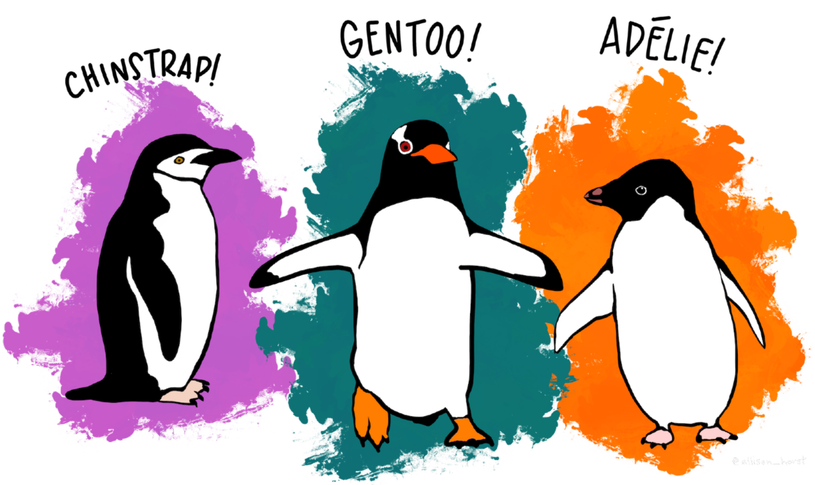
Teaching modeling in introductory statistics: A comparison of formula and tidyverse syntaxes

Amelia McNamara
University of St Thomas

https://arxiv.org/abs/2201.12960



Artwork by @allison_horst

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

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

Base syntax

```
mean(penguins$body_mass_g,
     na.rm = TRUE)
#> [1] 4201.754
```

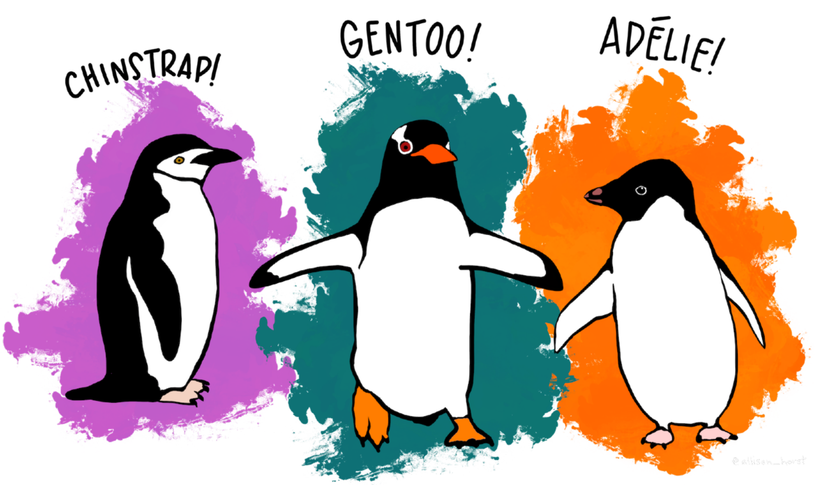
Formula syntax

```
library(mosaic)
mean(~body_mass_g, data = penguins,
     na.rm = TRUE)
#> [1] 4201.754
```

Tidyverse syntax

```
library(tidyverse)
penguins %>%
  drop_na(body_mass_g) %>%
  summarize(mean(body_mass_g))
#> # A tibble: 1 × 1
#>   `mean(body_mass_g)`
#>                 <dbl>
#> 1                4202.
```

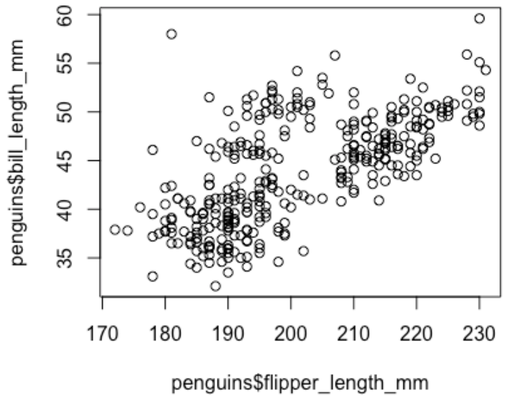
Syntax cheatsheet available from the RStudio contributed cheatsheets page

