

Principles of effective data visualization

California Ecology and Conservation

Summer 2024

**Design for the right audience,
accurately represent the data, and
keep it clear.**

Yan Holtz

An's personal data visualization heroes!



Meghan Harris



Ijeamaka Anyene



Allison Horst



Danielle Navarro



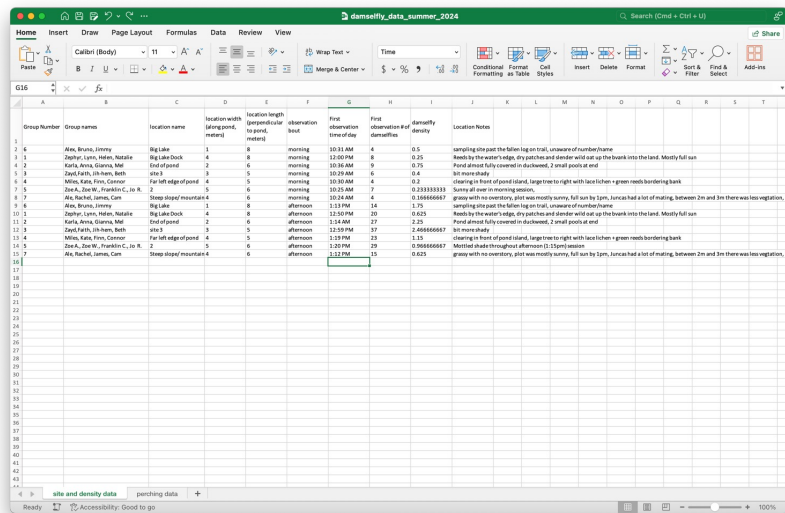
Nicola Rennie



Sam Csik

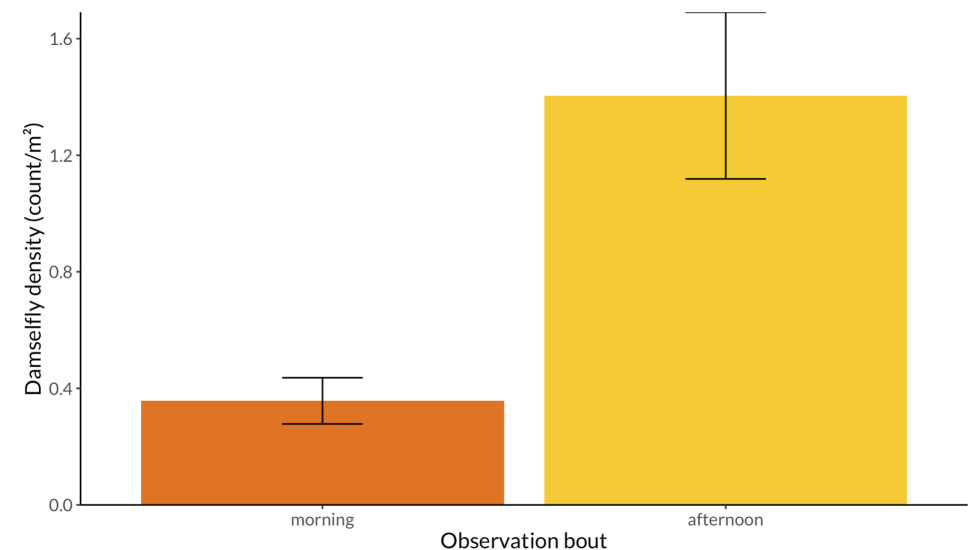
Why does data visualization matter?

It's hard to get people to care about this:



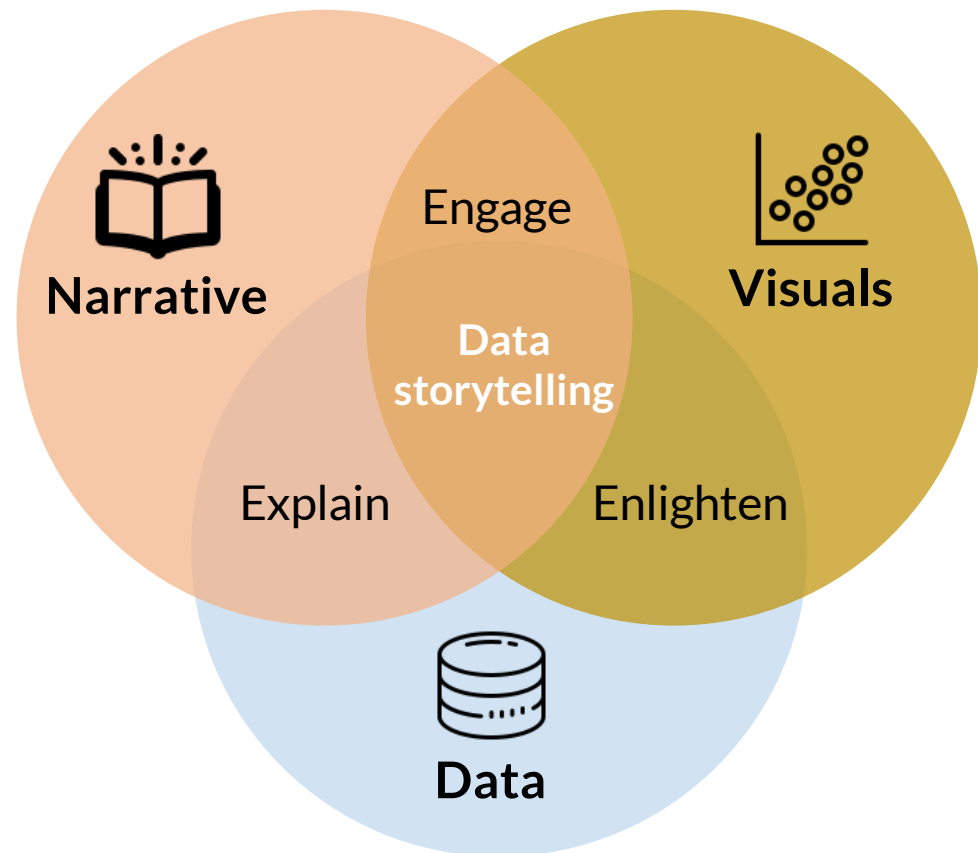
| Group Number | Group Name | Location Name | Location Width (along pond, meters) | Location Length (perpendicular to pond, meters) | Observation Hour | First Observation Time of Day | First Observation # | Damselfly Density | Location Notes |
|--------------|--------------------------------|-----------------------|-------------------------------------|---|------------------|-------------------------------|---------------------|--------------------|---|
| 1 | Alan, Brown, Jimmy | Big Lake | 1 | 8 | morning | 10:20 AM | 4 | 0.5 | Sampling the pond the following day on trail, unaware of number change |
| 2 | Zephur, Lynn, Melissa, Natalie | Big Lake Dock | 4 | 8 | morning | 10:20 AM | 8 | 0.25 | Reached by the water's edge, 40 patches and divider will sit up the bank into the land. Mostly full sun |
| 3 | Karla, Anna, Gloria, Beth | End of pond | 2 | 6 | morning | 10:20 AM | 9 | 0.75 | Pond almost fully covered in duckweed, 2 small ponds at end |
| 4 | Zach, Kelly, Will, Alex, Beth | side | 3 | 5 | morning | 10:20 AM | 6 | 0.4 | Got more closely |
| 5 | Mike, Kate, Finn, Connor | Far left edge of pond | 4 | 5 | morning | 10:20 AM | 4 | 0.2 | Observing in front of pond island, large tree to right with largeichen - green moss bordering bank |
| 6 | Alan, Brown, Jimmy | Big Lake | 1 | 8 | afternoon | 1:13 PM | 14 | 1.75 | Sampling the pond the following day on trail, unaware of number change |
| 7 | Zephur, Lynn, Melissa, Natalie | Big Lake Dock | 4 | 8 | afternoon | 1:13 PM | 20 | 0.625 | Reached by the water's edge, 40 patches and divider will sit up the bank into the land. Mostly full sun |
| 8 | Karla, Anna, Gloria, Beth | End of pond | 2 | 6 | afternoon | 1:14 PM | 27 | 2.25 | Pond almost fully covered in duckweed, 2 small ponds at end |
| 9 | Zach, Kelly, Will, Alex, Beth | side | 3 | 5 | afternoon | 1:15 PM | 17 | 2.0000000000000002 | Got more closely |
| 10 | Mike, Kate, Finn, Connor | Far left edge of pond | 4 | 5 | afternoon | 1:15 PM | 23 | 1.53 | Observing in front of pond island, large tree to right with largeichen - green moss bordering bank |
| 11 | Alan, Brown, Jimmy | Big Lake | 1 | 8 | afternoon | 1:20 PM | 29 | 0.9000000000000001 | Sampling the pond the following day on trail, unaware of number change |
| 12 | Zephur, Lynn, Melissa, Natalie | Big Lake Dock | 4 | 8 | afternoon | 1:23 PM | 15 | 0.625 | Reached by the water's edge, 40 patches and divider will sit up the bank into the land. Mostly full sun |

But they could care about this:



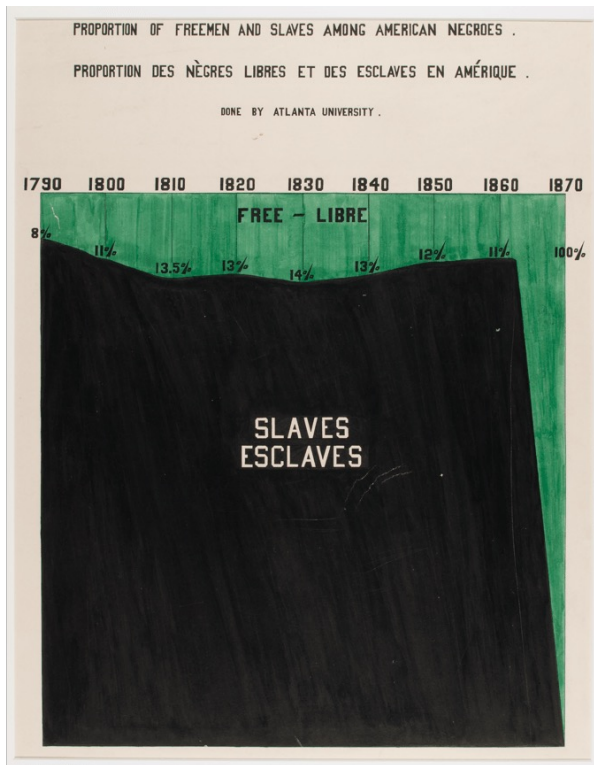
Storytelling is a crucial research skill!

