Data Glacier

Your Deep Learning Partner

G2M Insight For Cab Investment Firm

Atakan ÖZDİN

9, October 2021

Outline

- Problem Statement
- Datasets Information
- History of the Datasets
- EDA
- Conclusion

Problem Statement –G2M Cab Industry Case Study

- 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.

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

- <u>Cab Companies</u>

  Yellow Cab

  Pink Cab

Datasets Information

There are 4 datasets:

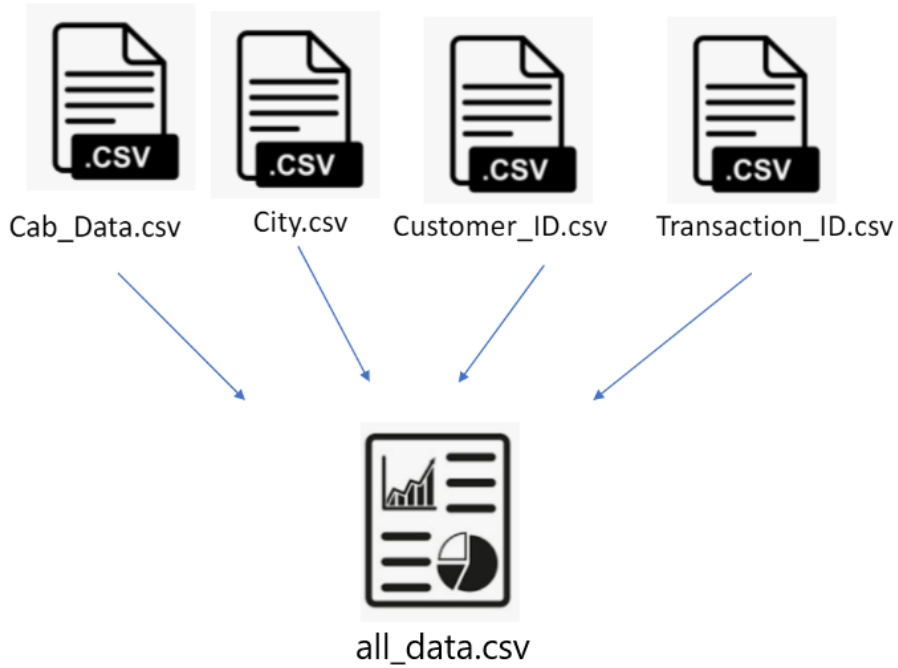
Cab_Data.csv - This file includes details of transaction for 2 cab companies.

City.csv - This file contains list of US cities, their population and number of cab users.

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.

- Timeframe of the data: **2016/01/02** to **2018/12/31**.
- Total data points : 356,392



History of the Dataset - Cab Data.csv

```
cab_df.info()
```

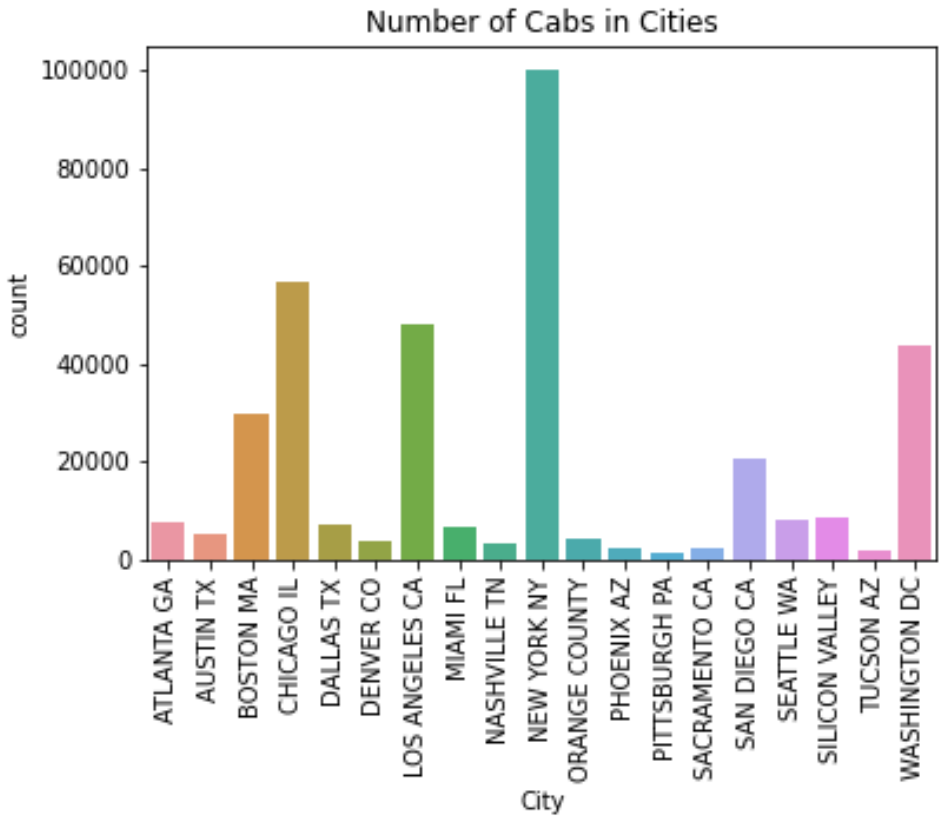
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 359392 entries, 0 to 359391
Data columns (total 7 columns):
 #   Column          Non-Null Count   Dtype         
---  ------          --------------   -----         
 0   Transaction ID  359392 non-null  int64         
 1   Date of Travel  359392 non-null  datetime64[ns]
 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       
dtypes: datetime64[ns](1), float64(3), int64(1), object(2)
memory usage: 21.9+ MB
```

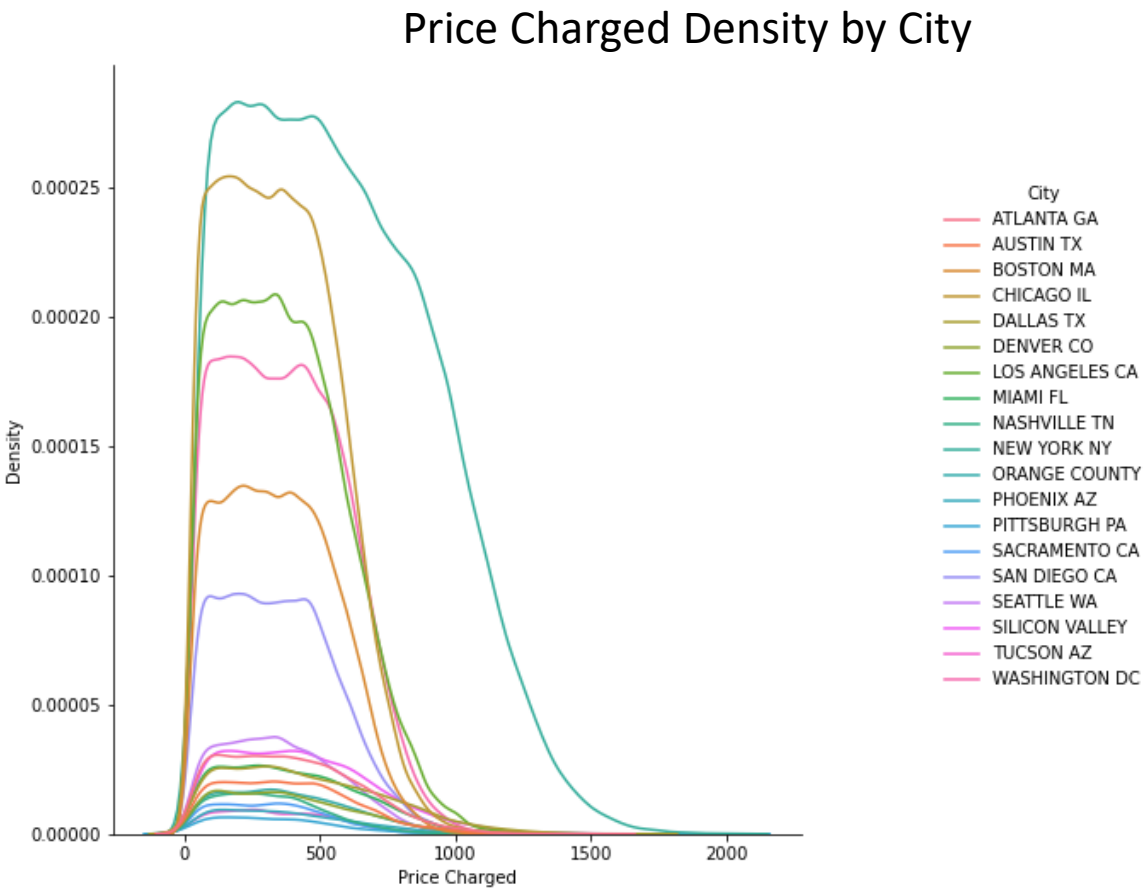
```
cab_df.isnull().sum()
```

