Some data viz advice

Daniel Anderson

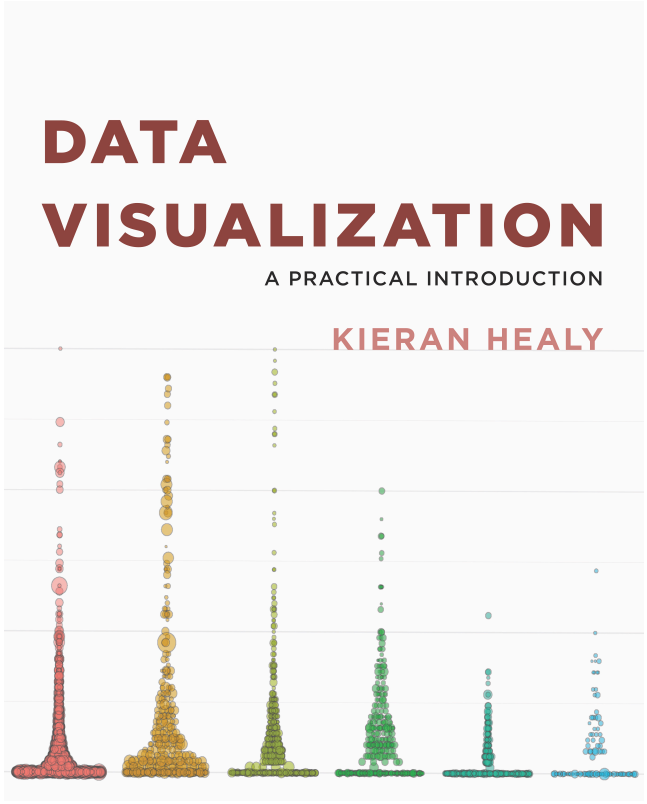
Thursday, April 8, 2021

whoami

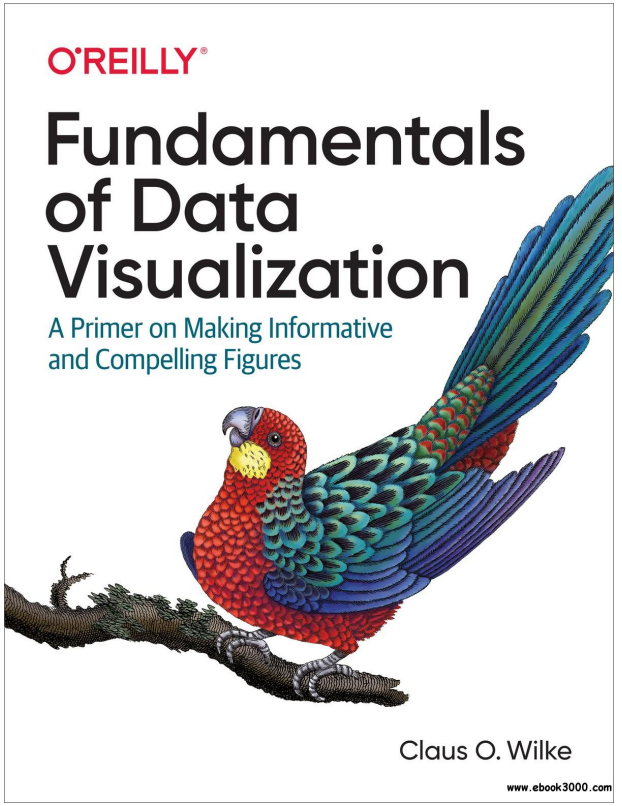
- Research Assistant Professor: Behavioral Research and Teaching
- Dad (two daughters: 8 and 6)
- Pronouns: he/him/his
- Primary areas of interest: 💗💗 R💗💗, computational research, achievement gaps, systemic inequities, and variance between educational institutions

Resources (free)

Healy



Wilke

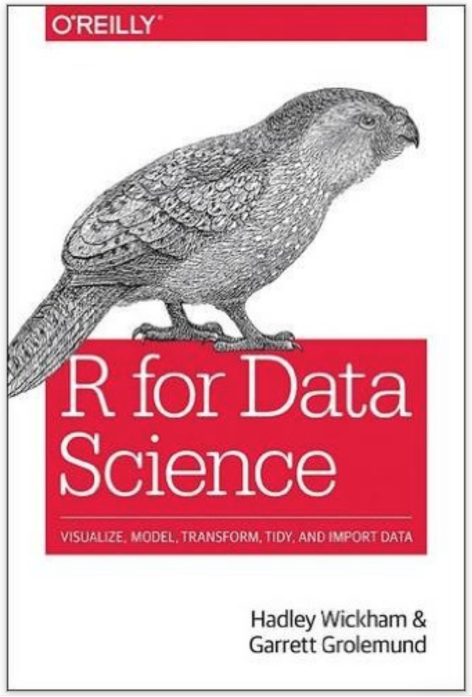


Other Resources

- 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

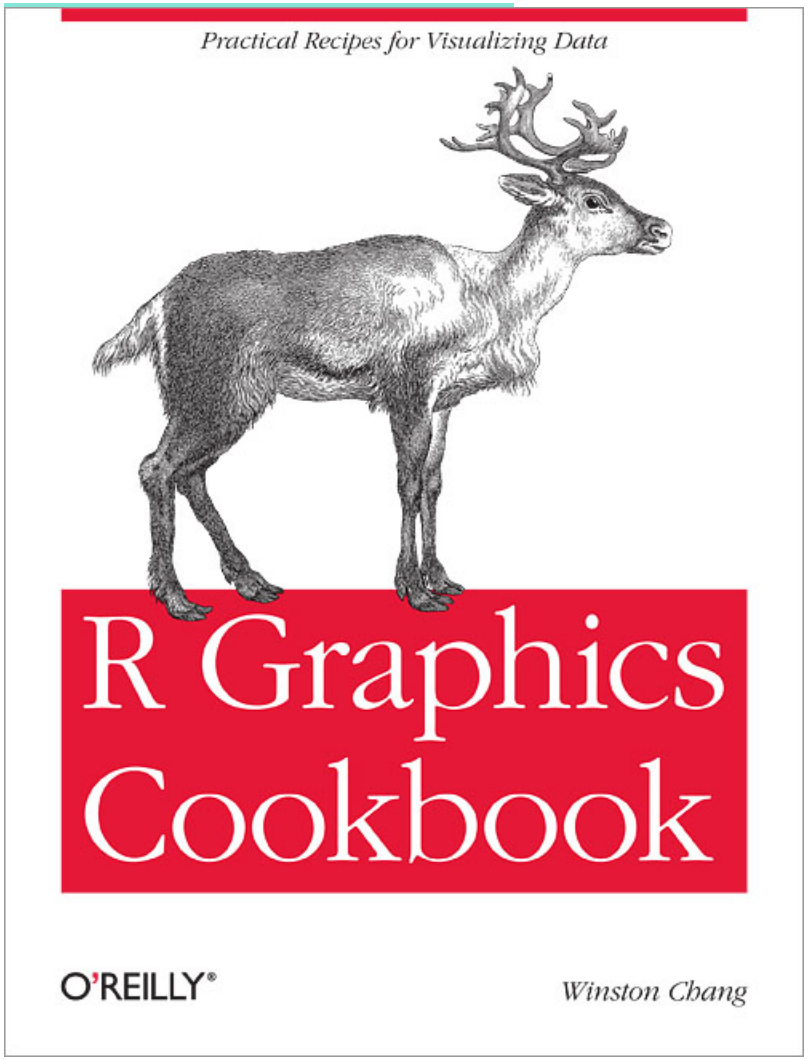
Where to start?

- I *really* recommend moving to R as quickly as possible

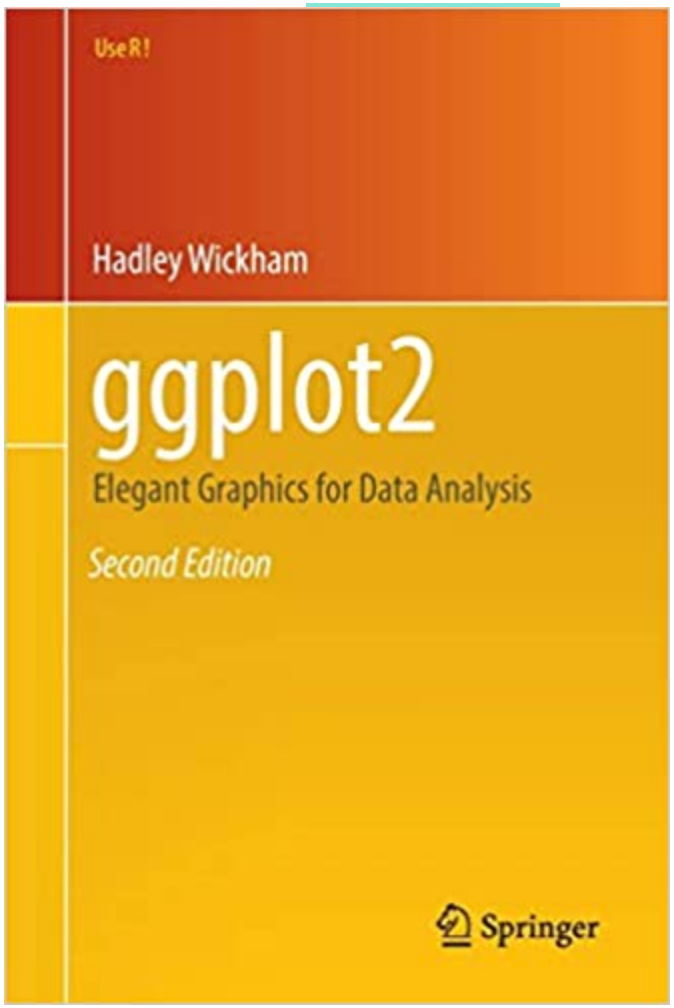


ggplot2!

https://r-graphics.org



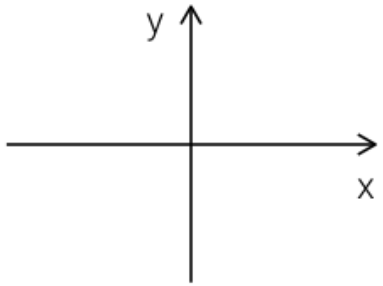
Third edition in progress!

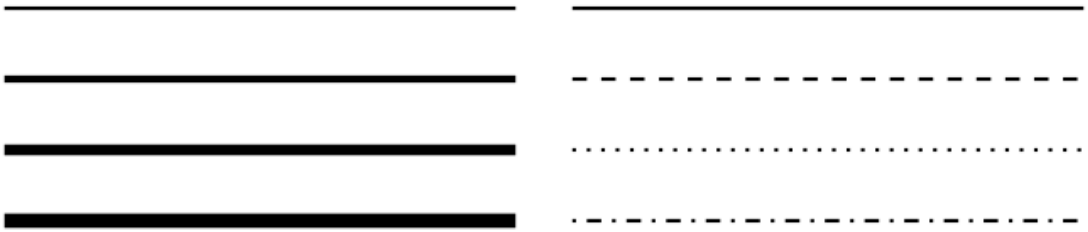


Last note before we really start

- 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





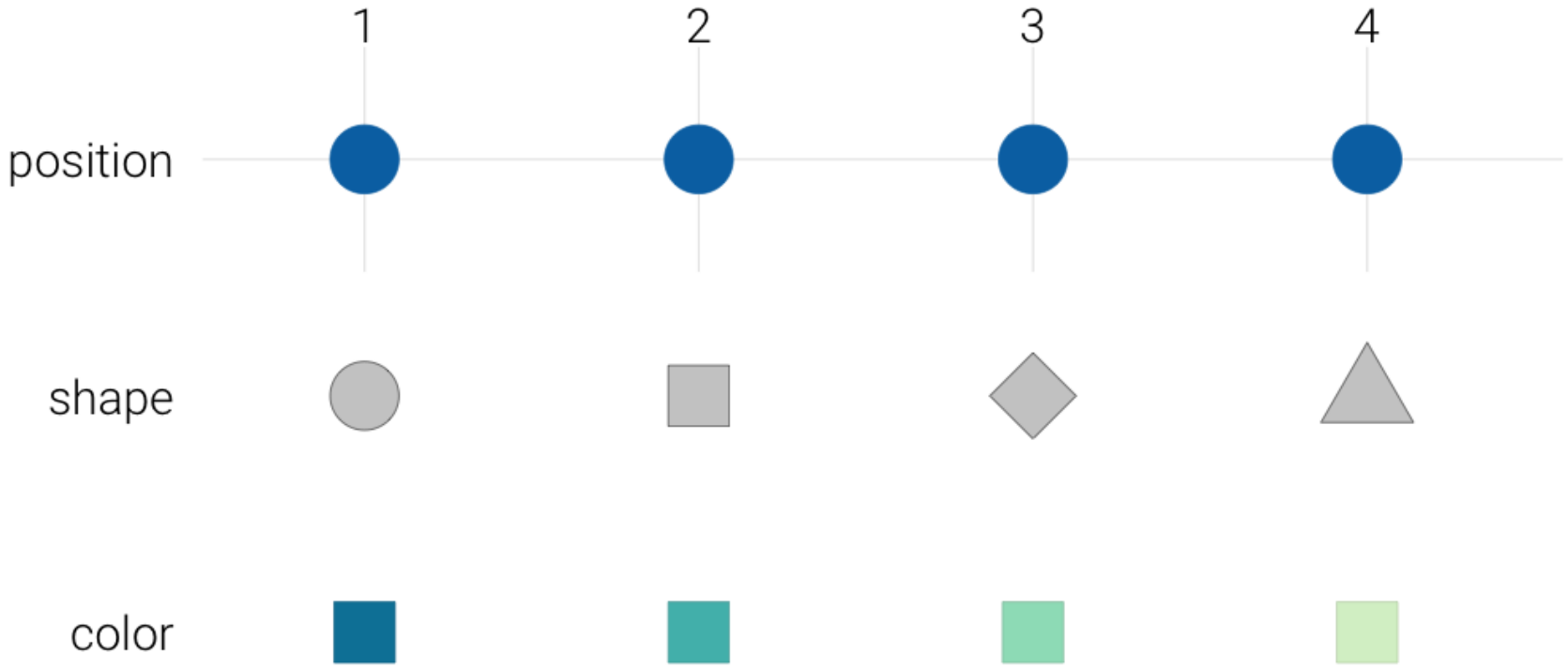
Other elements to consider

- 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?

Other elements to consider

- Type of data
  - Continuous/categorical
  - Which can be mapped to each aesthetic?
  - e.g., shape and line type can only be mapped to categorical data, whereas color and size can be mapped to either.

Basic Scales

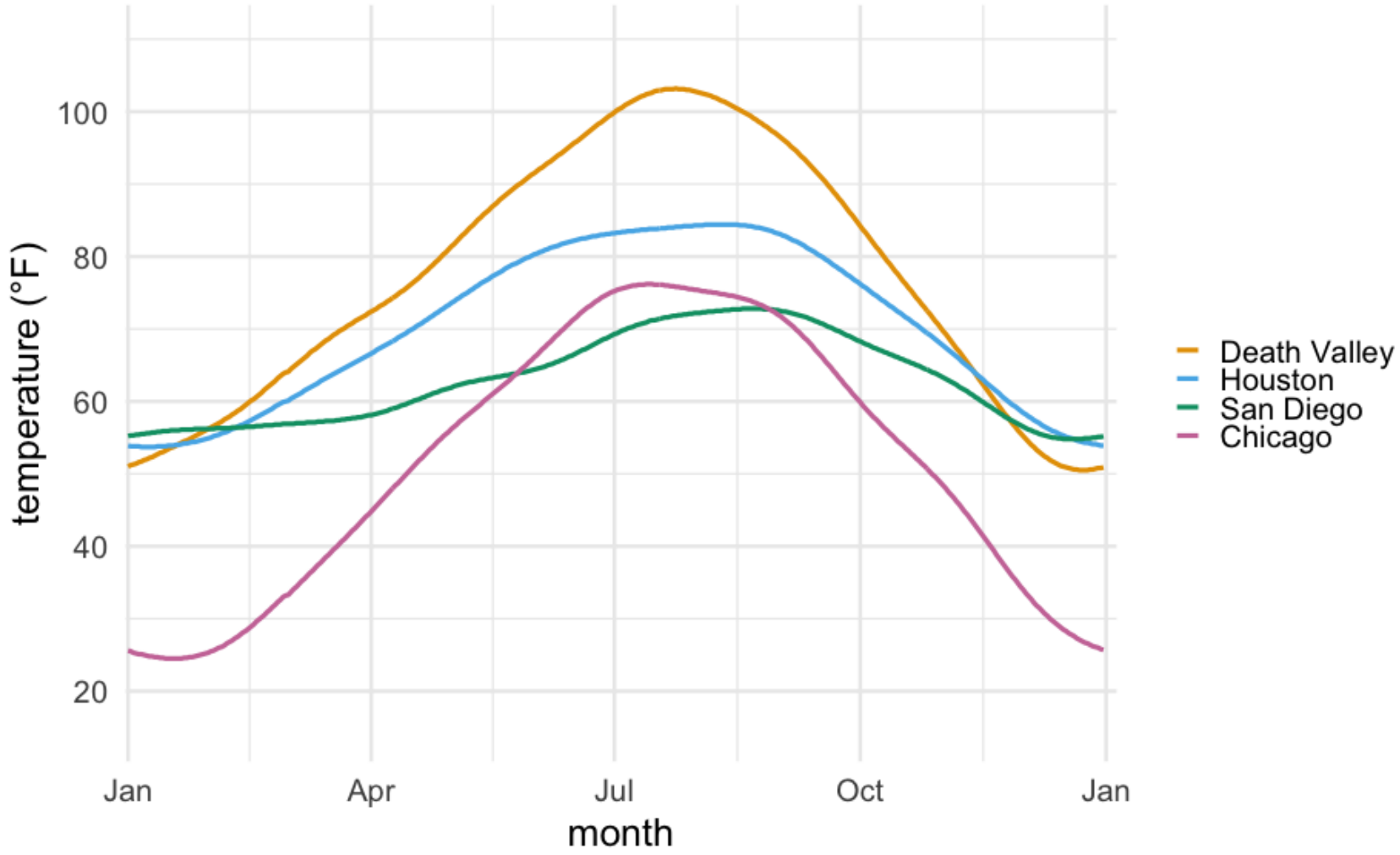


Talk with a neighbor

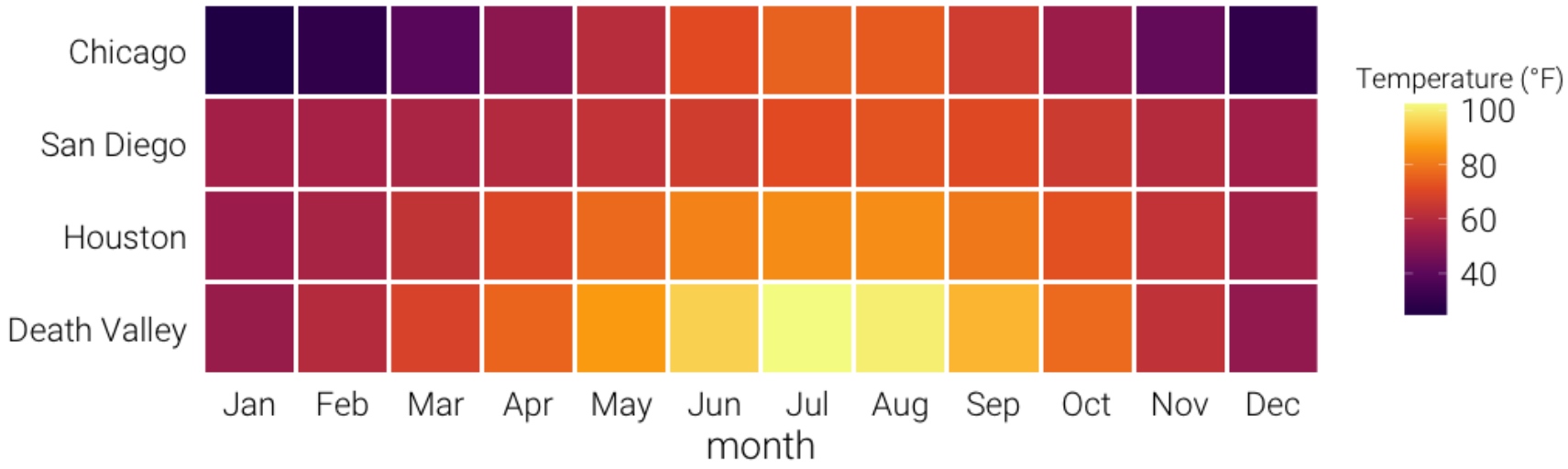
How would you encode these data into a display?

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

Putting it to practice



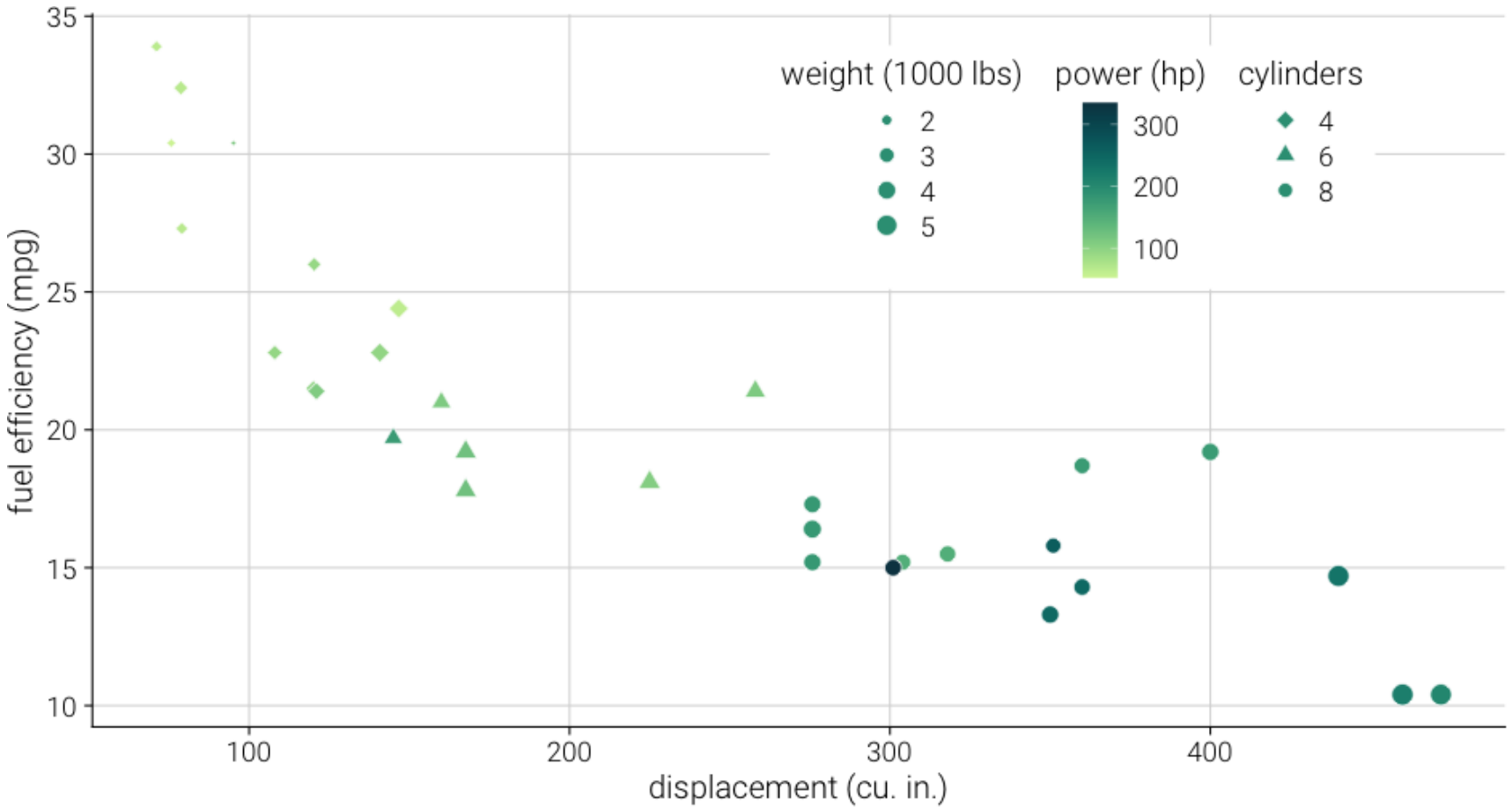
Alternative representation



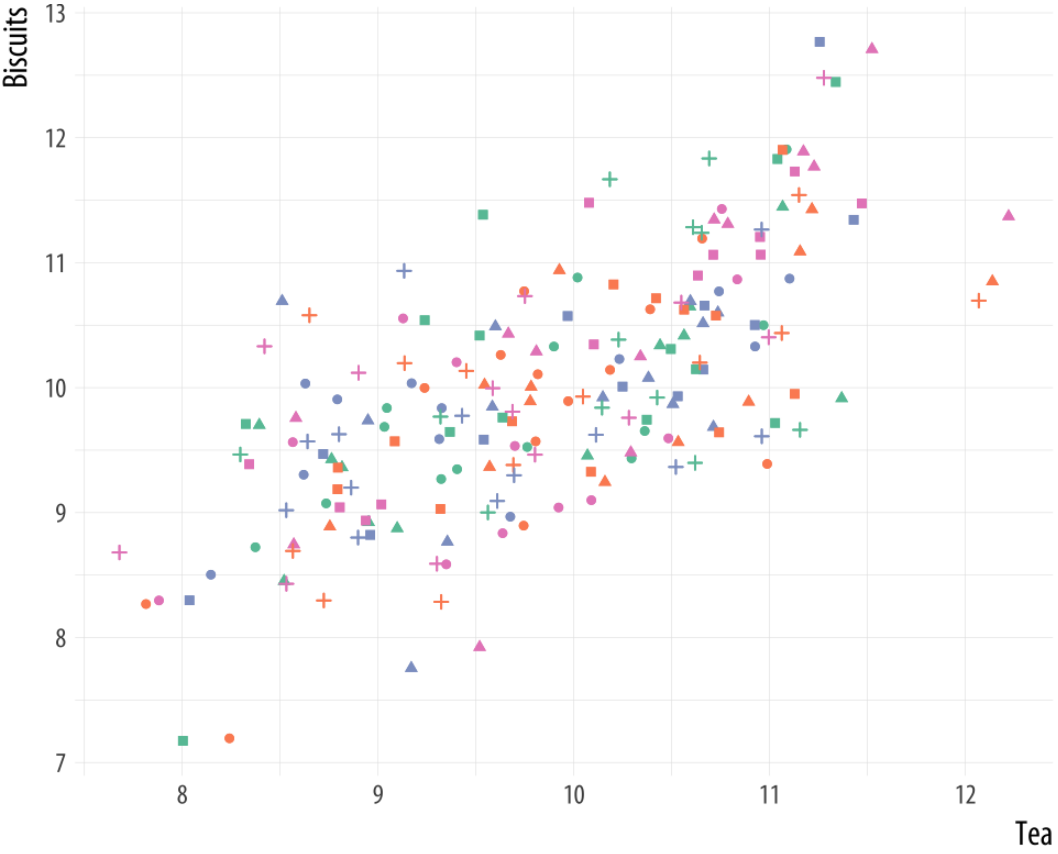
Comparison

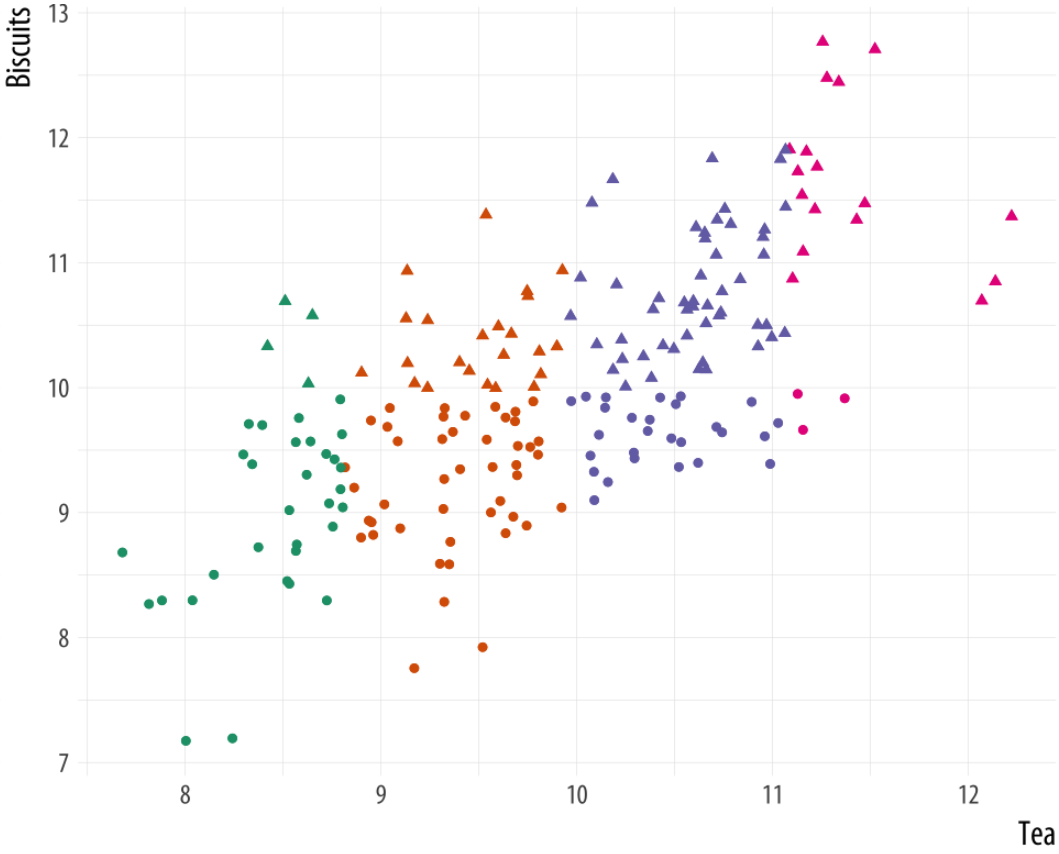
- Both represent three scales
    - Two position scales (x/y axis)
    - One color scale (categorical for the first, continuous for the second)

