**Big Data and Analysis: Midterm Project**

In this project, you will choose a dataset and investigate it using the tools we have used in 1st quarter. You will present a Jupyter Notebook with your findings, as well as an oral/visual presentation. This is considered a major project. It is assessed out of 50 points, but comprises a separate part of your grade (along with your future final project).

Part I: Choose your dataset (end of class 10/26)

Really, you can choose just about anything. The only guidelines are that it must be something you are actually interested in (I promise it will make your analysis better), it must be adequately large (ask me if you’re not sure), and you must cite its source. If you can’t, dive deeper, ask a librarian, or choose something else.

Two great sources for datasets are the FiveThirtyEight GitHub (go to FiveThirtyEight.com and scroll down to the bottom, click on the GitHub link) and the US Census (<http://censtats.census.gov/usa/usa.shtml> is a good place to start)

Part II: Choose your questions and plan your analysis (end of class 10/27)

Choose one central question to investigate, and a few related smaller questions. You can always add more or change them later, but this will give you a path to go on. Once you have your central question, plan (in writing, not code) how you will go about performing your analysis using Python.

Part III: Perform your analysis (end of class 11/4)

You must use at least 2 of the major techniques we covered in class. As a recap, these are:

-mean/median/mode

-Pearson R correlations/linear regression

-counting/sums

You must include at least 3 graphs, but you’ll most likely have more than that. You must provide analysis that is statistically and mathematically based and shows a depth of thought. If your original questions did not provide significant depth, revise them and continue to investigate.

Part IV: Write explanations/conclusions (end of class 11/7)

Be sure to include an introduction and a conclusion, subheadings that reflect the strands of your analysis, and significant explanation of your findings along the way. Please proofread, and be sure to use statistical and mathematical reasoning.

Part V: Create a presentation (start of class 11/8)

Though your Jupyter Notebook will not include explanations of code, your oral presentation should show the major points of your programming process. You should also include an explanation of your dataset, why you chose it and its source, and your initial hypothesis and findings. You will be presenting in class on 11/8 with a presentation of about 5 minutes.