Data Intake Report

Name: G2M Insight for Cab Investment Firm
Report date: July 20, 2025
Internship Batch: LISUM47
Version:<1.0>
Data intake by: Paula McCree-Bailey
Data intake reviewer:<intern who reviewed the report>
Data storage location: https://github.com/pmb-7684/DataGlacier_Internship/tree/main/Module2

Tabular data details:

	Cab_Data.csv
Total number of observations	359,392
Total number of files	1
Total number of features	7
Base format of the file	.csv
Size of the data	20,633 KB

	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	1 KB

	Customer_ID.csv
Total number of observations	49,171
Total number of files	1
Total number of features	4
Base format of the file	.csv
Size of the data	1,027 KB

	Transaction_ID.csv
Total number of observations	440,098
Total number of files	1
Total number of features	3
Base format of the file	.csv
Size of the data	8,788 KB

	Result_df.csv
Total number of observations	359,392
Total number of files	1
Total number of features	24
Base format of the file	.csv
Size of the data	8,788 KB

Proposed Approach:

- Dedup validation (identification)
  To check for duplicates in the dataset, `duplicates = result_df[result_df.duplicated()] ` which scans for rows that are exact duplicates of previous ones. Also used `duplicates = result_df[result_df.duplicated(subset=key_cols, keep=False)]` to make sure all duplicates are captured. There are no duplicate observations in our dataset.
- Assumptions:
  - The "Price Change" feature contains a lot of outliers. We are provided with the duration or rate. The decision was to retain the outliers.
  - "User feature represents the number of people taking a cab.
  - "Profit" feature is created as revenue (Price_charges) – cost of goods (Cost_of_trip)