Exploratory Data Analysis

G2M-insight-for-Cab-Investment-firm

19.06.2022

Agenda

Executive Summary

Problem Statement

Approach

EDA

EDA summary and Recommendations

Some hypothesis testing

Data Glacier

Your Deep Learning Partner

Executive Summary

- XYZ is a private 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 meaningful business insights to help XYZ firm in identifying the right company for making them the right investment.

Problem Statement

There is no one dataset to extract meaningful insights
So, before merging them(creating master dataset);

-Identify relationships across the files
-Field/feature transformations
-Determine which files should be joined versus which ones should be appended…

After creating master dataset:

- Forecasting profit and number of rides for each cab type
- Finding the most profitable Cab company
- Recommendations for investment

Approach

- 21 Features(including 6 derived features)
- Timeframe of the data: 2016-01-31 to 2018-12-31
- Total data points : 359,854

Normally 4 data set given as csv files;

-Cab_Data.csv

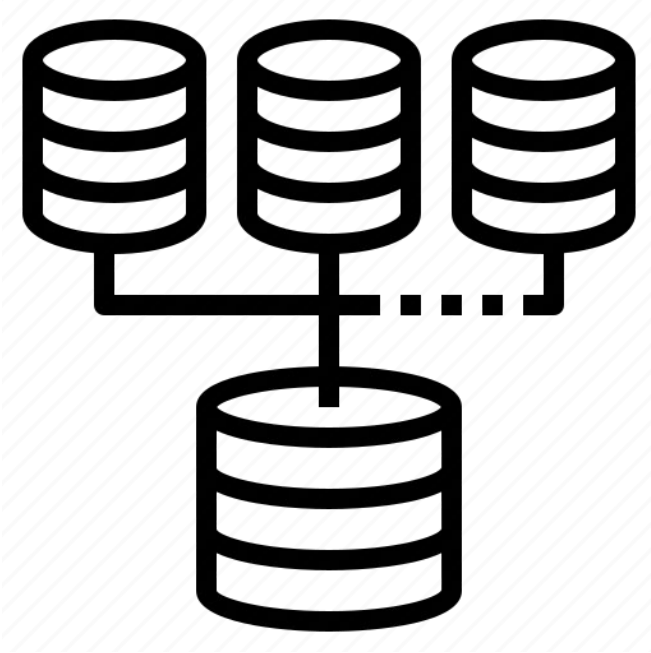
-Customer_ID.csv

-Transaction_ID.csv

-City.csv

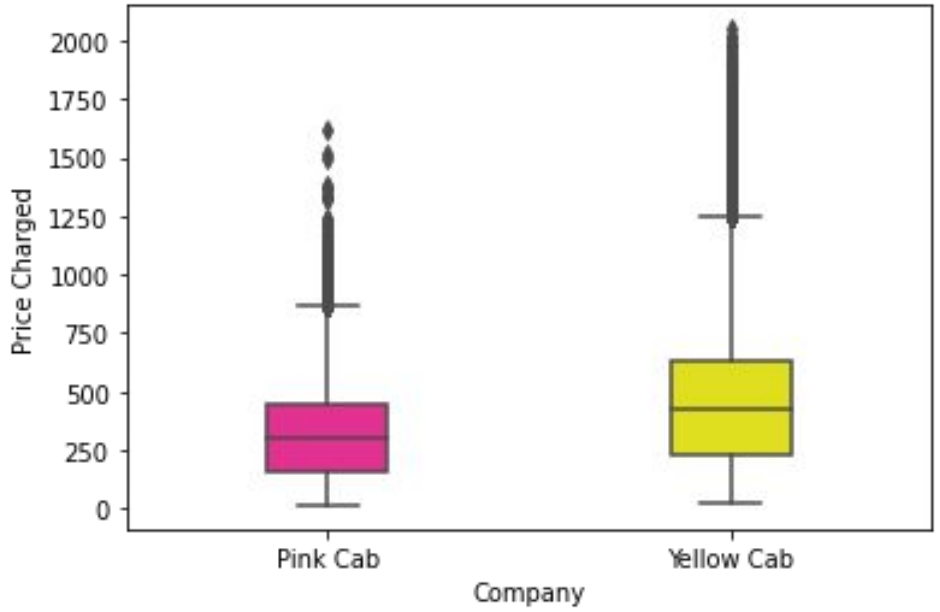
One external but related data set that is used;

-US Holiday Dates (2004-2021).csv

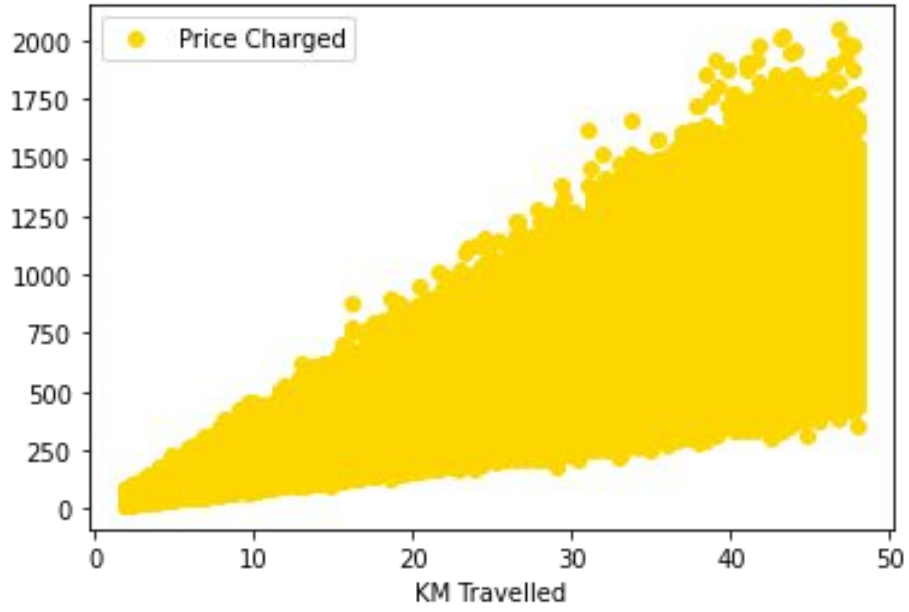


Merged Master data frame is created from these 5 files

EDA



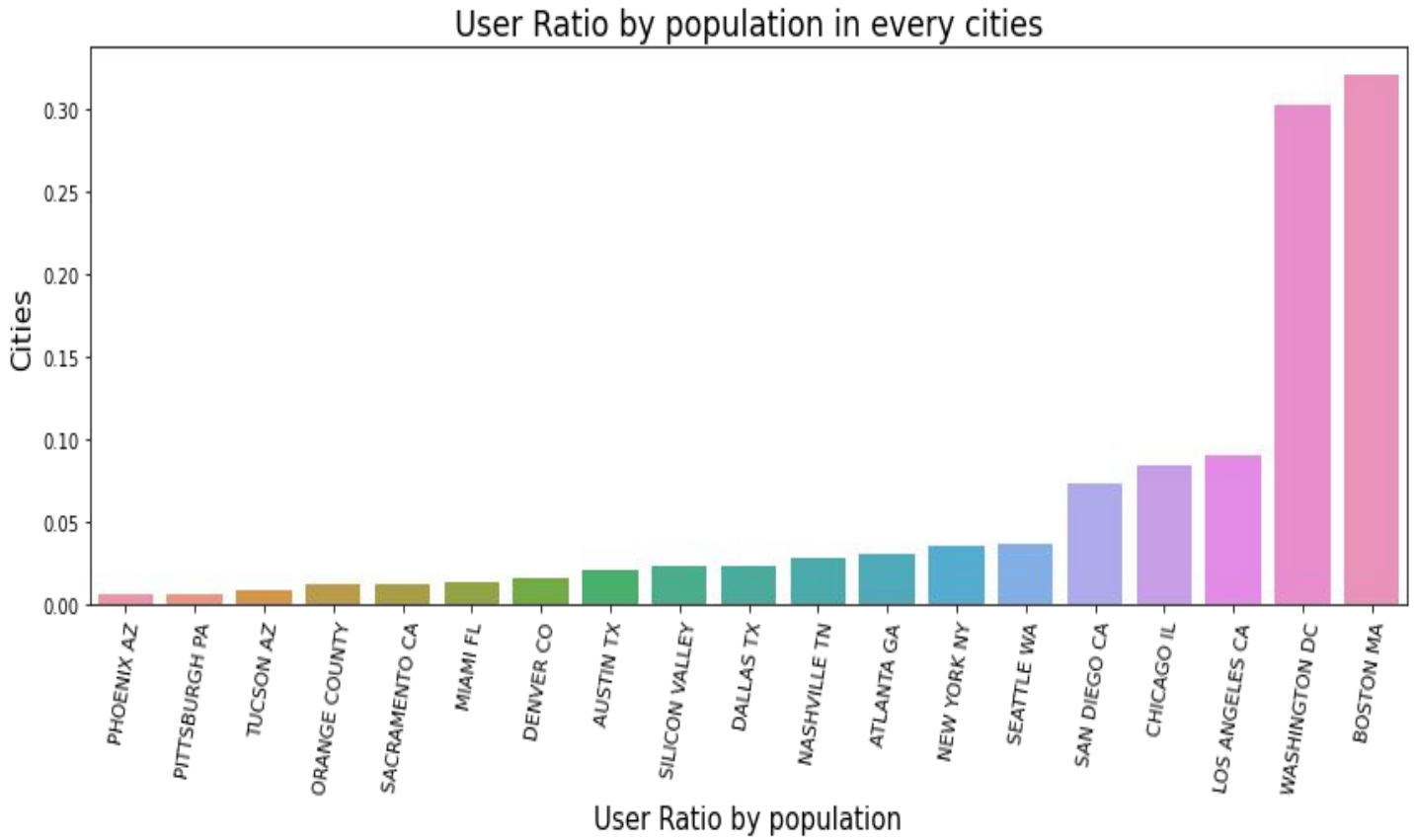
Plot box distributions of both companies, we can see that there are too many outliers



Correlation coefficient between Distance and Price charged:0.835…

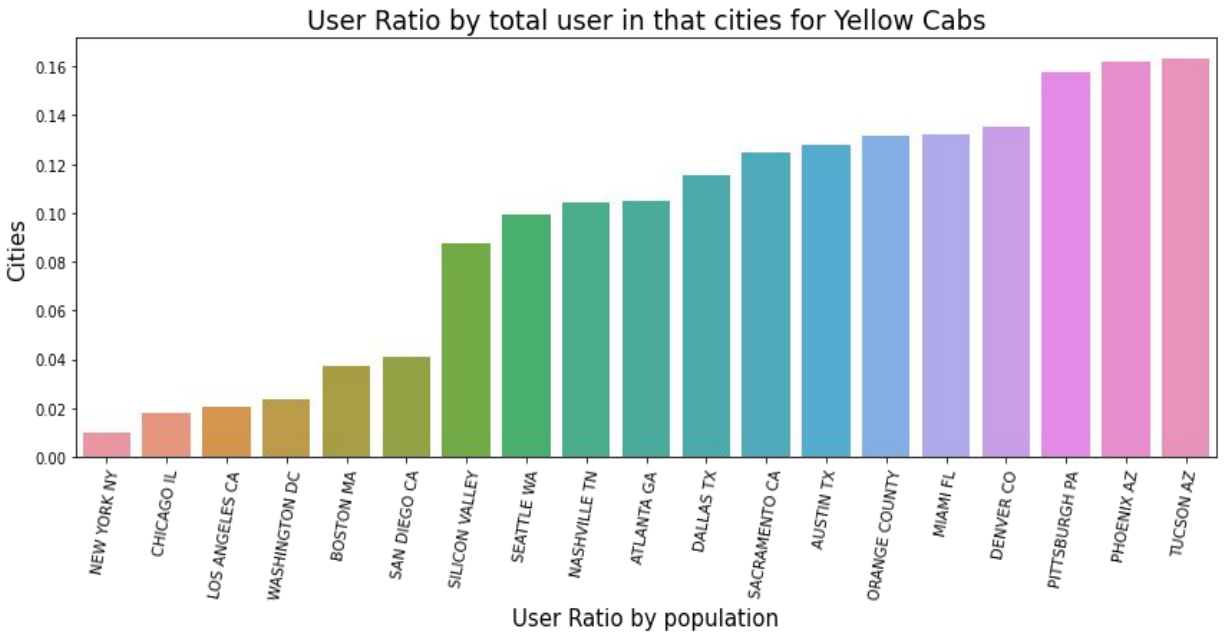
Since there is a high correlation coefficient and we can also see visually traveling distance is one of the reasons for high prices that are charged. Also, since we do not have any further information about other causes, it is better to not treat high prices as outliers in the "Price Charged" column.