More scales are possible



Additional scales 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

Some examples

Okabe Ito

ColorBrewer Dark2

ggplot2 hue

See more about the Okabe Ito palette origins here

Sequential scale examples

ColorBrewer Blues

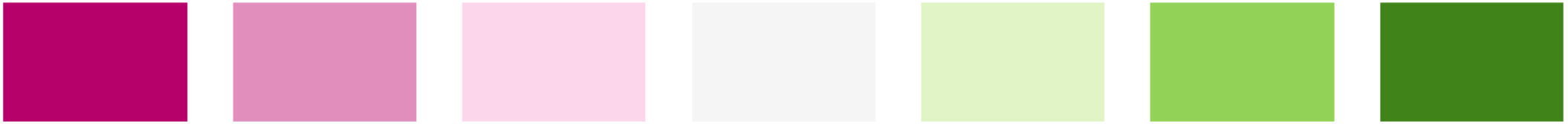
Heat

Viridis

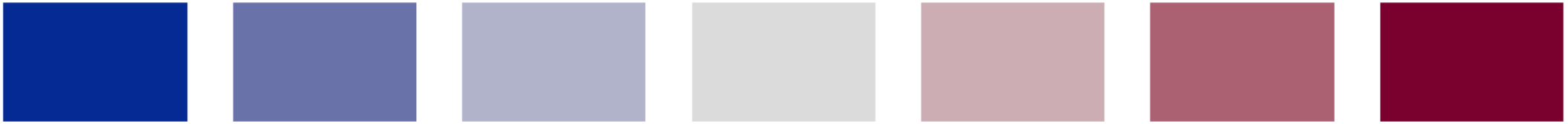
Diverging palettes

CARTO Earth

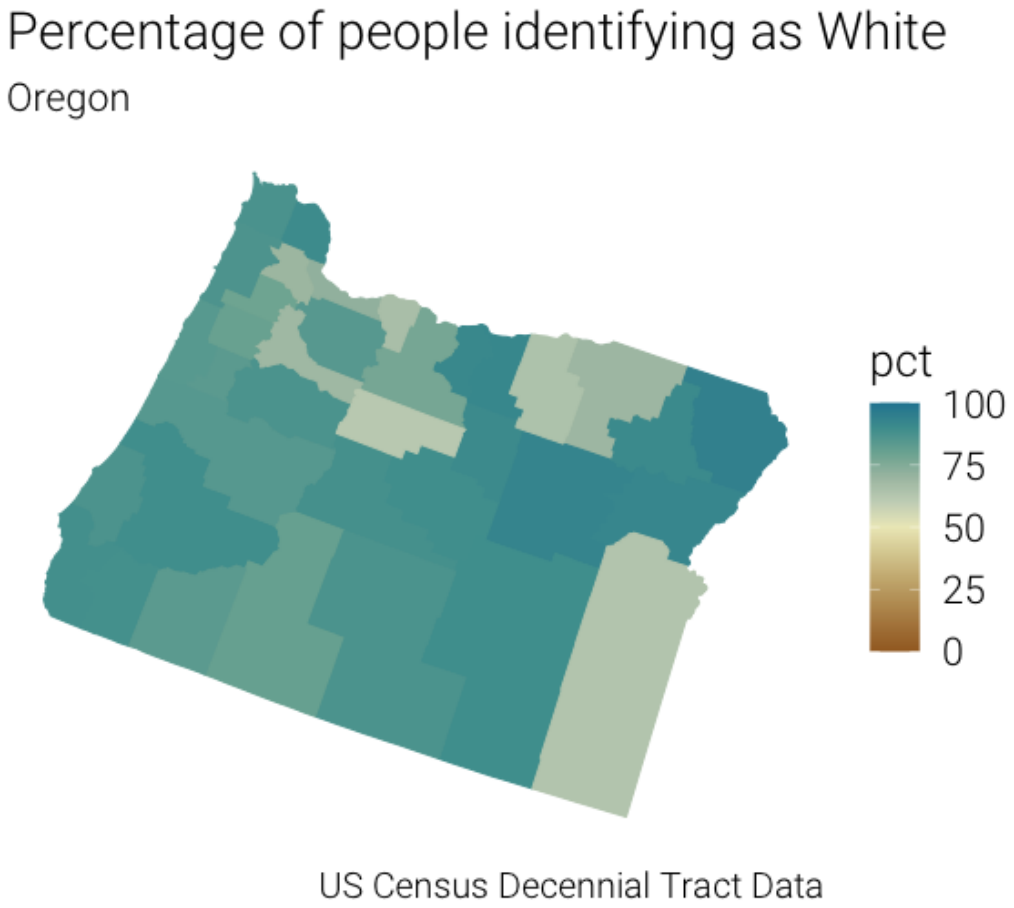
ColorBrewer PiYG



Blue-Red

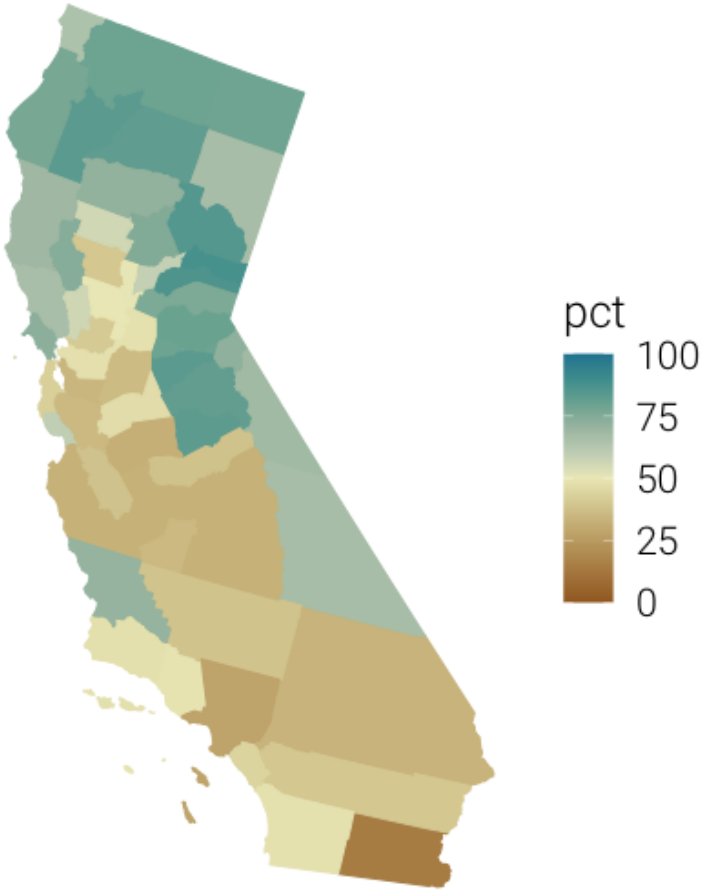


Earth palette



Percentage of people identifying as White

California

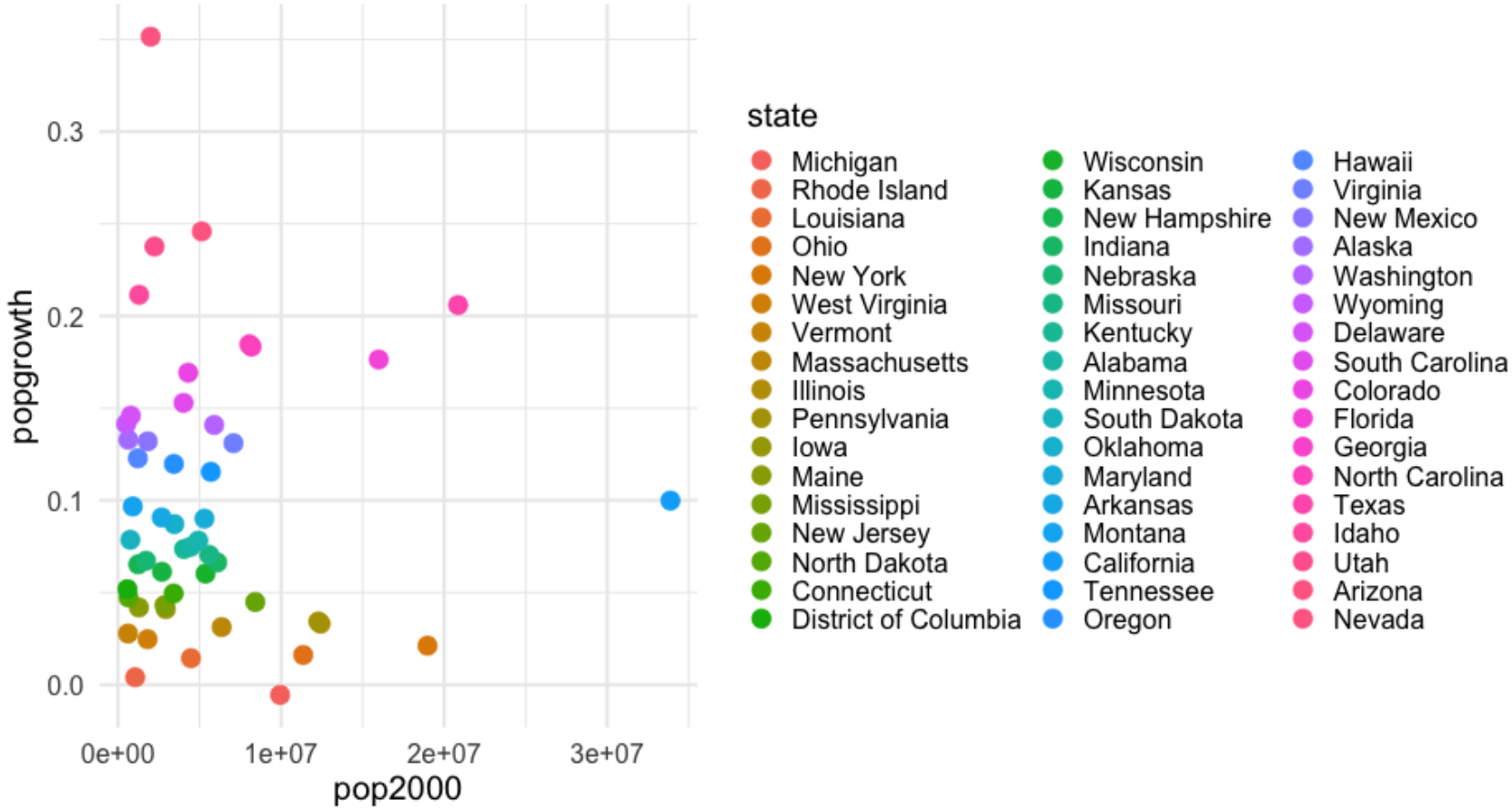


US Census Decennial County Data

Common problems

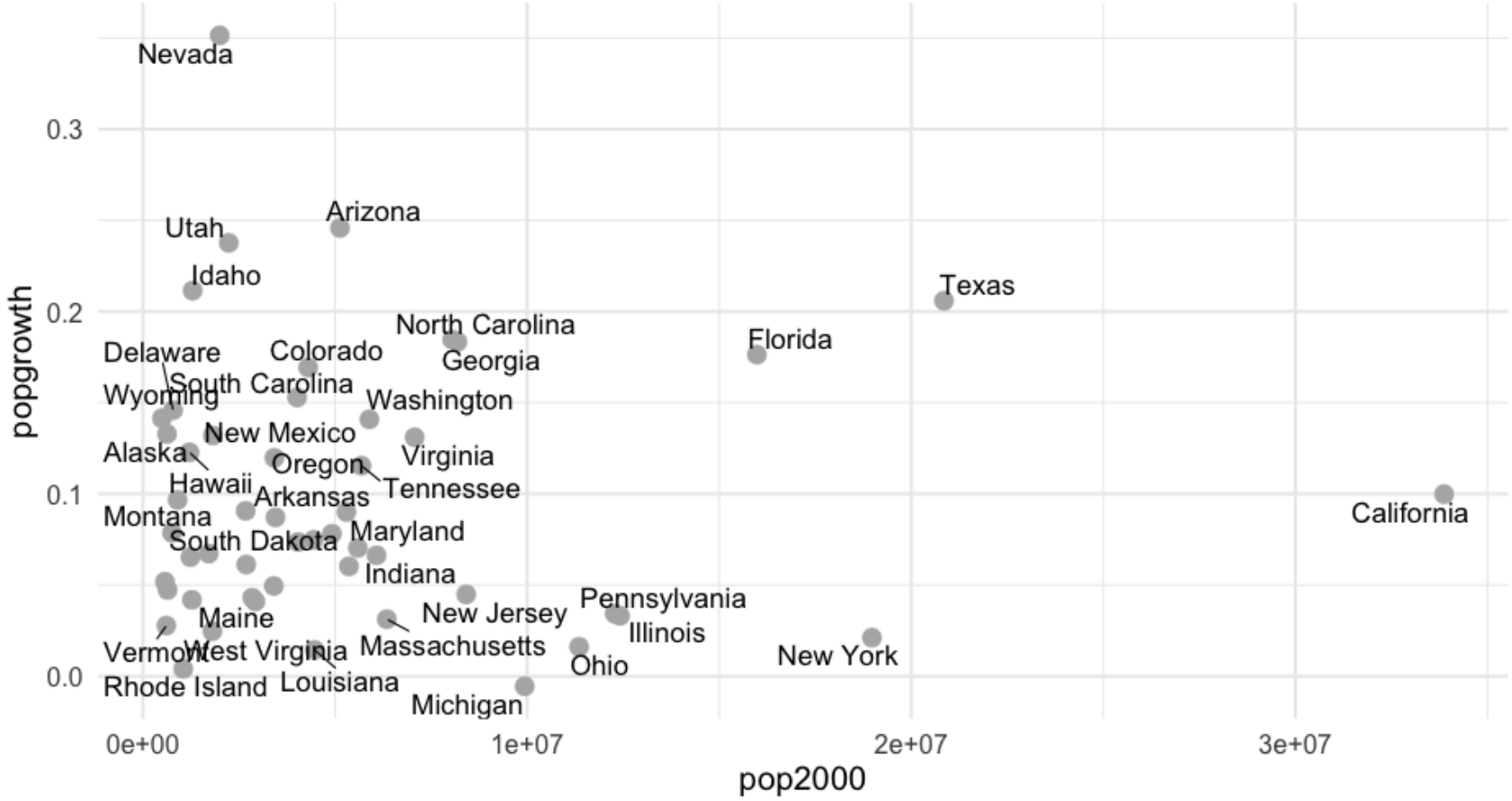
Too many colors

More than 5–ish generally becomes difficult to track



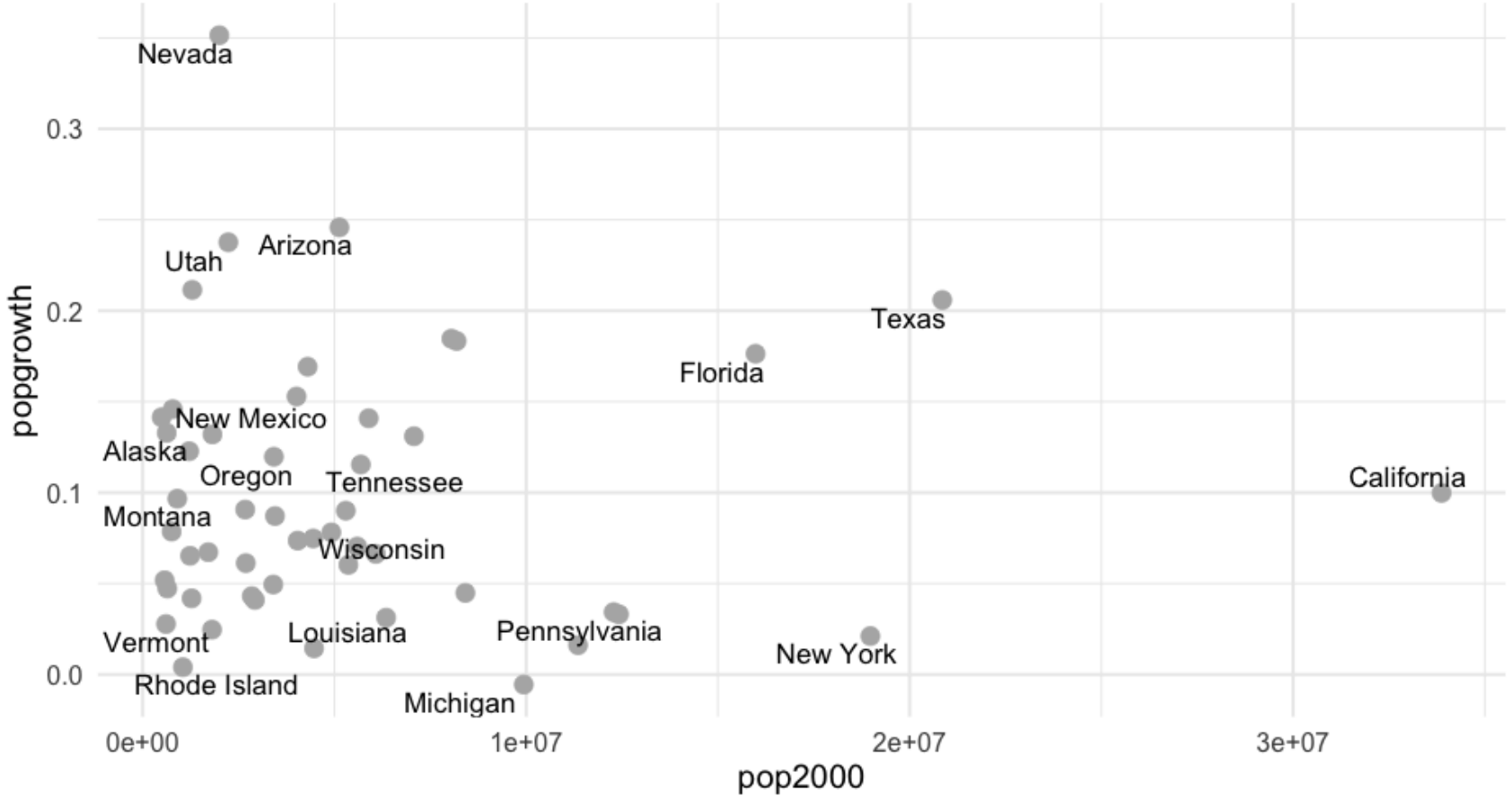
Use labels

still too many...



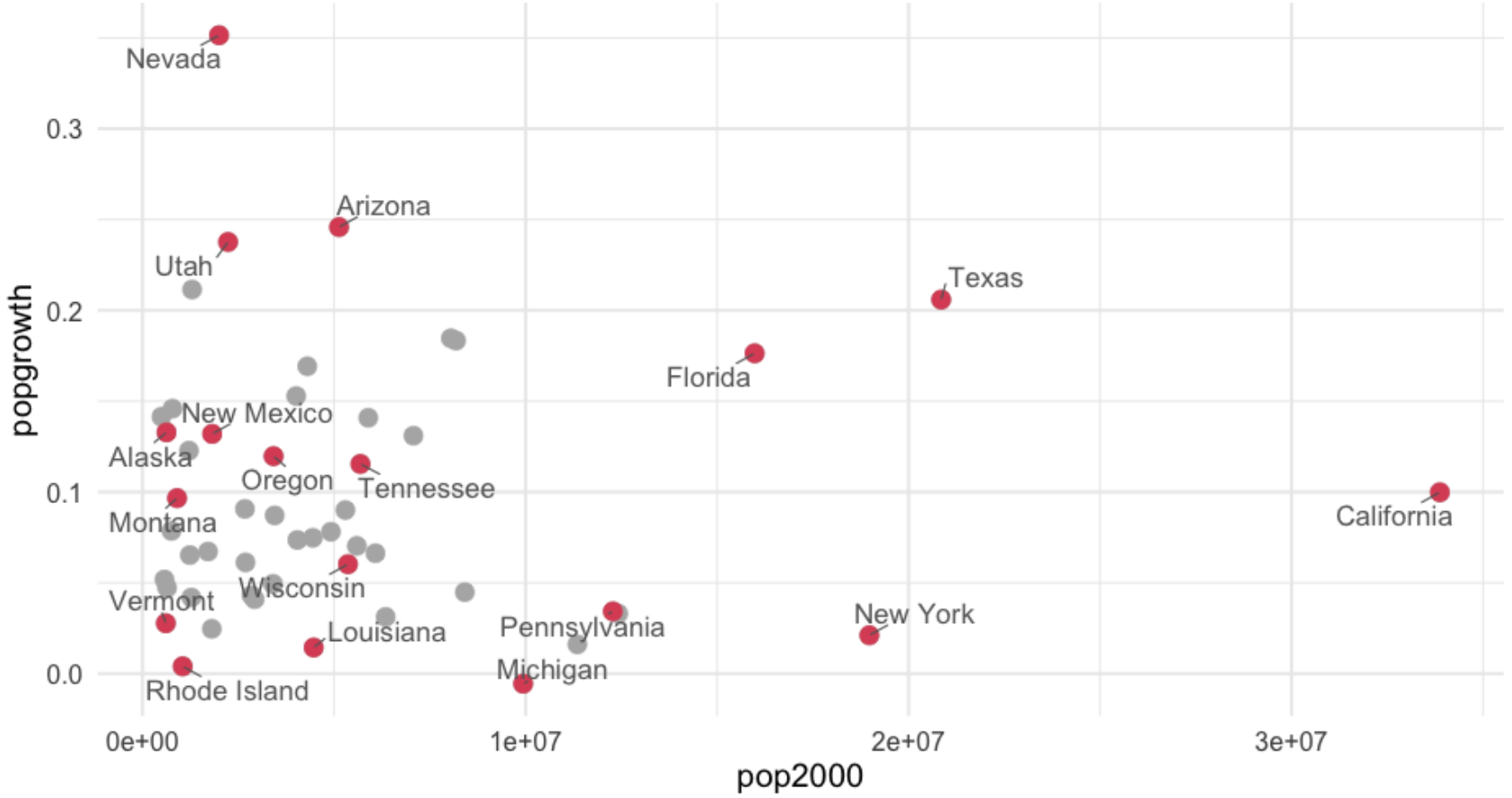
Better

Get a subset



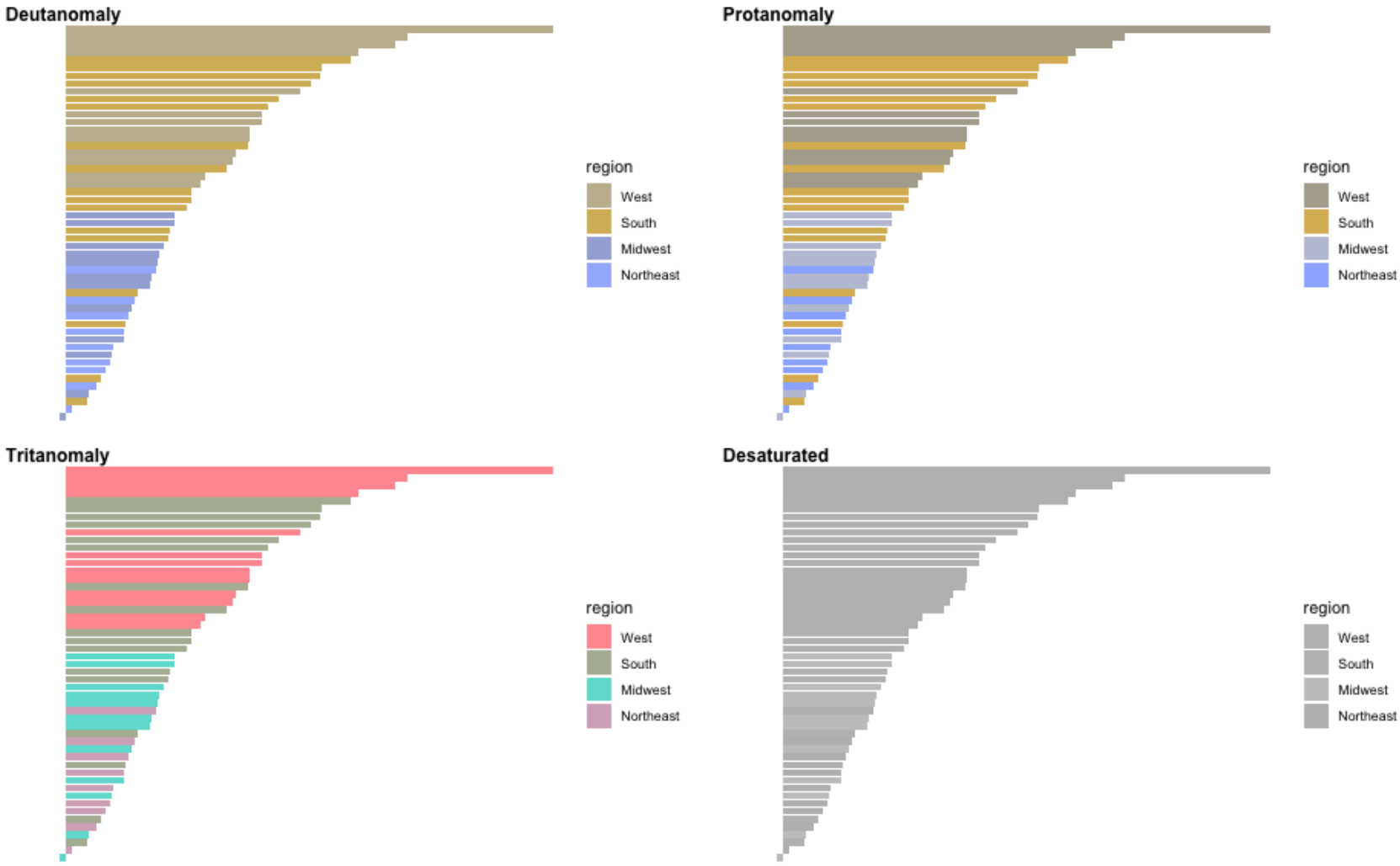
Best

(but could still be improved)

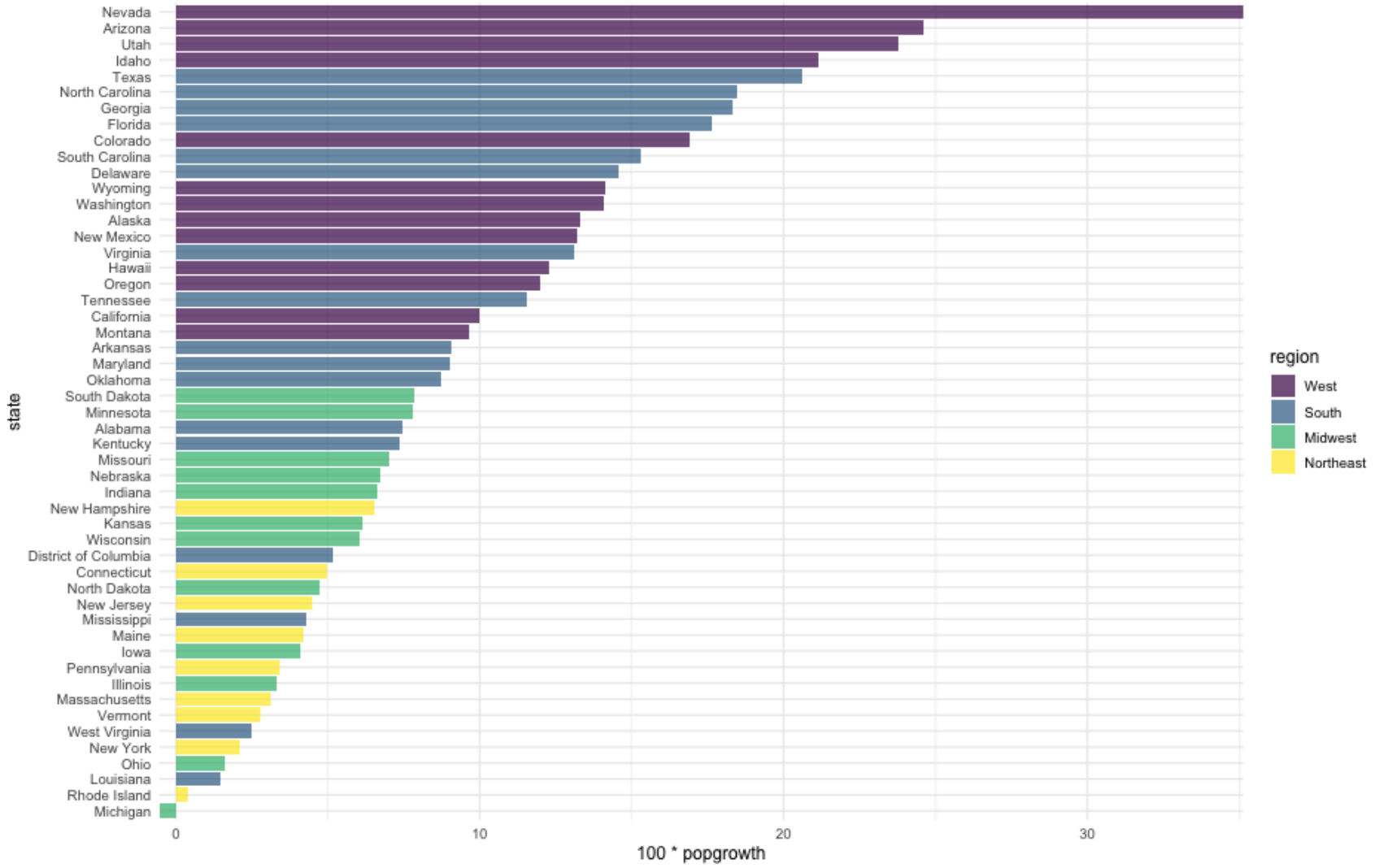


Problem w/ default palette

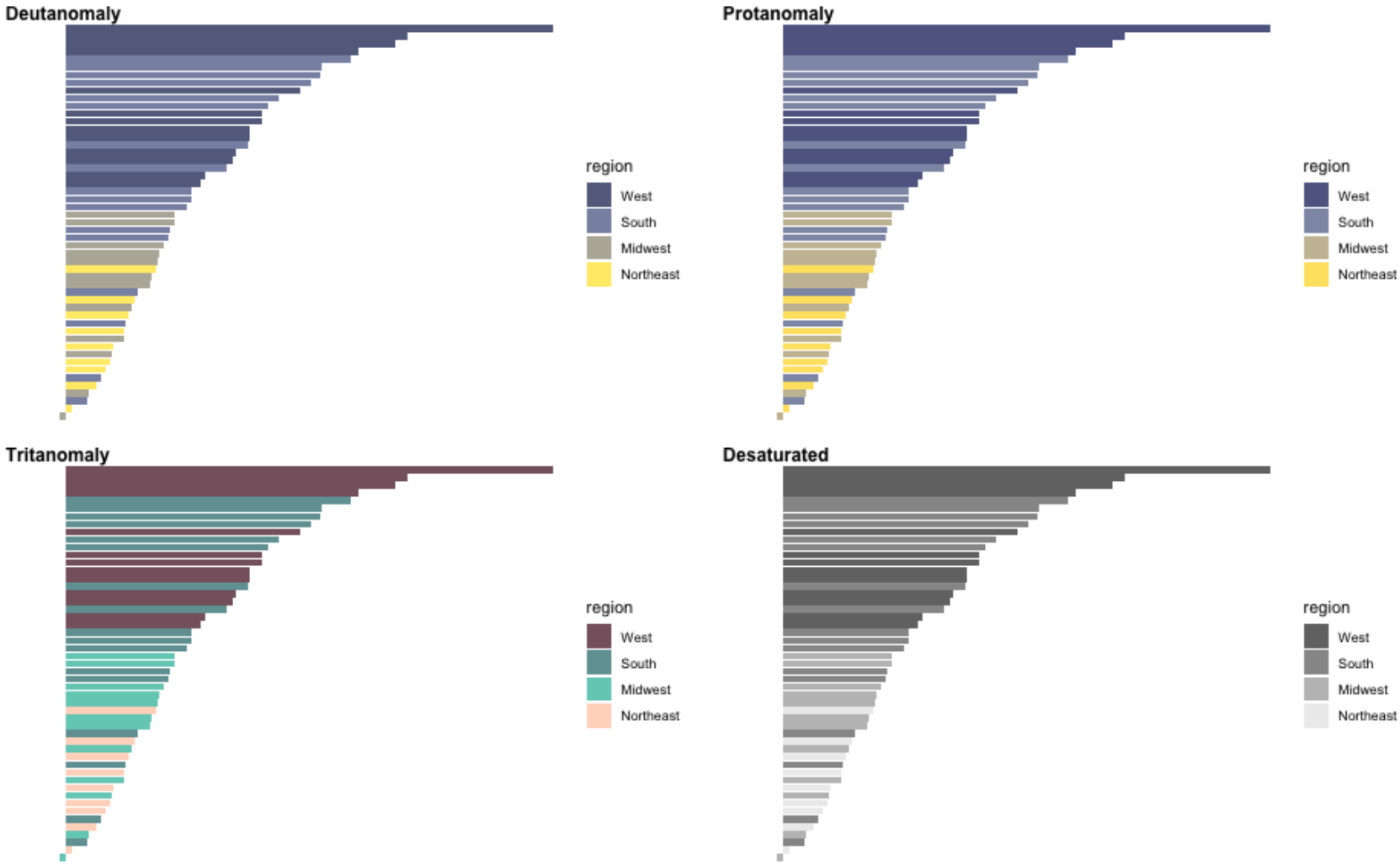
For {ggplot2}



Alternative: viridis



Revised version



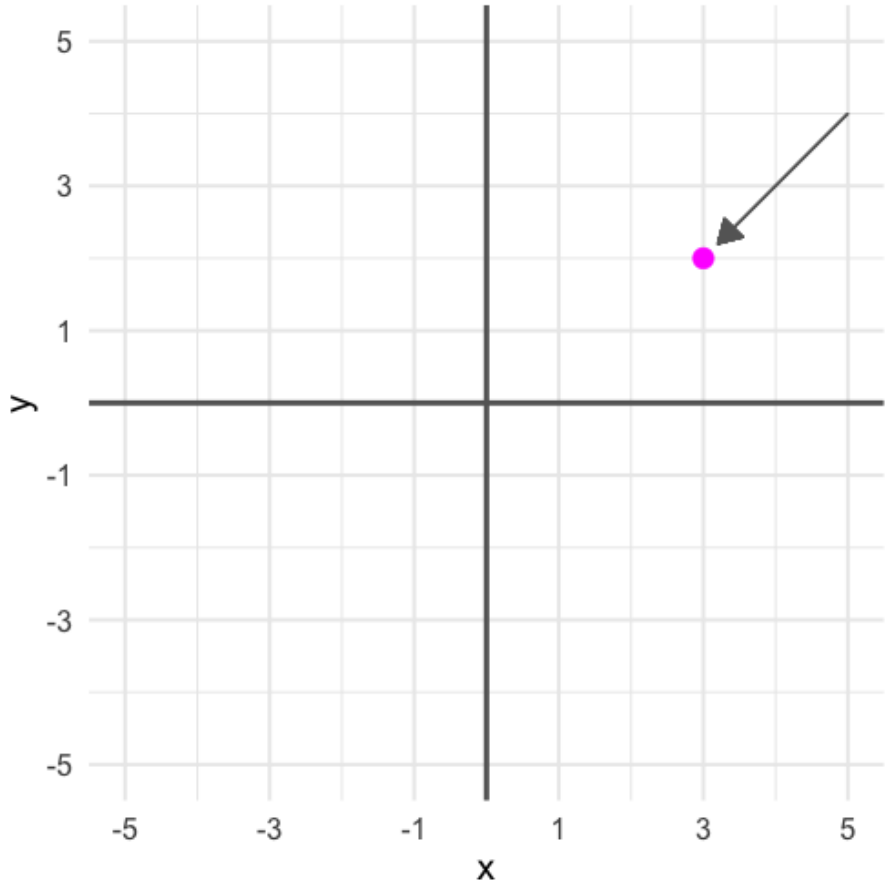
Last few note on palettes

- Do some research, find what you like and what tends to work well
- Check for colorblindness
- Look into http://colorbrewer2.org/

Visualizing Uncertainty

The primary problem

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

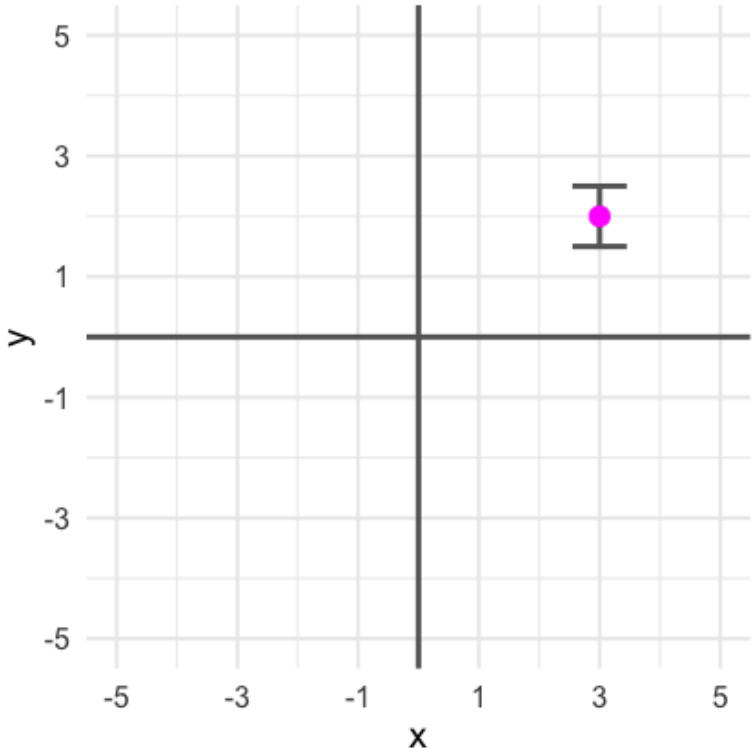


Some secondary problem

- 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



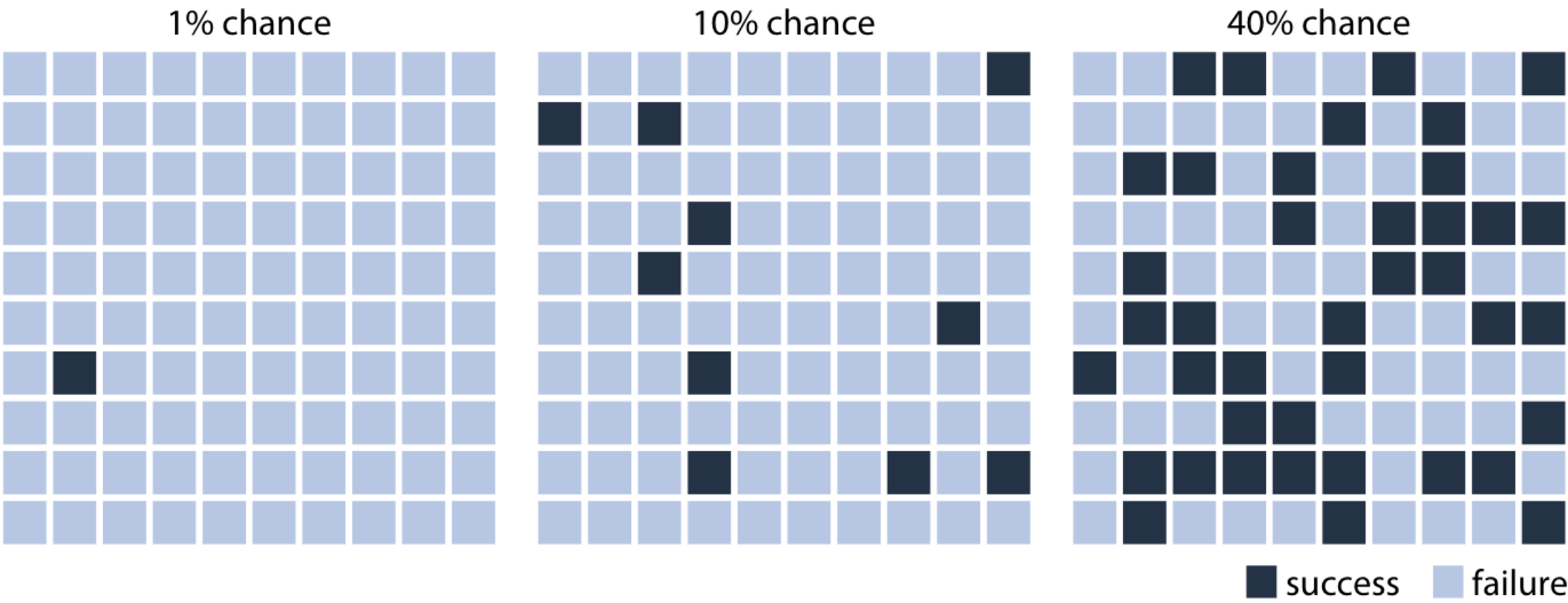
Thinking about uncertainty

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

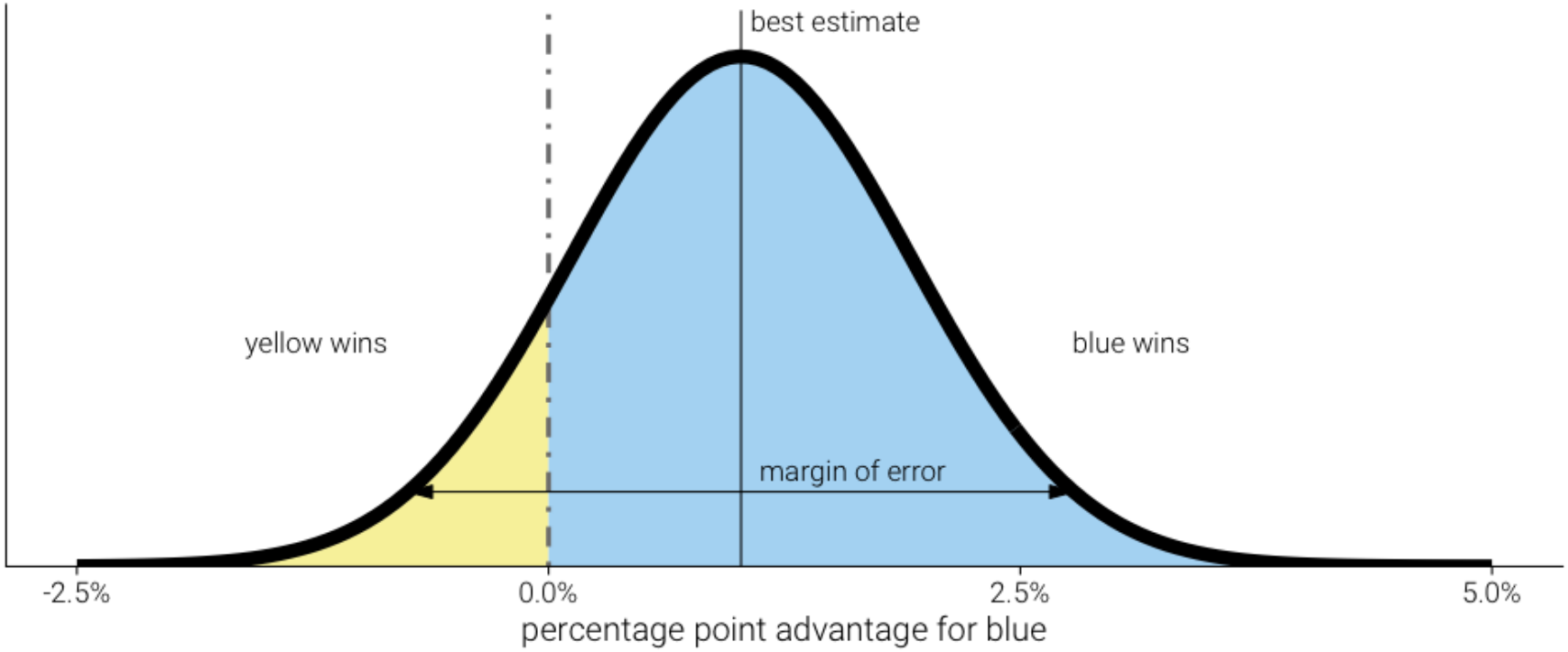


Framing a single uncertainty

Non–discrete probabilities

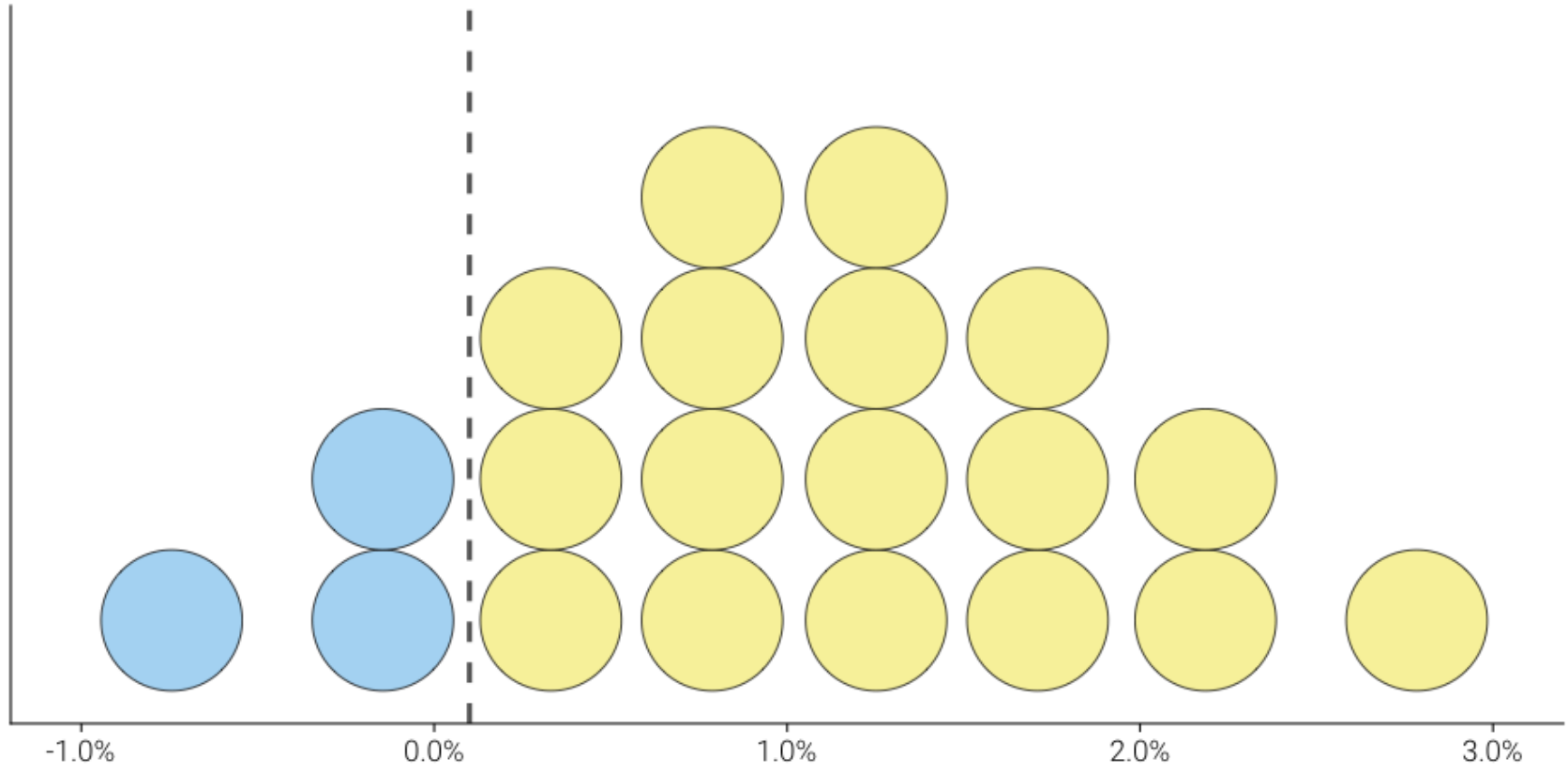
Blue party has 1% advantage w/ margin of error of 1.76 points

Who will win?



