In [1]:

```
import pandas as pd
import numpy as np
```

In [2]:

```
cab = pd.read_csv('Cab_Data.csv')
cab.head()
```

Out[2]:

|   | Transaction ID | Date of Travel | Company | City | KM Travelled | Price Charged | Cost of Trip |
|---|---|---|---|---|---|---|---|
| **0** | 10000011 | 42377 | Pink Cab | ATLANTA GA | 30.45 | 370.95 | 313.635 |
| **1** | 10000012 | 42375 | Pink Cab | ATLANTA GA | 28.62 | 358.52 | 334.854 |
| **2** | 10000013 | 42371 | Pink Cab | ATLANTA GA | 9.04 | 125.20 | 97.632 |
| **3** | 10000014 | 42376 | Pink Cab | ATLANTA GA | 33.17 | 377.40 | 351.602 |
| **4** | 10000015 | 42372 | Pink Cab | ATLANTA GA | 8.73 | 114.62 | 97.776 |

In [3]:

```
city = pd.read_csv('City.csv')
city.head()
```

Out[3]:

|   | City | Population | Users |
|---|---|---|---|
| **0** | NEW YORK NY | 8,405,837 | 302,149 |
| **1** | CHICAGO IL | 1,955,130 | 164,468 |
| **2** | LOS ANGELES CA | 1,595,037 | 144,132 |
| **3** | MIAMI FL | 1,339,155 | 17,675 |
| **4** | SILICON VALLEY | 1,177,609 | 27,247 |

In [4]:

```
customer = pd.read_csv('Customer_ID.csv')
customer.head()
```

Out[4]:

|   | Customer ID | Gender | Age | Income (USD/Month) |
|---|---|---|---|---|
| **0** | 29290 | Male | 28 | 10813 |
| **1** | 27703 | Male | 27 | 9237 |
| **2** | 28712 | Male | 53 | 11242 |
| **3** | 28020 | Male | 23 | 23327 |
| **4** | 27182 | Male | 33 | 8536 |

In [5]:

```
transaction = pd.read_csv('Transaction_ID.csv')
transaction.head()
```

Out[5]:

|   | Transaction ID | Customer ID | Payment_Mode |
|---|---|---|---|
| **0** | 10000011 | 29290 | Card |
| **1** | 10000012 | 27703 | Card |
| **2** | 10000013 | 28712 | Cash |
| **3** | 10000014 | 28020 | Cash |
| **4** | 10000015 | 27182 | Card |

After taking a look of each dataset, we will then see each attribute in each dataset

```
Cab_Data.csv
```

In [6]:
```
print(cab.shape)
```

```
(359392, 7)
```

In [7]:
```
print(cab.dtypes)
```

```
Transaction ID      int64
Date of Travel      int64
Company            object
City               object
KM Travelled      float64
Price Charged     float64
Cost of Trip      float64
dtype: object
```

```
City.csv
```

In [8]:
```
print(city.shape)
```

```
(20, 3)
```

In [9]:
```
print(city.dtypes)
```

```
City          object
Population    object
Users         object
dtype: object
```

```
Customer_ID.csv
```

In [10]:
```
print(customer.shape)
```

```
(49171, 4)
```

In [11]:
```
print(customer.dtypes)
```

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

```
Transaction_ID.csv
```

In [12]:
```
print(transaction.shape)
```

```
(440098, 3)
```

In [13]:
```
print(transaction.dtypes)
```

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

After seeing the info of variables, we are going to change the variable type and join the datasets

As we can see, the customer_ID and transaction_ID both have Customer ID as variable; thus we can join these two datasets on Customer ID

In [14]:
```
customer_tran = customer.merge(transaction, on='Customer ID', how='left')
customer_tran.head()
```

Out[14]:

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

After joining these two datasets, we can drop any na values and duplicates

In [15]:
```
print(customer_tran.shape)
```
```
(440098, 6)
```

In [16]:
```
cu_tran = customer_tran.dropna()
print(cu_tran.shape)
```
```
(440098, 6)
```

In [17]:
```
cu_tran_uni = cu_tran.drop_duplicates(subset=['Customer ID', 'Transaction ID'], keep='last')
print(cu_tran_uni.shape)
```
```
(440098, 6)
```

Next we can join Cab_user and city together since they both have city variable in common

In [18]:
```
cab_city = cab.merge(city, on='City', how='left')
cab_city.head()
```

Out[18]:

| | Transaction ID | Date of Travel | Company | City | KM Travelled | Price Charged | Cost of Trip | Population | Users |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 10000011 | 42377 | Pink Cab | ATLANTA GA | 30.45 | 370.95 | 313.635 | 814,885 | 24,701 |
| **1** | 10000012 | 42375 | Pink Cab | ATLANTA GA | 28.62 | 358.52 | 334.854 | 814,885 | 24,701 |
| **2** | 10000013 | 42371 | Pink Cab | ATLANTA GA | 9.04 | 125.20 | 97.632 | 814,885 | 24,701 |
| **3** | 10000014 | 42376 | Pink Cab | ATLANTA GA | 33.17 | 377.40 | 351.602 | 814,885 | 24,701 |
| **4** | 10000015 | 42372 | Pink Cab | ATLANTA GA | 8.73 | 114.62 | 97.776 | 814,885 | 24,701 |

