

R 4 Epidemiology

2024-06-06

Table of contents

Welcome	4
Acknowledgements	4
Introduction	5
Goals	5
Text conventions used in this book	6
Other reading	6
Contributing	7
Typos	7
Issues	14
License Information	14
About the Authors	15
Brad Cannell	15
Melvin Livingston	16
I Getting Started	17
1 Installing R and RStudio	18
1.1 Download and install on a Mac	18
1.2 Download and install on a PC	26
2 What is R?	34
2.1 What is data?	34
2.2 What is R?	39
2.2.1 Transferring data	40
2.2.2 Managing data	41
2.2.3 Analyzing data	42
2.2.4 Presenting data	43
II Coding Tools and Best Practices	45
3 Quarto Files	46

III	Collaboration	47
4	Using git and GitHub	48
IV	References	49
	References	50
	Appendices	51
A	Glossary	51

Welcome

Welcome to R for Epidemiology!

This electronic textbook was originally created to accompany the Introduction to R Programming for Epidemiologic Research course at the [University of Texas Health Science Center School of Public Health](#). However, we hope it will be useful to anyone who is interested in R, epidemiology, or human health and well-being.

Acknowledgements

This book is currently a work in progress (and probably always will be); however, there are already many people who have played an important role (some unknowingly) in helping develop it thus far. First, we'd like to offer our gratitude to all past, current, and future members of the R Core Team for maintaining this *amazing, free* software. We'd also like to express our gratitude to everyone at [Posit](#). You are also developing and *giving away* some amazing software. In particular, we'd like to acknowledge [Garrett Grolemund](#) and [Hadley Wickham](#). Both have had a huge impact on how we use and teach R. We'd also like to thank our students for all the feedback they've given us while taking our courses. In particular, we want to thank [Jared Wiegand](#) and Yiqun Wang for their many edits and suggestions.

This electronic textbook was created and published using [R](#), [RStudio](#), the [Quarto](#), and [GitHub](#).

Introduction

Goals

We're going to start the introduction by writing down some basic goals that underlie the construction and content of this book. We're writing this for you, the reader, but also to hold ourselves accountable as we write. So, feel free to read if you are interested or skip ahead if you aren't.

The goals of this book are:

1. **To teach you how to use R and RStudio as tools for applied epidemiology.¹** Our goal is not to teach you to be a computer scientist or an advanced R programmer. Therefore, some readers who are experienced programmers may catch some technical inaccuracies regarding what we consider to be the fine points of what R is doing “under the hood.”
2. **To make this writing as accessible and practically useful as possible without stripping out all of the complexity that makes doing epidemiology in real life a challenge.** In other words, We're going to try to give you all the tools you need to *do* epidemiology in “real world” conditions (as opposed to ideal conditions) without providing a whole bunch of extraneous (often theoretical) stuff that detracts from *doing*. Having said that, we will strive to add links to the other (often theoretical) stuff for readers who are interested.
3. **To teach you to accomplish common *tasks*,** rather than teach you to use functions or families of functions. In many R courses and texts, there is a focus on learning all the things a function, or set of related functions, can do. It's then up to you, the reader, to sift through all of these capabilities and decided which, if any, of the things that *can* be done will accomplish the tasks that you are *actually trying* to accomplish. Instead, we will strive to start with the end in mind. What is the task we are actually trying to accomplish? What are some functions/methods we could use to accomplish that task? What are the strengths and limitations of each?

¹In this case, “tools for applied epidemiology” means (1) understanding epidemiologic concepts; and (2) completing and interpreting epidemiologic analyses.

4. **To start each concept by showing you the end result** and then deconstruct how we arrived at that result, where possible. We find that it is easier for many people to understand new concepts when learning them as a component of a final product.
5. **To learn concepts with data** instead of (or alongside) mathematical formulas and text descriptions, where possible. We find that it is easier for many people to understand new concepts by seeing them in action.

Text conventions used in this book

- We will hyperlink many keywords or phrases to their [glossary](#) entry.
- Additionally, we may use **bold** face for a word or phrase that we want to call attention to, but it is not necessarily a keyword or phrase that we want to define in the glossary.
- **Highlighted inline code** is used to emphasize small sections of R code and program elements such as variable or function names.

Other reading

If you are interested in R4Epi, you may also be interested in:

- [Hands-on Programming with R](#) by Garrett Grolemund. This book is designed to provide a friendly introduction to the R language.
- [R for Data Science](#) by Hadley Wickham, Mine Çetinkaya-Rundel, and Garrett Grolemund. This book is designed to teach readers how to do data science with R.
- [Statistical Inference via Data Science: A ModernDive into R and the Tidyverse](#). This book is designed to be a gentle introduction to the practice of analyzing data and answering questions using data the way data scientists, statisticians, data journalists, and other researchers would.
- [Reproducible Research with R and RStudio](#) by Christopher Gandrud. This book gives you tools for data gathering, analysis, and presentation of results so that you can create dynamic and highly reproducible research.
- [Advanced R](#) by Hadley Wickham. This book is designed primarily for R users who want to improve their programming skills and understanding of the language.

Contributing

Over the years, we have learned so much from our students and colleagues, and we anticipate that there is much more we can learn from you – our readers. Therefore, we welcome and appreciate all constructive contributions to R4Epi!

Typos

The easiest way for you to contribute is to help us clean up the little typos and grammatical errors that inevitably sneak into the text.

If you spot a typo, you can offer a correction directly in GitHub. You will first need to create a free GitHub account: [sign-up at github.com](#). Later in the book, we will cover using GitHub in greater depth (See Chapter 4). Here, we're just going to walk you through how to fix a typo without much explanation of how GitHub works.

Let's say you spot a typo while reading along.

If you spot a typo, you can offer a correction directly in the easiest way to offer a correction is directly in GitHub. You will first need to create a free GitHub account: [sign-up at github.com](#). Later in the book, we will cover using GitHub in greater depth (See Chapter 4). Here, we're just going to walk you through how to fix a typo without much explanation of how GitHub works.

Uh, oh! The word “typo” should only have one “o”!

Let's say you spot a typoo while reading along.

Next, click the edit button in the toolbar as shown in the screenshot below.

 Edit this page

Report an issue

The first time you click the icon, you will be taken to the R4Epi repository on GitHub and asked to fork it. For our purposes, you can think of a GitHub repository as being similar to a shared folder on Dropbox or Google Drive.



You need to fork this repository to propose changes.

Sorry, you're not able to edit this repository directly. You need to fork it and propose your changes from there instead.

[Fork this repository](#)
Learn more about forks

Fork the Repository

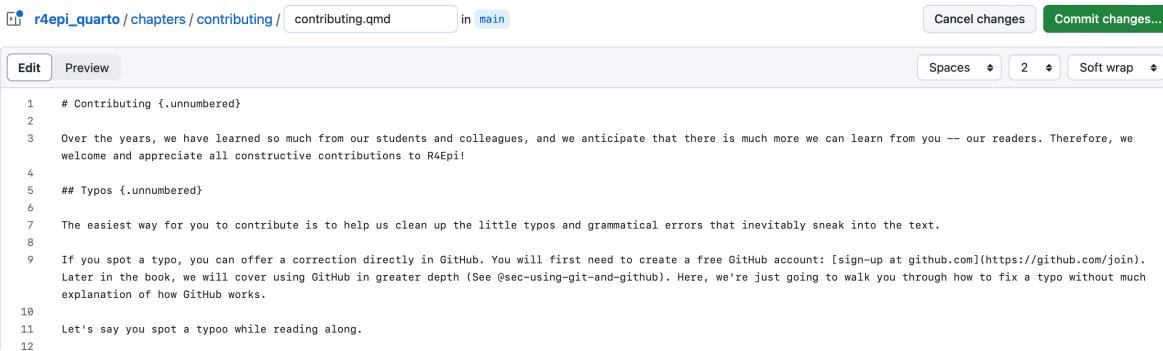
“Forking the repository” basically just means “make a copy of the repository” on your GitHub account. In other words, copy all of the files that make up the R4Epi textbook to your GitHub account. Then, you can fix the typos you found in your *copy* of the files that make up the book instead of directly editing the *actual* files that make up the book. This is a safeguard to prevent people from accidentally making changes that shouldn’t be made.

Note

Forking the R4Epi repository does not cost any money or add any files to your computer.

After you fork the repository, you will see a text editor on your screen.

You’re making changes in a project you don’t have write access to. Submitting a change will write it to a new branch in your fork arthur-epi/r4epi_quarto, so you can send a pull request.



The screenshot shows a GitHub text editor interface. At the top, there's a navigation bar with a 'r4epi_quarto / chapters / contributing / contributing.qmd' path, a 'main' tab, and 'Cancel changes' and 'Commit changes...' buttons. Below the navigation is a toolbar with 'Edit' (selected), 'Preview', 'Spaces', '2', and 'Soft wrap'. The main area contains the following Quarto code:

```
1 # Contributing {.unnumbered}
2
3 Over the years, we have learned so much from our students and colleagues, and we anticipate that there is much more we can learn from you -- our readers. Therefore, we
4 welcome and appreciate all constructive contributions to R4Epi!
5
6 ## Typo {.unnumbered}
7
8 The easiest way for you to contribute is to help us clean up the little typos and grammatical errors that inevitably sneak into the text.
9
10 If you spot a typo, you can offer a correction directly in GitHub. You will first need to create a free GitHub account: [sign-up at github.com](https://github.com/join). Later in the book, we will cover using GitHub in greater depth (See @sec-using-git-and-github). Here, we're just going to walk you through how to fix a typo without much
11 explanation of how GitHub works.
12 Let's say you spot a typo while reading along.
```

The text editor will display the contents of the file used to make the chapter you were looking at when you clicked the **edit** button. In this example, it was a file named `contributing.qmd`. The `.qmd` file extension means that the file is a Quarto file. We will learn more about Quarto files in Chapter 3, but for now just know that Quarto files can be used to create web pages and other documents that contain a mix of R code, text, and images.

Next, scroll down through the text until you find the typo and fix it. In this case, line 11 contains the word “typoo”. To fix it, you just need to click in the editor window and begin typing. In this case, you would click next to the word “typoo” and delete the second “o”.

You're making changes in a project you don't have write access to. Submitting a change will write it to a new branch in your fork arthur-epi/r4epi_quarto, so you can send a pull request.

