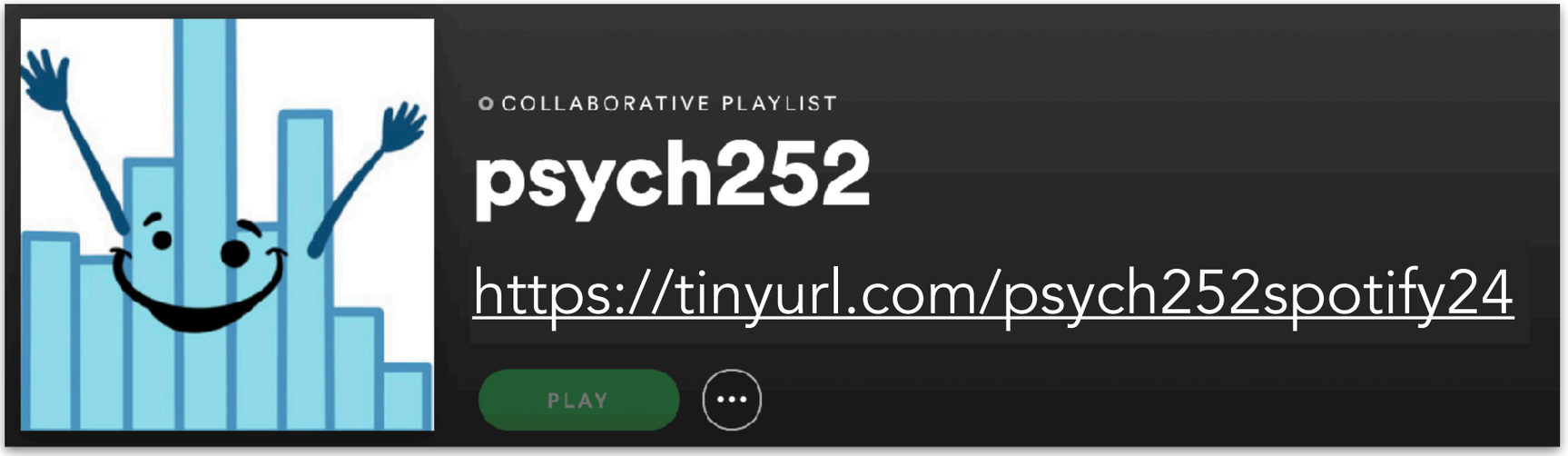
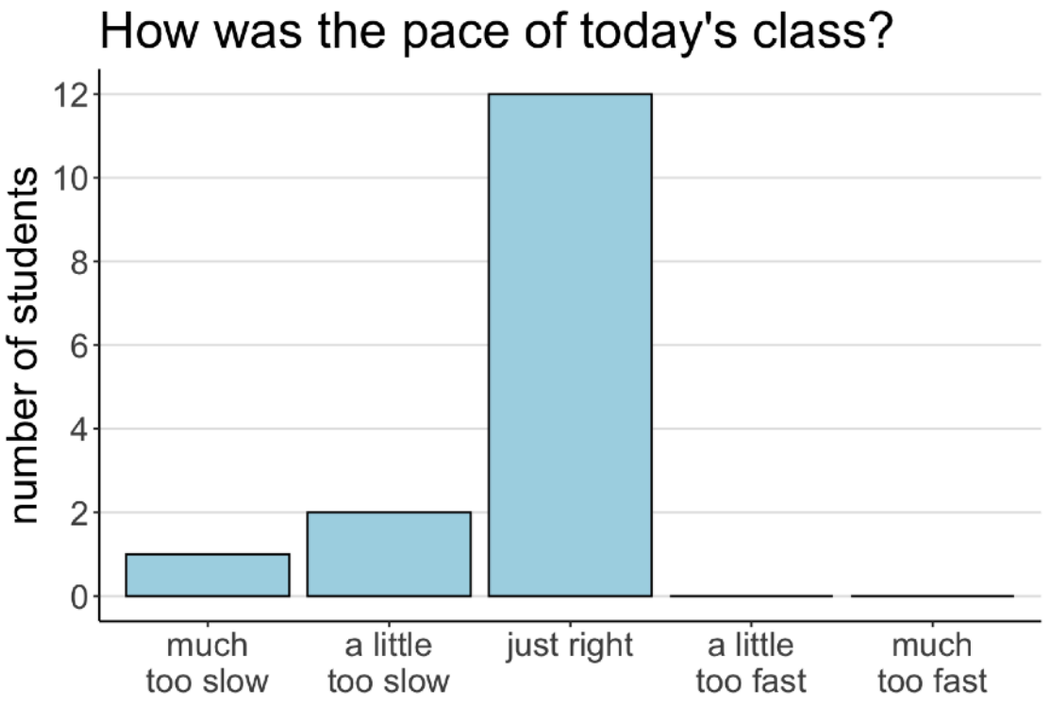
Data wrangling 2

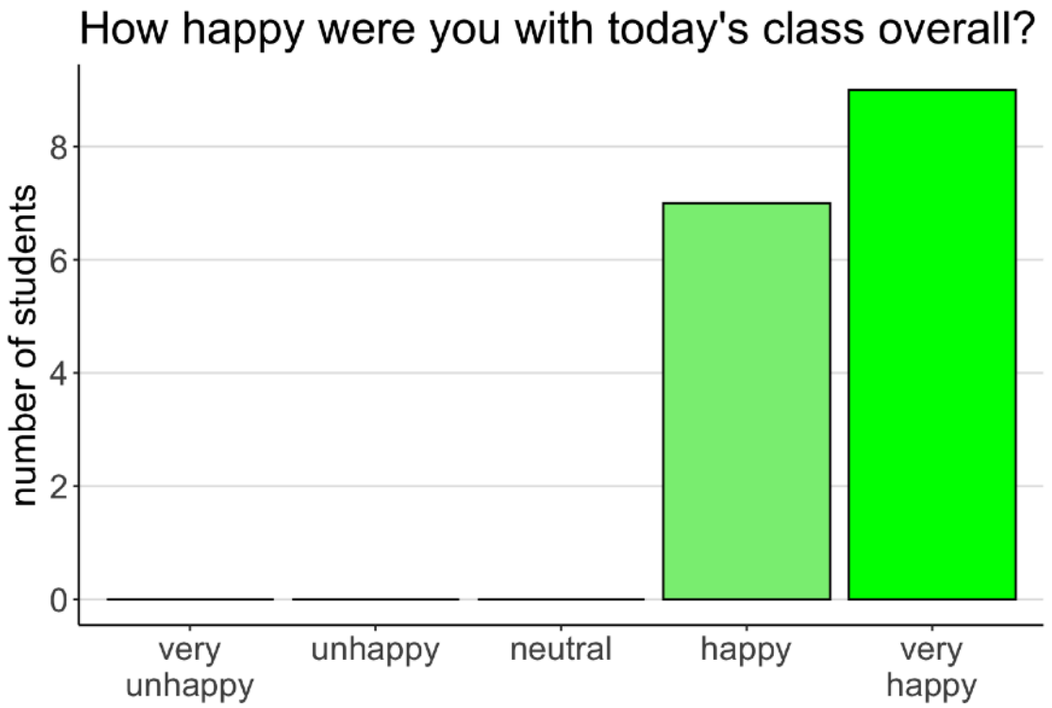
"Data don't make any sense,
we will have to resort to statistics."