Code taken from Wilke

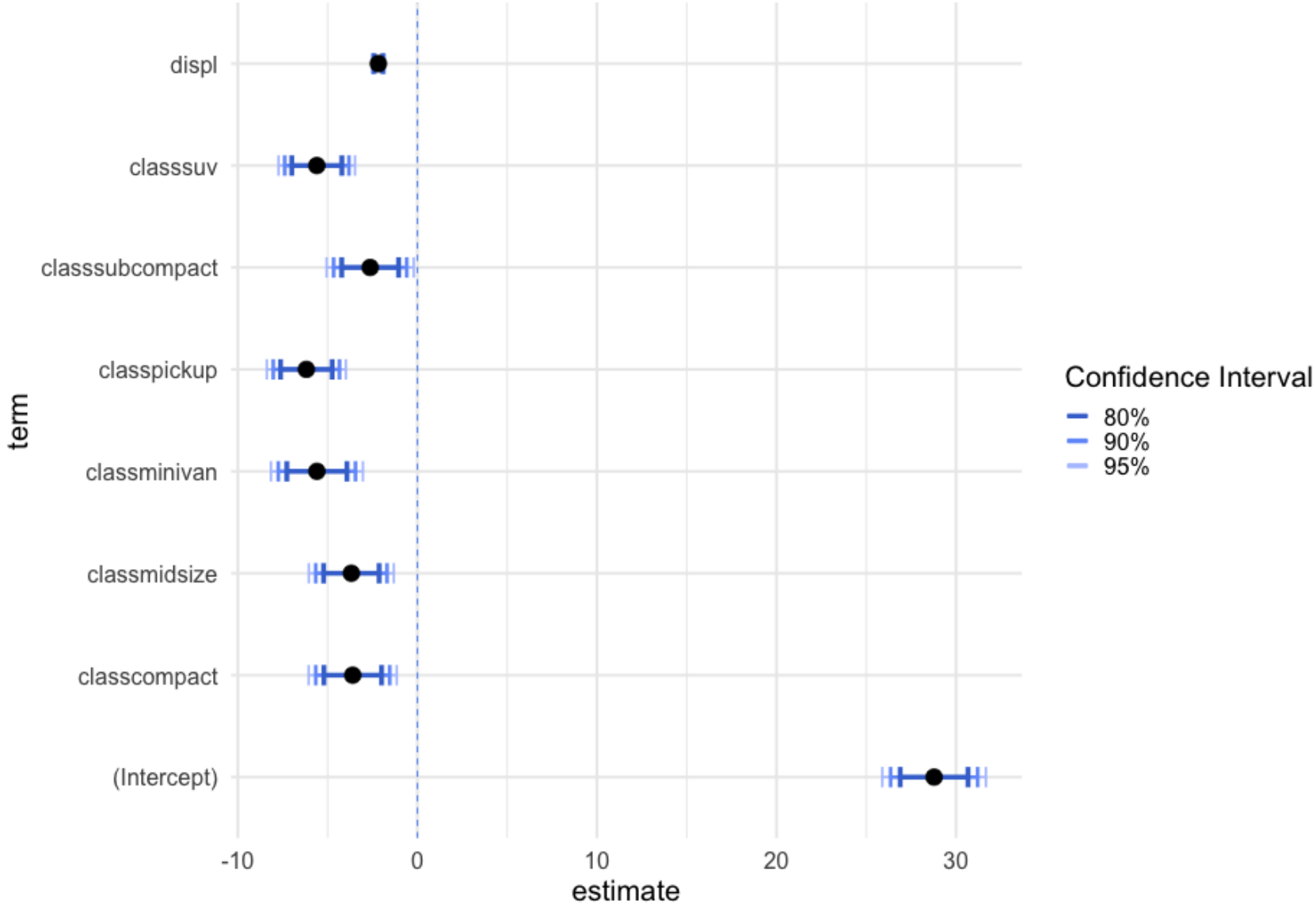
Descretized version



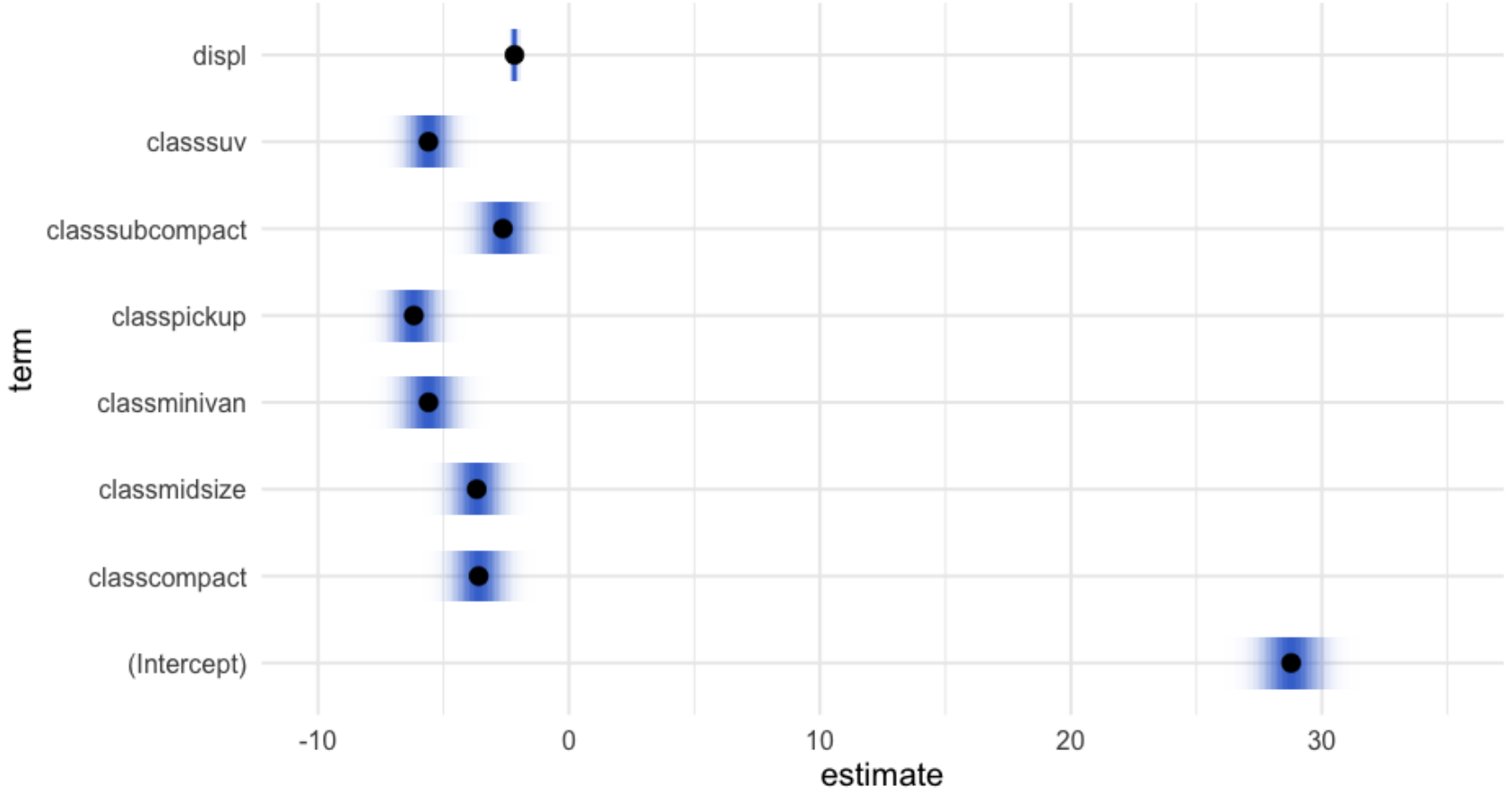
Point estimates

A few alternatives to error bars

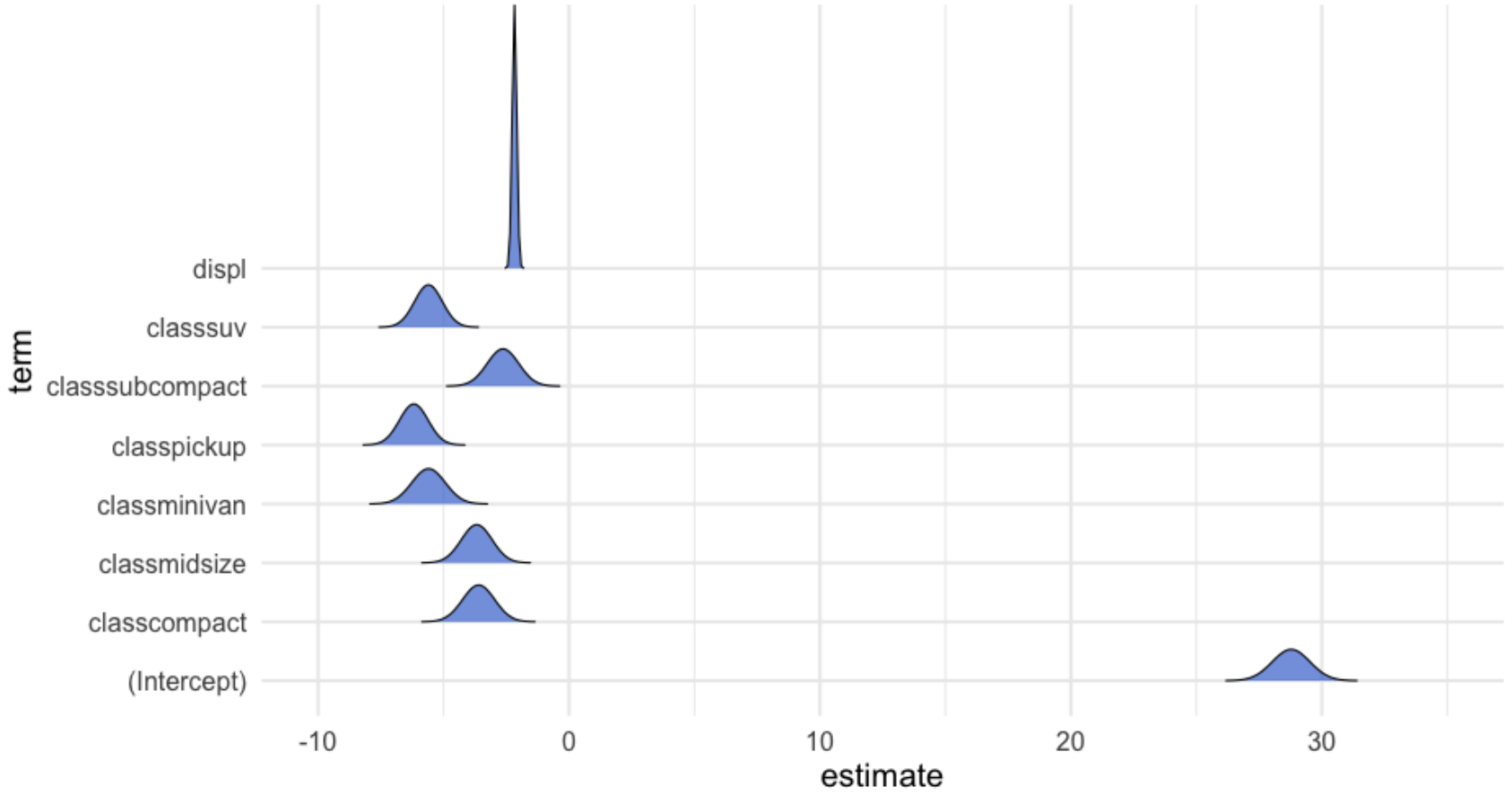
Multiple error bars



Density stripes



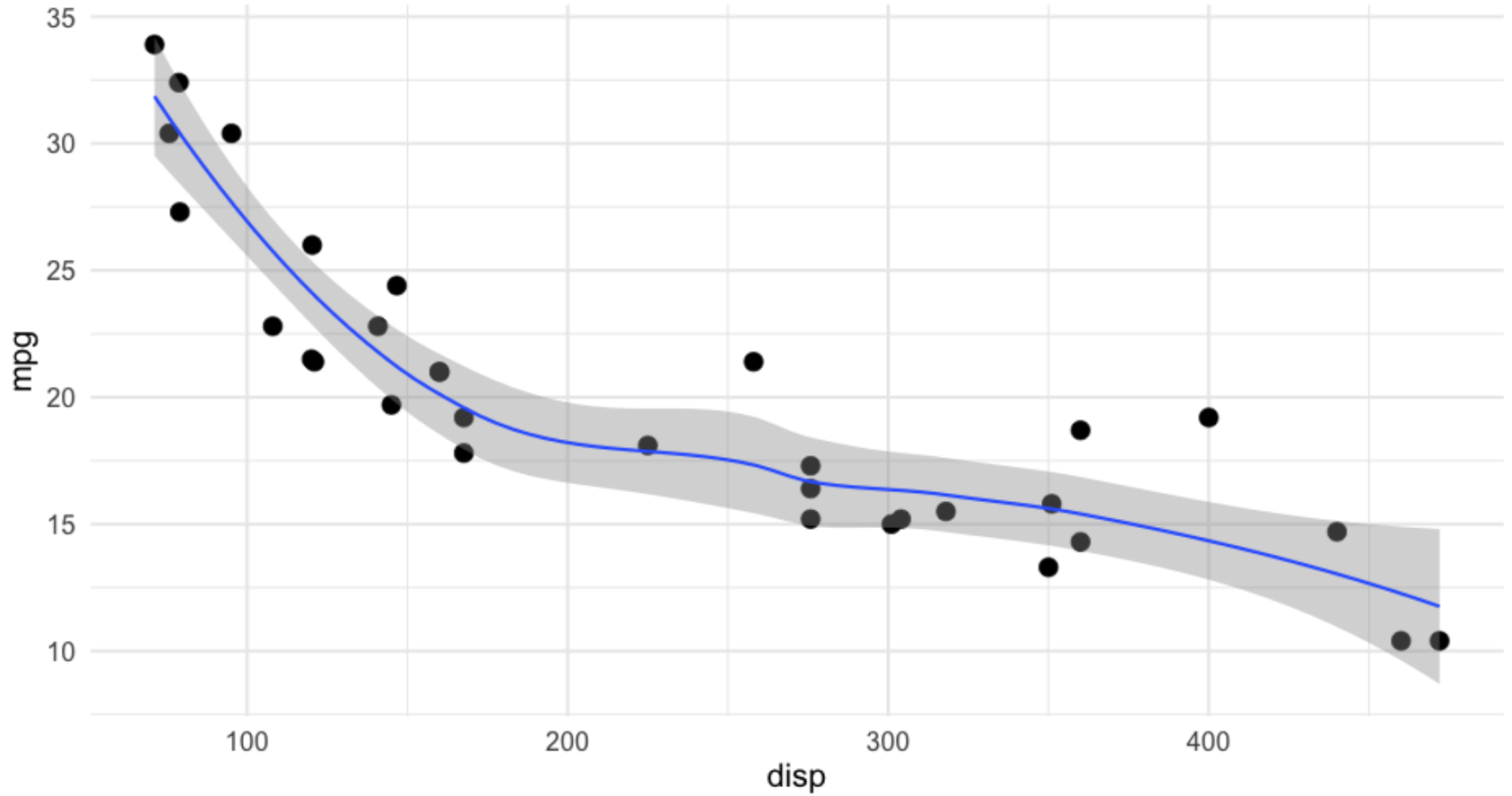
Actual densities



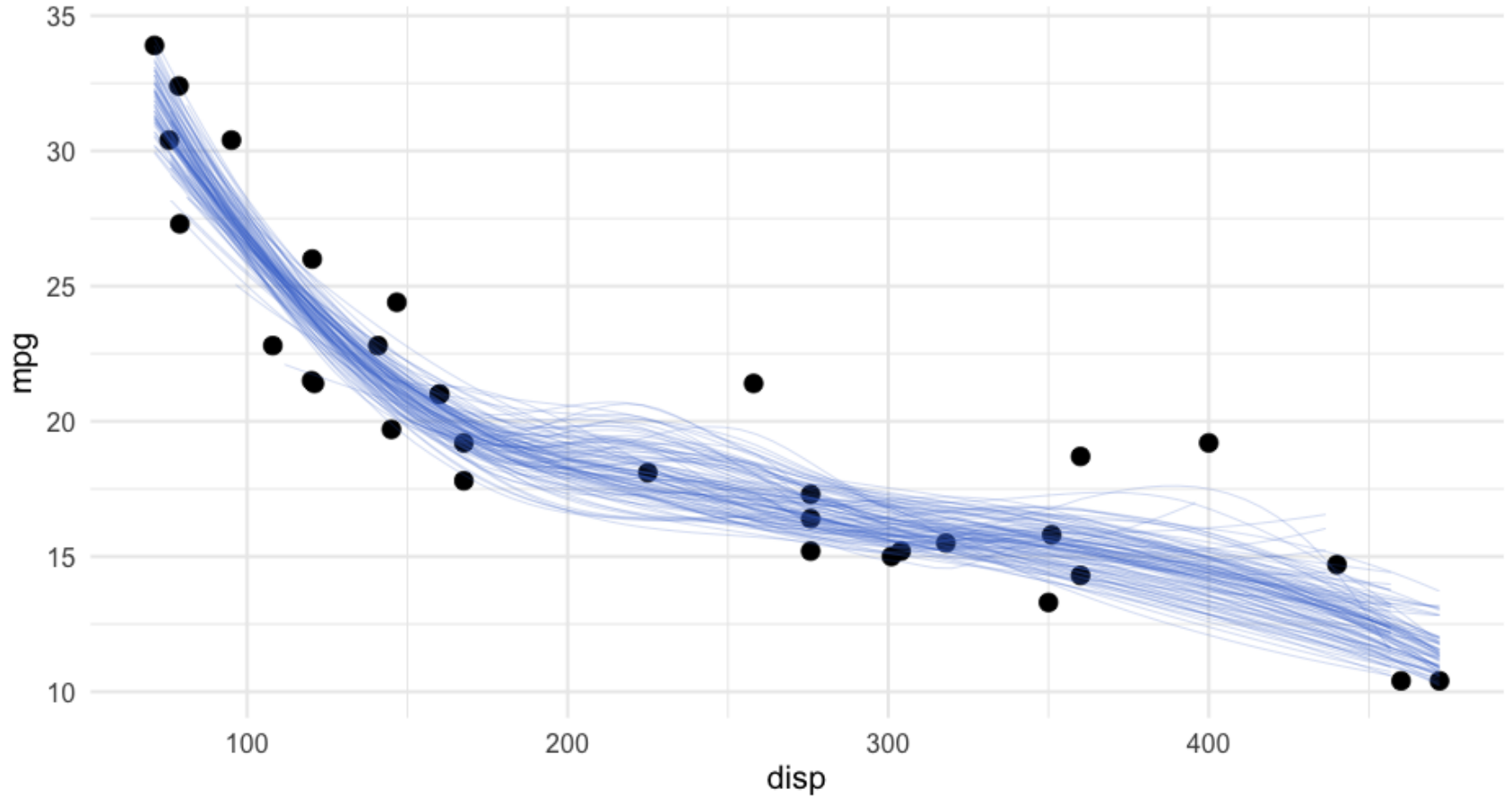
HOPs

Hypothetical Outcome Plots (and related plots)

Standard regression plot

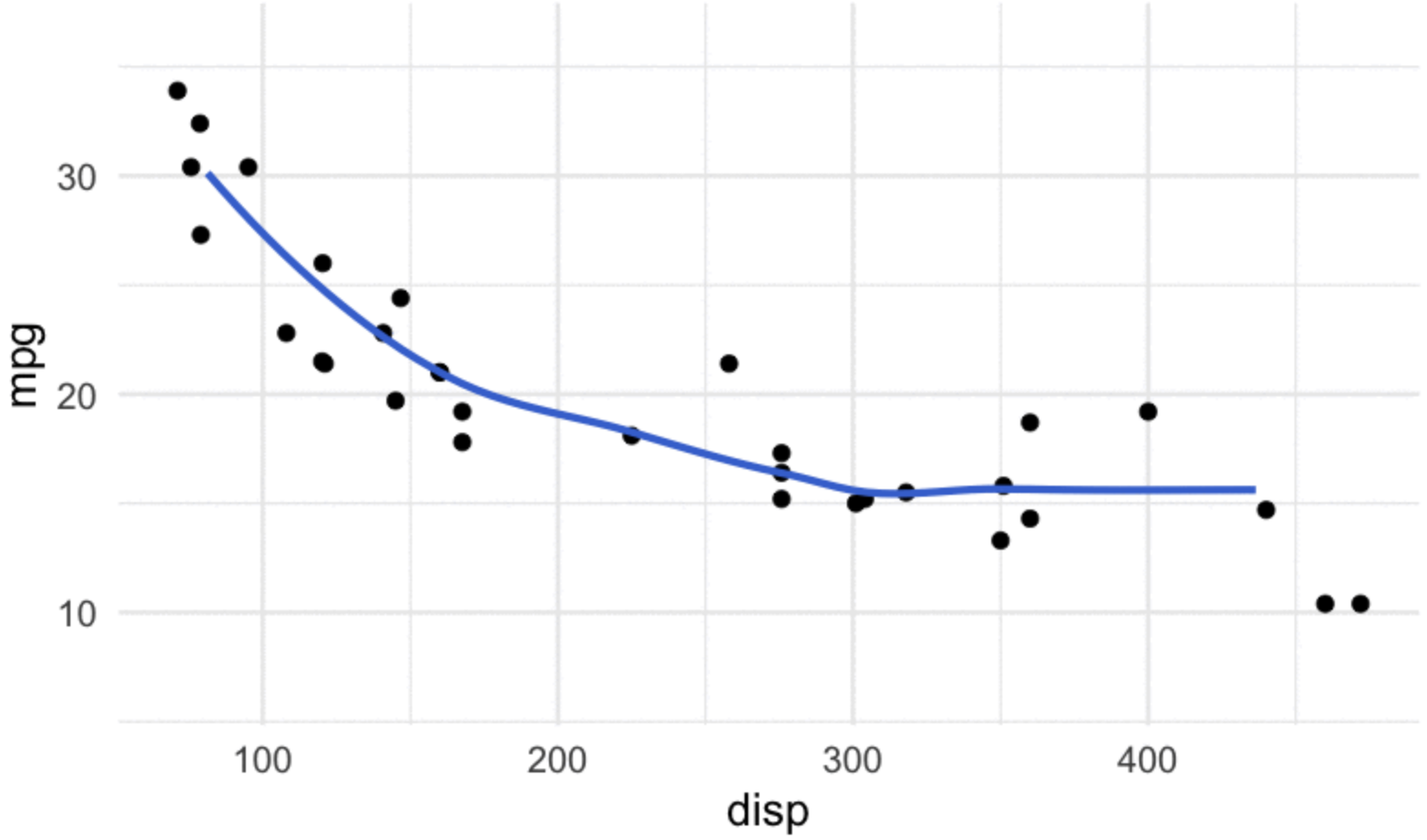


Alternative

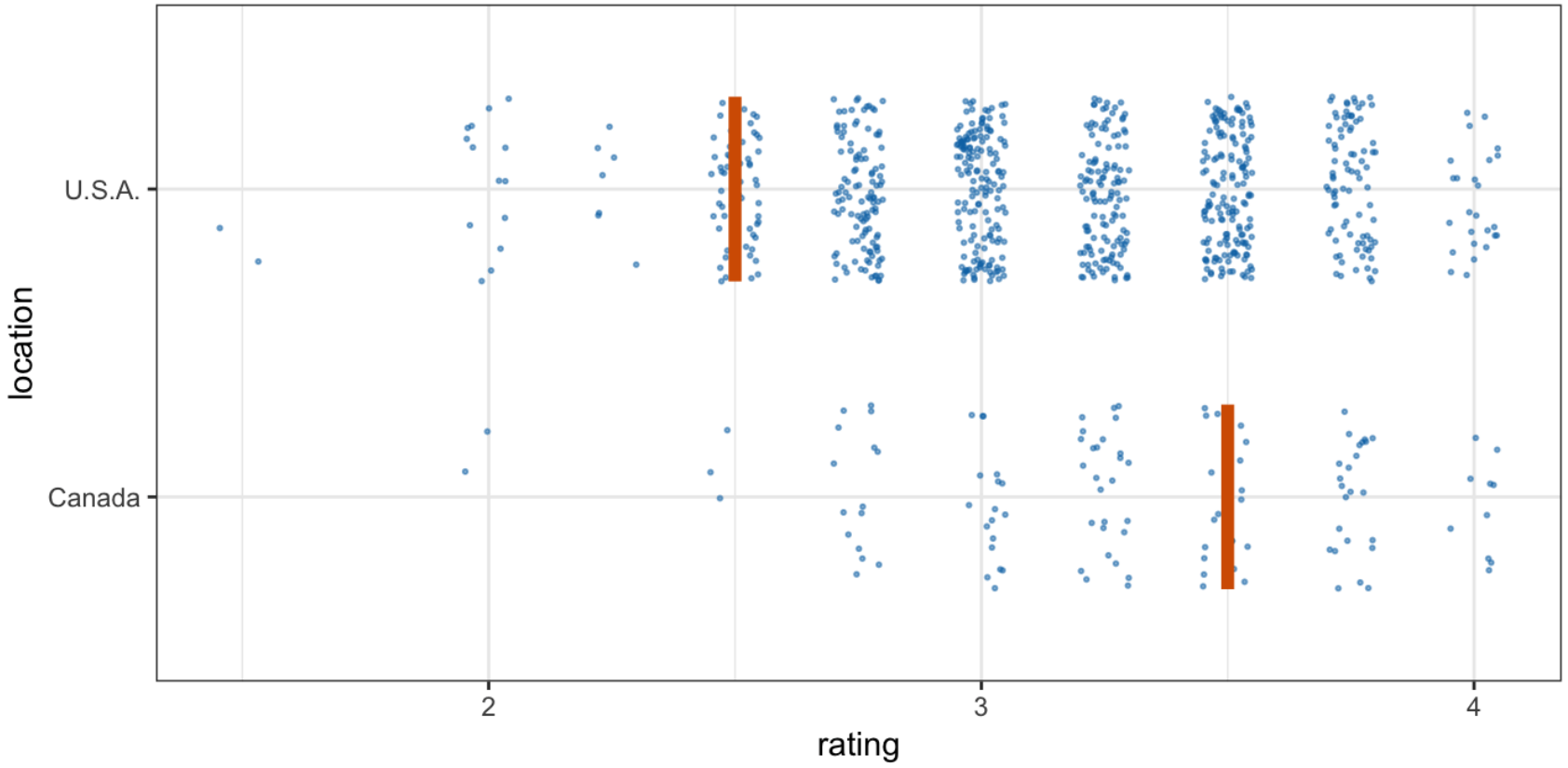


HOPs

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



Another example



Last few note 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

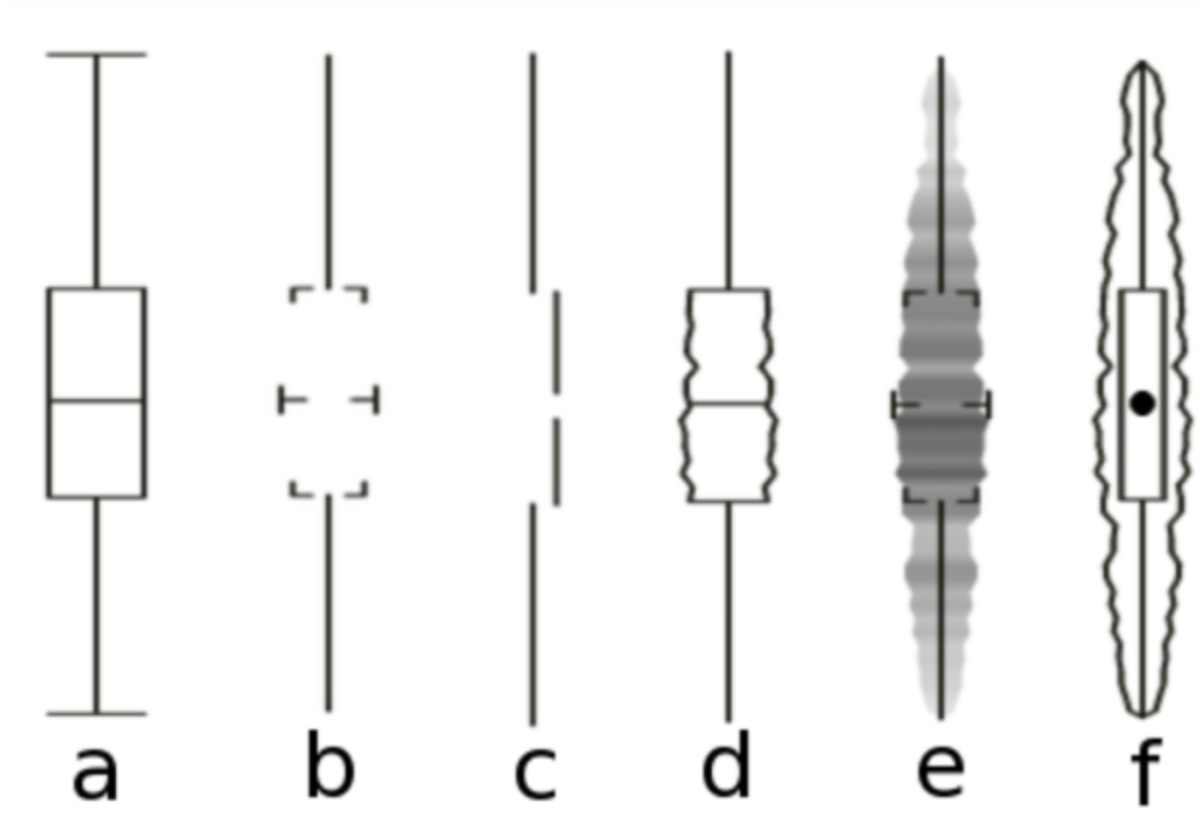
What is it?

> 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!

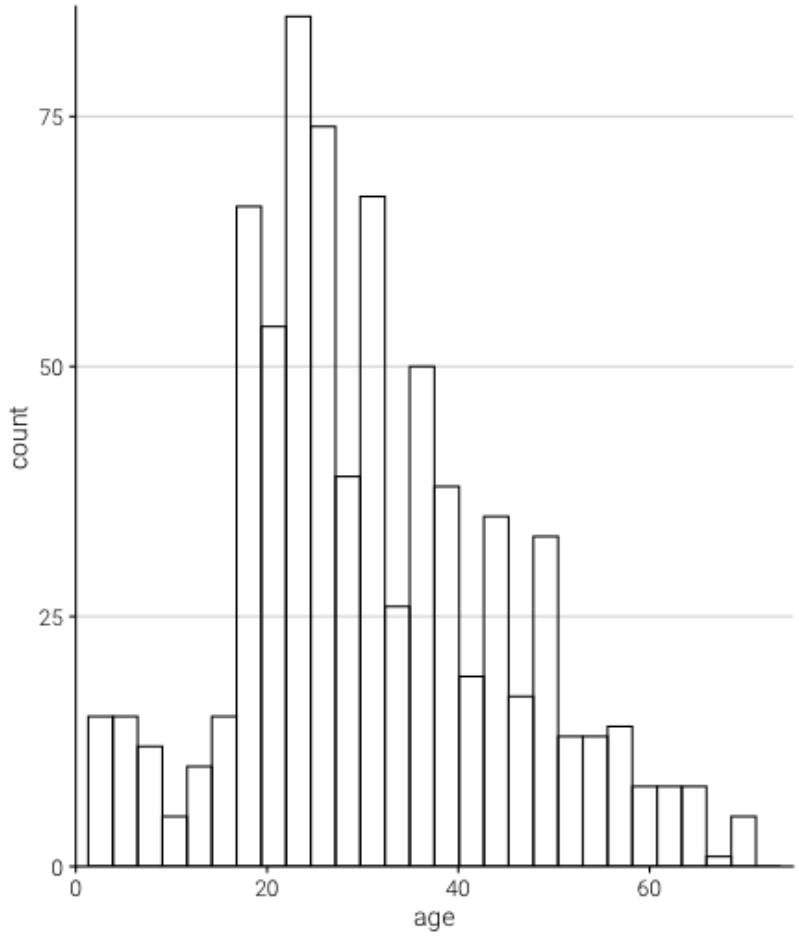

NOT SO FAST, MY FRIEND!

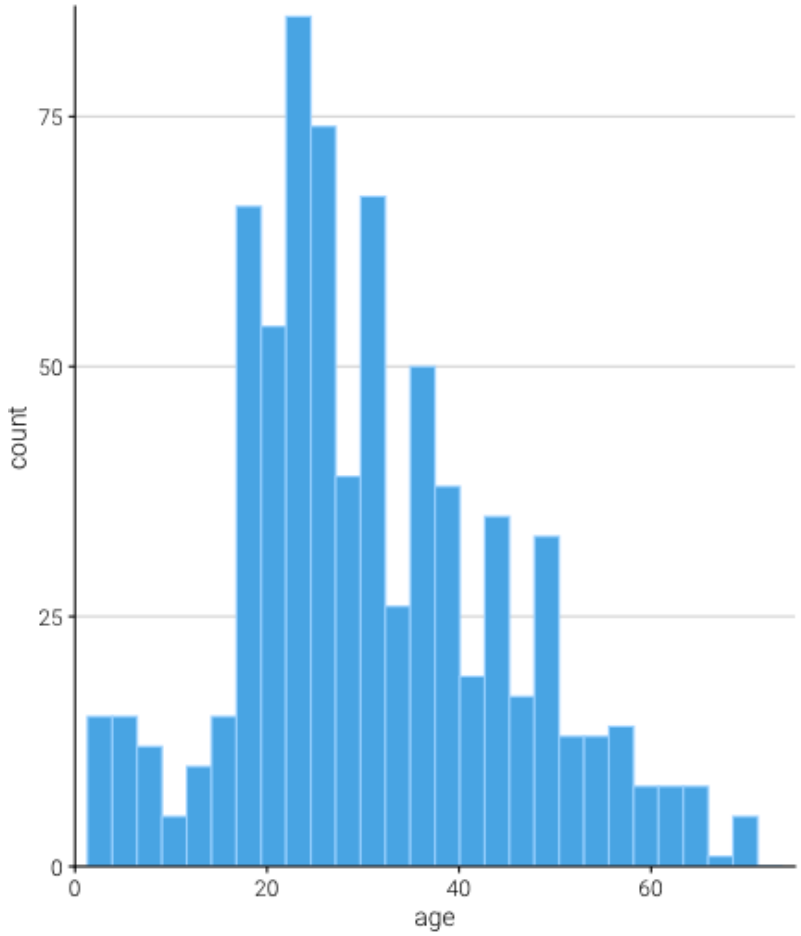
Minimize cognitive load

- 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?



