Data Glacier

Your Deep Learning Partner

Exploratory Data Analysis

G2M Case Study

Melis Tekin Akcin

26-June-2021

Outline

- ❑ Problem Statement
- ❑ History of Datasets
- ❑ EDA
- ❑ Recommendations

Problem Statement- G2M Cab Industry Case Study

- 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.

- Objective : XYZ is interested in using your actionable insights to help them identify the right company to make their investment.

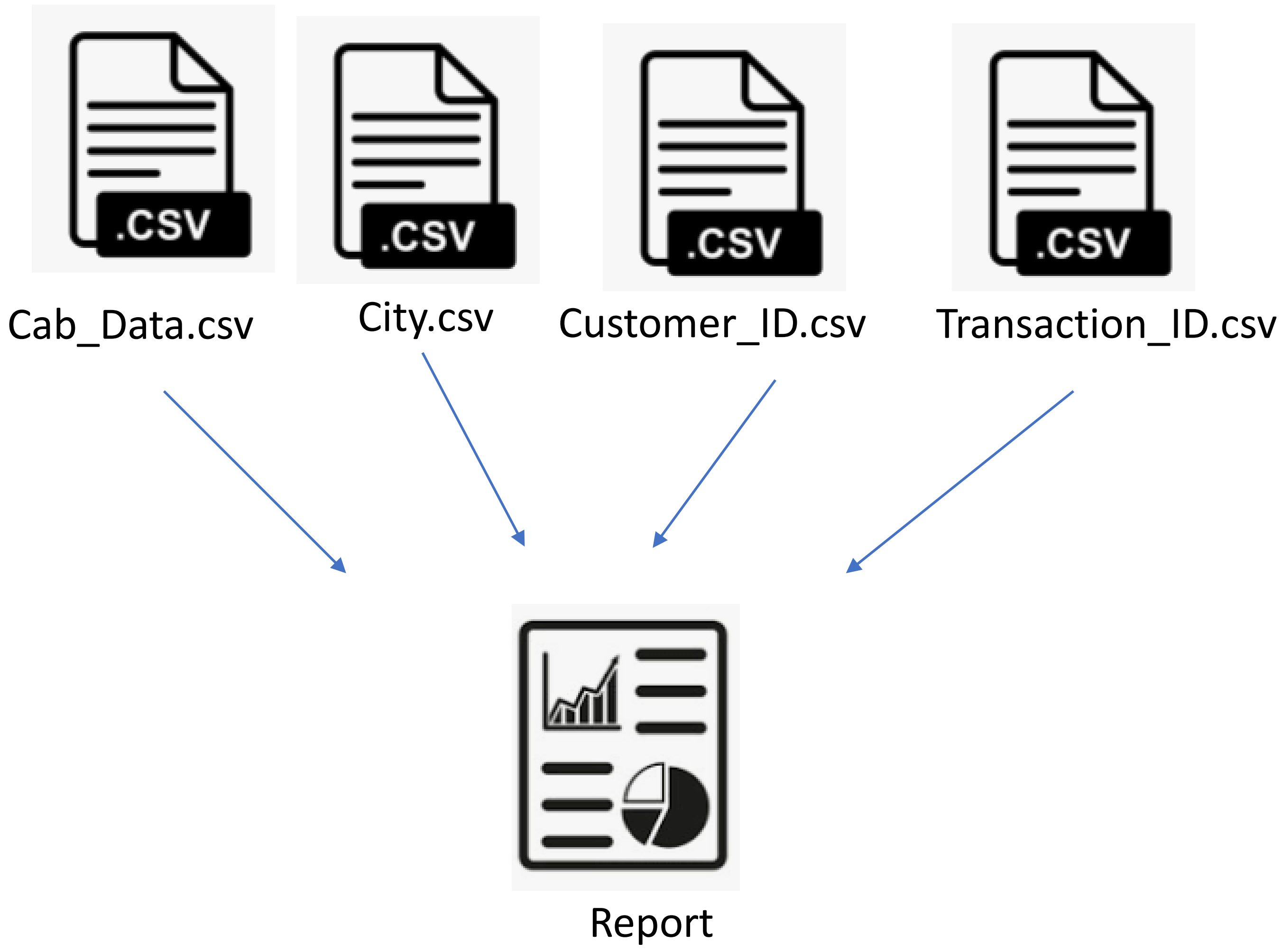
Problem Statement- G2M Cab Industry Case Study

The Analysis has performed as following:

- Understanding datasets,
- Identifying the relations between features,
- Finding the most preferred cab company,
- Recommendation for investment.

Dataset Information

- 4 different datasets.
- 16 different features (including 2 derived features).
- Time period of data is from **31/01/2016 to 31/12/2018.**



Cab_Data.csv

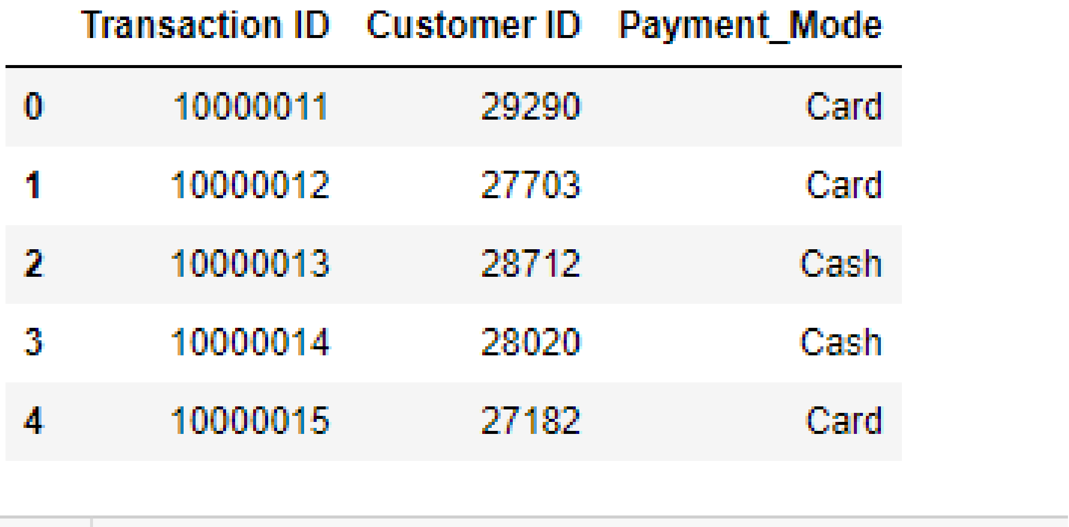
City.csv

Customer_ID.csv

Transaction_ID.csv

Report

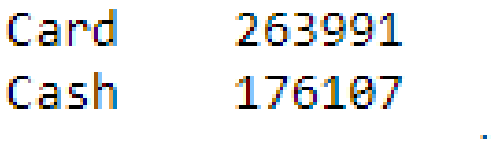
History of the datasets – Transaction_ID Dataset



	Transaction ID	Customer ID	Payment_Mode
0	10000011	29290	Card
1	10000012	27703	Card
2	10000013	28712	Cash
3	10000014	28020	Cash
4	10000015	27182	Card

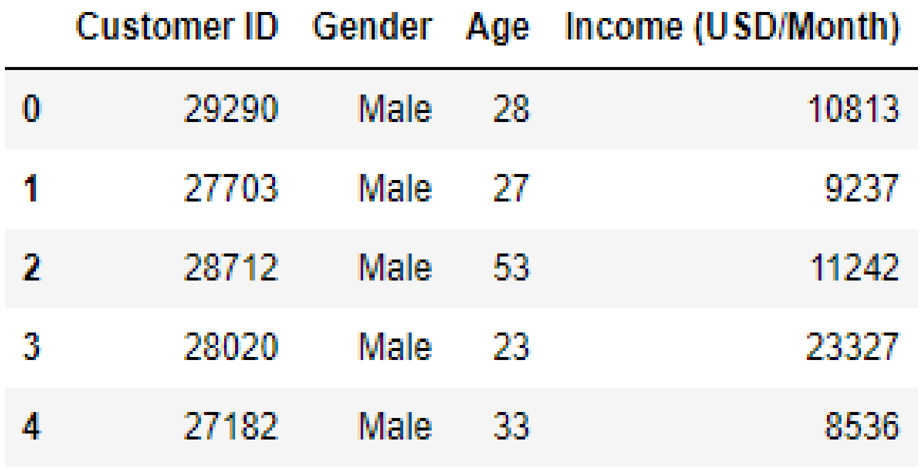
- Total Data Points : 440098
- There is no NA value
- There is no duplicate row

