Data Intake Report

Name: G2M insight for Cab Investment firm
Report date: 18 June, 2021
Internship Batch: LISUM01
Version: 1.0
Data intake by: Mohammad Shafiqul Islam
Data intake reviewer: Data Glacier
Data storage location: https://github.com/msishuvo/G2M-insight-for-Cab-Investment-firm

Tabular data details:

Total number of observations	344953
Total number of files	4
Total number of features	21
Base format of the file	.csv
Size of the data	67.9 MB

Cab data details:

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

City data details:

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 B

Customer_ID data details:

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

Transaction_ID data details:

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:

- Duplicate data were removed by using the drop_duplicate function from the pandas dataframe.
- The date had to be converted from an integer value to its original format.
- There were extra one-month data that contradicts the given time period. So they were removed.
- Outliers may present in the data, but due to the lack of other information, it is not considered.
- Apart from the original 15 features, some new columns such as profit, year, month, day, age group, weather are created using simple calculation.