StatR 101: Fall 2012
Homework 3
Eli Gurarie, October 9, 2012
Due Saturday, October 13, midnight

Instructions: All homework must by typed. For this homework, I would like to see the R code that you use, in particular for the functions that you write. Export plots as needed from R and incorporate them into your document. Upload the completed homework assignment into the course webpage drop-box. Much of this homework builds on the code in the in-class computer lab: `lab3.r`.

1. Write a function called `Median()` which calculates the median of a vector of quantitative measurements and is robust to NA's in the vector. Test that your function gives the same answer as `median()` for any of the columns in the Country data set (which is full of NA's). Note that you will have to test to see if the length of the vector is even or odd. You may want to use a combination of modular division: `%%` (see `?"%%"`) This is done using the `if(Logical Test)` command, which will run a line (or lines) of code if the logical test is true. An example is below:

   ```
   > a <- 1
   > if(a == 1)
   +   print("It's true!")
   [1] "It's true!"
   > if(a!=1)
   +   print("It's false!")
   > # note, no output here, because the test has failed
   > # note also, the "if(a!=1)" could have been simply replaced with "else"
   ```

   You can read more about `if()` in section 4.1.2 (p. 50) of the Braun Murdoch textbook.

2. We talked in lecture about "skewness" (often denoted γ) being a characteristic of a distribution of numbers related to its asymmetry.

   (a) Much as the mean and variance of a population can be calculated using the definition of population mean $\overline{X}$ and population variance s_x^2, the skewness g of a population is given by the formula below

   $$g = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^{3/2}} \qquad (1)$$

   Write a function called `Skew` which computes the skewness of a vector of measurements (and make it robust to NA's).

   (b) Qualitatively, positive skewness indicate that there is more weight on the right side of the distribution and negative skewness indicates that there is more weight on the left side of the distribution. Use your function to calculate the skewness of global GDP, literacy and birth rates. Which of these is most and least skewed, and in which direction. Does the skewness measure give you what you would expect, given what you know (and histograms you can plot) of these variables?

(c) Generate a single table that reports the skewness for GDP, literacy and birth rates for each of the 6 continental regions. Report any differences in the signs of the skewness between the global and regional results. How would interpret the wide difference in, e.g., the skewness of GDP?

(d) Produce boxplots of GDP, literacy and birthrate across continental regions. Make sure you label the y-axes and give a title to the boxplots. What visual feature of the boxplots seems most closely related to high or low values of skewness?

3. Figure 1 shows the "Gapminder World Map 2010", which displays the condition of countries along an axis of wealth (GDP per capita) and health (life expectancy at birth). I've gone through some trouble to download and organize the corresponding data for 2011,[1] and uploaded them here: http://faculty.washington.edu/eliezg/data. Recreate as closely as possible the Gapminder World Map using the `GDPLifeExpectancyRegion2011.csv` (note that the more or less matching region data is in the `Region2` column). In particular, try to include the following features:

   (a) The same x-y scaling,

   (b) Similar colors for the regional data, and similar types of circles (filled circles with black outlines).

   (c) Similar scaling for the sizes of the circles.

   (d) Similar grid lines (use the `abline(v = c(1,2,3))` and `abline(h = c(1,2,3))` functions).

   (e) Similar labeling text on the margins (use the `mtext()`) function.

   (f) Two legends: one identifying the colors with the regions, and one matching circles to particular population sizes.

