Data Manipulation with R

Second Edition

Efficiently perform data manipulation using the split-apply-combine strategy in R

Jaynal Abedin
Kishor Kumar Das

[PACKT] PUBLISHING open source* community experience distilled

Table of Contents

Data Manipulation with R Second Edition

Data Manipulation with R Second Edition

First published: January 2014

Second edition: March 2015

Production reference: 1250315

Published by Packt Publishing Ltd.

Livery Place

35 Livery Street

Birmingham B3 2PB, UK.

ISBN 978-1-78528-881-4

www.packtpub.com

Credits

Authors

Jaynal Abedin

Kishor Kumar Das

Reviewers

Erik M. Rodríguez Pacheco

Dr. Abbass Ismail Sharif

Dr. Brian J. Spiering

Jitendra Kumar Yadav

Commissioning Editor

Veena Pagare

Acquisition Editor

Sonali Vernekar

Content Development Editor

Manasi Pandire

Technical Editor

Utkarsha S. Kadam

Copy Editors

Khushnum Mistry

Karuna Narayanan

Vikrant Phadke

Adithi Shetty

Project Coordinator

Leena Purkait

Proofreaders

Stephen Copestake

Maria Gould

Paul Hindle

Jonathan Todd

Indexer

Monica Ajmera Mehta

Production Coordinator

Nilesh R. Mohite

Cover Work

Nilesh R. Mohite

About the Authors

Jaynal Abedin currently holds the position of senior statistician at the Centre for Communicable Diseases (CCD) at the International Centre for Diarrhoeal Disease Research, Bangladesh (http://www.icddrb.org/). He attained his bachelor's and master's degrees in statistics from the University of Rajshahi, Bangladesh. He has extensive experience in R programming and Stata, and has good leadership qualities. He has contributed to two books on R and also developed an R package named edeR, short for e-mail data extraction using R which is available at CRAN (http://cran.r-project.org/web/packages/edeR/index.html). He is currently leading a team of statisticians. He has hands-on experience in developing training material and facilitating training in R programming and Stata, along with statistical aspects in public health research. His primary areas of interest in research include causal inference and machine learning. He is currently involved in several ongoing public health research projects, and is a coauthor of nine peer-reviewed scientific papers. Moreover, he is involved in several work-in-progress manuscripts. He works as a freelance statistician in online marketplaces and has obtained a good reputation for his work.

Kishor Kumar Das is a statistician at the International Centre for Diarrhoeal Disease Research, Bangladesh, an internationally recognized organization that focuses mainly on public health research. He completed his MSc and BSc in applied statistics from the Institute of Statistical Research and Training, University of Dhaka, Bangladesh. He has extensively used R for data processing, statistical analysis, and graphs for more than 10 years. His research interests are survival analysis, machine learning, and statistical computing.

About the Reviewers

Erik M. Rodríguez Pacheco works as a manager in the business intelligence unit at Banco Improsa in San José, Costa Rica. He has 11 years of experience in the finance industry. He currently a professor of the Business Intelligence Specialization program at the Continuing Education Programs of Instituto Tecnológico de Costa Rica. Erik is an enthusiast of new technologies, particularly those related to business intelligence, data mining, and data scienc He holds a bachelor's degree in business administration from Universidad de Costa Rica, a specialization in business intelligence from Instituto Tecnológico de Costa Rica, a specialization in data mining from Promidat (Programa Iberoamericano de Formación en Minería de Datos), and a specialization in business intelligence and data mining from Universidad del Bosque, Colombia. He is currently enrolled in a specialization program in dat science from Johns Hopkins University. He can be reached at http://cr.linkedin.com/in/erikrodriguezp.

Dr. Abbass Ismail Sharif is an assistant professor of clinical data sciences and operations a the University of Southern California. He holds a PhD in statistics, an MS in computer scienc and an MS in instructional technology and learning sciences. Abbass does research in the field of statistical computing and data visualization. For this purpose, he extensively uses the R statistical environment. He has developed new multivariate visualization techniques for functional data, and is currently developing visualization techniques to study brain activity dat collected using Near-infrared spectroscopy (NIRS) technology.

Abbass has won a prestigious research award from the American Statistical Society for his doctoral work. He teaches both graduate and undergraduate statistics courses that range from introductory statistics and data analysis for decision-making to advanced modern statistical learning techniques, statistical computing, and data visualization.

