



Data Glacier

Your Deep Learning Partner

G2M Insight for
Cab Investment

3/1/2023

Agenda

Executive Summary

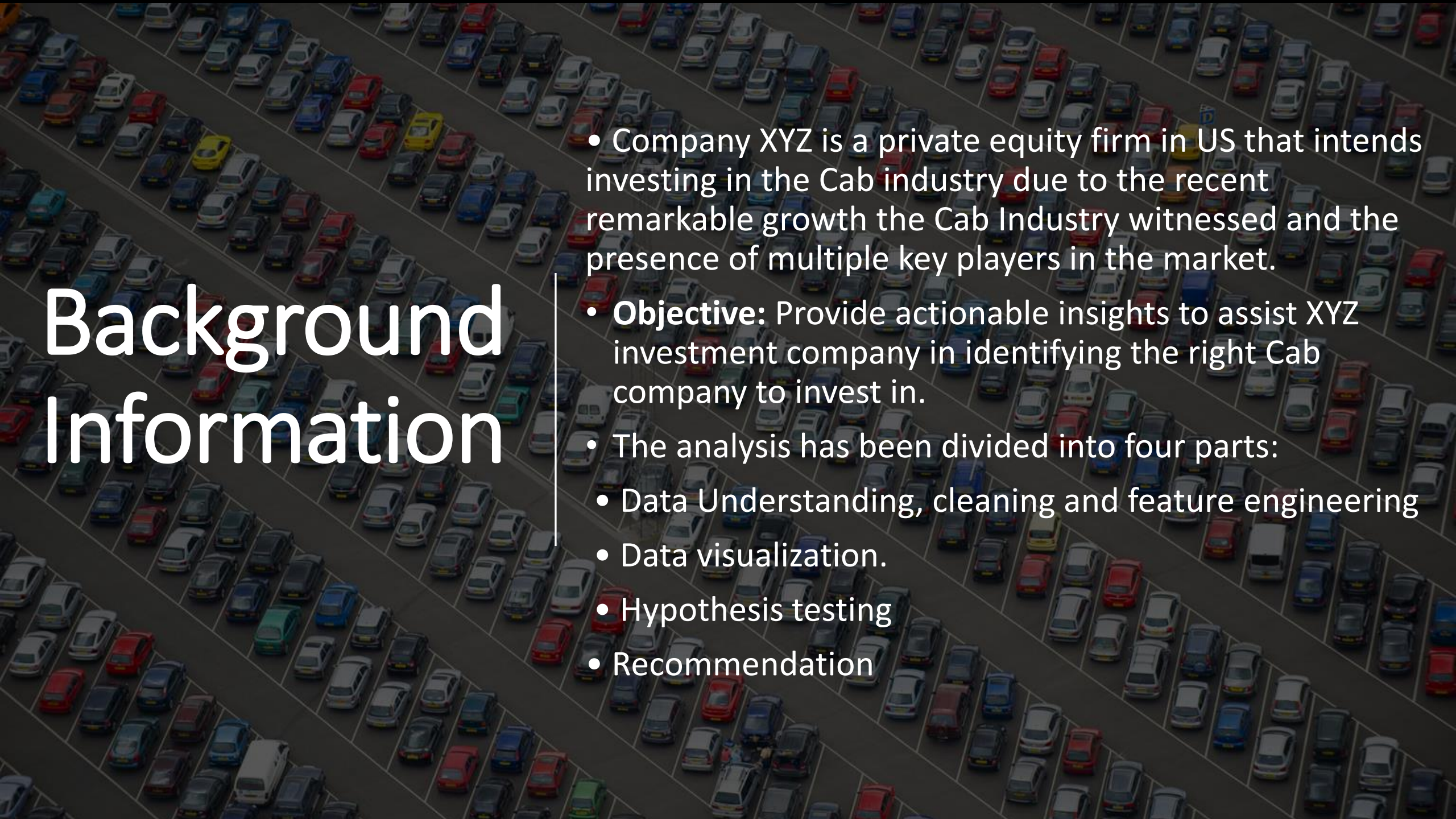
Problem Statement

Approach

EDA

EDA Summary

Recommendations




Background Information

- Company XYZ is a private equity firm in US that intends investing in the Cab industry due to the recent remarkable growth the Cab Industry witnessed and the presence of multiple key players in the market.
- **Objective:** Provide actionable insights to assist XYZ investment company in identifying the right Cab company to invest in.
- The analysis has been divided into four parts:
 - Data Understanding, cleaning and feature engineering
 - Data visualization.
 - Hypothesis testing
 - Recommendation


Data Preprocessing and Preparation

Four datasets were made available for the process:

- Cab_Data: This dataset includes information about two cab companies in the industry.
- Customer_ID: This contains demographic information about the customers with unique identifiers.
- Transaction_ID: this contains transaction information like payment mode and customer mapping
- City: This is the US cities information dataset with information like population and cab users.



Data Structure Analysis



Data date: ranges from 08/01/2016 to 02/01/2018.

Gender: Male and Female.

Payment Method: Card and Cash.

KM_Travelled: between 1.9 to 48.0.

Price Charged: between 15.6 to 2048.03.

Cost of Trip: between 19.0 to 691.2.

Age: between 18 to 65.

Income: between 2000 to 35,000.

Population: between 248968 to 8405837.

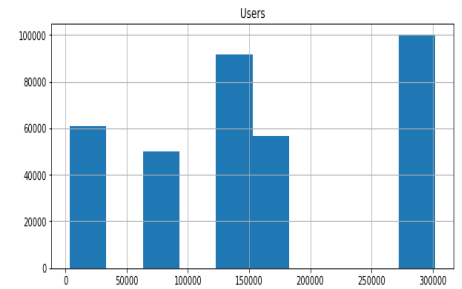
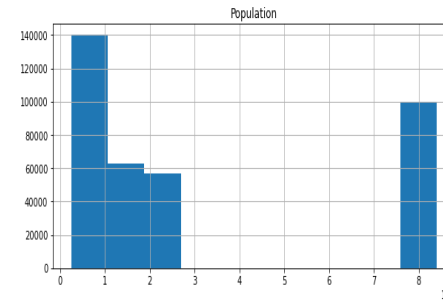
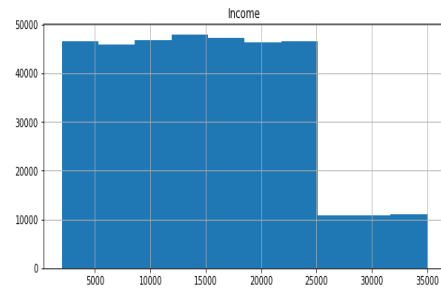
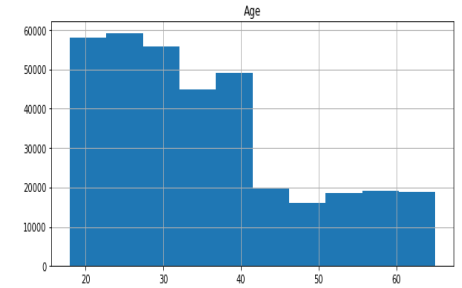
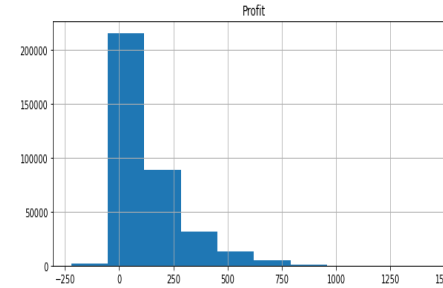
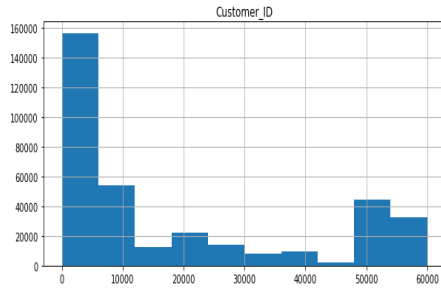
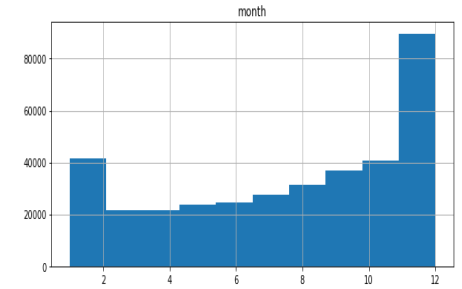
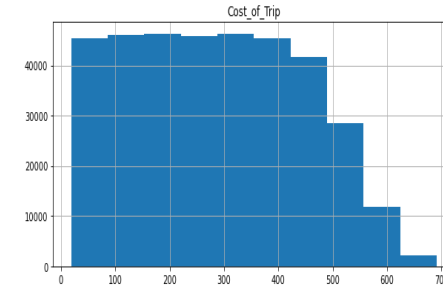
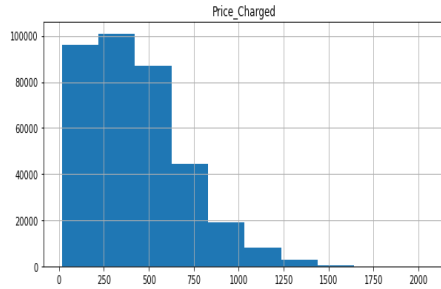
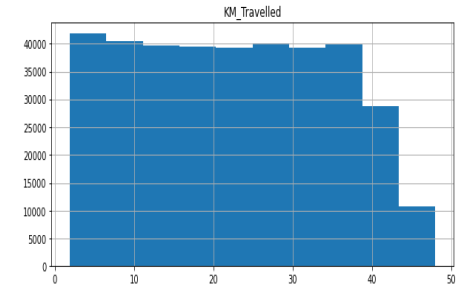
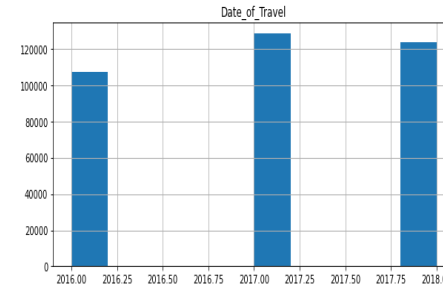
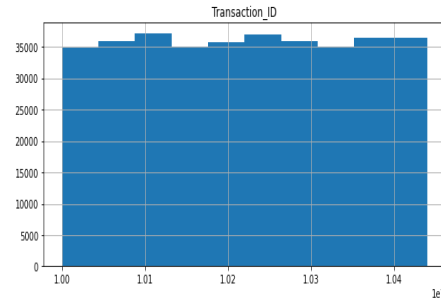
Cab Users: between 3643 to 302149.

