Data Visualization I

Designing effective charts

Simon A. Queenborough
Yale School of the Environment

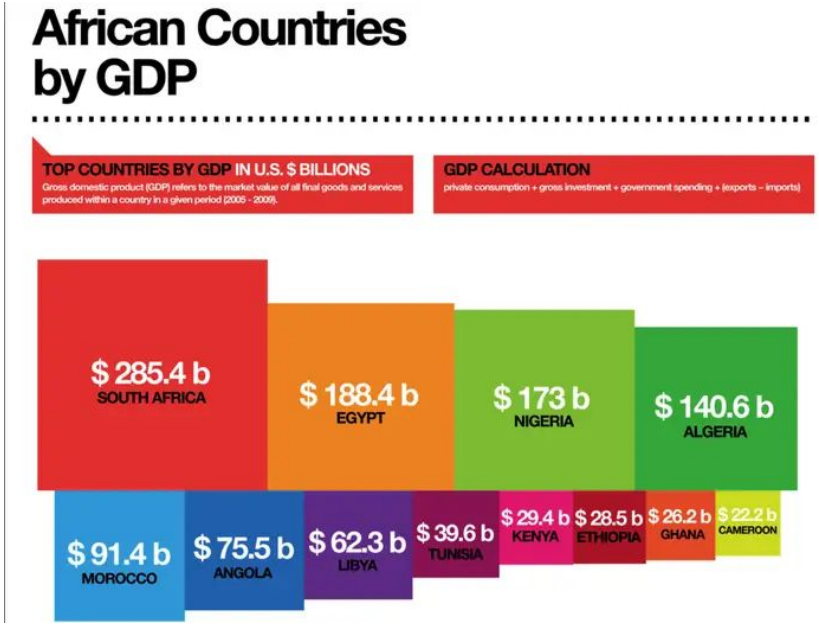
ELI CWA 303(d) Workshop 2020-06-09

Which of these two graphs do you prefer and why?

Mentimeter: Which graphic do you prefer and why?

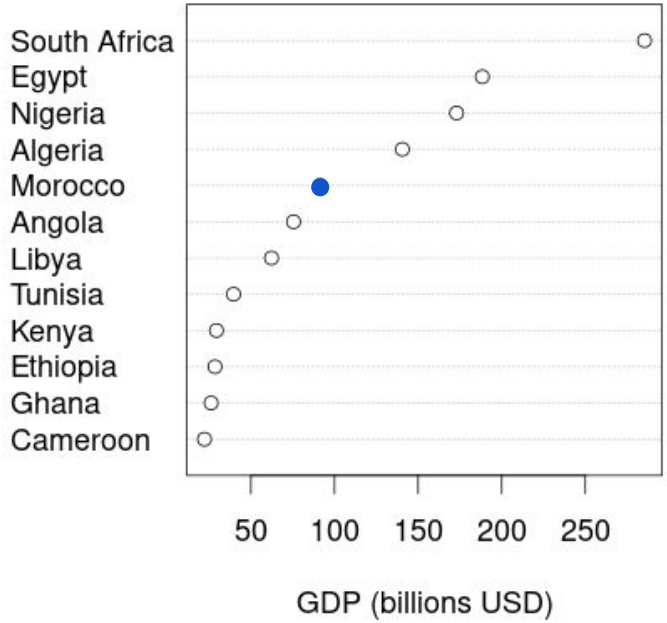
Scan here to comment

A



https://visual.ly/community/Infographics/economy/african-countries-gdp

B



Learning outcomes

Graphics is communication

Strengths and challenges of human perception

Three principles of effective communication:

- Have a clear purpose
- Show the data clearly
- Make the message obvious

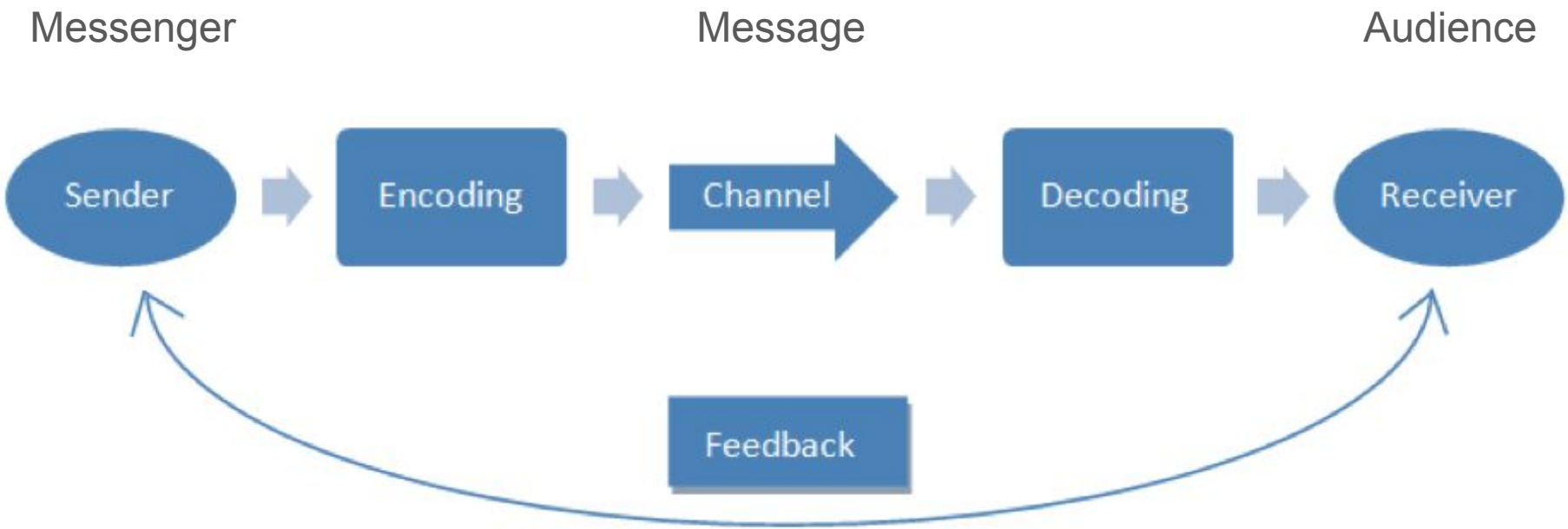
Apply knowledge in two mini-makeovers

The goal of graphics is **communication**

"A graph is **more effective** than another if its quantitative information can be **decoded** more **quickly** or more **easily** by most observers."

Nancy Robbins, *Creating More Effective Graphs*

The communication process



Graphics can speed up knowledge acquisition by:

A. **Playing to strengths of human brain function:**
   I. Pre-attentive visual processing
   II. Pattern recognition

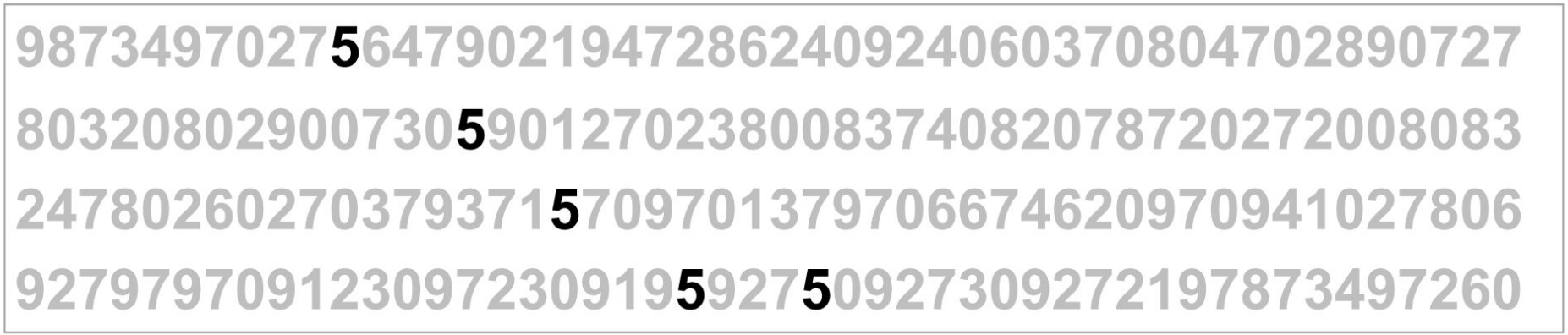
B. **Accounting for challenges of human perception**
   I. Perceive relative differences (not absolutes)
   II. Assign meaning depending on context (e.g., red and blue in US vs UK)
   III. Variation in ability (e.g., color vision deficiency)

A. Perception is sometimes **SLOW** …

9873497027564790219472862409240603708047028907278032080290073059012702380083740820787202720080832478026027037937157097013797066746209709410278069279797091230972309195927509273092721978734972607

How many times does **‘5’** appear above?

… and sometimes **FAST!**



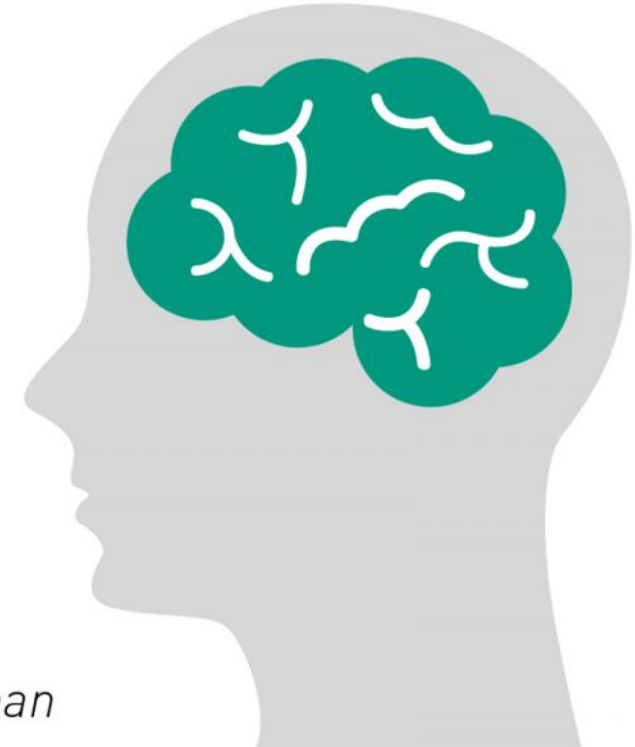
How many times does **‘5’** appear above?

SYSTEM 1

Intuition & instinct

95%

Unconscious
Fast
Associative
Automatic pilot



SYSTEM 2

Rational thinking

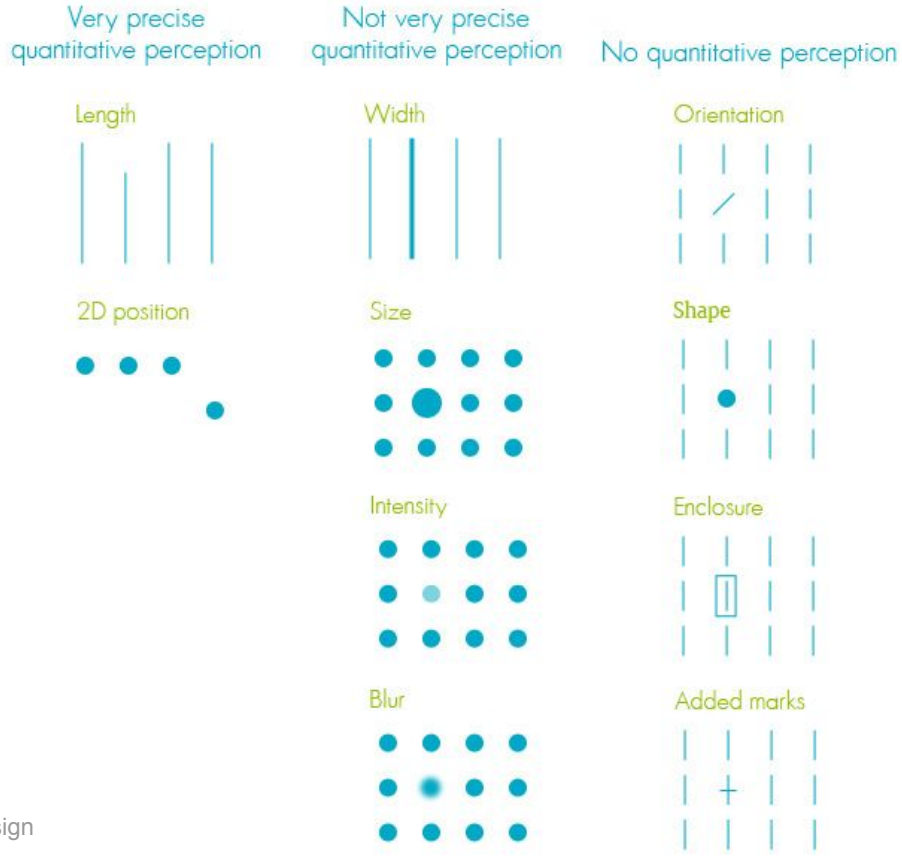
5%

Takes effort
Slow
Logical
Lazy
Indecisive

Source: Daniel Kahneman

A.i. Take advantage of pre-attentive visual processing

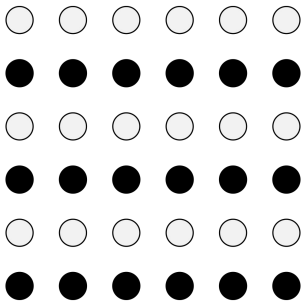
A.i. Take advantage of pre-attentive visual processing



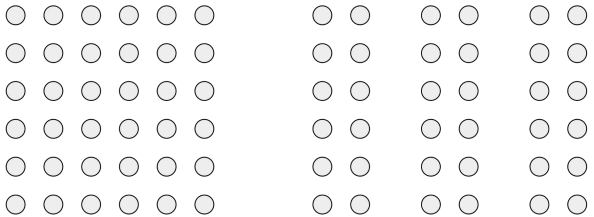
Stephen Few's Information Dashboard Design

A.ii. Humans seek patterns

SIMILARITY
Objects that **share similar attributes** (e.g., color or shape) are perceived as a group.



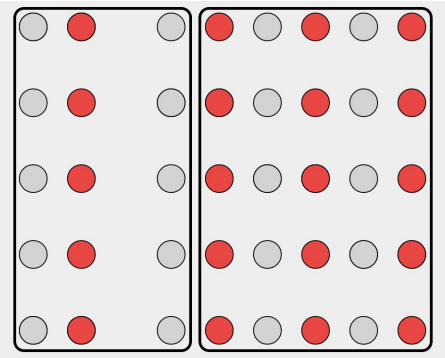
PROXIMITY
Objects that are **close together** are perceived as a group



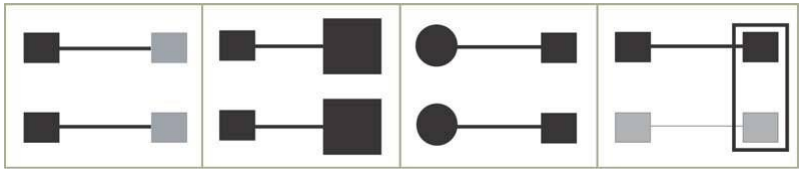
Wikipedia

A.ii. Humans seek patterns

ENCLOSURE
Objects that appear to have a **boundary around them** (e.g., formed by a line or area of common color) are perceived as a group.



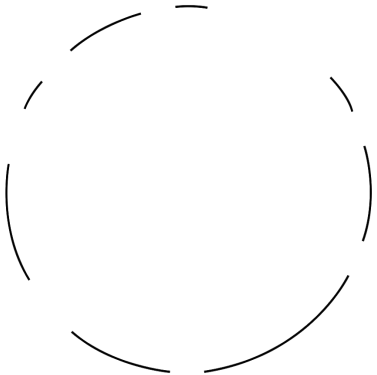
CONNECTION
Objects that are **connected** (e.g., by a line) are perceived as a group.



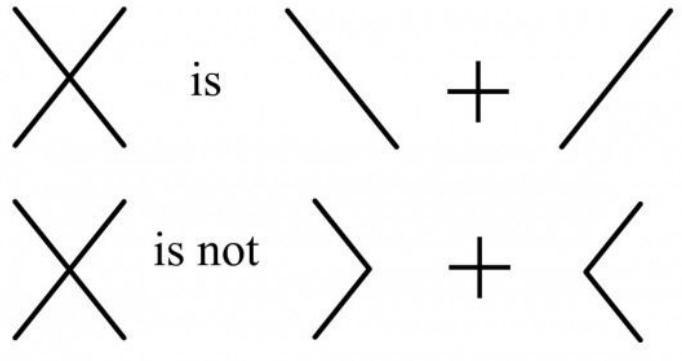
https://emeeks.github.io/gestaltdataviz/section1.html

A.ii. Humans seek patterns

CLOSURE
Open structures are perceived as closed, complete, and regular whenever there is a way that they can be reasonably interpreted as such.



