DSA 610, Homework #6, Spring 2025

Part I:
Instructions: Answer each discussion question in your own words. You may use posted resources or other online resources to answer the questions (**cite your sources**). Thoroughly explain your responses with a minimum of one paragraph (3-5 sentences) in length. Be thoughtful. We will discuss the answers in class.

1. Why would an organization want to reuse their own data? Why might they want to reuse data from another source?

2. What are some challenges associated with using data you did not collect yourself?

3. What are some sources of data that can be reused? List at least one private source and one publicly available source. What are the advantages and disadvantages of each.

4. What are some common model types and when would you use each? List at least three.

5. Of all the model types and tools in last week's lecture, choose one that you are unfamiliar with (or less familiar with). Investigate this model or tool and summarize it's key elements and describe a potential use case.

Part II:
Instructions: Use the attached dataset (**housing_hw6.xlsx** and **Ex5ID10000291_hw6.xlsx**) to complete the following tasks in Python. Report your answers to the questions on this homework sheet. Include your Jupyter notebook file along with your homework submission, saved as a PDF. Make sure that any graphs you create are quality graphs with a legend (as needed), axis titles, a descriptive title, appropriate ranges (bar graphs start at 0, etc.). They should be able to stand on their own. You may need to relabel some elements (such as replacing 0s and 1s with string labels).

Use Python to answer the following:

1. Create a scatterplot of the data in the Ex5 dataset. Use the same variables we did in the lecture example: Date and delta(H).

2. Construct and compare the following non-linear regression models:
   a. A polynomial model (you may need to experiment with degrees, but no more than degree 5)
   b. A gaussian process model (with noise)
   c. A spline model
   d. Optional: one model we did not set up in class is a LOESS model (which is popular model in ggplot, sometimes called LOWESS). You can find a statsmodels implementation here: https://www.statsmodels.org/dev/generated/statsmodels.nonparametric.smoothers_lowess.lowess.html or a special package implementation here: https://pypi.org/project/loess/
   e. Compute some metrics to compare and plot them on a graph with the data.

3. Switching to the housing data, construct a pair plot and correlation table to determine which variables have the highest correlation with selling price (omit the House variable).

4. Construct a one-variable linear regression model with the variable you identified in the previous question to predict selling price. Calculate appropriate metrics and evaluate the model. (Note: given the size of the dataset, do not do a test-train split here or in any of the remaining questions.)

5. Construct a multiple linear regression model to predict selling price from the other variables. Compare these results to a LASSO model and a Ridge model, and one other model of your choice (KNN, Random Forest, etc.) Compare the metrics for the models.

6. LASSO and Ridge regression are penalized regression models that depend on the size of the coefficients in the model. This can create sub-optimal results if the variables are not rescaled. Rescale the variables and then rerun these models (and KNN which can also be sensitive to scaling). Compare the results of your scaled versus unscaled results. Did rescaling improve model performance?