Data Intake Report

Name: Cab Industry Case Study

Report date: 20/6/2021

Internship Batch:2811291

Version:1.0

Data intake by:Colin Mburugu

Data intake reviewer:None

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

**Tabular data details:**

|  |  |
| --- | --- |
| **Total number of observations** | 359392 |
| **Total number of files** | 4 |
| **Total number of features** | 7 |
| **Base format of the file** | csv |
| **Size of the data** | 21.2mb |

|  |  |
| --- | --- |
| **Total number of observations** | 400098 |
| **Total number of files** | 4 |
| **Total number of features** | 3 |
| **Base format of the file** | csv |
| **Size of the data** | 9mb |

|  |  |
| --- | --- |
| **Total number of observations** | 49171 |
| **Total number of files** | 4 |
| **Total number of features** | 4 |
| **Base format of the file** | csv |
| **Size of the data** | 1.05mb |

|  |  |
| --- | --- |
| **Total number of observations** | 20 |
| **Total number of files** | 4 |
| **Total number of features** | 3 |
| **Base format of the file** | csv |
| **Size of the data** | 759B |

**Proposed Approach:**

* Mention approach of dedup validation (identification)

Used the df[column\_name].is\_unique () method to check if transaction id column of cab\_data and transaction\_id datasets to confirm each row was unique. Also applied the same method in customer\_id column in customer dataset. City data was so small just going through each row to confirm it’s a unique one.

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

Assumed the Date of Travel column range from 2016-01-02 to 2018-12-31.

Assumed the teen’s age range from 0-18,young adults 18-25,adults 25-35,mid-adults 35-55 ,the old 55 – 75 and senior citizens 75-100

Assumed annual income for low class ranges from 0-20000,middle class 20000- 45000,upper middle class 45000-140000,High income guys 140000-20000.