G2M Case Study – Cab Investment Firm

Data Glacier Virtual Internship

17th February, 2024

Data Glacier

Your Deep Learning Partner

Agenda

- Project Brief
- Data Overview
- Data Exploration
- Data Cleaning and Feature Engineering
- Statistical Analysis
- Data Visualization
- Recommendations

Project Brief – G2M Insights And Recommendations for Cab Investment Firm, XYZ

- Due to remarkable growth of the US Cab Industry in the last few years an investment firm, XYZ in the US, is interested in investing in the industry .
- Objective - Make actionable insights and recommendations that would assist firm XYZ in identifying the right cab company to invest in.
- We are going to be analyzing two cab companies, Pink Cab Company and Yellow Cab Company.

This is the approach we are going to follow to conduct our analyses and provide actionable insights:

- Data Overview
- Data Exploration
- Data Cleaning and Feature Engineering
- Statistical Analysis
- Data Visualizations
- Recommendations

SHALL WE BEGIN…..

Data Overview

We used 4 different sets for this project. These datasets are:

- Cab_data.csv - This dataset includes transaction details for 2 cab companies.
- Customer_ID .csv- This is a mapping table that contains a unique identifier which links the customer's demographic details.
- Transaction_ID.csv - this is a mapping table that contains transaction to customer mapping and payment mode.
- City.csv - this dataset is a list of US cities, their population and number of cab users.

Note: All monetary values are in USD

Data Exploration

After exploring the datasets, this is what I observed:

- The Cab_data.csv dataset has 7 columns and 358,547 rows. This dataset also has 3 missing values.
- The Customer_ID.csv dataset has 4 columns and 49,171 rows with no missing values.
- The Transaction_ID.csv dataset has 440,098 rows and 3 columns with no missing values.
- The City.csv dataset has 20 rows and 3 columns with no missing values.
- I also noticed the 'date of travel' column was in integer format instead of datetime format. This would affect our analysis and must be corrected.
- There was an entry in the City column of our Cab_Data.csv dataset called 'NEW'. This entry is not a city in the US.
- The 'Users' and 'Population' columns of the City.csv dataset are in string format instead of being in integer format.

We would have to do some data cleaning and feature engineering on our data to rectify some of the problems that were observed during the data exploration step.

Data Cleaning And Feature Engineering

After conducting our data cleaning and feature engineering process, I successfully did the following:

- Converted the 'Date of Travel' column to datetime format.
- Added date time parameters to our data.
- Renamed columns of the datasets.
- Removed missing values.
- Created a Master_Data dataframe by merging and joining our datasets.
- Created two new dataframes for both cab companies from the Master_Data dataframe.
- Created 12 new features (columns) which would very important for our analysis and insights.
- Converted our distance unit from kilometres to miles

Statistical Analysis

After conducting a thorough statistical analysis on the dataframes, I observed the following:

- There was a high presence of outliers in the dataframes.
- A high positive skewness in the 'Price_Charged' and 'Profit_Made' columns. This indicates a potential for large positive returns but poses a higher risk.
- A high kurtosis in some of our columns indicate a likelihood of extreme outcomes, whether positive or negative which would be very instrumental during risk assessment analysis and strategy selection.

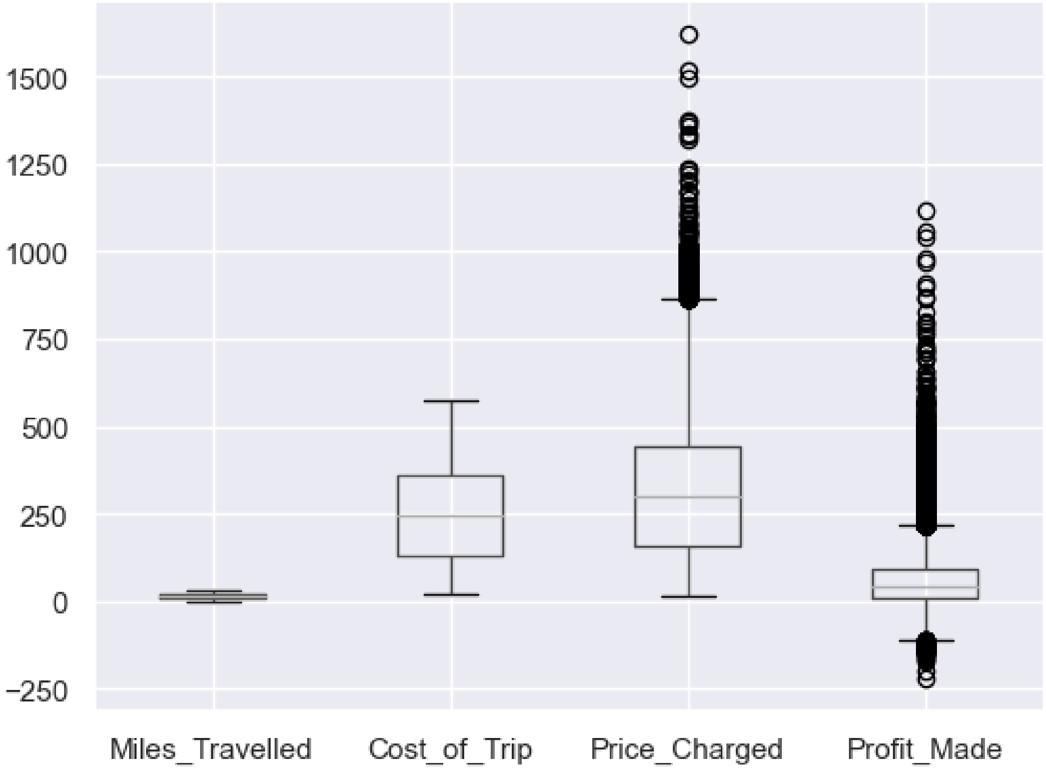
Data Visualization

In this section, I used graphs and visualizations of our data to conduct the following and make some actionable insights:

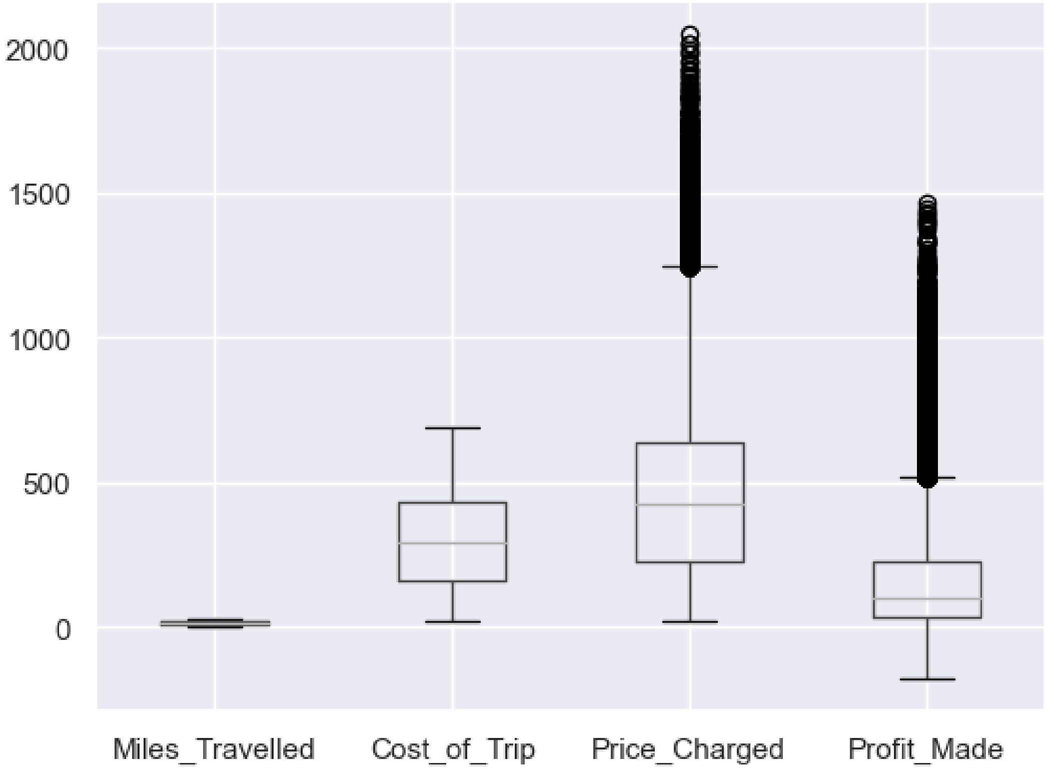
- Outlier Detection
- Correlation analysis
- Customer Analysis
- Rides Analysis
- Profit Analysis

Now let us go through each of them.....

Outlier Detection

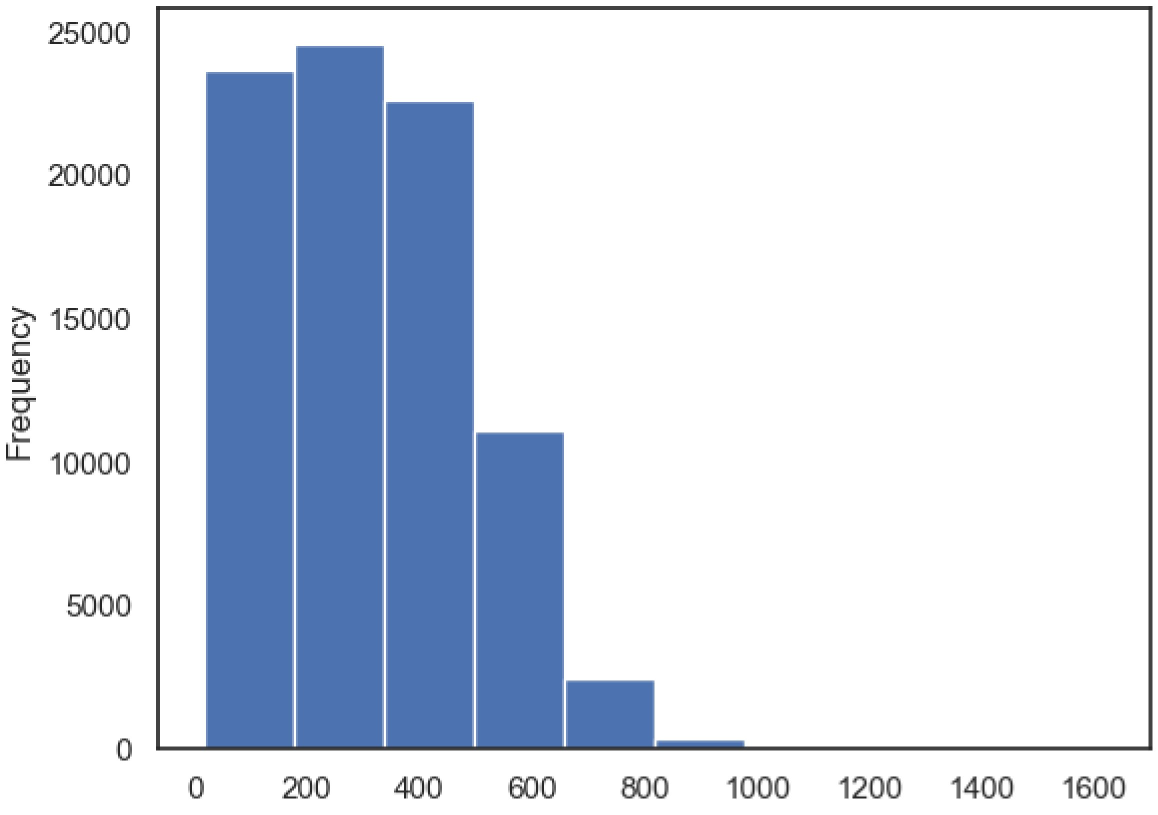


Pink Cab company

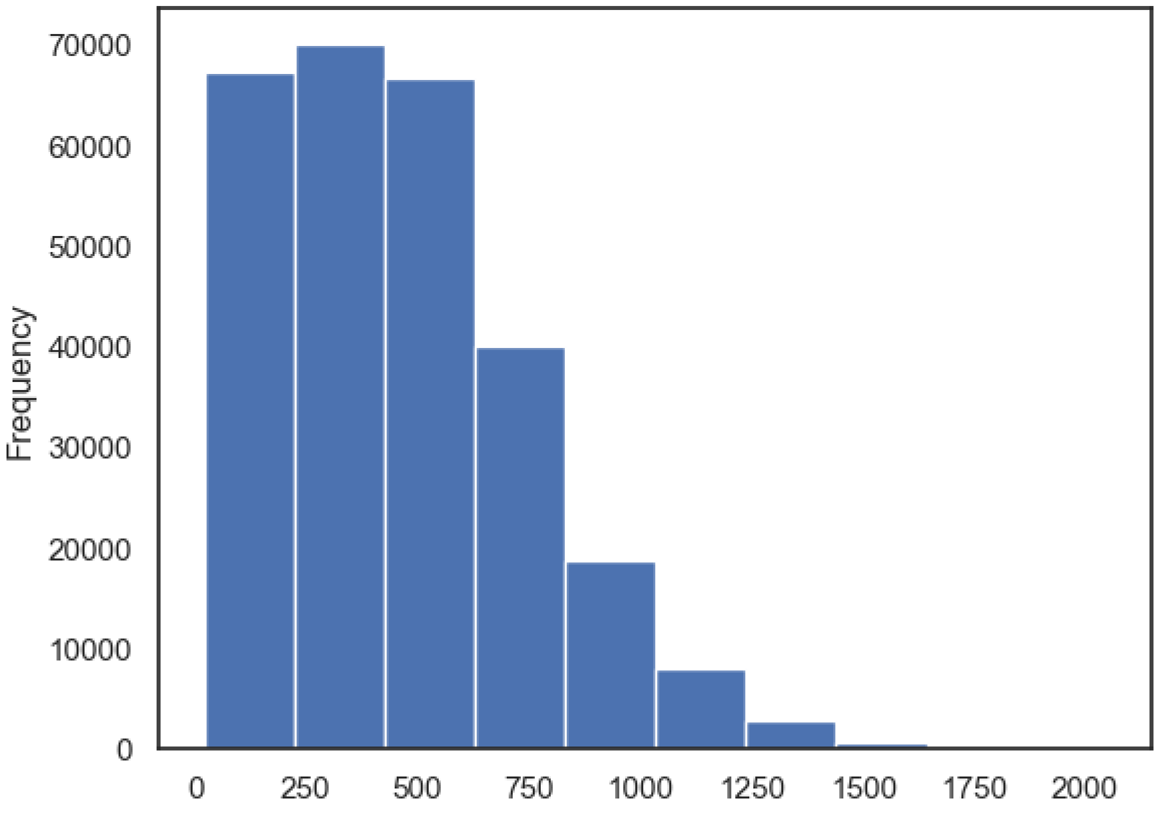


Yellow Cab company

- We can see from our visualizations that there are outliers present in our data for both cab companies.
- There are quite a number of outliers present in the 'Price_Charged' and 'Profit_Made' columns.