Number of Cash and Card Users:



```
Card      263991
Cash      176107
```

History of the datasets - Customer_ID Dataset

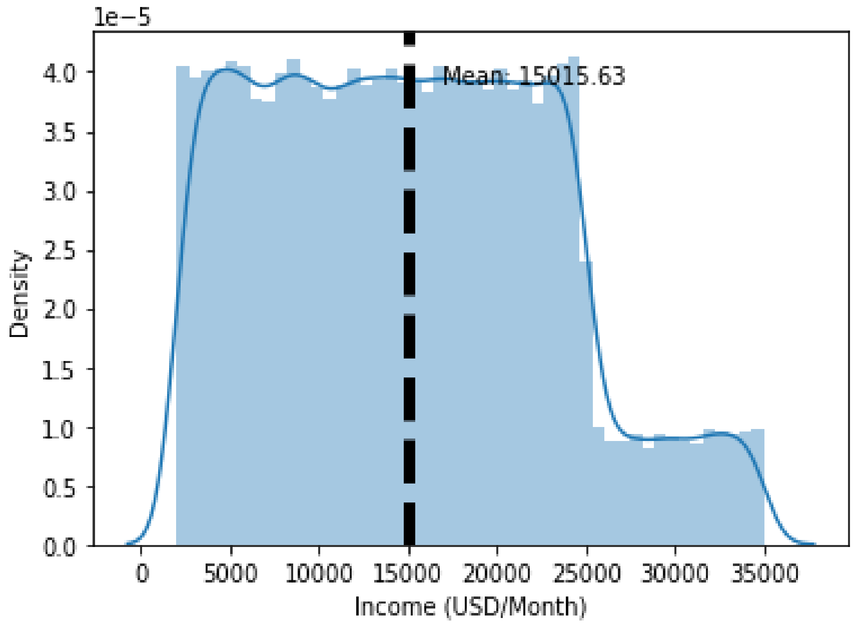


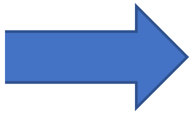
	Customer ID	Gender	Age	Income (USD/Month)
0	29290	Male	28	10813
1	27703	Male	27	9237
2	28712	Male	53	11242
3	28020	Male	23	23327
4	27182	Male	33	8536

- Total Data Points : 49171.
- There is no NA value.
- There is no duplicate row.
- There is no outliers.

History of the datasets – Customer_ID Dataset

Mean Value of the Income feature

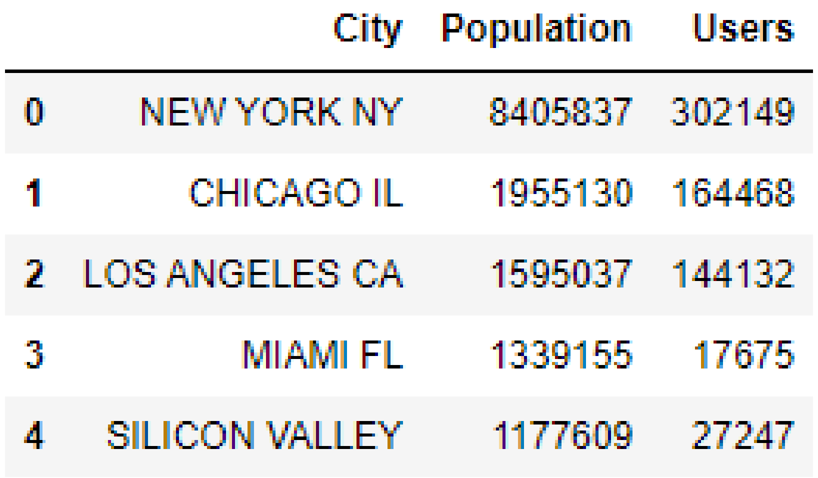




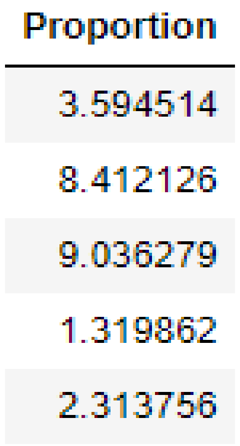
Assumptions:

- We will treat cab users whose salary is higher that 25000 as upper-class.
- We will treat cab users whose salary is between 10000 and 25000 as middle-class.
- We will treat those whose salary is lower than 10000 value as lower-class.

History of the datasets – City Dataset



	City	Population	Users
0	NEW YORK NY	8405837	302149
1	CHICAGO IL	1955130	164468
2	LOS ANGELES CA	1595037	144132
3	MIAMI FL	1339155	17675
4	SILICON VALLEY	1177609	27247



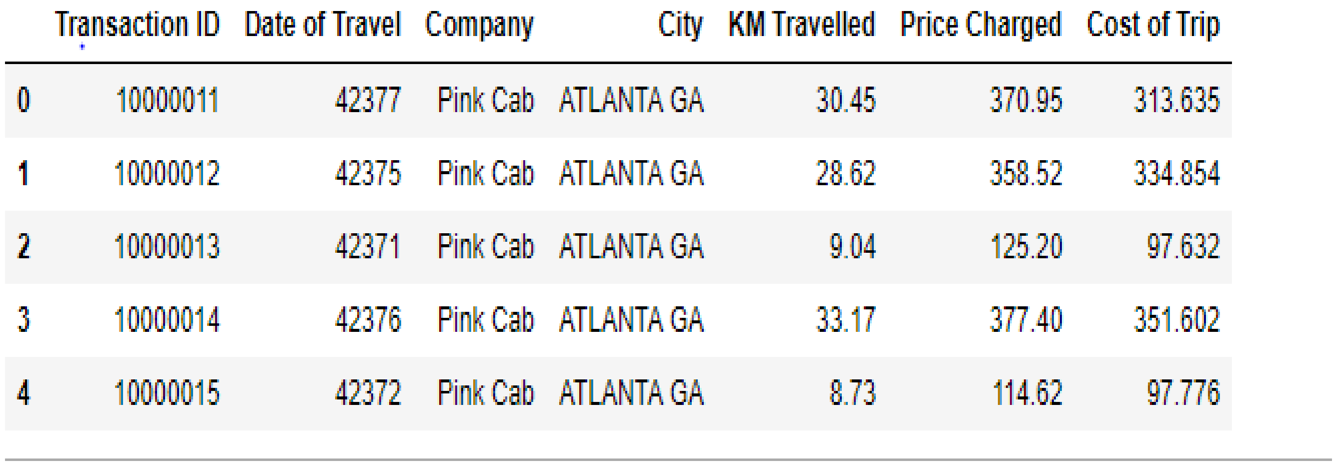
Proportion
3.594514
8.412126
9.036279
1.319862
2.313756

<u>Assumption:</u> There are two outliers in both Population and Users data. Since it will not affect our results, we are not treating them as outliers.

- Number of Features: 3.
- Total data Points: 20.
- No NA value.

