Tools for Reproducible Workflows in R

March, 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.

This initiative is funded by the following grant: R25GM141505 from the National Institute of General Medical Sciences (NIGMS). Except where otherwise indicated, the contents of this course are available for use under the Creative Commons Attribution 4.0 license. You are free to adapt and share the work, but you must give appropriate credit, provide a link to the license, and indicate if changes were made. Sample attribution: Tools for Reproducible Workflows in R by Fred Hutchinson Data Science Lab and University of Texas, Austin (CC-BY 4.0). You can download the illustrations by clicking [here](https://docs.google.com/presentation/d/1MNHf8JpolaEP_vQ_kB-1xRBF9wo3haCArRu117hBoHA/edit?usp=sharing).

## 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:
* R Studio tips and tricks for efficiency
* R Markdown to create reports
  + 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

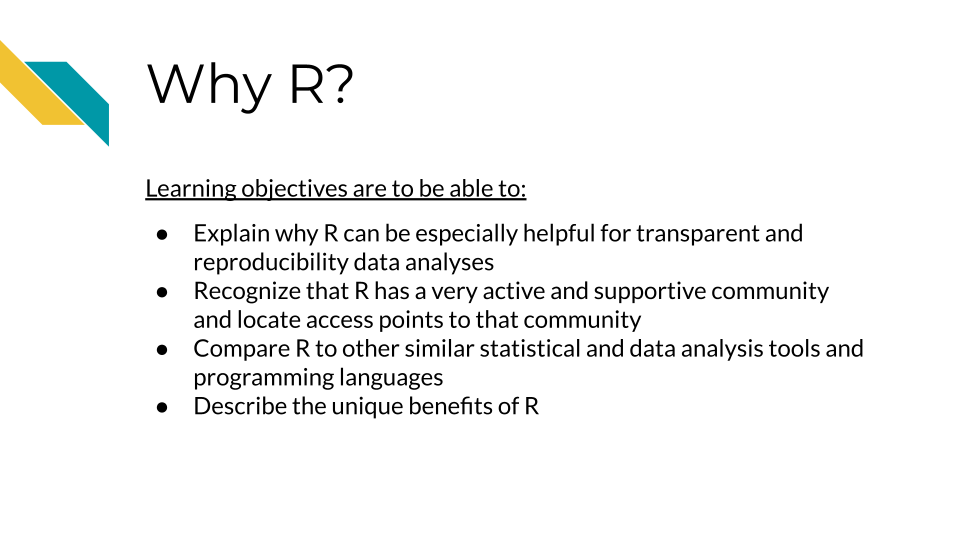
References will include Gillespie and Lovelace ([2021](#ref-gillespie_efficient_2021)), Riederer ([2020](#ref-riederer_column_2020)), Timbers, Campbell, and Lee ([2022](#ref-timbers_data_nodate)).

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 R for Reproducibility

## 2.1 Learning Objectives

Before we begin to jump into additional tools that R can help us with to be work more efficiently and in a more reproducible manner, it is helpful to first discuss why we should consider R in the first place. After completing this section you will be able to:

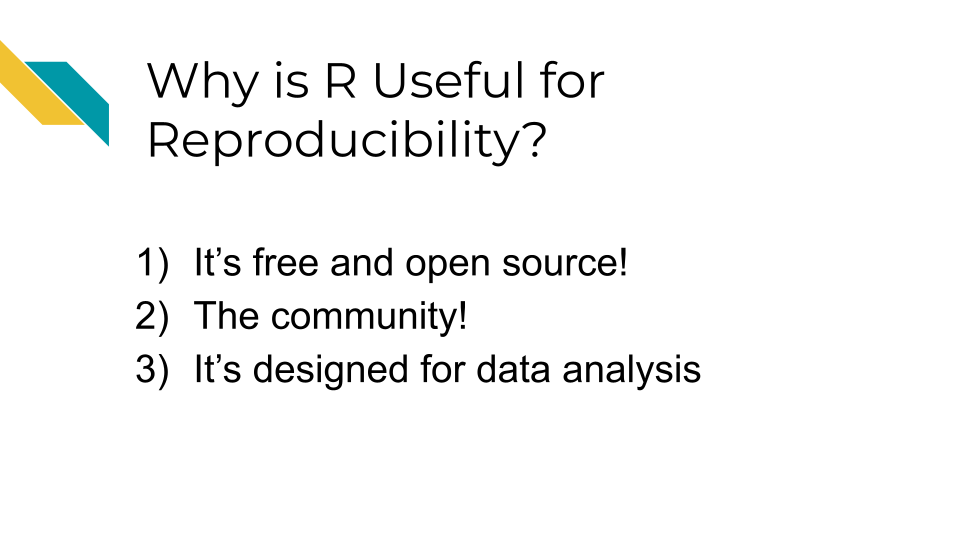


## 2.2 Why R



[R](https://www.r-project.org/) is a [programming language](https://en.wikipedia.org/wiki/Programming_language) for working with data, performing statistical analyses, and for creating plots and graphics that was developed in 1991 by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand ([“R (Programming Language)” 2023](#ref-r_2023); [“R: The R Project for Statistical Computing”](#ref-r_project)). Countless contributors have made R what it is today.

There are some especially useful aspects about R that make it a great option for creating reproducible data analyses.



## 2.3 It is free and open source

The first is that R is free and [open source](https://opensource.com/resources/what-open-source).



The term **open source** means that the code is publicly available. Thus all of the code involved in creating R is actually publicly available! This enables users to check what code is used in a particular **package** (a set of code that allows you to do various things) so that they can modify or build upon the code if they would like to.

In fact, many users create their own R **packages** to share their code with others. There are places such as the Comprehensive R Archive Network ([CRAN](https://cran.r-project.org/)) and elsewhere that allow users to publish their own packages for others to use.

* **programming language** - A specified set of notations to tell a computer what to do
* **R** - Programming language for working with data to perform statistical analyses and for creating plots and other graphics
* **open source** - Code is publicly available
* **R package** - A set of code that can be shared between users

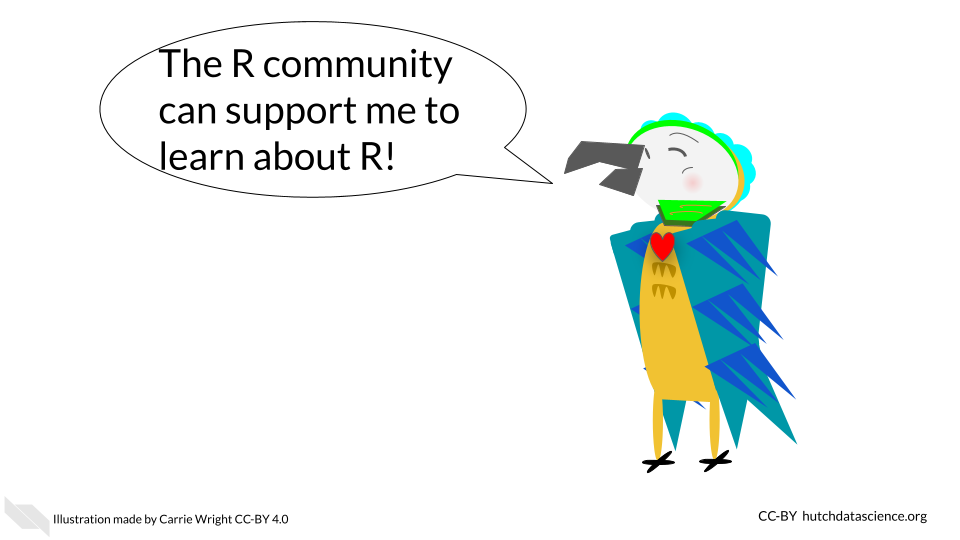
Why are these aspects good for reproducibility?

* Since R is free, it is accessible to anyone. Therefore, anyone could run your code if you shared it with them, without them needing to buy software.
* Since R is open source, if you use packages from others, people can determine what underlying code your code used (if you tell them what version you used - more on that later!)

