Exploratory Data Analysis

G2M Case Study

Sarra Othmani - Tunisia

21.07.2022

Agenda

Executive Summary

Problem Statement

Approach

EDA

EDA Summary

Recommendations

Agenda

Executive Summary

Problem Statement

Approach

EDA

EDA Summary

Recommendations

The cab Industry Use case

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.

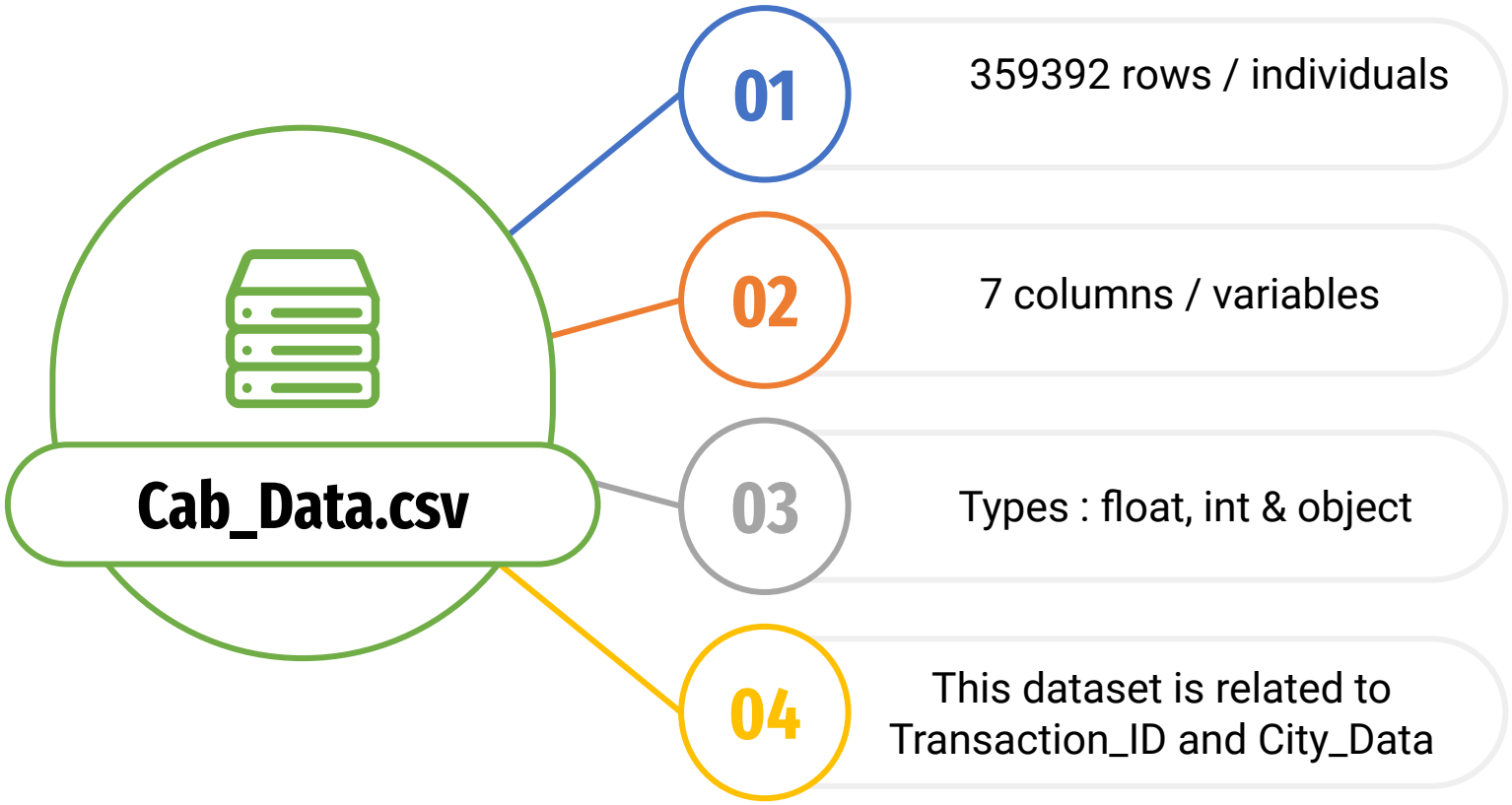
The outcome of the project's delivery will be an insight for the Cab investment firm, in order to take the right and best decision.

