G2M Insight For Cab Investment

Junfei Liu

Dec 21th 2022

Agenda

Problem Statement

Data Insights

EDA

Hypothesis Testing

Problem Statement:

- 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.

- Objective: Provide actionable insights to help XYZ firm in identifying the right company for making investment.

- Based on existing Cab Companies: Yellow Cab and Pink Cab

- The Analysis include :
  - Data understanding,
  - Data visualization,
  - Multiple hypothesis

Dataset Information:

- **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.

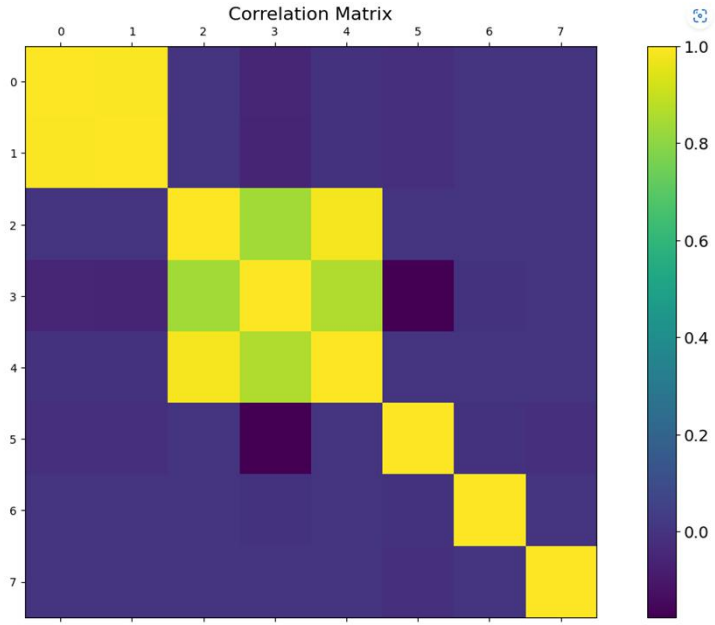
Dataset Merged

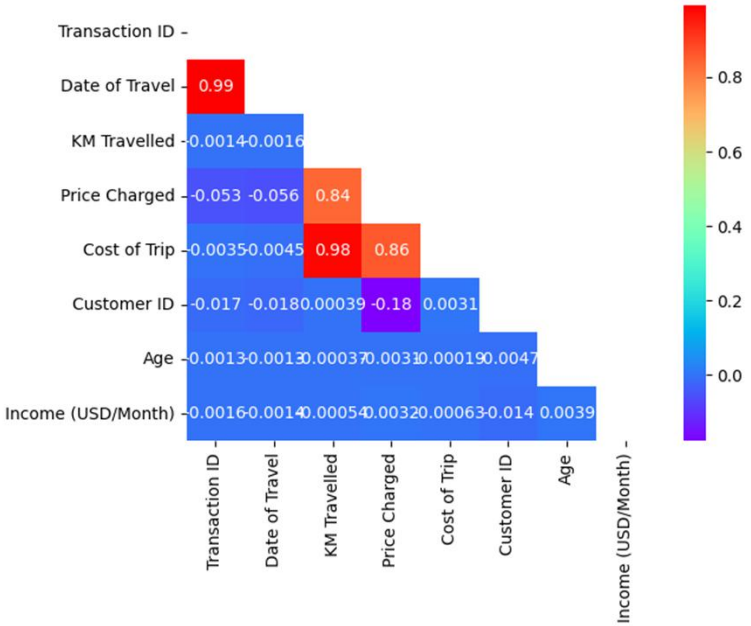
- All four datasets can be merged into one dataframe for better analysis.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 359392 entries, 0 to 359391
