Data Intake Report

Name: XYZ cab industry analysis
Report date: 13th June 2024
Internship Batch: LISUM34
Version:1.0
Data intake by: Biswadip Bhattacharyya
Data intake reviewer:
Data storage location:

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	21
Total number of files	
Total number of features	3
Base format of the file	.csv
Size of the data	1 kb

File name: Customer_ID

Total number of observations	49172
Total number of files	
Total number of features	4
Base format of the file	.csv
Size of the data	1 MB

File name:Transaction_ID

Total number of observations	440099
Total number of files	
Total number of features	3
Base format of the file	.csv
Size of the data	8.58 MB

Proposed Approach:

- Reviewed each dataset to understand the structure and relationship between the files.

- Identified key features in each dataset (eg, ‘Transaction ID’, ‘Customer ID’, ‘Date of travel’,’City’).
- Load each dataset into pandas dataframe for analysis.
- Checked for missing values and there was no missing values either.
- Extracted additional features such as ‘Day of Week’ and ‘Age Group’.
- Applied the duplicated() method to detect duplicate and confirmed that no duplicate records were found in any of the datasets.
- Merged the 4 datasets into a master dataset.
- Aggregated data to compute metrics such as total rides, average revenue per ride and total profit by company, city and day of the week.
- Prepared the data for hypothesis testing and visualization.

Assumptions:

- ‘Transaction ID’ and ‘Customer ID’ are assumed to be unique identifiers.
- Monetary values such as ‘Price charged’, ‘Cost of Trip’, and ‘Income’ are assumed to be in USD.