Data Intake Report

Name: G2M Case Study

Report date: <2022-07-21>

Internship Batch: LISUM11:30

Version:<1.0>

Data intake by:<Jilin He>

Data intake reviewer:<intern who reviewed the report>

Data storage location: <>

**Tabular data details:**

Cab\_Data

|  |  |
| --- | --- |
| **Total number of observations** | 359392 |
| **Total number of features** | 7 |
| **Base format of the file** | .csv |
| **Size of the data** | 20663kb |

City

|  |  |
| --- | --- |
| **Total number of observations** | 20 |
| **Total number of features** | 3 |
| **Base format of the file** | .csv |
| **Size of the data** | 1kb |

**Customer\_ID**

|  |  |
| --- | --- |
| **Total number of observations** | 49171 |
| **Total number of features** | 4 |
| **Base format of the file** | .csv |
| **Size of the data** | 1027kb |

**Transaction\_ID**

|  |  |
| --- | --- |
| **Total number of observations** | 440098 |
| **Total number of features** | 3 |
| **Base format of the file** | csv |
| **Size of the data** | 8788kb |

**Proposed Approach:**

* Mention approach of dedup validation (identification)

Use series.nuinque(), dataframe.duplicated() or dataframe.drop\_duplicates()

* Mention your assumptions (if you assume any other thing for data quality analysis)

In Cab\_Data’s Date of Travel column, I assume the time base is 1900-01-01 so that the dates will range from 2016-01-01 to 2018-12-31