Exploratory Data Analysis

G2M – Cab Investment Firm Project

20 April 2023

Executive Summary

The goal of this project is to help the private firm XYZ make an informed investment decision in the cab industry by analyzing multiple data sets related to two cab companies.

This presentation summarizes the analysis and recommendations on which company is performing better and is a better investment opportunity for XYZ.

Problem Statement

- XYZ is a private equity firm in the US. Due to remarkable growth in the Cab Industry in the last few years and multiple key players in the market, it is planning for an investment in the Cab industry.
- **Objective**: Provide actionable insights to help XYZ firm in identifying the right company for making an investment.

The analysis has been divided into four parts:

- Data Understanding and Visualization
- Finding the most users Cab company
- Finding the cheapest Cab company for users
- Finding the most profitable Cab company
- Multiple Hypothesis and Investigate

Datasets

- **Cab_Data.csv**
  This file includes details of transactions for 2 cab companies
- **Customer_ID.csv**
  This is a mapping table that contains a unique identifier that 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 a list of US cities, their population, and the number of cab users

Approach

This project mainly consists of three parts:

1. **Cleaning & Merging the Data**
   In this part, duplicate columns and rows are removed and a master dataset is created.

2. **Exploratory Data Analysis & Visualization**
   In the second part, a general analysis and visualization of the master data is done through EDA to notice possible correlations and generate some hypotheses.

3. **Hypothesis Testing**
   In the last part, the hypotheses that were generated are tested and visualized

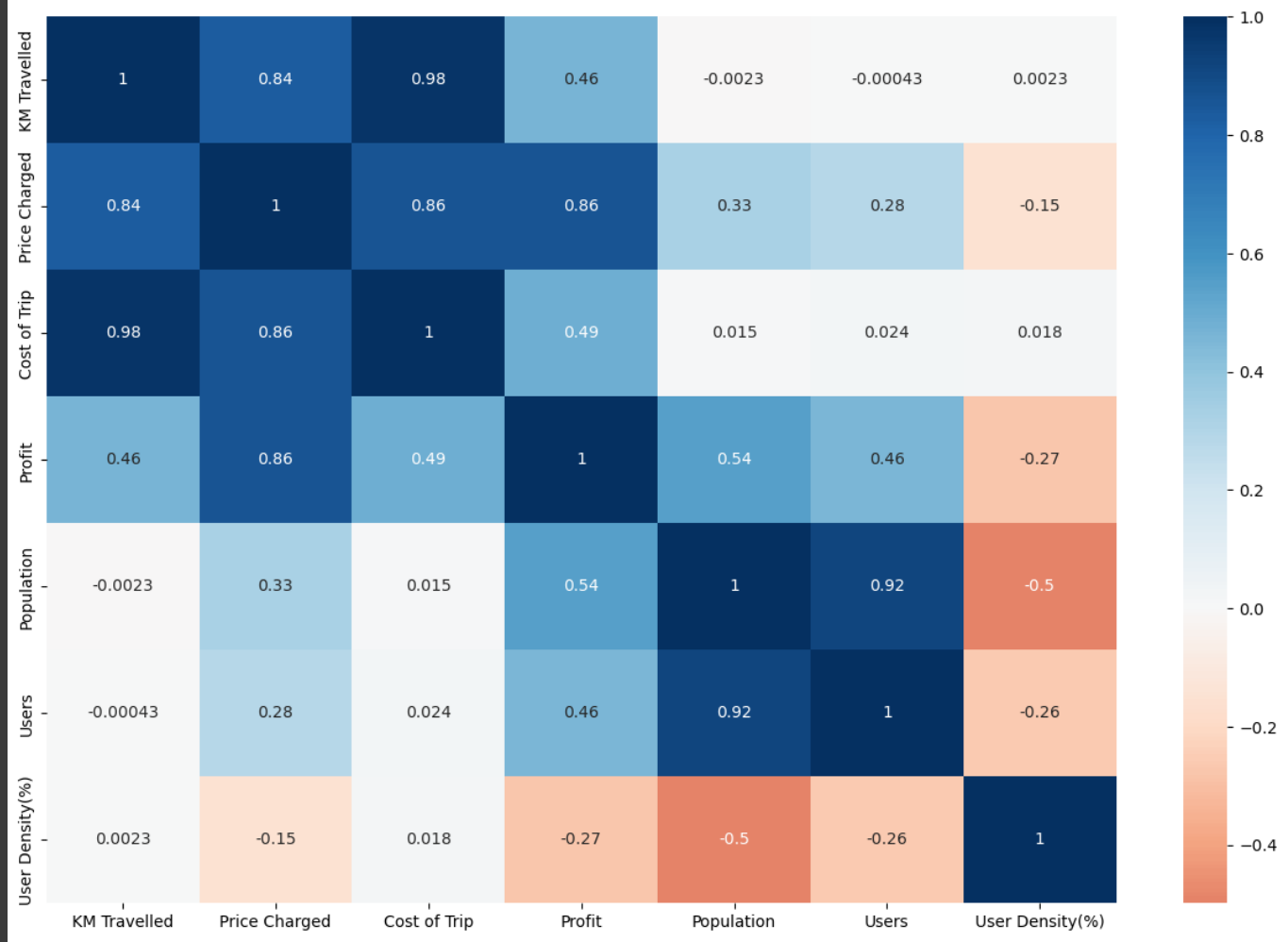
1. Cleaning and Merging

Changes Done

- Travel Date column is changed into actual date values from time periods.
- Profit column is added.
- Some columns' types are corrected.
- User Density column is added.
- Datasets are merged.
- Master data's columns are sorted.
- Master data is exported.
- Two separate files for the two companies are created.
- Company files are exported.

