invest in one of the Cab company.

Problem Statement

Analyze the Cab data to identify which Cab company is a better fit for investment for XYZ firm. XYZ firm wants to invest in a Cab company as it falls under the Go-To-Market industry segment.

We need to analyze which company customers prefer more. This is the foundation of G2M industry. Customers would prefer a particular cab based on its availability and price charged. Commuting is an essential part of everyone's life and finding a best fit for price charged and availability would attract more customers.

The Dataset consists information of two different cab companies – **Pink Cab** and **Yellow Cab**. EDA is performed to determine which Cab company is a better fit for XYZ firm's investment. Regression hypothesis is done to confirm the assumptions made from EDA.

Approach

- The data is for a time period of 31/01/2016 to 31/12/2018. Data consists information of the following –
  - Cab_Data – Covers transactions for two cab companies
  - Customer_ID – Consists details of customer with their location
  - Transaction_ID – Consists information of Customers and their payment mode
  - City – Consists of city information
- Profit is computed from the Cost of Trip and Price Charged data. The data does not have any null values. Hence, not much pre-processing is required.
- The data files have been merged based on the unique IDs. This helps to understand the relationships between the data and the features.

Data discovery –

US weather data can help us analyze the impacts of weather on cab profits and services.

US weather data has been recorded from 2016 – 2021 for different cities and has been categorized into Sever cold, fog, hail, rain, snow, storm and other precipitation. The weather data is provided by – Moosavi, Sobhan, Mohammad Hossein Samavatian, Arnab Nandi, Srinivasan Parthasarathy, and Rajiv Ramnath. "Short and Long-term Pattern Discovery Over Large-Scale Geo-Spatiotemporal Data." In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ACM, 2019.

EDA

The correlation table is computed and shown below –

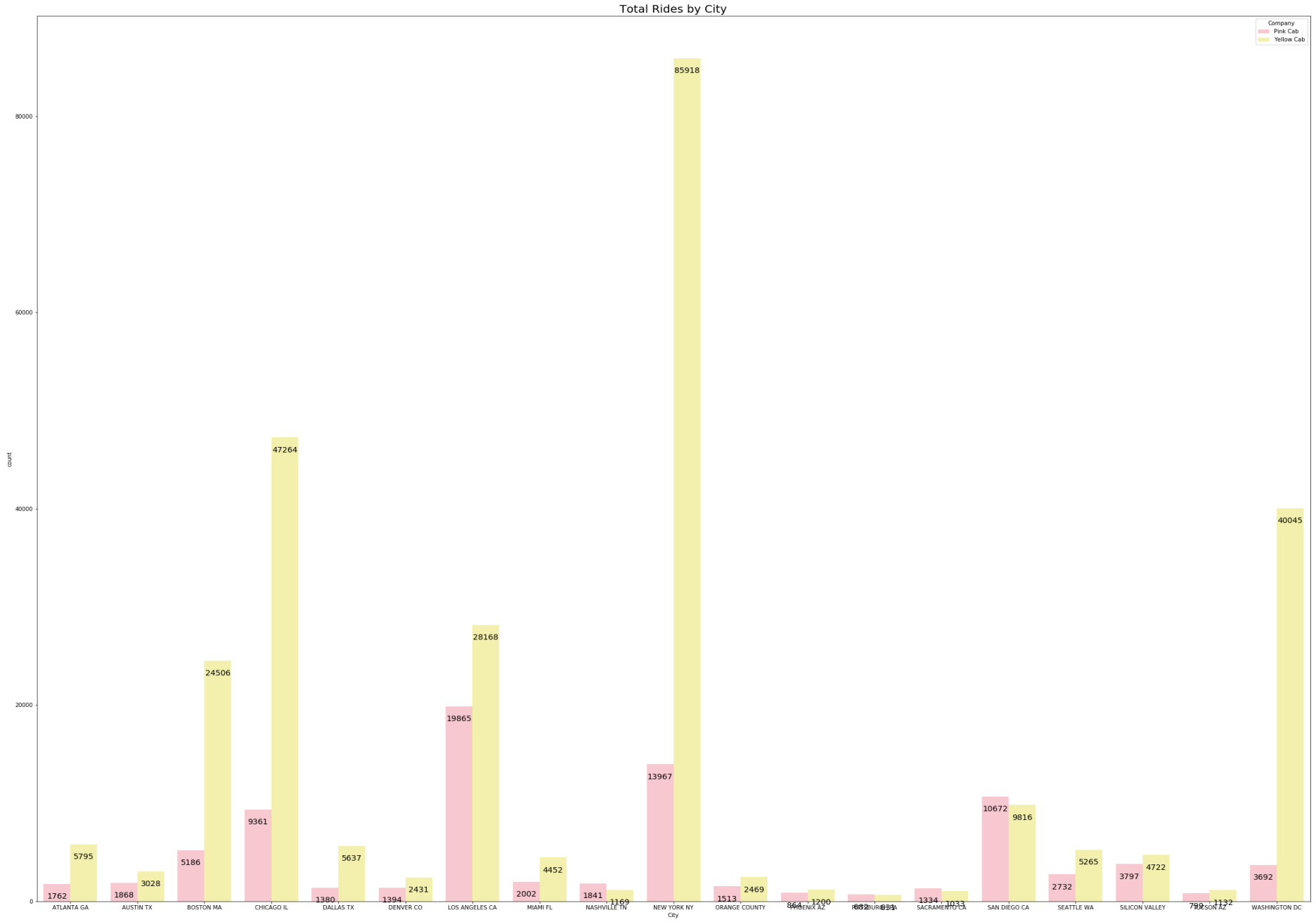
	Transaction ID	Date of Travel	KM Travelled	Price Charged	Cost of Trip	Customer ID	Age	Income (USD/Month)
Transaction ID	1.000000	0.993030	-0.001429	-0.052902	-0.003462	-0.016912	-0.001267	-0.001570
Date of Travel	0.993030	1.000000	-0.001621	-0.055559	-0.004484	-0.017653	-0.001346	-0.001368
KM Travelled	-0.001429	-0.001621	1.000000	0.835753	0.981848	0.000389	-0.000369	-0.000544
Price Charged	-0.052902	-0.055559	0.835753	1.000000	0.859812	-0.177324	-0.003084	0.003228
Cost of Trip	-0.003462	-0.004484	0.981848	0.859812	1.000000	0.003077	-0.000189	-0.000633
Customer ID	-0.016912	-0.017653	0.000389	-0.177324	0.003077	1.000000	-0.004735	-0.013608
Age	-0.001267	-0.001346	-0.000369	-0.003084	-0.000189	-0.004735	1.000000	0.003907
Income (USD/Month)	-0.001570	-0.001368	-0.000544	0.003228	-0.000633	-0.013608	0.003907	1.000000

