DATA INTAKE REPORT

NAME: VIVIAN KERUBO MOSOMI

Email: kerubomosomi7@gmail.com

COUNTRY: Kenya

There were 4 datasets:

1. Cab_Data.csv – Has details of transactions for 2 cab companies (Pink Cab and Yellow Cab)

2. Customer_ID.csv – Contains a unique identifier which links the customer's demographic details eg Age, Gender and Income

3. Transaction_ID.csv – Contains transaction to customer mapping and payment mode of the customer

4. City.csv – Has details of list of US cities, their population and number of cab users

1.Cab Data

Has 359392 rows and 7 columns

```
The info of cab_data:
<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
```

There were no null values and duplicates in the dataset

The only column that had outliers is Price Charged

2.City Data

Has 3 columns and 19 rows

The 3 columns are City, Population and Users

This dataset had no null values or duplicates

The data information is as follows:

```
<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
```

3. Customer Data

Has 49171 rows and 4 columns

This dataset has no duplicates and null values

```
Information of the dataset is:
<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
```

4. Transaction data

Has 440098 rows and 3 columns

Has no duplicate and null values

Has the following information:

```
<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
```