Data Glacier

Your Deep Learning Partner

G2M Case Study

G2M insight for Cab Investment firm

Name: Mudar Shullar

Date: 20.09.2022

Agenda

- Executive Summary
- Problem Statement
- Approach
- EDA
- EDA Summary
- Recommendations

Data Glacier
Your Deep Learning Partner

Executive Summary

- **The Client:**

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.

- **Project delivery:**

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

- **Areas to investigate:**

•Which company has maximum cab users at a particular time period?
•Does margin proportionally increase with increase in number of customers?
•What are the attributes of these customer segments?

Problem Statement

- XYZ‘s problem is that they cannot identify the right companies to invest in.

- 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.

Approach

- Datasets were imported Data Glacier’s GitHub
- Data was cleaned and the data types were checked (Understanding the field names and data types)
- After cleaning the analyzing the data, table relationships were found between each of the three table and were linked together with a primary/secondary key
- The user attribute of a city record is taken as the number of taxi users in that city, including yellow and pink taxi users
- Tables were modelled with Numpy library
- New column created called “Profit” using both "Price Charged” and “Cost of Trip"

Information about Datasets

Below are the list of datasets which are provided for the analysis:

- **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

Cab Dataset

Cab_Data_df

[3]:
```
Cab_Data_df.head(5)
```

t[3]:

	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

Data Types

```
Cab_Data_df.dtypes
```

```
Transaction ID      int64
Date of Travel      int64
Company            object
City               object
KM Travelled      float64
Price Charged     float64
Cost of Trip      float64
dtype: object
```

After returning the data types of each column in Cab_Data_df, we've noticed that the column "Date of Travel" should be changed. The array of dates in our example are dates extracted from an excel file, Each represents the days after the base_date (on/about 1899-12-29).

City Dataset

Data Info

```
City_df.head(5)
```

	City	Population	Users
0	NEW YORK NY	8,405,837	302,149
1	CHICAGO IL	1,955,130	164,468
2	LOS ANGELES CA	1,595,037	144,132
3	MIAMI FL	1,339,155	17,675
4	SILICON VALLEY	1,177,609	27,247

```
City_df.dtypes
```

```
City          object
Population    object
Users         object
dtype: object
```

After returning the data types for City_df, we see that we need to change the data types of both Population and Users from object to float64.

Customer Dataset

```
Customer_ID_df.head(5)
```

	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

```
Customer_ID_df.dtypes
```

```
Customer ID            int64
Gender                object
Age                    int64
Income (USD/Month)     int64
dtype: object
```

Transaction Dataset

Transaction_ID_df

```
Transaction_ID_df.head(5)
```

	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

```
Transaction_ID_df.dtypes
```

```
Transaction ID     int64
Customer ID        int64
Payment_Mode      object
dtype: object
```

Table Relationships in Datasets

Merge the data sets

```
df_merge1 = Cab_Data_df.merge(Transaction_ID_df, on = "Transaction ID")
df_merge1
```

	Transaction ID	Date of Travel	Company	City	KM Travelled	Price Charged	Cost of Trip	Customer ID	Payment_Mode
0	10000011	2017-01-06	Pink Cab	ATLANTA GA	30.45	370.95	313.6350	29290	Card
1	10000012	2017-01-04	Pink Cab	ATLANTA GA	28.62	358.52	334.8540	27703	Card
2	10000013	2016-12-31	Pink Cab	ATLANTA GA	9.04	125.20	97.6320	28712	Cash
3	10000014	2017-01-05	Pink Cab	ATLANTA GA	33.17	377.40	351.6020	28020	Cash
4	10000015	2017-01-01	Pink Cab	ATLANTA GA	8.73	114.62	97.7760	27182	Card
...	...	...	...	...	...	...	...	...	...

```
df_merge2 = df_merge1.merge(Customer_ID_df, on = "Customer ID")
df_merge2.head(5)
```

	Transaction ID	Date of Travel	Company	City	KM Travelled	Price Charged
0	10000011	2017-01-06	Pink Cab	ATLANTA GA	30.45	370.95
1	10351127	2019-07-20	Yellow Cab	ATLANTA GA	26.19	598.70
2	10412921	2019-11-22	Yellow Cab	ATLANTA GA	42.55	792.05
3	10000012	2017-01-04	Pink Cab	ATLANTA GA	28.62	358.52
4	10320494	2019-04-20	Yellow Cab	ATLANTA GA	36.38	721.10

```
df_master_data = df_merge2.merge(City_df, on = "City")
df_master_data.head(5)
```

	Transaction ID	Date of Travel	Company	City	KM Travelled	Price Charged	Cost of Trip
0	10000011	2017-01-06	Pink Cab	ATLANTA GA	30.45	370.95	313.6350
1	10351127	2019-07-20	Yellow Cab	ATLANTA GA	26.19	598.70	317.4228
2	10412921	2019-11-22	Yellow Cab	ATLANTA GA	42.55	792.05	597.4020
3	10000012	2017-01-04	Pink Cab	ATLANTA GA	28.62	358.52	334.8540
4	10320494	2019-04-20	Yellow Cab	ATLANTA GA	36.38	721.10	467.1192

Column Information (Merged Dataset)