Pink Cab Company

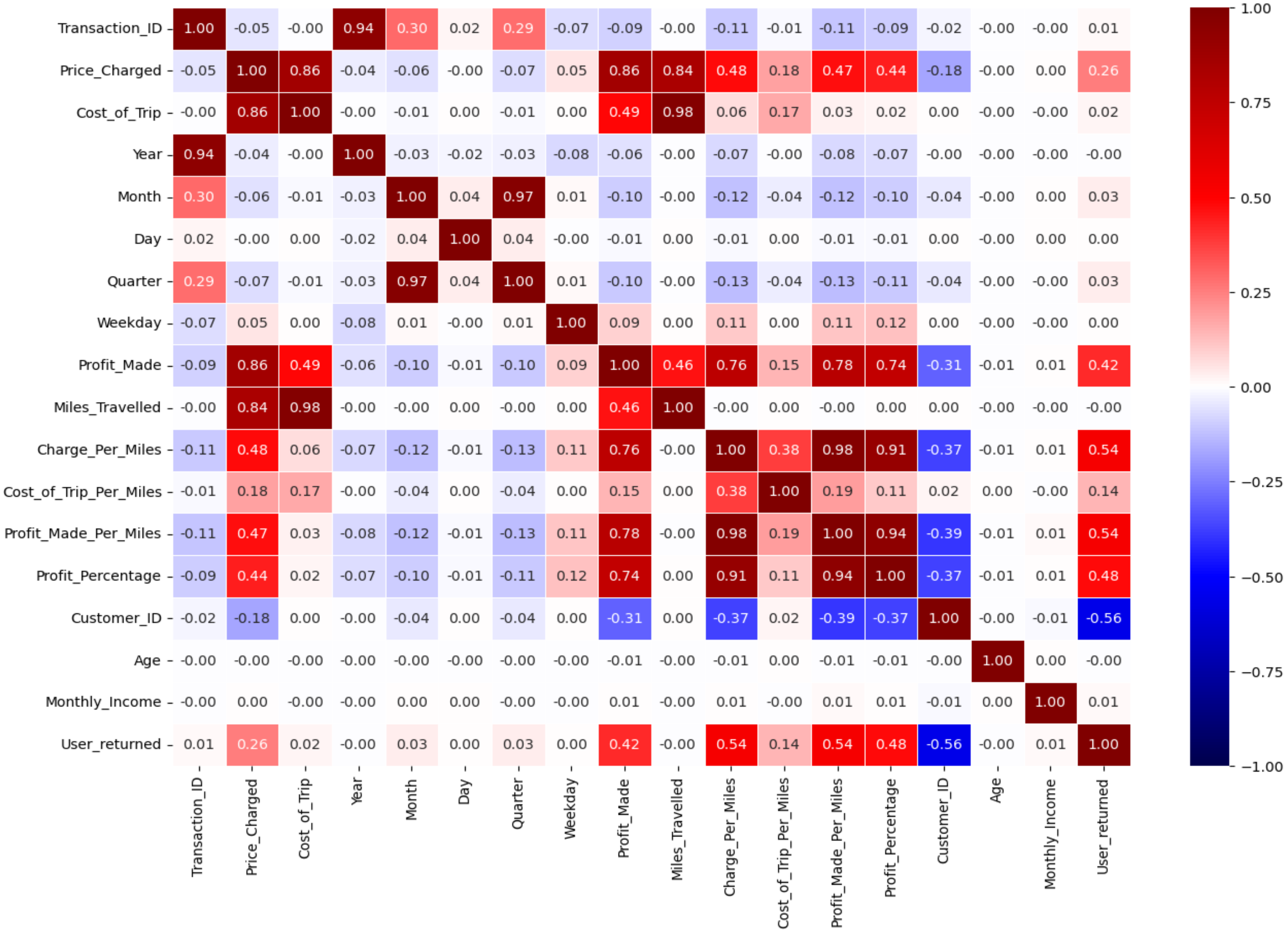


Yellow Cab Company

From the graphs above, we can see the following;

- The Yellow Cab company charges higher fares than the Pink Cab company.
- Also, both graphs indicate a right skewed distribution.
- Some of the prices charged by the Yellow cab company were extremely high.

Correlation Analysis

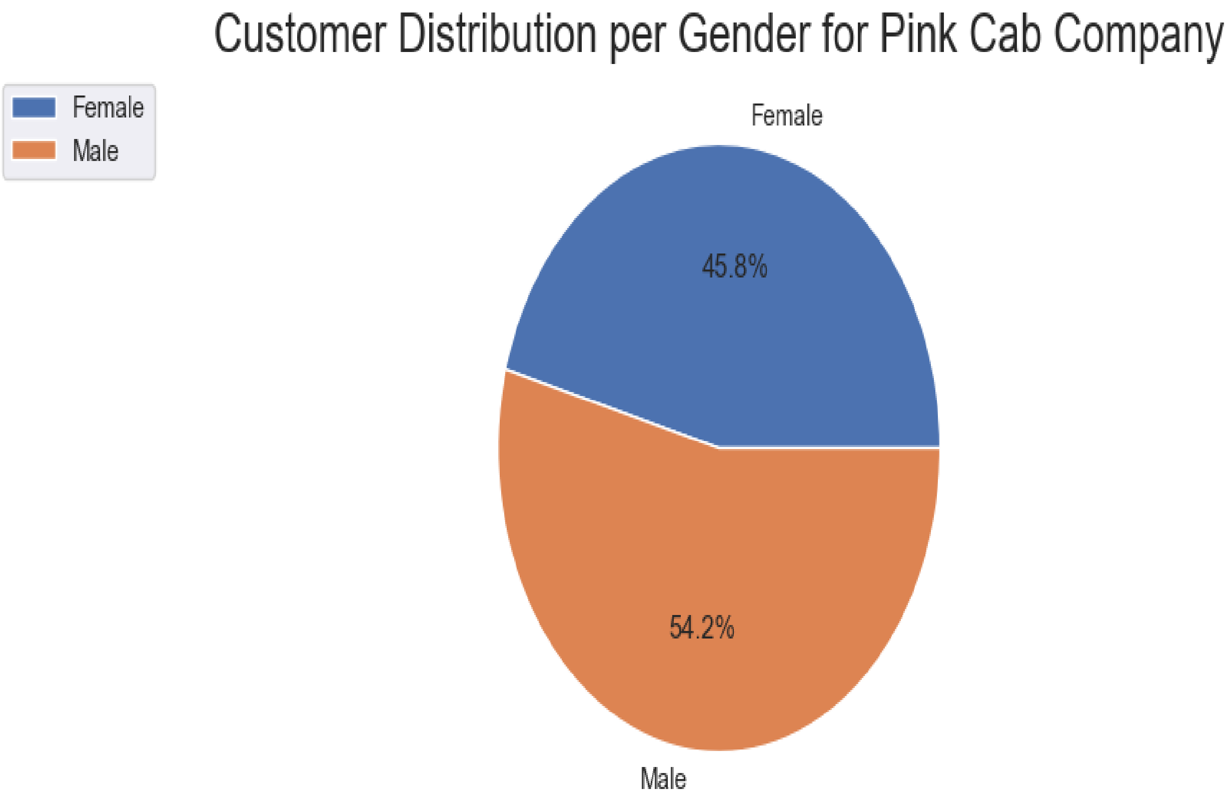


	Transaction_ID	Price_Charged	Cost_of_Trip	Year	Month	Day	Quarter	Weekday	Profit_Made	Miles_Travelled	Charge_Per_Miles	Cost_of_Trip_Per_Miles	Profit_Made_Per_Miles	Profit_Percentage	Customer_ID	Age	Monthly_Income	User_returned
Transaction_ID	1.00	-0.05	-0.00	0.94	0.30	0.02	0.29	-0.07	-0.09	-0.00	-0.11	-0.01	-0.11	-0.09	-0.02	-0.00	-0.00	0.01
Price_Charged	-0.05	1.00	0.86	-0.04	-0.06	-0.00	-0.07	0.05	0.86	0.84	0.48	0.18	0.47	0.44	-0.18	-0.00	0.00	0.26
Cost_of_Trip	-0.00	0.86	1.00	-0.00	-0.01	0.00	-0.01	0.00	0.49	0.98	0.06	0.17	0.03	0.02	0.00	-0.00	-0.00	0.02
Year	0.94	-0.04	-0.00	1.00	-0.03	-0.02	-0.03	-0.08	-0.06	-0.00	-0.07	-0.00	-0.08	-0.07	-0.00	-0.00	-0.00	-0.00
Month	0.30	-0.06	-0.01	-0.03	1.00	0.04	0.97	0.01	-0.10	-0.00	-0.12	-0.04	-0.12	-0.10	-0.04	-0.00	0.00	0.03
Day	0.02	-0.00	0.00	-0.02	0.04	1.00	0.04	-0.00	-0.01	0.00	-0.01	0.00	-0.01	-0.01	0.00	-0.00	0.00	0.00
Quarter	0.29	-0.07	-0.01	-0.03	0.97	0.04	1.00	0.01	-0.10	-0.00	-0.13	-0.04	-0.13	-0.11	-0.04	-0.00	-0.00	0.03
Weekday	-0.07	0.05	0.00	-0.08	0.01	-0.00	0.01	1.00	0.09	0.00	0.11	0.00	0.11	0.12	0.00	-0.00	-0.00	0.00
Profit_Made	-0.09	0.86	0.49	-0.06	-0.10	-0.01	-0.10	0.09	1.00	0.46	0.76	0.15	0.78	0.74	-0.31	-0.01	0.01	0.42
Miles_Travelled	-0.00	0.84	0.98	-0.00	-0.00	0.00	-0.00	0.00	0.46	1.00	-0.00	0.00	-0.00	0.00	0.00	-0.00	-0.00	-0.00
Charge_Per_Miles	-0.11	0.48	0.06	-0.07	-0.12	-0.01	-0.13	0.11	0.76	-0.00	1.00	0.38	0.98	0.91	-0.37	-0.01	0.01	0.54
Cost_of_Trip_Per_Miles	-0.01	0.18	0.17	-0.00	-0.04	0.00	-0.04	0.00	0.15	0.00	0.38	1.00	0.19	0.11	0.02	0.00	-0.00	0.14
Profit_Made_Per_Miles	-0.11	0.47	0.03	-0.08	-0.12	-0.01	-0.13	0.11	0.78	-0.00	0.98	0.19	1.00	0.94	-0.39	-0.01	0.01	0.54
Profit_Percentage	-0.09	0.44	0.02	-0.07	-0.10	-0.01	-0.11	0.12	0.74	0.00	0.91	0.11	0.94	1.00	-0.37	-0.01	0.01	0.48
Customer_ID	-0.02	-0.18	0.00	-0.00	-0.04	0.00	-0.04	0.00	-0.31	0.00	-0.37	0.02	-0.39	-0.37	1.00	-0.00	-0.01	-0.56
Age	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.01	-0.00	-0.01	0.00	-0.01	-0.01	-0.00	1.00	0.00	-0.00
Monthly_Income	-0.00	0.00	-0.00	-0.00	0.00	0.00	-0.00	-0.00	0.01	-0.00	0.01	-0.00	0.01	0.01	-0.01	0.00	1.00	0.01
User_returned	0.01	0.26	0.02	-0.00	0.03	0.00	0.03	0.00	0.42	-0.00	0.54	0.14	0.54	0.48	-0.56	-0.00	0.01	1.00

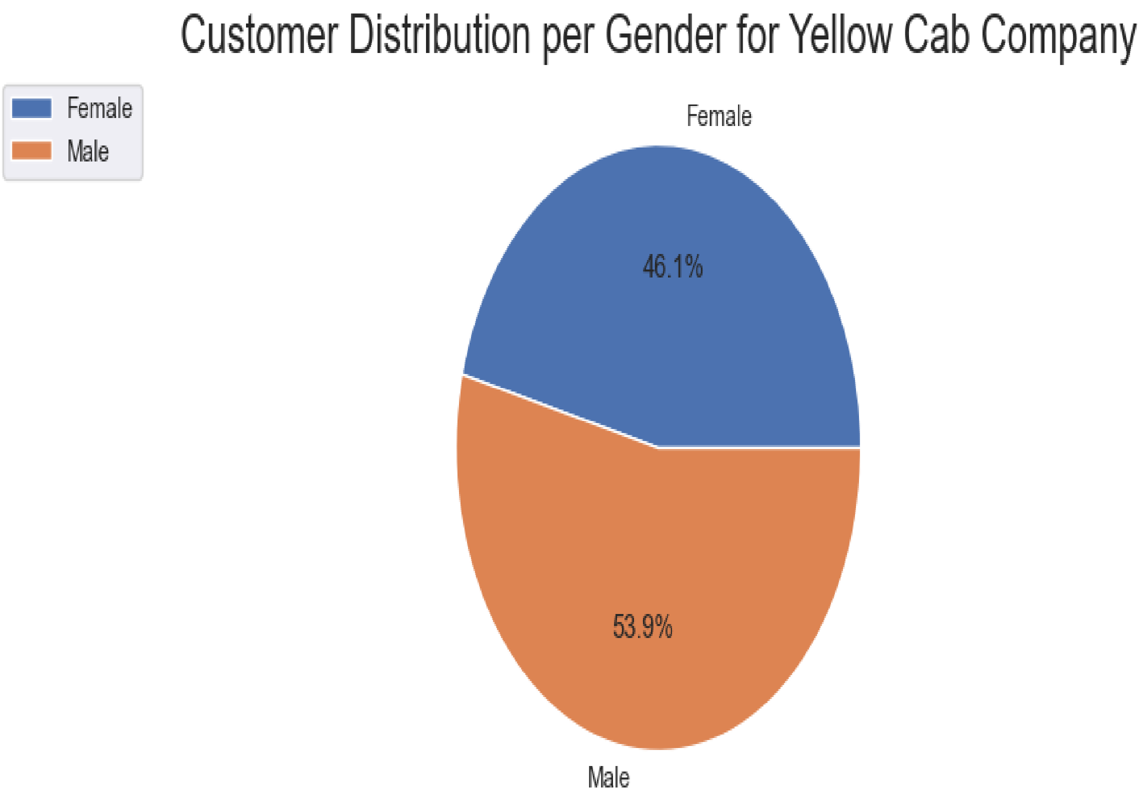
From the heatmap of the correlation matrix;

- I noticed a strong correlation between the price charged and the miles travelled. This indicates that the price charged for a trip depends on the distance travelled.
- Also, there was a strong correlation between the price charged and the profit made. This means the profit made depends on the price charged by the cab companies.
- Similarly, there was a strong correlation between the 'Cost_of_Trip' column and 'Miles_Travelled' column. This means the cost of a trip depends on the miles travelled.