Source documentation

4 individual datasets :

- **Cab_Data.csv** → details of transaction for 2 cab companies
- **Customer_ID.csv** → a mapping table that contains a unique identifier which links the customer's demographic details
- **Transaction_ID.csv** → a mapping table that contains transaction to customer mapping and payment mode
- **City.csv** → list of US cities, their population and number of cab users

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




```
<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
```

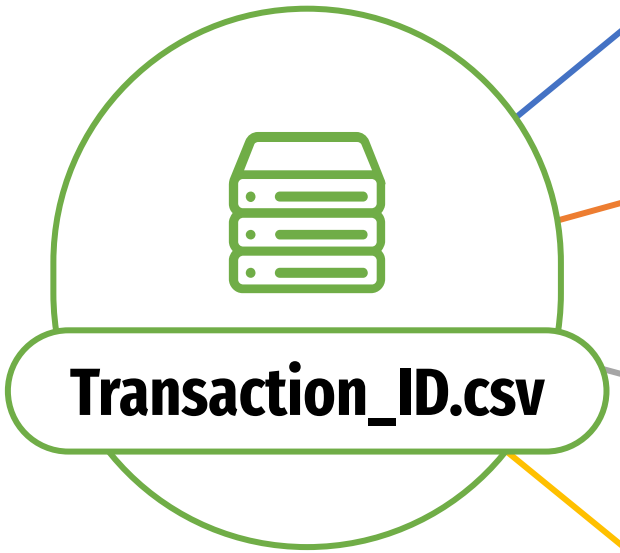
01 49171 rows / individuals

02 4 columns / variables

03 Types : int & object

04 This dataset is related to Transaction_ID

```
<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
```



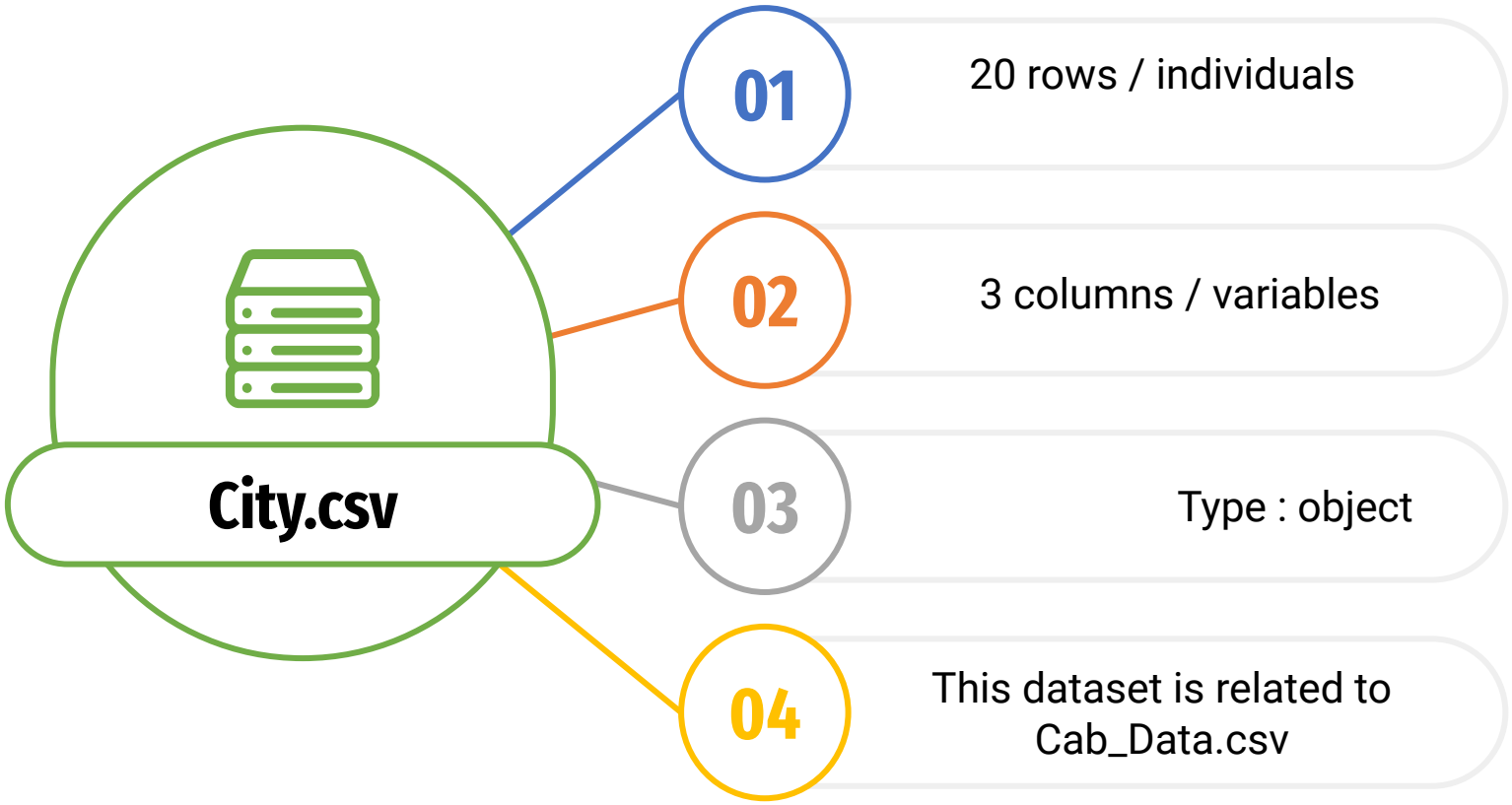
01 440098 rows / individuals

02 3 columns / variables

03 Types : int & object

04 This dataset is related to Customer_ID.csv

```
<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
```

Assumptions

- The exploratory analysis must be done on once and for all on the four provided dataset. Therefore, we should combine them in a global dataset that we will name df (data frame) by using the merge function : merge() -> MASTER DATA.
- The Price_Charged column presents outliers (outlier detection).
- Users of City dataset is a non-necessary feature as it’s not correlated or related to other columns. Therefore, we can drop this column in the analysis.