Cab Company: Yellow Cab & Pink Cab

Exploratory Data Analysis

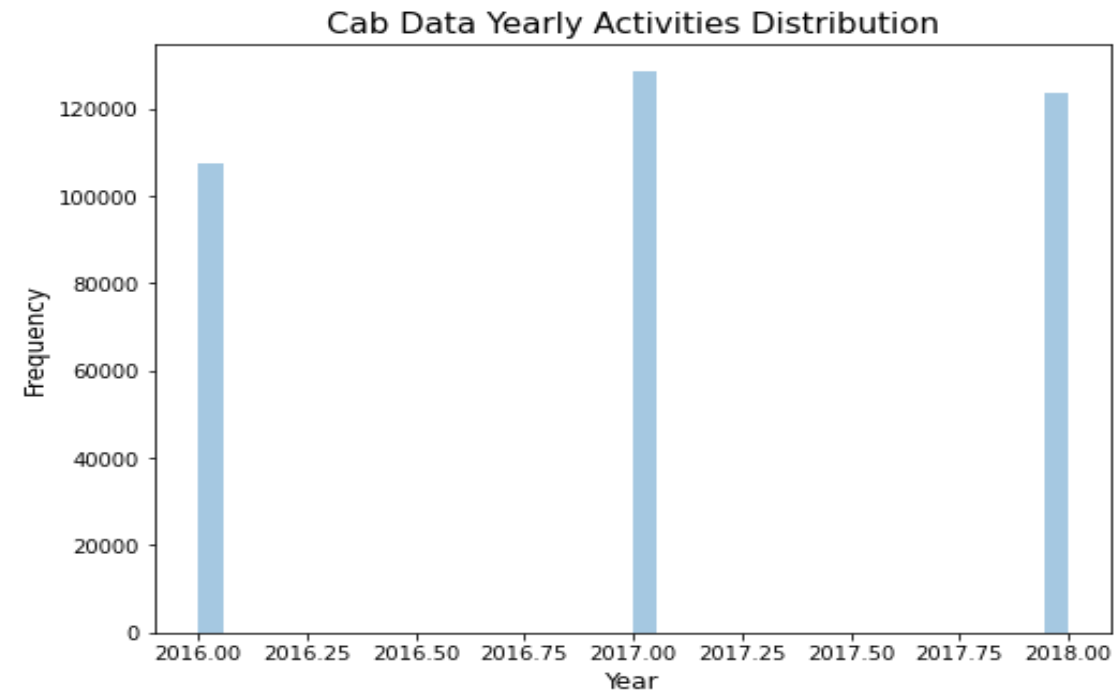
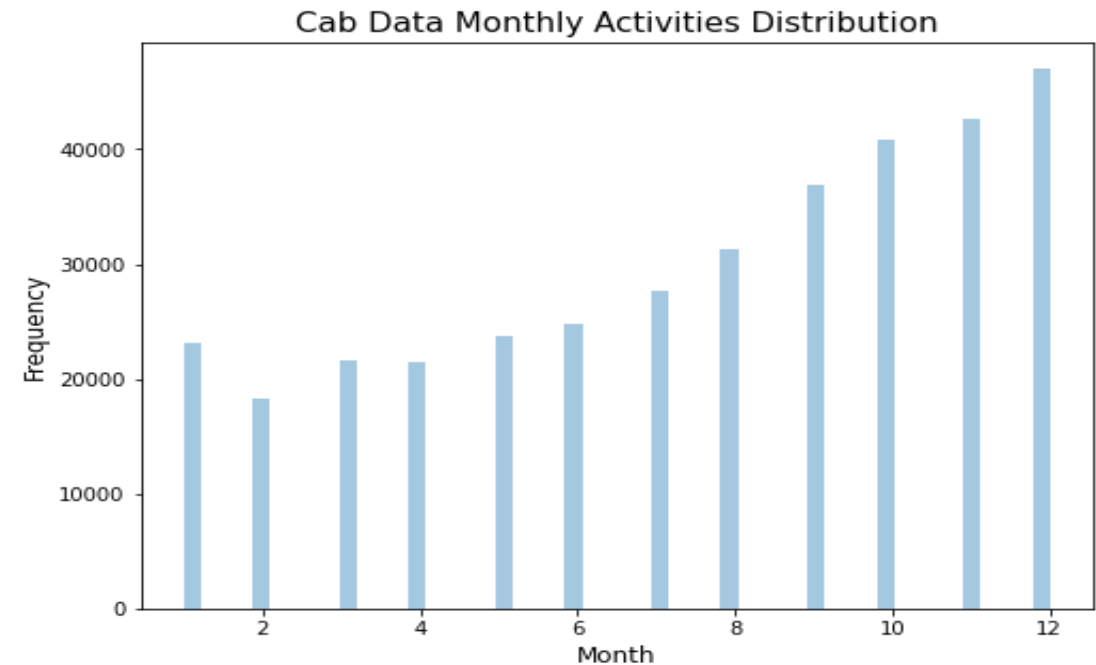
Distribution of Variables in the Complete Data

- We can see that all the numerical variables in the data are not uniformly distributed



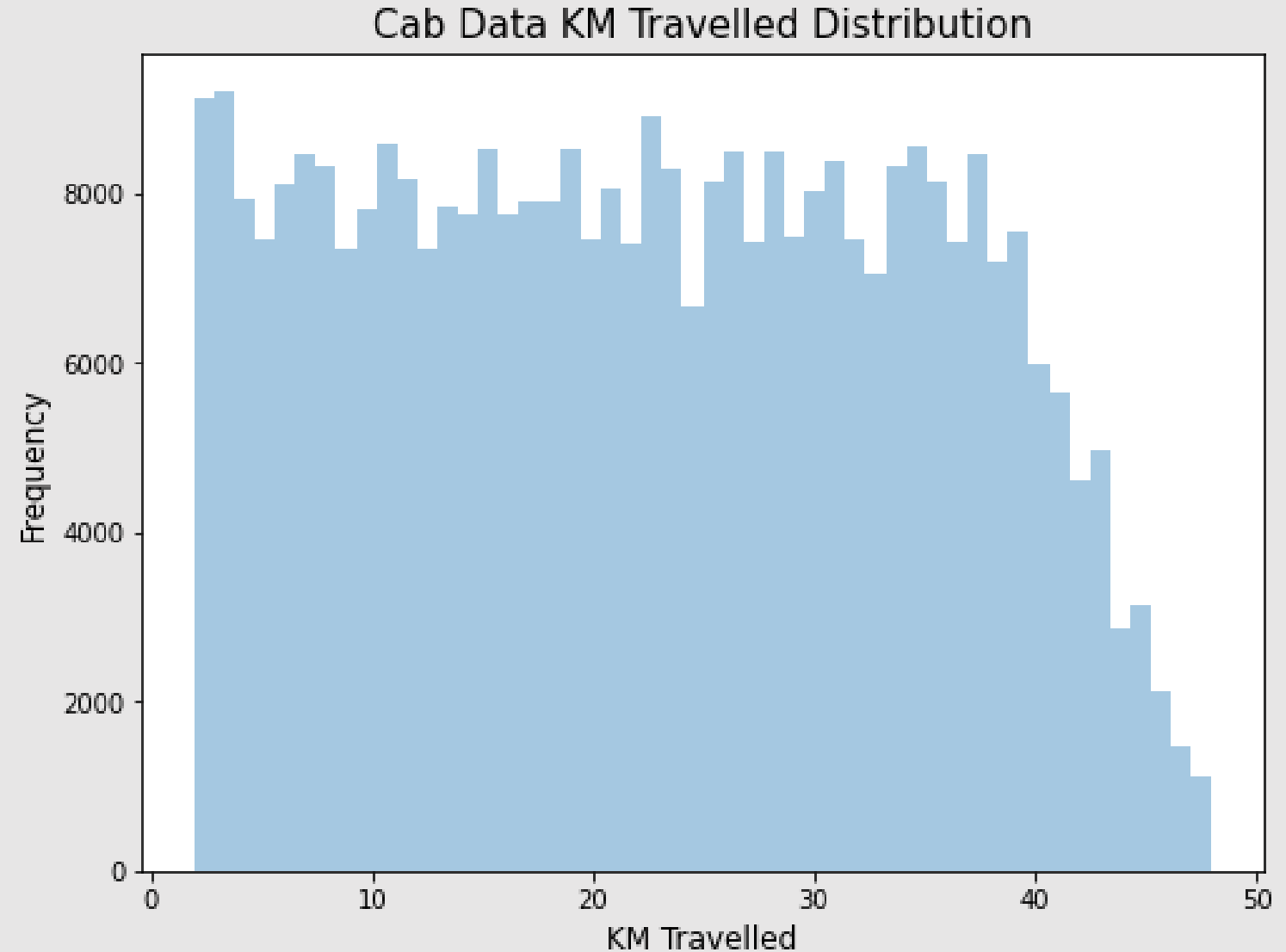
Monthly and Yearly Distribution of Both Cab Companies' Activities

- For both Cab companies, the chart illustrates that there is variation of activities across months and years. December and year 2017 recorded the highest activities.
- The average monthly chart illustrates that, aside from January, the trend of activities roughly witness ascension from February to December. The high growth in December would likely be due to Yuletide and this probably affected January positively.
- As for the yearly chart, 2016 had the lowest activities followed by 2018 and 2017 which had the highest.



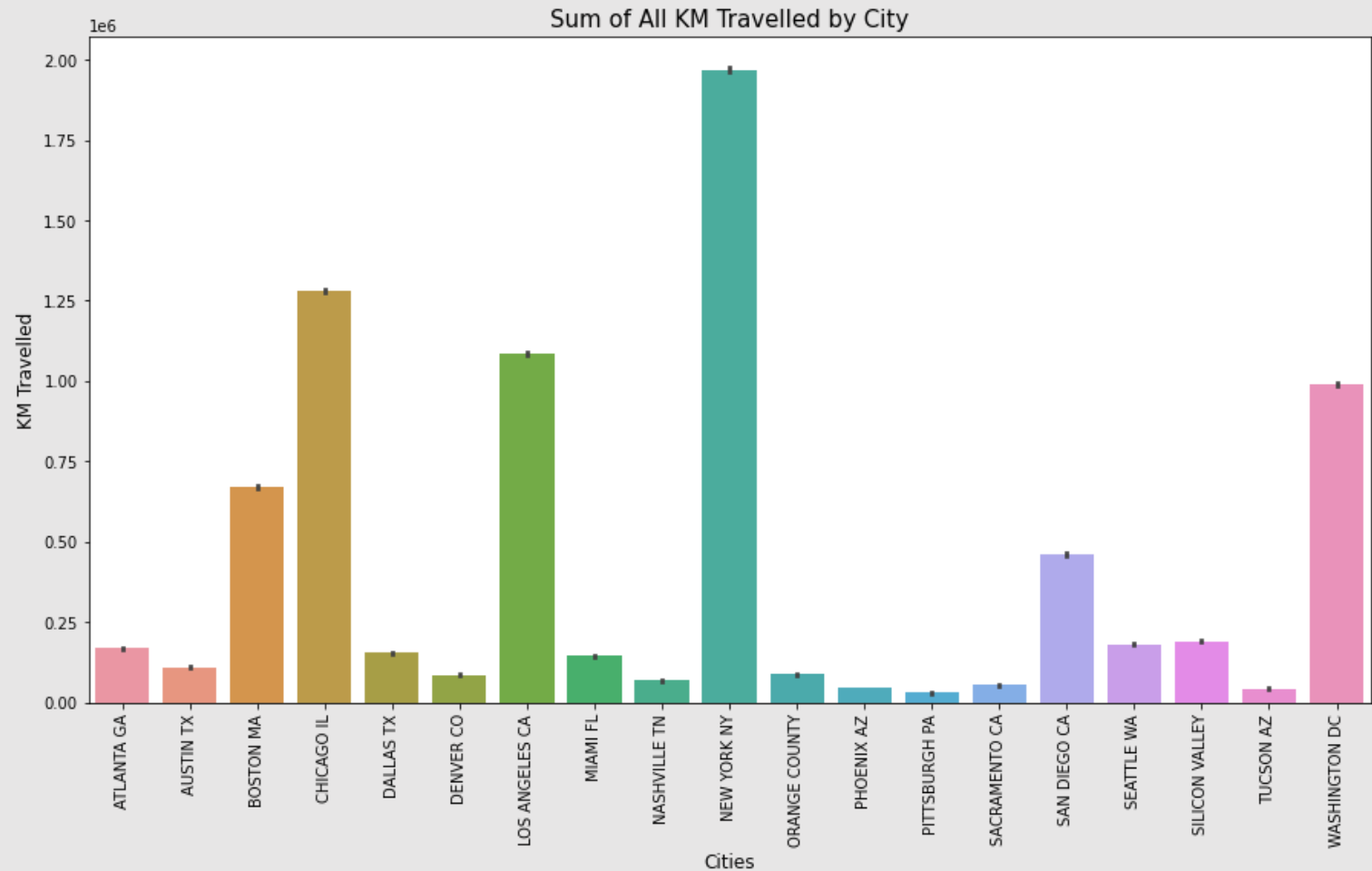
Frequency Distribution of KM Travelled by Both Cab Companies

- Lower travel distances have higher frequency. Thus, indicating that most customers hire cabs for shorter range travel.
- It shows that is more likely for customers to hire cab to travel within a city than travel across cities.



