Data Visualization: Techniques and Tools

Snyder Institute Science Communication Workshop

November 2nd, 2022

Instructor: Dr. Lindsay Hracs

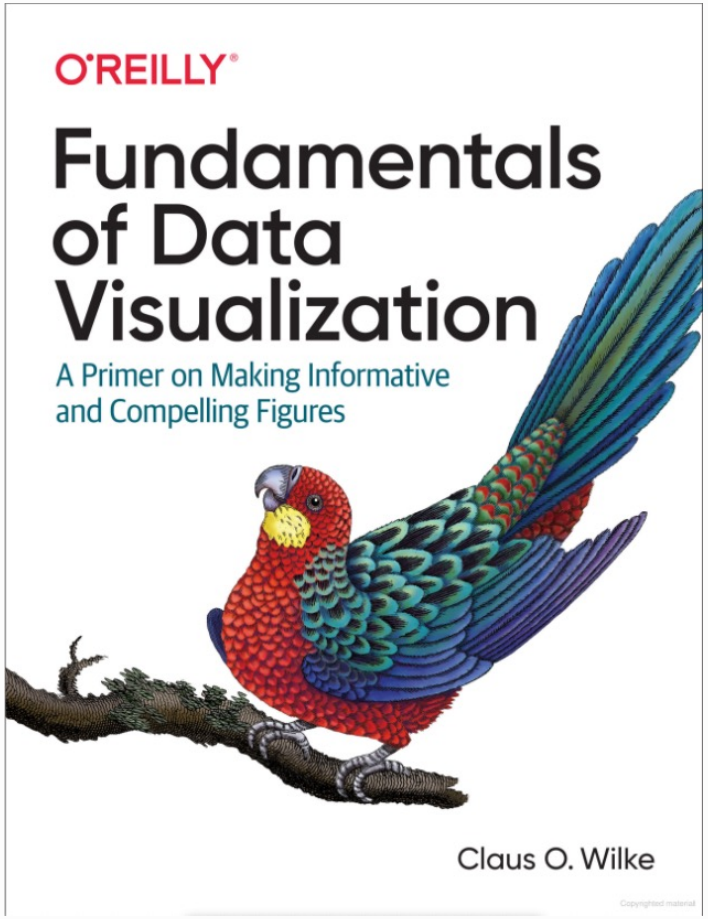
Territorial Acknowledgement

In the spirit of respect, reciprocity, and truth, I honour and acknowledge Moh'kins'tsis and the traditional territories of the people of the Treaty 7 region, which includes the Blackfoot confederacy, comprising the Siksika, Piikani, and Kainai First Nations, as well as the Tsuut'ina First Nation, and the Stoney Nakoda, including the Chiniki, Bearspaw, and Wesley First Nations. I acknowledge that this territory is home to the Métis Nation of Alberta, Region 3 within the historical Northwest Métis homeland. Finally, I acknowledge all Nations – Indigenous and non – who live, work, and play on this land, and who honour and celebrate this territory.

Finally, I highlight the necessity of action to go along with these words, including learning from those who choose to share their lived experiences, working to understand roles in colonialism, and challenging our biases related to race and identity as we move forward.

What is data visualization?

"Data visualization is part art and part science."
- Claus O. Wilke (2019)



- What are some different ways to visually represent data?

  Tables, bar plots, maps, histograms, scatter plots, line graphs, Venn diagrams, area graphs, pyramids, pie charts, box and whisker plots, radial bar plots, tally plots, cluster plots, density plots, heatmaps, violin plots, stem and leaf tables, stream graph, candlestick charts, radar charts, arc diagrams, network diagrams, tree diagrams, progress bars, word clouds, pictograms, error bars, flow charts, calendars, bubble charts...

  Question: How do you choose which visualization to use?

  Answer: It depends on the story you want to tell. (And your research questions.)

Why do we visualize data?

“The frustrating part of building a good visualization is that once you see the thing you needed to see, you don't need the visualization anymore.”

