Data Glacier

Your Deep Learning Partner

Exploratory Data Analysis

G2M insight for Cab investment firm

27 June 2021

Problem statement/Case Study

- **The Client 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 and as per their Go-to-Market(G2M) strategy they want to understand the market before taking final decision.**

- **Objective:**Provide actionable insights to help XYZ firm in identifying the right company for making investment.

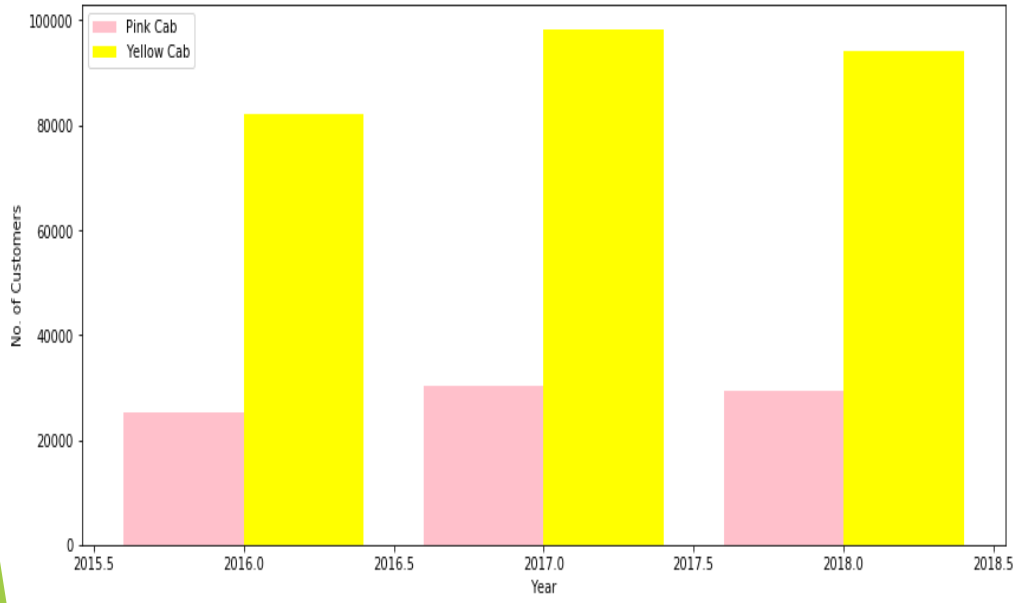
- **Cab Companies:**
- **Yellow Cab**
- **Pink Cab**
- **The analysis is divided into the following sections:**
- **Data understanding**
- **Exploratory data analysis**
- **Hypothesis testing**
- **Recommendations**

Data Understanding

- There are 4 datasets :
- **Cab_Data.csv-**The dataset contains **359392 observations/rows and 7 fields/columns**. This dataset contains transaction details for each cab type.
- **Customer_ID.csv-**The dataset contains **49171 rows/observations and 4 fields/columns**.This dataset contains demographic details of each customer.The column **Customer ID** is the unique identifier or sometimes called Primary Key for this dataset.
- **Transaction_ID.csv-**The dataset contains **440098 rows/observations and 3 fields/columns**.This dataset maps with the **Customer_ID.csv** dataset on the **Customer ID** field/column. **Column ID** is a Foreign Key to the **Customer_ID.csv** dataset and the **Transaction ID** column is a Primary Key.
- **City.csv-**The dataset contains **20 rows/observations and 3 fields/columns**. Its contains a list of cities, the population of the cities and the number of cab users in U.S.

