Data Intake Report

Name: G2M insight for Cab Investment firm
Report date: 14/05/2024
Internship Batch: LISUM33
Version:1.0
Data intake By: Sarima Iyayi
Data intake reviewer: Sarima Iyayi
Data storage location:
https://drive.google.com/drive/folders/1X5LmZTKmxtpsyrQ9CcVS37Nv_dTr7VBY?usp=sharing

Tabular data details:

Cab Data

Total number of observations	359,392
Total number of files	1
Total number of features	7
Base format of the file	.csv
Size of the data	20.2 MB

Customer Data

Total number of observations	49,171
Total number of files	1
Total number of features	4
Base format of the file	.csv
Size of the data	1 MB

Transaction Data

Total number of observations	440,098
Total number of files	1
Total number of features	3
Base format of the file	.csv
Size of the data	8.6 MB

City Data

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	759 BYTES

Holiday Data

Total number of observations	57
Total number of files	1
Total number of features	2
Base format of the file	.csv
Size of the data	15 KB

Proposed Approach:

Deduplication Validation (Identification):

- Identify duplicate records based on unique transaction identifiers and customer IDs.
- Check for consistency in key fields across datasets (e.g., transaction IDs, customer IDs, dates).

Assumptions for Data Quality Analysis:

- All date fields are assumed to be in a consistent format.
- No missing critical identifiers (e.g., transaction ID, customer ID).
- All monetary values are in the same currency.
- Population and user data are assumed to be accurate as provided.