A Hitchhiker's Guide to 'R for Data Science (2e)'

Wednesday, 15 Oct 2025

RPIrates: The RPI R Users Group
The Rensselaer Future of Computing Institute
Rensselaer Polytechnic Institute

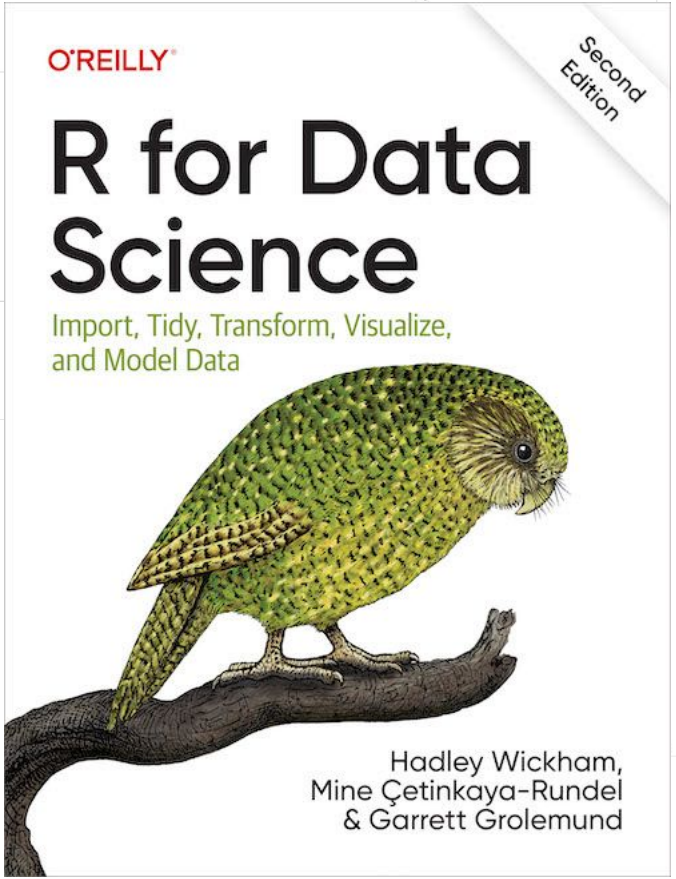
RPIrates
The RPI R Users Group

Tall

https://r4ds.hadley.nz/

How to do data science with R:

- Get your data into R...
- Get it into the most useful structure...
- Transform it...
- Visualize...
- Communicate!



https://github.com/hadley

Hadley Wickham

hadley · he/him

Unfollow

Chief Scientist at **@posit-pbc**

Pinned

tidyverse/dplyr Public

dplyr: A grammar of data manipulation

R 4.9k 2.1k

tidyverse/tidyverse Public

Easily install and load packages from the tidyverse

R 1.8k 292

r4ds Public

R for data science: a book

R 4.9k 4.4k

r-lib/devtools Public

Tools to make an R developer's life easier

R 2.5k 763

adv-r Public

Advanced R: a book

TeX 2.4k 1.7k

tidyverse/ellmer Public

Call LLM APIs from R

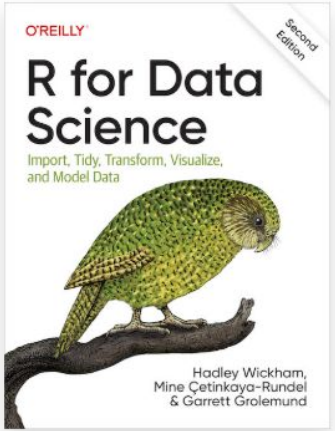
R 539 102

5,113 contributions in the last year

2025

R for Data Science (2e)

Welcome



This is the website for the 2nd edition of **"R for Data Science"**. This book will teach you how to do data science with R: You'll learn how to get your data into R, get it into the most useful structure, transform it and visualize.

