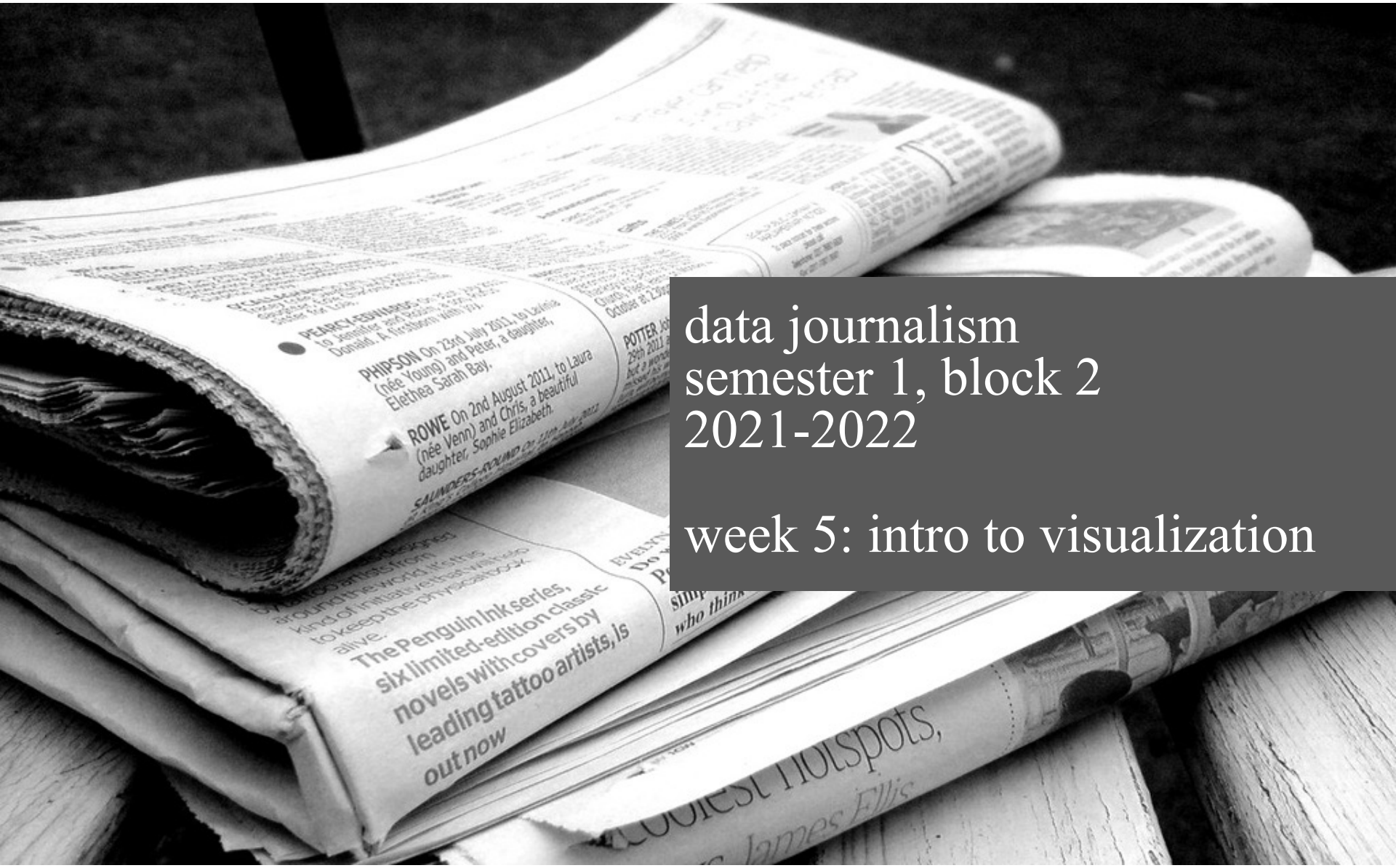


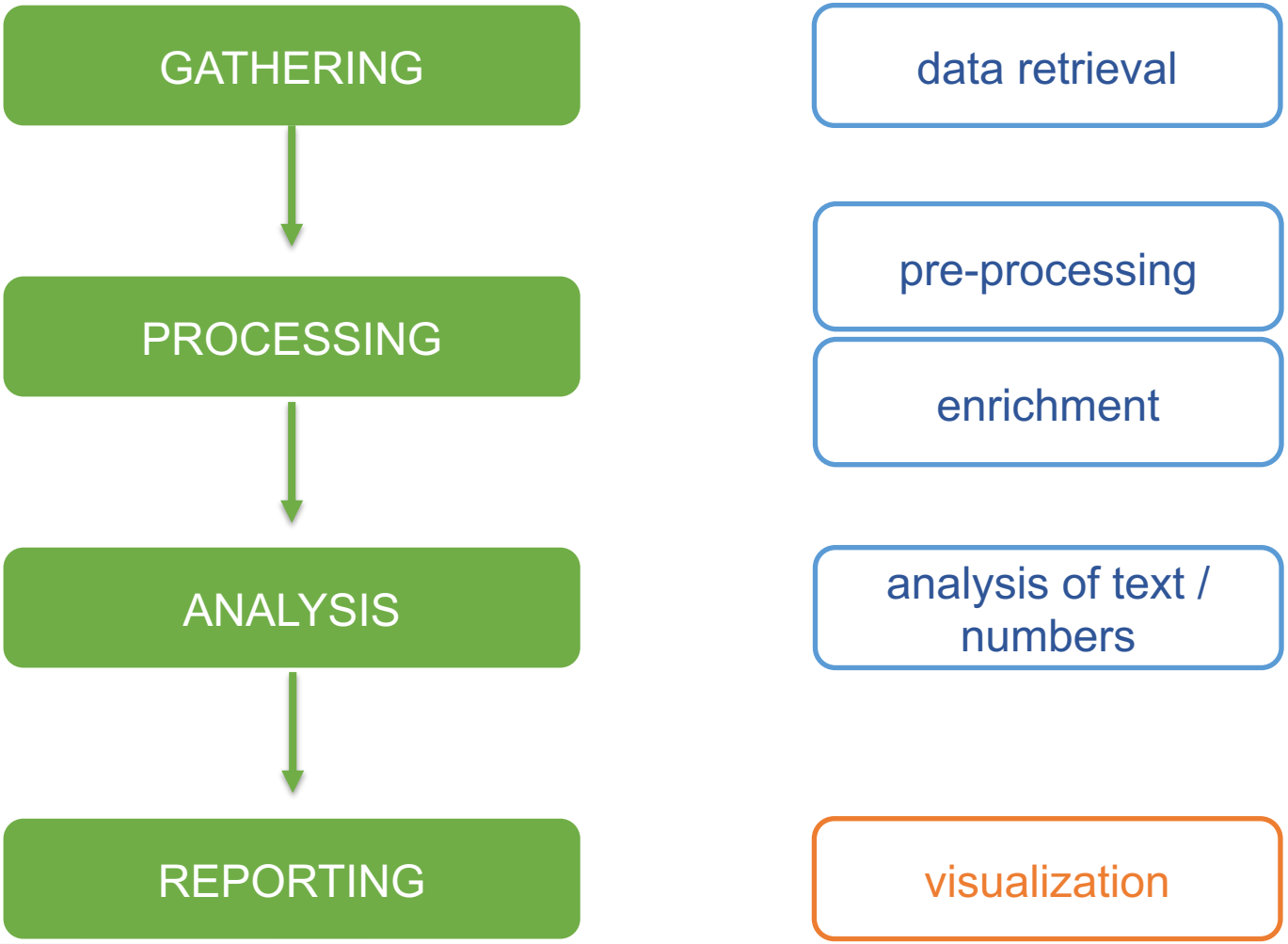
data journalism
semester 1, block 2
2021-2022

week 5: intro to visualization

first, a few announcements/tips

- rubric for final project published
- some extra office hours
  - 1 extra hour wednesdays
- but also: remember the discussion forum on canvas, for questions!
- reassurance about step 2 (let's look!)
  - if you haven't been able to get your data in order, probably time to choose something easier now
- remember to google. this is how a lot of your learning happens in this course.

DJ process & our workflow

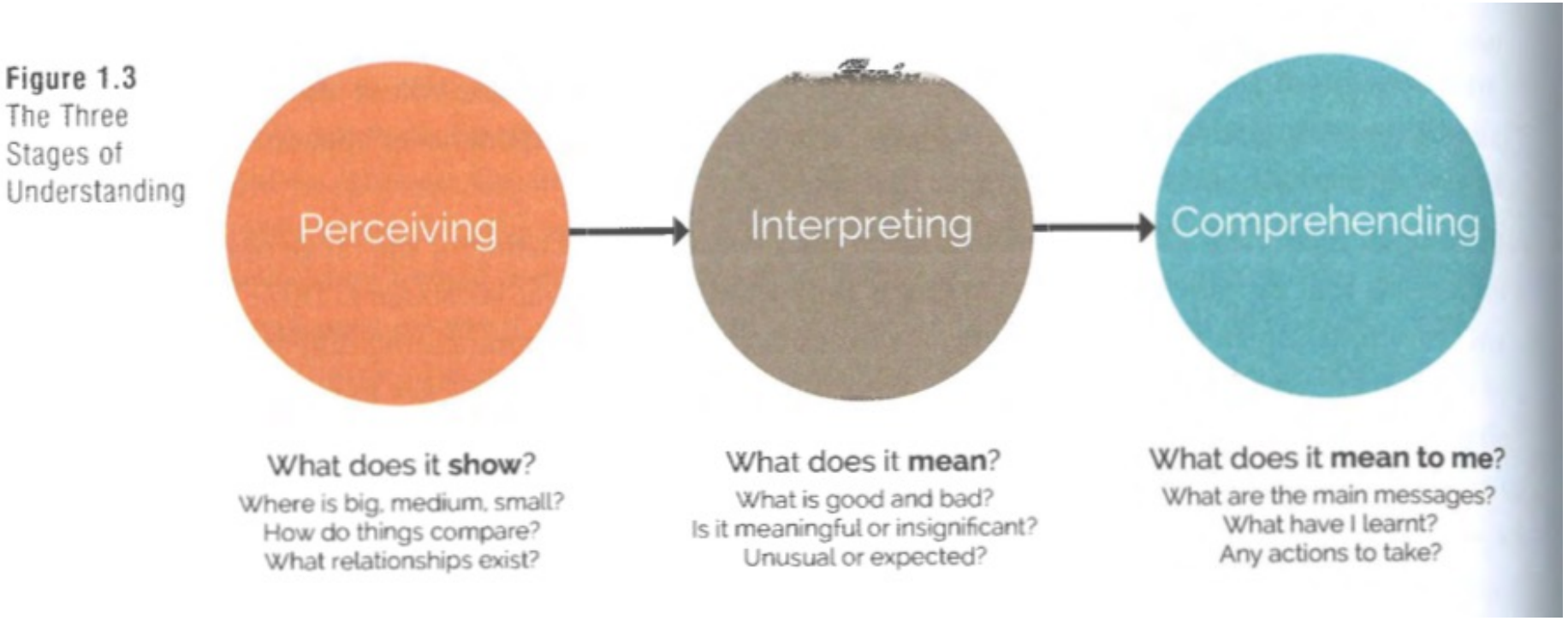


what is data visualization?

"The visual representation and presentation of data to facilitate understanding" (p. 15)

- **understanding** as a key element
- but what other goals might a journalist have?

understanding 'understanding'



Most visualizer control

Most viewer control

Kirk, A. (2016). *Data visualization. A handbook for data driven design*. London, UK: Sage.

effective visualizations are…

TRUSTWORTHY

- Is it reliable?
- Is the handling of the data reasonable & faithful?
- Does the design have integrity, and is it true?

