

Solving 8 visualisation challenges with ggplot2

November 2016

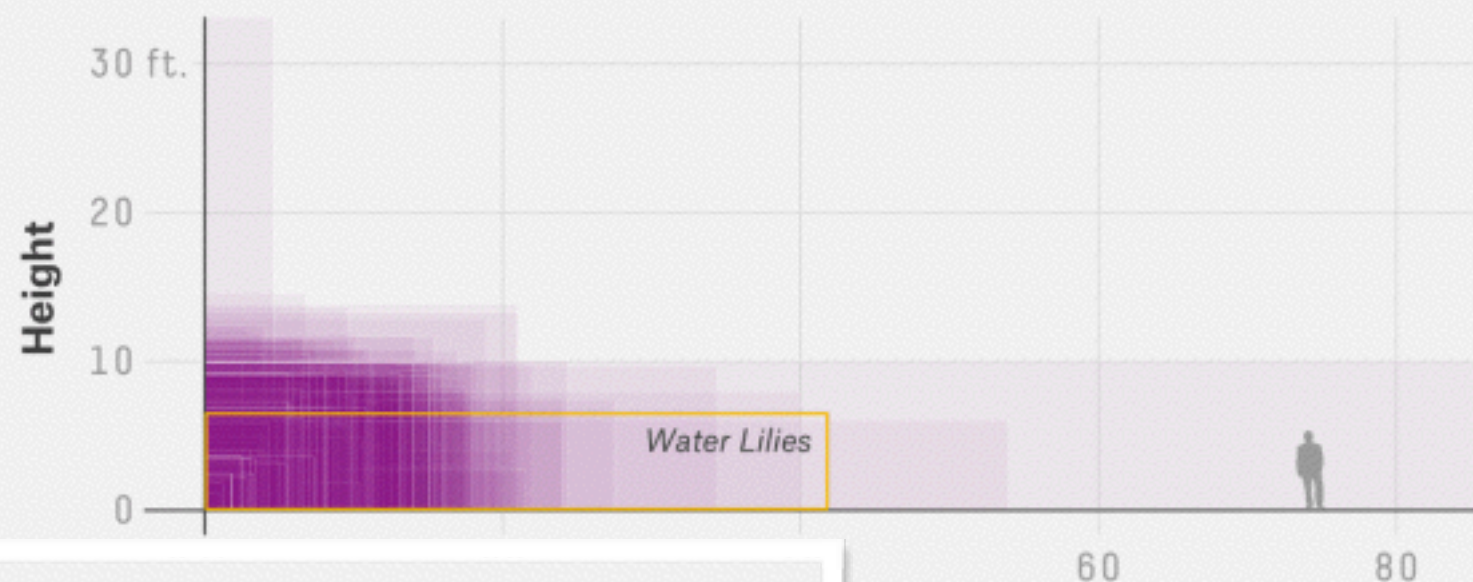
Hadley Wickham
[@hadleywickham](#)
Chief Scientist, RStudio

The Three Types Of Adam Sandler Movies

Box office gross in 2014 dollars vs. Rotten Tomatoes

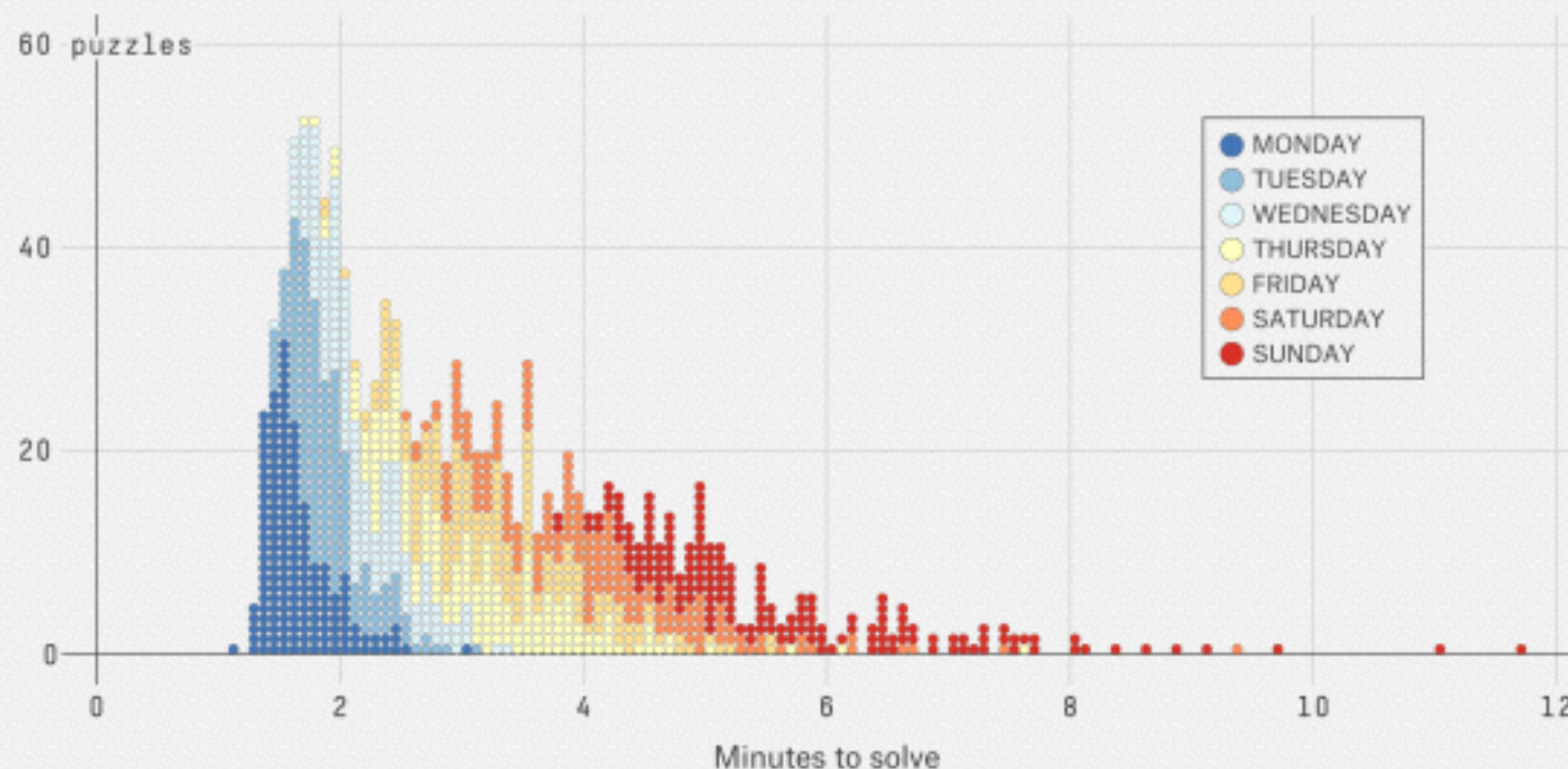


Every MoMA Painting By Size

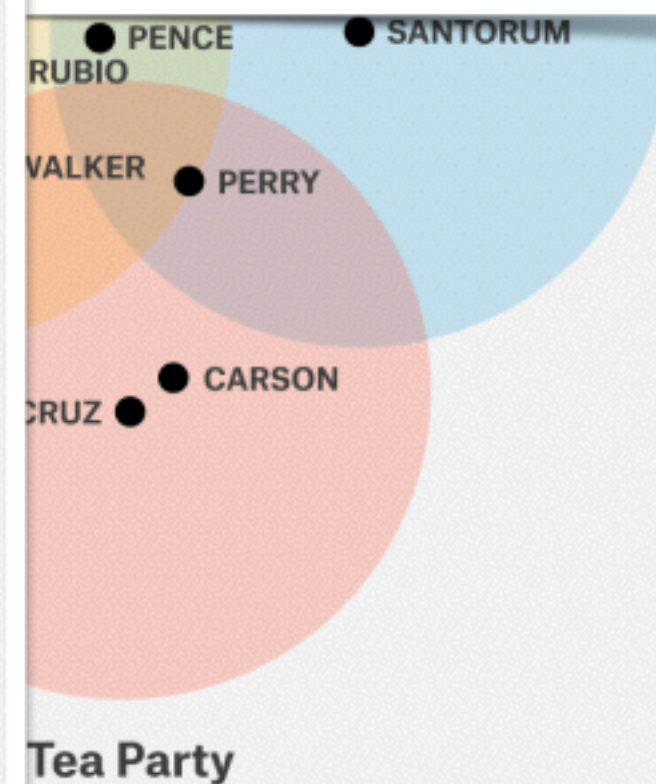


The Puzzling Speed Of Dan Feyer

Solve times for the past 1,208 New York Times crossword puzzles, by day of the week

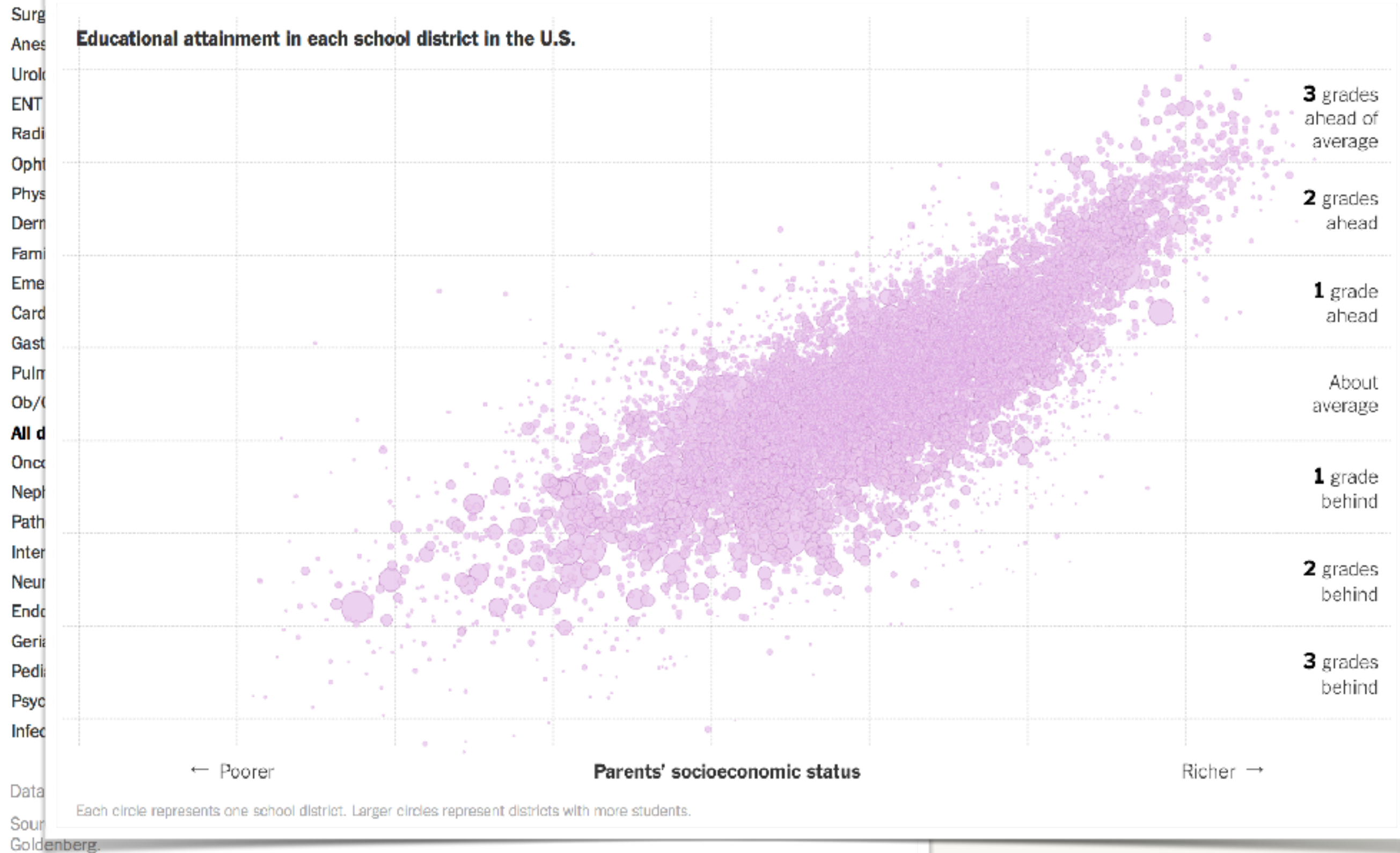


SOURCE: THE MUSEUM OF MODERN ART



Surgeons are Red, Psychiatrists are Blue

Percent of doctors who have a party registration who are Republicans





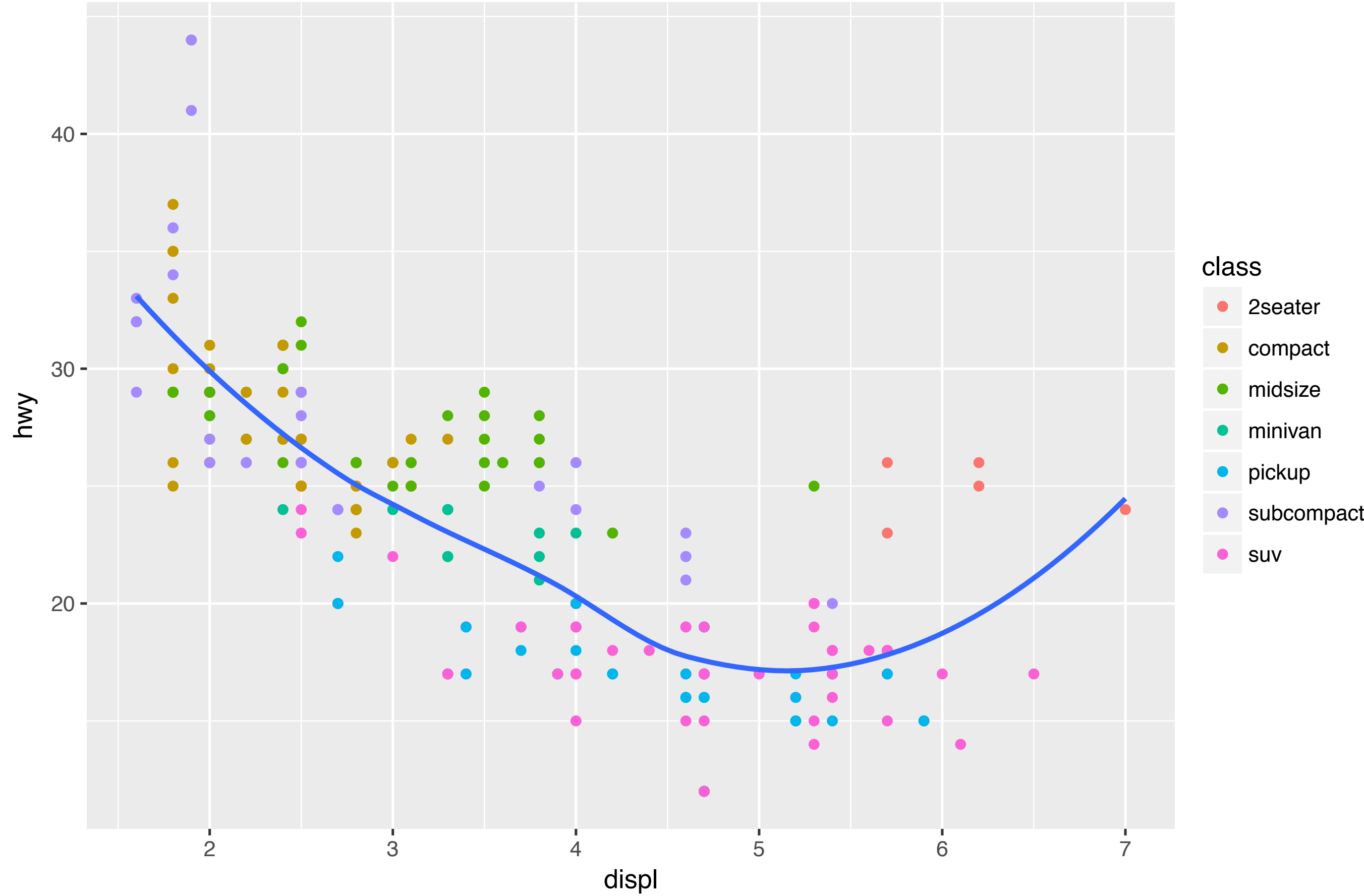
Labelling plots

A problem ignored for too long

Solved by Bob Rudis

Fuel efficiency generally decreases with engine size

Two seaters (sports cars) are an exception because of their light weight



Accessed with the labs() function

```
ggplot(mpg, aes(displ, hwy)) +  
  geom_point(aes(color = class)) +  
  geom_smooth(se = FALSE, method = "loess") +  
  labs(  
    title = "Fuel efficiency generally ...",  
    subtitle = "Two seaters (sports cars) ...",  
    caption = "Data from fueleconomy.gov"  
  )
```


2 Axes



two axes |

two axes **excel**
two axes **matlab**
two axes **of color marketing**
two axes **matplotlib**
two axes **excel 2010**
two axes **in r**
two axes **stata**
two axes **ggplot2**
two axes **crossed**
two axes **python**

Google Search

I'm Feeling Lucky

Stages of visualisation system popularity

1. Someone used it and complained about a bug 🙄
2. Someone used it in an academic paper 😊
3. Someone used it in a newspaper 😄
4. Someone used it to commit academic fraud 😱
5. So many people use it that google has auto completes for bad graphics ideas 🤔

It's not possible in ggplot2 because I believe plots with separate y scales (not y-scales that are transformations of each other) are fundamentally flawed. Some problems:

- They are not invertible: given a point on the plot space, you can not uniquely map it back to a point in the data space.
- They are relatively hard to read correctly compared to other options. See [A Study on Dual-Scale Data Charts](#) by Petra Isenberg, Anastasia Bezerianos, Pierre Dragicevic, and Jean-Daniel Fekete for details.
- They are easily manipulated to mislead: there is no unique way to specify the relative scales of the axes, leaving them open to manipulation. Two examples from the Junkcharts blog: [one](#), [two](#)
- They are arbitrary: why have only 2 scales, not 3, 4 or ten?