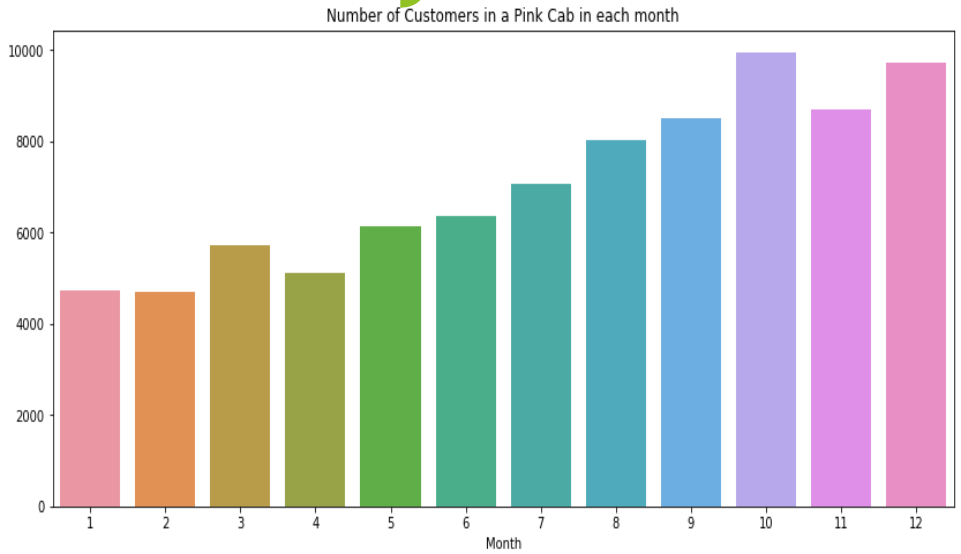
Exploratory Data Analysis

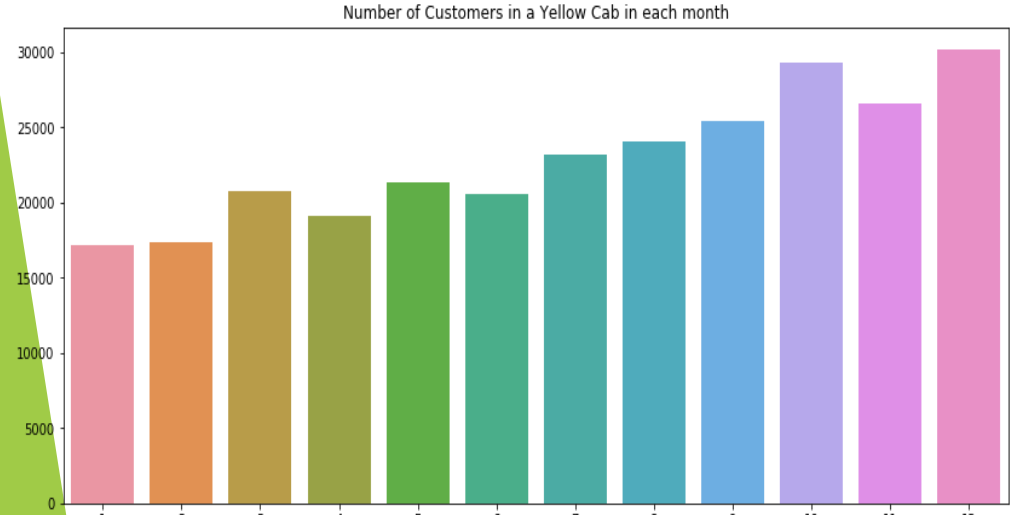
Analysis of a number of cab users on a yearly basis



- Higher number of cab users travelling in a Yellow cab.
- There is a very high range in between the number of cab users in a Yellow cab and a Pink cab.
- Pink cab ranges from 25080 to 30321 cab users from the year 2016 to the year 2018.
- Yellow cab ranges from 82239 to 98189 cab users.

