
# Exploratory Data Analysis

## G2M Case Study

**Prepared By: Maviya Shaikh**

**Date: August 5th , 2021**

# Agenda

- Executive Summary
- Problem Statement
- Datasets
- Data Preparation
- Testing Hypothesis
- Insights
- Recommendations

# Executive Summary

The demand for cabs are rising exponentially and due to its remarkable growth in US , XYZ company is planning for an investment in Cab industry. There are multiple players in the market and as per their Go-to-Market(G2M) strategy they want to understand the market before taking final decision.

# Problem Statement

XYZ company has two options for investment as per the data collected, first is the Yellow cab company and second is the Pink cab company. Time period of data is from 31/01/2016 to 31/12/2018.

**“Exploratory Data Analysis is conducted on the given datasets to determine whether XYZ company should invest in Yellow or Pink cab company”**

# Datasets

There are 4 datasets used to conduct the analysis which are as follows:

| Datasets | Number of records | Number of Features | Missing Values | Duplicate Values |
|---|---|---|---|---|
| Cab data | 359392 | 7 | 0 | 0 |
| City data | 20 | 3 | 0 | 0 |
| Customer data | 49171 | 4 | 0 | 0 |
| Transaction data | 440098 | 3 | 0 | 0 |

**Assumptions:**

- The data come from a reliable source and is accurate
- The outliers are not treated or removed from the dataset as all the data is considered to be the real values and there are not major outliers in this datasets that will deviate the analysis of the final result

# Data Preparation

Master dataset is prepared to conduct the further analysis in the following steps:

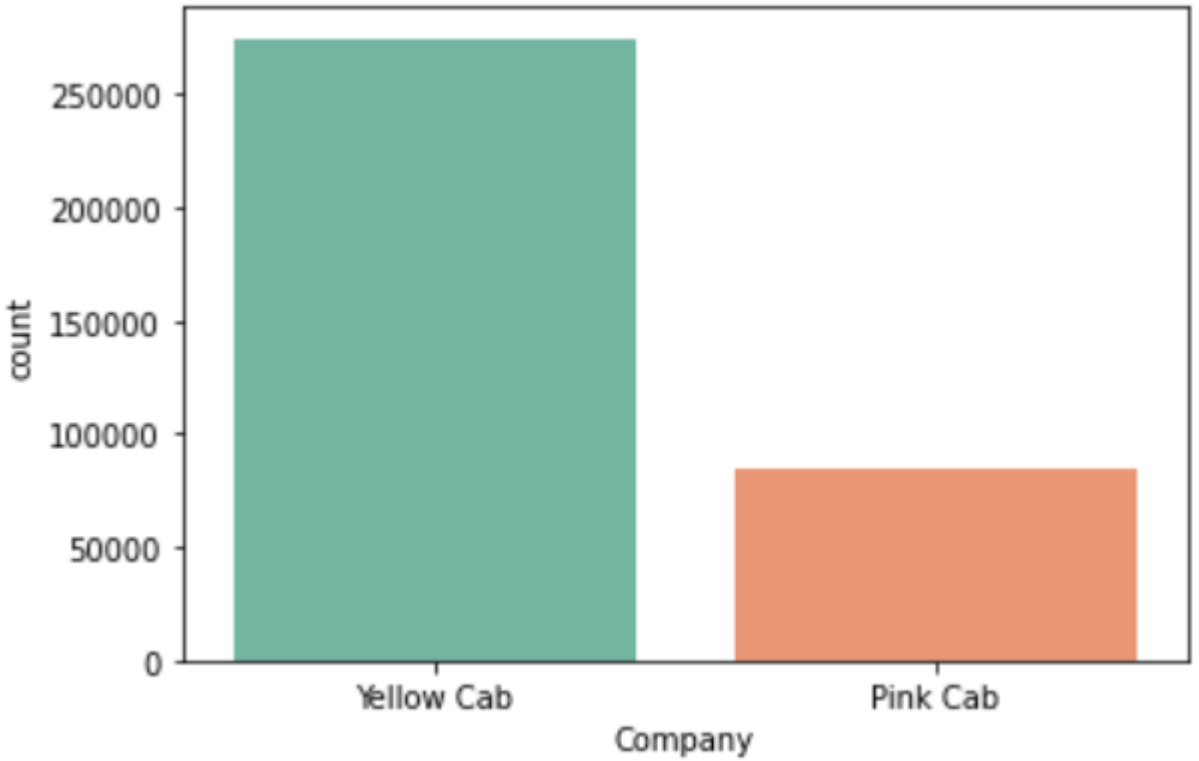
- Three datasets cab, customer and transaction are merged to built the master dataset, the city dataset is not merged because we get the city data from the cab dataset and we don’t need detail geographics in our analysis

- Feature Engineering is done because price charged and the cost of the trip will not tell us much, so two new features are created profit earned from that trip and the profit earned by kilometre to better analyse yellow and pink cab

- In total, we have 14 features to conduct the analysis of this case study. They are Transaction ID, Customer ID, Date of Travel, Company, City, Age, Gender, Income, KM Travelled, Price Charged, Cost of Trip, Payment Mode, Profit and Profit/km