You also might want to read Stephen Few's lengthy discussion on the topic [Dual-Scaled Axes in Graphs Are They Ever the Best Solution?](#).

share edit delete flag

edited Nov 5 '13 at 12:40

answered Jun 23 '10 at 13:10



[hadley](#)

56.3k ● 14 ● 112 ● 176

I see. I have heard from you in other post about adding a "transformed y-scale" on the right hand side of the chart, can it be used as a surrogate? how can I do it? Thanks! – [lokheart](#) Jun 24 '10 at 0:51

1 Agreed. The use of multiple y axes should be discouraged. – [Maiasaura](#) Jul 17 '10 at 4:55

21 Would you mind elaborate Your opinion? Not beeing enlightened , I think its a rather compact way of plotting two independent variables. It is also a feature that seems to be asked for, and it's been used widely. – [KarlP](#) Aug 12 '10 at 20:37

36 @hadley: Mostly I agree, but there is a genuine use for multiple y scales - the use of 2 different units for the same data, e.g., Celsius and Fahrenheit scales on temperature time series. – [Richie Cotton](#) Aug 25 '10 at 13:08

8 Yes, which is why that particular case is on the to do list. – [hadley](#) Aug 25 '10 at 21:16

66 this answer isn't very helpful without any explanation of what you mean by "fundamentally flawed". If it is well documented then cite the documentation – [KennyPeanuts](#) May 26 '11 at 17:17

13 Frequently done for exchange rates too. – [Brandon Bertelsen](#) Aug 8 '11 at 21:01

5 @Hadley In your opinion. Not in mine, nor many other scientists. Surely this can be achieved by putting a second plot (with a fully transparent background) directly over the first, so they appear as one. I just don't know how to ensure the corners of the bounding boxes are aligned / registered with each other. – [Nicholas Hamilton](#) Feb 13 '13 at 21:37

If The case mentioned by @KarlP is addressed, it means that many of the other situations can be addressed via simple data transformation. Easy peasy. – [Nicholas Hamilton](#) Feb 13 '13 at 22:47

2 @ADP You are welcome to implement it yourself, but given that I don't believe they're useful, I don't have any plans to. (Especially given that it's already trivial by rescaling each series independently before plotting) – [hadley](#) Feb 13 '13 at 22:59

1 @Hadley, Rescaling is trivial, i agree, what is the best method to put another set of labels and ticks? – [Nicholas Hamilton](#) Feb 13 '13 at 23:19

1 @KennyPeanuts finally added one pointer to the research on this topic. – [hadley](#) Feb 14 '13 at 13:27

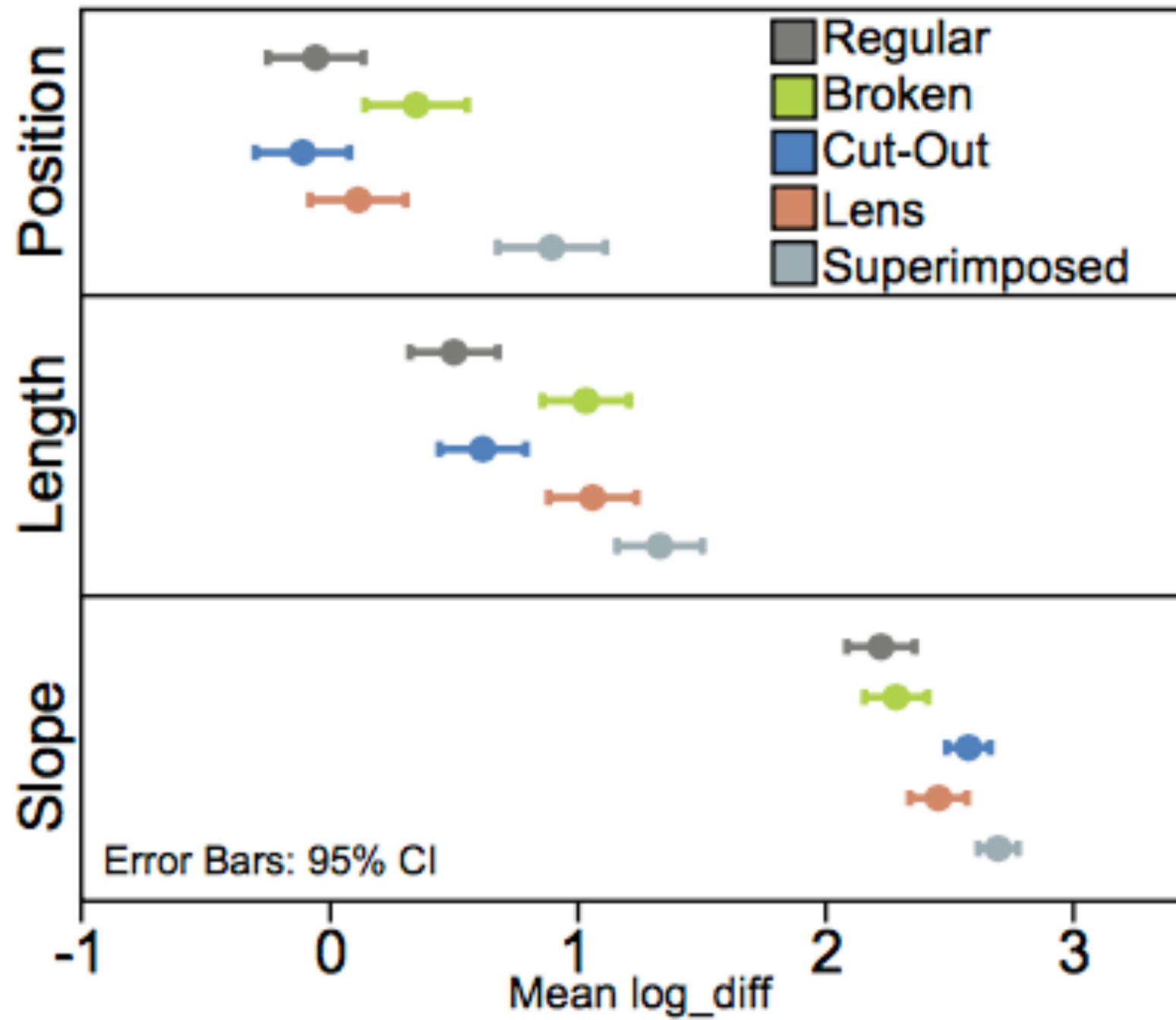
Wow, that's commitment! Thanks for the sources - interesting stuff. – [KennyPeanuts](#) Feb 15 '13 at 0:34

3 @hadley I agree that secondary axes should be avoided, but I'd love to be able to have a *primary* axis on the right hand side...particularly for plotting timeseries. – [seancarmody](#) Nov 5 '13 at 10:42

1 @hadley For example, in [Walther-Lieth Climate Diagrams](#), two y axes are commonly used. Since there is a

- 1 @hadley For example, in [Walther-Lieth Climate Diagrams](#), two y axes are commonly used. Since there is a fixed prescription how to do that the possible confusion is minimal... – [sebschub](#) Mar 25 '14 at 7:51
-
- 1 @sebschub with enough training you can overcome many problems in the underlying graphic – [hadley](#) Mar 25 '14 at 12:47
-
- 14 @hadley I am sorry, I do not see what is problematic with the given climate diagram. Putting temperature and precipitation in one diagram (with the fixed prescription), one gets a quick first guess whether it is humid or arid climate. Or the way around: what would be a better way to visualize temperature, precipitation and their "relation"? Anyway, thanks a lot for your work in ggplot2! – [sebschub](#) Mar 25 '14 at 14:11
-
- 4 -1 It is one thing to say a method is a bad way of doing it, but you don't propose an alternative that ggplot2 can do. As an answer to the question this is not helpful. – [Corone](#) Aug 29 '14 at 13:42
-
- Well, that is kind of dissapointing, because sometimes you need just to show some very basic stuff, where proper scales do not matter much - like y was moving in tandem with x and now it is not.:(– [flipper](#) Oct 6 '14 at 20:01
-
- 6 We do not currently manufacture cars with top speed in excess of 130km/h because that sort of speed would be unsafe. Oh wait... – [PatrickT](#) Dec 15 '14 at 11:30
-
- 11 @PatrickT I'd be happy to add that feature for you for \$80,000 ;) – [hadley](#) Dec 16 '14 at 6:14
-
- 44 A graphics package forcing an opinion on its users is fundamentally flawed. – [ROLO](#) Mar 26 '15 at 8:52
-
- @hadley First, thank you for the awesome work you've put into the package. Secondly, would highly appreciate your opinion concerning that [question](#) – [user5363218](#) Jan 12 at 10:28
-
- 12 ▲ @ROLO No, that's nonsense. I'm appalled at the number of upvotes this comment has received, because it shows a fundamental misunderstanding of API design: **every great API is opinionated**. That said, I agree that having dual x-axes can (very rarely!) be useful. – [Konrad Rudolph](#) Feb 12 at 18:17
-
- 9 ▲ Everyone crying about this is free to write their own solution to it. Acting like the author of a free graphics package owes you something is just weird. – [DWal](#) Mar 8 at 18:54
-
- Um... I wouldn't use more than 2 y-axes because there are only 2 dimensions. – [AmagicalFishy](#) Apr 8 at 20:18
-
- 1 Although not a good idea to use duel axis, @hadley you should let user to decide about that. – [TheRimalaya](#) Apr 23 at 21:02

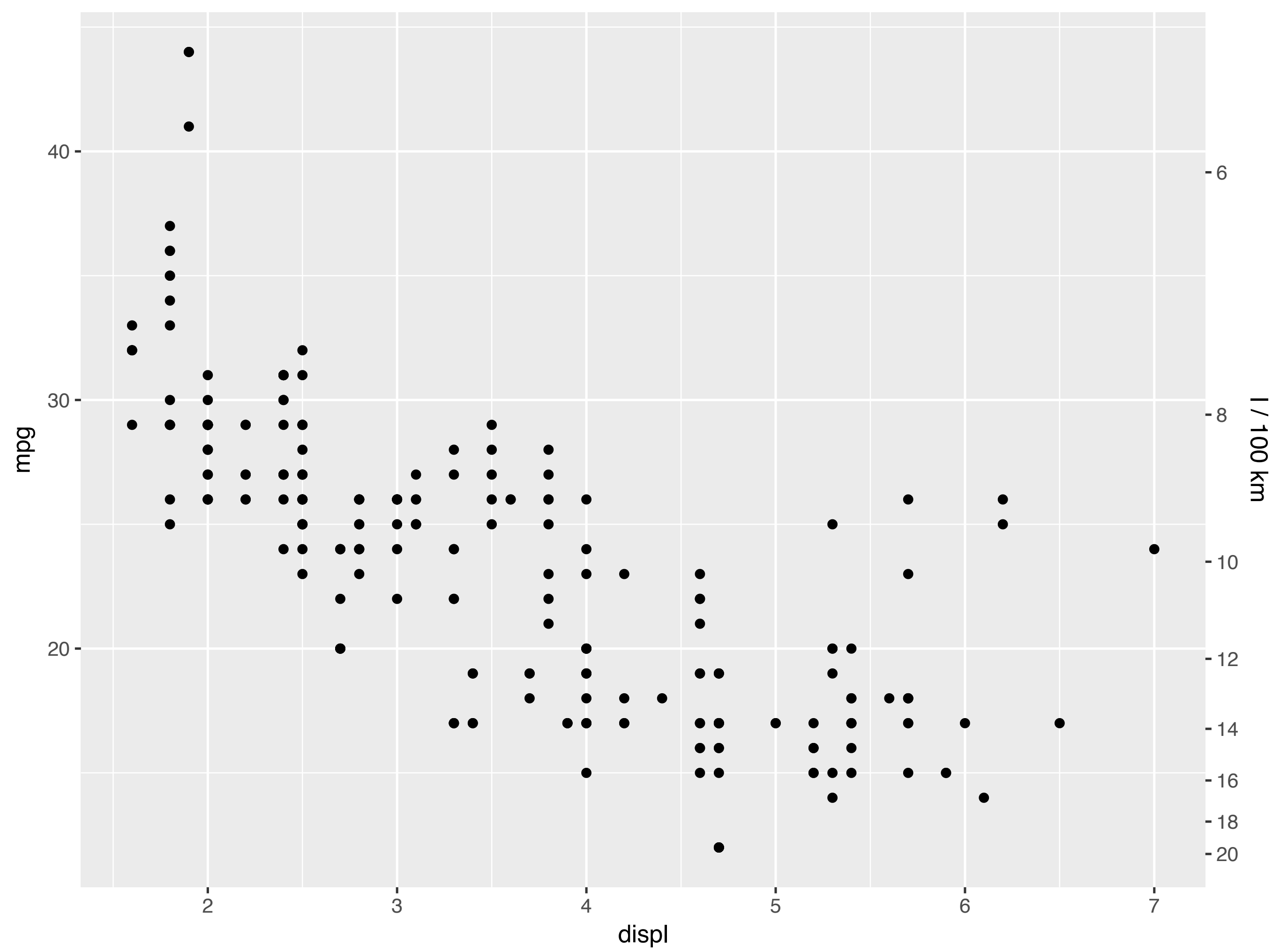
[add a comment](#)



Isenberg, Petra, et al. "A study on dual-scale data charts." *IEEE Transactions on Visualization and Computer Graphics* 17.12 (2011): 2469-2478.

