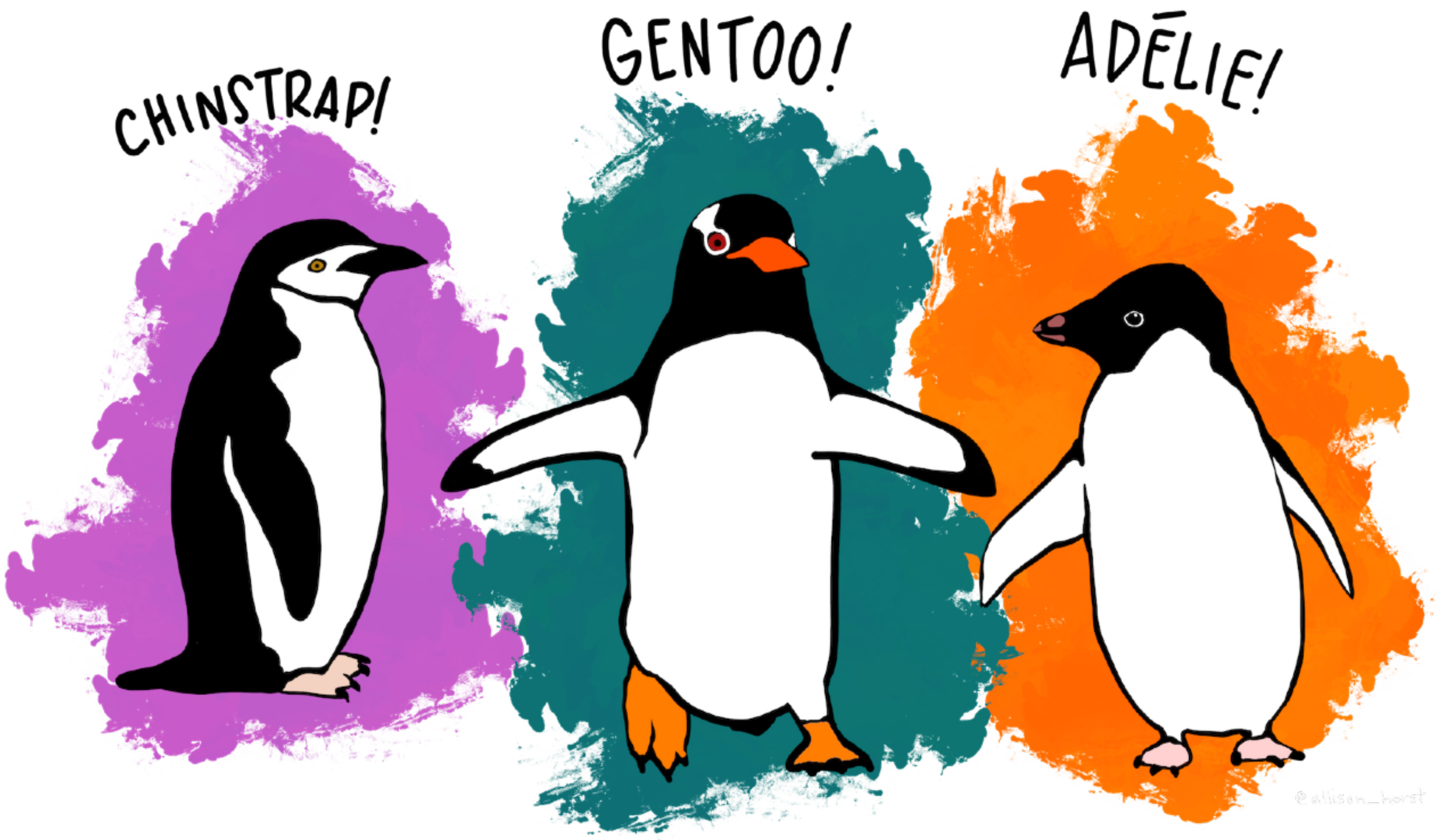
Teaching modeling in introductory statistics:
A comparison of formula and tidyverse syntaxes

Amelia McNamara
University of St Thomas

https://arxiv.org/abs/2201.12960



Horst AM, Hill AP, Gorman KB (2020). palmerpenguins: Palmer Archipelago (Antarctica) penguin data. R package version 0.1.0. https://allisonhorst.github.io/palmerpenguins/

Artwork by @allison_horst

```
library(palmerpenguins)
data("penguins")
```

Base syntax

```
mean(penguins$body_mass_g,
       na.rm = TRUE)
```

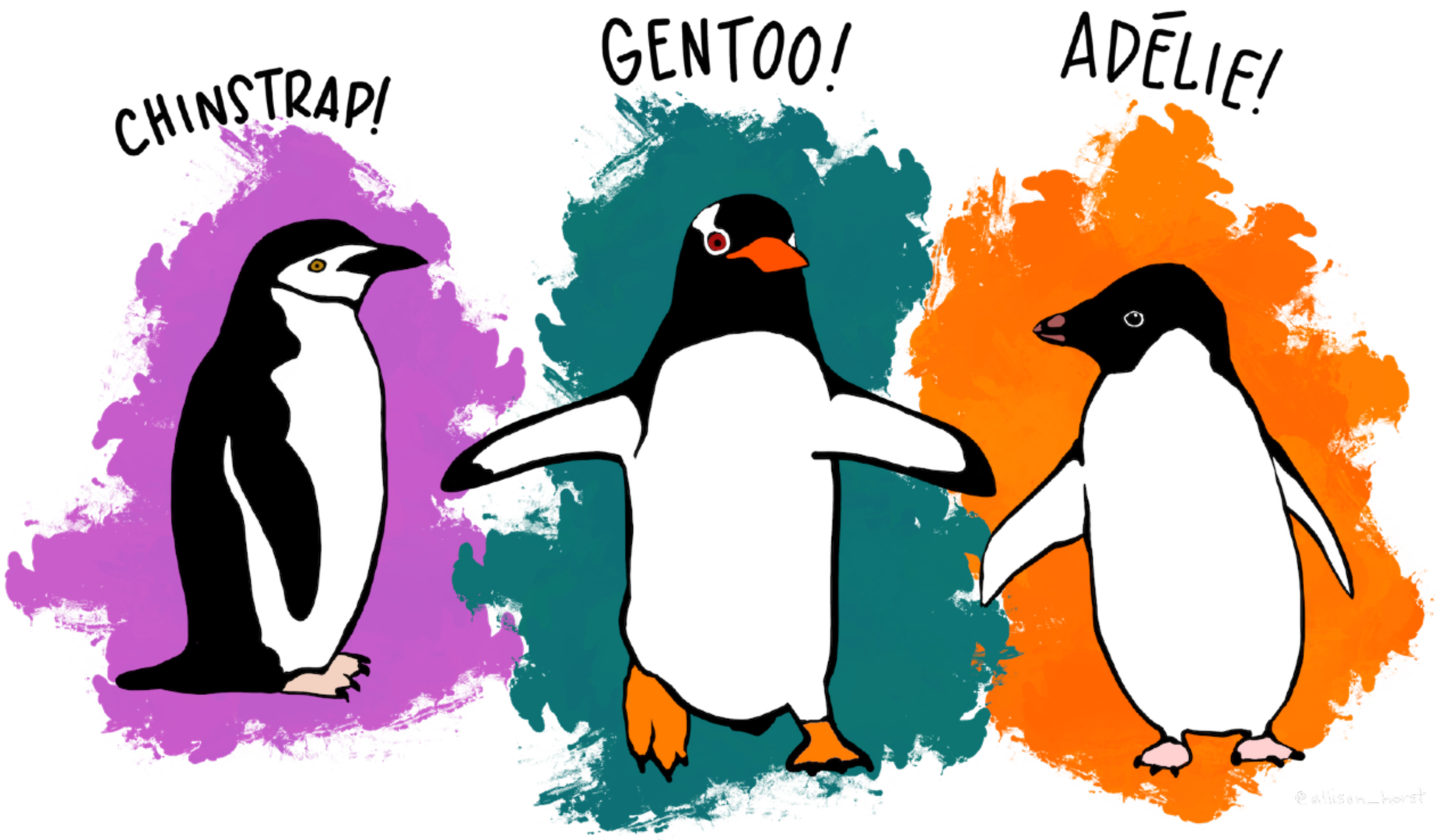
Formula syntax

```
library(mosaic)
mean(~body_mass_g, data = penguins,
       na.rm = TRUE)
```

