
# Project:G2M Insight for Cab Investment

## Data Glacier Virtual Internship

By: Riwaj Neupane

13th August 2023

# Agenda

- Problem Statement
- Dataset information
- EDA
- Hypothesis Testings
- Findings

# Problem Statement

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

Objective: Provide insight to help XYZ in identifying the right company to invest in.

Analysis done:

1. Problem Understanding
2. Finding users in companies
3. Finding profit for both companies
4. Finding cheaper company
5. Multiple hypothesis Testing

# DataSet Information

Data Glacier
Your Deep Learning Partner

The dataset contains 4 individual dataset

1. Cab_data: Contains 7 columns and 359392 observations, containing company type, price,km travelled.
2. City: This file contains 3 columns and 20 observations, containing city and their respective users.
3. customer_data: This file contains 4 columns and 49171 observations. containing users and their respective age, gender and income
4. Transaction_data: This file contains 3 columns and 440098 observations containing mode of payment for a particular customer.

Combining Dataset to form a complete merged dataset:

- Combining transaction and customer data on column ‘Customer ID’.
- Then combine it with cab data on column ‘Transaction ID’.
- Then combine it with city data on column ‘City’
- In combined dataset created a new column users per meaning dividing the users by total population.

# DataSet Information (contd…)

Data Glacier
Your Deep Learning Partner

Combined dataset Details:

Contains 20 columns and 359392 observations.

Data columns (total 19 columns):

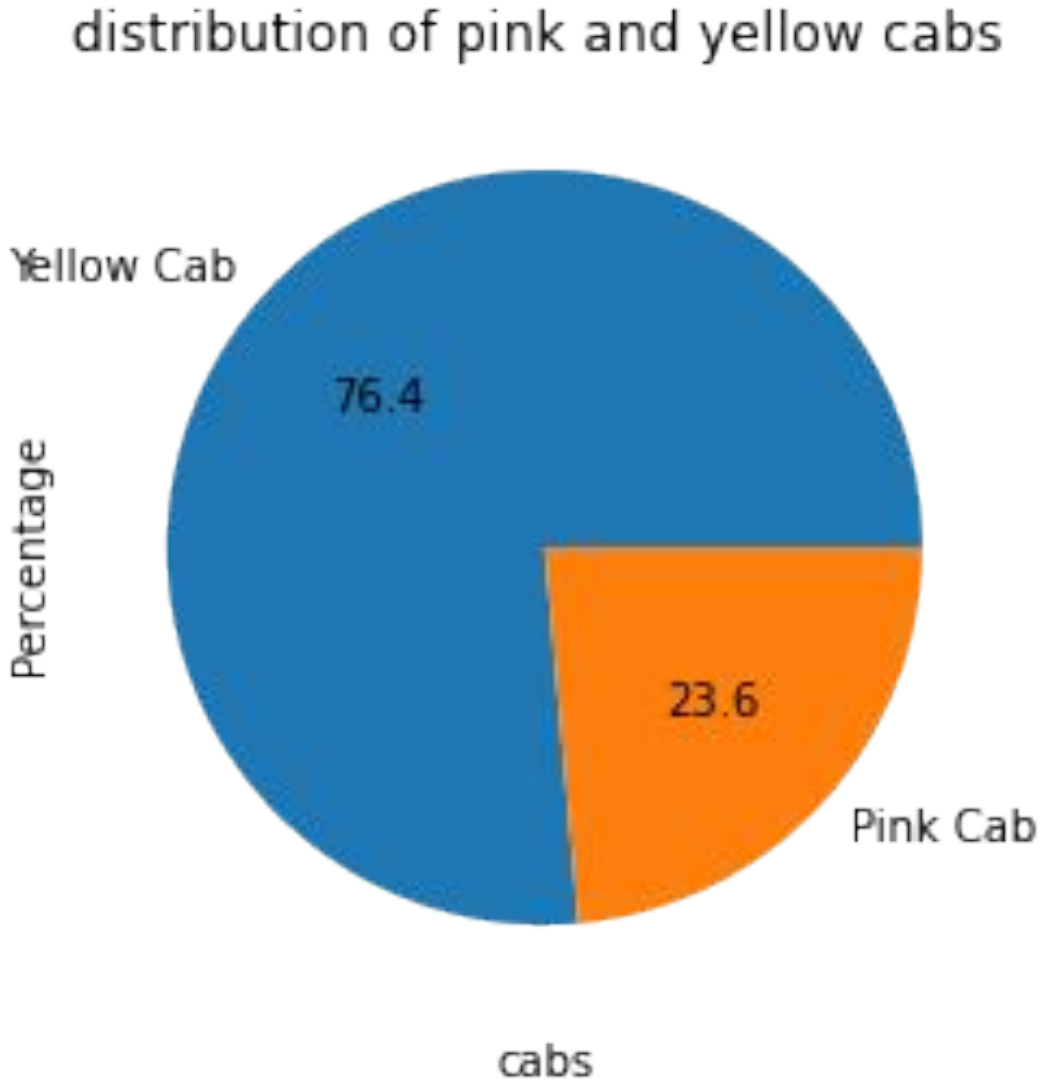
| # | Column | Non-Null Count | Dtype |
|---|---|---|---|
| 0 | Transaction ID | 359392 non-null | int64 |
| 1 | Customer ID | 359392 non-null | int64 |
| 2 | Payment_Mode | 359392 non-null | object |
| 3 | Gender | 359392 non-null | object |
| 4 | Age | 359392 non-null | int64 |
| 5 | Income (USD/Month) | 359392 non-null | int64 |
| 6 | Date of Travel | 359392 non-null | object |
| 7 | Company | 359392 non-null | object |
| 8 | City | 359392 non-null | object |
| 9 | KM Travelled | 359392 non-null | float64 |
| 10 | Price Charged | 359392 non-null | float64 |
| 11 | Cost of Trip | 359392 non-null | float64 |
| 12 | Population | 359392 non-null | int64 |
| 13 | Users | 359392 non-null | int64 |
| 14 | Year | 359392 non-null | int64 |
| 15 | Month | 359392 non-null | int64 |
| 16 | Day | 359392 non-null | int64 |
| 17 | Profit | 359392 non-null | float64 |
| 18 | users per | 359392 non-null | float64 |

dtypes: float64(5), int64(9), object(5)

# EDA

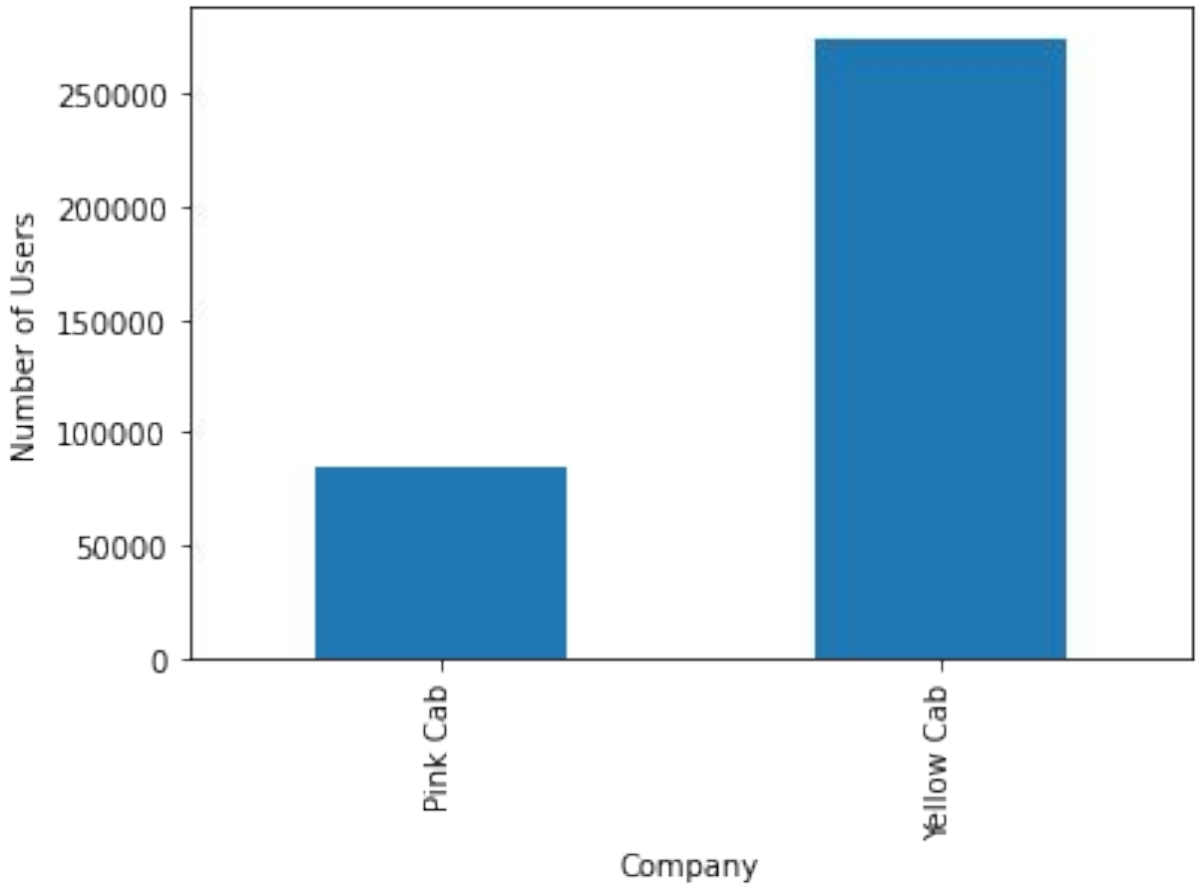
# Which company has more cabs ?