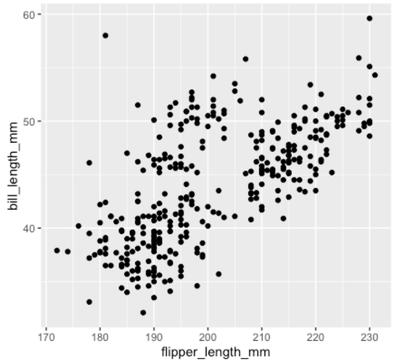



```
library(palmerpenguins)
data("penguins")
```

Base syntax

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



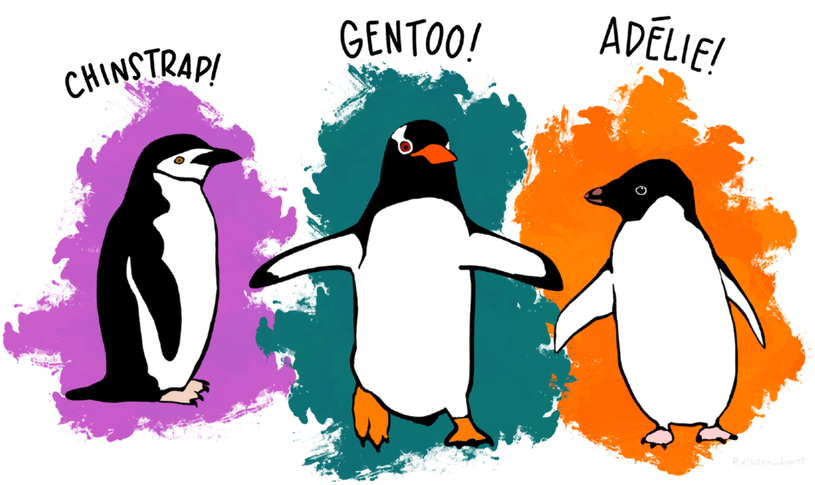


Tidyverse syntax

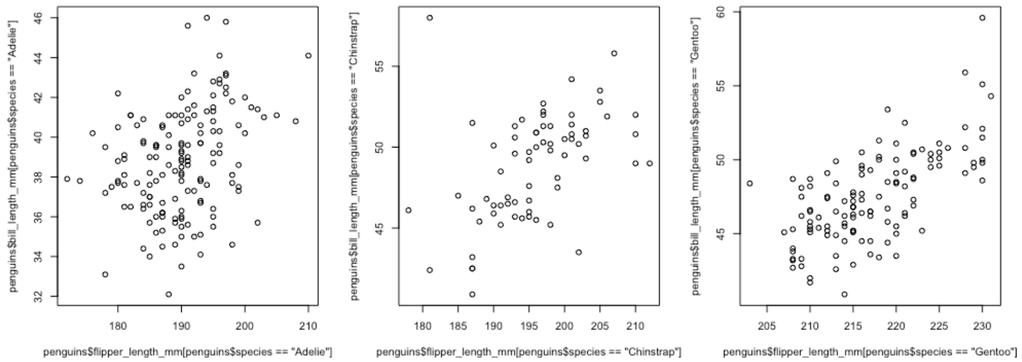
```
ggplot(penguins) +
  geom_point(aes(x = flipper_length_mm,
                 y = bill_length_mm))
```

