Data Intake Report

Name: G2M insight for Cab Investment firm
Report date: 13/12/2023
Internship Batch: LISUM28
Version:1.0
Data intake by: Mustansar Hussain
Data intake reviewer: Data Glacier
Data storage location: https://github.com/Musransar/G2M.git

Tabular data details:

Total number of observations	359392
Total number of files	5
Total number of features	16
Base format of the file	CSV
Size of the data	43.9+ MB

Proposed Approach:

The approach for deduplication validation involves:

- Data Profiling and Exploration:

Initial exploration of the dataset to understand its structure, features, and distributions.
Reviewing unique identifiers or key columns to identify potential duplicates.

- Identifying Duplicate Records:

Checking for identical rows that might indicate duplication using key columns or a combination of columns.
Utilizing methods like duplicated() in Pandas to identify rows with the same values in selected columns.
Checking for similar or nearly identical records that might represent duplicates but contain slight variations.

- Handling Assumptions:

Key Columns: Assuming certain columns or a combination of columns represent unique identifiers (like transaction IDs, unique customer IDs).
Threshold for Duplicates: Determining a threshold for similarity to consider records as duplicates (e.g., considering records with 90% similarity as duplicates).
Data Cleaning: Addressing issues like inconsistent formatting, misspellings, or variations in categorical data that might falsely identify records as duplicates.

- Documentation and Reporting:

Documenting the process, including the criteria used for deduplication and any assumptions made.
Reporting the number of identified duplicates, the methodology used for deduplication, and the resulting impact on the dataset's size and quality.