```
Transaction ID    0
Date of Travel    0
Company           0
City              0
KM Travelled      0
Price Charged     0
Cost of Trip      0
dtype: int64
```

- As you see, there are 359392 data points.
- There is no NA value.

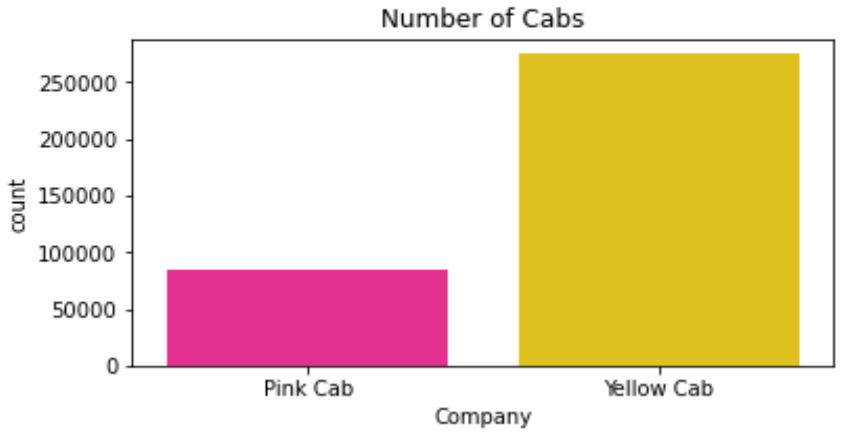
History of the Dataset - Cab Data.csv



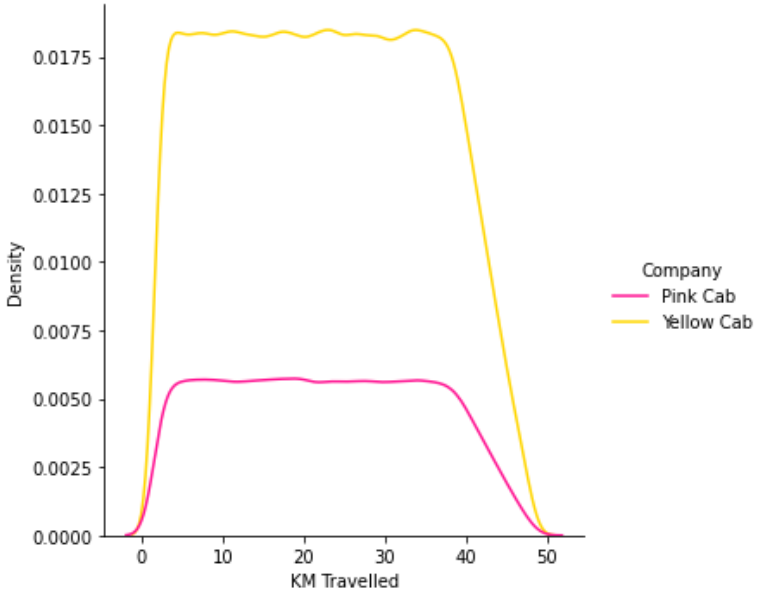


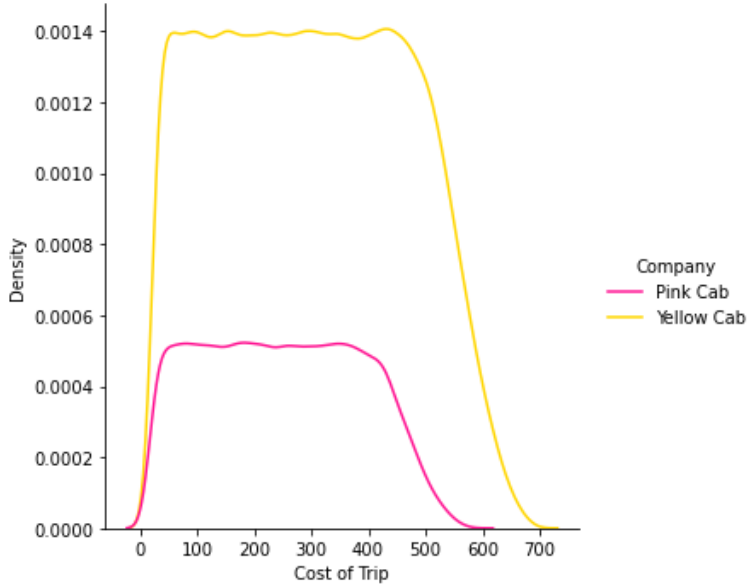
History of the Dataset - Cab Data.csv

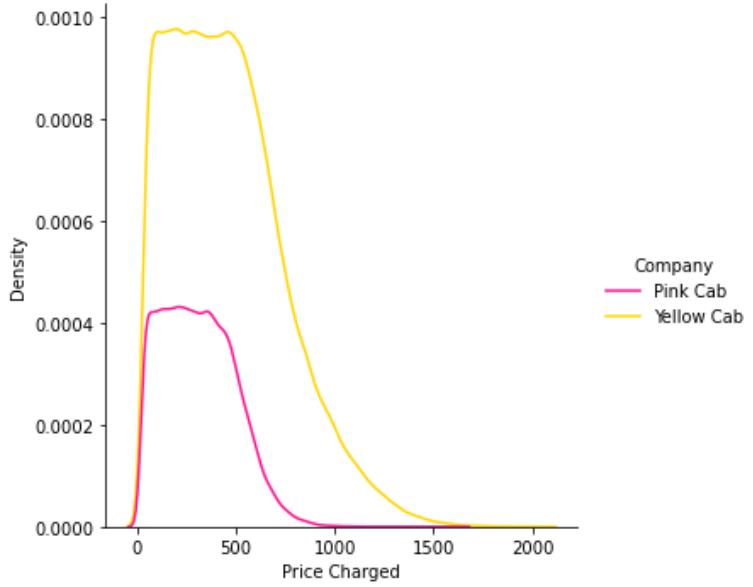
```
Yellow Cab     274681
Pink Cab        84711
Name: Company, dtype: int64
```



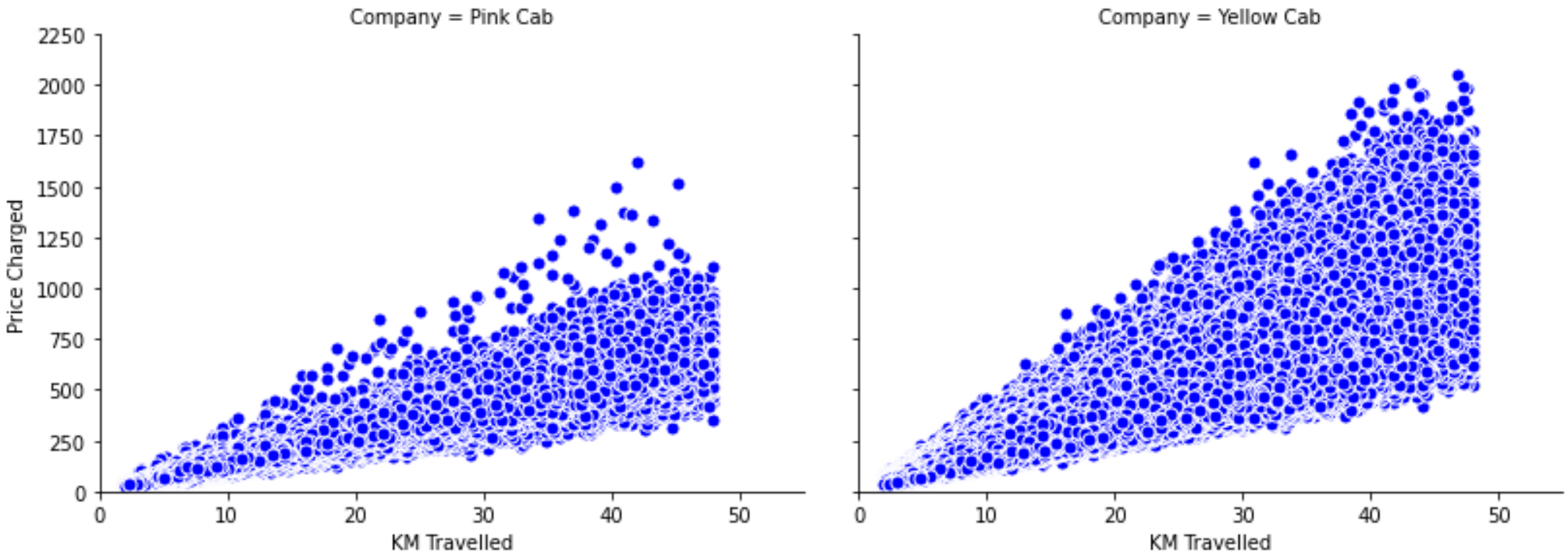
- The number of Yellow Cab is considerably higher than the Pink Cab.
- Due to the number of Yellow Cab it is seen that the number of kilometers traveled is more.







History of the Dataset - Cab Data.csv



In terms of the KM Travelled and Price Charged, it is seen that the yellow taxi travels more.

History of the Dataset - City Data.csv

```
city_df.info()
```

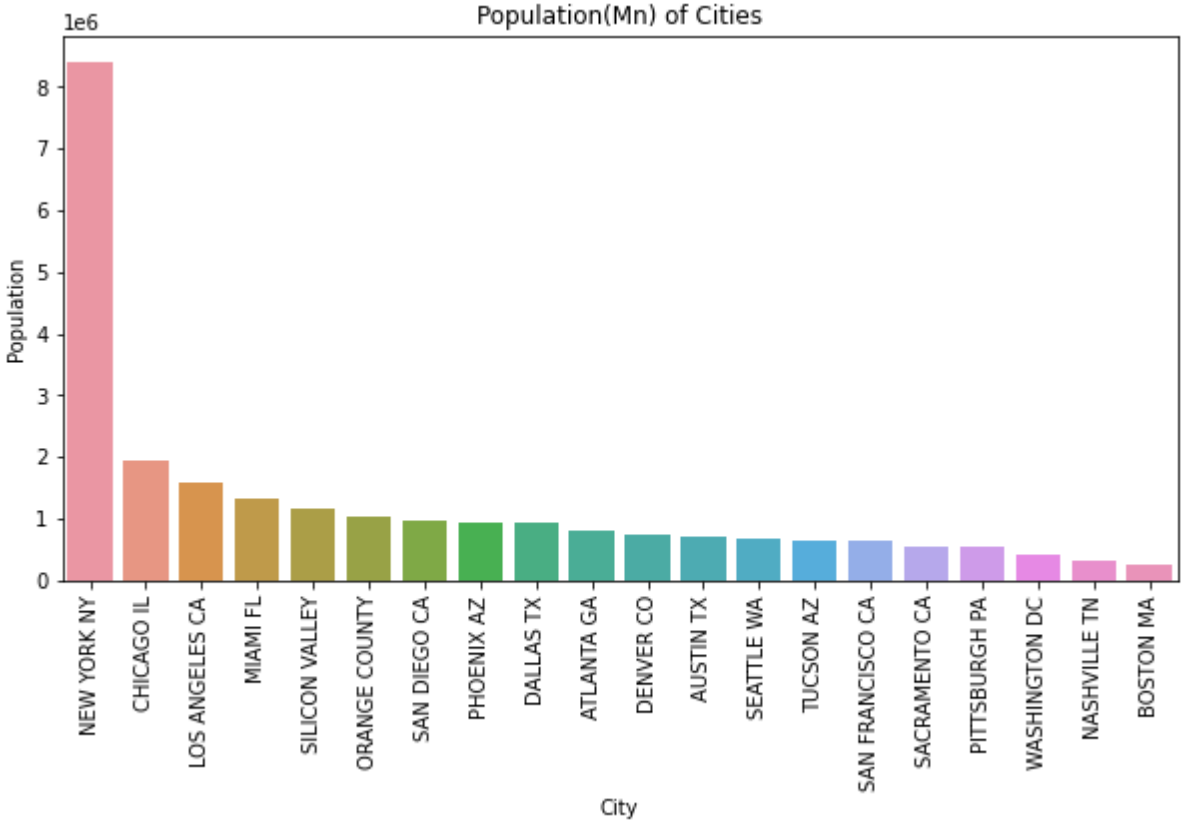
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 20 entries, 0 to 19
Data columns (total 3 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   City        20 non-null     object
 1   Population  20 non-null     object
 2   Users       20 non-null     object
dtypes: object(3)
memory usage: 640.0+ bytes
```

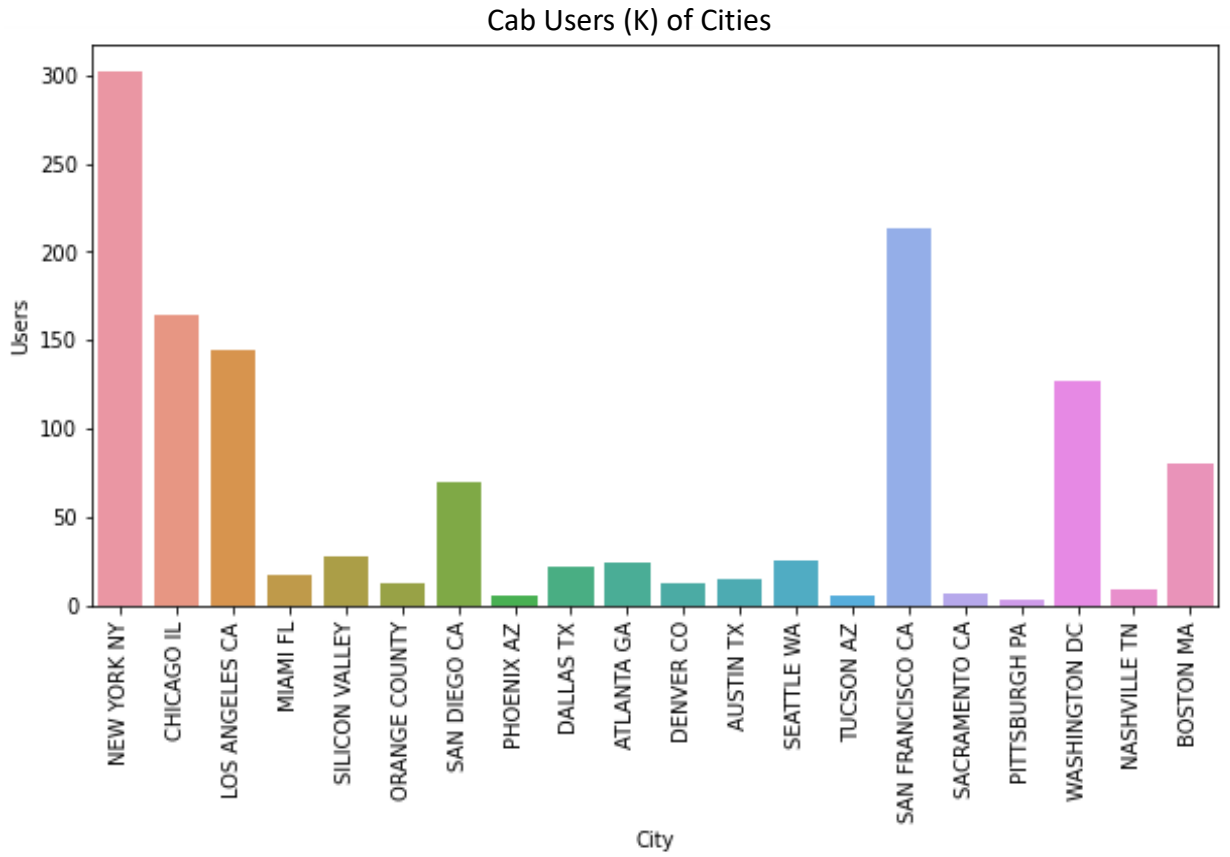
```
city_df.isnull().sum()
```

```
City          0
Population    0
Users         0
dtype: int64
```

- As you see, there are 20 data points.
- There is no NA value.