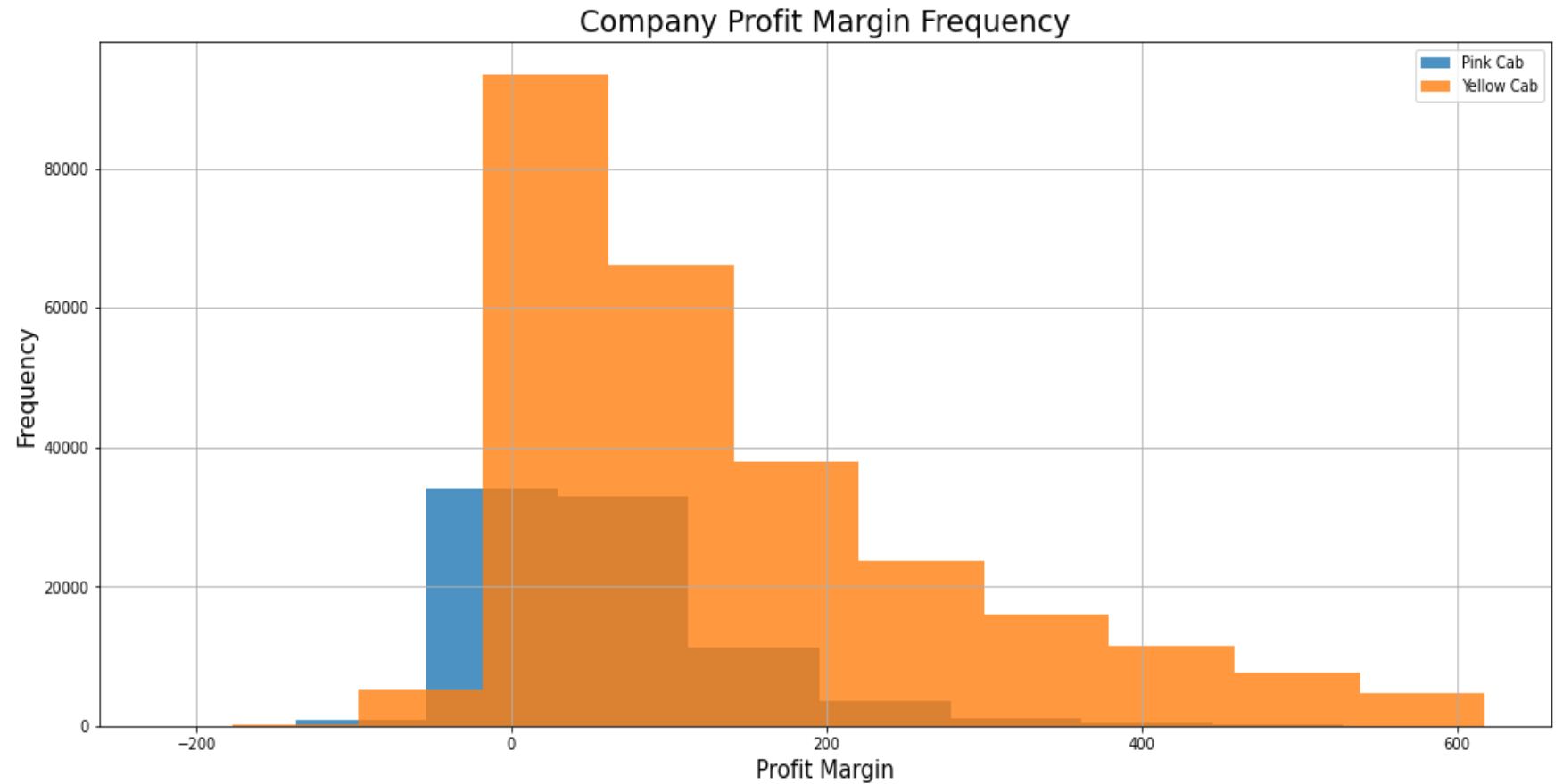
Frequency Distribution of KM Travelled by Both Cab Companies Across Cities

- Variation is also noticed in the KM travelled by both Cab companies across the different cities of USA.
- Being the largest city with mammoth population, New York city evidently had the highest travel distances while Pittsburgh, a city in Pennsylvania had the lowest. Hence the variable city may have effect on travel distance.



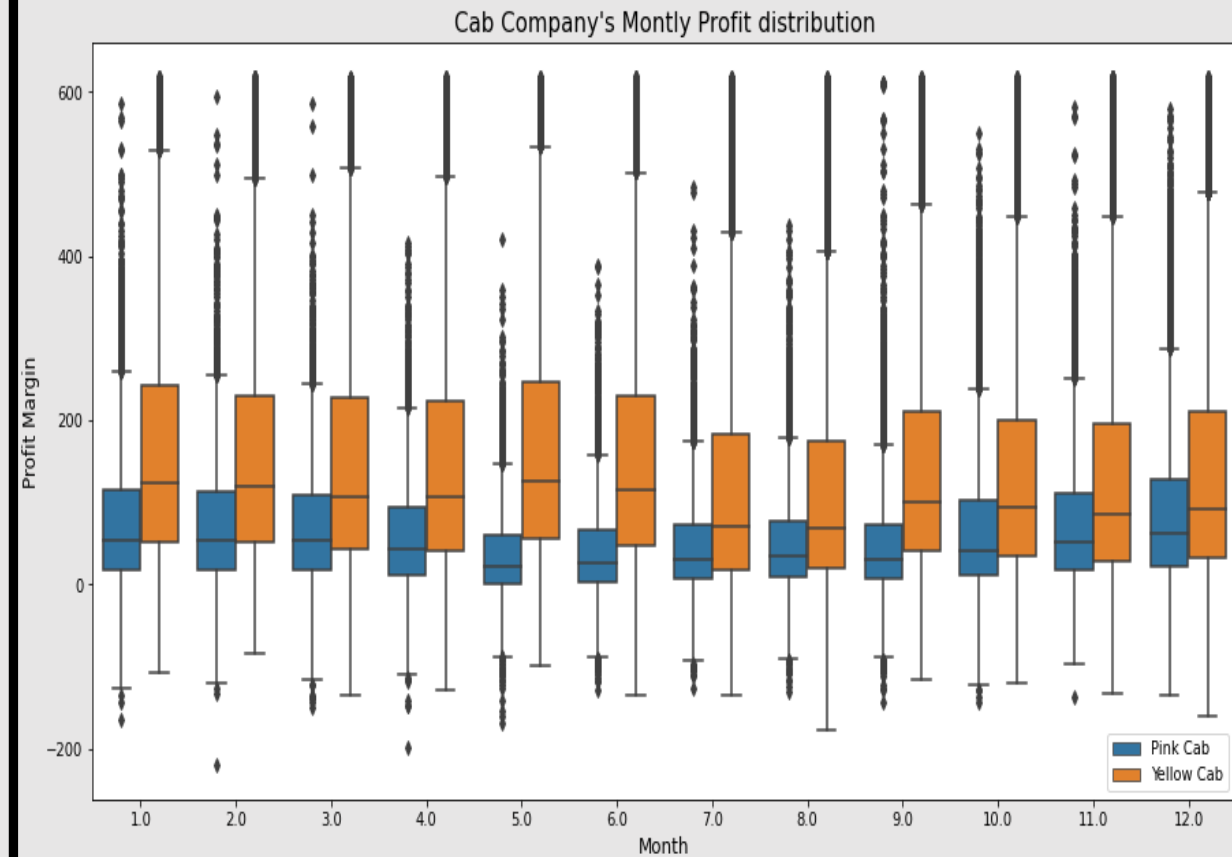
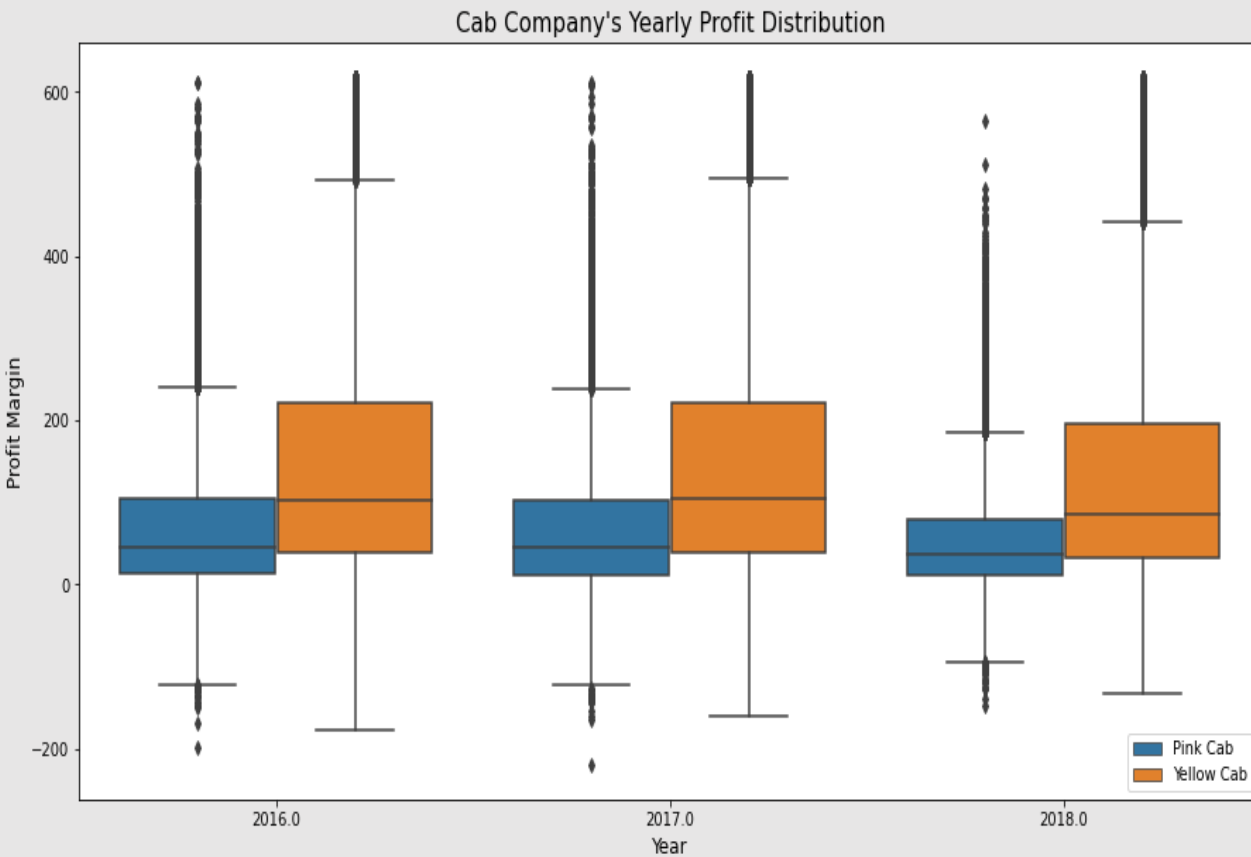
Comparing Result from Yellow and Pink Cab Company Using Exploratory Analysis