Yellow Cab



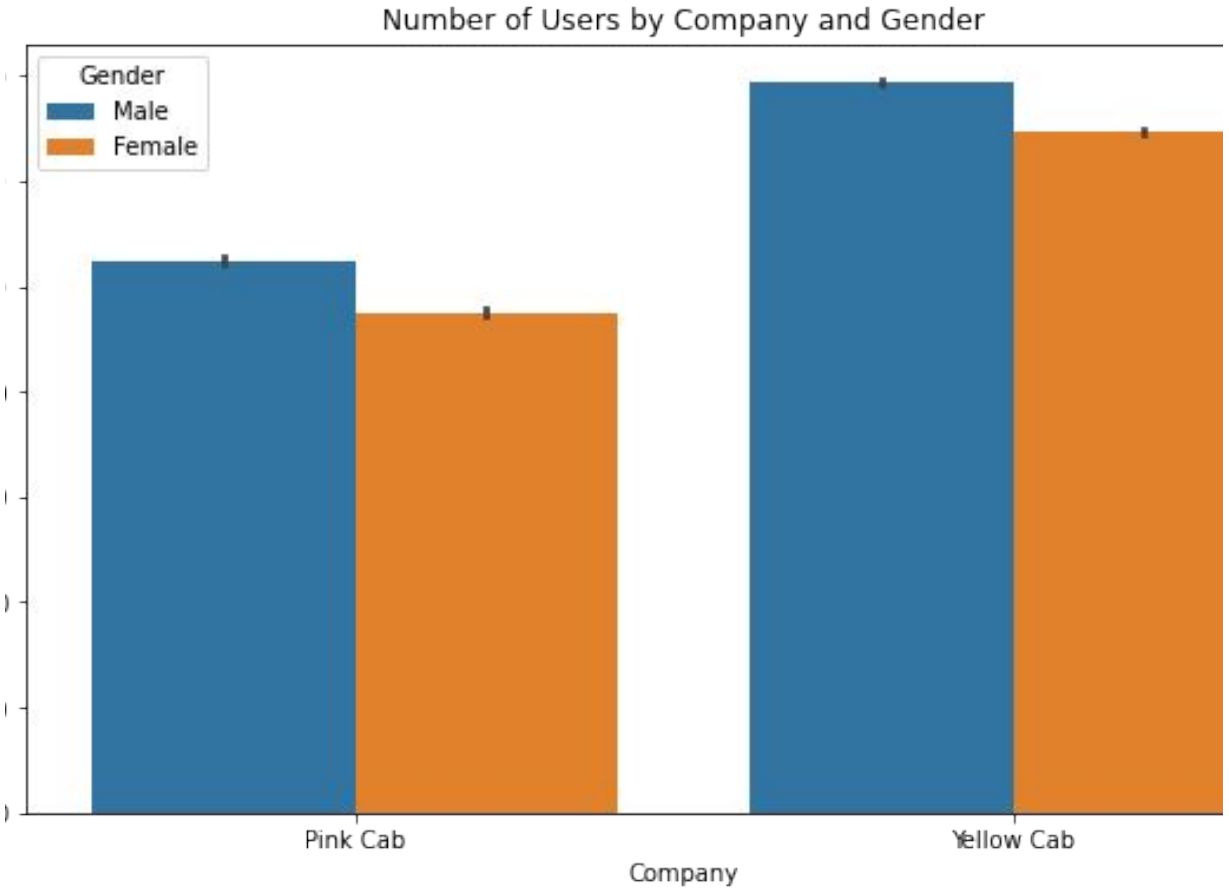
# Which company has more users?

Yellow Cab has more users



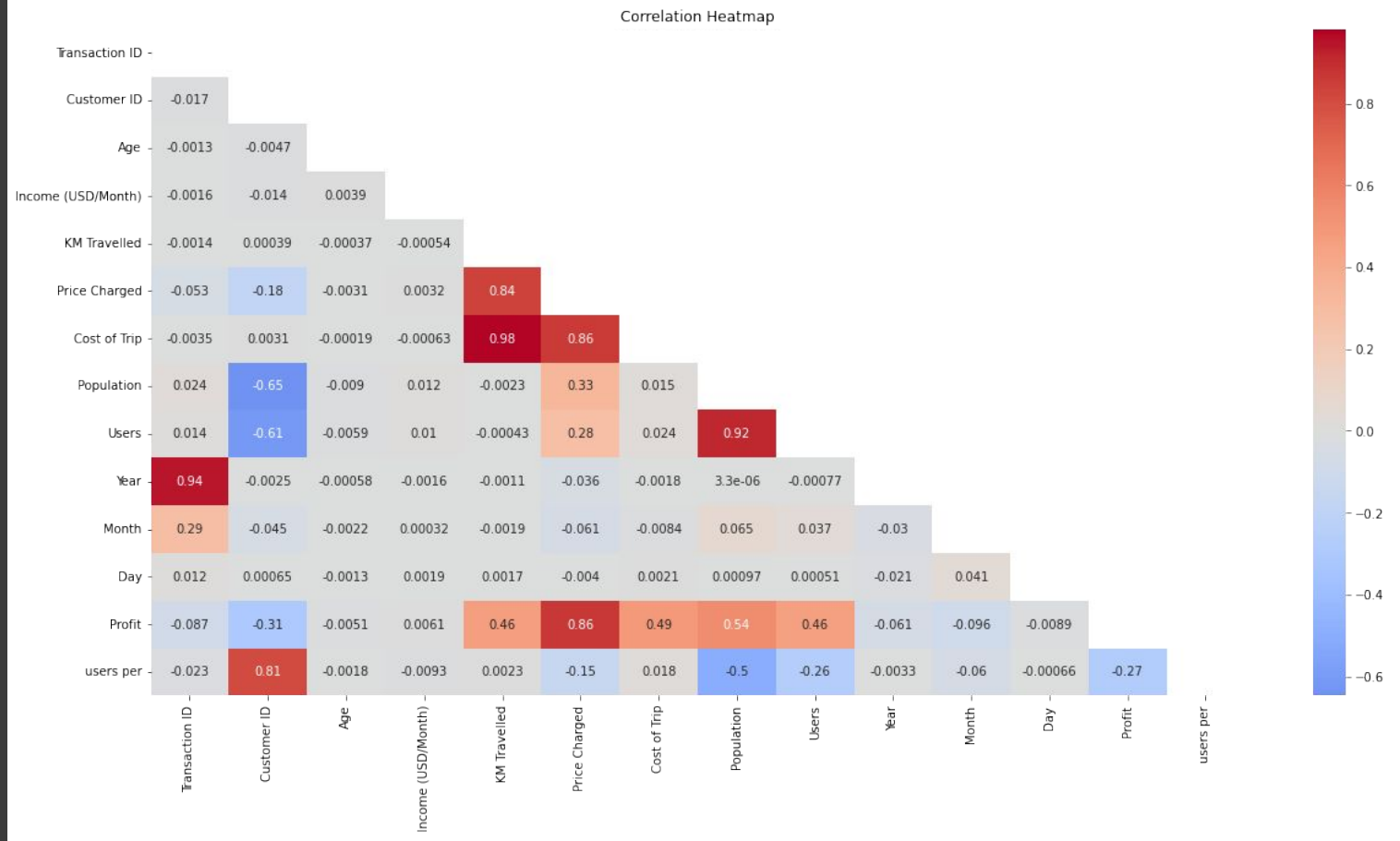
# Which company has more users w.r.t Gender?

There are more male than female users for both company



# Checking Correlation

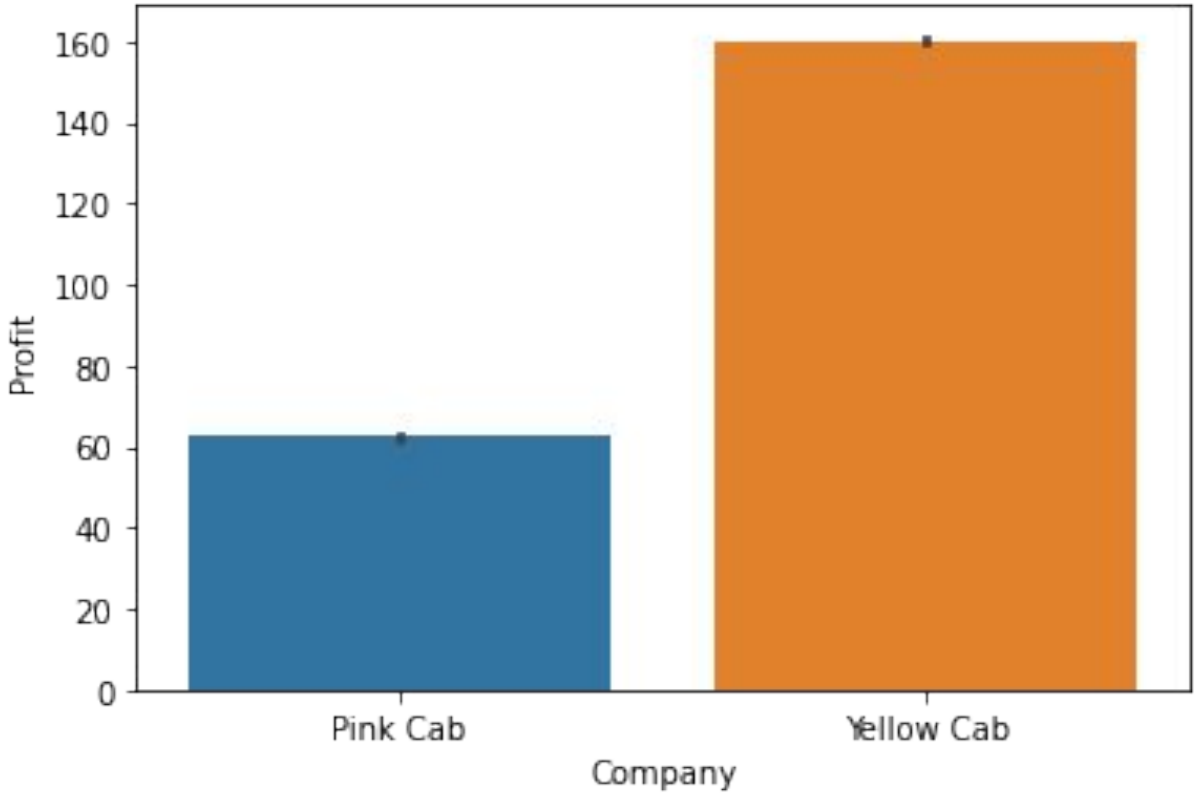
1. Population vs Users.
2. Price charged vs Cost of trip vs Profit
3. Km travelled vs Price Charged




