Exploratory Data Analysis

G2M insight for Cab Investment firm

17-Sep-2022

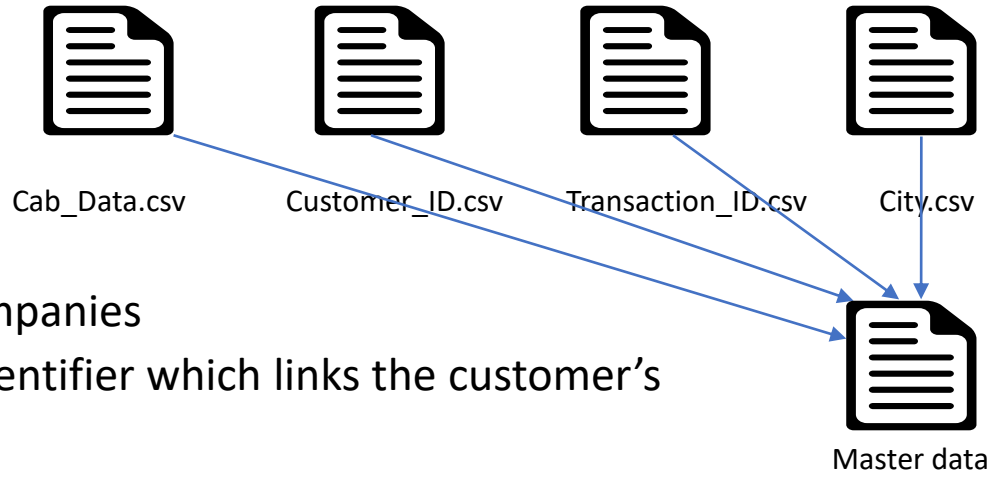
Agenda

- Background
- Data Preprocessing
- EDA
- EDA Summary
- Hypothesis and Test
- Recommendations and Insights

Background – G2M insight for Cab Investment firm

- 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 (Yellow/ Pink Cab) for making investment.



- Datasets:

4 individual data sets.

Time period of data is from 31/01/2016 to 31/12/2018.

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

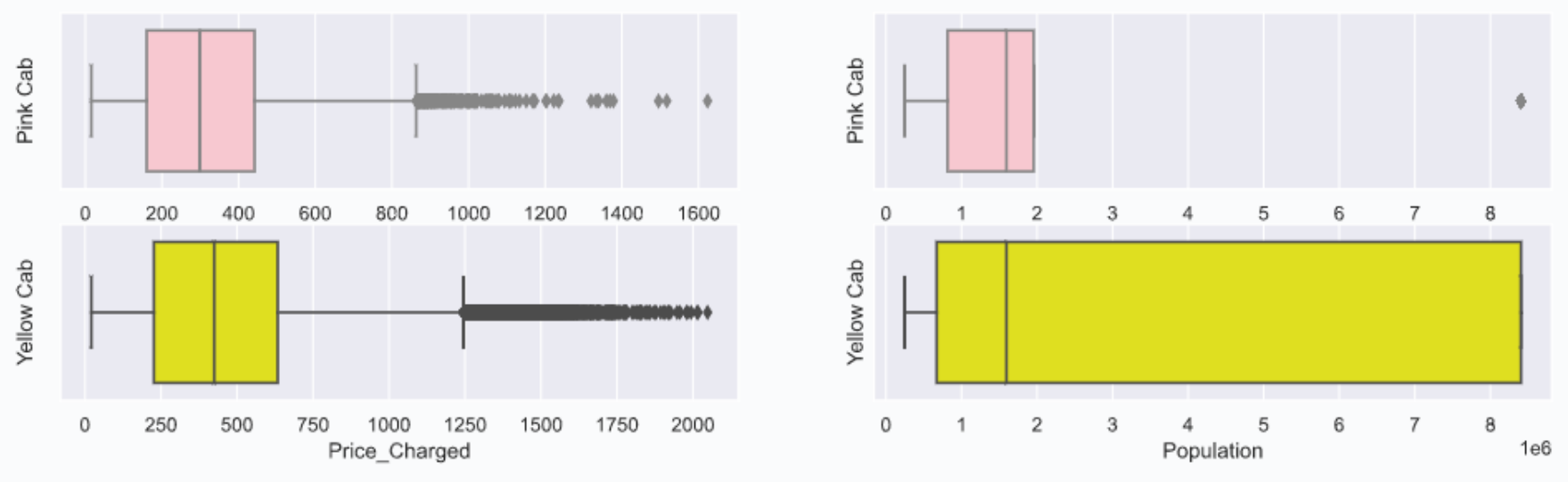
Data Preprocessing

- Executive Summary
  - Create Master Data: Identify relationships across the 4 files
  - Replacing spaces with '_' in column names
  - Convert the 'Date of Travel' column into datetime standard format
  - Add column of 'Profit' = Price_Charged - Cost_of_Trip
  - Removing ',' in population and users column values
  - Convert some columns from object to category

```
Transaction_ID              int64
Date_of_Travel      datetime64[ns]
Company                  category
City                     category
KM_Travelled              float64
Price_Charged             float64
Cost_of_Trip              float64
Customer_ID                int64
Payment_Mode             category
Gender                   category
Age                        int64
Income_(USD/Month)         int64
Population                 int64
Users                      int64
Year_of_Travel             int64
Month_of_Travel            int64
Profit                    float64
ProfitPercentage          float64
dtype: object
```