2.EDA

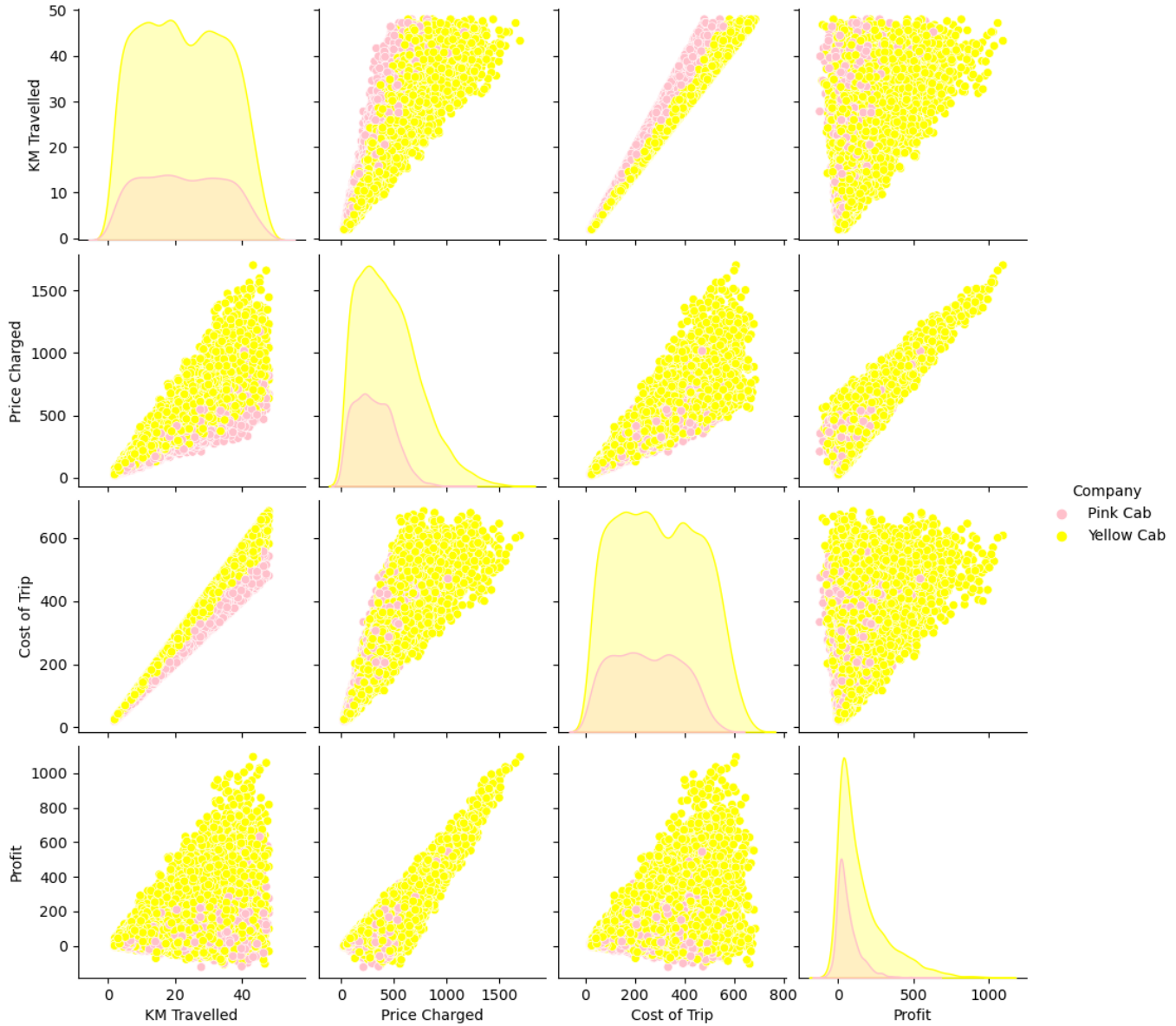
Correlation Matrix



Since trip cost and Km travelled are highly positive correlated, their relation doesn't give useful information.

