Data Glacier

Your Deep Learning Partner

Exploratory Data Analysis

G2M insight for Cab Investment firm

July 17 2022

Yining Liu

Agenda

Executive Summary

Problem Statement

Approach

EDA

EDA Summary

Recommendations

Problem Statement

- **Background**
- 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.

- We are anlaysing the relationship between different factors and how they affect the operation of cabs in major cities in the US

Executive Summary

- **Time period of data is from 31/01/2016** to **31/12/2018.** And there are four datasets:

1. Cab_Data.csv – this file includes details of transaction for 2 cab companies

2. Customer_ID.csv – this is a mapping table that contains a unique identifier which links the customer’ s demographic details

3. Transaction_ID.csv – this is a mapping table that contains transaction to customer mapping and payment mode

4. City.csv – this file contains list of US cities, their population and number of cab users

Executive Summary

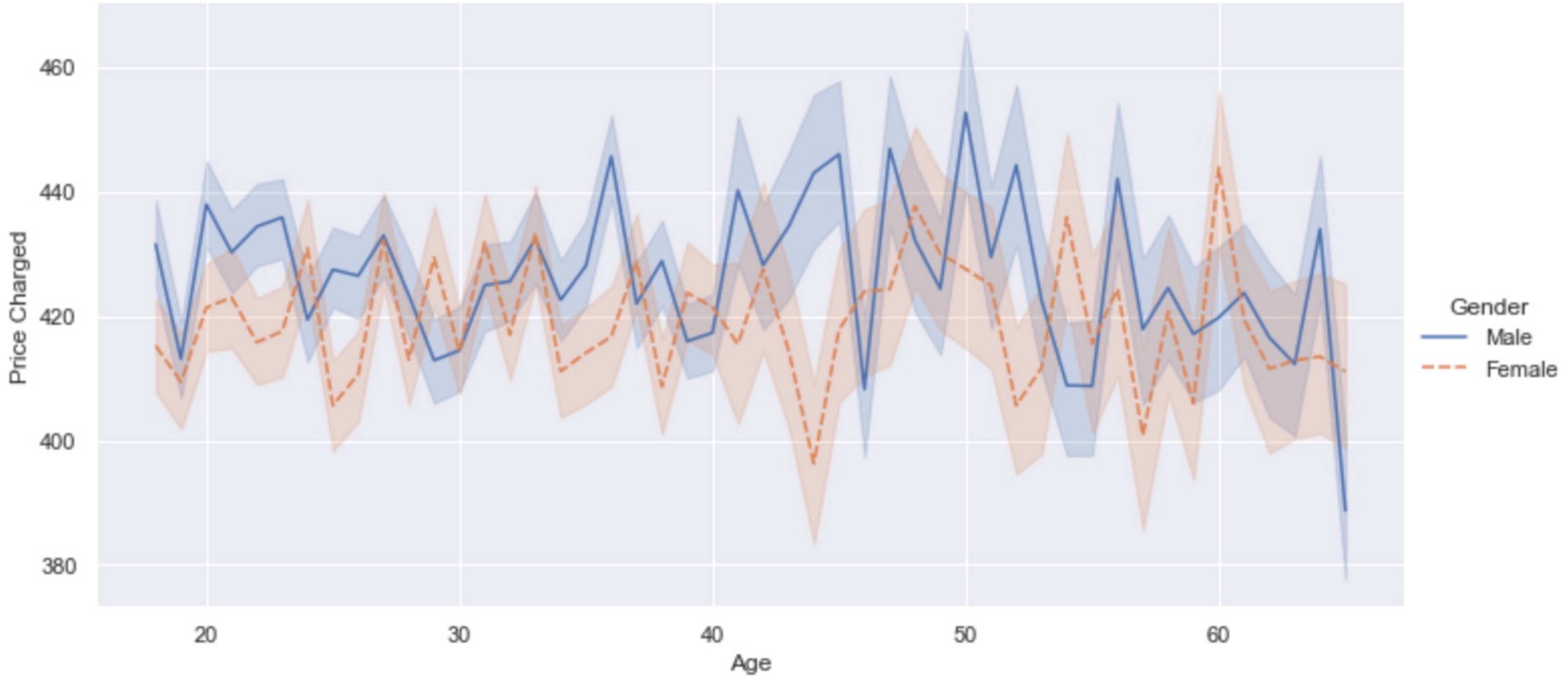
- There are 7 assumptions made base on the data set.
- Each assumption is shown with plots and descriptions below.

