STAT 133 Lec 1 2014
Homework 8
DUE Sun DEC 7, midnight

You will be working with 2004 election data and census data to predict the outcome of the 2012 election.

DATA MASHING

The data used comes from four different sources, we have created one data frame for you with all of the data (in all.Rda). For your information here is a description of the sources of the data.

Sources:

1. 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 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.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/Project2012/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>…
```

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

   The B01 file is 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

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>
```

4. 2004 Presidential Election results (county level) are available at http://www.stat.berkeley.edu/users/nolan/data/Project2012/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
```

For your information you have The data frame all.Rda has one row per county and includes data from all the files listed above. In addition to election results you have latitude and longitude as well as population information. The variables repVotes, demVotes and winner are from the 2012 election while bushVote, kerryVote and winParty are from the 2004 election.

```
> names(all)
 [1] "key"             "state"           "county"          "lat"
 [5] "lon"             "countyName"      "bushVote"        "kerryVote"
 [9] "winParty"        "repVotes"        "demVotes"        "winner"
[13] "GEO.id2"         "GEO.display.label.x" "totalPop"     "whitePop"
[17] "blackPop"
```

SUPERVISED LEARNING

Your goal here is to create two predictors for the 2012 election results using all these variables (except the actual 2012 results). You will use the 2004 election results (i.e. the winner in each county to train the predictors. Please select ONE of the following prediction methods to turn in. Two R-files will be pushed to you Monday Dec 1st and you will use the one for the method you select. For extra credit complete both.

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 winner of the 2004 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

PLOTS AND PREDICTION ASSESSMENT

Use the predictor you selected to predict the winner in the 2012 election.

A. Make plots or tables that showcase your findings. Include a caption for each plot/table with observations and comments. Save these in a file named hw8.pdf.

B. You can pair up with someone who used the other prediction method and include a comparison of the two methods.

FINAL TASK

C. Make a map similar to the NYT map shown below that compares the change in votes from 2004 to 2012. The length of the arrow is proportional to the vote shift from the 2008 to 2012 election. Your plot will be of the **vote shift** from 2004 to 2012. Notice that this plot does not depend on your prediction results, only the election results that are in all.Rda, and the location of each arrow are given by the longitude and latitude . Use the 'maps' package in R, map('usa') will plot a map of the US which you can then add to.