Journal of Statistical Software

October 2017, Volume 81, Book Review 2. doi: 10.18637/jss.v081.b02

Reviewer: Ulrike Grömping
Beuth University of Applied Sciences Berlin

Data Visualisation with R: 100 Examples

Thomas Rahlf
Springer-Verlag, Cham, 2017.
ISBN 978-3-319-49750-1. 400 pp. EUR 53.49 (H), EUR 41.64 (e).
http://www.datavisualisation-r.com/

I took to doing this review for mainly two reasons: I love visualization, because it allows to convey the message contained in data in the clearest and most concise way possible. And I am attracted by this book because it uses base R's very powerful toolbox for the creation of visualizations. The book expressly sets out to present ideas for finely tuned presentation graphics, as opposed to quick exploratory graphics or interactive visualizations.

A book on data visualization cannot possibly be "complete"; this book is devoted to the visualization of low-dimensional data, i.e., methods for visualizing multivariate data are only touched upon (using polar area charts), and ordination methods for multivariate data are not covered at all. Most of the examples are from business, economics or official statistics; this implies a focus on graphical tools for such data, including, e.g., tree maps for the subdivision of a quantitative total like a budget into contributions from several entities, or Lorenz curves for visualizing economic inequality. Interesting chart types from other fields of application, for example multi-vari charts that are used in the quality arena, are not included; for me, the most surprising omissions are the many possibilities for visualizing continuous quantitative data, like kernel density charts, violin or bean plots, or even the classical boxplot with all ingredients, while population pyramids are perhaps a little overemphasized. It is of course natural that the author sets his own priorities, and the examples are interesting and useful for a broad audience; thus, the above observations are not meant to be complaints.

The book starts with an initial overview chapter, followed by a general "Part I" of five chapters and a main "Part II" that presents (about) 100 examples, organized in six chapters. In the initial chapter, the author discusses the role of data visualization in today's world between science and data journalism. He recognizes pioneers of providing visualizations for everybody, like Hans Rosling (Gapminder, https://www.gapminder.org/) or Otto Neurath (an Austrian social scientist of the early 20th century who strived for making statistical facts accessible to illiterate or educationally disadvantaged people). Likewise, he acknowledges visualization experts like William S. Cleveland and Edward Tufte, who according to Thomas Rahlf set a standard for visualization. The book does not introduce any such standards in a systematic way; rather, in individual examples, reasons for design decisions occasionally relate

to principles brought forward by Tufte and others. In this context, I find it very agreeable that the book foregoes strict rules (like condemning pie charts) in favor of more cautious recommendations. The initial chapter also outlines how the examples are constructed: each uses unchanged real data, for which the entire code for the creation of the desired graphic is provided; the author states that he creates exactly the figure he previously envisioned in his mind (or as a sketch with pen and paper); this sounds idealized, but is certainly a pleasant contrast to letting the software decide.

"Part I" of the book (p. 6 to p. 93) introduces "Basics and Techniques" and, according to the author, can be skipped by R experts. I think that this part does have something to offer to almost every reader (including R experts): Chapter 2, "Structure and Technical Requirements", introduces the principal layout of visual displays of possibly several figures or charts, gives examples of the relation of graphical choices to the type of information that can be conveyed, and discusses typefaces, special symbols and colours. The (elegant and versatile) typeface family Lato is introduced here; variants of Lato are used in most examples; while this initially irritated me because it forces me to install the family for running the examples, I have quickly become a **Lato** fan. Chapter 3, "Implementation in R', explains important basics of R interspersed with specifics needed for visualizing data, like spatial data structures, the access to web data through APIs, graphical concepts of base graphics and packages and functions used in the book. Chapter 4, "Beyond R", presents tools outside of R that enable readers to incorporate visualizations into documents in exactly the layout that is desired (LaTeX, postprocessing and font creation with Inkscape); the font created in Section 4.2 is used in later examples. Chapter 5, "Regarding the Examples", gives an overview of the examples to come and how they are systematized.