Distribution of Profit Margin by the Two Cab Companies



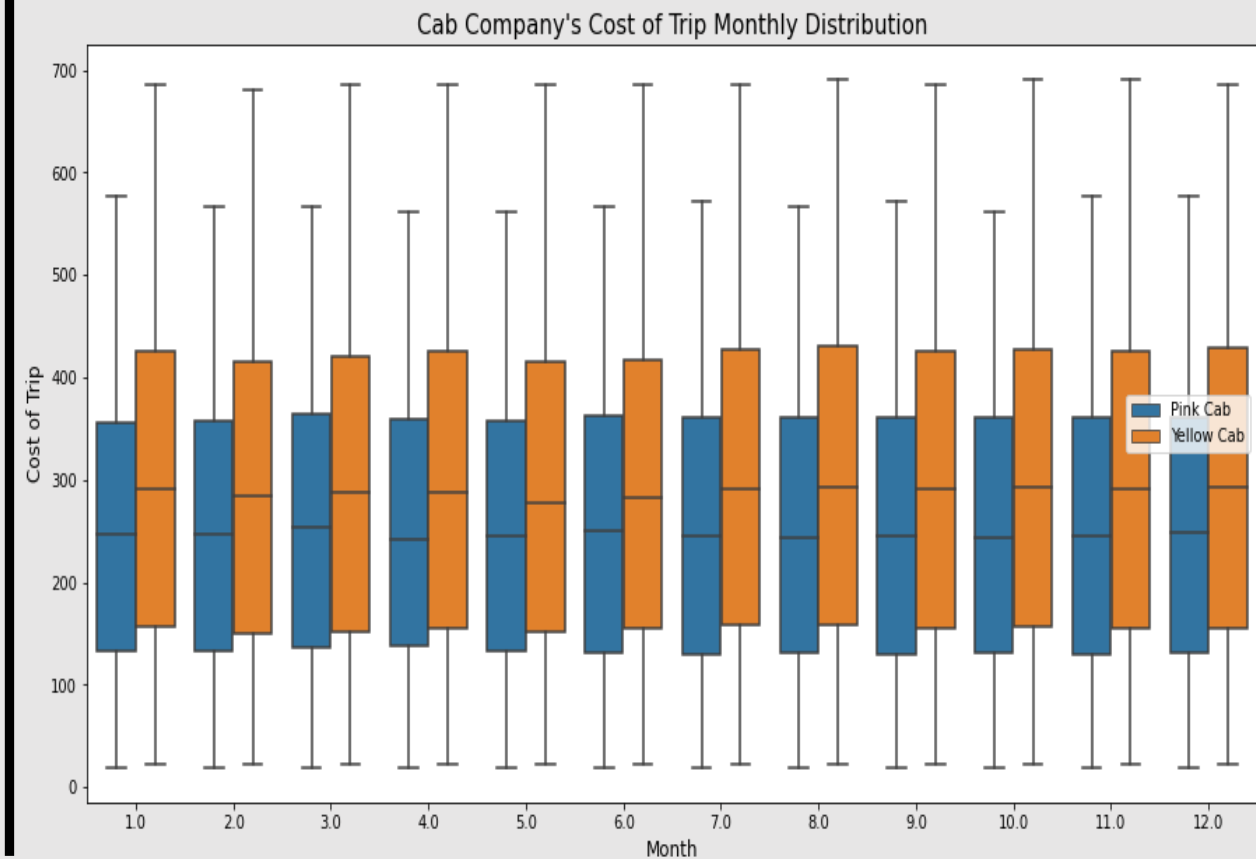
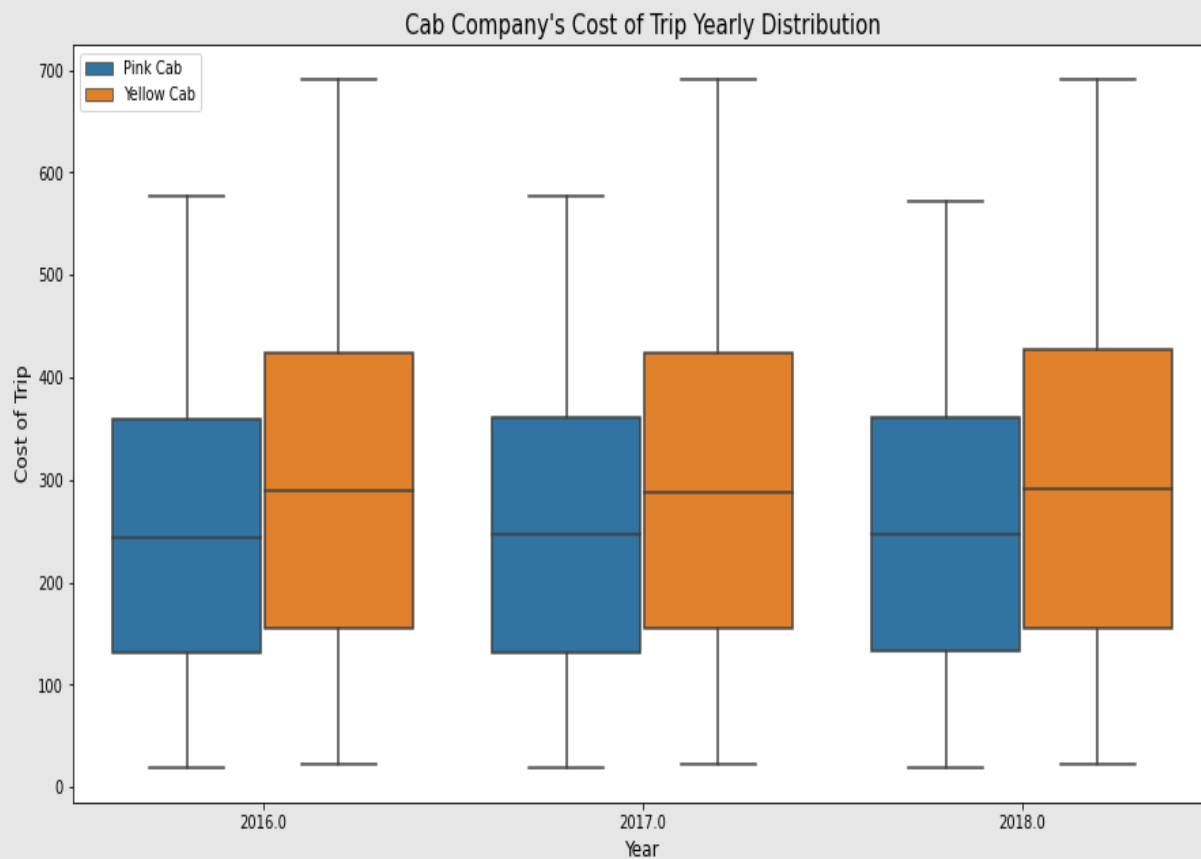
- According to the chart, the Yellow Cab company made lesser losses than the Pink Cab Company. On the other side, the Yellow Cab makes more profit than the Pink Can company.
- Profit Margin = Price Charged – Cost of Trip.

Yearly and Monthly Distribution of Cab Companies' Profit



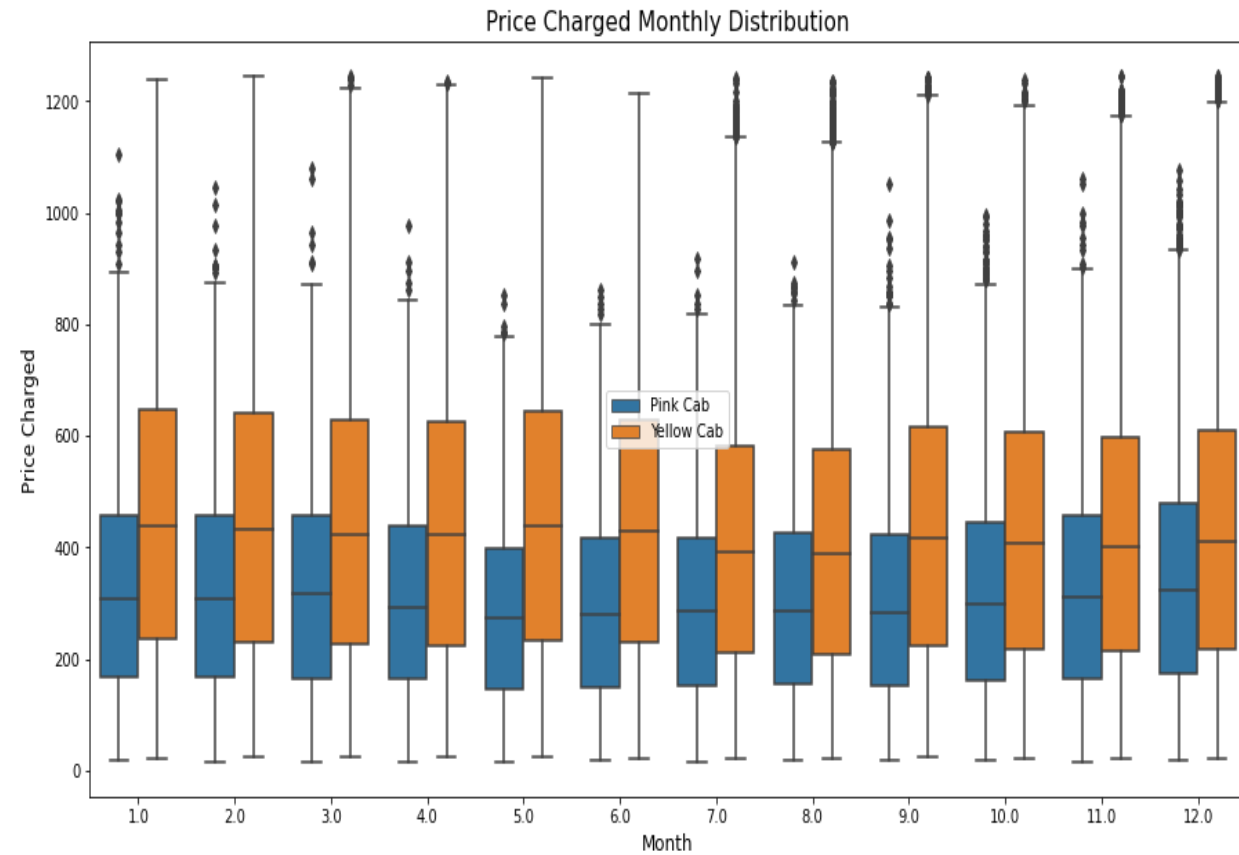
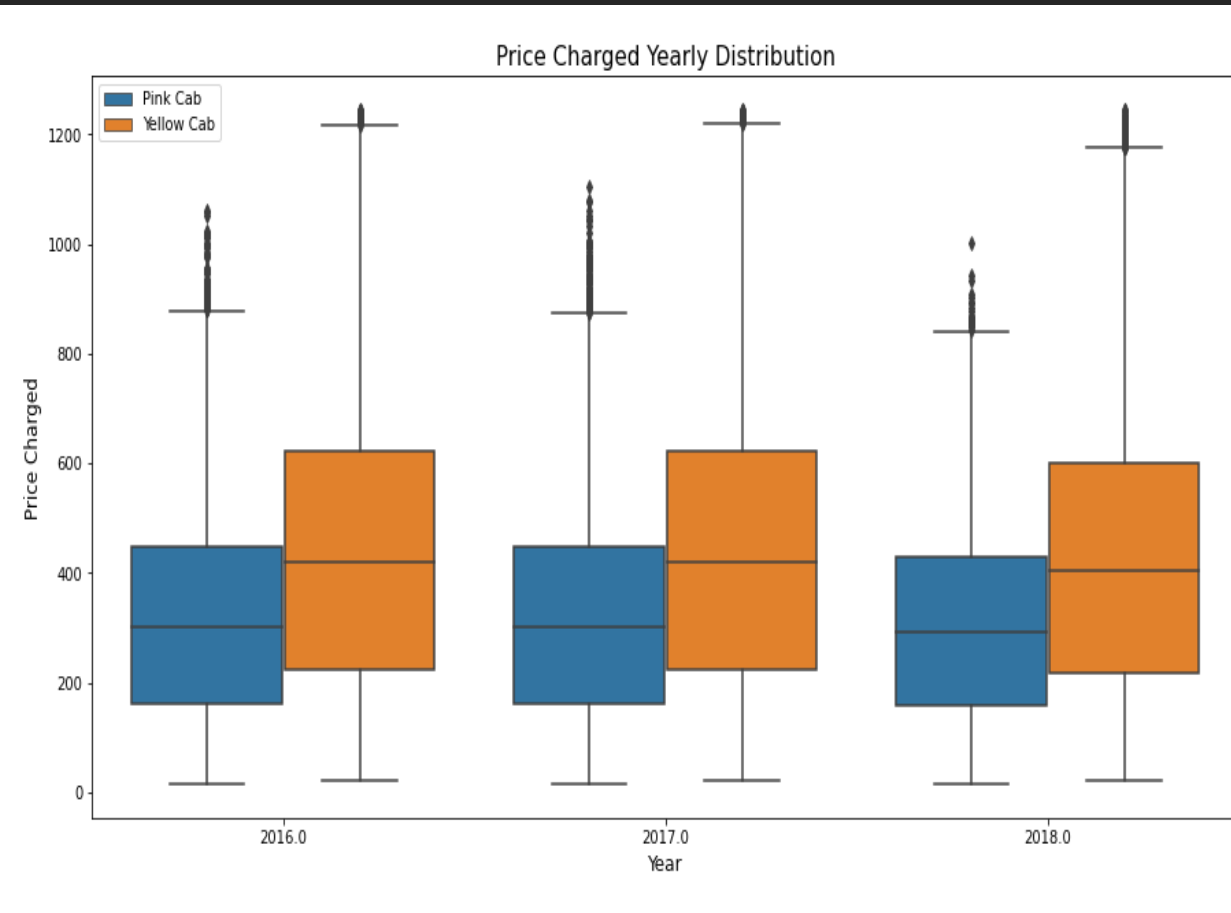
The chart shows the effect of month and years on profit variation for both companies.

