BSDS 100: Intro to Data Science with R
Final Class Project

Objective

Choose one of the projects below. Each of the projects are computational tasks that will require tools from what you've learned in this course. The aim of this final project is to give an opportunity for you to use and showcase all of the useful computational tools in R you've learned this semester. A major component of this project is to make any written code efficient, well-documented, and easy to use. Also any plots or output should be crisp, easily understood, and properly labeled.

You can use online resources for guidance but do not simply copy and paste some one else's code.

Grading Rubric

You will work in groups of 4. Each group will be required to turn in the following:

- A presentation deck: each group will present a 15 minute presentation on Thursday, November 30th or Tuesday, December 5th in class. You **must** put the final presentation in either Powerpoint or Latex beamer. Include the following in your presentation:
  - Description of the problem
  - A brief introductory description of any data that you use
  - An analysis of your data including any decisions you made along the way. Creativity counts here!
- A well-organized and comprehensive report consisting of a written analysis of your findings and the R code you used, knit into a .pdf file.
- Usual grading standards for HWs and Case Studies apply.

Due Dates

1. Enter your names, project choices (1, 2, 3, or 4), and your preferred date (not guarenteed) in the following Google Sheet

   Final Project Signup

   by **Tuesday, October 31st by 9:00 AM** .

2. Presentations are to be submitted on Canvas by **Thursday, November 30th by 9:00 AM**. Reports consisting of your code and written analysis will be due on **Sunday, December 10th at 11:59 PM**.

Project Choices

1. [**Case Study**] Choose a data set you are interested in and conduct a full analysis using the data science pipeline, applying the techniques in `R` you've learned in class. If you are having trouble finding a data set, you may use data from repositories such as `http://archive.ics.uci.edu/ml/`. Be creative! **Don't** use pre-loaded data in `R`.

   Discuss the challenges in the problem and the data set, and how you circumvented these problems. Consider issues of, for example, sparsity/missing values in the features and response, high numbers of features/dimensions, and the scalability issues of big data. For any problem, apply any method that you see fit, discuss the advantages and disadvantages of each method, and explain why you found this approach appropriate. Thoroughly explore and assess any inference that you make on the data and what lead to your analysis. In your presentation, explain the data, why it interested you, and your step-by-step analyses that lead to any final conclusions.

2. [**Exploring Machine Learning Topics**] Machine learning is an expansive field in data science with many topics that we have not yet covered in class. In this project, you will create a 15 minute lecture and demonstration of one of the following popular areas in machine learning to present to your classmates:

   - Clustering
   - Classification
   - Regularized Regression
   - Neural Networks
   - Image Segmentation
   - Natural language processing
   - Deep Learning

   The goal of this project is to research the topic chosen from above, present and implement at least one method in this area. Describe the topic, any challenges inherent in the area, and relationships with other topics that we have covered in the class. Apply your chosen method(s) to a data set of your choice, including any analyses and data-driven decisions from machine learning that can help you analyze the data. Remember you are the instructor of this topic, so present in a way that you wish someone would have taught you.