The features of Transaction ID and Date of Travel are highly correlated. The numbers closest to 1 have high correlation or they are related positively.

The numbers with negative correlation are correlated negatively.

EDA Summary

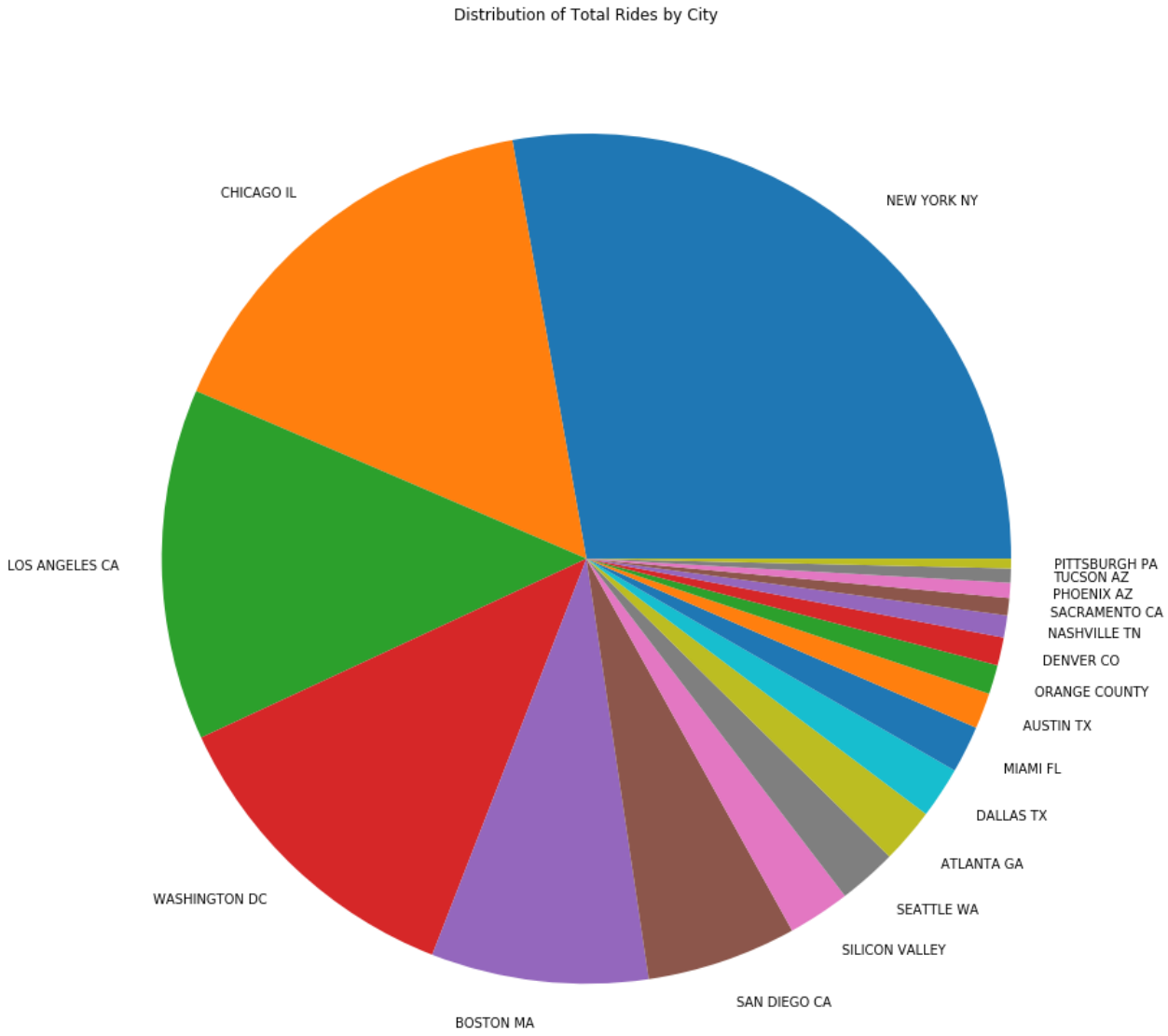
New York, Chicago, Washington are top 3 cities with highest yellow cab customers. Los Angeles, New York and San Diego are top 3 cities with highest Pink Cab customers.



