STAT 133
DUE THU DEC 8, 11pm – with intermediate deadlines, a 4-day grade period, and a hard deadline of MON DEC 12, noon

2016 Presidential Election Debrief Project

STEP 0. TEAM FORMING – DUE NOV 21

Create a project team consisting of 3 to 5 members. Each person must be working with at least 2 new team members. Once you have formed a team, enter your names on bcourses.

STEP 1. DATA WRANGLING – DUE DEC 1

Your goal here is to create one comprehensive data frame that consists of data from six sources.

Sources:

1. 2016 Presidential Election results reported at the county level. These are available at
   http://www.stat.berkeley.edu/users/nolan/data/voteProject/2016_US_County_Level_Presidential_Results.csv
   The original data are from tonmcg's github account at
   https://github.com/tonmcg/County_Level_Election_Results_12-16/blob/master/2016_US_County_Level_Presidential_Results.csv

2. 2012 Presidential Election results reported at the county level. The original data are available from http://www.politico.com/2012-election/map/#/President/2012/
   These data are now available at
   http://www.stat.berkeley.edu/users/nolan/data/voteProject/countyVotes2012/xxx.xml
   Where the xxx.xml is replaced by one of the following

```
alabama.xml                  louisiana.xml                oklahoma.xml
arizona.xml                  maine.xml                    oregon.xml
arkansas.xml                 maryland.xml                 pennsylvania.xml
california.xml               massachusetts.xml            rhode-island.xml
colorado.xml                 michigan.xml                 south-carolina.xml
connecticut.xml              minnesota.xml                south-dakota.xml
delaware.xml                 mississippi.xml              stateNames.txt
district-of-columbia.xml     missouri.xml                 tennessee.xml
florida.xml                  montana.xml                  texas.xml
georgia.xml                  nebraska.xml                 utah.xml
hawaii.xml                   nevada.xml                   vermont.xml
hrefs.txt                    new-hampshire.xml            virginia.xml
idaho.xml                    new-jersey.xml               washington.xml
illinois.xml                 new-mexico.xml               west-virginia.xml
indiana.xml                  new-york.xml                 wisconsin.xml
iowa.xml                     north-carolina.xml           wyoming.xml
kansas.xml                   north-dakota.xml
kentucky.xml                 ohio.xml
```

   These state names are available at

http://www.stat.berkeley.edu/users/nolan/data/voteProject/countyVotes2012/stateNames.txt

Here's snippet the Alabama.xml file:

```
<table>
<thead>
<tr>
<th scope="col" class="results-county">County</th>
<th scope="col" class="results-candidate">Candidate</th>
<th scope="col" class="results-party">Party</th>
<th scope="col" class="results-percentage">% Popular Vote</th>
<th scope="col" class="results-popular">Popular Vote</th>
</tr>
</thead>
<tbody id="county1001">
<tr class="party-republican race-winner">
<th rowspan="5" class="results-county">Autauga
<span class="precincts-reporting">100.0% Reporting</span>
</th>
<th scope="row" class="results-candidate">M. Romney</th>
<td class="results-party">
<abbr title="Republican">GOP</abbr>
</td>
<td class="results-percentage">72.6%</td>
<td class="results-popular">17,366</td>
</tr>
<tr class="party-democrat">
<th scope="row" class="results-candidate">B. Obama (i)
</th>
<td class="results-party">
<abbr title="Democratic">Dem</abbr>
</td>
<td class="results-percentage">26.6%</td><td class="results-popular"> 6,354
</td>
</tr>…
```

3. 2008 Presidential Election results (county level) are available from The Guardian in a Google Sheet at https://www.theguardian.com/news/datablog/2009/mar/02/us-elections-2008

   This sheet has been uploaded as an xlsx spreadsheet at http://www.stat.berkeley.edu/users/nolan/data/voteProject/countyVotes2008.xlsx

   Note that the spreadsheet has tabs for each state. You will need to export these data as CSV files (or some other delimited file) in order to merge them.

