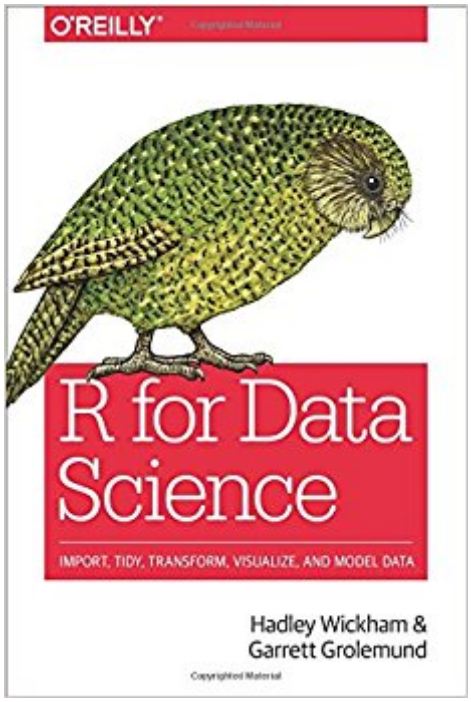
R For Data Science: Import, Tidy, Transform, Visualize, And Model Data



Synopsis

Learn how to use R to turn raw data into insight, knowledge, and understanding. This book introduces you to R, RStudio, and the tidyverse, a collection of R packages designed to work together to make data science fast, fluent, and fun. Suitable for readers with no previous programming experience, R for Data Science is designed to get you doing data science as quickly as possible.Authors Hadley Wickham and Garrett Grolemund guide you through the steps of importing, wrangling, exploring, and modeling your data and communicating the results. YouÃ¢â ¬â„¢ll get a complete, big-picture understanding of the data science cycle, along with basic tools you need to manage the details. Each section of the book is paired with exercises to help you practice what youÃ¢â ¬â„¢ve learned along the way.YouÃ¢â ¬â„¢ll learn how to:WrangleÃ¢â ¬â •transform your datasets into a form convenient for analysisProgramÃ¢â ¬â •learn powerful R tools for solving data problems with greater clarity and easeExploreÃ¢â ¬â •examine your data, generate hypotheses, and quickly test themModelÃ¢â ¬â •provide a low-dimensional summary that captures true "signals" in your datasetCommunicateÃ¢â ¬â •learn R Markdown for integrating prose, code, and results

Book Information

Paperback: 522 pages

Publisher: O'Reilly Media; 1 edition (January 5, 2017)

Language: English

ISBN-10: 1491910399

ISBN-13: 978-1491910399

Product Dimensions: 6 x 1 x 9 inches

Shipping Weight: 1.6 pounds (View shipping rates and policies)

Average Customer Review: 4.8 out of 5 stars 45 customer reviews

Best Sellers Rank: #2,657 in Books (See Top 100 in Books) #2 inÃ Â Books > Computers & Technology > Software > Mathematical & Statistical #2 inÃ Â Books > Computers & Technology > Databases & Big Data > Data Processing #20 inÃ Â Books > Textbooks > Science & Mathematics > Mathematics > Statistics

Customer Reviews

View larger From the Preface Data science is an exciting discipline that allows you to turn raw data into understanding, insight, and knowledge. The goal of R for Data Science is to help you learn the most important tools in R that will allow you to do data science. After reading this book,