EDA

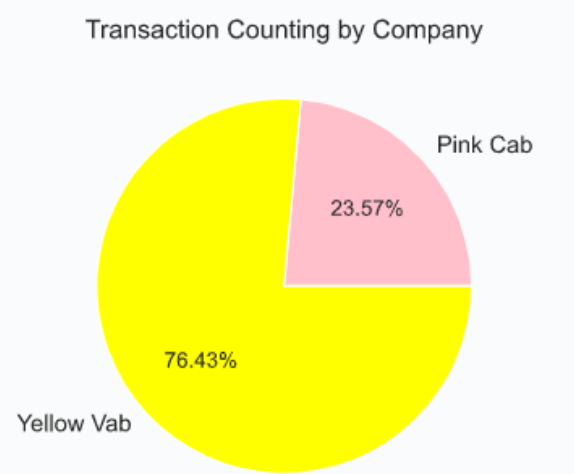
- Data Check
  - There are no missing values.
  - There are no duplicated rows.
  - Price_Charged and Population have outliers (statistically). Considering lack of further background information, so keep the data at this stage.



EDA

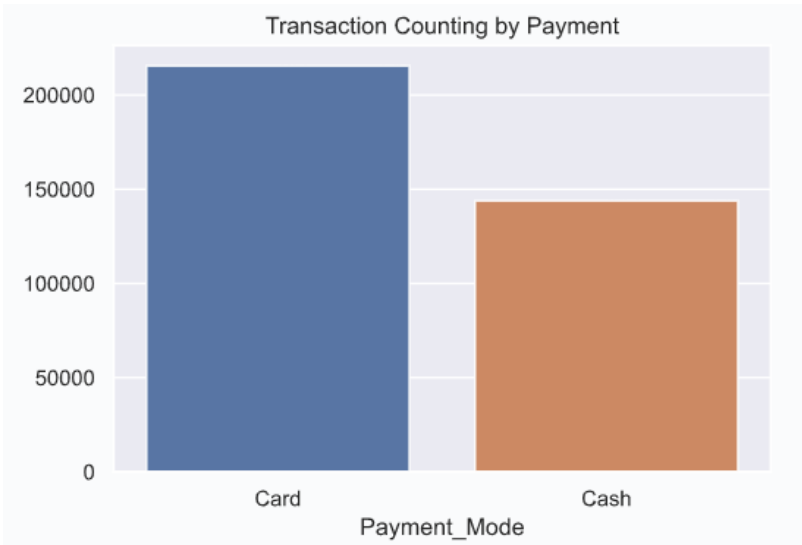
- Categorical variables

Company: Yellow Cab is used more than Pink Cab. 76.4% of the transactions are from Yellow Cab, approximately 3 times that of Pink Cab.



Company	
Pink Cab	84711
Yellow Cab	274681

Payment: Num of transactions paid by Card is 1.5 times num of those paid by cash.



