Journal of Statistical Software

July 2010, Volume 35, Book Review 1. http://www.jstatsoft.org/

Reviewer: Pedro M. Valero-Mora
University of Valencia

---

ggplot2: Elegant Graphics for Data Analysis

Hadley Wickham
Springer-Verlag, New York, 2009.
ISBN 978-0-387-98140-6. 216 pp. USD 59.95.
http://www.springer.com/978-0-387-98140-6

***ggplot2**: Elegant Graphics for Data Analysis* is a new addition to the UseR! series by Springer, probably the fastest expanding source of resources for computational statistics at the current moment. The books in this series are all linked with R, either presenting a new package developed by the own authors of the book or describing how to applying statistical techniques with the different packages available in R. **ggplot2** is an implementation in R of *The Grammar of Graphics* (Wilkinson 2005) a systematic approach to the specification of statistical graphics that was introduced in a book previously reviewed in the *Journal of Statistical Software* by Cox (2007). This implementation has been developed by Hadley Wickham, who is also the author of the book reviewed here.

ggplot2 is a wonderful piece of software that I have enjoyed learning and using. The basic use, supported by the command `qplot()` is enough for fulfilling the needs of those interested in creating the most common plots as seen in academic papers and reports, and, with only a bit of extra effort, other more advanced plots tuned up to specific problems. (A warning is needed here, **ggplot2** is a framework that allows making both meaningful or meaningless plots equally easily so it is on the user to know what makes a good graphic!)

The book has ten chapters. After an introductory chapter setting the principles and other preparatory stuff, Chapter 2 is devoted to the `qplot()` command. This command is intended as a replacement to `plot()`, the basic command for plotting in R, but it uses some aesthetic additions – a gray canvas with white grid lines is probably the most apparent but they are more – and different choices for defaults – e.g., `qplot(var)` will produce a histogram of `var` instead of the index plot obtained with `plot(var)`. Actually, `qplot()` is powerful enough for fulfilling many if not all the needs that somebody accustomed to using only a few standard graphics may have. Indeed, if you are one of these, you may regard yourself as fortunate as only 15 pages may be sufficient for satisfying your plotting needs.

Chapters 3 to 7 are devoted to explaining the most advanced features of **ggplot2**. Chapter 3 describes the components involved in drawing a simple plot. The example shows how the components can be changed to producing alternative plots. Then, the concept of layer is

introduced. Layers are responsible for creating the objects that we observe in a plot and provide a mechanism for tuning up plots to specific needs. The rest of the chapters in this group are devoted to explaining the components in more detail.

Chapter 4 shows how changing or adding components to a layer enables reusing a plot with different data, changing the different geometrical elements in it, its statistical summaries etc. Note that some of the examples are a little bit artificial, being more of a demonstration of the capabilities of the system than actual plots you might want to use in practice. Chapter 5 is more in the line of showing how to create real graphics organized according to common user goals such as producing "named" graphics like scatterplots or boxplots, displaying distributions, dealing with overplotting and so forth. Chapter 6 describes scales including mapping data to colors, legends and axes, and Chapter 7 is entitled "Positioning" and deals mainly with faceting, grouping, and coordinate systems.

Chapter 8 deals with polishing the final details of a plot for publication. I know some people that use R mainly because it draws plots that have "the right look" for academic publications – white background, a rectangle border, and no grid lines. Instead, **ggplot2** uses a different set of defaults – a borderless light gray background with a grid made of white lines – chosen with the intention of producing plots that look right both on the screen and on printing. Whereas I foresee many people preferring the classic style just out of habit, I would not be surprised to see **ggplot2** defaults edging its way into academic publications in the future. Meanwhile, **ggplot2** also includes the possibility of reverting the default appearance of each plot – via themes – to a more classic look.

Chapter 9 is an introduction to manipulating data. Many books assume that data are laid in a convenient shape so the analysis may start without any type of data preparation. However, as we all know very well, a lot of time typically goes into massaging the data in preparation for analysis. Consequently, this chapter gives an introduction to tools for data manipulation that work very conveniently in combination with **ggplot2**. There are functions for summarizing subsets of data, for reshaping time series data and for working with statistical objects.

Chapter 10 is a short chapter that introduces still more advanced techniques for **ggplot2** such as iteration, templates and how to create a function that will draw the same plot with different datasets. Finally, three appendixes give even more details about the links between the different syntaxes mentioned at the book – `qplot`, `ggplot`, and the grammar of graphics language (GGL) – and other technical details

It is my opinion that the best documentation for software is written sometimes by those that did not develop the software themselves. Commercial software companies, for example, usually hire professional writers for documenting the software they sell. For statistical software, writers are usually statisticians with the time and the cognitive distance necessary for focusing on only this part of the task. Programming and writing are both very demanding tasks and splitting up among different people should usually pay off.

ggplot2, the software, is clearly a very exciting development that has the potential of becoming one of the most used packages in R. However, **ggplot2**, the book, is not as good as the software. Thus, while certainly worth of being read at least one time, I found myself using the reference manual available at the internet site as the primary resource instead of using the book. However, as this was also rather insufficient, I got involved often in several trial and error loops before attaining satisfactory results. I am quite sure that **ggplot2** has the potential of becoming a very succesful package, and, as a consequence, I am also sure that other books will

be written in the future that will describe its power in a more transparent way. Meanwhile, however, I do recommend this book as the best current source for this otherwise exciting package for R.

References

Cox N (2007). "The Grammar of Graphics." *Journal of Statistical Software, Book Reviews*, **17**(3), 1–7. URL http://www.jstatsoft.org/v17/b03/.

Wilkinson L (2005). *The Grammar of Graphics*. 2nd edition. Springer-Verlag, New York.

Reviewer:

Pedro Valero-Mora
University of Valencia
Department of Methodology in the Behavioural Sciences
Valencia, Spain 46022
E-mail: valerop@uv.es
URL: http://www.uv.es/valerop/