Data Intake Report

Name: G2M Insight for Cab Investment firm

Report date: 30-12-2021

Internship Batch: LISUM05

Version: 1.0

Data intake by: R.Praneetha

Data intake reviewer: Data Glacier

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

**Tabular data details:**

**Cab\_Data.csv**

|  |  |
| --- | --- |
| **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** | 20.66MB |

**City.csv**

|  |  |
| --- | --- |
| **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** | 1KB |

**Transaction\_ID.csv**

|  |  |
| --- | --- |
| **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** | 8.78MB |

**Customer\_ID.csv**

|  |  |
| --- | --- |
| **Total number of observations** | 49171 |
| **Total number of files** | 1 |
| **Total number of features** | 7 |
| **Base format of the file** | .csv |
| **Size of the data** | 1.027MB |

**Proposed Approach:**

* Load all datasets and review them.
* Identify relationships between all datasets and merge accordingly.
* Find duplicates and remove them before applying any analysis.

**Note:** There are no duplicates found from the above-mentioned dataset.

* According to description of dataset it is mentioned that time period of data is from 31/01/2016 to 31/12/2018. Based on this time period convert numeric values (represents days) into actual dates by subtracting with actual numbers mentioned in the dataset.
* Two cab companies are mentioned mainly yellow cab and pink cab.
* In the program analysis is done to find profit in different cases. If profit is negative it is considered as loss.
* Note that, in day\_week, 0 is Monday and 6 is Sunday.
* Each User is accessed using unique User ID.
* All analysis part including maximum number of customers at particular time period is calculated and plotted in the “.ipynb” program file.