Data Glacier
Your Deep Learning Partner

EDA Summary

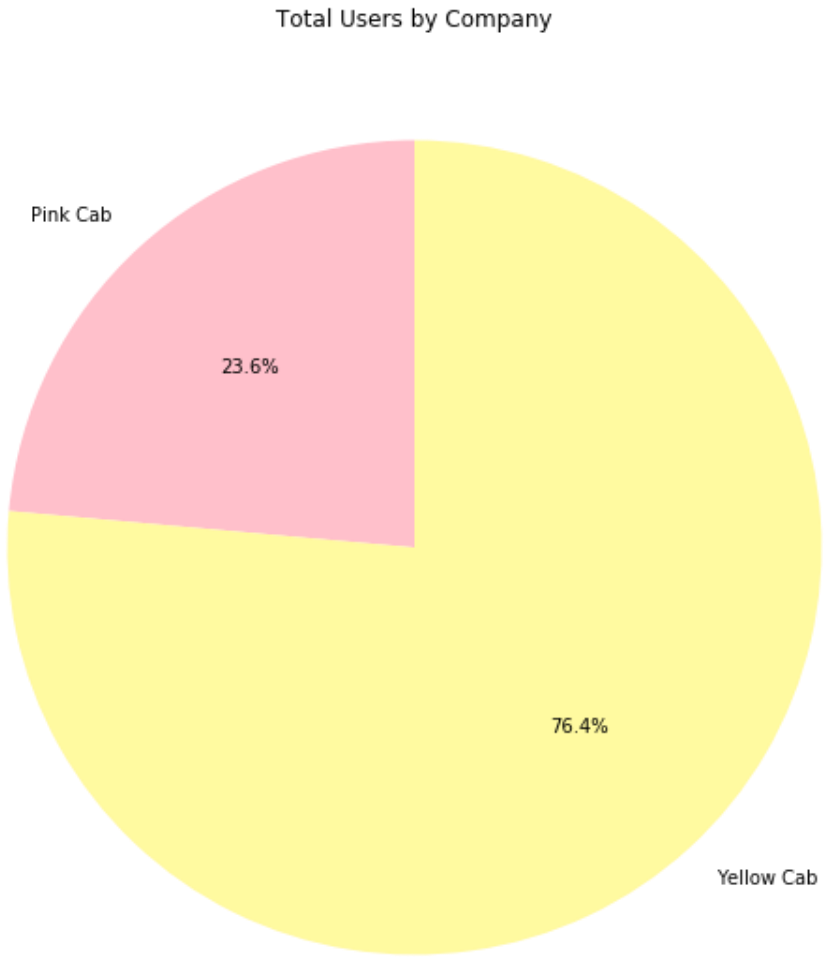
New York, Chicago, Los Angeles, Washington DC are the top city with total number of rides in US.



