Data Intake Report

Name: G2M Insight for Cab Investment Firm
Report date: September 19th 2022
Internship Batch: LISUM13:30
Version: 1.0
Data intake by: Devika Chandnani
Data intake reviewer:
Data storage location:
https://github.com/devikachandnani/G2M-Insight-for-Cab-Investment-Firm

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	21.2MB

Tabular data details: City

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 Bytes

Tabular data details: Customer_ID

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	1.1MB

Tabular data details: Transaction_ID

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	9MB

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

Proposed Approach:

- Deduplication not needed, data has no duplicates, verified when testing on Join
- Upon deeper look, identified that some cities don't have State information, that was fixed for each city in both Cab and City data which contain the City information.
- Date format needed to be converted.
- San Francisco not found in cab data, no transaction information available for cab rides in SF.
- Data Quality Analysis: Data is accurate, complete, consistent.