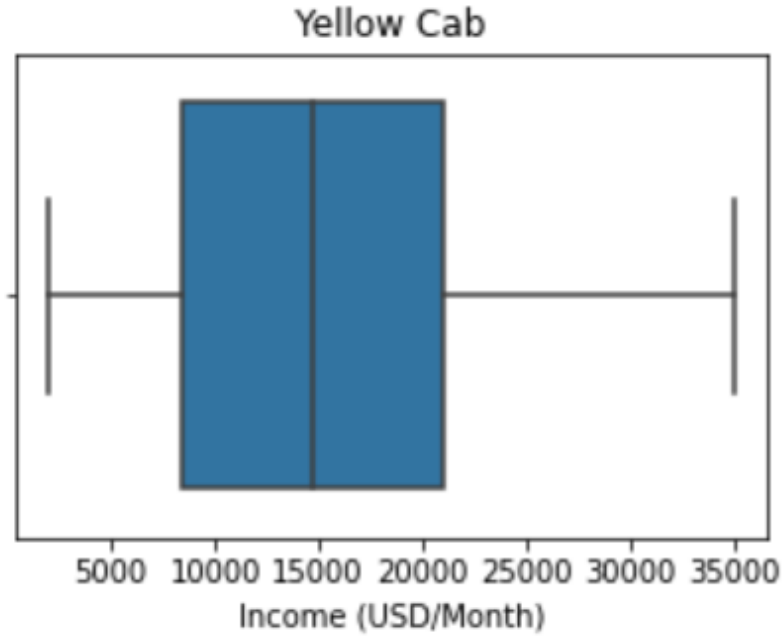
# Which company has more rides in the same time period?

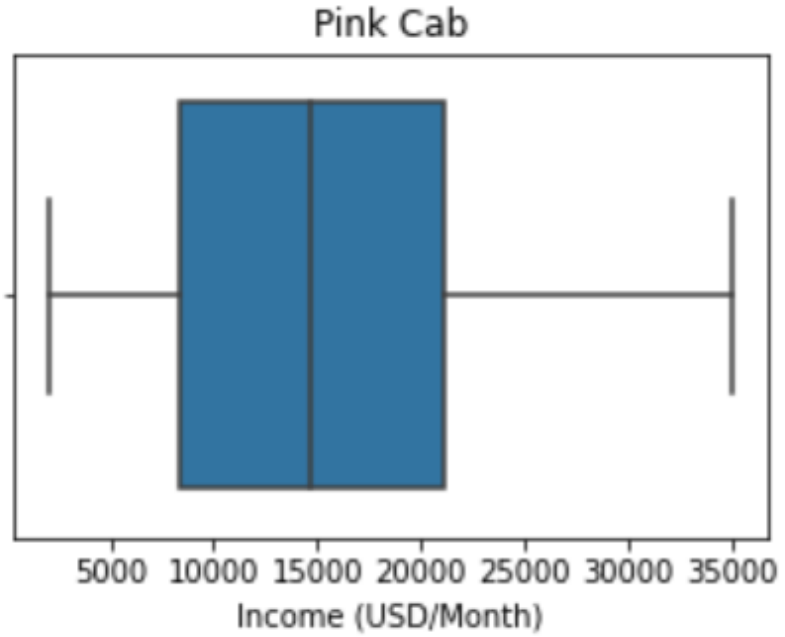
- There are 274681 data points for yellow cab and 844711 data points for pink cab
- In terms of percentages, Yellow cab company has 76% rides whereas Pink cab company has 24% rides from 31/01/2016 to 31/12/2018
- Thus, yellow cab company has 52% more rides than pink cab company



