Journal of Statistical Software

September 2018, Volume 86, Book Review 3. *doi: 10.18637/jss.v086.b03*

Reviewer: Ulrike Grömping
Beuth University of Applied Sciences Berlin

A Data Scientist's Guide to Acquiring, Cleaning and Managing Data in R

Samuel E. Buttrey, Lyn R. Whitaker
John Wiley & Sons, Hoboken, 2018.
ISBN 978-1-119-08002-2. 309 pp. USD 85.00 (H), USD 68.99 (e).
http://www.wiley.com/go/buttrey/datascientistsguide

"A Data Scientist's Guide to Acquiring, Cleaning and Managing Data in R" covers topics that have always been very important and time-consuming prerequisites of statistical data analysis, even before the term "Data Science" was coined. In the following, I will use "data handling" as a shorthand expression for the book's topics. The book's cover rightly states both tediousness and importance of data handling, as well as a lack of systematic education on the topic for many modelers; it promises no less than the "only how-to guide offering a unified, systematic approach to acquiring, cleaning, and managing data in R". I decided to review this book, because I currently prepare a lecture in which I, for the first time, want to teach efficient techniques for data handling – so far, my focus in teaching has been on R for statistical analysis, taking the usual nice and pre-cleaned data sets for examples, and I am one of those people who lack systematic training on data handling.

The book does not make any statements about the target audience, and even after reading through all of it I am not entirely sure for whom it is written. From the slow pace taken in the first few chapters, it seems to target R beginners, who also have little experience with data analysis; however, the book does not attempt to also provide a general introduction into R – which is good, because that would have to be over-ambitious for a small book like this one. While advanced R users will find a lot of information they know already, they can nevertheless benefit from the clear focus on data handling tasks, and from numerous details. For example, Section 6.1.4 provides an example, where reading tabular data fails because of an embedded delimiter, and explains steps to resolve this nasty issue. R beginners can also benefit from this book, though they will have to resort to other material in addition.

The book starts with an introductory chapter on R itself (Chaper 1, 19 pages), which describes specifics of R, gives hints on where beginners can find useful introductory material, provides a glossary of R related terms used in the book and gives recommendations on how to use the book. Chapters 2 to 4 introduce R data structures, starting with vectors (Chapter 2, 31 pages), proceeding to matrices, lists and data frames (Chapter 3, 45 pages) and finishing with text and factors (Chapter 4, 45 pages). Subsequently, the writing of functions and scripts (Chapter 5, 27 pages) and importing and exporting data (Chapter 6, 31 pages) are discussed.

Chapter 7 (33 pages) provides an extended practical example, which discusses the entire data handling process in broad context; aspects discussed include documentation, and also aspects not directly related to code, such as communication with data providers. The final chapter (18 pages) provides an extended exercise; solution hints for this exercise are given in the appendix. A list of references (many of which are for R packages) and an index complete the book. Each of the first seven chapters closes with a section called "Chapter Summary and Critical Data Handling Tools". These will be quite helpful for most readers: for R beginners, they ensure that the overall flow is not lost in the details, and for advanced readers, these may help to decide whether to read or to skip a chapter. Another helpful feature of the book: it provides various lists of common errors and pitfalls that are instructive especially for readers who are new to R or to handling data in R.

Many small code examples, scattered throughout the book, cover a lot of technical detail. The book authors are very experienced with handling data, and many of the code examples are presented as collected experiences, rather than as a systematic presentation of tools. Further topics are discussed on a high level only, which leaves users to have to work out specifics for themselves. For example, the introduction claims that XML and JSON are described in the book, and that it is described how data are acquired programmatically from web pages. I was looking forward to learning about these topics; however, the section on XML has less than two pages and can of course not provide any in-depth knowledge on XML, but remains very high-level. There is an example on reading websites with function `readHTMLTable` from package **XML**, which helped me to quickly obtain an example for my class; however, if an HTML table is more complex, e.g., with cells spanning several rows because several rows have the same value for a certain column, the result of that function becomes a mess; a web search led me to the small package **htmltab** whose function `htmltab` was able to read a much more complex table; such a hint would be a nice addition for the next edition of the book. Nevertheless, the book's example was a useful starting point. Likewise, for reading web data from an interactive form, using a documented API for which there is no custom R package, the book's example applies function `GET` from package **httr** to an example use case from the API documentation of a US government website for international trade data. Again, the example served as a starting point into exploring the topic; besides **httr**, the book mentions package **Rcurl** for the same purpose; the CRAN task view on "Web Technologies and Services" and other internet sources provided an impression of a larger field of R packages suitable for acquiring such data.

