G2M Case Study – Cab Investment Firm

Exploratory Data Analysis

2/20/2024

Agenda

Executive Summary

Problem Statement

Approach

EDA

EDA Summary

Recommendations

Background

XYZ is a private equity 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.

Objective: Provide actionable insights to help XYZ firm in identifying the right company for making investment.

The analysis will provide information about the following:

Data Understanding

Forecasting profit and number of rides for each cab type

Finding the most profitable Cab company

Recommendations for investment

Background – Assumptions & Datasets

- Timeframe of the data: 2016-01-01 to 2018-12-31
- Total data points: 355,032

Assumptions:

- Outliers are present in Price Charged feature but due to unavailability of trip duration details ,we are not treating this as outlier.
- Profit of rides are calculated keeping other factors constant and only Price Charged and Cost of Trip features used to calculate profit.
- Users feature of city dataset is treated as number of cab users in the city. We have assumed that this can be other cab users as well(including Yellow and Pink cab)

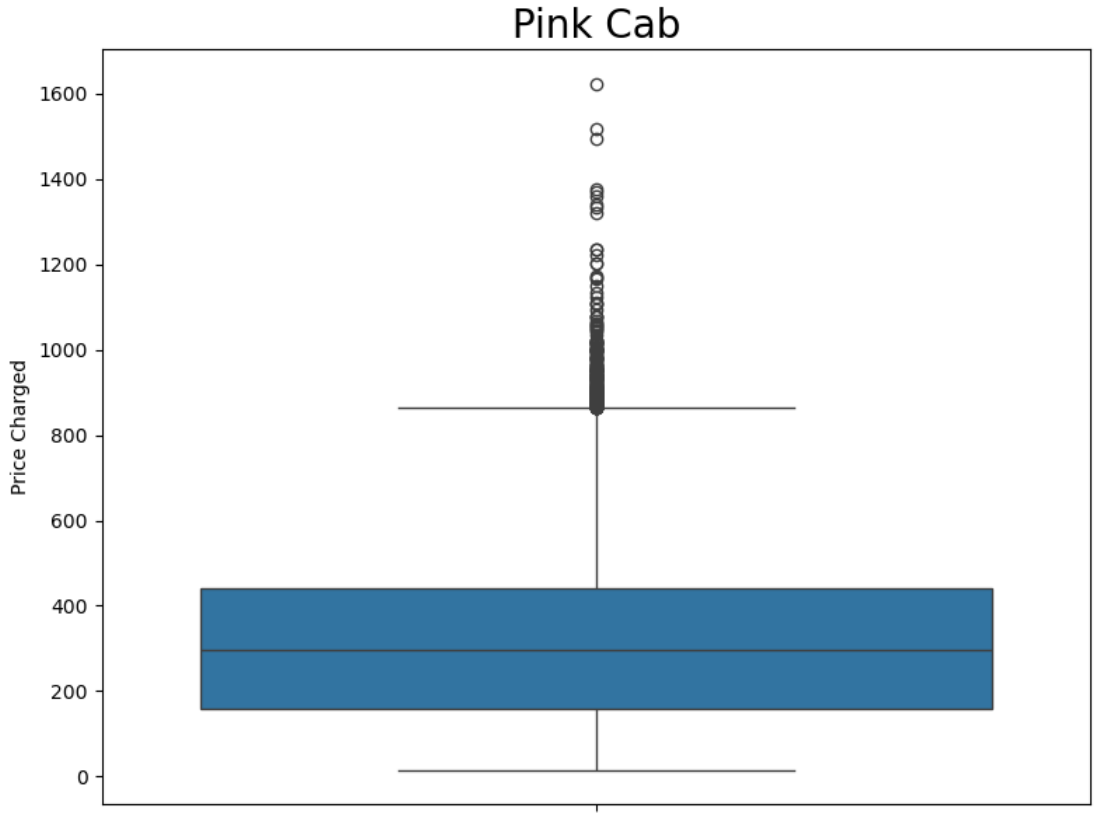
- Cab_Data.csv - details of transaction for 2 cab companies: yellow and pink cab company
- Customer_ID.csv - mapping table that contains customer's demographic details.
- Transaction_ID.csv - mapping table that contains transaction to customer mapping and payment mode.
- City.csv - list of US cities, their population and number of cab users.

