# CSCE/STAT 587 Clustering Homework
# Due: Tuesday, September 28

**Background:** The Abalone data set consists of 4177 rows of data in which each row corresponds to one of 4177 observations/instances. Each row consists of 9 columns. We will not be using columns 1 and 9 for this assignment.
You will cluster the observations using columns 2 through 8. These columns contain measurements of:

Length / continuous / mm / Longest shell measurement
Diameter / continuous / mm / perpendicular to length
Height / continuous / mm / with meat in shell
Whole weight / continuous / grams / whole abalone
Shucked weight / continuous / grams / weight of meat
Viscera weight / continuous / grams / gut weight (after bleeding)
Shell weight / continuous / grams / after being dried

Columns 1 and 9, which we will not use are:

Sex / nominal / -- / M, F, and I (infant)
Rings / integer / -- / +1.5 gives the age in years

**Step 0:** Review the material from the in-class lab we did on K-means clustering.

**Step 1:** Download the dataset "abalone.csv" from https://cse.sc.edu/~rose/587/CSV/abalone.csv using **wget**. Load this data set into rstudio using the "import data" button in the environment tab. Be sure to select "From Text (base)…" in the pulldown menu.

**Step 2:** Create a new data frame called "AbaloneFeatures" containing columns 2 through 8. Recall that we are NOT using columns 1 and 9 for this assignment

**Note:** You will be setting the random number generator seed to 555 using the command: set.seed(555). **Do this before each call to the kmeans() function so that the Grader can easily tell if your results are correct. We are doing this to avoid differences that might arise from different initial cluster seeds.**

**Step 3:** We want to explore different numbers of clusters in order to select a good value for K. We will cluster using all 7 numeric features in the AbaloneFeatures data frame. As in class, calculate the within-sum-of-squares values for k=1 to 20. Also, set the max number of iterations to be 15. The default is 10, but we want to allow at least 15.
Note: *Since we are using a for-loop to do this, be sure to set the random number generator seed* ***before*** *each call to kmeans().* Hint: the set.seed() call MUST be in the body of the loop and precede the call to kmeans().
Plot these sum-of-squares values. **Save the plot to a pdf file (use 8 inch by 8 inch canvas).**

**Step 4:** From step 3, above, it is clear that K=1 is not a good number of clusters? Choose the first ***reasonable*** K based on the results from step 3. (A ***reasonable*** K should be on the elbow.) Use the kmeans() function with this number for K. Plot the results such that *each cluster is plotted in a different color* (as we did in class). Observation: Since we are clustering on 7 features, R will create 2*choose(7,2) = 42 panels in your plot, i.e., all combinations of 2D projections. Note: you do not need to plot the cluster means, just the clusters themselves. **Save the plot to a pdf file (use 16 inch by 16 inch canvas). Note: this is a larger canvas than that used for step 3.**

**Step 5:** Normalize the data set using the normalization functions that we created during the K-means lab. Save the normalized data in a new data frame called “NormalizedAbaloneFeatures” Use the summary() to verify that the features range from 0 to 1.

**Step 6:** repeat steps 3 using the normalized data.

**Step 7:** repeat step 4 using the normalized data.

**Step 8. Compare plots from steps 3 and 6**. Did the plots change enough to cause you to select different values for k? Why or why not?

**Submit your plots from steps 3, 4, 6, and 7. Be sure to also document and submit your R-code for steps 2, 3, 4, 5, 6, and 7 in the form of a .R or .txt file. By document, I want you to at least label what step each group of commands or code correspond to. Finally, do not forgot to submit your analysis/explanation from step 8.**