**Group Problem Set – Simple Linear Regression**

The data set **PIRU\_BAI.xls** includes measurements of average annual basal area increment growth from red spruce plots across the northeastern US. We’ve gathered satellite imagery over the region and calculated a normalized difference water index (NDWI) for pixels at each plot location. We want to know if we can use this spectral vegetative water index to “predict” red spruce growth (BAI) without having to go out and core trees. This would allow us to see how red spruce growth varies across the landscape (beyond an assessment of just the trees we cored) and over time.

Your task is to create a simple linear regression model that can be applied to satellite imagery across the landscape to assess red spruce growth (BAI).

Use the following questions to walk you through a model development and evaluation. Note that we will be asking you to paste individual lines of code in addition to the outputs you get in this document.

1. Start by creating a scatterplot to visualize the relationship between Min\_NDWI and BAI.

* **Code:**
* **Figure:**

1. Build a linear regression model to predict BAI using NDWI as the independent variable. Summarize the output model so you can see the evaluation metrics.

* **Code:**
* **Equation (Y = *B1*+ *B2*X):**
* **r2:**
* **RMSE:**
* **p-value :**

1. Now let’s test the assumptions for the SLR:
   * Add the SLR line to your scatterplot for visualization of potential outliers.
   * Calculate and plot the residuals vs. predicted for all observations.
   * Test the residuals for normality.

* **Code:**
* **Figures (Line and Scatter, Residual vs Predicted):**
* **Based on these diagnostics, do you meet the assumptions for a SLR (justify your response for each:**
  + **Independent Samples?**
  + **Linear relationship?**
  + **Homoscedastic residuals?**
  + **Normally distributed residuals?**

1. Now let’s run some additional diagnostics. Calculate and plot the Cook’s D influence for all observations.
   * **Code:**
   * **Figure (Cook’s D):**
   * **Based on analysis of your figures and residuals, do you see any obvious outliers in your data? Justify your response.**
   * **Based on your Cook’s D do you see any obvious high leverage points that should be removed? Justify your response.**
2. Assume that any high leverage points are important data points to keep in….so you model stays the same as you just described. Evaluate your final model following the same steps you did in #2, with the addition of calculating the PRESS RMSE. Based on these evaluation metrics, answer the following:
   * **Is there a significant relationship between your vegetation index and BAI? Report the p-value and justify your response.**
   * **Is this an accurate model? Report the RMSE and percent error to justify your response.**
   * **Is this a “meaningful model”? Report the r2 and justify your response.**
   * **Is this model stable? Report the PRESS RMSE : RMSE to justify your response.**
   * **Over what range of values could this model be used? Report your working range.**
3. Do you think this model is useful to assess red spruce growth from satellite imagery across the study area? Justify your response.