r4epi_quarto / chapters / contributing / contributing.qmd in **main**

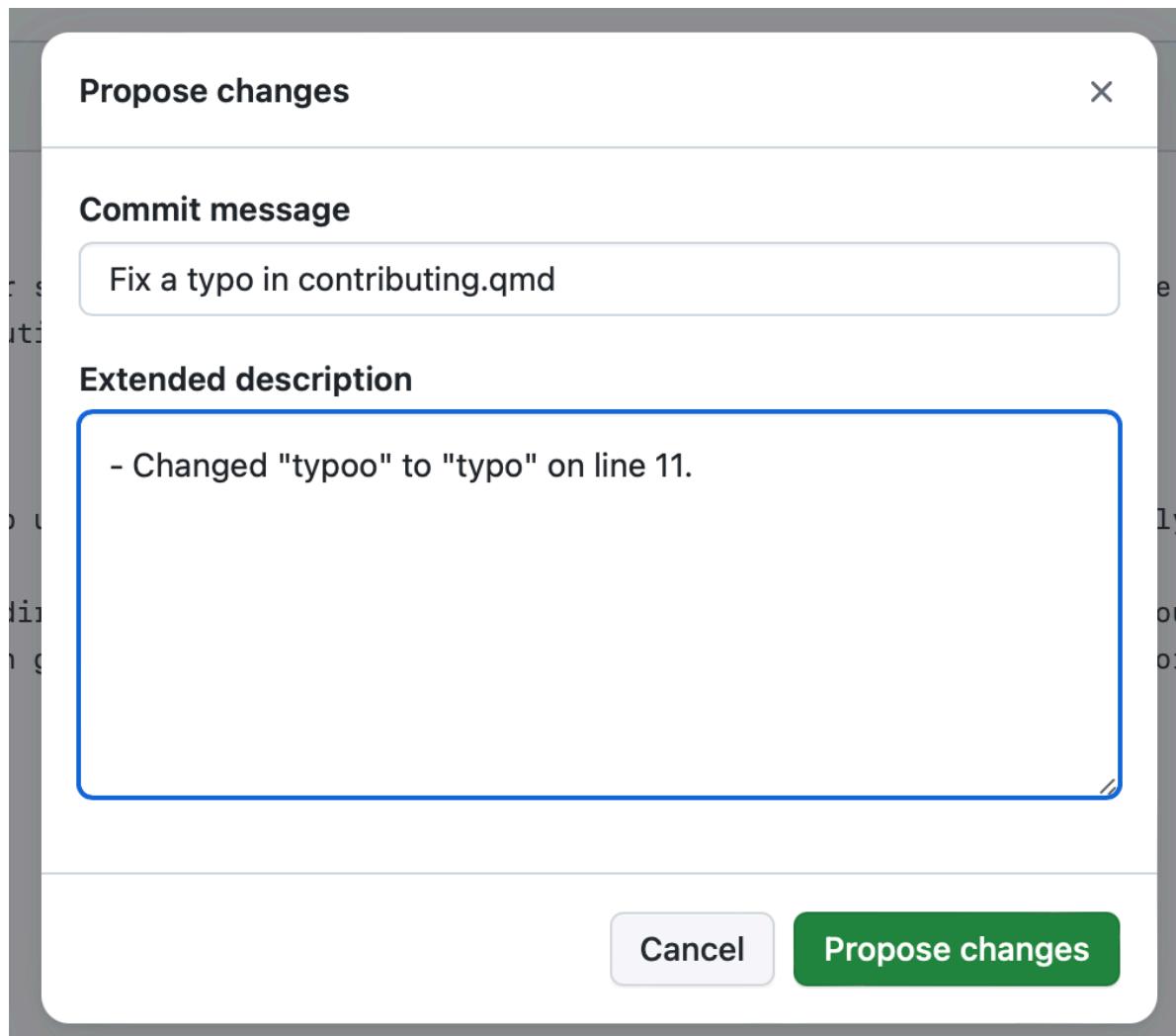
Commit changes...

Edit Preview Spaces 2 Soft wrap

```
1 # Contributing {.unnumbered}
2
3 Over the years, we have learned so much from our students and colleagues, and we anticipate that there is much more we can learn from you -- our readers. Therefore, we welcome and appreciate all constructive contributions to R4Epi!
4
5 ## Typos {.unnumbered}
6
7 The easiest way for you to contribute is to help us clean up the little typos and grammatical errors that inevitably sneak into the text.
8
9 If you spot a typo, you can off... You will first need to create a free GitHub account: [sign-up at github.com](https://github.com/join). Later in the book, we will cover using GitHub in greater depth (See @sec-using-git-and-github). Here, we're just going to walk you through how to fix a typo without much explanation of how GitHub works.
10
11 Let's say you spot a typo while reading along.
12
```

The screenshot shows a GitHub code editor interface. At the top, a message says "You're making changes in a project you don't have write access to. Submitting a change will write it to a new branch in your fork arthur-epi/r4epi_quarto, so you can send a pull request." Below this, the file path is "r4epi_quarto / chapters / contributing / contributing.qmd" and the branch is "main". There are two buttons: "Cancel changes" and a green "Commit changes..." button, which is highlighted with a red box. Below the buttons are "Edit" and "Preview" tabs, and "Spaces", "2", and "Soft wrap" settings. The main area contains a block of R code. Lines 7 and 9 contain annotations: "Deleted the extra 'o'" is written over line 7, and a red circle with a black arrow points to the word "typo" in line 9.

Now, the only thing left to do is propose your typo fix to the authors. To do so, click the green **Commit changes...** button on the right side of the screen above the text editor (surrounded with a red box in the screenshot above). When you click it, a new **Propose changes** box will appear on your screen. Type a brief (i.e., 72 characters or less) summary of the change you made in the **Commit message** box. There is also an **Extended description** box where you can add a more detailed description of what you did. In the screenshot below, shows an example commit message and extended description that will make it easy for the author to quickly figure out exactly what changes are being proposed.



Next, click the **Propose changes** button. That will take you to another screen where you will be able to create a pull request. This screen is kind of busy, but try not to let it overwhelm you.

Comparing changes

Choose two branches to see what's changed or to start a new pull request. If you need to, you can also [compare across forks](#) or [learn more about diff comparisons](#).

The screenshot shows a GitHub pull request interface. At the top, there are dropdown menus for 'base repository' (brad-cannell/r4epi_quarto), 'base' (main), 'head repository' (arthur-epi/r4epi_quarto), and 'compare' (patch-1). A green checkmark indicates 'Able to merge'. Below this, a message says 'These branches can be automatically merged.' A red box highlights the 'Create pull request' button. Another red box highlights the commit message 'Fix a typo in contributing.qmd'. A third red box highlights the unified diff view, specifically focusing on line 11 where 'typoo' was replaced by 'typo'.

For now, we will focus on the three different sections of the screen that are highlighted with a red outline. We will start at the bottom and work our way up. The red box that is closest to the bottom of the screenshot shows us that the change that made was on line 11. The word “typoo” (highlighted in red) was replaced with the word “typo” (highlighted in green). The red box in the middle of the screenshot shows us the brief description that was written for our proposed change – “Fix a typo in contributing.qmd”. Finally, the red box closest to the top of the screenshot is surrounding the Create pull request button. You will click it to move on with your pull request.

Open a pull request

Create a new pull request by comparing changes across two branches. If you need to, you can also [compare across forks](#). Learn more about diff comparisons [here](#).

The screenshot shows the GitHub interface for creating a pull request. At the top, there are dropdown menus for 'base repository' (brad-cannell/r4epi_quarto), 'base' (main), 'head repository' (arthur-epi/r4epi_quarto), and 'compare' (patch-1). A green checkmark indicates 'Able to merge'. Below this, there's a section to 'Add a title' with the placeholder 'Fix a typo in contributing.qmd'. To the right, 'Helpful resources' link to 'GitHub Community Guidelines'. Under 'Add a description', there's a rich text editor toolbar with options like H, B, I, etc. The main text area contains the commit message: '- Changed "typoo" to "typo" on line 11.' Below the editor, it says 'Markdown is supported' and 'Paste, drop, or click to add files'. At the bottom, there's a checkbox for 'Allow edits by maintainers' and a green 'Create pull request' button.

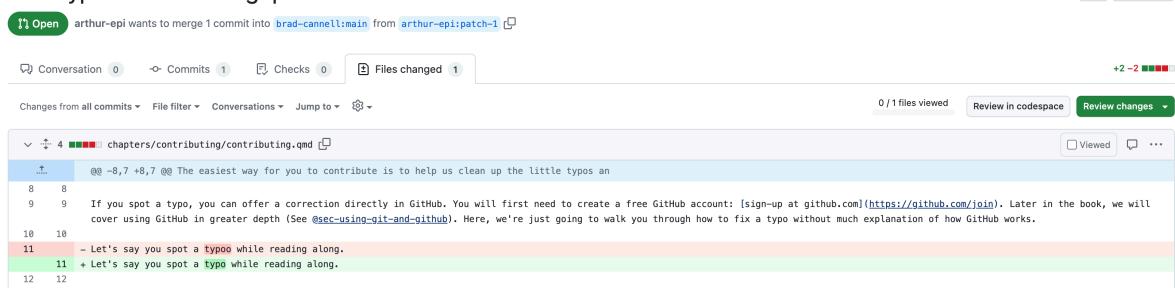
After doing so, you will get one final chance to amend the description of your proposed changes. If you are happy with the commit message and description, then click the **Create pull request** button one more time. At this point, your job is done! It is now up to the authors to review the changes you've proposed and “pull” them into the file in their repository.

In case you are curious, here is what the process looks like on the authors' end. First, when we open the R4Epi repository page on GitHub, we will see that there is a new pull request.

The screenshot shows the GitHub repository page for 'brad-cannell / r4epi_quarto'. The navigation bar at the top has tabs for 'Code', 'Issues 1', 'Pull requests 1' (which is highlighted with a red box), 'Actions', and 'Projects 1'. Below the navigation bar, there's a brief summary of the repository: 'Last updated 1 day ago · 13 commits · 1 branch · 1 pull request · 1 issue · 1 project · 1 dependency declaration'.

When we open the pull request, we can see the proposed changes to the file.

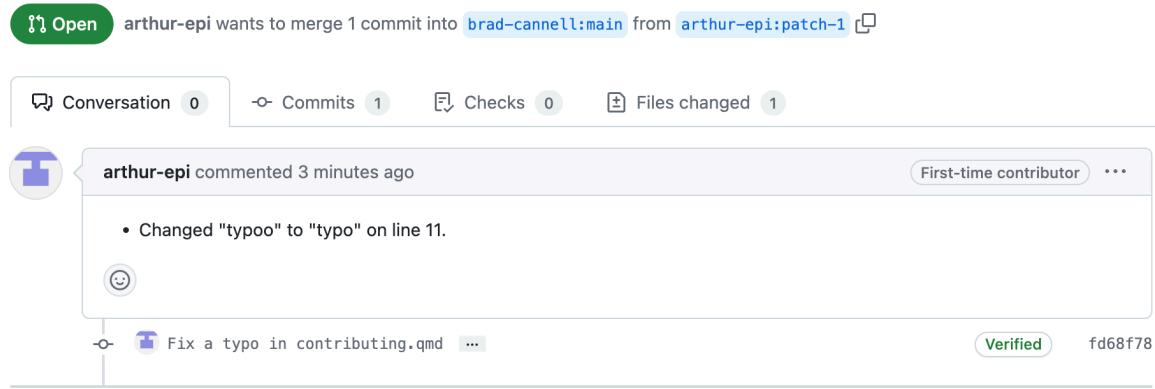
Fix a typo in contributing.qmd #7



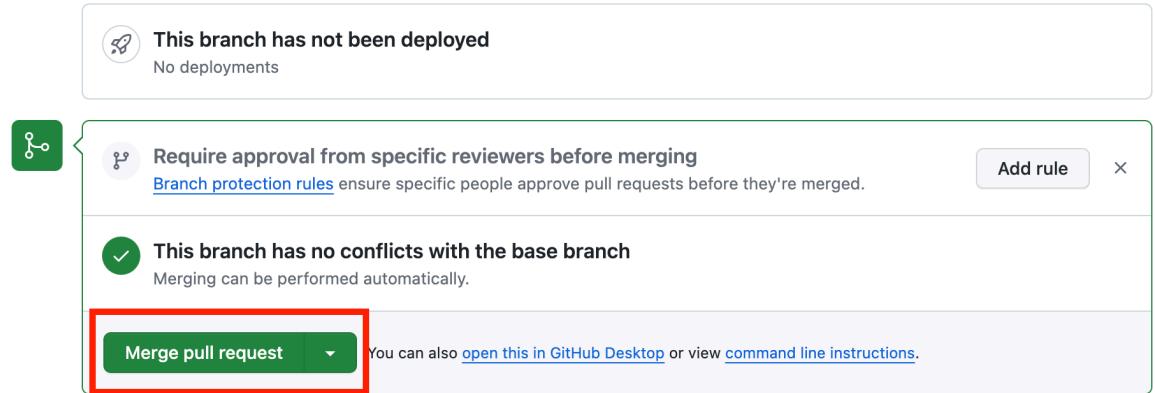
The screenshot shows a GitHub pull request interface. At the top, it says "arthur-epi wants to merge 1 commit into brad-cannell:main from arthur-epi:patch-1". Below this are tabs for Conversation (0), Commits (1), Checks (0), and Files changed (1). The "Files changed" tab is selected, showing a diff for "chapters/contributing/contributing.qmd". The diff highlights a change on line 11: "- Let's say you spot a typoo while reading along." is replaced by "+ Let's say you spot a typo while reading along.". The commit message "Fix a typo in contributing.qmd" is visible at the bottom of the commit list.

Then, all we have to do is click the `Merge pull request` button and the fixed file is “pulled in” to replace the file with the typo.

