Data Glacier

Your Deep Learning Partner

Exploratory Data Analysis

G2M Insight for Cab Investment Firm

By

Biswadip Bhattacharyya

Date : 21st June 2024

Agenda

- Problem Statement
- Approach
- EDA
- Hypothesis Testing
- Conclusion

Problem Statement

- XYZ is a private firm in the U.S 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 an actionable insights to XYZ company to make decision for investing in right company.

The analysis has been divided into 3 parts:

- Data information
- EDA
- Hypothesis Testing

Data Glacier
Your Deep Learning Partner

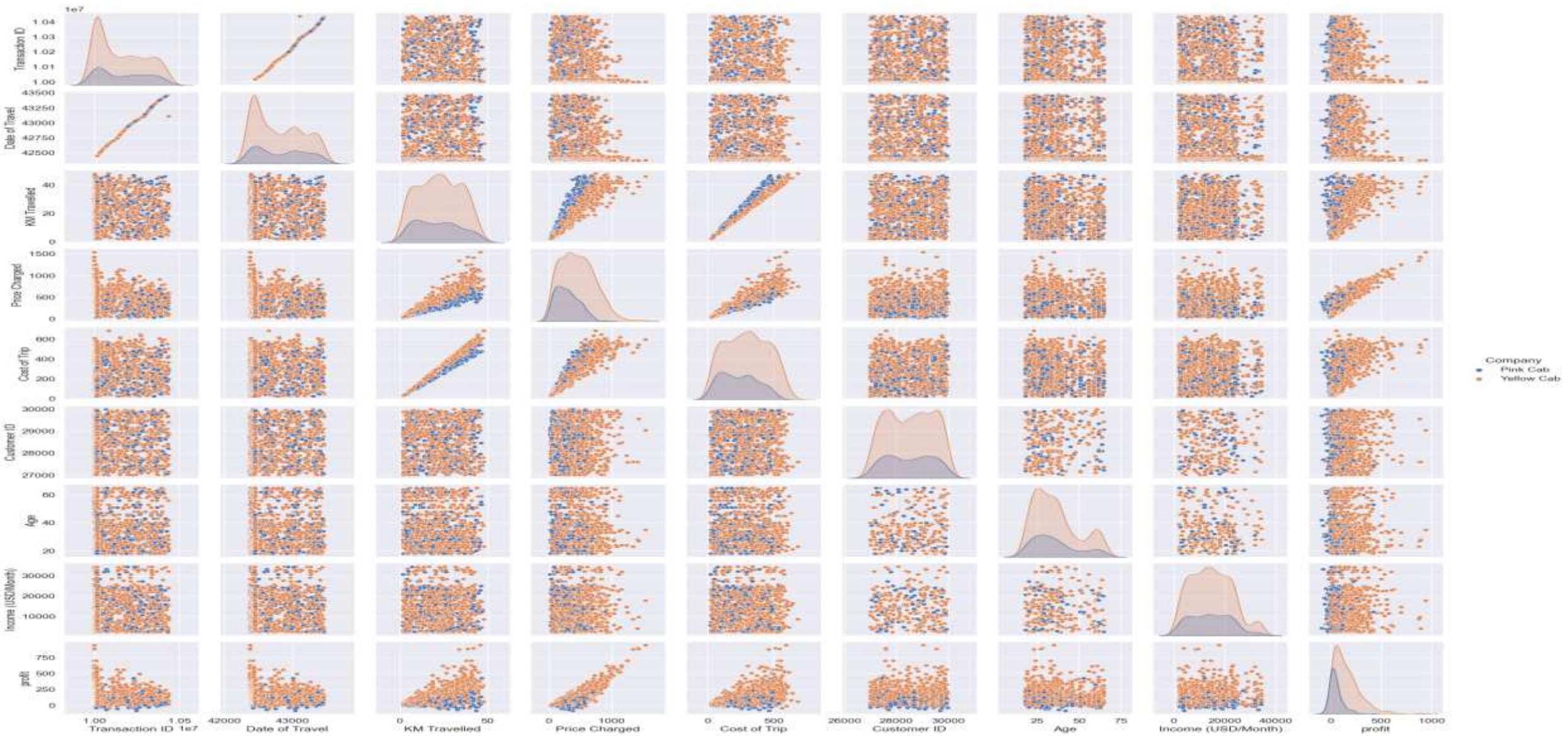
Data Information

Data Information

There are 4 datasets for this analysis:

- Cab_Data.csv- This dataset contains the details of transactions for two cab companies.
- Customer_ID.csv- This dataset contains the customer's demographic details.
- Transaction_ID.csv- This dataset contains the details of customer's payment mode.
- City.csv- This file contains the list of US cities, their populations and number of cab users.

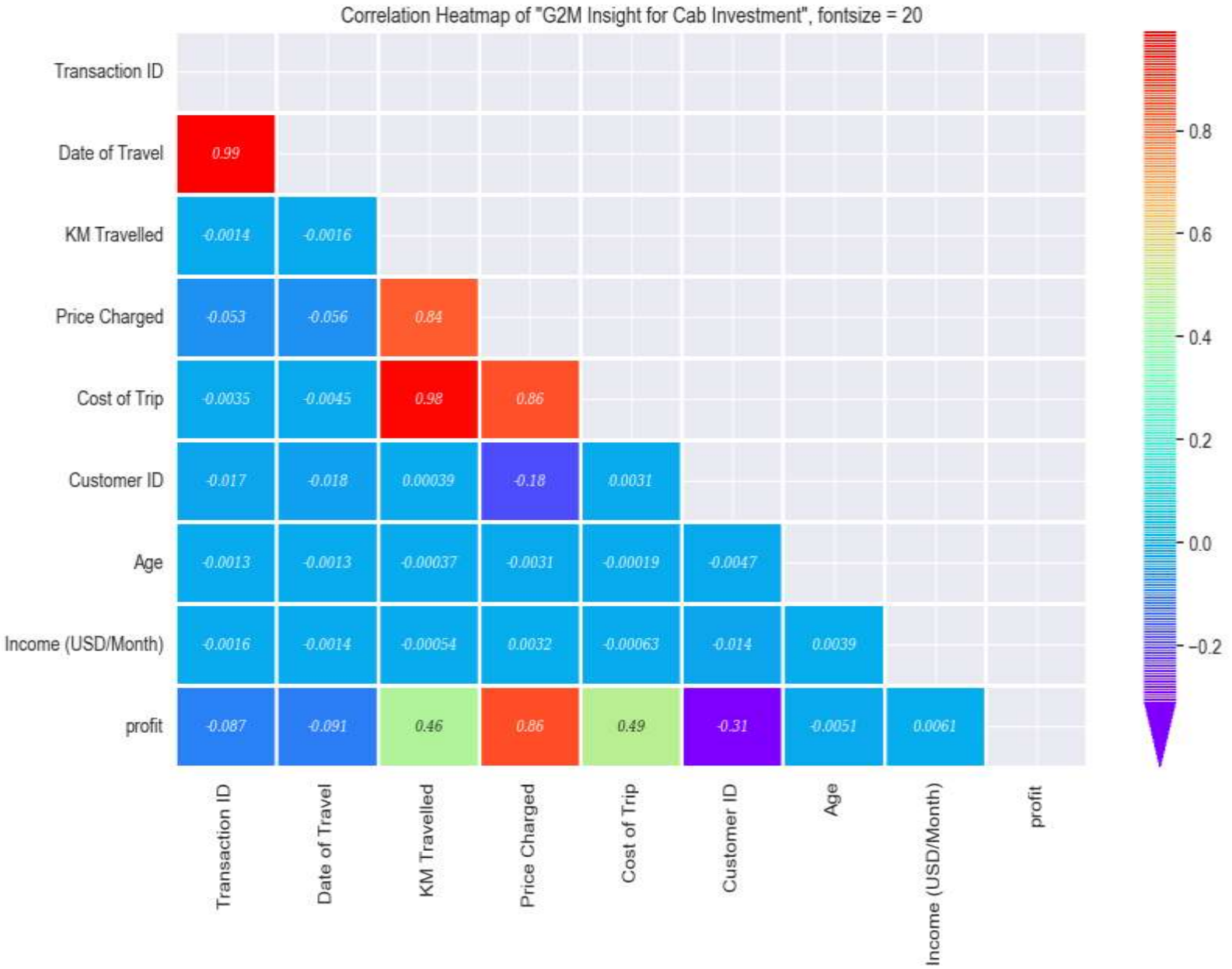
Relationship between variables



Correlation between variables

As we can see there is a strong correlation between:

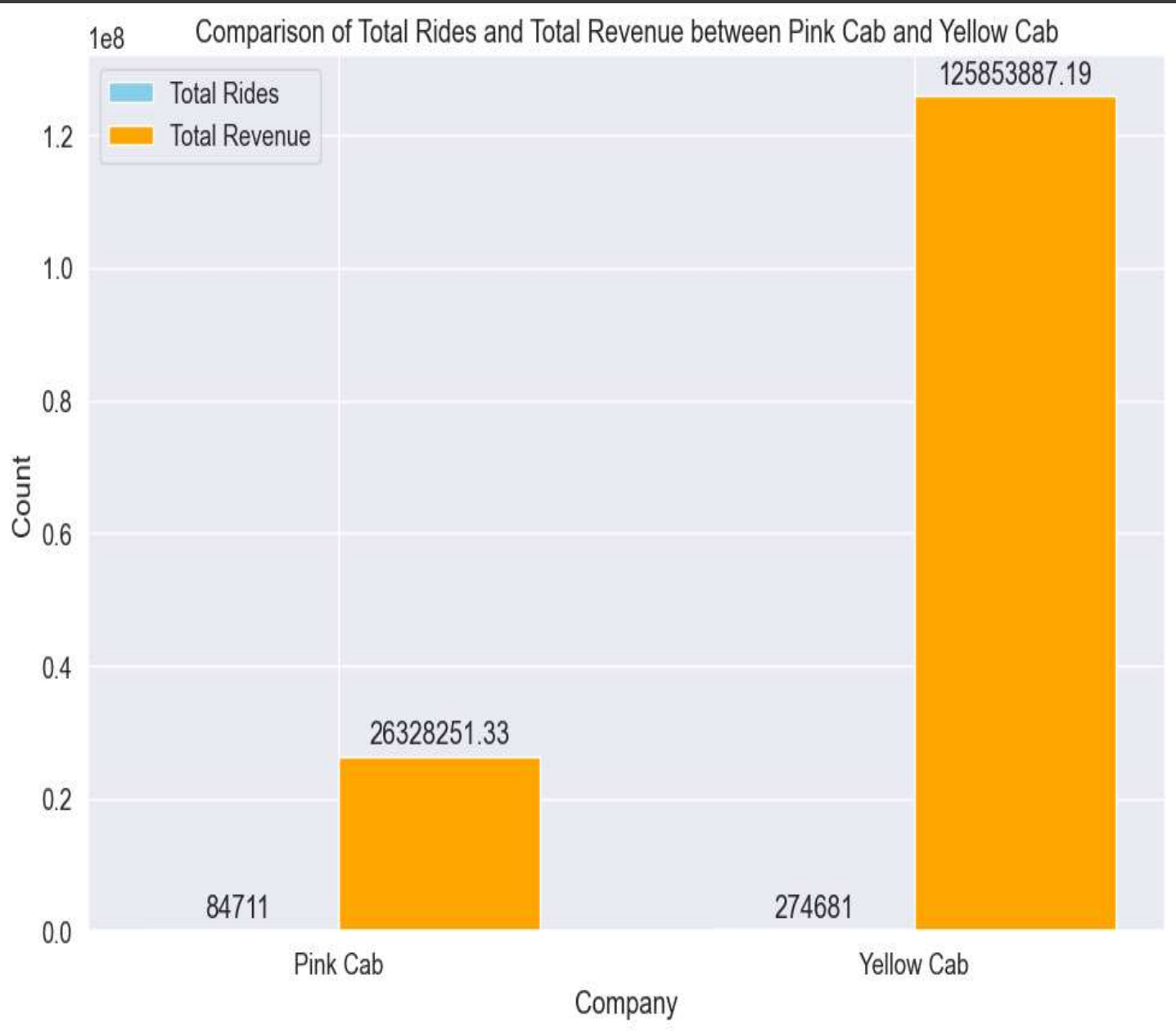
- Cost of trip VS KM travelled
- Cost of trip VS price Charged
- Price Charged VS profit



Exploratory Data Analysis (EDA)

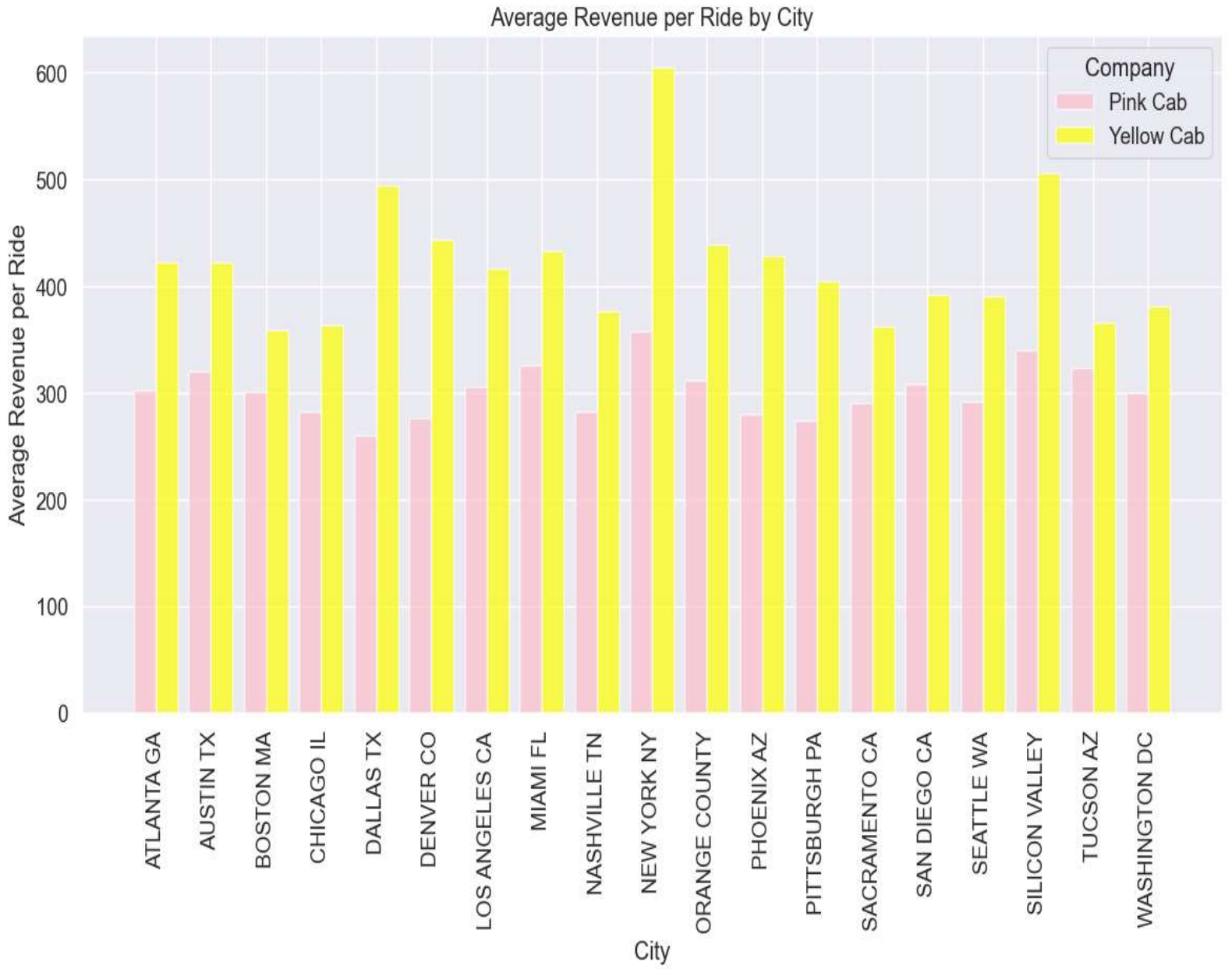
Comparison of total rides and total revenue between yellow can and pink cab

