Data Intake Report

Name: EDA Cab Companies
Report date: 6th of March 2023
Internship Batch: LISUM19
Version: N/A
Data intake by: Mario Rodrigues Peres
Data intake reviewer: N/A
Data storage location: GitHub

Tabular data details:

Dataset1 - 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	21.2 MB
Data type (unique values) of features	Transaction ID int64 (359392) Date of Travel int64 (1095) Company object (2) City object (19) KM Travelled float64 (874) Price Charged float64 (99176) Cost of Trip float64 (16291)

Dataset2 - 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	759 B

Data type (unique values) of features	City object (20) Population object (20) Users object (20)

Dataset3 - 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	1.1 MB
Data type (unique values) of features	Customer ID int64 (49171) Gender object (2) Age int64 (48) Income (USD/Month) int64 (23341)

Dataset4 - Transaction_ID

Total number of observations	440098
Total number of files	1
Total number of features	3
Base format of the file	csv
Size of the data	9 MB
Data type (unique values) of features	Transaction ID int64 (440098) Customer ID int64 (49171) Payment_Mode object (2)

Proposed Approach:

- There is no duplicated registers in any of the provided tables (checked using duplicated function of pandas library in python). Tables 1, 3 and 4 have unique IDs, whereas table 2 presents unique cities. Therefore, all registers are unique in all datasets.
- Data was merged into a single table and, basically, descriptive statistics was used to assess the dataset.