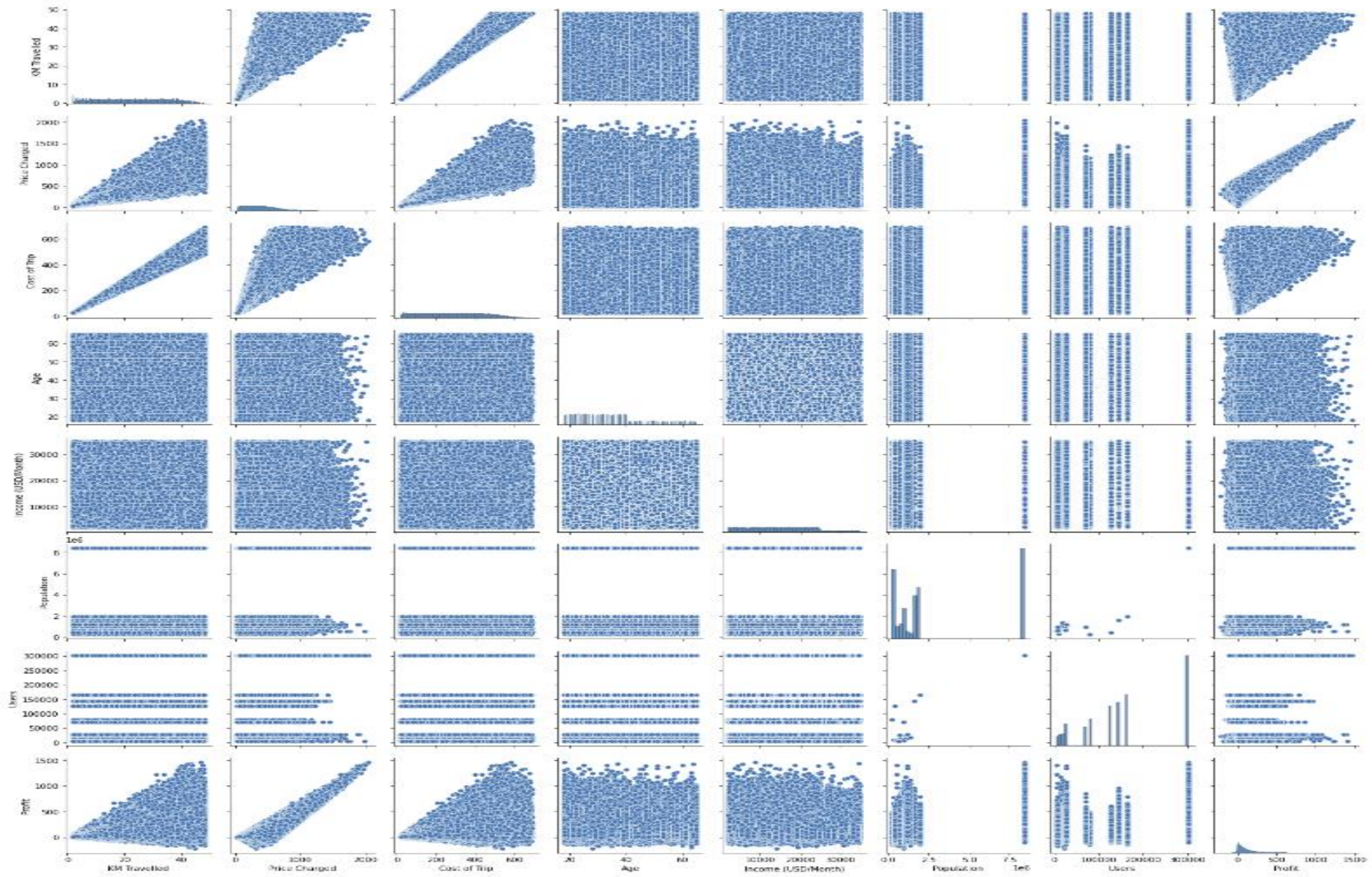
```
df_master_data.dtypes
```

```
Transaction ID          int64
Date of Travel         object
Company                object
City                   object
KM Travelled          float64
Price Charged         float64
Cost of Trip          float64
Customer ID             int64
Payment_Mode           object
Gender                 object
Age                     int64
Income (USD/Month)      int64
Population            float64
Users                 float64
Profit                float64
dtype: object
```