Data Glacier
Your Deep Learning Partner

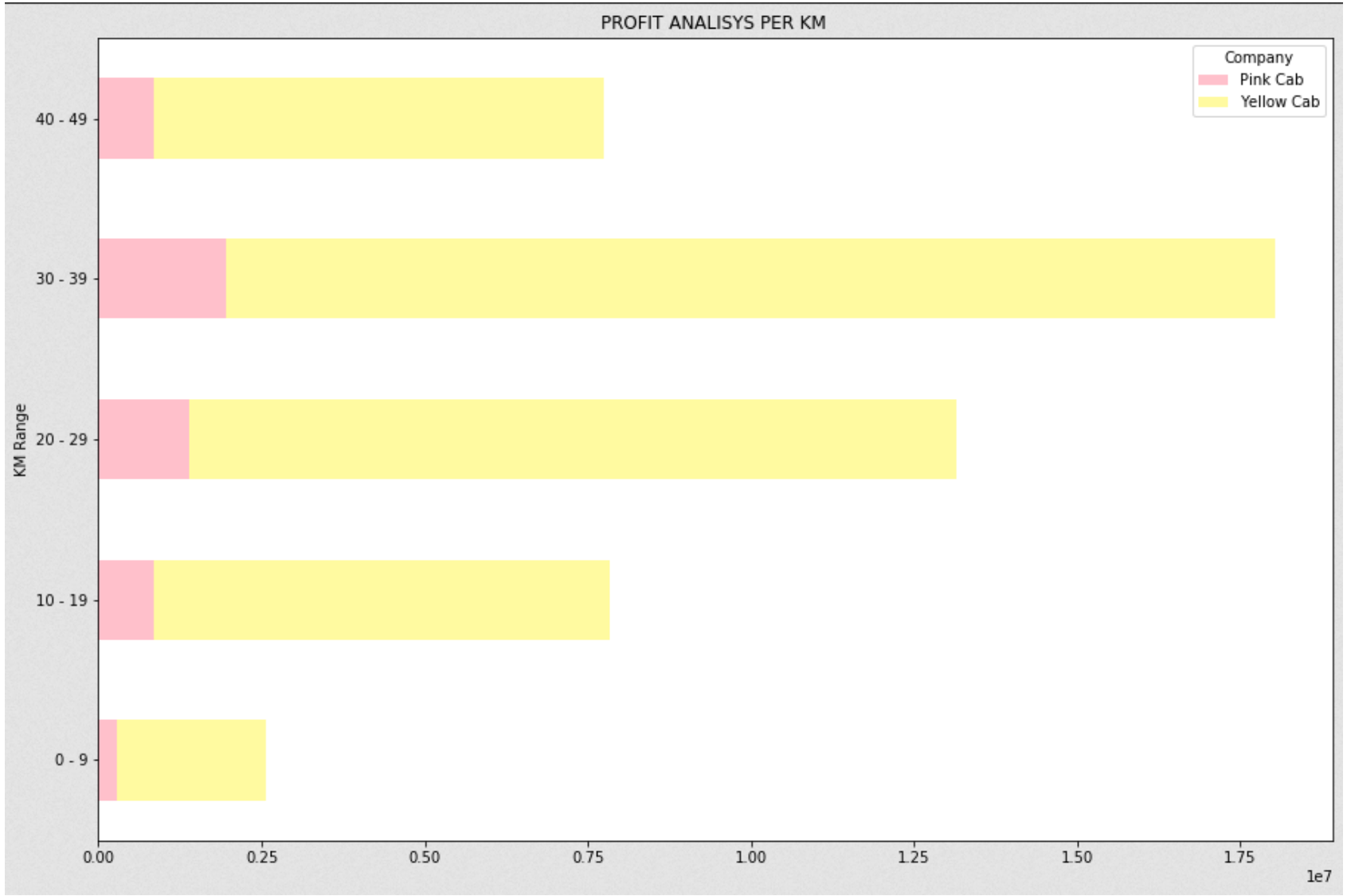
EDA Summary

Following chart shows the total number of users for both the cab companies. Yellow Cab has 76.4% Cab users whereas 23.6% people use Pink Cab.



EDA Summary

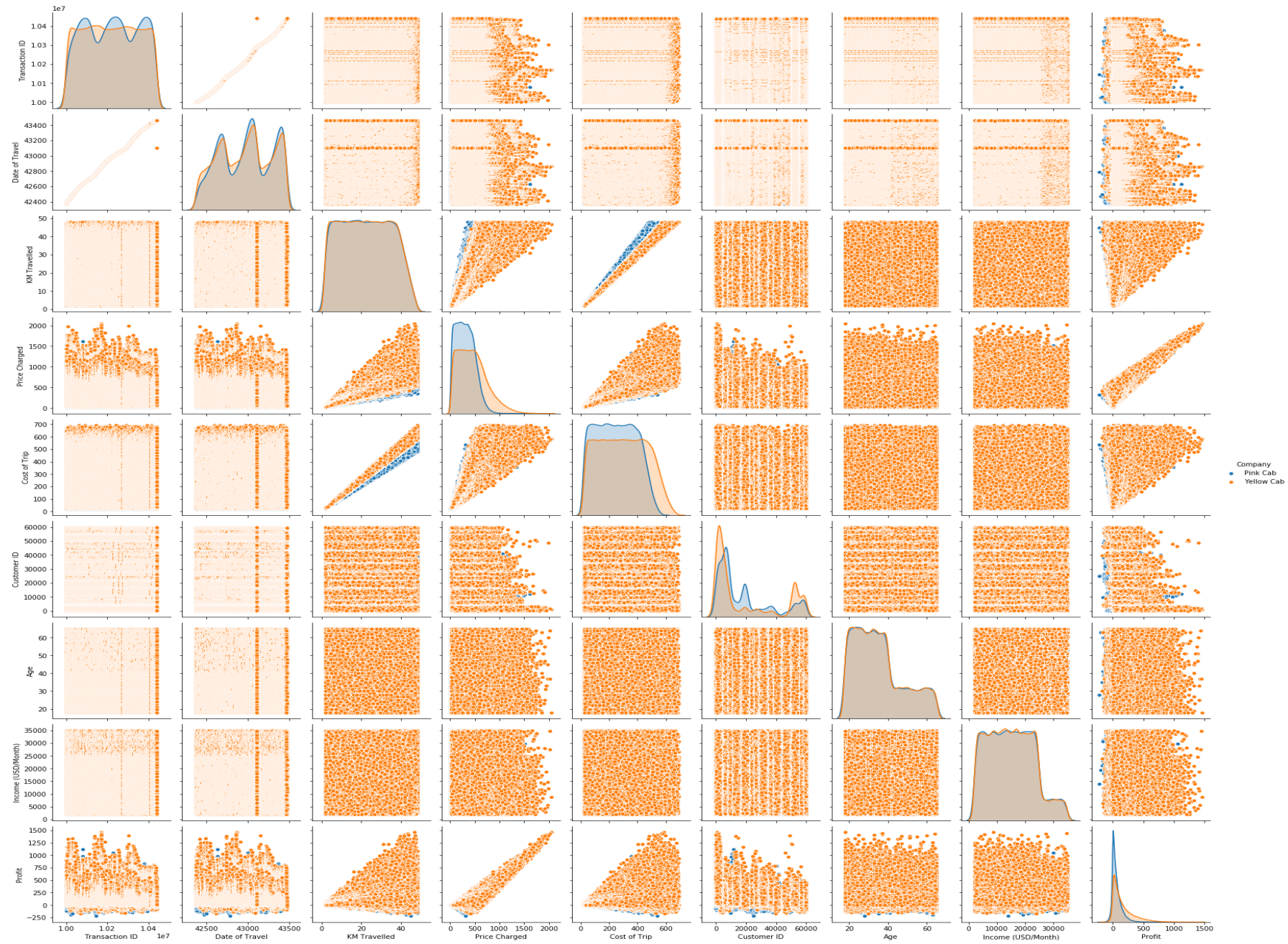
Following chart shows the total number of users for both the cab companies. Yellow Cab has 76.4% Cab users whereas 23.6% people use Pink Cab.