CONTINUITY
Objects that are **aligned** together or appear to be a **continuation** of one another are perceived as a group.



Wikipedia

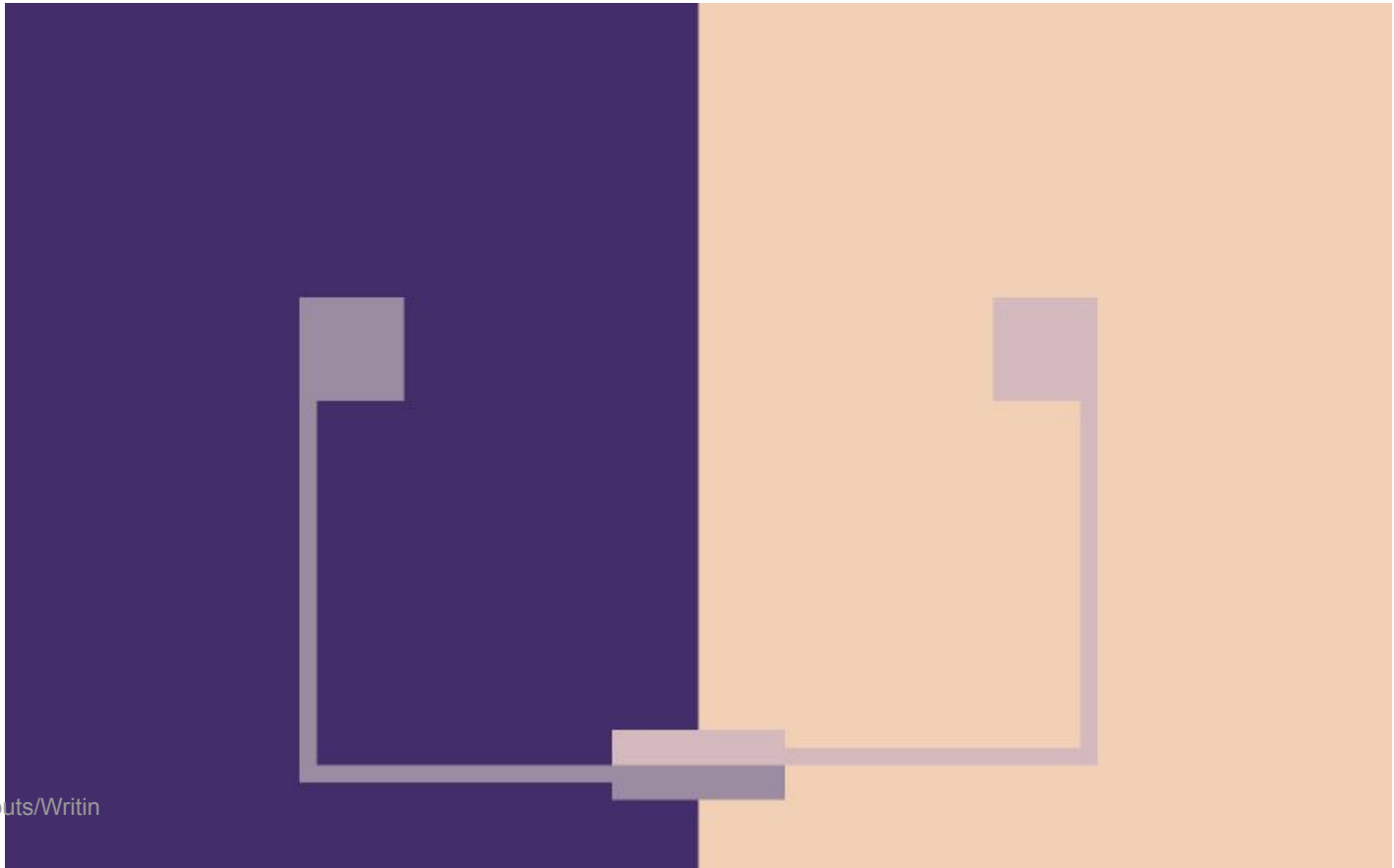
B.i. We perceive **relative difference**

Simultaneous contrast can make the **same** colors look **different**

http://www.engr.colostate.edu/ECE666/Handouts/WritingPapers/UsingColorEffectively.pdf

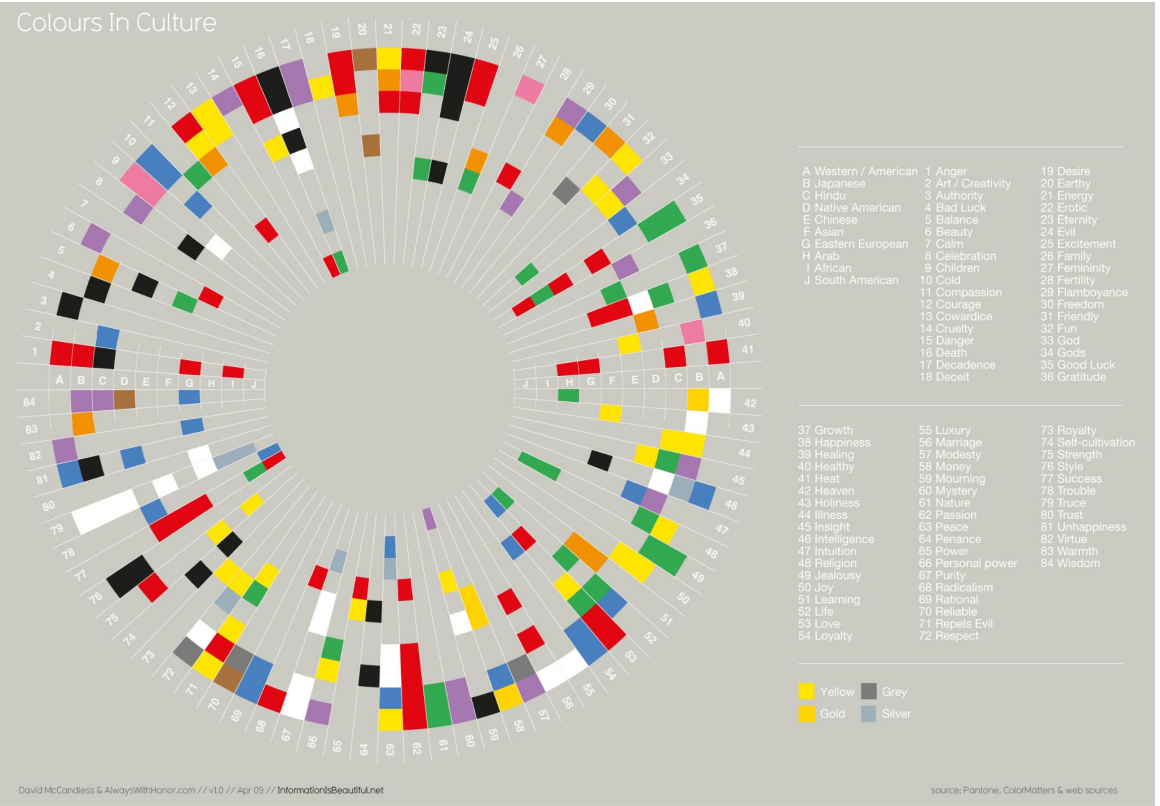
B.i. We perceive **relative difference**

Simultaneous contrast can make **different** colors look the **same**



http://www.engr.colostate.edu/ECE666/Handouts/WritingPapers/UsingColorEffectively.pdf

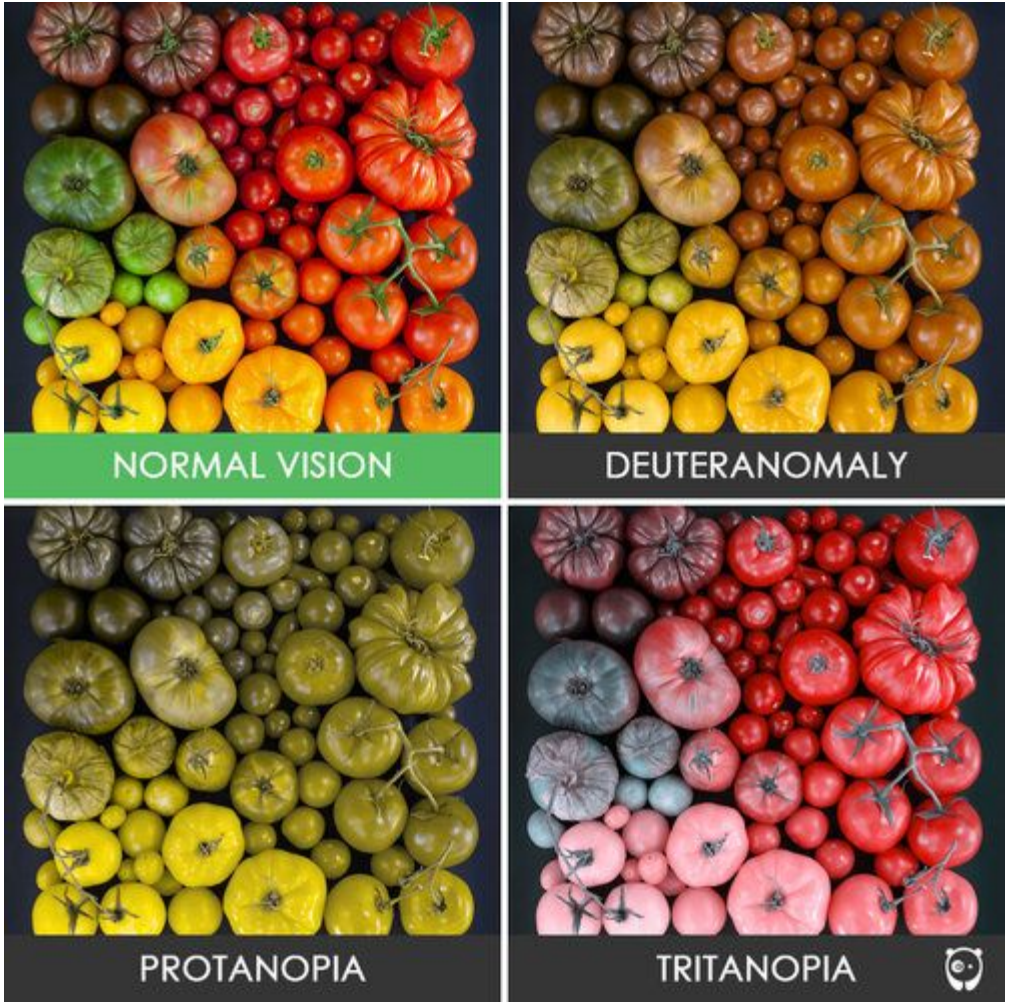
B.ii. Understand that **meaning** depends on **context**



https://www.informationisbeautiful.net/visualizations/colours-in-cultures/

B.iii. Account for **variation in perception**

Red–green color blindness affects up to **8% of males** and **0.5% of females** of Northern European descent.



https://www.storytellingwithdata.com/blog/2018/4/23/accessibility-considerations-for-visuals

THREE PRINCIPLES OF EFFECTIVE COMMUNICATION

1. Have a clear purpose
2. Show the data clearly
3. Make the message obvious

THREE PRINCIPLES OF EFFECTIVE COMMUNICATION

1. **Have a clear purpose**
2. Show the data clearly
3. Make the message obvious

1. Have a clear purpose

a) Understand the question you are trying to answer

b) Identify the quantitative evidence to answer that question

c) Know your audience and focus the design to support their needs

1a. Understand the question you are trying to answer

- Data exploration,
- Convey information,
- Deliver a message,
- Convince an audience,
- Support a decision?
- Any / all of the above .. ?

1b. Identify quantitative evidence to answer that question

Collect new data

Find existing data

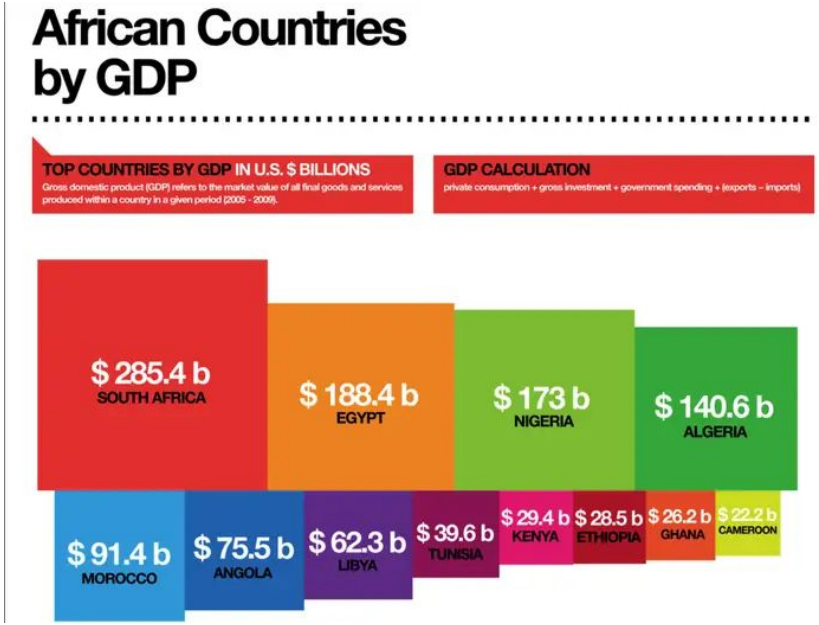
…

1c. Know your audience; focus design to support them

The audience is **not** you!

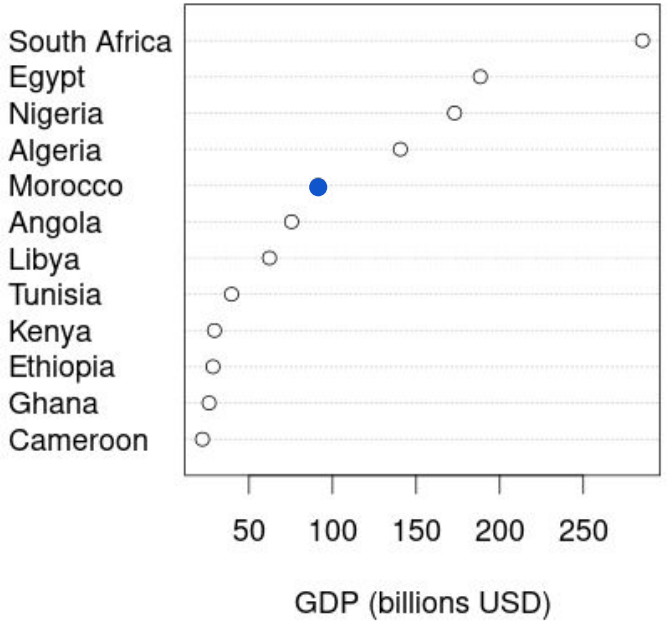
1c. Know your audience; focus design to support them

A. Public

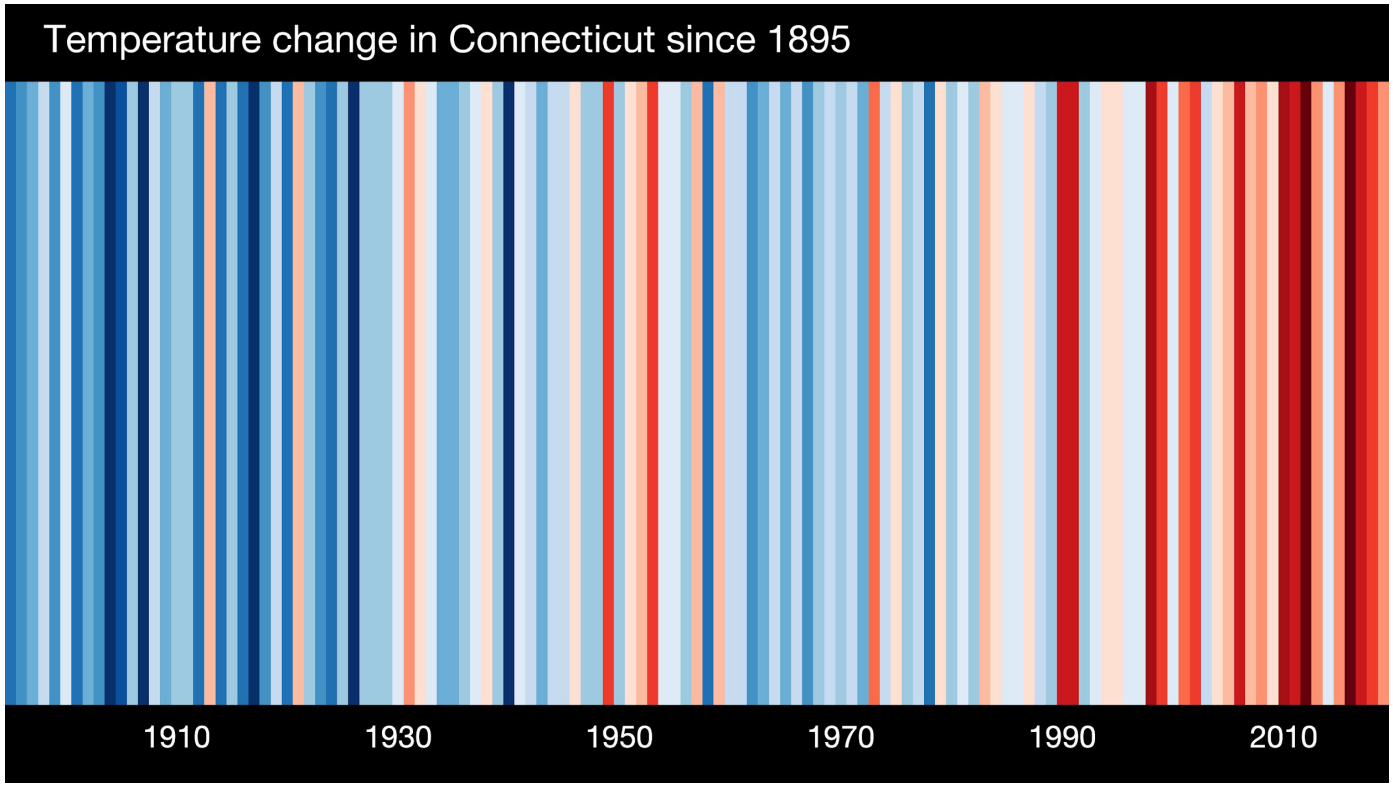


https://visual.ly/community/Infographics/economy/african-countries-gdp

B. Academic



Makeover #1



Mentimeter: Worst or best graphic ever … ? and why?