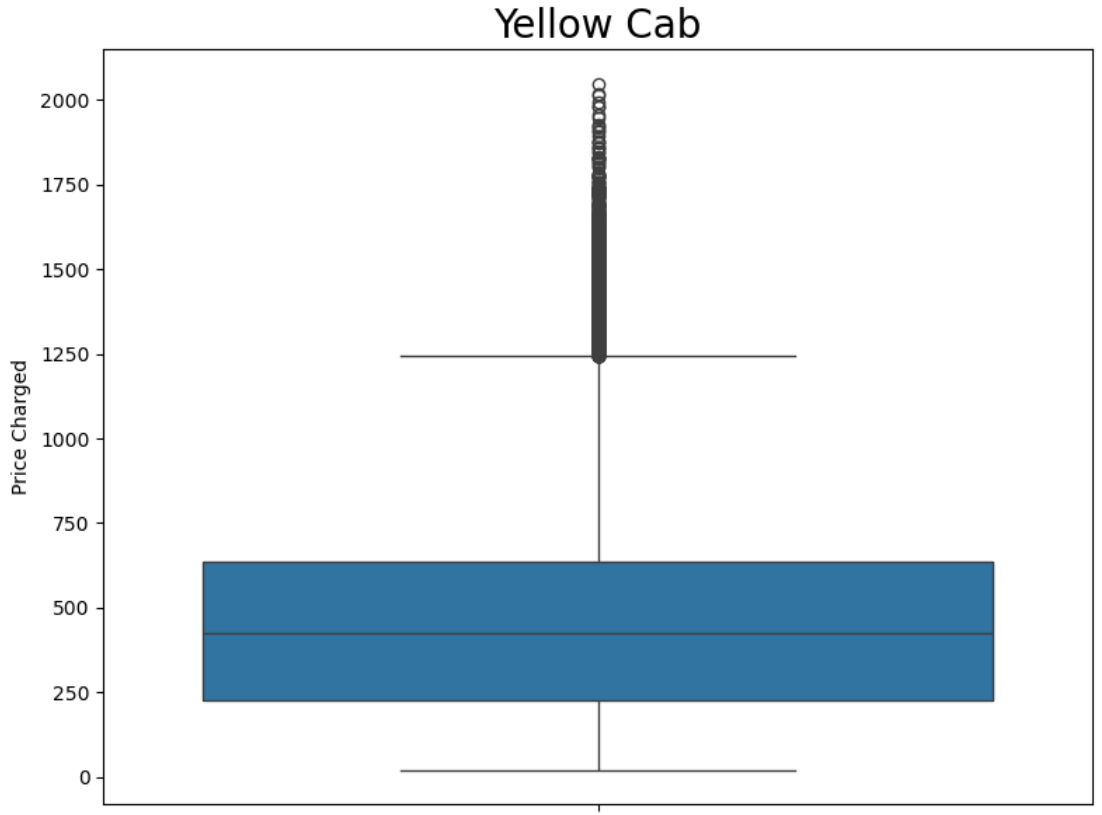
Hypothesis

- As there are a variety ranges of cars in terms of their market price, we expect to see a few outliers for the expensive/high-end cars or even the cheaper/low-end cars.
- We know that the majority of the population has an age range of roughly 20 to 40 years old. We expect to see the most amount of data at this age group as a bell curve distribution
- We expect to see similar applications to the income groups as the we expect incomes towards a potential bell curve distribution to have the most amount of usage of cabs.
- We expect there to be a higher amount of cab usage during the colder months of the year as people may tend to prefer said method of travel to avoid the cold.
- We expect to see a profit rate that is identical to the bell curve distribution for kilometers traveled.
- We expect to see more cab usage in the largest cities in America.
- Continuation, as there are more cab usage in the largest cities, we expect to see a higher amount of profit.

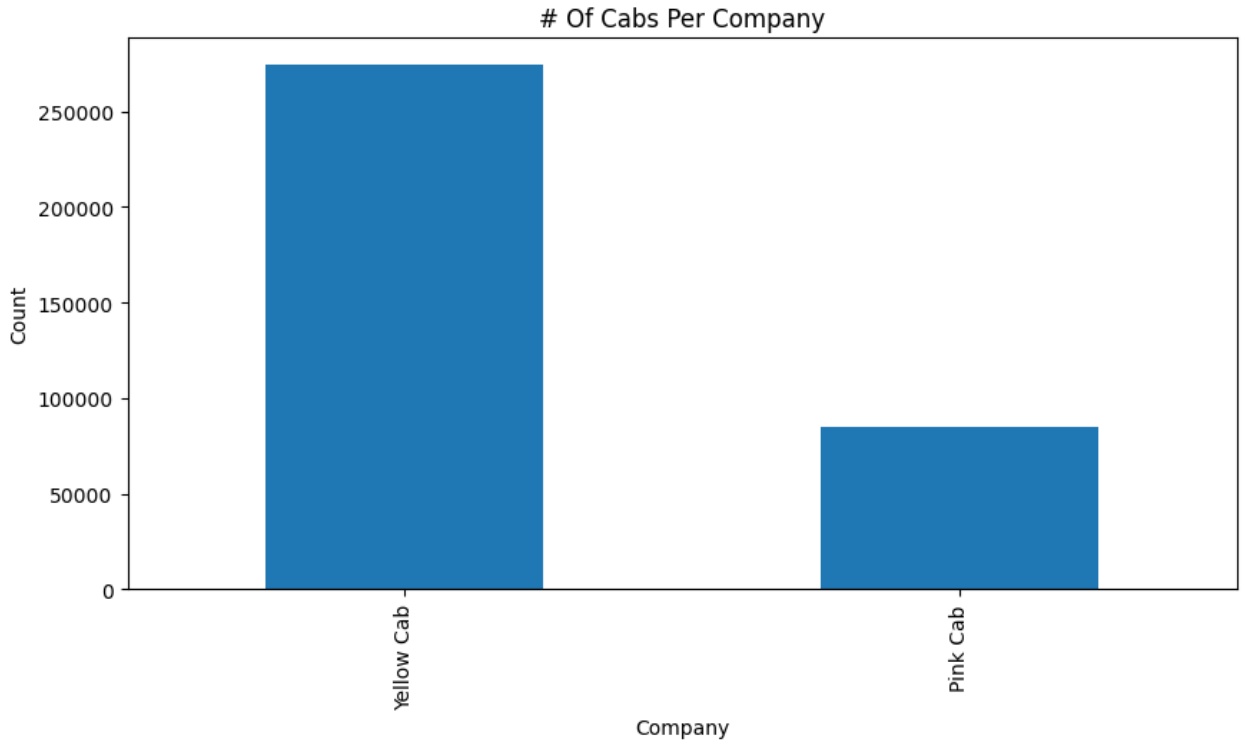
Exploratory Data Analysis – Price Changes & Outliers

- Higher price range for yellow cab company than pink cab company.
- The outliers are due to use of high-end cars.





