# Data Intake Report

## **Name:** G2M insight for Cab Investment firm

Report date: 12th of April 2023
Internship Batch: LISUM20
Version: 1.0
Data intake by: Nilsu Bozan
Data intake reviewer:

**Tabular data details:**

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

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

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

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

**Proposed Approach:**

**Approach of dedup validation (identification):** Identify the primary key or unique identifier for each dataset, check for duplicate records in each dataset using the primary key or unique identifier, check for missing values.

**My assumptions:**

Data is structured with consistent data types for each column.
Primary keys or unique identifiers are accurate and unique.
There are no missing or null values in critical columns.
Data has not been artificially manipulated or altered.