Book Review

Nathan Cruze¹

Benjamin S. Baumer, Daniel T. Kaplan, and Nicholas J. Horton. *Modern Data Science with R*. Boca Raton, FL: Chapman and Hall/CRC Press, Taylor and Francis Group, 2017. ISBN 978-1-4987-2448-7, 551pp, \$99.95.

Modern Data Science with R originated from materials that supported semester-long introductory and intermediate data science courses offered at Amherst, Smith, and Macalester Colleges. As such, this text is targeted to a rather general audience and assumes only a minimal background in introductory statistics and computer programming concepts as its starting point. The text is more conceptual rather than technical, and the numerous color-coded R code excerpts presented in the text are a useful aid for seasoned and first-time R programmers alike. The upper-level undergraduate data science student and the self-directed learner will both benefit from the contents and layout of the book. In the presented material, the authors draw an important distinction between the sometimes conflated terms “data science” and “big data”, and an exhaustive treatment of the latter is outside the scope of the book. While those looking for insights into handling truly massive data sets may be better served by other references, an invested reader should expect to be rewarded with a variety of foundational skills needed to handle, analyze, and interact with a wide variety of medium-sized data in R, i.e., data that may fit on one ordinary computer hard drive while being entirely too big to be stored in memory.

In addition to introducing the R programming language, the text is presented around the use of the RStudio front end, making extensive use of the `knitr` and `markdown` packages. The authors made this decision in the name of enhancing workflow and reproducibility, and emphasize these two points as important corner stones in the practice of data science. An additional R package developed by the authors, `mdsr`, installs most of the additional R packages necessary to carry out the analyses and exercises presented within the text and includes some of the smaller example data sets. A range of interesting larger data sets including airline data, Twitter data, and Federal Elections Committee data are drawn from online sources and referred to repeatedly throughout examples and exercises in the text. Several appendices provide additional information on R, RStudio, and `mdsr`.

Modern Data Science with R is clearly organized along three parts. Part I (Chapters 1–6) covers a variety of introductory topics in data science with an emphasis on data visualization and data wrangling, the process of acquiring and transforming a “raw” data format into a form more suitable for some end use. Exposition and demonstration of the

¹ USDA National Agricultural Statistics Service – Research and Development Division Room 6412A–South Building 1400 Independence Avenue, SW, Washington, District of Columbia 20250, U.S.A. Email: nathan.cruze@nass.usda.gov

former revolves around the `ggplot2` plotting paradigm (Wickham 2016). The latter is focused on data cleaning and data wrangling through the `dplyr` package. For the aims of this text, extensive use of these libraries make sense, although they are certainly not the only available options. While other potentially relevant and useful plotting approaches exist in R, e.g., base graphics (Murrell 2011) or lattice graphics (Sarkar 2008), the extensive and consistent use of the `ggplot2` grammar lowers some of the barriers to understanding other data types and visualizations presented in later supplemental chapters, e.g., mapping spatial data through the use of `ggmap` or `leaflet` packages. The use of `dplyr` (as opposed to `data.table`) naturally fits with many relational database backends, and it agrees with the syntax of other packages demonstrated in the text. In order to get the most out of the remainder of the book, a mastery of the graphics and data wrangling concepts presented in Chapters 3, 4 and 5 becomes essential. The thoughtful and refreshing discussion on professional ethics in Chapter 6 is a welcome addition.

Part II (Chapters 7–10) introduces a number of common statistical concepts and modeling approaches: sampling, bootstrapping, linear regression, statistical learning and prediction, unsupervised learning, and simulation. These are presented lucidly, and in a largely non-technical manner. (Note that a more mathematical treatment of regression models is given in Appendix E.) Entire textbooks have been written on each of these subjects, but the content in Part II may serve as useful introductory material or a supplement to other materials for those unfamiliar with these topics. Part II flows as a natural continuation of the aims and thought processes of the earlier chapters: that data may need structure to be analyzed, that visualization is a powerful tool for recognizing patterns and relationships, that the magnitude and association of a response variable with other explanatory variables may be quantified or predicted (regression, supervised learning), and that patterns in data may still emerge even when an outcome is not explicitly measured (unsupervised learning).

The collection of topics in Part III are essentially independent of the material in the first two parts of the book, but they may be of interest to serious students. Covered topics include the production of interactive data graphics (Chapter 11); SQL, databases, and database administration (Chapters 12–13); and spatial data and mapping, text mining, and network science (Chapters 14–16). At just fourteen pages long, the final chapter titled “Epilogue: Towards ‘big data’” can only hint at some of the issues, challenges, and techniques associated with harnessing big data.

A major strength of *Modern Data Science with R* is the sheer breadth of topics covered in a non-technical manner, and the generous use of code excerpts to help the reader walk through and reproduce important concepts. There is much to recommend this text to the advanced undergraduate data science students and general R users interested in improving their understanding of data wrangling, data visualization, and discovery.

References

- Murrell, P. 2011. *R Graphics*, 2nd ed. Boca Raton, FL: Chapman & Hall/CRC Press.

- Sarkar, D. 2008. *Lattice: Multivariate Data Visualization with R*. New York: Springer-Verlag.
- Wickham, H. 2016. *ggplot2: Elegant Graphics for Data Analysis*, 2nd ed. New York: Springer-Verlag.