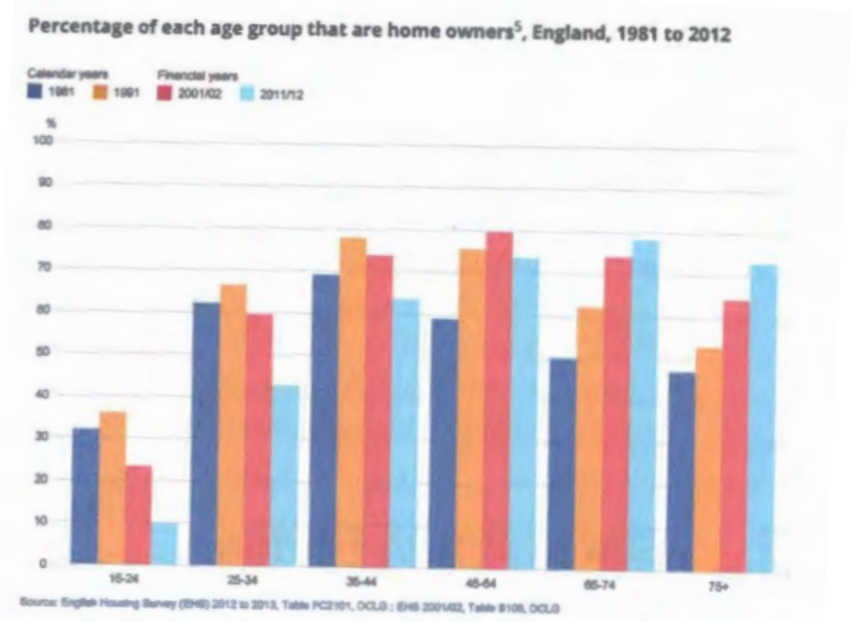


Figure 1.8 Housing and Home Ownership in the UK (ONS)

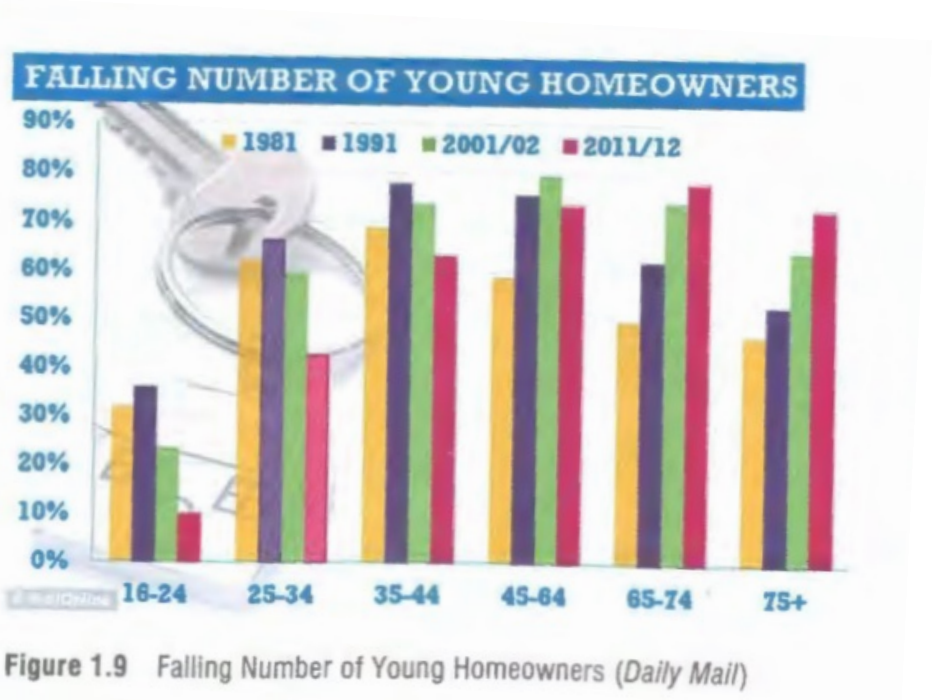


Figure 1.9 Falling Number of Young Homeowners (*Daily Mail*)

Kirk, A. (2019). *Data visualization. A handbook for data driven design.* 2nd edition. London, UK: Sage.

effective visualizations are…

TRUSTWORTHY

- Is it reliable?
- Is the handling of the data reasonable & faithful?
- Does the design have integrity, and is it true?

ACCESSIBLE

- Is it usable?
- Is the portrayal of the data/subject relevant?
- Is the representation suitably understandable?

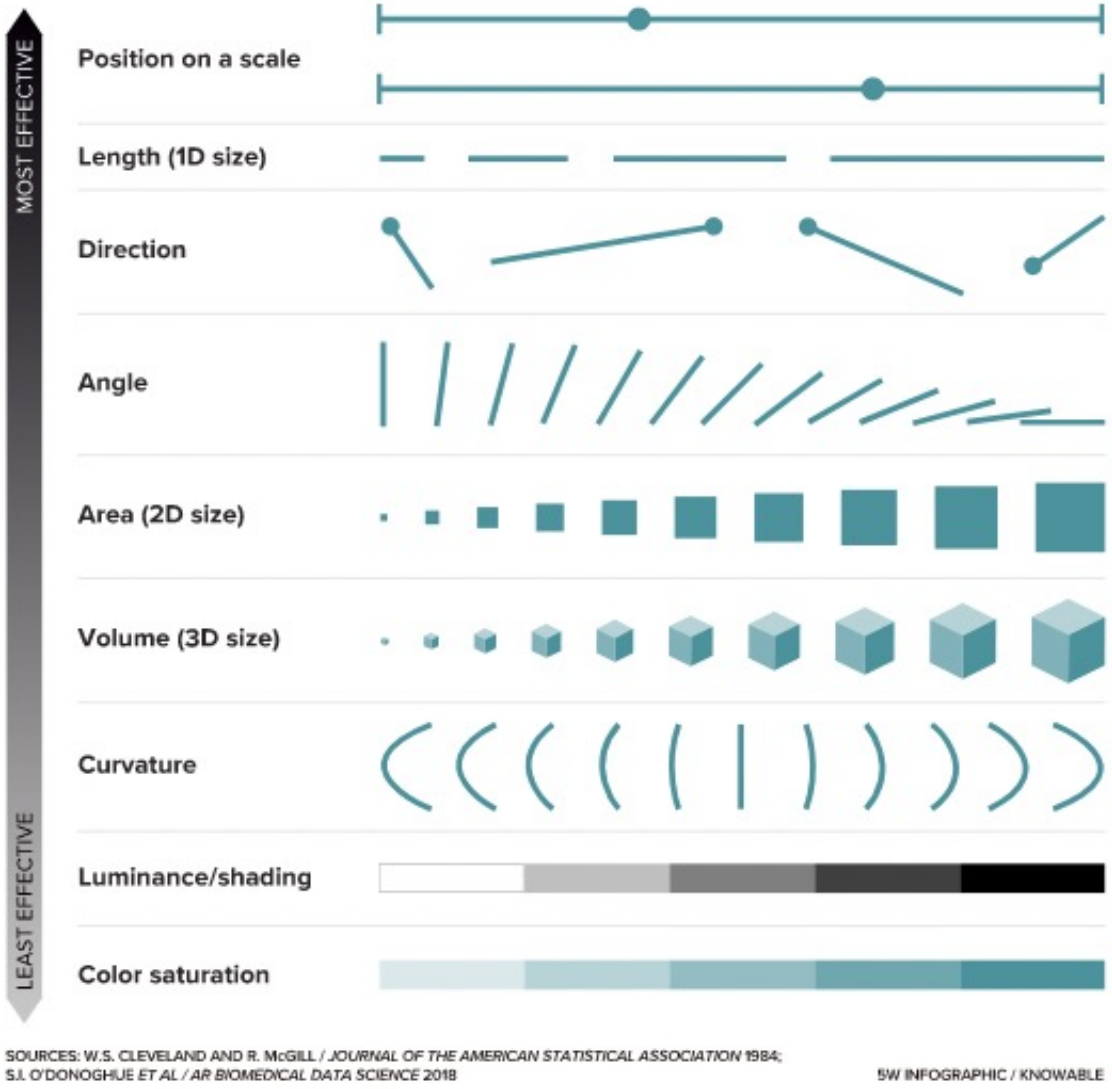
Kirk, A. (2019). *Data visualization. A handbook for data driven design.* 2nd edition. London, UK: Sage.

accessibility issues

- Voluntary versus necessary engagement
  - What do they want/need to learn?
- What domain knowledge is necessary?
  - How ‘literate’ can you expect your audience to be?
- Time available & placement
  - Is it a quick facts supplement, or a deeper exploration? When/where should audiences engage, and for how long?

Accessible scientific charts

- 1984/2018 studies
- What might have changed over time?



People are better at discerning subtleties in some types of visuals than others — the length of two lines, for example, or the direction of a line are easier to tell apart than shades of gray or the intensity of a color. Studies show that graphs using visual elements high on this list are easier to read and more effective than those near the bottom.

https://www.knowablemagazine.org/article/mind/2019/science-data-visualization

effective visualizations are…

TRUSTWORTHY

- Is it reliable?
- Is the handling of the data reasonable & faithful?
- Does the design have integrity, and is it true?

ACCESSIBLE

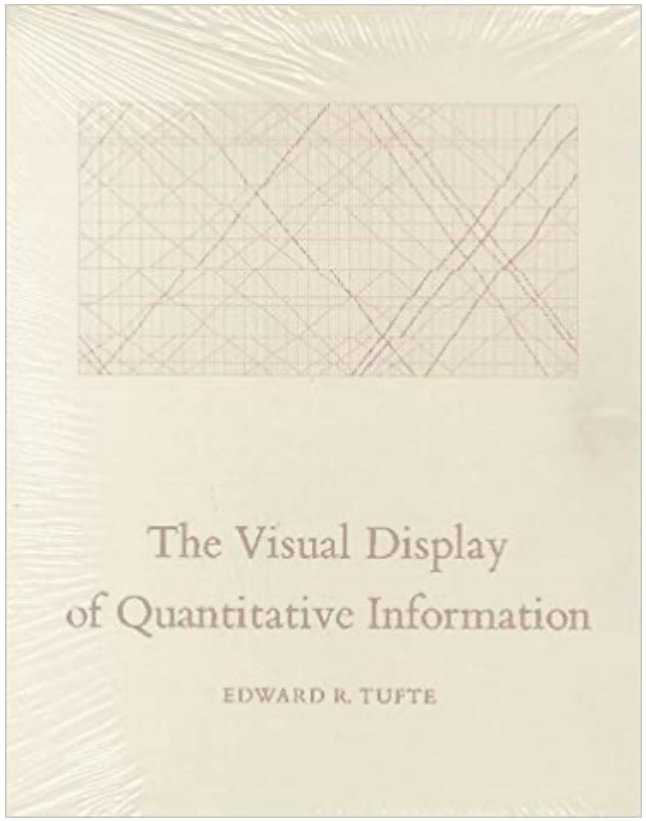
- Is it usable?
- Is the portrayal of the data/subject relevant?
- Is the representation suitably understandable?

ELEGANT

- Is it aesthetic?
- Does it eliminate the arbitrary?
- Does decoration enhance, not distract?

Kirk, A. (2016). *Data visualization. A handbook for data driven design.* London, UK: Sage.

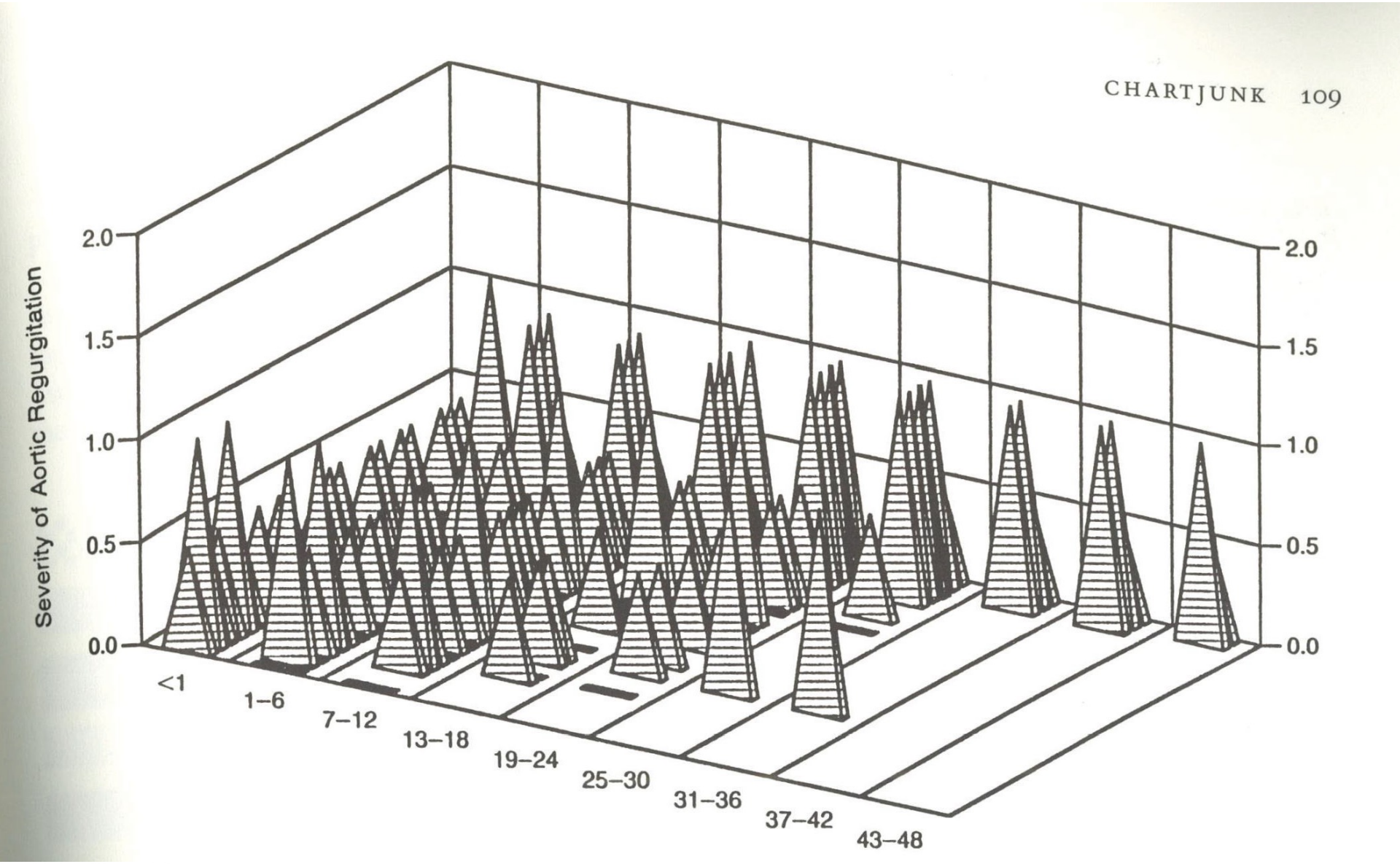
additional perspectives



A classic from 1982, still relevant today!

