Data Intake Report

Name: <Cab Data Industry>

Report date: <20-09-24>

Internship Batch:<Enter your batch code from Canvas course>

Version:<1.0>

Data intake by:<Usman Arif>

Data intake reviewer:<intern who reviewed the report>

Data storage location: <location URL eg: github, cloud>

**Tabular data details:**

Cab\_Data.csv

|  |  |
| --- | --- |
| **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** | 20.18 MB |

Transaction\_ID.csv

|  |  |
| --- | --- |
| **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** | 8.58 MB |

Customer\_ID.csv

|  |  |
| --- | --- |
| **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** | 1.0 MB |

|  |  |
| --- | --- |
| **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** | 759.0 B |

### **Proposed Approach: Deduplication Validation (Identification)**

For deduplication validation, the approach involves systematically checking for duplicate entries within each dataset using the .duplicated().sum() method provided by pandas. This method identifies whether a row has an identical copy elsewhere in the dataset.

1. **Identification**:
   * For each dataset (Cab Data, Customer Data, Transaction Data, City Data), duplicates were identified by checking for rows that were exact replicas across all columns.
   * Any detected duplicates were flagged, allowing us to decide if they should be dropped or further investigated based on the context of the data.
2. **Handling**:
   * In the event of duplicates, we would typically decide on whether to remove them or retain them based on business logic. For example, multiple identical cab trips might suggest data entry errors, whereas repeated customer entries could indicate multiple transactions by the same individual.

### **Assumptions:**

For data quality analysis, several assumptions were made to ensure robust validation and analysis:

1. **Assumption of Data Completeness**:
   * It was assumed that all relevant columns needed for the analysis (e.g., trip details, customer information, transaction data) are present and populated. Missing data, if any, would significantly affect the quality of the analysis and needed to be addressed through imputation or exclusion.
2. **Uniform Measurement Units**:
   * It was assumed that all measurements (e.g., KM Travelled, Price Charged, Cost of Trip) are recorded in consistent units across the dataset. This is critical to avoid misinterpretation when aggregating or comparing values.
3. **Correct Data Types**:
   * Data type corrections were assumed necessary where required, particularly for numerical fields like Population and Users in the city dataset. Without this, the analysis could encounter errors or incorrect calculations.
4. **Data Integrity**:
   * For transactions, it was assumed that there was integrity between linked datasets (e.g., Cab Data and Transaction Data), meaning that each transaction ID found in the cab data should have a corresponding entry in the transaction data, ensuring the relationships between datasets are valid.