Data Intake Report

Name: G2M insight for Cab Investment firm
Report date: 21-April-2024
Internship Batch: LISUM32
Version: 1.0
Data intake by: Monisha Shree Senthil Nathan
Data intake reviewer:
Data storage location: https://github.com/BigDataEngineer09/Internship-Data-Glacier/tree/main/Week2

Tabular data details:

1. 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	19.2+ MB

2. 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	608.0+ bytes

3. 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	1.5+ MB

4. 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	10.1+ MB

Proposed Approach:

- Dedup validation (identification) approach
  - Utilize unique identifiers such as Transaction ID or Customer ID to identify duplicate records within the dataset.
  - Use pandas functions like **duplicated()** and **drop_duplicates()** to identify and remove duplicate records based on the identified key fields.
  - Review the dataset before and after deduplication to ensure that duplicate records have been successfully identified and removed.

- Mention your assumptions (if you assume any other thing for data quality analysis)
  - Assume that the data is consistent across all records and fields, including consistent formatting, units, and conventions.
  - Assume that all necessary fields are populated for each record, and missing values may indicate data quality issues or incomplete data collection processes.
  - Assume that the data accurately reflects real-world entities and events, including accurate measurements, calculations, and representations of information.