CS 210 Project 1 (due April **7th - 11:59 pm**)

Analyzing Used Cars dataset
Using Pandas, Matplotlib, and Seaborn.

—

Overview

Here, we will use Python libraries to read and analyze a dataset for used cars. The dataset is fairly large (170,000 lines), so make sure you are not using your mobile data when you are downloading it :). At the same time, it's not really that large, as far as the datasets go – I resampled the original set, so this one is only about 10 Mb.

The dataset contains information about used car prices for several years (it goes up to 2017). Your tasks are concerned with some exploratory analysis and visualizations of this dataset.

You will have to work with documentation heavily, especially for visualization tasks. This is a normal workflow experience, it's practically impossible to remember all the options offered by different libraries, so don't be afraid to delve into documentation/Google search. Still feel free to ask questions, of course.

We are also introducing a **Seaborn library** -- a statistical visualization package, that wraps around `matplotlib` and packs many sophisticated visualization tools into simple commands. We'll discuss it on Wednesday and/or Friday, but it's not crucial for this project.

Task 1.

Download the dataset from the following link (used_cars_sample_2.csv) and place it (do not rename it!) in the subfolder named **`Data`** in your working folder. In other words, the folder that you work in will have your Project notebook in. The rest of the tasks should be implemented in your notebook, which you should name: **`YourLastName_Project1.ipynb`**

Task 2.

Read in the dataset into a dataframe. Skip the lines that have errors. Display the head of the dataframe and the summarizing information about it.

Remove the rows that have no 'Year' information about the car, but don't remove the lines that only have 'Model' information missing.

Task 3.

Find out which car **`Make`** is listed the most in this dataset. What's the most common **`Year`**? (Note: Here, I mean that you should write the code that will figure it out, and actually print it . In other words, decide this "programmatically", not by simply looking at it.)

Task 4.

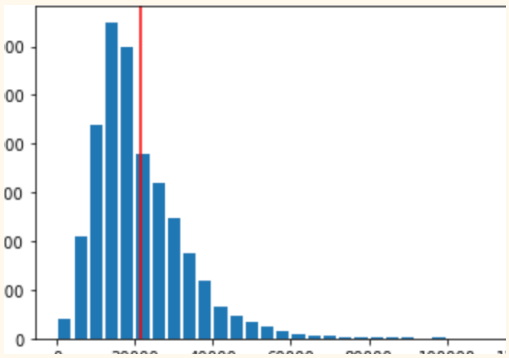
What is the average **`Price`** of all the cars listed in the dataset? Average **`Mileage`**?

Task 5.

Construct histograms for the **`Price`** distribution (at least 20 bins). Use matplotlib histogram here, not seaborn's. Place a vertical red line on top of that histogram to indicate the average price.

Do the same for **`Year`** and **`Mileage`** (histograms with vertical lines for average).
(Note: You might notice that the histogram for Price and Mileage are not very "interesting". To make it more interesting try filtering out "outliers")

Your plots in 5 should look something like that (not necessarily exactly like that!):



Task 6.

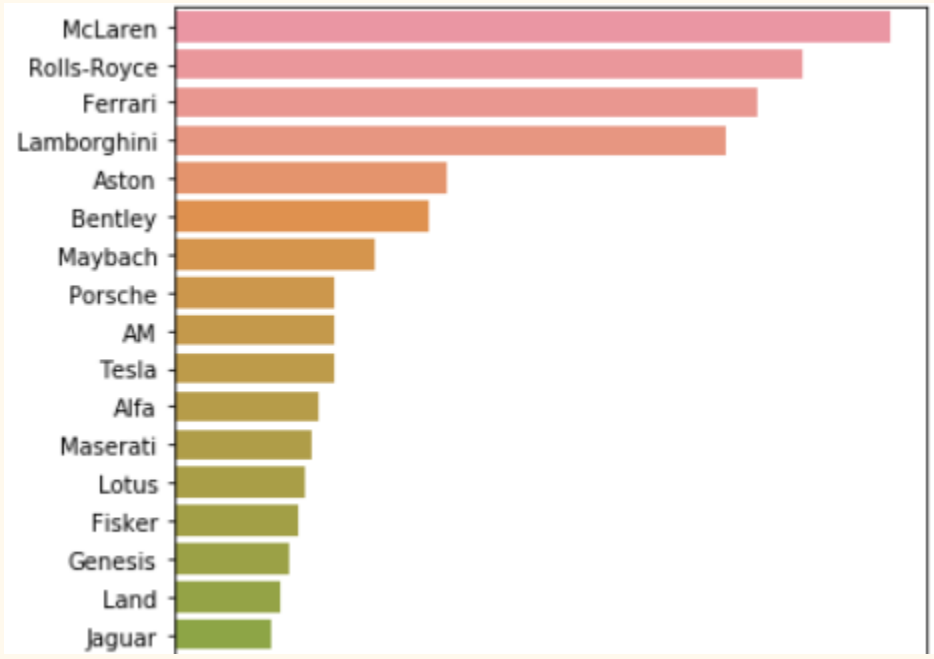
For each car **`Make`** find the average **`Price`**, and average **`Mileage`** and display this information.

