Biostatistics and Biometrics
Open Access Journal
ISSN: 2573-2633

Juniper Publishers
key to the Researchers

Mini Review
Volume 7 Issue 1 - May 2018
DOI: 10.19080/BBOAJ.2018.07.555704

Biostat Biometrics Open Acc J

A Brief Review of the Dook, R for Data Science

Dale Bowman and Lih-Yuan Deng*

Department of Mathematical Sciences, University of Memphis, USA

Submission: February 21, 2018; **Published:** May 25, 2018

***Corresponding author:** Lih Yuan Deng, Department of Mathematical Sciences, University of Memphis, Memphis, Tennessee, USA, Tel: 901-678-3134; Fax: 901-678-2480; Email: lihdeng@memphis.edu

Abstract

This article reviews a recent textbook that introduces a data scientist to the *tidyverse*, an environment in R that provides useful packages for importing data, tidying data, transforming data, visualization, modeling and communication of results. All of these tools are needed for managing a data science project.

Keywords: Tidy verse; Tidy data principles; Pipe; Tibble; GG Plot2; Communicate results; Data science; Exploration; Important packages; Environment; Data handling easier; Visualization; Data manipulation; Existing Variables; Outliers; Data Analysis; Graphical Techniques; Traditional data Fame; Operating system; Strings; Writing functions; Loops; R's Modeling functions

Introduction

R for Data Science

Import, Tidy, Transform, Visualize and Model Data is a recent text written by Hadley Wickham, developer of several widely used R packages including *ggplot2*, and Garrett Grolemund. The text was published by O'Reilly in 2016. The text is available online at [1] with source code for the examples at [2]. The text is divided into five parts illustrating the four-step model of the tools needed for a data science project:

I. Import, tidy, and transform the data,

II. Visualize and explore the data,

III. Model the data and

IV. Communicate results.

The book is not organized in the order in which a data science project would be completed, but rather in the order that the authors believe is the best way to learn the individual components. The first part of the book involves basic tools needed for exploration of data including *ggplot2* for data visualization, data transformations with *dplyr*, and basic concepts of exploratory data analysis. The second part of the text involves importation of the data, tidying of the data, and transforming the data as needed. This process is termed wrangling and the authors give a nice introduction to the problems that can arise in this stage – the initial one in the data science process. For the third part, the authors take the reader through some of the basic programming concepts needed to successfully complete the four-step model. The basics of modeling data for exploration is described in the fourth part of the text, with an emphasis on linear models. The final part of the book focuses on communicating results to decision makers and collaborators using R markdown.

Discussion

The textbook is designed to provide the reader with a solid foundation in some important packages in the *tidyverse* environment. Tidy data principles provide powerful tools for making data handling easier and more efficient. Tidy data has been formatted so that

a) Each column contains a single variable,

b) Each row contains a single observation, and

c) Each value has its own cell.

The basic *tidyverse* packages that are useful for working with data include: *ggplot2* for visualization, *dplyr* for data manipulation, *tidyr* for data tidying, *readr* for fast data import, *purr* for functional programming, and tibble a modern data frame.

The first part of the textbook focuses on exploration of an already wrangled data set. The graphics package in the tidyverse, *ggplot2* is investigated in some detail through applications to several data sets. The key functions of the data manipulation package, *dplyr*, allow the user to pick observations by value, reorder observations, pick variables by their name, create new variables as functions of existing variables, and summarize data by certain attributes. The final major component of this first part of the book takes the reader through the steps of exploratory data analysis, including examining variation, outliers, relationships between variables and examining patterns in data distributions using primarily graphical techniques. The idea of the

pipe (borrowed from UNIX Sell), one of the key components for working in the tidyverse, is introduced.

The second part of the book on data wrangling introduces the concept of a tibble, how to create one and how it differs from a traditional data frame. From there the *readr* package is discussed for importing rectangular data into R. The advantages of the *readr* functions over their traditional counter parts (read. csv() for example) include speed, independence from operating system, and the creation of *tibbles* instead of data frames. Following the import process, this part of the text moves on to tidying the data using functions in the *tidyr* package. Since most of the data a scientist will encounter will not be tidy, the text works through many different data structures that might occur and provide functions for *tidying* the data. Principles of how to work with relational data tables are covered using functions in *dplyr*. Techniques for working with strings, factors and data/time variables are also discussed [3].

The third part of the text includes more on using the *pipe* and other good programming practices including writing functions for code that will appear in several places in a program, the use of vectors in R as the component variables in a *tibble*, working with lists in R, and programming using for and while loops using purr. The material is covered thoroughly and is accessible to those without a programming background.

Modeling is introduced in the fourth part of the book with a focus on the linear model as the next step in the data science project. Emphasis is on modeling for hypothesis generation rather than confirmatory analysis. The *modelr* package is used in conjunction with base R's modeling functions so that they work with the *pipe*. The modeling process is first introduced on some simple simulated data to illustrate the components. Later models are built using real data. Methods of handling many models using *purrr* and *broom* packages are discussed. While no comprehensive discussion was given, many valuable references on the subject are given.

The final part of the book is focused on communicating results. R Markdown is introduced to integrate code, graphs, and text into reports for decision makers and for fellow analysts. The exploratory graphics discussed in the first part are transformed into expository graphics to aid decision makers for a better understanding the data. The reader can learn how to produce other output, such as websites, presentations and dashboards using R Markdown [4].

Conclusion

In summary, R for Data Science is a very useful text for introducing data scientists to the principles required to work in the *tidyverse*. Advantages of *tibbles* over data frames, *readr* over base R input functions, the pipe over traditional programming, and the advantages of *tidyr* and *dplyr* for data manipulation and tidying are clearly specified. The use of *ggplot2* to produce high quality graphics is well covered. The authors do not presume any previous programming experience in R nor extensive statistical knowledge, which makes this book suitable for introducing the *tidyverse* principles to most data scientists. However, the book does not overreach and indeed, as pointed out in the preface, it is not intended to cover all topics in data science. In particular, none of the data sets used through the text are big data sets and all are rectangular. The book is RStudio specific although RStudio is not a requirement for working in the *tidyverse*. Overall, it is a well written, gentle introduction to the *tidyverse* for programmers and non-programmers alike.

References

1. http://r4ds.had.co.nz
2. Wickham Hadley, Grolemund Garrett (2016) R for Data Science (1st edn) O'Reilly Media, Inc, USA.
3. Wickham Hadley (2009) ggplot2: Elegant Graphics for Data Analysis (1st edition), Springer-Verlag, New York, USA.

DOI: 10.19080/BBOJ.2018.07.555704

How to cite this article: Dale B, Lih-Yuan D. A Brief Review of the Dook, R for Data Science. Biostat Biometrics Open Acc J. 2018; 7(1): 555704. DOI: 10.19080/BBOJ.2018.07.555704.