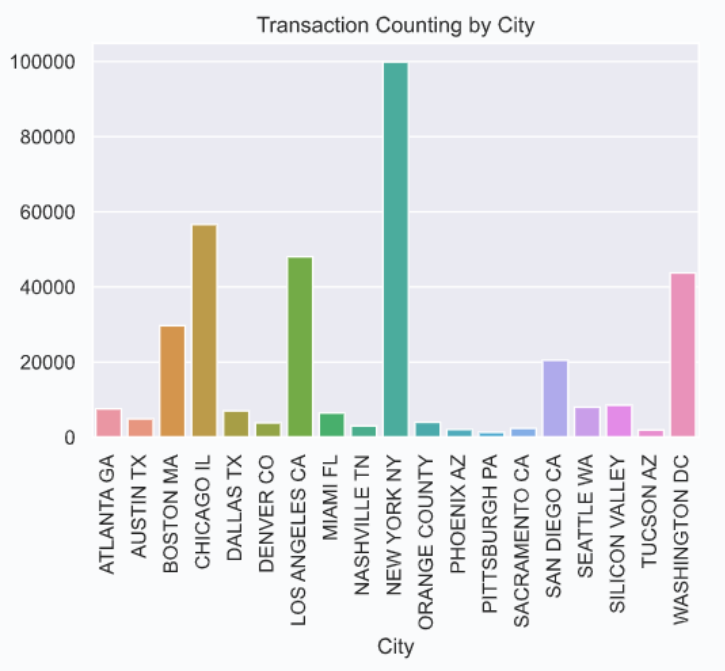
Payment_Mode	
Card	215504
Cash	143888

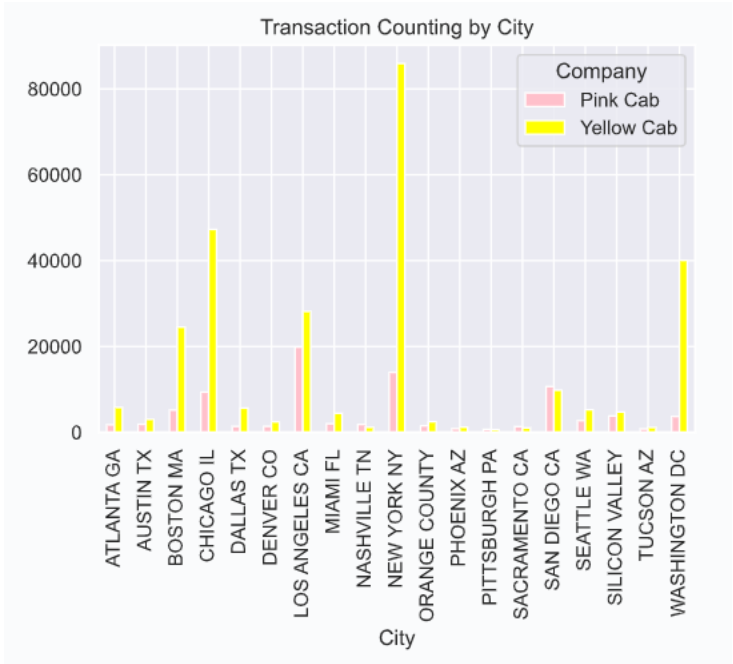
EDA

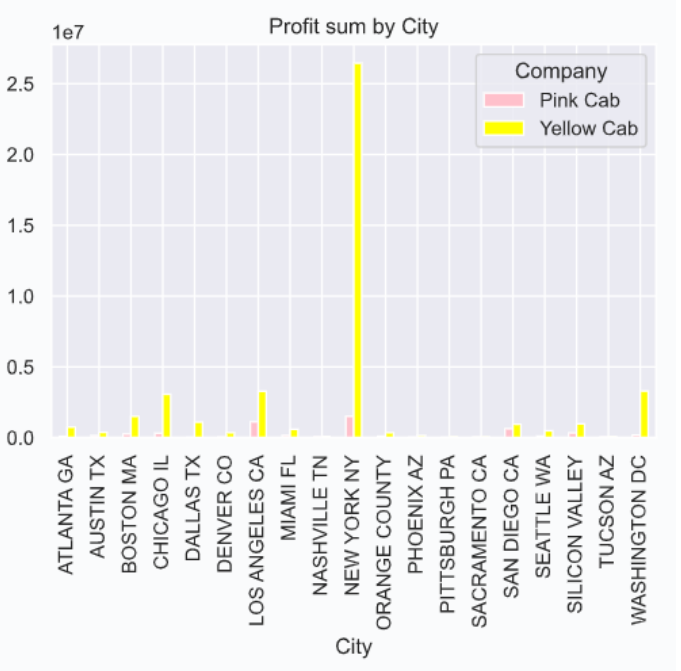
- Categorical variables

  City: NEW YORK NY has the most transactions while PITTSBURGH PA with least.
  - For Pink Cab: LOS ANGELES CA has the most transactions, followed by NEW YORK NY, SAN DIEGO CA. For
  - Yellow Cab: NEW YORK NY has the most transactions, followed by CHICAGO IL, WASHINGTON DC.

  Especially in New York, Yellow Cab's profit is extremely higher than Pink Cab.