History of the Datasets- Cab Dataset

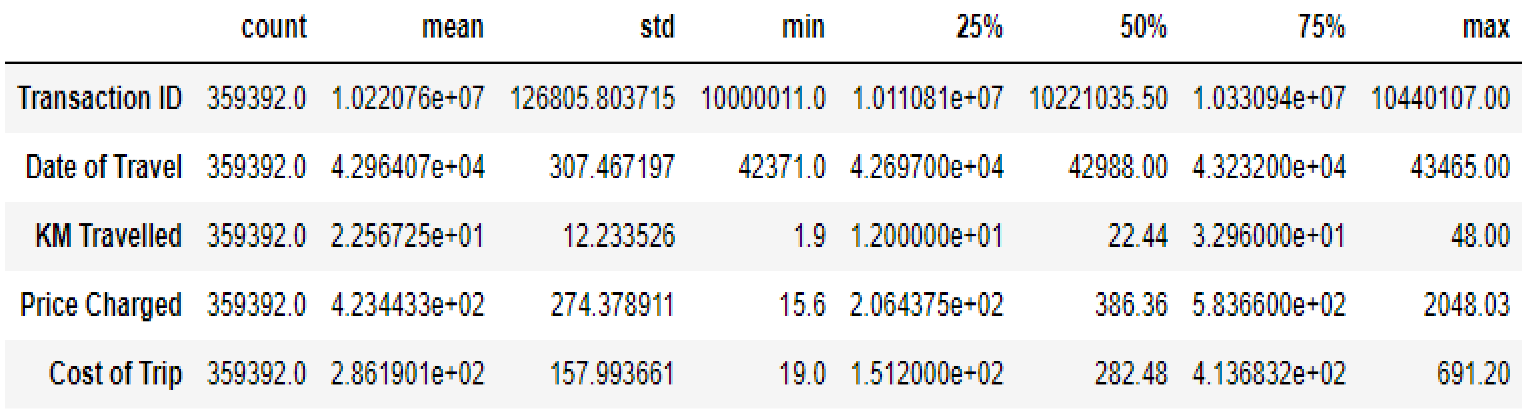
Cab dataset



	Transaction ID	Date of Travel	Company	City	KM Travelled	Price Charged	Cost of Trip
0	10000011	42377	Pink Cab	ATLANTA GA	30.45	370.95	313.635
1	10000012	42375	Pink Cab	ATLANTA GA	28.62	358.52	334.854
2	10000013	42371	Pink Cab	ATLANTA GA	9.04	125.20	97.632
3	10000014	42376	Pink Cab	ATLANTA GA	33.17	377.40	351.602
4	10000015	42372	Pink Cab	ATLANTA GA	8.73	114.62	97.776

- Number of Features: 7.
- Total data Points: 359392.
- No NA value.

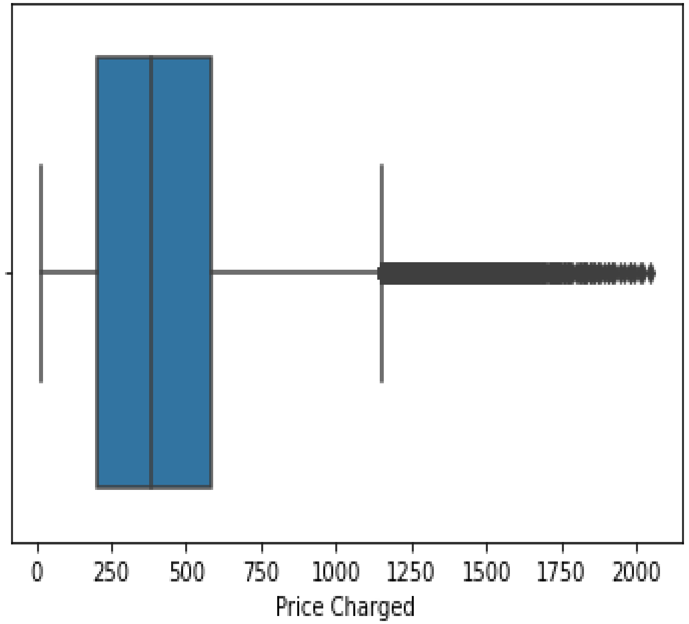
Descriptive Analysis for Cab Dataset

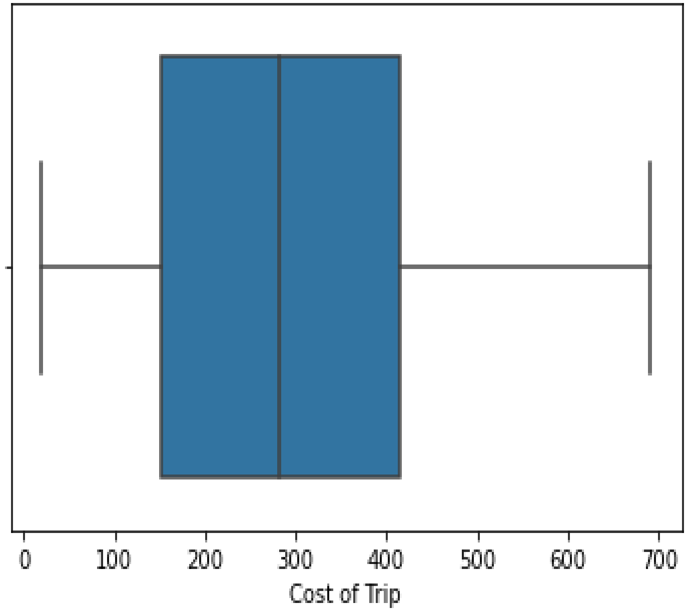


	count	mean	std	min	25%	50%	75%	max
Transaction ID	359392.0	1.022076e+07	126805.803715	10000011.0	1.011081e+07	10221035.50	1.033094e+07	10440107.00
Date of Travel	359392.0	4.296407e+04	307.467197	42371.0	4.269700e+04	42988.00	4.323200e+04	43465.00
KM Travelled	359392.0	2.256725e+01	12.233526	1.9	1.200000e+01	22.44	3.296000e+01	48.00
Price Charged	359392.0	4.234433e+02	274.378911	15.6	2.064375e+02	386.36	5.836600e+02	2048.03
Cost of Trip	359392.0	2.861901e+02	157.993661	19.0	1.512000e+02	282.48	4.136832e+02	691.20

There is a huge gap between the maximum values of Price Charged and Cost of Trip features. That suggests us to detect whether there are outliers.

Outlier Detection

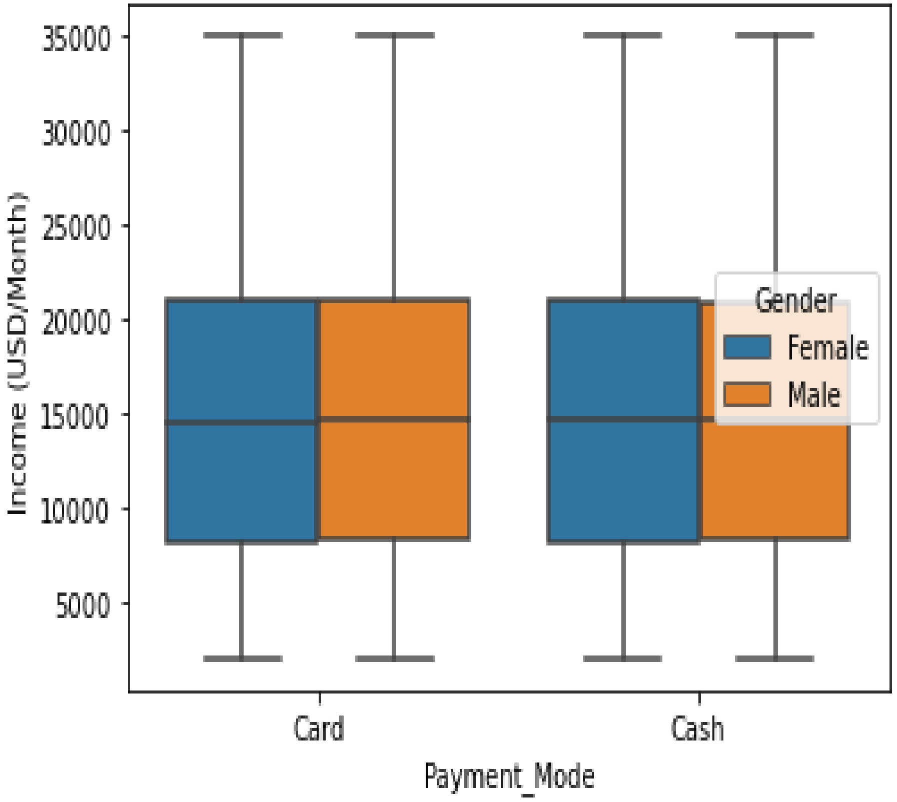




Since ignoring the outliers can change the structure of the data and the possibility that it can be the company's policy, we decided to correct the outliers.

<u>Data Correction: </u>We suppressed the outlier with the upper bound value.