Here is a screen shot of one state's (Wyoming) sheet

	A	B	C	D	E	F
1	County	Total Precinc	Precincts Re	Obama	McCain	Other
2	Albany	22	22	8,618	7,981	413
3	Big Horn	13	13	1,108	4,043	117
4	Campbell	35	35	2,986	13,001	238
5	Carbon	19	19	2,336	4,331	149
6	Converse	19	19	1,380	4,924	121
7	Crook	17	17	612	2,967	82
8	Fremont	32	32	6,016	11,082	383
9	Goshen	24	24	1,832	3,942	108
10	Hot Springs	4	4	618	1,834	70
11	Johnson	17	17	908	3,334	84
12	Laramie	60	60	16,070	24,549	808
13	Lincoln	18	18	1,823	6,485	217
14	Natrona	46	46	8,144	17,573	547
15	Niobrara	6	6	244	1,017	27
16	Park	29	29	3,757	10,838	305
17	Platte	13	13	1,407	2,993	126
18	Sheridan	29	29	4,450	10,169	275
19	Sublette	9	9	936	3,316	76
20	Sweetwater	36	36	5,762	10,360	440
21	Teton	18	18	7,472	4,567	216
22	Uinta	11	11	2,317	5,759	238
23	Washakie	5	5	1,042	2,956	62
24	Weston	8	8	658	2,618	92
25						

4. 2004 Presidential Election results (county level) are available at http://www.stat.berkeley.edu/users/nolan/data/voteProject/countyVotes2004.txt

Here's a snippet of those data:

```
"countyName" "bushVote" "kerryVote"
"arizona,apache" 8068 15082
"arizona,cochise" 24828 16219
"arizona,coconino" 20619 26513
"arizona,gila" 10494 7107
"arizona,graham" 7302 3141
"arizona,greenlee" 1899 1146
"arizona,la paz" 3158 1849
"arizona,maricopa" 539776 403882
"arizona,mohave" 29608 16267
"arizona,navajo" 16474 14224
```

5. Census data from the 2010 census available at http://factfinder2.census.gov/faces/nav/jsf/pages/searchresults.xhtml?refresh=t
   These data are available in three CSV files: `B01003.csv DP02.csv DP03.csv`
   These files each have an accompanying TXT file that describes the variables.
   `B01_metadata.txt DP02_metadata.txt DP03_metadata.txt`
   Not all variables described in the meta data files are available. The DP02 file contains socio-data, DP03 contains economic data, and B01 contains race information. For example the DP03 file contains information on:

   HC01_VC04, EMPLOYMENT STATUS - Population 16 years and over
   HC02_VC13, EMPLOYMENT STATUS - Percent Unemployed
   HC01_VC31, COMMUTING TO WORK - Public transportation
   HC01_VC42, OCCUPATION - Service occupations

Be careful with the B01 file as the data are organized differently than with DP02 and DP03. Here's a snippet:

```
GEO.id,GEO.id2,GEO.display-label,POPGROUP.id,POPGROUP.display-label, HD01_VD01,
HD02_VD01
0500000US01001,01001,"Autauga County, Alabama",001,Total population,53155,*****
0500000US01001,01001,"Autauga County, Alabama",002,White alone,42031,185
0500000US01001,01001,"Autauga County, Alabama",004,Black or African American alone,
9508,116
0500000US01003,01003,"Baldwin County, Alabama",001,Total population,175791,*****
0500000US01003,01003,"Baldwin County, Alabama",002,White alone,151453,831
0500000US01003,01003,"Baldwin County, Alabama",004,Black or African American alone,
16613,416
```

All six of these files are available at
http://www.stat.berkeley.edu/users/nolan/data/voteProject/census2010/xxx.csv

6. GML (Geographic Markup Language) data that contains the latitude and longitude for each county. These are available at
http://www.stat.berkeley.edu/users/nolan/data/voteProject/counties.gml

Here's a snippet from this file:

```
<?xml version="1.0"?>
<doc xmlns:gml="http://www.opengis.net/gml">
<state>
<gml:name abbreviation="AL"> ALABAMA </gml:name>
<county>
<gml:name> Autauga County </gml:name>
<gml:location>
<gml:coord>
<gml:X>  -86641472 </gml:X>
<gml:Y>  32542207  </gml:Y>
</gml:coord>
</gml:location>
</county>
```

Your data frame should contain one row per county. It should have data from all files. This means it should have at a minimum the following variables from the election results and the county locations:

- State
- County
- Trump votes and Clinton votes from 2016
- Obama votes and Romney votes from 2012
- Obama votes and McCain votes from 2008
- Bush votes and Kerry votes from 2004
- Latitude
- Longitude

In addition, select several variables from each of the three census files. For example Total Population and White alone from B01, Percent unemployed and Employed in service industry from DP03, etc. You will want 30-40 variables from these three files.

