Lab 0 (R warm-up)
Stat 215A, Fall 2014

Due: Tuesday September 9, in section (not graded)
(written by Yuval Benjamini, Jessica Li, Adam Bloniarz, Ryan Giordano)

Install R, learn it

- Go here, download and install R: http://cran.rstudio.com/
- Go here, download and install Rstudio: http://www.rstudio.com/ide/download/desktop
- Install ggplot. In an R session, type: `install.packages("ggplot2")`

If you've never used R before, here are a few resources for learning it:

- The Quick-R website: http://statmethods.net/
- *A Handbook of Statistical Analyses Using R* by Brian Everitt and Torsten Hothorn
- *Introductory Statistics with R* by Peter Dalgaard
- To learn ggplot, you can start at http://ggplot2.org. The book "Elegant Graphics for Data Analysis" is available as an ebook in the library

If you are more familiar with MATLAB or Python, there are numerous cheat sheets on the web to help you translate between the two. Here are a few examples:

- MATLAB to R: http://cran.r-project.org/doc/contrib/Hiebeler-matlabR.pdf
- NumPy to R: http://mathesaurus.sourceforge.net/r-numpy.html

In this class, we will ask you to follow the Google R style guide. Please read and follow the guidelines at https://google-styleguide.googlecode.com/svn/trunk/Rguide.xml

R warm-up exercises

1. Load the data

1. Load USArrests - type `data(USArrests)` in R
2. Download statecoord.txt data from bSpace
3. Load statecoord.txt into R using `read.table`
4. Merge them into a single data frame, and double check that you did this correctly.

2. Compare two variables

1. Try plotting "Murder" vs. "Assault" using `geom_point`. What do you see?
2. Try plotting "Rape" vs. urban population. There should be an outlier point. Mark it with a different color.
3. Now do the same plots using the names of states instead of the points (use `geom_text`). Do you see anything interesting?

3. Regression

You can fit a linear regression using `lm` function. You should also implement the linear algebra yourself (see chapter 3 of Freedman's textbook), and check that you get the same results.

1. Fit a linear regression of urban population on "Rape".
2. Plot predicted values versus the residuals. Do you see any trends?
3. Replot "Rape" vs. urban pop and draw a blue line with the predicted responses.
4. Now refit without the outlier, and add a red line on the same plot.
5. Compare the lines. Are the linear responses a good description of the data?
6. Make a publishable graph. Add a header (`ggtitle`), axis labels (`xlab` and `ylab`) and customize the legend (see `scale_color_manual` for example).