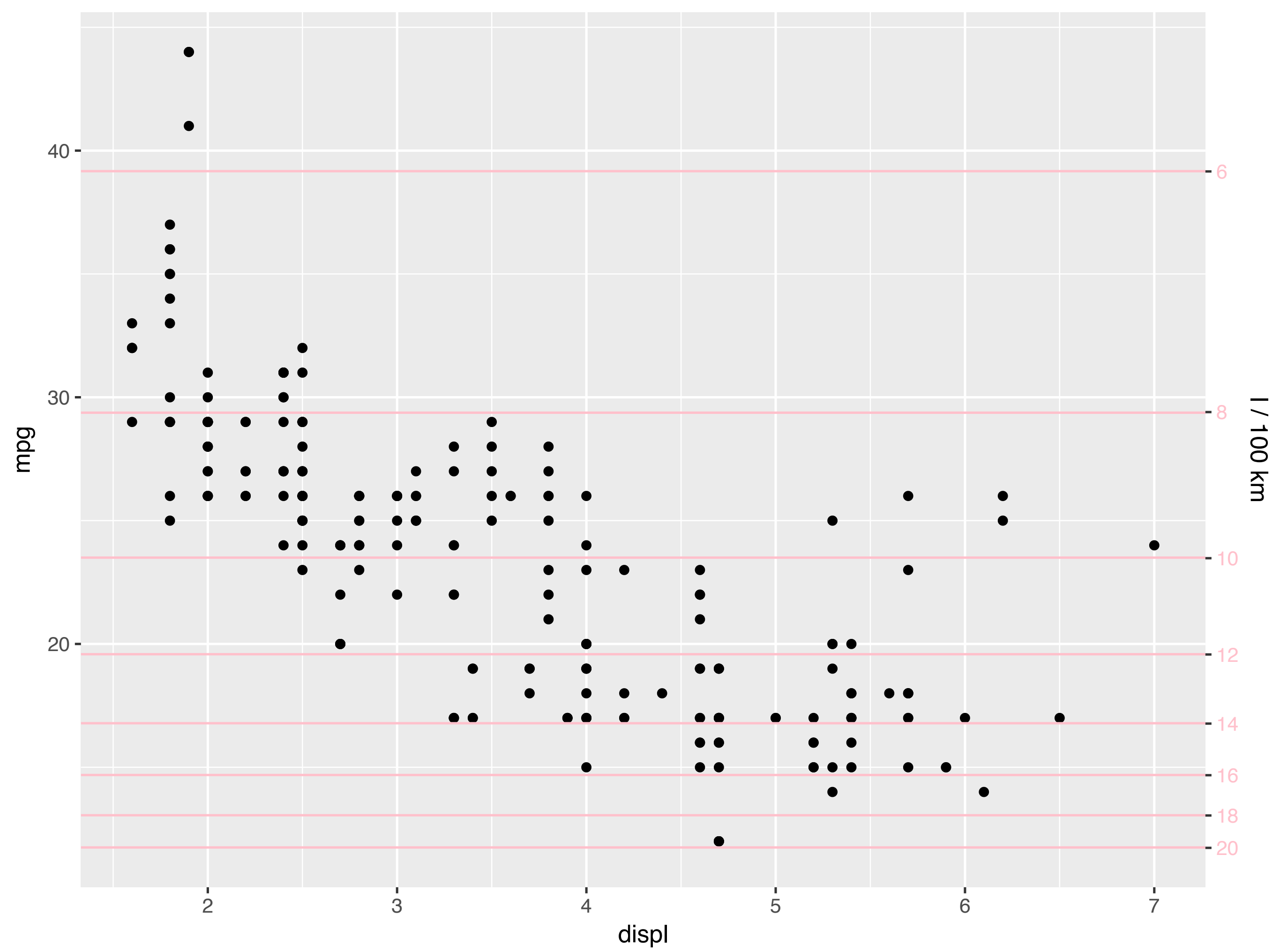
https://www.lri.fr/~isenberg/publications/papers/Isenberg_2011_ASO.pdf

But...



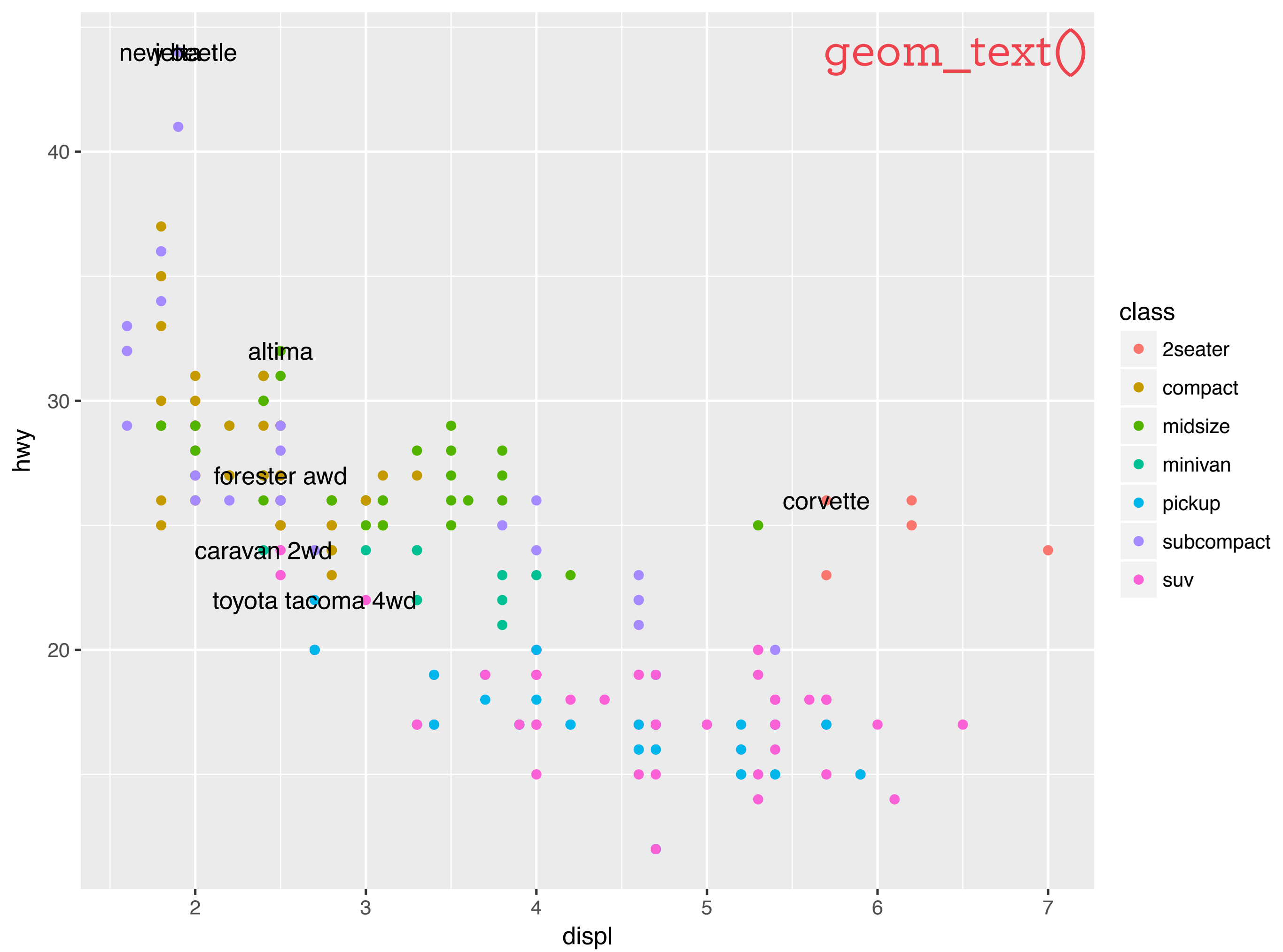
Only 1-to-1 transformations are allowed

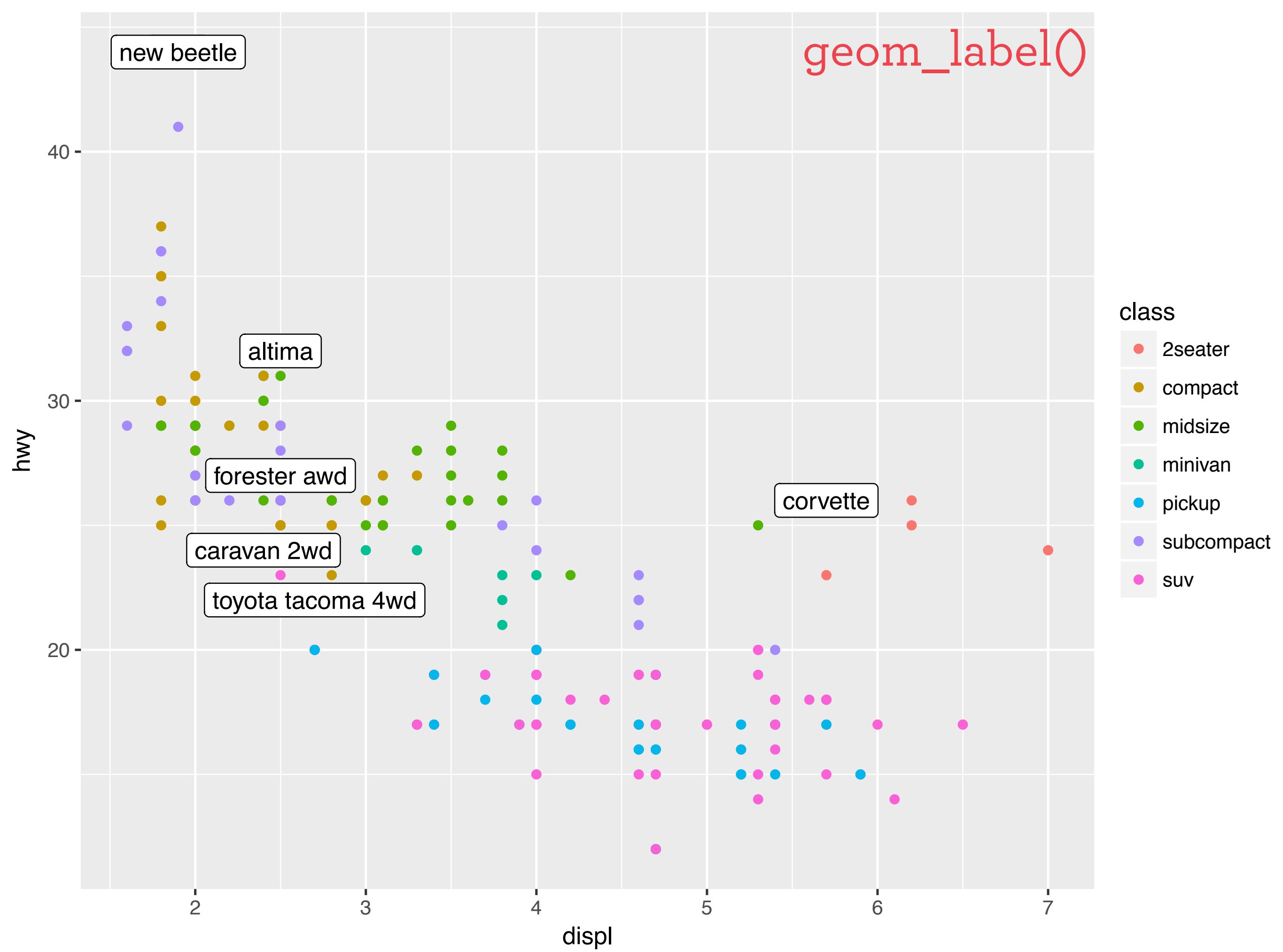
```
ggplot(mpg, aes(displ, hwy)) +  
  geom_point() +  
  scale_y_continuous(  
    "mpg",  
    sec.axis = sec_axis(  
      ~ 235 / . ,  
      name = "1 / 100 km",  
      breaks = function(x) {  
        235 / x  
      },  
      labels = 20, by = 2)  
    )  
  )
```





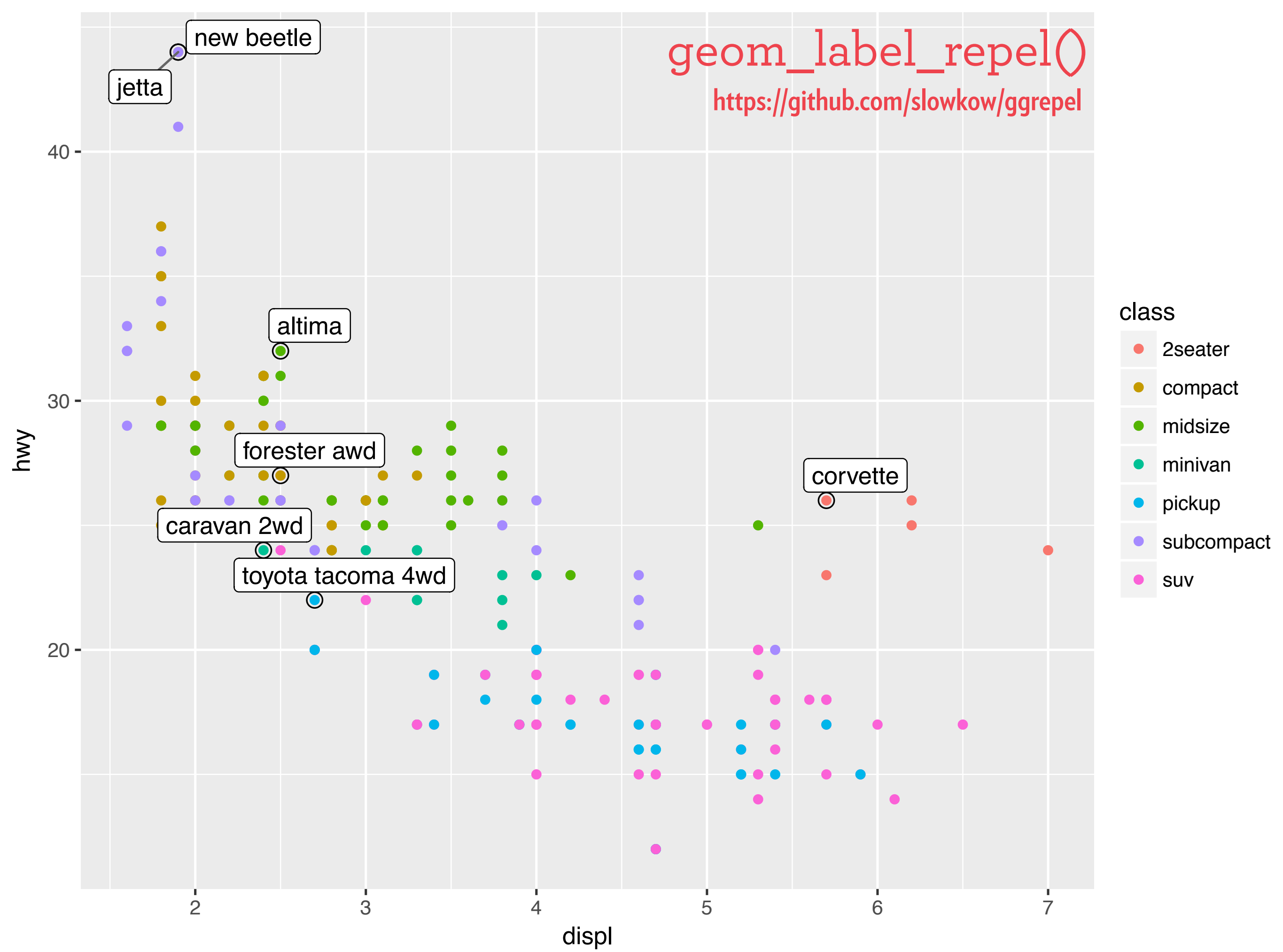
Labellirata

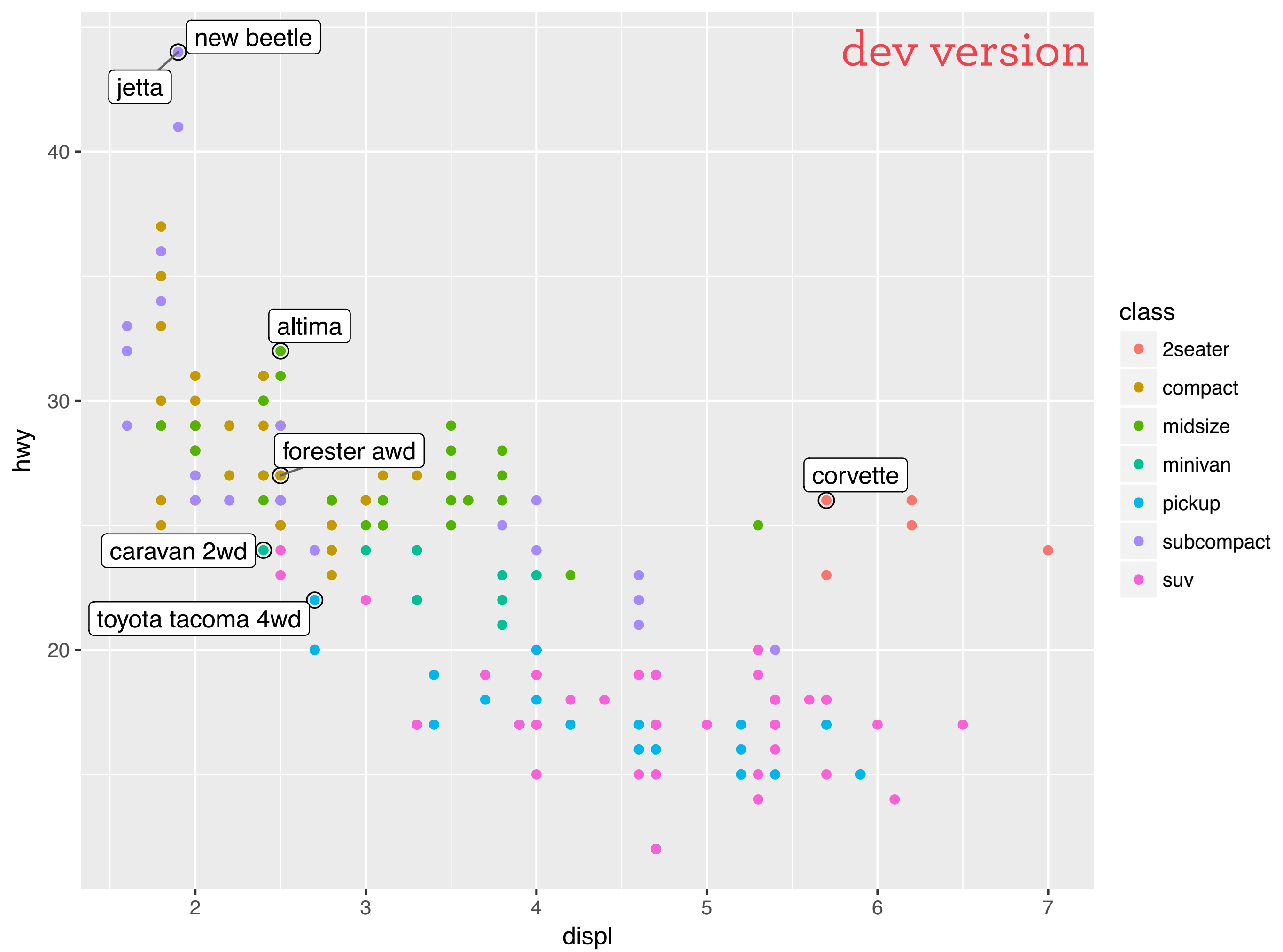




geom_label_repel()

