R FOR DATA SCIENCE

Getting started with Tidyverse R,
RStudio, and Quarto

John R Little
Duke University

January 10, 2023

R FOR DATA SCIENCE

Getting started with Tidyverse R, RStudio, and Quarto
John R Little
Center for Data & Visualization Sciences
Duke University
2023 January 10
Get code, data, and slides for today's workshop
- 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

R, THE TIDYVERSE, AND QUARTO

R is a programming language

1. A data-first programing 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

REPRODUCIBILITY

Reproducibility - Obtaining computational results using the same input data, computational steps, methods, code, and conditions of analysis. [1]

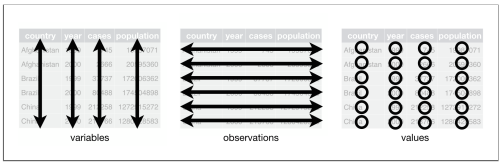
Replicability - Obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data.

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}

TIDY DATA

- Each variable is a column
- Each observation is a row
- Each value is a cell; each cell is a single value



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

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}

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)
```


ASSIGNMENT

An object name “gets value from” a data pipe

`<-`

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


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())`

NOTEBOOKS

Literate coding

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

PSHEW

Image credit: pixabay - https://pixabay.com/photos/rabbit-white-sleep-905971

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`

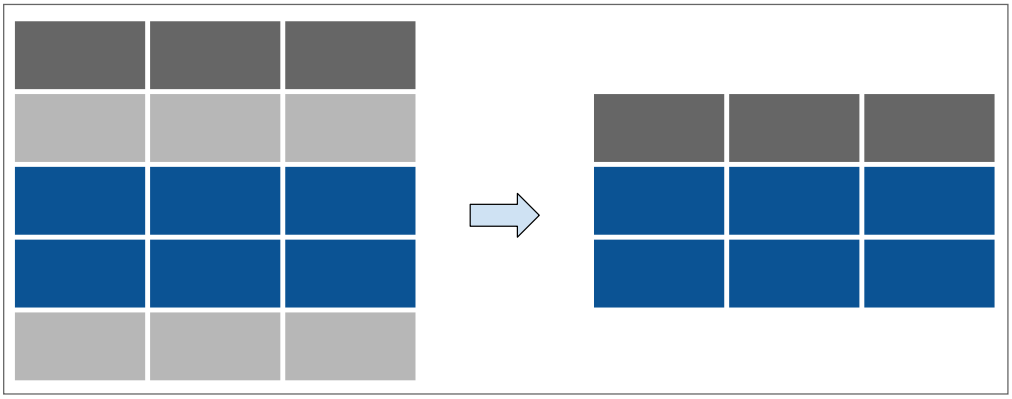
- subset by column
- subset by row
- sort
- generate new variables

- column totals (or subtotals with `group_by`)

filter

SUBSET ROWS BY VARIABLES

```
starwars %>% filter(eye_color == "orange")
```



CC BY 4.0 John R Little

select

SUBSET BY COLUMNS (VARIABLES)

```
starwars %>% select(hair_color, eye_color)
starwars %>% select(2:4)
starwars %>% select(name:mass, 10, 7, 4:6)
```

arrange

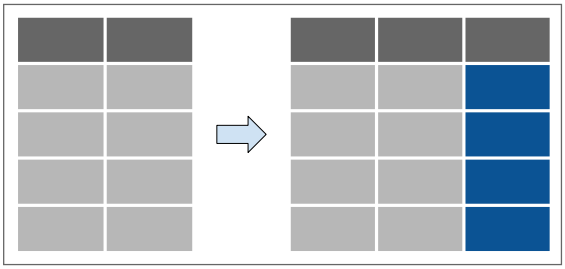
SORT ROWS BY VARIABLES

```
starwars %>% arrange(eye_color)
starwars %>% arrange(desc(eye_color))
starwars %>% arrange(desc(eye_color), hair_color)
```

mutate

CHANGE CELL VALUES

```
starwars %>% mutate(big_mass = mass * 100)
starwars %>% mutate(BMI = (mass / (height/100)^2))
starwars %>% mutate(
       nickname = str_c("Big", str_to_upper(hair_color),
                        sep = " "))
```



count

COUNT OBSERVATIONS BY GROUP

```
1  starwars %>% count(gender)
```


summarize

REDUCE MULTIPLE VALUES TO A SINGLE VALUE

```
starwars %>%
  drop_na(height) %>%
  summarise(n(), n_distinct(height), min(height), max(height))

starwars %>%
  drop_na(height) %>%
  group_by(gender) %>%
  summarise(Total = n(), n_distinct(height), min(height))
```


DOWNLOAD CODE

https://github.com/libjohn/rfun_flipped

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

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 **REP**roducible **EX**amples (REPREX.tidyverse.org)

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

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

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