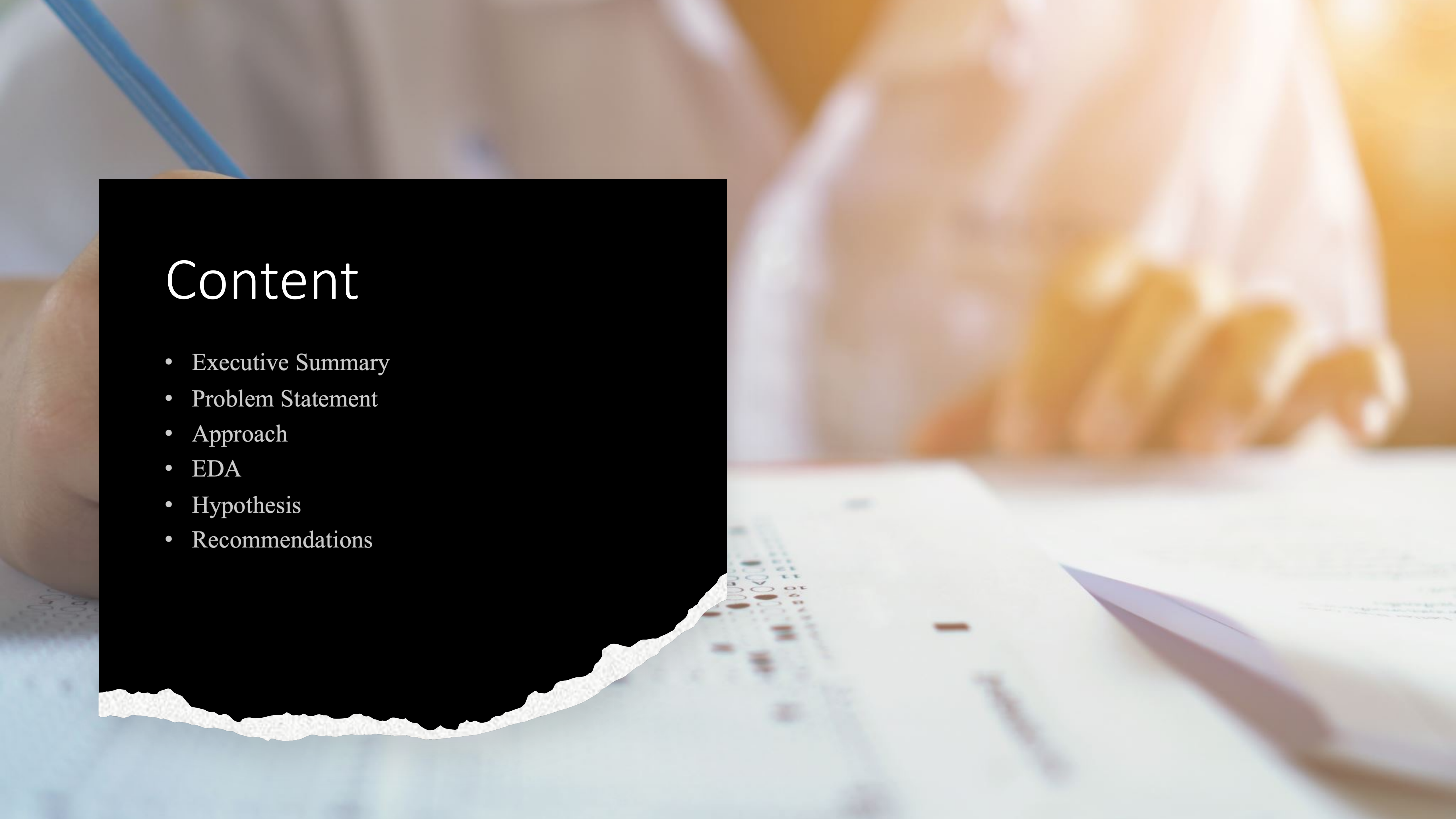
Exploratory Data Analysis

G2M Project for Data Analysts Internship

June 20th ,2023

Haoyue Chang



Content

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

Executive Summary

- XYZ, a private equity firm based in the United States, intends to invest in the Cab Industry due to its significant expansion in recent years and the presence of several prominent competitors in the market.
- Four datasets overall

- Cab_Data: 7 columns and 359,392 rows, consisting of float64(3), int64(2), and object(2) data types.
- Customer_ID: 4 columns and 49,171 rows, consisting of int64(3) and object(1) data types.
- Transaction_ID: 3 columns and 440,098 rows, consisting of int64(2) and object(1) data types.
- City: 3 columns and 20 rows, consisting of object(3) data types.

