Data Intake Report

Name: G2M insight for Cab Investment firm
Report date:
Internship Batch: LISUM14
Version: 1.0
Data intake by: Wen Li
Data intake reviewer:
Data storage location:

Tabular data details:

Cab_Data:

Total number of observations	359392
Total number of files	
Total number of features	7
Base format of the file	.csv
Size of the data	19.2+ MB

City:

Total number of observations	20
Total number of files	
Total number of features	3
Base format of the file	.csv
Size of the data	608/0+ bytes

Customer_ID:

Total number of observations	49171
Total number of files	
Total number of features	4
Base format of the file	.csv
Size of the data	1.5+ MB

Transaction_ID:

Total number of observations	440098
Total number of files	
Total number of features	3
Base format of the file	.csv
Size of the data	10.1+ MB

Proposed Approach:

- Mention approach of dedup validation (identification)

  We firstly merge customer and transaction data. Customer ID is the primary key of customer data and the foreign key of transaction data. 1 customer may have 1 or multiple transactions and 1 transaction only belongs to 1 customer.

  Secondly, we merge cab and city data. City is the primary key of city data and the foreign key of cab data. 1 city may have 1 or multiple transactions and 1 transaction only belongs to 1 city.

  We then merge above data together to get final data set and detect duplicate rows in the dataframe in order to avoid bias or the loss of accuracy.

  We also detect if there are missing values in the dataframe and remove outliers which can make a detrimental effect on model performance.

  One of the simplest methods for detecting outliers is box plots. A box plot is a graphical display for describing the distributions of the data. Box plots use the median and the lower and upper quartiles.

- Mention your assumptions (if you assume any other thing for data quality analysis)

  Is there any seasonality in number of customers using the cab service?

  Which gender uses the cab service the most?

  Which company is more profitable?

  Which city has more users?

  Is the income proportional to the use of cab service?

  Is the age proportional to the use of cab service?