<https://github.com/slowkow/ggrepel>





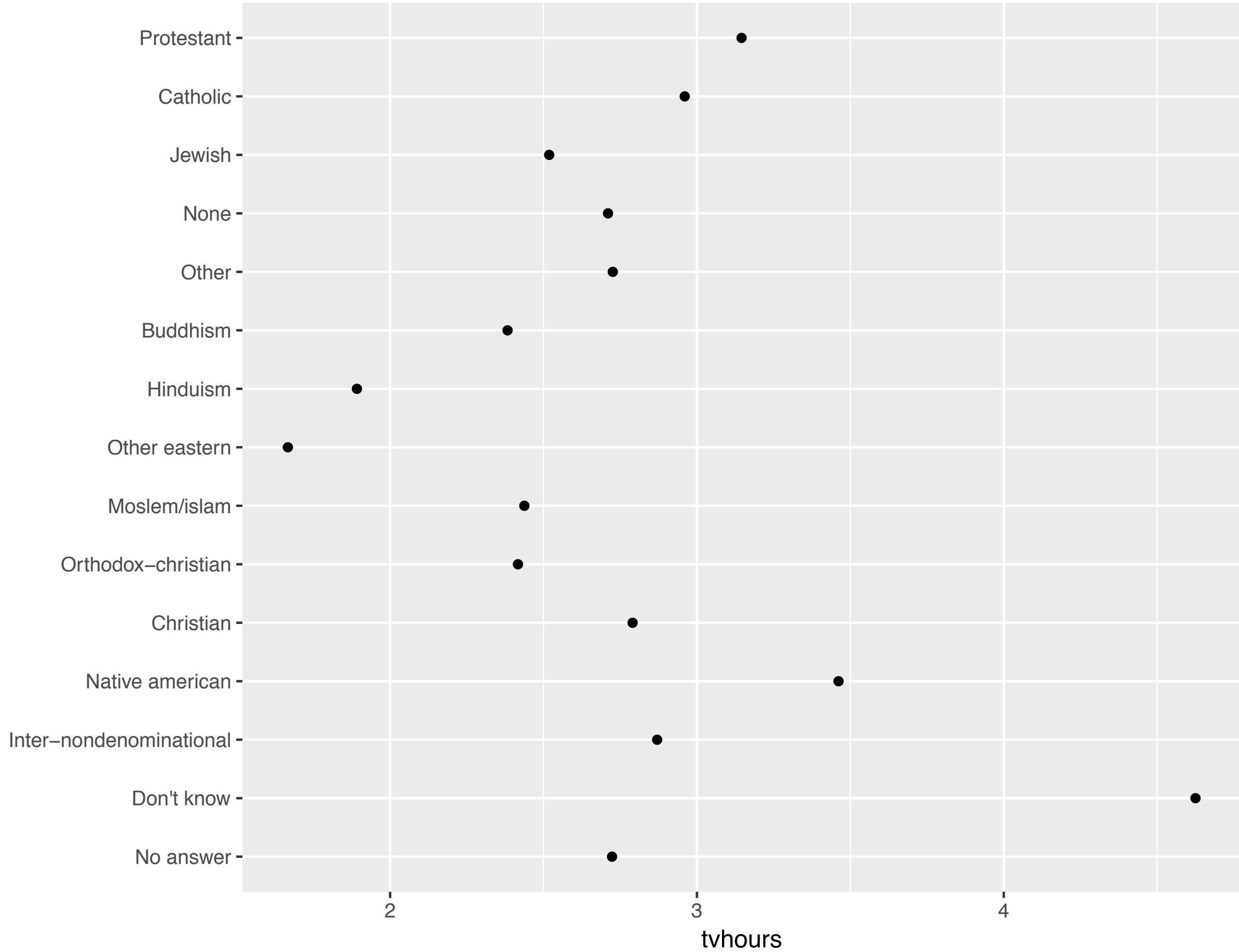


Two difference between a factor and a string:

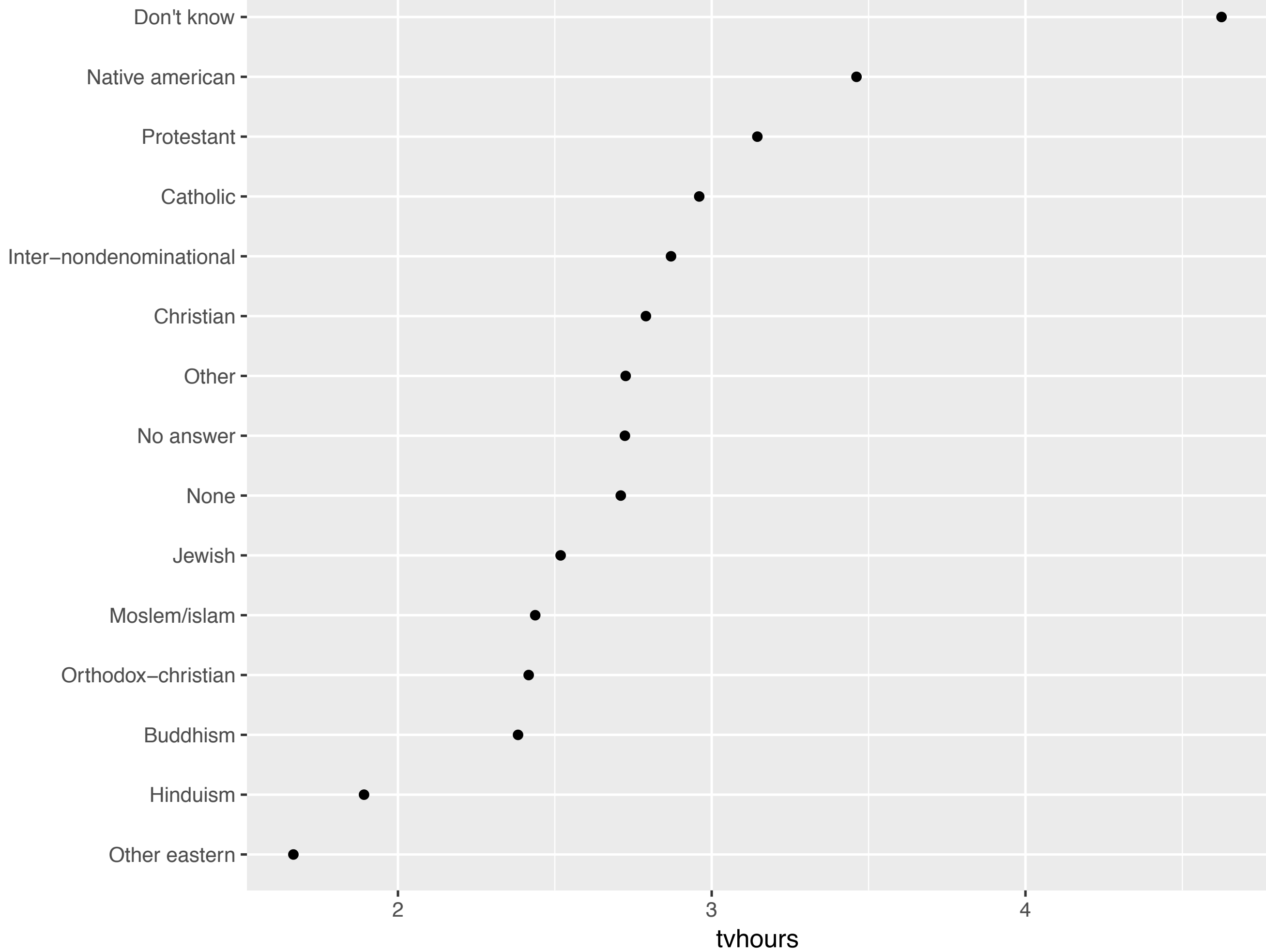
1. Fixed set of possible values
2. Arbitrary order

Some data from the general social survey

```
relig <- gss_cat %>%  
  group_by(relig) %>%  
  summarise(  
    tvhours = mean(tvhours, na.rm = TRUE),  
    n = n()  
  )
```

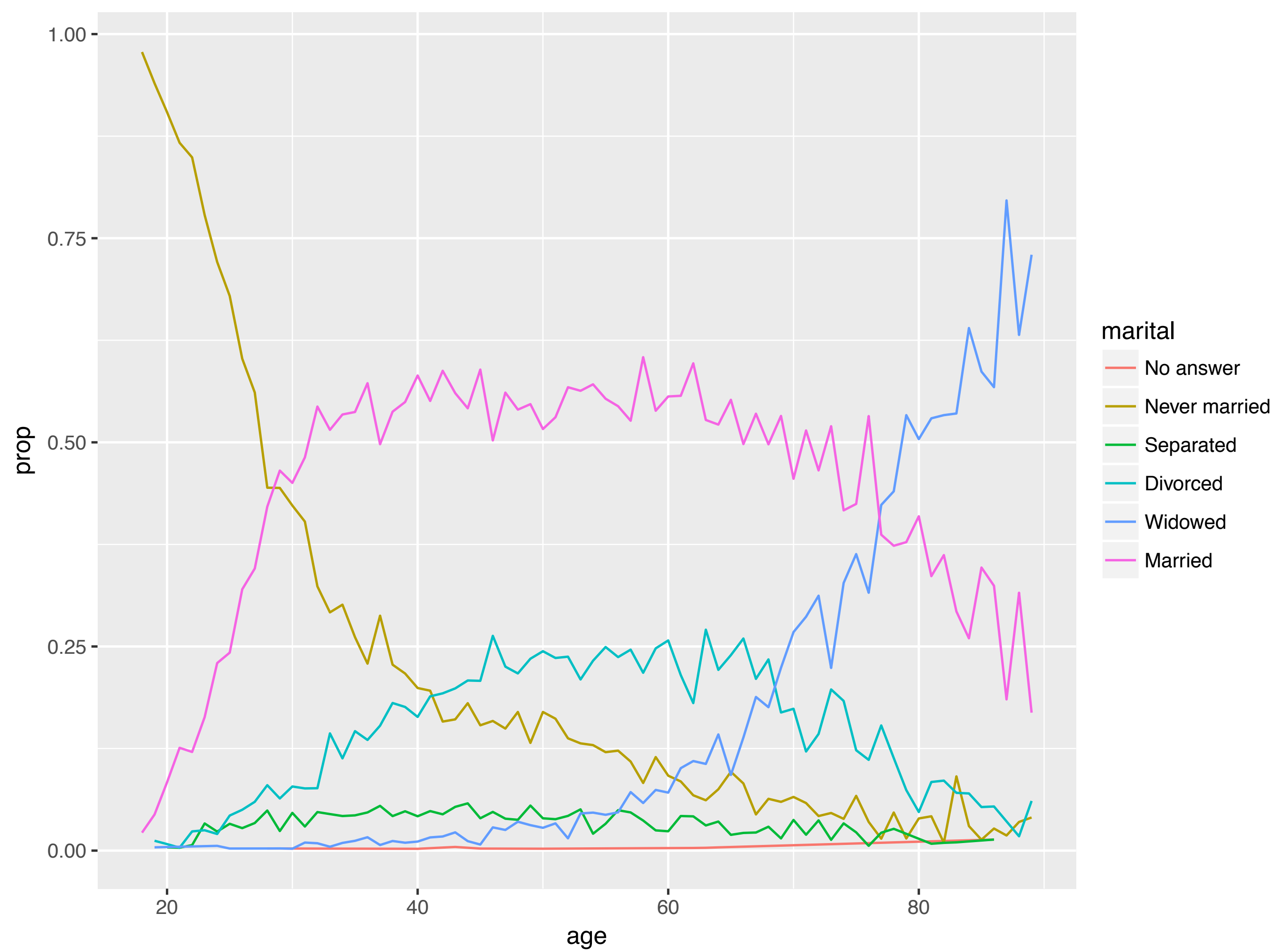


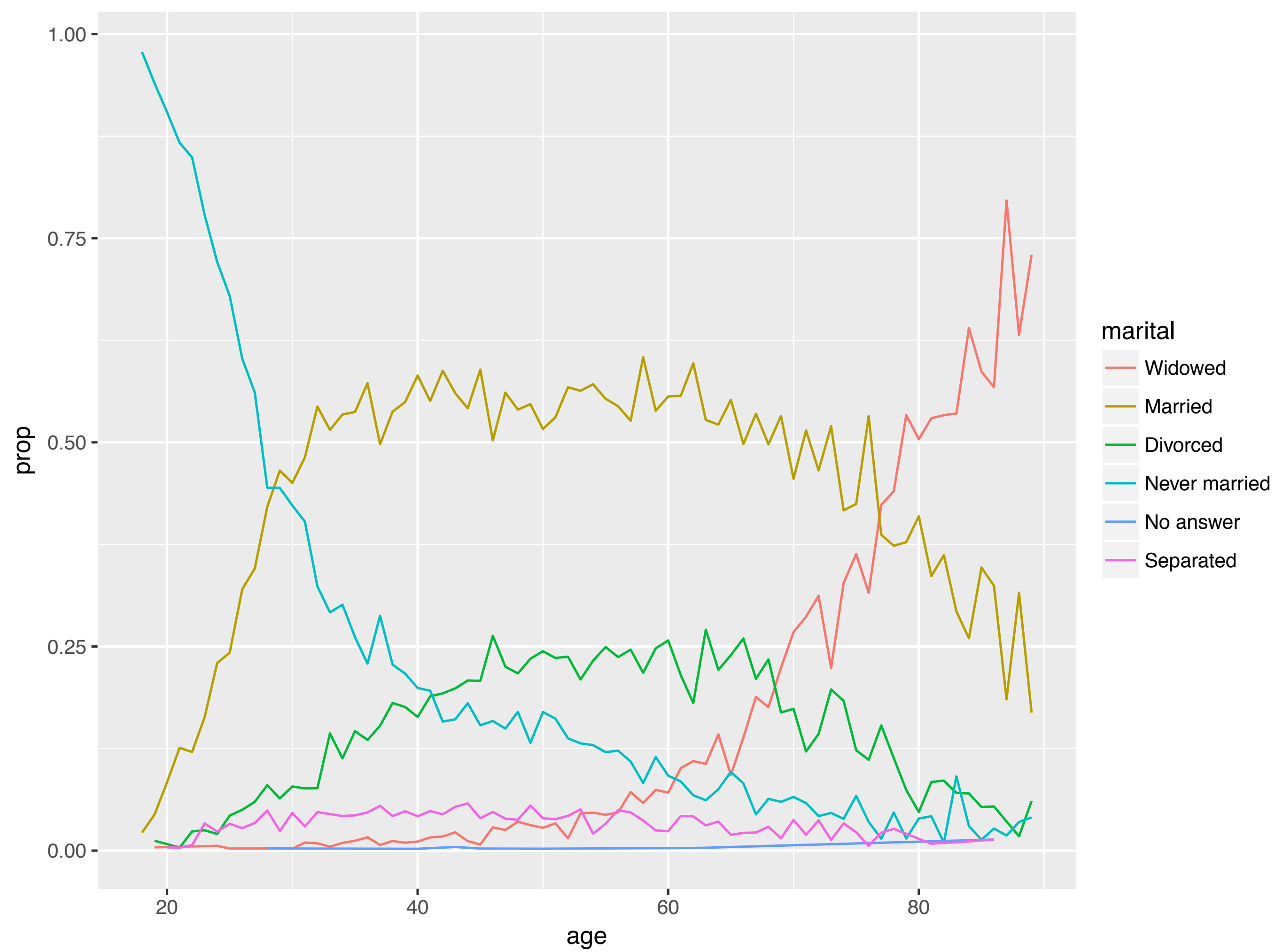
fct_reorder(relig, tvhours)



