Visualizing Data

Recap

- We have spent **a lot** of time working with `ggplot2`
- Everything covered this week can be applied with `ggplot2`
- Goal of this week is to get you to stop and think about your plots *before* you make them. What are the kinds of things you should be considering?

Stuff you already know

Data visualization is both an art *and* a science

Art:

- Aesthetically pleasing
- The story you're trying to tell

Science:

- Graphical representation of specific types of data
- The story you're trying to tell

Warning

Making figures for academic purposes $\neq$ making figures for pure data viz purposes

If you go into data science (and aren't constrained by academia), you'll want to check out R shiny (for interactive plots), web design, typography, etc.

Today

- What kinds of things should you be thinking about when it comes to data viz *for academic papers*?
- What **not** to do (I'm gonna rant a bit)
- Hopefully helpful resources (no memorizing!)

What should we be thinking about?

1. Telling your story
2. Contrast
3. Accessibility

Telling your story

- Raw data are not intuitive. For the most part, you can't look at a spreadsheet of numbers and decipher any patterns. Especially with really big spreadsheets!
- We need a way to graphically show the data so that our human eyes can try to make sense of the data.
- It is *so* easy to **LIE** with data! Balancing act of conveying your message and not lying.
- As our datasets become more complex and high dimensional, data visualization can become more challenging.
- The goal is NOT to show every bit of data you have collected. It's to show the relationship you care about in an honest manner.

What makes a good figure?

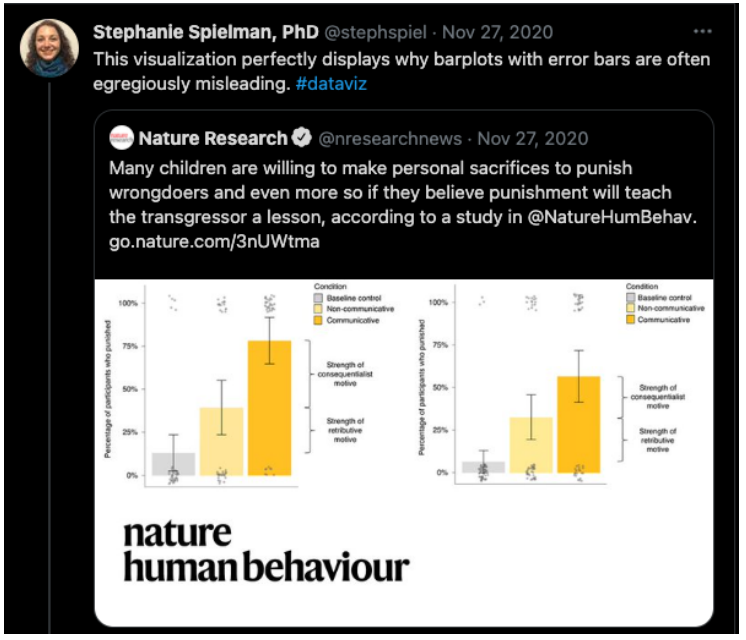
- Clear, descriptive title
- Axes are clearly labeled with variables and units of measurement
    - **Label. Your. Damn. Axes.**
- Scale is:
    - consistent across axis
    - easily interpreted
    - chosen so that data are evenly distributed (remember restricted range from correlation?)
- Data points are represented clearly, with a good key/legend if needed
- Graph is the *appropriate* type for your data (nominal, ordinal, interval, ratio etc.)

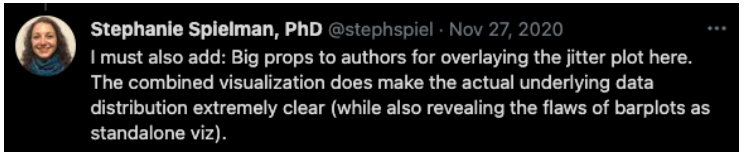
Appropriateness

There are no 100% right answers, but there are wrong ones...

- Pie charts are never the answer; 3D pie charts are the worst of the worse
  - Instead, try stacked bar plot or even better, stacked dot plot
- 3D bar plots are never the answer
- Typically, you should show either the raw data or at least a distribution
  - Shelly is anti bar plot, generally
  - Shelly is anti box plots on their own (nice when combined)
  - Shelly is skeptical of lines -- need to be careful

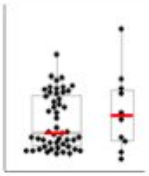
Why Bar Plots Can Suck





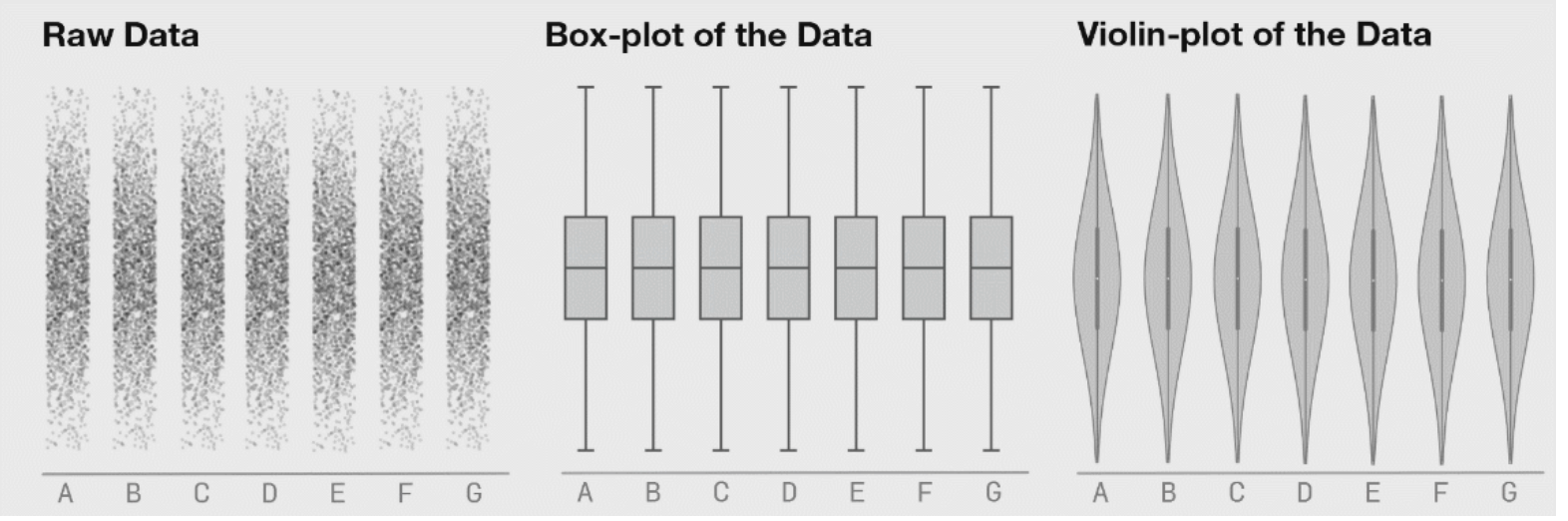
- Different alternatives exist, with dot plots being the best option probably
- Bars are more appropriate when you have proportions or counts; but even still -- dotplots
- Tracey Weissgerber has an excellent thread with many resources on why barplots suck and how to maximize the utility of dotplots: https://tinyurl.com/3e6222a4

What to do?