Dr. Brian J. Spiering started coding in his elementary school's computer laboratory, hacking BASIC to make programs that entertained his peers and annoyed the school authorities. Much later, he earned a PhD in psychology from the University of California, Santa Barbara, with emphasis on cognition, perception, and cognitive neuroscience. His research is focused on building mathematical and computer models of the human brain and behavior. He has taught biological psychology, data analysis, and statistics. Brian currently works as a data scientist and resides in San Francisco, California, USA.

Jitendra Kumar Yadav is a senior development architect working in research and development for product development and innovation. He is an expert in cloud and big data product development. He has contributed to the open source community in the form of code development and support for a variety of platforms based on big data, cloud technologies, virtualization, storage, networking, and cloud security. For this, he has used programming languages such as C++, Python, R, Java, Go, and Perl.

Jitendra loves to share his knowledge with fellow techies and others. He does so by publishin papers and books and attending corporate tech events. He has won several awards for his

excellent contributions to product development in the fields of cloud computing, big data, artificial intelligence, and virtualization. He has over 12 years of professional experience, and has spent most of his time in research and development.

Occasionally, when Jitendra needs to take a break, he spends his time traveling.

I'd like to thank those who nurtured me, my mom and dad, for all the hope, faith, love, and wise counseling. I would also like to thank those from the Packt Publishing team who made this book happen, especially Leena and Sarah, the reviewers, and the MODX community for an awesome open source development platform.

Support files, eBooks, discount offers, and more

For support files and downloads related to your book, please visit www.PacktPub.com.

Did you know that Packt offers eBook versions of every book published, with PDF and ePub files available? You can upgrade to the eBook version at www.PacktPub.com and as a print book customer, you are entitled to a discount on the eBook copy. Get in touch with us at <service@packtpub.com> for more details.

At www.PacktPub.com, you can also read a collection of free technical articles, sign up for a range of free newsletters and receive exclusive discounts and offers on Packt books and eBooks.

https://www2.packtpub.com/books/subscription/packtlib

Do you need instant solutions to your IT questions? PacktLib is Packt's online digital book library. Here, you can search, access, and read Packt's entire library of books.

Why subscribe?

- Fully searchable across every book published by Packt
- Copy and paste, print, and bookmark content
- On demand and accessible via a web browser

Free access for Packt account holders

If you have an account with Packt at www.PacktPub.com, you can use this to access PacktL today and view 9 entirely free books. Simply use your login credentials for immediate access

Preface

This book, *Data Manipulation with R*, is aimed at giving intermediate-to-advanced level users of R (who have knowledge about datasets) an opportunity to use state-of-the-art approaches in data manipulation. This book will discuss the types of data that can be handled using R an different types of operations for those data types. Upon reading this book, you will be able to efficiently manage and check the validity of your datasets with the effective use of R programming, including specialized packages for data management. You will come to know about the split-apply-combine strategy, which is a state-of-the-art approach in data management. You will also come to know the way to work with database software through ODBC with the help of very simple examples. This book ends with an introduction to text processing for text mining using R.

What this book covers

Chapter 1, *Introduction to R Data Types and Basic Operations*, discusses the way to get R, how to install it, and how to install various libraries. Upon introducing how to write commands in R, this chapter discusses different types of data used in R and their basic operations. Before introducing the data types in this chapter, we will highlight what an object in R is as w as their modes and classes. The mode of an object could be either numeric, character, or logical, whereas its class could be vector, factor, list, data frame, matrix, array, or others. Th chapter also highlights how to work with objects in different modes and how to convert from one mode to another and what caution should be taken during conversion. Missing values in and how to represent missing characters and numeric data types are also discussed here. Along with the data types and basic operations, this chapter sheds light on another important aspect, which is almost never mentioned in other textbooks—the object naming convention i R. We talk about popular object-naming conventions used in R.

Chapter 2, *Basic Data Manipulation*, introduces some special features where we need to take care during data acquisition. Then, an important aspect of factor manipulation is discussed, a well as subsetting a factor variable and how to remove unused factor levels. This chapter als includes coverage of vector and matrix operations. Date processing has been discussed usin an efficient R package: lubridate. Working with the date variable using the lubridate package much more efficient than using any other existing package that is designed to work with the date variable. Also, string processing has been highlighted, and the chapter ends with a description of subscripting and subsetting.