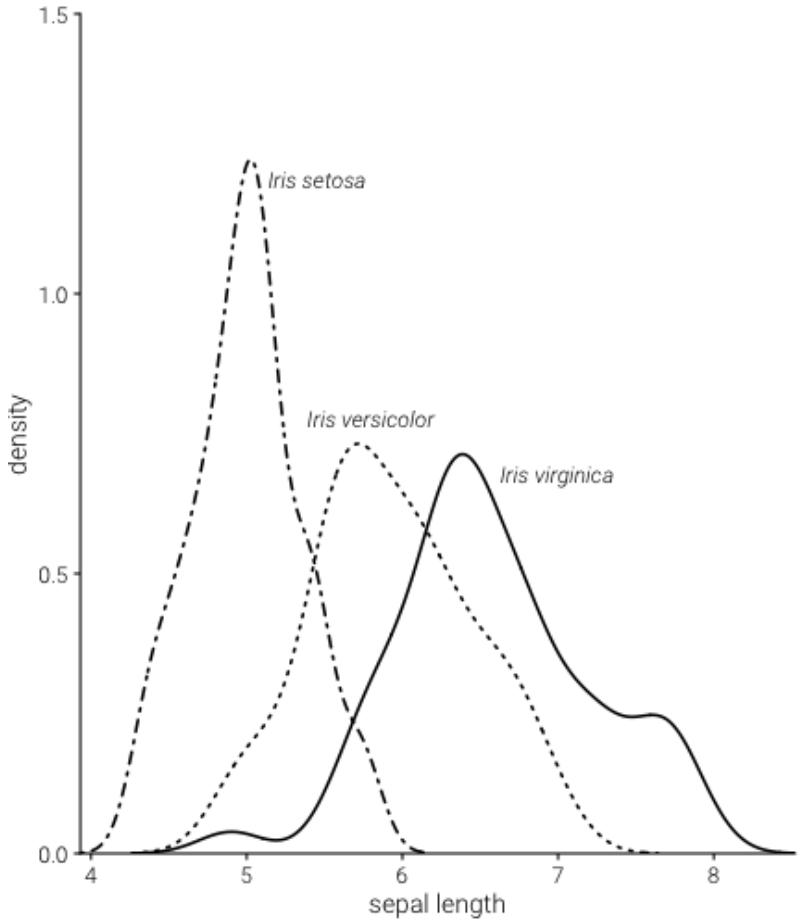


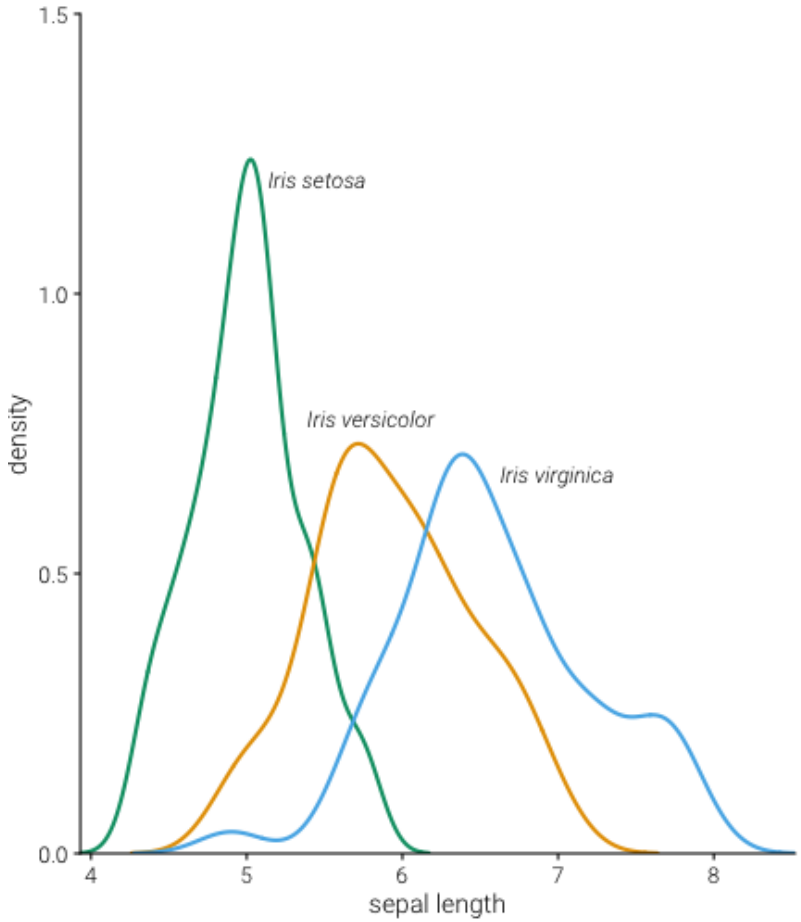
Advice from Wilke

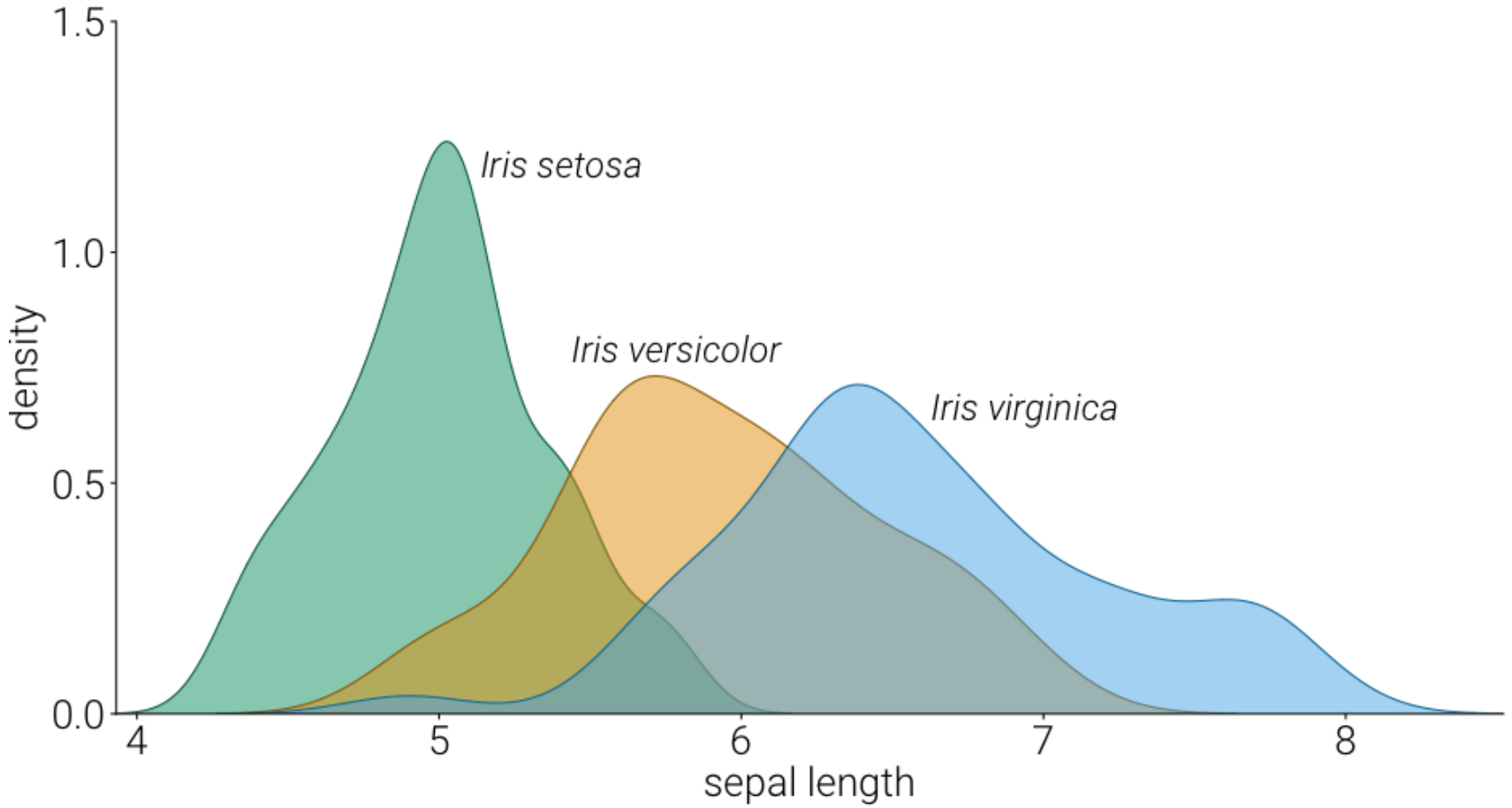
> Whenever possible, **visualize your data with solid, colored shapes** rather than with lines that outline those shapes. Solid shapes are more easily perceived, are less likely to create visual artifacts or optical illusions, and do more immediately convey amounts than do outlines.

emphasis added

Another example






1.5
1.0
0.5
0.0
density
Iris setosa
Iris versicolor
Iris virginica
4
5
6
7
8
sepal length

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

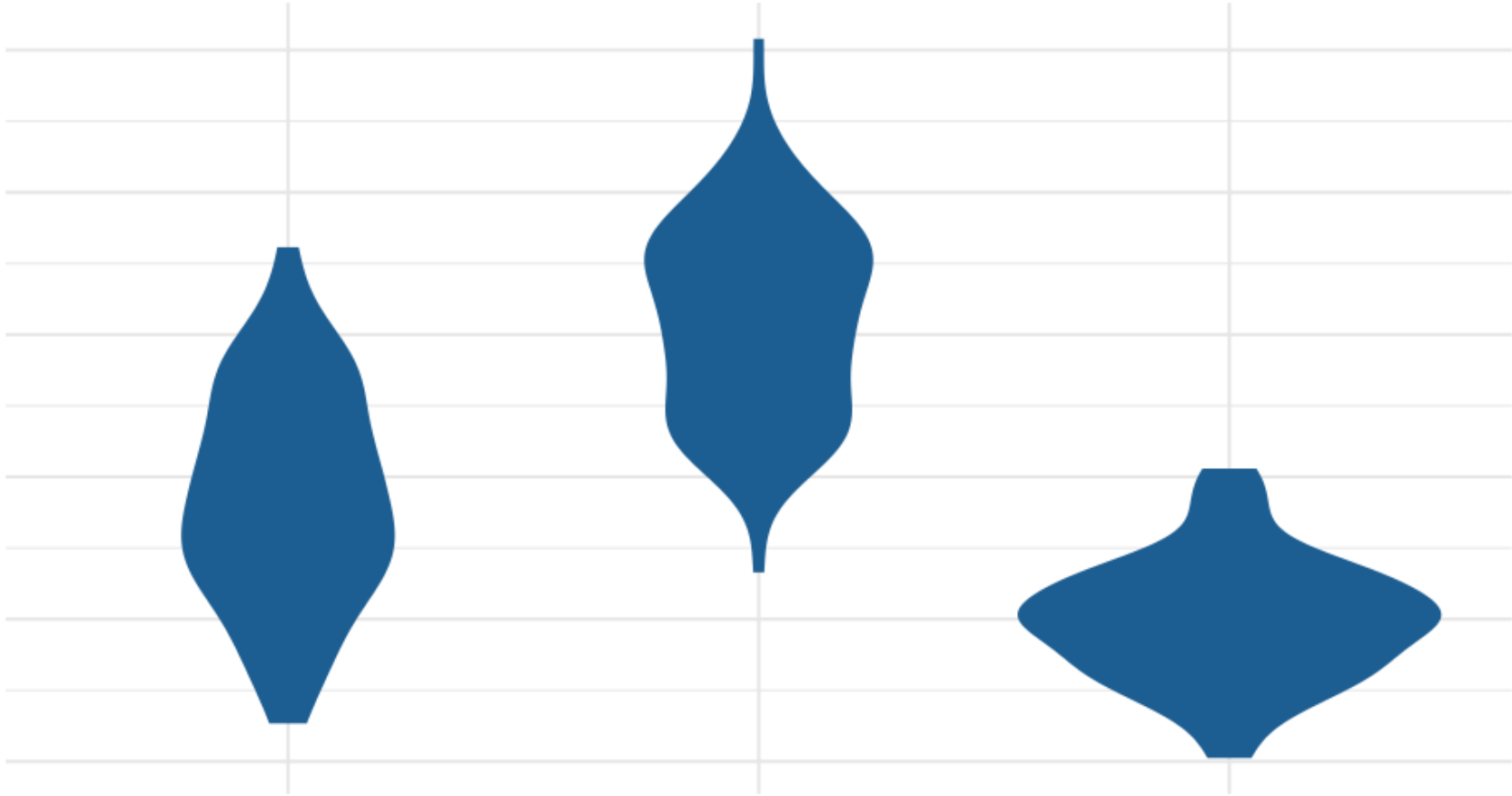
Grouped data

Distributions

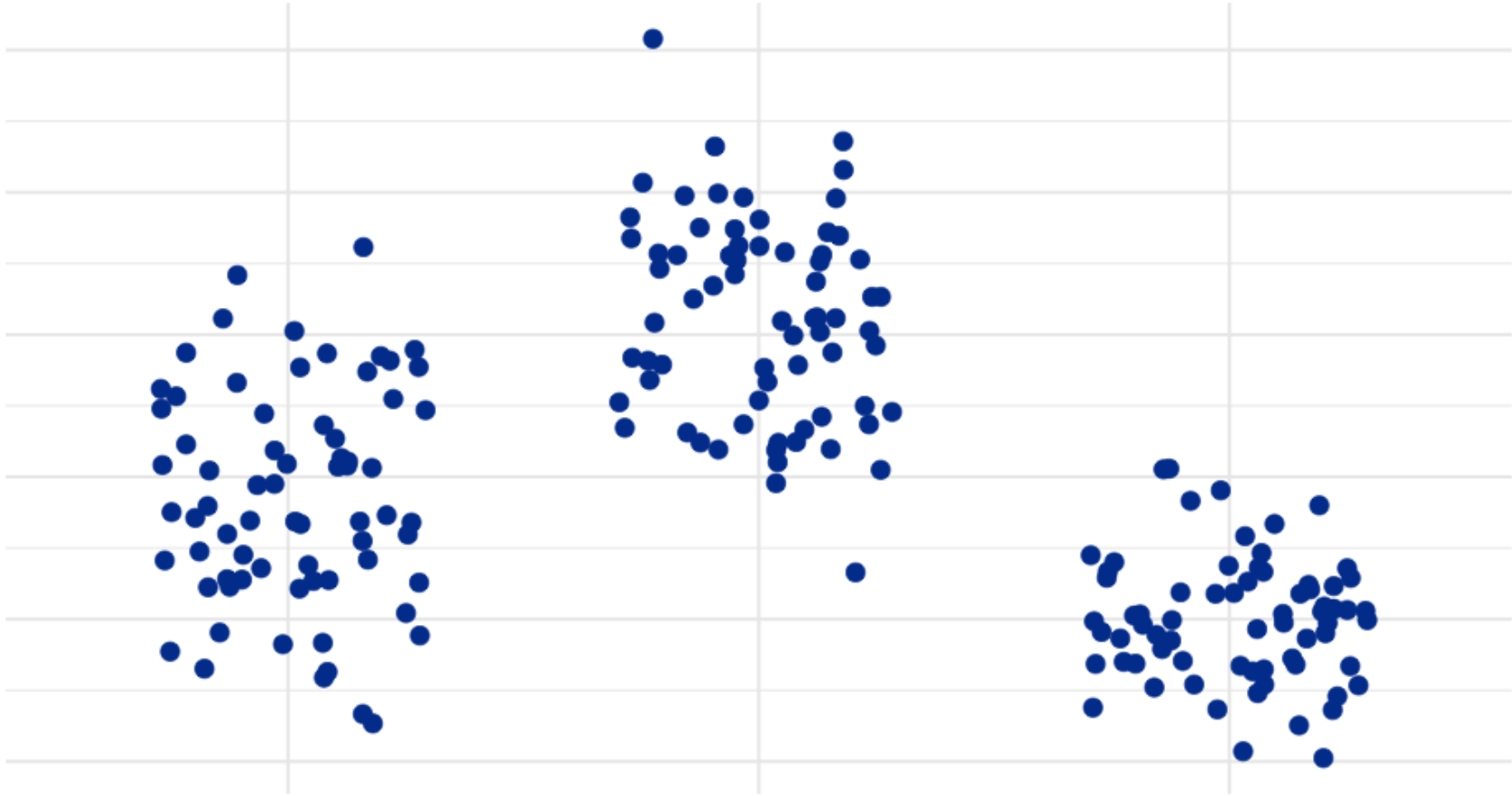
How do we display more than one distribution at a time?

Boxplots

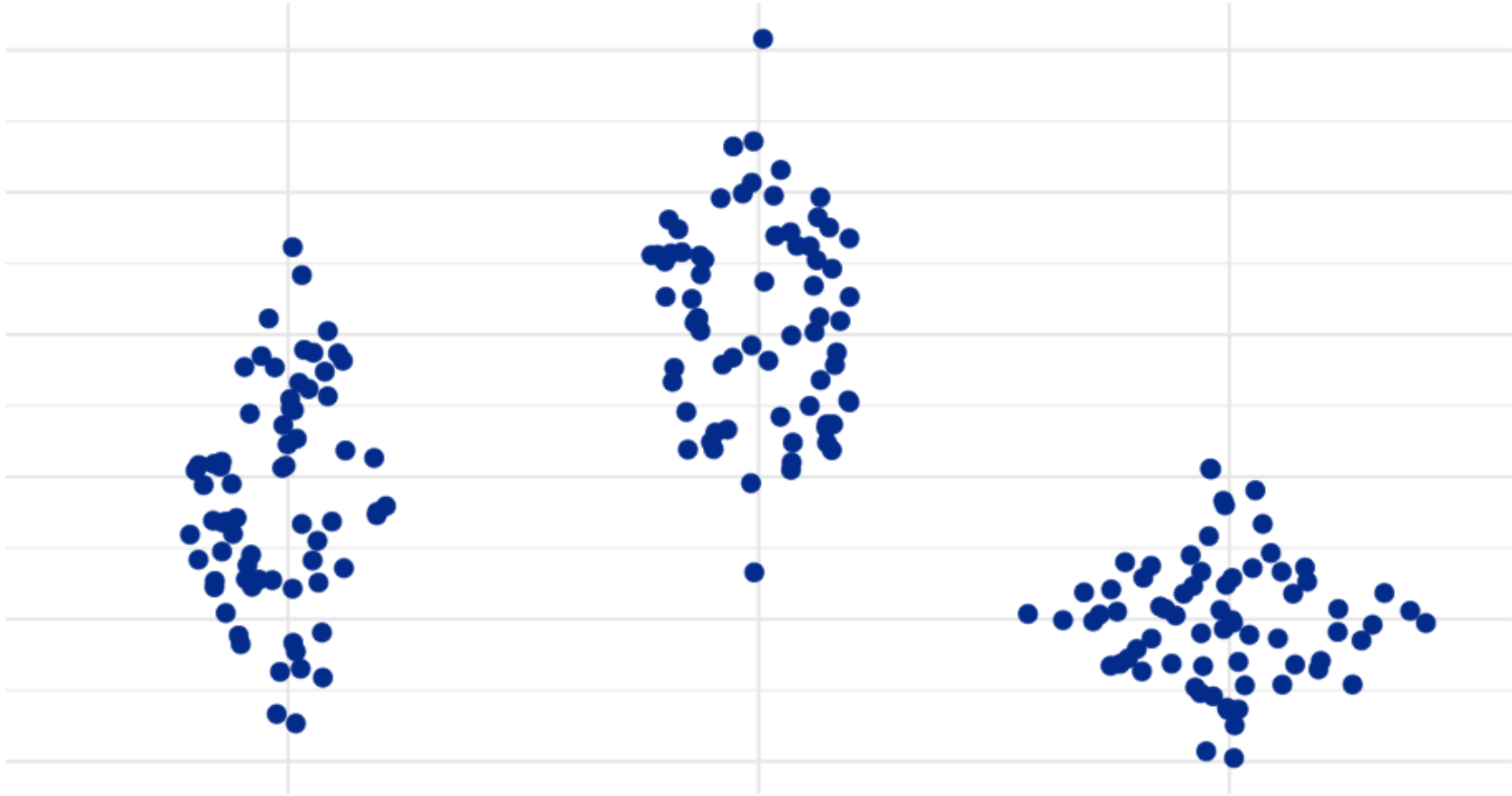
Violin plots



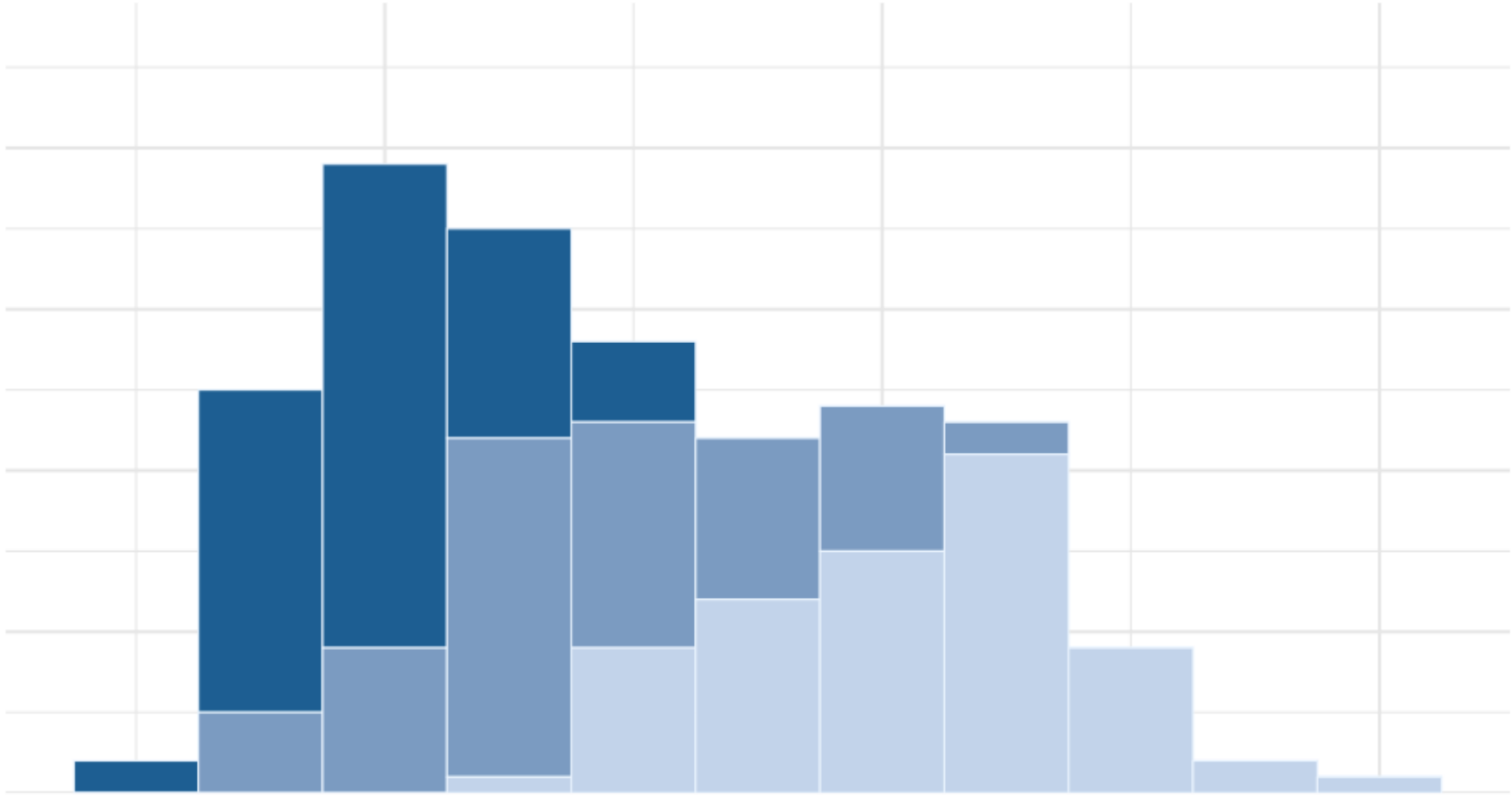
Jittered points



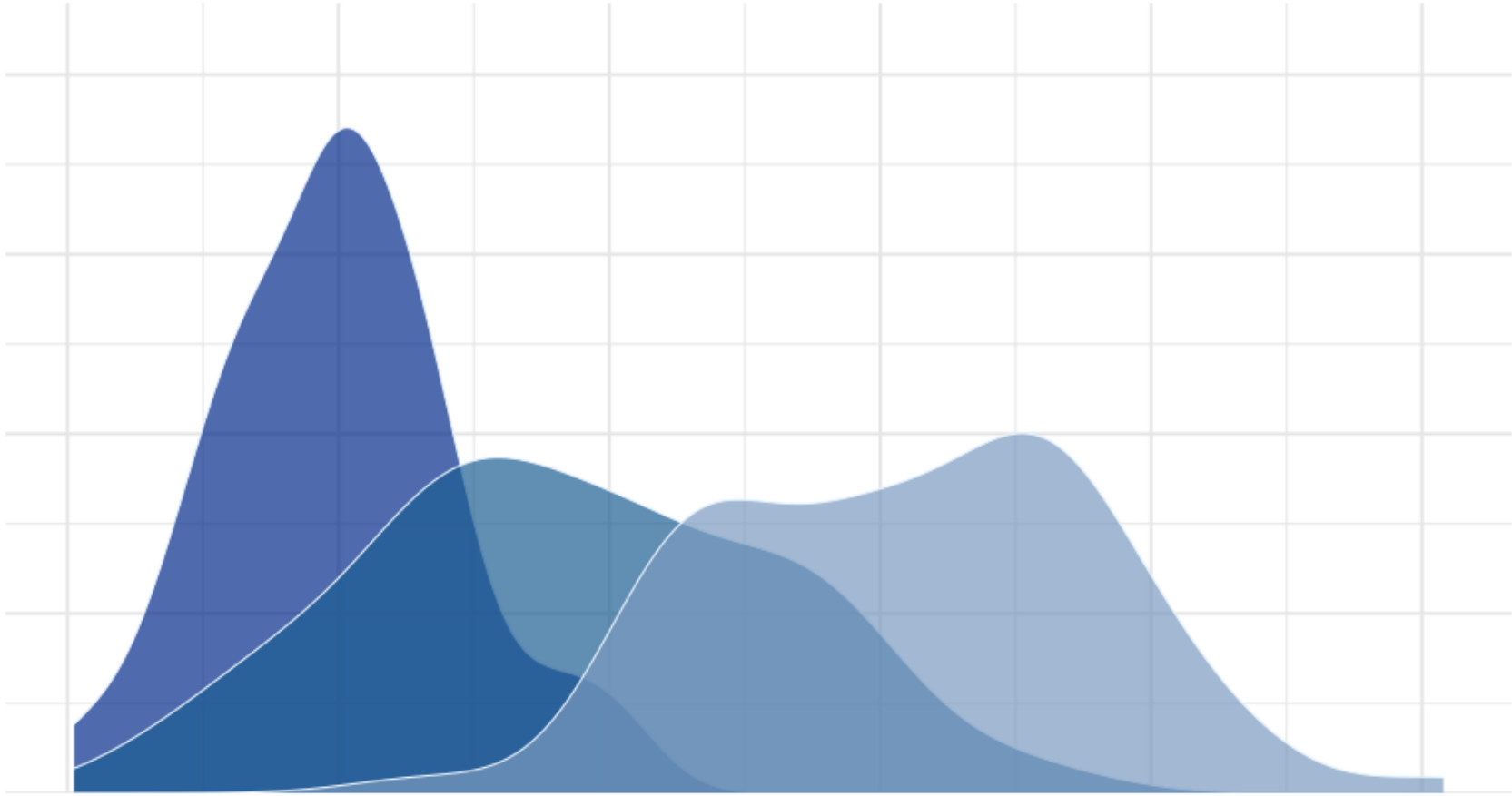
Sina plots



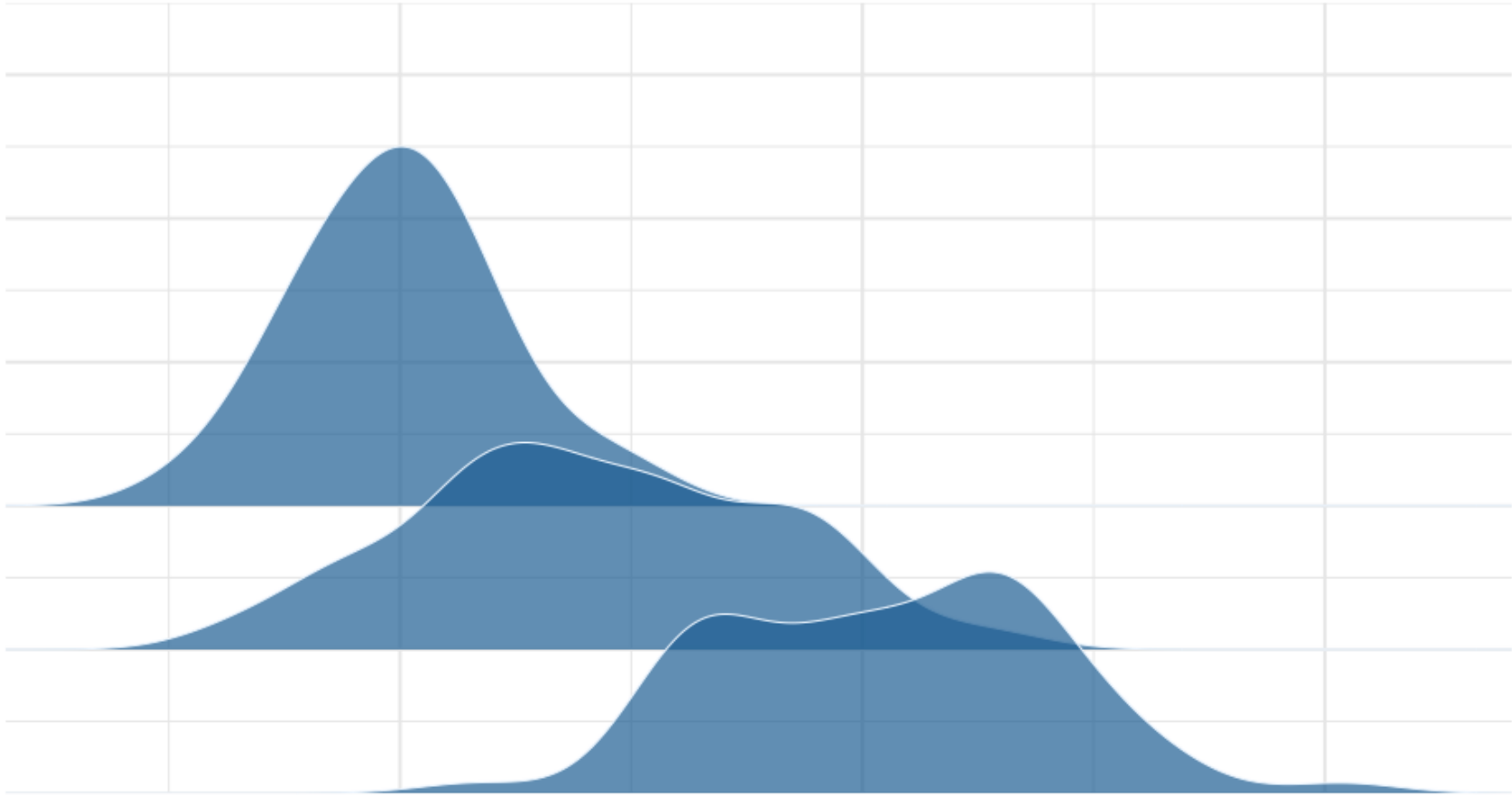
Stacked histograms



Overlapping densities



Ridgeline densities

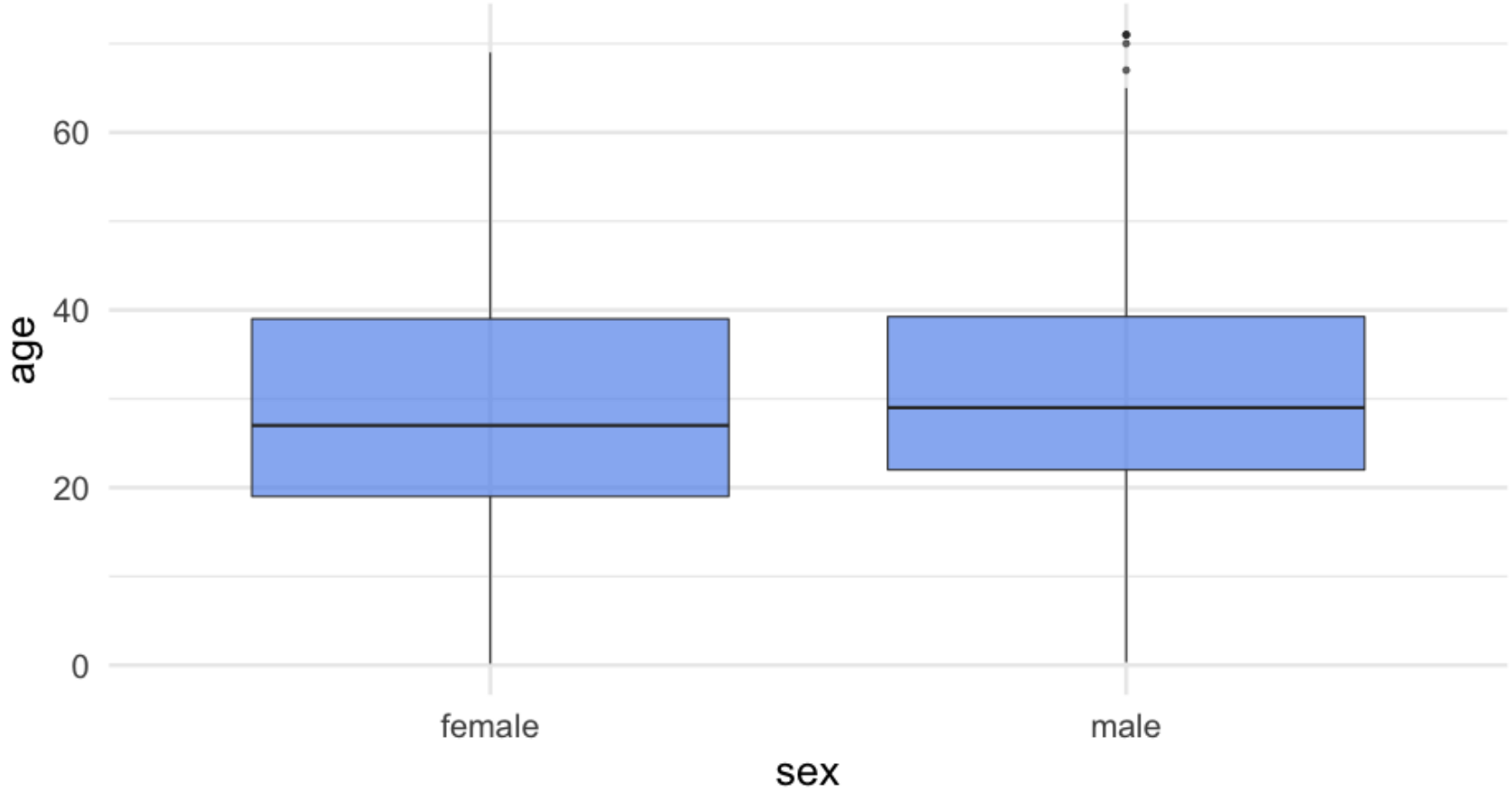


Quick empirical examples

Titanic data

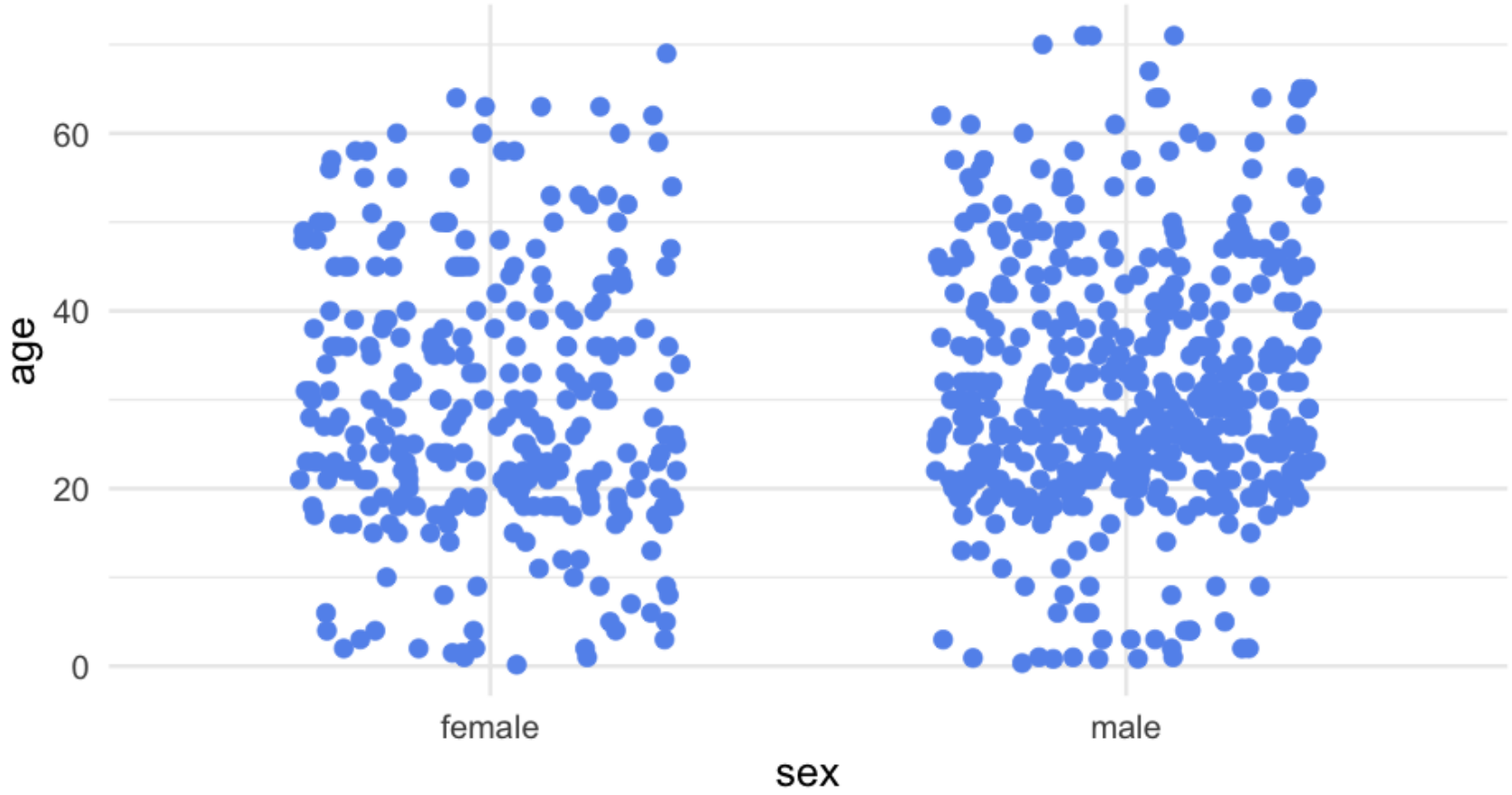
```
## # A tibble: 1,313 x 5
##   name                                            class   age sex    survived
##   <chr>                                           <chr> <dbl> <chr>     <int>
## 1 Allen, Miss Elisabeth Walton                    1st   29    female        1
## 2 Allison, Miss Helen Loraine                     1st    2    female        0
## 3 Allison, Mr Hudson Joshua Creighton             1st   30    male          0
## 4 Allison, Mrs Hudson JC (Bessie Waldo Daniels) 1st   25    female        0
## 5 Allison, Master Hudson Trevor                   1st    0.92 male          1
## 6 Anderson, Mr Harry                              1st   47    male          1
## # … with 1,307 more rows
```

