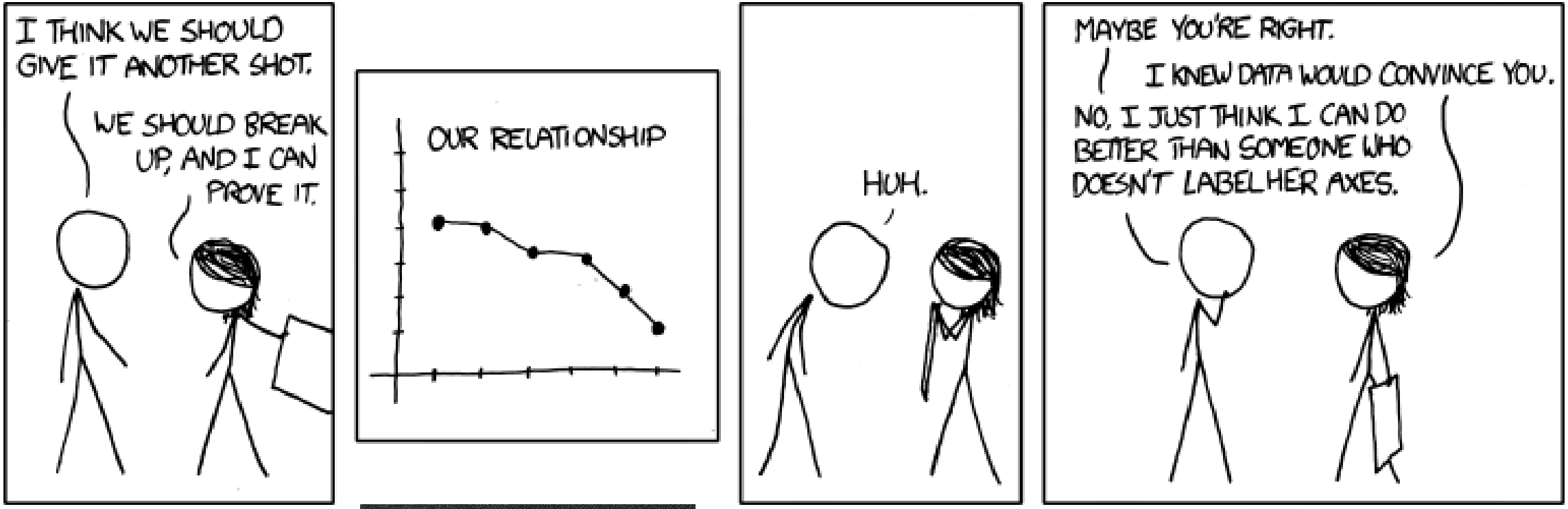
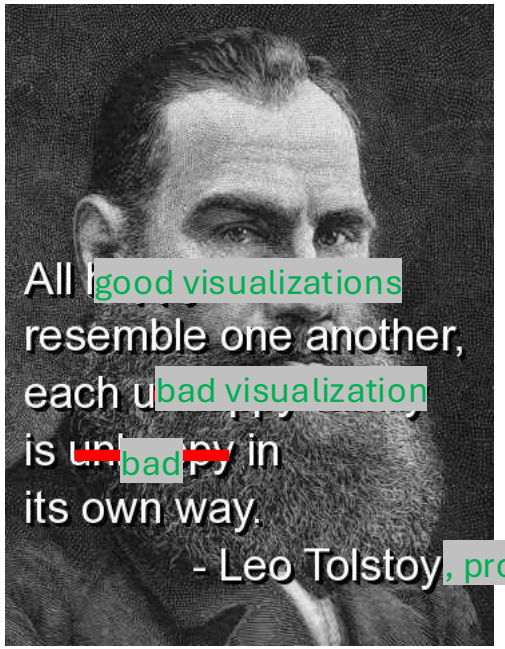
Principles of Visualization

HUGEN 2073

Jon Chernus

Adapted from HJ Park’s slides with permission


I THINK WE SHOULD GIVE IT ANOTHER SHOT.
WE SHOULD BREAK UP, AND I CAN PROVE IT.
OUR RELATIONSHIP
HUH.
MAYBE YOU'RE RIGHT.
I KNEW DATA WOULD CONVINCE YOU.
NO, I JUST THINK I CAN DO BETTER THAN SOMEONE WHO DOESN'T LABEL HER AXES.


All good visualizations
resemble one another,
each u bad visualization
is un bad py in
its own way.
- Leo Tolstoy, probably?

Why visualize? Two purposes in tension

- Exploring
  - Discovering questions and answers (iteratively)
  - Making decisions
  - Finding patterns
- Communicating
  - Presenting methods, arguments, conclusions, etc.
  - Recording these

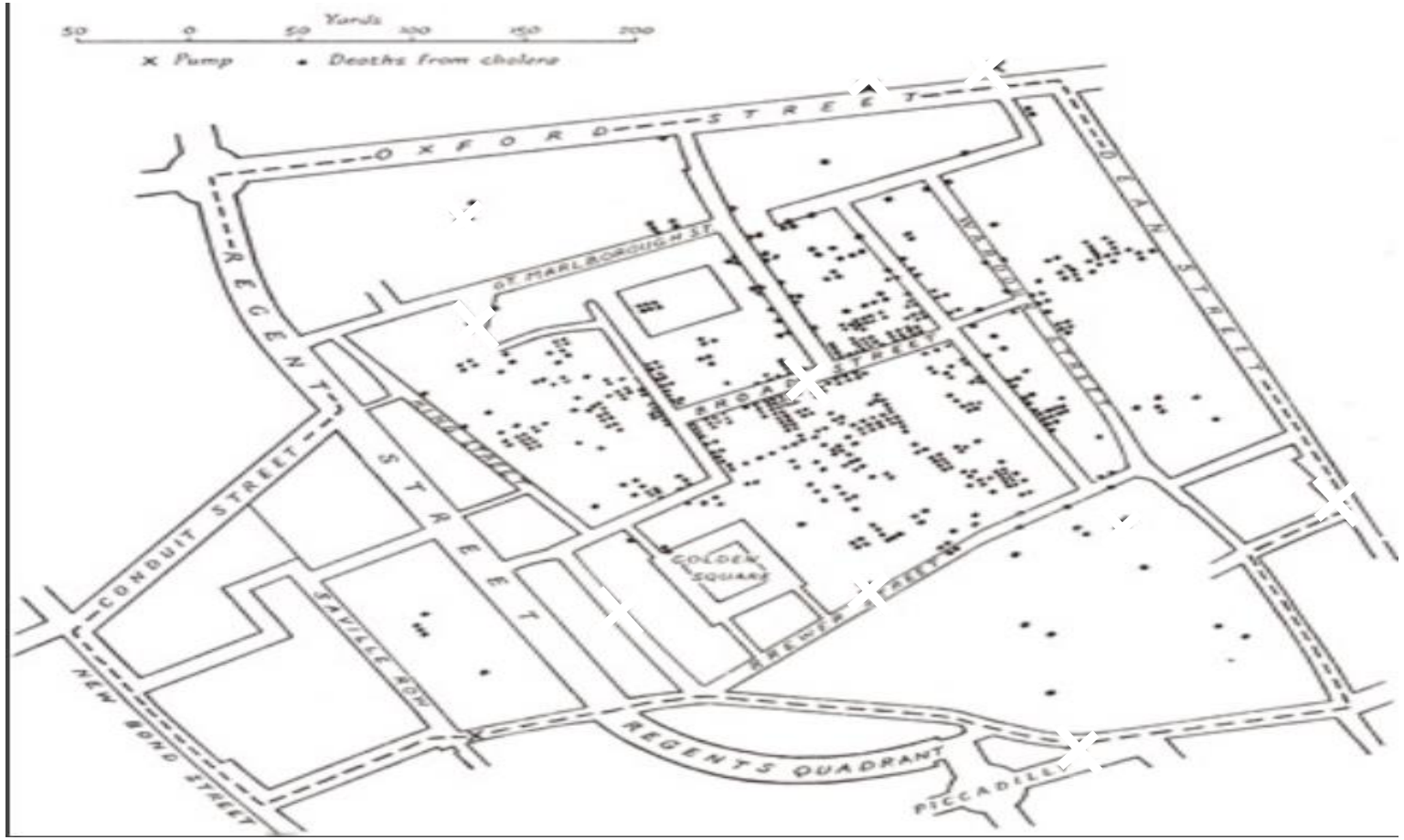
Example: exploring

The location of cholera deaths in central London, Sep. 1854

The Visual Display of Quantitative Information, E. Tufte, 2nd Ed. Graphics Press

Example: exploring

The location of cholera deaths in central London, Sep. 1854



The Visual Display of Quantitative Information, E. Tufte, 2nd Ed. Graphics Press

- The locations of water pumps marked
- An association between a water pump and the deaths
- Help make decisions efficiently

Example: exploring

Anscombe's Quartet: 11 (x, y) pairs

	Group 1		Group 2		Group 3		Group 4	
	X	Y	X	Y	X	Y	X	Y
1	10	8.04	10	9.14	10	7.46	8	6.58
2	8	6.95	8	8.14	8	6.77	8	5.76
3	13	7.58	13	8.74	13	12.74	8	7.71
4	9	8.81	9	8.77	9	7.11	8	8.81
5	11	8.33	11	9.26	11	7.81	8	8.47
6	14	9.96	14	8.1	14	8.84	8	7.04
7	6	7.24	6	6.13	6	6.08	8	5.25
8	4	4.26	4	3.1	4	5.39	19	12.5
9	12	10.84	12	9.13	12	8.15	8	5.56
10	7	4.82	7	7.26	7	6.42	8	7.91
11	5	5.68	5	4.74	5	5.73	8	6.89

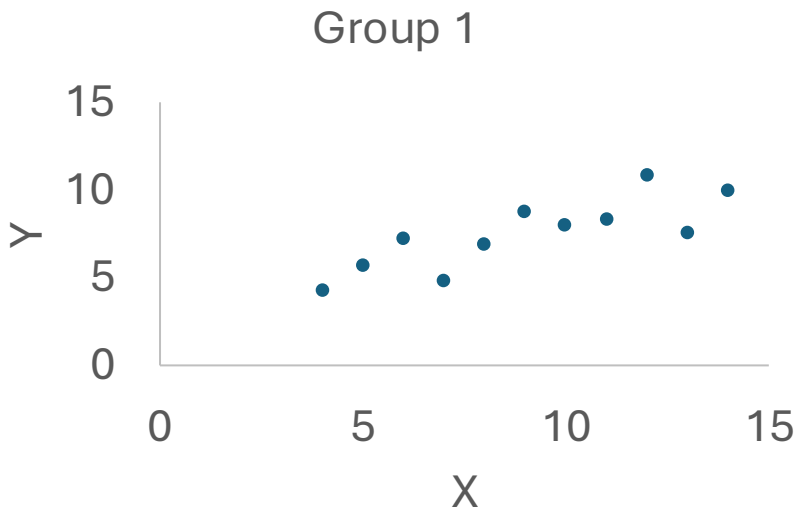
Example: exploring

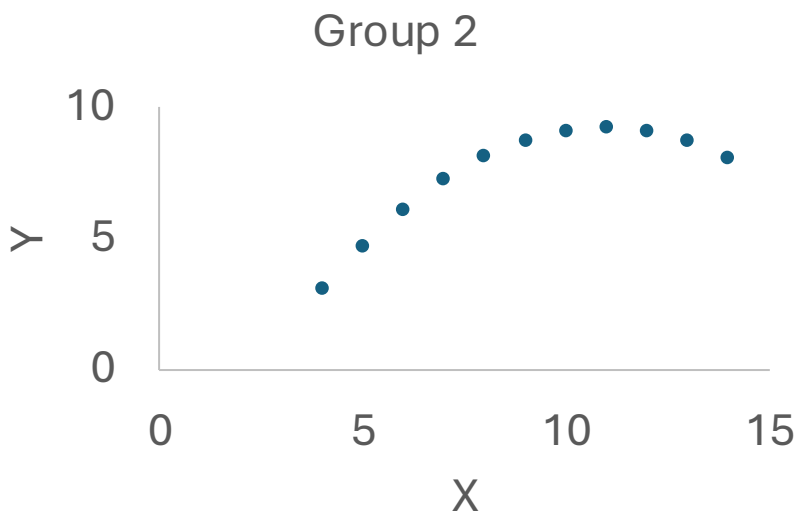
Anscombe's Quartet: 11 (x, y) pairs

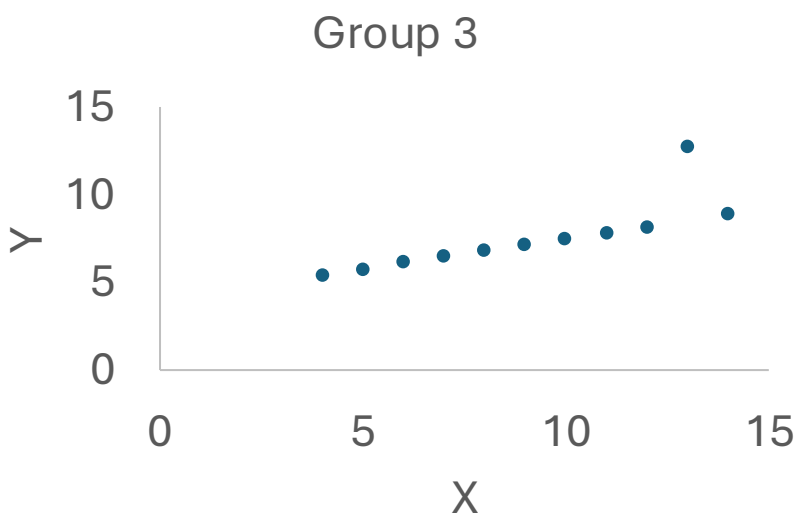
	Group 1		Group 2		Group 3		Group 4	
	X	Y	X	Y	X	Y	X	Y
1	10	8.04	10	9.14	10	7.46	8	6.58
2	8	6.95	8	8.14	8	6.77	8	5.76
3	13	7.58	13	8.74	13	12.74	8	7.71
4	9	8.81	9	8.77	9	7.11	8	8.81
5	11	8.33	11	9.26	11	7.81	8	8.47
6	14	9.96	14	8.1	14	8.84	8	7.04
7	6	7.24	6	6.13	6	6.08	8	5.25
8	4	4.26	4	3.1	4	5.39	19	12.5
9	12	10.84	12	9.13	12	8.15	8	5.56
10	7	4.82	7	7.26	7	6.42	8	7.91
11	5	5.68	5	4.74	5	5.73	8	6.89
Mean	9	7.5	9	7.5	9	7.5	9	7.5
Variance	9.1	3.4	9.1	3.4	9.1	3.4	9.1	3.4
Correlation	0.81		0.81		0.81		0.81	

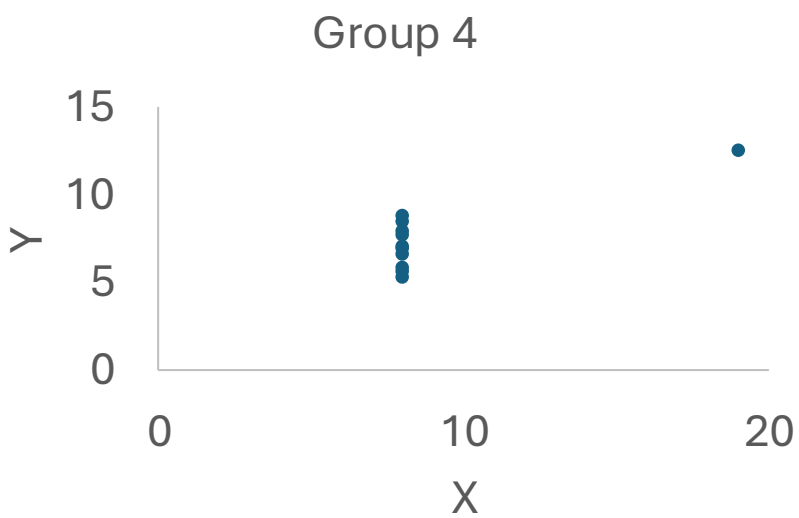
Statistics is used to find patterns… but are summary stats enough?