- beautiful illustrations
- theoretical arguments why to prefer some designs choices over others
- practical advice

Tufte: “chartjunk”

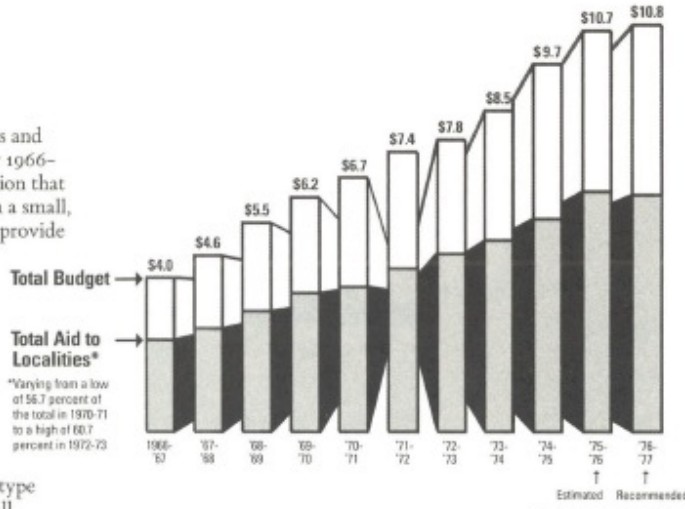


bad 3D, chartjunk, and trustworthiness

Despite the appearance created by the hyperactive design, the state budget actually did not increase during the last nine years shown. To generate the thoroughly false impression of a substantial and continuous increase in spending, the chart deploys several visual and statistical tricks—all working in the same direction, to exaggerate the growth in the budget. These graphical gimmicks:

These three parallelepipeds have been placed on an optical plane *in front* of the other eight, creating the image that the newer budgets tower over the older ones.

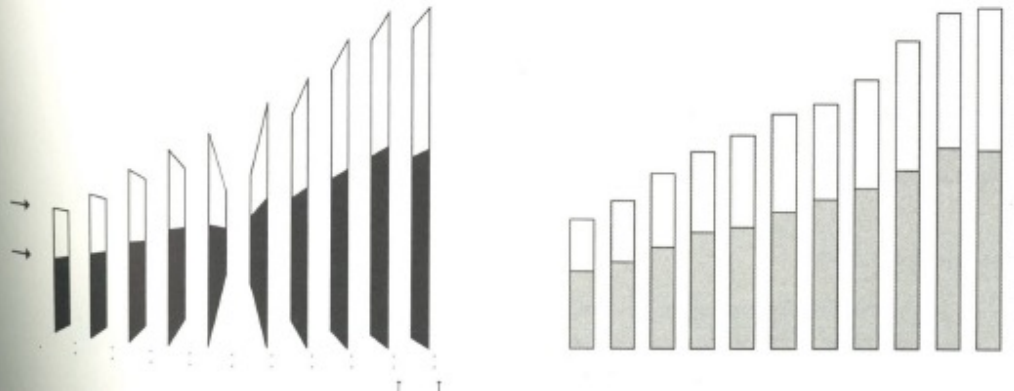
This cluster of type emphasizes and stretches out the low value for 1966–1967, encouraging the impression that recent years have shot up from a small, stable base. Horizontal arrows provide similar emphasis.



This squeezed-down block of type contributes to an image of small, squeezed-down budgets back in the good old days.

Arrows pointing straight up emphasize recent growth. Compare with horizontal arrows at left.

Leaving behind the distortion in the chartjunk heap at the left yields a calmer view:

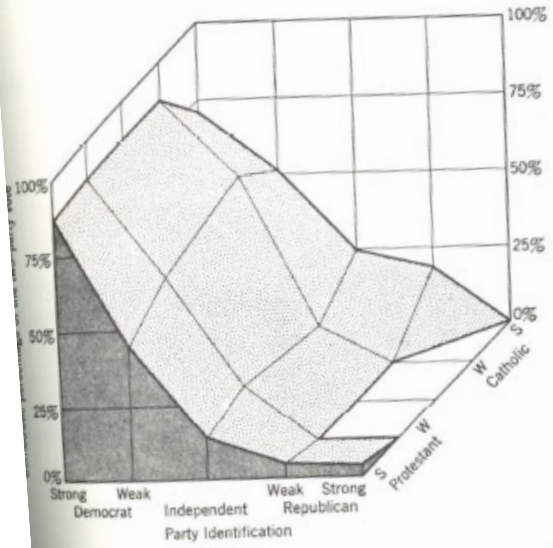


good 3D, and enhancement

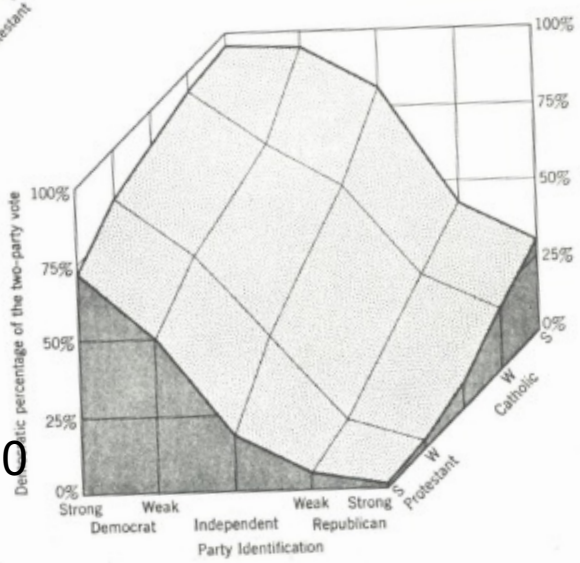
The grid that follows presents the data on the surface of the rock; on the sides, the grid is conventional. The two displays compare the effect of religion, taking into account party affiliation, on a person's vote for president in 1956 and in 1960 (when a Catholic ran for president). Note there is no reliable slope associated with religion in 1956, once party is controlled; in 1960, a systematic effect is found. Reading the slopes in the other direction shows the persistent effect of party in both elections:

Philip E. Converse, "Religion and Politics: The 1960 Election," in Angus Campbell, Philip E. Converse, Warren E. Miller, and Donald E. Stokes, *Elections and the Political Order* (New York, 1966), 102–103.