You have the same problem with more dimensions

```
by_age <- gss_cat %>%  
  filter(!is.na(age)) %>%  
  group_by(age, marital) %>%  
  count() %>%  
  mutate(prop = n / sum(n))
```







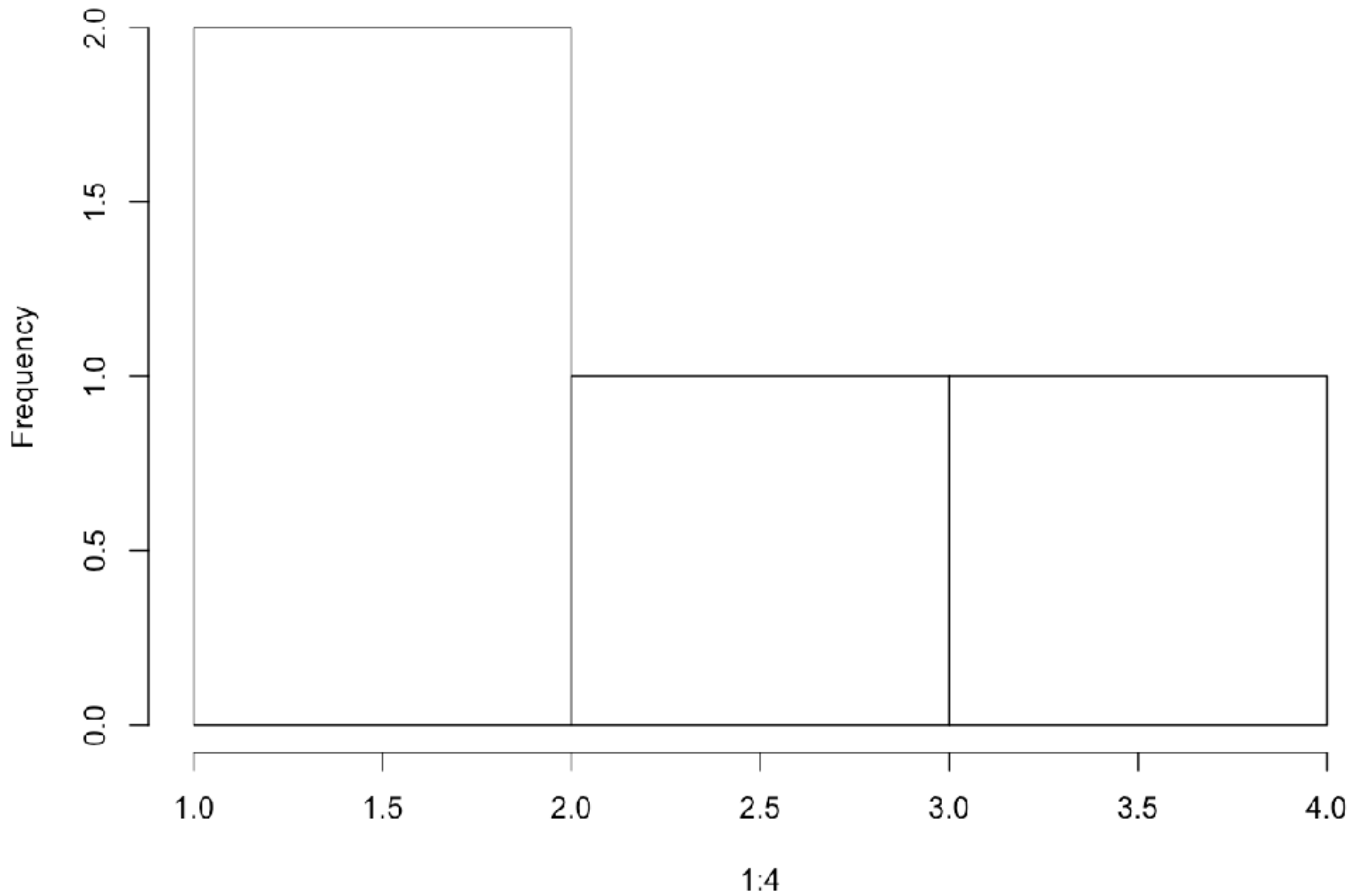
An **explicit** missing value (NA) is the presence of an absence; an **implicit** missing value is the absence of a presence.

Demo

Histograms

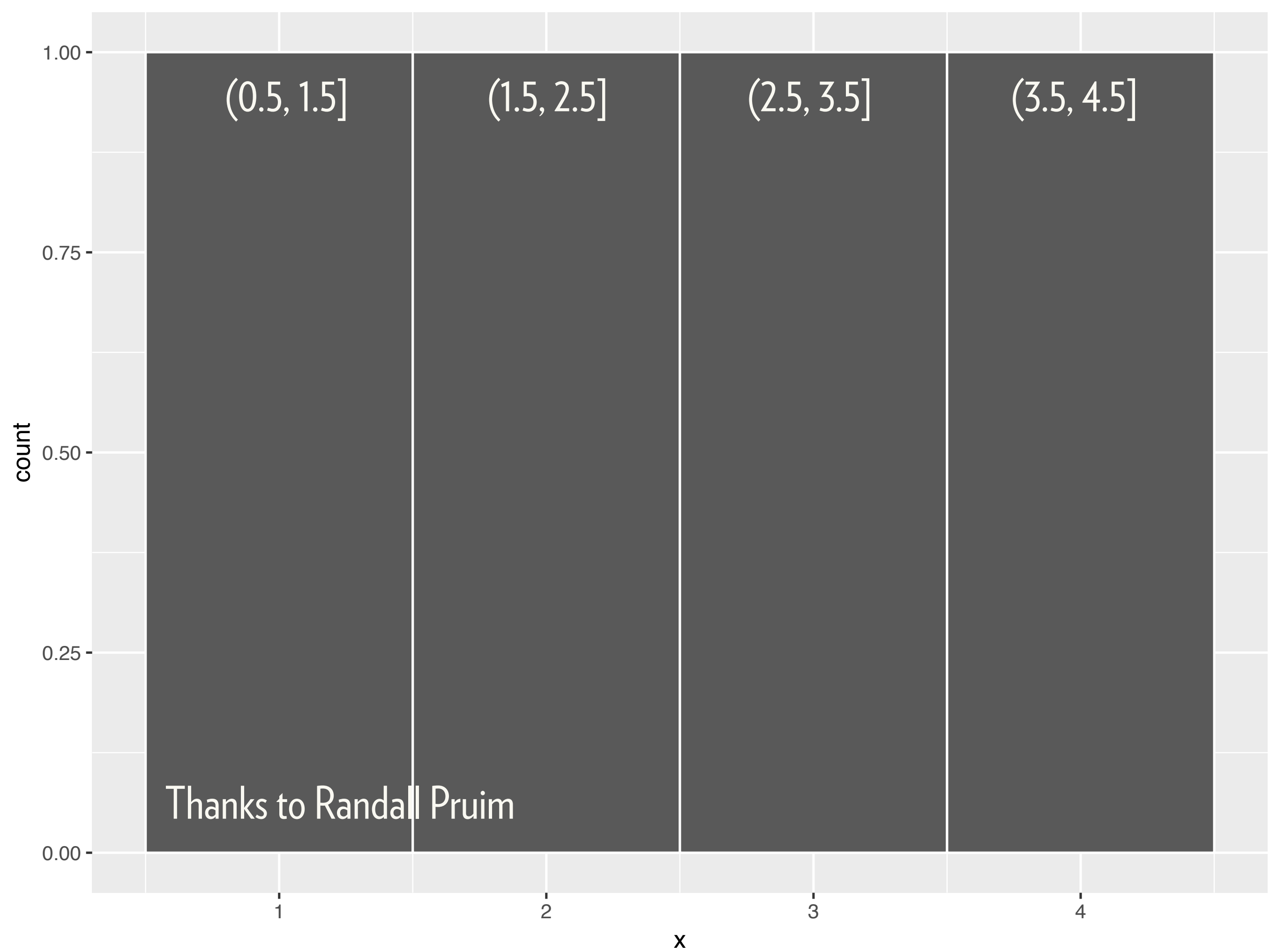


hist(1:4)



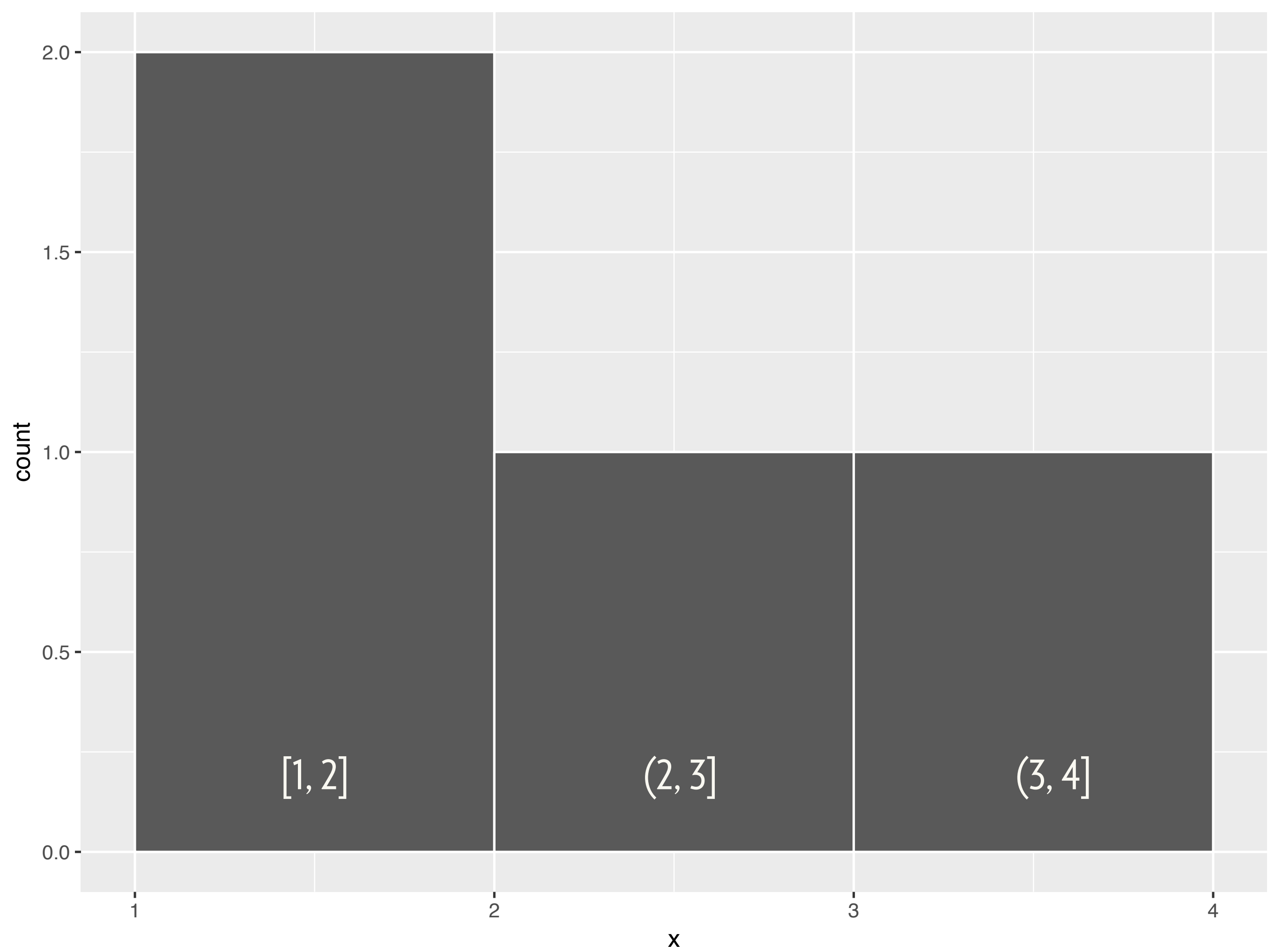
Equivalent ggplot2 code is a little longer

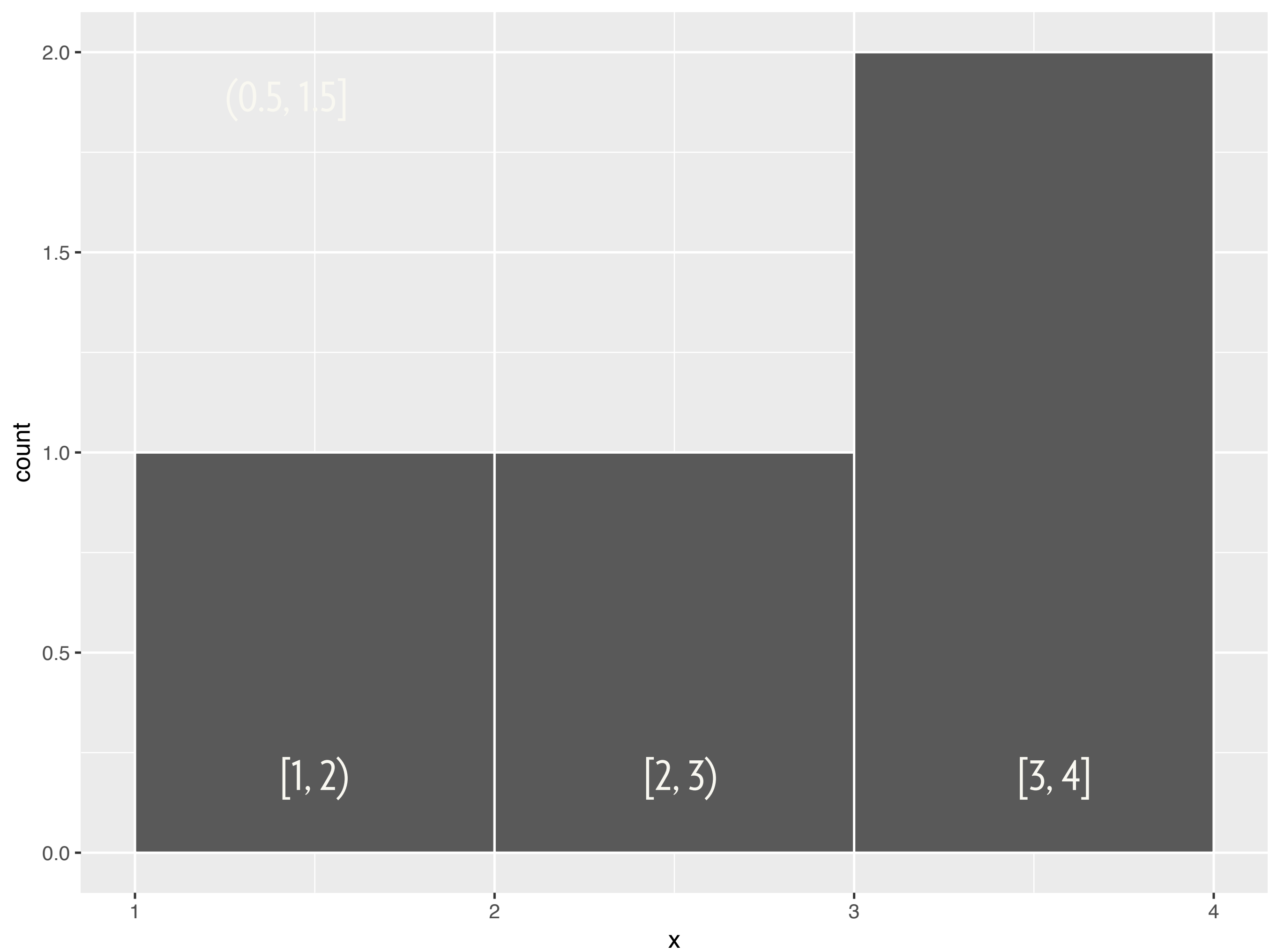
```
df <- tibble(x = 1:4)
df %>%
  ggplot(aes(x)) +
  geom_histogram(binwidth = 1)
```