Figure Types	Example	Type of Variable	What the Plot Shows	Sample Size	Data Distribution	Best Practices
Dot plot		Continuous	Individual data points & mean or median line Other summary statistics (i.e. error bars) can be added for larger samples	Very small OR small; can also be useful with medium samples	Sample size is too small to determine data distribution OR Any data distribution	• Make all data points visible - use symmetric jittering • Many groups: Increase white space between groups, emphasize summary statistics & de-emphasize points • Only add error bars if the sample size is large enough to avoid creating a false sense of certainty • Avoid "histograms with dots"
Dot plot with box plot or violin plot		Continuous	Combination of dot plot & box plot or violin plot (see descriptions above and below)	Medium	Any	• Make all data points visible (symmetric jittering) • Smaller n: Emphasize data points and de-emphasize box plot, delete box plot and show only median line for groups with very small n • Larger n: Emphasize box plot and de-emphasize points
Box plot		Continuous	Horizontal lines on box: 75th, 50th (median) and 25th percentile Whiskers: varies; often most extreme data points that are not outliers Dots above or below whiskers: outliers	Large	Do not use for bimodal data	• List sample size below group name on x-axis • Specify what whiskers represent in legend
Violin plot		Continuous	Gives an estimated outline of the data distribution. The precision of the outline increases with increasing sample size.	Large	Any	• List sample size below group name on x-axis • The violin plot should not include biologically impossible values
Bar graph		Counts or proportions	Bar height shows the value of the count or proportion	Any	Any	• **Do not use for continuous data**

