



---

# Journal of Statistical Software

August 2017, Volume 80, Book Review 2.

doi: 10.18637/jss.v080.b02

---

Reviewer: Tim Downie  
Beuth University of Applied Sciences, Berlin

---

## Modern Data Science with R

Benjamin S. Baumer, Daniel T. Kaplan, Nicholas J. Horton  
Chapman & Hall/CRC, Boca Raton, 2017.  
ISBN 978-1-4987-2448-7. 556 pp. USD 99.95.  
<https://www.crcpress.com/9781498724487>

---

*Modern Data Science with R* by Baumer, Kaplan and Horton is a comprehensive handbook to the current themes in the field of data science. The book sets out to give a comprehensive overview of the relevant topics. It does not attempt to cover each subject in fine detail, but is detailed enough to convey the important issues and to point the reader to appropriate resources for a more thorough understanding.

An important principle throughout the book is that datasets are routinely much larger than those presented in typical statistics courses, that the data are rarely in the right format to start analyzing, and that statistical methods, data handling and the presentation of results should not be considered separately. This approach is not new to practical data analysis and very little of the statistical methodology is new, but it does present these subjects coherently using up-to-date tools and examples.

The book stems from an advanced lecture course in data science. It would undoubtedly be useful to many postgraduate students of applied statistics. The handbook style will also be of use to statisticians who want to keep up to date in this area. In particular the book utilizes functions from many different R packages, and will be helpful for data analysts to keep their R skills up to date. Although one of the appendices covers an introduction to R ([R Core Team 2017](#)) and **RStudio** ([RStudio Team 2017](#)), realistically it is expected that the reader has some experience with R. Existing R users with no experience of **RStudio** might find the appendix useful, but **RStudio** is not required to work through this book.

Overall the book is well written, well structured and the general writing style is both objective and entertaining. The limitations in this book are readily acknowledged by the authors, who make it clear that it is impossible to cover all relevant topics in detail in one book. A notable appeal to the book is the interesting selection of data examples using a variety of datasets, e.g., the prevalence of HIV, baseball, baby names in the US and flights within in the US. Many chapters end with an *extended example* for the reader to work through, reinforcing the ideas introduced in the chapter, followed by well thought out exercises.

Throughout the methods are comprehensively illustrated using R code. All the examples use data that are directly available via R packages or a URL. The book comes with a scratch

panel to reveal a code to access the book online. This is very helpful for working through the R code examples.

The book is divided into three major parts, *Introduction to Data Science, Statistics and Modeling*, and *Topics in Data Science*, followed by six appendices. The rest of this review summarizes each part and chapter in detail, followed by a short conclusion.

## Part I – Introduction to Data Science

*Chapter 1: Prologue: Why data science?* starts with a discussion of what data science means illustrated with a short *Case Study: The evolution of sabermetrics*. “Sabermetrics” is the data analysis of baseball. For those who know the “Moneyball” story (Lewis 2003) there is little new information here. Those unacquainted with Moneyball will find the description too short.

*Chapter 2: Data visualization* starts with an example using funding data from the US federal elections in 2012. Starting with basic bar charts it is demonstrated how the message can be enhanced (or muddled) by the use of color and aggregation. These data in their initial form are misleading. Each bar corresponds to the “amount of money spent on each candidate”, but as we find out later, money spent by (for example) opponents criticizing Obama is classed as *money spent on Obama*. The authors highlight the problem, but the example is *too* extreme. Someone presenting a graph claiming that nearly USD 300 million was spent “on” Obama, without making it clear that over three-quarters was spent against him, is deliberately trying to mislead the reader.

The section moves on to different data presentation methods some well established (e.g., a scatter plot with a third variable indicated by the size of the point) and some highly complex diagrams, displaying a lot of information but requiring detailed explanation. Some often overlooked aspects of data visualization are mentioned such as the use of appropriate color schemes to accommodate colorblindness.

*Chapter 3: A grammar for graphics*. This chapter might have been incorporated into the data visualization chapter. The main distinction between the two is that Chapter 2 contains no R code, whereas coding is the main focus of Chapter 3. The concept of code as a structured taxonomy (a “grammar”) is expanded throughout this and the next chapter. Functions are the verbs, with data or variables taking the role of nouns, etc. The analogy is sometimes overdone, but is effective as a teaching tool as it emphasizes coding concepts independent from the software language.