Your feedback

Your feedback



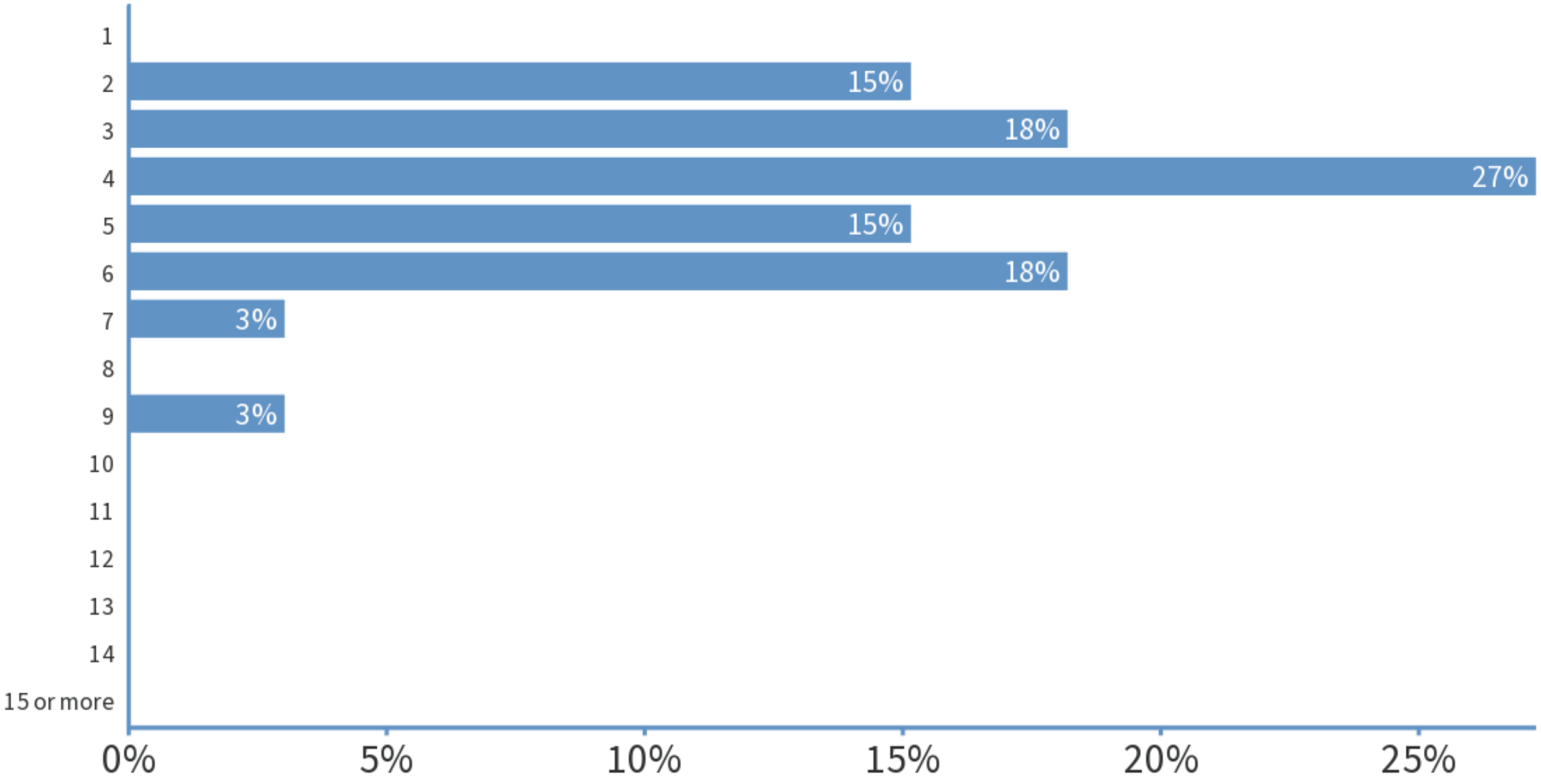


I like the style of the R tutorials that start from the very basics. These are helpful and encouraging to work in R

It was useful to go through R syntax (it felt slow for me but I imagine for those new to programming it was fast). Could be helpful to include secondary activities for people who already have the skill!

I really liked all the demonstrations in today's class! I feel like it was a great pace!

How many hours did it take you to complete Homework 1?



Things that came up

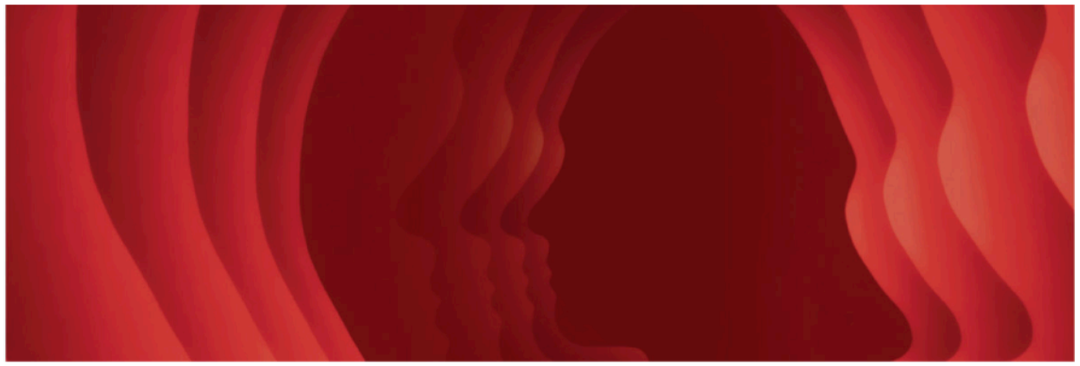
APS SPOTLIGHT

Vazire Outlines Goals for Transparency, Diversity in *Psychological Science*

December 19, 2023

TAGS: DIVERSITY | FEATURE | METHODS | PSYCHOLOGICAL SCIENCE | PUBLISHING

Log in to Save for Later



Under the Cortex

Getting Your Research Published: Insights on Academic Publishing with Simin…

0:00 -0:00

APS Fellow Simine Vazire is introducing new steps to ensure transparency, rigor, and diversity in the pages of *Psychological Science*, APS's flagship journal. Vazire became editor-in-chief of the publication Jan. 1, succeeding Patricia Bauer of Emory University, who continues as consulting editor for the publication.

Among the changes Vazire is bringing to *Psychological Science* is a new group of Statistics, Transparency, and Rigor (STAR) editors, whose core role will be confirming that articles contain sufficient information to ensure scientists can independently verify research claims. Tom E. Hardwicke, a research fellow at the University of Melbourne's School of Psychological Sciences, will lead the STAR team.

Vazire also plans to increase the number of people from underrepresented groups onto the publication's editorial board. She says she will continue the efforts that Bauer implemented to increase the diversity of research populations, authors, and subdisciplines in the journal's content.

Simine Vazire

https://www.psychologicalscience.org/observer/simine-vazire-psychological-science

Free access | Editorial | First published online December 27, 2023

Transparency Is Now the Default at *Psychological Science*

Tom E. Hardwicke and Simine Vazire View all authors and affiliations

OnlineFirst | https://doi.org/10.1177/09567976231221573

Contents | PDF / ePub | Cite article | Share options | Information, rights and permissions

