Data Intake Report

Name: G2M Insight for Cab Investment Firm
Report date: Nov. 2023
Internship Batch: LISUM25
Version:<1.0>
Data intake by: Seoyoung Kim
Data intake reviewer:
Data storage location: https://github.com/syoungk7/Exploratory-Data-Analysis.git

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	20.2 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	759 B

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 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	8.6 MB

Proposed Approach:

- Mention approach of dedup validation (identification)

    1. Check for duplicate customer IDs in Customer_ID.csv file and transaction IDs in Transaction_ID.csv file.
    2. Check for duplicate transaction IDs in Cab_Data.csv file.
    3. Check the uniqueness of city names or IDs in City.csv file.

- Mention your assumptions (if you assume any other thing for data quality analysis)

    1. Assume that the provided data is accurate.
    2. Assume that the datasets are complete and do not have missing essential information.
    3. Assume that the data is consistent across all the datasets, and there are no conflicting or contradictory entries.