

Ten simple rules for selecting an R package

Caroline J. Wendt ¹ , ² , G. Brooke Anderson ³ *

1 Department of Statistics, Colorado State University, Fort Collins, Colorado, United States of America

2 Department of Mathematics, Colorado State University, Fort Collins, Colorado, United States of America

3 Department of Environmental & Radiological Health Sciences, Colorado State University, Fort Collins, Colorado, United States of America

* Corresponding author: Brooke.Anderson@colostate.edu

Abstract

Write the abstract here.

Author summary

Write the author summary here.

Text based on plos sample manuscript, see

<http://journals.plos.org/ploscompbiol/s/latex>

Introduction

Explain what R is and how its package ecosystem works.

Points:

- Open source project, where many people contribute with their own extensions
- Large variation in the quality of different extensions (packages)
- That some R users, particularly new ones, struggle with finding and picking which packages to use.

Ideas of 10 things

Finding packages:

- CRAN task views
- Textbooks (“[x] with R”). May not be latest...
- Google searches, social media (#rstats)
- Conferences (and online streams of those). RStudio, UseR.

Picking a good package:

- On a public repository like CRAN or Bioconductor. Explain more about these repositories and what their standards are. Explain their role in the community. Give the alternative ways that R packages can be shared (GitHub, zipped file posted somewhere else). How these regularly check code and help with managing the web of dependencies.

- Quality of the documentation. Types of documentation (help files, vignette, packagedown website, bookdown book). 22
- Coverage by tests. Explain about unit testing and how it can help control quality. 23
- Peer review. ROpenSci. Associated with a peer reviewed paper. Associated with a book put out by a scientific publisher? 24
- Looking up package authors. Is there role in R development (RStudio, some big bio labs)? Is the work part of their work from an academic lab? Do they have a history of a lot of R development? GitHub profile. Google scholar profile. Also, is it a team of developers? Robust team? 25
- Evidence of established package. Lots of version. Clear NEWS providing explanations of changes. History of Issues and those being resolved. 26
- Exploring the code yourself. How open source framework provides this. GitHub mirror of CRAN if you don't want to download the zipped package file yourself. 27

Here are two sample references: [1,2]. 28

Introduction 29

R is a language and environment for statistical computing and graphics that was developed by statisticians and is collaboratively maintained by an international core group of contributors. Unlike many popular proprietary languages (e.g., MATLAB, SAS, SPSS), R is highly extensible, free and open-source software; the user can access and thus change, extend, and share code for desired applications. Accordingly, a vibrant community of R users has emerged, many of which engage in the development of extensions to the functionality of base R software known as packages. A prominent contributor in the R community, Hadley Wickham, views functional programming as analogous to following a recipe; to conceptualize packages, imagine R is the kitchen and packages are the special gadgets which allow you to cook and bake new recipes. R packages are coding delectables that enable the user to perform practical tasks (e.g., wrangling and cleaning data frames, designing interactive apps for visualizing data, performing dimensionality reduction) and solve problems (e.g., training regression and classification models, assessing the beta diversity of a population, analyzing gene expression microarray data) with interesting techniques. 30

As a natural consequence of the open-source nature of R, there is variation in the quality of different packages among the numerous choices that exist. The advanced R user—having developed an intuition for their workflow—may tend to be relatively confident when searching for and selecting packages. By contrast, a common experience that characterizes learning R at the outset is the struggle to 1) find a package to accomplish a particular task and 2) choose the best package to perform that task. Even so, there remain obscure and complicated problems that morph selecting an R package into a barrier despite experience. 31

In coding as in life, we endeavor to make choices that optimize outcomes. Just as one may go about shopping for shoes, deciding which graduate program to pursue, or conducting a literature review, there is a science behind selection. We inform our decisions by assessing, comparing, and filtering options based on indicators of quality such as utility, association, and reputation. Likewise, choosing an R package requires attending to similar details. We outline ten simple rules for finding and selecting R packages so that you will spend less time searching for the right tools and more time coding delicious recipes. 32

List of 10 rules	68
(currently in no particular order and not precisely worded)	69
1. Consider your purpose	70
• What do you want to use the package to accomplish?	71
• features	72
• functions	73
• organization	74
• package description	75
• compare similar options	76
2. Spend time searching; find and collect options	77
• internet searches (keyword "... in R")	78
• textbooks ("[x] with R" series)	79
• tutorials	80
• courses	81
• social media (#rstats)	82
• conferences (e.g., RStudio, useR!)	83
• consult collaborators	84
• CRAN task views	85
3. Check how it's shared	86
• check repository association	87
– CRAN	88
– Bioconductor	89
– GitHub	90
– GitLab (alternative to GitHub)	91
– ROpenSci (runs its own repository, only includes ones it has peer-reviewed)	92
– Self-hosted repositories (can be made with the <code>drat</code> package; see paper)	93
– purpose of repositories: mechanisms of quality control that regularly check code and manage webs of dependencies	94
– code and manage webs of dependencies	95
• alternative ways R packages can be shared (not repo)	96
– zipped file	97
– collaborators	98
4. Explore the availability and quality of help	99
• help files	100
• <code>help()</code>	101
• vignettes	102
• <code>DOCUMENTATION</code> file	103
• "cheatsheets" from RStudio	104
• <code>RDocumentation</code> (key word search, task views)	105
• websites (e.g., <code>packagedown</code>)	106
• <code>bookdown</code> books	107
• compare documentation completeness and resource quality	108
• find ways to get help beyond initial documentation	109
• listservs	110
• online communities	111

• Stack Overflow (frequency of questions and answers on the topic)	112
• See if GitHub repo for the package seems responsive to Issues	113
• Rcpp is an example of high-quality help	114
– associated book	115
– maintainer, Dirk, is known to be responsive to user questions (listserv)	116
– ample documentation including examples to get started	117
5. Verify the credibility of the author(s)	118
• team or single author (robust team?)	119
• associations (e.g., academia, industry, labs)	120
• expertise	121
• reputation	122
• experience (e.g., portfolio of packages, history of R development)	123
• role in R development (e.g., RStudio, regarded bio labs)	124
• profiles (e.g., GitHub, Google Scholar, Research Gate, Twitter)	125
6. Investigate the package development	126
• best practices	127
• unit testing (manage quality control)	128
• dependencies	129
• coverage by tests	130
• number of versions	131
• clarity of NEWS (explain updates and changes)	132
• GitHub Issues (history, resolution)	133
7. Read, research literature, seek evidence of peer review	134
• publications	135
• package itself	136
• papers about the package	137
• ROpenSci	138
• associations with books or publications from scientific publishers	139
8. Quantify how established the package is	140
• dependencies	141
• versions	142
• updates	143
• number of downloads	144
• popularity	145
• leaderboard	146
• ranking systems	147
9. Put the package to the test	148
• explore code	149
• interact with trial and error	150
• get a feel for using it in context of your goal	151
• open-source framework	152
• GitHub mirror of CRAN as an alternative to downloading zipped package file	153
10. Develop your own package	154
• necessity	155
• innovative idea	156
• novel approach or method	157
• unique and specialized purpose	158

References

1. Feynman RP, Vernon Jr. FL. The theory of a general quantum system interacting with a linear dissipative system. *Annals of Physics*. 1963;24: 118–173. doi:10.1016/0003-4916(63)90068-X
2. Dirac PAM. The lorentz transformation and absolute time. *Physica*. 1953;19: 888–896. doi:10.1016/S0031-8914(53)80099-6

159
160
161
162
163
164