Approach

- In this project, I used Python as the major language and also imported packages like numpy, matplotlib.pyplot, csv, pandas and seaborn.
- I made many plots to quantitatively describe the relationships between variables more intuitively, which also helped in making comparison.
- I mainly used boxplots in the analysis since they are more descriptive by showing median, upper/lower quartile and outliers.

EDA

- Assumption 1: Whether the PRICE people are willing to pay for a cab ride is related to the GENDER and AGE of the passenger

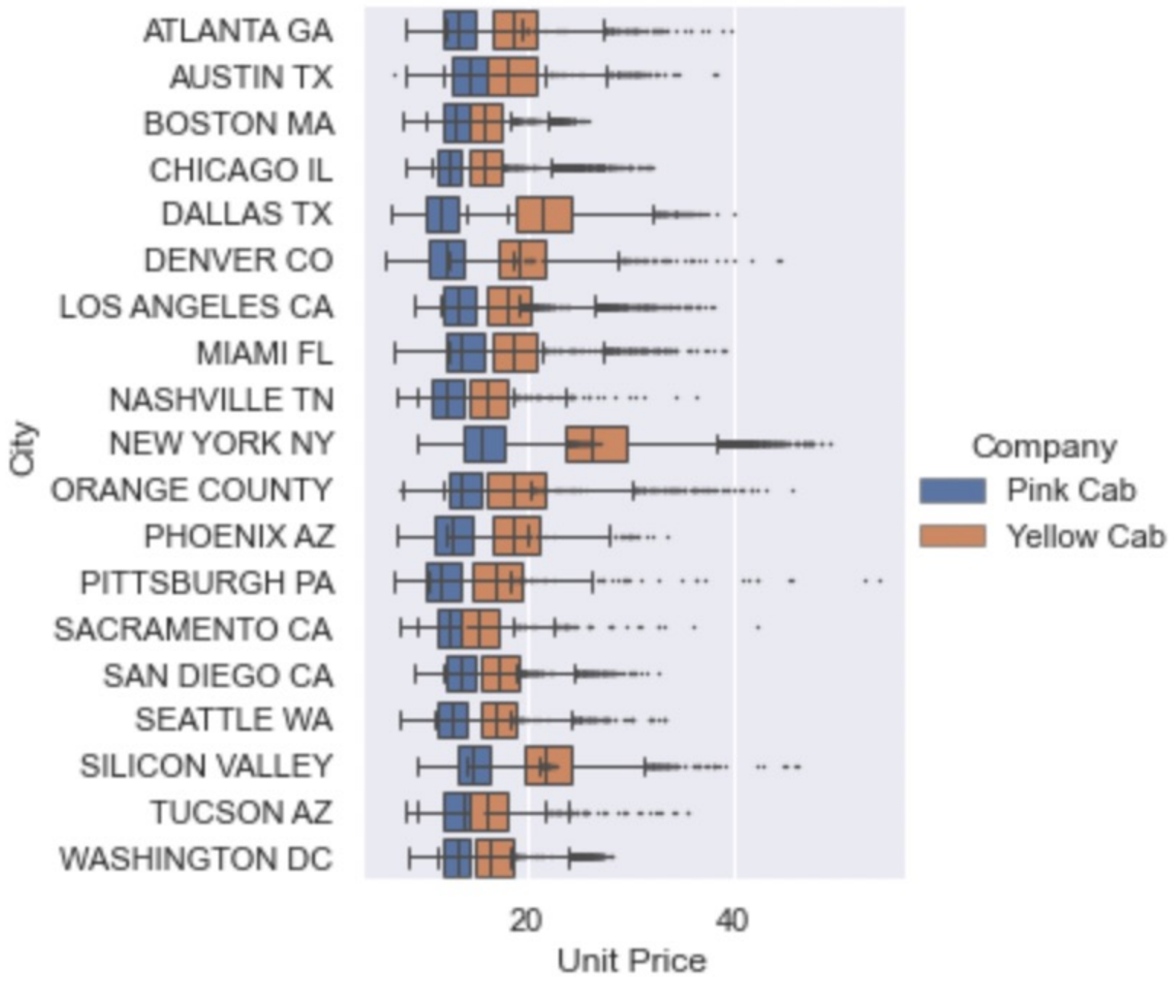


EDA Summary

- Summary 1:
- From the plot above we can observe that generally male are more likely to pay a higher price on each taxi ride no matter the age group. And there is also a slight trend that elder people would not like to spend more on a cab ride than the younger people.

EDA

- Assumption 2: The unit price per kilometer for a cab drive is dependent on then economic development. (The higher the development level, the higher the unit price would be.)