Fix a typo in contributing.qmd #7

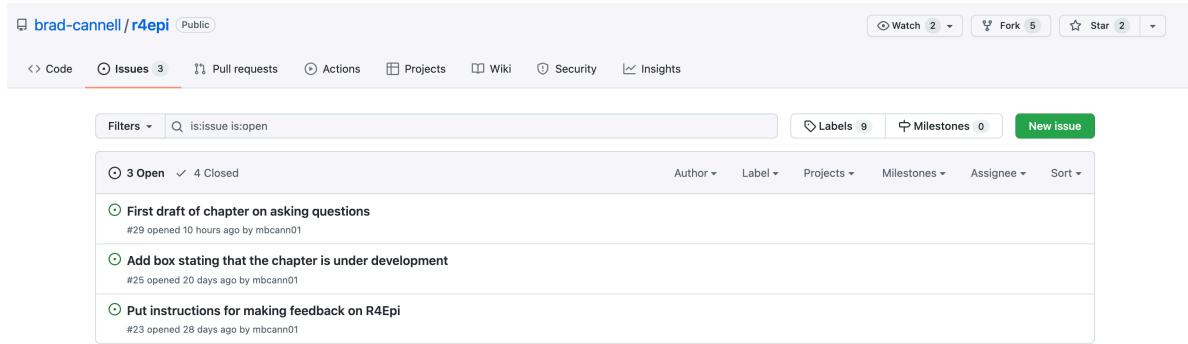


Add more commits by pushing to the [patch-1](#) branch on [arthur-epi/r4epi_quarto](#).



Issues

There may be times when you see a problem that you don't know how to fix, but you still want to make the authors aware of. In that case, you can create an [issue](#) in the R4Epi repository. To do so, navigate to the issue tracker using this link: <https://github.com/brad-cannell/r4epi/issues>.



The screenshot shows the GitHub interface for the 'r4epi' repository. The 'Issues' tab is active, displaying three open issues. The first issue is titled 'First draft of chapter on asking questions' and was opened 10 hours ago by 'mbcann01'. The second issue is 'Add box stating that the chapter is under development' and was opened 20 days ago by 'mbcann01'. The third issue is 'Put instructions for making feedback on R4Epi' and was opened 28 days ago by 'mbcann01'. At the top right, there are buttons for 'Watch 2', 'Fork 5', and 'Star 2'. Below the tabs, there are filters for 'Labels 9' and 'Milestones 0', and a prominent green 'New issue' button.

Once there, you can check to see if someone has already raised the issue you are concerned about. If not, you can click the green “New issue” button to raise it yourself.

Please note that R4Epi uses a [Contributor Code of Conduct](#). By contributing to this book, you agree to abide by its terms.

License Information

This book was created by Brad Cannell and is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

About the Authors

Brad Cannell

Michael (Brad) Cannell, PhD, MPH

Associate Professor

Elder Mistreatment Lead, UTHealth Institute of Aging

Director, Research Informatics Core, Cizik Nursing Research Institute

UTHealth Houston

McGovern Medical School

Joan and Stanford Alexander Division of Geriatric & Palliative Medicine

www.bradcannell.com

Dr. Cannell received his PhD in Epidemiology, and Graduate Certificate in Gerontology, in 2013 from the University of Florida. He received his MPH with a concentration in Epidemiology from the University of Louisville in 2009, and his BA in Political Science and Marketing from the University of North Texas in 2005. During his doctoral studies, he was a Graduate Research Assistant for the Florida Office on Disability and Health, an affiliated scholar with the Claude D. Pepper Older Americans Independence Center, and a student-inducted member of the Delta Omega Honorary Society in Public Health. In 2016, Dr. Cannell received a Graduate Certificate in Predictive Analytics from the University of Maryland University College, and a Certificate in Big Data and Social Analytics from the Massachusetts Institute of Technology.

He previously held professional staff positions in the Louisville Metro Health Department and the Northern Kentucky Independent District Health Department. He spent three years as a project epidemiologist for the Florida Office on Disability and Health at the University of Florida. He also served as an Environmental Science Officer in the United States Army Reserves from 2009 to 2013.

Dr. Cannell's research is broadly focused on healthy aging and health-related quality of life. Specifically, he has published research focusing on preservation of physical and cognitive function, living and aging with disability, and understanding and preventing elder mistreatment. Additionally, he has a strong background and training in epidemiologic methods and predictive analytics. He has been principal or co-investigator on multiple trials and observational studies in community and healthcare settings. He is currently the principal investigator on multiple data-driven federally funded projects that utilize technological solutions to public health issues in novel ways.

Contact

Connect with Dr. Cannell and follow his work.



Melvin Livingston

Melvin (Doug) Livingston, PhD

Research Associate Professor

Department of Behavioral, Social, and Health Education Sciences

Emory University Woodruff Health Sciences Center

Rollins School of Public Health

[Dr. Livingston's Faculty Profile](#)

Dr. Livingston is a methodologist with expertise in the application of quasi-experimental design principals to the evaluation for both community interventions and state policies. He has particular expertise in time series modeling, mixed effects modeling, econometric methods, and power analysis. As part of his work involving community trials, he has been the statistician on the long term follow-up study of a school based cluster randomized trial in low-income communities with a focus on explaining the etiology of risky alcohol, drug, and sexual behaviors. Additionally, he was the statistician for a longitudinal study examining the etiology of alcohol use among racially diverse and economically disadvantaged urban youth, and co-investigator for a NIAAA- and NIDA-funded trial to prevent alcohol use and alcohol-related problems among youth living in high-risk, low-income communities within the Cherokee Nation. Prevention work at the community level led him to an interest in the impact of state and federal socioeconomic policies on health outcomes. He is a Co-Investigator of a 50-state, 30-year study of effects of state-level economic and education policies on a diverse set of public health outcomes, explicitly examining differential effects across disadvantaged subgroups of the population.

His current research interests center around the application of quasi-experimental design and econometric methods to the evaluation of the health effects of state and federal policy.

Contact

Connect with Dr. Livingston and follow his work.



Part I

Getting Started

1 Installing R and RStudio

Before we can do any programming with [R](#), we first have to download it to our computer. Fortunately, R is free, easy to install, and runs on all major operating systems (i.e., Mac and Windows). However, R is even easier to use as when we combine it with another program called [RStudio](#). Fortunately, RStudio is also free and will also run on all major operating systems.

At this point, you may be wondering what R is, what RStudio is, and how they are related. We will answer those questions in the near future. However, in the interest of keeping things brief and simple, We're not going to get into them right now. Instead, all you have to worry about is getting the R programming language and the RStudio IDE (IDE is short for integrated development environment) downloaded and installed on your computer. The steps involved are slightly different depending on whether you are using a Mac or a PC (i.e., Windows). Therefore, please feel free to use the table of contents on the right-hand side of the screen to navigate directly to the instructions that you need for your computer.

 Note

In this chapter, we cover how to download and install R and RStudio on both Mac and PC. However, the screenshots in all following chapters will be from a Mac. The good news is that RStudio operates almost identically on Mac and PC.

Step 1: Regardless of which operating system you are using, please make sure your computer is on, properly functioning, connected to the internet, and has enough space on your hard drive to save R and RStudio.

1.1 Download and install on a Mac

Step 2: Navigate to the Comprehensive R Archive Network (CRAN), which is located at <https://cran.r-project.org/>.

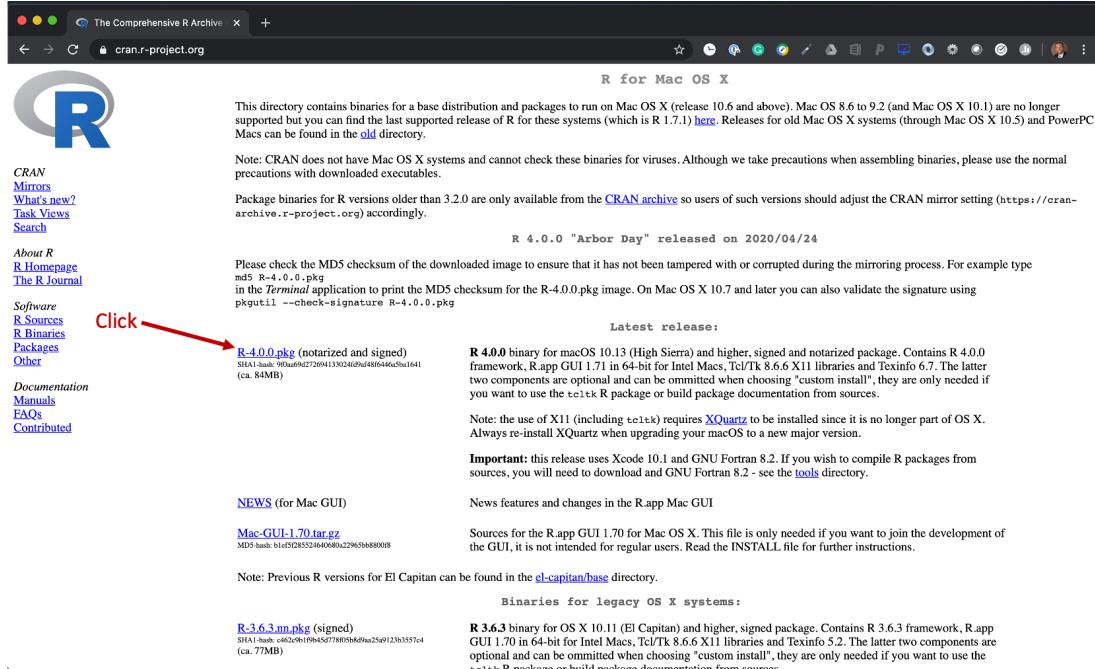
The screenshot shows the main page of the CRAN website. On the left, there's a sidebar with links like CRAN Mirrors, What's new?, Task Views, Search, About R, R Homepage, The R Journal, Software, R Sources, R Binaries, Packages, Other, Documentation, Manuals, FAQs, and Contributed. The main content area has a large title 'The Comprehensive R Archive Network'. Below it, a section titled 'Download and Install R' lists 'Precompiled binary distributions of the base system and contributed packages, Windows and Mac users most likely want one of these versions of R:'. It includes links for 'Download R for Linux', 'Download R for (Mac) OS X', and 'Download R for Windows'. A note below says 'R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.' Another section 'Source Code for all Platforms' provides links for the latest release (R-4.0.0.tar.gz), sources of R alpha and beta releases, daily snapshots, source code of older versions, and contributed extension packages. A 'Questions About R' section at the bottom has a link to 'answers to frequently asked questions'.

Step 3: Click on Download R for macOS.

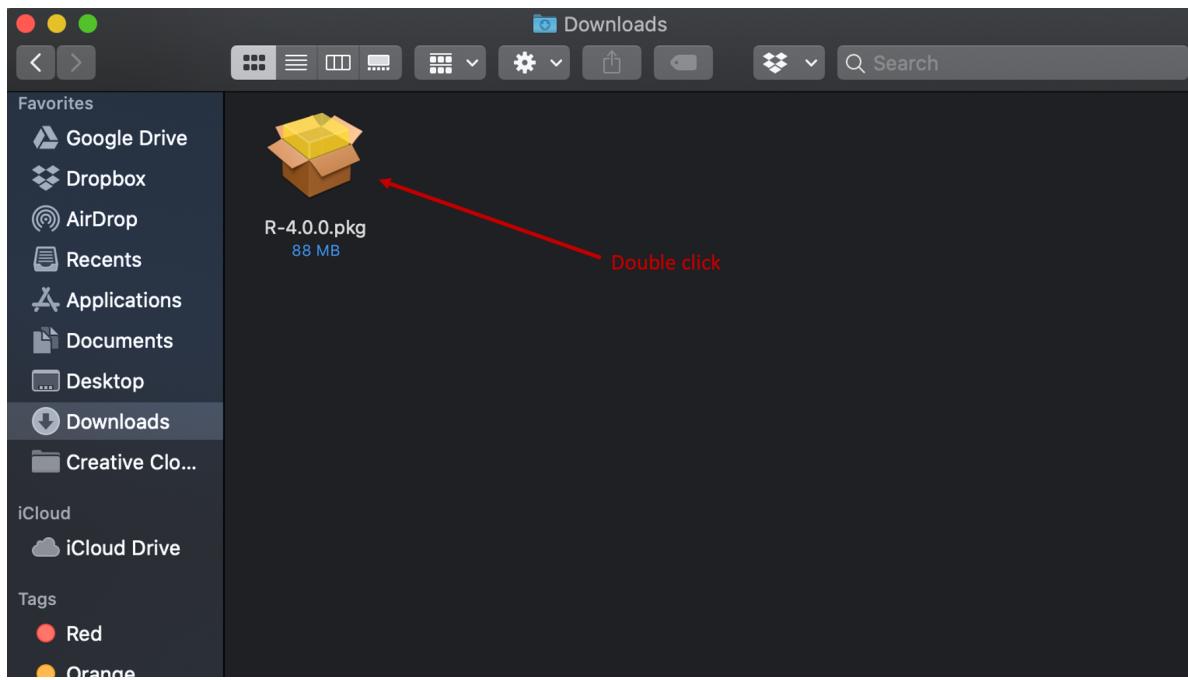
This screenshot is identical to the one above, showing the CRAN homepage. However, a red arrow points to the 'Download R for (Mac) OS X' link under the 'Download and Install R' section. This indicates the specific action required in Step 3.