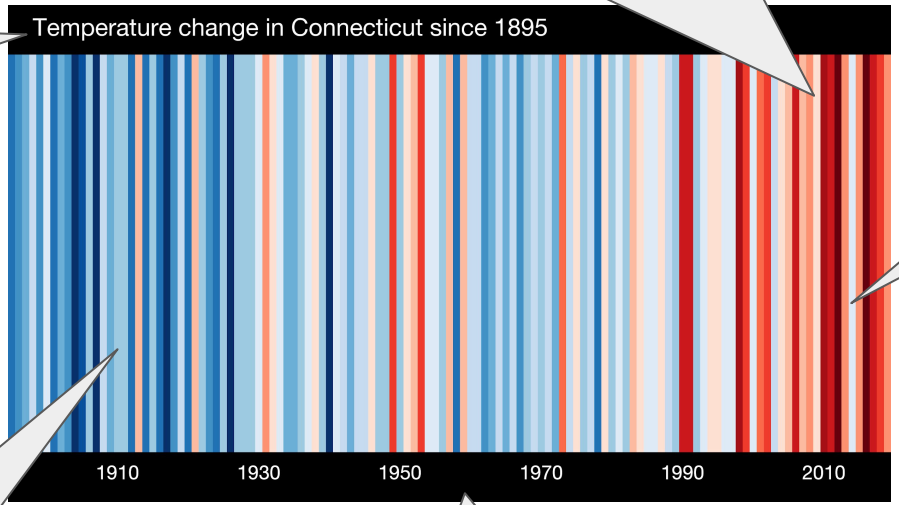
(1) Worst graph ever

(10) Most amazing graph ever

[GOOD] Communication to wide audience

Message: More red recently = hotter

Title: Clear, but no message

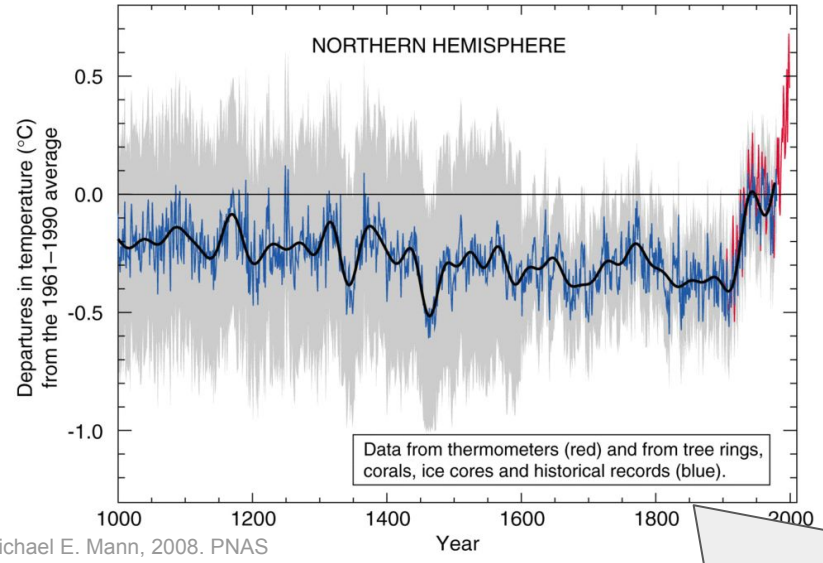


Color: Harder to make direct quantitative comparisons

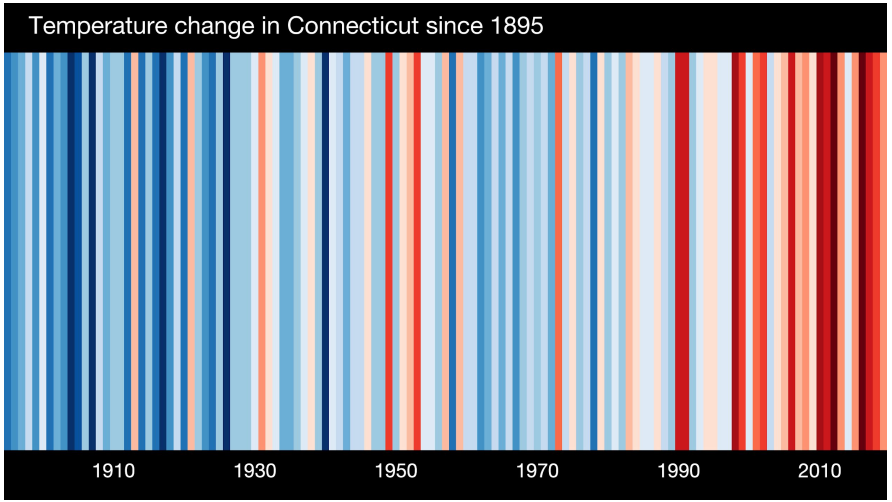
Good use of **diverging color scheme**

X axis:
- Clear annual scale

Compare scientific journal graphic with public-facing graphic



Michael E. Mann, 2008. PNAS



- Shows data on 2 axes
- Shows uncertainty
- Describes data sources
- Longer times series, maybe harder to understand pace of change?

THREE PRINCIPLES OF EFFECTIVE COMMUNICATION

1. Have a clear purpose
2. **Show the data clearly**
3. Make the message obvious

2. Show the data clearly

a) Choose the appropriate graph type to display your data

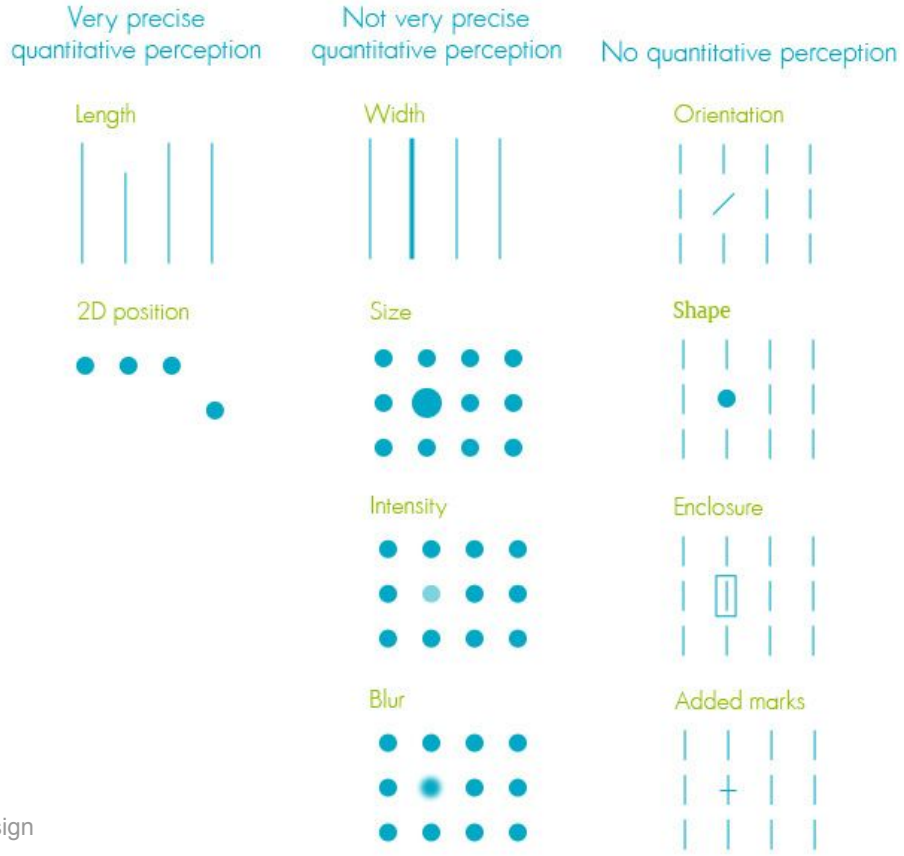
b) Avoid misrepresentation (use appropriate scales)

c) Maximize data to ink ratio (reduce distraction, less is more)

Encoding numeric vs categorical data

NUMERIC data

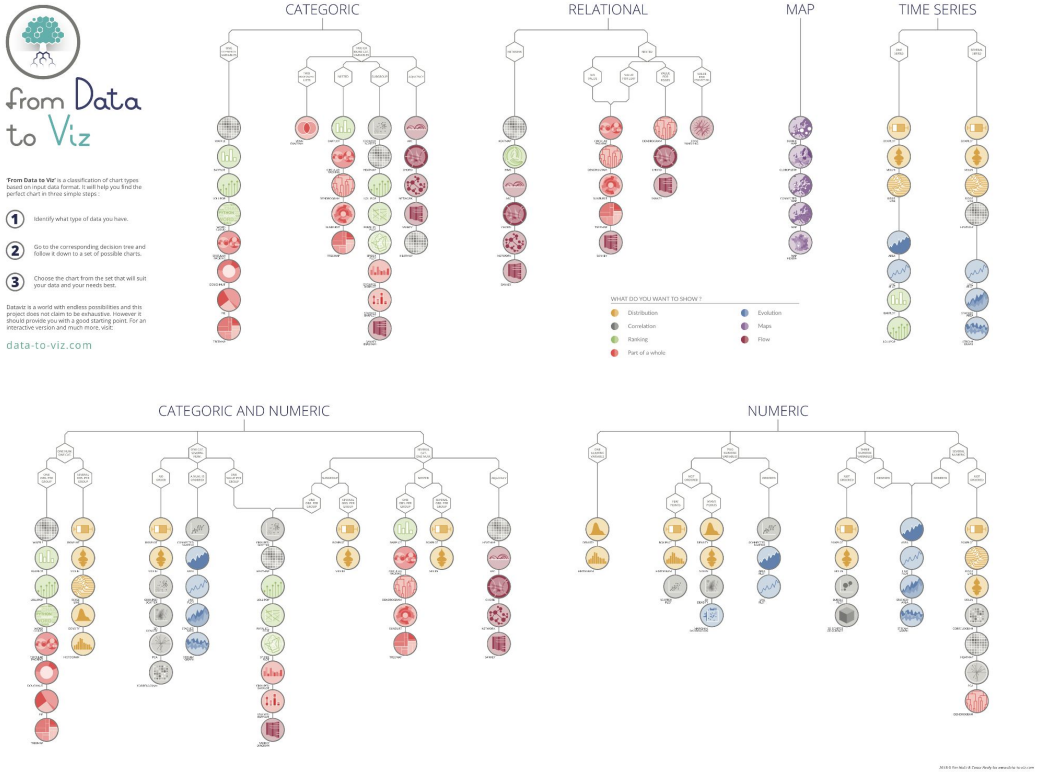
Points
Lines
Bars
Color



CATEGORICAL data

Shape/symbol
Line type
Color

2a. Choose appropriate graph type to display data



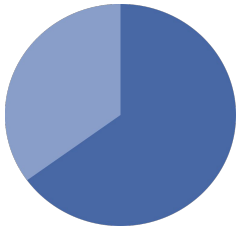
TOOL TIP!

https://www.data-to-viz.com/

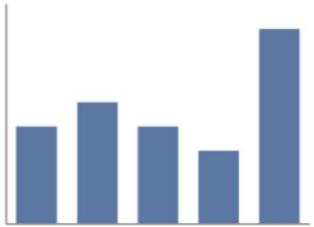
2a. Choose appropriate graph type to display data

TOOL TIP!

https://clauswilke.com/dataviz/directory-of-visualizations.html

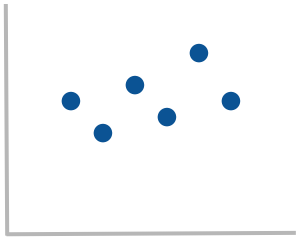


Pie



Vertical bar

Line



Scatter

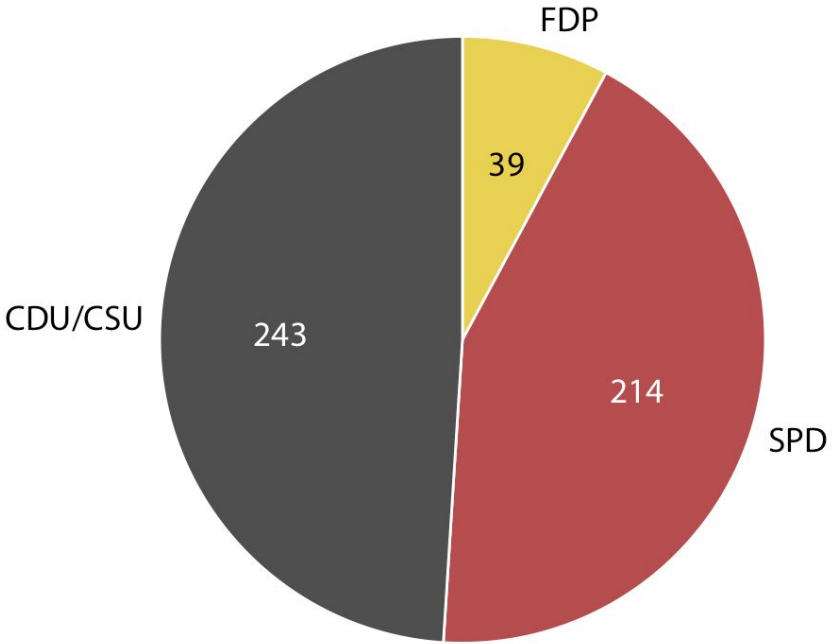
2a.i. Pie charts

Good if you have only few (2, 3) groups

Parts sum to whole (100%)

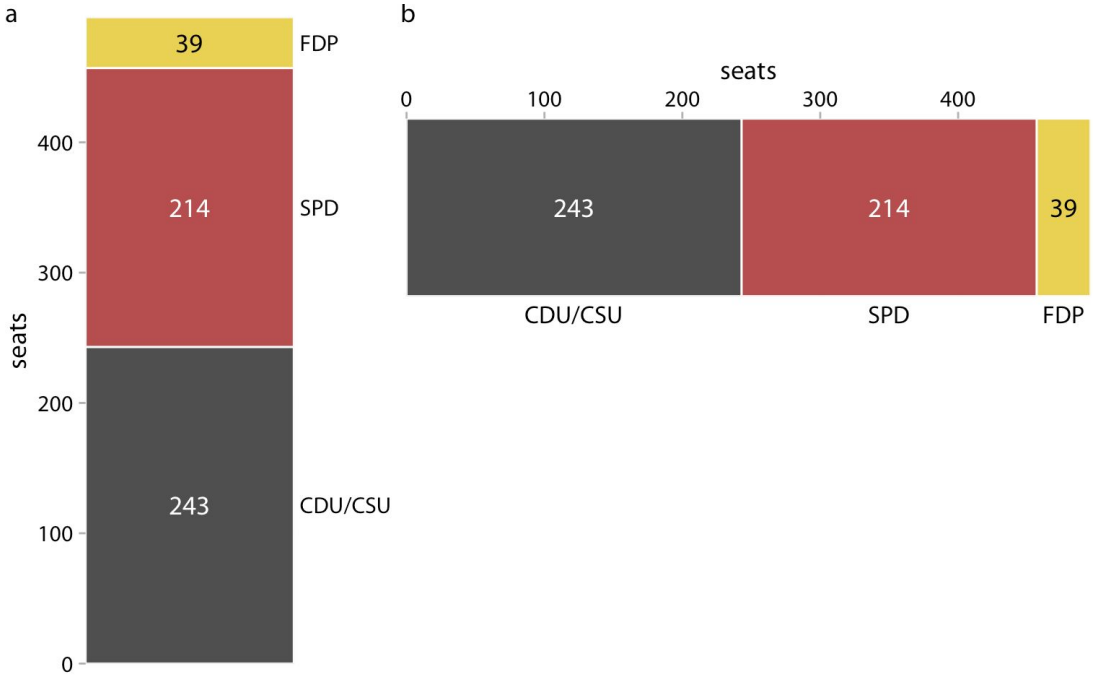
Start at 12 noon

Order parts



Party composition of the 8th German Bundestag, 1976–1980, visualized as a pie chart. Source: https://clauswilke.com/dataviz/

Stacked bar plots may be better than pies



Party composition of the 8th German Bundestag, 1976–1980, visualized as stacked bar charts. Source: https://clauswilke.com/dataviz/

2a.ii. Bar charts ...

