Math 475/575
Project #1:

Overall Description: In this Project, you will create a Jupyter notebook that illustrates some of the skills we have learned so far in this course. You will choose one of the following data sets:

- Miles-Per-Gallon (`https://archive.ics.uci.edu/ml/datasets/Auto+MPG`)
- Cuff-less Blood Pressure Prediction (`https://archive.ics.uci.edu/ml/datasets/Cuff-Less+Blood+Pressure+Estimation`)
- Online News Popularity (LARGE) (`https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity`)
- Residential Building Projects (`https://archive.ics.uci.edu/ml/datasets/Residential+Building+Data+Set`)
- A data set that you are interested in; please e-mail Dr. Penland to discuss by end of day Monday, October 12.

Point Value: In total, this assignment is worth 40 points (out of 500) towards your final grade. There are bonus points which make the assignment worth a total of 52 points.

General Rules on The Assignment

1. **Format:** Use the `M475-Project-1-Starter.ipynb` file as a baseline for starting your notebook. **Work** to create a **coherent report** that fulfills the requirements below. **Do not just check off the items with chunks of code.** (It's fine if you did this on past assignments, but not here.)
2. **Due Date:** This assignment should be turned in by the *end of the day* (i.e. midnight) on **Friday, October 16**. If you realize that you will not meet this deadline, you must e-mail Dr. Penland with a proposed extension utilizing the proposed number of your 5 "day-passes" for the semester. **If neither a submission nor an extension request is received by Friday, October 9, you will receive a "0" for the assignment.**

   (a) **General Statement on Academic Integrity**
   While individual competence is important, so is the ability to gather information and work as part of a team. The rules below are intended to represent this reality. **They are not "anything goes"** – notice that there are still clearly defined instances where Academic Dishonesty can occur. **If you are uncertain whether or not something is allowed, it is better to e-mail Dr. Penland and ask first.**

      i. **Rules on Collaboration**: You may work with up **two** co-authors on this project. Teams of co-authors will be **Please respond to the appropriate assignment with the name of your co-author no later than the end of the day on Friday, October 2.** Blackboard groups will be created for groups that indicate they wish to work together. Both people must be indicate co-authorship status for it to be accepted.

- **Example #1:** Alice and Danielle want to work together. Each one responds to the Coauthor Activity indicating this. They will be placed in a single group, submit a single file, and receive the same grade.
- **Example #2:** Bob really wants to work with Charlie, but Charlie would prefer to work by himself. Bob responds to the activity and indicates Charlie is his co-author. Charlie indicates he is working by himself. Bob and Charlie will not be co-authors; each one will work by himself.
- **Example #3:** Ryan, Samantha, and Timothy are good friends who often do projects together. Ryan lists Samantha as his co-author, Samantha lists Timothy as her co-author, and Timothy lists Ryan as his co-author. **This is invalid; each one will end up working by themself.**
- **Example #4:** Marilyn, James, and Elvis have a strange social dynamic. James indicates that he is working with Marilyn and Elvis. Marilyn and Elvis each indicate that they are working with each other. James will be a group of one; Marilyn and Elvis will be a group of two.

These examples should give you the idea to generalize how other situations will proceed. **If uncertain, please e-mail Dr. Penland.**

ii. **Rules On Communication:** You may discuss this project with anyone you wish, whether or not they are a co-author and whether or not they are in this class. Any such discussions must be documented in the notebook. You **may not** share code with people who are not listed as co-authors. **Sharing code in this way is Academic Dishonesty, whether or not the sharing is acknowledged.**
You **may** always communicate with Dr. Penland, but these should be acknowledged as well.