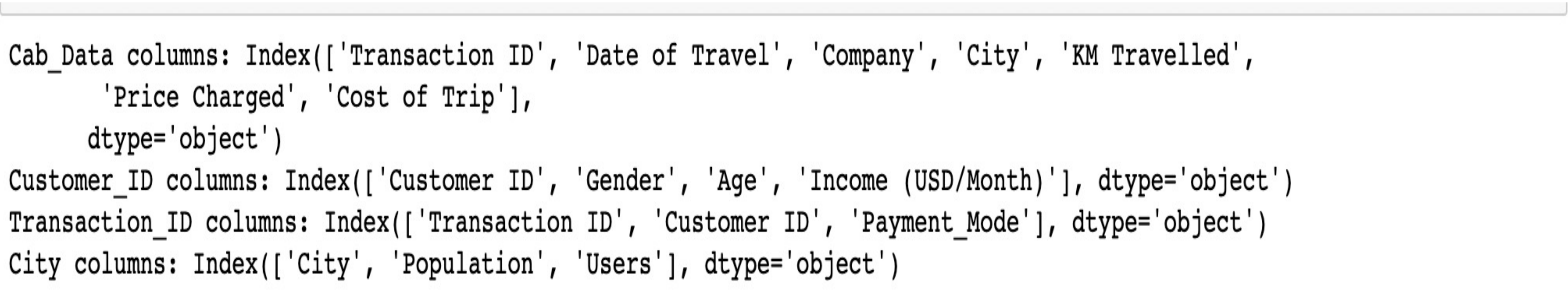
Problem Statement

Help XYZ company to choose a proper company to make an investment in Cab industry

Approach

- Data Loading
- Data Exploration
- Data Merging
- Exploratory Data Analysis (EDA) & Hypothesis

Data Loading



```
Cab_Data columns: Index(['Transaction ID', 'Date of Travel', 'Company', 'City', 'KM Travelled',
       'Price Charged', 'Cost of Trip'],
      dtype='object')
Customer_ID columns: Index(['Customer ID', 'Gender', 'Age', 'Income (USD/Month)'], dtype='object')
Transaction_ID columns: Index(['Transaction ID', 'Customer ID', 'Payment_Mode'], dtype='object')
City columns: Index(['City', 'Population', 'Users'], dtype='object')
```

Cab_Data contains ‘Transaction ID’, ‘Date of Travel’, ‘Company’, ‘City’, ‘KM Travelled’, ‘Price Charged’, ‘Cost of Trip’

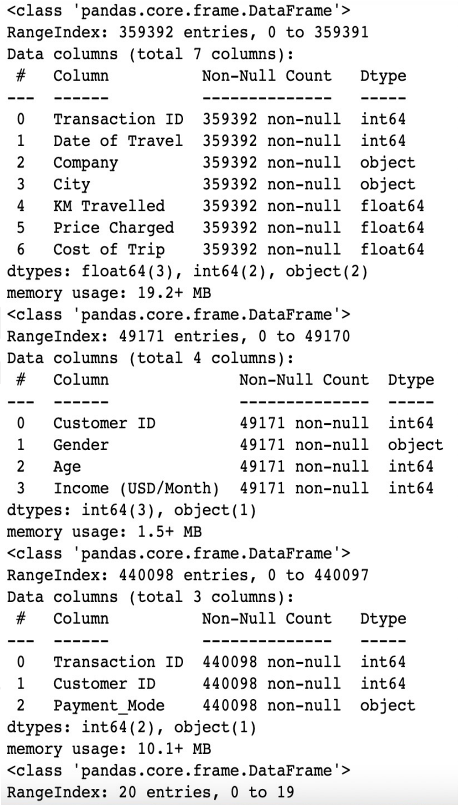
Customer_ID contains 'Customer ID', 'Gender', 'Age', 'Income (USD/Month)'

Transaction_ID contains 'Transaction ID', 'Customer ID', 'Payment_Mode'

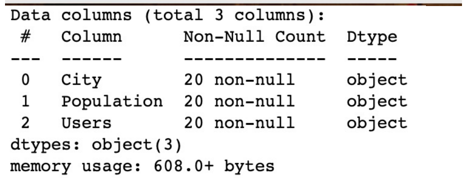
City contains 'City', 'Population', 'Users'