- Richard Biddle, Carleton University

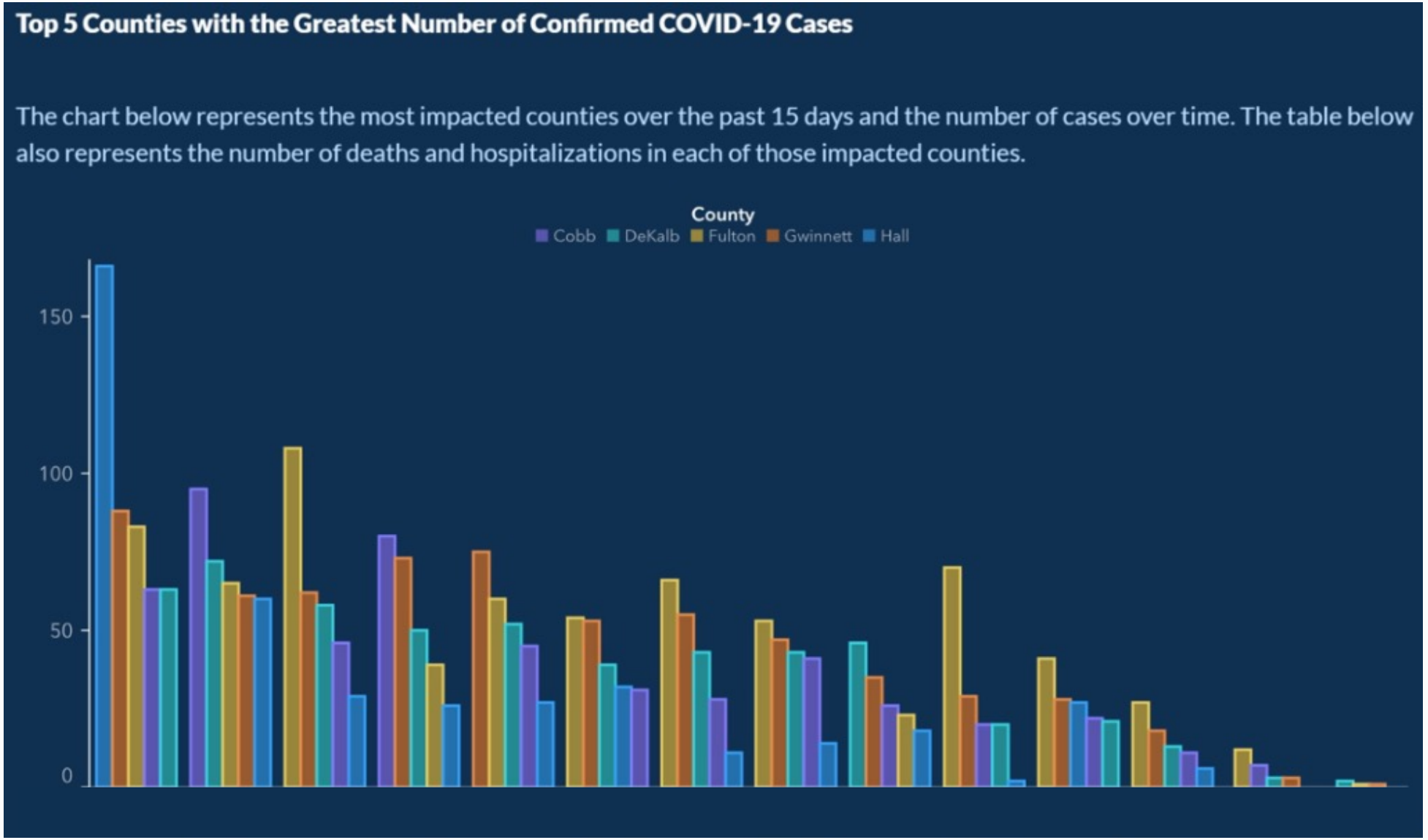
(John Aycock, UCalgary CPSC, p.c. 25-May-22)

We use data visualizations to...

- tell a (precise, accurate, and complete) story

- summarize numbers and text

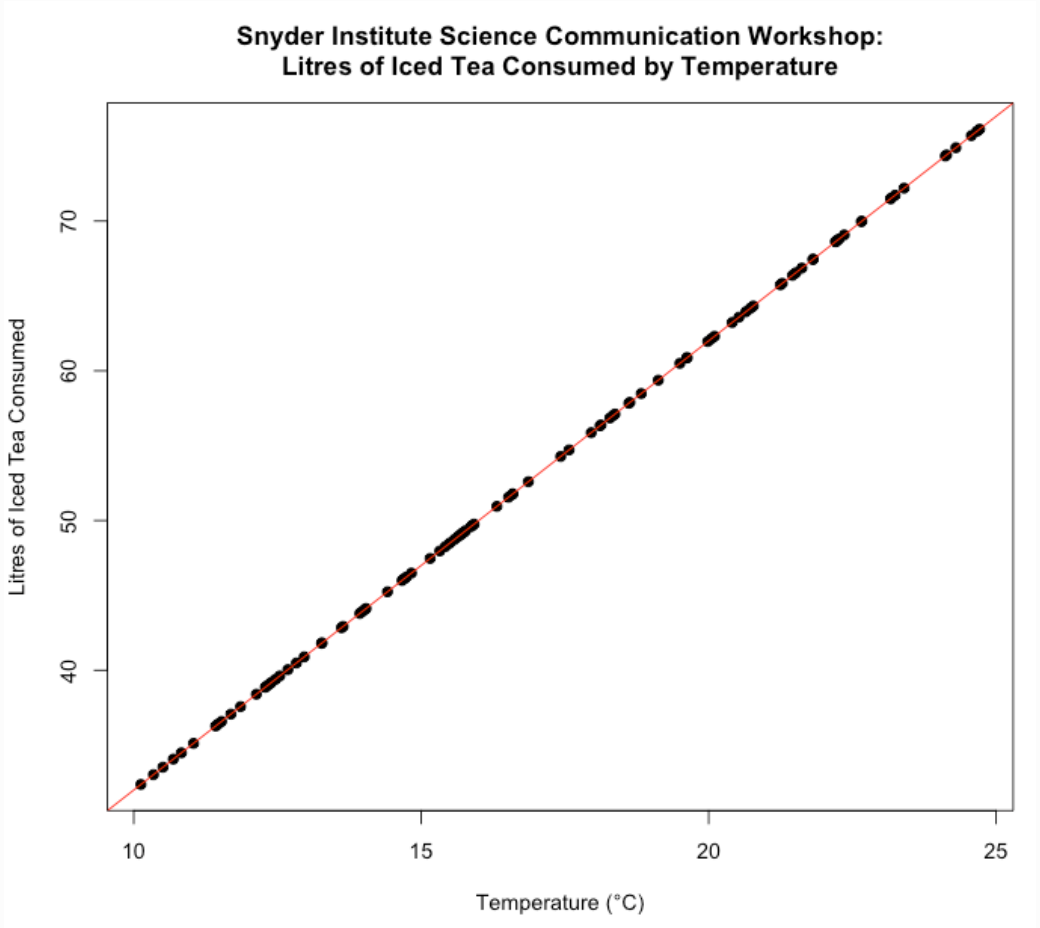
- take everything Joey taught us and make it obsolete (except maybe the exterminating zombies part...)
  - reduce structure; conserve energy
  - focus readers' attention
  - in reality, it should pair well with the text and enhance the overall discussion



Original source: Georgia Department of Public Health
Source: https://anderson-review.ucla.edu/graphic-presentation-of-covid-19-data-can-skew-perceptions-of-risk/

A good visualization should...

