Journal of Statistical Software

MMMMMM YYYY, Volume VV, Book Review II. doi: 10.18637/jss.v000.i00

Reviewer: Christopher J. Lortie
York University and NCEAS

Applied Time Series Analysis With R. Second Edition.

Wayne A. Woodward, Henry L. Gray, and Alan C. Elliott
CRC Press, USA, 2017.
ISBN 9781498734226. 522 pp. USD 109.95(P).
https://www.crcpress.com/Applied-Time-Series-Analysis-with-R-Second-Edition/Woodward-Gray-Elliott/p/book/9781498734226

Making time for time series

Introduce authors, time series and why important, and then explain the book.

The authors of this book are eminent data scientists within the R community particularly recognized for their recent work associated with **RStudio**. R is a statistical language and environment that functions independently of **RStudio** through support from the R Foundation and CRAN(The Comprehensive R Archive Network) to host packages, and it is licensed as free software. **RStudio** is an open-source development environment for R with a free desktop deployment but also has pro, server, and other options as paid products. The first author Hadley Wickham is the Chief Scientist at **RStudio**, and the second author Garrett Grolemund maintains shiny apps and focuses extensively on education for this organization. There are at least two major sources of consistent package development for R - **RStudio** and rOpenSci foundation. That said, there are 12,412 packages now available on CRAN for R (and more available directly from authors often via GitHub), but the developments forwarded both by **RStudio** and rOpenSci are unique in some respects because each have aligned sets of packages to support their respective missions. At rOpenSci, the packages are primarily associated with acquiring open data and handling it; whilst at **RStudio**, the package sets are primarily built for data wrangling and now reporting using its environment. Both have developed much, much more, but the purpose of the most commonly used packages differs (but are are not exclusive). All this is to say that the book is written as a general tool for better R use for a complete data science workflow but via specific packages and their respective grammar. The assumption of the authors is nonetheless that you will work in RStudio, and the book is written to take advantage of that specific environment and many of the packages are from RStudio affiliates. There is a tip icon interspersed throughout the book often associated with short-cuts for **RStudio**, but there are worked data science examples that do not have to be run in **RStudio**. However, the final section of the book entitled 'Communicate' is written for **RStudio** and describes how to use **RMarkdown** to effectively communicate code, data visualization, and reporting using this environment. It is wise to be cognizant of the fact that the

workflow and thinking described are designed to take advantage of the structure, logic, and function of the specific packages covered here. This is an appropriate and powerful set of tools for the data scientist, and the organization of the book appropriately reflects these targeted strengths. 'R for Data Science' thus describes a data science ecosystem. It is an incredibly useful ecosystem to consider because it provides a consistent and well-supported worflow. The workflow shifted my paradigm from data wrangling, statistics, then data visualization to data wrangling, data visualization, exploratory data analyses, then model fitting. The book provides a coherent and rapid mechanism to assess whether one should occupy this idea and code space for their statistical software needs.

Theory
A review of the theory

Code and case studies
A critical analysis of a statistical sofware book can address a wide range of questions from generic to programming language or software specific. There are at least three critical data science/general statistical questions that emerged in reviewing 'R for Data Science'.

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). Data science can provide data thinking tools (Baumer 2015). A data thinking tool can be the heuristic, semantics, and concepts needed to work with data. **Does this book (or any data science book for that matter) effectively communicate basic versus advanced data science concepts to the reader?** Criteria can include any of the following attributes: 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 followed. 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 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 excelled (but not Excel) 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. 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. My primary experience is in statistics and not data science. Consequently, some of the data science concepts were more challenging to grasp and link to higher-order ideas whilst model fitting was not. Some data science concepts align more readily with statistical worflows and semantics. 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 to statisticians with an intermediate level of experience. 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). **Does this book extend or improve upon previous resources particularly for the individual interested in using and learning data science to do statistics in R?** This is a facile question to address. Too much information, not too little is most likely the challenge for data scientists and statisticians whom use R face. 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, for now, that compiles a relatively consistent set of R contemporary packages together into a clean data science workflow appropriate for many purposes. This book is built up from extensive package development, and both R and its packages will continue to evolve. This book reframes and updates a ggplot2 book Wickham (2009) (that is due as a second edition May 2017 in print), and compiles documentation associated with tidydata that was not that extensive. This book does significantly advance the ecosystem of packages, its grammar, and the thinking into the domain of data science. 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 a rapid compiled set of resources and sample code that can 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. This book introduced me to a data science ecosystem that I did not fully appreciate in using individual packages and solving challenging as they emerged organically in my research. It also better prepared me for using contemporary R packages, and I hope ultimately do better statistics because I have a workflow that can support transformation, iteration, and commmunication with my peers.

There is no need to set up R versus **RStudio** as a dichotomy. One can work directly in both or the other. Nonetheless, there is an RStudio signal to the book, and this leads to the following question. **Can this book be read as a general data science book and by extension how much is this an R versus Rstudio book?** This book uses R packages, code, and associated grammar and logic to build a data science workflow. It teaches data science

through R and is thus best read by those specifically wanting to learn data science in R. This is obvious given the title but not necessarily trivial. Fluency in other languages common in data vizualization and statistics such as Python is important, and one can cursorily review books outside primary software language to assess strengths and limitations of current skills. Some sections of this book can be read by the generalist not interested in R, but this book uses R to teach data science. The purpose is not to advance data science theory explicitly but to highlight the tools that R provides in solving data challenges. The answer to the corollary question, R versus **RStudio**, is both. There are numerous subsections interspersed throughout the book contrasting specific packages to base R and explanations are also provided on how to interact with older and other code. Some of the strength of **RStudio** are highlighted within the context of data science, but it is not a gratitous product-placement scenario. This book can advance your competency in R coding and certainly advances an appreciation of the flexibility of this language through packages in tackling a wide-array of data science challenges and beyond. The book ends with an **RMarkdown** workflow - not with a bang but a reminder.

Conclusions

'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 is 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. It is not uncommon to be in the midst of a problem, rapidly look up a solution online, and move on. Time you enjoy wasting (on a technical book like this one), was not wasted.

References

Baumer B (2015). "A Data Science Course For Undergraduates: Thinking With Data." *The American Statistician*, **69**(4), 334–342. ISSN 0003-1305. doi:10.1080/00031305.2015.1081105. URL http://dx.doi.org/10.1080/00031305.2015.1081105.

Gandomi A, Haider M (2015). "Beyond the hype: Big data concepts, methods, and analytics." *International Journal of Information Management*, **35**(2), 137–144. ISSN 0268-4012. doi:http://dx.doi.org/10.1016/j.ijinfomgt.2014.10.007. URL http://www.sciencedirect.com/science/article/pii/S0268401214001066.

Marx V (2013). "Biology: The big challenges of big data." *Nature*, **498**(7453), 255–260. ISSN 0028-0836. doi:10.1038/498255a. URL http://dx.doi.org/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. ISSN 2150-8925. doi:10.1890/ES13-00359.1. URL http://dx.doi.org/10.1890/ES13-00359.1.

Wickham H (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-0-387-98140-6. URL http://ggplot2.org.

Reviewer:

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