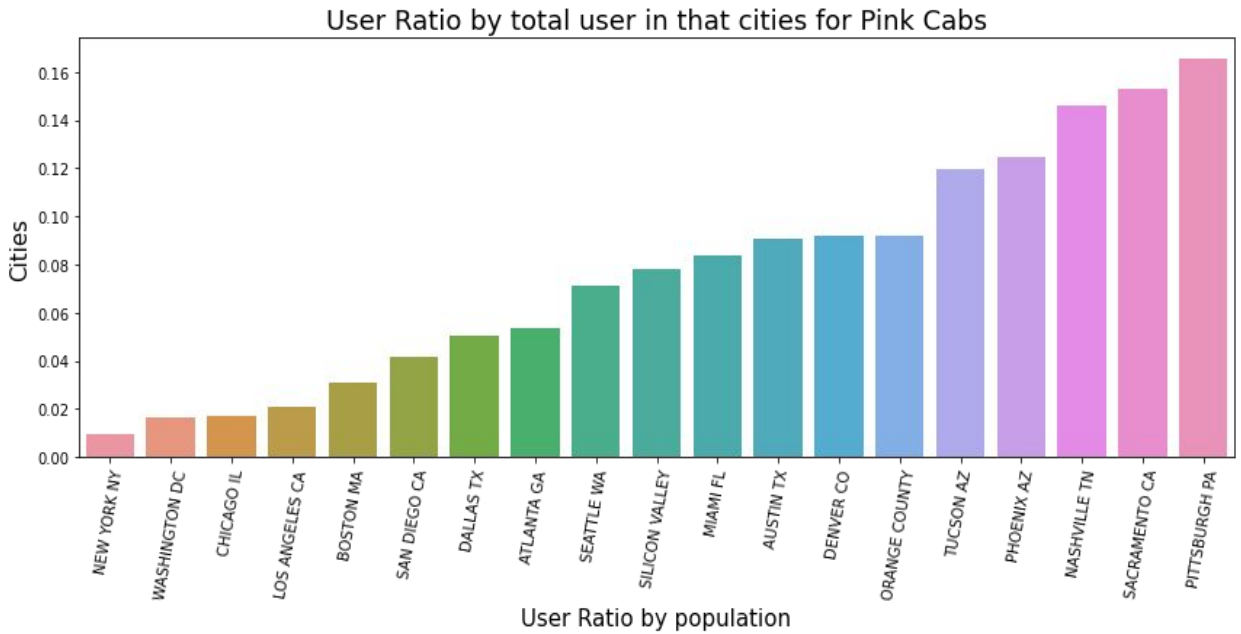
City based Analyses



This is just a general information that is extracted from data set, which represents how many of people(ratio) is using taxis, cabs in comparison with total population of that city.

City based Analyses



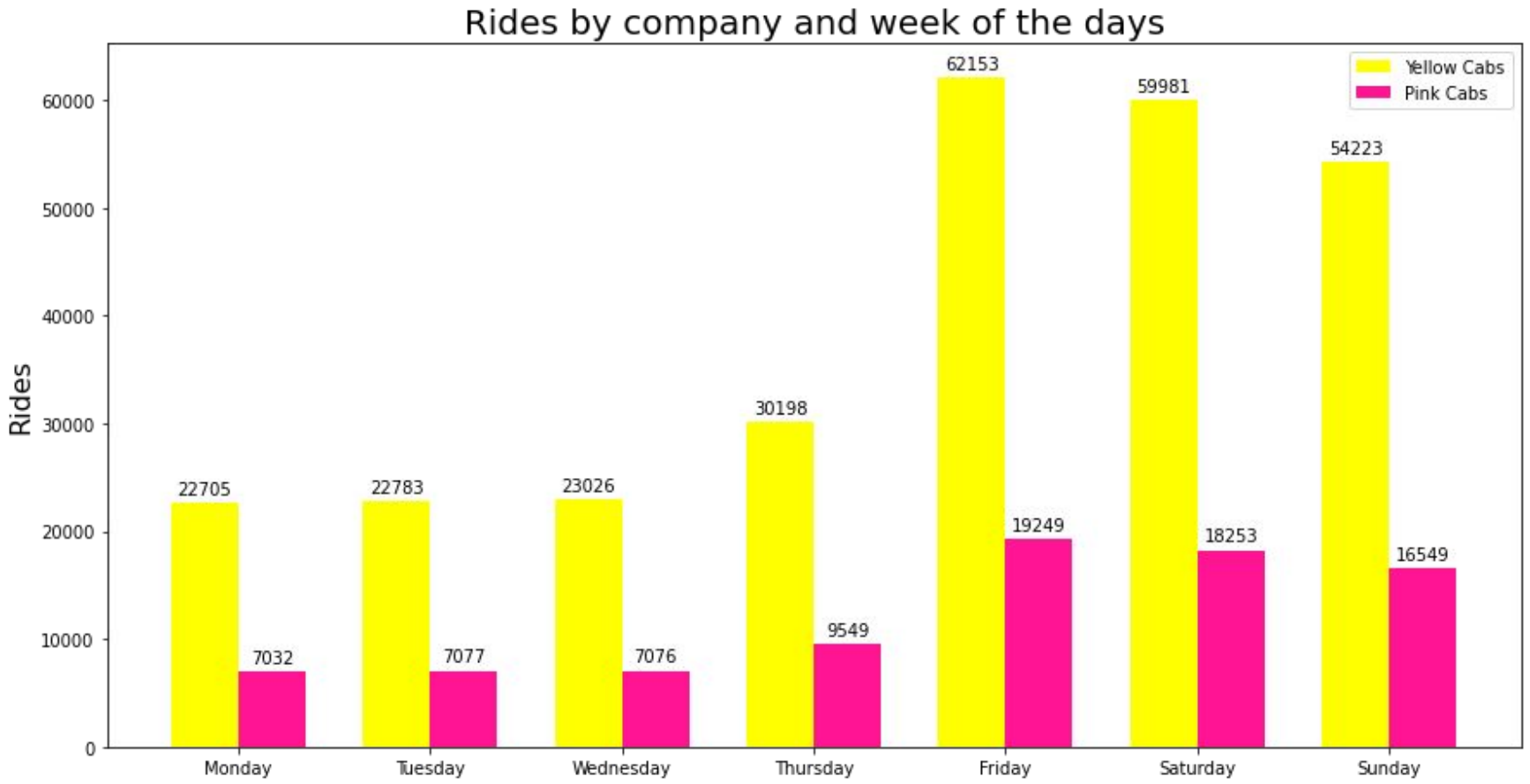


On average yellow cabs has the 9.4% of the total users where they have service for stated cities.

On average pink cabs has the 7.6% of the total users where they have service for stated cities.

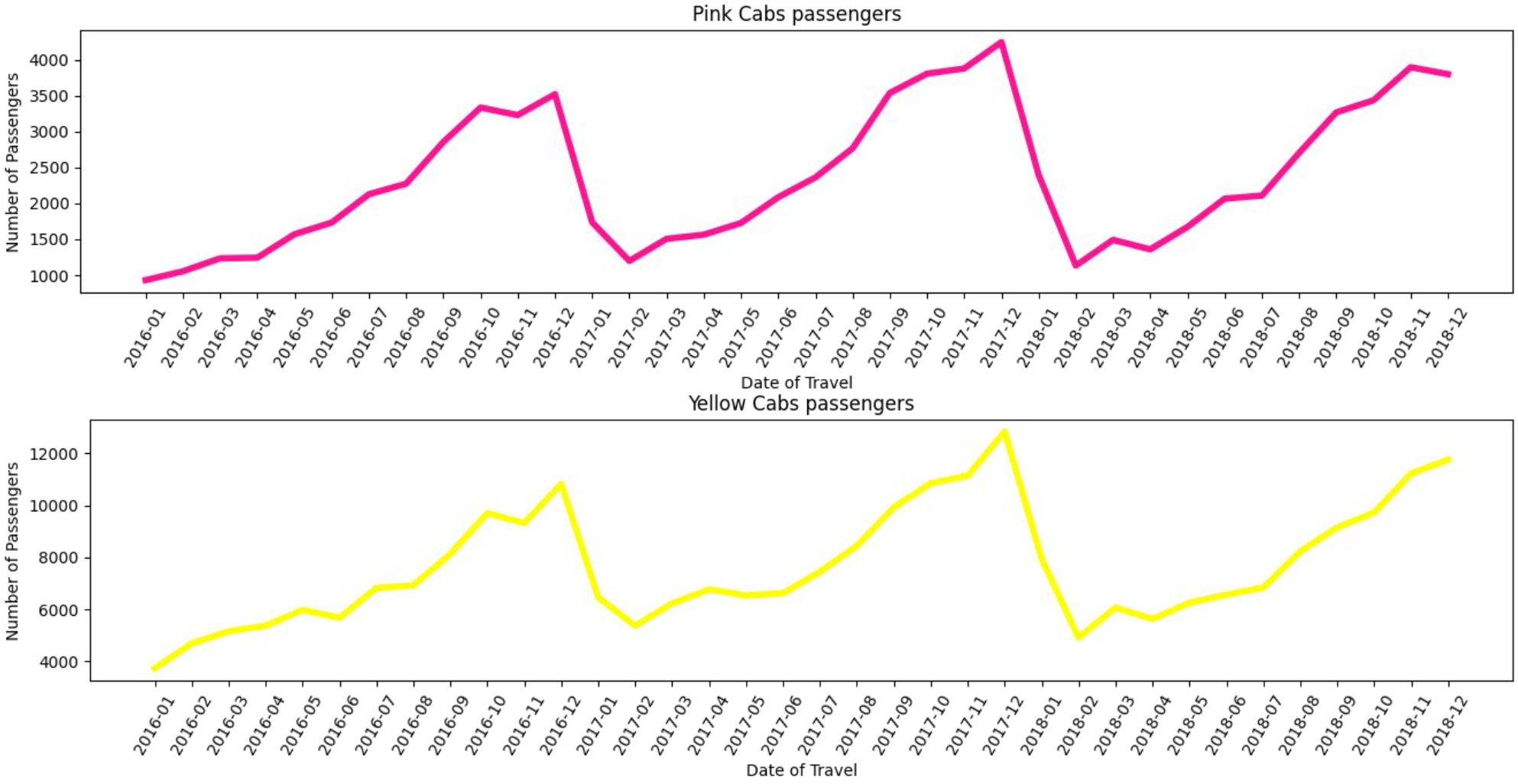
For both Yellow and Pink Cabs the cities that have least user proportion is common, they are "New York NY", "Chicago IL", "Los Angeles CA", "Washington DC", "Boston MA" and "San Diego CA". Assuming that these are very crowded and big cities and competition by other companies are also a very big threat.

Rides vs Days of the week analysis



Yellow cabs have the by far higher rides in total and for everyday, also from this graph we can see that Fridays Saturdays and Sundays are very significant for cab industry.

Seasonality analysis



Seasonality analysis

- **When we applied granger causality tests(to see if there is seasonality) to companies monthly passenger numbers:**

-Pink Cabs

ssr based F test: F=6.1813 , p=0.0183 ,
ssr based chi2 test: chi2=6.7608 , p=0.0093
likelihood ratio test: chi2=6.1813 , p=0.0129
parameter F test: F=6.1813 , p=0.0183

In the above case, the p-values are close to 0("zero") for all tests. So the 'month' indeed can be used to forecast the values for pink cabs.