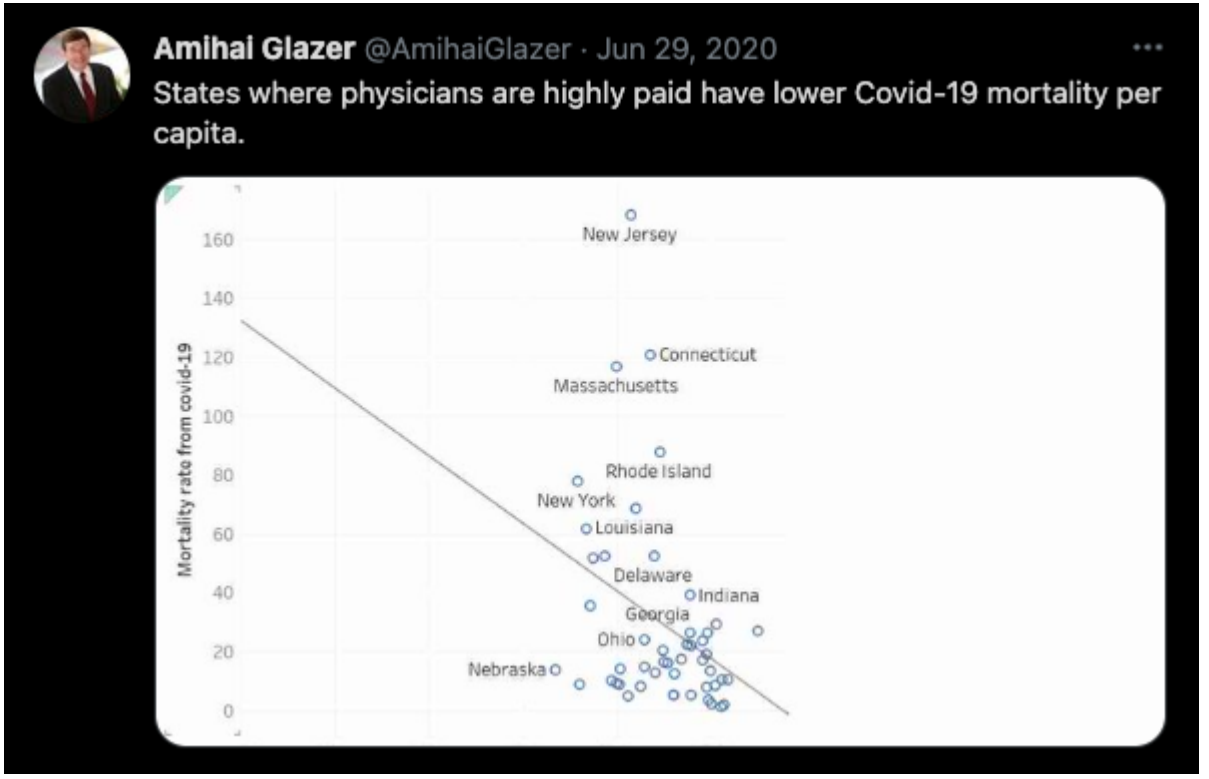
Source: @T_Weissgerber

Why box plots on their own can suck

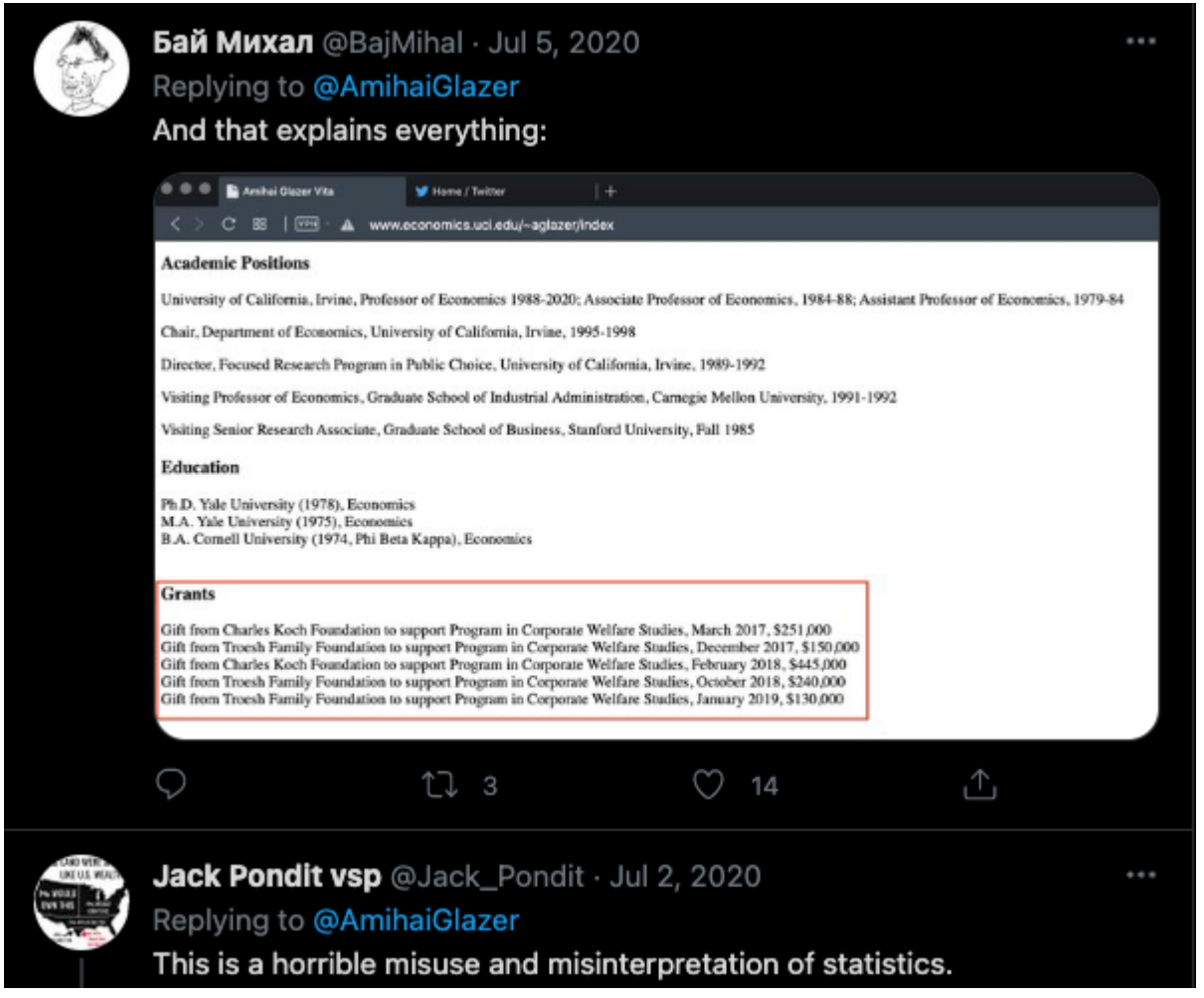


https://www.autodesk.com/research/publications/same-stats-different-graphs

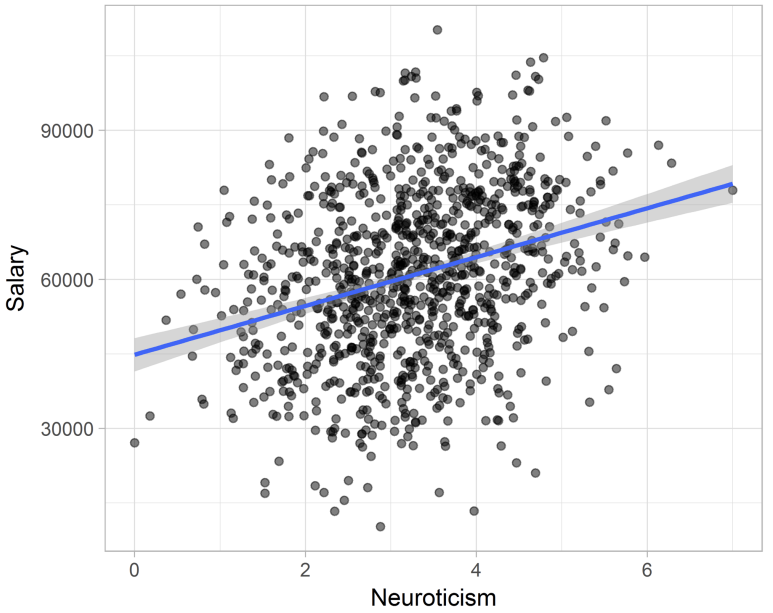
Why adding lines can be misleading

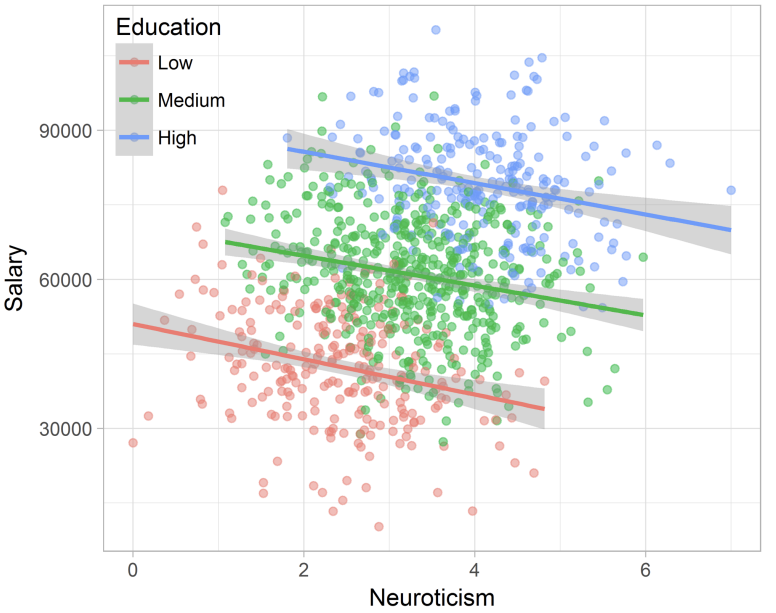


Lines Cont...


Бай Михал @BajMihal · Jul 5, 2020
Replying to @AmihaiGlazer
And that explains everything:
Amihai Glazer Vita
Home / Twitter
www.economics.uci.edu/~aglazer/index
Academic Positions
University of California, Irvine, Professor of Economics 1988-2020; Associate Professor of Economics, 1984-88; Assistant Professor of Economics, 1979-84
Chair, Department of Economics, University of California, Irvine, 1995-1998
Director, Focused Research Program in Public Choice, University of California, Irvine, 1989-1992
Visiting Professor of Economics, Graduate School of Industrial Administration, Carnegie Mellon University, 1991-1992
Visiting Senior Research Associate, Graduate School of Business, Stanford University, Fall 1985
Education
Ph.D. Yale University (1978), Economics
M.A. Yale University (1975), Economics
B.A. Cornell University (1974, Phi Beta Kappa), Economics
Grants
Gift from Charles Koch Foundation to support Program in Corporate Welfare Studies, March 2017, $251,000
Gift from Troesh Family Foundation to support Program in Corporate Welfare Studies, December 2017, $150,000
Gift from Charles Koch Foundation to support Program in Corporate Welfare Studies, February 2018, $445,000
Gift from Troesh Family Foundation to support Program in Corporate Welfare Studies, October 2018, $240,000
Gift from Troesh Family Foundation to support Program in Corporate Welfare Studies, January 2019, $130,000
3
14
Jack Pondit vsp @Jack_Pondit · Jul 2, 2020
Replying to @AmihaiGlazer
This is a horrible misuse and misinterpretation of statistics.

