Data Intake Report

Name: G2M insight for Cab Investment firm

Report date: Dec 14 2022

Internship Batch: LISUM16

Version:1.0

Data intake by: Yena Lee

Data intake reviewer:

Data storage location: http://github.com/yl4315/G2M\_insight\_for\_Cab\_Investment\_firm

**Tabular data details: 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** | 19.2+ MB |

**Tabular data details: Customer\_ID**

|  |  |
| --- | --- |
| **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** | 1.5+ MB |

**Tabular data details: Transaction\_ID**

|  |  |
| --- | --- |
| **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** | 10.1+ MB |

**Tabular data details: City**

|  |  |
| --- | --- |
| **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** | 608.0+ bytes |

**Proposed Approach:**

* No missing value and duplicate record are found.
* In Cab\_Data , range of “Date of Travel’ column exceeds 30 days than the period from 31/01/2016 to 31/12/2018. The range of date should be either '31/01/2016'-'31/01/2019' or '01/01/2016'-'31/12/2018'. Therefore, the assumption is violated, so changed the period of date to '01/01/2016'-'31/12/2018'.
* In Cab\_Data, all transaction ID is unique.
* In Customer\_ID, all customer ID is unique.
* In Transaction\_ID, all transaction ID is unique, and each customer is linked to one or more transaction IDs.