Income and Payment Mode Analysis in terms of Gender

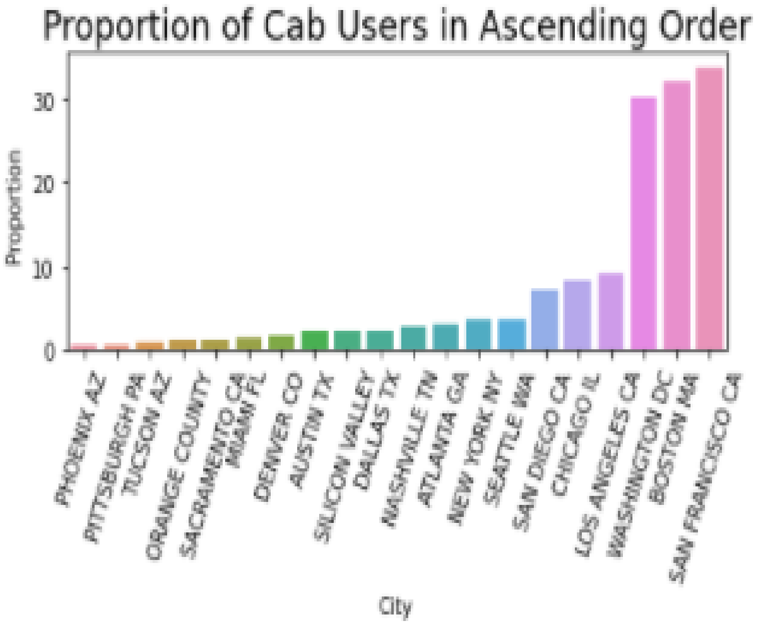


- If we consider Payment Mode and Income relation, the proportion of the male and female cab users are almost the same.

- In upper-class cab users:

%11 Male, %9 Female users prefer Pink cab company.
%42 Male, %36 Female users prefer Yellow cab company.

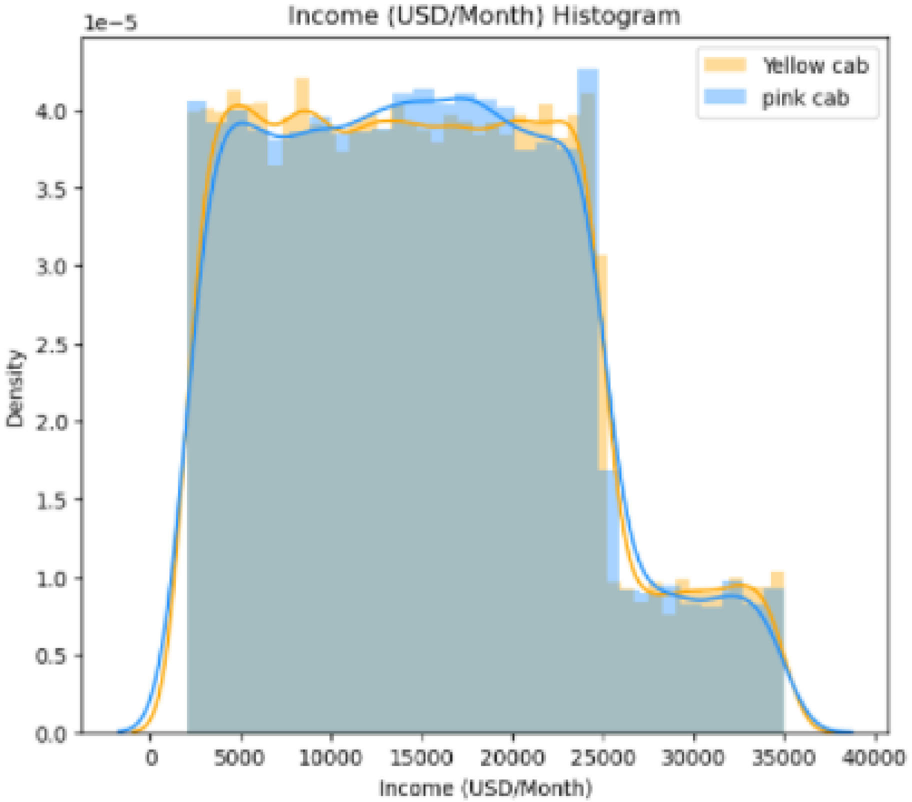
Cab Users Proportion in Different Cities



The first five cities of highest Proportion:

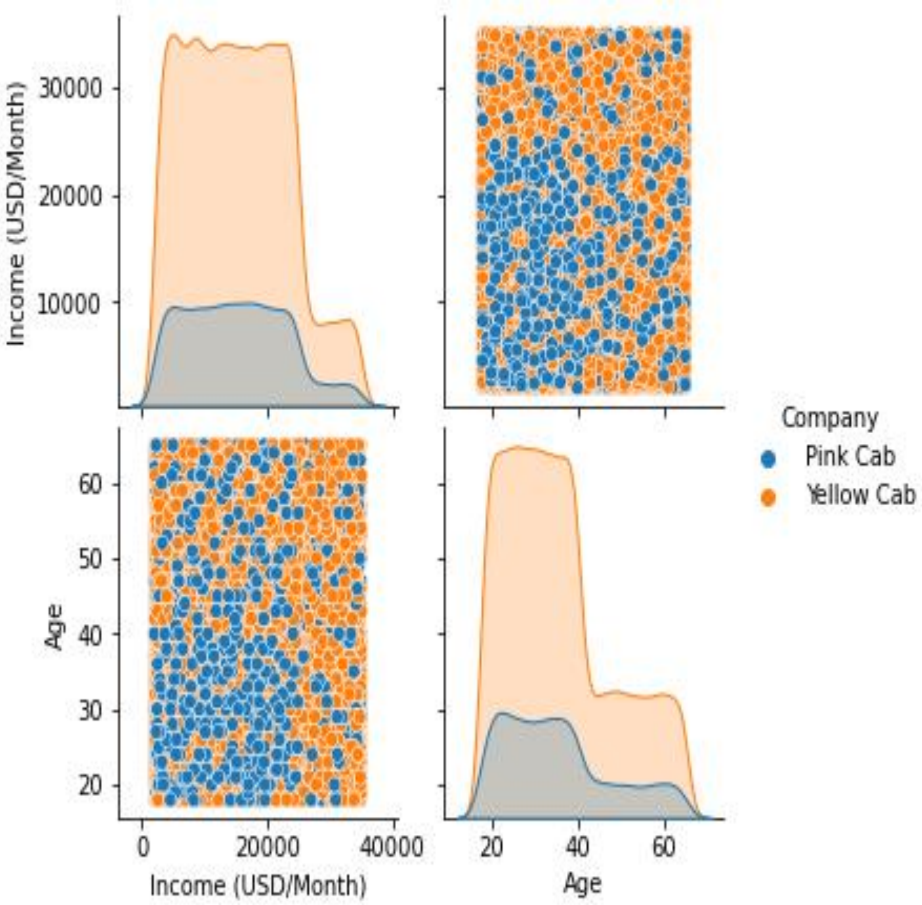
1) San Francisco CA,
2) Boston MA,
3) Washington DC,
4) Los Angeles CA
5) Chicago.

Relation between Income and Cab Company



There is no significant difference between the mean values of the incomes between Yellow and Pink cab companies.

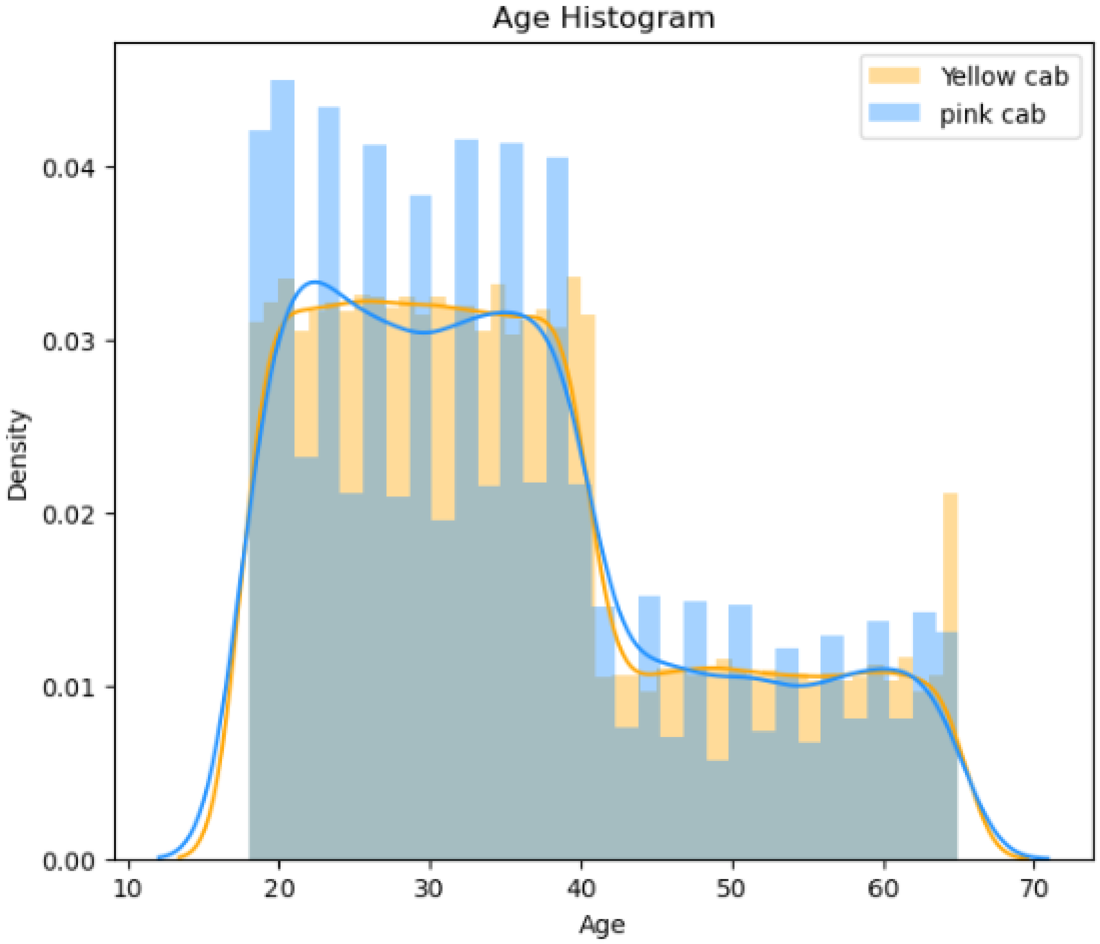
Relation between Income and Age in terms of Cab Company



