Data Intake Report

Name: G2m insight for Cab Investment firm
Report date: June 2021
Internship Batch: LISUM01
Version:1.0
Data intake by: Nathan Adam
Data intake reviewer:Nathan Adam
Data storage location: https://nbviewer.jupyter.org/github/N-A-ML/EDA/blob/main/EDA%20notebook.ipynb

Tabular data details:

Total number of observations	359392 for the combined file. 80706 rows were lost after merging.
Total number of files	1 (1 csv file created from 4 merged csv files)
Total number of features	16 (14 original, 2 created, some new dataframes and variables were also created when using .groupby and .resample with sum / mean etc in Python)
Base format of the file	.csv for all
Size of the data	35.8MB for the 1 combined file.

Proposed Approach:

- .duplicated() was used in Python to identify duplicates, none were found.

- Assumptions:
  1) Outliers were found for Price_Charged, but these were not removed since it's reasonable to believe that some trips were very long, and there was no data for trip duration.
  2) For some rows, for each company, there were some instances where the Price_Charged < Cost_of_Trip. We assume that there is no undercharging, and that this can be explained in another way. For example, perhaps the drivers were stuck in traffic for a long time.
  3) We assume that the 'Users' variable from the City dataset includes Yellow Cab and Pink Cab.
  4) The Cost of Trip variable includes literally all relevant costs for the business such as fuel, cab driver's wages, business running costs, and income tax and VAT, etc
  5) Profit per trip can be calculated with: Price_Charged – Cost_of_Trip.

Data was provided by Data Glacier. No authorization was required
The project involved investigating the datasets through exploratory data analysis, and recommending a company to invest in (based on the results)