Data Intake Report

Name: Analyzing what affects the number of cab rides in a given 2 year period to decide which company to invest in.

Report date: 04/13/2023

Internship Batch: LISUM20

Version: 1.0

Data intake by: Anish Mitra

Data intake reviewer: NA

Data storage location: NA

**Tabular data details:**

1. Cab Date

|  |  |
| --- | --- |
| **Total number of observations** | 359392 |
| **Total number of files** | N/A |
| **Total number of features** | 7 |
| **Base format of the file** | csv |
| **Size of the data** | 59.92 MB |

1. City Data

|  |  |
| --- | --- |
| **Total number of observations** | 20 |
| **Total number of files** | N/A |
| **Total number of features** | 3 |
| **Base format of the file** | csv |
| **Size of the data** | * 1. kB |

1. Customer Data

|  |  |
| --- | --- |
| **Total number of observations** | 49171 |
| **Total number of files** | N/A |
| **Total number of features** | 4 |
| **Base format of the file** | csv |
| **Size of the data** | 4.03 MB |

1. Transaction Data

|  |  |
| --- | --- |
| **Total number of observations** | 440098 |
| **Total number of files** | N/A |
| **Total number of features** | 3 |
| **Base format of the file** | csv |
| **Size of the data** | 32.32 MB |

1. Inflation Data

|  |  |
| --- | --- |
| **Total number of observations** | 123 |
| **Total number of files** | N/A |
| **Total number of features** | 3 |
| **Base format of the file** | csv |
| **Size of the data** | 9.61kB |

1. Customer Data

|  |  |
| --- | --- |
| **Total number of observations** | 499 |
| **Total number of files** | N/A |
| **Total number of features** | 8 |
| **Base format of the file** | csv |
| **Size of the data** | 59.21kB |

1. Transaction Data

|  |  |
| --- | --- |
| **Total number of observations** | 342 |
| **Total number of files** | N/A |
| **Total number of features** | 6 |
| **Base format of the file** | csv |
| **Size of the data** | 77.13kB |

**Note: Replicate same table with file name if you have more than one file.**

**Proposed Approach:**

* Mention approach of dedup validation (identification)

I ran the info() method to see if there were any null values which there were not in any of the dataframes and then I used value\_counts() method to see if there were any unwanted duplicates and there were not here either.

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

I don’t.