Data Glacier
Your Deep Learning Partner

# Which company has more profit

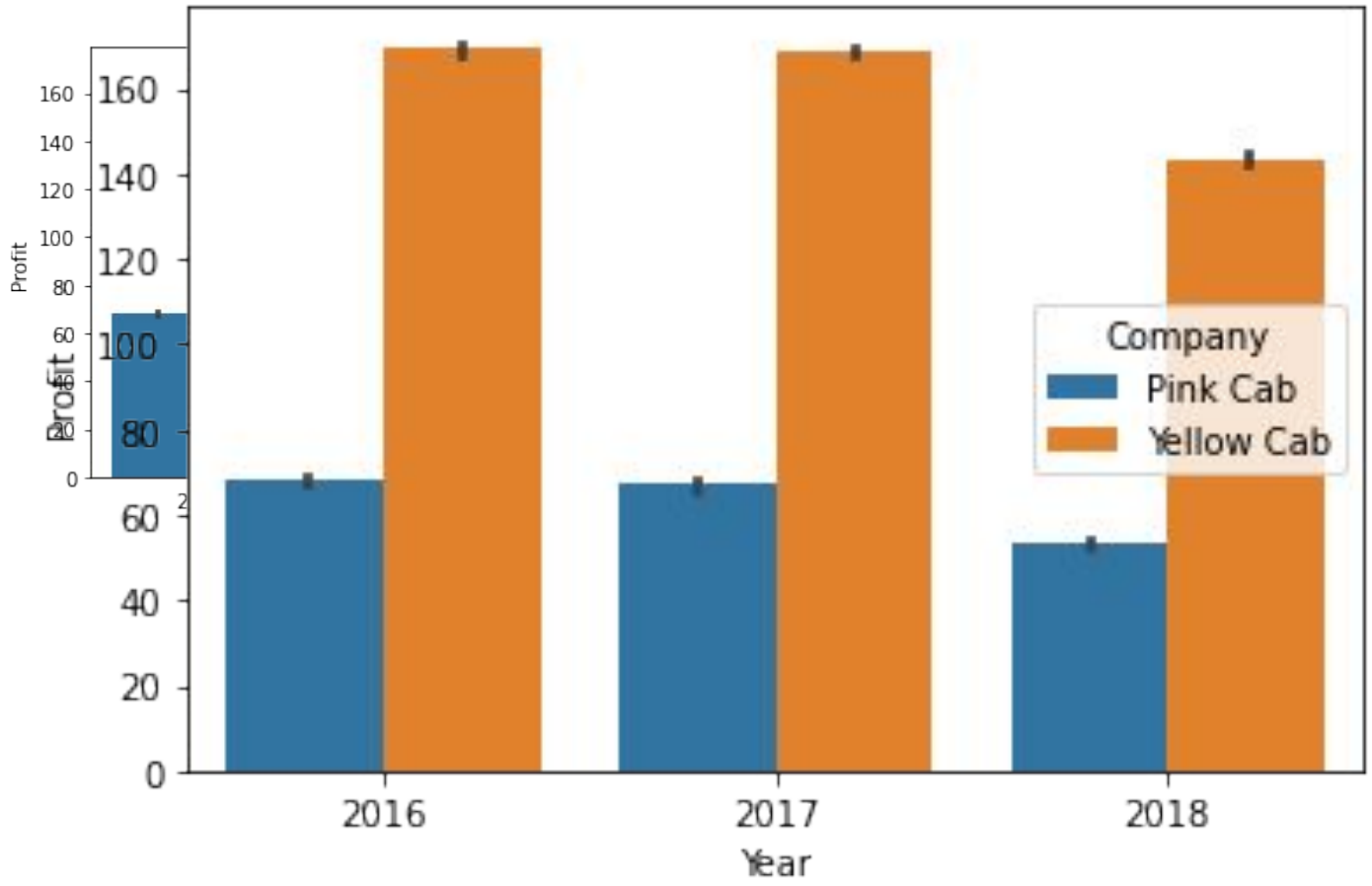
Yellow cab has more profit than pink cab




Data Glacier
Your Deep Learning Partner

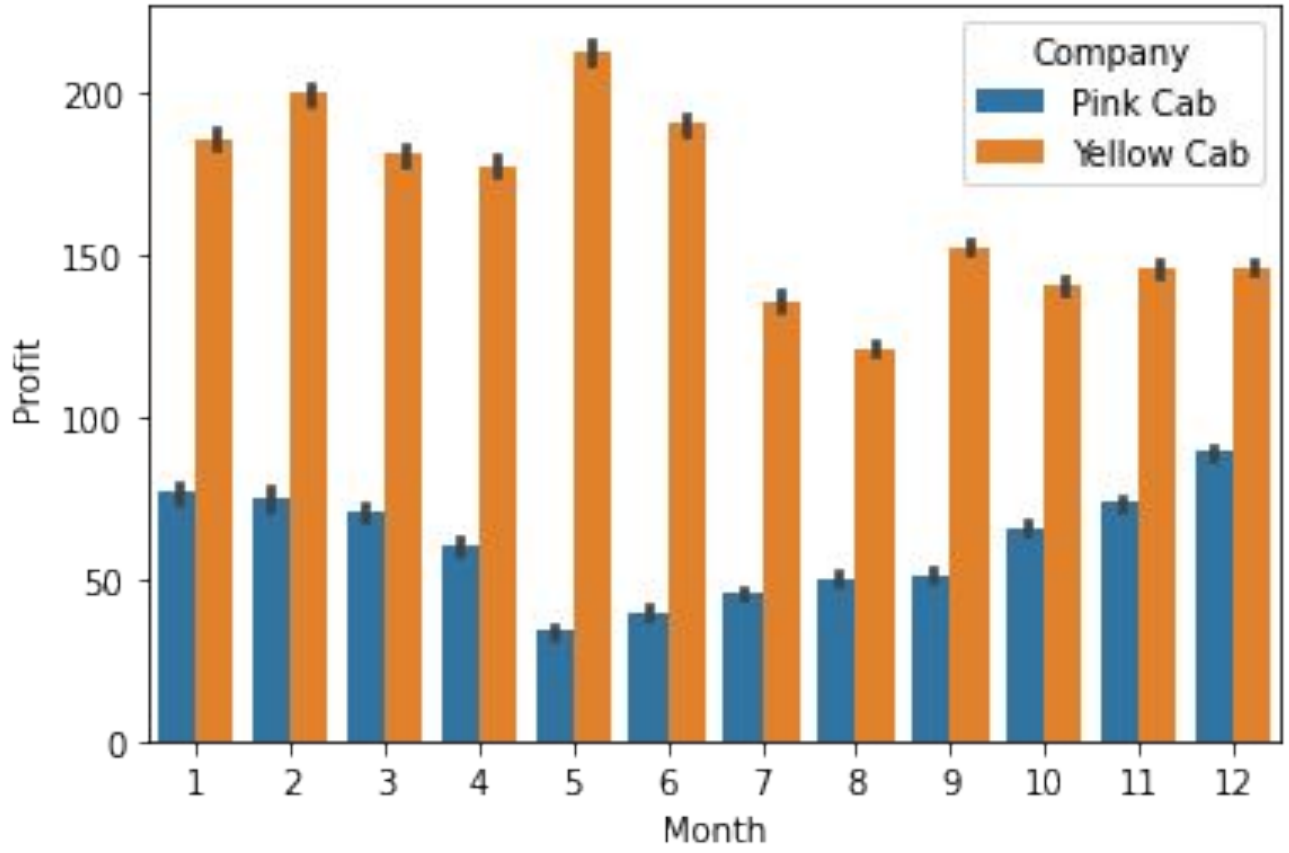
# Seeing the profit year wise for both the companies

2018 has least profit for both companies



# Seeing the profit month wise for both the companies

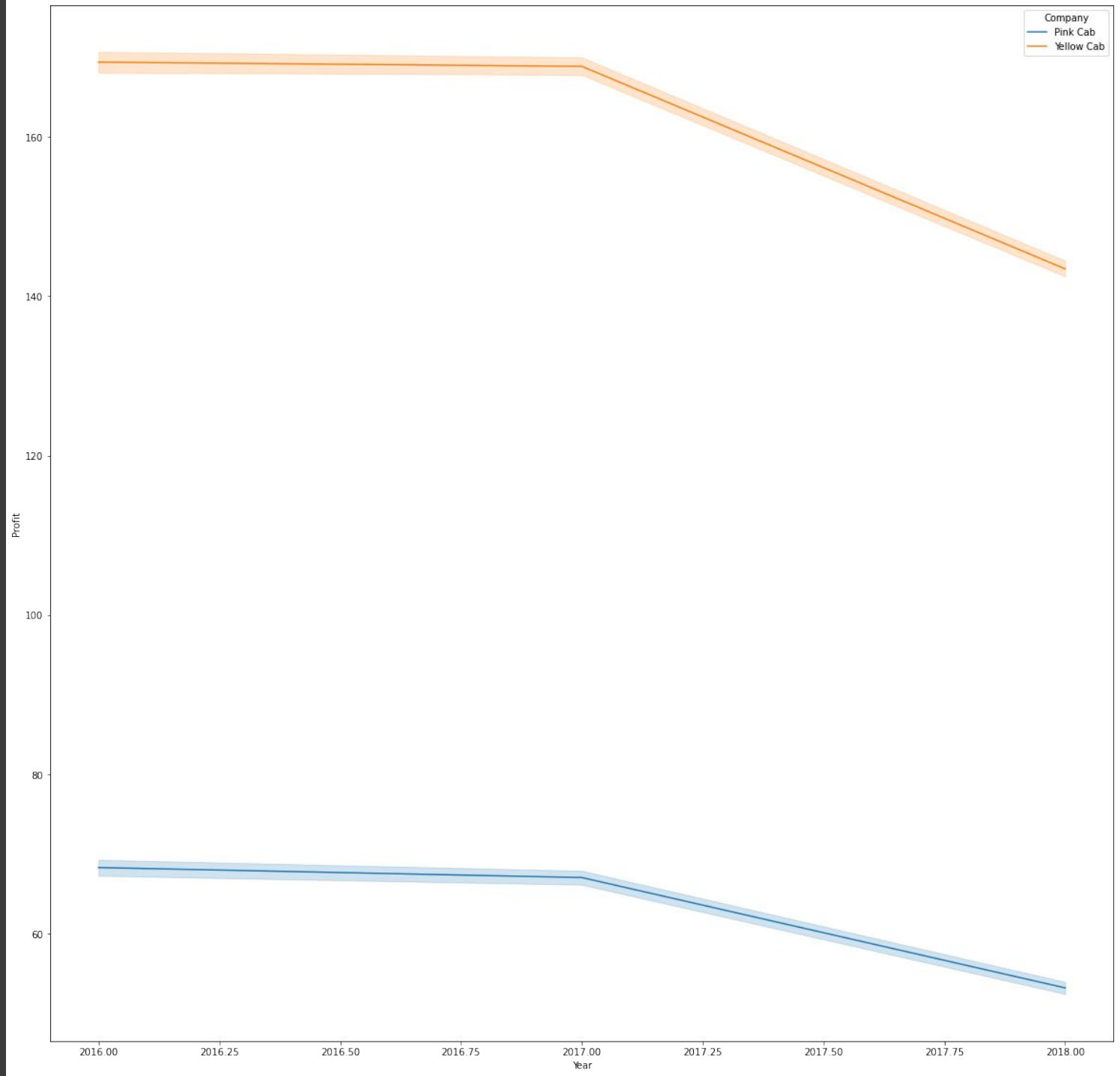
Yellow cab has highest profit for month of May, Pink cab has highest profit from month December