Customer Analysis

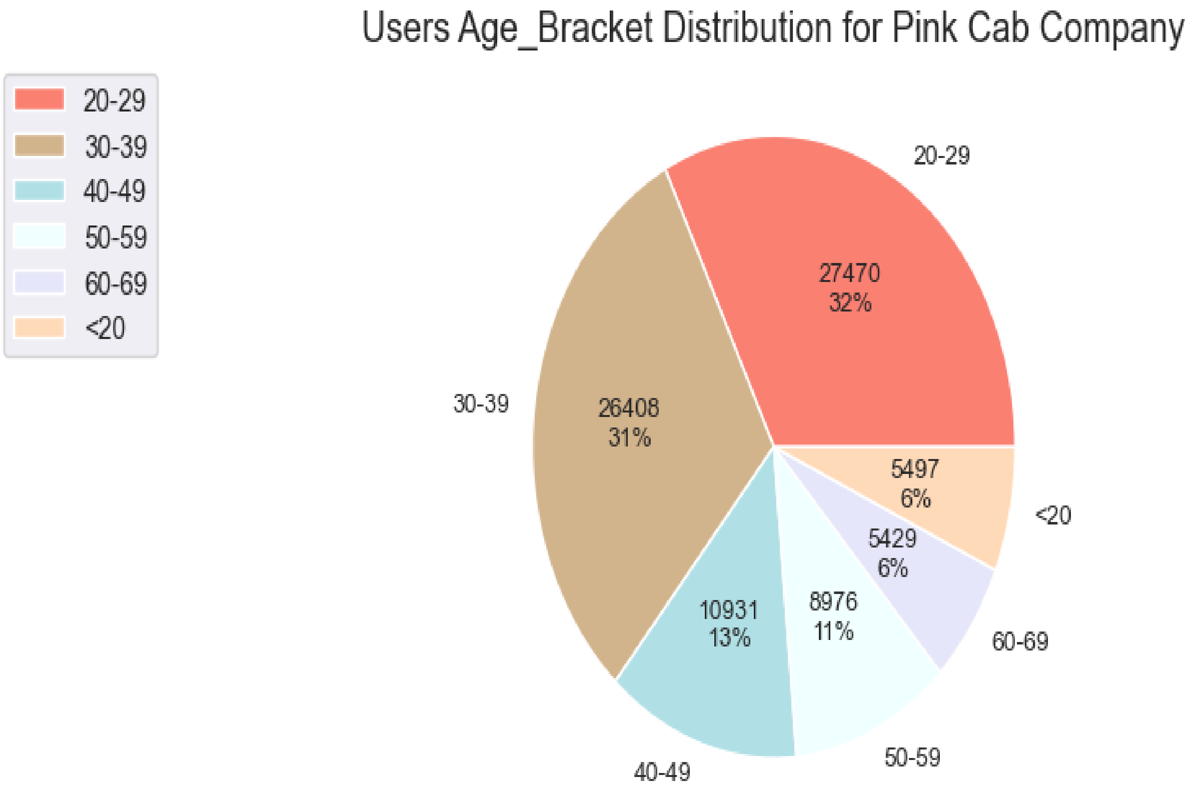


Customer Distribution per Gender for Pink Cab Company

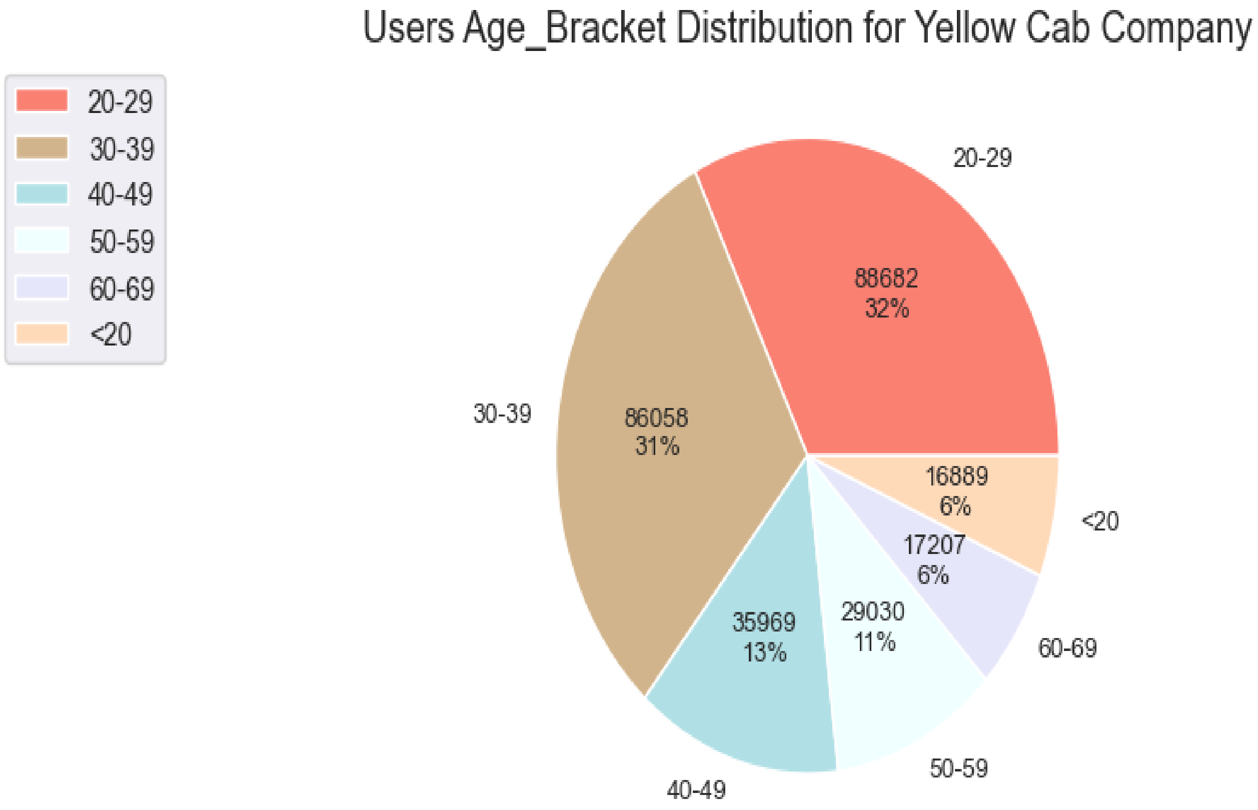


Customer Distribution per Gender for Yellow Cab Company

From the pie charts above we can see both cab companies have more male customers than female customers.



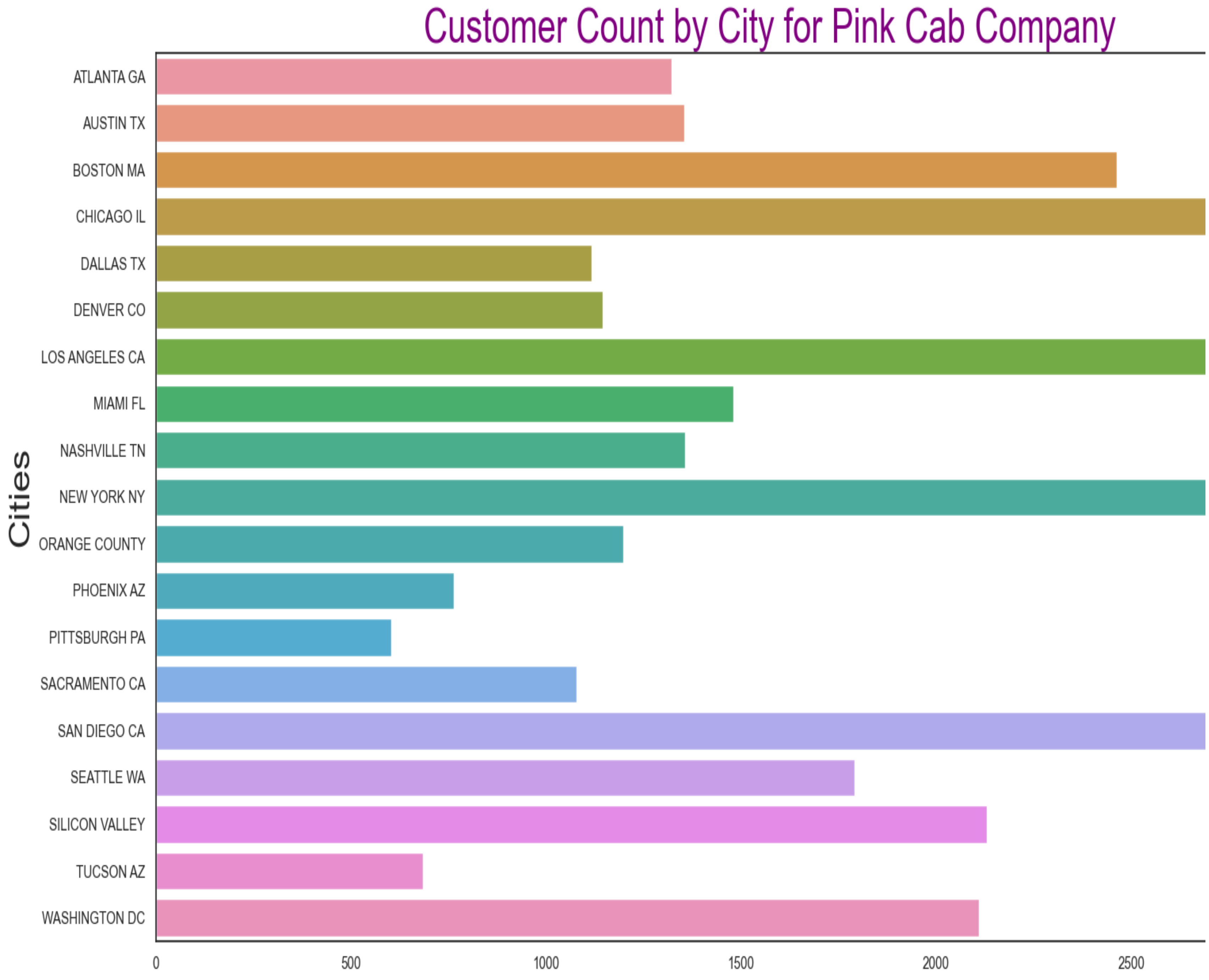
Users Age_Bracket Distribution for Pink Cab Company



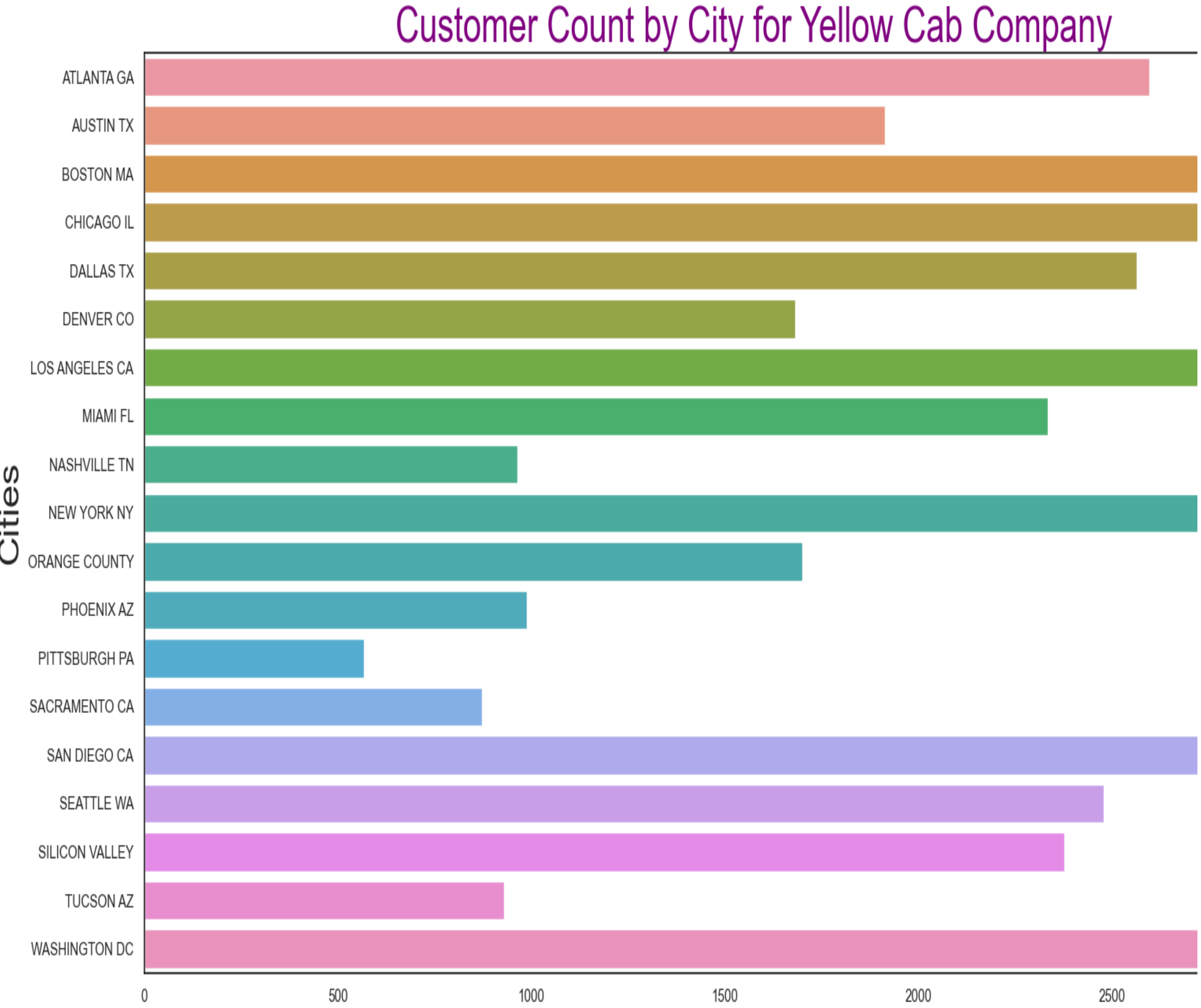
Users Age_Bracket Distribution for Yellow Cab Company

From our pie chart we can see;

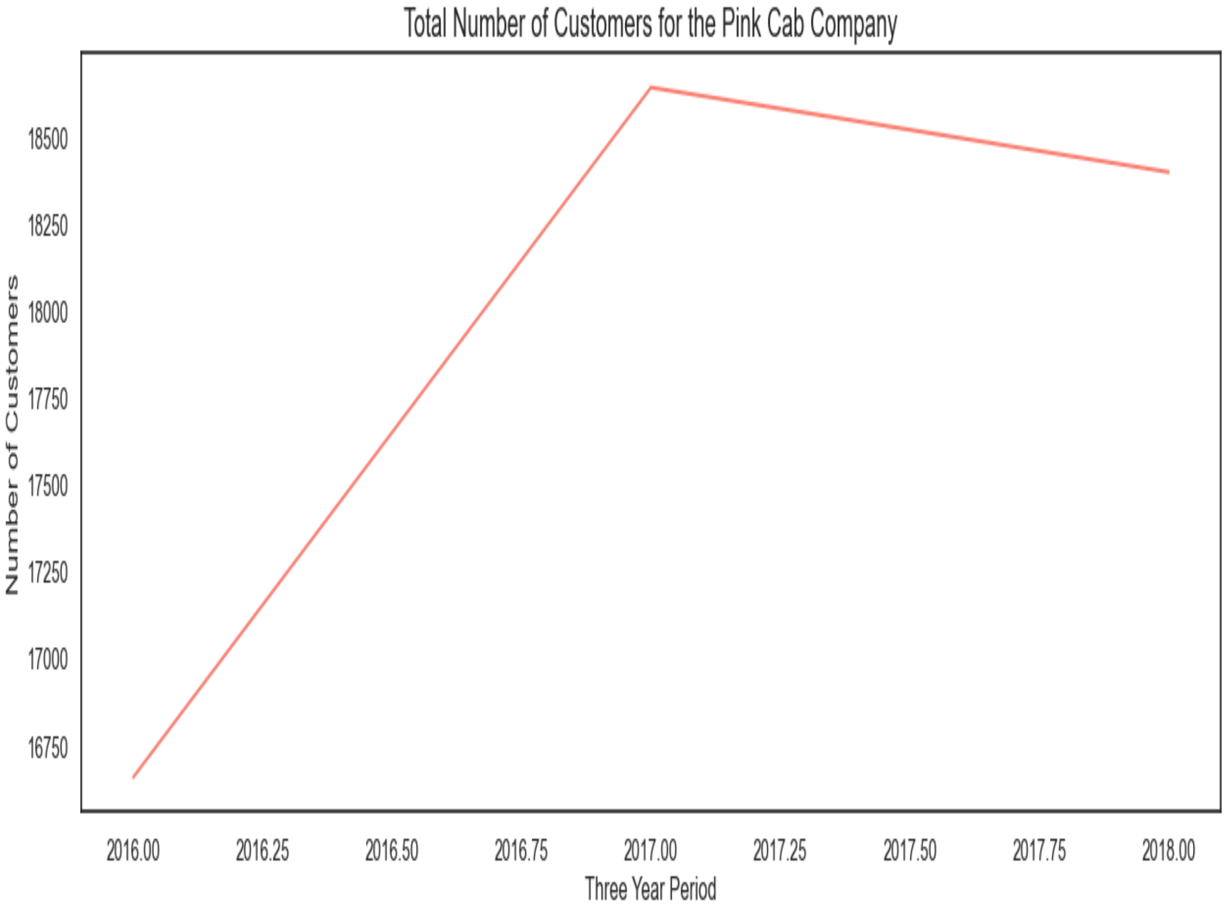
- People between the ages of 20-29 years and 30-39 years use both cab companies the most with percentages of 32% and 31% respectively.
- People less than 20 years and between 60-69 use the cab companies the least with each group having a percentage of 6%.

