GRADUATE TRAINING SERIES

PUBLICATION GRAPHICS AND DATA VISUALIZATION

Dr. Abby Benninghoff

Professor of Toxicology

Interim Head: Animal, Dairy and Veterinary Sciences

October, 2022 | 10:00 a.m. - 11:00 a.m.

Merrill-Cazier Library | Room 154 | Zoom

GrTS
Graduate Training Series

School of Graduate Studies
UtahStateUniversity®

Resources

- Edward R. Tufte, *The Visual Display of Quantitative Information* (1983)
- Claus O. Wilke, Fundamentals of Data Visualization (2019) Available as online resource @ https://clauswilke.com/dataviz/
- Kieran Healy, *Data Visualization – A Practical Introduction* (2019)
- Sarah Schwartz, “Powerful Publishable Plots” presentation, October 2018 at Utah State University, www.cehs.usu.edu/research/statstudio

Graphics reveal data, communicate complex ideas and dependencies with clarity, precision and efficiency

-Edward R. Tufte
"The Visual Display of Quantitative Information"

The BEST graph is one which:
"gives to the VIEWER
the greatest number of IDEAS
in the shortest TIME
with the least INK
in the smallest SPACE."

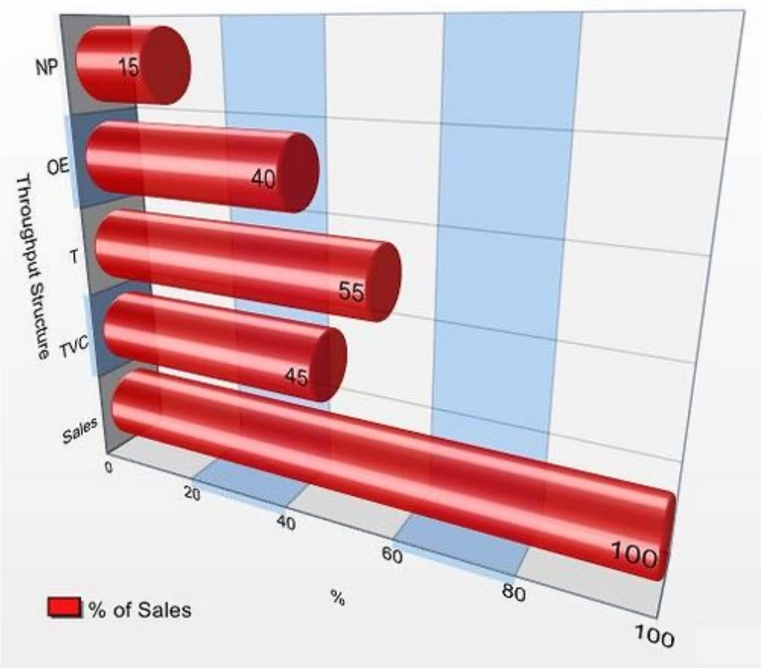
What makes figures bad?

Issues of bad graphic design

- Aesthetic
- Substantive
- Perceptual

Source: *Data Visualization, a Practical Introduction* by Kieran Healy

Bad taste



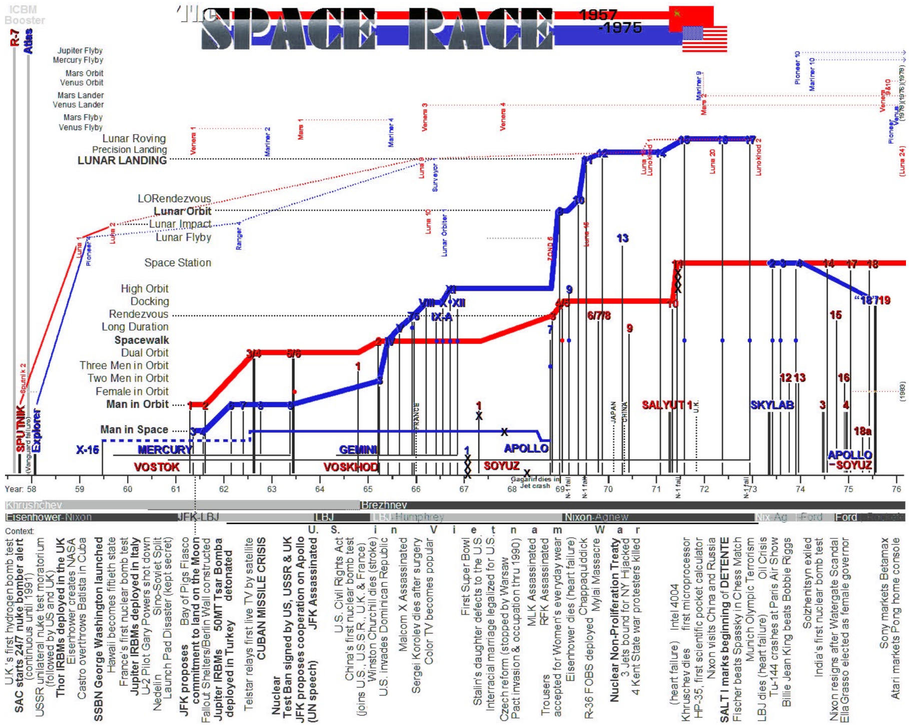
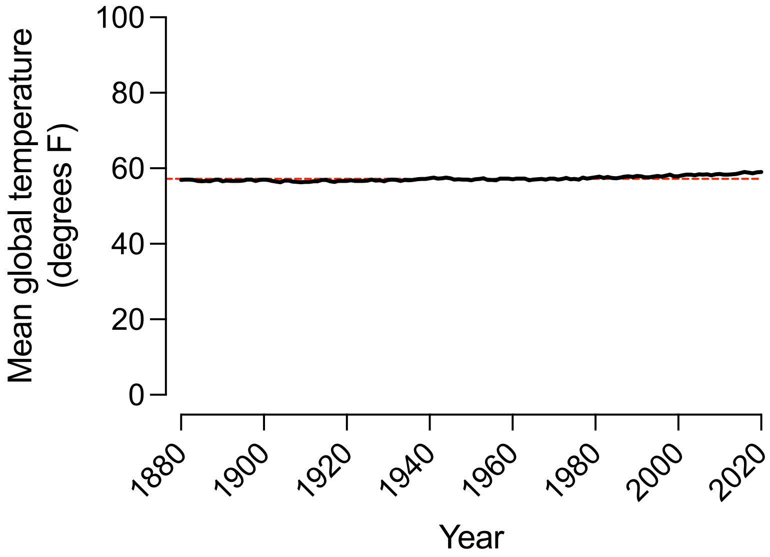


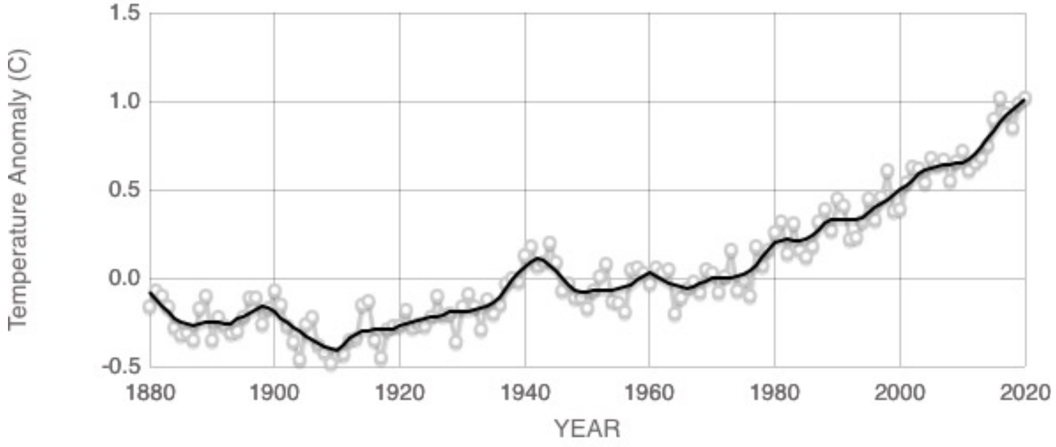
Image credit: “Throughput Accounting”, CC BY 3.0; “Moon Landing”, Wikimedia Commons

Substantive issues – bad data or misleading presentation



What problem do you see with this presentation?

Does the design of the plot introduce clear, intentional bias?

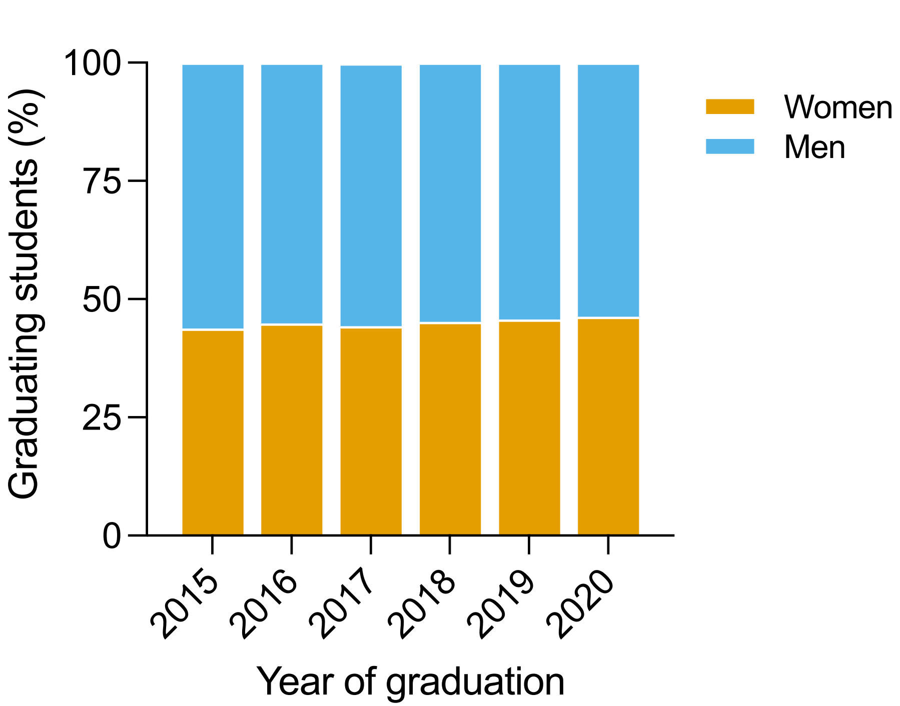


Source: climate.nasa.gov

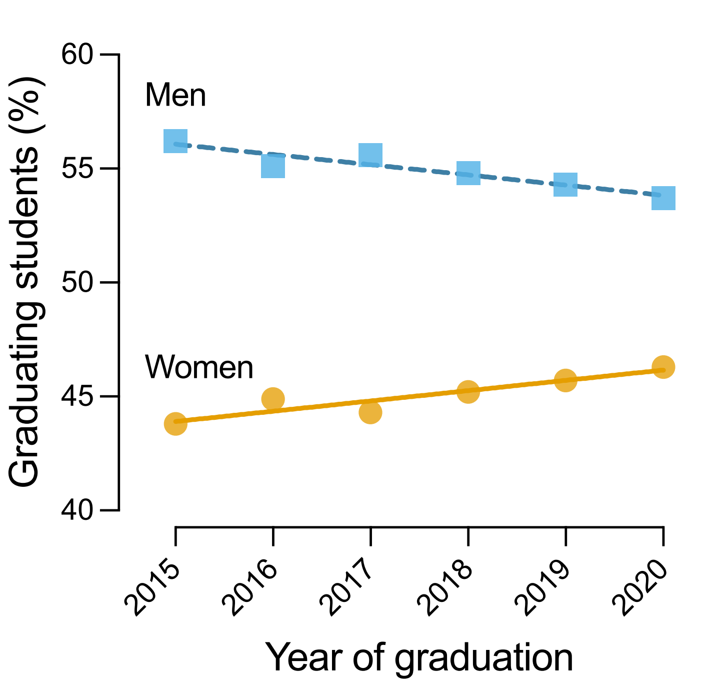
NASA and other climate groups convert the data to "temperature anomaly" (i.e., deviation from a reference temperature) to avoid problems with the scale and bias

Data source: NASA's Goddard Institute for Space Studies (GISS), Global Land-ocean Temperature Index, converted to Fahrenheit.

Perception issues



Graduates in STEM disciplines



Graduates in STEM disciplines

Think of the vastly different messages these two plots send.

Is one right or wrong?

Data Source: Utah State University, Office of Assessment, Analysis, and Accreditation

Remove

to improve

(the **data-ink** ratio)

Created by **Darkhorse Analytics**

www.**darkhorseanalytics**.com

A very brief review of graphics fundamentals

Linking data values to graphical aesthetics

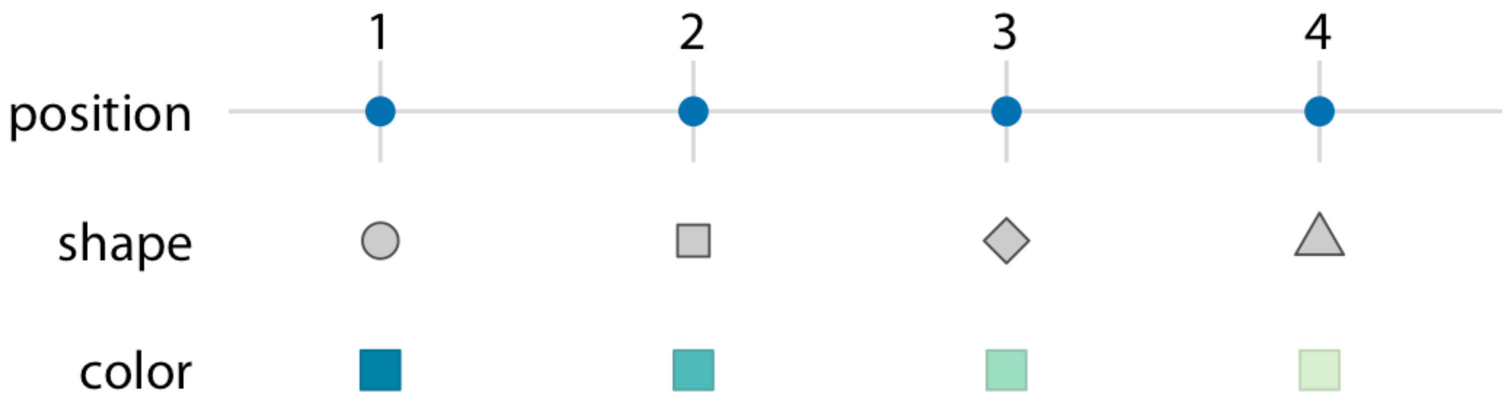


Image credit: https://clauswilke.com/dataviz/

High quality figure that links data in 5 scales using these three aesthetics

- **X** axis: displacement
- **Y** axis: fuel efficiency
- symbol **color**: power
- symbol **size**: weight
- symbol **shape**: cylinders

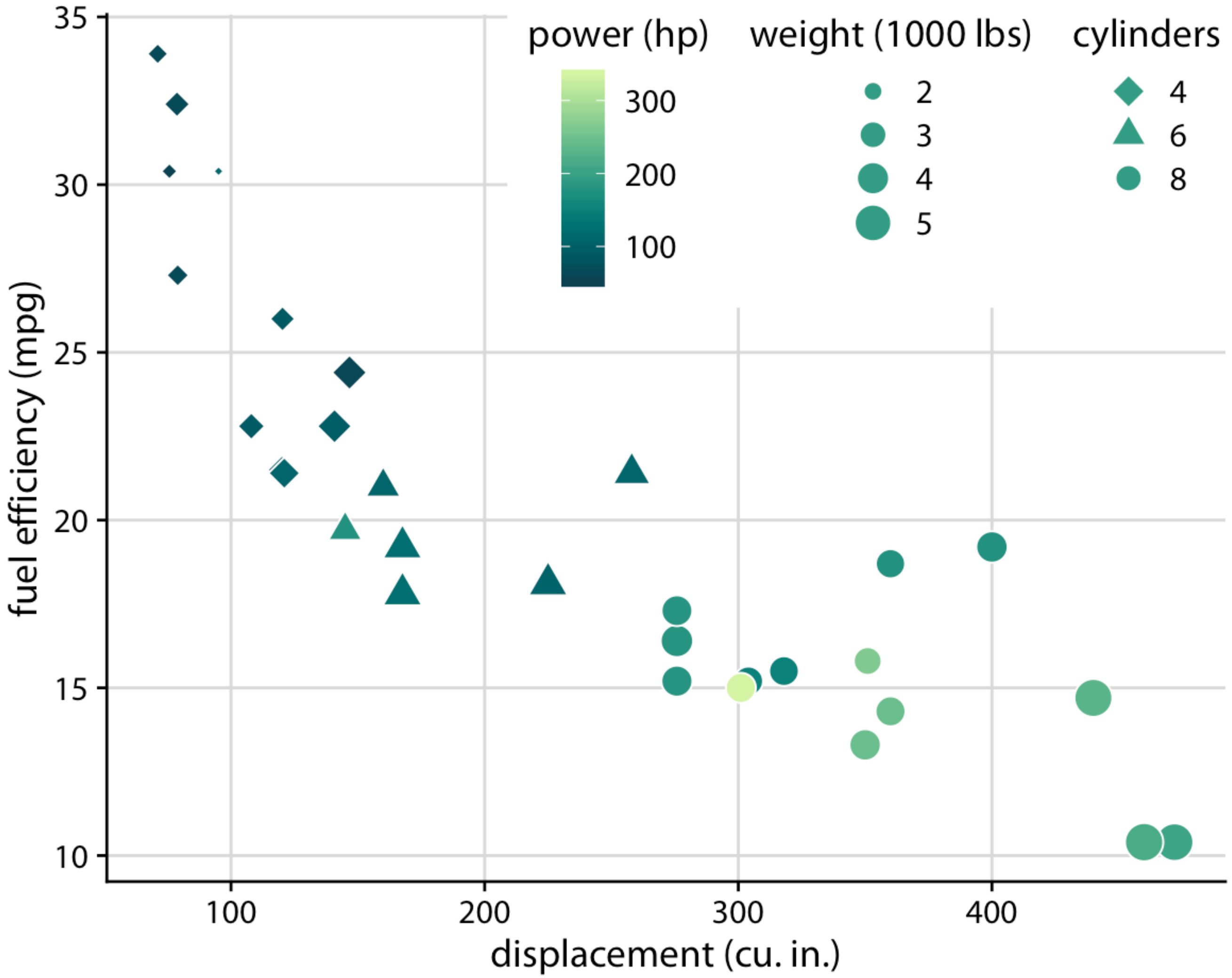
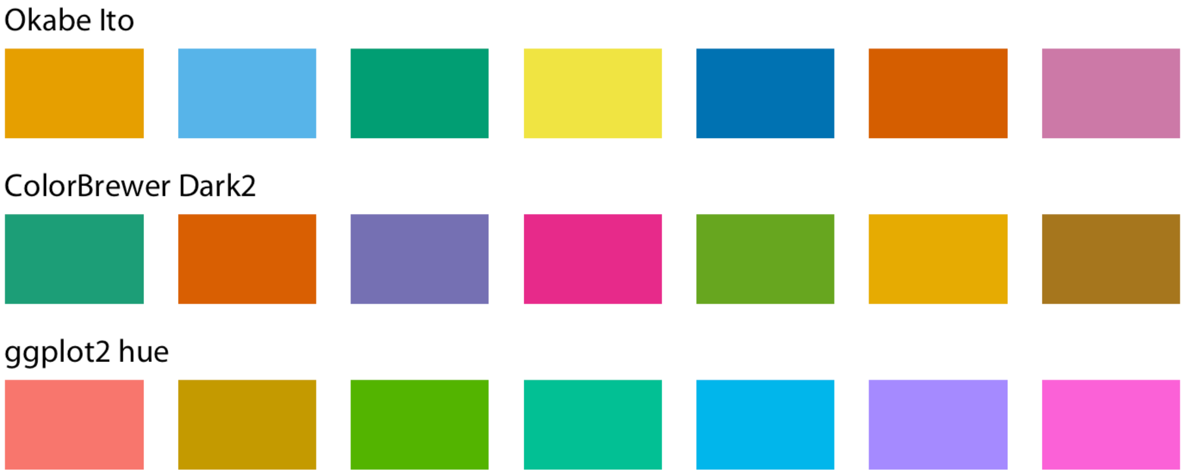