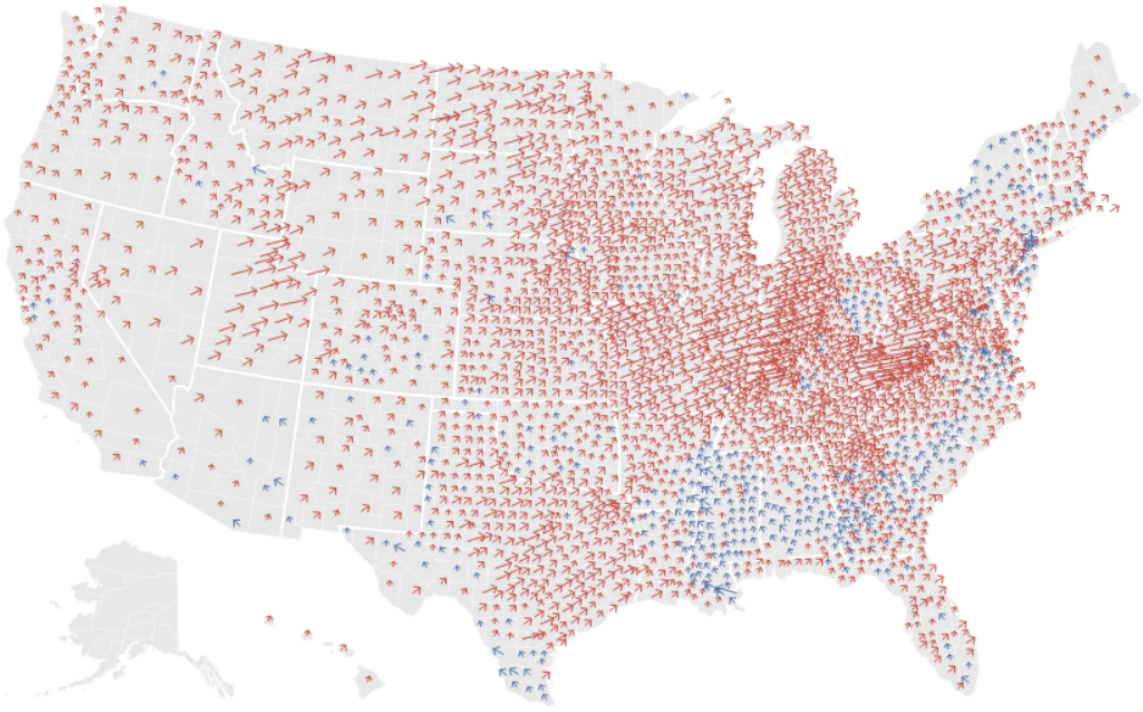
Each team member must contribute to the DATA MASHING STAGE. Make it clear in your code who has done which part.