Exploratory Data Analysis – Quantity of Cabs

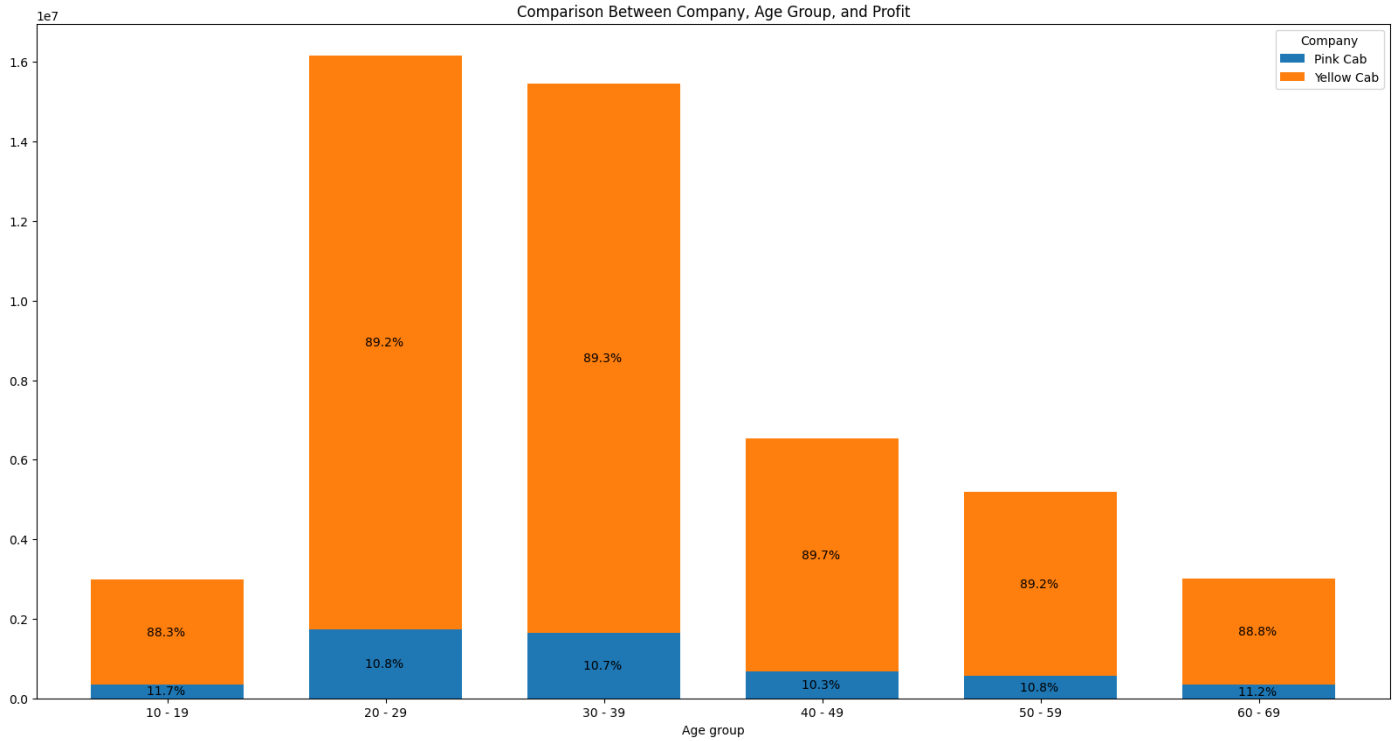


- Yellow cab company has a significantly higher count than pink cab company.

- Suggests yellow cab companies' size is much larger than pink cab company.