Image credit: https://clauswilke.com/dataviz/

Color as a tool

Qualitative color scales to distinguish



Sequential color scales to represent values

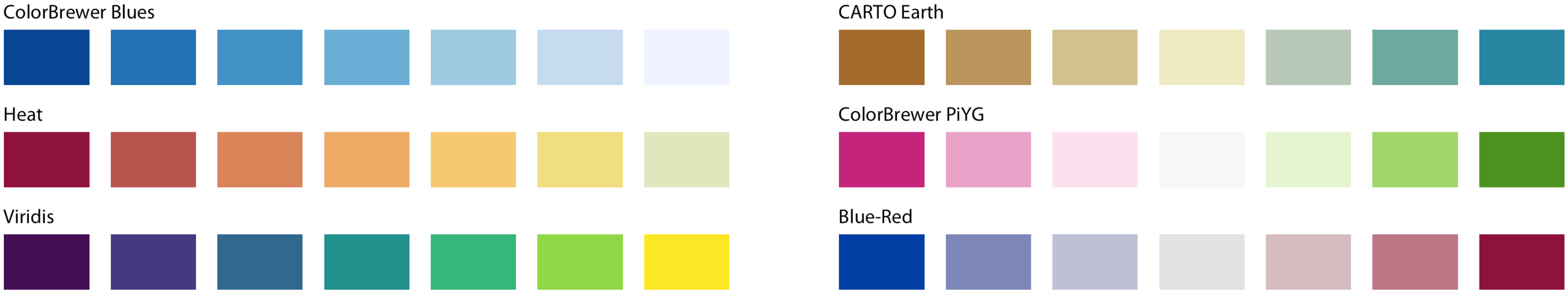


Image credit: https://clauswilke.com/dataviz/

Visualizing amounts

relationships between numeric and categorical variable

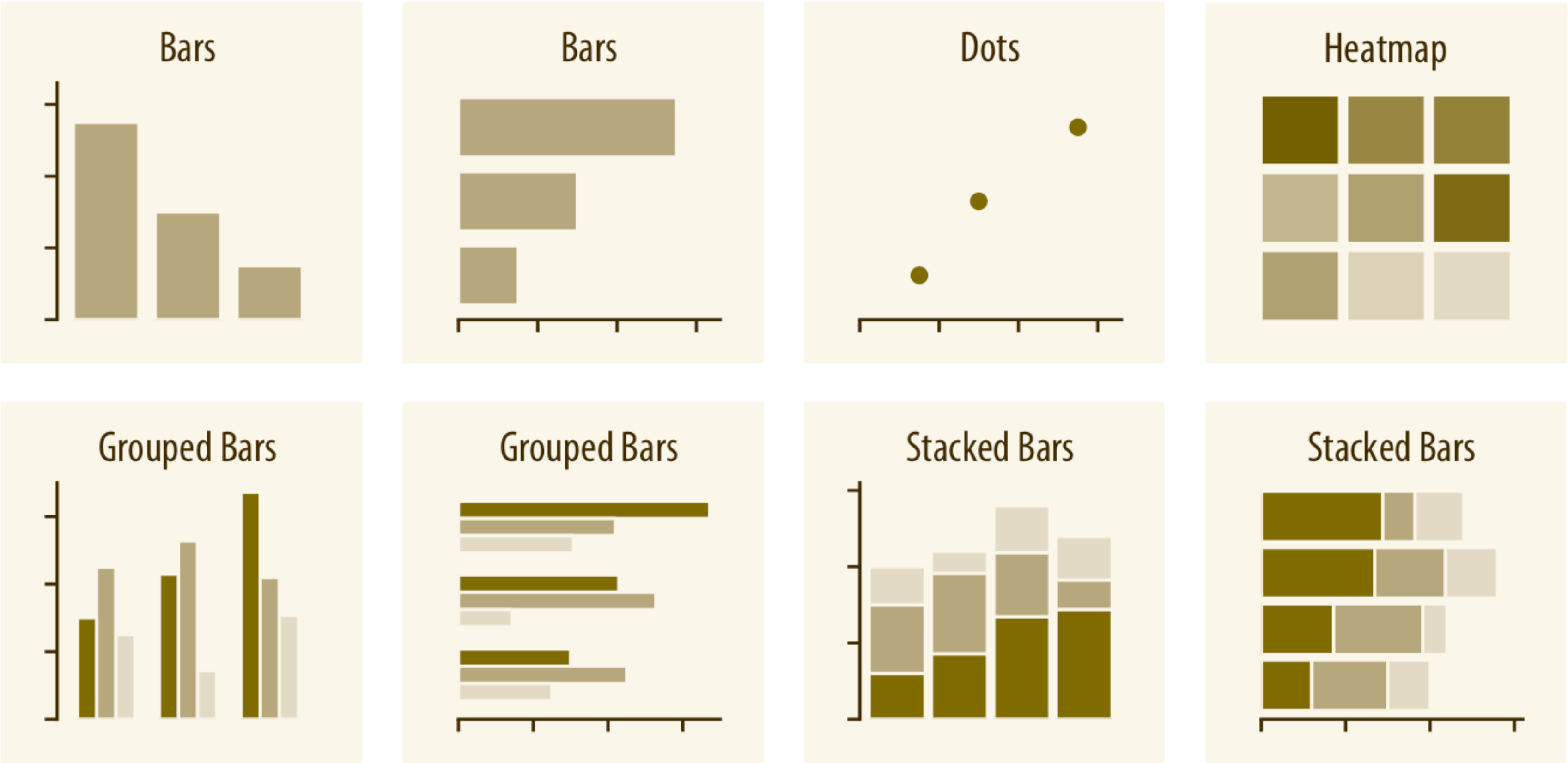


Image credit: https://clauswilke.com/dataviz/

XY plots

data sets with two or more continuous variables

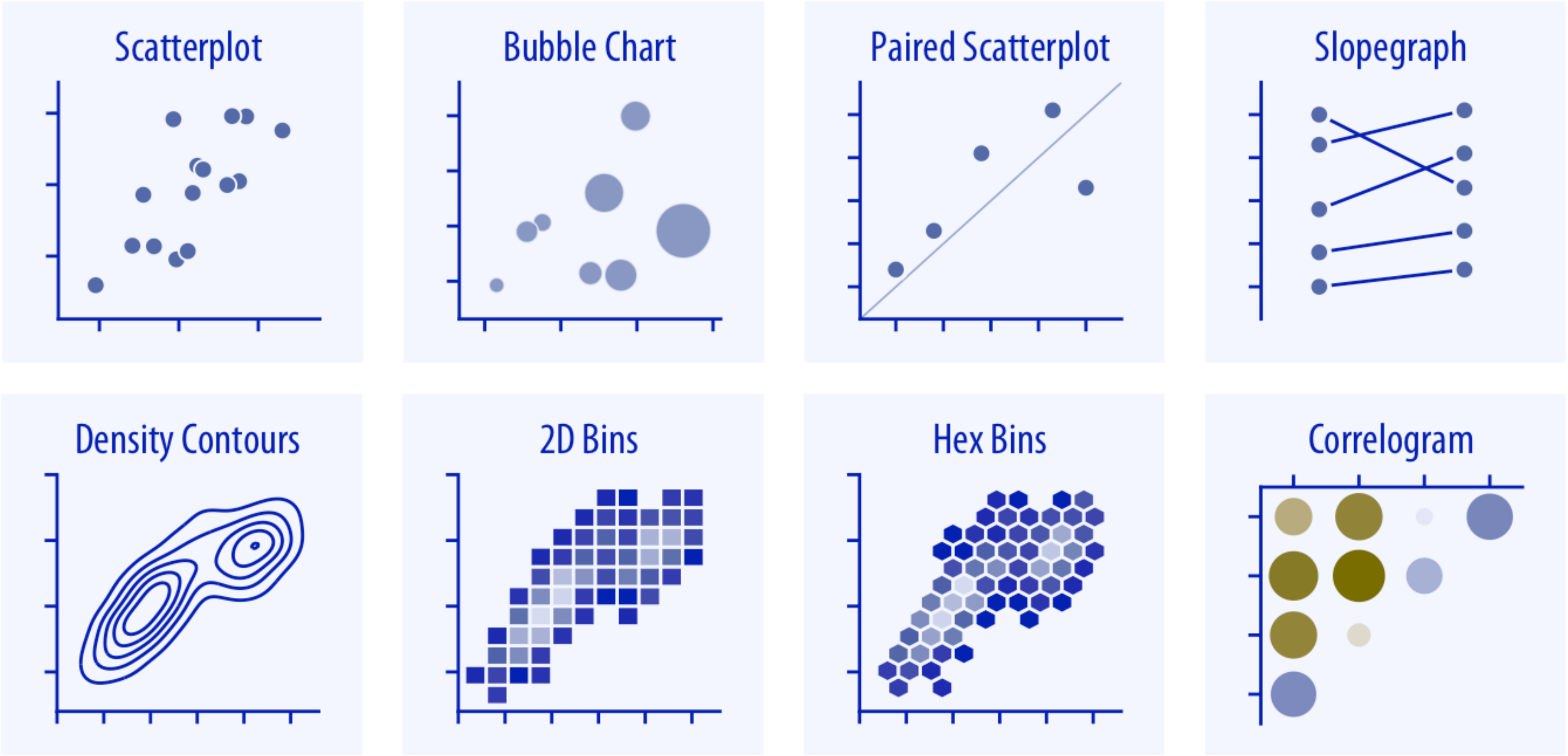


Image credit: https://clauswilke.com/dataviz/

Distributions

distribution of values within data sets

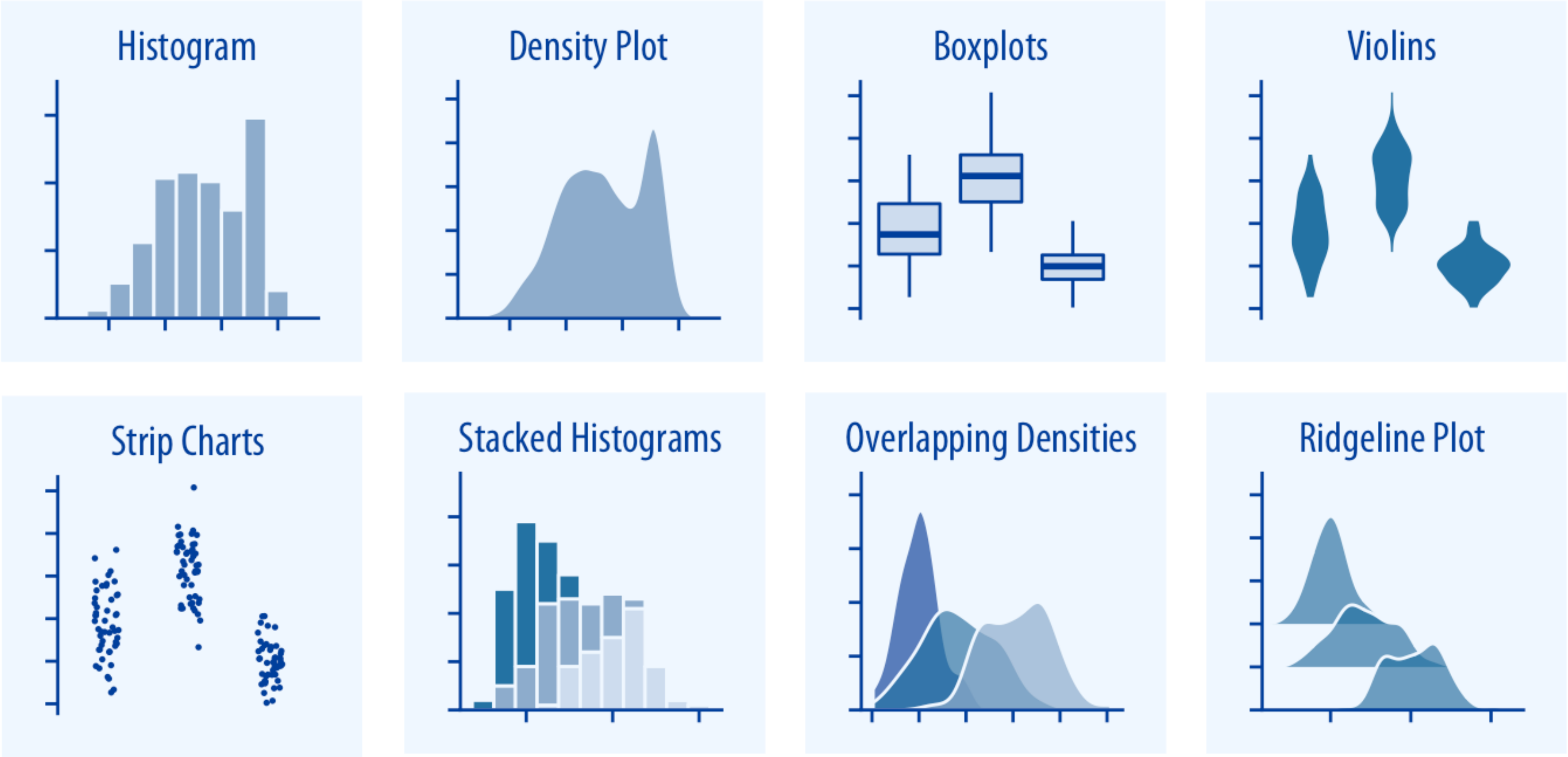


Image credit: https://clauswilke.com/dataviz/

Anatomy of a boxplot & violin plot

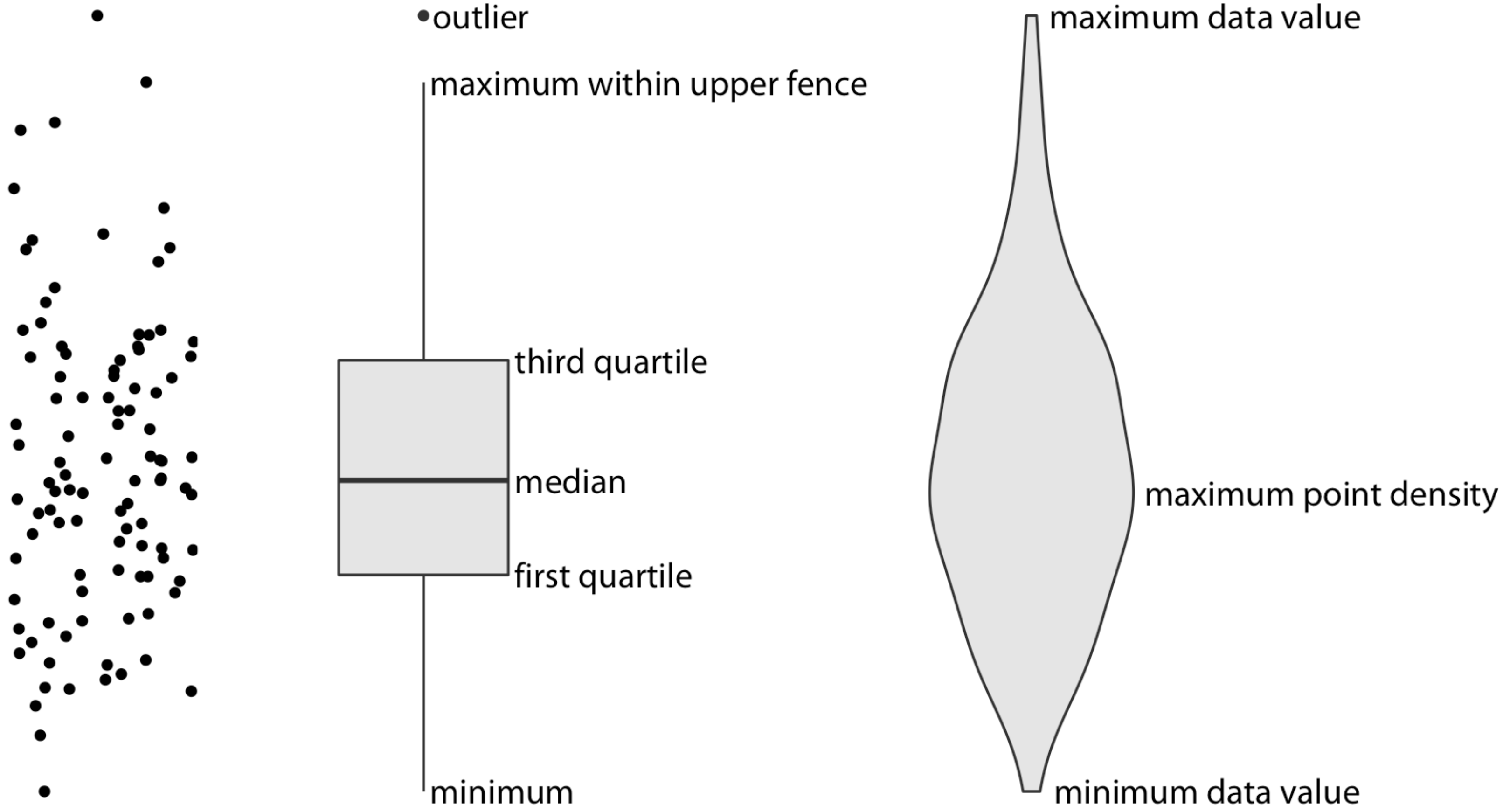


Image credit: https://clauswilke.com/dataviz/

Proportions

parts of a whole

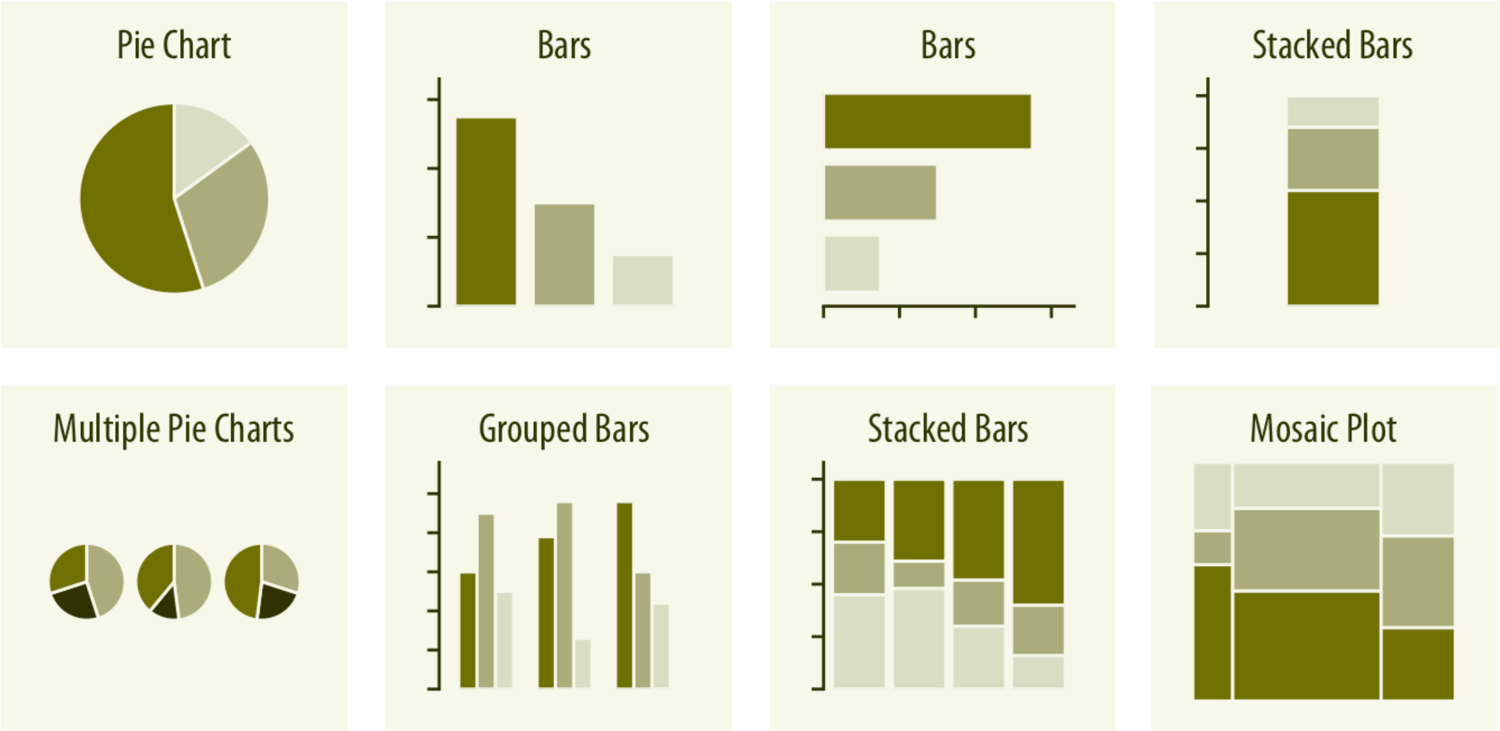


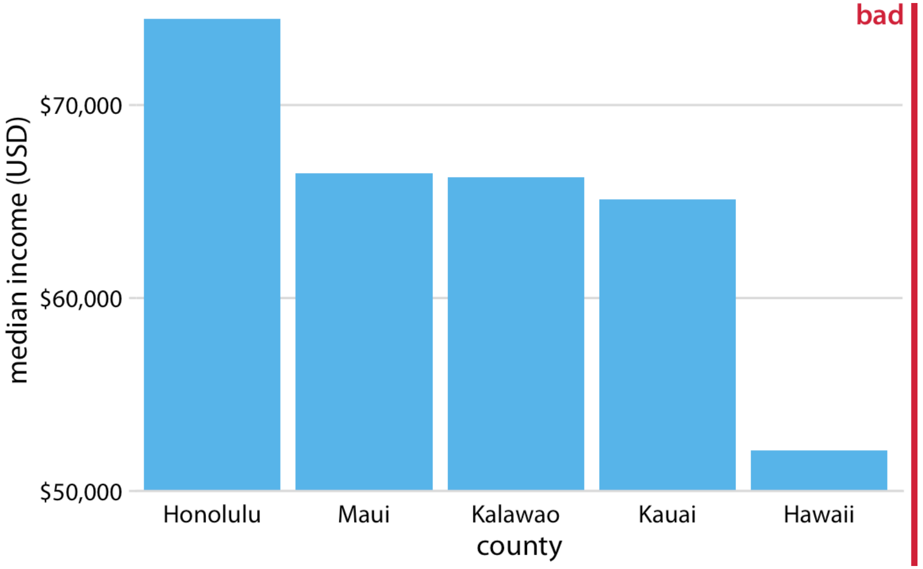
Image credit: https://clauswilke.com/dataviz/

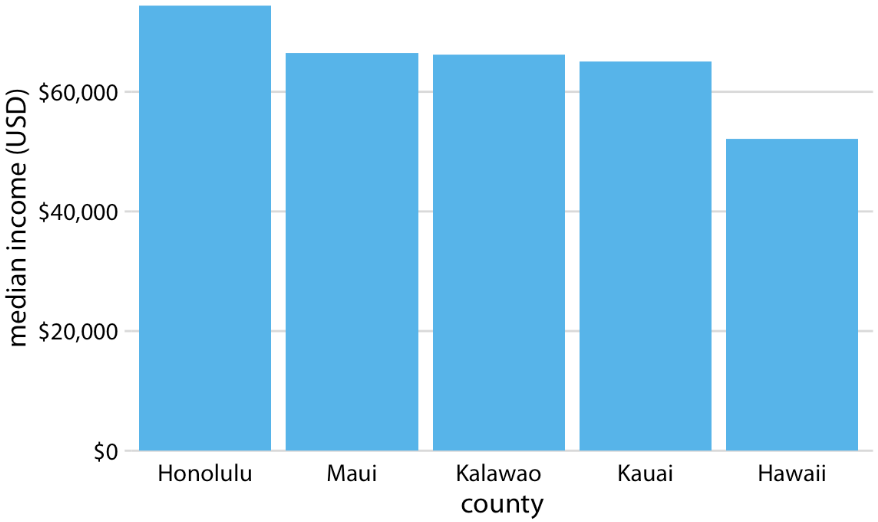
Graphics magic to improve data transparency

Tips & tricks

The principle of proportional ink

The sizes of shaded areas in a visualization need to be proportional to the data values they represent.

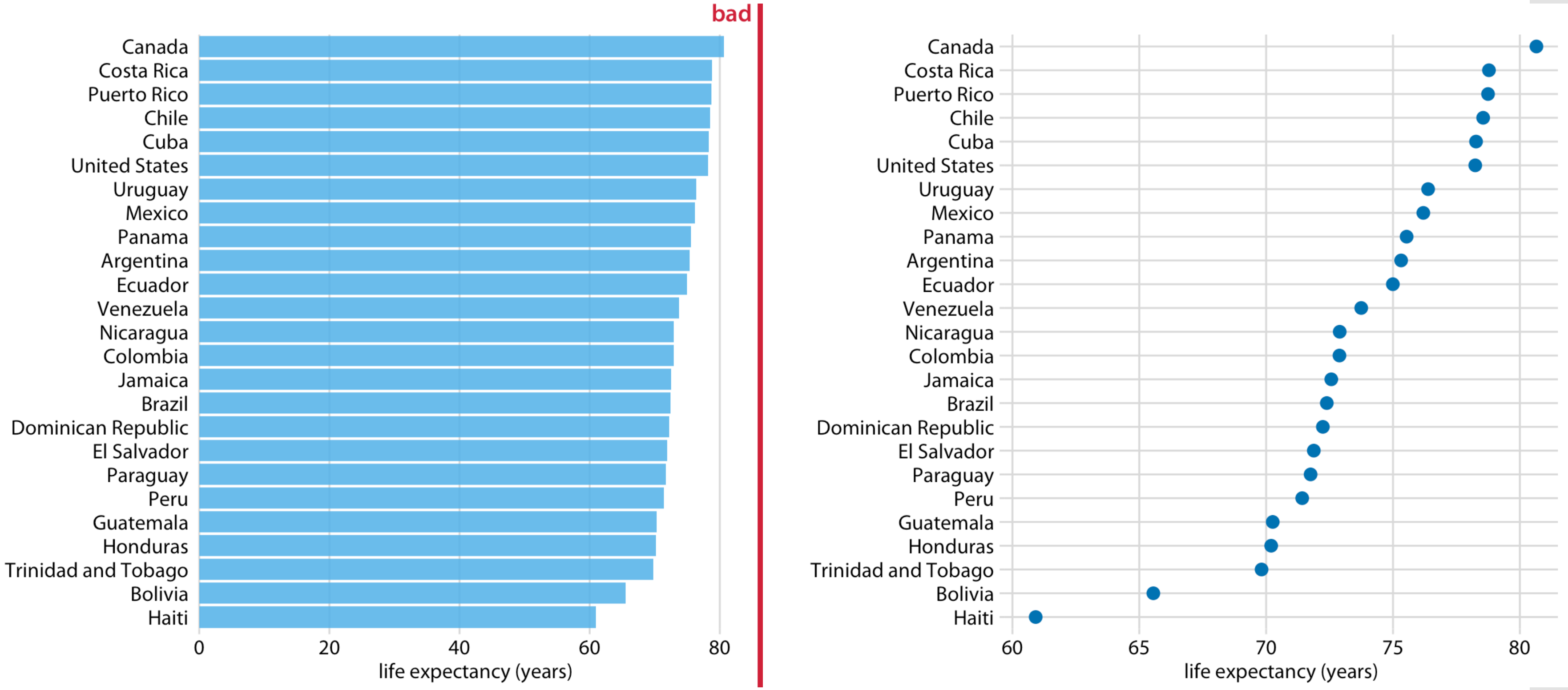




- We judge differences in these data by the areas of the bars.
