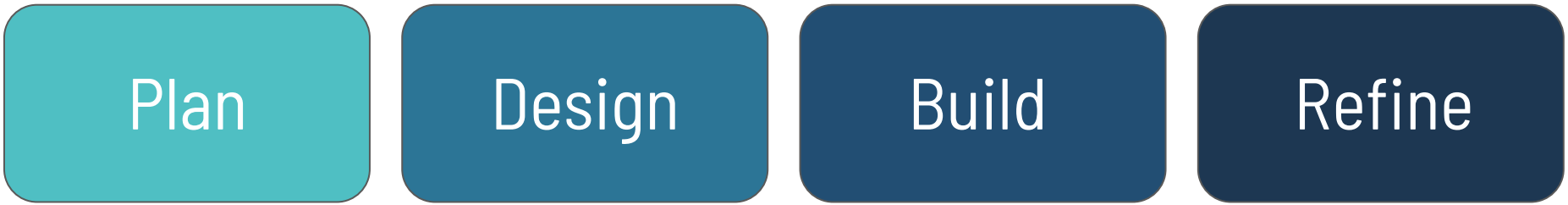
Data Science Basics in R

Day 4: Designing Data Visualizations

Goals for today

- **Learn** a step-by-step process for creating great data visualizations
- **Understand** your audience and your goals when visualizing data
- **Design** some fun and beautiful data visualizations
- **Get creative** and explore some new skills in R

Stages of creating a new data visualization



Plan

Design

Build

Refine

Define goals and audience

Ask yourself

Why are you doing this?
understand your data; support a decision; deliver a message; teach a topic

What are you trying to communicate?
today's status; change over time; comparing groups; highlighting extremes

Who is your audience?
what are their expectations; how much background do they have

When and where will the visualization be used?
peer-reviewed publication; poster; dashboard; slide-deck

Why this matters

Degree of effort
how much time you spend on the plot, based on goals, audience, timeline

Level of complexity
how visually simple or complex your plot is, including the type of plot

Style and formatting
how the data are presented visually, including colors, sizing, resolution

Context
how the data are described in writing and how they are presented overall

Plan | **Design** | Build | Refine

Brainstorm and sketch out ideas

Explore

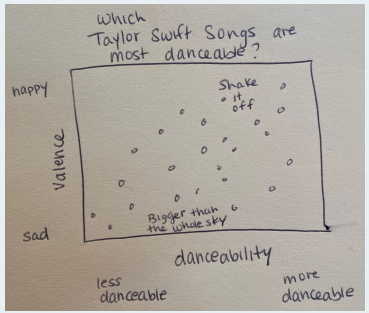
Data sources and types
do you already have the data, or do you also need to find/collect it

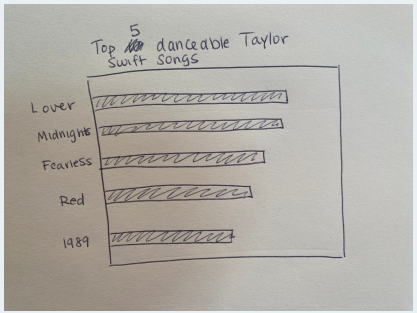
Different types of plots
visual elements (lines, points, size, shape, width) and their layouts

Analysis methods
different ways of summarizing and grouping data, where relevant

Color schemes
hues, color gradients in one direction, color gradients in two directions

Examples





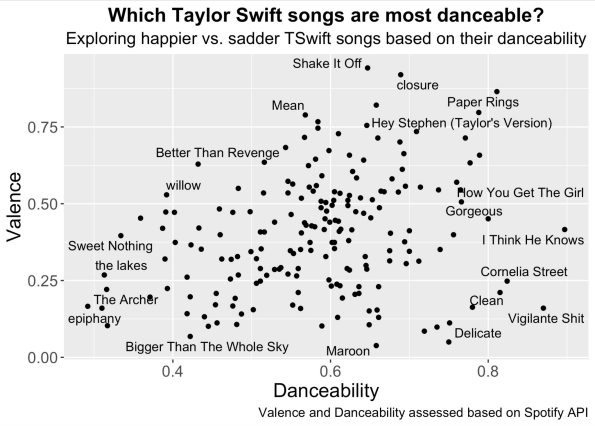
Plan | Design | **Build** | Refine

Create your data visualization

Code or sketch

```
ggplot(taylor_album_songs, aes(x = danceability, y = valence, label = label)) +
  geom_text_repel() +
  geom_point() +
  labs(x = "Danceability",
       y = "Valence",
       title = "Which Taylor Swift songs are most danceable?",
       subtitle = "Exploring happier vs. sadder TSwift songs based on their danceability",
       caption = "Valence and Danceability assessed based on Spotify API") +
    theme(plot.caption = element_text(size = rel(1)),
          plot.title = element_text(hjust = 0.5, size = rel(1.5), face = "bold"),
          plot.subtitle = element_text(hjust = 0.5, size = rel(1.3)),
          axis.text.x = element_text(size = rel(1.5)),
          axis.text.y = element_text(size = rel(1.5)),
          axis.title = element_text(size = rel(1.5)))
```

Examples



Plan

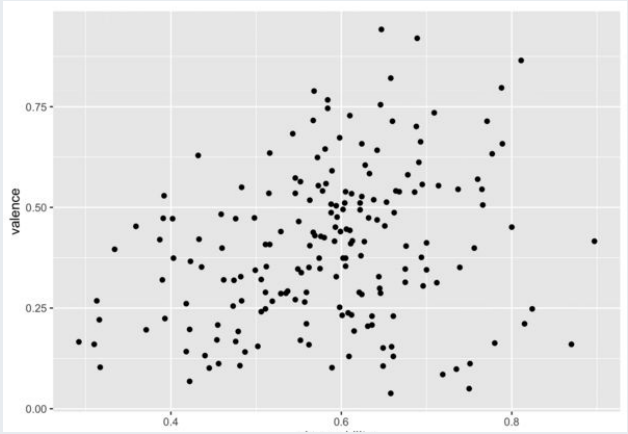
Design

Build

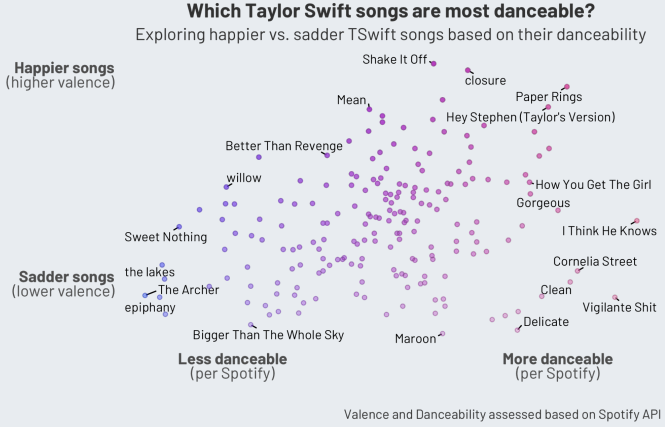
Refine

Review, update, and clarify

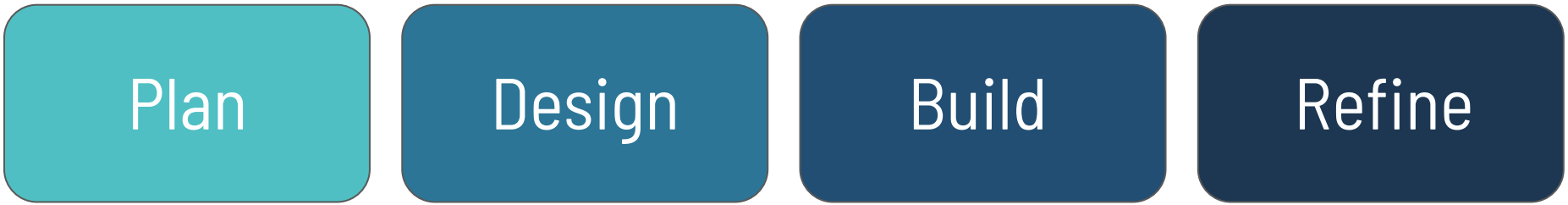
Iterations



Final design



Stages of creating a new data visualization



Stages of creating a new data visualization

Plan

Design

Build

Refine

Define goals and audience

Planning stage:
planning worksheet

Plan

Ask yourself

Why are you doing this?
understand your data; support a decision; deliver a message; teach a topic

What are you trying to communicate?
today's status; change over time; comparing groups; highlighting extremes

Who is your audience?
what are their expectations; how much background do they have

When and where will the visualization be used?
peer-reviewed publication; poster; dashboard; slide-deck

Reminder: Course datasets

Plan

Policy data

Categorical

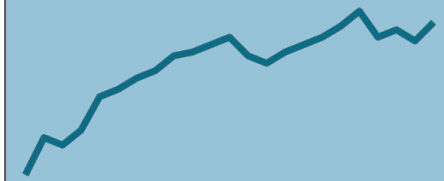
Current national policies related to measles vaccination, per country:

- Required
- Not required
- No data

Coverage data

Continuous

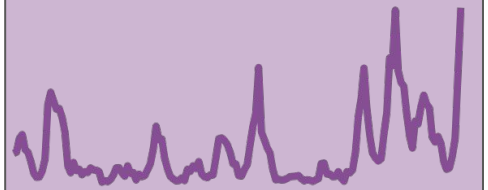
Yearly measles containing vaccine first-dose (MCV1) immunization coverage among 1-year-olds, per country, over time



Caseload data

Continuous

Monthly data on laboratory confirmed, epidemiologically linked, and/or clinical measles cases reported to WHO per country, over time



Your turn

planning worksheet

Plan

04:00

Why are you doing this?

understand your data; support a decision; deliver a message; teach a topic; something else?

What are you trying to communicate?

today's status; change over time; comparing groups; highlighting extremes; something else?

Who is your audience?

Who are they? What are their expectations? How much background do they have? Anything else?

When and where will the visualization be used?

peer-reviewed publication; poster; dashboard; website; slide deck (presented? shared as a deck?)
also note any known formatting constraints (e.g., size, image resolution, color schemes, fonts)

Plan

In-class analysis plan

planning worksheet

Why are you doing this?

understand your data; support a decision; deliver a message; teach a topic; something else?

understand our data and share what we've learned in a simple, beautiful graphic

What are you trying to communicate?

today's status; change over time; comparing groups; highlighting extremes; something else?

highlight the countries or types of countries that are most impacted by measles since 2022, as defined based on total measles caseload information

→ we will refine this as we learn more from the actual data, might include information on measles vaccination policy status, country size, region, or income group

Who is your audience?