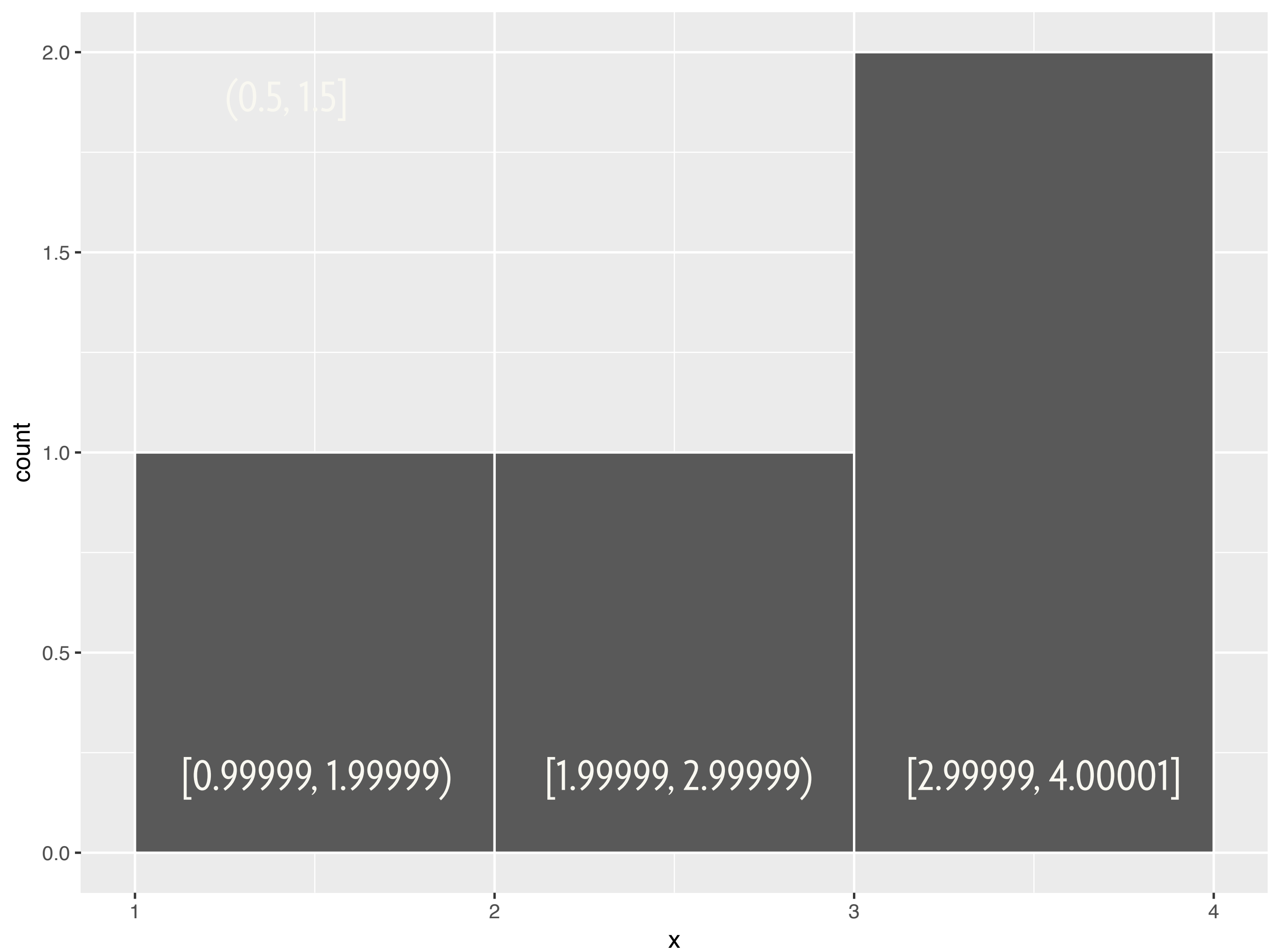



```
df %>%  
  ggplot(aes(x)) +  
  geom_histogram(  
    binwidth = 1,  
    boundary = 0  
  )
```

```
df %>%  
  ggplot(aes(x)) +  
  geom_histogram(  
    binwidth = 1,  
    boundary = 0,  
    closed = "left"  
  )
```







Bar



charts

```
ggplot(mpg, aes(class)) +  
  geom_bar(colour = "white")
```

count

60
40
20
0

2seater

compact

midsize

minivan

pickup

subcompact

suv

class



```
ggplot(mpg, aes(class, group = id)) +  
  geom_bar(col = "white")
```

count

60
40
20
0

2seater

compact

midsize

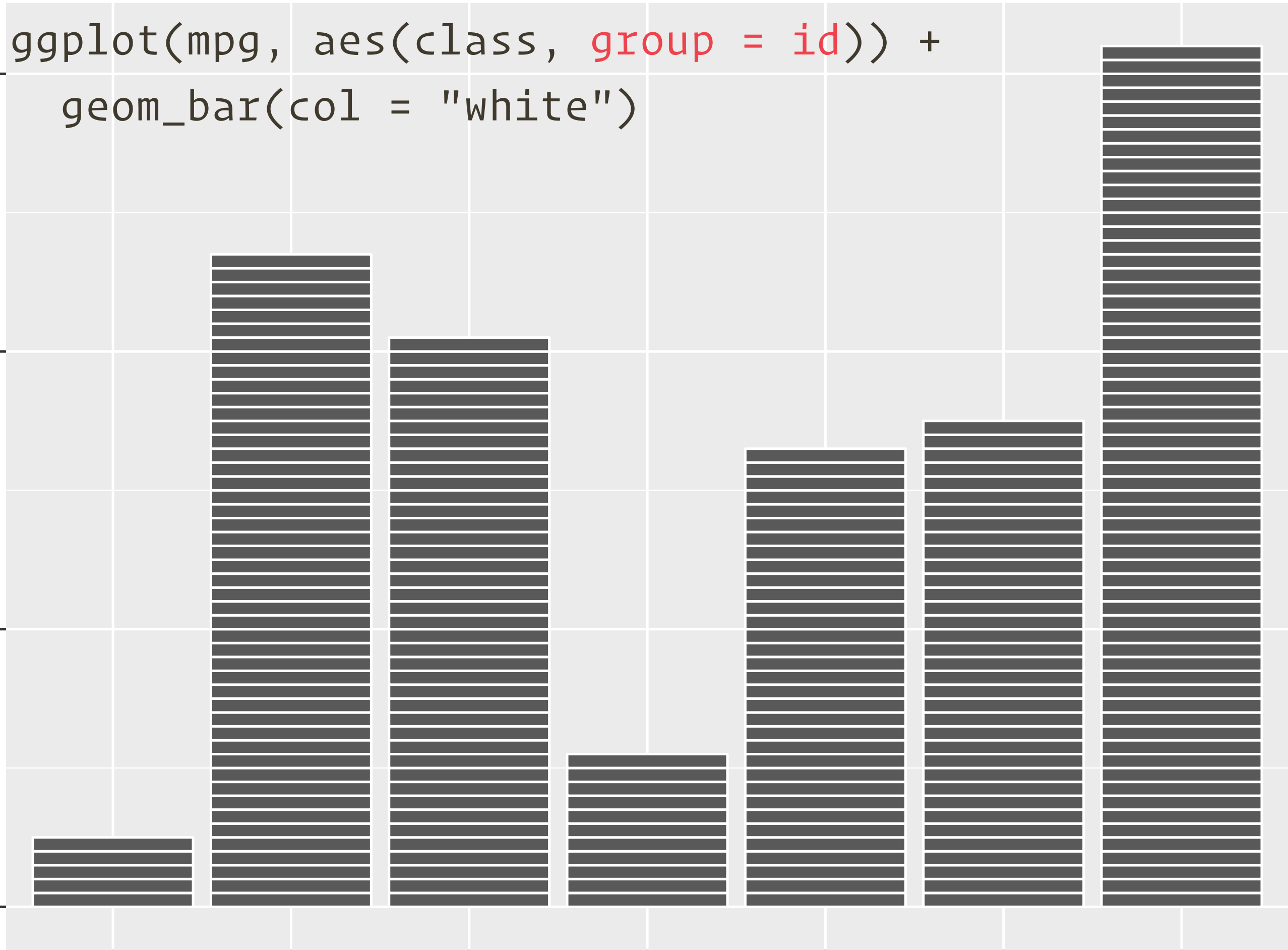
minivan

pickup

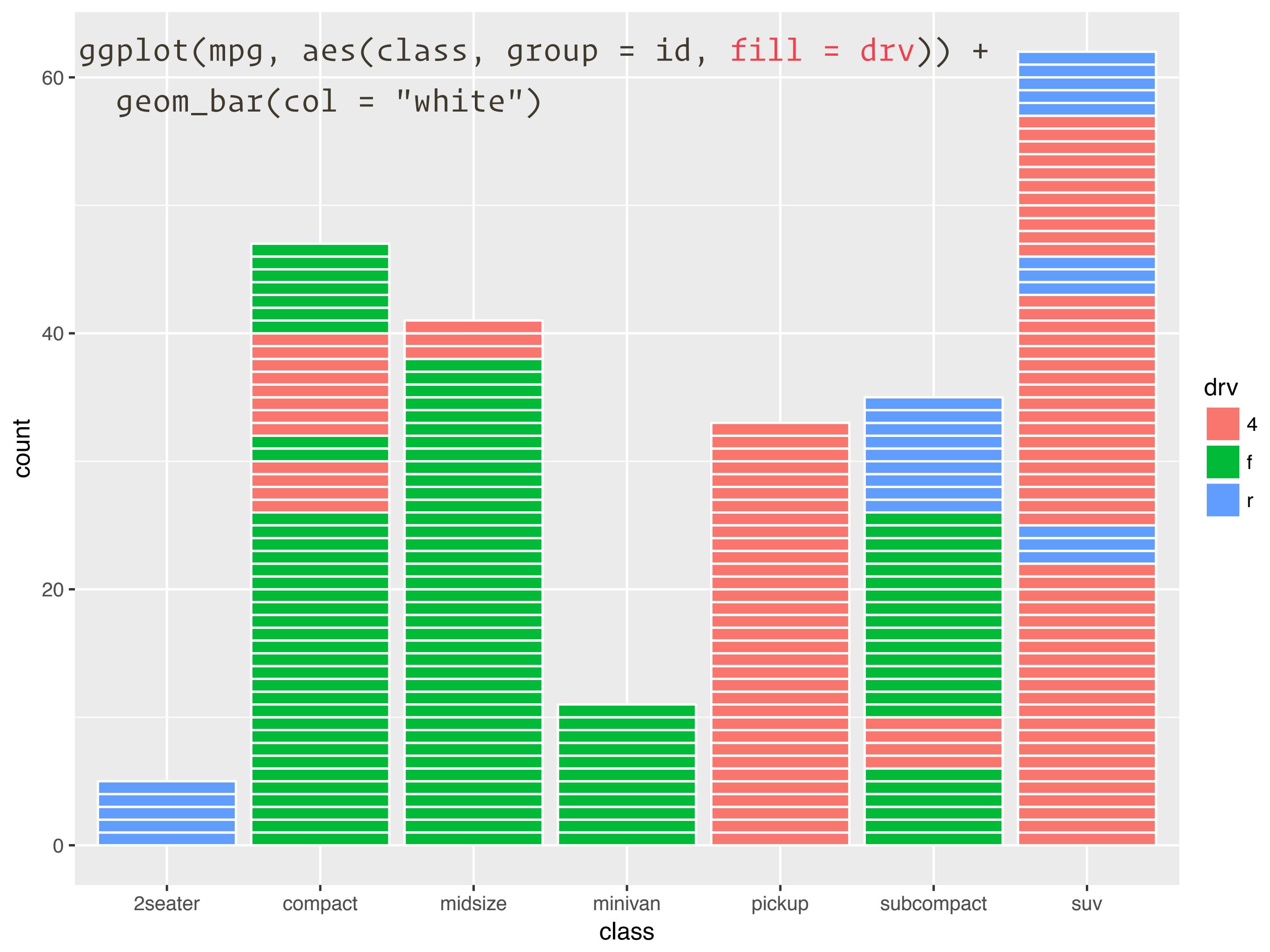
subcompact

suv

class



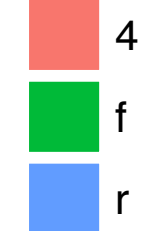

```
ggplot(mpg, aes(class, group = id, fill = drv)) +  
  geom_bar(col = "white")
```



```
ggplot(mpg, aes(class, fill = drv)) +  
  geom_bar(col = "white")
```

count

drv



60
40
20
0

2seater

compact

midsize

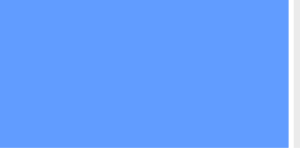
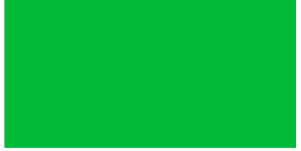
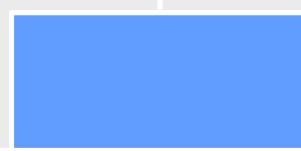
minivan