Data columns (total 14 columns):
 #   Column                Non-Null Count   Dtype  
---  ------                --------------   -----  
 0   Transaction ID        359392 non-null  int64  
 1   Date of Travel        359392 non-null  int64  
 2   Company               359392 non-null  object 
 3   City                  359392 non-null  object 
 4   KM Travelled          359392 non-null  float64
 5   Price Charged         359392 non-null  float64
 6   Cost of Trip          359392 non-null  float64
 7   Population            359392 non-null  object 
 8   Users                 359392 non-null  object 
 9   Customer ID           359392 non-null  int64  
 10  Payment_Mode          359392 non-null  object 
 11  Gender                359392 non-null  object 
 12  Age                   359392 non-null  int64  
 13  Income (USD/Month)    359392 non-null  int64  
dtypes: float64(3), int64(5), object(6)
memory usage: 41.1+ MB
```

	Transaction ID	Date of Travel	Company	City	KM Travelled	Price Charged	Cost of Trip	Population	Users	Customer ID	Payment_Mode	Gender	Age	Income (USD/Month)
0	10000011	42377	Pink Cab	ATLANTA GA	30.45	370.95	313.6350	814,885	24,701	29290	Card	Male	28	10813
1	10351127	43302	Yellow Cab	ATLANTA GA	26.19	598.70	317.4228	814,885	24,701	29290	Cash	Male	28	10813
2	10412921	43427	Yellow Cab	ATLANTA GA	42.55	792.05	597.4020	814,885	24,701	29290	Card	Male	28	10813
3	10000012	42375	Pink Cab	ATLANTA GA	28.62	358.52	334.8540	814,885	24,701	27703	Card	Male	27	9237
4	10320494	43211	Yellow Cab	ATLANTA GA	36.38	721.10	467.1192	814,885	24,701	27703	Card	Male	27	9237

Correlation Investigation