Boxplots

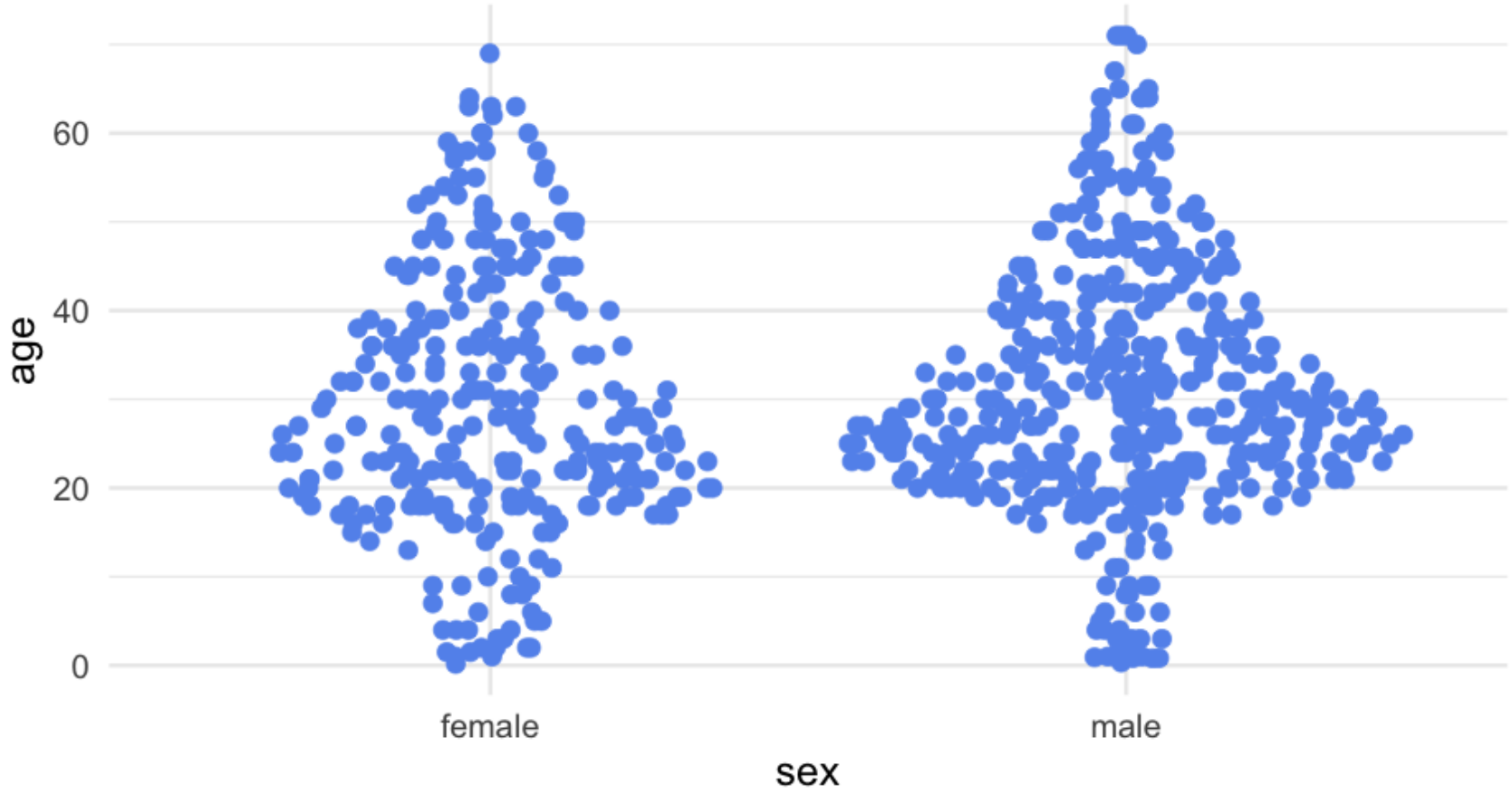


Violin plots

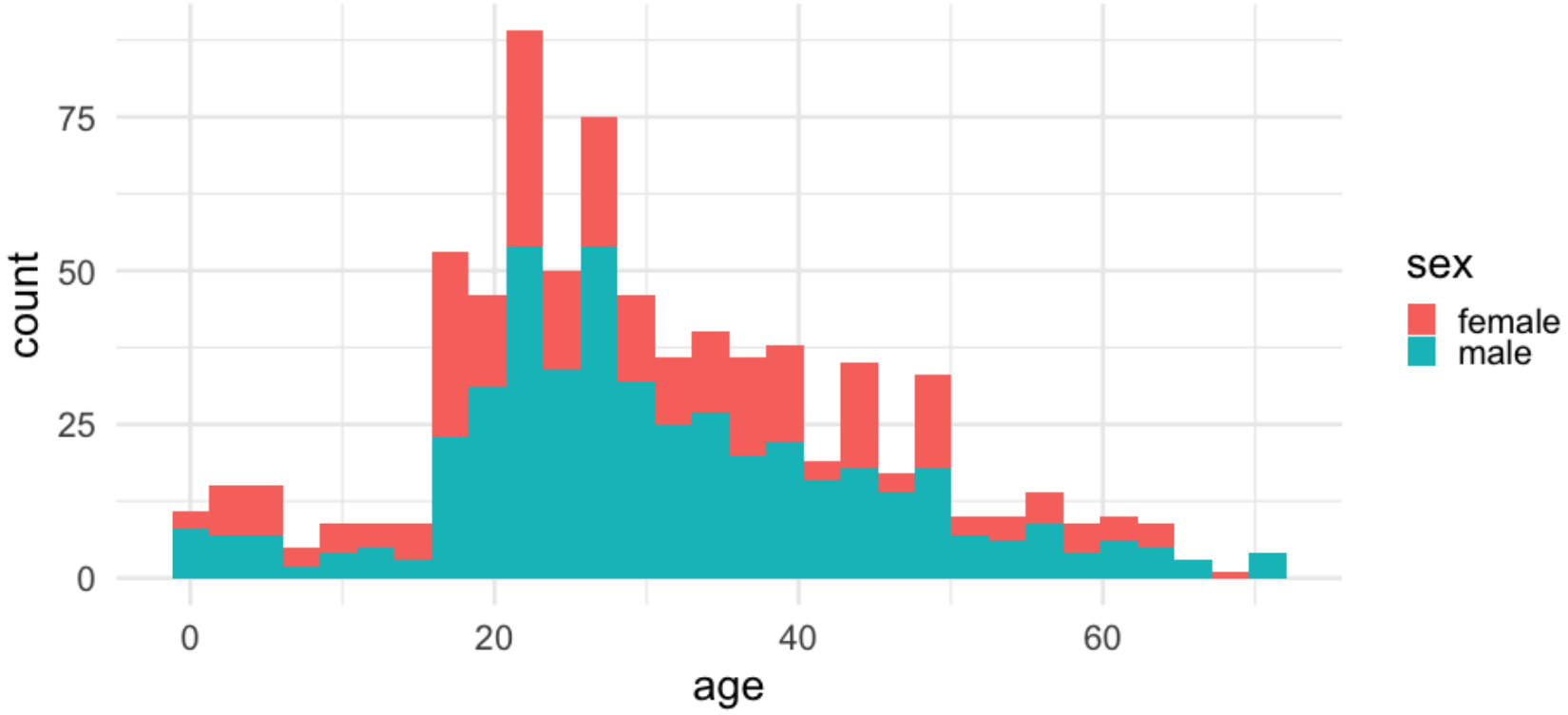
Jittered point plots



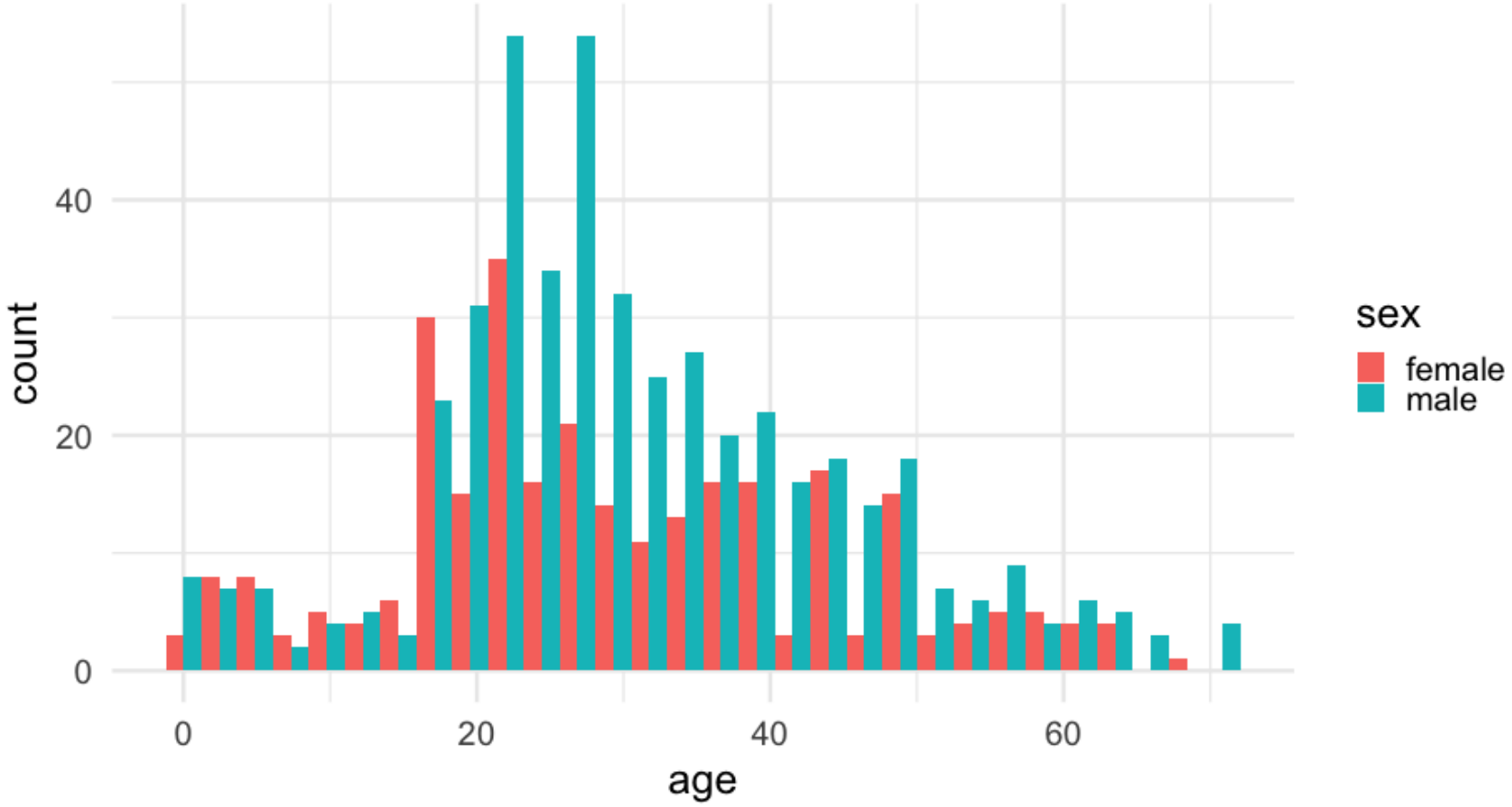
Sina plot



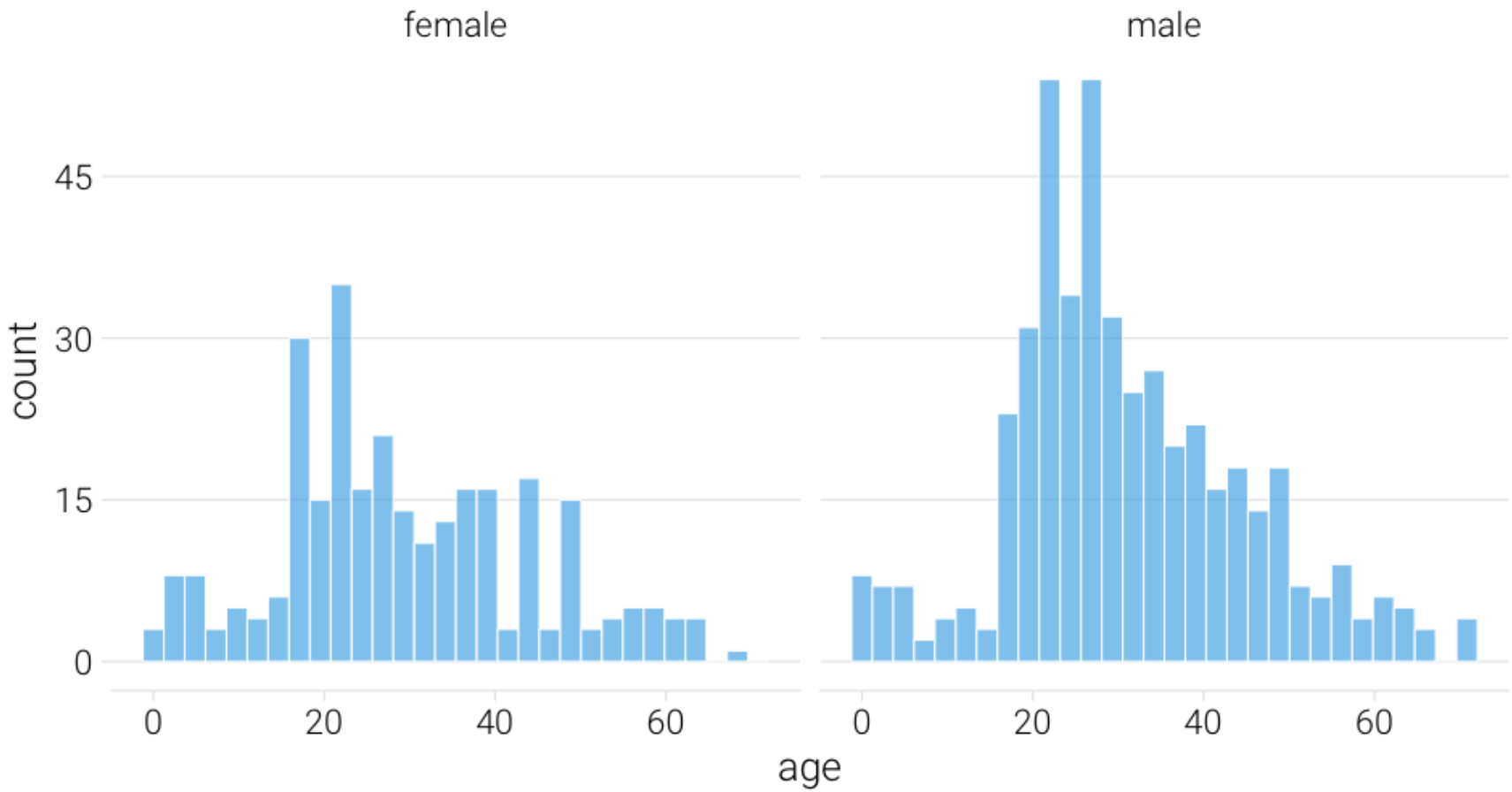
Stacked histogram



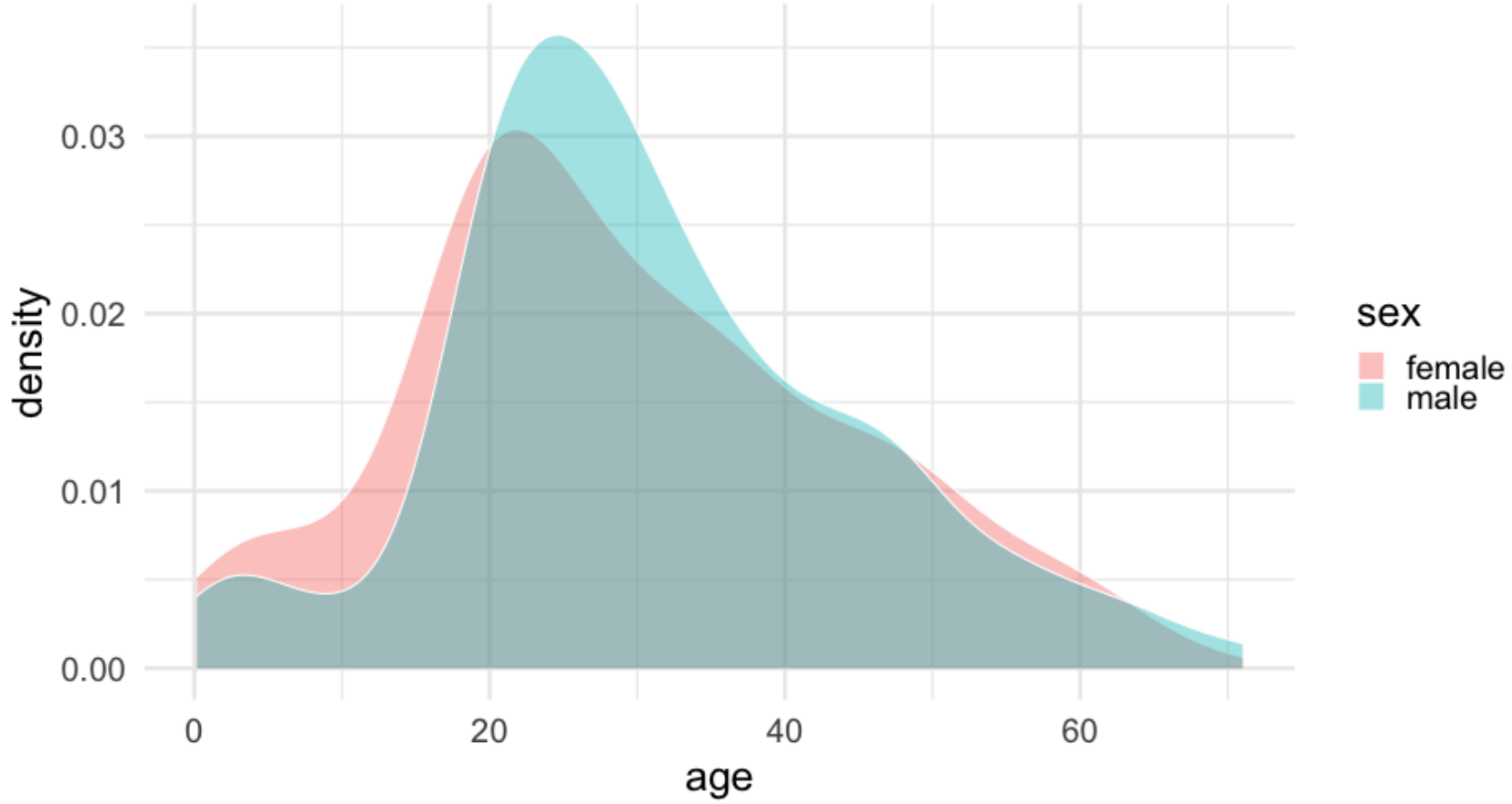
Dodged



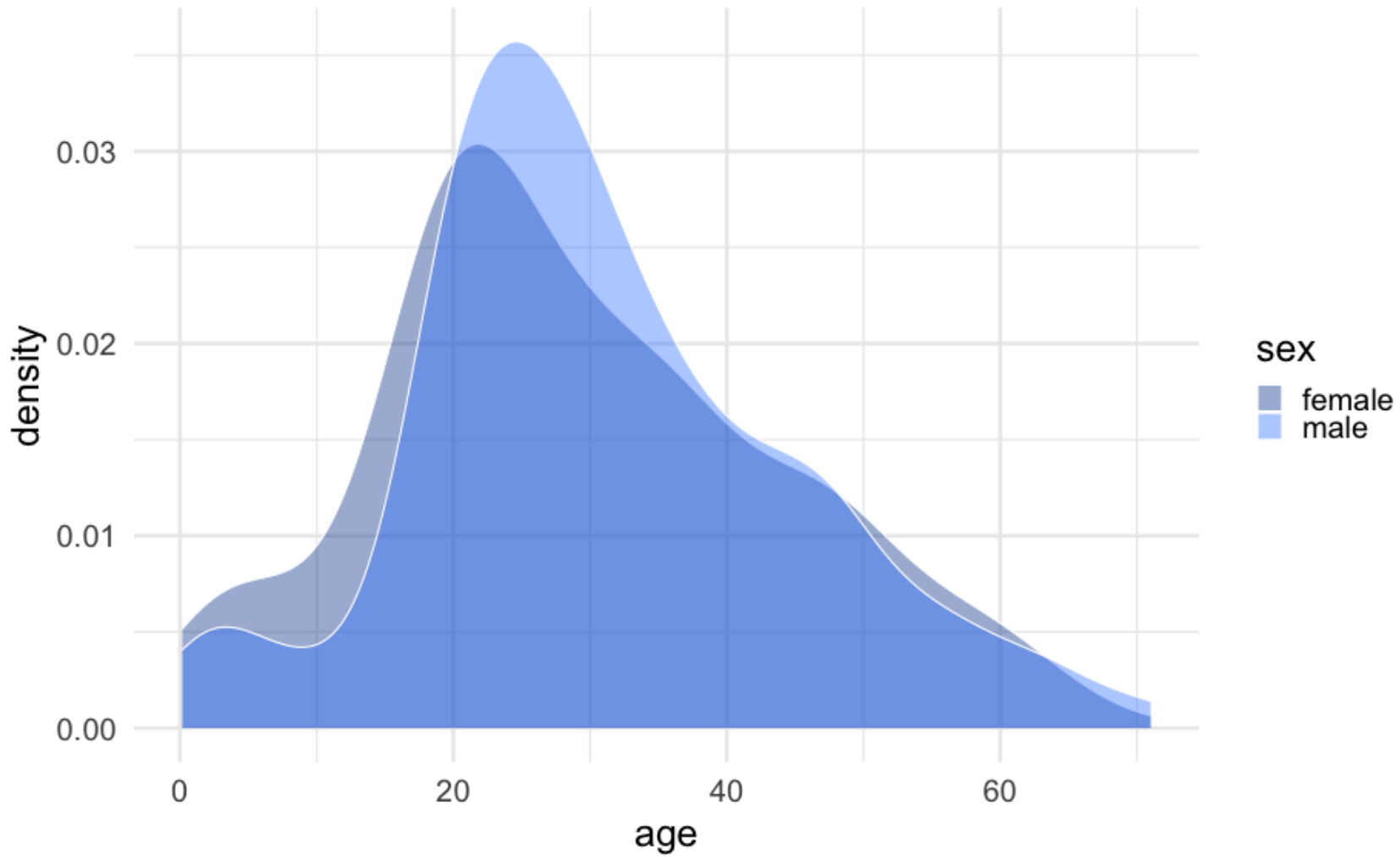
Better



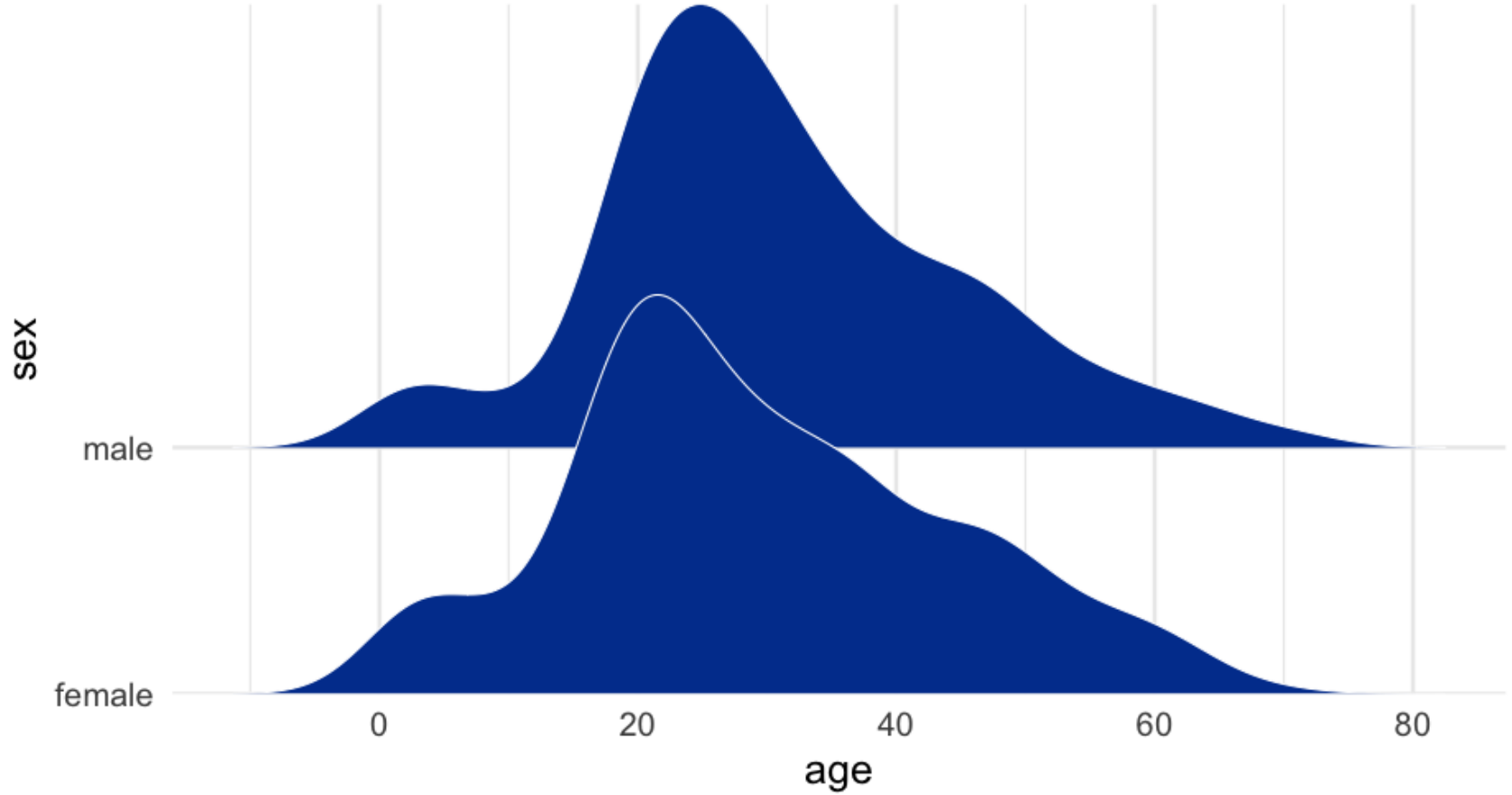
Overlapping densities



Note the default colors really don't work well in most of these

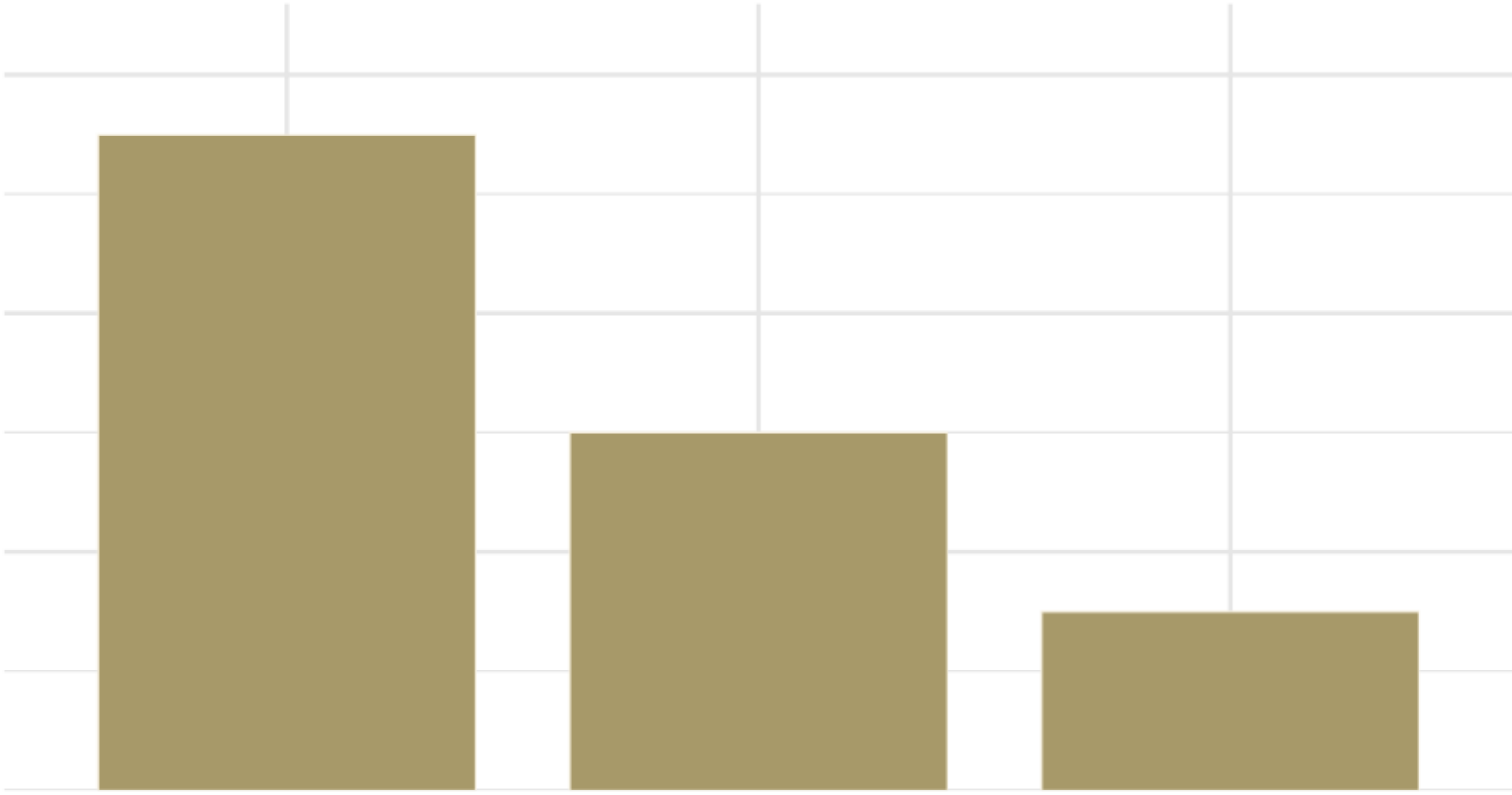

density
0.03
0.02
0.01
0.00
0
20
40
60
age
sex
female
male

Ridgeline densities

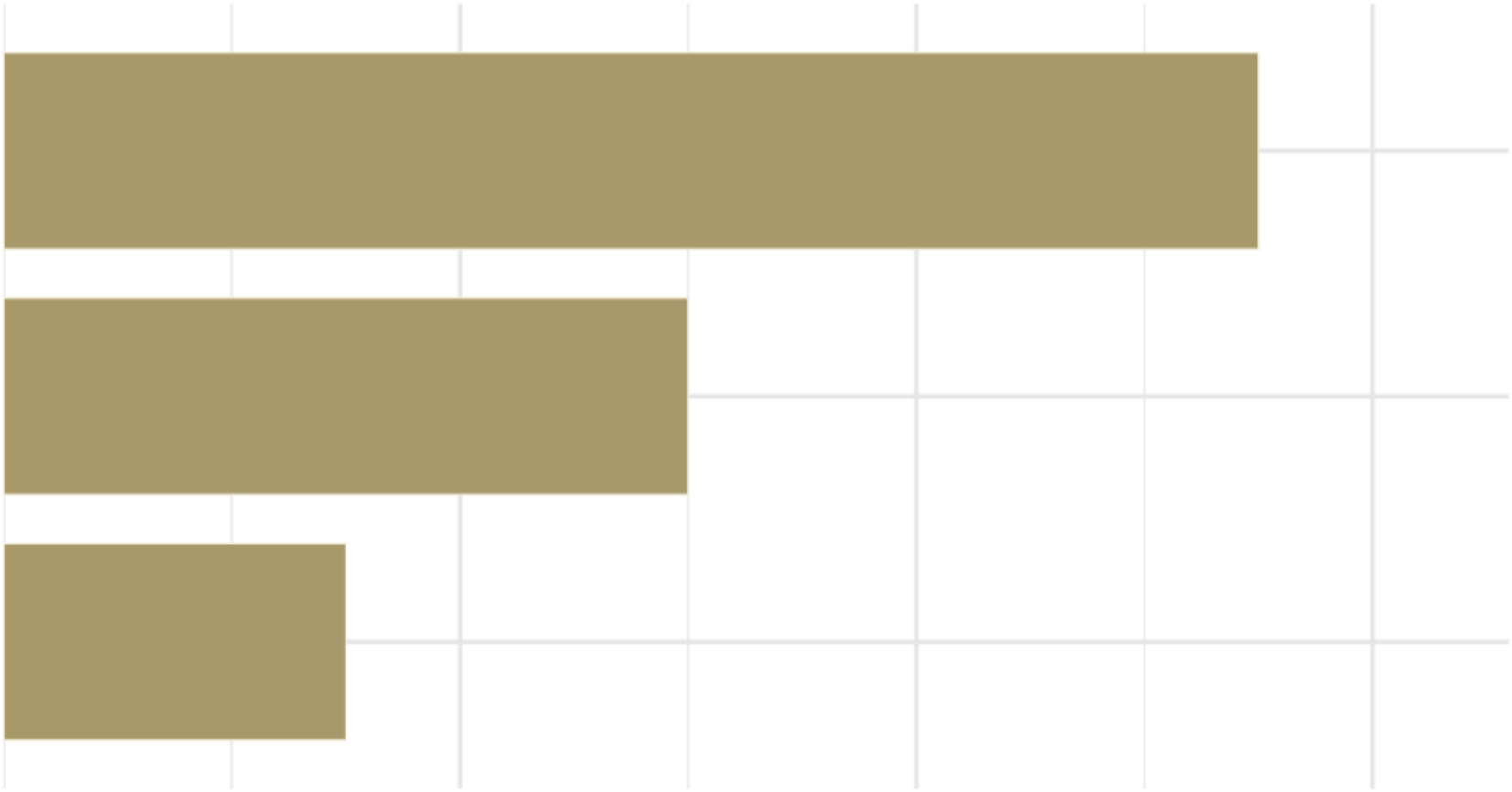


Visualizing amounts

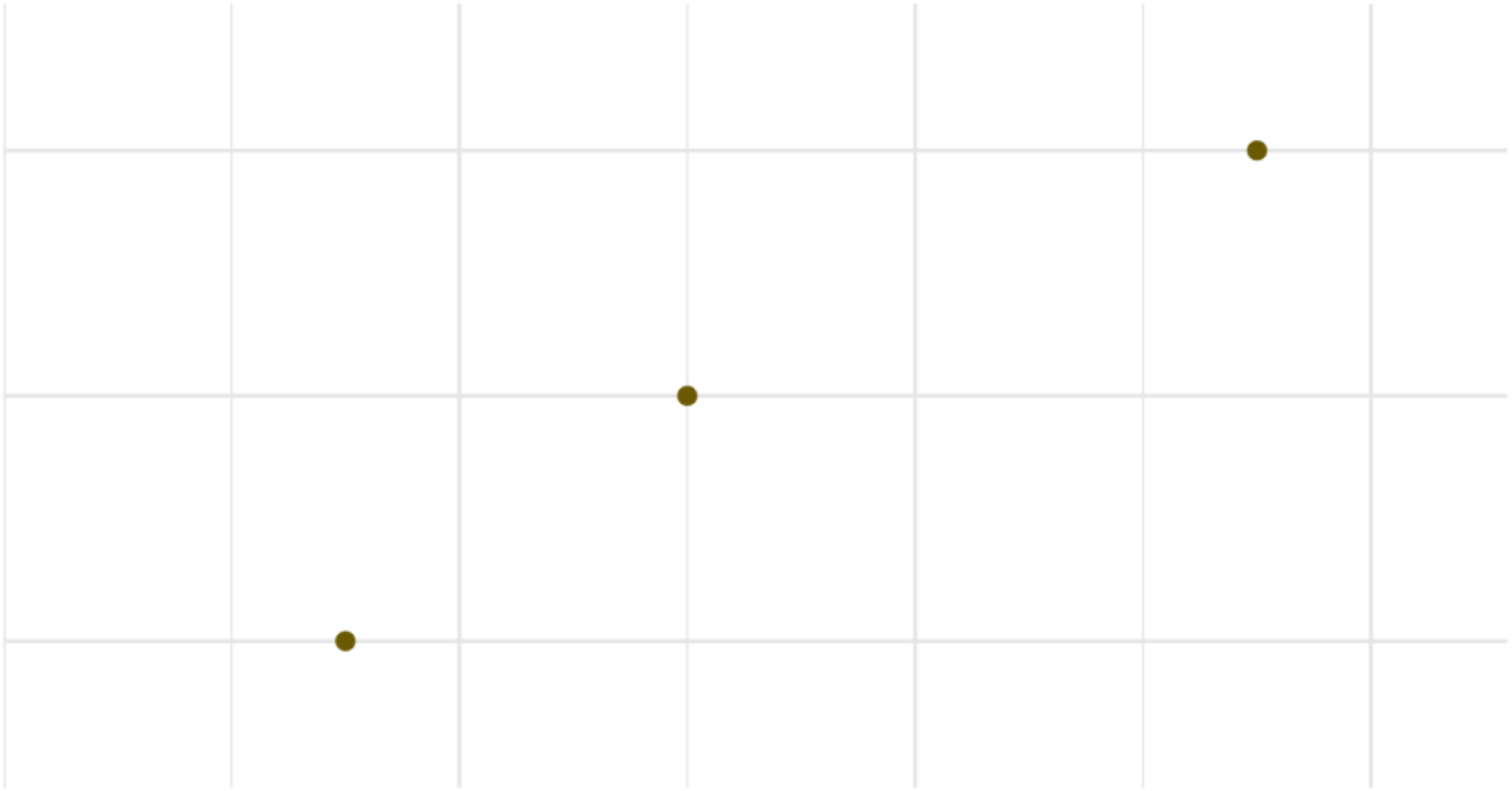
Bar plots



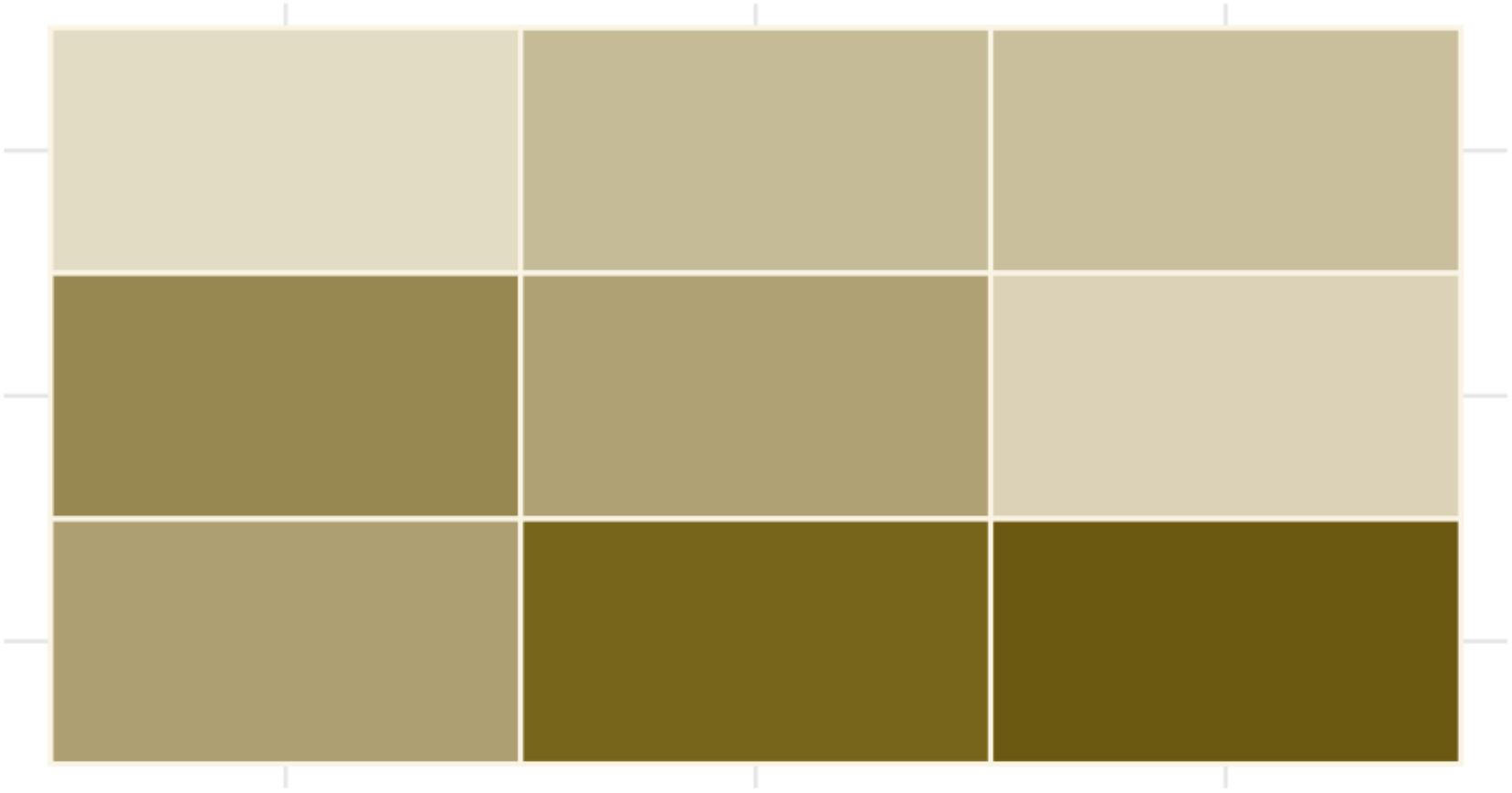
Flipped bars



Dotplot



