

# Using R Packages for Reproducible Workflows

Michael Dumelle

September 22, 2021



# Contents

<b>Welcome</b>	<b>5</b>
Schedule . . . . .	6
About Me . . . . .	6
Acknowledgements . . . . .	6
Getting Started . . . . .	7
<b>1 Building an R Package</b>	<b>9</b>
1.1 What is an R Package? . . . . .	9
1.2 Creating an R Package . . . . .	10
1.3 Debugging . . . . .	29
1.4 Additional Resources . . . . .	34
<b>2 Making an R Package a Research Compendium</b>	<b>35</b>
2.1 What is a Research Compendium? . . . . .	35
2.2 Why a Research Compendium? . . . . .	35
2.3 Why an R Package for a Research Compendium? . . . . .	36
2.4 Turning RPRW Into A Research Compendium . . . . .	36
2.5 Research Compendium Recap . . . . .	40
2.6 Sharing Your Research Compendium . . . . .	40
2.7 Hosting Your Research Compendium . . . . .	41
2.8 An Example R Package Reserach Compendium . . . . .	42
<b>3 Extensions</b>	<b>43</b>
3.1 R Projects . . . . .	43
3.2 R Markdown . . . . .	45
3.3 rticles . . . . .	47
3.4 Git and GitHub . . . . .	49
3.5 Continuous Integration . . . . .	50
3.6 Reproducible Examples . . . . .	51
3.7 Code Style . . . . .	52
3.8 Additional Tips . . . . .	52
3.9 Thank you! . . . . .	53

<b>4</b>	<b>Exercise Solutions</b>	<b>55</b>
	Building an R Package . . . . .	55
	Making an R Package a Research Compendium . . . . .	60
	Extensions . . . . .	60

# Welcome

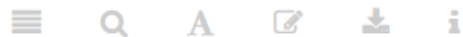
Welcome to the 2021 EPA R Workshop titled “Using R Packages for Reproducible Workflows” by me, Michael Dumelle – I am glad to have you here! You should have R and RStudio installed before starting this workshop. If not, no problem – just follow along with me as we proceed through the workshop and revisit this book later. R code and output from this workshop are in blocks with gray background like this

```
x <- c(1, 2, 3)
print(x)
```

```
#> [1] 1 2 3
```

You can copy and paste these blocks by hovering your cursor in the upper-right portion of the block and clicking the copy button (which looks like two pages that slightly overlap). You can either follow along throughout the workshop by copying and pasting the R code onto your machine and/or by reading the R code output the book provides. Sometimes when I use functions, I will prefix the function call `package::` so that it is clear what package the function is from. For example, `ggplot2::ggsave()` calls the `ggsave()` function in the `ggplot2` package.

There are two types of hyperlinks in this book: internal and external. Internal hyperlinks take you to the place in the book. External hyperlinks will take you to a webpage outside of the book. Both internal and external links will be described in text but the actual link will follow the relevant text and read (internal link) or (external link). There are several options in the upper-left toolbar of the book.



From left to right, they let you:

- Collapse or expand the table of contents
- Search throughout the document

- Adjust font size (small or large), font type (Serif or Sans), and color scheme (White, Sepia, or Night)
- Edit the book (you don't have access to this one)
- Download the book as a PDF or EPUB book (formatting warning: the book was built for a workshop format, not a PDF or EPUB format)
  - If you are unable to download the PDF via this link, it is also available [here](#) (external link)
- Use a Javascript keyboard shortcut helper

This workshop book was published using Bookdown [Xie, 2016]. The book will be maintained for an indefinite amount of time after the workshop, so feel free to check back in for a refresher whenever you would like!

## Schedule

Here is the tentative schedule for today:

- 1:00 - 1:15 EDT: Introductions and Getting Started (internal link)
- 1:15 - 2:00 EDT: Building an R Package Part I (internal link)
- 2:00 - 2:10 EDT: Break
- 2:10 - 2:50 EDT: Building an R Package Part II (internal link)
- 2:50 - 3:05 EDT: Break
- 3:05 - 3:50 EDT: Making an R Package a Research Compendium (internal link)
- 3:50 - 4:00 EDT: Break
- 4:00 - 4:30 EDT: Extensions (internal link)
- 4:30 - 4:40 EDT: Break
- 4:40 - 5:00 EDT: Questions

We will try to roughly adhere to this schedule, but it is possible that we stray slightly from it – this is my first time giving the workshop!

## About Me

I finished my PhD in statistics at Oregon State University in 2020. I now work in ORD-CPHEA-PESD-FEB as a statistician, primarily on the National Aquatic Resource Surveys (NARS) team. My main research interests are sampling and survey design, spatial statistics, and software development. To learn more about me and my work, visit my website (external link).

## Acknowledgements

I would like to thank Charlotte Wickham, Hadley Wickham, Jenny Bryan, and Yihui Xie for the immense impact their work has had on my programming journey. Much of this workshop draws from heavily their inspiration. I would

also like to thank everyone who helped me create and give this workshop. A special thanks to Scott Bohning for providing feedback after the first version of the workshop that significantly improved future versions.

## Getting Started

The first thing we need to do is install and load the devtools ([external link](#)) package by running

```
install.packages("devtools", dependencies = TRUE)
library(devtools)
```

The devtools package contains tools to make developing R packages easier, and we will use functions in devtools frequently throughout this workshop. devtools is closely linked to the usethis ([external link](#)) package, which is installed alongside devtools. The usethis package contains tools that make package development less tedious and more structured.

In this workshop we will build an R package named RPRW (**R Packages for Reproducible Workflows**) designed to quantify properties of rivers in North America. Then we will turn RPRW into a research compendium, a standard by which to organize a body of work.

If you don't want to build RPRW now and just want to look at the finished product, you can download in its entirety by running

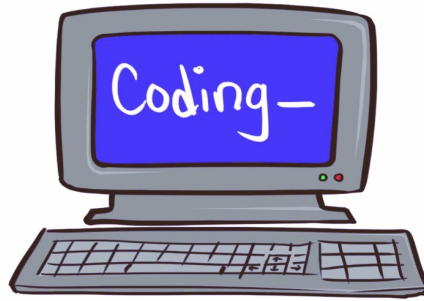
```
devtools::install_github(
  repo = "michaeldumelle/cRPRW",
  ref = "main",
  build_vignettes = TRUE
)
```

Notice that it has a “c” (short for “companion”) prefix to RPRW to avoid conflicts with the version you build yourself today.

Then you can load cRPRW by running

```
library(cRPRW)
```

Now that we have devtools installed and have talked about RPRW, let's get started building this R package!





# Chapter 1

## Building an R Package

### 1.1 What is an R Package?

An R package is a collection of code, data, documentation, and tests with a particular structure that can be shared with others. R packages are commonly downloaded from the Comprehensive R Archive Network (CRAN) ([external link](#)). You can install them from CRAN by running

```
install.packages("package_name")
```

load them by running

```
library(package_name)
```

and get help by running

```
?package_name
```

After loading packages, you can refer to functions within them by interactively running `function_name()`. To be explicit about which functions come from which packages, I tend to refer to them as `package_name::function_name()`. The package name prefix actually prevents the need to explicitly load the package using `library()`.

One of the many reasons R packages are so useful is because they are the fundamental vessel by which to share R code. If your code is in a package, others can easily download and use it – this is the essence of open-source software! Most R users have some experience with at least one package, so they will likely be familiar with how to use yours after you share it. But sharing code is not the only benefit of creating packages – a major benefit of using packages is that they enforce your code is built using a certain structure. Adhering to this structure:

1. saves you time – You don't need to think about how to organize your files

because packages tell you how these files must be organized!

- This rigid structure is especially helpful for me – before I started creating packages my R scripts would be saved in all sorts of locations with no particular organization structure. This made it *very challenging* to come back to my work later and find a particular script.
2. gives you access to a set of standardized tools – people have created many useful tools that work with R packages, so take advantage of them!
    - e.g. devtools and usethis
  3. forces you to frequently document your work
    - Before I started using R packages, when I would come back my old code, I was convinced someone else wrote it – I basically had to rewrite it all to understand it. R packages help prevent this.
  4. keeps your code, documentation, and package dependencies all in sync
  5. guides your research compendium
  6. can make your research compendium completely reproducible and easy to share
    - This is because packages are built upon R projects, which we discuss more later.

### 1.1.1 Exercises

1. What are some of your favorite R packages?
2. Of those we have talked about so far, what benefits of R packages are most appealing to you?

## 1.2 Creating an R Package

Next we will create the RPRW package together! If this is your first R package, then an extra special congratulations to you – this is a big milestone! For those interested in viewing the source code of the RPRW package, visit the GitHub repository [here](#) (external link).

### 1.2.1 The Motivating Dataset

Suppose we want RPRW to summarize the length (in kilometers) and discharge (meters per second cubed) of North American rivers. Suppose we also want RPRW to easily accommodate many subsets of rivers based on their name. Below are our data of interest:

```
#>           Missouri Mississippi Yukon Colorado Arkansas Columbia Red Canadian
#> length      3768          3544  3190      2330      2322      2000 1811      1458
#> discharge    1956         18400  6340        40      1004      7730  852       174
```

If the structure of `river` is atypical to you, see Exercise 3.7.1.1 (internal link).



### 1.2.2 The First Step

The first step is using devtools to create a package template for us. There are two ways to easily create this template: `usethis::create_package()` and the RStudio toolbar. In this class we will use `usethis::create_package()`. Suppose you want the R package to be contained in the folder “RPRW” at the location on your computer specified by “`path_to_RPRW_package`” (which must exist). To create the R package, run

```
usethis::create_package("path_to_RPRW_package/RPRW")
```

```
✓ Creating 'path_to_RPRW_package/RPRW/'
✓ Setting active project to 'path_to_RPRW_package/RPRW'
✓ Creating 'R/'
✓ Writing 'DESCRIPTION'
Package: RPRW
Title: What the Package Does (One Line, Title Case)
Version: 0.0.0.9000
Authors@R (parsed):
  * First Last <first.last@example.com> [aut, cre] (YOUR-ORCID-ID)
Description: What the package does (one paragraph).
License: `use_mit_license()`, `use_gpl3_license()` or friends to
  pick a license
Encoding: UTF-8
LazyData: true
Roxygen: list(markdown = TRUE)
RoxygenNote: 7.1.1
✓ Writing 'NAMESPACE'
✓ Writing 'RPRW.Rproj'
✓ Adding '^RPRW\\.Rproj$' to '.Rbuildignore'
✓ Adding '.Rproj.user' to '.gitignore'
✓ Adding '^\\.Rproj\\.user$' to '.Rbuildignore'
✓ Opening 'path_to_RPRW_package/RPRW/' in new RStudio session
✓ Setting active project to '<no active project>'
✓ Setting active project to 'path_to_RPRW_package/RPRW'
```

RStudio will automatically create a new window, and you can safely close the old RStudio window. As you can see in the “Files” tab of the bottom-right RStudio pane, this new window is “placed” within “`path_to_RPRW_package/RPRW`” and you can see several newly created files.