-Yellow Cabs

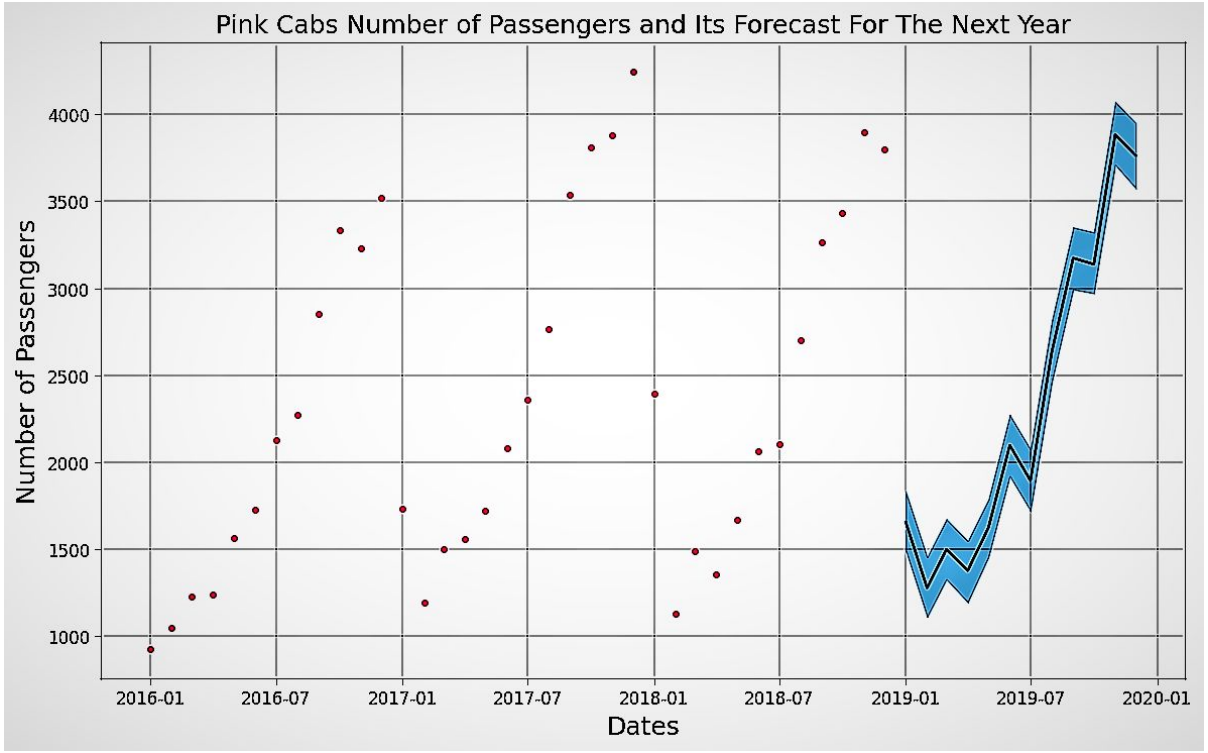
ssr based F test: F=10.0768 , p=0.0033
ssr based chi2 test: chi2=11.0215 , p=0.0009
likelihood ratio test: chi2=9.5816 , p=0.0020
parameter F test: F=10.0768 , p=0.0033

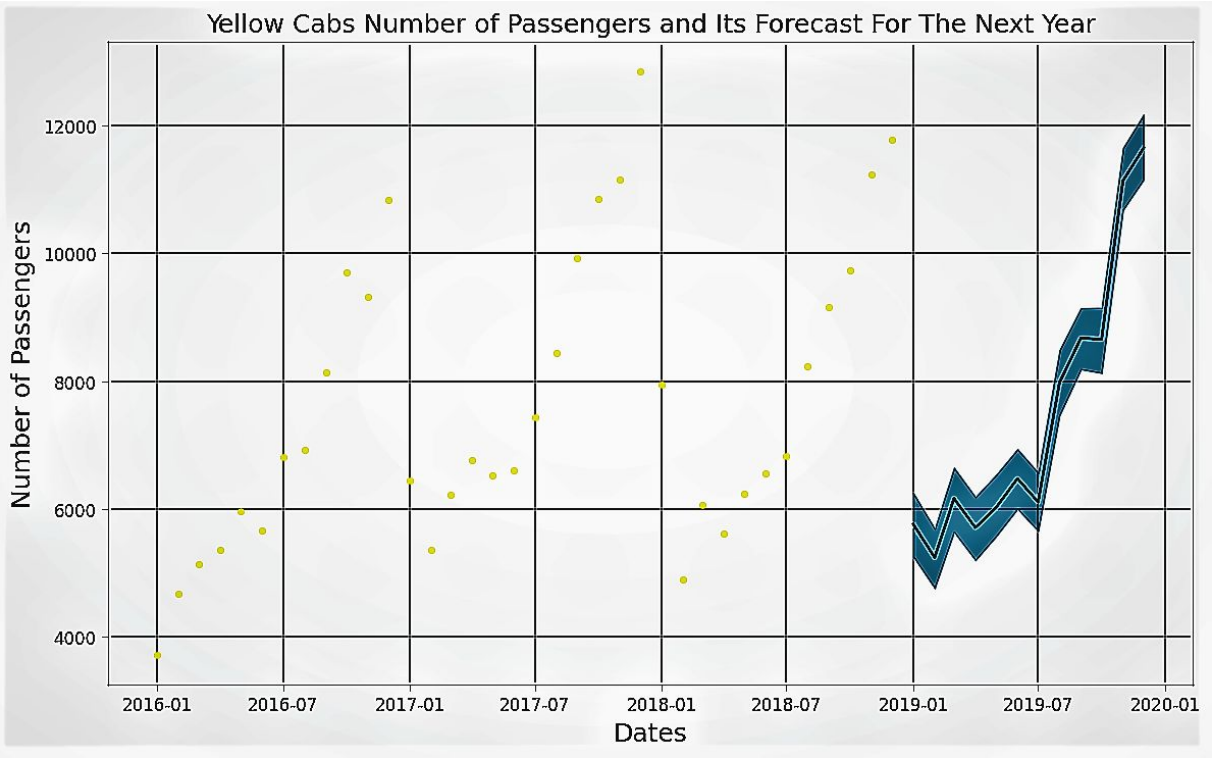
In the above case, the p-values are close to 0("zero") for all tests. So the 'month' indeed can be used to forecast the values for yellow cabs.

Most importantly both companies' data show seasonality as the test results reviewed

**For both companies multiplicative, additive decomposition and autocorrelation methods are applied and visualised the results in python codes as well.

Forecasting

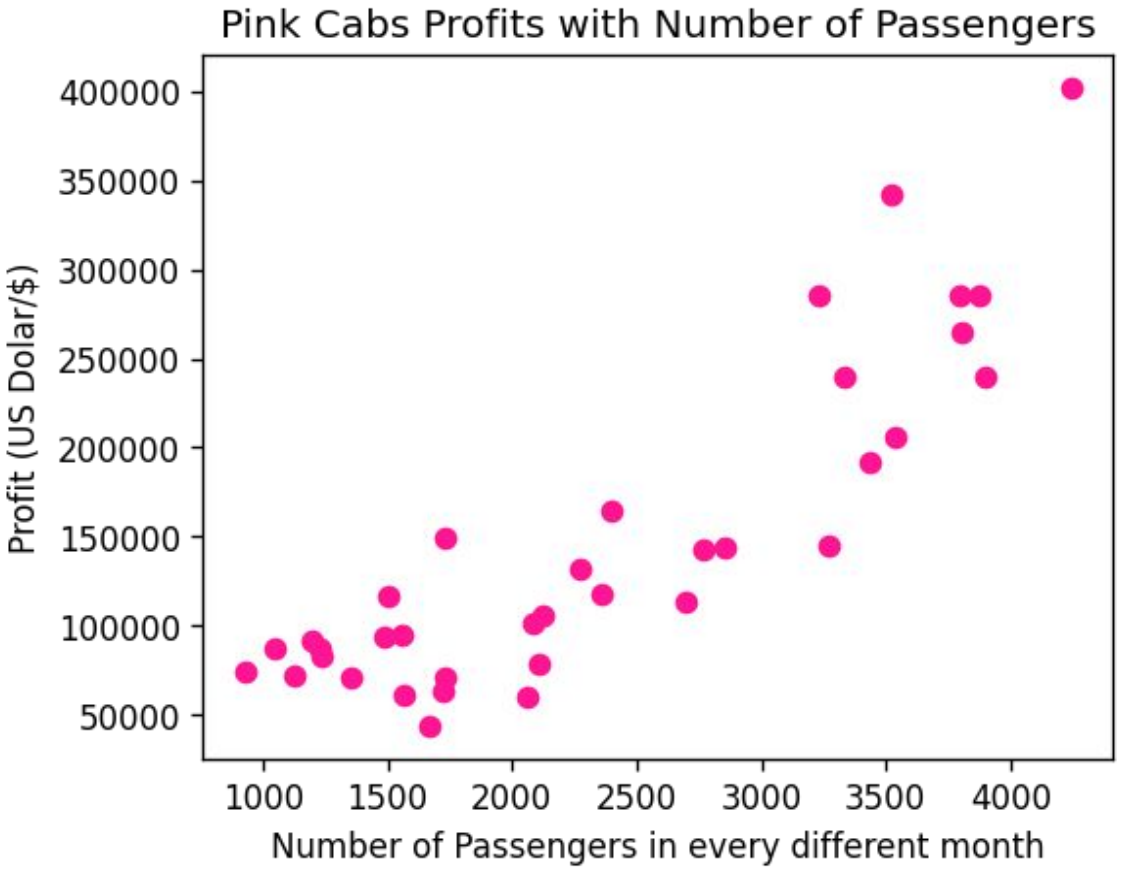




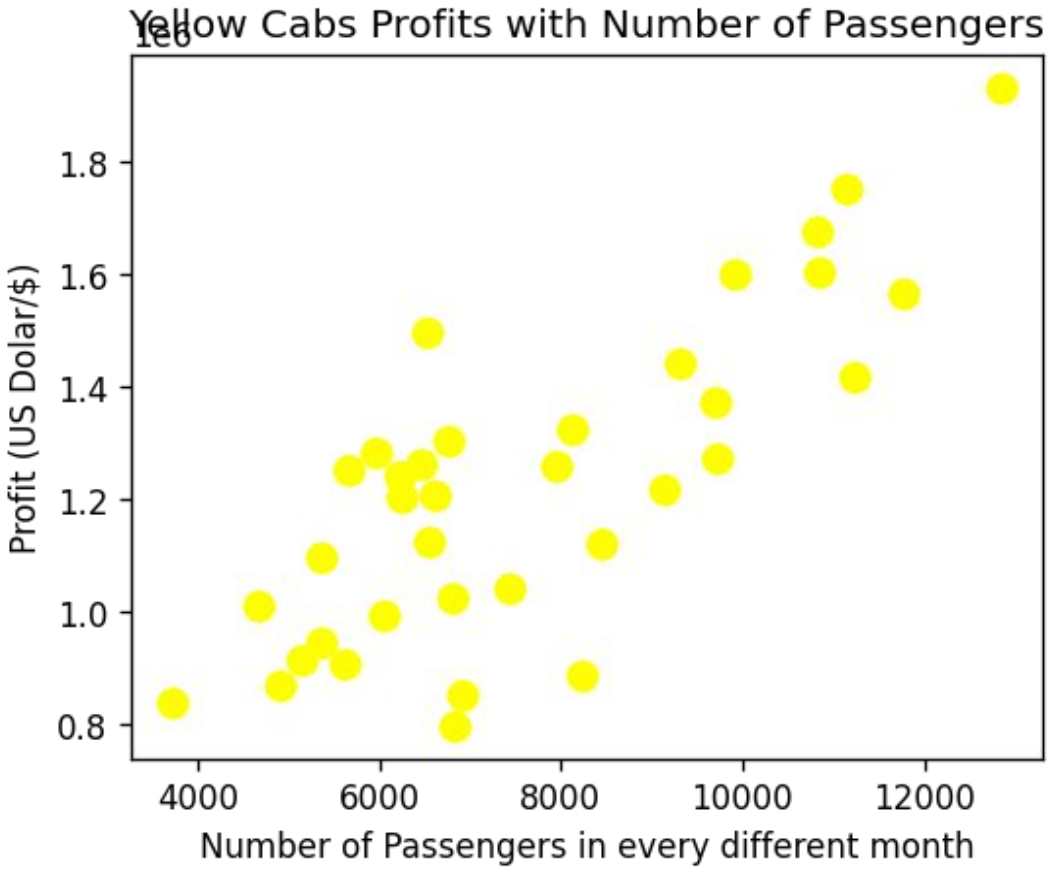
Since seasonality test we used are passed positively, we could try to forecast for the next year.
As it can be seen the number of passengers of Yellow cabs would have still around 2 times higher than the Pink ones.