# Does the rides depend on the income of the people?

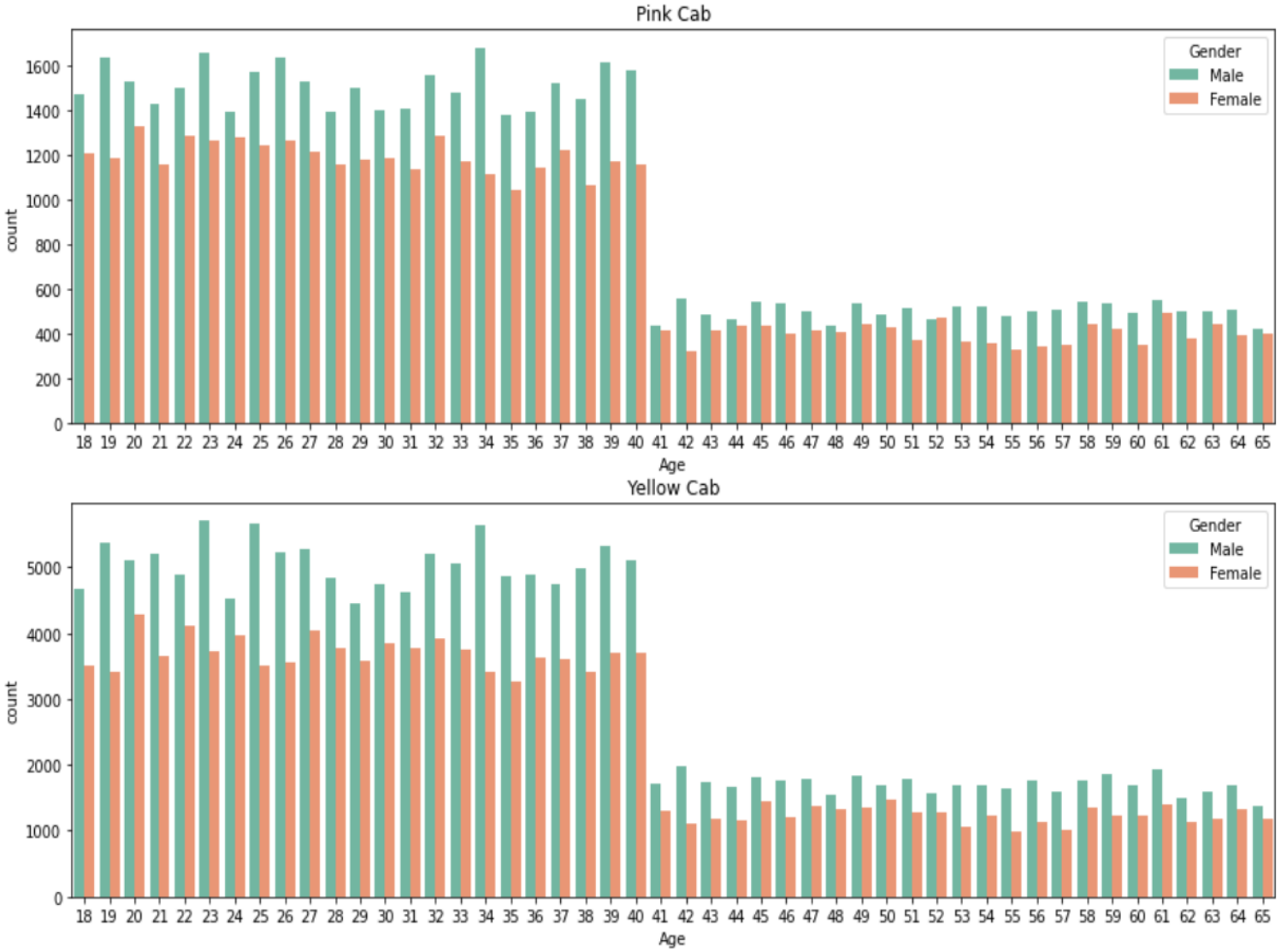
- For the income variable, the data points are normally distributed and there are no outliers.
- Most of the rides done by a people have an income in range of $10,000 to $22,000.
- The people with less than $10,0000 and more than $25,000 takes fewer rides
- So, the income of people does not matter while selecting the cab company for investment because the income distribution is similar for both the users of the cab company



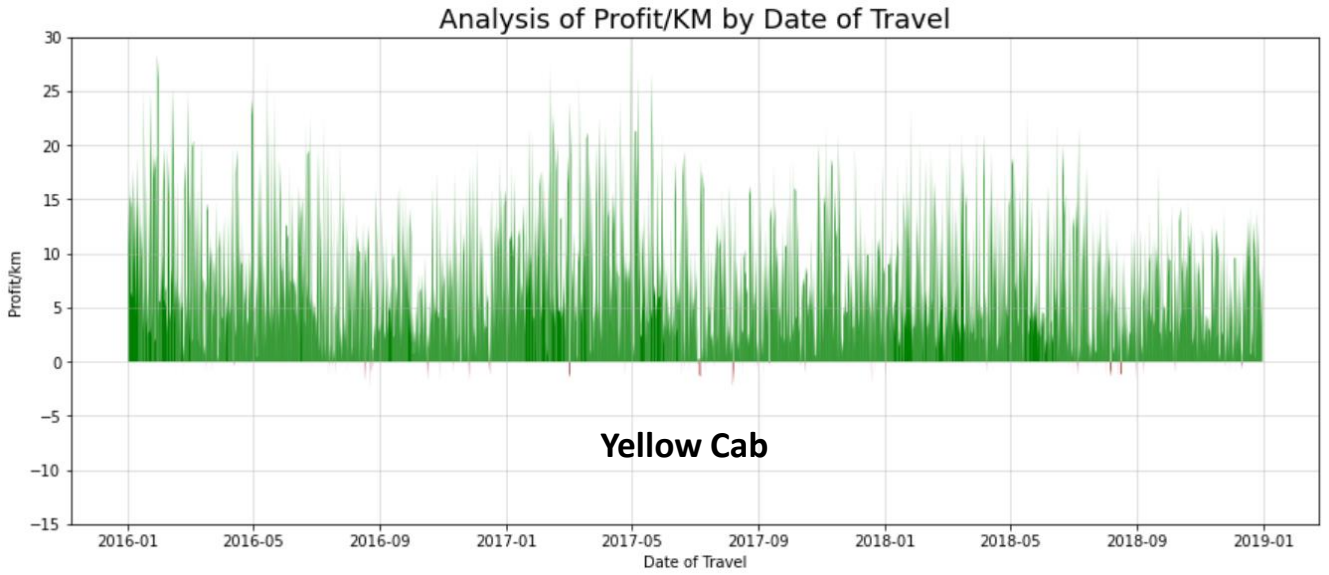
# Does Age and gender effect the company's customer base?