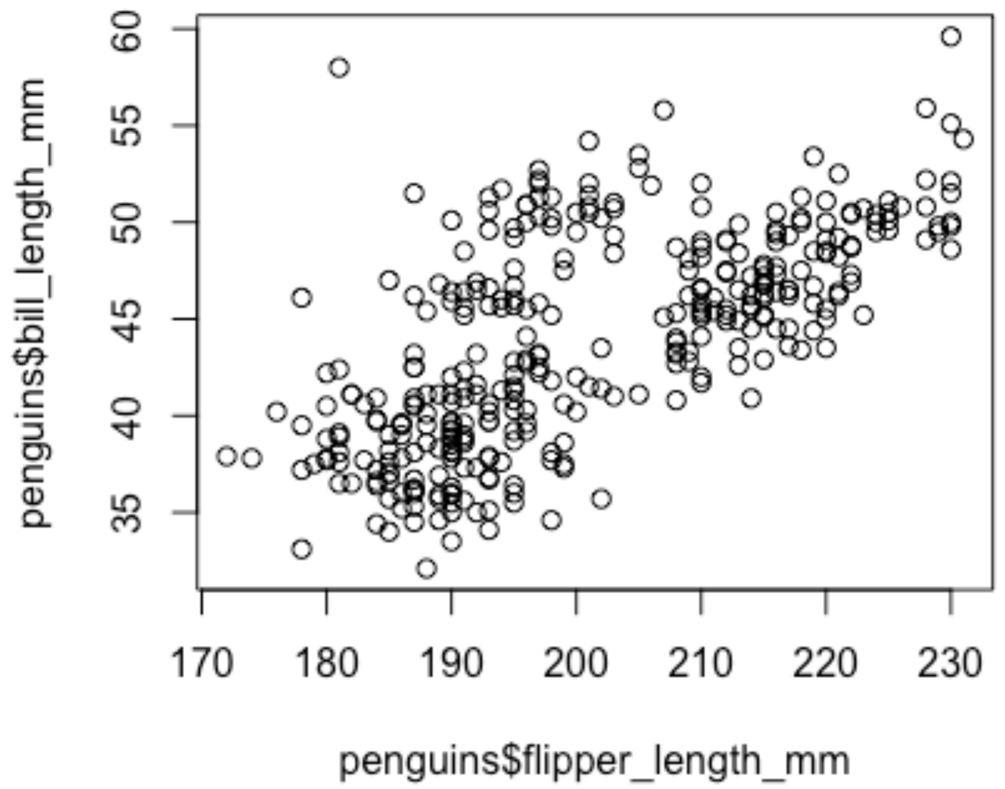
Tidyverse syntax

```
library(tidyverse)
penguins %>%
  drop_na(body_mass_g) %>%
  summarize(mean(body_mass_g))
```

Syntax cheatsheet available from the RStudio contributed cheatsheets page

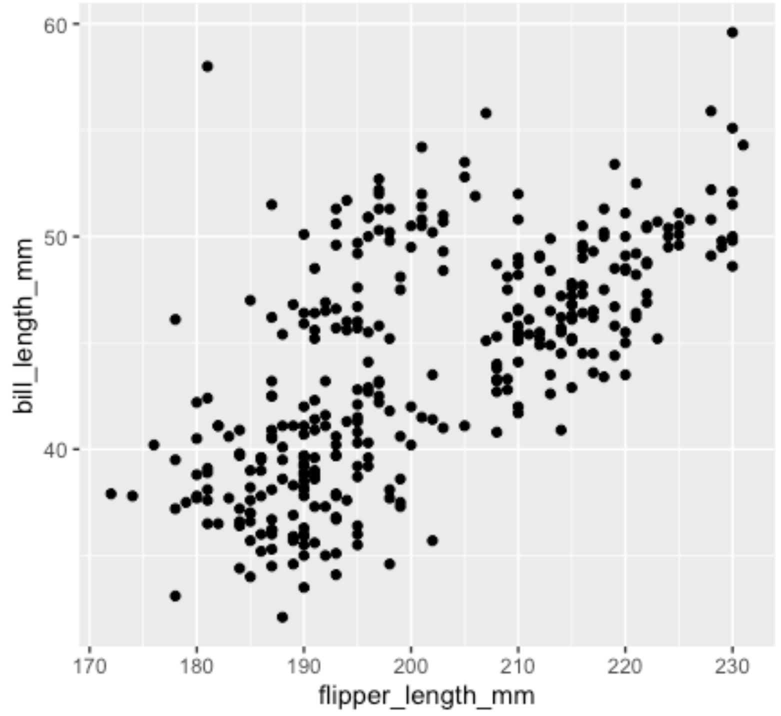


```
library(palmerpenguins)
data("penguins")
```



Base syntax

```
plot(penguins$flipper_length_mm,
     penguins$bill_length_mm)
```