Numbers have an important story to tell. They rely on you to give them a clear and convincing voice.

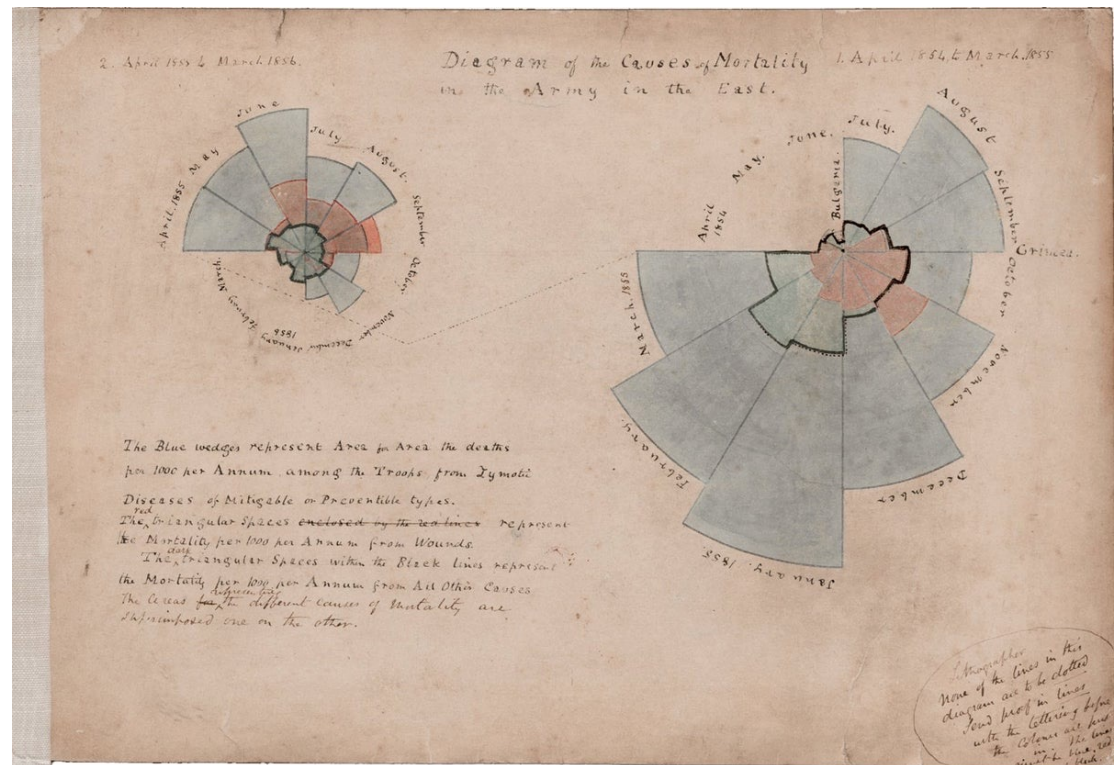


[Brent Dykes, *Forbes*, "Data Storytelling: The Essential Data Science Skill Everyone Needs"](#)

