Data Intake Report

Name: Data Glacier internship

Report date: Oct 10.2023

Internship Batch: LISUM26

Version:1.0

Data intake by: Mohamed Mohamed.

Data intake reviewer:< >

Data storage location: [Midohussien/week\_2-DataSets: Multiple data sets that contain information on 2 cab companies (github.com)](https://github.com/Midohussien/week_2-DataSets/tree/main)

**Tabular data details:**

* We have four datasets:

1. **Cab\_Data:**

|  |  |
| --- | --- |
| **Total number of observations** | 359392 rows |
| **Total number of files** | 1 |
| **Total number of features** | 7 columns |
| **Base format of the file** | csv |
| **Size of the data** | 20.1 MB |

1. **City:**

|  |  |
| --- | --- |
| **Total number of observations** | 20 rows |
| **Total number of files** | 1 |
| **Total number of features** | 3 columns |
| **Base format of the file** | .csv |
| **Size of the data** | 759 bytes |

1. **Customer\_ID:**

|  |  |
| --- | --- |
| **Total number of observations** | 49171 rows |
| **Total number of files** | 1 |
| **Total number of features** | 4 columns |
| **Base format of the file** | .csv |
| **Size of the data** | 1.00 MB |

4**.Transactions\_ID:**

|  |  |
| --- | --- |
| **Total number of observations** | 440098 rows |
| **Total number of files** | 1 |
| **Total number of features** | 3 columns |
| **Base format of the file** | .csv |
| **Size of the data** | 8.58 MB |

**Proposed Approach:**

1. **Objective**: Comparing the Pink Cab and Yellow Cab to figure out which one will be the good opportunity to invest in.
2. **Methodology:** using the different datasets given to find the best ways to compare between the two companies, as number of customers, number of transactions, net profit.
3. **Assumptions:**

* In the customer\_ID dataset, the customer\_ID column is the unique identifier for this dataset to link it to other datasets.
* In the Transaction\_ID dataset, the Transaction\_ID column is the unique identifier for this dataset to link it to other datasets.
* City dataset contain the city name that we’ll use with other datasets to find the demand in each city.