Who are they? What are their expectations? How much background do they have? Anything else?

people who are interested in health and who are familiar with time series data, but who have limited familiarity with measles specifically

When and where will the visualization be used?

peer-reviewed publication; poster; dashboard; website; slide deck (presented? shared as a deck?)
also note any known formatting constraints (e.g., size, image resolution, color schemes, fonts)

presented on an informal slideshow, featured on a github profile or maybe a LinkedIn post

Stages of creating a new data visualization

Define goals and audience

Brainstorm and sketch out ideas

Design

Design phase:
brainstorming and sketching

Explore

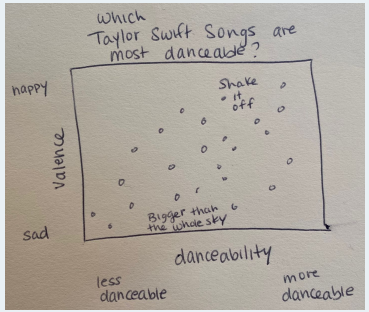
Data sources and types
do you already have the data, or do you also need to find/collect it

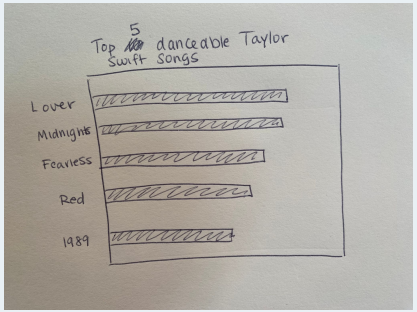
Different types of plots
visual elements (lines, points, size, shape, width) and their layouts

Analysis methods
different ways of summarizing and grouping data, where relevant

Color schemes
hues, color gradients in one direction, color gradients in two directions

Examples



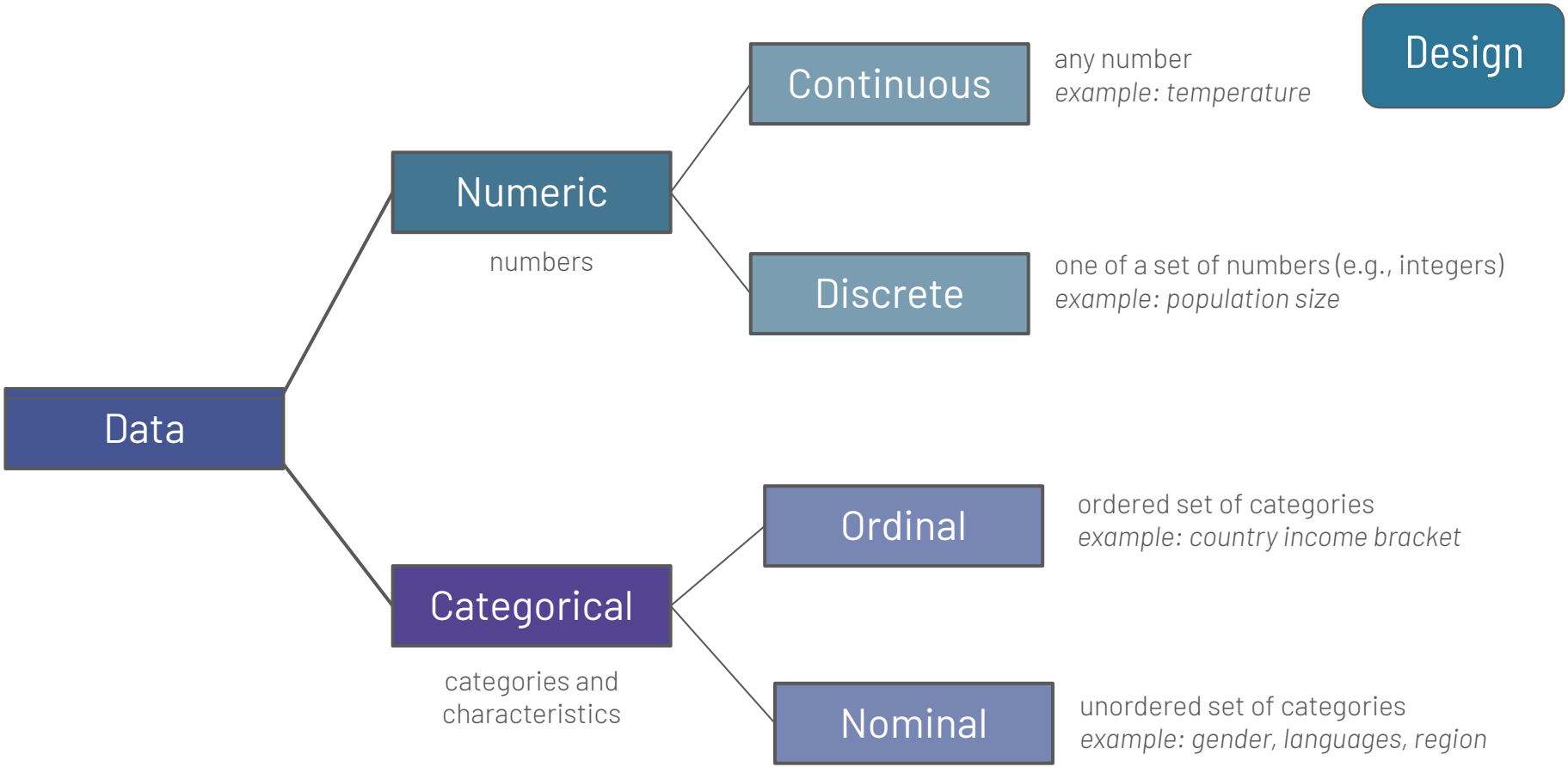


Design

Start with the data

Understand the data you have

- Understand your data, including strengths and limitations
  - How was it collected, for what purpose?
  - What types of data, in what format?
  - How much is missing?
  - Who is represented? Who is not?
- Explore general trends, ask questions, and learn


Design
Data
Numeric
numbers
Continuous
any number
example: temperature
Discrete
one of a set of numbers (e.g., integers)
example: population size
Categorical
categories and characteristics
Ordinal
ordered set of categories
example: country income bracket
Nominal
unordered set of categories
example: gender, languages, region

Cleveland McGill Scale

Design

How do we understand data, visually?

- 1985 study of how humans understand plots
- How accurately are people were able to interpret data?
  - experimented by asking subjects to guess data from graphics
  - varied plot design and details: position, shape, size, symbols
  - ranked most-to-least interpretable visual elements

Cleveland WS, McGill R. Graphical perception and graphical methods for analyzing scientific data. Science. 1985 Aug 30;229(4716):828-33.

Most accurate estimates

Least accurate estimates

Position

position along a specified scale
examples: barchart, scatterplot

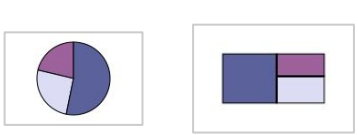
Length

length of a segment
examples: line segment, stacked barchart

Slope/Angle

slope or angle of a line or segment
examples: line chart, donut chart

Area



area of a shape or segment
examples: pie chart, treemap

Color intensity

intensity of a single color
example: varying color saturation or lightness

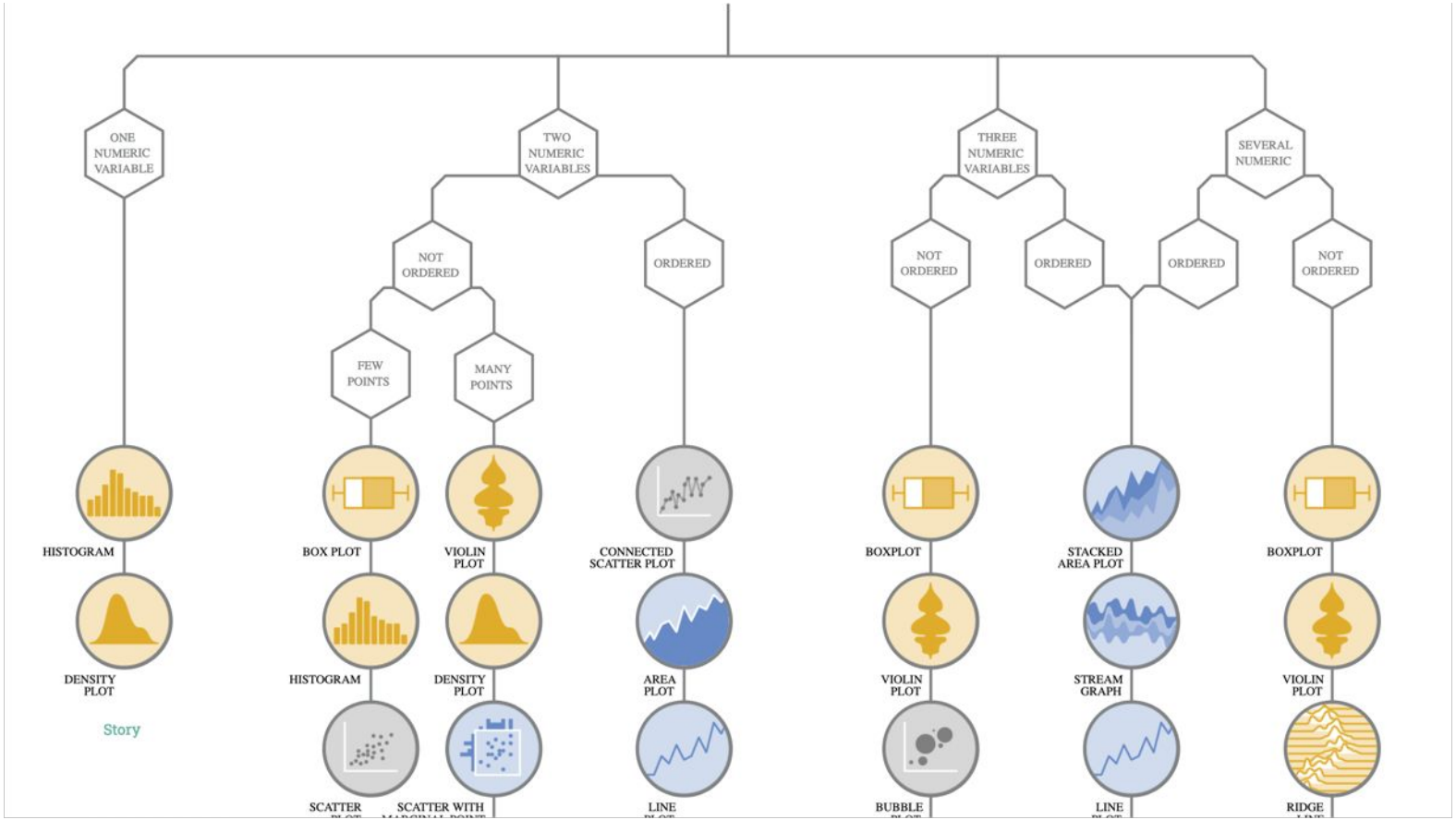
Color hue

sets of of multiple colors or a range of colors
example: multi-color gradient

Design

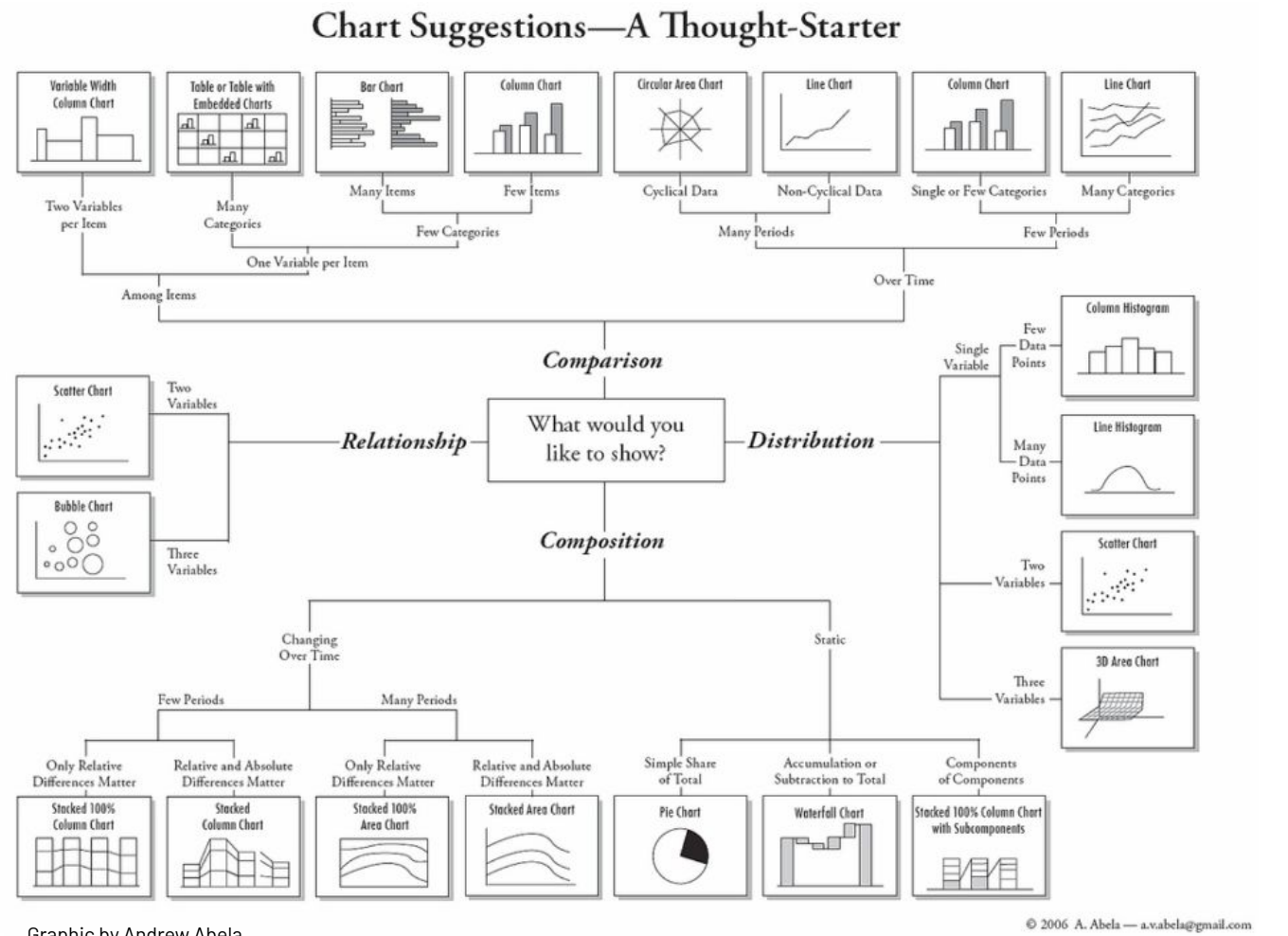
Cleveland WS, McGill R. Graphical perception and graphical methods for analyzing scientific data. Science. 1985 Aug 30;229(4716):828-33.