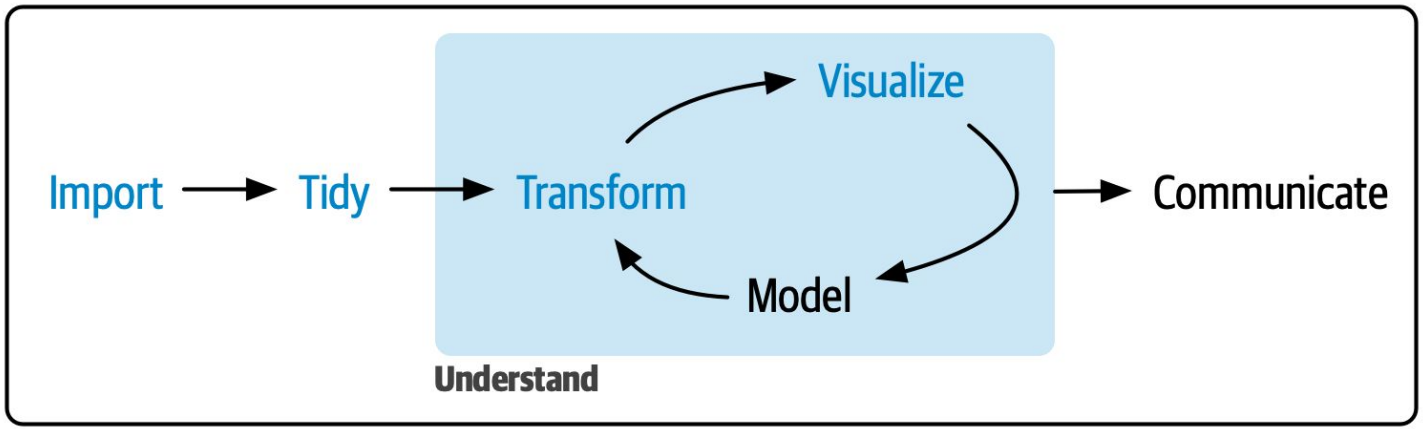
In this book, you will find a practicum of skills for data science. Just as a chemist learns how to clean test tubes and stock a lab, you'll learn how to clean data and draw plots—and many other things besides. These are the skills that allow data science to happen, and here you will find the best practices for doing each of these things with R. You'll learn how to use the grammar of graphics, literate programming, and reproducible research to save time. You'll also learn how to manage cognitive resources to facilitate discoveries when wrangling, visualizing, and exploring data.

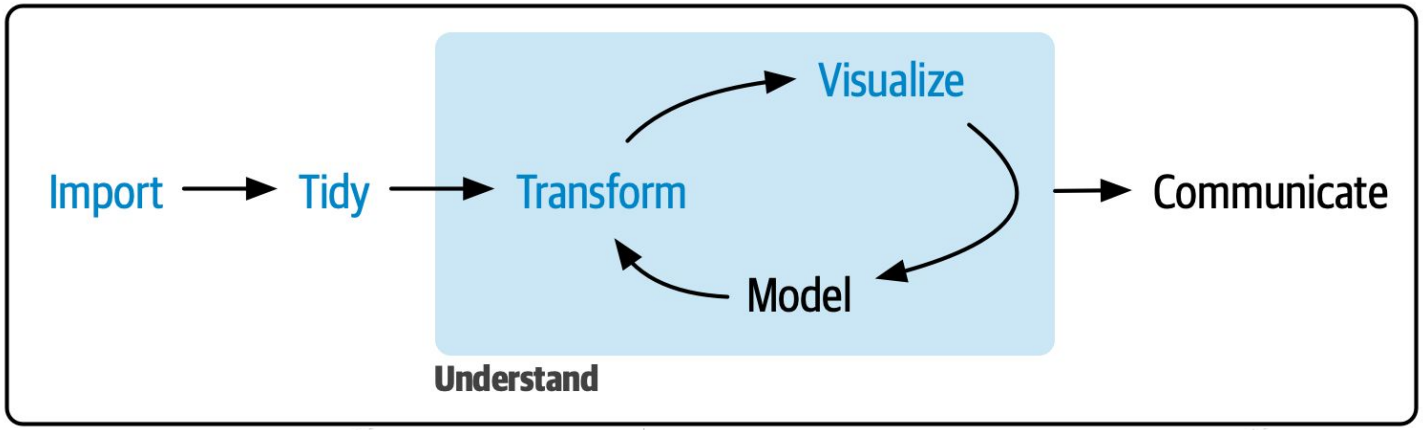
 If you'd like a physical copy of the book, you can order it on Amazon. If you appreciate reading the book for free and would like to give back, please make a donation to Kākāpō Recovery: the kākāpō (which appears on the cover of R4DS) is a critically endangered parrot native to New Zealand; there are only 244 left.

