Midterm Report
Due 04/05/2019 by beginning of class

Instructions:
You may work with others on these problems but you must turn in your own work. You may type your solutions in any way you like (LaTeX, Markdown, Office, etc...) as long as you present your work clearly and in an organized way.

Unless otherwise specified, you must hand in a printed copy of your work at the beginning of class.

The task:
This midterm assignment is designed to prepare you for your final project. In particular, you must begin your own open ended investigation into a data set of your choice. You may use data that you already work with outside of this class. **However, each individual in class must study a unique data set, different from each other individuals' data in this class.**

You will be expected to perform the following:

1. Select a data set of your choice, providing a URL/ DOI or other documentation of the authenticity and uniqueness with respect to the other projects in class. You may not use data from Faraway. If you use a public data set, you can sign up first-come, first-serve.
2. The data set must be registered here: `https://docs.google.com/spreadsheets/d/1G2GbiEEivGjEvj7Y8SIblxRpmTMjWrKpnMydON_ZutQ/edit?usp=sharing`. If this is data related to your research outside of class, please, still provide some documentation as above.
3. Perform exploratory analysis. You should interrogate the data for patterns, multi-modality, correlation between variables, summary statistics, trends, outliers, power laws (nonlinear scaling) and any points of interest.
4. Model at least one relationship you find in the data. This should be a "meaningful" relationship. This is subjective, but trivial relationships such as "there is a linear relationship between temperature in Celsius and temperature in Fahrenheit over the Atlantic" will be given zero credit... Use common sense.
5. Test the explanatory variables in the model for correlation and significance. Try to make the model as simple as possible, but without leaving anything important out. Justify your choices using hypothesis testing and confidence intervals for parameters.
6. Evaluate the goodness of fit of the model, and the major sources of uncertainty.
7. Evaluate the predictive and explanatory power of the model — particularly, how effective does the model appear to be at these tasks. How might these predictions be unreliable? What are the limits of the prediction power?
8. Evaluate issues that you think you might encounter with the assumptions we have utilized so far. For the final, we will diagnose the issues quantitatively, and take remedial measures to improve the model.

What to turn in:

This is to be in the form of a project proposal for the final. The proposal should be **no more than 5 pages, including figures**, but not including any references or appendices in this page count. **Concisely** address the points above. Describe only the most important and interesting parts of your analysis — **code snippets will not be necessary in the main text** but you should clearly describe your methodology for a general scientific audience that is familiar with these techniques. Contextualize your results with what sorts of tests you have made (e.g. for significance), what results you have left out and why these aren't included in your final analysis. Include figures and tables for the most important components of your analysis, and for explanation purposes.

Your corresponding code and work should be included in the final appendix, section 6; I reserve the right to request a copy of the original analysis. If there isn't sufficient documentation in the appendix and this cannot be provided by the student at request, the midterm will not receive any credit. Cases of plagiarism will be handled furthermore with respect to the policy on academic dishonesty.

Your proposal should be written clearly and structured as follows:

Section 1 Introduction. Discuss the data set, your opening research question and why this question is meaningful.

Section 2 Describe your exploratory analysis, including relevant tables and figures.

Section 3 Describe your model, how you arrived at it, its goodness of fit, its significance versus other choices of models, and its uncertainty. Describe the predictive power, and the uncertainty. Include relevant tables and figures.

Section 4 Describe your proposed research question for the final. How will you revise your original research question? What issues have you encountered so far? What assumptions do you think you need to (re-)evaluate?

Section 5 References to data sets, papers, books or other works consulted.

Section 6 An appendix including relevant code and work.

Whenever plotting:

- Your plot is clearly labeled in all axes, legends, and the plot includes a clear title.
- The plot must be sensible and easy to read.

Grading — You will be graded for each of the following items:

1. Introduction and motivation – 5 points;
2. Exploratory analysis – 5 points;
3. Model selection – 5 points;
4. Evaluating goodness of fit and sources uncertainty – 5 points;
5. Evaluating the predictive and explanatory power of the model – 5 points;

6. Evaluating issues of the model encountered so far, revised research question and conclusion – 5 points.

In addition, reports that fail to follow the instructions of this assignment, the structure for the proposal, or to meet standards of scientific writing will be subject to a loss of points.