- The top plot suggests a much greater relative difference in median income between Honolulu and Hawaii than there is in reality
- Bar plots like this should be anchored at zero.

Image credit: https://clauswilke.com/dataviz/

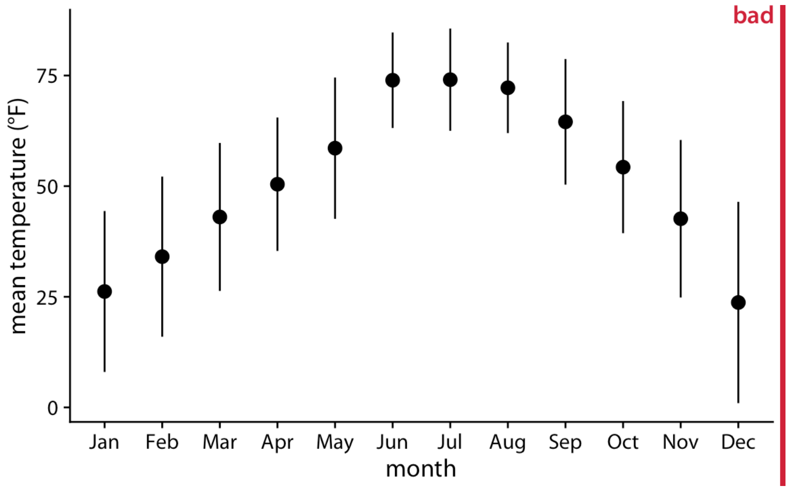
Alternative to the bar chart for showing amounts



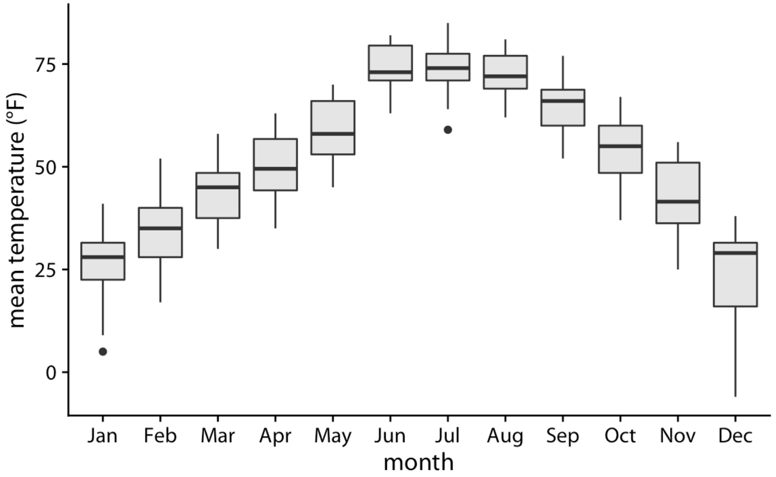
- Bar charts should be anchored at ZERO. Challenge when differences are small in magnitude
- Consider a **dot plot** as alternative
  - Can appropriately adjust axis to fit data range

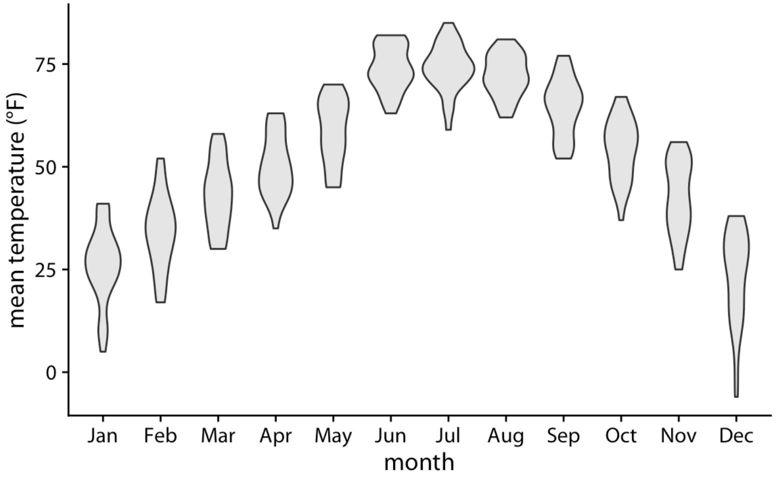
Image credit: https://clauswilke.com/dataviz/

Visualizing many distributions



- What is the intent of the error bars here?
- Error bars are typically used to visualize uncertainty of an estimate, not variability in a population



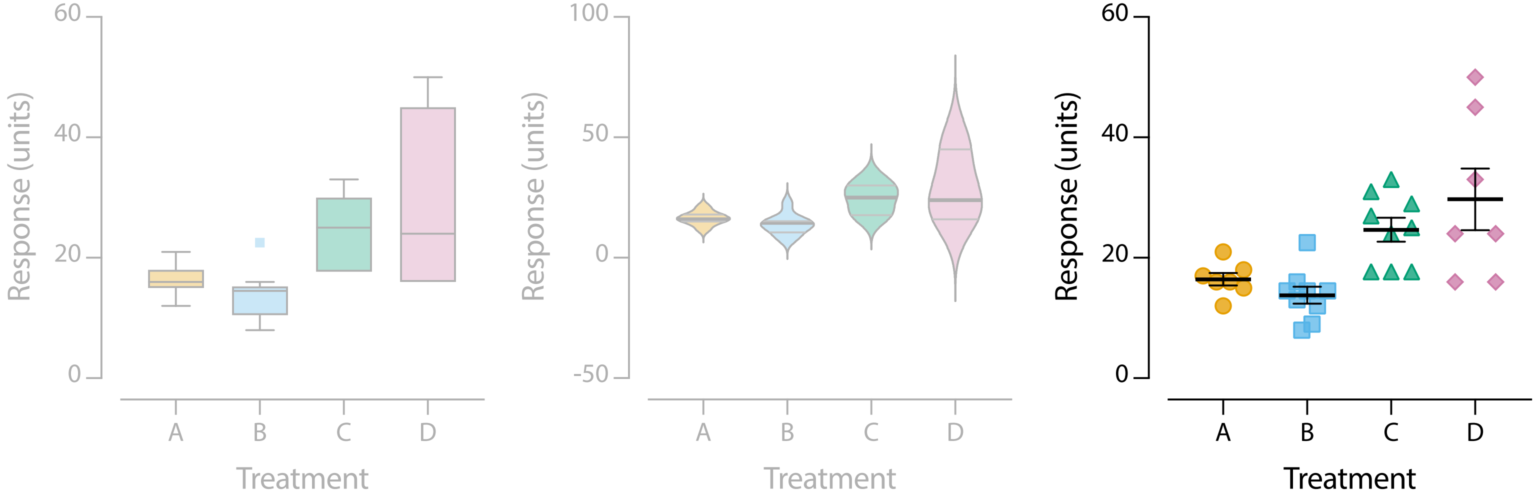


Box plots or violin plots show the variability in the population. Note that the distributions are not symmetric!

Image credit: https://clauswilke.com/dataviz/

Reveal your raw data for greater transparency

When the dataset is too sparse to justify visualization of a violin plot (or box plot), show the raw data as individual points instead (or in addition)



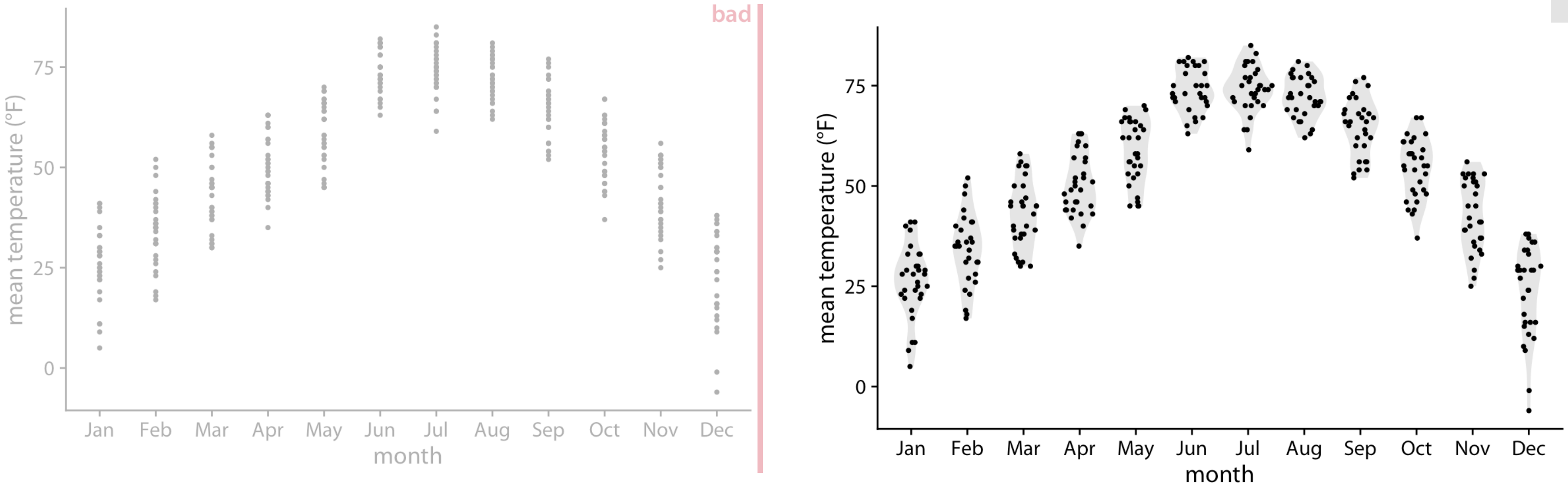
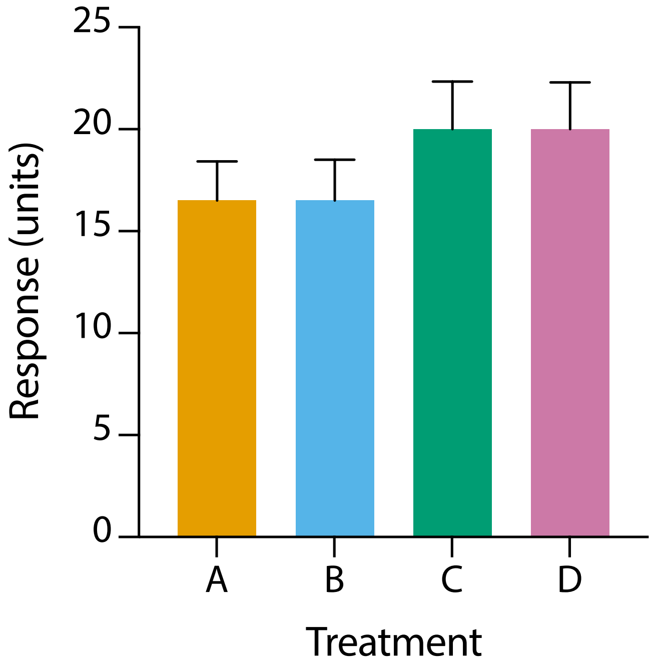
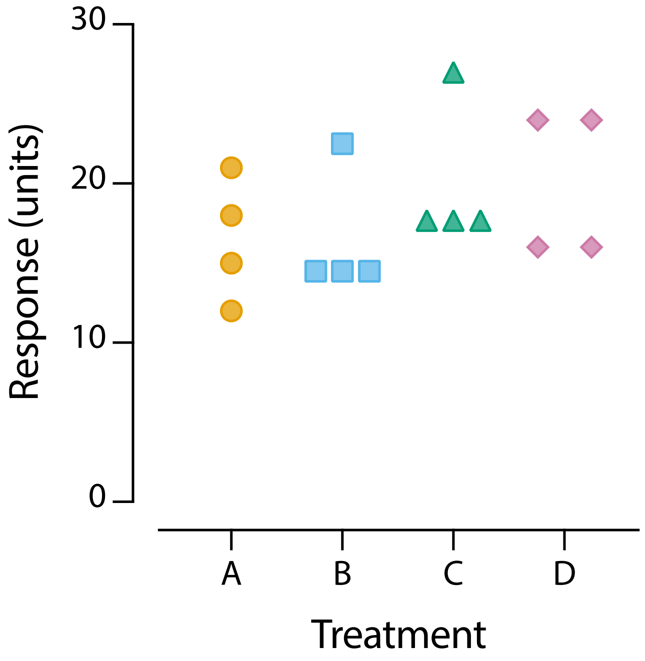