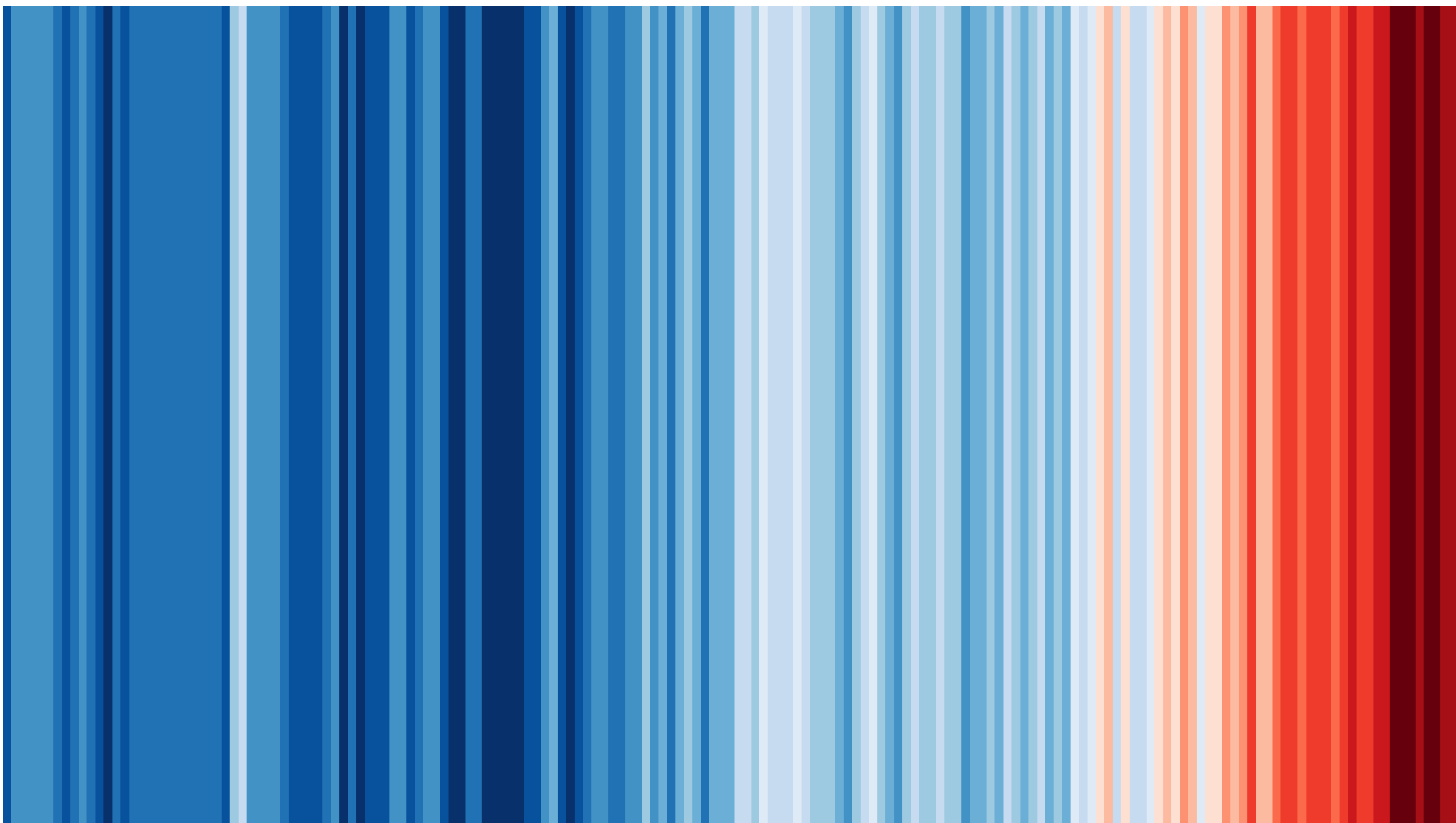
Data visualization has a rich history



WEB DuBois

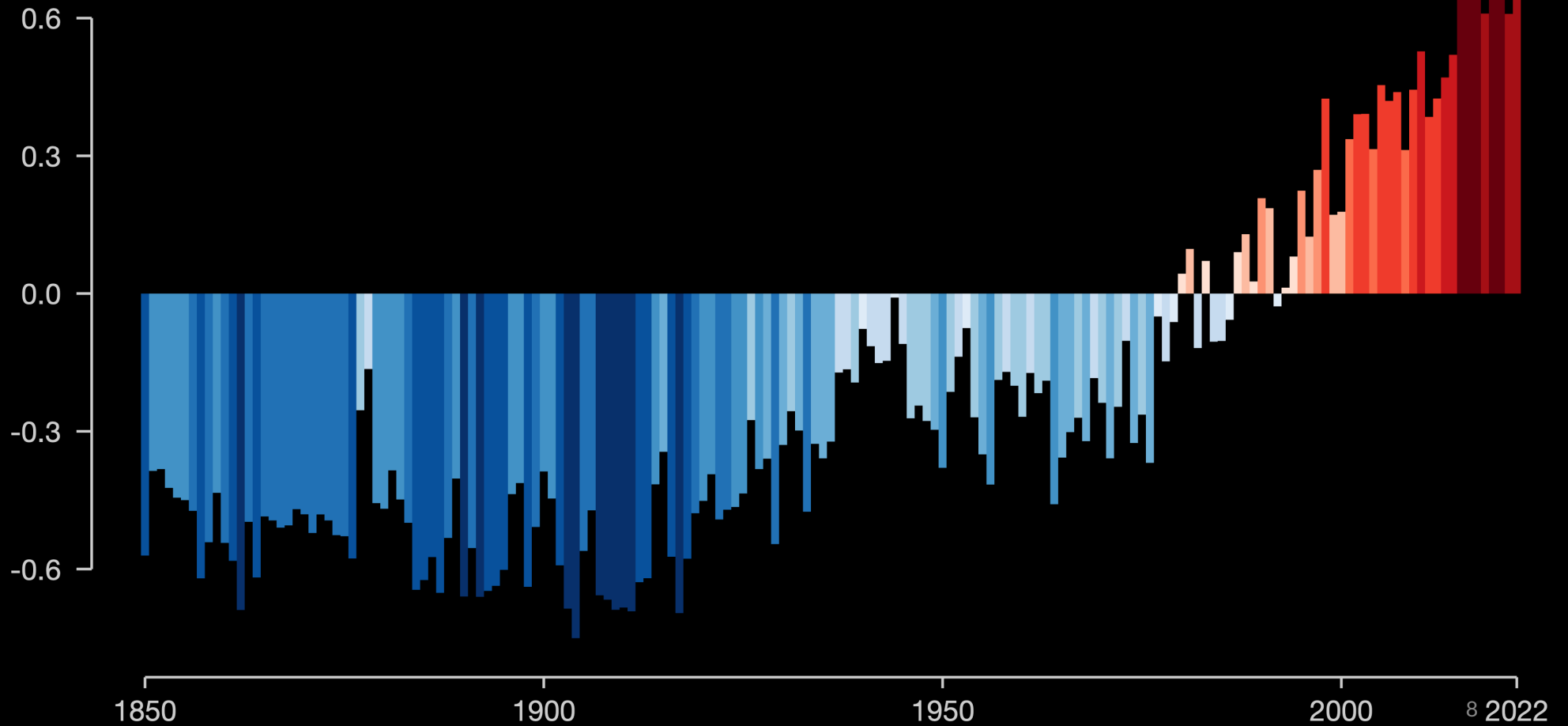


Florence Nightingale



Global temperature change

Relative to average of 1971-2000 [°C]



When making a graph, ask yourself these questions (in order of importance):

1. Are the data I'm showing correct?
2. Am I responsibly communicating the story?
3. Is it clear for the audience?
4. Does it look awesome?

When making a graph, ask yourself these questions (in order of importance):

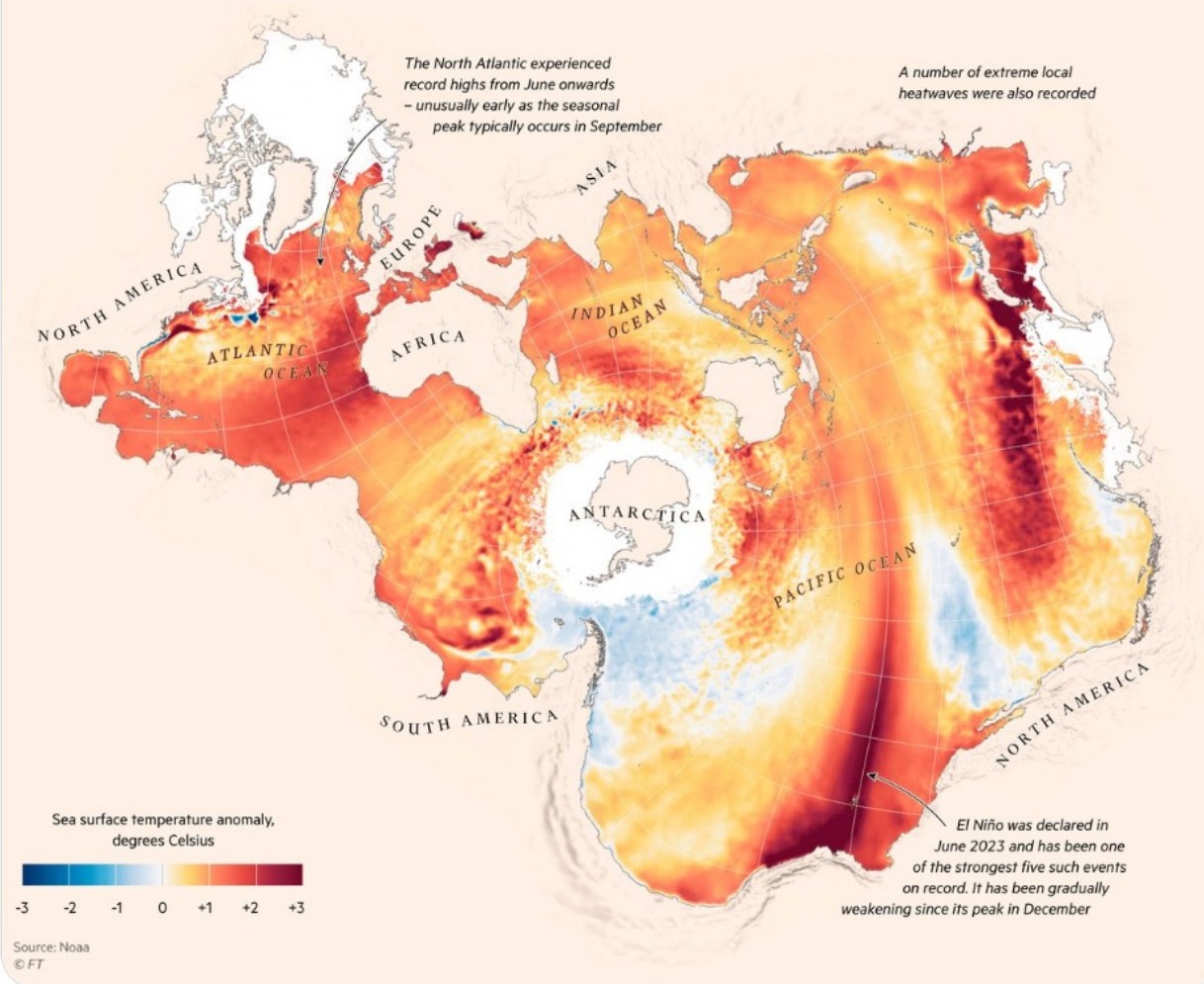
1. Are the data I'm showing correct?

Not part of the scope for today, but some tips:

- double check data collection (in the field) and data entry (in Excel)
- investigate outliers to make sure they're not typos, etc.

Exceptional ocean heat across the globe

Sea surface temperatures for March 2023–February 2024, compared with long-term average



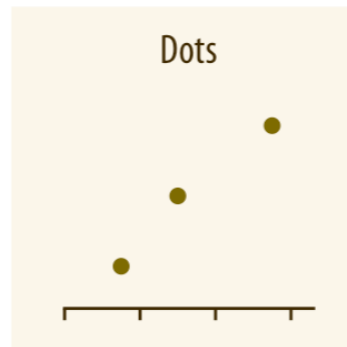
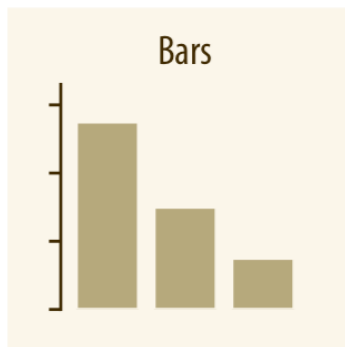
When making a graph, ask yourself these questions (in order of importance):

1. Are the data I'm showing correct?
2. Am I responsibly communicating the story?

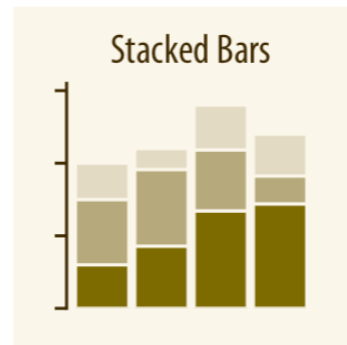
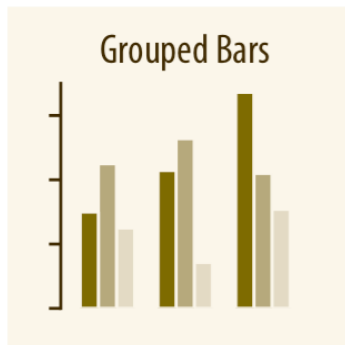
Am I responsibly communicating the story?

- Ask yourself: does my graph actually show what I want it to show?
- Solution: choose the right graph for the right variables

Visualizing amounts: how do groups differ in counts or measure?



y-axis: count or measure
x-axis: groups



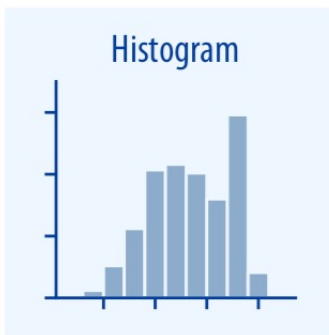
Example applications:

How does average plant height differ between shaded and non-shaded areas?

How does scrub jay count differ between restored and unrestored areas?

Visualizing distributions

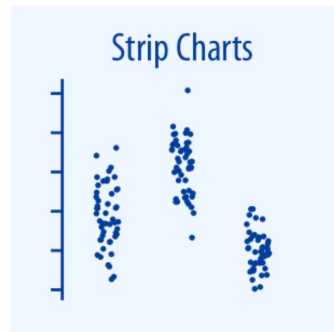
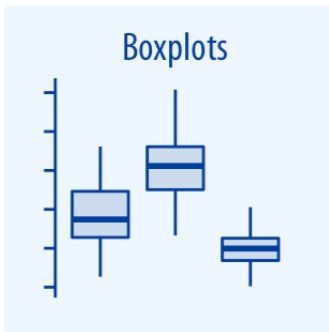
What is the distribution of a given variable?



y-axis: frequency
x-axis: variable of interest

Example application:
What is the distribution of damselfly count?

