Journal of Statistical Software

May 2022, Volume 102, Book Review 1. *doi: 10.18637/jss.v102.b01*

Reviewer: Ulrike Grömping
Berliner Hochschule für Technik

Pro Data Visualization Using R and JavaScript: Analyze and Visualize Key Data on the Web

Tom Barker, Jon Westfall
Apress, Berkeley, 2022.
ISBN 978-1-4842-7201-5 (paperback), 978-1-4842-7202-2 (e). 274 pp. EUR 48.14 (P), EUR 36.99 (e).
https://doi.org/10.1007/978-1-4842-7202-2

The book "Pro Data Visualization using R and JavaScript: Analyze and Visualize Key Data on the Web" by Barker and Westfall was a spur of the moment "deal" purchase following a promotion e-mail by Springer. It caught my attention, because I am about to teach another Data Visualization class, and I intend to increase my grasp on the popular JavaScript library **D3.js** (short: **D3**) by Mike Bostock for that purpose. The book is the second edition of a 2013 book by Tom Barker alone that came without the subtitle. For my personal background, the book's use of R instead of Python for analysis purposes is a great plus. This was my motivation for deciding to buy the ebook on a Friday evening with the intention of spending the weekend learning how to use **D3**. The book does not specify a target audience – we will come back to that.

The first chapter gives a very high-level introduction: it exemplifies basic types of visualization, covers well-known historic examples, and mentions some aspects of what the book calls the "modern landscape" of data visualization. It also covers the "why" of data visualization, gives a high level overview of the tools and the process, and discusses ethical aspects. Chapters 2 and 3 cover R, Chapter 4 introduces **D3**, including the very basics of JavaScript and information on SVG. Chapters 5 to 9 cover specific types of visualization in more detail, introducing related R and **D3** functionality on the way. The book is accompanied by a download page on GitHub. Locating that page (https://github.com/Apress/pro-data-visualization-r-javascript) was challenging; it took a phone call to Springer-Verlag in Heidelberg, who gave me contact data for someone at Apress who could provide the URL. This difficulty is an adverse consequence from the takeover of Apress by Springer combined with a redesign of the Springer website. Springer has been in the process of migrating materials for a long time (and does not handle that process in a user-friendly way). At the time of writing this review, there is still no hint on the book's homepage regarding the whereabouts of download material. The download material itself provides code and resources for Chapters 2 to 7. For the last two chapters, I am glad that I bought the ebook, so that I can copy-paste the code from the book.

I discuss the book chapters in the order I worked through them. Since I wanted to learn **D3** and trusted that my R would be sufficient, I started with Chapter 4 (Data Visualization with **D3**). Based on my understanding of HTML, and very basic knowledge of CSS, I found that the book allowed me to get started; however, some steps were left out. For example, I did not understand how to install **D3**, because I am more or less a JavaScript dummy. Hence, after downloading the tarball (which should have been a `zip` file according to a screenshot in the book), the only sentence "After that is installed, you can set up a project." left me in the dark. I overcame this obstacle by searching the web and referencing an online **D3** version instead of using a local installation. It is of course acknowledged that this type of information ages fast, but I would nevertheless have appreciated more information on handling this step. (After I got access to the book's download page, **D3** installation was no longer relevant, because the necessary **D3** files are available from there, although the book does not point readers to the download page for that purpose.). The chapter nicely builds up the necessary knowledge step by step. It finishes with a worked example that creates an HTML page with a small bar chart. Like in all subsequent **D3** examples, the code is presented in separate chunks related to aspects of chart creation, explaining the role of each chunk in interspersed paragraphs of text. The complete code is printed at the end of the chapter. The barchart from the four numbers 84, 62, 40, 109 (entered as an array in the script) is a simple technical example. For someone who is familiar with bar charts, it feels quite weird (for me to the point of being confusing) that the book initially places each bar at a horizontal position determined by its height instead of the usual equidistant positions; it would have been helpful for me if the book had initially explained that this will be rectified later on. The chapter proceeds by loading the data from a file rather than entering them as array in a script, which initially made the browser throw an error, for security reasons. While the book anticipated this error, its recommendation for resolving it was "It's advised to use a local web server to work around this issue while programming". Some hints on how to do that would have been appreciated, though a internet search eventually revealed a solution for my Windows 10 machine. Reading this chapter gave me the impression that the target audience the authors had in mind has more prior knowledge on web development than I do.

Chapter 5 (Visualizing Spatial Data from Access Logs) starts with historic maps by Minard and links to some current maps on the web (e.g., `http://hint.fm/wind/`, a map of current wind flow in the US). It goes on to introduce the format of web access logs, before processing data from this format with PHP and R. Large parts of the chapter strike me as off-topic: there are 15 pages of PHP processing of access logs – while I of course ackowledge the importance of data wrangling as a prerequisite of visualization, the strong focus on PHP (which was mentioned nowhere in the introduction or on the book cover) puzzled me. For me, using the PHP code would have required to make yet another install, since my Windows 10 webserver apparently did not come with PHP; for an R person like me (and for a book with R in the title), it would have been more natural to process the access log files in R. Presumably, the target audience the book authors had in mind is familiar with PHP. As there is no **D3**, JavaScript or SVG to be learnt in the chapter, I simply moved on to the next chapter. For people interested in wrangling log data with PHP (or even with R), the chapter offers interesting bits of advice, like how to check regular expressions with a web tool (available at `https://regexr.com/`). Visualizing spatial data, the topic of the chapter, is shown with a step-by-step example of how to display the data on a map using R. Readers can learn R technicalities from that example. However, the resulting map is disappointing in terms of data visualization: while the globe is displayed with US, Canada and Australia colored in light blue, this display does not reveal any information that couldn't have been obtained better by a simple univariate frequency table.