****Forecasting done by Prophet model in python**

Relation between profit and rides/customers

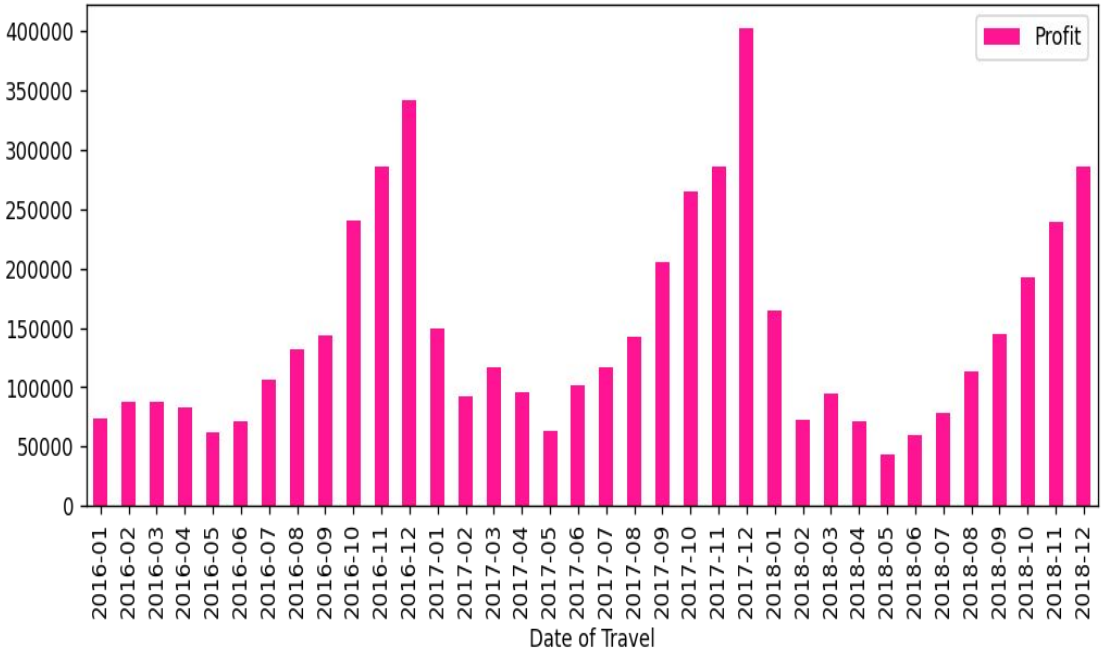


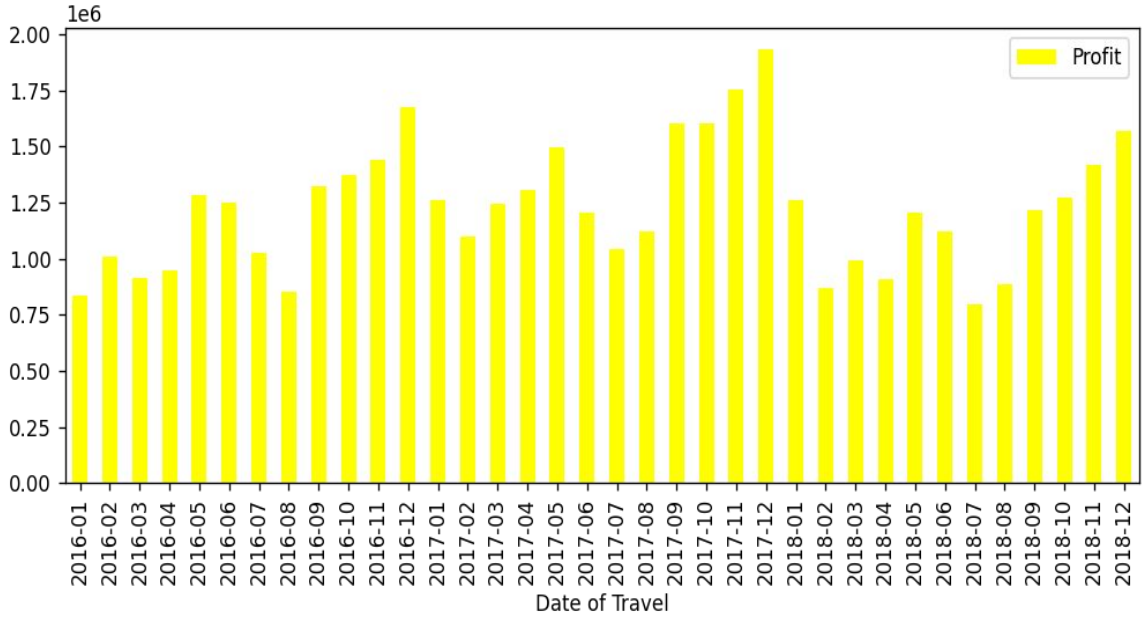
Correlation coefficient: 0.87



Correlation coefficient: 0.79

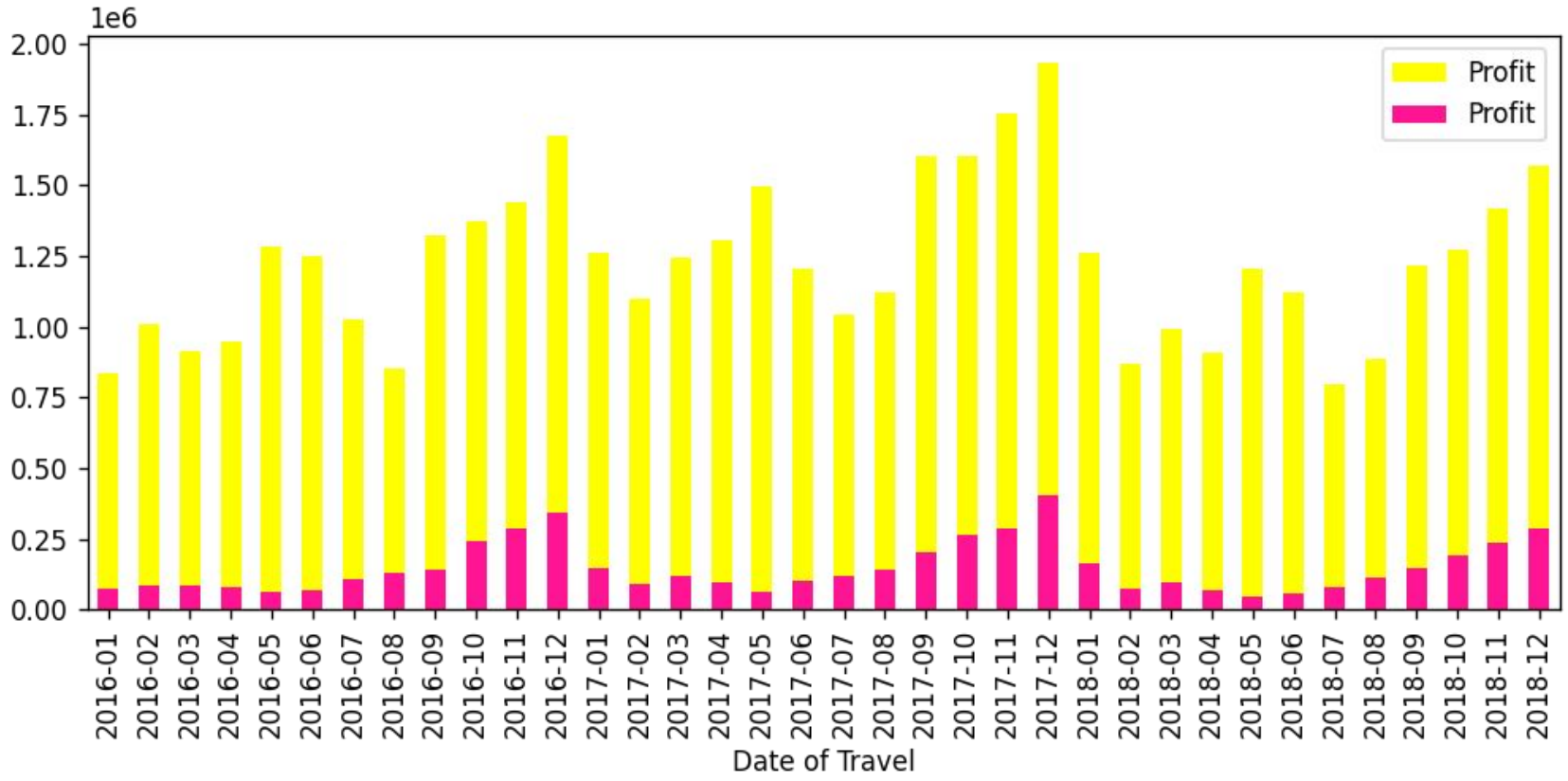
Profits





-Relative profit can be seen in the next slide

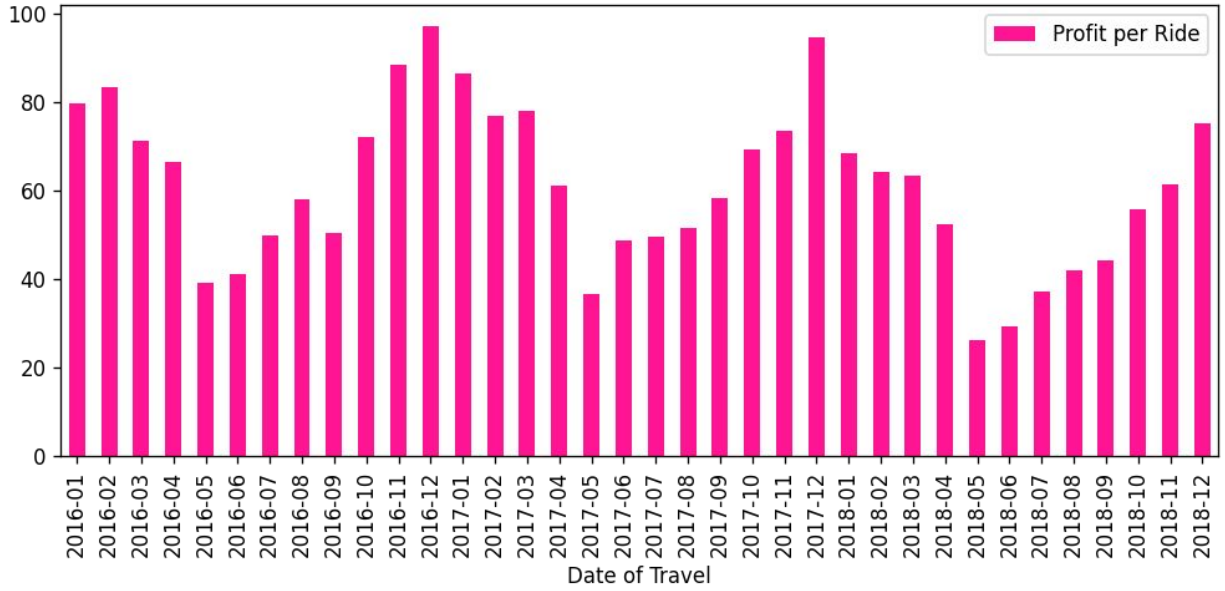
Profits

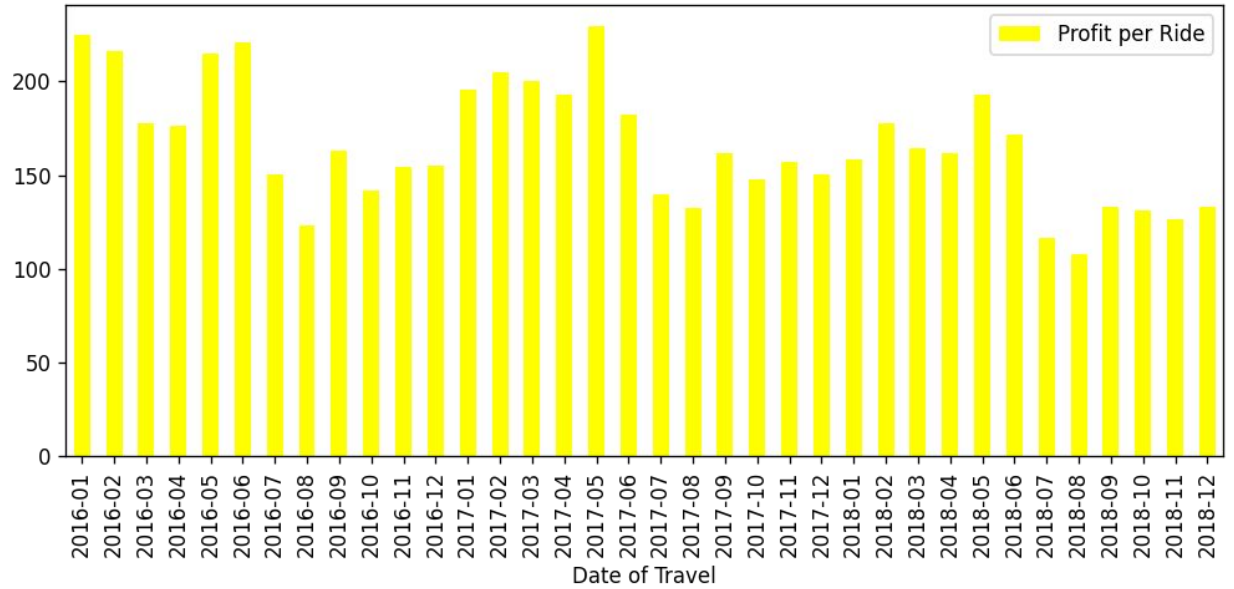


Average profit per month for Yellow cabs: $ 1,225,128.4
Average profit per month for Pink cabs: $ 147,522