Analysis of a number of cab users in a monthly basis

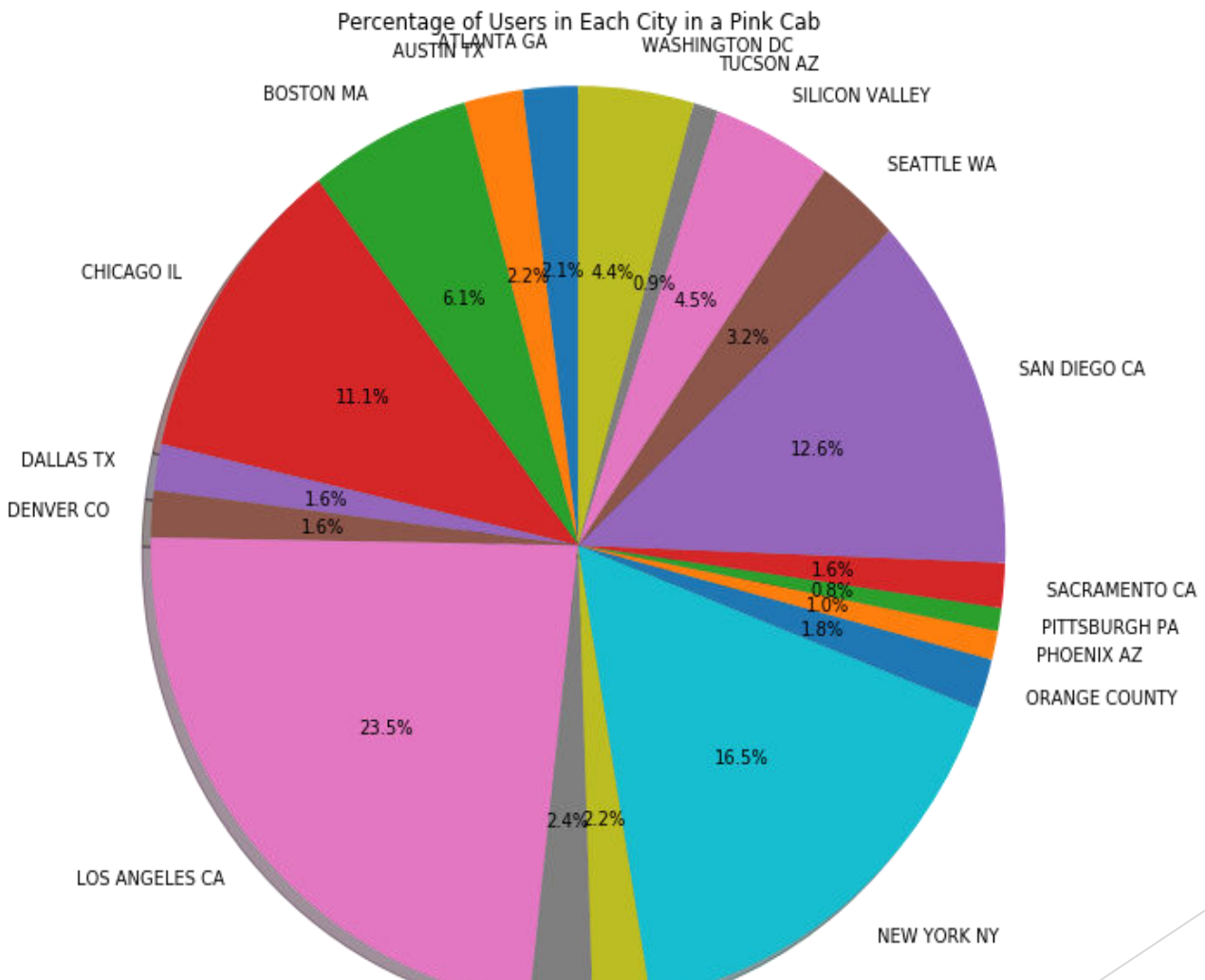




- From the graph ,we see the Yellow cab has a higher number of cab users.
- The range in a number of cab users for the Yellow cab is higher than that of the Pink cab.
- Yellow cab ranges from 17108 to 30135 cab users.