Simpson's Paradox

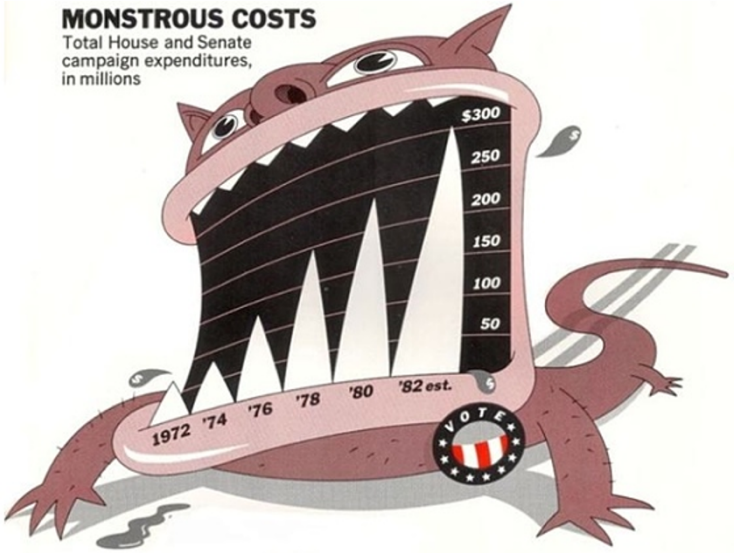




Things to avoid

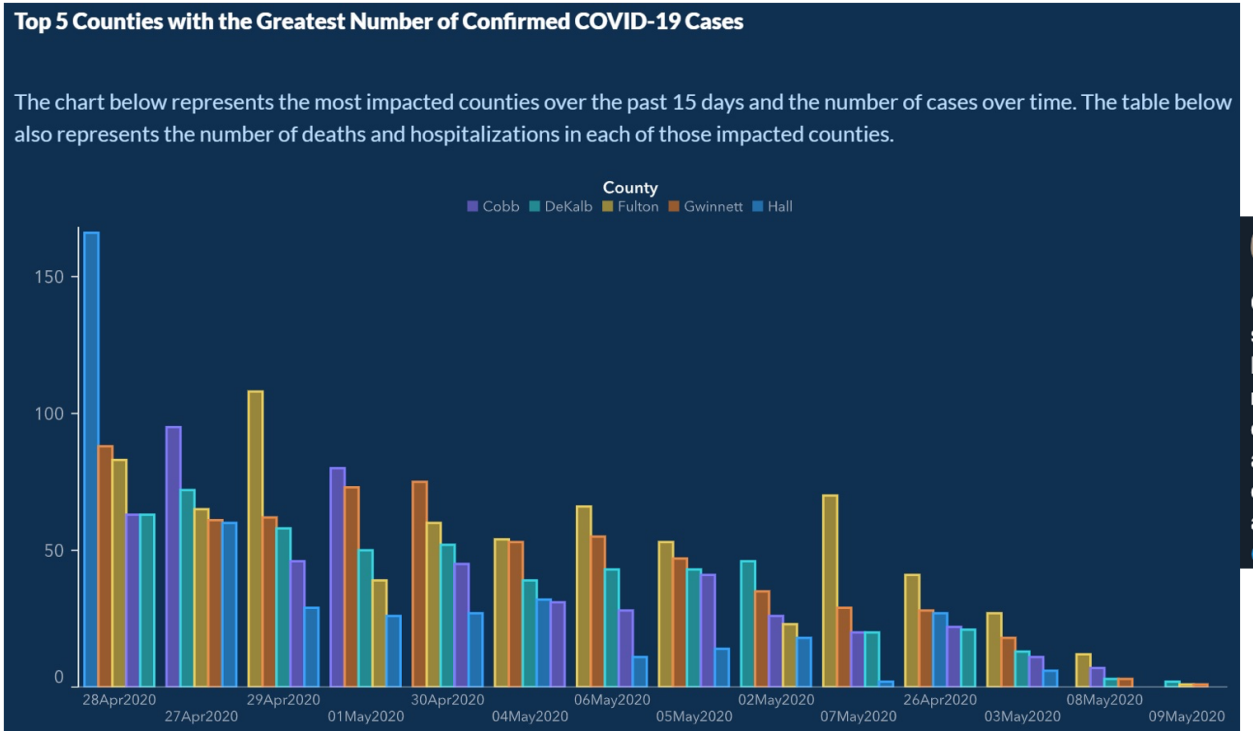
- Chartjunk
- Misleading text/axes
- Inaccurate plotting
- So many COVID-19 data visualizations...

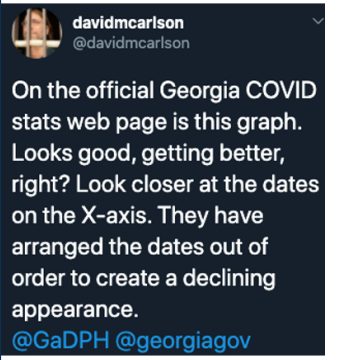
Chartjunk


MONSTROUS COSTS
Total House and Senate campaign expenditures, in millions
$300
250
200
150
100
50
1972
'74
'76
'78
'80
'82 est.
VOTE


THE BIG PICTURE
'TIS THE SEASON...
It's the best of times for retailers, especially department stores, some of which rake in as much as 50% of their yearly sales during the holidays. And this year looks good so far, retailers say.
UNADJUSTED DEPARTMENT STORE SALES BY QUARTER
20
15
10
5
0
I '92
II
III
IV
I '93
II
III
IV
EST.
▲ BILLIONS OF DOLLARS
DATA: JOHNSON REDBOOK SERVICE