1. increase the ability to discover **I**nsights about the data



A good visualization should...

1. increase the ability to discover **I**nsights about the data
2. generate **C**onfidence in the data
3. convey the **E**ssence of the data
4. minimize the **T**ime needed to answer questions about the data

→ **ICE-T** model for determining the value of a visualization

Wall E et al. A heuristic approach to value-driven evaluation of visualizations. *IEEE Trans. on Visualization and Computer Graphics*. 2019;25:491–500.

- we want to increase the ability of others interact with data while reducing the cognitive load placed on them, but we can also use visualizations as a tool in data exploration

  (and get the most out of meetings with your PIs and supervisors!)

  - trends in the data
  - before analysis (e.g., are your data normally distributed?)
    → quick and dirty functions in R: **hist()**, **qqplot()**, **qqline()**
  - identify outliers, help with data cleaning

What information can/should we visualize?

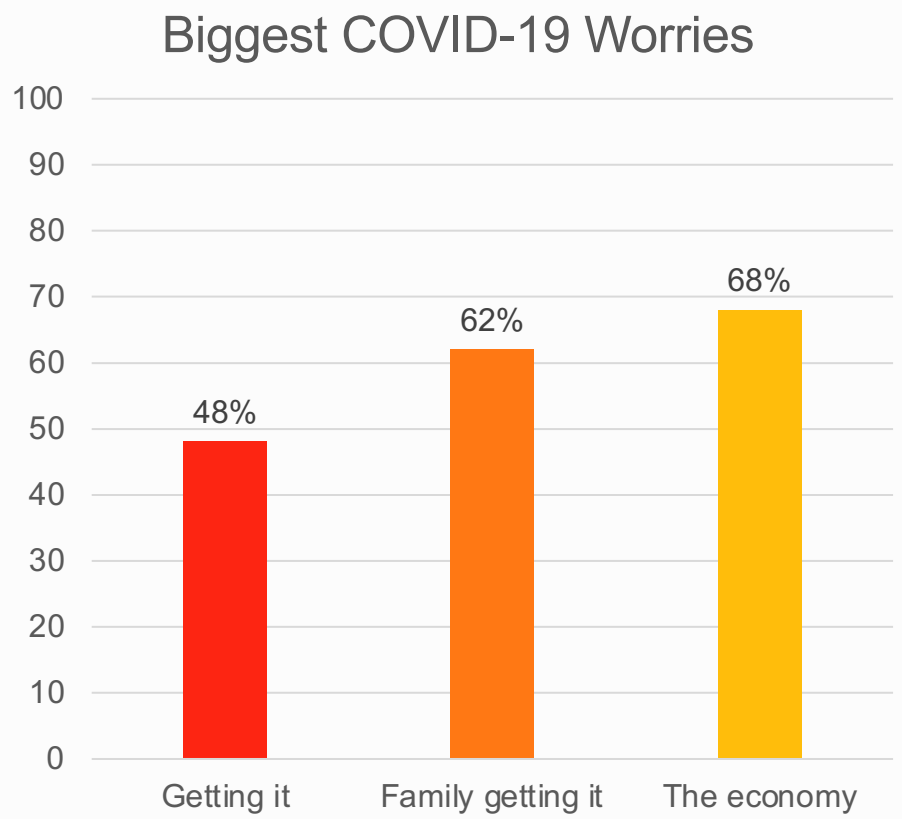
- there is probably a visualization for any type of information/data you have, but it is important to keep in mind:

GARBAGE IN, GARBAGE OUT (GIGO)