If you speak another language, you might be interested in the freely available translations of the 1st edition:

- Spanish
- Italian

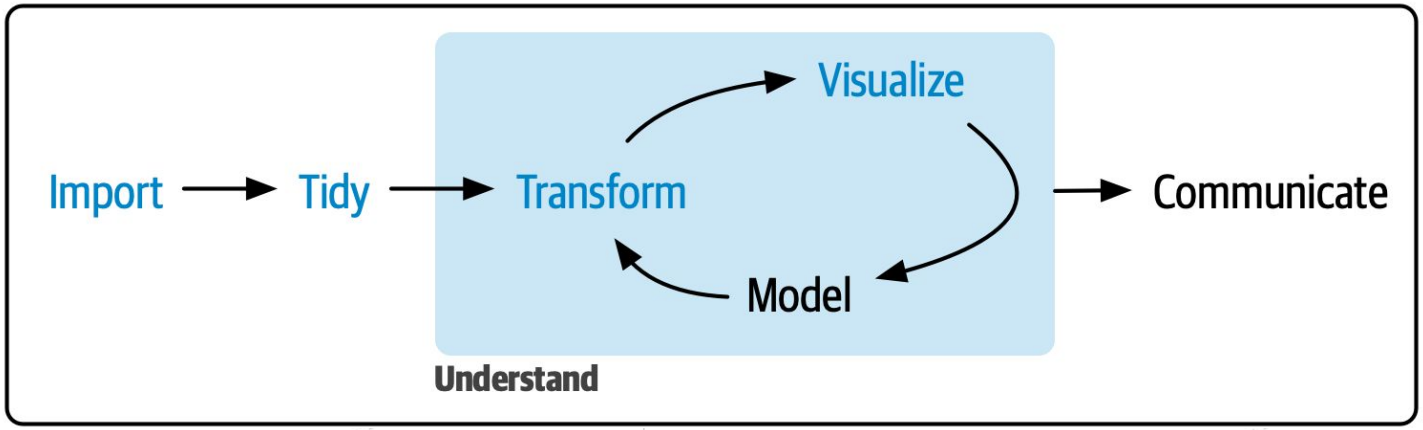
Introduction to R4DS: *Data Science in a Nutshell*





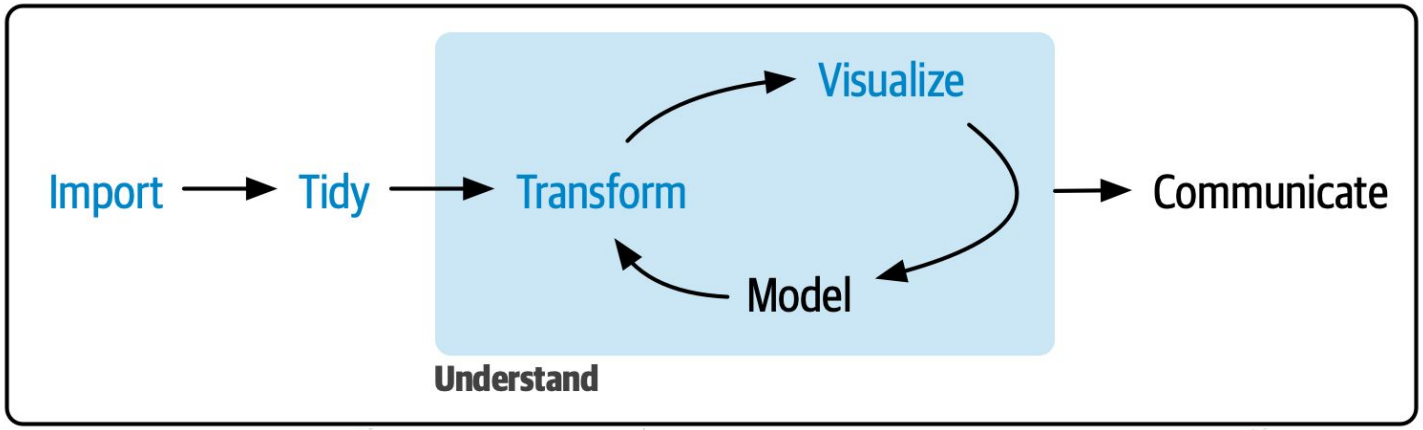
Import:

- Load data stored in a file, database, or obtained through a web API into an R dataframe
- "If you can't get your data into R, you can't do data science on it!"