Example: exploring





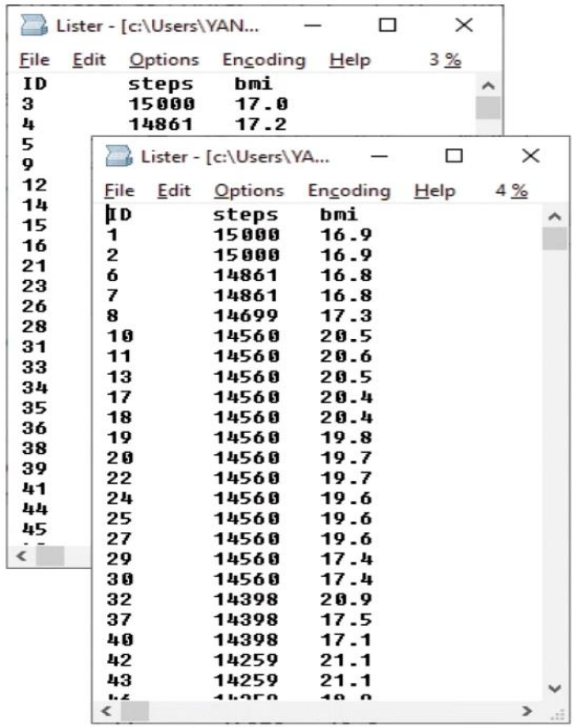




Example: exploring

The Gorilla Experiment:

an artificial dataset given to students with and without explicit hypotheses on the relationship between the variables

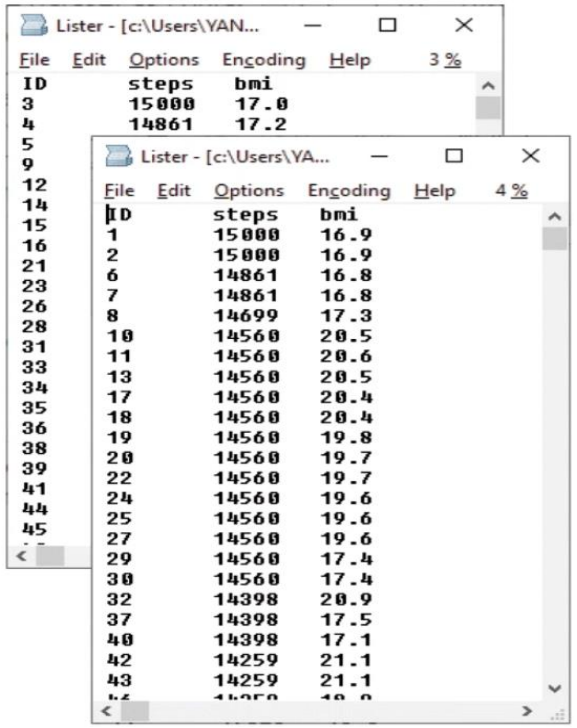


Yanai and Lercher, Genome Biology, 202

Example: exploring

- The Gorilla Experiment

An artificial dataset given to students <u>with</u> and <u>without</u> explicit hypotheses on the relationship between the variables



Lister - [c:\Users\YAN...

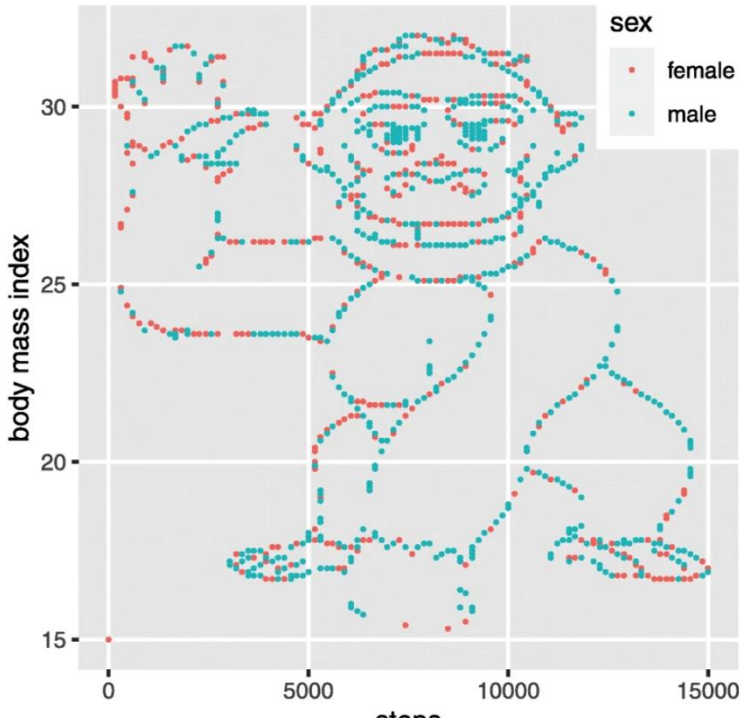
File Edit Options Encoding Help 3 %

ID	steps	bmi
3	15000	17.0
4	14861	17.2
5		
9		
12		
14		
15		
16		
21		
23		
26		
28		
31		
33		
34		
35		
36		
38		
39		
41		
44		
45		

Lister - [c:\Users\YA...

File Edit Options Encoding Help 4 %

ID	steps	bmi
1	15000	16.9
2	15000	16.9
6	14861	16.8
7	14861	16.8
8	14699	17.3
10	14560	20.5
11	14560	20.6
13	14560	20.5
17	14560	20.4
18	14560	20.4
19	14560	19.8
20	14560	19.7
22	14560	19.7
24	14560	19.6
25	14560	19.6
27	14560	19.6
29	14560	17.4
30	14560	17.4
32	14398	20.9
37	14398	17.5
40	14398	17.1
42	14259	21.1
43	14259	21.1



Yanai and Lercher, Genome Biology, 202

Example: exploring

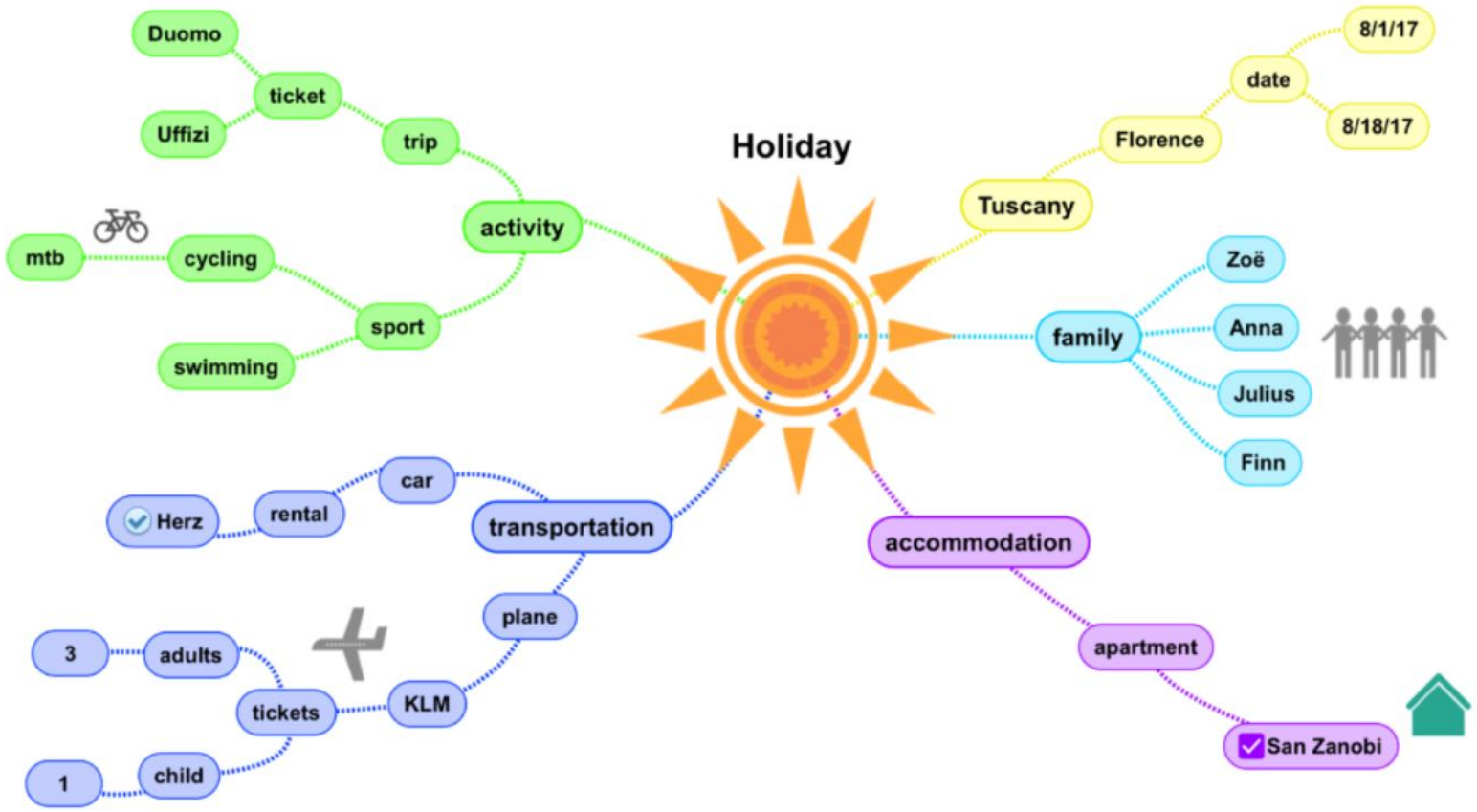
The Gorilla Experiment

An artificial dataset given to students with and without explicit hypotheses on the relationship between the variables

	Gorilla not discovered	Gorilla discovered
Hypothesis-focused	14	5
Hypothesis-free	5	9

Yanai and Lercher, Genome Biology, 202

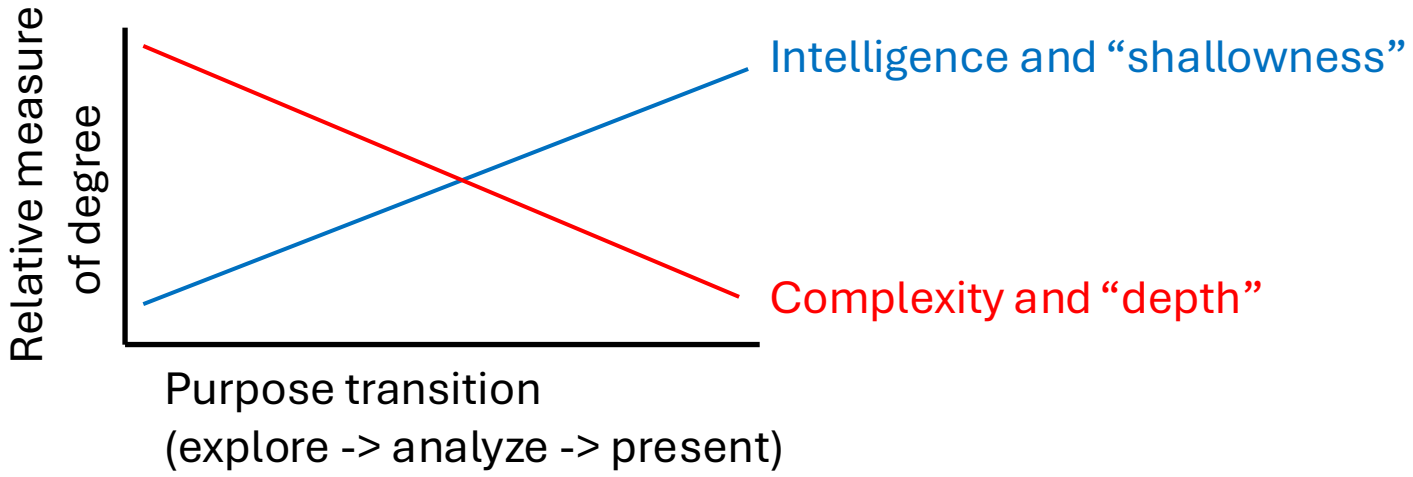
To explore? To present?



Good visualization has clear purposes in data science

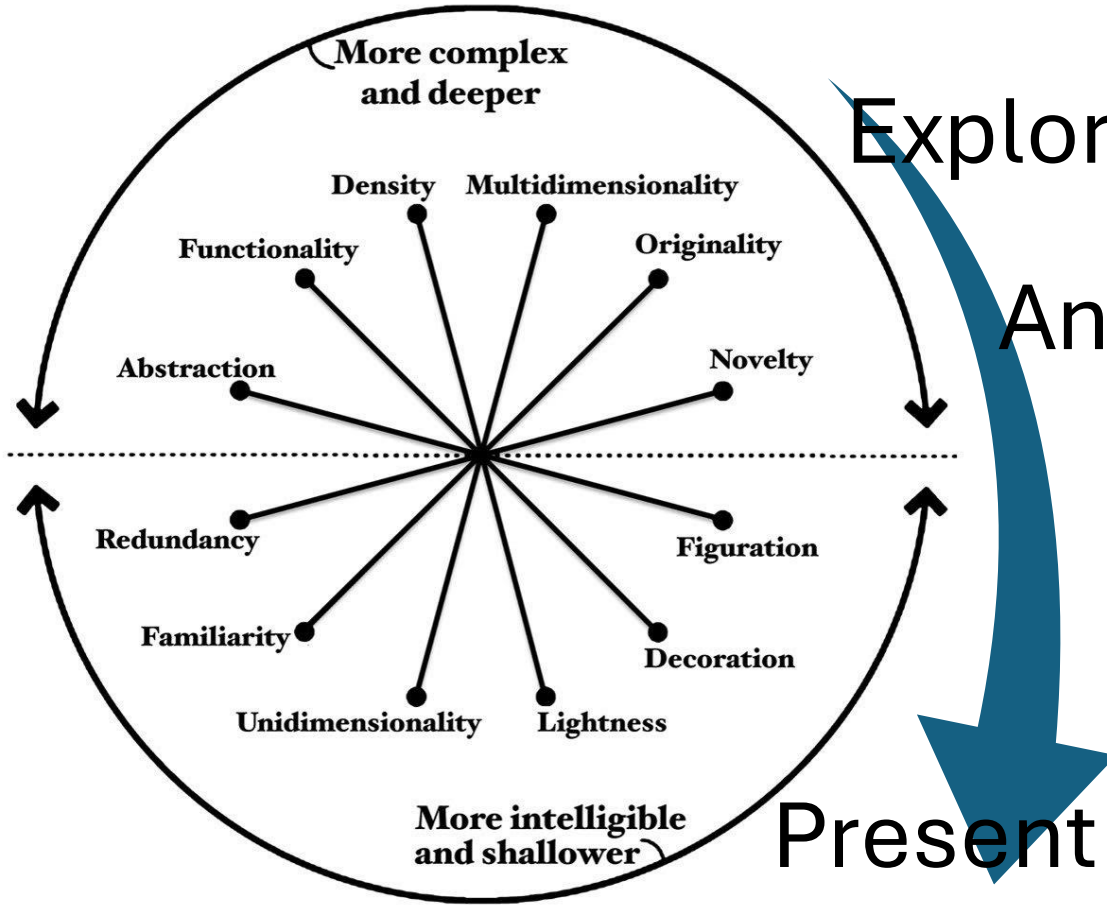
The purpose of visualization	To explore	To analyze	To present
	*Discover questions *Make decisions	*Answer the questions * Find patterns	*Present arguments