Data Glacier
Your Deep Learning Partner

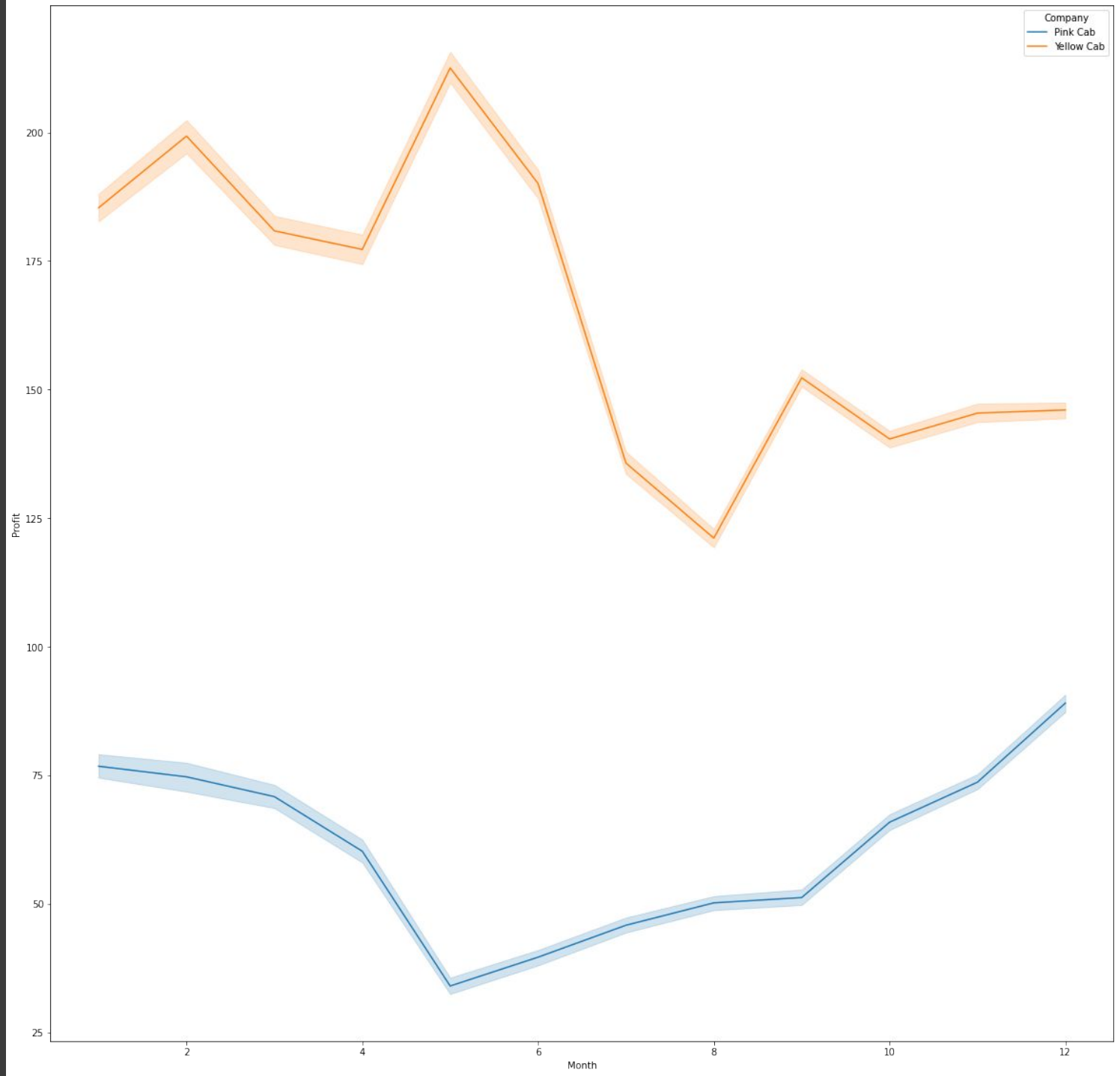
# Profit w.r.t Year

We can see that profit is decreasing for both company year wise




Data Glacier
Your Deep Learning Partner

# Profit w.r.t Month



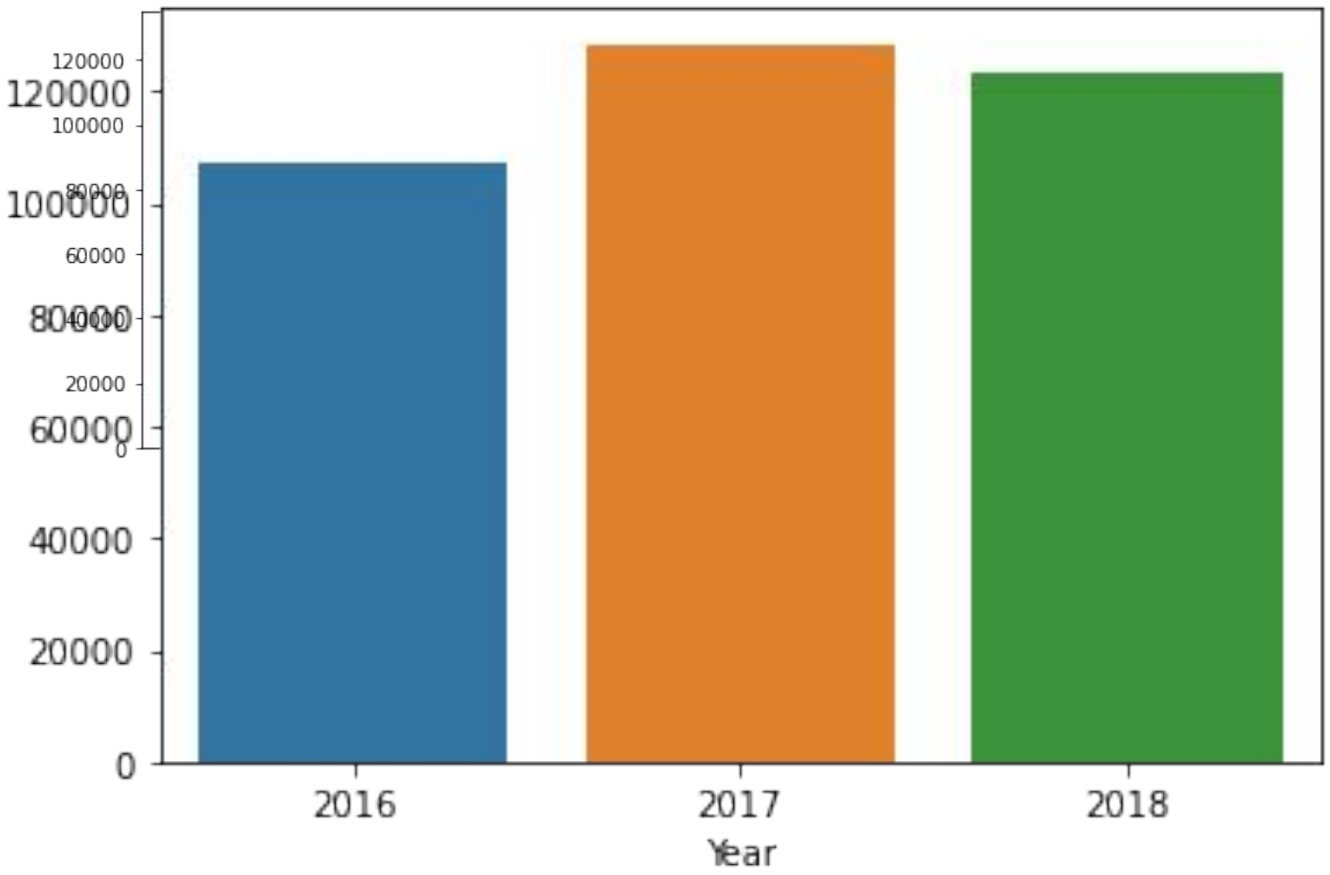

Data Glacier
Your Deep Learning Partner

# Payment mode for both companies

Equal users using both modes of payment