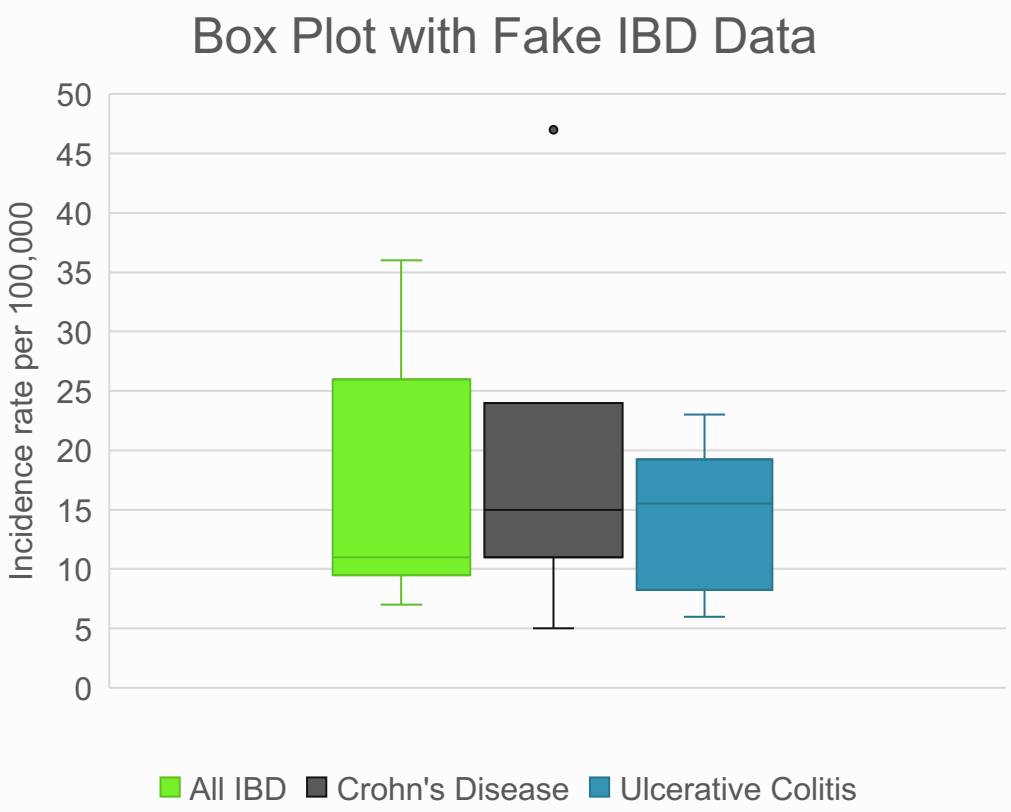
- you will likely spend most of your time prepping your data either outside or inside of the statistics/analysis/visualization software

Choosing the right visualization

Source: https://venngage.com/blog/misleading-graphs/



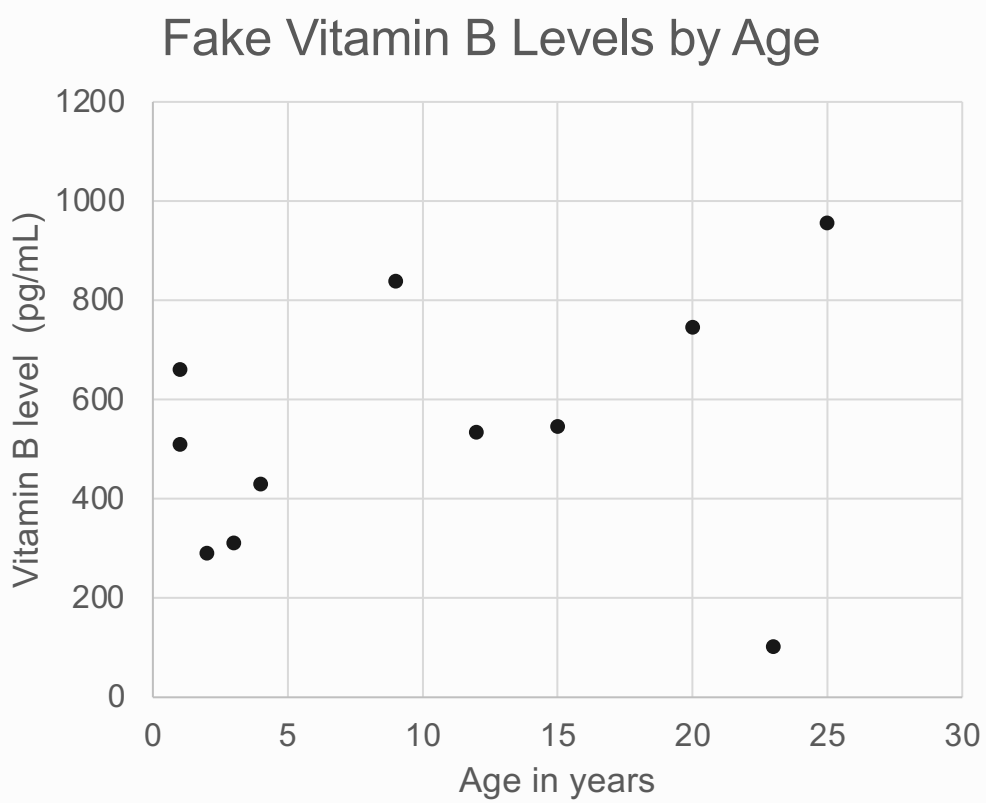
Box plots



- use when you want to compare distributions of a continuous variable that is split by category
- compare multiple distributions at a time