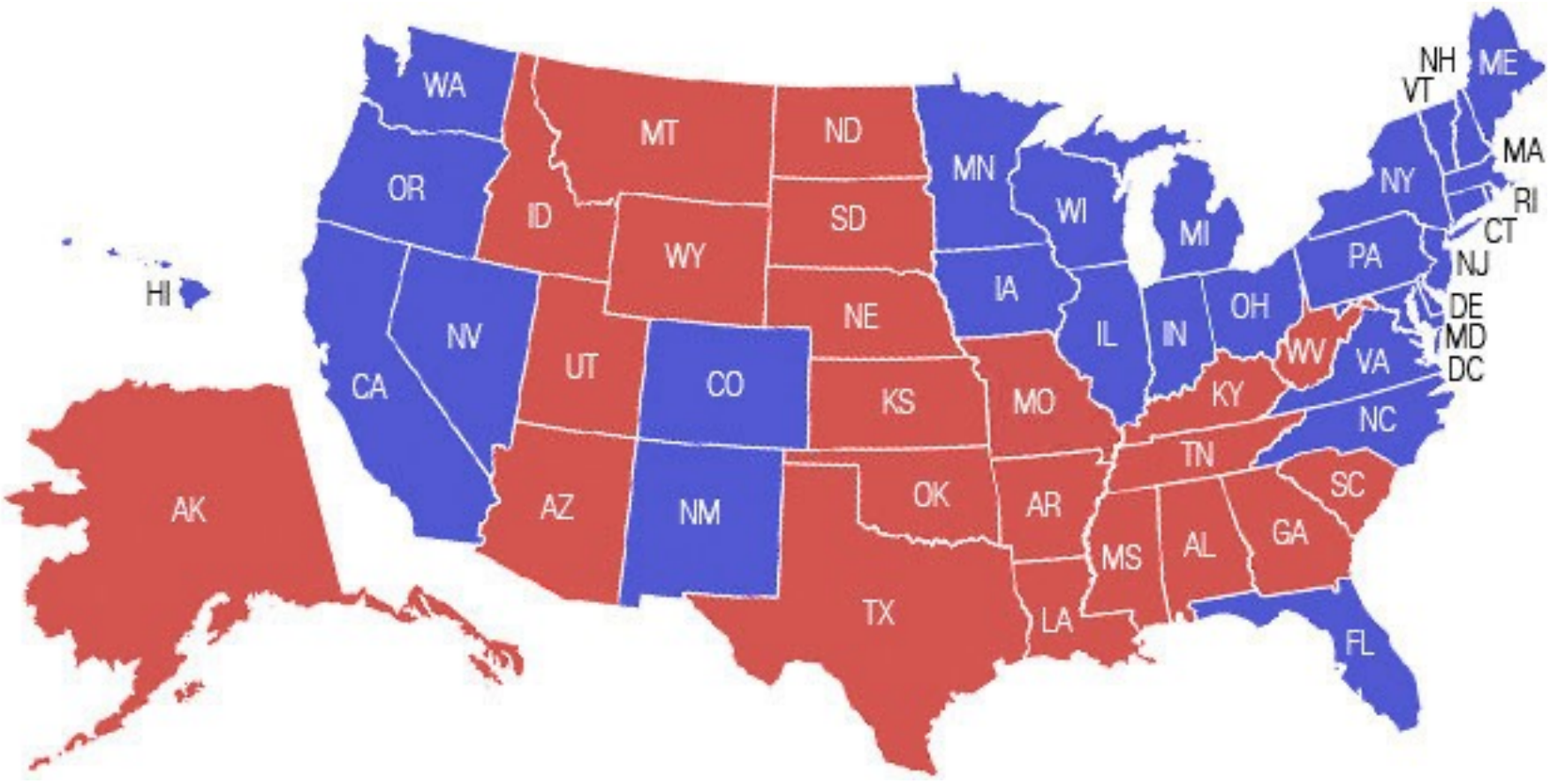
1956



1960

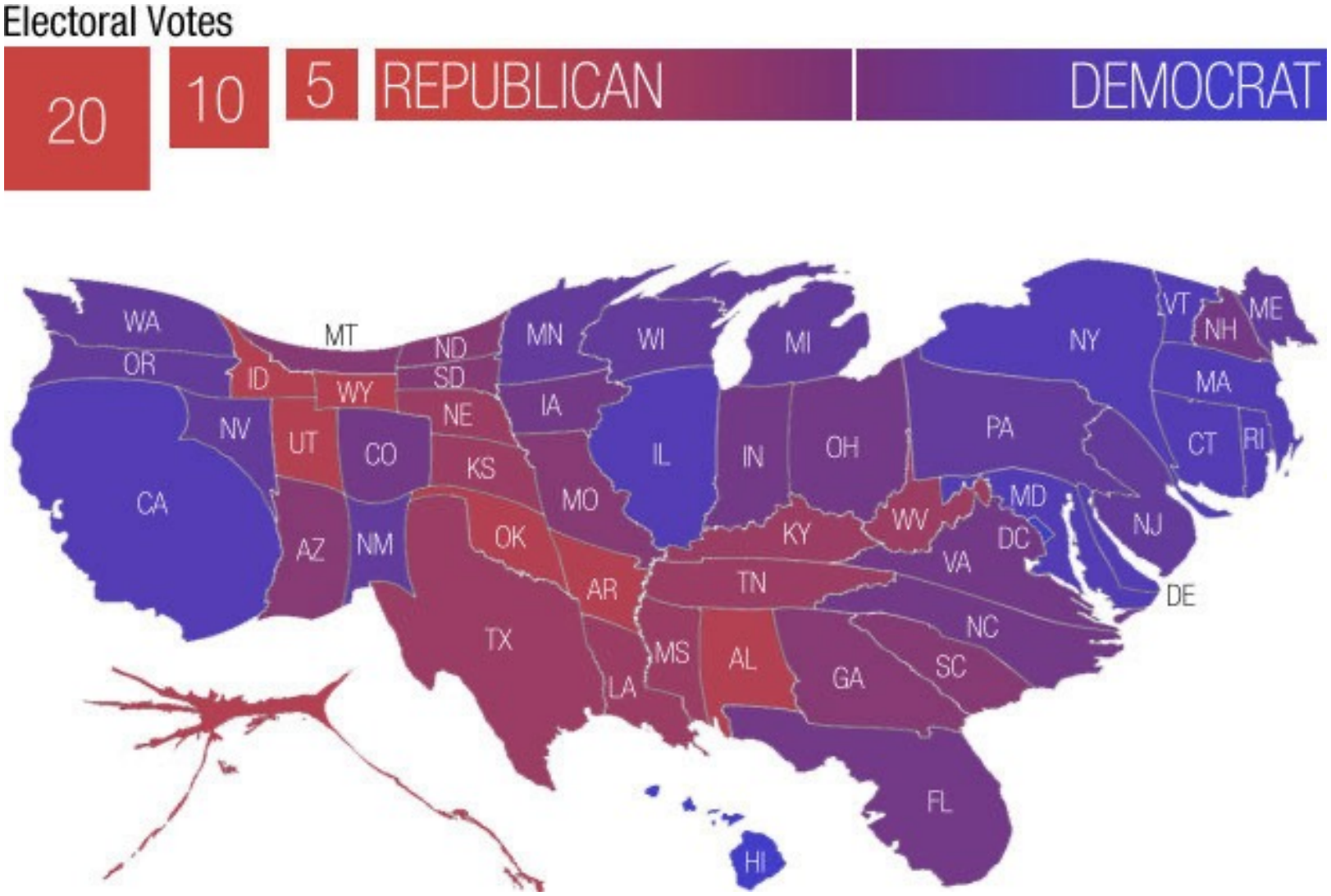


when distortion is actually trustworthy



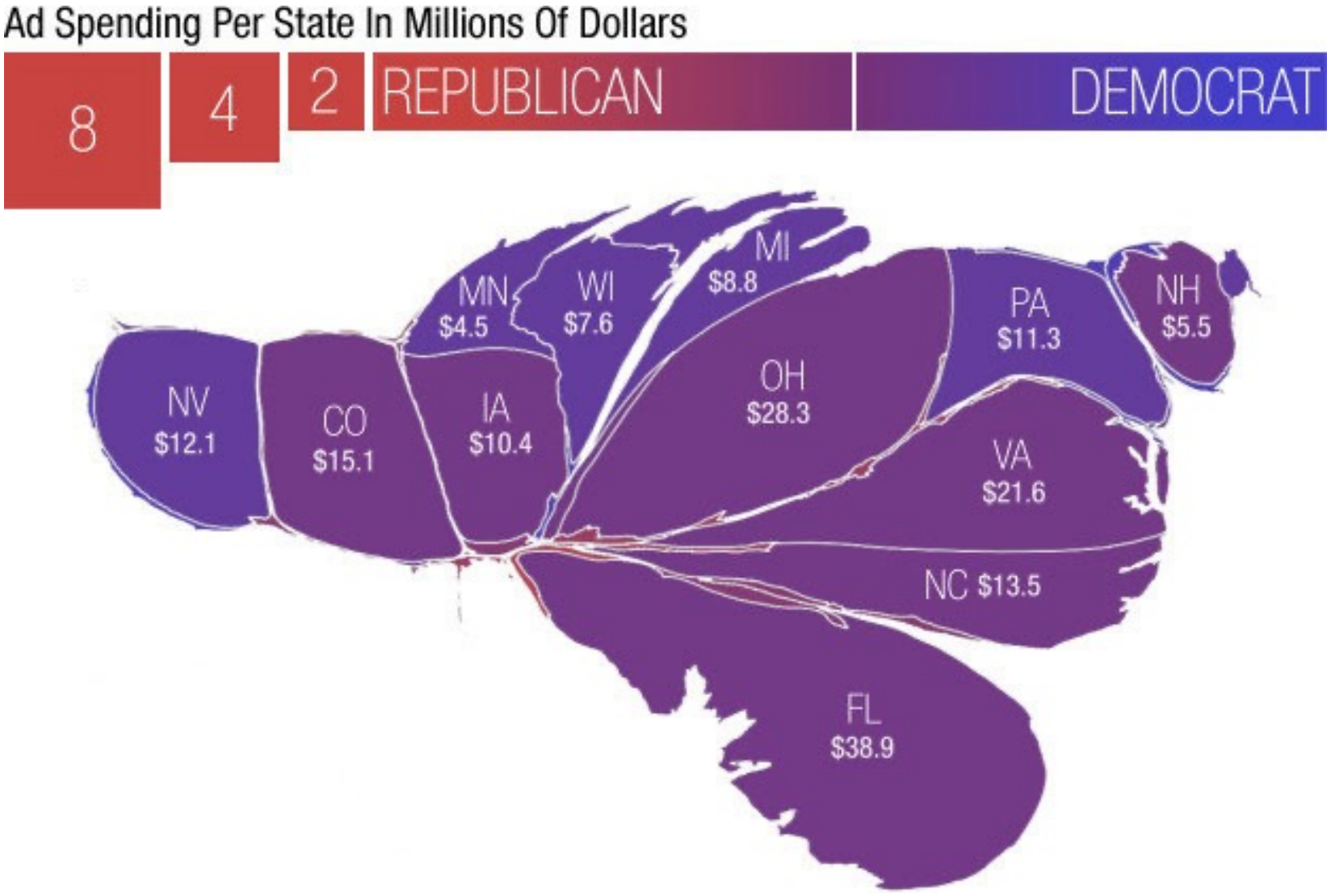
https://www.npr.org/sections/itsallpolitics/2012/11/01/163632378/a-campaign-map-morphed-by-money?t=1573564076759

when distortion is actually trustworthy



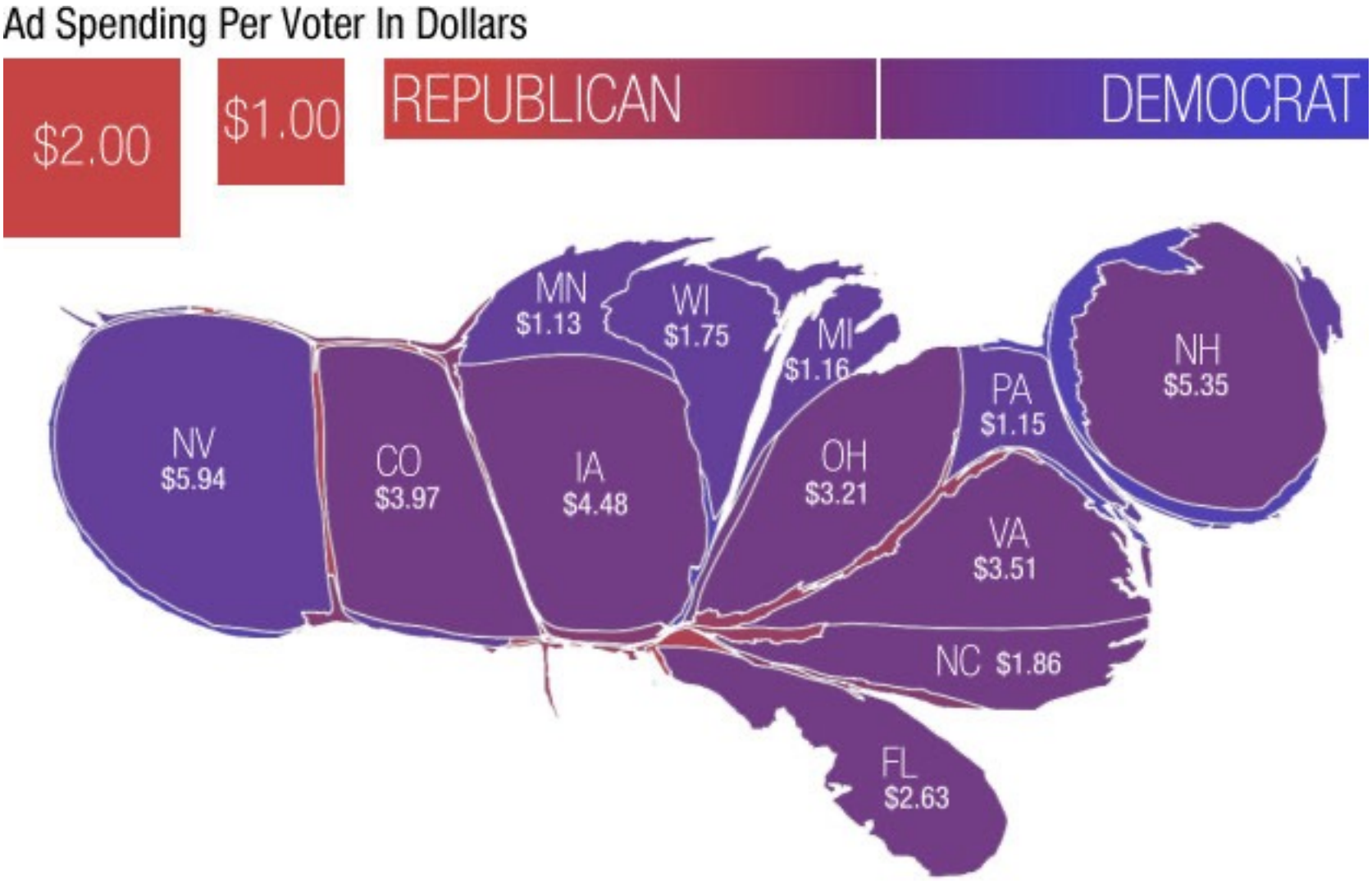
https://www.npr.org/sections/itsallpolitics/2012/11/01/163632378/a-campaign-map-morphed-by-money?t=1573564076759

when distortion is actually trustworthy



https://www.npr.org/sections/itsallpolitics/2012/11/01/163632378/a-campaign-map-morphed-by-money?t=1573564076759

when distortion is actually trustworthy



https://www.npr.org/sections/itsallpolitics/2012/11/01/163632378/a-campaign-map-morphed-by-money?t=1573564076759

and, awards for this

- https://www.informationisbeautifulawards.com/

- Also: https://informationisbeautiful.net/

Segel & Heer: the role of narrative & interactivity

- When the (causal) chain of events is important to your story
- Variety of genres found (in 2010... and now?)
- Author-driven versus reader-driven story types

Table 1. Properties of Author-Driven and Reader-Driven Stories. Most visualizations lie along a spectrum between these two extremes.

Author-Driven	Reader-Driven
Linear ordering of scenes	No prescribed ordering
Heavy messaging	No messaging
No interactivity	Free interactivity

- **Interactivity** can enhance, but shouldn't be goal in itself
  - Can allow some readers more info, while giving an overview to others
- No right answer, but variety of possible options

Segel, E., & Heer, J. (2010). Narrative visualization: Telling stories with data. *IEEE transactions on visualization and computer graphics*, 16(6), 1139-1148. DOI: 10.1109/TVCG.2010.179

Some additional tips, courtesy of DJH

- When NOT to use a visualization:
  - Story is better told through other means (multimedia, video, text)
  - Very few data points
  - Very little variation, or no clear trends/conclusions
  - When maps are misleading/space doesn't make sense
- And, don't forget about tables!
  - Good option for simple, clean presentation of simple data

Speaking of simplicity…

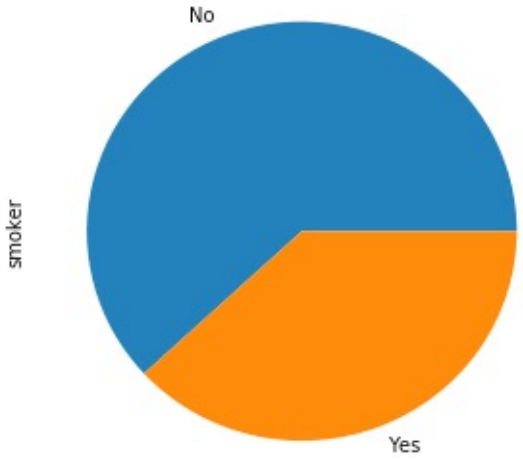
- For our course, we expect relatively simple visualizations
  - Should still be trustworthy, accessible, and elegant
  - But needn't be super fancy, interactive
- Can always "scale it up" if you want
  - Fancier modules
  - Group project focus

Let's look at some basic types

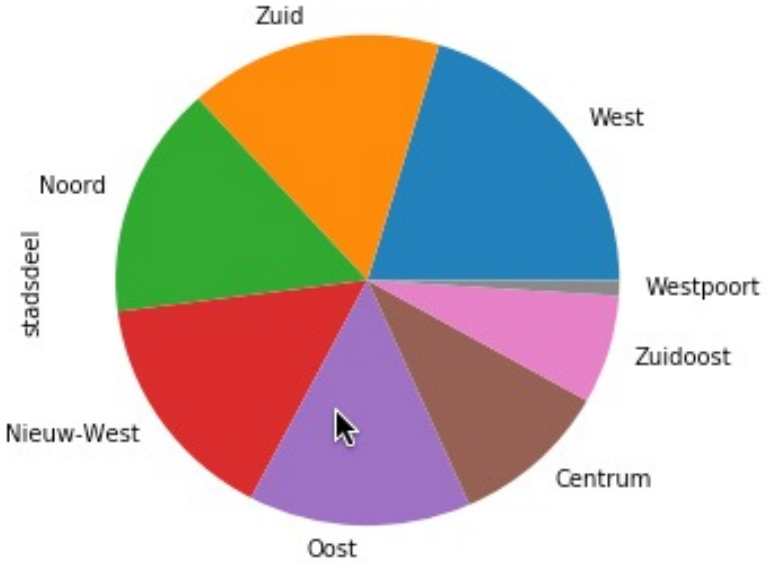
- Today, critique & think about utility of:
  - Pie charts
  - Bar charts
  - Point charts
  - Line charts
  - Scatterplots

- Thursday: learn how to make these (and others)

pie charts: when useful? when problematic?



