Data Intake Report

Name: G2M insight for Cab Investment Firm
Report date: 12/05/2023
Internship Batch: LISUM21
Version: 1.0
Data intake by: Zohra Bouchamaoui
Data intake reviewer: Zohra Bouchamaoui
Data storage location: Github

Tabular data details - Cab_Data.csv:

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	21.2 MB

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	1.1 MB

Tabular data details - 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	9 MB

Tabular data details - 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	759 Bytes

Proposed Approach:

- Mention approach of dedup validation (identification)

1. For dedup validation, I have first loaded the datasets and checked for any missing values using .isnull().sum().
2. I have also checked for any duplicates using .drop_duplicates(inplace=True)

- Mention your assumptions (if you assume any other thing for data quality analysis)

1. The date of travel provided in the Cab_Data.csv we’re not in the correct format. I used the datetime function in Python to convert the values in the ‘Date of Trip’ column and I assume that the dates generated by Python are correct (as they are within the range provided in the assignment details)
2. I assume that the data is accurate, meaning that it correctly represents the state of things during the given time period (cab fares, distances travelled, city population, etc.)
3. I assumed that the data is consistent, meaning that there are no conflicting or contradictory values within the dataset. I have checked for duplicates and any missing values but nothing was found.
4. I assume that the data provided by Data Glacier is an open source and that I am allowed to use it to complete this task.