Data Intake Report

Name: G2M Case Study
Report date: 04.10.2021
Internship Batch: LISUM04
Version: 1.0
Data intake by: Tolga Yaz
Data intake reviewer:
Data storage location:

Tabular data details:

Name of the File	Cab_Data.csv
Total number of observations	359 392
Total number of features	7
Base format of the file	.csv
Size of the data	20.10 MB

Name of the File	Customer_ID.csv
Total number of observations	49 171
Total number of features	4
Base format of the file	.csv
Size of the data	1.00 MB

Name of the File	Transaction_ID.csv
Total number of observations	440 098
Total number of features	3
Base format of the file	.csv
Size of the data	8.58 MB

Name of the File	City.csv
Total number of observations	20
Total number of features	3
Base format of the file	.csv
Size of the data	4.00 kB

Name of the File	dfMain.csv
Total number of observations	359 392
Total number of features	15
Base format of the file	.csv
Size of the data	37.30 MB

Proposed Approach:

- The data was transformed into Python Pandas Dataframes to be processed.
- The dataframes were observed carefully to identify the nulls but there was no null.
- .duplicated() method of Python Pandas Library was used to identify duplicates. No necessary duplicates were found to drop.
- The four .csv files that were provided were turned into one main file (dfMain.csv) after some necessary steps like dropping the unnecessary columns and adding some necessary columns. The changes can be observed in the Jupyter Notebook provided.