## 2.4 The community

R has a very rich and active community!

This makes it easier to reach out to others for help, find support, find tutorials, and more.



There are several R community groups that are especially helpful:

* [R Ladies](https://www.rladies.org/) - a support group that is not just for ladies, but is open to anyone who wants to improve their R skills! There are local chapters in many large cities that often have in-person meetings.
* There are lots of useful resources, such as the [R for Data Science book](https://r4ds.had.co.nz/) (written by two developers at Posit (formally called RStudio) which develops lots of core R packages), resources and online courses from the [Johns Hopkins Data Science Lab](https://jhudatascience.org/courses.html) including [Open Case Studies](https://www.opencasestudies.org/), resources and workshops from [Data Carpentry](https://datacarpentry.org/), [Dataquest](https://www.dataquest.io/v2/), [DataTrail](https://datatrail-jhu.github.io/DataTrail/) and more!

See this [link](https://jhudatascience.org/intro_to_r/resources.html) for more R resources.

Why is this rich community good for reproducibility?

* Overall your code has a better chance of being more accessible than if it were written in a language that is not open source or that has limited support.
* You can also find support to make sure your code does what you want it to, as well as support to make your code as reproducible as possible.

## 2.5 Designed for data

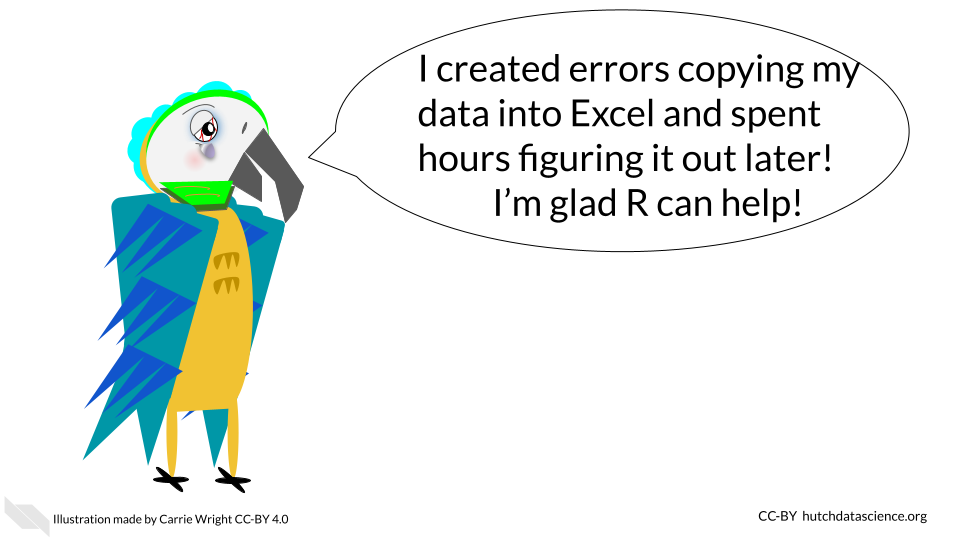
R is a statistical programming language, meaning it was designed to help you analyze data. It is the main focus of the language. This is one of the major advantages of using R over other programming languages that have more general purposes.

Because of this many people have designed useful packages that are especially relevant to:

1. Dealing with messy data in a systematic and reproducible way to get it into a state that is useful for data analysis
2. Producing statistical analysis of data
3. Creating effective plots of data

Although other options like [SPSS](https://www.ibm.com/products/spss-statistics) and [SAS](https://www.sas.com/) (which are not free!) can also be helpful for statistical analysis, R is especially powerful at getting messy data ready to analyze and for creating useful plots to represent patterns in data. Conveniently, R can do all of these steps in a data project and does not require users to switch between different programs to perform these tasks. R also helps create reports that can demonstrate to collaborators and others exactly how analysis was performed, aiding in the transparency of how the data was used from start to finish.

