Data Glacier

Your Deep Learning Partner

# Exploratory Data Analysis

## G2M – Cab Investment Firm Project

**20 April 2023**

**Salih Eren Yüzbaşıoğlu**

# Executive Summary

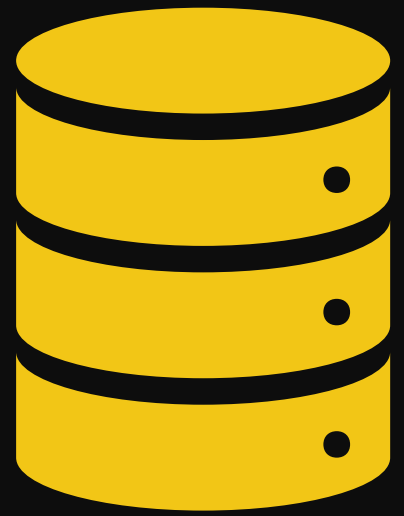
The goal of this project is to help the private firm XYZ make an informed investment decision in the cab industry by analyzing multiple data sets related to two cab companies.
This presentation summarizes the analysis and recommendations on which company is performing better and is a better investment opportunity for XYZ.

# Executive Summary

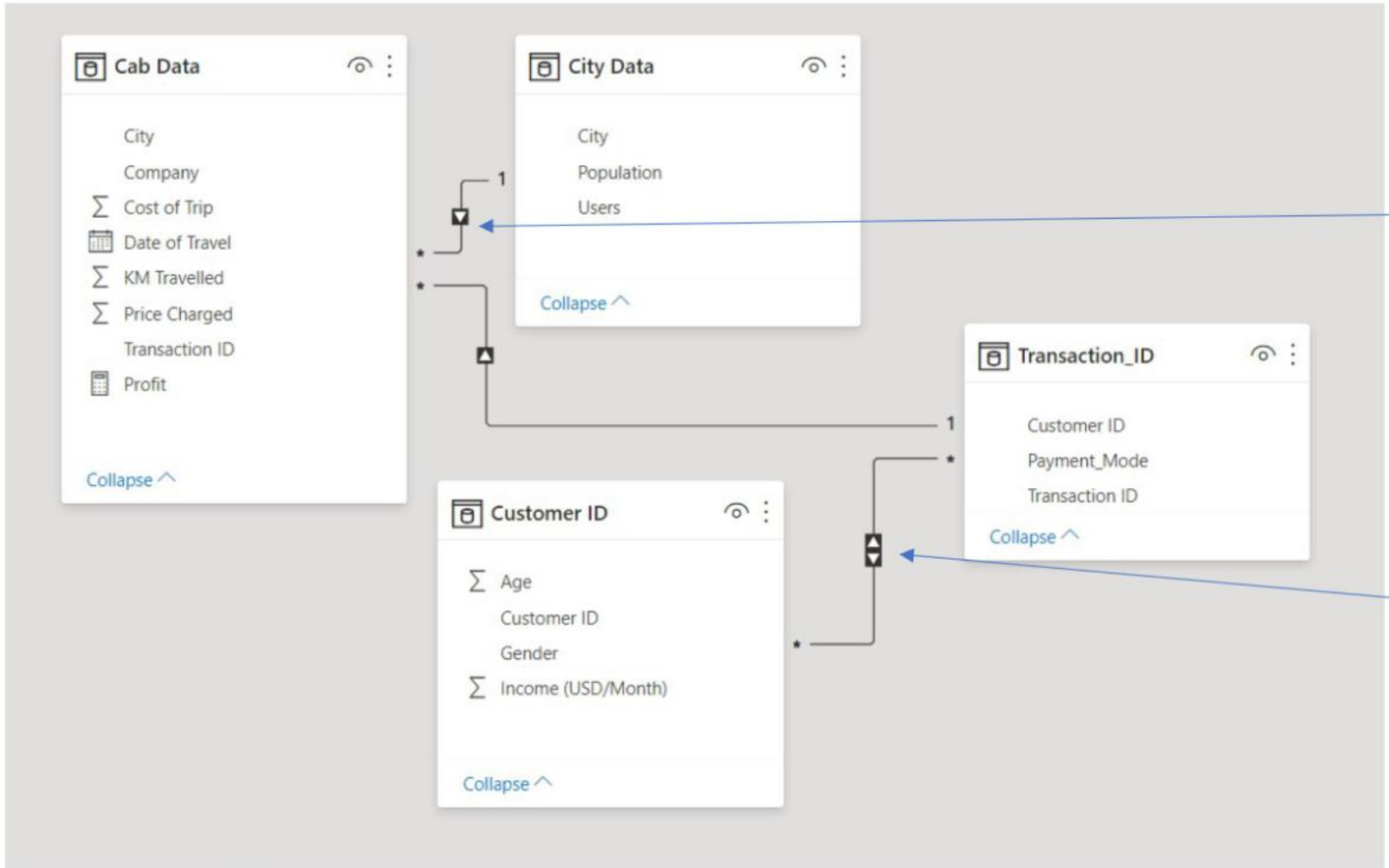
- XYZ is a private equity firm in the US. Due to remarkable growth in the Cab Industry in the last few years and key players in the market, it is planning for an investment in the Cab industry.
- **Objective**: Provide insights to help XYZ firm in identifying the right comp any for making an investment.

# Datasets

- **Cab_Data.csv**
  This file includes details of transactions for 2 cab companies
- **Customer_ID.csv**
  This is a mapping table that contains a unique identifier that 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 file contains a list of US cities, their population, and the number of cab users



# Table Relationship




Cab data and City Data tables are linked by similar fields city with one to many.
Similarly Cab Data and Transaction ID are linked by Transaction ID with one to many relationship

Table Transaction ID is linked to Customer ID with many to many relationship

# Approach

**Cleaning & Merging the Data**
**In this part, duplicate columns and rows** are removed and a master dataset is created.

**Exploratory Data Analysis & Visualization**
**In the second part, a general analysis and** visualization of the master data is done through EDA to notice possible correlations and generate some hypotheses.

**Hypothesis Testing**
**In the last part, the hypotheses that** were generated are tested and visualized

Data Glacier
Your Deep Learning Partner

# 1. Cleaning and Merging

# Changes Done

- Travel Date column is changed into actual date values from time periods.
- Transaction and Customer Id data frames merged.
- Users turned into integer.
- Profit, month, year, day columns added.
- Cab data merged.
- All data turned into 2 data frames named PinkCab and YellowCab.
- All data is exported.