Yearly and Monthly Distribution of Cab Companies' Cost of Trip



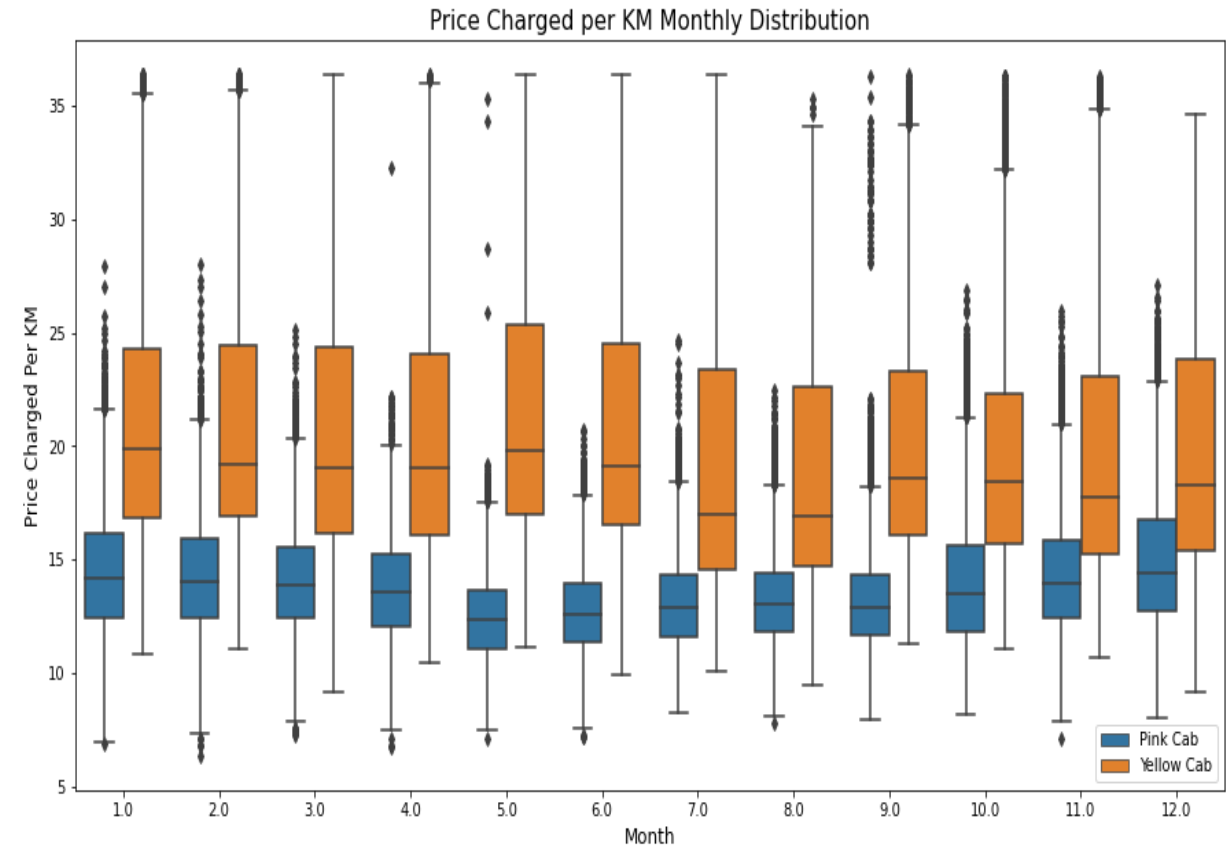
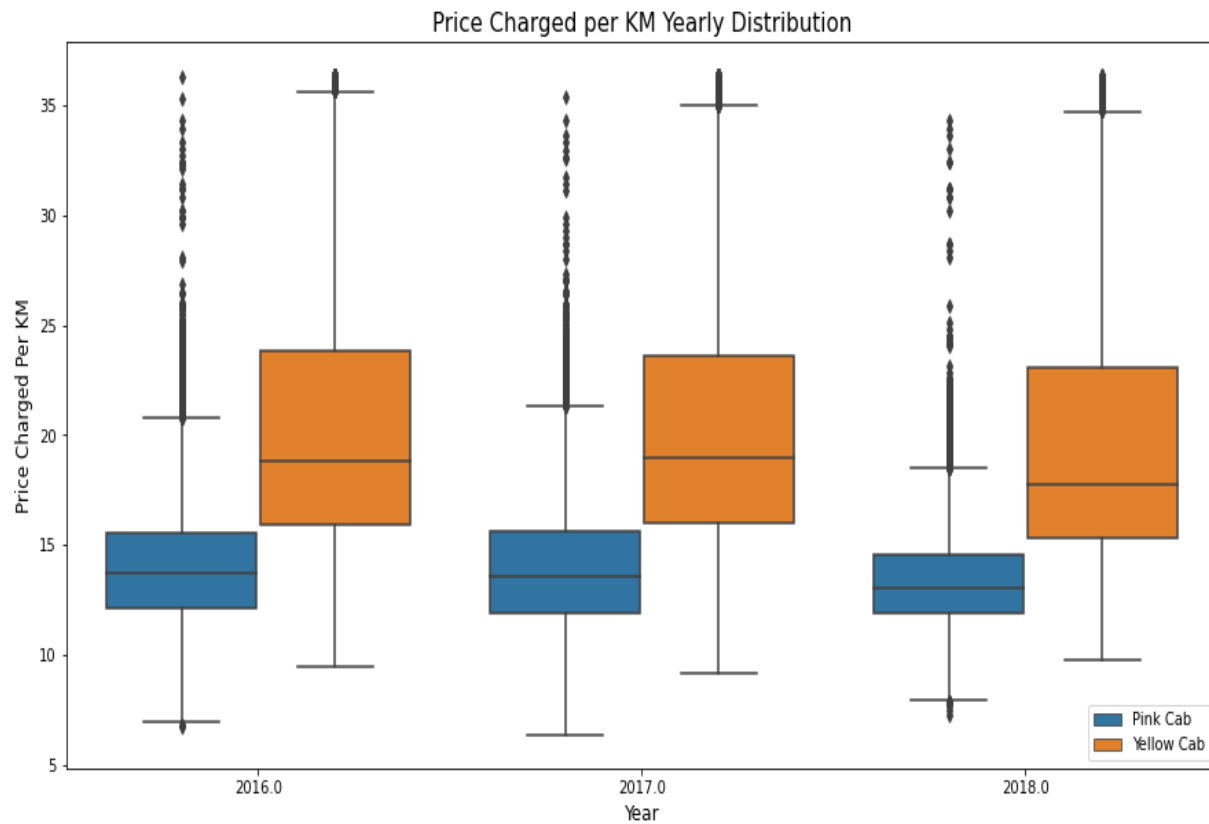
The chart shows that years and months has no effect on the cost of trip for both companies.

Yearly and Monthly Distribution of Cab Companies' Price Charged



The price charged by both companies also fluctuate as month and year varies.

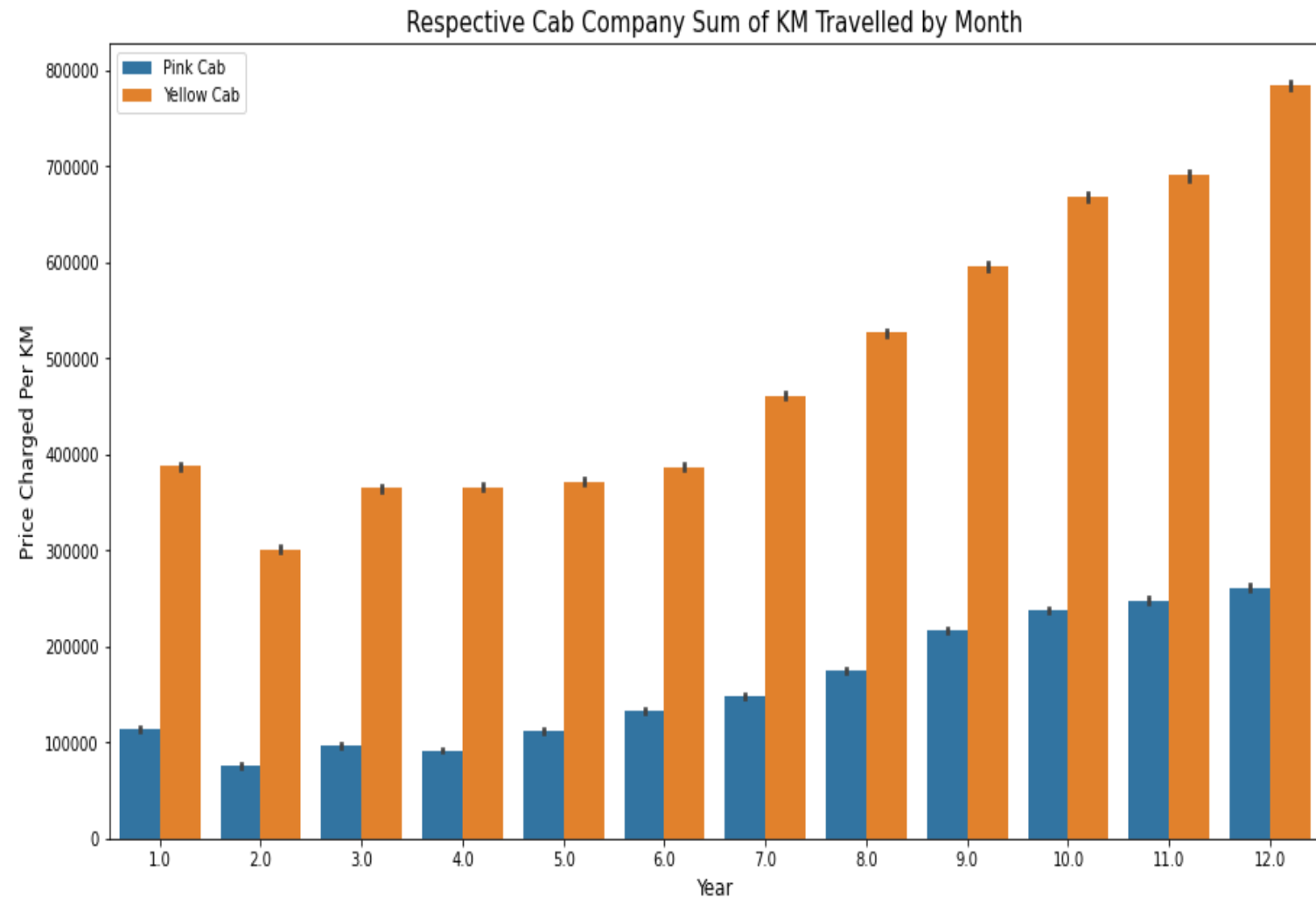
Yearly and Monthly Per KM Travel Charges by Both Companies



Per KM Trip charges by both companies is also determined by year and month time factors.

