Stat 133, Spring 2011
Homework 3: Graphics and Databases
Due Friday, Apr. 1 at 11:59PM

Graphics and Databases

1. At the web site http://www.fao.org/es/esc/prices/PricesServlet.jsp?lang=en is an applet that allows you to download data for prices of various commodities over ranges in time. Note that you can make multiple choices of commodities and dates by pressing the control key while clicking on the choices, or by clicking on a first choice and shift-clicking on a final choice. You can copy the link to the CSV file that appears after you click 'Submit Query' to download directly into R, or download the CSV file to your computer in the usual way.

   Choose between four and eight commodities of interest, and download the monthly data for as many years as are available. Convert the dates to one of R's date types, and plot each commodity's price over time in such a way that it's easy to compare the changes in the different commodity's prices over time.

2. At the web site http://www.dof.ca.gov/HTML/FS_DATA/STAT-ABS/Toc_xls.htm are links to several Excel spreadsheets containing statistical information about California.

   Choose a spreadsheet from that web page that interests you, download it, and read it into R using one of the methods presented in class. You should choose a spreadsheet that has at least one variable for every county in California. Choose a suitable gradient, and divide the variable into a reasonable number of groups, and create a map of California with the colors of the gradient representing the value of the variable for each county.

3. I have placed a copy of the SQLite albums database that was used in lecture at http://www.stat.berkeley.edu/classes/s133/albums.db. Use the database to answer the following questions:

   - List the names of the ten songs which appear the most times in the collection, along with the number of times they appear.
   - List the names of the ten artists who have the most albums in the collection, along with the number of albums.
   - List the names of the ten artists who have the most tracks in the collection, along with the number of tracks.
   - List the names of the ten artists who have the most total time (length) for all their tracks in the collection, along with the total time.

4. The next part of the assignment uses a baseball database which was downloaded from http://www.baseball-databank.org/. You will need to access the MySQL server running on `springer.berkeley.edu` using the username and password given in class, and, if you're not using an SCF machine, you'll have to set up an SSH tunnel. There are 25 data tables in the database, many of which you won't need to use. Answer the following three questions using at least one data summary and/or graph to support your answer.

   (a) Which college has produced the baseball players with the highest batting average? Batting average can be calculated by taking the number of hits (column `H` in the `Batting` table) and dividing by the number of at-bats (column `AB` in the `Batting` table). You may want to eliminate players who have had fewer than, say, 50 at-bats from the calculations.

   (b) Do players who hit more home runs receive higher salaries than players who hit fewer home runs?

   (c) Do players who have nick names (`nameNick` in the `Master` table) tend to receive higher salaries than players that don't have nick names? Can you think of a reason why this may be true?

   Notice that for the last two questions it's important to take time into account. In other words, you'll have to break down your analysis to handle each year separately.

Your submission should be contained in a *single* pdf file. Email this file to me (`s133@stat.berkeley.edu`), by 11:59PM (plus or minus a minute) on the due date. Make certain to save a copy of your email submission.