				
<input type="checkbox"/>		.gitignore	12 B	Sep 10, 2021, 9:07 AM
<input type="checkbox"/>		.Rbuildignore	34 B	Sep 10, 2021, 9:07 AM
<input type="checkbox"/>		DESCRIPTION	504 B	Sep 10, 2021, 9:07 AM
<input type="checkbox"/>		NAMESPACE	46 B	Sep 10, 2021, 9:07 AM
<input type="checkbox"/>		R		
<input type="checkbox"/>		RPRW.Rproj	436 B	Sep 10, 2021, 9:07 AM

Notice the `.Rproj` file, which indicates that your package is also an R project. We talk about R projects later, but the takeaway is that they provide a convenient default place for R to look for files (more formally, this default place is called a working directory). This is convenient because the functions in `devtools` and `usethis` we explore later will operate on the files in our working directory, which is forced to be the root of the “RPRW” folder by the package structure. Finally, notice that the top-right RStudio pane now has a “Build” tab, which contains some useful tools for your R package.

If you instead want to use the RStudio toolbar approach, navigate to “File -> New Project -> New Directory -> R Package -> Create Project” in the RStudio toolbar. With this approach, you will not see any output in the R console of your original RStudio window.

### 1.2.3 Some preliminaries

Because `usethis::create_package()` created a new RStudio window for us, we need to reload `devtools` (which also loads `usethis`):

```
library(devtools)
```

Next we will store the river data (the motivating dataset) in the global environment so we can work with it:

```
river <- data.frame(
  Missouri = c(3768, 1956),
  Mississippi = c(3544, 18400),
  Yukon = c(3190, 6340),
  Colorado = c(2330, 40),
  Arkansas = c(2322, 1004),
  Columbia = c(2000, 7730),
  Red = c(1811, 852),
  Canadian = c(1458, 174)
)
```

```
rownames(river) <- c("length", "discharge")
print(river)
```

```
#>           Missouri Mississippi Yukon Colorado Arkansas Columbia Red Canadian
#> length      3768      3544  3190    2330    2322    2000 1811    1458
#> discharge   1956    18400  6340     40    1004    7730  852     174
```

### 1.2.4 The First Function

R packages are composed of several functions that work together to accomplish a set of goals. Within your R package, the folder named “R” is where functions live. Let’s create our first function, `river_means()`, which finds the mean river length and discharge of desired rivers. To create an R script named `river_means.R` in the package’s R folder, run

```
usethis::use_r("river_means")
```

You will see some output in the console

```
* Modify 'R/river_means.R'
* Call `use_test()` to create a matching test file
```

and the R script is opened in the editor window of RStudio and placed in the R folder.



Generally, the name of your file should match the name of the function, and you should use separate files for separate functions. More experienced R users – it is okay to break this rule every once in a while depending on the context.

The body of `river_means()` looks like

```
river_means <- function(data, pattern) {
  desired_rivers <- grep(pattern, names(data))
  new_data <- data[, desired_rivers]
  rowMeans(new_data)
}
```

There are two arguments in `river_means()`:

1. **data**: A data frame (whose rows represent length or discharge and whose columns represent rivers)
  - Making **data** the first argument to `river_means()` makes it naturally pipeable ([external link](#))
2. **pattern**: A character string indicating the pattern by which to subset the rivers in **data** before calculating means (all rivers whose names satisfy the pattern are kept)

There are three lines of code in `river_means()`:

1. Find the variables (rivers) in `data` whose names contain the pattern.
2. Subset `data` to include only the rivers whose names contain the pattern
3. Find the mean length and discharge of the rivers whose names contain the pattern

By default, the function returns the object evaluated by its last line of code. So `river_means()` returns the output from `rowMeans()`. You can force the function to return a different object earlier using `return(object)`.

You can leave comments in your function by prefixing a line with `#`. Comments are used to communicate in plain language the intent of your code. Leaving comments is usually good practice because it helps others (including future you) better understand the goals of your code.

Paste this code into your version of `river_means.R` and click the save icon (or `Ctrl + S`).

### 1.2.5 `devtools::load_all()`

Now that we have written `river_means()`, it is natural to want to try it out. Though we wrote our function in `river_means.R`, to use it interactively we must tell R that it is an object we want to use. If you were not making an R package, you may store `river_means()` by running `river_means()` in “R/river\_means.R” script, copying and pasting `river_means()` into the console and pressing `Enter`, or by running

```
source("R/river_means.R")
```

These approaches put `river_means()` in the global environment (look at the environment tab in the upper-right pane of RStudio). Unfortunately messing around too much in the global environment can easily introduce errors in your functions or have them behave in unexpected ways.

To approach we use here is to leverage `devtools`, since after all, you are bothering to create a package, so you may as well use the tools designed for them! The `devtools` function `devtools::load_all()` emulates the process of building, installing, and loading a package. This has a few benefits. First, the functions are “sourced” into a better spot (not the global environment). Second, it makes all of your functions (and data, which we get to soon) available through a single line of code rather than running and/or sourcing several functions separately. I highly recommend you practice using `devtools::load_all()` instead of navigating the global environment.

```
devtools::load_all()
```

By running `search()` or selecting the drop-down “Global Environment” window in the “Environment” pane (upper-right pane in RStudio), you see `package:RPRW`, which indicates `devtools::load_all()` is treating RPRW like

a package. If you make any changes to your functions, make sure to save them and rerun `devtools::load_all()`.

Now we get to try out `river_means()`! Suppose we want to summarize the rivers whose names start with “Mi” (in `river`, this is only Missouri and Mississippi).

```
river_means(river, "Mi")
```

```
#>   length discharge
#>    3656      10178
```

Next suppose we want to summarize rivers whose names start with “Y” (Yukon) or “R” (Red)?

```
river_means(river, "Y|R")
```

```
#>   length discharge
#>  2500.5    3596.0
```

Hooray – our function works! Give yourself a congratulations :)! And for more on patterns and strings in R, visit [here](#) (external link) and [here](#) (external link).

### 1.2.5.1 Exercises

1. How many functions does `river_means()` call within the body of the function? What package are these functions in?
2. Try using `river_means()` with a few different patterns. Do you notice anything strange?

## 1.2.6 Creating Package Data

So far the `river` data has been stored in our global environment. What if you want to save this data in an easily accessible location and/or share it with others? Another package development tool to the rescue – `usethis::use_data()` (are you starting to notice the benefits of the additional tools `devtools` and `usethis` provide packages?).

```
usethis::use_data(river)
```

```
✓ Creating 'data/'
✓ Saving 'river' to 'data/river.rda'
```

You will see that at the root of your R package (inside “`path_to_RPRW_package/RPRW`”), there is now a folder called `data`. In `data`, there is a file called `river.rda`. `river_rda` contains the `river` data (in a compressed format built for R objects) and can now be loaded alongside your functions when running `devtools::load_all()`. For illustration, let’s remove `river` from our global environment and then load it using `devtools::load_all()`:

```
rm(river) # remove the river data from our global environment
devtools::load_all() # load the river data
print(river) # view the river data
```

```
#>           Missouri Mississippi Yukon Colorado Arkansas Columbia Red Canadian
#> length      3768          3544  3190      2330      2322      2000 1811      1458
#> discharge   1956          18400  6340        40      1004      7730  852      174
```

### 1.2.7 roxygen Comments

After writing `river_means()`, you should document it while its structure is fresh in your mind. This makes it easier for others (including future you) to understand how to use `river_means()`. Documenting R functions is a bit more complicated than leaving standard comments (`#`). R package documentation must be written in "R documentation (.Rd) format. The simplest way to write R documentation for functions is to use a special type of comment with your functions: *roxygen* comments (view the `roxygen2` (external link) package for more). *roxygen* comments are prefixed with `#'` (instead of `#`) and have special **tags** associated with them. *roxygen* comment templates can be automatically inserted into your R script by putting your cursor somewhere in the body of a function and do either of:

1. In the upper-left toolbar, go to "Code -> Insert Roxygen Skeleton"
2. Pressing `Ctrl/Cmd + Alt + Shift + R`

Inserting this *roxygen* template into `river_means()` yields

```
#' Title
#'
#' @param data
#' @param pattern
#'
#' @return
#' @export
#'
#' @examples
river_means <- function(data, pattern) {
  desired_rivers <- grep(pattern, names(data))
  new_data <- data[, desired_rivers]
  rowMeans(new_data)
}
```

The `@` indicates the tag. The template inserts four tags for us:

1. `@param`: to document function arguments
2. `@return`: to document function output
3. `@export`: to make the function available after installing the package



- You may not want to export every function in your package (e.g. helpers)

4. `@examples`: to provide examples of using the function

Let's document `river_means()`:

```
## Means of river lengths and discharges
##
## @param data A data frame with two rows. The first row indicates river length and
## second row indicates river discharge. The columns of data indicate river names.
## @param pattern A pattern by which to include only particular rivers
##
## @return The mean river length and mean river discharge for the desired rivers
## @export
##
## @examples
## data("river")
## river_means(river, pattern = "Mi")
river_means <- function(data, pattern) {
  desired_rivers <- grep(pattern, names(data))
  new_data <- data[, desired_rivers]
  rowMeans(new_data)
}
```

Next we need to turn these roxygen comments into a new format that powers the actual documentation. The details of this new format are not critical – the important take-away is that the roxygen comments are automatically transformed to the new format by running

```
devtools::document()
```

```
i Updating RPRW documentation
i Loading RPRW
Writing NAMESPACE
Writing river_means.Rd
```

`devtools::document()` creates a new folder at the root of the package called `man` that now contains a file called `river_means.Rd`

 ..

☐  `river_means.Rd` 658 B Sep 10, 2021, 11:45 AM

The `river_means.Rd` file is the result of transforming the roxygen comments to documentation that packages use. As you can see, `river_means.Rd` is a bit more complicated than the roxygen comments in `river_means()`. Luckily, `river_means.Rd` is automatically generated. The broader, important point of all this is that after running `devtools::document()`, you can view the documentation of `river_means()`!

**?river\_means**

river\_means {RPRW}

R Documentation

**Means of river lengths and discharges****Description**

Means of river lengths and discharges

**Usage**

```
river_means(data, pattern)
```

**Arguments**

**data** A data frame with two rows. The first row indicates river length and second row indicates river discharge. The columns of data indicate river names.

**pattern** A pattern by which to include only particular rivers

**Value**

The mean river length and mean river discharge for the desired rivers

**Examples**

```
data("river")
river_means(river, "Mi")
```

---

[Package RPRW version 0.0.0.9000 ]

Note that `devtools::document()` simultaneously updates documentation for all of the `.R` files in the `R` folder that have roxygen comments. From now on I'll stop reminding you every time, but again, notice the benefits of leveraging the tools that `devtools` and `usethis` provide.

For more on documentation in R packages (including documenting data), visit [here](#) (external link) and the vignettes [here](#) (external link). Though we skip documenting the river data here, the `RPRW` package available from GitHub does document it (the documentation can be viewed by running `?river` after `library(RPRW)`).

### 1.2.8 The Second Function

You may be wondering how to tell your package to use functions from other packages. `river_means()` uses a few other functions available in R (see Exercise 1.1.1.1 (internal link): `<-`, `grep()`, `names()`, `[]`, and `rowMeans()`). How does R know how to use the right `names()` function when executing `river_means()`? These questions related to an package's `NAMESPACE`, which we discuss a bit later.