youÃƒÂ¢Ã â ¬Ã â„¢ll have the tools to tackle a wide variety of data science challenges, using the best parts of R. What You Will Learn Data science is a huge field, and thereÃƒÂ¢Ã â ¬Ã â„¢s no way you can master it by reading a single book. The goal of this book is to give you a solid foundation in the most important tools. Our model of the tools needed in a typical data science project looks something like this: [image above]. First you must import your data into R. This typically means that you take data stored in a file, database, or web API, and load it into a data frame in R. If you canÃƒÂ¢Ã â ¬Ã â„¢t get your data into R, you canÃƒÂ¢Ã â ¬Ã â„¢t do data science on it. Once youÃƒÂ¢Ã â ¬Ã â„¢ve imported your data, it is a good idea to tidy it. Tidying your data means storing it in a consistent form that matches the semantics of the dataset with the way it is stored. In brief, when your data is tidy, each column is a variable, and each row is an observation. Tidy data is important because the consistent structure lets you focus your struggle on questions about the data, not fighting to get the data into the right form for different functions. Once you have tidy data, a common first step is to transform it. Transformation includes narrowing in on observations of interest (like all people in one city, or all data from the last year), creating new variables that are functions of existing variables (like computing velocity from speed and time), and calculating a set of summary statistics (like counts or means). Together, tidying and transforming are called wrangling, because getting your data in a form thatÃƒÂ¢Ã â ¬Ã â„¢s natural to work with often feels like a fight. Once you have tidy data with the variables you need, there are two main engines of knowledge generation: visualization and modeling. These have complementary strengths and weaknesses so any real analysis will iterate between them many times. Visualization is a fundamentally human activity. A good visualization will show you things that you did not expect, or raise new questions about the data. A good visualization might also hint that youÃƒÂ¢Ã â ¬Ã â„¢re asking the wrong question, or you need to collect different data. Visualizations can surprise you, but donÃƒÂ¢Ã â ¬Ã â„¢t scale particularly well because they require a human to interpret them. Models are complementary tools to visualization. Once you have made your questions sufficiently precise, you can use a model to answer them. Models are a fundamentally mathematical or computational tool, so they generally scale well. Even when they donÃƒÂ¢Ã â ¬Ã â„¢t, itÃƒÂ¢Ã â ¬Ã â„¢s usually cheaper to buy more computers than it is to buy more brains! But every model makes assumptions, and by its very nature a model cannot question its own assumptions. That means a model cannot fundamentally surprise you. The last step of data science is communication, an absolutely critical part of any data analysis project. It doesnÃƒÂ¢Ã â ¬Ã â„¢t matter how well your models and visualization have led you to understand the data unless you can also communicate your results to others. Surrounding all these tools is

programming. Programming is a cross-cutting tool that you use in every part of the project. You donÃƒÂ¢Ã â ¬Ã â„¢t need to be an expert programmer to be a data scientist, but learning more about programming pays off because becoming a better programmer allows you to automate common tasks, and solve new problems with greater ease. YouÃƒÂ¢Ã â ¬Ã â„¢ll use these tools in every data science project, but for most projects theyÃƒÂ¢Ã â ¬Ã â„¢re not enough. ThereÃƒÂ¢Ã â ¬Ã â„¢s a rough 80-20 rule at play; you can tackle about 80% of every project using the tools that youÃƒÂ¢Ã â ¬Ã â„¢ll learn in this book, but youÃƒÂ¢Ã â ¬Ã â„¢ll need other tools to tackle the remaining 20%. Throughout this book weÃƒÂ¢Ã â ¬Ã â„¢ll point you to resources where you can learn more.

Hadley Wickham is an Assistant Professor and the Dobelman FamilyJunior Chair in Statistics at Rice University. He is an active memberof the R community, has written and contributed to over 30 R packages, and won the John Chambers Award for Statistical Computing for his work developing tools for data reshaping and visualization. His research focuses on how to make data analysis better, faster and easier, with a particular emphasis on the use of visualization to better understand data and models.Garrett Grolemund is a statistician, teacher and R developer who currently works for RStudio. He sees data analysis as a largely untapped fountain of value for both industry and science. Garrett received his Ph.D at Rice University in Hadley Wickham’s lab, where his research traced the origins of data analysis as a cognitive process and identified how attentional and epistemological concerns guide every data analysis.Garrett is passionate about helping people avoid the frustration and unnecessary learning he went through while mastering data analysis. Even before he finished his dissertation, he started teaching corporate training in R and data analysis for Revolutions Analytics. He’s taught at Google, eBay, Axciom and many other companies, and is currently developing a training curriculum for RStudio that will make useful know-how even more accessible. Outside of teaching, Garrett spends time doing clinical trials research, legal research, and financial analysis. He also develops R software, he’s co-authored the lubridate R package--which provides methods to parse, manipulate, and do arithmetic with date-times--and wrote the ggsubplot package, which extends the ggplot2 package.