We can see that if we compare between the pink cab and yellow cab from 2016 - 2018, then we can clearly see that yellow cab has way more number of total rides and total revenue as compared to pink cab



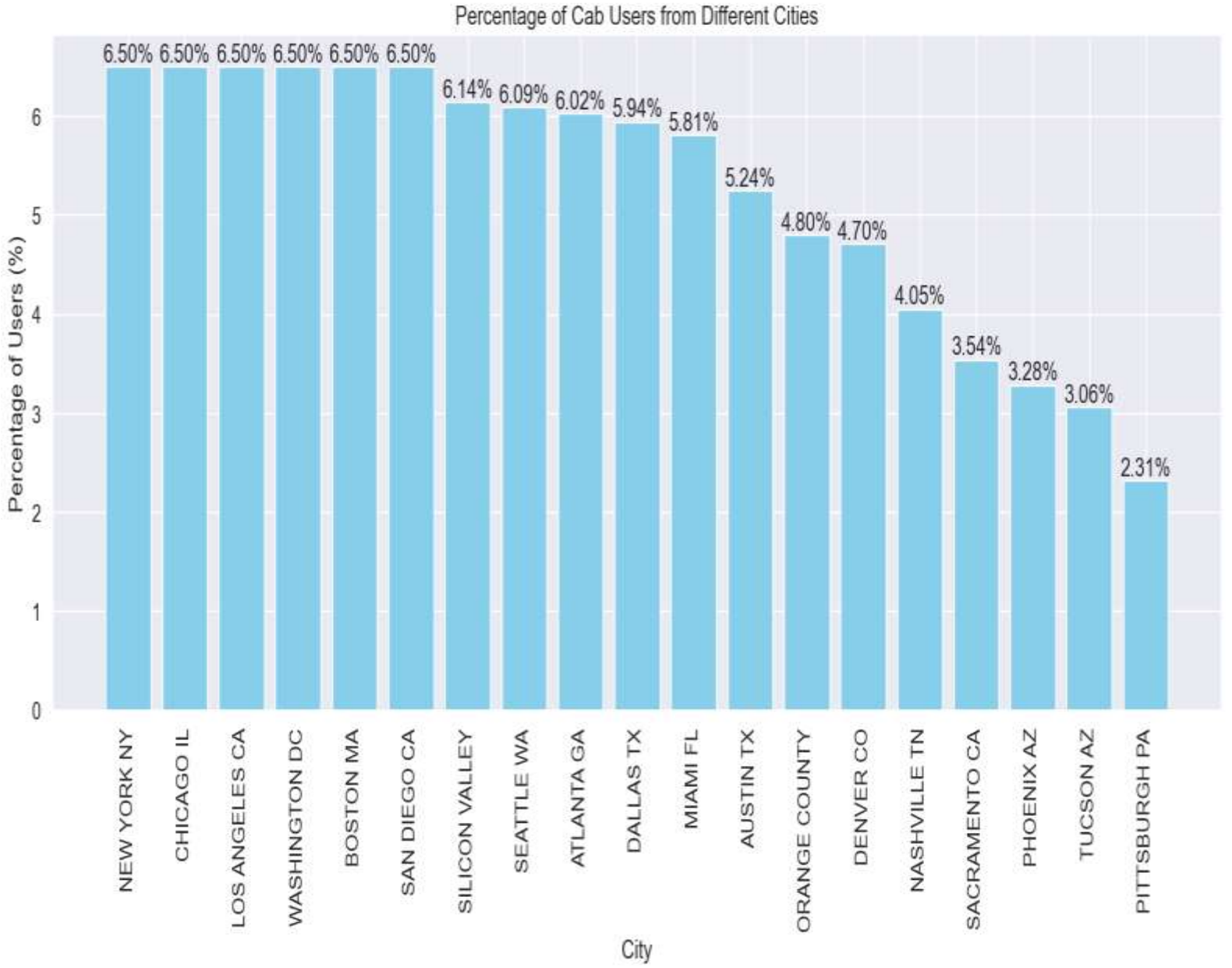
Average revenue per ride by city

According to this graph we can clearly see that yellow cab has more average revenue per ride in every city as compared to pink cab, and New York has the highest average revenue as compared to other cities for both cab companies.



