



**Data Glacier**

Your Deep Learning Partner

# **G2M insight for Cab Investment firm Data Analyst Virtual Internship**

**Submitted by : Anitha Venkatachalam**  
**Date:10/11/2022**

# Agenda

Executive Summary

Problem Statement

Approach

EDA

EDA Summary

Hypothesis Testing

Building Models

Recommendations

# Description:

- ❑ **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.**
- ❑ **Provide actionable insights to help XYZ firm in identifying the right company for making investment.**
- ❑ **Cab Companies:**
  - **Yellow Cab**
  - **Pink Cab**
- ❑ **The Analysis include :**
  - **Data Understanding,**
  - **Data Visualization,**
  - **Creating multiple hypothesis,**
  - **Building models and finding the best fit model based on Accuracy.**

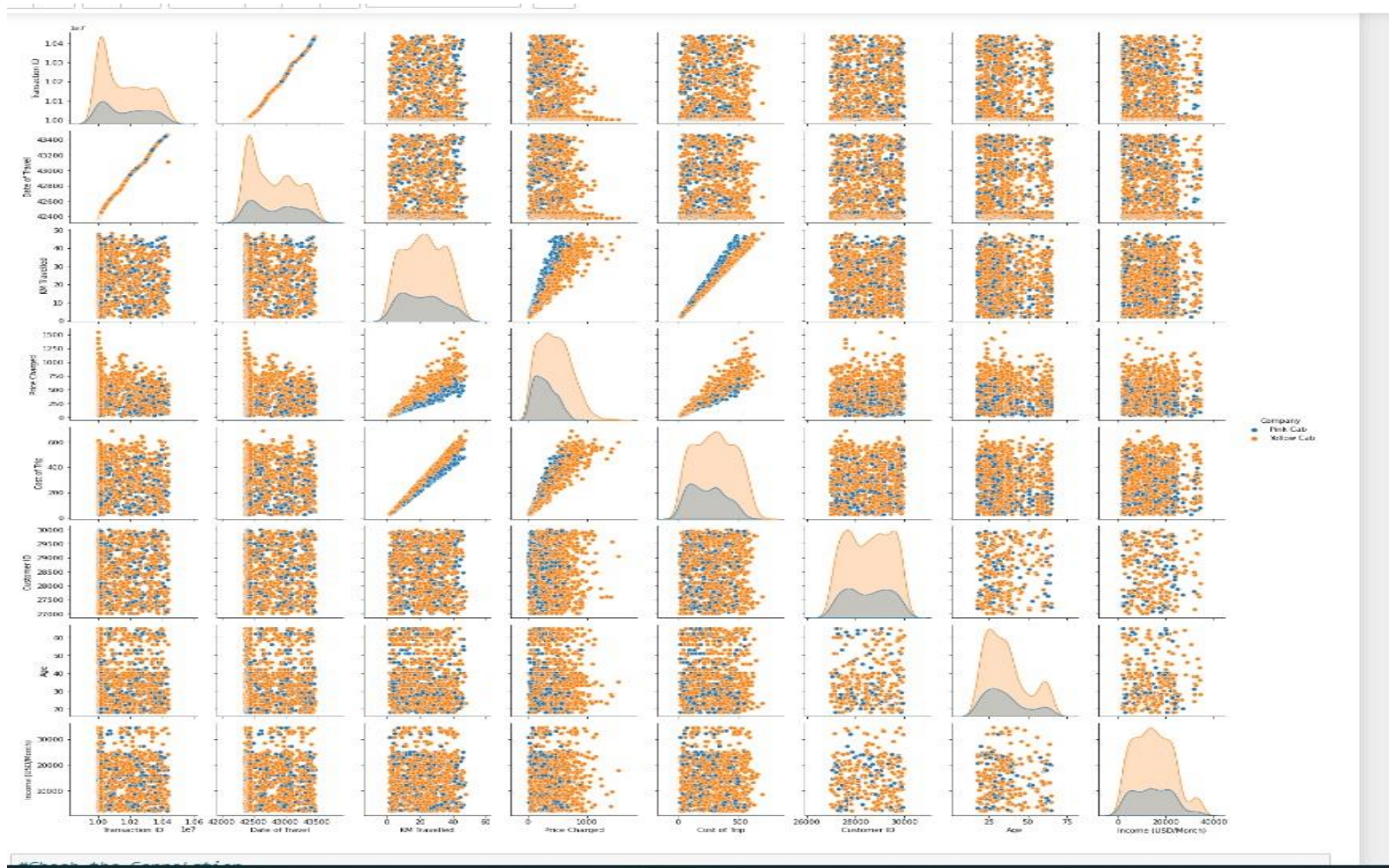
# Data Preparation:

## There are 4 datasets:

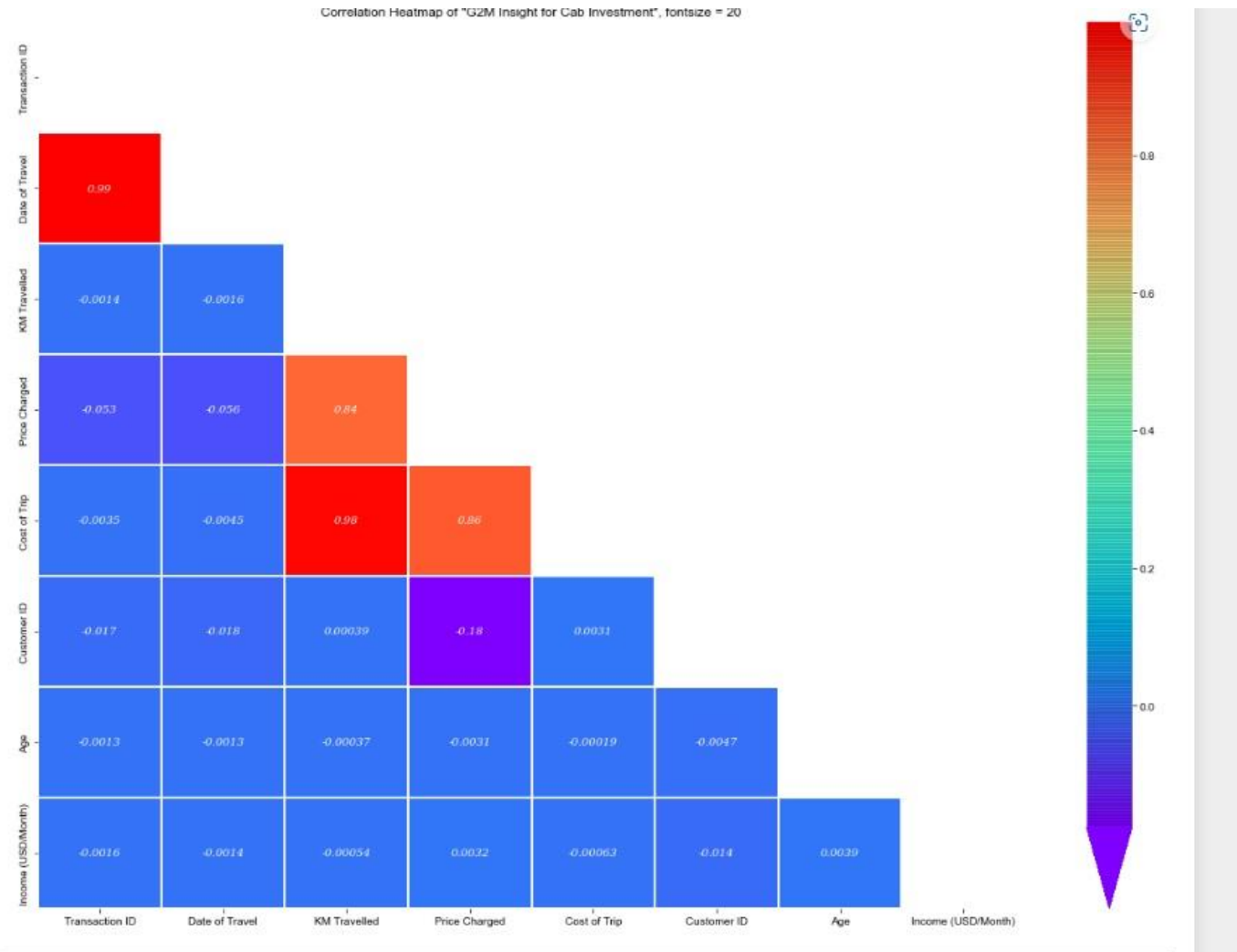
- **Cab\_Data.csv** – this file includes details of transaction for 2 cab companies.
- **Customer\_ID.csv** – this is a mapping table that contains a unique identifier which links the customer's demographic details.
- **Transaction\_ID.csv** – this is a mapping table that contains transaction to customer mapping and payment mode.
- **City.csv** – this file contains list of US cities, their population and number of cab users.

