Exploratory Data Analysis
<G2M Insight for Cab Investment Firm>

Date: 7/21/17
Name: Kyun Woo Hong

Agenda

Executive Summary

Problem Statement

Approach

EDA

EDA Summary

Recommendations

Data Glacier
Your Deep Learning Partner

Overview

- **The Client**: XYZ
  - The private firm in US.
  - Planning for an investment in Cab industry.

- **The Goal:** Present G2M strategy to the client.

- **Given Dataset:** 4 individual datasets from 1/31/2016 to 12/31/2018.
  - Cab_Data.csv
  - Customer_ID.csv
  - Transcation_ID.csv
  - City.csv

Overview

- **The Cab Companies**:
  - Yellow Cab
  - Pink Cab

- **The Analysis Including:**
  - Data Overview
  - EDA
  - Data Visualization
  - Answering multiple hypothesis

Data Overview

- **Cab_Data.csv**
  - Include detailed transactions for 2 cab companies.

- **Customer_ID.csv**
  - Map unique customer ID to customer's demographic detail

- **Transaction_ID.csv**
  - Map unique transaction ID to customer and payment type

- **City.csv**
  - Include lists of US cities with population and numbers of users.

Tabular data details: Cab_Data.csv

Total number of observations	359391
Total number of files	1
Total number of features	7
Base format of the file	.csv
Size of the data	20.2 MB

Tabular data details: Customer_ID.csv

Total number of observations	49171
Total number of files	1
Total number of features	4
Base format of the file	.csv
Size of the data	1 MB

Tabular data details: Transaction_ID.csv

Total number of observations	440098
Total number of files	1
Total number of features	3
Base format of the file	.csv
Size of the data	8.58 MB

Tabular data details: City.csv

| Total number of observations	20
Total number of files	1
Total number of features	3
Base format of the file	.csv
Size of the data	759 Bytes

EDA

- ## Profits : Yellow Cab vs Pink Cab
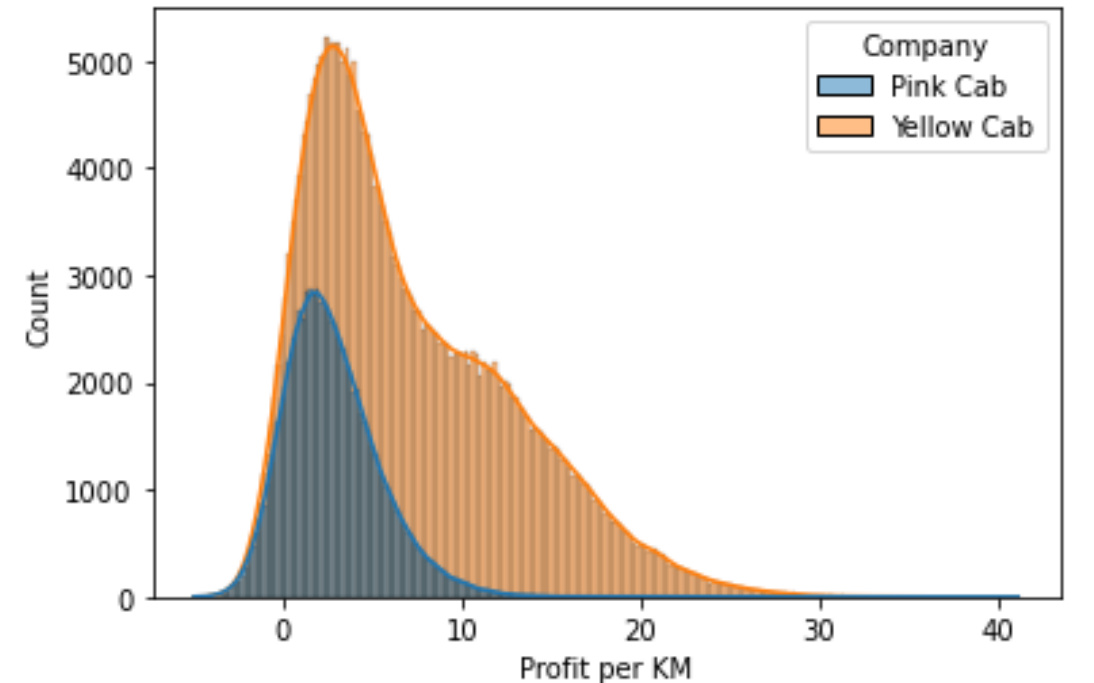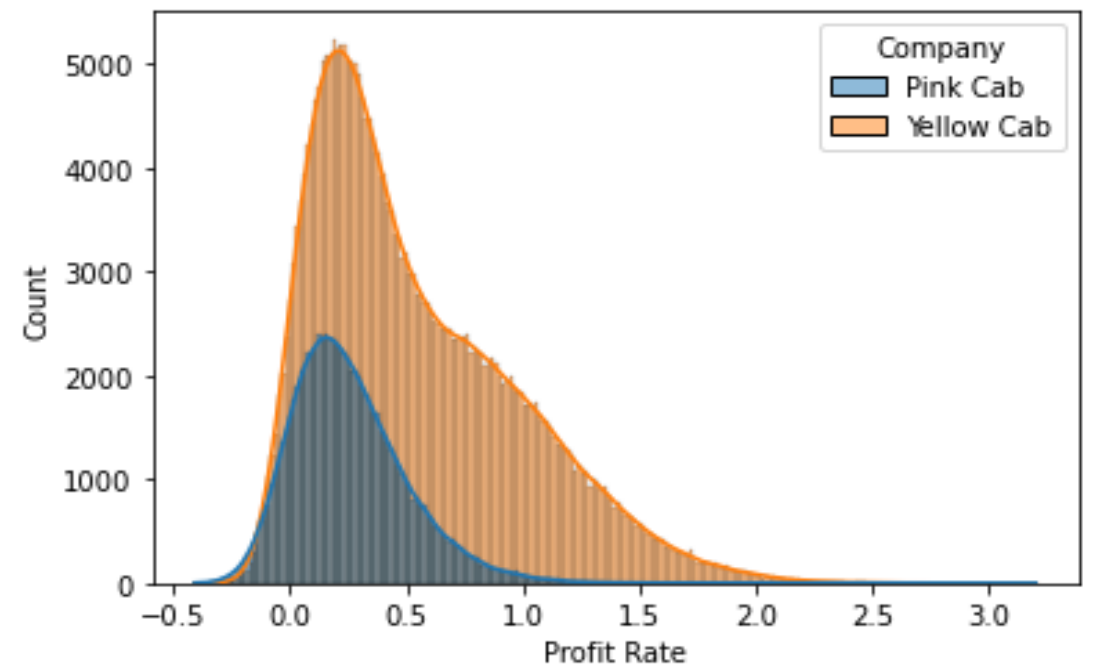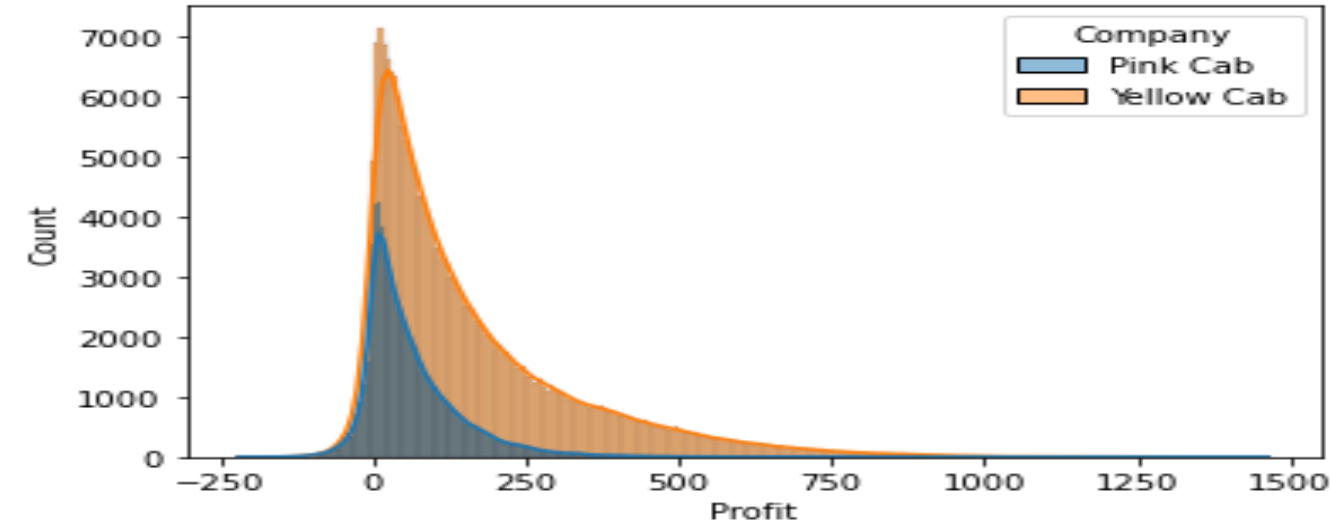
From above plot, Yellow Cab has more profit than Pink Cab. I also look at profit rate and profit per KM and both plots also show that Yellow Cab has more profit rate and profit per KM than Pink Cab. Interesting is that all three plots are right skewed.