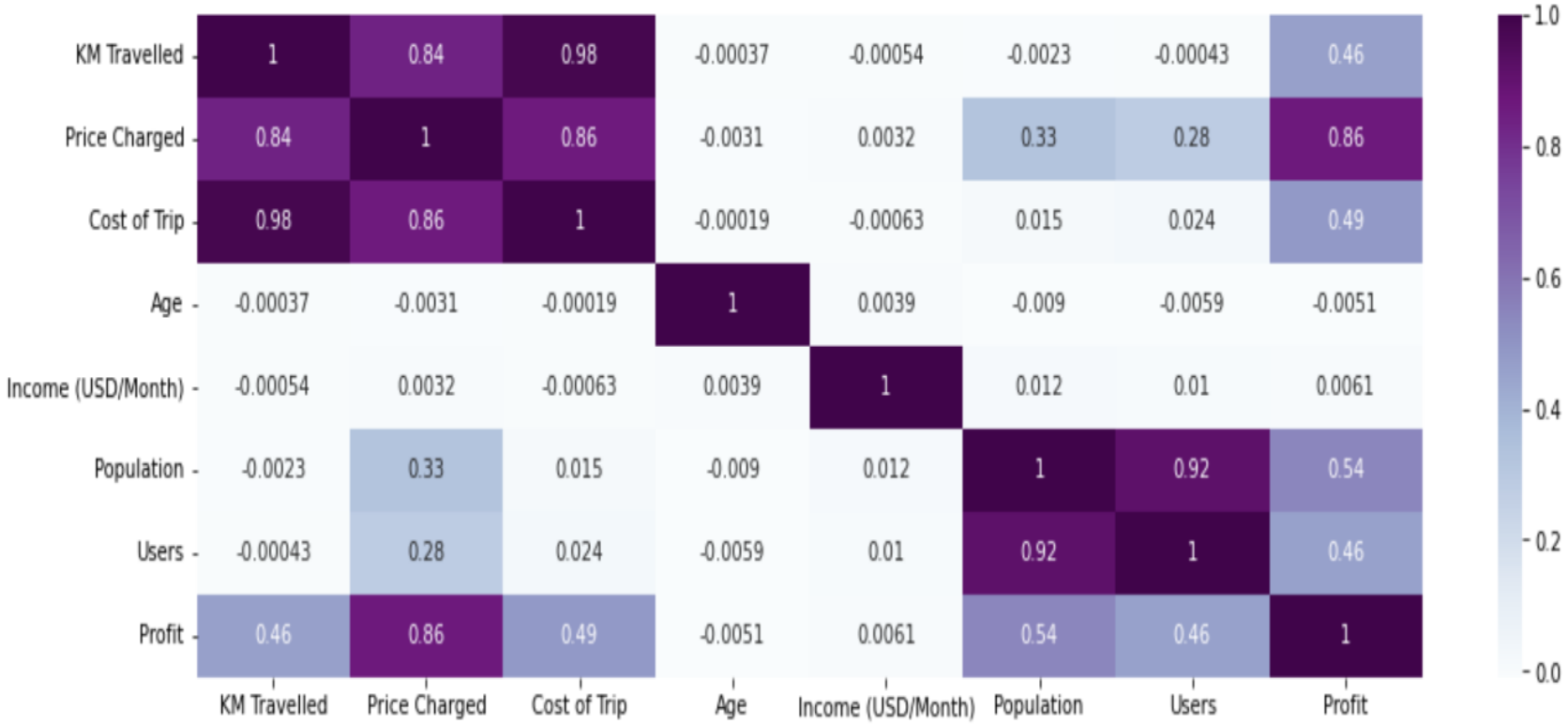
No NULL values are available in our newly merged data set.

Relationships Between Variables (Relationship Analysis)

```
[41]: sns.pairplot(data = df_master_data[["KM Travelled", "Price Charged", "Cost of Trip", "Age",
                                          "Income (USD/Month)", "Population", "Users", "Profit"]])
      plt.show()
```



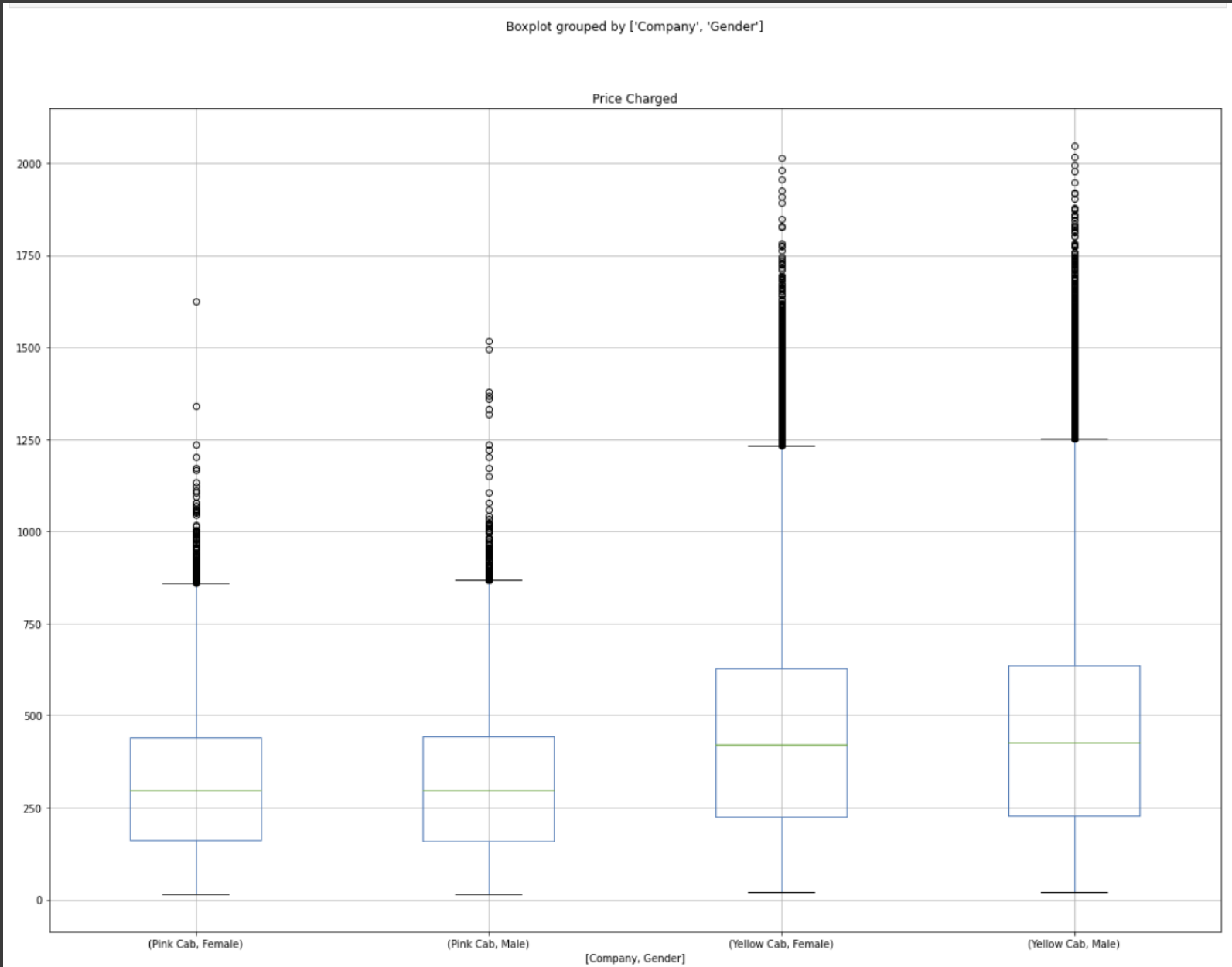
Correlation Between Variables (Relationship Analysis)