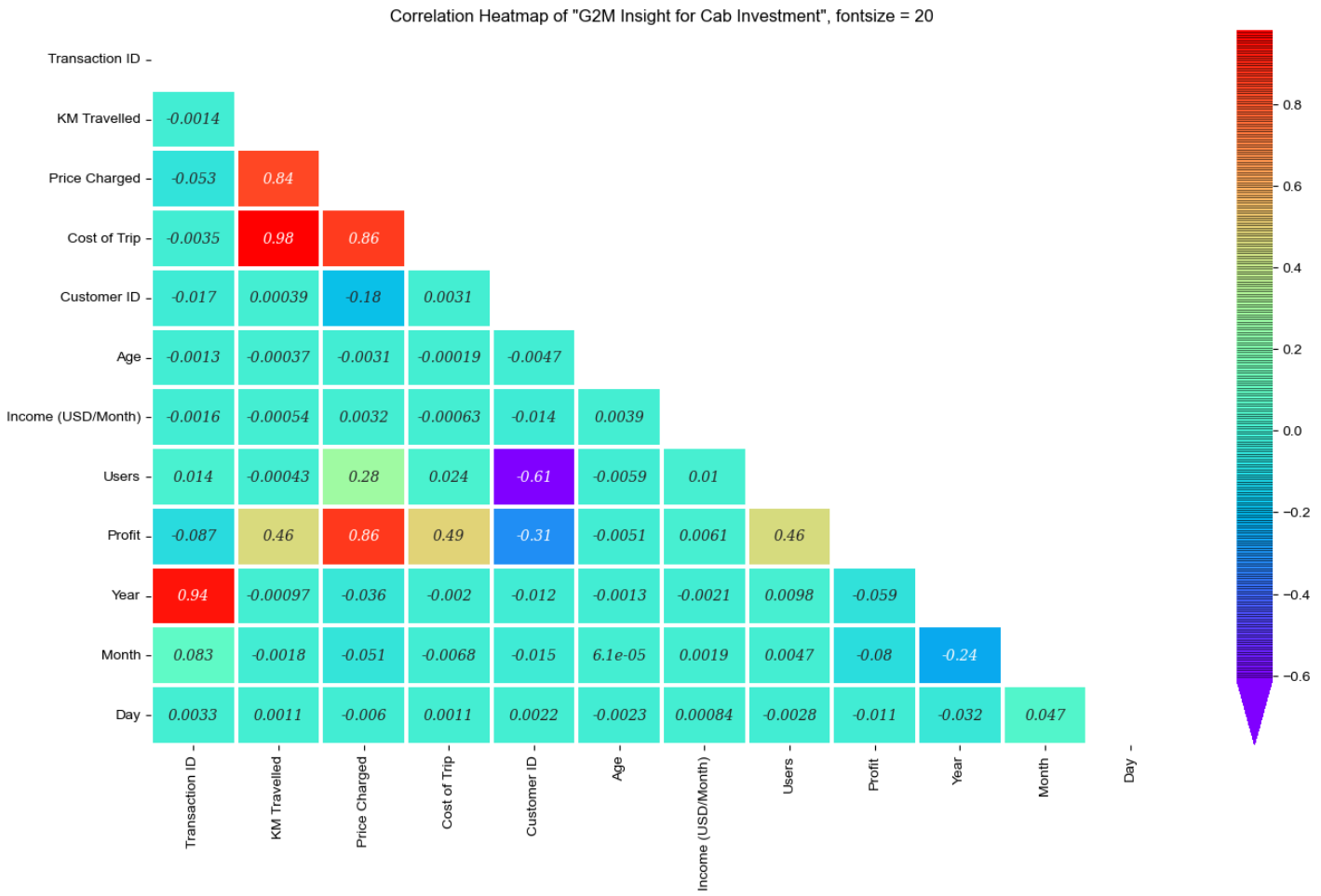
# 2.EDA AND VISUALIZATION

Data Glacier

Your Deep Learning Partner

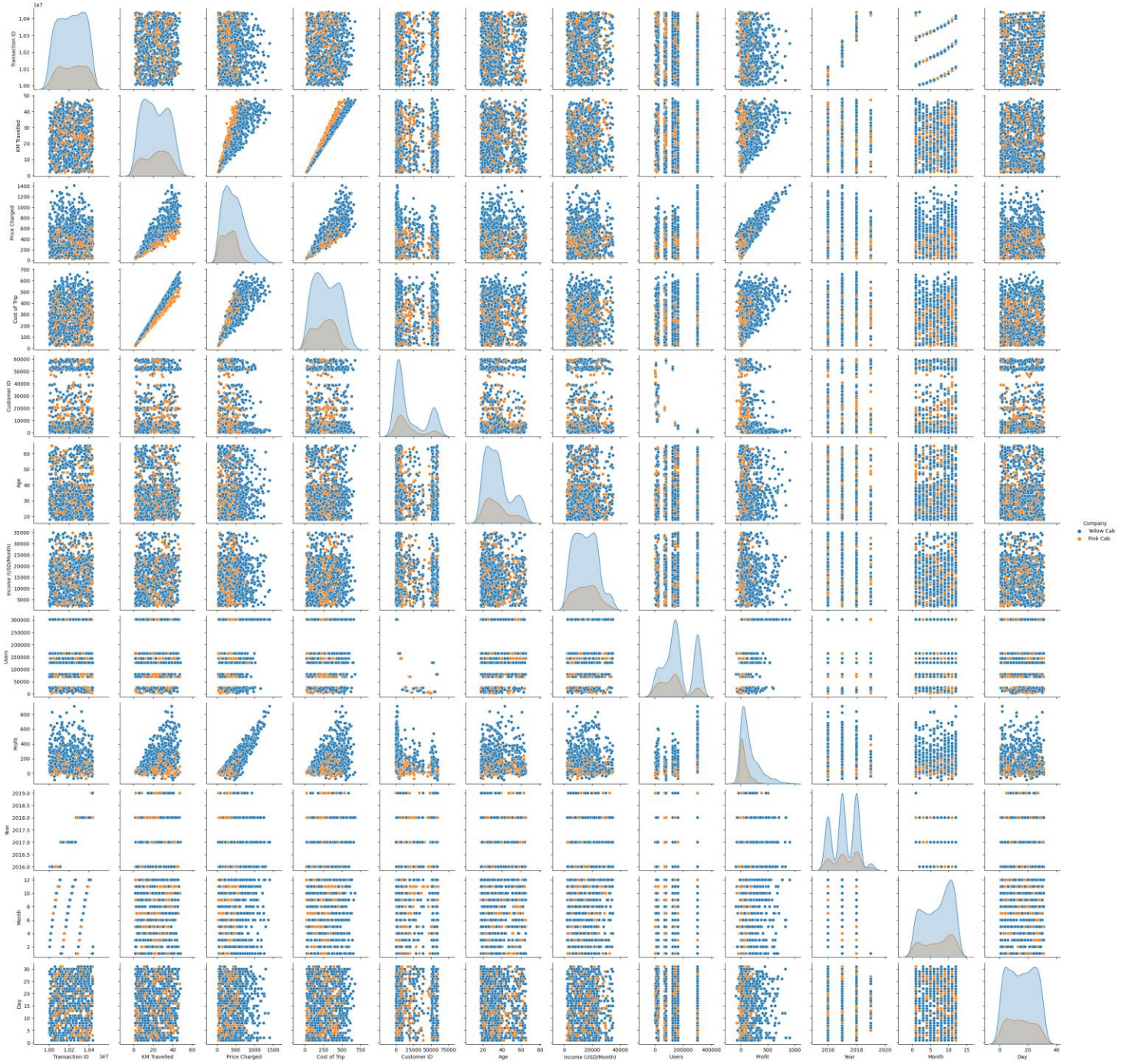
# HEATMAP

As we can see there is strong Correlation between
◉ Population vs Users ◉ Price Charged vs Cost of Trip vs KM Travelled