"Part II" is the core of the book: Chapters 6 to 10 provide examples for the visualization of categorical data (pp. 94–167), distributions of quantitative data (pp. 169–216), time series (pp. 217–279), bivariate quantitative data with scatter plots (pp. 281–304), or geography-related information with maps (pp. 305–351). The final Chapter 11 shows "Illustrative Examples", i.e., charts whose entertainment value is possibly higher than their information value. Polar area charts and charts with pictograms of people dominate this brief section. Notable exceptions are a variant of the famous figure about Napoleon's war activities that was popularized by Tufte (1983, Tufte called it "probably the best statistical graphic ever drawn"), and a bubble plot of life expectancy against GDP, which is inspired by a Gapminder representation. All visualizations from the book can be found at `http://www.datavisualization-r.com/`, which also links to the scripts for all examples and to many datasets. In the book, Thomas Rahlf explains the code for each example step by step, with more detail for earlier than for later examples. The code for each example is self-contained (with the exception that users are expected to have the Lato font installed and have created the symbols as explained in Chapter 4). The book's R code does not deserve an award for elegance, beauty, or briefness; readability could have been improved by simple things like introducing blanks around `<-`, and it would perhaps have been useful in some places to take steps for shortening repetitive code, like defining colors that are used in many places (e.g., `myClight <- rgb(191, 239, 255, 80), maxColorValue = 255)` would have allowed to shorten various statements on pp. 98 and 103). However, all examples follow a common structure, and repetitive code is aligned such that the repetition is immediately obvious, so that the printed code in the book is instructive, and readers can familiarize with R's rules for creating base graphics, while working through the examples. The simplicity of the book's R code does not detract from the proper-

ties of the figures, which are well-designed with a lot of attention to detail and thus deserve to be at the center of the reader's attention.

In the chapter on scatter plots, I was somewhat at odds with the book's choices of symbol sizes and coloring strategies: the symbols are very (unnecessarily) large, which causes so much overlap that I found it hard to perceive the number of symbols; also, in some figures symbols are colored because they are in a certain area, which seems redundant to me, and I consider it more logical to color the respective area instead of the symbols. I set out to *quickly* do that for one of these figures, and ended up with a very ugly visualization; I spent quite some effort (with my non-statistician husband as a critic) until I considered my version of the graphic attractive enough to compete with the book's. This certainly made me realize how much meticulous fine tuning is needed for obtaining not only an informative but also a beautiful graphical display. Figure 1 shows the book's design (top figure), and my attempt with coloring of quadrants; the color scheme used is the colorblind-friendly version of red(=bad)-yellow-green(=good) coloring; I decided against coloring the upper left and lower right quadrants, because I do not see an obvious preference between (relatively) low life expectancy and (relatively) low self-perceived health. While working on my own figure, I came to like the book's figure as well. It is quite beautiful, and it does not really loose any information by making the symbols as large as they are.

Throughout the book, there are nice examples of many small things that make a difference in making a figure beautiful: for example, there is a lot of information on incorporating fonts or special plot symbols, and reference lines are done in unobtrusive ways, e.g. in background color on top of the foreground, or by shading the background in two slightly different light colors. There is, however, one point that I do not like in the book's recommendations: using Inkscape for postprocessing a pdf file is in my view a desperate last resort that one should not normally use, when creating visualizations with R. The book recommends postprocessing for the placement of point labels; had I followed that recommendation for Figure 1, I would have had to repeat that postprocessing several times. I used calculated positions instead, including individual adjustment of text positions for certain points (Finland in both figures, in the top figure also Portugal).

The book is a translation of the German language title "Datendesign mit R" (data design with R), which was published in 2014 by Open Source Press (a second edition is scheduled to appear in the Springer Spektrum series next year under the title "Datenvisualisierung mit R"). Consequently, you will encounter the odd German label here and there because of occasional oversights in translation. There are also a few typos, but overall the book is carefully written and carefully edited. One thing I sorely miss is an index; it would be quite helpful to be able to look up functions used and example data and especially their first occurrence. That would be a nice improvement for a next edition.

Using real data is a strength of the book and at the same time creates difficulties: some of the data sets are not downloadable from the book's website because of license issues. While these can be downloaded elsewhere (with URLs to data providers listed in an appendix), not all of them are frozen and thus may be subject to change. For example, the 2011 data for Figure 1 can still be found at `http://www.oecdbetterlifeindex.org/`, but the obvious place holds the latest version of the data; the book's homepage now contains a section on data for pointing readers to difficult-to-find data sets like this one. I replicated various examples, using the easily found latest data versions instead of searching for the book's version of the data; this meant that I had to make small changes to the scripts in order to work with the