Data Exploration

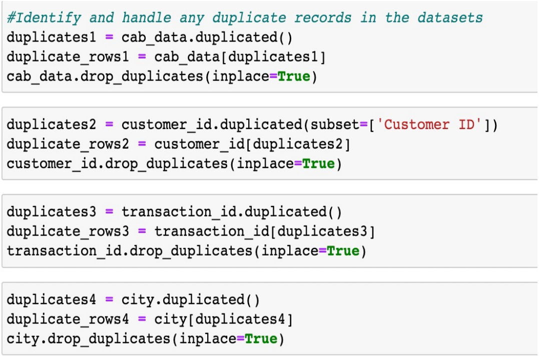
- cab_data:7 columns and 359392 rows,float64(3), int64(2), object(2);
- customer_id:4 columns, 49171 rows, int64(3), object(1)
- transaction_id:3 columns, 440098 rows, int64(2), object(1)
- city:3 columns, 20 rows, object(3)
- No missing values in all four datasets
- Dropped duplicates in datasets



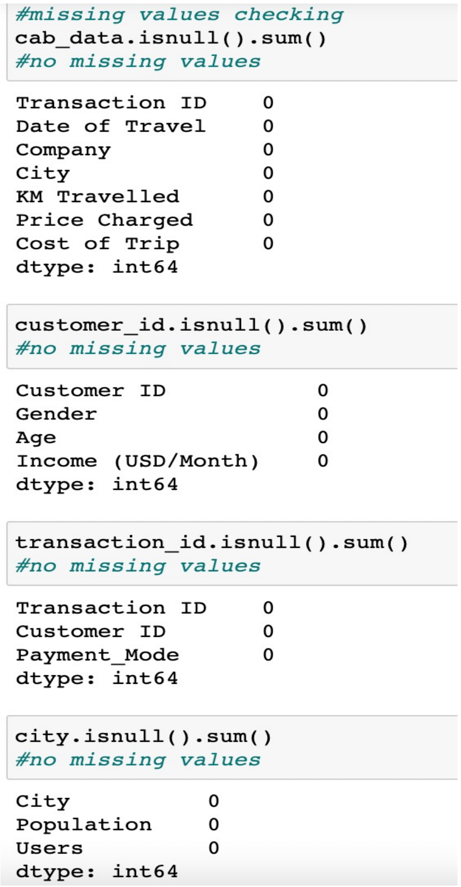
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 359392 entries, 0 to 359391
Data columns (total 7 columns):
 #   Column          Non-Null Count   Dtype  
---  ------          --------------   -----  
 0   Transaction ID  359392 non-null  int64  
 1   Date of Travel  359392 non-null  int64  
 2   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
dtypes: float64(3), int64(2), object(2)
memory usage: 19.2+ MB
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 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: 1.5+ MB
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 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: 10.1+ MB
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 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: 608.0+ bytes
```



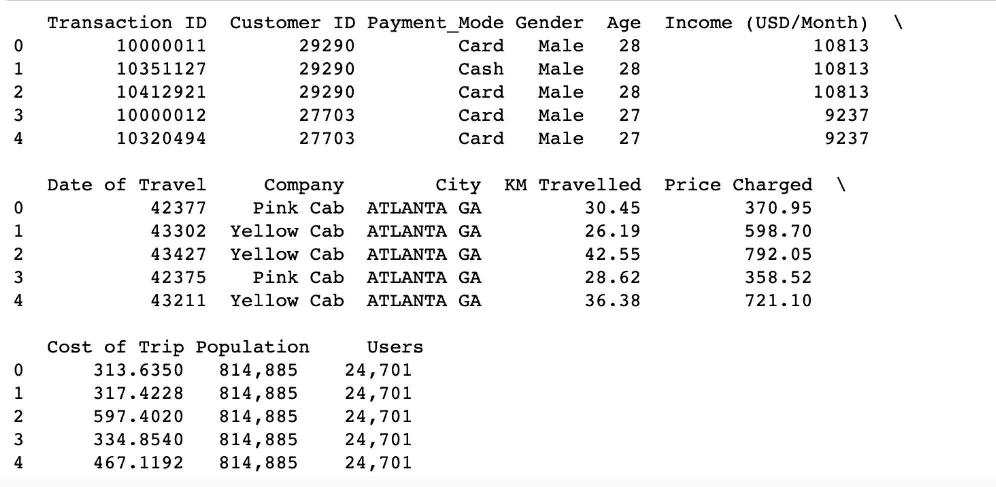
```
#Identify and handle any duplicate records in the datasets
duplicates1 = cab_data.duplicated()
duplicate_rows1 = cab_data[duplicates1]
cab_data.drop_duplicates(inplace=True)

duplicates2 = customer_id.duplicated(subset=['Customer ID'])
duplicate_rows2 = customer_id[duplicates2]
customer_id.drop_duplicates(inplace=True)

duplicates3 = transaction_id.duplicated()
duplicate_rows3 = transaction_id[duplicates3]
transaction_id.drop_duplicates(inplace=True)

duplicates4 = city.duplicated()
duplicate_rows4 = city[duplicates4]
city.drop_duplicates(inplace=True)
```