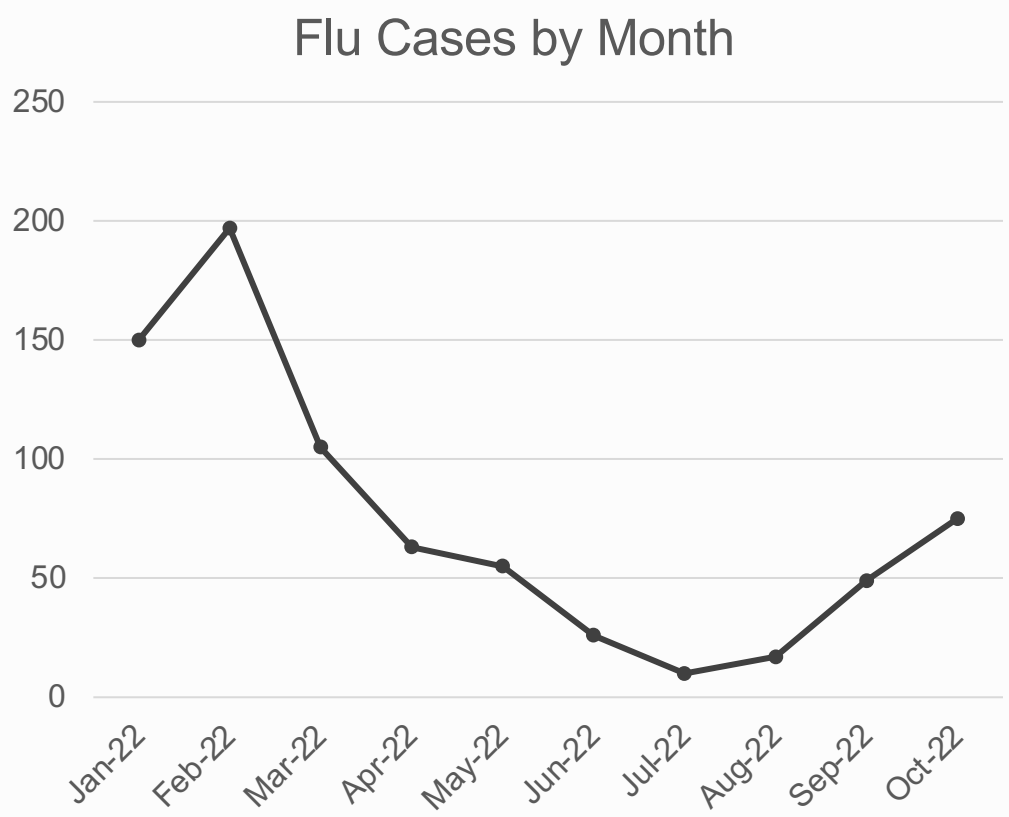
- line = median
- top of box = upper quartile
- bottom of box = lower quartile
- top whisker = upper extreme
- bottom whisker = lower extreme

Scatter plots



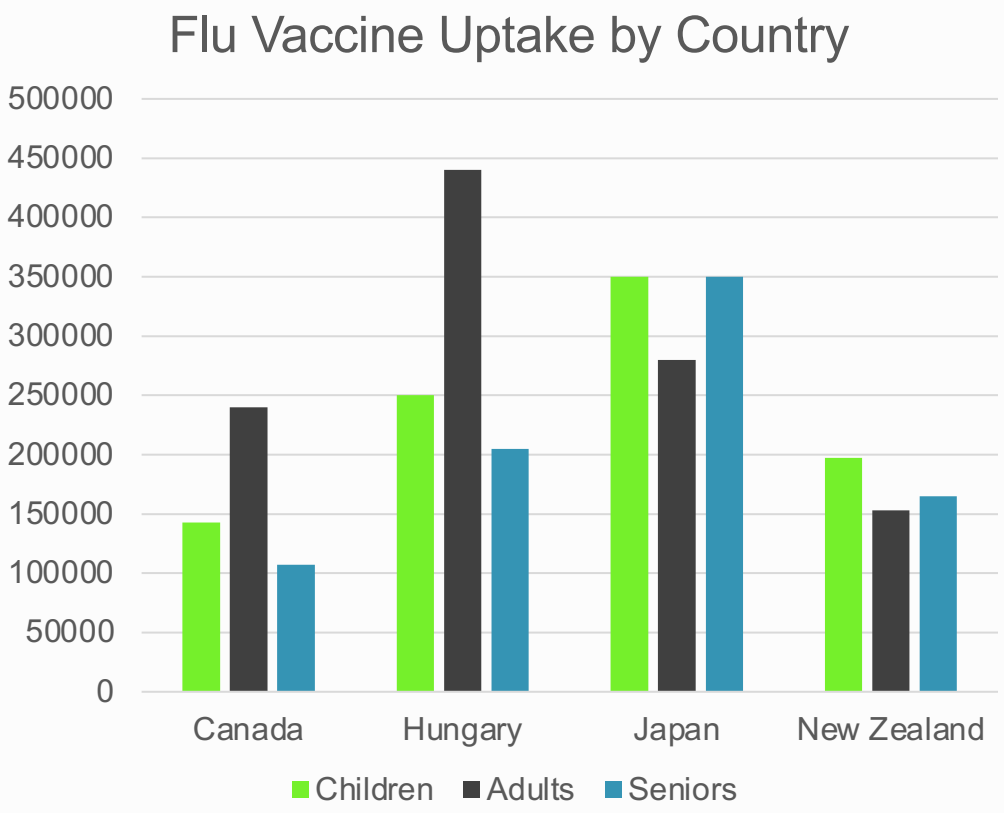
- use when you have two continuous variables
  - is *age* a continuous variable?
- you are interested in finding out whether there is a relationship between the variables