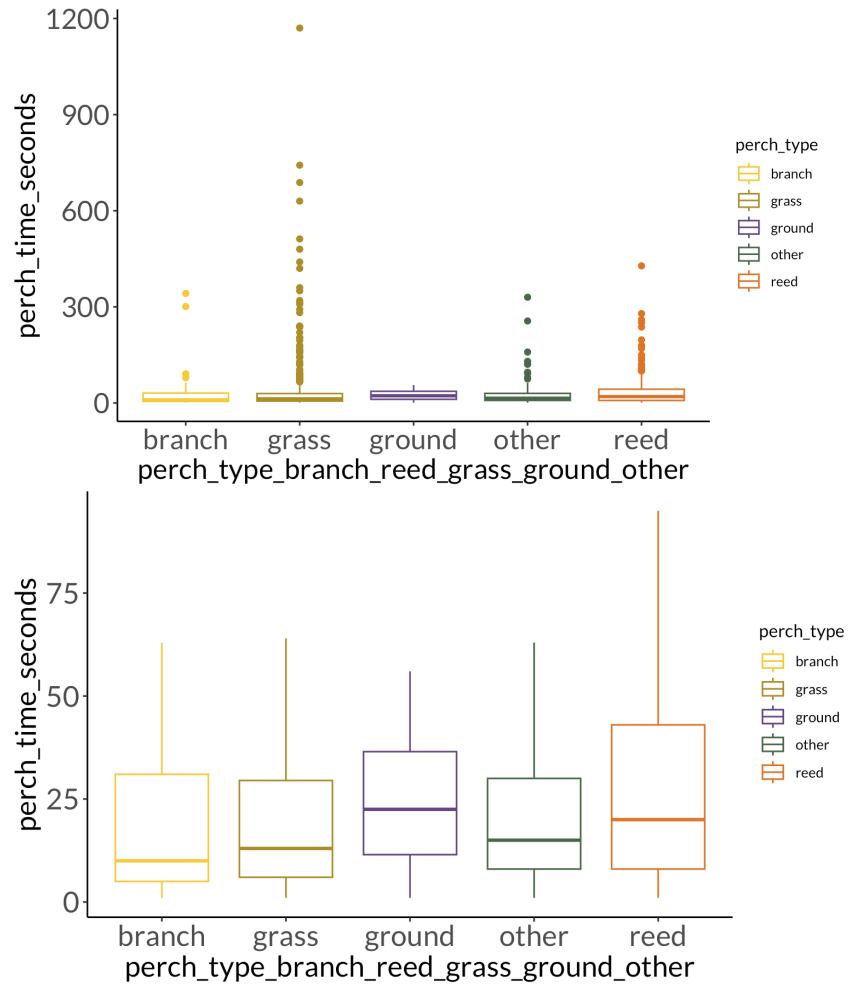
How do these groups differ in their distribution?



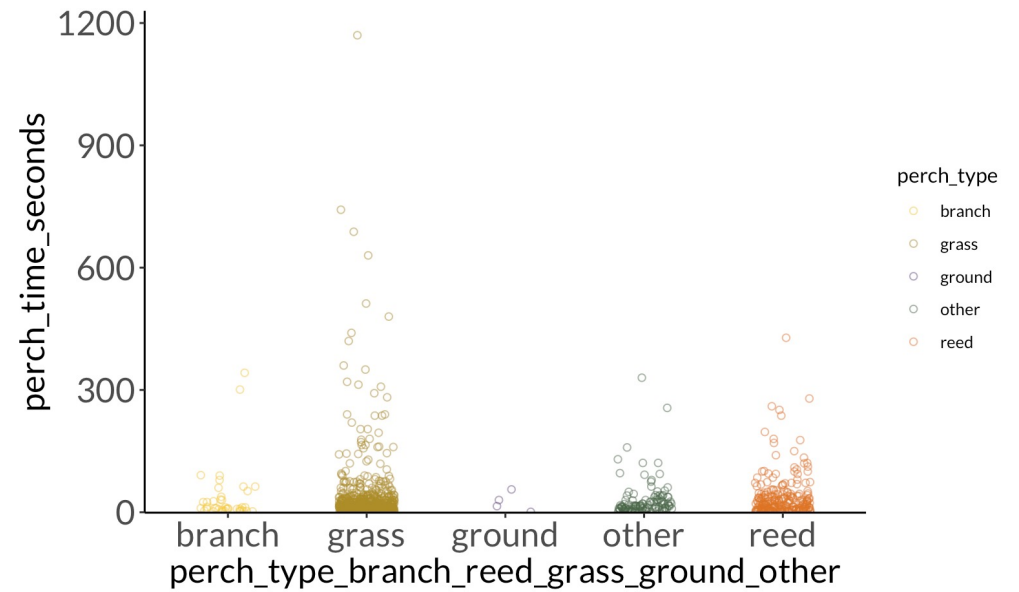
y-axis: count or measure
x-axis: groups

Example application:
What is the distribution of damselfly perching time between perch types?

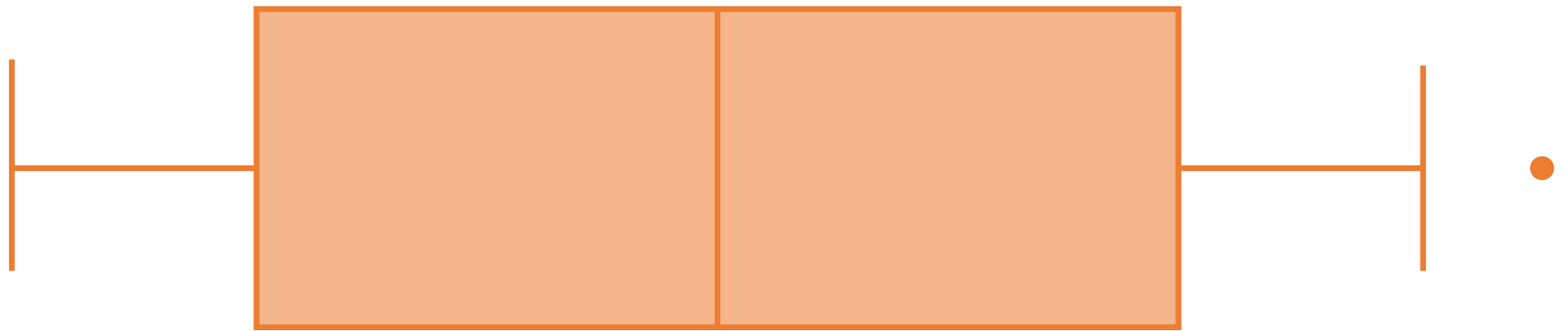
Boxplots



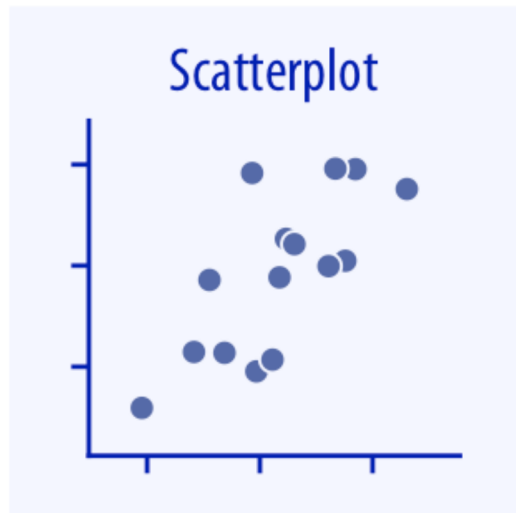
Strip chart



What's in a box-and-whisker?



Visualizing relationships: what is the relationship between two continuous or discrete variables?



Fundamentals of Data Visualization, Claus O. Wilke

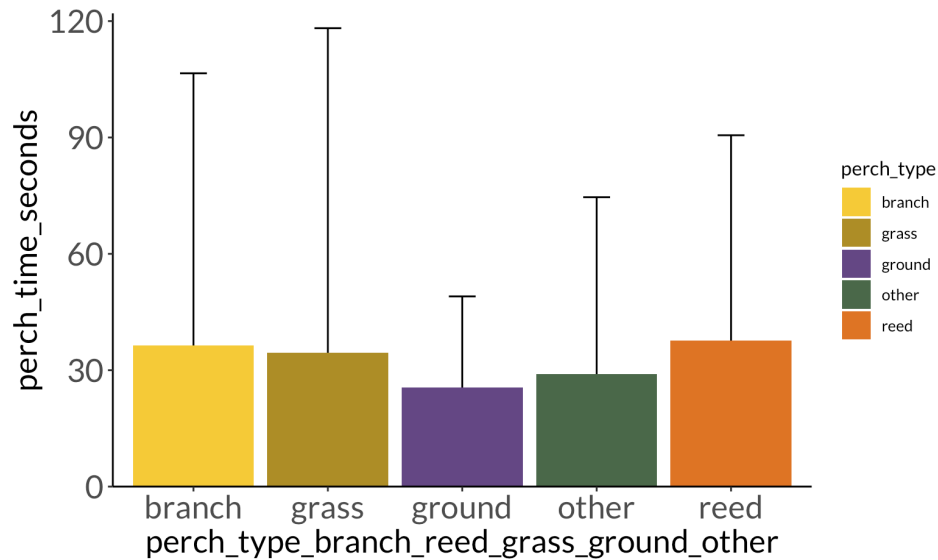
y-axis: response variable
x-axis: predictor variable

Example application:
What is the relationship between distance to
water and damselfly count?