```
#missing values checking
cab_data.isnull().sum()
#no missing values

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

customer_id.isnull().sum()
#no missing values

Customer ID           0
Gender                0
Age                   0
Income (USD/Month)    0
dtype: int64

transaction_id.isnull().sum()
#no missing values

Transaction ID    0
Customer ID       0
Payment_Mode      0
dtype: int64

city.isnull().sum()
#no missing values

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

Data Merging

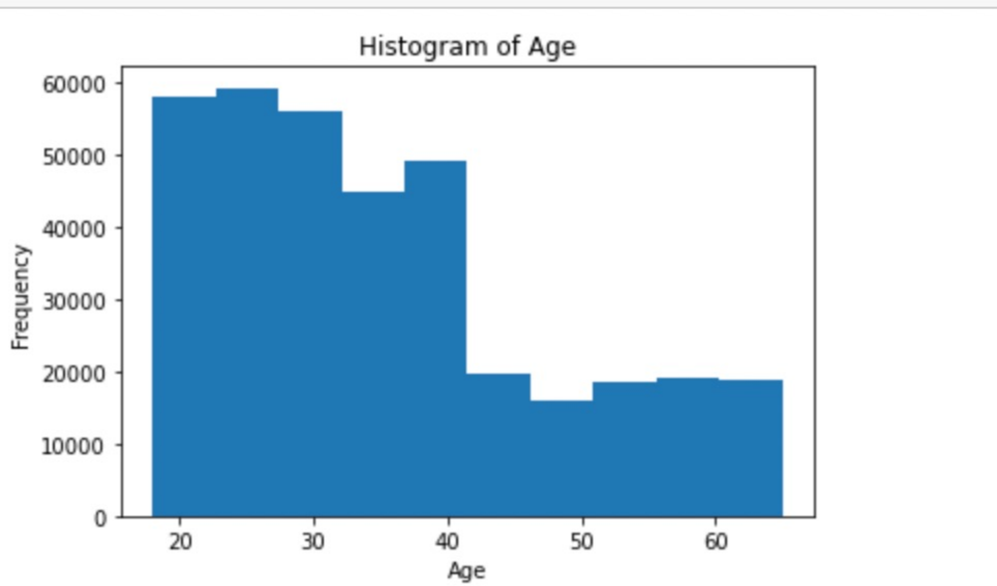
- Datasets were merged based on common columns to create a consolidated dataset for analysis
- Merged Transaction_Id.csv with Customer_ID.csv based on Customer ID
- Merged the new data with Cab_data.csv based on Transaction ID
- Merged the new data with City.csv based on City



```
   Transaction ID  Customer ID Payment_Mode Gender  Age  Income (USD/Month)  \
0        10000011        29290         Card   Male   28               10813
1        10351127        29290         Cash   Male   28               10813
2        10412921        29290         Card   Male   28               10813
3        10000012        27703         Card   Male   27                9237
4        10320494        27703         Card   Male   27                9237

   Date of Travel     Company        City  KM Travelled  Price Charged  \
0           42377    Pink Cab  ATLANTA GA         30.45         370.95
1           43302  Yellow Cab  ATLANTA GA         26.19         598.70
2           43427  Yellow Cab  ATLANTA GA         42.55         792.05
3           42375    Pink Cab  ATLANTA GA         28.62         358.52
4           43211  Yellow Cab  ATLANTA GA         36.38         721.10

   Cost of Trip Population     Users
0      313.6350    814,885    24,701
1      317.4228    814,885    24,701
2      597.4020    814,885    24,701
3      334.8540    814,885    24,701
4      467.1192    814,885    24,701
```