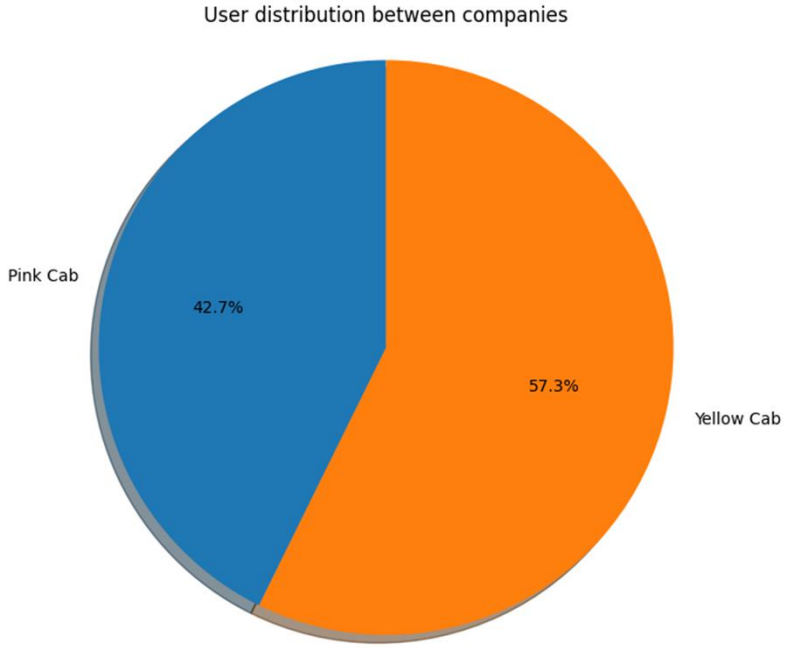


Two groups of strong correlations:

- Date of travel v.s. Transaction ID
- Price charged v.s. KM Travelled v.s. Cost of Trip

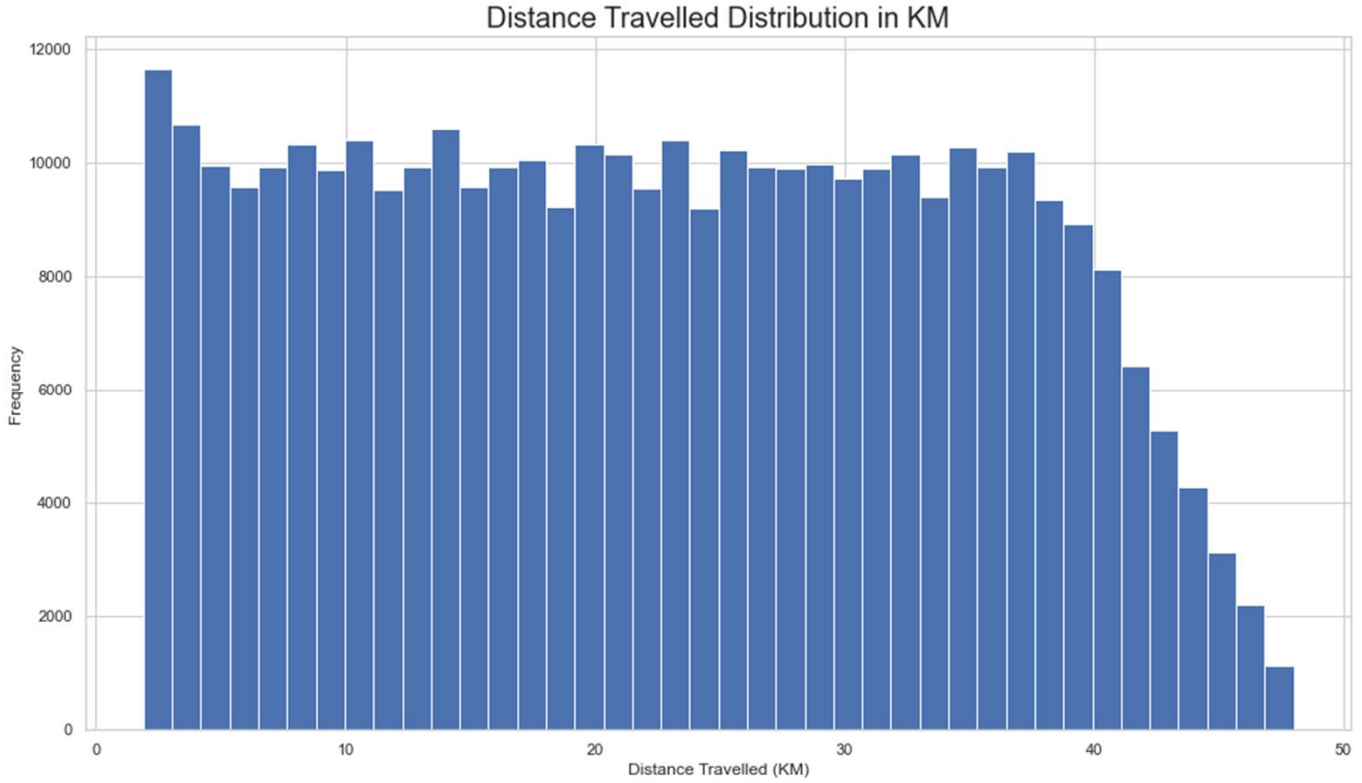
EXPLORATORY DATA ANALYSIS

1. Distribution of Users between Cabs Companies:



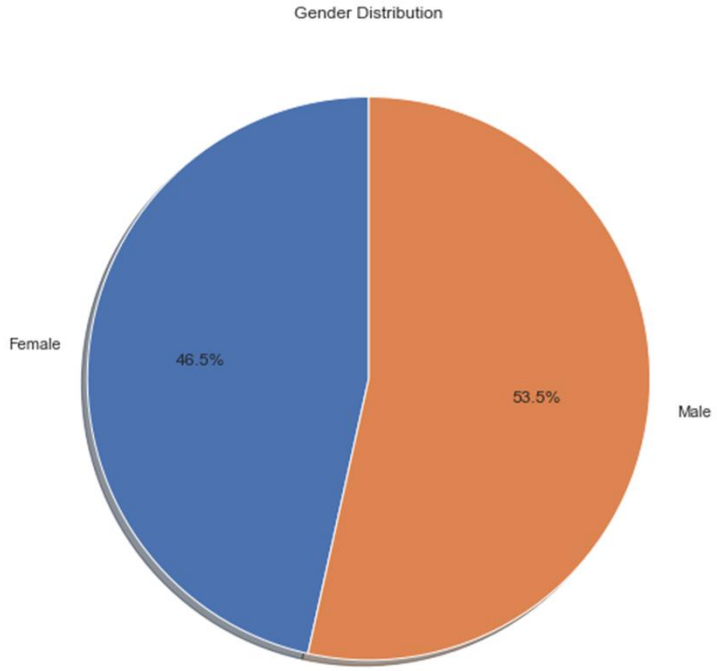
- From the figure above, we can see that yellow cab is more popular among customers.

2. Distribution of Distance Travelled in KM:



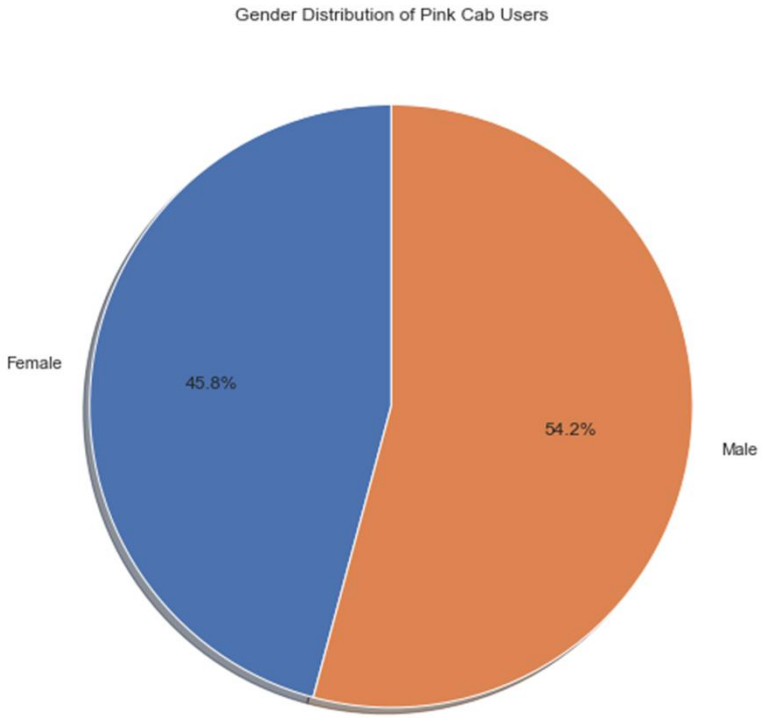
- From the figure above, most people have travelled between 2km and 40km, and we have a decreasing number of users with increasing distance travelled with over 40 km distance travelled.

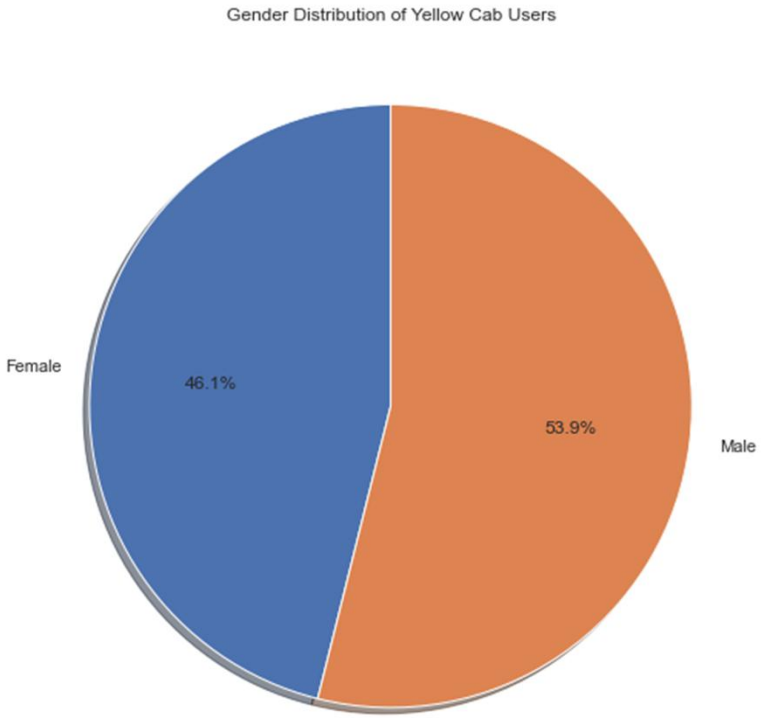
3. Distribution of Genders (total):



