Data Intake Report

Name: G2M insight for Cab Investment firm
Report date: 15/04/2025
Internship Batch: LISUM44
Version:
Data intake by: Ankita Roy
Data intake reviewer:
Data storage location: https://github.com/DataGlacier/DataSets

Tabular data details:

Total number of observations	359392
Total number of files	
Total number of features	13
Base format of the file	.csv
Size of the data	KB

Proposed Approach:

- Mention approach of dedup validation (identification)

  I started by extracting the raw datasets from the GitHub repository and tried to understand the business context first then accordingly combined multiple files into a single unified dataset by using the common columns. Once the dataset was merged, I tried to identify the key columns relevant to the problem.

  After understanding the structure and purpose of the data, I focused on deduplication validation. My approach was:
  - **Standardization:** I cleaned the data by removing inconsistencies such as leading/trailing spaces, case sensitivity, and irrelevant characters.
  - **Exact Duplicate Check:** I used .duplicated() across all columns and also on selected key features (e.g., names, addresses, IDs) to identify exact duplicates.
  - **Near Duplicate Identification:** I applied fuzzy matching techniques to catch near-duplicates-especially useful where types or small variations might exist.
  - **Manual Validation:** For borderline cases, I flagged the records for manual inspection or domain expert feedback to avoid incorrect deletions.

- Mention your assumptions (if you assume any other thing for data quality analysis)

- The combined dataset represents a complete and consistent snapshot from the source.

- Key columns (e.g., Name, Email, Phone) are consistent enough to detect duplicate entries.

- All required fields for deduplication were available in each record.
- The dataset does not include system-generated noise (timestamps, row IDs) as primary comparison fields.