Project: G2M Insight For Cab Investment Firm

Data science Virtual Internship

FATIMA EZZAHRA AMRAOUI

20-Jan-2021

Agenda

- Problem Statement
- Approach
- EDA
- Hypothesis testing
- Recommendations

Problem Statement

- 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 has been divided into four parts:

- Data Understanding
- Exploratory data analysis
- Finding the most profitable Cab company
- Hypothesis testing and Recommendations for investment

Approach

- Relationship between variables
- Descriptive statistics
- Data visualisation
- Hypothesis testing

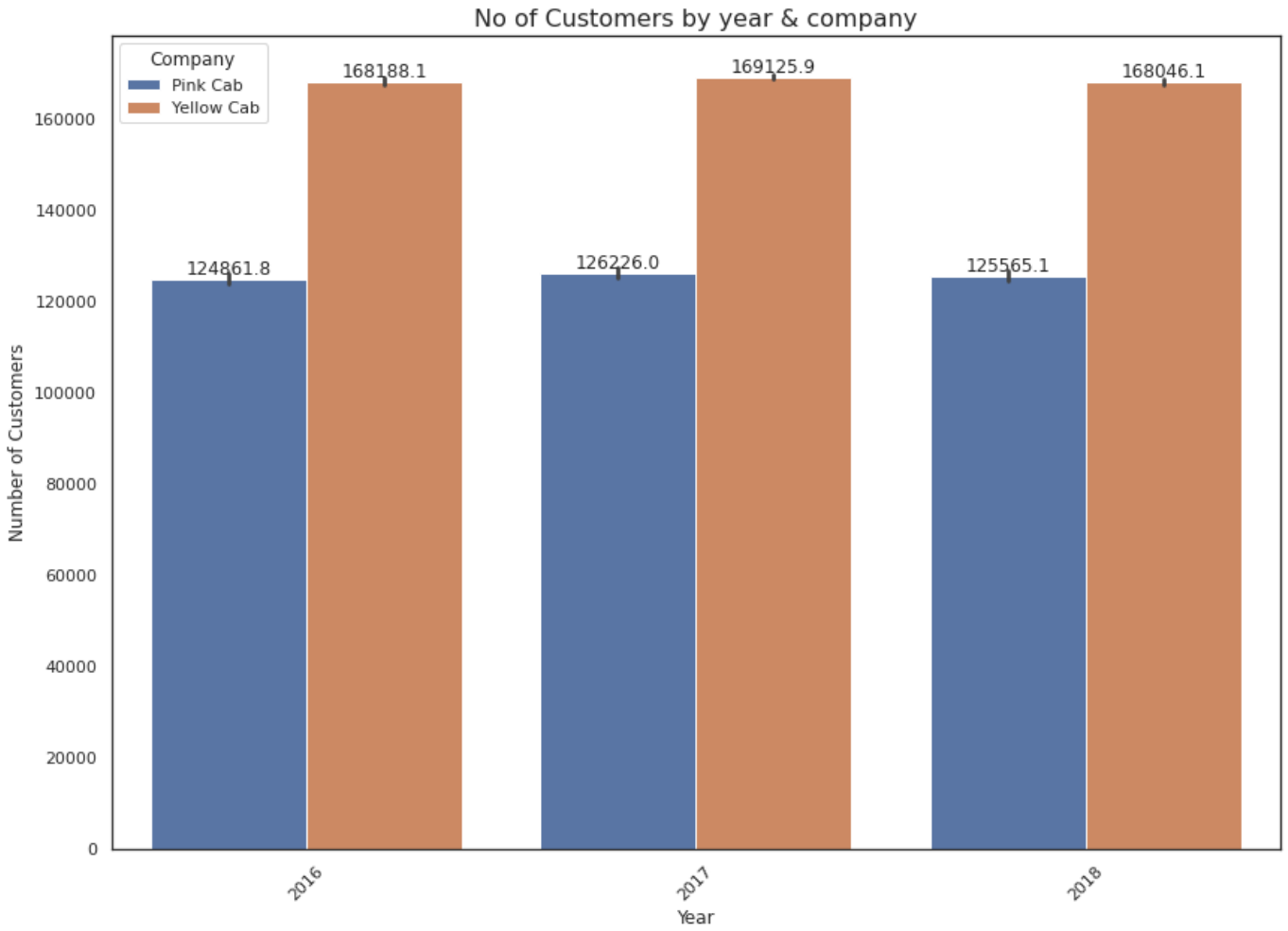
Data Exploration

- 4 datasets are merged to form the final dataset:

  Cab_Data.csv.
  Customer_ID.csv
  Transaction_ID.csv
  City.csv

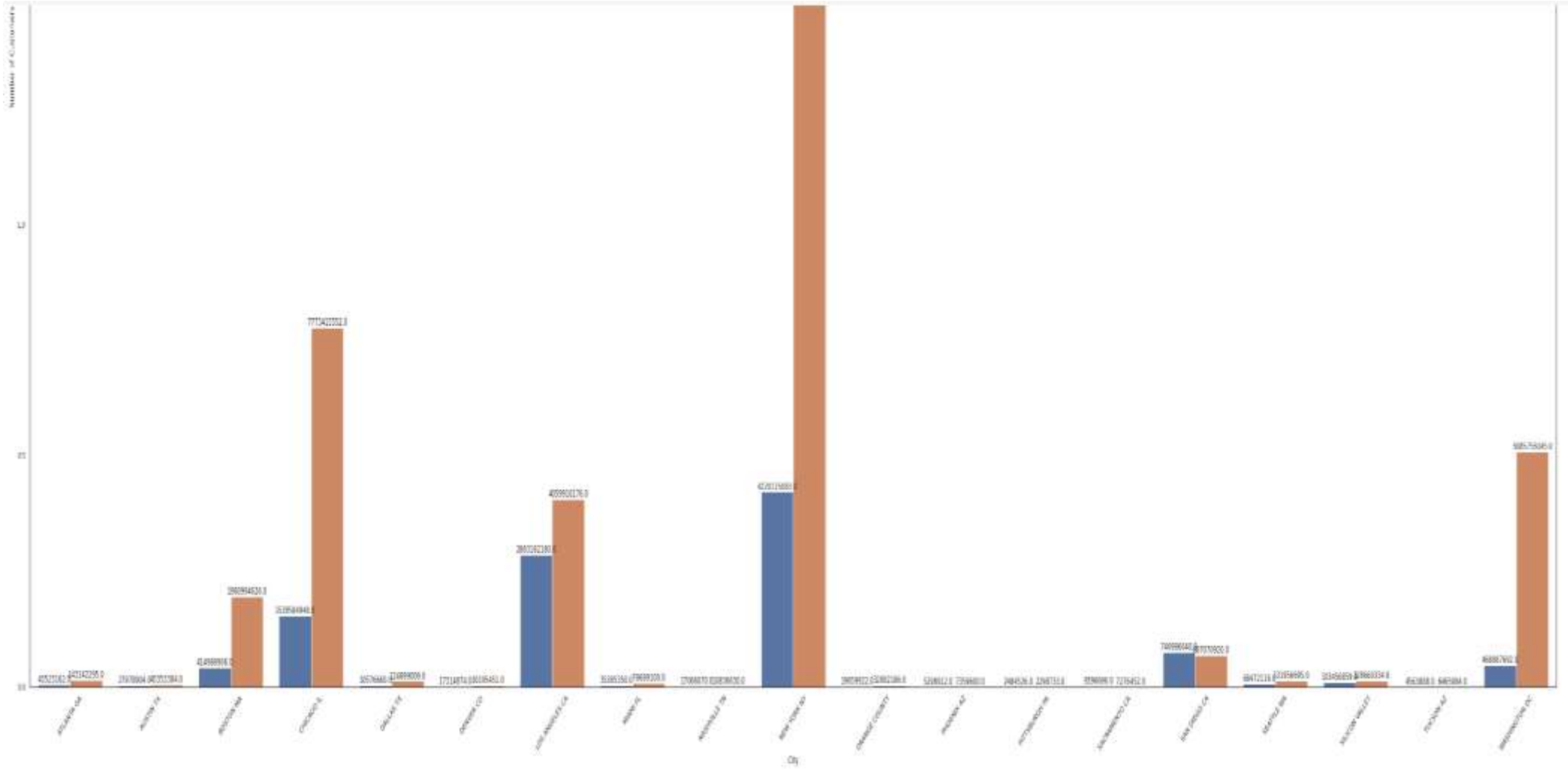
- Timeframe of the data: 2016-01-31 to 2018-12-31
- Total data points :359392

Customer Analysis



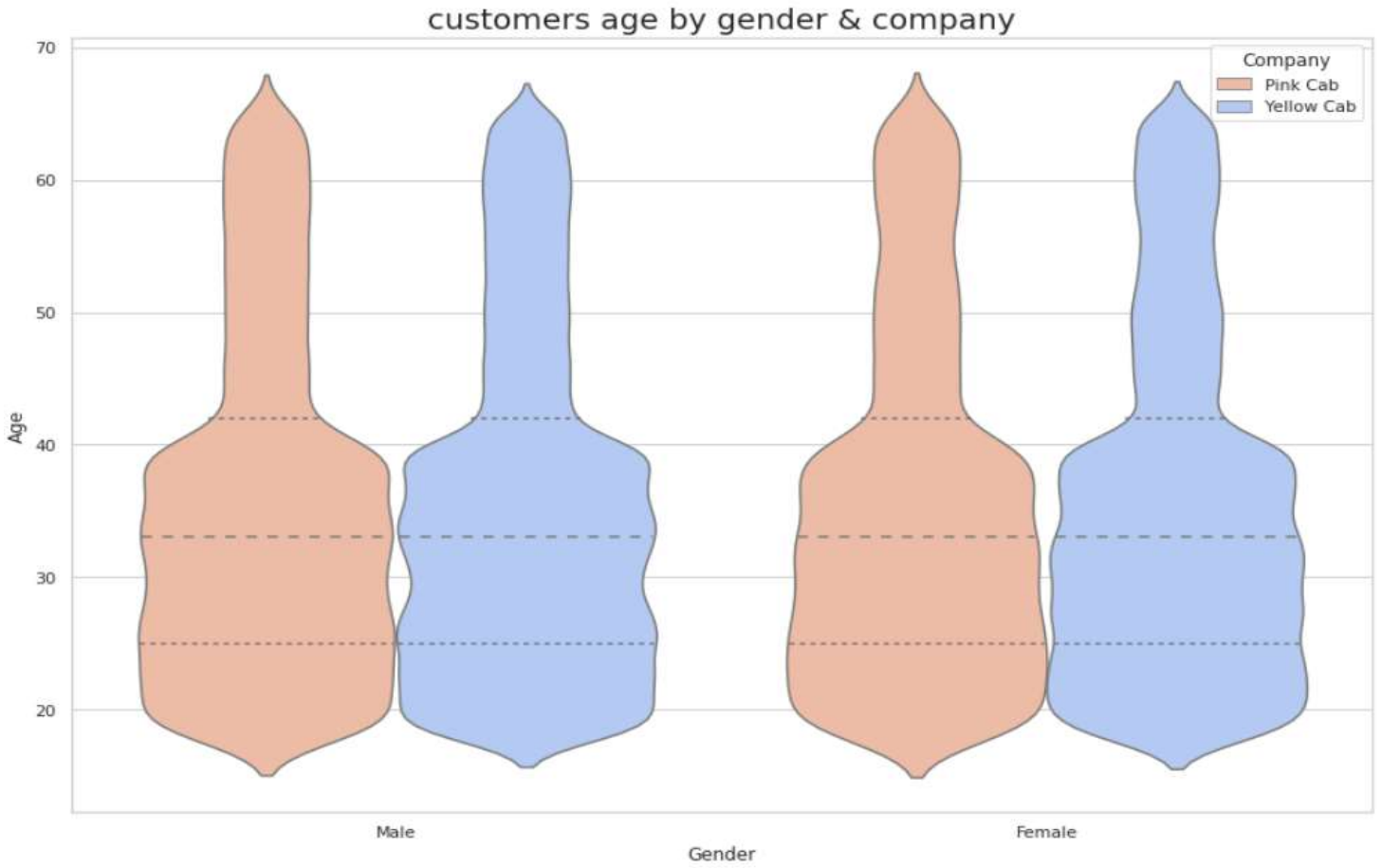
Yellow cab have more customers
Than yellow cab over the 3 years