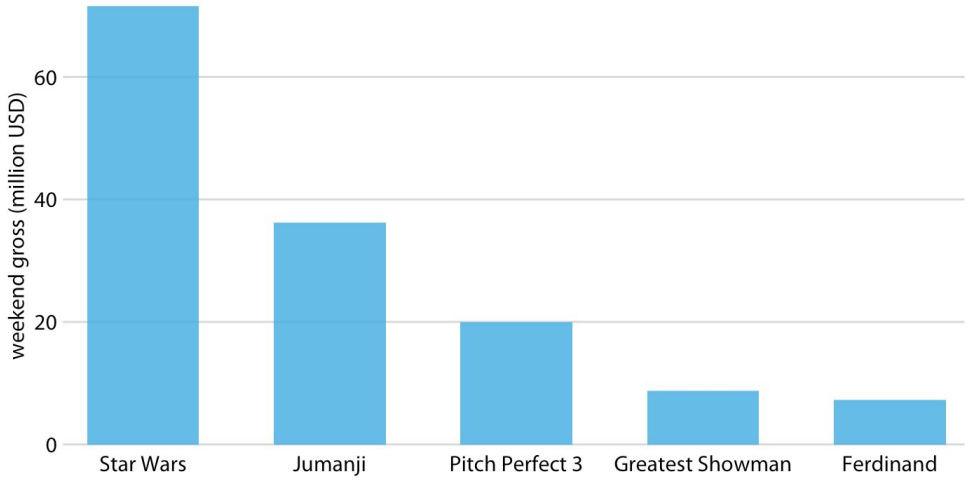
Rank bars by the same attribute

Baseline = 0

Wide bars (>2x white space)

Same color

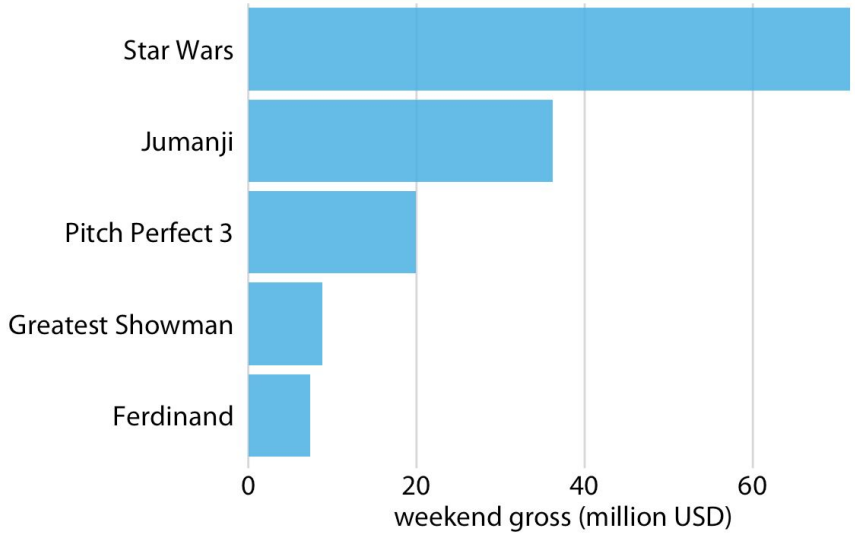
Horizontal text ...



Highest grossing movies for the weekend of December 22-24, 2017, displayed as a bar plot. Source: https://clauswilke.com/dataviz/

… Horizontal bar charts

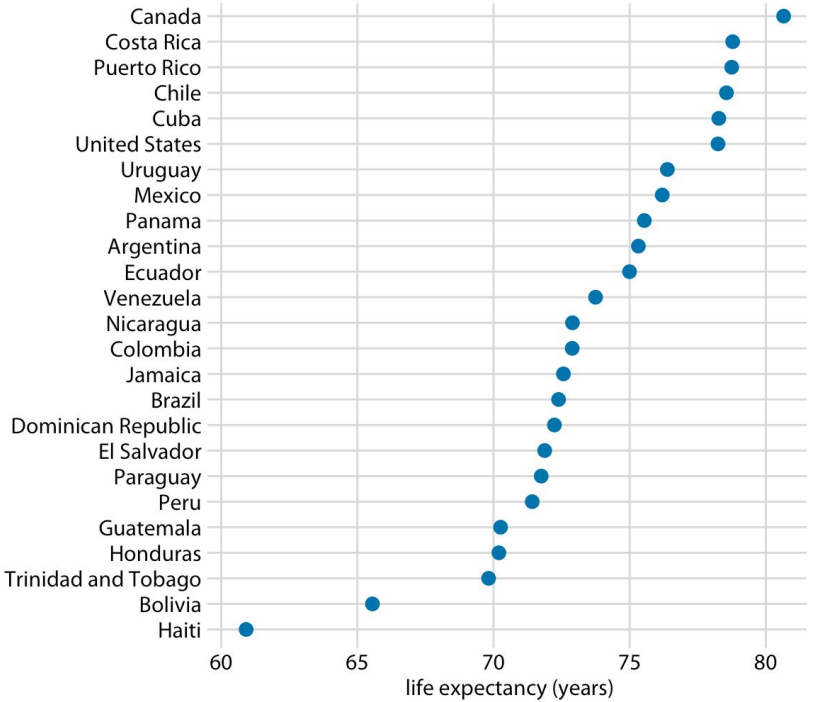
Plot negative bars to left



Highest grossing movies for the weekend of December 22-24, 2017, displayed as a bar plot. Source: https://clauswilke.com/dataviz/

Bar plots with many bars … dot plots

No need to start baseline at 0



Life expectancies of countries in the Americas, for the year 2007. Source: https://clauswilke.com/dataviz/

2a.iii. Line charts

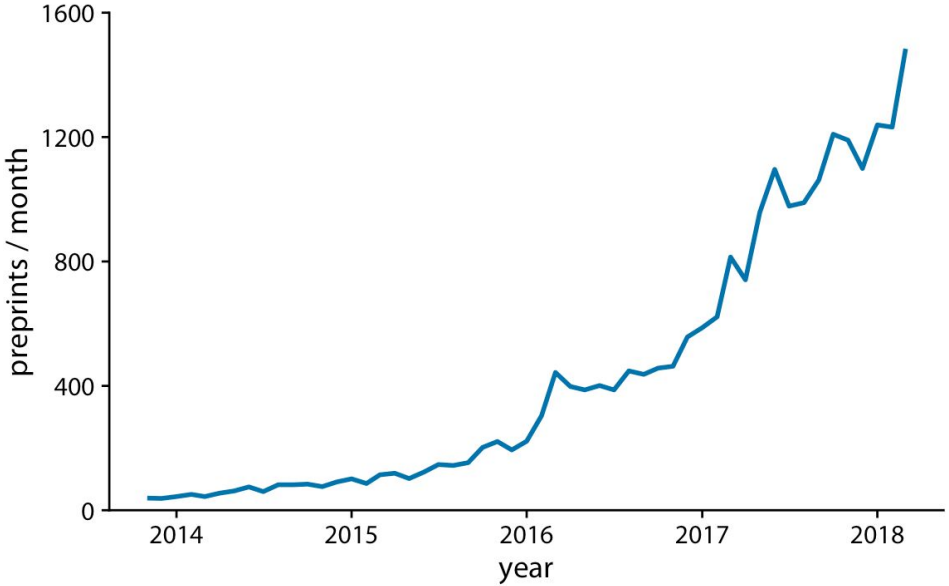
Use sensible y-axis: Line covers ⅔ y-axis range

Label lines directly

Only shade area below if baseline = 0

Use color and width for different lines

>4 lines = small multiples

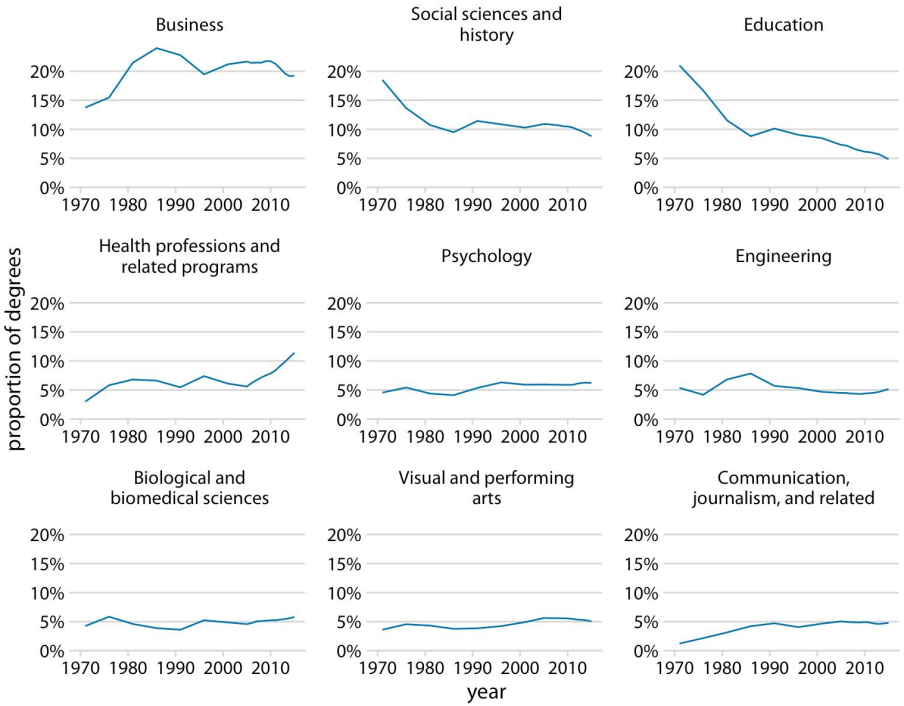


Monthly submissions to the preprint server bioRxiv, shown as a line graph without dots. Omitting the dots emphasizes the overall temporal trend while de-emphasizing individual observations at specific time points. It is particularly useful when the time points are spaced very densely. Source: https://clauswilke.com/dataviz/

Use small multiples

Align panels

Identical x- and y- axis range



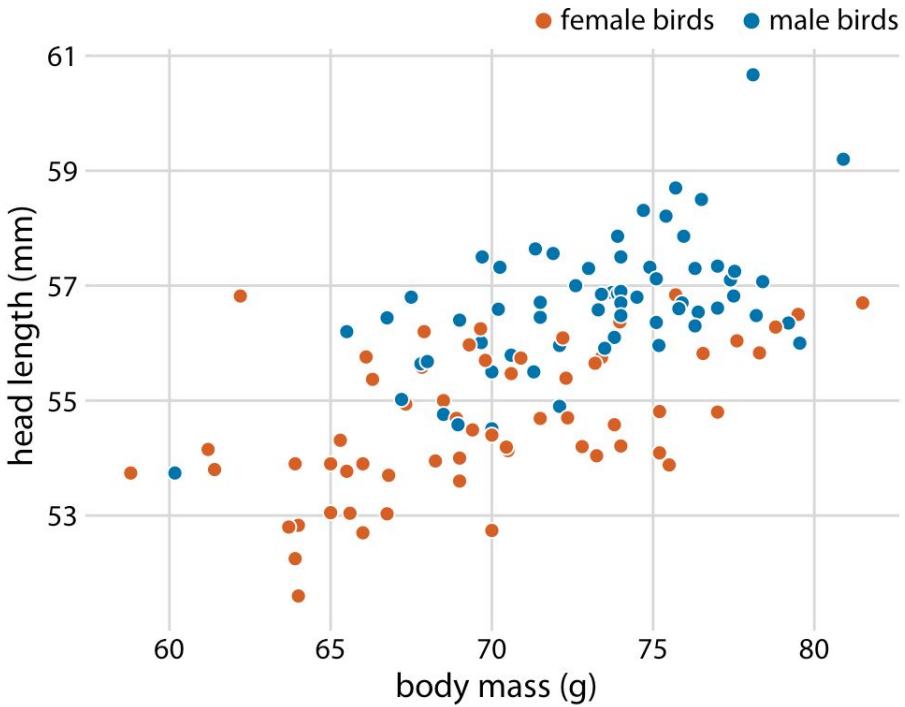
Trends in Bachelor's degrees conferred by U.S. institutions of higher learning. Shown are all degree areas that represent, on average, more than 4% of all degrees. Source: https://clauswilke.com/dataviz/