Design



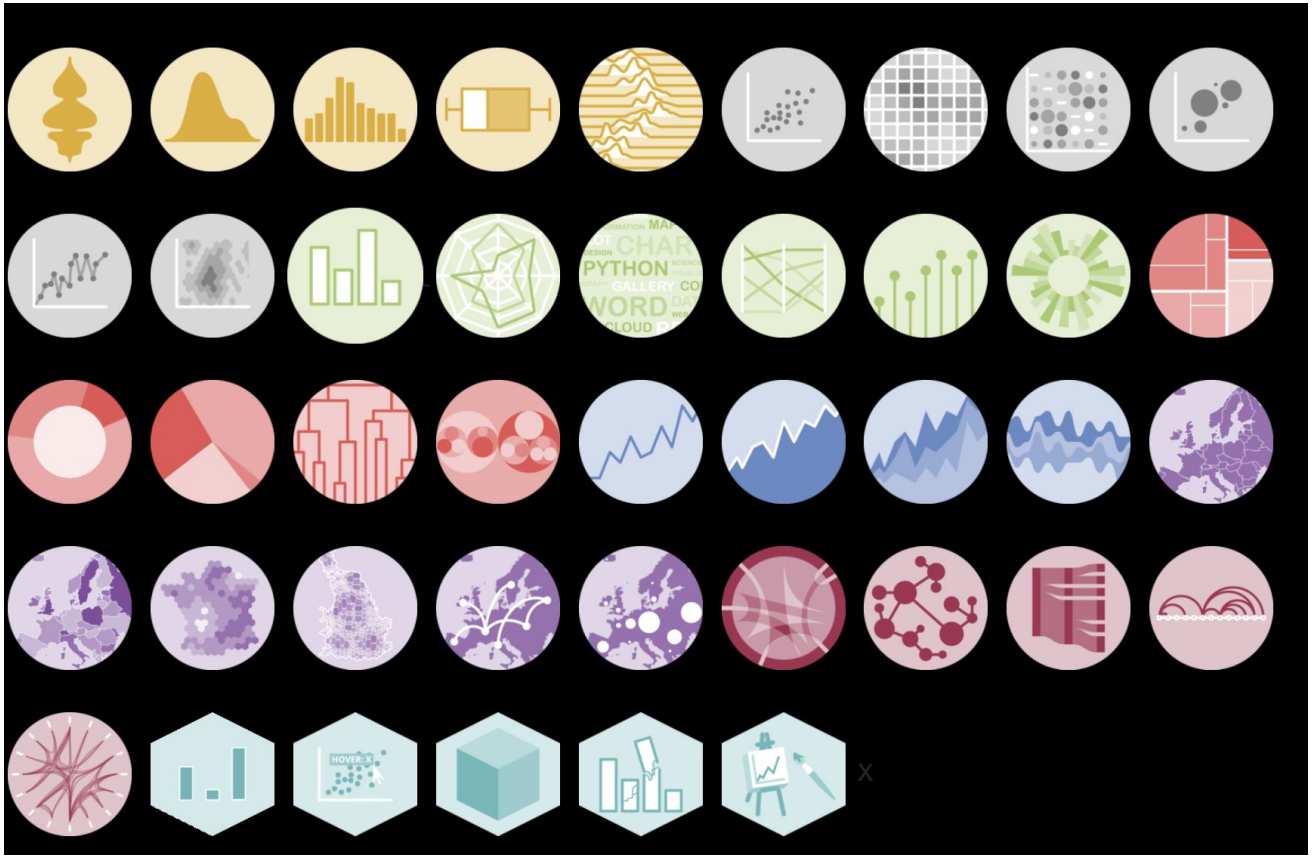
Graphic by Data to Viz
https://www.data-to-viz.com/

Design



Graphic by Andrew Abela
https://extremepresentation.typepad.com/blog/2006/09/choosing_a_good.html

Design



The R Graph Gallery
https://r-graph-gallery.com/index.html

Your turn!

Spend 10 minutes reviewing charts at
https://r-graph-gallery.com/index.html

After ten minutes, please be ready to share one example of a plot style that you think would be interesting to use as part of an analysis for your class, work, or this course (and why)!

