

# R FOR LUNCH

Import data and install RStudio / Tidyverse

John Little 

Duke University Libraries

Center for Data & Visualization Sciences

2024-01-17



# TODAY'S TOPICS

- How to import data
- Tour of RStudio IDE
- Coding notebooks

Preceded by [where to download RStudio and R](#)

# 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](mailto: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 YOR 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](#)*.

# DEFINITIONS



# R/TIDYVERSE/QUARTO

R/Tidyverse/Quarto represents the state of the art for practical reproducibility

# R & RSTUDIO

*R* is a data-first programming language

RStudio is an IDE

# REPRODUCIBILITY

- Independently and transparently achieve reliable results with the same data and the same workflow
  - Transparency with reproducible workflows
- Best workflow and ecosystem to achieve reproducible work is to “*do everything with code*”
  - Import data, analyze, visualize, and publish/share

# TIDYVERSE

- An opinionated set of packages for data manipulation and analysis
- A meta-package of eight symbiotic packages

## PACKAGES

- Extend R into your subject domain
- And/or make it easier to accomplish a computational task
- There are thousands
  - MetaCRAN, CRAN, BioConductor, GitHub

# QUARTO

works with R and Python

- A scientific publishing system (workflow)
  - dashboards, manuscripts, MSWord, slides, website, e-book, PDF
- **Coding Notebooks:** Code chunks interspersed with explanatory text (Natural language)
  - Render reproducible, shareable reports
- A next-gen (or modern) Markdown

# OPINIONATED

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

# TIDY DATA

# TIDY DATA<sup>1</sup>

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	127291272
China	2000	213766	128042583

variables

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	127291272
China	2000	213766	128042583

observations

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	127291272
China	2000	213766	128042583

values

1. A robust discussion of *tidy data* can be found in *R for Data Science* (Wickham, John R Little • Center for Data & Visualization Sciences • CC BY 4.0  
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

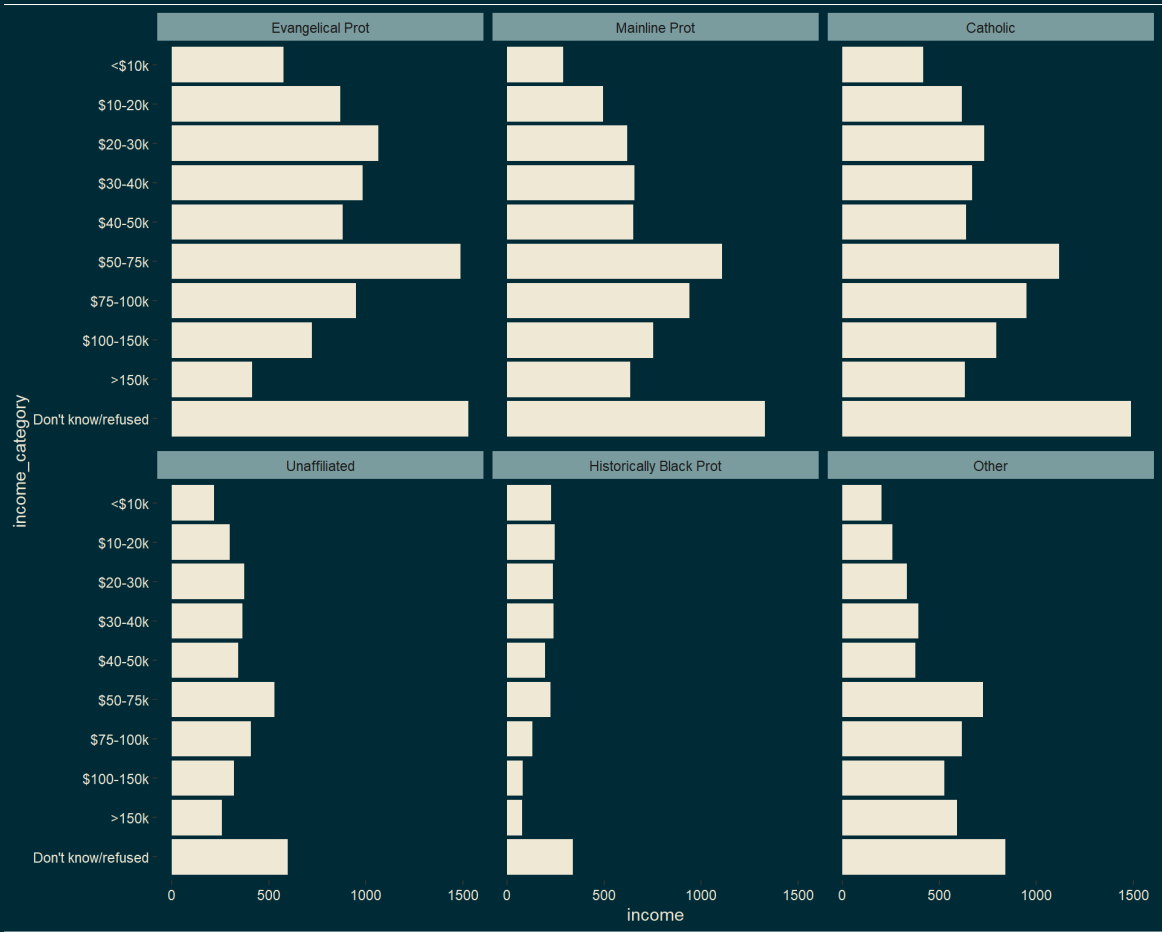


# TALL DATA

► 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

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

Image Credit: apreshill | CC BY 4.0 | [https://github.com/apreshill/teachthat/blob/master/pivot/pivot\\_longer\\_smaller.gif](https://github.com/apreshill/teachthat/blob/master/pivot/pivot_longer_smaller.gif)

# POLLS

# GRAMMAR (DATA AND GRAPHICS)

By next week you'll have the basic building blocks to

- Leverage reproducible data workflows: import data, analyze data, and generate visualizations.

Along the way

- Rendering reproducible reports (Quarto)
- Practical techniques
- Pro-tips that comprise a fluency of reproducible data analysis

# WE ARE HERE TO HELP

- [askData@duke.edu](mailto:askData@duke.edu)
- <https://library.duke.edu/data>
- <https://is.gd/littleconsult>

# LET'S DO IT



# THREE THINGS FOR TODAY

- Tour of the RStudio IDE (Projects)
- How to import data following
- Coding notebooks

# 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 **00\_import\_answers.qmd**
  - The answer file is in the RStudio **rforlunch\_exercises** project > **Files Tab** > **Answers folder**

# CLOSING

# PIPES AND ASSIGNMENTS

Operator	Operator Name	Keystore	Pnemonic
<-	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>