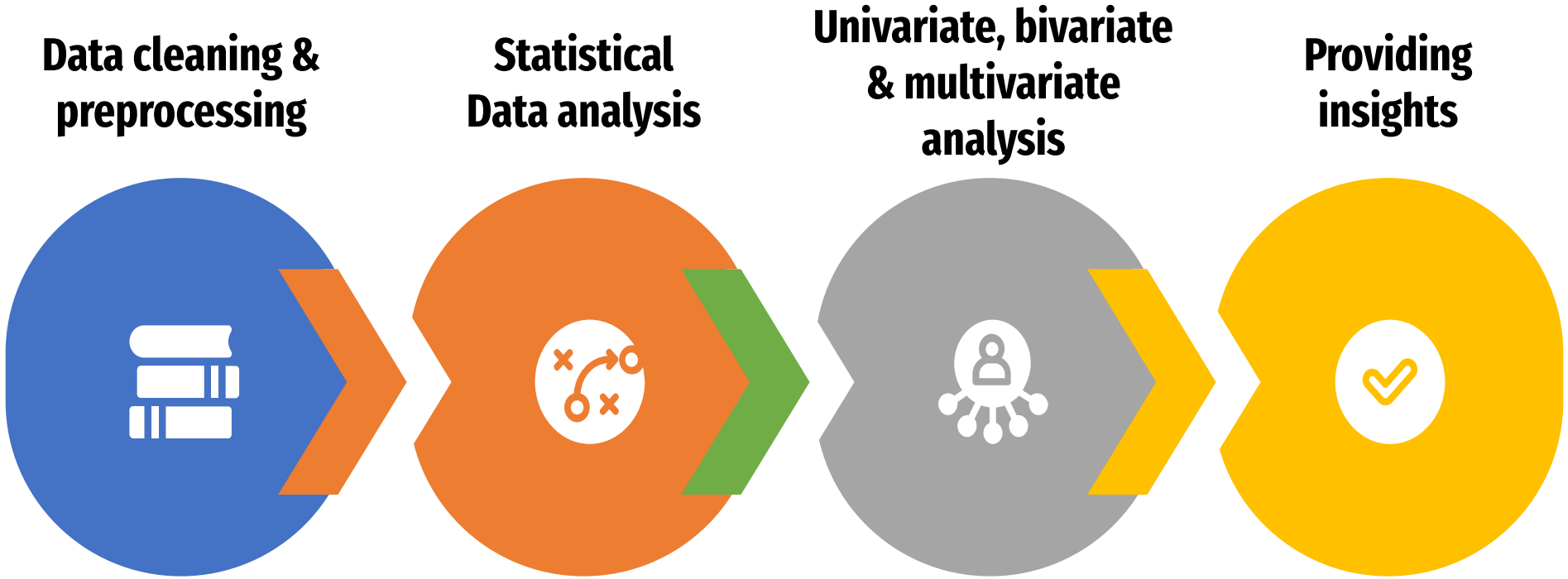
Field / Features transformations

Conversion of Date of Travel feature from xlrd to python date time format.

4 datasets are clean : no missing values & no duplicate rows.

Conversion of Population column type from object to float64.