Image credit: https://clauswilke.com/dataviz/

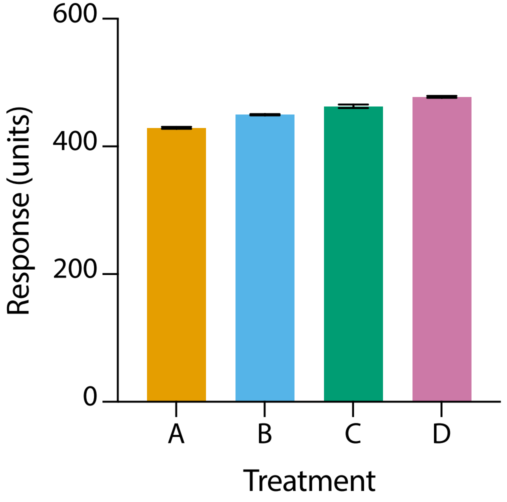
The notorious “dynamite plot”

- low data-to-ink ratio
- hide raw data
- assume symmetric SD, SEM, CI
- disguise the data distribution

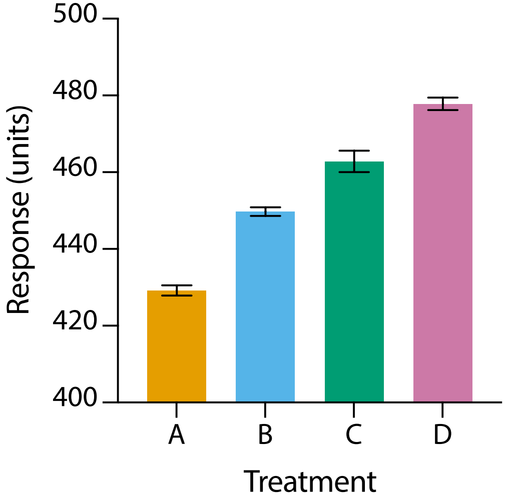




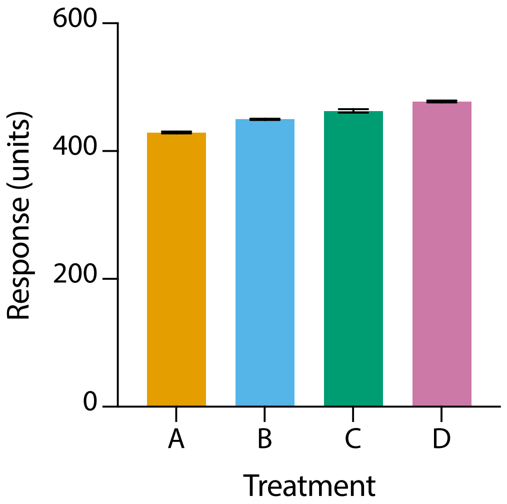
Transforming a bad “dynamite” plot



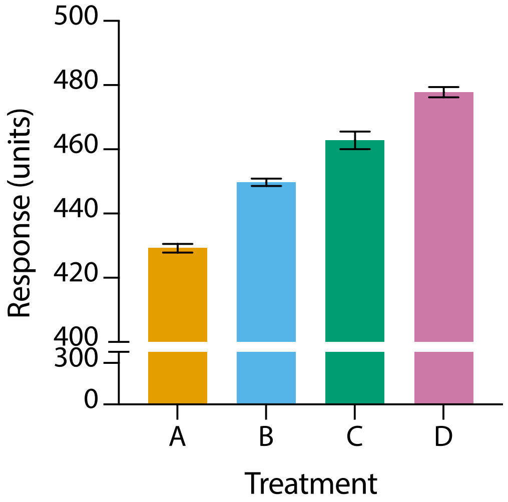
Adjust Y-axis range?



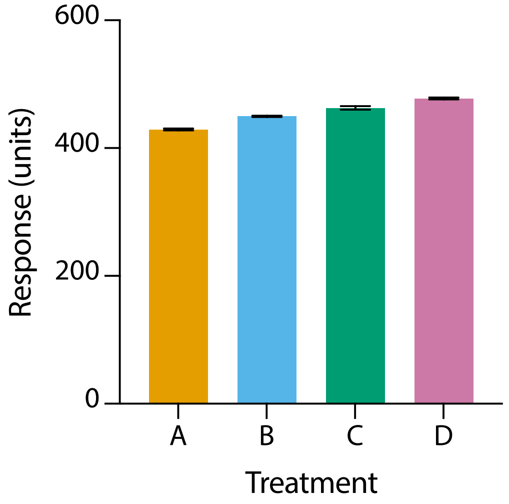
Transforming a bad “dynamite” plot



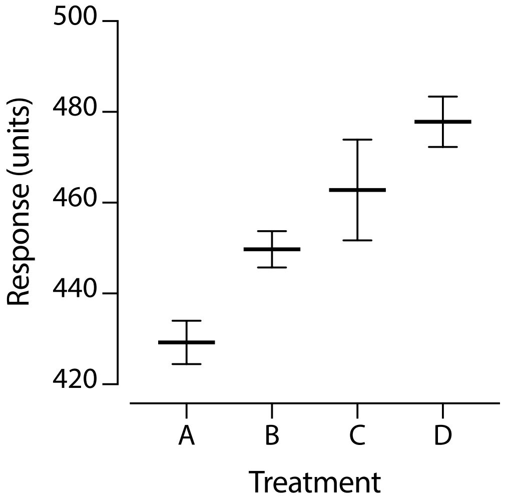
Add axis break?



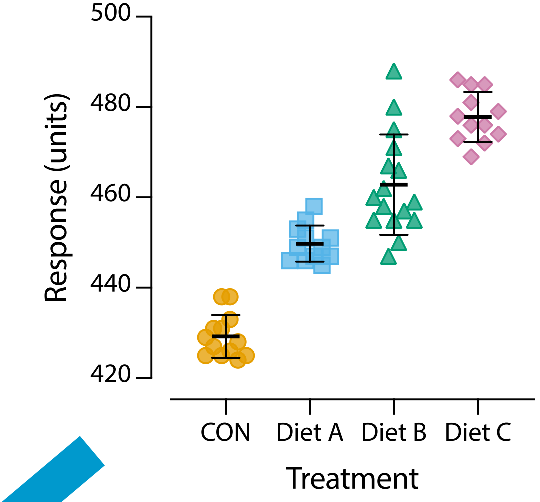
Transforming a bad “dynamite” plot



Remove data-ink?

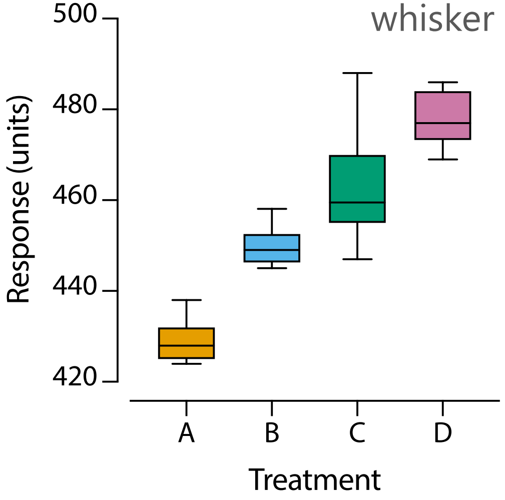


Show values

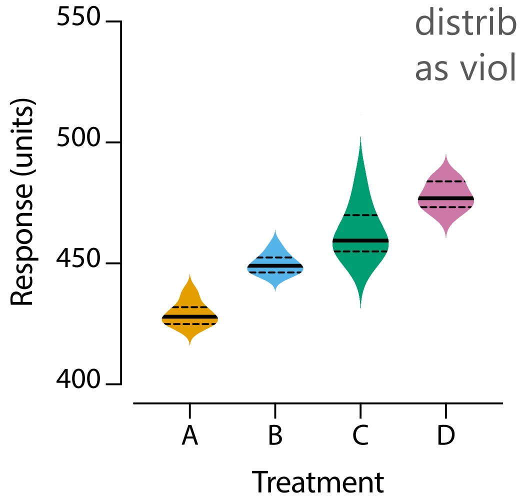


Publishable!

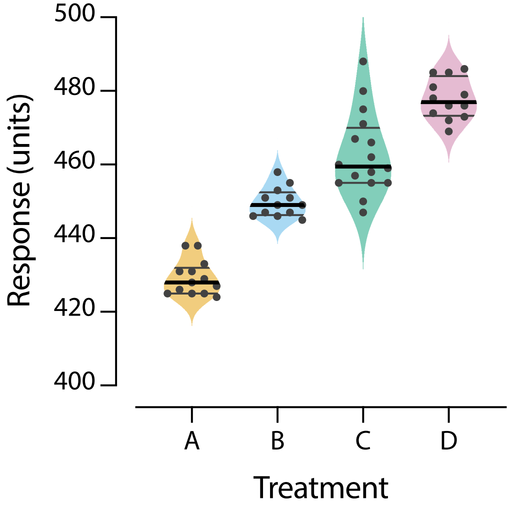
Summarize data as box-whisker



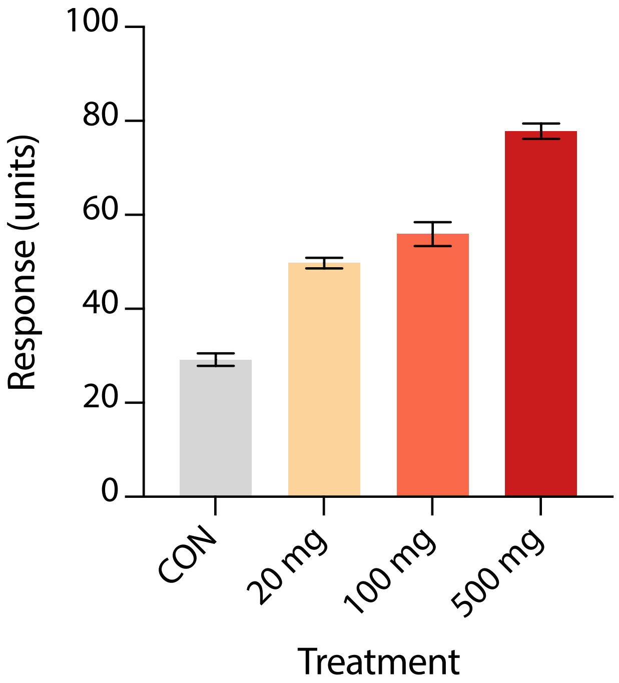
Show distribution as violin plot

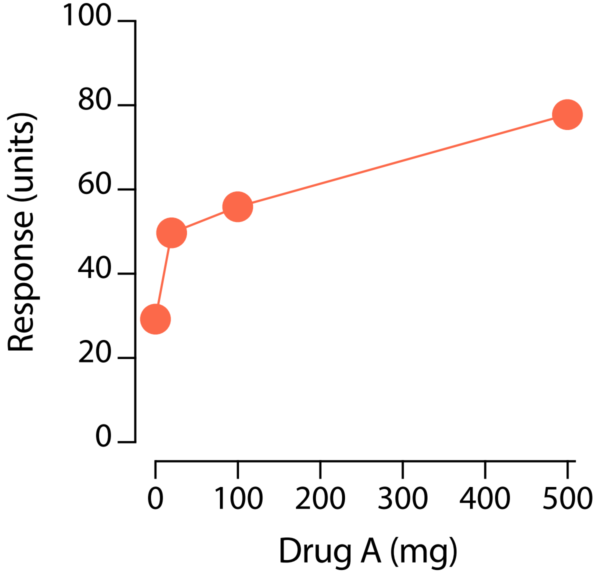


Reveal data & distribution

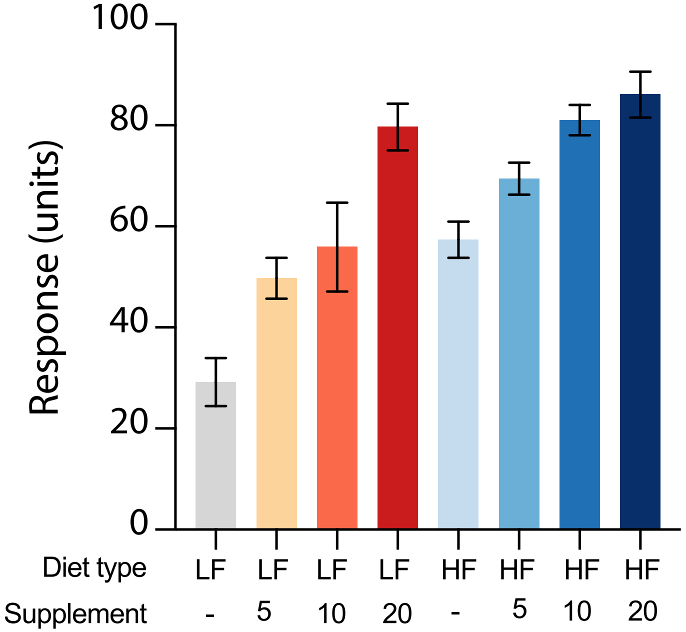


Use the right kind of plot for your data





- Data are plotted here as categorical. Are they really?
- These data appear to be continuous:
  - 0, 20, 100, 500 mg
- Plotting continuous data on categorical axis misrepresents the data
- Distance between continuous values may not be even, although often plotted as such
- Always plot continuous data on a continuous scale!



Study design main factors

- diet (LF, HF)
- supplement (none, 5, 10, 20) →

2x4 design → grouped plot!

(**0**, 5, 10, 20) . . . a continuous scale →

two continuous scales → XY plot!

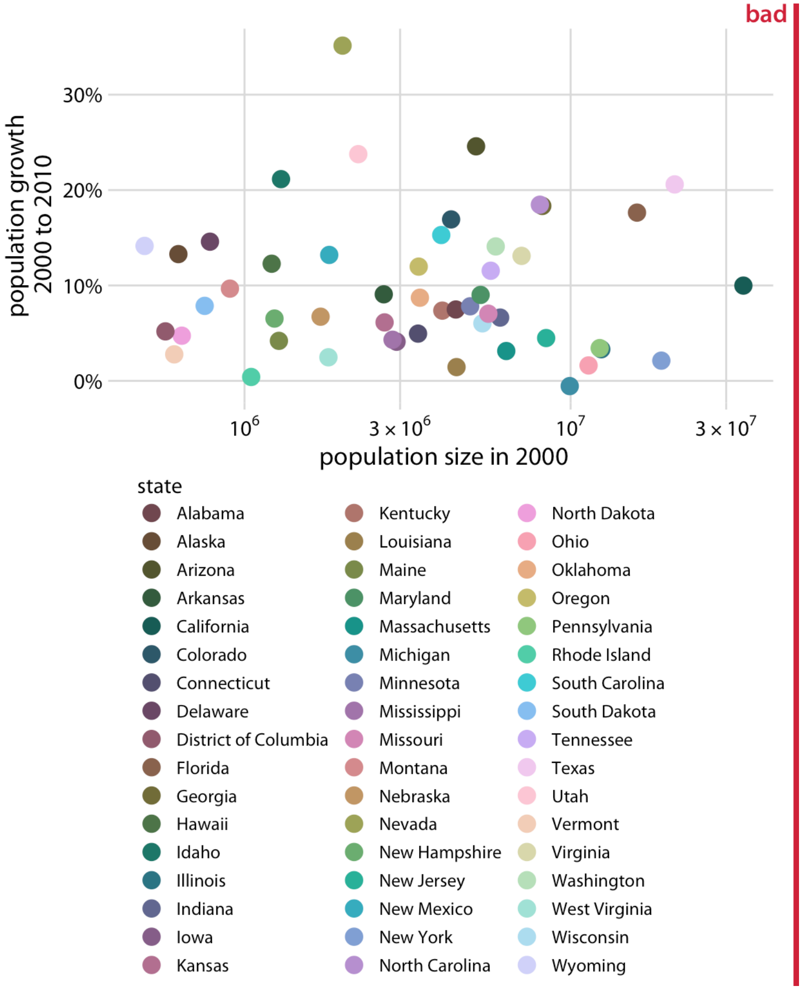
Reduce the non-data-ink

Notice how much cleaner this plot looks, which allows for the differences in trends to be apparent.

