Data Intake Report

Name: G2M Insight for Cab Investmnt firm

Report date: 09/14/2023

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

Version:1.0

Data intake by: Sali El-loh

Data intake reviewer: N/A

Data storage location: <https://github.com/DataGlacier/VC>

**Tabular data details:**

Cab\_Data.csv:

|  |  |
| --- | --- |
| **Total number of observations** | 359393 |
| **Total number of files** | 5 |
| **Total number of features** | 8 |
| **Base format of the file** | .csv |
| **Size of the data** | 24.8 MB |

Transaction\_ID.csv

|  |  |
| --- | --- |
| **Total number of observations** | 440099 |
| **Total number of files** | 5 |
| **Total number of features** | 3 |
| **Base format of the file** | .csv |
| **Size of the data** | 9 MB |

Customer\_ID.csv

|  |  |
| --- | --- |
| **Total number of observations** | 39761 |
| **Total number of files** | 5 |
| **Total number of features** | 4 |
| **Base format of the file** | .csv |
| **Size of the data** | 1.05 MB |

City.csv

|  |  |
| --- | --- |
| **Total number of observations** | 21 |
| **Total number of files** | 5 |
| **Total number of features** | 3 |
| **Base format of the file** | .csv |
| **Size of the data** | 759 B |

Asssignment\_1.pynb:

|  |  |
| --- | --- |
| **Total number of observations** | N/A |
| **Total number of files** | 5 |
| **Total number of features** | N/A |
| **Base format of the file** | .pynb |
| **Size of the data** | 83 Kb |

**Proposed Approach:**

* your Mention approach of dedup validation (identification)
* Mention assumptions (if you assume any other thing for data quality analysis)

**Proposed Approach (deduplication and assumptions):**

Deduplication Approach:

1. Identified the key fields for deduplication, such as transaction or customer identifiers.
2. Conducted a systematic comparison of these fields within the datasets.
3. Flagged or removed duplicate entries based on the unique identifiers.

Assumptions for Data Quality Analysis:

1. The provided data is representative of the actual market conditions during the specified time period.
2. Any missing data points can be reasonably imputed or omitted, depending on the analysis context.
3. Duplicates are considered as erroneous entries and should be removed or flagged for further investigation.
4. External data sources (e.g., holiday data) used for augmentation are accurate and relevant to the analysis.