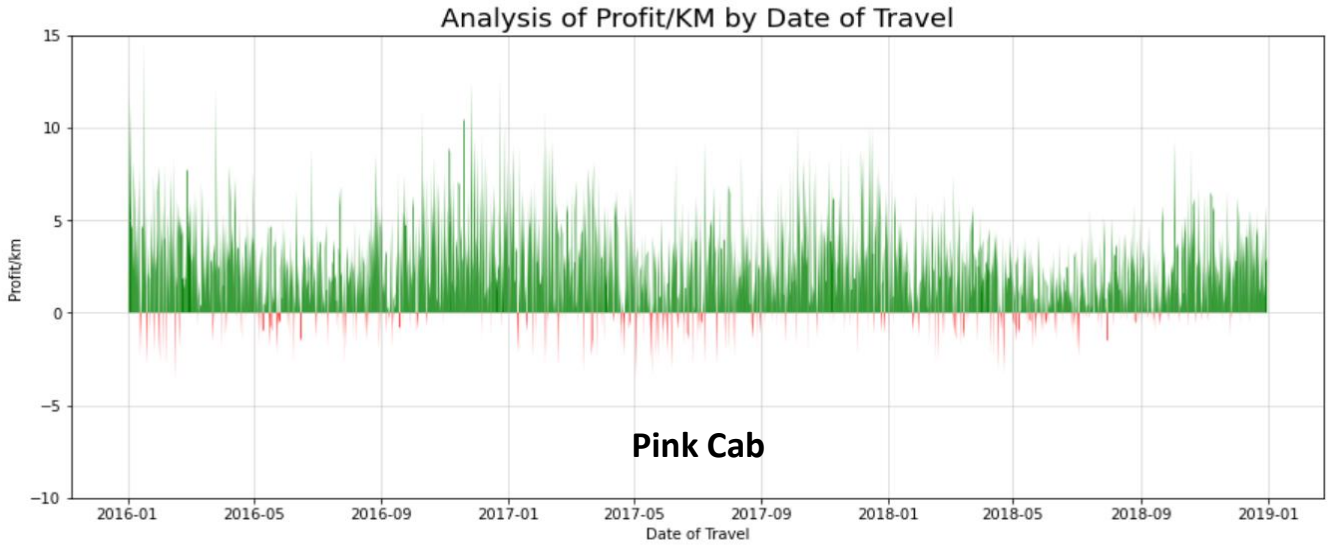
- The distribution of the  number of user based on age and gender almost looks similar, only the yellow cab has more users.
- Male users are comparatively more than female users
- The people in the age group of 18 to 40 tend to use more cabs then people in the age group of 41 to 65
- The age and gender of the person does not really make the difference while selecting the cab company for rides



# Analysis of both cab companies by profit per km

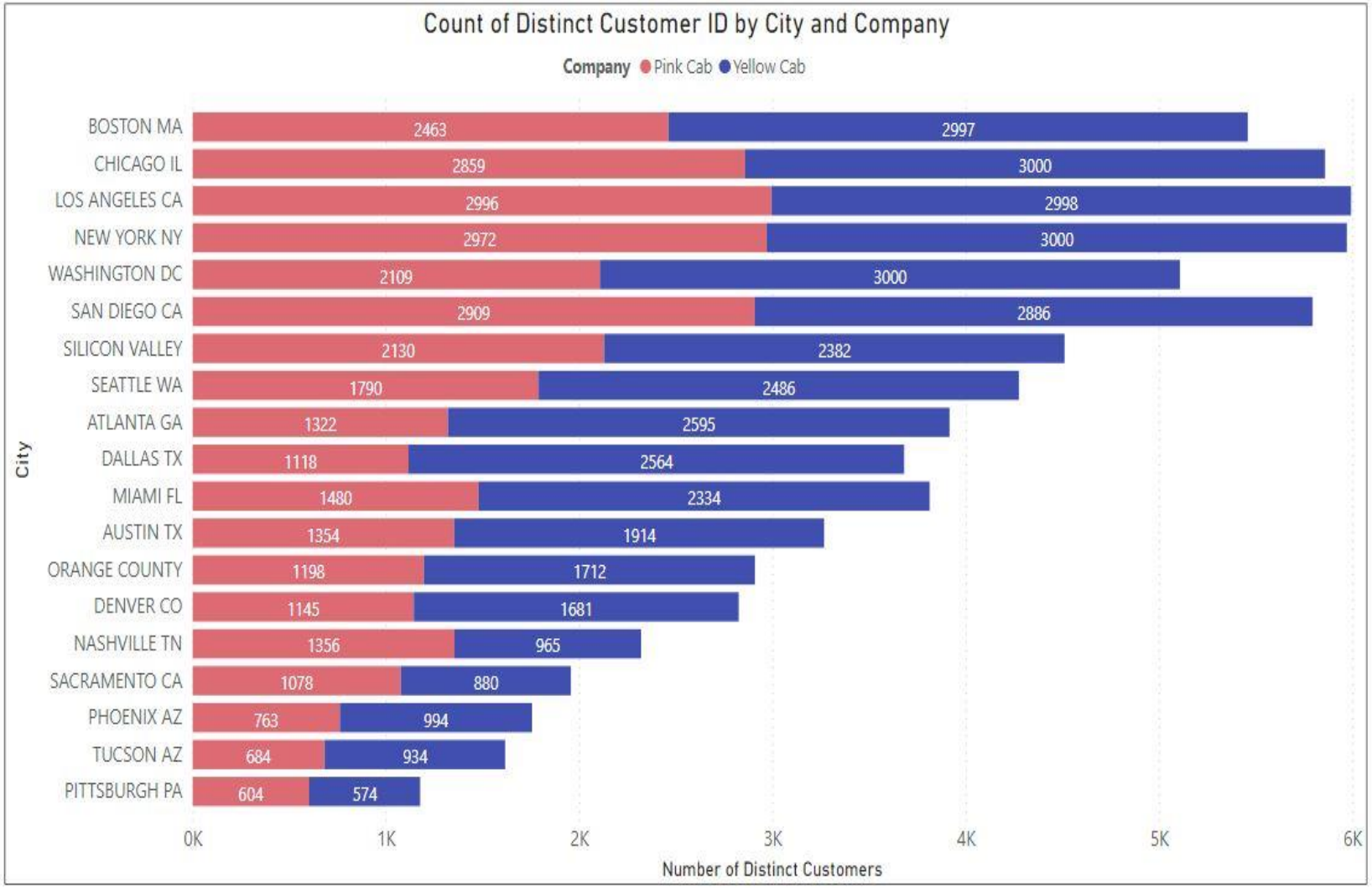
- Profit per km will give us broader view of understanding how each company is performing in earning profit in each ride throughout the time period.
- The green lines in the graph indicates there is a profit whereas red lines indicates there is a loss
- Pink Cab has relatively more red lines that means it has face many losses compare to yellow cab
- Clearly, yellow cab company is earning more profit per km



