Data Intake Report

Name: G2M insight for Cab Investment firm
Report date: 08/April/2024
Internship Batch: LISUM32
Version:1.0
Data intake by: Sidorela Mema
Data intake reviewer: Data Glacier
Data storage location: GitHub Link

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	21.2 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

Tabular data details: Customer_ID

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	1.1 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	9 MB

Proposed Approach: Mention approach of dedup validation (identification)

In the process of data exploration and preprocessing, no duplicate records or missing values were identified in the dataset. Therefore, deduplication validation was not necessary as there were no duplicate records to remove. Similarly, since there were no missing values, no imputation or handling of missing data was required. The dataset was found to be clean in terms of duplication and missingness, which ensures the integrity and reliability of the data for further analysis.