Lab 11: Money, It's A Hit

This lab shows how R interfaces with other programs, particularly SQL database management. We will use the package `RSQLite`, which not only provides an R interface but also installs a minimal library for database access.

Unless you are already a SQL ninja, the handout from Wednesday's lecture will be vitally important to this assignment.

Today, we will look trends in baseball team payrolls between the years 1985 and 2010. The data come from the Baseball Databank http://baseball-databank.org and is based in part on Lahman's Baseball Database. Information on the attributes in the database can be found at http://baseball1.com/files/database/readme58.txt. You will need to download the SQLite database file `baseball.db` (located at http://www.stat.cmu.edu/~cshalizi/statcomp/14/lectures/23/baseball.db) to your computer.

There is an R package, `lahman`, containing data frames with all of this data, which you can use after the lab session if you're curious. Using it in this session will be counter-productive.

1. Install the R packages `DBI`, `RSQLite` and `fImport`. Ensure that `plyr` is also installed if you wish to use those functions.

2. Here we will import payroll data from the database.

    a. Using DBI and RSQLite, setup a connection to the SQLite database stored in `baseball.db`. Use `dbListTables()` to list the tables in the database.
    b. Use the table that contains salaries and compute the payroll for each team in 2010. Use `dbReadTable()` to grab the entirety of the table, then select the relevant subset. Which teams had the highest payrolls?
    c. Repeat the previous step, but now do this using only `dbGetQuery()` and SQL. Verify that your answers are identical.
    d. Modify the SQL statement to compute the payroll for each team for each year from 1985 to 2010.

3. *Visualize the change in payrolls over time.* To do this sensibly, one needs to adjust for inflation. The following code snippet gets price levels (CPI, consumer price index) from FRED (the Federal Reserve Economic Data service).

```
library(fImport)
cpi <- fredSeries("CPIAUCSL",
              from = as.Date("1985-01-01"),
              to = as.Date("2011-01-01"))
cpi <- cpi[months(as.Date(rownames(cpi))) == "January"]
cpi <- cpi / cpi[length(cpi)]
```

    The CPI is measured monthly, but salaries are annual, so we arbitrarily take the price level each January as the level for the whole year. The end result is a vector, `cpi`, containing consumer price indices from 1985 to 2011, normalized so that 1 = \$1 in 2011. An expression like `y <- x/cpi[1990-1985+1]` will convert x 1990 dollars into y 2011 dollars.

    a. Plot the CPI as a function of time. Make sure that the horizontal axis is labeled with years, not the positions along the vector.
    b. Calculate the inflation-adjusted payroll of each baseball team over time. (Hint: You may find `plyr` helpful here.)
    c. Plot the inflation-adjusted payroll of each team over time. (There are many ways to do this, including `for` loops, `matplot`, etc.)

d. Plot the logarithm of inflation-adjusted payrolls over time.
e. Have payrolls generally kept up with inflation, outpaced it, or fallen behind? Are there teams or groups of teams whose payrolls have consistently been higher than the others? By what factor has the gap between the highest and the lowest payrolls grown (or shrunk) over time?

4. **Extra credit**: Expand your SQL query to also retrieve the number of games played, and the number of games won, by each team each year. Create a scatter-plot of the proportion of games won against the inflation-adjusted payroll.