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 regression, 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*.

Dependent Variable: The daily *average* poll numbers from RealClearPolitics, Huffington Post Pollster, and FiveThirtyEight can help you to generate your data points for the depen-

dent variable, but obviously they are all only daily estimates of the results of the Primaries themselves (the real polls, that is).

Note on the generation and coding of your dependent and independent variables: Code your dependent variable (the approval rates, prospects of a given candidate) based on the daily numbers from one of the three websites I included above. Use the complete average of all polls available (such an aggregate average is extant on some of the the aforementioned websites).[1] You can use the raw numbers for a given candidate or code it as a categorical dummy for `win/lose` for usage in a logistic regression analysis. Harvest the data for the two states you have in mind[2] for 60 or 70 days we have the data for. This variable is going to be your *Dependent Variable* or DV. For your independent variables, listed above, you need to use the techniques we have worked with during the class. Calculate the daily volume of the tweets from the state you are working on (this needs singling out geocoded tweets from that state), sentiment of the tweets per candidate per state per day, number of retweets etc. You will have to generate all of these variables for the states in question, and for the days you are considering. The linear fit on panel data you have generated has the following general format,

$$Y_{si} = AX_{si} + \epsilon_{si}. \tag{1}$$

In addition to your variables, generate an index dummy for the states ($s = 1, \ldots, K$), and one for time ($i = 1, \ldots, T$). Now you want to include your control variables for each state. These include the state's population, average income and alike, and are constant over all time units (i) in your dataset. Once you are done with constructing your panel of size KT, you can run your favorite regression model on the data you have produced. Note that you have to account for state specific characteristics of the data, that means you will have to "cluster your standard errors" on the state level, and use a Fixed Effects (FE) model with your regression. Finally, to account for time correlations from day to day, include a lagged (one day) version of your DV as one of the IVs. This should take care of the (Markov) time dependence between your data units from day i to day $i + 1$.

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

[1] Again note that these are mere estimates for real polls that happen at the primaries themselves.

[2] PA and MD here, although you can code one or two other competitive states to generate more data points.

results of your regression fits (DV vs. several IVs) (using `ggplot2` cmmands we dicussed), and coefficient plots (using `coefplot`).

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?