Biostat 561: Homework 3

Instructor: Amy Willis, Biostatistics, UW

17 April, 2019

Homework due April 24, 3:15pm

Link to Homework 3 submission: `https://classroom.github.com/a/EmuY_6-L`

A completed homework will involve multiple figures. There is no need to upload the figures – your `Rmd` and `pdf` files will suffice. 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 `lecture3` 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. (If, in future homeworks, you want analyze something other than microbial abundance data, please e-mail me data that you would like to analyse instead!)

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. This is a data wrangling exercise commonly observed in the wild! `stringr` is a useful package in the `tidyverse` for dealing with strings. Some of the functions starting with `str_` may help you.
- 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.

Question 4: Under five mortality

The `WHO` package allows you to download public health data from the World Health Organisation. Every datasheet has a code, which you can access via `get_codes()`. Once you have the code you want, you can download it using `get_data`.

Download the under-five mortality rate data from the `WHO` package (use `str_detect` to find the code you need).

What's the mean under-five mortality rate for each region? Show me *only* this information for the regions (you may need to exclude some data, missing or otherwise, from your table).

Plot the U5M rate across time for Australia, China, the United States, and the country that you grew up in. Show the rates for females, males, and both sexes in your plot. Make the plot as beautiful as you can!

Question 5: More integrating multiple data frames

The dataset `covariates.txt` provides information about the disease status of each sample in Question 3. 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?