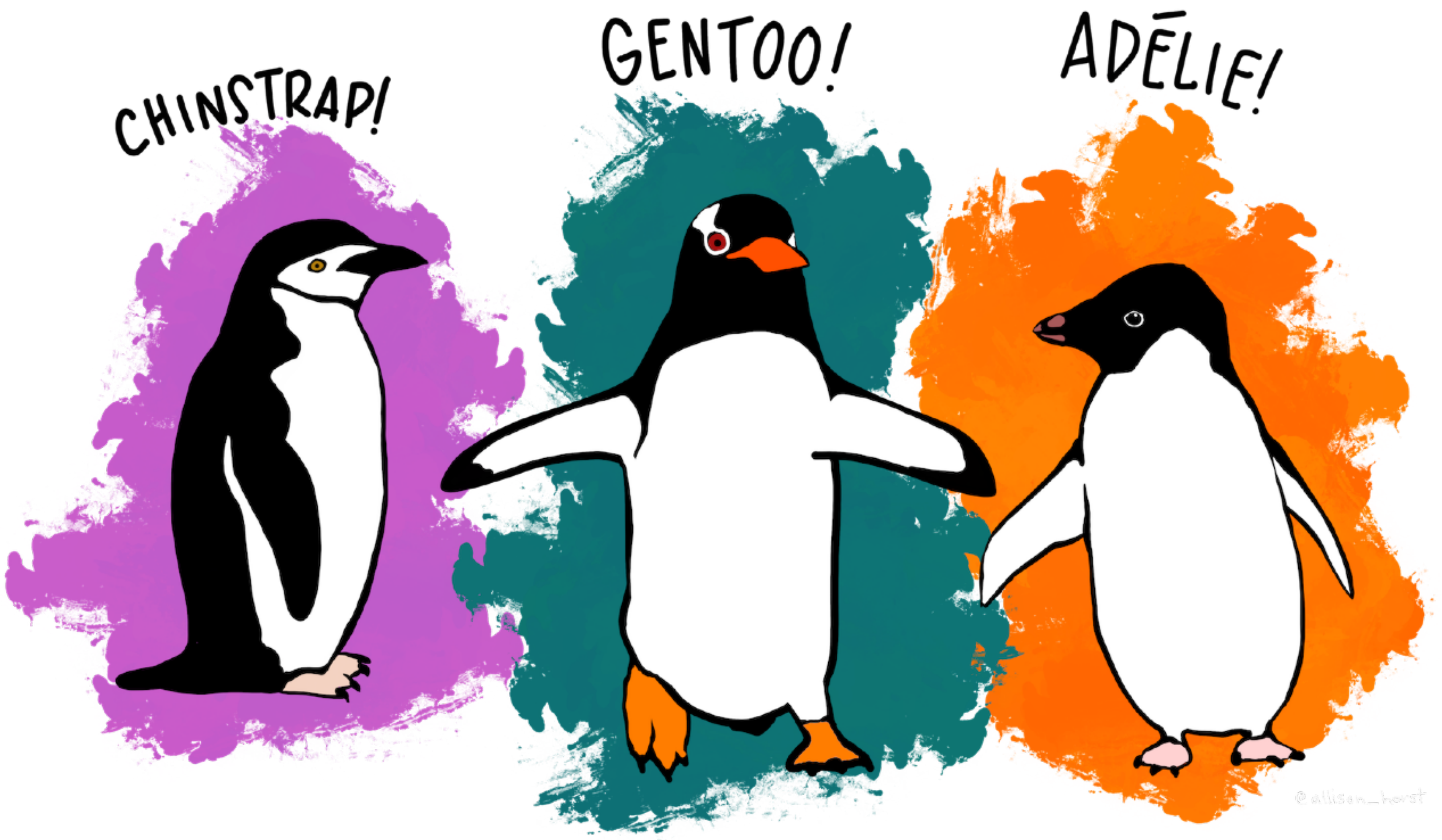


Tidyverse syntax

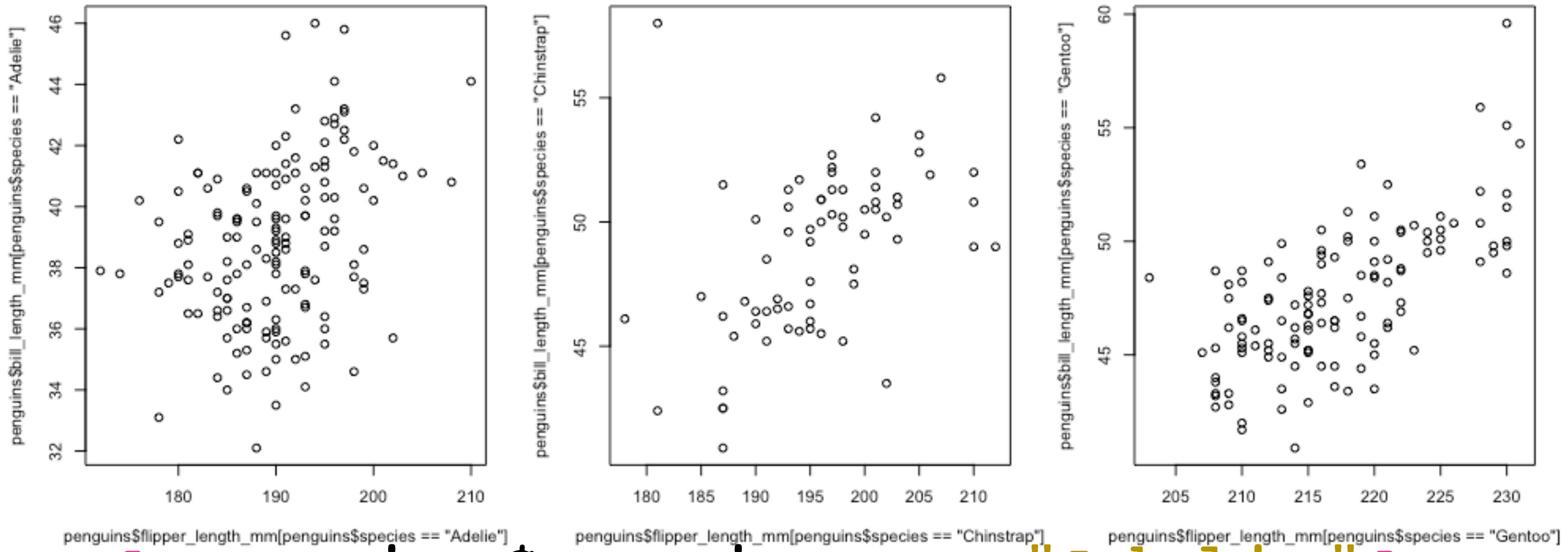
```
ggplot(penguins) +
    geom_point(aes(x = flipper_length_mm,
                   y = bill_length_mm))
```

Formula syntax

```
gf_point(bill_length_mm ~ flipper_length_mm,
         data = penguins)
```



```r
library(palmerpenguins)
data("penguins")
```



Base syntax

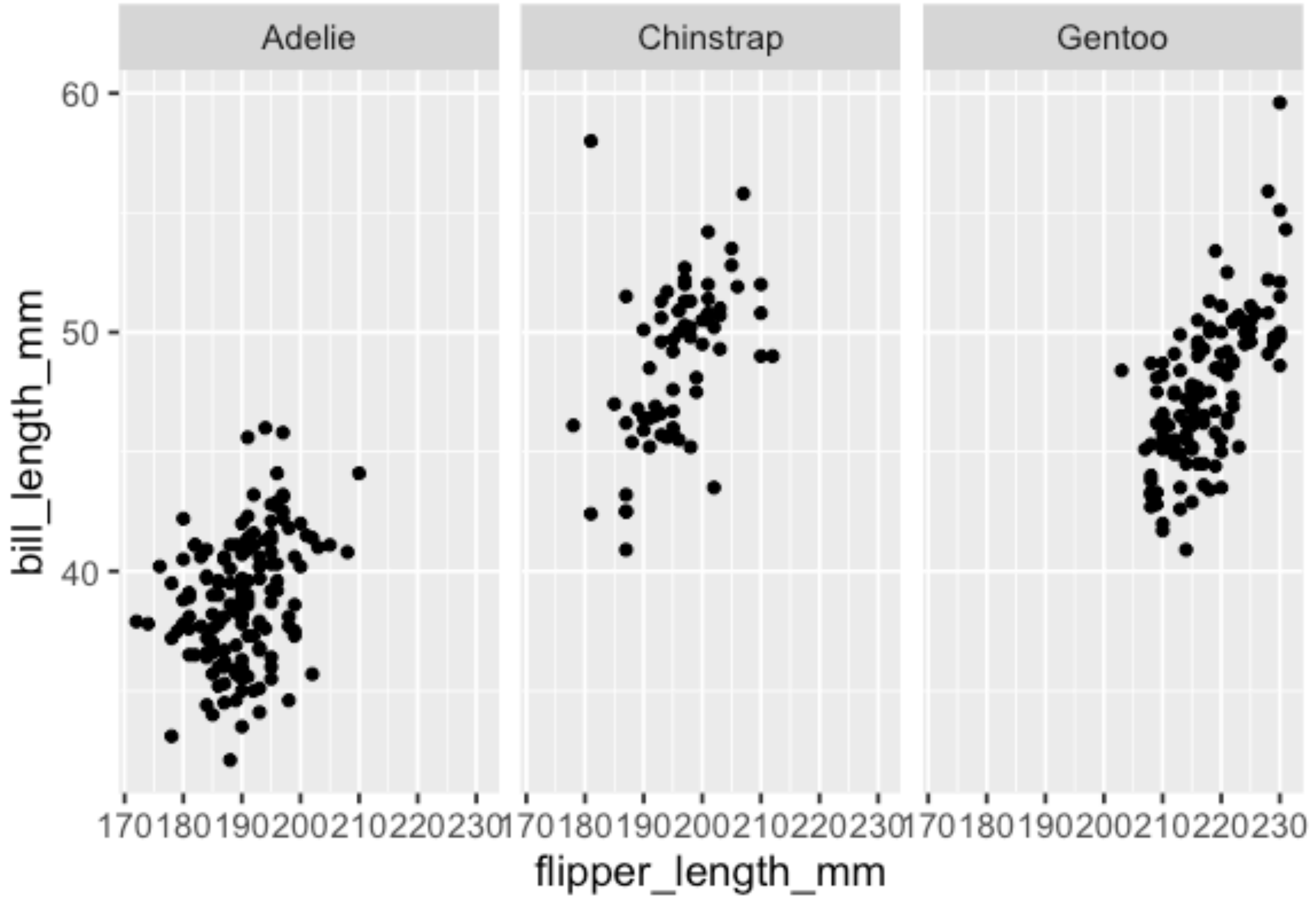
```r
par(mfrow = c(1, 3))
plot(penguins$flipper_length_mm[penguins$species == "Adelie"],
    penguins$bill_length_mm[penguins$species == "Adelie"])
plot(penguins$flipper_length_mm[penguins$species == "Chinstrap"],
    penguins$bill_length_mm[penguins$species == "Chinstrap"])
plot(penguins$flipper_length_mm[penguins$species == "Gentoo"],
     penguins$bill_length_mm[penguins$species == "Gentoo"])
```