The functions used in `river_means()` are all from the base package in R (external link). The base package is automatically installed alongside R. When

using functions from the base package in your package, you don't have to give R any special warning. This is the only package that operates this way, however. When using functions from other packages in your package, you need to let R know what package that function comes from (this is true even for other packages automatically installed alongside R - datasets, graphics, grDevices, methods, stats, and utils). To illustrate this process, suppose we want to create a new function in our package called `river_medians()` that operates like `river_means()` but computes medians instead of means. First we create the script in our R folder by running

```
usethis::use_r("river_medians")
```

Then we can write `river_medians()` (there is no `rowMedians()` function so we need to use `apply()` to summarize across rows – see `?apply` for more detail):

```
river_medians <- function(data, pattern) {
  desired_rivers <- grep(pattern, names(data))
  new_data <- data[, desired_rivers]
  apply(new_data, 1, median)
}
```

But wait! Run

```
?median
```

See anything different? You'll notice that `median()` is in the stats package (external link), not the base package. We need to let R know that when `river_medians()` uses `median()`, `median()` is from the stats package. To accomplish this, there are two steps to take.

The first step is to tell your package that at least one function from the stats package is required by running

```
usethis::use_package("stats")
```

```
✓ Adding 'stats' to Imports field in DESCRIPTION
* Refer to functions with `stats::fun()``
```

The package stats was added to the Imports field in the DESCRIPTION file at the root of your package. We will talk more about DESCRIPTION in a bit, but for now just remember that it contains metadata (high-level data) about your package. No matter how many times your package uses any function from stats, you only need to run `usethis::use_package("stats")` once. If you accidentally misspell a package name in here, you should delete the misspelled name in the Imports section of the DESCRIPTION file, which we cover later.

The second step is to tell the `river_medians()` function to use the `median()` function from stats. There are a few ways to do this. The way that some consider “best practice” is to preface any outside functions (aside from those in base) with `package::`. Incorporating this into `river_medians()` and inserting

roxygen documentation yields

```
#' Medians of river lengths and discharges
#'
#' @param data A data frame with two rows. The first row indicates river length and
#'   second row indicates river discharge. The columns of data indicate river names.
#' @param pattern A pattern by which to include only particular rivers
#'
#' @return The median river length and mean river discharge for the desired rivers
#' @export
#'
#' @examples
#' data("river")
#' river_medians(river, pattern = "Mi")
river_medians <- function(data, pattern) {
  desired_rivers <- grep(pattern, names(data))
  new_data <- data[, desired_rivers]
  apply(new_data, 1, stats::median)
}
```

This approach does add some extra typing and can be cumbersome if you are using many outside functions. The `package::` prefix can be avoided by importing `median()` from `stats` to `river_median()` using the roxygen tag `@importFrom`:

```
#' Medians of river lengths and discharges
#'
#' @param data A data frame with two rows. The first row indicates river length and
#'   second row indicates river discharge. The columns of data indicate river names.
#' @param pattern A pattern by which to include only particular rivers
#'
#' @return The median river length and mean river discharge for the desired rivers
#' @importFrom stats median
#' @export
#'
#' @examples
#' data("river")
#' river_medians(river, pattern = "Mi")
river_medians <- function(data, pattern) {
  desired_rivers <- grep(pattern, names(data))
  new_data <- data[, desired_rivers]
  apply(new_data, 1, median)
}
```

A third option is relevant if you are using several functions from `stats`. In this case, it may be easier to use the `@import` tag, which imports all functions from `stats` to `river_median()`.

```

#' Medians of river lengths and discharges
#'
#' @param data A data frame with two rows. The first row indicates river length and
#'   second row indicates river discharge. The columns of data indicate river names.
#' @param pattern A pattern by which to include only particular rivers
#'
#' @return The median river length and mean river discharge for the desired rivers
#' @import stats
#' @export
#'
#' @examples
#' data("rivers")
#' river_medians(rivers, pattern = "Mi")
river_medians <- function(data, pattern) {
  desired_rivers <- grep(pattern, names(data))
  new_data <- data[, desired_rivers]
  apply(new_data, 1, median)
}

```

Be careful when using `@importFrom` or `@import` (especially `@import`). If two functions from different packages have the same name and you try to use one it in your function, R may use the wrong one. In these contexts, using the `package::` prefix is crucial.

To finish creating the `river_medians()` documentation, run

```
devtools::document()
```

The `@importFrom` and `@import` tags affect the `NAMESPACE` file, which we get to later. This file is updated by running `devtools::document()`. Note that if you want to use `river_medians()` interactively, make sure to run `devtools::load_all()` again (for our example, this is not *strictly necessary* because `stats` comes loaded alongside R, but it is a good habit).

```
devtools::load_all()
```

### 1.2.8.1 Exercises

These exercises are challenging, so if they don't make sense now, that is okay! Make sure to re-review the solutions after the workshop.

1. Write a new function, `river_stats()`, that takes a data frame, `pattern`, and a general function by which to summarize river length and discharge. This general function should not be an actual function but rather a placeholder for a function a user inserts themselves.
2. Rewrite `river_stats()` from the previous exercise so that `river_stats()` also takes additional arguments to the summarizing function (hint: use

... as an argument)

### 1.2.9 DESCRIPTION

The `DESCRIPTION` file contains the R package's metadata (and was automatically created alongside the R package). The `DESCRIPTION` file lets you easily track version numbers, authorship, and additional R packages that your R package uses. There are two fields in `DESCRIPTION` that do most of the communication regarding how your R package uses additional R packages:

1. **Imports:** Packages here must be installed in order for your package to work. As a result, any package listed in `Imports` will be installed alongside your package. Packages in `Imports` help build the foundation of your package.
2. **Suggests:** Packages here enhance your package but are not required for your package to work. You might use suggested packages for enhanced plotting, additional data sets, or more. Packages in `Suggests` can add finishing touches to your package, but they are not part of your package's foundation – they are like a home's paint color, not its framing.

Other fields used to communicate how your R package uses additional R packages are `Depends`, `LinkingTo`, and `Enhances`. The difference between `Depends` and `Imports` is subtle – the general advice is to use `Imports` instead of `Depends`. For more information about these fields, see [Package Dependencies](#) (external link).

Editing the package, title, author, and description fields of this package yield a `DESCRIPTION` file that looks like

```
Package: RPRW
Title: R Packages for Reproducible Workflows
Version: 0.0.0.9000
Authors@R:
  person(given = "Michael",
         family = "Dumelle",
         role = c("aut", "cre"),
         email = "Dumelle.Michael@epa.gov")
Description: A companion R package for "Using R Packages for Reproducible Workflows"
  at the 2021 EPA R Workshop.
License: `use_mit_license()`, `use_gpl3_license()` or friends to pick a
  license
Encoding: UTF-8
LazyData: true
Roxygen: list(markdown = TRUE)
RoxygenNote: 7.1.1
Depends:
  R (>= 2.10)
```

```
Imports:
  stats
```

We will talk about some of the remaining parts of DESCRIPTION later.

### 1.2.10 NAMESPACE

While the DESCRIPTION file communicates which other packages your package *uses*, the NAMESPACE file communicates *how* your package uses these packages. More specifically, the NAMESPACE file controls which functions your package exports (making them available to others when your package is installed) and which packages from which functions must be available for your exported functions to work. The NAMESPACE file is automatically updated while running `devtools::document()` and should not be edited by hand (in fact, there is a warning in the NAMESPACE file to warn against editing the file by hand). This illustrates another advantage of using roxygen comments and `devtools::document`: documentation for a function from the .R file is used to generate .Rd documentation and the NAMESPACE imports and exports, so they all remain consistent.

In `river_median()`, if you called `median` using `stats::median`, your NAMESPACE file will look like

```
# Generated by roxygen2: do not edit by hand

export(river_means)
export(river_medians)
```

If you used the `@importFrom stats median` approach, your NAMESPACE will look like

```
# Generated by roxygen2: do not edit by hand

export(river_means)
export(river_medians)
importFrom(stats,median)
```

If you used the `@import stats` approach, your NAMESPACE will look like

```
# Generated by roxygen2: do not edit by hand

export(river_means)
export(river_medians)
import(stats)
```

### 1.2.11 testthat

Testing your code to make sure that it performs as intended is an important step in the package building process. Though this may seem like extra work upfront,

implementing a rigorous testing procedure for your package is essential for its continued success. In R, testing is incorporated into your package through the `testthat` (external link) package. To begin using `testthat`, run

```
usethis::use_testthat()
```

```
✓ Adding 'testthat' to Suggests field in DESCRIPTION
✓ Setting Config/testthat/edition field in DESCRIPTION to '3'
✓ Creating 'tests/testthat/'
✓ Writing 'tests/testthat.R'
```

The root of your package directory should look like

	..		
<input type="checkbox"/>	.gitignore	49 B	Sep 14, 2021, 8:47 AM
<input type="checkbox"/>	.Rbuildignore	55 B	Sep 17, 2021, 7:41 AM
<input type="checkbox"/>	.Rhistory	16 KB	Sep 17, 2021, 8:29 AM
<input type="checkbox"/>	data		
<input type="checkbox"/>	DESCRIPTION	608 B	Sep 15, 2021, 6:03 PM
<input type="checkbox"/>	inst		
<input type="checkbox"/>	man		
<input type="checkbox"/>	NAMESPACE	122 B	Sep 18, 2021, 8:57 AM
<input type="checkbox"/>	R		
<input type="checkbox"/>	RPRW.Rproj	436 B	Sep 18, 2021, 7:11 AM
<input type="checkbox"/>	tests		
<input type="checkbox"/>	vignettes		

The tests folder should look like

	..		
<input type="checkbox"/>	testthat		
<input type="checkbox"/>	testthat.R	52 B	Sep 13, 2021, 10:18 AM

Tests are generally written on a function-by-function basis. Tests for a function should be contained in an R script titled `test-function_name`. Luckily, `usethis::use_test()` creates an R script for you with the proper name! To start setting up tests for `river_means()`, run

```
usethis::use_test("river_means")
```

```
✓ Writing 'tests/testthat/test-river_means.R'
* Modify 'tests/testthat/test-river_means.R'
```

Your `testthat` folder should look like

	..		
<input type="checkbox"/>	test-river_means.R	838 B	Sep 13, 2021, 10:43 AM



Tests can be fairly detailed and cover many components of a function (such as input types, output types, function output, etc.). The tests in `testthat` are prefixed with `expect_`. Here we write a simple test that calculates whether our function, `river_means()`, yields output that we would expect if we calculated the means “by hand”. Copy and paste this into `test-river_means.R`:

```
test_that("the mean length is calculated correctly in a test case", {

  # calculate values required for the test for length

  ## calculate the means from the function
  river_means_val <- river_means(river, "Missouri|Mississippi")
  river_means_length <- river_means_val[[1]]

  ## calculate the means "by hand"
  raw_vec_length <- unlist(river["length", c("Missouri", "Mississippi")])
  raw_means_length <- mean(raw_vec_length)

  # perform the actual test for length

  ## check that the function and "by hand" output matches
  expect_equal(river_means_length, raw_means_length)
})
```

You should then see a “Run Tests” button in the editor window. Clicking on that button runs the tests and prints the results in the “Build” tab of the environment window, which should look like

```
==> Testing R file using 'testthat'
```

```
i Loading RPRW
```

```
== Testing test-river_means.R =====
[ FAIL 0 | WARN 0 | SKIP 0 | PASS 1 ] Done!
```

```
Test complete
```

If you have many tests, the `devtools::test()` function runs all of the tests in the `testthat` folder:

```
devtools::test()
```

```
i Loading RPRW
```

```
i Testing RPRW
```

```
✓ | OK F W S | Context
```

```
✓ | 1         | river_means
```

```
== Results =====
```

```
[ FAIL 0 | WARN 0 | SKIP 0 | PASS 1 ]
```

We can write a similar test for `discharge` and then repeat the process for `river_meadians()`. Then `devtools::test()` returns

```
devtools::test()
```

```
i Loading RPRW
i Testing RPRW
√ | OK F W S | Context
√ | 2         | river_means
√ | 2         | river_medians
```

```
== Results =====
[ FAIL 0 | WARN 0 | SKIP 0 | PASS 4 ]
```

All tests pass – hooray! I want to again emphasize how important testing is. I know that it seems like an extra chore, but in my experience, writing careful tests has always paid off...with interest.

