Assignment 5, Data Processing & Visualization QMSS G4063
Navid Hassanpour, `nh2519@columbia.edu`
Due Thu. April 21, 2016

Please submit your assignment on Courseworks and include links to your 1) code and 2) web-based visualizations in the report. If you have static visualizations, embed them in your submission file. There are bonus points for conducting the challenge part (note: it is optional) at the end of the assignment and uploading your Shiny app online.
Do not upload large data files, all coding should be executable on the data files made available to you online, no need for large file uploads.

Predicting the Results of the April 26 Primaries in Pennsylvania and Maryland, (1000 words, Regression Results Visualizations)

At this link, access a daily collection of tweets pertaining to the Primaries. Using the `.csv` files at this link and what you have learned in the course so far, *to the best of your abilities*, predict the results of the Primaries on April 26, in the states of *Pennsylvania*, and *Maryland*, for Democratic and Republican candidates.
Employ as many variables as you see relevant to this exercise. Your design should include a learning and classification mechanism, such as a relevant regression analysis, OLS or logistic, or any other scheme of classification and prediction you have in mind. Note that your dependent variable can be the share of the vote in a given poll. You can also apply logistic regressions with binary dependent variables (losing or winning the elections).
Some of the independent variable you can incorporate to your prediction models include, but are not limited to,

1. Volume of the tweets
2. Sentiment of the tweets pertaining to a specific candidate
3. Number of retweets in tweets pertaining to a specific candidate
    (a) Exploratory: Other network parameters such as *clustering coefficient*, *average degree*, *average path length*

Use the above parameters and your design to estimate the prospects of a candidate (`Clinton` v. `Sanders`, and `Cruz` v. `Trump`). Note that your data points are *daily*.
Note that the daily *average* poll numbers from RealClearPolitics and FiveThirtyEight can help you to generate extra data points, but obviously they are secondary to the results of the Primaries themselves (the real polls, that is).

If you can use enough geolocated tweets, then you can apply controls for each state to counter the state-specific factors. Try to account for temporal factors (dependence across consecutive days) and location specific effects (for each state). How do you ameliorate these concerns? Report the results of your findings along with your design and analysis. Do you find any of your independent variables to be highly predictive of the poll results? Discuss those that are significant, and outline the choice of your control variables. Include visualizations for the results of your regression fits (DV vs. several IVs), and coefficient plots.

Extra points: Fine Tuning Your Design Strategy with The New York Primaries on April 19 (500 extra words, visualizations if needed) Apply the same prediction strategy for the New York Primaries (both Democratic and Republican). You will have a chance to compare your predictions (done on April 18) to the real results. How far from real numbers were your predictions? What do you think caused that discrepancy? How did you make necessary adjustments to ameliorate the shortcomings/amplify the good predictors for the April 26 Primaries in Pennsylvania and Maryland?