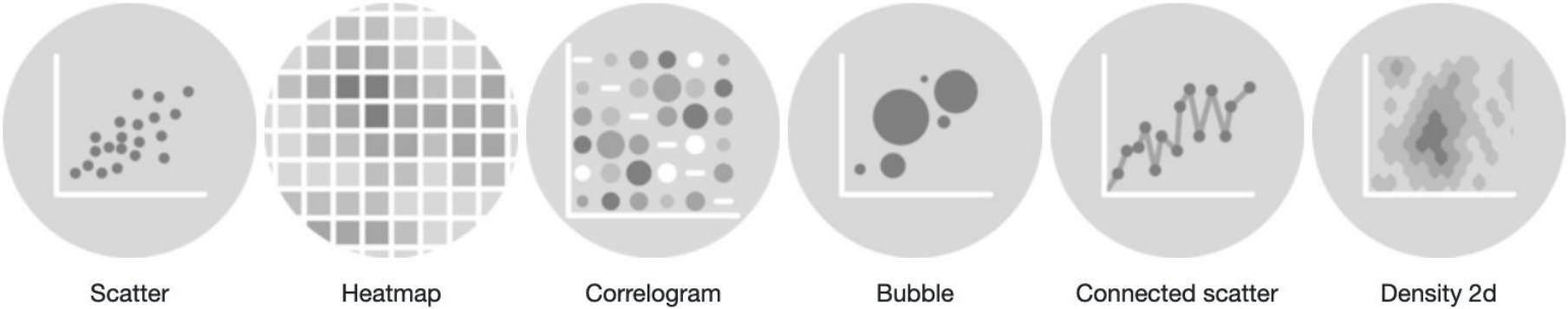
Basic plot styles
distribution

Design

The R Graph Gallery
https://r-graph-gallery.com/index.html

Basic plot styles
correlation

Design



The R Graph Gallery
https://r-graph-gallery.com/index.html

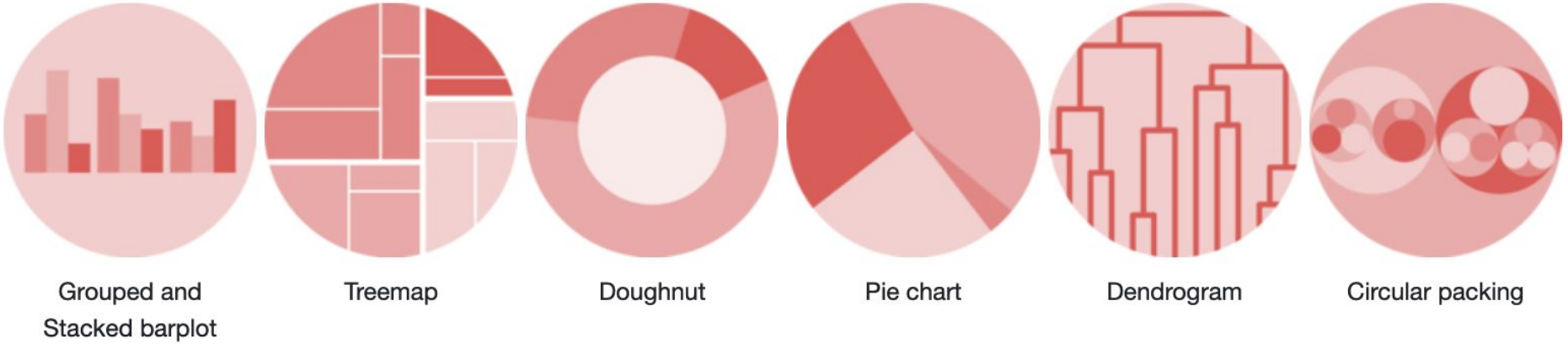
Basic plot styles
ranking

Design

The R Graph Gallery
https://r-graph-gallery.com/index.html

Basic plot styles
part of a whole

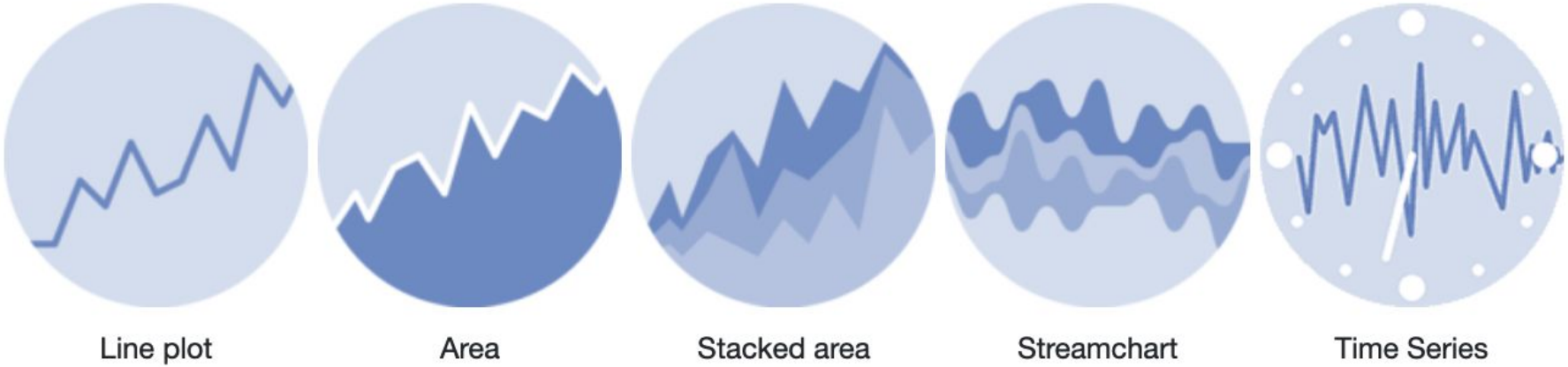
Design



The R Graph Gallery
https://r-graph-gallery.com/index.html

Basic plot styles
change over time

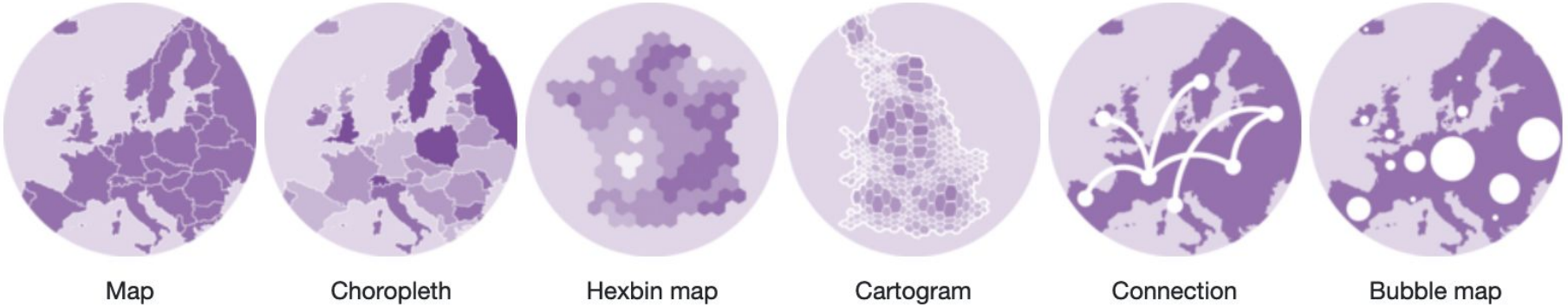
Design



The R Graph Gallery
https://r-graph-gallery.com/index.html

Basic plot styles
maps

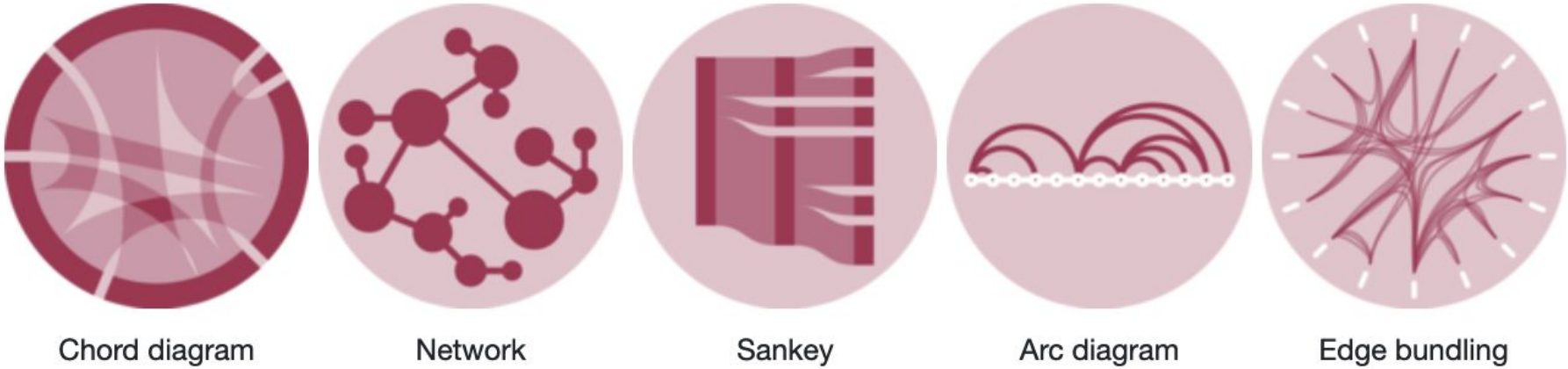
Design



The R Graph Gallery
https://r-graph-gallery.com/index.html

Basic plot styles
networks or flow

Design



The R Graph Gallery
https://r-graph-gallery.com/index.html

Design

Your turn

brainstorming and sketching

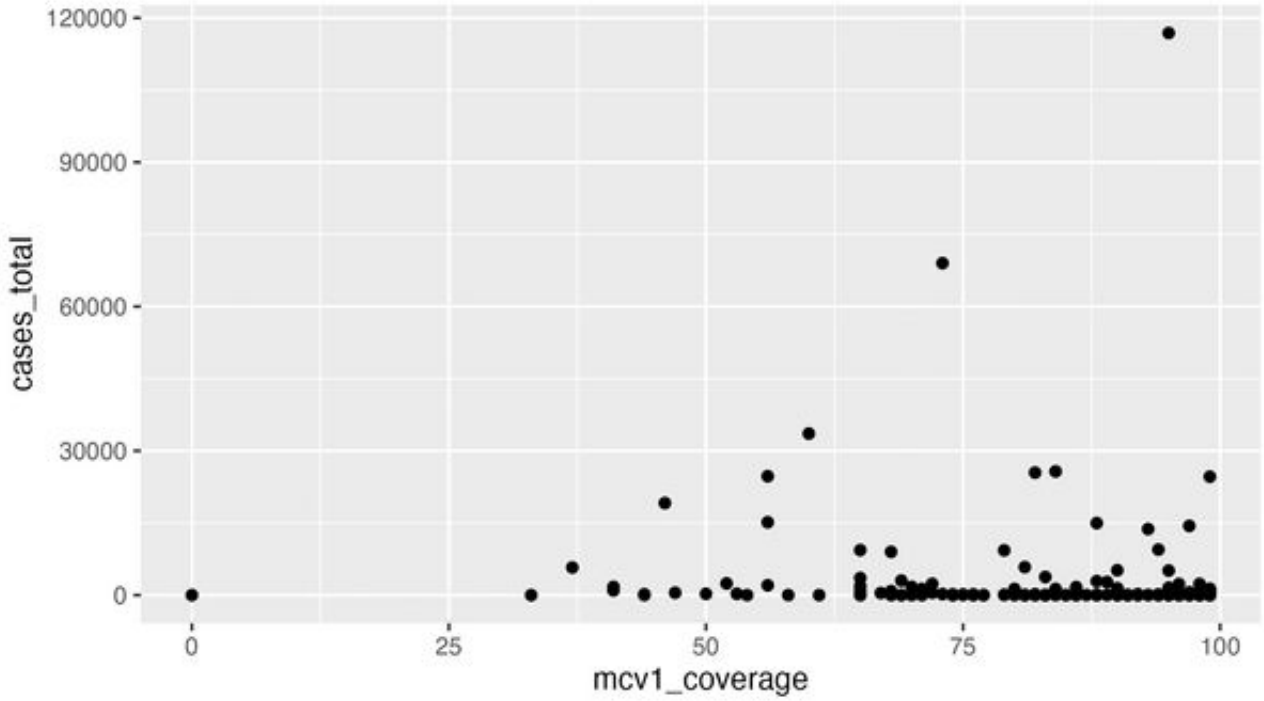
Stages of creating a new data visualization

Define goals and audience

Brainstorm and sketch out ideas

Create your data visualization

Build

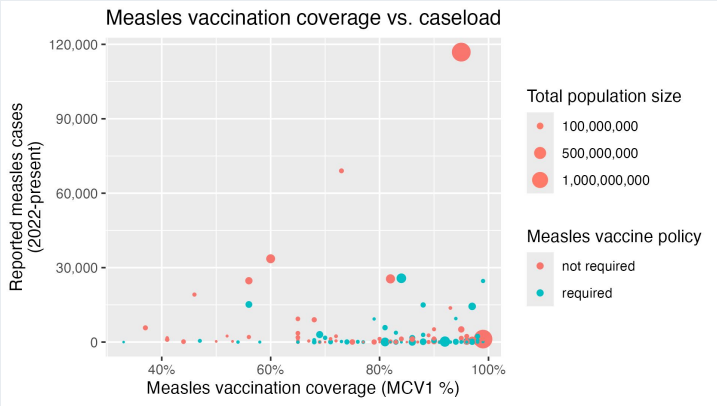

120000
90000
60000
30000
0
cases_total
0
25
50
75
100
mcv1_coverage

Build

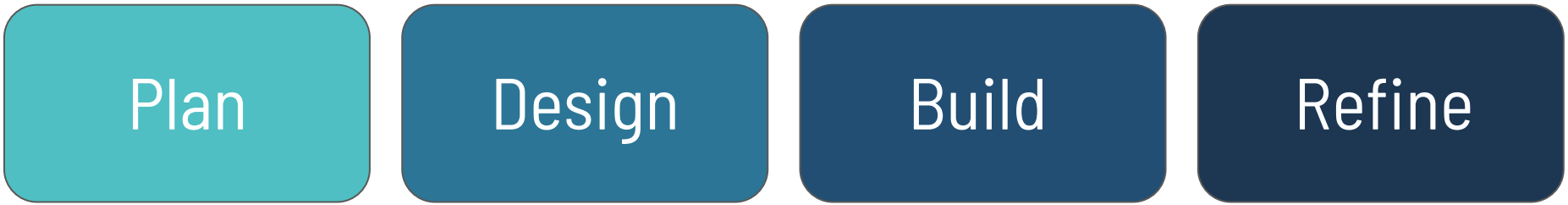
Code

```
coverage_cases %>%
  filter(country_name != "North Korea") %>%
  filter(measles_vaccine_policy != "no data") %>%
  ggplot(aes(x = mcv1_coverage/100,
             y = cases_total,
             color = measles_vaccine_policy,
             size = total_population)) +
  geom_point() +
  labs(title = "Measles vaccination coverage vs. caseload",
       x = "Measles vaccination coverage (MCV1 %)",
       y = "Reported measles cases\n(2022-present)",
       color = "Measles vaccine policy",
       size = "Total population size") +
  scale_x_continuous(labels = scales::percent) +
  scale_y_continuous(labels = scales::comma) +
  scale_size_continuous(labels = scales::comma,
                        breaks = c(100000000, 500000000, 1000000000),
                        range = c(.05, 5)) +
  guides(size = guide_legend(override.aes = list(colour = "#FF7A69"))) +
  theme(legend.position = "right")
```

Figure



Stages of creating a new data visualization



Design

Refine your visualization

Adjust colors, labels, and visual styles

- Use **color** intentionally
- Write clear and relevant **text** with simple **fonts**
- Adjust **sizing and style** based on your audience

Colors

use colors intentionally

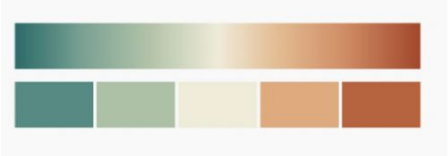
Refine

Sequential

Numbers or categories from low to high
Emphasize *range* of data

example: % of people with internet

Diverging



Scale from negative to positive
Emphasize extreme values

% votes in a two-party system

Unordered categories

Unordered categories
Easily distinguishable colors

gender

Color range graphics and best practices from Lisa Charlotte Muth.
https://blog.datawrapper.de/diverging-vs-sequential-color-scales/#1-Use-diverging-shades-if-there%E2%80%99s-a-meaningful-middle-point

Refine

Colors

consider accessibility

- Utilize **visual contract**, including between text and colors
  - check contrast with coolors.co/contrast-checker/
- Take into account the most common forms of **colorblindness**
  - Red-green color deficiency (reds/green, red/black, yellow/green)
  - Check construct with coolers.co/ (click sunglasses icon)
- Consider **using text**, in addition to color, to convey meaning

Cleveland Clinic Colorblindness.
https://my.clevelandclinic.org/health/diseases/11604-color-blindness

Refine

Colors

borrow from existing palettes

Ask about “design systems” or “brand guidelines”

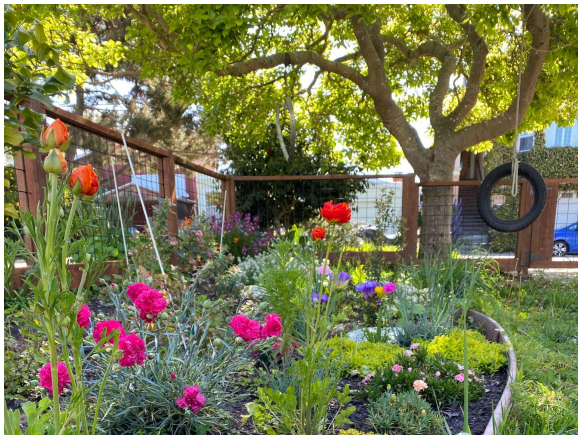
World Health Organization: apps.who.int/gho/data/design-language/design-system/colors/
Global Fund: brandpad.io/the-global-fund-brand-guide-for-partners/
UNAIDS: unaids.org/en/brandbuilder/colour
UNICEF: unicef.github.io/design-system/design-guidelines.html
US HHS: hhs.gov/web/services-and-resources/icon-and-widget-library/index.html
IBM: ibm.com/design/language/color/

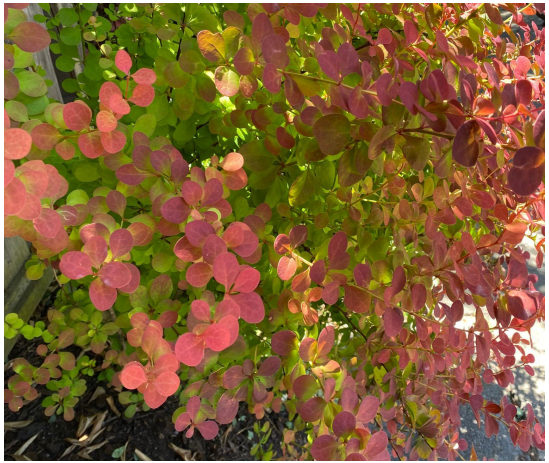
Refine

Colors

borrow from real-life

Canva Color Palette: Find colors in photos you can upload
https://www.canva.com/colors/color-palette-generator/





Refine

Colors

resources for choosing colors

- ColorSpace: Generate various palettes based on one starting color
  https://mycolor.space/

  Chroma.js Color Helper: Generate scales from multiple colors
  https://www.vis4.net/palettes/

- Data Color Picker: Pick equidistant colors from two ends of a scale
  https://www.learnui.design/tools/data-color-picker.html#palette