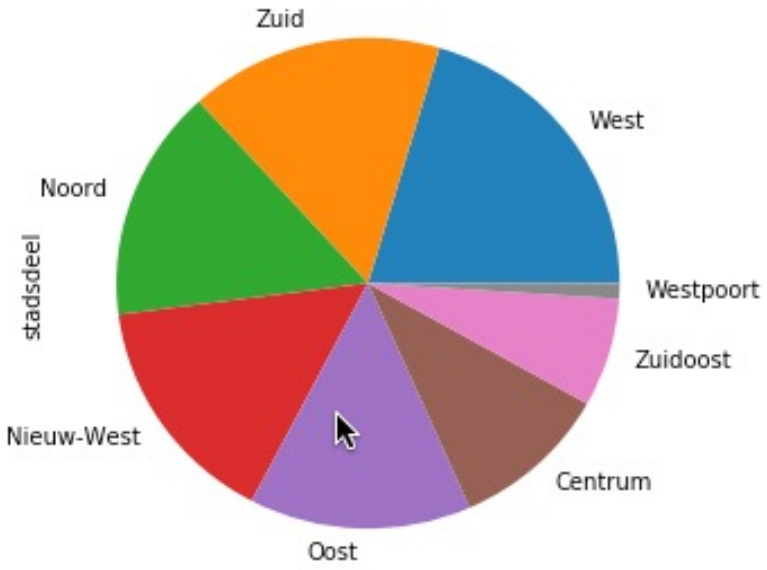
versus



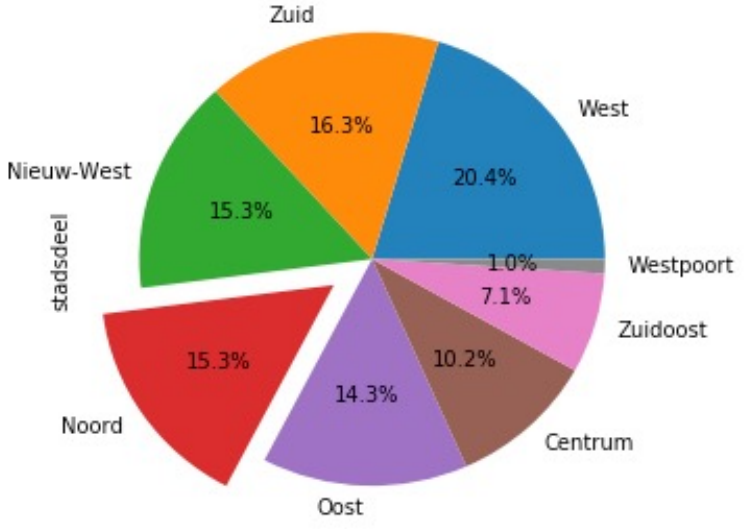
When useful?

- Proportions
- Categorical (nominal) data
- Only makes sense if add up to 100%!

pie charts: level of detail: helpful or not?

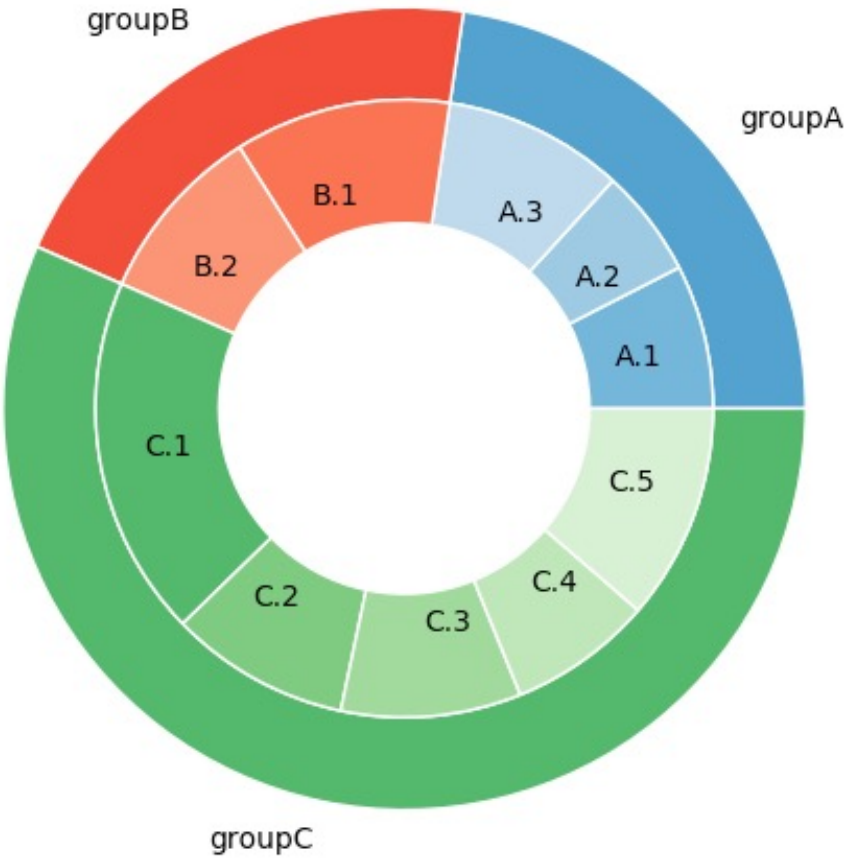


versus



All single lines of code, using matplotlib

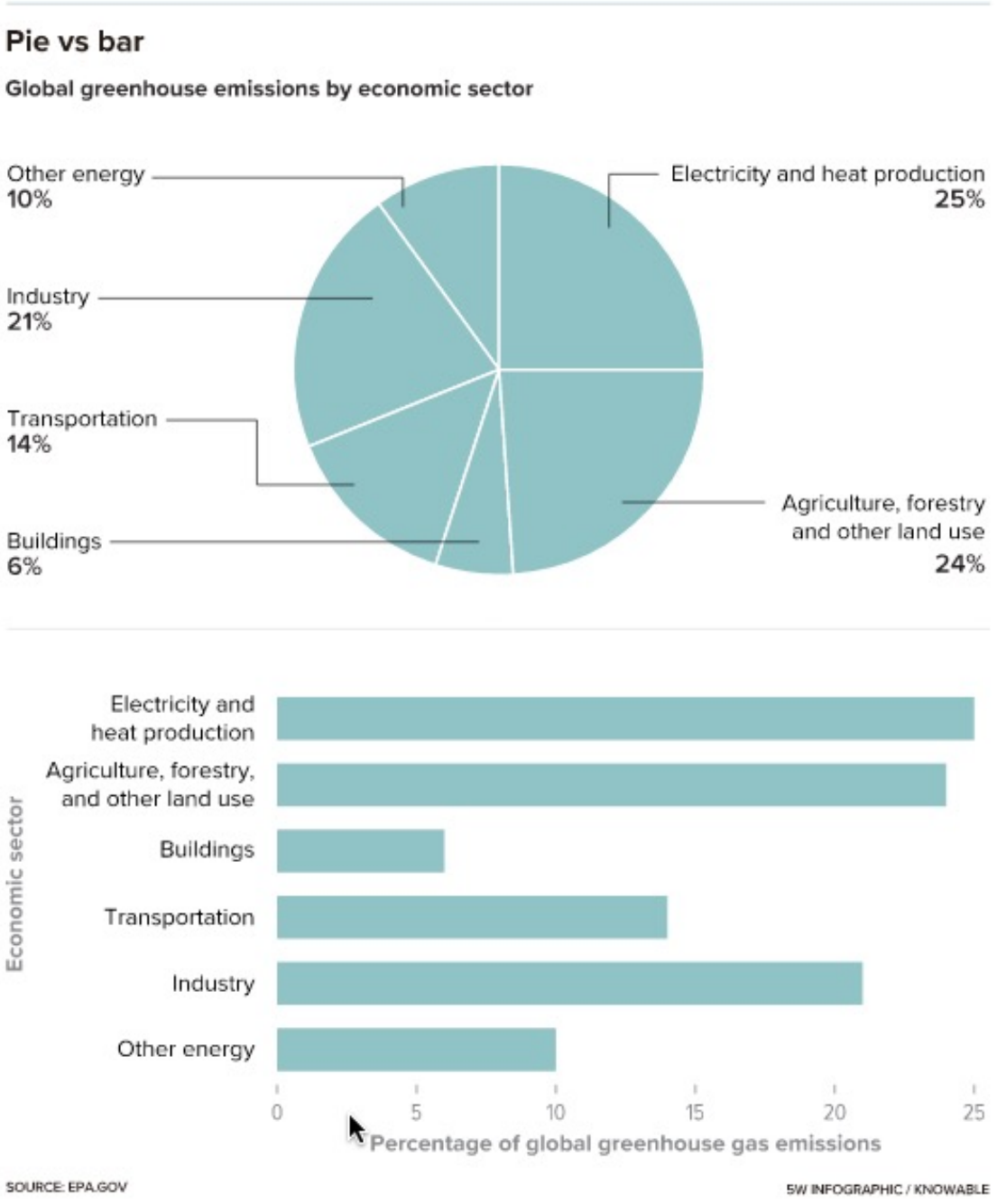
related option: donut plots



https://python-graph-gallery.com/163-donut-plot-with-subgroups/

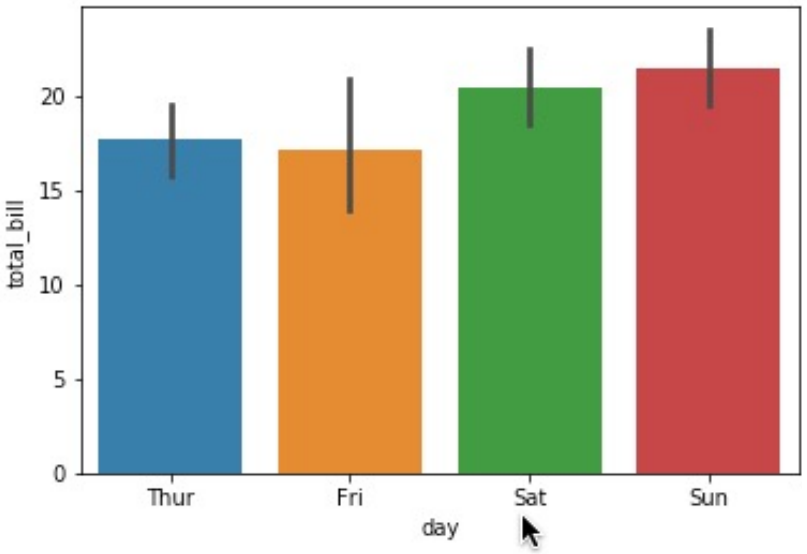
pie vs. bar

Which is better, for what purpose?



https://www.knowablemagazine.org/article/mind/2019/science-data-visualization

bar charts

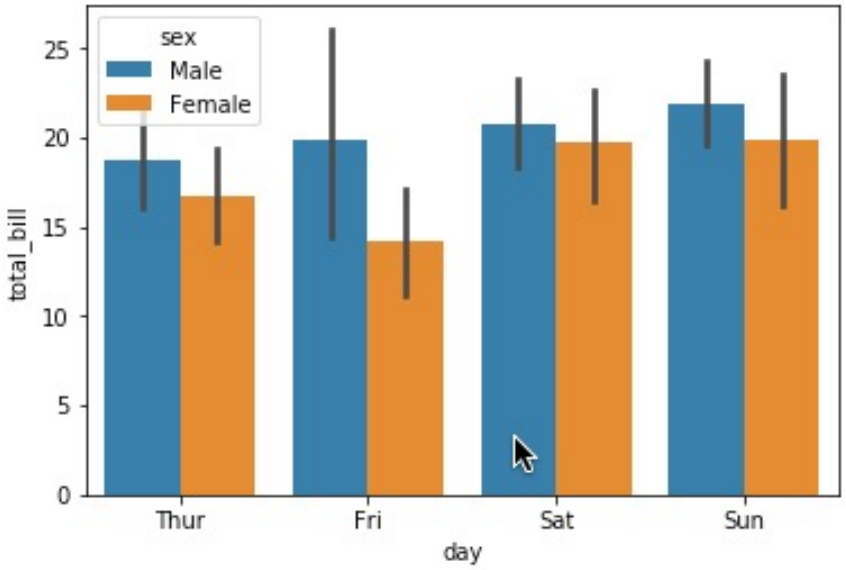


When useful?