History of the Dataset - City Data.csv





History of the Dataset - Customer_ID.csv

```
customer_df.info()
```

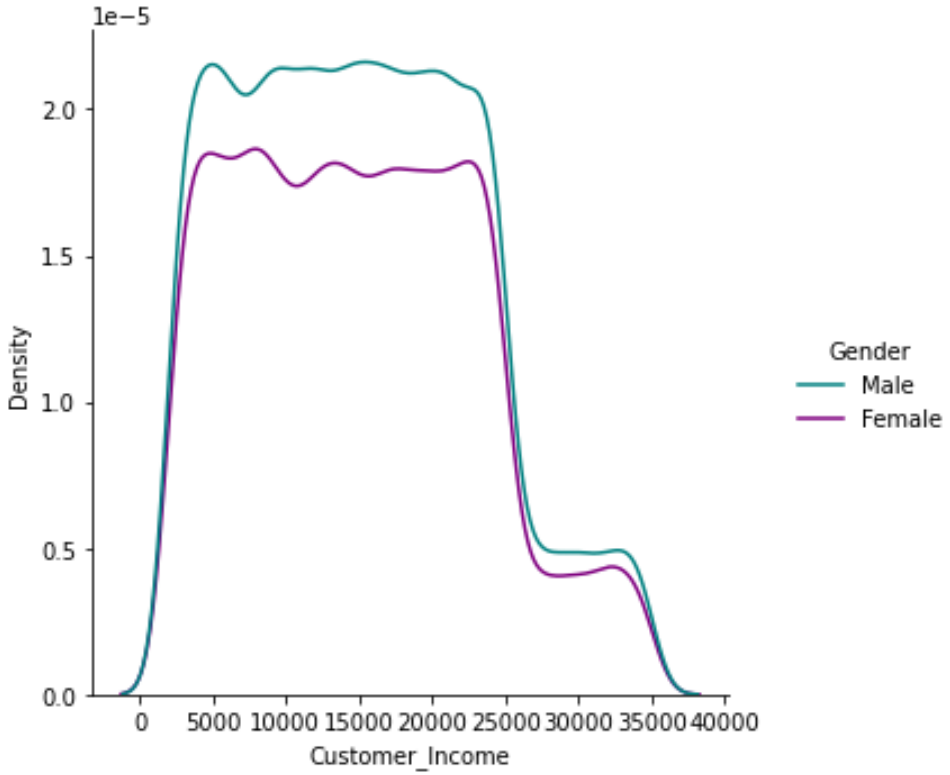
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 49171 entries, 0 to 49170
Data columns (total 4 columns):
 #   Column              Non-Null Count  Dtype 
---  ------              --------------  ----- 
 0   Customer ID         49171 non-null  int64 
 1   Gender              49171 non-null  object
 2   Age                 49171 non-null  int64 
 3   Income (USD/Month)  49171 non-null  int64 
dtypes: int64(3), object(1)
memory usage: 1.9+ MB
```

```
customer_df.isnull().sum()
```

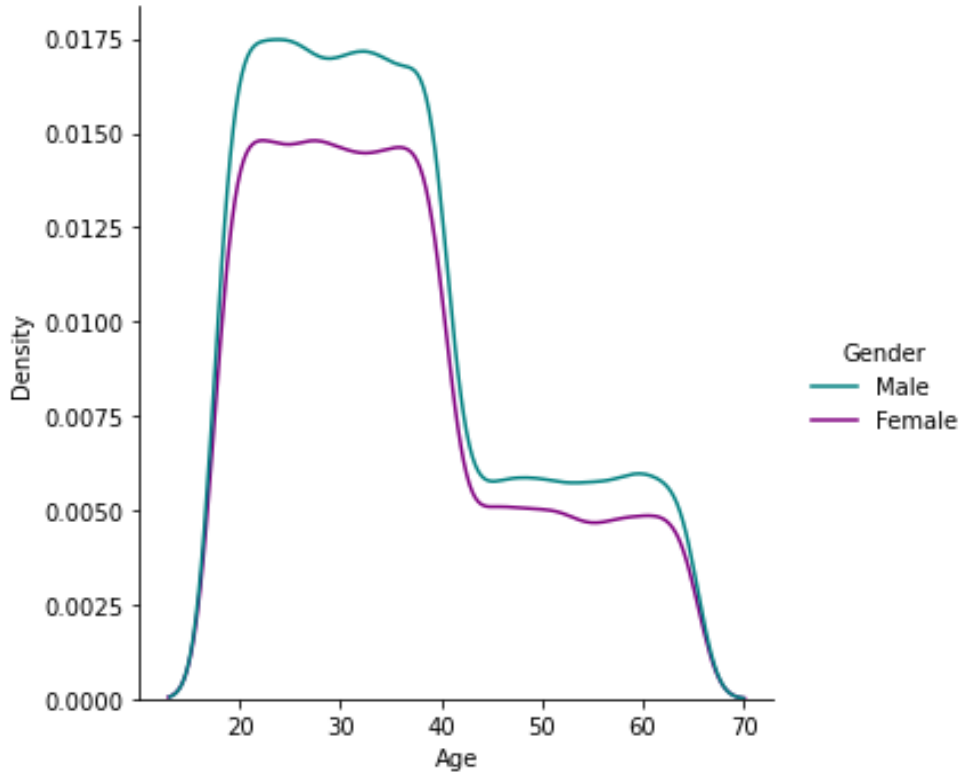
```
Customer ID         0
Gender              0
Age                 0
Customer_Income     0
dtype: int64
```

- As you see, there are 49171 data points.
- There is no NA value.

History of the Dataset - Customer_ID.csv



- Income of customers using cabs is between 4K and 27K.
- In general, it is seen that male users are more.



- Age of customers using cabs is approximately between 20 and 42.
- In general, it is seen that male users are more.

History of the Dataset - Transaction_ID.csv

```
transaction_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 440098 entries, 0 to 440097
Data columns (total 3 columns):
 #   Column          Non-Null Count   Dtype 
---  ------          --------------   ----- 
 0   Transaction ID  440098 non-null  int64 
 1   Customer ID     440098 non-null  int64 
 2   Payment_Mode    440098 non-null  object
dtypes: int64(2), object(1)
memory usage: 13.4+ MB
```

```
transaction_df.isnull().sum()
```

```
Transaction ID    0
Customer ID       0
Payment_Mode      0
dtype: int64
```

- As you see, there are 440098 data points.
- There is no NA value.

History of the Dataset – all_data.csv

```
all_df.info()
<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  datetime64[ns]
 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   Customer ID      359392 non-null  int64
 8   Payment_Mode     359392 non-null  object
 9   Gender           359392 non-null  object
 10  Age              359392 non-null  int64
 11  Customer_Income  359392 non-null  int64
 12  Population       359392 non-null  float64
 13  Users            359392 non-null  float64
dtypes: datetime64[ns](1), float64(5), int64(4), object(4)
memory usage: 41.1+ MB
```

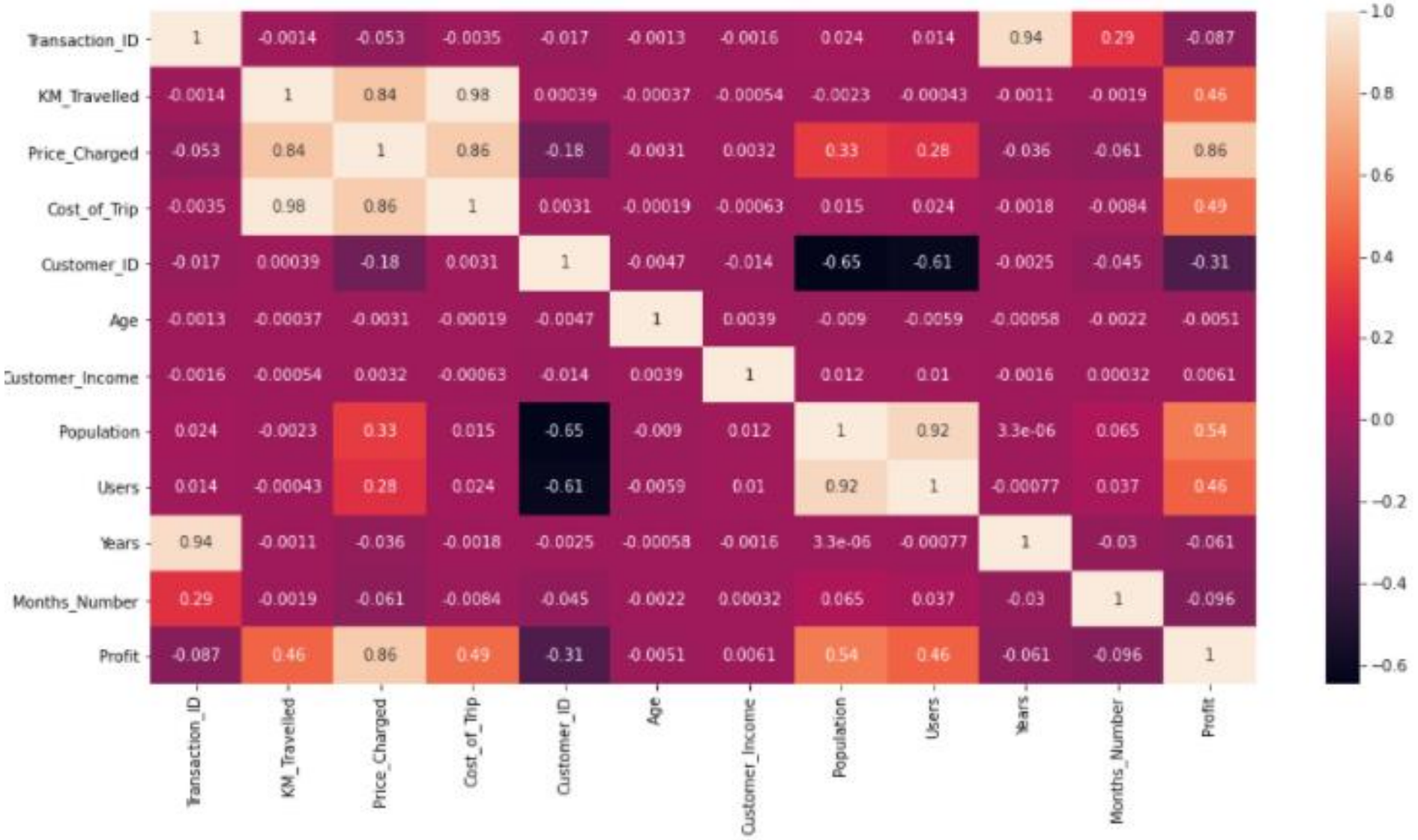
```
all_df.isnull().sum()
Transaction ID     0
Date of Travel     0
Company            0
City               0
KM Travelled       0
Price Charged      0
Cost of Trip       0
Customer ID        0
Payment_Mode       0
Gender             0
Age                0
Customer_Income    0
Population         0
Users              0
dtype: int64
```