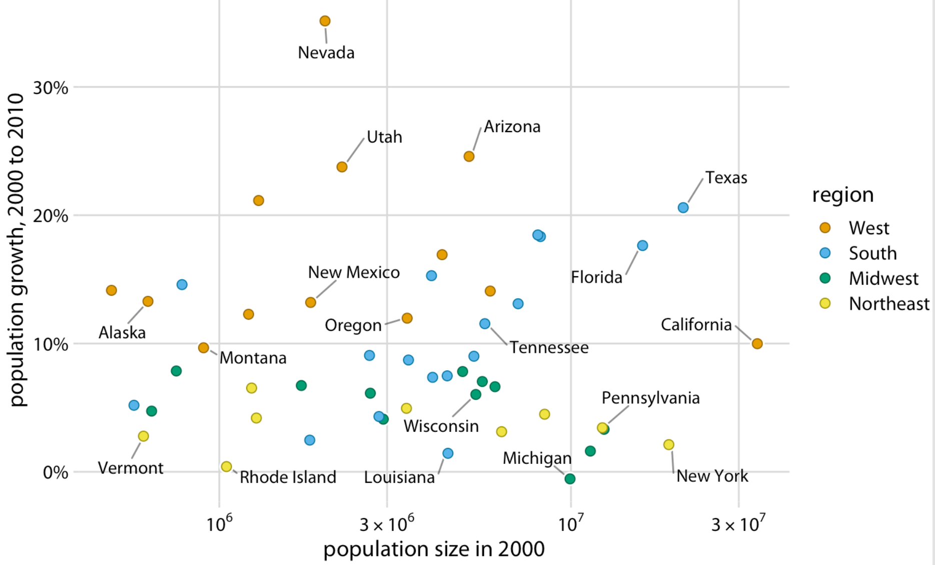
The supplement appears more effective in subjects consuming a low-fat diet

Design pitfalls to avoid

Encoding too much or irrelevant information



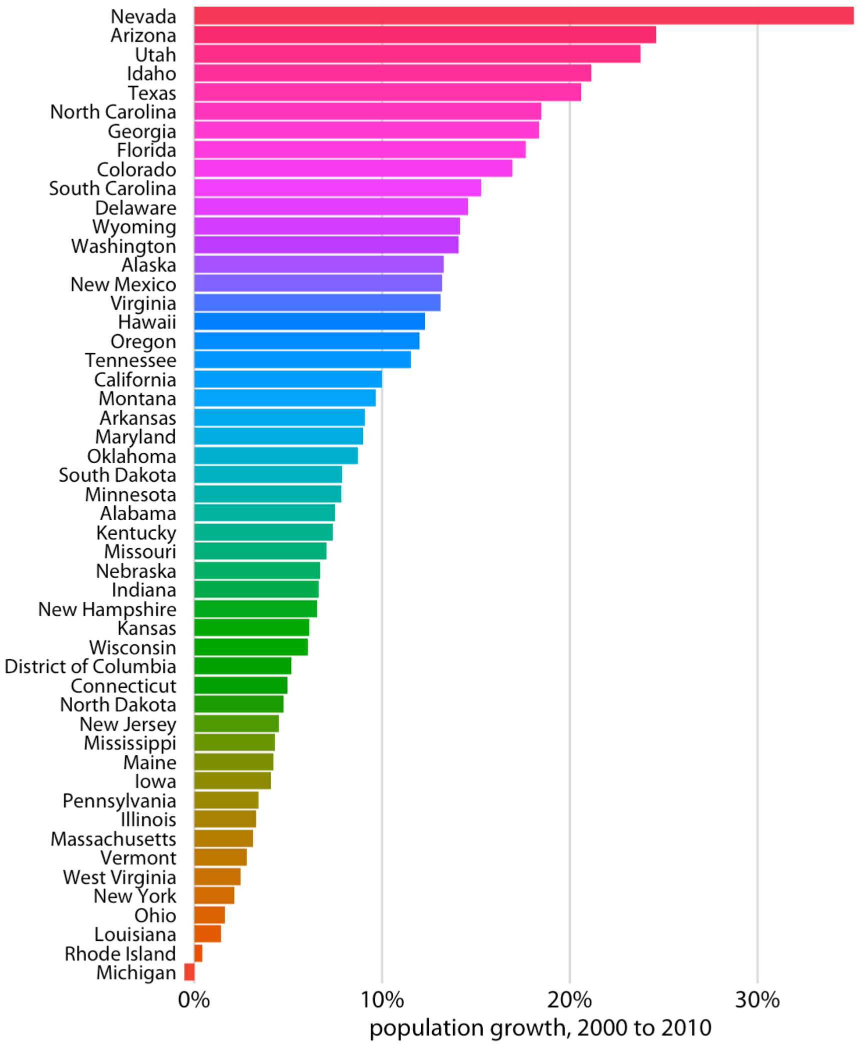
Can you compare the location of Colorado vs. Connecticut on this plot?



Solution: Group colors by region, highlight key points of interest

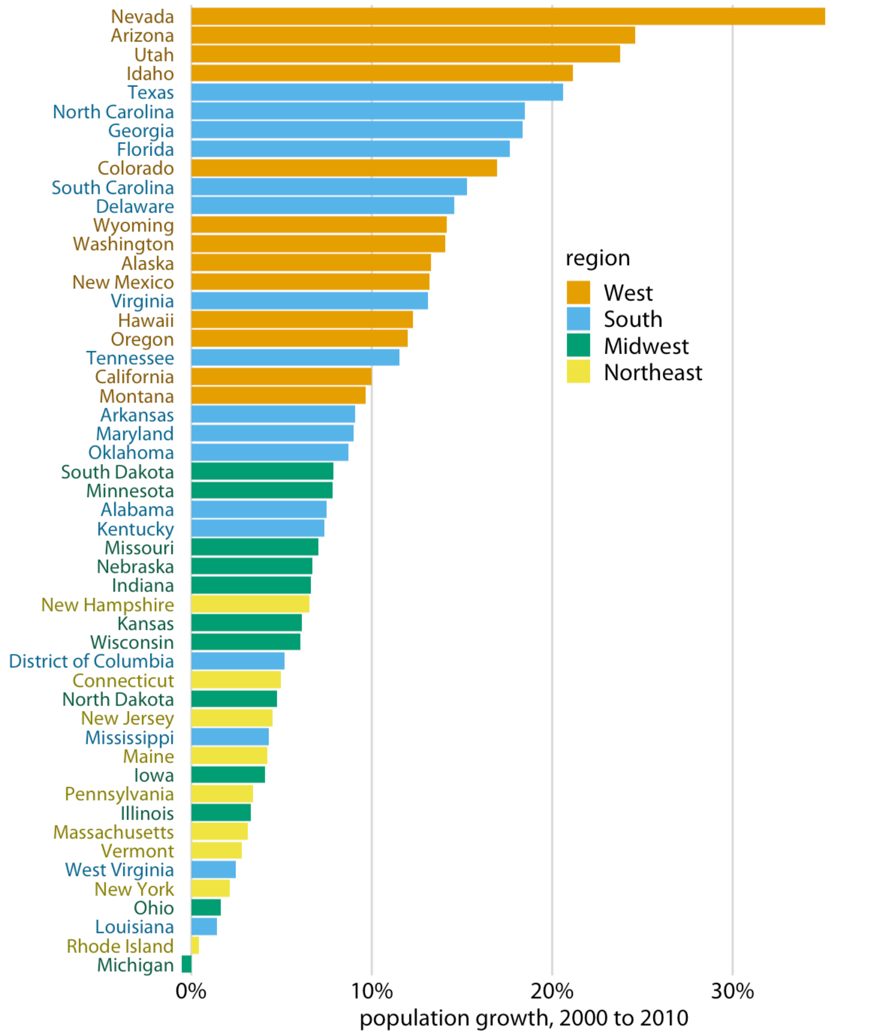
Image credit: https://clauswilke.com/dataviz/

Color for the sake of color



What is the message of this plot?

What function does this rainbow color scheme play?

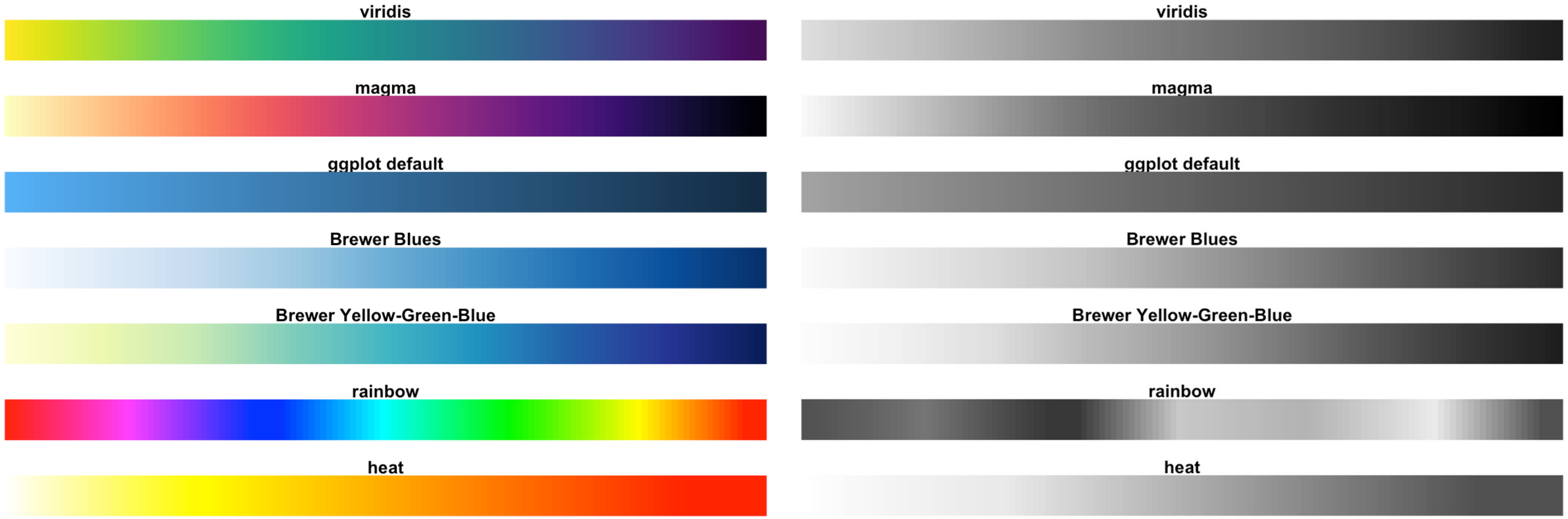


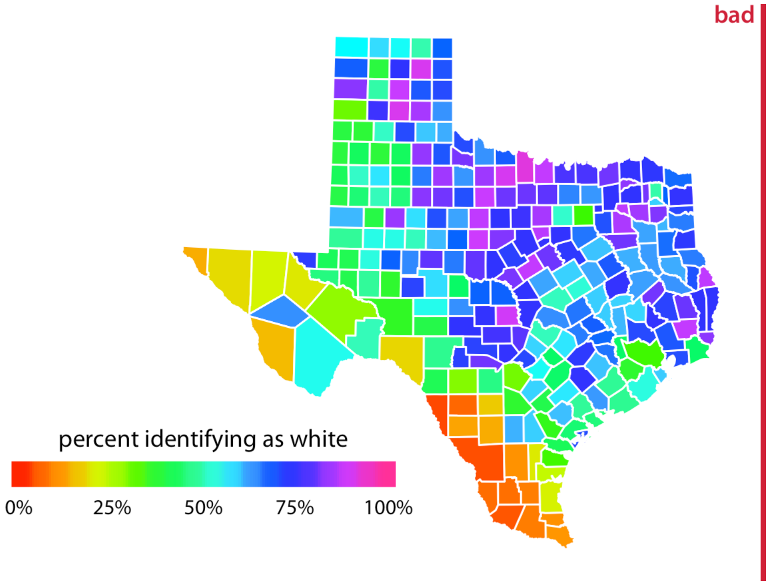
Solution: Grouping color by region reveals that the Western and Southern states experienced more rapid population growth than Northeast and Midwest

Image credit: https://clauswilke.com/dataviz/

Use monotonic color scales

Certain colors stand out in the traditional rainbow scale, which emphasizes the wrong data





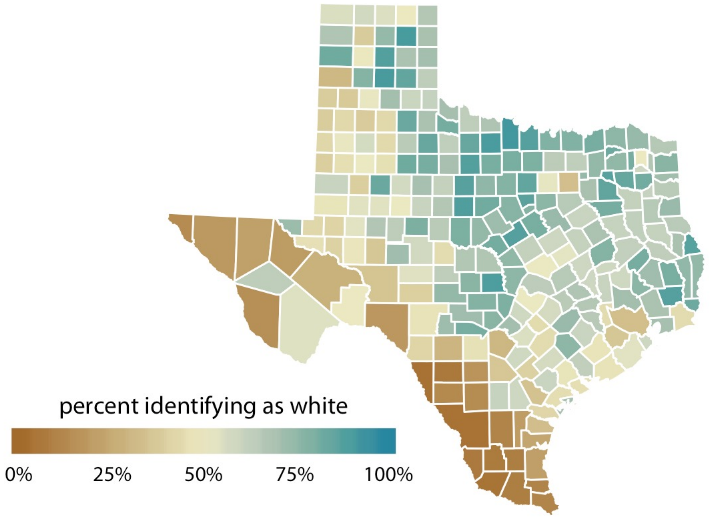
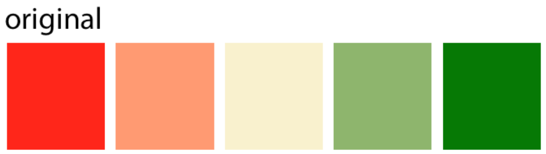
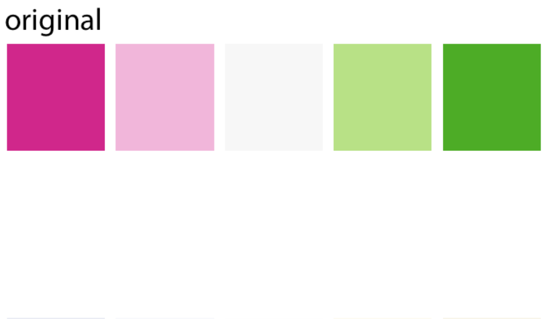


Image credit: https://clauswilke.com/dataviz/

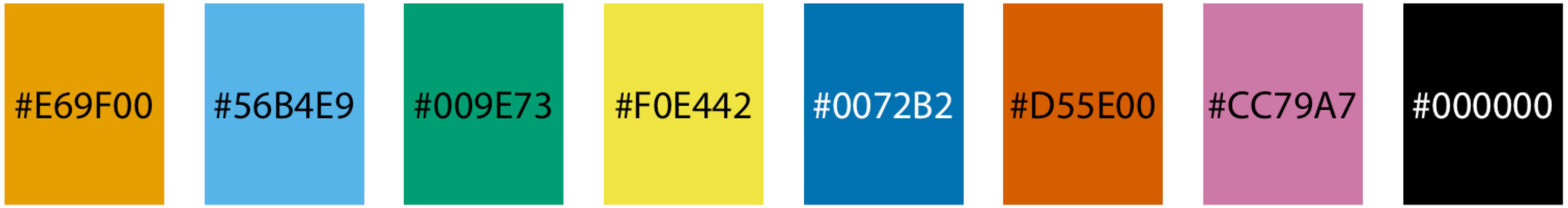
Design with visually impaired in mind



A scheme with red and green is difficult for some colorblind to visualize



This pink/green scheme (R color brewer) works for all types of color blindness

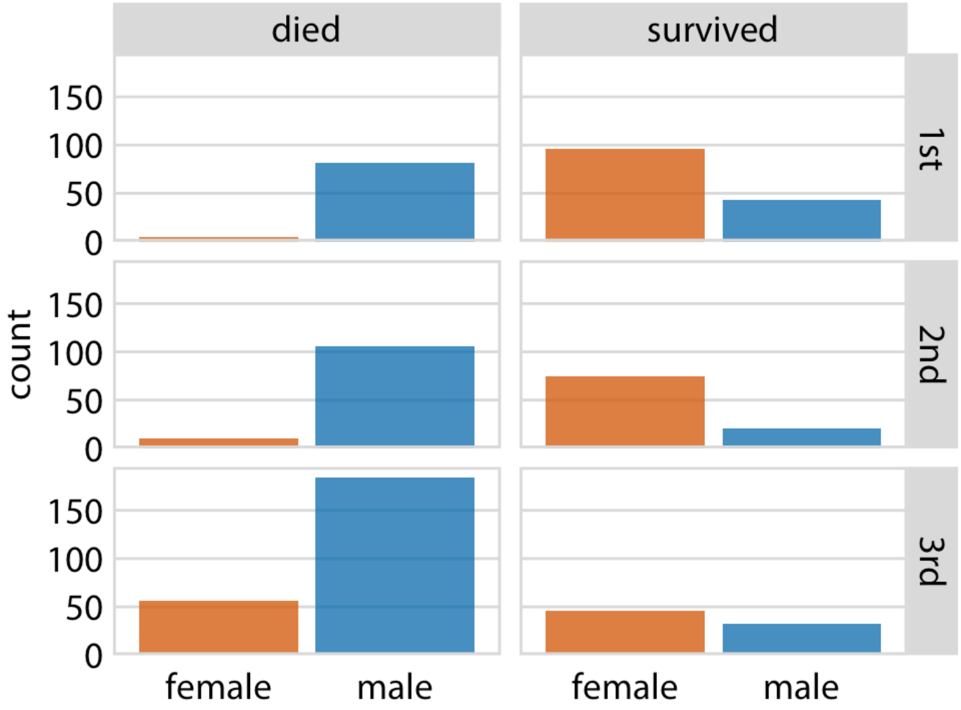


Recommended color palette for all color-vision deficiencies (https://jfly.uni-koeln.de/color/). Hexadecimal codes are shown

Image credit: https://clauswilke.com/dataviz/

Multi-panel plots

Small multiples are a powerful tool to visualize very large amounts of data at once

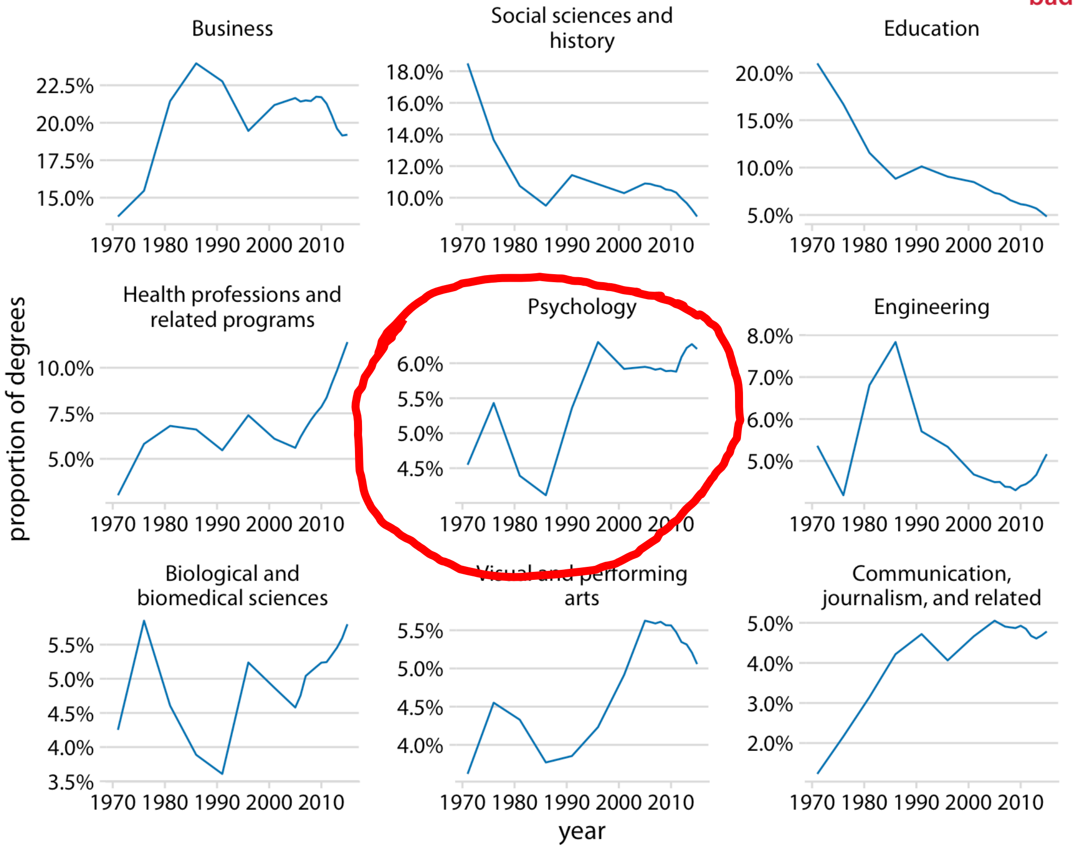


Status: alive vs. dead
Cabin class: 1st, 2nd, or 3rd
Gender: female vs. male

This data set depicting the fate of the passengers of the Titanic works nicely as a multi-panel plot, created using "faceting" in R.