Data Glacier
Your Deep Learning Partner

EDA Summary

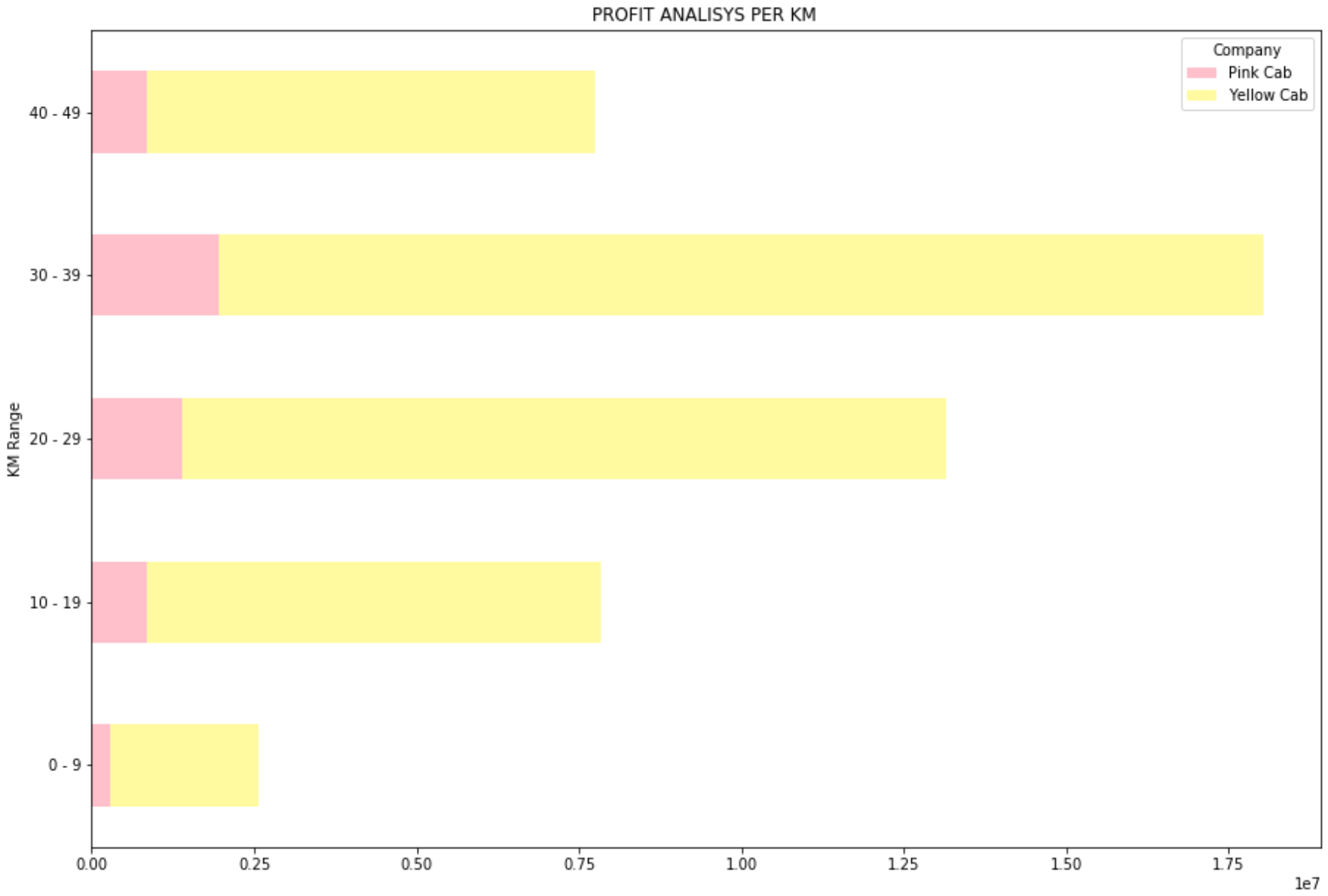
Pair plot for all the features and Profit. Profit is computed from Price Charged and Cost of Trip.



