Intermediate Data Science with R: Scope, insights, and materials for a second course in R.

Kelly Bodwin
Department of Statistics
California Polytechnic State University
kbodwin@calpoly.edu

Tyson Barrett
Highmark Health and
Utah State Univeristy
tyson.barrett@usu.edu

Abstract

This paper proposes a definition of scope for learners and educators at the intermediate level using R for data science. We also correspondingly provide an online repository of teaching and learning resources, including a publicly hosted textbook.

Keywords: R, statistical programming, data science education

1 Introduction

For new learners of R Statistical Software (R Core Team, 2024), the landscape of learning materials and classroom opportunities is extensive. Perhaps the most well-known resource is the free online textbook *R for Data Science* (Wickham et al., 2023); but a plethora of other resources exists for different syntactical approaches, use cases, etc. (Crawley, 2012; Verzani, 2004; Navarro, 2015). Additionally, college programs frequently offer Introduction to R courses, and professional conferences regularly offer pre-conference workshops. Despite the many different approaches to introductory R education, there is a loose general consensus on what constitutes "Beginner" or "Introductory" material: language skills like function calling and scripting; data skills like import, wrangling, summarizing, and visualization; and analytical skills like basic statistical tests or models.

However, after foundational R skills have been established - whether through self-learning, a workplace, or coursework - there is little consensus on next steps. Learning materials and opportunities at the "Intermediate-to-Advanced" R level tend to be specialized; focusing deeply on a particular application area. In particular, textbooks purporting to be at this level (Matloff, 2011; Wickham, 2019) are often extremely development-focused, teaching users about the architecture of the language itself and/or how to design and create software. Other approaches limit their scope to topics like visualization (Sievert, 2020; Wilke, 2019) , statistical modeling (Hastie et al., 2009) or particular domain spaces (Gentleman et al., 2005).

In this project, we propose and justify a standard set of topics that are typically untaught at introductory levels, that we believe should be considered the backbone of intermediate R education in data science. We also provide a free, online, open-source resource of educational materials for these topics.

2 Methods

We identify five problem spaces in data computing that data scientists commonly address, that require skills outside of the typical introductory curriculum:

2.1 Data types and sources beyond comma-separated files.

Introductory R materials nearly always limit their scope to data read in from local *.txt* or *.csv* files. While convenient for streamlining teaching, this structure rarely mimics the true needs of a data scientist.

Increasingly, data science work is done on data stored in databases, whether locally (e.g. *duckdb* (Raasveldt and Mühleisen, 2019)) or remotely (e.g. on an SQL server). For an R learner to move to an intermediate practitioner, they must have some understanding of the database structure, as well as fluency in R tools for interfacing, such as *arrow* (Richardson et al., 2024), *DBI* (R Special Interest Group on Databases (R-SIG-DB) et al., 2024) and *odbc* (Hester et al., 2024).

Additionally, it is crucial for a practitioner to be able to interface with APIs, using tools like *jsonlite* (Ooms, 2014) or *XML* (Temple Lang, 2024) to handle pulled data that is not in csv format. It is also often of interest to extract data from non-API online sources, necessitating training in webscraping using tools like *rvest* (Wickham, 2024) or *htmltools* (Cheng et al., 2024).

2.2 Advanced and dynamic data visualization.

As data visualization is a common topic for a full course or textbook, a wealth of resources is already in place; we refer to (Wilke, 2019; Sievert, 2020; Sarkar, 2008) and others for coverage of these topics.

2.3 Complexities of unclean or unstructured data.

While introductory courses do often touch upon data cleaning and wrangling, the complexities of real data are often far more convoluted than a beginner student is equipped to address. We propose the following for a second sequence curriculum:

Handling of missing data. Learners should be taught more sophisticated methods to recognize, assess, and replace missing data. They may use tools like *naniar* (Tierney and Cook, 2023) to find and visualize trends in missingness, and learn common methods for imputation.

Multi-pivot pipelines. Data reshaping is rarely a one-step procedure; learners will practice scenarios that require two or more pivots on a given data frame in sequence.

Joins with repeats. A common source of data glitches is when joins are performed many-to-many match scenarios, resulting in repeat cases in the dataset. An intermediate course should teach strategies for error checking these scenarios and troubleshooting them responsibly.

Manual cleaning via regular expressions. Although intermediate learners may have seen string methods and possibly even regular expressions in a first course, they have likely not experienced using regular expressions for bespoke data cleaning. A focus on more complex regular expressions should be incorporated in a more advanced course.

