Data Intake Report

Name: G2M insight for Cab Investment firm
Report date: 12/12/2023

Internship Batch: LISUM28

Version: 1.0

Data intake by: Parthkumar Narotambhai Sutariya

Data intake reviewer:

Data storage location: https://github.com/parthsutariya31/Cab_Data_Analysis.git

Tabular data details:

Total number of observations

359392

Total number of files

Total number of features

7

Base format of the file

.csv

Size of the data

20MB

Total number of observations

20

Total number of files

Total number of features

3

Base format of the file

.csv

Size of the data

1KB

Total number of observations

49171

Total number of files

Total number of features

4

Base format of the file

.csv

Size of the data

1MB

Total number of observations

440098

Total number of files

Total number of features

3

Base format of the file

.csv

Size of the data

8.5MB

Proposed Approach:

- At first, I have merged every dataset into master dataset on the basis of similar featurefrom each dataset.
- Then, To Check dedupe I have only checked 'Transaction_ID' feature from master dataset because transaction must be different for every observation. So, if transaction id isdifferent than that particular observation is different from all other observations. I checked it with pandas.nunique() method to count unique transaction id. Master dataset have total 359392 observation and it also has 359392 different Transaction_ID. So, everyobservation is different from each other.

- I noticed outliers in ‘Price_Charged’ and ‘Population’. We can remove it ifwe want totrain this data.
- Moreover,I have found positive correlation between some feature such as
    - Km_Travelled - Price Charged
    - KM_Travelled - Cost_of_Trip
    - Price_Charged - Cost_of_Trip
    - Population – Users