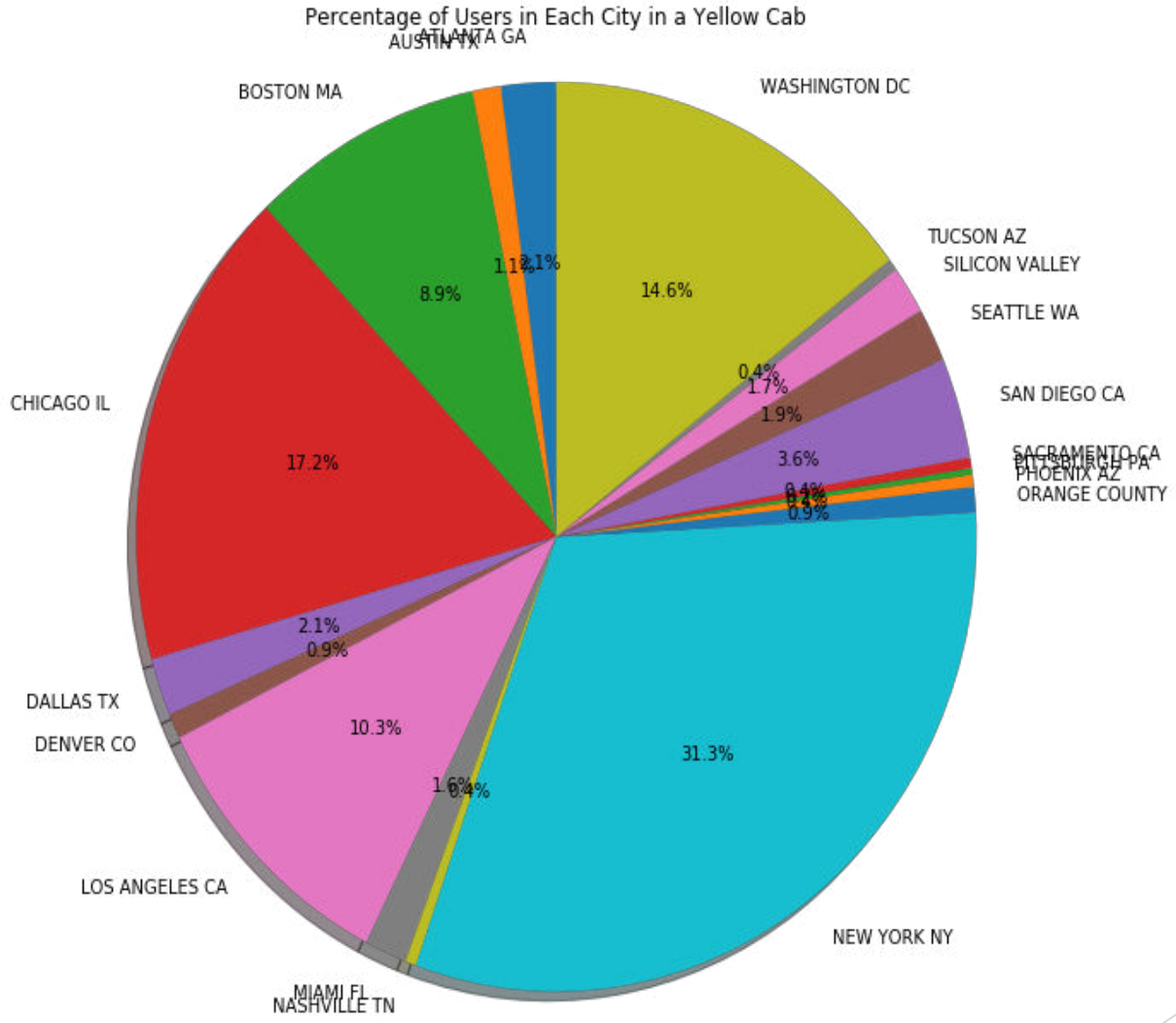
Pink cab ranges from 4734 to 9729 cab users.

