Data Intake Report

Name: G2M insight for Cab Investment firm
Report date: 14th-Aug-23
Internship Batch: LISUM24
Version:1.0
Data intake by: Anusha Asim
Data intake reviewer: N/A
Data storage location: https://github.com/DataGlacier/DataSets/blob/main/README.md

Tabular data details:

File 1: Cab_Data.csv

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	20.10 MB

File 2: 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	4.00 KB

File 3: Customer_ID.csv

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	1.00 MB

File 4: 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	8.58 MB

Proposed Approach:

- Implement deduplication validation to ensure data accuracy using Pandas drop_duplicates() method.

- One-hot encode categorical variables before conducting statistical tests.

- Conduct exploratory data analysis (EDA) to understand data distribution and relationships.

- Perform statistical tests for hypothesis validation.

- Utilize data visualization techniques to present insights effectively.

- Assumptions:

    - The variables in the dataset follow a normal distribution for statistical tests.
    - The assumption of homoscedasticity holds, indicating equal error variances in statistical analyses.
    - Data preprocessing includes removal of missing values and outliers.
    - Dataset is complete and representative of the specified time period.