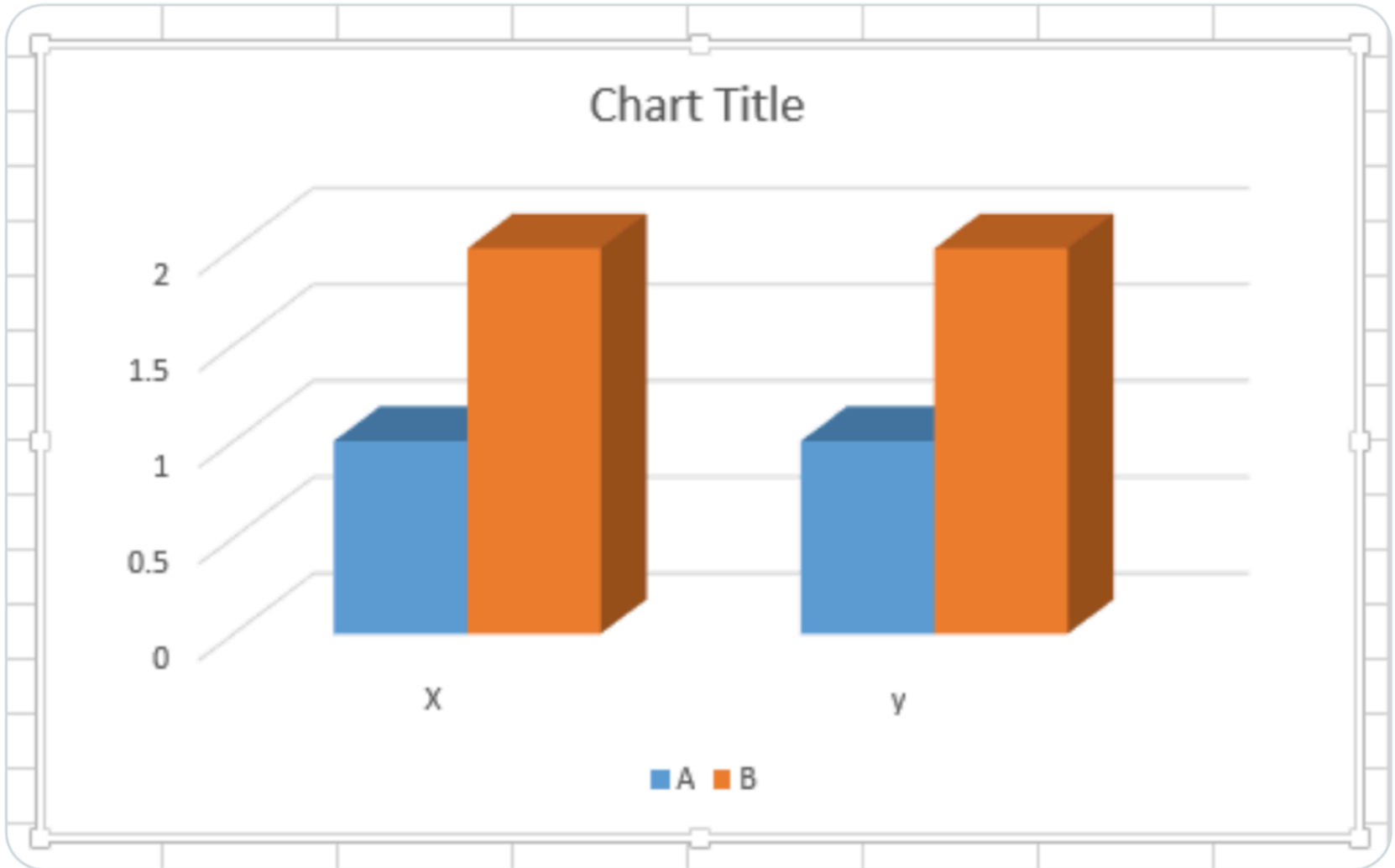
In the early 2010s, serious concerns about statistics, transparency, and rigor triggered an unprecedented period of introspection and change in the field of psychology. *Psychological Science* was quick to respond: In a 2013 editorial titled "Business Not as Usual," then–editor in chief Erich Eich introduced several new initiatives to raise standards, including removing word counts for Method sections and introducing "badges" to encourage preregistration and sharing of data and materials (Eich, 2014). Subsequent editors Stephen Lindsay and Patricia Bauer advanced these policies and the journal currently (2023) requires a range of transparent research practices during peer review and encourages them postreview (Bauer, 2020, 2021, 2022; Lindsay, 2017, 2019).

A decade on, it is clear that things are moving in the right direction. In 2022, 69% of articles published in *Psychological Science* had an open data badge, 55% had an open materials badge, and 43% had a preregistration badge (Bauer, 2023). However, a series of metaresearch studies have highlighted gaps between the theoretical goals of research transparency and its implementation. For example, sharing data does not guarantee that an independent scientist can reuse them or reproduce the reported results (Hardwicke et al., 2018; Towse et al., 2020), and the extent to which preregistrations effectively constrain analytic flexibility in practice varies considerably because of a lack of detail and undisclosed deviations from the original plan (Bakker et al., 2020; Claesen et al., 2021; van den Akker et al., 2023). Most pertinently, three studies have reported difficulties reproducing the results of articles published in *Psychological Science* (Crüwell et al., 2023; Hardwicke et al., 2021; Homewood, 2023). In short, transparency is improving, but there is much work left to do.

https://journals.sagepub.com/doi/10.1177/09567976231221573

Daniel Gould, MD
@DJGould94

3D bar graphs.
Do they add any valuable information? No
But do they look cool? Also no



3:17 AM · Jan 15, 2024 · **70.4K** Views

19 78 1K 17

Data wrangling time ...

Tidy data

"TIDY DATA is a standard way of mapping the meaning of a dataset to its structure."

–HADLEY WICKHAM

In tidy data:

- each variable forms a column
- each observation forms a row
- each cell is a single measurement

each column a variable

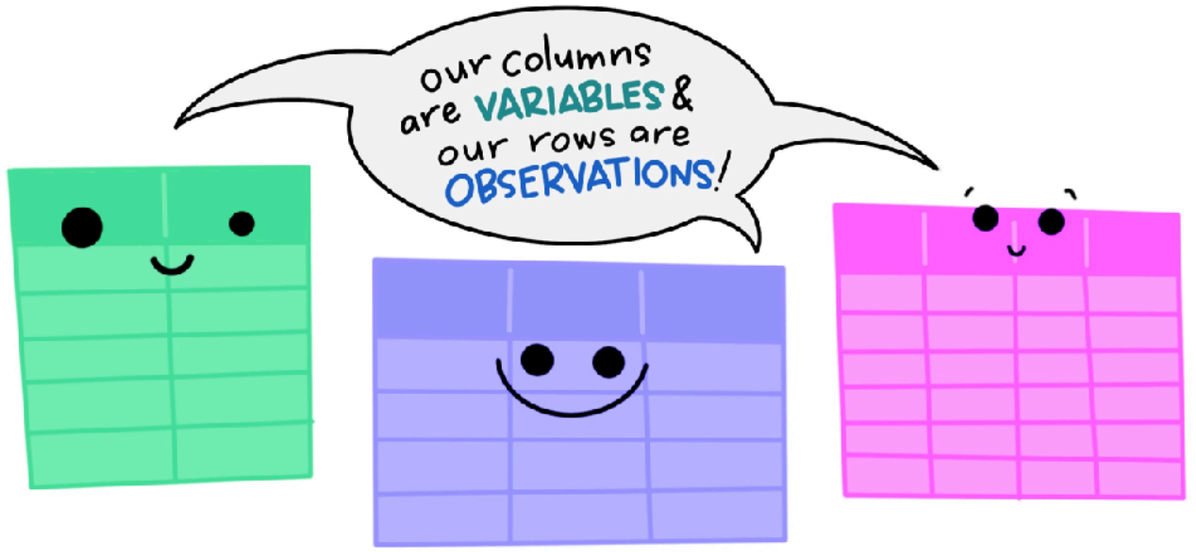
id	name	color
1	floof	gray
2	max	black
3	cat	orange
4	donut	gray
5	merlin	black
6	panda	calico

each row an observation

Wickham, H. (2014). Tidy Data. Journal of Statistical Software 59 (10). DOI: 10.18637/jss.v059.i10

