Data Intake Report

Name: G2M insight for Cab Investment firm
Report date: 13.08.2023
Internship Batch: LISUM24
Version:<1.0>
Data intake by: Elena Stojanovska
Data intake reviewer:<intern who reviewed the report>
Data storage location: https://github.com/elensto/VC

Tabular data details:

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	19.2 MB

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	0.00061 MB

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	1.5 MB

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	10.1 MB

Proposed Approach:

1. **Data Profiling and Understanding:**

- Getting know the 4 datasets and the columns within them.
- Checking for total number of observations and number of unique values in each column.
- Merging the 4 datasets into one

2. **Data Cleaning:**

- Checking for missing values, converting data types, and formatting data:
    - For this case it was necessary to format (remove unwanted characters) in columns "Population" and "Users" and convert the data type for columns "Population", "Users" and "Date of Travel".

3. **Deduplication Criteria**

- Identifying the columns that are relevant for deduplication:
    - In this case columns of interest would be: "Date of Travel", "City", "KM Travelled", "Price Charged", "Cost of Trip", additionally customer's "Gender", "Age", "Income (USD/Month)" can also be considered.

4. **Deduplication Approach and Implementation:**

- Hashing: selected due to the length of the dataset. It calculates hash values for records based on selected columns and compares hash values.
- Implementation: on the cleaned dataset, considering the following columns: "Date of Travel", "City", "KM Travelled", "Price Charged", "Cost of Trip"

5. **Review and Validation:**

- Results: 16 potential duplicates were found, where the values for "Date of Travel", "City", "KM Travelled", "Price Charged", "Cost of Trip" were an exact match. However, the additional criteria columns that relate to customer identification did not match.
- Action: Given the mismatch with the additional criteria and that the number of duplicates is insignificant when compared to the total number of observations in the dataset, it was decided that the potential duplicates should stay in the dataset.