Journal of Statistical Software

July 2022, Volume 103, Book Review 2. *doi: 10.18637/jss.v103.b02*

Reviewer: Christopher J. Lortie
York University

Python and R for the Modern Data Scientist

Rick J. Scavetta, Bayan Angelov
O'Reilly Media, Sebastopol, 2021.
ISBN 978-1492093404. 198 pp. USD 39.49 (P).
https://www.oreilly.com/library/view/python-and-r/9781492093398/

Context

Computation in many fields including those that use statistical software is increasingly driven by needs that can be addressed in many programming ecosystems. In projects that require statistical analyses, both R and Python comprise two frequent resources. In ecology, R is the most frequently used (Lai, Lortie, Muenchen, Yang, and Ma 2019). In bioinformatic gene set analyses, R is also more frequently used in peer-reviewed publications, but Python is still an important statistical resource depending on the specific project (Xie, Jauhari, and Mora 2021). Python outcompetes other languages in use for machine learning and some forms of factor analyses (Hao and Ho 2019; Persson and Khojasteh 2021; Raschka, Patterson, and Nolet 2020). However, the relative frequency that a tool is used for statistical analyses is only one metric of importance and not necessarily a proxy for its merit or its capacity to support innovation and efficient in analyses for practitioners (Zhao, Yan, and Li 2018). It is thus critical that we explore contrasts of at least these two common software languages that support statistics because data scientists can become isolated or polarized within their specific competencies, ideologies, and workflows. A high-level discussion of strengths and weaknesses specific to data endeavors with statistics is germane to both decisions on specific projects and on competency development as a scientist.

Content

The need for a balanced, informed discussion of R and Python to support basic data science endeavors for statistics was recently published (Scavetta and Angelov 2021). This is the first edition of this book, and the primary focus is to highlight the 'best of both worlds' in these two languages for a diversity of statistical challenges. It is assumed that the reader has basic to intermediate competency in either R or Python but not both. That said, the principles, highlights, and clarity of this book support a wide audience of readers competent in general

programming software provided statistics is the main focus of a project. These principles and learning opportunities with explicit, structured contrasts are developed throughout the book in a total of 7 chapters organized into 4 sections.

The first section of the book provides context and history for both R and Python whilst the second section comprises a very clear presentation of each language for the reader familiar with the alternate tool. The third section offers a modern context for handling data in either ecosystem and proposes a workflow that can incorporate both depending on whether the purpose is exploratory data analysis, machine learning, data engineering, or reporting. This section is also a roadmap to enable choice in a few representative statistical instances. It is proposed that R is likely better for time series and spatial analyses whilst Python is likely more effective for machine learning and imagery data (Scavetta and Angelov 2021).

Clear criteria are developed for all analyses including the availability and quality of third-packages to augment the core features of each language. Worked examples with additional resources are provided in each instance, and parallel contrasts for each ecosystem provided where appropriate. This is very instructive and can support decision making for other projects. The final section provides a more in-depth discussion explanation of how interoperability between these two languages is possible if one elected to use both, and concludes with a thorough, sufficiently detailed example for a real-world dataset. The book also provides a 'bilingual' dictionary that shows how to do the same task in R and Python sorted into tables with embedded code chunks. This book uses differentiation in fonts to indicate new terms, denote program names, and show commands with code chunks. Small icon animal elements with textboxes also provide tips and suggestions, general notes to consider, and warning or cautions. There is also an online supplement on GitHub that includes code, examples, data, and exercises.

Critique

This is a highly readable, accessible textbook appropriate for the intended audience and beyond. The chapters are well written and organized. The overarching organization of the book is also logical and supports reading sequentially or use of specific chapters individually as needed. Terms and labels with titles and subtitles effectively organize the content and provide scaffolding to connect key concepts throughout the book (Jumaat and Tasir 2014). The examples and code chunks support learning and do not disrupt the flow in reading and logically reasoning through the arguments provided for a specific context. Cognitive overload is effectively avoided by clear goals, principles, and criteria provided as needed to enable reasonable contrasts at a high-level of abstraction (Ou, Henriques, Senthilnathan, Ke, Grainger, and Germain 2022). The semantics and examples for each programming language are well tuned to a wide audience of readers competent in either R or Python. A reader will not get 'lost in the weeds' with this book because of the clarity and organization of the content. The writing is also enjoyable, light, and includes fun facts and humor interspersed with challenges for the attentive reader.

The introduction to each language, its history, and how to use each is also invaluable - even for your preferred go-to choice - before engaging with the content that contrasts each tool. The history is also an ideal springboard for relatively new students to statistical programming (Auker and Barthelmess 2020). The concept of dialects and the incredible capacity for both R and Python to get job done through varied base language or package choices is also introduced

and well articulated. The code examples are very current and highlight the salient principles associated with different data structures, associated semantics, and style nuances for each ecosystem. Some of the individual cognitive benefits of how to think about problems based on dialect or language are provided (Scherer, Siddiq, and Sánchez-Scherer 2021), and the benefits to different workflow choices are proposed as well. The packages and statistical concepts used to support a diverse data scientist toolbox for statistics is impressive and interesting. The index provided in the book is in sufficient detail because it includes a list of all packages and statistical tests that enable lookups for subsequent use as a potential methodological resource.