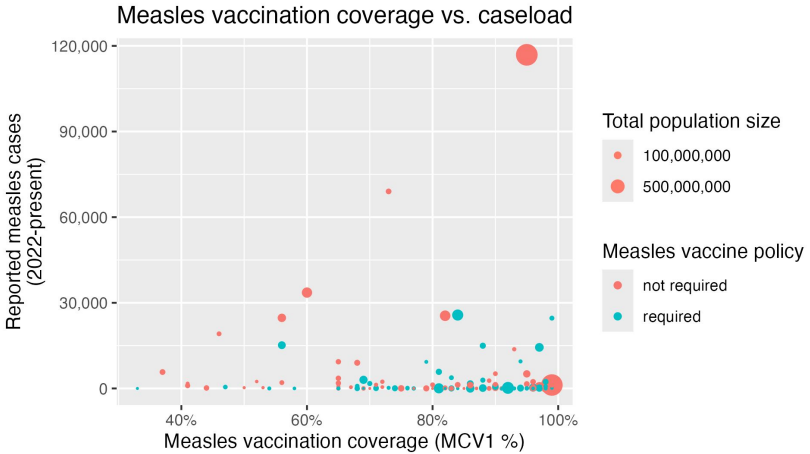
- ColorBrewer: Pre-defined colorblind friendly palettes
  https://colorbrewer2.org/

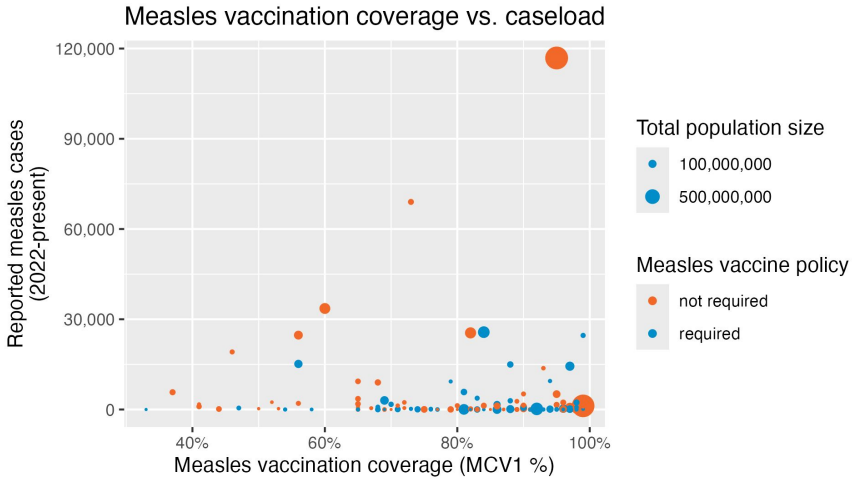
My three most frequently used resources

Colors

refine color selection for our plot

Refine

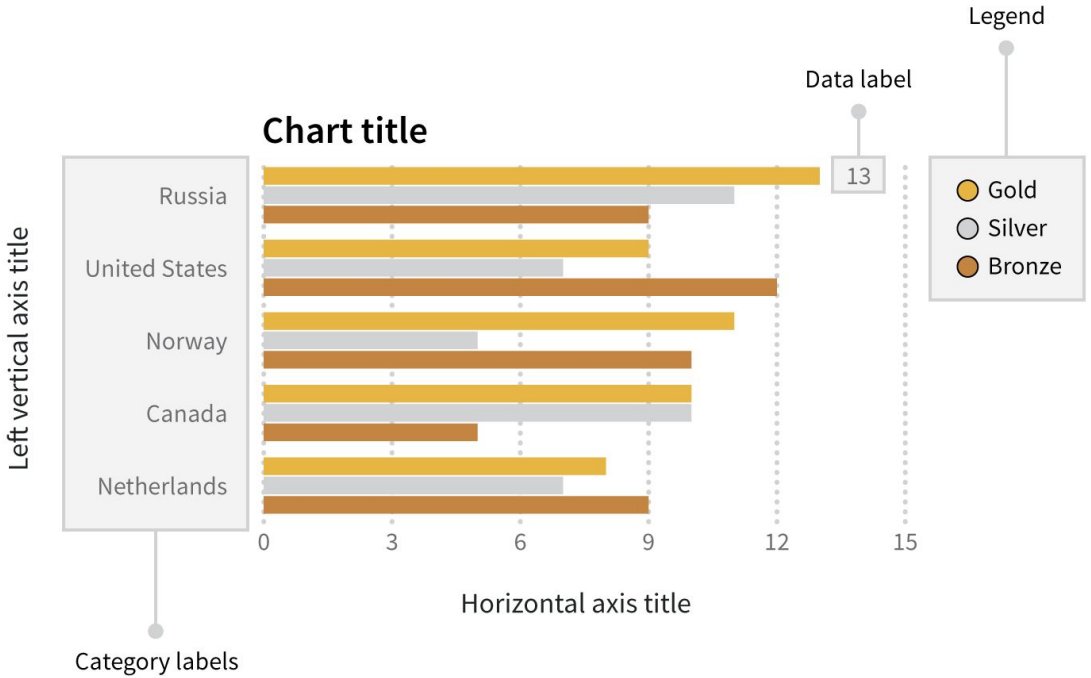




Refine

Text and labels

use clear and informative axis titles



Amanda Makulec, Data + Design Chapter 14. https://trinachi.github.io/data-design-builds/ch14.html

Refine

Text and labels

consider size, style, and positioning

- Which line of text did you read first?
- Even though English readers typically read from top to bottom, left to right, the combination of size, style, and color probably meant that you read “Dracula” first

Refine

Text and labels

highlight key information with text or color

68349014573829476859347589430674930584392056439504832594036 7839

68349014573829476859347589430**6**74930584392056439504832594036 7839

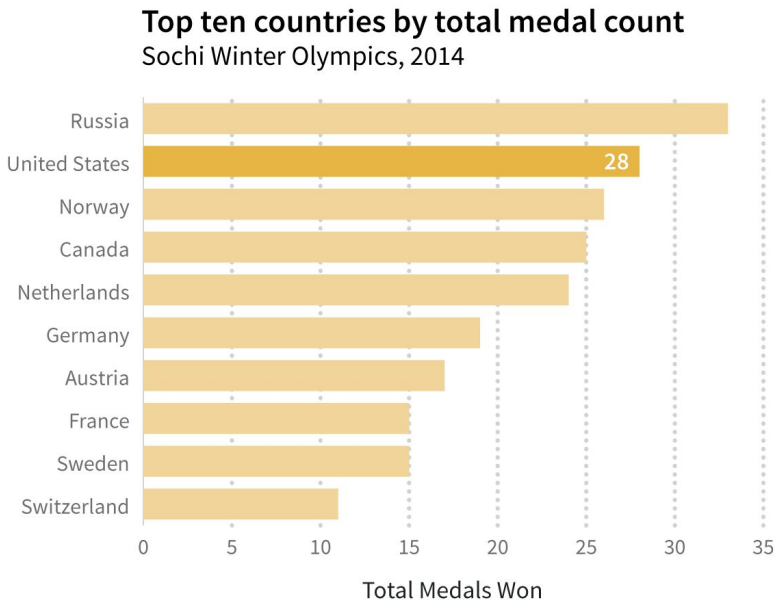
- How long does it take you to find all the “6”s in the top line of text vs. the bottom?
- We can use size (e.g., larger), style (e.g., bold), and color to emphasize results in text labels and in a figure itself

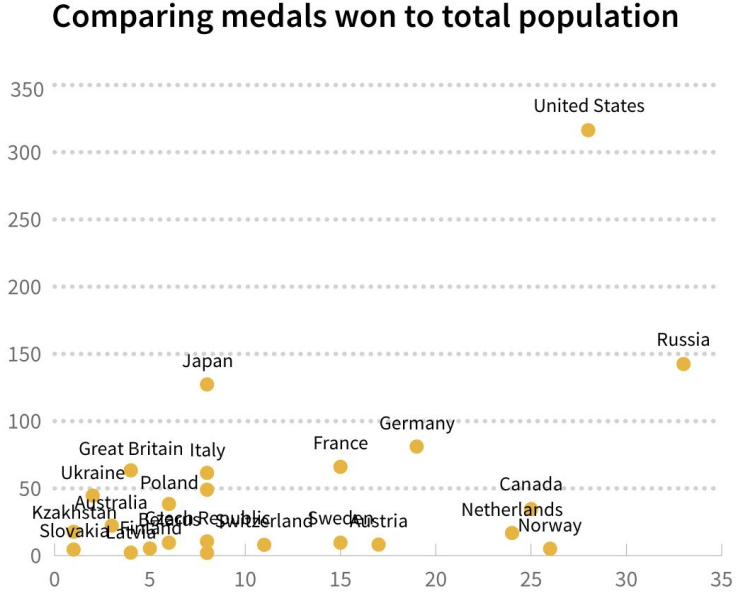
Example from Few S. Show me the numbers. Analytics Pres. 2004.

Text and labels

highlight key information with text or color

Refine



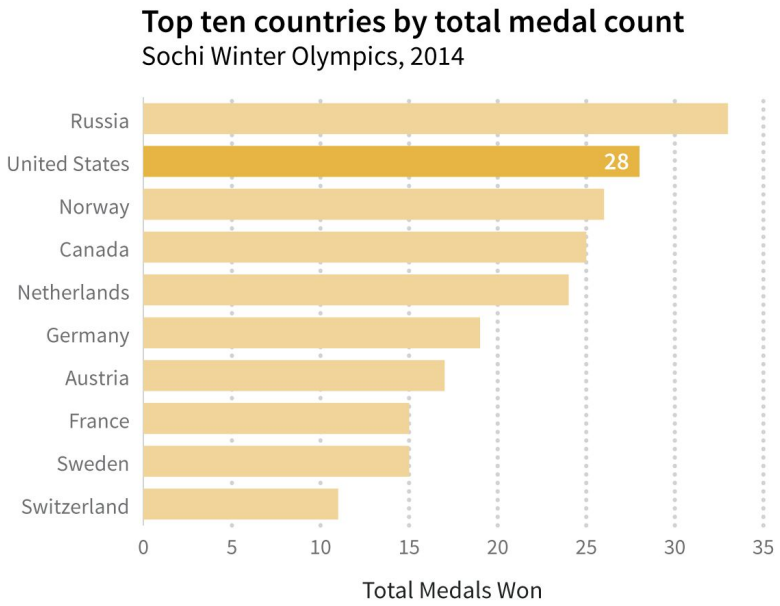


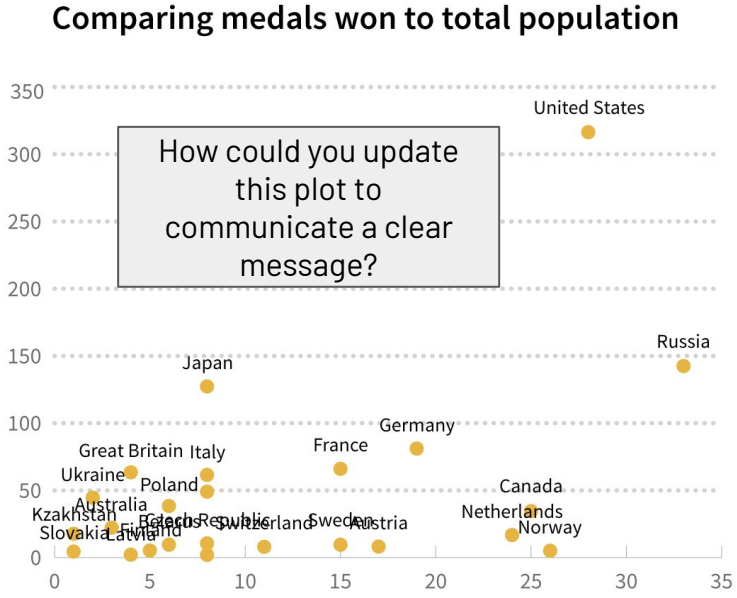
Amanda Makulec, Data + Design Chapter 14. https://trinachi.github.io/data-design-builds/ch14.html