Ratio between them Yellows/Pinks: $ 8.3
Total profit for Yellow cabs: $ 44,104,621
Total profit for Pink cabs: $ 5,310,794.7

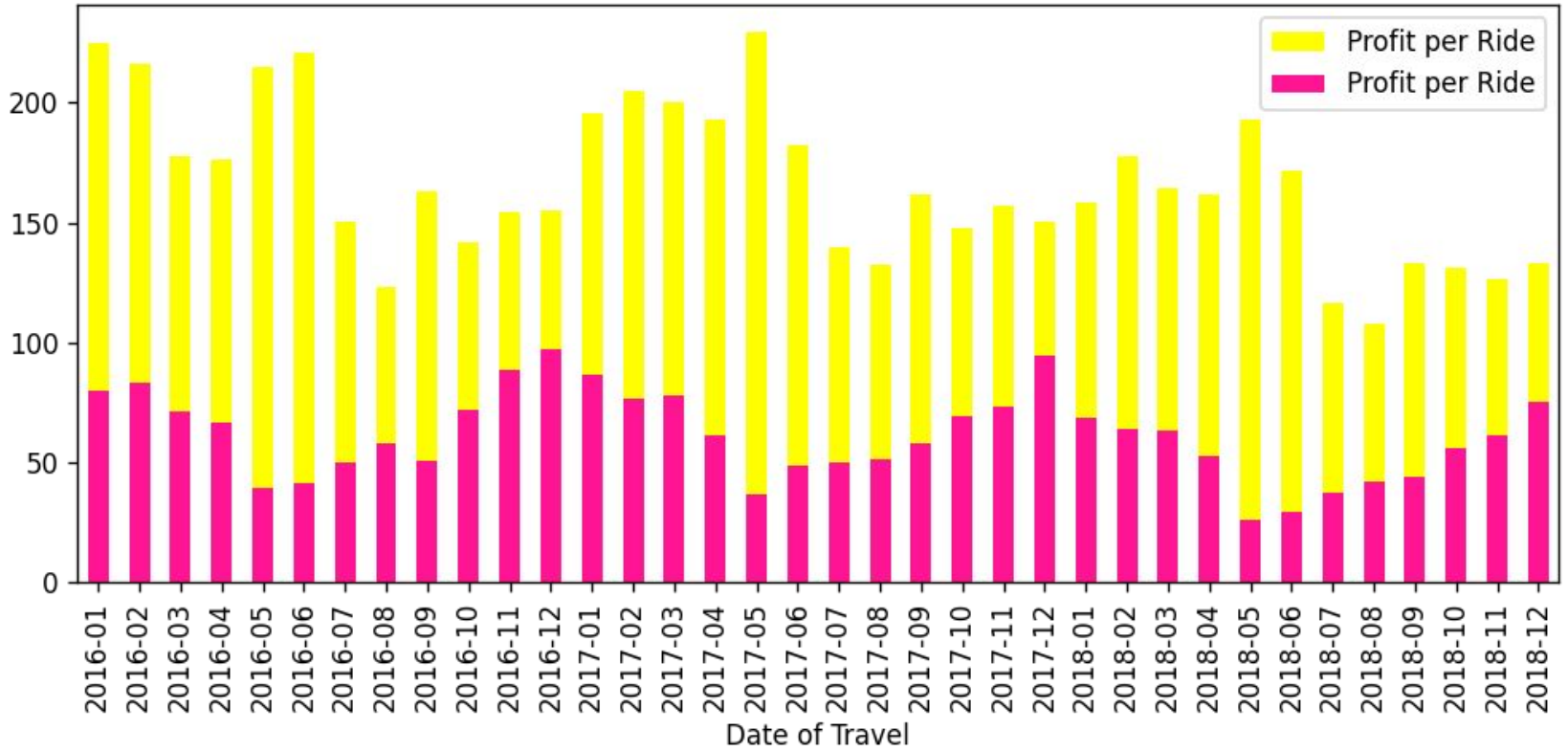
Profits per Ride





-Relative profit/ride can be seen in the next slide

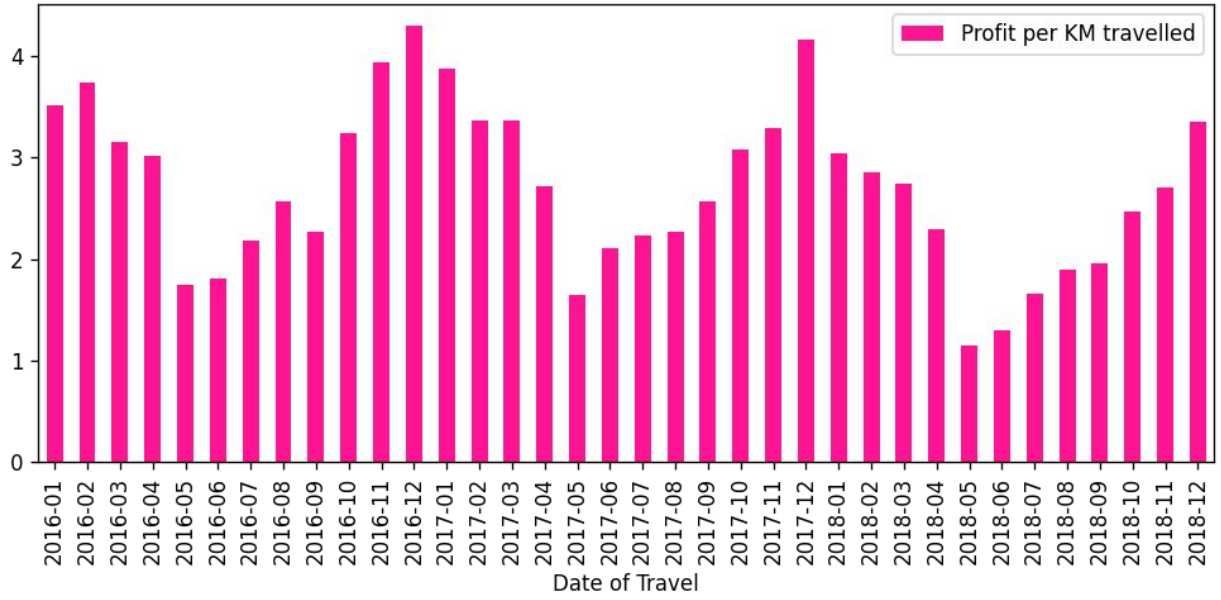
Profits per Ride

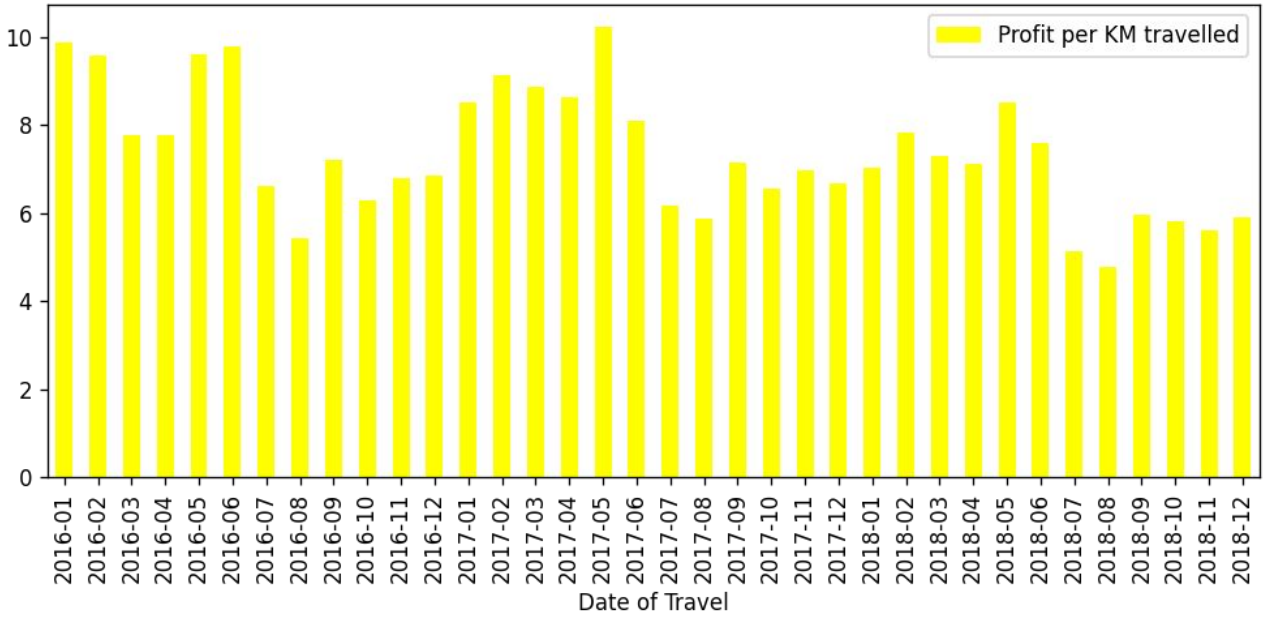


Average profit per ride and month for Yellow cabs: $ 166.26
Average profit per ride and month for Pink cabs: $ 61.18

Ratio between them Yellows/Pinks: $ 2.7

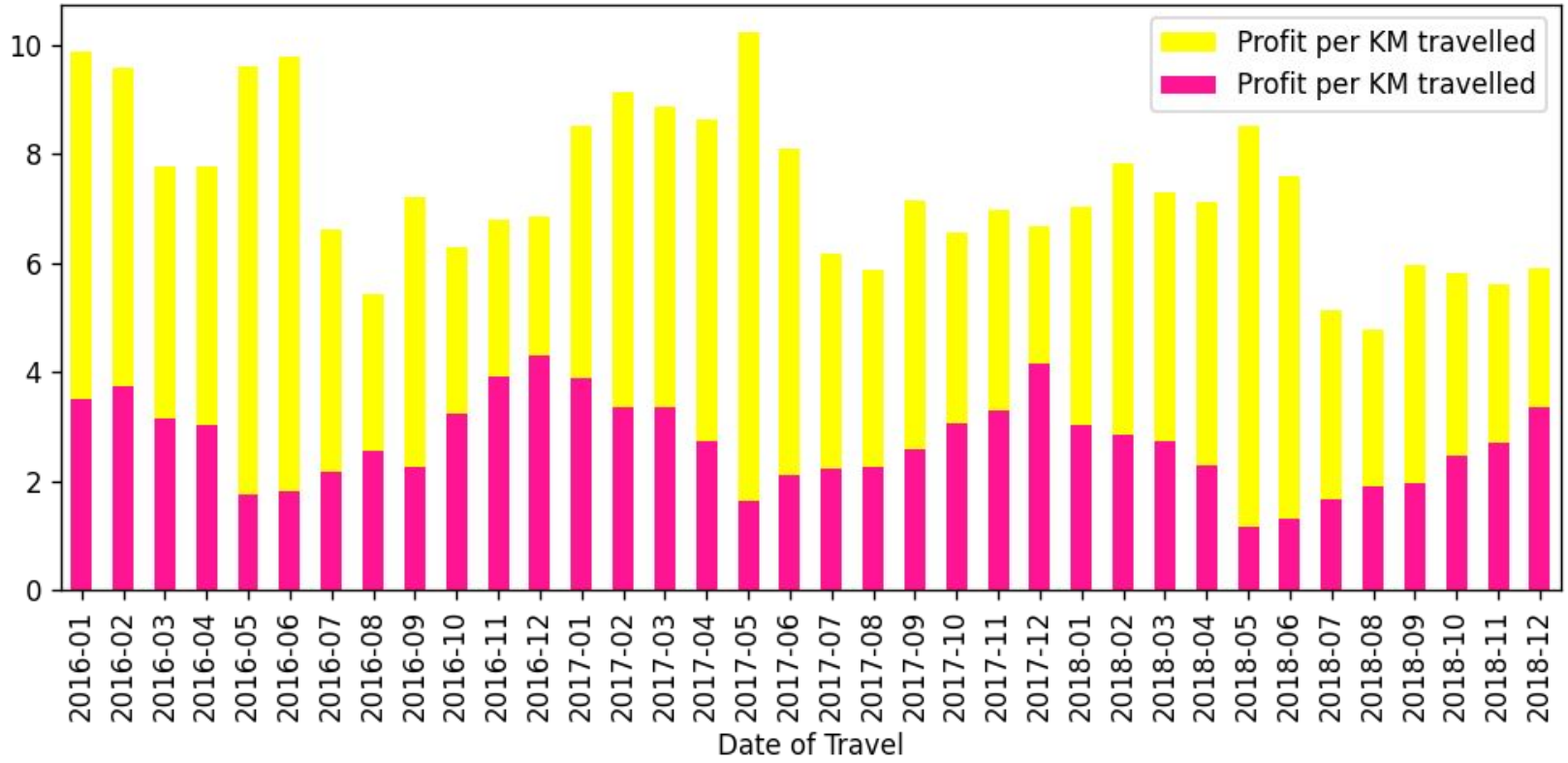
Profits per Distance(KM)