Percentage of cab users in each city travelling in Pink cab



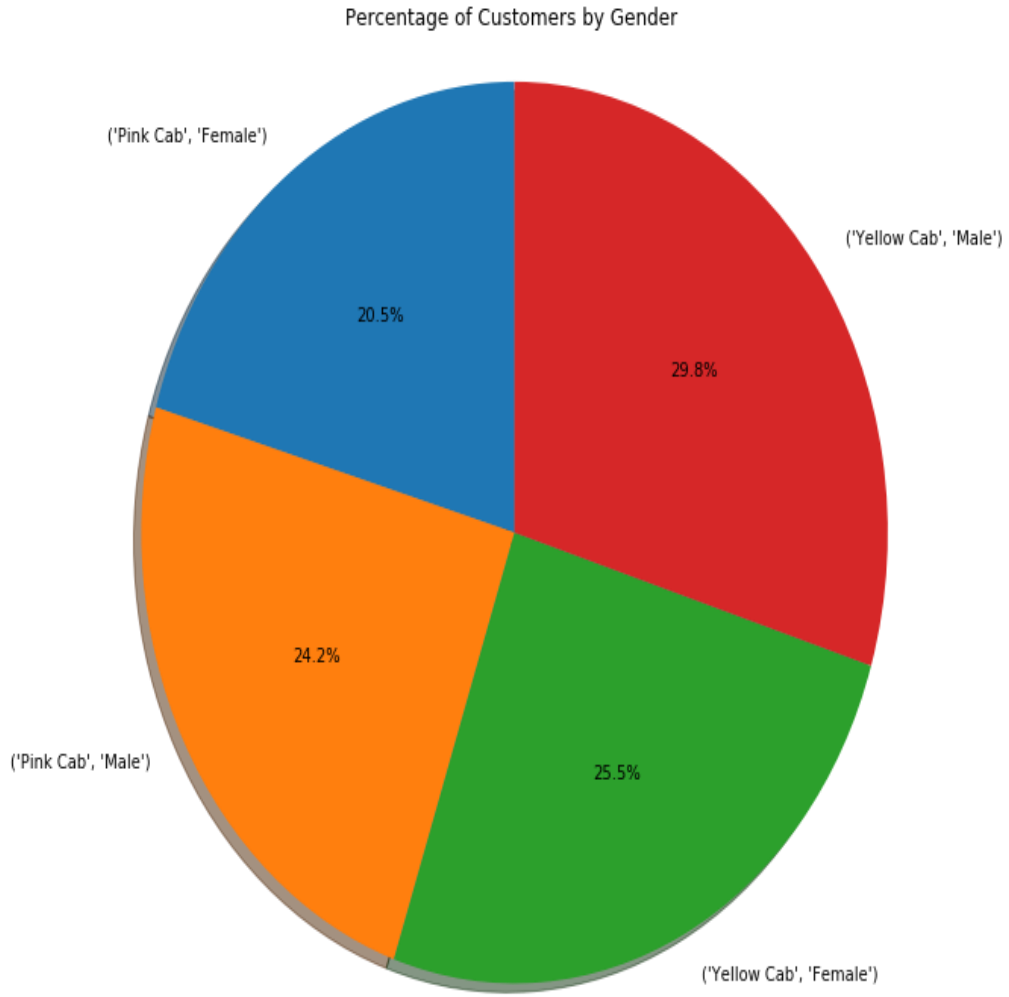
- In the LOS ANGELES city the highest percentage of cab users are travelling in a Pink cab with 23.5%.

Percentage of cab users in each city travelling in a Yellow cab



- In the New York city the highest percentage of cab users are travelling In a Yellow cab with 31.3%.

