Data Intake Report

Name: G2M insight for Cab Investment firm
Report date: May 14, 2023
Internship Batch: LISUM33
Version: 1.0
Data intake by: Paul Junver Soriano
Data intake reviewer:<intern who reviewed the report>
Data storage location: Github

Tabular data details:

Cab_Data.csv

Total number of observations	359,352
Total number of files	1
Total number of features	7
Base format of the file	.csv
Size of the data	19.2 MB

City.csv

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 bytes

Customer_ID.csv

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	1.5 MB

Transaction_ID.csv

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:

- Mention approach of dedup validation (identification)
    - Use pandas .duplicated() method
- Mention your assumptions (if you assume any other thing for data quality analysis)
    - Assume that date is accurate