

# Data Intake Report

Name: G2M insight for Cab Investment firm

Report date: 13 July 2022

Internship Batch: LISUM11

Version: 1.1

Data intake by: Disha Lamba

Data intake reviewer:

Data storage location: <https://github.com/dldisha/G2M-insight>

## Tabular data details: Cab\_Data

<b>Total number of observations</b>	359392
<b>Total number of files</b>	1
<b>Total number of features</b>	7
<b>Base format of the file</b>	.csv
<b>Size of the data</b>	20.1 MB

## Tabular data details: City

<b>Total number of observations</b>	20
<b>Total number of files</b>	1
<b>Total number of features</b>	3
<b>Base format of the file</b>	.csv
<b>Size of the data</b>	759 bytes

## Tabular data details: Customer\_ID

<b>Total number of observations</b>	49171
<b>Total number of files</b>	1
<b>Total number of features</b>	4
<b>Base format of the file</b>	.csv
<b>Size of the data</b>	1 MB

## Tabular data details: Transaction\_ID

<b>Total number of observations</b>	440098
<b>Total number of files</b>	1
<b>Total number of features</b>	3
<b>Base format of the file</b>	.csv
<b>Size of the data</b>	8.58 MB

**Proposed Approach:**

My first step will be to check for the source of the dataset like how it was collected and if it is authentic.

**a) Data source:**

Check if data is representative of our target situation. If not then look for other data sources

**b) Null Values:**

To deal with null and default values, I would replace them with mean/median/mode as appropriate but in the 4 datasets there are **no null values**.

**c) Duplicate Values:**

Prevent and fix duplicate data by using matching algorithms or some software and deleting the duplicate rows. But there are **no duplicate values** in the 4 datasets.

**d) Join tables:**

To create master data, I would join the 4 tables. First, I would join *Cab Data* with *Customer\_ID* and *Transaction\_ID* w.r.t 'transaction\_ID' and 'Customer\_ID' columns respectively. Next, I would join this table with *City* table w.r.t 'City' column respectively.

**e) Data Transformation:**

For data transformation, I would transform the column 'Date of travel' from *Cab Data* to a more understandable format for better analysis. I would further divide it into Month and Year column and drop Date of Travel column.

If required I would also normalize my datasets.

**f) Outliner detection:**

Outliers will be removed from numerical fields so that they don't affect the analysis. But some outliers because of holidays or seasonal breaks.

**g) Check for Data Leakage:**

If there's data leakage then performing data preparation within cross-validation folds or holding back a validated dataset.

My second step will be to perform data visualizations and understand relationships and distributions among the data which would help me with feature selection for the target situation.

In the end, 100% accuracy and completeness don't exist. My objective will be to have data quality and analysis to an acceptable threshold.