EDA
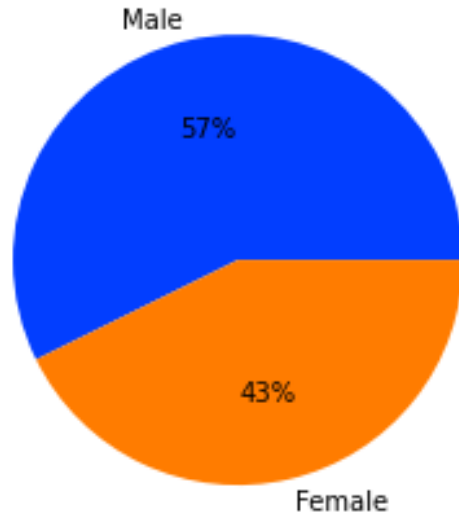
- **Monthly and Yearly Profits : Yellow Cab vs Pink Cab**
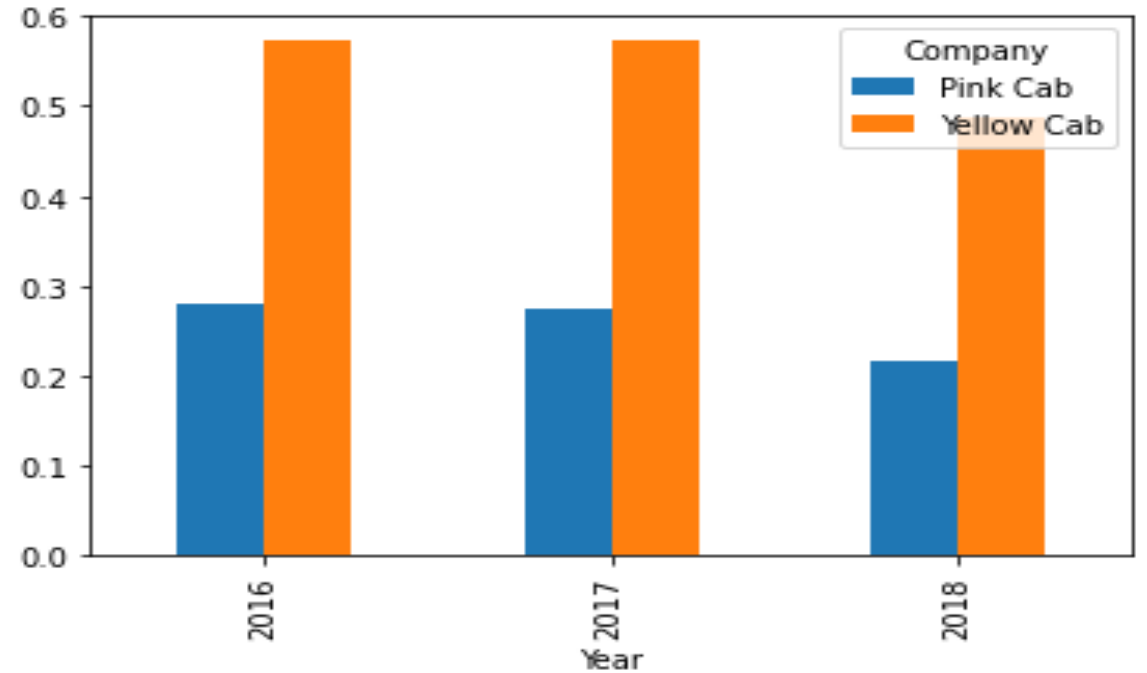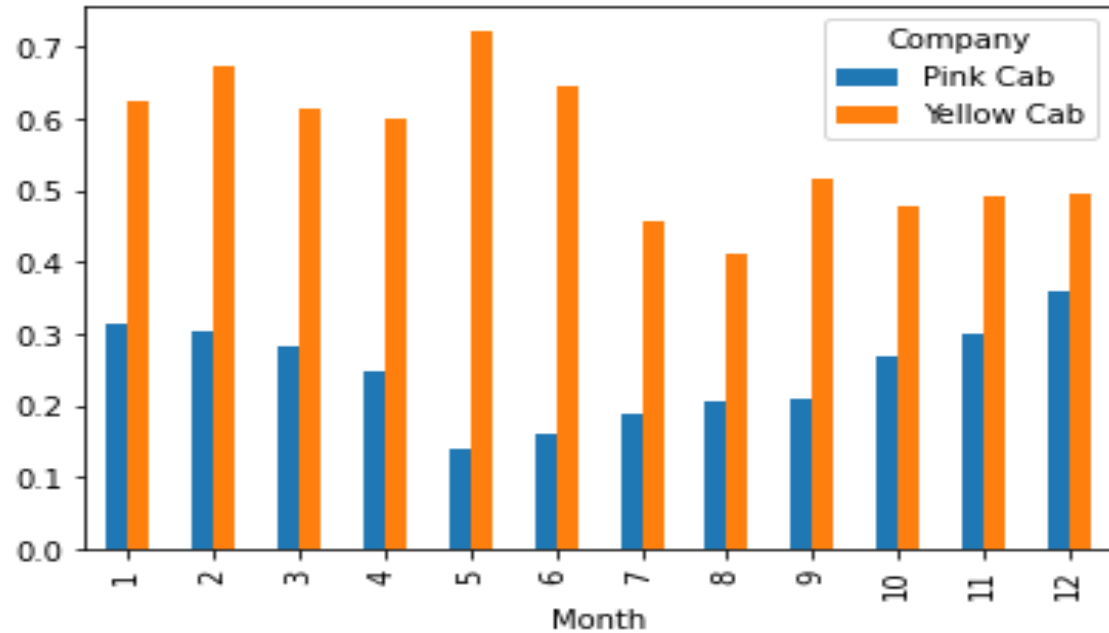
I also look at monthly and yearly profit rate for each company. Profit rates in 2016 and 2017 do not change but profit rates