Step 4: Click on the link for the latest version of R. As you are reading this, the newest version may be different than the version you see in this picture, but the location of the newest version

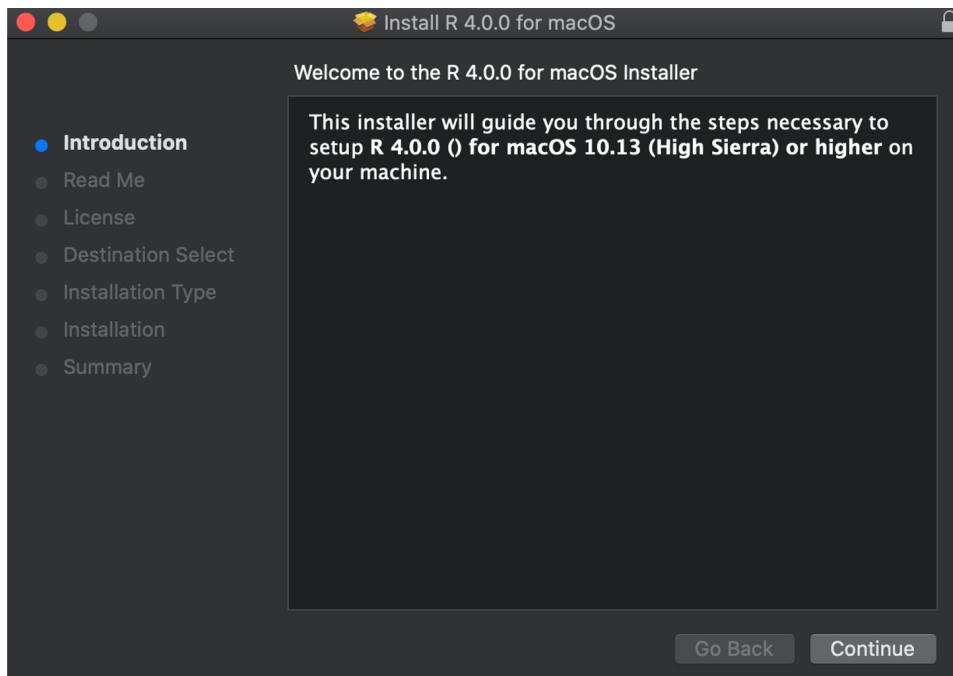
should be roughly in the same place – the middle of the screen under “Latest release:”. After clicking the link, R should start to download to your computer automatically.



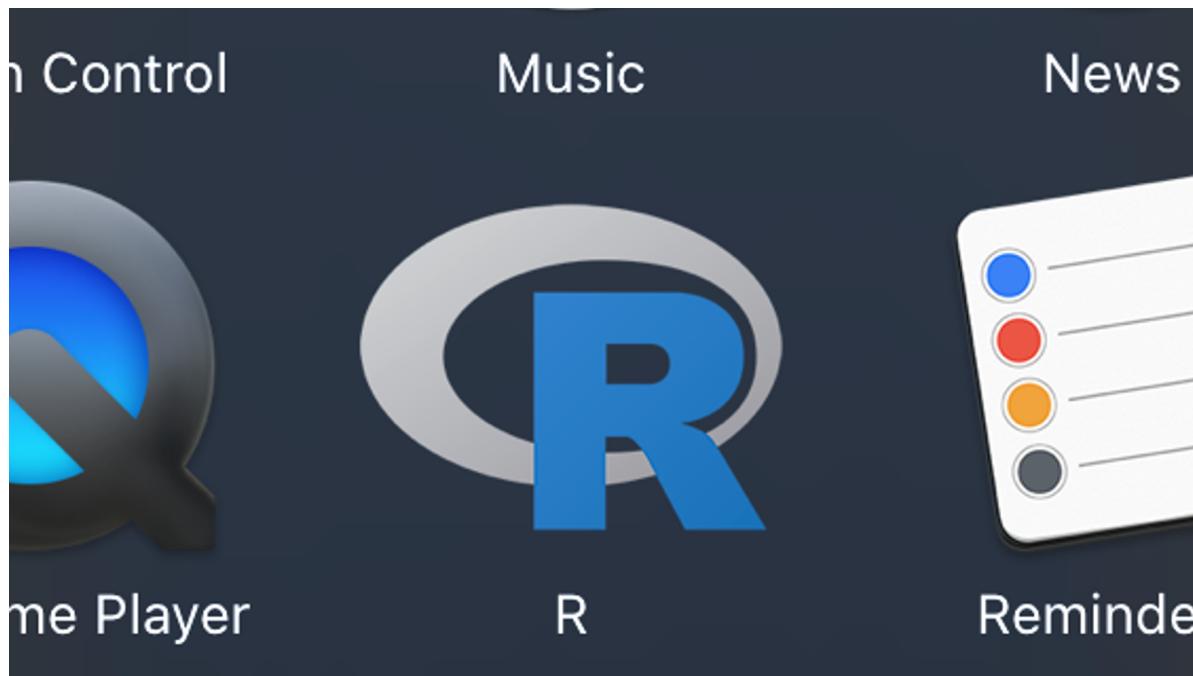
Step 5: Locate the package file you just downloaded and double click it. Unless you've changed your download settings, this file will probably be in your “downloads” folder. That is the default location for most web browsers. After you locate the file, just double click it.



Step 6: A dialogue box will open and ask you to make some decisions about how and where you want to install R on your computer. We typically just click “continue” at every step without changing any of the default options.



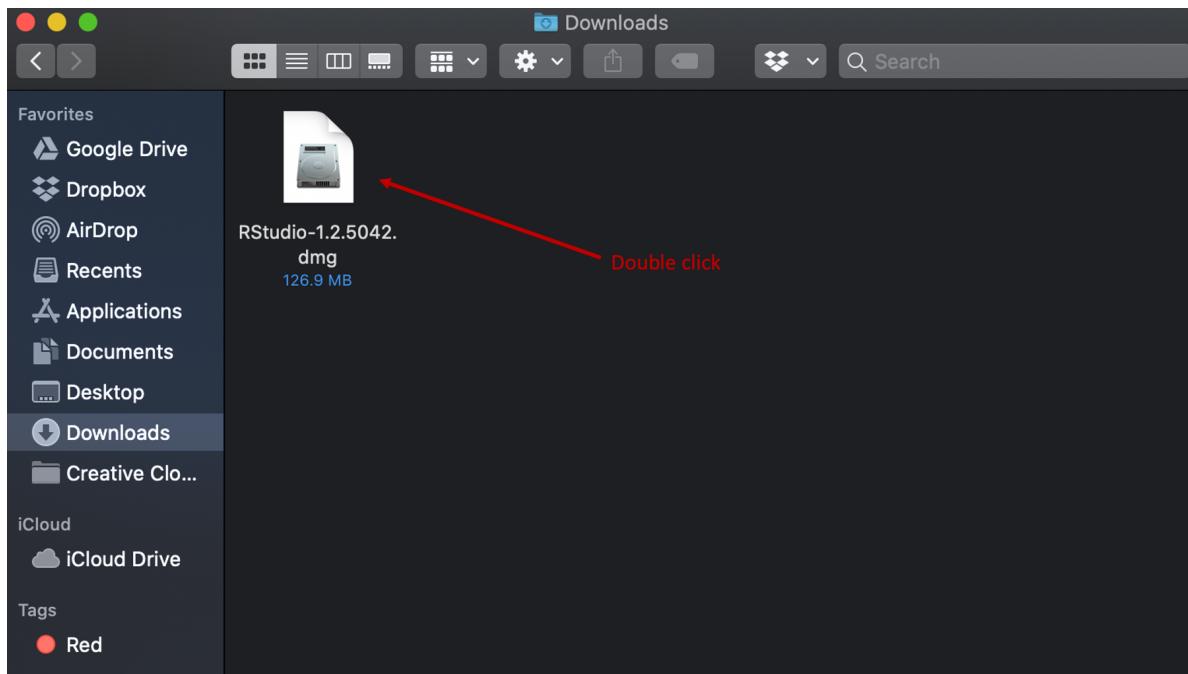
If R installed properly, you should now see it in your applications folder.



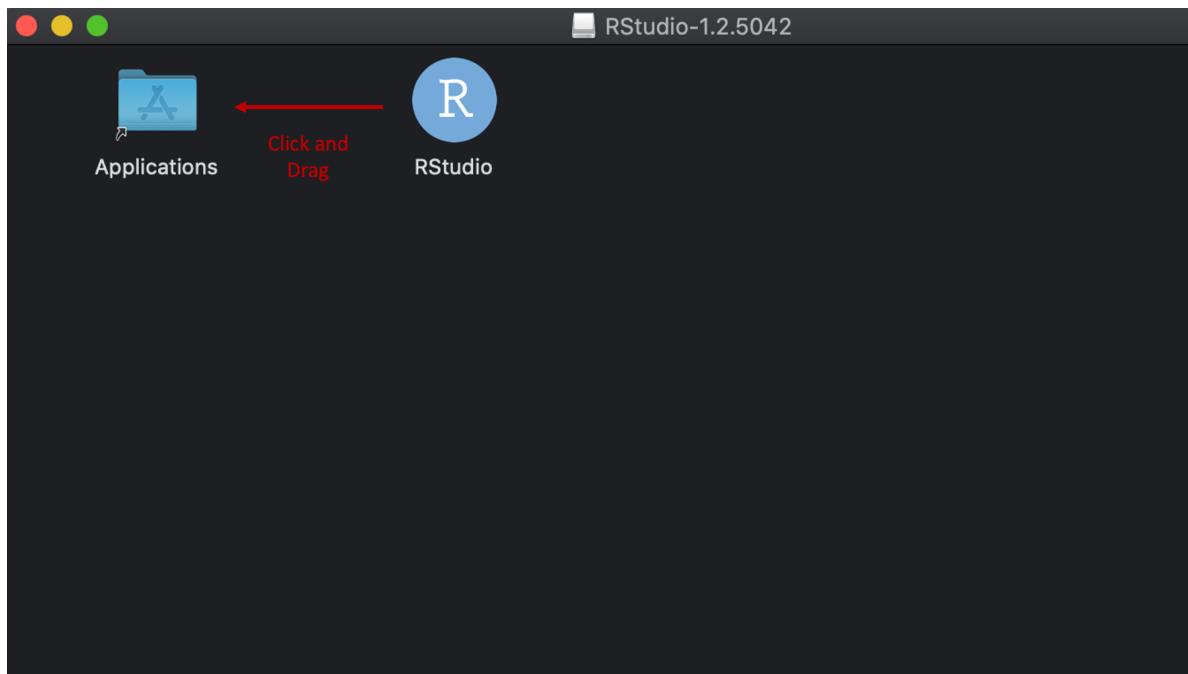
Step 7: Now, we need to install the RStudio IDE. To do this, navigate to the RStudio desktop download website, which is located at <https://posit.co/download/rstudio-desktop/>. On that page, click the button to download the latest version of RStudio for your computer. Note that the website may look different than what you see in the screenshot below because websites change over time.