2.4 Speed and efficiency concerns for large or repeated analyses.

In real data settings, R programmers are often asked to handle larger data and/or many duplicate analysis processes. It is critical that an intermediate user be exposed to methods of code streamlining, such as matrix operations, as well as tools built for speed and computational efficiency in these contexts. Most important among these is the *data.table* (Barrett et al., 2024) package, known for being the most optimized code for in-memory data manipulation. Much of this project will pertain to developing learning materials for adoption of the *data.table* package, which are largely lacking in the education space.

2.5 Workflow and reproducibility for long-term collaborative projects.

Practical data science requires use of collaborative tools including version control and professional communication. At present, we recommend teaching Quarto notebooks (Allaire and Dervieux, 2024) and git filetracking (Chacon and Straub, 2014) with GitHub or GitLab hosting.

Additionally, code projects must be structured in a way that is efficient and collaborative. An intermediate course should teach custom function writing, scripting, and possibly package-based workflow (Wickham and Bryan, 2023).

3 Discussion

The work in this project will serve as a foundational starting point for much-needed Intermediate R curriculum in data science. Learners will have the opportunity to work through the open-source "course-in-a-box" textbook and activities, and educators may use and build upon these resources as they create their own materials.

References

JJ Allaire and Christophe Dervieux, 2024. *quarto: R Interface to 'Quarto' Markdown Publishing System*. R package version 1.4.

Tyson Barrett, Matt Dowle, Arun Srinivasan, Jan Gorecki, Michael Chirico, and Toby Hocking, 2024. *data.table: Extension of 'data.frame'*. R package version 1.15.4.

Scott Chacon and Ben Straub. 2014. *Pro git*. Apress.

Joe Cheng, Carson Sievert, Barret Schloerke, Winston Chang, Yihui Xie, and Jeff Allen, 2024. *htmltools: Tools for HTML*. R package version 0.5.8.1.

Michael J Crawley. 2012. *The R book*. John Wiley & Sons.

Robert Gentleman, Vincent Carey, Wolfgang Huber, Rafael Irizarry, and Sandrine Dudoit. 2005. *Bioinformatics and computational biology solutions using R and Bioconductor*. Springer Science & Business Media.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. An introduction to statistical learning.

Jim Hester, Hadley Wickham, and Oliver Gjoneski, 2024. *odbc: Connect to ODBC Compatible Databases (using the DBI Interface)*. R package version 1.4.2.

Norman Matloff. 2011. *The art of R programming: A tour of statistical software design*. No Starch Press.

Danielle Navarro. 2015. Learning statistics with r: A tutorial for psychology students and other beginners (version 0.6). *Sydney, Australia: University of New South Wales*.

Jeroen Ooms. 2014. The jsonlite package: A practical and consistent mapping between json data and r objects. *arXiv:1403.2805 [stat.CO]*.

R Core Team, 2024. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

R Special Interest Group on Databases (R-SIG-DB), Hadley Wickham, and Kirill Müller, 2024. *DBI: R Database Interface*. R package version 1.2.3.

Mark Raasveldt and Hannes Mühleisen. 2019. Duckdb: an embeddable analytical database. In *Proceedings of the 2019 International Conference on Management of Data*, pages 1981–1984.

Neal Richardson, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow, 2024. *arrow: Integration to 'Apache' 'Arrow'*. R package version 18.1.0.

Deepayan Sarkar. 2008. Multivariate data visualization with r. *Use R*.

Carson Sievert. 2020. *Interactive web-based data visualization with R, plotly, and shiny*. Chapman and Hall/CRC.

Duncan Temple Lang, 2024. *XML: Tools for Parsing and Generating XML Within R and S-Plus*. R package version 3.99-0.17.

Nicholas Tierney and Dianne Cook. 2023. Expanding tidy data principles to facilitate missing data exploration, visualization and assessment of imputations. *Journal of Statistical Software*, 105(7):1–31.

John Verzani. 2004. *Using R for introductory statistics*. Chapman and Hall/CRC.

Hadley Wickham and Jennifer Bryan. 2023. *R packages*. " O'Reilly Media, Inc.".

Hadley Wickham, Mine Çetinkaya-Rundel, and Garrett Grolemund. 2023. *R for data science*. " O'Reilly Media, Inc.".

Hadley Wickham. 2019. *Advanced r*. chapman and hall/CRC.

Hadley Wickham, 2024. *rvest: Easily Harvest (Scrape) Web Pages*. R package version 1.0.4.

Claus O Wilke. 2019. *Fundamentals of data visualization: a primer on making informative and compelling figures*. O'Reilly Media.