# **EXPLORATORY DATA ANALYSIS**

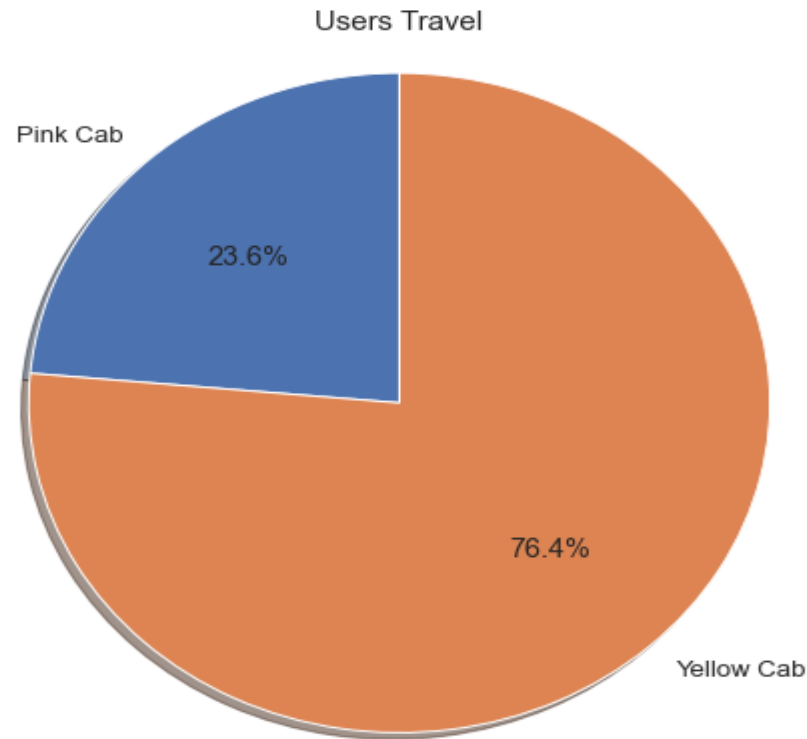
# Relationships Between Variables:



# Correlation Between Variables:



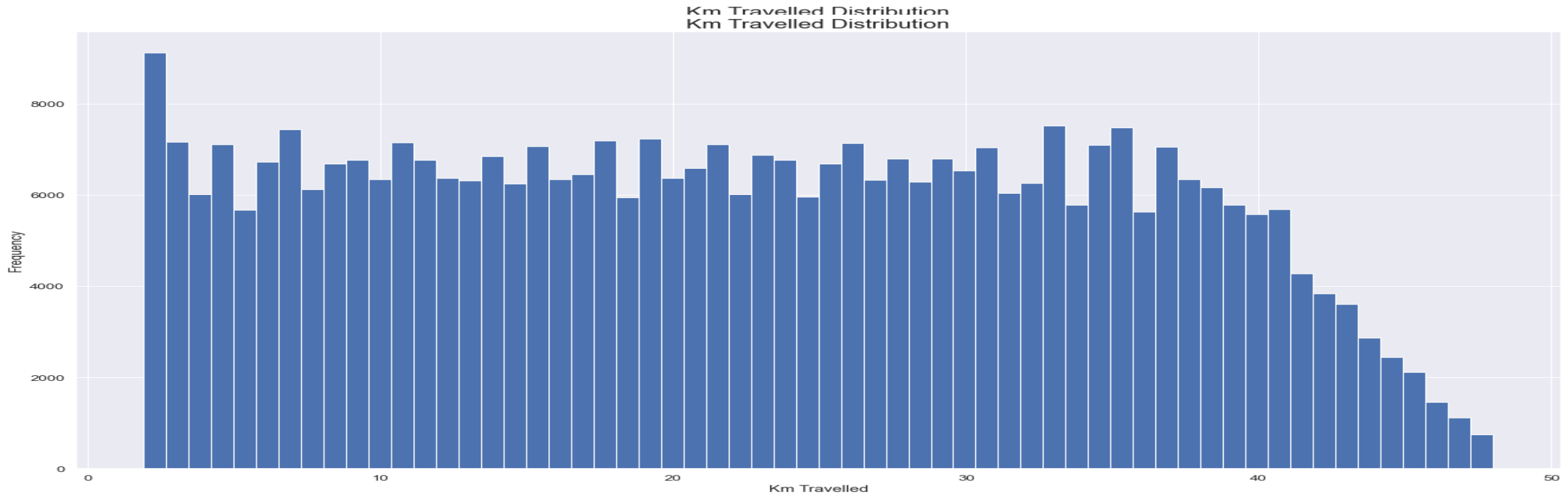
# Which cab have more users :