Formula syntax

```r
gf_point(bill_length_mm ~ flipper_length_mm | species,
         data = penguins)
```

Tidyverse syntax

```r
ggplot(penguins, aes(x = flipper_length_mm,
                     y = bill_length_mm)) +
    geom_point() +
    facet_grid(~species)
```



Head-to-head comparison

- Students enrolled in the same lecture class (60-90 students)
- Lecture was broken into three smaller sections for lab material
- I taught two of the sections, and both were designated as using R
- Using random assignment (coin flip) I chose one to use **tidyverse syntax** and one to use **formula syntax**
- Lots of data:
  - Pre- and post-survey
  - RMarkdown documents and associated code
  - YouTube analytics
  - RStudio Cloud analytics

https://arxiv.org/abs/2201.12960

Both sections

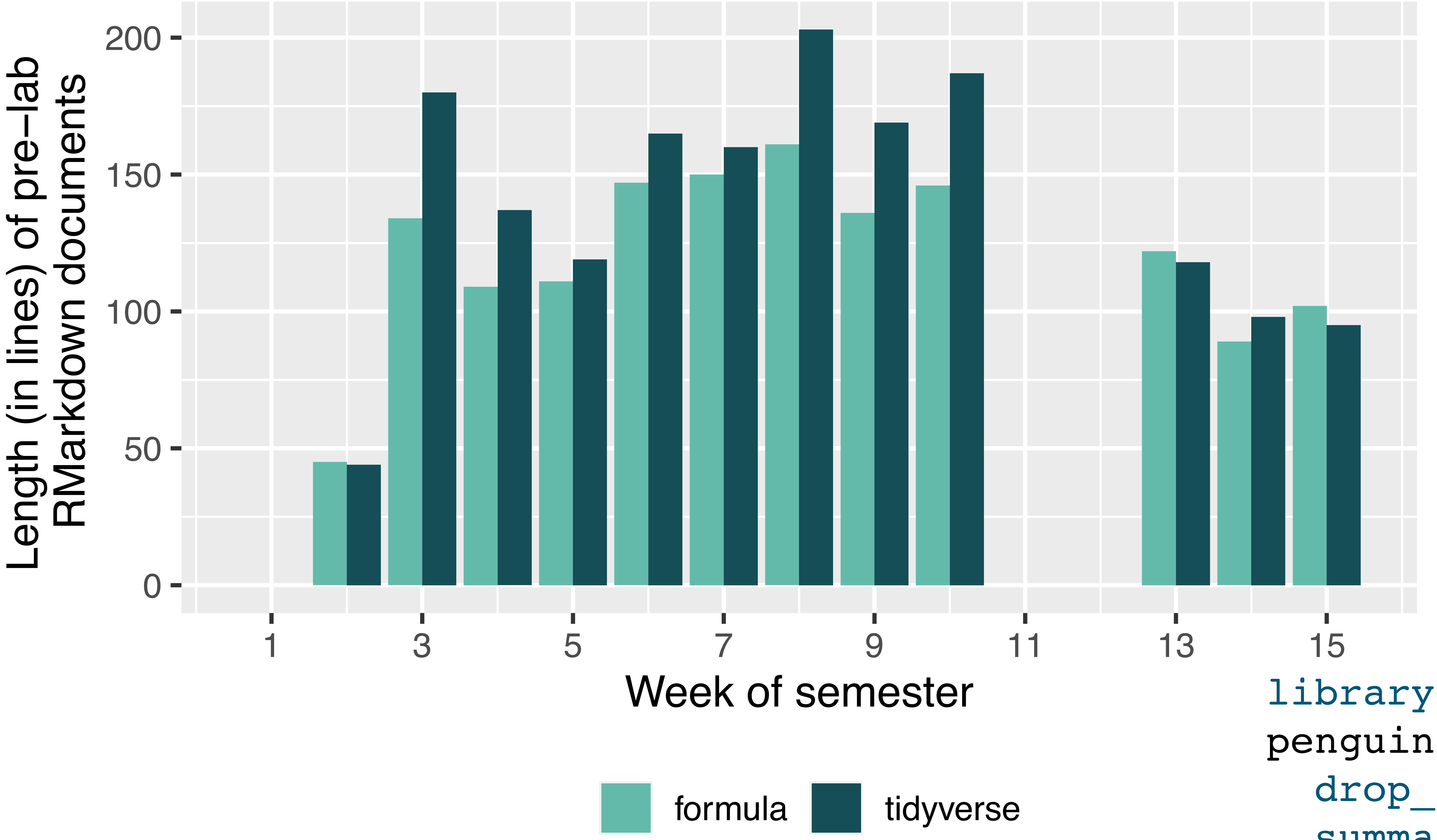
- Consisted of 21 students (fewer took pre/post survey)
- Were comprised mostly of Business majors
- Had similar prior programming experience
- Were given a pre-lab RMarkdown document and associated YouTube video(s) for the material of the week
- Met synchronously to ask questions on the real lab assignment
- Completed the actual lab in a templated RMarkdown document

Prior programming experience

	formula	tidyverse
No	10	9
Yes, but not with R	2	4

https://arxiv.org/abs/2201.12960

tidyverse labs slightly longer



Length of pre-lab documents (in lines) each week.