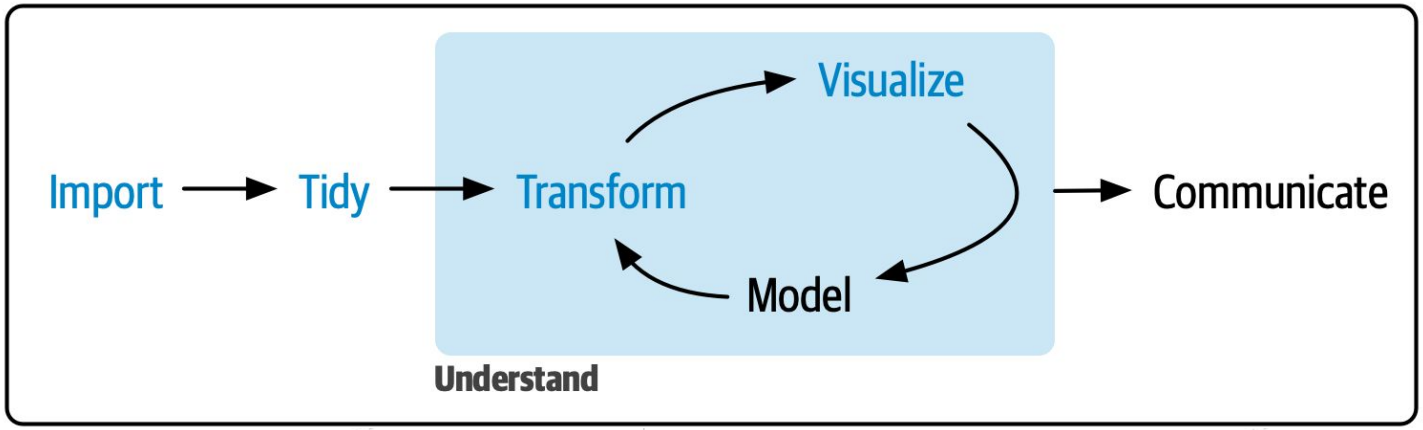
Tidy:

- Store it in a consistent form that matches the semantics of the dataset with how it is stored.
- When your data is tidy, each column is a variable and each row is an observation.
- Tidy data is important because the consistent structure lets you focus your efforts on answering questions about the data, not fighting to get the data into the right form for different functions.



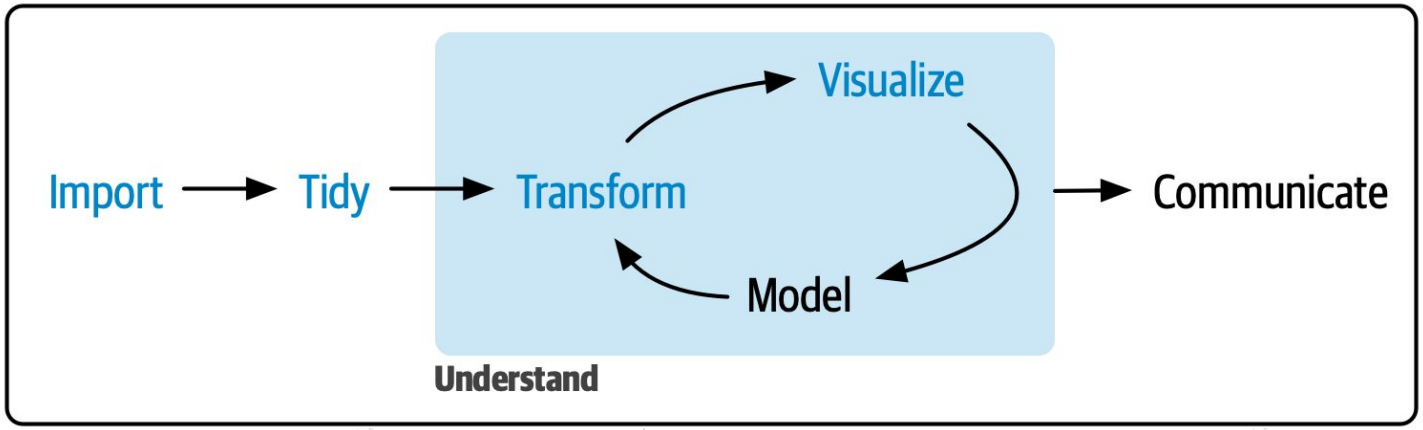
Transform:

- Focus on observations of interest (like *all people in one city* or *all data from the last year*)
- Create new variables from existing variables (e.g. compute *speed* from *distance* and *time*)
- Calculate a set of summary statistics (like *counts* or *means*)
- tidying and transforming together are called **wrangling**