- counts per (discrete) category
- or: other statistical property (e..g, mean) per (discrete) category

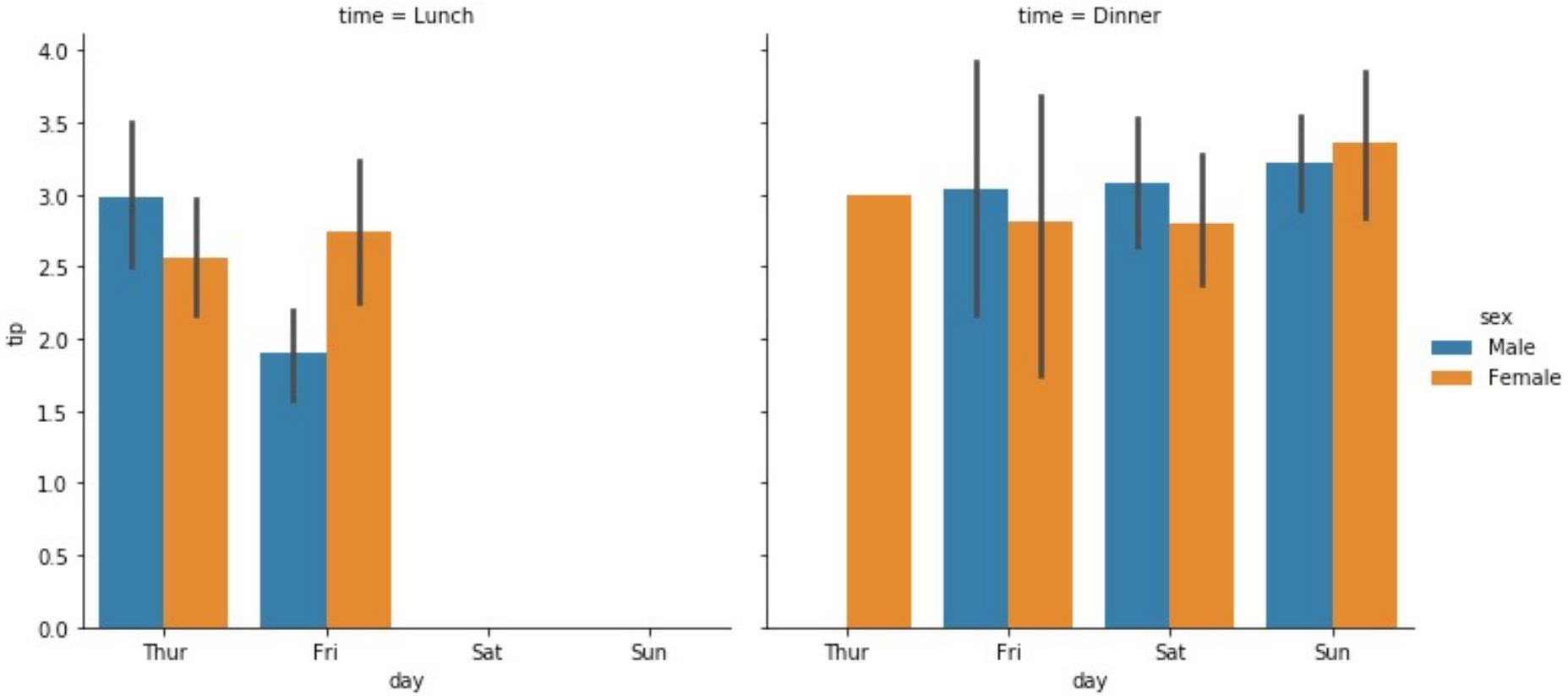
Here, using Seaborn 'barplot'

- Confidence intervals – necessary?
- Simple to group by, as well:



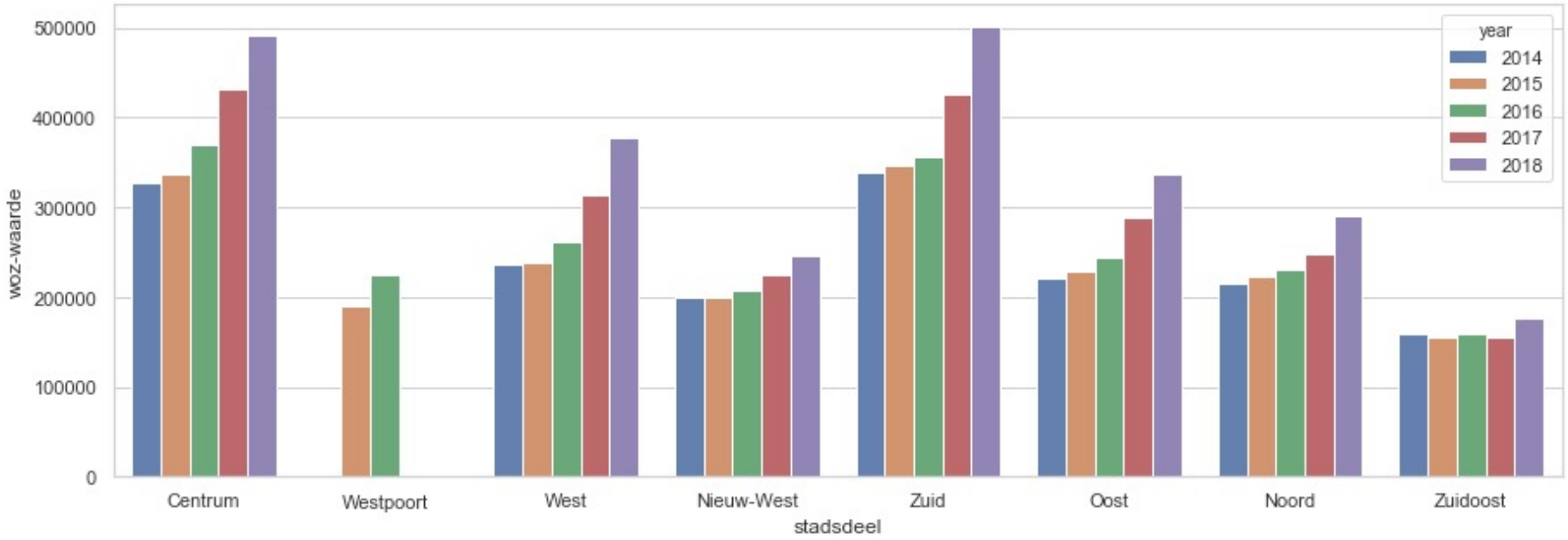
side-by-side plots for comparison

- Using Seaborn ‘catplot’



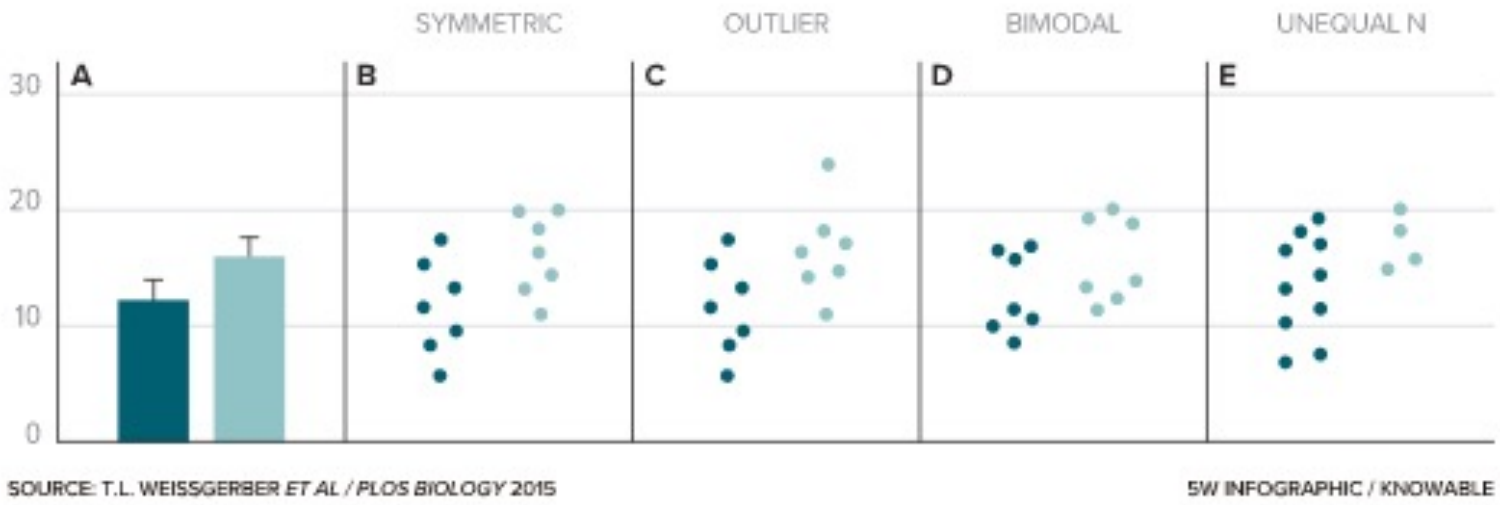
and clustered bar charts

- What types of data would be well-represented here?
- What are the limitations of bar charts?



Hidden in the bars

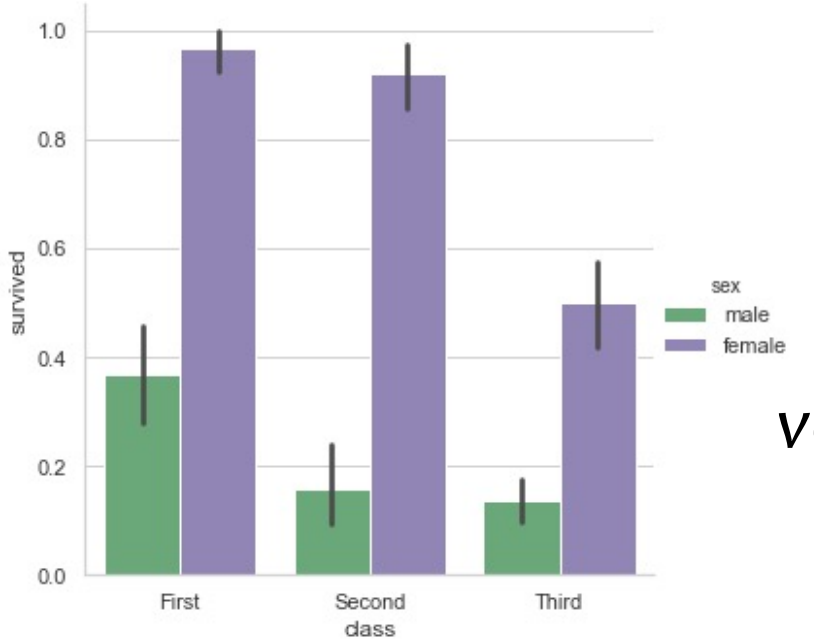
Data revealed in scatterplots may be masked within a bar chart.



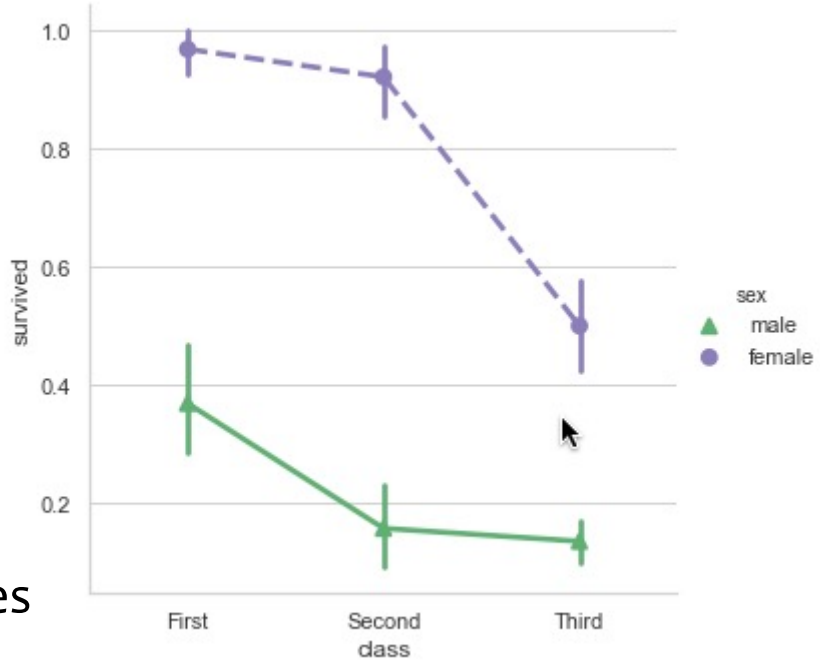
Every one of the four sets of data on the right can be accurately represented by the same bar graph on the left, illustrating how bar graphs can obscure important details about the data, possibly misleading readers.

https://www.knowablemagazine.org/article/mind/2019/science-data-visualization

point charts (e.g., deaths on the titanic by ticket class)



