

Some data viz advice

Daniel Anderson

Monday, May 13, 2021

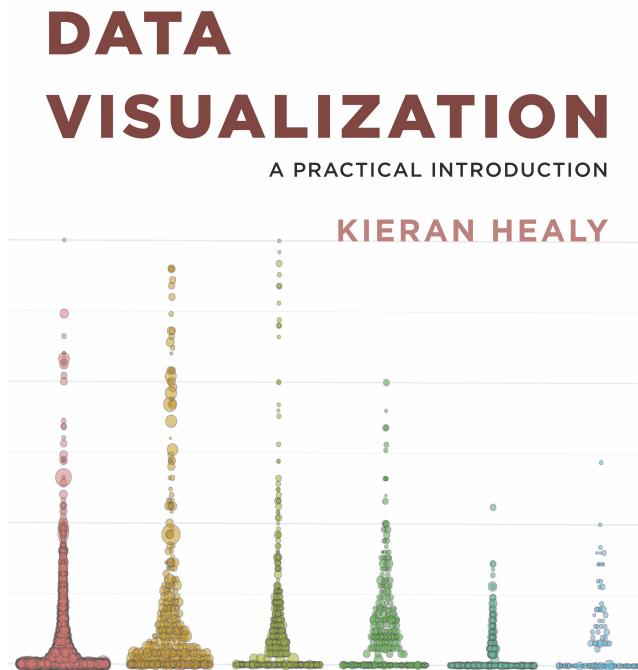
whoami

- Research Assistant
Professor: Behavioral
Research and Teaching
- Dad (two daughters: 8
and 6)
- Pronouns: he/him/his
- Primary areas of interest:
   R  ,
computational research,
systemic inequities in
educational systems and
variance between
educational institutions

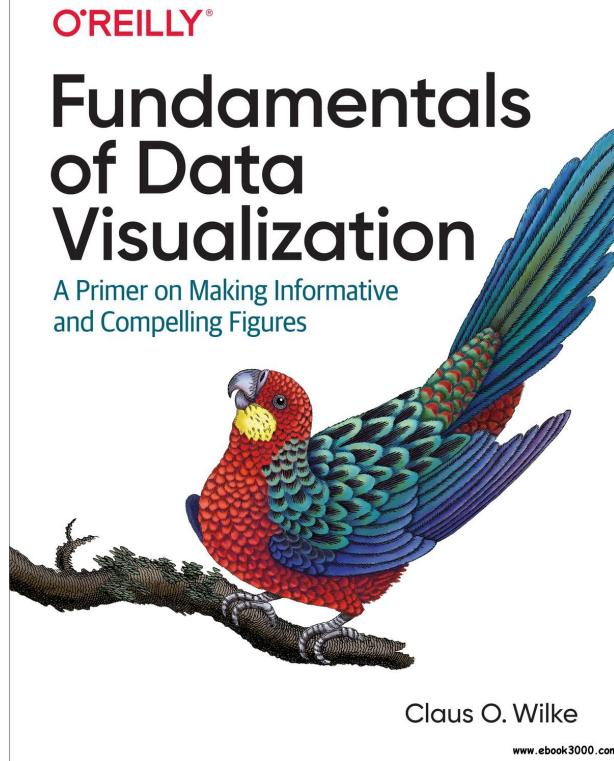


Resources (free)

Healy



Wilke



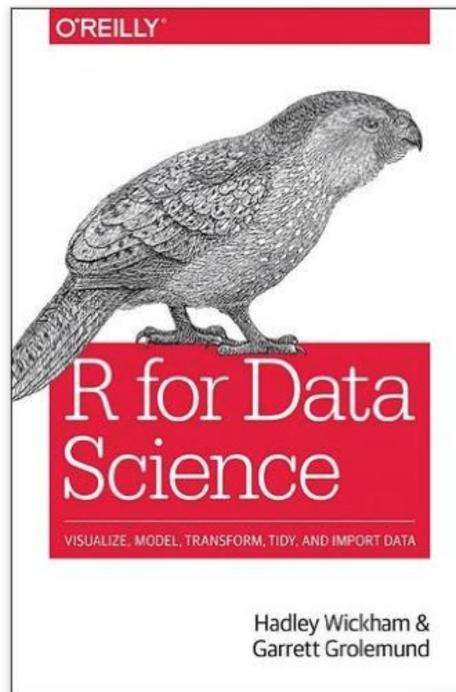
Other Resources

- My classes!
- Sequence
 - EDLD 651: Introductory Educational Data Science (EDS)
 - EDLD 652: Data Visualization for EDS
 - EDLD 653: Functional Programming for EDS
 - EDLD 654: Machine Learning for EDS
 - Capstone

Where to start?

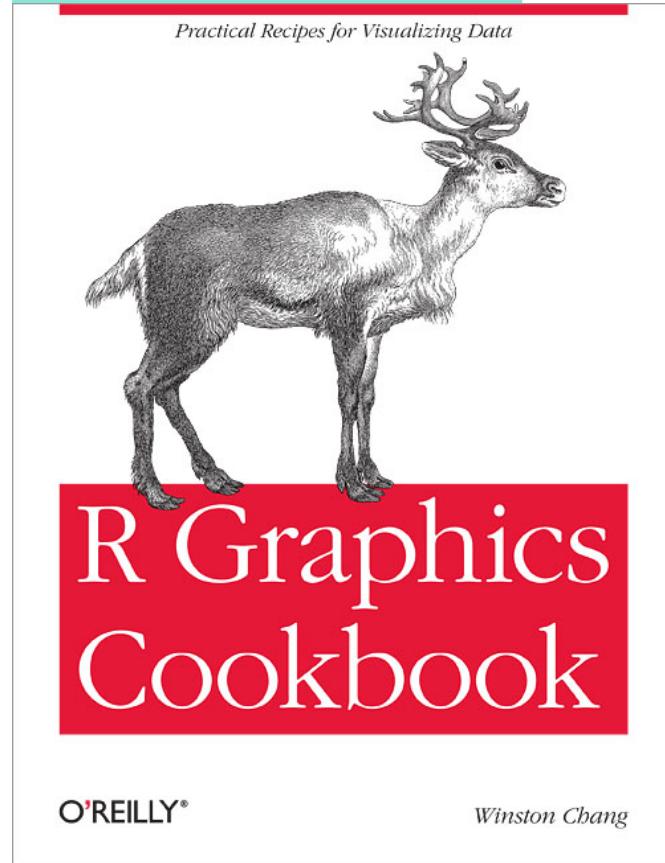
- I *really* recommend moving to R as quickly as possible, if you haven't already

R for Data Science



ggplot2!

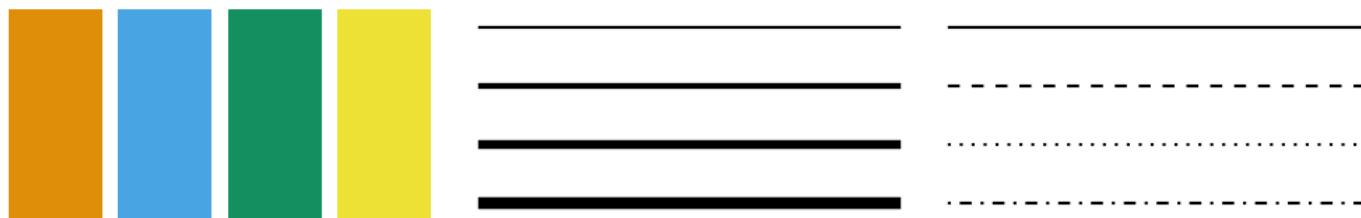
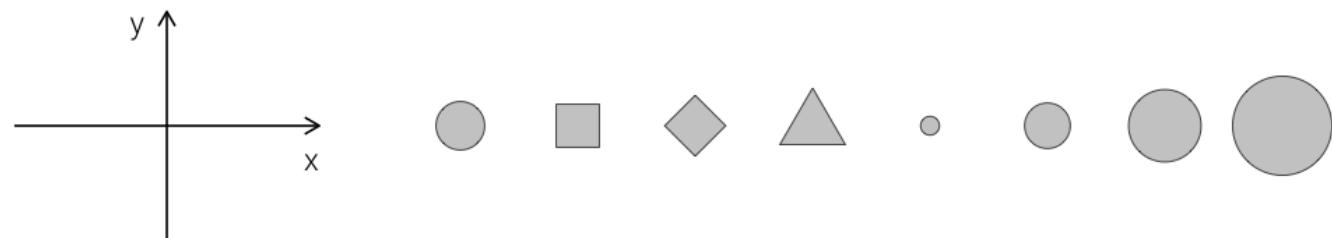
<https://r-graphics.org>



Last note before we really start

- These slides were produced with R
- See the source code [here](#)
- The focus of this particular talk is not on the code itself

Different ways of encoding data



Other elements to consider

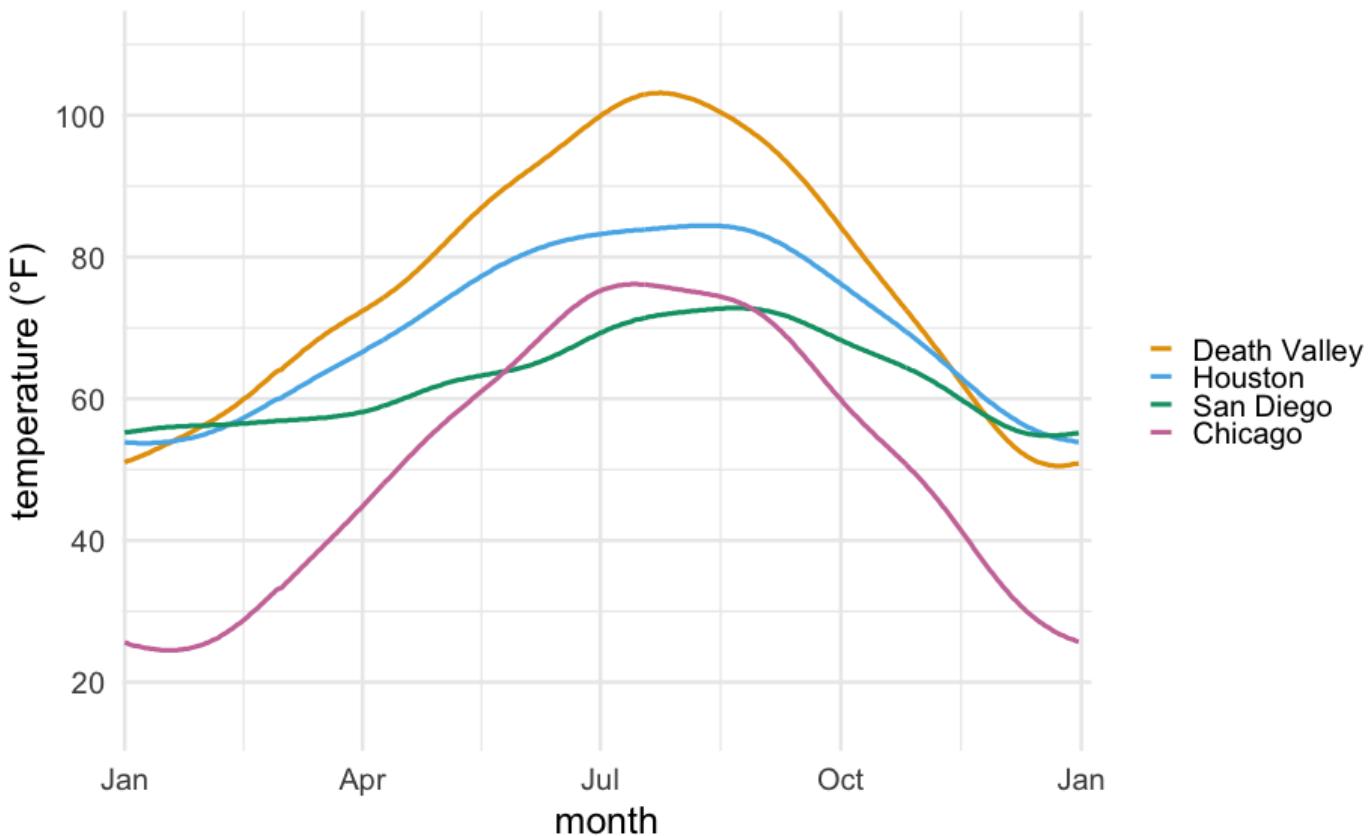
- Text
 - How is the text displayed (e.g., font, face, location)?
 - What is the purpose of the text?
- Transparency
 - Are there overlapping pieces?
 - Can transparency help?

Encoding data?

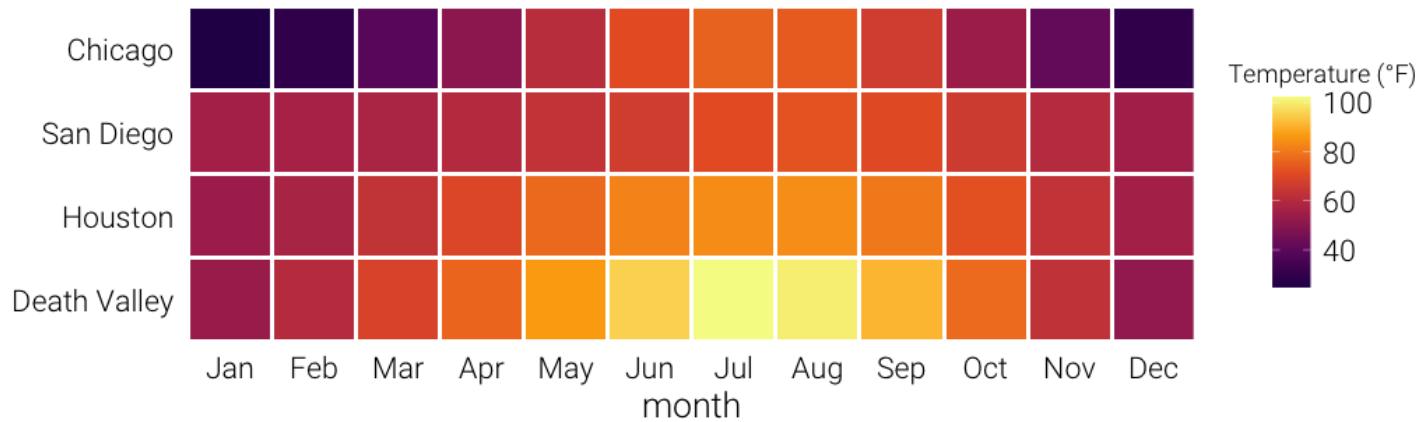
How might we encode these data into a display?

| Month | Day | Location | Temperature |
|-------|-----|--------------|-------------|
| Jan | 1 | Chicago | 25.6 |
| Jan | 1 | San Diego | 55.2 |
| Jan | 1 | Houston | 53.9 |
| Jan | 1 | Death Valley | 51.0 |
| Jan | 2 | Chicago | 25.5 |
| Jan | 2 | San Diego | 55.3 |
| Jan | 2 | Houston | 53.8 |
| Jan | 2 | Death Valley | 51.2 |
| Jan | 3 | Chicago | 25.3 |

One option



Alternative representation

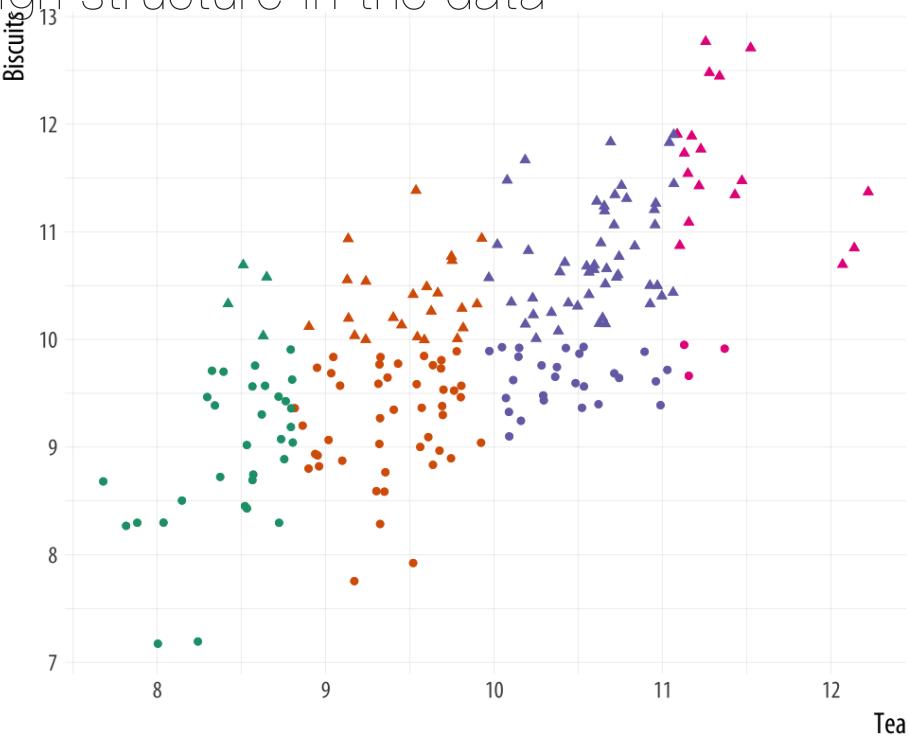
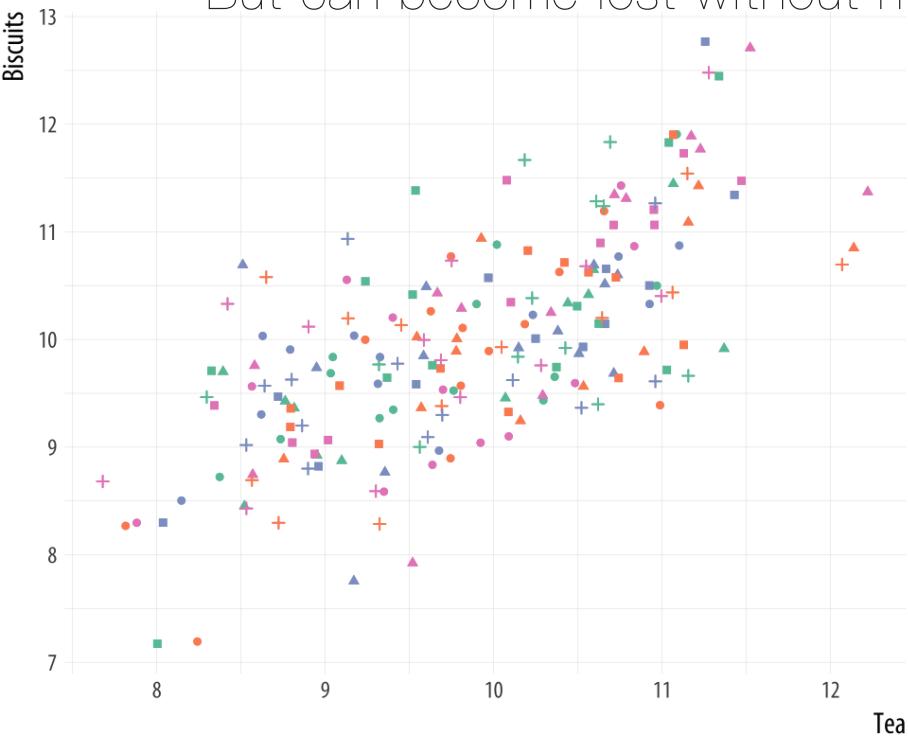


Comparison

- Both represent three scales
 - Two position scales (x/y axis)
 - One color scale (categorical for the first, continuous for the second)

More scales are possible

But can become lost without high structure in the data



Thinking
more about
color

Three fundamental uses

1. Distinguish groups from each other
2. Represent data values
3. Highlight

Discrete items

Often no intrinsic order

Qualitative color scale

- Finite number of colors
 - Chosen to maximize distinctness, while also being *equivalent*
 - Equivalent
 - No color should stand out
 - No impression of order

Examples

Okabe-Ito



Color Brewer: Dark2



Colorspace: Dynamic



See more about the Okabe Ito palette origins [here](#)

Sequential scale examples

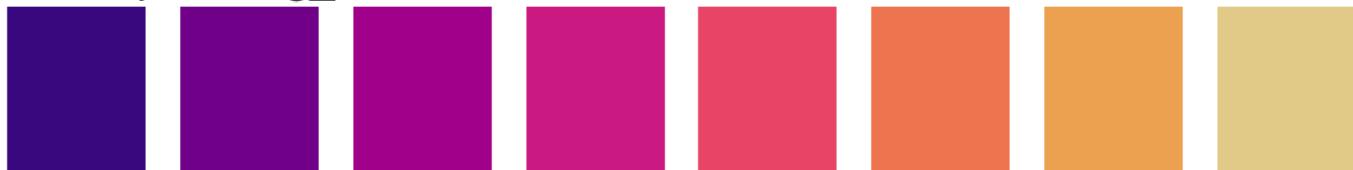
Cividis



Color Brewer: Purples



Colorspace: ag_Sunset



Diverging palettes

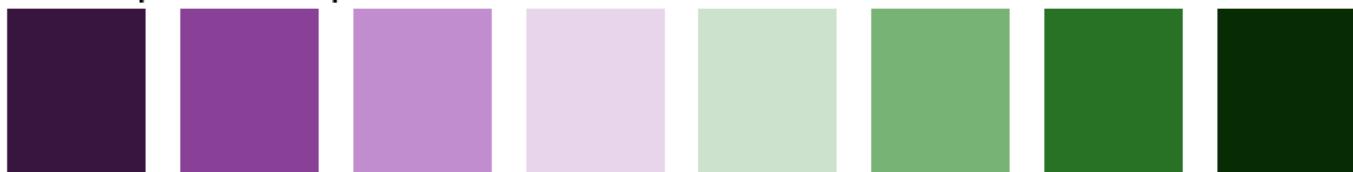
scico: vik



Color Brewer: PuOr

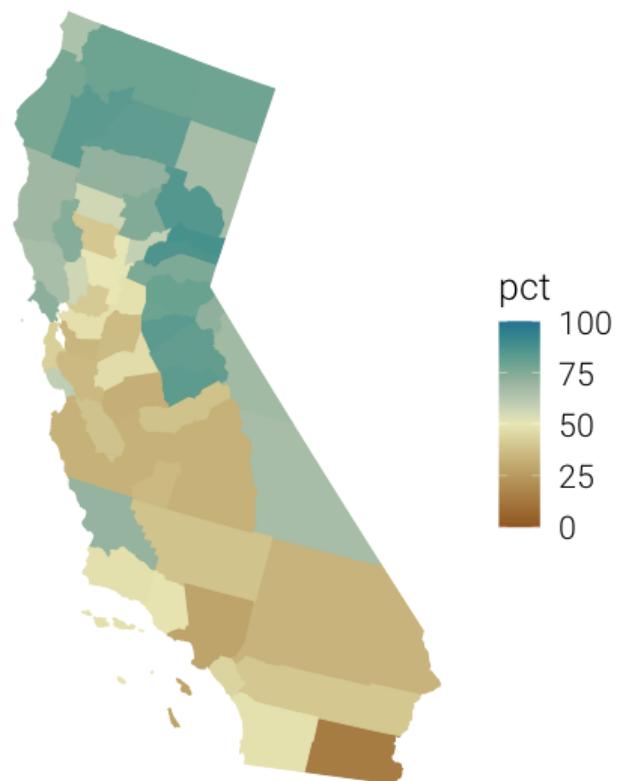


Colorspace: Purple-Green



Earth palette

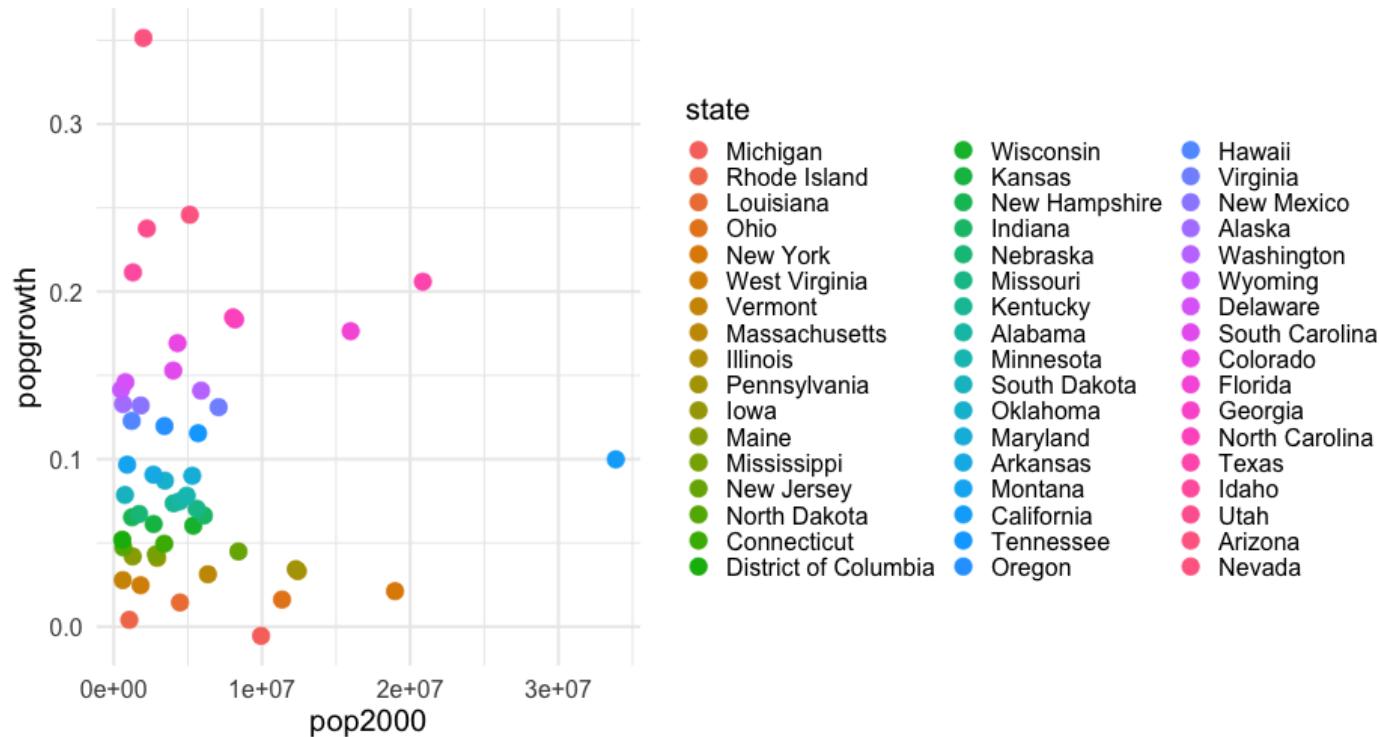
Applied example: Percentage of people coded "white" in California according to the 2010 decennial census.



