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

by Abbie M. Popa (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 someone else's code for the entire product. You may use data from repositories such as `http://archive.ics.uci.edu/ml/`, kaggle `https://www.kaggle.com/datasets`, or google's dataset search tool `https://toolbox.google.com/datasetsearch`. Be creative! **Don't** use pre-loaded data in R.

Grading

You will work in groups of 3 or 4. Choose one of the projects below. Each group will be required to turn in the following:

- A knit PDF report including your code, output and paragraphs describing the following:
  - Description of the problem
  - A brief description of any data that you analyze
  - Analysis of your data including any decisions you made along the way. Creativity counts here!
  - Conclusions you can draw, particularly with regard to the original problem you presented

Rubric

Each item is worth 15 points, for a total possible score of 75 points.

- **Question or Problem**: Does the project include a thorough, accurate, specific, and clear explanation of the question or problem being asked or solved?
- **Data Display**: Does the project include appropriate, well-labelled, accurate displays (graphs and/or tables) of the data or method?
- **Data Analysis**: Does the project include appropriate, accurate, logical, and thorough analysis of the data?
- **Conclusion or Product**: If an analysis project is chosen, are the conclusions logical and well-supported? If the team builds an ML method, does the method work as stated and account for edge cases?
- **Presentation and Professionalism**: Was the project presented well, with a neat, easy to read pdf. A neat easy to read pdf will include all text reports in plain text (outside of code blocks) with headers alerting the reader to where they are in the project. Similarly, code will be readable and not run off the page (exception, if running read.csv or read.table and the address of the file is too long to fit on the page this is acceptable as there is no easy way to correct it). The paragraphs will not be graded on grammar and spelling per se, they should be thorough and readable. If their are sufficient mistakes to decrease readability, points will be deducted. Plots should also be easy to read, with labelled axes and titles, and jitter or color when it is helpful for interpretation.

Project Choices

1. **Case Study** Choose a data set 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. 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 report, explain the data, why it interested you, and your step-by-step analyses that lead to any final conclusions.
2. **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 report, give instructions for the reader to follow and that demonstrate how the app works given 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.

3. **ML Topics** Machine learning is an expansive field with many topics that we have not yet covered in class. 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, report on and implement at least one method in this area using `R`. 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 write your report in a way that you wish someone would have taught you.

Due Dates

1. You must form groups, pick your project, and select your dataset by **Thursday, November 1st**. Enter this information in the google sheet linked on Canvas.

2. Reports (including `R` code) are to be submitted on Canvas by **Friday, December 7** at 11:59 PM.