

Amelia McNamara [@AmeliaMN](#)

Implications of R Syntax in Intro Stats

A minimal reproducible... statistics course

R syntax

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

base

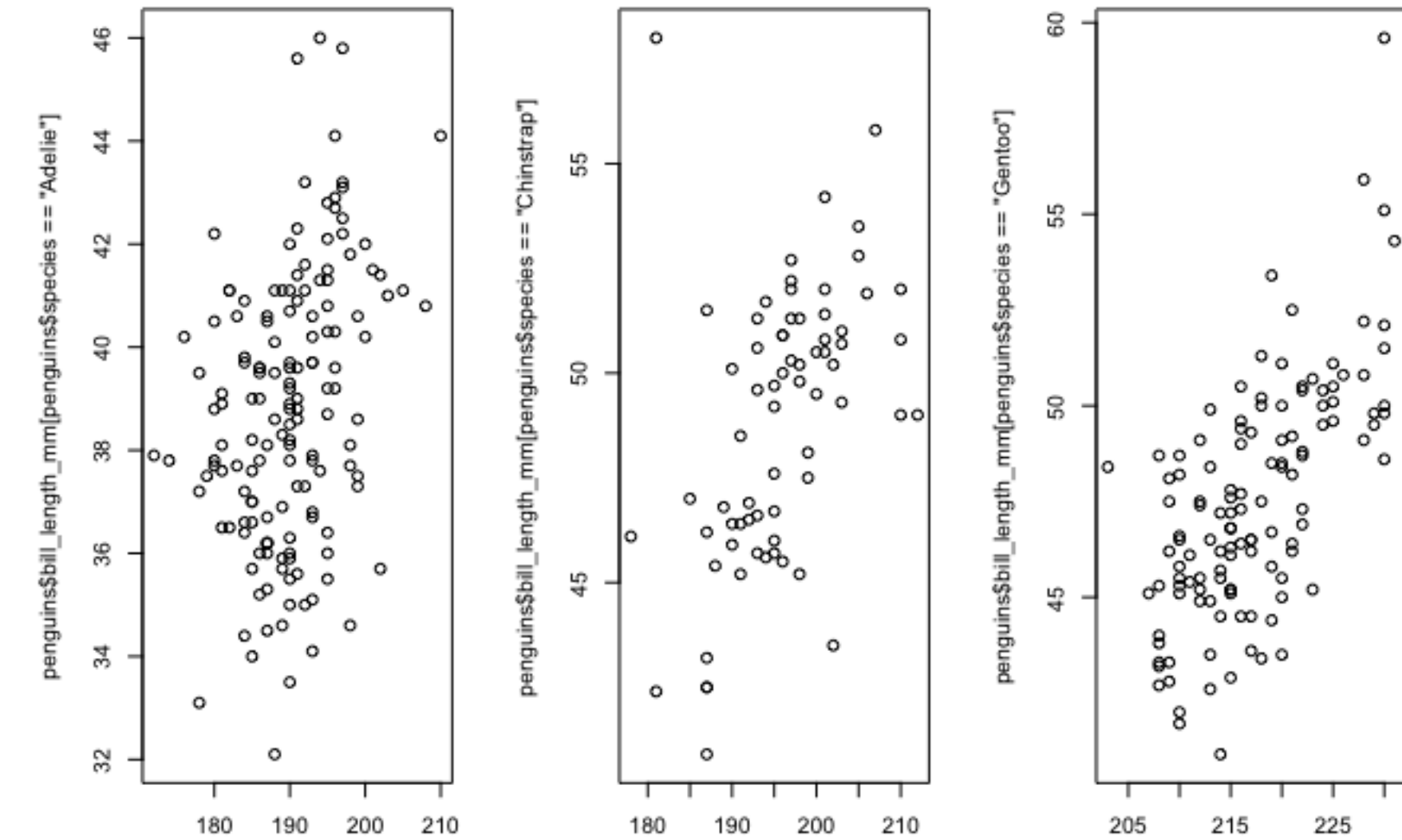
```
par(mfrow = c(1, 3))  
plot(penguins$flipper_length_mm[penguins$species == "Adelie"],  
     penguins$bill_length_mm[penguins$species == "Adelie"])  
plot(penguins$flipper_length_mm[penguins$species == "Chinstrap"],  
     penguins$bill_length_mm[penguins$species == "Chinstrap"])  
plot(penguins$flipper_length_mm[penguins$species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

formula