- ❑ From the above graphs, we can see that for Yellow Cab most of the rides are in the range of 76.4% and pink cab is 23.6%

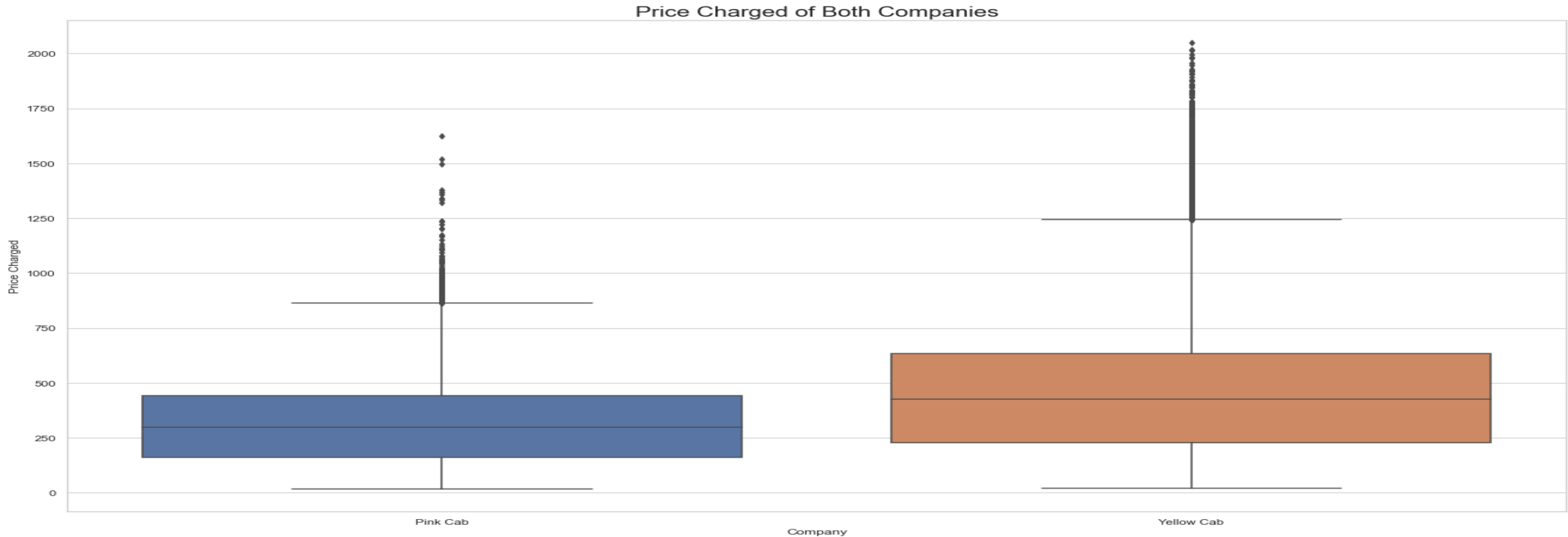