# Which company has larger customer base based on cities?

- The graph shows the unique number of users in different cities of US.
- The pink cab company has comparatively less distinct users than yellow cab company except 4 cities Pittsburgh, Sacramento, Nashville and San Diego.
- Overall, yellow cab company has more unique users in most of the cities of US

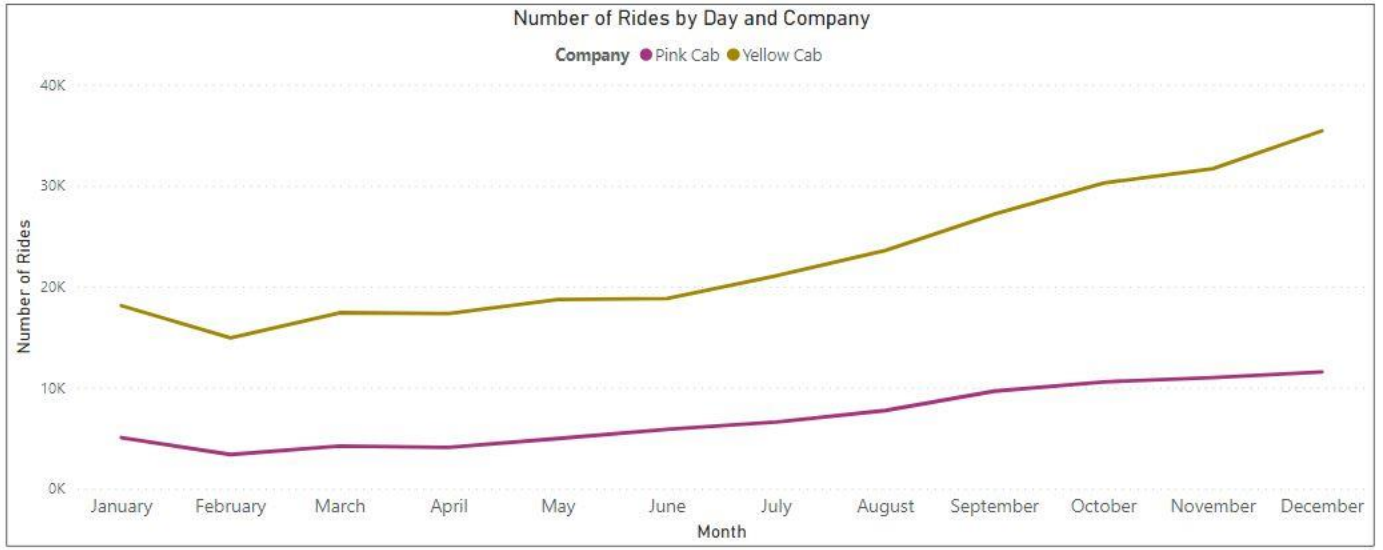


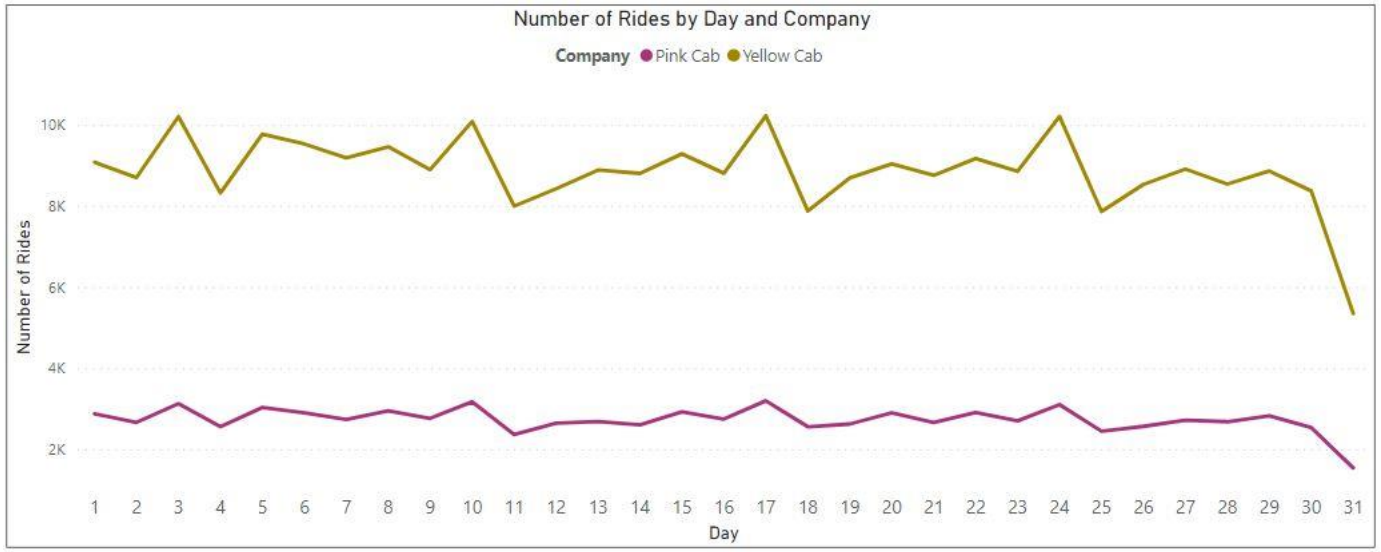
# Analysis of number of rides by month and day

- The first line graph shows the frequency of rides by month and second by day of the month