* Record



Tradeoffs in visualization

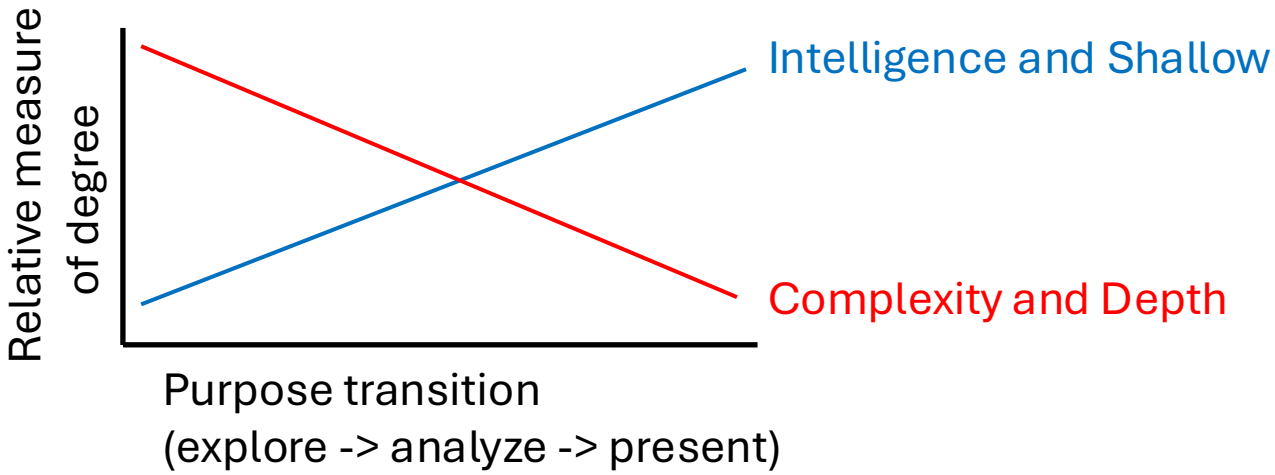
Cairo's Visualization Wheel



Explore

Analyze

Present



https://ryanwingate.com/visualization/guidelines/visualization-wheel/

The embedded “instructions” to read scientific visualization

The three “elements” of visualization

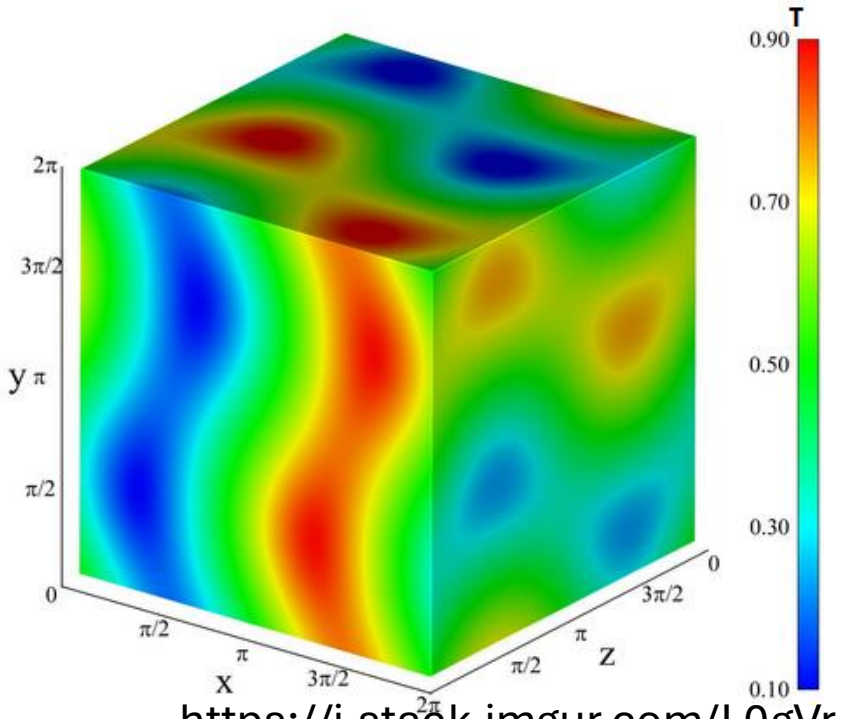
- The spatial substrate
- The graphics content
- The graphics design

Figure-specific instructions will be discussed throughout the course for each figure style

Spatial substrate

How space is used to represent data

- 2D along x-axis and y-axis
- 3D along x-, y-, and z-axis
- 4D along x-, y-, z-, and an additional axis (hard to understand)
  - Time?
  - Color?



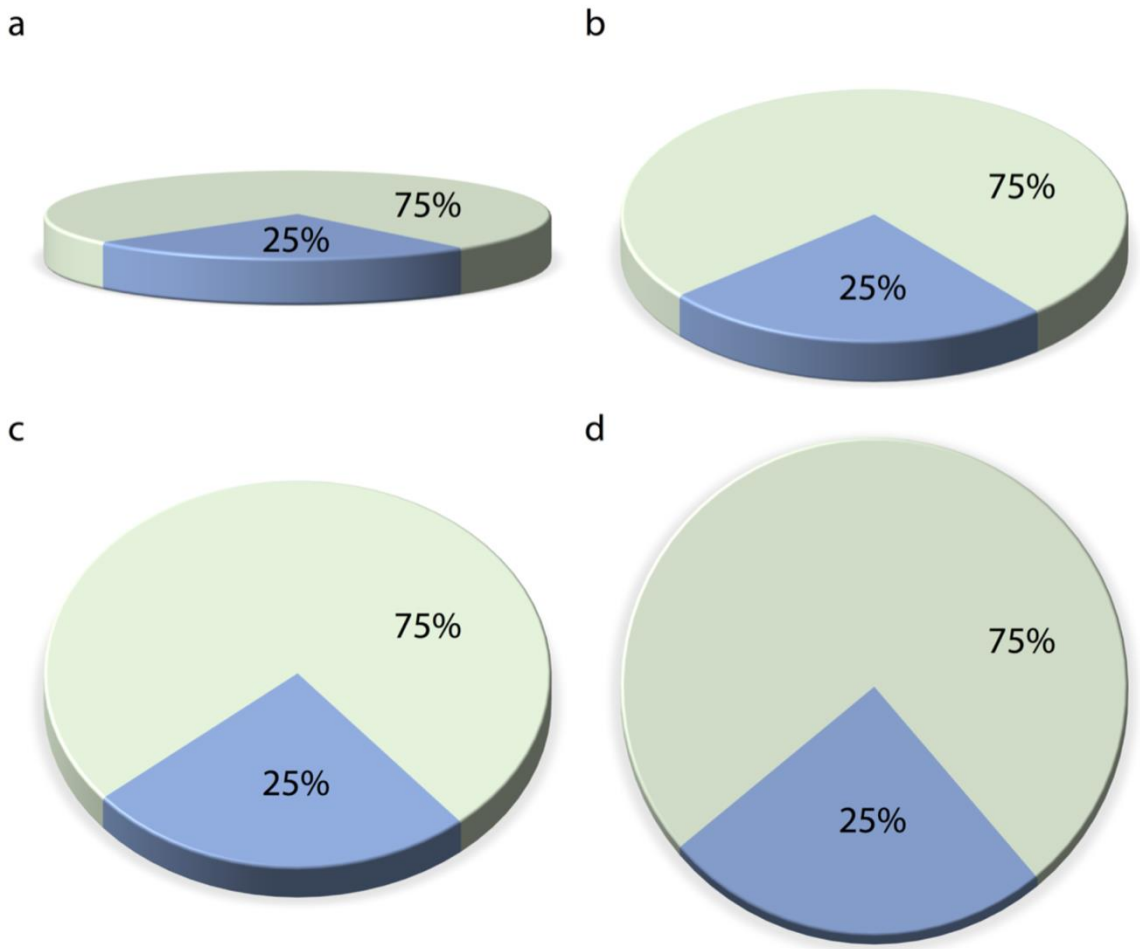
https://i.stack.imgur.com/L0gVr.pn

Implications for 3D

3D projected into 2D

- The projection distorts the data and the human visual system cannot fully correct for the distortion
- 3-D often <u>only to decorate – bad</u>

The same pie plot from four different angles

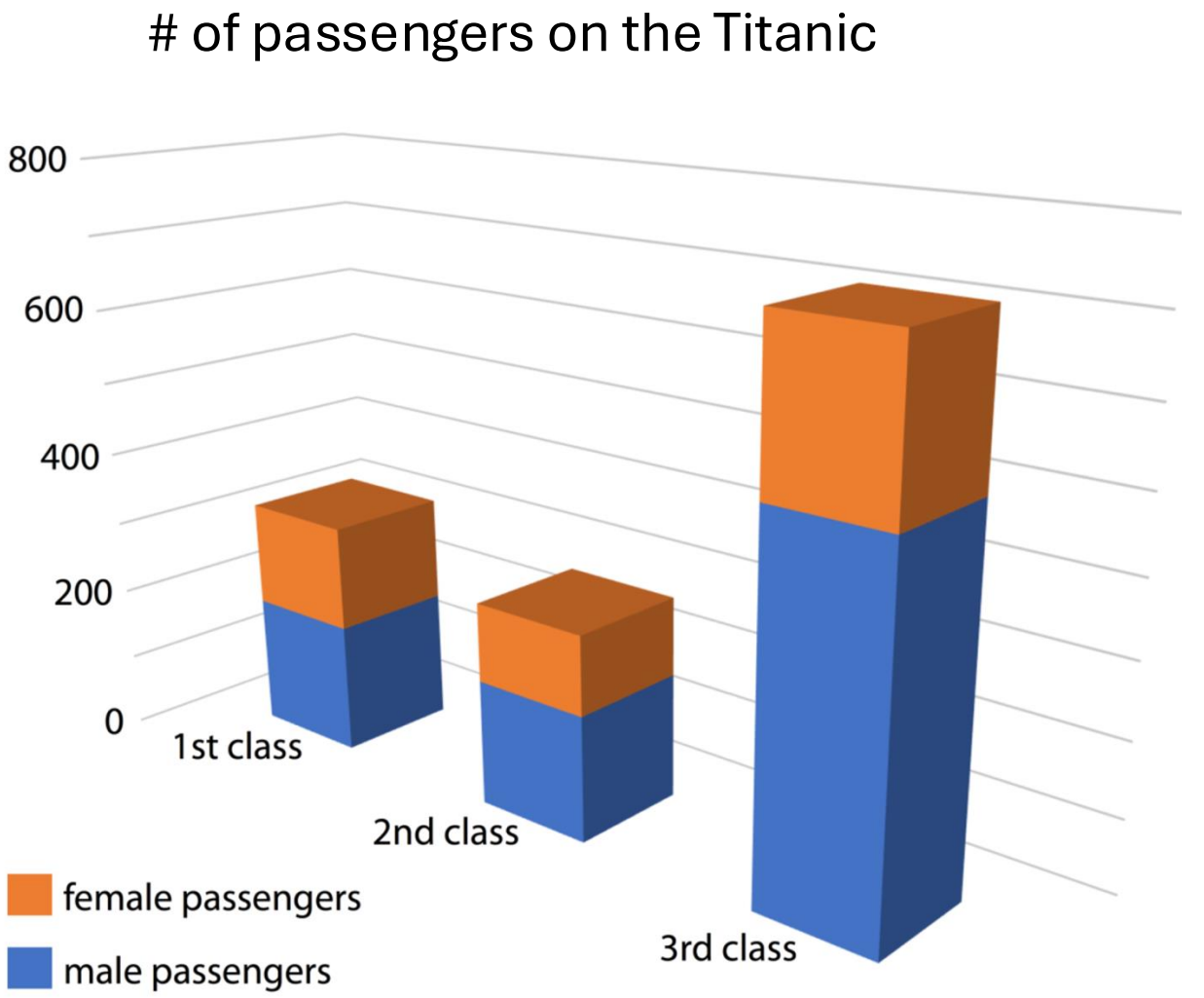


Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed.

Implications for 3D

3D projected into 2D

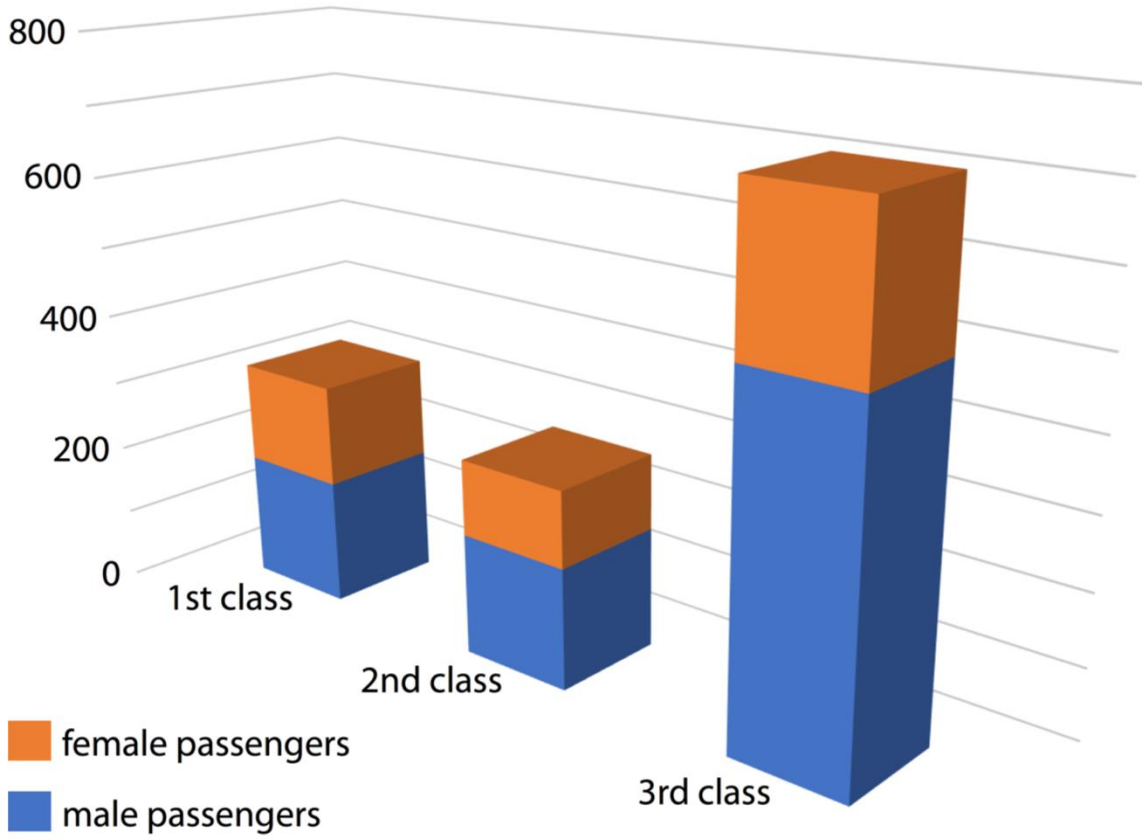
- The projection distorts the data and the human visual system cannot fully correct for the distortion
- 3-D often <u>only to decorate – bad</u>



Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed

2D: generally better

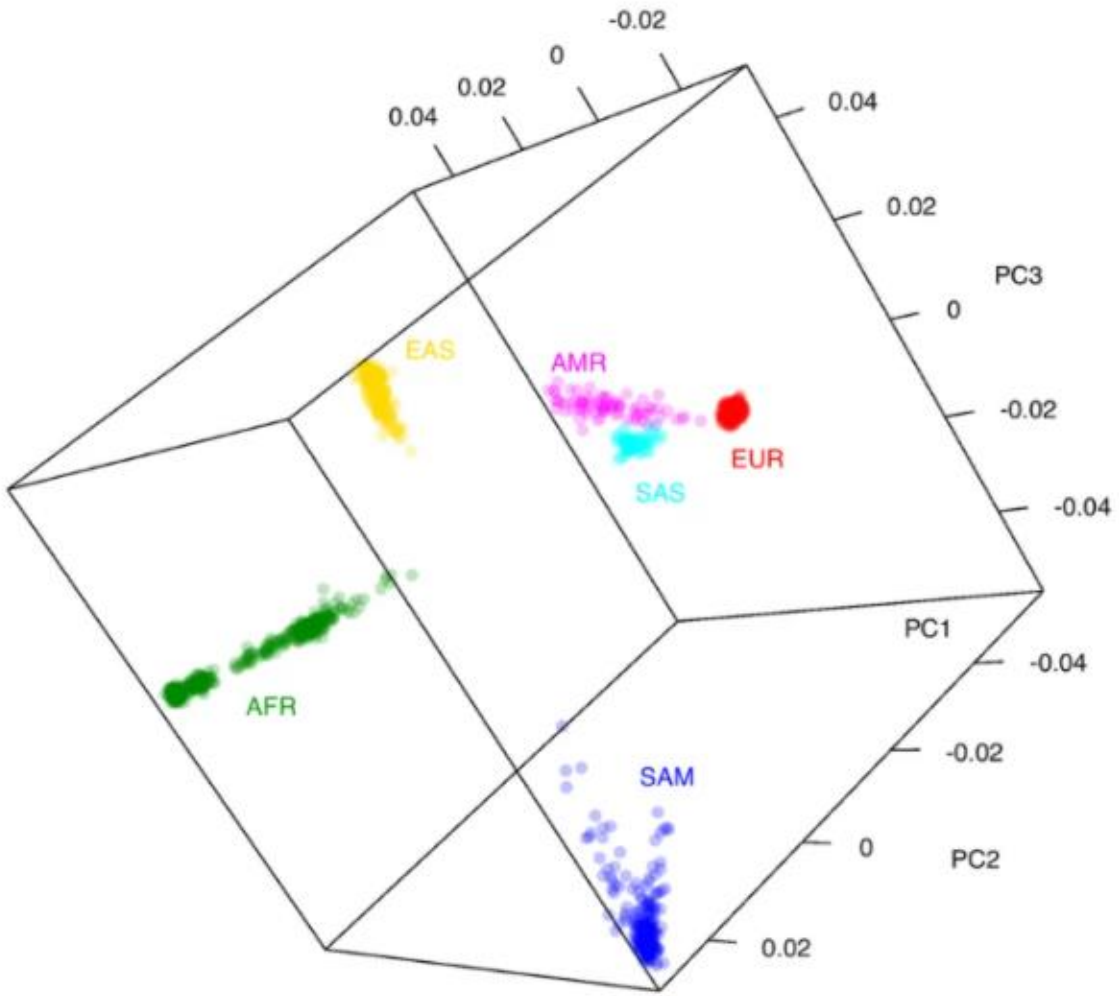
of passengers on the Titanic



Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed

3D still sometimes helps

PCA plot on genome data of Samoan and other populations



Minster et al. Nat. Gen. 2016

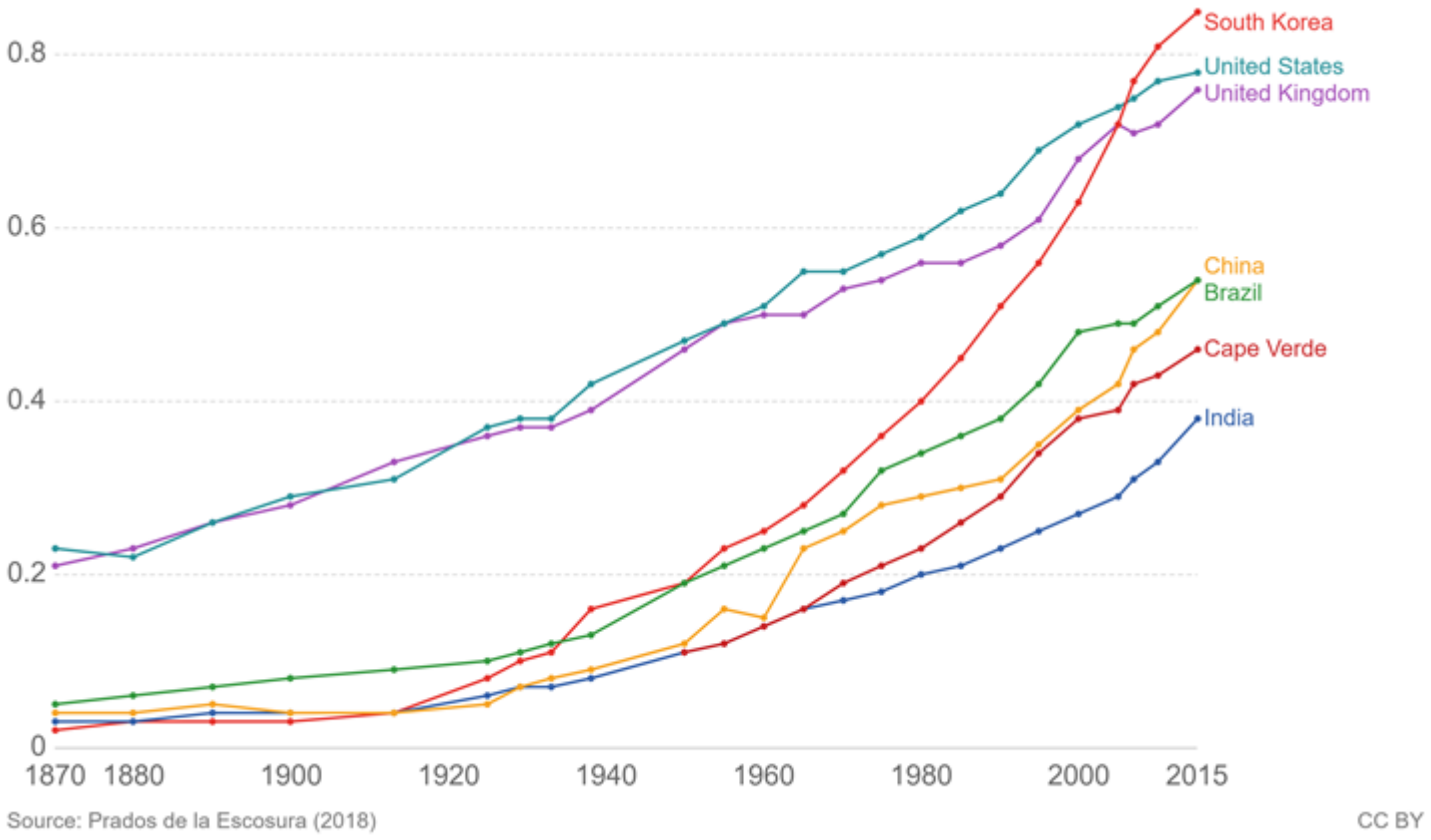
Graphics content (example)

Points to represent data, e.g.