#### 1.2.11.1 Exercises

1. Write similar tests for `river_means()` (`discharge`), `river_medians()` (`length`), and `river_medians()` (`discharge`).
2. Check out the `covr` ([external link](#)) package, which quantifies how much of your package's code is used by your tests.

#### 1.2.12 License

At some point, your package needs a license. The license places restrictions on how your package can be shared with others. Licensing can be rather technical and complicated, so I refer you [here](#) for more information. For illustration purposes, we will use a GPL-3 license for this package:

```
usethis::use_gpl_license()
```

```
√ Setting License field in DESCRIPTION to 'GPL (>= 3)'
√ Writing 'LICENSE.md'
√ Adding '^LICENSE\\.md$' to '.Rbuildignore'
```

`usethis::use_gpl_license()` adds the license to `DESCRIPTION` and creates a `LICENSE.md` file at the root of your package that contains information about the license. Approved licenses for EPA repositories include CC0, MIT, GPL-3, and Apache 2.

#### 1.2.13 Vignettes

Vignettes act as high-level user guides for your package. They are the glue that binds together several components of your package to illustrate how the package

can be used to solve a particular problem. Vignettes tend to be R Markdown documents. Though we cover R Markdown documents in more detail later, for now we note that they interweave text and R code to create reproducible documents. You can recognize whether a document is an R markdown document by the `.Rmd` extension that R documents have.

The `ggplot2` (external link) package is a popular package for visualizing data. After installing `ggplot2` (you can check to see if it is already installed by running `"ggplot2" %in% installed.packages()`), you can view its available vignettes by running

```
vignette(package = "ggplot2")
```

after which a file will pop up alongside your R scripts with the contents

```
Vignettes in package 'ggplot2':

ggplot2-specs                Aesthetic specifications (source, html)
extending-ggplot2            Extending ggplot2 (source, html)
ggplot2-in-packages          Using ggplot2 in packages (source, html)
```

To then view a specific vignette, run `vignette("topic", "package")`. For example, to view the vignette regarding aesthetic specifications, run

```
vignette("ggplot2-specs", "ggplot2")
```

The vignette can then be viewed in the “Help” tab of the bottom-right pane of RStudio. Vignettes are also available on a package’s CRAN page – for the `ggplot2` aesthetic specifications vignette viewable on its CRAN page, see here (external link).

Suppose we want to include a vignette in RPRW that shows how to compute river statistics. Running

```
use_vignette("river-statistics", "River Statistics")
```

```
✓ Adding 'knitr' to Suggests field in DESCRIPTION
✓ Setting VignetteBuilder field in DESCRIPTION to 'knitr'
✓ Adding 'inst/doc' to '.gitignore'
✓ Creating 'vignettes/'
✓ Adding '*.html', '*.R' to 'vignettes/.gitignore'
✓ Adding 'rmarkdown' to Suggests field in DESCRIPTION
✓ Writing 'vignettes/river-statistics.Rmd'
* Modify 'vignettes/river-statistics.Rmd'
```

does a few things. It alters your `DESCRIPTION` file, most notably adding the `knitr` (external link) and `rmarkdown` packages to the `Suggests` field. It also creates a vignette folder at the root of your package and adds to that folder an `rmarkdown` vignette named `river-statistics.Rmd` (notice the `.Rmd` extension) whose title is “River Statistics”.

While we won't create a vignette for our package during the workshop, I have added a vignette to the RPRW package. To view it (after installation, which we discuss later), run

```
vignette("river-statistics", "RPRW")
```

#### 1.2.14 devtools::check()

At this point we have built R package and are ready to share it with the world! But we should probably check to make sure we did not make any small mistakes. Does devtools have a way to do this? Of course – introducing `devtools::check()`! `devtools::check()` evaluates several automated tests on your package to make sure it can be properly installed and shared – this can take a few minutes. The `devtools::check()` output lists errors, warnings, and notes associated with your package. Though the warnings and notes are important, it is most crucial to address the errors immediately.

Hopefully after running `devtools::check()`, your output looks like

```
-- R CMD check results ----- RPRW 0.0.0.9000 ----
Duration: 34.4s
```

```
0 errors ✓ | 0 warnings ✓ | 0 notes ✓
```

If not, do your best to fix the errors and address the warnings and notes!

#### 1.2.15 devtools::install()

After `devtools::check()` identifies zero errors (and hopefully zero warnings and zero notes), you can install your package by running

```
devtools::install()
```

After installing RPRW, you can use `library(RPRW)` to load it as you do for any other R package. To uninstall, you can run

```
remove.packages("RPRW")
```

##### 1.2.15.1 Exercises

1. Find the argument to `devtools::install()` that identifies whether vignettes are built alongside installation. What is the default value of this argument?

#### 1.2.16 Congratulations

Congratulations on building an R package! If this was your first one, then an extra special congratulations to you! Hopefully stepping through the process together made it all less intimidating.



### 1.2.17 A Recap

Here is a table that walks through the tools we discussed today, step-by-step (omitting the `devtools::` and `usethis::` prefixes):

Task	Command
Creation	<code>create_package()</code>
Add Function	<code>use_r()</code>
Add Package	<code>use_package()</code>
Load Package	<code>load_all()</code>
Add Data	<code>use_data()</code>
Documentation	<code>document()</code>
Testing (Setup)	<code>use_testthat()</code>
Testing (Create)	<code>use_test()</code>
Testing (Perform)	<code>test()</code>
License (GPL-3)	<code>use_gpl_license()</code>
Vignette	<code>use_vignette()</code>
Check	<code>check()</code>
Install	<code>install()</code>

## 1.3 Debugging

Even the best of programmers write code that may fail in unintended ways. This unintended failing of code is referred to as a “bug,” and the process of fixing the “bug” is known as “debugging”. Don’t expect to always write perfect code – do expect to have the tools necessary to track down bugs and remedy them.

Though we went through a rigorous documentation and testing procedure when creating `river_means()` and `river_medians()`, there are still bugs present in these functions.

### 1.3.1 A Mysterious Error Message

We have already used `river_means()` and `river_medians()` to successfully find means and medians for several patterns. But running `river_means(river, "R")` yields a mystifying error:

```
river_means(river, "R")
```

```
#> Error in rowMeans(new_data): 'x' must be an array of at least two dimensions
```

Something is wrong – and we need to figure out what. A good first step is to copy and paste the error into a Google search engine and see if anyone has solved the problem yet. If you are lucky, this approach will identify the bug for you. If you are not so lucky, you need to try a different approach. Fortunately, R has built-in tools that can help you track down bugs – we will explore a few of these tools next.

### 1.3.2 `traceback()`

The `traceback()` function is generally run in the console after receiving an error. `traceback()` identifies where the error occurred in the previous function call. Here we run `traceback()` after `river_means(river, "R")`:

```
traceback()
```

```
3: stop("'x' must be an array of at least two dimensions")
2: rowMeans(new_data) at #4
1: river_means(river, "R")
```

We now know that the error occurs in Line 4 of `river_means()` (the `rowMeans(new_data)` step). While this is very useful information, we still don't know exactly why the error occurs.

### 1.3.3 `browser()`

In my early stages of debugging, I would save my arguments to a function in the global environment and then run through each line of the function returning the error, effectively “recreating” the environment in which the error occurred. Perhaps some of you have done this too. The bad news is that this approach is often error-prone and inefficient. The good news is that there is a better way – `browser()`. Inserting `browser()` into your code and then running it lets you interactively step into the function, starting at wherever `browser()` is located. No more saving arguments to the global environment and running code line-by-line! Let's try this out with by changing the body of `river_means()` to be:

```

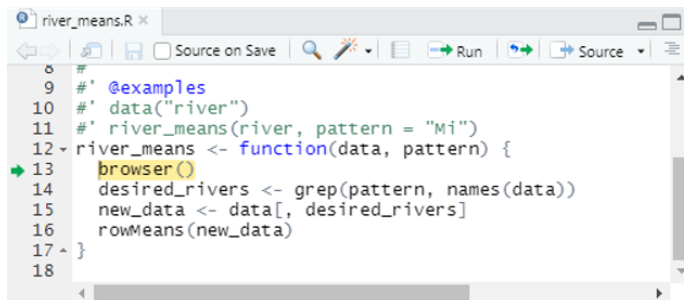
river_means <- function(data, pattern) {
  browser()
  desired_rivers <- grep(pattern, names(data))
  new_data <- data[, desired_rivers]
  rowMeans(new_data)
}

```

After using `devtools::load_all()`, calling `river_means()` will let you step into the function before the first line of code. For example,

```
river_means(river, "R")
```

opens a new file in the editor (upper left-hand) window that looks like



which highlights the most recently executed line in the file, as seen above. The Environment tab shows objects in the function's current environment (in this case, the argument values and any objects the function defines).

You will also see a few buttons in the R console (lower left-hand) window:



These five buttons help you interactively navigate through `river_means()`. From left to right, they are

- Next (n), which executes the next line of the code
- Step (s), which steps into the function called by the current line of code
- Finish (f), which finishes execution of the current function
- Continue (c), which leaves the interactive context and continues execution of the function
- Stop (Q), which leaves the interactive context and terminates execution of the function

Before navigating through any lines of `river_means()`, we see `data` and `pattern` are defined:

```
print(data)
```

```

#>           Missouri Mississippi Yukon Colorado Arkansas Columbia Red Canadian
#> length      3768      3544  3190      2330      2322      2000 1811      1458

```

```
#> discharge      1956      18400  6340      40      1004      7730  852      174
print(pattern)
```

```
#> [1] "R"
```

Pressing **n** (Next) executes `browser()`. Pressing **n** (Next) again executes

```
desired_rivers <- grep(pattern, names(data))
```

Inspecting `desired_rivers`, we see

```
print(desired_rivers)
```

```
#> [1] 7
```

This seems correct, as the seventh river, “Red”, does match the pattern “R”. So the error isn’t here. Let’s press **n** (Next) to evaluate the next line

```
new_data <- data[, desired_rivers]
```

Inspecting `new_data`, we see

```
new_data
```

```
#> [1] 1811 852
```

Well this seems weird – `new_data` does not look like a data frame. Let’s inspect the structure

```
str(new_data)
```

```
#>  num [1:2] 1811 852
```

`new_data` is in fact not a data frame, but rather it is a numeric vector. Because `rowMeans()` requires an array of two or more dimensions (e.g. a matrix or data frame), the next line of code, `rowMeans()`, fails. Pressing **n** again runs

```
rowMeans(new_data)
```

and returns the error while removing you from the interactive mode.

This is a good example of one of the reasons debugging can be so challenging: often, an error occurs because code from a few lines back runs in an unintended way (even though it runs successfully). Note that a similar error occurs when running `river_medians(river, "R")` because `apply()` requires an array of two or more dimensions.

