

# R FOR DATA SCIENCE

Getting started with Tidyverse R,  
RStudio, and Quarto

John R Little

Duke University

January 5, 2023

CC BY 4.0 John R Little



# R FOR DATA SCIENCE

Getting started with Tidyverse R, RStudio, and Quarto

John R Little

Center for Data & Visualization Sciences

Duke University

2023 January 05

Get code, data, and slides for today's workshop

- [https://github.com/libjohn/rfun\\_flipped](https://github.com/libjohn/rfun_flipped)

- <https://github.com/libjohn/intro2r-code>

15:00

CC BY 4.0 John R Little



## TOPICS

- How to use R
- How reproducibility is easily accomplished
- How to learn R efficiently
  - Part 1 (today): focus on data wrangling with dplyr
  - Part 2: visualization with ggplot2, briefly: EDA, interactive plots, linear regression models
  - Part 3: iterations & custom functions
  - Part 4: case study in wrangling and visualization by ingesting multiple excel worksheets from multiple excel workbooks

CC BY 4.0 John R Little



## R, THE TIDYVERSE, AND QUARTO

R is a programming language

1. A data-first programming language → computational thinking
2. The Tidyverse (and Tidymodels) is designed for humans
3. Quarto Notebooks: a publishing system
  - a. Publish high-quality articles, reports, presentations (slide-decks), websites, blogs, and books in HTML, PDF, MS Word, ePub, and more.
  - b. Works with other languages and IDEs

CC BY 4.0 John R Little



## REPRODUCIBILITY

**Reproducibility** - Obtaining computational results using the same input data, computational steps, methods, code, and conditions of analysis.<sup>1</sup> **Replicability** - Obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data.

Goals of a tool-based, first-class approach

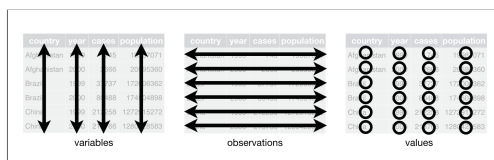
- Do as much as possible with code
- Integrate prose with code; **visualize inline**
- Generate reports for target audience
- Iterate efficiently; {purrr}

CC BY 4.0 John R Little



## TIDY DATA

- Each variable is a column
- Each observation is a row
- Each type of observational unit is a table



Citation: <https://doi.org/10.18637/jss.v059.i10>

Preprint: <https://vita.had.co.nz/papers/tidy-data.pdf>

See more in **R for Data Science** by Wickham and Grolemund

CC BY 4.0 John R Little



## TIDYVERSE

- A dialect of R
- Easier to learn because of consistency and documentation
- Assumptions
  - Data have semantic meaning that can be documented grammatically
  - Tidy data are wrangled, visualized, and iterated easily via grammar
  - 50-80% of any data project is data wrangling  
{dplyr} & {tidyr}

CC BY 4.0 John R Little



## PIPE | DATA PIPE | DATA SENTENCE

A conjunction (“and then”), read left to right, creating a “data sentence”

`%>%` ({magrittr}) | `{tidyverse}` or `|>` (base R)

```
1 Starwars |>
2   select(name, skin_color, homeworld)
```

CC BY 4.0 John R Little



## ASSIGNMENT

An object name “gets value from” a data pipe

<-

```
1 small_df <- starwars |>
2   filter(gender == "feminine")
```

CC BY 4.0 John R Little



## PROJECT

- Keep stuff organized in the same directory  
e.g. data, analysis, scripts, documentation, and outputs
- In the notebook, refer to subdirectories via relative paths  
better than `setwd()`
- Shareability, portability, legibility, and **reproducibility**  
Use Restart-R-and-Run instead of `rm(list=ls())`

CC BY 4.0 John R Little



## NOTEBOOKS

Literate coding

- Intersperse prose and code
- Integrated outputs with analysis
- Render reports from code

CC BY 4.0 John R Little



## {dplyr}

`library(dplyr)` or `library(tidyverse)`. Use {dplyr} to wrangle data

`select`

`filter`

`arrange`

`mutate`

`group_by`

`summarize`

subset by column

subset by row

sort

generate new variables

column totals (or subtotals with `group_by`)

CC BY 4.0 John R Little



## **UPCOMING WORKSHOPS**

- Visualization with ggplot2 (and interactive graphics) & Modeling (syntax)
- Iteration and custom functions
- Quarto and Observable interactivity

CC BY 4.0 John R Little



## **RFUN RESOURCES**

- Web scraping
- Slide decks
- Text mining, sentiment analysis, etc.
- Dashboards ; interactivity
- DBI: i.e. SQL without knowing SQL (working with databases)
- git/GitHub

CC BY 4.0 John R Little



## NEXT STEPS

Best way to learn *and/or* **consultations**

- Take a small subset of a project you know well then recreate it in R
- If you get stuck, schedule me for a free consultation walk-ins welcome
- Documentation: {package-name}.tidyverse.org (<https://dplyr.tidyverse.org>)
- Ask questions at RStudio Community ; R for DS online learning community ; R Ladies RTP
- Formulate questions as **RE**producible **EX**amples ([REPREX.tidyverse.org](https://reprex.tidyverse.org))

CC BY 4.0 John R Little



**FIN**



John R Little  
Data Science Librarian  
Center for Data & Visualization Sciences  
Duke University Libraries

<https://JohnLittle.info> • <https://Rfun.library.duke.edu> • <https://Library.duke.edu/data>

CC BY 4.0 John R Little







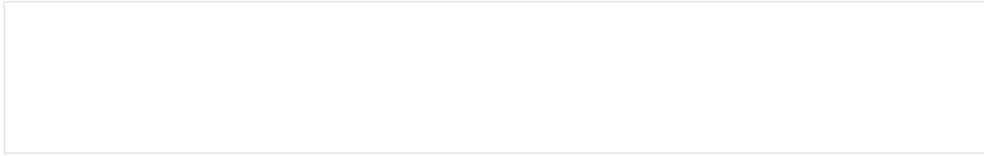
Creative Commons: Attribution 4.0 • <https://creativecommons.org/licenses/by-nc/4.0>

# 1. National Academies of Sciences, Engineering, and Medicine (NASEM)

Image Credit: <https://r4ds.hadley.nz>

**Error**

×



CC BY 4.0 John R Little