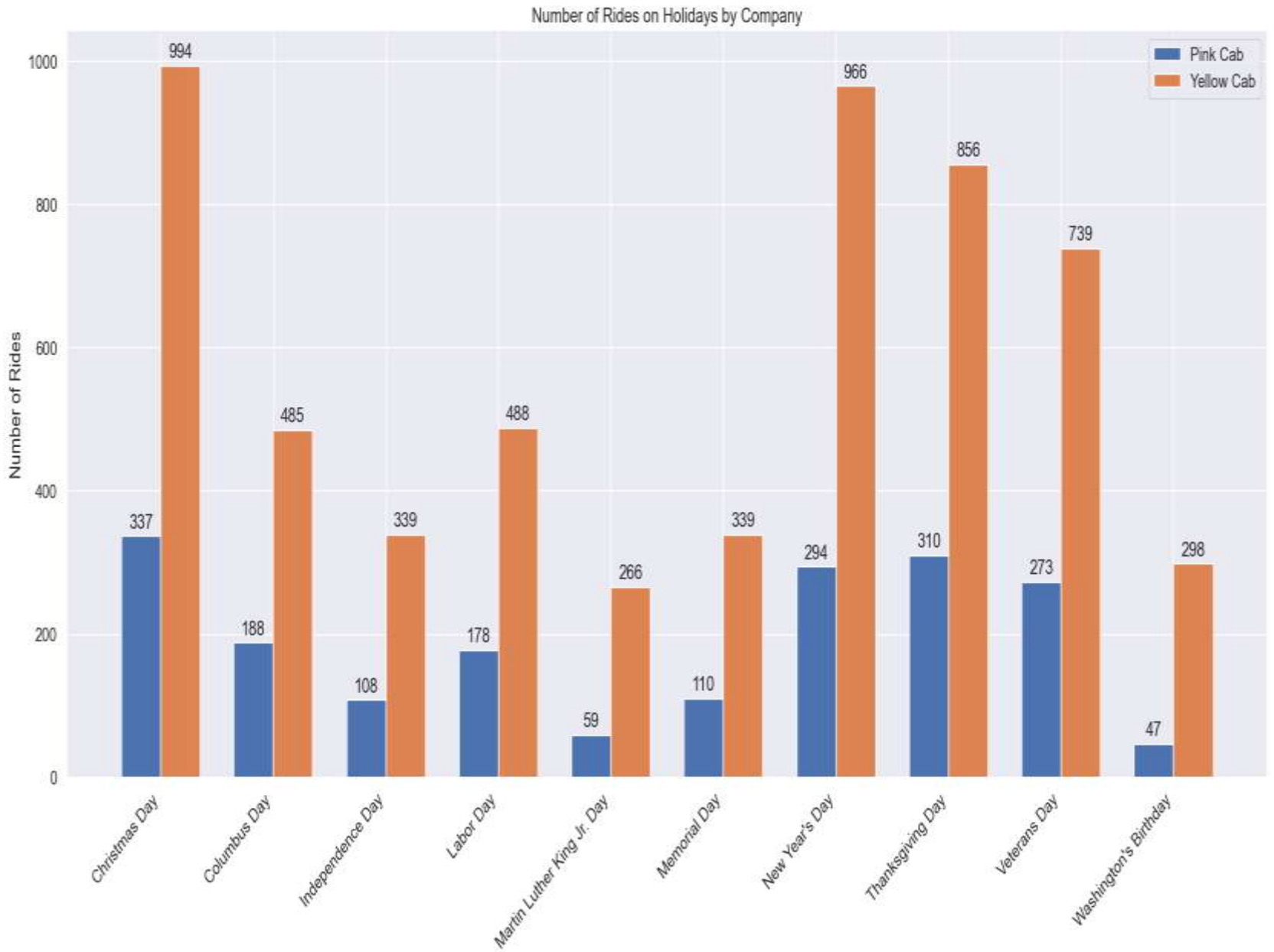
Percentage of cab users from different cities

We can see from this graph that New York, Chicago, Los Angeles, Washington DC, Boston and San Diego has the most cab users as compared to different cities, whereas Pittsburgh has lowest cab users.



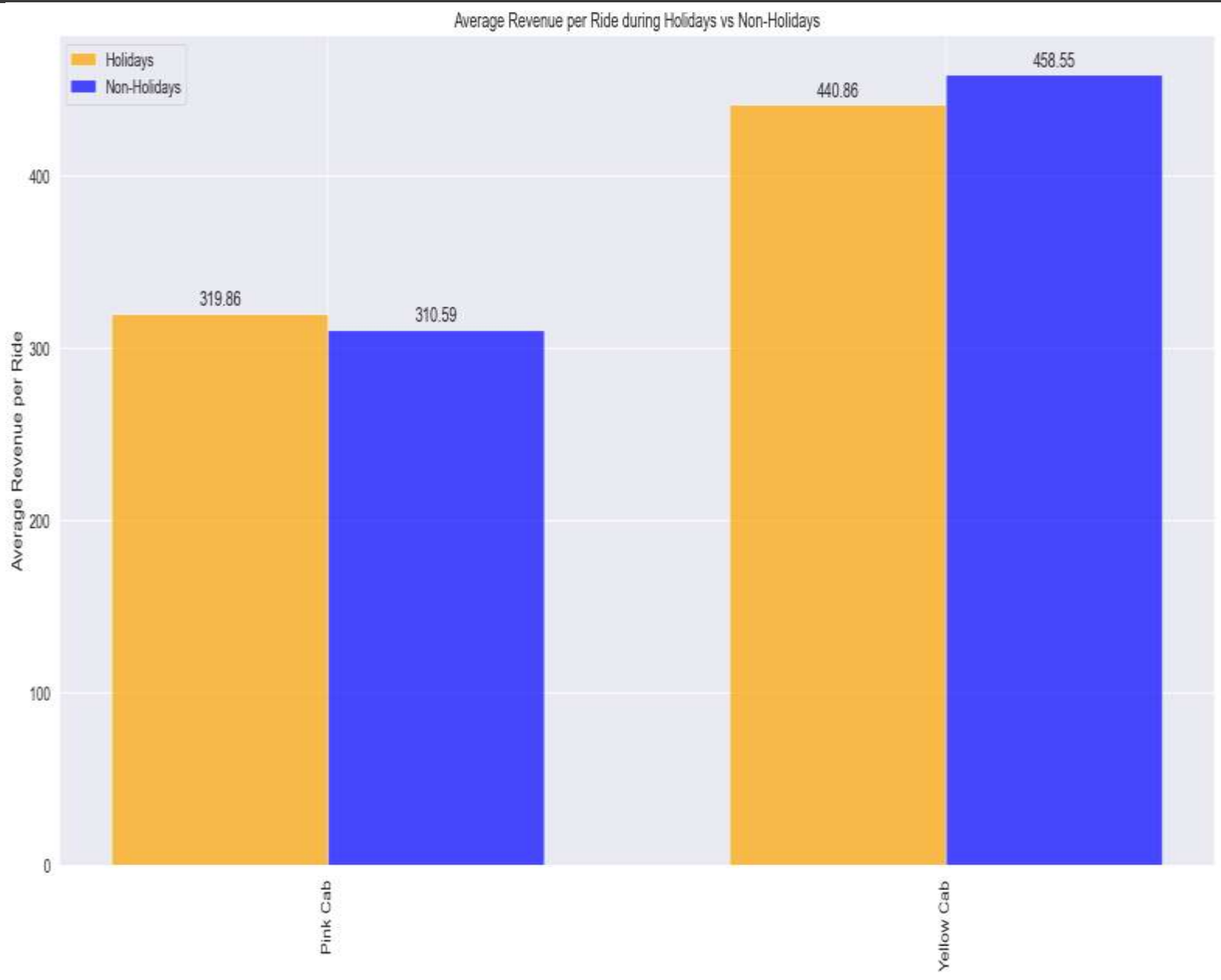
Number of rides on holidays by company

From this bar plot we can clearly see that yellow cab dominates in every holidays and in Christmas day we can see that the number of rides is highest as compared to other holidays for both the cab companies.