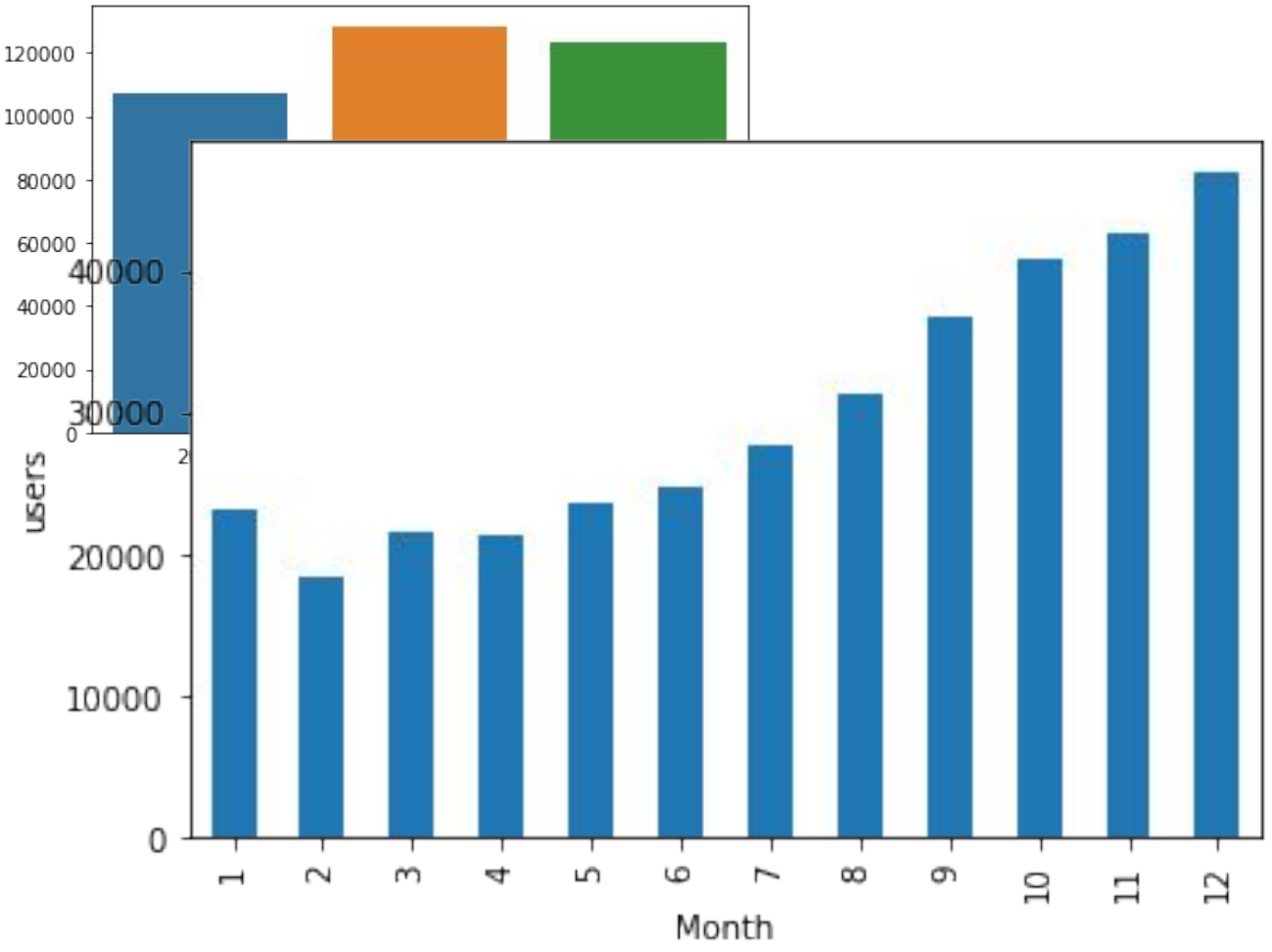
# Travel frequency per year

2017 has highest tavel



# Travel frequency per month

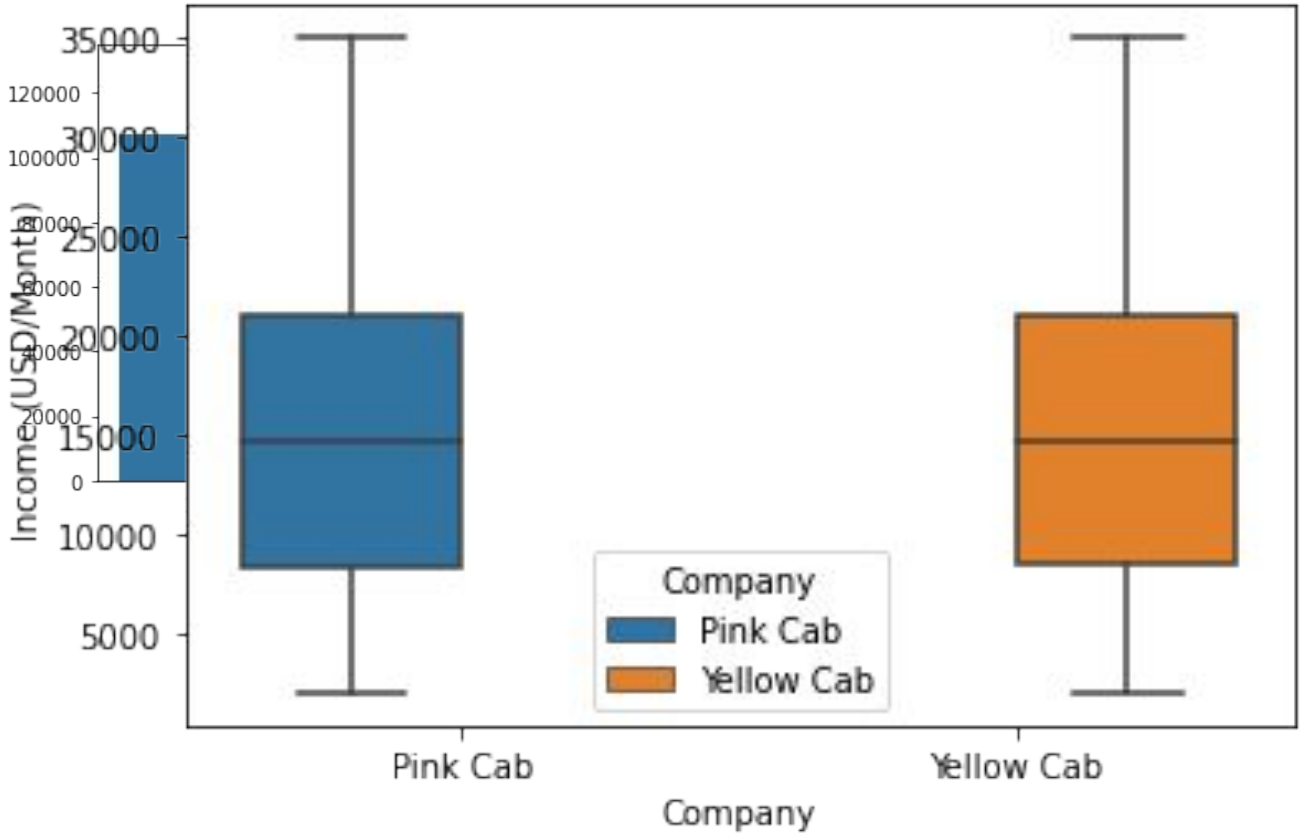
December has highest tavel




Data Glacier
Your Deep Learning Partner

# Distribution of income of users

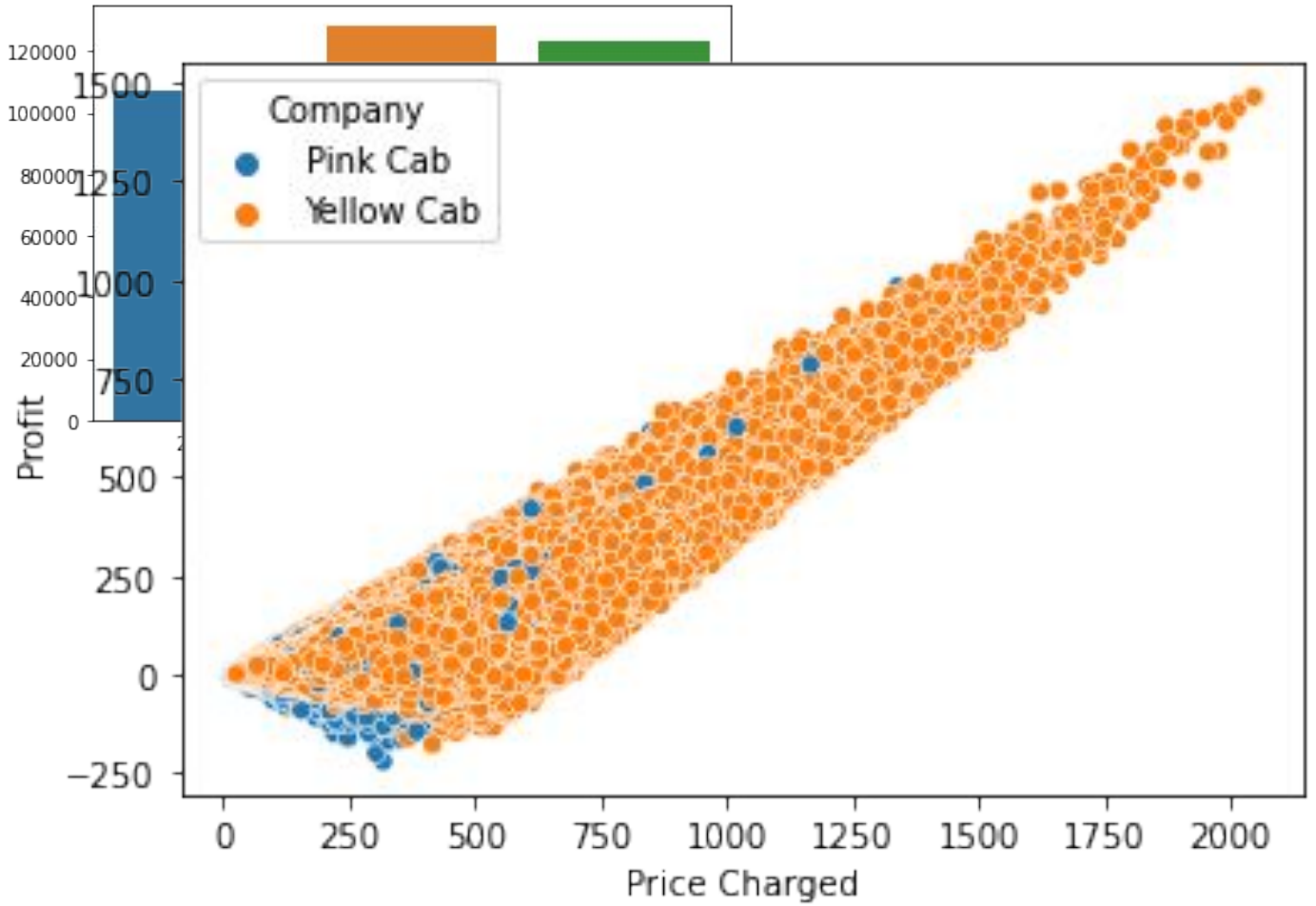
Both companies users have a average income of around 15k