drop in 2018. Yellow Cab has the most profit rates in May but Pink Cab has the most profit in December.
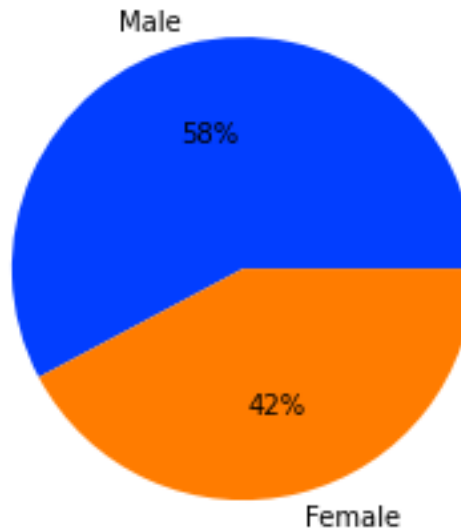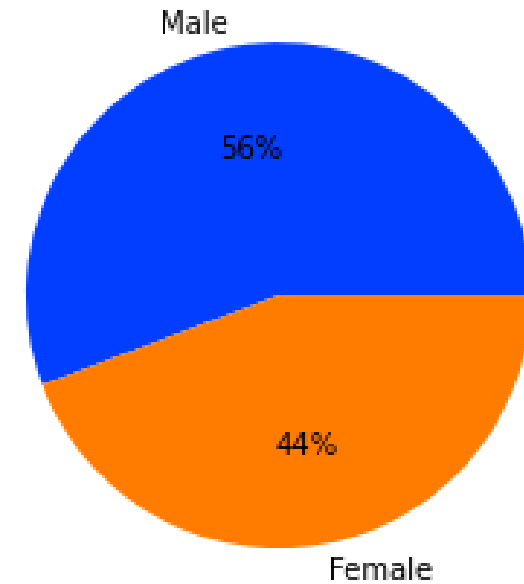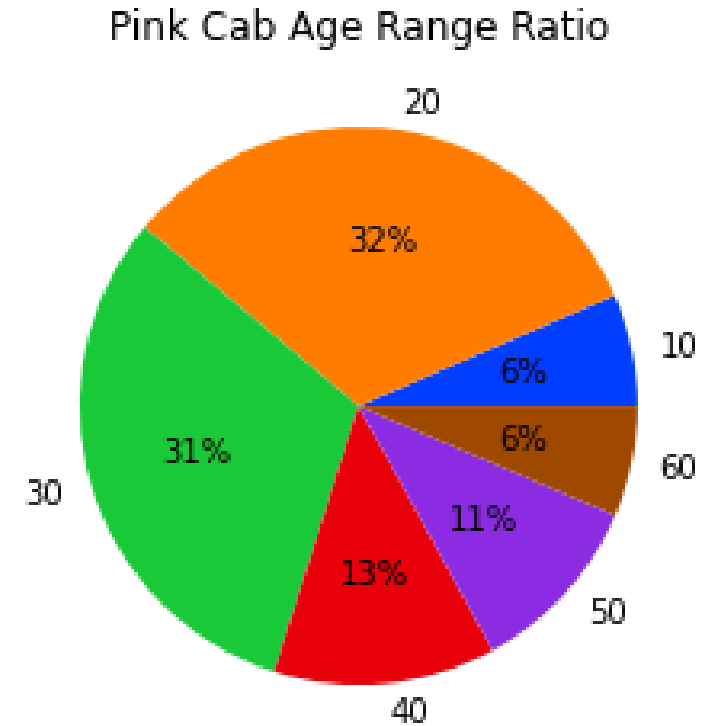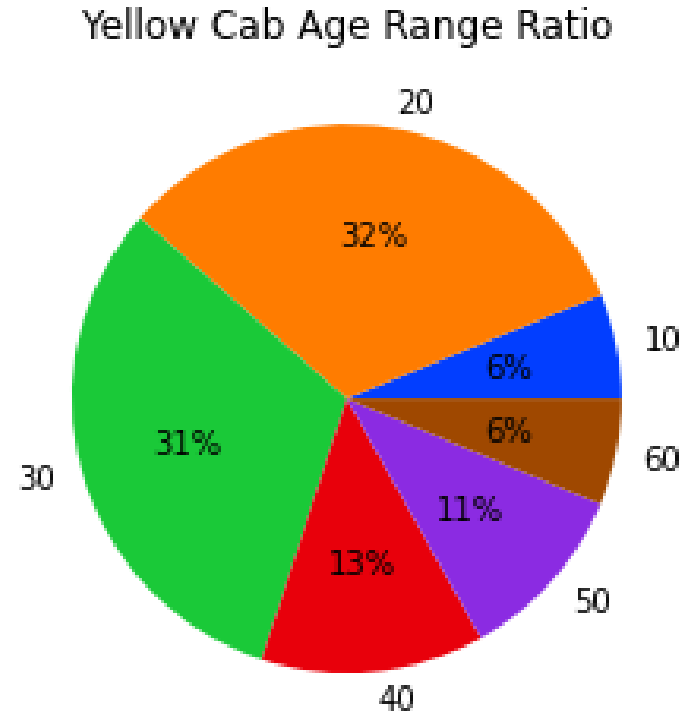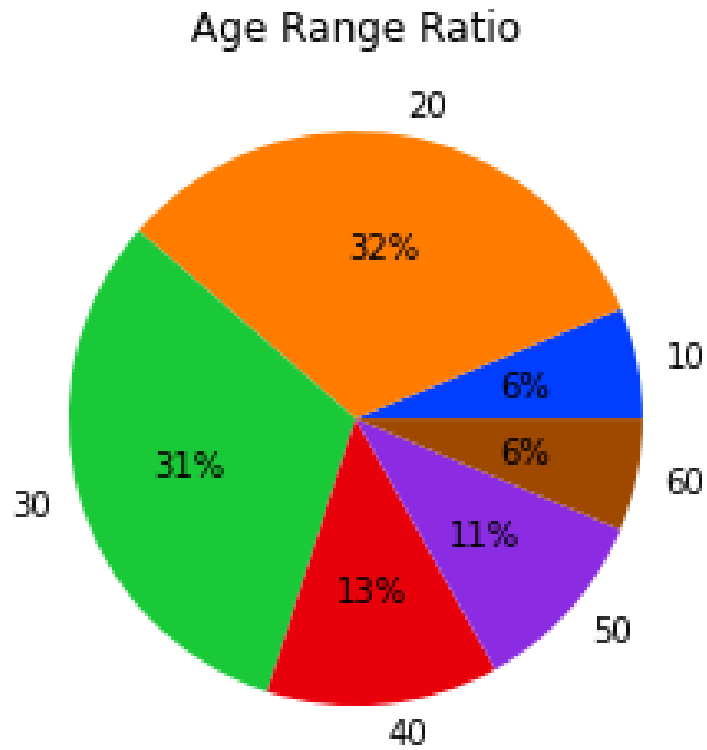
EDA

