# Data Intake Report

Name: G2M Insight for Cab Investment Firm
Report date: 14th January, 2025
Internship Batch: LISUM41: 30 December, 2024 – 31 March, 2025
Version: 1.0
Data intake by: Mahnoor Farhat
Data intake reviewer: N/A
Data storage location: https://github.com/mahnoor-farhat/cab-firm-analysis

**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** | 20.1 MB |

**Tabular data details: City Data**

| | |
|---|---|
| **Total number of observations** | 19 |
| **Total number of files** | 1 |
| **Total number of features** | 3 |
| **Base format of the file** | .csv |
| **Size of the data** | 759 B |

**Tabular data details: Customer Data**

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

**Tabular data details: Transaction Data**

| | |
|---|---|
| **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** | 8.58 MB |

**Note: Replicate same table with file name if you have more than one file.**

**Proposed Approach:**

- Each file/dataframe was checked for duplicates but no duplicates were found, after which all of the data was merged as master dataframe.
- It was assumed that each table with column ID is unique and acts as a primary identifier across all datasets.
- Critical columns such as Transaction ID, Date of Travel, Customer ID, Company, and City were assumed to have no missing values.
- It was assumed that numerical fields, such as Price Charged, Cost of Trip, and KM Travelled, contain only non-negative values.
- Categorical variables such as Company, City, Payment_Mode, and Gender were assumed to have consistent and valid entries.