Tools for Reproducible Workflows in R

January, 2023

Table of Contents

# About this Course

Reproducibility of data analyses can be enhanced through the use of tools designed to manage the complexity involved in any data analysis designed to address an important scientific question. We focus on a few software tools that aid in project organization, collaboration, auditability of analyses, and maintaining the integrity of data and code. In this course, we view a data analysis as a complex system with many integrated parts that together produce analytic results. The tools we focus on here allow data analysts to diagnose unexpected results, quickly identify problems with data and code, and provide a basis for managing the dynamic nature of data analysis.

## 0.1 Available course formats

This course is available in multiple formats which allows you to take it in the way that best suites your needs. You can take it for certificate which can be for free or fee.

* The material for this course can be viewed without login requirement on this [Bookdown website](LINK%20HERE). This format might be most appropriate for you if you rely on screen-reader technology.
* This course can be taken for [free certification through Leanpub](LINK%20HERE).
* This course can be taken on [Coursera for certification here](LINK%20HERE) (but it is not available for free on Coursera).
* Our courses are open source, you can find the [source material for this course on GitHub](LINK%20HERE).

# 1 Introduction

In this course we will explore a variety of tools that can assist with data analysis from a broad range of fields. The tools we will cover may take some time to get used to, but the payoff will be immeasurable. Not only are these skills valuable for career advancement, they will also make your work-life easier. The tools will enhance your ability to reproduce your work across similar projects, stay organized, collaborate with others effectively, and more.

## 1.1 Motivation

Many researchers are self-taught when it comes to computer science. However, data analysis has become a requirement for most researchers. The ability to smoothly work in a reproducible manner not only makes for easier more maintainable workflows, it also improves scientific rigor and transparency.

This course will help learners to use tools that will make their data analytic workflows more organized, more understandable to collaborators (and your future self!), and ultimately more efficient.

## 1.2 Target Audience

This course is intended for people conducting data analyses at the level of a graduate student or higher. The course is designed so that the majority of the material is presented in a high-level manner that should be applicable to researchers working in a broad range of areas. The course is centered around the R programming language, a widely used statistical analysis software package.

## 1.3 Curriculum

The course covers…

## 1.4 Learning Objectives

* Implement basic project organization tools:
  + Setup and configure RStudio/RStudio projects for data analysis (here package and file structure/paths)
  + Install and configure ProjectTemplate package for formalizing and automating workflows
* Apply the pointblank package for validation of tabular data
* Write functions and package them
* Apply the testthat package for building software unit tests
* Setup and use Git repositories for version control of code
* Interface with GitHub to share Git repositories for collaboration; execute GitHub-based workflows
  + Pull Requests
  + Code review
  + Issues
  + Discussions

References will include Gillespie and Lovelace ([2021](#ref-gillespie_efficient_2021)), Riederer ([2020](#ref-riederer_column_2020)), Timbers, Campbell, and Lee ([2022](#ref-timbers_data_nodate)).

Code review references will include “About Scientific Code Review” ([n.d.](#ref-hutchdatascience_code_review)), Radigan ([n.d.](#ref-radigan_what_nodate)), Parker ([2017](#ref-parker_opinionated_2017)), Bodner ([2018](#ref-bodner_10_2018)).

# 2 A new chapter

If you haven’t yet read the getting started Wiki pages; [start there](https://www.ottrproject.org/getting_started.html).

To see the rendered version of this chapter and the rest of the template, see here: <https://jhudatascience.org/OTTR_Template/>.

Every chapter needs to start out with this chunk of code:

## 2.1 Learning Objectives

Every chapter also needs Learning objectives that will look like this:

This chapter will cover:

* {You can use <https://tips.uark.edu/using-blooms-taxonomy/> to define some learning objectives here}
* {Another learning objective}

## 2.2 Libraries

For this chapter, we’ll need the following packages attached:

\*Remember to add [any additional packages you need to your course’s own docker image](https://github.com/jhudsl/OTTR_Template/wiki/Using-Docker#starting-a-new-docker-image).

library(magrittr)

## 2.3 Topic of Section

You can write all your text in sections like this, using ## to indicate a new header. you can use additional pound symbols to create lower levels of headers.

See [here](https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf) for additional general information about how you can format text within R Markdown files. In addition, see [here](https://pandoc.org/MANUAL.html#pandocs-markdown) for more in depth and advanced options.

### 2.3.1 Subtopic

Here’s a subheading (using three pound symbols) and some text in this subsection!

