Data Intake Report

Name: G2M insight for Cab Investment firm

Report date: 27/12/23

Internship Batch: LISUM28

Version:<1.0>

Data intake by: Yusuf Shamsi

Data intake reviewer:<intern who reviewed the report>

Data storage location: <https://github.com/Yushamsi/Data-Glacier-Week-2---3>

**Tabular data details:**

**Cab\_Data:**

|  |  |
| --- | --- |
| **Total number of observations** | 359392 |
| **Total number of files** | 4 Total |
| **Total number of features** | 7 |
| **Base format of the file** | .csv |
| **Size of the data** | 21.2 MB |

**City:**

|  |  |
| --- | --- |
| **Total number of observations** | 20 |
| **Total number of files** | 4 Total |
| **Total number of features** | 3 |
| **Base format of the file** | .csv |
| **Size of the data** | 759 Bytes |

**Customer\_ID:**

|  |  |
| --- | --- |
| **Total number of observations** | 49171 |
| **Total number of files** | 4 Total |
| **Total number of features** | 4 |
| **Base format of the file** | .csv |
| **Size of the data** | 1.1 MB |

**Transaction\_ID:**

|  |  |
| --- | --- |
| **Total number of observations** | 440098 |
| **Total number of files** | 4 Total |
| **Total number of features** | 3 |
| **Base format of the file** | .csv |
| **Size of the data** | 9 MB |

**Note: Replicate same table with file name if you have more than one file.**

**Proposed Approach:**

**Approach for Deduplication Validation**

1. **Identifying Unique Identifiers:**
   * **For each dataset, I identified columns that can act as unique identifiers. For example, Transaction ID in Cab\_Data.csv and Transaction\_ID.csv, and Customer ID in Customer\_ID.csv.**
2. **Checking for Duplicate Records:**
   * **I used these unique identifiers to check for any duplicate records within each dataset. This involved using pandas functions like duplicated() to flag and count any repeated entries.**

**Assumptions for Data Quality Analysis**

1. **Data Completeness**:
   * I assumed that the datasets provided are complete representations of the transactions, customers, and cities involved. No external data sources were used for the basic structural integrity of the data.
   * I assumed that the "Date of Travel" column in the Cab\_Dta.csv was originally in the Excel serial date format.

**Note: Convert this doc in pdf and provide the link of pdf file in your dashboard.**

**Please do not forget to remove this section while converting the file into pdf.**