So what is happening here? We know the code works when the pattern yields at least two matches in `river`. But now, with one match, the code is failing. Let’s start with the `[` function called by `data[, desired_rivers]`, where we receive unintended output. Let’s look at the documentation of `[` by running



```
? `[`
```

A close look at the documentation reveals further documentation for the behavior of `[` when applied to a data frame

```
? `[.data.frame`
```

Notice the `drop` argument in this documentation. The `drop` argument, when `TRUE`, coerces the object to the lowest possible dimension. When subsetting a data frame by one column, `drop` is `TRUE` by default and data frame structure is turned into a vector structure, which only has one dimension – this is why the error is occurring. So we need to force our subsetting to always keep the data frame structure by setting `drop = FALSE`. Accommodating this change in `river_means()` and `river_medians()` yields functions that look like

```
river_means <- function(data, pattern) {
  desired_rivers <- grep(pattern, names(data))
  new_data <- data[, desired_rivers, drop = FALSE]
  rowMeans(new_data)
}

river_medians <- function(data, pattern) {
  desired_rivers <- grep(pattern, names(data))
  new_data <- data[, desired_rivers, drop = FALSE]
  apply(new_data, 1, median)
}
```

Now running `river_means()` and `river_medians()` with patterns only having one match behaves as intended

```
river_means(river, "R")
```

```
#>    length discharge
#>    1811          852
```

```
river_medians(river, "R")
```

```
#>    length discharge
#>    1811          852
```

For more information about debugging in R, watch this (external link) and/or read this (external link).

### 1.3.4 Another Error – No Mysterious Message

We previously received and fixed an error message that occurred when the pattern only matched one river. So we know our code works when the pattern matches at least one river. But what happens when the pattern matches zero rivers?

```
river_means(river, "ZZZ")
```

```
#>    length discharge
#>      NaN      NaN
```

The length and discharge means are `NaN`, which stands for “Not a Number” – this is certainly unintended (note that `river_medians(river, "ZZZ")` returns similar output.). These types of bugs are especially pernicious because there is no error message associated with them – we can’t use `traceback()`. Say `river_means()` was a small part of a much larger function, – this bug could be the root of a confusing error several lines later. Whenever function output is unexpected and you can’t immediately figure out why, use `browser()` to diagnose the bug. In this context, `desired_rivers` is a length-zero vector, which causes problems in the remaining parts of the function. To guard against these types of bugs, identify potentially problematic scenarios for your function and program defensively, forcing the function to return an informative error message when a problematic scenario is identified. Here we force `river_means()` to return an error message when the pattern matches zero rivers:

```
river_means <- function(data, pattern) {
  desired_rivers <- grep(pattern, names(data))
  if (length(desired_rivers) == 0) {
    stop("This is an error message that stops the function.")
  }
  new_data <- data[, desired_rivers, drop = FALSE]
  rowMeans(new_data)
}
```

```
river_means(rivers, "ZZZ")
```

```
#> Error in river_means(rivers, "ZZZ"): This is an error message that stops the function
```

#### 1.3.4.1 Exercises

1. Rewrite `river_medians()` so that it stops and returns an informative error message when the pattern matches zero rivers.

## 1.4 Additional Resources

- R Packages (external link) by Hadley Wickham (external link) and Jenny Bryan (external link) [Wickham, 2015]
- Writing an R package from scratch (external link) by Hilary Parker (external link)
- Writing R Extensions (external link) by CRAN (this resource is very technical)

## Chapter 2

# Making an R Package a Research Compendium

### 2.1 What is a Research Compendium?

Marwick et al. [2018] state that the goal of a research compendium is to provide a standard and easily recognizable way to organize the digital materials of a project to enable others to inspect, reproduce, and extend the work. Three generic principles define a research compendium:

1. Files should be organized according to the prevailing conventions of a broader community. This helps members of the community recognize the structure of the project and make tools that utilize the structure.
2. There should be a clear separation of data, methods, and output. For example, raw data should be kept separate from the code that cleans the data so that others can access the raw data.
3. The computational environment for the methods should be clearly specified. At the most basic level, this means recording the names and version numbers of software. At the most detailed level, this means completely reproducing the computing environment.

#### 2.1.1 Exercises

1. What principles of a research compendium seem most beneficial to you?

### 2.2 Why a Research Compendium?

Marwick et al. [2018] (and references therein) give several benefits of a research compendium:

- A convenient way to publicly share data and code
- Work with publicly available data sets may receive higher numbers of citations than work with private data sets (note that publicly available data sets also tend to be easier to clear through EPA platforms)
- Data sharing is associated with higher publication productivity
  - Of 7,040 NSF and NIH awards studied, the median number of publications associated with each research grant was five when the data were private and 10 when the data were public
- Structured and simplified file management and workflows
- More defense against errors
- Easier to communicate with others (and future you)

### 2.2.1 Exercises

1. Can you identify any benefits to a research compendium that are not already on this list?

## 2.3 Why an R Package for a Research Compendium?

As previously mentioned, a research compendium requires a specific structure. Earlier in this workshop, we got some experience with something in R that also requires a specific structure.....R packages! A research compendium can significantly benefit from adopting the structure of an R package. Organization is simpler (as you already have a structure), writing and documenting functions and data helps guard against errors in an analysis, and devtools and usethis have several development tools to help ensure your R package performs as intended.


We have previously discussed how to include data in an R package, but many of you may be wondering how we include additional pieces of a research compendium like analysis scripts, output, or a manuscript itself. If the structure of an R package is so rigid – how do these pieces fit? We discuss one such approach next.

## 2.4 Turning RPRW Into A Research Compendium

If a folder named `inst` is placed at the root of an R package's directory, all of the raw contents in `inst` will be installed upon installation of the R package. This is where we can put the pieces of our research compendium that are not explicitly related to the previous implementation of our R package. We can use this approach to turn RPRW into a research compendium.





Suppose we are using RPRW to supplement a manuscript about rivers. Let's first add the `inst` folder – the root of your package's directory should look

similar to

	..		
<input type="checkbox"/>	 .gitignore	49 B	Sep 14, 2021, 8:47 AM
<input type="checkbox"/>	 .Rbuildignore	44 B	Sep 14, 2021, 8:07 AM
<input type="checkbox"/>	 .Rhistory	38 B	Sep 10, 2021, 9:57 AM
<input type="checkbox"/>	 data		
<input type="checkbox"/>	 DESCRIPTION	608 B	Sep 15, 2021, 6:03 PM
<input type="checkbox"/>	 inst		
<input type="checkbox"/>	 LICENSE.md	34.1 KB	Sep 14, 2021, 8:07 AM
<input type="checkbox"/>	 man		
<input type="checkbox"/>	 NAMESPACE	122 B	Sep 15, 2021, 5:56 PM
<input type="checkbox"/>	 R		
<input type="checkbox"/>	 RPRW.Rproj	436 B	Sep 15, 2021, 6:01 PM
<input type="checkbox"/>	 tests		
<input type="checkbox"/>	 vignettes		

Let's add three folders to `inst`:

1. **analysis** for our analysis scripts – these scripts contain code that is not related to the functions in the package but is related to other parts of the research compendium (e.g. generating output for the manuscript)
2. **output** for our analysis output
3. **manuscript** for our manuscript

	..
<input type="checkbox"/>	 analysis
<input type="checkbox"/>	 manuscript
<input type="checkbox"/>	 output

First we will make an R script to put in `analysis`. Recall that in Exercise 1.2.7.1.2 (internal link), we generated `river_stats()`, whose body looks like

```
river_stats <- function(data, pattern, FUN, ...) {
  desired_rivers <- grep(pattern, names(data))
  new_data <- data[, desired_rivers]
  apply(new_data, 1, FUN, ...)
}
```

`river_stats()` works like `river_means()` and `river_medians()` but takes a general function instead of computing the mean or median specifically. Suppose our goal is to find the minimum length and discharge for rivers whose names contain "C". Our script to put in `analysis` will look like

```
# find some summary statistics on rivers with pattern "C"

## load RPRW -- we need to do this because the script is separate from the
## functions installed in the RPRW package
library(RPRW)
```

```
## find minimums
river_min <- river_stats(river, "C", min)

## make a data frame of the minimums
min_df <- data.frame(
  data = "river",
  pattern = "C",
  length_min = river_min[[1]],
  discharge_min = river_min[[2]]
)

## write out the minimums (without row names)
write.csv(min_df, "inst/output/min_df", row.names = FALSE)
```

This script loads RPRW and finds the minimum length and discharge for rivers whose names start with C. The script then saves the contents to a CSV file in the `output` folder. These contents are used when building the manuscript. The manuscript is in the `manuscript` folder and is a completely reproducible R Markdown document (using an `rticles` template, which we discuss later). Though we won't discuss the details of this R Markdown document here unless we have time, the files can be viewed on your machine (after installing RPRW) at the location provided by

```
system.file("manuscript", package = "RPRW")
```

A compiled PDF of the manuscript is available [here](#) (external link).

### 2.4.1 Exercises

1. There are some folder names within `inst` that should be avoided – what are they?

### 2.4.2 Adding a Citation

Often times your research compendium may be centered around a particular publication. You may build an R package that deals with specific tasks and then write your manuscript using the research compendium structure. In this event, you may want your publication to be easily citable. You can add a citation to your R package research compendium by running

```
use_citation()
```

```
✓ Writing 'inst/CITATION'
* Modify 'inst/CITATION'
```

The citation file is placed in the `inst` folder. It looks a little intimidating at first, but it automatically creates a text version and LaTeX version of your citation

with relatively little ease. Here is raw code I used to create the citation in the RPRW package

```
citHeader("To cite RPRW in publications use:")

citEntry(
  entry    = "Manual",
  title    = "Using R Packages for Reproducible Workflows",
  author   = personList(as.person("Michael Dumelle")),
  journal  = "EPA 2021 R Workshop",
  url      = "https://github.com/michaeldumelle/R-Packages-Reproducible-Workflows-Book",
  textVersion =
  paste(
    "Michael Dumelle.",
    "(2021).",
    "Using R Packages for Reproducible Workflows.",
    "EPA 2021 R Workshop.",
    "URL https://github.com/michaeldumelle/R-Packages-Reproducible-Workflows-Book."
  )
)
```

To view the citation, run

```
citation(package = "RPRW")

#>
#> To cite RPRW in publications use:
#>
#> Michael Dumelle. (2021). Using R Packages for Reproducible Workflows.
#> EPA 2021 R Workshop. URL
#> https://github.com/michaeldumelle/R-Packages-Reproducible-Workflows-Book.
#>
#> A BibTeX entry for LaTeX users is
#>
#> @Manual{,
#>   title = {Using R Packages for Reproducible Workflows},
#>   author = {Michael Dumelle},
#>   journal = {EPA 2021 R Workshop},
#>   url = {https://github.com/michaeldumelle/R-Packages-Reproducible-Workflows-Book},
#> }
```

If left unedited, the package citation defaults to something that looks like

To cite package 'RPRW' in publications use:

```
Michael Dumelle (2021). RPRW: R Packages for Reproducible Workflows.
R package version 0.0.0.9000.
```

A BibTeX entry for LaTeX users is

```
@Manual{,
  title = {RPRW: R Packages for Reproducible Workflows},
  author = {Michael Dumelle},
  year = {2021},
  note = {R package version 0.0.0.9000},
}
```

## 2.5 Research Compendium Recap

The research compendium now visibly contains:

- **data**: a folder containing data
- **DESCRIPTION**: a file containing metadata about our package
- **inst**: a folder containing auxiliary files to be downloaded alongside the package's installation
  - **analysis**: a folder containing R scripts that are separate from the R package but used as part of the research compendium
  - **CITATION**: a file containing citation information
  - **manuscript**: a file containing the manuscript and auxiliary files
  - **output**: a file containing the CSV output
- **LICENSE.md**: a file containing metadata about our license (GPL-3)
- **man**: a folder containing R function documentation from the package
- **NAMESPACE**: a file containing exports and imports
- **R**: a folder containing R functions from the package
- **RPRW.Rproj**: a file containing R Project metadata
- **tests**: a folder containing R function testing
- **vignettes**: a folder containing the package vignette

You can also add a **README.md** file to give a broad overview of your package (**README.md** is useful whether or not you use Git and GitHub). The **README.md** file is a Markdown (external link) file (notice the **.md** extension and notice the connection between R Markdown (**.Rmd**) and Markdown (**.md**)) in the **RPRW** package that looks like

```
# RPRW
```

```
Companion R package for "Using R Packages for Reproducible Workflows" at the 2021 EPA I
```

## 2.6 Sharing Your Research Compendium

So now you have a research compendium that combines your R functions with your data, code, and manuscript to create a reproducible product that is easy to share with your colleagues. But how do we share it? Running



```
devtools::build()
```

will build an R package for you. By default, the package will be located in the same folder as the folder containing your R package with name “package-name\_version-number.tar.gz”. Note that this is one level above the location of your R package’s directory. For example, if folder “A” holds “RPRW”, which is the folder containing all of the files associated with my R package, then using `devtools::build()` will install “RPRW\_0.0.0.9000.tar.gz” in folder “A”. If there are any files in your package folder that you do not present in the installed version of the package, place the file paths in the `.Rbuildignore` file. For more on `.Rbuildignore`, visit [here](#) (external link).

