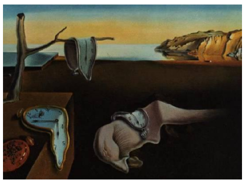

Journal of Statistical Software

April 2017, Volume 77, Book Review 2. *doi: 10.18637/jss.v077.b02*

Reviewer: Virgilio Gómez-Rubio
Universidad de Castilla-La Mancha

ggplot2 – Elegant Graphics for Data Analysis (2nd Edition)

Hadley Wickham
Springer-Verlag, New York, 2016.
ISBN 978-3-319-24275-0. 260 pp. EUR 51.99 (S).
http://www.springer.com/gp/book/9783319242750

ggplot2 is one of the main R packages for graphics and visualization of statistical data. The book has been written by Hadley Wickham, a popular and prolific R developer who has created a good number of widely used R packages for data analysis and visualization, including the **ggplot2** package.

The book is divided into three main sections: Getting started, the Grammar and Data analysis. The first section covers the basics of **ggplot2** and shows how to produce standard statistical plots. The second section focuses on the underlying 'grammar of graphics' (Wilkinson 2005), upon which the **ggplot2** package is structured. Finally, the last section discusses data preprocessing for analysis and visualization using several other packages from the *tidyverse* (https://blog.rstudio.org/2016/09/15/tidyverse-1-0-0/).

Chapter 1 is a short introduction to the book. Here, the author introduces the grammar of graphics (Wilkinson 2005), which outlines the main components of a plot, such as data, geometric objects (points, lines, etc.), scales, etc. I believe that by making these elements explicit it is easier to think about graphics and how to display statistical information in a plot. At the end of the chapter the reader will find a summary of the main differences between **ggplot2** and other R graphics packages.

The first examples on **ggplot2** are in Chapter 2, where the main types of plots are described. The author emphasizes three key components in a plot: data, aesthetic mappings and other layers. Explaining how plots are conceived in **ggplot2** and the grammar of graphics is one of the strongest points of the book, as this way of regarding plots can be taken when producing graphics with other plotting frameworks.

Chapter 3 goes deeper into some of the ideas of building a layered plot. In particular, it covers adding statistical summaries of the data and other metadata to the plot. This chapter builds on the previous one and provides a comprehensive description of the different options to create plots with **ggplot2**. In addition to describing how to customize labels and annotations, this chapter covers how to display spatial and temporal data. Adding statistical summaries, such as density estimates, histograms or binned data, is also discussed in this chapter. Finally, a few remarks on overplotting are given to avoid placing too many elements in a plot and create plots that are easier to understand.

The grammar of graphics (Wilkinson 2005) is described in more detail in Chapters 4. Here, the book goes through all the different elements and layers required to produce a plot with **ggplot2** and introduces was it to come in the next chapters.

The layered mechanism that is built in **ggplot2** is fully described in Chapter 5. In particular, this chapter emphasizes five components that are important when producing a plot: data, aesthetic mappings, geometric objects, statistical transformations and positioning of different elements. Each component is discussed separately, and some of them are described in more detail later in the book.

Chapter 6 deals with scales, axes and legends. Many options to set these three elements of the plot for different types of data are discussed. I have found the discussion about how **ggplot2** handles colors very enlightening. The author discusses the HCL color space, based on hue, chroma and luminance, as opposed to the RGB (red, green, blue) system. He also gives many tips on how to design palettes for continuous and discrete data, and highlights the importance of creating palettes with color blindness in mind.

Positioning of the different elements in a **ggplot2** plot are covered in Chapter 7. This is particularly useful when using faceted plots (i.e., conditional plots or trellis plots).

Chapter 8 concludes the second section of the book with a description of the theme system in **ggplot2**. This is useful to produce plots that have a similar look as well as to producing plots for printed publications. At the end of the chapter the reader will find a few remarks about the best way to save plots produced with **ggplot2**.

The last section of the book deals with data handling to organize them in a structure suitable for plotting. Chapter 9 focuses on the **tidyr** package, that includes functions or 'verbs' to reorganize information in a `data.frame` to make it 'tidy'. Shortly, a data is 'tidy' if variables are in columns and observations by row.

Chapter 10 introduces another package in the 'tidyverse': the **dplyr** package. This package allows the user to manipulate an existing `data.frame` in a number of ways. In particular, filtering, creating new variables and providing group summaries are tackled in this chapter. The pipe operator `%>%`, from the **magrittr** package, is discussed at the end of the chapter for creating pipelines to perform several operations on a data set.

A short introduction to data modeling is in Chapter 11. Here, the author describes how to use the **broom** package to obtain summary statistics from linear models that can be added to plots created with **ggplot2**. This is particularly interesting as these summaries can be included in the plot to annotate it, provide an insight of trends in the data or visualize models.

Finally, Chapter 12 gives an overview of how to program new components for **ggplot2**. This is particularly useful to create new types of visualizations using the grammar of graphics and that provide extensions the basic `geoms` in **ggplot2**. Those users that are interested in creating new visualizations for a particular type of data or to extend the basic types will find this chapter helpful.

The book contains a plethora of code chunks and plots so that the reader can easily understand the options and subtleties of creating plots with **ggplot2**. The book includes exercises in all chapters, so that the reader can practice. The exercises are usually placed together with the contents they try to reinforce and they are a valuable resource for teaching (and learning).

However, I usually prefer to place exercises at the end of the chapter so that they will not interfere with the main contents.

In a nutshell, this book does a great job at introducing and describing the **ggplot2** package to produce graphics and this is obviously the main reference. It will serve both the occasional user and more advanced users with a need to produce customized plots for their own packages. In addition, I also found the hints about how to organize the different elements in a plot quite helpful. I believe that most of these ideas can be taken when producing graphics using any other plotting framework in R or any other statistical software. Not surprisingly, this book is listed among Springer-Verlag's 100 bestselling books among all topics (https://www.springer.com/gp/booksellers/bestsellers).

Finally, printed copies of the book can be obtained from the publisher but the book is available online at https://github.com/hadley/ggplot2-book, where users can contribute and help to fix typos and bugs in the code.

References

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

Reviewer:

Virgilio Gómez-Rubio
Department of Mathematics
Universidad de Castilla-La Mancha
Avda. España s/n
02071 Albacete, Spain
E-mail: Virgilio.Gomez@uclm.es
URL: http://www.uclm.es/profesorado/vgomez/

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
April 2017, Volume 77, Book Review 2 *Published:* 2017-04-03
doi:10.18637/jss.v077.b02