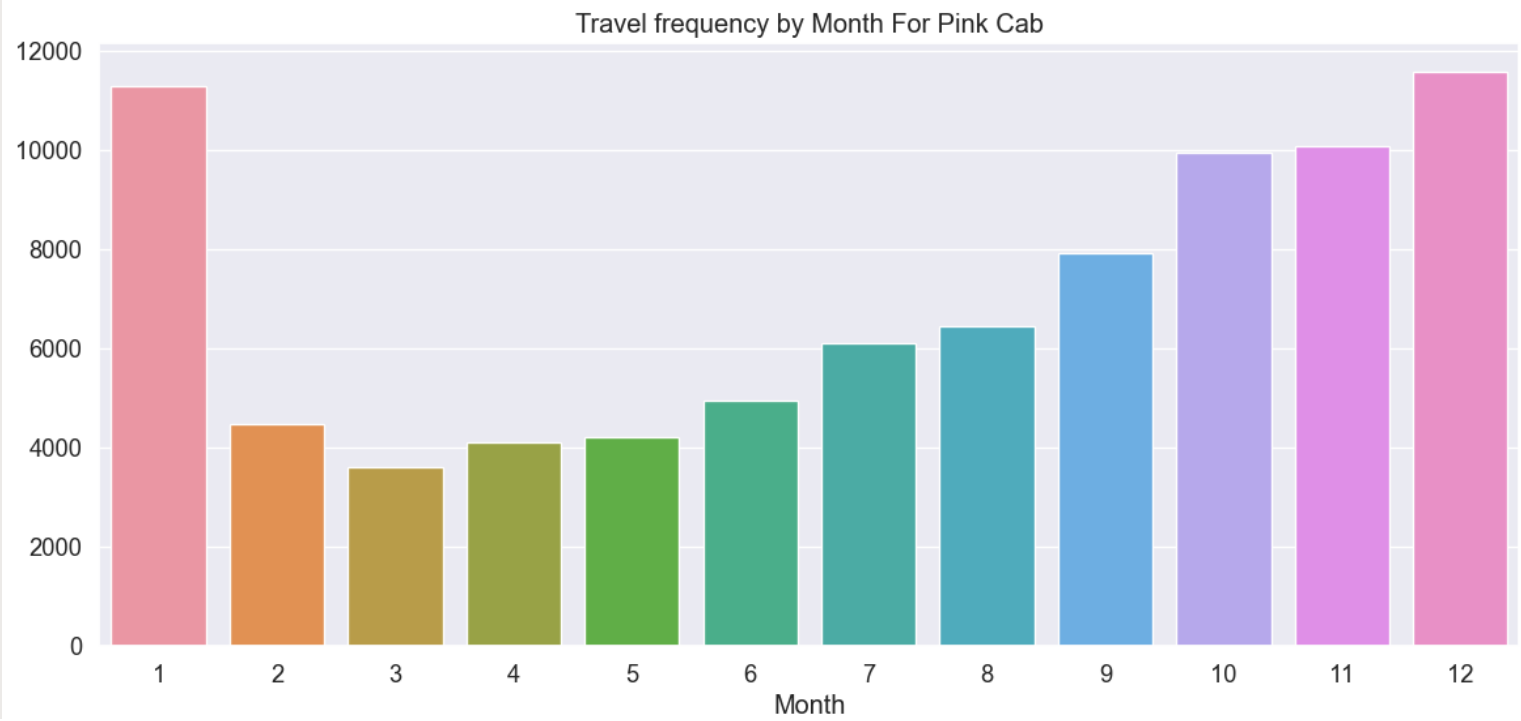


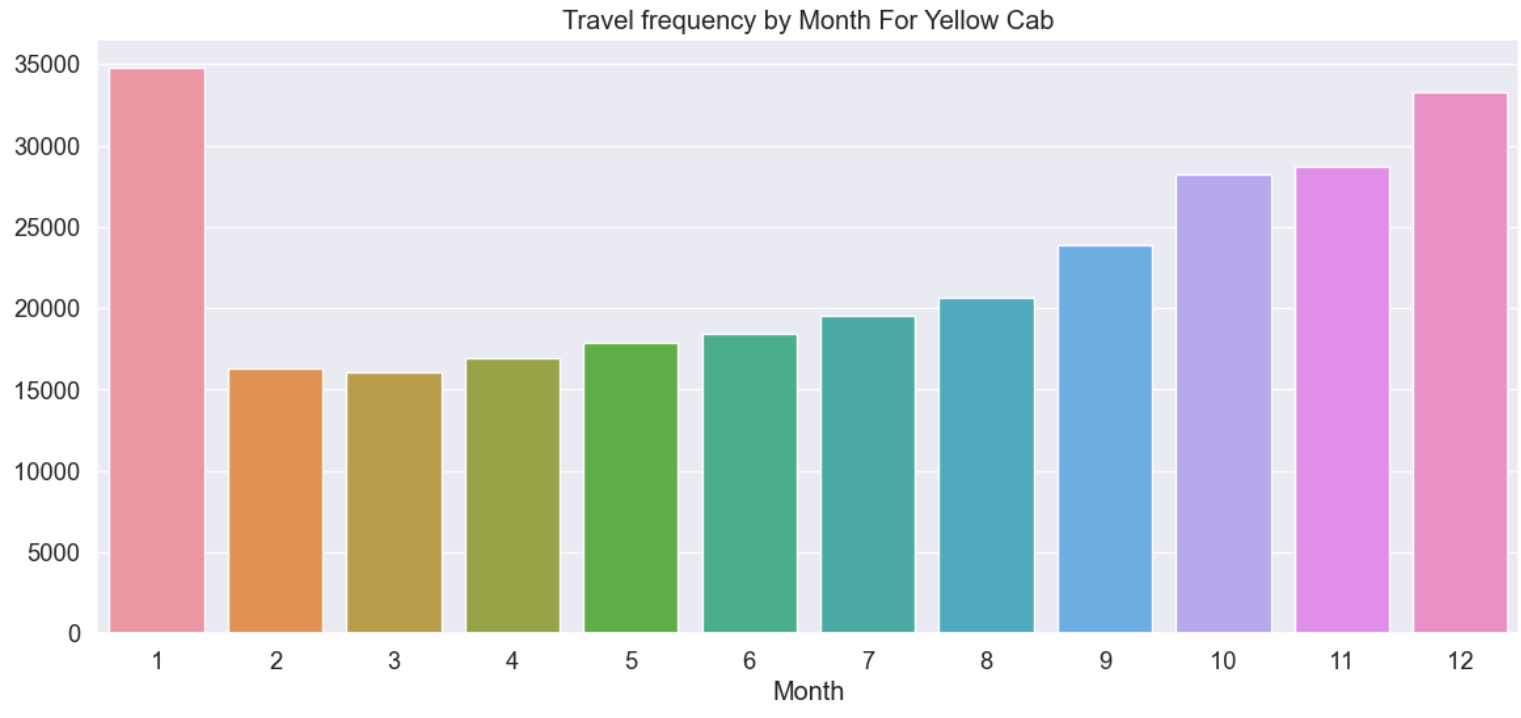

Data Glacier

# Relationships Between Variables



# Travel frequency by Months

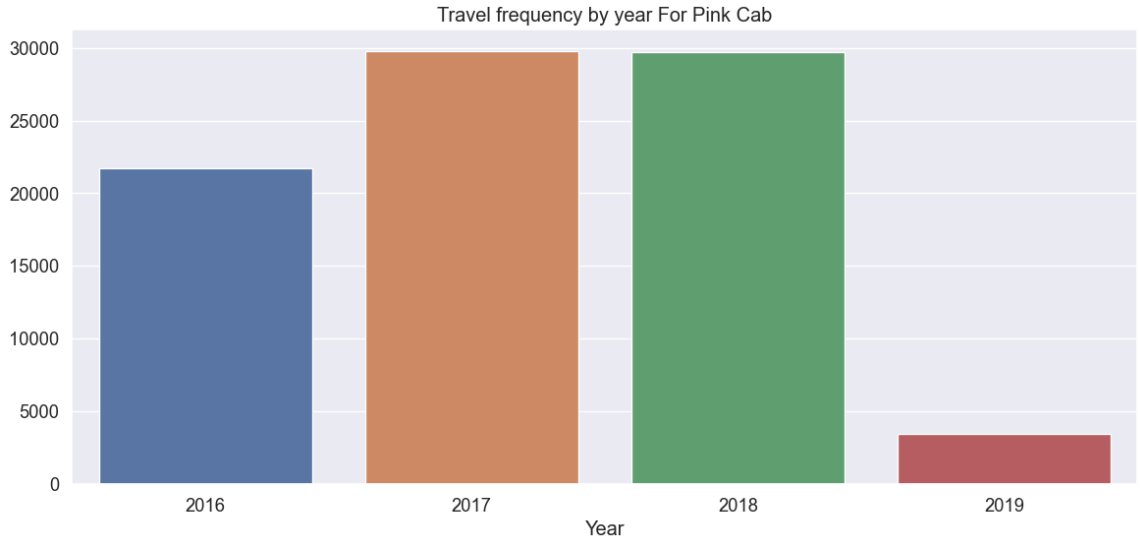


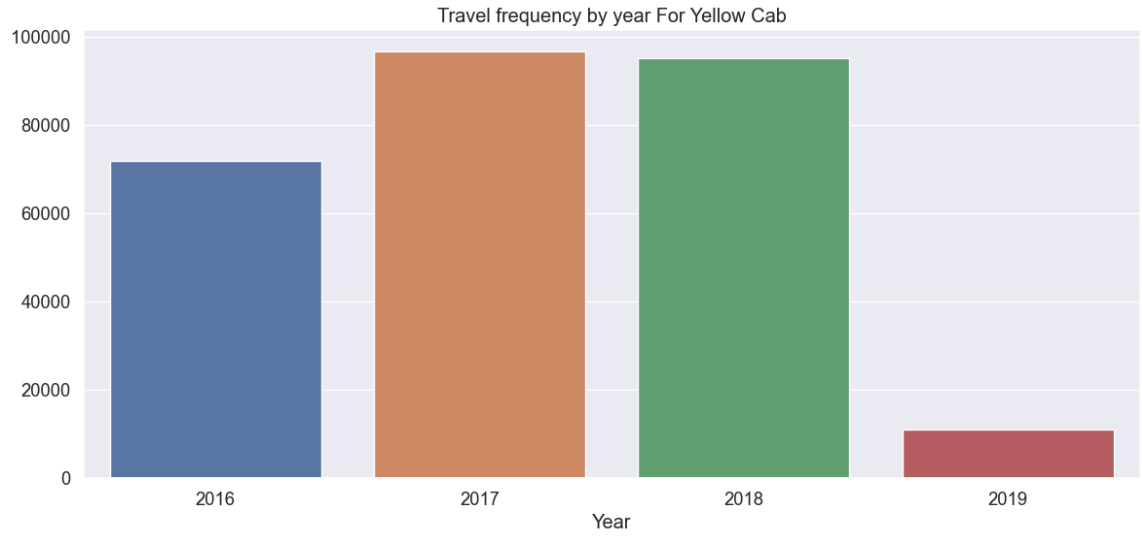




Data Glacier

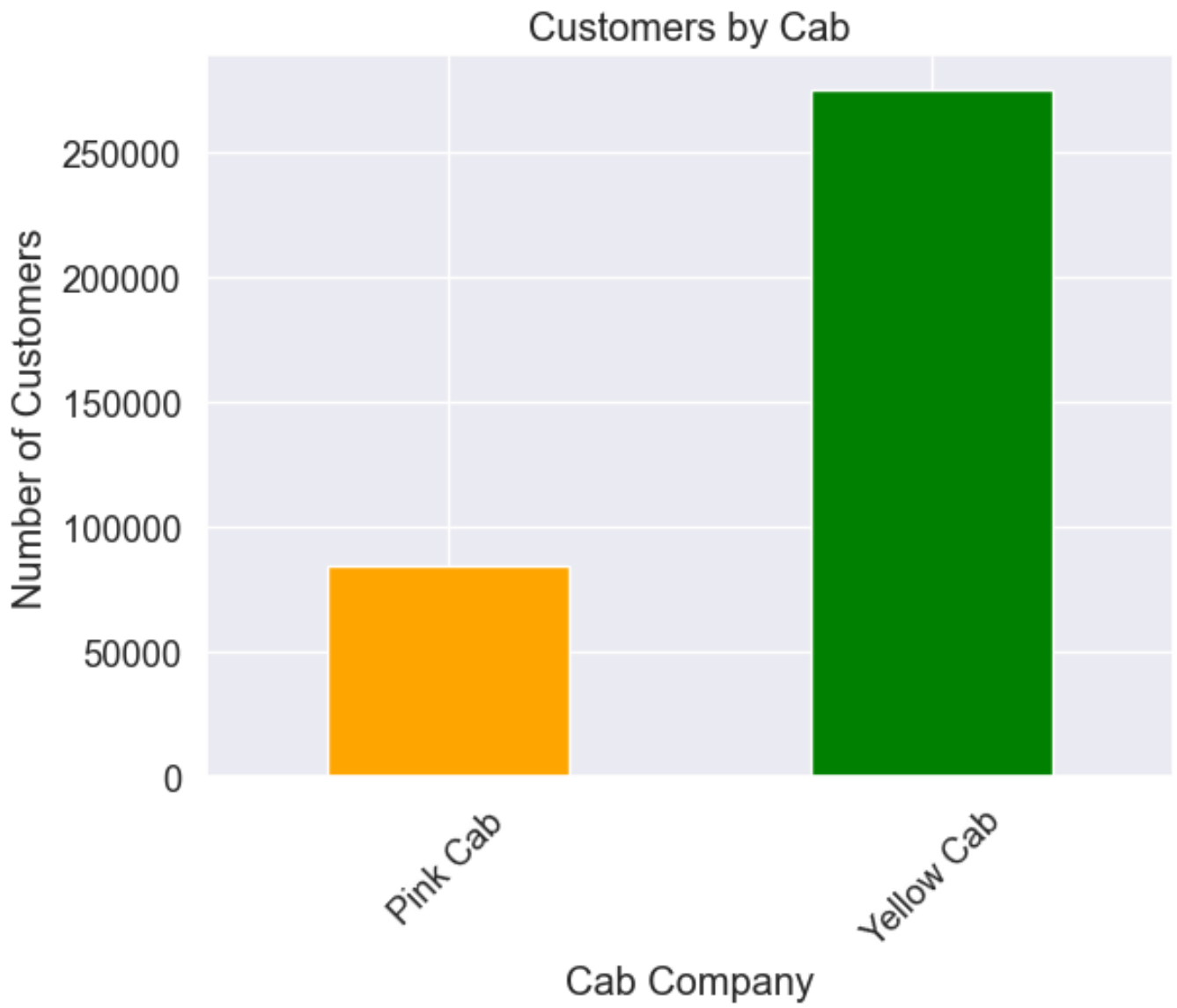
# Travel Frequency by Year

- **From the graph it shows that on yearly basis no. of transactions for Yellow cab is higher than Pink cab. Also 2019 low because it is only for early months.**