EDA

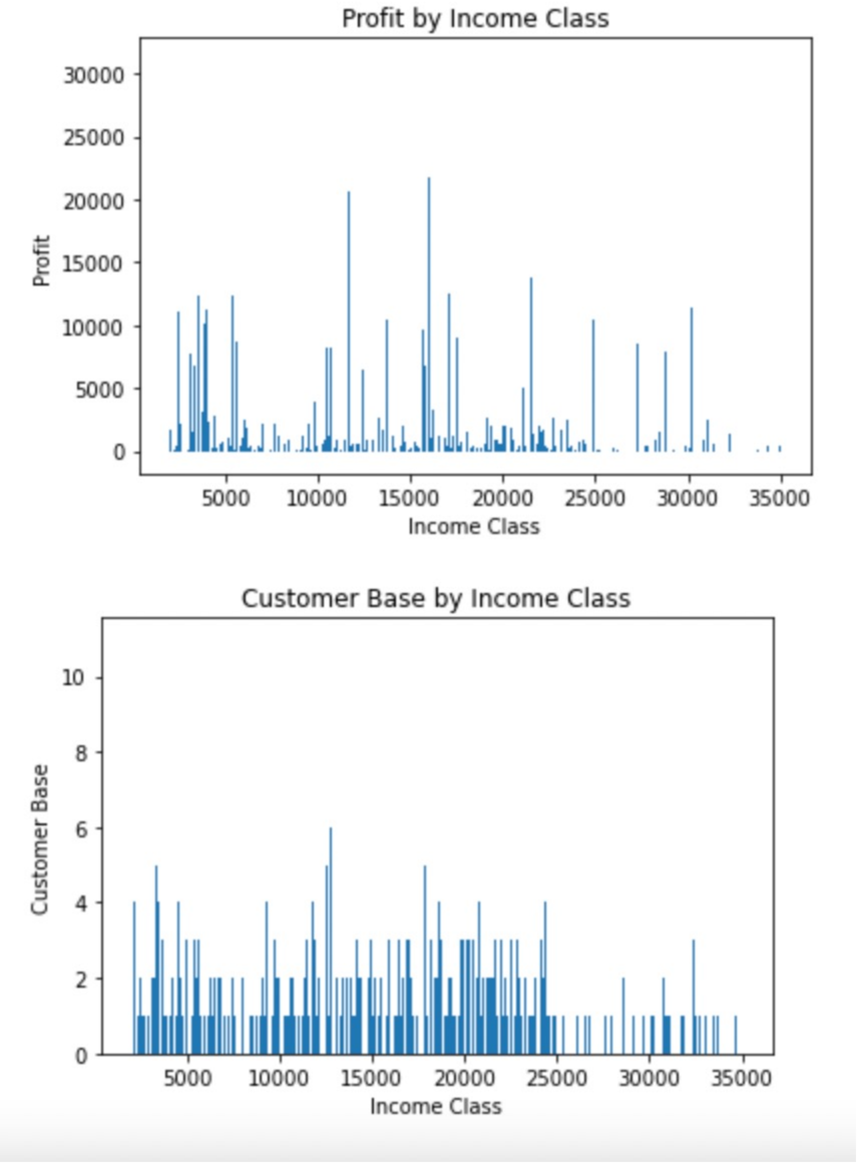
- The distribution of ages appears to be skewed towards the younger age range, indicating a potential trend of younger individuals utilizing cab services more frequently compared to older age groups.
- A significant portion of the cab users in the dataset falls within the younger age brackets.



EDA

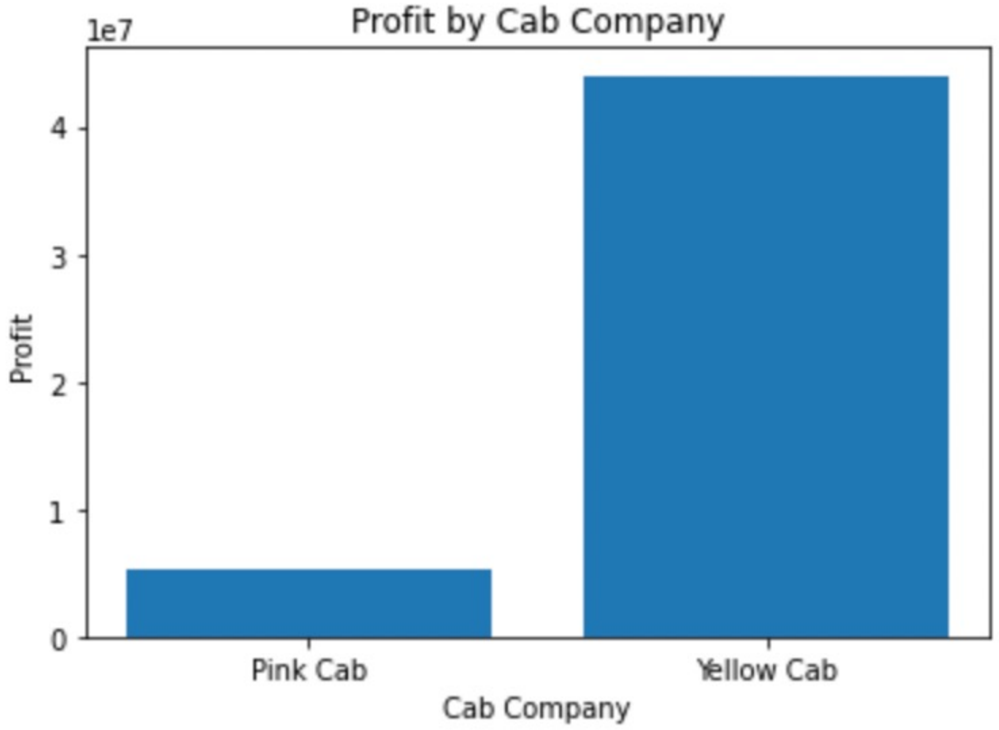
- The bar plot shows the distribution of profit across different income classes.
- Higher income classes tend to generate higher profits, as indicated by the taller bars for higher income classes.

