Journal of Statistical Software

April 2017, Volume 77, Book Review 1. doi: 10.18637/jss.v077.b01

Reviewer: Christopher J. Lortie
York University and NCEAS

R for Data Science

Hadley Wickham, Garrett Grolemund
O'Reilly, Canada, 2016.
ISBN 978-1-4919-1039-9. 522 pp. USD 39.11 (P).
http://r4ds.had.co.nz/

Data science is a complex domain, and decisions associated with wrangling big and little data are non-trivial (Gandomi and Haider 2015; Peters, Havstad, Cushing, Tweedie, Fuentes, and Villanueva-Rosales 2014; Marx 2013). This book is written as a general resource for R by providing a complete data science workflow, i.e., a set of steps for specific packages. The workflow or set of steps is the anchor for the book and is developed immediately within the preface. Import, tidy, transform-visualize-model iteratively, followed by communicate. The workflow is described in text, illustrated, and sets of chapters are linked to the workflow throughout the book. This structure provides an excellent backbone to the content and facilitates its use as a resource because one can easily revisit a specific chapter for reference when working through a real problem. The meaning of each step is self-explanatory but nonetheless well defined and demonstrated by worked examples throughout the book.

The efficacy of communication for this statistical software and implementation book was evaluated using the following criteria: clarity of writing, supporting visuals that make complex data science concepts accessible, and an appropriate balance between detail and general understanding of process. 'R for Data Science' was successful in all three potential dimensions of communication. The writing is direct. Most chapters lead with code, examples, then the description follows. This exposes the reader more rapidly to the relevant material needed to grasp and do the data science. The book is primarily written in a show-then-tell format, and this approach reduces the need for the reader to process large chunks of description (introductions are very brief in each chapter). Telling one how to do something versus showing it directly can of course be appropriate in some contexts, and readers have different learning styles. Nonetheless, showing the data science first engages and challenges the reader to read the R code and learn the grammar. Reading code others have written is an important skill and considering a problem before seeing the solution stimulates deeper learning. If anything, there could have been even more development of the problem-solution model in the writing, but I recognize that this can sometimes come at the cost of clarity and can tax the patience of readers at different levels. There are exercises provided to consolidate learning and they are pitched at the right level consistent with each chapter. The supporting visuals excel (but not

Excel, pardon the pun) at visualizing the layered grammar of graphics in **ggplot2**, relational data with **dplyr**, and subsetting with vectors. Visual learners will appreciate the concepts illustrated, use of color, and a certain to be favorite – the pepper shaker, with pepper packet in it, with pepper in the packet – to illustrate subsetting of lists of lists. Most chapters balance detail and general understanding of process well. This it not to say that the details of coding were never a challenge to reconcile with the big picture. Many data science and coding concepts are complex. The 'Iteration with **purr**' chapter was a challenge in merging and contrasting the details between different options such as 'for loops' versus functionals. However, later chapters such as those in the model section struck a better balance. This difference can in part be due to an audience experience bias such as one's background in statistics versus data science. This suggests that different audiences will be able to better capitalize on the show-then-tell approach depending on their experience. The book is thus well pitched for beginner to intermediate data scientists and likely for statisticians with an intermediate level of experience with data science concepts and approaches. The communication and writing style is accessible and not unduly technical for all readers.

There is extensive support for R available in the form of documentation (documentation for R directly and reference manuals and vignettes for CRAN packages), FAQs, StackOverflow, blogs, webinars, workshops, and many books (and many are also free). Too much information, not too little is most likely the challenge for data scientists and statisticians working in R. For the R community in particular, the breadth and scope of packages, discussion, and documentation are unparalleled. Typically, this is a benefit in solving a problem, and frequently, there is no one single solution but many. However, processing and parsing responses, solutions, and code from different sources is time consuming and, at times, overwhelming. 'R for Data Science' is a logical, contemporary entry point that compiles a relatively consistent set of current R packages together into a clean data science workflow appropriate for many purposes. The book is built up from extensive package development, and both R and its packages will continue to evolve. The book reframes and updates a **ggplot2** book (Wickham 2009) and complements the updated book (Wickham 2016). It explains the philosophy and grammar of this package succinctly. It also further develops the concept of 'tidydata' (i.e., columns as variables, rows as observations, Wickham 2014). The concept of this mapping of data is not unique to the 'tidyverse', but this ecosystem offers functions to easily deal with some frequent types of inconvenient data and to readily wrangle and specify what constitutes a variable and an observation so that the concept of tidydata makes sense. Tidydata thus set up dataframes for more efficient processing. This ecosystem of packages, its grammar, and the thinking are better situated within the domain of data science through the book. The novelty in this book is a coherent workflow across different concepts and packages. It is a solid foundation for the statistician interested in learning and improving data handling skills. For the data scientist versed in the extensive resources distributed online for R, it is an integrated set of resources and sample code that can readily provide and affirm a literate, reproducible philosophy of data science. It is not about efficient programming or coding in R, it is about efficient data science.

'R for Data Science' is an excellent resource. If you are already familiar with this ecosystem of packages and ideas, it is nonetheless still valuable. You may be reading about many of the approaches and tools you already use or have seen, but in seeing them organized and described, in many instances by the authors of the packages, one gains novel insights. Even if you do not agree with the assumptions in full, the documentation and logic described provides

a more complete sense of how data science needs, package development in R, and the goal of integration are useful for statistical languages. Open science development can rapidly provide us with new packages, but sometimes connecting and understanding them is a challenge. This book is thus an excellent example of the value of documentation beyond vignettes that facilitates deeper learning and appreciation of the landscape and not just the details of the moment. When using R, it is not uncommon to be in the midst of a problem, rapidly look up a solution online (from whatever resource works), and move on. The solution may or may not come from a book, and if it does, one captures the relevant code or explanation from the snippet only. This begs the question of investing in a complete book. For this book, I recommend the investment: time you enjoy wasting (on a technical book like this one) is not wasted.

References

Gandomi A, Haider M (2015). "Beyond the Hype: Big Data Concepts, Methods, and Analytics." *International Journal of Information Management*, **35**(2), 137–144. doi:10.1016/j.ijinfomgt.2014.10.007.

Marx V (2013). "Biology: The Big Challenges of Big Data." *Nature*, **498**(7453), 255–260. doi:10.1038/498255a.

Peters DPC, Havstad KM, Cushing J, Tweedie C, Fuentes O, Villanueva-Rosales N (2014). "Harnessing the Power of Big Data: Infusing the Scientific Method with Machine Learning to Transform Ecology." *Ecosphere*, **5**(6), 1–15. doi:10.1890/es13-00359.1.

Wickham H (2009). ***ggplot2**: Elegant Graphics for Data Analysis*. Springer-Verlag, New York.

Wickham H (2014). "Tidy Data." *Journal of Statistical Software*, **59**, 1–23. doi:10.18637/jss.v059.i10.

Wickham H (2016). ***ggplot2**: Elegant Graphics for Data Analysis*. 2nd edition. Springer-Verlag, New York.

Reviewer:

Christopher J. Lortie
York University and NCEAS
Biology
Toronto, Canada, M3J1P3
E-mail: lortie@yorku.ca
URL: http://www.christopherlortie.info/

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 1 *Published:* 2017-04-03
doi:10.18637/jss.v077.b01