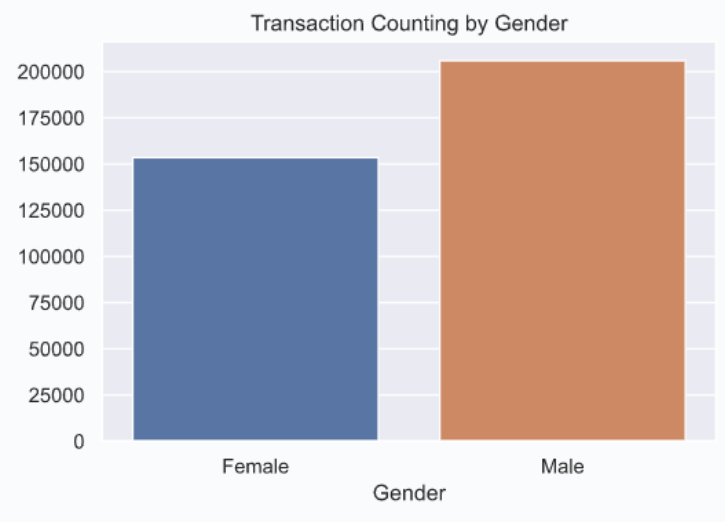




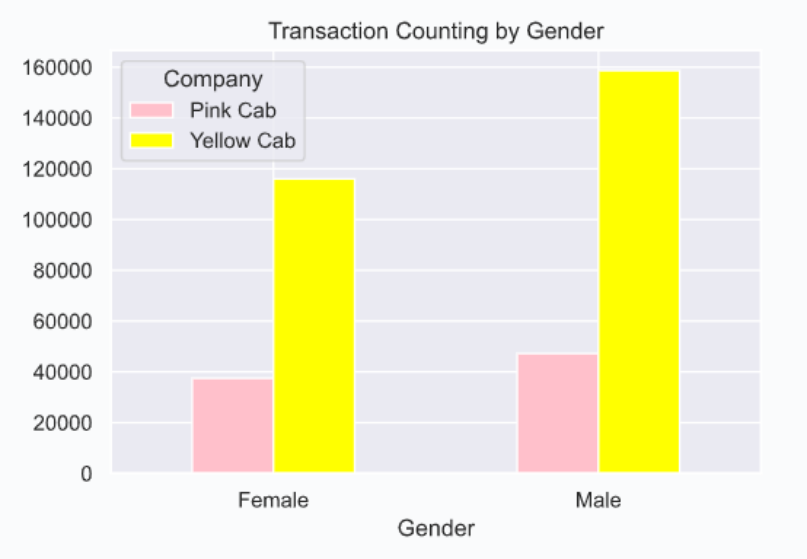
EDA

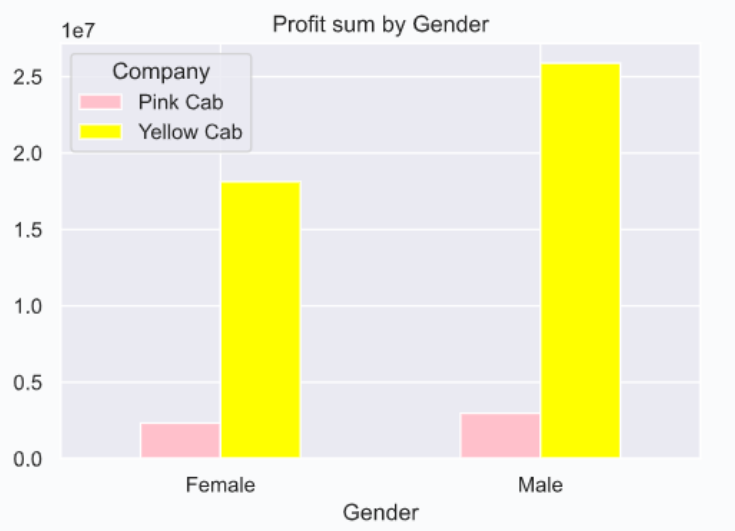
- Categorical variables

  Gender: Male use cab more frequently than female.
  - Yellow Cab has bigger gap in gender difference than Pink Cab, but they both have more male users.
  - Looks there is more obvious gender difference in profit of Yellow Cab. Later wee can further raise hypothesis and test.




```
Gender
Female    153480
Male      205912
```