- The bar plot illustrates the distribution of the customer base across income classes.
- The bars represent the number of unique customers in each income class.



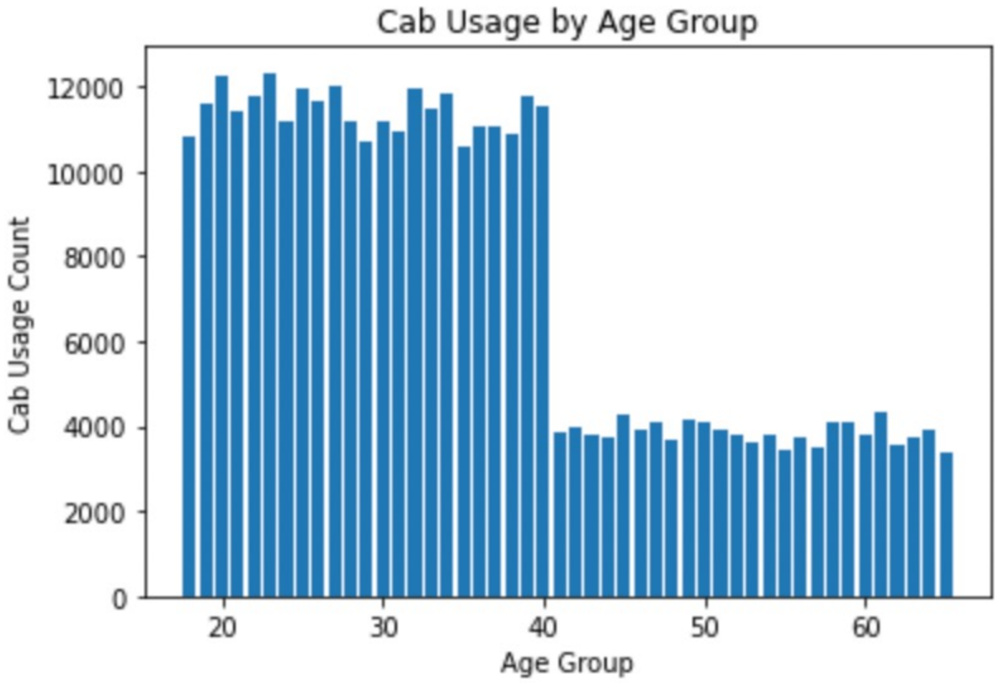
Hypothesis 1: Profitability varies based on the cab company

- The bar plot visualizes the profit by the cab company
- X-axis represents the cab companies
- Y-axis represents the profit
- It is evident that profitability does vary based on the cab company