- ## Customer Summary: Gender

Gender ratio is similar to both cap companies. Male is about 57% and Female is about 43%.
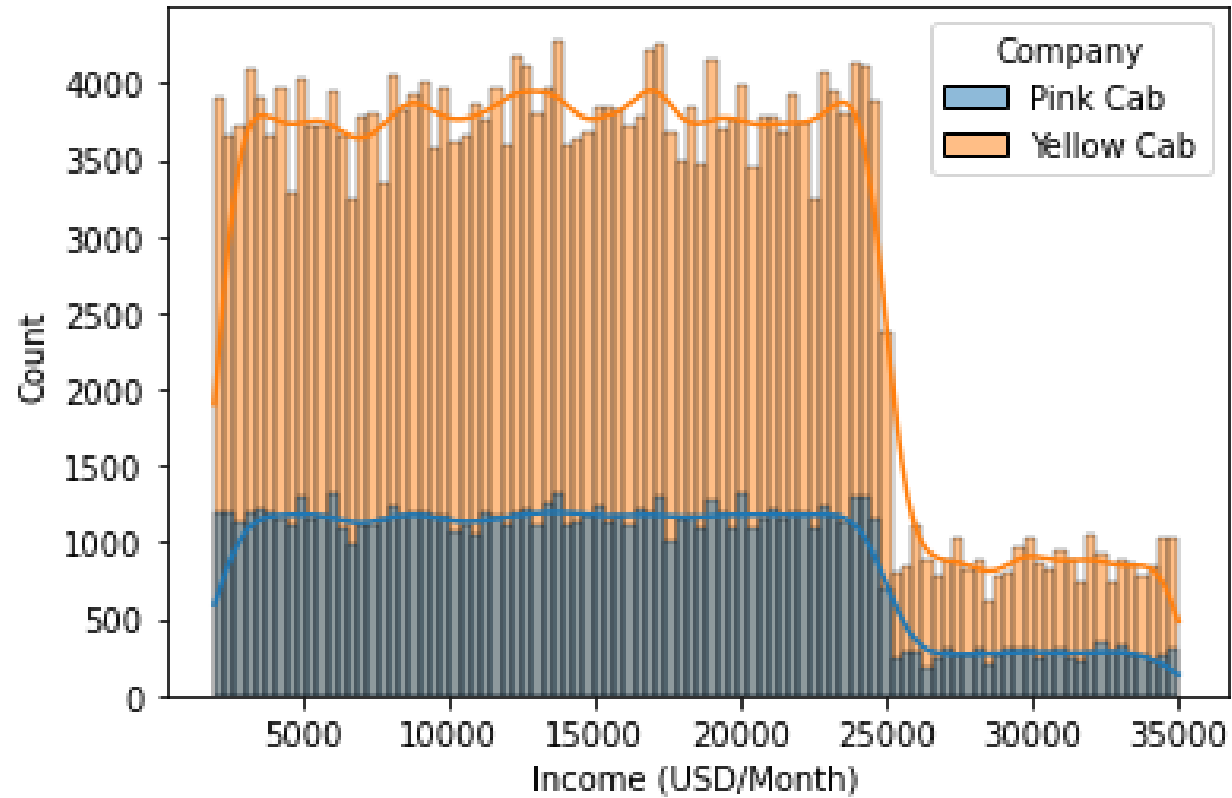
EDA

- **Customer Summary: Age**

Age ratio is exactly same for both cab companies. No difference found.