```
Missing data
        Company:        0 (0.0%)
        KM_Travelled:   0 (0.0%)
        Price_Charged:  0 (0.0%)
        Users:  0 (0.0%)
        Profit:         0 (0.0%)
        Cities:         0 (0.0%)
        Years:  0 (0.0%)
        Payment_Mode:   0 (0.0%)
        Age:    0 (0.0%)
        Gender:         0 (0.0%)
        Customer_ID:    0 (0.0%)
        Transaction_ID:         0 (0.0%)
```

- As you see, there are 359392 data points.
- There is no NA value.

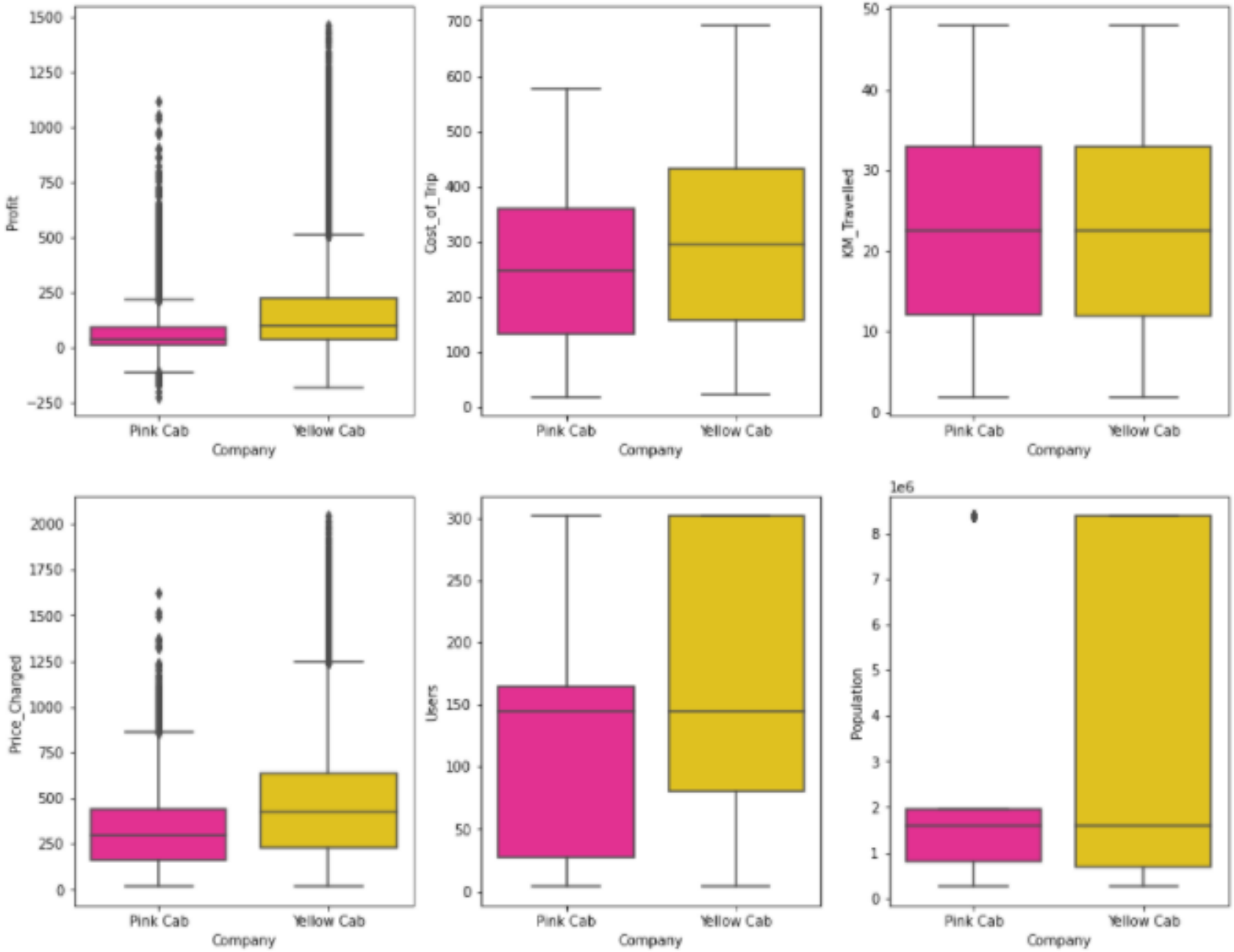
- 4 datasets are combined in this file.

Correlation of numeric variables in all_data

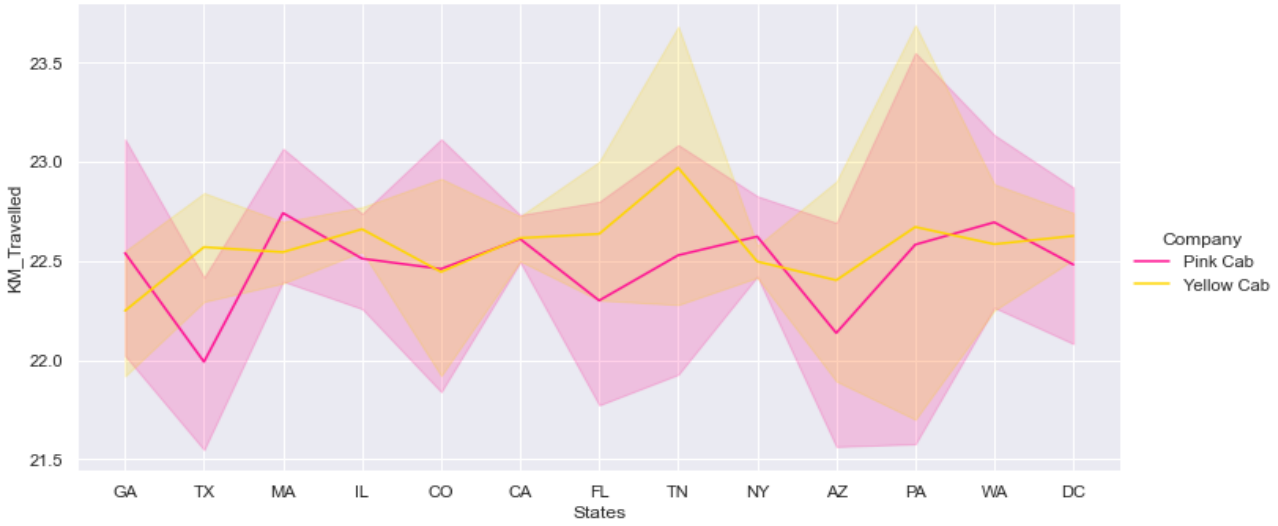


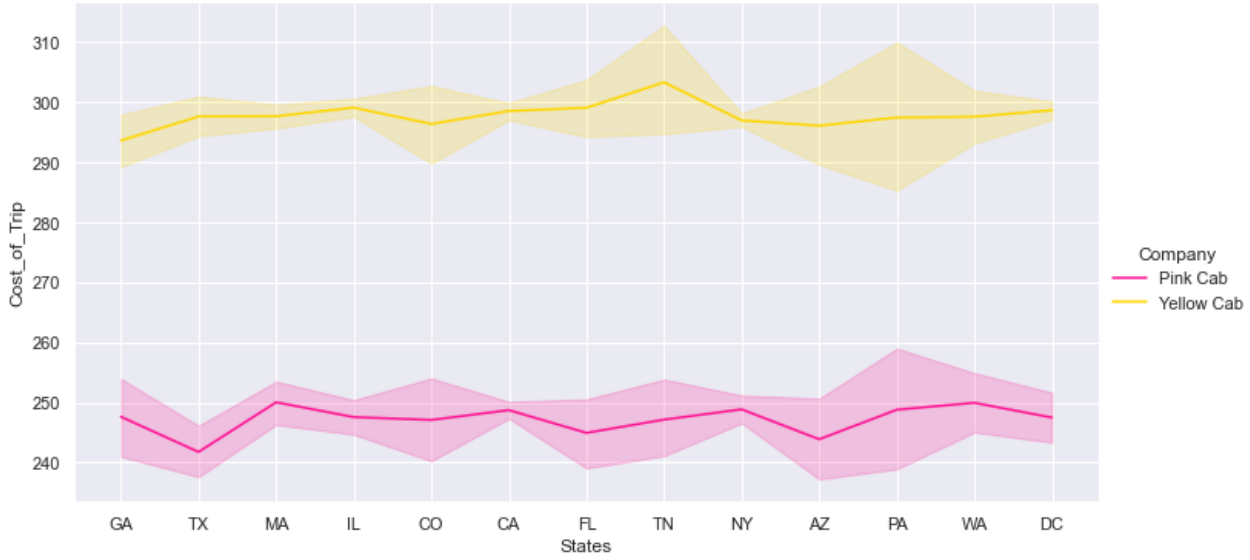
Box Plot Analysis

- The appearance of outliers explains some of the deviations.