Common problems

Too many colors

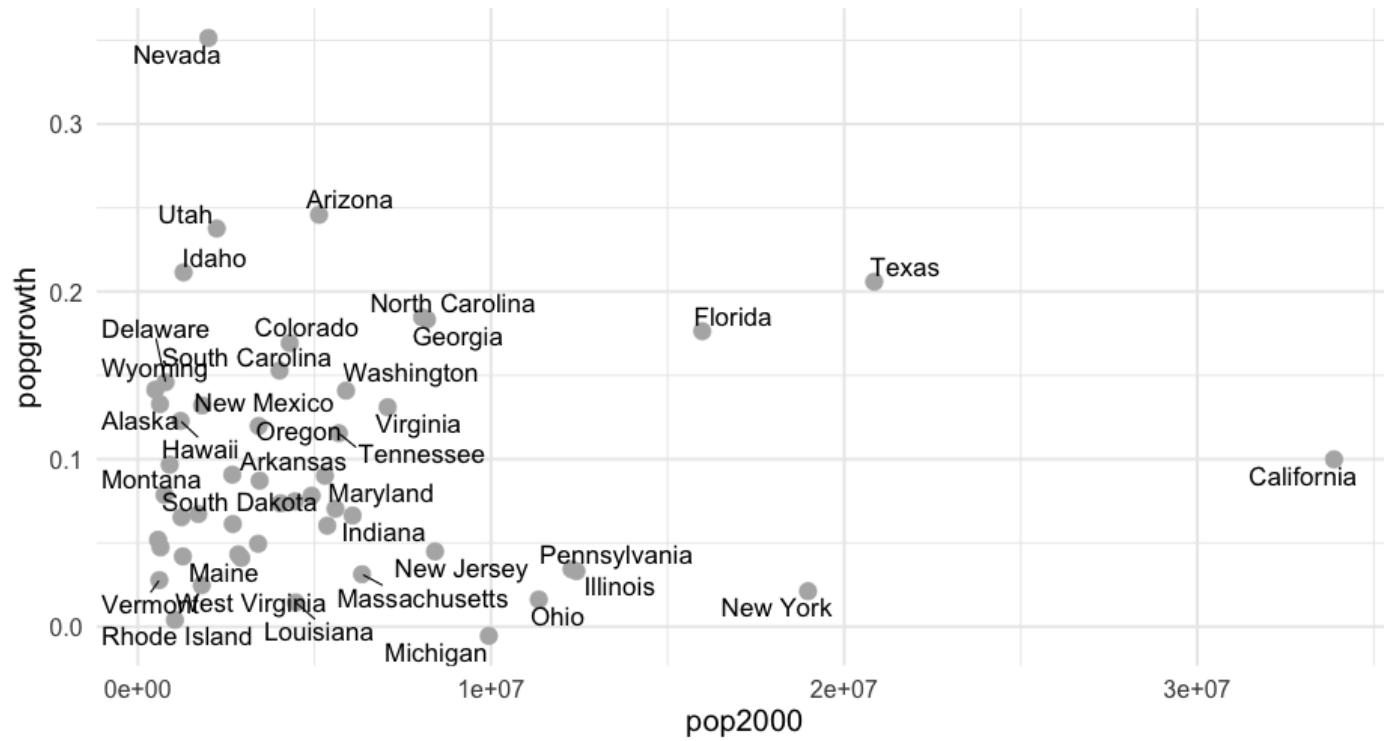
More than 5-ish generally becomes difficult to track



Data from Wilke

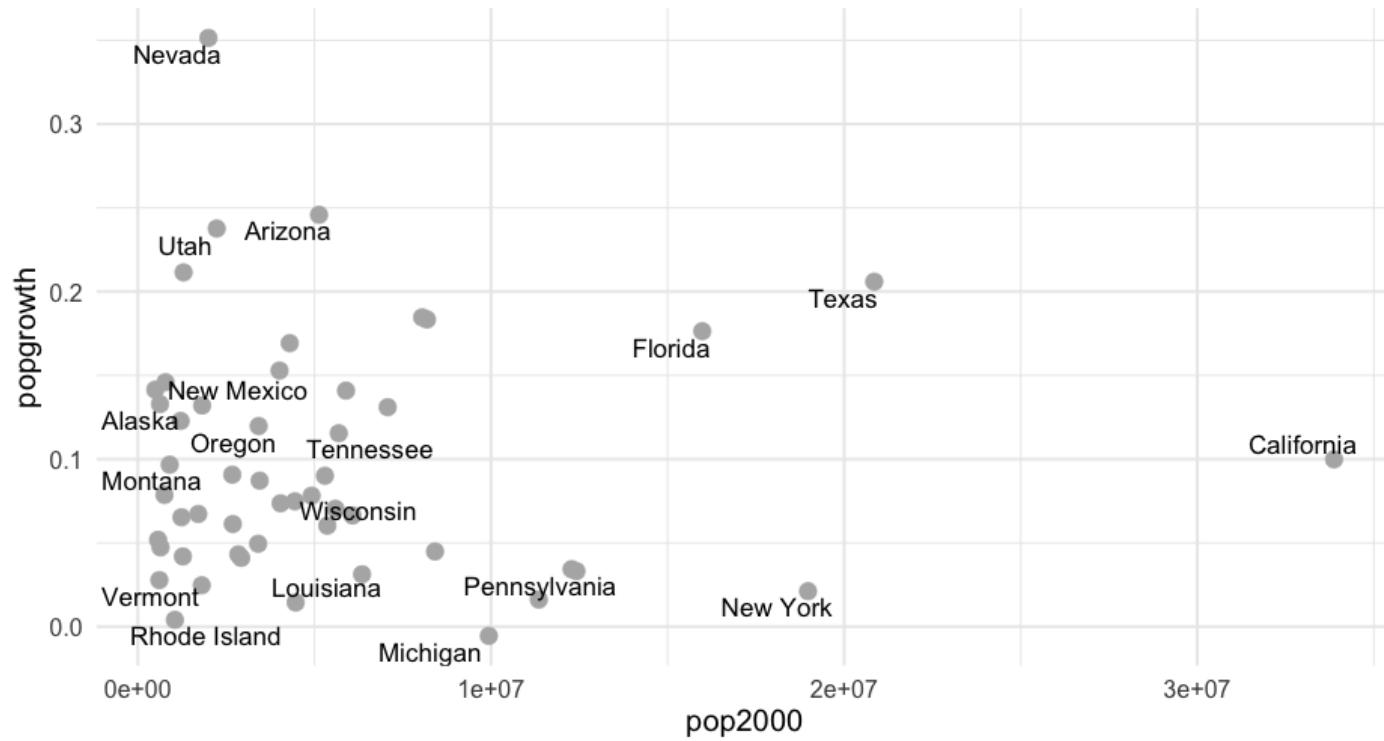
Use labels

still too many...



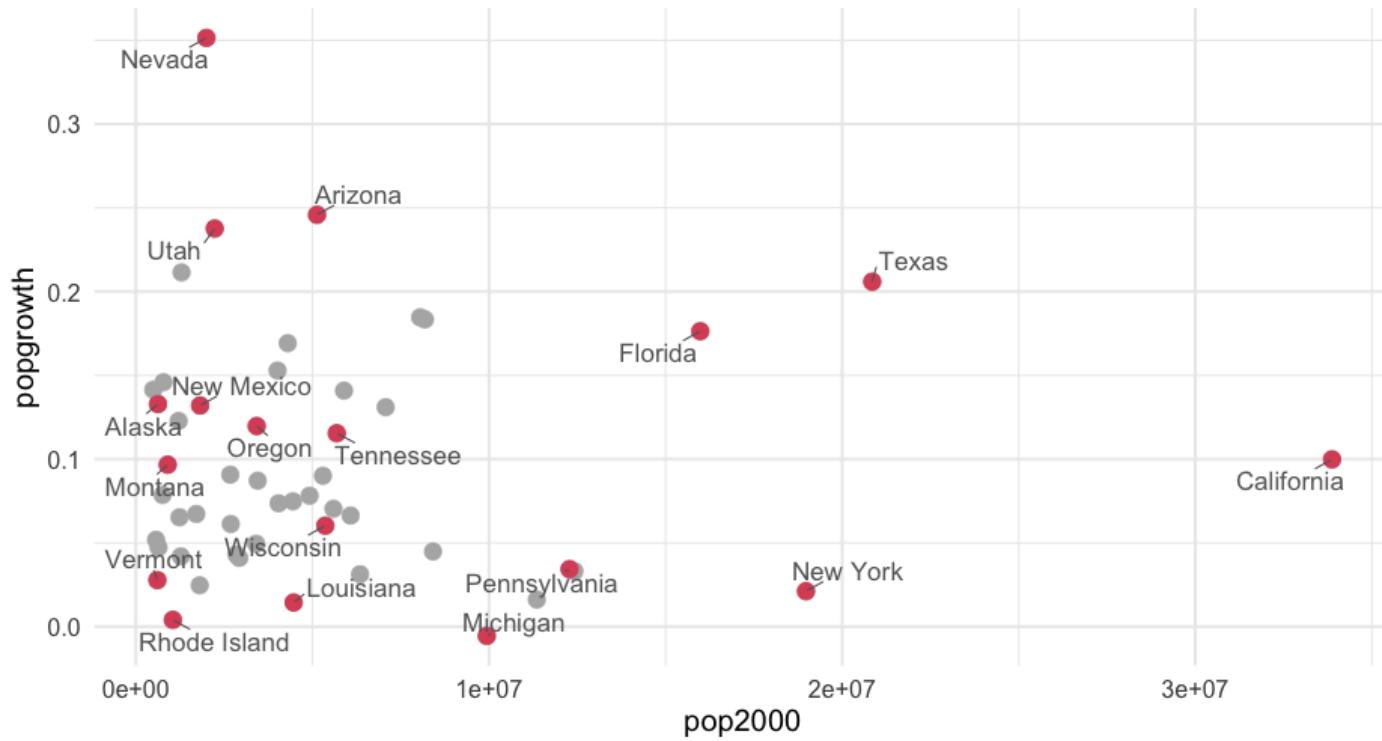
Better

Get a subset



Best

(but could still be improved)



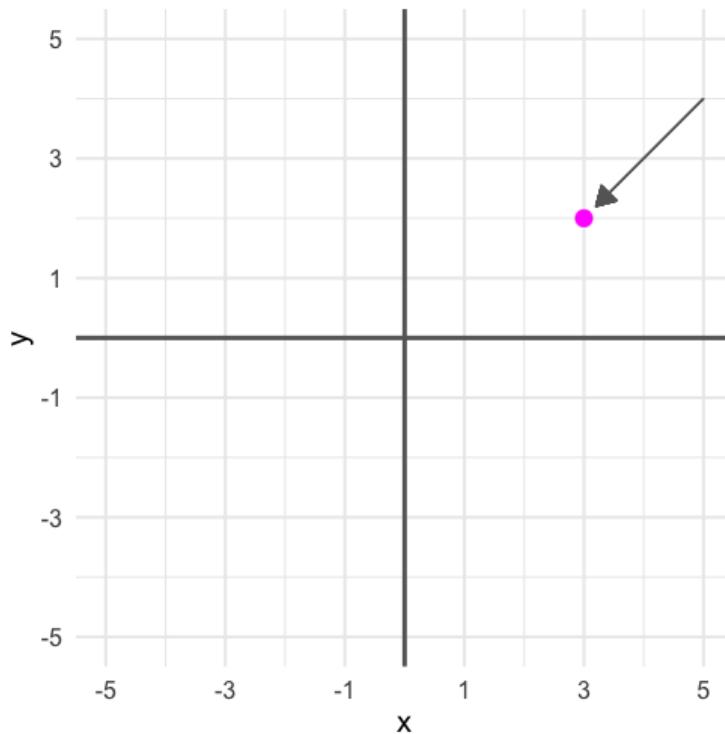
Last few note on palettes

- Do some research, find what you like **and** what tends to work well
- Check for colorblindness (many options, but consider [here](#))
- Look into <http://colorbrewer2.org/>

Visualizing Uncertainty

The primary problem

- When we see a point on a plot, we interpret it as **THE** value.

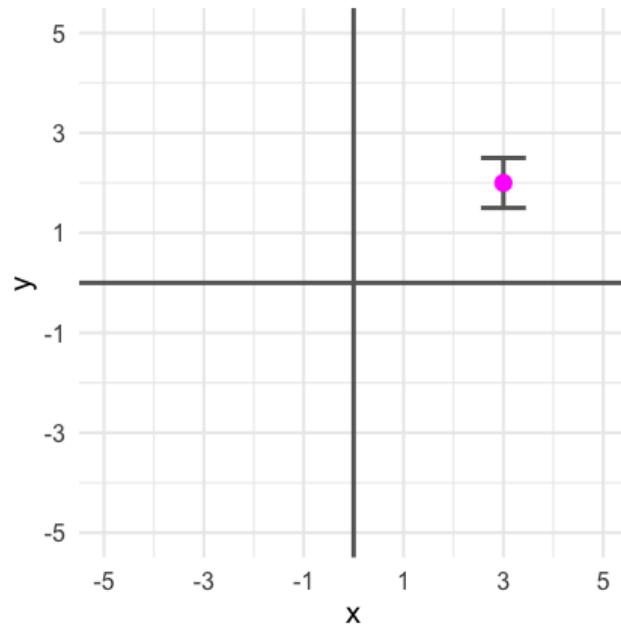


Some secondary problem

- We're not great at understanding probabilities
- We regularly round probabilities to 100% or 0%
- As probabilities move to the tails, we're generally worse

How do we typically communicate uncertainty?

- Error bars



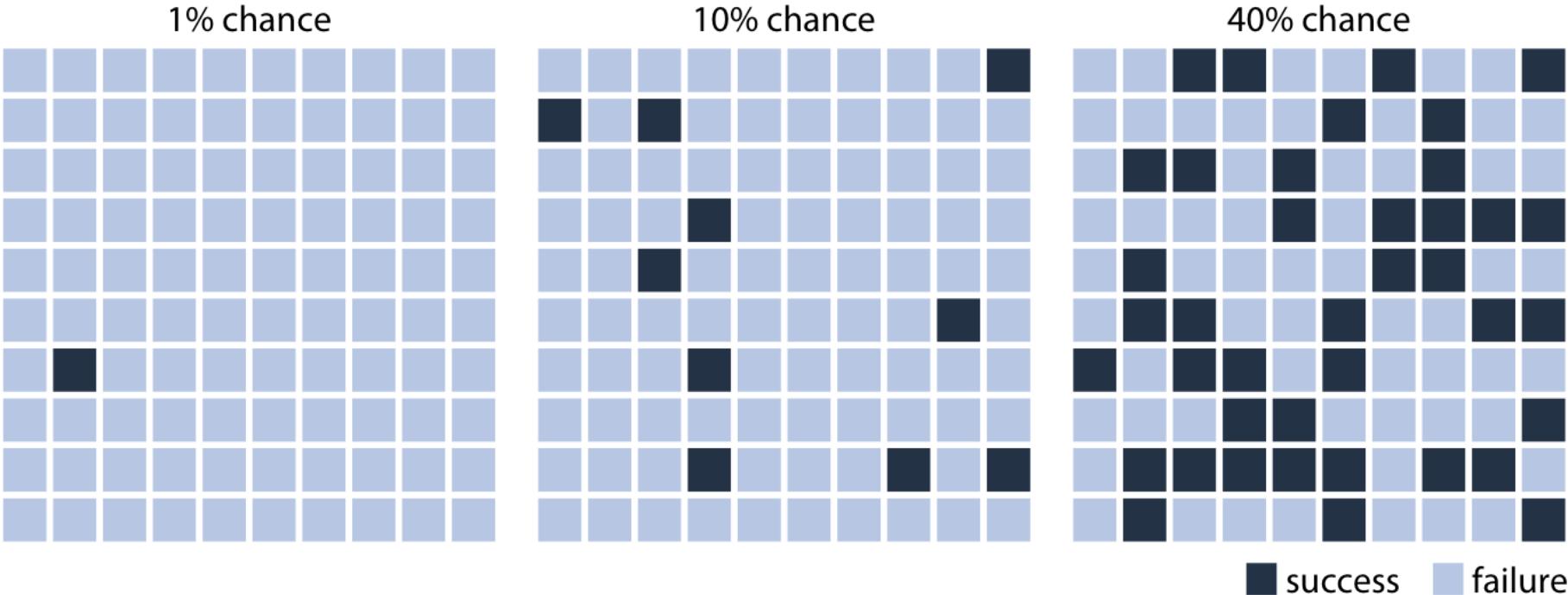
Thinking about uncertainty

Uncertainty means exactly what it sounds like – we are not 100% sure.

- We are nearly always uncertain of future events (forecasting)
- We can also be uncertain about past events
 - I saw a parked car at 8 AM, but the next time I looked at 2PM it was gone. What time did it leave?

Quantifying uncertainty

- We quantify our uncertainty mathematically using probability
- Framing probabilities as frequencies is generally more intuitive



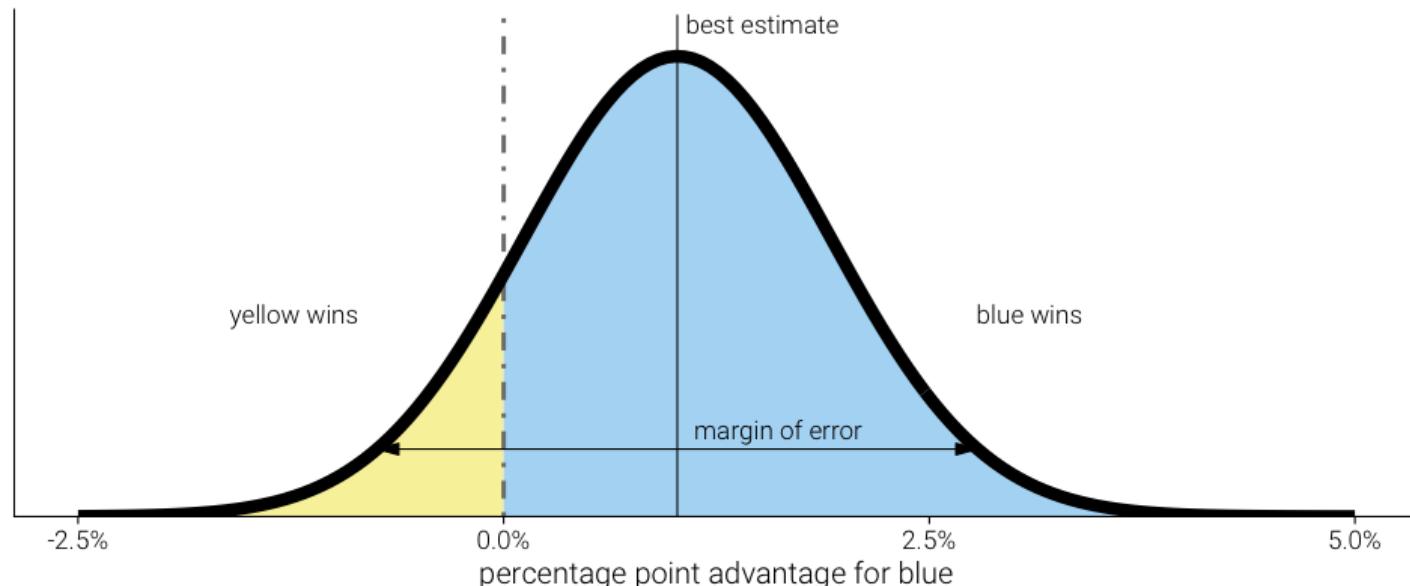
Framing a single uncertainty

Example from Wilke

Non-discrete probabilities

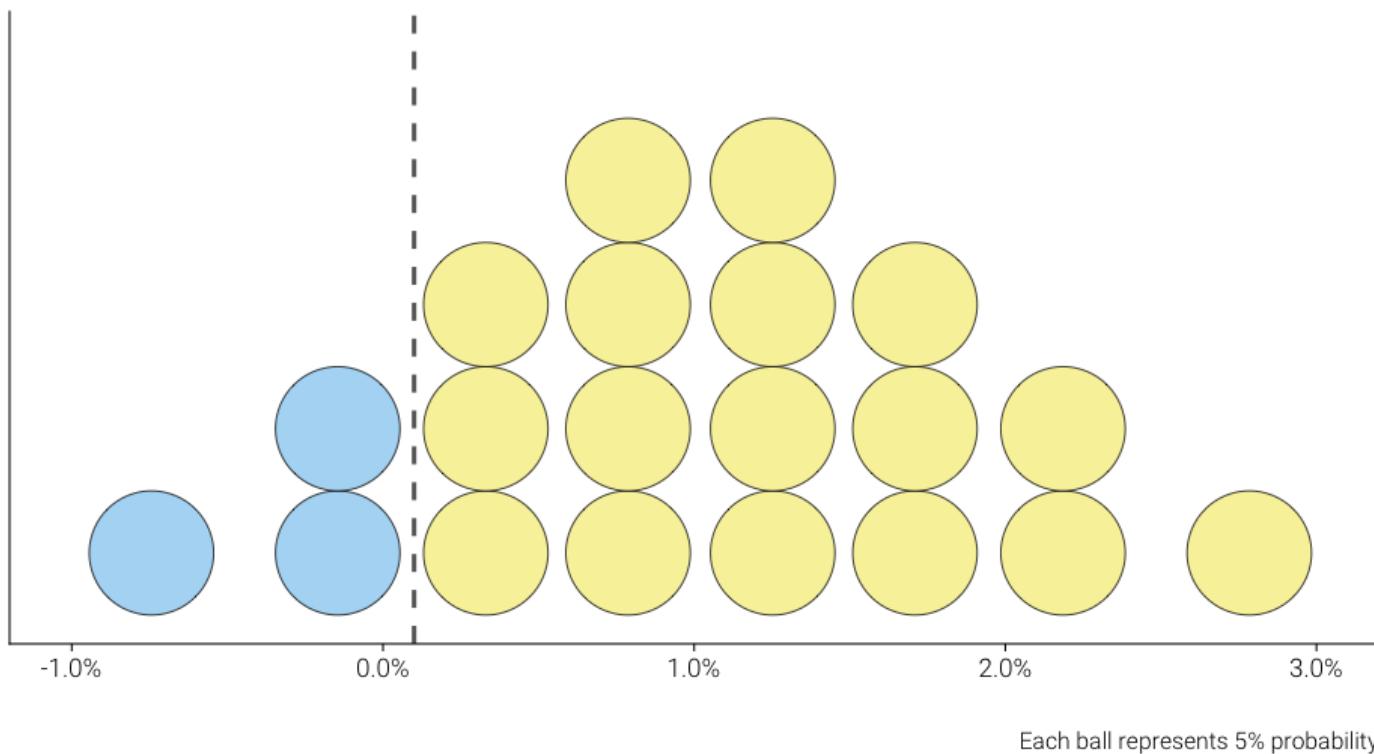
Blue party has 1% advantage w/ margin of error of 1.76 points

Who will win?



Code taken from [Wilke](#)