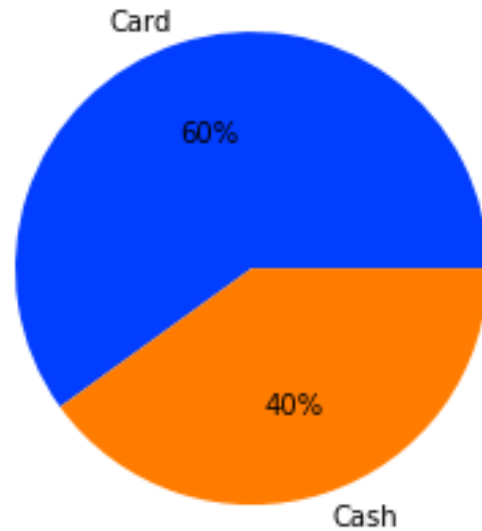# EDA

- **Customer Summary: Income**

I do not find any significant difference on customer's income ratio between two cab companies
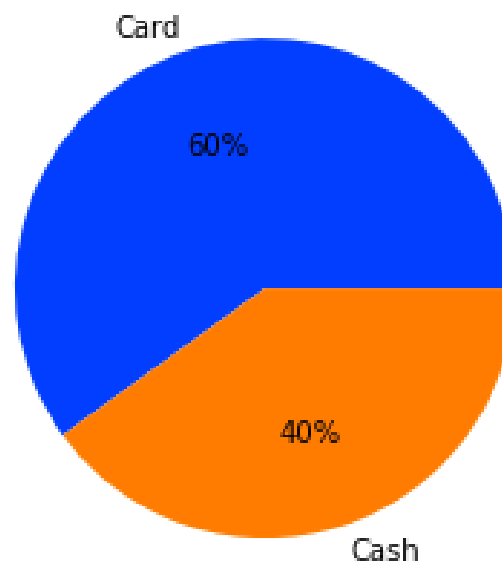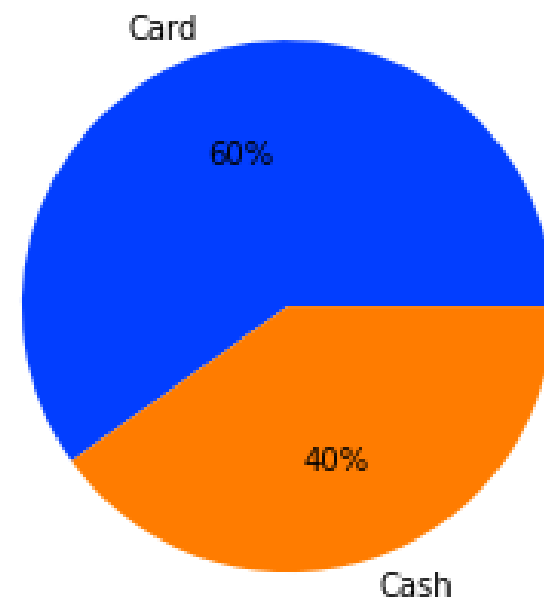
EDA

- ## Customer Summary: Payment Type

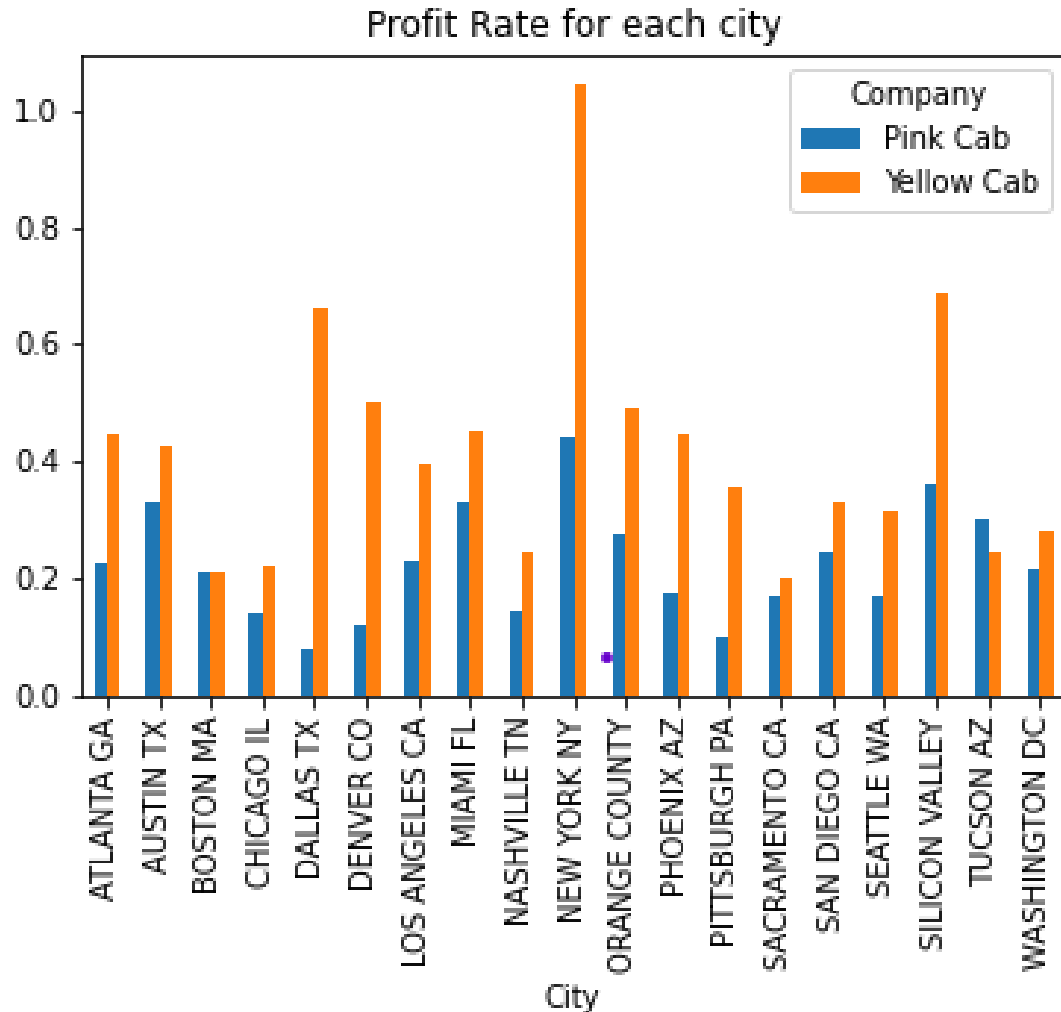| Payment Type Ratio (Combined two cab companies)	Yellow Cap Payment Type Ratio	Pink Cap Payment Type Ratio