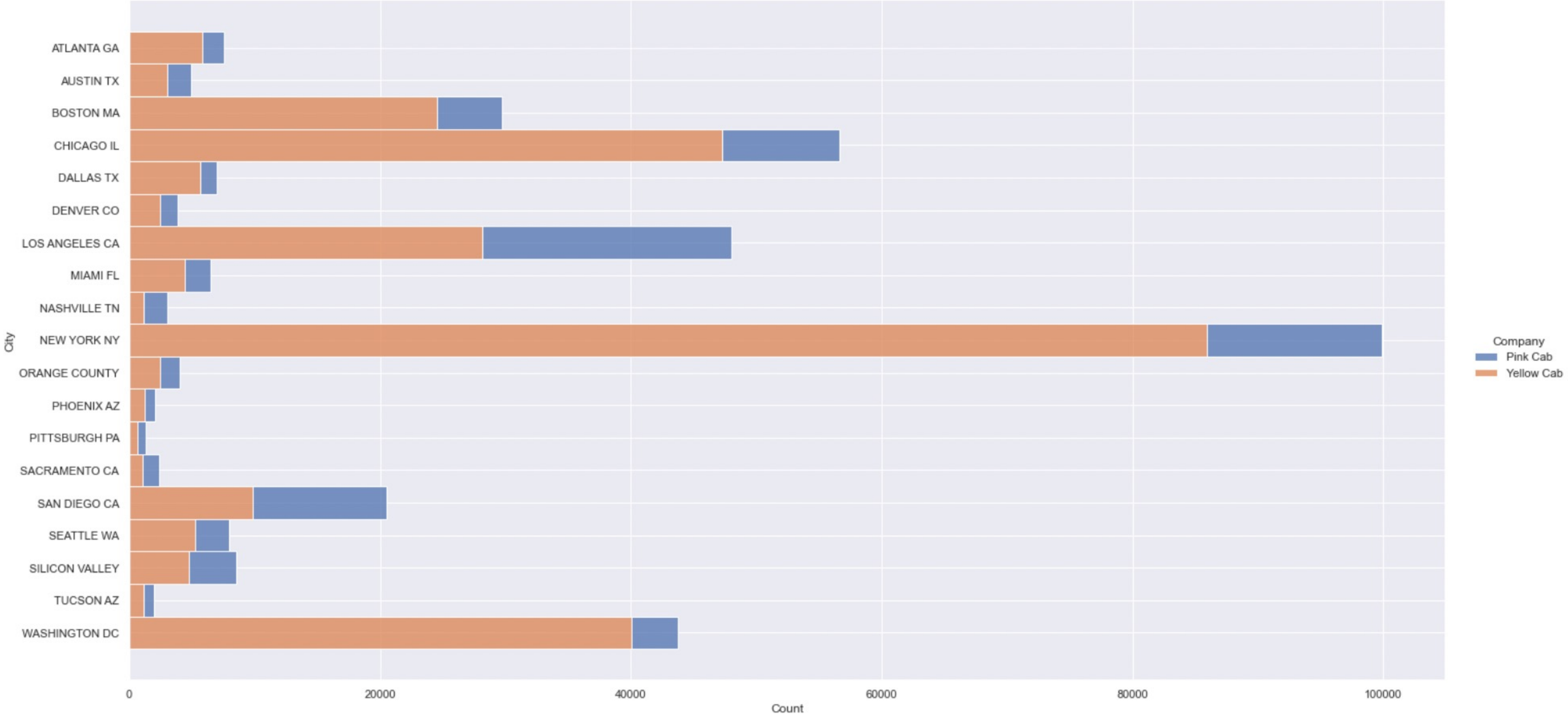


EDA Summary

- Summary 2:
- From the boxplot above, it is obvious that in all cities the price per kilometer of Yellow Cab is much higher than that in Pink Cab.
- In addition, unit price of cab rides in New York city is much higher than the prices in other cities, which makes sense since it is one of the most developed city in this world so there's a higher traffic demand and the cost of everything is New York is just higher than other cities.
- And we can also see that large cities like dallas silicon valley have higher unit price for the same reason above.
