# Data Intake Report

Name: G2M insight for Cab Investment firm
Report date: 02/28/25
Internship Batch: LISUM41
Version:<1.0>
Data intake by: Sophonie Sidrac
Data intake reviewer: Sophonie Sidrac
Data storage location: https://github.com/1Sophani/DataGlacier-Internship/tree/main/Week%202

**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** | 20.2 MB |

**City:**

| | |
|---|---|
| **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** | 759 B |

**Customer_ID:**

| | |
|---|---|
| **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 |

**Transaction_ID:**

| | |
|---|---|
| **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** | 8.6 MB |

**EDA:**

| | |
|---|---|
| **Total number of observations** | 359392 |
| **Total number of files** | 1 |
| **Total number of features** | 15 |

| | |
|---|---|
| **Base format of the file** | .ipynb |
| **Size of the data** | 1 MB |

**Proposed Approach:**

1. **Data Cleaning:**
   - **Duplicate Handling:** The data will be sorted based on transaction ID, or any unique identifier present in the dataset. Duplicates will be identified based on these unique identifiers and removed.
   - **Handling Missing Data:** Rows with missing data will be flagged and analyzed. on the nature and quantity of the missing data, imputation or removal methods might be applied.
   - **Data Transformation:** Convert categorical variables into a format suitable for analysis, possibly using encoding techniques. Normalize numerical data if required.
2. **Exploratory Data Analysis (EDA):**
   - **Demographic Analysis:** Identify which demographic group utilizes taxis the most. This will involve grouping the data by demographic and then calculating the sum or count of taxi uses.
   - **Company Popularity:** Assess which taxi company is more popular in each city. This will involve grouping by city and company and then comparing the count of uses.
3. **Assumptions:**
   - **Missing Data as Invalid:** Any missing data is considered as an invalid entry unless proven otherwise.
   - **Duplicate Transactions:** Any transaction ID appearing more than once is considered a duplicate.
   - **Data Format:** Non-numeric values in numeric fields are considered errors and will be treated or removed.

| |
|---|
| **Submitted by: Sophonie Sidrac** |
| **Submitted to: Data Glacier** |
| **Submission Date: 03/03/25** |