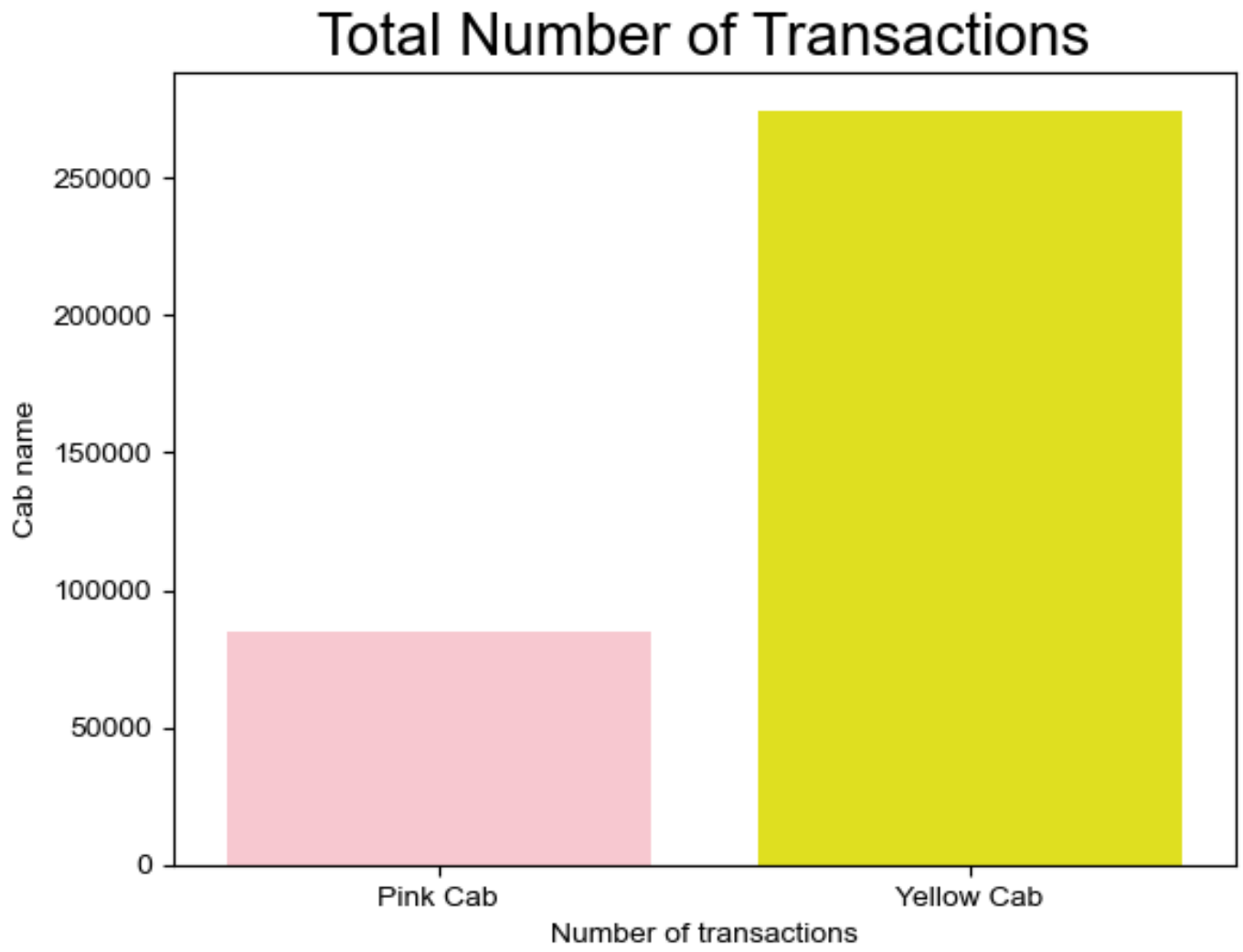
Data Glacier
Your Deep Learning Partner

Pairplot

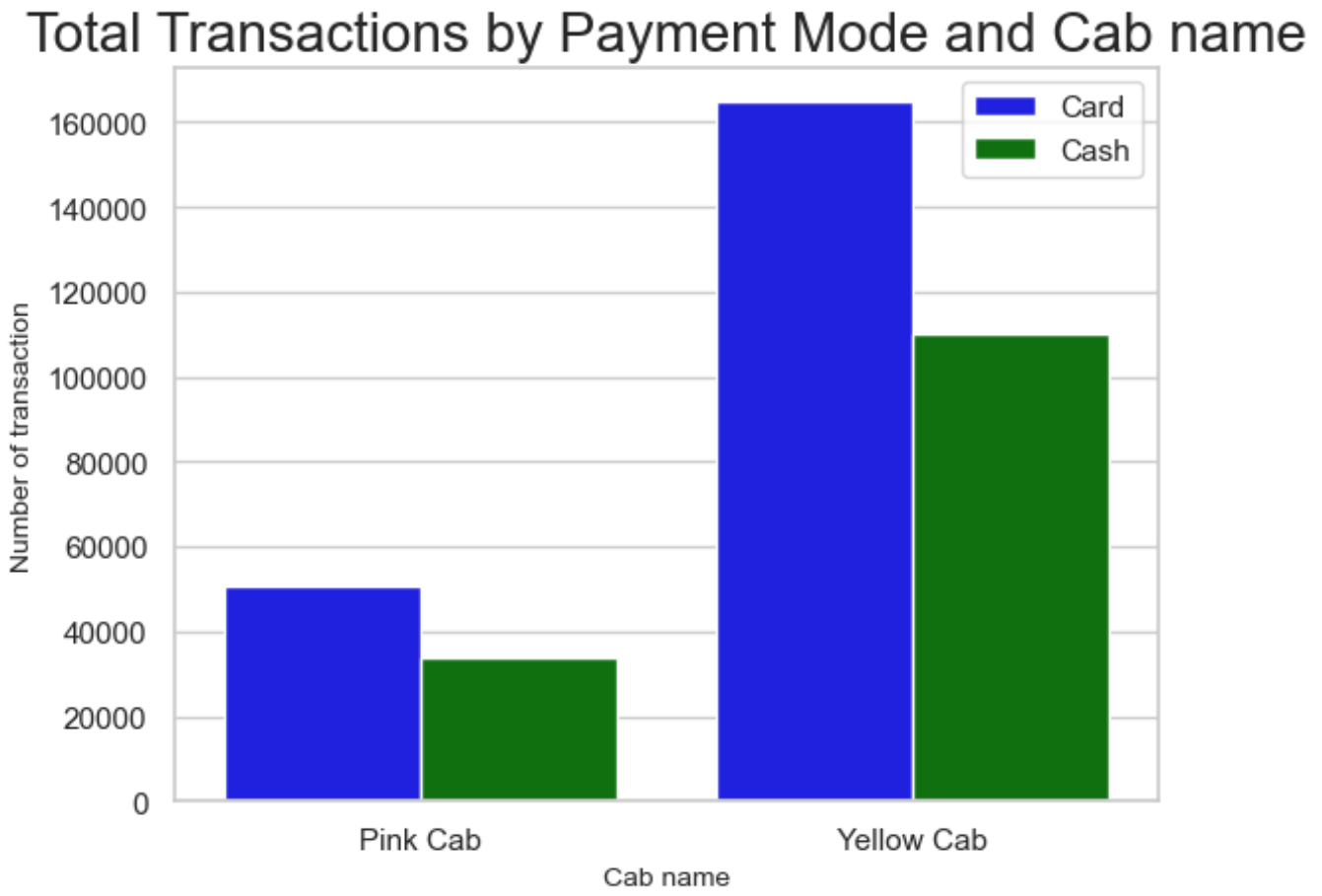


Data Glacier
Your Deep Learning Partner

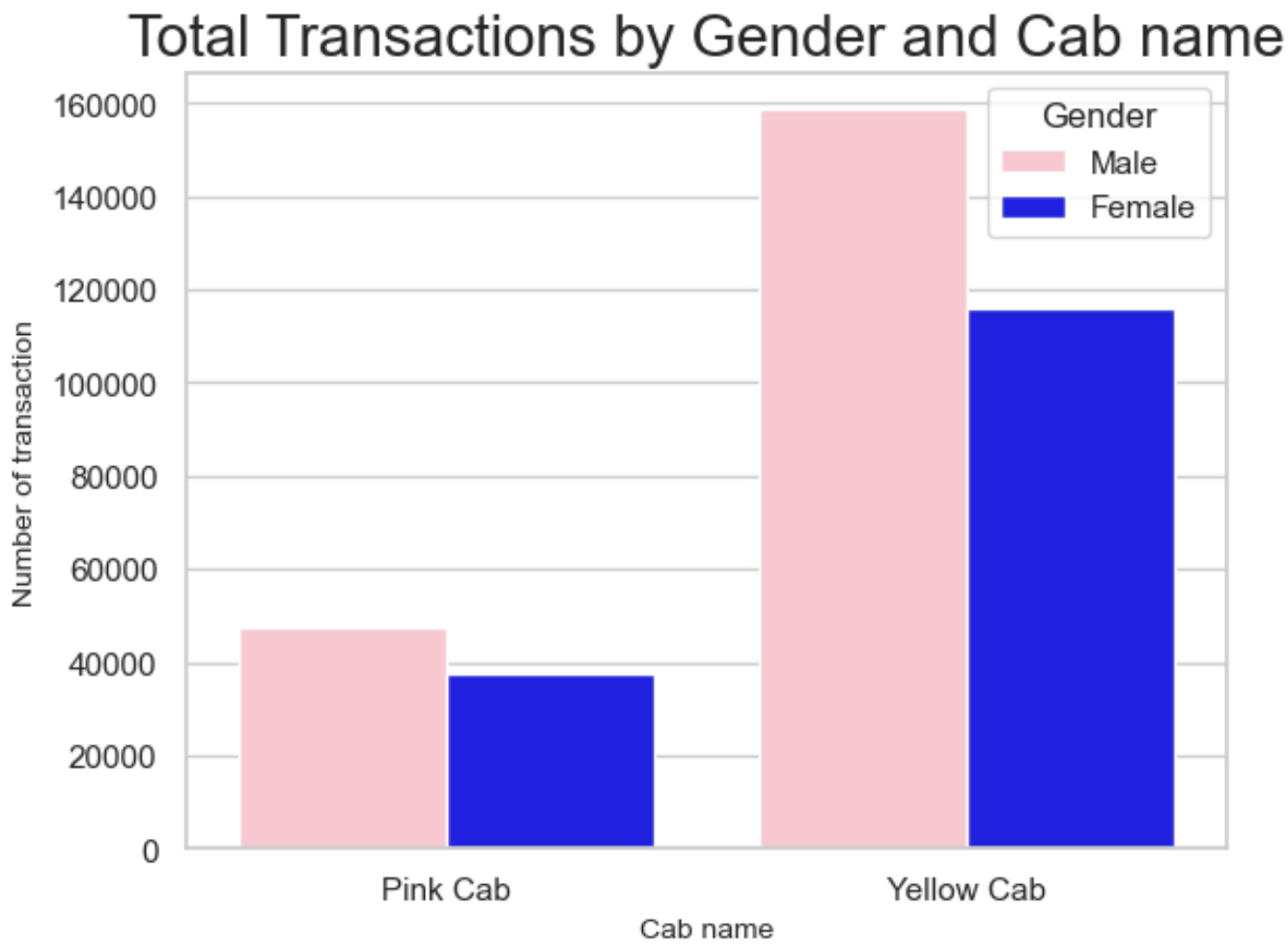
Transactions



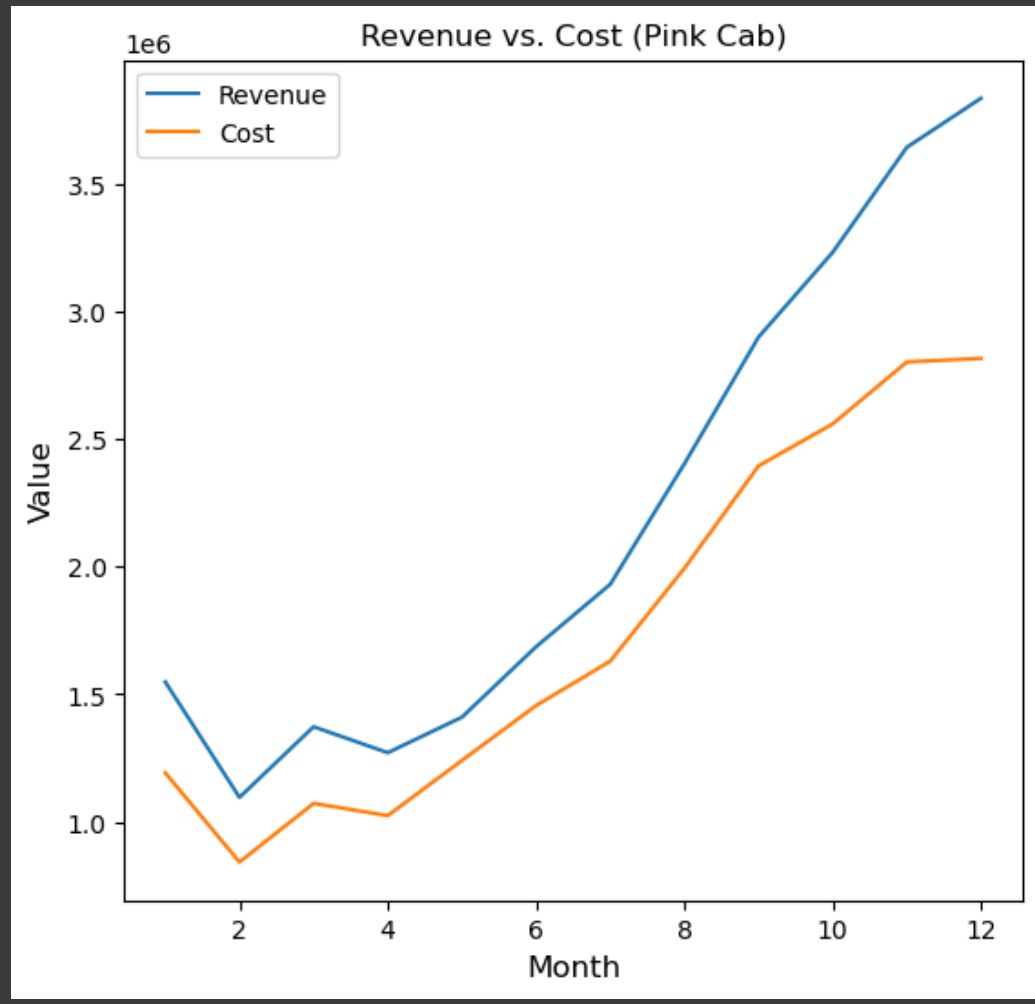
Transactions by Payment Mode

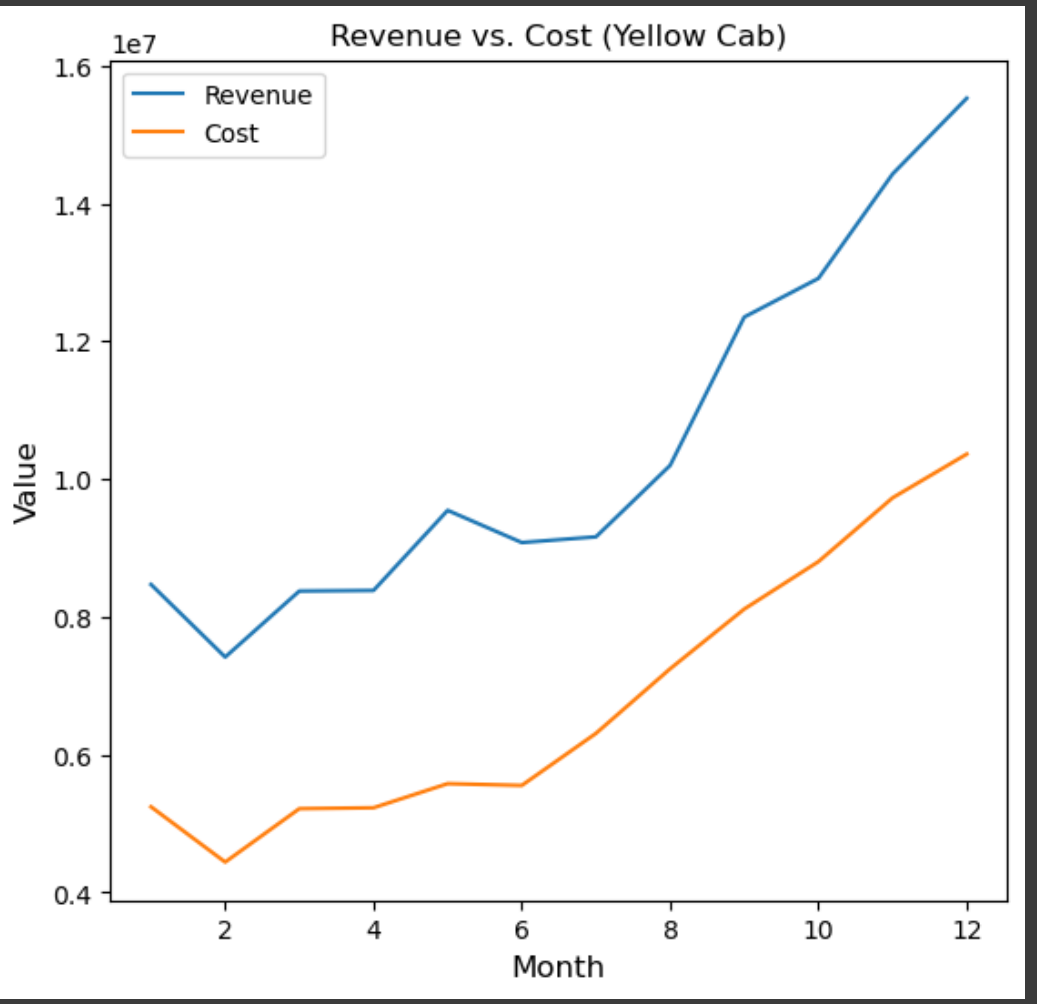


Transactions by Gender



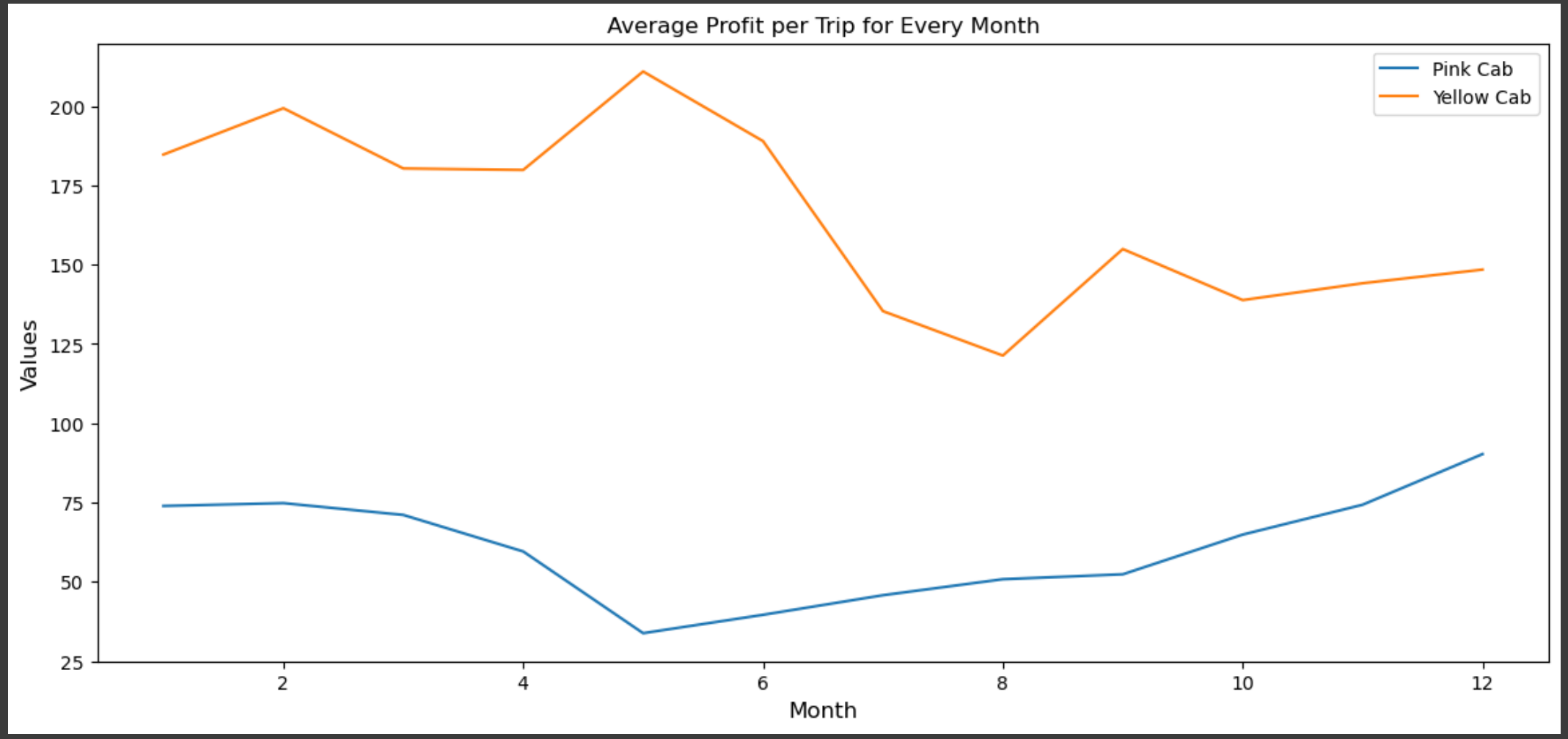
Revenue vs. Cost



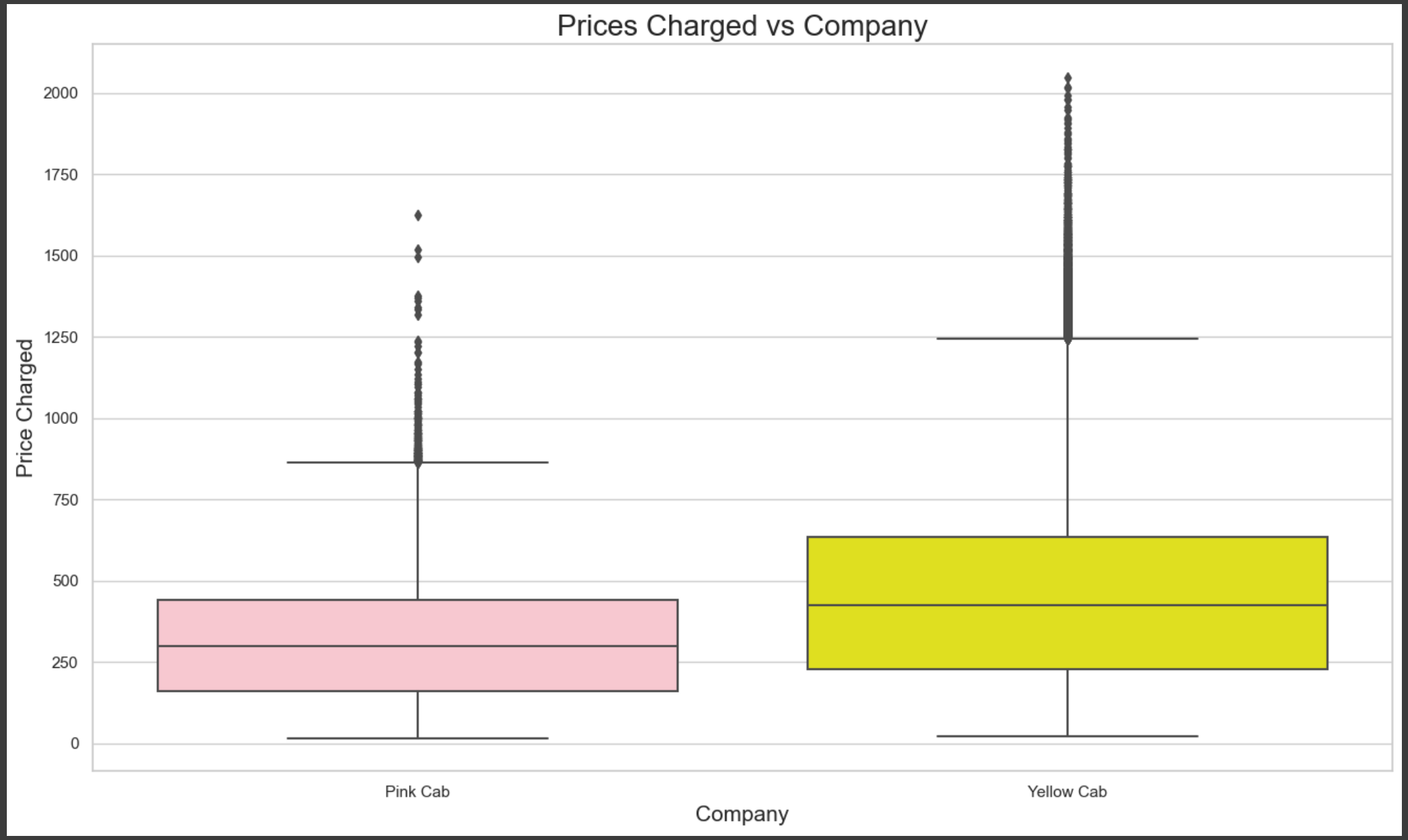


Data Glacier
Your Deep Learning Partner