Line plots



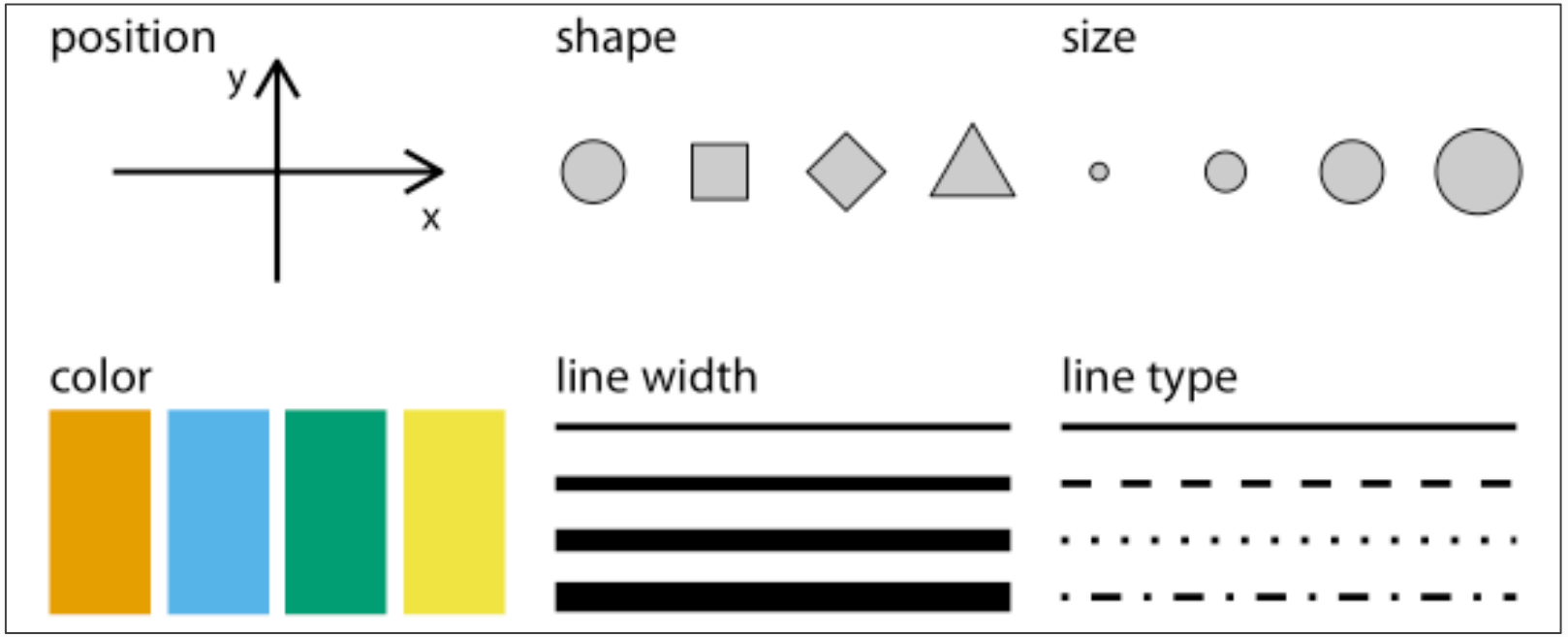
- use when you have two continuous variables that are connected or even grouped somehow
- you are interested in finding out whether there is a relationship between the variables
- often used with dates or times on the x-axis

Bar plots



- similar use case to box plots, but does not give you distributional information
- use when you are interested in comparing counts or percentages of a categorical variable

How does visual information convey meaning?



Wilke (2019: 8)

- all graphical elements have a *size*, a *shape*, and a *colour*
  - on plots, graphical elements also have a *position* (cartesian, sequential, etc.)
  - lines have a *width* and a *type*
  - text has a font *family* and *face*
- aesthetics are a combination of personal preference and best practices

- Wilke argues that aesthetics fall into two categories: those that can represent continuous data, and those that cannot

