Data Intake Report

Name: G2M insight for Cab Investment firm
Report date: 14-November-2024
Internship Batch: LISUM39
Version:1.0
Data intake by: Rohan Khatri
Data intake reviewer: Data Glacier
Data storage location:
https://github.com/rohankhatri7/DataGlacier-Internship/tree/main/Week%202
Tabular data details: Cab_Data

Total number of observations	359392
Total number of files	1
Total number of features	7
Base format of the file	csv
Size of the data	20.1 MB

Tabular data details: Transaction_ID

Total number of observations	440098
Total number of files	1
Total number of features	3
Base format of the file	csv
Size of the data	8.58 MB

Tabular data details: Customer_Data

Total number of observations	49171
Total number of files	1
Total number of features	4

Base format of the file	csv
Size of the data	1 MB

Tabular data details: City

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	4 KB

Proposed Approach:

Deduplication Validation Strategy:

- **Primary Key Verification:** Each dataset will be examined for unique identifiers to ensure data consistency:
    - **Cab_Data.csv:**
        - I will identify and remove any records with duplicate identifiers.
    - **City.csv:**
        - I will check for unique entries by using city names or other city-specific details.
    - **Customer_ID.csv:**
        - Customer ID will be used as the primary key to detect and eliminate any duplicate records in this dataset.
    - **Transaction_ID.csv:**
        - I will use Transaction ID as the primary key to find and address any duplicate transaction records.

Assumptions:

- The Unique Identifiers are consistent since each data set has their own use of unique identifiers that do not change over time.