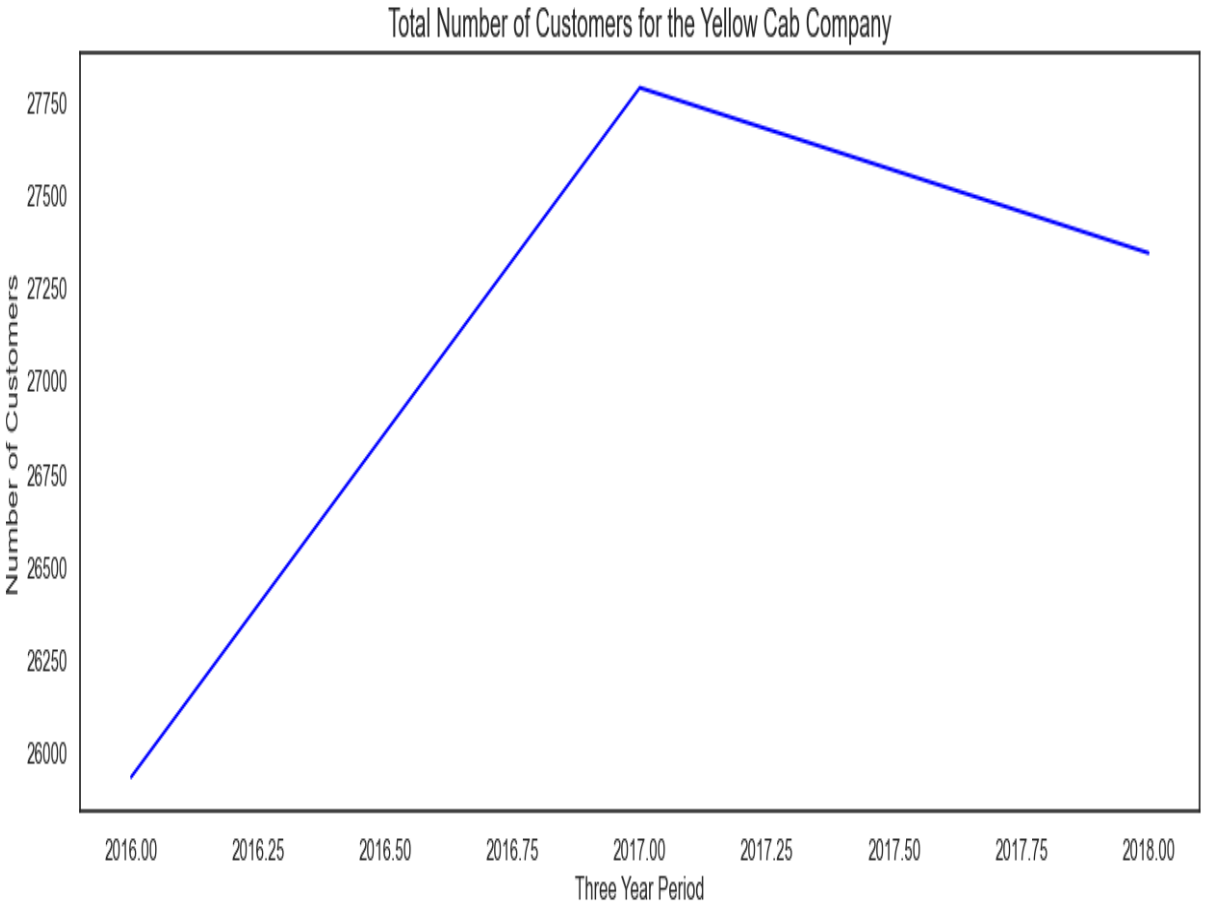


Looking at our graph, you will notice Los Angeles, Denver and San Diego have very high number of Pink Cab customers with Pittsburgh having the least number of customers

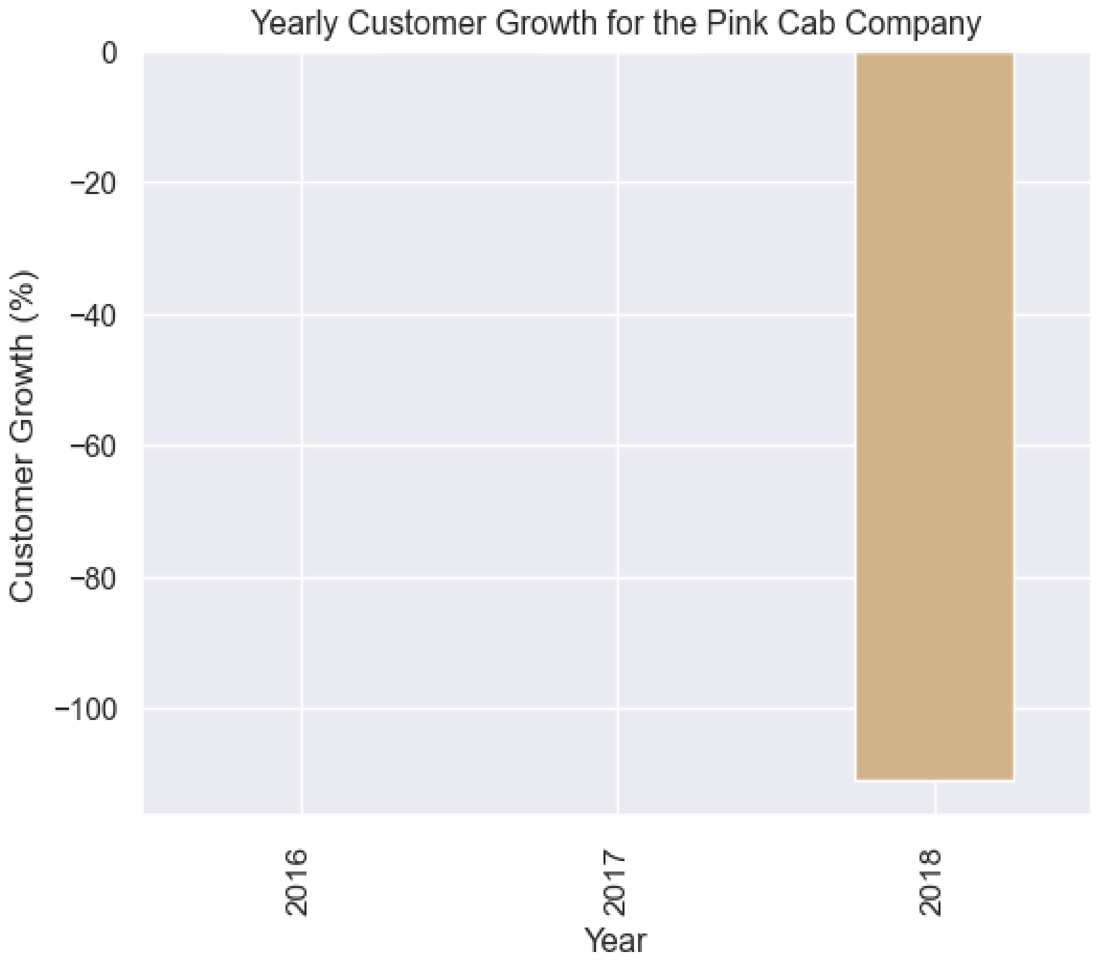


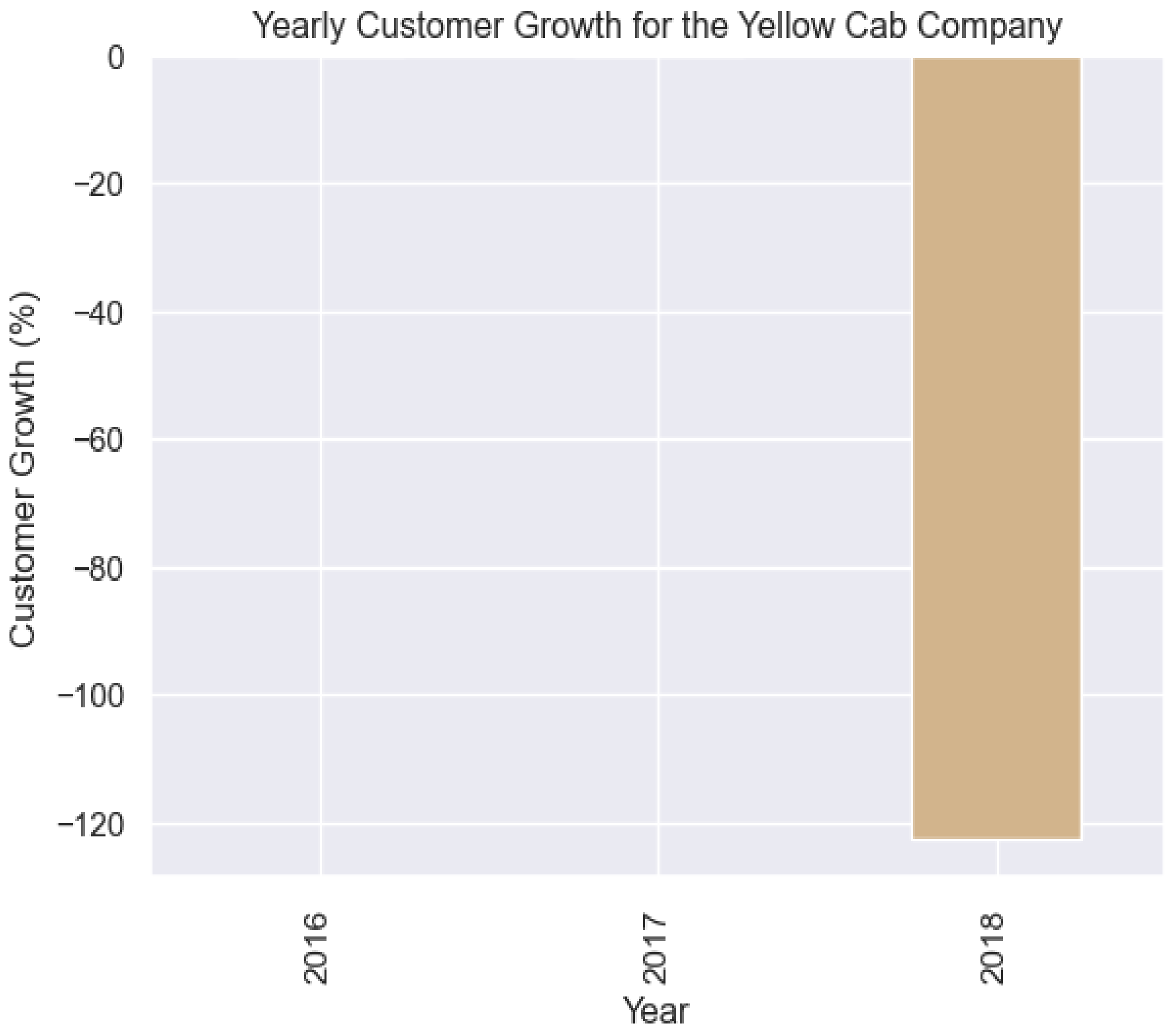
Looking at our graph, you will notice Boston, Los Angeles, New York and Washington DC have very high number of Yellow Cab customers with Pittsburgh having the least number of customers.



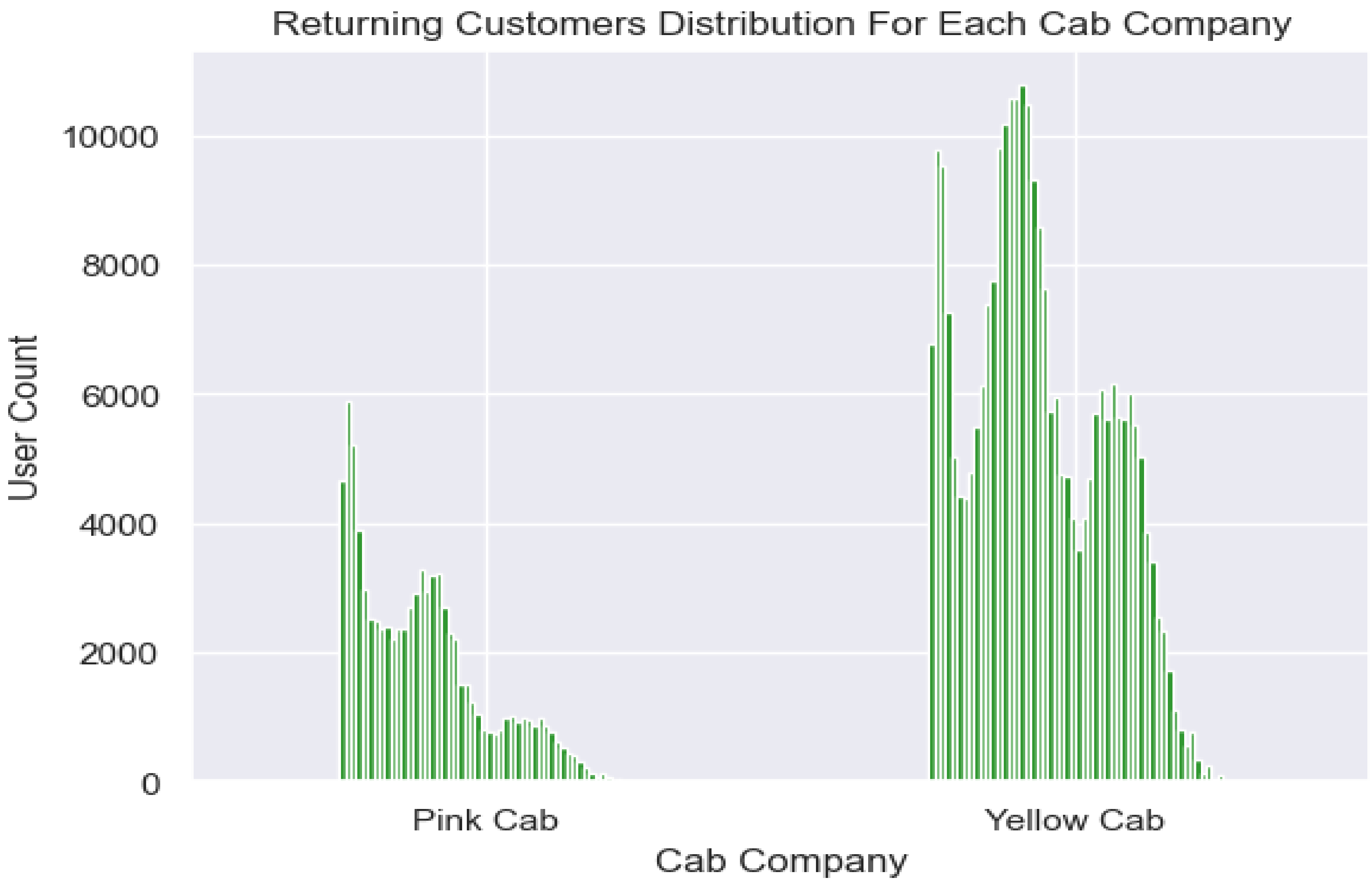


From the graphs, we can observe a steady growth in customers from 2016 - 2017 for both cab companies. However, this growth declines from 2017 -2018.





From the graphs above, we can see both cab companies experienced a negative customer growth in 2018. This is because both companies lost customers in the year 2018 after experiencing a significant customer growth in 2017.