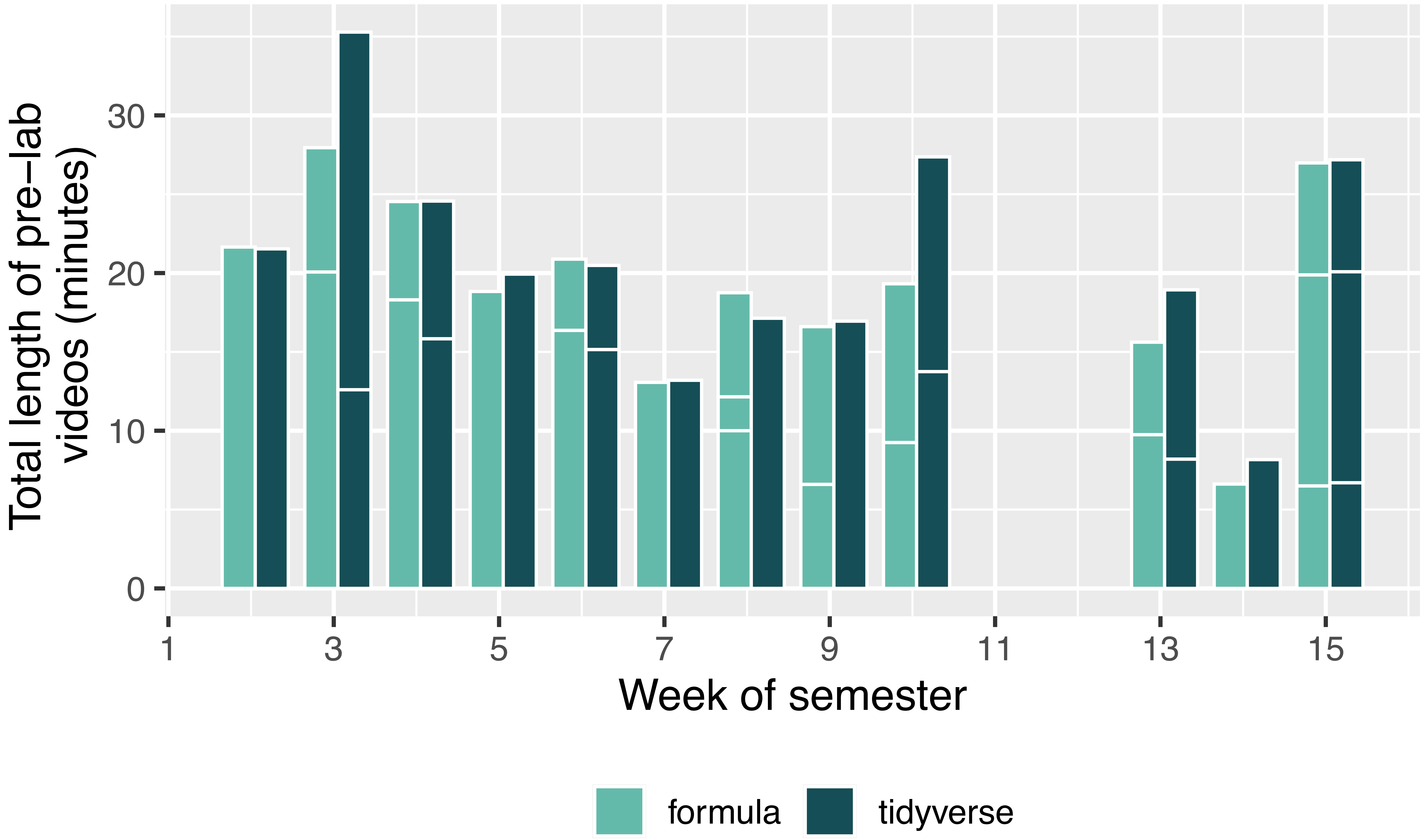
Tidyverse labs tended to be longer, on average 16 lines longer or 18% longer. This makes sense given how the tidyverse is written.

```
library(tidyverse)
penguins %>%
  drop_na(body_mass_g) %>%
  summarize(mean(body_mass_g))
```

```
library(mosaic)
mean(~body_mass_g, data = penguins, na.rm = TRUE)
```

https://arxiv.org/abs/2201.12960

tidyverse labs slightly longer

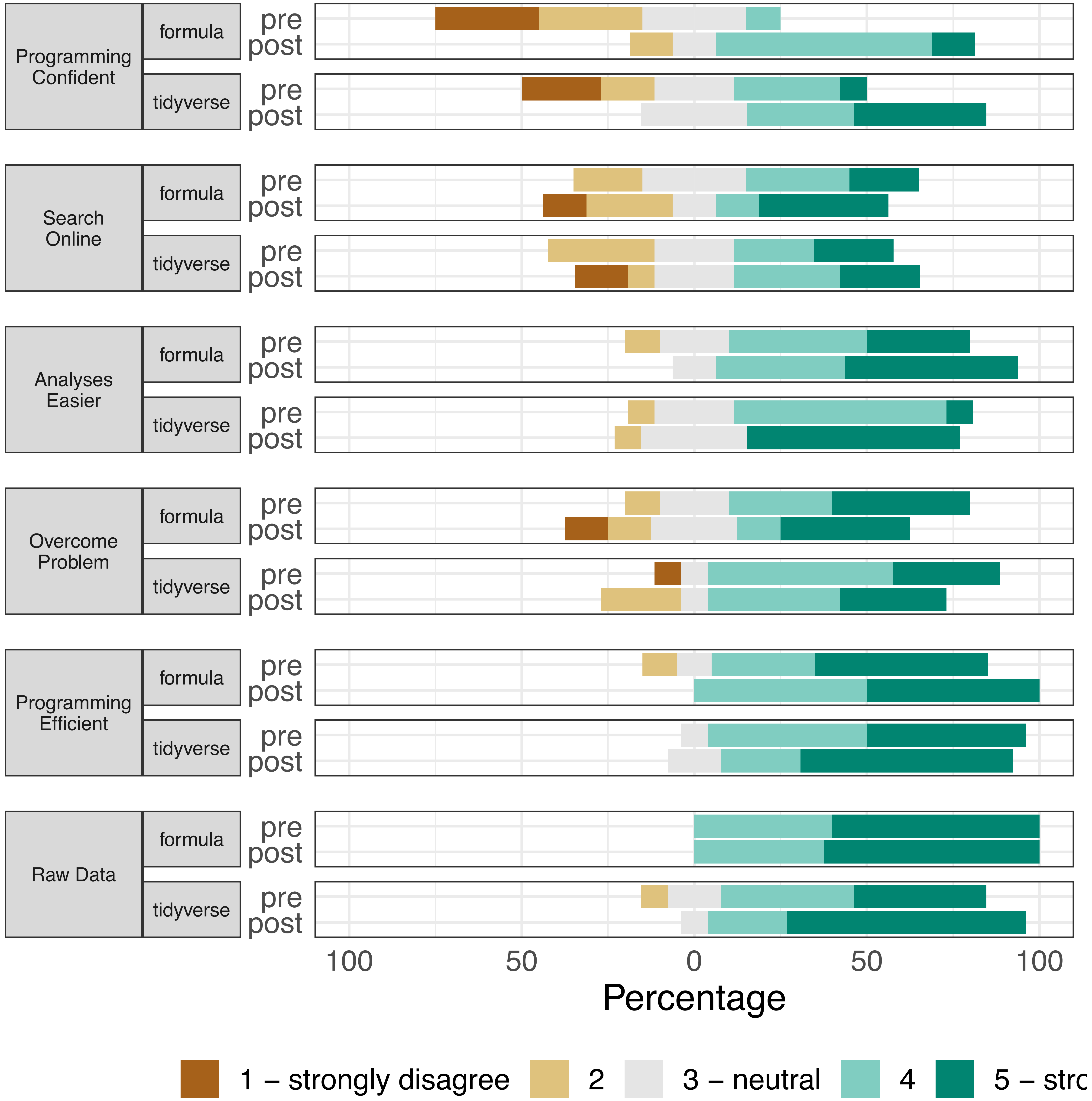


Length of pre-lab videos each week. Outlines help delineate multiple videos for a single week.

