Data Intake Report

Name: G2M insight for Cab Investment firm
Report date: 12th June 2021
Internship Batch: LISUM01
Version:1.0
Data intake by: Samuel Alejandro Cueva Lozano
Data intake reviewer: Samuel Alejandro Cueva Lozano
Data storage location: https://github.com/DataGlacier/DataSets

Tabular data details:

Name of the file: Cab_Data.csv

Total number of observations	359393
Total number of files	1
Total number of features	7
Base format of the file	.csv
Size of the data	21.2MB

Name of the file: Customer_ID.csv

Total number of observations	49172
Total number of files	1
Total number of features	4
Base format of the file	.csv
Size of the data	1.1MB

Name of the file: Transaction_ID.csv

Total number of observations	440099
Total number of files	1
Total number of features	3
Base format of the file	.csv
Size of the data	9.0MB

Name of the file: City.csv

Total number of observations	21
Total number of files	1
Total number of features	3
Base format of the file	.csv
Size of the data	759 bytes

Proposed Approach:

Duplicate values

To consider two o more records as duplicates these must have the same values in the most of the attributes (excluding unique identifier) or very similar values in the case that they are strings (e.g. ‘Samuel Cueva Lozano’ = ‘Samy Cueva Loz’ and the others attributes be the same).

A record in the data set will be considered duplicate if there is a transaction with the same kilometers traveled, the same city, the same company and the same day.

Join tables

The table *Cab_Data.csv* will be joined with the table *Customer_ID.csv* by the table *Transaction_ID.csv* since the latter has the *Transaction ID* (from *Cab_Data.csv*) and Customer ID (from *Customer_ID.csv)* attributes, then the table *City.csv* will also join them.

Null values

The *Empty values, NaN values* and values like *“?”* will be treated as ***Null values***

Fields with ***too many null values*** would be removed or filled according to the following:

- Fields like *Age, Gender, Income, Company* and *Users* would be filled because they have missing values that depends on their hypothetical values or depend on other attributes.
- The rest of fields would be removed because they have missing values due to bad configuration, issues with data collection, or untraceable random reasons.

After cleaning at the field level, rows with null values would be removed.

Outlier detection

Outliers will be removed from numerical fields so that they don’t negatively affect the analysis.

A fixed threshold would be used for the *Age, Population* and *Users* attribute to avoid inconsistent data, then the IQR Score will be used to filter out the outliers in all attributes.

Data Transformation

The field *Date of Travel* from the table *Cab_Data.csv* could be transformed to a more readable format for better understanding.