Text and labels

highlight key information with text or color

Refine



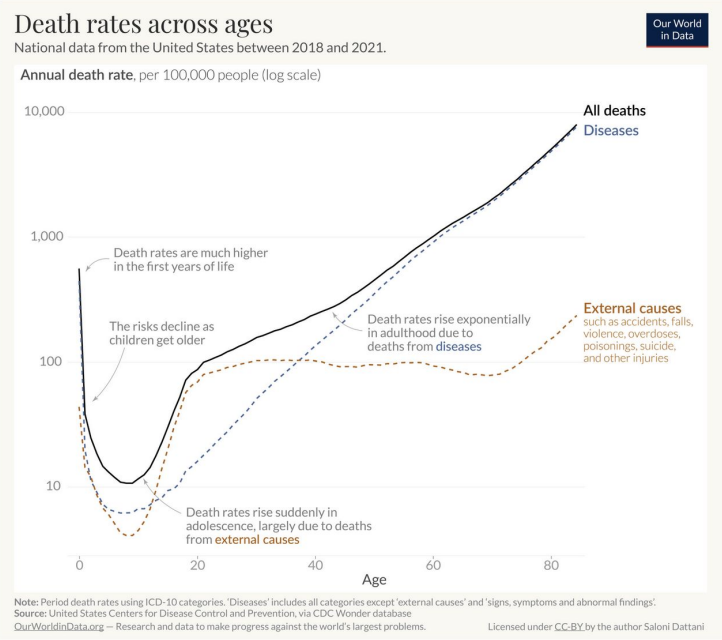


Amanda Makulec, Data + Design Chapter 14. https://trinachi.github.io/data-design-builds/ch14.html

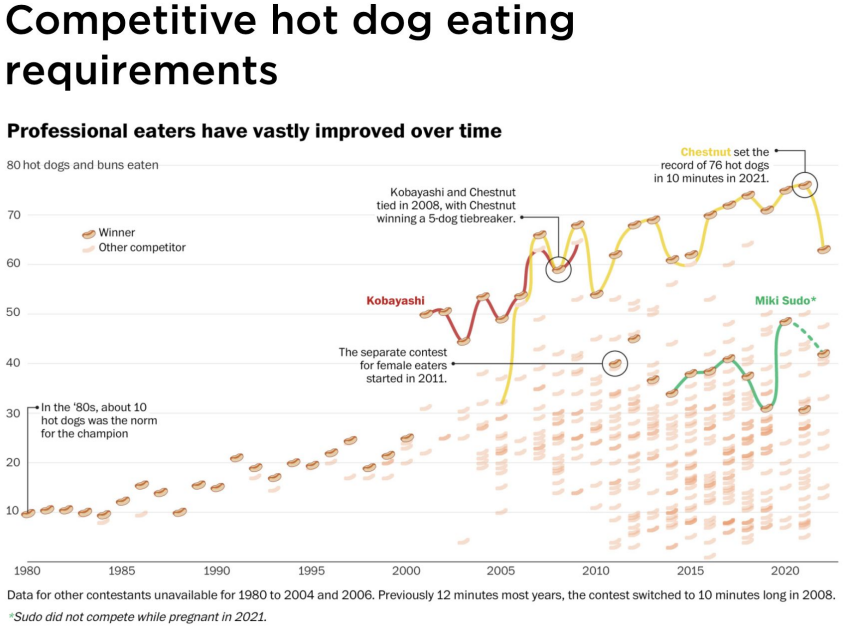
Text and labels

highlight key information with text or color

Refine



Saloni Dattani. Our World in Data.
How does the risk of death change as we age – and how has this changed over time?
https://ourworldindata.org/how-do-the-risks-of-death-change-as-people-age



Washington Post, via Flowing Data.
https://flowingdata.com/2023/07/03/competitive-hot-dog-eating-requirements/

Text and labels

choose a font intentionally

Refine

Serif

A font **with** serifs – lines or marks at the end of a letter's stroke

Examples:

Times
EB Garamond
Merriweather
Source Serif Pro
Lora
Bitter
Crimson Pro

Rule of thumb:
Good for titles, stylized fonts, or to convey emotion

Sans Serif

A font **without** ("sans") serifs on the letters, with clearly defined edges of each letter

Examples:

Open Sans
Roboto
Poppins
Noto Sans
Work Sans
Epilogue
Barlow

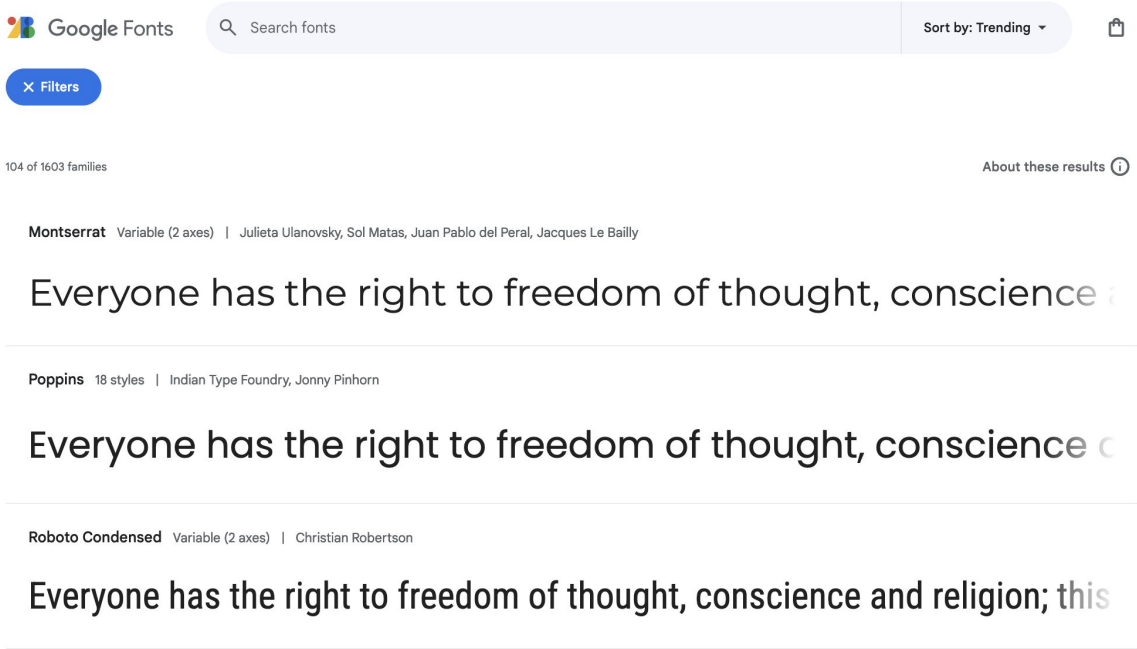
Rule of thumb:
Good for plot labels, projecting on a screen with lower resolution

Google Fonts Knowledge Glossary
https://fonts.google.com/knowledge/glossary/sans_serif

Text and labels

explore different font options

Refine

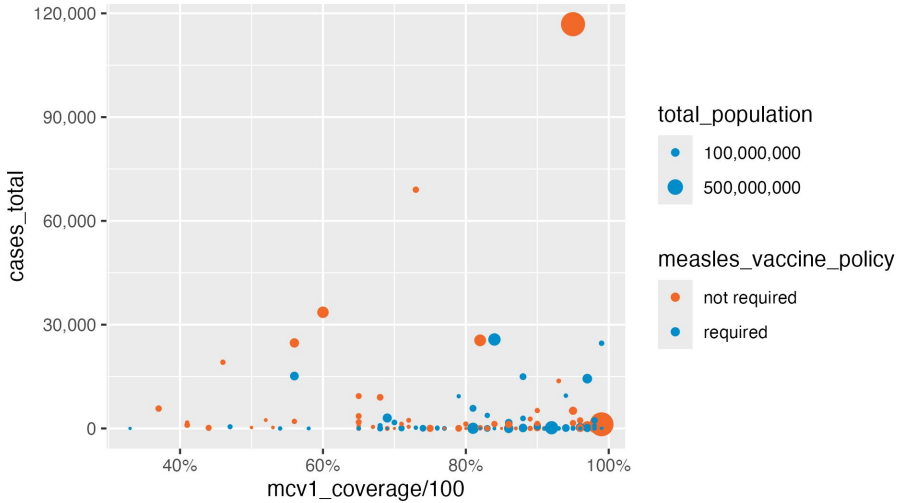


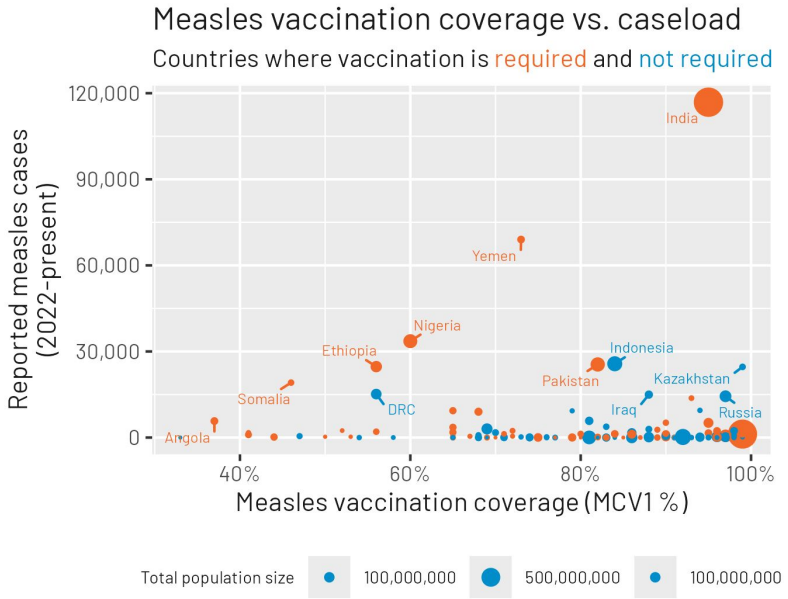
Google Fonts.
https://fonts.google.com/?stylecount=14

Text and labels

refine text selection for our plot

Refine

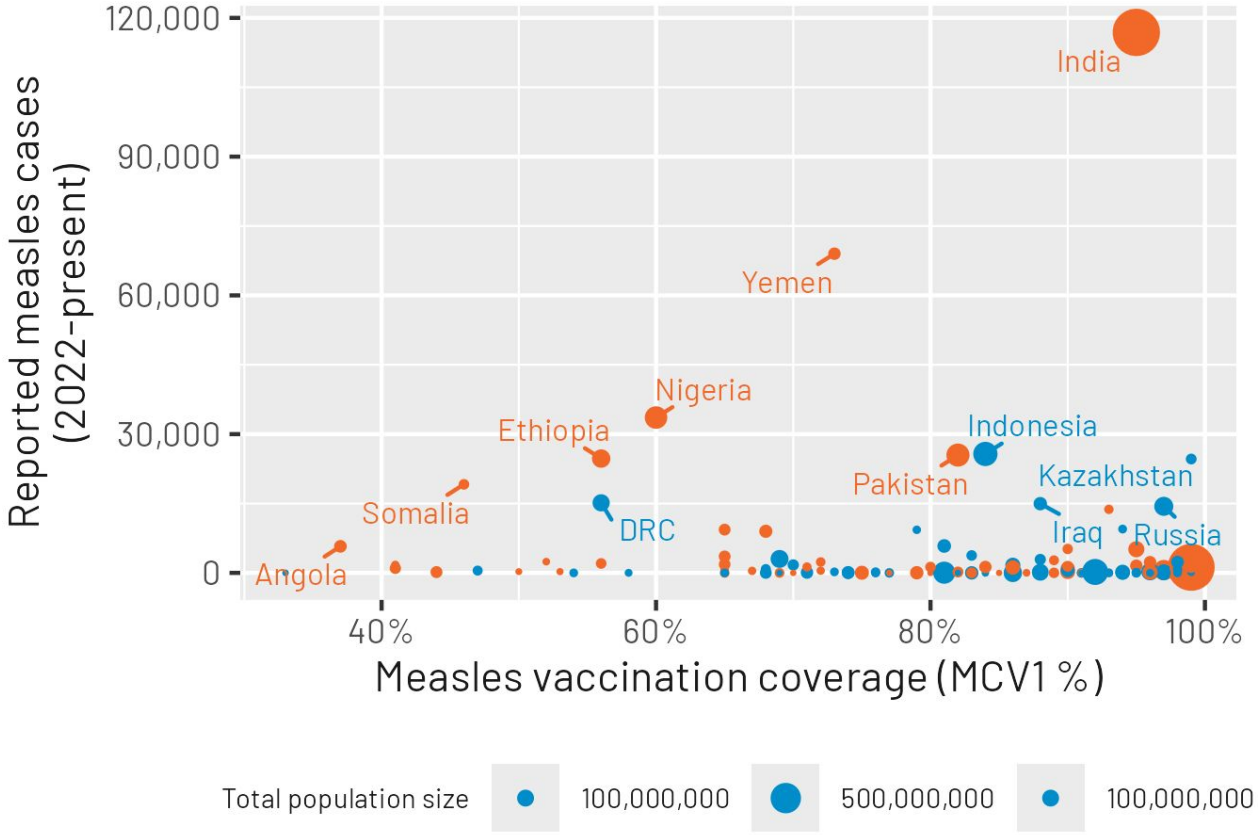




Measles vaccination coverage vs. caseload

Countries where vaccination is required and not required

Refine



OPTIONAL homework

review some beautiful and creative data visualizations

- One of the most helpful ways to improve my data visualization skills is to learn from the amazing work that other people are already doing.
    - Cara Thompson's portfolio
- What are your favorite visualizations (from this list or elsewhere)? Why?
- What do they communicate?

