Data Intake Report

Name: G2M insight for Cab Investment firm
Report date: 11/14/2024
Internship Batch: LISUM39
Version:1.0
Data intake by: Ethan Dy
Data intake reviewer: Data Glacier
Data storage location: https://github.com/ethan05d/DataGlacier-Internship/tree/main/Week%202

Tabular data details: Cab_Data.csv

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	20.10 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	4.00 KB

Tabular data details: Customer_ID.csv

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	1.00 MB

Tabular data details: Transaction_ID.csv

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

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

Proposed Approach:

- **Approach for Deduplication Validation:**
  - Primary Key Matching: For each dataset, I will check for unique identifiers:
    - Cab_Data.csv:
      - I will check for any records with matching identifiers.
    - City.csv:
      - Since this dataset has a small number of observations of 20 rows, I'll look for unique entries based on city names or other city-specific identifiers.
    - Customer_ID.csv:
      - I'll use Customer ID as the primary unique identifier to check for duplicate records in this dataset.
    - Transaction_ID.csv:
      - Using Transaction ID as the primary unique identifier, I'll identify any duplicate transaction entries.
- **Assumptions:**
  - Unique Identifiers are consistent, since each dataset has a consistent use of unique identifiers, such as Customer ID or Transaction ID, and that these identifiers do not change over time for the same entities.