Data Intake Report

Name: G2M insight for Cab Investment firm
Report date: 09 November 2023
Internship Batch: LISUM27
Version:1.0
Data intake by: Mufunwa Nemushungwa
Data intake reviewer:
Data storage location:

Cab data details:

Total number of observations	359392
Total number of files	
Total number of features	7
Base format of the file	.csv
Size of the data	21.2 MB

City data details:

Total number of observations	20
Total number of files	
Total number of features	3
Base format of the file	.csv
Size of the data	759 B

Customer ID data details:

Total number of observations	49171
Total number of files	
Total number of features	4
Base format of the file	.csv
Size of the data	1.1 MB

Transaction ID data details:

Total number of observations	440098
Total number of files	
Total number of features	3
Base format of the file	.csv
Size of the data	9 MB

Proposed Approach:

- The deduplication validation approach is methodical and comprehensive across all datasets. Firstly, within the Customer IDs dataset, the uniqueness of each ID is verified, aiming to detect any inconsistencies or duplications. Any duplicate customer IDs found will be merged to ensure a single, accurate record. Similarly, the approach extends to the Transactions IDs dataset, the City dataset concerning city names, and the Cab dataset for transaction IDs. This systematic process ensures the identification and resolution of duplicates across multiple key datasets, thereby enhancing data integrity and accuracy.

- The following assumptions were made:
  1. All the datasets have unique identifiers (e.g., customer ID, transaction ID and city names) that can be used to link and merge the datasets.
  2. Assuming consistency in data formats and structures across all datasets involved.
  3. Assuming accuracy in critical fields like IDs, city name and other unique identifiers used for deduplication.
  4. Assuming minimal missing or null values that might impact the deduplication process.