Card — 60%
Cash — 40%

There is no difference between payment type ratio on both cab companies.

EDA

- **City Summary: Profit Rate for Each City**


Profit Rate for each city

From the profit rate per city plot, NYC has the highest profit rate for both Pink and Yellow Cabs and Silicon Valley is the next highest profit rate for both cab companies. However, I can find a few cities that are interested in: Dallas, Denver, Pittsburgh, Tucson, and D.C. Yellow cab has such a high profit rate on three cities: Dallas, Denver, and Pittsburgh. Tucson is only city Pink Cab has higher profit rate than Yellow Cab.

EDA

- **Holiday Effect : Holiday - Profit Correlation**
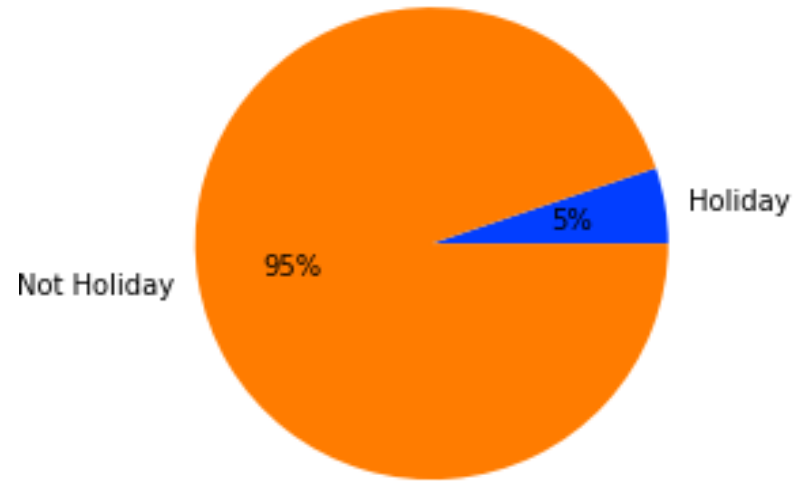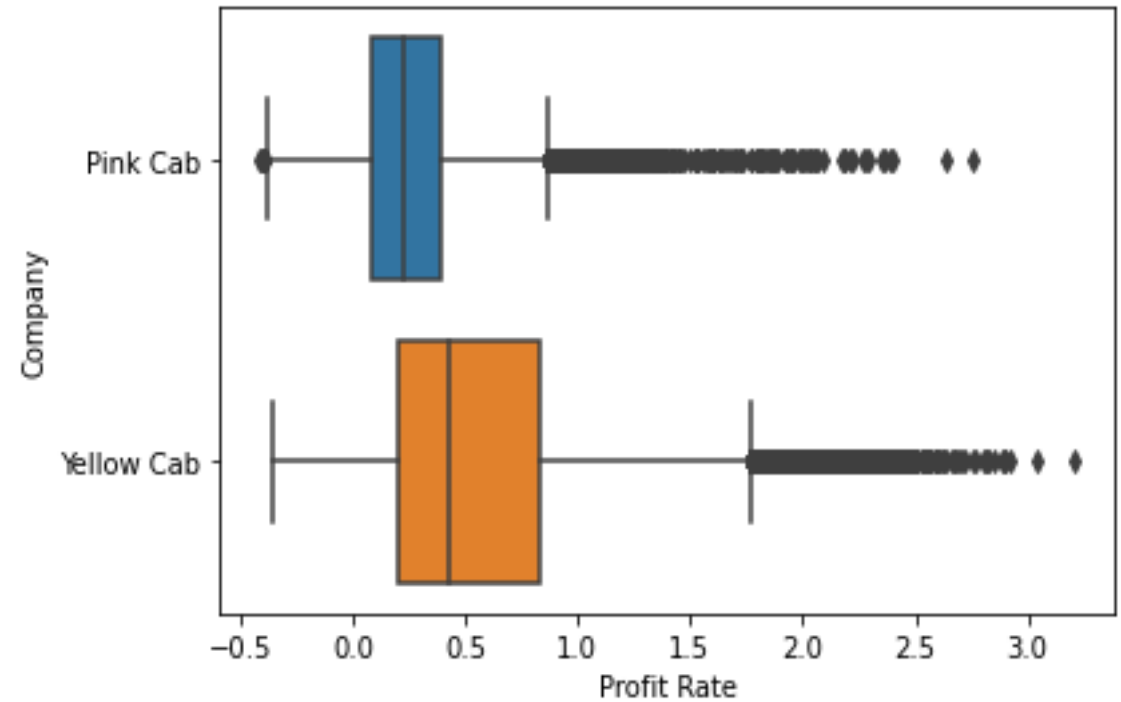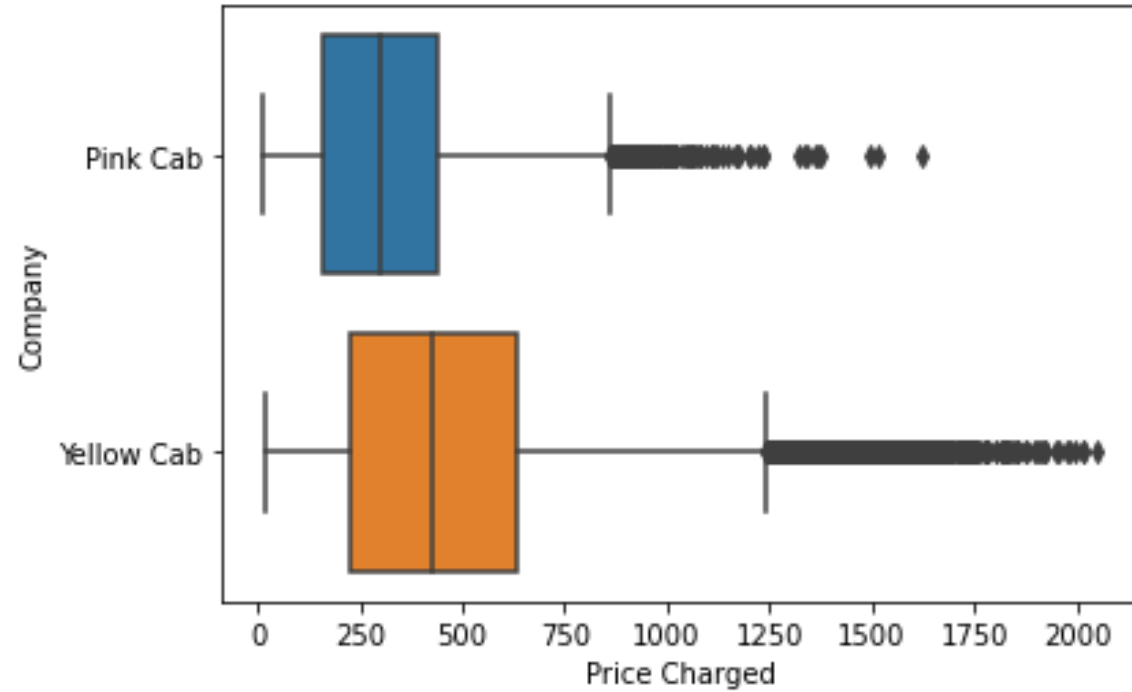


```
holiday  Company
0        Pink Cab        0.257055
         Yellow Cab      0.541681
1        Pink Cab        0.226796
         Yellow Cab      0.565926
Name: Profit Rate, dtype: float64
```

