Data Intake Report

Name: <G2M insight for Cab Investment firm>
Report date: <14/03/2024>
Internship Batch:<LISUM31>
Version:<1.0>
Data intake by:<Noella Mutuku>
Data intake reviewer:<intern who reviewed the report>
Data storage location: <GitHub>

Tabular data details, Master Data:

Total number of observations	359,392
Total number of files	4
Total number of features	18
Base format of the file	csv
Size of the data	0.05 GB

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

Proposed Approach:

Approach of dedup validation (identification):

- We identified unique identifiers such as Transaction ID and Customer ID within the dataset to detect duplicate records.
- Utilizing pandas functions like duplicated() and drop_duplicates(), we identified and removed duplicate rows based on these selected columns.
- Visual inspection of the duplicate records revealed no significant patterns or anomalies, confirming the accuracy of the deduplication process.

Assumptions

- We assumed that the data follows a consistent format across all columns, including date format, numerical format, and categorical format.
- The data entry process was assumed to be accurate, minimizing errors and inconsistencies in the dataset.
- The dataset was assumed to be complete, with no missing values that could impact the analysis.
- The source of the data was assumed to be reliable and trustworthy, reducing the risk of data inaccuracies or biases.