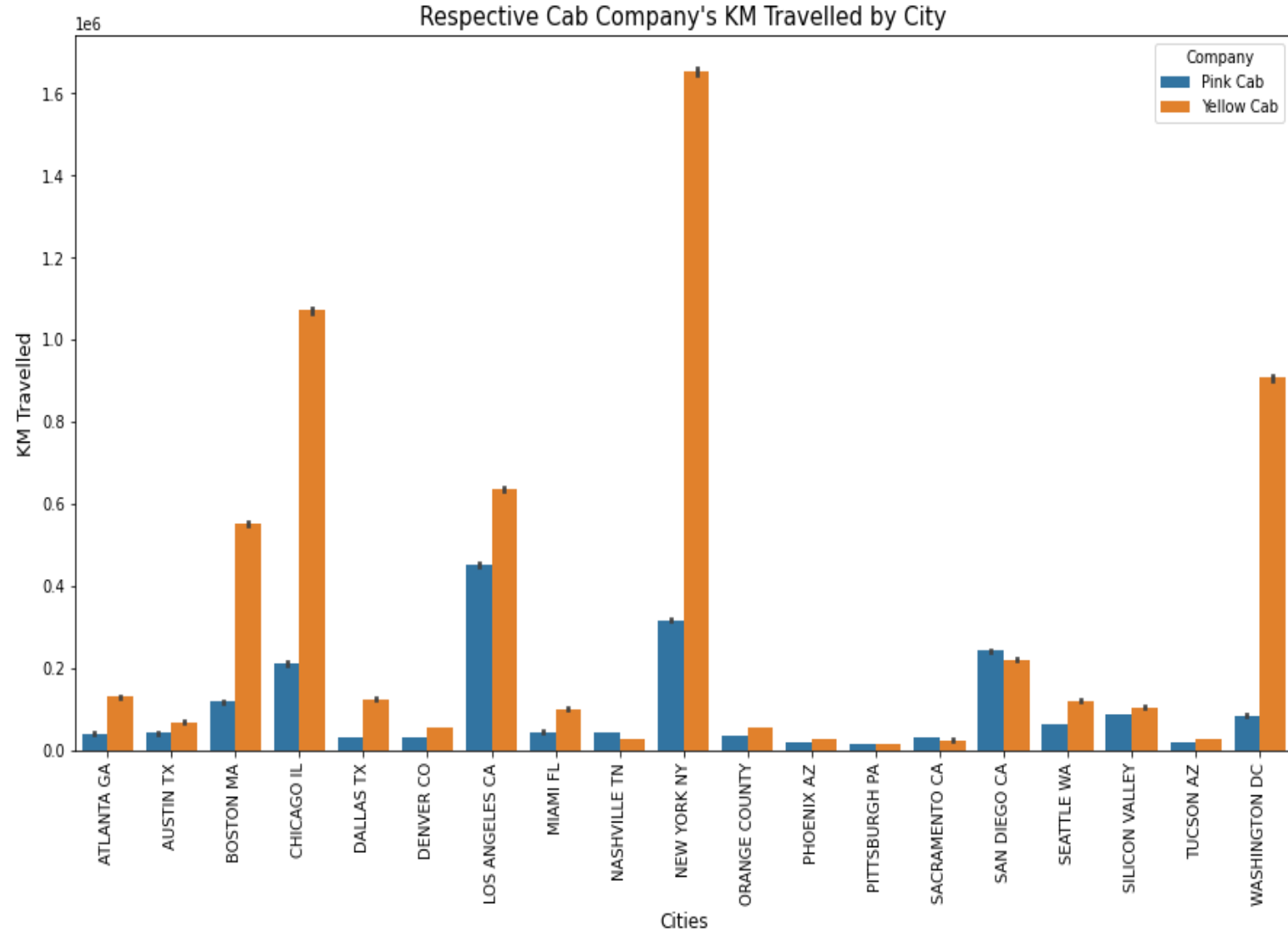
Sum of Monthly Travel Distances in KM

- Both companies witness the similar trend across most months. In March, April and May, we can see slight variation in the trend of the two companies.



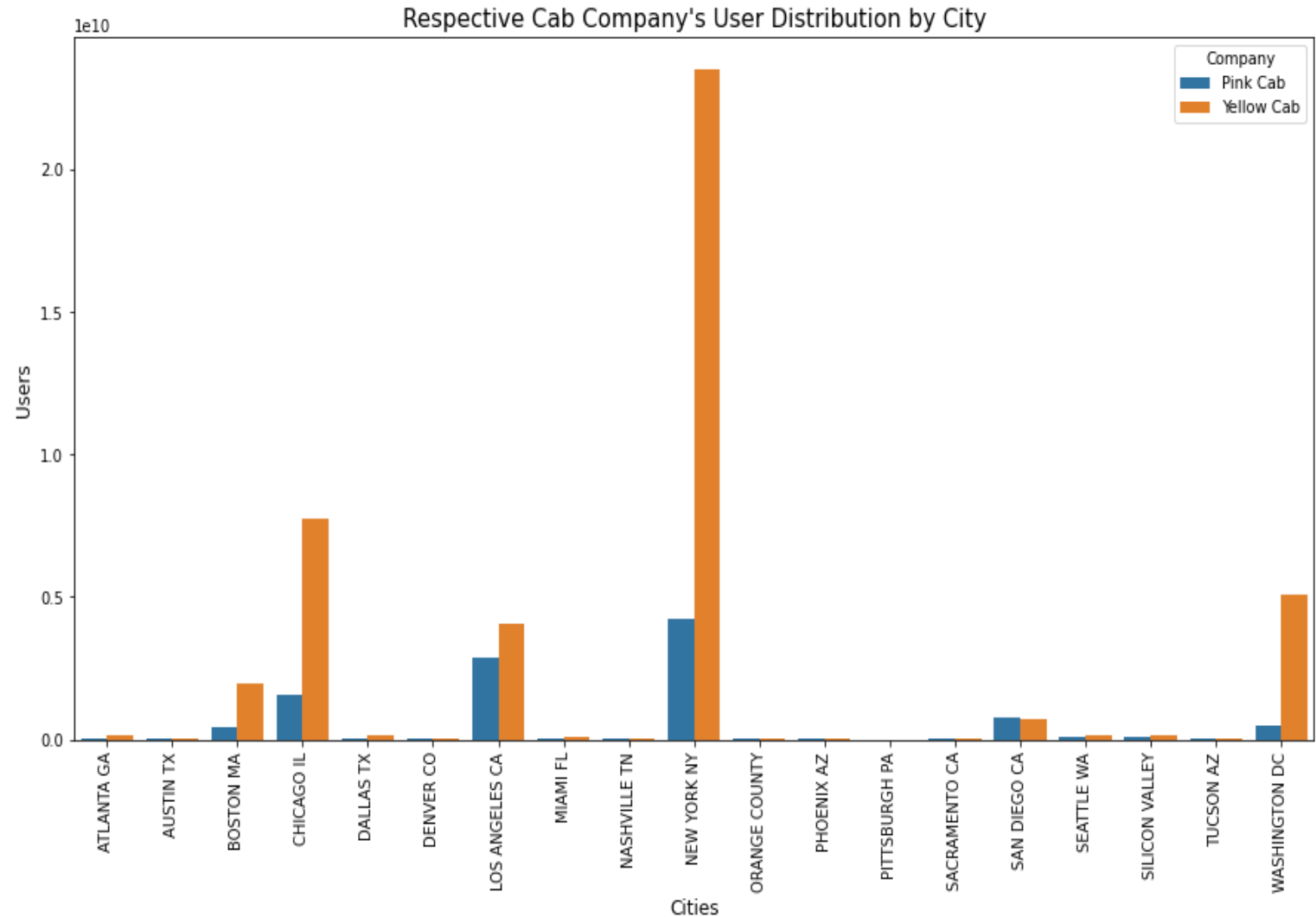
Distribution of Total KM Travelled across cities.

- As seen in the overall data for KM x Cities chart, travel distances of both companies is affected by cities. However, the trend across the cities for both companies differ. While Los Angeles represent the highest for Pink Cab, Yellow Cab maintains New York for the highest distance travelled. Both companies presented their lowest figure in Pittsburgh, with almost the same value
- Pink Cab surpassed Yellow Cab in only San Diego.



