Exercises 3

(1) Life expectancy and economic development

After completing the "House prices" walkthrough, download the data on life expectancy (LifeExpectancy.csv) from the class website. Life expectancy is often used as an indicator for the well-being of a country. Experts on economic development are interested in the relationship between a country's life expectancy and its economic well-being.

This data set has the following variables:
Country: the name of the country
PPGDP: per-person gross domestic product in US dollars
LifeExp: life expectancy at birth in that country
Group: whether the country is in the OECD, Africa, or other

To clarify the "group" variable, the OECD is the Organization for Economic Cooperation and Development:

> The Organisation for Economic Co-operation and Development (OECD) . . . is an international economic organisation of 34 countries founded in 1961 to stimulate economic progress and world trade. It is a forum of countries describing themselves as committed to democracy and the market economy, providing a platform to compare policy experiences, seeking answers to common problems, identify good practices and coordinate domestic and international policies of its members.[1]

[1] Wikipedia, http://en.wikipedia.org/wiki/Organisation_for_Economic_Co-operation_and_Development, accessed 7 Feb 2015.

Build a regression model that relates life expectancy (the response) to GDP. Use a transformation if necessary, and think carefully about whether the "Group" variable seems to modulate the relationship between GDP and life expectancy. Address two questions. 1) What would you predict the life expectancy to be for an OECD country with a GDP of $20,000 per person? 2) What about for an African country with a GDP of $1000 per person? Make sure to provide an interval prediction (not just a point prediction) and to provide some measure of the accuracy/coverage of your interval.

(2) Solder skips

Finish the case study on quality control in the supply chain for circuit-board manufacturing: https://github.com/jgscott/learnR/blob/master/cases/solder/solder.md.

Read about the problem and work your way through the introductory commands I've posted. Then address the question I pose at the bottom of the page: build a model to predict solder skips using these

three predictor variables [Opening, Solder, and Mask], and explain what you have learned fro your data analysis. Write a short report summarizing your analysis and conclusions, including whatever figures or model output that you deem appropriate.

Note: writing up the results of a data analysis is not a skill that anyone is born with. It requires practice and—at least here in the beginning—a bit of help. I have posted some guidelines here: http://jgscott.github.io/teaching/writeups/write_ups/.

(3) Project data sources

There's nothing to turn in here, but I am carving out a homework problem specifically to encourage you to **spend 30-60 minutes thinking about and researching potential data sets for your project after spring break.**[2] I'll give more details on the project in due course, but the basic idea is simple: pose an interesting question; collect a relevant data set; and use the data, in conjuction with the statistical modeling tools we have learned in class, to answer the question you have posed. Make sure to quantify any uncertainty that arises in answering your question, and to address any shortcomings in the answer provided by your data and analysis. This assignment is purposely open-ended, allowing you considerable freedom to follow a path dictacted by your own intellectual curiosity. You will be evaluated both on the technical correctness and the overall intellectual quality of your presentation.

[2] Due the third Friday after spring break.

Where can one find data? Everywhere! These websites in particular have long lists of data sources: https://github.com/caesar0301/awesome-public-datasets and http://stats-for-change.github.io/data.html. If you want to branch out even further, here's a short list of other sources you might consider: major newspapers, the U.S. census, the Federal Reserve, academic journals, the Economist, Twitter, the World Bank, ESPN.com or other sports sites, Craigslist, Amazon prices, EBay, the Bureau of Labor Statistics, Facebook, the World Economic Forum, the OECD Factbook, the CIA World Factbook, the Securities and Exchange Commission, Yahoo finance, Google Public Data Explorer, your own vital signs, your own experiment or survey, your favorite blogs, your other classes, and your friends. If you know how to write a program that will scrape a website, your options are almost limitless here. As template questions, you should recall some of those we've used for in-class activities so far (house prices; the price of a gallon of milk; used car prices; and so forth). You might even consider recapitulating a similar analysis using a different data set.