Customer Analysis



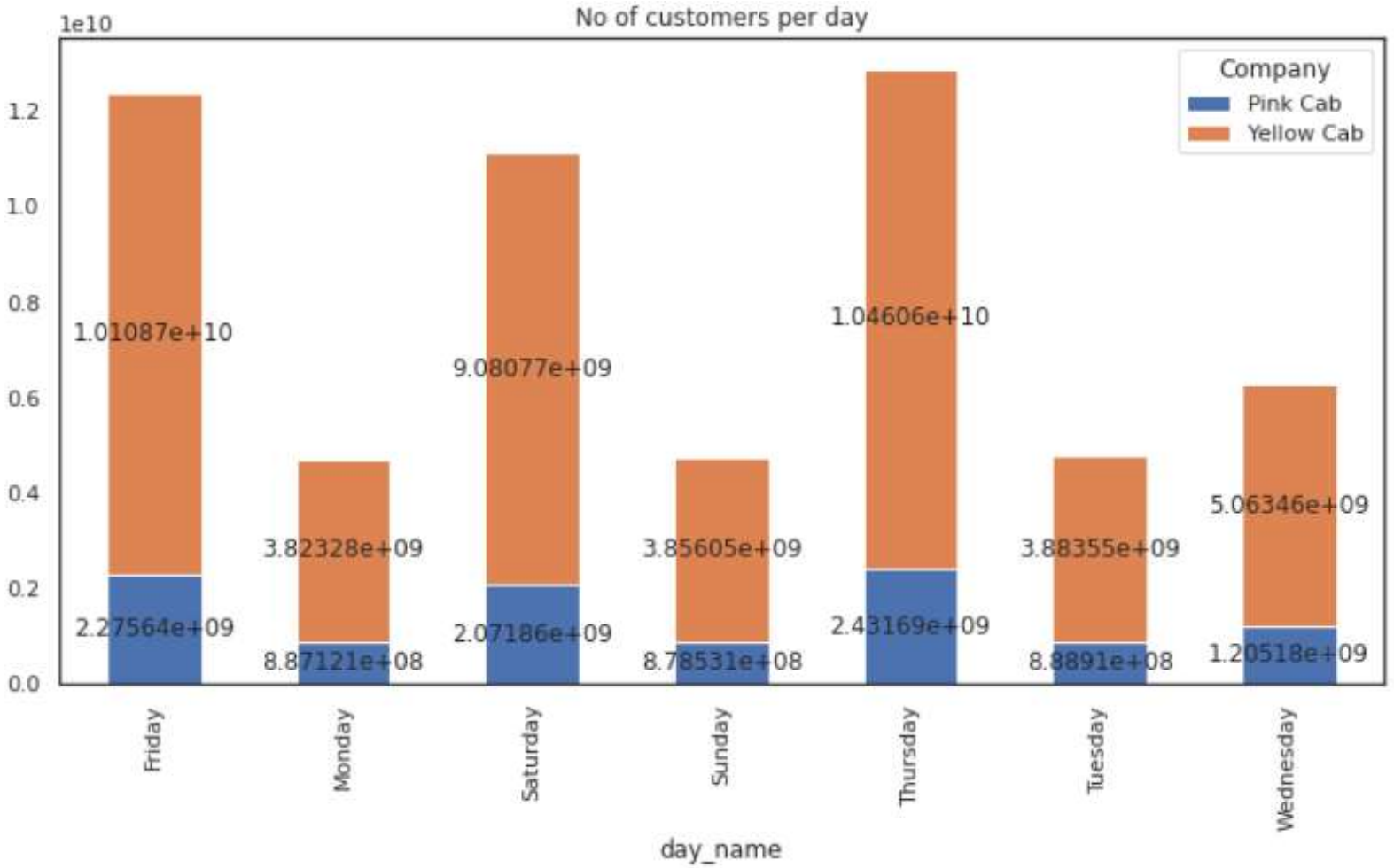
This plot display No of customers by city,cities that have more customers are : Bostonlos angeles,new york,chicago & washigthon. Where yellow cab exceed pink cab

