STAT 133
Final Project
DUE Fri Dec 5, 11:55pm – with intermediate deadlines

STEP 1. TEAM FORMING – DUE Thursday Nov 13

Create a project team consisting of 4 or 5 members.

STEP 2. DATA MASHING – DUE 11:55pm Tuesday Dec 2: This is a hard deadline just like a regular HW assignment.
Your goal here is to create one comprehensive data frame that consists of data from three sources. You should think of a team name and one person from the team should submit a file YourTeamNameStep2.R. The first lines of the file should be comments with the names of the people in your group, one name per line. Try to be unique in your choice of team name so that everyone has a different team name…

Sources:

1. 2012 Presidential Election results reported at the county level. The original data were from politico.com. These data are available at http://www.stat.berkeley.edu/users/nolan/data/Project2012/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.xml                    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/Project2012/countyVotes2012/stateNames.txt

   You only need the data for % Democrat and % GOP, and there is no data for Alaska. You'll need county information to match with the other datasets.

   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>…
```

2. 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/Project2012/census2010/xxx.csv
   Include all variables in these files, with NAs for missing data. A lot of it won't be useful in making predictions in STEP 3 and 4 but for now just include everything.

3. 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/Project2012/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>
```

STEP 3. SUPERVISED LEARNING – due 11:55pm DEC 5. This is a soft deadline, and you won't be penalized as long as you turn in everything by 11:55pm DEC 12.

You can turn in the code for STEP 2-4 all together in one file called YourTeamNameFinal.R. It should run from start to finish, so it will first mash the data as you did in STEP 2 and then use that data. If you want to change your code from STEP 2 to make it work differently, that's fine, we won't look back at the STEP 2 code anyway.

Your goal here is to create two predictors for the 2012 election results using all these variables (except the actual 2012 results). I recommend 10-20 predictors. You can choose them based on an organized method of looking at correlations or plots, or you can go through and choose a few that you think will work well. You will use the 2004 election results (i.e. the winner in each county (Rep or Dem)) to train the predictors. If I had been a bit more organized I would have had you put that in the data.frame originally in STEP 2, but it's pretty straightforward to match once you've done STEP 2 already. The data is at `http://www.stat.berkeley.edu/~nolan/data/Project2012/countyVotes2004.txt`
You are trying to solve the real world problem of predicting election results before the election happens, so you can only use data from before the election (i.e. the variables from the census as well as previous election results) to build the model.

A. *Recursive Partitioning* (rpart() in rpart package) – Read the documentation carefully and make sure that your data are of the correct types for use by rpart(). The method is "class". Play around with the parameters for fitting the tree until you have a tree that you are satisfied with. 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
   Make a plot of your tree.

B. *Nearest Neighbor* – Use *k* nearest neighbors (the knn() function in R) to predict the winners of the 2012 election. A neighbor should be determined by geography (latitude and longitude) plus a few other features of a county. Play around with various values of *k* and with which variables to include in the distance calculation. The `train` set and `test` set will be the same – the data frame of longitude, latitude, and the other variables that you have chosen to include. The `cl` argument contains the winning party for the 2004 election. Ask for the proportion of votes among the *k* neighbors to be returned so that you can use this in determining the winner.

Have 2-3 people in your group work on A and 2-3 work on B. Indicate who does what by commenting in your code.

STEP 4. PLOTS AND PREDICTION ASSESSMENT – DEC 5 (soft deadline, same as for STEP 3)

Prepare a document in html that contains a set of plots with captions.
Use the two predictors developed in STEP 3 to predict winners in the 2012 election. Compare the two models. Did they do well in the same places? Dig deeper and explore where your model did well and where it did poorly.

A. Make plots that showcase your findings. Turn in 3 to 5 plots. (For example, show a plot like in HW 8 that shows misclassification for knn based on different values of k, plots for different numbers of variables used as predictors, etc.)

B. Fancy plot: make a map similar to the NYT map shown on the next page that compares the change in votes from 2004 to 2012. The length of the arrow is proportional to the vote shift from the 2004 to the 2012 election. If you don't want to make this map, you can do something else of comparable difficulty, possibly involving new data and/or data from 2014 (non-presidential) election results (you'll need to find the data).

Write captions for each of your plots describing the main features and how they make your points. Turn in an html document containing your plots, captions, and any other observations/comments.