Problem Set 1

Due Wednesday, September 9, 2015, 23:59

For now, copy and paste your commands and results (including figures) into a word document.

1. Suppose you track your commute times for two weeks (10 days) and you find the following times in minutes: 17 16 20 24 22 15 21 15 17 22

    - Use the function max to find the longest commute time, the function mean to find the average and the function min to find the minimum.
    - Oops, the 24 was a mistake. It should have been 18. How can you fix this without rewriting the vector? Do so, and then find the new average.
    - How many times was your commute 20 minutes or more? Hint: try using sum() and a logical statement. Read here about logical operators
    - What percent of your commutes are less than 17 minutes?
    - Replace the 5th element of the vector with NA. Calculate the new mean.

2. The Crimean War

    - Install the HistData package and load Nightingale. Read about it here
    - How many people died of Wounds? How many people died of Disease?
    - Create a piechart using the pie() command. Make sure to label all the slices. Give your graph a title. If you wish, change the colors.

3. Loading data into R

- We didn't get to this in class, but loading data into R will be central to your life as a data analysist.
- Download the following files into your working directory (hint: use getwd())
    - If you want to make it more challenging, read it directly from the web (without downloading it to your working directory).
    - https://raw.githubusercontent.com/ylelkes/R_wav/master/data%20examples/Countries-Europe.csv
    - https://github.com/ylelkes/R*wav/blob/master/data%20examples/child*data.sav?raw=true
    - https://github.com/ylelkes/R_wav/blob/master/data%20examples/GaltonFamilies.dta?raw=true

- Using read.csv, the foreign package, or haven load each one of these datasets into R.
- For: "Countries-Europe.csv" (let's call that object europe)
    - What is the median population of Europe?
    - What is the mean population/land area
    - If you replace X and Y in the following, you will get most populated country in Europe: `europe$X[europe$Y==max(europe$population)]`
    - What variable is X? What variable is Y?
    - In your own words, describe what this command is doing..

- For GaltonFamilies.dta:

    - using the cor() command, what is the correlation between a child's height and his mother's height and what is the correlation between the child's height and the father's height?
    - Use a logical statement, get R to confirm that the first correlation does not equal the second correlation.
    - Using lm(), is there a relationship between the father's height and the number of children he has?

- For child_data.sav
    - What is the memory span of child with the highest IQ?
    - Create a correlation table for the entire dataset
    - From the correlation table, extract the correlation between memory span & IQ, the correlation between age and reading ability, and the correlation between IQ and reading ability into a vector
    - write that vector to a csv file.