We can conclude from the executed heatmap on our master data, that there's a strong positive correlation between the variables "KM Travelled", "Price Charged" and "Cost of Trip", also there is a strong positive correlation between "Users" and "Population". There's also a strog correlation between "Price Charged" and "Profit", which is normal in real life situations. We can also conclude that "Age" does not correlate to higher "Income(USD/Month)"

Box-Plot (Price Charged)

We conclude from this boxplot that Yellow Cab's prices are higher than Pink Cab's drivers and male taxi drivers at Yellow Cab demand higher taxi fares than female drivers at the same company.

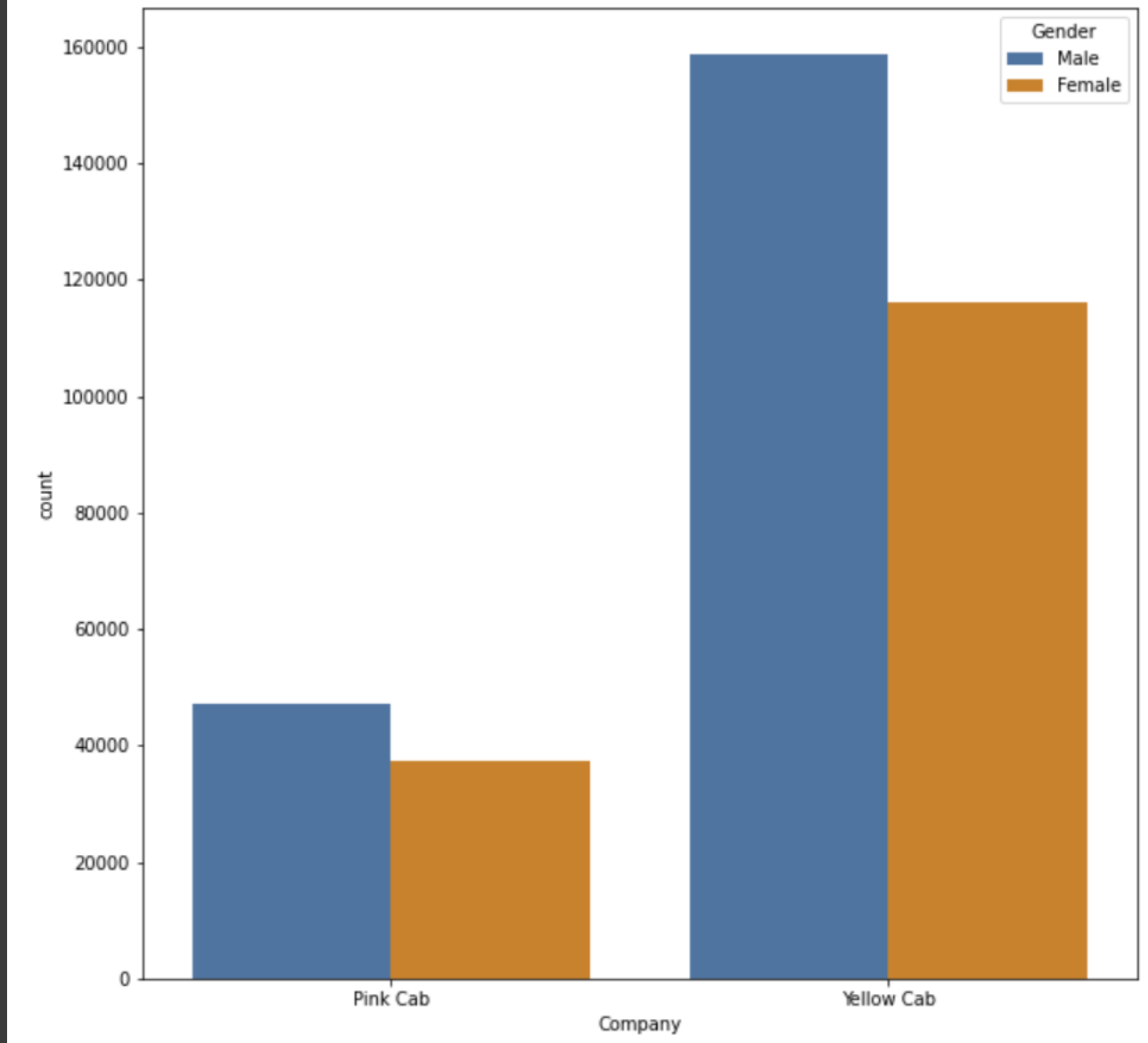


We conclude from this boxplot that Yellow Cab's prices are higher than Pink Cab's drivers and male taxi drivers at Yellow Cab demand higher taxi fares than female drivers at the same company.