In [19]:
```
print(cab_city.shape)
```
```
(359392, 9)
```

drop na and duplicates

In [20]:
```
ca_ci = cab_city.dropna()
ca_ci_uni = ca_ci.drop_duplicates(subset=['Transaction ID', 'Date of Travel'], keep='last')
print(ca_ci_uni.shape)
```
```
(359392, 9)
```

Finally, join these two together

```
In [21]: df = cab_city = cu_tran_uni.merge(ca_ci_uni, on='Transaction ID', how='left')
         print(df.shape)

         (440098, 14)
```

```
In [22]: df.head()
```

Out[22]:

| | Customer ID | Gender | Age | Income (USD/Month) | Transaction ID | Payment_Mode | Date of Travel | Company | City | KM Travelled | Price Charged | Cost of Trip | Po |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 29290 | Male | 28 | 10813 | 10000011 | Card | 42377.0 | Pink Cab | ATLANTA GA | 30.45 | 370.95 | 313.6350 | |
| **1** | 29290 | Male | 28 | 10813 | 10351127 | Cash | 43302.0 | Yellow Cab | ATLANTA GA | 26.19 | 598.70 | 317.4228 | |
| **2** | 29290 | Male | 28 | 10813 | 10412921 | Card | 43427.0 | Yellow Cab | ATLANTA GA | 42.55 | 792.05 | 597.4020 | |
| **3** | 27703 | Male | 27 | 9237 | 10000012 | Card | 42375.0 | Pink Cab | ATLANTA GA | 28.62 | 358.52 | 334.8540 | |
| **4** | 27703 | Male | 27 | 9237 | 10320494 | Card | 43211.0 | Yellow Cab | ATLANTA GA | 36.38 | 721.10 | 467.1192 | |

```
In [23]: df.isnull().values.any()  # check null value

Out[23]: True
```

```
In [24]: df1 = df.dropna()
         print(df1.shape)

         (359392, 14)
```

```
In [25]: df1.dtypes

Out[25]: Customer ID             int64
         Gender                 object
         Age                     int64
         Income (USD/Month)      int64
         Transaction ID          int64
         Payment_Mode           object
         Date of Travel        float64
         Company                object
         City                   object
         KM Travelled          float64
         Price Charged         float64
         Cost of Trip          float64
         Population             object
         Users                  object
         dtype: object
```

After we have the final dataframe, we can then do some manipulations for this dataframe

First we can drop the columns that we don't need for analysis

```
In [26]: df2 = df1.drop(['Population','Users'], axis=1)
```

In [27]:
```
df2.head()
```

Out[27]:

| | Customer ID | Gender | Age | Income (USD/Month) | Transaction ID | Payment_Mode | Date of Travel | Company | City | KM Travelled | Price Charged | Cost of Trip |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 29290 | Male | 28 | 10813 | 10000011 | Card | 42377.0 | Pink Cab | ATLANTA GA | 30.45 | 370.95 | 313.6350 |
| **1** | 29290 | Male | 28 | 10813 | 10351127 | Cash | 43302.0 | Yellow Cab | ATLANTA GA | 26.19 | 598.70 | 317.4228 |
| **2** | 29290 | Male | 28 | 10813 | 10412921 | Card | 43427.0 | Yellow Cab | ATLANTA GA | 42.55 | 792.05 | 597.4020 |
| **3** | 27703 | Male | 27 | 9237 | 10000012 | Card | 42375.0 | Pink Cab | ATLANTA GA | 28.62 | 358.52 | 334.8540 |
| **4** | 27703 | Male | 27 | 9237 | 10320494 | Card | 43211.0 | Yellow Cab | ATLANTA GA | 36.38 | 721.10 | 467.1192 |

Since we have price charged and cost, we can calculate the benefit

In [28]:
```
df2["benefit"] = df2["Price Charged"] - df2["Cost of Trip"]
df2.head()
```

Out[28]:

| | Customer ID | Gender | Age | Income (USD/Month) | Transaction ID | Payment_Mode | Date of Travel | Company | City | KM Travelled | Price Charged | Cost of Trip | I |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 29290 | Male | 28 | 10813 | 10000011 | Card | 42377.0 | Pink Cab | ATLANTA GA | 30.45 | 370.95 | 313.6350 | 5 |
| **1** | 29290 | Male | 28 | 10813 | 10351127 | Cash | 43302.0 | Yellow Cab | ATLANTA GA | 26.19 | 598.70 | 317.4228 | 28 |
| **2** | 29290 | Male | 28 | 10813 | 10412921 | Card | 43427.0 | Yellow Cab | ATLANTA GA | 42.55 | 792.05 | 597.4020 | 19 |
| **3** | 27703 | Male | 27 | 9237 | 10000012 | Card | 42375.0 | Pink Cab | ATLANTA GA | 28.62 | 358.52 | 334.8540 | 2 |
| **4** | 27703 | Male | 27 | 9237 | 10320494 | Card | 43211.0 | Yellow Cab | ATLANTA GA | 36.38 | 721.10 | 467.1192 | 25 |

