Stat 133, Spr 2015
Homework 6: XML
Due Sunday, April 12 at 11:55pm

Your task is to create a visualization (scatter plot) of 4 variables, including infant mortality and population for countries around the world. The source for the data is the 2013 CIA Factbook, which is an XML file that I have placed on my Website at
`http://www.stat.berkeley.edu/users/nolan/data/factbook/factbook.xml`

Use your browser to search for these two variables. It's a very large file and is slow to load into the browser. Search the file for "Infant Mortality" to find the node that has the desired values. Examine the node and its children carefully to figure out how to extract the information that you want. The population appears in a similar format in the document. Choose two other variables that interest you and extract them as well. By the way, the table in the appendix of the factbook contains county codes and country name.

Use the functions in the XML package, such as xpathSApply, xmlValue, xmlGetAttr, etc. to extract the desired information. Note that infant mortality values for each country do not appear in the same order as population values. This is because the countries are arranged in decreasing order for each quantity. That is, the most populous country does not have the highest infant mortality. Additionally, it may be that values are not available for all of countries for some of the variables. So, when you extract the infant mortality variable, also extract the country code. You will need it to merge the variables together into one data frame.

Here are the steps:

1. Extract a vector of infant mortality rates and a vector of corresponding country codes from the factbook. Convert the rates to numeric and bind the two vectors into a data frame.

2. Repeat the previous step for population and two other variables of interest to you and that you think might make a good plot.

3. Use the merge() function to merge the 4 data frames into one data frame. Be sure to read about the by, by.x, by.y, all, all.x, and all.y parameters to merge to figure out how best to combine the data frames.

4. Make a beautiful scatter plot that incorporates 4 variables. You may want to use color, plotting symbol size, plotting symbol, etc. to represent some of the variables. Consider whether a transformation of scale makes a more informative plot. Label axes, provide a legend, etc. Write a caption for you plot. (Note the cut() function may be helpful if you want to discretize a numeric variable to use for color or size.)

What to turn in the Rmd file with your code to create the data frame and make the plot and a paragraph/caption for your plot, and turn in the knitted Rmd file.