## 2.4 Code examples

You can demonstrate code like this:

output\_dir <- file.path("resources", "code\_output")  
if (!dir.exists(output\_dir)) {  
 dir.create(output\_dir)  
}

And make plots too:

hist\_plot <- hist(iris$Sepal.Length)



You can also save these plots to file:

png(file.path(output\_dir, "test\_plot.png"))  
hist\_plot

## $breaks  
## [1] 4.0 4.5 5.0 5.5 6.0 6.5 7.0 7.5 8.0  
##   
## $counts  
## [1] 5 27 27 30 31 18 6 6  
##   
## $density  
## [1] 0.06666667 0.36000000 0.36000000 0.40000000 0.41333333 0.24000000 0.08000000  
## [8] 0.08000000  
##   
## $mids  
## [1] 4.25 4.75 5.25 5.75 6.25 6.75 7.25 7.75  
##   
## $xname  
## [1] "iris$Sepal.Length"  
##   
## $equidist  
## [1] TRUE  
##   
## attr(,"class")  
## [1] "histogram"

dev.off()

## png   
## 2

## 2.5 Image example

How to include a Google slide. It’s simplest to use the ottrpal package:



But if you have the slide or some other image locally downloaded you can also use HTML like this:

## 2.6 Video examples

You may also want to embed videos in your course. If alternatively, you just want to include a link you can do so like this:

Check out this [link to a video](https://www.youtube.com/embed/VOCYL-FNbr0) using markdown syntax.

### 2.6.1 Using knitr

To embed videos in your course, you can use knitr::include\_url() like this: Note that you should use echo=FALSE in the code chunk because we don’t want the code part of this to show up. If you are unfamiliar with [how R Markdown code chunks work, read this](https://rmarkdown.rstudio.com/lesson-3.html).

## PhantomJS not found. You can install it with webshot::install\_phantomjs(). If it is installed, please make sure the phantomjs executable can be found via the PATH variable.

### 2.6.2 Using HTML

## 2.7 File examples

You can again use simple markdown syntax to just include a link to a file like so:

[A file](https://www.bgsu.edu/content/dam/BGSU/center-for-faculty-excellence/docs/TLGuides/TLGuide-Learning-Objectives.pdf).

Alternatively you can embed files like PDFs.

### 2.7.1 Using knitr

### 2.7.2 Using HTML

## 2.8 Website Examples

Yet again you can use a link to a website like so:

[A Website](https://yihui.org)

You might want to have users open a website in a new tab by default, especially if they need to reference both the course and a resource at once.

[A Website](https://yihui.org)

Or, you can embed some websites.

### 2.8.1 Using knitr

This works:

### 2.8.2 Using HTML

If you’d like the URL to show up in a new tab you can do this:

<a href="https://www.linkedin.com" target="\_blank">LinkedIn</a>

## 2.9 Citation examples

We can put citations at the end of a sentence like this ([Allaire et al. 2021](#ref-rmarkdown2021)). Or multiple citations Xie, Allaire, and Grolemund ([2018](#ref-Xie2018)).

but they need a ; separator ([Allaire et al. 2021](#ref-rmarkdown2021); [Xie, Allaire, and Grolemund 2018](#ref-Xie2018)).

In text, we can put citations like this Allaire et al. ([2021](#ref-rmarkdown2021)).

## 2.10 Stylized boxes

Occasionally, you might find it useful to emphasize a particular piece of information. To help you do so, we have provided css code and images (no need for you to worry about that!) to create the following stylized boxes.

You can use these boxes in your course with either of two options: using HTML code or Pandoc syntax.

### 2.10.1 Using rmarkdown container syntax

The rmarkdown package allows for a different syntax to be converted to the HTML that you just saw and also allows for conversion to LaTeX. See the [Bookdown](https://bookdown.org/yihui/rmarkdown-cookbook/custom-blocks.html) documentation for more information ([Xie, Dervieux, and Riederer 2020](#ref-Xie2020)). Note that Bookdown uses Pandoc.

::: {.notice}  
Note using rmarkdown syntax.  
  
:::

Note using rmarkdown syntax.

As an example you might do something like this:

Please click on the subsection headers in the left hand navigation bar (e.g., 2.1, 4.3) a second time to expand the table of contents and enable the scroll\_highlight feature ([see more](introduction.html#scroll-highlight))

### 2.10.2 Using HTML

To add a warning box like the following use:

<div class = "notice">  
Followed by the text you want inside  
</div>

This will create the following:

Followed by the text you want inside

Here is a <div class = "warning"> box:

Note text

Here is a <div class = "github"> box:

GitHub text

Here is a <div class = "dictionary"> box:

dictionary text

Here is a <div class = "reflection"> box:

reflection text

## 2.11 Dropdown summaries

You can hide additional information in a dropdown menu

Here’s more words that are hidden.