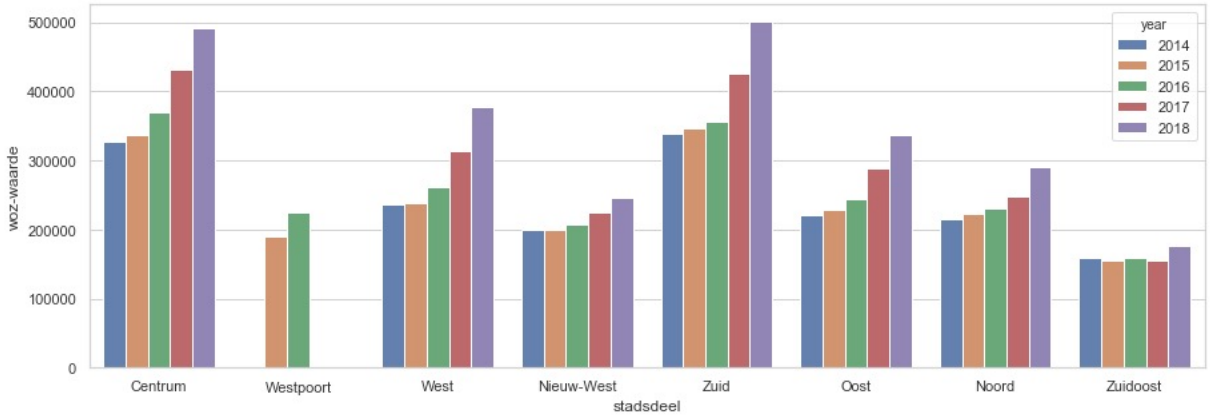
versus



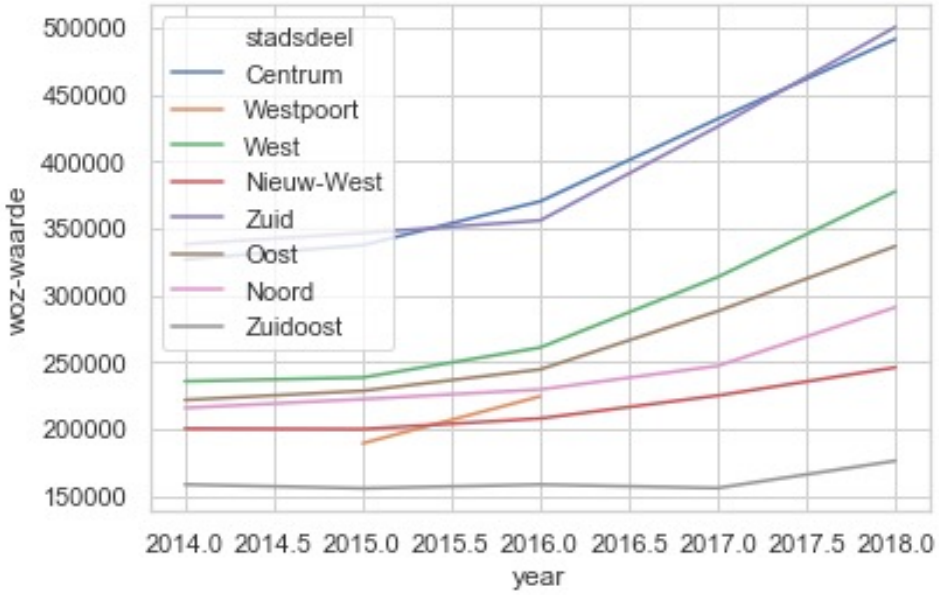
Which do you prefer and why?
Why are these useful?

- Can look cleaner
- Can give a clearer message if the categories can be meaningfully ordered (narrative!)

line charts



versus

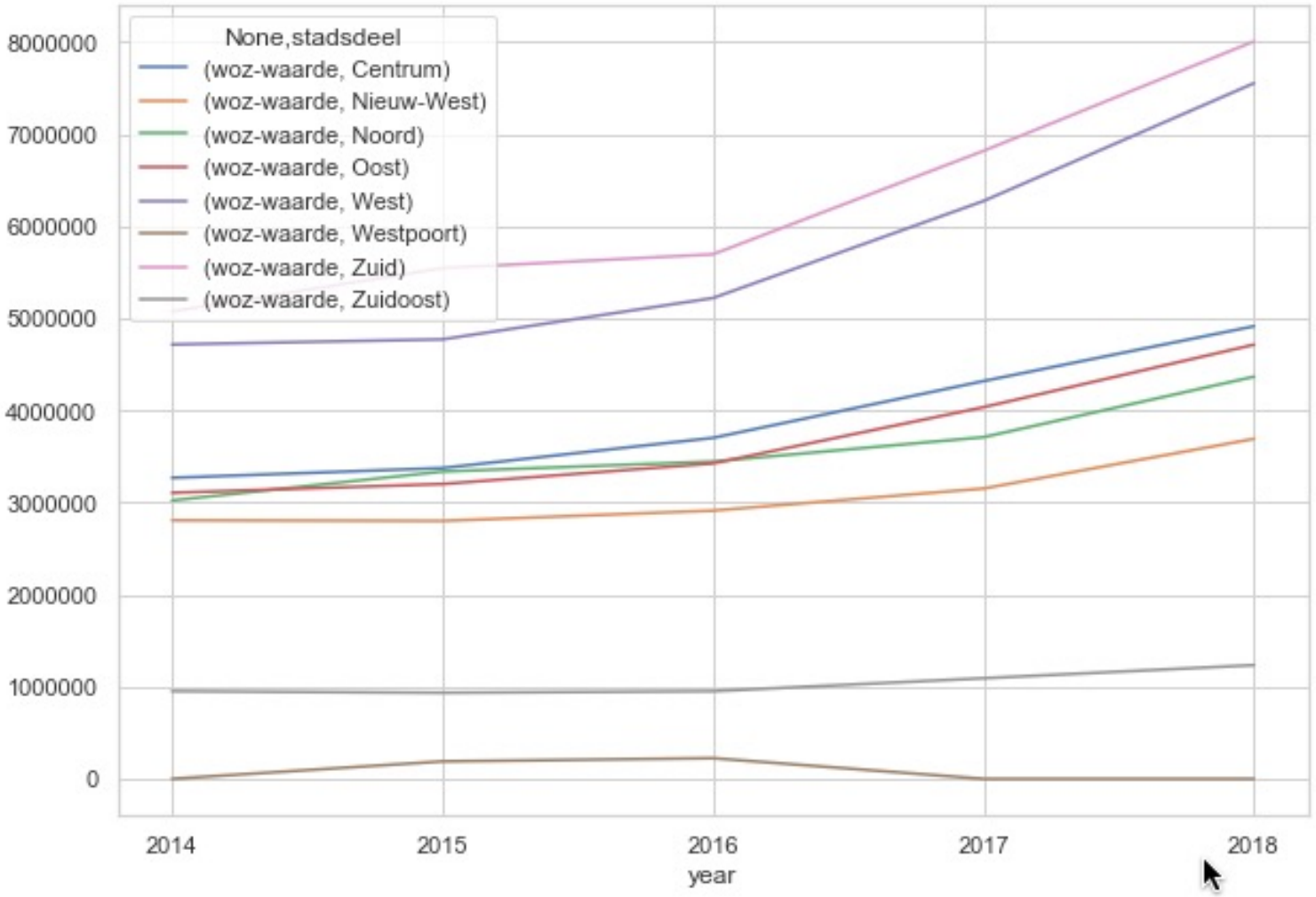


When are these useful?

- When x-axis values ordered, evenly spaced (typically)
- When x–axis has many measurements
- Most typical: over time plots

- Which preferred here?

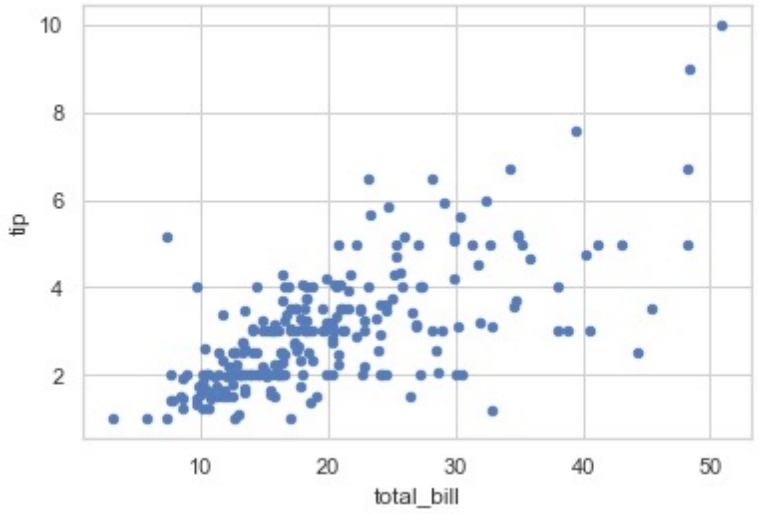
a prettier version

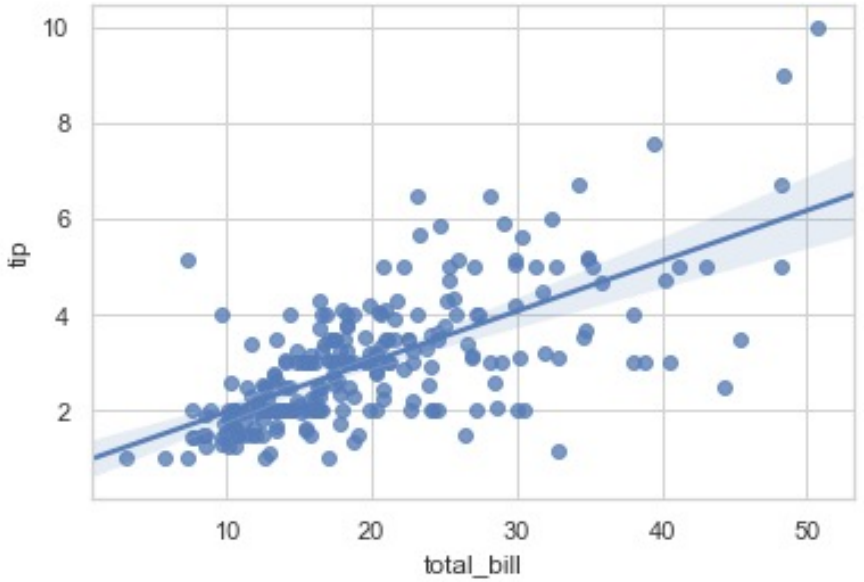


Using matplotlib, slightly more complex in this case

last one today: scatterplots

- For bivariate relationships, sometimes this is most effective

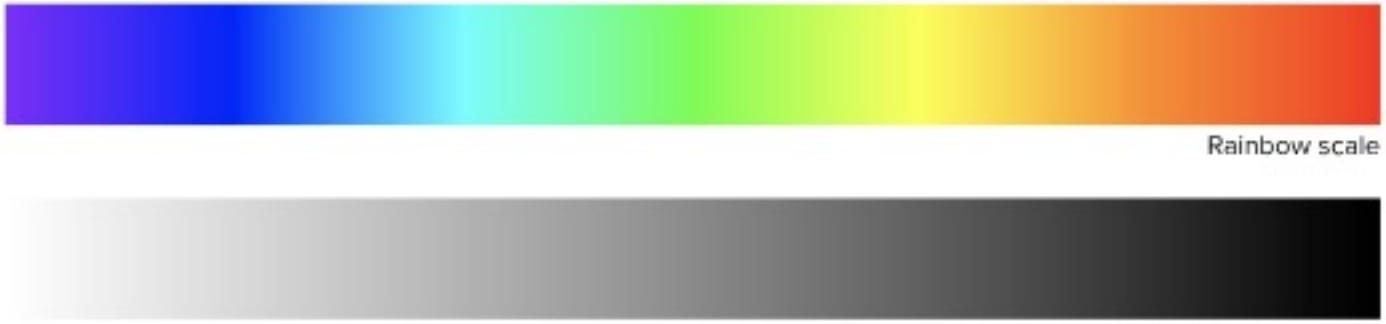




- Simple ≠ unsophisticated
- Can still add regression lines, CIs

considering color, and avoiding rainbows?

- Not always intuitive, either; relationship between colors needs to be clear

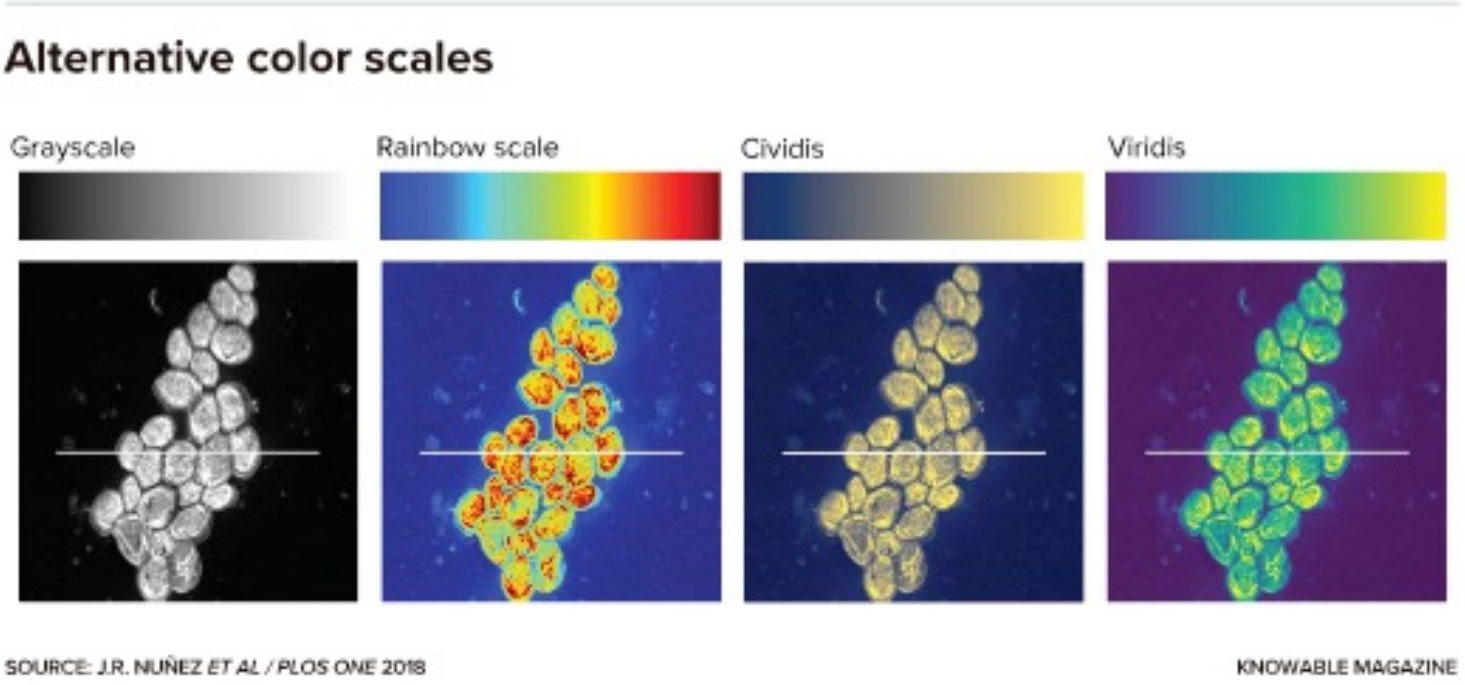


The grayscale may look dull, but it is intuitive. It's very clear how each individual shade on the scale relates to the others. This is not true for the rainbow scale, which is one of the reasons cartographers and data visualization experts avoid it.

