Data Intake Report

# Name: <G2M insight for Cab Investment firm>

Report date: <7/2/2022>

Internship Batch:<LISUM10: 30>

Version:<1.0>

Data intake by:<Yawo Amen Eklou>

Data intake reviewer:<intern who reviewed the report>

Data storage location: <https://github.com/yeklou/xyz.git>

**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** | 21.2 MB |

**Tabular data details: 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 |

**Tabular data details: 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** | 1.05 MB |

**Tabular data details: 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** | 9 MB |

**Tabular data details: data\_cleaned**

|  |  |
| --- | --- |
| **Total number of observations** | < 268741> |
| **Total number of files** | <1> |
| **Total number of features** | <10> |
| **Base format of the file** | <.csv> |
| **Size of the data** | <…> |

**Proposed Approach:**

**Unique values**

**Graphical user interface, text, application

Description automatically generated**

* Mention approach of dedupe validation (identification)

The “drop\_duplicate” function is used to remove any duplicate value from the dataset. This allows the accuracy of the database and improve the view of the data by the customer

**Table

Description automatically generated with medium confidence**

* Mention your assumptions (if you assume any other thing for data quality analysis)

Data wrangling was performed by merging ‘Cab Data', 'Customer\_ID', 'Transaction\_ID' and appended to City file to create ‘data’ dataset.

Annoying variables such as: 'Customer ID', 'Transaction ID' have been then removed from the ‘data’ dataset because they have no correlation with other variables.

NA and outliers’ detection analyses have been performed to prevent a use of biased records in the data.

In this process outliers have been identified and removed for every single features are columns in the data to create a final dataset named ‘data\_cleaned’ which is stored as a ‘csv’ file. This solves the issue of any eventual bias in the data that can cause bias in the data analysis.

**Graphical user interface

Description automatically generated with medium confidence**