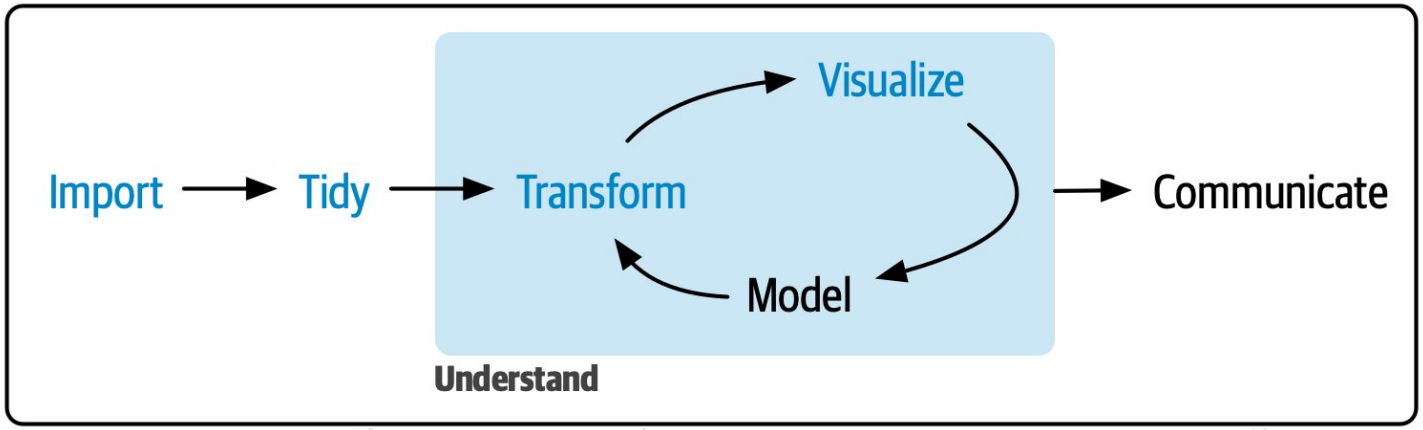
Visualize:

- Good visualizations will show you things you did not expect or raise new questions about the data.
- Good visualizations might hint that you're asking the wrong question or that you need to collect different data.
- Visualizations can surprise you, but don't scale particularly well because they require a human to interpret them.



Model:

- Models are complementary tools to visualization.
- Once your questions are sufficiently precise, you can use a model to answer them.
- Models are mathematical or computational tools, and they generally scale well.
- Every model makes assumptions, and by its very nature, a model cannot question its own assumptions.
- A model cannot fundamentally surprise you.



Communicate:

- An absolutely critical part of any data analysis project.
- It doesn't matter how well your models and visualization have led you to understand the data unless you can also communicate your results to others!

But that's not how the book* is organized!!

* And website...

What's in 'R for Data Science (2ed)'

- Visualization [Ch. 1]
  - The basics of ggplot2
  - Immediate payoff!!
- Data Transformation [Ch. 3]
  - The basics of dplyr--and the tidyverse!
  - How to: **select** important variables; **filter** out key observations; create (**mutate**) new variables; and **summarise** variables by **group**.
- Data Tidying [Ch. 5]
  - The underlying principles of having "tidy" data, making transformation, visualization, and modelling easier.
- Data Import [Ch. 7]
  - Before you can transform and visualize data, you need to get your data into R!

What's in 'R for Data Science (2ed)'

- Visualization [Ch. 1]
  - The basics of ggplot2
  - Immediate payoff!!
- Data Transformation [Ch. 3]
  - The basics of dplyr--and the tidyverse!
  - How to: **select** important variables; **filter** out key observations; create (**mutate**) new variables; and **summarise** variables by **group**.
- Data Tidying [Ch. 5]
  - The underlying principles of having "tidy" data, making transformation, visualization, and modelling easier.
- Data Import [Ch. 7]
  - Before you can transform and visualize data, you need to get your data into R!
- Workflow: Basics [Ch. 2]
  - The basics of coding in R
  - Assignments, comments, calling functions, etc.
- Workflow: Code Style [Ch. 4]
  - Introduce the most important points of the tidyverse style guide
- Workflow: Scripts and Projects [Ch. 6]
  - R scripts and the RStudio Script Editor
  - Creating and managing R projects
- Workflow: Getting Help [Ch. 8]
  - Online sources for R help
  - Making a reprex (reproducible example)

https://r4ds.hadley.nz/whole-game.html

What's in 'R for Data Science (2ed)'

- Visualization [Ch. 1]
  - The basics of ggplot2
  - Immediate payoff!!
- Data Transformation [Ch. 3]
  - The basics of dplyr--and the tidyverse!
  - How to: **select** important variables; **filter** out key observations; create (**mutate**) new variables; and **summarise** variables by **group**.
- Data Tidying [Ch. 5]
  - The underlying principles of having "tidy" data, making transformation, visualization, and modelling easier.
- Data Import [Ch. 7]
  - Before you can transform and visualize data, you need to get your data into R!