Image credit: https://clauswilke.com/dataviz/

Multi-panel plots should be consistent in scaling for easy and accurate interpretation



bad

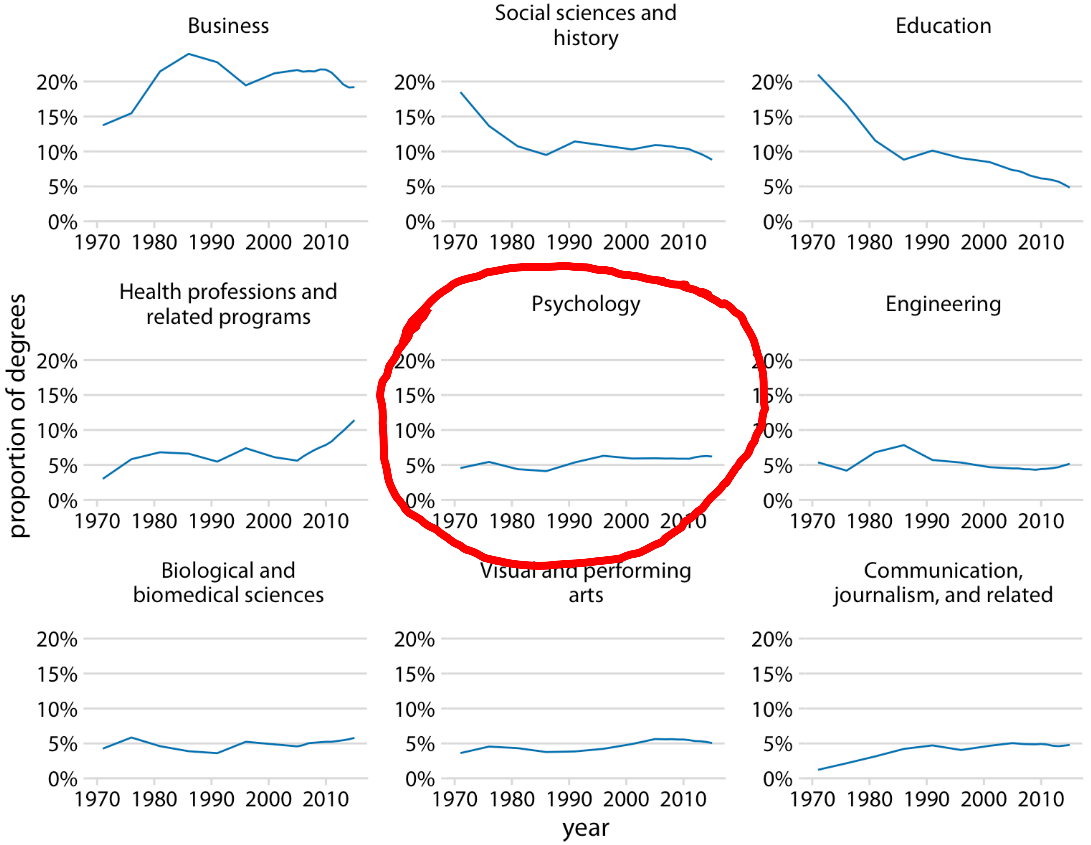


Image credit: https://clauswilke.com/dataviz/

With mixed multi-panel plots, be consistent in some aesthetic attributes

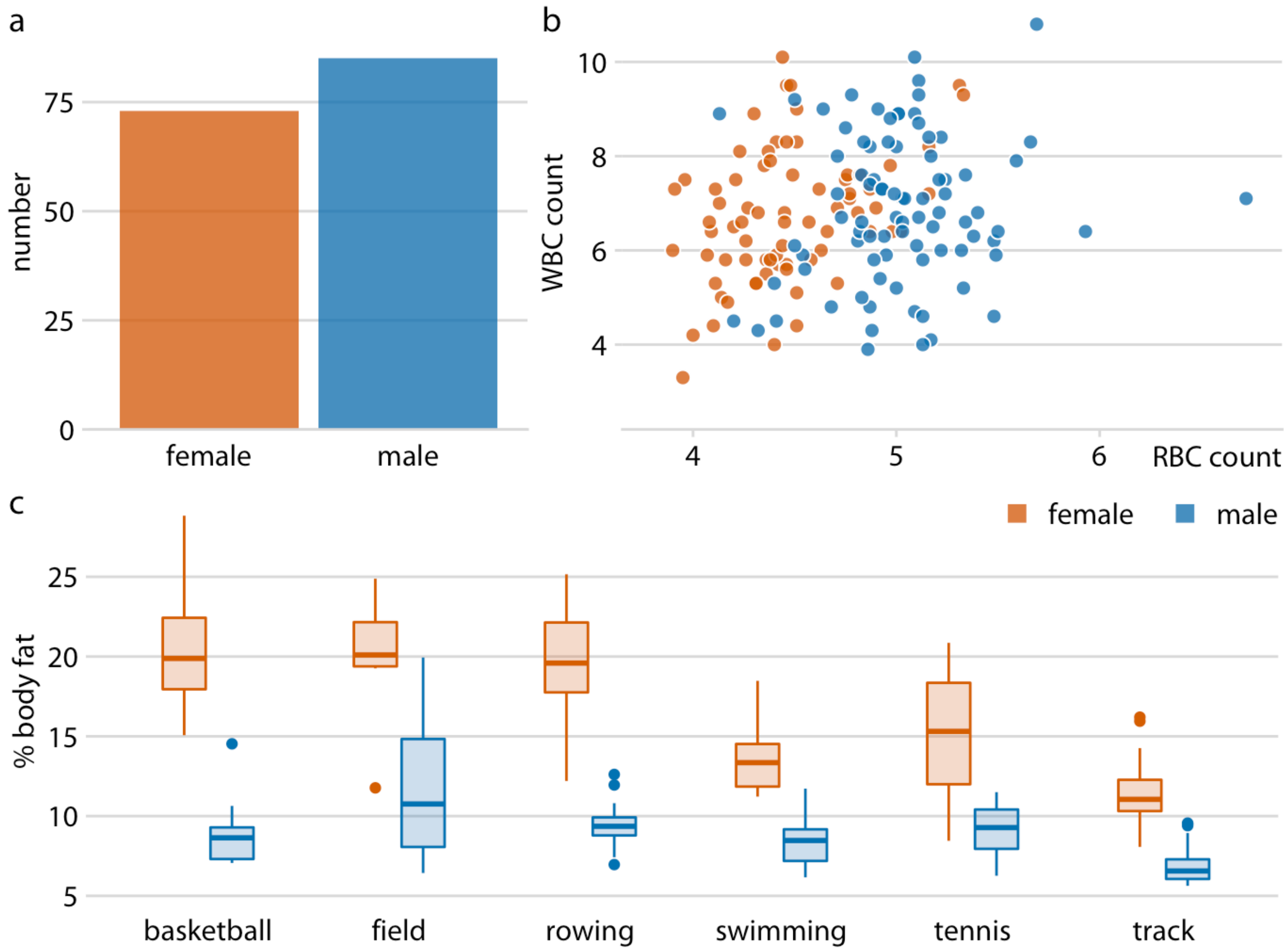
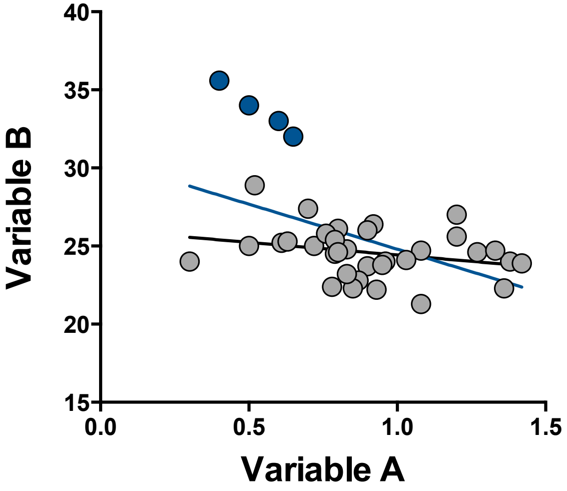


Image credit: https://clauswilke.com/dataviz/

Including/excluding data

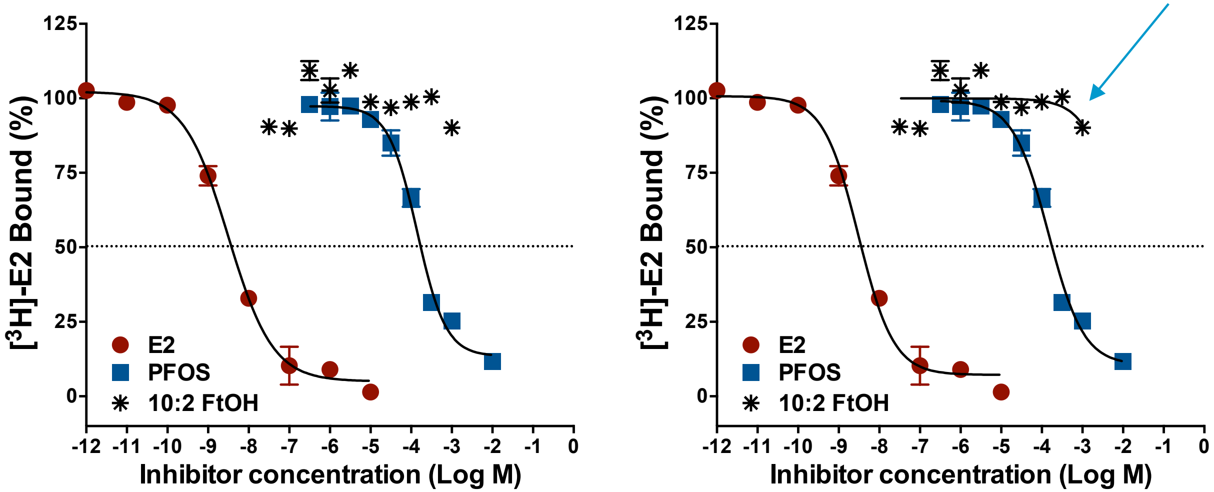
Do not arbitrarily delete data points without scientific justification

Don’t “massage” line fits or change parameters post-hoc to best fit your data



Consider how the four points colored blue influence the regression fit.

Are these outliers? Do they have an outsized impact on the apparent trend?

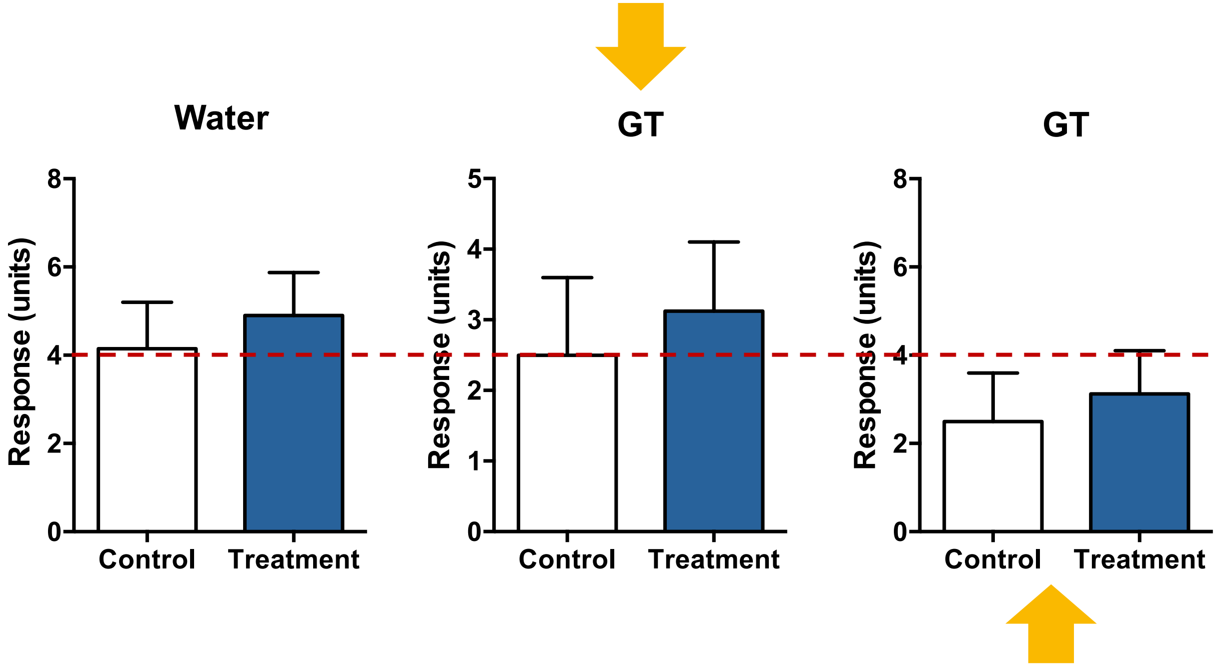


A switch from 4-parameter to 3-parameter curve allows for fit of 3rd data set, but is this post-hoc change in analysis appropriate or meaningful?

Presenting multiple figures

Jointly presented data should be on the same scale (most of the time)

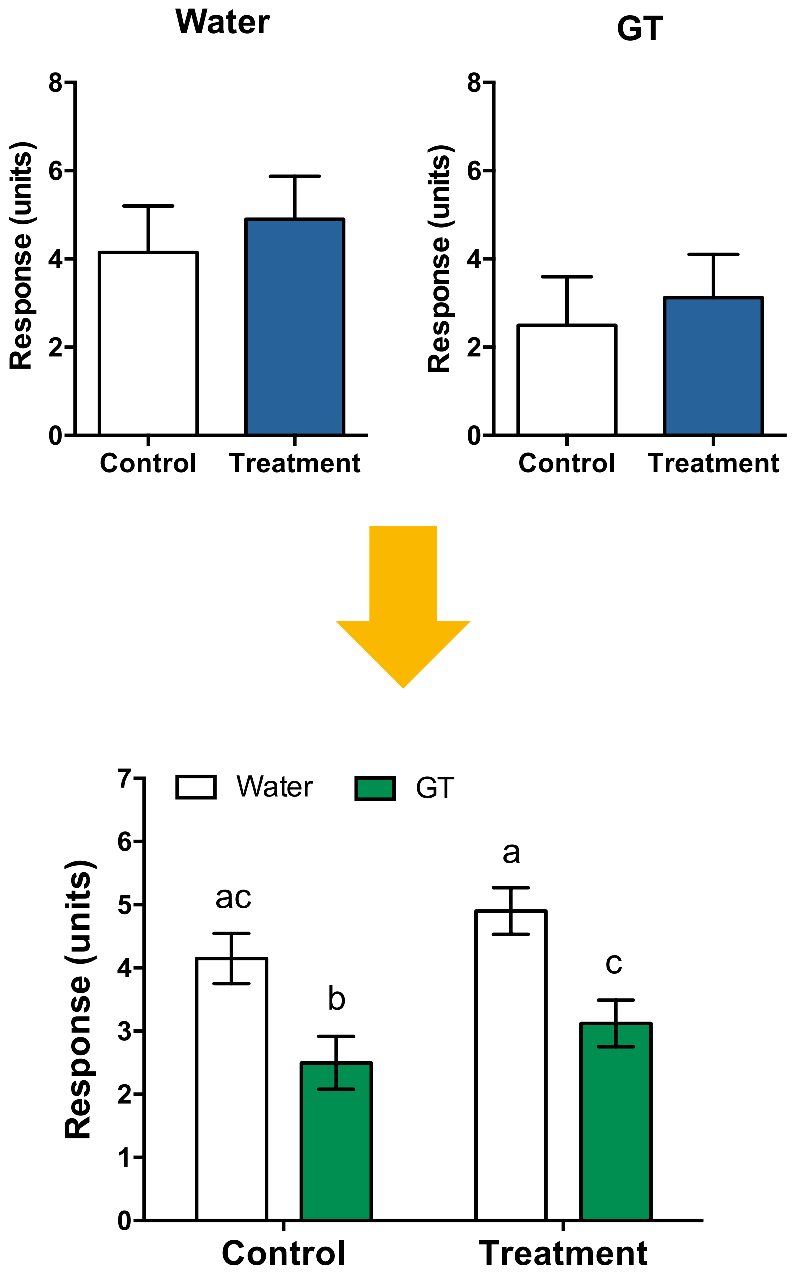
The center panel uses a different scale for the same type of data as in the left panel. Clearly, an comparison between "Water" and "GT" is intended. But the different scales obfuscate the comparison.



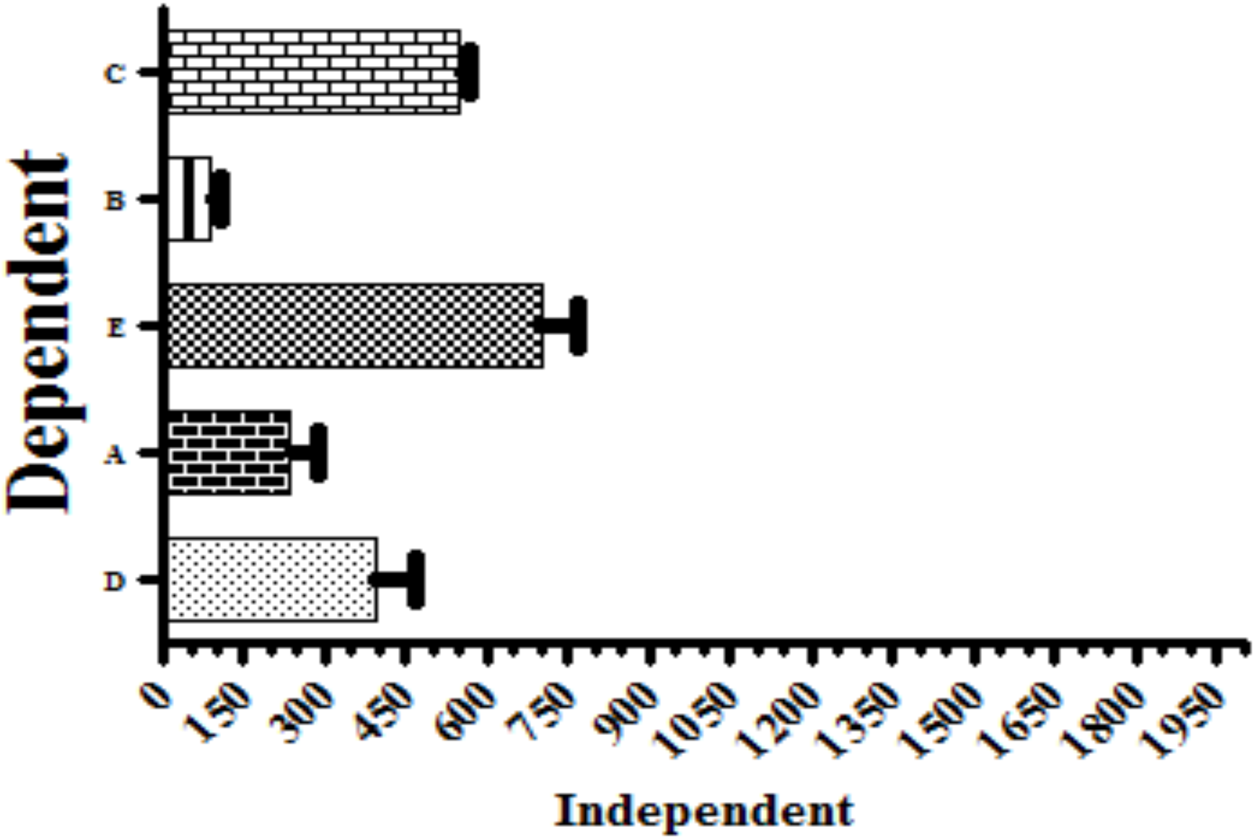
The right panel corrects this problem. Now you can see more easily that the response was overall a bit lower for both control and treatment in the GT group compared to the Water group. The dashed red line added helps you visualize the impact of the selected scale.

