Data Intake Report

Name: Insight for Cab Investment Firm
Report date: May 13 2023
Internship Batch: LISUM21
Version:<1.0>
Data intake by: Zongdao Wen
Data intake reviewer:<intern who reviewed the report>
Data storage location: Github

Tabular data details:

Transaction_ID

Total number of observations	440098
Total number of files	4
Total number of features	3
Base format of the file	csv
Size of the data	9MB

Customer_ID

Total number of observations	49171
Total number of files	4
Total number of features	4
Base format of the file	csv
Size of the data	1.1MB

City

Total number of observations	440098
Total number of files	4
Total number of features	3
Base format of the file	csv
Size of the data	759bytes

Cab Data

Total number of observations	440098
Total number of files	4
Total number of features	8
Base format of the file	csv
Size of the data	21.9MB

Note: Replicate same table with file name if you have more than one file.

Proposed Approach:

The 4 CSV files are combined using pd.merge. First, Transcation_ID and Cab_Data are merged based on Transactio ID column. The resulting df is then merged with Customer_ID

based on Customer ID column. Finally, the resulting df is mergered with City based of City column. Rows with n/a values are dropped.

The date column of the final combined data frame is then being separated into two columns: Year and Month. In addition, profit per km column is created and added to the data frame for every observation. The resulting data frame contains 359392 rows and 17 columns.

Assumptions:

Profits are calculated strictly by finding the difference between Price_charged and Cost_of_Trip.

Rows with n/a values are dropped. The resulting data has no outliers. Every ride has a Customer ID and Transaction ID.