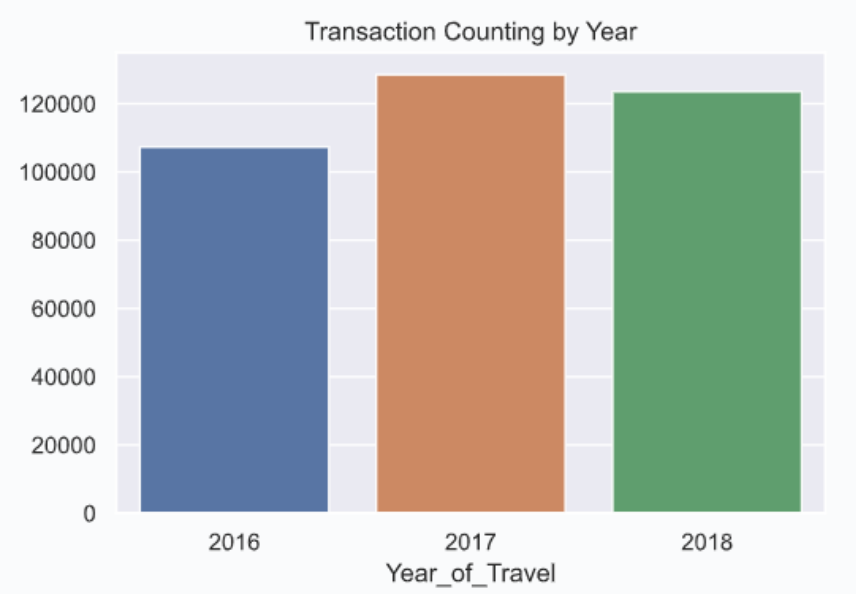




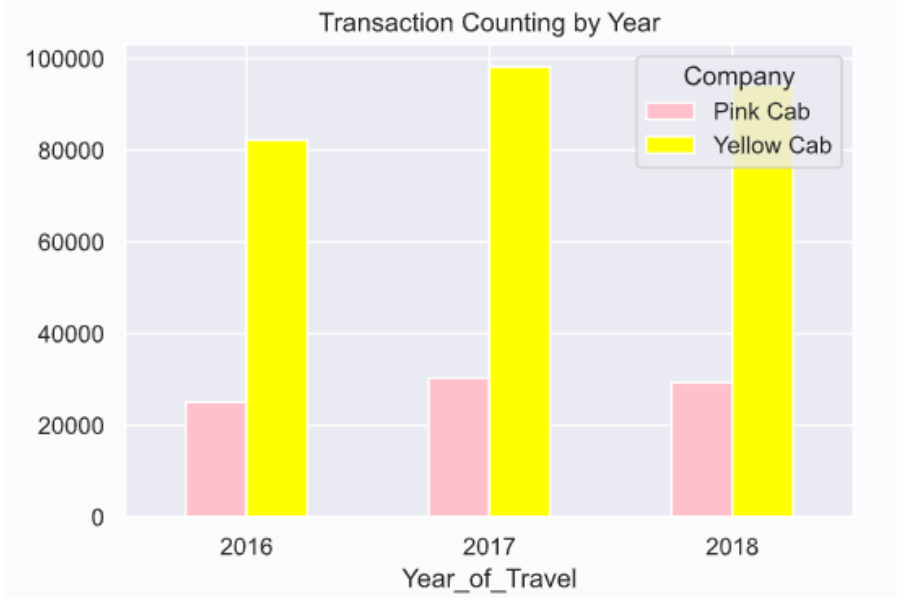
EDA

- Numerical variables

  Year: 2017 has the most transactions but the ‘Year’ column approximately has fairly distributed data.



Year_of_Travel	
2016	107319
2017	128510
2018	123563

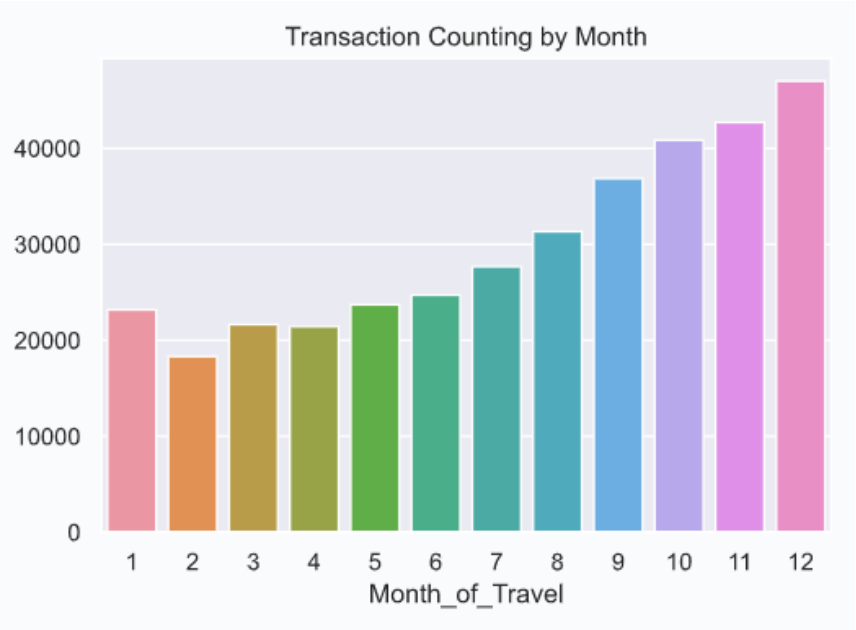


Company	Pink Cab	Yellow Cab
Year_of_Travel		
2016	25080	82239
2017	30321	98189
2018	29310	94253

EDA

- Numerical variables

  Month: December has the most transactions while February with the least.
  - It can be said that transactions increase by month. No seasonality can be observed.
  - Same trends in both Pink Cab & Yellow Cab.