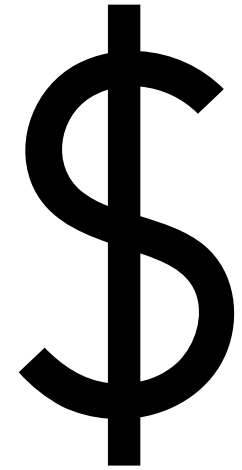



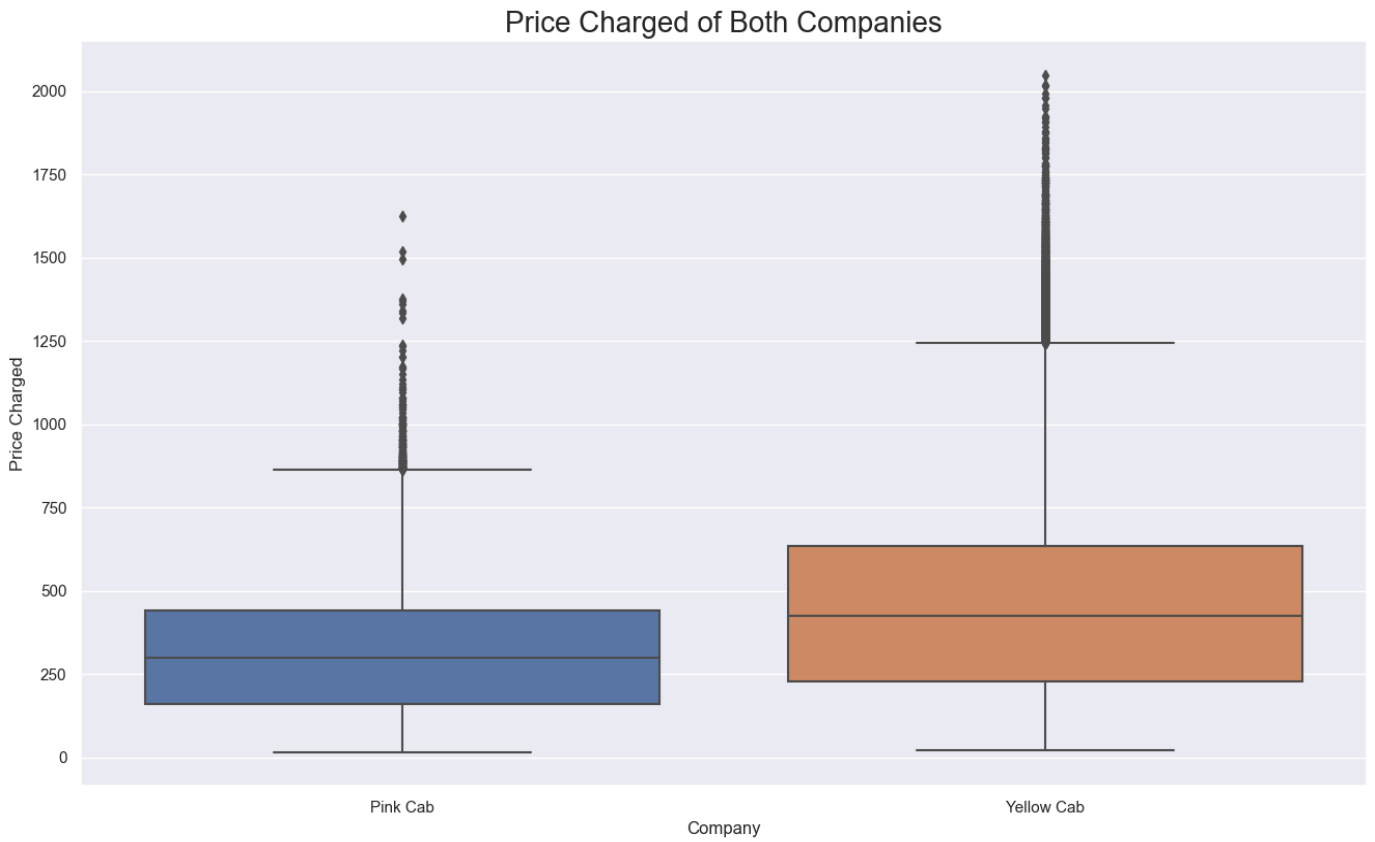
# Customers by Cab



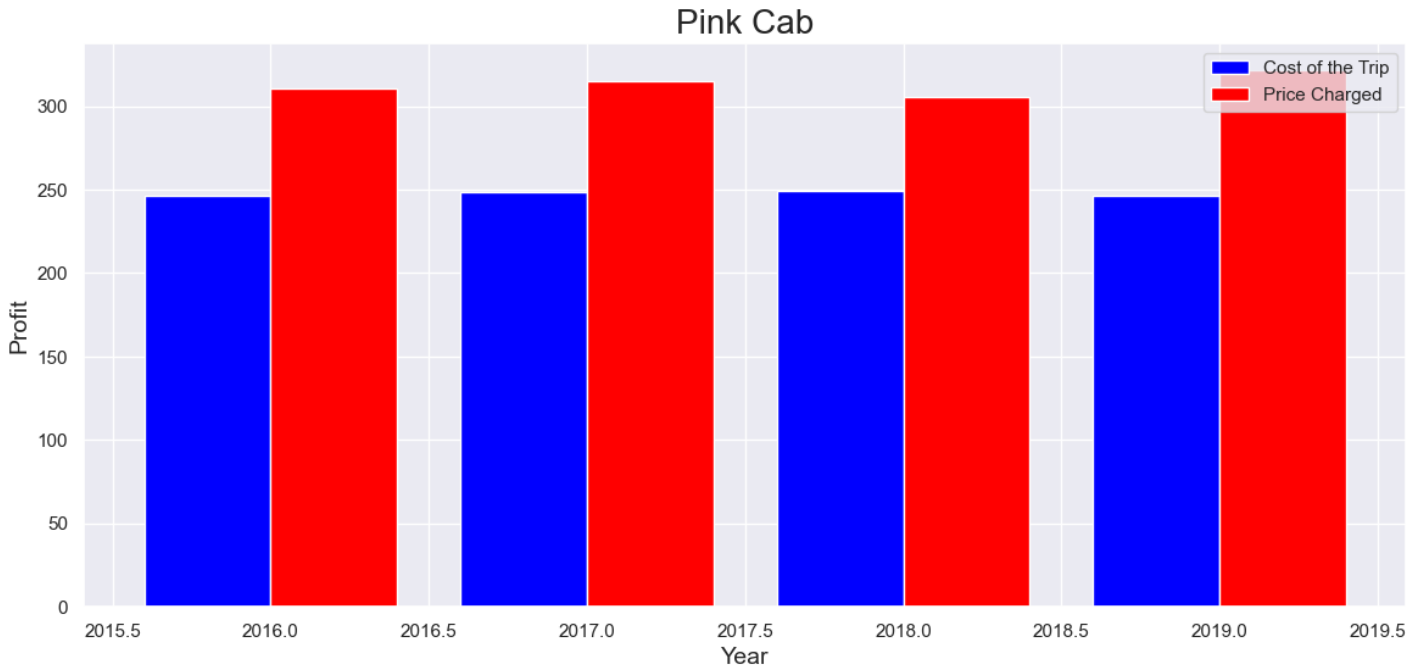
# Price Charged for Both Companies

- **As we can see Price Charged of Yellow Cab is highest as compared to Pink Cab**