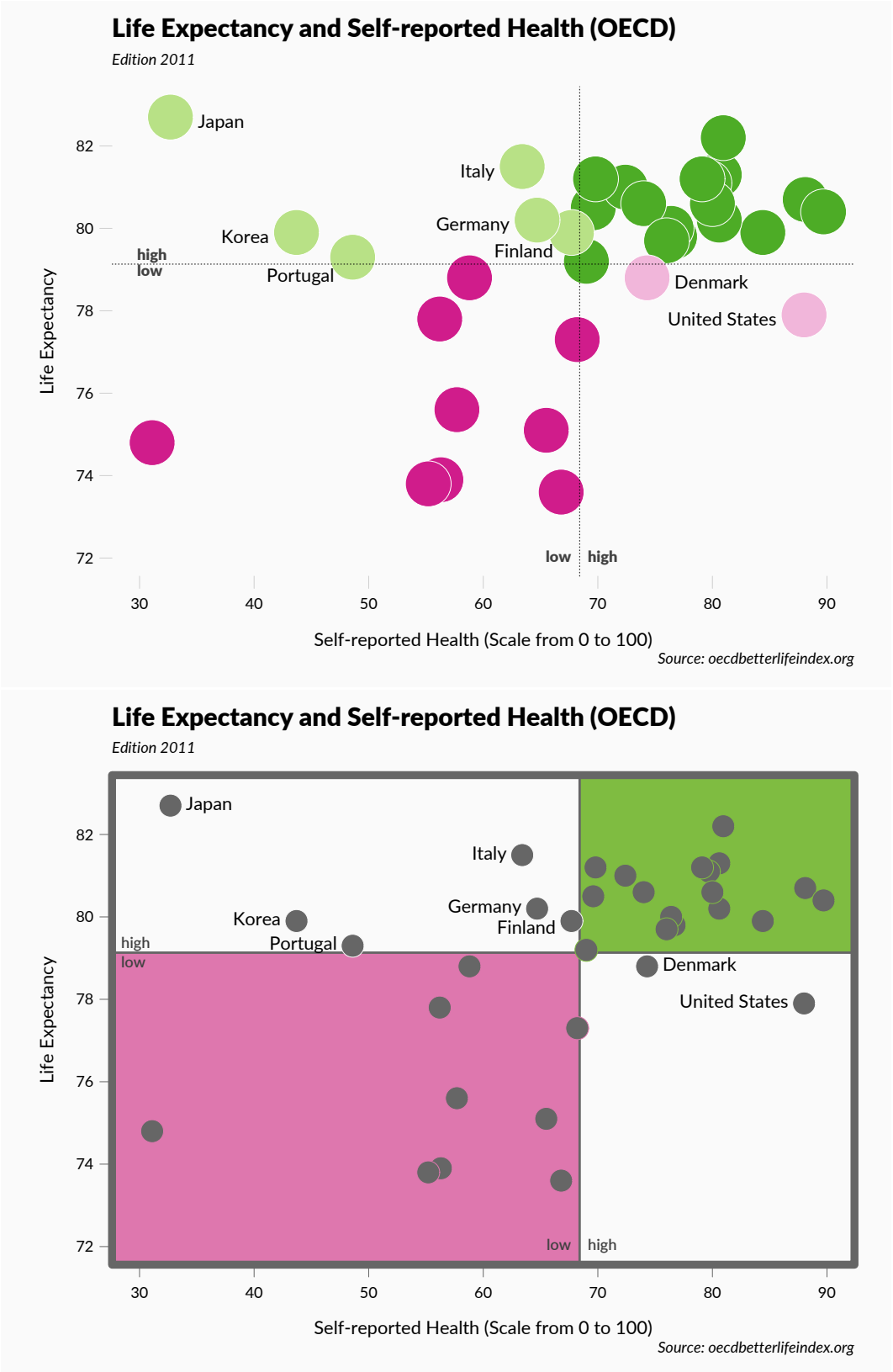

Figure 1: Coloring points according to location vs. coloring the quadrant.

current versions of the data. For example, as there were 41 instead of 39 states in the latest version of the International Social Survey Programme file, the 8x5 layout of the panel of pyramids shown in the book had to be changed into a 7x6 layout in order to accommodate all pyramids on one page. Where the variable structure in an SPSS file had changed (ZA4800 as the successor of ZA4753), I could not read the data any more with the script from the book (which uses package **memisc**), but the function `read.spss` from package **foreign** (strangely not mentioned in the book) gave me access to the data and allowed to run the script after a few modifications.

I definitely recommend this book. The example-based approach is very successful for introducing readers to R's graphical capabilities; readers can learn proficiency in using base R graphics for obtaining exactly the static presentation figure they envision – including ambitious infographics. The focus is on printed reports, where "printed" includes the creation of pdf files. The insights gained from the book can also be used for other purposes, e.g. for creating graphics to be incorporated into an animation created by R package **animation** or **magick** (not covered in the book). Last but (by far) not least, the book can be used as a collection of ideas for useful, informative and beautiful graphical displays.

References

Tufte E (1983). *The Visual Display of Quantitative Information.* Graphics Press, Cheshire.

Reviewer:

Ulrike Grömping
Beuth University of Applied Sciences Berlin
Department II
D-13353 Berlin
E-mail: groemping@bht-berlin.de
URL: http://prof.beuth-hochschule.de/groemping/

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
October 2017, Volume 81, Book Review 2 *Published:* 2017-10-27
doi:10.18637/jss.v081.b02