

Data Intake Report

Name: G2M Insight for Cab Investment Firm

Report date: 13/06/2024

Internship Batch: LISUM34

Version: 1.0

Data intake by: Nazri

Data intake reviewer: Data Glacier

Data storage location: <https://github.com/NazriJasmal/G2M-Cab-Investment-Firm--Datasets.git>

Tabular data details: Cab_Data

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	20663KB

Tabular data details: 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	1027KB

Tabular data details: 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	8788KB

Tabular data details: 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

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. Then we merge the 'city_data' dataset to it with the 'city' column and hence form 'final_cab_data' dataset.
- 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