- So our assumption is right. Higher development level means that it's more expensive for a taxi ride.

EDA

- Assumption 3: The number of customers in different cities of the two companies in relation to the local economic situation.

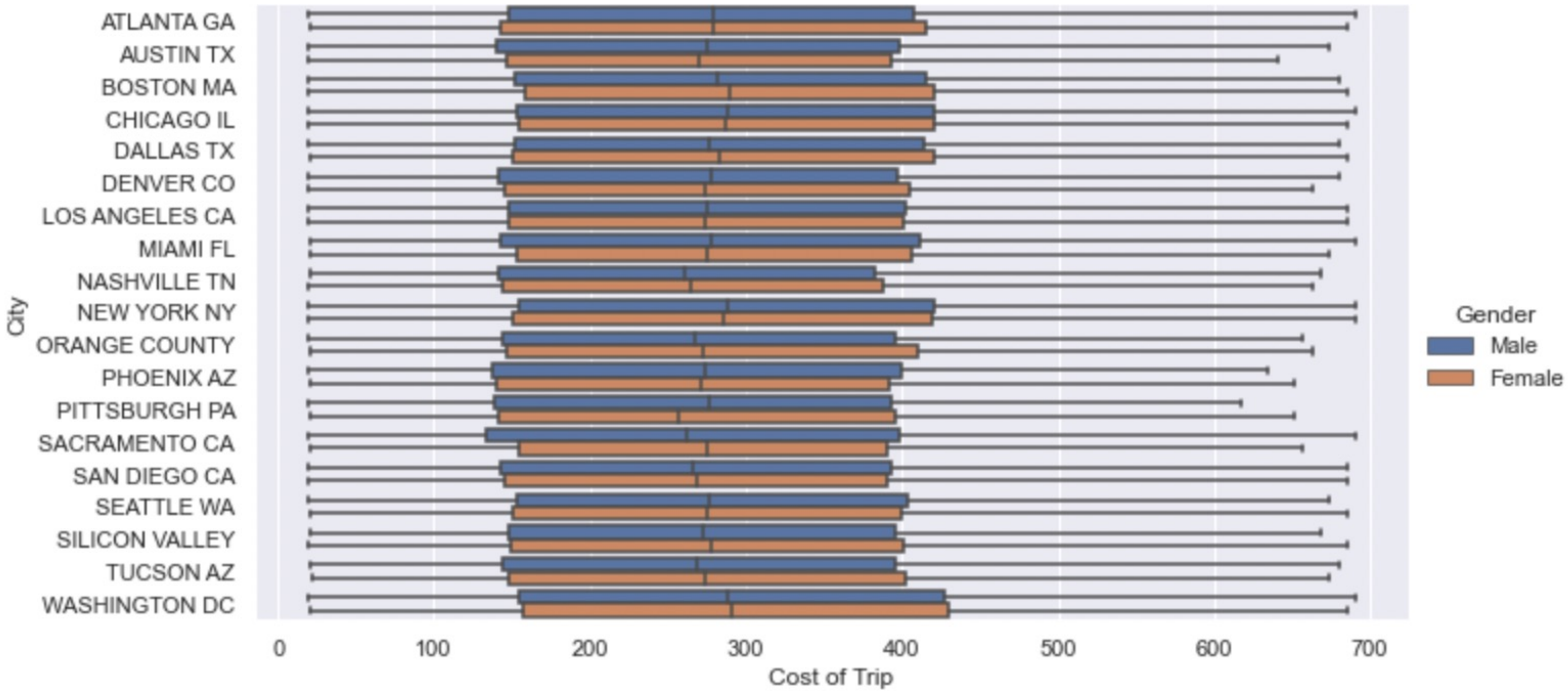


EDA Summary

- Summary 3:
- From the histogram above we can easily point out that in most of the cities, the costumer volume of Yellow Cab is much greater than that of the Pink Cab.
- But there are also 2 exceptions: Nashiville, Sacramento. In San Diego, 2 companies have the similar customer volume.