- Due to the high number of yellow taxis, its dominance can be seen in all values.
- The negative profit could not be explained by the available data.

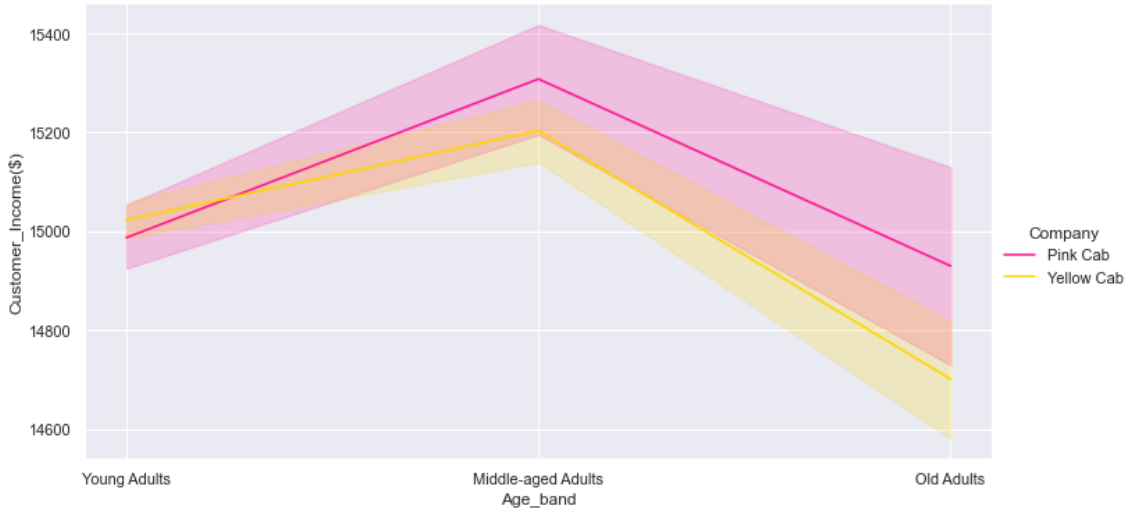


Profit Analysis

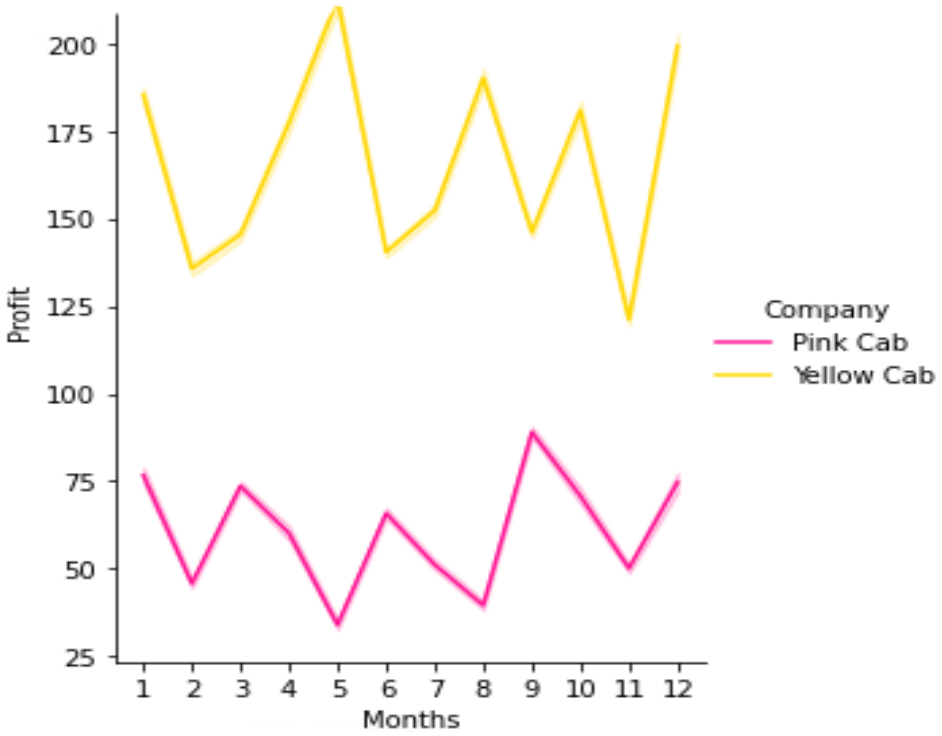




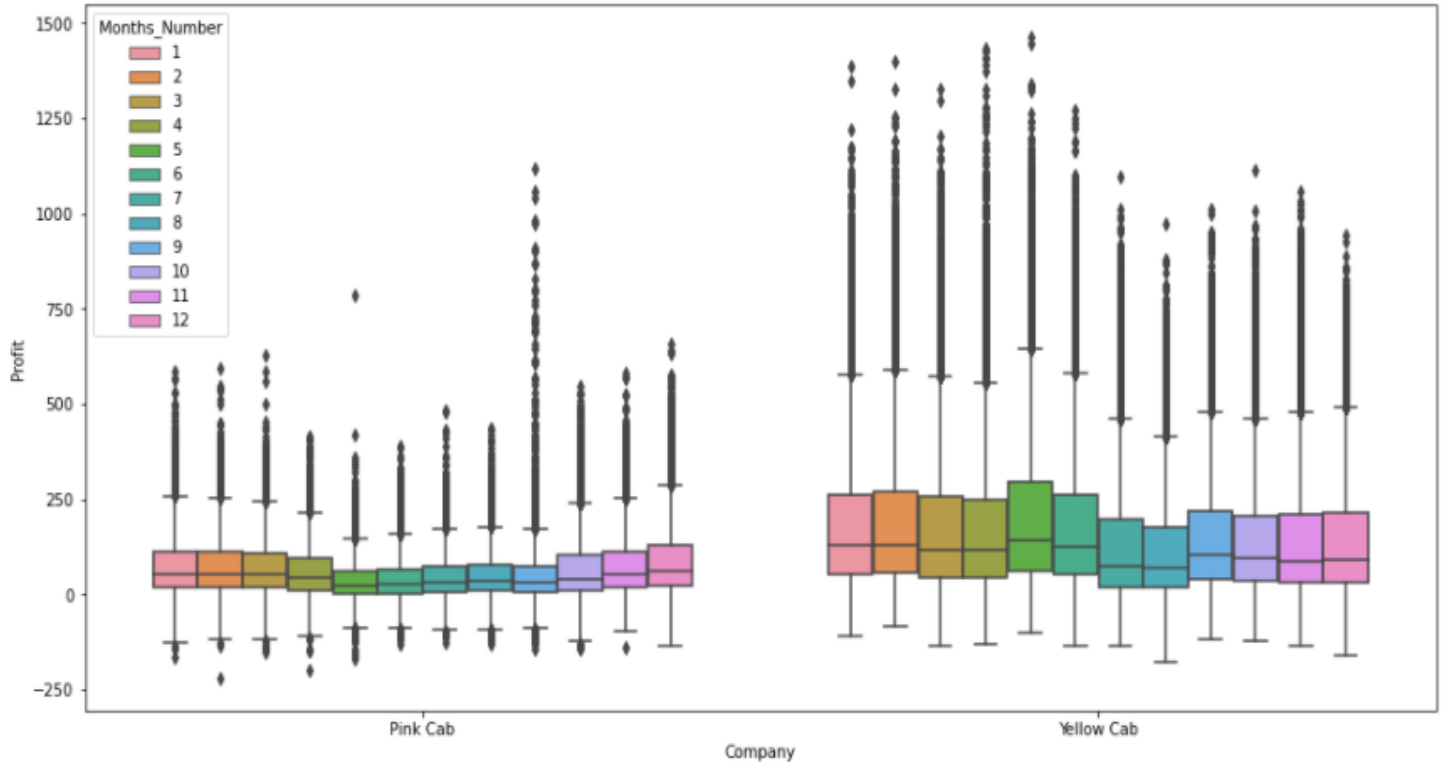
- The Pink Cab has done more KM_ Travelled in some States (GA, MA, NY, WA). As can be seen the Cost_of_Trip is low.
- In addition most people who use the Pink Cab seem to have a very high income.



Profit Analysis

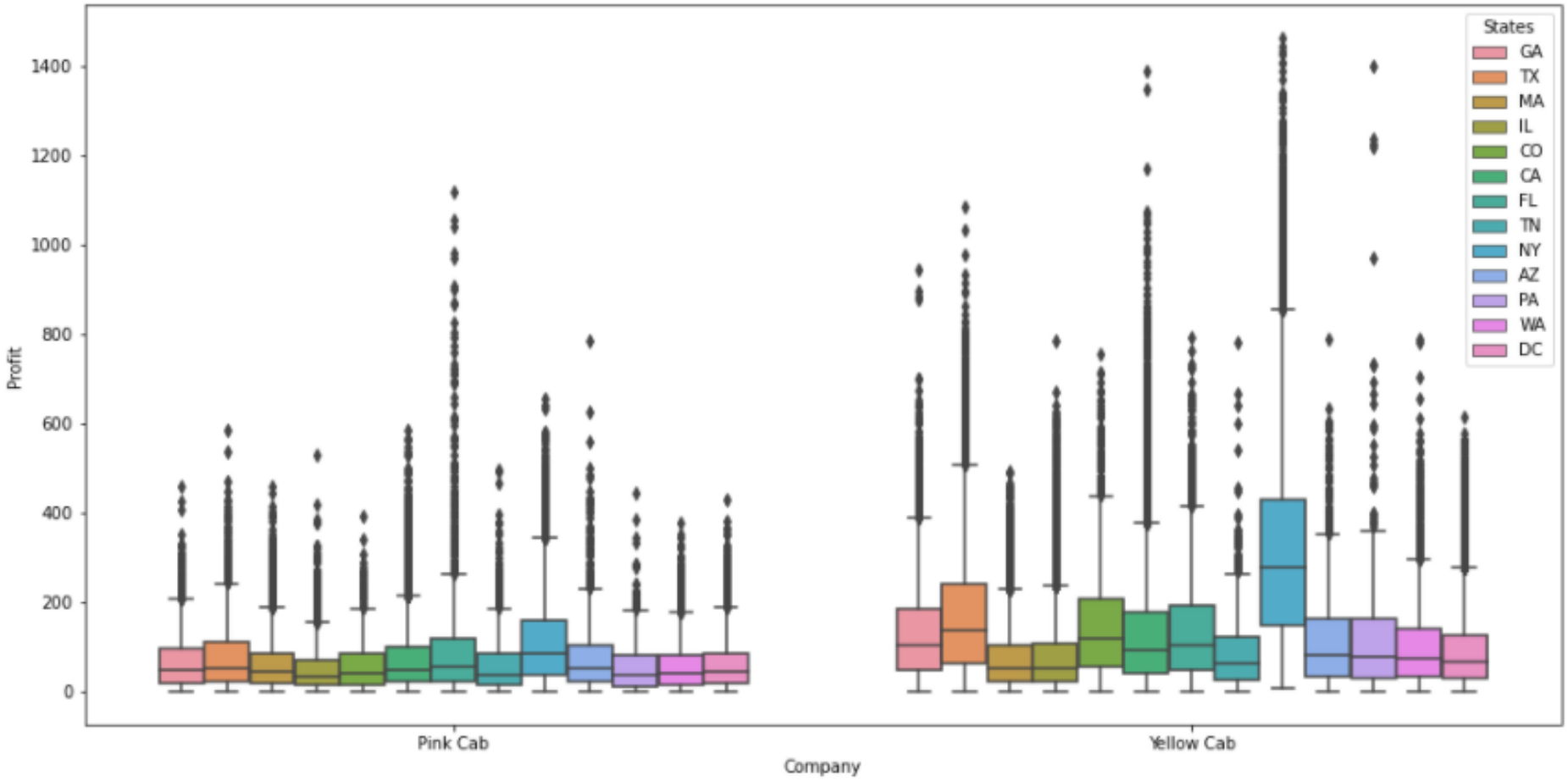


- Both companies made a profit.
- It is seen that the highest profit of the Yellow Cab is at the 5th month and the highest profit of the Pink Cab is at the 9th month.



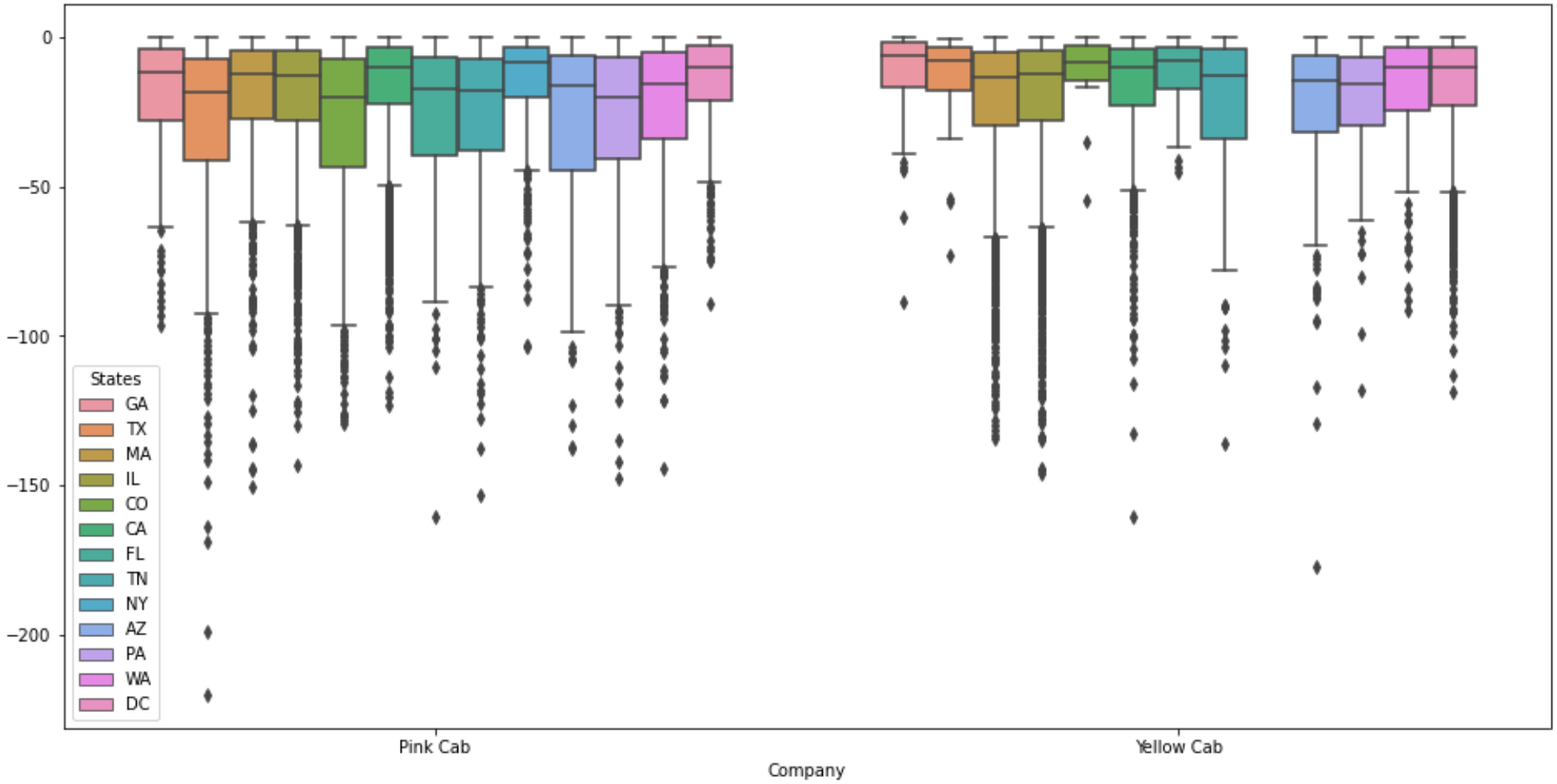
