Journal of Statistical Software

September 2010, Volume 36, Book Review 3. http://www.jstatsoft.org/

Reviewer: Hadley Wickham and Garrett Grolemund
Rice University

Graphics for Statistics and Data Analysis with R

Kevin J. Keen
Chapman and Hall/CRC, Boca Raton, FL, 2010.
ISBN 9781584880875. 489 pp. USD 69.95 (H).
http://www.crcpress.com/product/isbn/9781584880875

A proliferation of misused graphics has followed the proliferation of cheap statistical and graphing software. To quell this epidemic, we must develop our graphicacy skills, something that is rarely taught in school. Keen offers this book, "Graphics for Statistics and Data Analysis with R", as a lesson in the subject.

A good graph can communicate information more directly and immediately than numbers or words, but it is not easy to make a good graph. Keen begins in Chapter 1 with an outline of the principles behind good graphics. A good graph should be tailored to its audience and the audience must have adequate time to understand it. A good graph should also have a thoughtfully chosen scale and thoughtfully chosen design elements (which Keen calls the "graphical frame"). The guiding principle behind these choices should be to avoid ambiguity through clear labelling.

Keen subscribes to a mental model where people must access a "graphical icon" in their long term memory before they can understand a graph. When a viewer encounters a novel graph type, it takes him longer to interpret it because he must first build a new graphical icon. By this reasoning, Keen recommends that analysts stick to commonly used graphs. This becomes the organizing principle for the remainder of the text:

- single discrete: basic (Chapter 2), and advanced (Chapter 3); dot, bar, pie, stacked bar, pictogram.
- single continuous: exploration (Chapter 4), diagnostics (Chapter 5), nonparametric density estimation (Chapter 6), parametric density estimation (Chapter 7); dotplot, stemplot, boxplot, ECDF, QQ plot, probability plot, histogram, density plots, normal density estimation, Person's curves, etc.
- two variables: both discrete (Chapter 8), one discrete and one continuous (Chapter 9) and both continuous (Chapter 10); grouped and side-by-side dot (with and without whiskers), bar, box, and pie chart, mosaic plot, sunflower plot, bagplot, 2D histogram, etc.

- models: simple linear regression (Chapter 11) and polynomial regression (Chapter 12); residual plots, influence plots, etc.
- multivariate data (Chapter 13); glyphs, 3D charts, scatterplot matrices, etc.

The book is methodical and complete, but focuses too much on plotting one or two variables from a small data set. Too much space is taken up by minor variations of a plot (especially the first eight color figures, pp. 71–74 and 88–89), and too little (only 44 pages out of 435) on multivariate data. This is particularly problematic given that almost all real data have more than two variables!

This is a result of Keen's model of graphical reading: we must explicitly list every named graphic, rather than learning how they are generated from underlying principles. There are two powerful and general techniques that receive short shrift because of this: there is no discussion of color and only a tiny discussion of conditioned/trellised plots. (The author also incorrectly describes a casement plot as a scatterplot matrix, not, correctly, as a trellised plot.)

For a book on graphics, the details of the graphic display could be better. The graphics all look a little large (especially for the amount of data they display), and the text on the graphics is substantially larger than the text on the page. Many graphics lack frames, and have poor figure-ground contrast, making it hard to distinguish guides from the data.

Reading this book will give you the ability to recognize and create the majority of the named graphics of statistics, but named graphics are only a small portion of interesting and useful graphics. I would recommend this book if you were interested in a detailed survey of 1D and 2D graphics (a better title would be "1D and 2D graphics in R"), but it will not provide you with the skills to explore and understand large, modern datasets, nor to create new types of graphic tailored for a specific purpose.

Reviewer:

Hadley Wickham
Rice University
Department of Statistics
Houston, TX, United States of America
E-mail: hadley@rice.edu
URL: http://had.co.nz/