Using R (RStudio + Tidyverse)

Arend M Kuyper

## Introduction

The purpose of this chapter is to provide you with an introduction to the free statistical software R along with RStudio and the tidyverse packages. While R can have a steep learning curve and be intimidating to new users, especially those new to coding, RStudio and the tidyverse packages make R much more accessible. Beyond improving accessibility, these tools are intentionally designed to make users more productive with R and improve the reproducibility of their work. Not surprisingly, this chapter is focused on the “how to” of foundational data work. We will highlight and demonstrate essential best practices for using R with RStudio and the tidyverse for quantitative research. However, don’t let this chapter’s focus on first steps of data work deceive, R and its ecosystem of extension packages allow for the implementation of all the statistical techniques found in this book and much, much more.

## R

R is a powerful and highly flexible statistical analysis tool. It provides a wide array of statistical techniques and methods while also providing highly developed graphics capabilities. R’s ability to create publication-quality graphics has been, and continues to be, one of its greatest strengths. It does all of this free of charge with a dedicated community of developers.

The R project is a GNU project, which means it is a free software, mass collaboration project. Knowing that R is open-source and is actively developed and maintained through mass collaboration provides important context for users concerning its basic structure and potential resources. R can be considered as being made up of two 2 parts:

1. the base R system that is downloaded from the Comprehensive R Archive Network, also known as CRAN, and
2. a large ecosystem of extension packages, sometimes called libraries.

The base R system is actively maintained and updated for various operating systems by the R Core Team. Typically there is 1 major update along with 2 minor updates per year. Having an active release schedule like this is critical for the success of open-source software. This ensures up-to-date compatibility with operating systems and signals to users that it won’t be abandoned.

While the base R system adequately covers most statistical needs and functionality, it is arguably the large ecosystem of extension packages that has contributed to R’s growth. Packages can provide implementations of methodologies not currently in base R, improve the usability of R, or provide tools that allow you to do non-statistical tasks (i.e. sending emails or building websites). In particular, the tidyverse packages have been very influential and have made working with R significantly more accessible. We will take a closer look at the tidyverse packages later in this chapter.

As of this writing, the CRAN package repository features over 21,000 contributed packages. The CRAN repository checks all packages for compatibility and expects the packages to be maintained, which means users can expect packages from the CRAN repository to work with R. This work is done by a network of volunteers, called the CRAN team, and it is a testament to the size and dedication of the R community that this is possible. Though CRAN is the primary R package repository, users can find packages through:

* CRAN-like repositories such as BioConductor and R-forge;
* GitHub and BitBucket;
* Personal websites.

While packages outside of the CRAN repository aren’t vetted by the CRAN team for compatibility, they can be very useful. They may implement cutting edge statistical techniques or provide tools for more bespoke analyses. Going through the CRAN submission process can be daunting, time intensive, and restrictive so it is not uncommon to find very useful packages not hosted on CRAN.

### Using R

A common roadblock for many new R users is that it requires the users to write code or commands. This can be a significant hurdle for many, but there are several free software options that make working with R much more user-friendly. The most popular being RStudio, which we will discuss in more detail later on in the chapter. The need to write code or commands isn’t removed, but it is made much more intuitive and accessing help is made easier. Using R and having to write code is a net positive for increasing the reproducibility of research, at least for computational and analysis work.

The value of learning to write R code is significantly enhanced by following best practices for coding and setting up workflows. When users are first learning it can seem unnecessary to follow such advice, but it is important to avoid developing bad and inefficient habits. RStudio and the tidyverse are specifically designed to guide users to follow and implement best practices. We will be highlighting and demonstrating some of these best practices in the following sections, but readers wanting more guidance should see the **Suggested further readings**.

## RStudio

Description of RStudio with images, setting global options for best practices, and using projects as a best practice

## Tidyverse

The tidyverse is an opinionated collection of R packages designed for data science. All packages share an underlying design philosophy, grammar, and data structures. (*Tidyverse*, 2024)

Overview of core packages and adjacent packages

Comment about piping

### Dataset

Quick overview of dataset to be used in demo.

### dplyr: data wrangling

Demo(s) with comments and a few tips for best practices

### ggplot2: data visualization

Demo(s) with comments and a few tips for best practices

## Comment: AI & R coding

*Comparable Wage Index for Teachers (CWIFT)* (2024)

Cornman, Nixon, Spence, Taylor, & Geverdt (2019)

## Conclusion

## Research essentials

## Questions for further investigation

## Suggested further reading

Hadley Wickham Danielle Navarro & Pedersen (2024)

Hadley Wickham Mine Çetinkaya-Rundel & Grolemund (2024)

## References

*Comparable wage index for teachers (CWIFT)*. (2024). NCES Education Demographic; Geographic Estimates (EDGE); <https://nces.ed.gov/programs/edge/Economic/TeacherWage>.

Cornman, S. Q., Nixon, L. C., Spence, M. J., Taylor, L. L., & Geverdt, D. E. (2019). *Education demographic and geographic estimates (EDGE) program: American community survey comparable wage index for teachers (ACS-CWIFT)* (No. NCES 2018130). Washington, DC: U.S. Department of Education; National Center for Education Statistics.

Hadley Wickham, Danielle Navarro, & Pedersen, T. L. (2024). *ggplot2: Elegant graphics for data analysis (3e)*. <https://ggplot2-book.org/>.

Hadley Wickham, Mine Çetinkaya-Rundel, & Grolemund, G. (2024). *R for data science (2e)*. <https://r4ds.hadley.nz/>.

*Tidyverse*. (2024). <https://www.tidyverse.org/>.