EDA

- Assumption 4: In different states, is there a difference in price acceptance between the different genders of passengers？

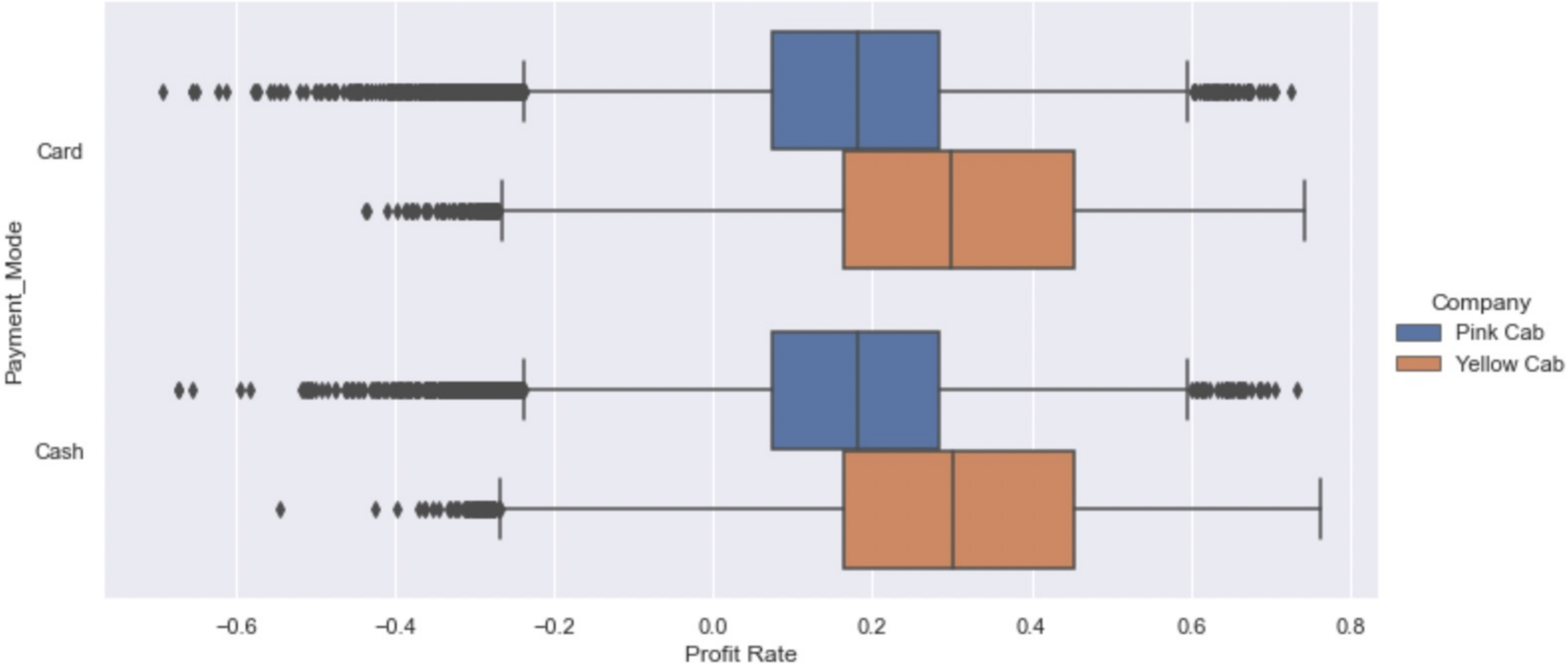


EDA Summary

- Summary 4
- From the plot above we can see that in different states, there shows no obvious tendencies that passenger in different genders have different acceptability on the price of each taxi ride.
- Gender is not a factor deciding the price a passenger would pay for a taxi drive.