- From the figure above, there are close percentage of male and female users based on the user pool of both companies.

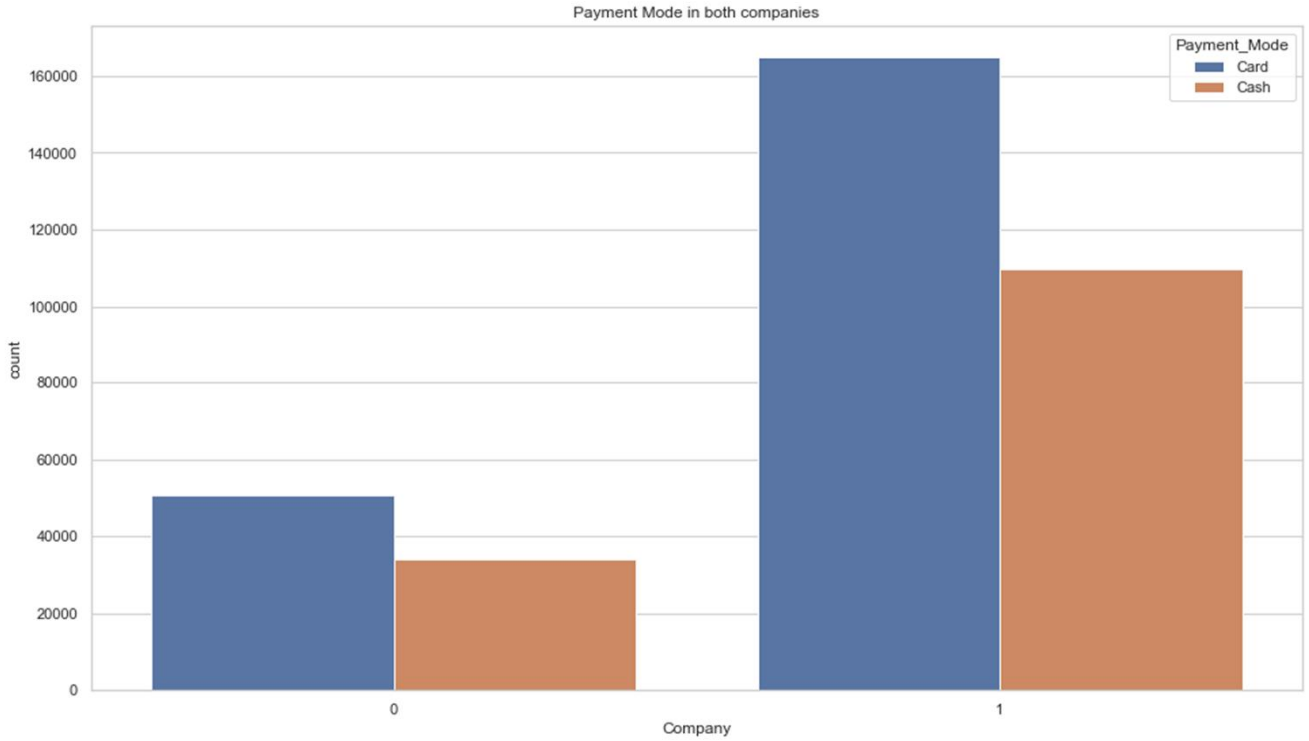
3.1. Distribution of Genders between Companies:





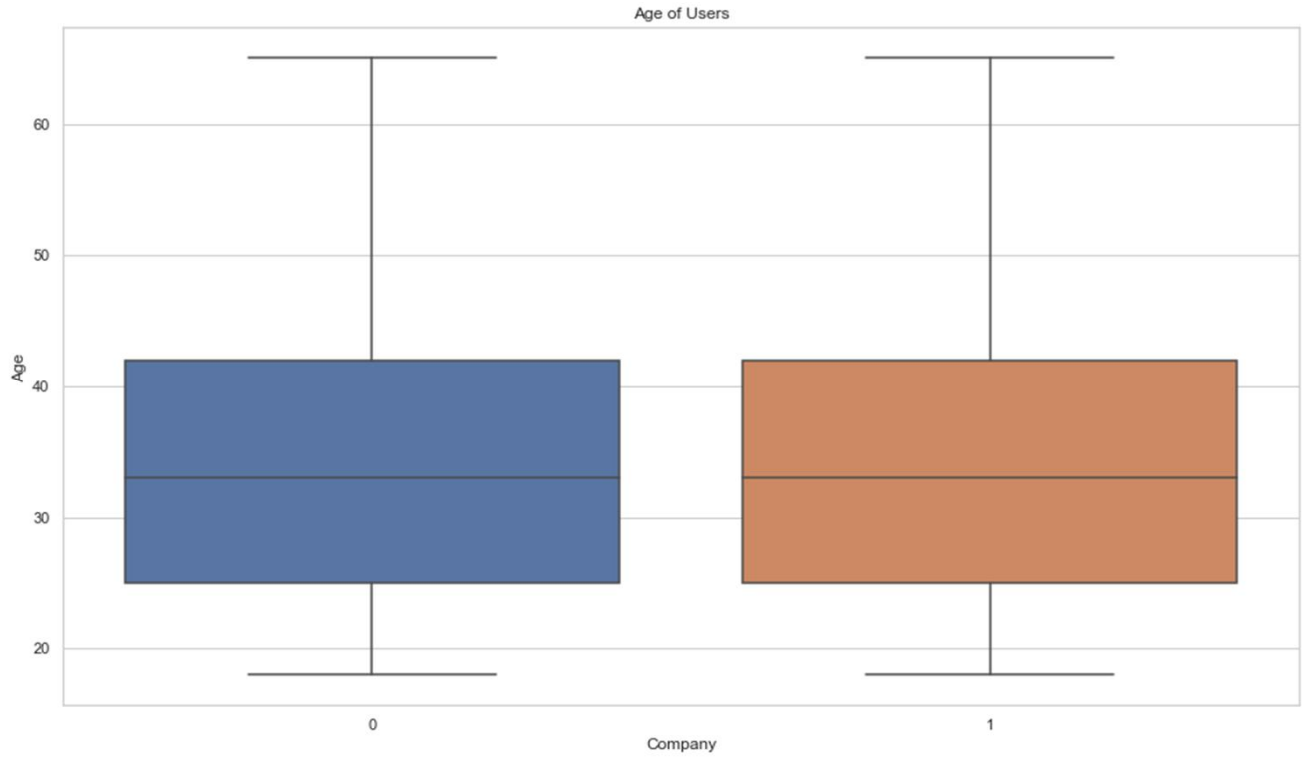
- As we can see, the gender distributions are similar for both companies.

4. Distribution of Payment Type:



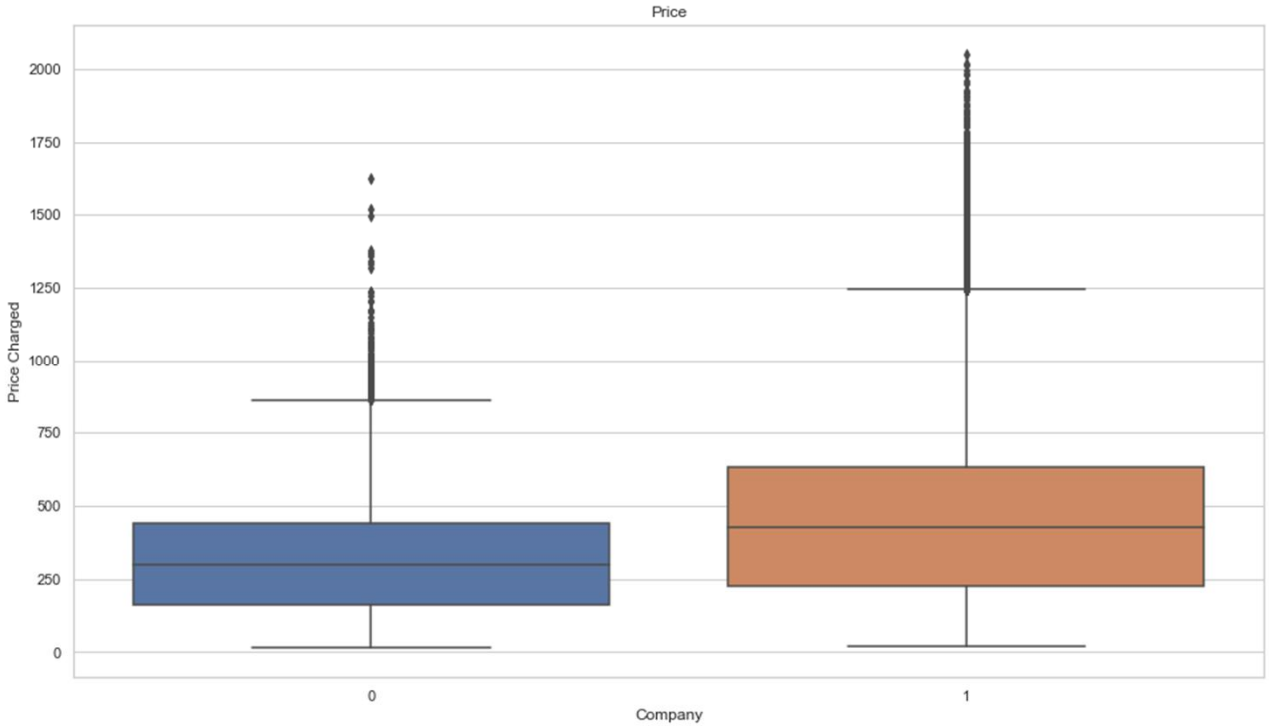
- From the figure above, we can see users prefer to pay with card than cash for both companies.

5. Distribution of User Age:



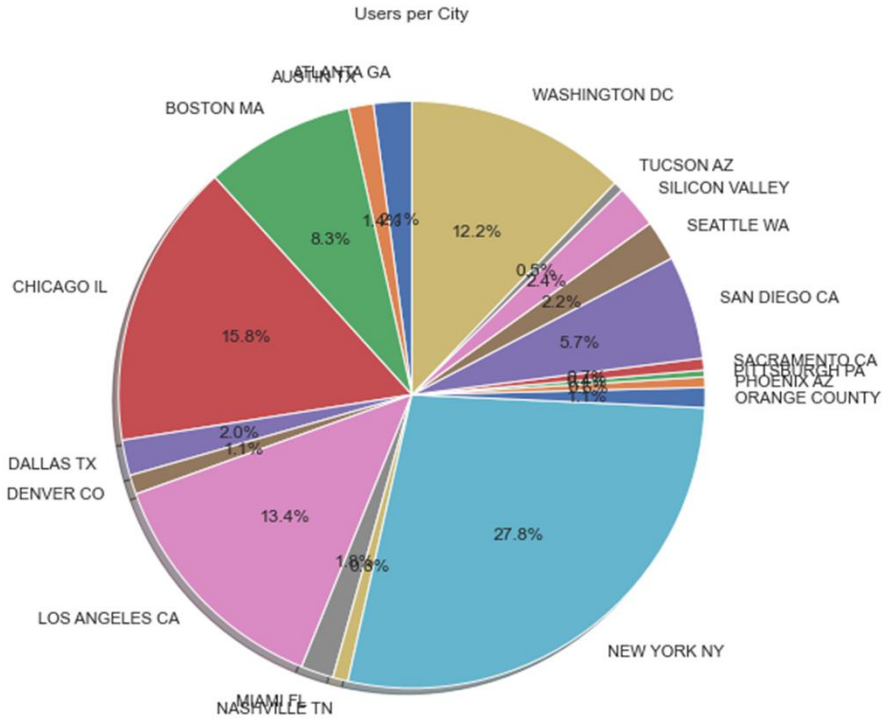
- The distribution of user ages are similar for both companies with a range from under 20 to over 60 and an average of around 35.

6. Distribution of Price Charged:



- From the figure above, most people have travelled between 2km and 40km, and we have a decreasing number of users with increasing distance travelled with over 40 km distance travelled.

7. Distribution of Cities:



- From the figure above, we can see the most users come from New York city, Chicago, Los Angeles, and Washington DC.

HYPOTHESIS TESTING

Hypothesis 1: will the profits be different for different genders?

H0: The profits will not be different for genders.
H1: The profits will be different for genders.

```
P value is  5.9218848213269 77e-37
We accept alternative hypothesis (H1)
```

- We can see that the profits are indeed different for two genders.
- Is it because both firms employ different price strategies for different genders, or is one company's strategy dominating the result?

Hypothesis 1.1: Will the test results for hypotheses 1 be different for two cab companies?

Test on Yellow Cab:

```
P value is  6.060473042494056e-25
We accept alternative hypothesis (H1)
```

Test on Pink Cab:

```
P value is  0.1151530590042 58
We accept null hypothesis (H0)
```

- We can see the profits earned by yellow cab are different for two genders, yet the profits earned by yellow cab are not significantly different for two genders.

- Through hypothesis 1, we found out different strategies of two companies on different genders.

Hypothesis2: will the profits be different for different payment modes?

H0: The profits will not be different for payment modes.
H1: The profits will be different for payment modes.

```
P value is  0.4454195660215633
We accept null hypothesis (H0)
```

- We can see that the profits are not significantly different for two payment modes.

Thank You