Data Intake Report

Name: Exploratory Data Analysis for ‘Yellow’ and ‘Pink’ Cab Companies

Report date: 26th of June 2021

Internship Batch: LISUM01

Version:<1.0>

Data intake by: Enias Vontas

Data intake reviewer:

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

**Tabular data details:**

|  |  |
| --- | --- |
| **Total number of observations** | 359392 |
| **Total number of files** | 1, **Cab\_Data.csv** |
| **Total number of features** | 7 |
| **Base format of the file** | .csv |
| **Size of the data** | 20.1 MB |

|  |  |
| --- | --- |
| **Total number of observations** | 49171 |
| **Total number of files** | 1, **Customer\_ID.csv** |
| **Total number of features** | 4 |
| **Base format of the file** | .csv |
| **Size of the data** | 1 MB |

|  |  |
| --- | --- |
| **Total number of observations** | 440098 |
| **Total number of files** | 1, **Transaction\_ID.csv** |
| **Total number of features** | 3 |
| **Base format of the file** | .csv |
| **Size of the data** | 8.58 MB |

|  |  |
| --- | --- |
| **Total number of observations** | 20 |
| **Total number of files** | 1, **City.csv** |
| **Total number of features** | 3 |
| **Base format of the file** | .csv |
| **Size of the data** | 1 KB |

No authorization was required to access the datasets and no missing values were found.

**Proposed Approach:**

* For possible duplicates we set as dictionaries (‘dict’ in Python) the ID columns of our master dataset and compared with the original columns of the datasets. We did not spot duplicates, but the number of customers changed from 49.171 to 46.020 unique values, meaning that 3.151 Customer Ids did not make any transactions and so are not included into our master dataset.