Data Intake Report

Name: G2M Insight for Cab Investment Firm
Report date: 09/21/2023
Internship Batch: LISUM25
Version:<1.0>
Data intake by: Alison Jing March
Data intake reviewer: Alison Jing March
Data storage location: https://github.com/alisonjing/DataSets

Tabular data details: Cab_Data.csv

Total number of observations	359,392
Total number of files	1
Total number of features	7
Base format of the file	csv
Size of the data	20.2 MB

Tabular data details: City.csv

Total number of observations	20
Total number of files	1
Total number of features	3
Base format of the file	csv
Size of the data	759 Bytes

Tabular data details: Customer_ID.csv

Total number of observations	49,171
Total number of files	1
Total number of features	4
Base format of the file	csv
Size of the data	1.00 MB

Tabular data details: Transaction_ID.csv

Total number of observations	440,098
Total number of files	1
Total number of features	3
Base format of the file	csv
Size of the data	8.58 MB

Note: Replicate same table with file name if you have more than one file.

Proposed Approach:

- First import datasets one by one in the Jupyter notebook and check any empty(null) as well as duplicated values presented, then observe the data columns and check unique identifiers.
- Join transaction_id dataset with customer_id dataset by “Customer ID”. Join cab_data with city by “City”, then lastly join the two new dataframes together into the complete dataframe df.
- Convert data types: convert data types in the dataframe:
    - **df['Population'] = df['Population'].str.replace(',','')**
    - **df['Users'] = df['Users'].str.replace(',','')**
- Convert objects to integer datatypes
    - **df['Population'] = df['Population'].astype(int)**
    - **df['Users'] = df['Users'].astype(int)**
- Perform EDA and statistical analysis an generate summary table, plots and moving average trend.
- **Assumptions:**
    - We use profits as the main indicator for determining a successful cab investment.
    - The dataset is complete and pertaining to the stats between 2016-2018
    - Determine the future trend based on the plotted total profits and profit moving average across the duration of the data.