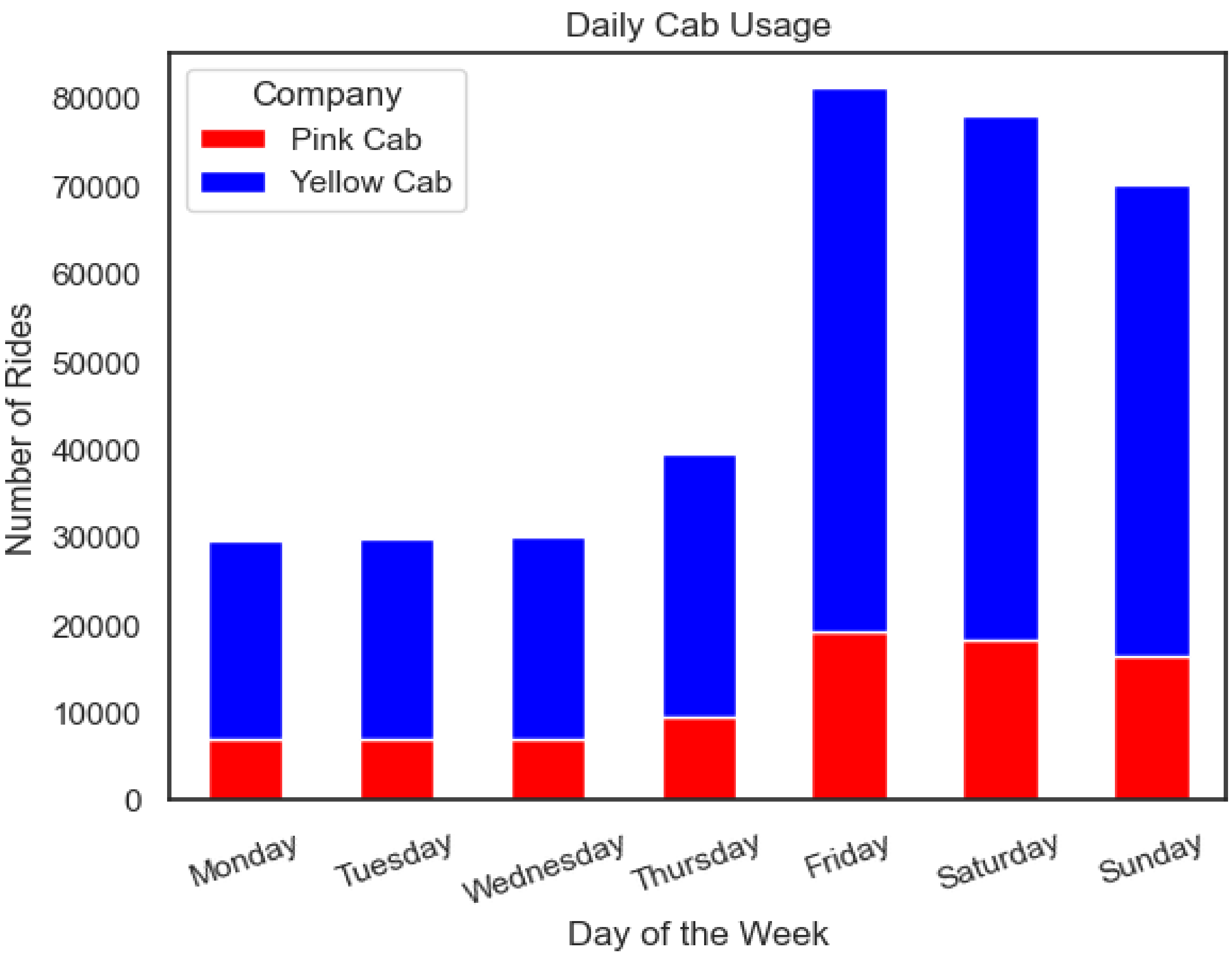


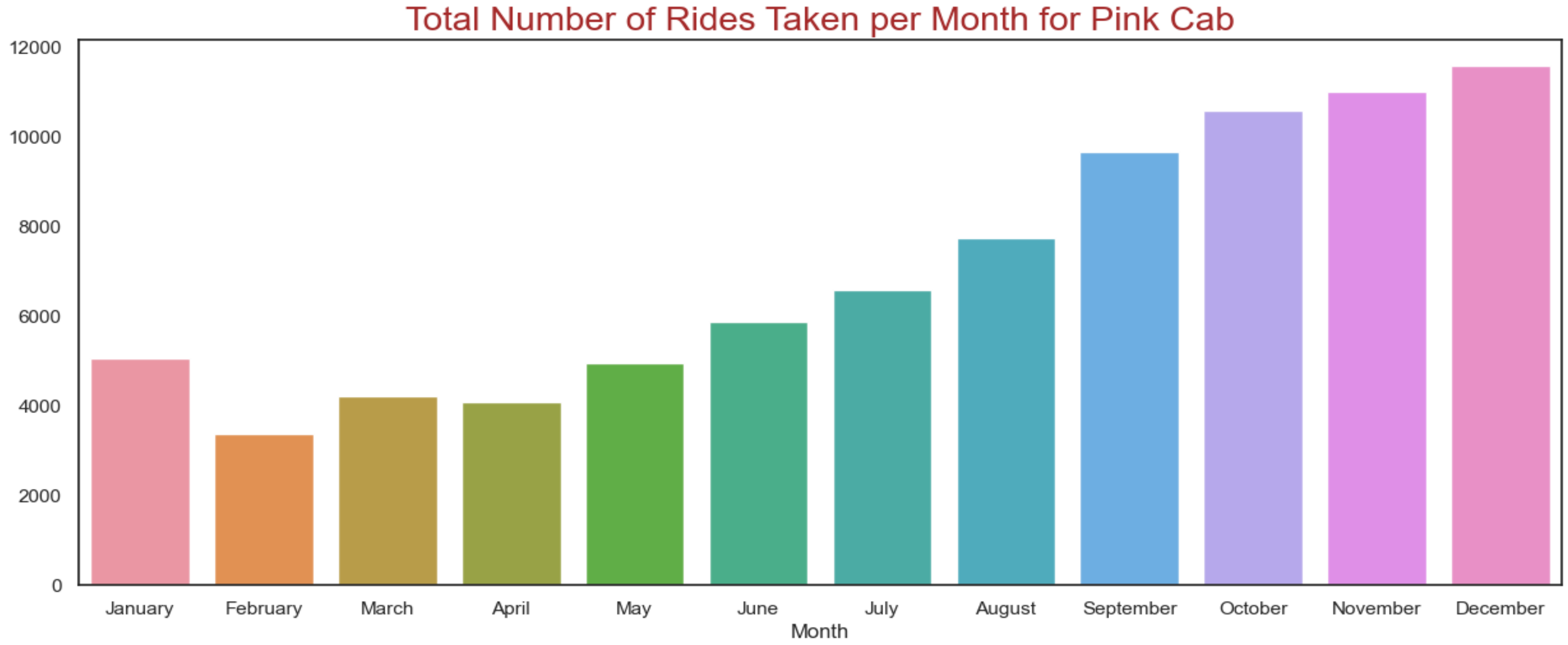
From the visualization, it is evident more users returned to use the Yellow Cab company after their first experience with the company. Very few people returned to patronize the Pink Cab company

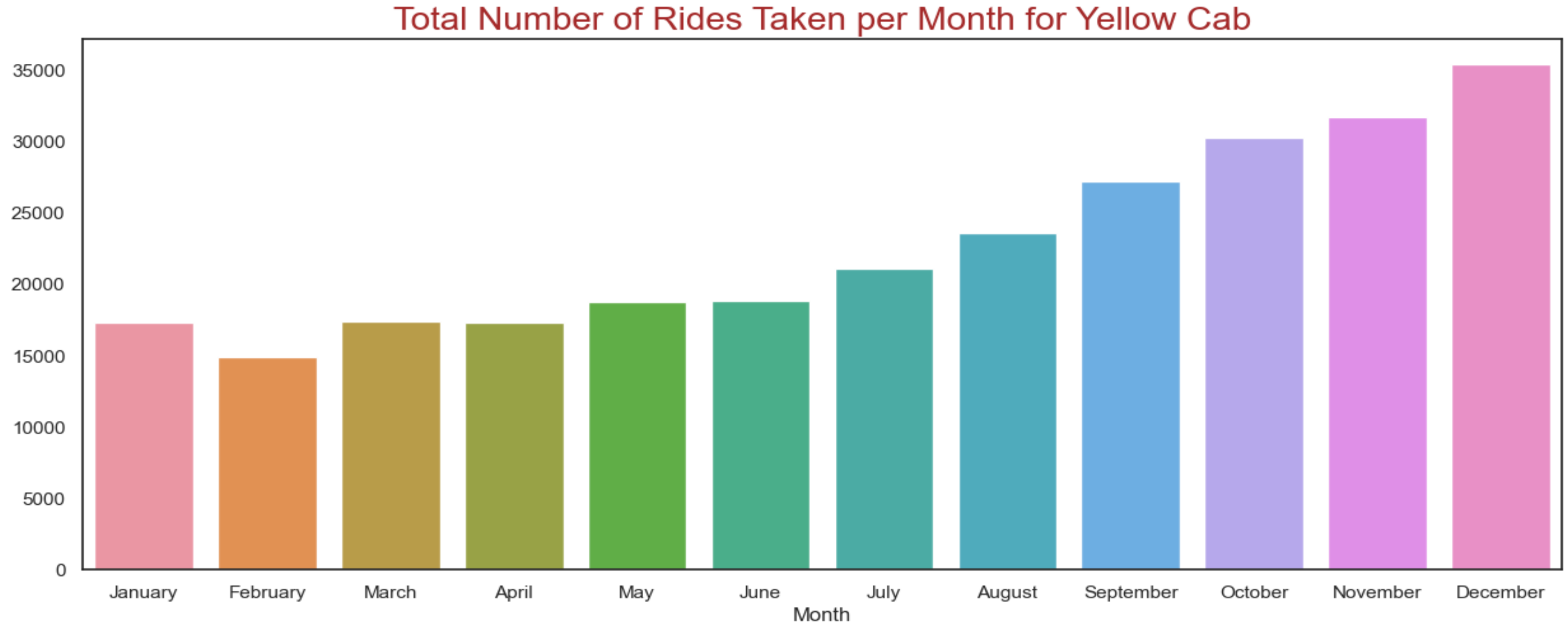
Rides Analysis

From our bar graph;

- It is clear the Yellow Cab company has a higher daily cab usage compared to the Pink Cab company.
- Also, the cab companies are patronized more on the weekends compared to weekdays.
- Additionally, both cab companies are patronized the most on Fridays

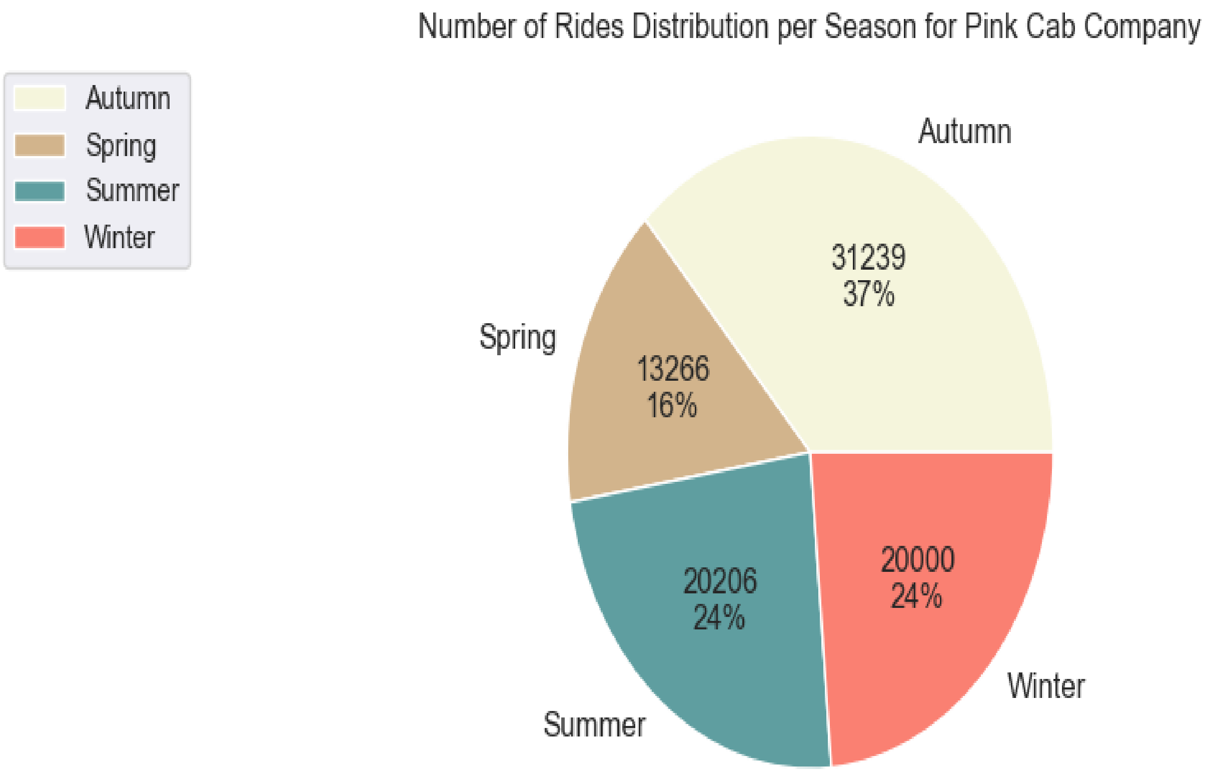


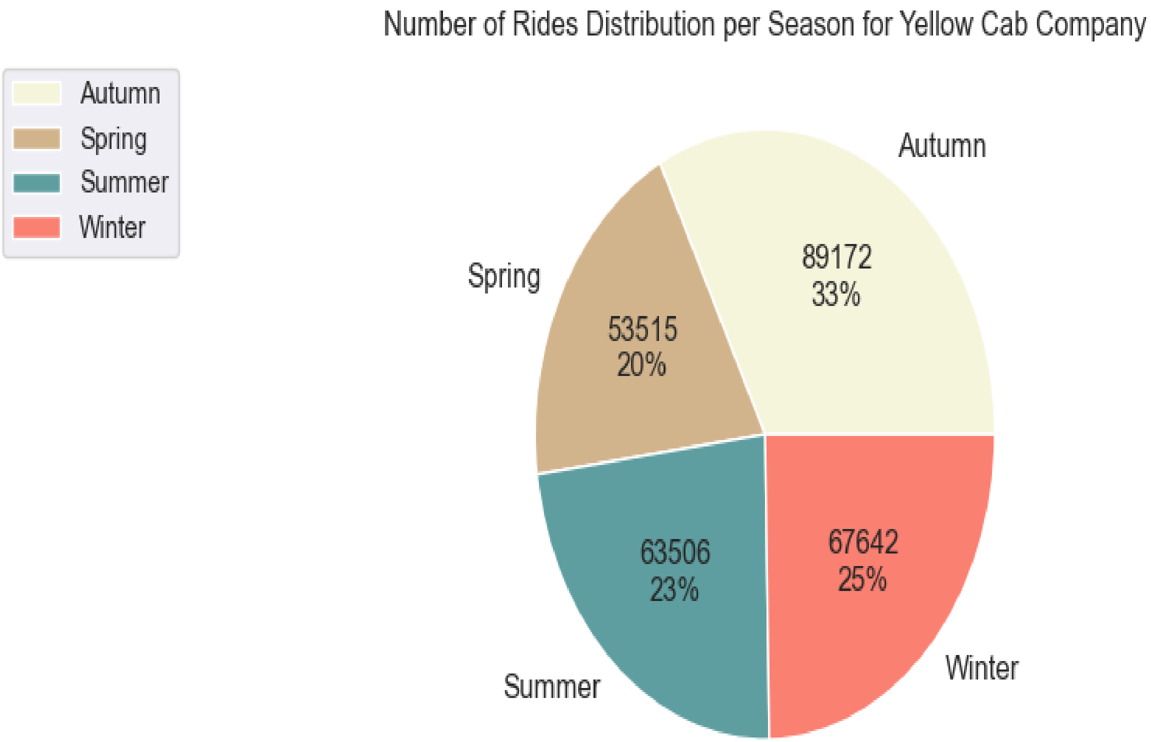




We can see from our visualizations that;

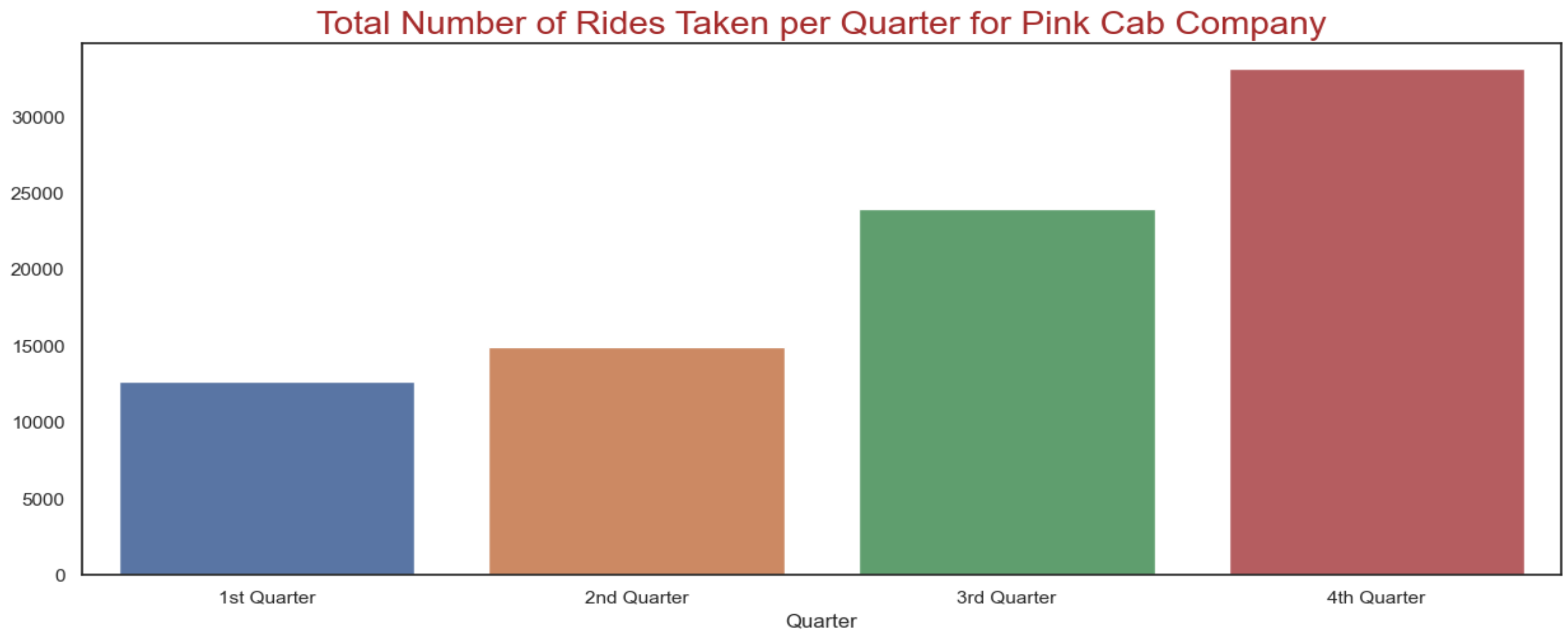
- December experiences the highest patronage for both cab companies.
- Both cab companies are patronized the least in February.

