Data Intake Report

Name: G2M insight for Cab Investment firm
Report date: 14 July 2022
Internship Batch: LISUM11:30
Version:1.0
Data intake by: Loh Wan Teng
Data intake reviewer: Data Glacier
Data storage location: https://github.com/DataGlacier/DataSets

Tabular data details: Cab_Data

Total number of observations	359392 rows
Total number of files	1 file
Total number of features	7 columns
Base format of the file	.csv
Size of the data	19.2+ MB

Tabular data details: City

Total number of observations	20 rows
Total number of files	1 file
Total number of features	3 columns
Base format of the file	.csv
Size of the data	0.608 + KB

Tabular data details: Customer_ID

Total number of observations	49171 rows
Total number of files	1 file
Total number of features	4 columns
Base format of the file	.csv
Size of the data	1.5+ MB

Tabular data details: Transaction_ID

Total number of observations	440098 rows
Total number of files	1 file
Total number of features	3 columns
Base format of the file	.csv
Size of the data	10.1+ MB

Proposed Approach:

- Import libraries and datasets
- Validate data type of each column of each dataset (convert to suitable datatype if necessary)
- Check missing value (no null values and NA detected)
- Check duplicated rows (no duplicated rows detected)
- Find common columns of each dataset
- Merge datasets into one main dataset
- Drop meaningless or unnecessary columns
- Understanding basic information of main dataset
- Outliers Treatment
- Visualisation and Analysis
- Hypothesis Testing and Analysis
- Recommendations