Data Intake Report

Name: <Cab Insight G2M Project>
Report date: <May 17, 2025>
Internship Batch:<LISUM44>
Version:<1.0>
Data intake by:<Victoria Yang>
Data intake reviewer:<intern who reviewed the report>
Data storage location: <https://github.com/vicxiya24>

Tabular data details:
File Name: **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	<20.2 MB>

File Name: 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	<8.6 MB>

File Name: **Cab_Data.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 MB>

File Name: **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 B>

Proposed Approach:

- Population and User counts in City.csv were cleaned by removing commas and converting to integers
- City names were normalized (uppercased and stripped) before merging
- All date values in Cab_Data.csv were converted from Excel serial format using origin="1899-12-30" in pd.to_datetime
- Used .drop_duplicates() on all merged datasets to ensure uniqueness
- Primary Keys of Transaction ID and Customer ID
- Verified consistency during joins using .merge(validate='one_to_one') to ensure no unexpected duplication