# Distribution of KM Travelled for both Cabs:

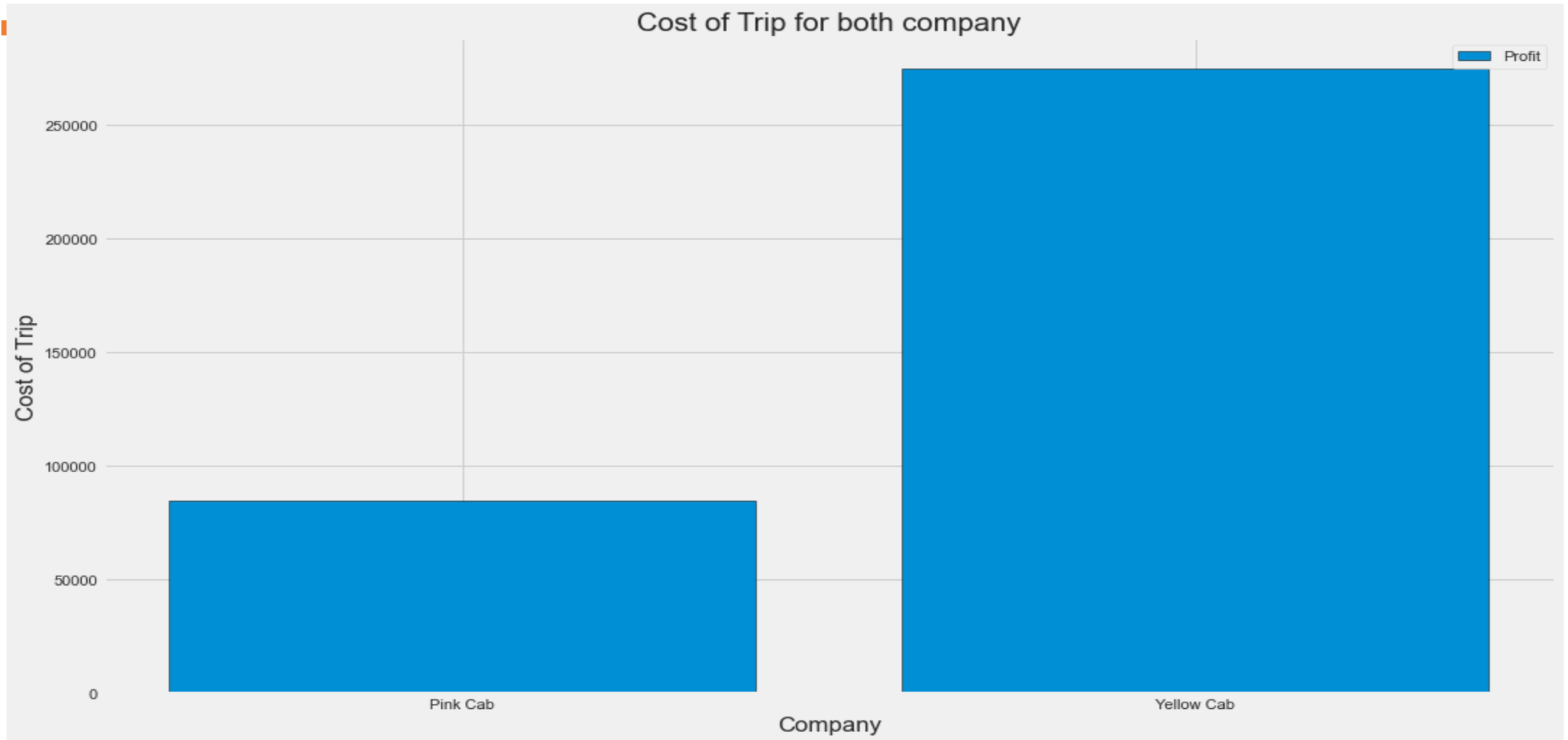


❑ From the above graphs, we can see that for both Pink and Yellow Cab most of the rides are in the range of approximately 2 to 48 KM.