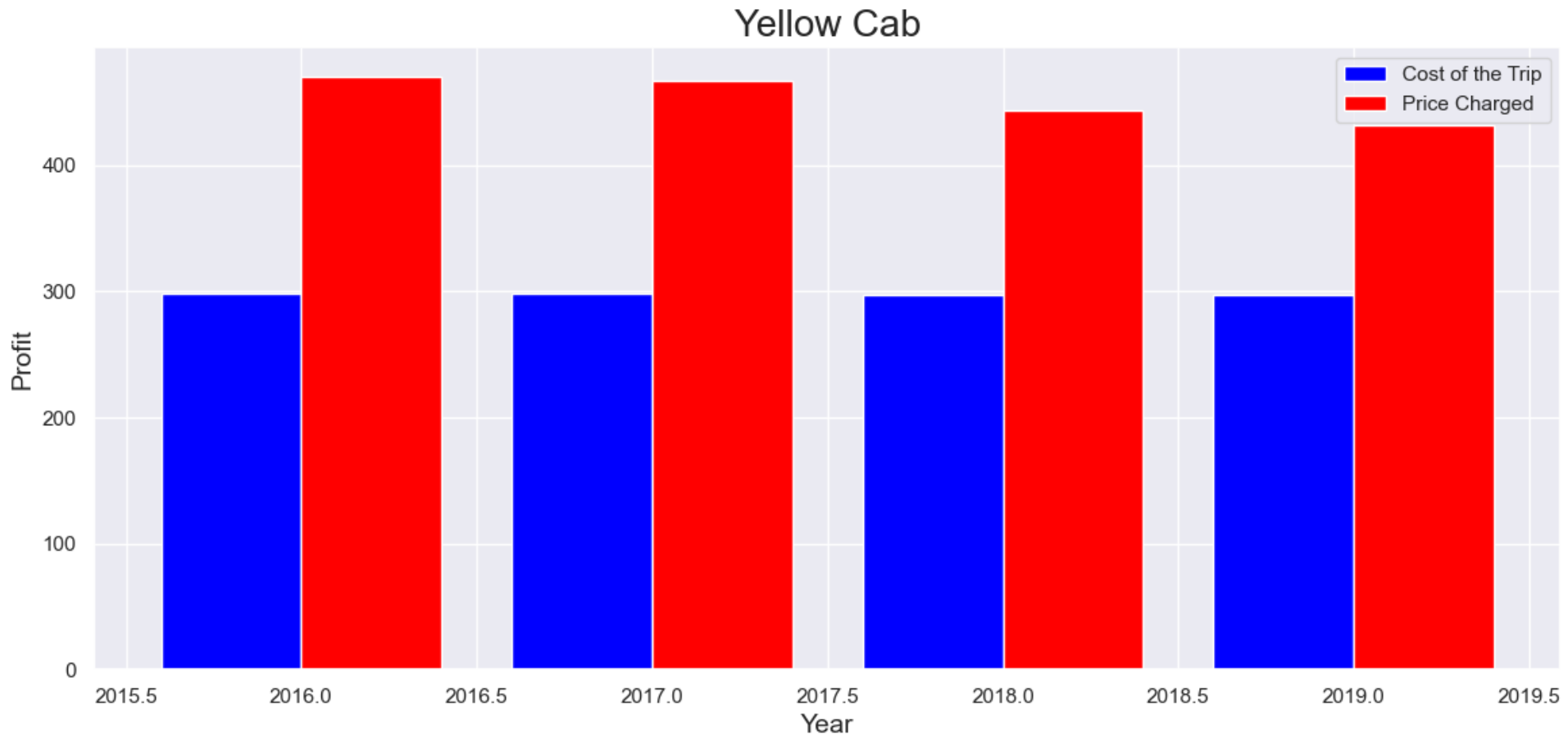




# Price Margins for Both Companies

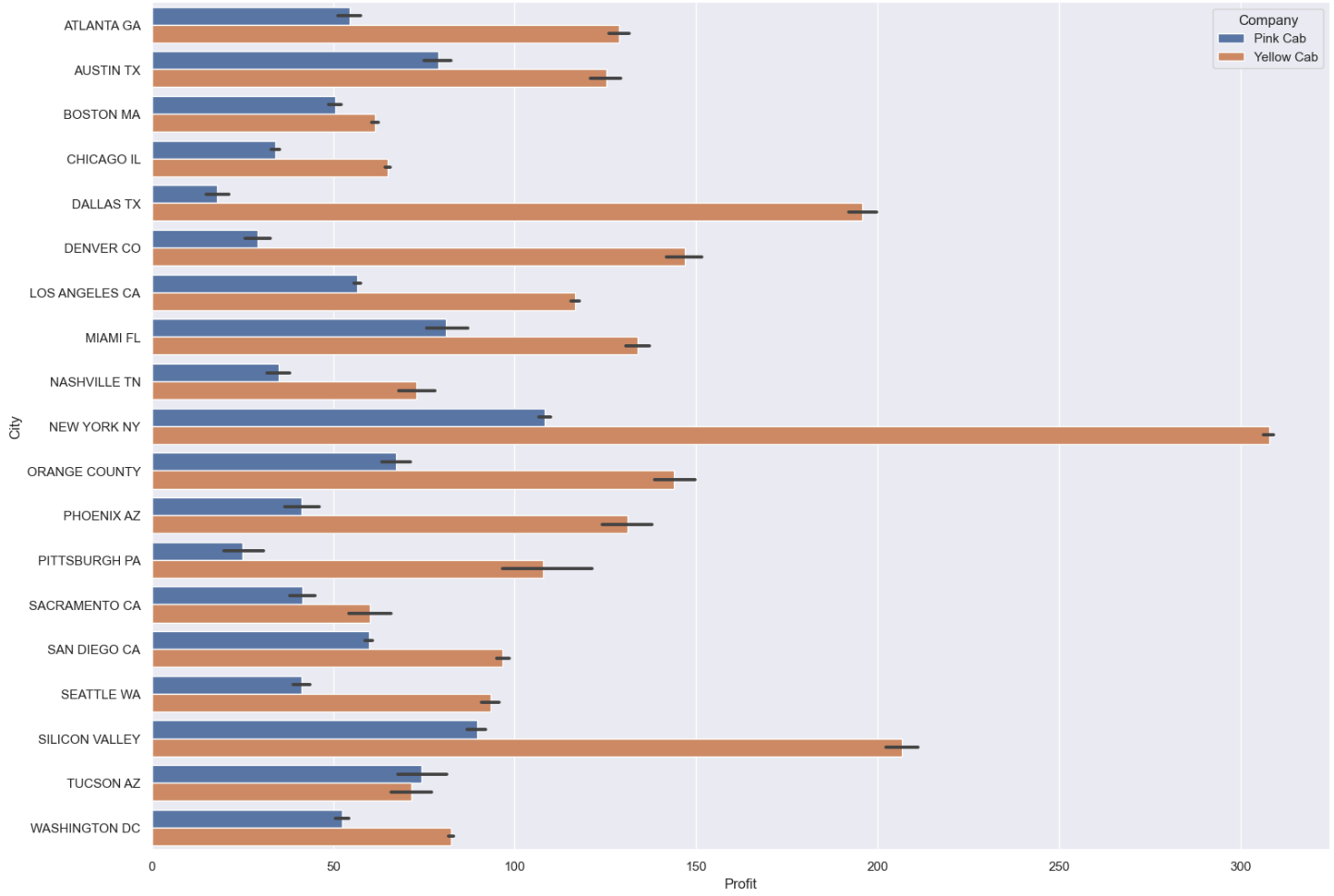


# The Yellow cab has a higher Profit Margin (Price Charged - Cost of Trip) compared to Pink cab

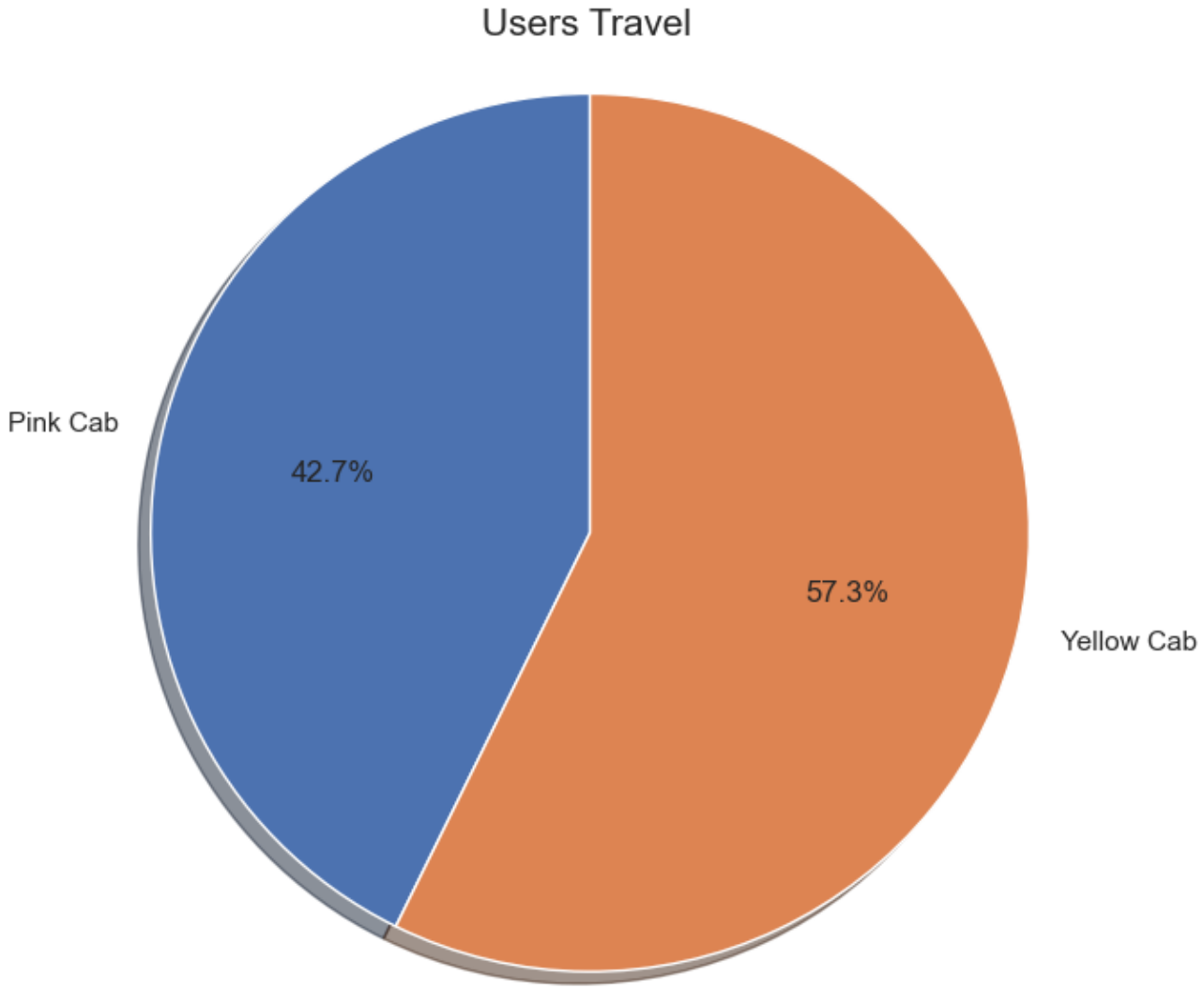


# Profit With Cities for Both Companies

- Almost in all cities Yellow Cab has more profit.