customer Analysis



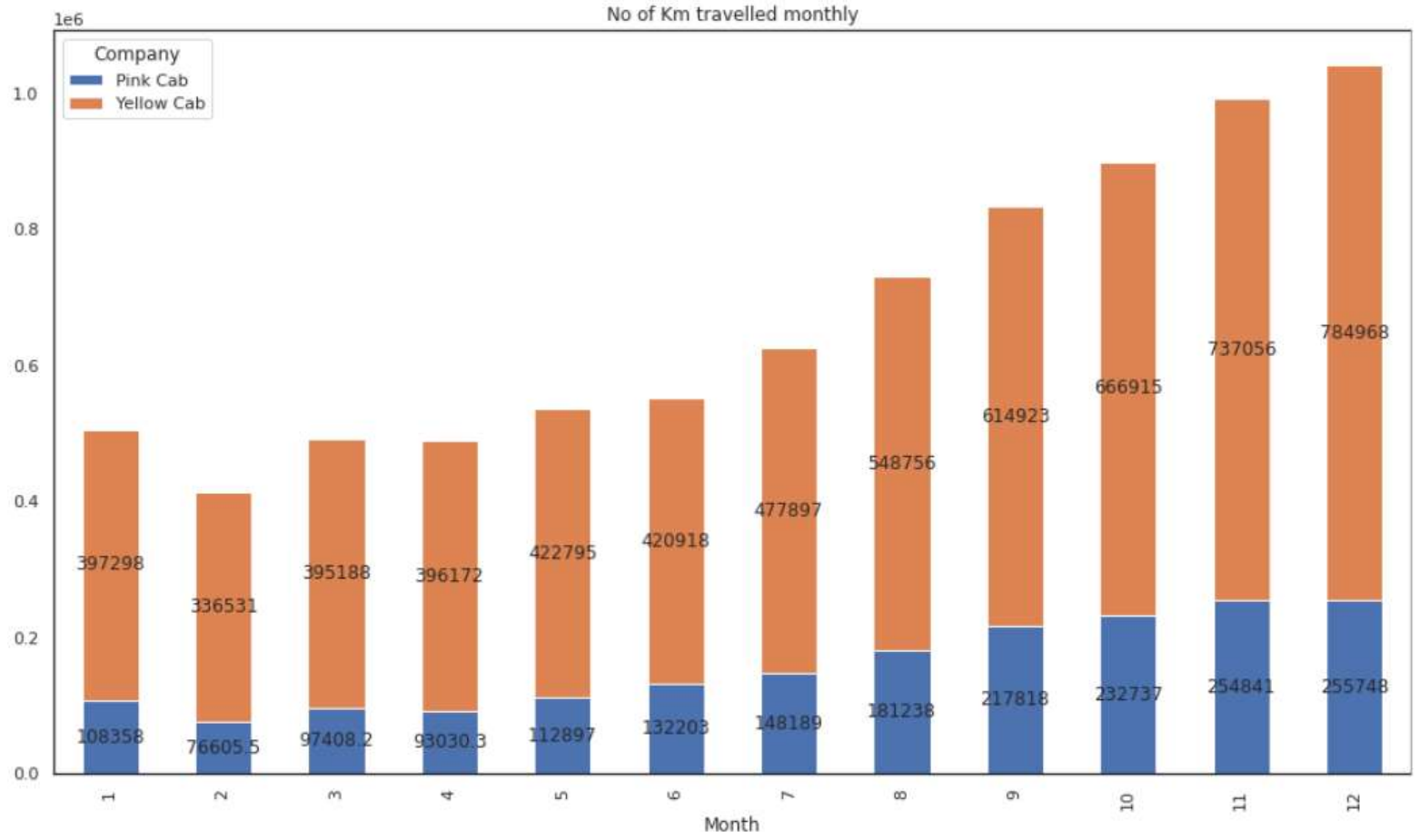
35 si the average age of both male and female who use cab services