Each R4DS chapter is an excellent tutorial!

g functions, etc.

- Introduce the most important points of the tidyverse style guide

- Workflow: Scripts and Projects [Ch. 6]
  - R scripts and the RStudio Script Editor
  - Creating and managing R projects
- Workflow: Getting Help [Ch. 8]
  - Online sources for R help
  - Making a reprex! (reproducible example)

https://r4ds.hadley.nz/whole-game.html

What's NOT in 'R for Data Science (2ed)'

- Modelling
  - See esp.: Tidy Modelling with R (Max Kuhn and Julia Silge)
  - Modelling in the tidyverse style
- "Big Data"
  - See: data.table
  - A high-performance version of base R's data.frame, optimized for larger dat
  - Not consistent with the tidyverse...
- Shiny
  - See; https://shiny.posit.co/
- Deep R development including creating R packages
  - See: Advanced R (2ed) (Hadley Wickham)
- Python, Julia, and friends...

R4DS Deep Dives: **Visualize** [Ch. 9-11]

- Layers [Ch. 9]
  - Aesthetic mappings
  - Geometric objects
  - Facets
  - Statistical transformations
  - Position adjustments
  - Coordinate Systems
  - "The layered grammar of graphics"
- Exploratory Data Analysis [Ch. 10]
  - Variation
  - Unusual values
  - Covariation
  - Patterns and models
- Communication [Ch. 11]
  - Labels
  - Annotations
  - Scales
  - Themes
  - Layout

R4DS Deep Dives: **Transform** [Ch. 12-19]

- Logical vectors [Ch. 12]
  - How to create them in a variety of ways
- Numbers [Ch. 13]
  - Tools for vectors of numbers
- Strings [Ch. 14]
  - The tools to work with strings: slice them, dice them, stick them back together again.
  - Mostly focuses on the stringr package
- Regular Expressions [Ch. 15]
  - A powerful tool for manipulating strings
- Factors [Ch. 16]
  - The data type that R uses to store categorical data
- Dates and times [Ch. 17]
  - The key tools for working with dates and date-times
  - With the help of the lubridate package, you'll learn to how to overcome the most common challenges.
- Missing Values [Ch. 18]
  - Missing values in depth
- Joins [Ch. 19]
  - Tools to join two (or more) data frames together
  - Grappling with the idea of keys, and think about how you identify each row in a dataset

R4DS Deep Dives: **Import** [Ch. 20-24]

- Spreadsheets [Ch. 20]
  - How to import data from Excel spreadsheets and Google Sheets
- Databases [Ch. 21]
  - How to get data out of a database and into R
  - …and a little about how to get data out of R and into a database
- Arrow [Ch. 22]
  - A powerful tool for working with out-of-memory data, particularly when it's stored in the "parquet" format
- Hierarchical Data [Ch. 23]
  - How to work with hierarchical data
  - Esp. deeply nested lists produced by data stored in JSON!
- Web Scraping [Ch. 24]
  - The art and science of extracting data from web pages

R4DS Deep Dives: **Program** [Ch. 25-27]

- Functions [Ch. 25]
  - Avoid copying and pasting more than twice!
  - Repeating yourself in code is dangerous because it can quickly lead to errors and inconsistencies
  - Learn to write functions that let you extract out repeated tidyverse code so that it can be easily reused
- Iteration [Ch. 26]
  - You often need to repeat the same actions on different inputs
  - You need tools for iteration that let you do similar things again and again
  - These tools include for loops and functional programming
- "A field guide to base R" [Ch. 27]
  - You'll often see code that doesn't use the tidyverse
  - Here you'll learn some of the most important base R functions that you'll see in the wild

R4DS Deep Dives: **Communicate** [Ch. 28-29]

- Quarto [Ch. 28]
  - A unified authoring framework for data science, combining code, results, and prose.
  - Quarto documents are fully reproducible and support dozens of output formats, like PDFs, Word files, presentations, and more
  - Think, "R Markdown on steroids!"

- Quarto Formats [Ch. 29]
  - The many varieties of Quarto outputs: dashboards, websites, even books!