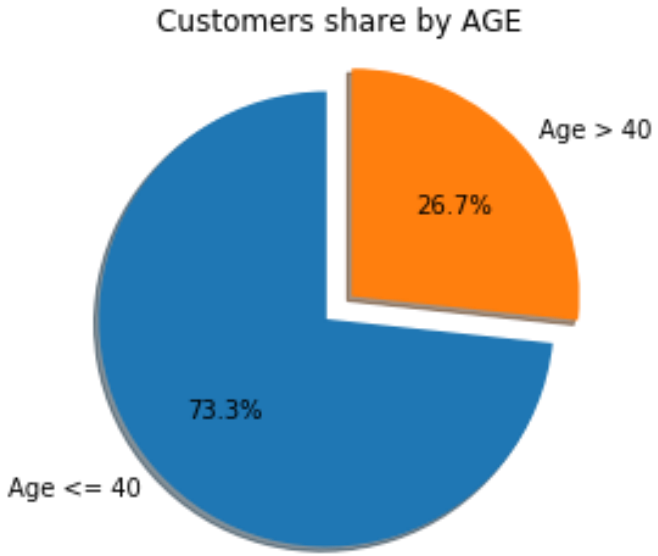
Gender Analysis



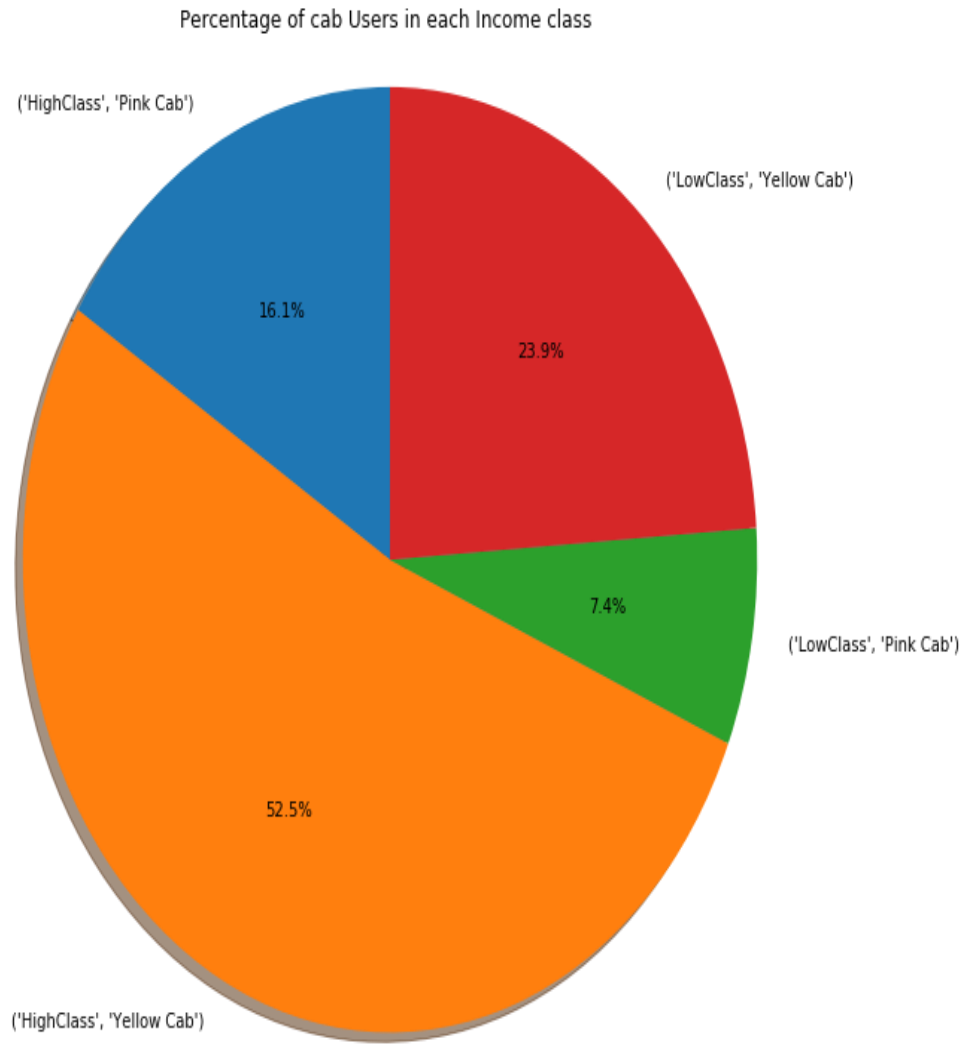
- Gender is one of the characteristics in determining customer behavior.
- Decision needs to be taken on what gender groups preference.
- Therefore, insights are required to reveal this.
- As shown on a pie chart illustrated the following is revealed:
  - In the Pink cab there are 20.5% females and 24.2% males.
  - In the Yellow cab there are 29.8% males
  - and 25.5% females.
- There is a higher percentage of males travelling in both Yellow and Pink cab.
- This means the male gender group is dominant.