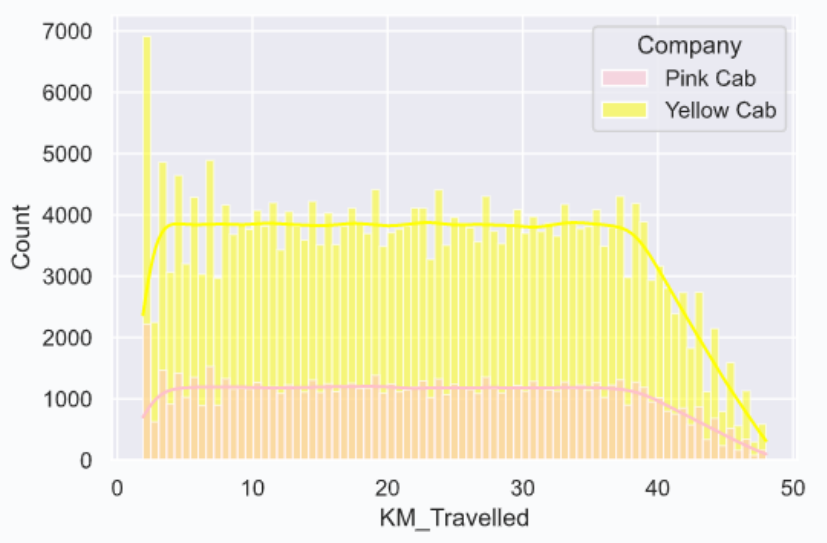


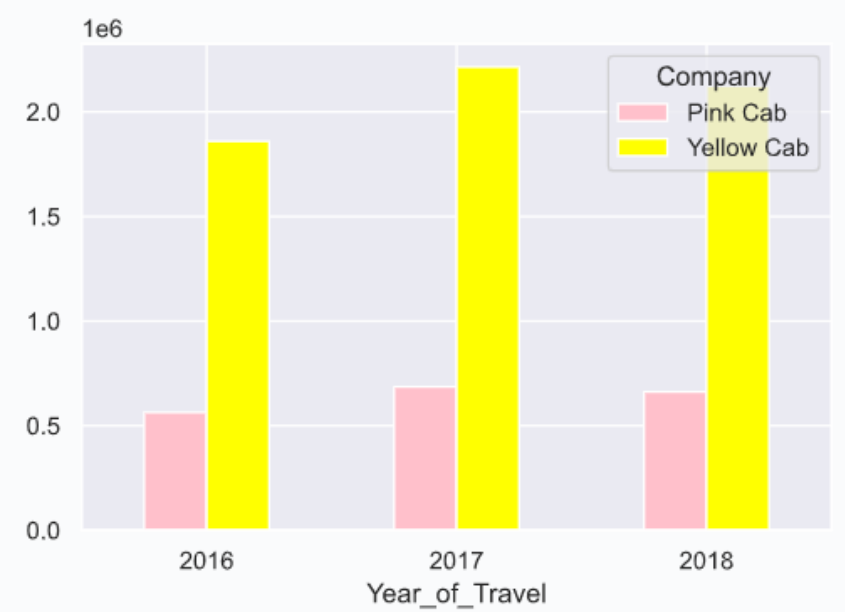
EDA

- Numerical variables

KM_Travelled:
For both Pink Cab & Yellow Cab:

- Most transactions travel 2~40 KM. The most transactions are short-distance travel.
- 2017 has the most KM_Travelled sum.