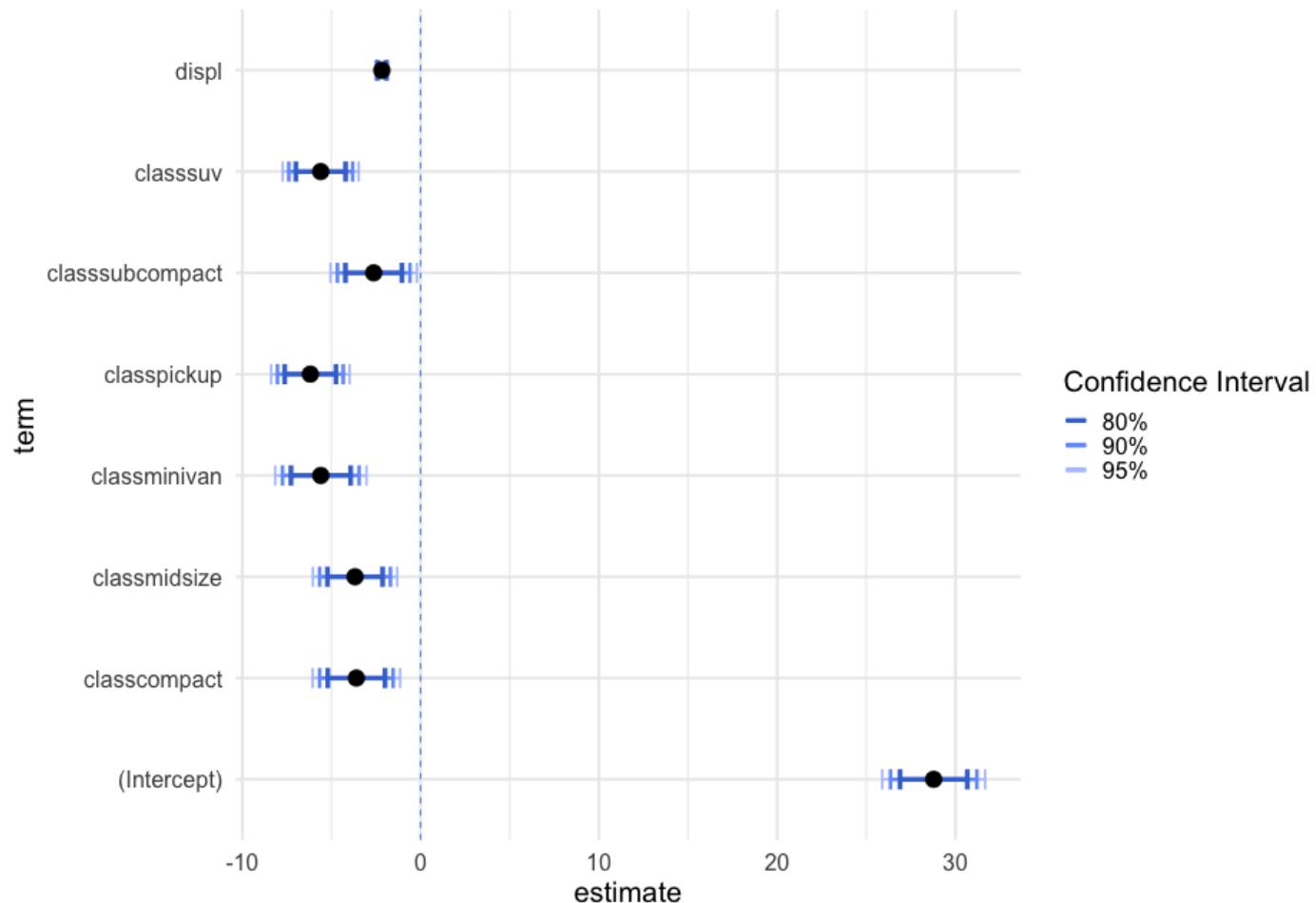
Descretized version



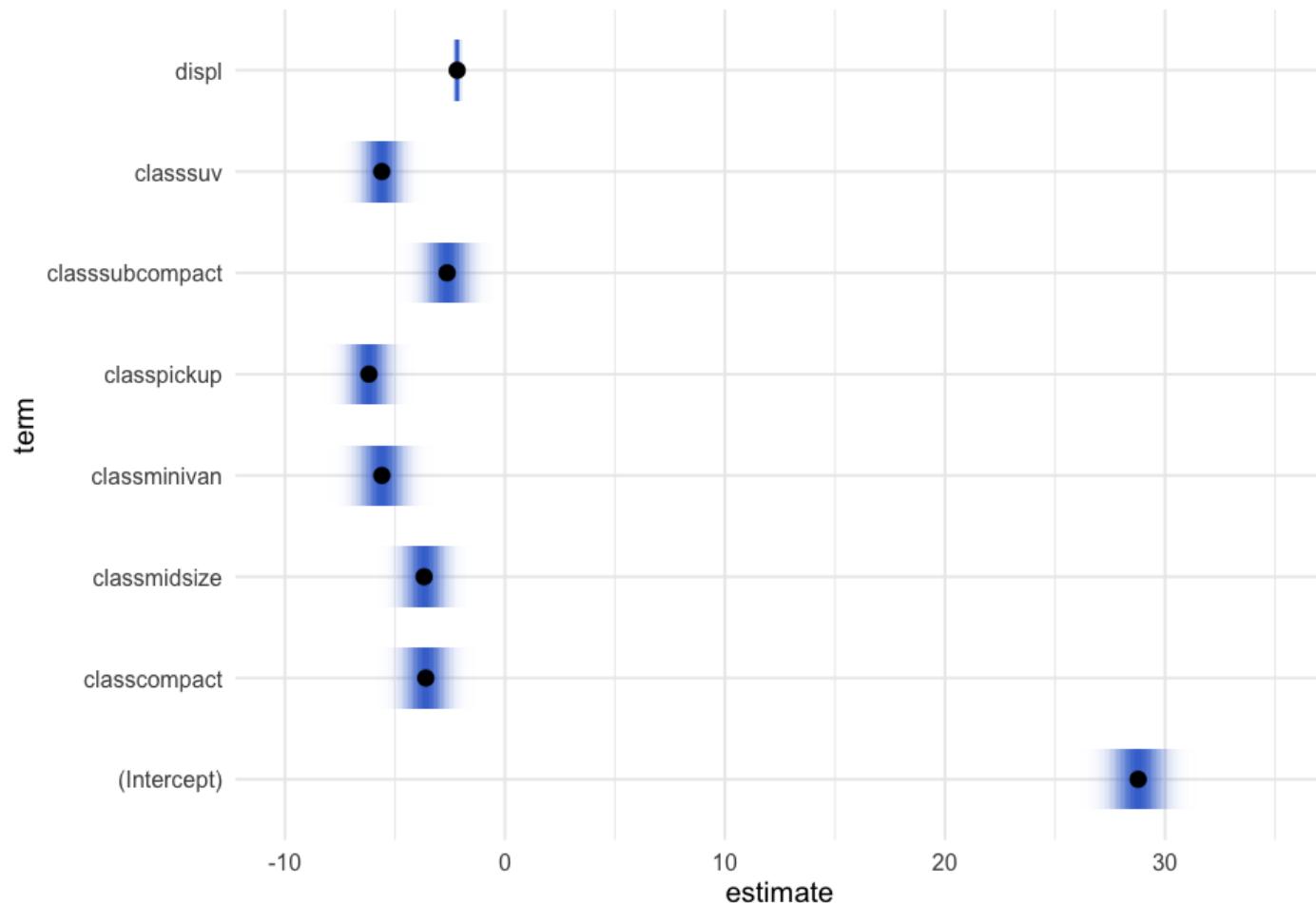
Point estimates

A few alternatives to error bars

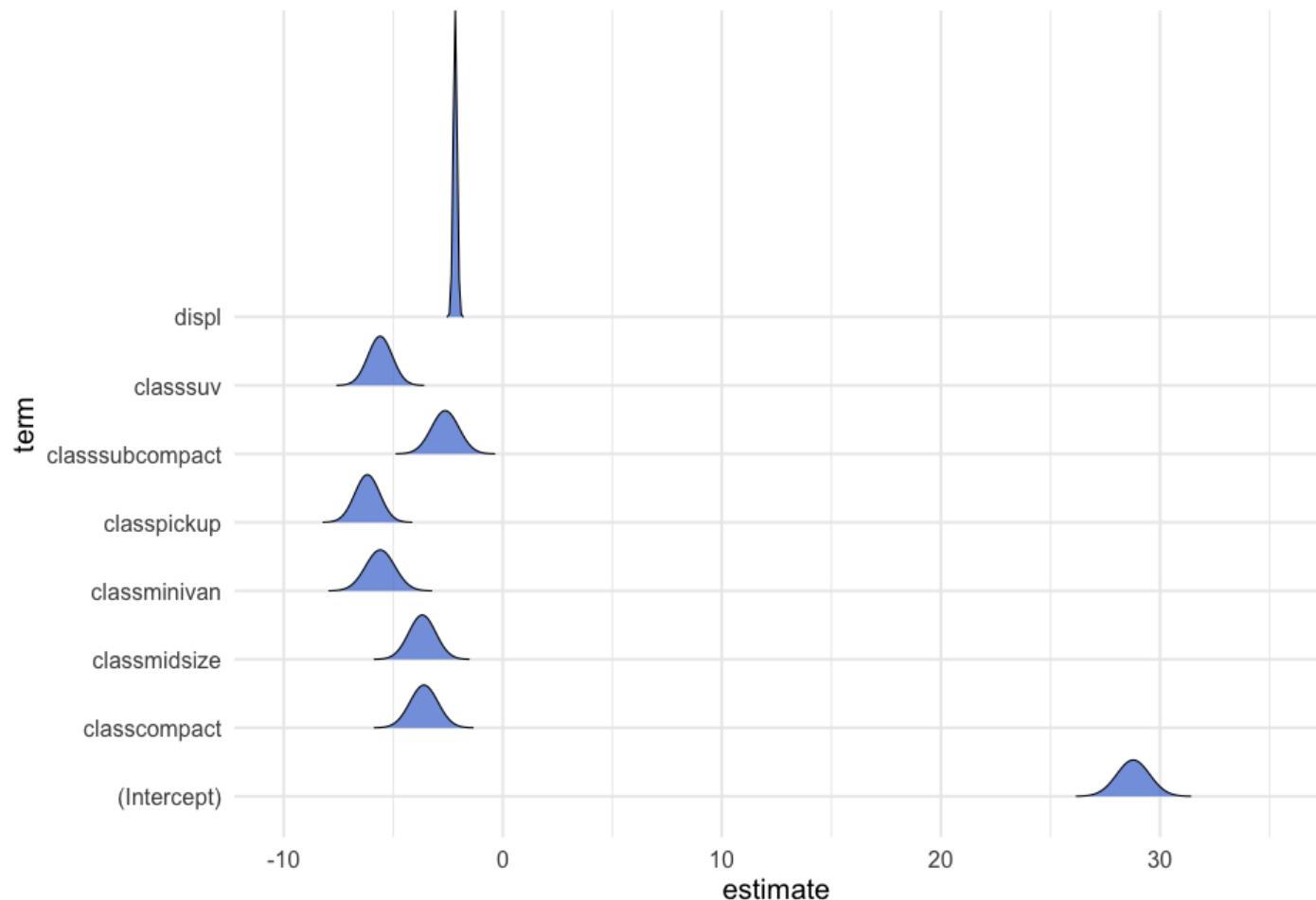
Multiple error bars



Density stripes



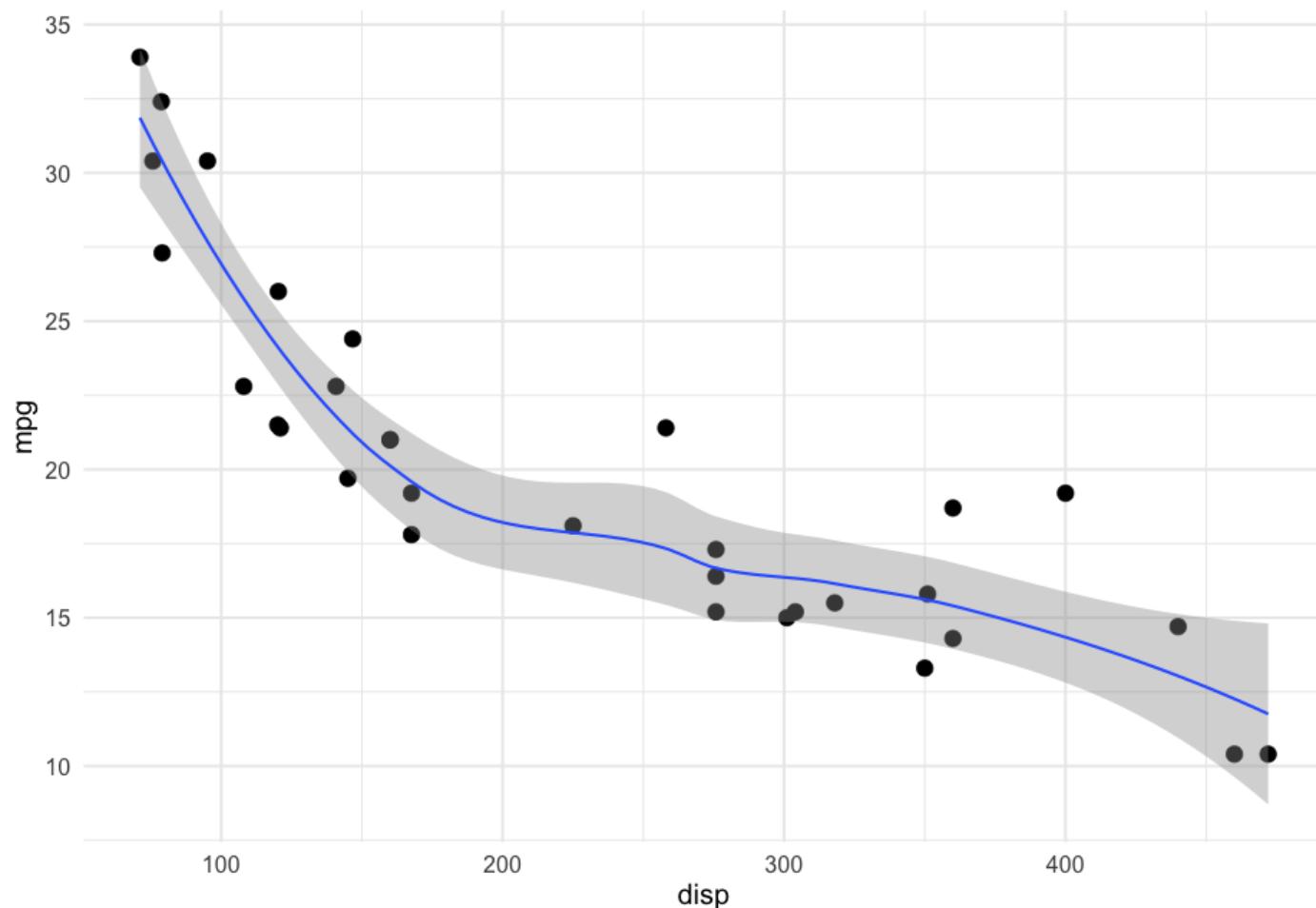
Actual densities



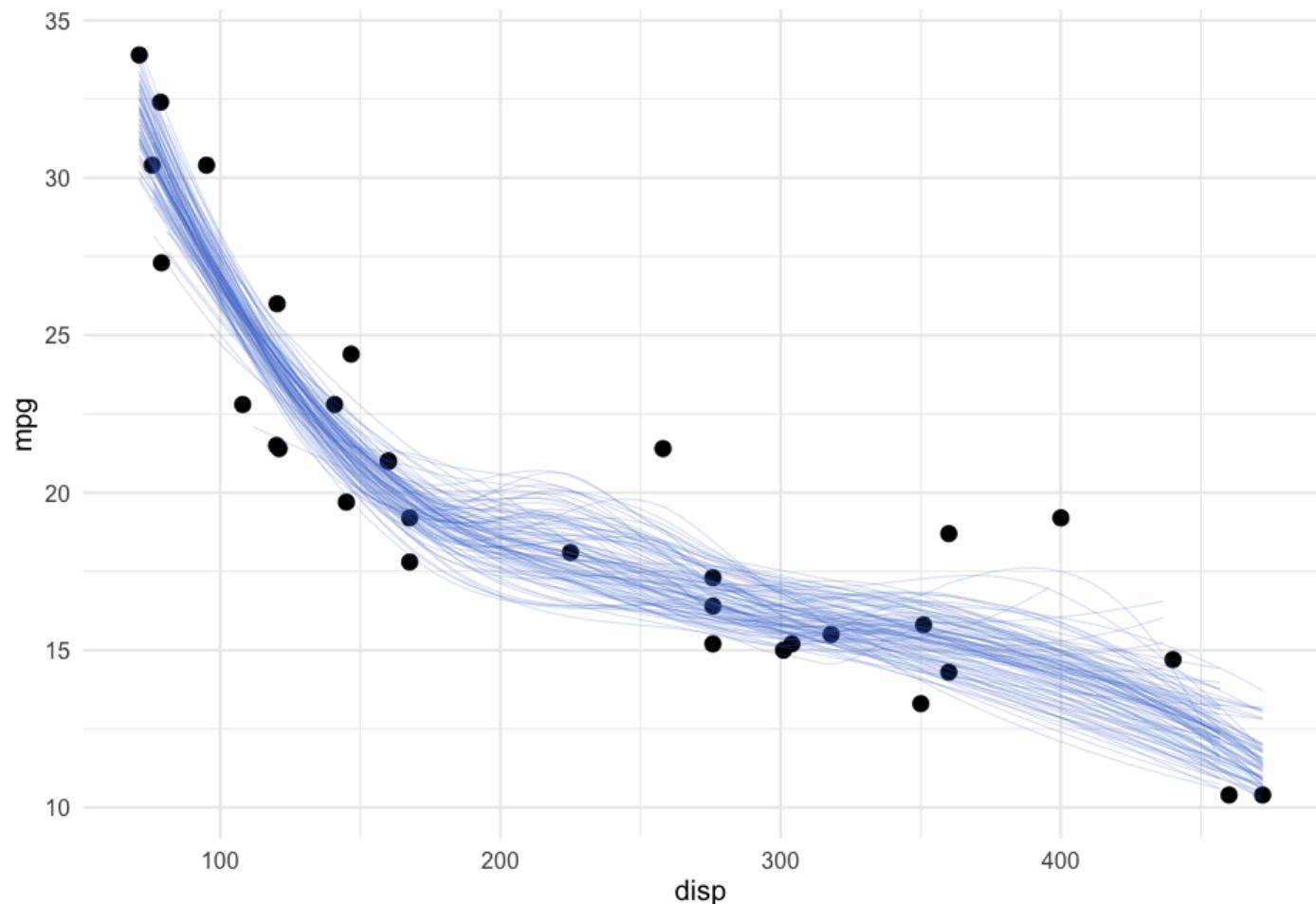
HOPs

Hypothetical Outcome Plots (and related plots)

Standard regression plot

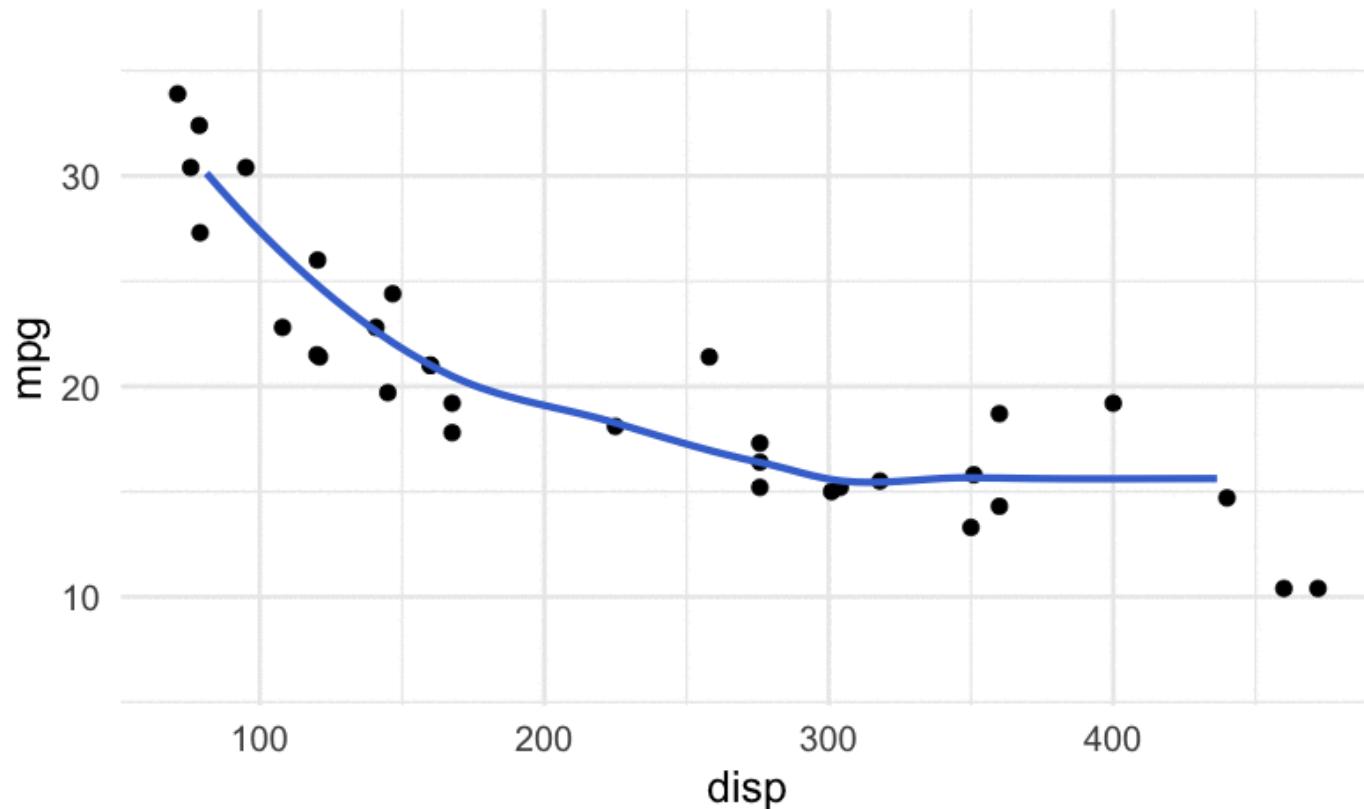


Alternative



HOPs

Hops animate the process, so you can't settle on one "truth"



Last few notes on uncertainty

- Do try to communicate uncertainty whenever possible
- I'd highly recommend watching [this](#) talk by Matthew Kay on the topic

Data ink ratio

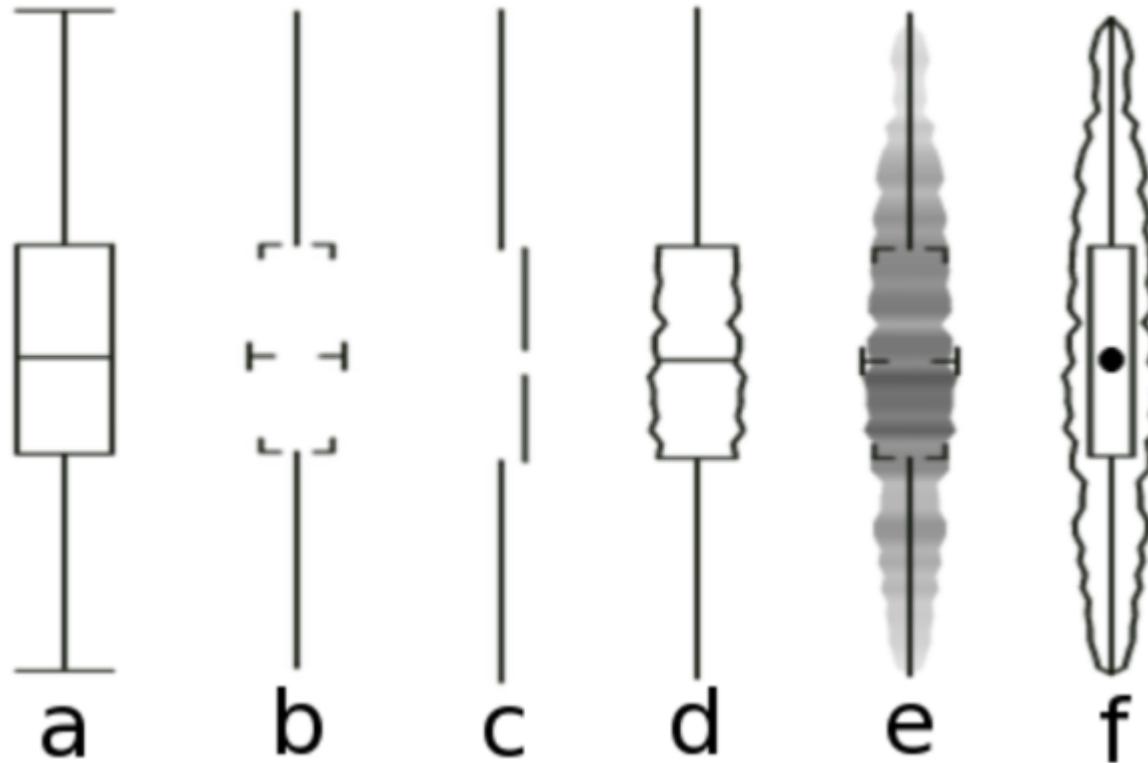
What is it?

Above all else, show the data

—Edward Tufte

- Data-Ink Ratio = Ink devoted to the data / total ink used to produce the figure
- Common goal: Maximize the data-ink ratio

Example



- First thought might be... Cool!

A color photograph of an older man with light brown hair, wearing a dark suit jacket, a white shirt, and a red patterned tie. He has a wide-open mouth and appears to be shouting or speaking very forcefully. A large, white, hand-drawn-style speech bubble is positioned above his head, containing the text "NOT SO FAST, MY FRIEND!" in black, all-caps, sans-serif font.

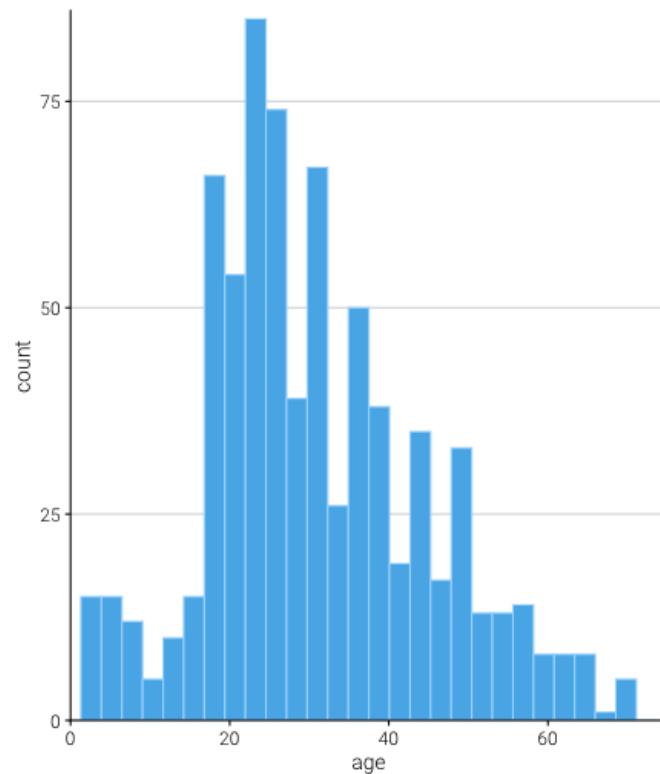
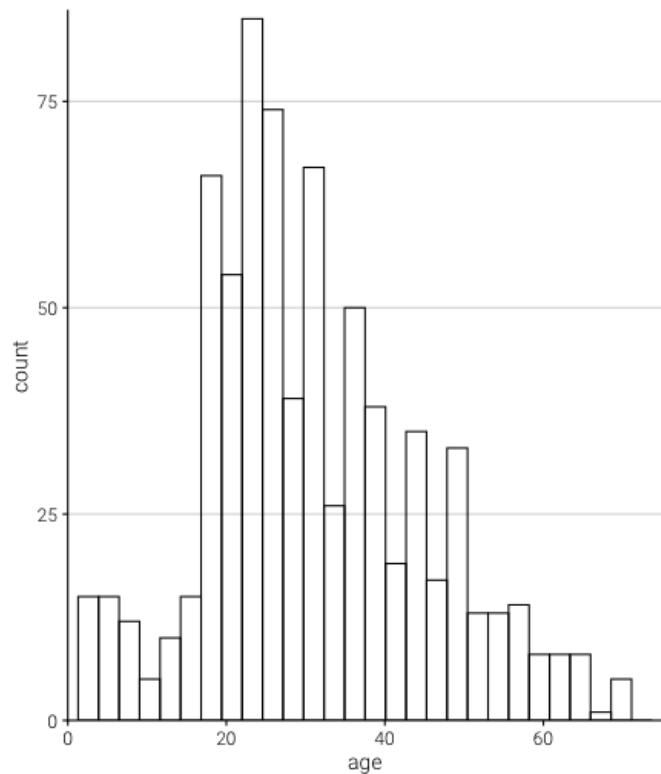
NOT SO
FAST, MY
FRIEND!

Minimize cognitive load

- Empirically, Tufte's plot was the most difficult for viewers to interpret.
- Visual cues (labels, gridlines) reduce the data–ink ratio, but can also reduce cognitive load.

An example

Which do you prefer?