- Although it seems that both companies make a profit according to the months, in fact there are losses.
- The reason for this could not be found with the available data.

Profit Analysis



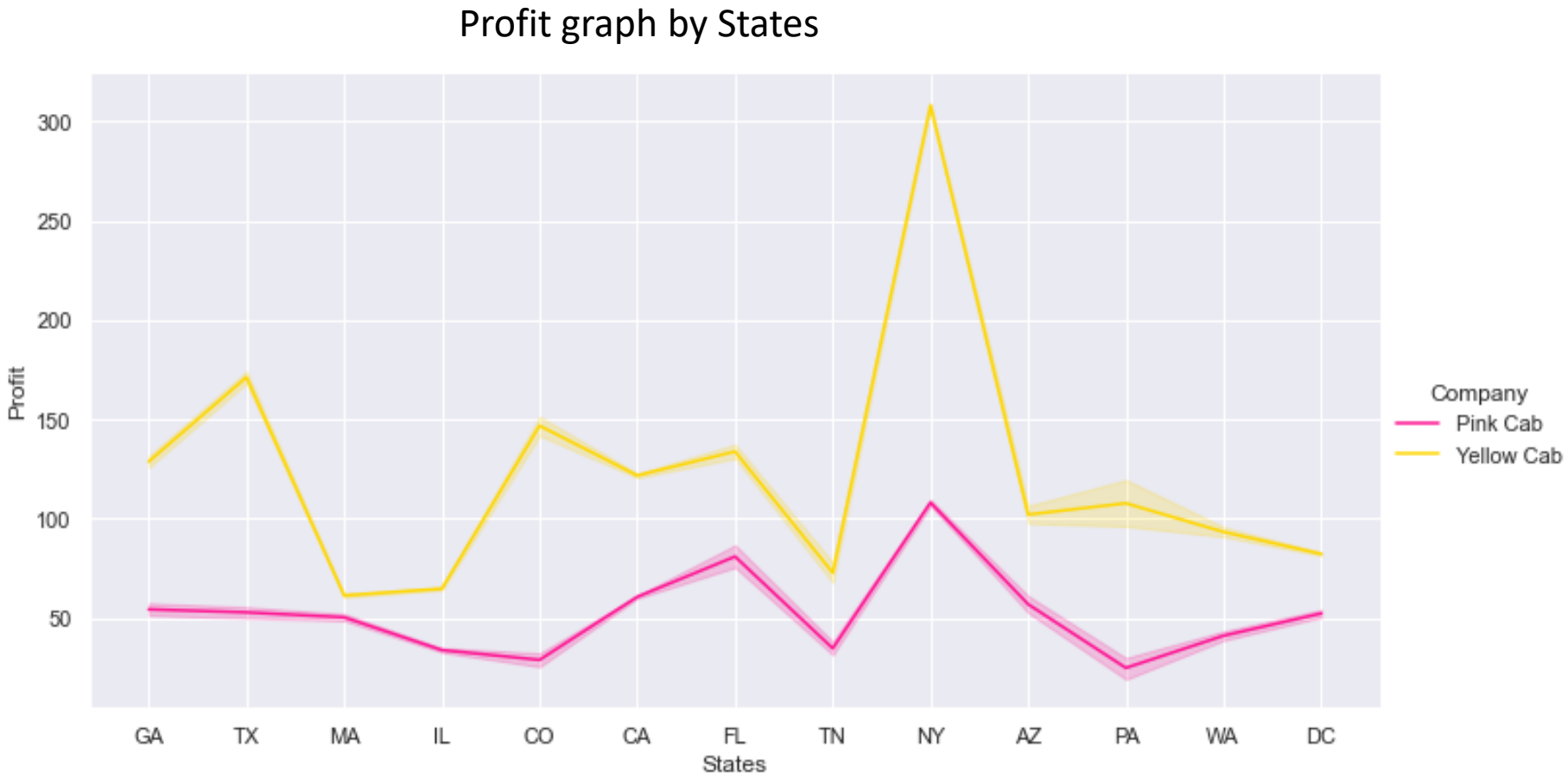
- Yellow Cab earned its highest income from NY.
- On the other hand, Pink Cab earned its highest income from FL.

Profit Analysis



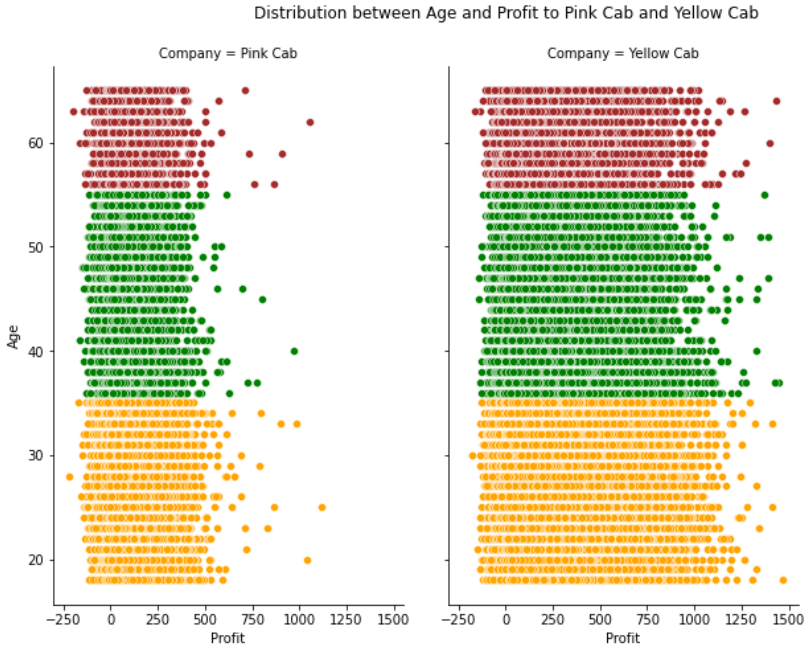
- The graph shows the losses of both companies by States. But it noteworthy that the Yellow Cab did not lose in NewYork (NY).

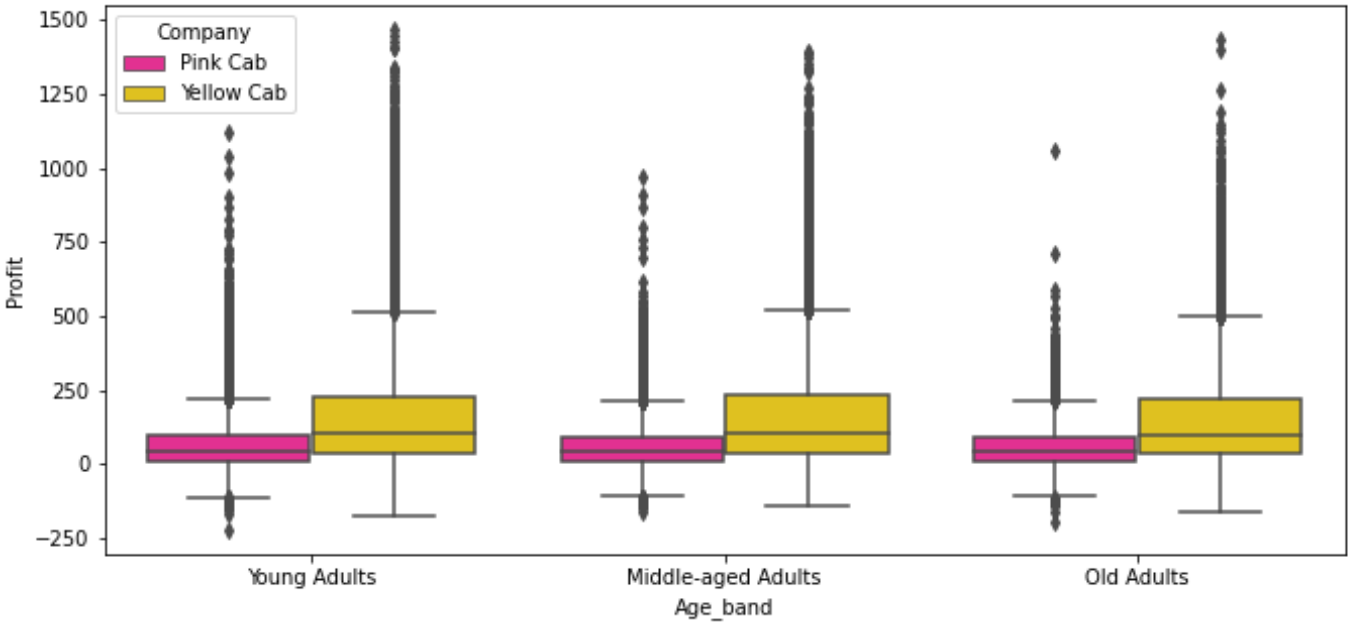
Profit Analysis



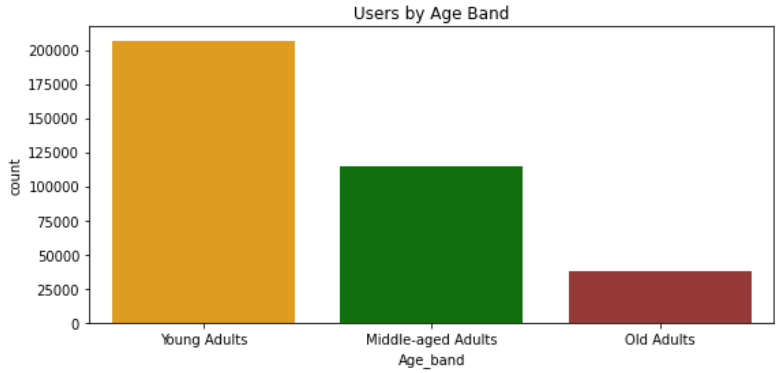
- The 3 most profitable States in the Yellow Cab are NY, TX and CO.
- In a Pink Cab is NY, FL and AZ.

Profit Analysis



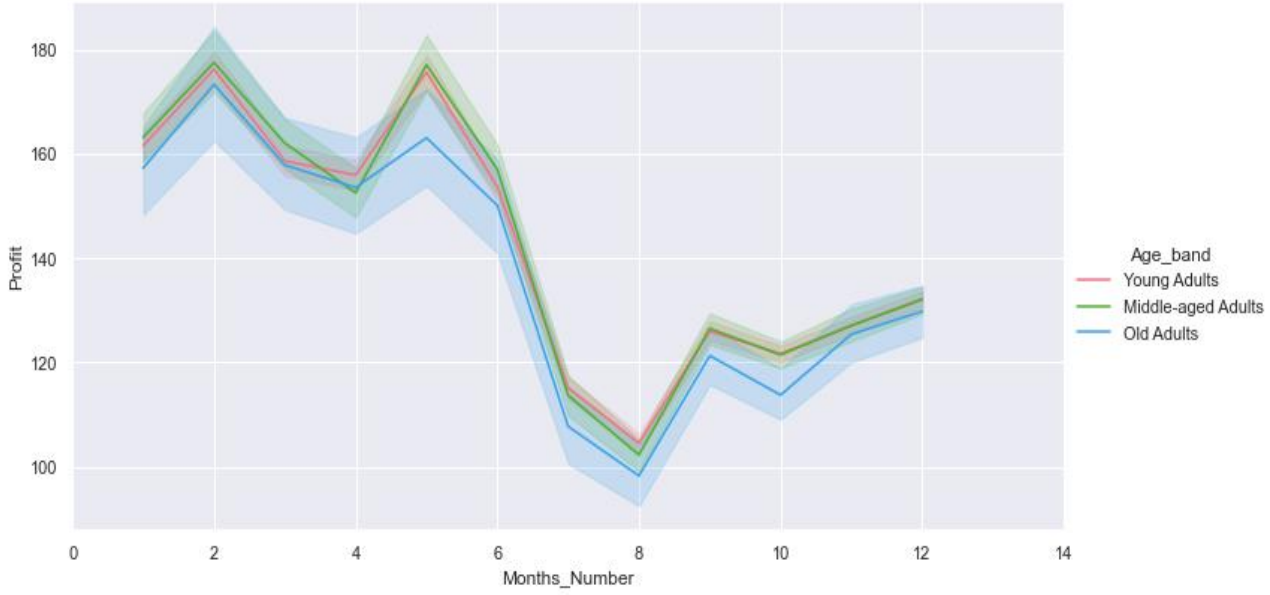


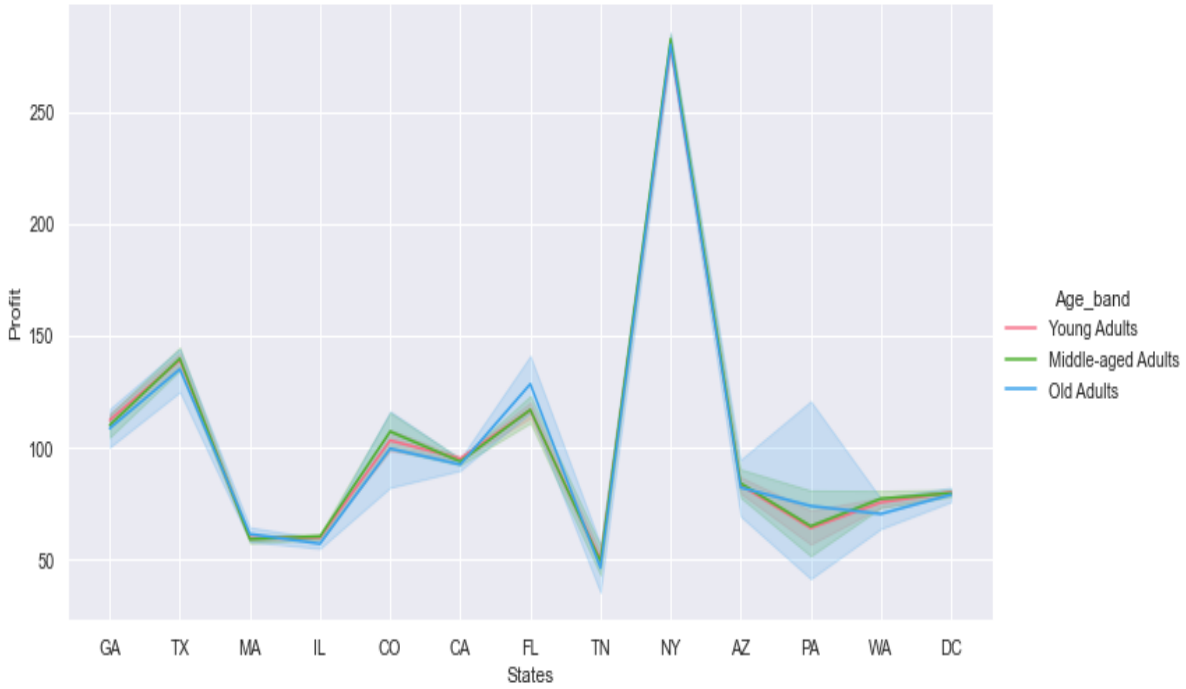