Continuous Data Aesthetics

position

size

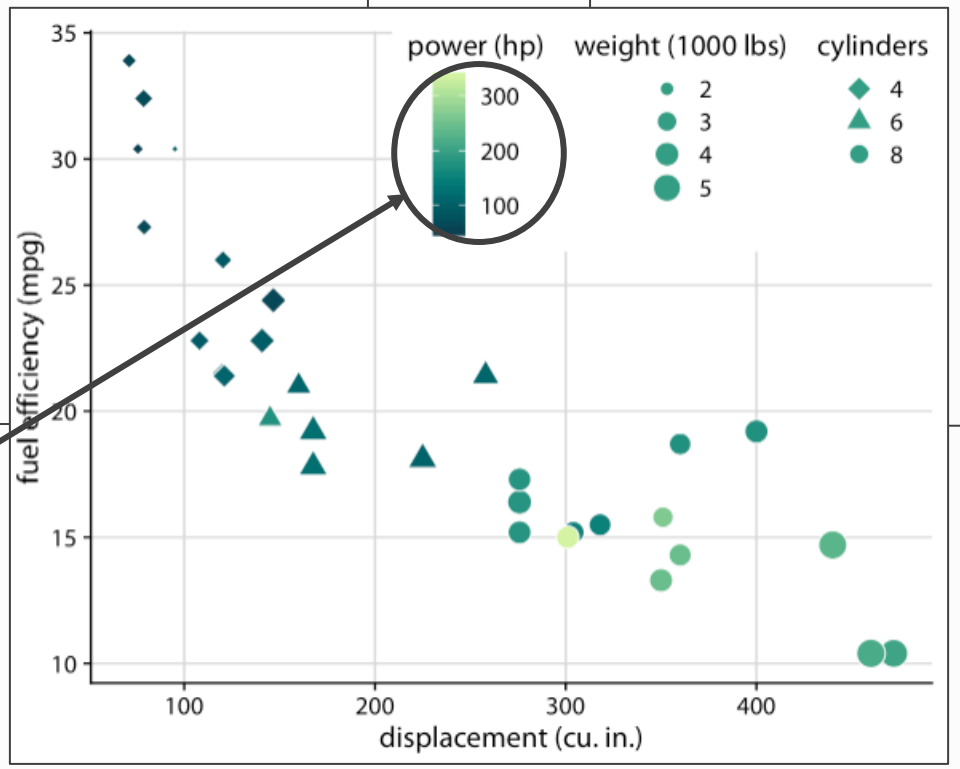
colour

line width

Discrete Data Aesthetics

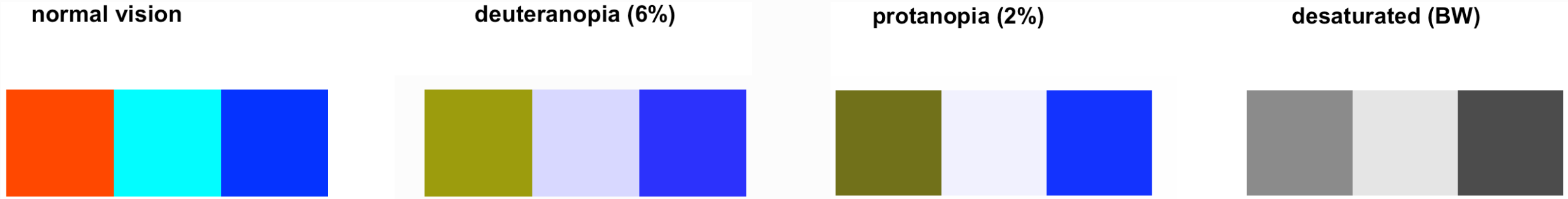
shape

line type



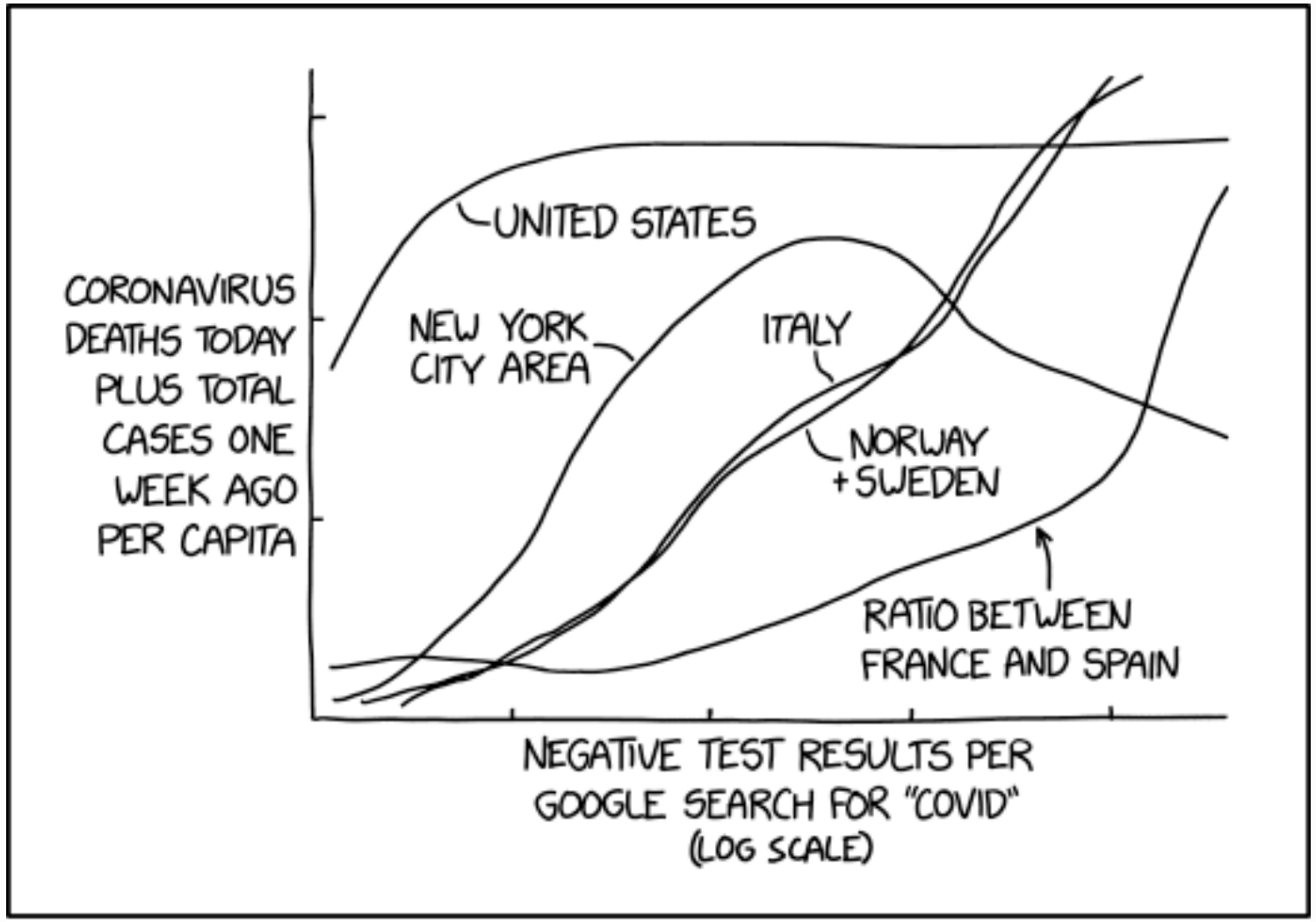
NOTE: darker shades and more saturated colours should be used for larger quantities!

