Info 271B
Fall 2013

Lab 1: Exploratory Dataset Analysis

Introduction.

In this lab, you will work in groups of two to write your own R script and perform an analysis on the Countries dataset, which is available on the course website. Your group should submit two files:

1. A script file (.R) containing your program code
    a. Include a title section containing your full names, program description, approximate time it took to complete the lab, and whether you mind me using your results as a class example.
    b. Include comments before each logical section of the program, explaining what it does
    c. Label your variables clearly
    d. Name this file your-last-names-lab1.R

2. A commented log / output file containing your results
    a. Please use .doc or .pdf format
    b. Include the important output from running your script with comments, as well as your discussion, and graphs that you generate
    c. Name this file your-last-names-lab1.doc or .pdf

Place both files in a zip folder, named your-last-names-lab1.zip, and email them to agswigart at ischool.berkeley.edu. Please title your email "lab1 submission"

All files are due at 3:30pm on Thursday, Oct 10.

Program Requirements.

Write a well-commented R script to perform each of the following tasks

1. Variable Manipulation

    a. Create a variable, internet_growth, that equals the fractional increase in the percentage of internet users from 2010 to 2011. Compute the median, mean (and any other numerical statistics you want) for this variable.

    b. Based on the variable you just created, find the country with the highest level of internet growth.

c. Create a Boolean variable, high_gdp, that equals TRUE if a country's 2012 gdp is higher than the mean. Compute how many countries are above the mean, and how many below. Explain your result in terms of the shape of the distribution.

d. Compute the mean level of internet_growth for countries with above-average gdp, and for countries with below-average gdp. Comment on which one is greater and why you believe that is.

e. Recode the region variable into dummy variables (one for each region of the world). Make sure R treats your variables as factors.

2. Data import

   a. Find one new metric country-level variable from some public source, and merge it into your dataset. Possible sources include the World Bank (http://data.worldbank.org/indicator), the United Nations (http://data.un.org), and the World Health Organization (http://apps.who.int/gho/data/view.main). I would be especially happy, however, if you find a new data source I don't know about. Please identify your source clearly in comments. Most likely, you will need to import your data as a csv or tab-delimited text file. You will then need to view your new dataframe to diagnose any errors that happened during the import. Most common problems can be corrected by opening the file in a spreadsheet program and editing some entries.

3. Graphs

   a. Create a histogram for the new variable you imported in part 2 and describe its shape. Comment on whether it appears normal, and provide numerical summaries to support your point. If your variable is not normally distributed, check if taking the logarithm helps correct this. Comment on whether this is a sensible thing to do for your variable.

   b. Create a scatterplot between fertility rate and the percent of internet users (2011). Describe the relationship between these variables. What transformation can you perform to make the relationship more linear? Comment on whether this is a sensible thing to do.

   c. Create one more graph that you think is interesting that involves the new variable you imported in part 2. Write a few sentences about what the graph shows.