Age Analysis

- We can see that there is 73.3% customers of age less than 40 years and 26.7% customers of age more than 40 years.



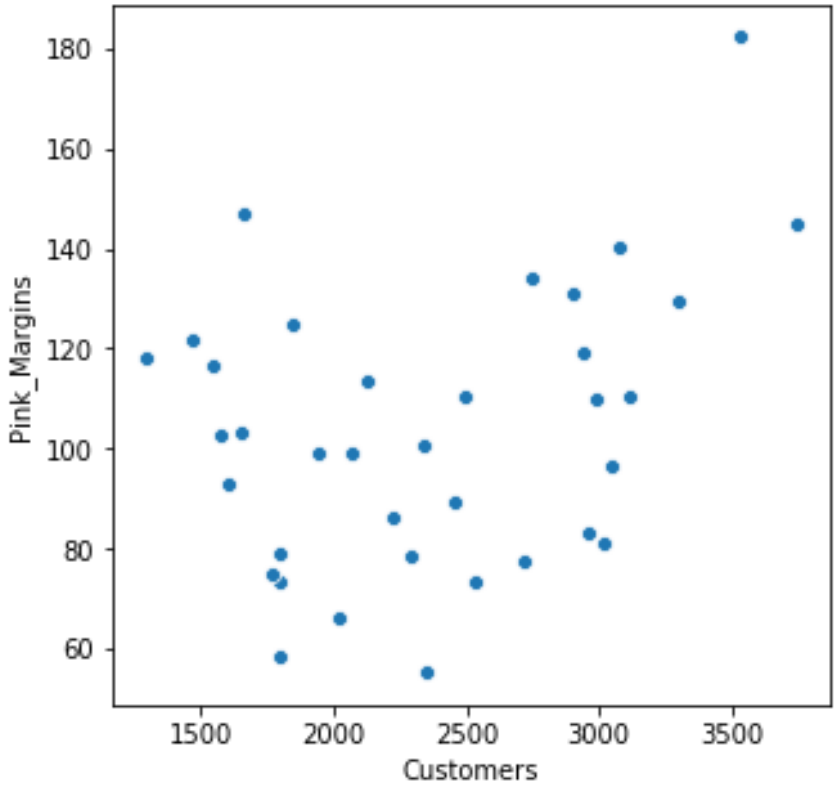
Percentage of cab users in each Income group



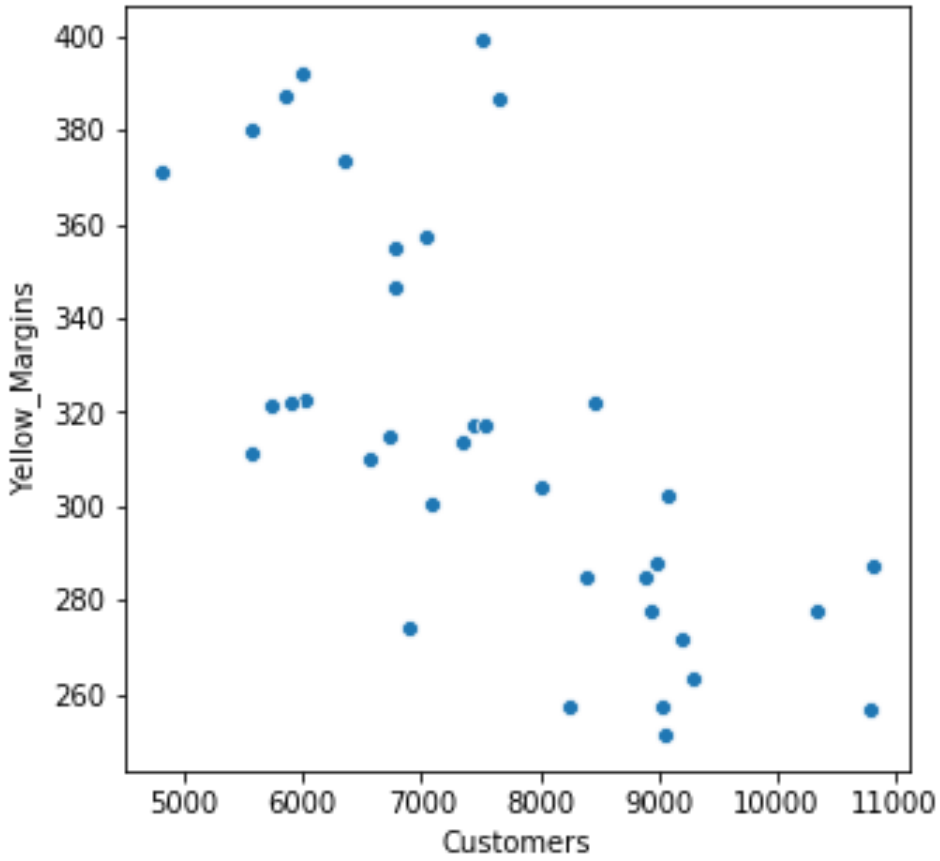
- Looking at the pie chart as illustrated, we can see that there Is a majority of Income group of cab users belonging to an Income group of a High class travelling in Yellow cab.