pickup

subcompact

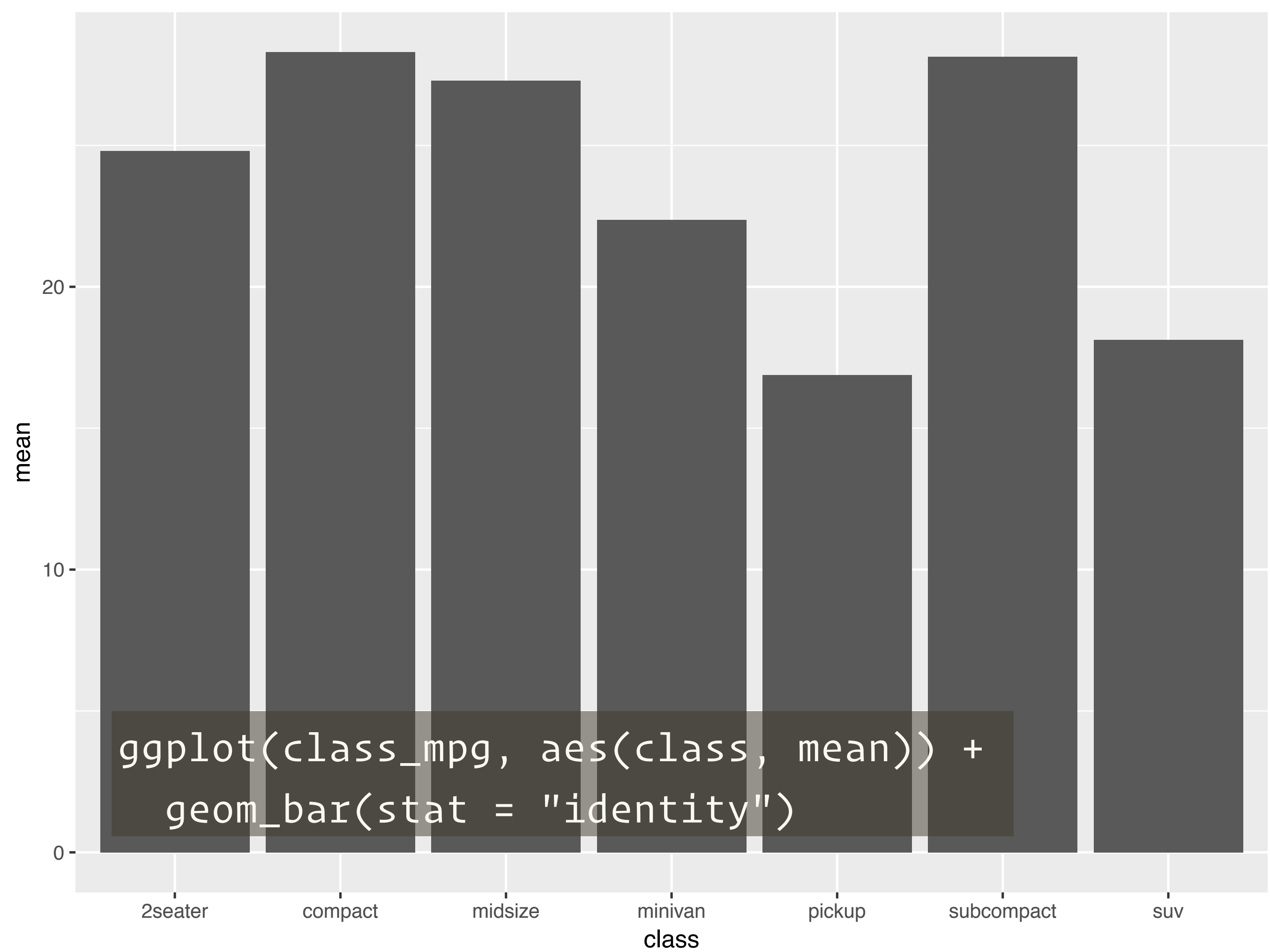
suv

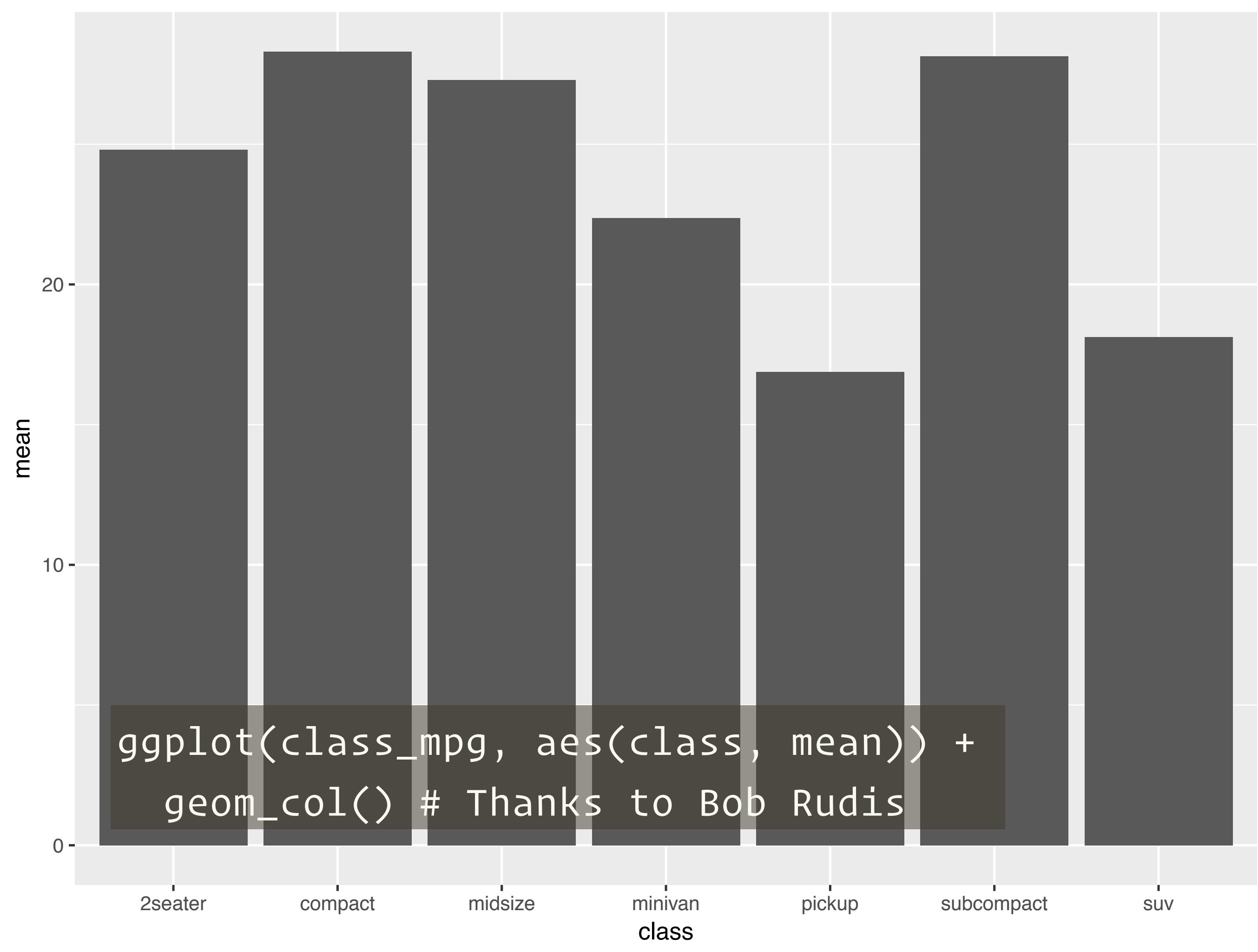
class

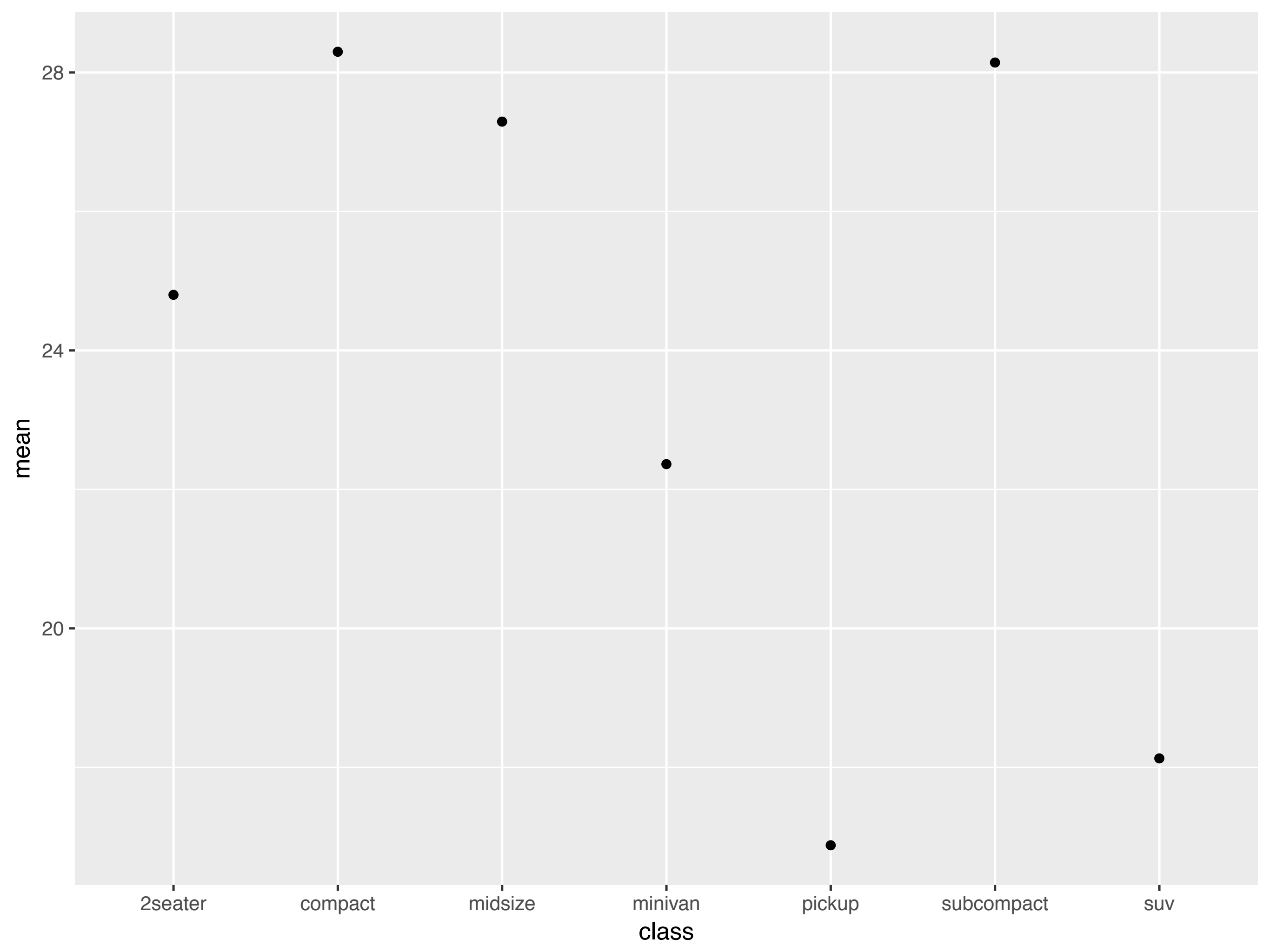


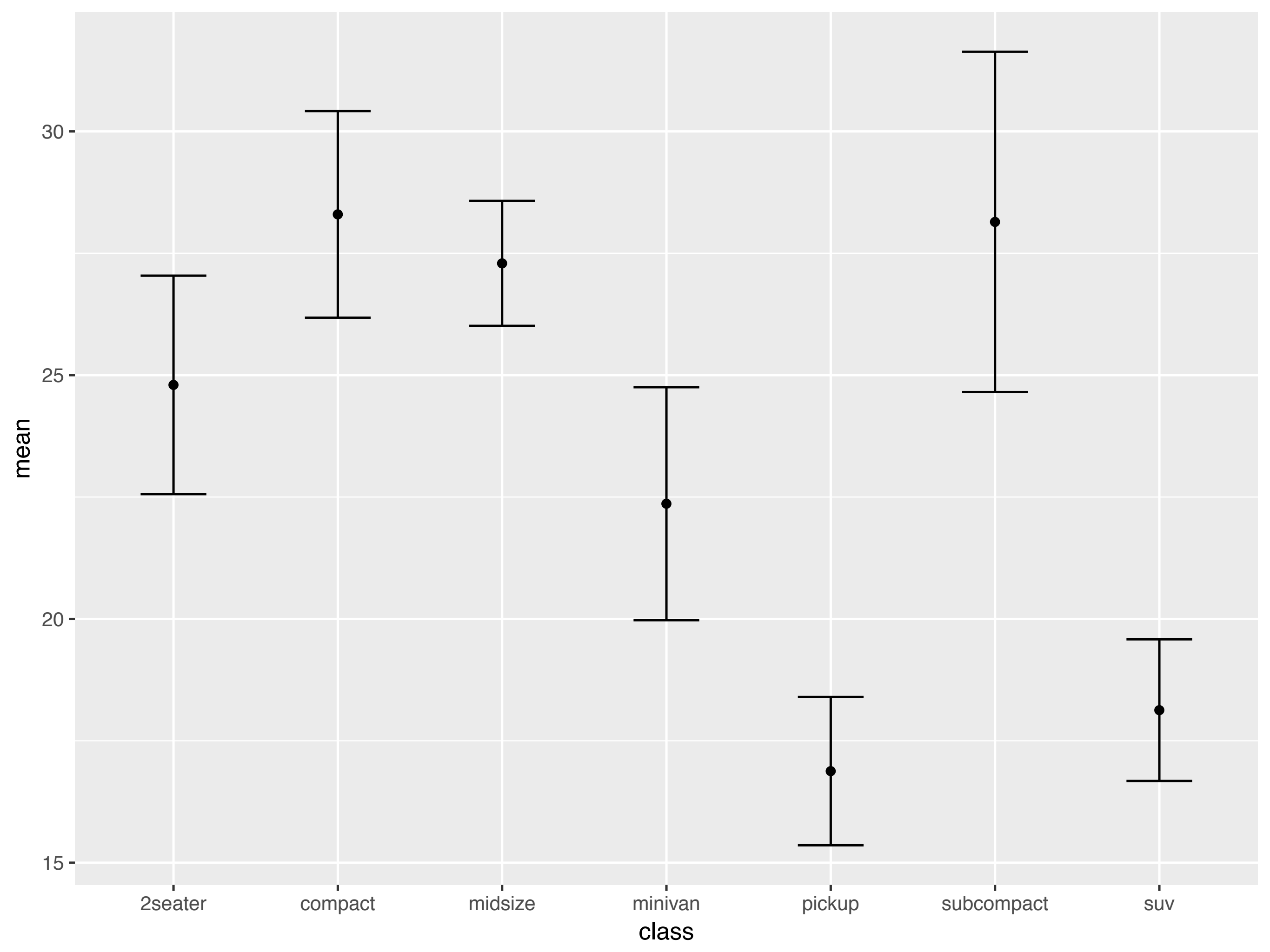
Another type of bar chart displays summaries

```
class_mpg <- mpg %>%  
  group_by(class) %>%  
  summarise(  
    mean = mean(hwy),  
    se = 1.96 * sd(hwy) / sqrt(n())  
  )
```