-Relative profit/ride can be seen in the next slide

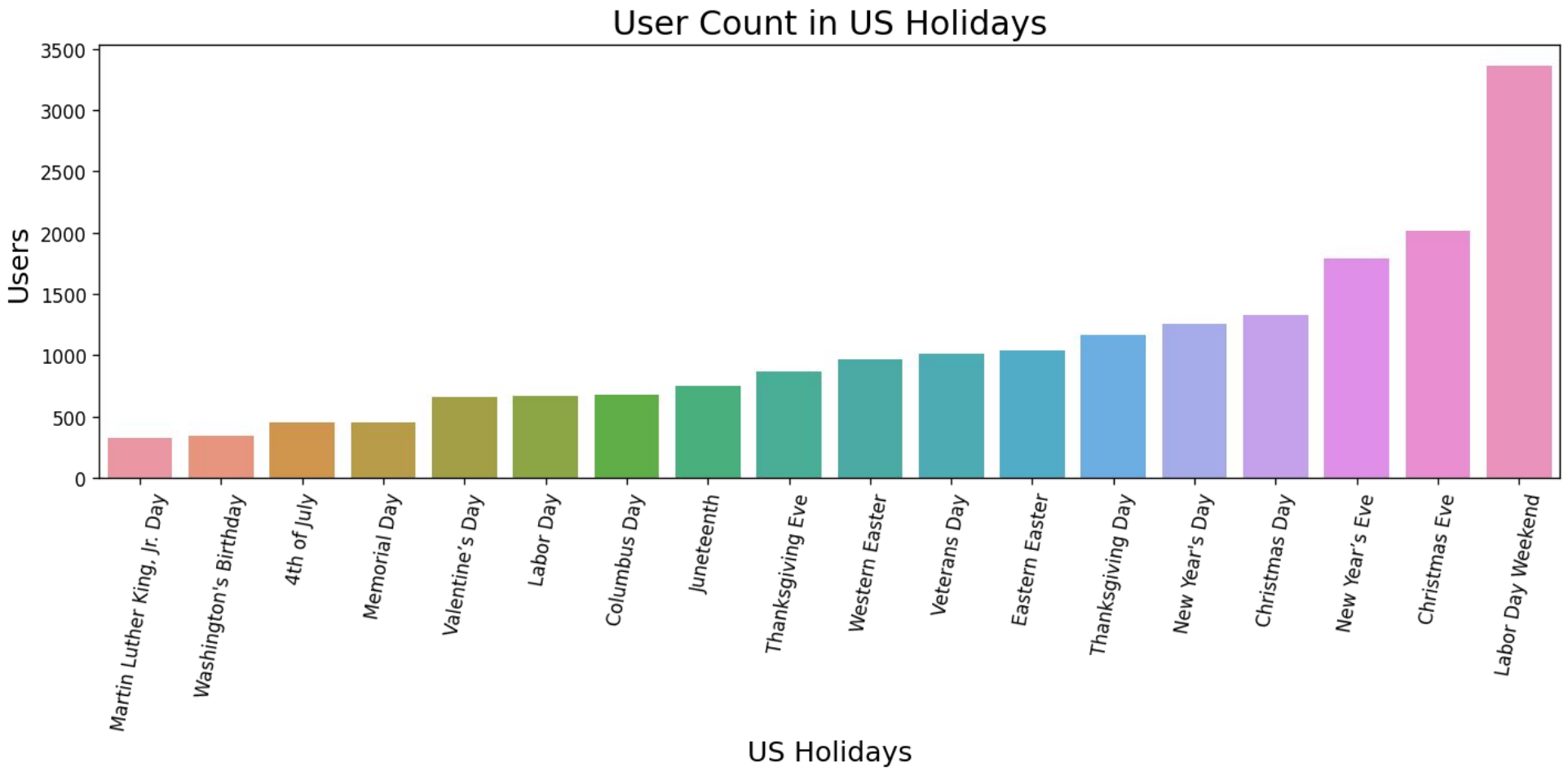
Profits per Ride



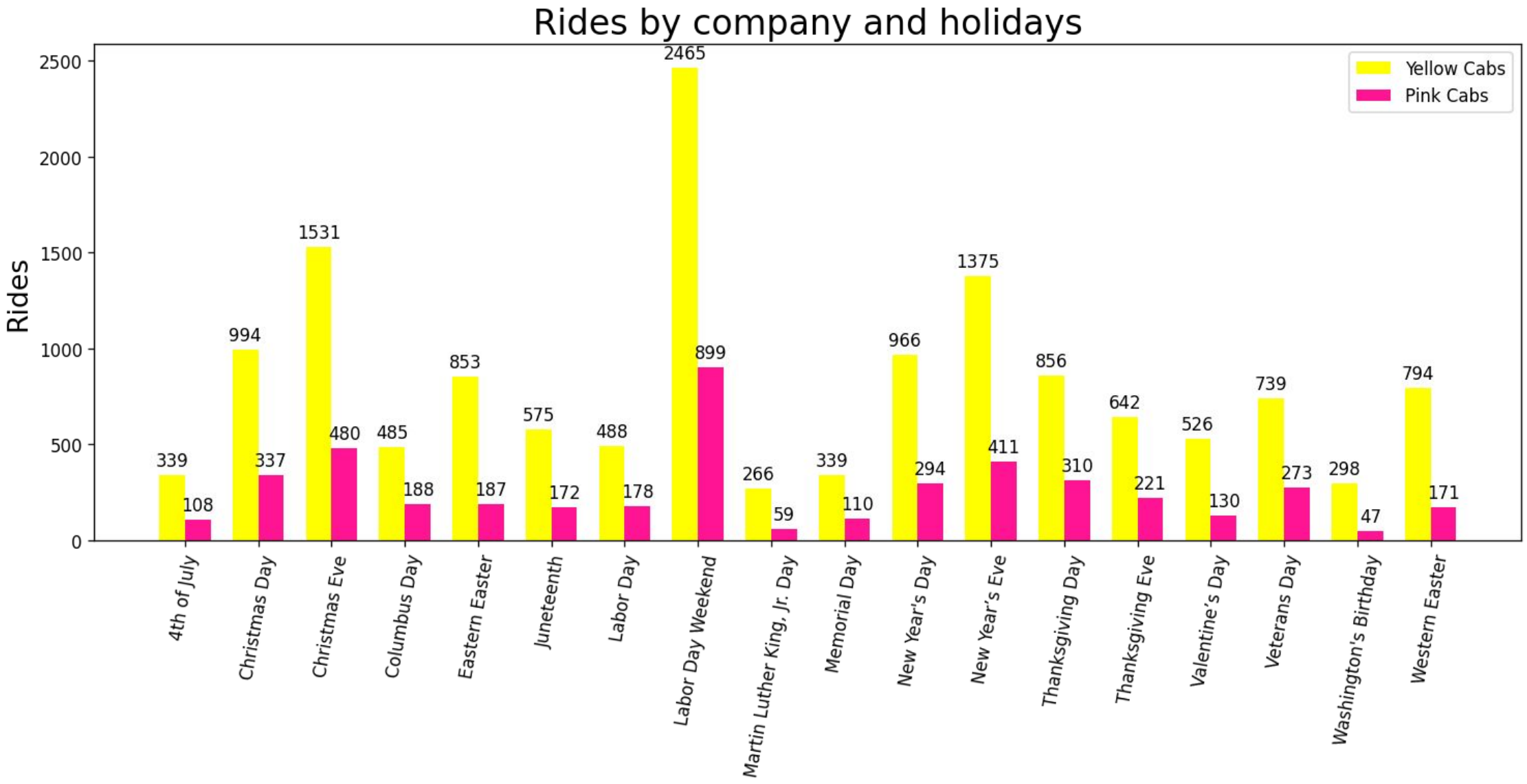
Average profit per Km that travelled for Yellow cabs: $ 7.36
Average profit per Km that travelled for Pink cabs: $ 2.71