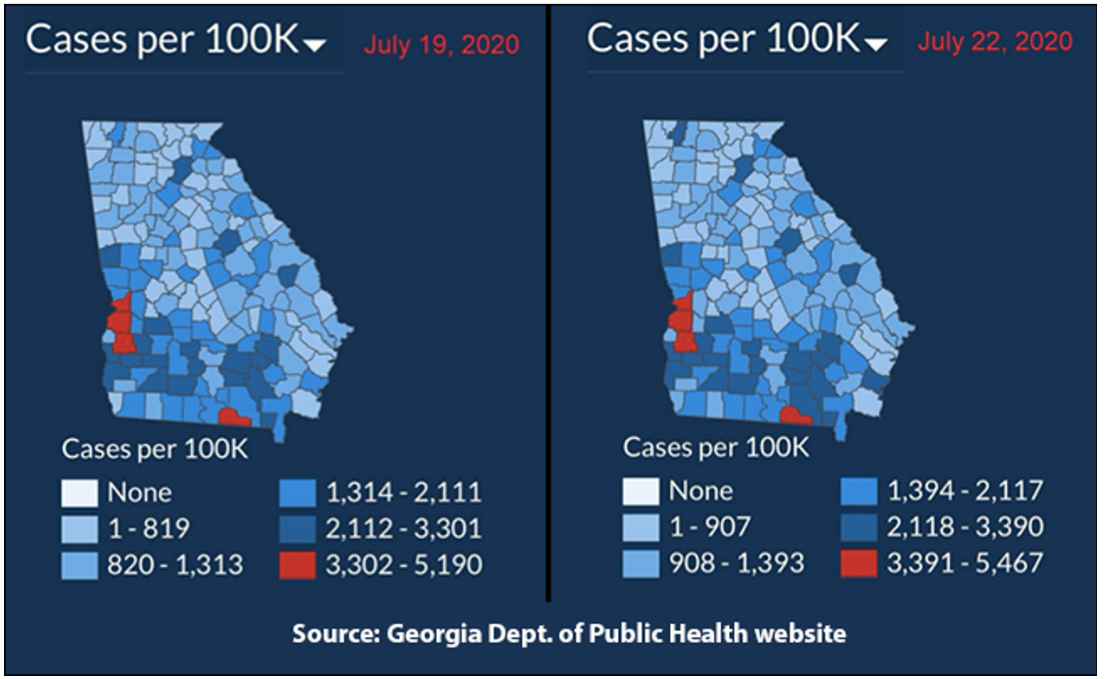
Misleading Text/Axes



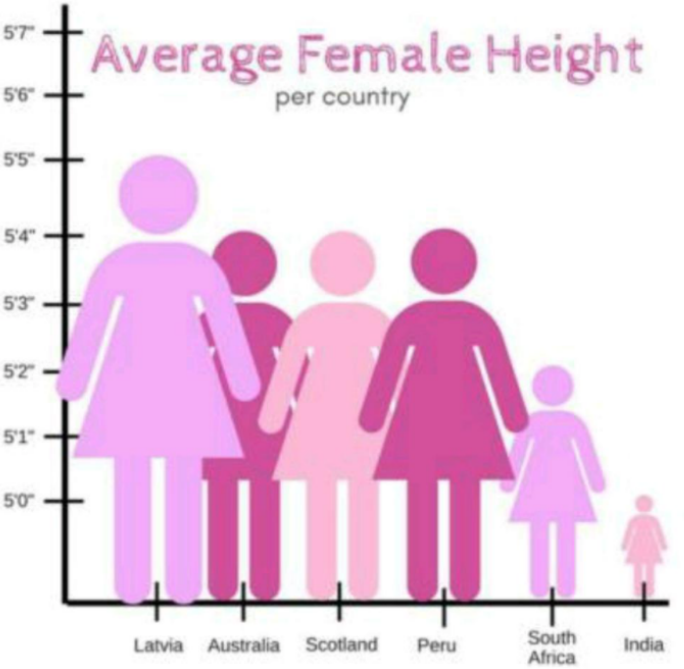


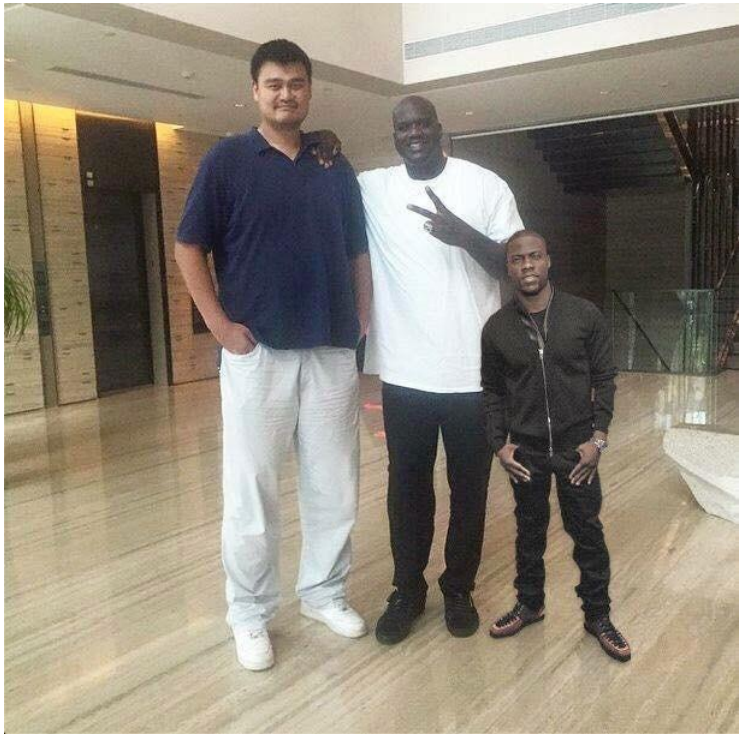
Inaccurate Plotting

This strange plot was put out by Georgia's Health Department. It's trying to show that basically there haven't been any real changes in COVID-19 statewide. But look at the values in the legend...They've changed them to basically keep the same graph. WUT?!



Misleading Shapes





- Kevin Hart: 5'2
- Shaquille O'Neal: 7'1
- Yao Ming: 7'6

Accessibility is IMPORTANT

- Colorblindness sucks. ~1 in 12 men are colorblind (much lower in women)
- People have poor vision (glasses, anyone?)
- Journals scale your figure sizes down so that it fits within the article (like within a column of text)

What can we do?

- Colors
- Contrast
- Big text size (bigger/bolder you can get away with less color contrast)

Color Palettes

Also super helpful! `RColorBrewer` and `ggsci` are great. But there are millions of others.

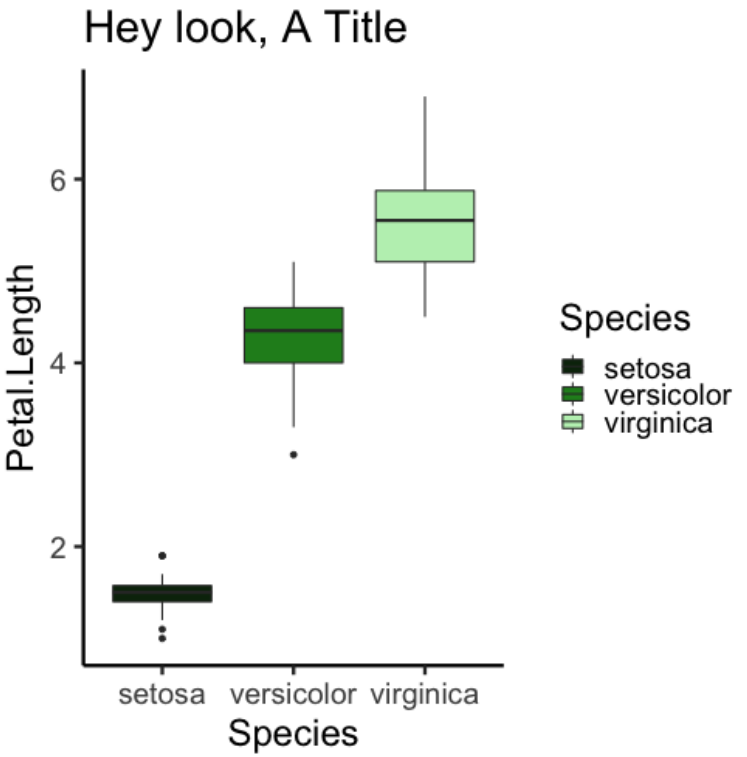
Ex: the color palette for all of the slides on this website? The Aussie color palette from https://flatuicolors.com/palette/au

All you need are hex codes (6 digits, alphanumeric). This is true for all color palettes (including monochromatic).

Different types of palettes (unordered, sequential, divergent etc.). Look at this blog post to learn more about these

Contrast

When you have something side-by-side, you can have different colors. OR you can have the same color but a different shade/tone/tint.