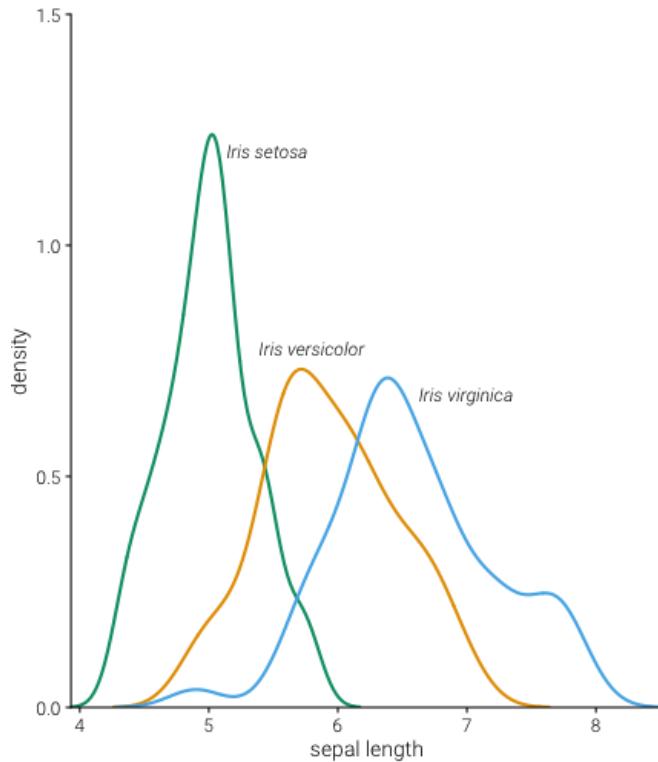
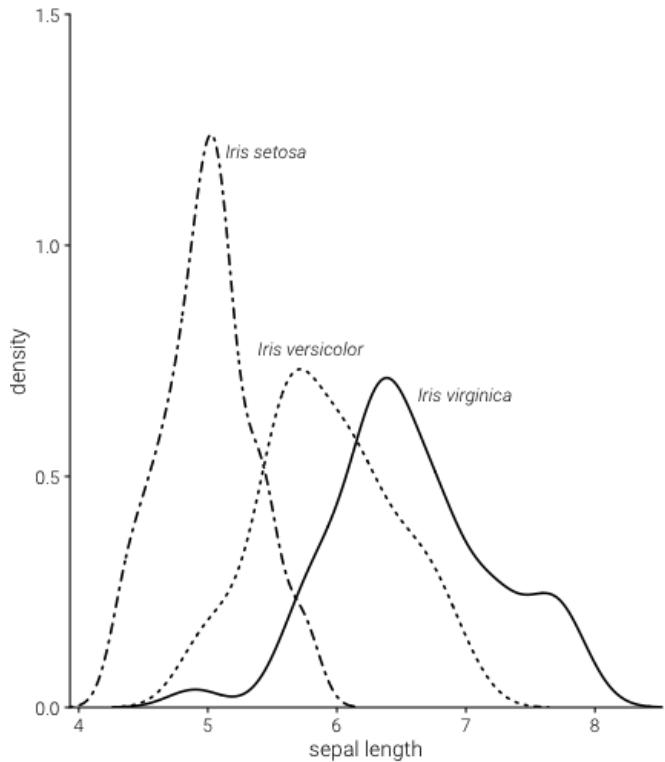
Code/example taken from [Wilke](#)

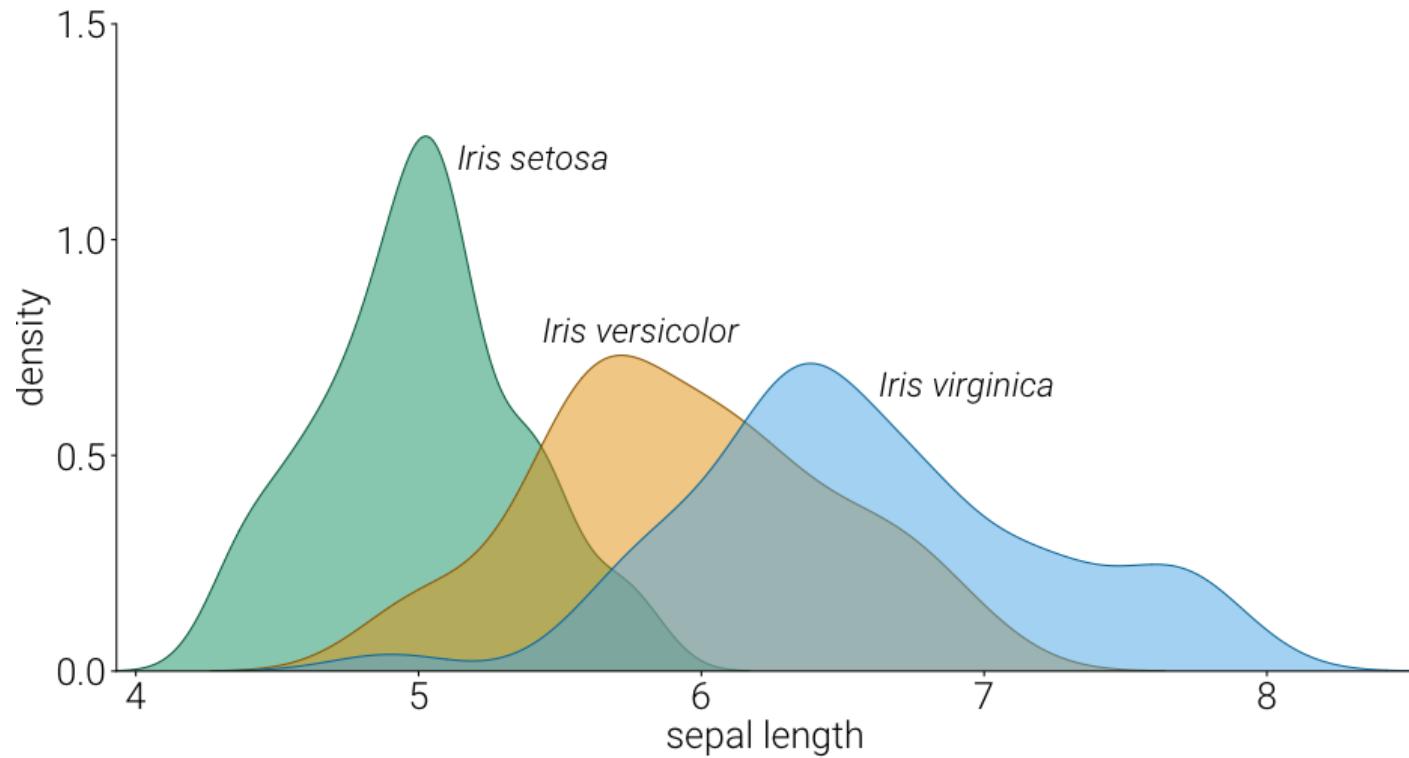
Advice from Wilke

Whenever possible, **visualize your data with solid, colored shapes** rather than with lines that outline those shapes. Solid shapes are more easily perceived, are less likely to create visual artifacts or optical illusions, and do more immediately convey amounts than do outlines.

emphasis added

Another example





Labels in place of legends

Prior slide is a great example of when annotations can be used in place of a legend to

- reduce cognitive load
- increase clarity
- increase beauty
- maximize the figure size

Practical
advice so far

Include uncertainty wherever possible

Avoid line drawings

Maximize the data–ink ratio within reason (but preference reduction of cognitive load)

Use color to your advantage (and think critically about the palettes you choose)

Consider plot annotations over legends

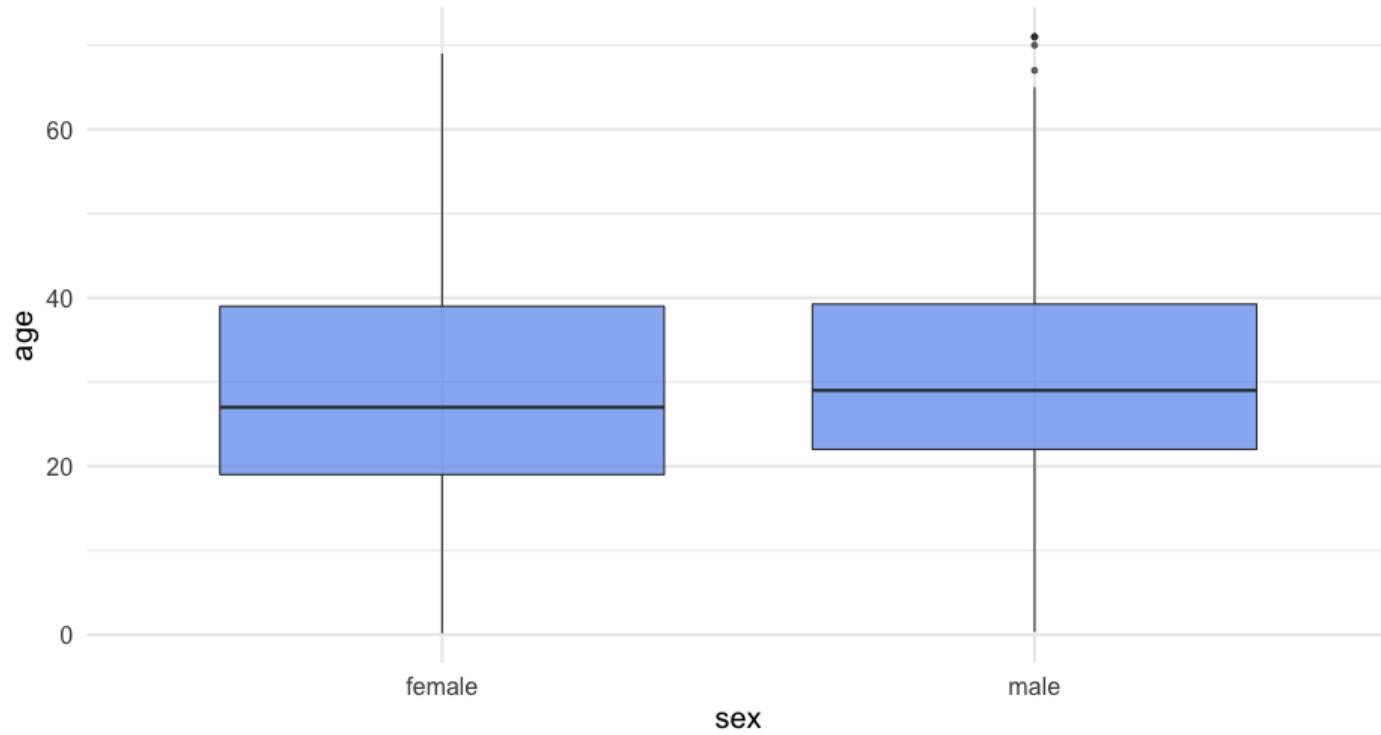
Dealing with grouped data

Quick empirical example

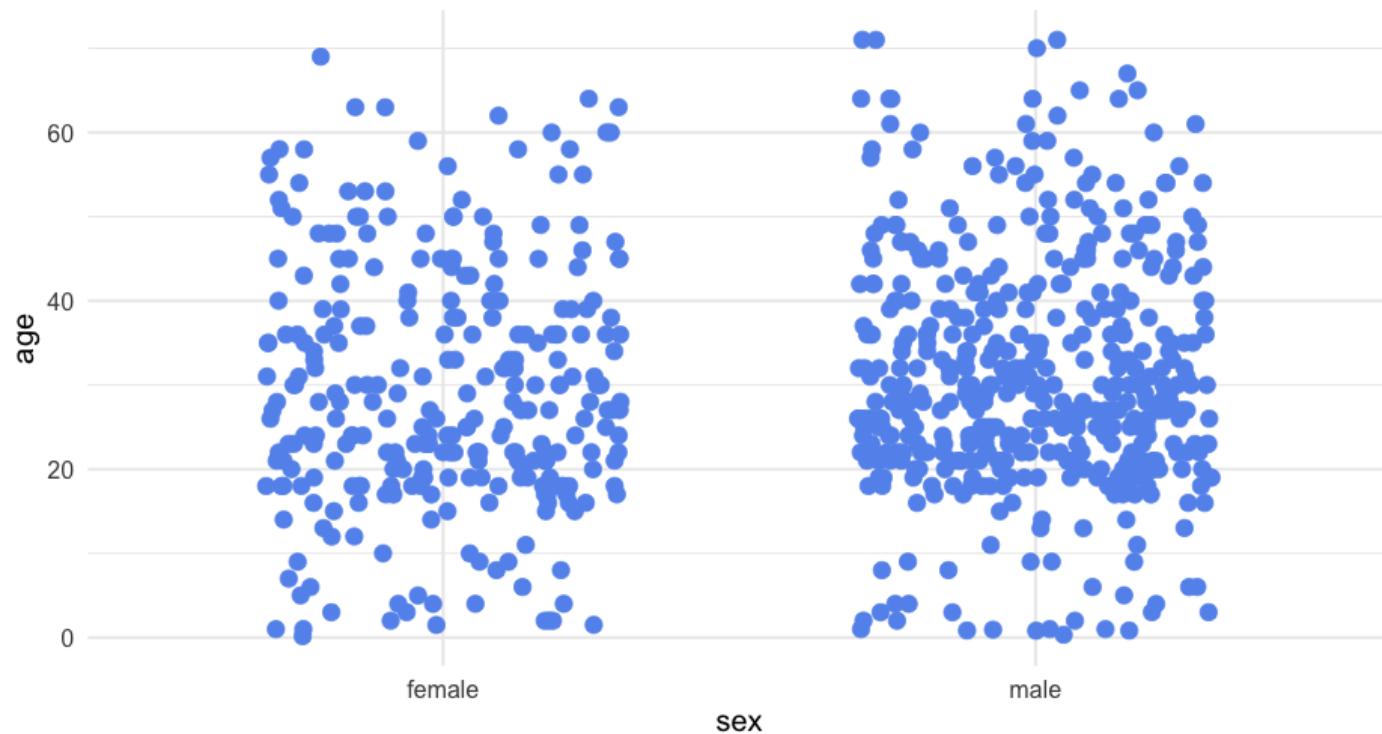
Titanic data

```
## # A tibble: 1,313 x 5
##   name                               class     age sex   survi
##   <chr>                            <chr> <dbl> <chr>  <i
## 1 Allen, Miss Elisabeth Walton    1st      29   female
## 2 Allison, Miss Helen Loraine    1st      2    female
## 3 Allison, Mr Hudson Joshua Creighton 1st      30   male
## 4 Allison, Mrs Hudson JC (Bessie Waldo Daniels) 1st      25   female
## 5 Allison, Master Hudson Trevor   1st      0.92 male
## 6 Anderson, Mr Harry              1st      47   male
## # ... with 1,307 more rows
```