Visualizing spread or uncertainty

Standard deviation

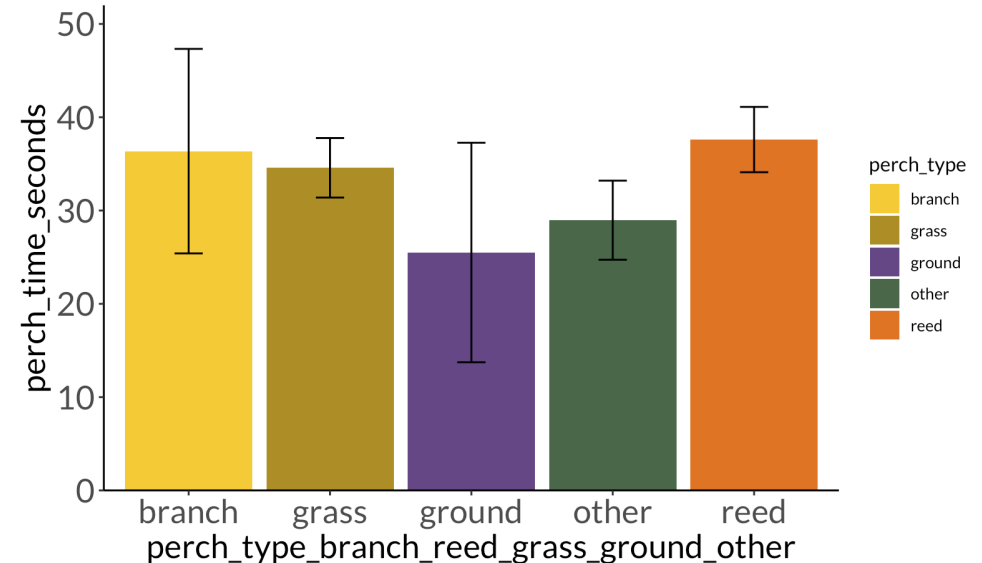
How spread out from the mean is your variable?



$$s = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n - 1}}$$

Standard error

How precise is your sample? How well does your sample capture the population it represents?



$$SE_{\bar{y}} = \frac{s}{\sqrt{n}}$$

What kind of figure would you make?

You want to determine how plant biomass (measured in g) between soil nitrogen (measured as high, medium, and low) treatment plots.

Write a hypothesis.

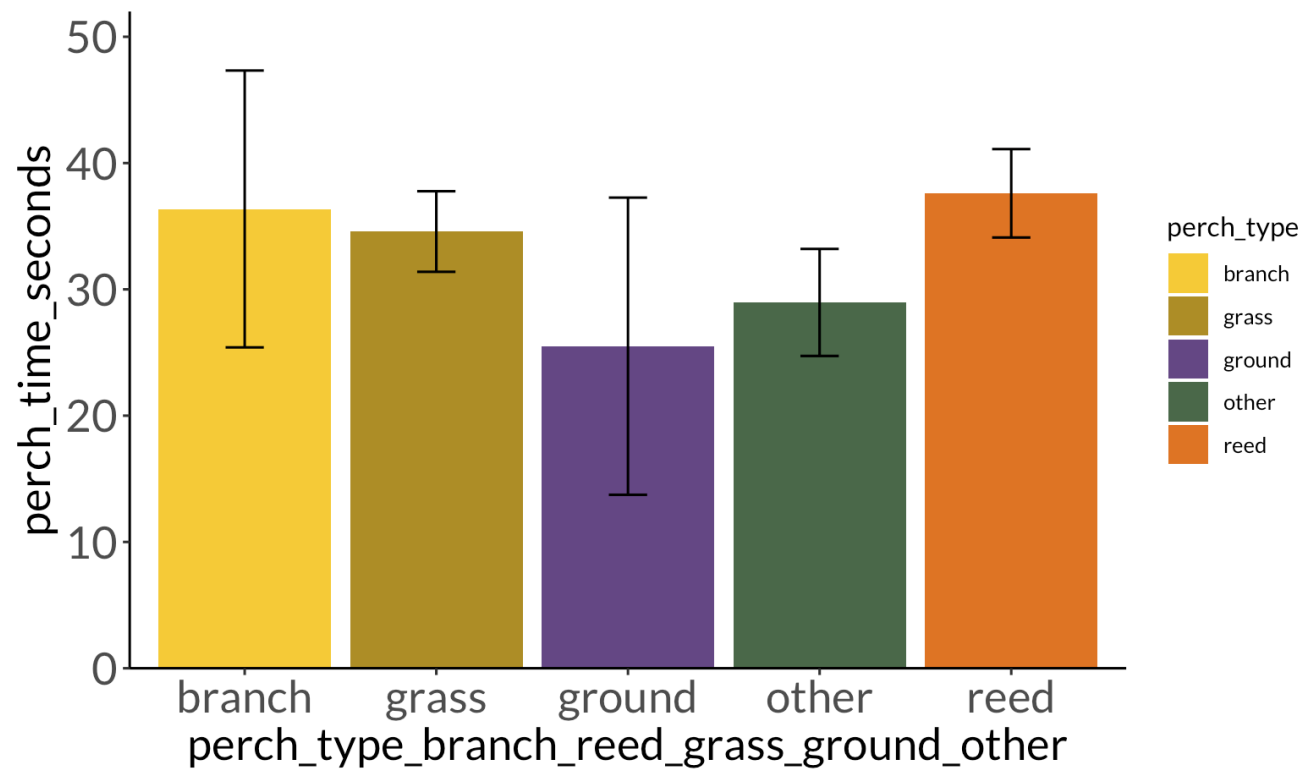
Then, draw the plot that represents your hypothesis.

When making a graph, ask yourself these questions (in order of importance):

1. Are the data I'm showing correct?
2. Am I responsibly communicating the story?
Solution: choose the right graph for your variables!
3. Is it clear for the audience?

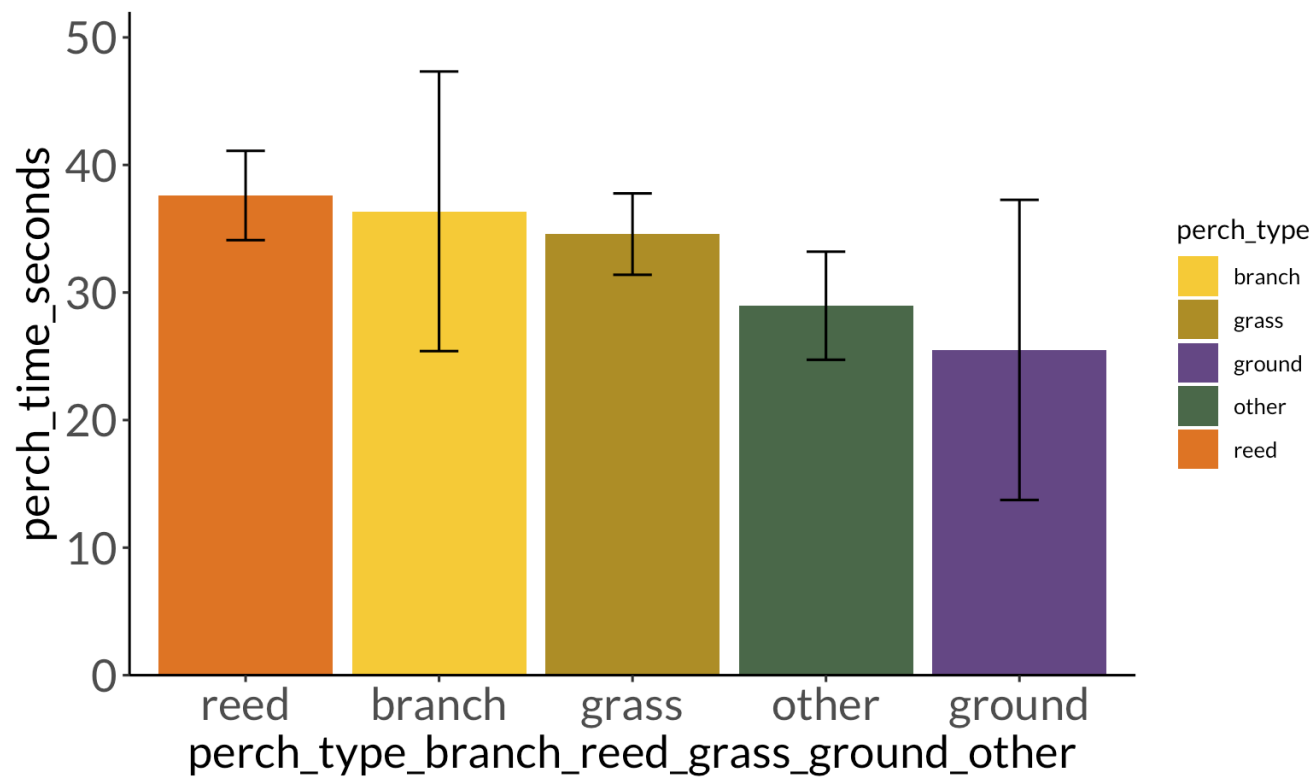
Is it clear for the audience?

- You control what people see (colors, shapes, lines) and the order in which they see them
 - what should be viewed together?
 - what should be picked out?
 - what should be seen in order?
- Ask yourself: what is the “main message” of my graph?



How is the x-axis ordered?

Solution: reorder the axis!

