Data Intake Report

Name: Go-to-Market Exploratory Data Analysis

Report date: 10/14/22

Internship Batch: LISUM14

Version: 1.0

Data intake by: Pedram Doroudchi

Data intake reviewer:

Data storage location:

**Tabular data details:**

Cab\_Data.csv

|  |  |
| --- | --- |
| **Total number of observations** | 359,392 |
| **Total number of files** |  |
| **Total number of features** | 7 |
| **Base format of the file** | .csv |
| **Size of the data** | 21.2 MB |

Customer\_ID.csv

|  |  |
| --- | --- |
| **Total number of observations** | 49,171 |
| **Total number of files** |  |
| **Total number of features** | 4 |
| **Base format of the file** | .csv |
| **Size of the data** | 1.1 MB |

Transaction\_ID.csv

|  |  |
| --- | --- |
| **Total number of observations** | 440,098 |
| **Total number of files** |  |
| **Total number of features** | 3 |
| **Base format of the file** | .csv |
| **Size of the data** | 9 MB |

City.csv

|  |  |
| --- | --- |
| **Total number of observations** | 19 |
| **Total number of files** |  |
| **Total number of features** | 3 |
| **Base format of the file** | .csv |
| **Size of the data** | 759 bytes |

**Proposed Approach:**

* Used Pandas DataFrame duplicated() method to find zero observations with duplicate transaction ID’s
* Noted some large outlier values with respect to “price\_charged” feature but assumed this was due to some external factor like trip duration since there are about 13,500 values over 1,000 and 3 values over 2,000