Data Glacier
Your Deep Learning Partner

# Is there a relation between price charged and profit

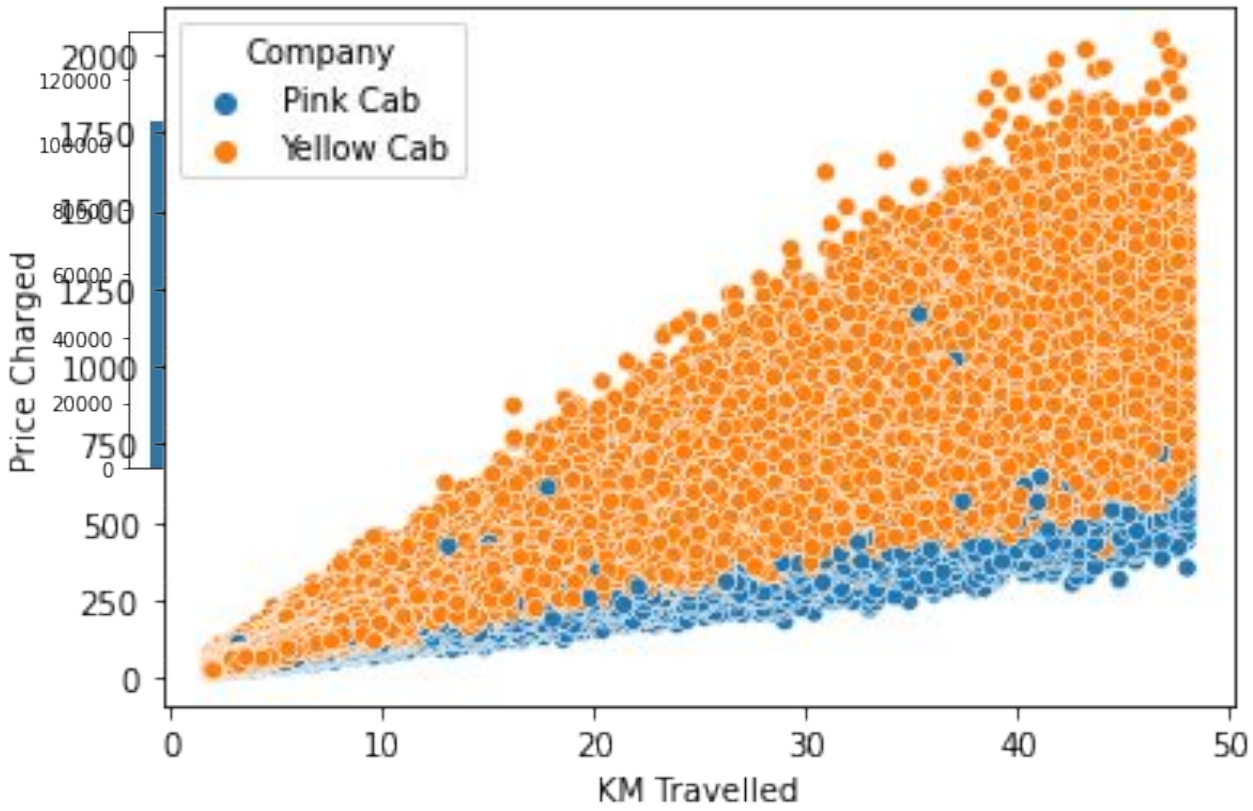
clearly we can see that there is a linear relation between price charged and profit earned




Data Glacier
Your Deep Learning Partner

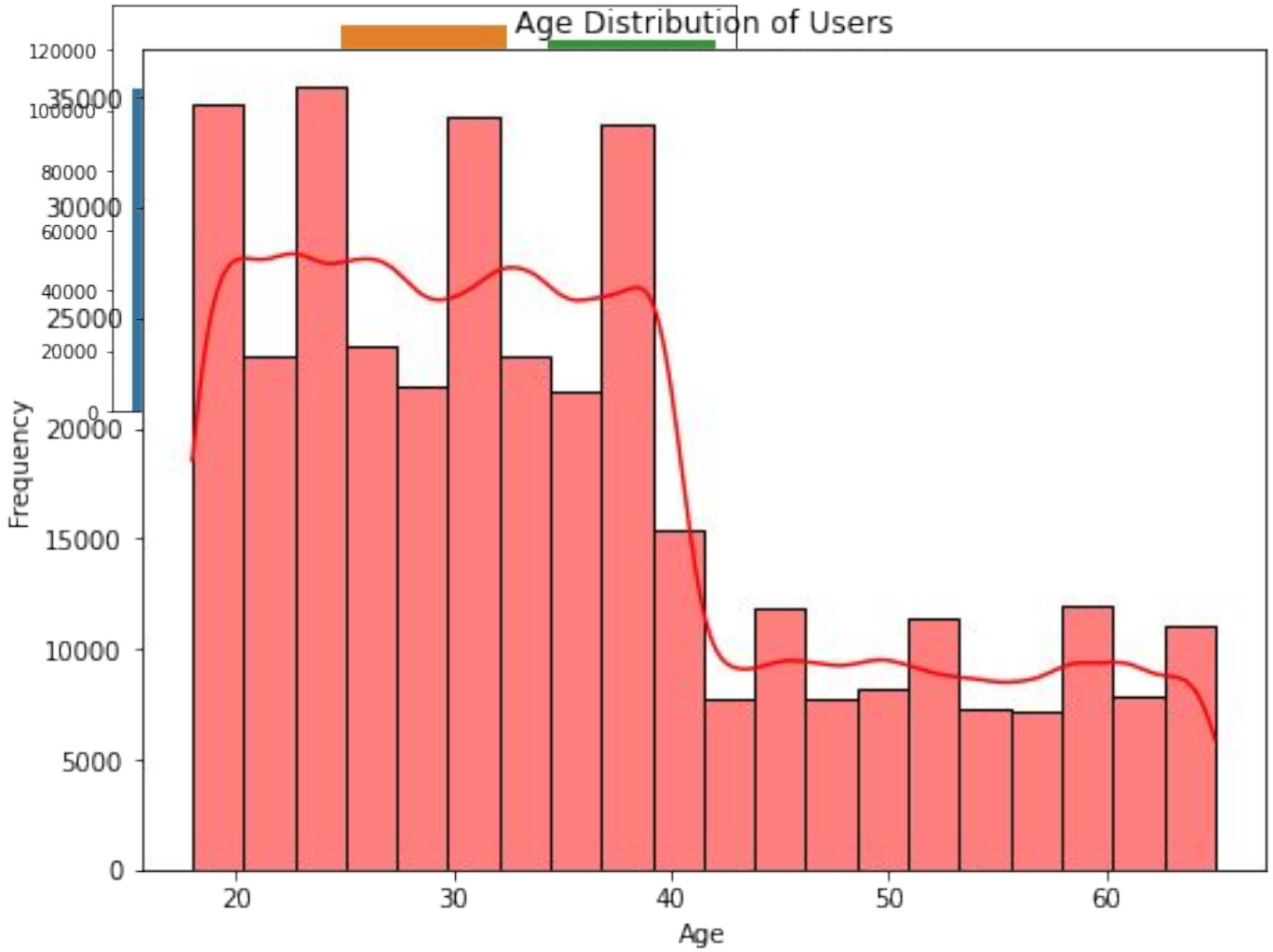
# relationship between km travelled and price charged