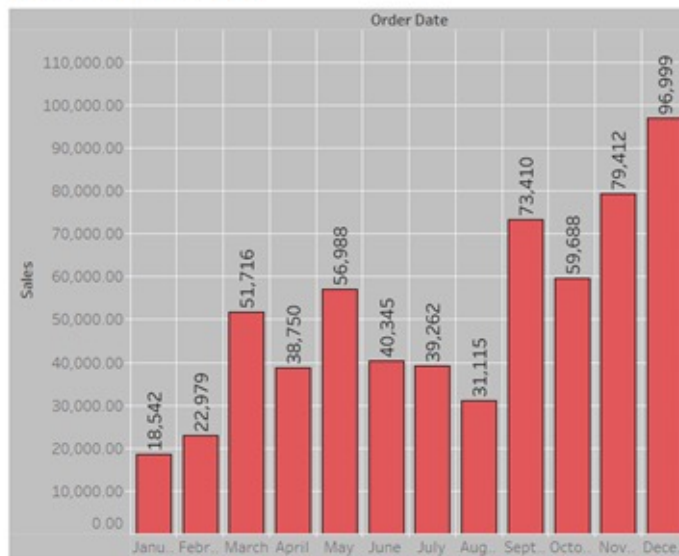


Visual clutter: the data to ink ratio

ratio of elements in a visualization conveying information to the total elements in the image

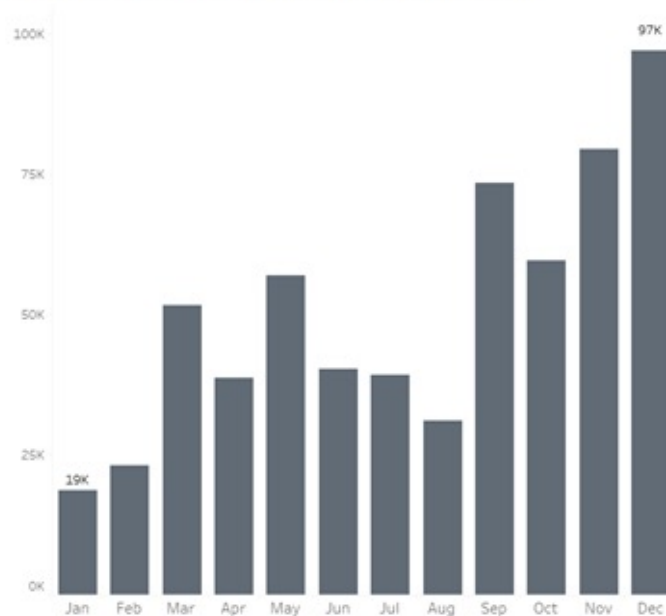
Low data:ink ratio

Monthly Sales Analysis of a USA Superstore: Unveiling Revenue Trends and Seasonal Patterns for a Successful Business Year in 2020

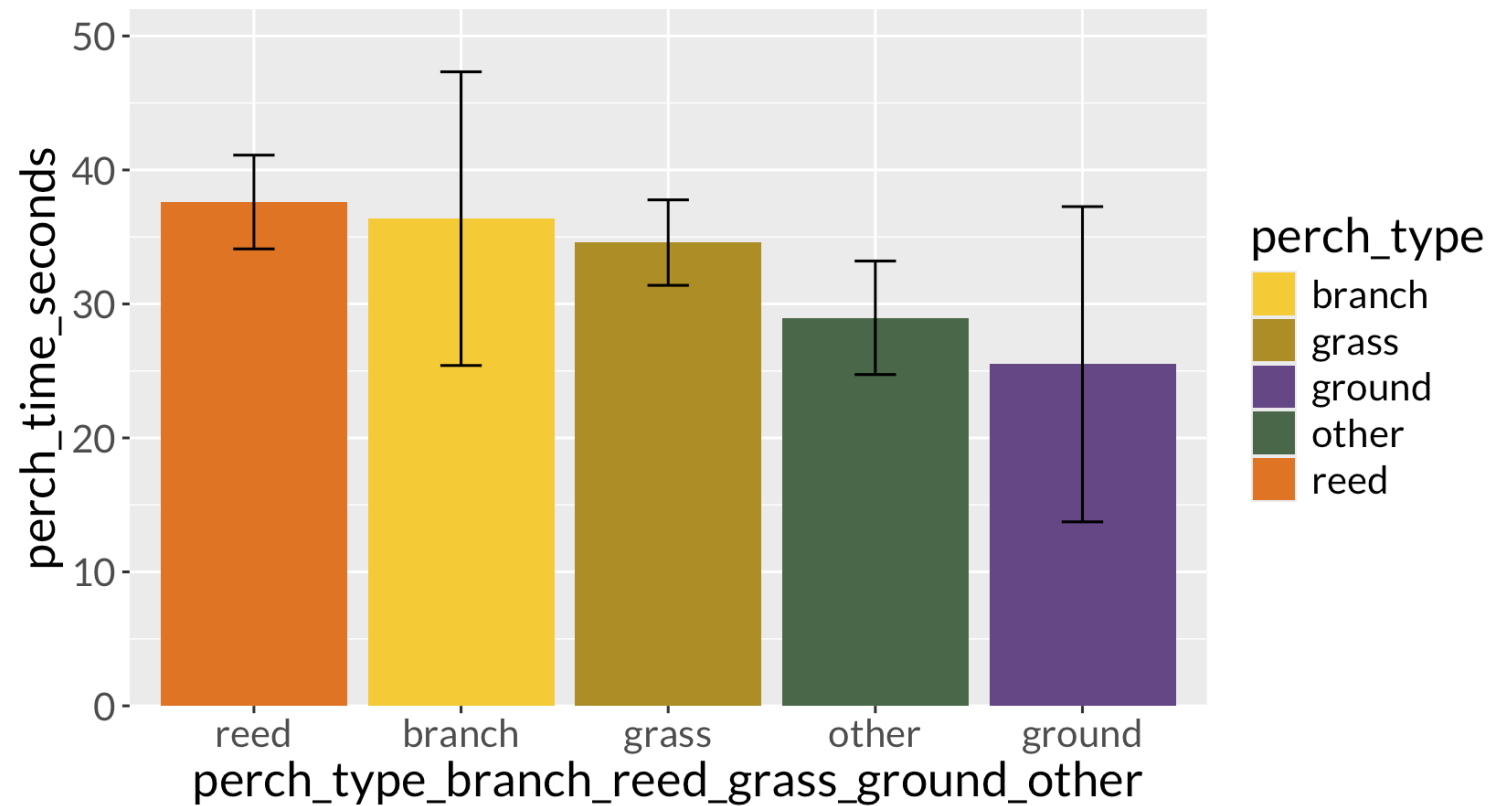


High data:ink ratio

USA Superstore Monthly Sales by Months, 2020

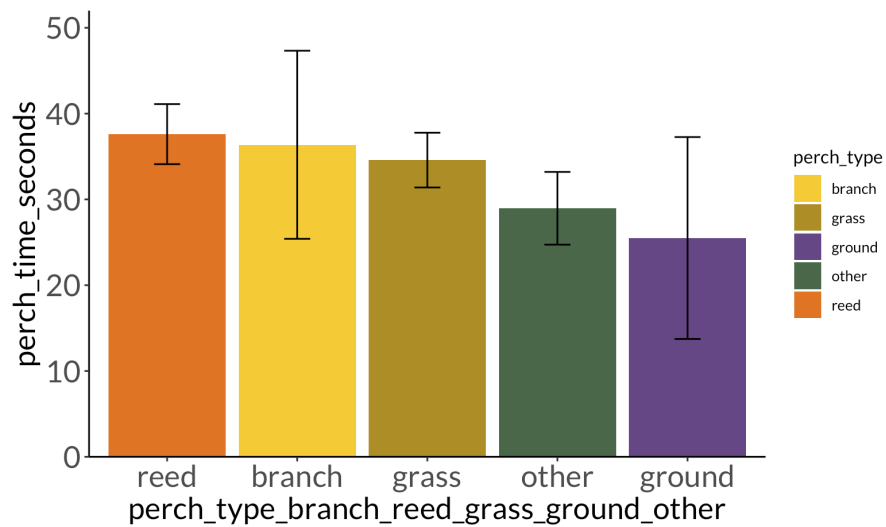


Does it help audience understanding? If not, take it out!



Improving data:ink ratio with a different kind of plot

“Dynamite plot”: bar with whiskers
(derogatory)



Dot and whisker

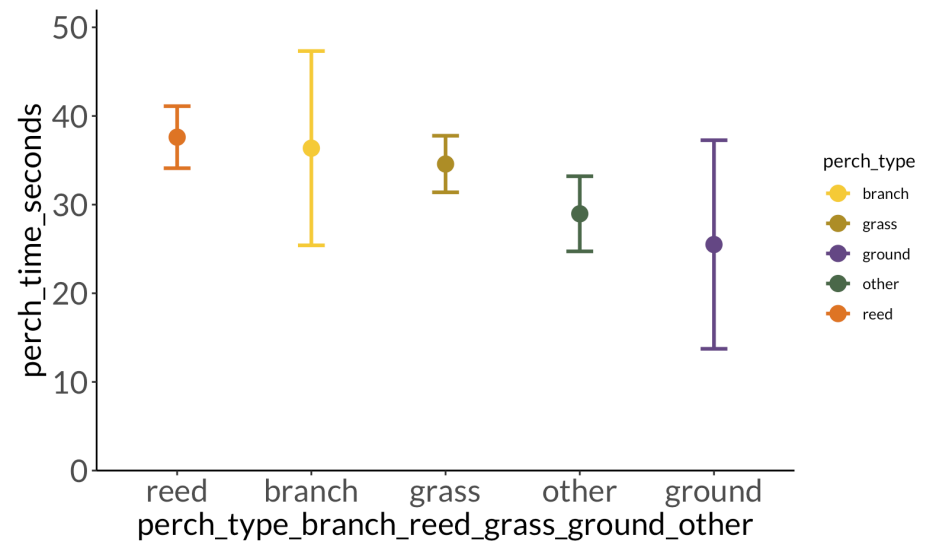
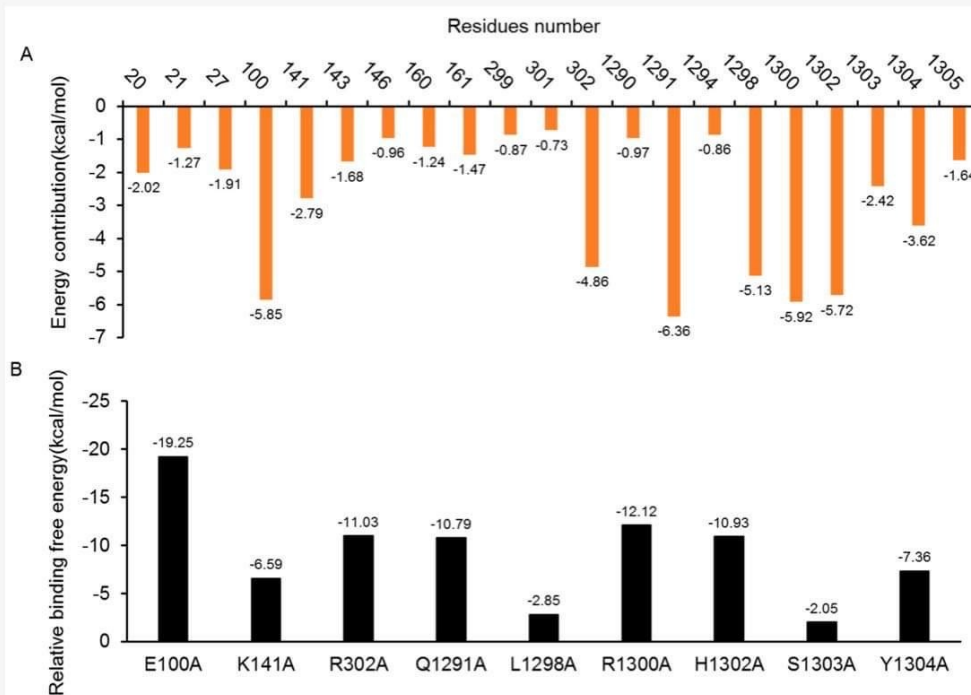