- additional design considerations to make:
  - add Image Alternative Text using *knitr* package in R; search best practices for Image Alt Text
    - limitations regarding which output types *knitr* works with, but since R is open-source this is actively being worked on
  - use colour-blind friendly colour palettes from packages such as *colorBlindness, munsell*, *viridis*, *RColorBrewer*, *dichromat*, *colorblindr*, *shades*, or *ggsci*



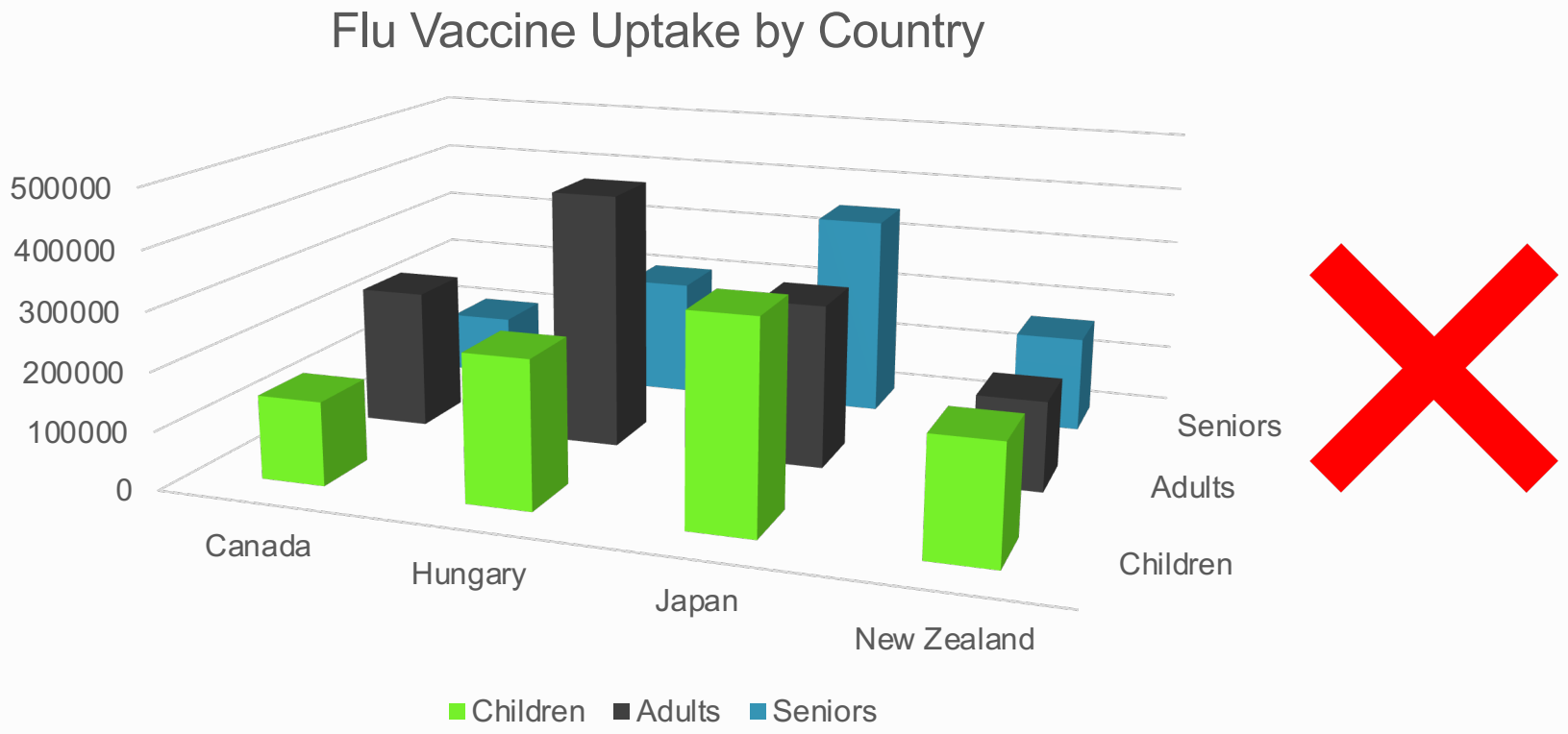
What else can we do with data visualization?

(Hint: Not this.)



https://xkcd.com/2294/

- three-dimensional (3D) plots?



- the rule of thumb is to avoid 3D visualizations when using a 2D medium (e.g., journal article, PowerPoint presentation, etc.)

- web-based interactive visualizations increase usability of 3D plots

- *shiny* applications for dynamic, interactive data visualization

  STOP COVID-19 in IBD dashboard

  Global Hospitalization Trends for Crohn's Disease and Ulcerative Colitis in the 21st Century

Additional Resources

Fundamentals of Data Visualization
https://clauswilke.com/dataviz/

knitr
https://cran.r-project.org/web/packages/knitr/index.html
https://yihui.org/knitr/

colorBlindness Guide
https://cran.r-project.org/web/packages/colorBlindness/vignettes/colorBlindness.html

For the second half of today's workshop, you will want to access files here:

https://github.com/kaplan-gi/Snyder-Institute-Science-Communication-Workshop

(type out the URL somewhere and keep it handy)

With any remaining time today I would like to learn more about your data and discuss what you would like to do with the second Data Visualization workshop.