Data Intake Report

Name: <G2M\_Case\_Study>

Report date: <21-May-2024>

Internship Batch:< LISUM33>

Version:<1.0>

Data intake by:<Shahad Ali Aldawsari>

Data intake reviewer:<intern who reviewed the report>

Data storage location: <https://github.com/DataGlacier/DataSets>

**Tabular data details:**

|  |  |
| --- | --- |
| **Total number of observations** | <359392> |
| **Total number of files** | <4> |
| **Total number of features** | <14> |
| **Base format of the file** | <.csv> |
| **Size of the data** | <31.478KB> |

**Note: Replicate same table with file name if you have more than one file.**

|  |  |
| --- | --- |
| **File name** | <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** | <20.663KB> |

|  |  |
| --- | --- |
| **File name** | <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** | <1KB> |

|  |  |
| --- | --- |
| **File name** | <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** | <1.027KB> |

|  |  |
| --- | --- |
| **File name** | <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** | <8.788KB> |

**Proposed Approach:**

* First, I started reading datasets files from GitHub.
* Then I merged datasets files using commune columns 'except city dataset' so that each row indicates unique ride information.
* After that I started tchicking columns data types, nan values count, and memory usage, using info() function.
* I decided to change 2 columns data type (Customer ID and Transaction ID) from int to str.
* I tchicked same statistical information such (count, mean, standard deviation, minimum, maximum, lower, 50 and upper percentiles using describe() function for numeric columns.
* I also tchicked same statistical information such (count, unique, top, and frequency)
* After understanding dataset, I used same features to generate other useful features such as (Profit, Profit rate, age categories, and income categories).
* Finally, I made several assumptions, tested them, and drew conclusions.