STAT 133
Project 3
DUE MON DEC 10, 11pm – with intermediate deadlines

STEP 1. TEAM FORMING – DUE NOV 21

Create a project team consisting of 3 to 4 members. Once you have formed a team, go to the Forum section of bspace and enter the following information for your team:

- team name,
- team member names,
- indicate which team member will be the official point of contact for the project

Once you have supplied your team information, the point of contact should send an email to the instructor and GSI, ccing all team members, with the following subject line:

STAT 133 Final Project Team: YourTeamName

Put your actual team name in for "YourTeamName".

STEP 2. DATA MASHING – DUE DEC 5

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

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 now 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.xml
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
```

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/census2012/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/census2012/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
```

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

- State
- County
- Obama votes
- Romney votes
- Bush votes

- Kerry votes
- Latitude
- Longitude

In addition, select several variable 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 3. SUPERVISED LEARNING – DEC 10

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 (Rep or Dem) to train the predictors.

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

Have 1-2 people in your group work on A and two work on B. Indicate which in your group.

STEP 4. PLOTS AND PREDICTION ASSESSMENT – DEC 10

Prepare a document that contains a set of plots with captions.

Use the two predictors developed in STEP 3 to predict the winner in the 2012 election. Assess both of your models accuracy. 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.

B. 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 can be started now, it doesn't depend on earlier work.

Write captions for each of your plots describing the main features and how they make your points.