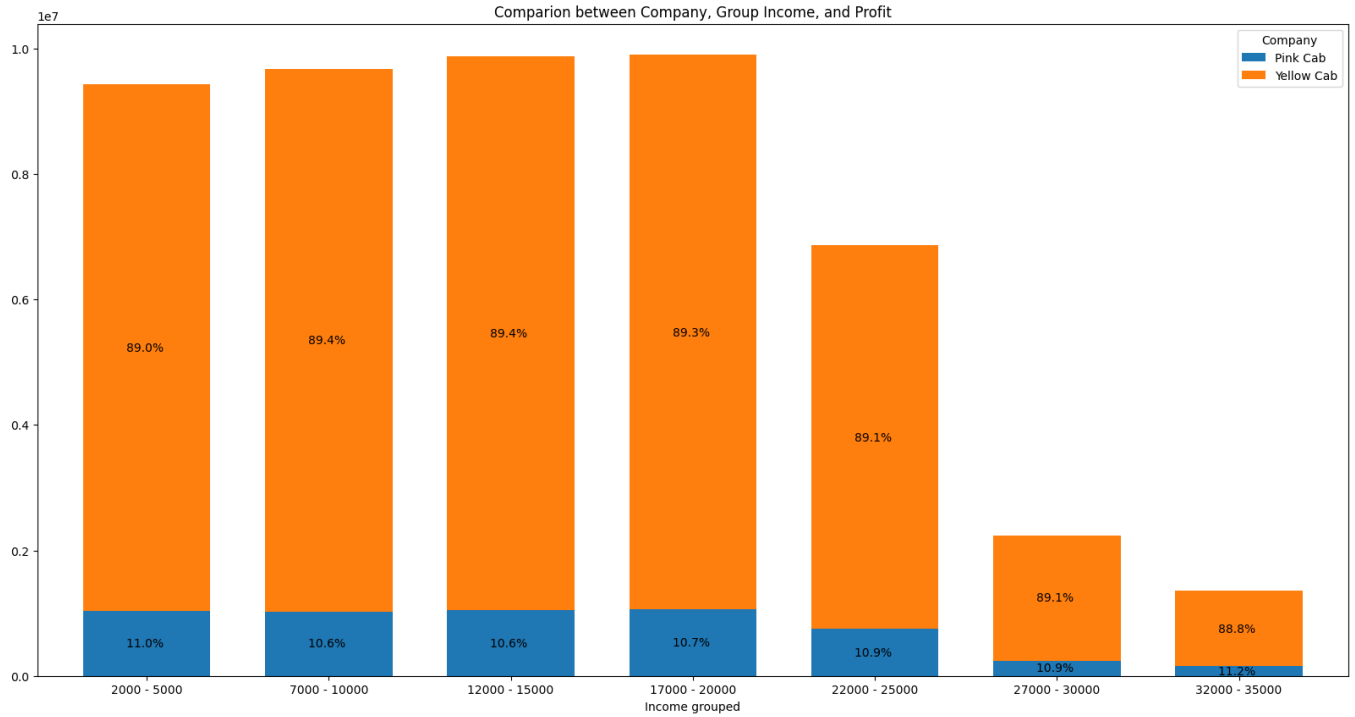
EDA: Profit – Comparison between Company & Age group

- We see that the yellow cab company makes nearly 90% of the profit regardless of age when compared to the pink cab company.



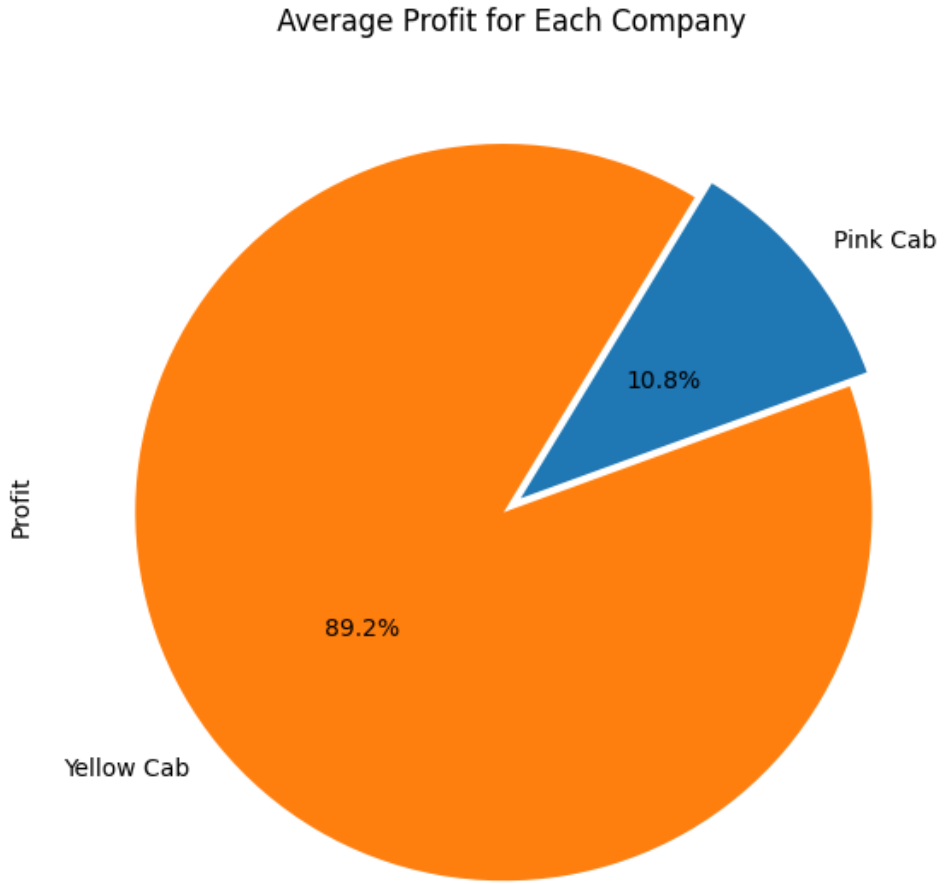
EDA: Profit – Comparison between Company & Group Income

- We achieve similar results to the previous plot. We see that the yellow cab company makes nearly 90% of the profit regardless of group income when compared to the pink cab company.