Chapter 6 (Visualizing Data over Time) caters to the time series needs of IT project developers: bug tracking over time. Here, the book provides example data right at the beginning, processes and visualizes them in R and then uses **D3** for showing them with interactive elements in an HTML file. The use case for **D3** is to add tooltips to time series charts. This could be done from within R, e.g., by using **ggplot2** in combination with **ggiraph**. But I want to learn how to use **D3**. From this chapter onwards, the book uses an older **D3** version, because version 3 of **D3** "allowed for a bit more flexibility and step-by-step building"; for obtaining that version, readers are referred to the book's download page. The chosen visualization makes sense for the chosen application: daily bug counts are presented in a time series, and a mouseover event shows the bug identifiers for the selected day.

The example for Chapters 7 and 8 is taken from agile product development: Chapter 7 (Bar Charts) analyses production incidents and Chapter 8 (Correlation Analysis with Scatter Plots) development velocity. Chapter 7 introduces standard bar charts, stacked bar charts and grouped bar charts. Subsequently, it provides standard and stacked bar charts in R, after adequately preprocessing the data. Afterwards, a very simple bar chart with interactivity is created with **D3**: two bars are subdivided into stacked bars on mouseover. Chapter 8 covers scatter plots and bubble charts. Again, these are first created in R before presenting an interactive web chart. The R visualizations themselves are not exactly examples: Arguably, the axes of the scatter plot should be swapped, and it does not appear adequate to encode a binary qualitative feature by bubble size (why not, e.g., color?). The interactive web chart has two dropdowns for selecting the features for the two axes; these are created by JavaScript. Chapter 9 (Visualizing the Balance of Delivery and Quality with Parallel Coordinates) initially explains parallel coordinate plots using a built-in dataset from R; it also explains using grey shades for a third feature. The chapter continues by creating parallel coordinate plots for the agile product development data from the previous two chapters, as well as for an additional effort data set that is provided as a printout in the book. The authors emphasize that the purpose of their use of parallel coordinates lies in looking at the actual data points for each development team and comparing the two plots. The subsequent part of the chapter creates a brushable **D3** parallel coordinate plot, for which a mouseover fades all lines except the selected one, which of course makes sense for the stated purpose.

Let us also have a brief look at Chapters 2 and 3, for which I am an expert rather than a learner. Chapter 2 (R language primer) introduces the R console on a Mac and basic commands like installing a package, ways to read data and the most basic data structures, including matrices and data frames, but not lists. The term "list" is used, but in a non-R way, which I find a bit unfortunate. The book does not introduce simple element-wise matrix operations, but it does introduce the `apply` function, using the example of adding one to each element of a numeric matrix `m` with the expression `apply(m, 2, function(x) x <- x+1)`. Presenting this complicated solution without even hinting at the much simpler `m+1` as an alternative seems inadequate for readers who are not familiar with R. The good use cases for the `apply` function are beyond this chapter, except for the application of a scalar function to each element of a matrix, for which the book also gives an example. Chapter 3 (A deeper dive into R) introduces the use of classes (S3 and S4), and continues to present the `summary` and `sd` functions and `sapply` for obtaining the standard deviation for all columns of a data frame with numeric variables. The chapter finishes with an introduction to **RStudio**, **rmarkdown** and Rpubs. R content regarding the creation of graphics is not covered here, but is provided in Chapters 5 to 9.

So, what do I think of this book? Let me start with what I like: The book is very hands-on. Building up the code step by step, with explanations of the code's purpose, helps the reader to understand the final code and encourages experimentation, and there is always a data example to which the code is applied. Thus, there are positive points regarding the teaching of **D3** coding. Unfortunately, most chapters use a very old **D3** version, which is a severe problem with **D3**, because it does not even attempt to be backwards compatible. Readers who do not know any R are presented with a somewhat strange selection of R features, in my view. But it may well be that they can learn some R from the book, if they combine it with web sources. However, be warned that the "About the book" text on the book homepage heavily oversells the book's qualities: I nowhere found anything that could be related to "work with the R **plumber** API generator, **shiny**, and more", and some other claims are also a bit exaggerated. The real downside of the book are the visualizations: the entire book does not contain one really good visualization (apart from the aforementioned external examples). In many cases very simple changes would have made for substantial improvements. For example, in Chapter 6, the R code for visualizations could be improved by taking care of the order of factor levels (where the default alphabetic order does not correspond to the ordinal severity variable), or by placing the legend on white space instead of having it hide part of a line. In summary, I would not recommend the book to other statisticians with my inclination of learning how to use **D3**, and I would neither recommend it to anyone who wants to learn about decent visualizations.

Reviewer:

Ulrike Grömping
Berliner Hochschule für Technik
Department II
D-13353 Berlin, Germany
E-mail: ulrike.groemping@bht-berlin.de
URL: http://prof.bht-berlin.de/groemping/

Journal of Statistical Software https://www.jstatsoft.org/
published by the Foundation for Open Access Statistics https://www.foastat.org/
May 2022, Volume 102, Book Review 1 *Published:* 2022-05-25
doi:10.18637/jss.v102.b01