Chapter 2
Introduction to R

A language for data analysis and graphics. This definition of R was used by Ross Ihaka and Robert Gentleman in the title of their 1996 paper (Ihaka and Gentleman 1996) outlining their experience of designing and implementing the R software. It's safe to say this remains the essence of what R is; however, it's tough to encapsulate such a diverse programming language into a single phrase.

During the last decade, the R programming language has become one of the most widely used tools for statistics and data science. Its application runs the gamut from data preprocessing, cleaning, web scraping and visualization to a wide range of analytic tasks such as computational statistics, econometrics, optimization, and natural language processing. In 2012 R had over two million users and continues to grow by double-digit percentage points every year. R has become an essential analytic software throughout industry; being used by organizations such as Google, Facebook, New York Times, Twitter, Etsy, Department of Defense, and even in presidential political campaigns. So what makes R such a popular tool?

2.1 Open Source

R is an *open source* software created over 20 years ago by Ihaka and Gentleman at the University of Auckland, New Zealand. However, its history is even longer as its lineage goes back to the S programming language created by John Chambers out of Bell Labs back in the 1970s.[1] R is actually a combination of S with lexical scoping semantics inspired by Scheme (Morandat and Hill 2012). Whereas the resulting language is very similar in appearance to S, the underlying implementation and semantics are derived from Scheme. Unbeknownst to many the S language has been a popular vehicle for research in statistical methodology, and R provides an *open source* route to participate in that activity.

[1] Consequently, R is named partly after its authors (Ross and Robert) and partly as a play on the name of S.

© Springer International Publishing Switzerland 2016

B.C. Boehmke, *Data Wrangling with R*, Use R!, DOI 10.1007/978-3-319-45599-0_2

Although the history of S and R is interesting,[2] the principal artifact to observe is that R is an *open source* software. Although some contest that open-source software is merely a "craze",[3] most evidence suggests that open-source is here to stay and represents a *new*[4] norm for programming languages. Open-source software such as R blurs the distinction between developer and user, which provides the ability to extend and modify the analytic functionality to your, or your organization's needs. The data analysis process is rarely restricted to just a handful of tasks with predictable input and outputs that can be pre-defined by a fixed user interface as is common in proprietary software. Rather, as previously mentioned in the introduction, data analyses include unique, different, and often multiple requirements regarding the specific tasks involved. Open source software allows more flexibility for you, the data analyst, to manage how data are being transformed, manipulated, and modeled "under the hood" of software rather than relying on "stiff" point and click software interfaces. Open source also allows you to operate on every major platform rather than be restricted to what your personal budget allows or the idiosyncratic purchases of organizations.

This invariably leads to new expectations for data analysts; however, organizations are proving to greatly value the increased technical abilities of open source data analysts as evidenced by a recent O'Reilly survey revealing that data analysts focusing on open source technologies make more money than those still dealing in proprietary technologies.

2.2 Flexibility

Another benefit of open source is that anybody can access the source code, modify and improve it. As a result, many excellent programmers contribute to improving existing R code and developing new capabilities. Researchers from all walks of life (academic institutions, industry, and focus groups such as RStudio[5] and rOpenSci[6]) are contributing to advancements of R's capabilities and best practices. This has resulted in some powerful tools that advance both statistical and non-statistical modeling capabilities that are taking data analysis to new levels.

[2] See Roger Peng's *R programming for Data Science* for further, yet concise, details on S and R's history.

[3] This was recently argued by Pollack, Klimberg, and Boklage (2015) which was appropriately rebutted by Boehmke and Jackson (2016).

[4] Open-source is far from new as it has been around for decades (i.e. A-2 in the 1950s, IBM's ACP in the '60s, Tiny BASIC in the '70s) but has gained prominence since the late 1990s.

[5] https://www.rstudio.com

[6] https://ropensci.org/packages

Many researchers in academic institutions are using and developing R code to develop the latest techniques in statistics and machine learning. As part of their research, they often publish an R package to accompany their research articles.[7] This provides immediate access to the latest analytic techniques and implementations. And this research is not soley focused on generalized algorithms as many new capabilities are in the form of advancing analytic algorithms for tasks in specific domains. A quick assessment of the different task domains[8] for which code is being developed illustrates the wide spectrum—econometrics, finance, chemometrics and computational physics, pharmacokinetics, social sciences, etc.

Powerful tools are also being developed to perform many tasks that greatly aid the data analysis process. This is not limited to just new ways to wrangle your data but also new ways to visualize and communicate data. R packages are now making it easier than ever to create interactive graphics and websites and produce sophisticated HTML and PDF reports. R packages are also integrating communication with high-performance programming languages such as C, Fortran, and C++ making data analysis more powerful, efficient, and posthaste than ever.

So although the analytic mantra "*use the right tool for the problem*" should always be in our prefrontal cortex, the advancements and flexibility of R is making it the right tool for many problems.

2.3 Community

The R community is fantastically diverse and engaged. On a daily basis, the R community generates opportunities and resources for learning about R. These cover the full spectrum of training—books, online courses, R user groups, workshops, conferences, etc. And with over two million users and developers, finding help and technical expertise is only a simple click away. Support is available through R mailing lists, Q&A websites, social media networks, and numerous blogs.

So now that you know how awesome R is, it's time to learn how to use it.

Bibliography

Ihaka, Ross, and Robert Gentleman. "R: A language for data analysis and graphics." Journal of Computational and Graphical Statistics 5, no. 3 (1996):299–314.

Morandat, Floréal, Brandon Hill, Leo Osvald, and Jan Vitek. "Evaluating the design of the R language." In European Conference on Object-Oriented Programming, pp. 104–131. Springer Berlin Heidelberg, 2012.

Pollack, R. D., Klimberg, R. K., and Boklage, S.H. "The true cost of 'free' statistical software." OR/MS Today, vol. 42, no. 5 (2015):34–35.

Boehmke, Bradley C. and Jackson, Ross A. "Unpacking the true cost of 'free' statistical software." OR/MS Today, vol. 43, no. 1 (2016):26–27.

[7] See *The Journal of Statistical Software* and *The R Journal*.

[8] https://cran.r-project.org/web/views/

http://www.springer.com/978-3-319-45598-3

Data Wrangling with R

Boehmke, B.

2016, XII, 238 p. 24 illus., 10 illus. in color., Softcover
ISBN: 978-3-319-45598-3