Project 3 Instructions
SDS348/385
Due Thursday May 3th, 2018 by 7:00 pm on Canvas

For Project 3 you will ask and answer a question based on a data set of your choosing obtained from the NCBI Entrez database. This project will have two parts, one in python and one in R. The python part will use biopython to extract your dataset from the Entrez database, and the R part will analyze the data and visualize the results.

Undergraduates registered for SDS 348 can work in pairs on this project. However, 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. Graduate students registered for SDS 385 must complete the project on their own.

Your project submission must consist of 4 files:

1. Jupyter Notebook (.ipynb) file
2. Jupyter Notebook converted to pdf
3. R Markdown (.Rmd) file
4. R Markdown converted to pdf

The Jupyter Notebook file should download data from the NCBI Entrez database, do processing of the data as appropriate, and save the output as a .csv file. The R Markdown file should read in the .csv file and then perform the plotting and statistical analysis, as described below.

We must be able to reproduce your analysis by placing the .ipynb file and the .Rmd file into the same folder and then first executing all cells in the .ipynb file and then knitting the .Rmd file.

Specific instructions for Part I

Part I should start with a brief introduction providing background and motivation and then state the overarching conceptual question that you are going to answer. The introduction should also explain the data source and variables you will use to answer the question. (2 paragraphs total, plus question)

Following the introduction, your .ipynb notebook should contain code that downloads data from the NCBI, possibly does some processing of the data as appropriate (e.g., counting genes, or calculating nucleotide compositions, or identifying unique co-authors of an article author), and then writes the results into a .csv file.

Grading for this part will be based on the following criteria:

- The code must be properly commented and readable.
- The code must run and produce an appropriate output .csv file.
- Part I of the project must use the provided .ipynb template.

Project 3 Instructions
SDS348/385
Due Thursday May 3th, 2018 by 7:00 pm on Canvas

Specific instructions for Part II
In Part II, we expect **two (and only two!)** meaningful figures and **one** statistical analysis to answer the overarching question. The project should conclude with a meaningful discussion that explains how your overarching question is answered by the two figures and one statistical analysis.

Grading for this part will be based on the following criteria:

- The statistical analysis has to be multivariate (consider at least 3 variables at once). We accept descriptive analyses such as PCA or clustering if they provide useful insight into the question.
- In your statistical analysis, 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()`.
- The two figures need to be of different types. E.g., one plot could be a box-plot and one a scatter plot.
- At least one of the two figures has to be multivariate (consider at least 3 variables at once).
- If you do a PCA analysis, a PC1 vs PC2 scatter plot and a rotation plot will jointly count as **one** figure. You will have to make one more, other figure.
- Both the figures and the analysis need to be clear, understandable, and jointly provide a clear answer to the overarching question. Negative results are acceptable.
- Part II of the project must use the provided .Rmd template.

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 two plots in the context of the question you asked. (2-4 sentences)
- Answer the question you have asked.

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.
- When reading or writing .csv files, using directory paths that exist only on your computer

..