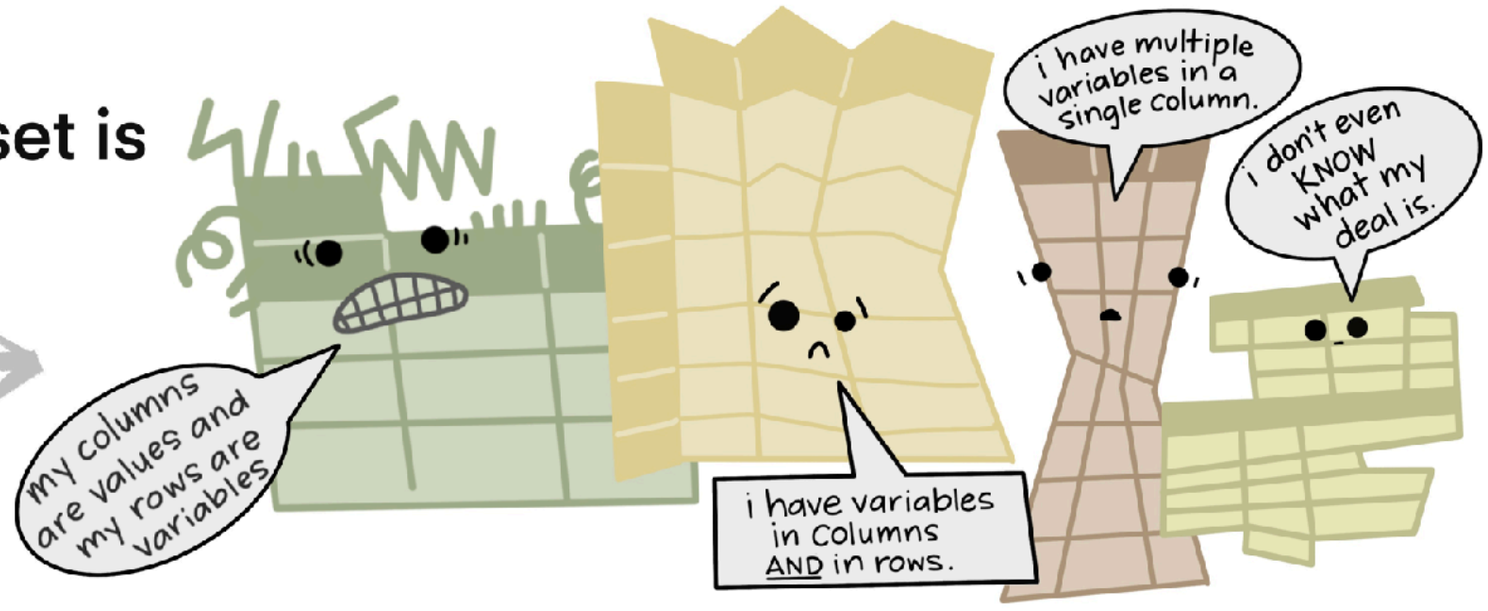
Tidy data

The standard structure of tidy data means that "tidy datasets are all alike..."



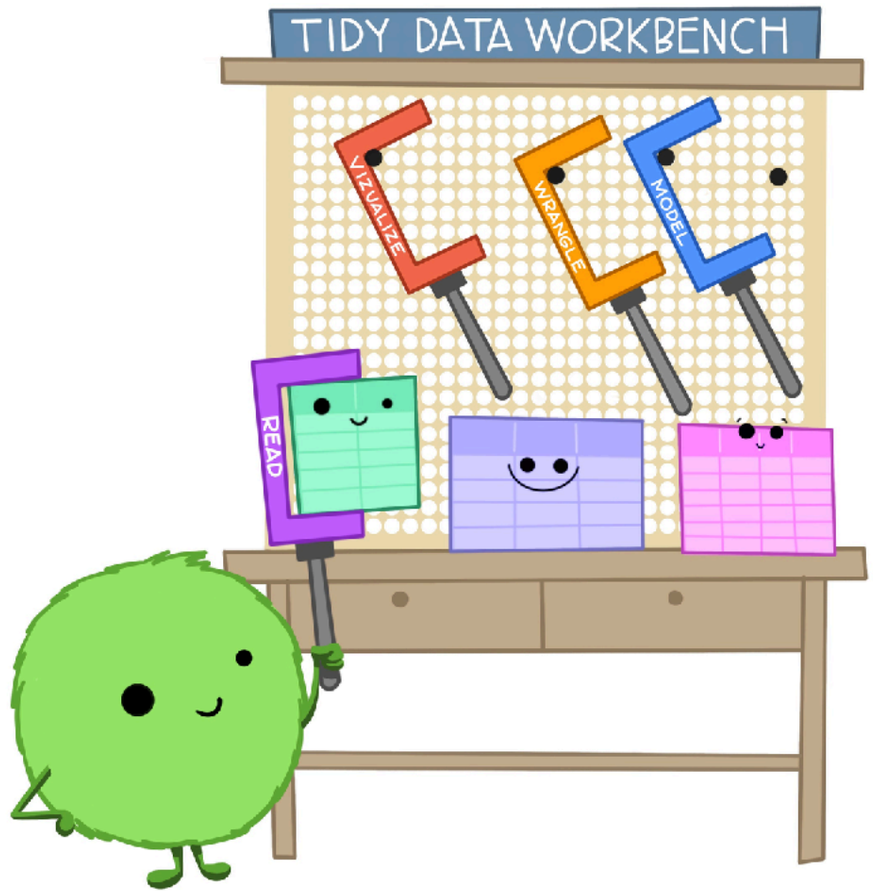
"...but every messy dataset is messy in its own way."

–HADLEY WICKHAM

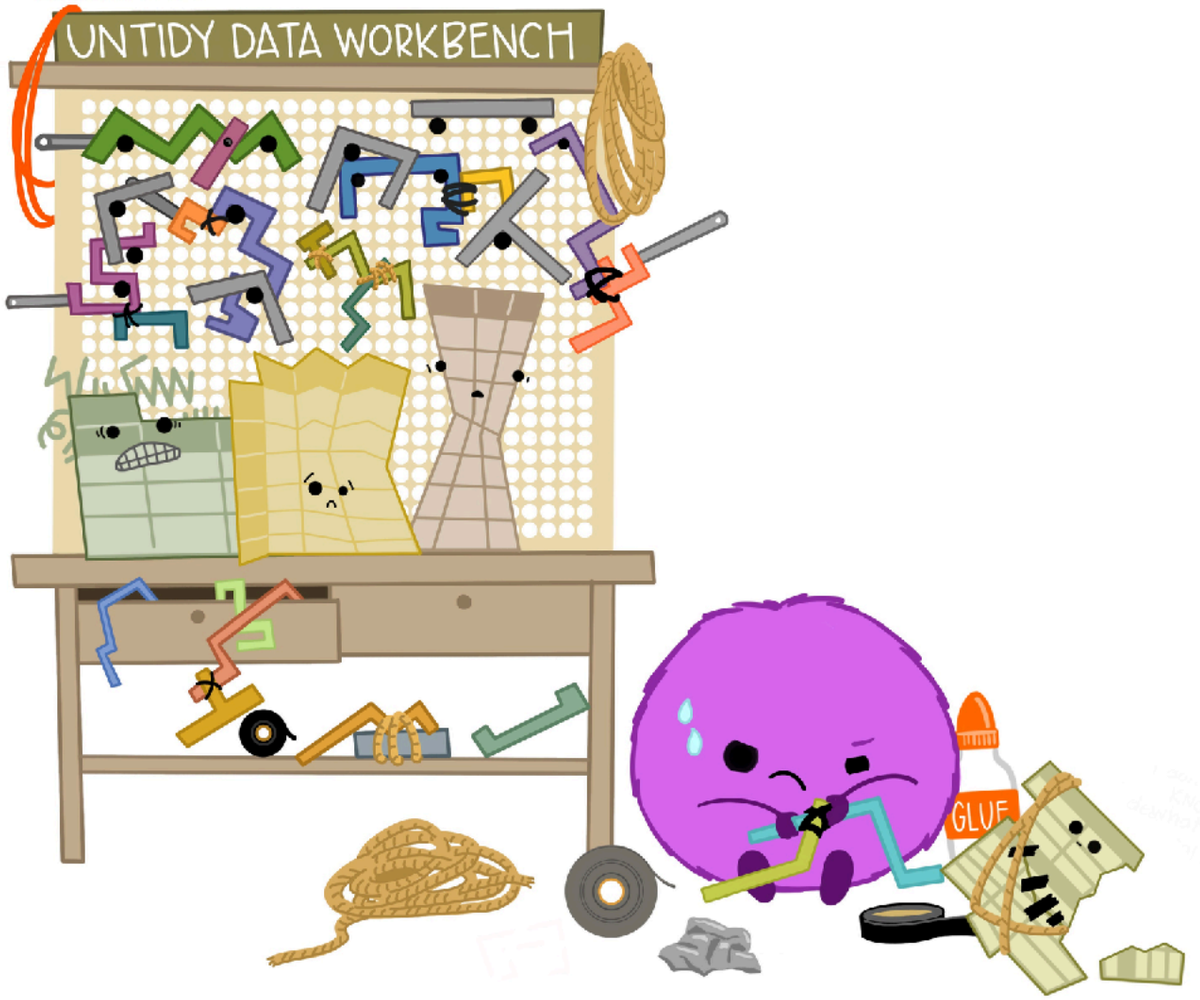


Tidy data

When working with tidy data, we can use the same tools in similar ways for different datasets...



