Data Intake Report

Name: G2M Insight for Cab Investment Firm
Report date: 13/10/2023
Internship Batch: LISUM26
Version: 1.0
Data intake by: Nazri
Data intake reviewer:<intern who reviewed the report>
Data storage location: DataGlacier/DataSets (github.com)

Tabular data details: Cab_Data

Total number of observations	359393
Total number of files	1
Total number of features	7
Base format of the file	.csv
Size of the data	20663KB

Tabular data details: Customer_ID

Total number of observations	49172
Total number of files	1
Total number of features	4
Base format of the file	.csv
Size of the data	1027KB

Tabular data details: Transaction_ID

Total number of observations	176108
Total number of files	1
Total number of features	3
Base format of the file	.csv
Size of the data	8788KB

Tabular data details: City

Total number of observations	21
Total number of files	1
Total number of features	3
Base format of the file	.csv
Size of the data	1 KB

Proposed Approach:

- **Here we merge the datasets 'cab_data' and 'transaction_data' with the 'Transaction ID' column and the customer_data dataset will be merged to it with the 'Customer ID' column. We will form a new dataset 'final_cab_data' with these merged dataset**
- **Here we are not merging city_data dataset to 'final_cab_data' dataset as this won't give any useful insights. Merging makes the values of 'population of the city' and 'no: of cab users in the city' repeated each time in the final_cab_data dataset against the city column**
- **We then perform Exploratory data Analysis in the final_cab_data dataset by understanding the data and its structure.**
- **Review the dataset's columns and their significance.**
- **Identifying the duplicate records and null values in the final_cab_data dataset**