customer Analysis



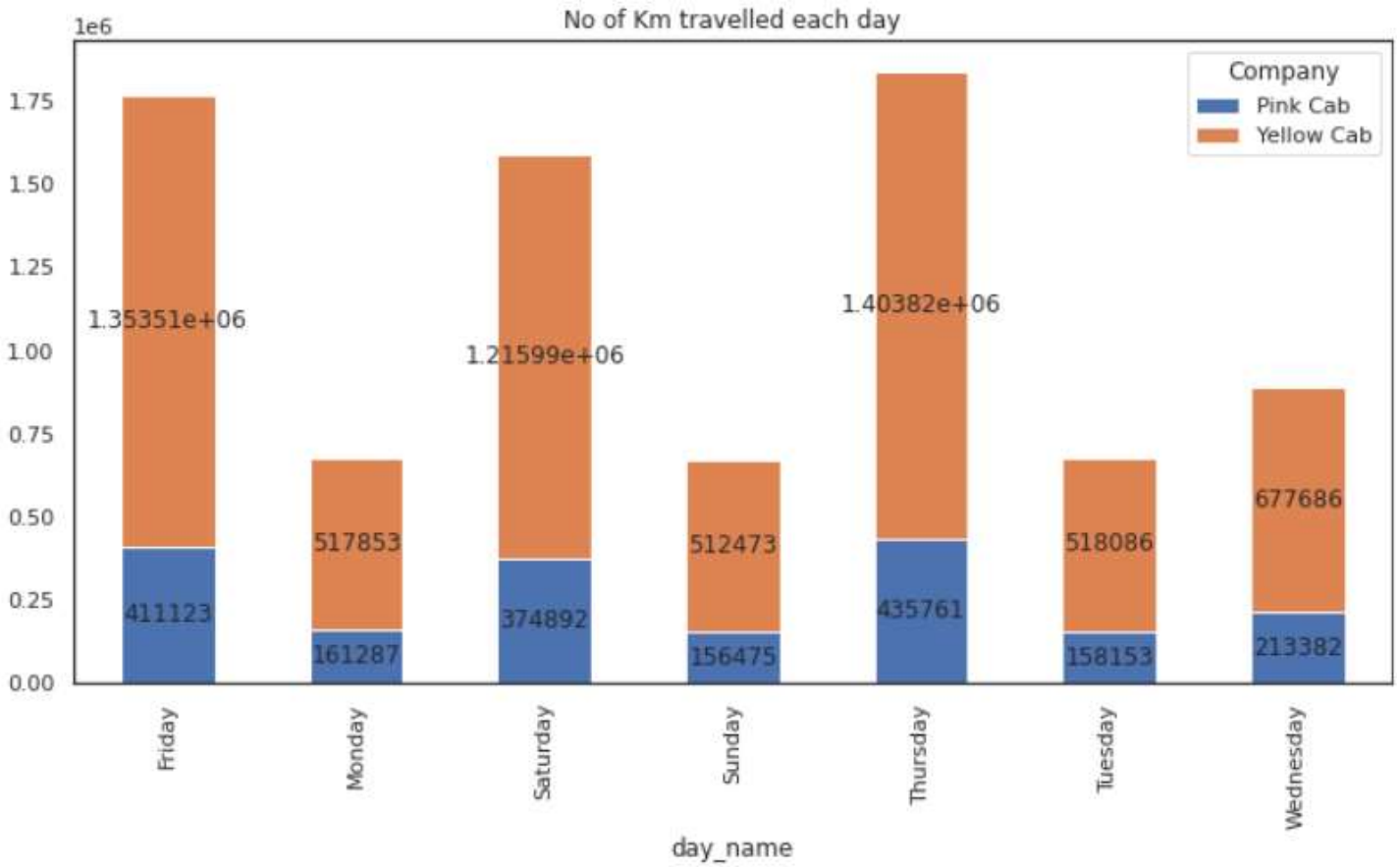
Friday ,Saturday and Thursday are the days that reach high amount of customers with yellow cab exceed pink cab

Km travelled Analysis



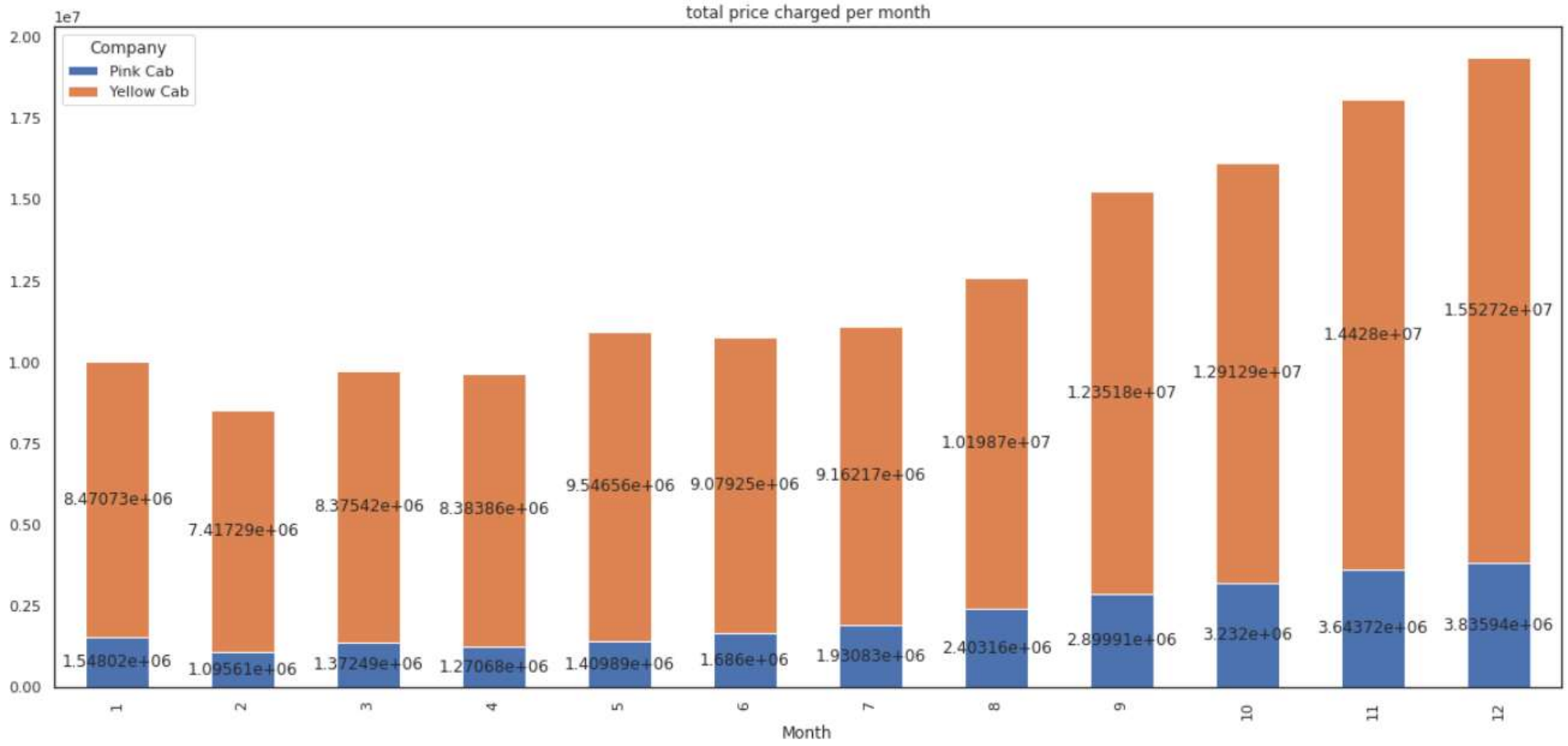
In december customers ride on cabs for long Km compared to other months.
We can notice that no of Km travelled gradually grows over month.