Ratio between them Yellows/Pinks: $ 2.7

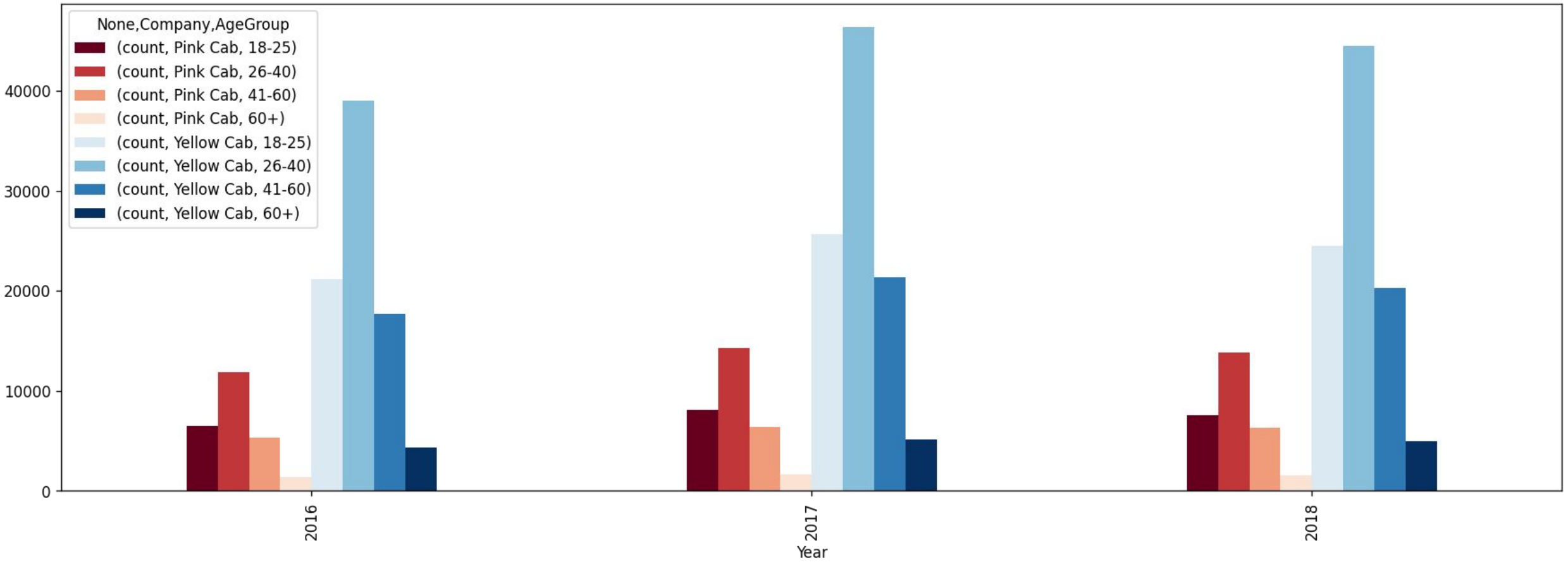
Rides in the US holidays



Rides in the US holidays



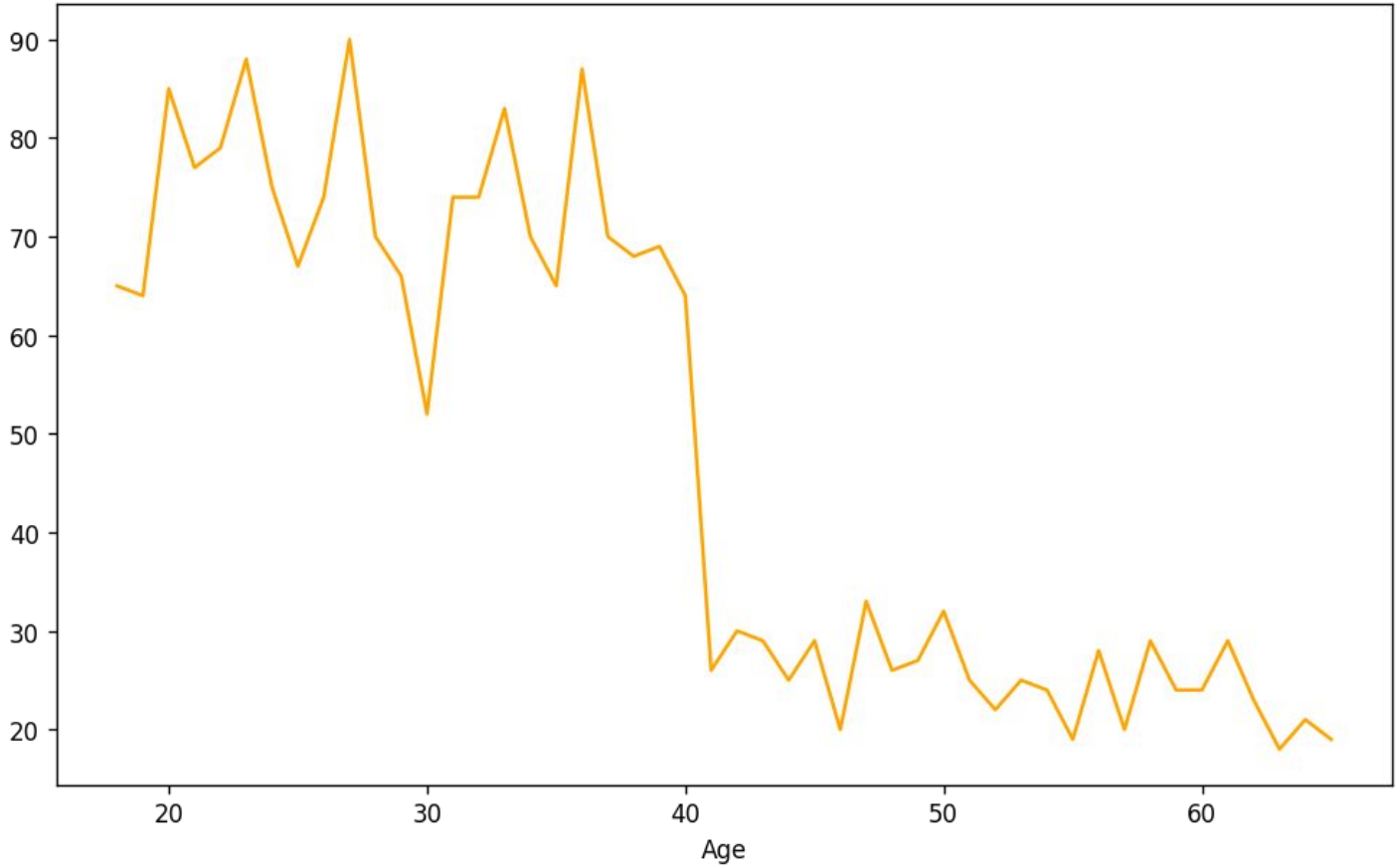
Customer Segmentation(Age)



-The most number of customers come from 26-40 age group for the two companies for every year

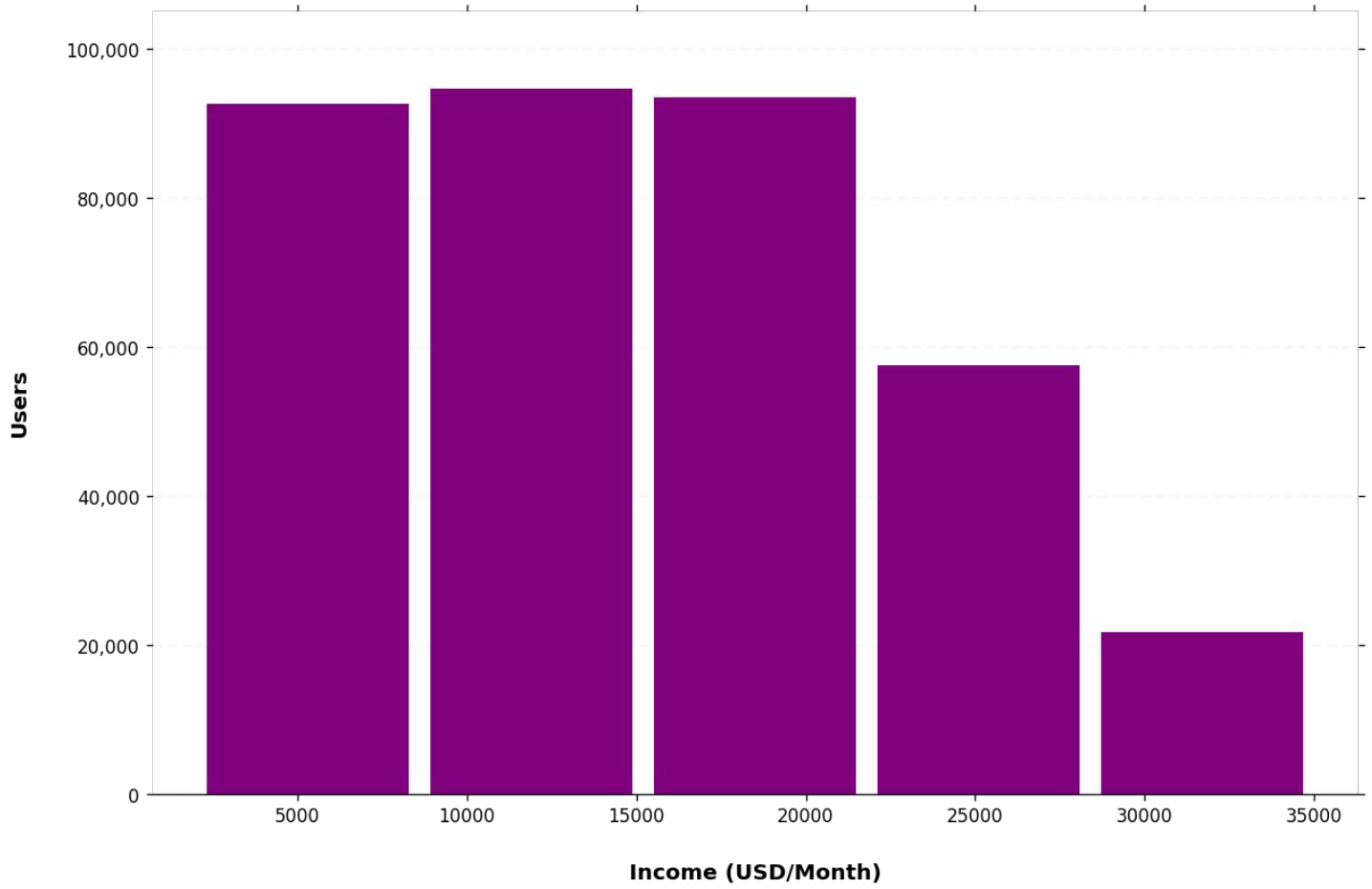
Customer Segmentation(Age)

#Customers amount that used cabs more than 30 times in given 3 years according to their age



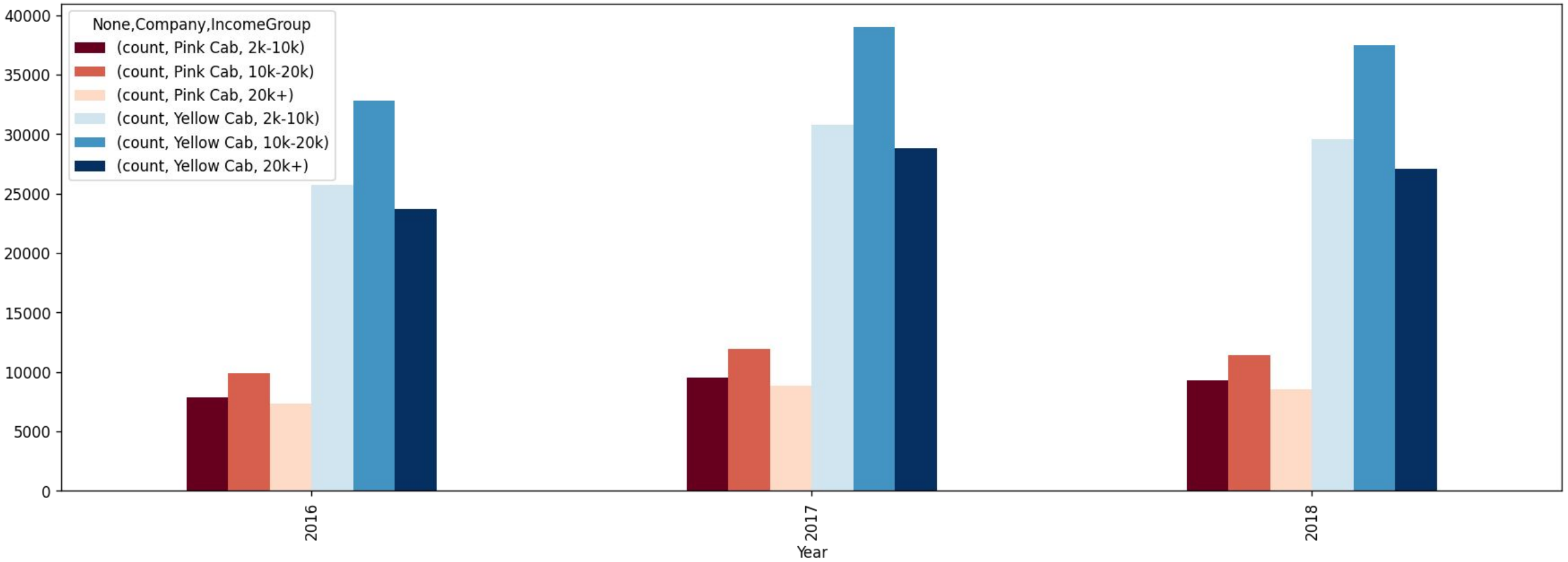
-Customer retention is significantly decreasing after 40 years old

Customer Segmentation(Income)



