R FOR LUNCH

Data wrangling with {dplyr}

John Little
Duke University Libraries
Center for Data & Visualization Sciences

2024-01-17

TODAY'S TOPICS

- Five essential {dplyr} data wrangling verbs
- Data pipes inside code-chunks

YESTERDAY (VIDEO)

- Import data
- Tour of RStudio IDE
- Coding notebooks (Quarto)

HOUSEKEEPING

- Drew / Lauren / breakout rooms
- CDVS
  - Themes
    - Data Management (Plans, Reproducibility, Repositories)
    - Data Science
    - Data Visualization
    - GIS and Spatial Analysis
    - Data Sources

HOUSEKEEPING CONTINUED

- Website - https://library.duke.edu/data
- Workshops
  - https://library.duke.edu/data/workshops
- Consulting in the Lab
  - askData@duke.edu
  - my schedule: https://is.gd/littleconsult

R FOR LUNCH AS A SERIES

R for Lunch is a series that meets 8 times (till EOM Feb.)
After today it will meet regularly on Thursdays at noon.

- Sign-up for each workshop individually
- Each episode has a unique zoom link

EAT YOUR OWN DOG FOOD

Model how R can work for practical reproducible workflows

- Code in RStudio
- One kind of report is these slides (GitHub)
- Another report is the *Introduction to R/Tidyverse/Quarto* text.

PIPES AND ASSIGNMENTS

Operator	Operator Name	Keystore	Pnuemonic
`<-`	assignment	Alt-dash	“Gets value from”
`|>` or `%>%`	pipe	Ctrl-Shift-M	“And then”

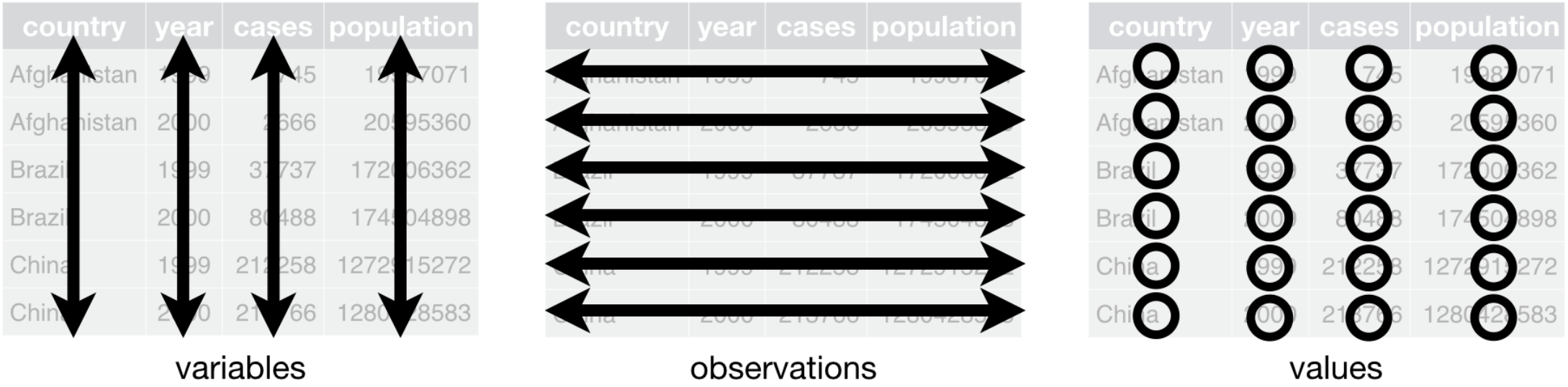
TIDYVERSE AND TIDY DATA

John R Little • Center for Data & Visualization Sciences •

FOUNDATION

Tidyverse and Quarto is the most practical and well developed, reproducible, scientific analysis and publishing workflow available.

TIDY DATA[1]



1. A robust discussion of *tidy data* can be found in *R for Data Science* (Wickham, Cetinkaya-Rundel, and Grolemund 2023): *https://r4ds.had.co.nz/tidy-data.html*

TIDY DATA

- Every row is a single observation
- Every column is a variable
- The cells are single data values

