Data Intake Report

Name: G2M insight for Cab Investment firm

Report date: 14/03/2025

Internship Batch: LISUM43 28 Feb 2025 – 31 Feb 2025

Version:1.0

Data intake by: Mohammadistiyak Shaikh

Data intake reviewer: Not Applicable

Data storage location: https://github.com/iamistiyak/Data\_Glaciers\_Internship

**Tabular data details:**

|  |  |
| --- | --- |
| **File name** | Cab\_data |
| **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** | 20.1 MB |

|  |  |
| --- | --- |
| **File name** | transaction\_ID |
| **Total number of observations** | 440098 |
| **Total number of features** | 3 |
| **Base format of the file** | .csv |
| **Size of the data** | 8.58 MB |

|  |  |
| --- | --- |
| **File name** | City\_data |
| **Total number of observations** | 20 |
| **Total number of features** | 3 |
| **Base format of the file** | .csv |
| **Size of the data** | 759 bytes |

|  |  |
| --- | --- |
| **File name** | Cutomer\_ID |
| **Total number of observations** | 49171 |
| **Total number of features** | 4 |
| **Base format of the file** | .csv |
| **Size of the data** | 1 MB |

**Proposed Approach:**

**Deduplication Validation (Identification Approach)**

To ensure data integrity, we will follow these steps for **duplicate identification and validation**:

1. **Check for Exact Duplicates:**
   * Identify and remove any rows that are fully identical across all columns within each dataset.
   * Use methods such as df.duplicated() in Python (Pandas) to find and remove duplicates.
2. **Check for Duplicate Primary Keys:**
   * Verify if unique identifiers (e.g., Customer\_ID, Transaction\_ID) appear multiple times when they are expected to be unique.
   * If duplicates exist, further investigate if they represent data entry errors or valid repeated transactions.
3. **Cross-File Duplication Check:**
   * Ensure no duplicate records appear across multiple datasets.
   * Validate relationships between different files (e.g., Transaction\_ID linking transactions to customers) to check for inconsistencies.
4. **Handling Partial Duplicates:**
   * Identify records with the same key attributes but minor differences in fields (e.g., case sensitivity, extra spaces).
   * Apply fuzzy matching or similarity checks to detect near-duplicates.

**Assumptions for Data Quality Analysis**

We assume the following conditions for our analysis:

1. **Data Completeness Assumption:**
   * Each dataset is assumed to be complete, meaning all necessary columns for the analysis exist without major missing values.
   * If critical fields (e.g., Customer\_ID, Transaction\_ID) have missing values, they will be treated as errors or outliers.
2. **Consistency Assumption:**
   * We assume that all transactions in Cab\_Data.csv correspond to valid Transaction\_IDs in Transaction\_ID.csv and valid Customer\_IDs in Customer\_ID.csv.
   * If there are missing mappings, we will analyze whether they result from data loss or valid missing relationships.
3. **Data Format Assumption:**
   * All categorical fields (e.g., city names, payment methods) follow a consistent format and spelling.
   * All date fields are correctly formatted (e.g., YYYY-MM-DD) and represent valid dates.
4. **Time Consistency Assumption:**
   * The transaction timestamps are correctly recorded and within the dataset's date range (31/01/2016 – 31/12/2018).
   * If any transaction falls outside this period, it will be treated as a potential error.
5. **No Assumed Bias in Data Collection:**
   * We assume that the provided datasets fairly represent cab usage patterns and customer demographics, without systematic biases in data collection.