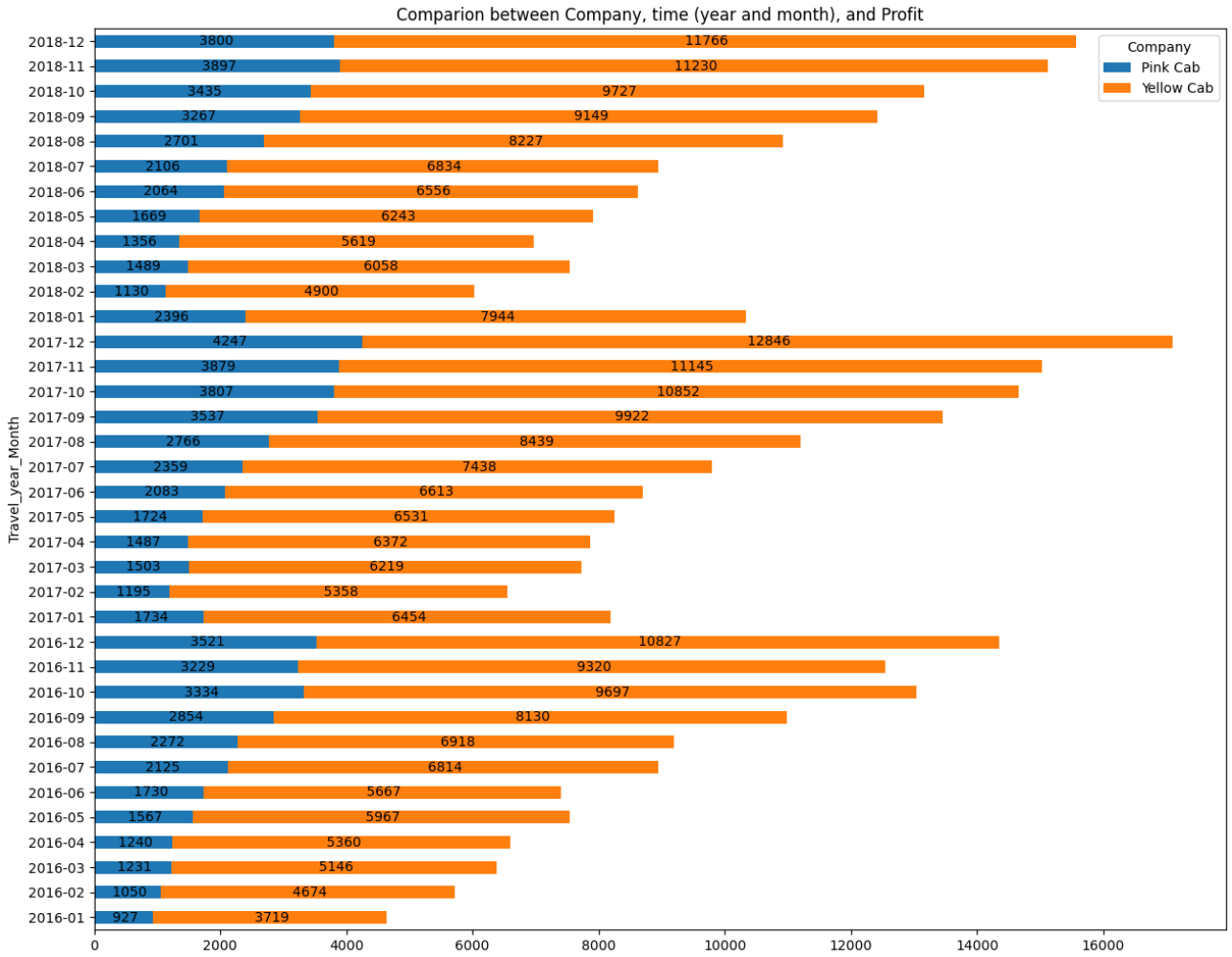
EDA: Profit – Average

- We wanted to see the exact average of profit for each company. We see that the yellow cab company makes 89.2% average profit compared to the pink cab company.