Wickham and Grolemund have produced an excellent book that would help a beginning R user become very efficient in explanatory analysis. Unsurprisingly the approach that they expound utilises the "hadleyverse" a collection of packages (ggplot2 for visualisation, tidyr for reshaping, dplyr for selecting and filtering, purrr for functional programming, broom for linear models etc) that

dramatically speed up most of the common steps involved in an analysis. One benefit of Wickham's involvement in these packages has been a coherent philosophy that sits behind them. It can be a little tricky when learning this philosophy, but the long term benefits are enormous.The book is broken up into a number of sections that effectively builds up the ability to ingest, transform, visualise and model datasets. A good portion of the book is available in an online version, to give you a taste of how it is written. Many have been following it as it was written. I have passed on copies of the book to a number of colleagues who were just starting out and the response has been uniformly positive. In my own case I was familiar with some of the these packages; ggplot2, dplyr, tidyr, but found the book taught me purrr and how to better use the packages together.Probably my two biggest caveats to readers are that there are situations where packages from outside the "hadleyverse" maybe required. The authors do a great job of pointing this out, but it does pay in my experience to know data.table and lattice for example. Both because they can occasionally fit a problem better but also because you inevitably come across other people's code where these packages are used. The other caveat is that the modelling is a little rudimentary. Most of the examples are just fitting independent regression models, whereas it seems to me that a hierarchical model would be a better fit. Still these are small things and it would be silly to expect a single book to cover all of these areas.In short this is the book I would give to someone who was keen to learn about how to use R for data science. It reads really well building up the different components whilst still being a valuable reference if you just need a reminder of a particular package (what is the difference between tibbles and data frames again?). Even though a good portion of the book is available online, it is well worth it to have the full thing on your bookshelf (digital or otherwise). On a broader note with Max Kuhn (author of the excellent "Applied Predictive Modelling" with Kjell Johnson) joining Wickham and Grolemund at RStudio, it is a great time to start your R journey.

In my opinion, this is the best book written about using R for data analysis! Very easy to read and follow, even some bit entertaining.There is some typos and errors in the book, like the wrong graph in the wrong place or typos. But you can look at the online electronic version of the book, it is the most updated and corrected. And it's free! Also, the author provided answer keys for the exercises. Just google it, you will find it. It helps you to compare your answer to the author's and learn more!In summary, I love love love this book! I now know why Hadley Wickham is famous for a reason.

A great introduction to the Tidyverse. Coming from SAS programming with only a little R experience this provided a good introduction to me to using R and the Tidyverse in a consistent manner that is

focused on getting you programming and giving you useful tools.

Required reading. This is my go-to resource for basic questions. I bought copies for two of my students since they were working on a research project with me -- they love it!

Makes R easier to use. I am, overall, a fan of Hadley Wickham’s work, and this doesn’t disappoint.

Fantastic, great intro and builds on each concept.

This book is really helpful if you are interested in systematically learning the data wrangling and getting to know the up-to-date ggplot2 package.

My one complaint. The errata page listed by O’reilly does not exists!!!See page 126 and 127 where the examples listed will not work.read_csv() inlineread_csv() col_names=FALSEWill result in errors.

Processing TIDYING UP: ORGANIZE & DECLUTTER 10 MINUTE CLEANING HABIT:SIMPLE STRESS-FREE HABIT TO SPARK JOY WHEN YOU TIDY UP Text Mining with R: A Tidy Approach Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data Data Analytics and Python Programming: 2 Bundle Manuscript: Beginners Guide to Learn Data Analytics, Predictive Analytics and Data Science with Python Programming Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking Spartan Fit!: 30 Days. Transform Your Mind. Transform Your Body. Commit to Grit.