Data Intake Report

Name: **G2M insight for Cab Investment firm**
Report date: **09-13-2023**
Internship Batch: **LISUM25**
Version: 1.0
Data intake by: **Sharon Akoth Okech**
Data intake reviewer:
Data storage location: **GitHub**

Cab_Data details:

Total number of observations	359392
Total number of files	4
Total number of features	7
Base format of the file	.csv.
Size of the data	19.2+ MB

Transaction_ID details

Total number of observations	440098
Total number of files	4
Total number of features	3
Base format of the file	.csv.
Size of the data	10.1+ MB

City Details

Total number of observations	20
Total number of files	4
Total number of features	3
Base format of the file	.csv.
Size of the data	608.0+ bytes

Customer_ID details

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	1.5+ MB

Proposed Approach:

1. Deduplication: Examined the data for duplicate records. This was accomplished by finding a unique identifier, such as a customer ID or transaction ID, and determining whether any values were duplicated. If duplicates are discovered, examine further to see whether they are genuine or the result of data input errors.

2. Assumptions: After checking for missing values and null values, It’s assumed that the data value in every field is in the corresponding data type of the field.

3. Missing Values: Determined whether there are any missing values in the data. Missing values could imply incorrect data entry or incomplete records. Determine whether the missing values are important and whether or not they can be handled or imputed. A substantial number of missing values in specific fields may indicate data quality issues.

4. Outliers: Examined the data for any outliers. Outliers are extraordinary results that differ dramatically from the rest of the data. Outliers may suggest data entry problems or data quality issues. Investigate outliers to determine whether they are valid data points or should be reported for additional investigation.

5. Data Inconsistencies: Examined the data for any inconsistencies. Inconsistencies in formatting, unit types, or name practices. Inconsistencies can complicate data processing and lead to errors or misleading conclusions. To maintain consistency across all records, standardize the data.

6. Data Integrity: Verified the data's integrity. This entails determining if the data adheres to defined data rules, such as data types, range validation, or referential integrity. Ensuring data integrity contributes to data accuracy and reduces errors in future data processing.
7. Documentation: Keep a record of any data quality issues or concerns that arise during the analysis. This will aid in the tracking and resolution of concerns, as well as serve as a reference for future analysis.