EDA

- Assumption 5: The 2 cab companies have different profitability rates on different payment method.

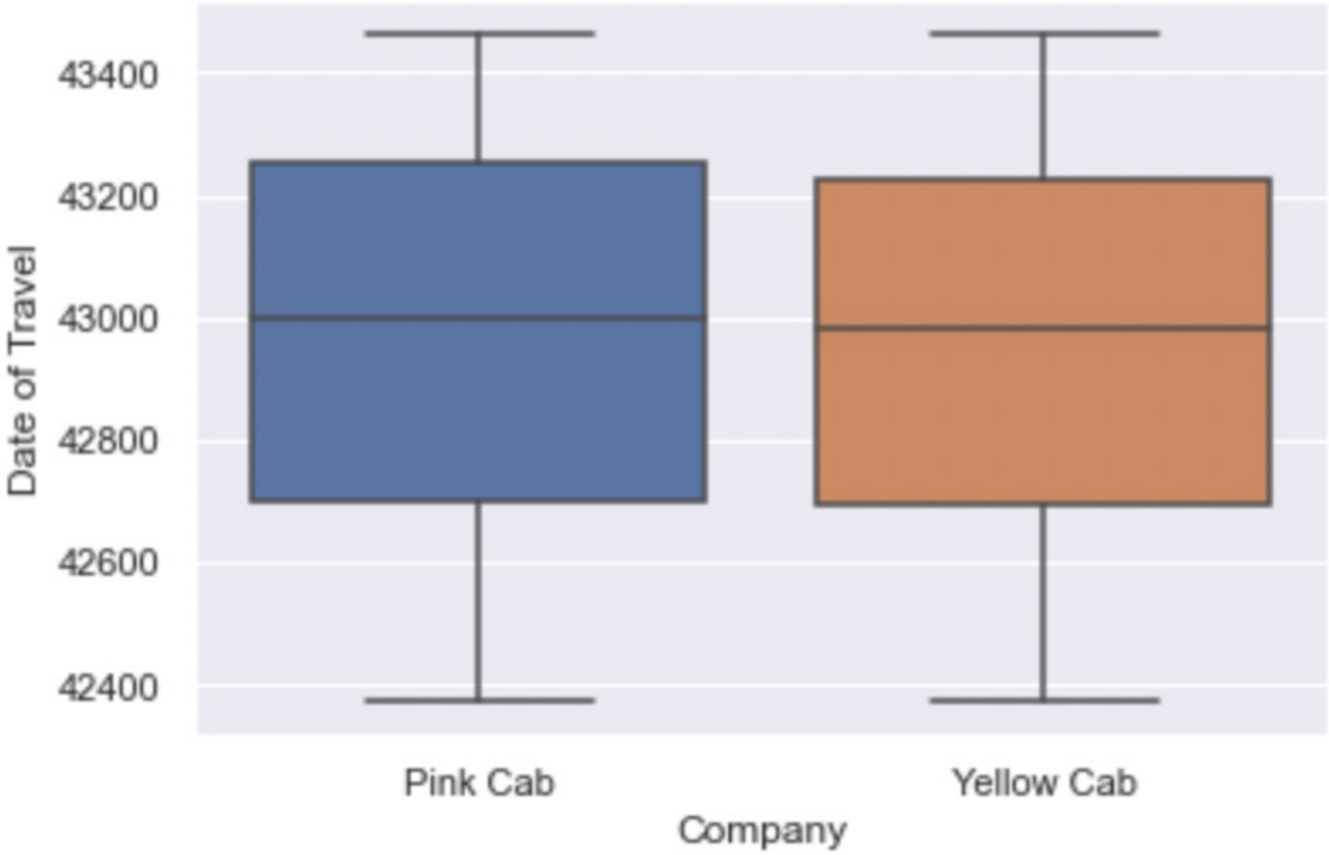


EDA Summary

- Summary 5:
- From the boxplot above we can observe that generally payment method does NOT affect the profit rate of each company.
- But generally speaking Yellow Cab makes more money than Pink Cab, whcih mean that the Yellow Cab has a higher profit rate.

