Data Intake Report

Name: EDA G2M Insight for Cab Investment Firm
Report date: 5/14/2024
Internship Batch: LISUM33
Version:1.0
Data intake by: Jacob Farrington
Data intake reviewer: N/A
Data storage location: https://github.com/S0n0f1saac/G2M-Insight

Tabular data details:

Cab_Data.csv

Total number of observations	440097
Total number of files	1
Total number of features	7
Base format of the file	.csv
Size of the data	20,663 KB

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	1 KB

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	1,027 KB

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	8,788 KB

Proposed Approach:

- I began by viewing the csv files provided to familiarize myself with the data sets’ structure, I then began to explore the data via descriptive statistics and data visualization to gain insight and potentially identify duplicates. I performed exact match deduplication on the columns I identified and then compared the two files to one another. I remo9ved 8705 duplicates from key columns.