Km travelled Analysis



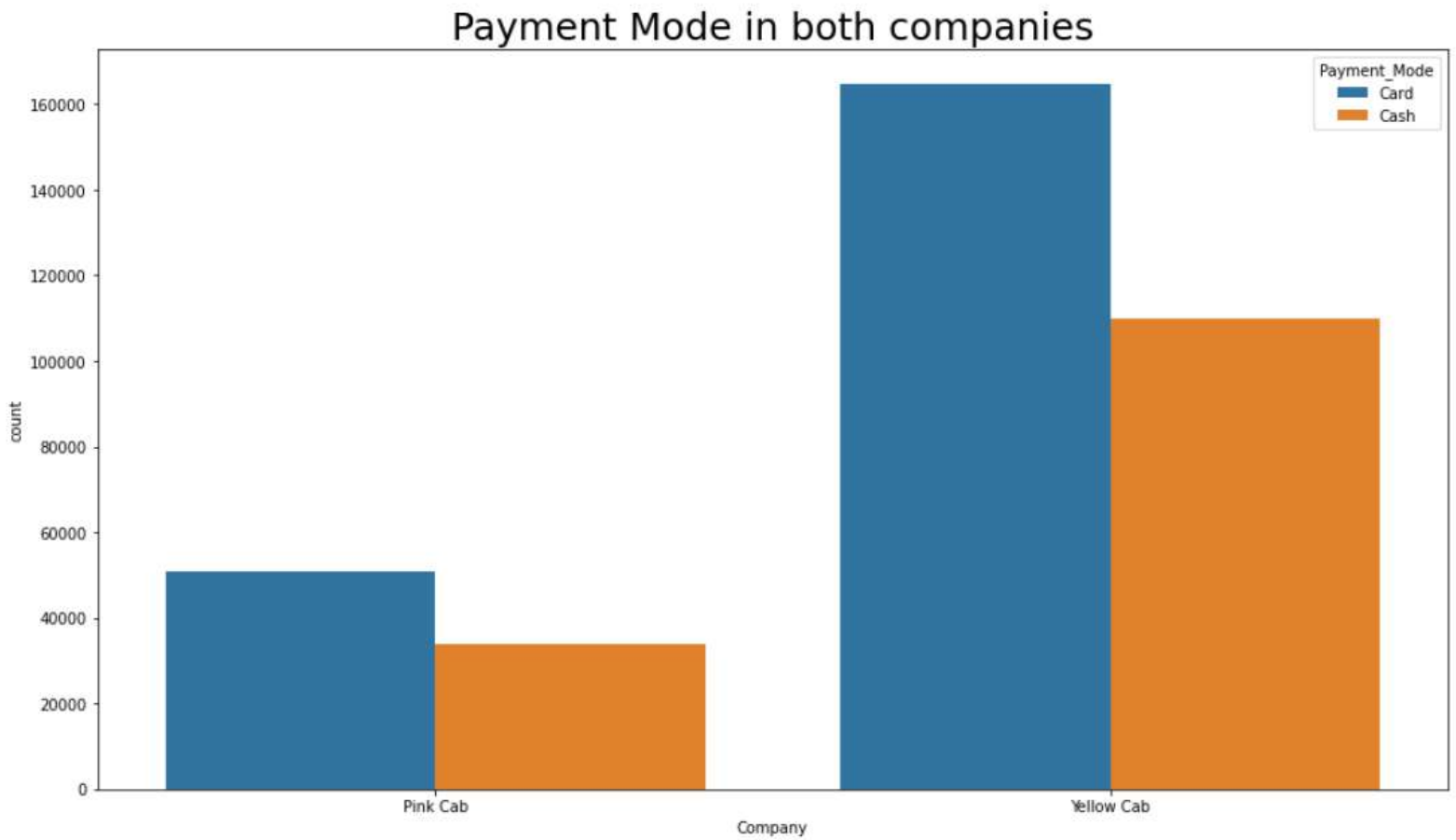
No of Km travelled is higher on Friday , Saturday and Thursday compared to other days.