EDA

- Assumption 6: Will the consumers choose different cab companies on different dates of travel?

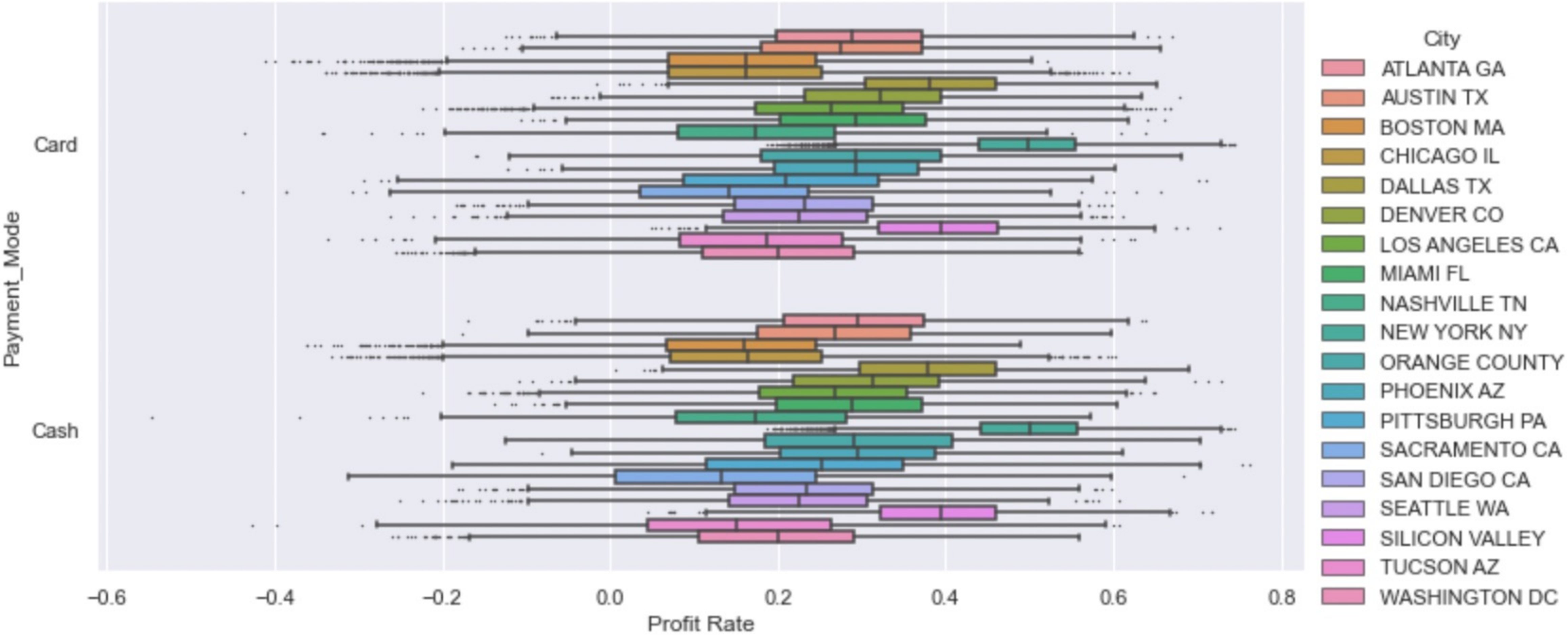


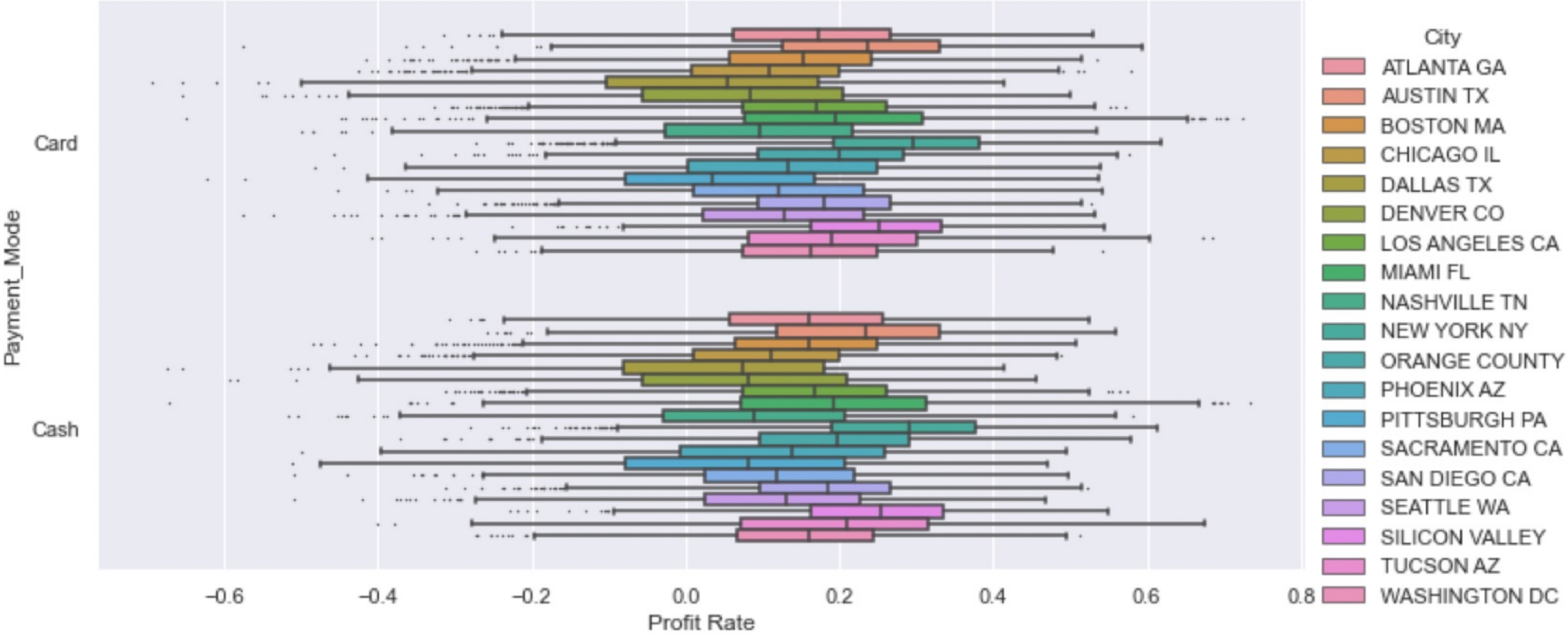
EDA Summary

- Summary 6:
- From the plot above we can see that the median, upper/lower interquartile are pretty similar. And the boundary of maximum and minimum value are at the same level. So there is no relation between the date of travel and the cab companies chosen, which means that consumers take taxi randomly without much special preference on the companies.

EDA

- Assumption 7: Profit rate of each company does NOT varies state-wisely.




Card
Cash
Payment_Mode
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
Profit Rate
City
ATLANTA GA
AUSTIN TX
BOSTON MA
CHICAGO IL
DALLAS TX
DENVER CO
LOS ANGELES CA
MIAMI FL
NASHVILLE TN
NEW YORK NY
ORANGE COUNTY
PHOENIX AZ
PITTSBURGH PA
SACRAMENTO CA
SAN DIEGO CA
SEATTLE WA
SILICON VALLEY
TUCSON AZ
WASHINGTON DC

EDA Summary

- Summary 7:
- Obvious that the patterns have no difference when it comes to different payment methods and cab companies.
- And in each state, profit rate of both Yellow Cab and Pink Cab are similar to each other.

Recommendations

- From the above data and images we can conclude that, in general, the cab industry is differently developed in different cities due to their different levels of development. The more developed the city's economy is, the greater the turnover of the cab industry.
- However, regardless of the city, price is not the main factor that determines the industry. For example, Yellow Cab's unit price is much higher than Pink Cab's, but Yellow Cab's business is much better than Pink Cab's. Therefore, no matter where passengers are, their choice of whether to take a cab is based on the need or not, and price is not price is not a factor that prevents them from taking a cab. Also, the dates did not have any significant impact on the cab companies' business

Thank You

Data Glacier

Your Deep Learning Partner