Data Intake Report

Name: G2M insight for Cab Investment firm Case Study
Report date: 2024.7.15
Internship Batch: LISUM35
Version:1.0
Data intake by: Kua Hong Rui
Data intake reviewer:
Data storage location:

Tabular data details:

Cab_Data.csv

Total number of observations	359, 392
Total number of files	
Total number of features	7
Base format of the file	.csv
Size of the data	21.5 MB

City.csv

Total number of observations	22
Total number of files	
Total number of features	3
Base format of the file	.csv
Size of the data	4.00 KB

Customer_ID.csv

Total number of observations	49, 171
Total number of files	
Total number of features	4
Base format of the file	.csv
Size of the data	1.00 MB

Transaction_ID.csv

Total number of observations	440, 091
Total number of files	
Total number of features	3
Base format of the file	.csv
Size of the data	8.58 MB

Proposed Approach:

- Check missing values for every column in each data set, found no missing values for every data set.

- Used drop_duplicates function on python to remove duplicate rows after merging each data sets into one final data set.
- The profit generated from each ride is calculated while holding other variables constant. Only the Price_Charged and Cost_of_Trip features are utilized to determine the profit.
- In this analysis, the user feature of the city dataset is defined as the number of cab users in the city (Including Yellow and Pink Cabs users and other cab users).