Formula syntax

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




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



Base syntax

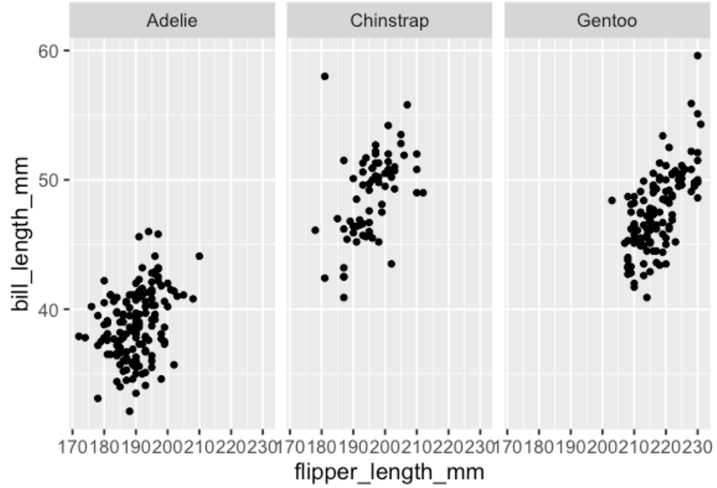
```
par(mfrow = c(1, 3))
plot(penguins$flipper_length_mm[penguins$species == "Adelie"],
     penguins$bill_length_mm[penguins$species == "Adelie"])
plot(penguins$flipper_length_mm[penguins$species == "Chinstrap"],
     penguins$bill_length_mm[penguins$species == "Chinstrap"])
plot(penguins$flipper_length_mm[penguins$species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

Formula syntax

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

Tidyverse syntax

```
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

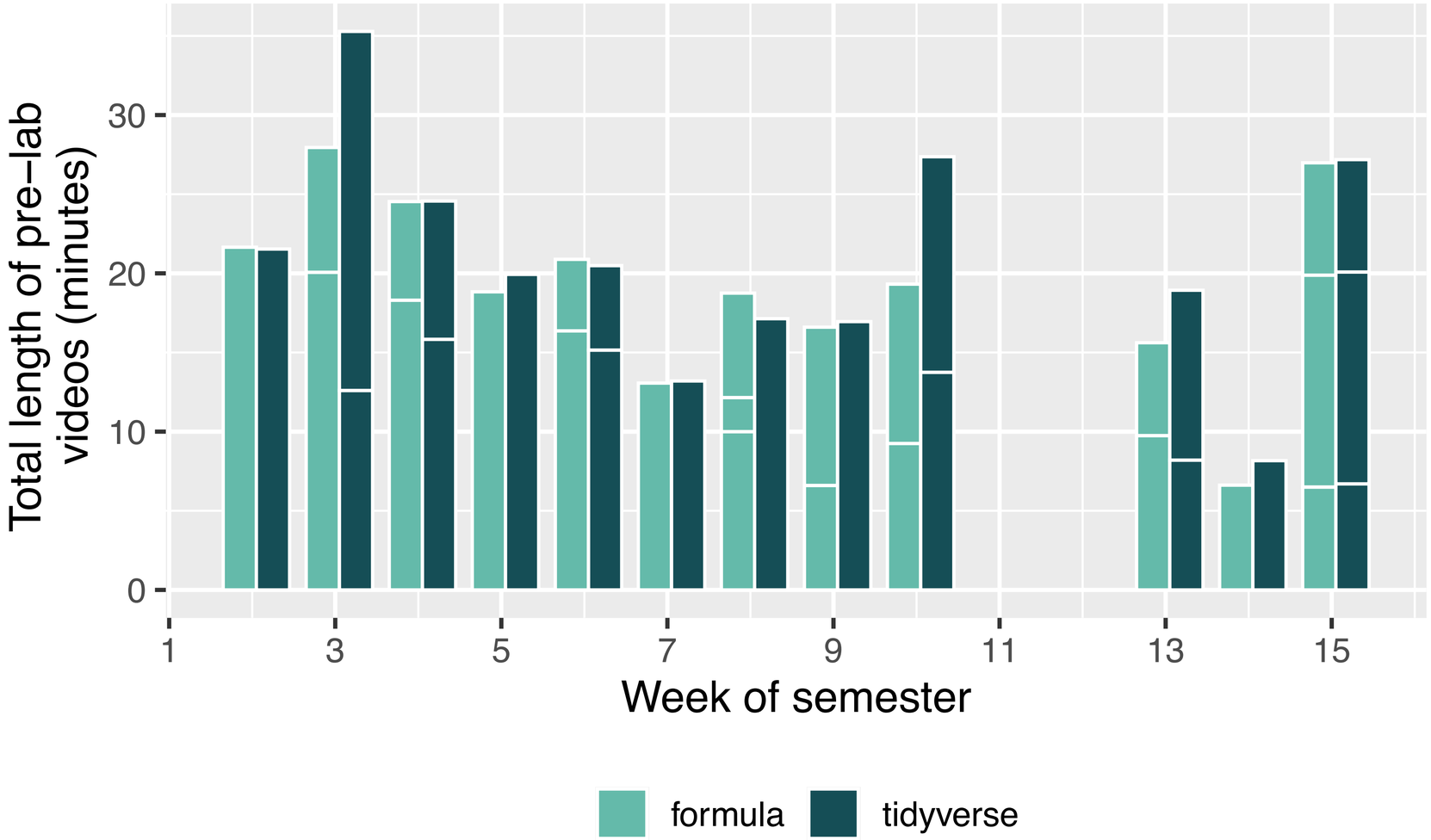
- Consisted of 21 students (fewer took pre/post survey)
- Were comprised mostly of Business majors
- Had similar prior programming experience
- Were given a pre-lab RMarkdown document and associated YouTube video(s) for the material of the week
- Met synchronously to ask questions on the real lab assignment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

