Data Intake Report

Name: G2M insight for Cab Investment Firm
Report date: 20th June 2021
Internship Batch: LISUM01
Version:<1.0>
Data intake by: Reeha Khan
Data intake reviewer: -
Data storage location: <location URL eg: github, cloud>

Tabular data details:

Total number of observations	359392
Total number of files	4
Total number of features	14
Base format of the file	.csv
Size of the data	Total: 38.3 MB

Each file details:

Cab_Data:

Total number of observations	359392
Total number of files	1
Total number of features	14
Base format of the file	.csv
Size of the data	20,663 KB

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	1 KB

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	1,027 KB

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	8,788 KB

Proposed Approach:

- Mention approach of dedup validation (identification)

  Duplicates are found by using

  ```
  df.duplicated().sum().any()
  ```

  which returns False, meaning there are no duplicates in the merged file which contains all the data.

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

  Null data is found by using

  ```
  df.isnull().sum().any()
  ```

  which also returns False.