- **Example #1:** Bob is stuck on how to change the color in a scatterplot using `matplotlib`. He asks Charlie. Charlie sends him a link to a relevant section of the `matplotlib` documentation. This is acceptable, **as long as Bob acknowledges Charlie's help in the appropriate section on the Jupyter notebook.**
- **Example #2:** Bob is stuck on how to change the color in a scatterplot using `matplotlib`. He asks Charlie. Charlie refers Bob to the portion in the Blackboard lecture videos where this was discussed. This is acceptable, **as long as Bob acknowledges Charlie's help in the appropriate section on the Jupyter notebook.**
- **Example #3:** Bob is stuck on how to change the color in a scatterplot using `matplotlib`. He asks Charlie. Charlie writes an example, screenshots it, and sends it to Bob. **This is unacceptable and constitutes Academic Dishonesty.**
- **Example #4:** Bob is stuck on how to change the color in a scatterplot using `matplotlib`. He asks Charlie. Charlie sends a Jupyter notebook to Bob that includes this part of code. **This is unacceptable and constitutes Academic Dishonesty, even if Bob says "I got the code for how to change color in a scatterplot from Charlie".**

iii. **Rules On References:** You may utilize any coding resources you wish, except for code that has been created by other students in the class who are not your co-authors. **Any reference** from which you use information must be cited using

the appropriate format. **Any code that is utilized from a reference must be cleared for the appropriate license, and acknowledgment must be given in the notebook close to the place where the code is listed.** You will probably need to modify code examples that you reference; the reference should still be included.

Rubric: The points will be assigned based on the following criteria.

- **(0 points total on the assignment; all other items irrelevant.)** on the assignment for any Academic Dishonesty violations as outlined above. **This includes no information given for acknowledgments or resources.**
- **(0 points total on the assignment; all other items irrelevant)** if you include a code cell which does not work.
- **(BONUS: +3 points)** if you complete the Project by yourself.
- **(BONUS: +3 points)** if you complete the Project with a co-author whom you did not know before October 1, 2020.
- **(BONUS: +3 points)** if you use the **LARGE** data sets listed above.
- **(BONUS: +3 points)** if you achieve a model with a score above 0.8 on the test data set from your chosen data.
- **(BONUS: +3 points)** if you ask a question that helps Dr. Penland improve the clarity of the instructions, or improve future projects in some way.
- An Acknowledgments section is completed that refers to any communications with anyone who was not a co-author (including Dr. Penland).
- **(+3 points)** The document is organized into sections that perform the appropriate tasks.
- **(+3 points)** The document has a table of contents with hyperlinks to the relevant sections.
- **(+2 points)** The choice of data set is explained.
- **(+2 points)** The modeling problem is well-explained.
- **(+2 points)** The data is read appropriately.
- **(+3 points)** At least one appropriate plot is created to understand an aspect of the data.
- **(+3 points)** The information from the plot is summarized in a sentence.
- **(+2 points)** The data is split into training and testing set with test set size equal to 0.3.
- **(+3 points)** An sklearn model is trained on the training set. The data should be preprocessed appropriately (dealing with missing values, categorical variables, and scaled if needed). Feature selection may be included, but this is optional.
- **(+3 points)** The model is scored on both the training set and the testing set using an appropriate command from sklearn.
- **(+3 points)** At least one residual plot is created and analyzed.
- **(+3 points)** The performance of the model is analyzed based on the previous item two items. Does the model appear to be underfit/overfit? Is there any information that is left out?

- **(+7 points)** An alternative model is trained and analyzed according the steps outlined above. **This model should include some degree of feature selection, whether automated (using a library) or chosen based on human interpretation.**
- **(+4 points)** A coherent explanation of which model, if either, you would recommend for implementation. You do not need to recommend a model. If you do not recommend a model, propose the next model that you would try.
- **(+3 points)** A reflection on three specific things that you have learned or strengthened through the process of completing this project.
- **(+3 points)** Three questions that you might ask as a result of this experience. These include things you want to know, as well questions that you have answered but might be helpful to other students. They can be procedural (e.g. "How do we change the color of the points in matplotlib?") or conceptual (e.g. "How can we recognize that Decision Trees are likely to work well for a data set?") (**Please do not use either of these questions.**