...but working with untidy data often means reinventing the wheel with one-time approaches that are hard to iterate or reuse.



Using functions with `tidyverse` verbs

```
1 df.starwars %>%
2   select(where(fn = is.numeric))
```

my recommendation

- `fn = is.numeric`
- `fn = "is.numeric"`
- `fn = function(x){is.numeric(x)}`
- `fn = ~ is.numeric(.)`

flexible, short, works well with other verbs we'll learn about later

- `fn = ~ !is.numeric(.)`

select all variables that are not numeric

Reprex

- best way to get help is by posting a **reprex**
- **reprex =** reproducible example

reprex

CRAN 0.2.1 | build passing | build passing | codecov 78% | lifecycle stable

Overview

Prepare reprexes for posting to GitHub issues, StackOverflow, or Slack snippets. What is a `reprex`? It's a reproducible example, as coined by Romain Francois.

Given R code on the clipboard, selected in RStudio, as an expression (quoted or not), or in a file ...

- run it via `rmarkdown::render()`,
- with deliberate choices re: arguments and setup chunk.

Get resulting runnable code + output as

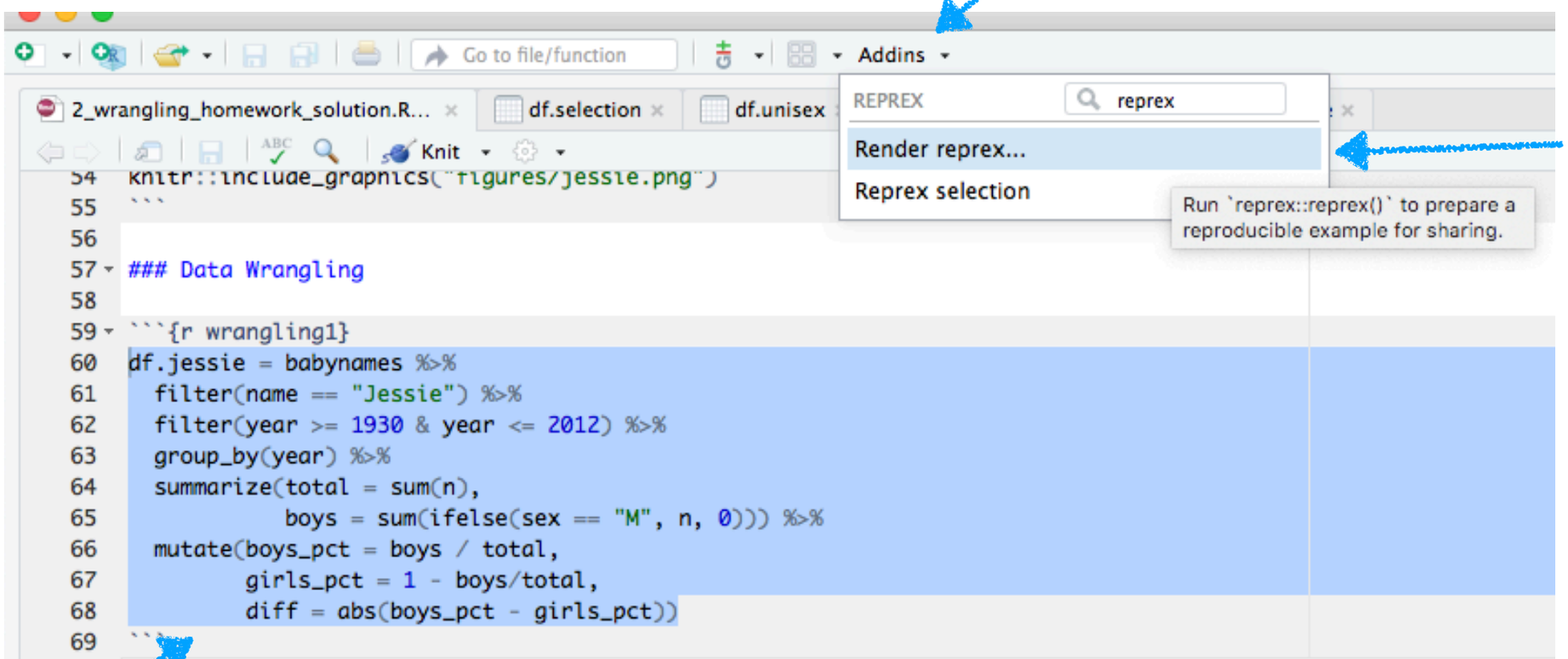
- Markdown, formatted for target venue, e.g. `gh` or `so`, or as
- R code, augmented with commented output.

Result is returned invisibly, placed on the clipboard, and written to a file. Preview an HTML version in RStudio viewer or default browser.