R can also import data from many different sources that other statistical software can’t handle (including scraping data from websites or [PDFs](https://www.adobe.com/acrobat/about-adobe-pdf.html)). This allows users much more flexibility to use data as close to the source as possible. This can enable users to stop copy and pasting data and reduce the risk of human error. If you are interested, see [Open Case Studies](https://www.opencasestudies.org/) for more guidance on importing many different kinds of data.



Why are these design features especially helpful for creating reproducible analyses?

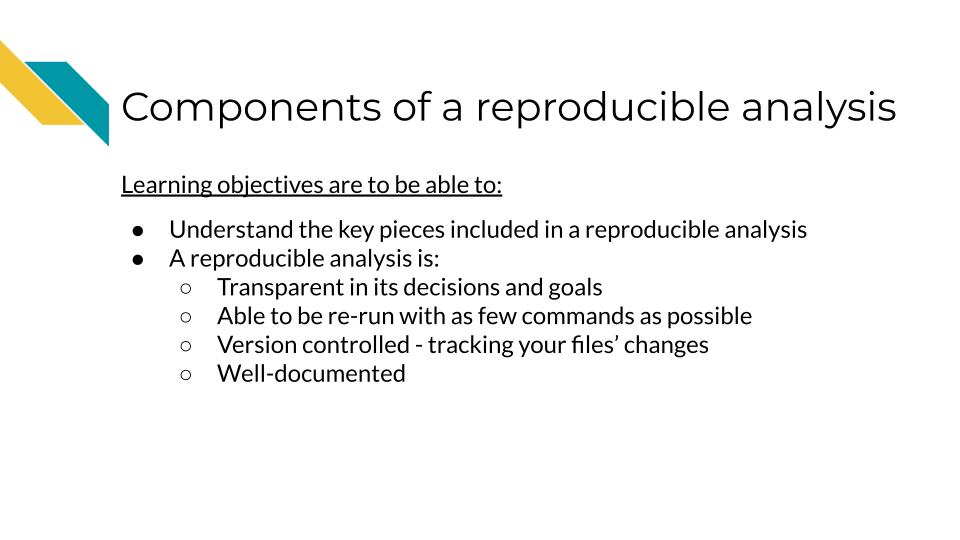
1. It enables users to work with messy data and get it ready for analysis, as opposed to requiring users to use other programs. The tidyverse a suite of very helpful packages has many data wrangling packages that are especially intuitive for others to read and understand your code.
2. Users can create effective plots using the same program as for data prep and analysis. The ggplot2 package is famous for making really effective and customizable plots.
3. It helps create reports that can show the entire data analysis process from importing the data to making plots. R Markdown reports are very helpful for this.
4. It is easier to import data closer to the original source, rather than converting files or copy and pasting data, which can result in accidental modifications of the data.

## 2.6 Conclusion

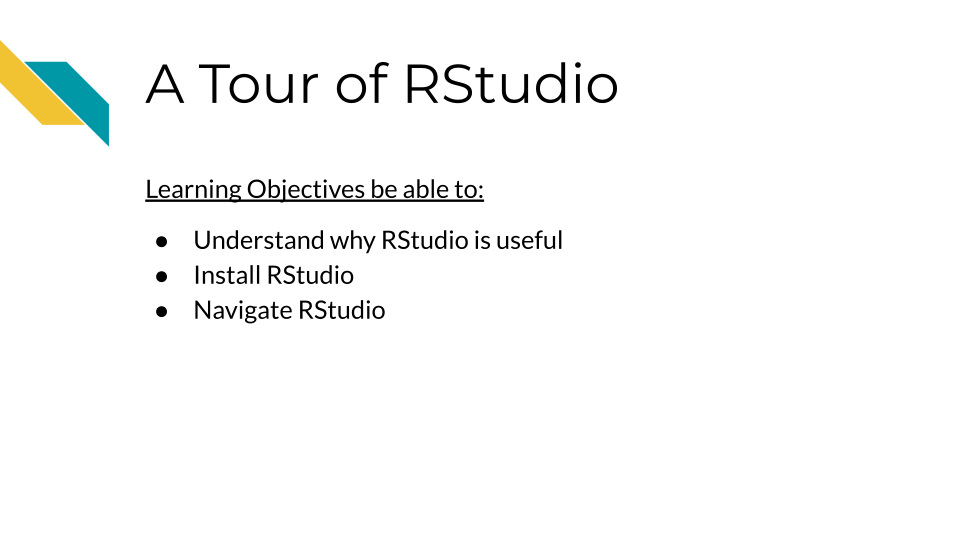
In summary, R can be especially useful if you want to make your data analyses more transparent and reproducible for the following reasons:

1. It is free and open source, meaning that code that you might incorporate in your analyses is accessible to anyone. Secondly, others can use your code without needing to buy software.
2. There is a rich R community that can help you make the most out of your code and learn how to write your code in a more reproducible manner.
3. R is particularly powerful for preparing data for analysis and for creating visual representations of data. Beyond being free, these unique benefits make R a particularly good statistical tool.
4. R is especially designed to analyze data and for the entirety of the process, which makes it great for creating transparent information about how you actually worked with data from start to finish.

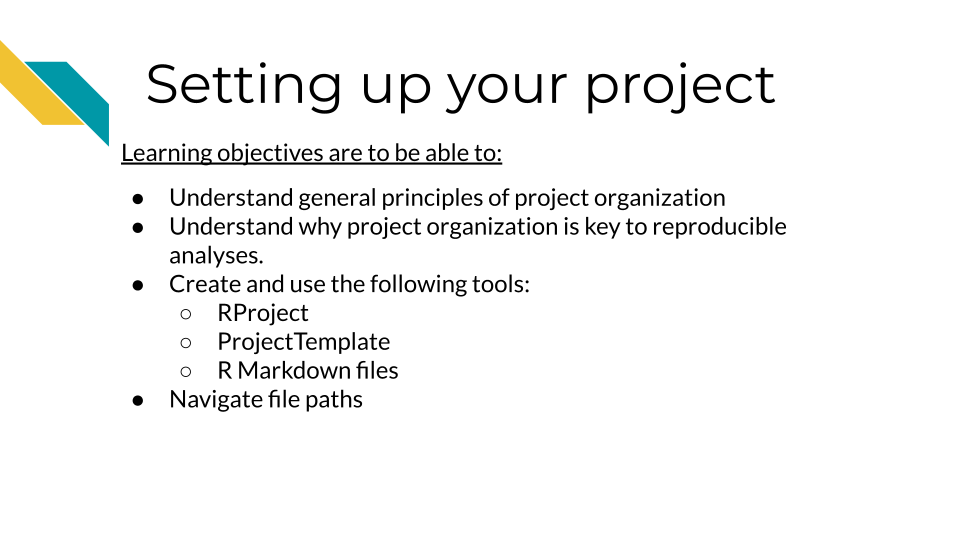
# 3 Components of a reproducible analysis