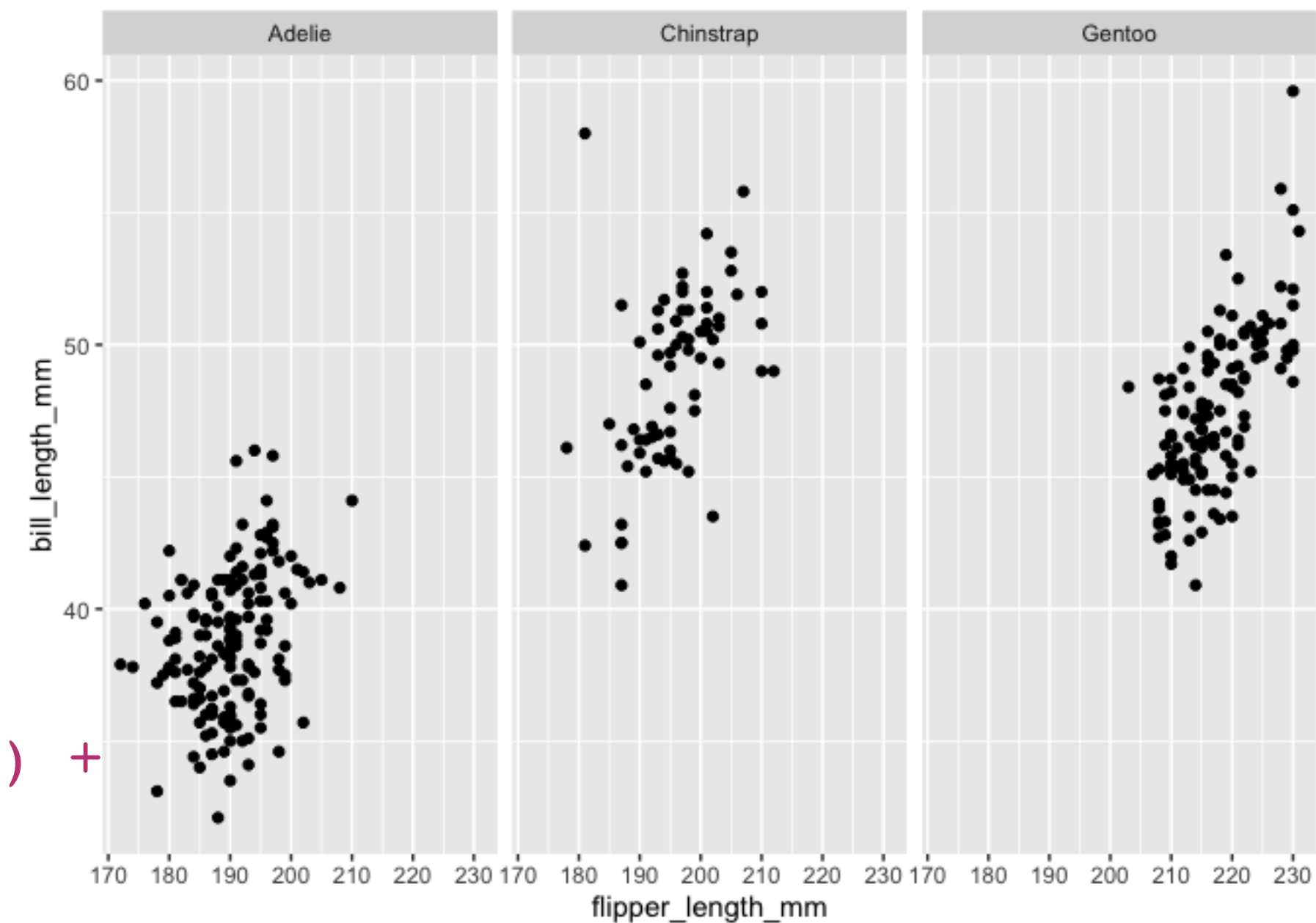
```
library(mosaic)  
gf_point(bill_length_mm ~ flipper_length_mm | species,  
         data = penguins)
```

tidyverse

```
library(ggplot2)  
ggplot(penguins, aes(x = flipper_length_mm, y = bill_length_mm)) +  
  geom_point() +  
  facet_grid(~species)
```



iguins\$flipper_length_mm[penguins\$species == uins\$flipper_length_mm[penguins\$species == "guins\$flipper_length_mm[penguins\$species ==



R syntax

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

base

```
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"])
```

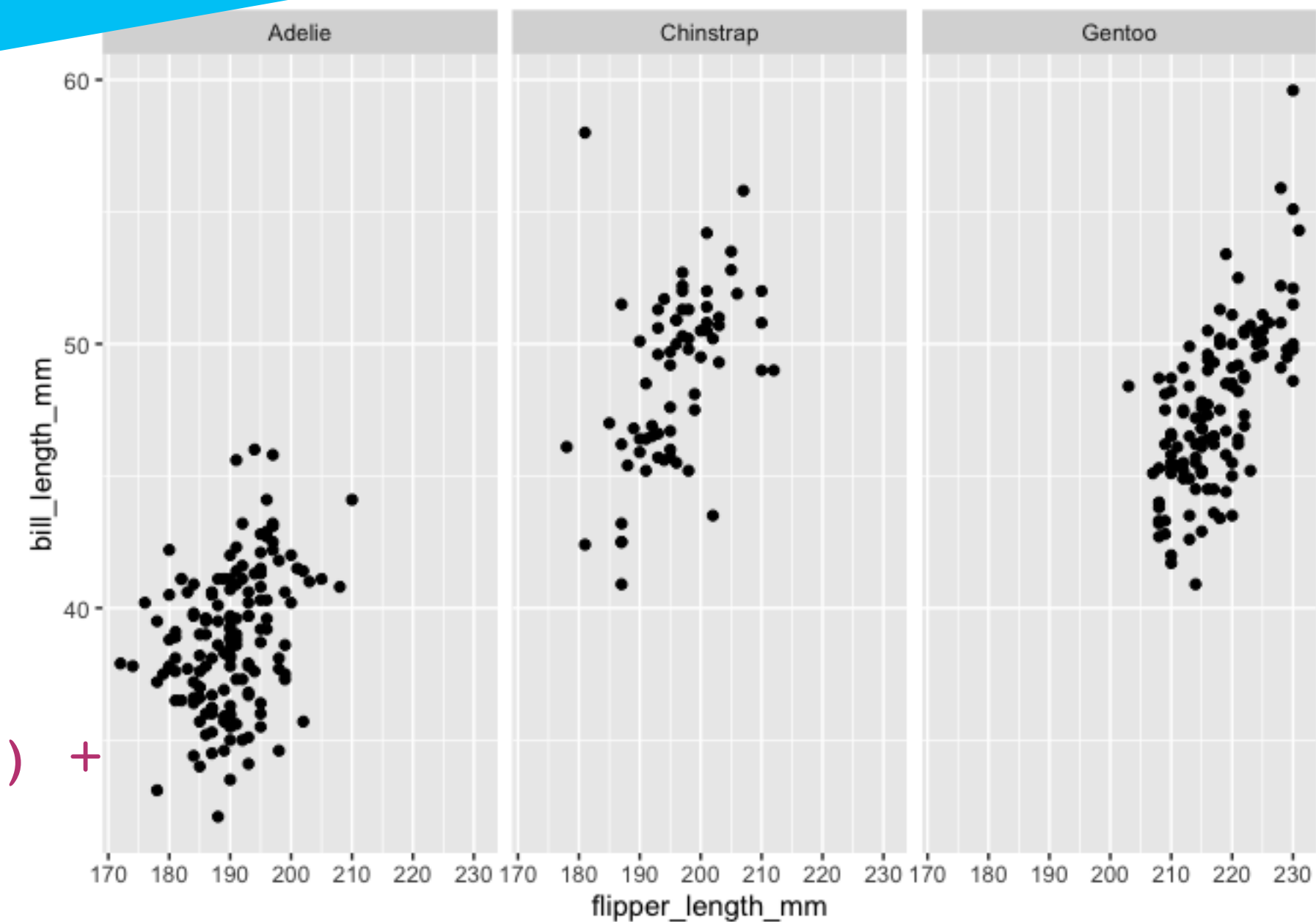


WHICH SHOULD WE TEACH?

```
library(ggplot2)  
ggplot(penguins, aes(x = flipper_length_mm, y = bill_length_mm)) +  
  geom_point() +  
  facet_grid(~species)
```

tidyverse

```
library(ggplot2)  
ggplot(penguins, aes(x = flipper_length_mm, y = bill_length_mm)) +  
  geom_point() +  
  facet_grid(~species)
```



**I did a semester-long, head-to-head comparison
of formula and tidyverse syntaxes**

**I did a semester-long, head-to-head comparison
of formula and tidyverse syntaxes**

Why?

I did a semester-long, head-to-head comparison of formula and tidyverse syntaxes

Why?

- To get some data

I did a semester-long, head-to-head comparison of formula and tidyverse syntaxes

Why?