Profit per Trip



Prices Charged



EDA Summary

- Yellow cab seems to have more transactions but Pink Cab’s profit margin is increasing faster.
- Yellow Cab has a higher percentage of female customers.
- Yellow Cab has a higher percentage of customers who pay with credit card.
- Yellow Cab generally charges more.

3. Hypothesis Testing

Hypothesis 1

H0 : Yellow Cab has a higher proportion of customers who use cash payment modes compared to Pink Cab.

H1 : Yellow Cab doesn't have a higher proportion of customers who use cash payment modes compared to Pink Cab.

```
h1p1 = pink_cab[pink_cab.Payment_Mode=="Cash"].Payment_Mode.count()
h1p2 = pink_cab.Payment_Mode.count()
h1p = h1p1/h1p2

h1y1 = yellow_cab[yellow_cab.Payment_Mode=="Cash"].Payment_Mode.count()
h1y2 = yellow_cab.Payment_Mode.count()
h1y = h1y1/h1y2

if h1y>h1p:
    print("We accept the null hypothesis (H0) that Yellow Cab has a",
          "higher proportion of customers who use cash payment modes",
          "compared to Pink Cab.")
else:
    print("We accept the alternative hypothesis (H1) that Yellow Cab",
          "doesn't have a higher proportion of customers who use cash",
          "payment modes compared to Pink Cab.")
```

```
We accept the alternative hypothesis (H1) that Yellow Cab doesn't have a higher proportion of customers who use cash payment modes compared to Pink Cab.
```

Pink cab has a higher proportion of customers who use cash payment modes compared to Yellow Cab.

Hypothesis 2

H0 : Pink Cab customers travel longer distances in average than Yellow Cab.

H1 : Pink Cab customers don't travel longer distances in average than Yellow Cab.

```
h2p = pink_cab["KM Travelled"].sum() / pink_cab["Transaction ID"].count()

h2y = yellow_cab["KM Travelled"].sum() / yellow_cab["Transaction ID"].count()

if h2p>h2y:
    print("We accept the null hypothesis (H0) that Pink Cab customers",
          "travel longer distances in average than Yellow Cab.")
else:
    print("We accept the alternative hypothesis (H1) that Pink Cab customers",
    "don't travel longer distances in average than Yellow Cab.")
```