Graph style should reflect experiment design

Data groupings should infer what type of analysis was performed.



A poorly formatted graph

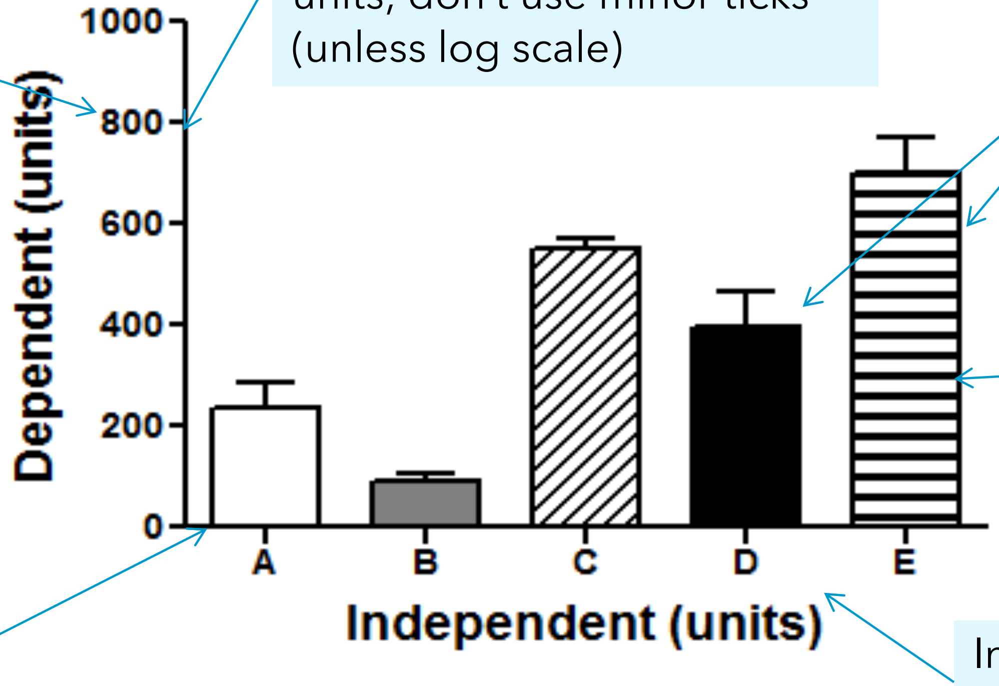


Use large font for axis titles, two sizes smaller for axis labels; be consistent in font sizes

Scale is more appropriate; subdivide Y-axis into big units; don't use minor ticks (unless log scale)

Use solid fills and big patterns to distinguish bars

Dependent variable (measurement) on the Y-axis



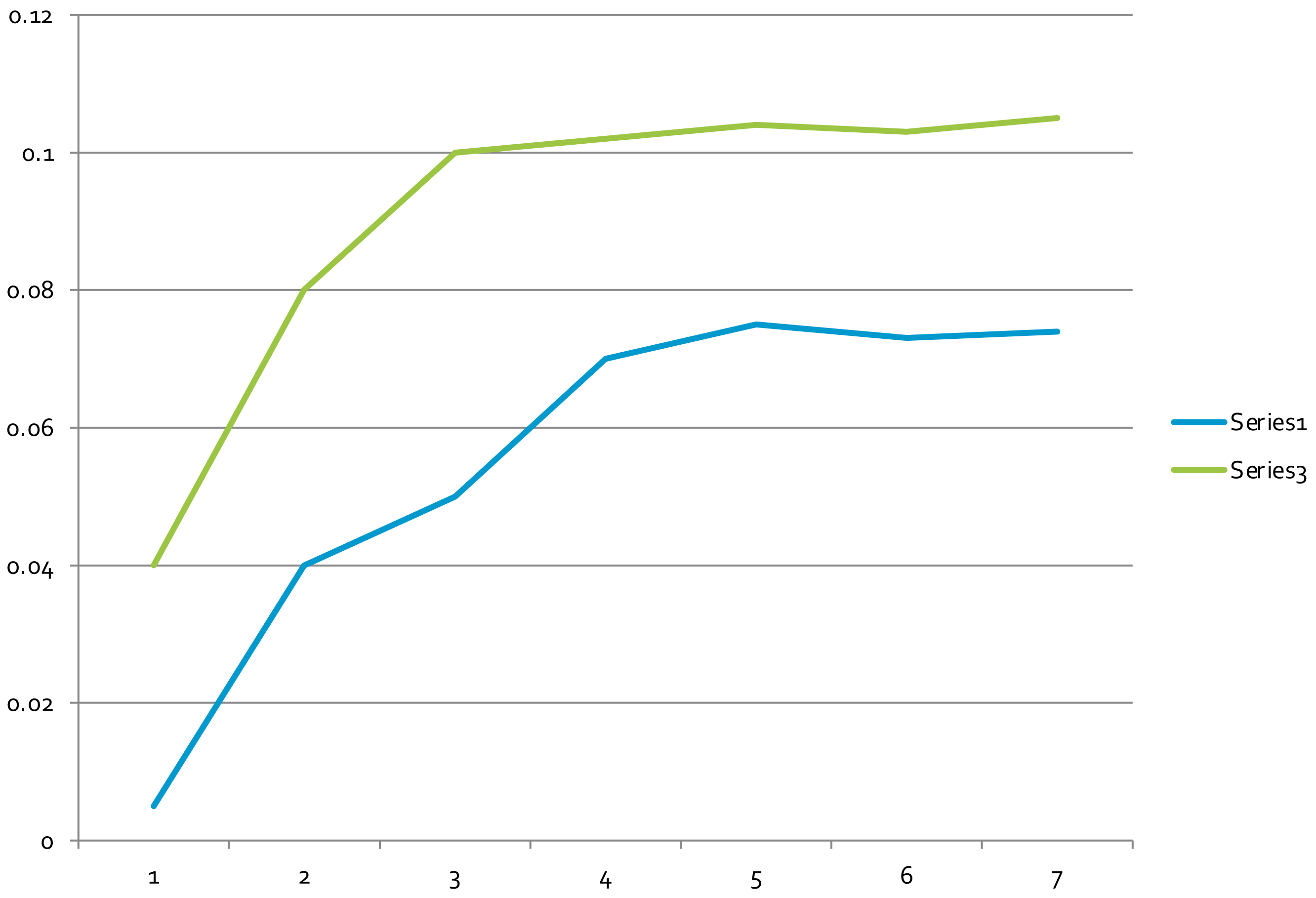
Use the same thickness for axes, bar outlines and error bars – thick enough to be seen but not overly thick

Order your categories so that they are easily interpreted, tell the story; use white for control or reference group which should be placed at the left-most position

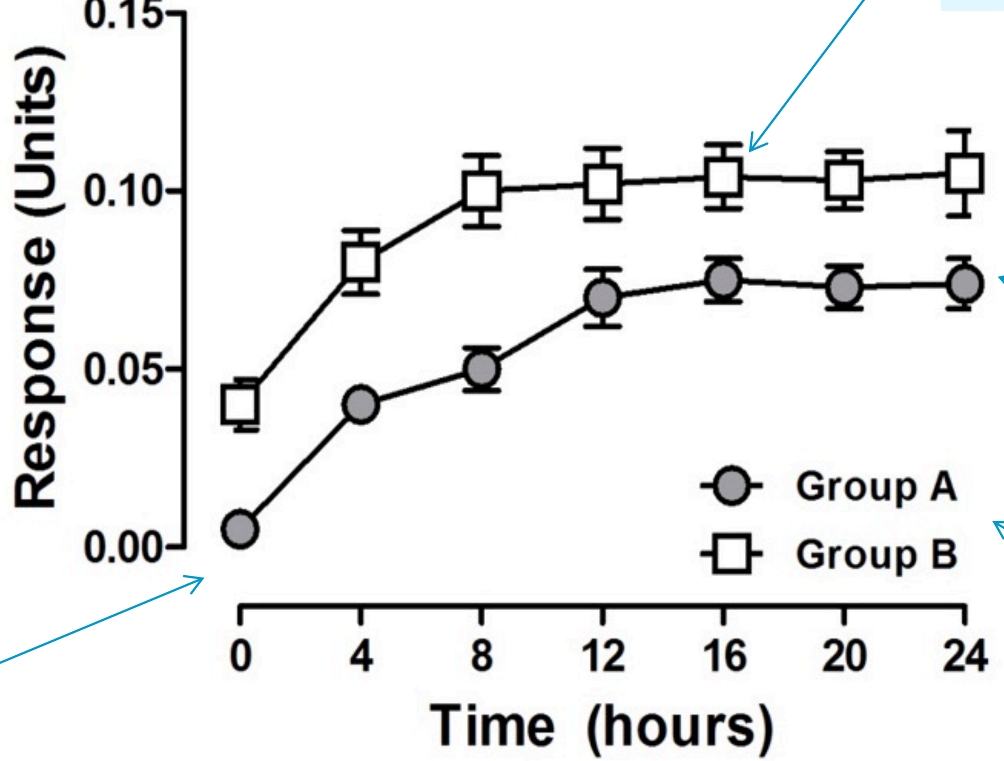
Independent variable (treatments) on the X-axis

Keep the figure looking clean and easily readable

Poorly formatted line graph



Follow previous recommendations for font sizes, axes thickness, etc. Also use the same thickness for symbol outline.



Keep the field behind the data clear of any grid lines (unless needed to show reference measurement, such as for normalized data)

Use big symbols, easy to see when graph is reduced for publication. Use shading to help distinguish symbols.

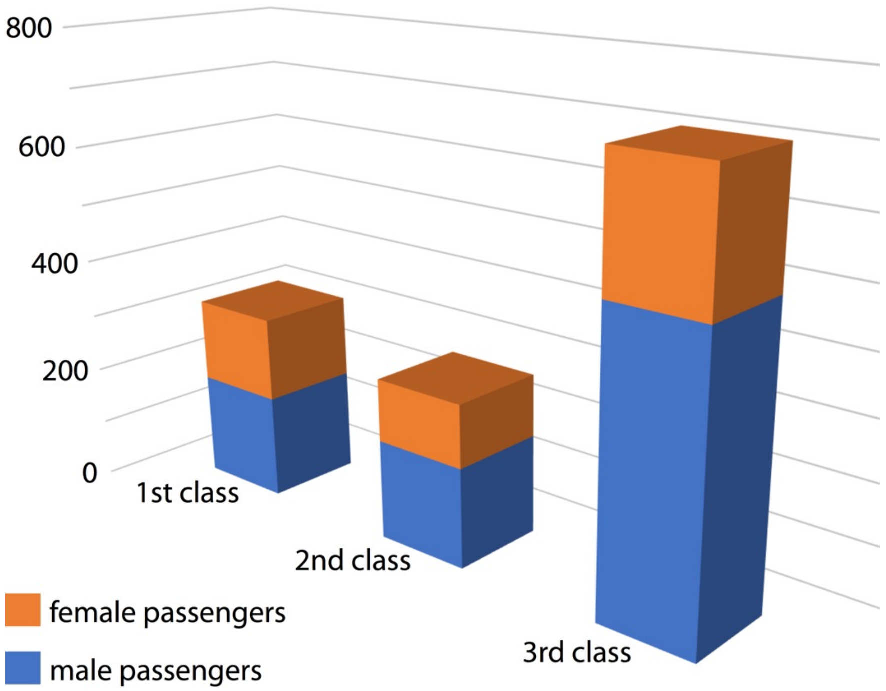
If the first data point falls on top of the Y-axis, then offset the axes for clarity.

Include a legend to identify symbols

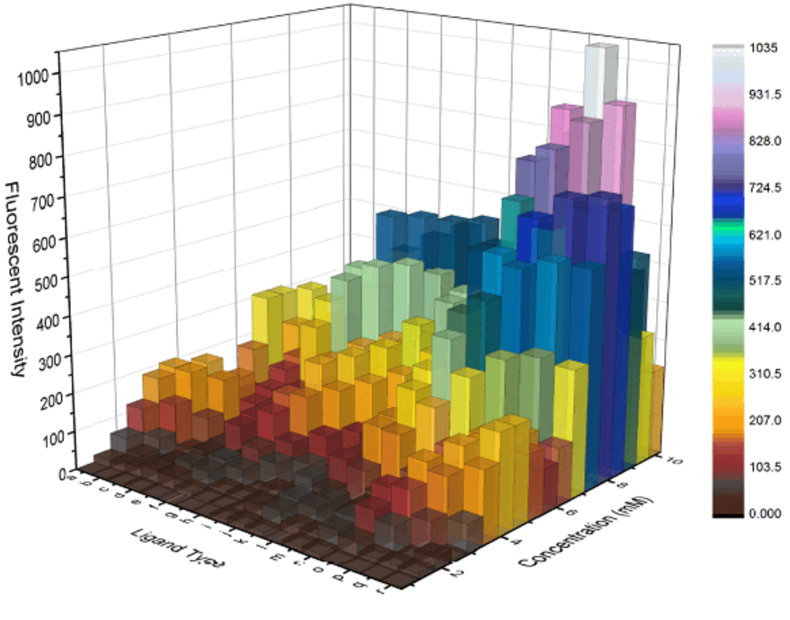
Avoid the 3D temptation

Do not use 3D unless absolutely necessary.

Many other ways to show 3 levels of data without using a rotated 3D graph



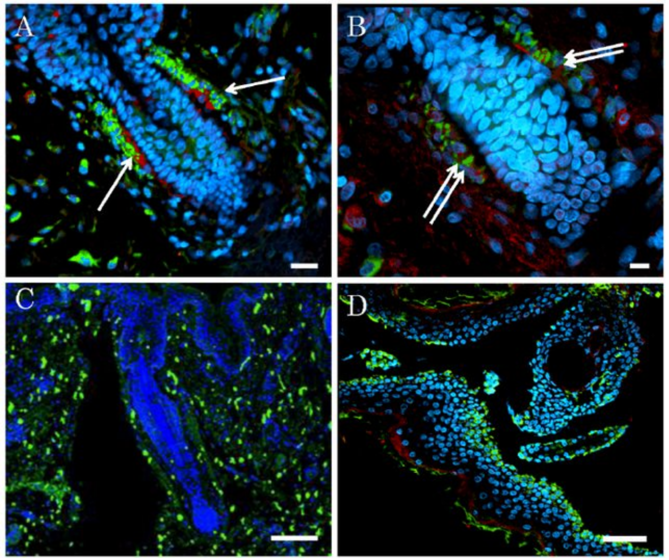
clauswilke.com



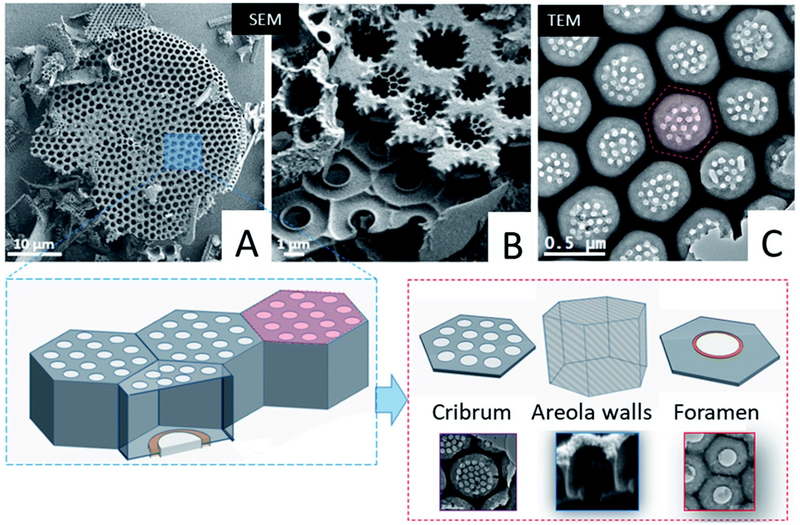
www.originlab.com

Pictures as figures

- Do not assume that anyone knows what is in a picture
  - Use arrows, markers to identify features
  - Specify magnification or Include a scale bar and define in the legend
  - Specify meanings of colors in figure or in legend
  - Include key explanations in figure legend or footnote

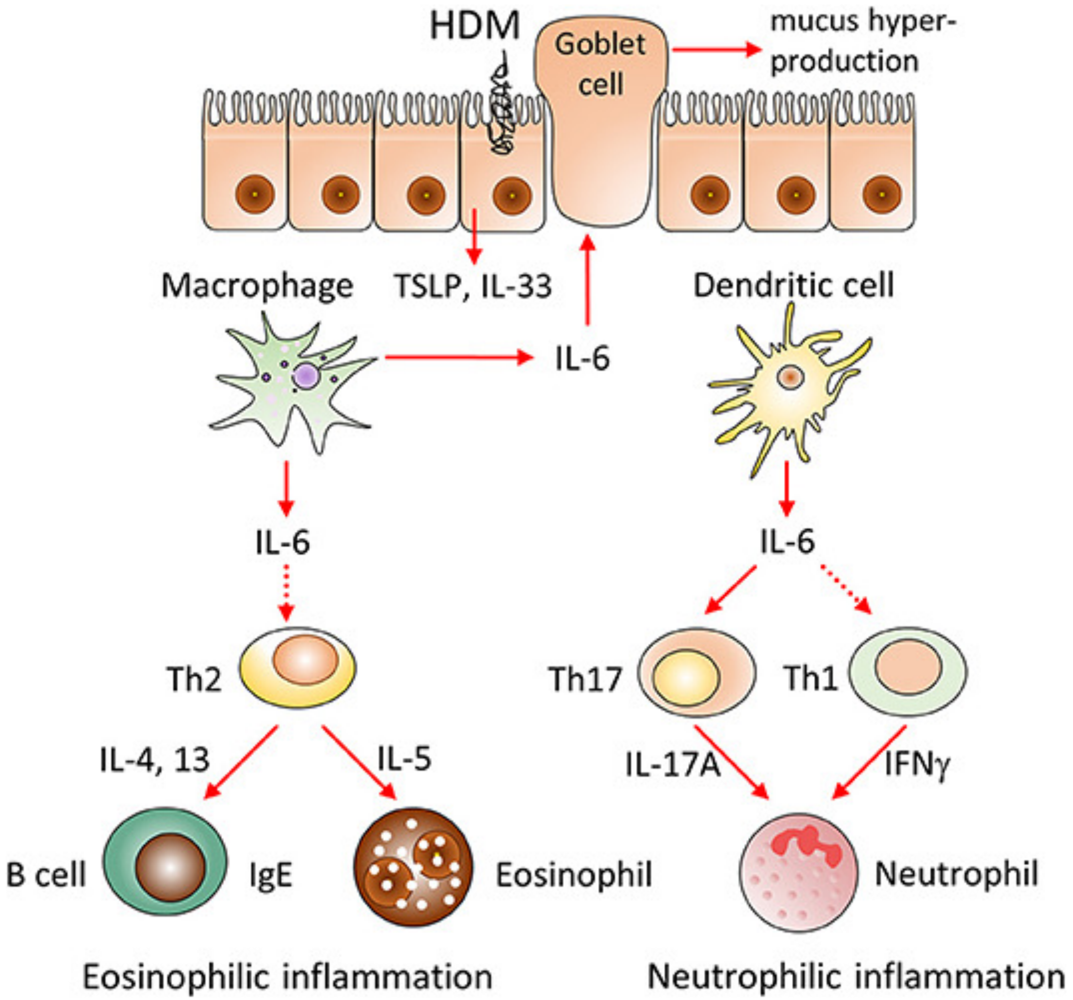


Kanno et al 2013 Intl J Mol Sci



Darouich, et al 2022. Nanoscale Adv

Use diagrams to convey complex ideas, but keep them simple!