After building the package, you can send it to a colleague, they can save it on their machine, and then they can install it by running

```
install.packages(path_to_package, repos = NULL, type = "source")
```

If your colleague doesn’t remember where they saved the file, they can interactively search for it by running

```
install.packages(file.choose(), repos = NULL, type = "source")
```

Because R packages leverage R projects your colleague can immediately run any of your code on their machine (a reminder that we discuss R projects soon)!

## 2.7 Hosting Your Research Compendium

You can host your research compendium in an open-access repository so that a broader audience may access it. One approach is to host it on GitHub, which we discuss more later. This approach has the advantage of being easy to install by running

```
devtools::install_github("GitHub_user_name/repository_name", "repository_branch")
```

You can also put your research compendium on Zenodo ([external link](#)). Zenodo is an open-access repository designed to permanently archive research compendia. GitHub and Zenodo compliment each other nicely: GitHub acts as the “easy to access” version of the research compendium while Zenodo is the more permanent version. For an example of a research compendium that combines these approaches, visit the GitHub repository [here](#) (external link) and Zenodo archive [here](#) (external link).

### 2.7.1 Exercises

1. Run `?devtools::build` to look at some additional arguments – which seem useful to you?

## 2.8 An Example R Package Reserach Compendium

A recent publication of mine used this R package research compendium structure – you can view the GitHub repository [here](#) (external link).

## Chapter 3

# Extensions

### 3.1 R Projects

When I first started coding in R, I would try to share my code with others and encountered a problem – my colleagues would copy, paste, and run my code but it would not work on their machine. Often this was because my scripts read in files from different file paths (locations) on my computer – and these file paths were unique to my machine!. To get my code to work, my colleagues would have to edit my code that used file paths so that they pointed to the correct files on their machine. Not only is this a lot to ask of someone, it also makes it incredibly error-prone. More often than not, I would have to meet with someone to get the code I sent them up and running. This was a waste of time for everyone involved.

R looks for files to load in a working directory. This is helpful because looking for files in a working directory prevents you from having to type out the full path to a file every time you want to load something. For example, suppose you are not using a working directory and you have several files to load that live in `a/long/path/with/s p a c e s/and/cH-a_r--aCt-er$!!!`. To load `my_cool_file.csv` and `my_other_cool_file.csv`, you have to run

```
read.csv("a/long/path/with/s p a c e s/and/cH-a_r--aCt-er$!!!/my_cool_file.csv")
```

and then retype (or copy/paste) the path to run

```
read.csv("a/long/path/with/s p a c e s/and/cH-a_r--aCt-er$!!!/my_other_cool_file.csv")
```

This workflow is tedious and error-prone. If you set your working directory to `a/really/long/path/with/s p a c e s/and/cH-a_r--aCt-er$!!!`, then to load `my_cool_file.csv` and `my_other_cool_file.csv`, you have to run

```
read.csv("my_cool_file.csv")
read.csv("my_other_cool_file.csv")
```

You can set working directories in R using `setwd()`, but this isn't really a good idea for various reasons (you get to discuss them in the exercises). There is a better way to control your working directory – R projects!

R projects automatically set your working directory to live wherever the project is stored on your machine. So if I have `my_cool_file.csv` and `my_other_cool_file.csv` in the root (top-most level) of my R project, I can load them by running

```
read.csv("my_cool_file.csv")
read.csv("my_other_cool_file.csv")
```

while inside the R project. This is so powerful because that means I can bundle up my R project, send it to my colleague, and then the R project will set my colleague's working directory to live wherever they save the project. That means that they can load `my_cool_file.csv` and `my_other_cool_file.csv` using the exact same code I used. No more file path problems – this is a vastly improved workflow. It may not seem like a big deal now, but learning how to use R projects makes sharing your work substantially easier.

A side benefit of R projects is that they provide a convenient structure by which to organize all the files associated with a particular task. To learn more about creating and maintaining R packages, visit [here](#) (external link) and [here](#) (external link). If you do not use R projects yet, I recommend that you start.

### 3.1.1 Exercises

1. What are some drawbacks of running `setwd()` whenever you want to set a working directory?
2. Look into the [here](#) (external link) package, designed to help solve working directory problems (that can even occur within R Projects).

### 3.1.2 renv

The [renv](#) (external link) package brings R package version management to your projects. `renv` prevents code in your R project from returning separate results based on the version of the R packages installed. It works by saving the specific versions of the R packages in your project and then makes it easy for others to install those same versions before running code in the project.

## 3.2 R Markdown

R Markdown documents provide a convenient way to combine text, R code, and results into a fully reproducible document that compiles (knits) to one of several output types (HTML, PDF, word, slide decks, etc.). An R Markdown document has file extension `.Rmd` and can be created in RStudio by clicking “File -> New File -> R Markdown”, which brings up the following options

The screenshot shows the 'New R Markdown' dialog box. On the left, a sidebar lists four options: 'Document' (with a document icon), 'Presentation' (with a presentation icon), 'Shiny' (with the R logo), and 'From Template' (with a folder icon). 'Document' is selected. The main panel has a 'Title' field containing 'Untitled' and an 'Author' field containing 'Michael Dumelle'. Below these is the 'Default Output Format' section. It has three radio buttons: 'HTML' (selected), 'PDF', and 'Word'. Under 'HTML' is the text: 'Recommended format for authoring (you can switch to PDF or Word output anytime)'. Under 'PDF' is the text: 'PDF output requires TeX (MiKTeX on Windows, MacTeX 2013+ on OS X, TeX Live 2013+ on Linux)'. Under 'Word' is the text: 'Previewing Word documents requires an installation of MS Word (or Libre/Open Office on Linux)'. At the bottom of the dialog are three buttons: 'Create Empty Document', 'OK', and 'Cancel'.

Pressing OK automatically opens a file with several contents

```
---  
title: "Untitled"  
author: "Michael Dumelle"  
date: "9/17/2021"  
output: html_document  
---  
  
```${r setup, include=FALSE}  
knitr::opts_chunk$set(echo = TRUE)  
```
```

```
## R Markdown
```

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML documents.

When you click the **Knit** button a document will be generated that includes both content and a graph of the output of this document.

```
```{r cars}
summary(cars)
```
```

```
## Including Plots
```

You can also embed plots, for example:

```
```{r pressure, echo=FALSE}
plot(pressure)
```
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R output.

- The part surrounded by `---` is the YAML header, which contains meta-data about your document – big picture options are controlled here. For more on YAML headers, check out the `yaml` package and its associated vignettes [here](#) (external link).
- The parts prefixed by `#` are first-level (section-level) headers, and the parts prefixed by `##` are second level (subsection-level) headers.
- The parts surrounded by ````` are code chunks – these are the engine that powers R Markdown. Code chunks let you run and display R code and output in your document. More generally, code chunks look like

