Data Intake Report

Name: G2M insight for Cab Investment firm
Report date: 14th/02/2024
Internship Batch: LISUM30:30 Jan 24 – 30Apr 24
Version: 1.0
Data intake by: David Paul Nalumenya, nalumenyad@gmail.com
Data intake reviewer:<intern who reviewed the report>
Data storage location: https://github.com/DataGlacier/DataSets

Tabular data details:

File name: Cab_Data.csv

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	21.8MB

File name: City.csv

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	0.759MB

File name: Customer_ID.csv

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	1.0MB

File name: Transaction_ID.csv

Total number of observations	440099
Total number of files	1
Total number of features	3
Base format of the file	.csv
Size of the data	8.58MB

Note: Replicate same table with file name if you have more than one file.

Proposed Approach:
For data deduplication and validation, I will select the features that are key in identification of duplicates within my data, identify possibility of data duplication as well as duplicated columns using in-built python functions for detection of duplicates, harmonize the columns in to a single file. Visualization will be done using pandas to visually identify duplicates, correlation and data distribution. Missing values will be handled appropriately through imputation or deletion.

An assumption of data completeness is made initially however for the case of 5-10% missing values, selected rows will be removed as long as it does not affect representativeness of the data, this decision will be more likely made basing off domain knowledge and understanding of the data.

An assumption of data consistency across the different data files, the main data is not in a single data file (.csv), however caution is taken to determine completeness and consistency between the different data files that somewhat have the same features that can be concatenated together (this will be done by observation and merging using inbuilt python scripts).

An assumption of data uniqueness is important especially in columns or features that must not hold duplicates, presence of duplicates will be considered an error and therefore careful assessment of duplicates, missing values and data integrity will be done using visualizing, descriptive analytics.