There is an interesting fact that Pink Cab actually has lower profit rate on Holiday unlike Yellow Cab.
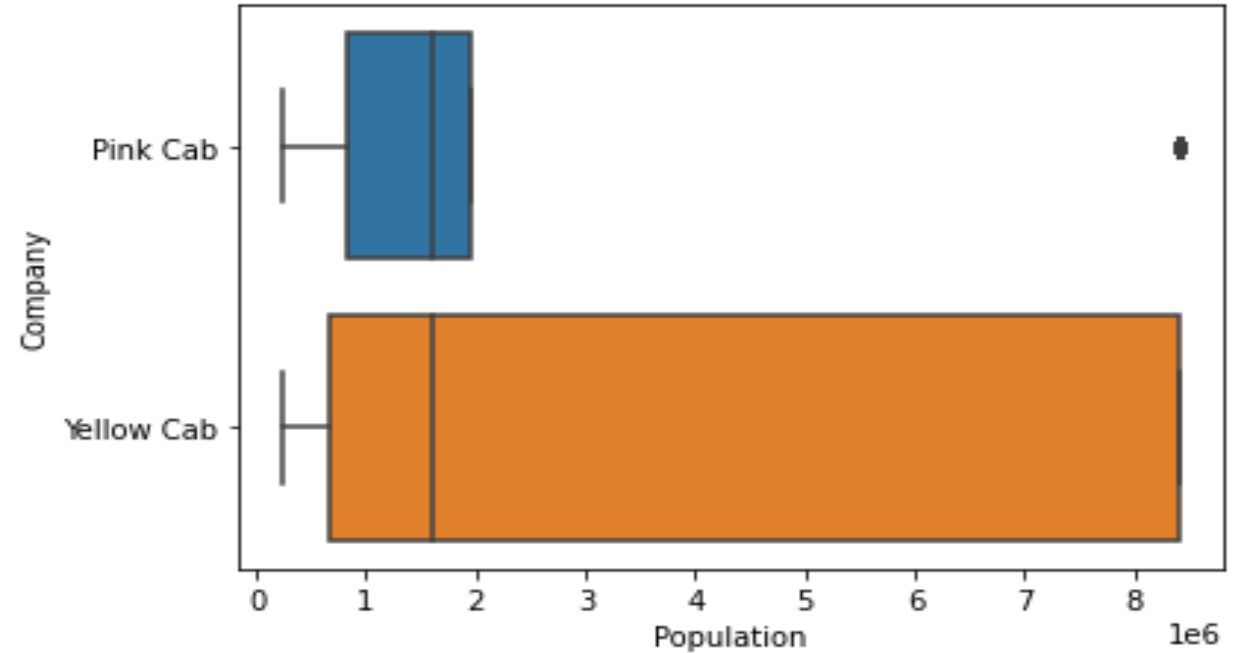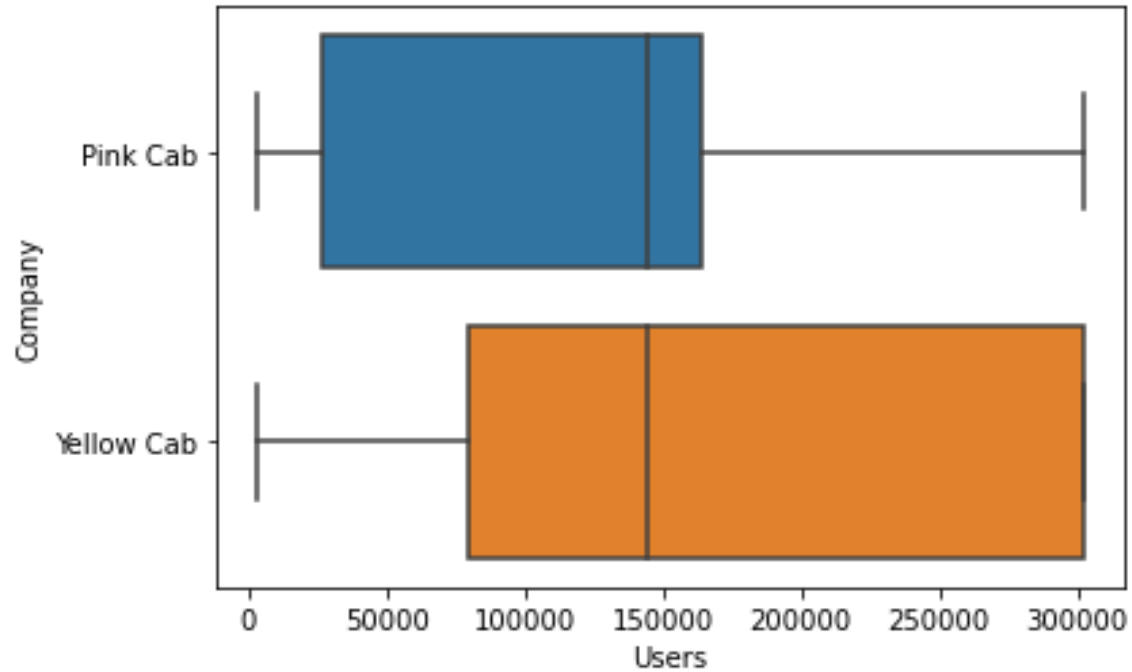
EDA Summary