Overall: not much difference

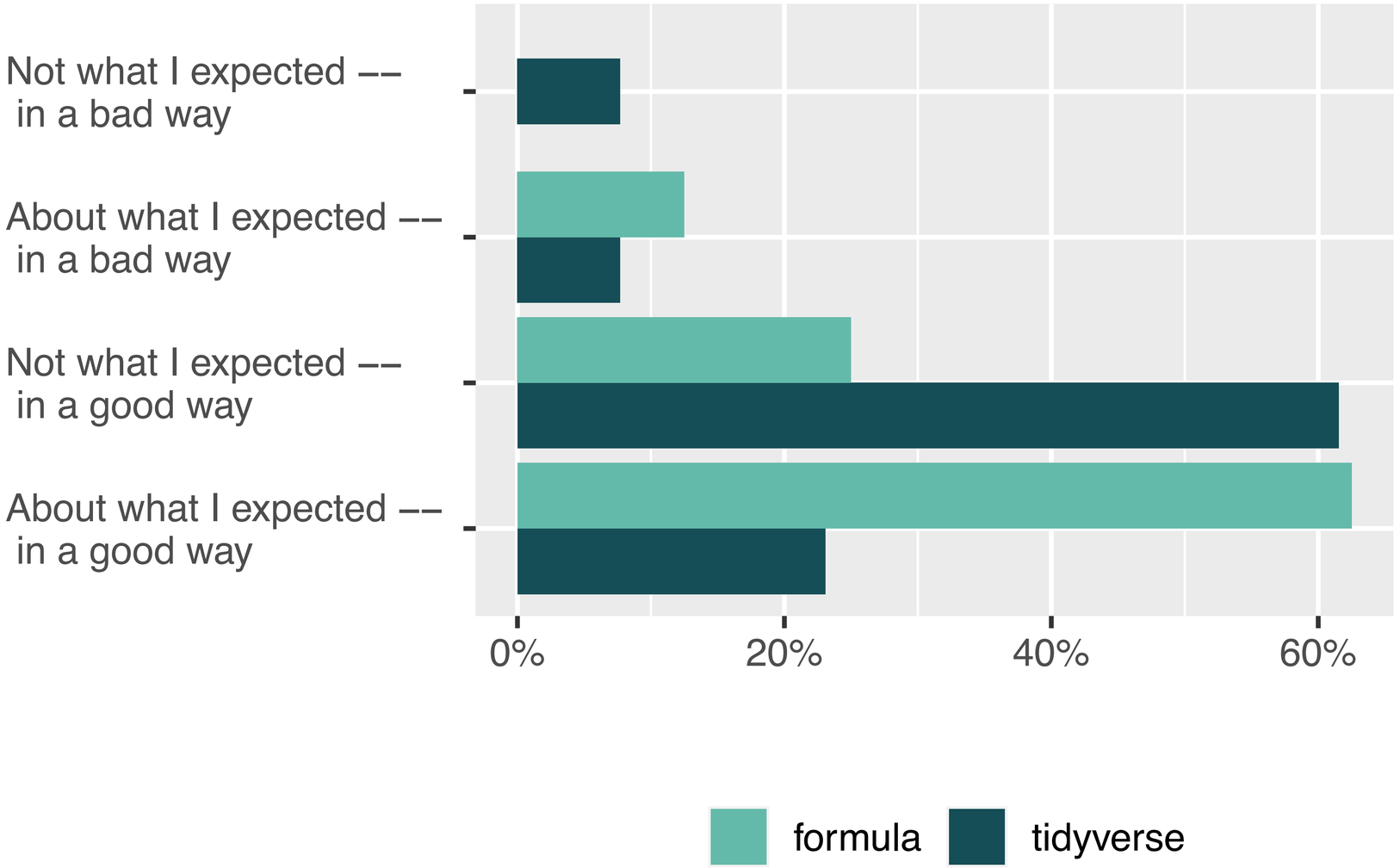


Length of pre-lab videos each week. Outlines help delineate multiple videos for a single week.

https://arxiv.org/abs/2201.12960

Overall: not much difference

How was the experience of learning to program in R?