Data Glacier
Your Deep Learning Partner

EDA Summary

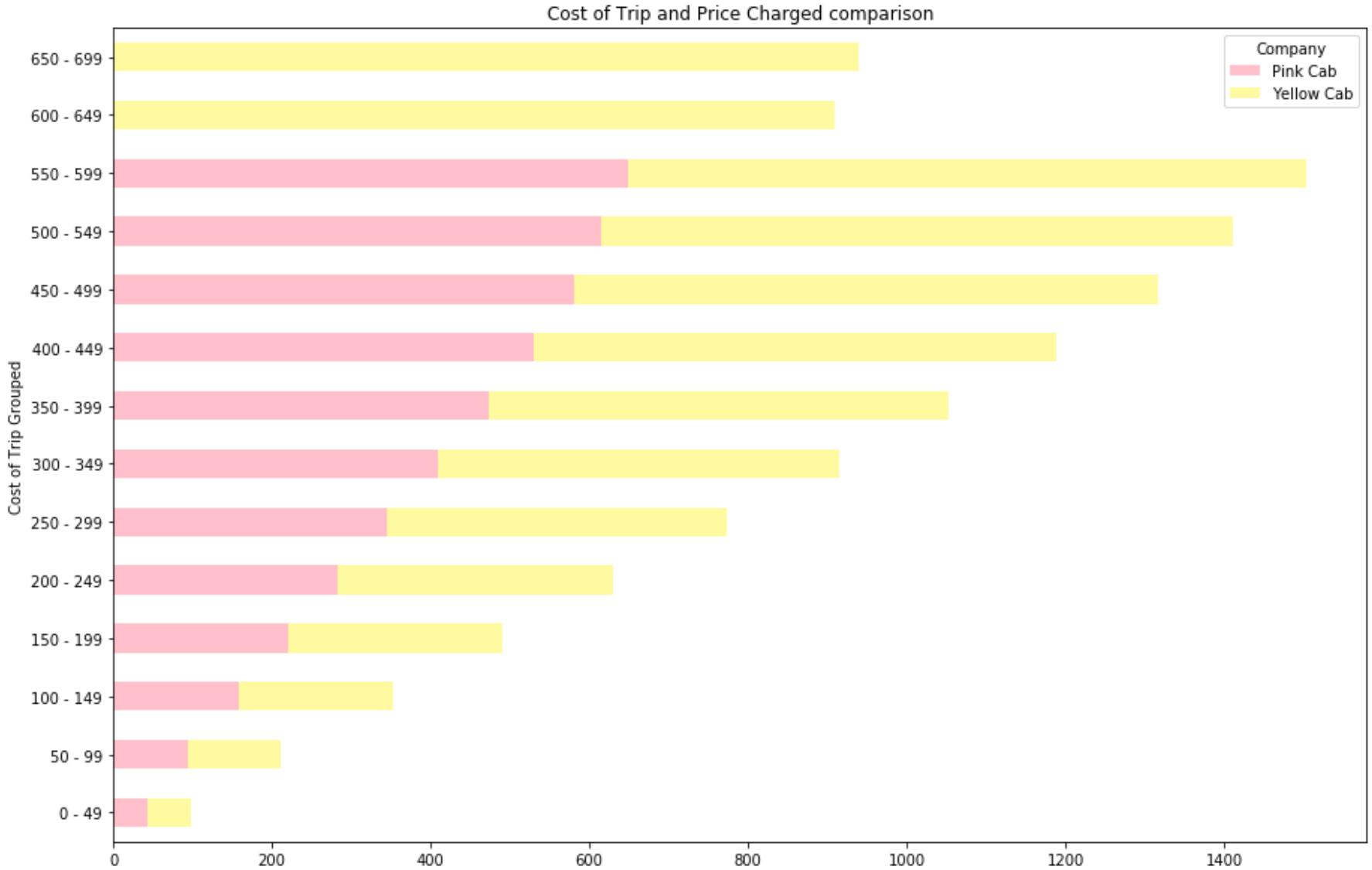
Analysis of Profit per KM travelled. The KM is categorized in categories of every 10 KM range. The maximum profit is earned for KM in the 30-39 KM range. The next category for profit is 20-29 KM range. Both companies have least profit in 0-9 KM range.