- Markers or colors to indicate different groups

Lines

- Real line to represent trends



The graphics content

Points to represent data

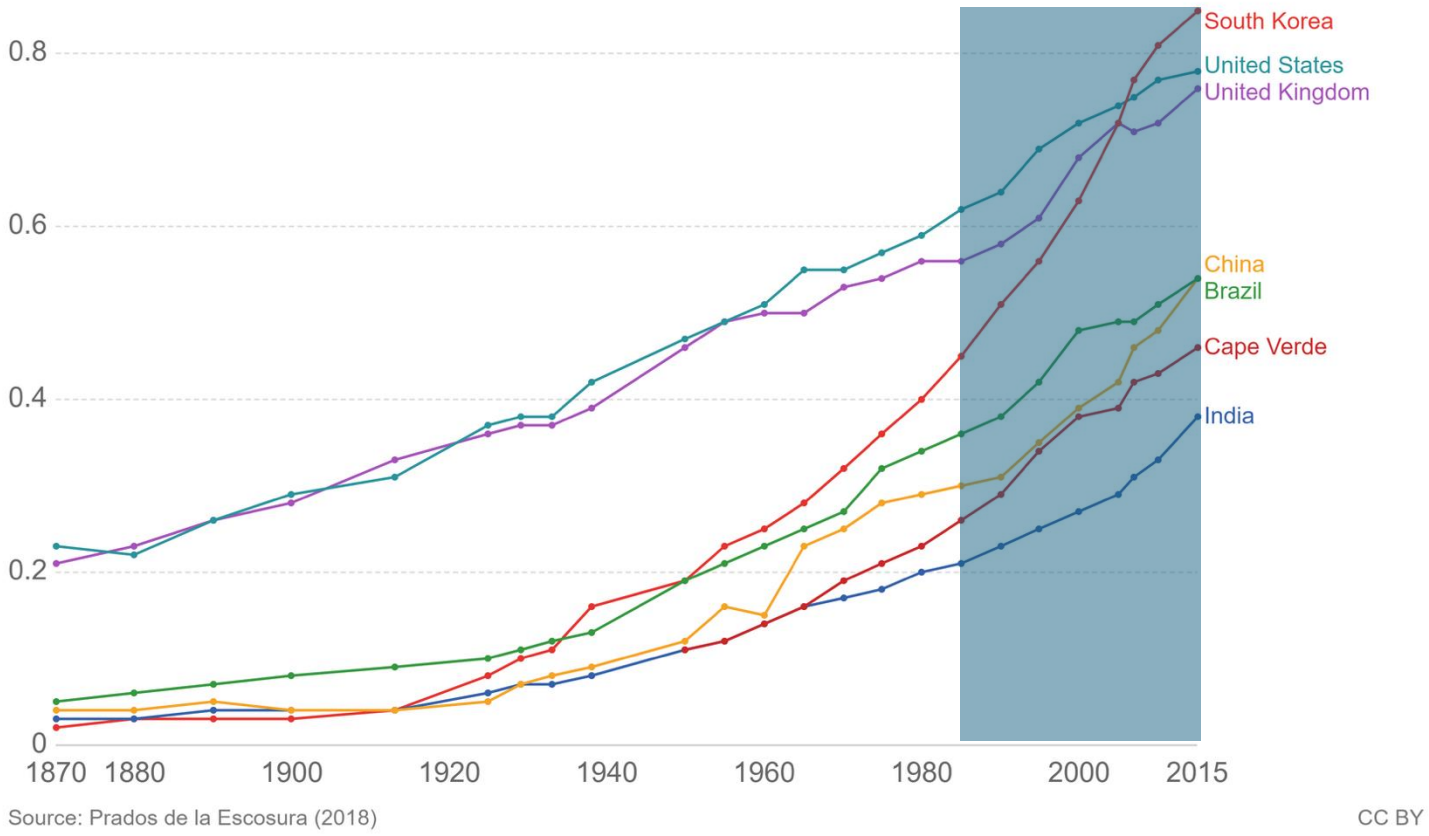
- Marker to indicate different groups

Lines

- Real line to represent trends
- Dotted line to help interpret the trends

e.g. HIHD at 1940

- Arrow when there is directionality
- Shapes to highlight an area



The graphics content

Points to represent data

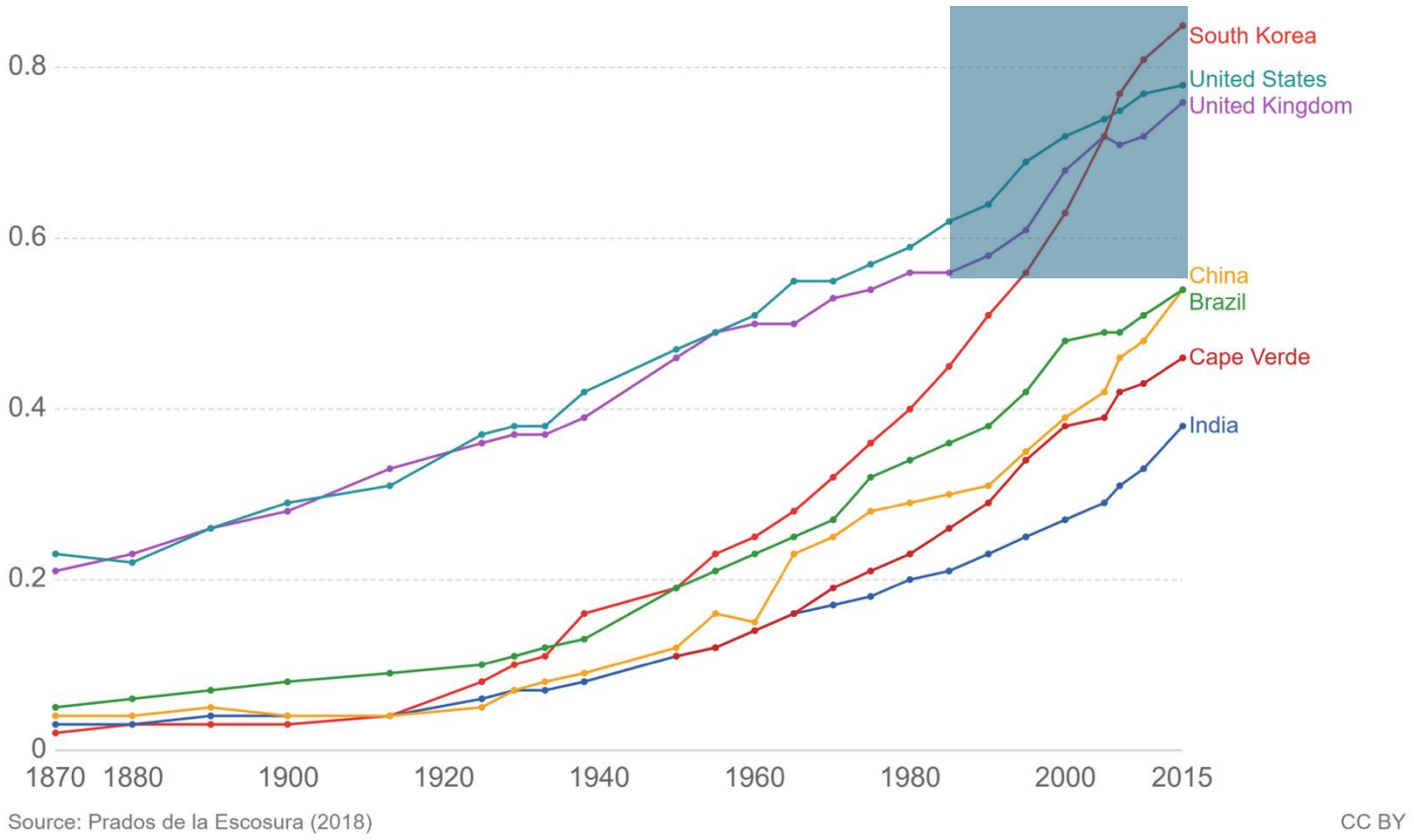
- Marker to indicate different groups

Lines

- Real line to represent trends
- Dotted line to help interpret the trends

e.g. HIHD at 1940

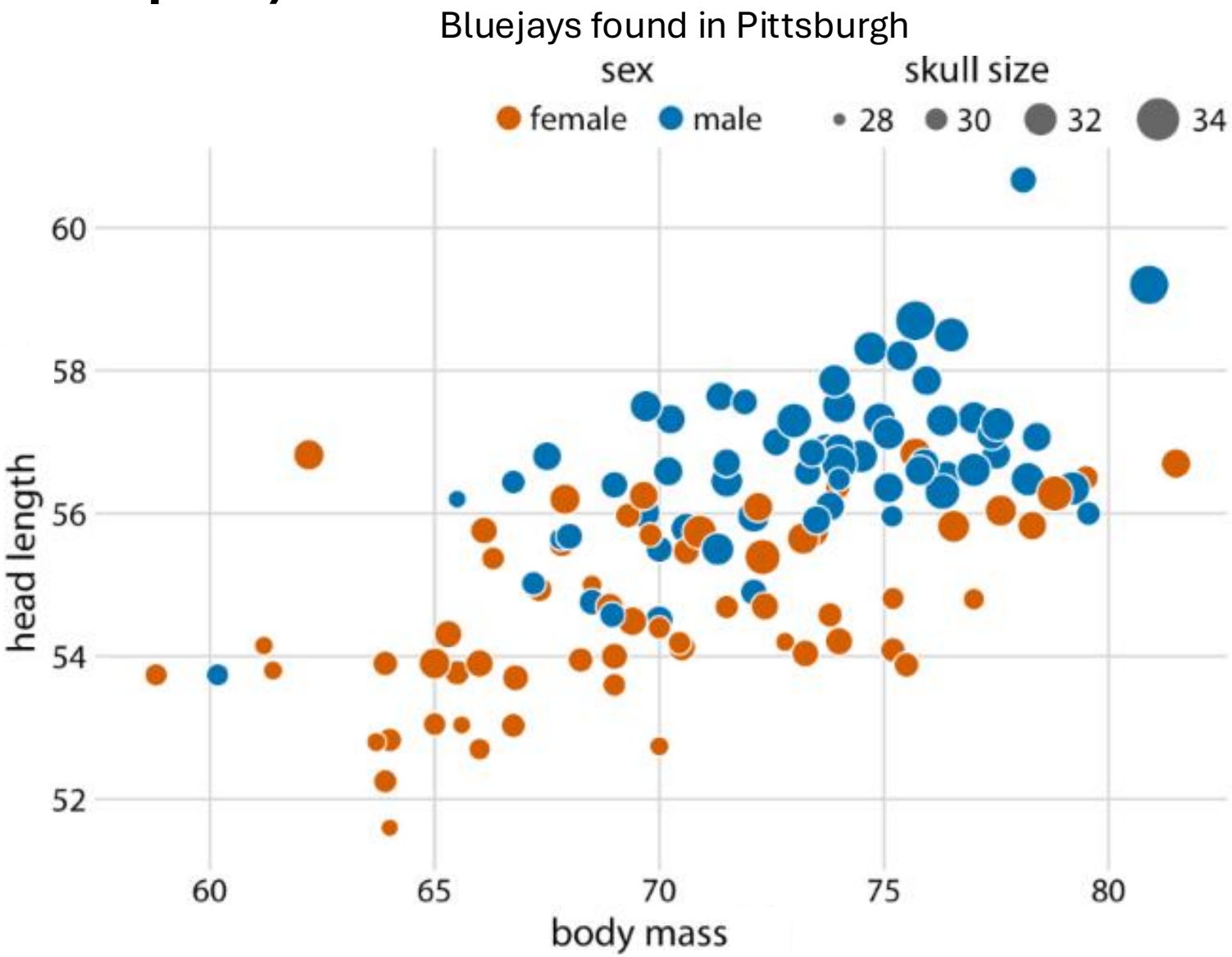
- Arrow when there is directionality
- Shapes to highlight an area



Graphics design (example)

X- and y-axis

- Ticks, tick-labels and axis-labels
  - Needs units whenever applicable

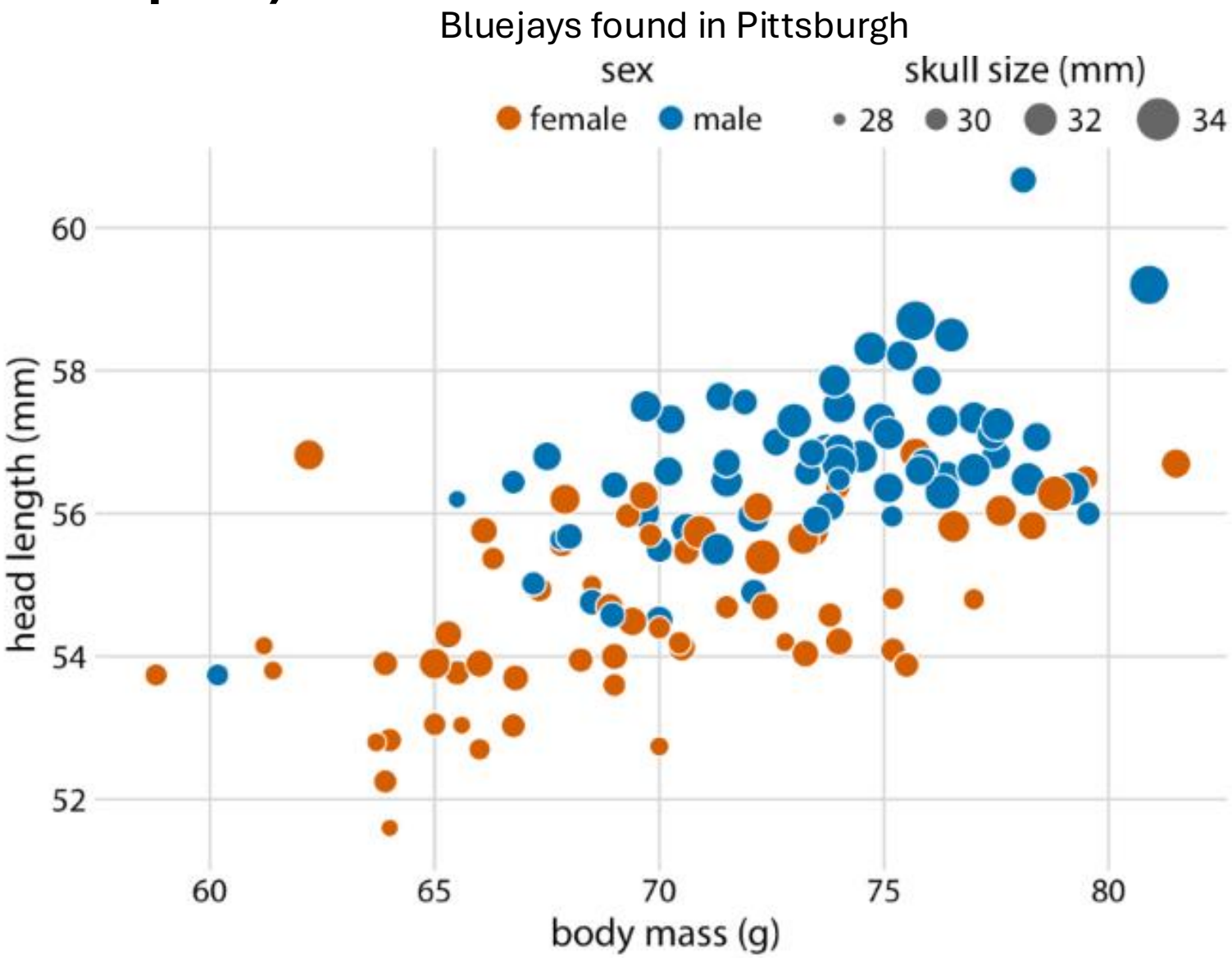


Adapted from Fundamentals of Data Visualization, Wilke, O'Reilly, 1st

Graphics design (example)

X- and y-axis

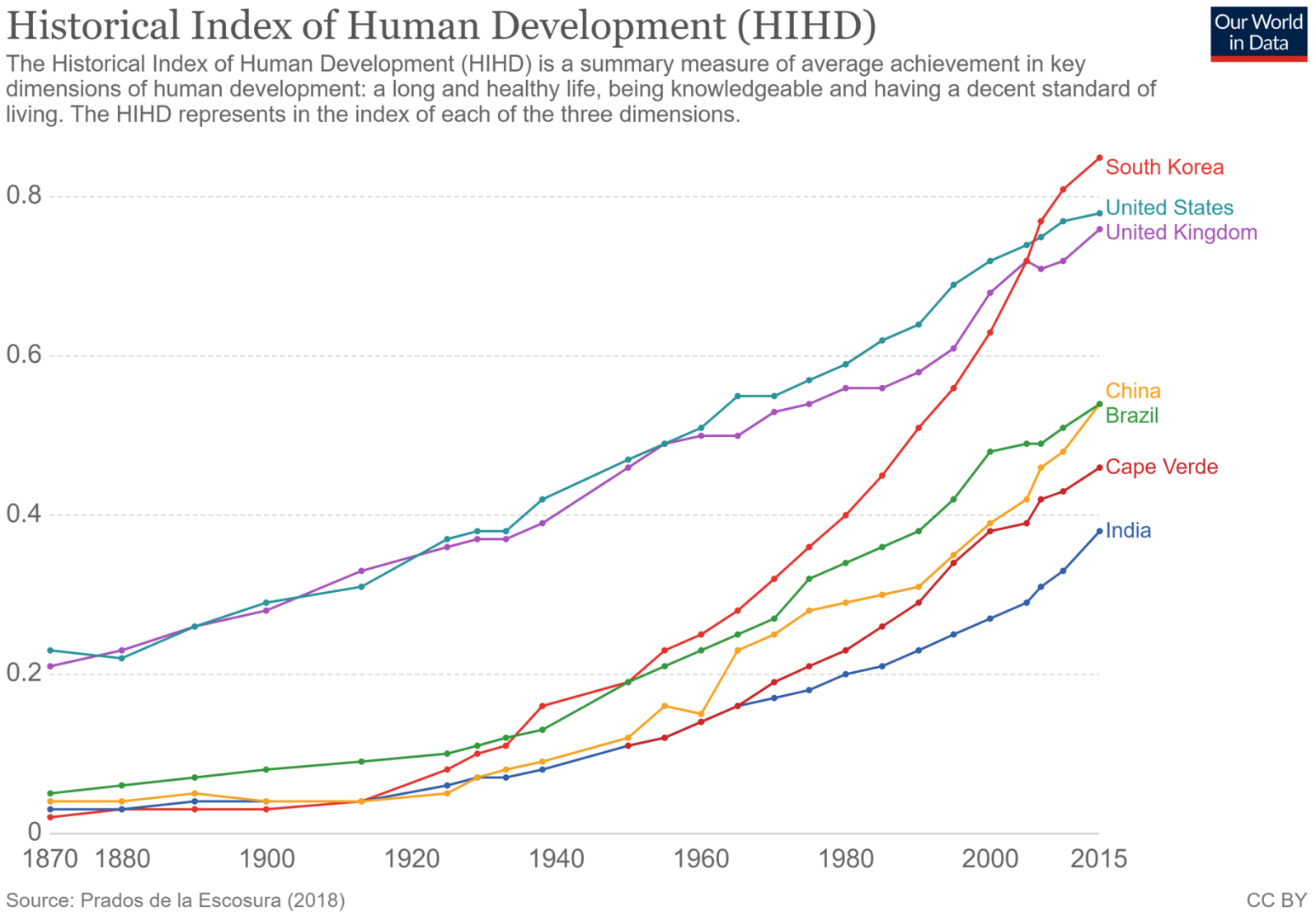
- Ticks and labels
  - Needs units whenever applicable



Adapted from Fundamentals of Data Visualization, Wilke, O'Reilly, 1st

Graphics design (example)

Title (what the figure is about) and caption (what point it makes)

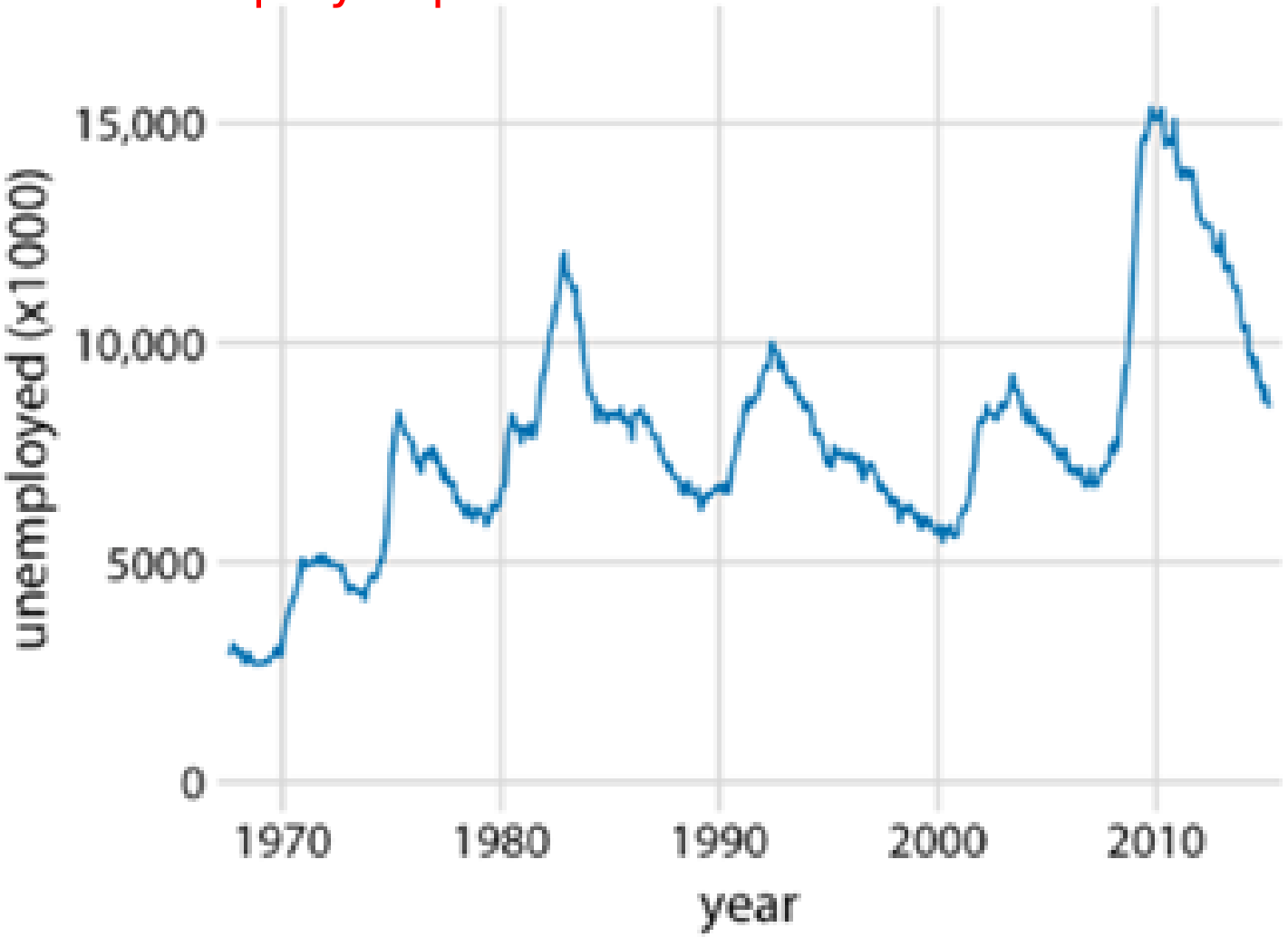


Graphics design (example)