Task 7.

Construct a barplot with bars reflecting the information in the previous task. I.e., each bar should have a name of the **`Make`**, and the height (or length) of the bar should be reflecting the average **`Price`** for that model. For better understanding, sort these values in descending order.

First, do this using `matplotlib` alone, then do the same using `seaborn` library.

Your plot should look something like that (not necessarily exactly like that!):



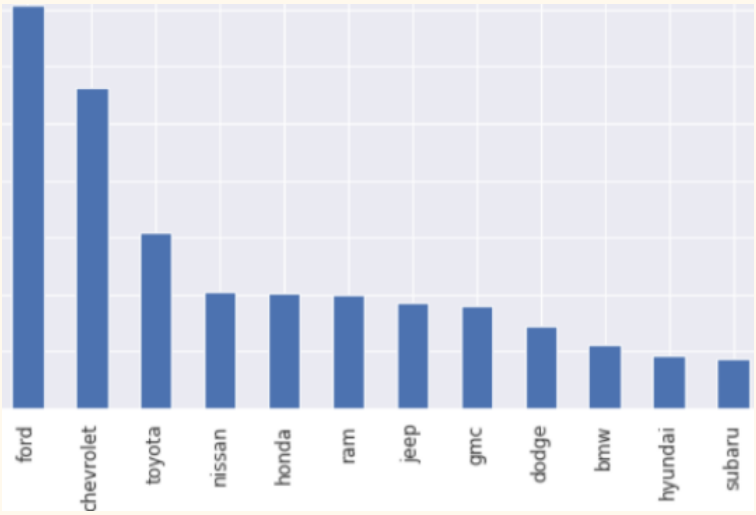
(Note: Make sure though, that the figure is large enough to fit the bars and all the labels are clearly visible.)

You should then have a similar barplot for average **`Mileage`** for each **`Make`**.

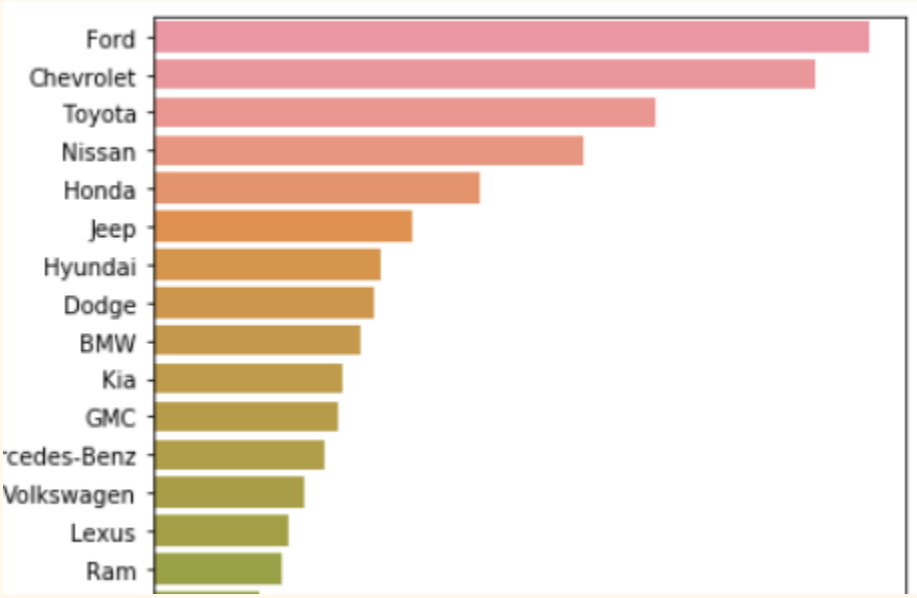
Task 8.

Construct a similar barplot as above, except instead of an average price for each **`Make`**, use the count (how many are listed). Make sure that you also have the numbers reflected on the corresponding axis.