- Lower and Middle class customers of Age less than or equal 40 prefer to use Pink Company.

- In addition to being a preferred company in middle class, Yellow cab is the most preferred cab company in upper- class Users.

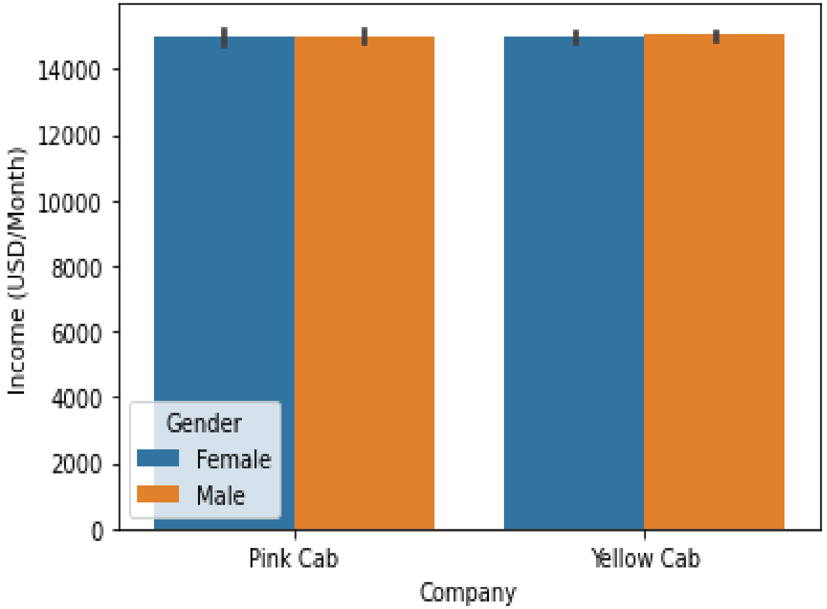
Company Preference in terms of Age



Yellow Cab's Age average is equal to Pink Cab's Age average.

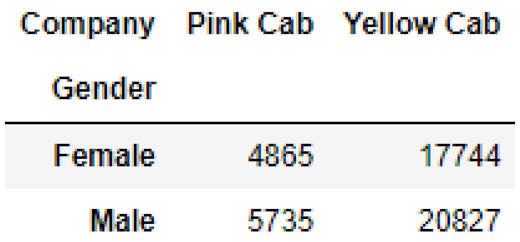
- Yellow Cab’s Age Average is 35.38.
- Pink Cba’s Age Average is 35.28.

Company Preference in terms of Gender



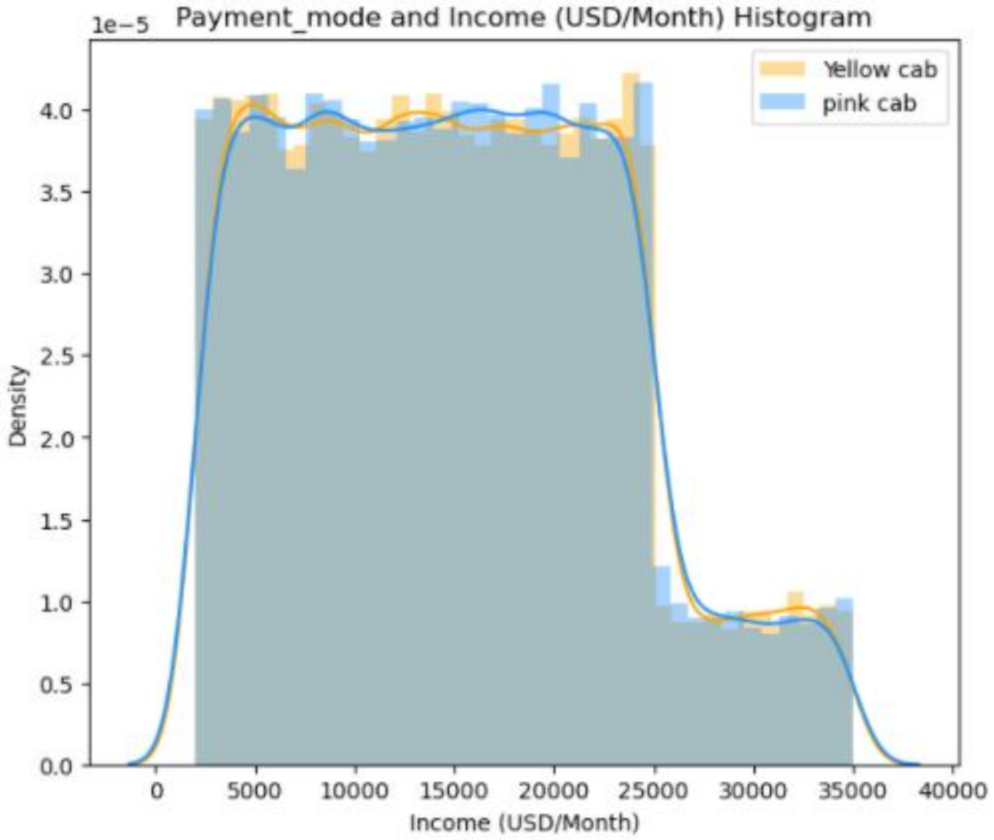
The gender rate that prefers Yellow Cab is equal to the gender rate of the Pink Cab.

Gender-Company Cross Table



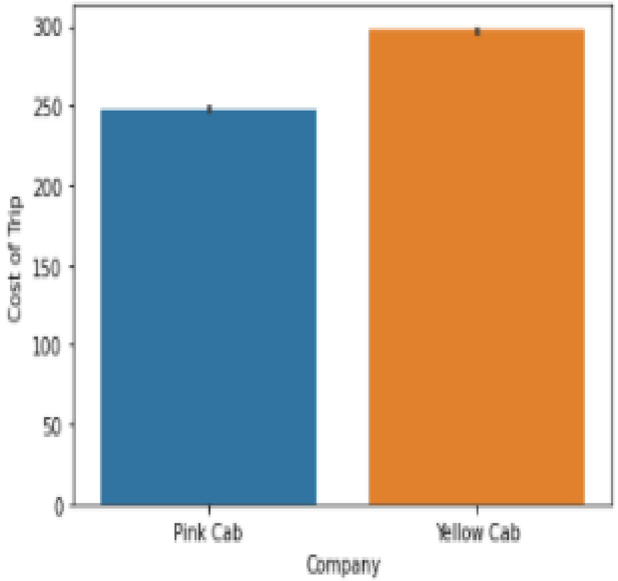
Company	Pink Cab	Yellow Cab
Gender		
Female	4865	17744
Male	5735	20827

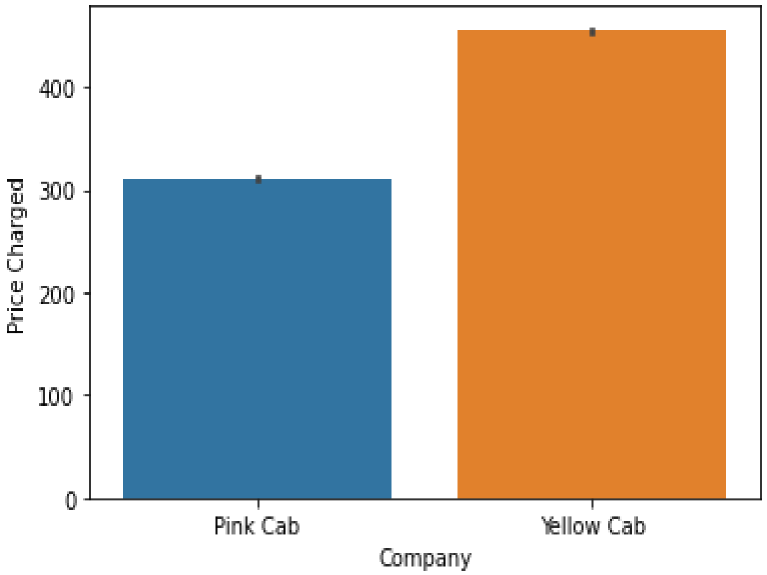
Payment Mode and Income Relation in Cab Companies



As income increases, there is no significant difference between Card and Cash Users in Yellow and Pink Company.