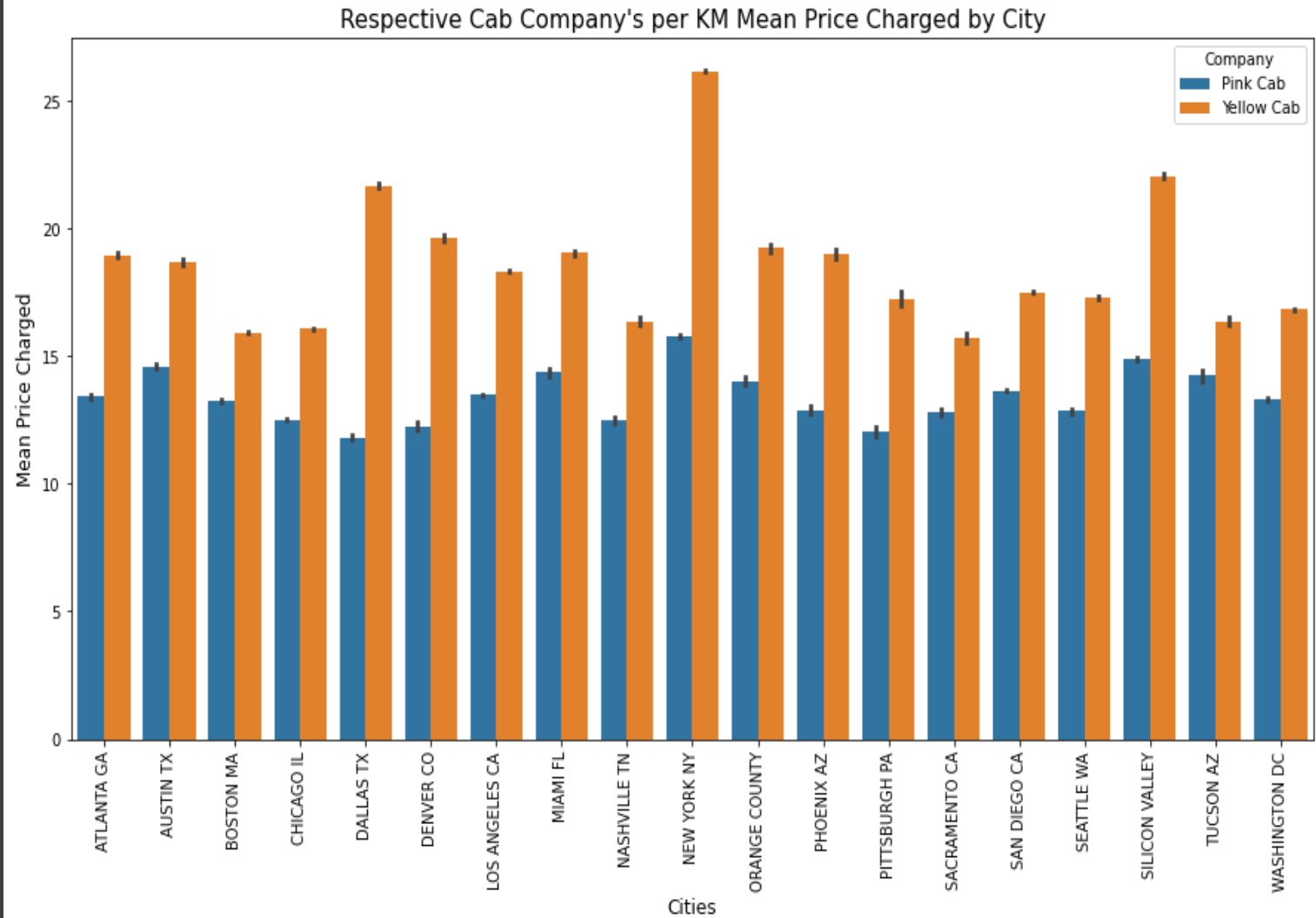
Distribution of Total Users across cities.

- As seen in the KM per City chart, the cab companies' users also vary across the cities. However, the user per city variation differs from the KM per city variation. For both companies, New York city presented the user number of users, followed by Los Angeles and Chicago.



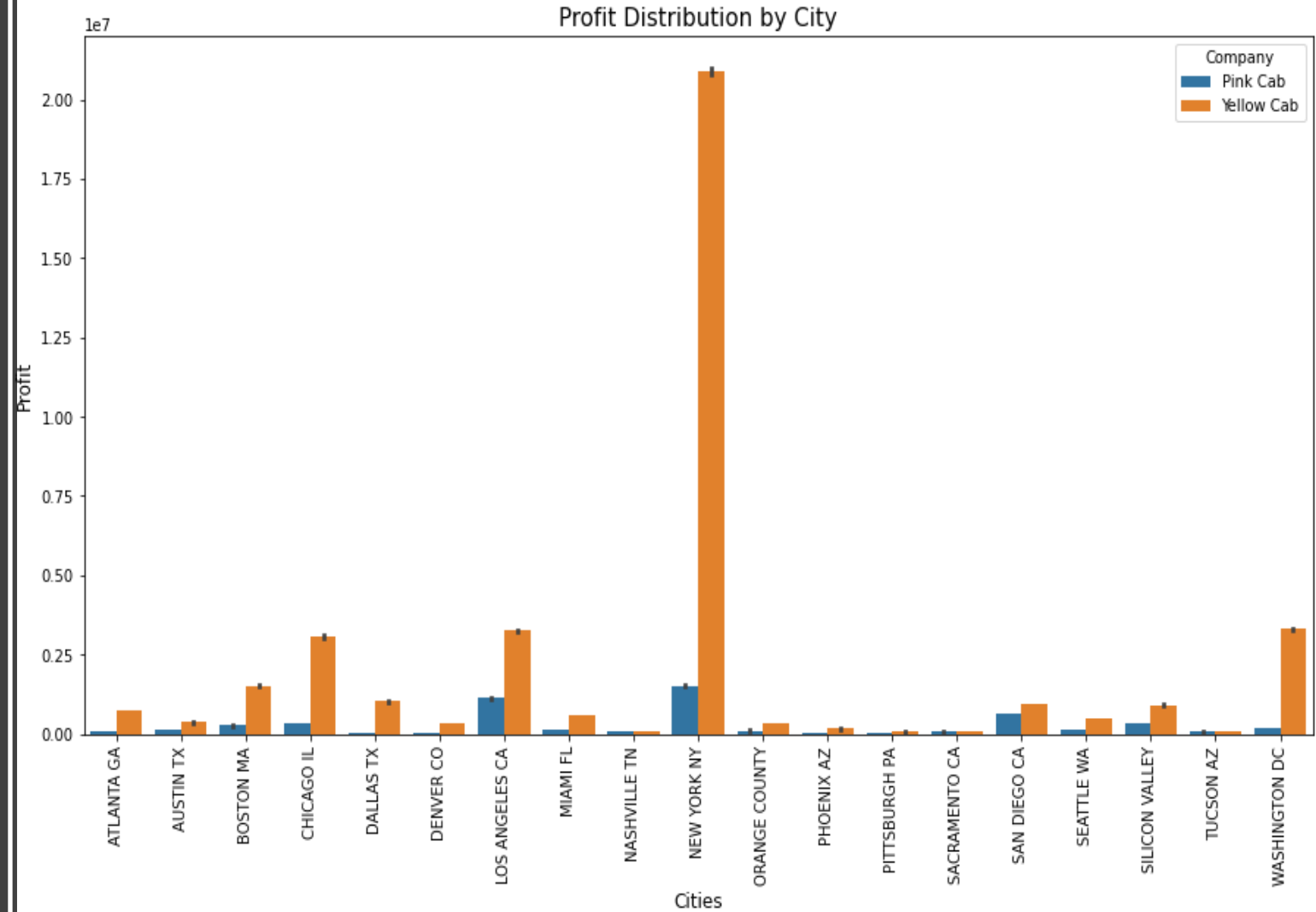
Mean Per KM Price Charged Distribution by City.

- The city variable also affected the price charged by both Cab company. This is probably due to population, standard of living and the demand for services among other factors.



Total Profit Distribution by City.

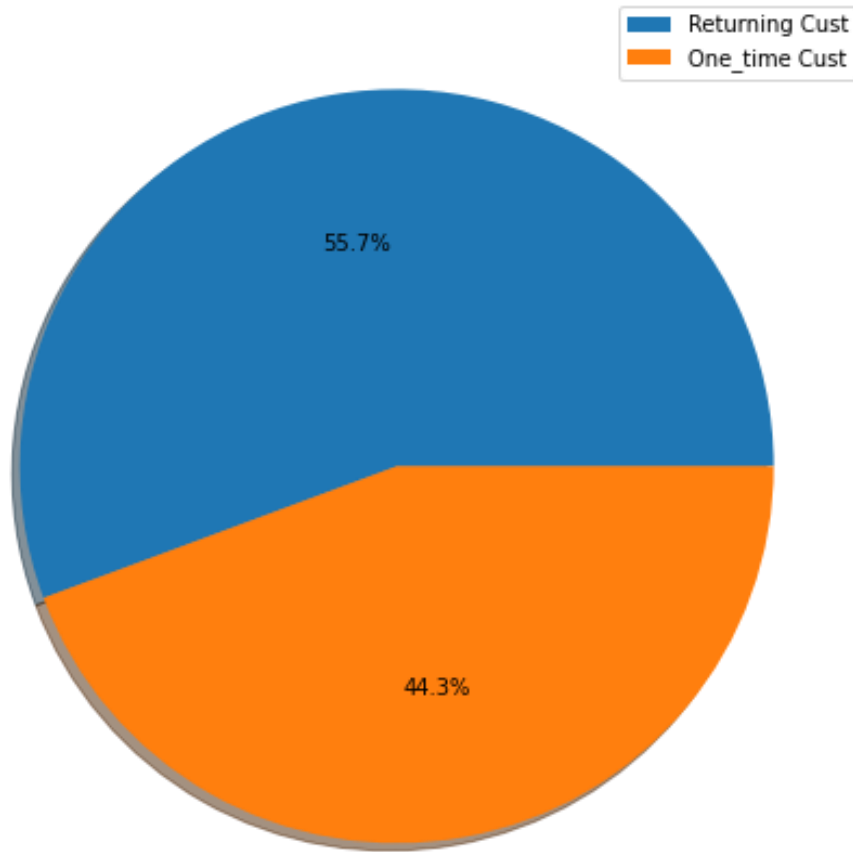
- Just as in the case of KM and User distribution per city, New York presented the highest profit for bot companies.



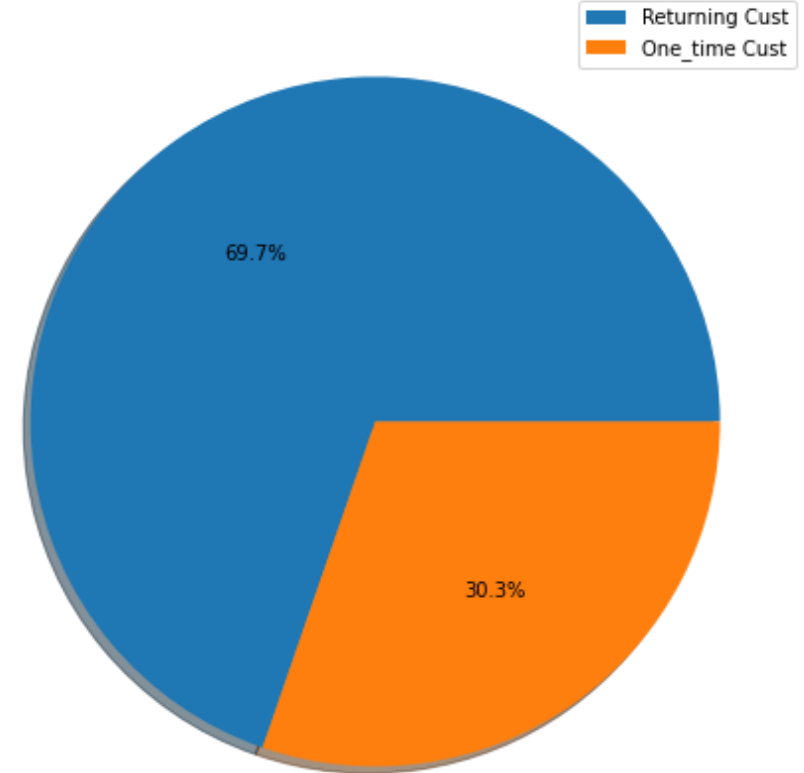
Percentage Returning and One-Time Customers

- Yellow Cab has higher returning customers, 14% more than Pink Cab Company. On the other hand, Pink Cab has higher one-time customers, 14% more than yellow Cab.