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

https://arxiv.org/abs/2201.12960

Slight difference in number of functions

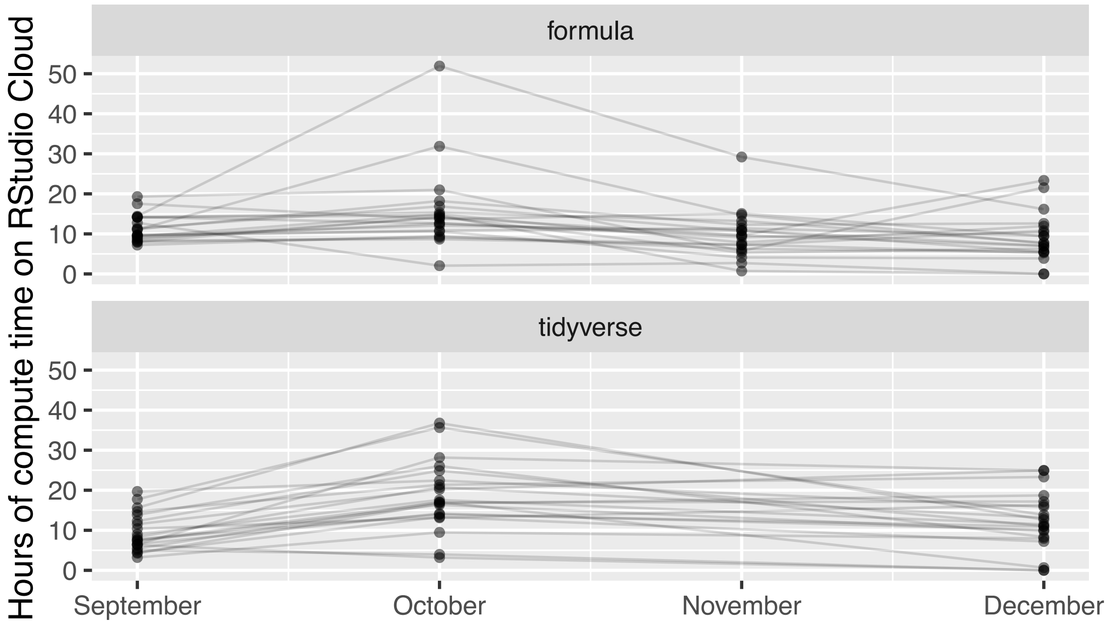
The formula section saw a total of **37 functions** and the tidyverse section saw **50**, with an **overlap of 18 functions** between the two sections.