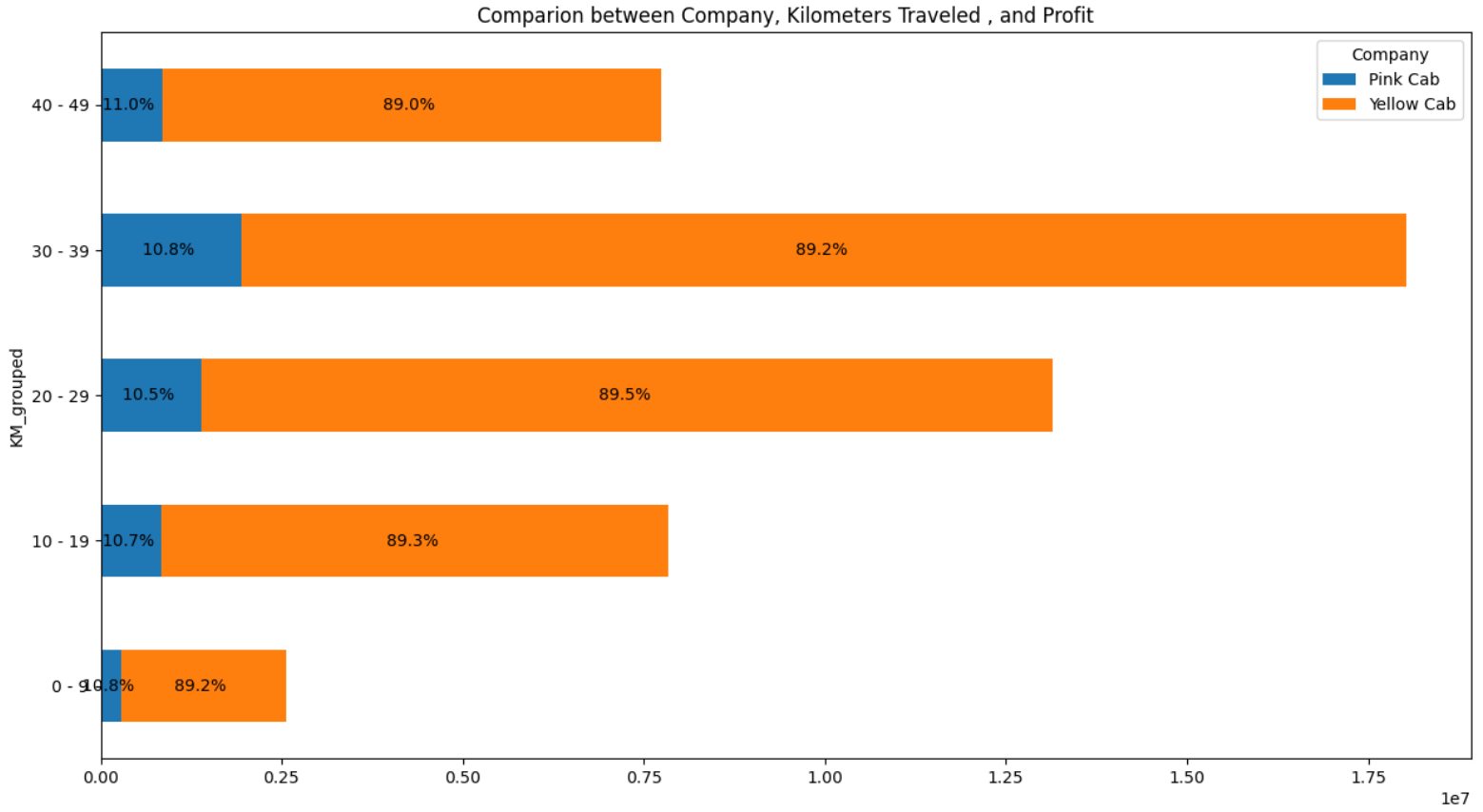
EDA: Profit – Month and Year

- We see that regardless of the year or month, the yellow cab company surpasses the amount of profit for the pink cab company
- It is worth noting the patterns in months. We see that there is a higher amount of profit towards the end of the year for both companies.



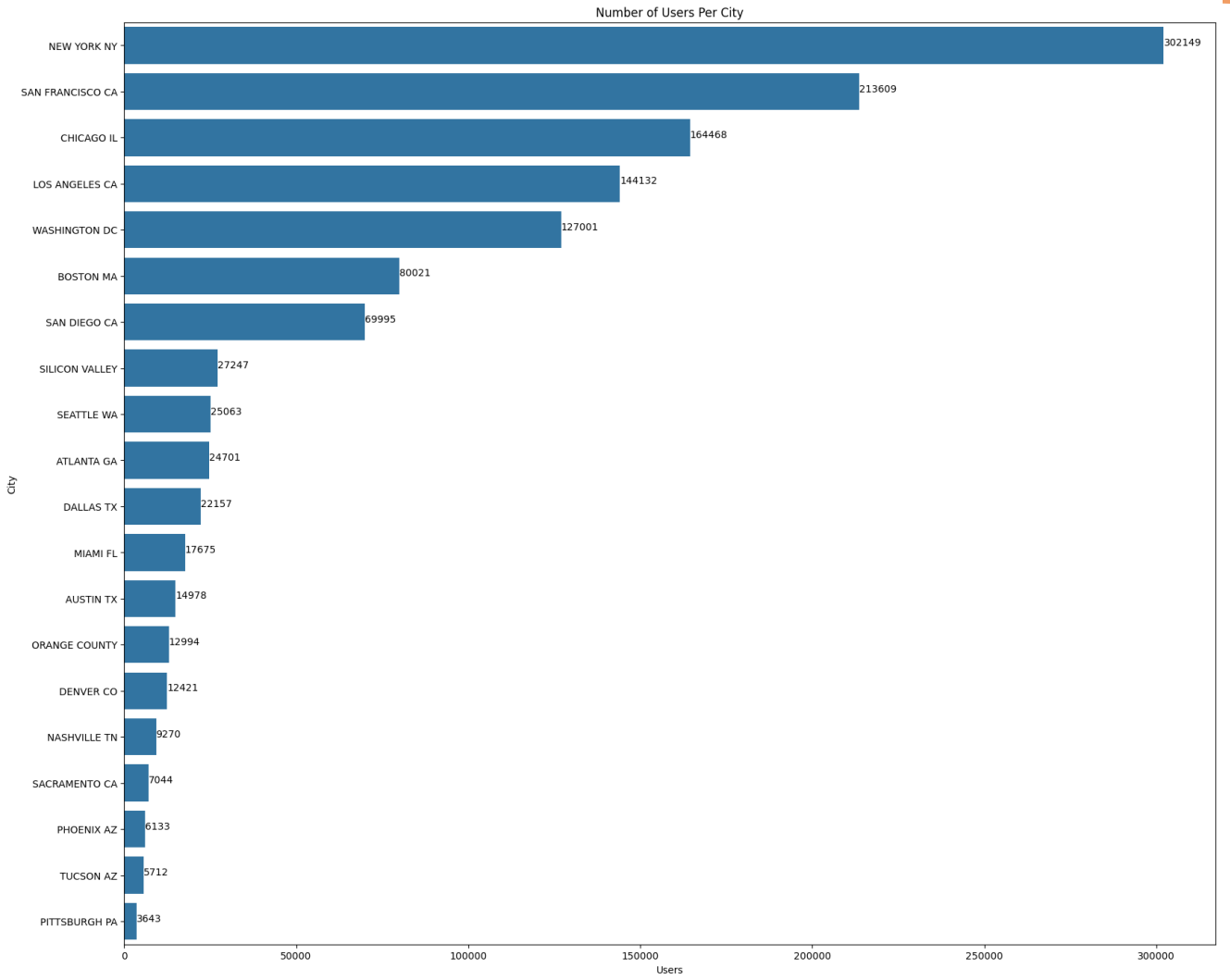
EDA: Profit – Comparison between Company & Kilometers traveled

- We see that again; the yellow cab company makes nearly 89% to 90% of the profit compared to the pink cab company regardless of kilometers traveled.