8

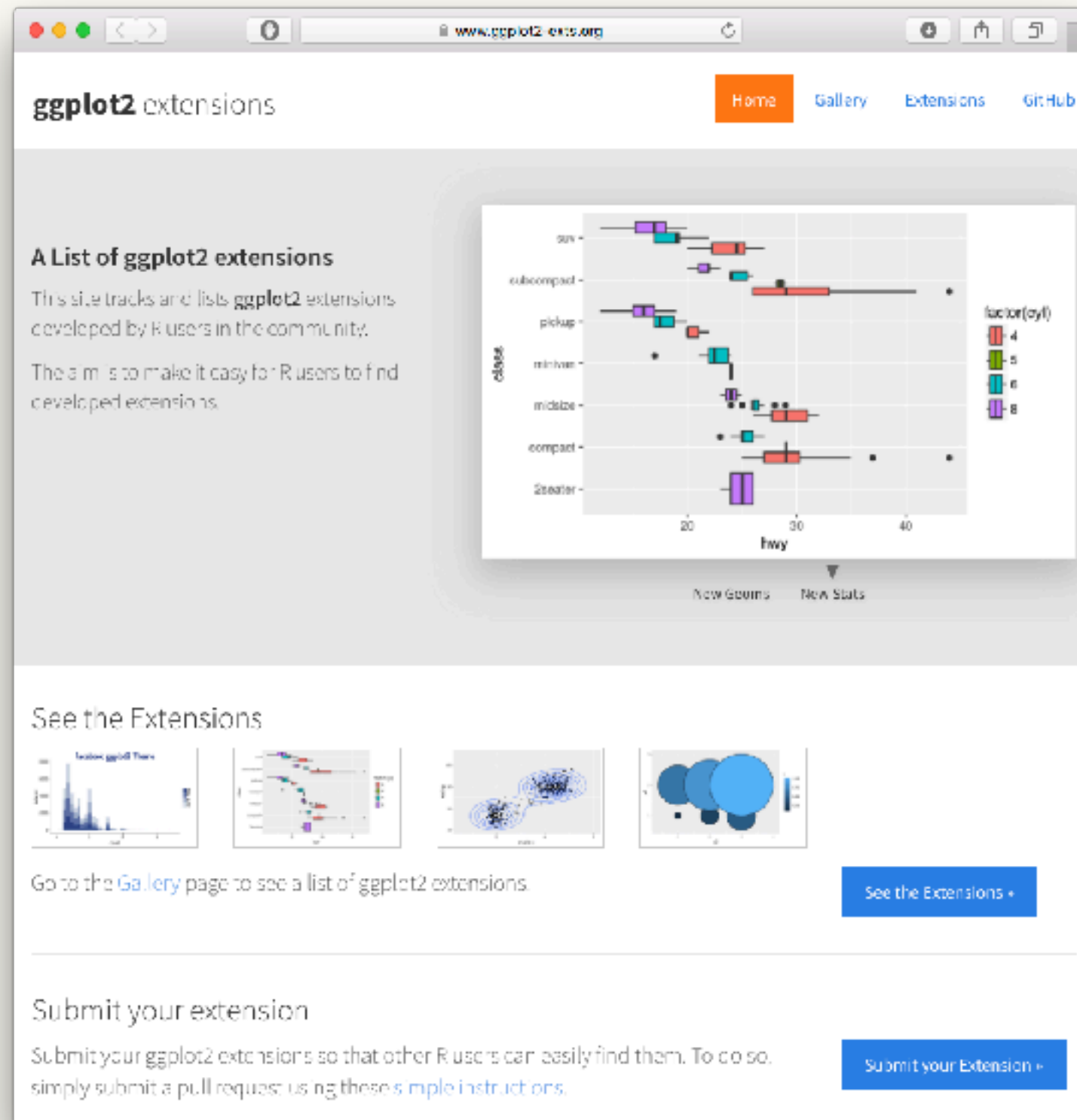
9

10

11

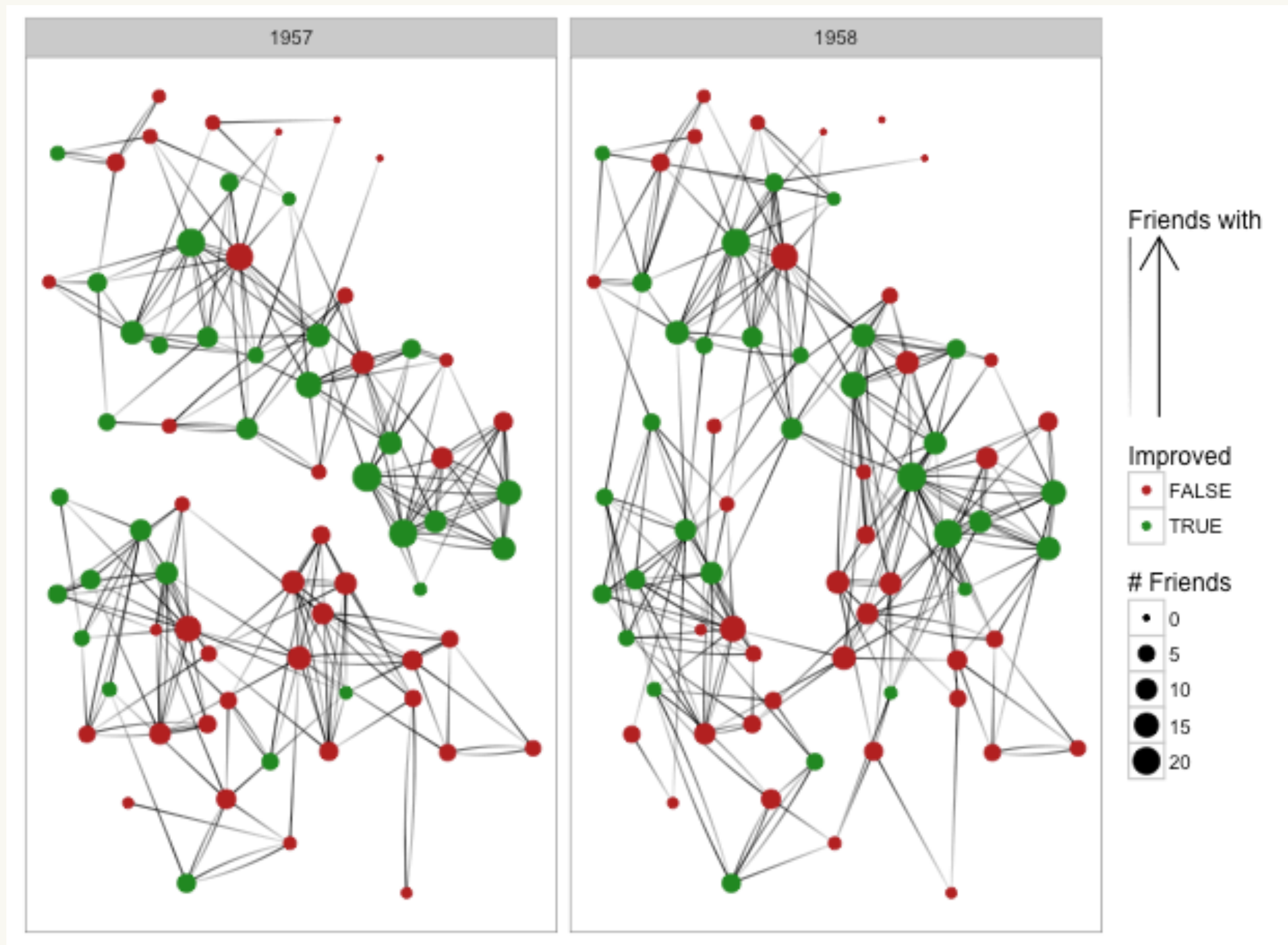
ggplot2
extensions

2.1.0 introduced a formal extension mechanism



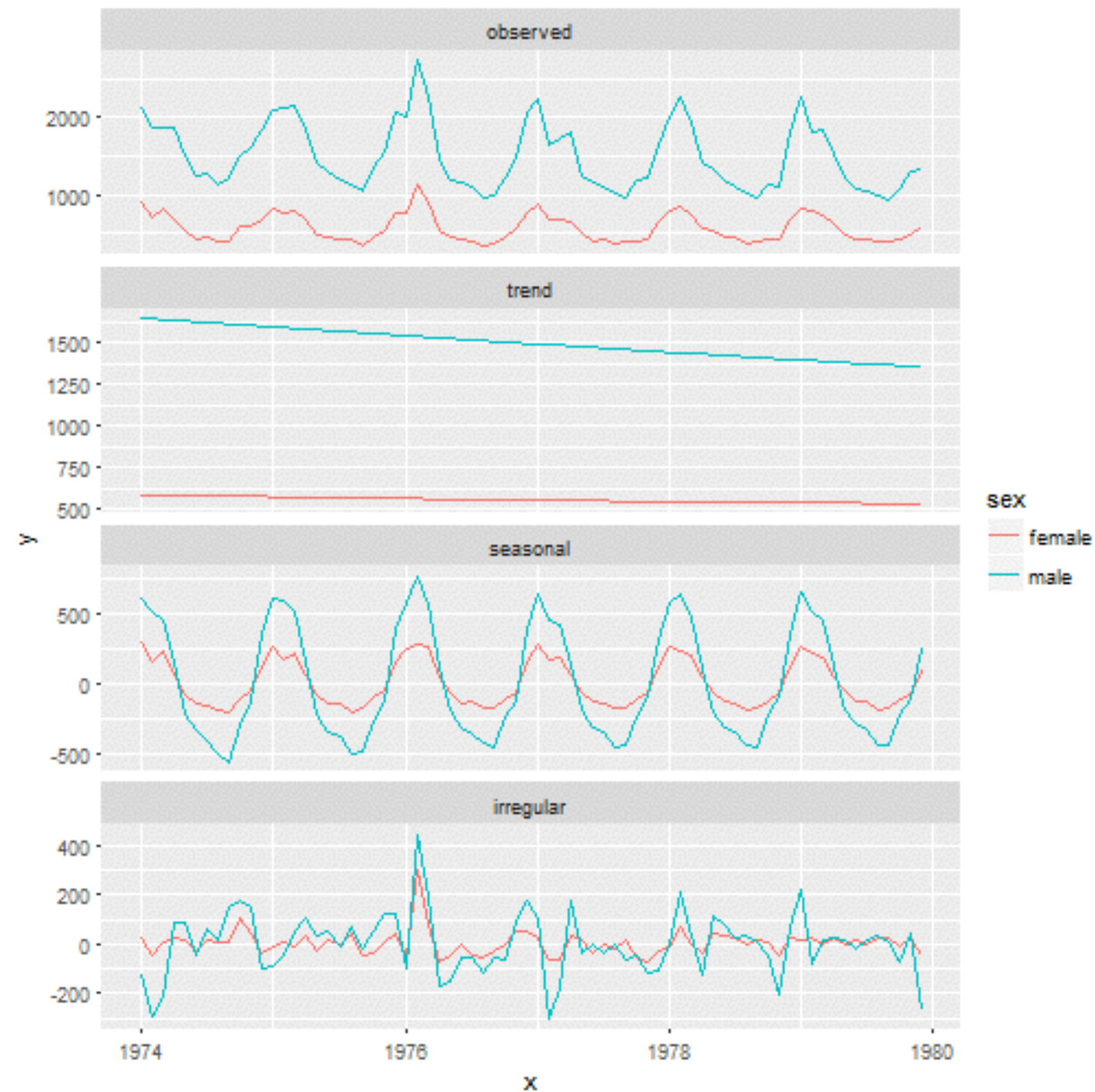
<https://www.ggplot2-exts.org>, by Daniel Emaasit

ggraph, by Thomas Lin Pedersen

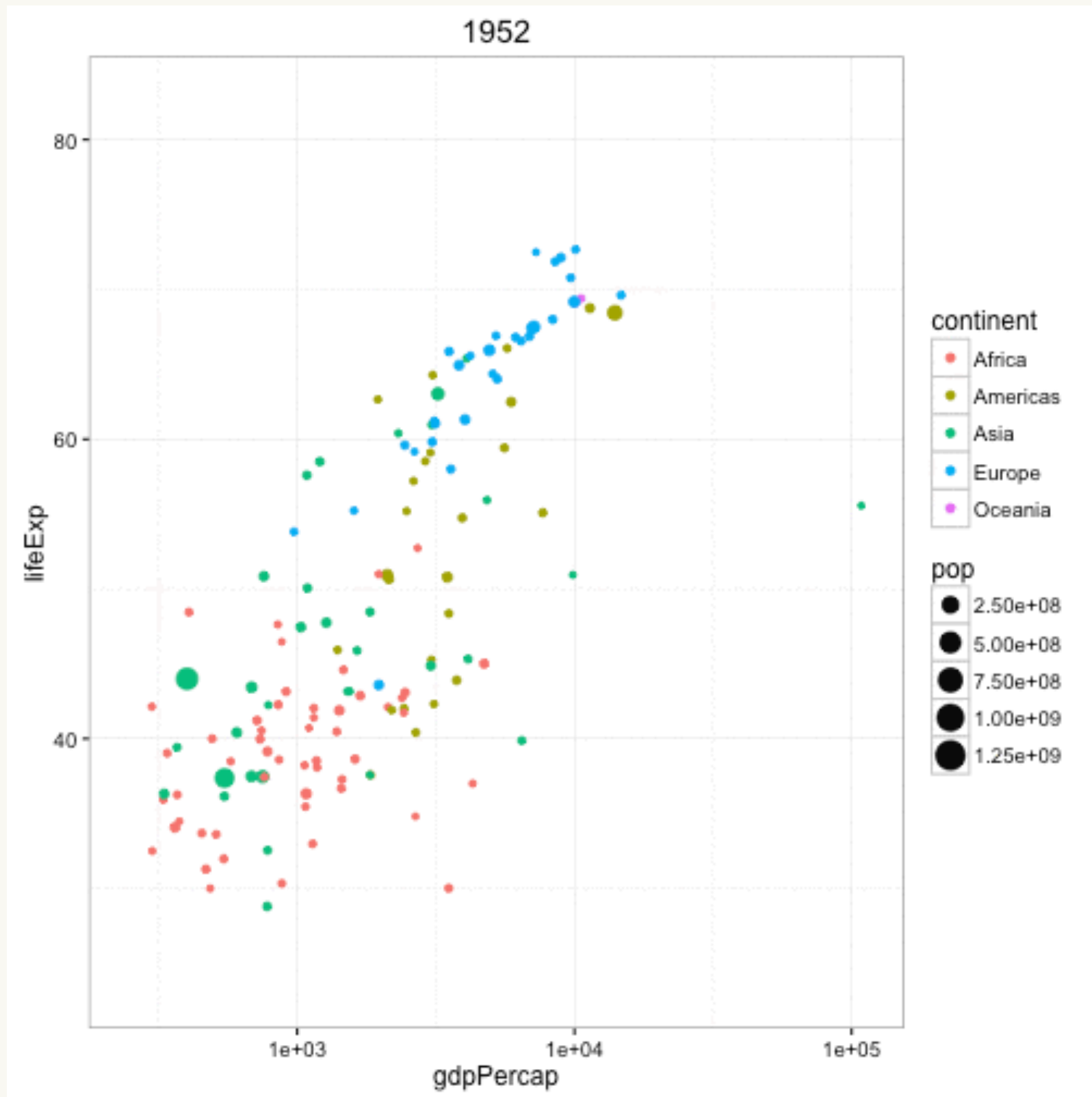


ggseas by Peter Ellis

Uses X13-SEATS-ARIMA
in seasonal package



gganimate by David Robinson



Conclusion



Labelling plots

A problem ignored for too long

Solved by Bob Rudis

2 Axes



Labellirata





Histograms



Bar



charts

8

9

10

11

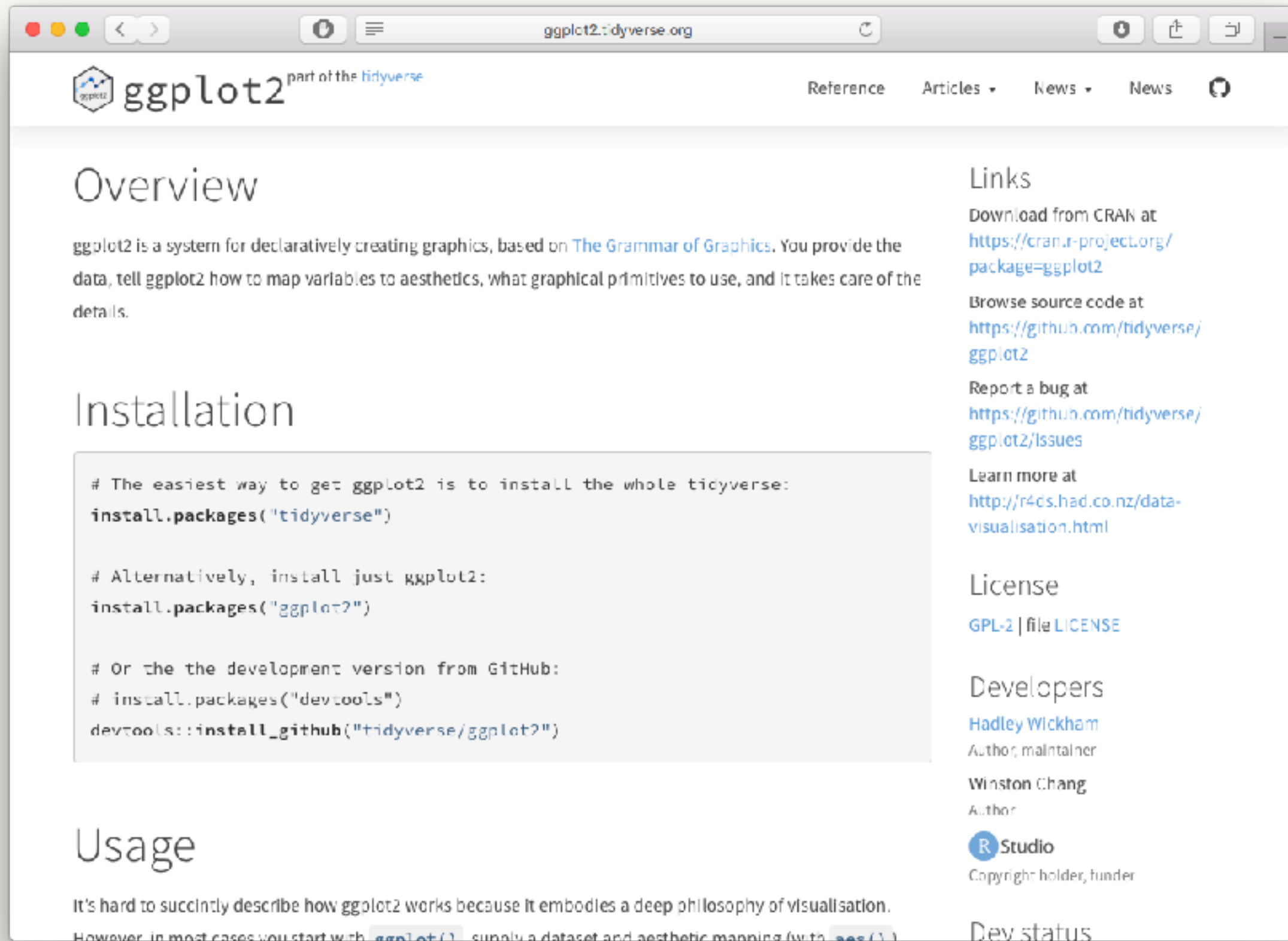
ggplot2
extensions

Many of the features I discussed here have been added in recent versions of ggplot2.

See the **release notes** for more detail.



http://ggplot2.tidyverse.org



The screenshot shows the homepage of the ggplot2 website. The browser's address bar displays 'ggplot2.tidyverse.org'. The website header includes the 'ggplot2' logo, a note 'part of the tidyverse', and navigation links for 'Reference', 'Articles', 'News', and 'News'. The main content area is divided into two columns. The left column contains sections for 'Overview', 'Installation', and 'Usage'. The 'Installation' section features a code block with three methods of installing the package. The right column contains sections for 'Links', 'License', 'Developers', and 'Dev status'. The 'Links' section provides URLs for downloading from CRAN, browsing source code on GitHub, reporting bugs, and learning more. The 'License' section links to the GPL-2 license. The 'Developers' section lists Hadley Wickham as the author and maintainer, and Winston Chang as another author. The 'Dev status' section is partially visible at the bottom.

Overview

ggplot2 is a system for declaratively creating graphics, based on [The Grammar of Graphics](#). You provide the data, tell ggplot2 how to map variables to aesthetics, what graphical primitives to use, and it takes care of the details.

Installation

```
# The easiest way to get ggplot2 is to install the whole tidyverse:  
install.packages("tidyverse")  
  
# Alternatively, install just ggplot2:  
install.packages("ggplot2")  
  
# Or the the development version from GitHub:  
# install.packages("devtools")  
devtools::install_github("tidyverse/ggplot2")
```

Usage

It's hard to succinctly describe how ggplot2 works because it embodies a deep philosophy of visualisation. However, in most cases you start with `ggplot()`, supply a dataset and aesthetic mapping (with `aes()`).

Links

Download from CRAN at <https://cran.r-project.org/package=ggplot2>

Browse source code at <https://github.com/tidyverse/ggplot2>

Report a bug at <https://github.com/tidyverse/ggplot2/issues>

Learn more at <http://r4ds.had.co.nz/data-visualisation.html>


License

[GPL-2](#) | file [LICENSE](#)

Developers

[Hadley Wickham](#)
Author, maintainer

[Winston Chang](#)
Author

 **Studio**
Copyright holder, funder

Dev status