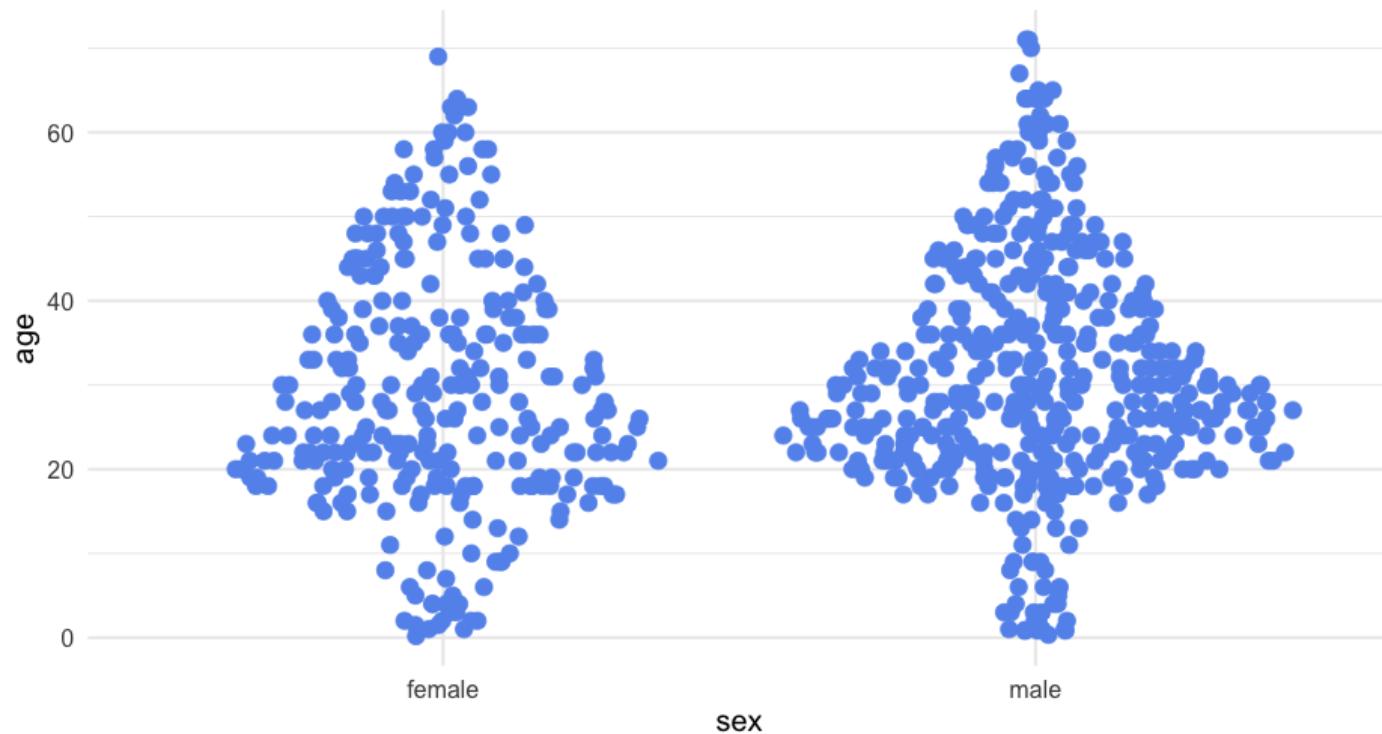
Boxplots



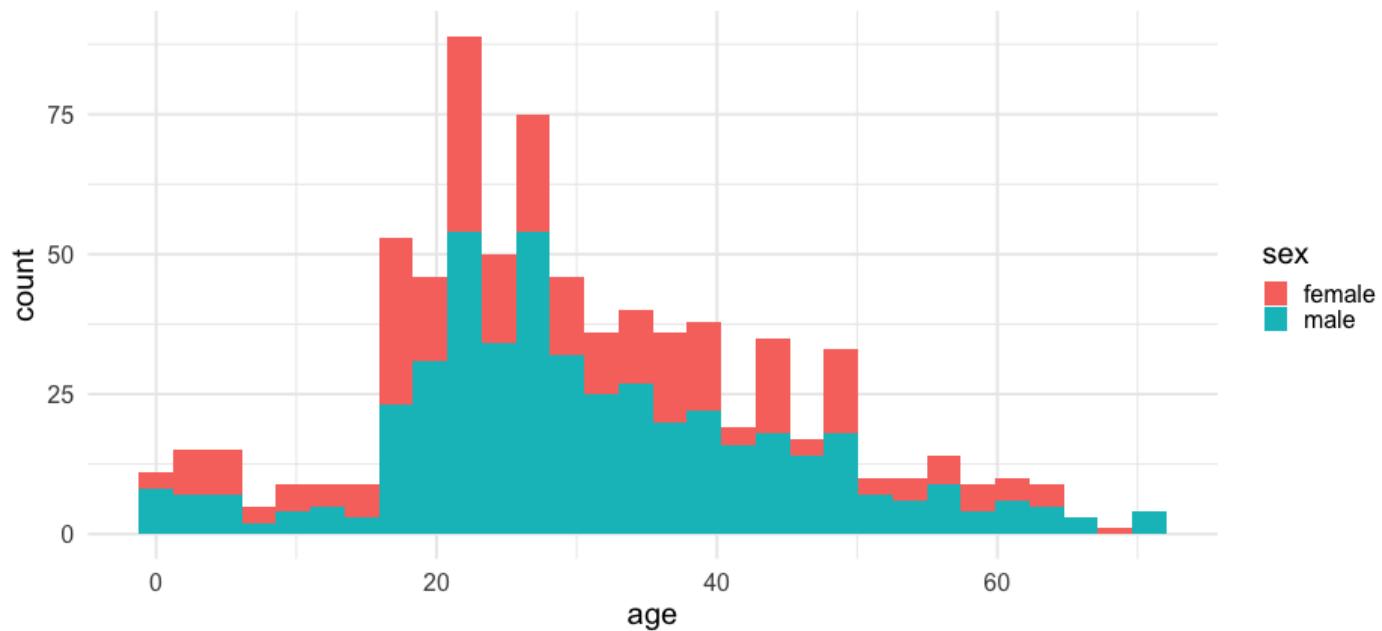
Jittered point plots



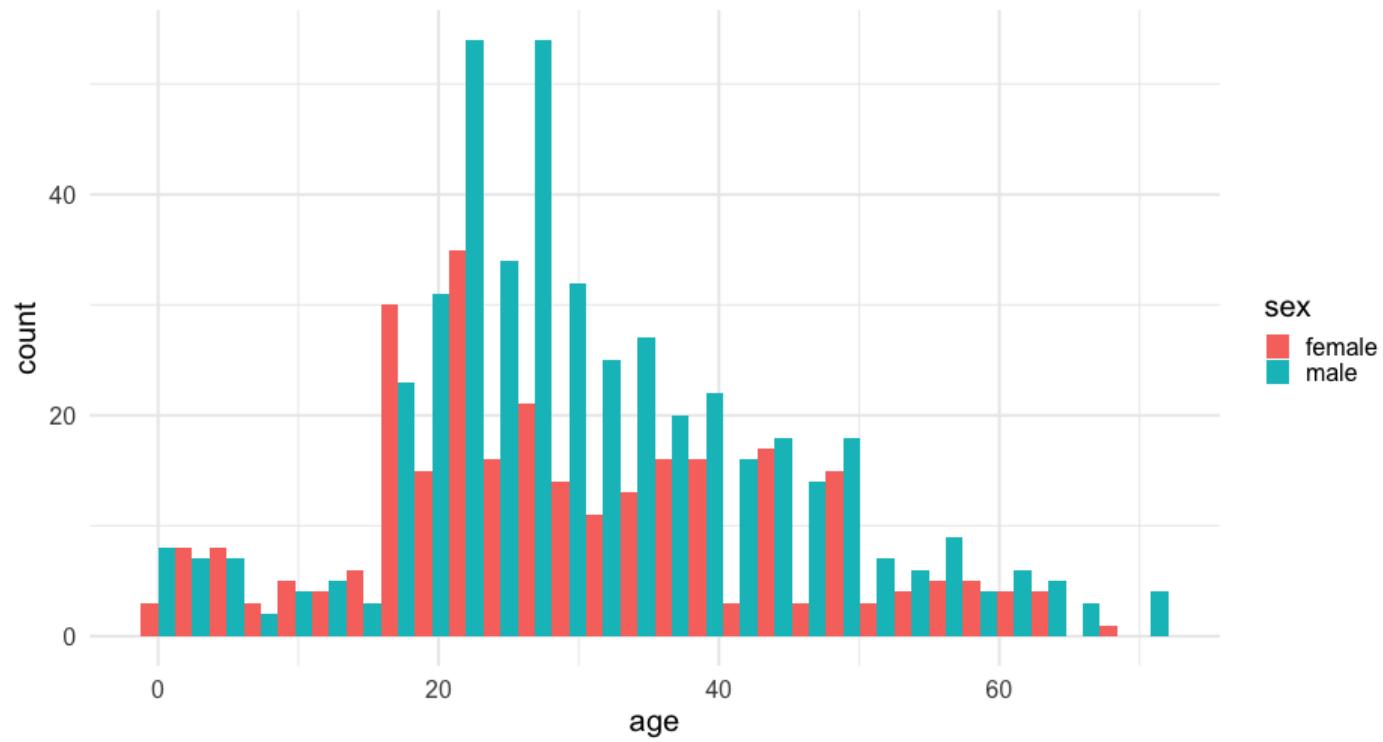
Sina plot



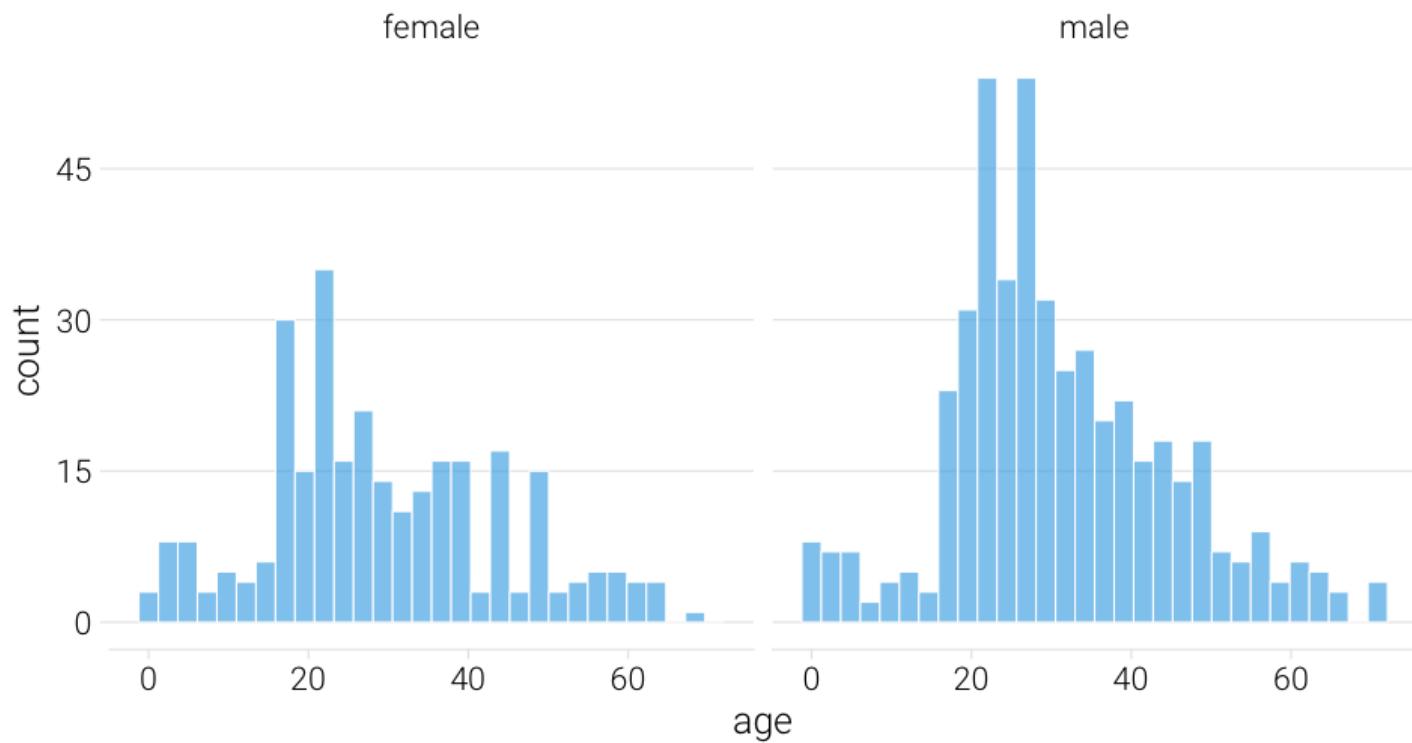
Stacked histogram



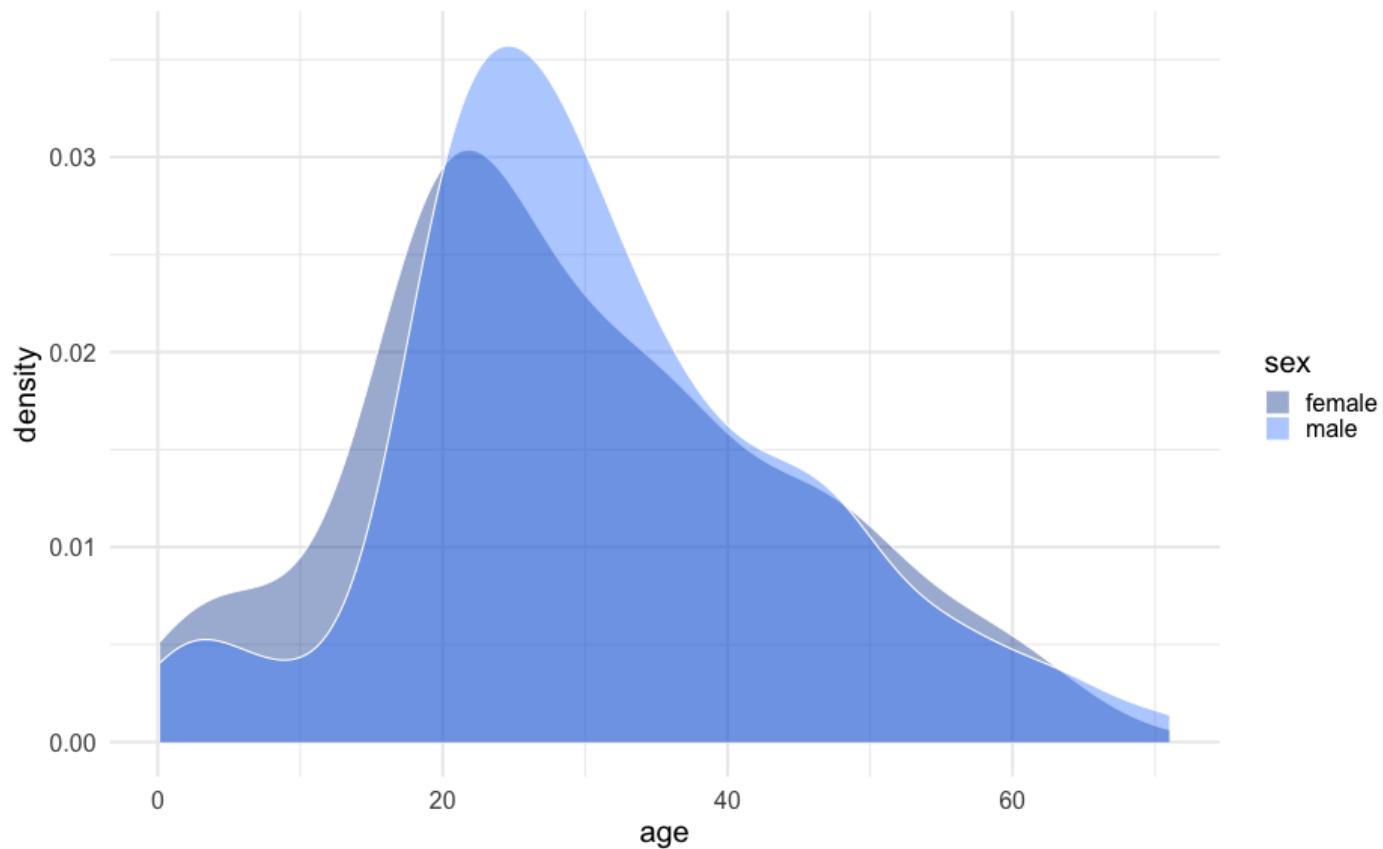
Dodged



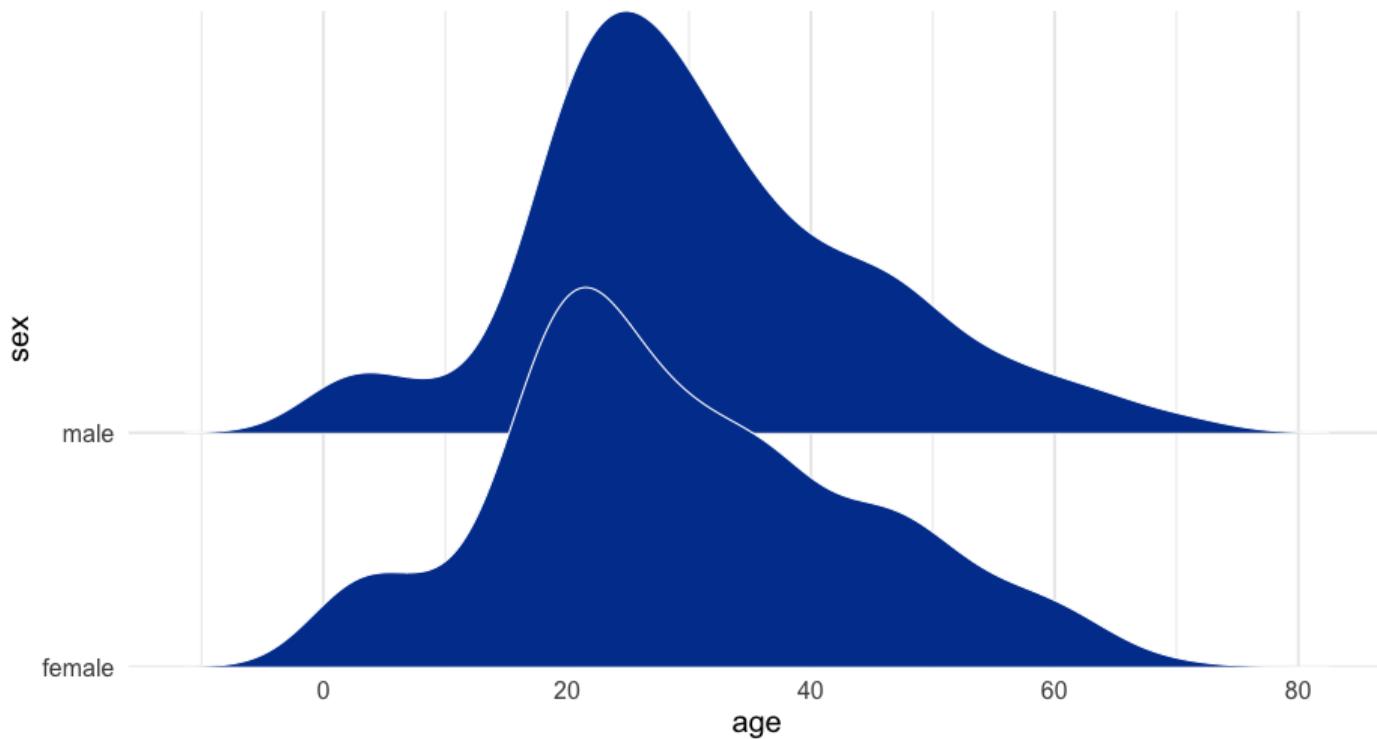
Better



Overlapping densities



Ridgeline densities



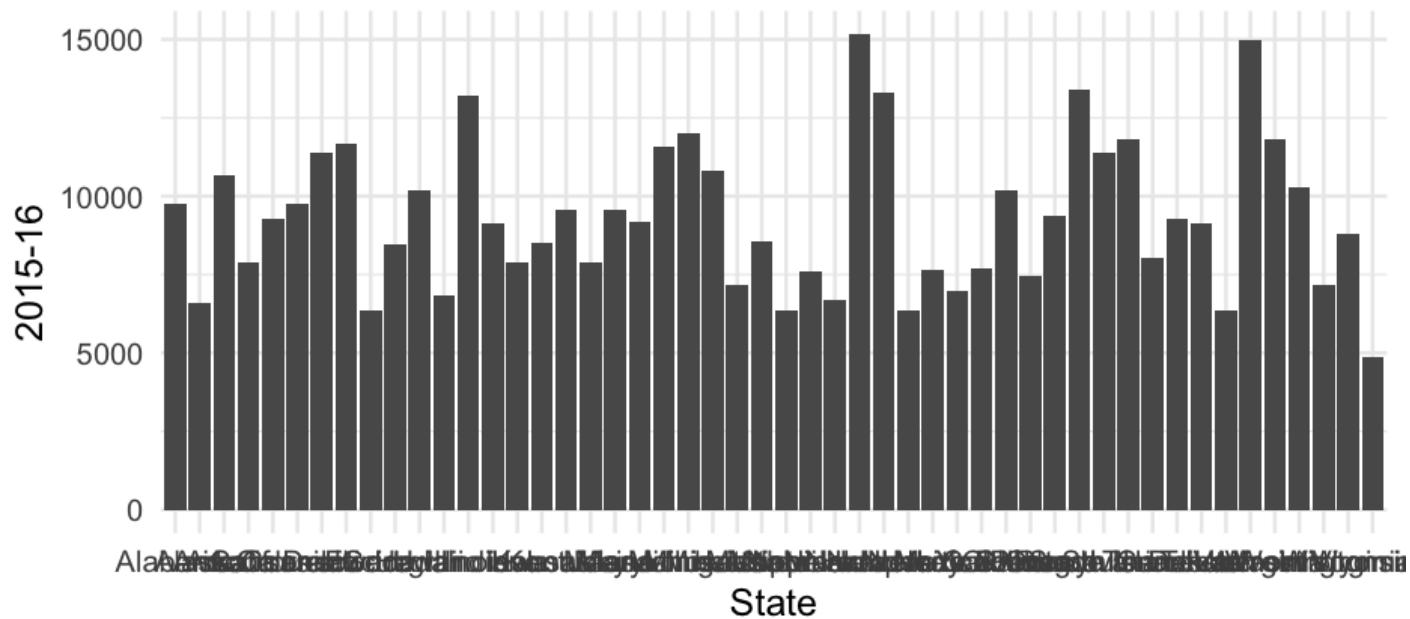
Visualizing amounts

How much does college cost?

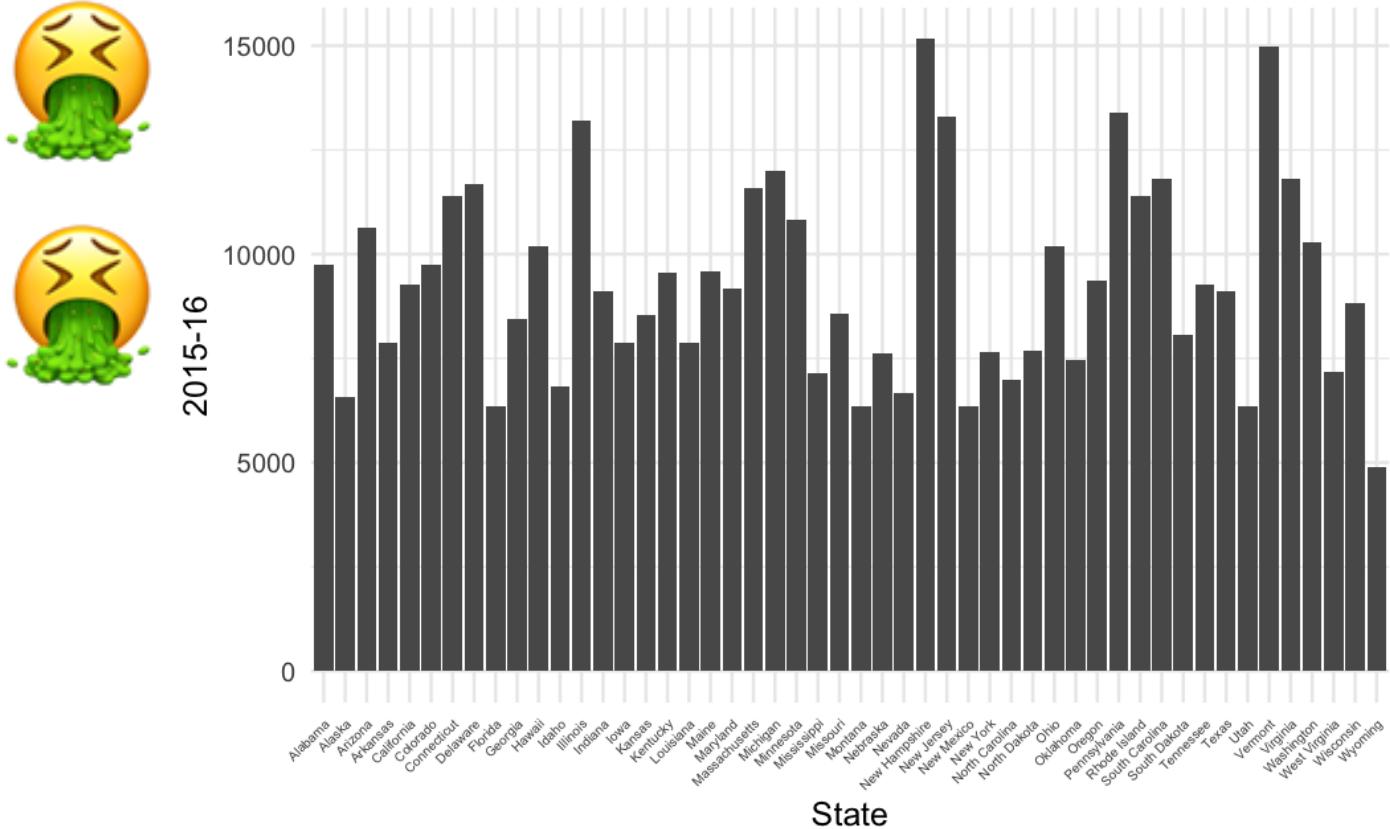
Tuition data

```
## # A tibble: 6 x 13
##   State    `2004-05` `2005-06` `2006-07` `2007-08` `2008-09` `2009-10` ...
##   <chr>     <dbl>     <dbl>     <dbl>     <dbl>     <dbl>     <dbl> ...
## 1 Alabama  5682.838  5840.550  5753.496  6008.169  6475.092  7188.954
## 2 Alaska   4328.281  4632.623  4918.501  5069.822  5075.482  5454.607
## 3 Arizona  5138.495  5415.516  5481.419  5681.638  6058.464  7263.204
## 4 Arkansas 5772.302  6082.379  6231.977  6414.900  6416.503  6627.092
## 5 California 5285.921  5527.881  5334.826  5672.472  5897.888  7258.771
## 6 Colorado  4703.777  5406.967  5596.348  6227.002  6284.137  6948.473
## # ... with 6 more variables: 2010-11 <dbl>, 2011-12 <dbl>, 2012-13 <dbl>,
## #   2013-14 <dbl>, 2014-15 <dbl>, 2015-16 <dbl>
```

