# Data Intake Report

Name: Insight For Cab Investment Firm (Week 2 Case Study)
Report date: 9/11/2023
Internship Batch: LISUM25
Version: 1.0
Data intake by: Connor Bryson
Data intake reviewer:
Data storage location:

**Tabular data details:**

**NOTE: The number of rows provided in the tables for each dataset are the amount of rows of the original dataset without removal of duplicate rows. The EDA and presentation will be based on the datasets that have removed duplicate rows.**

**Transaction_ID**

| | |
|---|---|
| **Total number of observations** | 440098 observations (rows) |
| **Total number of files** | |
| **Total number of features** | 3 features (columns) |
| **Base format of the file** | .csv |
| **Size of the data** | 8.58 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
```

## Customer_ID

| | |
|---|---|
| **Total number of observations** | 49171 observations (rows) |
| **Total number of files** | |
| **Total number of features** | 4 features (columns) |
| **Base format of the file** | .csv |
| **Size of the data** | 1 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
```

## City

| | |
|---|---|
| **Total number of observations** | 20 observations (rows) |
| **Total number of files** | |
| **Total number of features** | 3 features (columns) |
| **Base format of the file** | .csv |
| **Size of the data** | 759 Bytes |

```
<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
```

## Cab_Data

| | |
|---|---|
| **Total number of observations** | 359392 observations |
| **Total number of files** | |
| **Total number of features** | 7 features (columns) |
| **Base format of the file** | .csv |
| **Size of the data** | 20.1 MB |

```
<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
```

**Proposed Approach:**

- For removing duplicate rows, I decided to use the pandas package “drop_duplicates” package with drops duplicates based on columns. This will allow every row to be tidy and unique so data analysis is efficient and accurate.
- My assumption about the data is that the data has been given to me without any care or previous analysis. This means that there could be NA values or duplicate rows to the data.
- My goal with the data is to get a better understanding of the data by performing EDA and data cleaning to provide an accurate report of the data.