Reprex

```
install.package("reprex")
```

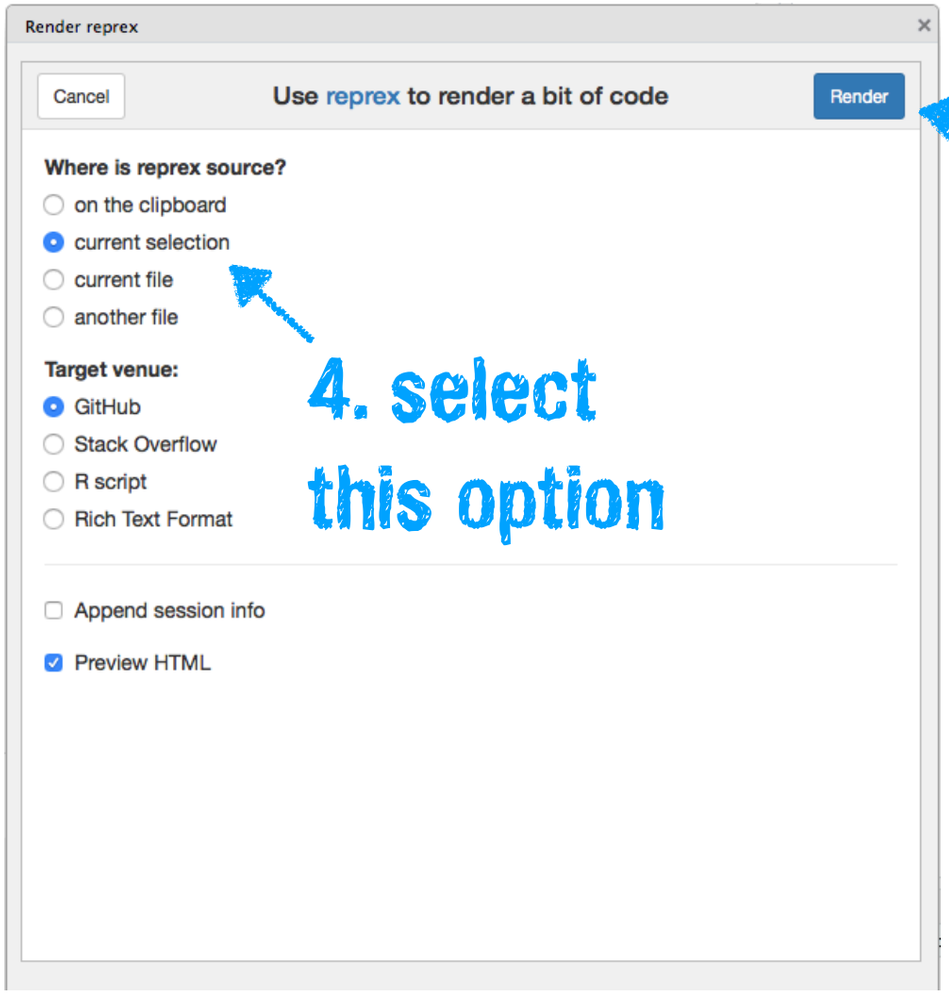
2. click on Addins



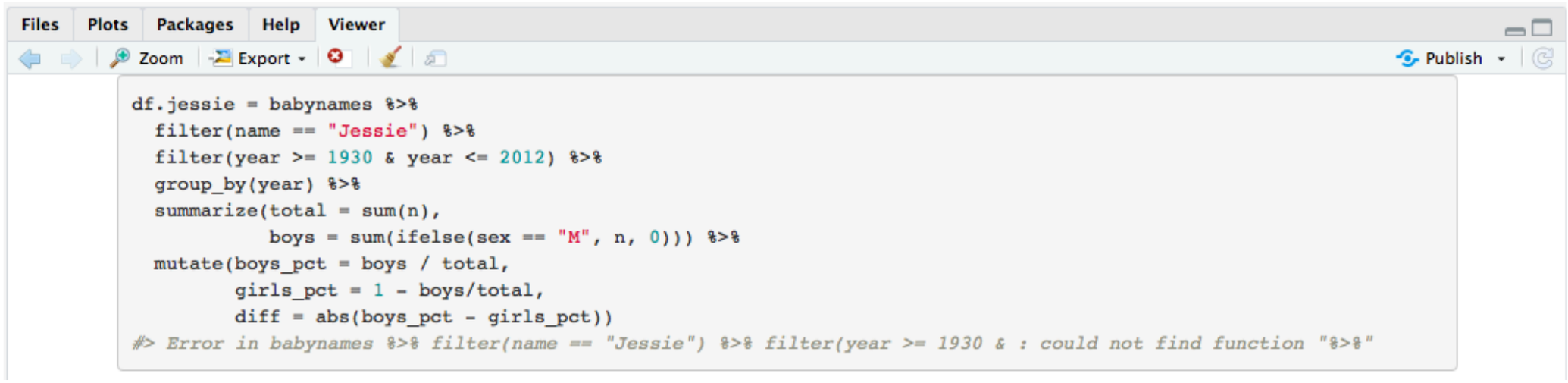
3. Render reprex

1. select the text

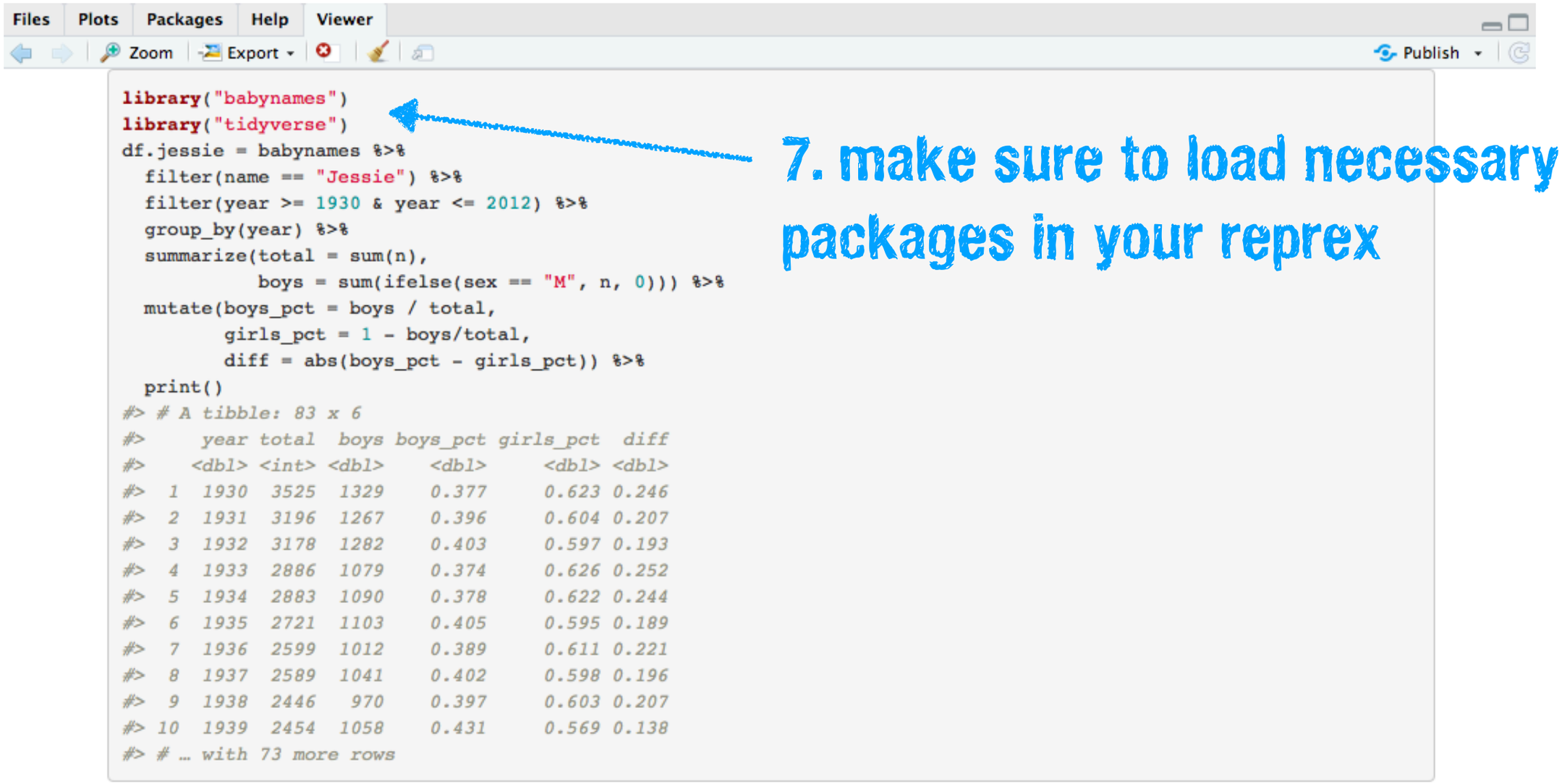
Reprex



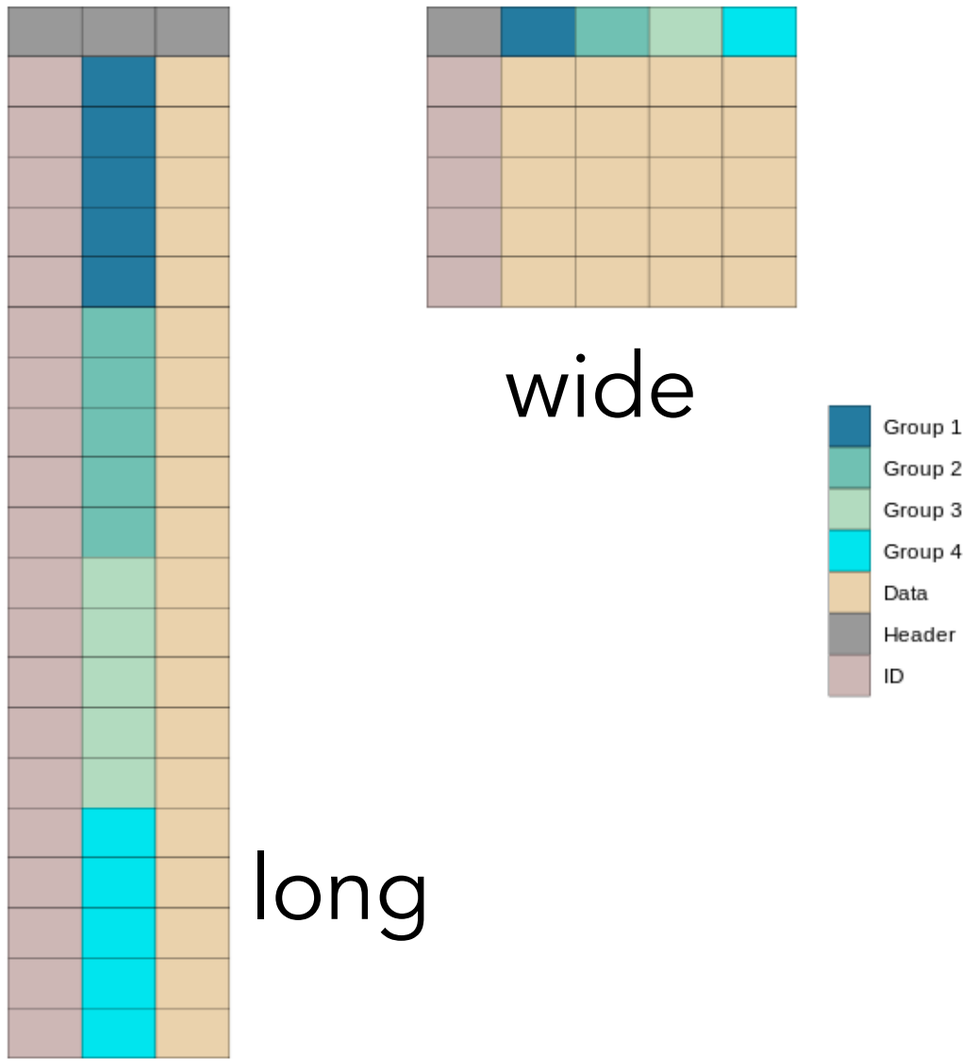
6. copy and paste from the viewer



Reprex

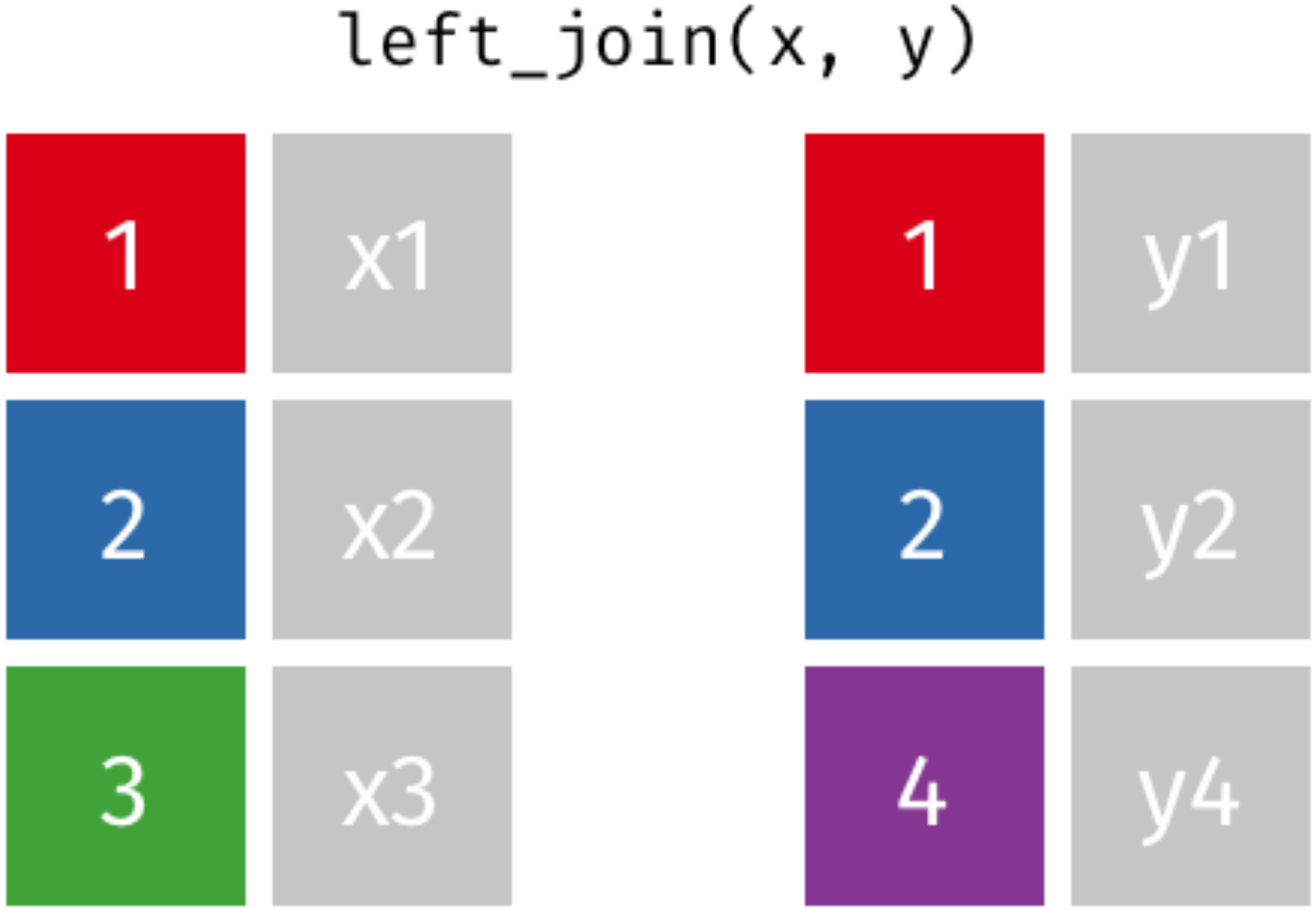

Files
Plots
Packages
Help
Viewer
Zoom
Export
Publish
library("babynames")
library("tidyverse")
df.jessie = babynames %>%
filter(name == "Jessie") %>%
filter(year >= 1930 & year <= 2012) %>%
group_by(year) %>%
summarize(total = sum(n),
boys = sum(ifelse(sex == "M", n, 0))) %>%
mutate(boys_pct = boys / total,
girls_pct = 1 - boys/total,
diff = abs(boys_pct - girls_pct)) %>%
print()
#> # A tibble: 83 x 6
#> year total boys boys_pct girls_pct diff
#> <dbl> <int> <dbl> <dbl> <dbl> <dbl>
#> 1 1930 3525 1329 0.377 0.623 0.246
#> 2 1931 3196 1267 0.396 0.604 0.207
#> 3 1932 3178 1282 0.403 0.597 0.193
#> 4 1933 2886 1079 0.374 0.626 0.252
#> 5 1934 2883 1090 0.378 0.622 0.244
#> 6 1935 2721 1103 0.405 0.595 0.189
#> 7 1936 2599 1012 0.389 0.611 0.221
#> 8 1937 2589 1041 0.402 0.598 0.196
#> 9 1938 2446 970 0.397 0.603 0.207
#> 10 1939 2454 1058 0.431 0.569 0.138
#> # … with 73 more rows
7. make sure to load necessary packages in your reprex
Created on 2019-01-24 by the reprex package (v0.2.1)

pivot_longer()