Data Glacier
Your Deep Learning Partner

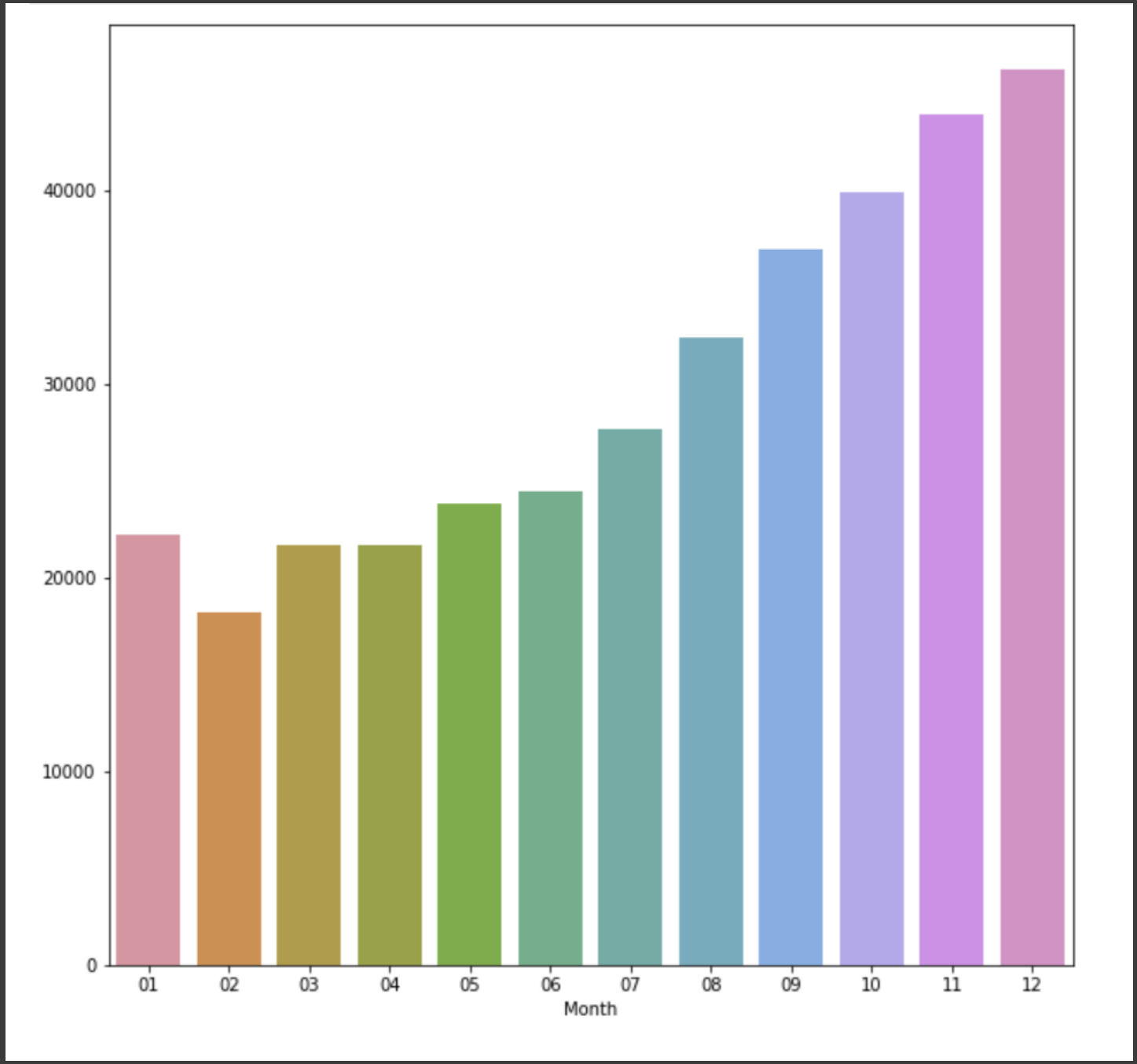
Gender Ratio By Each Company

```
plt.figure(figsize = (10, 10))
ax = sns.countplot(x = "Company", hue = "Gender", data = df_master_data)
plt.show()
```