Figure 6. (A) Per-residue binding energy decomposition of predicted GluN2B-CT_{1290–1310}/DA PK1 complex **1**. The energy contribution (the absolute value) larger than 0.60 kcal/mol to at least one of the studied residues for the binding of GluN2B-CT_{1290–1310}/DA PK1 are displayed. The orange bar shows the residues with an absolute binding free energy value of more than 0.60 kcal/mol. **(B)** Alanine scanning analyses of predicted GluN2B-CT_{1290–1310}/DA PK1 complex **1**.



What is wrong with this figure?

What would you do to fix it?

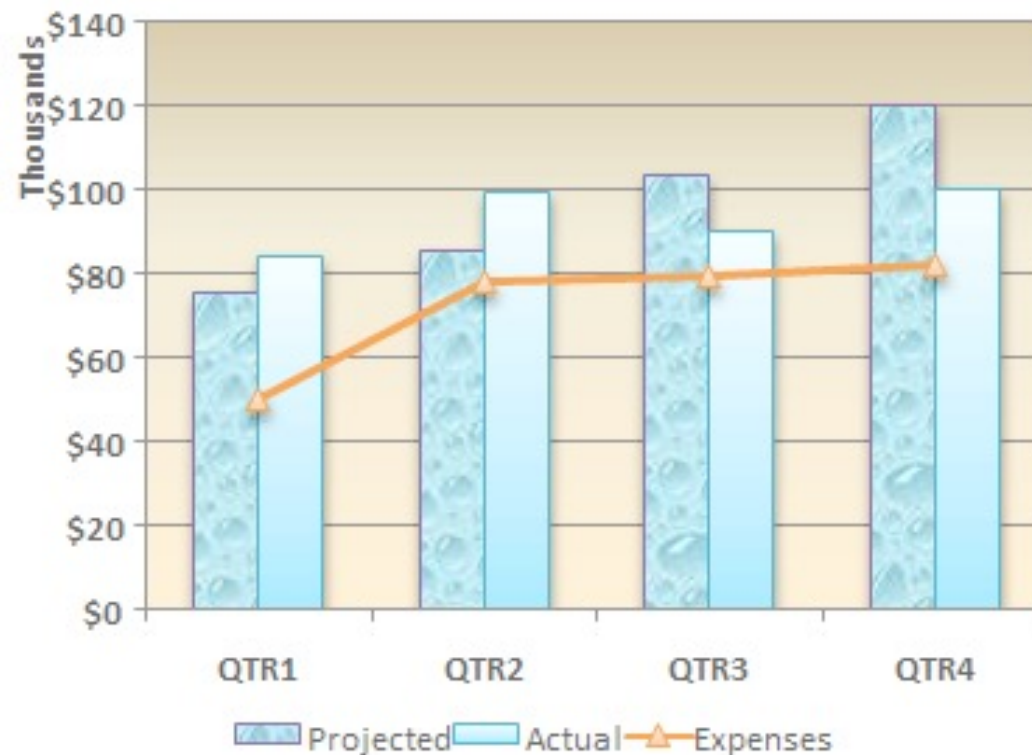
When making a graph, ask yourself these questions (in order of importance):

1. Are the data I'm showing correct?
2. Am I responsibly communicating the story?
Solution: choose the right graph for your variables!
3. Is it clear for the audience?
Solution: get rid of visual clutter!
4. Does it look awesome?

Does it look awesome?



Colors, patterns, etc. are fun – but what do they add?



3 major components of “color”



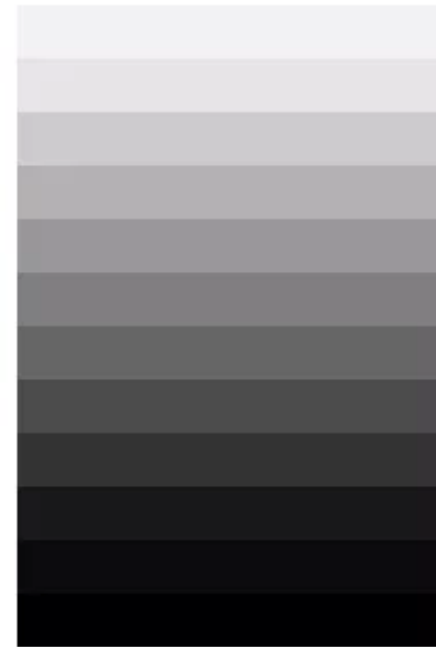
Hue

different colors (e.g. red, blue, purple)



Saturation

color intensity, vivid → neutral



Value

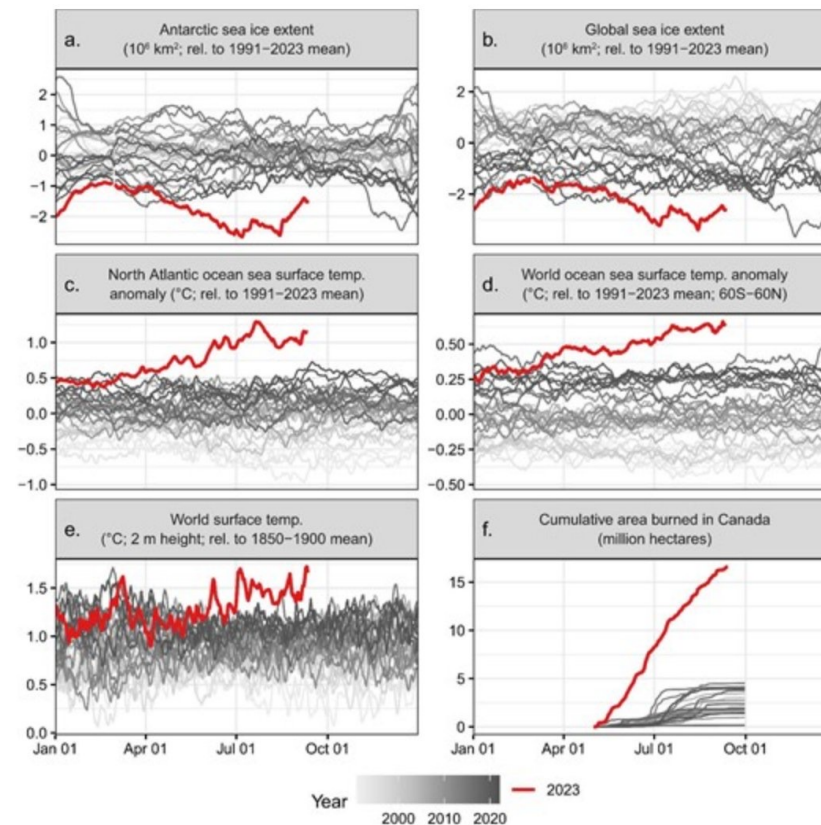
lightness or darkness of a hue

Solution: use different colors, transparencies, hues

- highlighted colors: different color than everything else
- transparencies: highlight lines of best fit or summary statistics while showing underlying data
- hues: show differences between groups

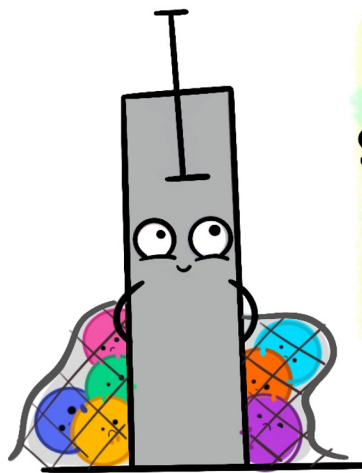
Highlight main points with different hues, saturations, or values

Figure 1.



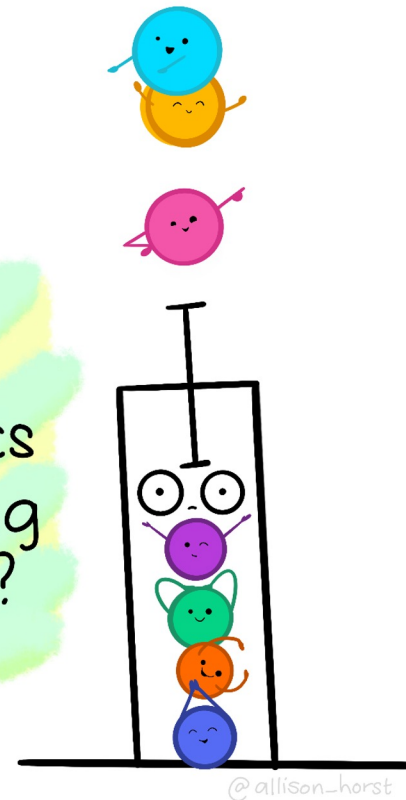
How is Earth's climate different from what it was before?