- We can see from the first graph that the rides for yellow cab company is increasing rapidly towards the end of the year whereas for pink cab company, it is increasing slowly. December is the month which has the highest user, this is probably because of Christmas holidays.

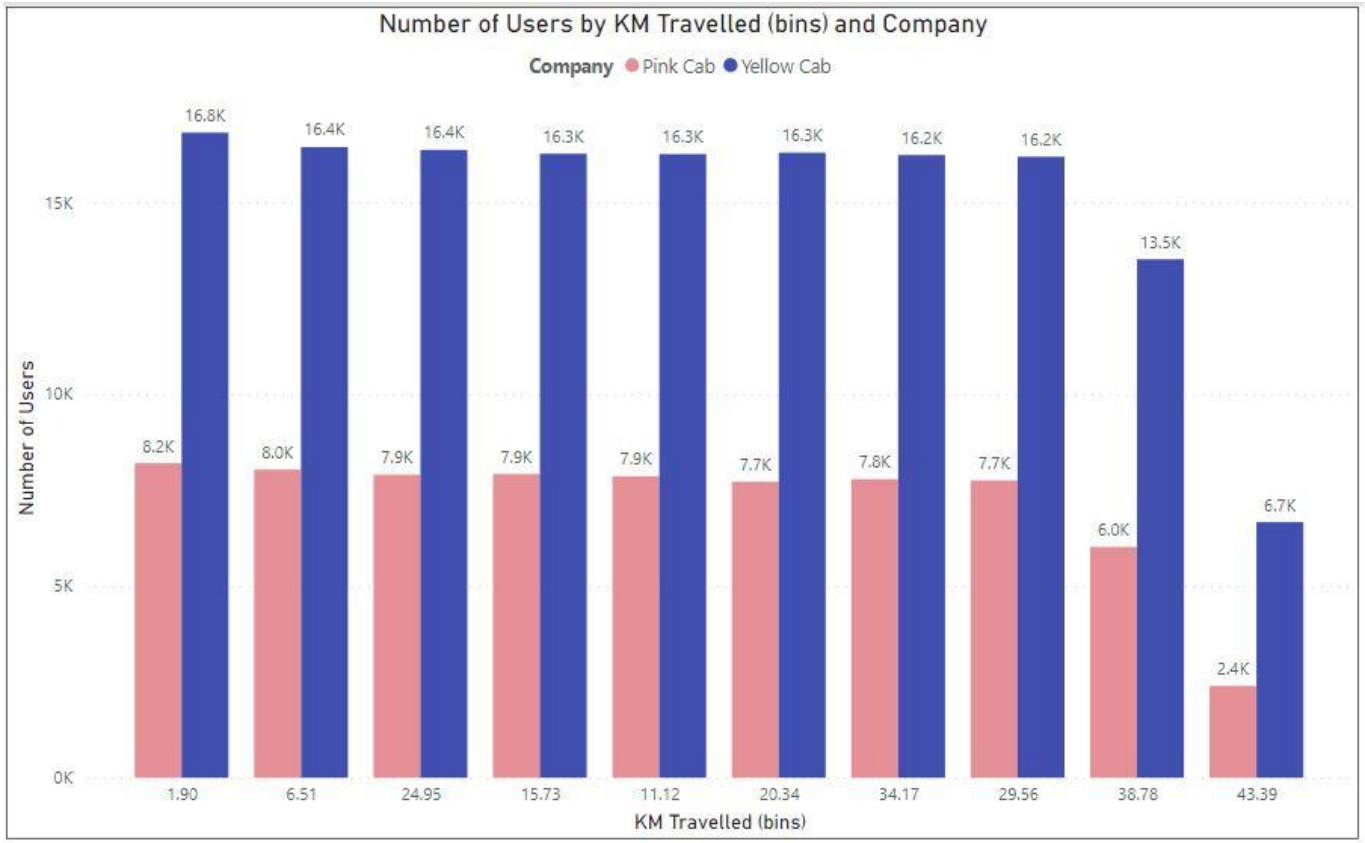
- Both the companies had almost similar kind of pattern for second graph. 3rd, 10th, 17th and 24th day of the month have the more number of rides compare to other days.



