Data Intake Report

Name: G2M Case Study
Report date: June 14, 2024
Internship Batch: LISUM34
Version:<1.0>
Data intake by: Hassan Eisa
Data intake reviewer: N/A
Data storage location: Github

Tabular data details: Customer_ID.csv

Total number of observations	49,170 rows
Total number of files	1
Total number of features	4 columns
Base format of the file	.csv
Size of the data	1 MB

Tabular data details: City.csv

Total number of observations	18 rows
Total number of files	1
Total number of features	3 columns
Base format of the file	.csv
Size of the data	758 bytes

Tabular data details: Transaction_ID.csv

Total number of observations	440,000 rows
Total number of files	1
Total number of features	3 columns
Base format of the file	.csv
Size of the data	8.9 MB

Tabular data details: Cab_Data.csv

Total number of observations	359,390 rows
Total number of files	1
Total number of features	7 columns
Base format of the file	.csv
Size of the data	21 MB

Proposed Approach:

- Utilized Pandas methods (‘df.duplicated()’ and ‘df.drop_duplicates()’) to identify and manage duplicates based on specified columns.
- Assumed duplicates are unwanted and implement deduplication strategies to maintain data integrity and ensure accurate analysis outcomes.