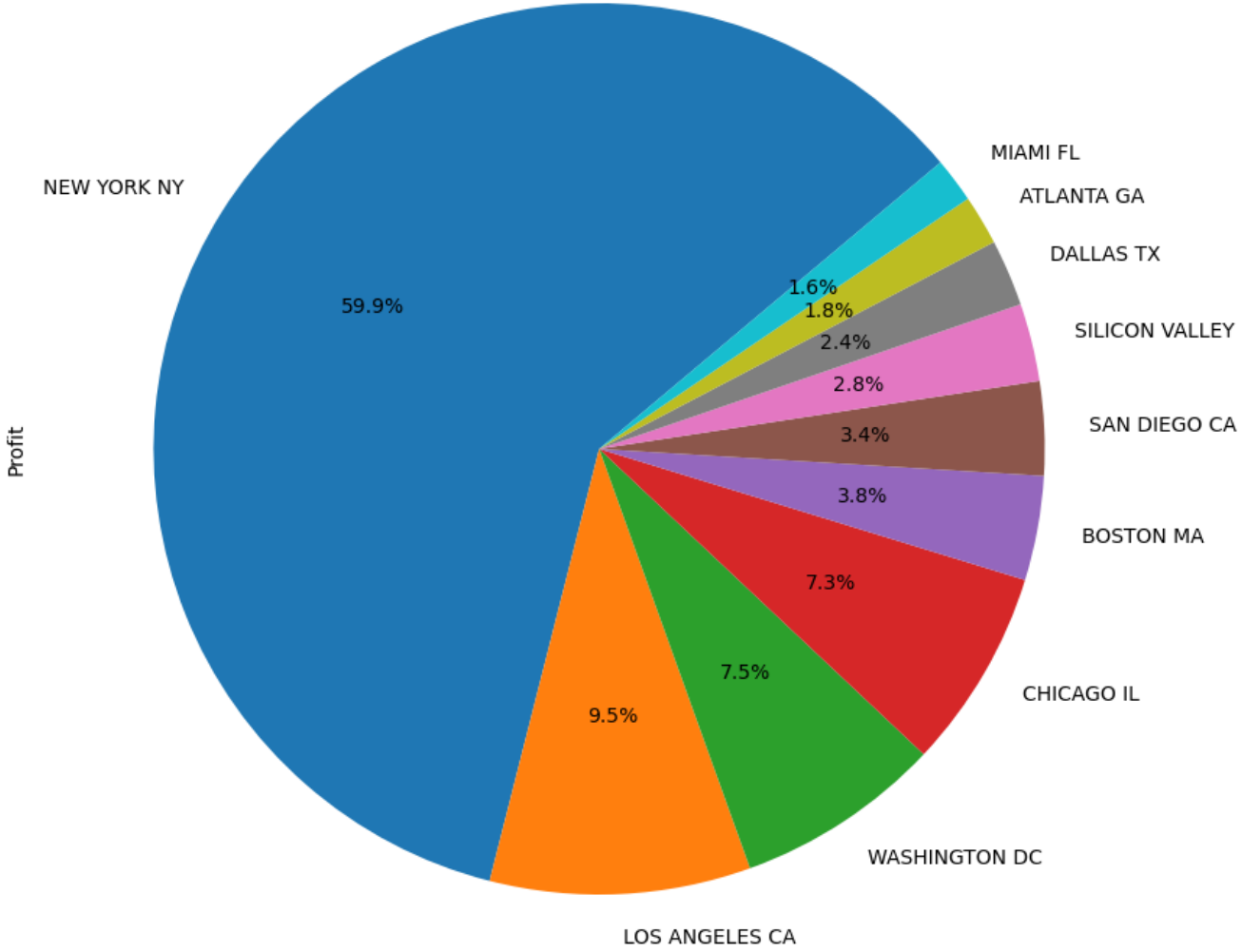


EDA: Number of uses per city

- We see that New York, San Francisco, Chicago, Los Angeles, and Washington DC make up most of the pink and yellow cab company usage.