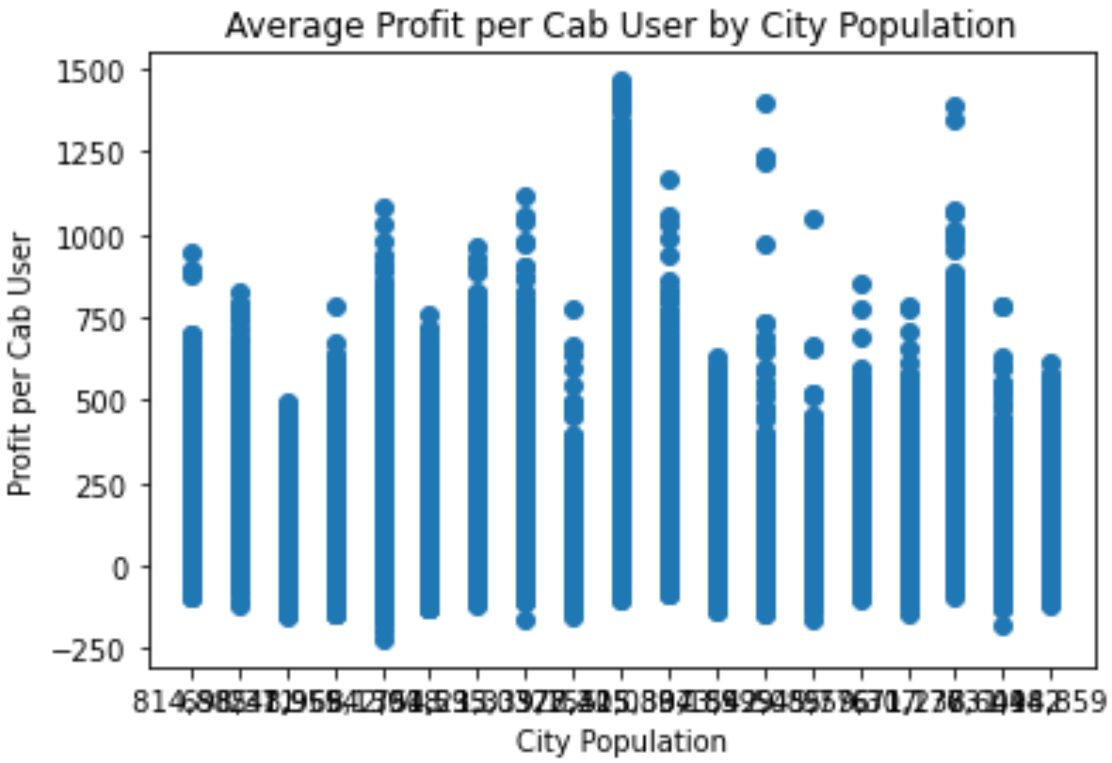
Hypothesis 2: Customer demographics impact cab usage

- The bar plot visualizes the profit by the cab company
- X-axis represents the cab companies
- Y-axis represents the profit
- It is evident that profitability does vary based on the cab company



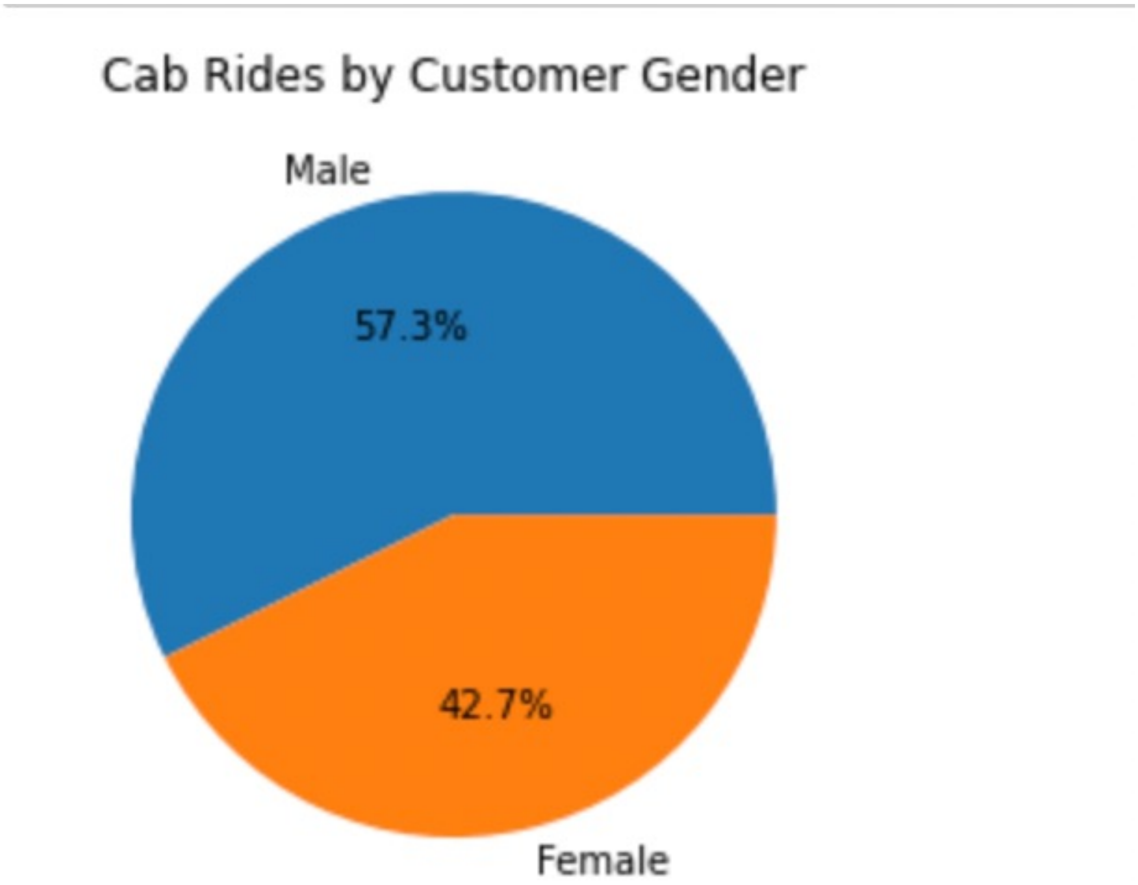
Hypothesis 3:Average profit per cab user varies by city population