There is a linear relationship between KM travelled and price charged



# distribution of ages of users in dataset

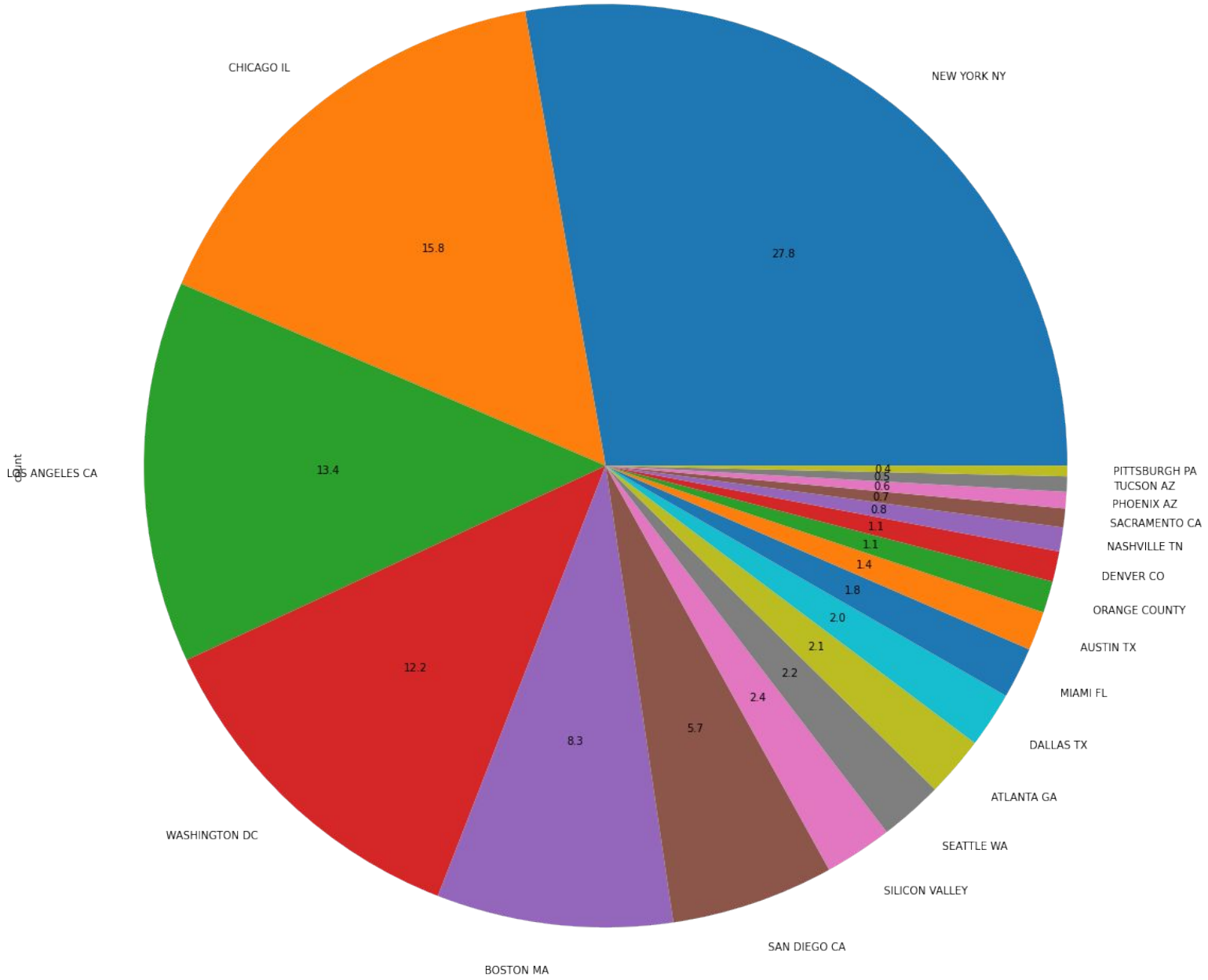
Most users lie between age of 20 and 40




Data Glacier
Your Deep Learning Partner

# Users Per city

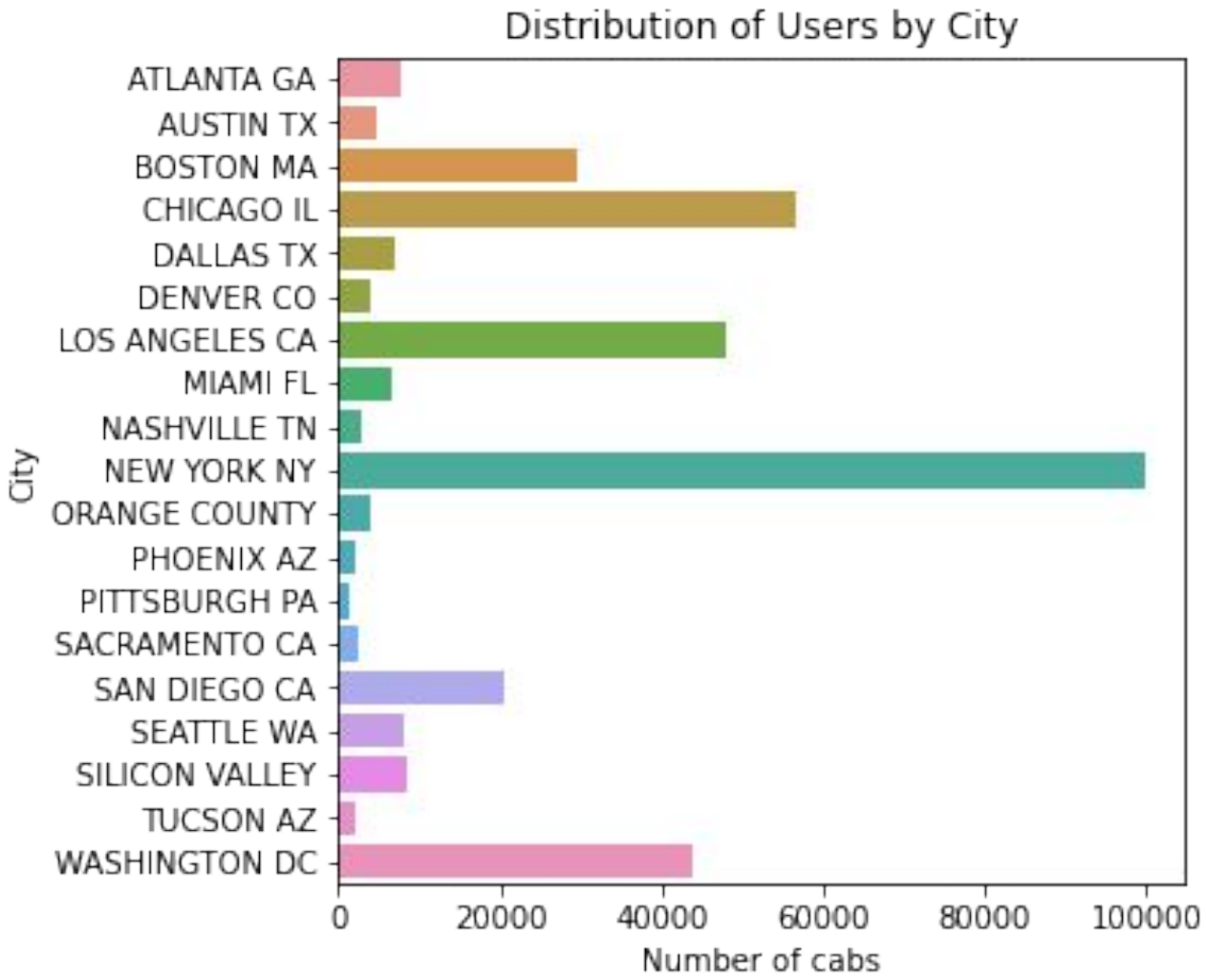
NY has more users than others




Data Glacier
Your Deep Learning Partner

# see cabs in cities

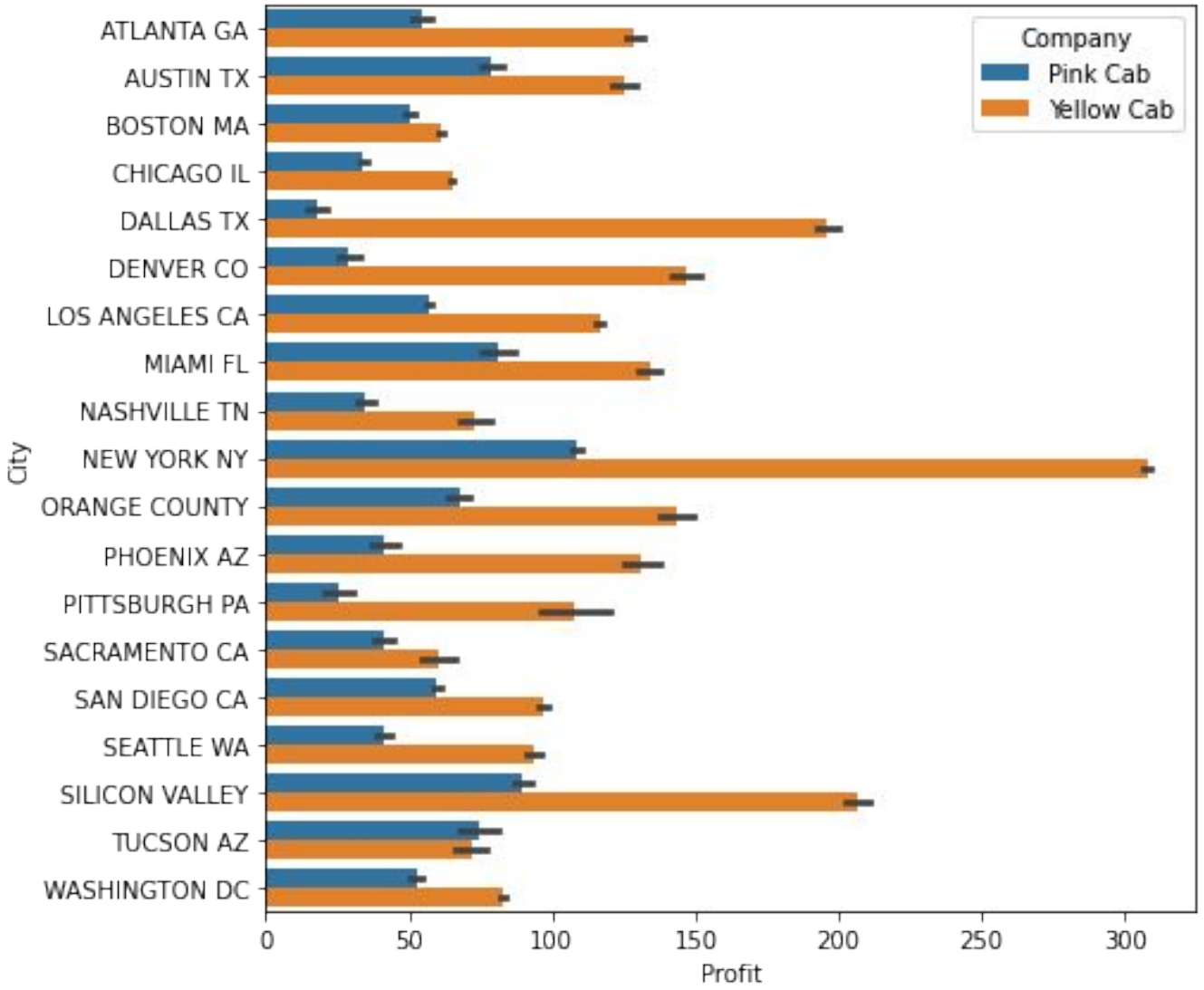
NY has highest number of cabs




Data Glacier
Your Deep Learning Partner

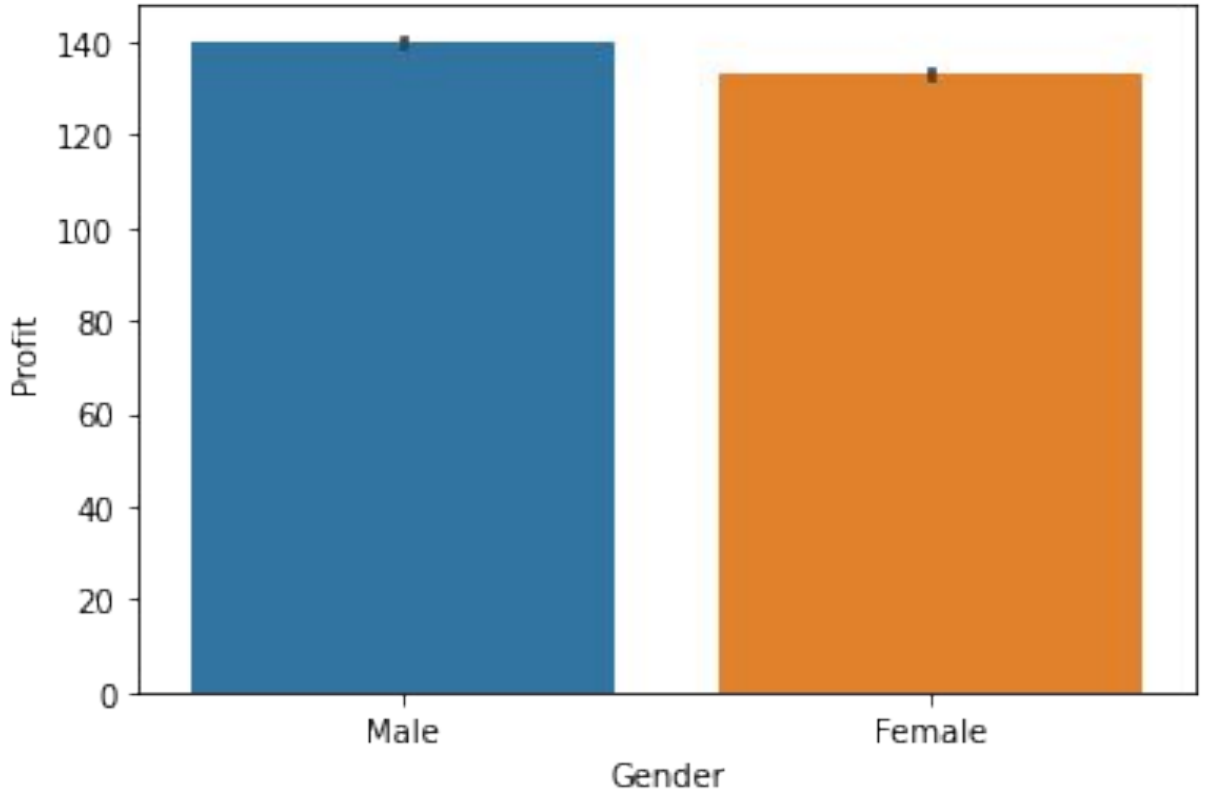
# see profit of cabs in cities

NY has highest profit for both cabs



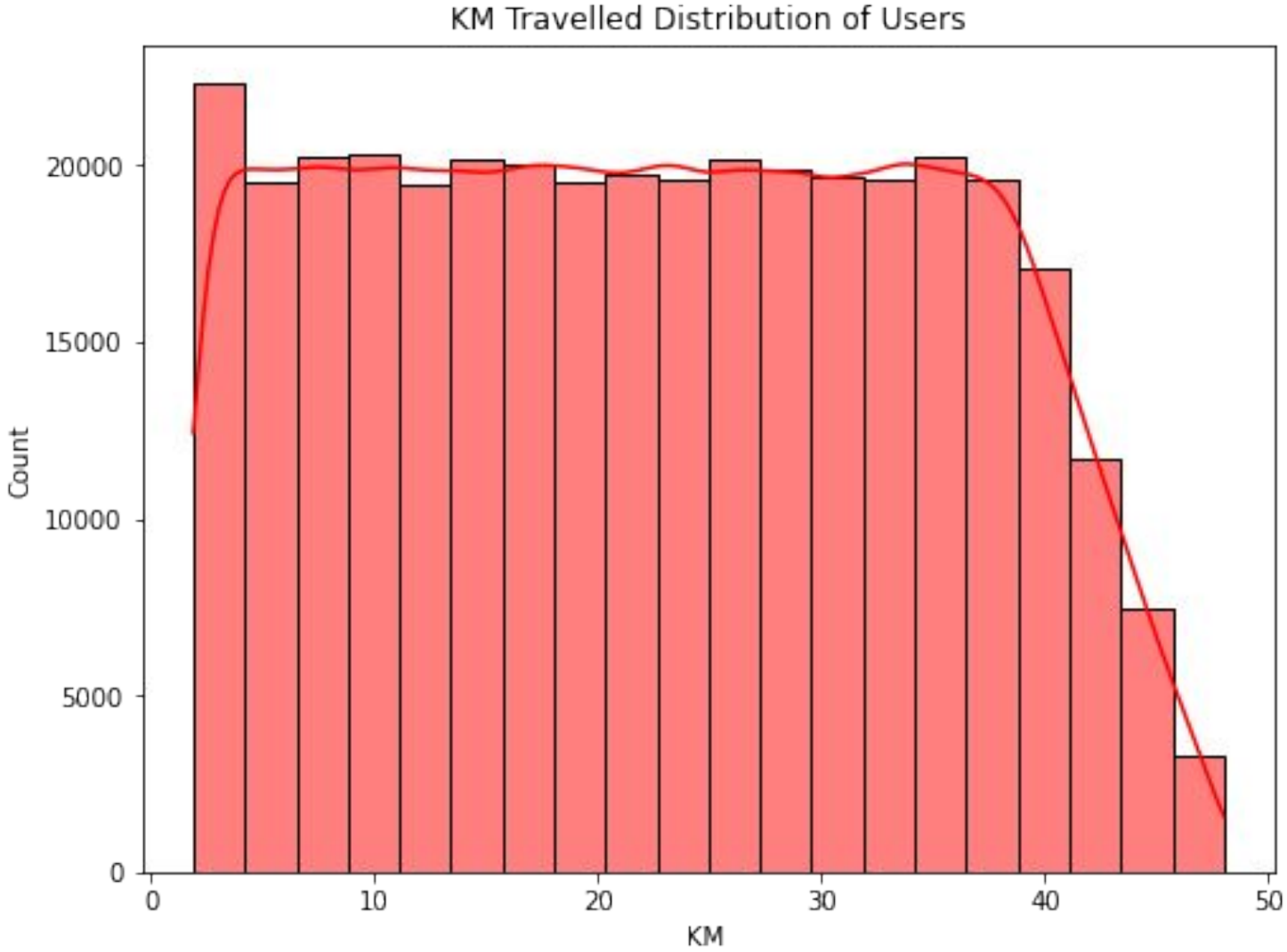
# Does profit vary with gender

Male users result in higher profit




Data Glacier
Your Deep Learning Partner

# KM travelled distribution




Data Glacier
Your Deep Learning Partner

# Hypothesis Testing

“Is there any seasonality in number of customers using the cab service?”

H0: There is seasonality
H1: There is no seasonality

For Pink cab:

```
P value for Pink Cab: 0.3753760105985561
Pink Cab: The time series is not stationary. Seasonality might not be present.
```

For yellow cab:

```
P value is  0.28017298125790135
Yellow Cab: The time series is not stationary. Seasonality might not be present.
```

# Is there any difference regarding profit and gender

H0: there is no difference in profit regarding gender

H1: there is a difference in profit regarding gender

For Pink cab:

```
p value is: 0.11515305900425798
For Pink Cab: We accept null hypothesis (H0) that there is no difference in profit regarding gender.
```

For yellow cab:

```
p value is: 6.060473042494144e-25
For Yellow Cab: We accept alternative hypothesis (H1) that there is a difference in profit
```

Data Glacier
Your Deep Learning Partner

# Is there any difference regarding city and profit

H0: there is no difference in profit regarding city

H1: there is difference in profit regarding city

For Pink cab:

```
p value is: 0.0
For Pink Cab: We reject the null hypothesis (H0) that there is no difference in profit regarding city.
```

For yellow cab:

```
p value is: 0.0
For Yellow Cab: We reject the null hypothesis (H0) that there is no difference in profit regarding city.
```

# Is there any difference in payment mode and profit

H0: there is no difference in profit regarding payment mode

H1: there is difference in profit regarding payment mode

## For Pink cab:

```
p value is: 0.7900465828758374
For Pink Cab: We accept the null hypothesis (H0) that there is no difference in profit regarding payment mode.
```

## For yellow cab:

```
p value is: 0.293306063875188