Bad use of colors

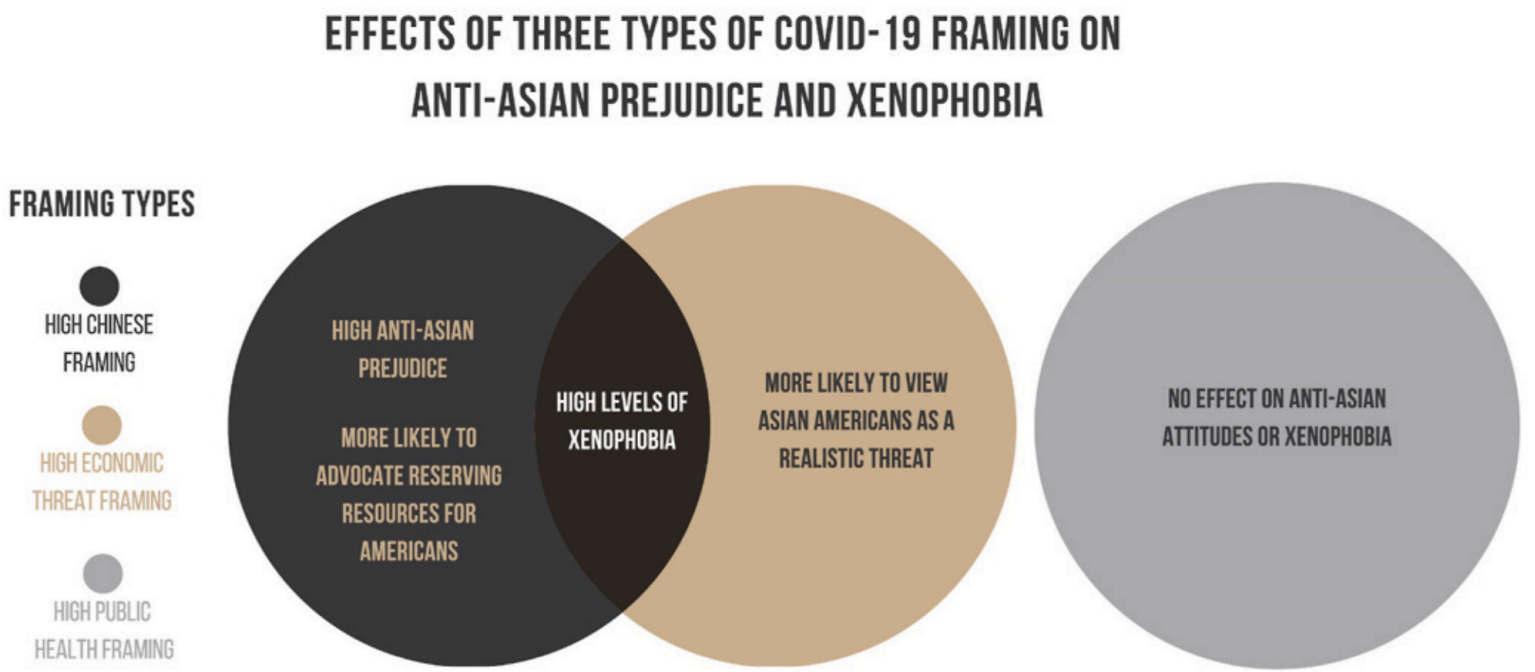
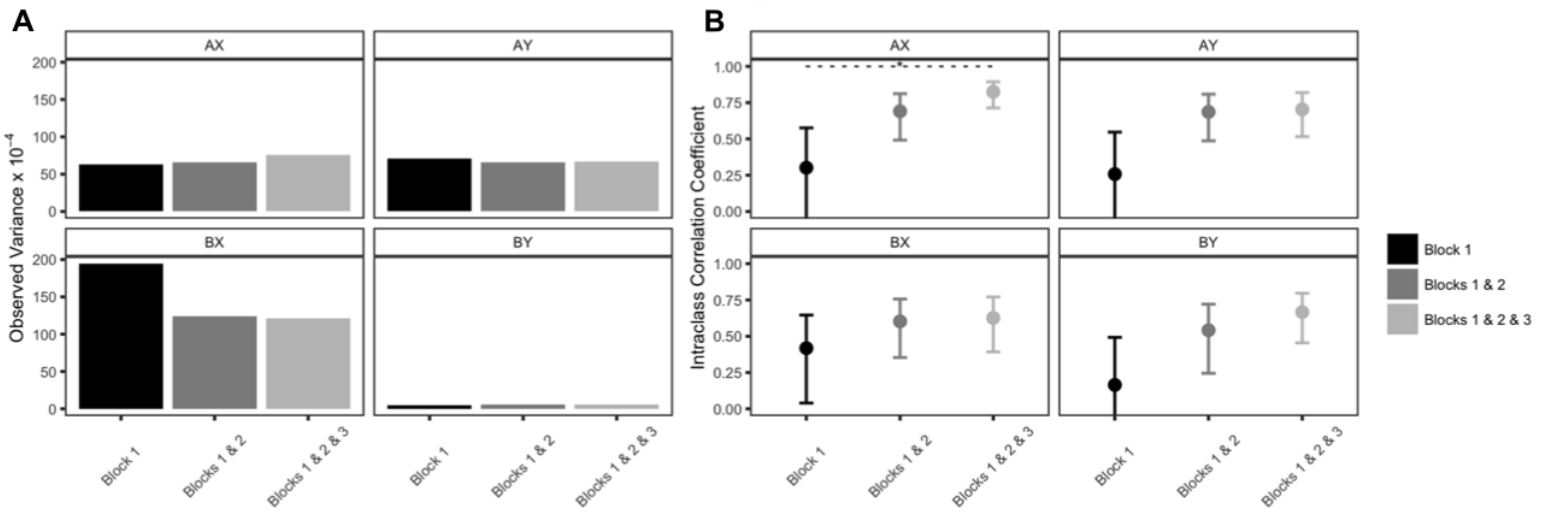


Fig. 1. Summary of primary study findings.

Grayscale

The most obvious of this is grayscale (technically, it's not a hue, and you're dealing with saturation, but that's completely unimportant for this intro):



Providing Contrast is Important

It lets the reader *easily* extract meaningful information.

Colors, shapes, size (as in bubble plots), sometimes transparency etc.

Need help picking different shades/tones/tints of the same color? A ton of websites can help! Ex: https://www.colorhexa.com/

Text Size

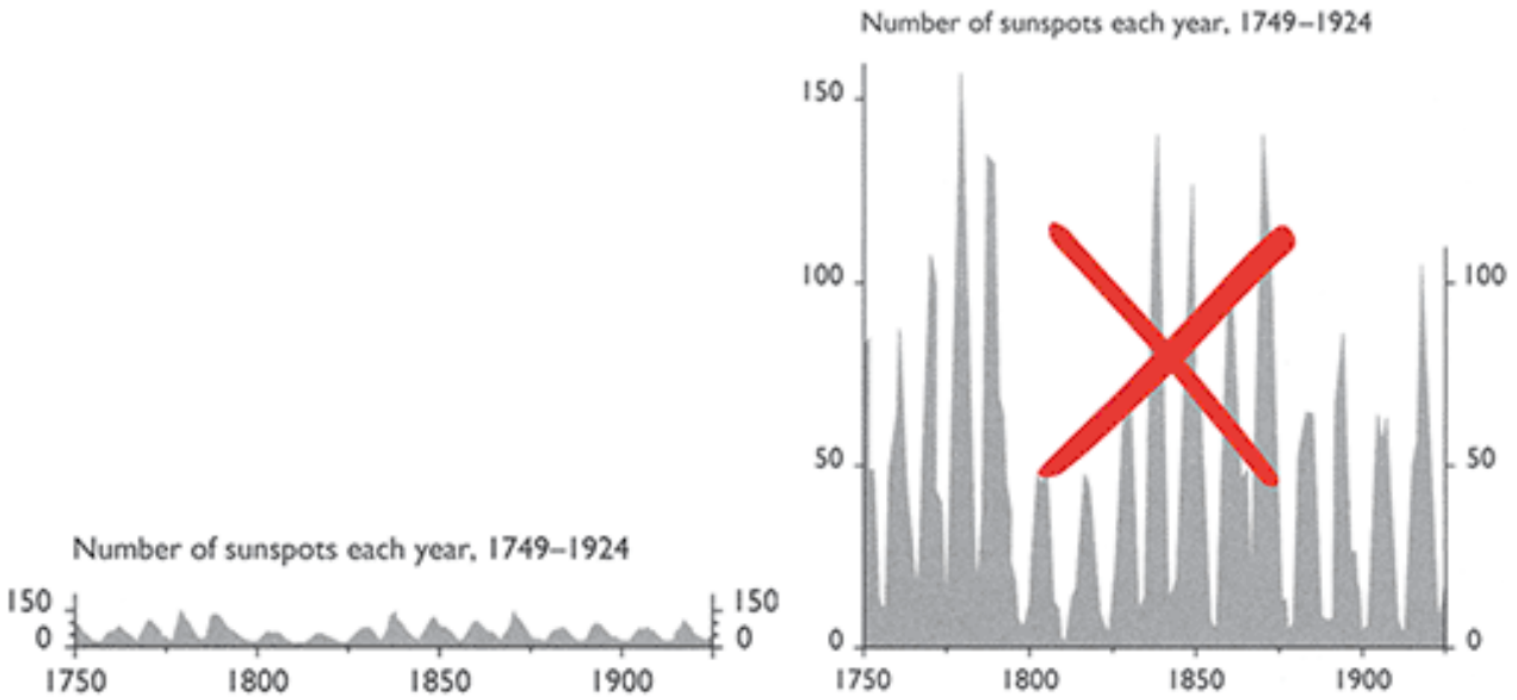
Don't make your font size super uber small!

Some of us are getting old and our eyesight is fading *(I'm not bitter...yes I am...)*

Also, academic journals scale down your figures. Better to make the text size larger so that when it gets scaled down, it's still readable.

