MATH216 HW 1
Due: Friday, September 22nd at 11:59pm
Please submit an electronic copy of your assignment to: *alyford@middlebury.edu* using the subject line: MATH216 - HW1

This homework will use the *Plants.csv* data set attached to this email. These data are a subset of data collected from an experimental study conducted on three species of Hydrangeas, a flowering plant native to much of America. The hydrangeas were planted in a 10x10 grid, and no two hydrangeas were planted in the same plot. There are six variables of interest in this data set.

1) **BiomassT1**: A measure of the hydrangea's biomass (the total mass of the hydrangea) in grams at the first measuring time halfway through the experiment.
2) **BiomassT2**: A measure of the hydrangea's biomass in grams at the final measuring time at the end of experiment.
3) **Row**: A number indicating the row in which the hydrangea was planted.
4) **Column**: A number indicating the column in which the hydrangea was planted.
5) **Fert**: The brand of fertilizer used. There are three different brands of fertilizer, (for purposes of anonymity, here called A, B, and C).
6) **Species**: This refers to the species of hydrangeas used in this experiment: species 1, 2, and 3.

Based on the data provided, answer the following questions.

(1) This experiment was conducted by biologists, not statisticians. As such (no offense), you should first check to see if the design seems reasonable. Although we have not formally defined a reasonable experimental design, examples of an unreasonable design would include only using fertilizer A on hydrangea species 1 (and not any of the other fertilizers) or only placing species 1 in column 1 (and not any of the other columns). Consider using the table() function in R to check this. This function creates a contingency table of two (or more) categorical variables. For example, table(x,y) would create a table of the number of times each level of x occurs with each level of y. In your opinion, based on the data, does the experimental design presented here seem reasonable? Why or why not?

(2) Next, determine if there is any difference in final biomass (*BiomassT2*) between the species. Construct a graph showing boxplots of the final biomasses for each species. Based solely on the graph, does there appear to be a difference between the species? If so, which ones?

(3) Support your answer in (2) by comparing the mean final biomass of each species. Do these numbers appear close?

(4) Perform a similar analysis for fertilizer type. Does there appear to be a difference in final biomass for different fertilizer brands? If so, which ones, and calculate means to support your answer.

(5) Now produce a graph that looks at final biomass (quantitative), fertilizer brand (qualitative), and species (qualitative) simultaneously. Comment about what this graph tells you about the relationship between the three aforementioned variables.

6) Thus far, you have only used final biomass as the quantitative variable of interest. Construct a scatter plot of BiomassT1 versus BiomassT2 and comment about the relationship. Describe what you see.

7) Now color the dots for the plot you made in (6) by the fertilizer brand used, and set the shape of the points to be equal to the species type. Does this graph agree or disagree with your findings in (5)? Do you notice anything strange about this plot (look in the lower left-hand corner)? Describe what you see and hypothesize why this might have occurred (in the context of the data).

(8) In an attempt to discover the cause of the 'strangeness' seen in the graph in (7), determine if row position or column position had any effect on final plant biomass. Produce a graph (or graphs) to support your answer.

(9) Since no one ever reads ahead, I'm sure you didn't read this and learn that you can construct heatmaps in ggplot! (If you did read ahead, please answer number 6 without making a heatmap). Use the geom_tile() function to construct a heatmap of final plant biomass as a function of row position and column position

Make a guess about what happened to create the pattern you see here.

(10) Create a new data set that removes the 'problem' points, whichever you deem those to be. Recreate all of the graphs in (1) to (9) using this new data set (you do not need to change your written responses, only the graphs). Comment about the new graphs and note if anything has changed. *Hint: Using R Markdown, this can be done in one line! However, there are many ways to achieve this without any copy/paste or rewriting of code.*

(11) Suppose the researchers who conducted the study believe that the best indicator for the hydrangea's growth is the average of the two biomass measurements. Create a new variable that averages the biomass measured at time 1 with the biomass measured at time 2, and use this variable and accompanying graphs to answer the following question: Do the effects of fertilizer, species, and/or positioning change if the average biomass is used instead of final biomass?

(12) Finally, based on your analysis of these data, tell the researchers what you have learned about how to produce hydrangeas with the largest biomass. Be sure to address if this advice depends on the species.