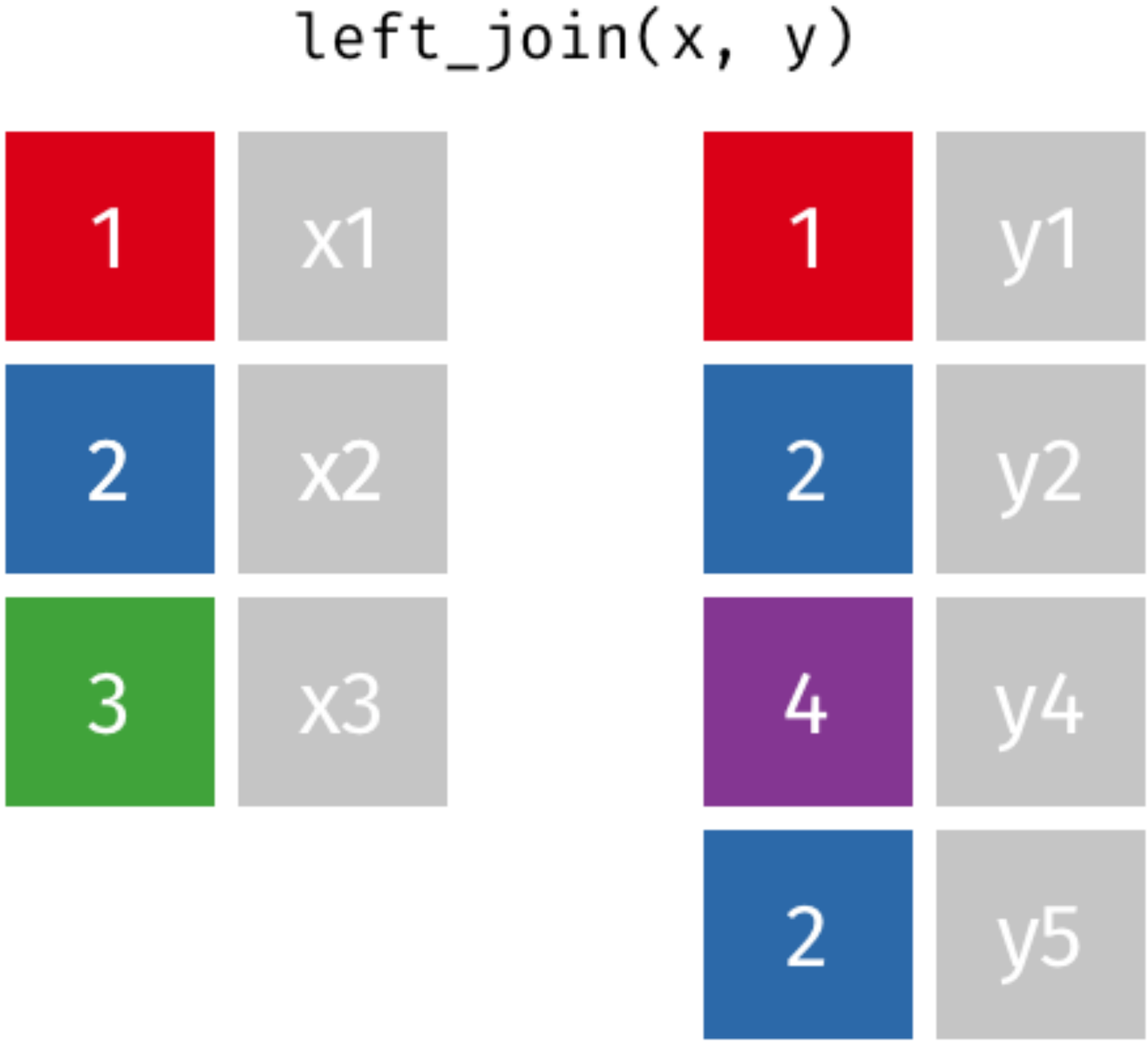


pivot_wider()

left_join()



left_join()



Timer

05 : 00

I'm done.

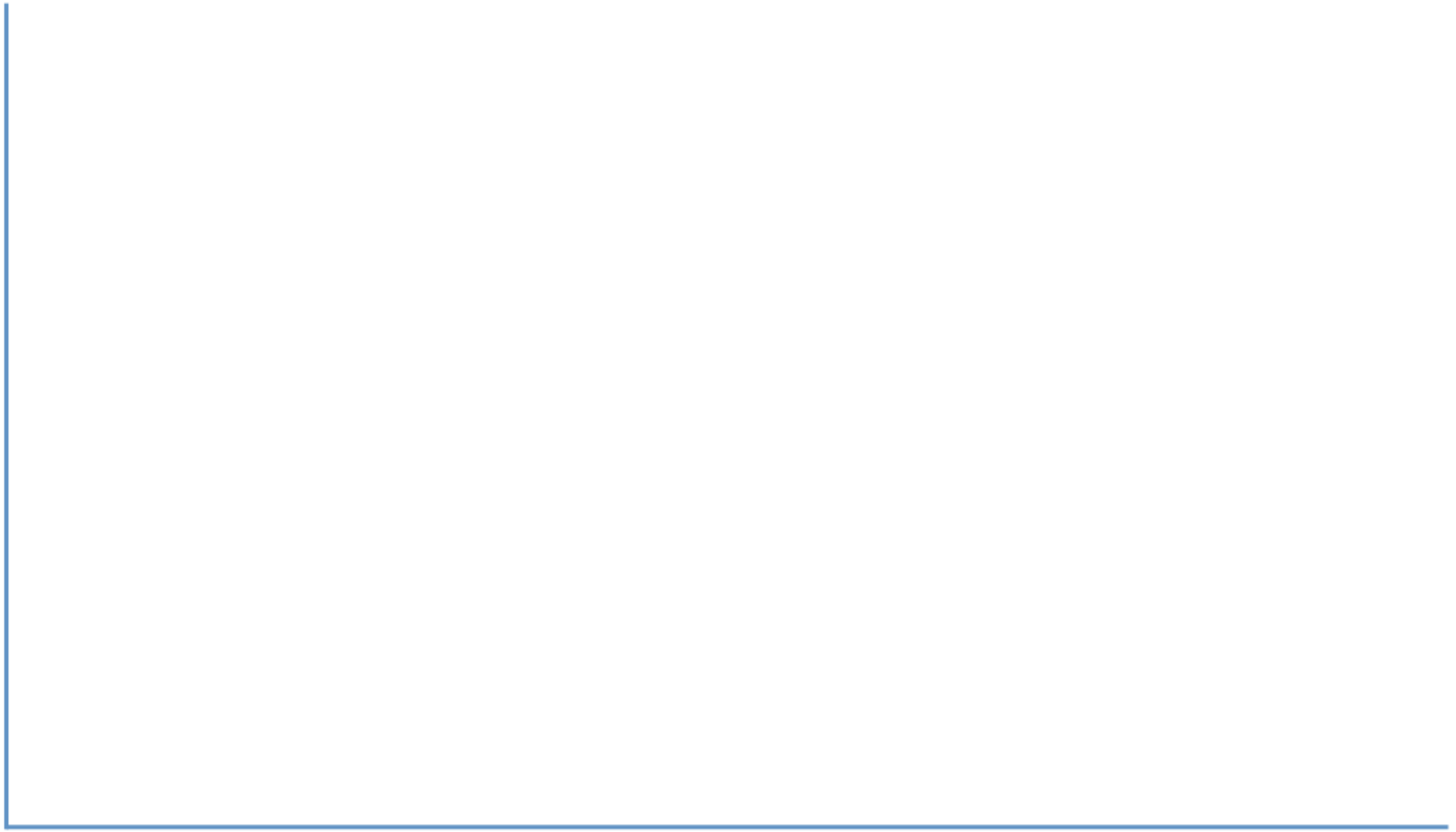
blue

Please help.

pink

Feedback

How was the pace of today's class?



How happy were you with today's class overall?

What did you like about today's class? What could be improved next time?

Thank you!