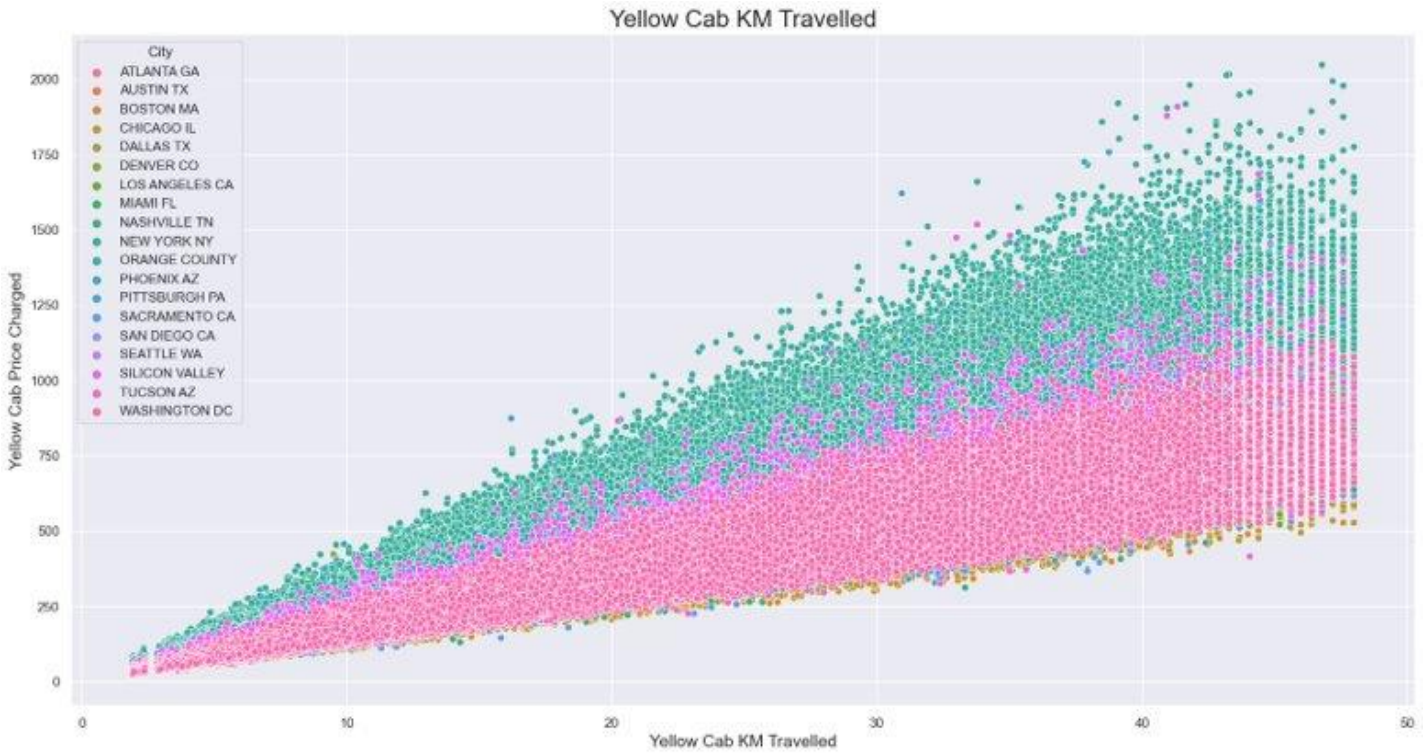


# Another Way To Show User Travel



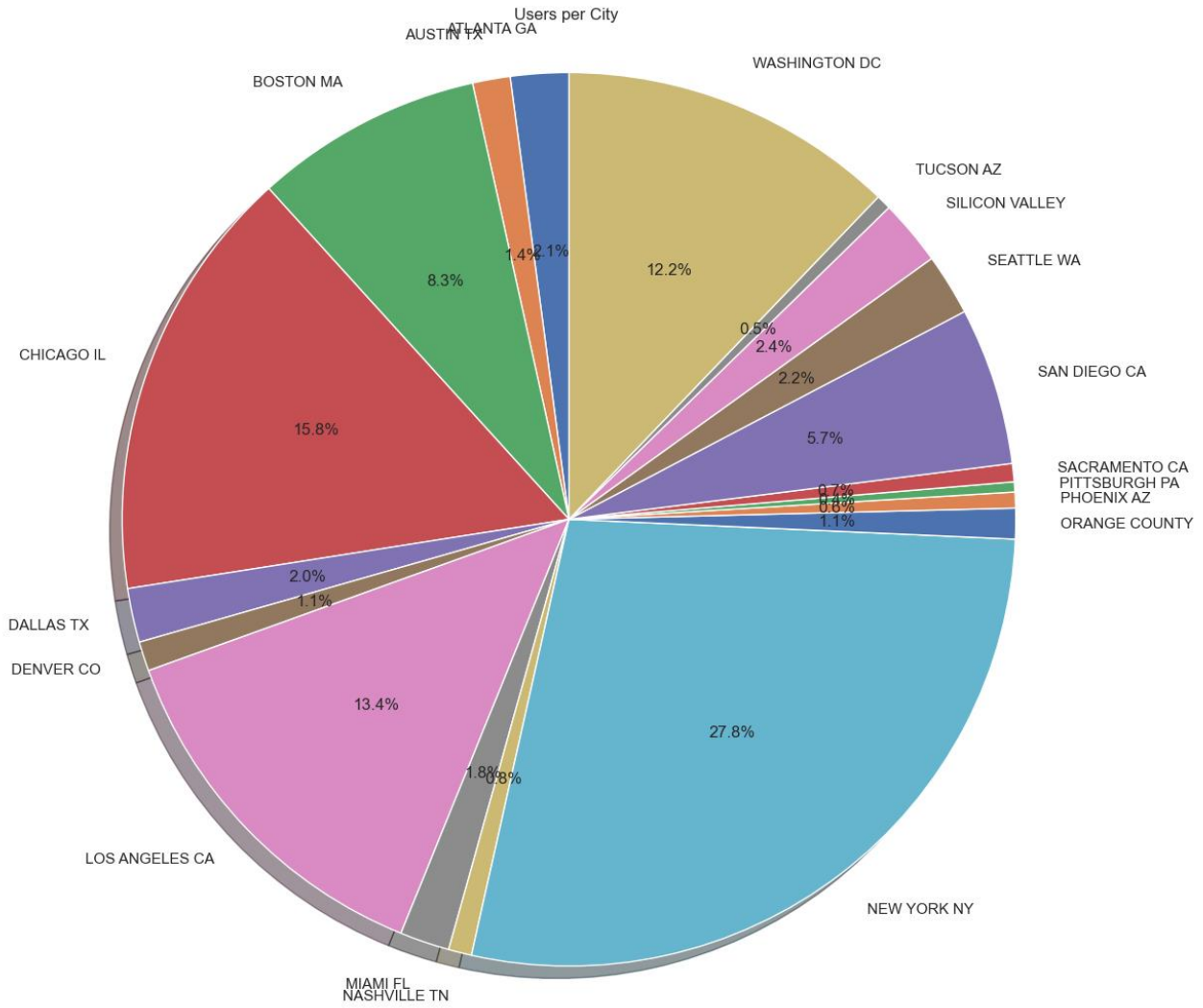

Pink Cab KM Travelled
City
ATLANTA GA
AUSTIN TX
BOSTON MA
CHICAGO IL
DALLAS TX
DENVER CO
LOS ANGELES CA
MIAMI FL
NASHVILLE TN
NEW YORK NY
ORANGE COUNTY
PHOENIX AZ
PITTSBURGH PA
SACRAMENTO CA
SAN DIEGO CA
SEATTLE WA
SILICON VALLEY
TUCSON AZ
WASHINGTON DC
Pink Cab Price Charged
0
200
400
600
800
1000
1200
1400
1600
0
10
20
30
40
50
Pink Cab KM Travelled

>From the graphs we see that for New York City the Yellow Cab price charged is more in comparison to the other cities.
and for Pink cab all the cities have the same increase in prices with increase in distance
the outliers exist in both the graphs which may be due to high end cars or weather.

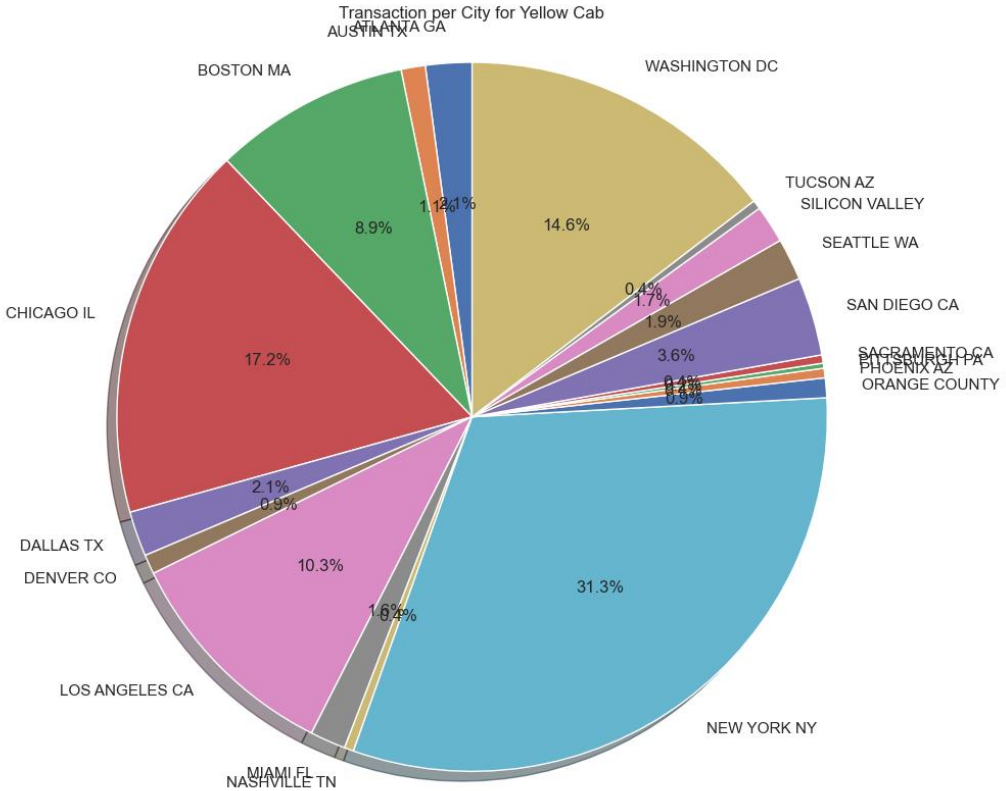


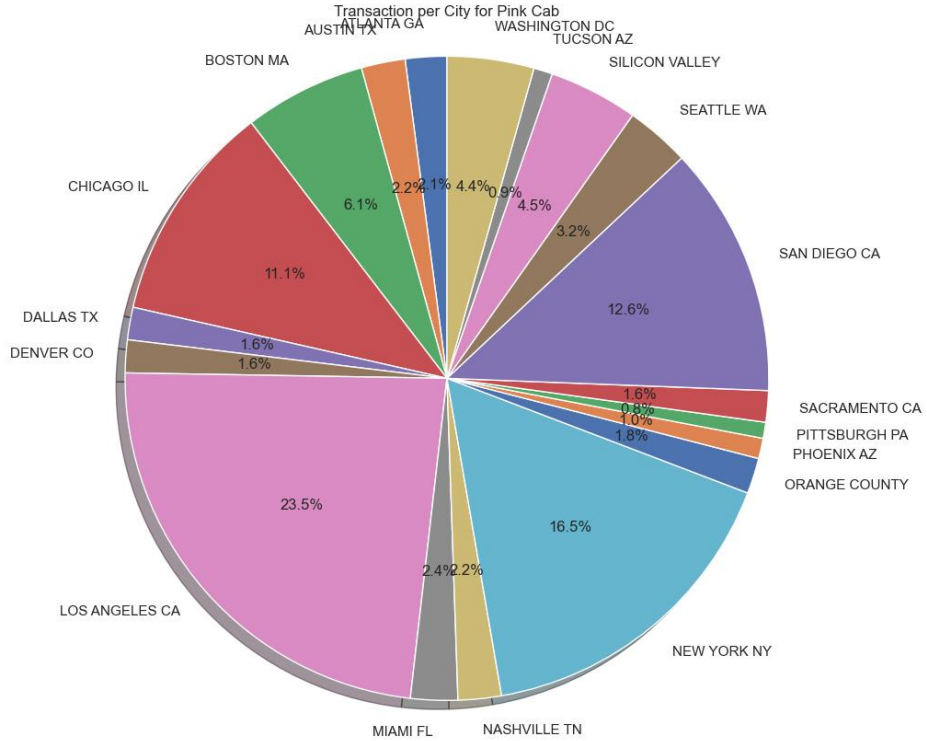
# Users Per City

- **New York City has the highest Cab users with 28% followed by Chicago with 16% and Los Angeles with 13%**



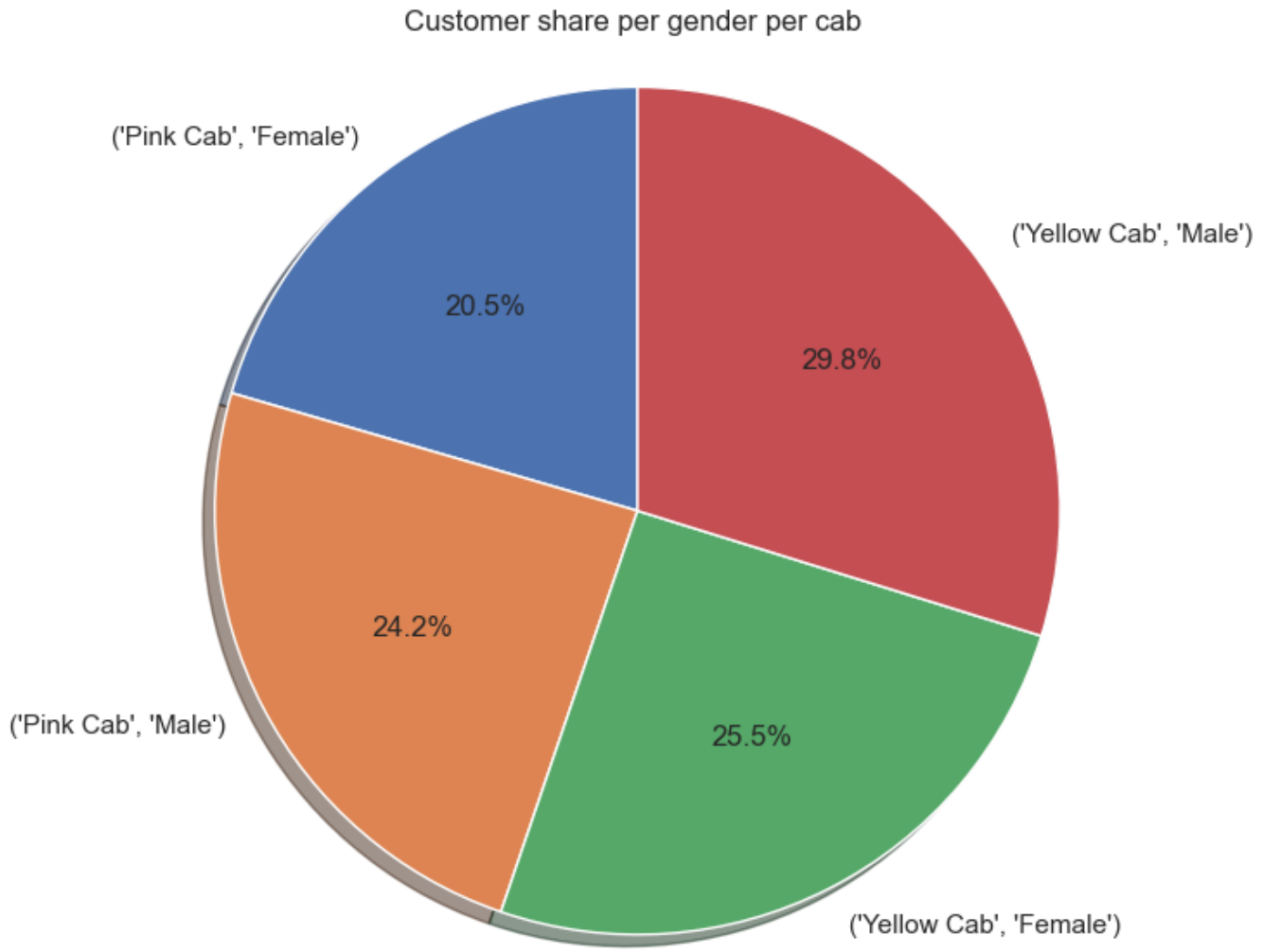
# Transaction for Yellow Cab is highest in New York City which has the highest Cab Users of 28%
# Transaction for Pink Cab is highest in Los Angeles City