2a.iv. Scatterplots

Use sensible axes, usually low to high

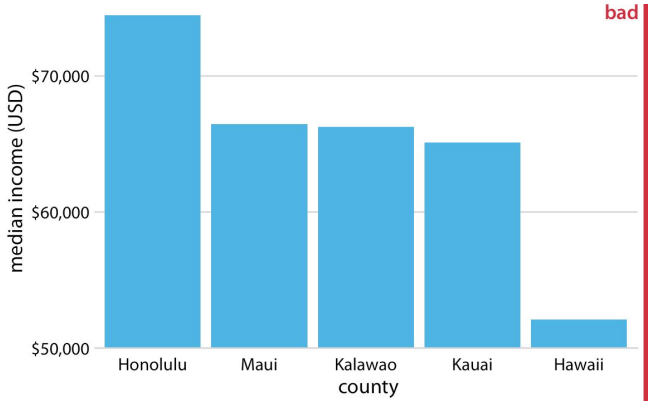
Large points, identified

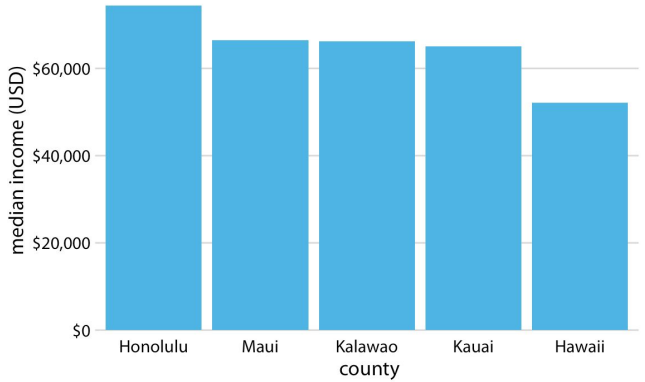
Clear axis labels, with units



https://clauswilke.com/dataviz/

2b. Avoid misrepresentation: start bars at 0

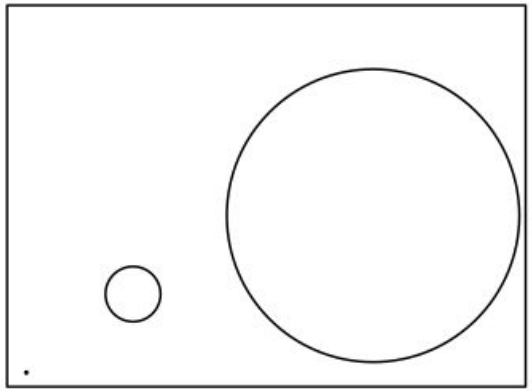




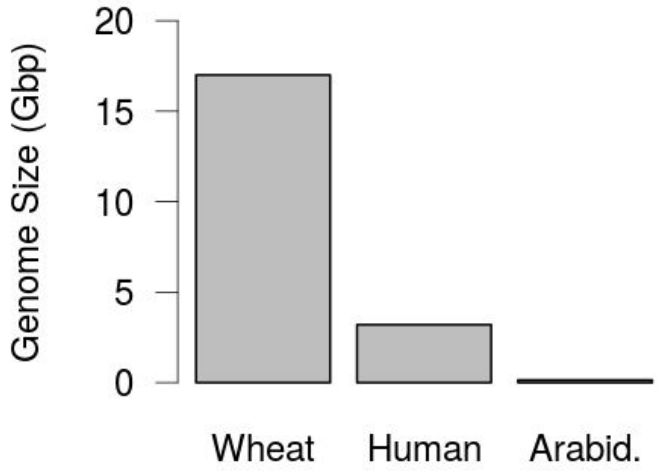
https://clauswilke.com/dataviz/

2b. Avoid misrepresentation: don’t distort data

Genome size (Gbp) ~ Area

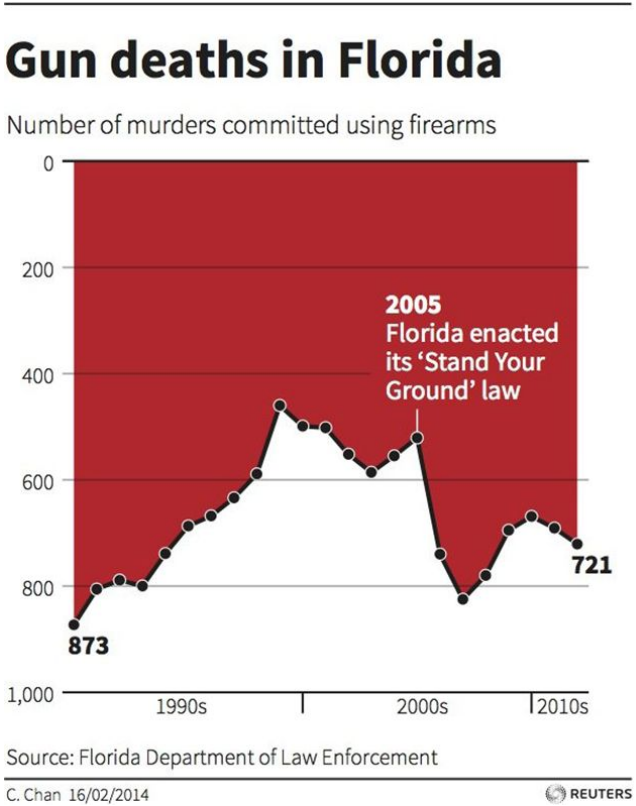


Genome size (Gbp) ~ Length



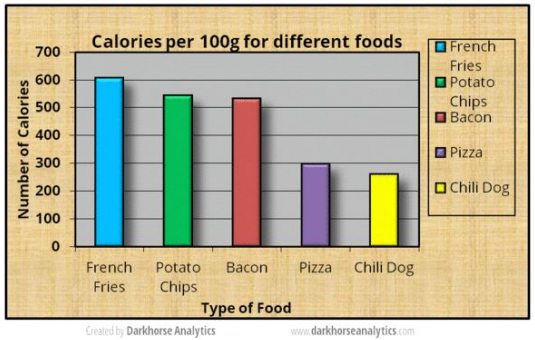
2b. Avoid misrepresentation: don't go against convention

Here, y-axis goes **DOWN ..**



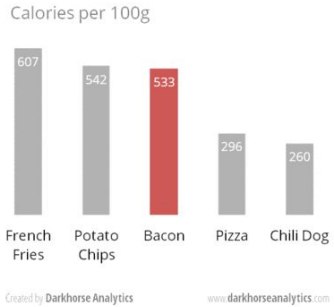
2c. Maximize data:ink, within reason

Before



Remove
to improve
(the **data-ink** ratio)

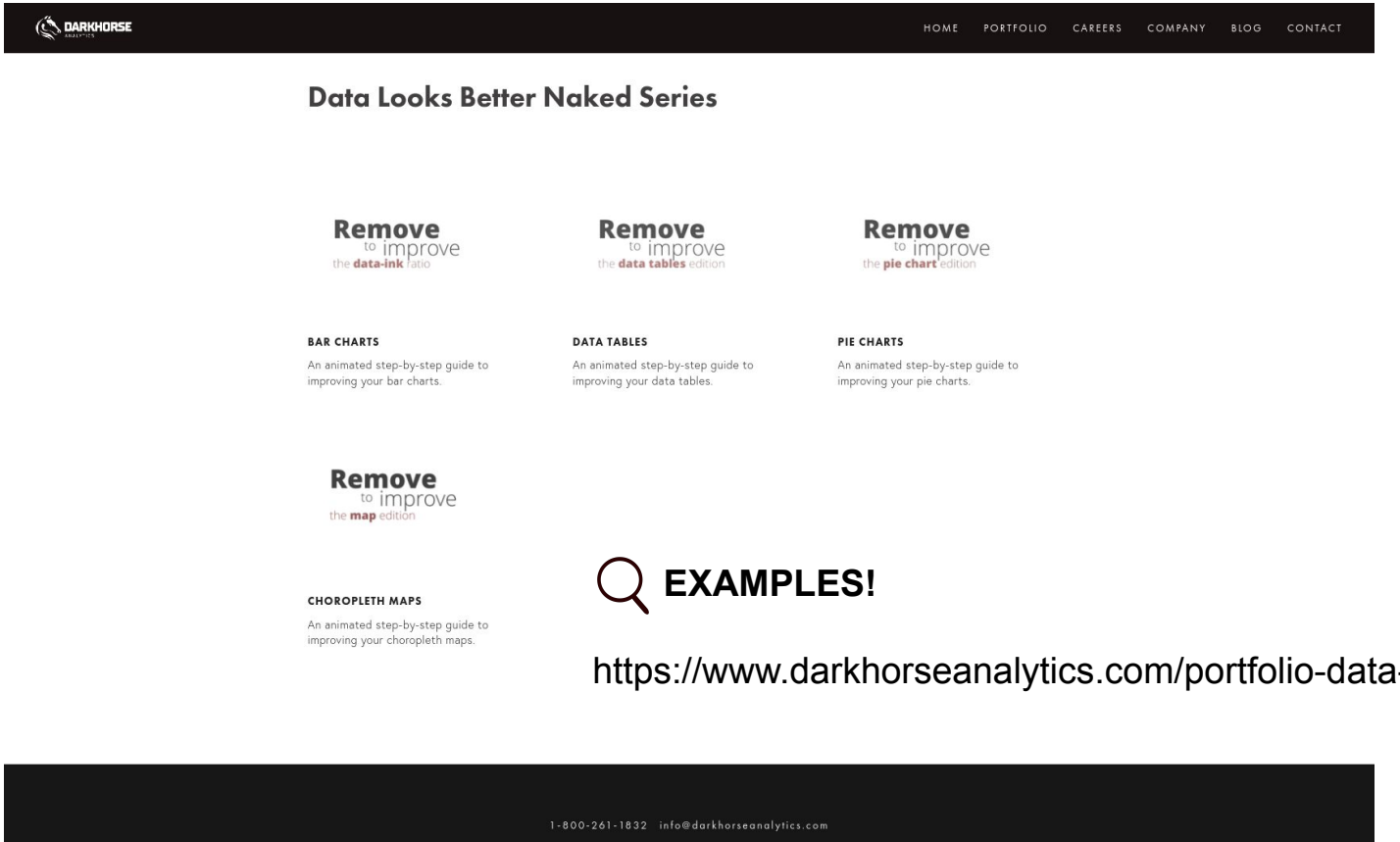
After



Created by Darkhorse Analytics www.darkhorseanalytics.com

https://www.darkhorseanalytics.com/portfolio/data-looks-better-naked-bar-charts

2c. Maximize data:ink (reduce distraction, less is more)

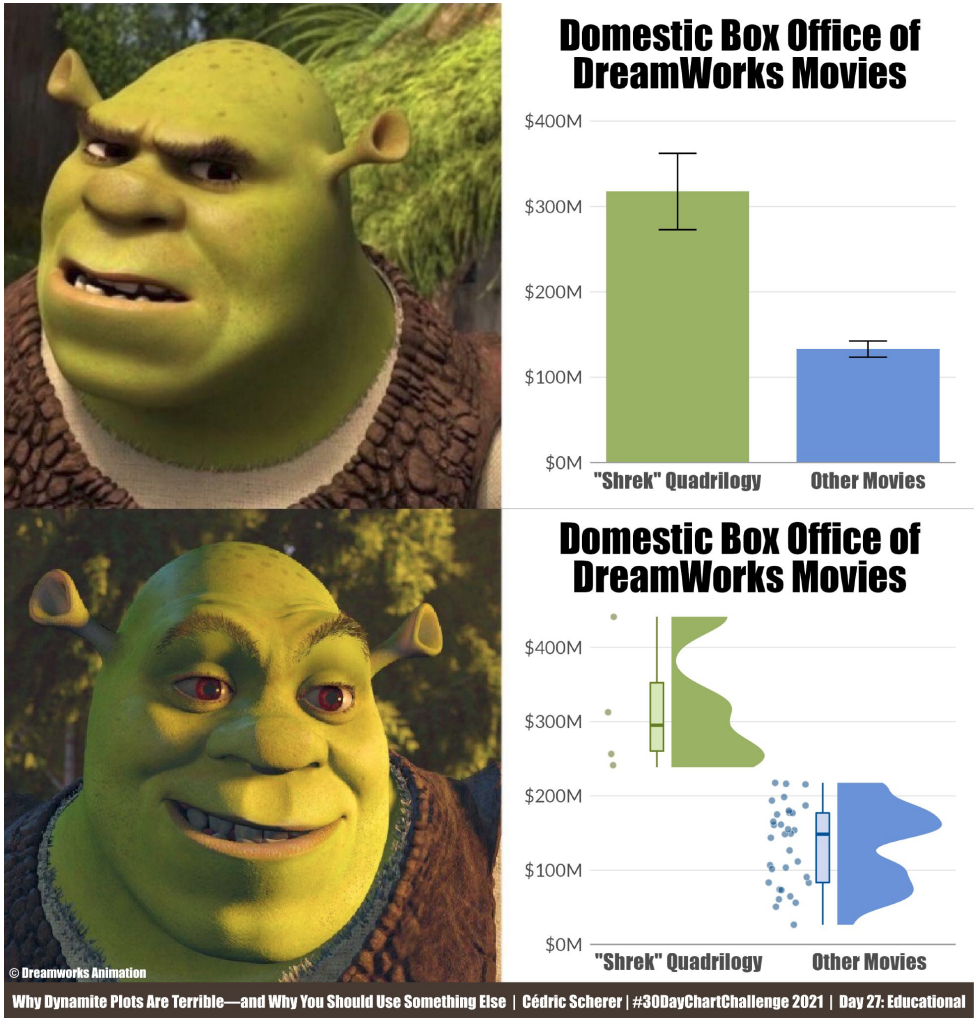


EXAMPLES!

https://www.darkhorseanalytics.com/portfolio-data-looks-better-naked

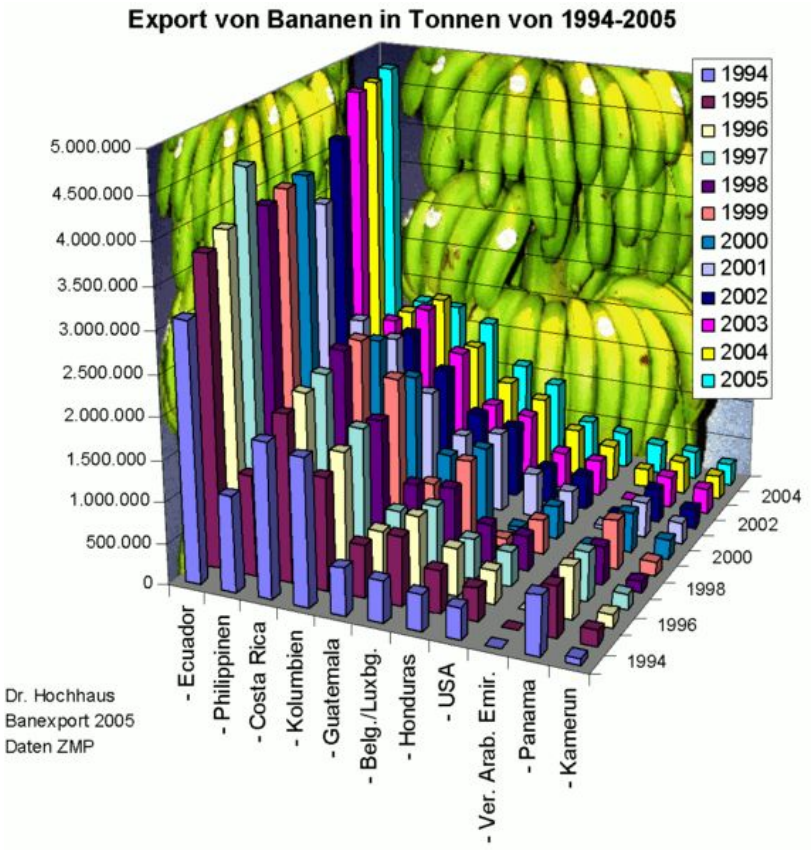
2. Display the data clearly

Try and show as much raw data as possible



Makeover #2

https://de.wikipedia.org/wiki/K%C3%BChlschifffahrt

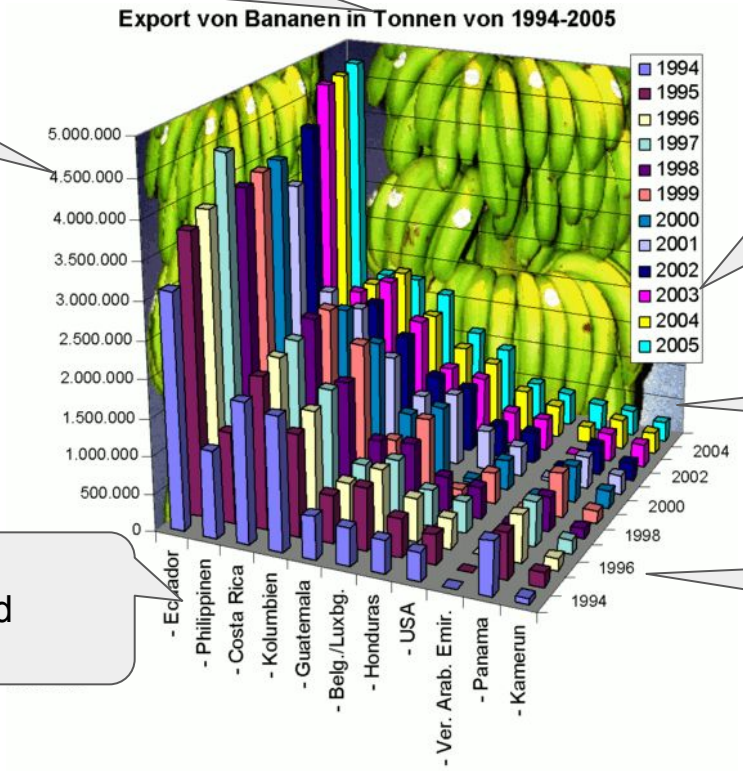


Mentimeter: Worst or best graphic ever … ? and why? What improvements would you make?

(1) Worst graph ever **(10) Most amazing graph ever**

[BAD!] Main problem: more color than information

Message: What is the message?



