Data Intake Report

Name: G2M Insight for Cab Investment Firm

Report date: 24/06/2021

Internship Batch: LISUM01

Version:1.0

Data intake by: Manoj Kumar Thangaraj

Data intake reviewer:

Data storage location: https://github.com/manojthangaraj92/G2M-CabInvestmentFirm.git

**Tabular data details:**

**Table Name**: Cab\_data

|  |  |
| --- | --- |
| **Total number of observations** | 359392 |
| **Total number of files** | 1 |
| **Total number of features** | 7 |
| **Base format of the file** | Csv file |
| **Size of the data** | 20.63MB |

**Table Name: City**

|  |  |
| --- | --- |
| **Total number of observations** | 20 |
| **Total number of files** | 1 |
| **Total number of features** | 3 |
| **Base format of the file** | Csv file |
| **Size of the data** | 1KB |

**Table Name: Customer\_ID**

|  |  |
| --- | --- |
| **Total number of observations** | 49171 |
| **Total number of files** | 1 |
| **Total number of features** | 4 |
| **Base format of the file** | Csv file |
| **Size of the data** | 1.02MB |

**Table Name: Transaction\_ID**

|  |  |
| --- | --- |
| **Total number of observations** | 440098 |
| **Total number of files** | 1 |
| **Total number of features** | 3 |
| **Base format of the file** | Csv file |
| **Size of the data** | 8.7MB |

**Proposed Approach:**

The four individual datasets are merged using their unique keys in each column. Then the data pre-processing has been done to check for duplication of rows and null values etc.,

Descriptive statistics has been carried out on the dataset to develop useful insights.

Time series analysis has been carried out to develop actionable insights from the dataset.

Various hypothesis generated and tested against this dataset to check the credibility.

**Assumptions:**

1. The start date is considered to 02/01/2016(DD/MM/YYYY), as per datetime code provided. It contradicts with the start date and end date given in the document by one month. Therefore, I considered the end date by document as true and start date is brought forward by 1 month.
2. The cost of the trip is assumed as normal or estimated cost of trip as in base charge.
3. The Price charged for the trip is assumed as actual coast incurred.
4. The difference between the price charged and the cost of the trip is assumed to be profit per trip.
5. Profit per KM is calculated by using profit and KM travelled as there is no information about the time included.