Across, title, caption, and labels

- Check redundancy

# of unemployed persons from 1970 to 2015

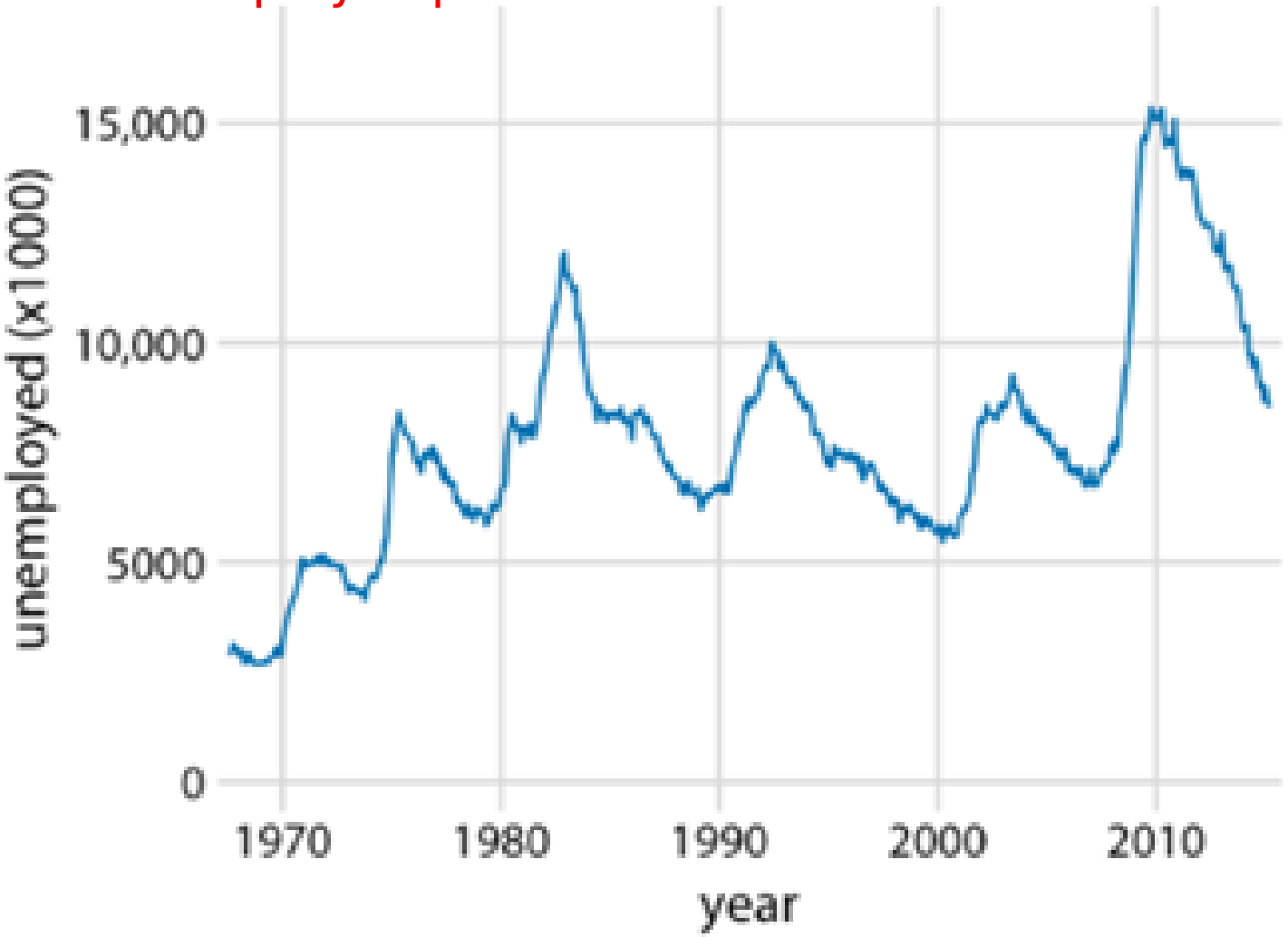


Graphics design (example)

Across, title, caption, and labels

- Check redundancy

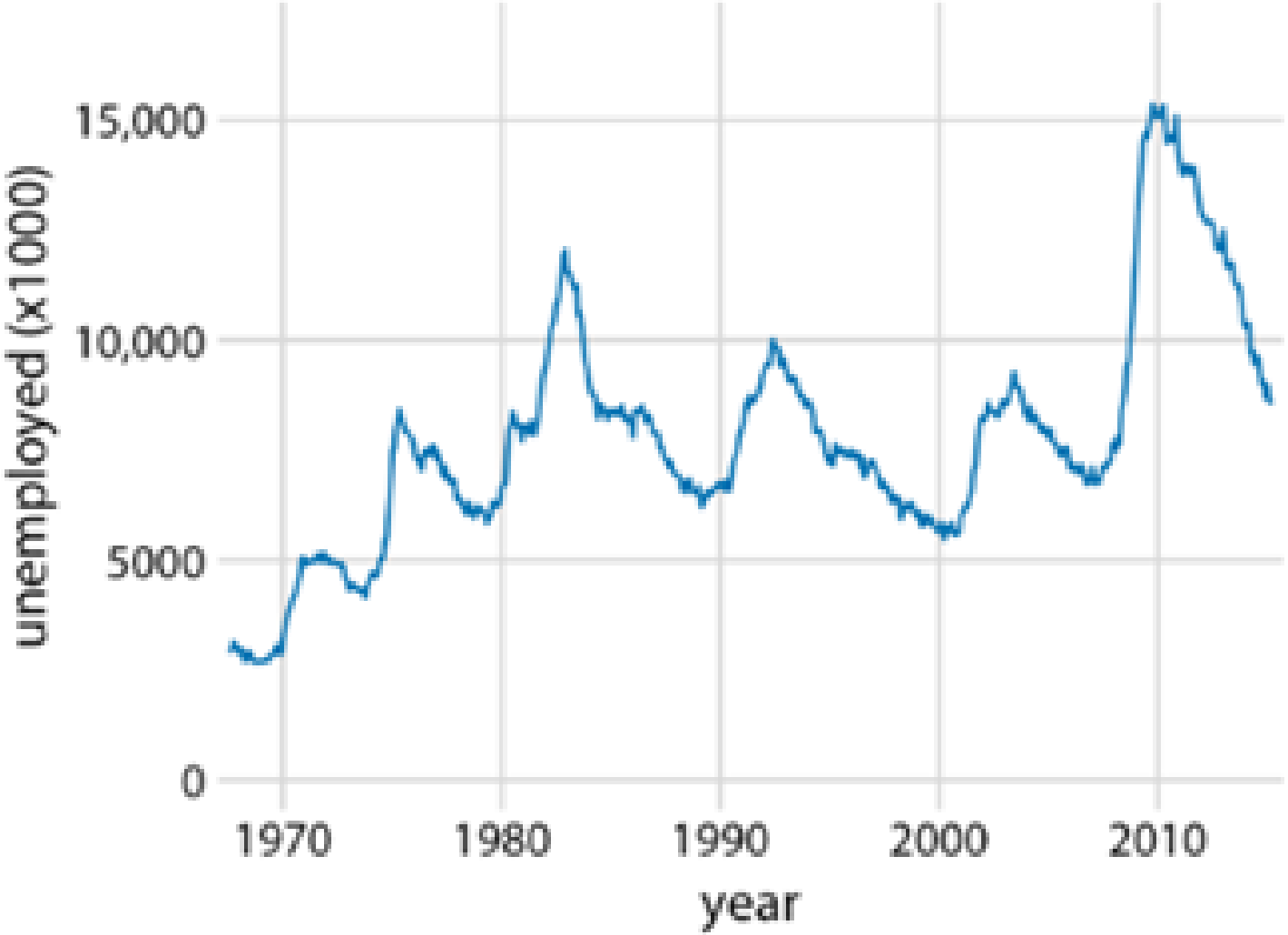
# of unemployed persons



Graphics design (example)

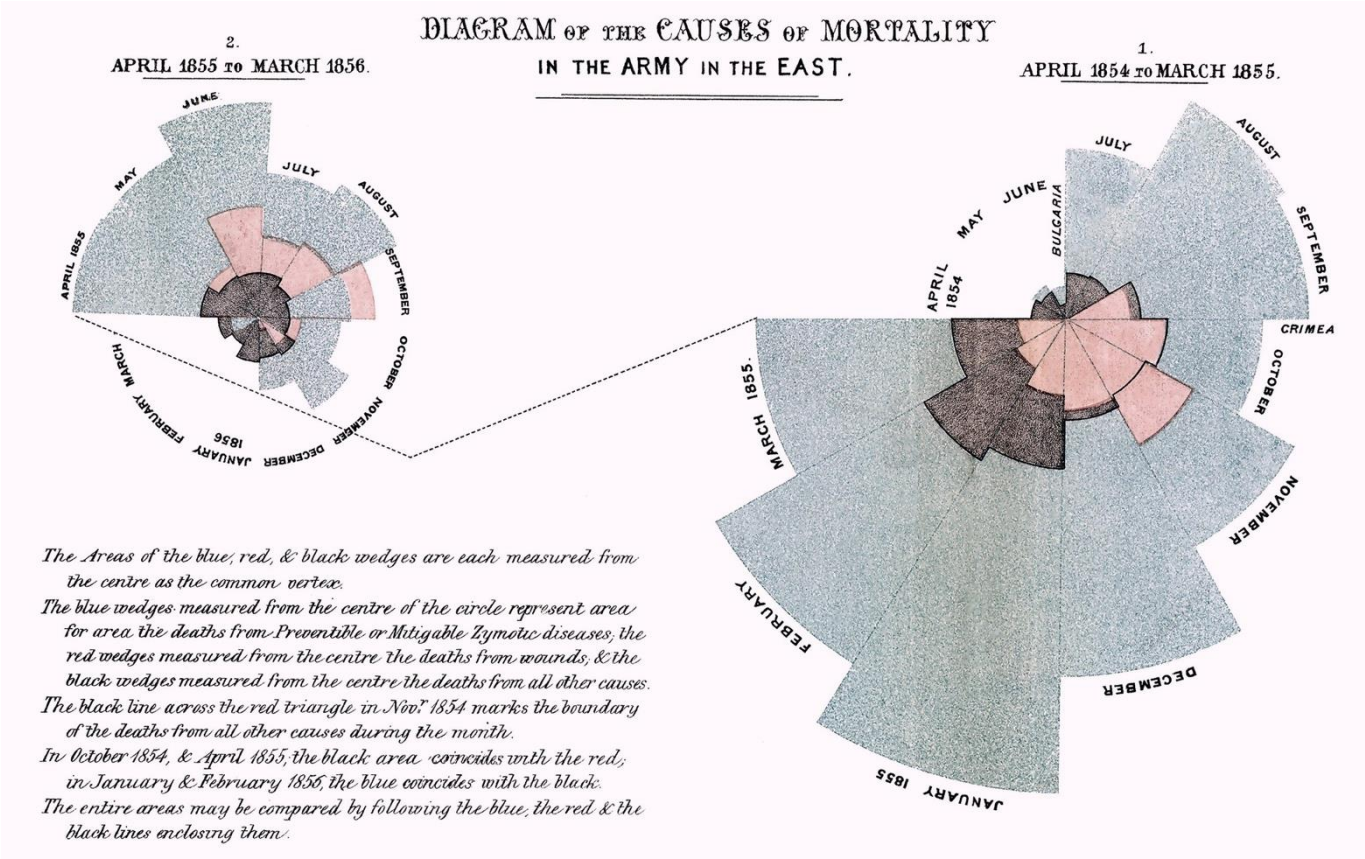
Across, title, caption, and labels

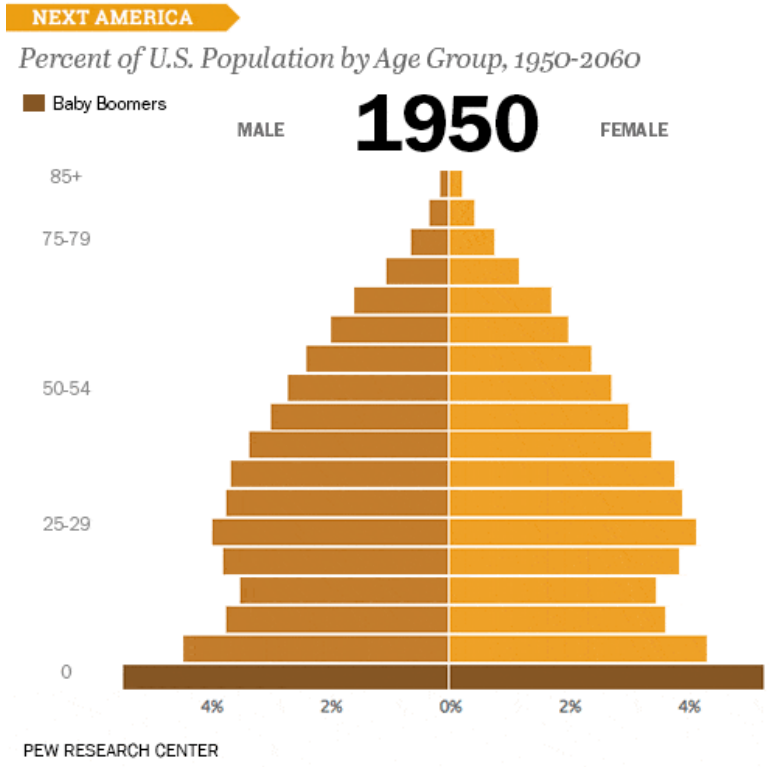
- Check redundancy

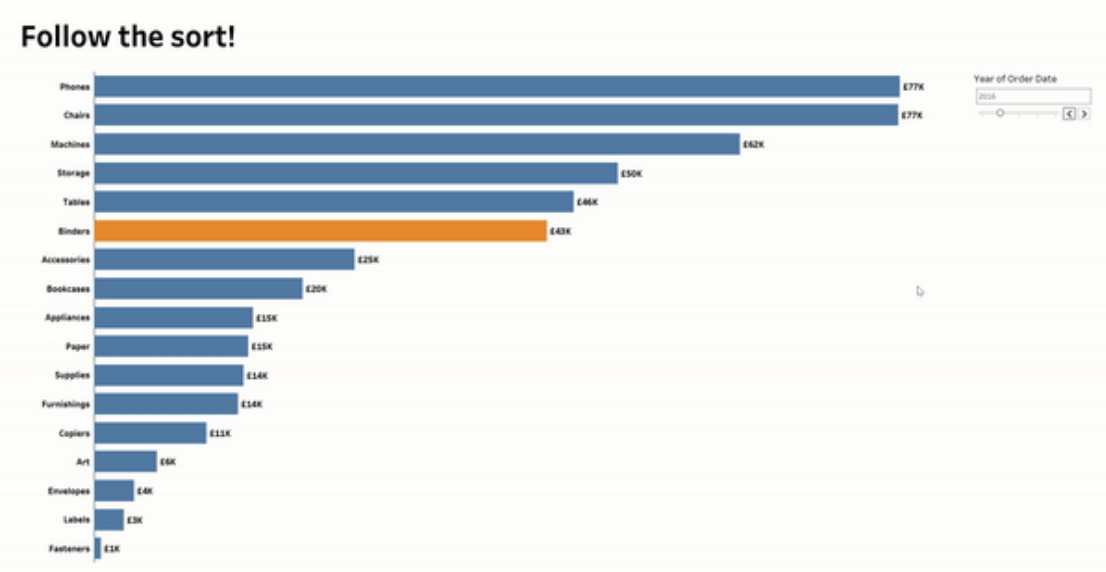


Example

Identify the visualization elements for each







Some principles for good visualizations

- Choose substrate/content judiciously
  - Depends on data/variable type
  - We will discuss this indetail for many visualization/data types
- Keep data-ink ratio high
- Avoid clutter/“chartjunk”
- Use color judiciously
- Avoid deception
- Highlight your message/have a clear purpose!
- Don’t trust the defaults

