Data Intake Report

Name: G2M insight for Cab Investment firm
Report date: 7th October 2021
Internship Batch: LISUM04
Version: <1.0>
Data intake by: Mazen Hawwa
Data storage location: https://github.com/xotofloyt/G2M-insight-for-cab-investment.git

Tabular data details:

Cab_Data.csv

Total number of observations	359392
Total number of features	7
Base format of the file	CSV
Size of the data	20.6 MB

Transaction_ID.csv

Total number of observations	440098
Total number of features	3
Base format of the file	CSV
Size of the data	8.5 MB

Customer_ID.csv

Total number of observations	49171
Total number of features	4
Base format of the file	CSV
Size of the data	1 MB

City.csv

Total number of observations	20
Total number of features	3
Base format of the file	CSV
Size of the data	1 KB

Proposed Approach:

- Each file was read into pandas dataframe, rows and features examined for duplicates and missing values.
- Duplicates dropped, converting some datatypes from string to int and merging all 4 files with inner join.
- Main data file is Cab_Data.csv, remaining files are mapping files to provide information about transactions, users and city information that is useful for analysis.
- Given is that data reading is from 2016.1.31 till 2018.12.31
- After reading in the datasets:
    - Cab data was found to contain 8 duplicated entries(when we don't take Transaction IDs into account) and they were dropped.
    - In Cities data, Population and Users column had to be converted from text to int.
    - True range of the data was from 2016.1.4 till 2nd 2019.1.2
    - 6 features were created for the various analysis performed:

      1- Profit/KM - to determine Profit per distance for each trip

      2- Cost/KM - to determine Cost per distance for each trip

      3- Users % Population - to determine the percentage of the population of each city that are registered users

      4- Age Group - to divide the age of users into 5 groups and analyze by age group.

      5- % Profit of Company Total - the percentage of profit of each trip of the total company profit

      6- Income Group - to divide the users into 5 income groups and analyze by income group.