Data Intake Report

Name: G2M Insight for Cab Investment firm
Report date: 3/14/2025
Internship Batch: LISUM43
Version: 1.0
Data intake by: Saif Amer
Data intake reviewer: Saif Amer
Data storage location: https://github.com/Saif-Amer/Data-Glacier/tree/main/week2

Tabular data details:

Cab_Data.csv

Total number of observations	359393
Total number of files	4
Total number of features	7
Base format of the file	.csv
Size of the data	20.2 MB

City.csv

Total number of observations	21
Total number of files	4
Total number of features	3
Base format of the file	.csv
Size of the data	759 Bytes

Customer_ID.csv

Total number of observations	49172
Total number of files	4
Total number of features	5
Base format of the file	.csv
Size of the data	1MB

Transactoin_ID.csv

Total number of observations	440099
Total number of files	4
Total number of features	3
Base format of the file	.csv
Size of the data	8.58 MB

Proposed Approach:

- Mention approach of dedup validation (identification)
- Mention your assumptions (if you assume any other thing for data quality analysis)

Handling Missing Values:

- **No major missing values were found.**

Checking for Duplicates:

- **Removed duplicate records based on Transaction ID.**

Data Type Adjustments:

- **Converted Date of Travel to datetime format.**

Merging Datasets:

- **Joined Cab_Data, Customer_ID, Transaction_ID, and City into a Master Dataset.**