The extended example investigates the year of birth distribution of first names in the US population. The reader is guided through a graphical data analysis, generating many composite diagrams, ending with a diagram presenting the age quartiles of the 25 most frequent US female names.

*Chapter 4: Data wrangling* is a modern word for data processing, using database terminology, such as *filter* to mean subsetting the rows of a data matrix. The R code from this chapter onwards extensively uses the functions and concepts in the package **dplyr** (*A Grammar of Data Manipulation*, Wickham, François, Henry, and Müller 2017). Notation not in the base S language is introduced, in particular the use of the operators between R functions such as `+` and `%>%`.

*Chapter 5: Tidy data and iteration*. The term *tidy data* is defined and means tidy in the

sense of “in a format ready for data analysis” rather than “cleaned of errors”. The classic data matrix format is usually a tidy dataset. An example of an error free data structure, which is nonetheless “non-tidy” is a spreadsheet containing records sorted into groups, with a subtotal row after each group. There is a short section on avoiding `for` loops, covering the S language `apply` functions followed by iteratively reapplying the same code to subsets of data using the `do` function.

*Chapter 6: Professional ethics.* This is an important and very interesting subject, but when reading through the book, the sudden change in subject matter is jolting. I can see that in a lecture course, this would be an appropriate place to change the theme and style, but in the book this subject would not be out of place as an appendix.

## Part II – Statistics and Modeling

Part II covers subjects that one would expect to take center stage in a text book on data science. The models presented are methods which are scalable and do not rely on distributional or sampling assumptions. Being the central theme of data science, I had expected this part to be more substantial than it is.

*Chapter 7: Statistical foundations* briefly introduces the idea that a dataset is a sample from a population before covering resampling methods, handling outliers and linear models. The book does not cover traditional statistical models in detail, most being left to an appendix (e.g., logistic regression). Multiple linear regression appears in this section as a prelude to machine learning methods in later sections.

Almost all statisticians will know the shortcomings of decision making based on  $p$ -values. The authors express their opinion in an interesting way:  *$p$ -values below 0.05 provide a kind of “certificate” for the test, but as the  $p$ -value conveys much less information than usually supposed, the “certificate” might not be worth the paper it’s printed on.*

*Chapter 8: Statistical learning and predictive analytics* is a detailed introduction to supervised learning, including, for example, classifiers, random forests, ensemble methods and cross-validation. There is clearly a trade off between brevity and completeness. On the whole I think that the authors have chosen the right compromise, but sometimes the text is quite terse, and occasionally assumes a prior knowledge that might be challenging to many master’s level students.

*Chapter 9: Unsupervised learning* focuses on clustering methods and is noticeably shorter than the previous chapter. There is no specific extended example, although a detailed example analyzing Scottish parliament voting in the subsection *Dimension reduction* serves as an extended example.

*Chapter 10: Simulation* should be covered somewhere in a modern master’s program in statistics and is relevant to data science. Most statisticians will find little new in this section other than perhaps the modern coding style.

## Part III – Topics in Data Science

This part tackles a variety of very different subjects; each chapter is independent of the others. A reader who has little interest in one chapter could easily skip it without hindrance when reading other chapters.

*Chapter 11: Interactive data graphics* were once the domain of specialized statistical software. Now they are routinely accessible for all via the user’s browser using *JavaScript* and *Widgets*. Different types of interactive graphics are presented. Each of the covered methods is briefly explained and many, but not all, have example code. It is particularly important for the user to replicate the R examples, as the interactive properties are not available in a printed book. For example the screenshot of a so called *dygraph* appears in the book to be just a normal time series diagram, because the slider underneath the diagram has been cropped out. Using the slider the user can explore the time series over different time scales.

A very quick overview of the **shiny** “web application framework” (Chang, Cheng, Allaire, Xie, and McPherson 2017) is presented using a graphical web app as an example. The extended example in this chapter reproduces a complex diagram from Nathan Yau and his excellent blog on data visualization *Flowing Data* (Yau 2011). It is unclear however, what this has to do with interactive graphics, as the final diagram is not interactive.

