Data Intake Report

Name: G2M insight week 2
Report date: 1/12/23
Internship Batch: LISUM17
Version:<1.0>
Data intake by: Serena Yuan
Data intake reviewer:<intern who reviewed the report>
Data storage location: https://github.com/DataGlacier/DataSets.git

Tabular data details:

Cab Data:

Total number of observations	359392
Total number of files	1
Total number of features	7
Base format of the file	csv
Size of the data	21.2 MB

City Data:

Total number of observations	20
Total number of files	1
Total number of features	3
Base format of the file	csv
Size of the data	759 B

Customer Data:

Total number of observations	49171
Total number of files	1
Total number of features	4
Base format of the file	csv
Size of the data	1.05 MB

Transaction Data:

Total number of observations	440098
Total number of files	1
Total number of features	3
Base format of the file	csv
Size of the data	9 MB

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

Proposed Approach:

The approach of data deduplication I am using involves using pandas drop duplicates method in pandas dataframes but using subsets of column names that do not involve their ids which are unique. This involves assuming for cab data that the transaction id are not included in the subset for dropping duplicates.

This method reduced the number of rows of the cab data.