In general, the book is well-written and contains fewer typos or mistakes than many other first editions. Nevertheless, there is the odd mistake: for example, the book claims that the first element on R's search path is the "current `.RData` file (although it carries the confusing name `.GlobalEnv`)". This is at least confusing itself. Also, the provision of code and data could have been handled in a more professional way: the book claims that there is an R package **cleaningBook** on the Comprehensive R Archive Network (CRAN, p. 19), which is not the case. Instead, the book's website provides the data for Chapter 7 and data and code for the book's examples, including solution code for the extended exercise of Chapter 8; however, the code is very inconvenient to use, because the continuation lines of all multi-line R statements start with a comment character. It can be hoped that the book's website will be maintained, e.g., by uploading improved code files, or by providing errata/updates. Updates are particularly useful, where examples rely on web content: for example, even after the brief time period since the book was written, the aforementioned API for international trade data on the US government website has already changed, so that the code for this example requires changing.

As mentioned before, the book's content is clearly focused on data handling tasks. This implies that chapter content sometimes differs from what I would have expected. For example, the section called "R Data, Part 1: Vectors" contains parts on tables, vectors as sets, finding duplicates or identification of run lengths (function `rle`), since these help in data handling tasks, if the data comes as vectors. Structure-wise, I am not always happy with the authors' choices; for example, it would seem more natural to me to discuss parsing commands or retrieving / assigning objects by a calculated name in the chapter "Writing Functions and Scripts", rather than in the chapter called "R Data, Part 3: Text and Factors". Of course, it may be quite subjective what is considered "natural"; luckily, there is an index so that one can locate things that one remembers to have seen somewhere, as long as one can think of a suitable search expression. Some structuring tool, similar to a flow chart or "cheat sheet", might come in handy, however; this might be another potential improvement for a second edition.

In spite of the many small code examples, there are only few proper examples, and no exercises except for the last chapter. Both the small code examples and the proper examples are quite helpful. In particular, the extended practical example in Chapter 7 is illustrative. While reading it, I wished that the bigger picture provided with this example would have been explained earlier in the book. However, I understand the authors' decision to place it close to the end, because it is very convenient to be able to rely on all the content of the earlier chapters in the presentation of the example. I did not work on the extended exercise; I imagine that users who do invest the time will learn a lot about data handling with R through that experience.

"A Data Scientist's Guide to Acquiring, Cleaning and Managing Data in R" has been written by two statisticians who were familiar with R long before the relatively recent omnipresence of data science, which brought along R packages with a strong focus on data handling, including the so-called tidyverse or the package **data.table**. The book mainly presents classical R functionality and only rarely refers to the more recent tools (e.g., for **httr**). I appreciate the classical focus of this book. Nevertheless, in some cases, inclusion of additional packages would have been helpful. For example, package **lubridate** would have been worth mentioning in the section on date and time data, or function `fread` of package **data.table** for convenient and fast reading of large rectangular data files. There is one other thing I missed: in my opinion, literate programming (e.g., with packages like `Sweave` or **rmarkdown**) should have been included in the recommendations for achieving reproducibility of data handling.

Overall, the book was worth the time spent with it. I recommend it particularly for people like me who in principle know their R but have never spent very much thought on the data handling side of it. It will also be a good read for R beginners who want to benefit from the authors' treasure trove of experience and their view on the big picture of data handling in context. However, in spite of the claim on the book's cover, one should not expect an entirely systematic approach to data handling.

Reviewer:

Ulrike Grömping
Beuth University of Applied Sciences Berlin

Department II
D-13353 Berlin, Germany
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/
September 2018, Volume 86, Book Review 3 *Published:* 2018-09-30
doi:10.18637/jss.v086.b03