It should look something like that (but with numbers):



Or like that:

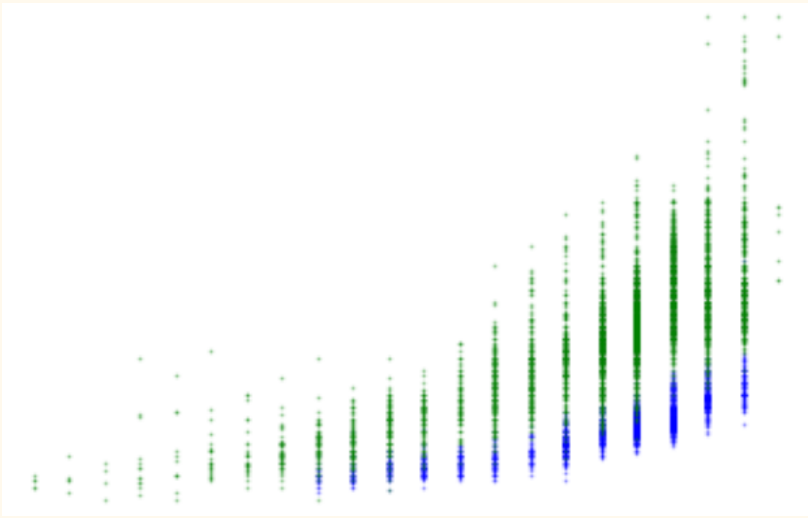


Task 9.

Use `matplotlib` to construct a scatterplot of **`Years`** vs **`Price`** for all Honda Civics. Within the same axes also construct a similar scatterplot for Ford F-150 trucks, using different color markers. Notice here, that you should count all Fords that have 'F-150' within the **`Model`** name, not just those that are exactly 'F-150'. For example, 'F-1504WD' should be counted.

Remove the obvious outliers.

Your plot should look something like the picture below. Additionally, of course include numbers on the axes, axes names, legends and a title.

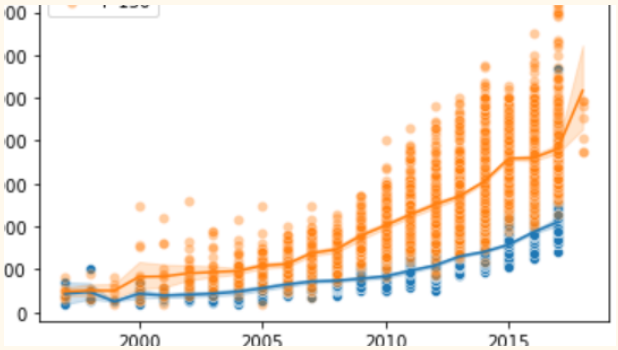


After, you are done, see if you can do this easier using the seaborn package. That means, that you are required to also do it using seaborn, and then observe whether it was easier or not

Task 10.

Add lines connecting average prices for that year for both Civics and F-150s, on top of the previous plot. (So two "average" lines, each point corresponding to a given year and representing the average price of all vehicles under that Make/Model for the given year).

Something like that (but not necessarily exactly like that!):

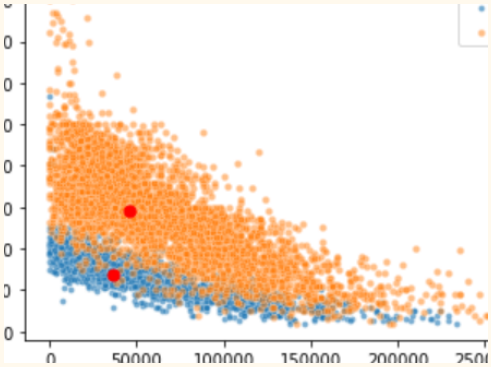


Use either matplotlib or seaborn plots here, whichever you think works better.

Task 11.

Do a similar scatter plot as in Task 9, but this time it should be **`Mileage`** vs **`Price`**, again for Civics and F-150s. Then, add a thick red dot at (median Mileage, median Price) for each of the two models. Use either matplotlib or seaborn plots here, whichever you think works better.

Something like that (but not necessarily exactly like that!):



Task 12.

Investigate if there is any truth to the belief that the cars in the northern states "get older" faster than the cars in the south. The "theory" is that because of the salt on the roads in winter, the cars in the north get rusty and lose their value faster. Check if there is any truth to it.

Don't necessarily run any fancy statistics. Instead, produce a visualization (that does rely on some simple stats of course) supporting your conclusion, whatever it is.

Suggestion: pick several models, pick a set of "northern states" and a set of "southern states", and compare the difference in prices for older models,

Task 14.

Pick some fancy looking visualization from the Seaborn library gallery and apply it to this dataset. Make sure, it makes some sense, and explain it. (Note: It's more or less up to you, what a 'fancy' looking graph is, as long as it's not a simple scatterplot or smth similar. Think: "it should have multiple operations rolled in one". Use seaborn gallery as a quide: https://seaborn.pydata.org/examples/)

Seaborn library: To complete some of the tasks, you would need to install a **`seaborn`** library, if you don't have it installed already. You can easily do it via Anaconda (go to 'Environments', type 'seaborn' to search 'All' packages, and if it is not installed, check the checkbox, and click 'Apply'). Or you can simply type `!pip install seaborn` in your notebook cell and run it.

Email me your project (as an attached Jupyter Notebook ipynb file). Make the subject of the e-mail exactly **CS210 Project1**

ADDITIONAL INFORMATION:

Some additional information may be added here, if needed. Check frequently.

1. ...