# Data Intake Report

Name: G2M insight for Cab Investment firm
Report date: 20/03/2023
Internship Batch:<Enter your batch code from Canvas course>
Version:<1.0>
Data intake by: Ian Kihara Wangui
Data intake reviewer: Ian Kihara Wangui
Data storage location: https://github.com/Iandavidk/DataGlacier/tree/main/Week-2-G2M-insight-for_cab-Investment-firm

**Tabular data details: Cab_Data.csv**

| | |
|---|---|
| **Total number of observations** | 359392 |
| **Total number of files** | 4 |
| **Total number of features** | 7 |
| **Base format of the file** | .csv |
| **Size of the data** | 20.2 MB |

**Tabular data details: City.csv**

| | |
|---|---|
| **Total number of observations** | 20 |
| **Total number of files** | 4 |
| **Total number of features** | 3 |
| **Base format of the file** | .csv |
| **Size of the data** | 759 Bytes |

**Tabular data details: Customer_ID.csv**

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

**Tabular data details: Transaction_ID.csv**

| | |
|---|---|
| **Total number of observations** | 440098 |
| **Total number of files** | 4 |
| **Total number of features** | 3 |
| **Base format of the file** | .csv |
| **Size of the data** | 8.58 MB |

**Proposed Approach:**

- There were no missing values in the data.
- There were no duplicated values in the data.
- I detected the presence of outliers using the z-scores method in the Price Charged column which I will remove later.
- To create master data I will perform a cross merge as follows: first will be cab and city data whose output will be merged with the transaction data and then the output of that will be merged with the customer data.
- The ‘Date of Travel’ field will be transformed inorder to enrich the analysis of the data.