Data Intake Report

Name: Cab Industry Analysis

Report date:13/02/2024

Internship Batch:LISUM30

Version:<1.0>

Data intake by:Omar Mahmoud Hamdan

Data intake reviewer:Data Glacier

Data storage location: omarhamdaan/Internship_EDA_W2 (github.com)

Tabular data details:

Transaction_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	8.5MB

Tabular data details:

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	1KB

Tabular data details:

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	20MB

Tabular data details:

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	1027KB

Deduplication Validation (Identification):

1. Utilized Python's pandas library to identify and remove duplicate entries based on key identifiers such as Transaction ID and Customer ID.
2. Assured unique entries for accurate analysis.

Assumptions for Data Quality Analysis:

1. Assumed that each entry in the Transaction and Cab Data files represents a unique ride.
2. The 'Date of Travel' fields were considered accurate and were converted to a datetime format for time series analysis.
3. Population and Users data in the City.csv file were assumed to be the latest and relevant to the period of the cab data.