# Problem Set 4

Due April 2, 10:00 AM (Before Class)

## Instructions

1. The following questions should each be answered within an R script. Be sure to provide many comments in the script to facilitate grading. Undocumented code will not be graded.
2. Work on git. Fork the repository found at https://github.com/domlockett/PDS-PS3 and add your code, committing and pushing frequently. Use meaningful commit messages – these may affect your grade.
3. You may work in teams, but each student should develop their own R script. To be clear, there should be no copy and paste. Each keystroke in the assignment should be your own.
4. If you have any questions regarding the Problem Set, contact the TAs or use their office hours.
5. For students new to programming, this may take a while. Get started.
6. You will need to install `ggplot2` and `dplyr` to complete this dataset.

## Sample Statistics

1. Load the following data: http://politicaldatascience.com/PDS/Datasets/GSS-data.csv. \

   The variable `poleff11` asks participants to rate their level of agreement with the statement "People like me don't have any say about what the government does" (see the codebook for more information on all variables in this dataset at: http://politicaldatascience.com/PDS/Datasets/gss_codebook.csv).

   - Convert this variable into a numeric where higher values indicate higher levels of political efficacy (1- strongly agrees with the statement; 5- strongly disagrees with the statment) and all other values ('Cant choose' etc.) become NA's.
   - What is the proportion of individuals from the entire sample who do not feel as though they have a say in the government? The proportion of those who do feel as though they have a say?
   - Using a sample of **25** what is the average level of efficacy individuals feel on the one to five scale? At 100? 500?
   - Pull a random sample of 25 from the `poleff11` data and calculate the mean. Now repeat this process 500 times and store these values in a variable called `trials_25`.
   - Now create a variable called `trials_100` where we do 500 trials with n=100 instead of 25.
   - Draw a histogram of the sampling distribution for the two trials you just conducted. Give the plots meaningful titles and axis labels. Save these plots in your repository.
   - What notable difference occur when we use a larger sample size in our trials?

## Bootstrapping

2. **Run the command `set.seed(25)`.** Create a new variable `eff_subset` which draws a sample of 300 observations from the original `poleff11` data.

   - Draw a sample of 25 from `eff_subset` and calculate the mean. Repeat this task 500 times and store the results in a variable named `trials`.
   - What is the mean of the 500 trials? 1000 trials? How do these values compare to the mean of the actual `poleff11` variable?

- Notice that with pulling random samples of only 25 observation **from a fraction of the data** we are able to approximate the mean of the entire data quite well. That is the logic of bootstrapping, with only a small sample (here our population is the entire dataset), one can make reasonable approximations about the population!

## Supervised Learning

1. -
   -