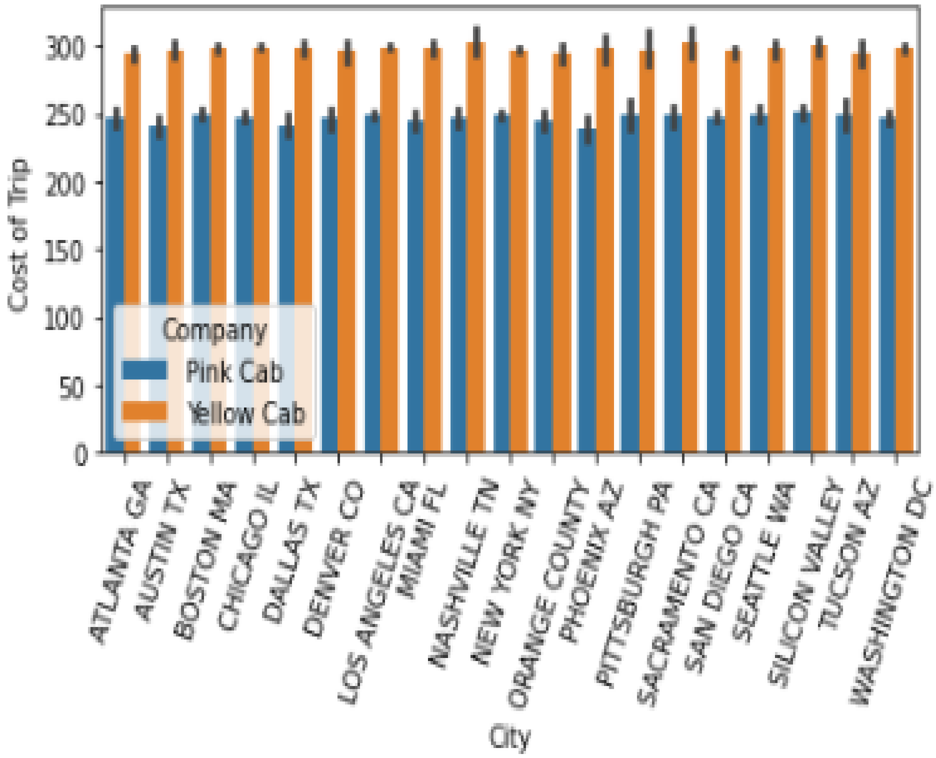
Cost of Trip and Price Charged Relation in Cab Companies





If we consider the Cost of Trip and Price Charged values, Yellow cab company adds more profits on top of their prices.

Cost of Trip in Different Cities and Cab Companies

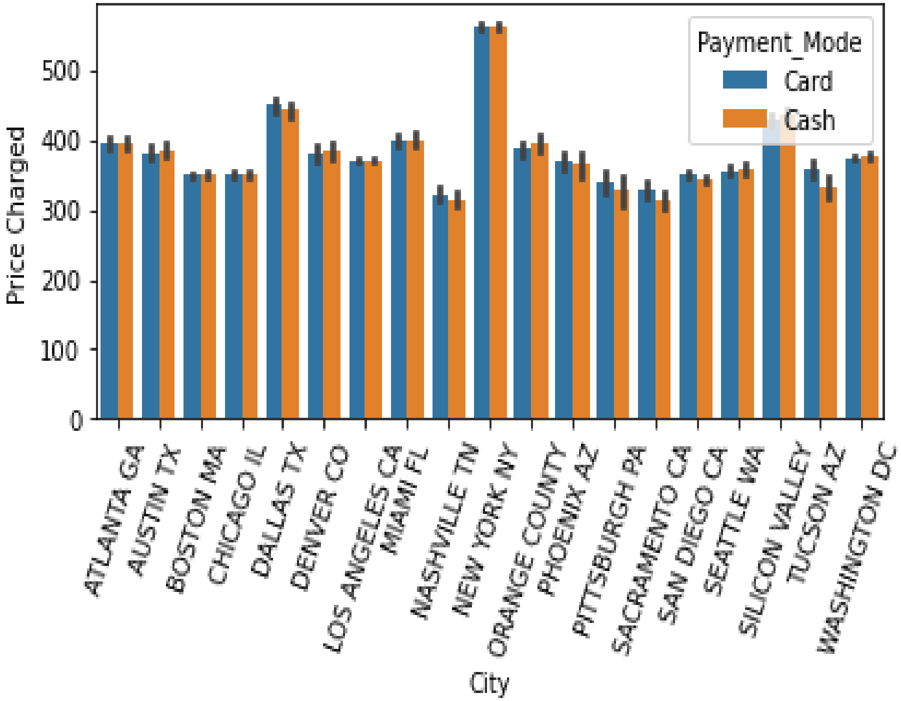


- The overall cost of trip in Yellow Company is much higher than Pink Company.

- Yellow Cab has the highest cost of Trip in Nashville TN.

- **Yellow Cab Company dominates the majority of the Market in all the 20 cities.**

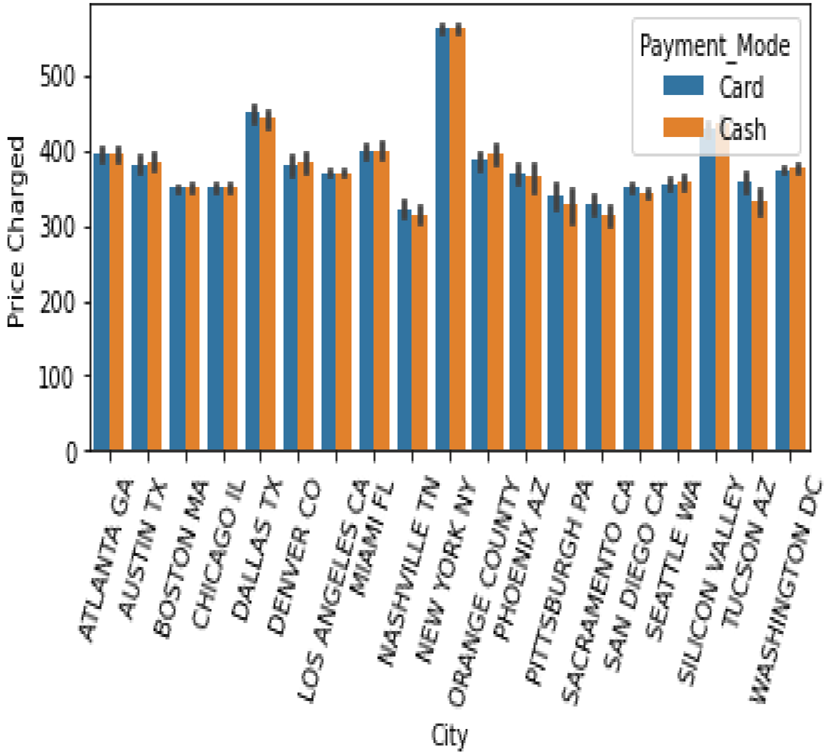
Price Charged in Different Cities and Cab Companies



- The overall charged price in Yellow Company is much higher than Pink Company.

- New York NY has the highest difference between Pink and Yellow Cab Company.