Y axis:
- Too many 0s … 4.5M

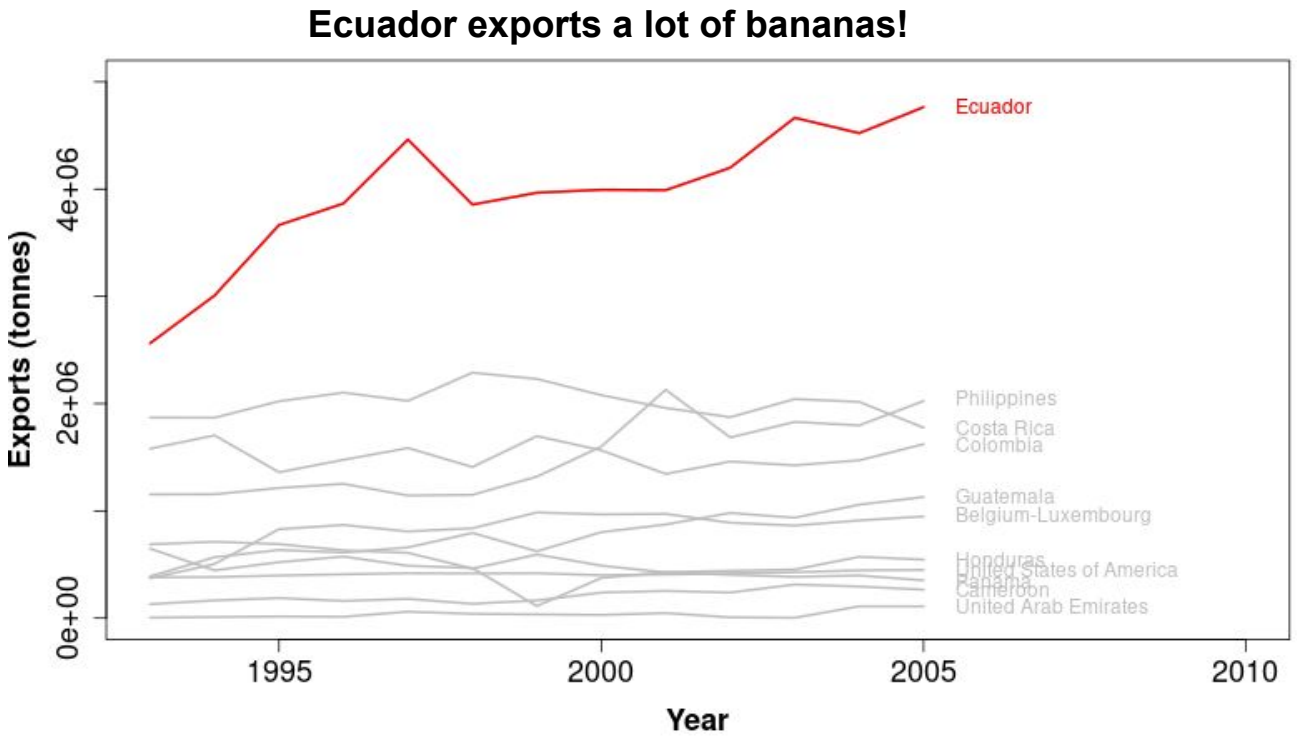
Data:
- **Time series** data should usually be **lines** and not bars
- 3D bar graphs are hard to interpret and see all the data

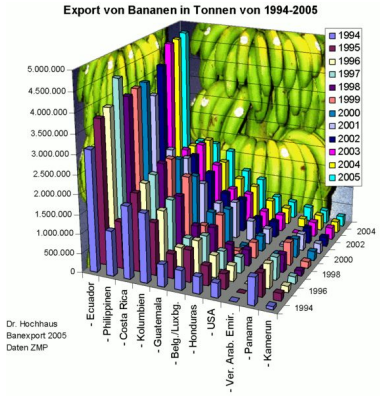
Image:
- Too many bananas!

X axis:
- Vertical text is hard to read
- Why use the hyphen?

Color:
- Years are color-coded; should be country

Remake original …





THREE PRINCIPLES OF EFFECTIVE COMMUNICATION

1. Have a clear purpose
2. Show the data clearly
3. **Make the message obvious**

3. Make the message obvious

a) Minimize mental arithmetic

b) Use proximity and alignment to aid in comparisons

c) Use colors and annotations to highlight important details

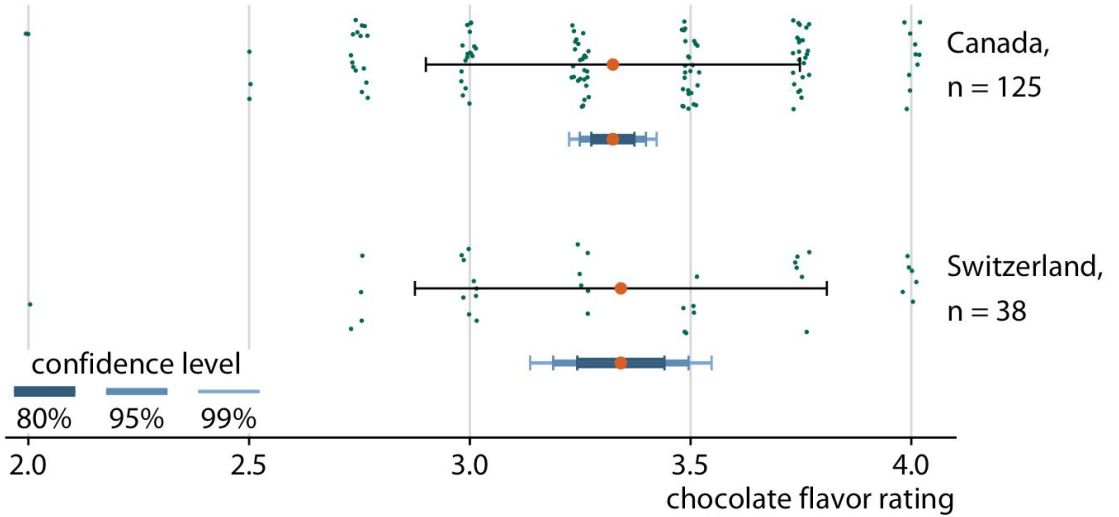
3a. Add meaningful information to tell the whole story

Reference lines,

Benchmark effects,

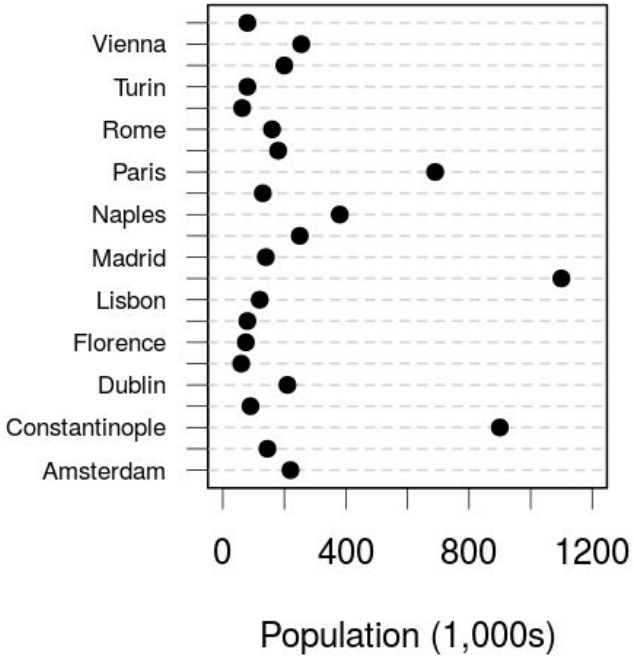
Inferences,

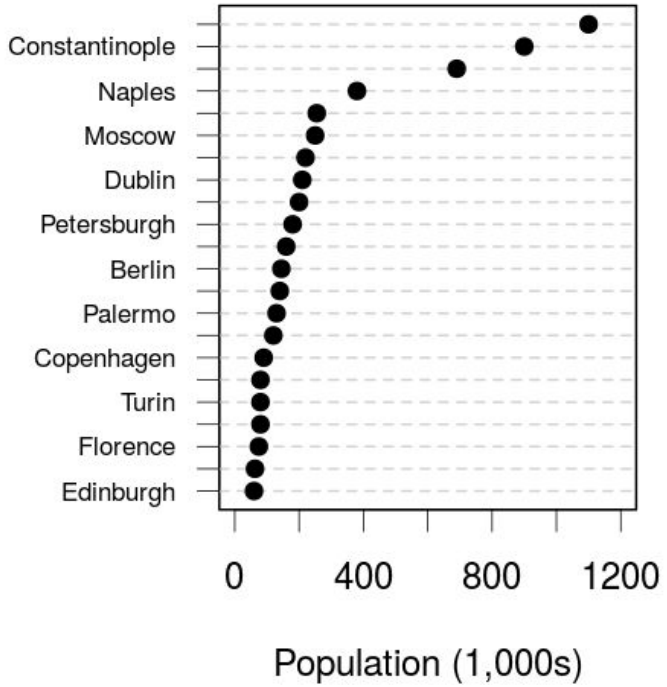
Variation, etc.



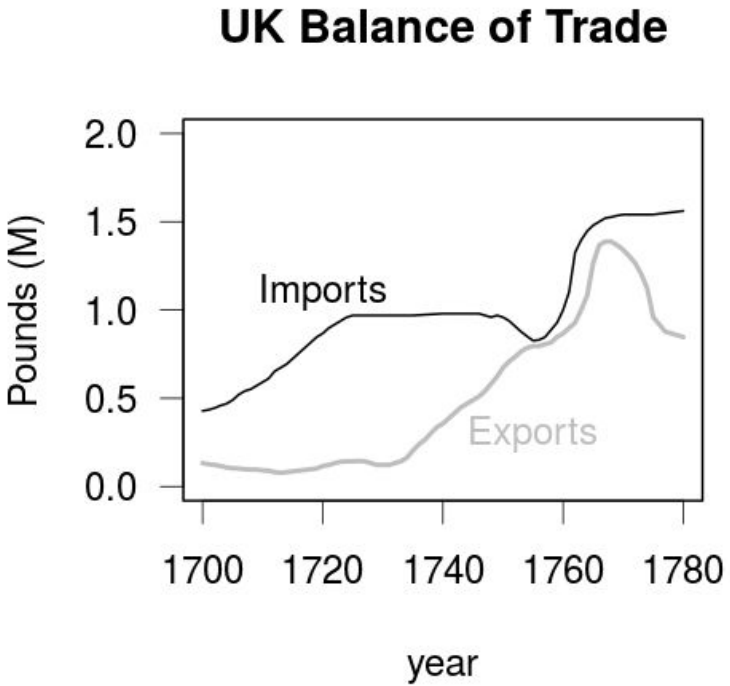
Confidence intervals widen with smaller sample size. Chocolate bars from Canada and Switzerland have comparable mean ratings and comparable standard deviations (indicated with simple black error bars). However, over three times as many Canadian bars were rated as Swiss bars, and therefore the confidence intervals (indicated with error bars of different colors and thickness drawn on top of one another) are substantially wider for the mean of the Swiss ratings than for the mean of the Canadian ratings

3a. Minimize mental arithmetic: Aid comparisons

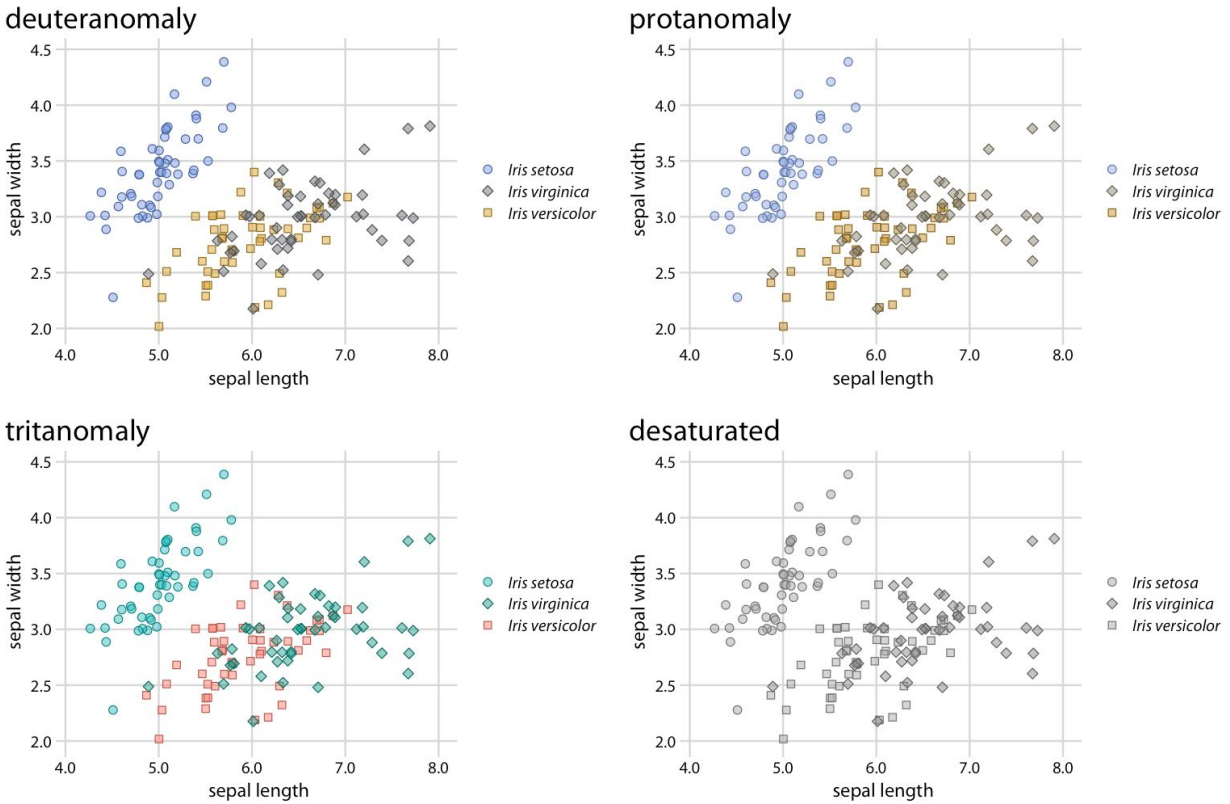




3a. Minimize mental arithmetic: Label data directly



3a. Minimize mental arithmetic: Use effective redundancy



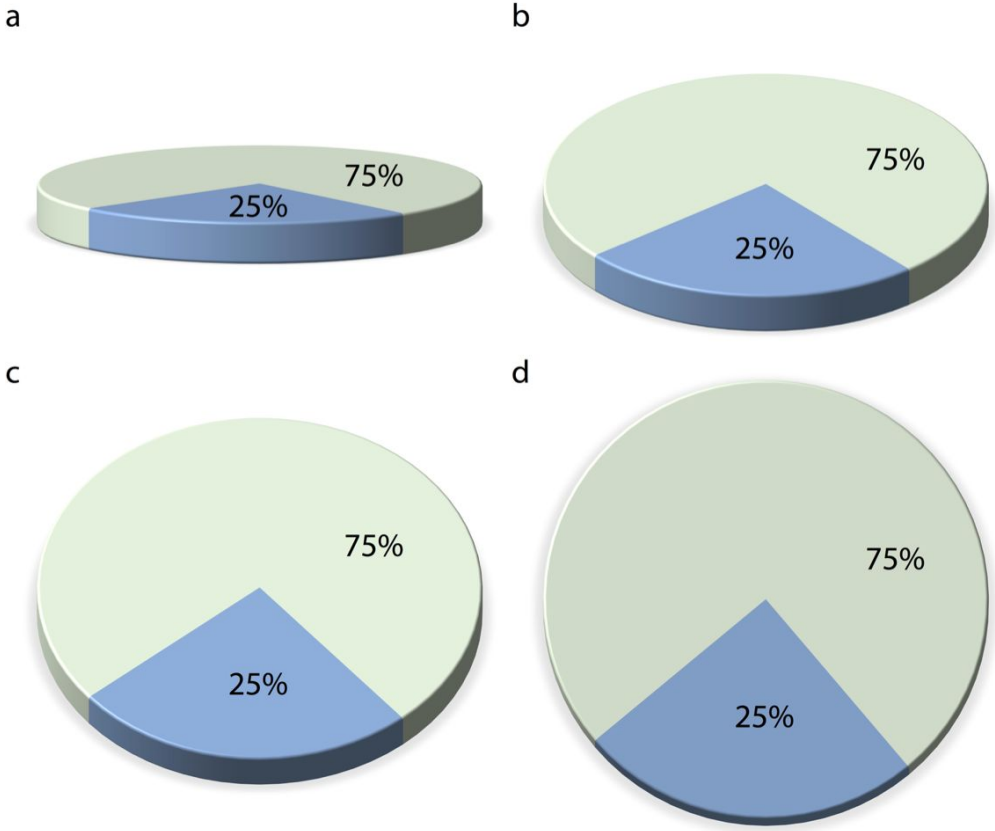
3a. Minimize mental arithmetic: Use clearly different colors, symbols, etc.

3a. Minimize mental arithmetic: Avoid 3D

3D distorts the data

The same 3D pie chart shown from four different angles.

Rotating a pie into the third dimension makes pie slices in the front appear larger than they really are and pie slices in the back appear smaller.