Pink Cab Returning and One-Time Customers

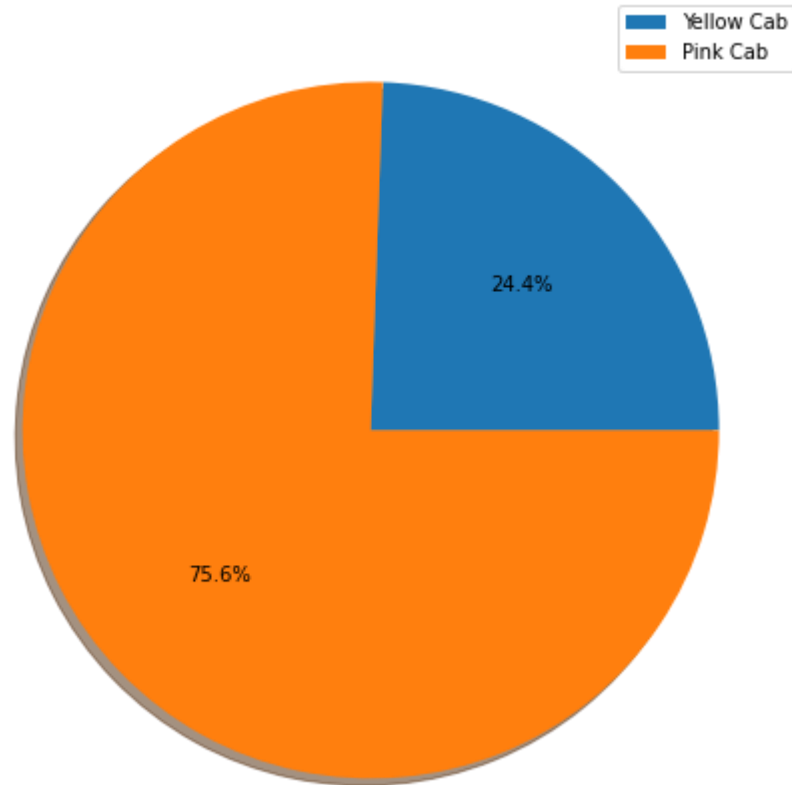


Yellow Cab Returning and One-Time Customers



Share of KM Travelled by Both Companies.

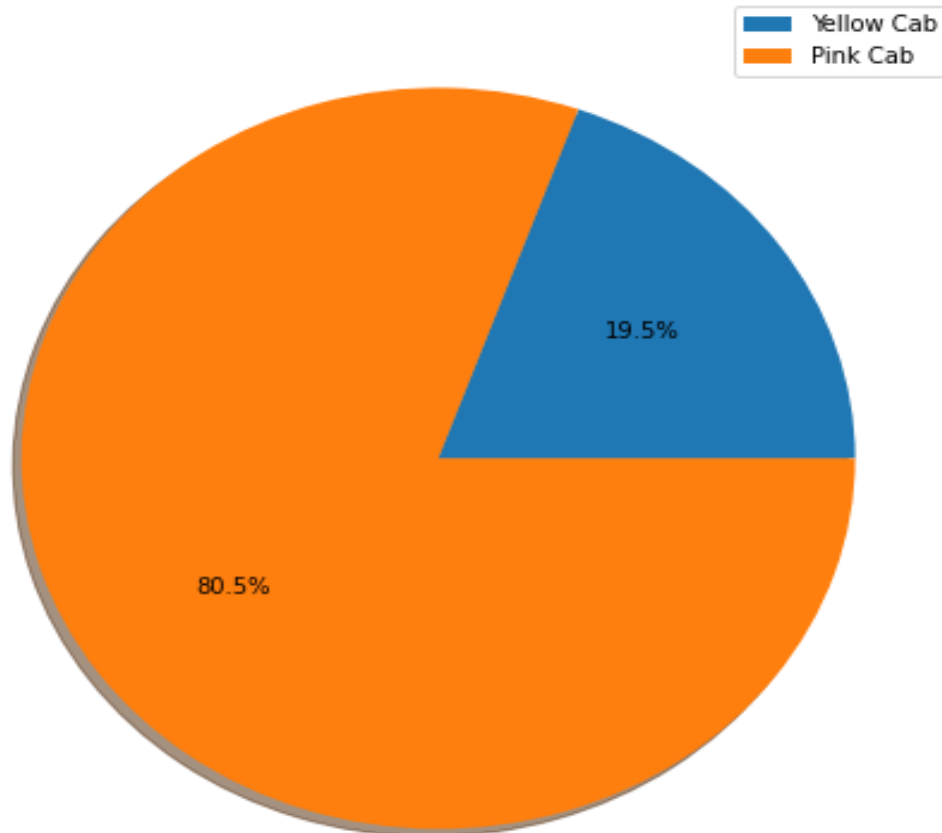
Cab Companies Share of Total KM Travelled



- The Yellow Cab company was involved in more activities than the Pink Cab company, recording 75.6% of the travelled distances (51.2%) more than Pink Cab company which account for only 24.4% of the travelled distances.

Share of User by Both Companies

Cab Companies Share of Total User



- In terms of total number of users in the cab industry as presented in data, Yellow Cab company has more user than Pink Cab Company, 80.5% to 19.5% respectively.

Result Comparison Between Yellow Cab and Pink Cab's Data

	Total Profit	Avg. Price Charged per KM	Total Users
Yellow Cab	38,171,480.	19.8	43,790,218,386
Pink Cab	5,283,231	13.7	10,637,545,967



Results and Discussion



Yellow Cab	Pink Cab
<ul style="list-style-type: none">• Yellow Cab accounted for 76% of all the travel activities in the data• Yellow Cab Price Charged ranges from \$20.7 to \$1245• Yellow Cab total profit is 44,020,373• Yellow Cab Avg. price charge per km is 19.87• Yellow Cab total users is 43,790,218,386• Yellow Cab has 69.7% returning customers	<ul style="list-style-type: none">• Pink Cab accounted for 24% of all the travel activities in the data• Pink Cab Price Charged ranges from \$15.6 to \$1105• Pink Cab total profit is 5,307,328• Pink Cab Avg. price charged per km is 13.76• Pink Cab total users is 10,637,545,967• Pink Cab has 69.7% returning customers



The different charts above illustrates that most variables in the data witnessed variation across the city and time (year and month), except from variable 'Cost of Trip'.



This is expected because even if the price charged vary, the cost it takes the driver to deliver it services should not vary across the cities and months except factors like fuel price also vary across cities and months.



Hence, this implies that groups in variable city and month would have effect on the outcome of the variables. However, using charts only, one cannot affirm that the effect of the groups within the two variables have is significant unless tested.

The slide features a solid orange background. A large white circle is centered on the page. Inside this circle, the words "Hypothesis" and "Testing" are written in a black, sans-serif font, stacked vertically. To the left of the white circle, there is a dashed yellow arc. At the bottom right of the white circle, there is a small solid blue circle.

Hypothesis Testing

Hypothesis: there is no significant difference in the effect of each group in variables 'City, Gender, Payment Mode, Age, Month on KM Travelled

Yellow Cab Company

	df	sum_sq	mean_sq	F	PR(>F)
City	18.0	9.774106e+04	5430.058777	37.237796	1.704644e-130
Gender	1.0	1.915868e+02	191.586801	1.313848	2.517002e-01
Payment_Mode	1.0	2.889139e+00	2.889139	0.019813	8.880610e-01
Age	1.0	1.428405e+01	14.284047	0.097956	7.542968e-01
Month	1.0	9.911650e+03	9911.650454	67.971275	1.666375e-16
Residual	266064.0	3.879776e+07	145.821165	NaN	NaN

Pink Cab Company

	df	sum_sq	mean_sq	F	PR(>F)
City	18.0	2.842911e+03	157.939507	1.056014	0.391336
Gender	1.0	5.158323e+01	51.583230	0.344895	0.557019
Payment_Mode	1.0	1.238437e+02	123.843709	0.828043	0.362842
Age	1.0	2.187628e+02	218.762752	1.462690	0.226506
Month	1.0	6.411699e+01	64.116992	0.428698	0.512630
Residual	84657.0	1.266147e+07	149.561971	NaN	NaN

For variables Gender and Payment mode, we fail to reject the null hypothesis that there is significant difference in the effect each group in the variables on KM Travelled for both companies. While we reject the null hypothesis for variable City in Yellow Cab, indicating that each group in the variable has significant effect on the outcome of KM travelled, each group in variable City shows no significant effect on the outcome of KM in Pink Cab.

Hypothesis: there is no significant difference in the effect of each group in variables 'City, Gender, Payment Mode, Age, Month on Profit

Yellow Cab Company

	df	sum_sq	mean_sq	F	PR(>F)
City	18.0	9.774106e+04	5430.058777	37.237796	1.704644e-130
Gender	1.0	1.915868e+02	191.586801	1.313848	2.517002e-01
Payment_Mode	1.0	2.889139e+00	2.889139	0.019813	8.880610e-01
Age	1.0	1.428405e+01	14.284047	0.097956	7.542968e-01
Month	1.0	9.911650e+03	9911.650454	67.971275	1.666375e-16
Residual	266064.0	3.879776e+07	145.821165	NaN	NaN

Pink Cab Company

	df	sum_sq	mean_sq	F	PR(>F)
City	18.0	2.842911e+03	157.939507	1.056014	0.391336
Gender	1.0	5.158323e+01	51.583230	0.344895	0.557019
Payment_Mode	1.0	1.238437e+02	123.843709	0.828043	0.362842
Age	1.0	2.187628e+02	218.762752	1.462690	0.226506
Month	1.0	6.411699e+01	64.116992	0.428698	0.512630
Residual	84657.0	1.266147e+07	149.561971	NaN	NaN

For the variables Gender and Payment mode, we fail to reject the null hypothesis that there is significant difference in the effect each group on the variables on Profit made by the two companies. However, there is variation in the effect of the groups in variable City of the profit made. Hence, profits made by both companies will vary across the cities.

Hypothesis: there is no significant difference in the effect of each group in variables 'City, Gender, Payment Mode, Age, Month on Users

Yellow Cab Company

	df	sum_sq	mean_sq	F	PR(>F)
City	18.0	9.774106e+04	5430.058777	37.237796	1.704644e-130
Gender	1.0	1.915868e+02	191.586801	1.313848	2.517002e-01
Payment_Mode	1.0	2.889139e+00	2.889139	0.019813	8.880610e-01
Age	1.0	1.428405e+01	14.284047	0.097956	7.542968e-01
Month	1.0	9.911650e+03	9911.650454	67.971275	1.666375e-16
Residual	266064.0	3.879776e+07	145.821165	NaN	NaN

Pink Cab Company

	df	sum_sq	mean_sq	F	PR(>F)
City	18.0	2.842911e+03	157.939507	1.056014	0.391336
Gender	1.0	5.158323e+01	51.583230	0.344895	0.557019
Payment_Mode	1.0	1.238437e+02	123.843709	0.828043	0.362842
Age	1.0	2.187628e+02	218.762752	1.462690	0.226506
Month	1.0	6.411699e+01	64.116992	0.428698	0.512630
Residual	84657.0	1.266147e+07	149.561971	NaN	NaN

For the variables Gender and Payment mode, we fail to reject the null hypothesis that there is significant difference in the effect each group on the variables on Profit made by the two companies. However, there is variation in the effect of the groups in variable City of the profit made. Hence, profits made by both companies will vary across the cities.



Recommendation

Recommendation

- Customer Reach: The Yellow cab company has higher customer reach than Pink cab with over 80% share of the market. It also has 14% more returning customers than Pink cab company, indicating that its customers are more satisfied. It is also evident from the charts that Yellow cab company has more users than Pink in most of the city, indicating their spatial reach.
- Travel Distance: Yellow cab company accounts for over 75% of the travel distance covered by the two companies. According to the hypothesis tested, distance travelled by the Yellow cab company is significantly influenced by the different cities.
- It was observed that the Yellow Cab company charges more and makes more profit per KM travelled than the Pink Cab Company
- Yellow Cab company travelled more distance than the Pink Cab company and the former accounted for more activities than the latter in the data.
- Based on the prediction made, both companies will witness loss in profit. However, the Yellow cab company will get lesser loss (0.002%) compared to the Pink Cab company (-0.0052).
- On the basis of above points, I can recommend that if invested into the Yellow cab company, it would likely yield more profit than the Pink Cab company, hence it would be worth it to invest in the Yellow Cab company over Pink Cab.

Thank You