Project 3 Instructions
SDS348/385
Due Thursday May 4th, 2017 by 7:00 pm on Canvas

For 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 you proposed.

Each student must turn in his or her own project. If you are working with a partner, both you and your partner must turn in a final project, but these project documents may be the same. Be sure to include both partners' names in the final project.

You should write your project in a single knitted (and converted to PDF) RMarkdown file[1], organized into three main sections: Introduction, Question 1, and Question 2. You should incorporate all R code used into the main RMarkdown document, as done on previous projects. If you used any python code in your project, then it needs to be provided as a separate Jupyter Notebook(s), converted to PDF. Be sure to organize your python code properly so that it is clear which code pertains to which question. Also, in your document, clearly point out any 3rd-party packages or libraries that are required to run your analysis.

You must additionally turn in the raw RMarkdown file and/or Jupyter Notebook file, as well as all external data you used. We need to be able to rerun your entire analysis and reproduce your results.

Your project must fall into one of the two following categories, which have somewhat different requirements:

1. Statistics-based projects
2. Programming-based projects

1. Statistics-based projects

For statistics-based projects, you take an existing dataset and answer your questions by running a statistical analysis and make one or more plots. For such projects, each of the following criteria needs to be met:

- Part of the answer for each question has to be a statistical analysis of some sort. We accept descriptive analyses such as PCA or clustering if they provide useful insight into the question.
- At least one of the two questions needs to be answered with a multivariate analysis.
- The analyses for the two questions need to be different.
- The main plots for the two questions need to be different. E.g., one plot could be a box-plot and one a scatter plot. If you make more than one plot for each

[1] If using Python exclusively, write your project into a single Jupyter Notebook. Follow the same format as you would an RMarkdown document.

question, then it is Ok if you reuse some plot types. However, the main plot answering Q1 needs to be different from the main plot answering Q2.
- If you run multiple tests, such as multiple *t* tests among groups, you need to correct your *p* values for multiple testing, e.g. using `p.adjust()`.
- If your dataset contains obvious potentially confounding variables, then you need to either use multivariate linear models instead of univariate tests (i.e., *t* tests or correlations) or explain why you think the univariate tests are appropriate.

2. Programming-based projects

For programming-based projects, you will write computer code (in either R or python) that answers your questions by extracting relevant information from a suitable source of raw data. For each question, the final product from your code should be either a plot or a small, humanly-readable table that answers your question. For such projects, each of the following criteria needs to be met:

- There needs to be a non-trivial amount of programming. If you proposed a programming-based project in your Homework 9 and we approved it then you can assume that your project meets this criterion.
- You need to use two different programming approaches for the two questions. This will generally mean that the code for Question 1 is obviously different from the code for Question 2.
- Your final results need to be presented as a plot, unless your final data tables are very small and highly informative. (If your final tables have more than a few rows or columns then the data probably need to be plotted.)

Instructions for all projects

The Introduction should be relatively short (1-2 paragraphs) and should contain the source of and a brief description of your dataset. Make sure you describe the key features of the dataset that you use in your project, including what all the variables are and (where appropriate) in what units they are measured.

The section for each question should begin with the question you are asking, followed by the analysis. If you conducted any aspect of your analysis in python, you should indicate in which file the code is located, by directly referencing the attached Jupyter Notebook file. After your analysis has been conducted, you will display your plot(s), and discuss/interpret the results. The discussion **must** accomplish the following:

- Describe your chosen analysis and provide a brief overview of your methods (3-5 sentences)
- Explicitly justify why this analysis makes sense for the question you have asked (1-3 sentences)
- Interpret your plot(s) in the context of the question you asked (2-4 sentences)
- Answer the question you have asked

Project 3 Instructions
SDS348/385
Due Thursday May 4th, 2017 by 7:00 pm on Canvas

Please bear in mind that you will lose points for any of the following:

- No comments in your code (either R or python)
- Code which produces an error message (we need to be able to re-run your entire analysis)
- Missing code and/or reporting results without corresponding code
- Extraneous code or plots which do not contribute to your final analysis or discussion
- Not turning in your dataset
  NOTE: If your dataset is built into R or an available R package, then you do not need to turn in a file with the data. You must, however, state in your Introduction in which R package the dataset can be found.