Data Intake Report

Name: G2M insight for Cab Investment firm
Report date: 13/07/2025
Internship Batch: LISUM47
Version: 1.0
Data intake by: Keval Jaysukhbhai Savaliya
Data intake reviewer:
Data storage location: https://github.com/DataGlacier/DataSets

Tabular data details:

File Name	Cab_Data
Total number of observations	359392
Total number of files	-
Total number of features	7
Base format of the file	.csv
Size of the data	20.1 MB

File Name	City
Total number of observations	20
Total number of files	-
Total number of features	3
Base format of the file	.csv
Size of the data	759 bytes

File Name	Customer_ID
Total number of observations	49171
Total number of files	-
Total number of features	4
Base format of the file	.csv
Size of the data	1.00 MB

File Name	Transaction_ID
Total number of observations	440098
Total number of files	-
Total number of features	3
Base format of the file	.csv
Size of the data	8.58 MB

Objective:

This report documents the quality checks and understanding of the 4 provided datasets, as part of XYZ's investment analysis in the U.S. cab industry. The datasets cover cab transactions, customer profiles, transaction mappings, and city-level demographic data.

Proposed Approach:

1. **Data Cleaning:**
    a. **Null Value Check:**
        i. **Approach for check:**
            - Used **.isnull().sum()** in Pandas to identify any row has null values or not.
            - If duplicates were found, we would drop them using **.dropna()**

    b. **Duplicate Value Check:**
        i. **Approach for check:**
            - Used **.duplicated().sum()** in Pandas to identify any full-row or key-based duplicates.
            - If duplicates were found, we would drop them using **.drop_duplicates()**

2. **Column-Wise Format Fixes:**
    a. In Cab Data, 'Date of Travel' was in Excel Format, so it was converted to datetime using the origin 1899-12-30.
    b. In City Data, 'Population' and 'Users' were in objects, so I removed commas and converted to int64.

3. **Merging Approach:**
    - Cab_Data.csv and Transaction_ID.csv merged on Transaction ID, merged with Customer_ID.csv on Customer ID, merged with City.csv on City field.

4. **Assumptions:**
    - Transaction ID is the primary key for cab rides and is unique.
    - Customer ID is consistent across all tables.
    - City name spelling is standardised and matches across files.
    - Users in the City.csv file refer to the total cab users, not just Yellow or Pink.

5. **Outlier detection**
    - Outliers were both visually and statistically recognised in 'Price Charged' and 'Population'.

6. **Correlation Observations**
    - From my correlation (.corr()) analysis, I found the following positive relation observed in the features:
      - KM Travelled - Price Charged
      - KM Travelled - Cost of Trip
      - Price Charged - Cost of Trip
      - Population - Users