## 2.12 Print out session info

You should print out session info when you have code for [reproducibility purposes](https://jhudatascience.org/Reproducibility_in_Cancer_Informatics/managing-package-versions.html).

devtools::session\_info()

## ─ Session info ───────────────────────────────────────────────────────────────  
## setting value   
## version R version 4.0.2 (2020-06-22)  
## os Ubuntu 20.04.3 LTS   
## system x86\_64, linux-gnu   
## ui X11   
## language (EN)   
## collate en\_US.UTF-8   
## ctype en\_US.UTF-8   
## tz Etc/UTC   
## date 2023-01-26   
##   
## ─ Packages ───────────────────────────────────────────────────────────────────  
## package \* version date lib source   
## assertthat 0.2.1 2019-03-21 [1] RSPM (R 4.0.3)   
## bookdown 0.24 2022-02-15 [1] Github (rstudio/bookdown@88bc4ea)   
## callr 3.4.4 2020-09-07 [1] RSPM (R 4.0.2)   
## cli 2.0.2 2020-02-28 [1] RSPM (R 4.0.0)   
## crayon 1.3.4 2017-09-16 [1] RSPM (R 4.0.0)   
## curl 4.3 2019-12-02 [1] RSPM (R 4.0.3)   
## desc 1.2.0 2018-05-01 [1] RSPM (R 4.0.3)   
## devtools 2.3.2 2020-09-18 [1] RSPM (R 4.0.3)   
## digest 0.6.25 2020-02-23 [1] RSPM (R 4.0.0)   
## ellipsis 0.3.1 2020-05-15 [1] RSPM (R 4.0.3)   
## evaluate 0.14 2019-05-28 [1] RSPM (R 4.0.3)   
## fansi 0.4.1 2020-01-08 [1] RSPM (R 4.0.0)   
## fs 1.5.0 2020-07-31 [1] RSPM (R 4.0.3)   
## glue 1.6.1 2022-01-22 [1] CRAN (R 4.0.2)   
## highr 0.8 2019-03-20 [1] RSPM (R 4.0.3)   
## hms 0.5.3 2020-01-08 [1] RSPM (R 4.0.0)   
## htmltools 0.5.0 2020-06-16 [1] RSPM (R 4.0.1)   
## httr 1.4.2 2020-07-20 [1] RSPM (R 4.0.3)   
## knitr 1.33 2022-02-15 [1] Github (yihui/knitr@a1052d1)   
## lifecycle 1.0.0 2021-02-15 [1] CRAN (R 4.0.2)   
## magrittr \* 2.0.2 2022-01-26 [1] CRAN (R 4.0.2)   
## memoise 1.1.0 2017-04-21 [1] RSPM (R 4.0.0)   
## ottrpal 0.1.2 2022-02-15 [1] Github (jhudsl/ottrpal@1018848)   
## pillar 1.4.6 2020-07-10 [1] RSPM (R 4.0.2)   
## pkgbuild 1.1.0 2020-07-13 [1] RSPM (R 4.0.2)   
## pkgconfig 2.0.3 2019-09-22 [1] RSPM (R 4.0.3)   
## pkgload 1.1.0 2020-05-29 [1] RSPM (R 4.0.3)   
## prettyunits 1.1.1 2020-01-24 [1] RSPM (R 4.0.3)   
## processx 3.4.4 2020-09-03 [1] RSPM (R 4.0.2)   
## ps 1.3.4 2020-08-11 [1] RSPM (R 4.0.2)   
## purrr 0.3.4 2020-04-17 [1] RSPM (R 4.0.3)   
## R6 2.4.1 2019-11-12 [1] RSPM (R 4.0.0)   
## readr 1.4.0 2020-10-05 [1] RSPM (R 4.0.2)   
## remotes 2.2.0 2020-07-21 [1] RSPM (R 4.0.3)   
## rlang 0.4.10 2022-02-15 [1] Github (r-lib/rlang@f0c9be5)   
## rmarkdown 2.10 2022-02-15 [1] Github (rstudio/rmarkdown@02d3c25)  
## rprojroot 2.0.2 2020-11-15 [1] CRAN (R 4.0.2)   
## sessioninfo 1.1.1 2018-11-05 [1] RSPM (R 4.0.3)   
## stringi 1.5.3 2020-09-09 [1] RSPM (R 4.0.3)   
## stringr 1.4.0 2019-02-10 [1] RSPM (R 4.0.3)   
## testthat 3.0.1 2022-02-15 [1] Github (R-lib/testthat@e99155a)   
## tibble 3.0.3 2020-07-10 [1] RSPM (R 4.0.2)   
## usethis 2.1.5.9000 2022-02-15 [1] Github (r-lib/usethis@57b109a)   
## vctrs 0.3.4 2020-08-29 [1] RSPM (R 4.0.2)   
## webshot 0.5.2 2019-11-22 [1] RSPM (R 4.0.3)   
## withr 2.3.0 2020-09-22 [1] RSPM (R 4.0.2)   
## xfun 0.26 2022-02-15 [1] Github (yihui/xfun@74c2a66)   
## yaml 2.2.1 2020-02-01 [1] RSPM (R 4.0.3)   
##   
## [1] /usr/local/lib/R/site-library  
## [2] /usr/local/lib/R/library