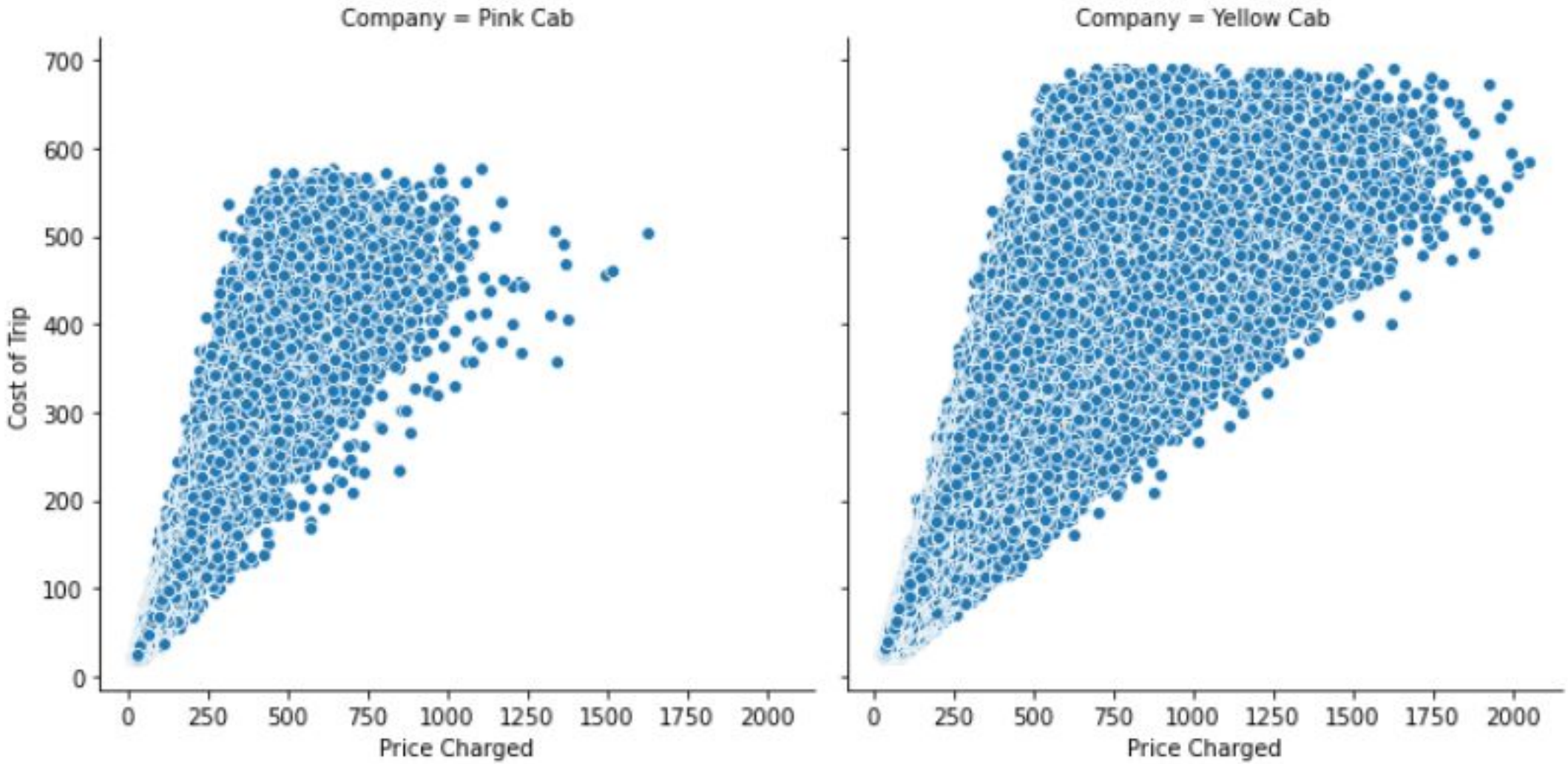
Data Analysis | The process



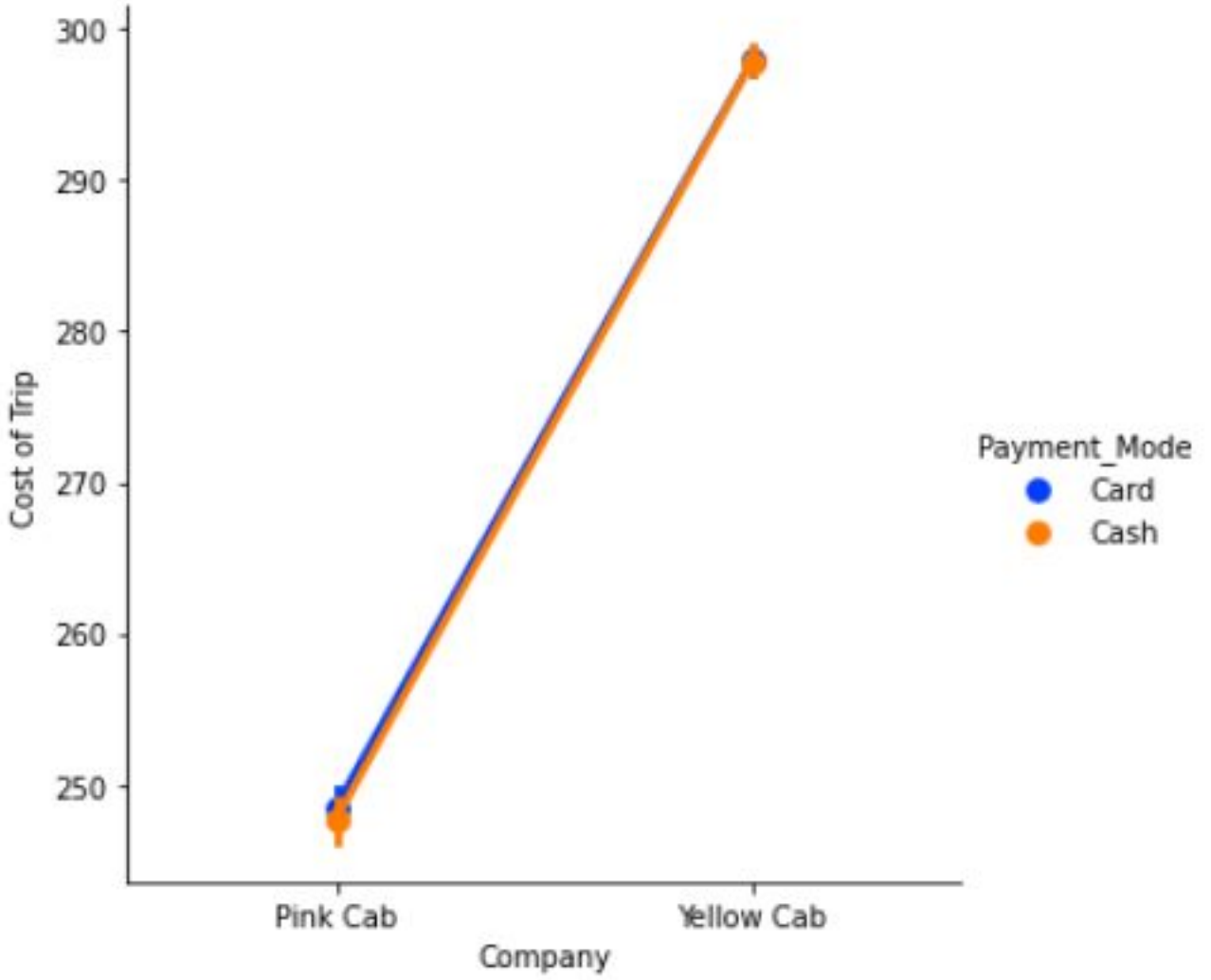
Correlation between all features

	Transaction ID	Date of Travel	KM Travelled	Price Charged	Cost of Trip	Customer ID	Age	Income (USD/Month)
Transaction ID	1.000000	0.993030	-0.001429	-0.052902	-0.003462	-0.016912	-0.001267	-0.001570