```
```{r label, chunk_options}
R code
```
```

- The line

```
knitr::opts_chunk$set()
```

lets you set default options for your code chunks. For example, `knitr::opts_chunk$set(echo = TRUE)` sets `echo = TRUE` for all all code chunks, unless a particular code chunk sets `echo = FALSE`. More information on available code chunk options is available [here](#) (external link).

- The rest of the document contains the body of the document – the plain text.

For more on the structure of R Markdown documents, visit [here](#) (external link).

R Markdown documents are useful for a variety of reasons, two of which we will focus on next:

1. R Markdown lets you create fully reproducible documents by combining R code and text (these are often called **dynamic** documents). What if there is a slight change to the data? No problem, the tables or figures are reproduced when re-knitting. What if you want to use a new output style (like a PDF or Word Document)? No problem, just change `output:` and re-knit.
2. This approach is less error-prone and usually less work compared to the alternative approach of remaking tables, figures, etc. “by hand” and the inserting them manually into your document..

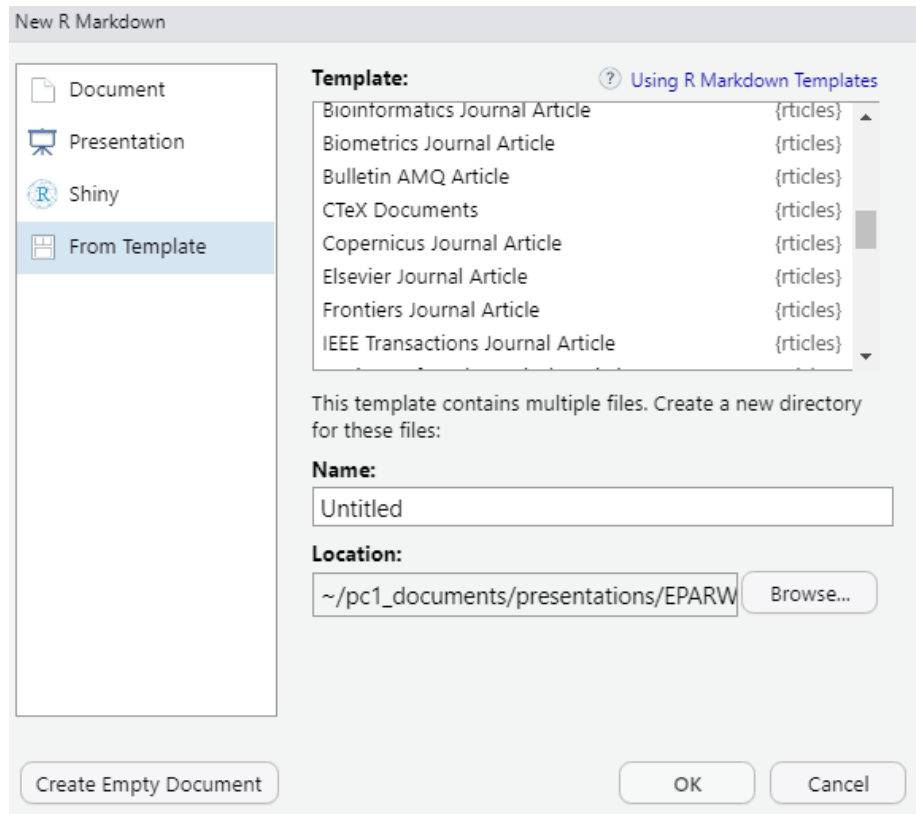
For much, much more on R Markdown, read [this \(external link\)](#) and [this \(external link\)](#). I highly recommend you get some experience with it, as its tools are immensely powerful.

### 3.2.1 Exercises

1. What benefits do you see to R Markdown documents?
2. Save the basic R Markdown template we just discussed and knit it to see its contents.

## 3.3 *rticles*

The *rticles* package is an extension of R Markdown designed to make adhering to journal style requirements easier. Some journal types *rticles* accommodates are Elsevier, PLOS, Sage, Springer, and Taylor & Francis, among others. After installing *rticles* (`install.packages("rticles")`), you can access article templates by starting a new R Markdown document and selecting the template tab



The rcticles templates have complicated YAML headers, but they generally have pretty clear inline instructions for how to edit them. The YAML headers change between templates, but the body of the document is just standard R Markdown (potentially with some LaTeX). This is very convenient because the body of the document is the same across templates (potentially with minor LaTeX discrepancies that are generally explained in the template's inline instructions). Decide you want to switch from an Elsevier journal to a Taylor & Francis journal? No problem, just change the template type! For this reason, rcticles is one my favorite R packages.

For a brief introduction LaTeX, see this (external link). While some working knowledge of LaTeX is not necessarily required to successfully use rcticles, it can be immensely helpful.

### 3.3.1 Exercises

1. Visit this (external link) to see all the journals that rcticles has a template for.



## 3.4 Git and GitHub

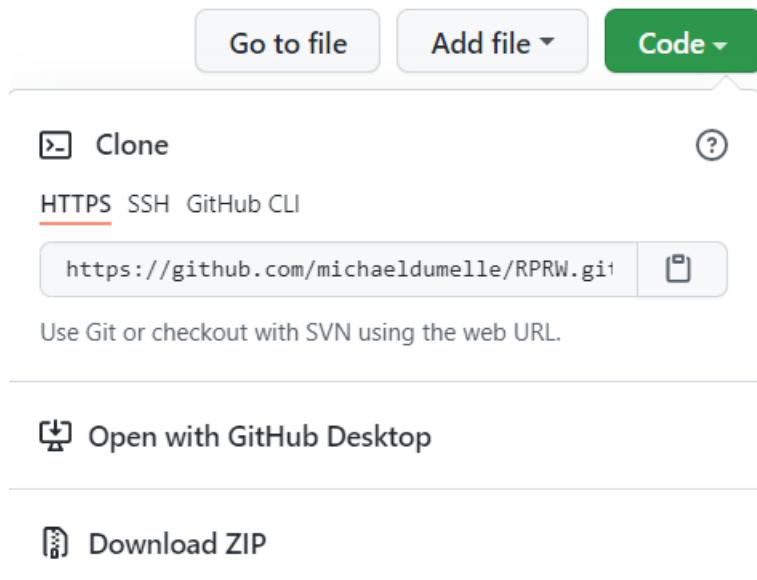
I highly recommend you learn how to use Git and GitHub. Though intimidating, frustrating, and time-consuming at first, Git and GitHub are well worth the effort. Git ([external link](#)) is a version control system. A version control system keeps track of all changes made to your code in a Git repository (which is a `.git/` folder somewhere in your code – note the `.git/` folder is usually hidden). Git is useful on its own, but it benefits greatly from integration with GitHub. GitHub ([external link](#)) is a website that lets you easily share your code, collaborate with others, track changes to your code, and backup your code, among other features. Git and GitHub is by far the most popular version control system for R packages, and access to development versions of R packages is usually only available through GitHub.

```
devtools::install_github("username/packageName")
```

Though we won't get into details of Git and GitHub here (aside from a quick example we discuss next), luckily RStudio has many tools to enable R to communicate with Git and GitHub. For an intro to Git and GitHub through RStudio, visit this ([external link](#)). For a thorough book about using Git and GitHub through RStudio, visit this ([external link](#)). For a video overview of using Git and GitHub through RStudio, watch this ([external link](#)).

### 3.4.1 Cloning a GitHub Repository in RStudio

After ensuring GitHub and R Studio can talk to one another, one GitHub and RStudio workflow is to create a repository on GitHub and then clone it using RStudio. After creating the repository on GitHub, copy the HTTPS URL:



Then go to “File -> New Project -> Version Control -> Git”. Copying and pasting the HTTPS URL in the “Repository URL:” field and pressing “Create Project” clones the repository and creates an R project in a folder with name given by the “Project directory name:” field (which defaults to the repository name) located at the “Create project as subdirectory of:” field. You can then use `usethis::create_package()` to turn the repository into an R package (note that `usethis::create_package(".")` creates a package in the current working directory).

### 3.4.2 .gitignore

The `.gitignore` file in your package files deserves some special attention. Similar to `.Rbuildignore`, `.gitignore` contains file paths that are stored locally but not meant to be put on GitHub. For more on `.gitignore`, visit [here](#) (external link).

## 3.5 Continuous Integration

Continuous integration (often abbreviated CI) is the process of performing automated checks on code anytime it is updated. This may seem cumbersome, but it is a really good idea because it forces you to check for errors with every new version of code. One option for continuous integration in your R package’s GitHub repository is GitHub Actions ([external link](#)). `devtools` has several tools to help you set up GitHub actions. I recommend using their “standard” GitHub actions template, which can be added to your R project by running

```
use_github_action(name = "check-standard")
```

```
✓ Creating '.github/'
✓ Adding '^\\.github$' to '.Rbuildignore'
✓ Adding '*.html' to '.github/.gitignore'
✓ Creating '.github/workflows/'
✓ Writing '.github/workflows/check-standard.yaml'
* Learn more at <https://github.com/r-lib/actions/blob/master/examples/README.md>
```

You will notice that this code created a new folder at the root of our package named `.github`. The `.github` folder may be hidden so you may need to enable the viewing of hidden files to look through its contents on your machine (in the files pane of R studio click more and check “show hidden files”. Then every time you push to GitHub, GitHub actions will simulate `devtools::check()` on several operating systems (Windows, Mac, Linux) and return the results. To view this approach implemented for the RPRW package, visit this (external link). Another option for continuous integration is Travis CI (external link).

## 3.6 Reproducible Examples

If you want to solicit help regarding your code from others, it is helpful if your code is a reprex. A reprex is a reproducible example. Creating a reprex gives others the ability to look at and run your code with the intent of solving the problem you have. Conversations about code are more productive with

1. code that actually runs (the code should work on your machine as intended)
2. code that I don't have to run (I should be able to see your output without actually having to run the code)
3. code that I can easily run (I should be able to run the code on my machine and obtain the same results as you) To accomplish all three above, the code should
  - Be run in a new R session
  - Load all necessary packages
  - Create all necessary objects
  - Be self-contained (don't use anything from outside the reprex)
  - Contain the code as well as the output
  - Run properly on both our machines

How do you ensure your example is a reprex (satisfies the above properties)? Use the `reprex` package! See here (external link) for more information about `reprex` and here (external link) and for some do's and don'ts of a reprex. Finally, see here (external link) for a video introduction to `reprex` with Jenny Bryan, the `reprex` maintainer.

### 3.7 Code Style

The tidyverse style guide ([external link](#)) provides a set of code-writing guidelines that help make your code as easy to read as possible. The tidyverse style guide claims “Good coding style is like correct punctuation: you can manage without it, but it sure makes things easier to read.”

The styler ([external link](#)) package lets you restyle text, files, or entire projects. It includes an RStudio add in and is integrated with usethis. For example,

```
usethis::use_tidy_style()
```

will automatically restyle all functions in your package’s R folder to conform to the tidyverse style guide.

The lintr ([external link](#)) package performs automated checks to help you conform to a style guide. The default style guide is the tidyverse style guide.

### 3.8 Additional Tips

- RStudio has cheat sheets for many commonly used packages available here ([external link](#))
- To view all R Studio keyboard shortcuts, go to “Tools (editor/upper-left window) -> Keyboard Shortcuts Help” or press **Alt + Shift + K**
- The most recent R Studio version (1.4) is a large update with several wonderful new features – if you have not requested it yet, I recommend you do so. Two of these features are Visual R Markdown ([external link](#)) and the command palette ([external link](#)).
- I recommend you try and use the tidyverse ([external links](#)) for data analyses. The tidyverse is a collection of R packages designed to make data analyses easier to code and easier to comprehend. To learn more, read this ([external link](#)) cover-to-cover – it is one of my favorite books.
- If you are building an R package you plan to submit to CRAN (i.e. the primary purpose of the R package is for others to download and use it – the primary purpose is NOT to be a research compendium), I recommend you try to only import absolutely necessary packages. The more packages your package imports, the higher the risk your package breaks when a package you depend on introduces an update / bug / etc. For example, if plan to use the purrr package (which is part of the tidyverse) for the `purrr::map()` function, consider using `lapply()` from the base package instead. If there is a base-R equivalent of a function you want in your package, use the base-R version, even if it means you have to relearn the base-R version
  - When I say base-R here, I mean any of the packages that come installed alongside R: the packages base, datasets, graphics, grDevices, methods, stats, and utils.
- Visit this ([external link](#)) for more R tips and tricks

- Sign up for R Weekly (external link) to get weekly R updates.
- You don't have to learn how to do all of this all at once! Move at a pace you are comfortable with.
- Everyone has their own coding style, and my recommendations may differ from your coding preferences – that is okay! All I hope is that you consider them.

### 3.8.1 Exercises

1. A data set is called “tidy” if it has three qualities: 1) each variable has its own column, 2) each observation has its own row, and 3) each value has its own cell. Is `river` tidy? Why or why not?
2. Make `river` tidy.

## 3.9 Thank you!

Thank you so much for attending this workshop! I do genuinely appreciate you taking some time out of your day to take this journey with me, and I hope that you got something useful out of this workshop.

A stylized, handwritten-style orange graphic that says "Thank you!". The word "Thank" is on the top line and "you!" is on the bottom line. There are decorative orange lines and dots around the text, particularly around the exclamation mark and the bottom of the word "you".



## Chapter 4

# Exercise Solutions

### Building an R Package

**Exercise 1.1.1.1** (internal link) What are some of your favorite R packages?

Answer 1.1.1.1 This is for you to answer! I have many favorites, but a few are devtools, purrr, rlang, rticles, styler, and usethis.

**Exercise 1.1.1.2** (internal link) Of those we have talked about so far, what benefits of R packages are most appealing to you?

Answer 1.1.1.2 This is for you to answer! One of the the most appealing benefits to me is the structure that helps keep me organized.

**Exercise 1.2.4.1.1** (internal link) How many functions does `river_means()` call within the body of the function? What package are these functions in?

Answer 1.2.4.1.1 There are five functions: `<-`, `grep()`, `names()`, `[`, and `rowMeans()`. They are all in the “base” package – this fact be can seen in each function’s documentation

```
?`<`  
?grep  
?names  
?`[`  
?rowMeans
```

**Exercise 1.2.4.1.2** (internal link) Try using `river_means()` with a few different patterns. Do you notice anything strange?

Answer 1.2.4.1.1 Strange behavior occurs when the pattern matches only one river or zero rivers. If the pattern matches only one river, there is an error:

```
river_means(river, "R")
```

```
#> Error in rowMeans(new_data): 'x' must be an array of at least two dimensions
```

If the pattern matches zero rivers, the mean is NaN (which stands for not a number)