OPTIONAL reading
learn more about **colors**

Lisa Charlotte Muth. A detailed guide to colors in data vis style guides
https://blog.datawrapper.de/colors-for-data-vis-style-guides/

Cara Thompson. Five tips for creating bespoke colour schemes
https://www.cararthompson.com/talks/nhsr2022-palatable-palettes/

Lisa Charlotte Muth. When to use sequential and diverging color scales
https://blog.datawrapper.de/diverging-vs-sequential-color-scales/

Color palettes and accessibility features for data visualization
medium.com/carbondesign/color-palettes-and-accessibility-features-for-data-visualization-7869f4874fca

OPTIONAL reading
learn more about **fonts**

Elliot Jay Stocks. Making sense of typographic classifications
fonts.google.com/knowledge/introducing_type/making_sense_of_typographic_classifications

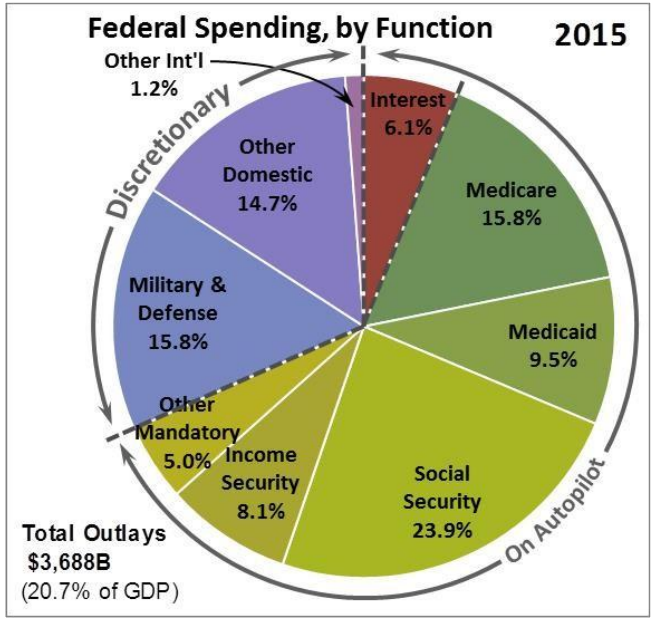
Elliot Jay Stocks. Pairing typefaces
onts.google.com/knowledge/choosing_type/pairing_typefaces

WCAG2 Accessibility by Design
wcag2.com/accessible-typography-and-style/

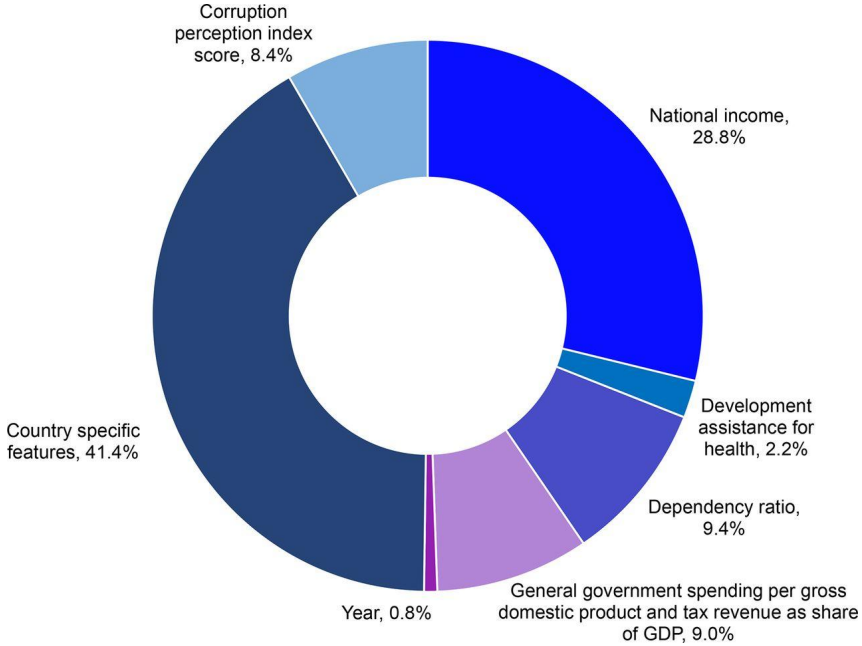
Appendix

Plot styles

pie or donut charts

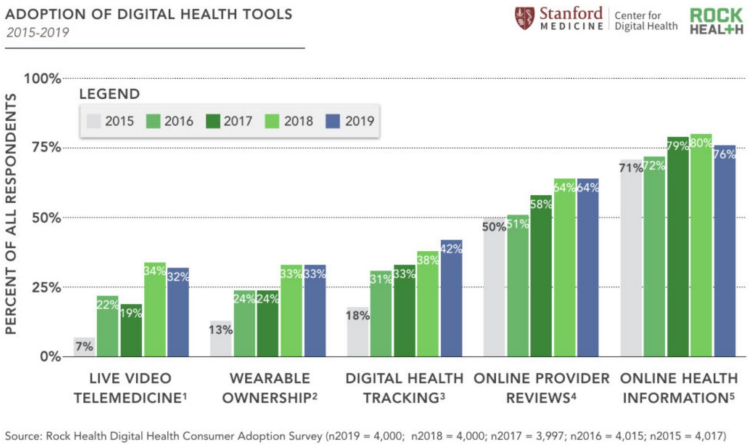


Source: Randall Bolten
https://www.linkedin.com/pulse/last-pie-chart-actually-says-something-important-randall-bolten



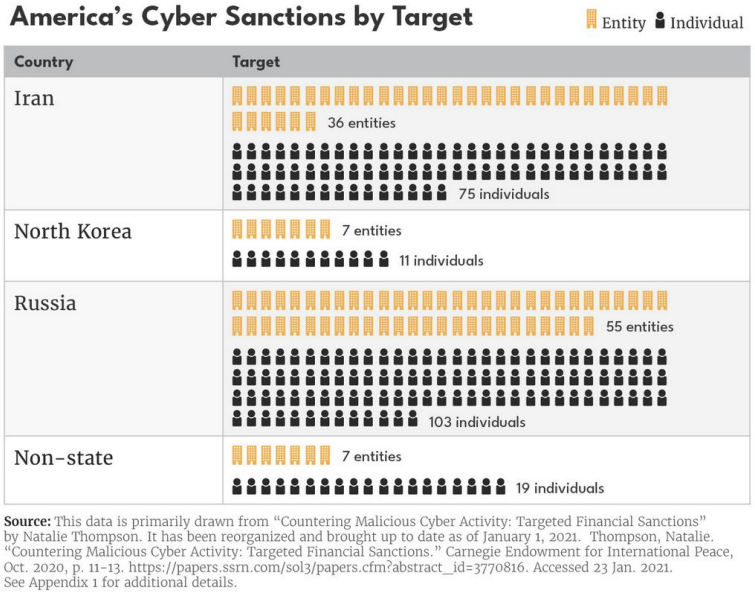
Source: Micah AE, Chen CS, Zlavog BS, Hashimi G, Chapin A, Dieleman JL. Trends and drivers of government health spending in sub-Saharan Africa, 1995–2015. BMJ global health. 2019 Jan 1;4(1):e001159.

Plot styles
barcharts

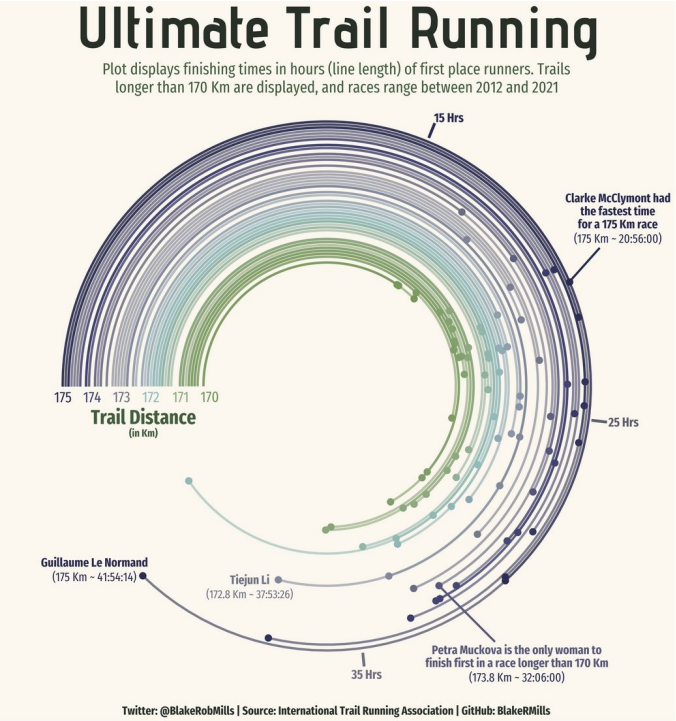


Source: Rock Health
https://rockhealth.com/insights/digital-health-consumer-adoption-report-2019/

Plot styles
variations of barcharts



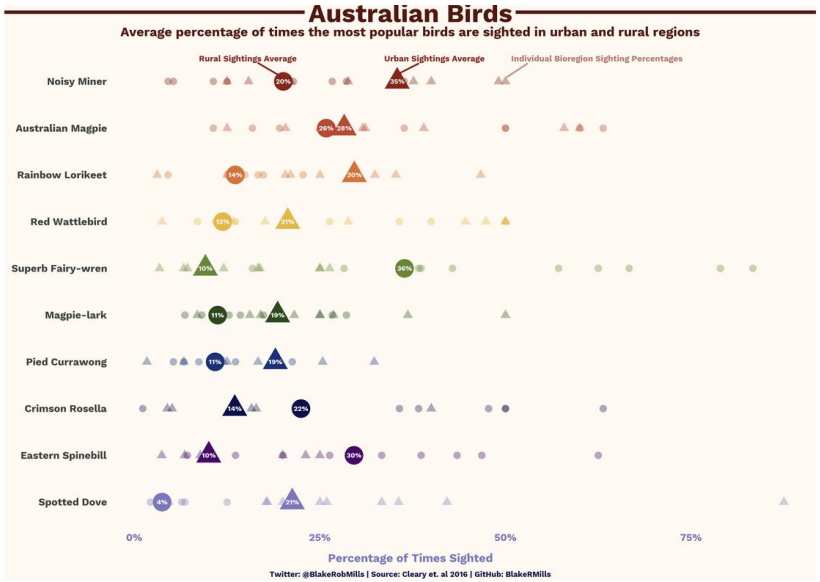
Source: Third Way
https://www.thirdway.org/memo/unpacking-us-cyber-sanctions