Again, tidyverse videos tended to be slightly longer, but only slightly! 2 minutes longer on average, or 9% longer.

https://arxiv.org/abs/2201.12960

Pre/post survey mostly inconclusive

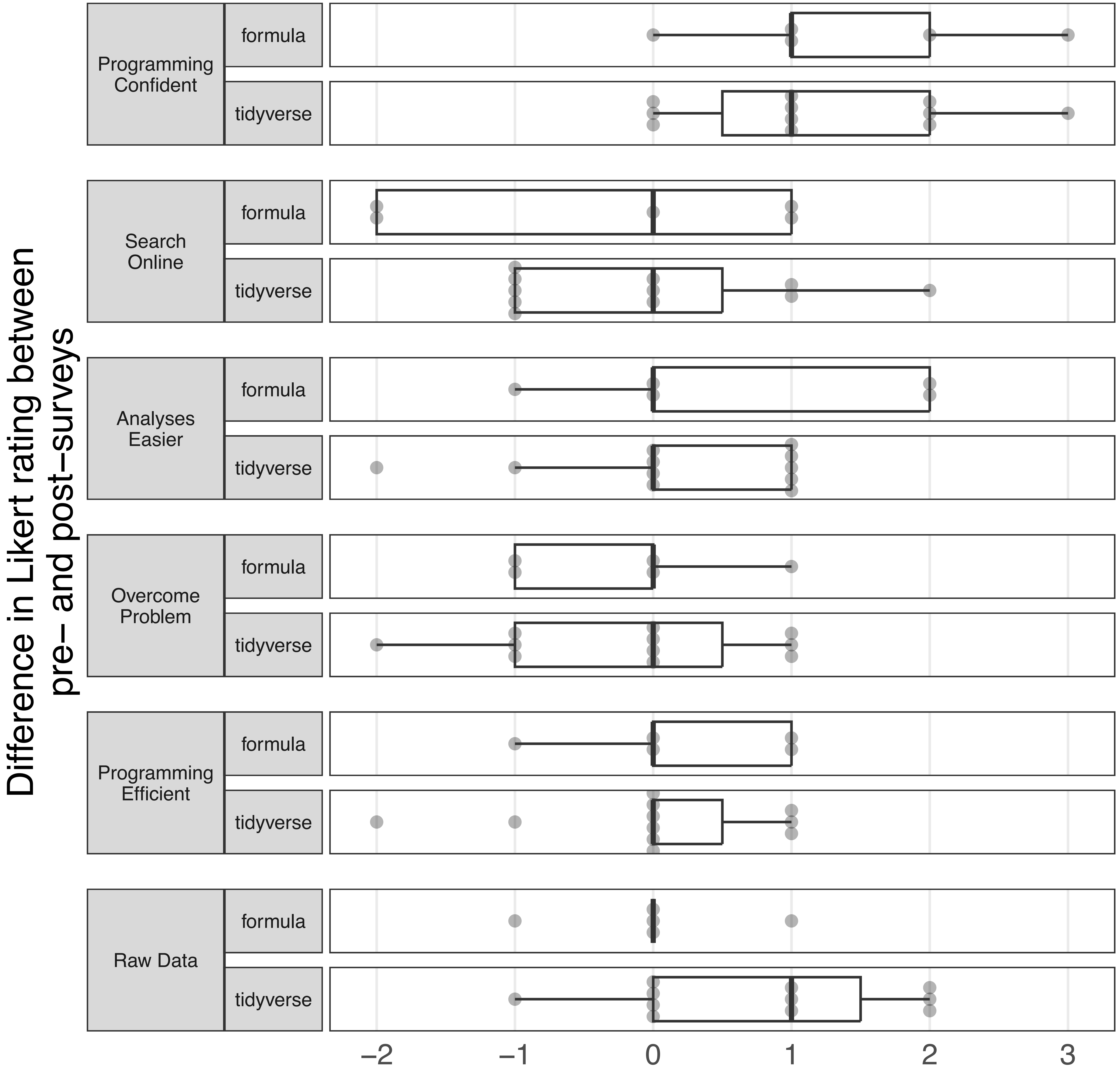


Pre and post responses to Likert-scale questions. Most questions show some level of improvement, such as the first question, ‘I am confident in my ability to make use of programming software to work with data.’ but others show no change or even a decline in agreement.

Questions from The Carpentries probably weren’t appropriate for this class and context.

https://arxiv.org/abs/2201.12960

Pre/post survey mostly inconclusive

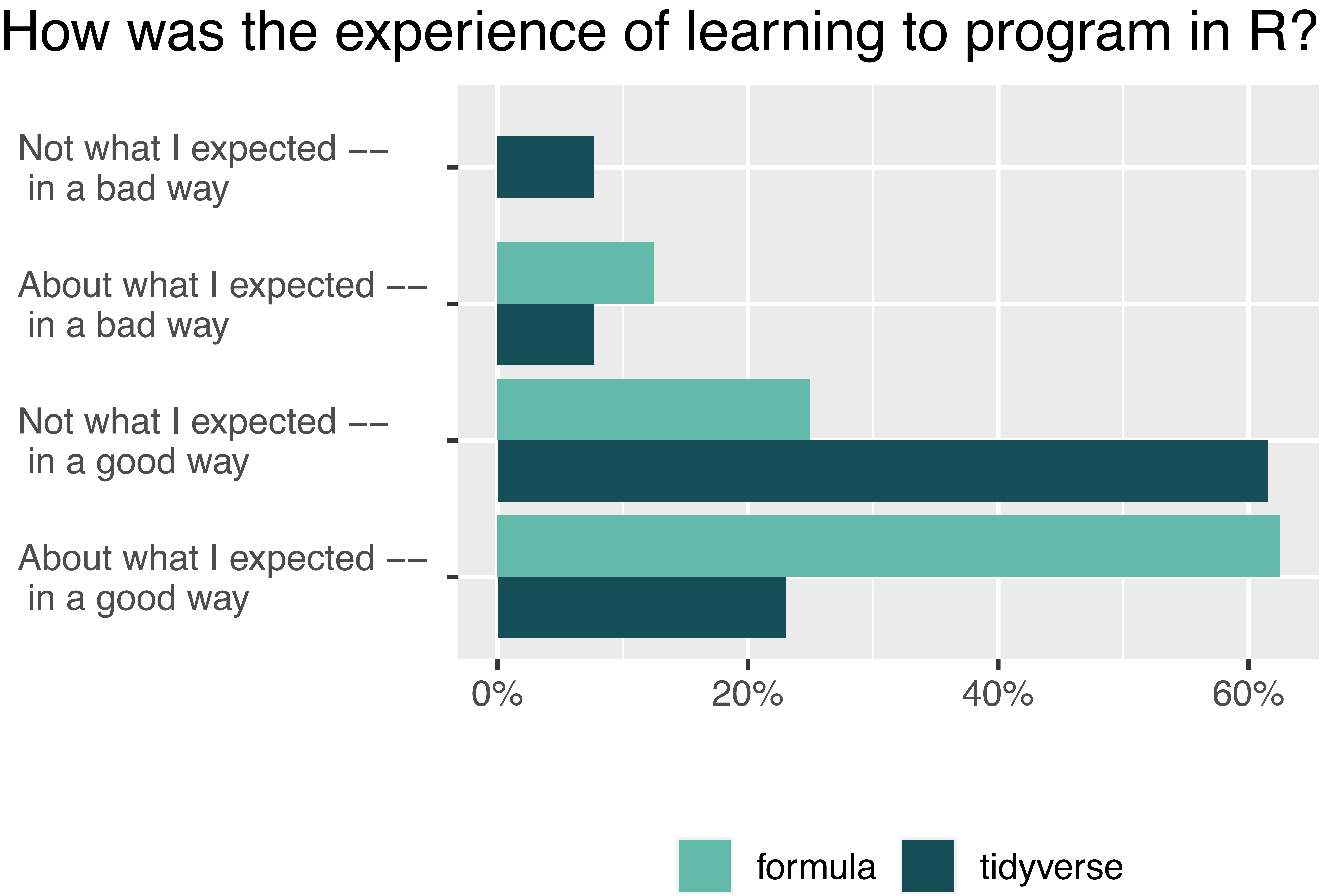


Pairing helps show some differences better.