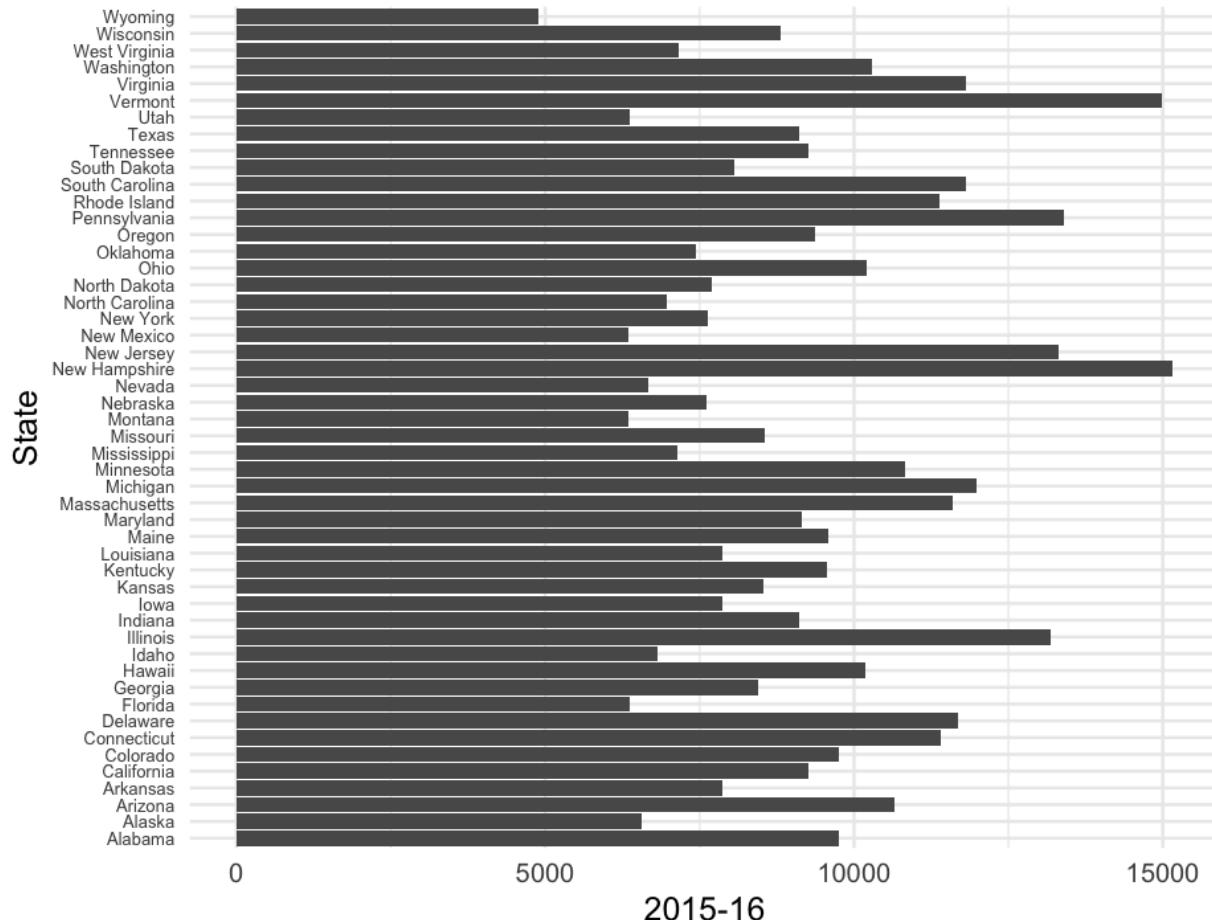
By state: 2015–16



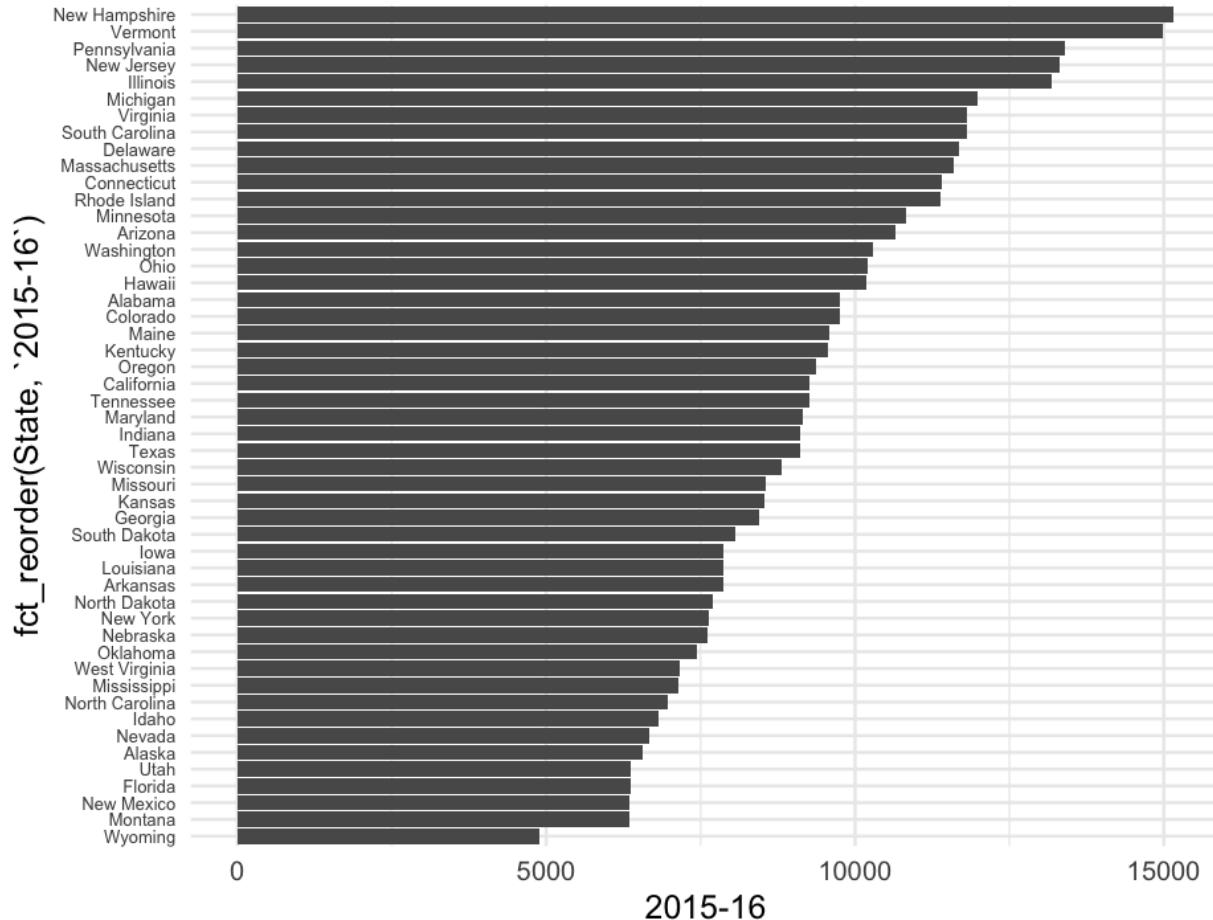
Two puke emoji version



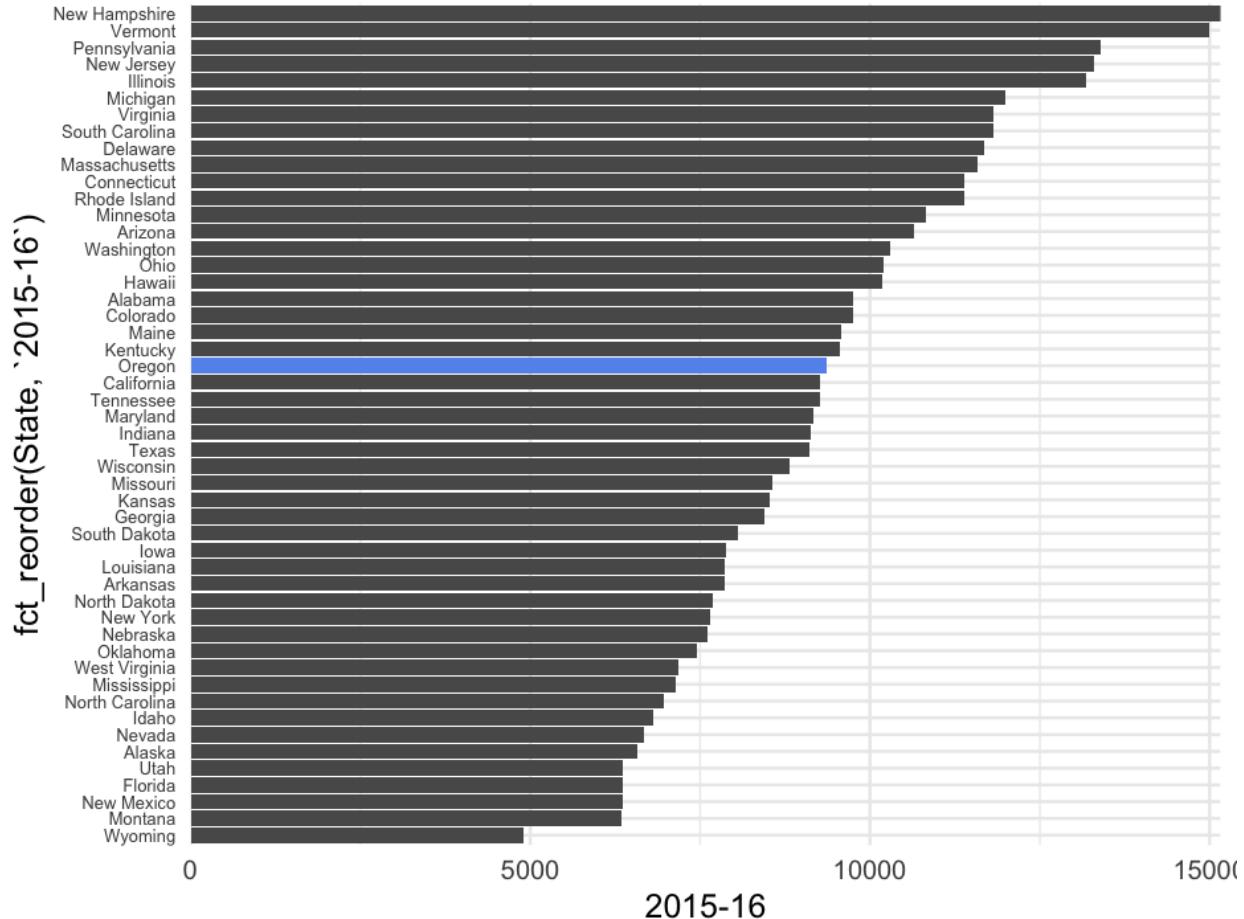
One puke emoji version



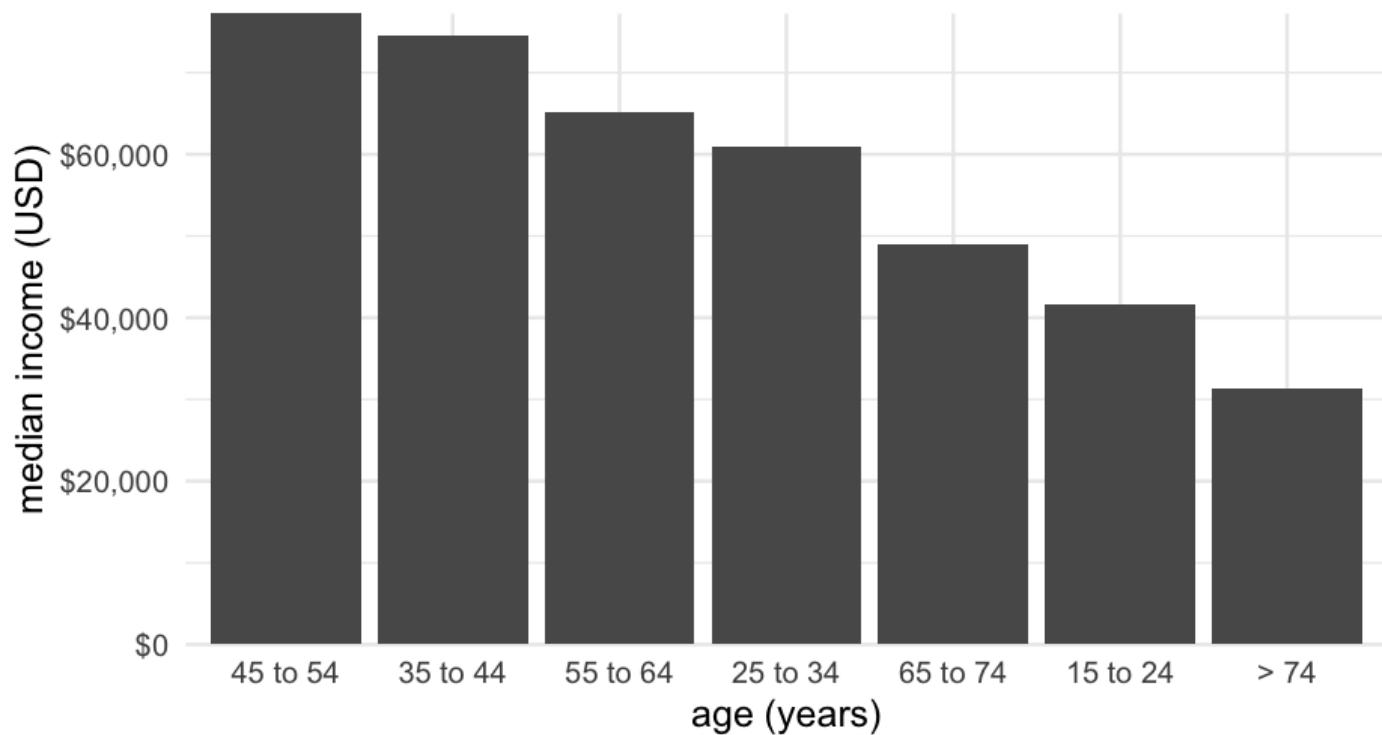
Kinda smiley version



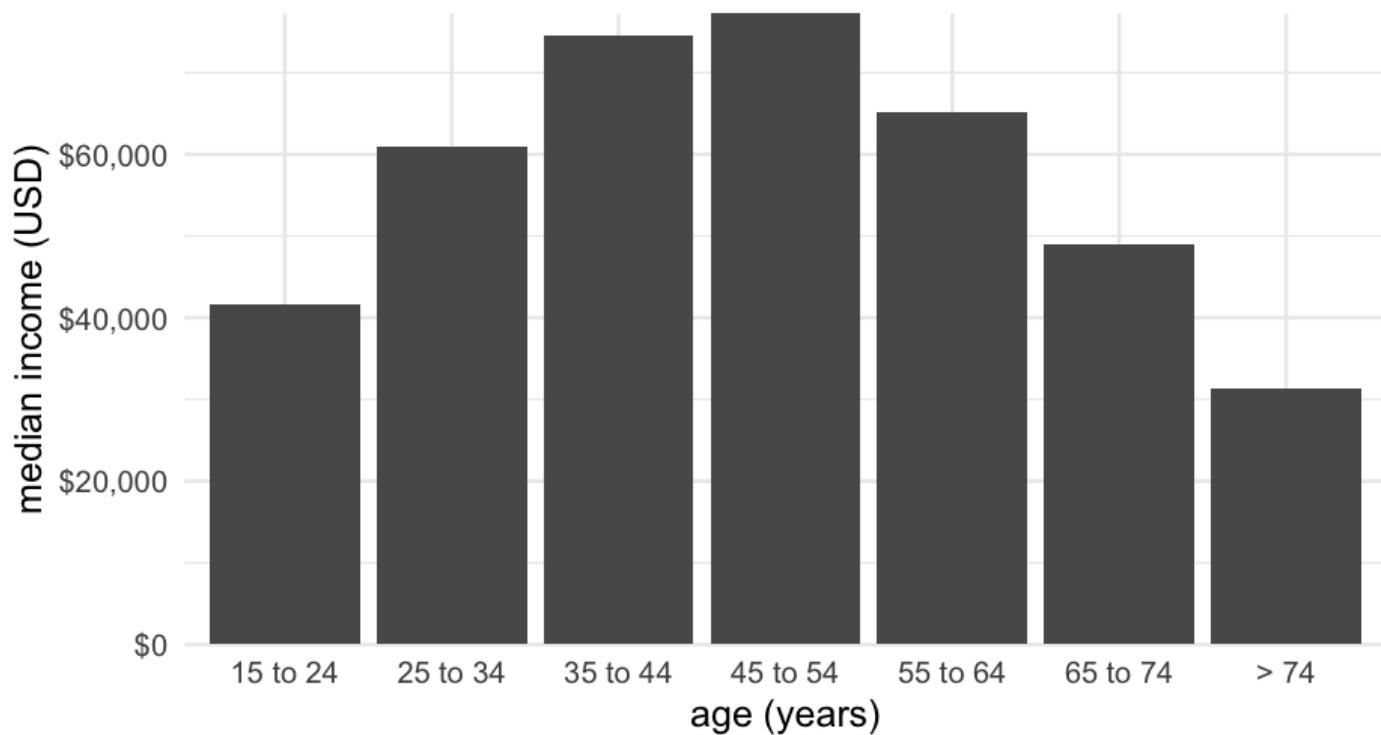
Highlight a state



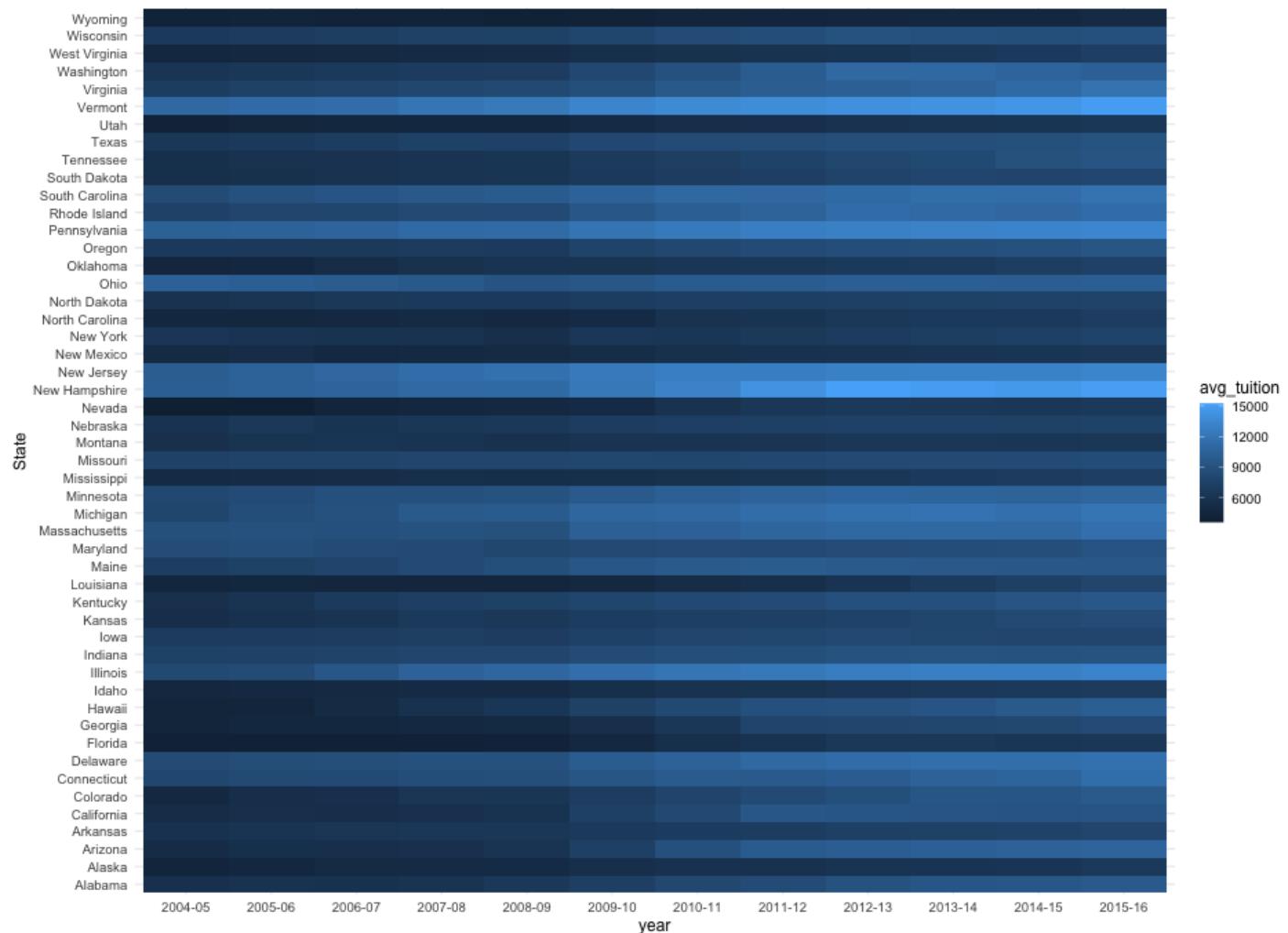
Not always good to sort



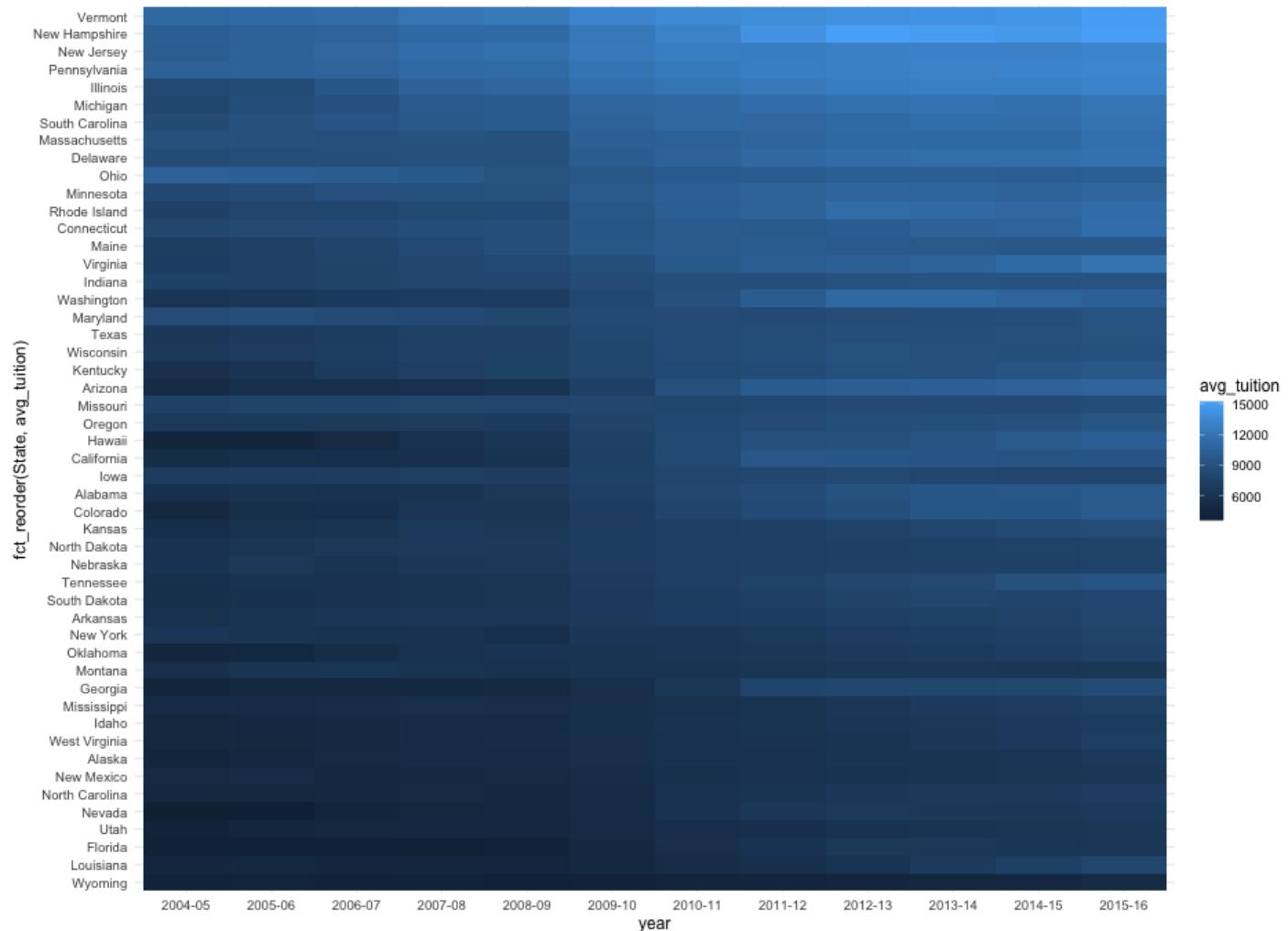
Much better



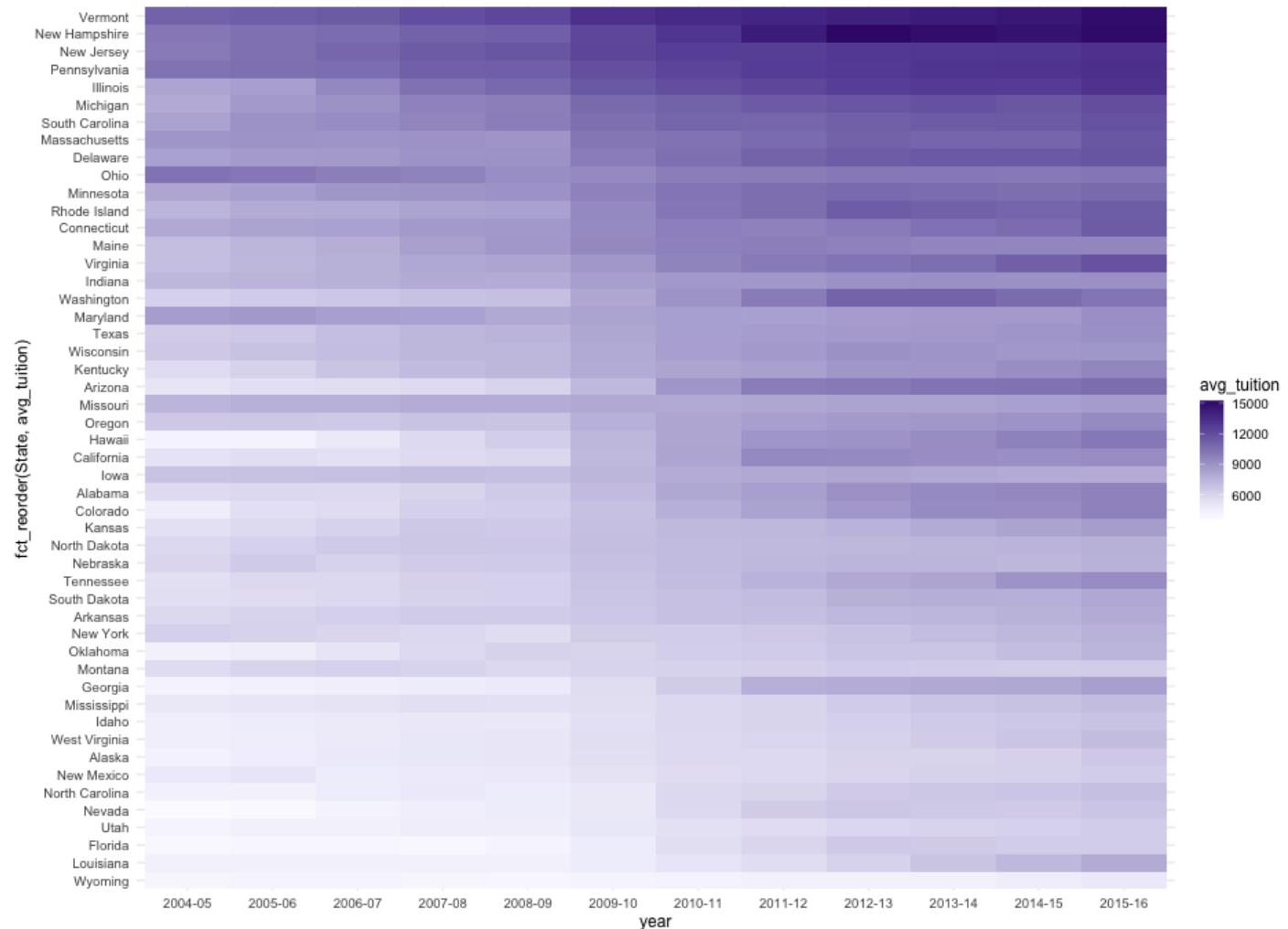
Heatmap



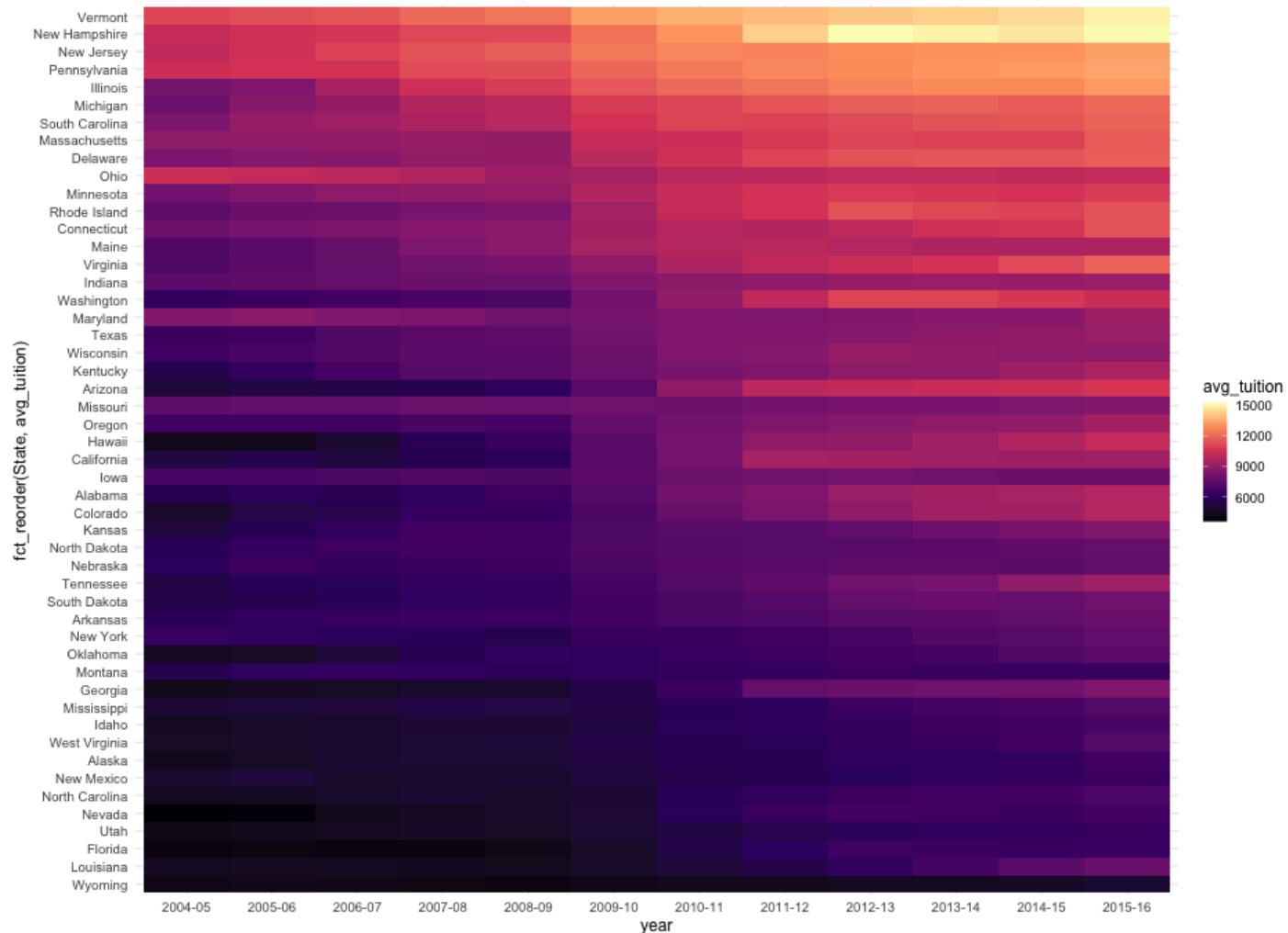
Better heatmap

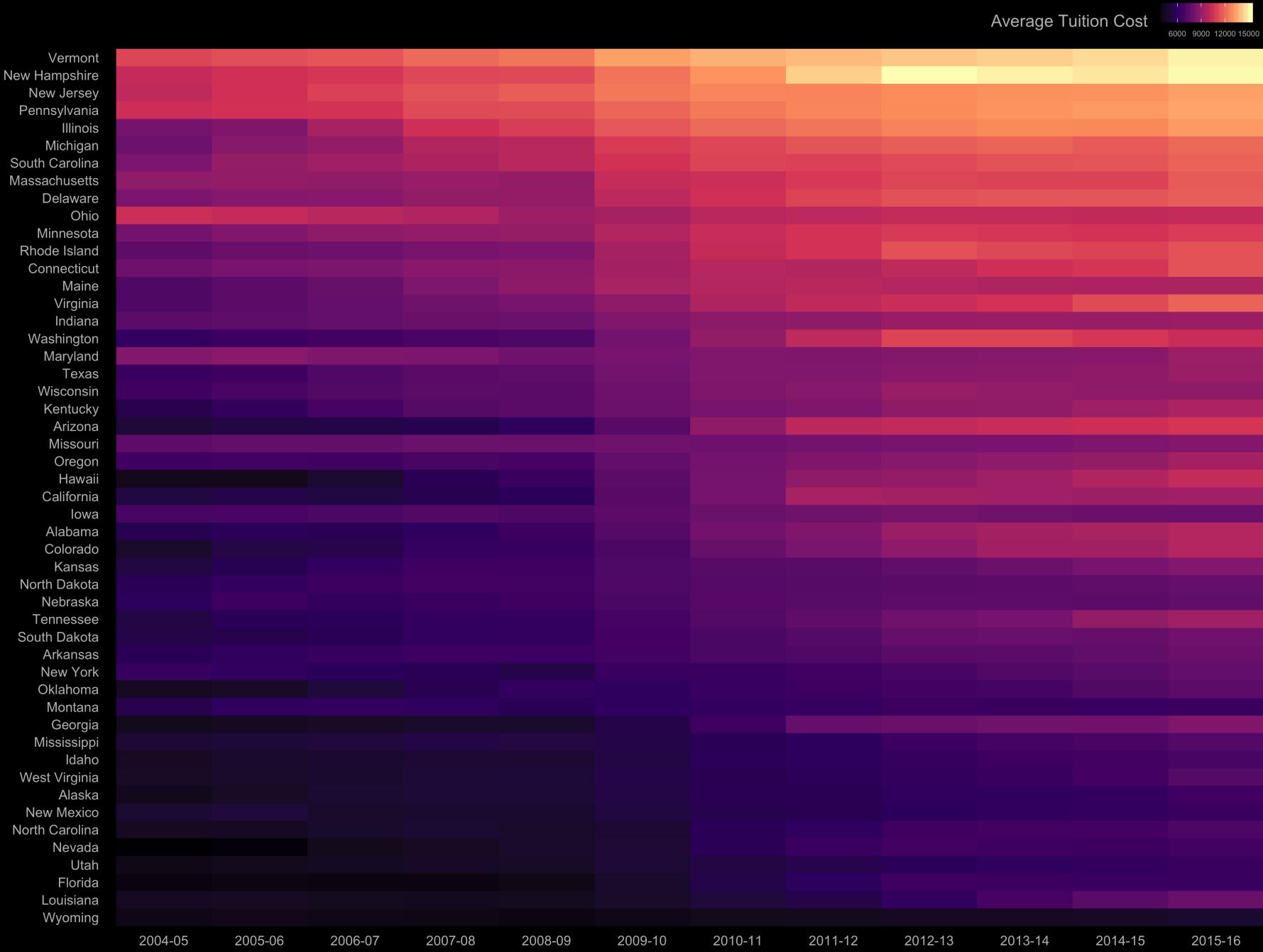


Even better?



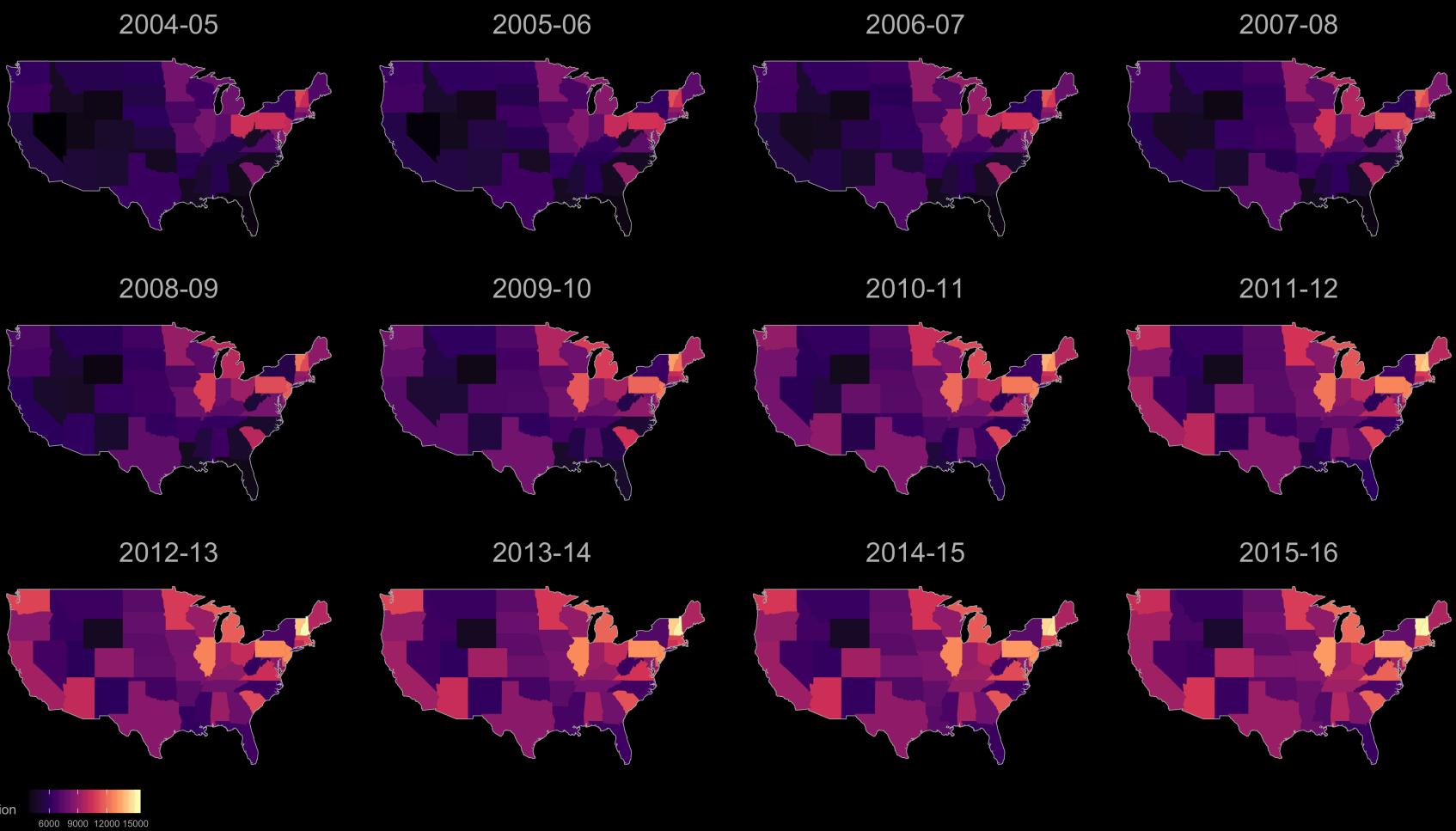
Or maybe this one?