Source: Blake Mills, TidyTuesday
https://github.com/BlakeRMills/TidyTuesday

Plot styles

dot plots

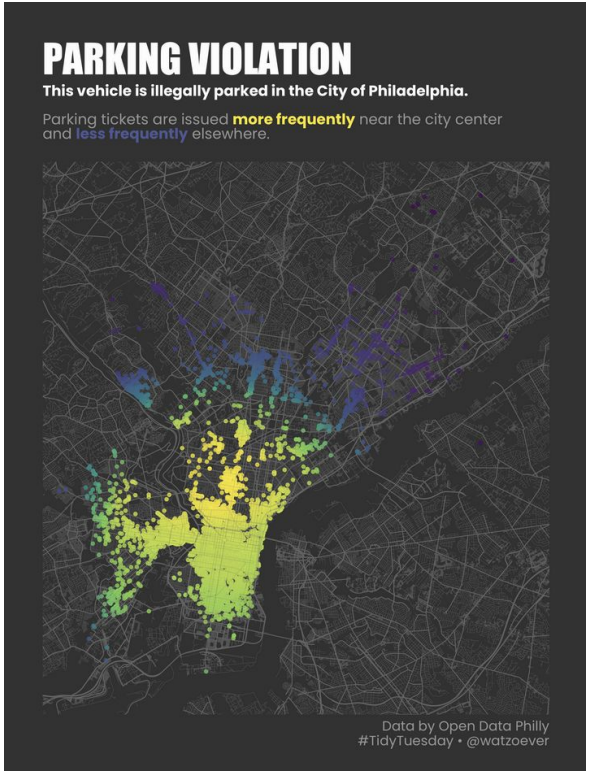


Source: Blake Mills, TidyTuesday
https://github.com/BlakeRMills/TidyTuesday

Plot styles
heatmaps

	Cassidy-Graham	Skinny repeal	Partial repeal	Repeal and replace	House-passed plan
Individual mandate	Repeal	Repeal	Repeal	Repeal	Repeal
Employer mandate	Repeal	Repeal	Repeal	Repeal	Repeal
Subsidies for out-of-pocket costs	Repeal	Keep	Repeal	Repeal	Repeal
Tax credits for premiums	Repeal	Keep	Repeal	Change	Change
Medicaid expansion	Repeal	Keep	Repeal	Change	Change
Essential health benefits	Up to states	Keep	Keep	Up to states	Up to states
Prohibitions on annual and lifetime limits	Up to states	Keep	Keep	Up to states	Up to states
Pre-existing conditions policy	Up to states	Keep	Keep	Up to states	Up to states
Restrictions on charging more for older Americans	Up to states	Keep	Keep	Up to states	Up to states
Taxes created under Obamacare	Change	Change	Change	Change	Repeal
Health savings account	Change	Change	Change	Change	Change
Dependent coverage until 26	Keep	Keep	Keep	Keep	Keep
Vote results	**Expected next week**	**Failed 49-51**	**Failed 45-55**	**Failed 43-57**	**Passed 217-213**
Increase in the number of uninsured in 10 years	**No score**	**16 million**	**32 million**	**No score**	**23 million**

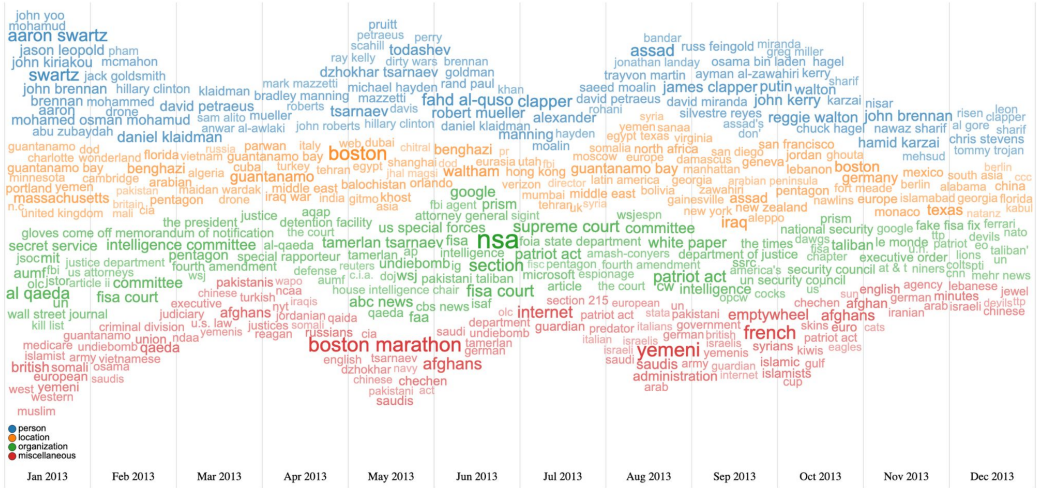
Source: New York Times
https://www.nytimes.com/interactive/2017/09/22/us/republican-health-plan-comparison.html



Source: Julia Watzek TidyTuesday
https://github.com/jwatzek/tidytuesday

Plot styles
word clouds

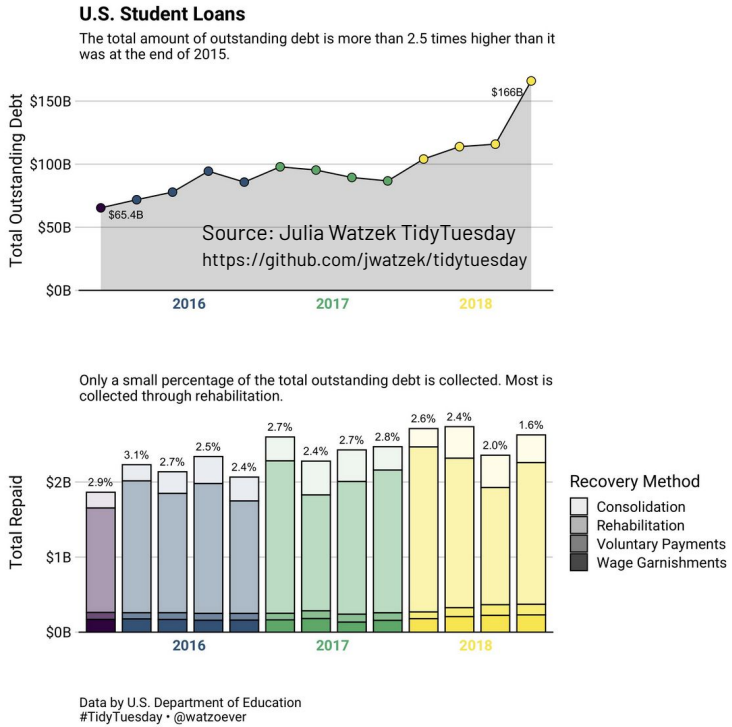
Design

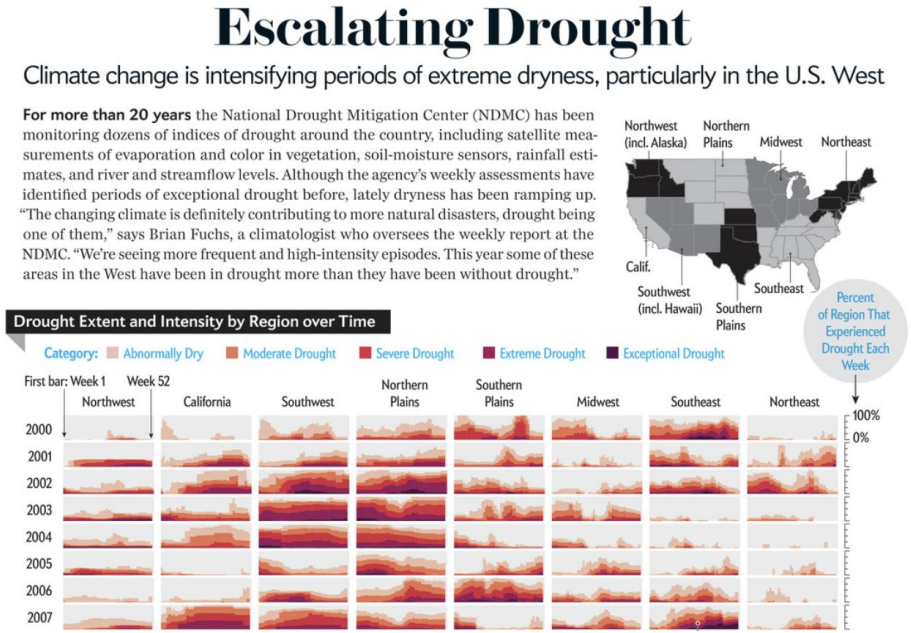


Source: WordStream
https://idatavisualizationlab.github.io/WordStream/examples.html?

Plot styles

combining plot styles

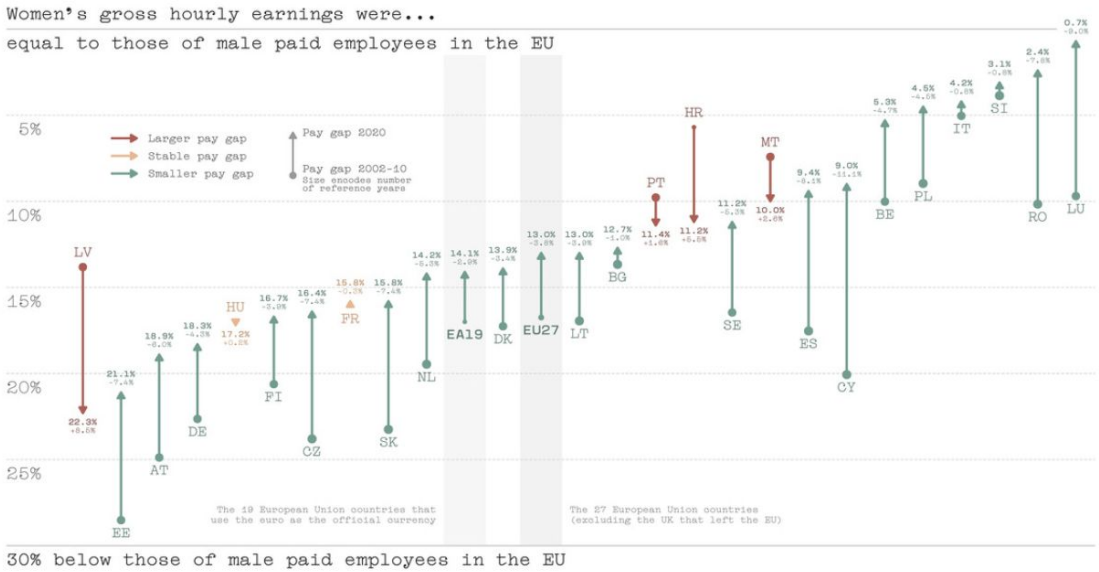




Source: Cedric Scherer
https://www.cedricscherer.com/top/dataviz/

Other plot styles

Cleveland dot plot



Source: Cedric Scherer
https://www.cedricscherer.com/top/dataviz/

Get the details right

Axes matter

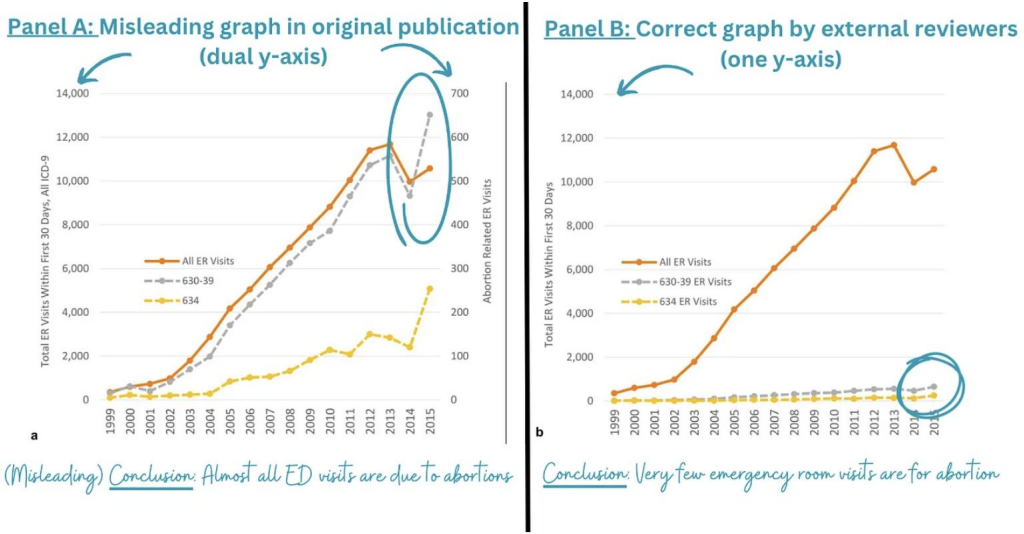


Figure Source: Upadhyay et al, 2024; Annotations by YLE

Source: Your Local Epidemiologist Substack.
https://yourlocalepidemiologist.substack.com/p/two-retracted-studies-at-the-supreme