Heatmap



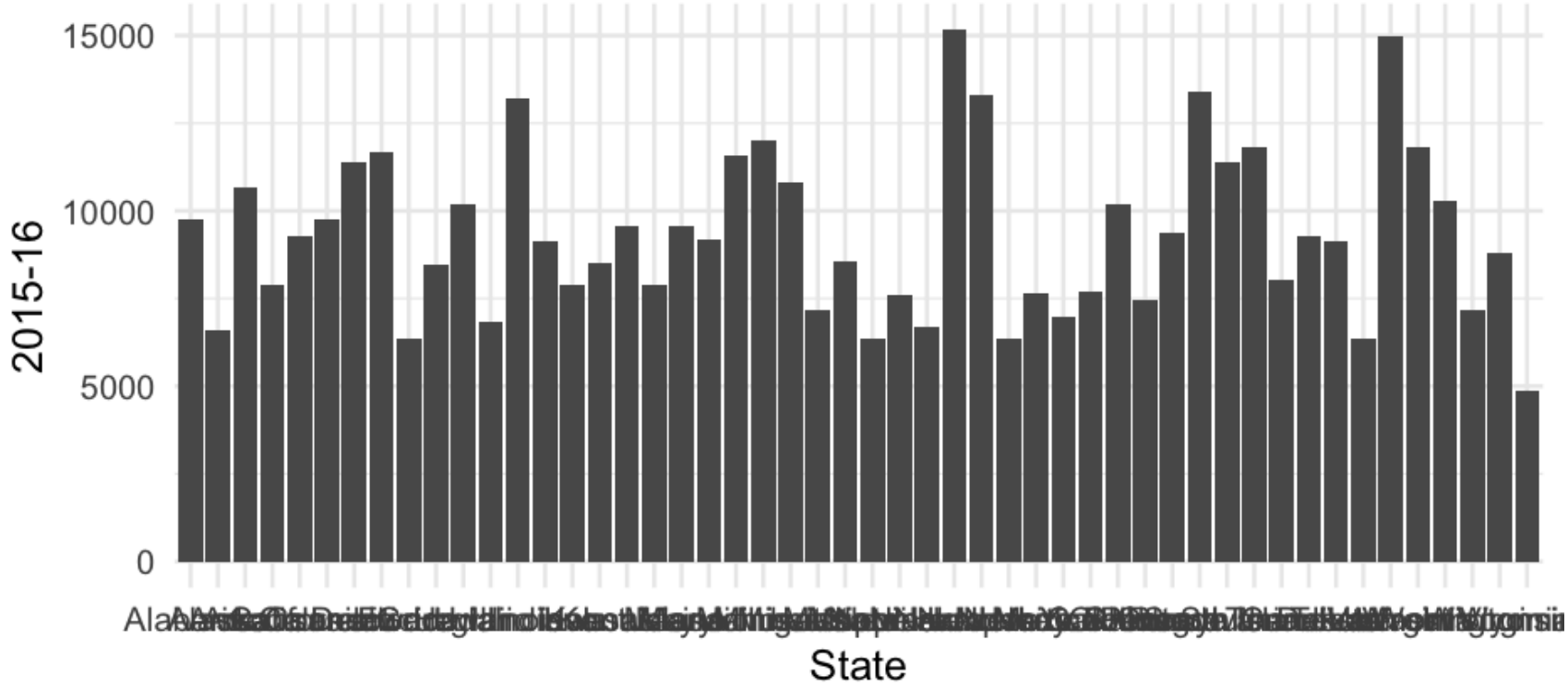
A short journey

How much does college cost?

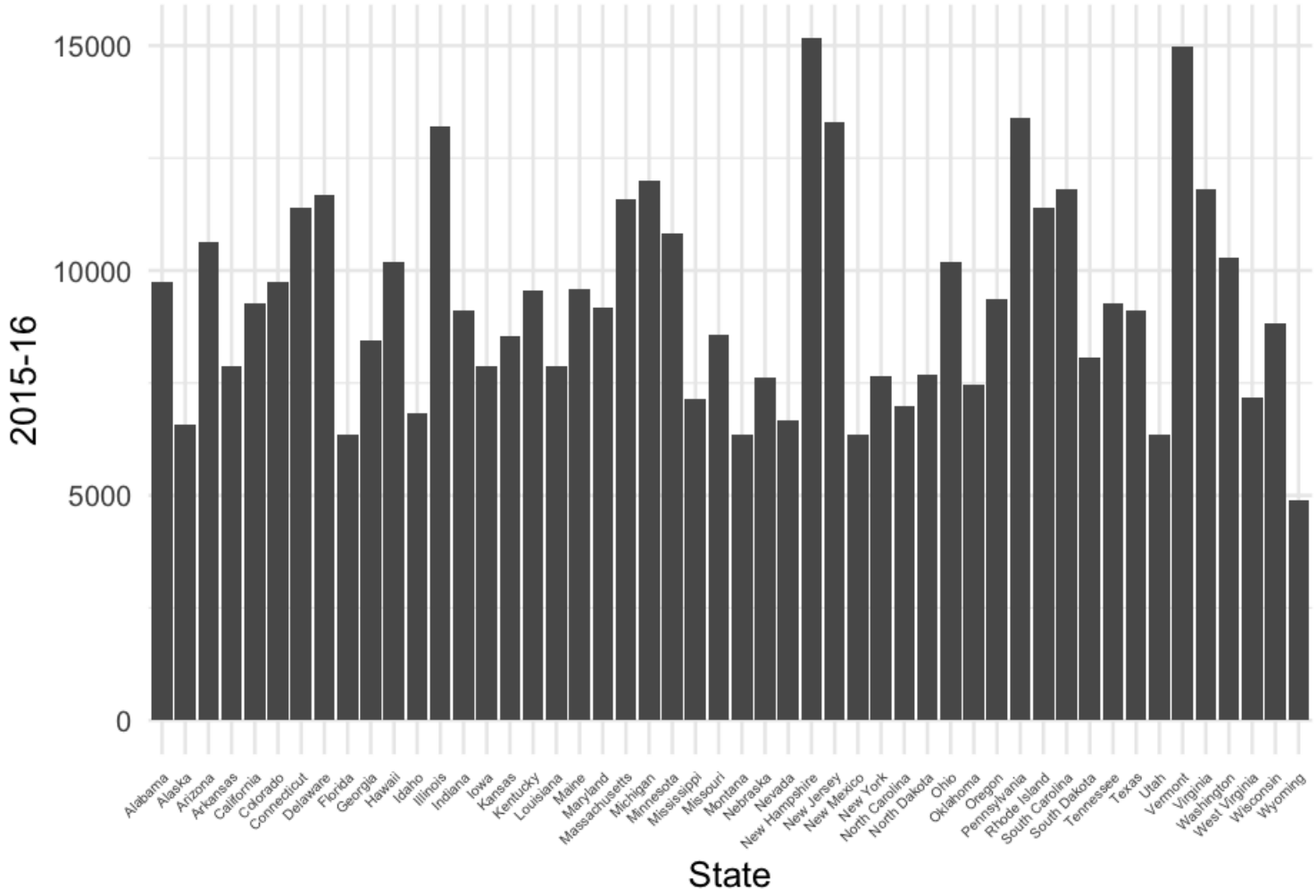
Tuition data

```
## # 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>
```

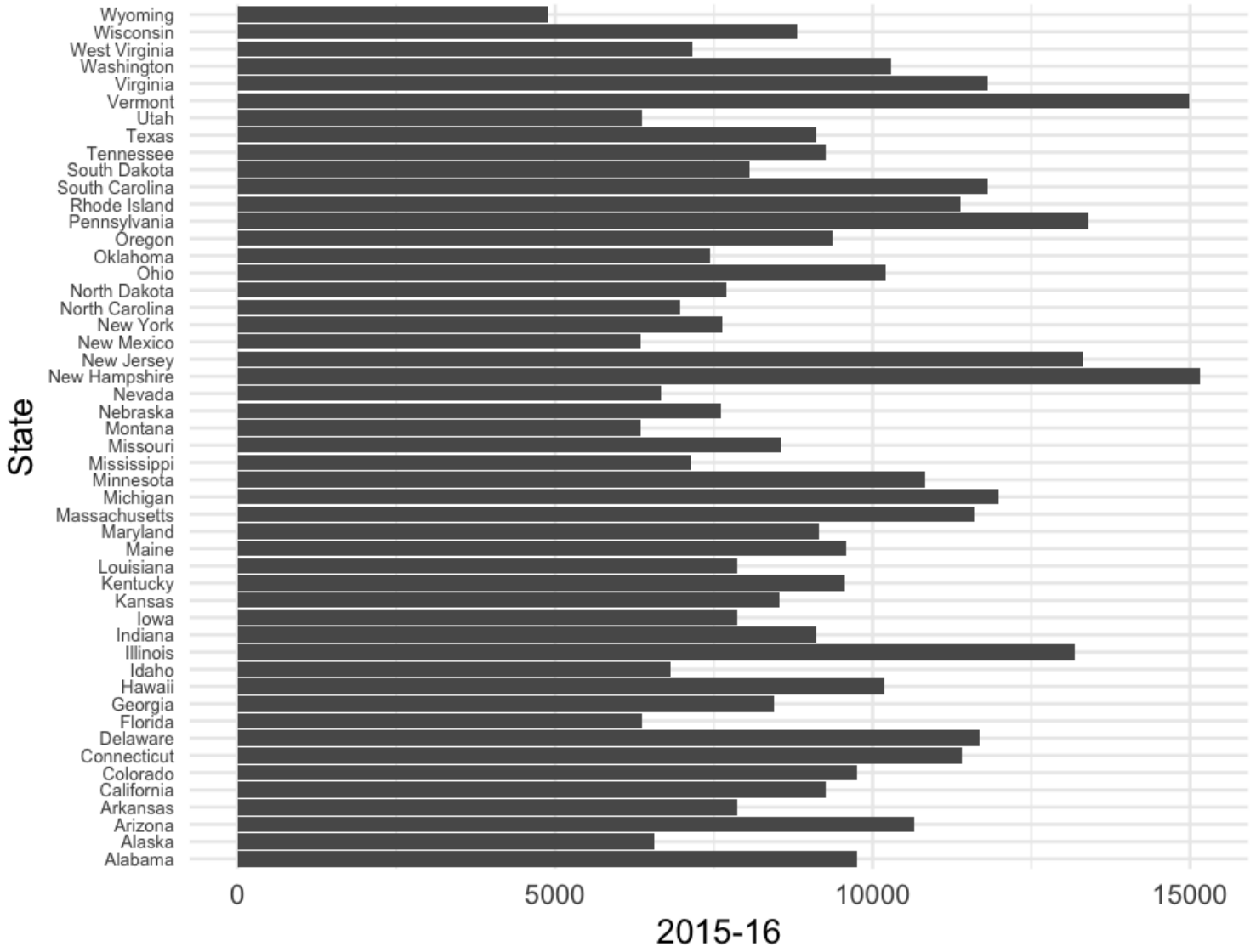
By state: 2015–16



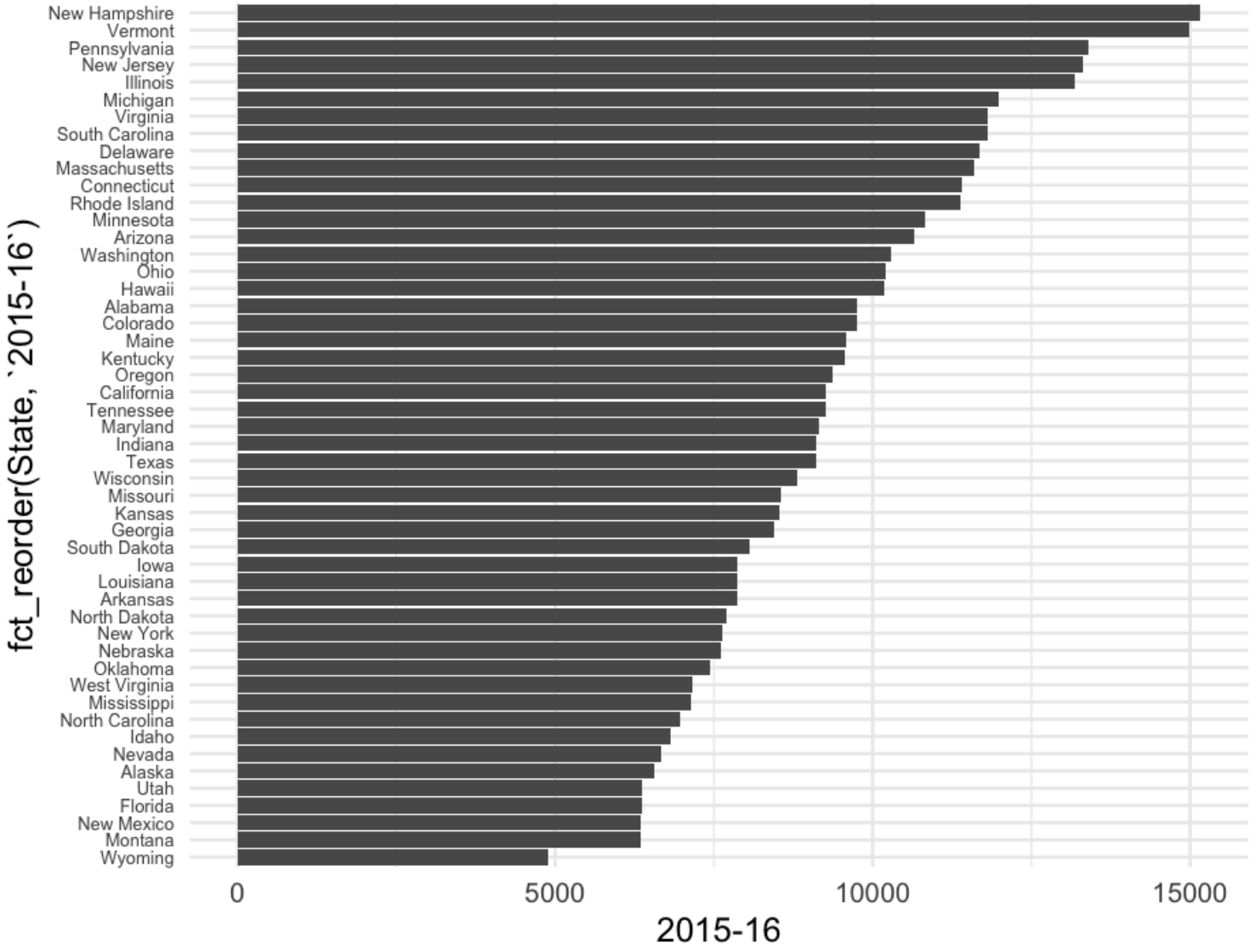
Two puke emoji version



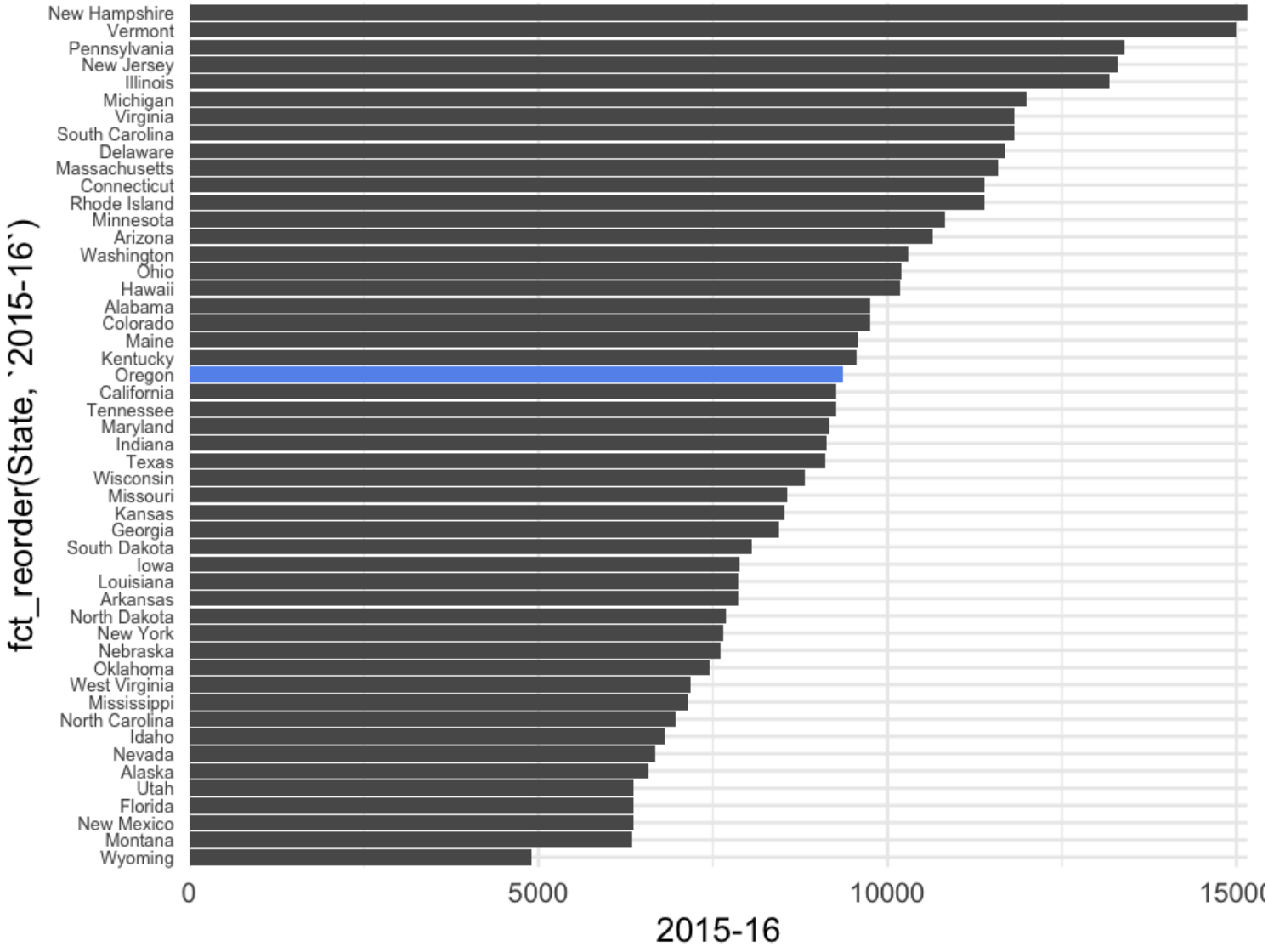
One puke emoji version



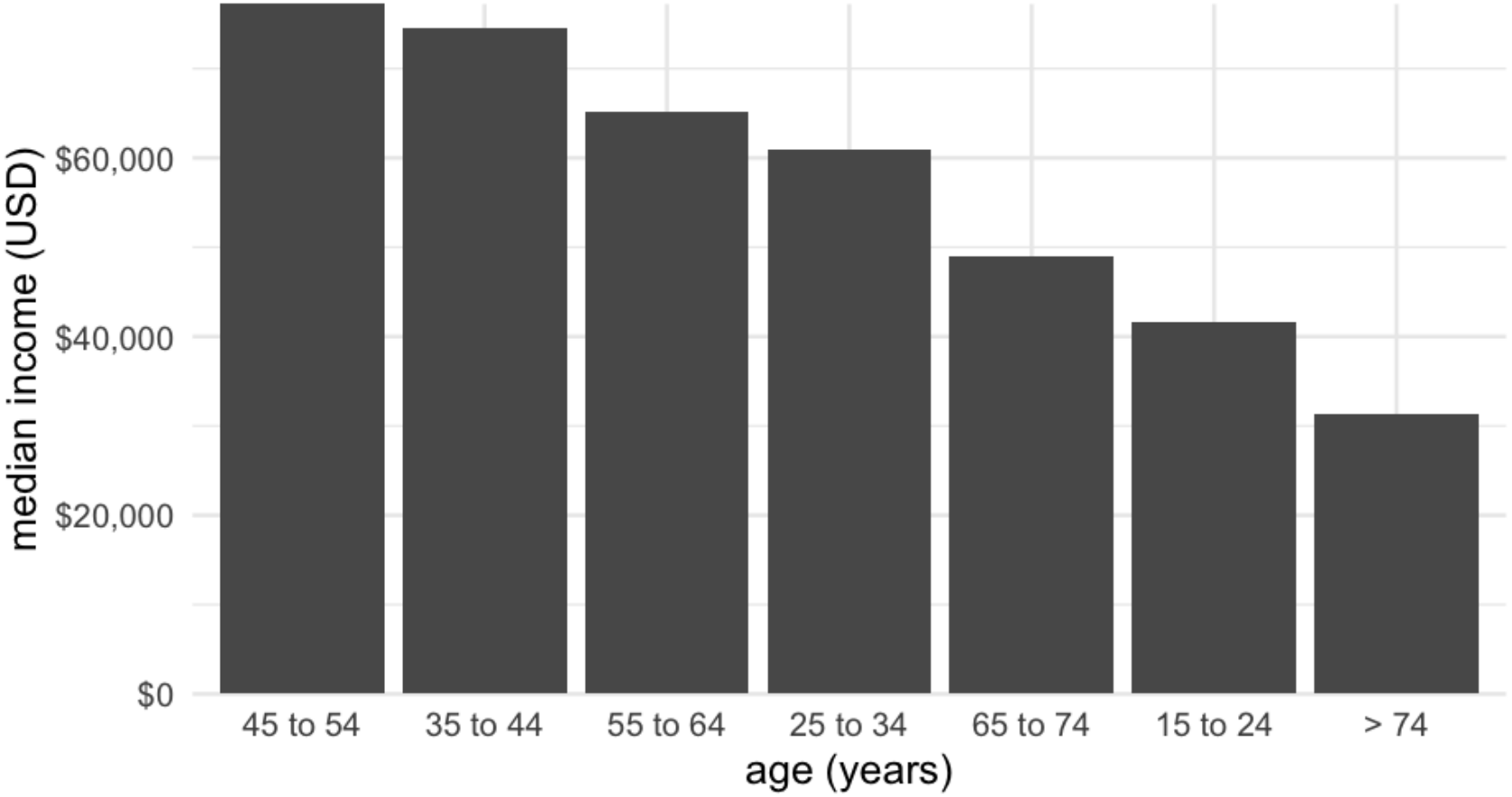
Kinda smiley version



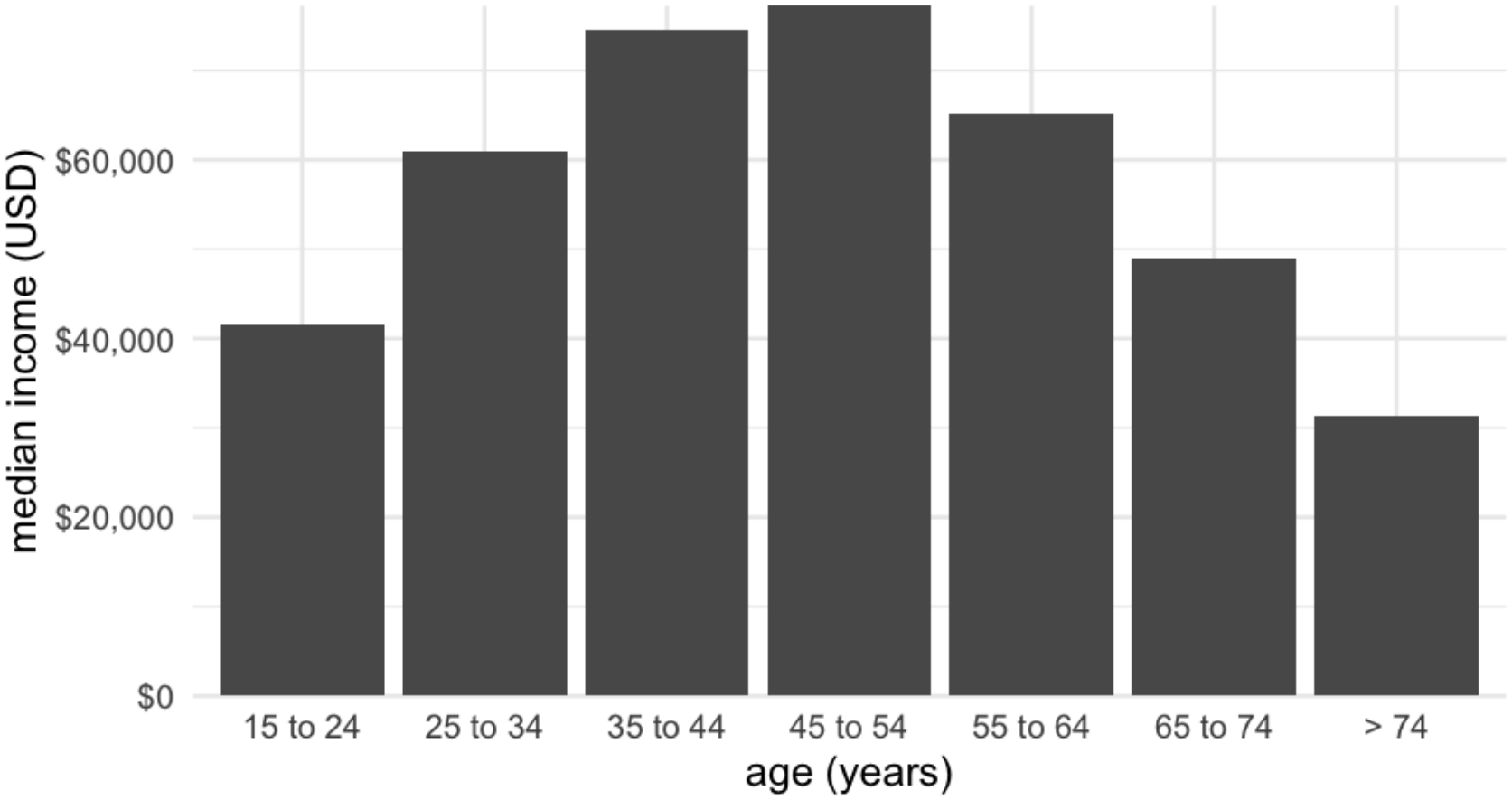
Highlight Oregon



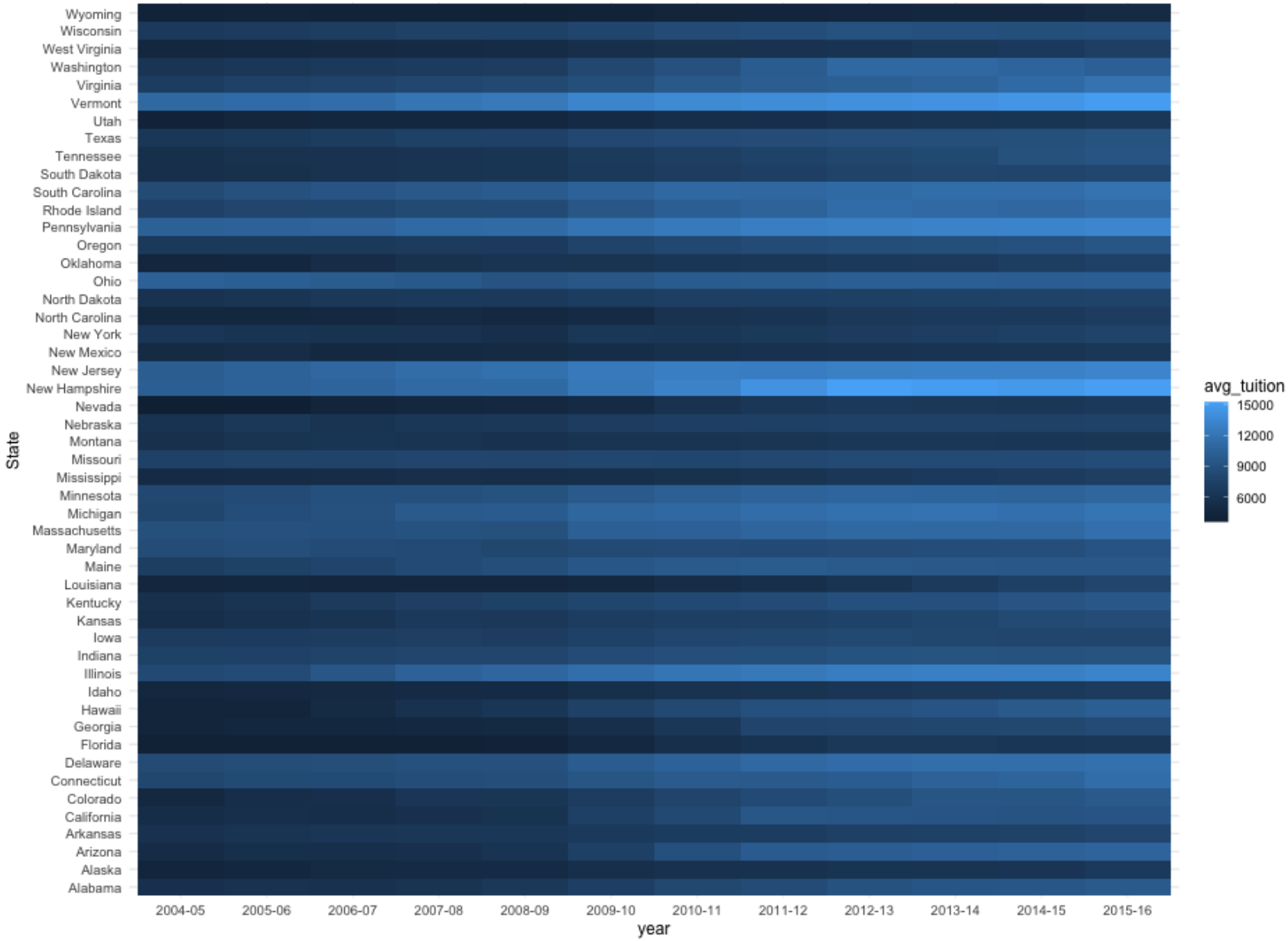
Not always good to sort



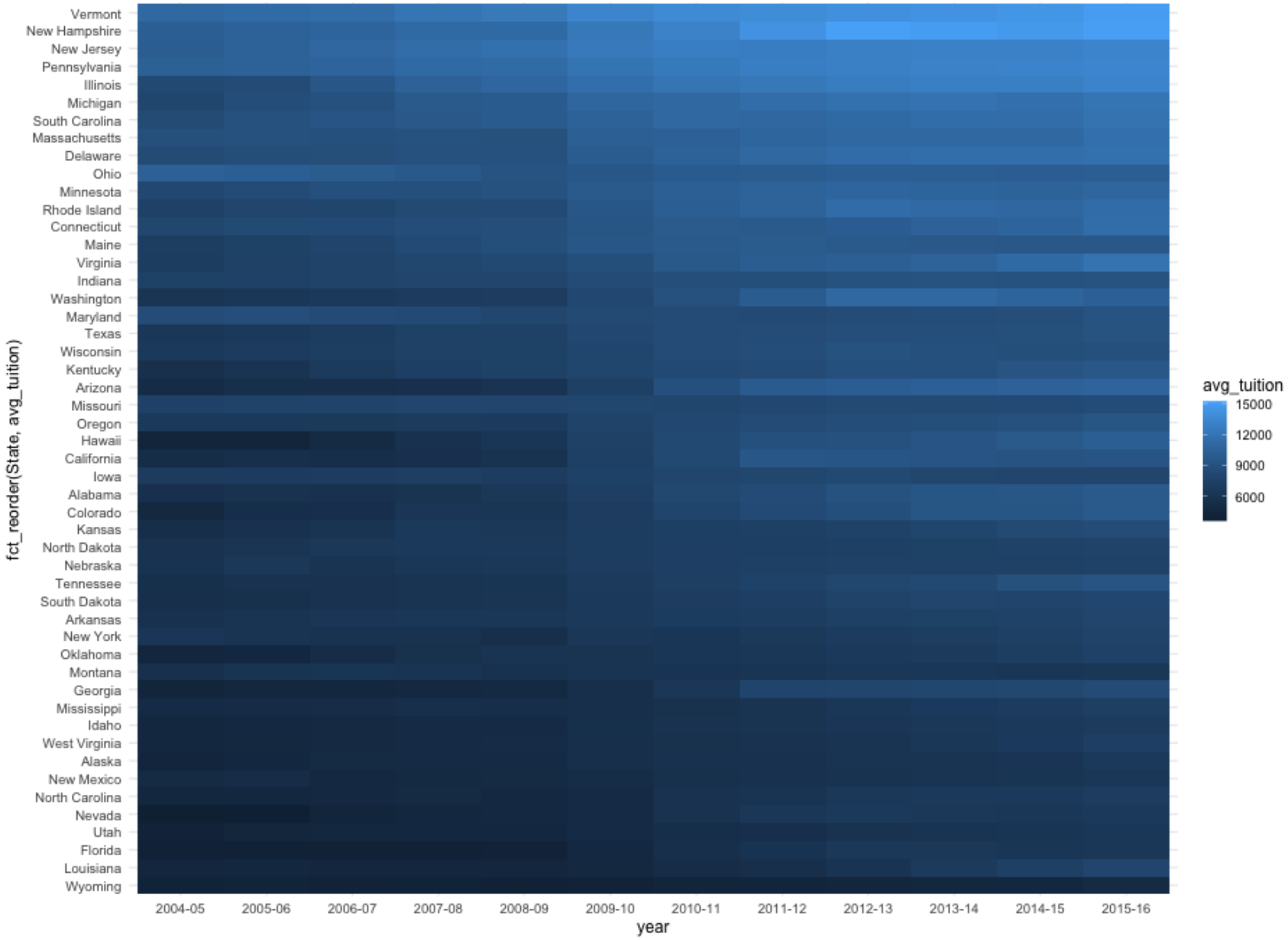
Much better


median income (USD)
$60,000
$40,000
$20,000
$0
15 to 24
25 to 34
35 to 44
45 to 54
55 to 64
65 to 74
> 74
age (years)