Good visualization keeps the data-ink ratio high

"Above all else, show the data."
-Edward Tufte, 1983

$$\text{Data-ink ratio} = \frac{\text{Data-ink}}{\text{Total ink used to print the graphic}}$$

(=the non-erasable core of a graphic)

= **proportion of a graphic's ink devoted to the non-redundant display of data-information**

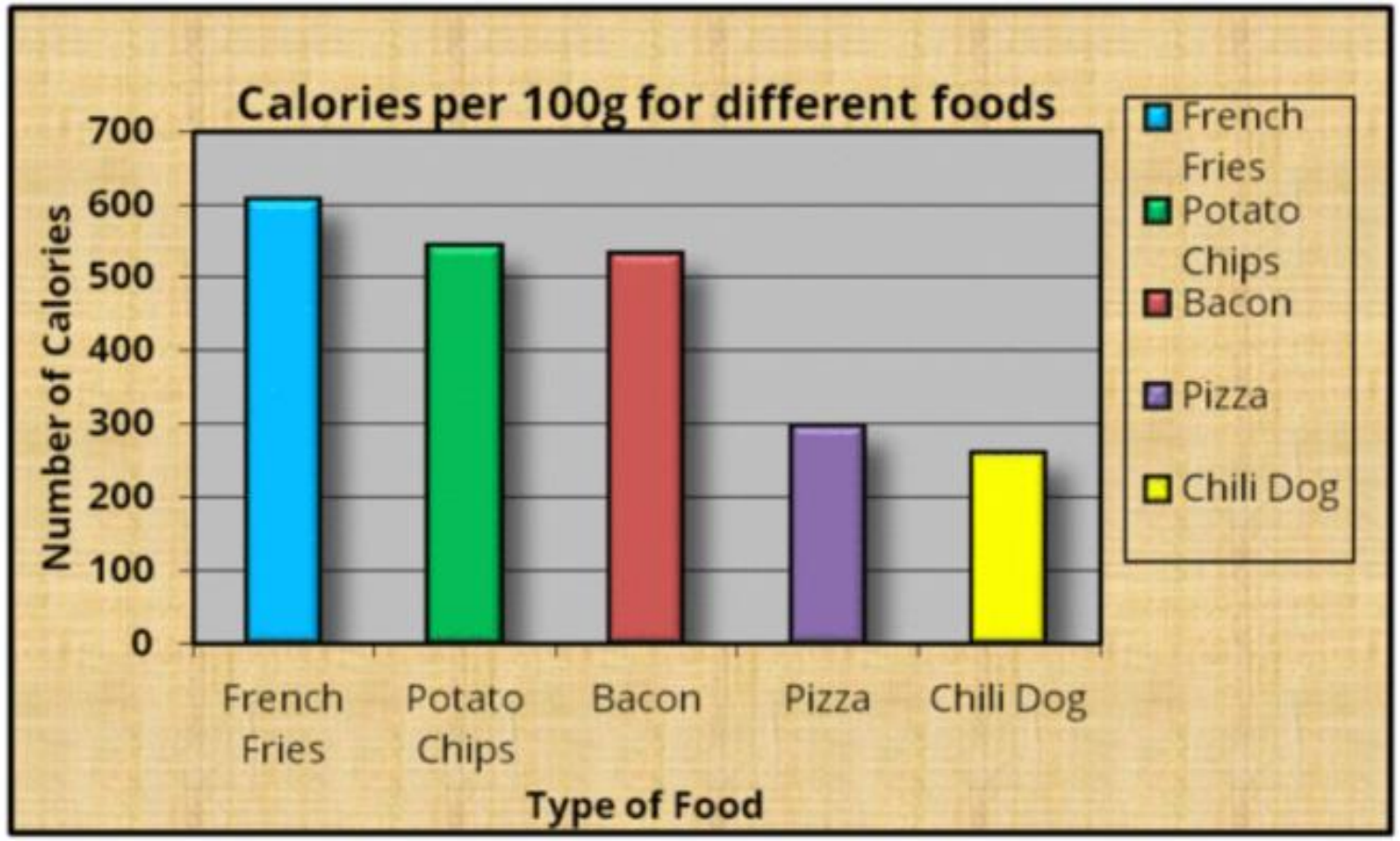
= **1.0 – proportion of a graphic that can be erased**

= proportion of the non-erasable core of a graphic
= amount of information/amount of attention needed

https://infovis-wiki.net/wiki/Data-Ink_Ratio

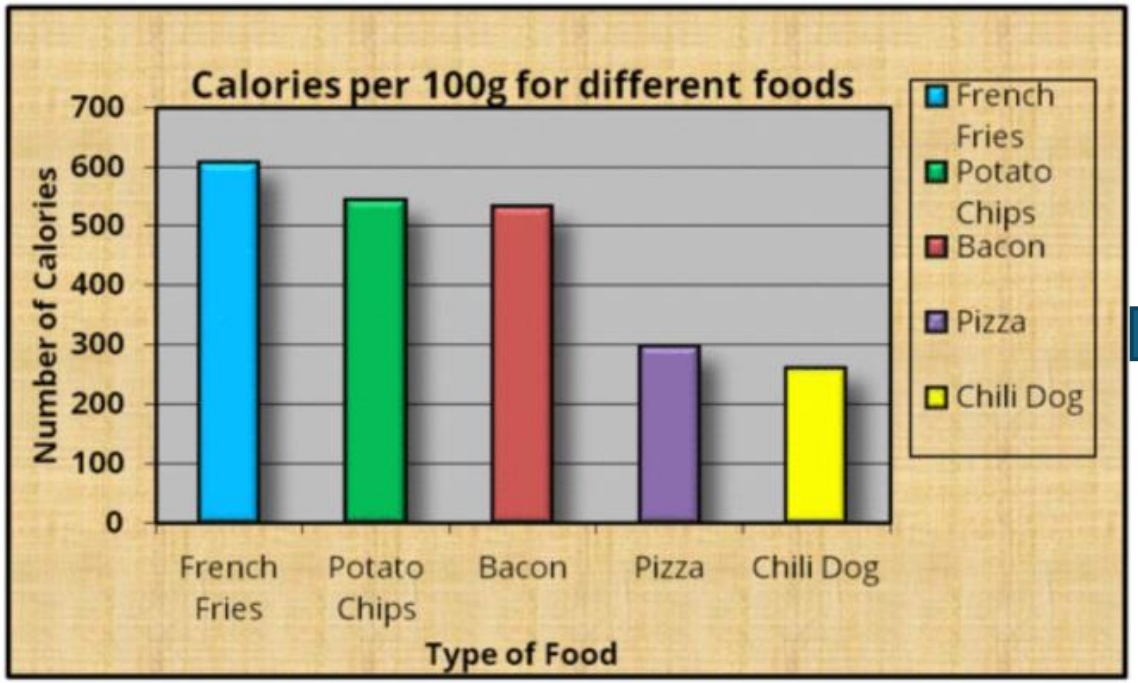
Exercise to maximize the data-ink ratio

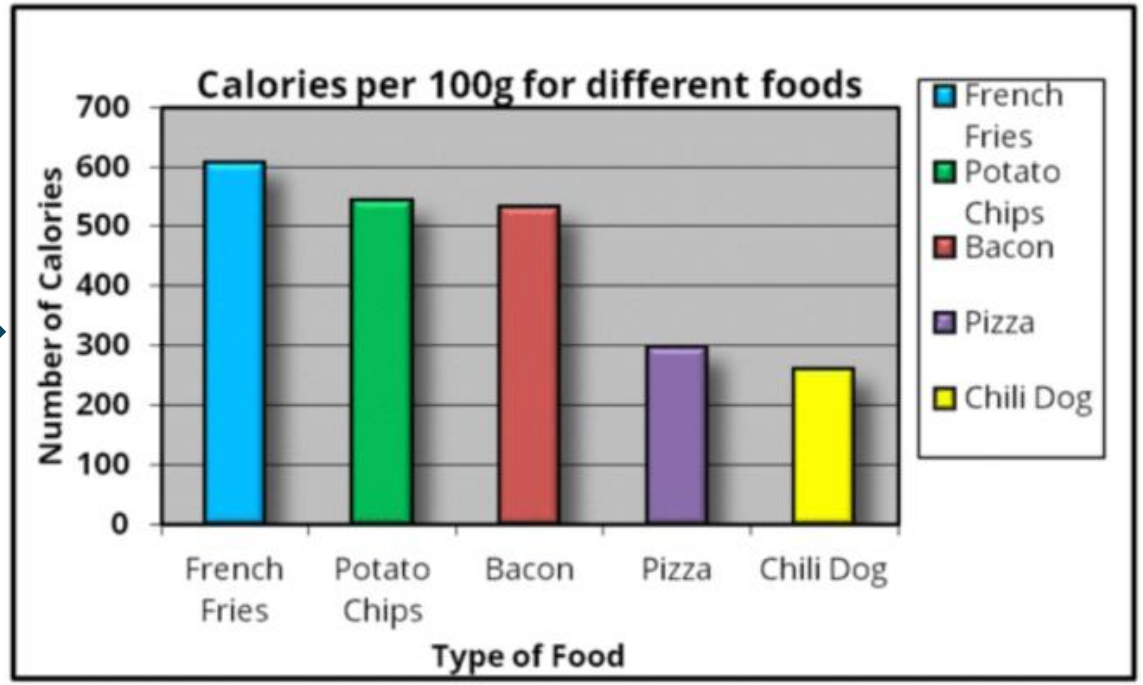
- Suppose the message is "bacon has more Calories than pizza or a chili dog but less than French fries and potato chips"
- How to improve this figure?