Data Glacier
Your Deep Learning Partner

EDA Summary

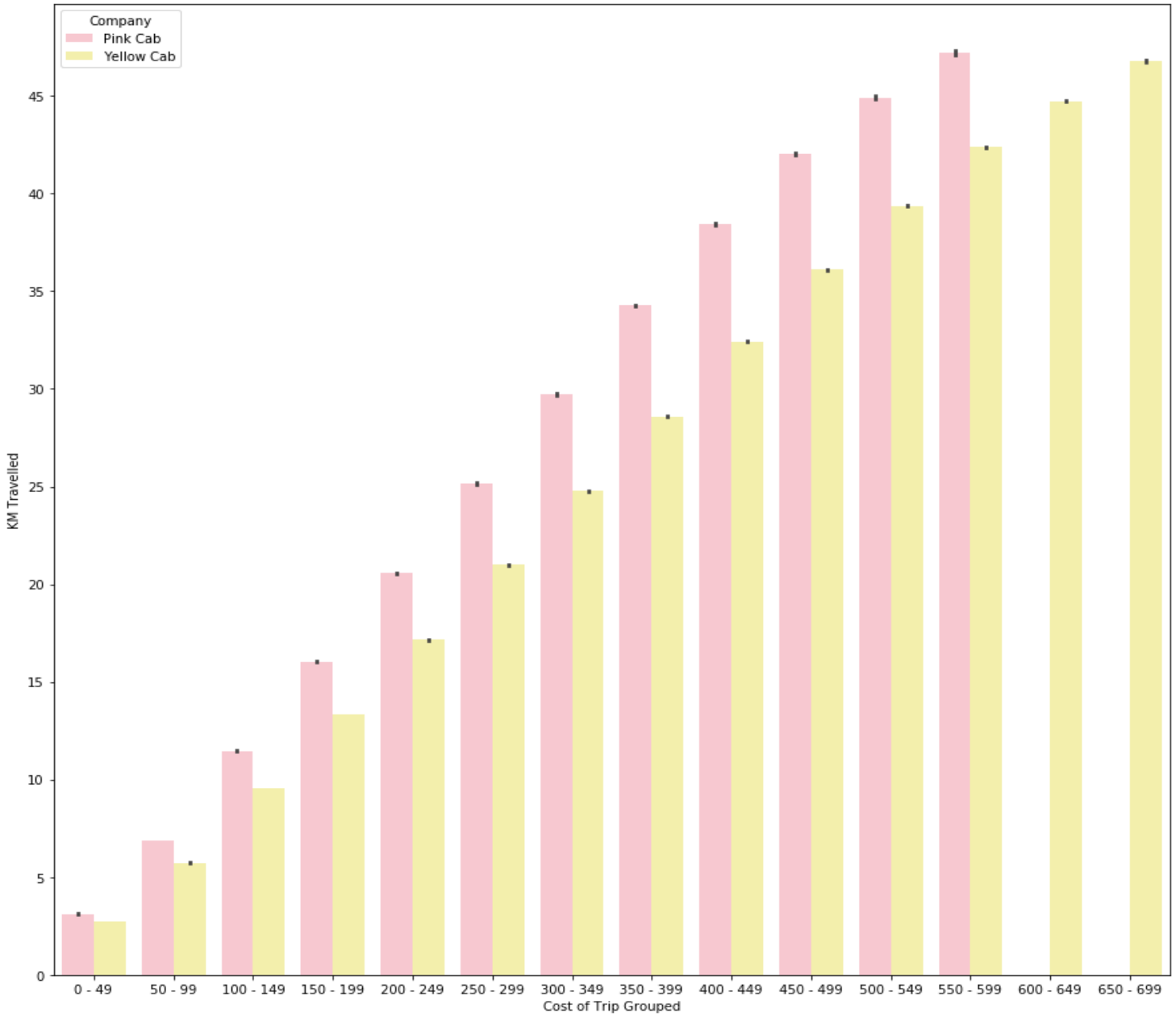
Analysis of Cost of Trip and Price Charged. Both companies seem to have a similar trend. However, yellow cab has outliers for 600-700 price range. There is no data available for pink cab for that range.



Data Glacier
Your Deep Learning Partner

EDA Summary

Analysis of Cost of Trip for KM travelled. Pink cab has more cost for KM travelled when compared to Yellow cab.



