Data Intake Report

Name: Cab Investment Firm Analysis
Report date: 03/15/2025
Internship Batch: LISUM43
Version: 1.0
Data intake by: Duong Duc
Data intake reviewer:
Data storage location: GitHub Repository

Tabular data details:

Cab_Data.csv

Total number of observations	359392
Total number of files	4
Total number of features	7
Base format of the file	.csv
Size of the data	59.92 MB

City.csv

Total number of observations	20
Total number of files	4
Total number of features	3
Base format of the file	.csv
Size of the data	0.00MB

Customer_ID.csv

Total number of observations	49171
Total number of files	4
Total number of features	4
Base format of the file	.csv
Size of the data	4.03 MB

Transaction_ID.csv

Total number of observations	440098
Total number of files	4
Total number of features	3
Base format of the file	.csv
Size of the data	32.32 MB

Proposed Approach:

- Deduplication and Validation:
  - Duplicates will be identified and removed based on Transaction ID and Customer ID.
  - Columns with missing values will be removed.
  - Outliers in Price_Charged and Cost_of_Trip will be flagged for further reviewed.
- Assumptions:
  - Profit is calculated only using Priced_Charged and Cost_of_Trip to calculate profit.
  - Because there is no data on trip duration, we cannot pinpoint outliers in the Price_Charged feature.

Note: Convert this doc in pdf and provide the link of pdf file in your dashboard.
Please do not forget to remove this section while converting the file into pdf.