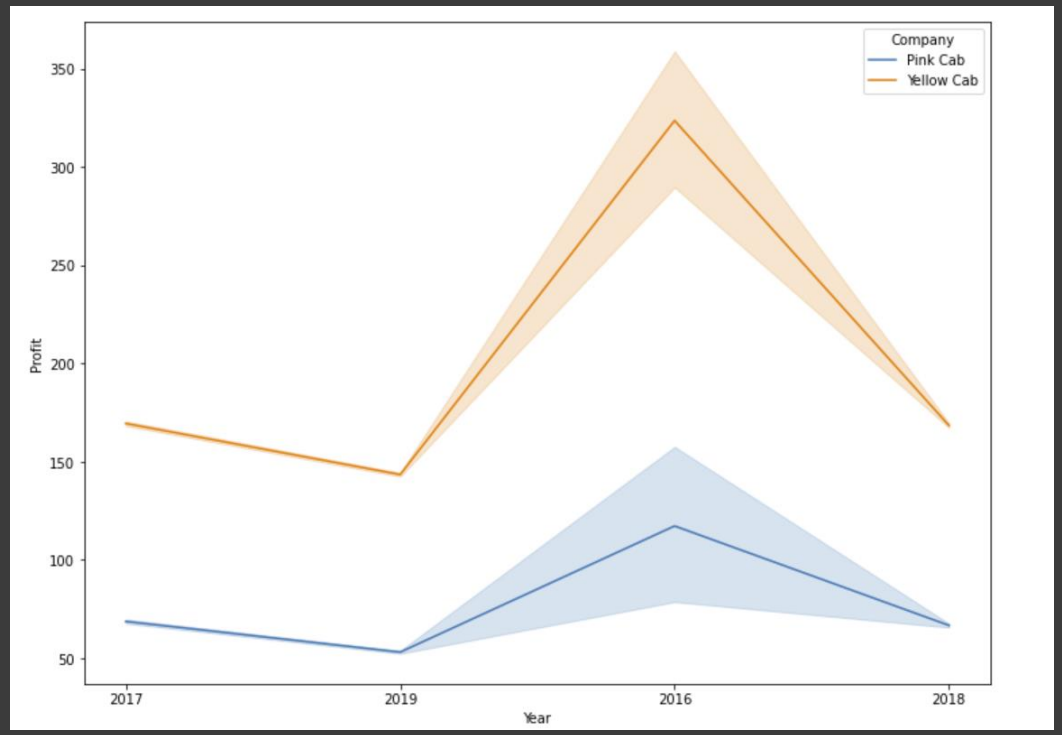
Data Glacier
Your Deep Learning Partner

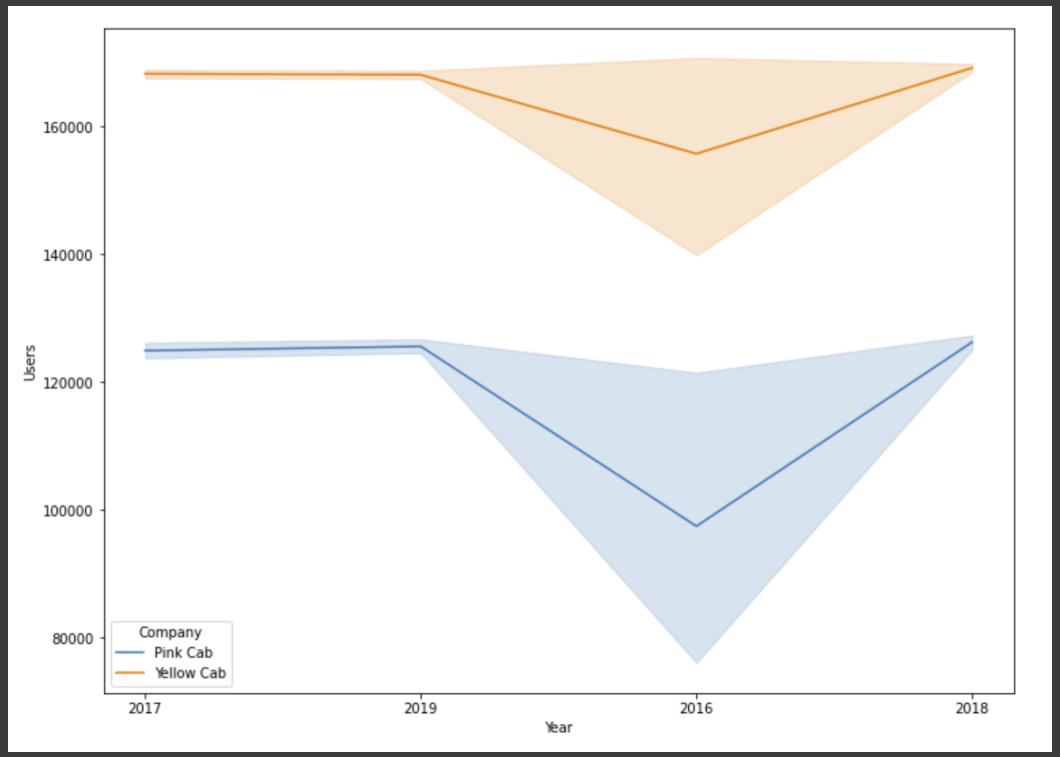
Which company has maximum cab users at a particular time period?



Data Glacier
Your Deep Learning Partner

Does margin proportionally increase with increase in number of customers?





As we can see here, margin does not proportionally increase with increase in number of customer in both taxi companies.

Data Glacier
Your Deep Learning Partner

What are the attributes of these customer segments? e.g., gender and age

```
yearly_cal = df_master_data.groupby(["Year","Company"])["Users"].count().to_frame()
yearly_cal
```

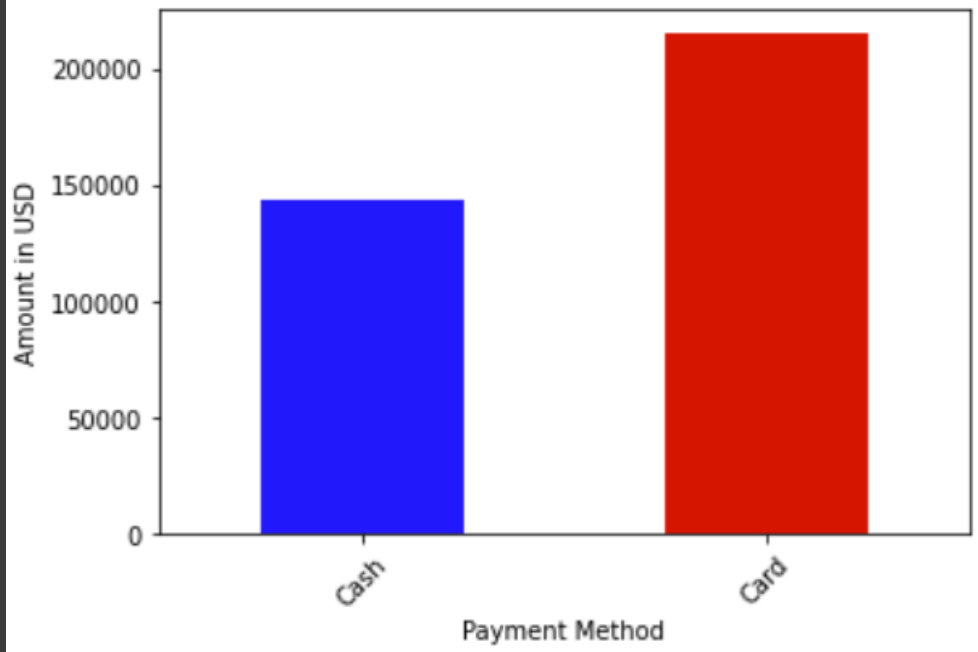
		Users
Year	Company	
2016	Pink Cab	41
	Yellow Cab	140
2017	Pink Cab	25244
	Yellow Cab	82797
2018	Pink Cab	30205
	Yellow Cab	97759
2019	Pink Cab	29221
	Yellow Cab	93985

```
gender_cal = df_master_data.groupby(["Gender","Company"])["Users"].count().to_frame()
gender_cal
```