Chapter 3, *Data Manipulation Using plyr and dplyr*, introduces the state-of-the-art approach called split-apply-combine to manipulate datasets. Data manipulation is an integral part of da cleaning and analysis. For a large dataset, it is always preferable to perform operations withi the subgroup of a dataset to speed up the process. In R, this type of data manipulation can be done with base functionality, but for large datasets, it requires a considerable amount of coding and eventually takes longer to process. In the case of large datasets, we can split the dataset performing the manipulation or analysis and then combine them again into a single output. This chapter contains a discussion of the different functions in the plyr package that are used for group-wise data manipulation and also for data analysis. This chapter also contains examples and discussions of the dplyr package to work with data frames. Working with data frames using dplyr is much more efficient and intuitive. You will have a very good understanding of data frame processing through the examples of this chapter.

Chapter 4, *Reshaping Datasets*, deals with the orientation of datasets. Reshaping data is a common and tedious task in real-life data manipulation and analysis. A dataset might come with different levels of grouping, and we need some reorientation to perform certain types of analysis. To perform statistical analysis, we sometimes require wide data and sometimes lon data, and in this case, we need to be able to fluently and fluidly reshape data to meet the requirements of statistical analysis. Important functions from the reshape2 package have been discussed in this chapter with examples.

Chapter 5, *R and Databases*, talks about dealing with database software and R. One of the major problems in R is that its memory is bound by the system virtual memory, and that is why working with a dataset requires the data to be smaller than its memory. However, in reality, the dataset is larger than the virtual memory and sometimes the length of arrays or vectors exceeds the maximum addressable range. To overcome these two limitations, R can be utilized with databases. Interacting with databases using R and dealing with large datasets with specialized packages and data manipulation with sqldf have been discussed with examples in this chapter.

Chapter 6, *Text Manipulation*, covers the processing of text data for text mining. This chapter introduces various sources of text data and the process of obtaining that data. This chapter also discusses processing text data for text mining purposes by using various relevant packages.

What you need for this book

Knowledge about statistical data is required. You are expected to have basic knowledge of R
To run the examples from this book, R should be installed, and it can be found at http://www.r-project.org. The example files are produced on R 3.0.2.

Who this book is for

This book is for intermediate-to-advanced level users of R who have knowledge about datasets, and also for those who regularly work with different research data, including but no limited to public health, business analysis, and the machine learning community.

Conventions

In this book, you will find a number of styles of text that distinguish between different kinds of information. Here are some examples of these styles, and an explanation of their meaning.

Code words in text, database table names, folder names, filenames, file extensions, pathnames, dummy URLs, user input, and Twitter handles are shown as follows: "Once we have an R object, we can easily assess its mode by using `mode()`."

A block of code is set as follows:

```
num.obj <- seq(from=1,to=10,by=2)
logical.obj<-c(TRUE,TRUE,FALSE,TRUE,FALSE)
character.obj <- c("a","b","c")

is.numeric(num.obj)
[1] TRUE

is.logical(num.obj)
[1] FALSE

is.character(num.obj)
[1] FALSE
```

When we wish to draw your attention to a particular part of a code block, the relevant lines or items are set in bold:

```
# Calling xlsx library
library(xlsx)
# importing xlsxanscombe.xlsx
anscombe_xlsx <- read.xlsx2("xlsxanscombe.xlsx",sheetIndex=1)
```

New terms and **important words** are shown in bold. Words that you see on the screen, in menus or dialog boxes for example, appear in the text like this: "Click on the **Add...** button and select an appropriate ODBC driver and then locate the desired file and give a data sourc name."

Note

Warnings or important notes appear in a box like this.

Tip

Tips and tricks appear like this.

Reader feedback

Feedback from our readers is always welcome. Let us know what you think about this book–what you liked or disliked. Reader feedback is important for us as it helps us develop titles that you will really get the most out of.

To send us general feedback, simply e-mail <feedback@packtpub.com>, and mention the book's title in the subject of your message.

If there is a topic that you have expertise in and you are interested in either writing or contributing to a book, see our author guide at www.packtpub.com/authors.

Customer support

Now that you are the proud owner of a Packt book, we have a number of things to help you t
get the most from your purchase.

Downloading the example code

You can download the example code files from your account at http://www.packtpub.com for
all the Packt Publishing books you have purchased. If you purchased this book elsewhere,
you can visit http://www.packtpub.com/support and register to have the files e-mailed directly
to you.

Errata

Although we have taken every care to ensure the accuracy of our content, mistakes do
happen. If you find a mistake in one of our books—maybe a mistake in the text or the code—
we would be grateful if you could report this to us. By doing so, you can save other readers
from frustration and help us improve subsequent versions of this book. If you find any errata
please report them by visiting http://www.packtpub.com/submit-errata, selecting your book,
clicking on the **Errata Submission Form** link, and entering the details of your errata. Once
your errata are verified, your submission will be accepted and the errata will be uploaded to
our website or added to any list of existing errata under the Errata section of that title.