```
Proportion of Total Young Adults: 57.5 %
Proportion of Total Middle-aged Adults: 31.9 %
Proportion of Total Old Adults: 10.6 %
```



- Seventy percent of cabs users are Young_Adults.
- Yellow Cab with the highest profits also by Age_band.

Profit Analysis



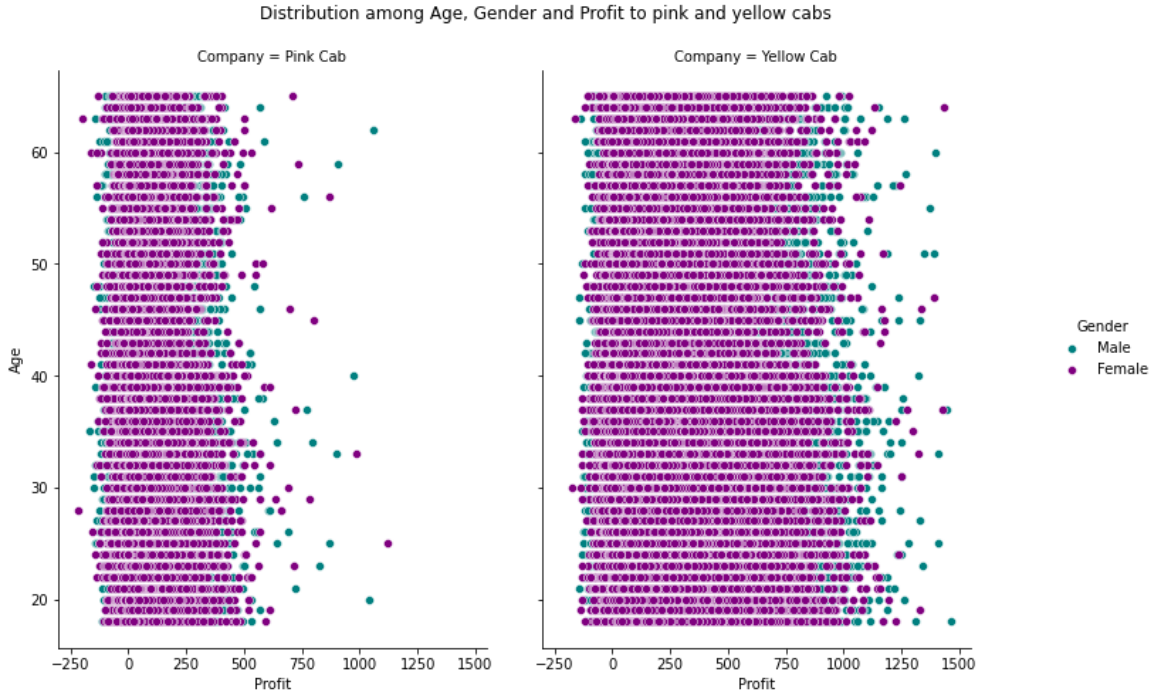


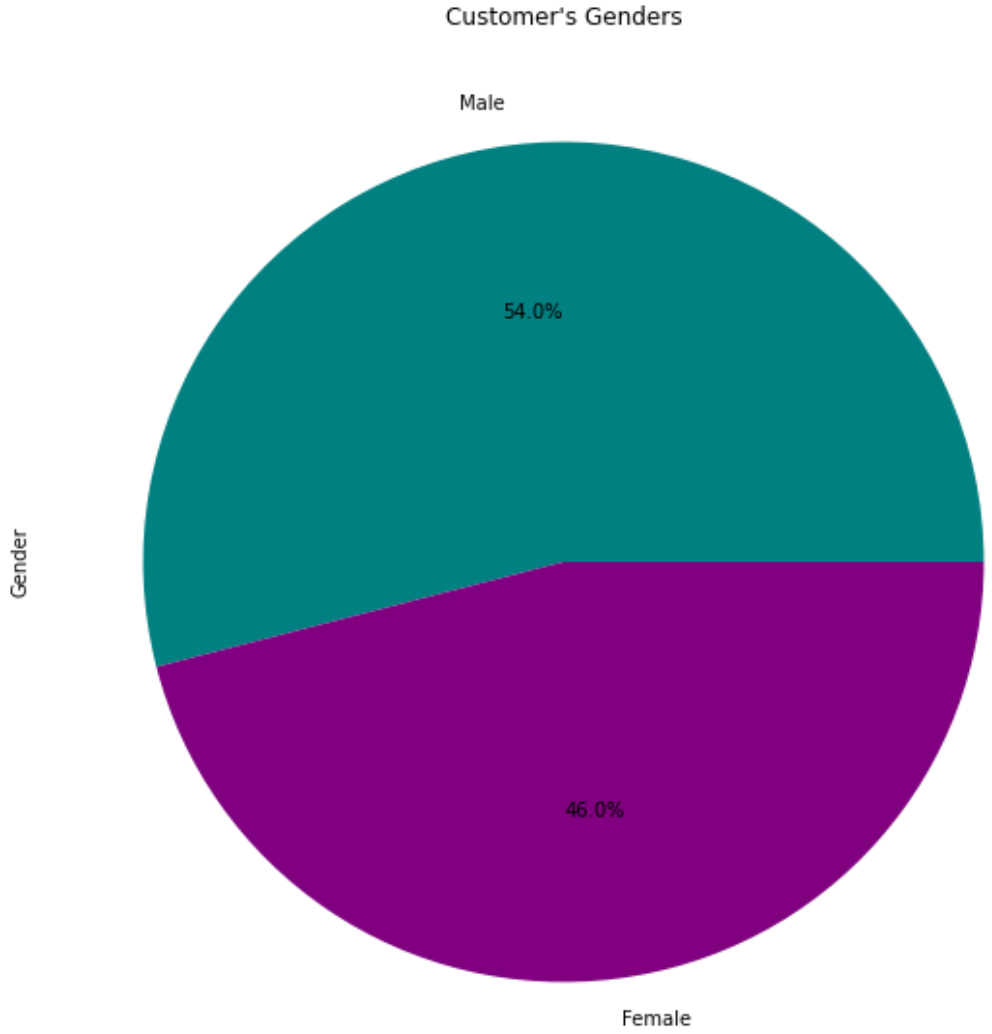
- When the profit is analyzed by months, the highest income Middle_aged_Adults in 2 and 5 months.
- When looked at by state, the highest profit comes from Older_Adults in NY.

Profit Analysis

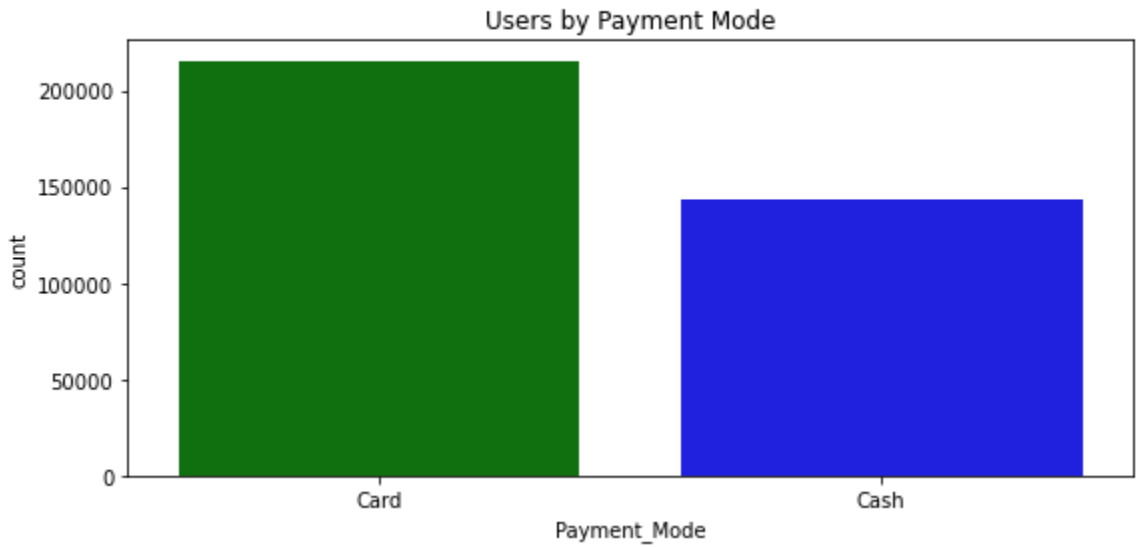
	Company	Gender	Gender_Count
0	Pink Cab	Female	37480
1	Pink Cab	Male	47231
2	Yellow Cab	Female	116000
3	Yellow Cab	Male	158681

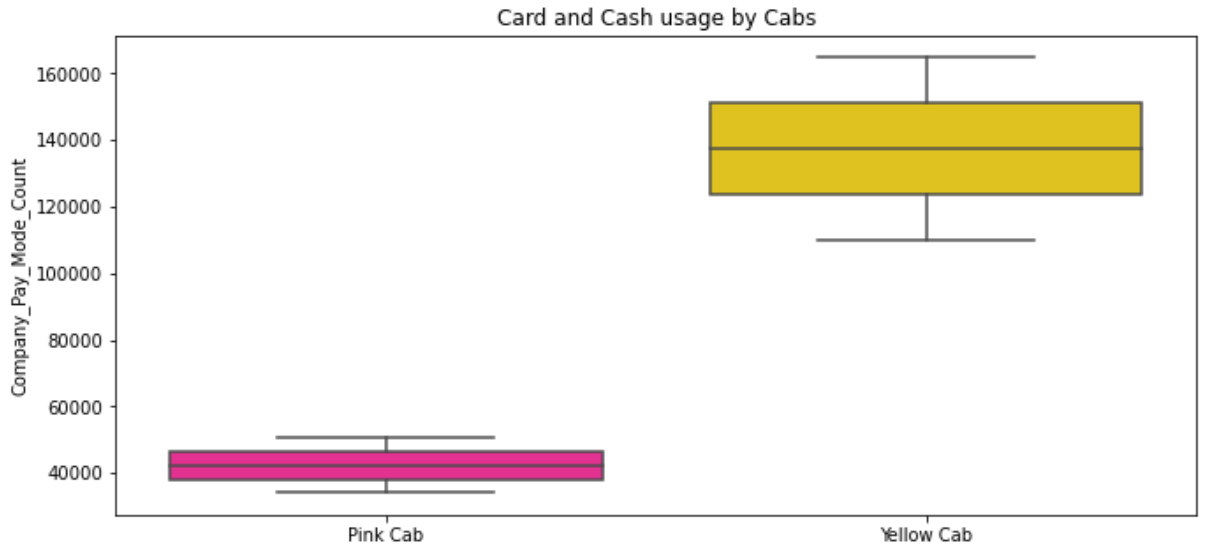
- In both cabs have more male users.



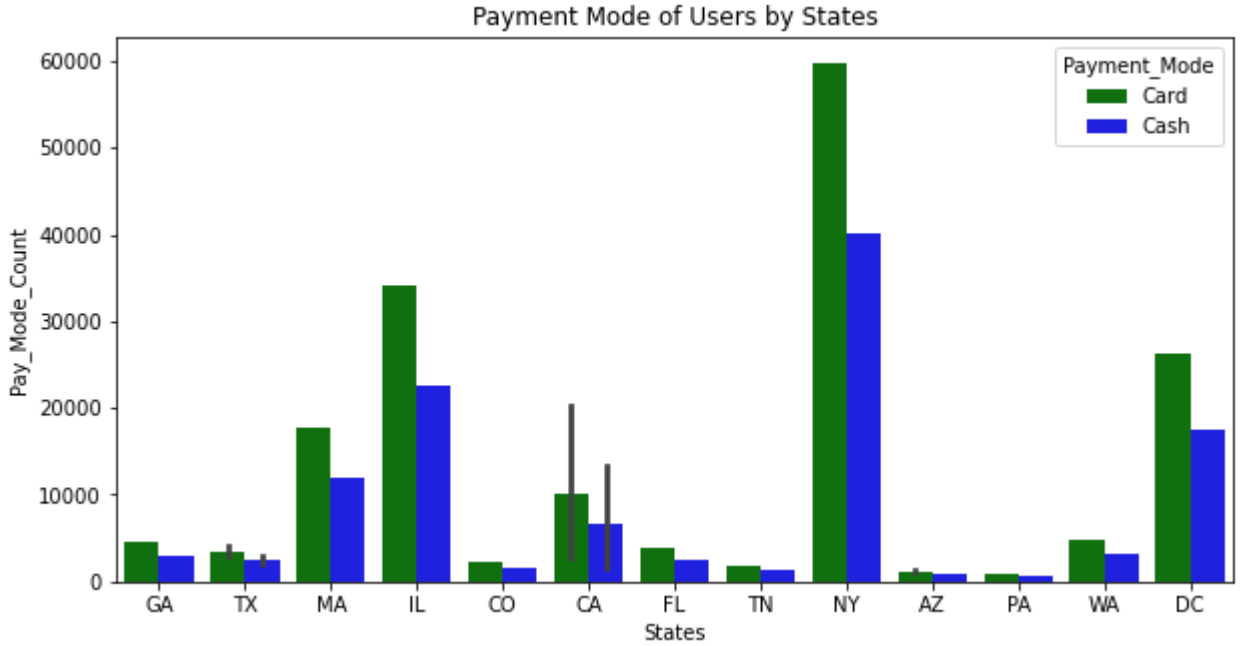