OS	Download	Size	SHA-256
Windows 10/11	RSTUDIO-2024.04.1-748.EXE	263.07 MB	44C8797C
macOS 12+	RSTUDIO-2024.04.1-748.DMG	566.51 MB	A5EDA699
Ubuntu 20/Debian 11	RSTUDIO-2024.04.1-748-AMD64.DEB	194.71 MB	505311AE
Ubuntu 22/Debian 12	RSTUDIO-2024.04.1-748-AMD64.DEB	197.00 MB	88D485CD
OpenSUSE 15	RSTUDIO-2024.04.1-748-X86_64.RPM	197.21 MB	D25315A4
Fedora 34/Red Hat 8	RSTUDIO-2024.04.1-748-X86_64.RPM	219.99 MB	A97A28A7
Fedora 36/Red Hat 9	RSTUDIO-2024.04.1-748-X86_64.RPM	211.10 MB	69580324

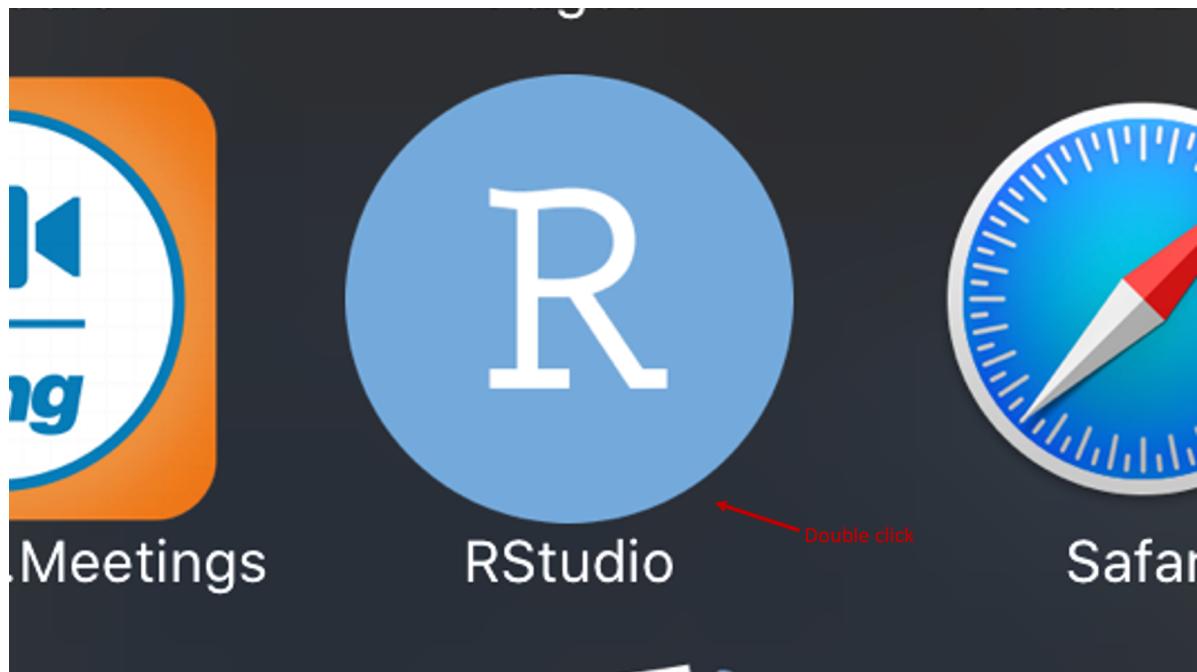
Step 8: Again, locate the DMG file you just downloaded and double click it. Unless you've changed your download settings, this file should be in the same location as the R package file you already downloaded.



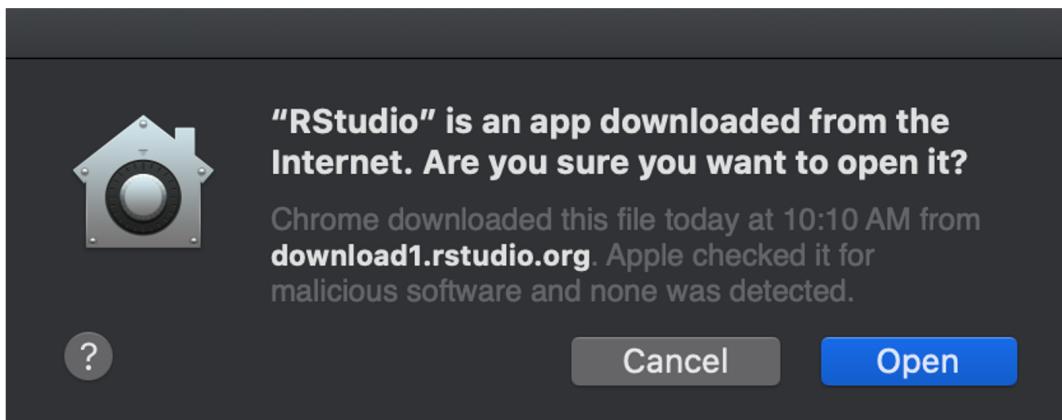
Step 9: A new finder window should automatically pop up that looks like the one you see below. Click on the RStudio icon and drag it into the Applications folder.



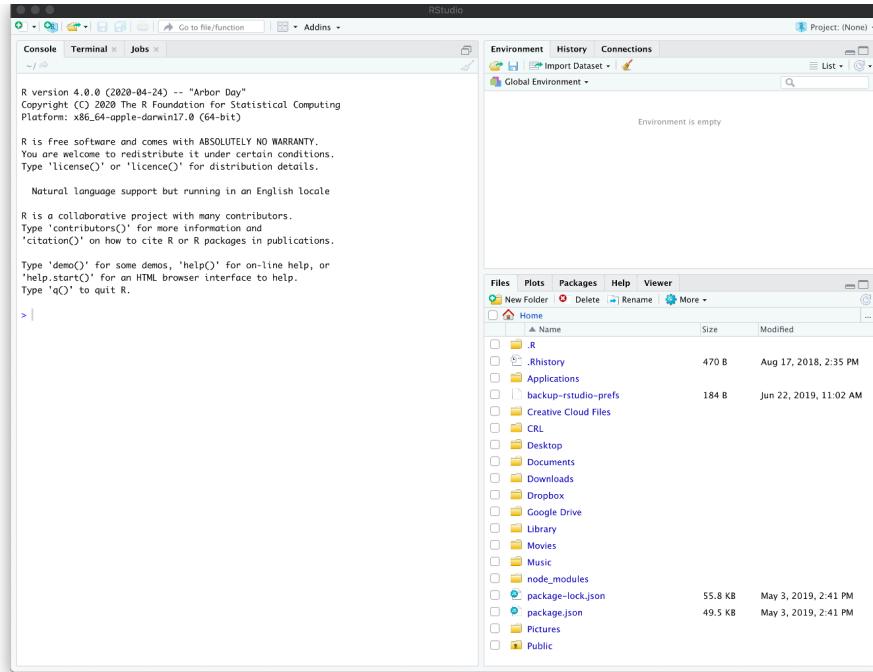
You should now see RStudio in your Applications folder. Double click the icon to open RStudio.



If this warning pops up, just click Open.



The RStudio IDE should open and look something like the window you see here. If so, you are good to go!



1.2 Download and install on a PC

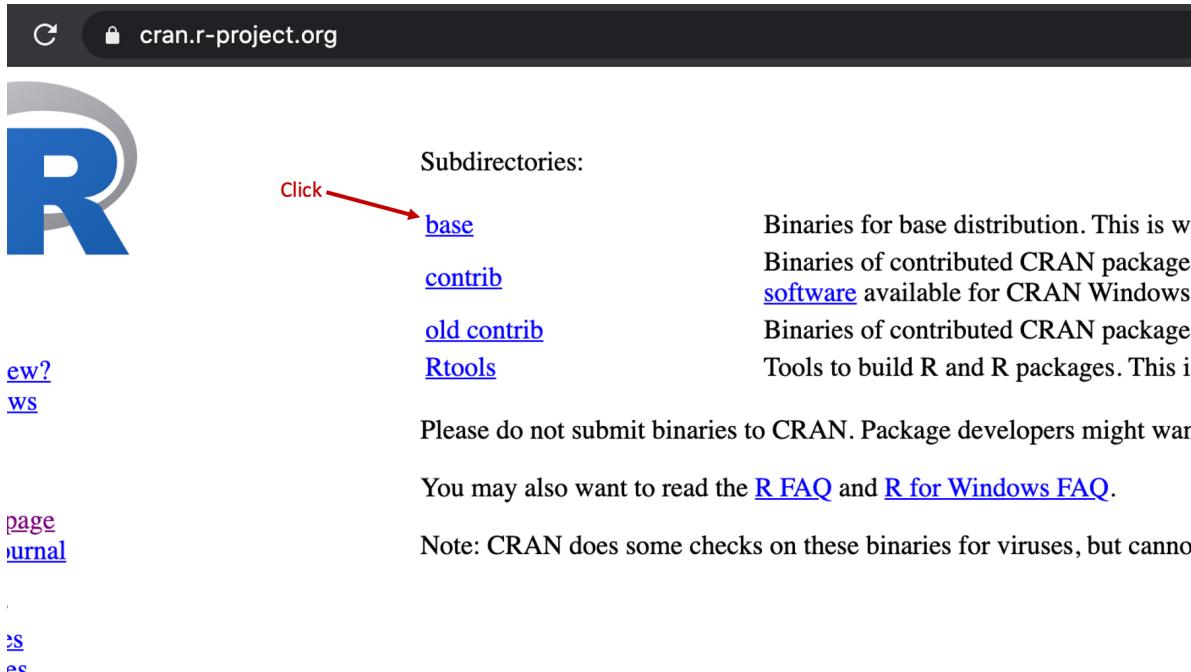
Step 2: Navigate to the Comprehensive R Archive Network (CRAN), which is located at <https://cran.r-project.org/>.

The screenshot shows the main page of the CRAN website. On the left, there's a sidebar with links like CRAN, Mirrors, What's new?, Task Views, Search, About R, R Homepage, The R Journal, Software, R Sources, R Binaries, Packages, Other, Documentation, Manuals, FAQs, and Contributed. The main content area has a large title 'The Comprehensive R Archive Network'. Below it, a section titled 'Download and Install R' lists precompiled binary distributions for Windows and Mac users. It also mentions that R is part of many Linux distributions and provides source code for all platforms. A 'Questions About R' section at the bottom includes a link to frequently asked questions.

Step 3: Click on Download R for Windows.

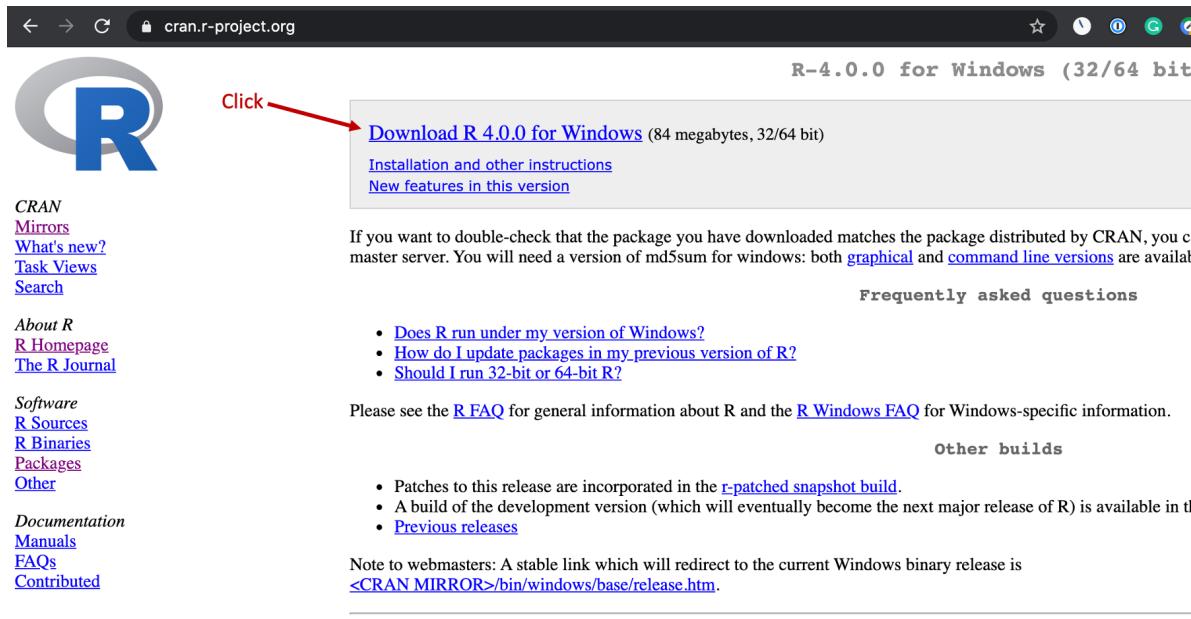
This screenshot is identical to the one above, but it includes a red arrow pointing to the 'Download R for Windows' link within the 'Download and Install R' section. This indicates the specific action required in the step-by-step guide.