# About the Authors

These credits are based on our [course contributors table guidelines](https://www.ottrproject.org/more_features.html#giving-credits-to-contributors).

| Credits | Names |
| --- | --- |
| **Pedagogy** |  |
| Lead Content Instructor(s) | [FirstName LastName](link%20to%20personal%20website) |
| Lecturer(s) (include chapter name/link in parentheses if only for specific chapters) - make new line if more than one chapter involved | Delivered the course in some way - video or audio |
| Content Author(s) (include chapter name/link in parentheses if only for specific chapters) - make new line if more than one chapter involved | If any other authors besides lead instructor |
| Content Contributor(s) (include section name/link in parentheses) - make new line if more than one section involved | Wrote less than a chapter |
| Content Editor(s)/Reviewer(s) | Checked your content |
| Content Director(s) | Helped guide the content direction |
| Content Consultants (include chapter name/link in parentheses or word “General”) - make new line if more than one chapter involved | Gave high level advice on content |
| Acknowledgments | Gave small assistance to content but not to the level of consulting |
| **Production** |  |
| Content Publisher(s) | Helped with publishing platform |
| Content Publishing Reviewer(s) | Reviewed overall content and aesthetics on publishing platform |
| **Technical** |  |
| Course Publishing Engineer(s) | Helped with the code for the technical aspects related to the specific course generation |
| Template Publishing Engineers | [Candace Savonen](https://www.cansavvy.com/), [Carrie Wright](https://carriewright11.github.io/) |
| Publishing Maintenance Engineer | [Candace Savonen](https://www.cansavvy.com/) |
| Technical Publishing Stylists | [Carrie Wright](https://carriewright11.github.io/), [Candace Savonen](https://www.cansavvy.com/) |
| Package Developers ([ottrpal](https://github.com/jhudsl/ottrpal)) [Candace Savonen](https://www.cansavvy.com/), [John Muschelli](https://johnmuschelli.com/), [Carrie Wright](https://carriewright11.github.io/) |  |
| **Art and Design** |  |
| Illustrator(s) | Created graphics for the course |
| Figure Artist(s) | Created figures/plots for course |
| Videographer(s) | Filmed videos |
| Videography Editor(s) | Edited film |
| Audiographer(s) | Recorded audio |
| Audiography Editor(s) | Edited audio recordings |
| **Funding** |  |
| Funder(s) | Institution/individual who funded course including grant number |
| Funding Staff | Staff members who help with funding |