Price charged per month



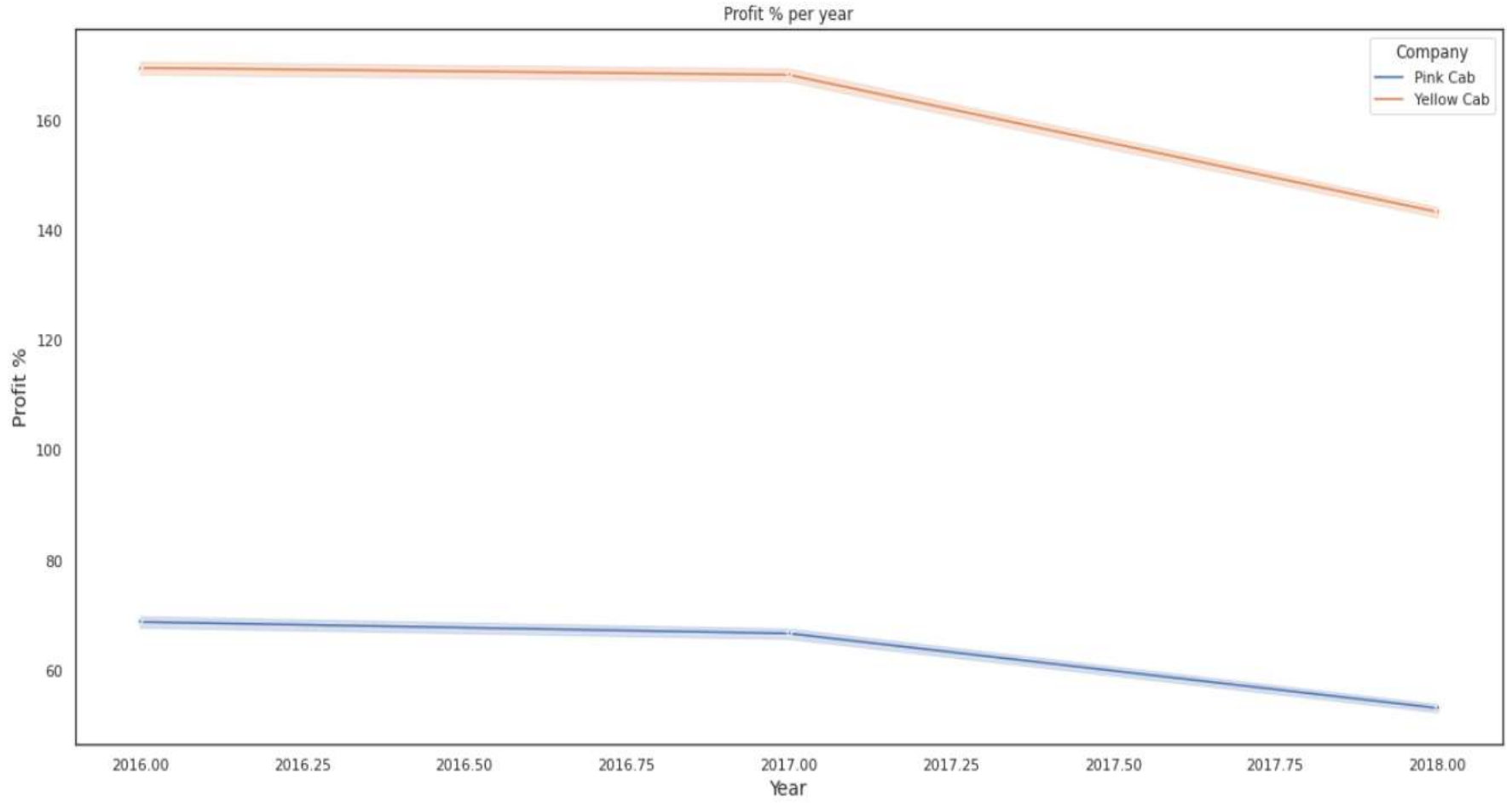
- Price charged increase gradually towards December where it reaches Its maximum .

Payment mode for both company



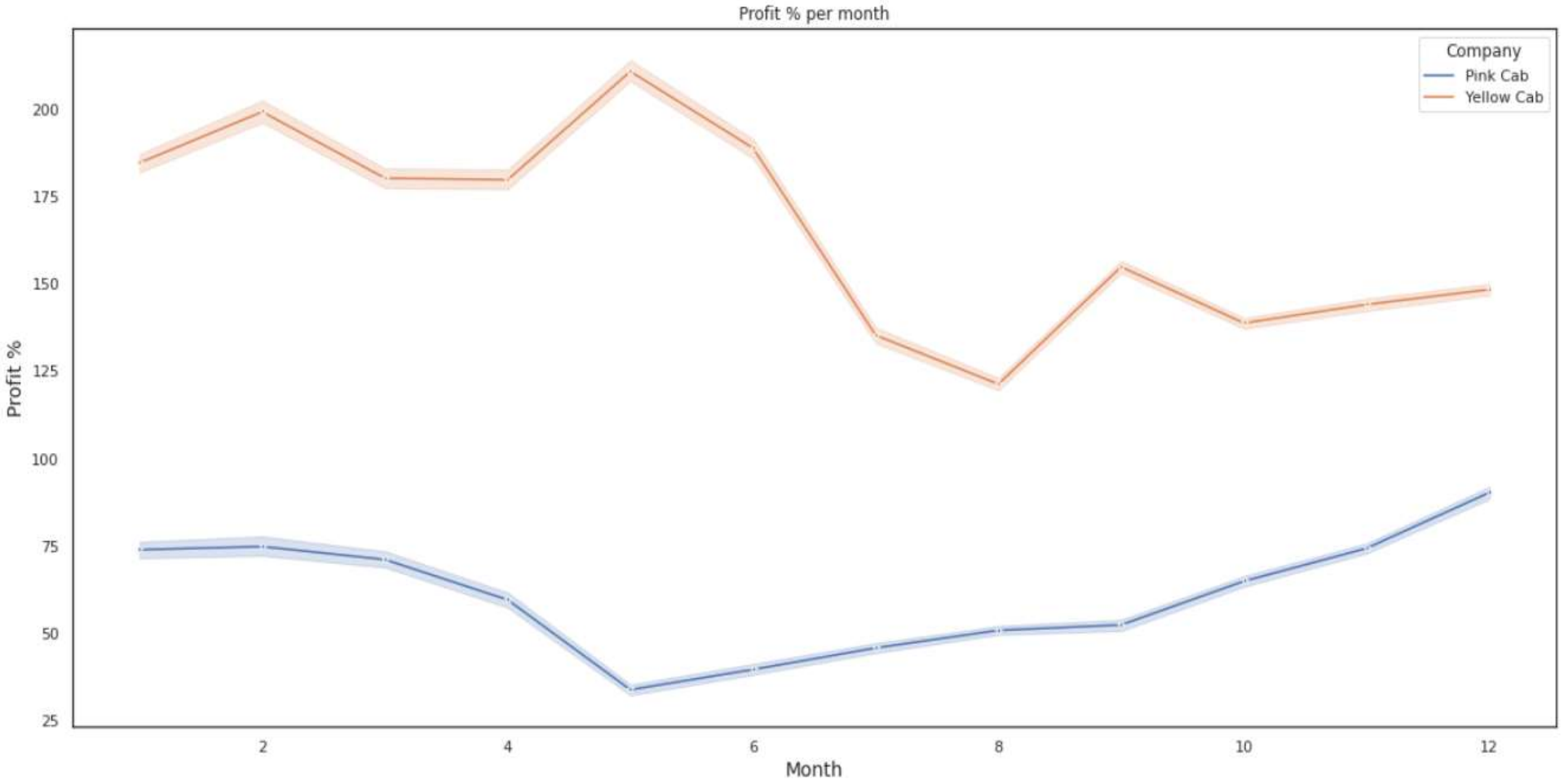
Most of people pay with card rather than cash.