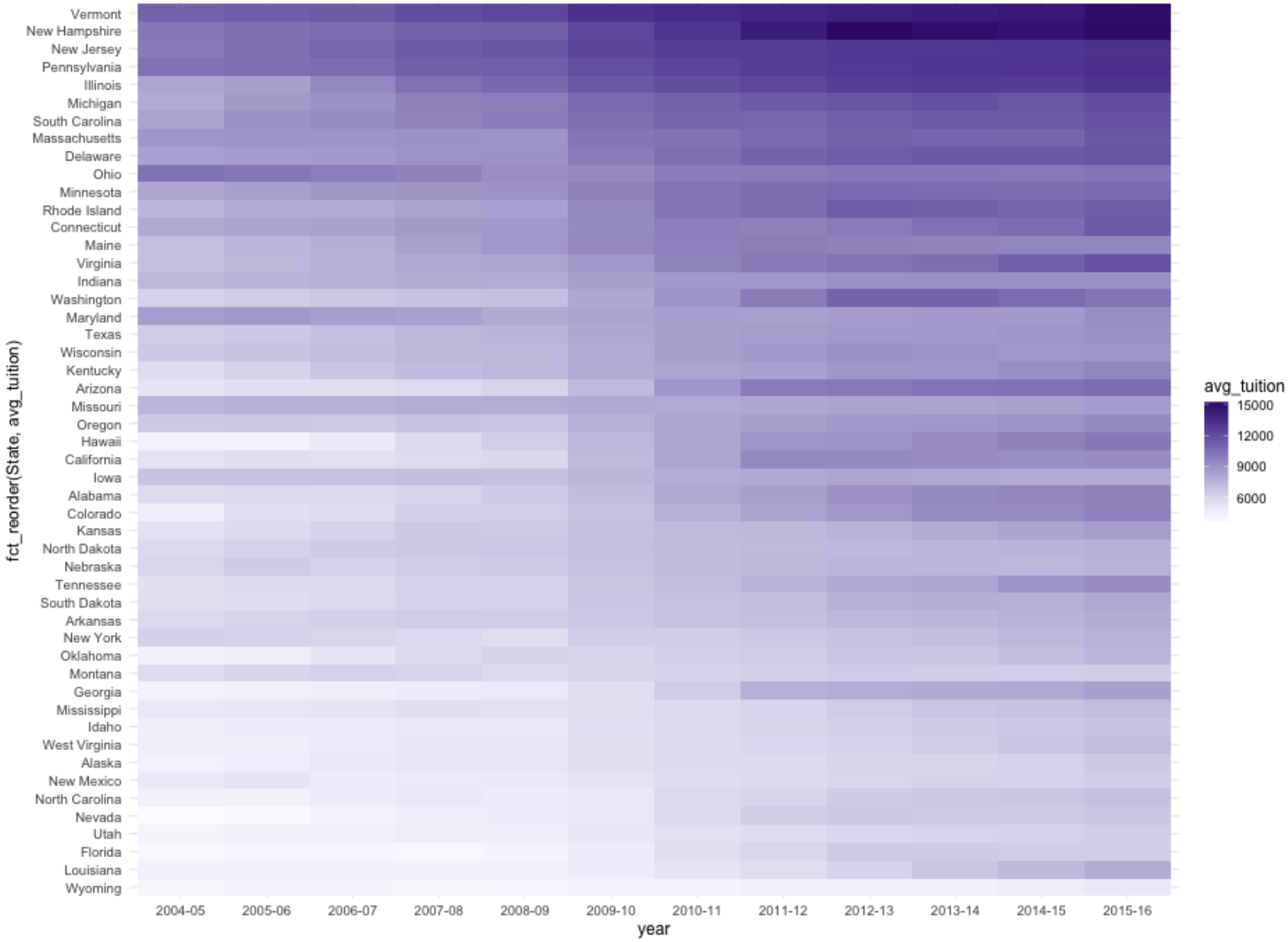
Heatmap


Wyoming
Wisconsin
West Virginia
Washington
Virginia
Vermont
Utah
Texas
Tennessee
South Dakota
South Carolina
Rhode Island
Pennsylvania
Oregon
Oklahoma
Ohio
North Dakota
North Carolina
New York
New Mexico
New Jersey
New Hampshire
Nevada
Nebraska
Montana
Missouri
Mississippi
Minnesota
Michigan
Massachusetts
Maryland
Maine
Louisiana
Kentucky
Kansas
Iowa
Indiana
Illinois
Idaho
Hawaii
Georgia
Florida
Delaware
Connecticut
Colorado
California
Arkansas
Arizona
Alaska
Alabama
State
2004-05
2005-06
2006-07
2007-08
2008-09
2009-10
2010-11
2011-12
2012-13
2013-14
2014-15
2015-16
year
avg_tuition
15000
12000
9000
6000

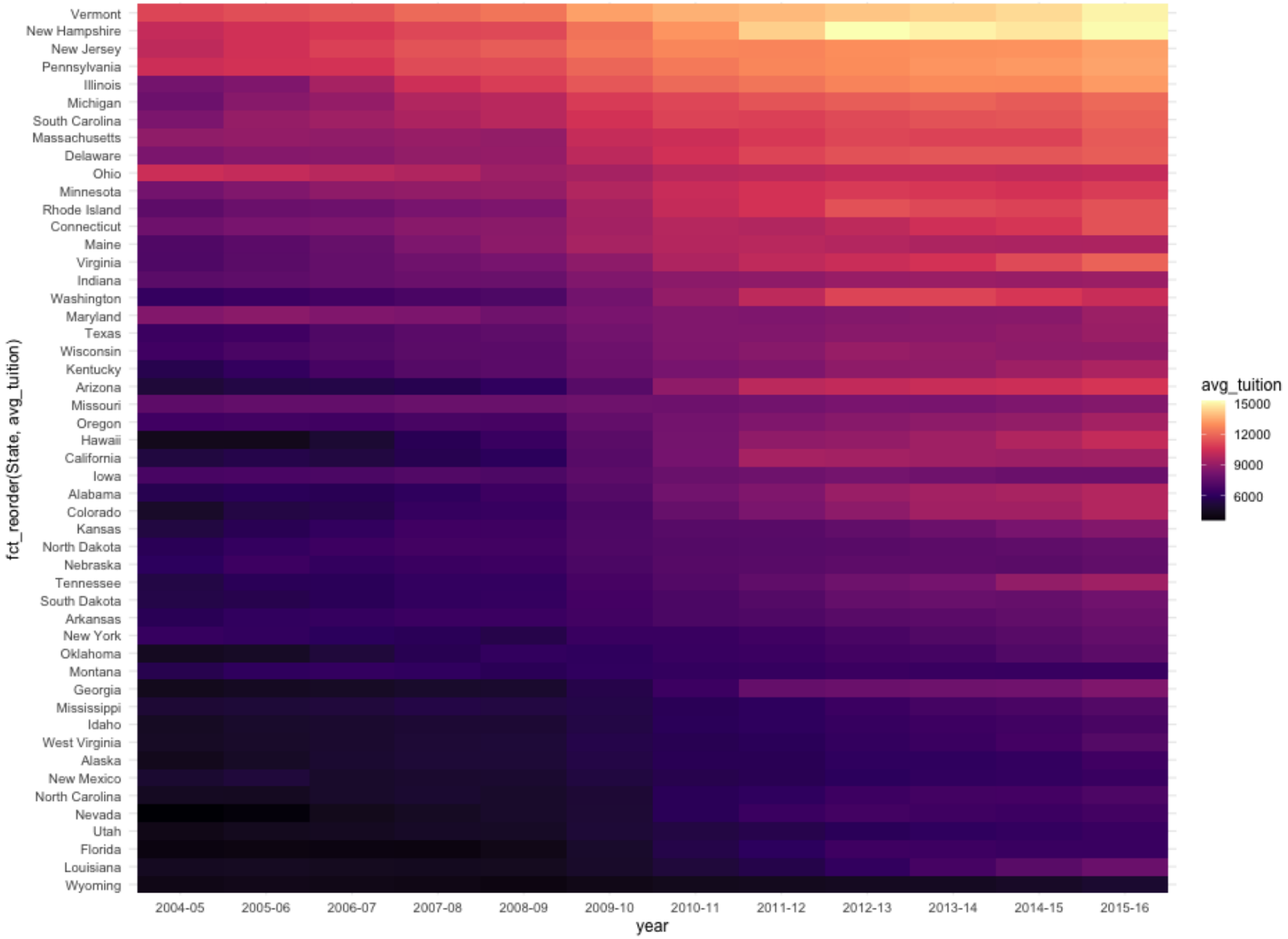
Better heatmap

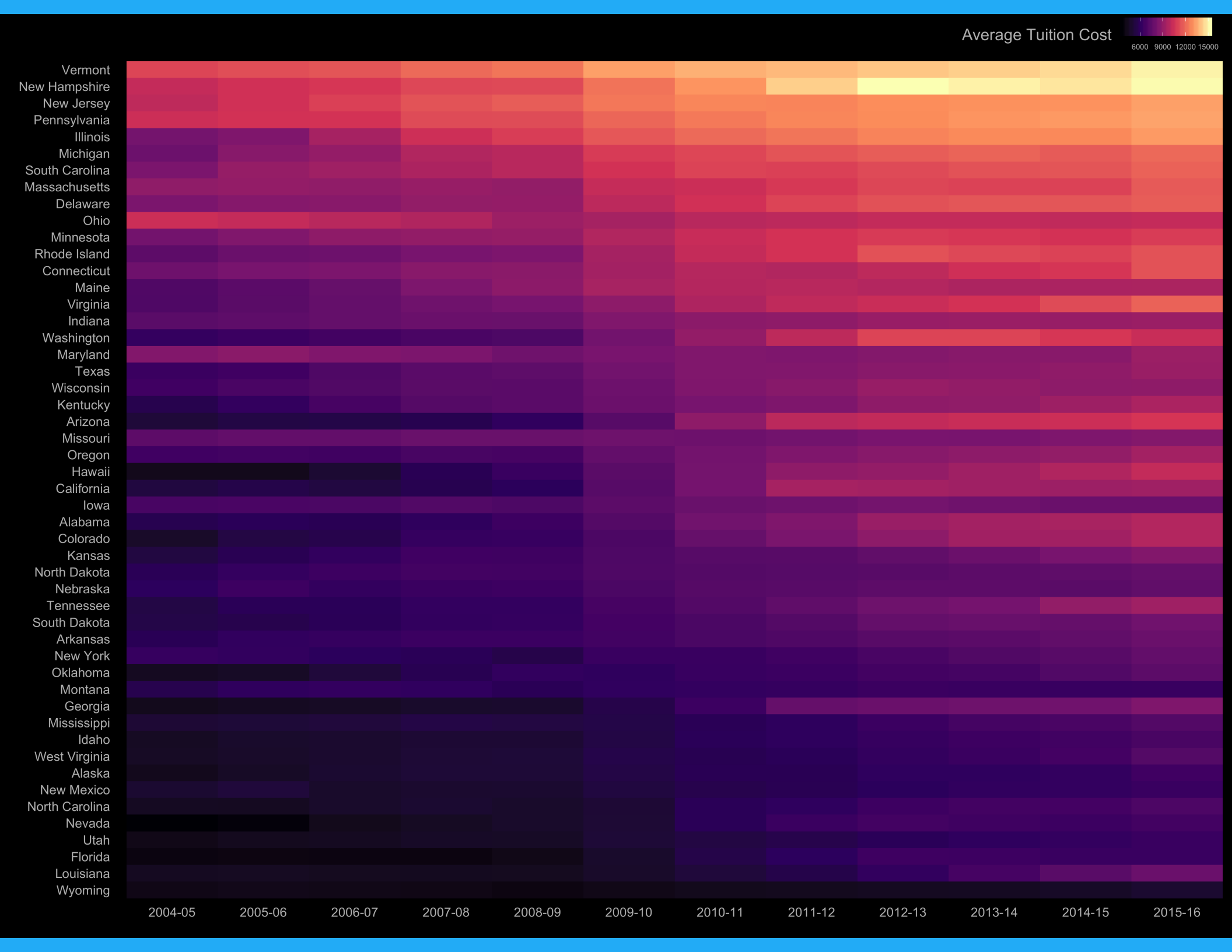


Even better?


Vermont
New Hampshire
New Jersey
Pennsylvania
Illinois
Michigan
South Carolina
Massachusetts
Delaware
Ohio
Minnesota
Rhode Island
Connecticut
Maine
Virginia
Indiana
Washington
Maryland
Texas
Wisconsin
Kentucky
Arizona
Missouri
Oregon
Hawaii
California
Iowa
Alabama
Colorado
Kansas
North Dakota
Nebraska
Tennessee
South Dakota
Arkansas
New York
Oklahoma
Montana
Georgia
Mississippi
Idaho
West Virginia
Alaska
New Mexico
North Carolina
Nevada
Utah
Florida
Louisiana
Wyoming
fct_reorder(State, avg_tuition)
2004-05
2005-06
2006-07
2007-08
2008-09
2009-10
2010-11
2011-12
2012-13
2013-14
2014-15
2015-16
year
avg_tuition
15000
12000
9000
6000

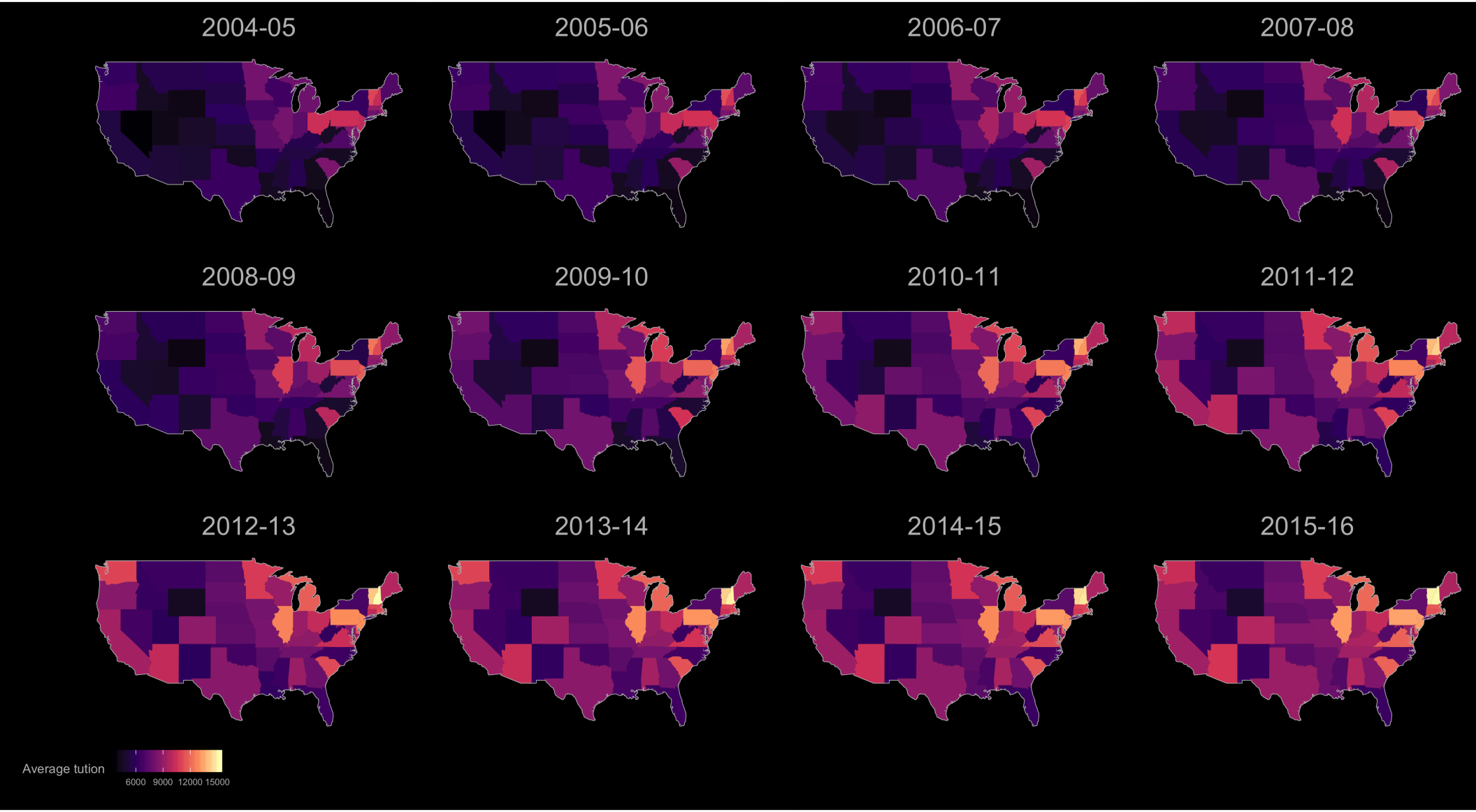
Or maybe this one?




Average Tuition Cost
6000 9000 12000 15000
Vermont
New Hampshire
New Jersey
Pennsylvania
Illinois
Michigan
South Carolina
Massachusetts
Delaware
Ohio
Minnesota
Rhode Island
Connecticut
Maine
Virginia
Indiana
Washington
Maryland
Texas
Wisconsin
Kentucky
Arizona
Missouri
Oregon
Hawaii
California
Iowa
Alabama
Colorado
Kansas
North Dakota
Nebraska
Tennessee
South Dakota
Arkansas
New York
Oklahoma
Montana
Georgia
Mississippi
Idaho
West Virginia
Alaska
New Mexico
North Carolina
Nevada
Utah
Florida
Louisiana
Wyoming
2004-05
2005-06
2006-07
2007-08
2008-09
2009-10
2010-11
2011-12
2012-13
2013-14
2014-15
2015-16

Quick aside

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

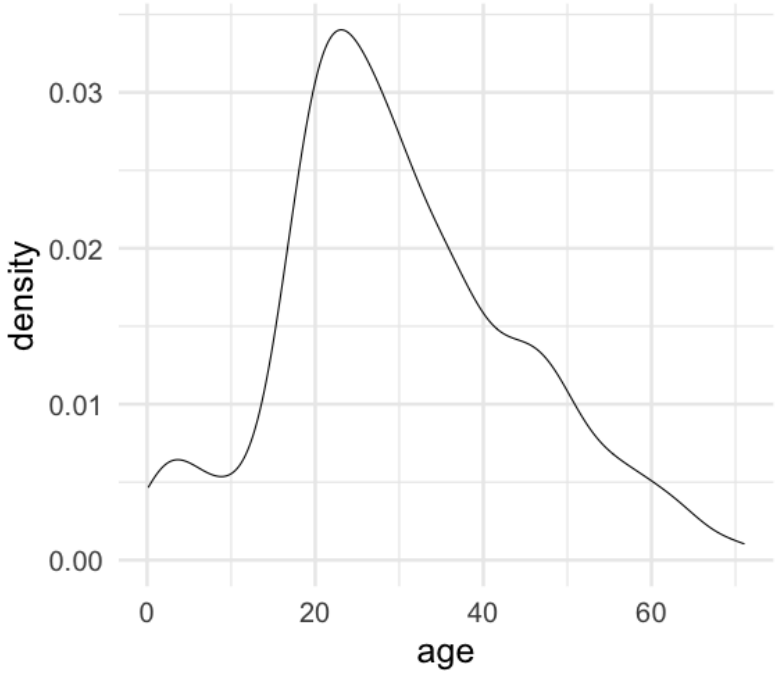

2004-05
2005-06
2006-07
2007-08
2008-09
2009-10
2010-11
2011-12
2012-13
2013-14
2014-15
2015-16
Average tution
6000 9000 12000 15000

Some things to avoid

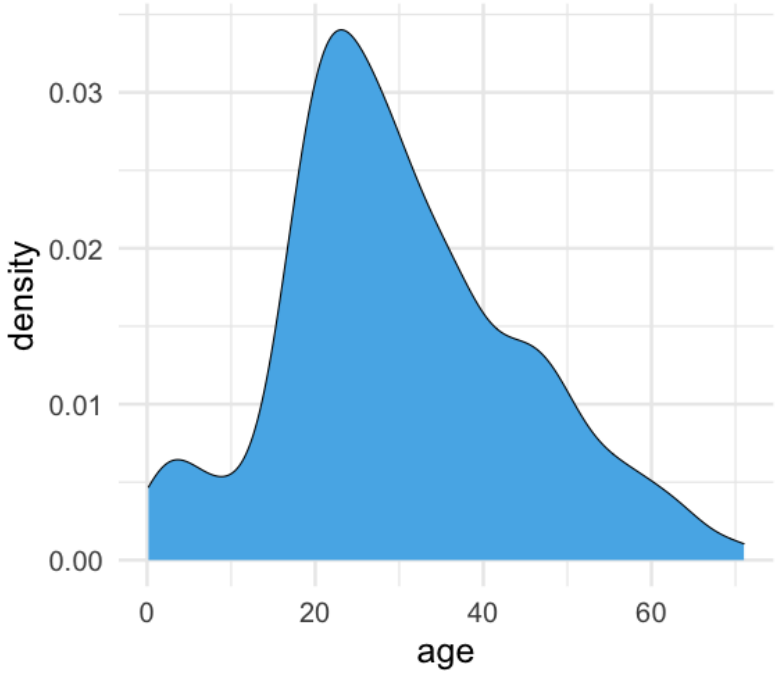
Line drawings

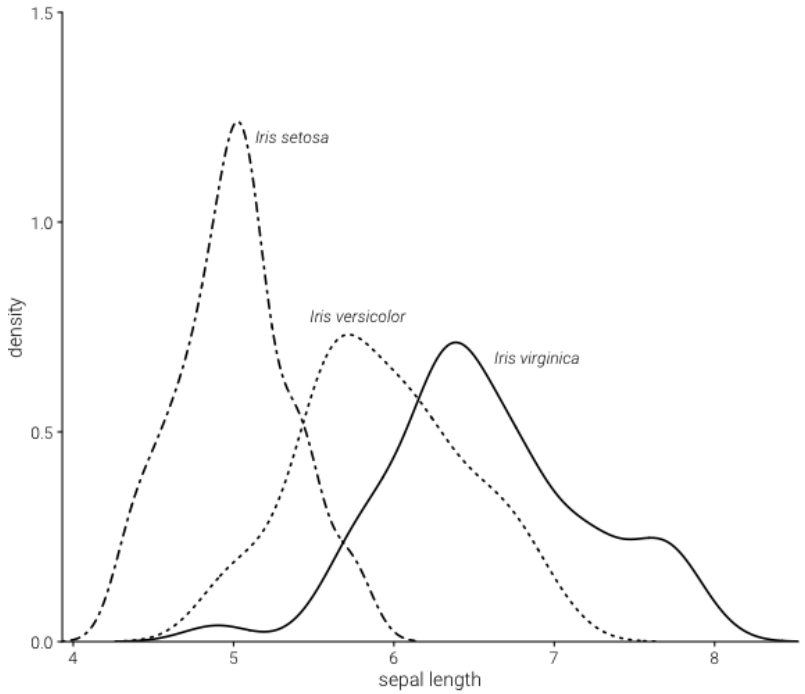
As discussed earlier

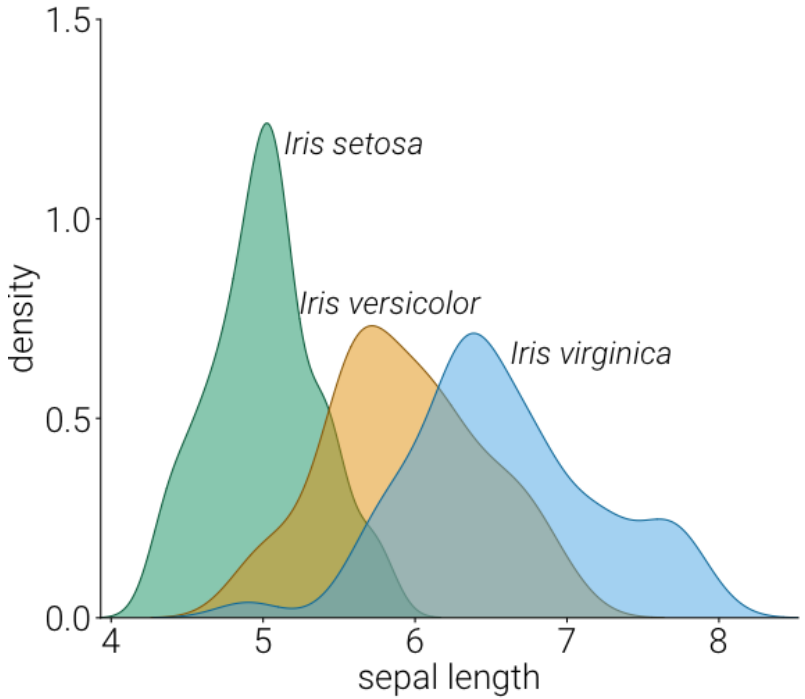
😫



Change the fill

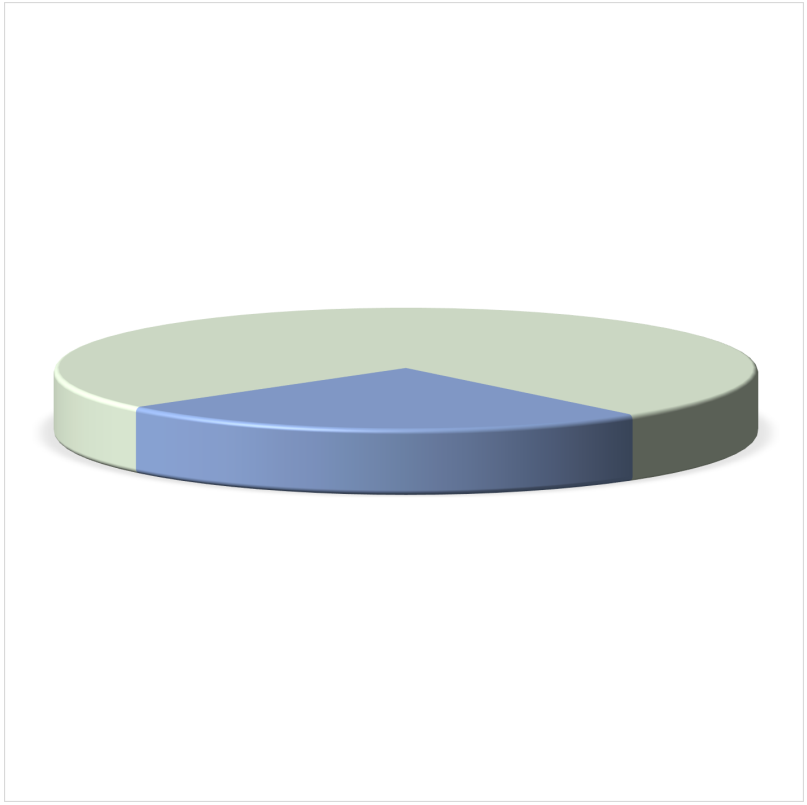


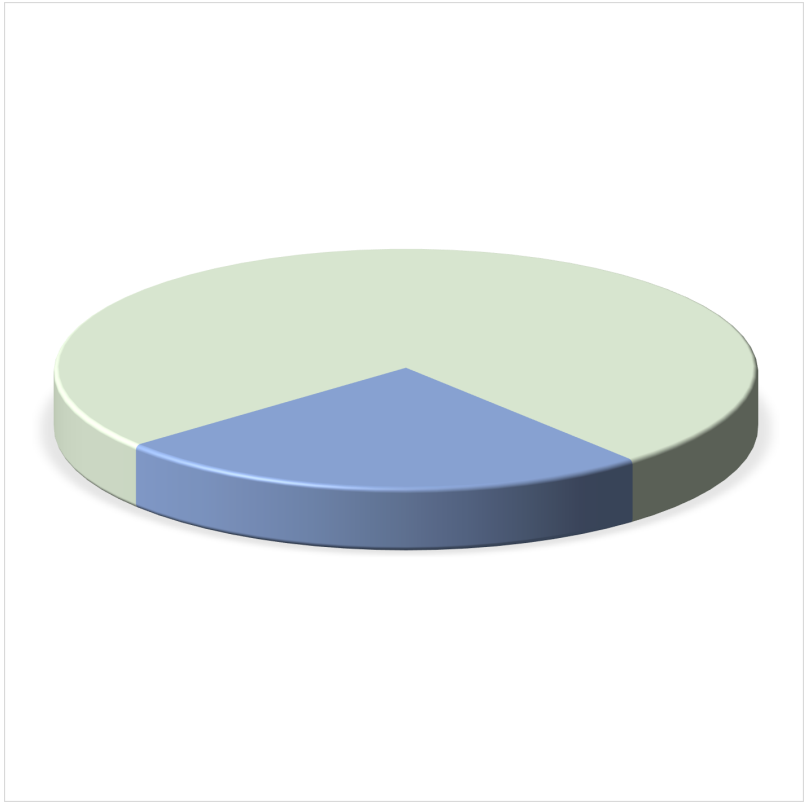

1.5
1.0
0.5
0.0
density
Iris setosa
Iris versicolor
Iris virginica
4
5
6
7
8
sepal length


1.5
1.0
0.5
0.0
density
Iris setosa
Iris versicolor
Iris virginica
4
5
6
7
8
sepal length