Summary statistics don't tell the whole story!



Artwork by @allison_horst

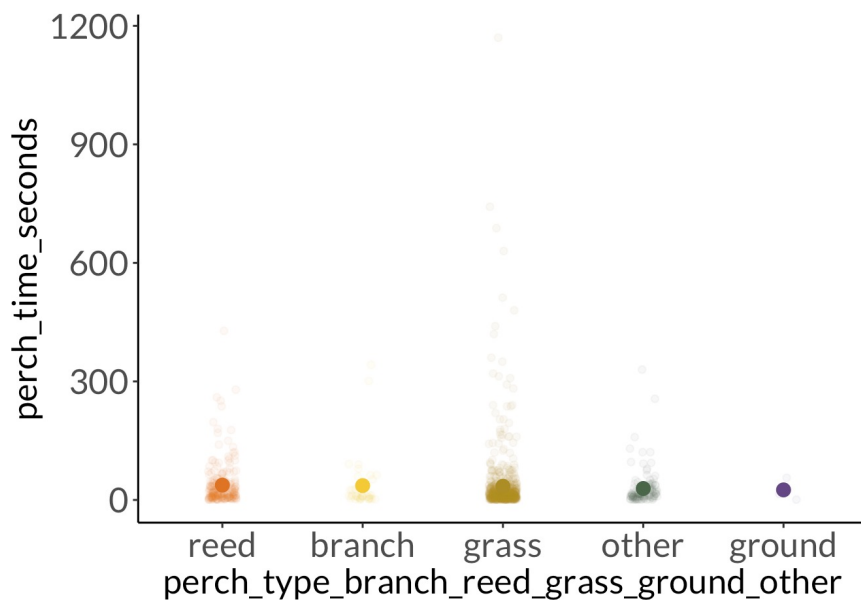
are your
summary statistics
hiding something
interesting?



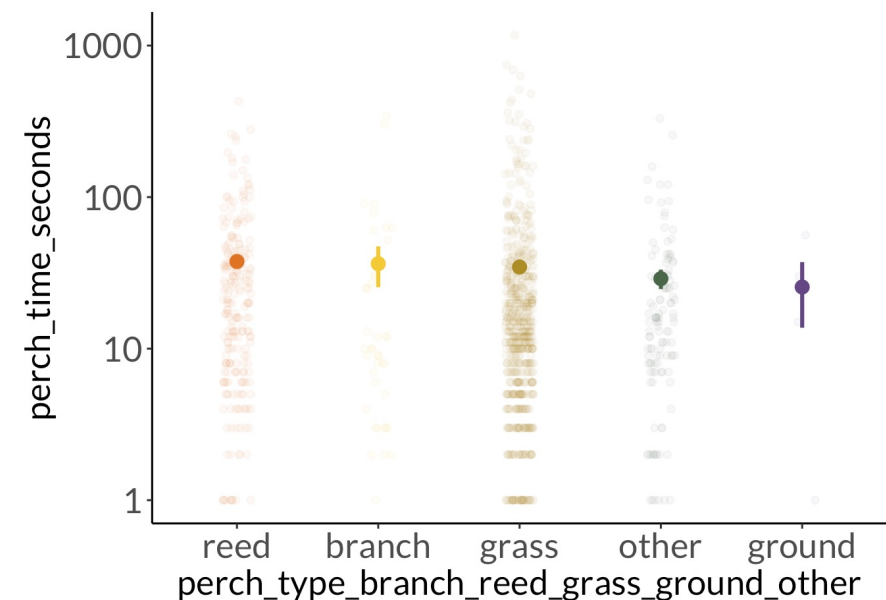
@allison_horst

Improving summaries: show data with different transparencies

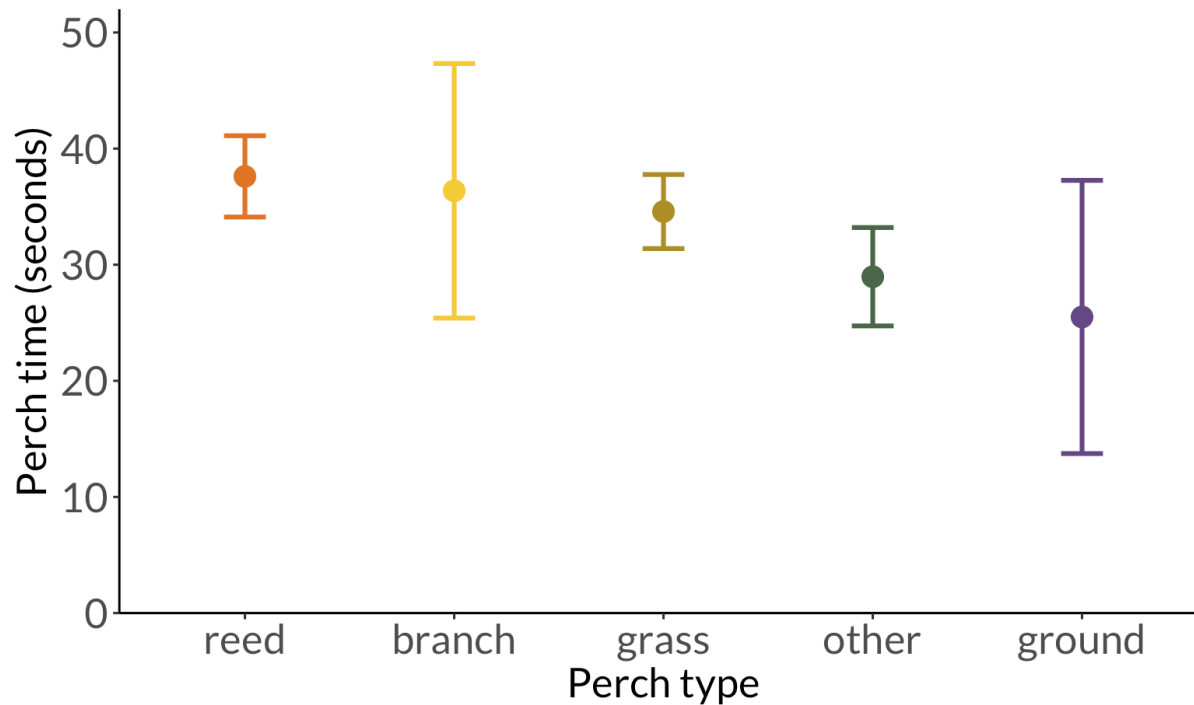
With the axis on the seconds scale:



With the axis on a log10 scale:



Another fix: using full labels instead of acronyms and/or direct labelling



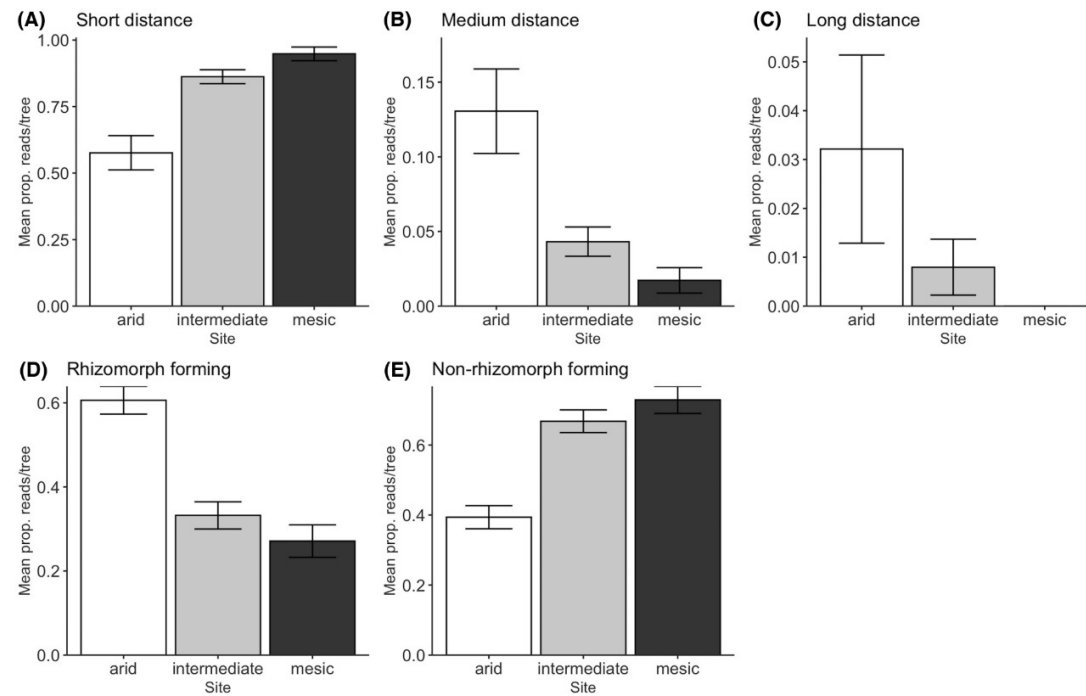
When making a graph, ask yourself these questions (in order of importance):

1. Are the data I'm showing correct?
2. Am I responsibly communicating the story?
Solution: choose the right graph for your variables!
3. Is it clear for the audience?
Solution: get rid of visual clutter!
4. Does it look awesome?
Solution: use hue, saturation, value within colors and clean up visual clutter!

Breaking the rules is ok

- Does data visualization have rules or does it all just depend?
- Master the rules – then break them
- Why you sometimes need to break the rules

People are capable of change!



Bui et al. 2020