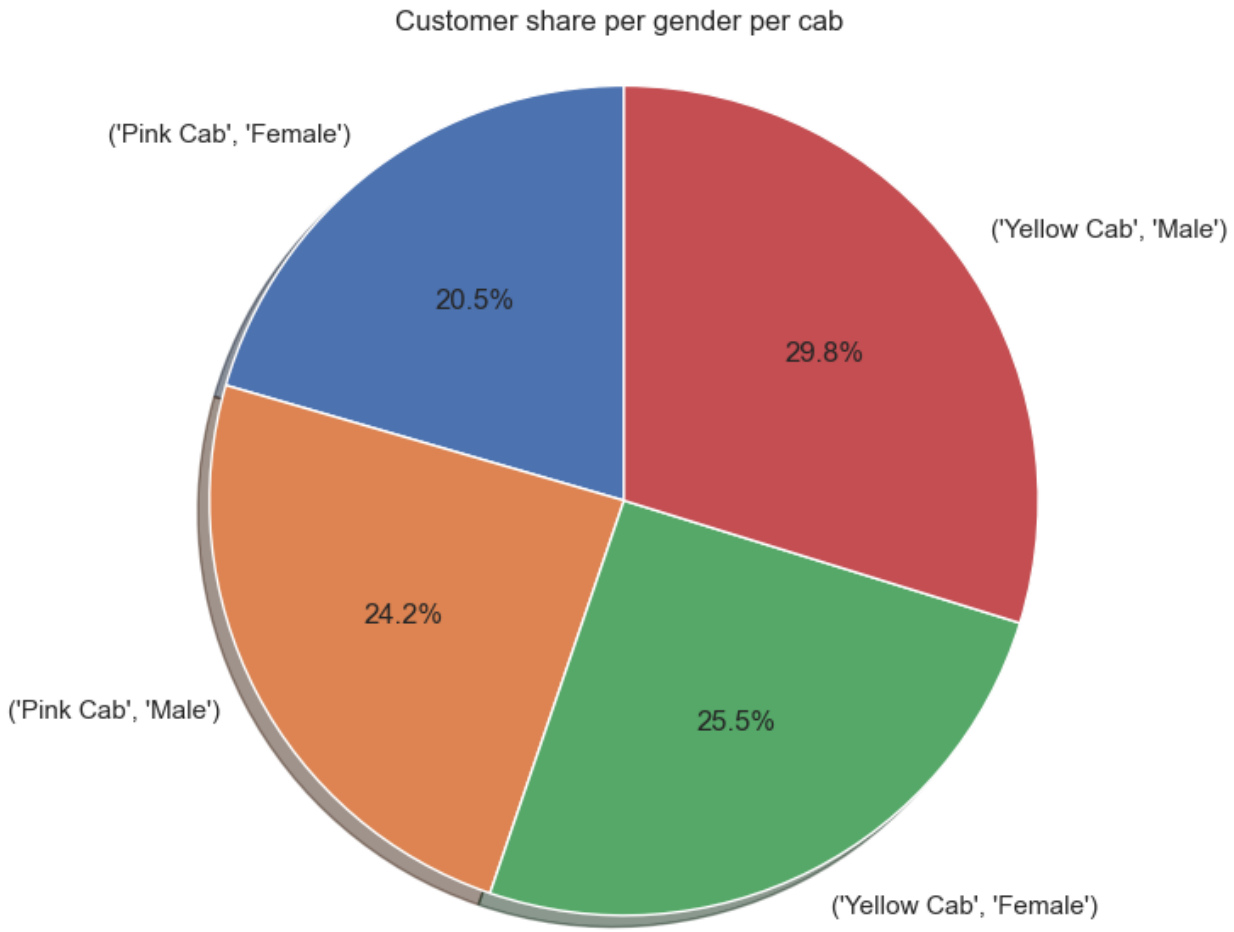
```
We accept the alternative hypothesis (H1) that Pink Cab customers don't travel
longer distances in average than to Yellow Cab.
```

Yellow Cab customers travel longer distances in average than to Pink Cab.

Hypothesis 3

H0 : Yellow Cab has higher number of female customers than Pink Cab.

H1 : Yellow Cab doesn't have higher number of female customers than Pink Cab.



We accept the null hypothesis (H0) that Yellow Cab has higher number of female customers than Pink Cab.

Yellow Cab has higher number of female customers than Pink Cab.

Data Glacier
Your Deep Learning Partner

Hypothesis 4

H0 : Yellow Cab generally has older customers.

H1 : Yellow Cab generally doesn't have older customers.

```
h4p = pink_cab.Age.mean()
h4y = yellow_cab.Age.mean()
print("Average customer age of Pink Cab:", h4p)
print("Average customer age of Yellow Cab:", h4y)

if h4y>h4p:
    print("We accept the null hypothesis (H0) that",
          "Yellow Cab generally has older customers.")
else:
    print("We accept the alternative hypothesis (H1) that",
          "Yellow Cab generally doesn't have older customers.")
```

```
Average customer age of Pink Cab: 35.322413854162974
Average customer age of Yellow Cab: 35.34111205361856
We accept the null hypothesis (H0) that Yellow Cab generally has
older customers.
```

Altough noticably insignificant, Yellow Cab generally has older customers.

Hypothesis 5

H0 : Price charged is correlated with the Income of the customer.

H1 : Price charged is not correlated with the Income of the customer.

```
h5 = master["Price Charged"].corr(master["Income (USD/Month)"])

if h5>0.5 or h5<-0.5:
    print("We accept the null hypothesis (H0) that Price",
          "charged is correlated with the Income of the customer.")
else:
    print("We accept the alternative hypothesis (H1) that Price",
          "charged is not correlated with the Income of the customer.")
```

```
We accept the alternative hypothesis (H1) that Price charged is not correlated
with the Income of the customer.
```

Price charged is not correlated with the Income of the customer.

From Ahmet Metin Zengin,

Thank You