

Data Intake Report

Name: Cab Investment Analysis

Report date: 08/14/2024

Internship Batch: LISUM36

Version:1.0

Data intake by: Brittney Smith

Data intake reviewer: Self

Data storage location: [github](#).

Tabular data details:

Cab_Data.csv

Total number of observations	10,888
Total number of files	1
Total number of features	7
Base format of the file	..csv
Size of the data	Approximately 1.4 MB

Customer_ID.csv

Total number of observations	4913
Total number of files	1
Total number of features	3
Base format of the file	.csv
Size of the data	Approximately 163KB

Transaction_ID.csv

Total number of observations	21,671
Total number of files	1
Total number of features	3
Base format of the file	.csv
Size of the data	Approximately 489 KB

City.csv

Total number of observations	49
Total number of files	1
Total number of features	3
Base format of the file	.csv
Size of the data	Approximately 3 KB

Note: Replicate same table with file name if you have more than one file.

Proposed Approach:

- Mention approach of dedup validation (identification)
 - We checked for duplicate entries by identifying unique combinations of key fields such as Transaction ID in the Transaction_ID.csv file. Any rows with identical Transaction ID values were considered duplicates and were removed to ensure data quality
- Mention your assumptions (if you assume any other thing for data quality analysis)
 - We assumed that all datasets were recorded consistently and that all necessary fields were present. Additionally, we assumed that the population data in City.csv was up to date and accurate for the time period of the analysis.
 - Another assumption was that the format of the Price Charged and Cost of Trip fields were consistent and free of currency symbols or commas, which could affect numerical analysis.