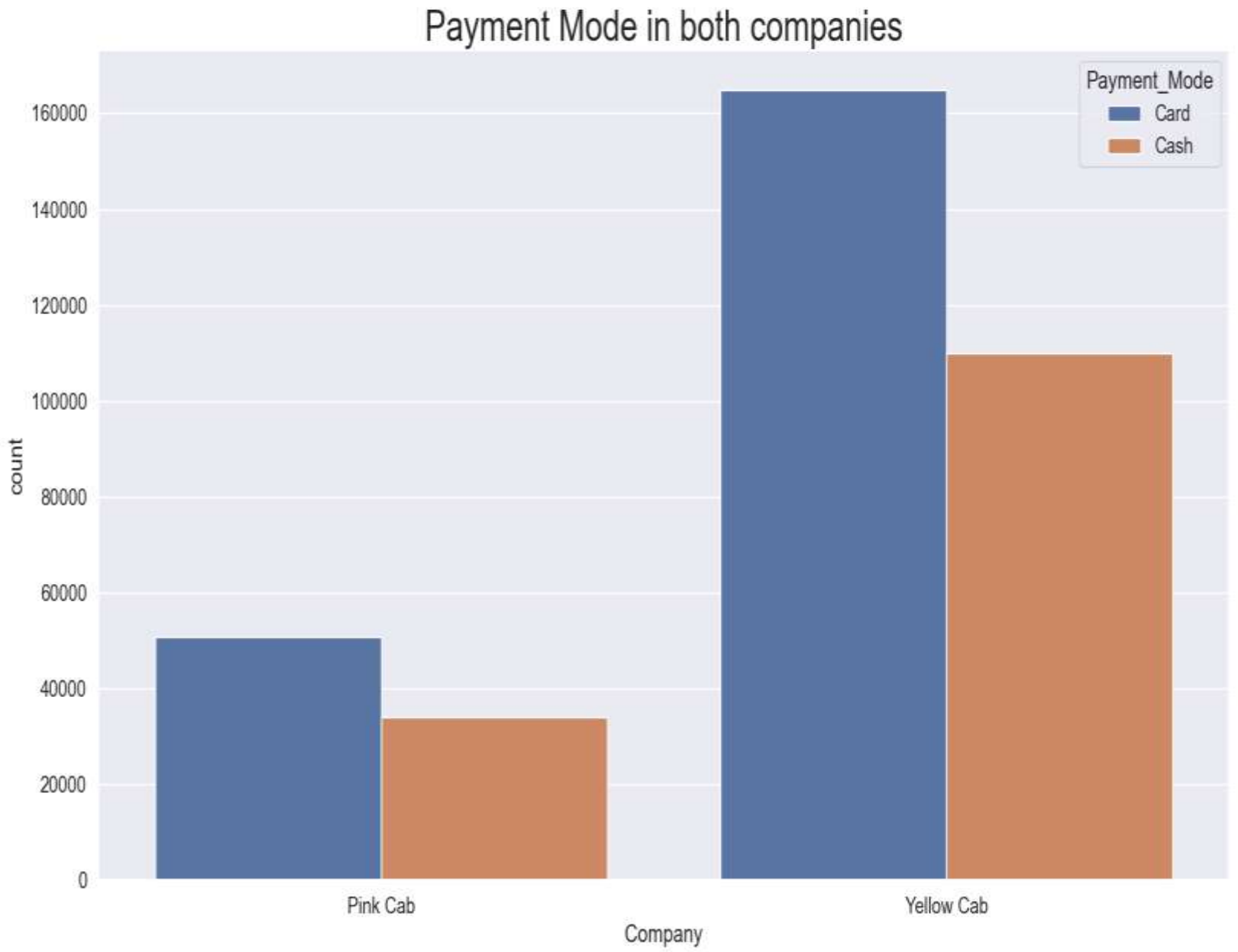
Average revenue per ride during holidays vs non-holidays

As we can see from this plot, yellow cab has the highest average revenue per ride in both holidays and non-holidays and have slightly higher average in non-holidays than holidays, whereas pink cab has higher average revenue in holidays than non-holidays.



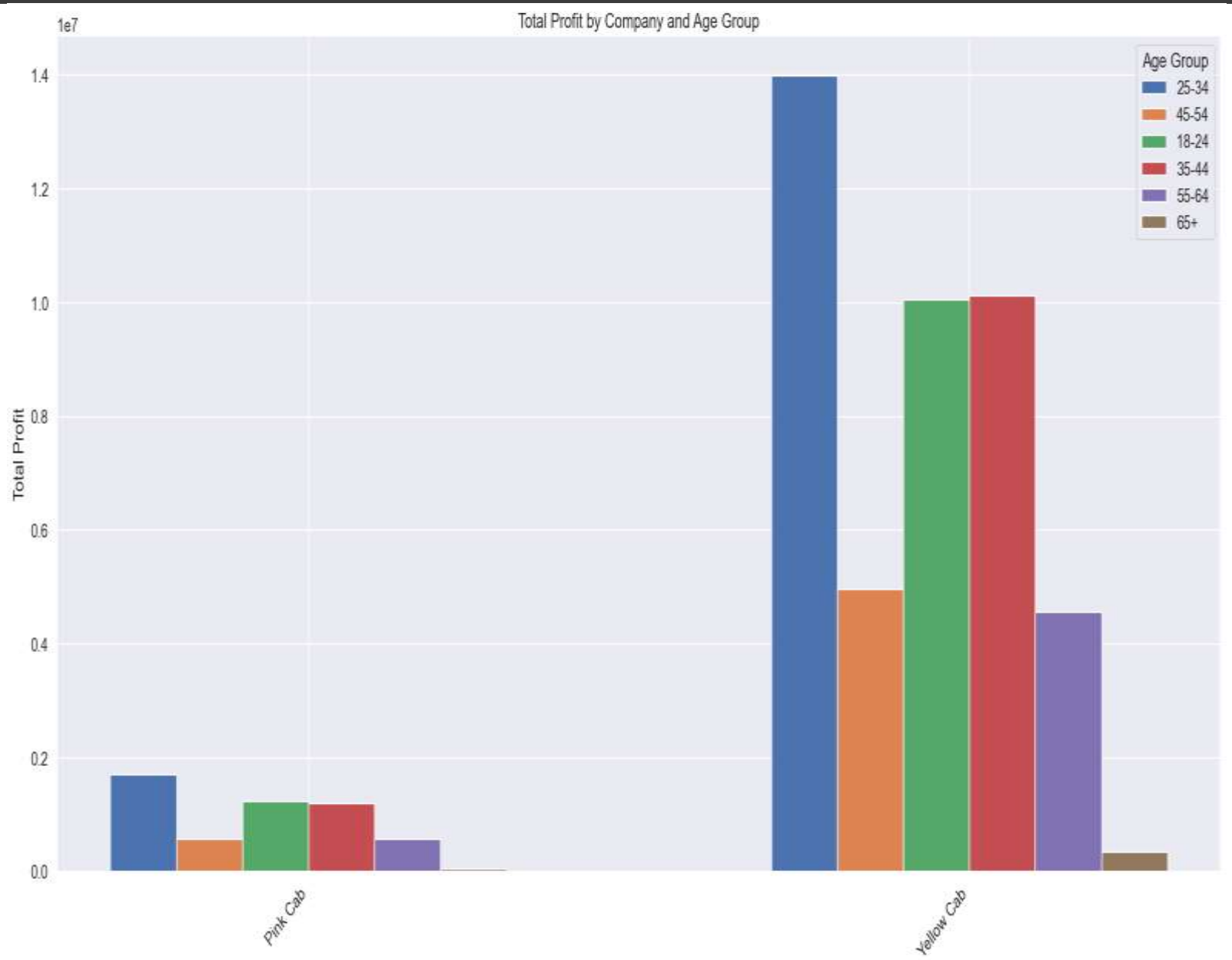
Payment mode in both companies

In case of payment mode, customer use card more than cash in case of both cab companies.



