Data Intake Report

Name: EDA of G2M Cab Datasets
Report date: 07/19/2024
Internship Batch: LISUM35
Version: 1.0
Data intake by: Haipei Xu

Tabular data details:

Cab_Data

Total number of observations	359392
Total number of features	7
Base format of the file	csv
Size of the data	20.18 MB

City

Total number of observations	20
Total number of features	3
Base format of the file	csv
Size of the data	759 bytes

Customer_ID

Total number of observations	49172
Total number of features	4
Base format of the file	csv
Size of the data	1 MB

Transactio_ID

Total number of observations	440098
Total number of features	3
Base format of the file	csv
Size of the data	8.58 MB

Provided with these four datasets—Cab_Data.csv (transaction details for two cab companies), Customer_ID.csv (customer demographics), Transaction_ID.csv (transaction-customer mapping and payment modes), and City.csv (US cities' population and cab user numbers). we are going to analyze these datasets and present visuals, analysis, and recommendations to XYZ's Executive team, helping them identify the best investment option.

Proposed Approach:

- Mention approach of dedup validation (identification): If there are two same rows in one table, then remove the duplicate
- Mention your assumptions (if you assume any other thing for data quality analysis): Nope