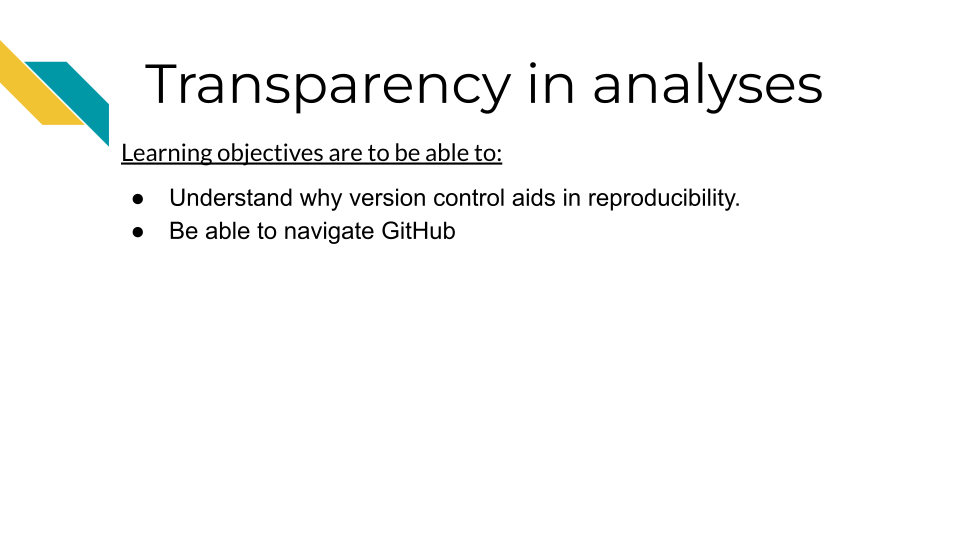
# 4 A Tour of RStudio



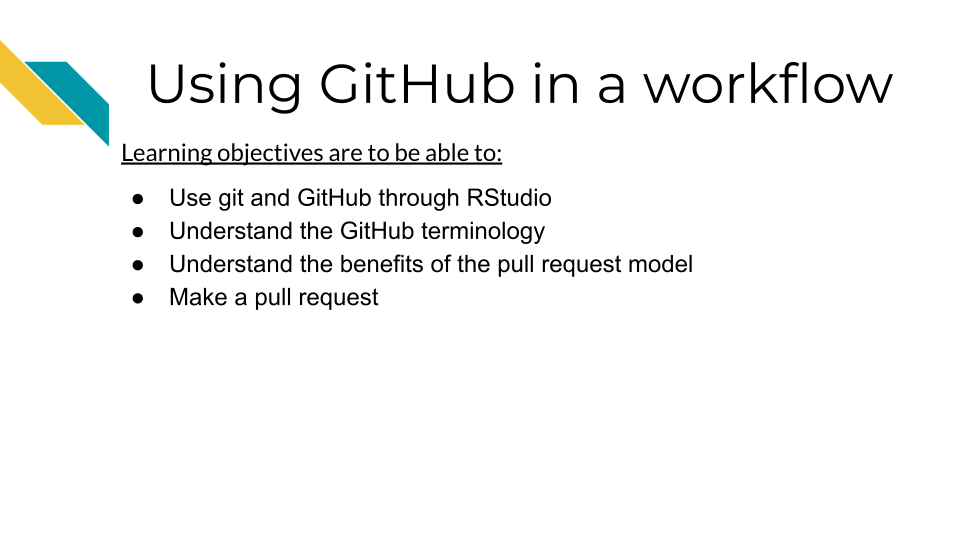
# 5 Setting up your project



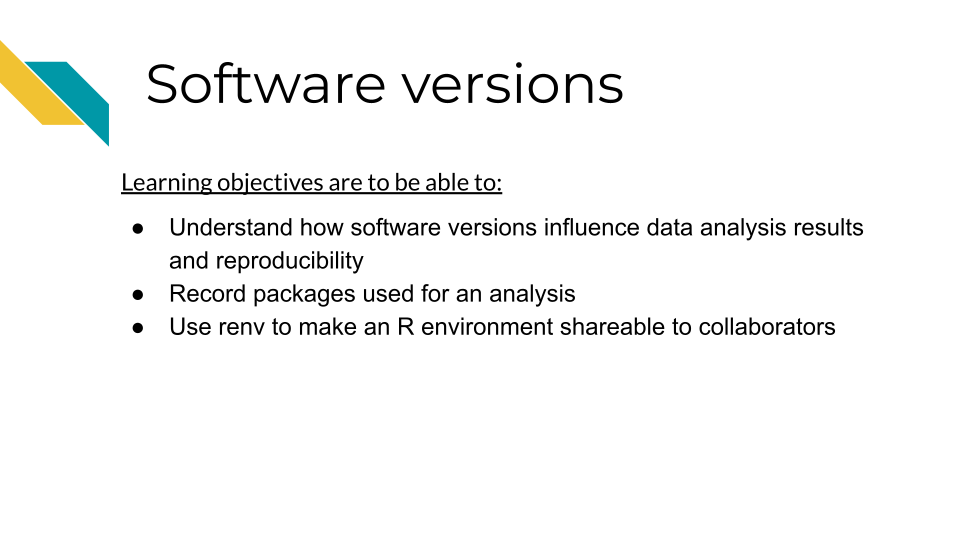
# 6 Transparency in Analyses



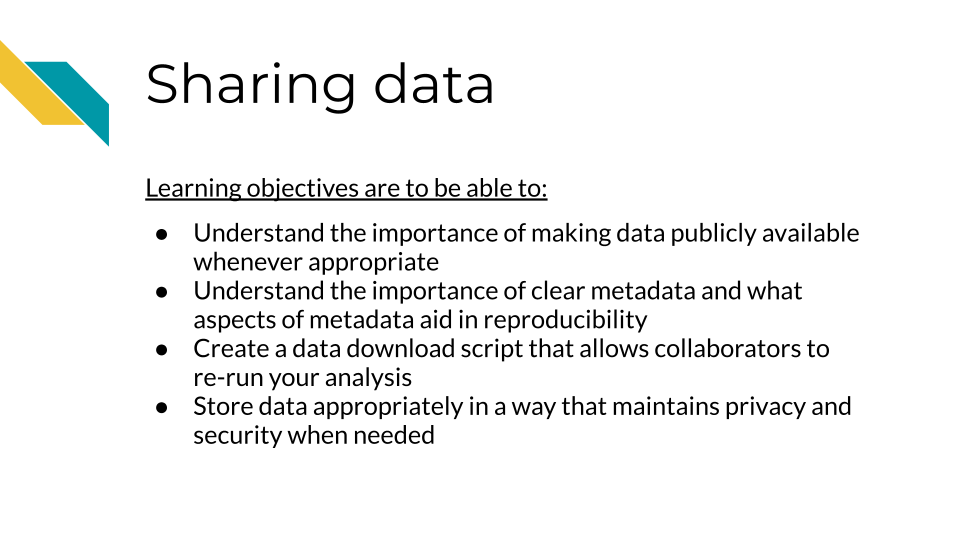
# 7 Using GitHub in a workflow



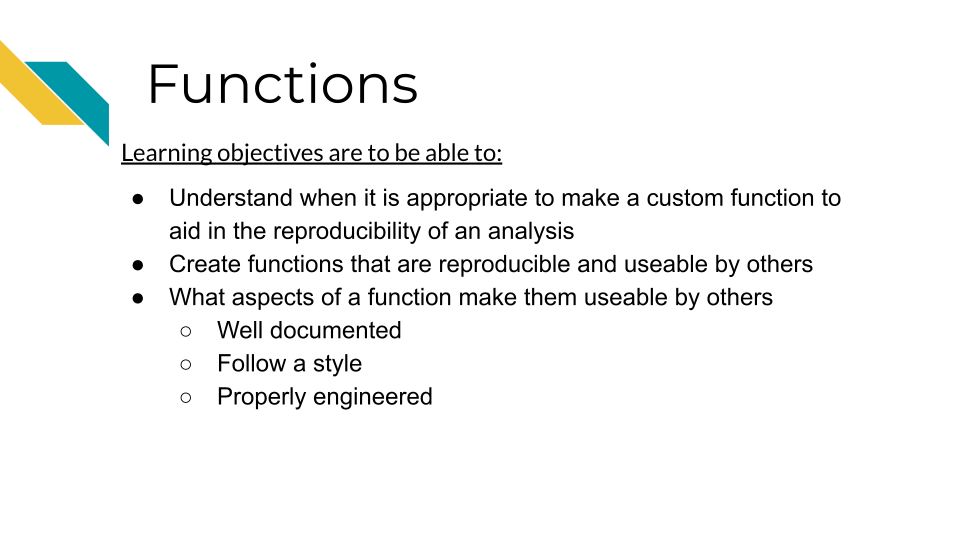
# 8 Software versions



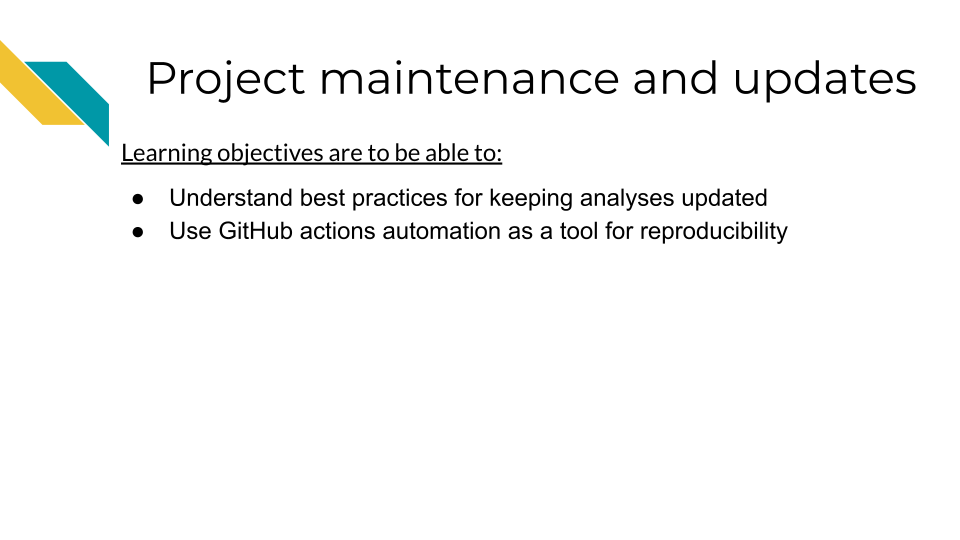
# 9 Sharing Data



# 10 Functions



# 11 Project maintenance and updates



# 12 Collaborations through GitHub



# 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 20.04.5 LTS   
## system x86\_64, linux-gnu   
## ui X11   
## language (EN)   
## collate en\_US.UTF-8   
## ctype en\_US.UTF-8   
## tz Etc/UTC   
## date 2023-03-15   
##   
## ─ Packages ───────────────────────────────────────────────────────────────────  
## package \* version date