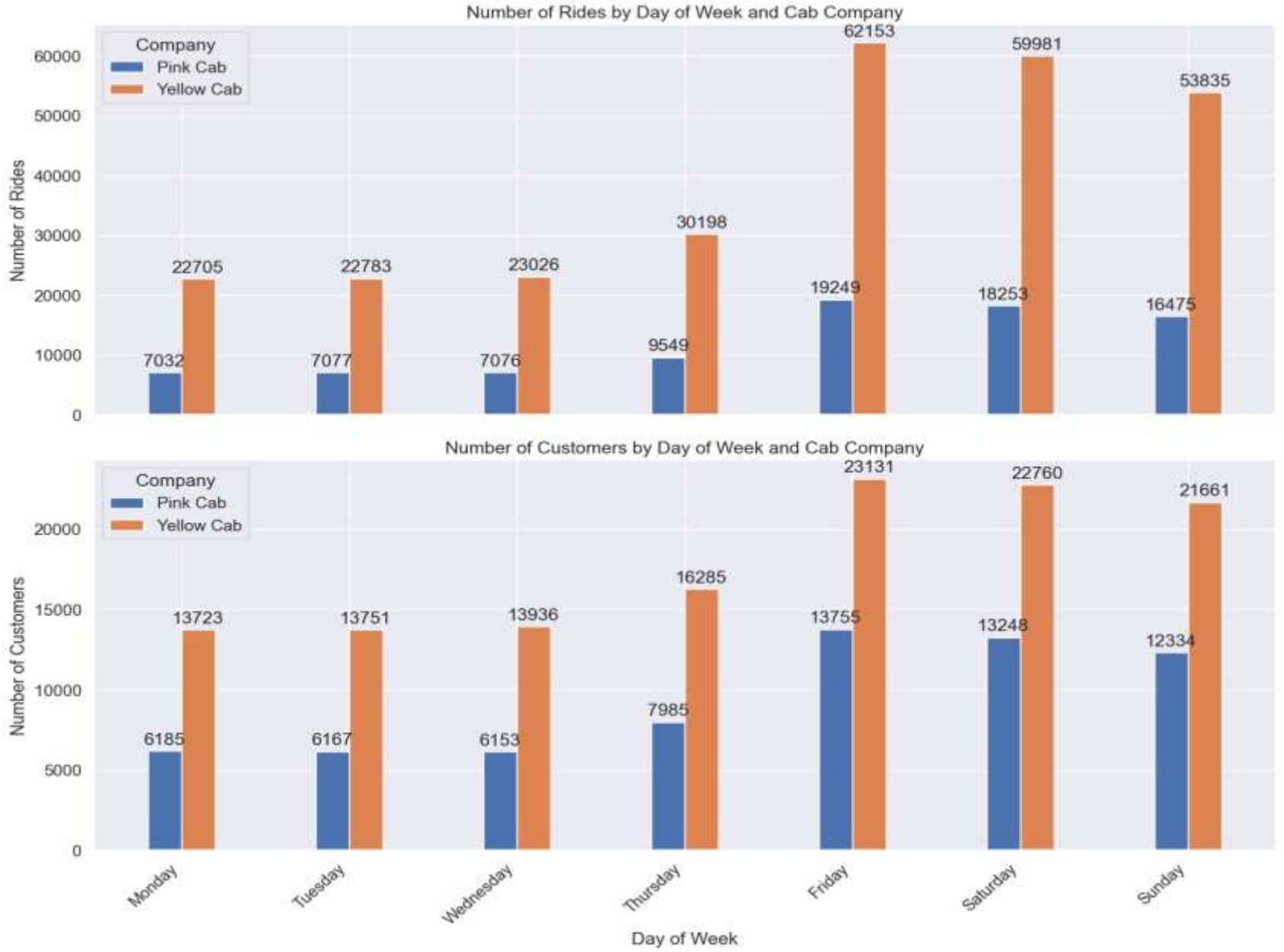
Total profit by company and age group

In this case we can clearly see that yellow cab dominates in total profit in all age groups. For yellow cab, age group 25-34 has the highest profit, whereas age group 65+ is the lowest.
For pink cab age group 25-34 has highest profit and age group 65+ has lowest profit just like yellow cab.



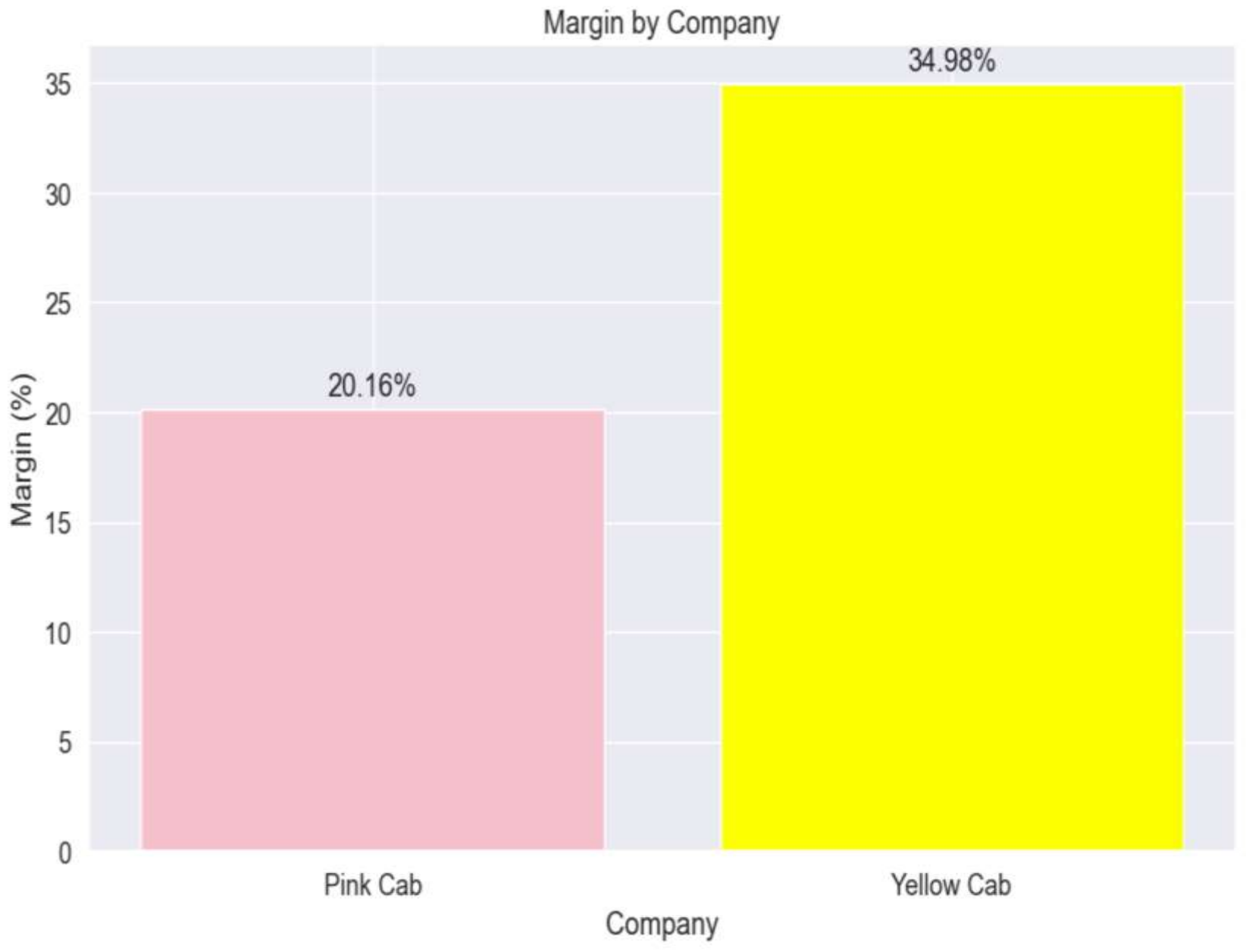
Number of rides and customers by day of week

From this plot we can see that customers use cab more in Friday, Saturday and Sunday where Friday being the busiest day for both the cab companies.



