Homework 9
SDS348/385
Due Tuesday April 11th, 2017 by 7:00pm

For this homework, you will write a proposal for Project 3. On Project 3, you will analyze a data set of your choosing with the specific goal of answering **two** questions about the data set. You should address each question computationally, using R and/or python, and produce a plot for each question illustrating the trend of interest. You will then answer your questions and interpret the plot, in the context of the questions your proposed.

Graduate students will have to write and submit their own proposals. Undergraduate students may either work in pairs or work alone for this project. If you will work in pairs, both partners may submit the same proposal, with both names at the top.

Your proposal should contain the following:

1. Identify and describe a data set you will analyze. Describe either how you intend to collect and/or generate the data, or alternatively if you intend to use an existing data set, give the source of the data. (1-3 sentences)

   If you are having difficulty finding a data set, the following list of built-in R data sets is a good place to start:

   https://vincentarelbundock.github.io/Rdatasets/datasets.html

2. Propose two questions about this data set that you will ask and address in your project. These questions should be of similar scope to questions you addressed on previous projects (nothing too complicated!).

   As a reminder, your questions should be **conceptual** and not procedural. (For example, "What is the distribution of ages?" is a procedural question because all it asks you to do is plot a distribution, but, "Is incidence of diabetes higher in older women or in younger women?" is a **conceptual** question because you have to determine which type of plot is appropriate for the question and interpret that plot.)

3. Explain how you will analyze the data in order to address your proposed questions. You should indicate how you will answer each question, including which computer language(s) (e.g. R or Python) you will use, which methods you will use, what your plot for each question will be, and how you will create the plot (e.g. using ggplot2). (~ 2-4 sentences per question)