Data Intake Report

Name: G2M Insight for Cab Investment Firm
Report date: 13/4/2025
Internship Batch: LISUM44
Version: 1.0
Data intake by: Yash Parmar
Data intake reviewer: N/A
Data storage location: <location URL eg: github, cloud>

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

- Customer_ID.csv

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

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

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

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

Proposed Approach:

- Mention approach of dedup validation (identification)
    - I removed any duplicate rows from the merged data frame, using "drop_duplicates()" on the data frame. I think this would be the best option since I assumed that Transaction ID would be unique to each transaction, therefore there would not be many duplicates, even if there were at all.
- Mention your assumptions (if you assume any other thing for data quality analysis)
    - Other than the fact that each Transaction ID was unique to a transaction, and therefore there would not be any duplicates, for dropping duplicate rows. I also assumed that the smallest date number provided was the first date, and everything after that would be numerically ordered. It's also assumed that the data is accurate, and does not have any discrepancy between units, such as kilometers or USD/hour. One last thing that I assumed was that the outliers were due to real and meaningful circumstances such as charging more for more people in the car, not due to some racial profiling, car accident, or any circumstances outside the norm.