Data Intake Report

Name: Week 2 EDA
Report date:
Internship Batch: LISUM18
Version:<1.0>
Data intake by:Drew Springfield
Data intake reviewer:
Data storage location:

Tabular data details:

City.csv

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	1kB

Transaction_ID.csv

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	8.788MB

Customer_ID.csv

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	1.027MB

Cab_Data.csv

Total number of observations	359392
Total number of files	1
Total number of features	8
Base format of the file	.csv
Size of the data	20.663MB

Proposed Approach:

My approach to exploring this dataset is to combine the 4 csv files into one master list. The benefit of this is querys become more efficient by not traversing over or comparing multiple files. I chose to dedupe the master list instead of each file individually because duplicates in certain files may not actually be duplicates when combined with other data. My goal is to find which investment opportunity creates a higher expected ROI.