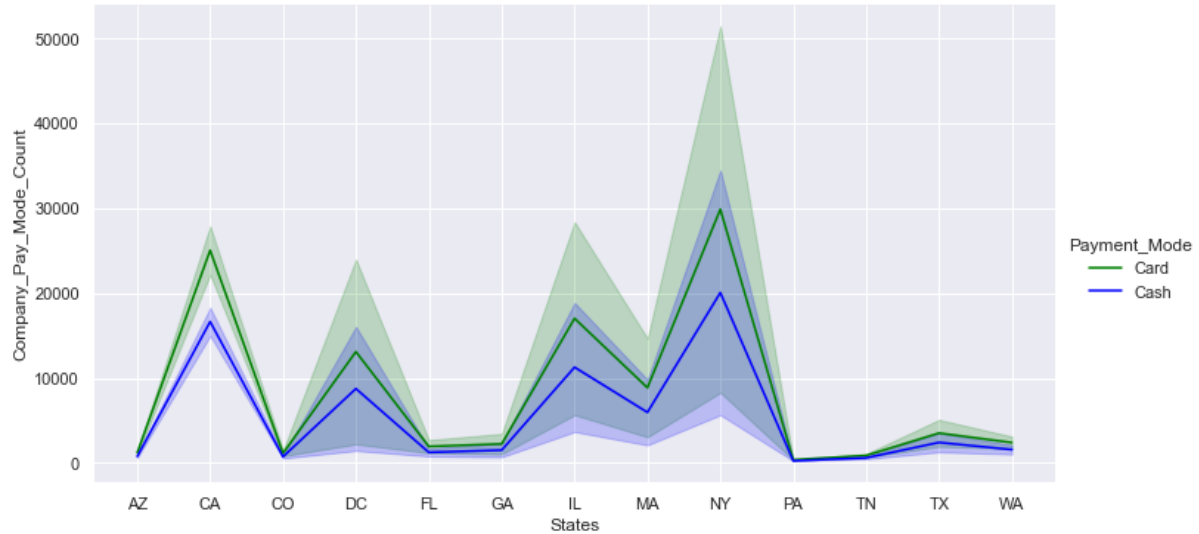
```
Proportion of Total Users Card: 59.963 %
Proportion of Total Users Cash: 40.037 %
```

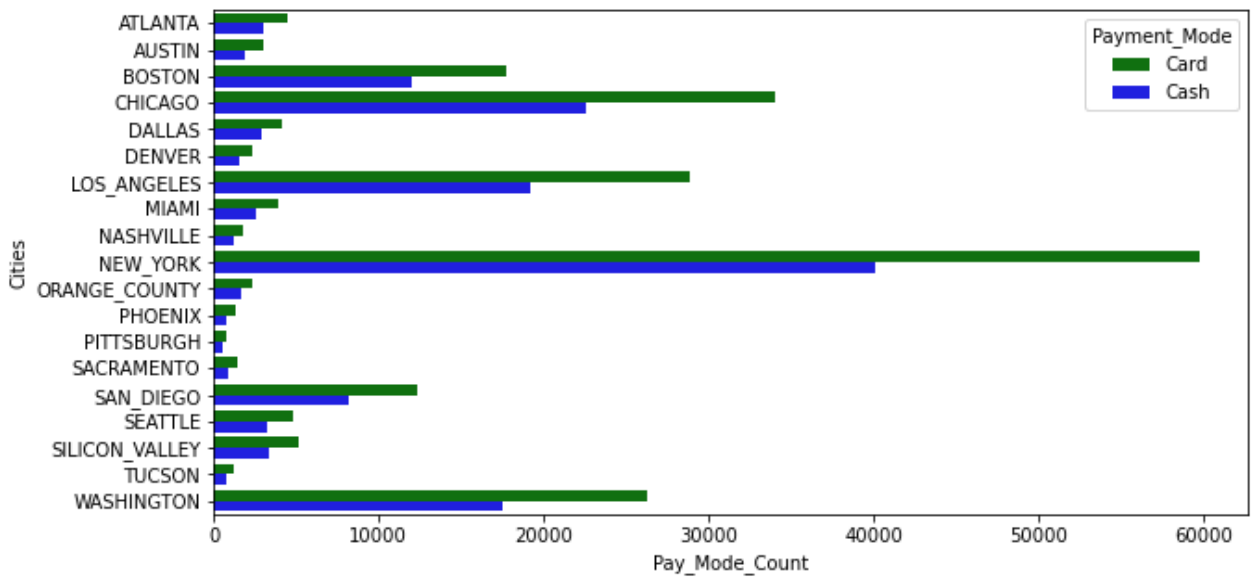


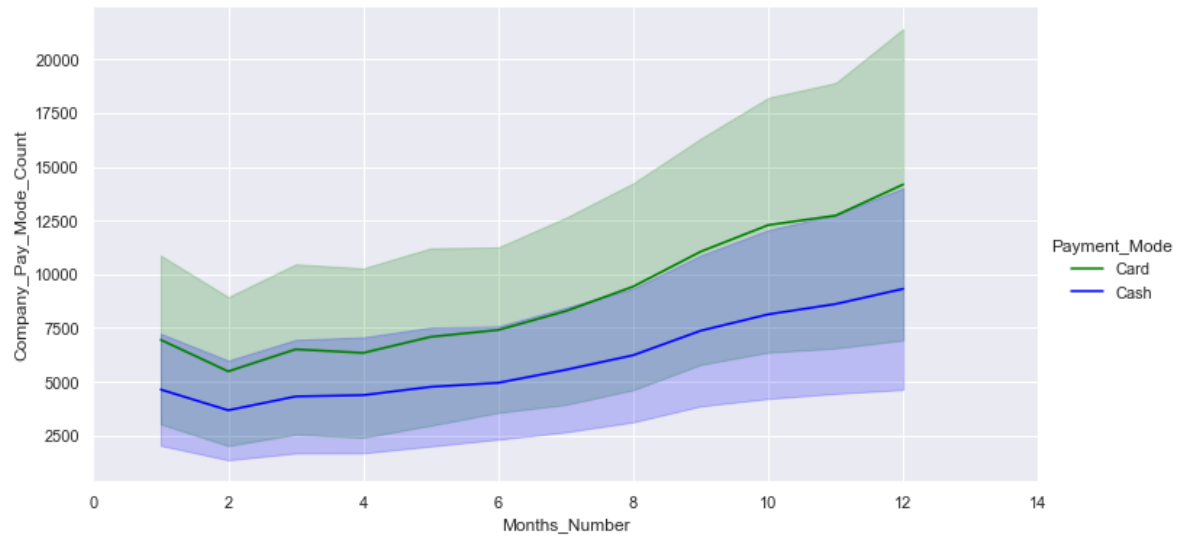


- In both cabs have more Card users.









- Payment is generally made by card.

Conclusion

- **When the datasets of 3 years were examined, it was seen that the most profitable company was the Yellow Cab compared to the Pink Cab.**
- **Yellow cab's average profit per KM is almost three times the average profit per KM of the Pink cab.**
- **In general, the Payment_Mode is card.**
- **According to the detailed analysis, XYZ firm should invest in YELLOW CAB. Given the losses, XYZ firm should invest more to Yellow Cab in the New York.**

THANK YOU !

Data Glacier

Your Deep Learning Partner