		Users
Gender	Company	
Female	Pink Cab	37480
	Yellow Cab	116000
Male	Pink Cab	47231
	Yellow Cab	158681

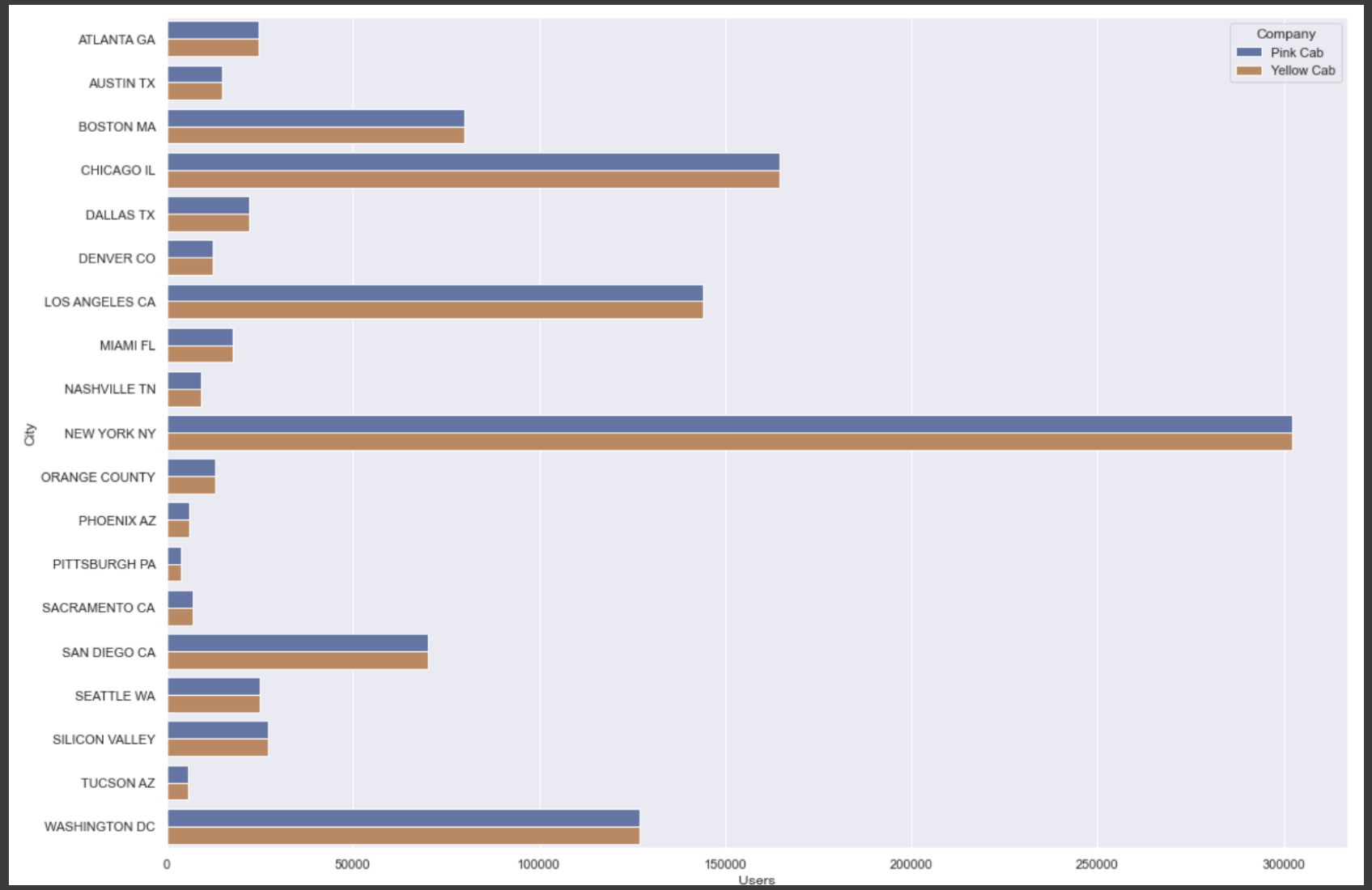
Payment Method

```
df_master_data["Payment_Mode"].value_counts(ascending=True).plot(kind='bar', color=["blue", "red"], rot=45)
plt.xlabel("Payment Method")
plt.ylabel("Amount in USD")
plt.show()
```