Questions from The Carpentries probably weren't appropriate for this class and context.

https://arxiv.org/abs/2201.12960

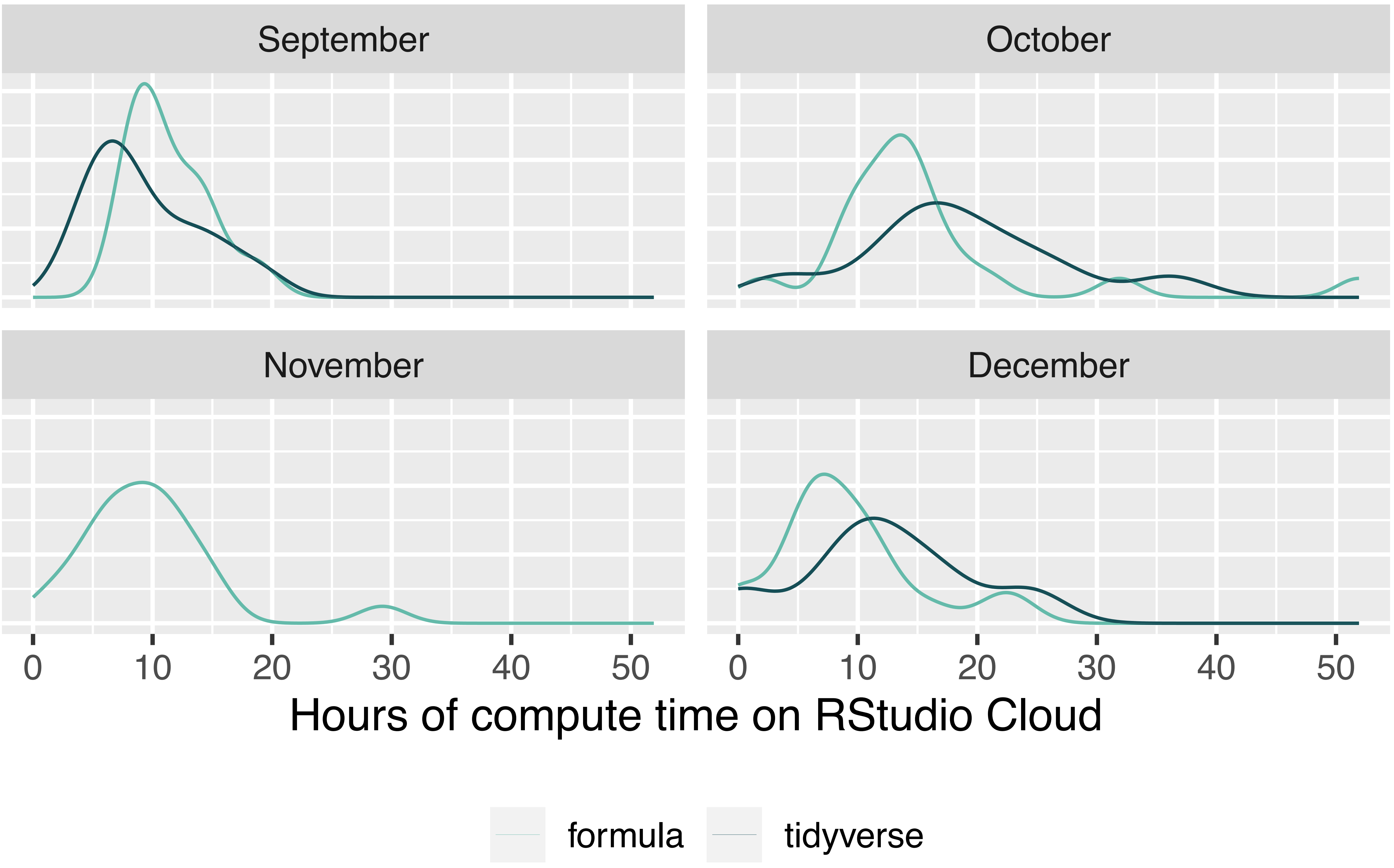
Overall: students don't hate R

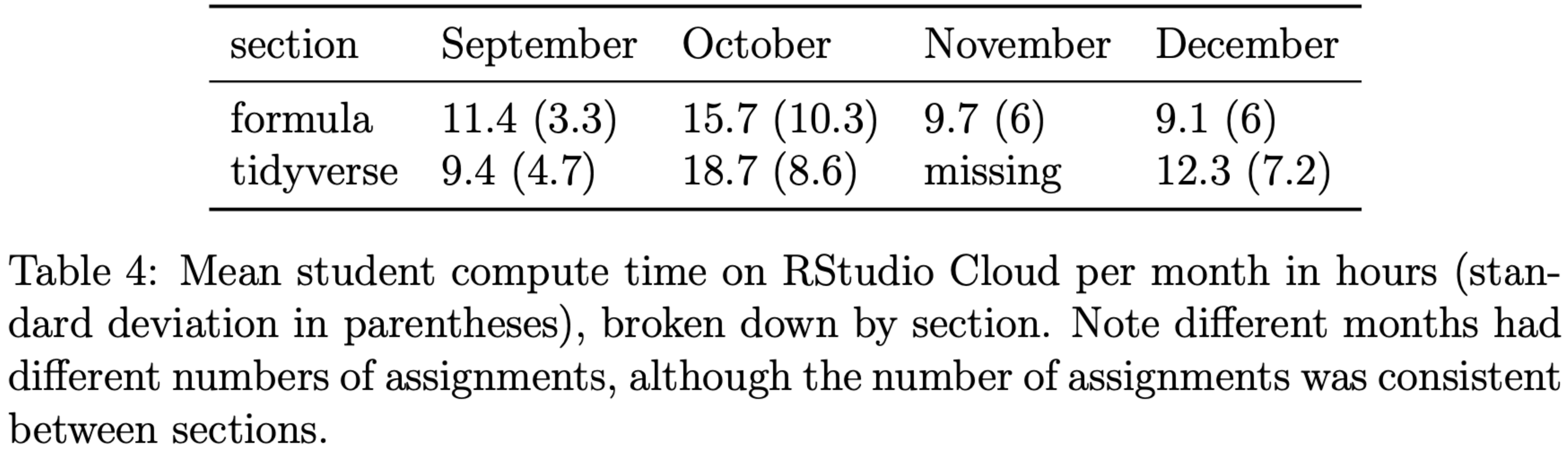


Responses to the question, "How was the experience of learning to program in R?"

