

## Data Intake Report

Name: G2M Strategy Case Study

Report date: 12/28/2021

Internship Batch: LISUM05

Version:1.0

Data intake by: Anup Pudasaini

Data intake reviewer: None.

Data storage location:

"https://raw.githubusercontent.com/DataGlacier/DataSets/main/Cab\_Data.csv"

"https://raw.githubusercontent.com/DataGlacier/DataSets/main/Customer\_ID.csv"

"https://raw.githubusercontent.com/DataGlacier/DataSets/main/Transaction\_ID.csv"

"https://raw.githubusercontent.com/DataGlacier/DataSets/main/City.csv"

"https://raw.githubusercontent.com/seeanup/Week2\_3/main/Temp\_2016-2018.csv"

(Note: Data from above URLs are combined to come up with number of observations/features)

### Tabular data details:

<b>Total number of observations</b>	359392
<b>Total number of files</b>	5
<b>Total number of features</b>	15
<b>Base format of the file</b>	.csv
<b>Size of the data</b>	MB

**Note:** Analysis were performed in Google Colab. Files from URLs are imported directly to Colab in Raw format!

### Proposed Approach:

- Combined different datasets into one. Checked any missing/null values.
- Converted some features into a suitable datatype for analysis purposes and feature engineering performed.
- Finally looked at overall distribution to see presence of any outliers that can be filtered out. The decision was to keep them since no reasoning can be found to remove them.
- Data analysis using some statistics, manipulation, tables and plot were performed afterwards.
- Machine Learning Modelling was also explored in the end after Exploratory Data Analysis was completed.