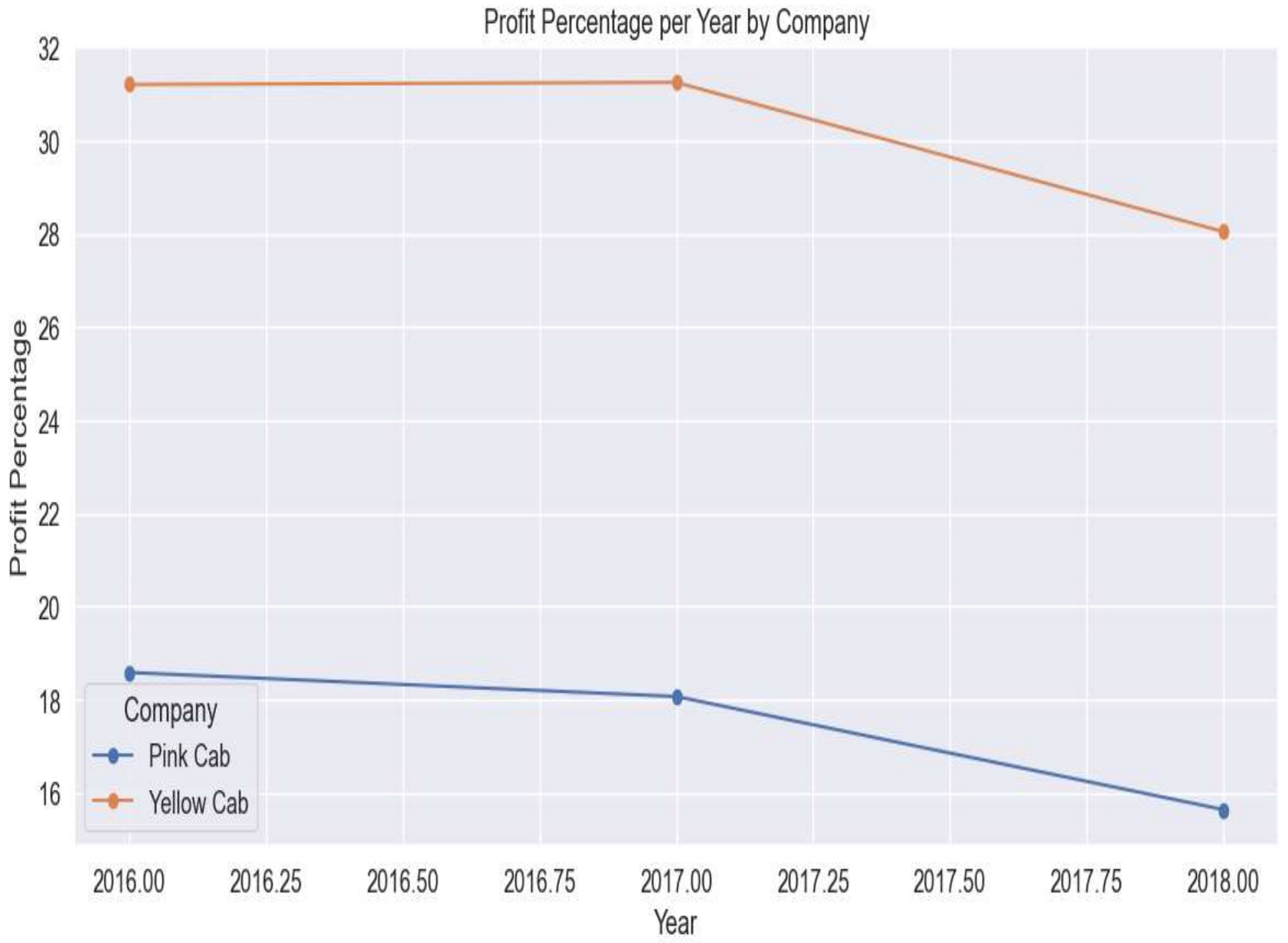
Profit Margin

From this bar plot we can conclude the yellow cab has the highest profit margin of 34.98% as compared to the pink cab of profit margin of 20.16%.



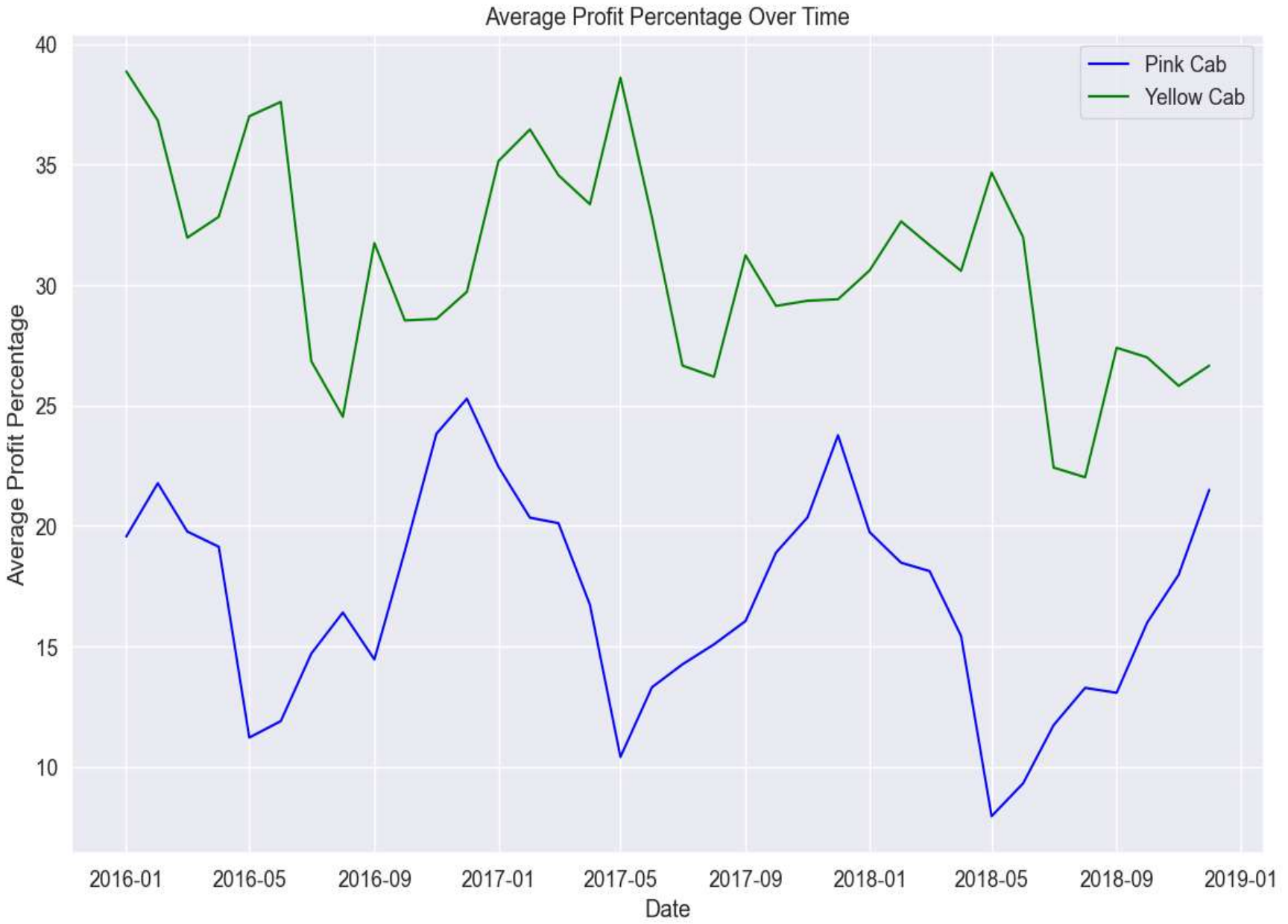
Profit percentage per year

We can see that profit percentage for both cab companies were slightly stable until 2017, after which it falls drastically.



