Data Intake Report

Name: G2M Insight for Cab Investment Firm
Report date: 13/09/2024
Internship Batch: LISUM37
Version: 1.0
Data intake by: Bristy
Data intake reviewer:
Data storage location:

Tabular data details:

Cab Data Dataset:

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	20663 KB

City Dataset:

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

Customer ID Dataset:

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	1027 KB

Transaction ID Dataset:

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	8788 KB

Proposed Approach:

1. Deduplication Validation Approach: First, I used the `.duplicated()` method to identify duplicate rows in each dataset, checking for any entries that may have been repeated.
2. I ensured the primary keys (`Transaction_ID`, `Customer_ID`, etc.) do not contain duplicates. I also merged the datasets on these keys and validate consistency across different datasets.
3. After identifying duplicates, I removed them using `drop_duplicates()`and further investigated to understand if they reflect valid repeated transactions.
4. For deeper validation, I looked for other types of redundancies, like similar entries that vary in only non-significant ways (e.g., name variations due to case sensitivity).

Assumptions for Data Quality Analysis:

1. I assumed that all dates are consistently formatted and use standard date-time formats across all datasets. But it wasn't, so I cleaned the date formats to ensure consistency.
2. I assumed there are no missing critical values in key fields such as `Transaction_ID`, `Customer_ID`, and other unique identifiers. If there are missing values, I removed the affected rows, depending on their significance.
3. The datasets have already been pre-validated for structural consistency (correct columns, data types, etc.), but I double-checked for any unexpected anomalies.
4. I assumed the `Population` and `Users` columns in the `City` dataset reflect accurate data, and that there are no mismatches between the city names across the datasets.
5. I also assumed that the financial data (`Price Charged`, `Cost of Trip`) is accurate and consistent across the merged datasets. If discrepancies arise, further investigation or external validation may be necessary.