Date of Travel	0.993030	1.000000	-0.001621	-0.055559	-0.004484	-0.017653	-0.001346	-0.001368
KM Travelled	-0.001429	-0.001621	1.000000	0.835753	0.981848	0.000389	-0.000369	-0.000544
Price Charged	-0.052902	-0.055559	0.835753	1.000000	0.859812	-0.177324	-0.003084	0.003228
Cost of Trip	-0.003462	-0.004484	0.981848	0.859812	1.000000	0.003077	-0.000189	-0.000633
Customer ID	-0.016912	-0.017653	0.000389	-0.177324	0.003077	1.000000	-0.004735	-0.013608
Age	-0.001267	-0.001346	-0.000369	-0.003084	-0.000189	-0.004735	1.000000	0.003907
Income (USD/Month)	-0.001570	-0.001368	-0.000544	0.003228	-0.000633	-0.013608	0.003907	1.000000

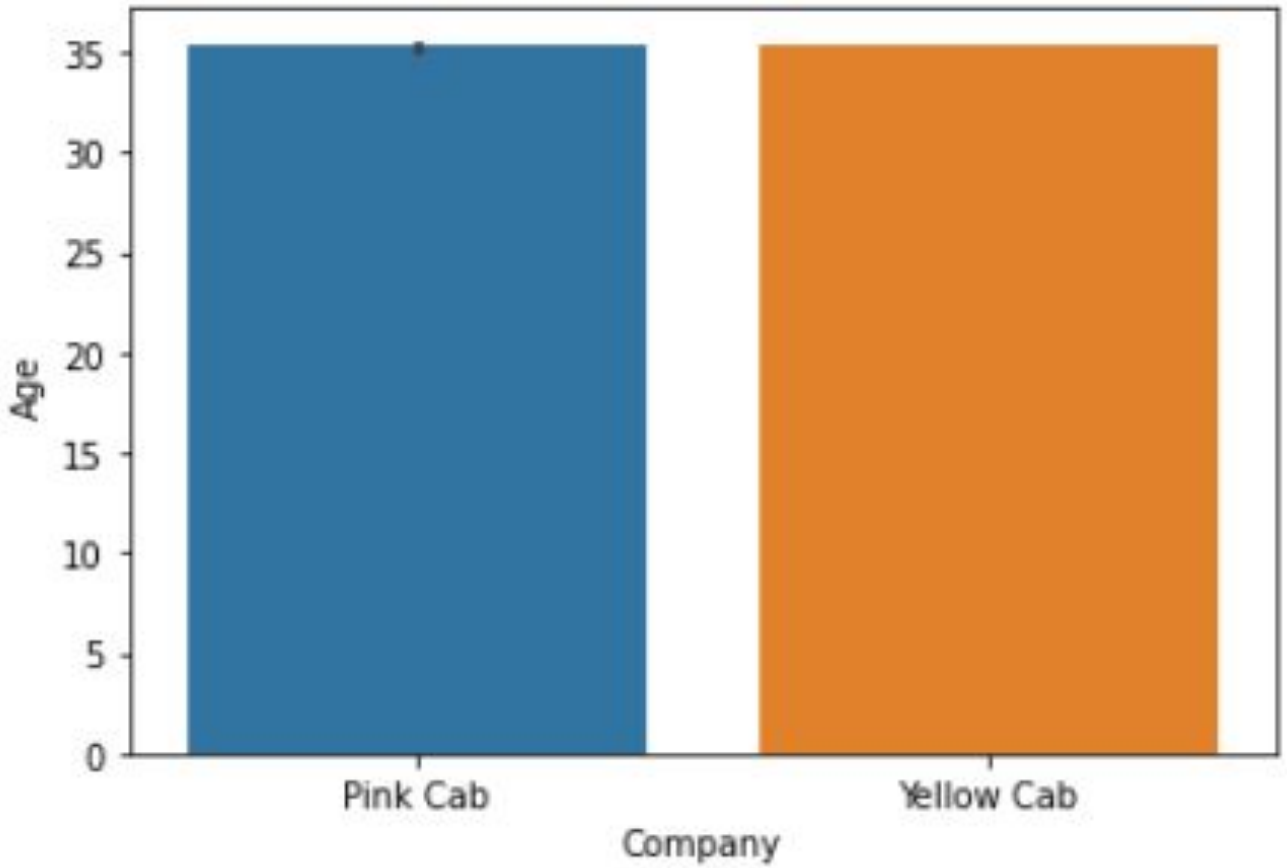
Price and cost of Pink & Yellow cabs analysis