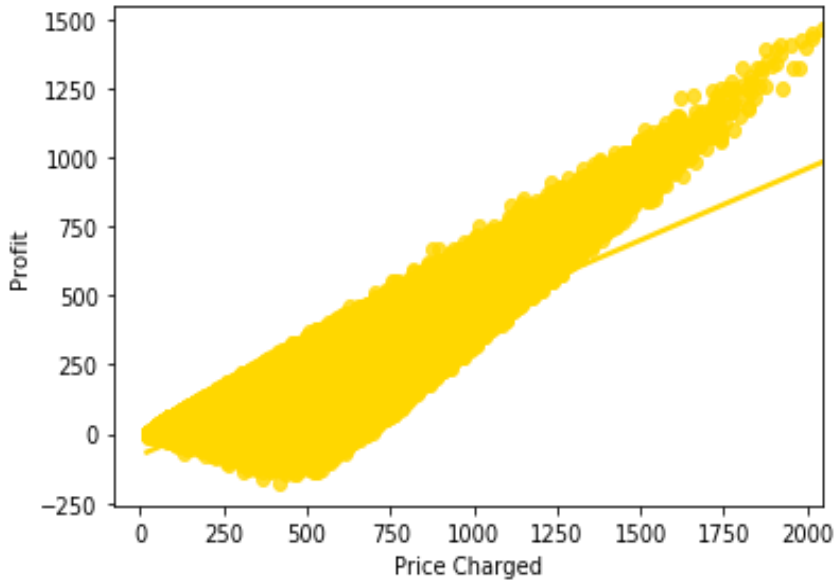
Data Glacier
Your Deep Learning Partner

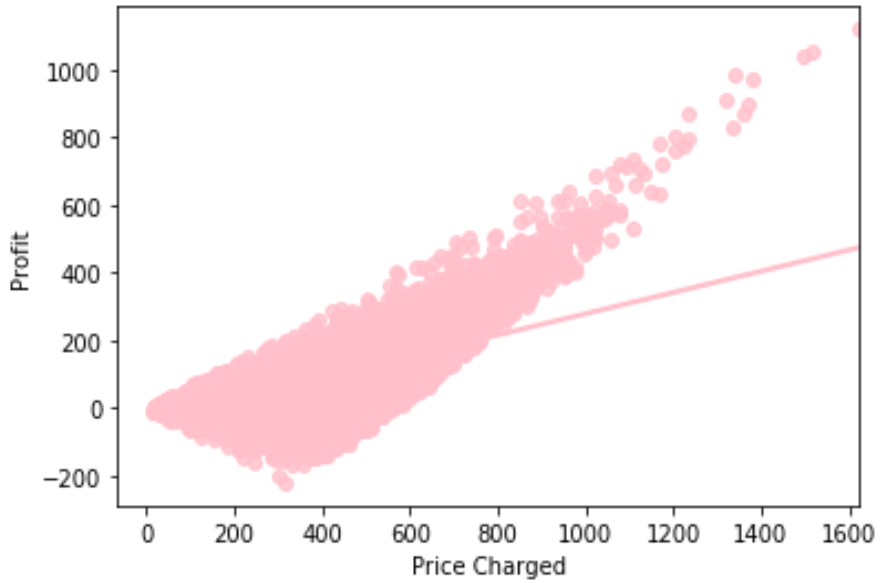
EDA Summary

Hypothesis

Regression Hypothesis for Profit and Price charged.

Yellow cab covers more points on the regression line. Pink cab does not cover as many points on the regression line and has many outliers.





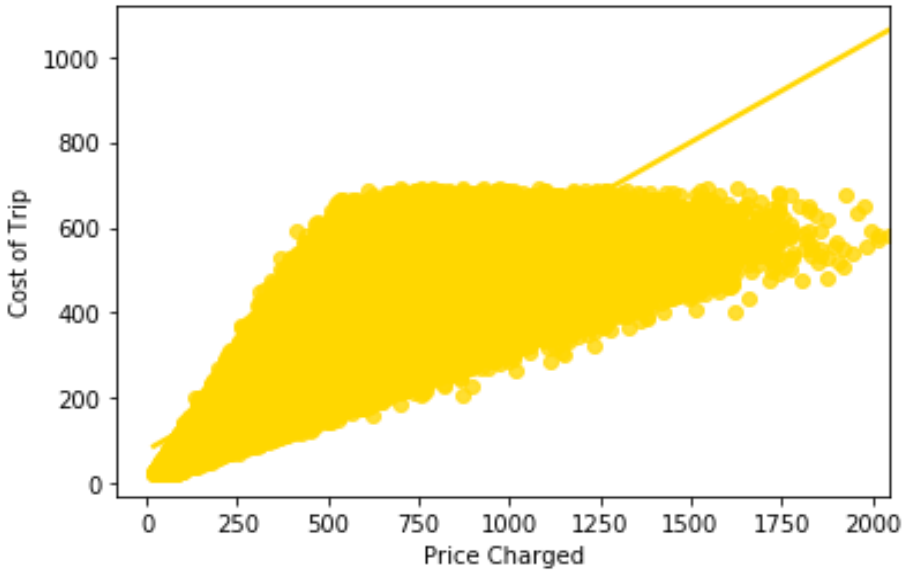
Data Glacier
Your Deep Learning Partner

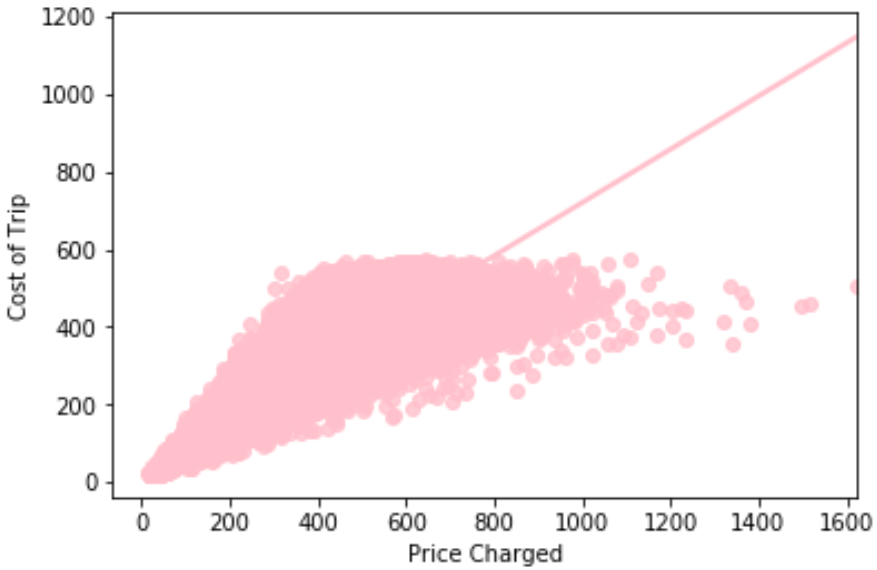
EDA Summary

Hypothesis

Regression Hypothesis for Cost of Trip and Price charged.

Yellow cab covers more points on the regression line. Pink cab does not cover as many points on the regression line and has many outliers.





Recommendations

The conclusion after analysis of both cab companies is as follows –

- Pink Cab charges higher cost for KM travelled compared to Yellow Cab
- Yellow cab has better customer reach than Pink Cab.
- Yellow cab has better profits compared to Pink Cab.
- Yellow Cab also travels to higher KM range than pink cab. We can assume that they want to grow their business or expand their areas of service.

Yellow Cab company is a better fit for XYZ company investment from all the analysis and Regression Hypothesis.

Thank You