- To get some data
- Constraints breed creativity

The results

- Some things are easy (RMarkdown, inference)
- Some things are hard
 - Formula: dealing with/explaining missing data

```
options(na.rm = TRUE)
mean(body_mass_g ~ species, data = penguins, na.rm = TRUE)
cor(body_mass_g ~ species, data = penguins, use = "complete.obs")
```

- Tidiverse: dealing with/explaining two categorical variables

```
penguins %>%
  group_by(sex, island) %>%
  summarize(n = n()) %>%
  mutate(prop = n / sum(n))
#> # A tibble: 4 × 4
#> # Groups:   sex [2]
#>   sex      island      n prop
#>   <fct>   <fct>   <int> <dbl>
#> 1 female Biscoe     80 0.567
#> 2 female Dream      61 0.433
#> 3 male   Biscoe     83 0.572
#> 4 male   Dream      62 0.428

library(infer)
penguins %>%
  prop_test(
    response = island,
    explanatory = sex,
    alternative = "two-sided",
    order = c("female", "male")
  )
#> # A tibble: 1 × 6
#>   statistic chisq_df p_value alternative lower_ci upper_ci
#>   <dbl>      <dbl> <dbl> <chr>          <dbl>    <dbl>
#> 1  1.78e-30      1  1.00 two.sided    -0.127  0.117
```


The results

- A minimal reproducible... statistics course doesn't need many functions
 - formula section saw 37 functions
 - tidyverse section saw 50

function	times
<code>library()</code>	30
<code>set()</code>	18
<code>mean()</code>	17
<code>gf_histogram()</code>	14
<code>read.csv()</code>	14

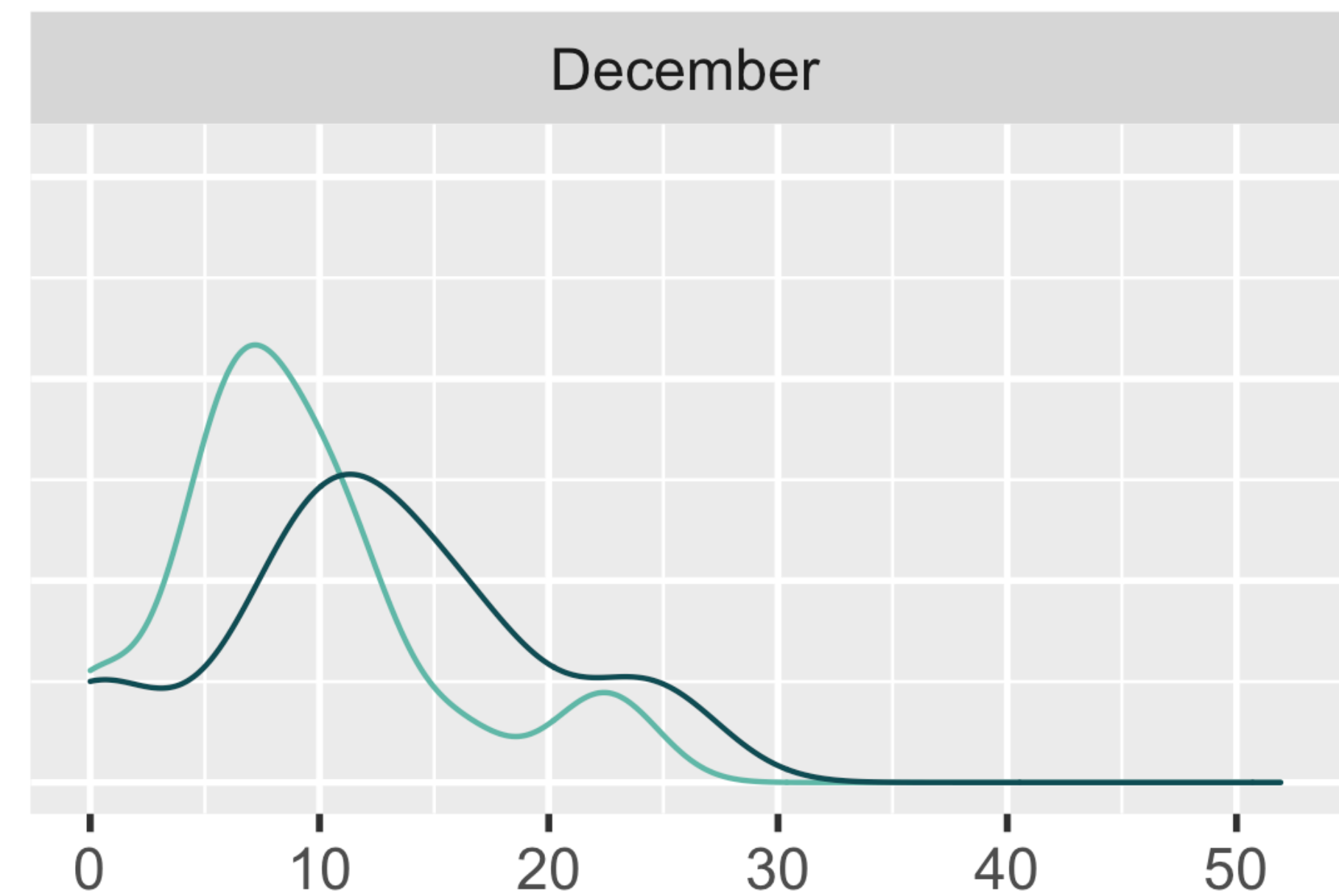
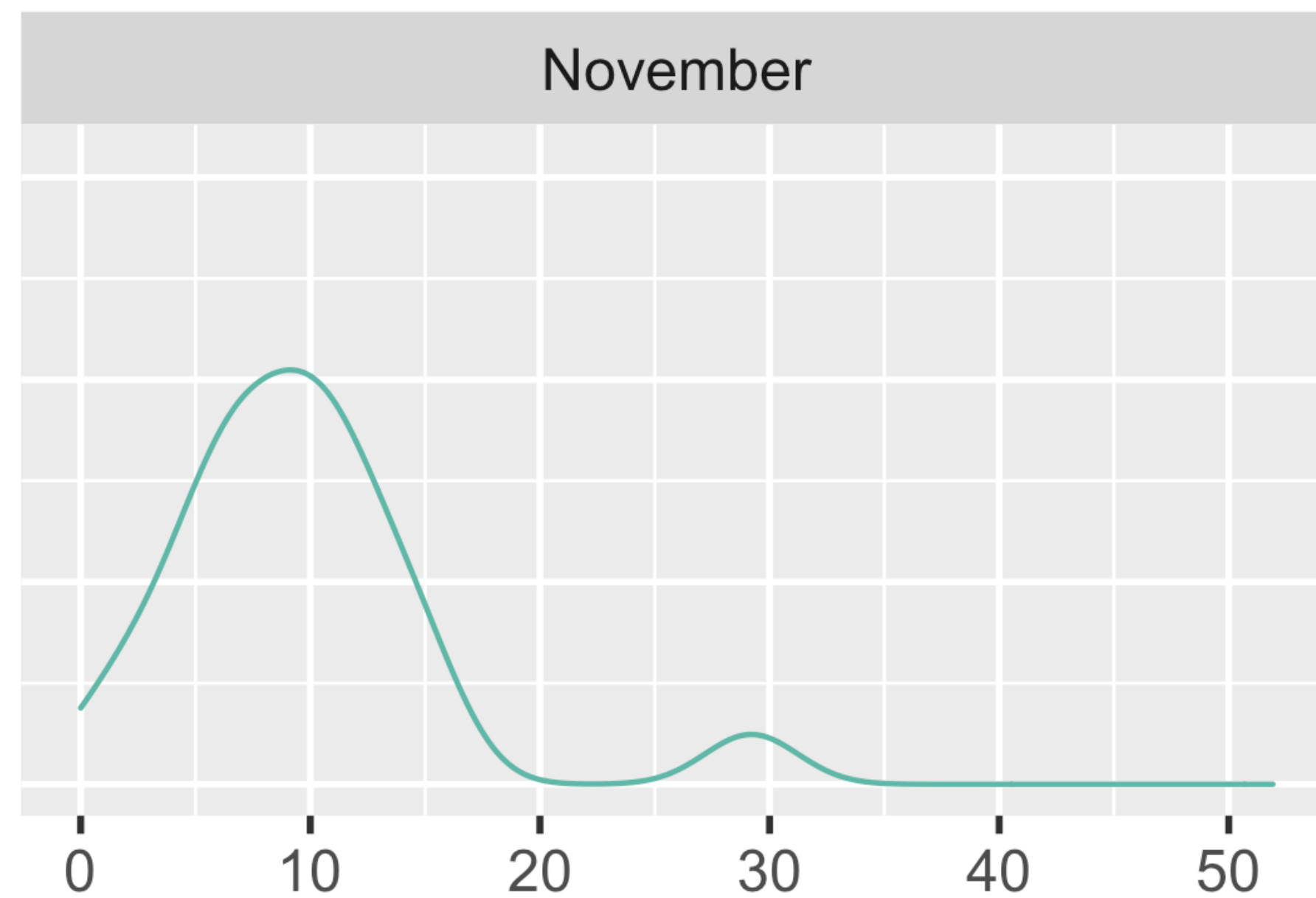