You can also use animation to build up complex diagrams

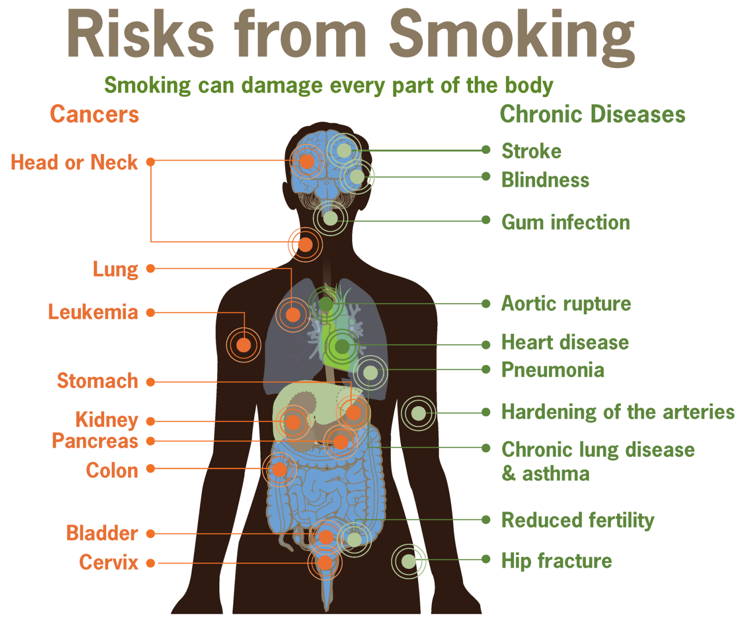
Image credit: Gubernatorova, et al (2018) Front Immunol

What is an infographic?

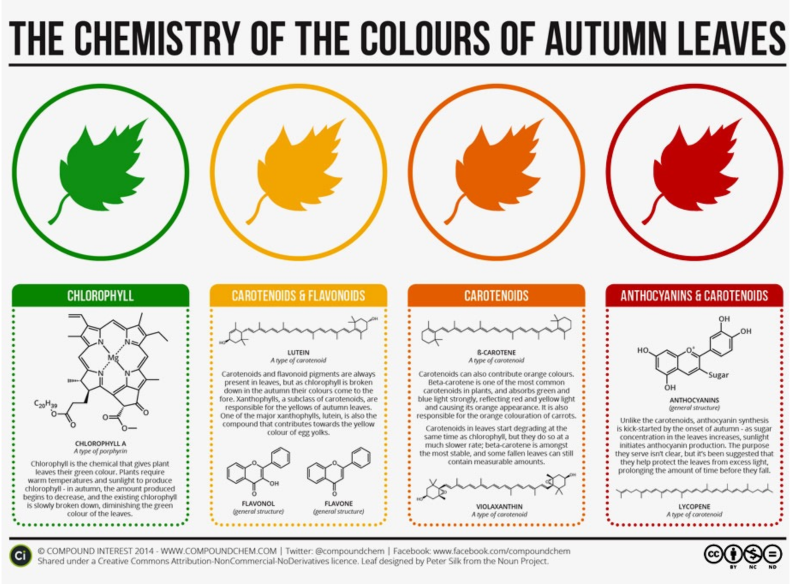
A visual image used to represent information or data

Designed to be accessible for non-experts

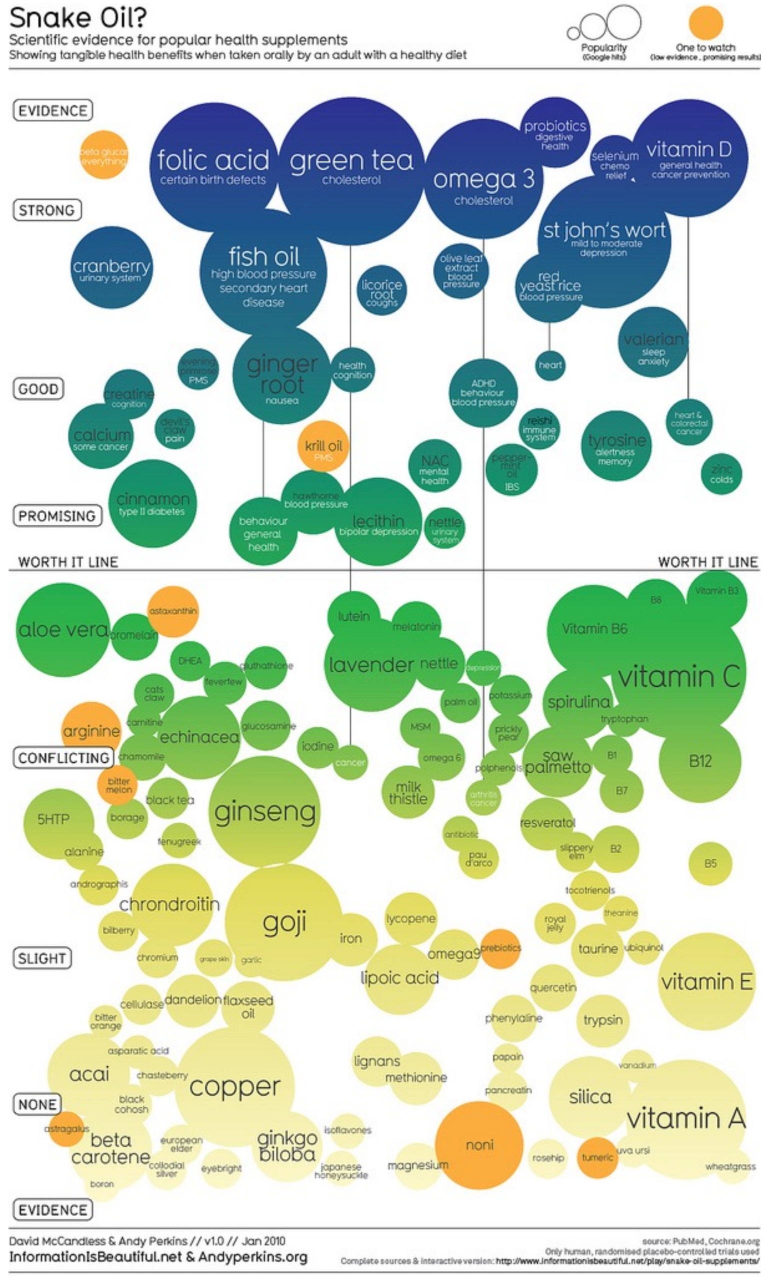
Keep it simple for an oral presentation



public domain


THE CHEMISTRY OF THE COLOURS OF AUTUMN LEAVES
CHLOROPHYLL
CAROTENOIDS & FLAVONOIDS
CAROTENOIDS
ANTHOCYANINS & CAROTENOIDS
CHLOROPHYLL A
LUTEIN
ß-CAROTENE
ANTHOCYANINS
FLAVONOL
FLAVONE
VIOLAXANTHIN
LYCOPENE
© COMPOUND INTEREST 2014 - WWW.COMPOUNDCHEM.COM | Twitter: @compoundchem | Facebook: www.facebook.com/compoundchem
Shared under a Creative Commons Attribution-NonCommercial-NoDerivatives licence. Leaf designed by Peter Silk from the Noun Project.

CC BY-NC-ND 4.0



InformationIsBeautiful.net

When to use infographic style?

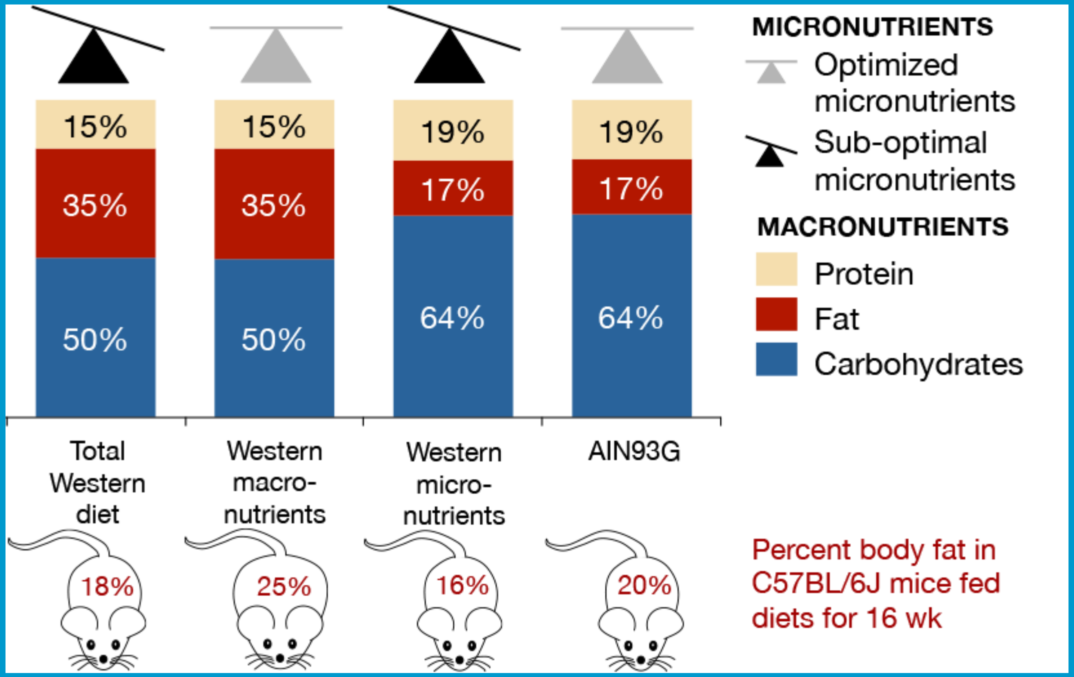
Excellent for oral presentations, poster presentations

Coordinating complex ideas presented in introduction or conclusion

Excellent for public presentation

Infographics excel at **distilling complex data into a simple visual format**

Not generally appropriate as main figures in journal articles, but could be very effective as graphical abstract



Benninghoff et al 2016 Nutr Res

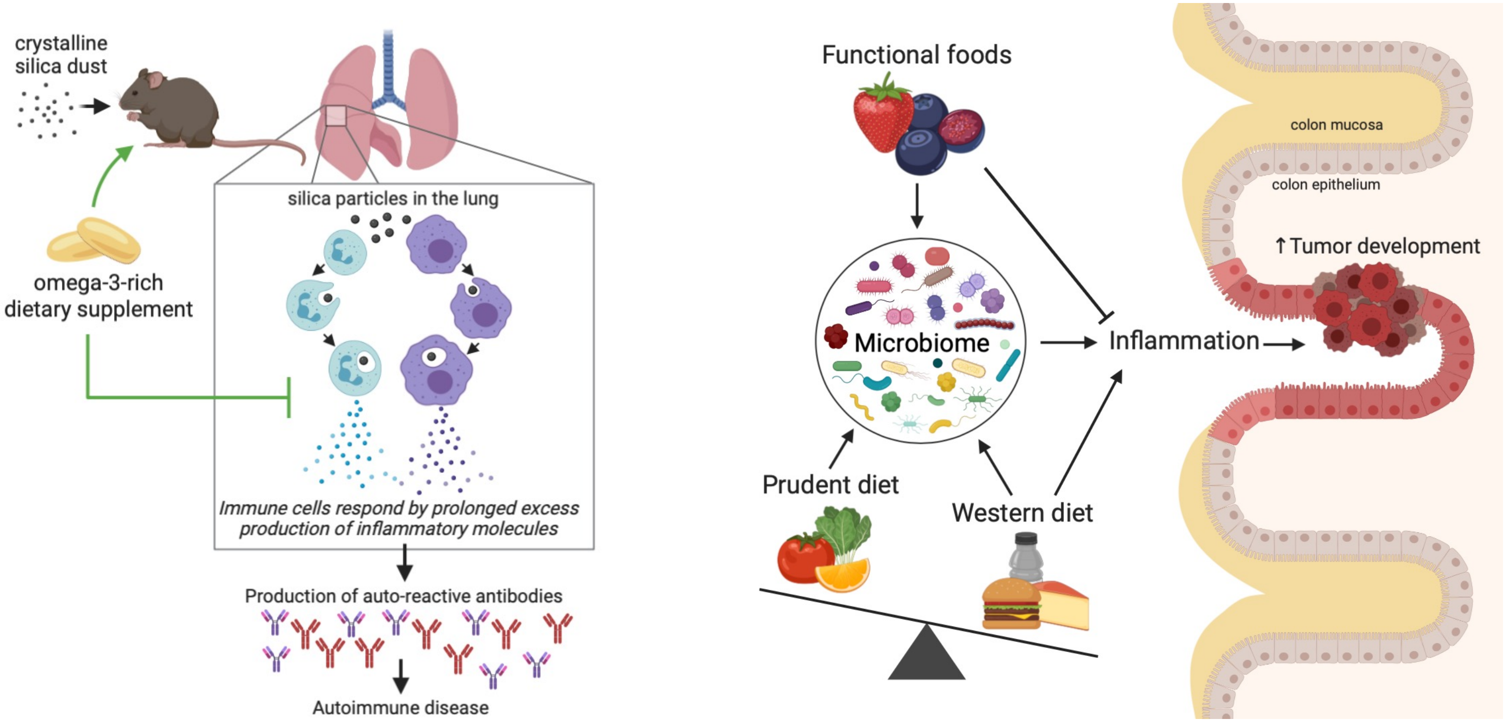
Graphical abstracts

avoid lots of text, complicated plots

focus on “take-home” message

make sure readable a published size

- Scribendi – Top 10 tips for designing graphical abstracts (https://tinyurl.com/4z2dk6kp)
- BioRender – Top 5 design Tips for winning graphical abstract YouTube video https://www.youtube.com/watch?v=35x2nPMzWbE


crystalline silica dust
omega-3-rich dietary supplement
silica particles in the lung
Immune cells respond by prolonged excess production of inflammatory molecules
Production of auto-reactive antibodies
Autoimmune disease
Functional foods
Microbiome
Inflammation
Prudent diet
Western diet
colon mucosa
colon epithelium
↑Tumor development

Graphical abstracts created using bioRender.com

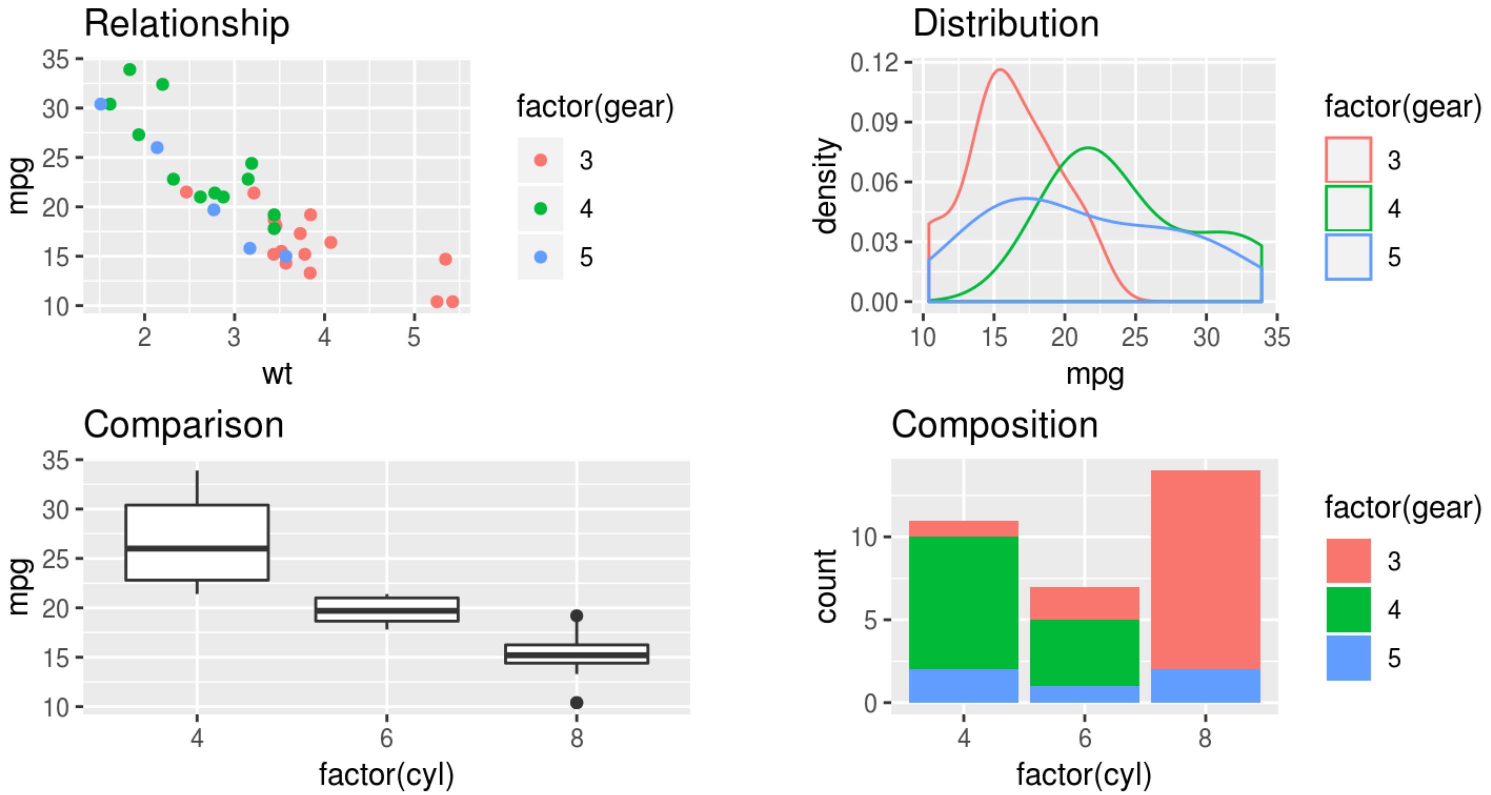
Software options

- Excel is just <u>awful</u> for making science graphs
  - Especially bad for multi-panel figures
  - Thinks too much for you
- Try out other software that specializes in scientific data presentation
  - SigmaPlot http://www.sigmaplot.com/
    - Haven’t worked with this one in years
    - Advanced graphics with somewhat steep learning curve
  - **GraphPad Prism** http://www.graphpad.com/
    - Easy to use interface
    - Integrates statistics with graphing
    - Drawback – pricey for individuals (~$100/yr); bulk licensing available
    - 30-day trial available

GraphPad
Software

A few words about R plots

ggplot2



- Highly versatile, free software for data visualization and analysis
- Steep learning curve
- Default settings do need some tweaking to make suitable for presentations, publications

Graphics reveal data, communicate complex ideas and dependencies with clarity, precision and efficiency

-Edward R. Tufte
“The Visual Display of Quantitative Information”

The BEST graph is one which:
“gives to the VIEWER
the greatest number of IDEAS
in the shortest TIME
with the least INK
in the smallest SPACE.”