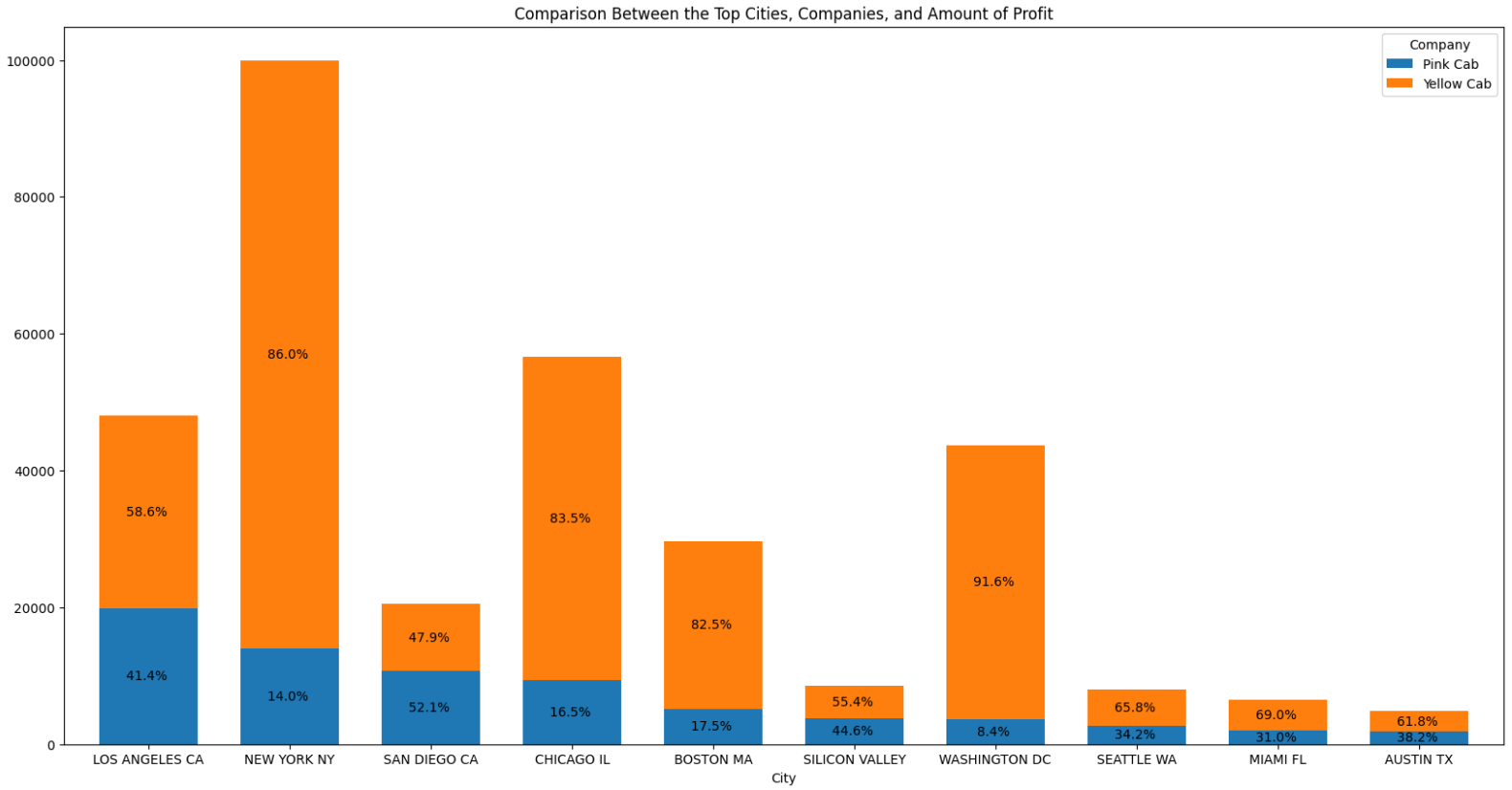
Different perspective of each city vs amount of profit



EDA: Profit - Comparison between Companies & Cities

Looking at the top 5 cities:

- We see that New York, Chicago, and Washington DC had the most profit in the yellow cab company
- Los Angeles and San Diego has an even amount of profit between the two companies



Conclusion (1/2)

After comparing both cab companies to determine which company should the client invest in, we discovered that:

- We did see a few outliers which was ignored due to said outliers potentially being the high-end/expensive cars.

- We saw that the yellow cab company has nearly three times the number of cabs when compared to the pink cab company.

- We saw the greatest amount of data between the age range of 20 to 40 years old which corresponds with our bell curve distribution theory.

  We also saw that the yellow cab company makes nearly 90% of the profit in all age groups when compared to the pink cab company.

- For income groups, we saw a skewed (to the left) bell curve distribution, thus our theory was wrong in this regard.

  Again, we saw that the yellow cab company makes nearly 90% of the profit in all age groups when compared to the pink cab company.

- We calculated the average of profit between the two companies regarding age and income and achieved 89.2% profit for the yellow cab company and 10.8% for the pink cab company.

Conclusion (2/2)

- We see an exponential increase of cab usage towards the end of the years and the first month of the year which suggest that our hypothesis that people prefer to take the cab towards the colder months of the year.

  We also see that the yellow cab company makes the majority of the profit throughout the year which is expected given our previous tests.

- Our hypothesis about a bell curve distribution for kilometers traveled was partially correct. Although not a perfect bell curve, it still resembles the shape.

  As before and expected, we see nearly 90% of the profit towards the yellow cab company.

- We were correct about more cab usage in the largest cities in America. New York, San Francisco, Chicago, Los Angeles, and Washington DC make up the majority of pink and yellow cab company usage.

- Looking at the top 5 cities, surprisingly we see roughly equal profit in Los Angeles between the two companies.

  Unsurprisingly, we see most of the profit going towards the yellow cab company for the other 4 cities.

Conclusion

Objective:

Provide actionable insights to help XYZ firm in identifying the right company for making investment.

Solution:

We heavily encourage your firm to invest in the yellow cab company as the profits of the yellow cab company are significantly higher than the pink cab company. As we were given a three-year sample, we also see that the yellow cab company has been stable and consistent with their profit throughout all the years. With that being said, our investigation leads to the yellow cab company being the winner.

Thank You