This means that the High Class Contributes to high investment.

Does margin proportionally increase with increase in number of customers?



- As illustrated on the diagram, we see that the Pink cabs increase their margins with an increase in number of customers.



- Here as we see, the Yellow cabs decrease their margins with an increase in number of customers.

Is there a difference in margin/profit between male and female customers for Yellow cabs?

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

```
18394 21502
We accept alternate hypothesis that there is a statistical difference
P value is  0.005345131177690902
```

- There is a difference in margin/profit for the Yellow cab between male and female customers,and therefore we accept the alternate hypothesis.

Is there a difference in margin/profit between male and female customers for Pink cabs?

- We accept the null hypothesis,and therefore there is no difference in margin/profit for Pink cabs between male and female customers.

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

```
14819 17511
We accept null hypothesis that there is no statistical difference
P value is  0.5889424154257326
```

Is there a difference in margins/profit due to age of customers for Pink cab?

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

```
23721 8609
We accept null hypothesis that theres no difference
P value is  0.7840727156471817
```

- There is no difference is no difference in margin/profit for Pink cab due to the age of customers,and therefore we accept the null hypothesis.
- It doesn't make any difference whether the customer is less than equal to 40 or greater than 40.

Is there a difference in profit/margin due to the age of customers for Yellow cabs?

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

```
29254 6295
We accept null hypothesis that theres no difference
P value is  0.5813353417655228
```

- There is no difference in profit/margin for Yellow cabs due to the age of customers,so we accept the null hypothesis.
- It also doesn't make any difference whether the customer is less than equal to 40 or greater than 40.

Is there a difference between gender and KM Travelled for Yellow cabs?

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

```
18394 21502
We accept null hypothesis that theres no difference
P value is  0.20078753614825767
```

- There is no difference between gender and KM Travelled for yellow cab, so we accept the null hypothesis.

Is there a difference between gender and KM Travelled for Pink cab?

- There is a difference between gender and KM Travelled for Pink cabs.

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

```
14819 17511
We accept alternate hypothesis that theres a difference
P value is  0.045565869972749515
```

Recommendations

- To make a precise decision in which company would be a better investment opportunity, we need to clearly review our figures revealed from the exploratory data analysis.
- The first exploration was determining the number cab users periodically(monthly and yearly) and our insights revealed as follows:
- The Yellow cab resulted in a higher cab users than that of the Pink cab travelling on a monthly basis. The Yellow cab revealed a higher range of cab users than that of the Pink cab.
- The Yellow cab also resulted in a higher cab users than that of the Pink cab travelling on a yearly basis and revealed a higher range of cab users.
- So far, the Yellow cab company has been excelling and therefore is a better investment opportunity for XYZ.

Thank You