- There is no clear pattern or trend indicating a strong relationship between city population and average profit per cab user
- There is no significant correlation or variation between city population and average profit per cab user.



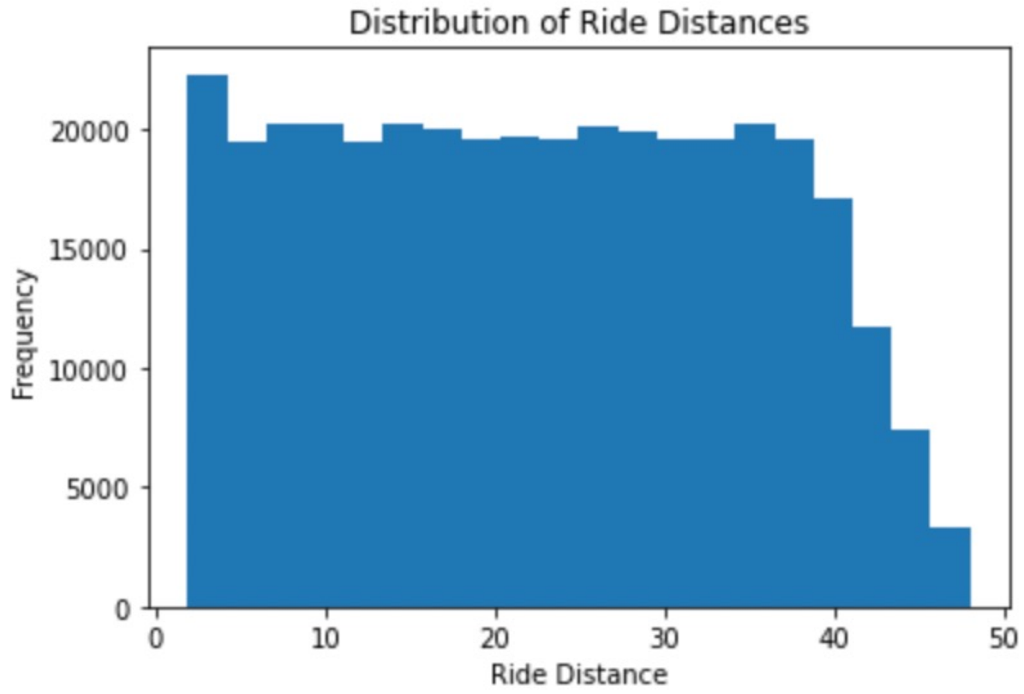
Hypothesis 4: per cab user by customer gender

- It is evident that cab rides do vary based on customer gender



Hypothesis 5: Customer analysis varies based on ride distance

- X-axis represents the ride distance, divided into different bins,
- Y-axis represents the frequency or count of ride distances falling within each bin.
- We can gain an understanding of the spread and concentration of ride distances



Recommendation

- Cab Company Profitability: It's clear that profitability varies among cab companies. XYZ firm should invest in companies that have shown higher profitability.
- Customer Demographics and Usage: Demographic factors like age and gender impact cab usage. XYZ firm should customize marketing strategies and services to cater to specific customer segments.
- Customer Preferences and Ride Distance: Ride distances reveal valuable insights into customer preferences. XYZ firm should use this information to optimize operations and service offerings.
- Exploring New Markets: The analysis highlights untapped potential among certain income groups. XYZ firm should target these segments by offering tailored services, discounts, or loyalty programs.
- Continuous Data Analysis: Regularly analyzing data and monitoring market trends, customer preferences, and competition is crucial for XYZ firm to make informed investment decisions and identify emerging opportunities.
- Strategic Partnerships and Acquisitions: XYZ firm can consider forming strategic partnerships or acquiring cab companies that align with their investment goals and have a strong market presence. This can facilitate expansion, access to new customer segments, and synergistic growth.

Thank you