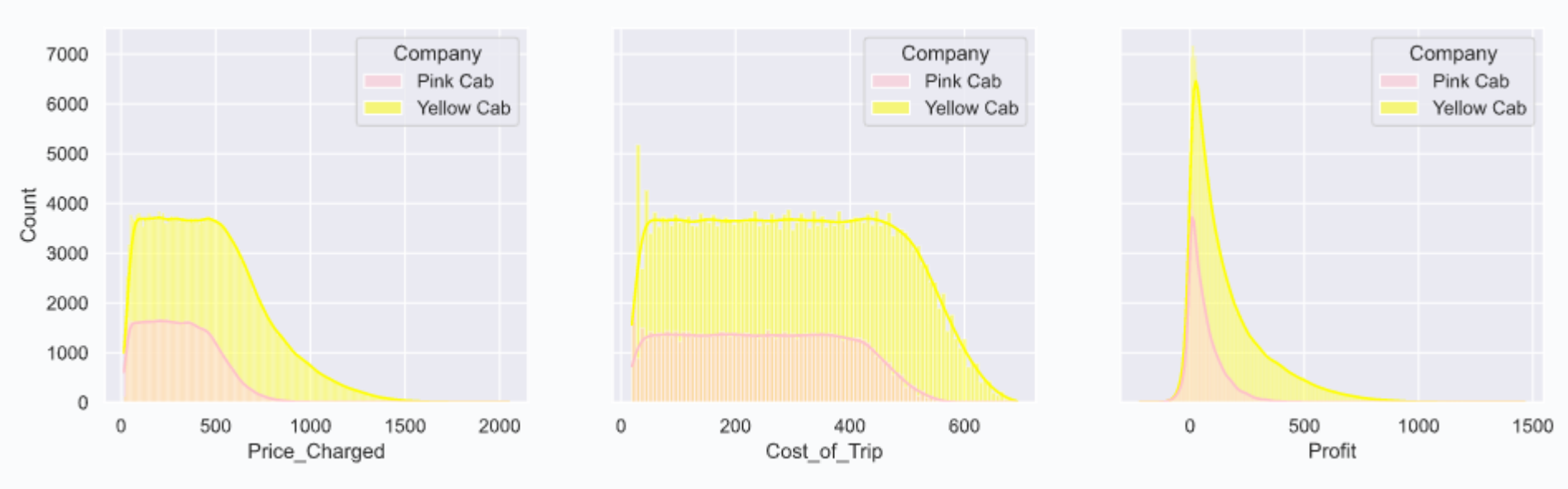


Company	Pink Cab	Yellow Cab
Year_of_Travel		
2016	563509.67	1859978.21
2017	685823.52	2214879.02
2018	661739.92	2124560.24

EDA

- Numerical variables

  Price_Charged, Cost_of_Trip, Profit: For both Pink Cab & Yellow Cab, Profit are not too high. Maybe since the most transactions are short-distance travel.

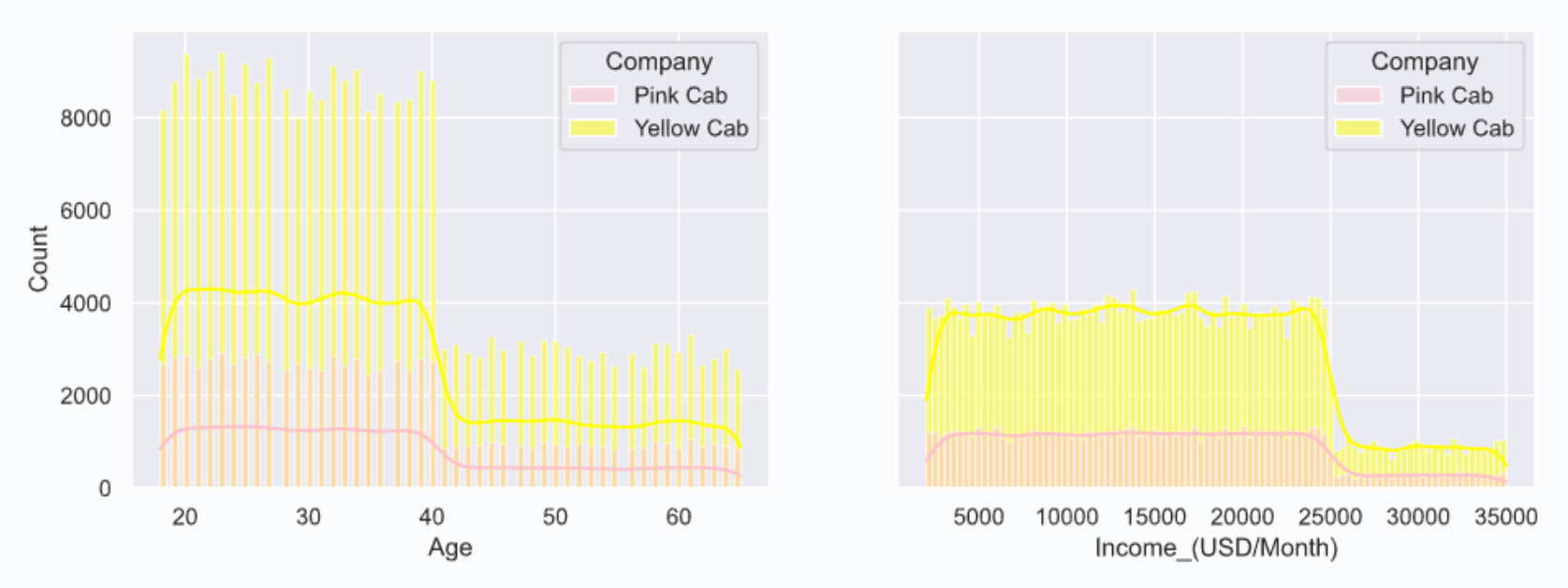


EDA

- Numerical variables

Age, Income_(USD/Month):

- People ages in the range of 18~40 use cab more frequently than those whose age is over 40.
- People's income in the range of 5000~25000 USD/Month use cab more frequently than those whose income is over 25000.