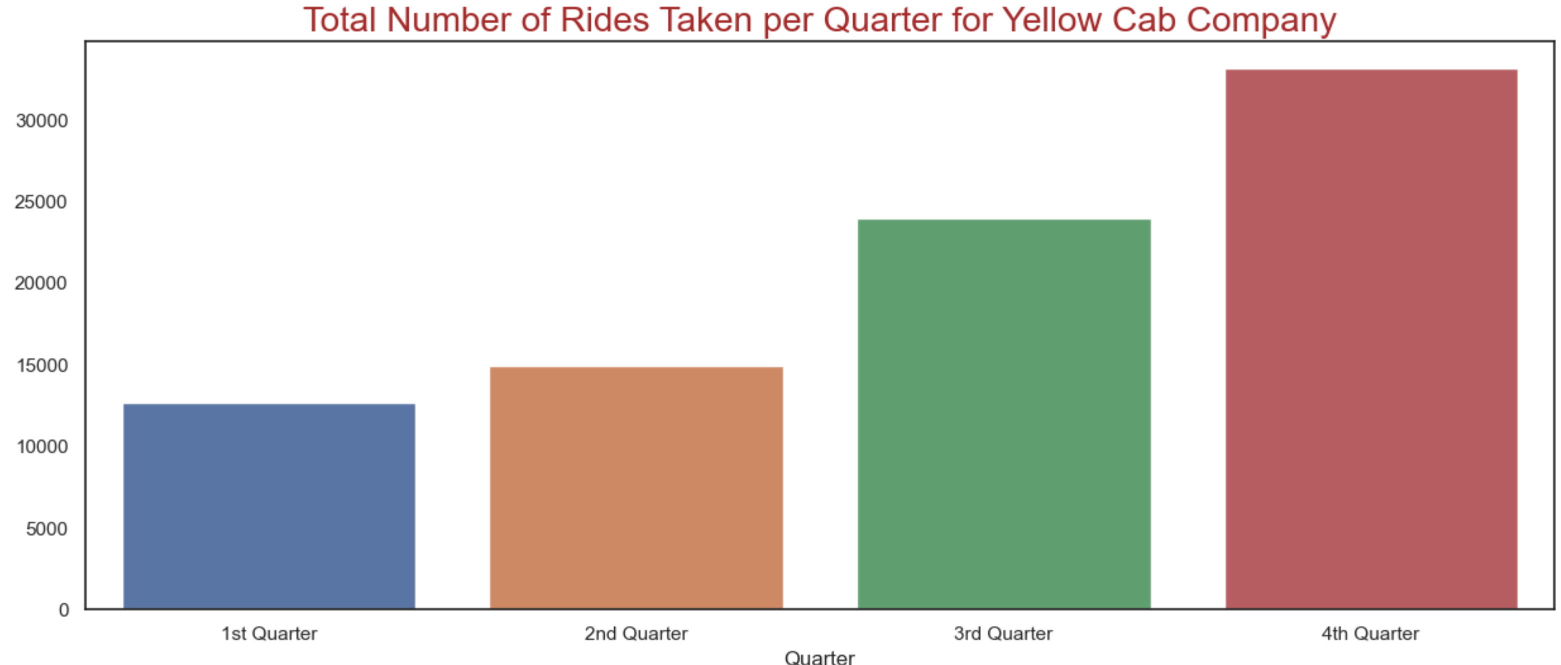




Looking at the visualizations of both cab companies;

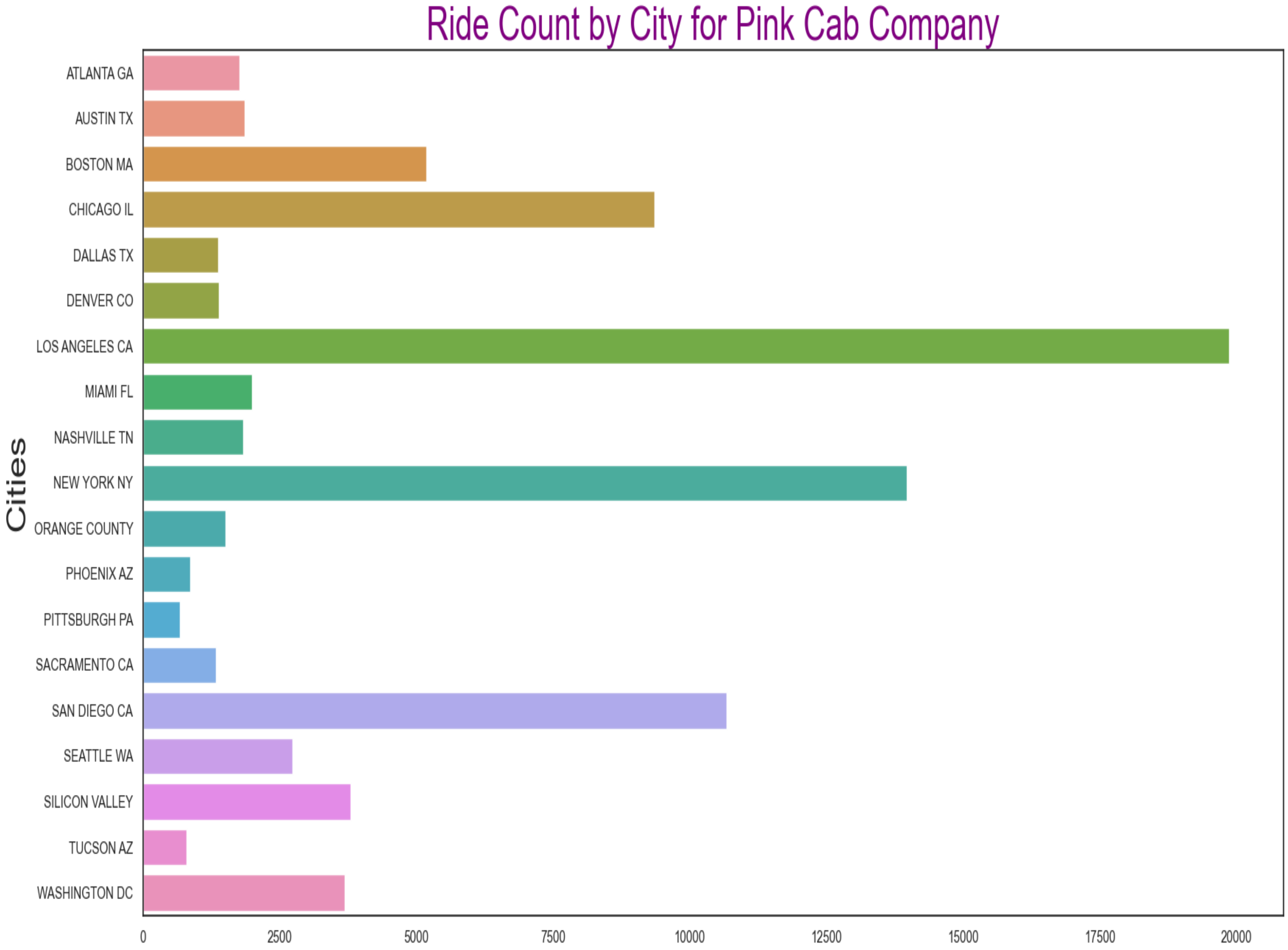
- They are used the most in Autumn season (September to November)
- During Spring time (March to May), they are used the least.





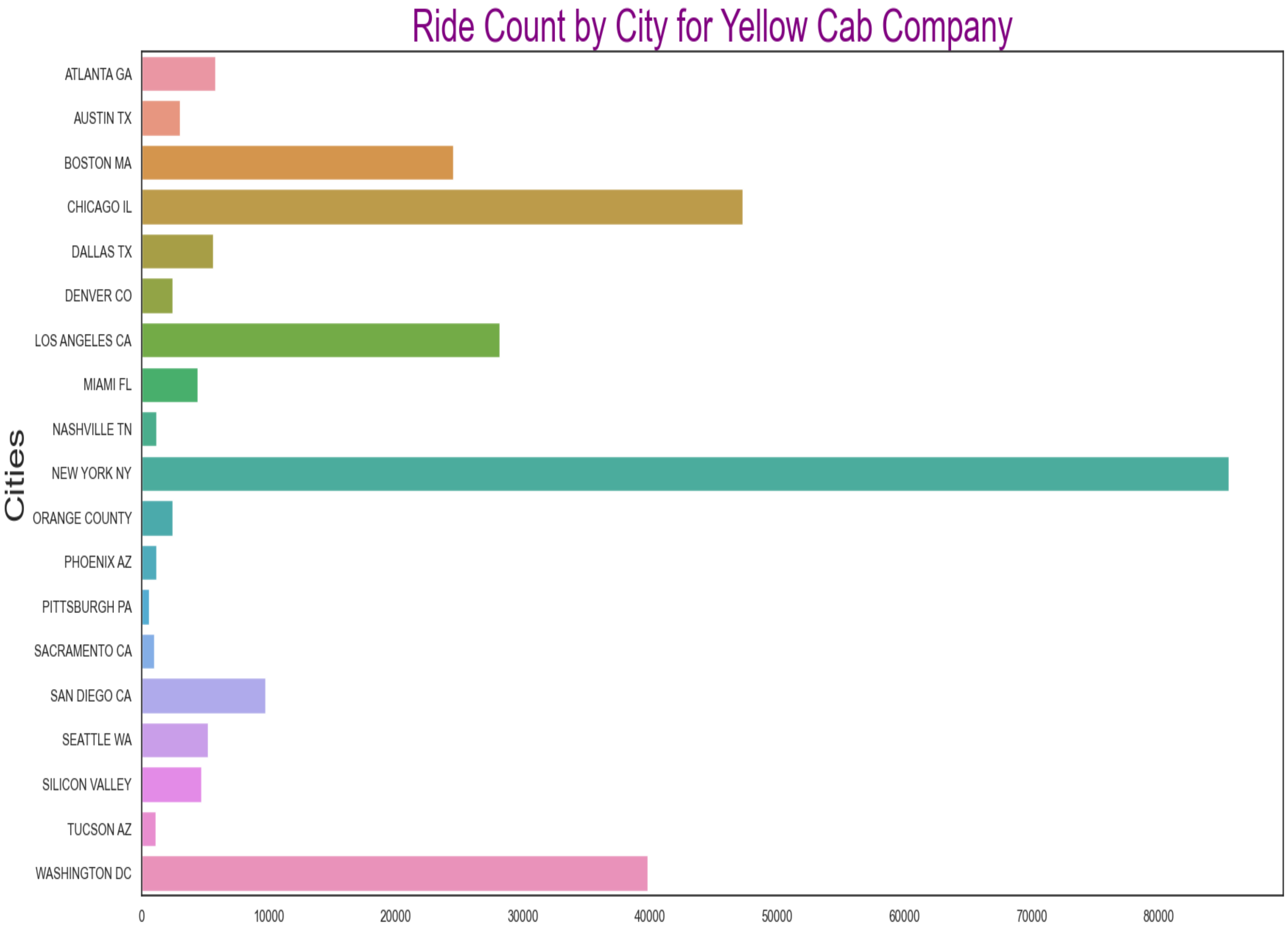
From our bar graphs, we can see;

- Both cab companies experience the highest usage in the 4th quarter of the year. Hence, more profit is made in the 4th quarter of the year
- The lowest cab usage and profit made by both cab companies is seen in the 1st quarter



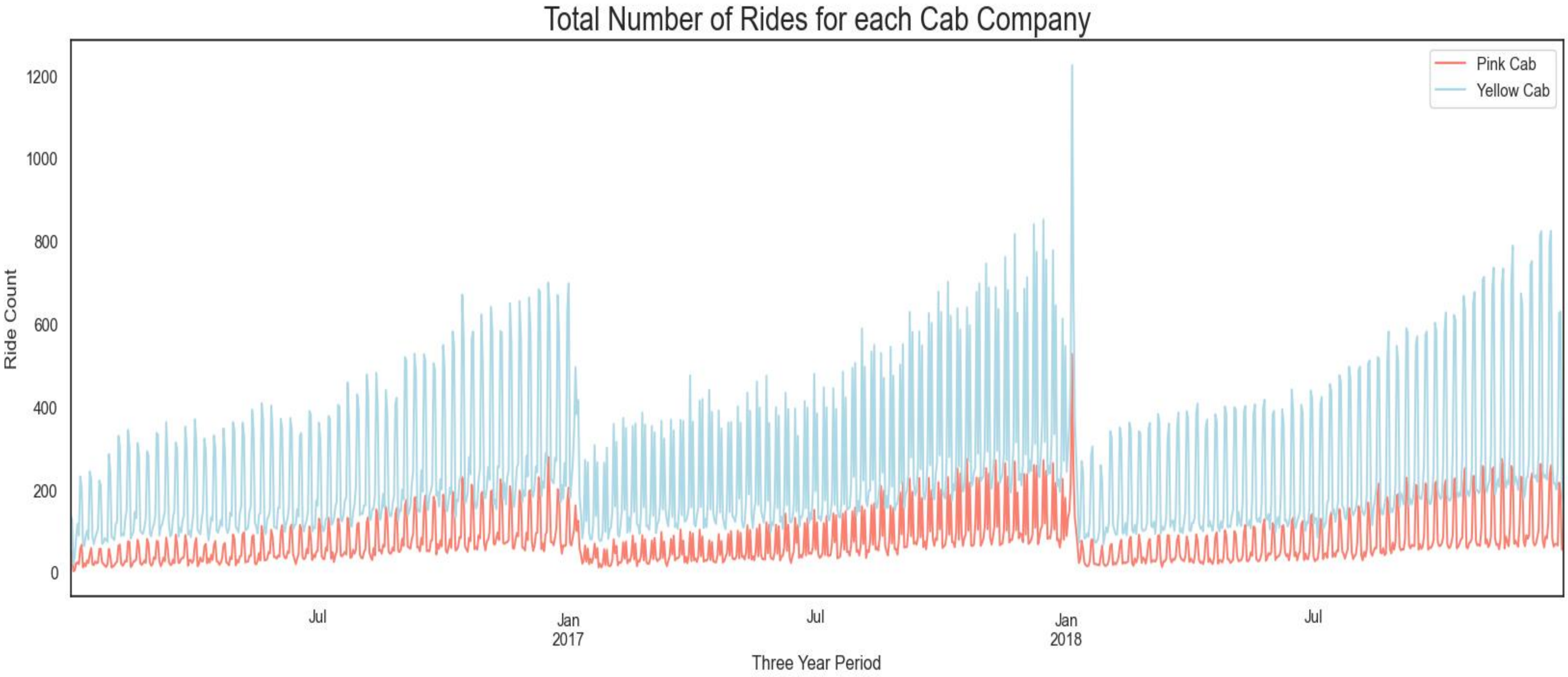
- From the graph we can see more Pink Cabs are taken in Los Angeles, CA compared to other cities

- Also, Pittsburgh, PA has the least Pink Cabs taken



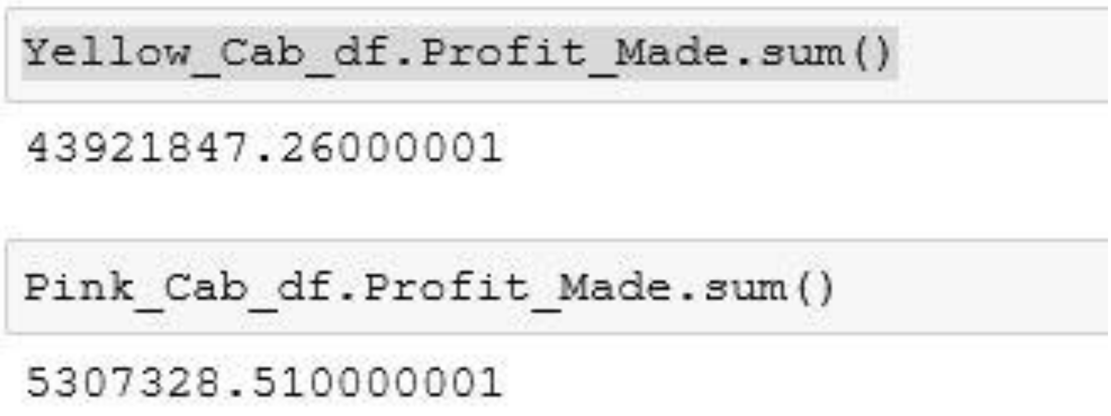
- From our bar plot, it is evident New York City, NY has the highest number of Yellow Cab rides taken.