-Income distribution of customers

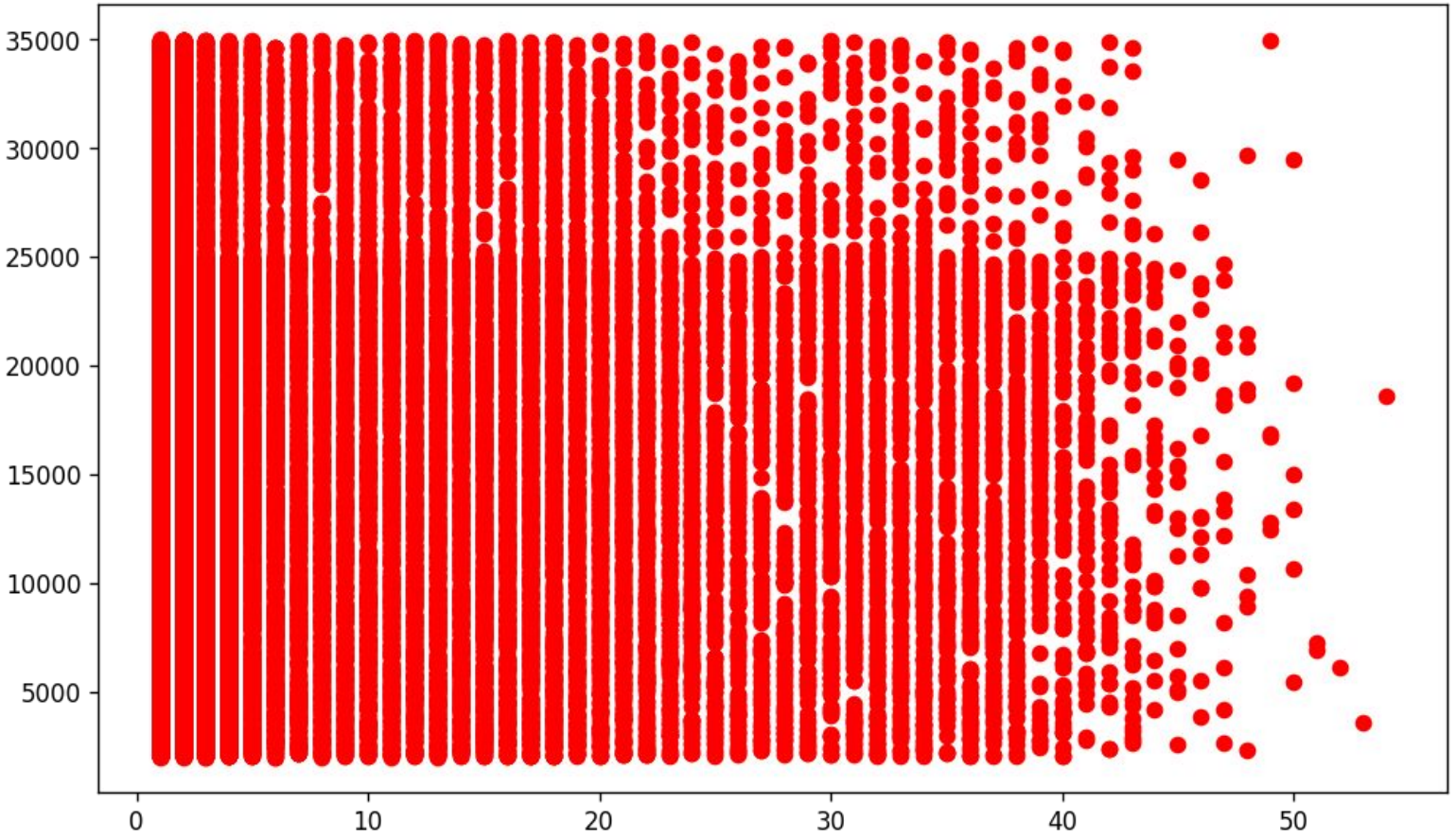
Customer Segmentation(Income)



-Total rides according to income groups of customers

Customer Segmentation(Income)

Income per month USD($)

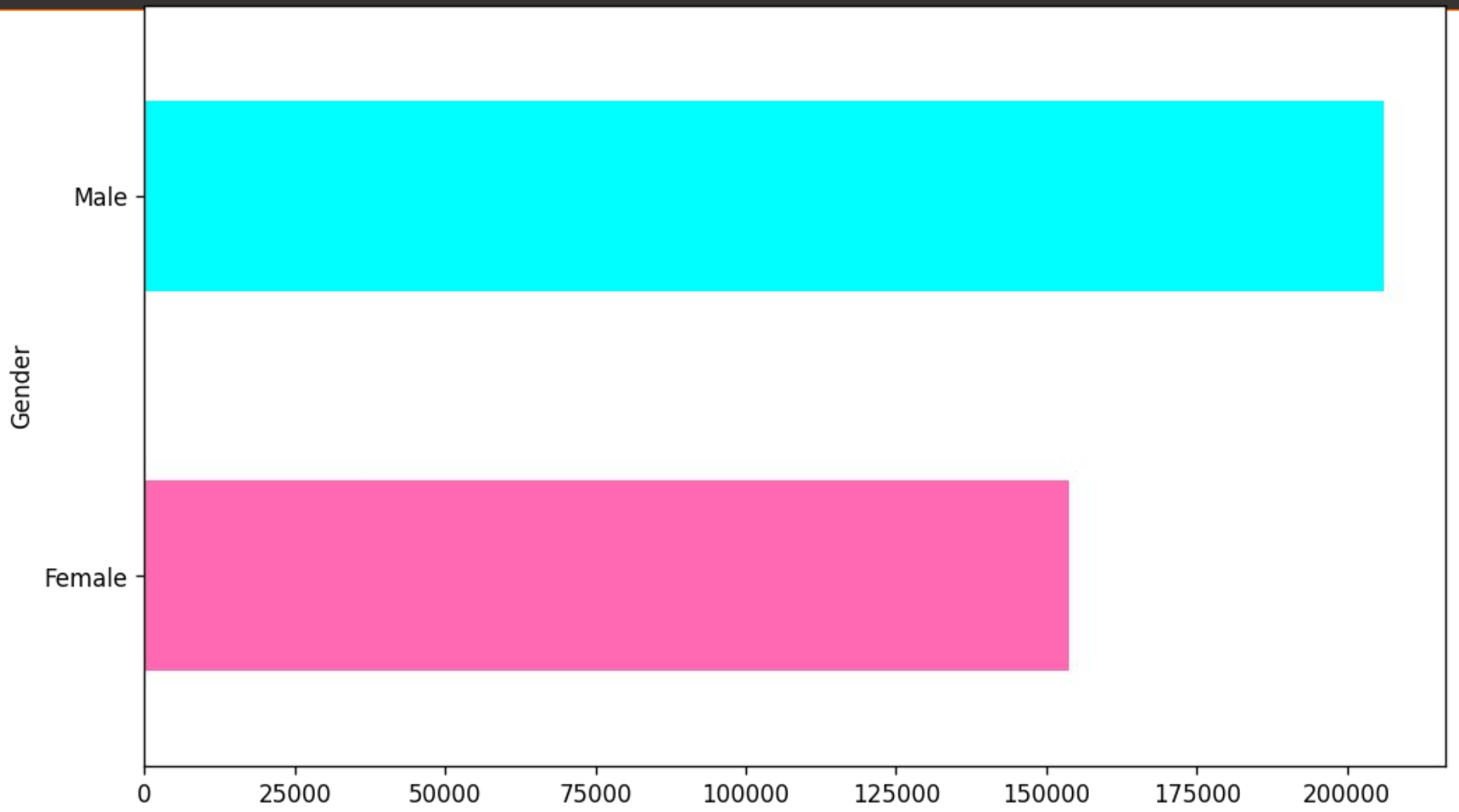


Total rides by customers in 3 years

-As we can see there is no correlation between income level and customer retention.

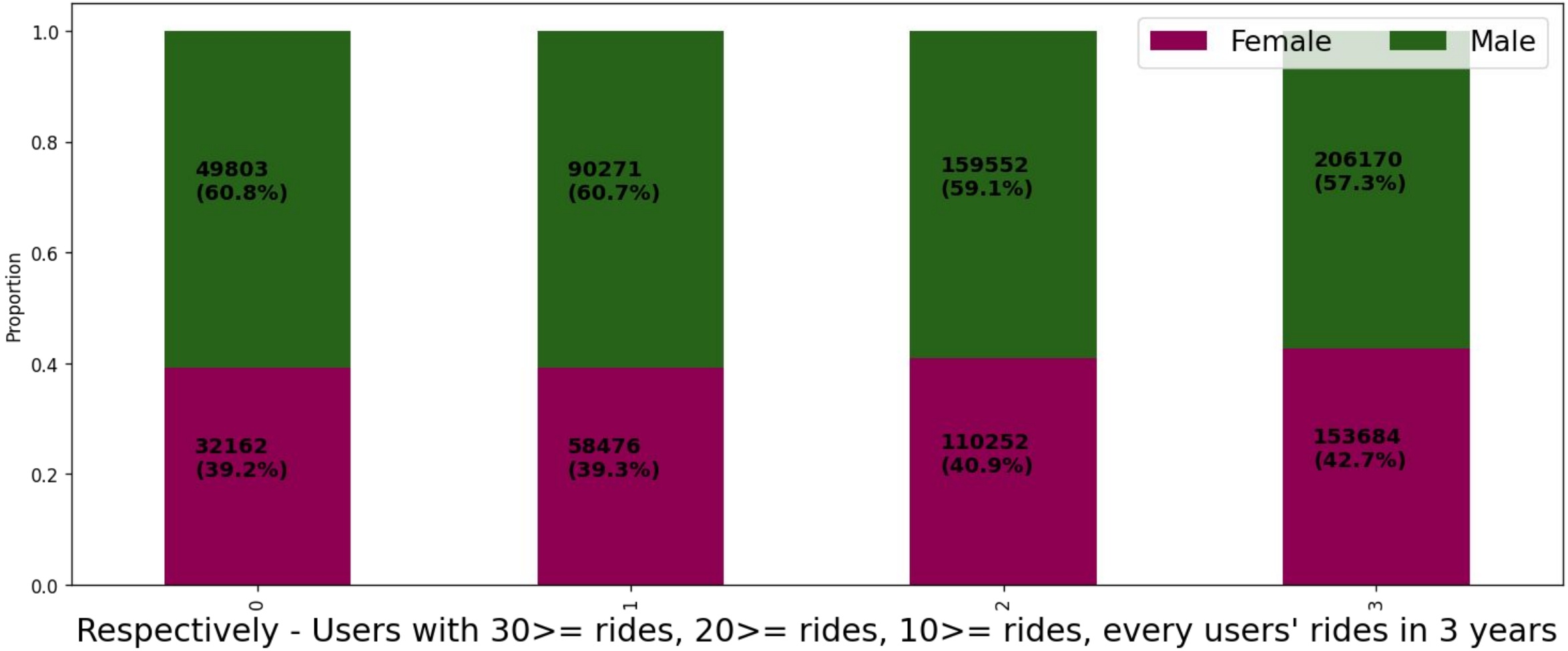
****Correlation coefficient 0.0053**

Retention of customers vs Gender



	Female	Male
Rides	153684	206170

Retention of customers vs Gender



**As we can see if we filter users by their retention to these two firms by their gender, there is almost no change in proportion of gender distribution

EDA summary and Recommendations

Yellow cabs have much more customers, much more profitable per ride, per distance(KM) and in total profits. Even both companies year 2019 profits can be foreseen or forecasted since there is seasonality. This one also suggesting that Yellow cabs are much superior, so there are ratios of profitability etc in analyses, if XYZ company wants to invest in one of these companies(Pink or Yellow), they should go for Yellow.

- **Extra recommendation for business**

-Customer retention is significantly decreasing after 40 years old **Slide 23

Cab companies can create special campaign for people that are older than 40 in this case to attract old people

Some hypothesis testings and results

Test	Result
Testing to see whether age of the customer affects the payment method(cash-card)	There is no difference in payment mode for both Yellow and Pink Cabs according to customer ages
Testing to see whether income of the customer affects the payment method(cash-card)	There is no difference in payment mode for both Yellow and Pink Cabs according to customer incomes

**T-test is used and p value(significance taken as 5%, 0.05)

Some hypothesis testings and results

Test	Result
Testing for Profits per distance(KM) and Age (Choosing age "40" as a threshold)	There is no difference regarding being younger/older than 40 years old for both Yellow Cab and Pink cabs
Testing for Profits per distance(KM) and Age (Choosing age "60" as a threshold)	There is no difference regarding being younger/older than 60 years old for Pink Cab There is significant difference regarding being younger/older than 60 years old for Yellow Cab

**T-test is used and p value(significance taken as 5%, 0.05)

Some hypothesis testings and results

We have seen, there is significant difference regarding being younger/older than 60 years old for Yellow Cabs, in previous slide

Average profit from per KM, from younger people(<= 60) : $ 7.121
Average profit from per KM, from younger people(<= 60) : $ 6.825

Difference between profits per KM that are traveled for the people that older or younger 60 years old: $ 0.295

Average KM traveled in one ride(All rides): 22.57 KM

Even it seems there is not significant profit difference between rides of older people when we look at the rides by their "per KM", if we multiply by average distance 22.5 KM with 0.3, almost 7$ profits are changing for yellow cabs.

Some hypothesis testings and results

Test

Testing to see if two companies are taking different distances on average by the cities

Result

There is no difference in distance taken between two companies by cities

 # Average distance taken by Pink Cabs: 22.44 KM

 # Average distance taken by Yellow Cabs: 22.57 KM

****T-test is used and p value(significance taken as 5%, 0.05)**

Some hypothesis testings and results

Test

Testing to see if two companies have different age levels of customers by cities

Result

There is no difference in ages between two companies by cities

 # Average age of Pink Cabs customers: 35.37 KM

 # Average age of Yellow Cabs customers: 35.31 KM

****T-test is used and p value(significance taken as 5%, 0.05)**

Thank You