https://arxiv.org/abs/2201.12960

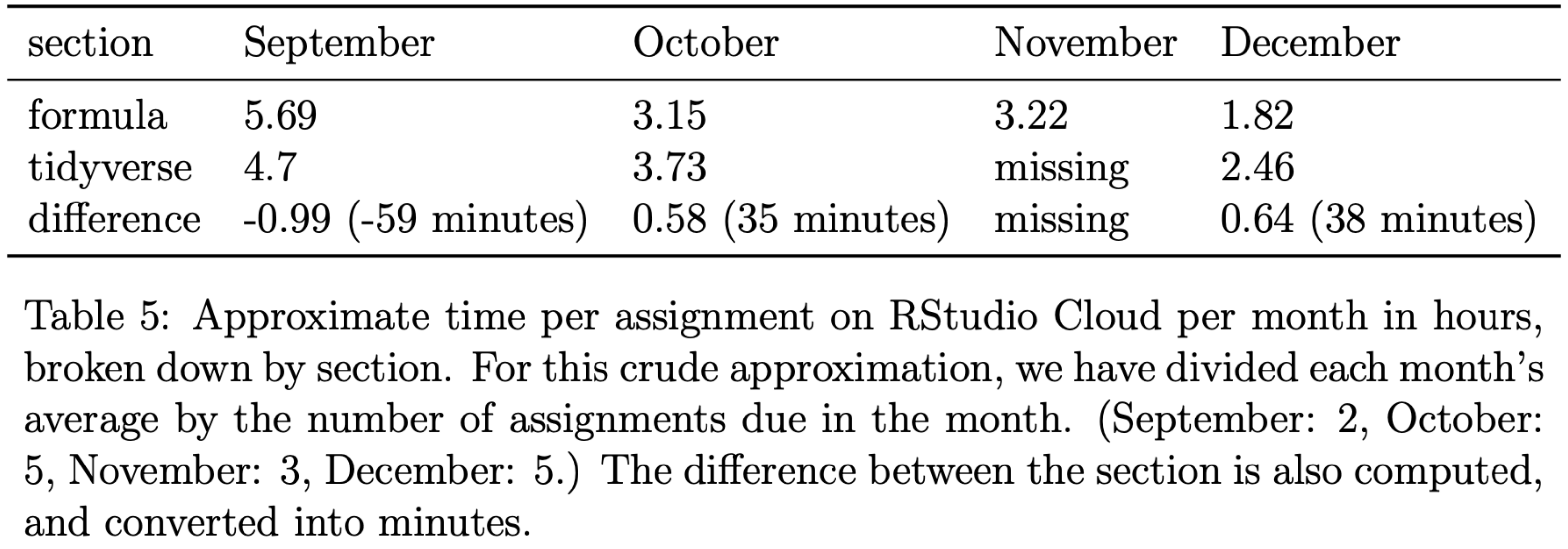
Compute time was different





section	September	October	November	December
formula	11.4 (3.3)	15.7 (10.3)	9.7 (6)	9.1 (6)
tidyverse	9.4 (4.7)	18.7 (8.6)	missing	12.3 (7.2)

Table 4: Mean student compute time on RStudio Cloud per month in hours (standard deviation in parentheses), broken down by section. Note different months had different numbers of assignments, although the number of assignments was consistent between sections.



section	September	October	November	December
formula	5.69	3.15	3.22	1.82
tidyverse	4.7	3.73	missing	2.46
difference	-0.99 (-59 minutes)	0.58 (35 minutes)	missing	0.64 (38 minutes)

Table 5: Approximate time per assignment on RStudio Cloud per month in hours, broken down by section. For this crude approximation, we have divided each month's average by the number of assignments due in the month. (September: 2, October: 5, November: 3, December: 5.) The difference between the section is also computed, and converted into minutes.

https://arxiv.org/abs/2201.12960

Slight difference in number of functions

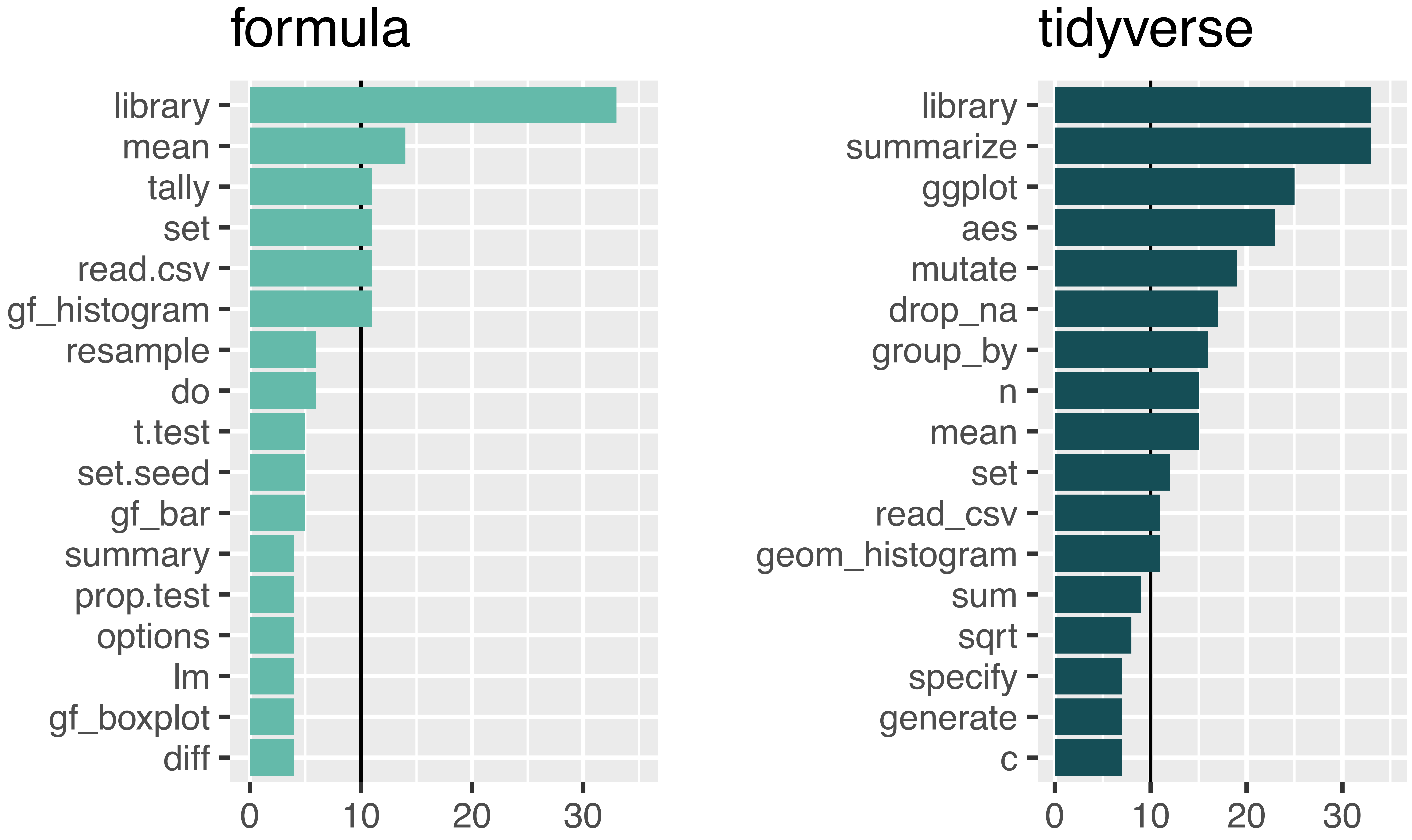
The formula section saw a total of **41 functions** and the tidyverse section saw **52**, with an **overlap of 21 functions** between the two sections.

Neither of these numbers are very large!

The functions both sections of students saw included helper functions like `library()`, `set.seed()`, and `set()` (a function in the knitr options included in the top of each RMarkdown document), statistics like `mean()`, `sd()`, and `cor()`, and modeling-related functions like `aov()`, `lm()`, `summary()` and `predict()`.

https://arxiv.org/abs/2201.12960

Slight difference in number of functions



No big surprises here, has to do with how code is written

https://arxiv.org/abs/2201.12960

Challenges/differences

- Summary statistics for two categorical variables (`tally()` versus `group_by()` and `summarize()`)

```
tally(species ~ island, data = penguins, format
= "percent")
#>            island
#> species        Biscoe     Dream Torgersen
#>   Adelie     26.19048  45.16129 100.00000
#>   Chinstrap   0.00000  54.83871   0.00000
#>   Gentoo     73.80952   0.00000   0.00000
```

- Summary statistics for quantitative variables (NA behavior)

```
penguins %>%
  group_by(island, species) %>%
  summarize(n = n()) %>%
  mutate(prop = n / sum(n))
#> # A tibble: 5 × 4
#> # Groups:   island [3]
#>   island    species       n  prop
#>   <fct>     <fct>     <int> <dbl>
#> 1 Biscoe    Adelie       44 0.262
#> 2 Biscoe    Gentoo      124 0.738
#> 3 Dream     Adelie       56 0.452
#> 4 Dream     Chinstrap    68 0.548
#> 5 Torgersen Adelie       52 1
```

- Inference for two categorical variables (`mosiac::prop.test()` versus `infer::prop_test()`)

Big takeaways

- Consider syntax
- Be consistent!
- Try counting the functions you show students
-

Materials are available

- https://arxiv.org/abs/2201.12960
- https://github.com/AmeliaMN/ComparingSyntaxForModeling
- https://github.com/AmeliaMN/STAT220-labs