Data Intake Report

Name: XYZ Cab Investment Analysis
Report date: 14-01-2025
Internship Batch: LISUM41
Version: 0.1
Data intake by: Shubham More
Data intake reviewer: -
Data storage location: -

Tabular data details:

1. Cab_Data

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	21.8 MB

2. Customer_ID

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

3. Transactions_ID

Total number of observations	44098
Total number of files	1
Total number of features	3
Base format of the file	.csv
Size of the data	8.58 MB

4. City

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

5. Master

Total number of observations	359392
Total number of files	1
Total number of features	14
Base format of the file	.csv
Size of the data	42.1 MB

Approach for Deduplication Validation (Identification)

- We checked for duplicate entries in all datasets, including the combined master dataset.
- No duplicates were found, meaning every row in the data is unique.
- This ensures that our analysis will be accurate and not affected by repeated information.

Assumptions for Data Analysis

- The provided datasets are assumed to cover the complete range of transactions for the specified time period.
- Outliers in features like **`Price_Charged`** or **`Cost_of_Trip`** are assumed to be valid unless they appear to be obvious data entry errors.
- The **`Users`** column in the **`city`** dataset represents all cab users in the city, not just the two companies being analyzed.