# Distribution of Price Charged for both Cabs:

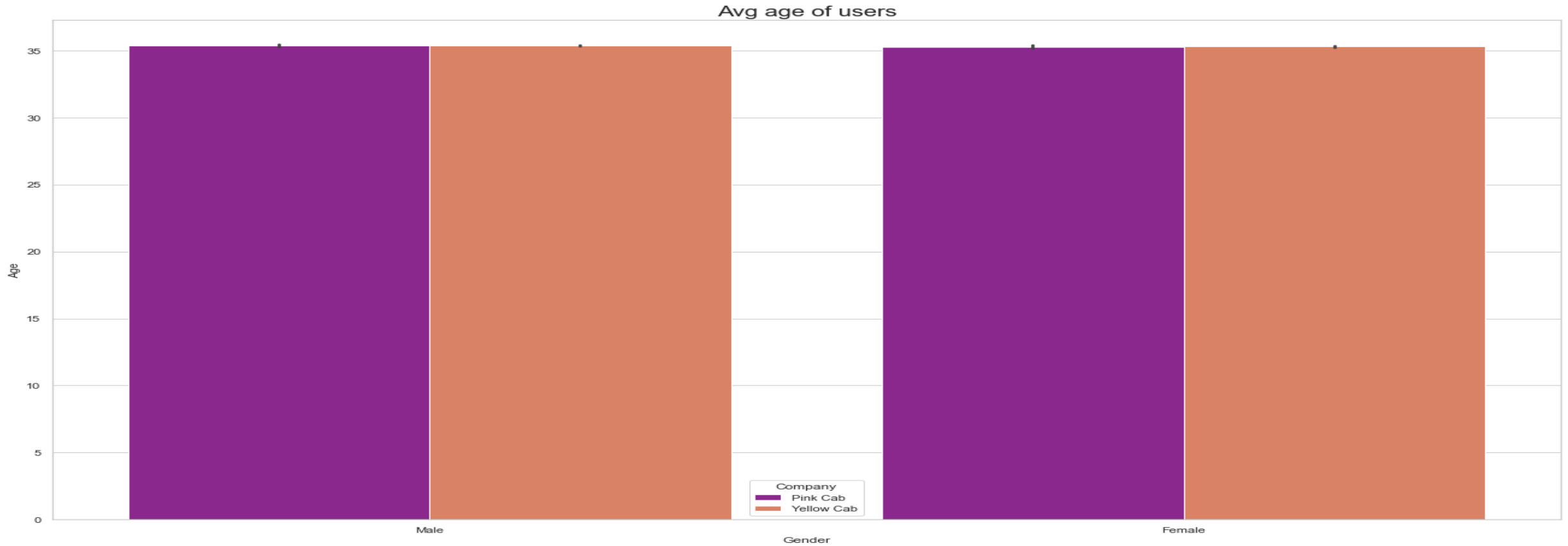


- ❑ The Price Charge range for Yellow cab is more than the Pink cab.
- ❑ The outliers are due to use of high-end cars.



**❑ Yellow Cab has higher cost of trip compared to Pink Cab .**

# Average Age per users:



❑ From the graph it shows 35 Average age of female and male who use cab service.

# Average Income of Users:



❑ From the graph it shows Average income is around \$15000 who use cab service.

# EDA Summary:

## Pink Cab

- Rides are in the range of approximately 2 to 48 KM.
- Price Charge range from 150 to 450 dollars.
- The pink has 23.6% of users.
- The maximum cost of Trip around 250 dollars.

## Yellow Cab

- Rides are in the range of approximately 2 to 48 KM.
- Price Charge range from 250 to 600 dollars.
- The pink has 76.4% of users.
- The maximum cost of Trip around 300dollars.

# Hypothesis Testing

# Hypothesis 1: There is no difference gender in pink cab

- Pink cab: We accept null hypothesis that there is no difference regarding gender for pink cab
- **Pink Cab Result:**
- 47231 37480
- P value is 0.11515305900425798
- We accept null hypothesis that there is no difference regarding gender for Pink Cab



## Hypothesis 2: There is difference gender in yellow cab

- Yellow Cab : We accept alternative hypothesis that there is a difference regarding gender for yellow cab
- **Yellow cab Result:**
- 158681 116000
- P value is  $6.060473042494144e-25$
- We accept alternative hypothesis that there is a difference regarding gender for Yellow Cab

## Hypothesis 3: There is no difference regarding age for pink cab

- Pink cab: We accept null hypothesis that there is no difference regarding age for Pink Cab
- **Pink Cab:**
- 75629 9894
- P value is 0.6363803611221691
- We accept null hypothesis that there is no difference regarding age for Pink Cab

## Hypothesis 4: There is difference regarding age for yellow Cab

- Yellow Cab: We accept alternative hypothesis that there is a difference regarding age for Yellow Cab
- **Yellow cab:**
- 245687 31624
- P value is  $1.0156079682838574e-07$
- We accept alternative hypothesis that there is a difference regarding age for Yellow Cab

## Hypothesis 5: There is no difference in payment mode for Pink Cab

- Pink Cab: We accept null hypothesis that there is no difference in payment mode for Pink Cab
- **Pink Cab:**
- P value is 0.7900465828793288
- We accept null hypothesis that there is no difference in payment mode for Pink Cab

## Hypothesis 6: There is no difference in payment mode for yellow Cab

- Yellow Cab: We accept null hypothesis that there is no difference in payment mode for Yellow cab
- **Yellow Cab Result:**
- P value is 0.2933060638298729
- We accept null hypothesis that there is no difference in payment mode for Yellow Cab

# Thank You



**Data Glacier**

Your Deep Learning Partner