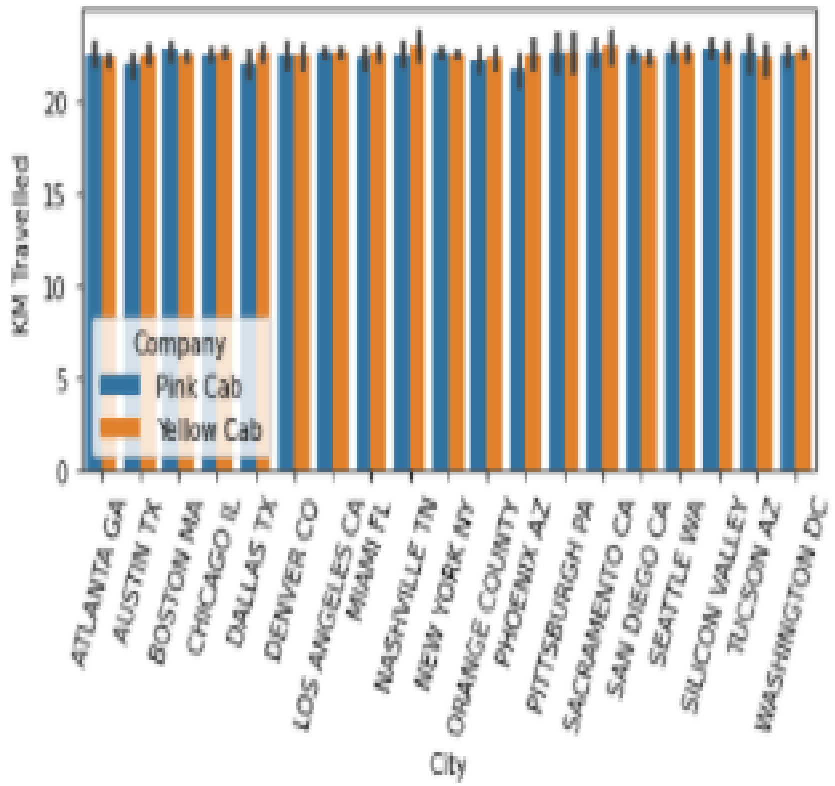
Price Charged and Payment Mode Relation in Different Cities



- There is no significant difference in Card and Cash Users as charged price increases.

- Users prefer to pay by card slightly higher in Dallas TX, Nashville TN, San Diego CA, Sacramento CA, Phoenix AZ, Tucson AZ.

Ride Distance in Different Cities and Cab Companies

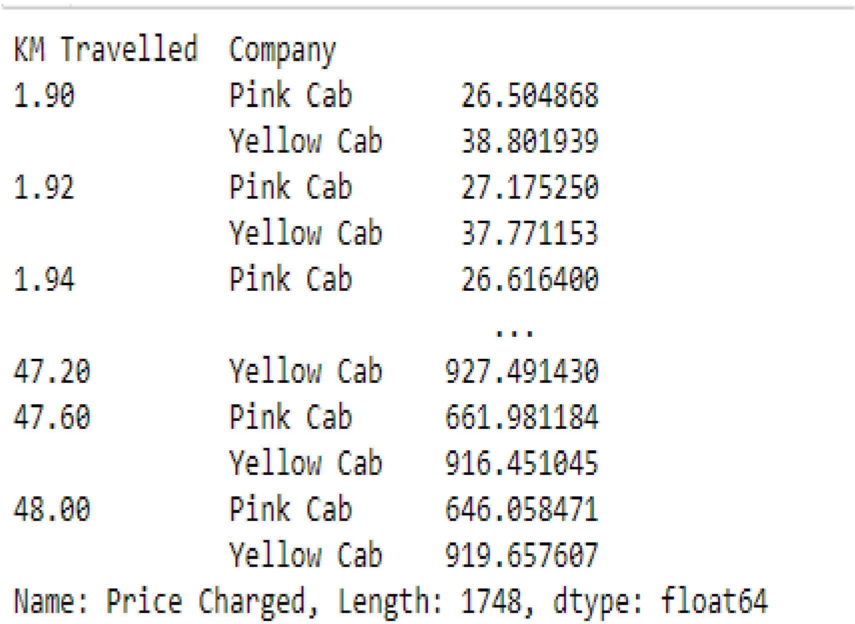


- There is no significant difference in KM Travelled between two cab companies**.**

- In Atlanta City, Boston MA, New York NY, Silicon Valley and Tucson AZ, the KM Travelled is lightly higher in Pink Cab Company.

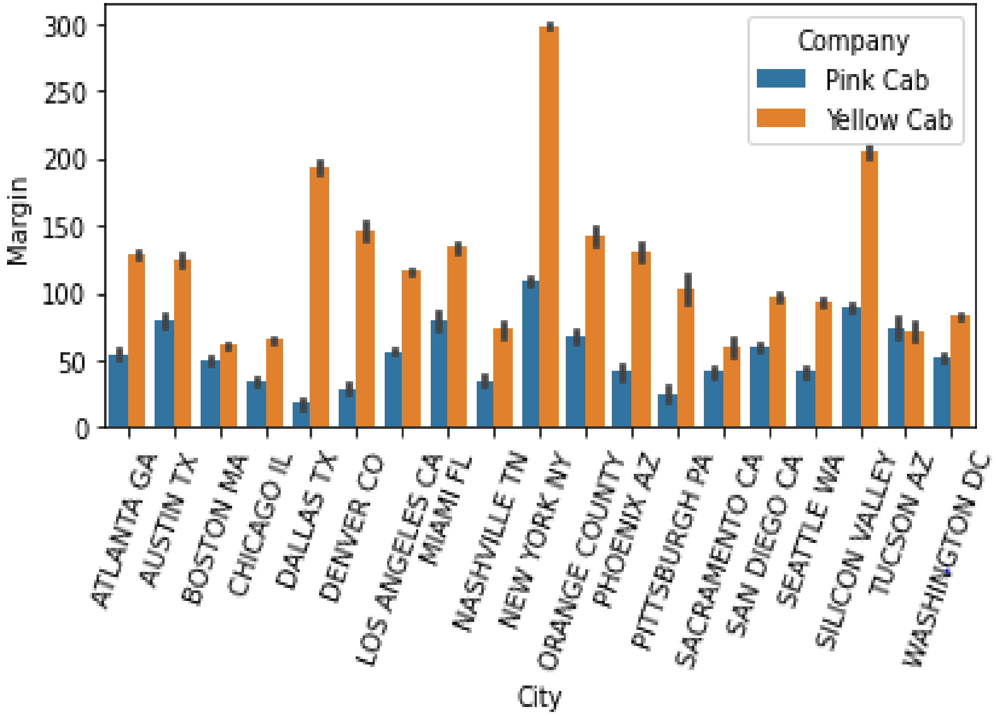
Customer Preference between Pink and Yellow Company

Pink Company is more customer friendly if we consider KM Travelled and Price Charged features. Does being customer friendly affect the profit rate?



```
KM Travelled  Company
1.90          Pink Cab       26.504868
              Yellow Cab     38.801939
1.92          Pink Cab       27.175250
              Yellow Cab     37.771153
1.94          Pink Cab       26.616400
                               ...
47.20         Yellow Cab    927.491430
47.60         Pink Cab      661.981184
              Yellow Cab    916.451045
48.00         Pink Cab      646.058471
              Yellow Cab    919.657607
Name: Price Charged, Length: 1748, dtype: float64
```