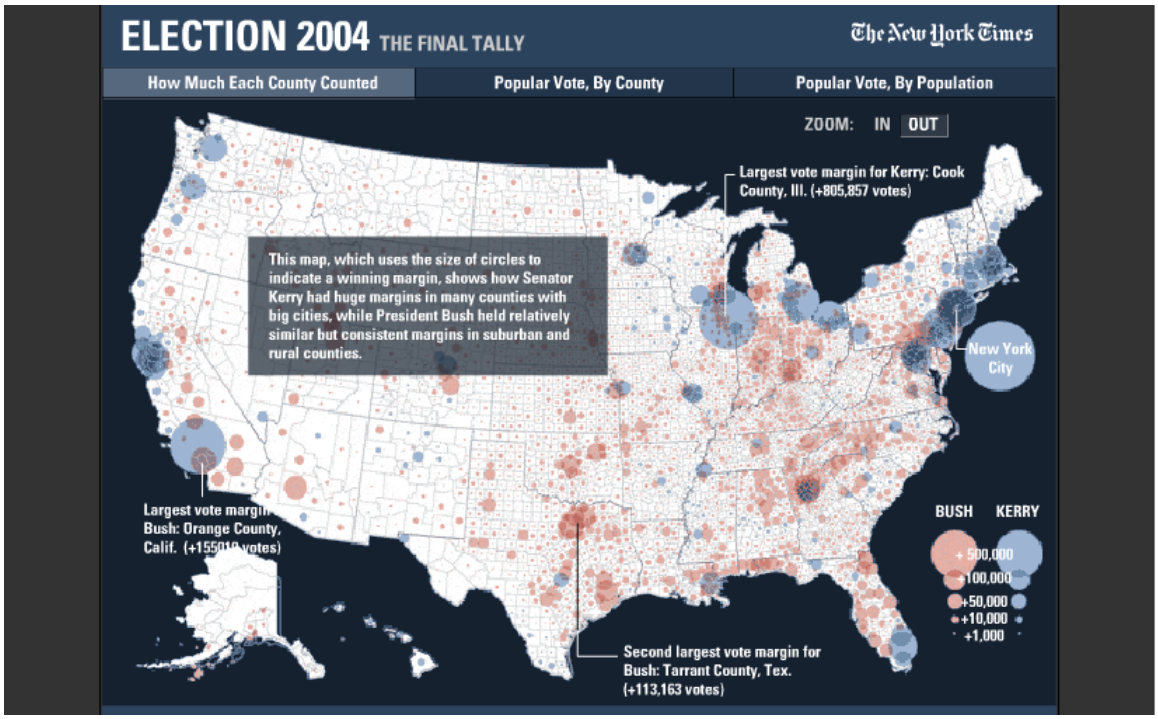
STEP 2. EXPLORATION

Your goal here is to carry out preliminary explorations that can help you in the further stages of analysis. Make several plots and describe your findings.

STEP 3. MAP MAKING

Your goal here is to make an informative map describing the election results. These may be put in the context of previous elections. To help you, consider the following maps. One shows the change in votes from 2004 to 2008, where the length of the arrow is proportional to the vote shift. Another shows the total vote in each county for the candidates. Feel free to gain inspiration from maps that you find online, but be sure to acknowledge your sources.





STEP 4. MODELING

Your goal here is to create two predictors for the 2016 election results using the variables you create in the merge. One predictor is for the 2016 election results and one is for the change from 2012 to 2016. Assess the accuracy of your predictors. Compare the predictors. Did they do well in the same places? Explore where each did well and where it did poorly.

Use different methods for the two predictors. For example, K-NN, Classification trees, Naïve Bayes. If you are familiar with other methods such as logistic regression or SVM, you may use them.

Recursive Partitioning – Read the documentation carefully and make sure that your data are of the correct types for use by rpart(). Consider various values for the control parameters in fitting the tree. To figure out how to do this, read the help for the rpart.control() function. Arguments to this function can be passed in the call to rpart() through its … argument. You may find the following documentation helpful: http://www.statmethods.net/advstats/cart.html in addition to the package documentation at http://cran.r-project.org/web/packages/rpart/rpart.pdf Be sure to make a plot of your tree.

Nearest Neighbor – Use *k* nearest neighbors (the knn() function in R) to predict the winner of the 2004 election. A neighbor can be determined by geography (latitude and longitude) plus a few other features of a county. If some variables are continuous and others are dichotomous then consider creating two distance matrices (one ofr each set of

variables) and combining the distances. Consider various values of *k* and with which variables to include in the distance calculation.

Training your predictor. Consider a few options for training your predictor. For example, consider training your data based on the 2012 election results, and then testing your data with the 2016 results. Or, consider 2 or 3-fold cross-validation where you split each states counties into 2 or 3 groups at random to create your folds.

STEP 5. FINAL REPORT – DEC 8 … 12

Communication is an important part of this project. Prepare a report that discusses your findings. The report should have the following sections:

1. Introduction
2. Data description
3. Map
4. Predicting the 2016 results
5. Predicting the change from 2012 to 2016
6. Discussion
7. References

In the section on the map provide a discussion of the patterns and trends that you see in your map. Connect it to the findings in the predictor sections. In your sections on the predictors, include plots that explain your predictor, including where it works well and where it doesn't.

In the discussion section, synthesize your findings from the map and the two models. Compare the two models. Did they do well in the same places?

Make plots that showcase your findings in these sections. Include plots in the data description section that help set the stage for the analysis. Turn in at least 6 and no more than 12 plots. Write captions for each plot that describes the main features and how they make your points.

For the map making and modeling, your team can split up and have 1 to 2 people work on each aspect. Be sure to put the names of the person(s) that worked on each part in the title of that section of the report.

What to turn in:

1. The final data frame that you use to analyze the election results.
2. The Rmd file for your report
3. A knitted version of your report that hides the code chunks
4. A knitted version of your report that displays the code chunks.