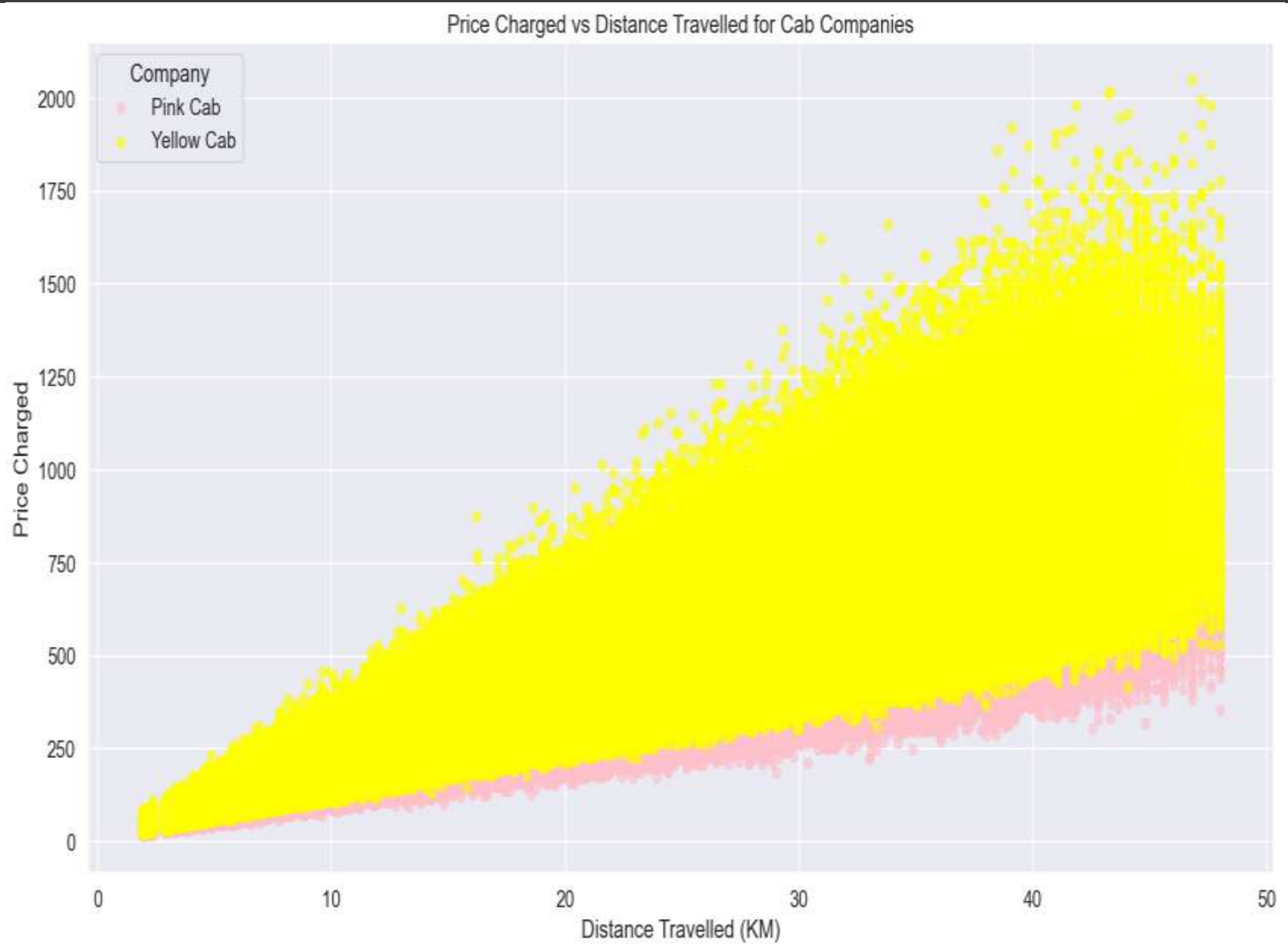
Average profit percentage over time

- From this plot we can see that Yellow Cab shows a general pattern of volatility with several peaks and troughs.
- There is a noticeable drop around early 2017, but the profits recover and stabilize between 30% and 35% towards the end of 2018.
- Pink Cab's profits also fluctuate but show an upward trend towards the end of the period, particularly after mid-2018, reaching around 20%.



Price charged vs Distance

- Yellow Cab charges higher prices overall compared to Pink Cab. This is evident as the yellow points are spread higher along the price axis than the pink points.
- For both companies, the price charged increases as the distance traveled increases. However, the rate at which the price increases differs significantly between the two companies.
- Yellow Cab shows a steeper increase in price with distance, indicating higher fares per kilometer traveled compared to Pink Cab.



Hypothesis Testing

Hypothesis 1-Number of rides varies by Days of the week

H0: Number of rides doesn’t varies by days of the week

H1: Number of rides varies by days of week

From the result we can see that p value is less than 0.05 which means that it rejects the null hypothesis i.e the number of rides varies by the days of the week.

```
ANOVA results for Number of Rides by Day of the Week:
F-statistic: 13.761060336399861
p-value: 0.00298304141605062 55
```

Hypothesis 2- "Profit Margin Proportionally Increases with Number of Customers"

H0: Profit margin doesn’t proportionally increase with number of customers
H1: Profit margin proportionally increase with number of customers

```
ANOVA results for Margin by Number of Customers:
F-statistic: 505.8577222909265
p-value: 3.933886834964757e-108
```

From the result we can see that p value is less than 0.05 which means that it rejects the null hypothesis i.e Profit margin proportionally increase with number of customers

Hypothesis 3-" Is there any difference in Profit regarding Age"

H0: There is no difference in profit regarding age
H1: There is difference in profit regarding age

From the result we can see that p value is less than 0.05 which means that it rejects the null hypothesis i.e there is difference in profit regarding age.

```
ANOVA results for Profit by Age Group:
F-statistic: 16.503495089979438
p-value: 2.5049474611508 42e-16
```

Hypothesis 4-"Profitability is Affected by the Distance Travelled"

H0: Profitability doesn’t affect by the distance travelled
H1: Profitability does affect by the distance travelled

From the result we can see that p value is less than 0.05 which means that it rejects the null hypothesis i.e profitability does affect by the distance travelled

```
ANOVA results for Profitability by Distance Travelled:
F-statistic: 37079.89312150977
p-value: 0.0
There is a significant difference in profitability across different distance categories.
```

Hypothesis 5: Customer Age Affects the Average Revenue per Ride

H0: Customer age doesn’t affect the average revenue per ride
H1: Customer age does affect the average revenue per ride.

```
ANOVA Results for Average Revenue per Ride across Age Groups:
F-statistic: 4.101200209625385
p-value: 0.0025217933585060873

There is a significant difference in average revenue per ride across age groups.
```

From the result we can see that p value is less than 0.05 which means that it rejects the null hypothesis i.e Customer age affect the average revenue per ride

Conclusion

After all this analysis we can say that Yellow cab is better than Pink cab based on this factor

- Higher profit margin
- Has more users
- Has more transactions per year

Thank You