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

by James D. Wilson (University of San Francisco)

Objective

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

You can use *some* code from online resources but do not simply copy and paste some one else's code for the entire product. 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.

What to Turn In

You will work in groups of up to 4. Choose one of the projects below. Each group will be required to turn in the following:

- A write-up that is done in an R markdown file (that has been knit and saved as a .pdf) that answers all the questions in the "Components" section below. You must print any plots and show the code for anything necessary to answer the questions.
- A 15 minute presentation (cannot go over 15 minutes!), which will be presented on either *Monday, December 5th or Wednesday, December 7th.* In your presentation, include a summary / discussion of everything in the "Components" section below. This does not mean to show all plots, rather, give a presentation that conveys the message and the details of what you would present in a business meeting. Imagine you have one shot to speak in front of a CEO of a company that wants to know what you've been working on.

Due Dates

- The presentation (a .ppt or .pdf submitted to Canvas) is due **Monday, December 5th at 9:00 AM**.
- The write-up (a knit .pdf file submitted to Canvas) is due on the last day of the final exams on **Thursday, December 15th at 12:00 PM**.

Components of the write-up and presentation

- Statement of question/topic that you want to answer and what motivated you to study the question/topic.
- If applicable, what data will you analyze? Identify at least one data source (can be from Kaggle, etc.)
- What challenges do you face in analyzing this data?
- What packages were needed for this case study?
- Discussion: What did you learn from this experience? What more could you do with this project in the future?

Grading

Grading will be done as follows. This is based on a 100 point maximum.

- The presentation is worth 30 points and the write-up is worth 30 points. For this, each of the components to be included in the write-up and presentation is worth 6 points a piece.
- Having an R markdown that has knitted - worth 15 points
- Having a clear, concise report and presentation. For example, no extraneous plots, no going over 15 minutes for a presentation. - worth 15 points.
- Being present for all presentations and contributing to the presentation for your group - worth 10 points.

Project Choices

1. [**Additional ML Topics**] Machine learning is an expansive field with many topics that we have not yet covered in class (though many more topics will be discussed next module). In this project, you will first choose one of the following popular areas in machine learning:
    - Clustering
    - Classification
    - Regression
    - Natural language processing
    - Deep Learning
    - Image Segmentation
    - Neural Networks
    - Semi-supervised 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.

2. [**Case Study**] Choose your favorite data set or one from a Kaggle competition at Kaggle.com to which you can apply computation techniques in R described in class. Discuss the challenges in the problem and the data set, and how you circumvented these problems. Consider issues of, for example, sparsity in the features and response, high dimensions, and the scalability issues of BIG data. For any problem, apply any method that you see appropriate and discuss the advantages and disadvantages of each method and why you found them 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.

3. [**R Shiny Application**] One way to provide a user-friendly environment to apply R code and any other coded functions is to create a graphical user interface with R Shiny. In this assignment, create a aesthetically pleasing application that performs a task of your choice. Your interface must contain the following components:

    - Data input and output
    - A Function that was written by you (that contains at least 20 lines of code and is properly written and well-documented).
    - Visualization of the data using *ggplot2*
    - Concise summary of Results

    For the presentation, show how your app works and give an example with a data set of your choice. To get started, see https://shiny.rstudio. com and https://shiny.rstudio.com/tutorial/ for a tutorial of how to create a Shiny app.

First Step: Sign Up!

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

Final Project Signup

by **October 26th** by 9:00 AM.