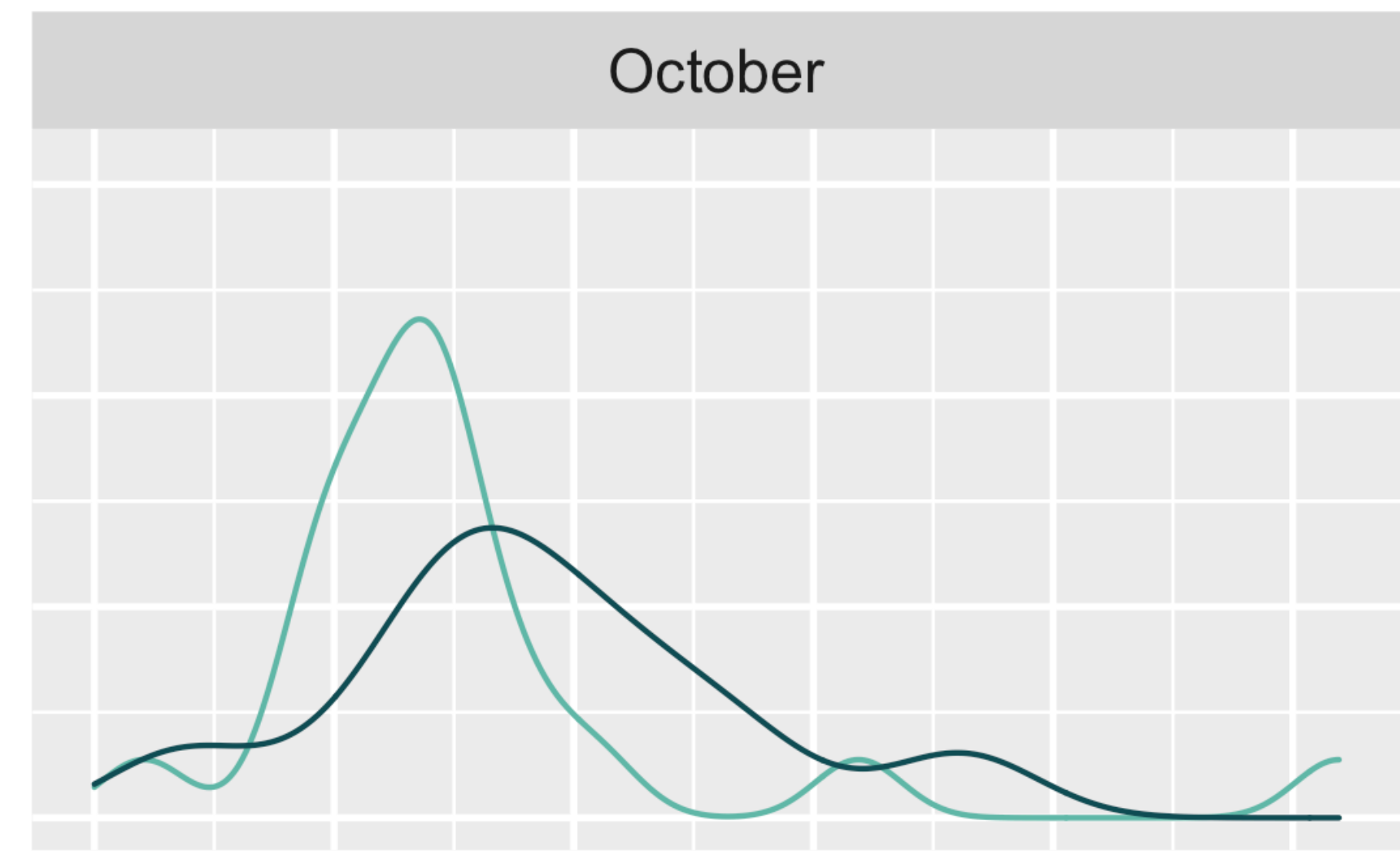
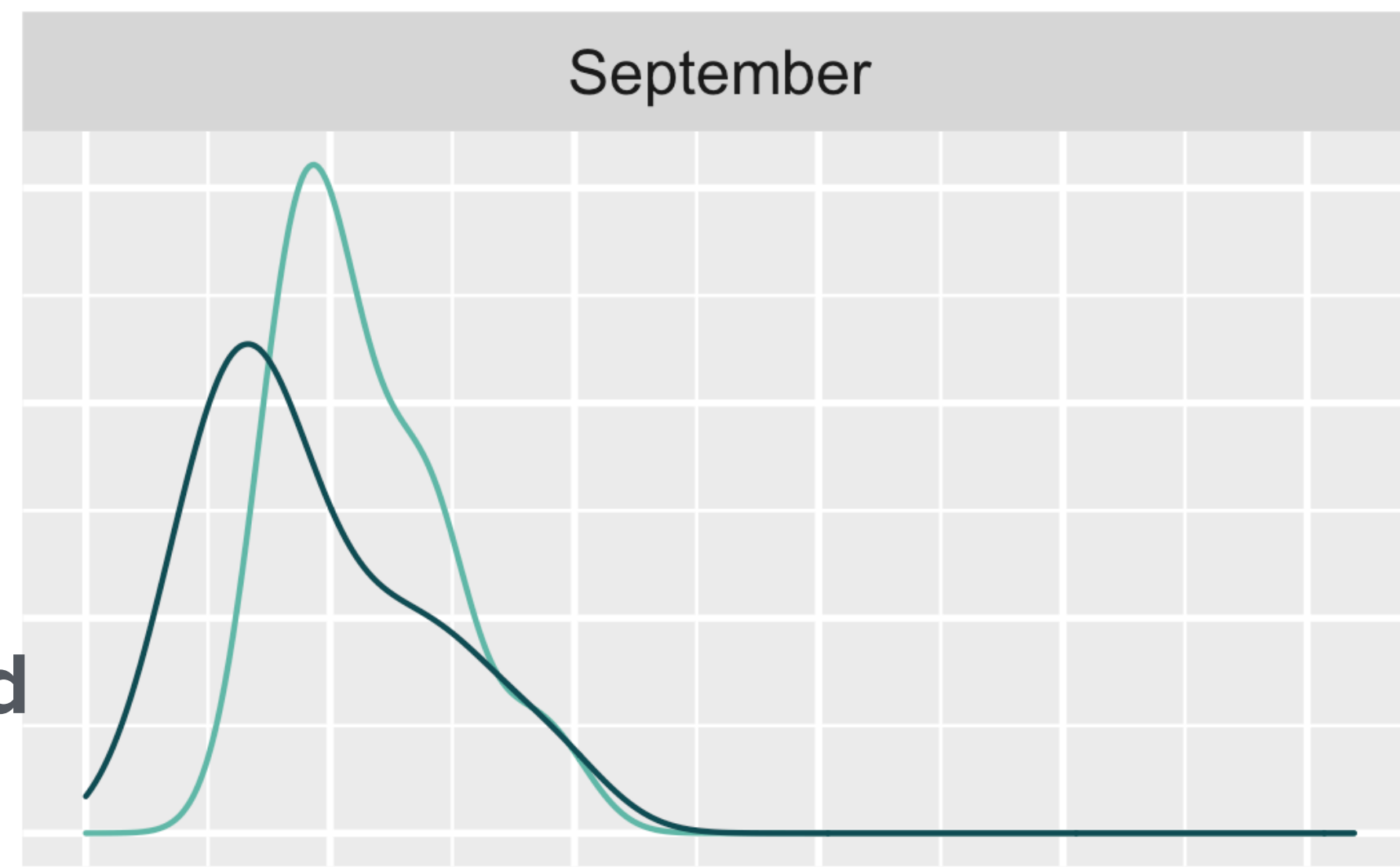
formula section

function	times
<code>summarize()</code>	36
<code>library()</code>	30
<code>ggplot()</code>	29
<code>aes()</code>	28
<code>drop_na()</code>	23

tidyverse section

The results

- Students spent more time on RStudio.cloud in the tidyverse section
- Why? ˘(ツ)˘

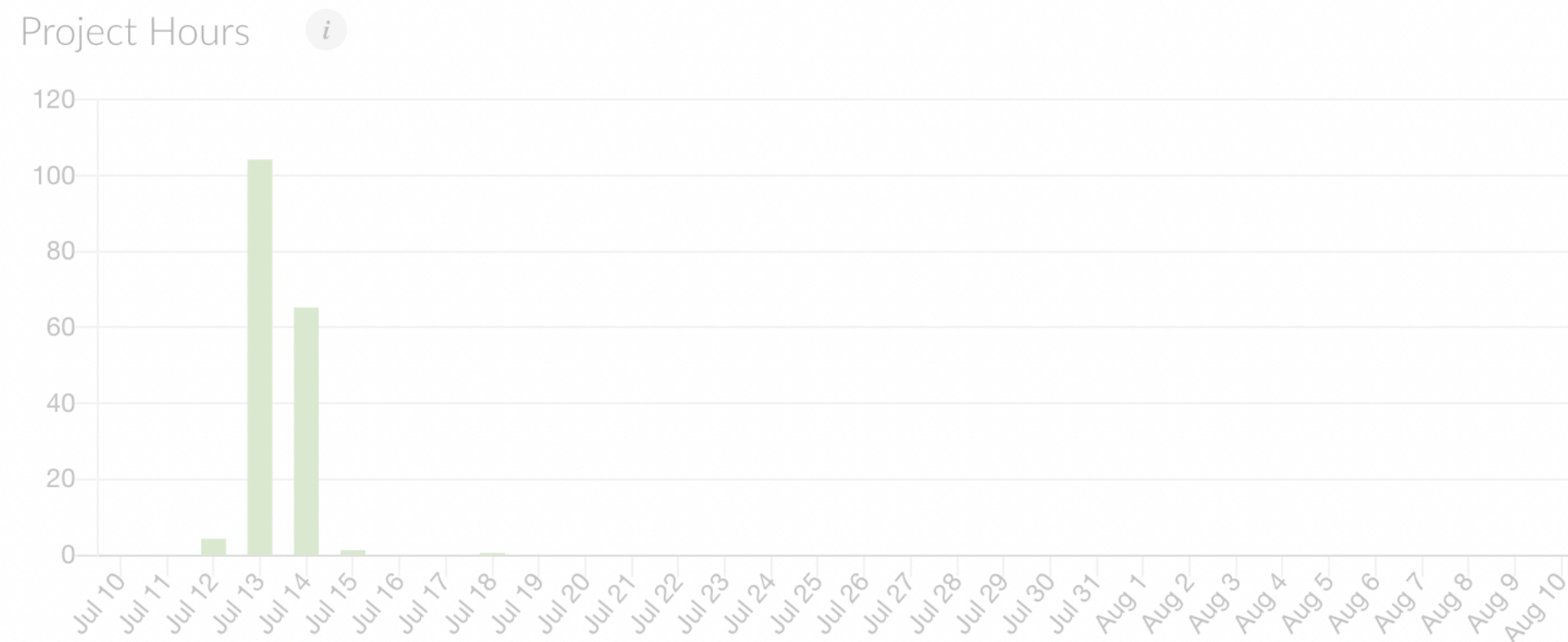


Hours of compute time on RStudio Cloud

— formula — tidyverse

Want to try this yourself?