# Analysis of KM Travelled by number of users for both companies

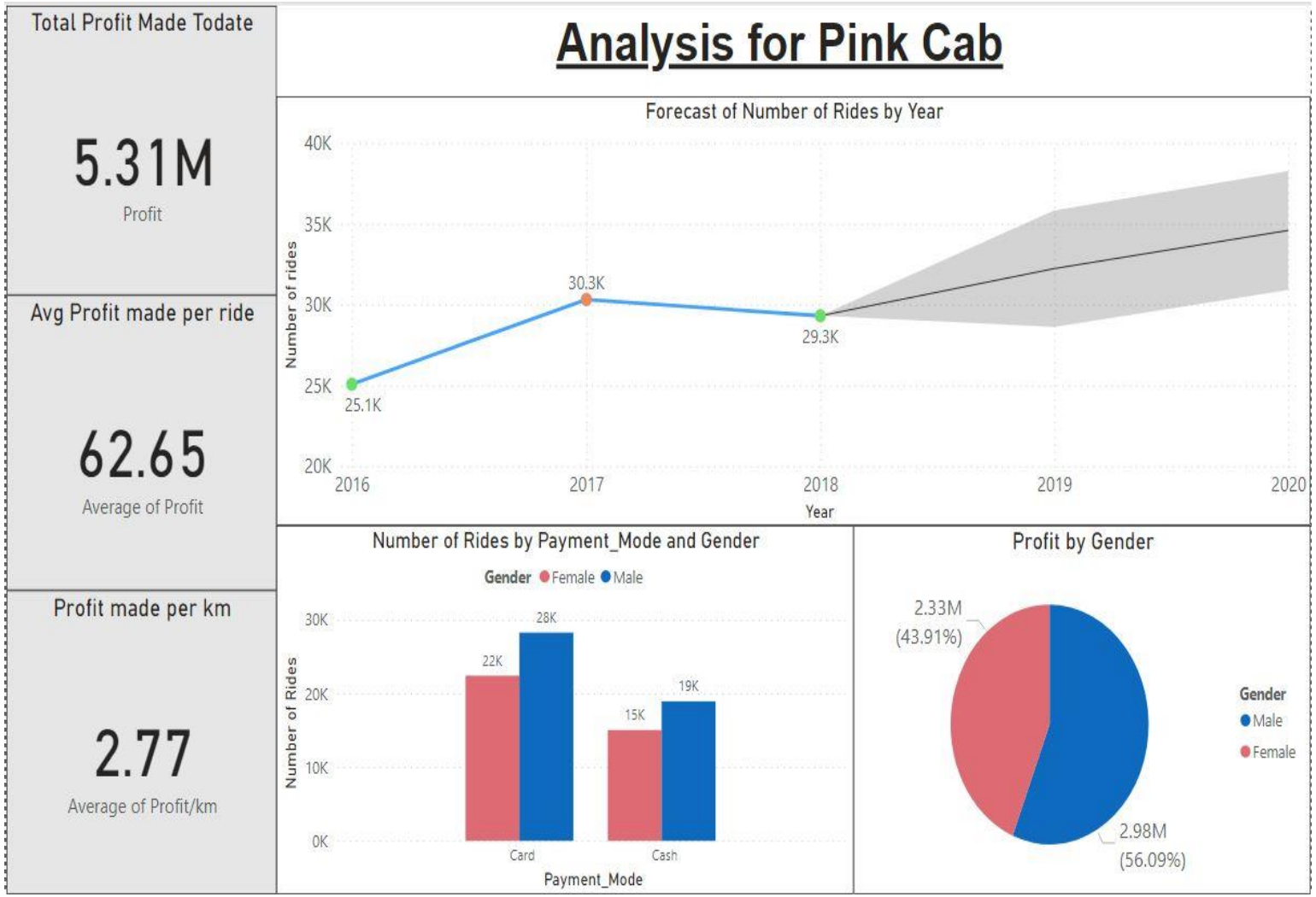
- The distribution for Kilometre travelled by users is almost similar for pink and yellow cab
- From the graph, we can see that the highest number of users preferred short rides up to 5 to 6 kms.
- Moreover, there are approximately same number of users for 6 to 25 km. There are slight decrease in number of users with 30 to 40 km and sudden decrease in number of users for more than 40kms.