Quick aside

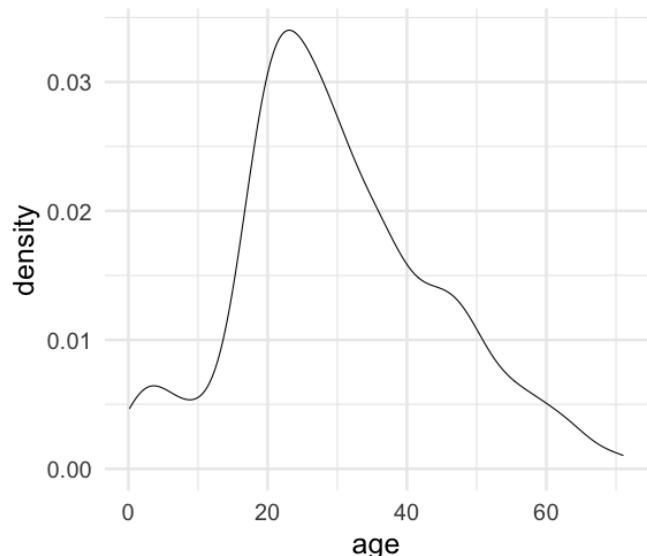
- Think about the data you have
- Given that these are state–level data, they have a geographic component



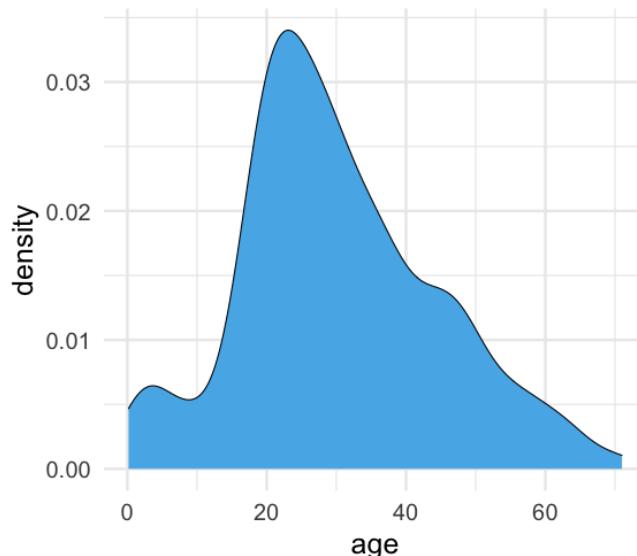
Some things
to avoid

Line drawings

As discussed earlier

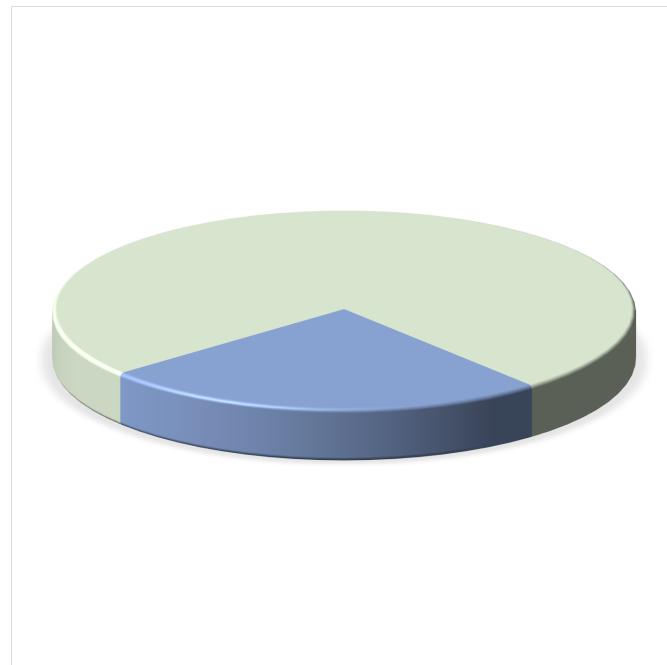
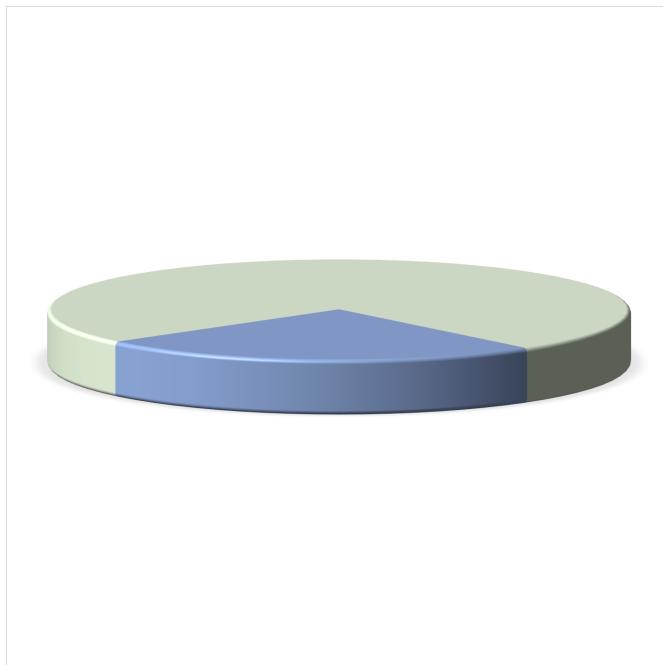


Change the fill



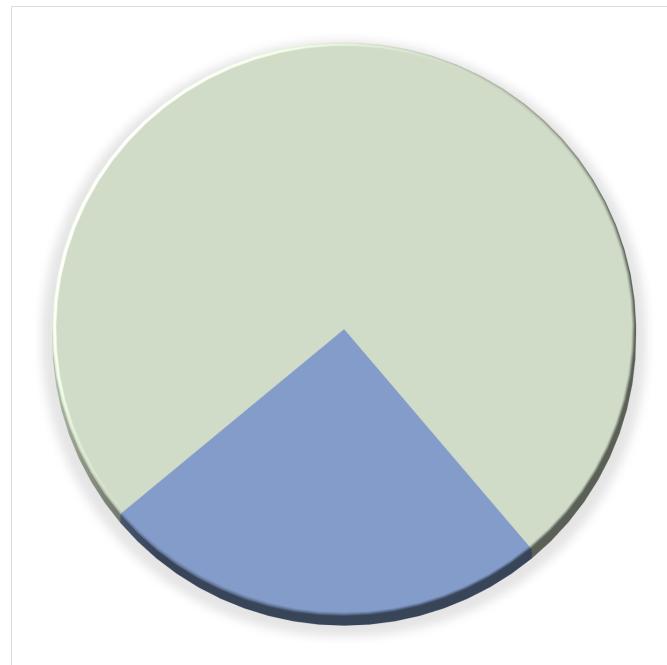
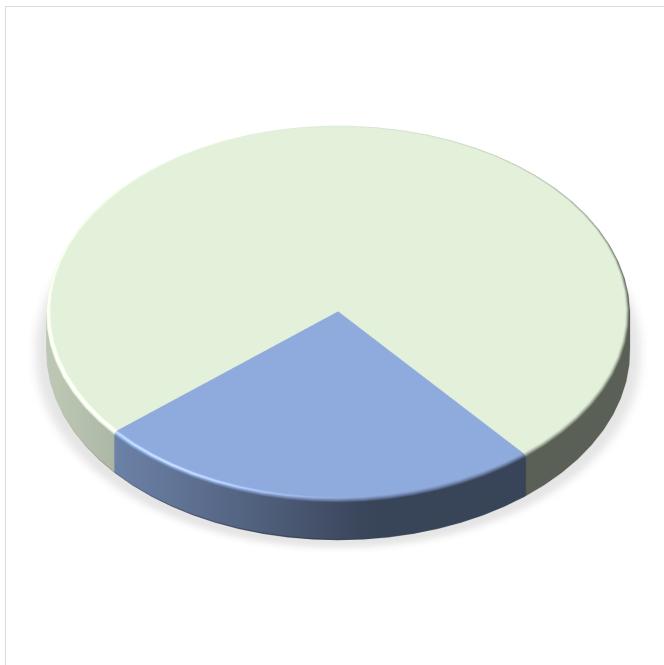
Much worse

Unnecessary 3D



Much worse

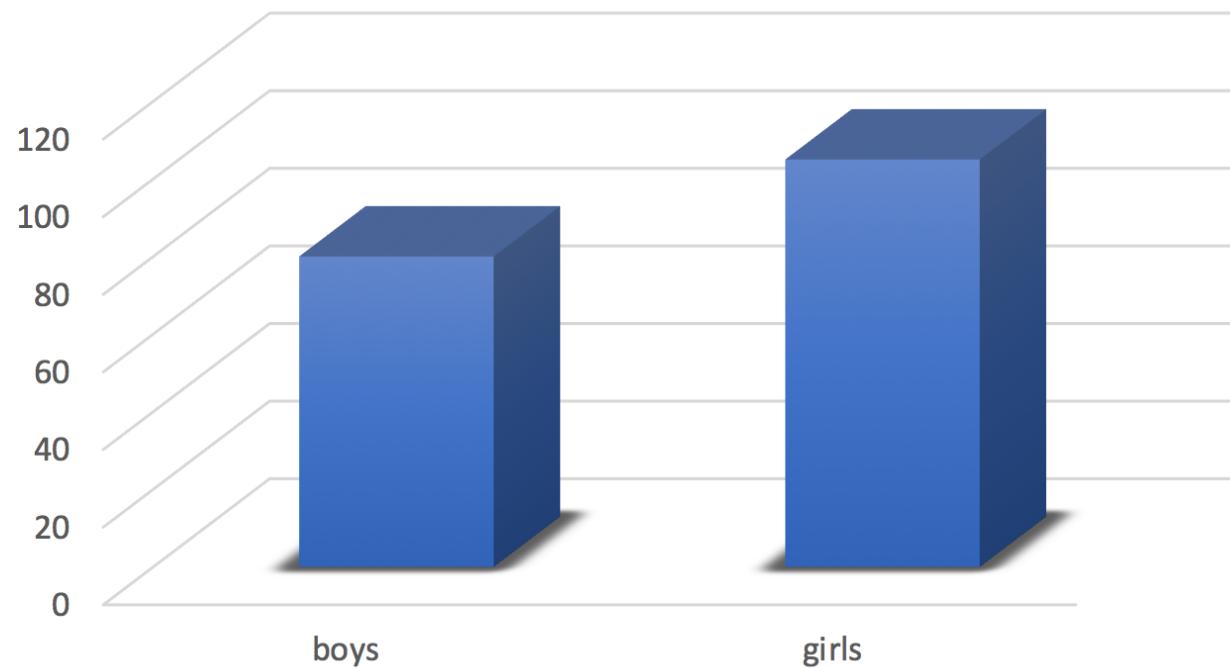
Unnecessary 3D



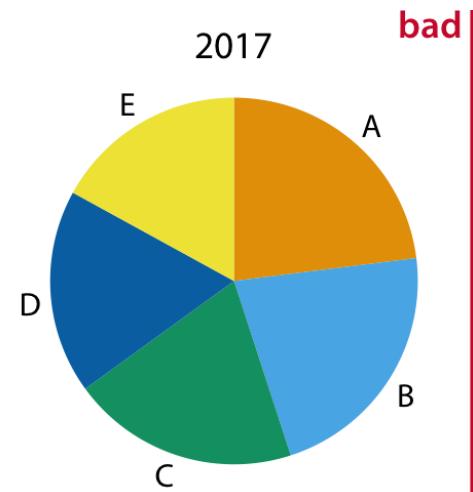
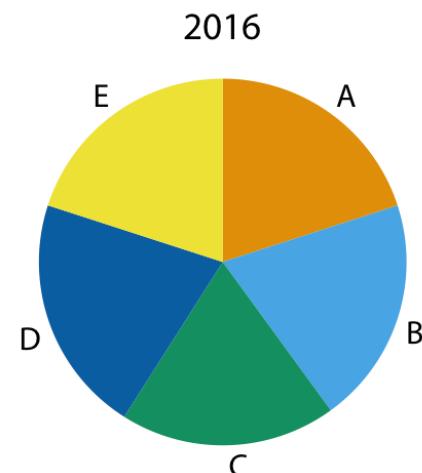
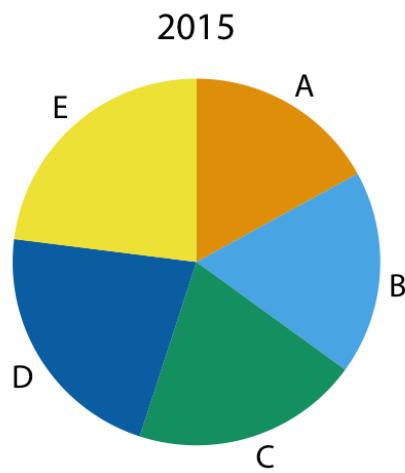
Horrid example

Used relatively regularly

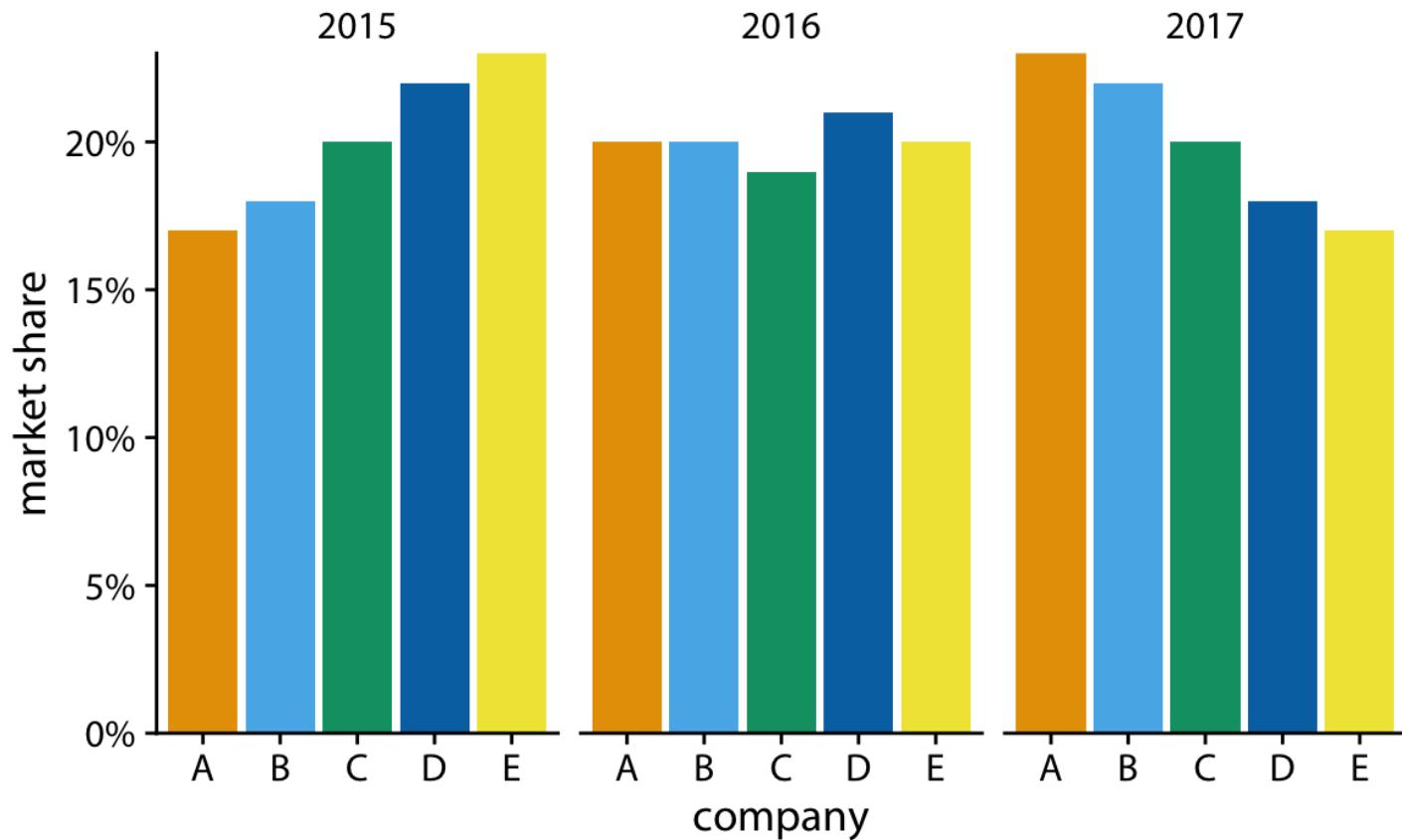
The left bar is 80; the right is 105



Pie charts w/lots of categories

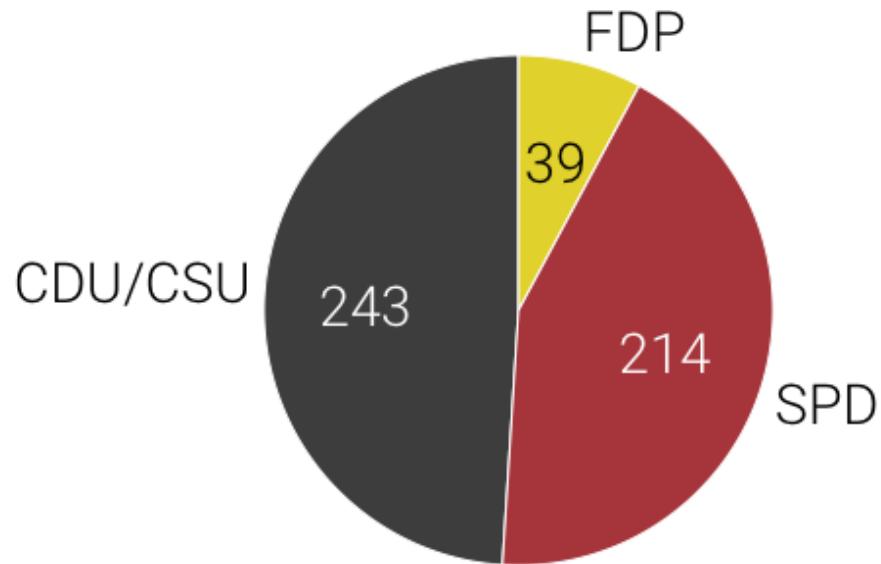


Alternative representation



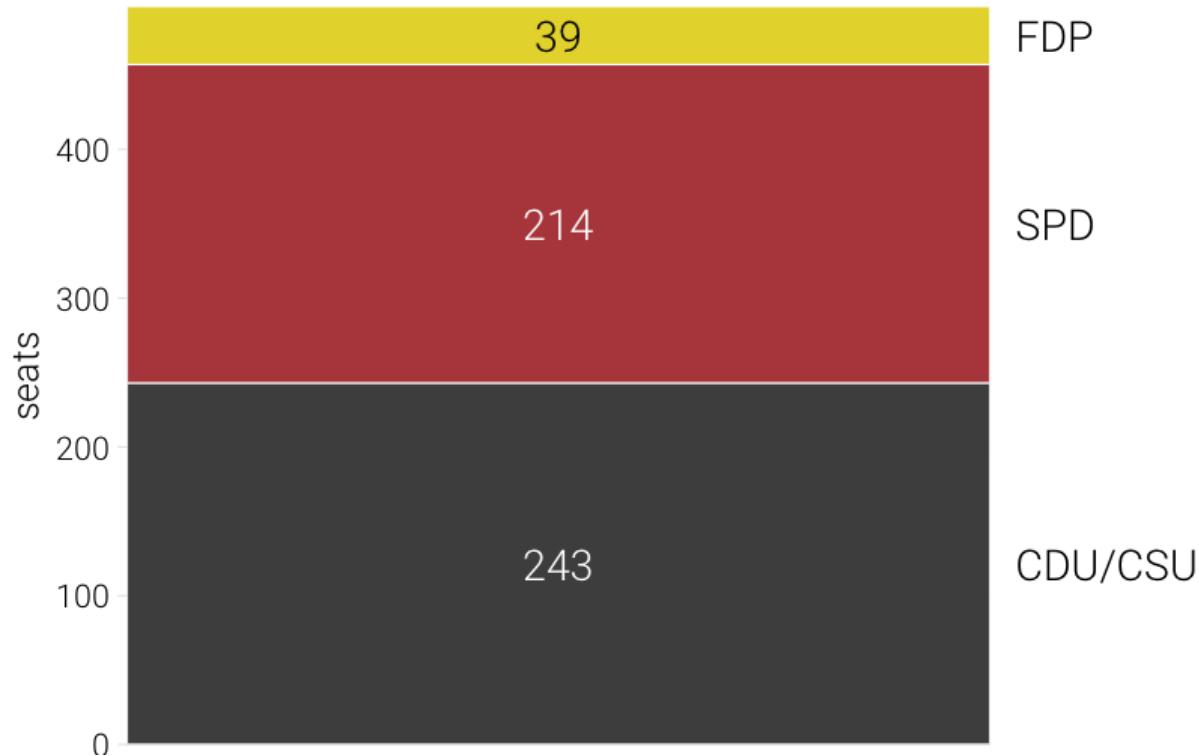
A case for pie charts

- n categories low,
- differences are relatively large
- familiar for some audiences

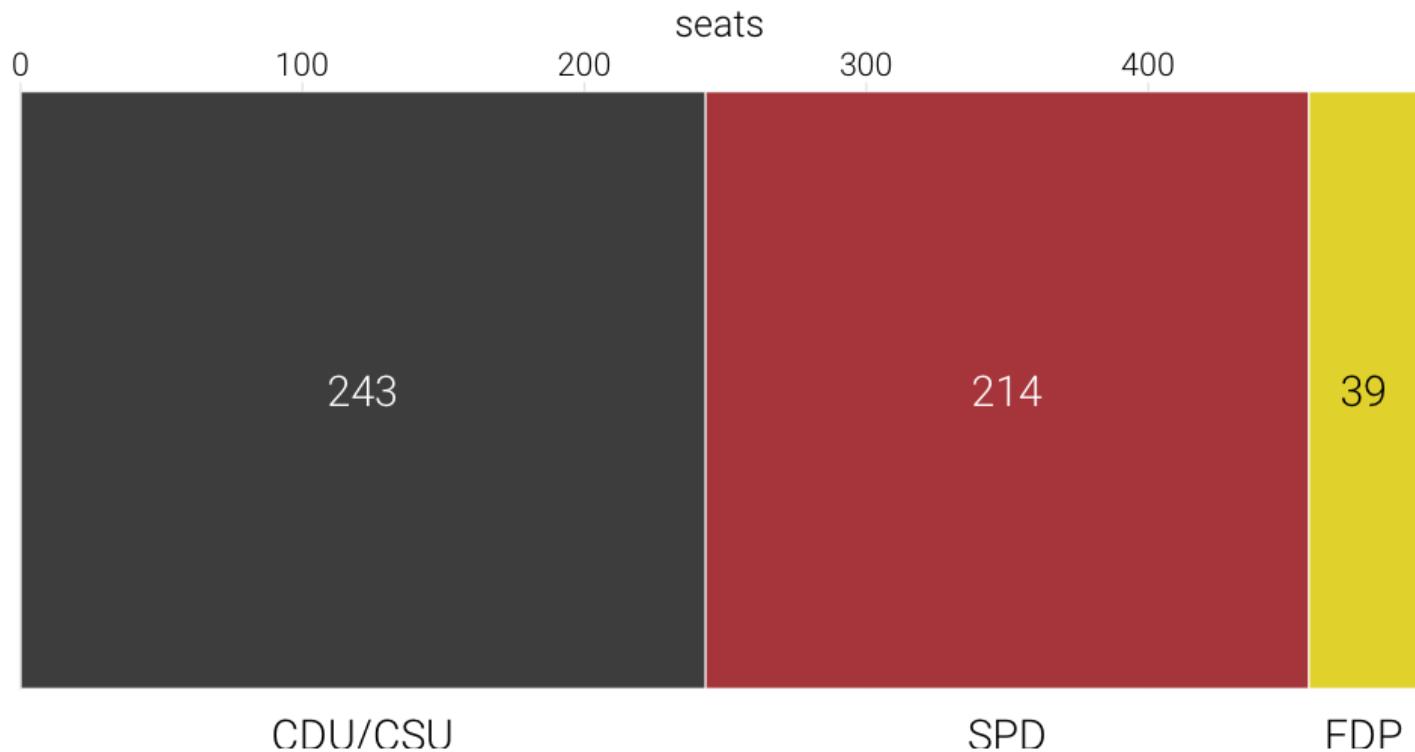


The anatomy of a pie chart

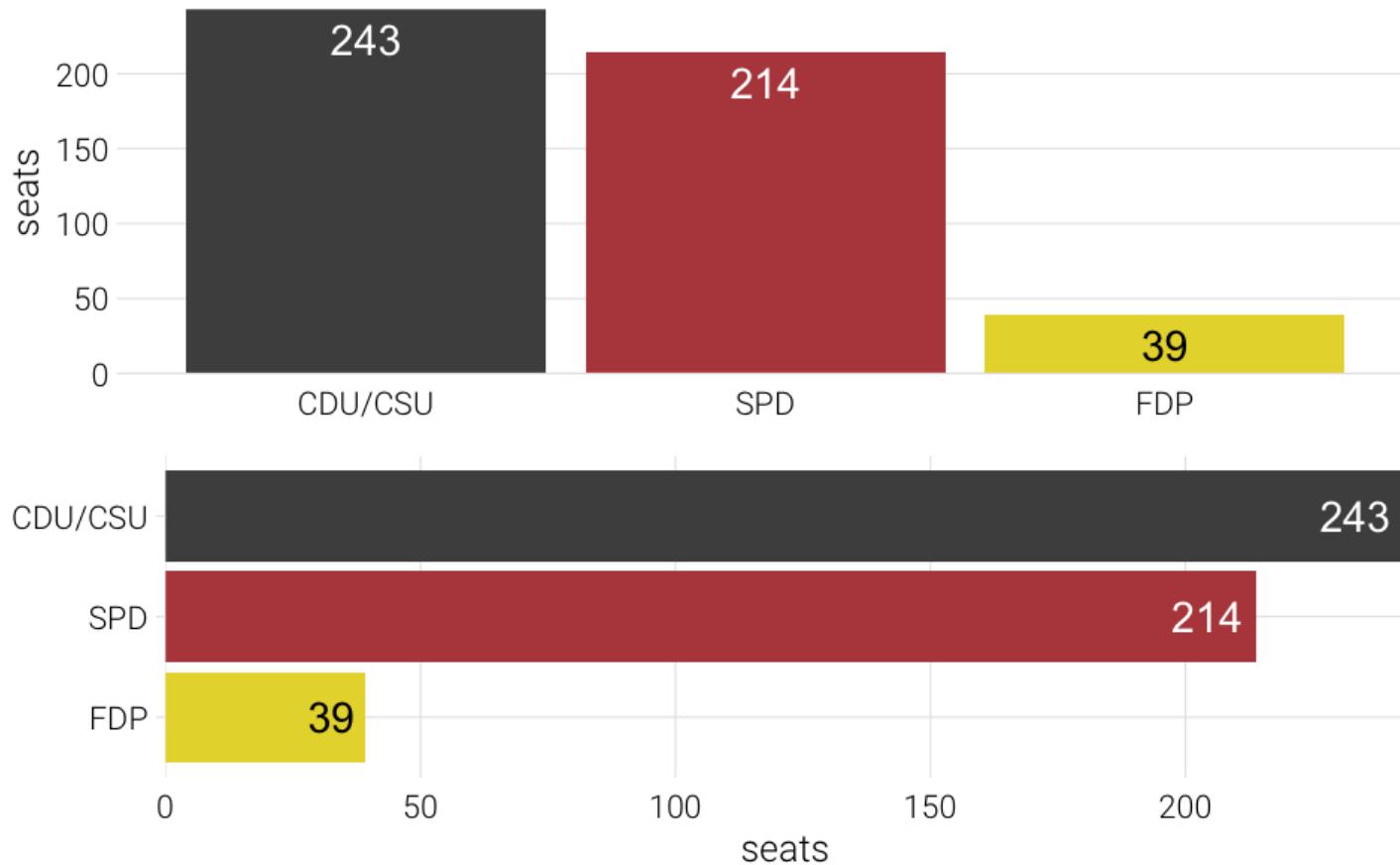
Pie charts are just stacked bar charts with a radial coordinate system



Horizontal

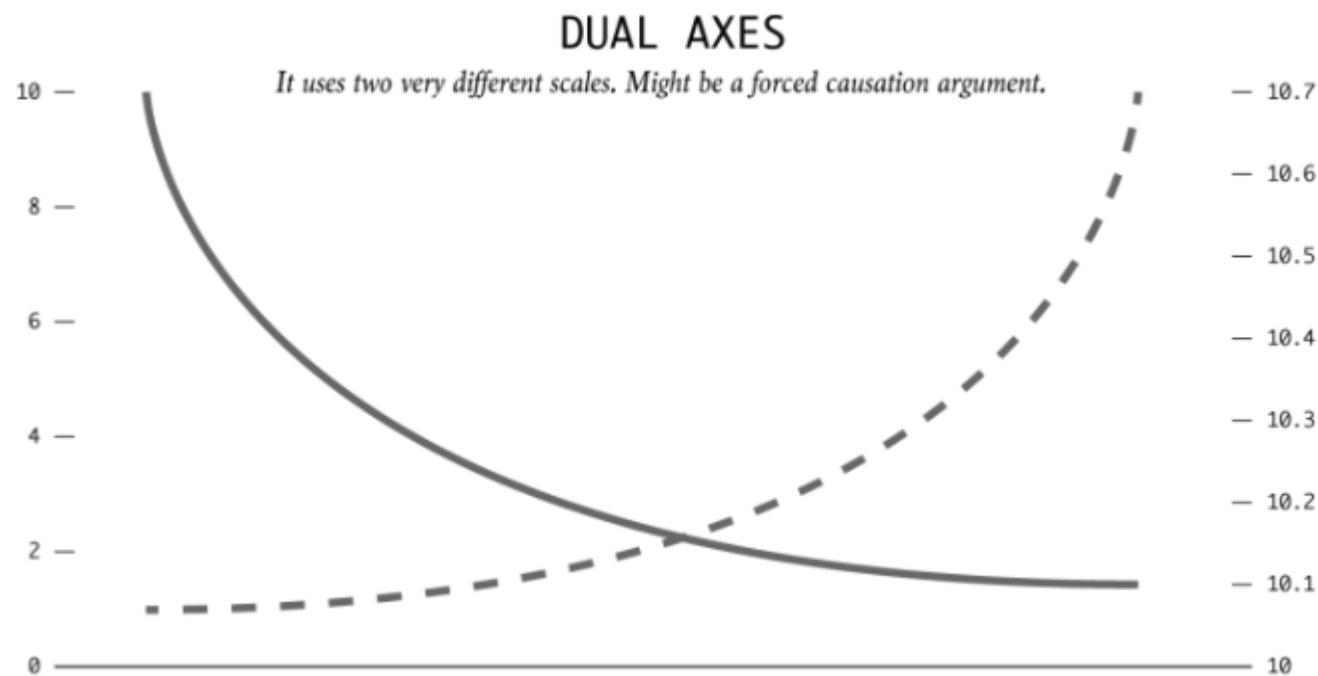


My preference, generally



Dual axes

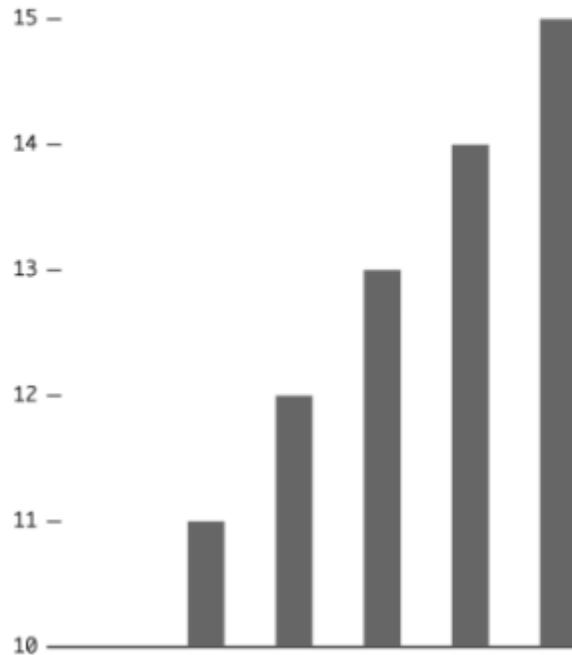
- Exception: if second axis is a direct transformation of the first



