Data Intake Report

Name: G2M insight for Cab Investment firm
Report date: 14 November 2023
Internship Batch: LISUM27
Version:<1.0>
Data intake by: Venkata Sai Bharadwaj Velamakanni
Data intake reviewer:<intern who reviewed the report>
Data storage location: https://github.com/bharadwajvvs/G2M_insight-for-Cab-Investment-firm/tree/main

Tabular data details:

Cab_Data.csv

Total number of observations	359393
Total number of files	4
Total number of features	7
Base format of the file	.csv
Size of the data	21.2 mb

City.csv

Total number of observations	21
Total number of files	4
Total number of features	3
Base format of the file	.csv
Size of the data	759 bytes

Customer_ID.csv

Total number of observations	49172
Total number of files	4
Total number of features	4
Base format of the file	.csv
Size of the data	1.1 mb

Transction_ID.csv

Total number of observations	440099
Total number of files	4
Total number of features	3
Base format of the file	.csv
Size of the data	9 mb

Proposed Approach:

- Dedup Validation (Identification):
  To identify and remove duplicate records in the cab data, we can use the following approach:
  1. Identify unique identifiers: Identify unique identifiers for each cab trip, such as the transaction ID and Date of Travel.
  2. Remove duplicate records: Remove duplicate records from the data, keeping only one record for each unique trip.
- Assumptions:
  The unique identifiers are accurate and complete.
- Additional Assumptions for Data Quality Analysis:
  1. The data is clean and free of errors.
  2. The data is representative of the overall cab usage population.
  3. The data is consistent over time.