- Also, Pittsburgh, PA has the least Yellow Cabs taken



From the graph above, it is evident the Yellow Cab company has more rides taken than the Pink Cab company over the three year period.

Profit Analysis

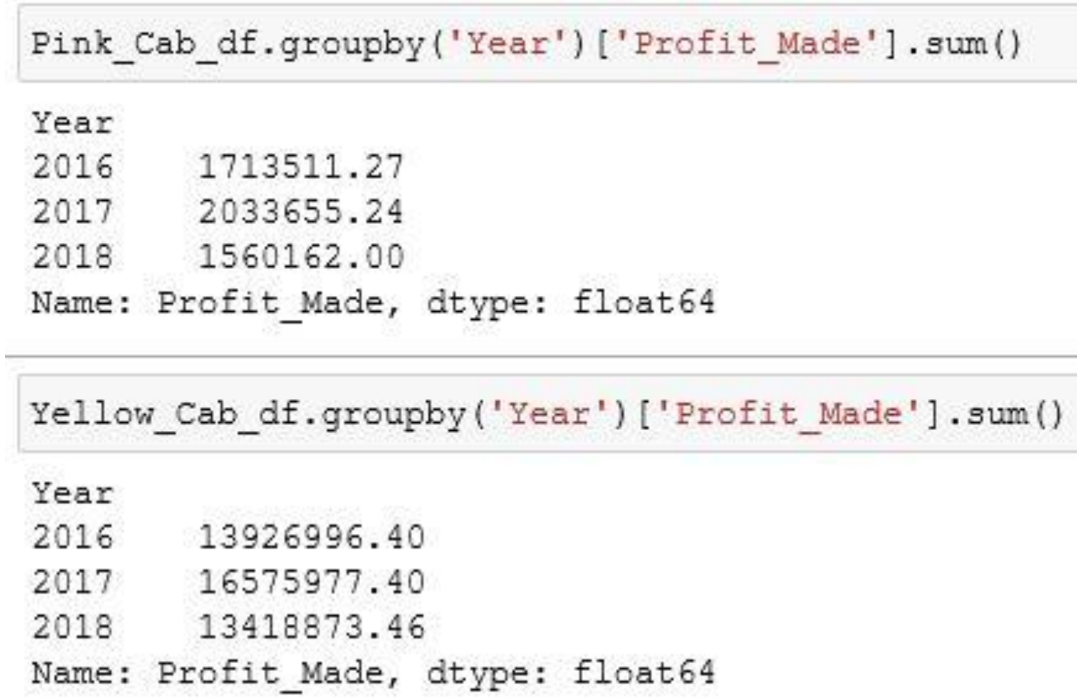


```
Yellow_Cab_df.Profit_Made.sum()
```

```
43921847.26000001
```

```
Pink_Cab_df.Profit_Made.sum()
```

```
5307328.510000001
```



```
Pink_Cab_df.groupby('Year')['Profit_Made'].sum()
```

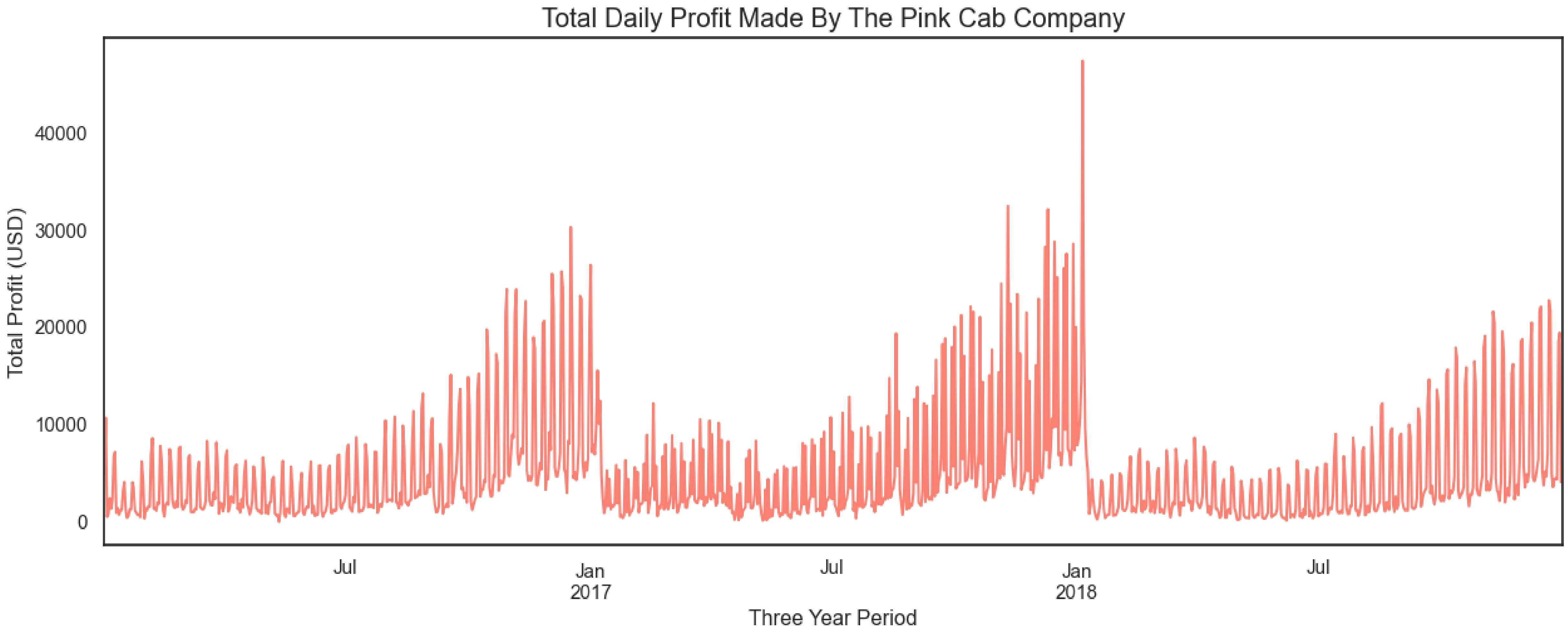
```
Year
2016    1713511.27
2017    2033655.24
2018    1560162.00
Name: Profit_Made, dtype: float64
```

```
Yellow_Cab_df.groupby('Year')['Profit_Made'].sum()
```

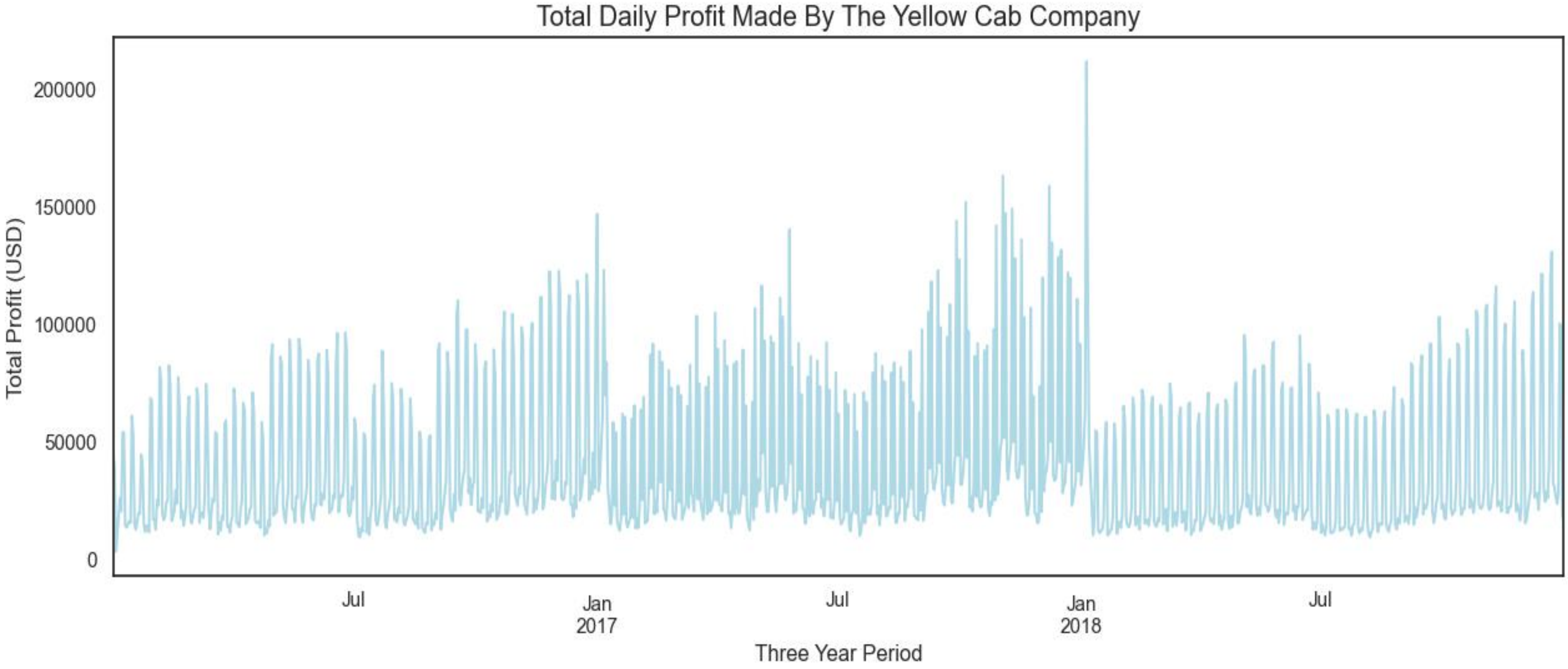
```
Year
2016    13926996.40
2017    16575977.40
2018    13418873.46
Name: Profit_Made, dtype: float64
```

It is evident that the Yellow Cab company makes more profit than the Pink Cab company with a total sum of profit made of $43,921,847 from 2016-2018.

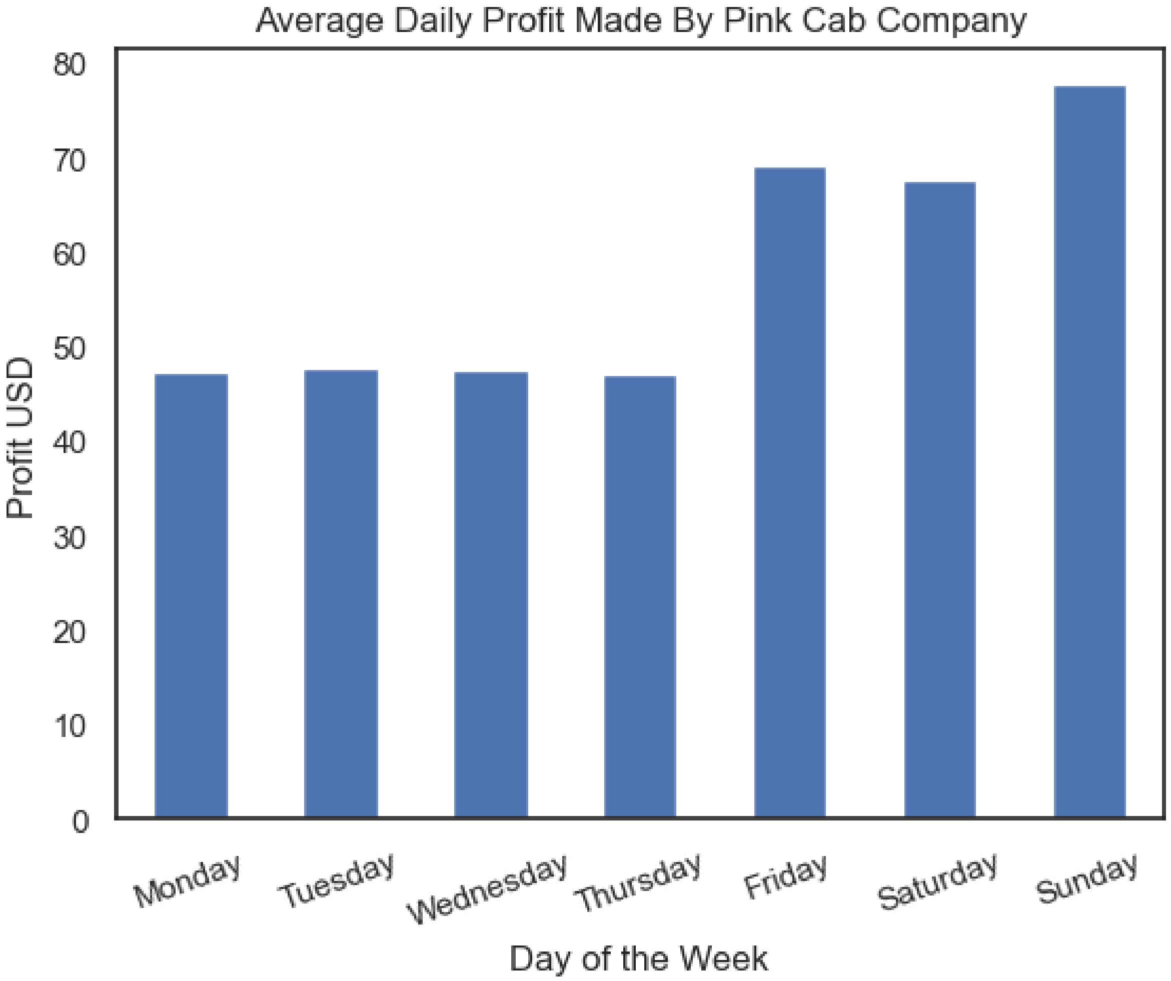
Also, the both cab companies experience their highest profits in 2017 In 2018, both cab companies experienced their lowest profits.

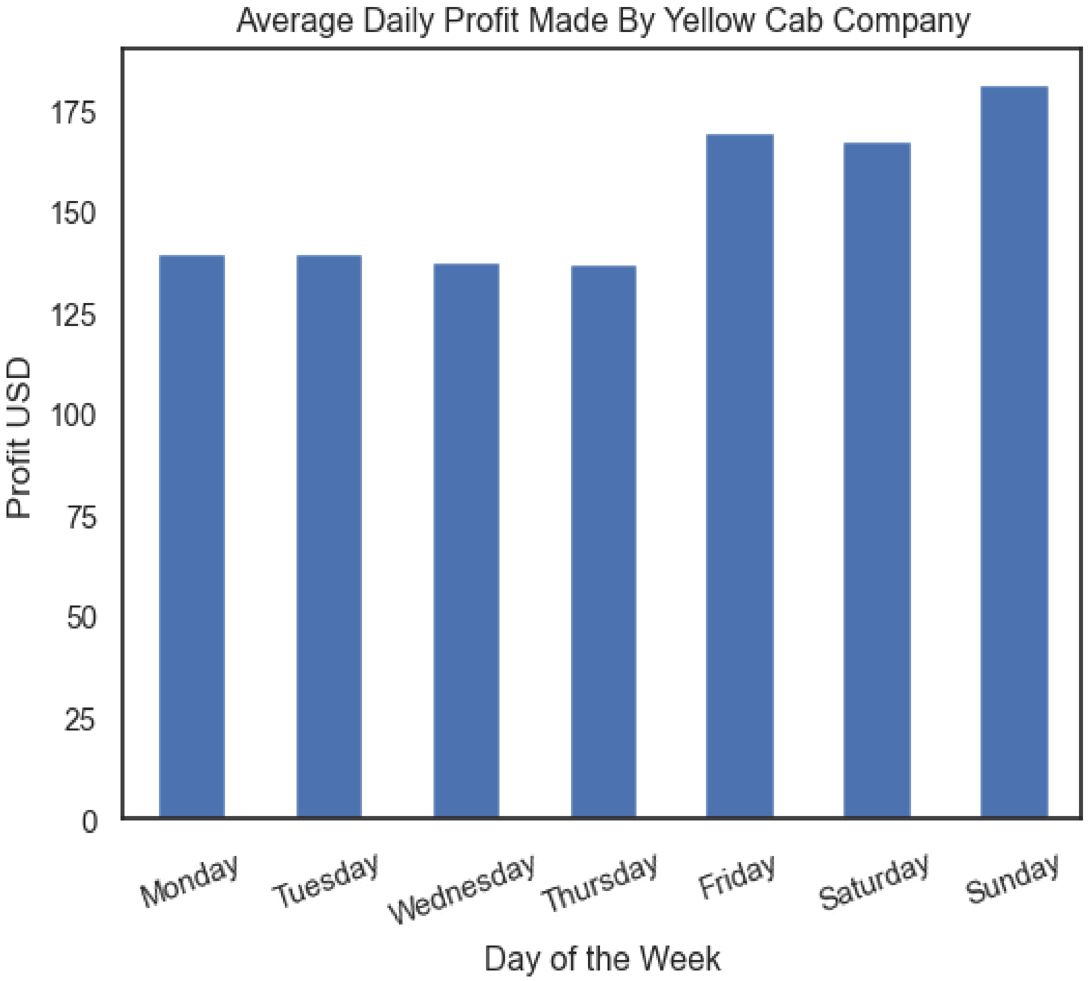


- Looking at our graph, we can see the Pink Cab makes its highest profit at the end of 2017.
- However, 2018 shows a very poor performance with the company experiencing its lowest profits made over the three year period.