```
river_means(river, "ZZZ")
```

```
#>    length discharge
#>      NaN         NaN
```

These are both unintended and are discussed more in [Debugging](#) (internal link).

**Exercise 1.2.7.1.1** (internal link) Write a new function, `river_stats()`, that takes a data frame, pattern, and a general function by which to summarize river length and discharge. This general function should not be an actual function but rather a placeholder for a function a user inserts themselves.

Answer 1.2.7.1.1

```
#' Summary statistics of river lengths and discharges
#' 
#' @param data A data frame with two rows. The first row indicates river length and
#'   second row indicates river discharge. The columns of data indicate river names.
#' @param pattern A pattern by which to include only particular rivers
#' @param FUN A function to summarize the rivers
#' 
#' @return The summarized river length and mean river discharge for the desired rivers
#' @export
#' 
#' @examples
#' data("river")
#' river_stat(river, "Mi", min)
river_stats <- function(data, pattern, FUN) {
  desired_rivers <- grep(pattern, names(data))
  new_data <- data[, desired_rivers]
  apply(new_data, 1, FUN)
}
river_stats(river, "Mi", min)
```

```
#>    length discharge
#>      3544         1956
river_stats(river, "Mi", max)
```

```
#>    length discharge
#>      3768         18400
```



```

river_stats(river, "Mi", mean)

#>   length discharge
#>    3656      10178

river_stats(river, "Mi", stats::median)

#>   length discharge
#>    3656      10178

```

**Exercise 1.2.7.1.2** [internal link](#)) Rewrite `river_stats()` from the previous exercise so that `river_stats()` also takes additional arguments to the summarizing function (hint: use `...` as an argument)

Answer 1.2.7.1.2

```

' Summary statistics of river lengths and discharges
'
' @param data A data frame with two rows. The first row indicates river length and
'   second row indicates river discharge. The columns of data indicate river names.
' @param pattern A pattern by which to include only particular rivers
' @param FUN A function to summarize the rivers
' @param ... Additional arguments to pass to \code{FUN}
'
' @return The summarized river length and mean river discharge for the desired rivers
' @export
'
' @examples
' data("river")
' river_stat(river, "Mi|C", mean, trim = 0.5)
river_stats <- function(data, pattern, FUN, ...) {
  desired_rivers <- grep(pattern, names(data))
  new_data <- data[, desired_rivers]
  apply(new_data, 1, FUN, ...)
}
river_stats(river, "Mi|C", mean, trim = 0)

#>   length discharge
#>    2620      5660

river_stats(river, "Mi|C", mean, trim = 0.5)

#>   length discharge
#>    2330      1956

```

**Exercise 1.2.10.1** [internal link](#)) Write similar tests for `river_means()` (discharge), `river_medians()` (length), and `river_medians()` (discharge).

Answer 1.2.10.1

```
testthat::test_that("the mean discharge is calculated correctly in a test case", {

  # calculate values required for the test for discharge

  ## calculate the means from the function
  river_means_val <- river_means(river, "Missouri|Mississippi")
  river_means_discharge <- river_means_val[[2]]

  ## calculate the means "by hand"
  raw_vec_discharge <- unlist(river["discharge", c("Missouri", "Mississippi")])
  raw_means_discharge <- mean(raw_vec_discharge)

  # perform the actual test for discharge

  ## check that the function and "by hand" output matches
  expect_equal(river_means_discharge, raw_means_discharge)
})
```

```
#> -- Error (<text>:16:3): the mean discharge is calculated correctly in a test case --
#> Error: could not find function "expect_equal"
```

```
testthat::test_that("the median length is calculated correctly in a test case", {

  # calculate values required for the test for length

  ## calculate the medians from the function
  river_medians_val <- river_medians(river, "Missouri|Mississippi")
  river_medians_length <- river_medians_val[[1]]

  ## calculate the medians "by hand"
  raw_vec_length <- unlist(river["length", c("Missouri", "Mississippi")])
  raw_medians_length <- median(raw_vec_length)

  # perform the actual test for length

  ## check that the function and "by hand" output matches
  expect_equal(river_medians_length, raw_medians_length)
})
```

```
#> -- Error (<text>:16:3): the median length is calculated correctly in a test case --
#> Error: could not find function "expect_equal"
```

```
testthat::test_that("the mean discharge is calculated correctly in a test case", {

  # calculate values required for the test for discharge

  ## calculate the medians from the function
```

```

river_medians_val <- river_medians(river, "Missouri|Mississippi")
river_medians_discharge <- river_medians_val[[2]]

## calculate the medians "by hand"
raw_vec_discharge <- unlist(river["discharge", c("Missouri", "Mississippi")])
raw_medians_discharge <- median(raw_vec_discharge)

# perform the actual test for discharge

## check that the function and "by hand" output matches
expect_equal(river_medians_discharge, raw_medians_discharge)
})

```

```

#> -- Error (<text>:16:3): the mean discharge is calculated correctly in a test case --
#> Error: could not find function "expect_equal"

```

**Exercise 1.2.10.2** (internal link) Check out the `covr` (external link) package, which quantifies how much of your package's code is used by your tests.

Answer 1.2.10.2 NA

**Exercise 1.2.14.1.1** (internal link) Find the argument to `devtools::install()` that identifies whether vignettes are built alongside installation. What is the default value of this argument?

Answer 1.2.14.1.1 Running

```
?devtools_install
```

we see that `build_vignettes` controls whether vignettes are build alongside installation. The default argument for `build_vignettes` is `FALSE`.

**Exercise 1.3.4.1.1** (internal link) Rewrite `river_medians()` so that it stops and returns an informative error message when the pattern matches zero rivers.

Answer 1.3.4.1.1

```

river_medians <- function(data, pattern) {
  desired_rivers <- grep(pattern, names(data))
  if (length(desired_rivers) == 0) {
    stop("The pattern provided does not match any rivers in the data provided.")
  }
  new_data <- data[, desired_rivers, drop = FALSE]
  rowMeans(new_data)
}
river_medians(rivers, "ZZZ")

```

```
#> Error in river_medians(rivers, "ZZZ"): The pattern provided does not match any rivers in the d
```

The error messages in the RPRW package are more informative but more difficult

to code.

## Making an R Package a Research Compendium

**Exercise 2.1.1.1** (internal link) What principles of a research compendium seem most beneficial to you?

Answer 2.1.1.1 This is for you to answer! To me, the file organization, file separation, and computational details all seem like wonderful principles!

**Exercise 2.2.1.1** (internal link) Can you identify any benefits to a research compendium that are not already on this list?

Answer 2.2.1.1 This is for you to answer! To me, another benefit is leveraging a structure that facilitates reproducibility.

**Exercise 2.4.1.1** (internal link) 1. There are some folder names within `inst` that should be avoided – what are they?

Answer 2.4.1.1 You should not name a folder in `inst` called `build`, `data`, `demo`, `doc`, `docs`, `exec`, `help`, `html`, `inst`, `libs`, `Meta`, `man`, `po`, `R`, `src`, `tests`, `tools`, or `vignettes`. The package building process automatically creates these folders upon installation – so also putting them in `inst` creates conflicts. For those that use GitHub, `inst/doc` is automatically placed in `gitignore`.

**Exercise 2.6.1.1** (internal link) Run `?devtools::build` to look at some additional arguments – which seem useful to you?

Answer 2.6.1.1 This is for you to answer! But `vignettes` and `manual` seem important to be aware of.

## Extensions

**Exercise 3.1.1.1** (internal link) What are some drawbacks of running `setwd()` whenever you want to set a working directory?

Answer 3.1.1.1 \* You have to remember to run `setwd()` anytime you want to re-run a script \* It is easy to run `setwd()` and point to the wrong working directory \* It means that your code is harder to share because someone else will have to change the line of code calling `setwd()` \* Can you think of more?

**Exercise 3.1.1.2** (internal link) Look into the `here` (external link) package, designed to help solve working directory problems (that can even occur within R Projects).

Answer 3.1.1.2 NA

**Exercise 3.2.1.1** (internal link) What benefits do you see to R Markdown documents?

Answer 3.1.1.1 This is for you to answer! One of my favorite benefits is making reproducible figures, tables, and output within the document.

**Exercise 3.2.1.2** (internal link) Save the basic R Markdown template we just discussed and knit it to see its contents.

Answer 3.1.1.1 Go to “File -> New File -> R Markdown” and create an R Markdown template. Then in the editor (upper-left) window, click “Knit”. Save the file to a location and then the document will knit to an HTML document and pop up in a default viewer. To knit to other output types, click on the down arrow to the right of the “Knit” button.

**Exercise 3.3.1.1** (internal link) Visit this (external) to see all the journals that articles has a template for.

Answer 3.1.1.1 NA

**Exercise 3.7.1.1** (internal link) 1. A data set is called “tidy” if it has three qualities: 1) each variable has its own column, 2) each observation has its own row, and 3) each value has its own cell. Is `river` tidy?

Answer 3.7.1.1 `river` is not tidy. The river names are spread across columns. One could also argue that `length` and `discharge` are both values of a `metric` argument. Then a tidy `river` data set would look like

`river_tidy`

```
#> # A tibble: 16 x 3
#>   name      metric  value
#>   <chr>    <chr>   <dbl>
#> 1 Missouri length    3768
#> 2 Mississippi length    3544
#> 3 Yukon     length    3190
#> 4 Colorado length    2330
#> 5 Arkansas length    2322
#> 6 Columbia length    2000
#> 7 Red       length    1811
#> 8 Canadian length    1458
#> 9 Missouri discharge  1956
#> 10 Mississippi discharge 18400
#> 11 Yukon     discharge  6340
#> 12 Colorado discharge    40
#> 13 Arkansas discharge  1004
#> 14 Columbia discharge  7730
#> 15 Red       discharge   852
#> 16 Canadian discharge   174
```

**Exercise 3.7.1.2** (internal link) Make `river` tidy

Answer 3.7.1.1

```

# approach 1 to loading packages
library(tibble) # run install.packages("tibble") if needed
library(tidyr) # run install.packages("tidyr") if needed
library(dplyr) # run install.packages("dplyr") if needed

# approach 2 to loading packages
library(tidyverse) # run install.packages("tidyverse") if needed

river_tidy <- river %>%
  tibble::rownames_to_column(var = "metric") %>%
  tidyr::pivot_longer(Missouri:Canadian) %>%
  dplyr::select(name, metric, value)

river_tidy

#> # A tibble: 16 x 3
#>   name      metric  value
#>   <chr>    <chr>   <dbl>
#> 1 Missouri length    3768
#> 2 Mississippi length   3544
#> 3 Yukon    length    3190
#> 4 Colorado length    2330
#> 5 Arkansas length    2322
#> 6 Columbia length    2000
#> 7 Red      length    1811
#> 8 Canadian length    1458
#> 9 Missouri discharge  1956
#> 10 Mississippi discharge 18400
#> 11 Yukon    discharge  6340
#> 12 Colorado discharge    40
#> 13 Arkansas discharge  1004
#> 14 Columbia discharge  7730
#> 15 Red      discharge   852
#> 16 Canadian discharge   174

```

# Bibliography

Ben Marwick, Carl Boettiger, and Lincoln Mullen. Packaging data analytical work reproducibly using r (and friends). *The American Statistician*, 72(1): 80–88, 2018.

Hadley Wickham. *R packages: organize, test, document, and share your code*. ” O’Reilly Media, Inc.”, 2015.

Yihui Xie. *Bookdown: authoring books and technical documents with R mark-down*. CRC Press, 2016.