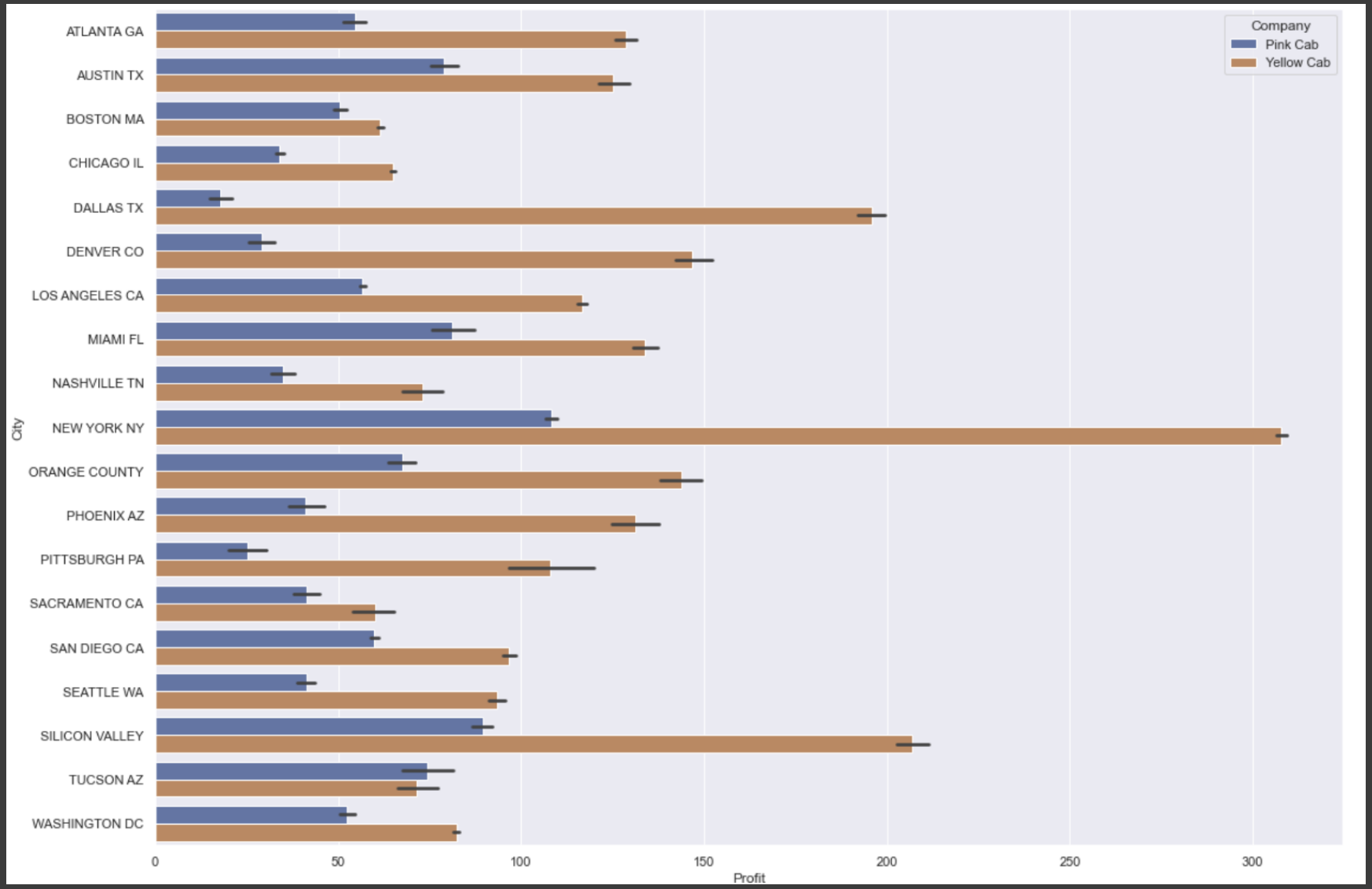
More customer pay with card than cash.

User Distribution in Each City



Data Glacier
Your Deep Learning Partner

Profit Ratio by City



Data Glacier
Your Deep Learning Partner

Price Charged With Respect to Distance

As we can see there is a linear relationship between KM traveled and Price Charged as we expected. However, Yellow Cab has high charges compared to Pink.

Data Glacier
Your Deep Learning Partner

Hypothesis 1

- **Hypothesis 1 → Is there any difference in profit regarding the time of the year?**

H0: There is no difference in profit regarding time of the year for Pink Cab company

H1: there is difference in profit regarding time of the year for Pink Cab company

```
27356 17516
P value is  5.4550802023130026e-55
Hypothesis H1 is accepted i.e. there is difference in profit regarding time of the year for Yellow Cab company.
```

```
9647 4228
P value is  2.0442096649570663e-38
Hypothesis H1 is accepted i.e. there is difference in profit regarding time of the year for Pink Cab company.
```

Hypothesis 2

- Hypothesis 2 → Is there any difference in profit regarding age?

H0: There is no difference in profit regarding age.

H1: There is difference in profit regarding age.

```
0 272123
P value is  nan
Hypothesis H0 is accepted i.e. there is no difference in profit regarding age for Yellow Cab company.
```

```
0 83890
P value is  nan
Hypothesis H0 is accepted i.e. there is no difference in profit regarding age for Pink Cab company.
```

Hypothesis 3

Hypothesis 3 → Is there any difference in profit regarding gender?

H0: There is no difference in profit regarding gender.

H1: There is difference in profit regarding gender.

```
47231 0
P value is  nan
Hypothesis H0 is accepted i.e. there is no difference in profit regarding gender for Pink Cab company.
```

```
158681 116000
P value is  6.060473042494144e-25
Hypothesis H1 is accepted i.e. there is difference in profit regarding gender for Yellow Cab company.
```

EDA Summary

Yellow cab company is a better investment choice than the pink cab company, due to the following reasons:

- Profit Margin
- More Users
- More transactions per Year

Thank You