- **Price Charged and Profit Rate**

For Price Charged and Profit Rate, Yellow Cab has higher median value and higher proportion with higher price which cause higher profit rate.
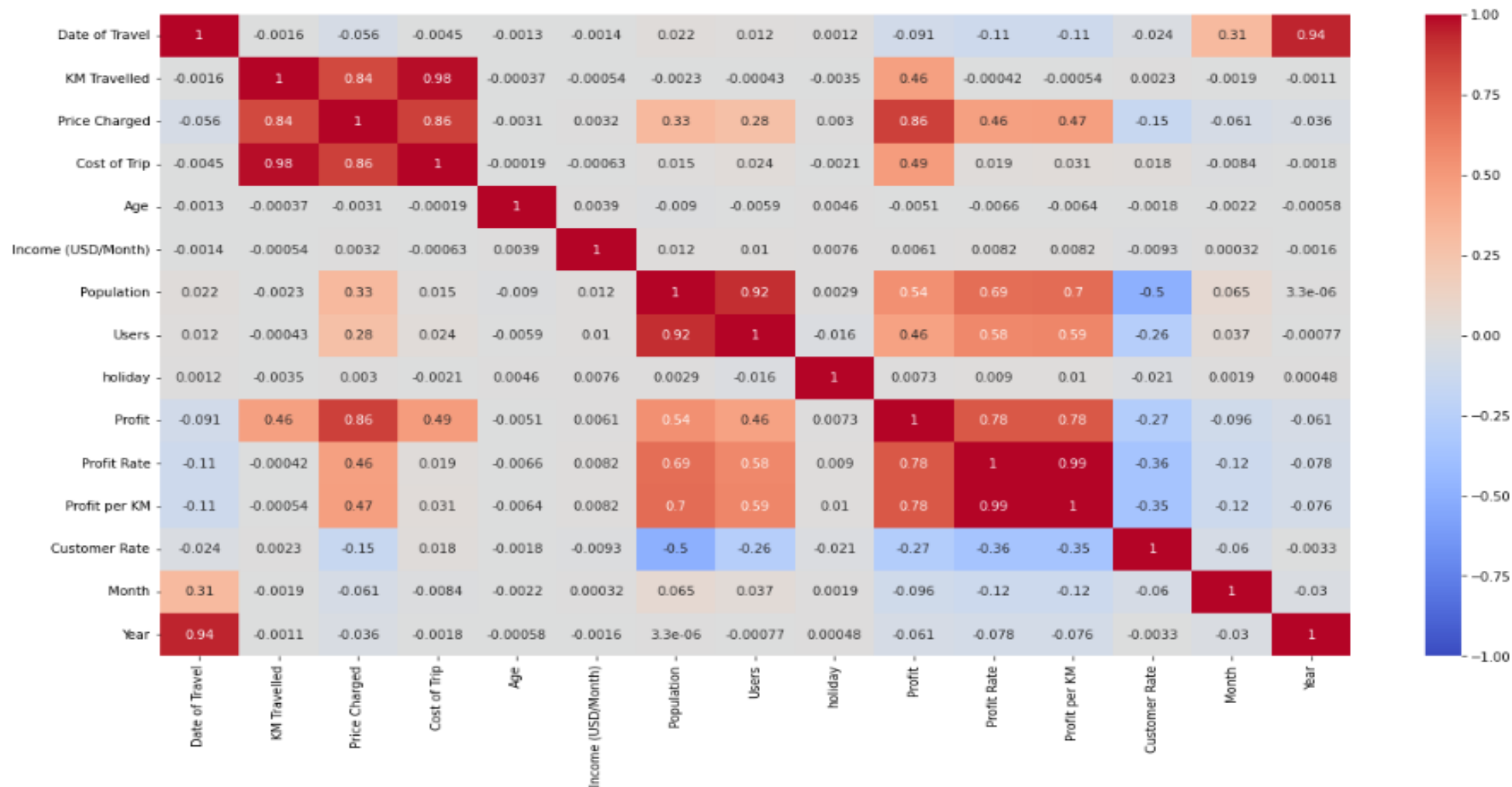
EDA Summary

- **User and Population**

Users and population box plots indicate that Pink Cab is mainly running smaller cities and Yellow cab is running mainly bigger cities.

EDA Summary

- **Correlation**

Recommendations

- Transaction: Even though the profit rates per year is almost even, the numbers of transaction per year from 2016 to 2018 is twice on Yellow Cab than Pink Cab.

- Profit: Yellow Cab is higher throughout the years from 2016 to 2018 than Pink Cab

- Thus, I recommend to invest Yellow Cab over Pink Cab.

Kyun Woo Hong

Thank You

Data Glacier

Your Deep Learning Partner