CREDIT: 5W INFOGRAPHIC / KNOWABLE

https://www.knowablemagazine.org/article/mind/2019/science-data-visualization

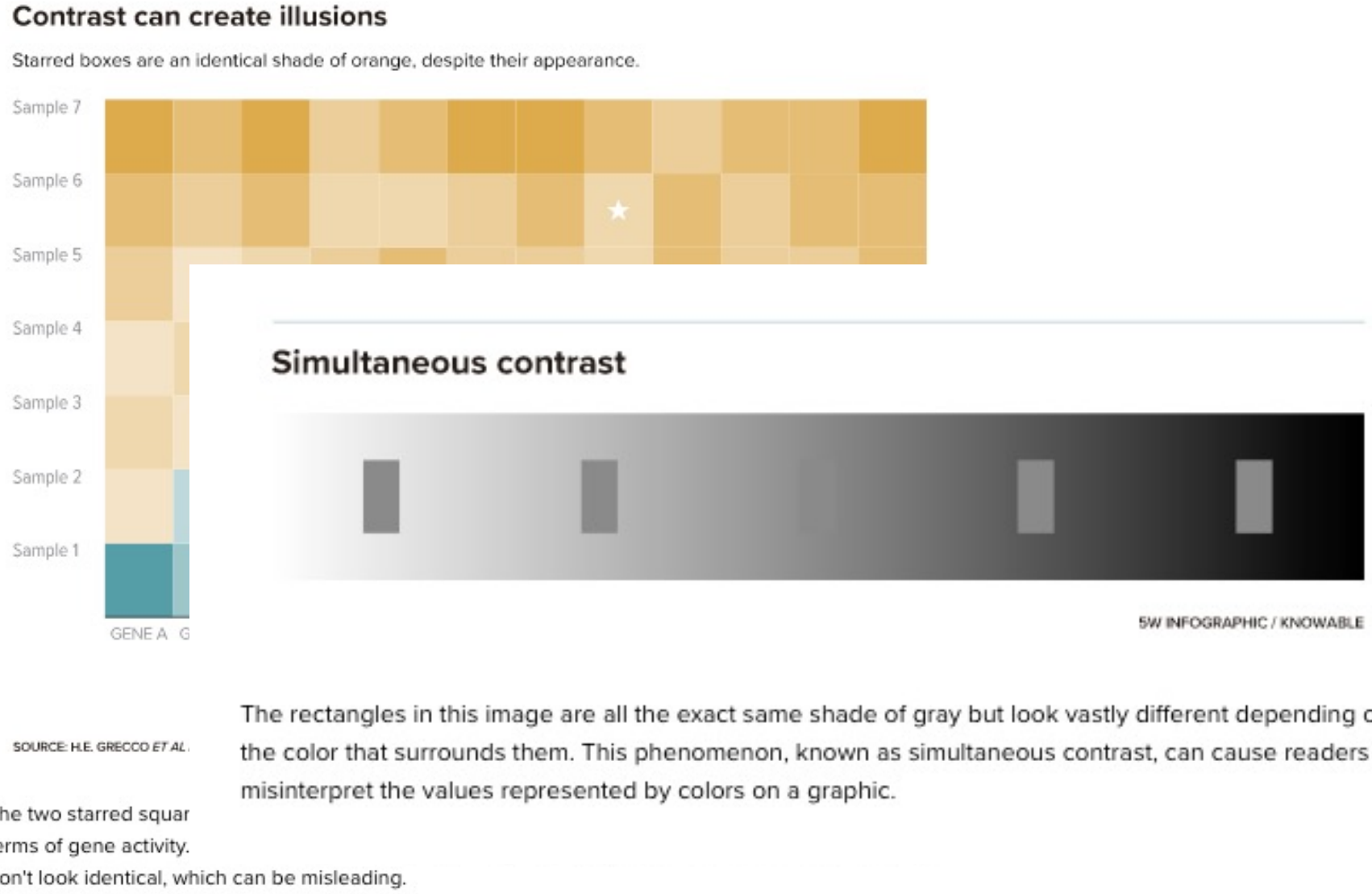
considering color, and avoiding rainbows?



A microscopic image of yeast cells rendered with different color scales highlights the counterintuitive nature of the rainbow scale. Both the viridis and cividis color scales are intended to better represent the underlying data and are easier to read. Cividis was specifically designed to be legible for color-blind people as well.

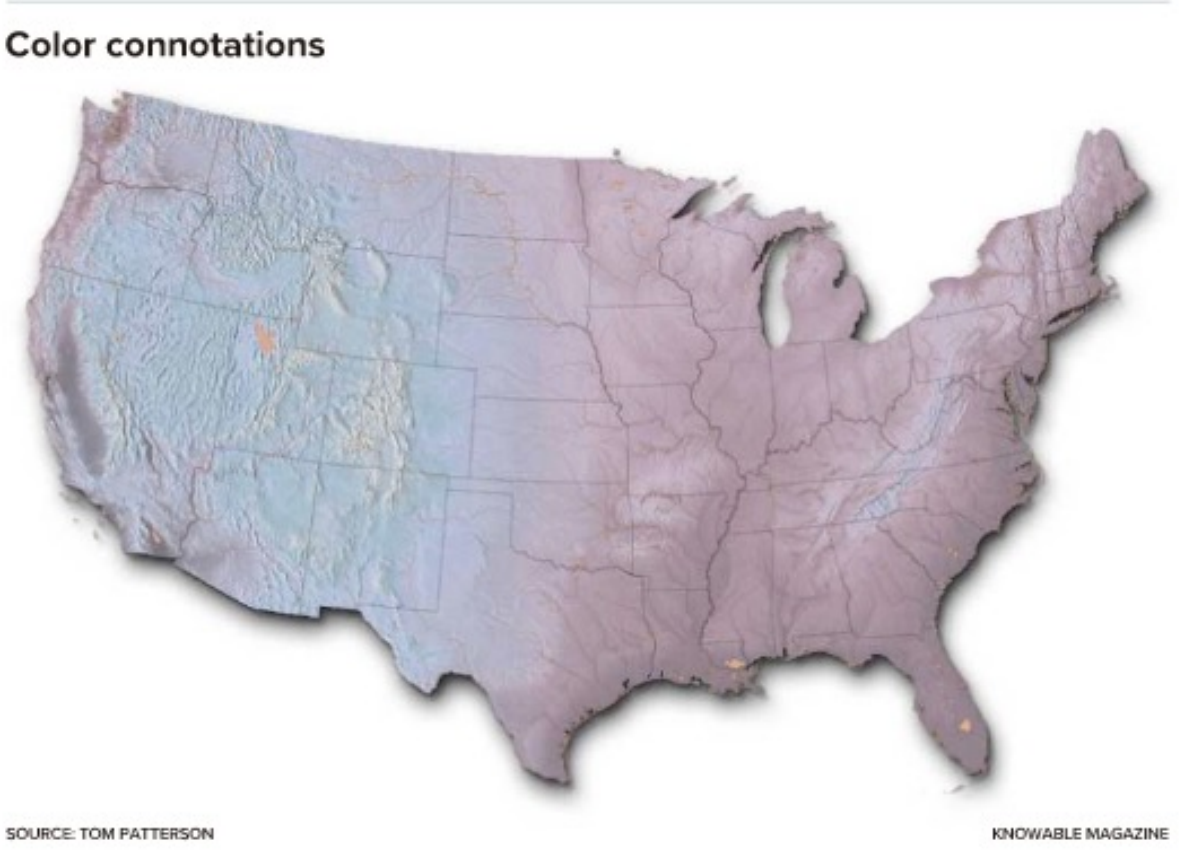
https://www.knowablemagazine.org/article/mind/2019/science-data-visualization

and, beware of simultaneous contrast and heatmaps



https://www.knowablemagazine.org/article/mind/2019/science-data-visualization

and don’t go too far off the grid, color-wise

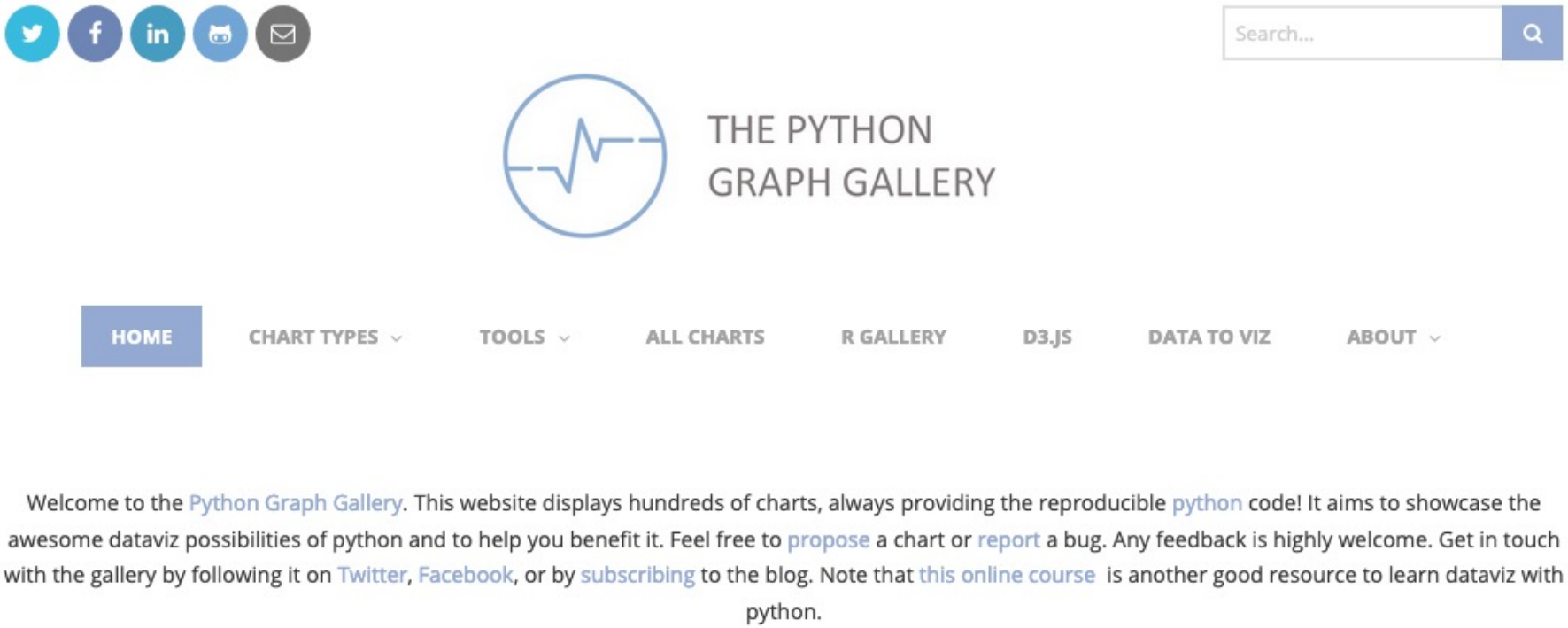


Readers have culturally defined expectations about what different colors mean. Violating such expectations makes graphs, maps and other illustrations more difficult to decipher, as this color-shifted relief map of the United States demonstrates.

https://www.knowablemagazine.org/article/mind/2019/science-data-visualization

also, this!

- https://python-graph-gallery.com/



If time: small group activity

- Work in groups of 2-3, find a visualization example from a **major news outlet**

- Discuss its trustworthiness, accessibility, and elegance
  - What does it do well?
  - What should be improved?
  - Does it enhance **understanding** of the story? Why (not)?

- Post a link to a really good/bad example on Canvas