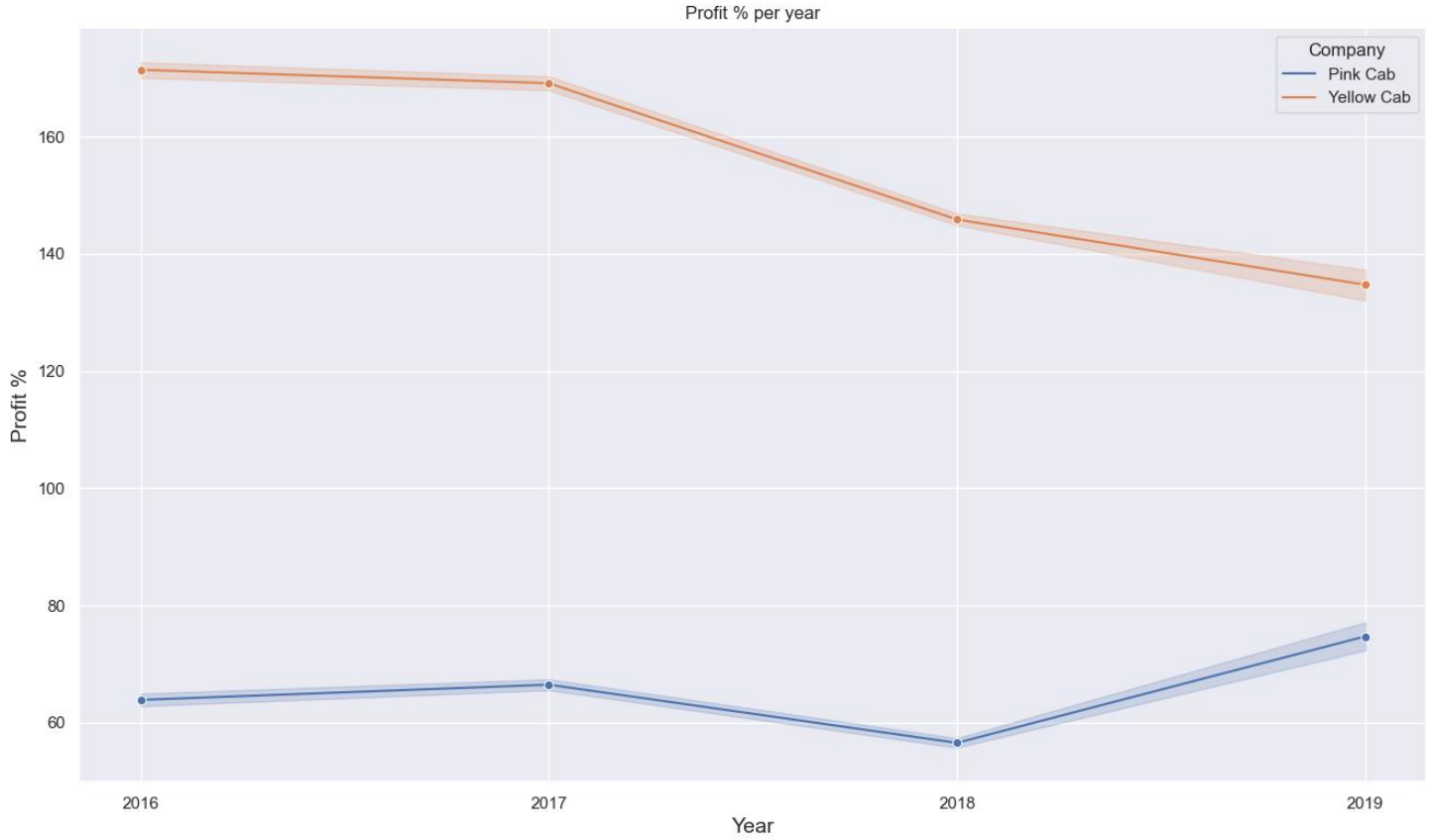
# Gender Data

- ***Female Customers in Yellow Cab is higher compared to Male customer**



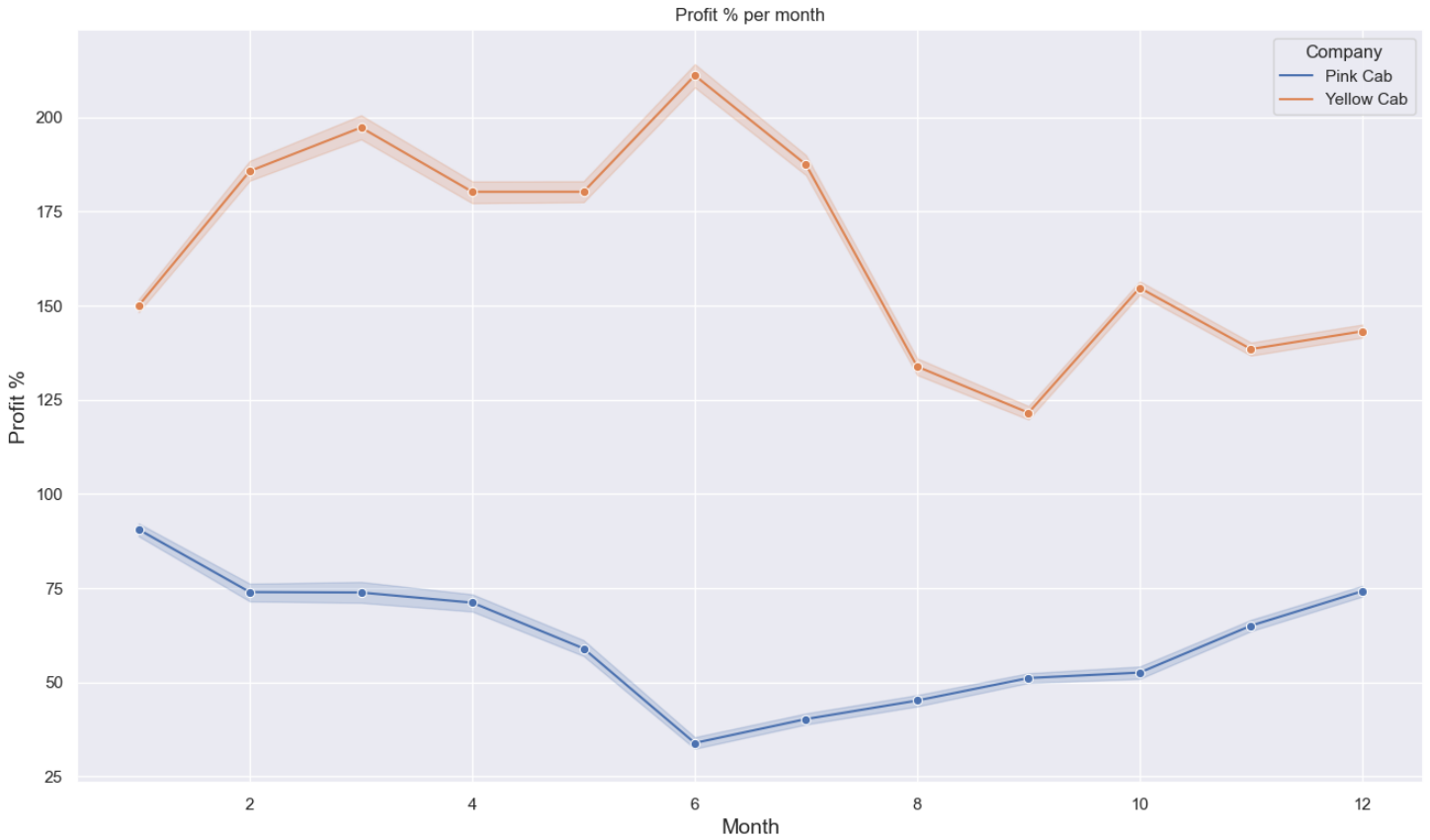
# Profit And Time

- While Profit Margin of Yellow Cab Decreases Over Time Pink Cab Actually Increased