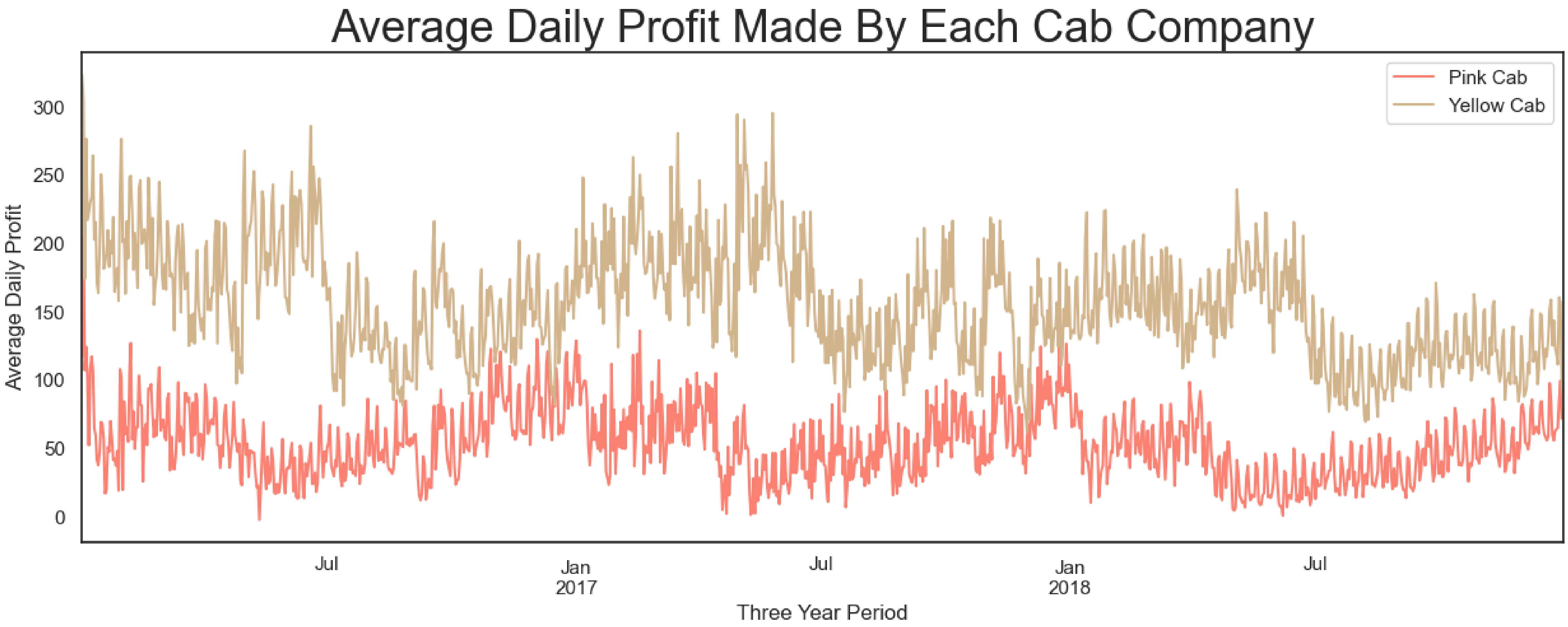


- Looking at our graph, we can see the Yellow Cab makes its highest profit at the end of 2017.
- However, 2018 shows a very poor performance with the company experiencing its lowest profits made over the three year period.

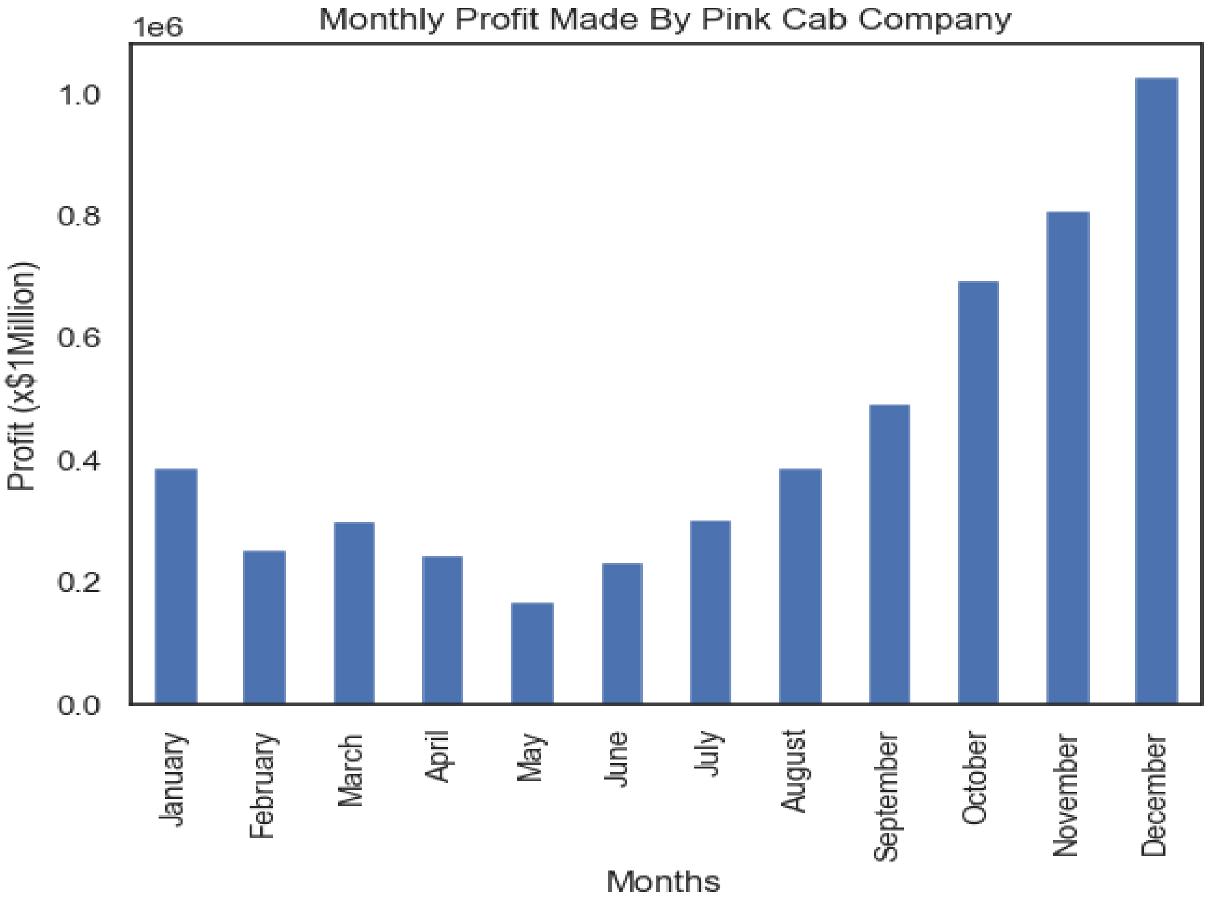


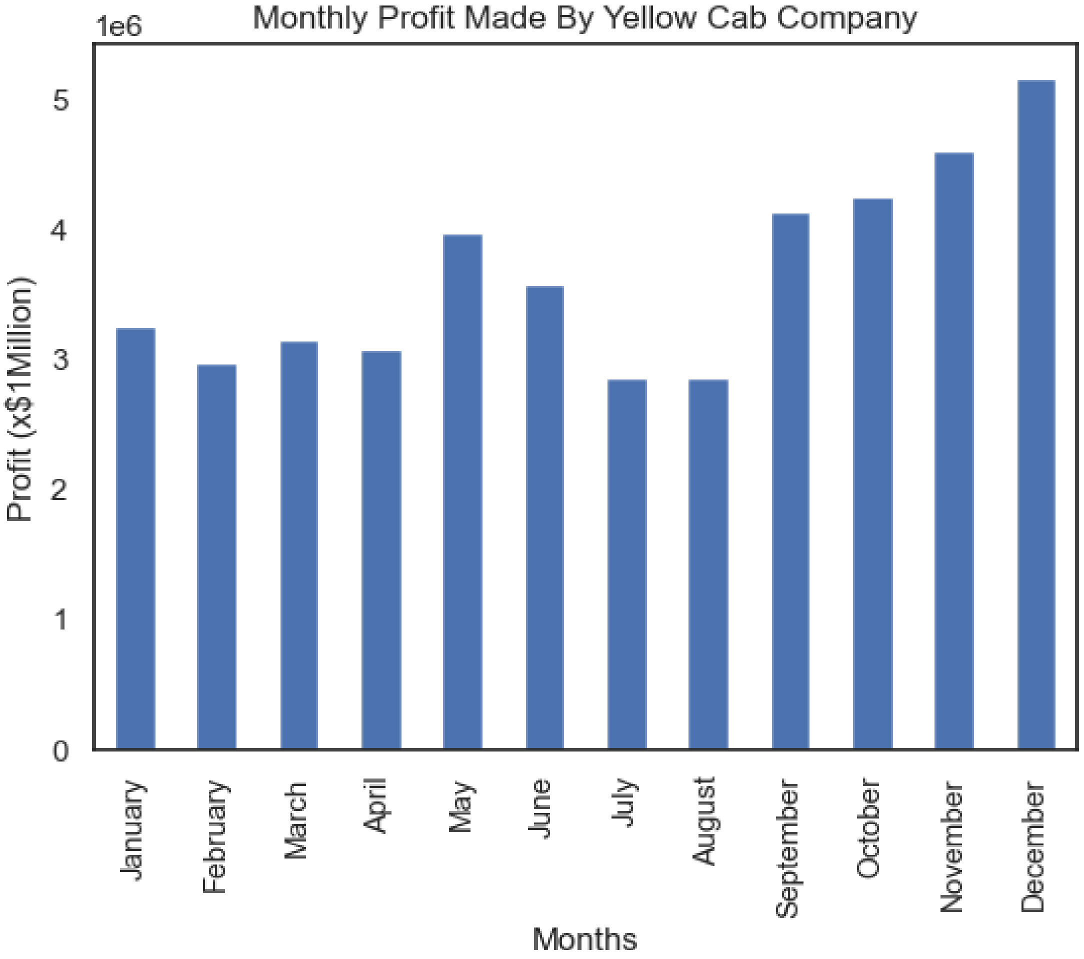


Looking at the graphs above, we can see both cab companies make their highest daily profits on Friday, Saturday and Sunday

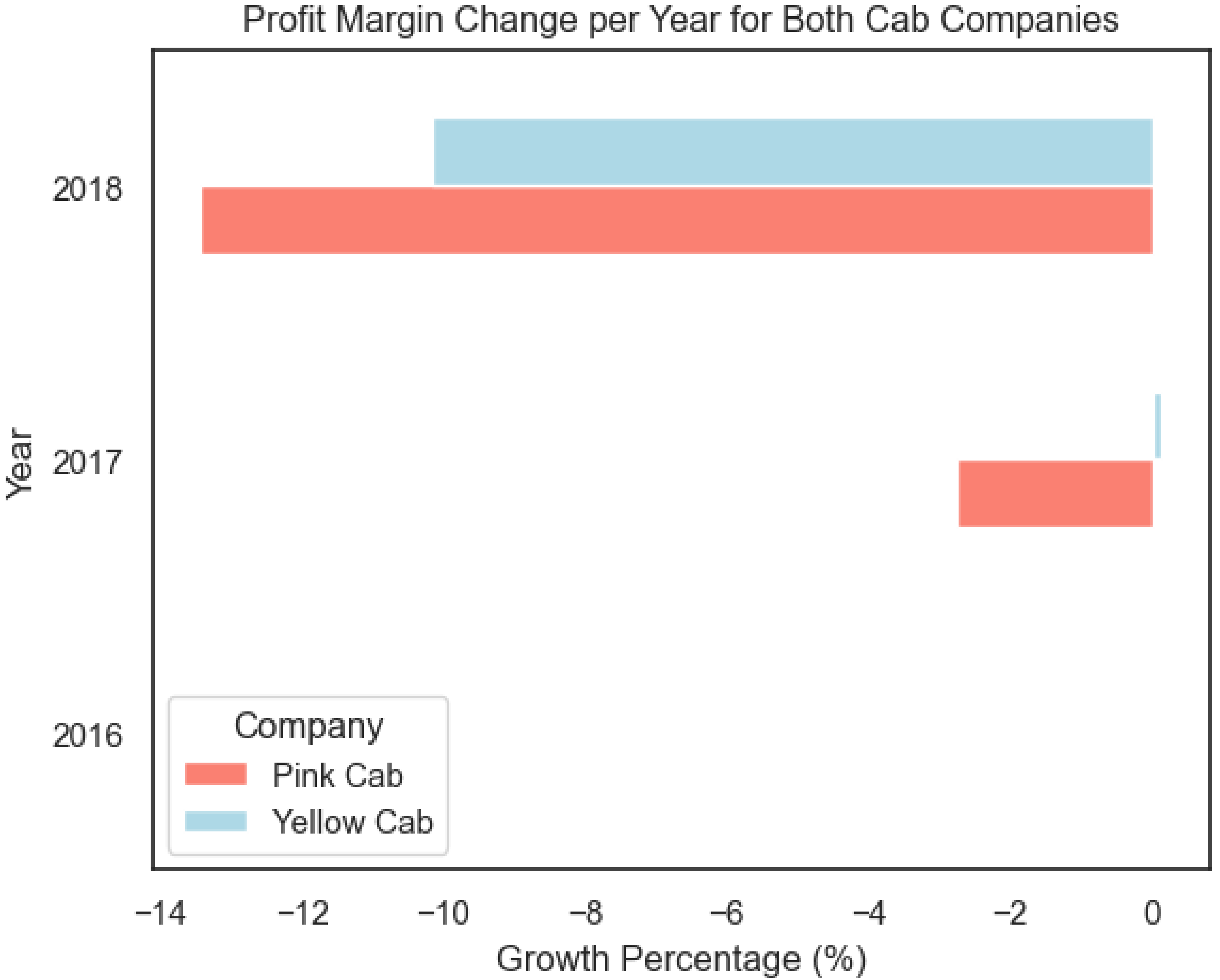


From the graph above, it is clear that the Yellow Cab company has a higher average daily profit than the Pink Cab company.





From the graphs above, we can see both Cab companies make their highest average profit in the month of December. This makes sense because in an earlier graph, December was the month with the highest number of rides taken.



From the bar chart, we can see both cab companies experience a negative profit margin change with the Pink Cab company experiencing the highest. This indicates a larger loss of customers and very low patronage of the Pink Cab company compared with the Yellow Cab company.

Recommendations

After conducting a thorough Exploratory Data Analysis, Statistical Analysis and Data Visualization, I recommend the Yellow Cab company as a better investment option due to the following reasons:

- Customer Retention - The Yellow Cab saw more customers returning to use its services. This indicates a low churn rate.
- Cab Usage - The Yellow Cab company experiences a higher daily, monthly and yearly can usage than the Pink Cab company. Also, the Yellow Cab company has a higher total number of rides taken.
- Charge - The Yellow Cab company charges higher fares than the Pink Cab company.
- Profit - The Yellow Cab company makes more profit on a daily, monthly and yearly basis. Also, the Yellow Cab company makes more profit over a three year period. Thus I strongly believe the Yellow Cab company is a more profitable business to invest in.

However, a risk assessment analysis should be conducted to select the best business strategy. This is because of a presence of high positive skewness which suggests large positive returns but a higher risk. Also, the Yellow Cab company experiences a negative value as its profit margin change in 2018

Thank You

Data Glacier

Your Deep Learning Partner