Data Intake Report

Name: G2M insight for Cab Investment Firm

Report date: 13-09-2023

Internship Batch: LISUM25

Version: 0.1

Data intake by: Ishwarya Rajaiya

Data intake reviewer:

Data storage location:

**Tabular data details:**

|  |  |
| --- | --- |
| **File Name** | Cab Data |
| **Total number of observations** | 359392 |
| **Total number of files** |  |
| **Total number of features** | 7 |
| **Base format of the file** | .csv |
| **Size of the data** | 22.96 MB |

|  |  |
| --- | --- |
| **File Name** | City Data |
| **Total number of observations** | 20 |
| **Total number of files** |  |
| **Total number of features** | 3 |
| **Base format of the file** | .csv |
| **Size of the data** | 759 B |

|  |  |
| --- | --- |
| **File Name** | Customer ID |
| **Total number of observations** | 49171 |
| **Total number of files** |  |
| **Total number of features** | 4 |
| **Base format of the file** | .csv |
| **Size of the data** | 1.05 MB |

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

**Proposed Approach:**

**Column data types:**

|  |  |
| --- | --- |
| **CITY DATA** | |
| **Columns** | **Data type** |
| City | Object |
| Population | Object |
| Users | object |

|  |  |
| --- | --- |
| **CUSTOMER ID** | |
| **Columns** | **Data type** |
| Customer ID | int64 |
| Age | Object |
| Gender | int64 |
| Income (USD/Month) | int64 |

|  |  |
| --- | --- |
| **CAB DATA** | |
| **Columns** | **Data type** |
| Transaction ID | int64 |
| Date of Travel | Object |
| Company | object |
| City | Object |
| KM Travelled | float64 |
| Price Charged | float64 |
| Cost of Trip | float64 |

|  |  |
| --- | --- |
| **TRANSACTION ID** | |
| **Columns** | **Data type** |
| Transaction ID | int64 |
| Customer ID | int64 |
| Payment mode | Object |

**DATA ANALYSIS:**

* The dataset has no duplicates, so deduplication approach is not necessary.
* Also, there are no missing values in the dataset indicating complete data.
* Gender data shows the gender of customers, with 'Male' being predominant.
* The data appears to be of good quality, with no apparent errors or inconsistencies.