*Chapter 12: Database querying using SQL.* The authors refer to the baseball data as *small data* even though the data library contains many datasets, one with over 136 000 observations. The term “small” does not directly refer to the number of observations and variables, rather that the entire data can be loaded and processed in memory on a standard computer. The authors introduce the name *medium data*, meaning data which fits on a typical user’s hard drive, but is too big to be held in memory. The standard approach is to store data in a database and use the SQL language to preprocess and load only the data needed for each current task. SQL code is introduced in this chapter, but more usefully R commands in the **dplyr** package are presented, which circumvent the need to write SQL code. The authors do emphasize, however, the importance of knowing some SQL, as the translation from R code to SQL using **dplyr** is somewhat limited. At the end of the chapter there is a helpful comparison of accessing databases via SQL and R.

*Chapter 13: Database administration* This chapter follows up the SQL theme from Chapter 12 concentrating on the construction of a Database.

*Chapter 14: Working with spatial data.* The emphasis in this chapter is not on developing spatial models but on the data handling and visualization of spatial data. The first subsection uses the famous John Snow Cholera example, and point out that this story is often simplified to make it more impressive. The section *Making maps* uses the John Snow dataset and creates a diagram with the modern London streets and quality graphics using **ggplot**, **get\_map** and a specific projection of earth coordinates (Wickham 2009). This is a long example but the fine details are warranted; the authors do not gloss over the finicky complications of creating map-based diagrams. There are two further extended examples in the chapter, election results of US congressional districts and historical airline routes.

*Chapter 15: Text as data* includes some useful resources on parsing and analyzing text, especially useful after scraping data from websites designed to be read by humans. Included in this chapter is a subsection on accessing and analyzing twitter feeds.

*Chapter 16: Network science* introduces the general principle of graphs, including some specific issues with plotting graphs. An extended example investigates the *six degrees of separation* hypothesis using the IMDB data for film actors. Later in the chapter is an overview of Google’s search engine method as the motivation for the *page rank metric*.

*Chapter 17: Epilogue: Towards “big data”.* “Big data is an exceptionally hot topic, but not so well defined.” The book does not attempt to cover the analysis of genuinely “big

data” but does explain some of the issues involved, such as parallel processing and compiled programming.

Part III is followed by six *appendices*. There are some important topics in the appendices such as using R packages, algorithmic thinking, writing user defined R functions and ANCOVA and GLM models, but many topics will be well known to most readers. It would be advisable for a reader who is new to this field to cast an eye over the appendices, before starting on the main part of the book, in case some of the background knowledge needs to be filled in.

## Miscellaneous comments and conclusions

The book has a very North American perspective. To give just a few examples: temperatures are given on the Fahrenheit scale; despite an extremely brief explanation of baseball hidden in a footnote, the analysis of the baseball data will be too esoteric for many outside of North America; “only 15% of people named Robin are male” is nowhere near true in other English speaking countries.

In a couple of instances, the code in the online book does not exactly match the code in the printed version. This should not be a problem for experienced R users, but might be confusing for newish users.

Finally, the topics in this book are currently highly relevant in a field which is rapidly developing. Some topics will soon become obsolete, other topics are or will become essential to data science. For these topics the code presented in the book will at sometime become outdated, but hopefully the overall methodology and discussions will remain relevant for many years.

In conclusion, I recommend this book as a course companion to a master’s level course in data analysis and to statisticians who want to keep their skills in the field of data science up to date. It should be used as a handbook covering no topic in fine detail but giving very good getting-started examples to a variety of different subjects in the field.

## References

- Chang W, Cheng J, Allaire J, Xie Y, McPherson J (2017). *shiny: Web Application Framework for R*. R package version 1.0.1, URL <https://CRAN.R-project.org/package=shiny>.
- Lewis MM (2003). *Moneyball: The Art of Winning an Unfair Game*. Norton paperback. W.W. Norton.
- R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- RStudio Team (2017). *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston. URL <http://www.RStudio.com/>.
- Wickham H (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag. URL <http://ggplot2.org/>.
- Wickham H, François R, Henry L, Müller K (2017). *dplyr: A Grammar of Data Manipulation*. R package version 0.7.2, URL <https://CRAN.R-project.org/package=dplyr>.

Yau N (2011). *Visualize This: The FlowingData Guide to Design, Visualization, and Statistics*. John Wiley & Sons. doi:10.1002/9781118722213.

**Reviewer:**

Tim Downie  
Beuth University of Applied Sciences  
Department II  
D-13353, Berlin, Germany  
E-mail: [tim.downie@beuth-hochschule.de](mailto:tim.downie@beuth-hochschule.de)