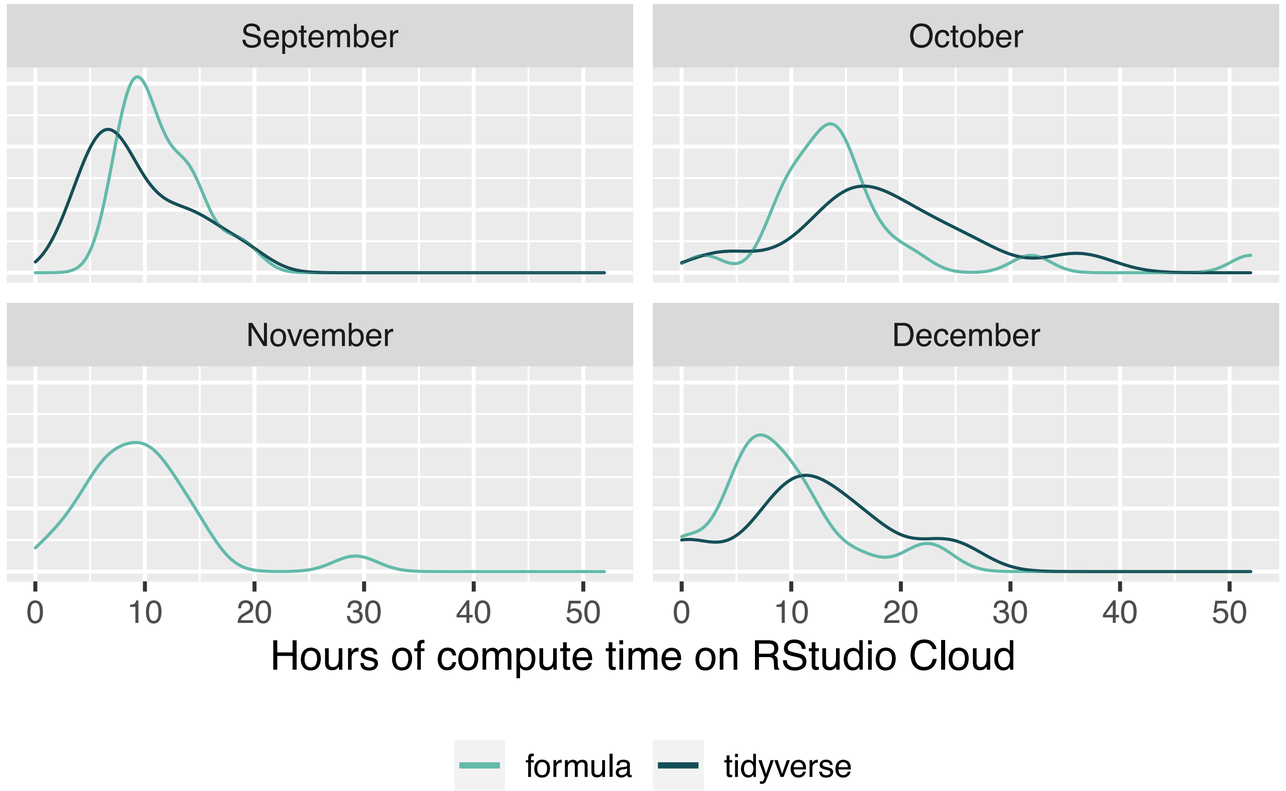
Neither of these numbers are very large!

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 RMark- down document), statistics like `mean()`, `sd()`, and `cor()`, and modeling-related functions like `aov()`, `lm()`, `summary()` and `predict()`.

https://arxiv.org/abs/2201.12960

Compute time was different





https://arxiv.org/abs/2201.12960

Materials are available

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