Data Intake Report

Name: G2M Insight for Cab Investment Firm
Report date: 14-10-2023
Internship Batch: LISUM26: 30 Sep – 30 Dec 2023
Version:<1.0>
Data intake by: David Anthony Adeleye
Data intake reviewer:<intern who reviewed the report>
Data storage location: https://github.com/DataGlacier/DataSets.git

Tabular data details:

Total number of observations	359392
Total number of files	4
Total number of features	7
Base format of the file	csv
Size of the data	21MB

Total number of observations	20
Total number of files	4
Total number of features	3
Base format of the file	csv
Size of the data	759B

Total number of observations	49171
Total number of files	4
Total number of features	4
Base format of the file	csv
Size of the data	1.1MB

Total number of observations	440098
Total number of files	4
Total number of features	3
Base format of the file	csv
Size of the data	9MB

Proposed Approach:

In my analysis of the cab data, I have implemented a deduplication validation approach to identify and handle duplicate records effectively. This process involves identifying and removing duplicate entries in the dataset to ensure data integrity and accuracy in my analysis.

I also carried out thorough univariate and bivariate analysis to obtain insights about the data. Feature engineering was done to better understand the reason for some observed trends.