Step 4: Click on the base link.



The screenshot shows the CRAN homepage with the R logo on the left. To the right, there are several navigation links: [Subdirectories](#), [base](#), [contrib](#), [old contrib](#), [Rtools](#), [ew?](#), [ws](#), [page](#), [urnal](#), [?S](#), and [?C](#). A red arrow points from the text "Click" to the [Subdirectories](#) link. Below these links, there is a section titled "Please do not submit binaries to CRAN. Package developers might want to read the [R FAQ](#) and [R for Windows FAQ](#). Note: CRAN does some checks on these binaries for viruses, but cannot guarantee they are safe."

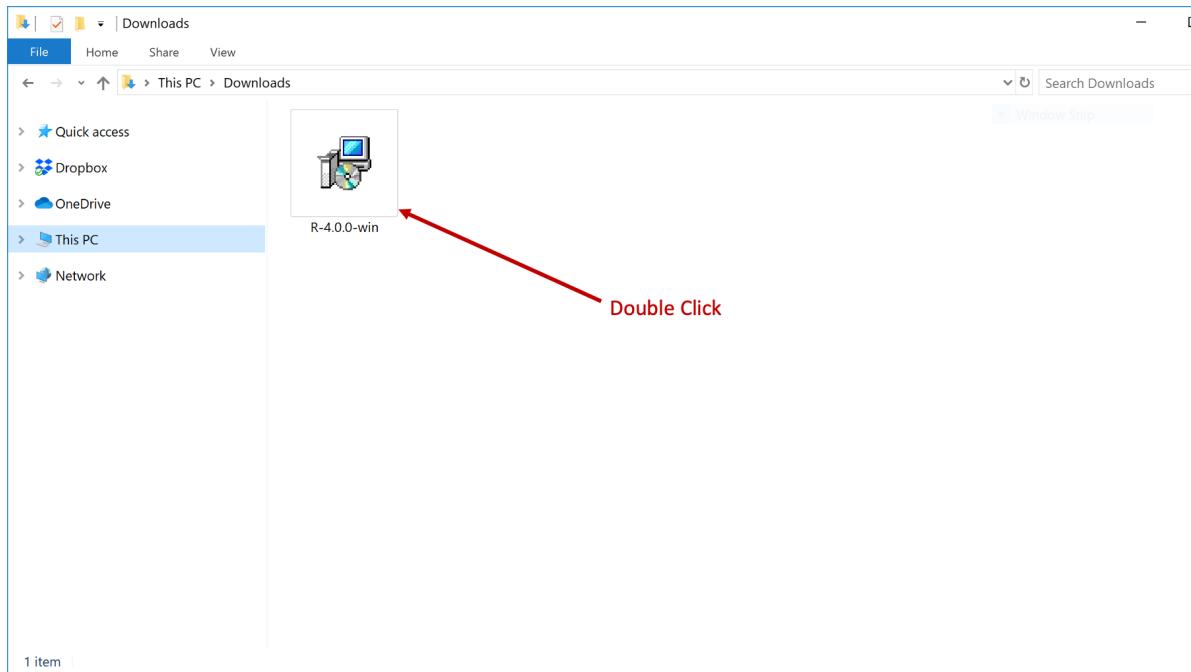
Step 5: Click on the link for the latest version of R. As you are reading this, the newest version may be different than the version you see in this picture, but the location of the newest version should be roughly the same. After clicking, R should start to download to your computer.



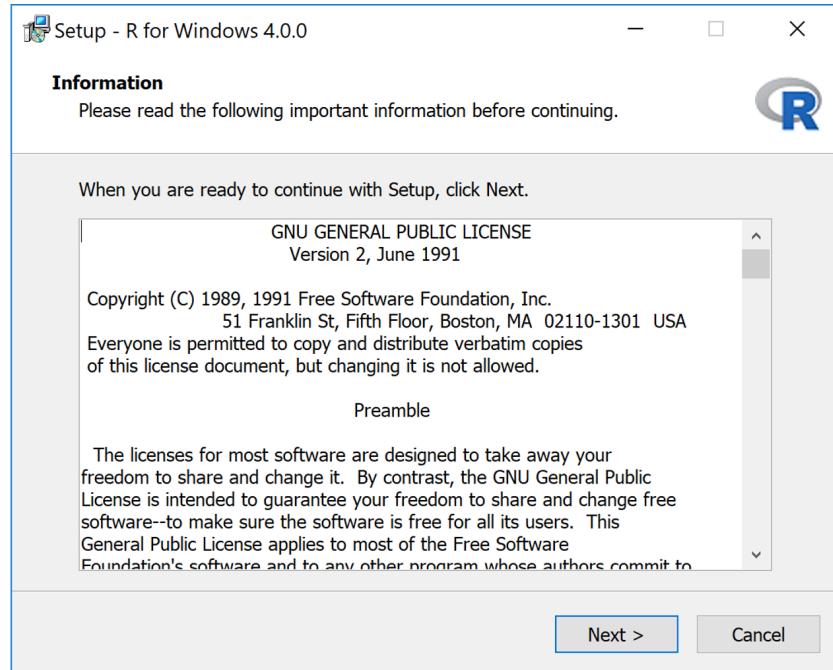
The screenshot shows the CRAN R-4.0.0 for Windows page. At the top, it says "R-4.0.0 for Windows (32/64 bit)". On the left, there's a sidebar with links like CRAN, Mirrors, What's new?, Task Views, Search, About R, R Homepage, The R Journal, Software, R Sources, R Binaries, Packages, Other, Documentation, Manuals, FAQs, and Contributed. A red arrow points from the text "Click" to the [Download R 4.0.0 for Windows](#) link, which is highlighted with a box. Below the download link, there are links for Installation and other instructions and New features in this version. The main content area contains text about double-checking the package against the master server and links for Frequently asked questions, Patches to this release, and Previous releases. It also notes a stable link to the current Windows binary release.

Step 6: Locate the installation file you just downloaded and double click it. Unless you've

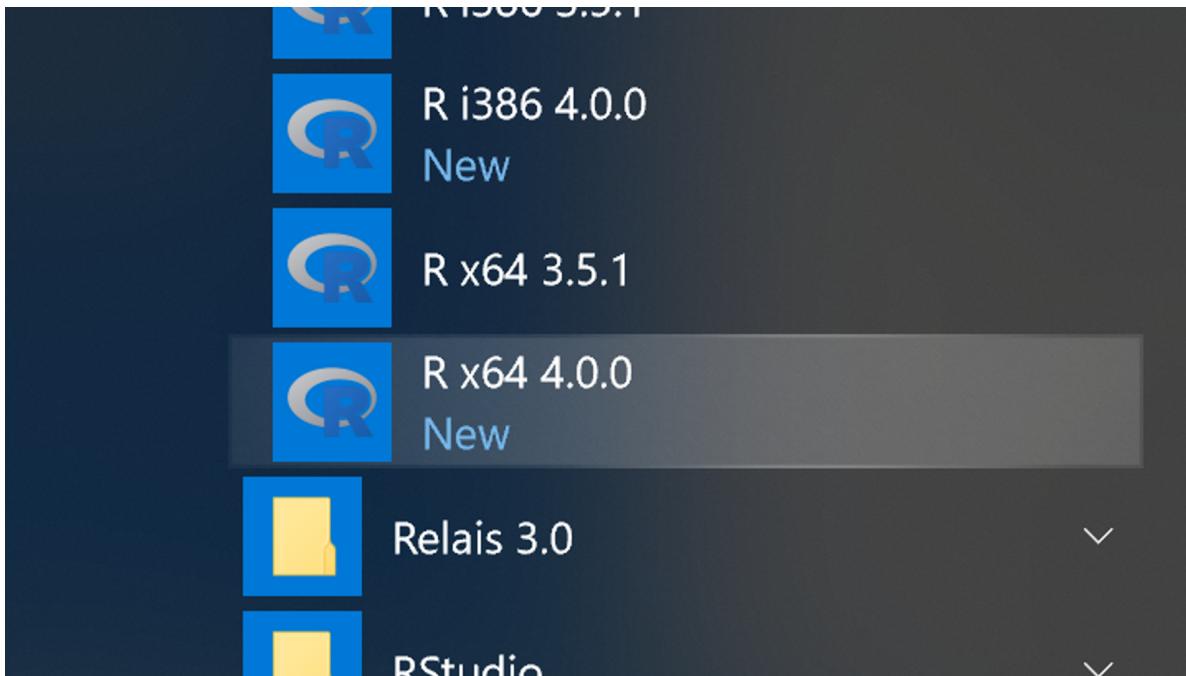
changed your download settings, this file will probably be in your downloads folder. That is the default location for most web browsers.



Step 7: A dialogue box will open that asks you to make some decisions about how and where you want to install R on your computer. We typically just click “Next” at every step without changing any of the default options.

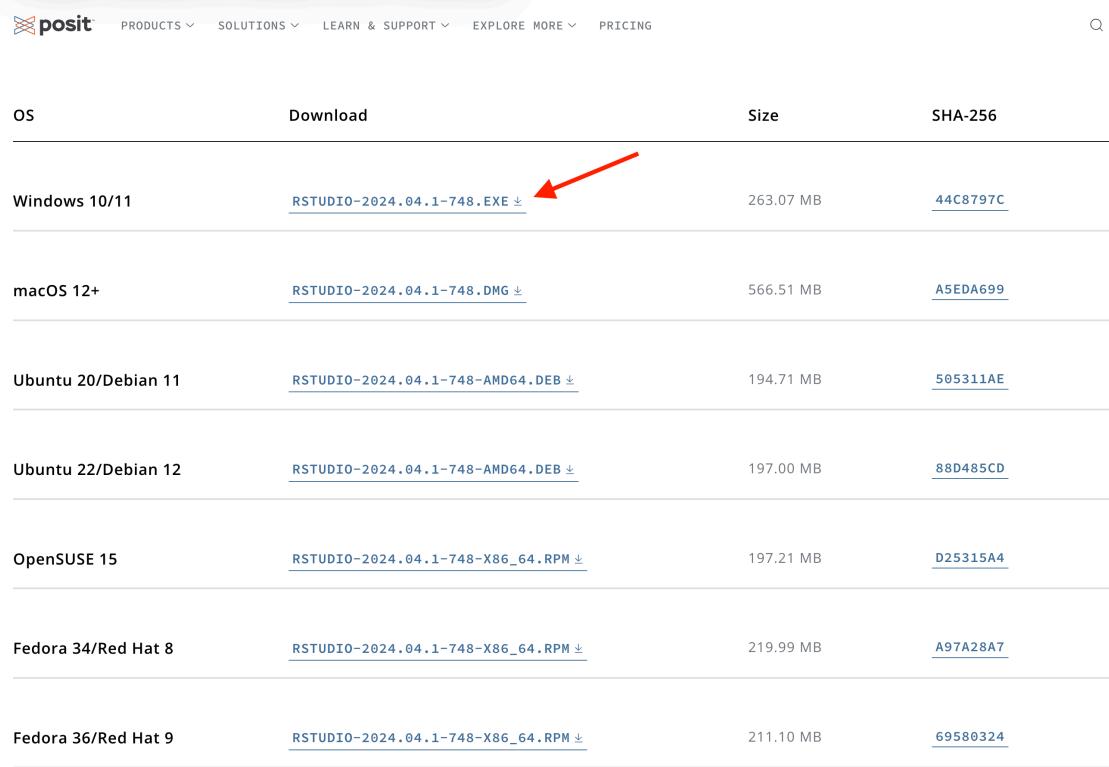


If R installed properly, you should now see it in the Windows start menu.