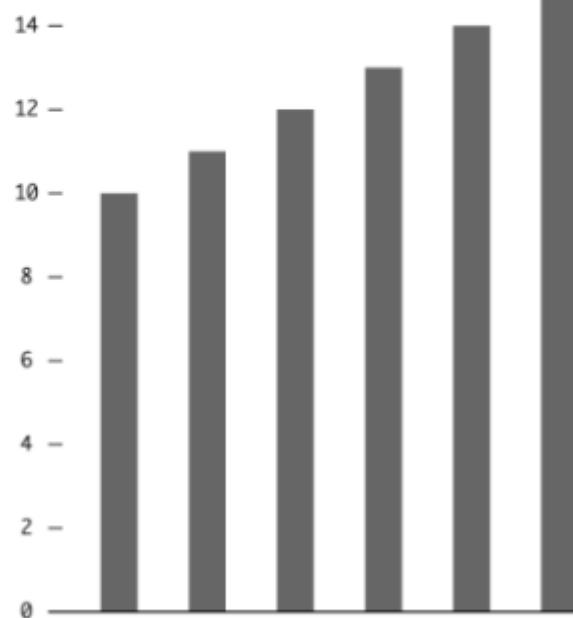
Truncated axes

TRUNCATED AXIS

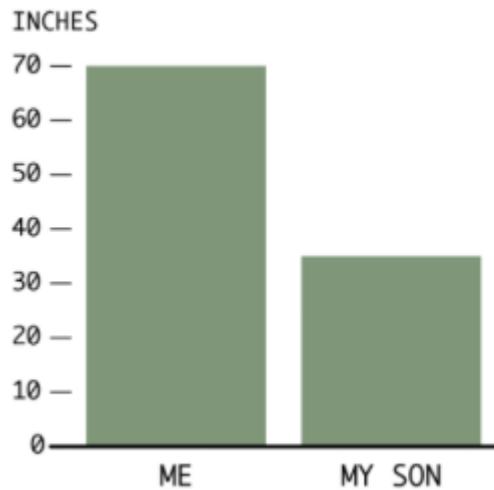
The value axis starts at ten. Liar, liar, pants on fire.



The value axis starts at zero. Good.

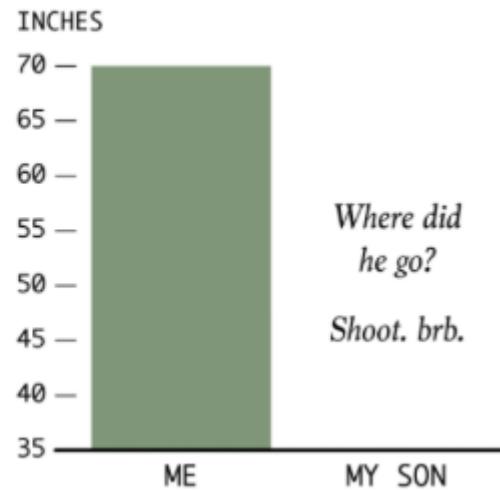


Height



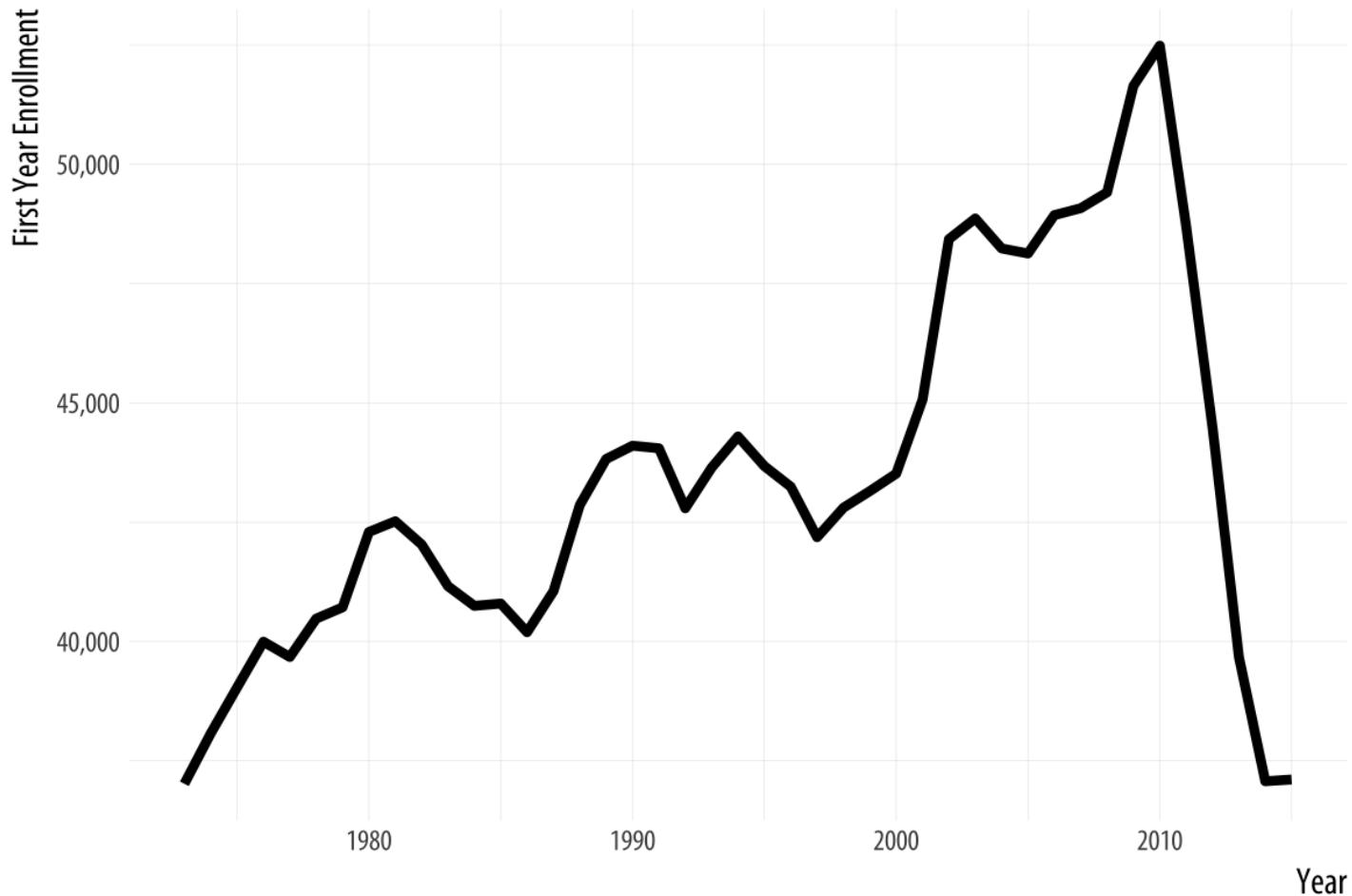
VS.

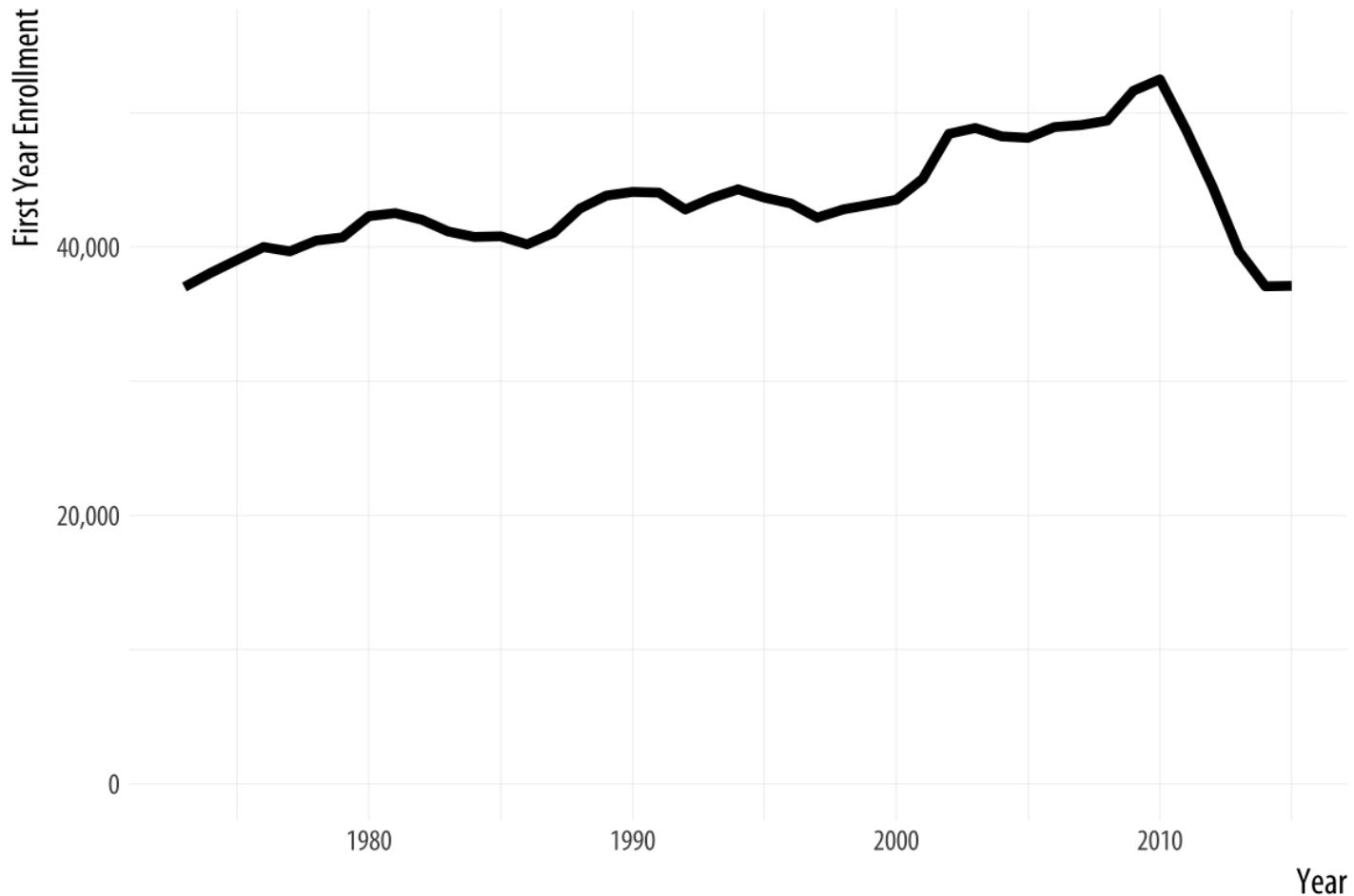
Height



Not always a bad thing

It is tempting to lay down inflexible rules about what to do in terms of producing your graphs, and to dismiss people who don't follow them as producing junk charts or lying with statistics. But being honest with your data is a bigger problem than can be solved by rules of thumb about making graphs. In this case there is a moderate level of agreement that bar charts should generally include a zero baseline (or equivalent) given that bars encode their variables as lengths. But it would be a mistake to think that a dot plot was by the same token deliberately misleading, just because it kept itself to the range of the data instead.



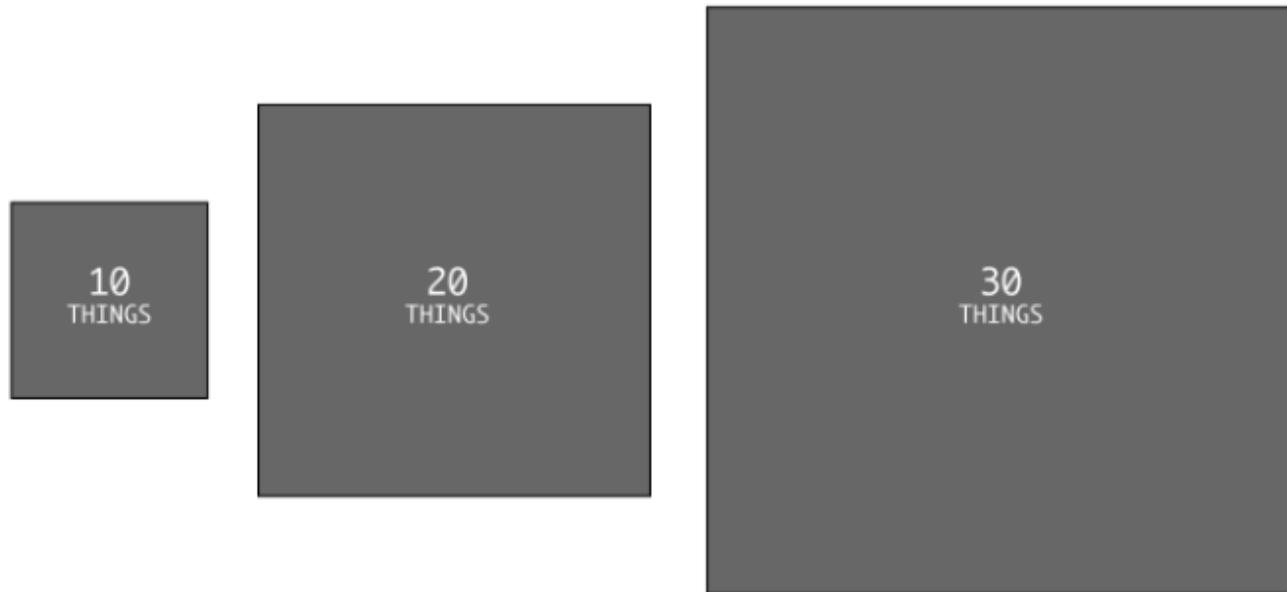


Scaling issues

AREA SIZED BY SINGLE DIMENSION

Thirty is three times ten, but that third rectangle looks a lot bigger than the first.

Might be trying to inflate significance.



Poor binning choices

ODD CHOICE OF BINNING

*Two bins. What's really in the 1+ category?
Might be hiding something.*



That's better. It can show more variation.



Example from flowing data

Conclusions

Practical takeaways to make
better visualizations

1. Avoid line drawings
2. Sort bar charts in ascending/descending order as long as the other axis does not have implicit meaning
3. Consider dropping legends and using annotations, when possible
4. Use color to your advantage, but be sensitive to color-blindness, and use the right kind of palette
5. Consider double-encoding data (shapes and color)
6. Make your labels bigger! Didn't talk about this one much but it's super common and really important

Some things to avoid

- Essentially never
 - Use dual axes (produce separate plots instead)
 - Use 3D unnecessarily
- Be wary of
 - Truncated axes
 - Pie charts (particularly with lots of categories)

Mediums for communication

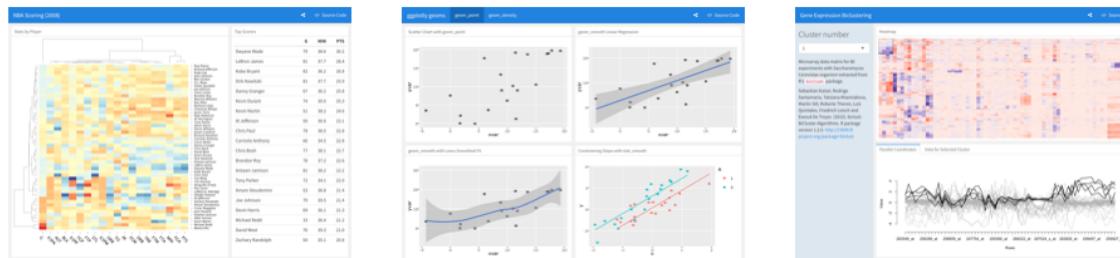
Probably mostly out of time but...

Flex Dashboards!

flexdashboard for R Home Using Shiny Layouts Examples

flexdashboard: Easy interactive dashboards for R

- Use [R Markdown](#) to publish a group of related data visualizations as a dashboard.
- Support for a wide variety of components including [htmlwidgets](#); base, lattice, and grid graphics; tabular data; gauges and value boxes; and text annotations.
- Flexible and easy to specify row and column-based [layouts](#). Components are intelligently re-sized to fill the browser and adapted for display on mobile devices.
- [Storyboard](#) layouts for presenting sequences of visualizations and related commentary.
- Optionally use [Shiny](#) to drive visualizations dynamically.



Getting Started

Install the **flexdashboard** package from CRAN as follows:

```
install.packages("flexdashboard")
```

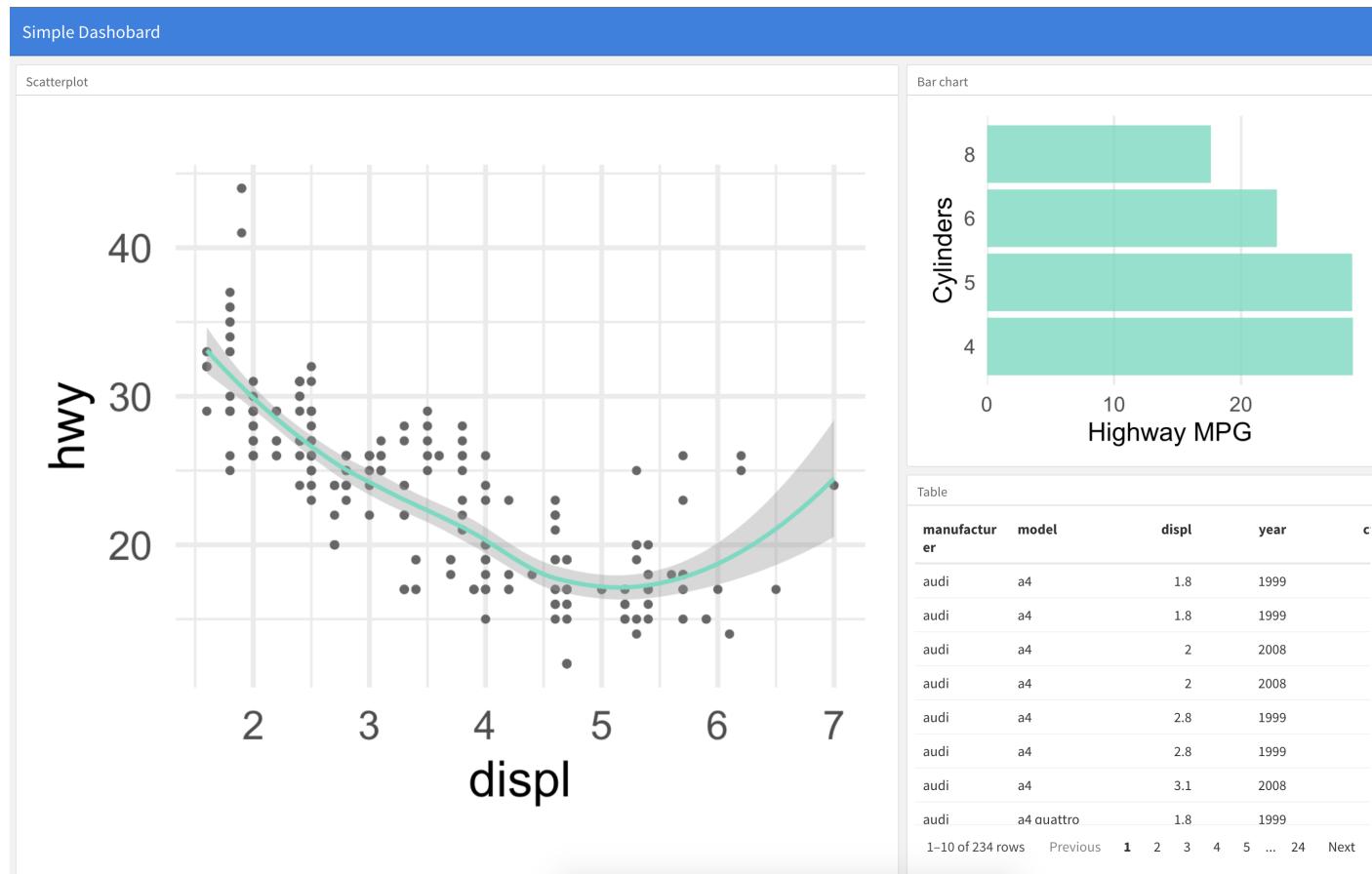
Example code



The screenshot shows the RStudio interface with a file named "example-dashboard.Rmd" open. The code is written in R Markdown, combining R code chunks and Rmd syntax.

```
1 ---  
2 title: "Simple Dashboard"  
3 output:  
4   flexdashboard::flex_dashboard:  
5     orientation: columns  
6     vertical_layout: fill  
7 ---  
8  
9 ```{r setup, include=FALSE}  
10 library(flexdashboard)  
11 ```  
12  
13 Column {data-width=650}  
14 -----  
15  
16 ### Scatterplot  
17  
18 ```{r}  
19 library(reactable)  
20 library(tidyverse)  
21 theme_set(theme_minimal(25))  
22  
23 ggplot(mpg, aes(displ, hwy)) +  
24   geom_point(color = "gray40") +  
25   geom_smooth(color = "#57DBC1")  
26 ```  
27  
28 Column {data-width=350}  
29 -----  
30  
31 ### Bar chart  
32  
33 ```{r}  
34 mpg %>%  
35   group_by(cyl) %>%  
36   summarize(mean = mean(hwy)) %>%  
37   ggplot(aes(mean, factor(cyl))) +  
38   geom_col(fill = "#57DBC1", alpha = 0.8) +  
39   labs(y = "Cylinders",  
40         x = "Highway MPG") +  
41   scale_x_continuous(expand = c(0, 0)) +  
42   theme(panel.grid.minor = element_blank(),  
43         panel.grid.major.y = element_blank())  
44 ```  
45  
46 ### Table  
47  
48 ```{r}  
49 reactable(mpg)  
50```  
51 --
```

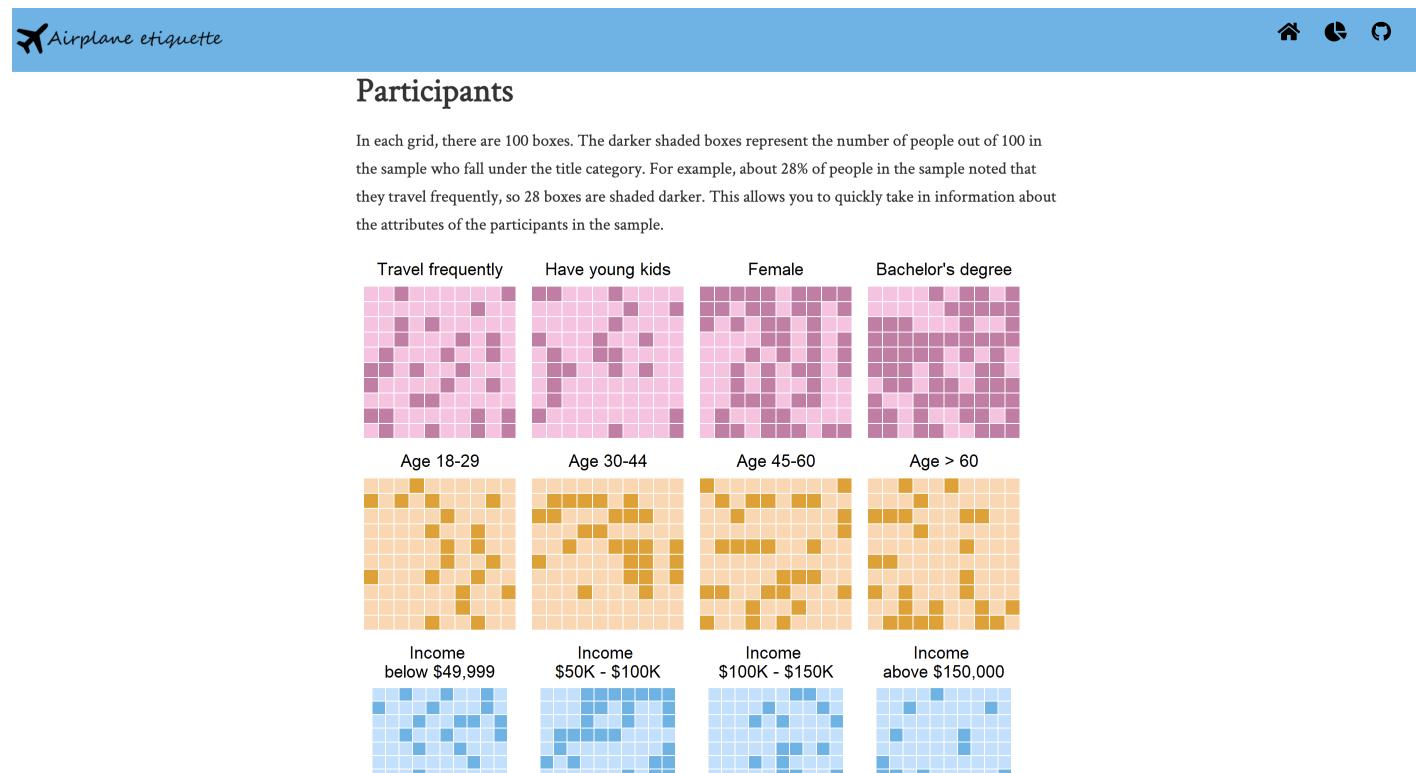
Example output



Other formats (R centric)

Blog posts through [distill](#)

Example from a former student.



Blogdown

For full customization, go [blogdown](#) – but buyer beware
(distill is probably better for most academic sharing)

See [Alison's](#) site for a great example, along with great
tutorials for getting started

Thanks!

Questions?

 @datalorax_

 @datalorax

 website

 daniela@uoregon.edu