Biostat 561: Homework 4

Instructor: Amy Willis, Biostatistics, UW

October 19, 2017

Homework due October 26, 2 p.m.

Link to Homework 4 submission: `https://classroom.github.com/a/VtY_d3vi`

A completed homework will involve multiple figures. Please prepare your submission as a single pdf file – do not upload figures individually – and make sure you show your code in an easy-to-read format. All plots that you submit with this homework should be publication-worthy, i.e. informative legends and titles are mandatory.

Question 1: The Rules

Edward Tufte, data illustration extraordinaire, has 10 rules for effectively illustrating information. All rules apply to the illustration of both quantitative and qualitative information. Read through the rules here.

http://www.sealthreinhold.com/school/tuftes-rules/rule_one.php

Find 3 figures that display quantitative information, preferably with some qualitative information as well, from different sources and in different styles. These can be from the statistical literature or scientific literature (eg. preprint or journal article that you're reading), from a periodical (eg. newspaper or magazine), or from popular culture (eg. a blog post or advertisement). Critique the 3 figures, explaining what they each do well, and what they each do poorly. Explain how you would improve them. Explicitly refer to Tufte's rules in your answer, detailing which rules are broken and upheld by the figures.

Question 2: Multiple plots

Based on the `gapminder` dataset in the package `gapminder`, make 2 plots: the first showing the population of a particular country over time, and the second showing the per-capita GDP of that same country over time (you can pick whichever country is most of interest to you). Following the instructions available at

https://cran.r-project.org/web/packages/gridExtra/vignettes/arrangeGrob.html

arrange them into a single figure using the packages `grid` and `gridExtra`. Is it more natural to plot these 2 figures horizontally or vertically? Why?

Question 3: Data wrangling with dates & legends

The dataset `abundances.txt` in the `lecture4` folder on github contains the abundances of different microbial taxa in different samples. The rows give the taxon name, and the columns give the sample, i.e. the element in row i and column j gives the number of times that taxon i was observed in sample j. The sample names give the dates that the samples were collected.

Make a stacked bar plot showing how the compositions of the taxa change across time (e.g. if I look at three taxa (1, 2, 3) who have counts (10, 20, 10), then the stacked bar plot will show 25% for taxon 1, 50% for taxon 2 and 25% for taxon 3. Since there are 448 different taxa in the sample, choose a reasonable number of taxa and only plot the relative abundances of these taxa.

Tips:

- The dates are in an inconsistent form, and need to be cleaned up before they can be used. In particular, the column names give the dates in the following form: X + day + . + month + . + year + . + sample ID number. Not all dates have a sample ID number, and the year format differs. Be sure to adjust for this.
- You will need to convert the abundances from wide to long format before you can use `geom_bar`.
- To confirm that your figure is correct, choose a specific sample/date and make sure that the relative abundances match those in the figure.

If you are relatively new to R, you are welcome to find a classmate and submit this question together. Please work according to an honour policy: both members in the group should contribute to the solution, and both should be able to reproduce the plot individually before it is submitted.

Bonus Question: Integrating multiple data frames

It is not mandatory to complete this question, though it is excellent practice for research and data wrangling. I strongly encourage every one of you to make a reasonable attempt.

The dataset `covariates.txt` provides information about the disease status of each sample. Repeat Question 3, producing 2 plots – one for positive disease status, the other for negative disease status. Can you see a difference between the 2 states with respect to the taxonomic abundances?