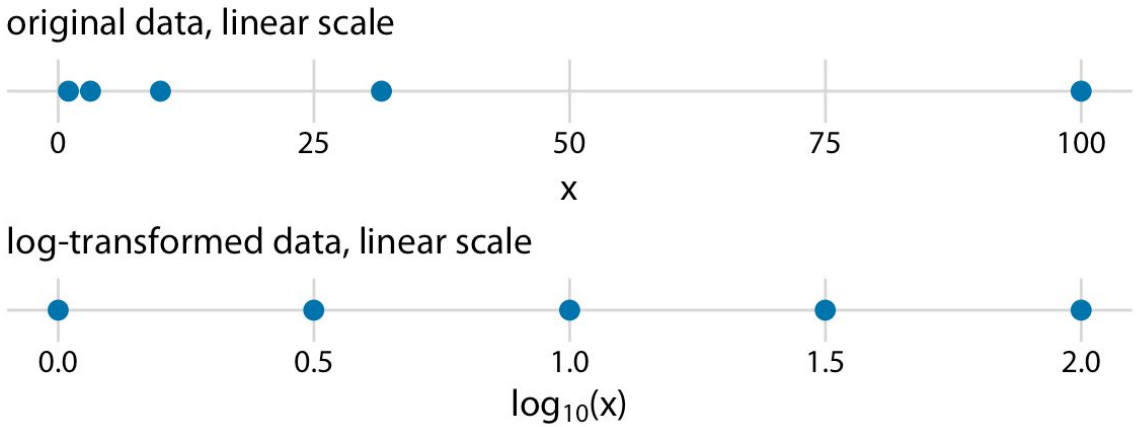
3a. Minimize mental arithmetic: Use sensible axes

Tick marks at natural increments (e.g., 0.2, 0.25, 1, 2, 5, 10, 25, 50, 100, …)

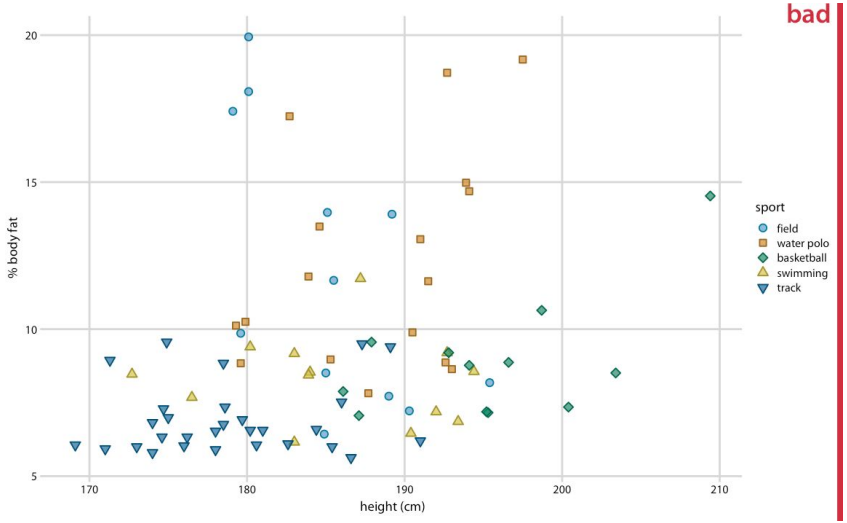
Keep to same decimal place

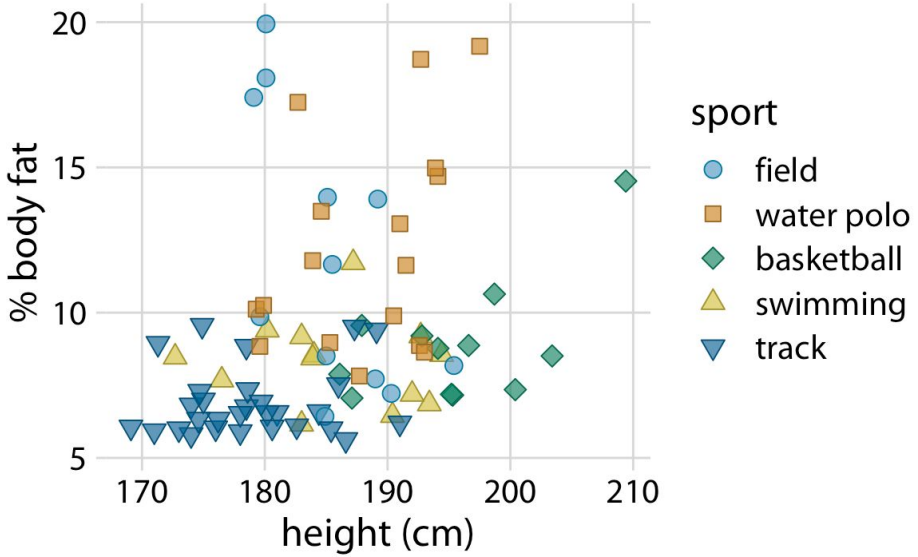
Maybe include hundred, millions, etc in axis title

Use scales that people understand …



3a. Minimize mental arithmetic: Use larger axis labels



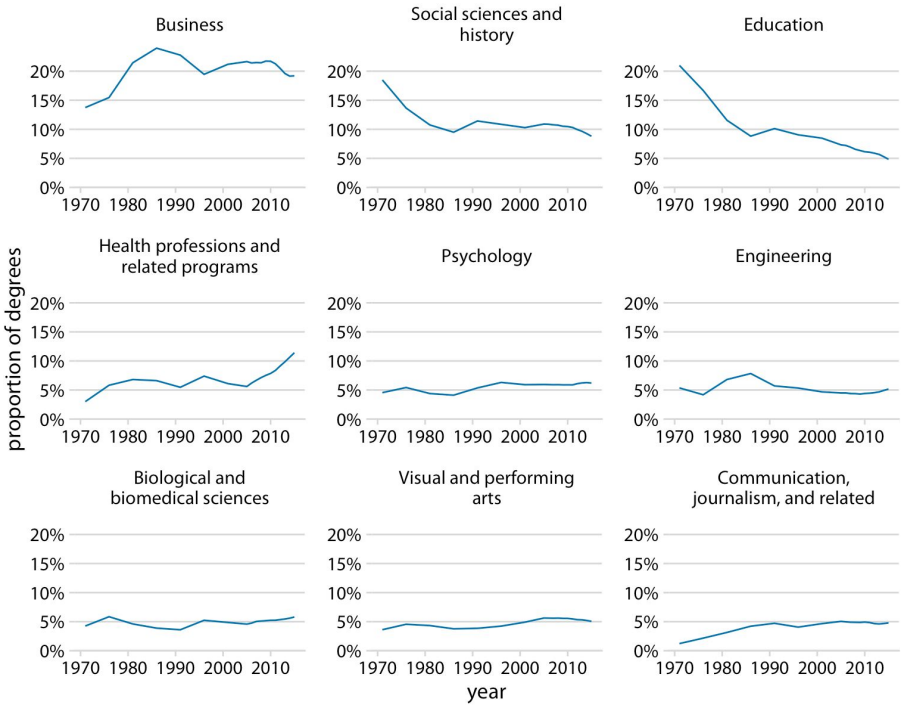


3b. Use proximity and alignment to aid in comparisons

Align small multiples

Same axes and limits on all panels

Same line color and type (and symbol, points, etc) across panels/figures



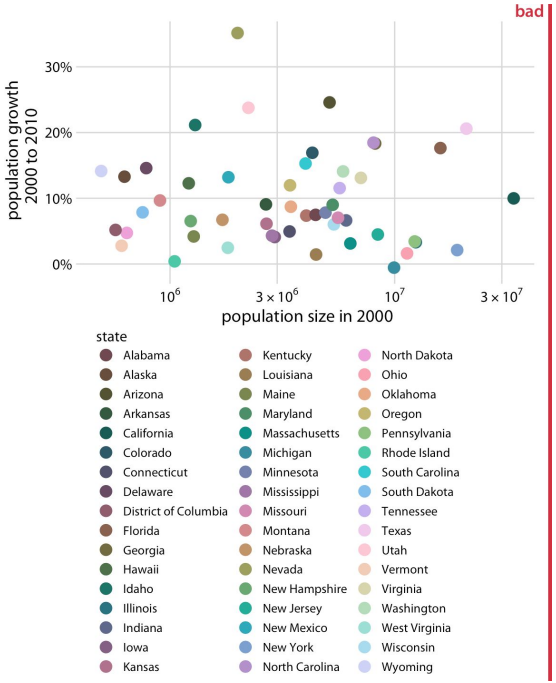
https://clauswilke.com/dataviz/

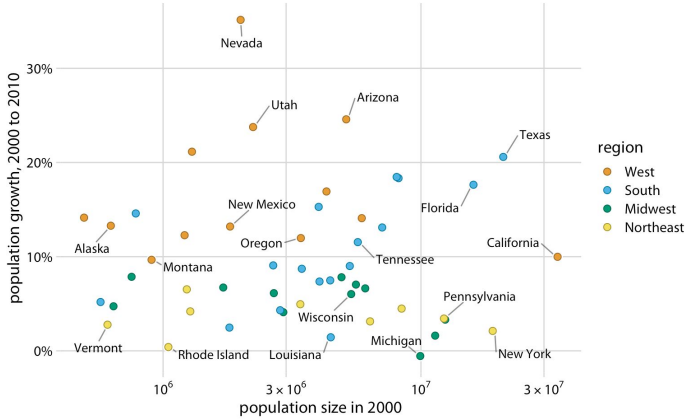
3c. Use colors to highlight important details

> “
> avoiding **catastrophe**
> becomes the
>
> first principle in bringing color to information:
>
> ”
> **Above all, do no harm.**

- Envisioning Information, Edward Tufte, Graphics Press, 1990

Do not encode too much or irrelevant information



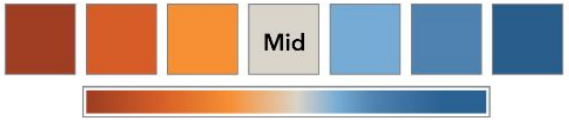


https://clauswilke.com/dataviz/

The Use of Color in Data Visualization

Sequential
color is ordered from low to high

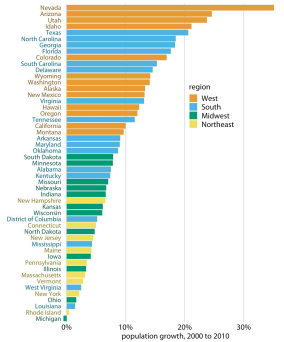
Diverging
two sequential colors with a neutral midpoint

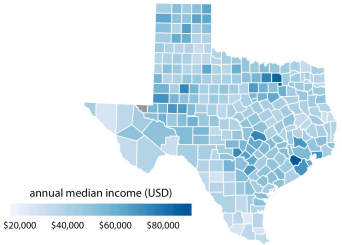


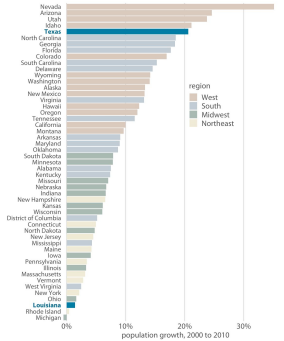
Categorical
contrasting colors for individual comparison

Highlight
color used to highlight something

Alert
color used to get reader's attention





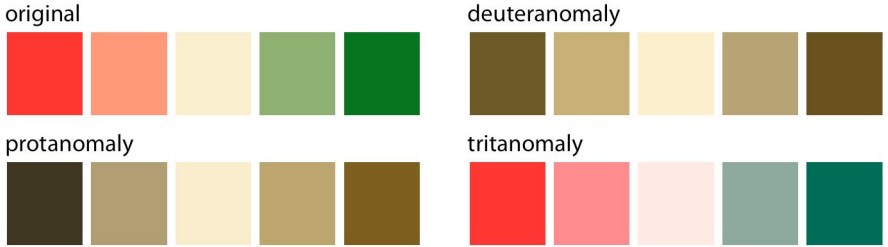


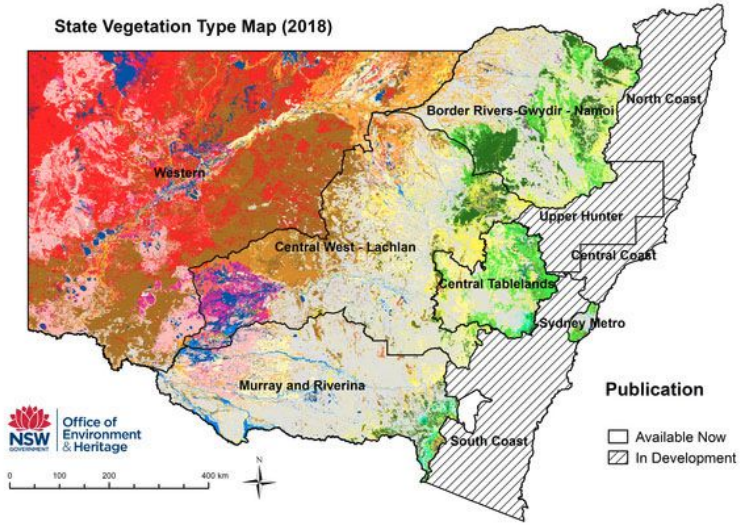
https://clauswilke.com/dataviz/

Do account for **color-vision deficiencies**

Approximately **8% of males** and **0.5% of females** suffer from some sort of color-vision deficiency.

A red–green contrast becomes indistinguishable under red–green cvd (deuteranomaly or protanomaly):

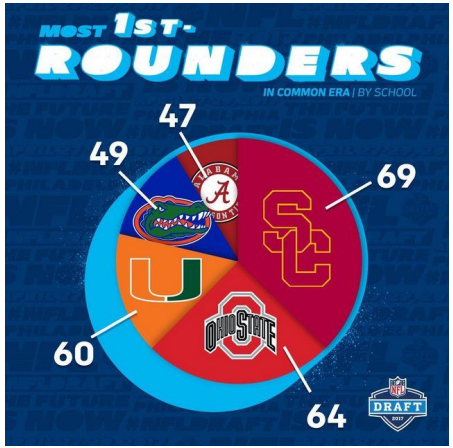


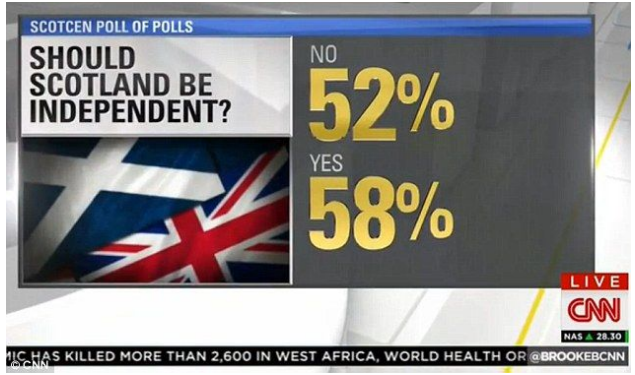


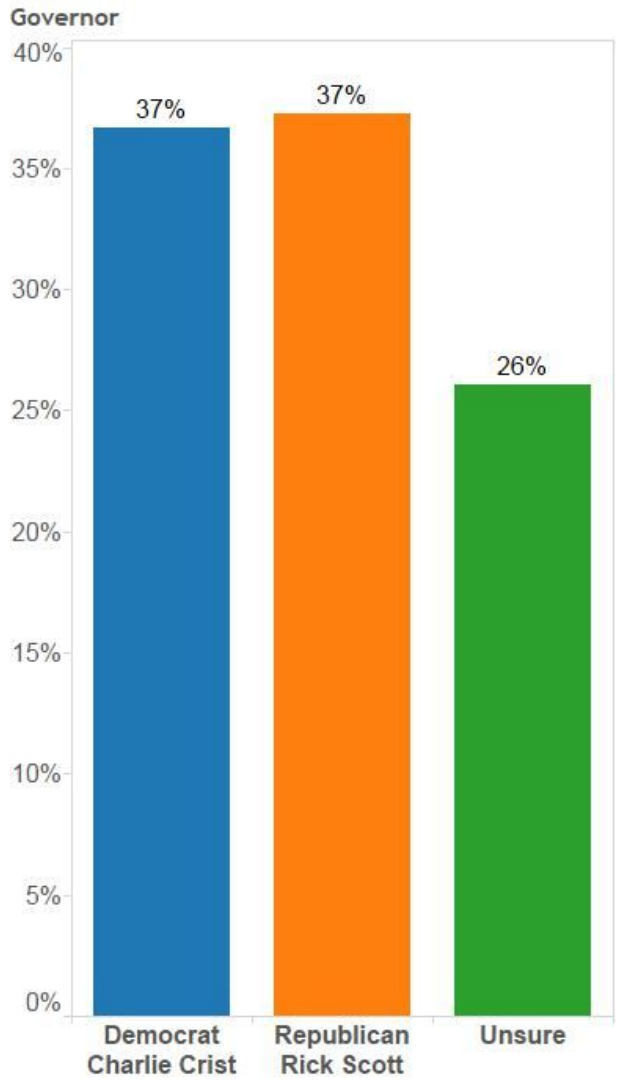
https://clauswilke.com/dataviz/color-pitfalls.html#not-designing-for-color-vision-deficiency

https://www.environment.nsw.gov.au/vegetation/state-vegetation-type-map.htm

Please, just don’t …



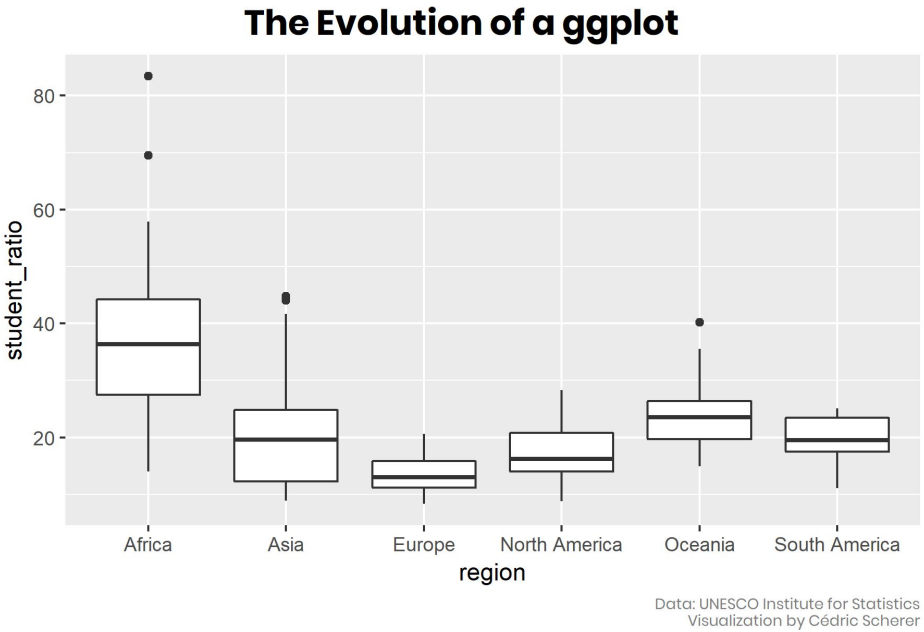




THREE PRINCIPLES OF EFFECTIVE COMMUNICATION

1. Have a clear purpose
2. Show the data clearly
3. Make the message obvious

R CODE: The evolution of a ggplot



CODE TIP!

