

Data Intake Report

Name: G2M Case Study
Report date: July 4, 2023
Internship Batch: LISUM21
Version: 1.0
Data intake by: Parwinder Singh
Data intake reviewer: N/A
Data storage location: Github

Tabular data details: Cab_Data.csv

Total number of observations	359,392 rows
Total number of files	1
Total number of features	7 columns
Base format of the file	.csv
Size of the data	21.2 MB

Tabular data details: City.csv

Total number of observations	19 rows
Total number of files	1
Total number of features	3 columns
Base format of the file	.csv
Size of the data	759 bytes

Tabular data details: Customer_ID.csv

Total number of observations	49,171 rows
Total number of files	1
Total number of features	4 columns
Base format of the file	.csv
Size of the data	1.1 MB

Tabular data details: Transaction_ID.csv

Total number of observations	440,098 rows
Total number of files	1
Total number of features	3
Base format of the file	.csv
Size of the data	9 MB

Proposed Approach:

- Ran certain code to check for duplicates within each dataset using code like `.duplicated().sum()` which resulted in no duplicates for each dataset.
- I assumed middle-class 'Income (USD/Month)' individuals having a majority within the cab market.
- Assumed 'Cost of Trip' accounts for all expenses for both companies.