lib source   
## assertthat 0.2.1 2019-03-21 [1] RSPM (R 4.0.5)   
## bookdown 0.24 2023-03-10 [1] Github (rstudio/bookdown@88bc4ea)   
## cachem 1.0.7 2023-02-24 [1] CRAN (R 4.0.2)   
## callr 3.5.0 2020-10-08 [1] RSPM (R 4.0.2)   
## cli 3.6.0 2023-01-09 [1] CRAN (R 4.0.2)   
## 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 0.20 2023-01-17 [1] CRAN (R 4.0.2)   
## fastmap 1.1.1 2023-02-24 [1] CRAN (R 4.0.2)   
## fs 1.5.0 2020-07-31 [1] RSPM (R 4.0.3)   
## glue 1.4.2 2020-08-27 [1] RSPM (R 4.0.5)   
## htmltools 0.5.4 2022-12-07 [1] CRAN (R 4.0.2)   
## knitr 1.33 2023-03-10 [1] Github (yihui/knitr@a1052d1)   
## magrittr 2.0.3 2022-03-30 [1] CRAN (R 4.0.2)   
## memoise 2.0.1 2021-11-26 [1] CRAN (R 4.0.2)   
## 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 1.4.0 2020-10-07 [1] RSPM (R 4.0.2)   
## 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 1.0.6 2022-09-24 [1] CRAN (R 4.0.2)   
## rmarkdown 2.10 2023-03-10 [1] Github (rstudio/rmarkdown@02d3c25)  
## rprojroot 2.0.3 2022-04-02 [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 2023-03-10 [1] Github (R-lib/testthat@e99155a)   
## usethis 1.6.3 2020-09-17 [1] RSPM (R 4.0.2)   
## withr 2.3.0 2020-09-22 [1] RSPM (R 4.0.2)   
## xfun 0.26 2023-03-10 [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

# 13 References

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

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>.

“R (Programming Language).” 2023. *Wikipedia*. <https://en.wikipedia.org/w/index.php?title=R_(programming_language)&oldid=1144363470>.

“R: The R Project for Statistical Computing.” <https://www.r-project.org/>.

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>.