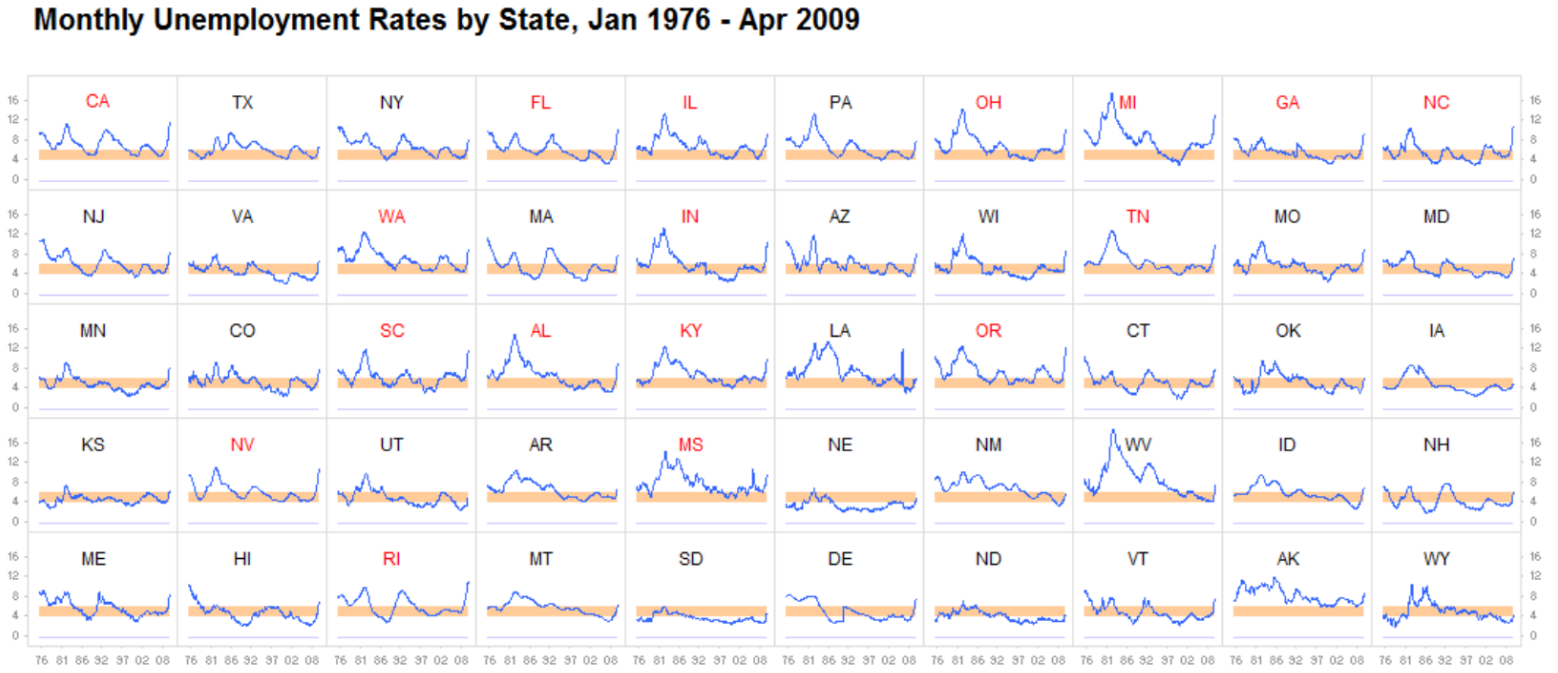
Aspect Ratios

You want to display your figures faithfully, but you don't want to take up extra space you don't need. Think about the aspect ratio of your plots!



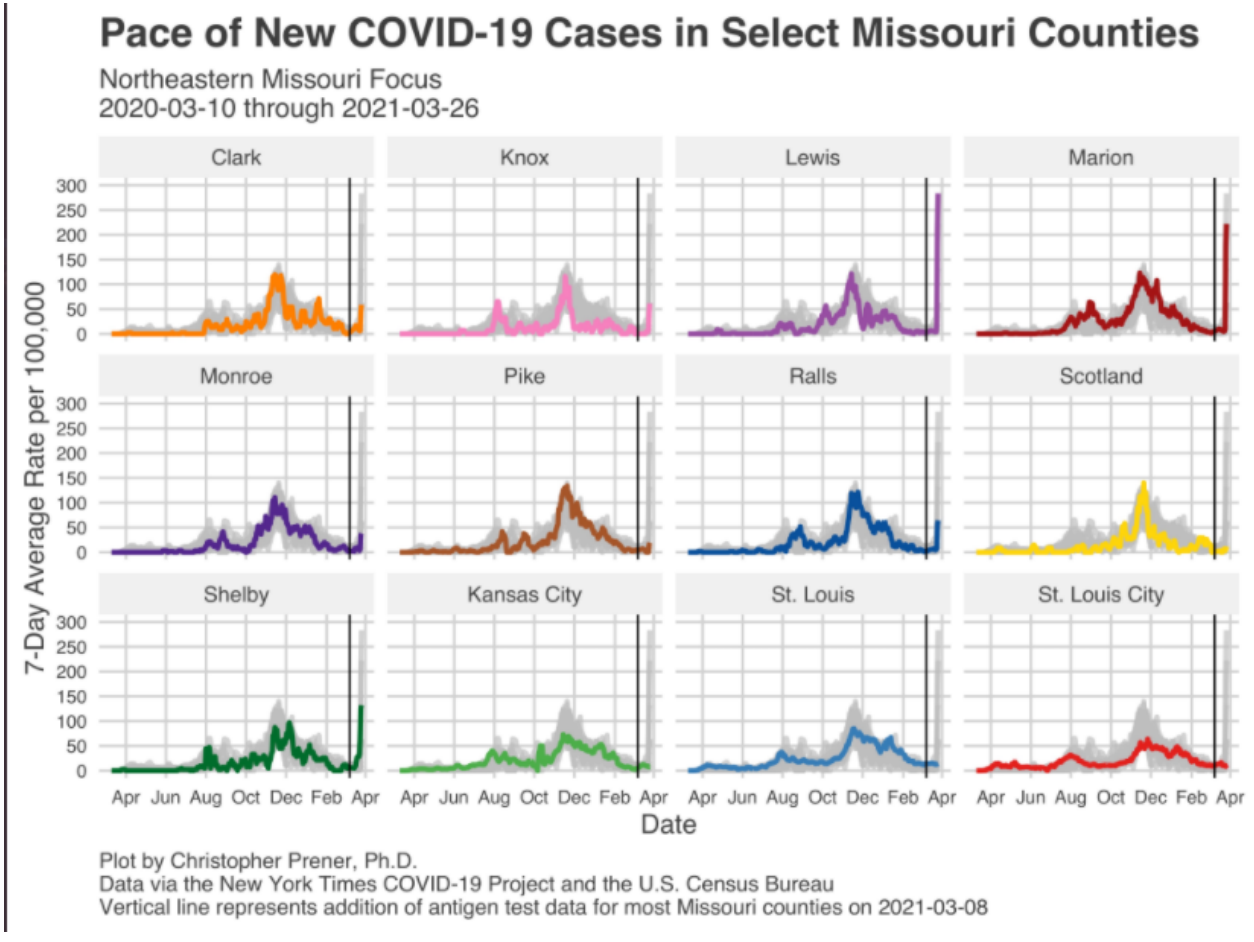
Tufte, Visual Explanations. Graphics Press, Cheshire, Conn, 1997