# Dashboard for Pink Cab

Below are the insights for Pink cab company:

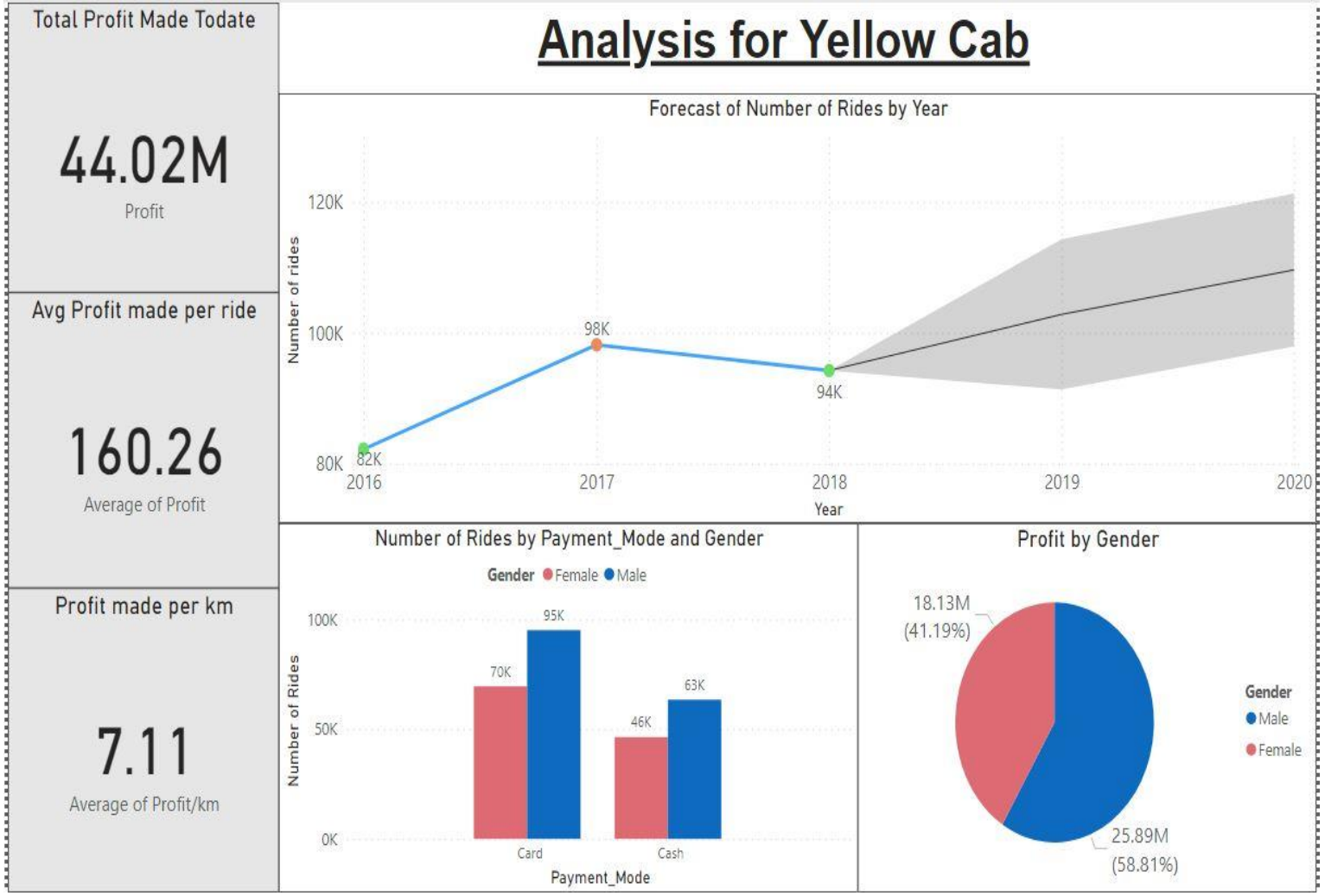
- They have made $ 5.31 million profit between the time frame given
- Average profit they made per ride is $62.65
- On average per km they make $2.22 profit
- The number of rides forecasted for the year 2019 and 2020 is roughly around 28k to 38k
- More people prefer to pay through card then cash
- 44% of profit comes from female users while 56% of profit comes from male users



# Dashboard for Yellow Cab

Below are the insights for Pink cab company:

- They have made $44.02 million profit between the time frame given
- Average profit they made per ride is $160.26
- On average per km they make $7.11 profit
- The number of rides forecasted for the year 2019 and 2020 is roughly around 90k to 120k
- More people prefer to pay through card then cash
- 41% of profit comes from female users while 59% of profit comes from male users



# Insights

Based on the following hypothesis, the performance of yellow and pink company is evaluated:

- **Number of Rides** – Yellow cab company has 52% more rides compare to Pink cab company
- **Income, Age and Gender** - Both the companies shows approximately same distribution for this features. So, it is not helping in selecting the best company to invest
- **Profit or loss by Profit/km** – Pink Cab company had incur more losses by km compare to Yellow cab company
- **Customer Base by cities** – Yellow cab companies has higher customer base than pink cab company in most of the cities of US
- **Number of Rides by month and day** – Both the companies has similar pattern for this feature only that yellow can company has higher number of users compare to pink cab company
- **KM Travelled** – The distribution of users with respect to km travelled is same but yellow cab company has almost double users than pink cab company
- **Payment method** – More people preferred to pay with the card rather than cash for both the companies
- **Average profit per rides and average profit per km**– Yellow cab company has approximately 2.6 times more profit per ride and per km than yellow cab company
- **Forecasting of rides** - Yellow cab company has three times higher forecasted rides than pink cab company

# Recommendation

- I would recommend XYZ company to **invest in yellow cab company** based on their larger customer base, more profits and less loss per rides. Also, average profit per km and forecasting of rides for next two years for yellow cab company is approximately three times higher than pink cab company.

# Thank You