Much worse

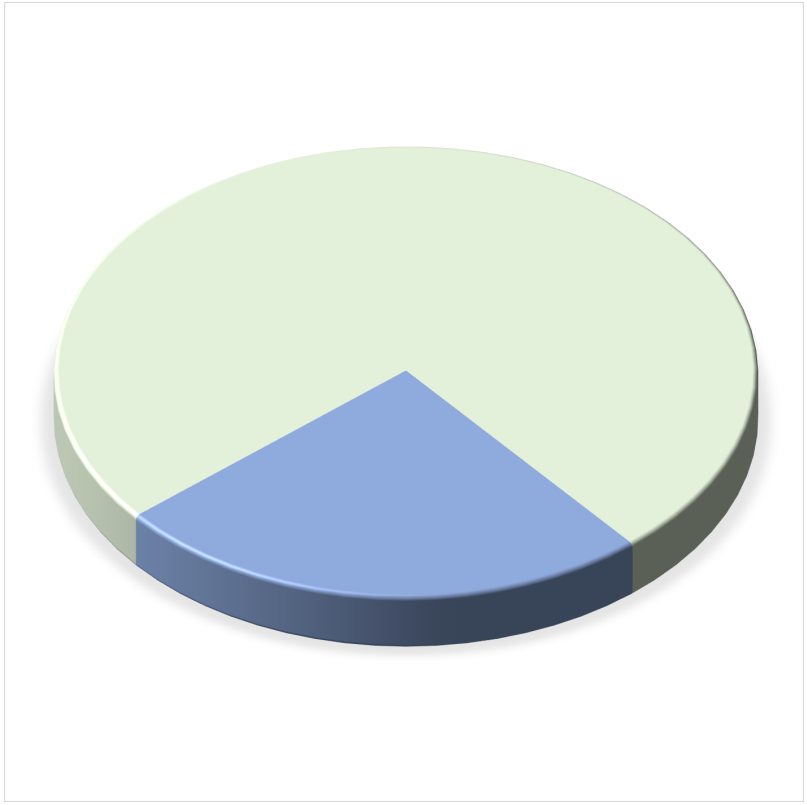
Unnecessary 3D

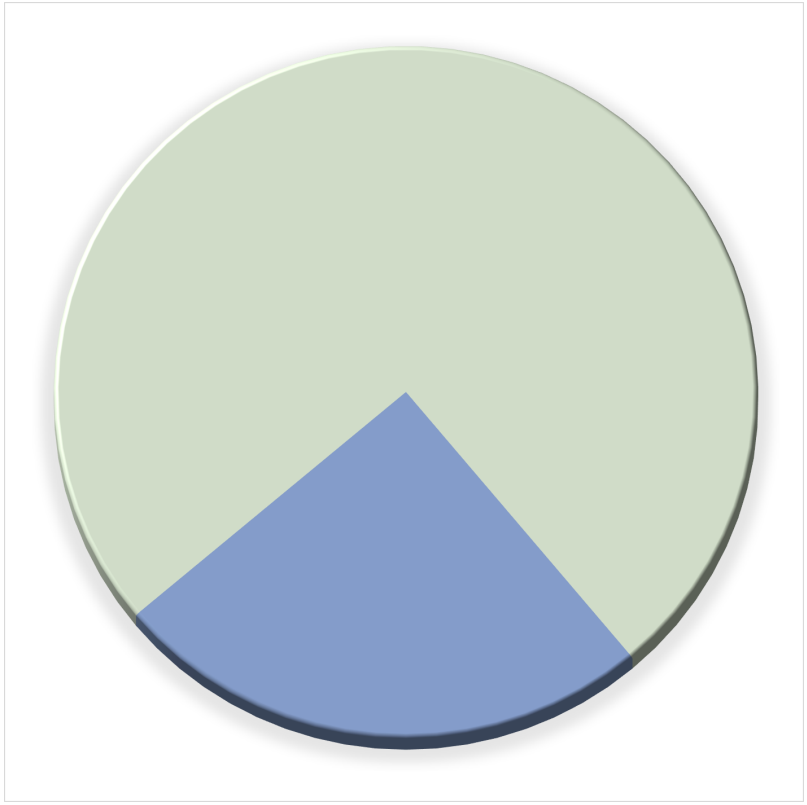




Much worse

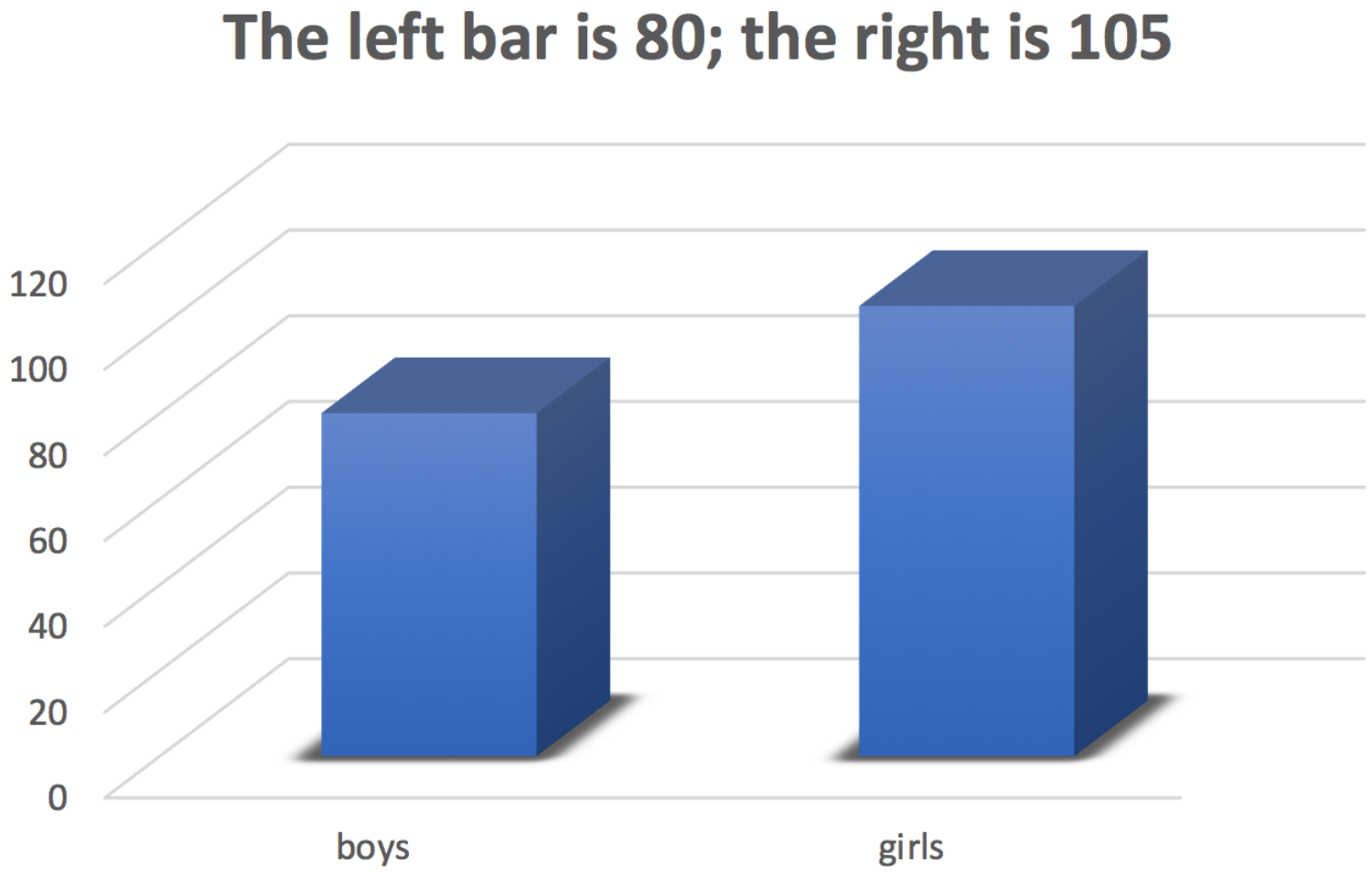
Unnecessary 3D





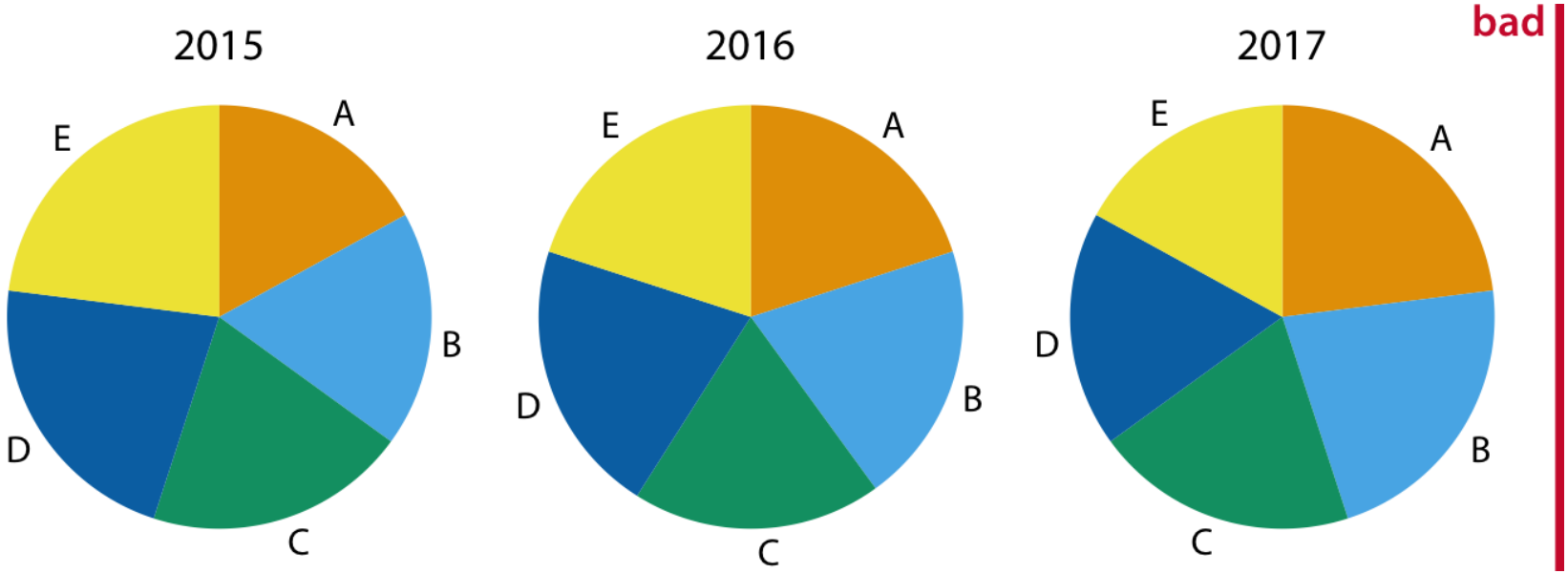
Horrid example

Used relatively regularly

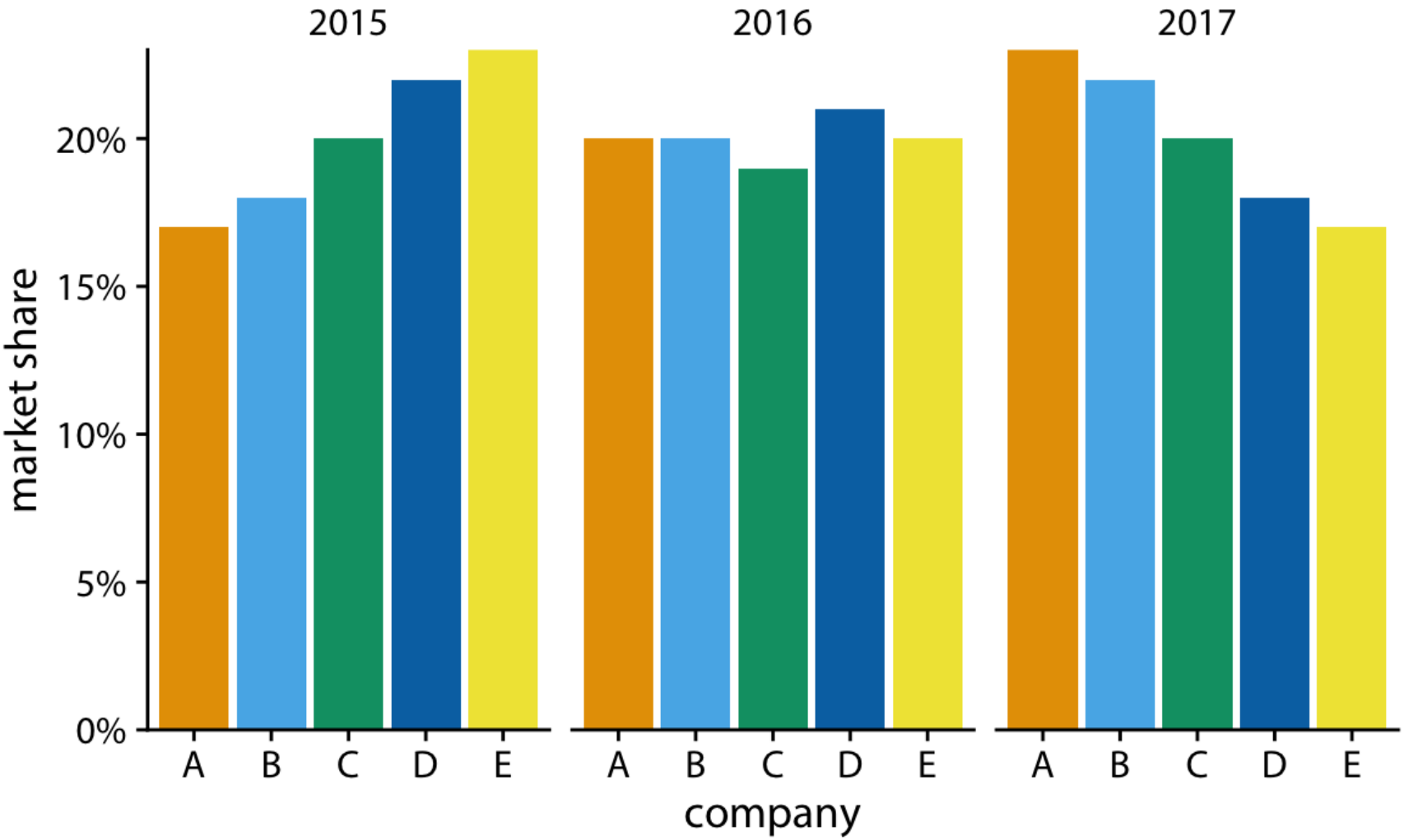


Pie charts

Especially w/lots of categories

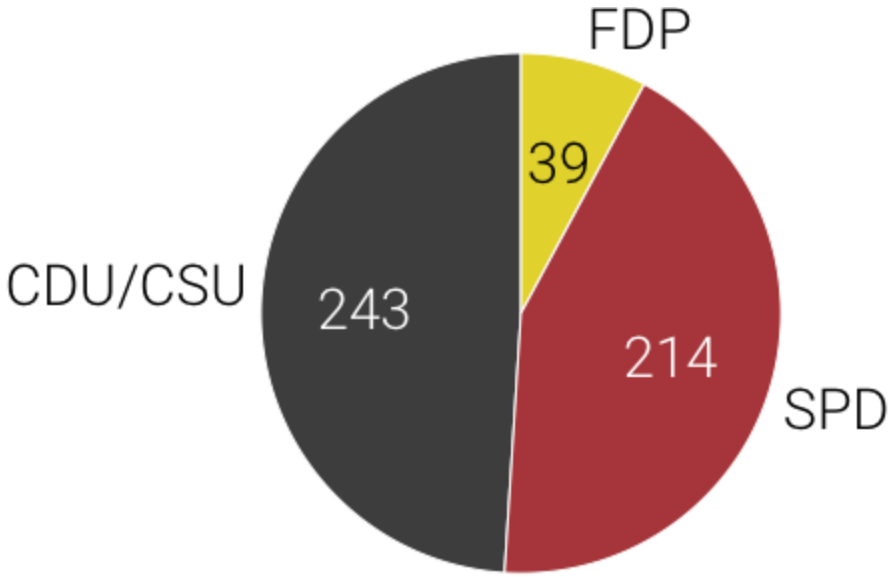


Alternative representation



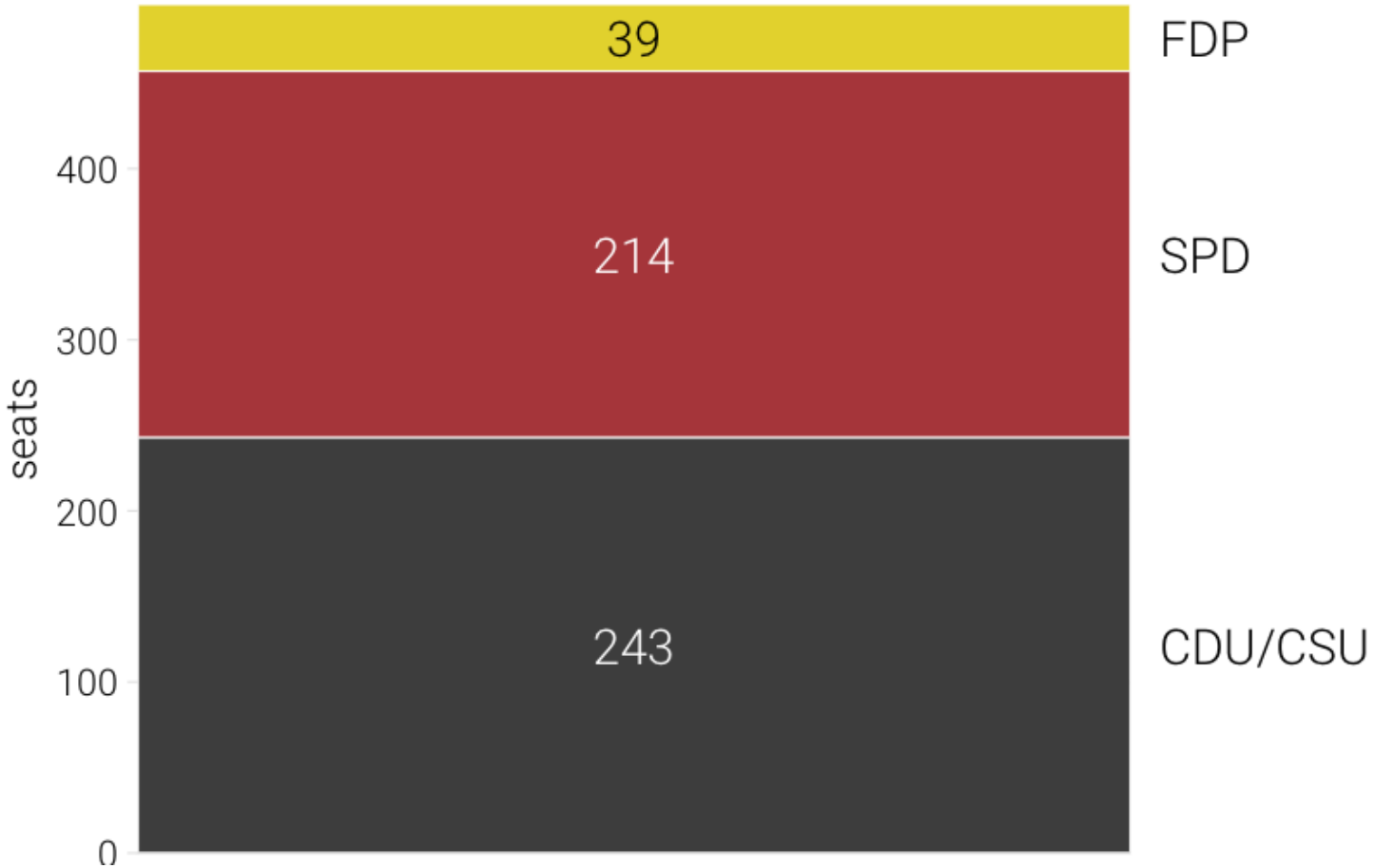
A case for pie charts

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

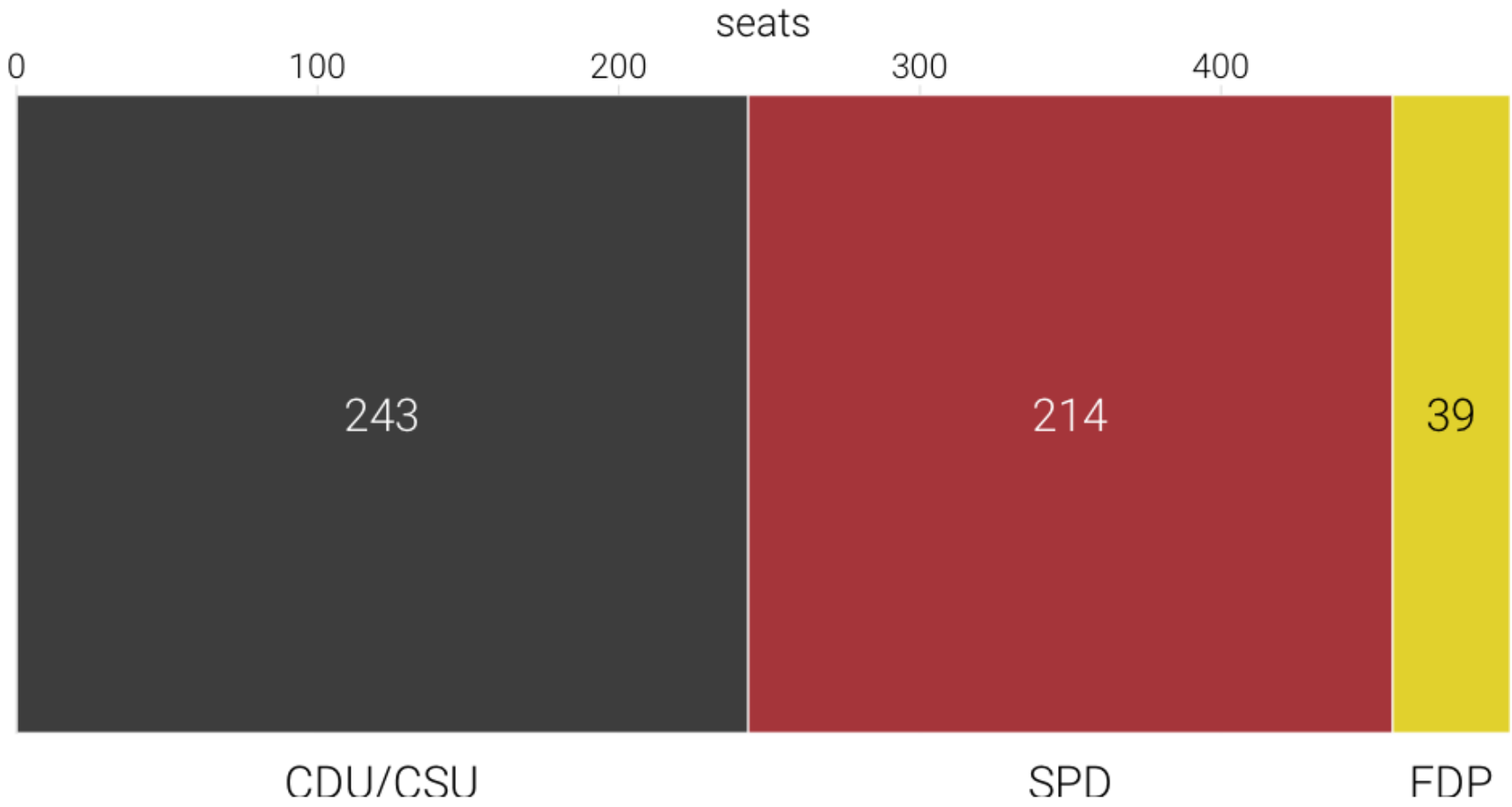


The anatomy of a pie chart

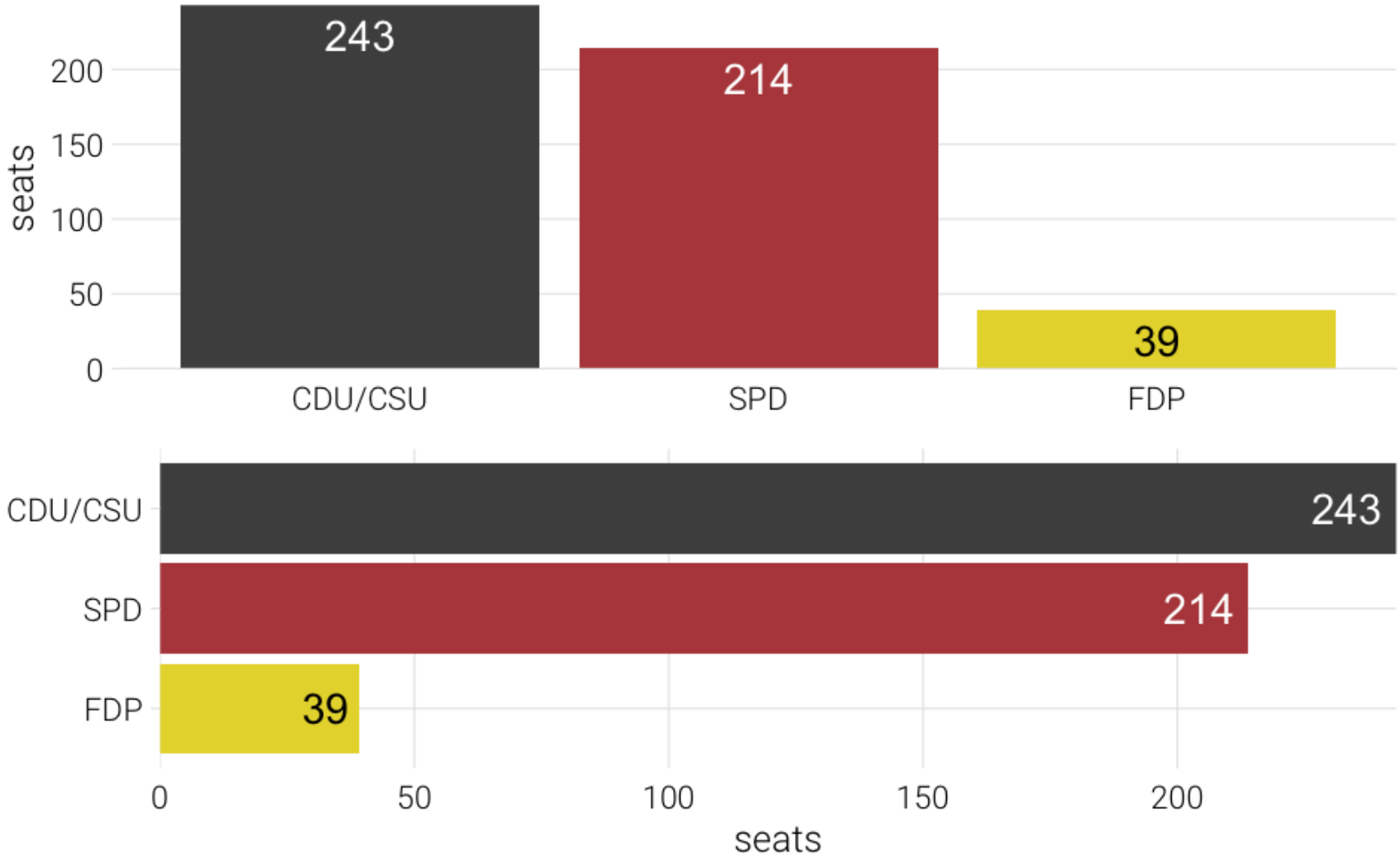
Pie charts are just stacked bar charts with a radial coordinate system



Horizontal

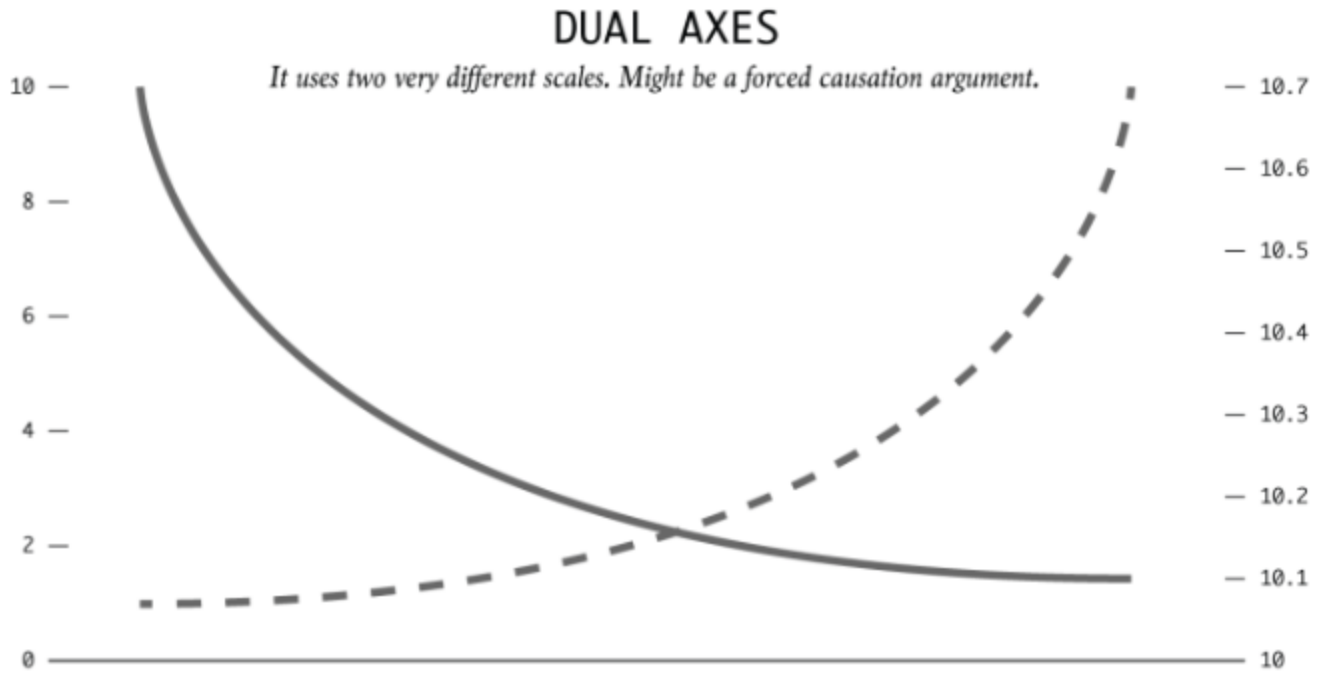


My preference



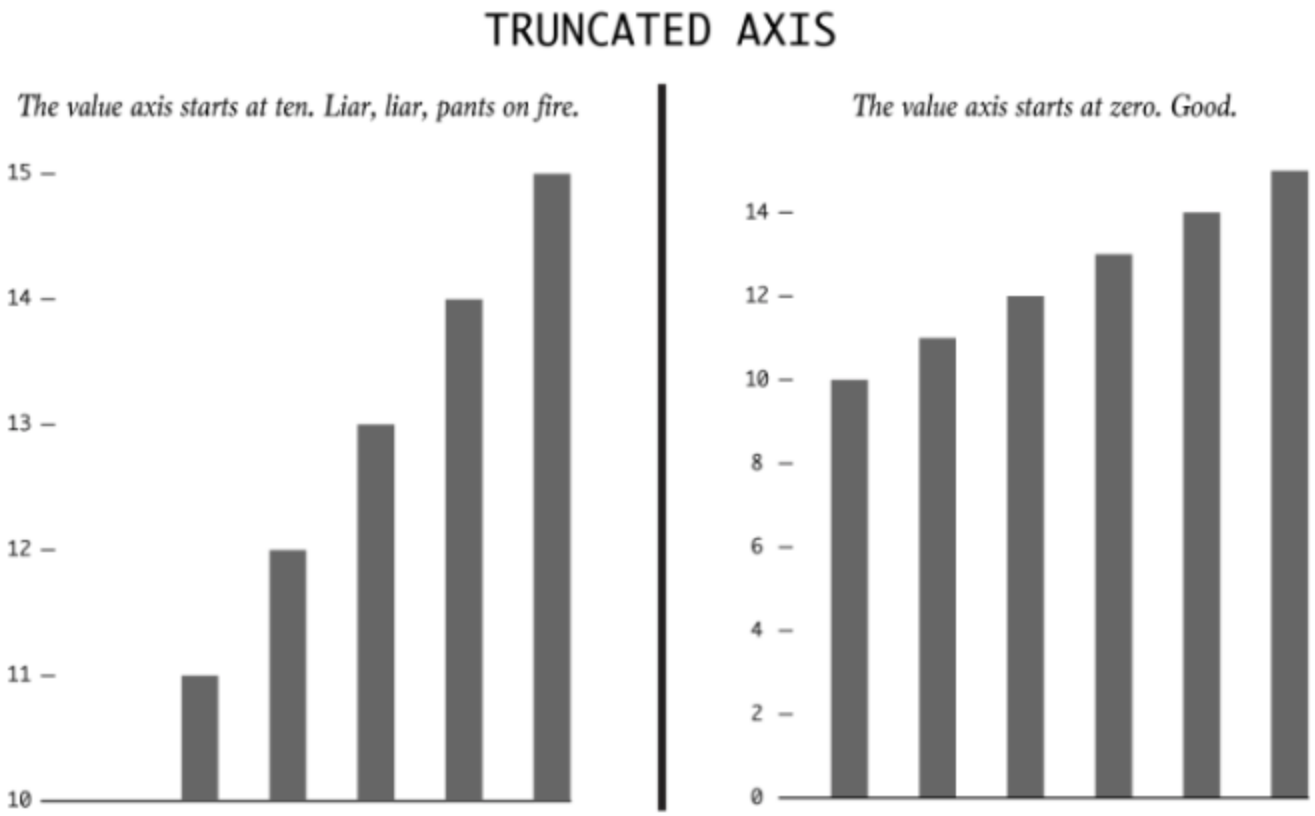
Dual axes

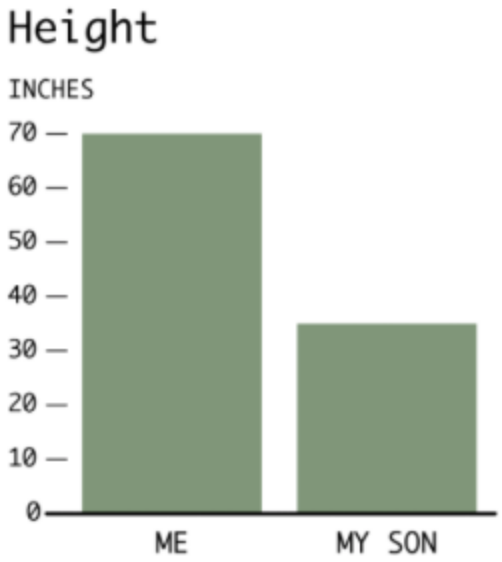
- Exception: if second axis is a direct transformation of the first



.footnote[See many examples here

Truncated axes



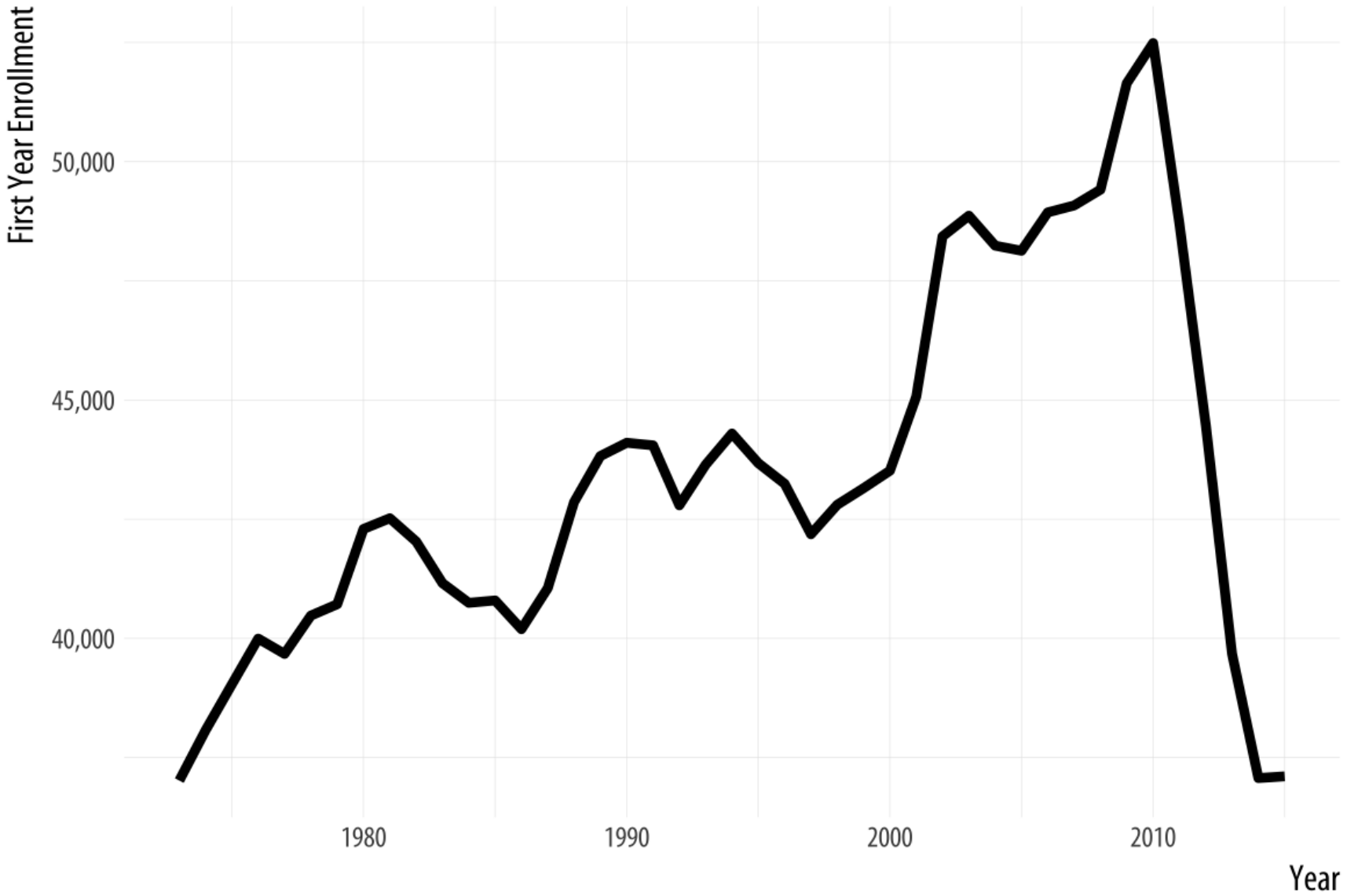

Height
INCHES
70
60
50
40
30
20
10
0
ME
MY SON

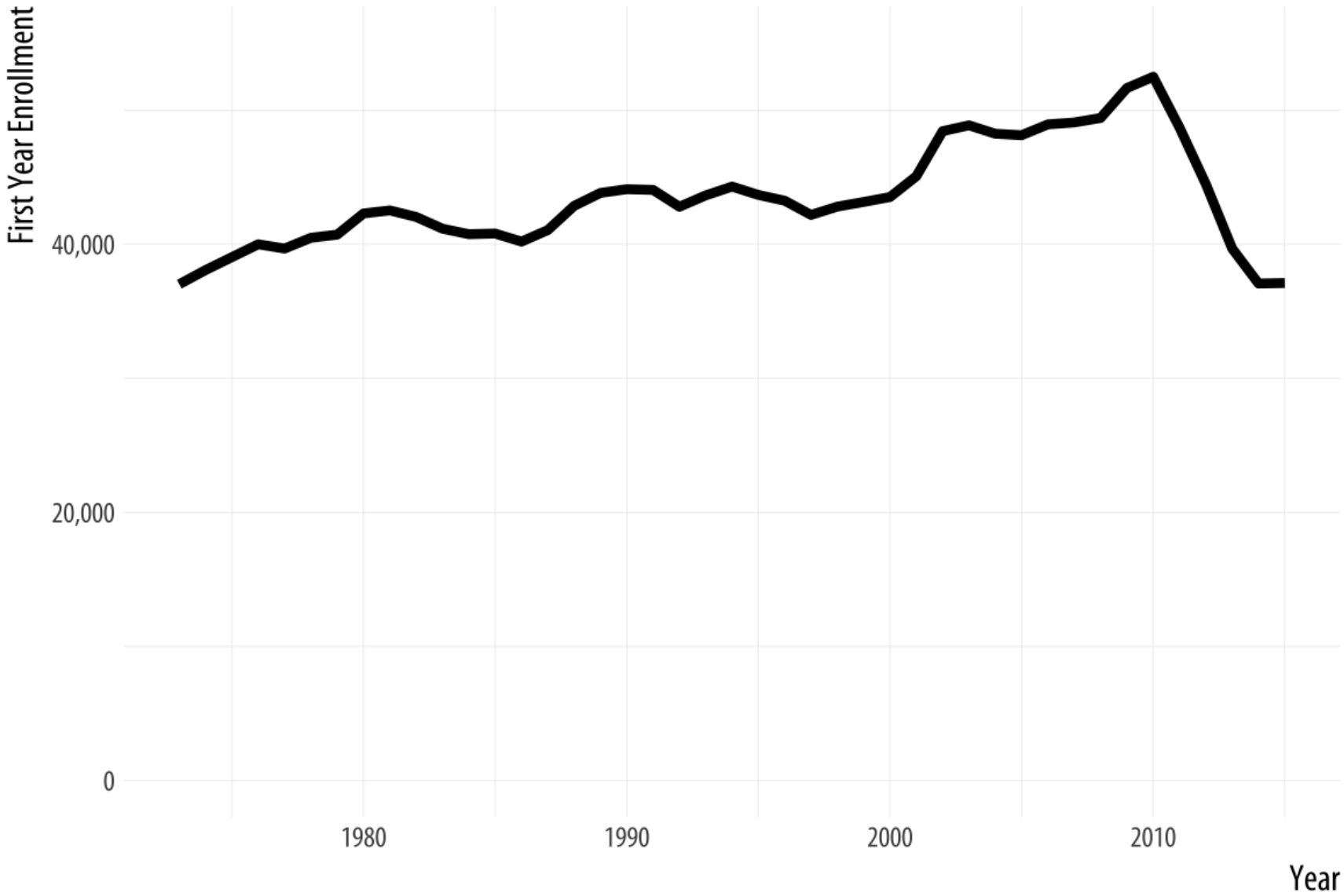
vs.


Height
INCHES
70
65
60
55
50
45
40
35
ME
MY SON
Where did he go?
Shoot. brb.

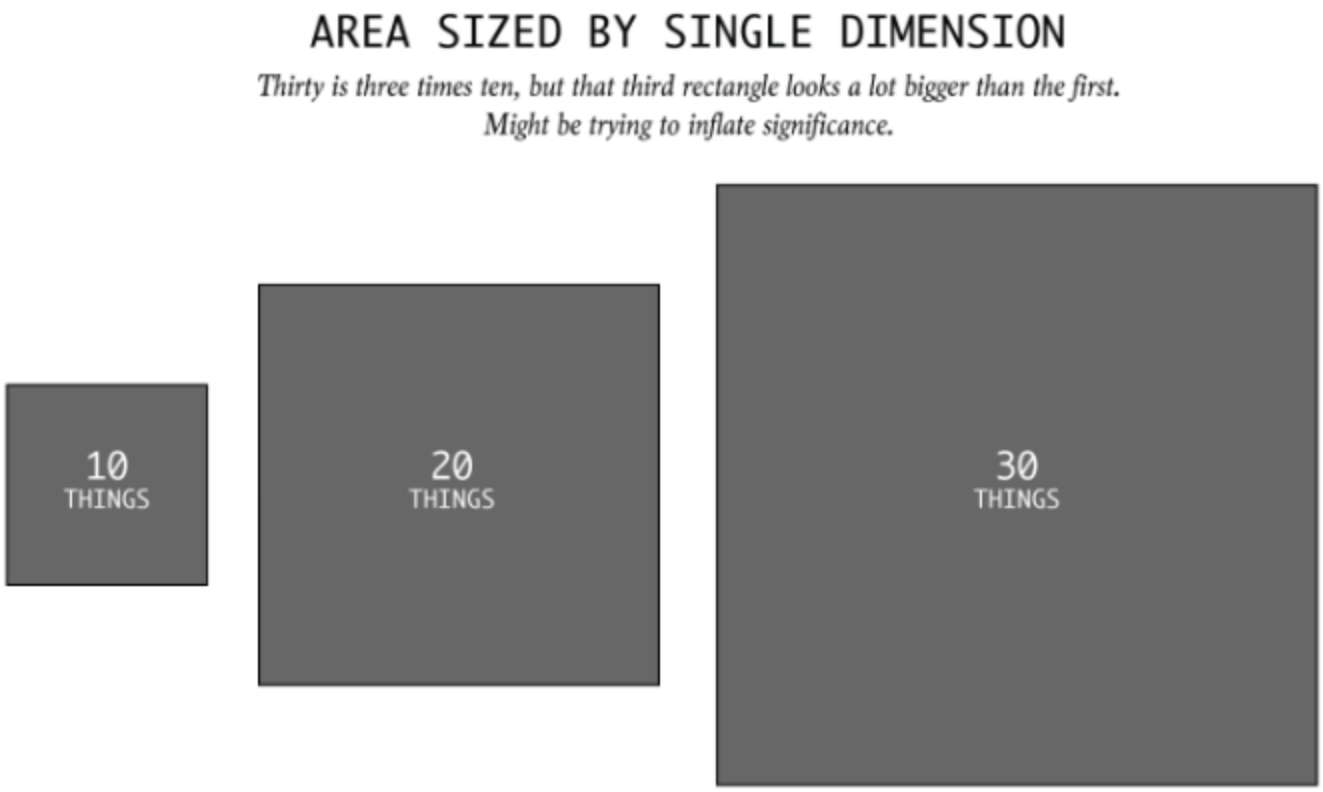
Not always a bad thing

> 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.

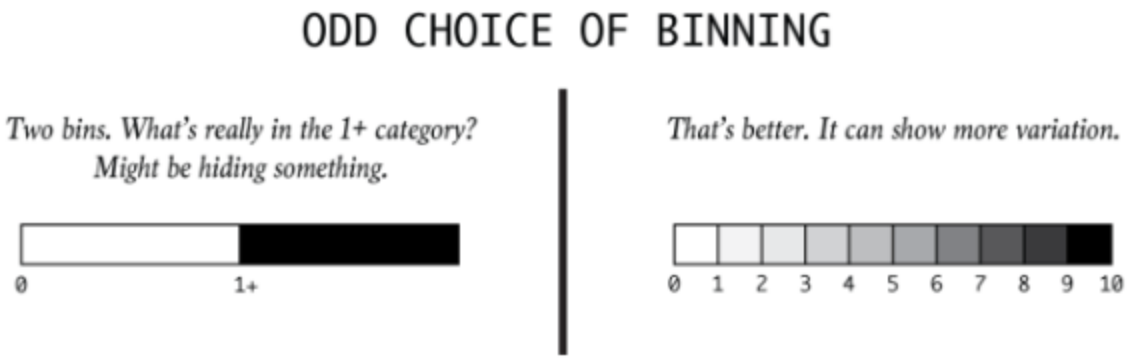

First Year Enrollment
50,000
45,000
40,000
1980
1990
2000
2010
Year


First Year Enrollment
40,000
20,000
0
1980
1990
2000
2010
Year

Scaling issues



Poor binning choices



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)

Thanks!

Questions?

@datalorax_

@datalorax

website

daniela@uoregon.edu