In [29]:
```
df2['benefit(+/-)'] = np.where(df2['benefit'] > 0, 'Positive', 'Negative')
df2.head()
```

Out[29]:

| | Customer ID | Gender | Age | Income (USD/Month) | Transaction ID | Payment_Mode | Date of Travel | Company | City | KM Travelled | Price Charged | Cost of Trip | I |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 29290 | Male | 28 | 10813 | 10000011 | Card | 42377.0 | Pink Cab | ATLANTA GA | 30.45 | 370.95 | 313.6350 | 5 |
| **1** | 29290 | Male | 28 | 10813 | 10351127 | Cash | 43302.0 | Yellow Cab | ATLANTA GA | 26.19 | 598.70 | 317.4228 | 28 |
| **2** | 29290 | Male | 28 | 10813 | 10412921 | Card | 43427.0 | Yellow Cab | ATLANTA GA | 42.55 | 792.05 | 597.4020 | 19 |
| **3** | 27703 | Male | 27 | 9237 | 10000012 | Card | 42375.0 | Pink Cab | ATLANTA GA | 28.62 | 358.52 | 334.8540 | 2 |
| **4** | 27703 | Male | 27 | 9237 | 10320494 | Card | 43211.0 | Yellow Cab | ATLANTA GA | 36.38 | 721.10 | 467.1192 | 25 |

Next since we have the date of travel variable, we can change it to normal datetime format

In [38]:
```
from datetime import datetime
```

In [39]:
```
df3 = df2
```

In [40]:
```
df3["Date of Travel"] = df3["Date of Travel"].astype("int")
df3["Date of Travel"] = df3["Date of Travel"].apply(datetime.fromordinal)
df3.head()
```

Out[40]:

| | Customer ID | Gender | Age | Income (USD/Month) | Transaction ID | Payment_Mode | Date of Travel | Company | City | KM Travelled | Price Charged | Cost of Trip |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 29290 | Male | 28 | 10813 | 10000011 | Card | 0117-01-09 00:00:00 | Pink Cab | ATLANTA GA | 30.45 | 370.95 | 313.6350 |
| **1** | 29290 | Male | 28 | 10813 | 10351127 | Cash | 0119-07-23 00:00:00 | Yellow Cab | ATLANTA GA | 26.19 | 598.70 | 317.4228 |
| **2** | 29290 | Male | 28 | 10813 | 10412921 | Card | 0119-11-25 00:00:00 | Yellow Cab | ATLANTA GA | 42.55 | 792.05 | 597.4020 |
| **3** | 27703 | Male | 27 | 9237 | 10000012 | Card | 0117-01-07 00:00:00 | Pink Cab | ATLANTA GA | 28.62 | 358.52 | 334.8540 |
| **4** | 27703 | Male | 27 | 9237 | 10320494 | Card | 0119-04-23 00:00:00 | Yellow Cab | ATLANTA GA | 36.38 | 721.10 | 467.1192 |

In [41]:
```
df3["Date of Travel"] = df3["Date of Travel"].astype("str")
df3["year"] = df3["Date of Travel"].str[:4]
df3.year.unique()  # we can see the number of year is 4 not 3, so one of them is outlier that should be
```

Out[41]:
```
array(['0117', '0119', '0118', '0120'], dtype=object)
```

In [42]:
```
print(len(df3[df3["year"]=="0120"]))  # we only have 513 data points for this year, so we have to drop
```

```
513
```

In [43]:
```
df3 = df3.drop(df3[(df3.year == "0120")].index)
df3.shape
```

Out[43]:
```
(358879, 15)
```

In [44]:
```
df3 = df3.drop(['Date of Travel'], axis=1)
df3['year'] = df3['year'].replace(['0117', '0118','0119'], ['2016', '2017','2018'])
df3.head()
```

Out[44]:

| | Customer ID | Gender | Age | Income (USD/Month) | Transaction ID | Payment_Mode | Company | City | KM Travelled | Price Charged | Cost of Trip | benefit | b |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 29290 | Male | 28 | 10813 | 10000011 | Card | Pink Cab | ATLANTA GA | 30.45 | 370.95 | 313.6350 | 57.3150 | |
| **1** | 29290 | Male | 28 | 10813 | 10351127 | Cash | Yellow Cab | ATLANTA GA | 26.19 | 598.70 | 317.4228 | 281.2772 | |
| **2** | 29290 | Male | 28 | 10813 | 10412921 | Card | Yellow Cab | ATLANTA GA | 42.55 | 792.05 | 597.4020 | 194.6480 | |
| **3** | 27703 | Male | 27 | 9237 | 10000012 | Card | Pink Cab | ATLANTA GA | 28.62 | 358.52 | 334.8540 | 23.6660 | |
| **4** | 27703 | Male | 27 | 9237 | 10320494 | Card | Yellow Cab | ATLANTA GA | 36.38 | 721.10 | 467.1192 | 253.9808 | |

In [45]:
```
df3.to_csv('eda.csv')
```

In [ ]:
```

```