https://www.cedricscherer.com/2019/05/17/the-evolution-of-a-ggplot-ep.-1/

Resources

Examples of good charts

https://www.cedricscherer.com/2021/05/09/contributions-30daychartchallenge-2021/

Books and guides

Healy, K. 2019 Data Visualization: A practical introduction. https://socviz.co/

Wilke, C. 2019. Fundamentals of Data Visualization. https://clauswilke.com/dataviz/

Wong, D. 2013. The Wall Street Journal Guide to Information Graphics. ~$17 (Amazon) ISBN# 978-0393347289

Data Visualization 101 ebook: https://visage.co/content/data-visualization-101/

Wickham & Grolemund. R for Data Science. Chapter 3: Data Visualization https://r4ds.had.co.nz/data-visualisation.html

Tools for making graphs: R

https://swirlstats.com/

https://r-graphics.org/

https://www.r-bloggers.com/2016/05/free-e-book-effective-graphs-with-microsoft-r-open/

https://stat545.com/graphics-overview.html

https://www.cedricscherer.com/2019/08/05/a-ggplot2-tutorial-for-beautiful-plotting-in-r/

https://www.cedricscherer.com/2019/05/17/the-evolution-of-a-ggplot-ep.-1/

Tools for making graphs: Others

Most software has horrible defaults. Datawrapper is actually ok!

Data Wrapper: https://www.datawrapper.de/

Tableau: https://www.tableau.com/academic/students

Excel: https://stephanieevergreen.com/how-to/

Canva: https://www.canva.com/

Color

Comprehensive list of color palettes in r:
https://github.com/EmilHvitfeldt/r-color-palettes

Palettes: https://colorbrewer2.org/

Scales: https://clauswilke.com/dataviz/color-basics.html

Pitfalls: https://clauswilke.com/dataviz/color-pitfalls.html

Cheatsheets

Design Principles

https://github.com/GraphicsPrinciples/CheatSheet/blob/master/NVSCheatSheet.pdf

https://stephanieevergreen.com/data-visualization-checklist/

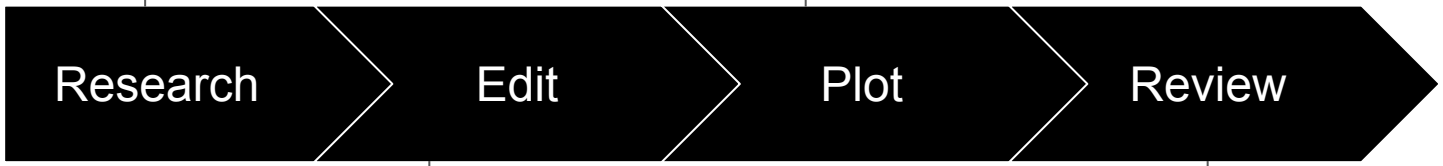
ggplot

https://rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf

How do you get actionable insights?

Identify problem, need, & data

Apply design principles to choose chart, add layers, text, refine



Understand, clean, & shape your data

Check for errors, areas of confusion, opportunities for improvement

Research

Who is your audience?

What do they need?

What is your question?

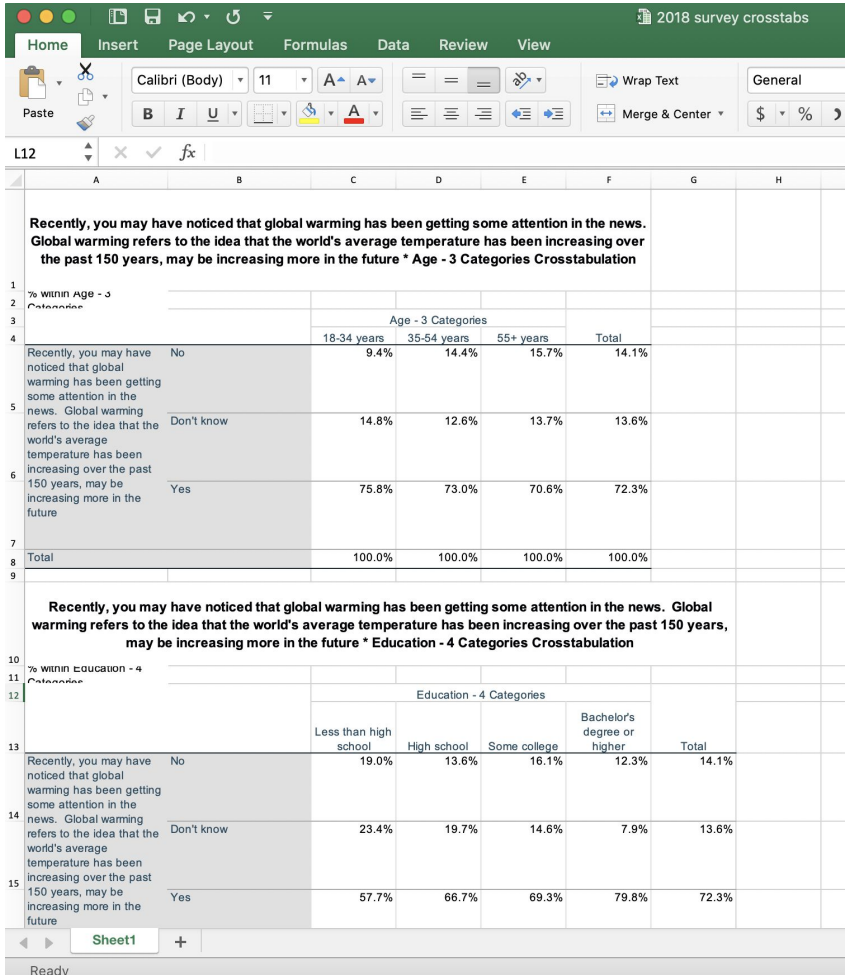
Is it timely? Is it relevant?

Can you answer it?

Is there data? How good is it?

Can you get it? How quickly? At what cost?

Edit



Recently, you may have noticed that global warming has been getting some attention in the news. Global warming refers to the idea that the world's average temperature has been increasing over the past 150 years, may be increasing more in the future * Age - 3 Categories Crosstabulation

% within Age - 3 Categories

		Age - 3 Categories			
		18-34 years	35-54 years	55+ years	Total
Recently, you may have noticed that global warming has been getting some attention in the news. Global warming refers to the idea that the world's average temperature has been increasing over the past 150 years, may be increasing more in the future	No	9.4%	14.4%	15.7%	14.1%
	Don't know	14.8%	12.6%	13.7%	13.6%
	Yes	75.8%	73.0%	70.6%	72.3%
Total		100.0%	100.0%	100.0%	100.0%

Recently, you may have noticed that global warming has been getting some attention in the news. Global warming refers to the idea that the world's average temperature has been increasing over the past 150 years, may be increasing more in the future * Education - 4 Categories Crosstabulation

% within Education - 4 Categories

		Education - 4 Categories				
		Less than high school	High school	Some college	Bachelor's degree or higher	Total
Recently, you may have noticed that global warming has been getting some attention in the news. Global warming refers to the idea that the world's average temperature has been increasing over the past 150 years, may be increasing more in the future	No	19.0%	13.6%	16.1%	12.3%	14.1%
	Don't know	23.4%	19.7%	14.6%	7.9%	13.6%
	Yes	57.7%	66.7%	69.3%	79.8%	72.3%

Plot

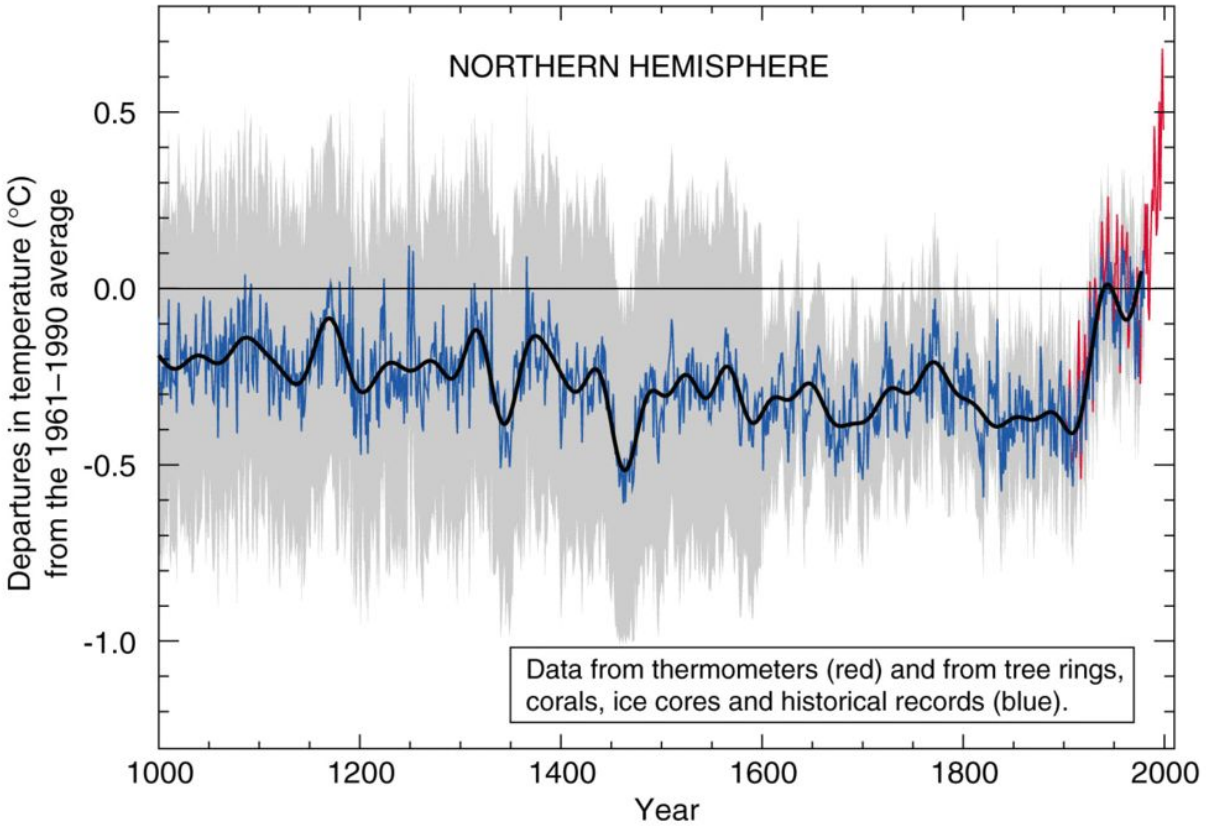
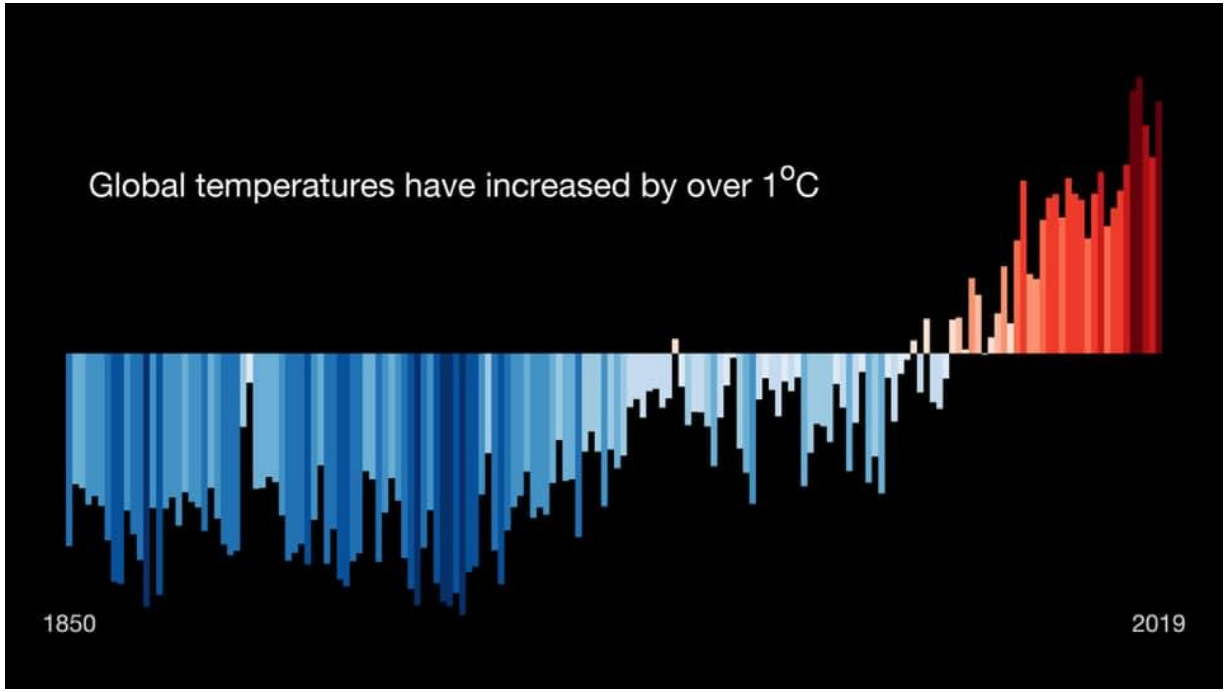


IMAGE: IPCC AND MICHAEL MANN/PENN STATE

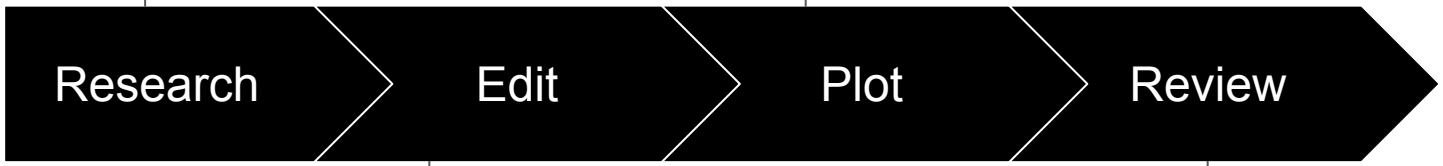
Review



How do you get actionable insights?

Identify problem, need, & data

Apply design principles to choose chart, add layers, text, refine



Understand, clean, & shape your data

Check for errors, areas of confusion, opportunities for improvement