To view the previously submitted errata, go to
https://www.packtpub.com/books/content/support and enter the name of the book in the
search field. The required information will appear under the **Errata** section.

Piracy

Piracy of copyrighted material on the Internet is an ongoing problem across all media. At
Packt, we take the protection of our copyright and licenses very seriously. If you come acros
any illegal copies of our works in any form on the Internet, please provide us with the locatio
address or website name immediately so that we can pursue a remedy.

Chapter 1. Introduction to R Data Types and Basic Operations

R is an object-oriented programming language and an environment that is a variation of the S language written by Ross Ihaka and Robert Gentlemen (hence, the name R). What can we do using R? The answer is we can do anything we can think of that is logical and/or structural. With R, we can perform data processing, write functions, produce graphs, perform complex data analysis, and also produce our own customized packages (a collection of functions to perform specified tasks) to solve specific problems. We can develop up-to-date statistical techniques through R packages. Most importantly, R is open source and is a freely available software that will remain free.

Assuming that readers have very preliminary or no knowledge of R, the layout of this chapter is divided in to two major sections; the first one will be an introduction to R, and the second major section will relate to data types and basic operations.

The following are the reasons to use R:

- **R is free**: It comes with a license, but we do not have to pay anything to get it. It is not only free, but also open source. We can see the source code, change it as per our own requirements, and also distribute it without violating the license. Academicians across different disciplines around the world reviewed the core of the R system and also contributed to make it better.
- **R is a powerful software**: It is used to perform data processing and data analysis, and to produce a variety of graphs. All the necessary functions for data processing are available in R. It has a substantial collection of libraries (a library is a collection of functions to perform certain types of task), which are written by researchers working in a variety of fields. That is why, whether you are a statistician, biologist, environmentalist, or data scientist, you should find a set of functions that serves your purpose. The graphic system in R is one of the most powerful tools in this era. We have full control over every part of graphs produced in R.
- **R is up-to-date**: R is now one of the standard platforms to implement our research work. We should be able to find an R package suitable for the most recent developments, whatever our field is.
- **R is a community**: R is being developed by a team of volunteers. Also, it includes large communities that are writing new functions every day and that can help us out if we face any problem.
- **R is the language of communication**: R is now becoming a prominent way of sharing new findings with other researchers in this field.

Here is a summary of why we should use R:

- R is free, and it will remain free.
- It involves up-to-date implementation of recent statistical techniques.

- There is flexibility. The user has control over each and every part of a dataset and each component of each output.
- It is customizable based on the user's need.
- It has a large number of built-in libraries.
- It has a cloud-computing feature.
- It has rich graphics.
- It has a wide range of flexible data structures.
- It intelligently handles missing values.

Getting different versions of R

The source code, documentation, and other related files are maintained in the **Comprehensive R Archive Network** (**CRAN**), which can be found at http://cran.r-project.org/. CRAN is a collection of websites that contain identical materials consisting of the R distributions, contributed extensions, and documentation for R and binaries. The user can select anyone of the CRAN sites to download the R software. The user can download the software that is compatible to their computer's platform such as Windows, Mac, and Linux.

To download binaries for different platforms, anyone can use the following links:

- For Linux, http://cran.at.r-project.org/bin/linux/
- For Mac OS X, http://cran.at.r-project.org/bin/macosx/
- For Windows, http://cran.at.r-project.org/bin/windows/

The preceding links are applicable to download the most recent version of R. The latest R Version 3.1.2 (Pumpkin Helmet) was released on October 31, 2014.

To get the old version of R, Windows users can look at the various releases at http://cran.r-project.org/bin/windows/base/old/, and Mac users can look at http://cran.r0-project.org/bin/macosx/old/ to download the desired one.

Installing R on different platforms

To install R on various platforms, the first requirement is to download appropriate binaries tha are compatible with the relevant platform. In this section, we will briefly discuss installation on the Windows platform and will refer users to http://cran.r-project.org/doc/manuals/r-release/R admin.html for the documentation for alternative platforms.

Installing R under Windows is as easy as installing any other software. After downloading the binary file for Windows (it comes with an `.exe` file), the name is for example, `R-3.1.2-win.exe`. This executable file contains binaries for a base distribution and a large number of add-on packages from CRAN. Users can install it just by double-clicking on the file and following the on-screen instructions. There is no special care that needs to be taken during installation; just go with the default selections.