For Yellow Cab: We accept the null hypothesis (H0) that there is no difference in profit regarding payment mode.
```

# is there any difference in income and profit

H0: there is no correlation between income and profit.

H1: there is correlation between income and profit

For Pink cab:

```
p value is: 0.21074604704768066
For Pink Cab: We accept the null hypothesis (H0) that there is no correlation between income and profit.
```

For yellow cab:

```
p value is: 0.00018411176745302524
For Yellow Cab: We reject the null hypothesis (H0) that there is no correlation between income and profit.
```

Data Glacier
Your Deep Learning Partner

# Is there any difference between KM Travelled and Profit

H0: there is no correlation between KM Travelled and profit

H1: there is correlation between KM Travelled and profit

## For Pink cab:

```
p value is: 0.0
For Pink Cab: We reject the null hypothesis (H0) that there is no correlation between KM Travelled and profit.
```

## For yellow cab:

```
p value is: 0.0
For Yellow Cab: We reject the null hypothesis (H0) that there is no correlation between KM Travelled and profit.
```

# Is there any difference between Cost of Trip and Profit

H0: there is no correlation between Cost of Trip and profit

H1: there is correlation between Cost of Trip and profit

## For Pink cab:

```
p value is: 0.0
For Pink Cab: We reject the null hypothesis (H0) that there is no correlation between Cost of Trip and profit.
```

## For yellow cab:

```
p value is: 0.0
For Yellow Cab: We reject the null hypothesis (H0) that there is no correlation between Cost of Trip and profit.
```

# Is there any difference between Profit and Year

H0: there is no difference in profit based on the year

H1 : there is difference in profit based on the year.

For Pink cab:

```
p value is: 1.3845866439321786e-145
For Pink Cab: We reject the null hypothesis (H0) that there is no difference in profit based on the year.
```

For yellow cab:

```
p value is: 5.3455921384708516e-301
For Yellow Cab: We reject the null hypothesis (H0) that there is no difference in profit based on the year.
```

Data Glacier
Your Deep Learning Partner

# Findings

1. Yellow cab has more users
2. There are more yellow cabs than pink cabs
3. There are more male users than female users for both the companies
4. Yellow cab has more profit than pink cabs
5. Profit for both companies is decreasing year by year
6. 2017 has highest travel
7. December month has the highest travel
8. Both companies users have a average income of around 15k
9. Most users lie between age of 20 and 40
10. NY has more users than others
11. NY has highest profit for both cabs

# Thank You!!!

www.thebodytransformation.com