EDA

- Correlation

There is strongly positive correlation between KM_Travelled & Price_Charged , KM_Travelled & Cost_of_Trip, Price_Charged & Cost_of_Trip, Users & Population.

```
Pearson correlation coefficient:
KM_Travelled & Price_Charged : 0.8357531580209331
KM_Travelled & Cost_of_Trip : 0.9818483823189854
Price_Charged & Cost_of_Trip : 0.859811726291571
Users & Population : 0.915490344475287
```

EDA Summary

Market

1. How is the market sharing of two companies?
Yellow Cab is used more than Pink Cab. 76.4% of the transactions are from Yellow Cab, approximately ***3 times*** that of Pink Cab.

2. Do the market situation vary in cities?
NEW YORK NY has the most transactions while ***PITTSBURGH PA*** with least.
For ***Pink Cab***: ***LOS ANGELES CA*** has the most transactions, followed by ***NEW YORK NY, SAN DIEGO CA.***
For ***Yellow Cab***: ***NEW YORK NY*** has the most transactions, followed by ***CHICAGO IL, WASHINGTON DC.***

Customers

1. Is there gender differences of customers?
Male use cab more frequently than female.
Yellow Cab has bigger gap in gender difference than Pink Cab, but they both have more male users.
Looks there is more obvious ***gender difference*** in ***profit*** of ***Yellow Cab.***

2. Is there any obvious features of customers' age & income?
People ages in the range of ***18 ~ 40*** use cab more frequently than those whose age is over 40.
People's income in the range of ***5000 ~ 25000*** USD/Month use cab more frequently than those whose income is over 25000.

EDA Summary

Transactions

1. Is there any seasonality in transactions of cab service?

2017 has the most transactions but the 'Year' column approximately has fairly distributed data.
December has the most transactions while ***February*** with the least.
It can be said that transactions increase by month. ***No*** seasonality can be observed.

2. How is the travel distance (KM) of transactions?

For both Pink Cab & Yellow Cab: ***2017*** has the most KM_Travelled sum.
For both Pink Cab & Yellow Cab: Most transactions travel ***2 ~ 40*** KM. The most transactions are ***short-distance*** travel.
For both Pink Cab & Yellow Cab: Profit are ***not*** too high. Maybe since the most transactions are ***short-distance*** travel.

3. Which payment method is prefered?

Num of transactions paid by ***Card*** is 1.5 times num of those paid by cash.
Yellow Cab has bigger gap in payment method than Pink Cab, but they both have more transactions paid by card.

Others

Is there any correlation amongst attributions?

There is strongly positive correlation between ***KM_Travelled & Price_Charged*** , ***KM_Travelled & Cost_of_Trip***, ***Price_Charged & Cost_of_Trip***, ***Users & Population***.

Hypothesis and Test

1. Null Hypothesis: There is no gender difference in profit.

```
Test for Pink Cab:
P value is  0.11515305900425186
We accept null hypothesis that there is no statistical gender difference in profit.
```

```
Test for Yellow Cab:
P value is  6.060473042494144e-25
We accept alternate hypothesis that there is statistical gender difference in profit.
```

Result: For ***Yellow Cab*** - There is statistical gender difference in profit.

2. Null Hypothesis: There is no age difference in profit.

```
Test for Pink Cab:
P value is  0.15128344738584695
We accept null hypothesis that there is no statistical age difference in profit.
```

```
Test for Yellow Cab:
P value is  0.9031038421935373
We accept null hypothesis that there is no statistical age difference in profit.
```

Result: For both cabs - There is ***no*** statistical age difference in profit.

3. Null Hypothesis: There is no distance difference in profit.

```
Test for Pink Cab:
P value is  0.0
We accept alternate hypothesis that there is statistical distance difference in profit.
```

```
Test for Yellow Cab:
P value is  0.0
We accept alternate hypothesis that there is statistical distance difference in profit.
```

Result: For both cabs - There ***is*** statistical age difference in profit.

Recommendations and Insights

Briefly, based on the analysis, ***Yellow Cab*** is more worthy to make investment.

- Yellow Cab is more popular and has more market sharing.
- In many cities (especially New York, Washington, Chicago, Boston, Los Angeles, etc), Yellow Cab has overwhelming expansion advantage.
- What's more, in New York, Yellow Cab's profit is extremely higher than Pink Cab.
- Though target customers have similar features (age range: 18 to 40, income range: 5000 to 25000), Yellow Cab is more appeal to males and has higher profit gained from this group.
- Though short-distance travel are the main market, Yellow Cab has the ability to deal with long-distance travel, which can directly lead to higher profit.

Thank You