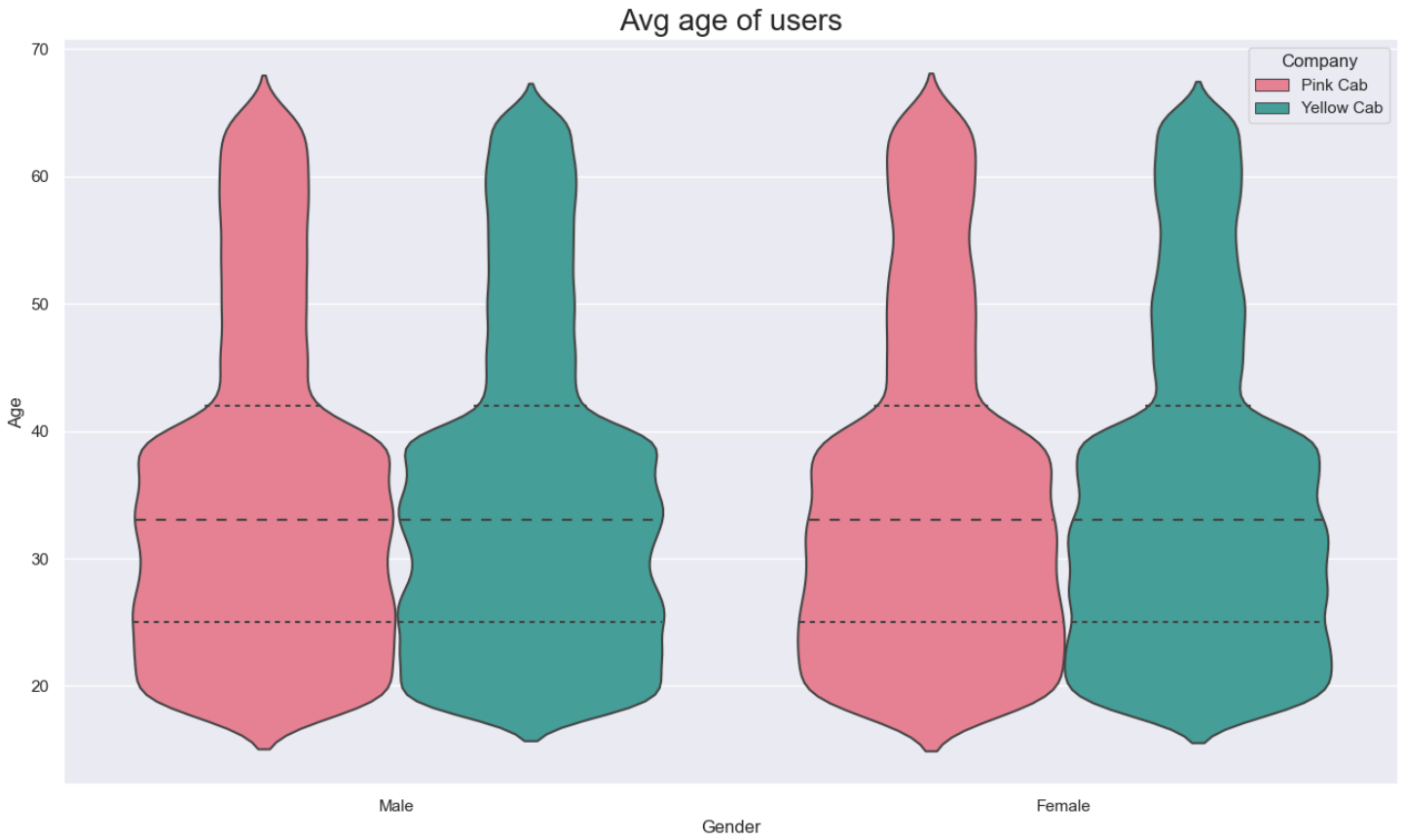
# Profit per Month

- **For both companies profit margin varies between months but important to note here that they don't show the same increments or decrements**

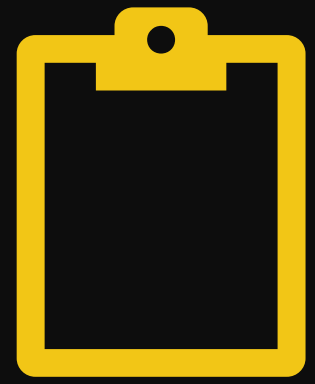


# Average Age of Users

- **As we can see 34 is Average age of Female and Male who use Cab service**



# EDA SUMMARY



- Yellow Cab seems to have more customers and profit but it's profit margin is decreasing while Pink Cab's profit margin is increasing.
- Yellow Cab has a higher percentage of female users.
- Yellow Cab generally charges more.

# 3.HYPOTESIS TESTING

# Hypothesis 1: There is a difference regarding travel length between companies

```
a = PinkCab['KM Travelled'].mean()
b = YellowCab['KM Travelled'].mean()
difference = abs(b-a)
ratio = (difference/b)*100
str1= str(ratio)[:4]
if ratio < 10 :
    print(f"We accept null hypothesis that there is no difference, ratio is equal to: {str1} ")
else:
    print(f"We accept alternate hypothesis that there is no difference, ratio is equal to: {str1} ")
```

```
We accept null hypothesis that there is no difference, ratio is equal to: 0.04
```

# Hypothesis 2: Is there difference in margins for Card payer and Cash payers

**Hypothesis 2: Is there difference in margins for Card payer and Cash payers**

```
#PinkCab
a = PinkCab[PinkCab.Payment_Mode=='Cash'].groupby('Transaction ID').Profit.mean()
b = PinkCab[PinkCab.Payment_Mode=='Card'].groupby('Transaction ID').Profit.mean()


_, p_value = stats.ttest_ind(a.values,b=b.values,equal_var=True)
if(p_value<0.05):
    print('We accept alternate hypothesis that theres a difference')
else:
    print('We accept null hypothesis that theres no difference')

print('P value is ', p_value)
```

```
We accept null hypothesis that theres no difference
P value is  0.7900465828793286
```

```
#YellowCab
a = YellowCab[YellowCab.Payment_Mode=='Cash'].groupby('Transaction ID').Profit.mean()
b = YellowCab[YellowCab.Payment_Mode=='Card'].groupby('Transaction ID').Profit.mean()


_, p_value = stats.ttest_ind(a.values,b=b.values,equal_var=True)
if(p_value<0.05):
    print('We accept alternate hypothesis that theres a difference')
else:
    print('We accept null hypothesis that theres no difference')

print('P value is ', p_value)
```

```
We accept null hypothesis that theres no difference
P value is  0.2933060638298729
```

# Hypothesis 3: Is there any difference in Profit regarding Age

## Yellow Cab

```
a = YellowCab[YellowCab.Age <= 60].groupby('Transaction ID').Profit.mean()
b = YellowCab[YellowCab.Age >= 60].groupby('Transaction ID').Profit.mean()
print(a.shape[0],b.shape[0])

_, p_value = stats.ttest_ind(a.values,
                             b.values,
                             equal_var=True)

print('P value is ', p_value)

if(p_value<0.05):
    print('We accept alternative hypothesis (H1) that there is a difference regarding age for YellowCab Cab')
else:
    print('We accept null hypothesis (H0) that there is no difference regarding age for YellowCab Cab')
```

```
260356 17257
P value is  6.328485471267631e-05
We accept alternative hypothesis (H1) that there is a difference regarding age for YellowCab Cab
```

## Pink Cab

```
a = PinkCab[PinkCab.Age <= 60].groupby('Transaction ID').Profit.mean()
b = PinkCab[PinkCab.Age >= 60].groupby('Transaction ID').Profit.mean()
print(a.shape[0],b.shape[0])

_, p_value = stats.ttest_ind(a.values,
                             b.values,
                             equal_var=True)

print('P value is ', p_value)

if(p_value<0.05):
    print('We accept alternative hypothesis (H1) that there is a difference regarding age for Pink Cab')
else:
    print('We accept null hypothesis (H0) that there is no difference regarding age for Pink Cab')
```

```
80125 5429
P value is  0.4816748536155635
We accept null hypothesis (H0) that there is no difference regarding age for Pink Cab
```

# Hypothesis 4: Price charged is correlated with the Income of the customer

**Hypothesis 4: Price charged is correlated with the Income of the customer**

```
h5 = All_Data["Price Charged"].corr(All_Data["Income (USD/Month)"])

if h5>0.5 or h5<-0.5:
    print("We accept the null hypothesis (HQ) that Price",
    "charged is correlated with the Income of the customer.")
else:
    print ("We accept the alternative hypothesis (H1) that Price",
    "charged is not correlated with the Income of the customer.")
```

```
We accept the alternative hypothesis (H1) that Price charged is not correlated with the Income of the customer.
```

# Hypothesis 5: Is the average cost of a trip with Pink Cab is lower than that of Yellow Cab.

**Hypothesis 5: Is the average cost of a trip with Pink Cab is lower than that of Yellow Cab.**

```
a = YellowCab.groupby('Transaction ID')['Cost of Trip'].mean()
b = PinkCab.groupby('Transaction ID')['Cost of Trip'].mean()
print(a.shape[0], b.shape[0])

from scipy import stats
_, p_value = stats.ttest_ind(a.values,
                             b.values,
                             equal_var=True)

print('P value is ', p_value)

if p_value < 0.05:
    print('We accept alternative hypothesis (H1) that the average cost of a trip with Pink Cab is lower than that of Yellow Cab.'
else:
    print('We accept null hypothesis (H0) that there is no significant difference in the average cost of a trip between Pink Cab
```

```
274681 84711
P value is  0.0
We accept alternative hypothesis (H1) that the average cost of a trip with Pink Cab is lower than that of Yellow Cab.
```

# Hypotesis 6: The number of rides taken by customers in a particular city is correlated with the population of that city.

```
from scipy.stats import pearsonr
All_Data['Population'] = All_Data['Population'].apply(lambda x: str(x))

All_Data['Population'] = All_Data['Population'].apply(lambda x: convert_int(x))
# Calculate the correlation between the number of rides and the population
pop_counts = All_Data.groupby('Population').size()
pop_counts = pop_counts.sort_index()
corr, p_value = pearsonr(pop_counts.index, pop_counts.values)

# create a bar chart of the population counts

# Print the results
if p_value < 0.05:
    print(f'We accept the alternative hypothesis (H1) that the number of rides is correlated with the population.\np value: {p_va
else:
    print('We accept the null hypothesis (H0) that there is no correlation between the number of rides and the population.')
```

```
We accept the alternative hypothesis (H1) that the number of rides is correlated with the population.
p value: 3.6168115765213676e-05
```

# Conclusions

Based on the results of our analysis, it is clear that Pink Cab is still in the process of establishing itself in the market, as evidenced by its smaller market share in comparison to Yellow Cab. This is further supported by the fact that Pink Cab has a lower price per KM, likely in an effort to incentivize market disruption and attract customers away from more established competitors.

It is also worth noting that while Yellow Cab has been experiencing a decrease in profits over the years, Pink Cab's profits have been increasing. This trend suggests that Pink Cab has the potential for growth, which is a key consideration for investors seeking to maximize returns.

Despite Pink Cab's perceived disadvantages in terms of its market share and features, we recommend this company as the better investment opportunity. This recommendation takes into account the potential for growth and the fact that the company is still in transition, which presents an opportunity for investors to get in on the ground floor and potentially reap greater returns in the long run.

From
Salih Eren
Yüzbaşıoğlu

Data Glacier
Your Deep Learning Partner

# Thank You