- Function data: `getParseData()`
- Rstudio.cloud data



Members (27)

Name	Project Hours	Active Projects
Amelia McNamara	16.67	5

Filter: Invert Hide data URLs

All | Fetch/XHR JS CSS Img Media Font Doc WS Wasm Manifest Other

Has blocked cookies Blocked Requests 3rd-party requests

20000 ms 40000 ms 60000 ms 80000 ms 100000 ms

Name Headers Payload Preview Response >>

data:image/svg+xml;...
 data:image/svg+xml;...
 data:image/svg+xml;...
 data:image/svg+xml;...
 space_member_max
 series?from=165749131900...
 usage?from=165749131900...
 space_member_max
 series?from=165749131900...
 usage?from=165749131900...
 space_content_max
 usage?fr...
 space_cc...
 usage?fr...
 account...
 account...
 cloud
 cloud
 pubac5d...
 compute...
 compute...
 pubac5d...
 pubac5d...
57 requests

General

Request URL: https://api.rstudio.cloud/v1/spaces/262068/usage?from=1657491319000&until=1660169719000&space_id=262068&groupby=user_id

Request Method: GET

Status Code: 200

Remote Address: 54.88.184.94:443

Referrer Policy: strict-origin-when-cross-origin

Response Headers

access-control-allow-credentials: true

io.cloud

UTF-8

ea4cb9fa"

491319000&

groupby=user_

Open in new tab
Clear browser cache
Clear browser cookies
Copy
Block request URL
Block request domain
Replay XHR
Sort By
Header Options
Save all as HAR with content

Copy link address
Copy response
Copy stack trace
Copy as PowerShell
Copy as fetch
Copy as Node.js fetch
Copy as cURL
Copy all as PowerShell
Copy all as fetch
Copy all as Node.js fetch
Copy all as cURL
Copy all as HAR

Pre-print

- Teaching modeling in introductory statistics: A comparison of formula and tidyverse syntaxes
- <https://arxiv.org/abs/2201.12960>

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

Amelia McNamara *

Department of Computer & Information Sciences, University of St Thomas

May 13, 2022

Abstract

This paper reports on a head-to-head comparison run in a pair of introductory statistics labs, one conducted fully in the formula syntax, the other in tidyverse. Analysis of incidental data from YouTube and RStudio Cloud show interesting distinctions. The formula section appeared to watch a larger proportion of pre-lab YouTube videos, but spend less time computing on RStudio Cloud. Conversely, the tidyverse section watched a smaller proportion of the videos and spent more time computing. Analysis of lab materials showed that tidyverse labs tended to be slightly longer in terms of lines in the provided RMarkdown materials, but not in minutes of the associated YouTube videos. The tidyverse labs exposed students to slightly more distinct R functions, but both labs relied on a quite small vocabulary of consistent functions, which can provide a starting point for instructors interested in teaching introductory statistics in R. Analysis of pre- and post-survey data show no differences between the two labs, so students appeared to have a positive experience regardless of section. This work provides additional evidence for instructors looking to choose between syntaxes for introductory statistics teaching.

Keywords: R language, instruction, data science, statistical computing

*amelia.mcnamara@stthomas.edu

THANK YOU

[@AmeliaMN](https://twitter.com/AmeliaMN)

www.amelia.mn

<https://arxiv.org/abs/2201.12960>

<https://github.com/AmeliaMN/STAT220-labs>

<https://github.com/AmeliaMN/ComparingSyntaxForModeling>