Most importantly however, this book is a potential stand-out general offering that is relatively unique. There is an extensive set of writings in journals and books to support either R or Python but not both in parallel. In examining both in concert, a cogent argument is definitively supported for modern data scientists that use statistics to consider becoming fluent (to some extent) in more than one programming language. The definition of modern data science as collective, simple, accessible, generalizable, outward facing, and ethical is profound (Scavetta and Angelov 2021). Criteria for tools within this framework are also advanced including that it must be open source, feature-complete, and well maintained. This paradox of choice for solution sets from multiple, related package offerings is non-trivial (Lortie, Braun, Filazzola, and Miguel 2020), and this book provides a functional framework for choice between dialects, dataset formats, languages, and workflows. This comprehensive yet accessible paradigm challenges many assumptions about simply getting the work done, once, at the potential expense of replication science and collaboration. This echos the repeated call for better coding practices in many ecosystems (Mathin 2008), and it eclipses the first step of beginning with a style guide (Wickham 2021) by proposing concepts and workflows that can be embraced through a pluralism of tools and strategies. The book is an excellent example of balance between detail and big-picture thinking. It further serves as an enlightening illustration of the capacity for data science and its tools to promote lucid statistical reasoning.

References

Auker LA, Barthelmess EL (2020). "Teaching R in the Undergraduate Ecology Classroom: Approaches, Lessons Learned, and Recommendations." *Ecosphere*, **11**(4), e03060. doi:10.1002/ecs2.3060.

Hao J, Ho TK (2019). "Machine Learning Made Easy. A Review of Scikit-Learn Package in Python Programming Language." *Journal of Educational and Behavioral Statistics*, **44**(3), 348–361. doi:10.3102/1076998619832248.

Jumaat NF, Tasir Z (2014). "Instructional Scaffolding in Online Learning Environment: A Meta-analysis." In *2014 International Conference on Teaching and Learning in Computing and Engineering*, pp. 74–77. doi:10.1109/LaTiCE.2014.22.

Lai J, Lortie CJ, Muenchen RA, Yang J, Ma K (2019). "Evaluating the Popularity of R in Ecology." *Ecosphere*, **10**(1), e02567. doi:10.1002/ecs2.2567.

Lortie CJ, Braun J, Filazzola A, Miguel F (2020). "A Checklist for Choosing between R Packages in Ecology and Evolution." *Ecology and Evolution*, **10**, 1098–1105. doi:10.1002/ece3.5970.

Mathin RC (2008). *Clean Code: A Handbook of Agile Software Craftsmanship*. O'Reilly, Boston.

Ou WJA, Henriques GJB, Senthilnathan A, Ke PJ, Grainger TN, Germain RM (2022). "Writing Accessible Theory in Ecology and Evolution: Insights from Cognitive Load Theory." *BioScience*, **72**(3), 300–313. doi:10.1093/biosci/biab133.

Persson I, Khojasteh J (2021). "Python Packages for Exploratory Factor Analysis." *Structural Equation Modeling: A Multidisciplinary Journal*, **28**(6), 983–988. doi:10.1080/10705511.2021.1910037.

Raschka S, Patterson J, Nolet C (2020). "Machine Learning in Python: Main Developments and Technology Trends in Data Science, Machine Learning, and Artificial Intelligence." *Information*, **11**(4). doi:10.3390/info11040193.

Scavetta RJ, Angelov B (2021). *Python and R for the Modern Data Scientist. The Best of Both Worlds*. O'Reilly Media, Sebastopol.

Scherer R, Siddiq F, Sánchez-Scherer B (2021). "Some Evidence on the Cognitive Benefits of Learning to Code." *Frontiers in Psychology*, **12**. doi:10.3389/fpsyg.2021.559424.

Wickham H (2021). *The **tidyverse** Style Guide*. RStudio, GitHub. URL https://style.tidyverse.org/.

Xie C, Jauhari S, Mora A (2021). "Popularity and Performance of Bioinformatics Software: The Case of Gene Set Analysis." *BMC Bioinformatics*, **22**(1), 191. doi:10.1186/s12859-021-04124-5.

Zhao M, Yan E, Li K (2018). "Data Set Mentions and Citations: A Content Analysis of Full-Text Publications." *Journal of the Association for Information Science and Technology*, **69**(1), 32–46. doi:10.1002/asi.23919.

Reviewer:

Christopher J. Lortie
York University *and* NCEAS
Biology
Toronto, Canada, M3J1P3
E-mail: chris@ecoblender.org
URL: https://www.nceas.ucsb.edu/about-us/our-people

Journal of Statistical Software https://www.jstatsoft.org/
published by the Foundation for Open Access Statistics https://www.foastat.org/
July 2022, Volume 103, Book Review 2 *Published:* 2022-07-13
doi:10.18637/jss.v103.b02