WIDE DATA

- Code

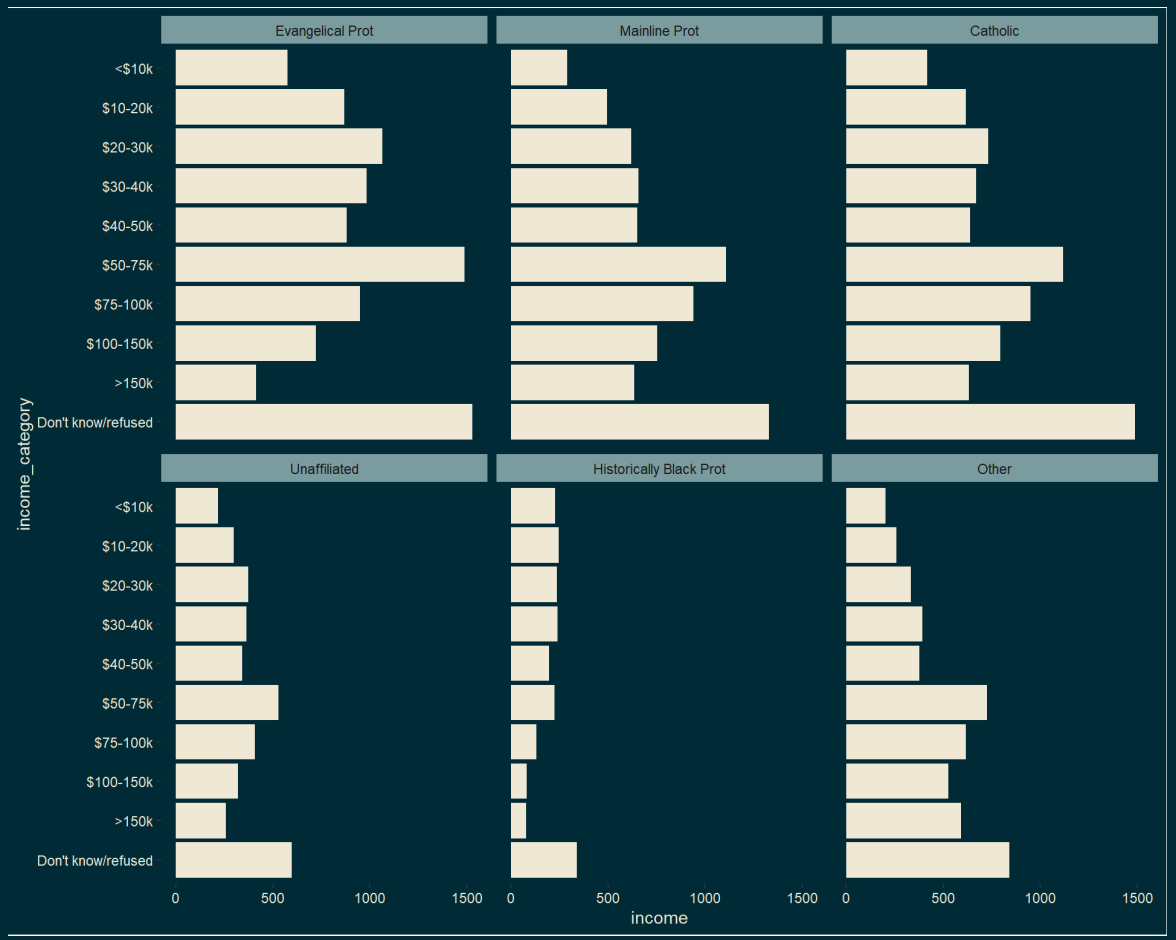
	RELIGION	<$10K	$10-20K	$20-30K	$30-40K	$40-50K	$50-75K	$75-100K	$100-150K	>150K	DON'T KNOW/REFUSED
1	Agnostic	27	34	60	81	76	137	122	109	84	96
2	Atheist	12	27	37	52	35	70	73	59	74	76
3	Buddhist	27	21	30	34	33	58	62	39	53	54
4	Catholic	418	617	732	670	638	1116	949	792	633	1489
5	Don’t know/refused	15	14	15	11	10	35	21	17	18	116
6..17											
18	Unaffiliated	217	299	374	365	341	528	407	321	258	597

TALL DATA

▶ Code

	RELIGION	INCOME_CATEGORY	INCOME
1	Agnostic	<$10k	27
2	Agnostic	$10-20k	34
3	Agnostic	$20-30k	60
4	Agnostic	$30-40k	81
5	Agnostic	$40-50k	76
6..179			
180	Unaffiliated	Don't know/refused	597

▶ Code



CODE

```
relig_income |>
  pivot_longer(cols = -religion, names_to = "income_category") |>
  ggplot(aes(value, income_category)) +
  geom_col() +
  facet_wrap(vars(religion))
```

Image Credit: apreshill | CC BY 4.0 | https://github.com/apreshill/teachthat/blob/master/pivot/pivot_longer_smaller.gif]

POLLS

DPLYR

https://intro2r.library.duke.edu/wrangle.html

WE ARE HERE TO HELP

- askData@duke.edu
- https://library.duke.edu/data
- https://is.gd/littleconsult

LET'S DO IT

TWO THINGS FOR TODAY

- five essential {dplyr} data wrangling verbs
- data pipes inside code-chunks
- https://intro2r.library.duke.edu/wrangle.html

EXERCISES

1. https://intro2r.library.duke.edu/ > Exercises > Link out > Green **Code** button > Download ZIP
2. Then, Unzip (i.e. Expand) the folder (on your local file system)
3. Then, double click the `rforlunch_exercises.Rproj` file
4. From **RStudio** the Files tab, open the `01_dplyr.qmd`
    - The answer file is in the **RStudio** `rforlunch_exercises` project > `Files Tab > Answers` folder

CLOSING

PIPES AND ASSIGNMENTS

Operator	Operator Name	Keystore	Pnuemonic
`<-`	assignment	Alt-dash	“Gets value from”
`|>` or `%>%`	pipe	Ctrl-Shift-M	“And then”

CITATION MANAGEMENT

RStudio > Quarto Notebook > Insert > Citation

Example DOI: 10.18637/jss.v059.i10

AI-PAIRED CODING

- Data science concepts: **Microsoft copilot** (“More precise” setting)
- Code completion: **GitHub copilot** and RStudio (IDE) or VSCode (IDE)

BYE FOR NOW

- askData@duke.edu
- https://is.gd/littleconsult
- https://library.duke.edu/data