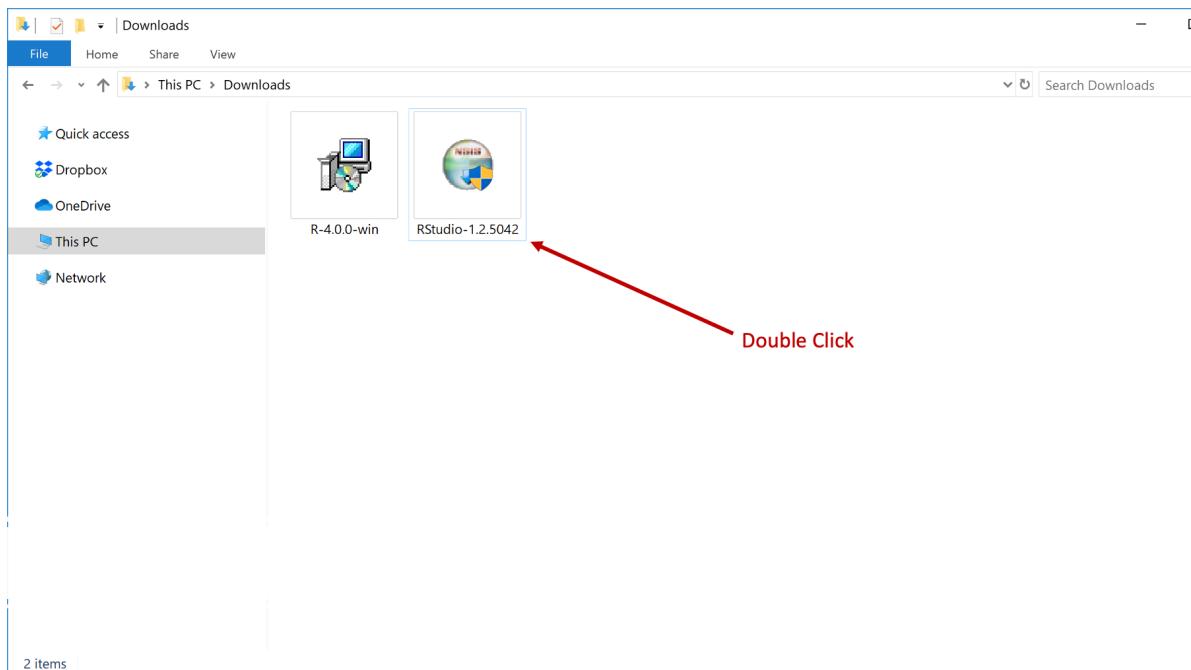
Step 8: Now, we need to install the RStudio IDE. To do this, navigate to the RStudio desktop download website, which is located at <https://posit.co/download/rstudio-desktop/>. On that page, click the button to download the latest version of RStudio for your computer. Note that

the website may look different than what you see in the screenshot below because websites change over time.

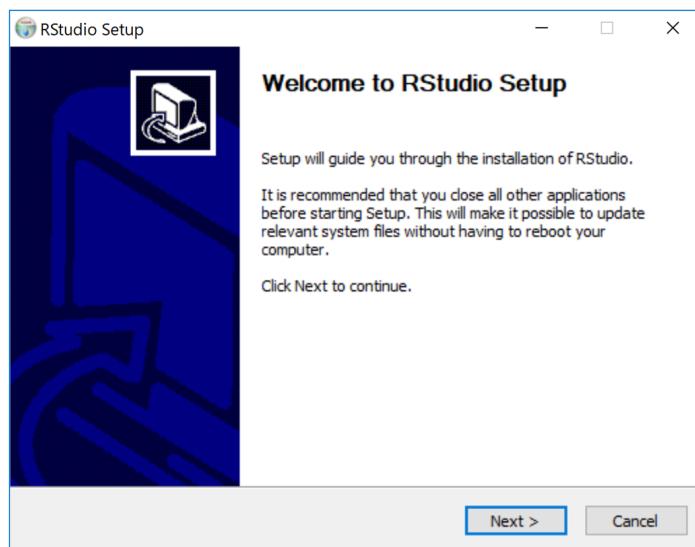


OS	Download	Size	SHA-256
Windows 10/11	RSTUDIO-2024.04.1-748.EXE	263.07 MB	44C8797C
macOS 12+	RSTUDIO-2024.04.1-748.DMG	566.51 MB	A5EDA699
Ubuntu 20/Debian 11	RSTUDIO-2024.04.1-748-AMD64.DEB	194.71 MB	505311AE
Ubuntu 22/Debian 12	RSTUDIO-2024.04.1-748-AMD64.DEB	197.00 MB	88D485CD
OpenSUSE 15	RSTUDIO-2024.04.1-748-X86_64.RPM	197.21 MB	D25315A4
Fedor a 34/Red Hat 8	RSTUDIO-2024.04.1-748-X86_64.RPM	219.99 MB	A97A28A7
Fedor a 36/Red Hat 9	RSTUDIO-2024.04.1-748-X86_64.RPM	211.10 MB	69580324

Step 9: Again, locate the installation file you just downloaded and double click it. Unless you've changed your download settings, this file should be in the same location as the R installation file you already downloaded.

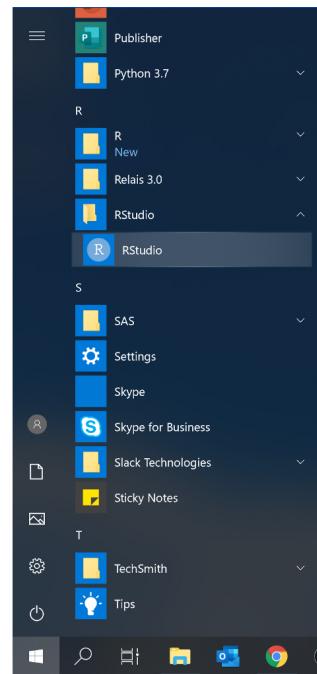


Step 10: Another dialogue box will open and ask you to make some decisions about how and where you want to install RStudio on your computer. We typically just click “Next” at every step without changing any of the default options.

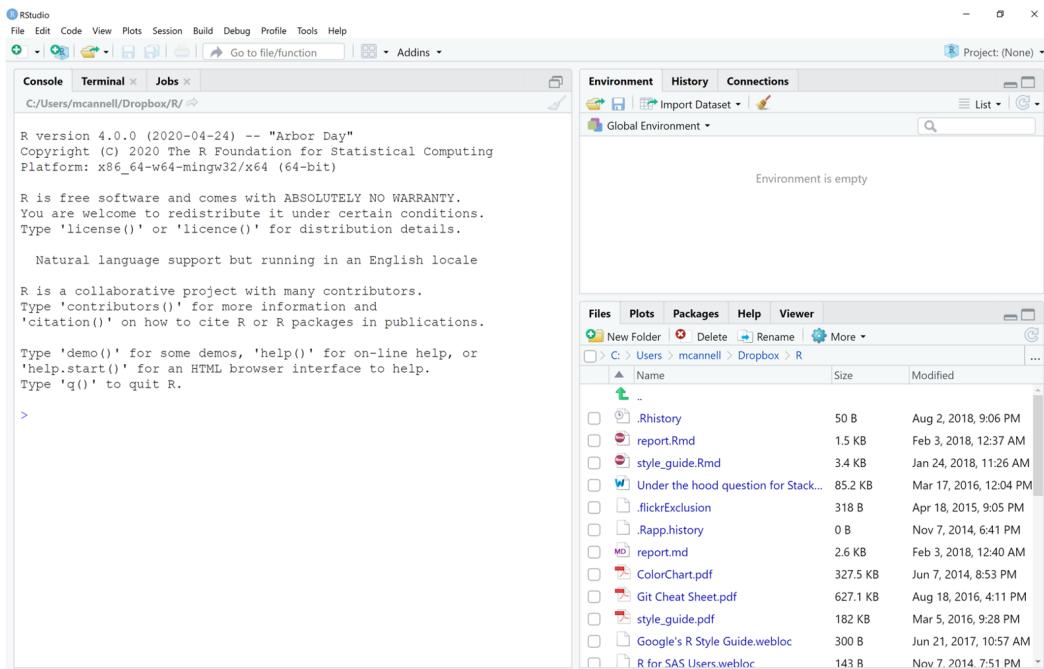


When RStudio is finished installing, you should see RStudio in the Windows start menu. Click

the icon to open RStudio.



The RStudio IDE should open and look something like the window you see here. If so, you are good to go!



2 What is R?

At this point in the book, you should have installed R and RStudio on your computer, but you may be thinking to yourself, “I don’t even know what R is.” Well, in this chapter you’ll find out. We’ll start with an overview of the R language, and then briefly touch on its capabilities and uses. You’ll also see a complete R program and some complete documents generated by R programs. In this book you’ll learn how to create similar programs and documents, and by the end of the book you’ll be able to write your own R programs and present your results in the form of an issue brief written for general audiences who may or may not have public health expertise. But, before we discuss R let’s discuss something even more basic – data. Here’s a question for you: What is data?

2.1 What is data?

Data is information about objects (e.g., people, places, schools) and observable phenomenon (e.g., weather, temperatures, and disease symptoms) that is recorded and stored somehow as a collection of symbols, numbers, and letters. So, data is just information that has been “written” down.

Here we have a table, which is a common way of organizing data. In R, we will typically refer to these tables as **data frames**.

ID	Gender	Height	Weight
001	Male	71	190
002	Male	69	176
003	Female	64	130
004	Female	65	154

Each box in a data frame is called a **cell**.

ID	Gender	Height	Weight
001	Male	71	190
002	Male	69	176
003	Female	64	130
004	Female	65	154

Moving from left to right across the data frame are **columns**. Columns are also sometimes referred to as **variables**. In this book, we will often use the terms columns and variables interchangeably. Each column in a data frame has one, and only one, type. For now, know

that the type tells us what kind of data is contained in a column and what we can *do* with that data. You may have already noticed that 3 of the columns in the table we've been looking at contain numbers and 1 of the columns contains words. These columns will have different types in R and we can do different things with them based on their type. For example, we could ask R to tell us what the average value of the numbers in the height column are, but it wouldn't make sense to ask R to tell us the average value of the words in the Gender column. We will talk more about many of the different column types exist in R later in this book.

ID	Gender	Height	Weight
001	Male	71	190
002	Male	69	176
003	Female	64	130
004	Female	65	154

The information contained in the first cell of each column is called the **column name** (or variable) name.

R gives us a lot of flexibility in terms of what we can name our columns, but there are a few rules.

1. Column names can contain letters, numbers and the dot (.) or underscore (_) characters.
2. Additionally, they can begin with a letter or a dot – as long as the dot is not followed by a number. So, a name like “.2cats” is not allowed.
3. Finally, R has some reserved words that you are not allowed to use for column names. These include: “if”, “else”, “repeat”, “while”, “function”, “for”, “in”, “next”, and “break”.

ID	Gender	Height	Weight
1. Numbers and the dot (.) or underscore (_) characters	Male	71	190
2. Begins with a letter or a dot as long as the dot is not followed by a number	Male	69	176
3. No reserved words			
003	Female	64	130
004	Female	65	154

Moving from top to bottom across the table are **rows**, which are sometimes referred to as records.

ID	Gender	Height	Weight
001	Male	71	190
002	Male	69	176
003	Female	64	130
004	Female	65	154

Finally, the contents of each cell are called **values**.

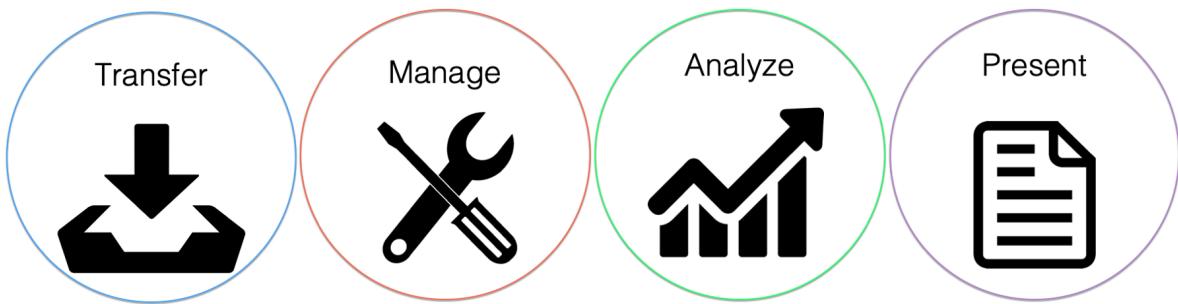
ID	Gender	Height	Weight
001	Male	71	190
002	Male	69	176
003	Female	64	130
004	Female	65	154

You should now be up to speed on some basic terminology used by R, as well as other analytic, database, and spreadsheet programs. These terms will be used repeatedly throughout the course.

