Data Intake Report

Name: G2M insight for Cab Investment firm
Report date: 9/15/2023
Internship Batch: LISUM25
Version: 1.0
Data intake by: Haseeb Javed
Data intake reviewer:
Data storage location: https://github.com/haseebjaved4652/DataSet

Tabular data details:

Accumulated Cost of Trip

Total number of observations	359392
Total number of files	1 file: “Cab_Data.csv”
Total number of features	3 columns: “Date of Travel”, “Company”, “Cost of Trip”
Base format of the file	ipynb
Size of the data	KB

Mileage by Company

Total number of observations	359392
Total number of files	1 file: “Cab_Data.csv”
Total number of features	3 columns: “Date of Travel”, “Company”, “KM Travelled”
Base format of the file	ipynb
Size of the data	60 KB

Gender Distribution

Total number of observations	359392 + 49171 + 440098 = 848661
Total number of files	3 files: “Cab_Data.csv”, “Customer_ID.csv”, “Transaction_ID.csv”
Total number of features	3 columns: “Gender”, “Payment Mode”, “Date of Travel”
Base format of the file	ipynb
Size of the data	53 KB

Number of Transaction per City

Total number of observations	359392
Total number of files	1 file: “Cab_Data.csv”
Total number of features	3 columns: “Date of Travel”, “City”, “Transaction ID”
Base format of the file	ipynb
Size of the data	86 KB

User-to-Population Ratio

Total number of observations	20 rows
Total number of files	1 file: “City.csv”
Total number of features	3 columns: “City”, “Population”, “Users”
Base format of the file	ipynb
Size of the data	116 KB

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

Deduplication Validation Approach: The code finds and removes duplicate entries.

Assumptions: The code handles missing or invalid data by converting them to NaN. It also avoids division by zero by replacing zero population values with one.

Note: Convert this doc in pdf and provide the link of pdf file in your dashboard.
Please do not forget to remove this section while converting the file into pdf.