## ─ Session info ───────────────────────────────────────────────────────────────  
## setting value   
## version R version 4.0.2 (2020-06-22)  
## os Ubuntu 20.04.3 LTS   
## system x86\_64, linux-gnu   
## ui X11   
## language (EN)   
## collate en\_US.UTF-8   
## ctype en\_US.UTF-8   
## tz Etc/UTC   
## date 2023-01-26   
##   
## ─ Packages ───────────────────────────────────────────────────────────────────  
## package \* version date lib source   
## assertthat 0.2.1 2019-03-21 [1] RSPM (R 4.0.3)   
## bookdown 0.24 2022-02-15 [1] Github (rstudio/bookdown@88bc4ea)   
## callr 3.4.4 2020-09-07 [1] RSPM (R 4.0.2)   
## cli 2.0.2 2020-02-28 [1] RSPM (R 4.0.0)   
## crayon 1.3.4 2017-09-16 [1] RSPM (R 4.0.0)   
## desc 1.2.0 2018-05-01 [1] RSPM (R 4.0.3)   
## devtools 2.3.2 2020-09-18 [1] RSPM (R 4.0.3)   
## digest 0.6.25 2020-02-23 [1] RSPM (R 4.0.0)   
## ellipsis 0.3.1 2020-05-15 [1] RSPM (R 4.0.3)   
## evaluate 0.14 2019-05-28 [1] RSPM (R 4.0.3)   
## fansi 0.4.1 2020-01-08 [1] RSPM (R 4.0.0)   
## fs 1.5.0 2020-07-31 [1] RSPM (R 4.0.3)   
## glue 1.6.1 2022-01-22 [1] CRAN (R 4.0.2)   
## htmltools 0.5.0 2020-06-16 [1] RSPM (R 4.0.1)   
## knitr 1.33 2022-02-15 [1] Github (yihui/knitr@a1052d1)   
## lifecycle 1.0.0 2021-02-15 [1] CRAN (R 4.0.2)   
## magrittr 2.0.2 2022-01-26 [1] CRAN (R 4.0.2)   
## memoise 1.1.0 2017-04-21 [1] RSPM (R 4.0.0)   
## pkgbuild 1.1.0 2020-07-13 [1] RSPM (R 4.0.2)   
## pkgload 1.1.0 2020-05-29 [1] RSPM (R 4.0.3)   
## prettyunits 1.1.1 2020-01-24 [1] RSPM (R 4.0.3)   
## processx 3.4.4 2020-09-03 [1] RSPM (R 4.0.2)   
## ps 1.3.4 2020-08-11 [1] RSPM (R 4.0.2)   
## purrr 0.3.4 2020-04-17 [1] RSPM (R 4.0.3)   
## R6 2.4.1 2019-11-12 [1] RSPM (R 4.0.0)   
## remotes 2.2.0 2020-07-21 [1] RSPM (R 4.0.3)   
## rlang 0.4.10 2022-02-15 [1] Github (r-lib/rlang@f0c9be5)   
## rmarkdown 2.10 2022-02-15 [1] Github (rstudio/rmarkdown@02d3c25)  
## rprojroot 2.0.2 2020-11-15 [1] CRAN (R 4.0.2)   
## sessioninfo 1.1.1 2018-11-05 [1] RSPM (R 4.0.3)   
## stringi 1.5.3 2020-09-09 [1] RSPM (R 4.0.3)   
## stringr 1.4.0 2019-02-10 [1] RSPM (R 4.0.3)   
## testthat 3.0.1 2022-02-15 [1] Github (R-lib/testthat@e99155a)   
## usethis 2.1.5.9000 2022-02-15 [1] Github (r-lib/usethis@57b109a)   
## withr 2.3.0 2020-09-22 [1] RSPM (R 4.0.2)   
## xfun 0.26 2022-02-15 [1] Github (yihui/xfun@74c2a66)   
## yaml 2.2.1 2020-02-01 [1] RSPM (R 4.0.3)   
##   
## [1] /usr/local/lib/R/site-library  
## [2] /usr/local/lib/R/library

# 3 References

“About Scientific Code Review.” n.d. Accessed January 26, 2023. <https://hutchdatascience.org/code_review/>.

Allaire, JJ, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, Winston Chang, and Richard Iannone. 2021. *Rmarkdown: Dynamic Documents for r*. <https://github.com/rstudio/rmarkdown>.

Bodner, Herbert. 2018. “10 Reasons Why Code Reviews Make Better Code and Better Teams.” *Simple Programmer*. <https://simpleprogrammer.com/why-code-reviews-make-better-code-teams/>.

Gillespie, Colin, and Robin Lovelace. 2021. *Efficient R Programming*. O’Reilly Media Inc. <https://csgillespie.github.io/efficientR/>.

Parker, Hilary. 2017. “Opinionated Analysis Development.” Preprint. PeerJ Preprints. <https://doi.org/10.7287/peerj.preprints.3210v1>.

Radigan, Dan. n.d. “What Is a Code Review & How It Can Save Time.” Atlassian. Accessed January 26, 2023. <https://www.atlassian.com/agile/software-development/code-reviews>.

Riederer, Emily. 2020. “Column Names as Contracts.” Emily Riederer. September 2020. <https://emilyriederer.netlify.app/post/column-name-contracts/>.

Timbers, Tiffany, Trevor Campbell, and Melissa Lee. 2022. *Data Science*. <https://datasciencebook.ca>.

Xie, Yihui, J. J. Allaire, and Garrett Grolemund. 2018. *R Markdown: The Definitive Guide*. Boca Raton, Florida: Chapman; Hall/CRC. <https://bookdown.org/yihui/rmarkdown>.

Xie, Yihui, Christophe Dervieux, and Emily Riederer. 2020. *R Markdown Cookbook*. Boca Raton, Florida: Chapman; Hall/CRC. <https://bookdown.org/yihui/rmarkdown-cookbook>.