Margin Rates in Different Cities and Cab Companies

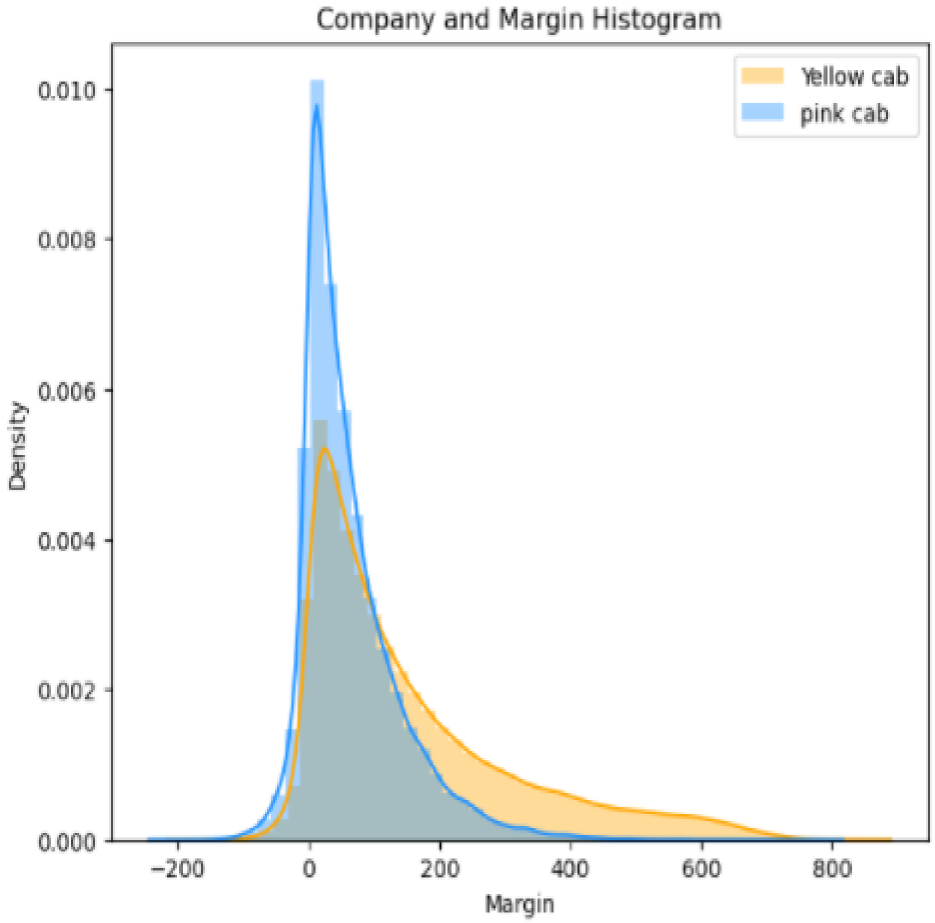


- Yellow Cab’s margin is highest in New York NY.

- Pink Cab’s Margin is higher than Yellow Cab’s Margin in Tucson AZ. And it is the only city that Pink Cab is doing better than the Yellow Cab.

- Total Margin of Yellow Cab: 274681
  Total Margin of Pink Cab: 84711.

Margin Proportionality Between Companies



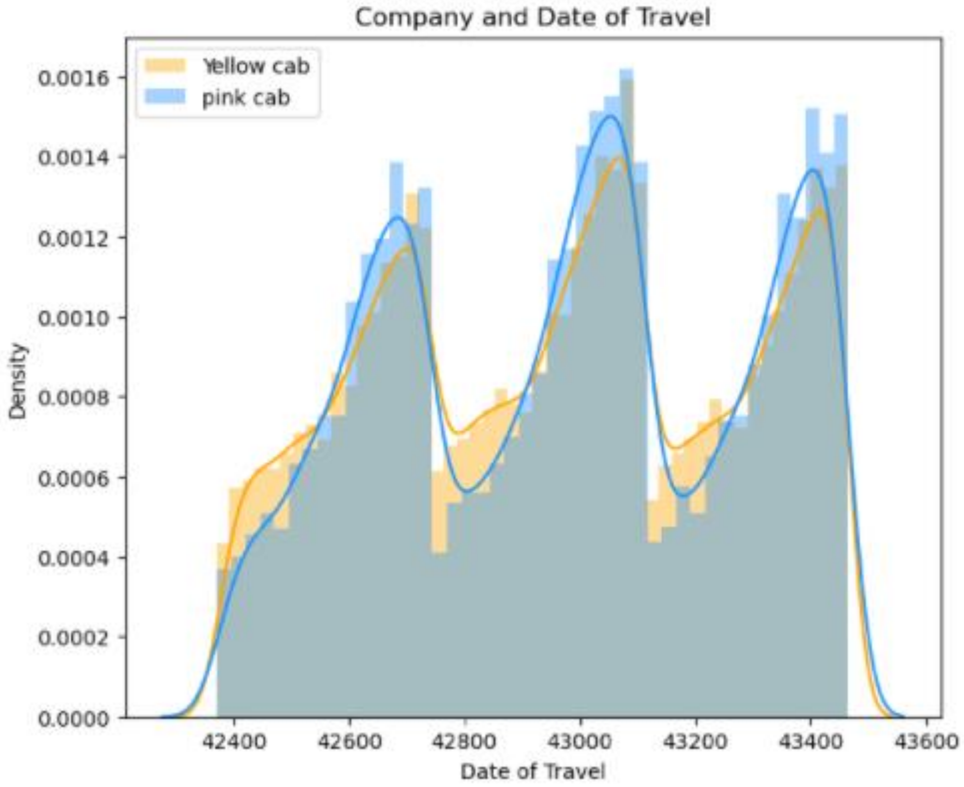
- Yellow Cab Company’s margin is higher that the Pink Cab Company.

- As Yellow Cab Company’s Users increase, then so is the margin.

Assumptions:

- While calculating the profits of the cab companies, we only consider the Cost of Trip and Price Charged features.

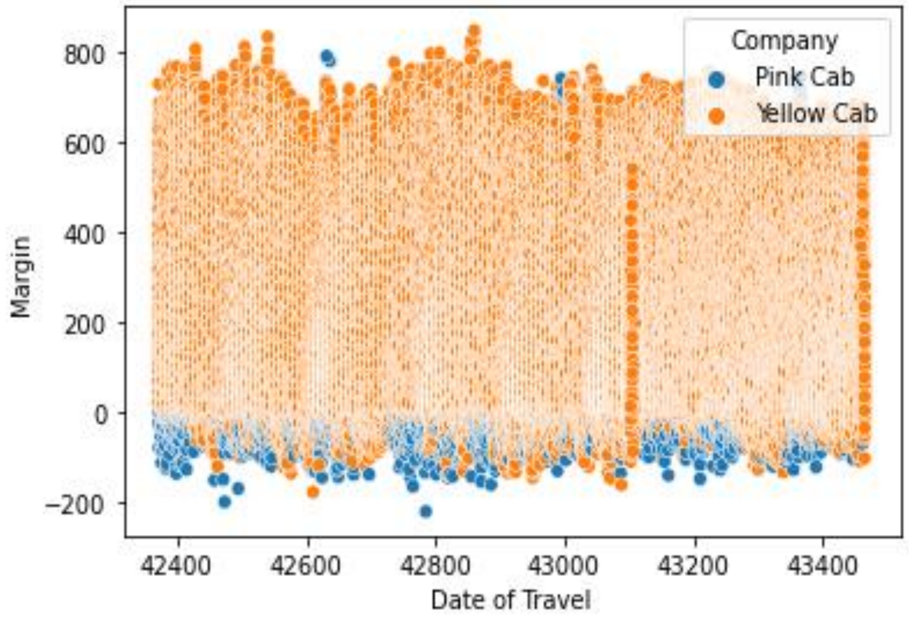
Relation between seasonality and companies



- Seasonality affects the demand to use cabs in both companies.

- It repeats similar pattern in both companies.

Relation between seasonality and companies



- As margin increases, Yellow cab is dominating the company. The cab users prefer to use Yellow Cab mostly.

Recommendations for Investment

We have examined the Yellow and Pink Cab Companies. We have the following results:

- **Customer Base:** Yellow cab company is more preferred company across 20 cities. If we consider the difference between cost of trip and charged price, Pink cab seems more customer friendly. But, unfortunately, Yellow cab is still more inclusive across customers.

- **Age and Income Base Customer Segments:** The age average in both Yellow and Pink cab are close to each other. On the other hand, lower and middle class customers of age less than or equal to 40 prefer to use Pink company. This can be the affect of being customer friendly. Yellow cab seems more inclusive in middle and upper class customers of age 40+.

Recommendations for Investment

- **Customer Satisfaction :** Yellow cab is more reliable among customers despite the fact that it has an expensive pricing policy. The percentage of the female users that prefers Yellow Cab is almost 4 times the percentage of the female users of Pink Cab Company. Similarly, the percentage of the male of Yellow Cab users is 3 times of the Pink Cab users.

- **Profit Rates**: When we compare the KM travelled values and the margin, Yellow Cab Company is doing much better than Pink Cab Company. Even in places Pink Cab Company's KM travelled values are higher than Yellow Cab Company's, the margin of Yellow Cab Company is much more higher except Tucson AZ. Yellow Cab's total margin is almost 3 times than of Pink Cab's total margin.

Based on the above discussions, we conclude that Yellow Cab dominates the majority of the market and hence, we will recommend Yellow Cab for investment.

THANK YOU

Data Glacier
Your Deep Learning Partner