Stat 133, Spr '08
Primary Project: Part I
Due Friday May 9, midnight

For this project you will use data from several sources. All of the data sets are available on bspace in the Resources section of the class website.

In the first part of the project, you are to pull together the data from the various sources to create a data frame for use in the analysis. Recall that the analysis will involve fitting a model to predict whether a county is an Obama-majority or Clinton-majority county and a map to represent your findings.

In this stage you will massage the data into common format, and perform exploratory analysis to confirm that your data frame is correct, to see if the data have any problems that need to be addressed, and to inform your model fitting.

Census Data County level information from the 200 census. These data are provided as Excel spreadsheets. They were taken from the Internet at
`http://blueprod.ssd.census.gov/statab/ccdb/`
You are provided with three tables:

- **cc07_tabB1.xls** has data for population, land area, and population density for all counties .
- **cc07_tabB3.xls** has data for age (under 5, 5 to 14, 15 to 24, ... 65 to 74, 75 and over), race (white alone, black alone, asian alone, ...), and gender (males per 100 females).
- **cc07_tabB4.xls** has data for educational attainment (HS degree or higher, Bachelor's degree or higher), foreign born population, and speaking language other than english at home, Hou sehold income $75,000 or more, persons in poverty.

2004 Presidential Election The plain text file, **countyVotes2004.txt** contains county level votes for the 2004 presidential election. These results were scrapes from a CNN Website in 2005.

2008 Democratic Primary Election results from the 2008 Democratic primaries are available in **DemPrimary2008.txt**. Note that not all states are present as these data were acquired from the CNN Website on April 25, before all states had held their primary/caucus. Also note that some other states are not available. These data are from
`http://www.cnn.com/ELECTION/2008/primaries/results/county`

Geographic data The gml file **counties.gml** provides the latitude and longitude of the center for each county, In addition, the file **codes02.csv** contains information about the US Regions, e.g. which states belong to which regions. It is from TBD.

Other Data These additional data sets may prove useful in your analysis. The file **stateAbbrevs** contains the full state name and the two letter abbreviations for the state names. The **top30Cities** file contains the name, longitude, and latitude of the 30 largest cities in the US.

Report For this part of the project, you are to create a data frame, where each row refers to a county in the US. The variables in the data frame should include, state abbreviation, county name, census data (all of the variables that were used in the NY Times analysis plus a couple more), the region (Northeast, South, Midwest, and West), latitude, longitude, votes for Bush and Kerry in the 2004 election, votes for Obama, and Clinton in the 2008 primary, when available, an indicator for whether the county went for Clinton or Obama in the primary/caucus.

You are to turn in the code used to create the data frame as a plain text file. You are also to write a 2-3 page report on your work preparing the data for future analysis. Be sure to include a discussion of the problems that you encountered and how you remedied them (or not),

Note If you ned an extra day or two to complete Part I of the assignment, please discuss this with me prior to the deadline. Part II will have a due date of a week later, and Part III will be due the day of the final exam. These parts will be posted this week for those groups who wish to complete the assignment earlier.