Profit Analysis



Profit margin decrease for both companies

Profit Analysis



Profit varies by months

Hypothesis Testing

*Based on the heatmap displayed above we found that there is a strong linear relationship between profit and price charged with r=0.86. Now let's calculate the p-value of the **pearson test** to decide which hypothesis to accept.*

H0 : there is no linear relationship between the two variables.

H1 : there is a linear relationship between the two variables.

```
[ ]  from scipy.stats.stats import pearsonr
     from scipy import stats

     #calculation correlation coefficient and p-value between x and y
     pearsonr(df['Price Charged'], df['Profit'])
```

```
(0.864153946750663, 0.0)
```

Since the **p-value = 0.0<0.05** so we reject the null hypothesis and accept the alternative hypothesis that states that there is a linear relationship between profit and price charged(when one increase the other increase too and vise verca)

Hypothesis Testing

Does profit differe based on gender?

*For that we will run **t-test** because we want to compare only two groups (male and female) and because one is categorical and the other is numerical*

Lets set the hypothesis:

H0 : there is no difference between the groups in profit.

H1 : there is a difference between the groups in profit

```
[ ]  from scipy import stats

     x = df[(df.Gender=='Male')&(df.Company=='Pink Cab')].groupby('Transaction ID').Profit.mean()
     y = df[(df.Gender=='Female')&(df.Company=='Pink Cab')].groupby('Transaction ID').Profit.mean()
     _, p_value = stats.ttest_ind(x.values,
                                  y.values,
                                  equal_var=True)

     print('P value = ', p_value)

     if(p_value<0.05):
         print('We accept alternative hypothesis (H1) that states that there is a difference between male and female in profit for Pink Cab')
     else:
         print('We accept null hypothesis (H0) that  states that there is no difference between male and female in profit for Pink Cab')
```

```
P value =  0.11515305900425798
We accept null hypothesis (H0) that  states that there is no difference between male and female in profit for Pink Cab
```

Hypothesis Testing

```
[ ]
    x = df[(df.Gender=='Male')&(df.Company=='Yellow Cab')].groupby('Transaction ID').Profit.mean()
    y = df[(df.Gender=='Female')&(df.Company=='Yellow Cab')].groupby('Transaction ID').Profit.mean()
    _, p_value = stats.ttest_ind(x.values,
                                 y.values,
                                 equal_var=True)

    print('P value = ', p_value)

    if(p_value<0.05):
        print('We accept alternative hypothesis (H1) that states that there is a difference between male and female in profit for Yellow Cab')
    else:
        print('We accept null hypothesis (H0) that  states that there is no difference between male and female in profit for Yellow Cab')
```

```
P value =  6.060473042494144e-25
We accept alternative hypothesis (H1) that states that there is a difference between male and female in profit for Yellow Cab
```

Hypothesis Testing

Does profit differe based on age?

```
[ ]  x = df[(df.Age <= 40)&(df.Company=='Pink Cab')].groupby('Transaction ID').Profit.mean()
     y = df[(df.Age >= 40)&(df.Company=='Pink Cab')].groupby('Transaction ID').Profit.mean()
     _, p_value = stats.ttest_ind(x.values,
                                  y.values,
                                  equal_var=True)

     print('P value = ', 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')
```

```
P value =  0.09093510590632374
We accept null hypothesis (H0) that there is no difference regarding age for Pink Cab
```

Hypothesis Testing

```
[ ]  a = df[(df.Age <= 40)&(df.Company=='Yellow Cab')].groupby('Transaction ID').Profit.mean()
     b = df[(df.Age >= 40)&(df.Company=='Yellow Cab')].groupby('Transaction ID').Profit.mean()
     _, p_value = stats.ttest_ind(a.values,
                                  b.values,
                                  equal_var=True)

     print('P value = ', p_value)

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

```
P value =  0.44246196729249976
We accept null hypothesis (H0) that there is no difference regarding age for yellow Cab
```

Recommendations

We have evaluated both the cab companies on following points and found Yellow cab better than Pink cab:

- **Customer Reach :** Yellow cab has higher customer. We have also observed that Yellow cab is doing good in covering other cab users as compared to Pink cab.
- **Profit :** yellow cab reaches higher profit compared o pink cab
- **Transaction :** yellow cab have more transaction per year than pink cab

On the basis of above point , we will recommend Yellow cab for investment.

Thank You