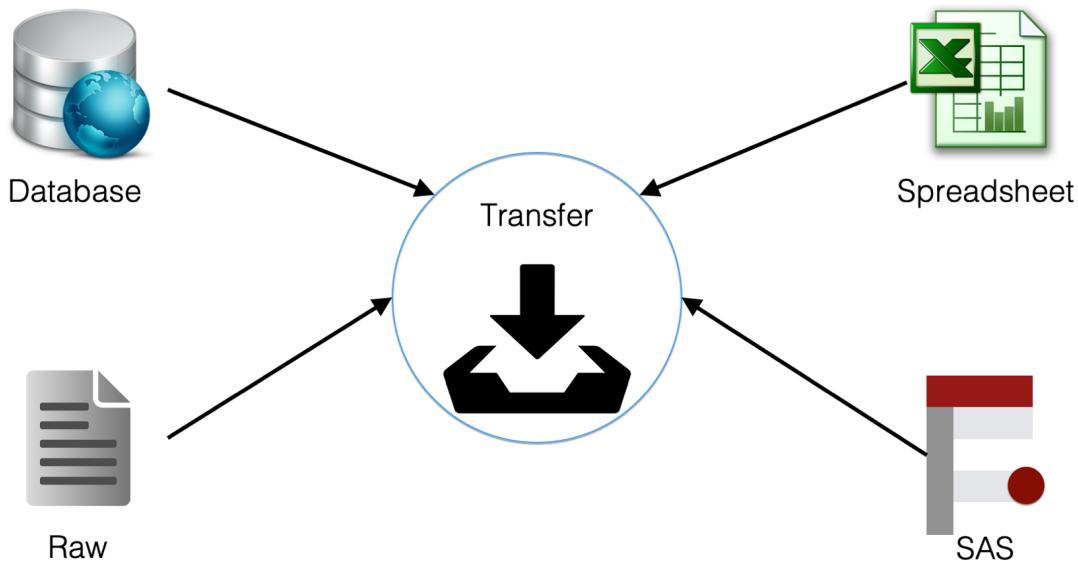
2.2 What is R?



So, what is R? Well, R is an **open source** statistical programming language that was created in the 1990's specifically for data analysis. We will talk more about what open source means later, but for now, just think of R as an easy (relatively) way to ask your computer to do math and statistics for you. More specifically, by the end of this book you will be able to independently use R to transfer data, manage data, analyze data, and present the results of your analysis. Let's quickly take a closer look at each of these.



2.2.1 Transferring data



So, what do we mean by “transfer data”? Well, individuals and organizations store their data

using different computer programs that use different file types. Some common examples that you may come across in epidemiology are database files, spreadsheets, raw data files, and SAS data sets. No matter how the data is stored, you can't do anything with it until you can get it into R, in a form that R can use, and in a location that you can reach. In other words, transferring your data. Therefore, among our first tasks in this course will be to transfer data.

2.2.2 Managing data



This isn't very specific, but managing data is all the things you may have to do to your data to get it ready for analysis. You may also hear people refer to this process as data wrangling or data munging. Some specific examples of data management tasks include:

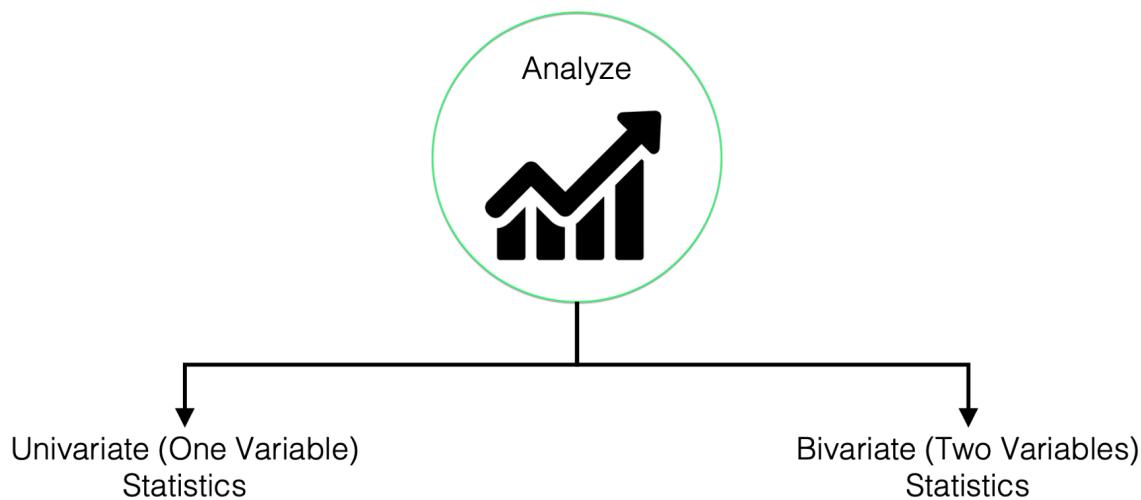
- Validating and cleaning data. In other words, dealing with potential errors in the data.
- Subsetting data. For example, using only some of the columns or some of the rows.
- Creating new variables. For example, creating a BMI variable in a data frame that was sent to you with height and weight columns.
- Combining data frames. For example, combining sociodemographic data about study participants with data collected in the field during an intervention.

You may sometimes hear people refer to the 80/20 rule in reference to data management. This “rule” says that in a typical data analysis project, roughly 80% of your time will be spent on data management and only 20% will be spent on the analysis itself. We can’t provide you with any empirical evidence (i.e., data) to back this claim up. But, as people who have been involved in many projects that involve the collection and analysis of data, we can tell you anecdotally that this “rule” is probably pretty close to being accurate in most cases.

Additionally, it’s been our experience that most students of epidemiology are required to take one or more classes that emphasize methods for analyzing data; however, almost none of them have taken a course that emphasizes data management!

Therefore, because data management is such a large component of most projects that involve the collection and analysis of data, and because most readers will have already been exposed to data analysis to a much greater extent than data management, this course will heavily emphasize the latter.

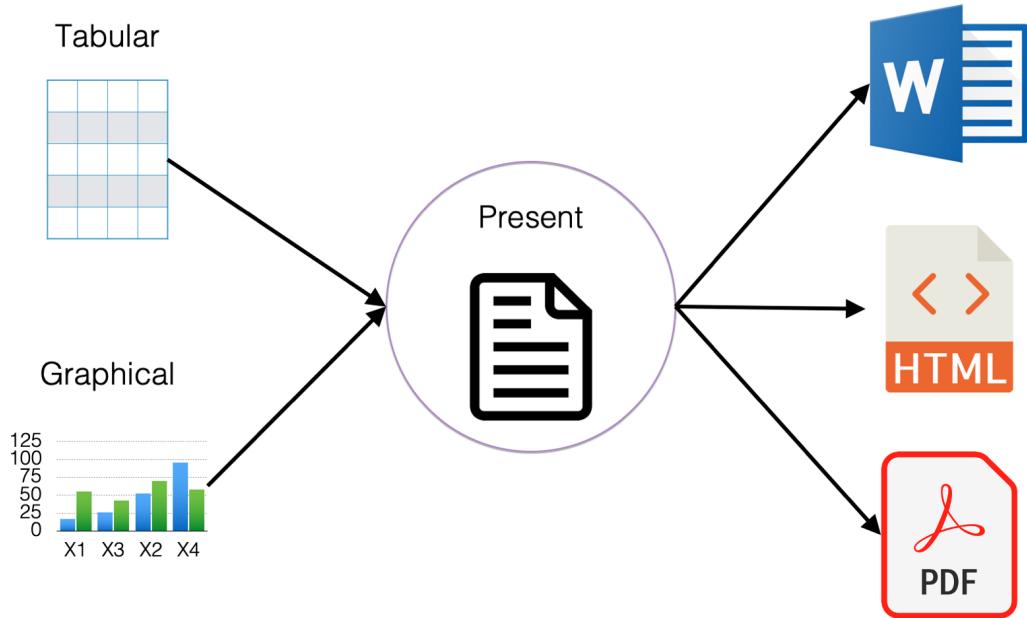
2.2.3 Analyzing data



As just discussed, this is probably the capability you most closely associate with R, and there is no doubt that R is a powerful tool for analyzing data. However, in this book we won’t go beyond using R to calculate basic descriptive statistics. For our purposes, descriptive statistics include:

- Measures of central tendency. For example, mean, median, and mode.
- Measures of dispersion. For example, variance and standard error.
- Measures for describing categorical variables. For example, counts and percentages.
- Describing data using graphs and charts. With R, we can describe our data using beautiful and informative graphs.

2.2.4 Presenting data



And finally, the ultimate goal is typically to present your findings in some form or another. For example, a report, a website, or a journal article. With R you can present your results in many different formats with relative ease. In fact, this is one of our favorite things about R and RStudio. In this class you will learn how to take your text, tabular, or graphical results and then publish them in many different formats including Microsoft Word, html files that can be viewed in web browsers, and pdf documents. Let's take a look at some examples.

1. **Microsoft Word documents.** [Click here](#) to view an example report created for one of my research projects in Microsoft Word.
2. **PDF documents.** [Click here](#) to view a data dictionary we created in PDF format.

3. **HTML files.** Hypertext Markup Language (HTML) files are what you are looking at whenever you view a webpage. You can use R to create HTML files that others can view in their web browser. You can email them these files to view in their web browser, or you can make them available for others to view online just like any other website. [Click here](#) to view an example dashboard we created for one of my research projects.
4. **Web applications.** You can even use R to create full-fledged web applications. View the [RStudio website](#) to see some examples.

Part II

Coding Tools and Best Practices

3 Quarto Files

Part III

Collaboration

4 Using git and GitHub

Part IV

References

References

1. GitHub. *About Issues*. Github; 2024.
2. R Core Team. *What Is r?* R Foundation for Statistical Computing; 2024.
3. GitHub. About repositories. Published online December 2023.
4. RStudio. RStudio. Published online 2020.

A Glossary

Console. Coming soon.

Data frame. For our purposes, data frames are just R’s term for data set or data table. Data frames are made up of columns (variables) and rows (observations). In R, all columns of a data frame must have the same length.

Functions. Coming soon.

- **Arguments:** Arguments always go *inside* the parentheses of a function and give the function the information it needs to give us the result we want.
- **Pass:** In programming lingo, you *pass* a value to a function argument. For example, in the function call `seq(from = 2, to = 100, by = 2)` we could say that we passed a value of 2 to the `from` argument, we passed a value of 100 to the `to` argument, and we passed a value of 2 to the `by` argument.
- **Returns:** Instead of saying, “the `seq()` function *gives us* a sequence of numbers...” we could say, “the `seq()` function *returns* us a sequence of numbers...” In programming lingo, functions *return* one or more results.

Global environment. Coming soon.

Issue (GitHub) GitHub’s documentation says issues are “items you can create in a repository to plan, discuss and track work. Issues are simple to create and flexible to suit a variety of scenarios. You can use issues to track work, give or receive feedback, collaborate on ideas or tasks, and efficiently communicate with others.”¹

Objects. Coming soon.

R R’s documentation says “R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues.”² R is open source, and you can download it for free from The Comprehensive R Archive Network (CRAN) at <https://cran.r-project.org/>.

Repository GitHub’s documentation says “a repository contains all of your code, your files, and each file’s revision history. You can discuss and manage your work within the repository.”³ A repository can exist *locally* as a set of files on your computer. A repository can also exist *remotely* as a set of files on a sever somewhere, for example, on GitHub.

RStudio RStudio is an integrated development environment (IDE) for R. It includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging and workspace management. RStudio is available in open source and commercial editions and runs on the desktop (Windows, Mac, and Linux) or in a browser connected to RStudio Server or RStudio Server Pro (Debian/Ubuntu, Red Hat/CentOS, and SUSE Linux).⁴