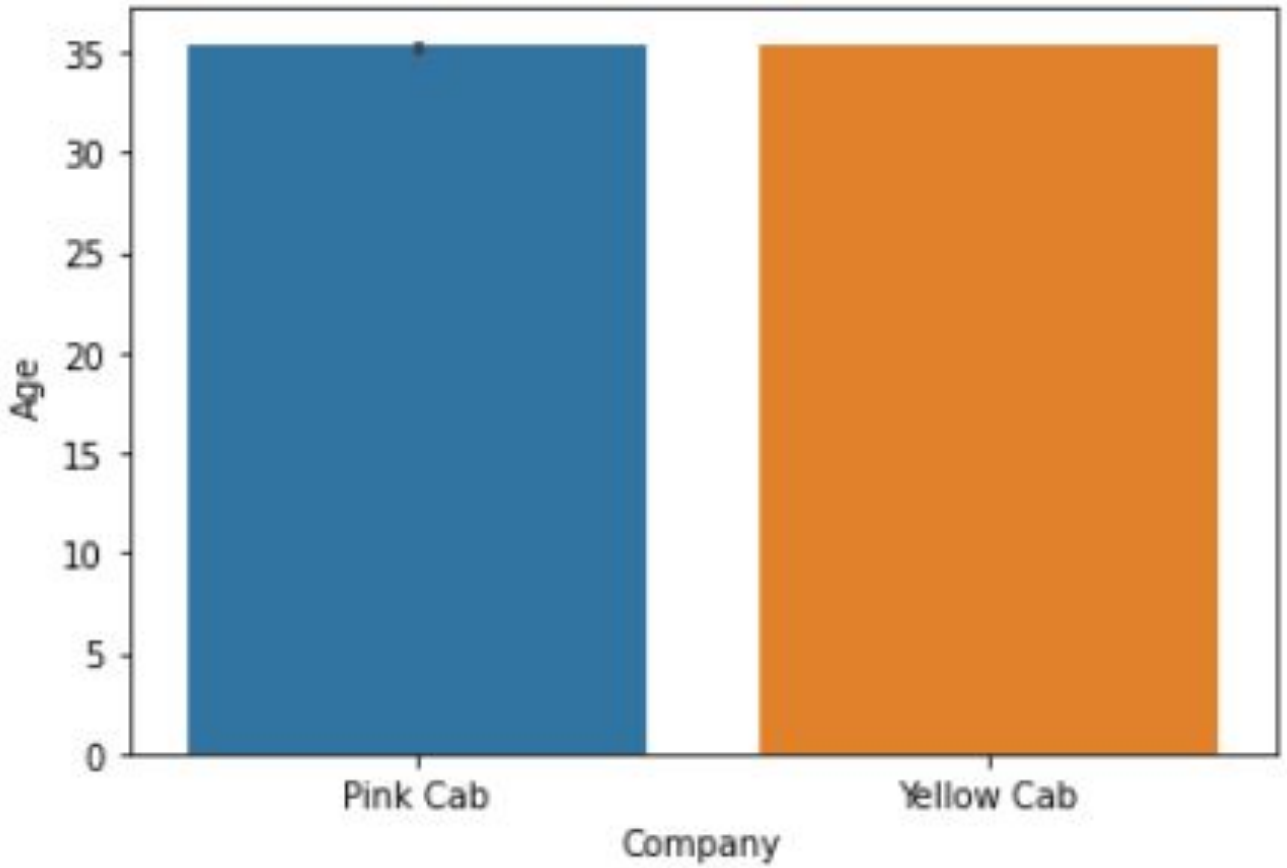
Company analysis based on mode of payment



Company analysis based on age



Company analysis based age



Recommendation :

Yellow Cab firm is the best decision for investment.

Thank You