DS311
Group Project Documentation
Online Cars Sale Marketplace Data Set Analysis Plan

This dataset contains the car sale listings on a popular online auto marketplace. Students can explore the dataset to identify some interesting facts about cars sales on this online marketplace platform. The dataset has many categorical variables and some variables that need to be altered/cleaned before using them for any type of data analysis. For those that would like some practice with real-world data, here is a list of data cleaning objectives you can try to accomplish with this data:

1. Remove the symbols from the 'price' variable and convert it into an integer. Also remove any rows that have a value 'not priced'.
2. In the 'New/Used' column there are many different types of 'certified' values. Change all these to be displayed as 'Certified'.
3. Remove rows from the data where the value for 'drivetrain' is '-' . These were most likely cases where the seller did not input the drivetrain information. (You can also impute this data if you feel comfortable doing so).
4. Get creative! Explore the data and see what else you can do with it.

Here are 3 high-level questions you should answer with this dataset.

I. What is the most popular car listing on this platform?
   - Define what is popular car listing. Is it number of listings of certain type of car? Rating? Or something else?
   - How to uniquely define a type of car? By country made, brand, year, etc?

II. Is there a premium for dealers selling their cars on this platform compared to the private sellers?
   - Analyze the listing price between private sellers and dealers.
   - How to compare the listing prices apple-to-apple?

III. Where do we find the most used car or new car listing from this platform?
   - Define the geographical level in your analysis, state, city, or zipcode?
   - Visualize the date on a map.

Your team should answer the above high-level questions in the project. The second part of the project is to come up with three other high-level questions and break them down into smaller subset of questions to answer.

For each question your team is answering in this project, the answers must be supported by the data, which can be tables or graphs. However, you should be aware that you never have a chance in the real-world presentation to flash out 15 plots and 20 tables because no stakeholders will wait for your explanations to all the plots and tables. Be wise on selecting the right presenting material and make sure telling a story from each of them.