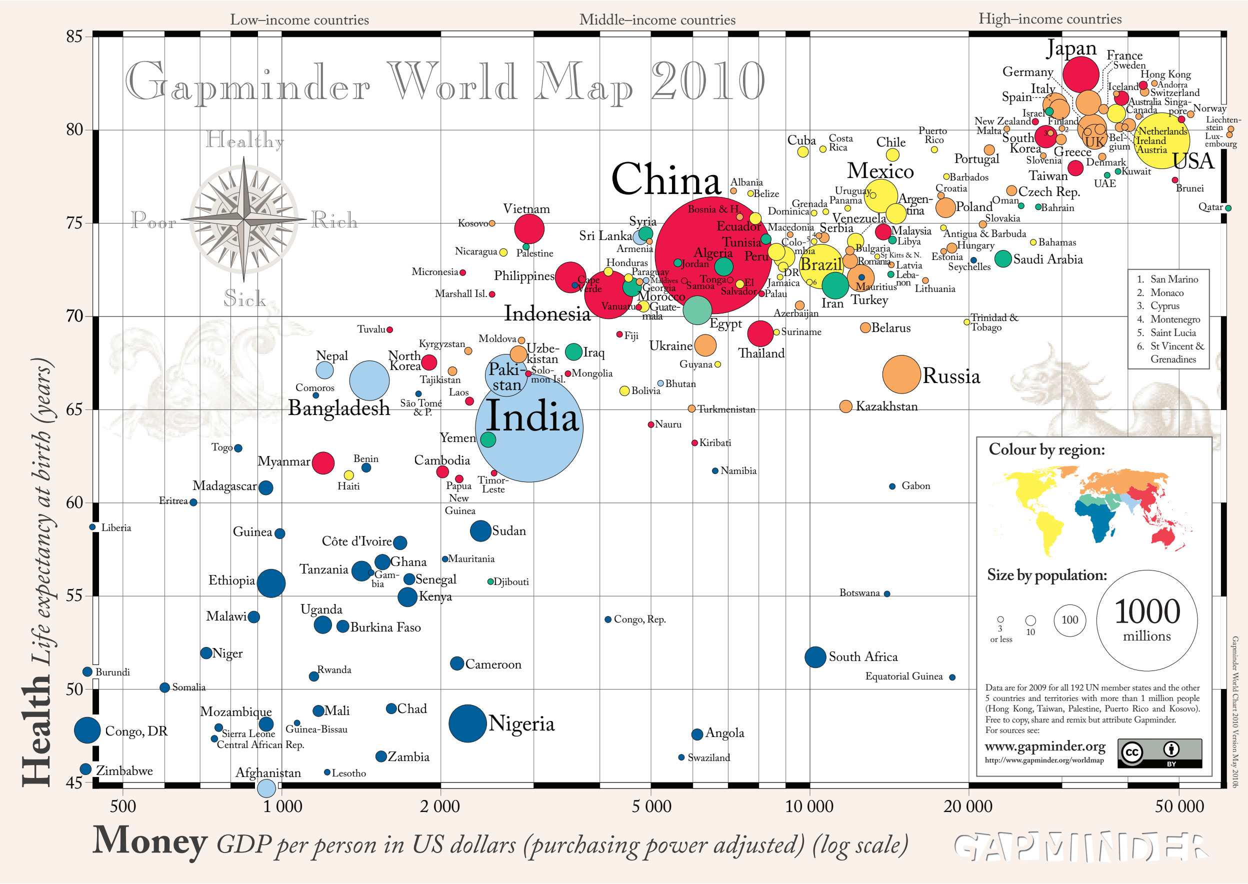
   (g) Label the 10 largest countries.

   Note that you can download data into R directly from a web address. For example, the following code should work:
   `read.csv("http://faculty.washington.edu/eliezg/data/GDPLifeExpectancyRegion2011.csv")`.

Bonus problem: Thing about some data that you may like to analyze for your final project and make some summary tables and figures showing, for example, how a continuous variable might vary under different factors.

[1] Data sources: http://www.gapminder.org/data/ and http://www.iucnredlist.org/technical-documents/data-organization/countries-by-regions

Figure 1: Gapminder World Map 2010


Gapminder World Map 2010
Low–income countries
Middle–income countries
High–income countries
Healthy
Sick
Poor
Rich
Health Life expectancy at birth (years)
Money GDP per person in US dollars (purchasing power adjusted) (log scale)
China
India
Japan
USA
Mexico
Brazil
Russia
Nigeria
Indonesia
Bangladesh
Pakistan
Vietnam
Philippines
Thailand
Iran
Turkey
Egypt
Ukraine
Ethiopia
Congo, DR
South Africa
Sudan
Kenya
Tanzania
Uganda
Afghanistan
Zimbabwe
Angola
Swaziland
Botswana
Equatorial Guinea
Gabon
Qatar
1. San Marino
2. Monaco
3. Cyprus
4. Montenegro
5. Saint Lucia
6. St Vincent & Grenadines
Colour by region:
Size by population:
3 or less
1000 millions
Data are for 2009 for all 192 UN member states and the other 5 countries and territories with more than 1 million people (Hong Kong, Taiwan, Palestine, Puerto Rico and Kosovo). Free to copy, share and remix but attribute Gapminder. For sources see:
www.gapminder.org
http://www.gapminder.org/worldmap
Gapminder World Chart 2010 Version Map 2010b
GAPMINDER