Small Multiples



https://www.juiceanalytics.com/writing/better-know-visualization-small-multiples

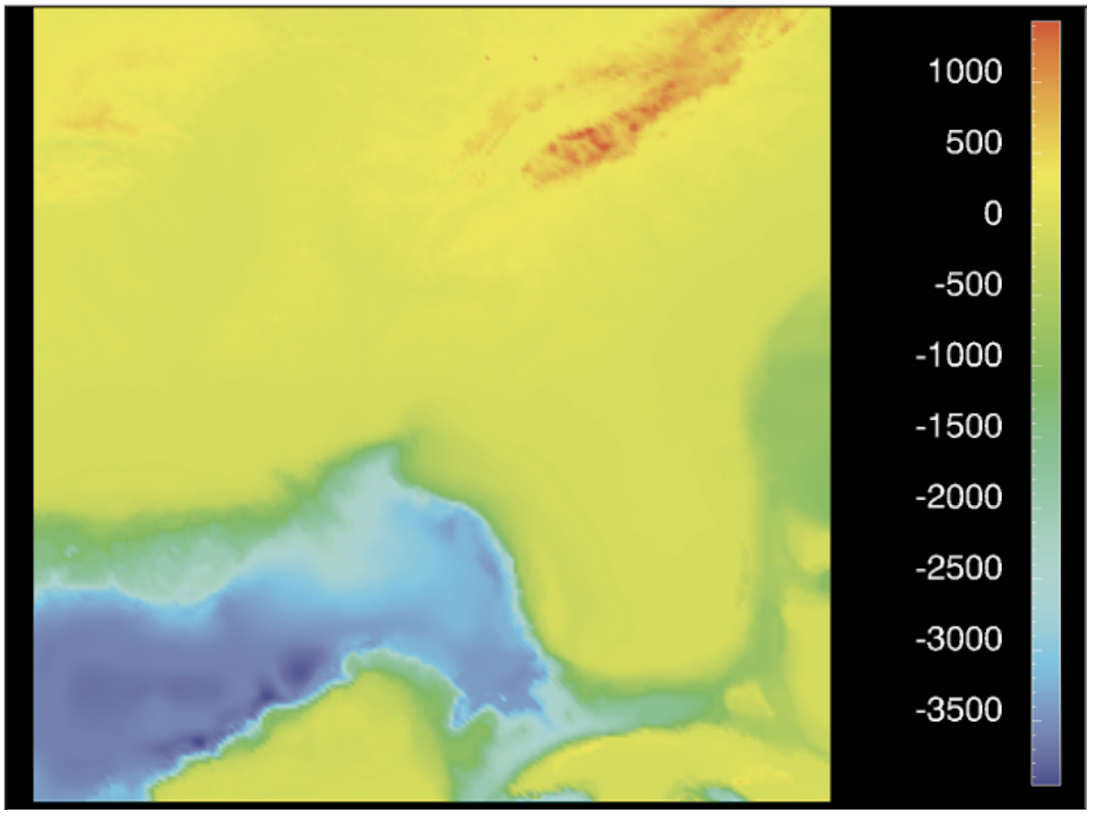
Small Multiples with COVID



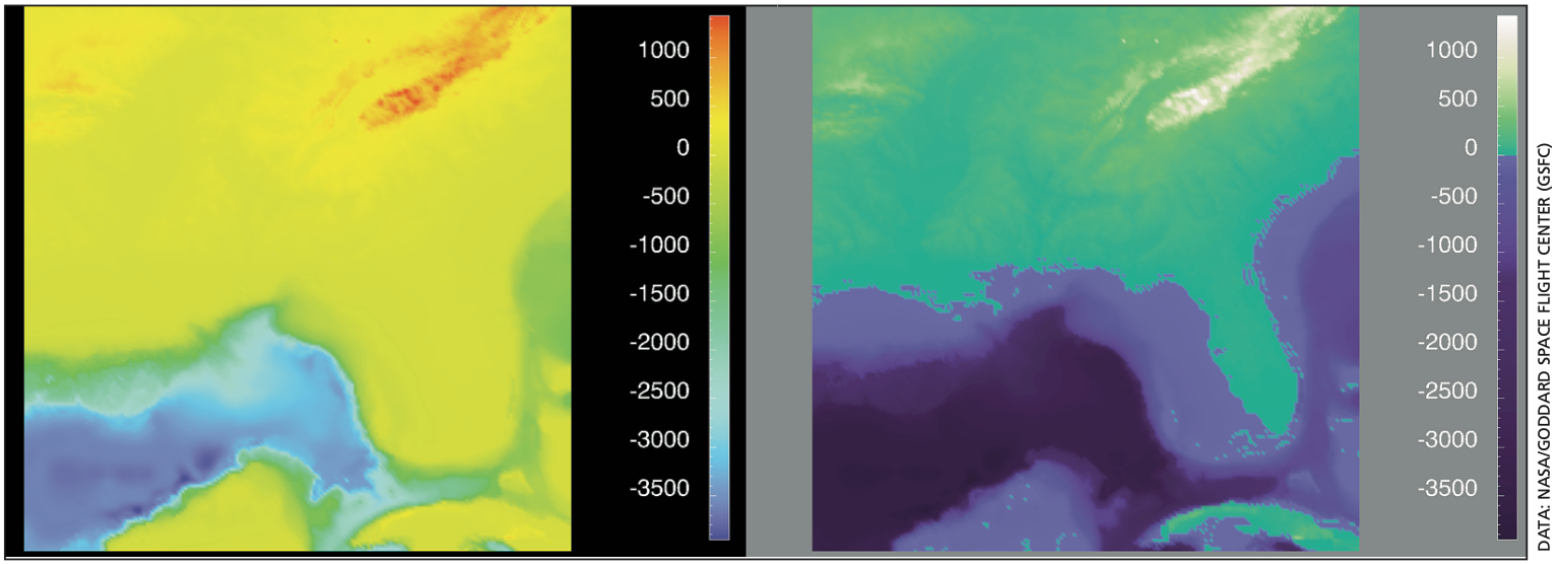
https://twitter.com/chrisprener/status/1375937857840353280/photo/2

Rainbow Colormaps

What is this?



Don't Use Rainbow Colormaps



Rogowitz & Treinish, IEEE Spectrum, 35(12):52-59. 1998

Self-evaluation

How do you know if you've made a good figure?

- Does it **EASILY** communicate what you want?
- Do readers need to read and re-read your figure legend, or is your message clear?
- Is it accessible to people with poor eyesight or colorblindness?
- Does it ***faithfully*** reflect your data? Beauty + truth

Rules

- Don't make a plot when a table will do
- Represent data with appropriate significant figures
- Use appropriate plot types for the data types
- **Label your axes**
- Title your figures
- Whenever possible, show all the data (or at least the distribution)
- Don't rely on a legend or caption/text
- Don't rely on default plotting conventions
- If possible, show outliers rather than removing them
- Sort categorical data accordingly
- Exploit small multiples to great effect (in R, use faceting)
- Strive to maintain the same color conventions/palettes across all figures
- Start with what you want your plot to look like, then work backwards
- **SHOW. YOUR. CODE.**

Helpful Things in R

- `geom_dotplot`; better than bar plots, typically (see earlier slide)
- Revisit our section on `ggplot2`; I bet you missed a lot...
- Raincloud plots; blog post here
- Using the same theme modifications for all plots? Make a function to store that theme and call it later (more on functions soon!)
- **Google is your friend; Google is your professor**
- Interested in making generative art in `R`? Check out the Twitter accounts of Danielle Navarro and Ijeamaka Anyene; then look at their websites!
- Want examples of good COVID plotting with open `R` code? Check out Chris Prener's Twitter account. If you'd like this as a weekly newsletter, check out his River City Data tracking site!