https://www.darkhorseanalytics.com/blo

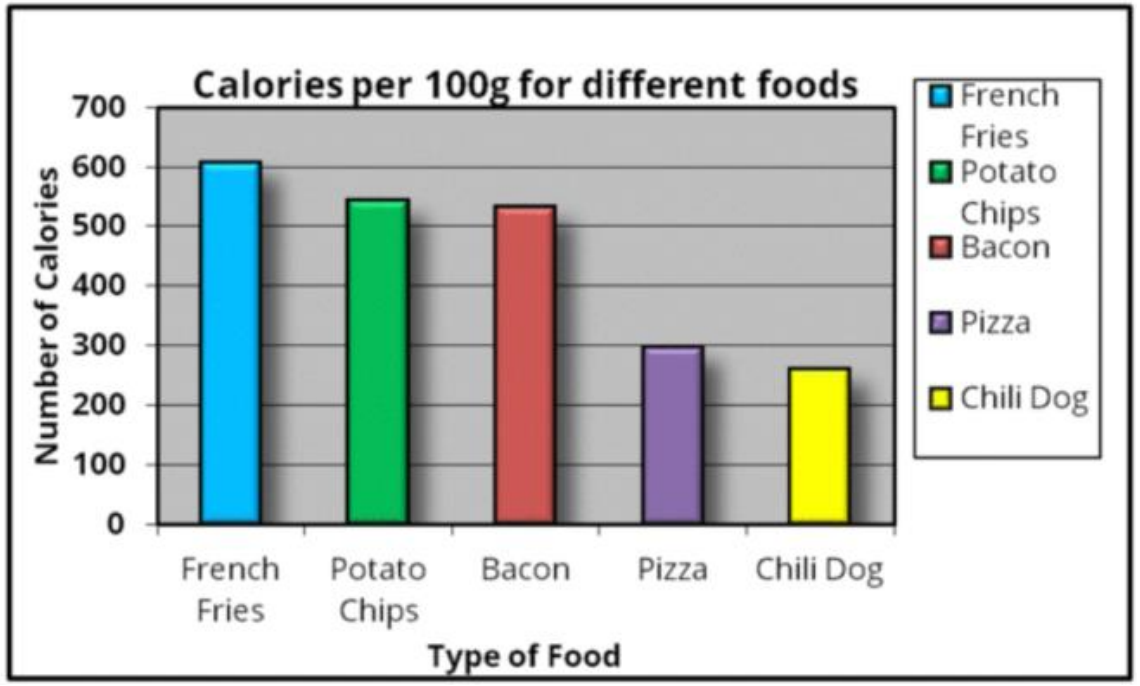
Exercise to maximize the data-ink ratio

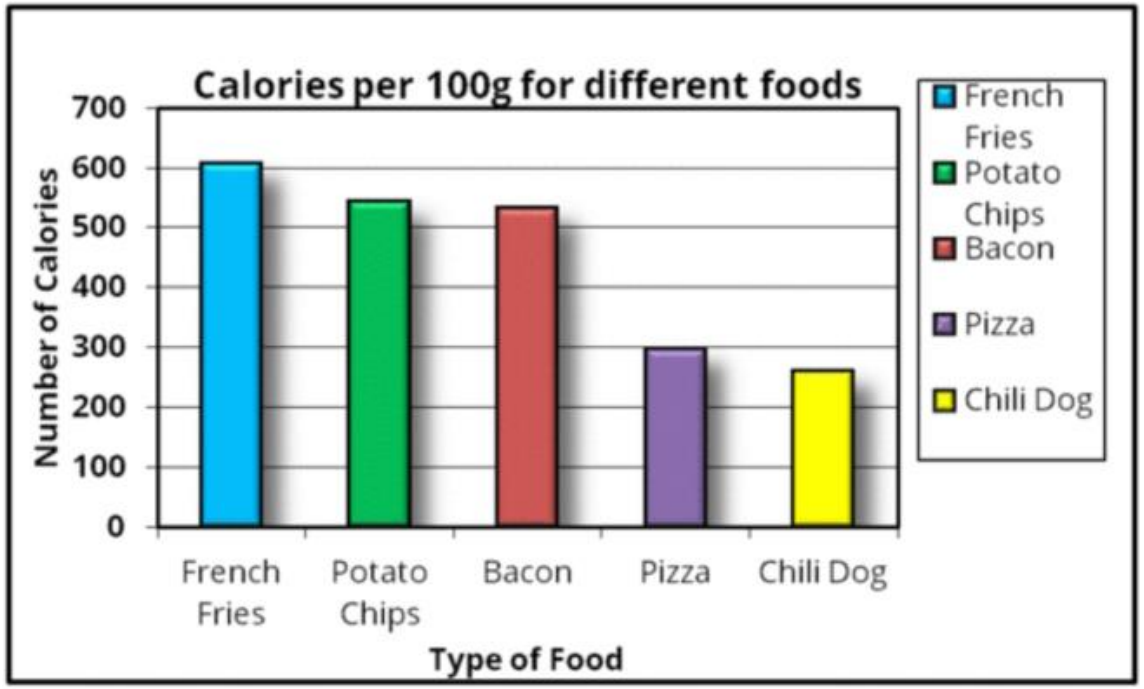




https://www.darkhorseanalytics.com/blo

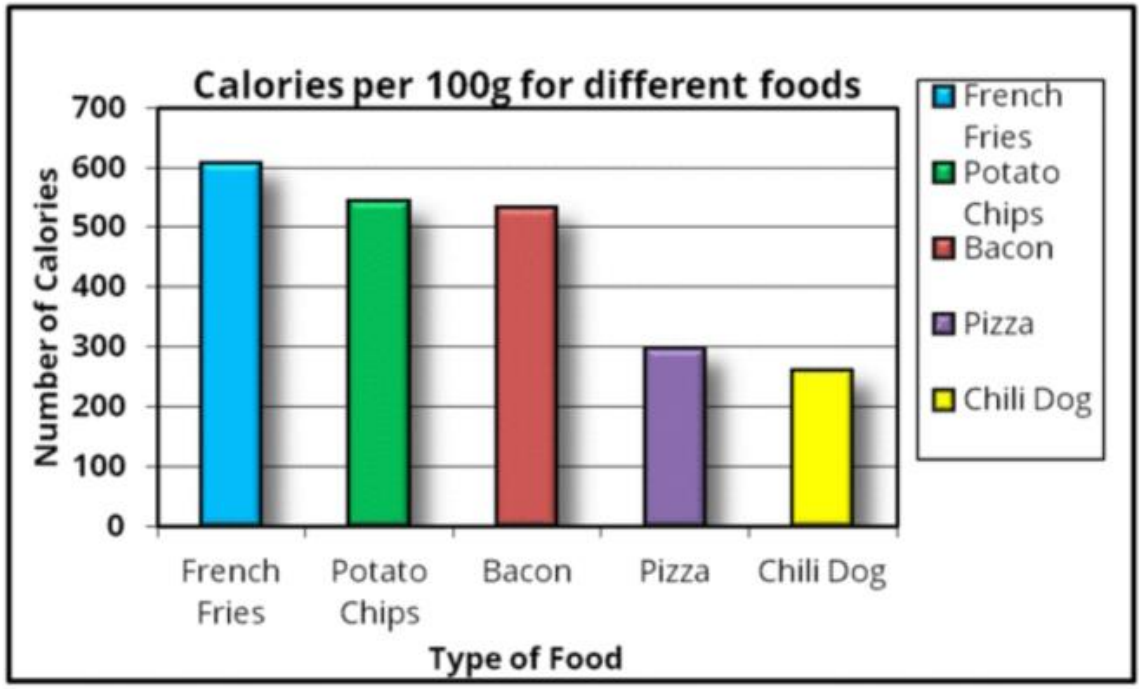
Exercise to maximize the data-ink ratio

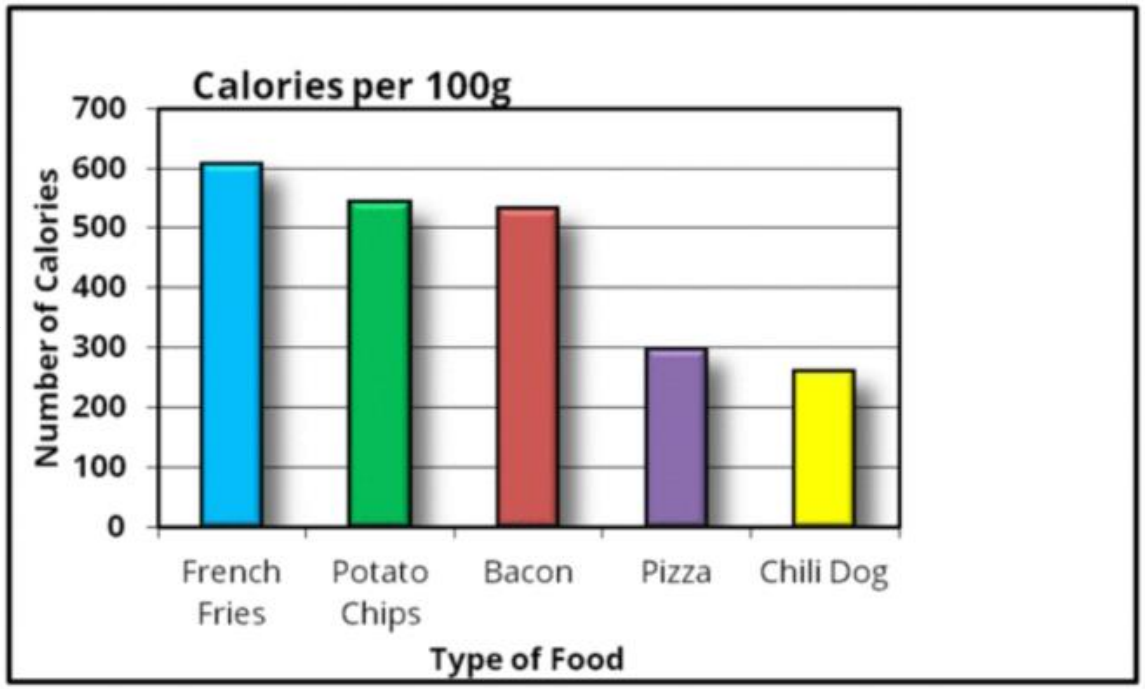




https://www.darkhorseanalytics.com/blo

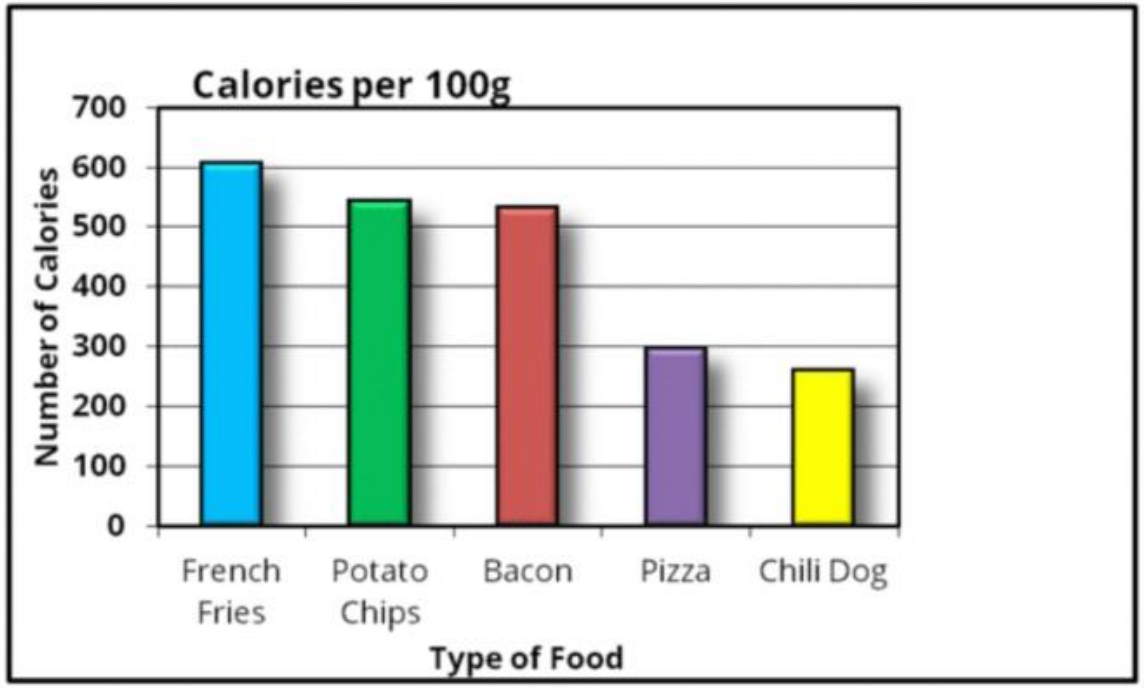
Exercise to maximize the data-ink ratio

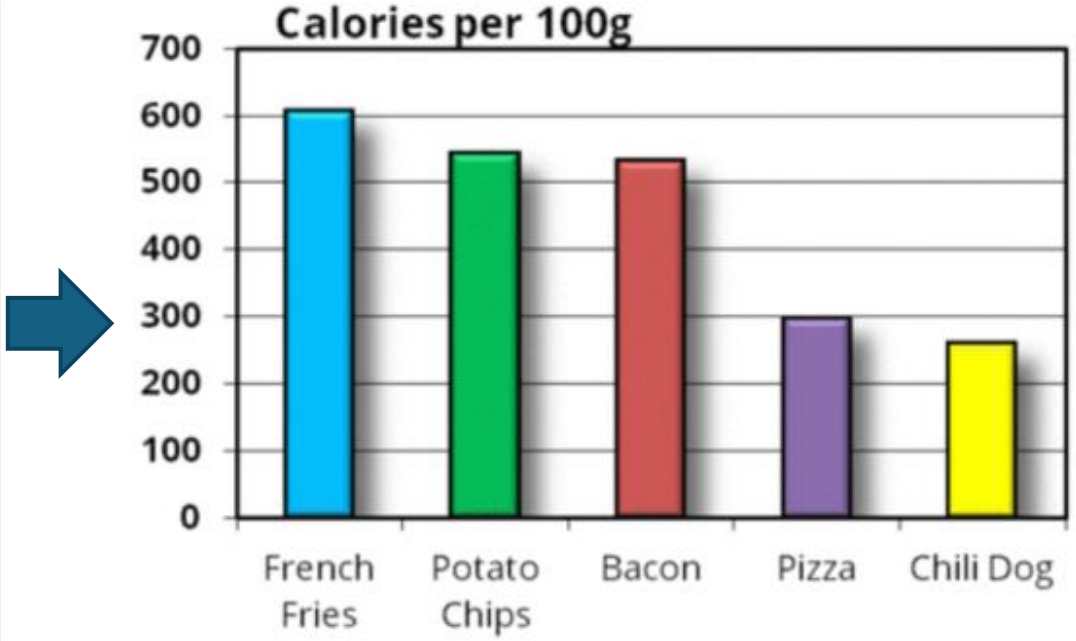




https://www.darkhorseanalytics.com/blo

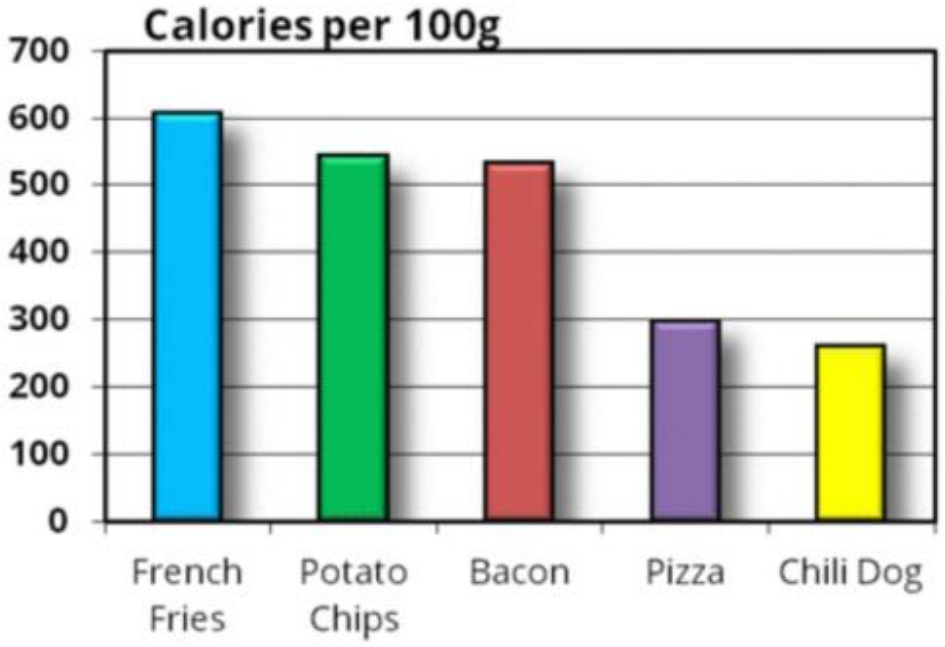
Exercise to maximize the data-ink ratio

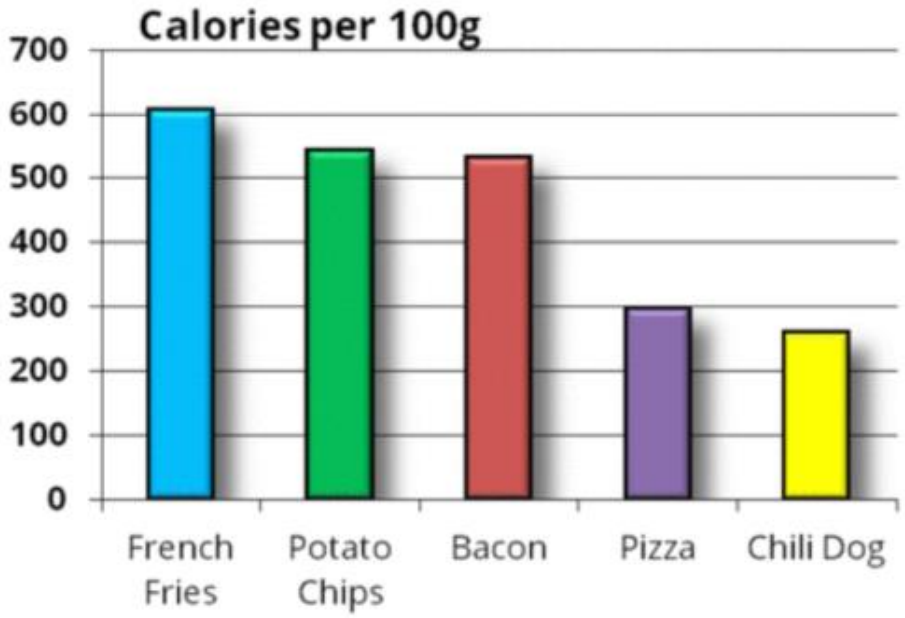




https://www.darkhorseanalytics.com/blo

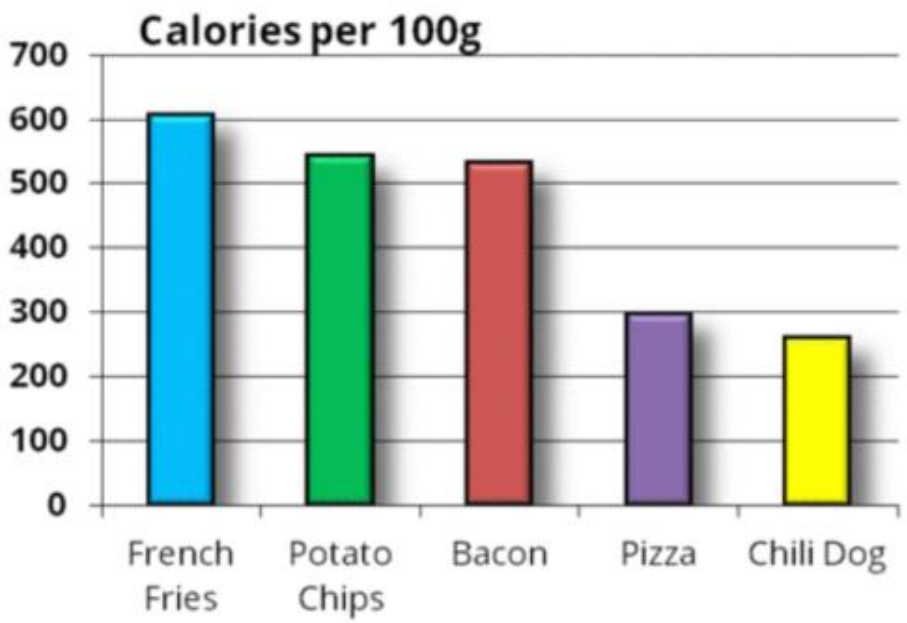
Exercise to maximize the data-ink ratio

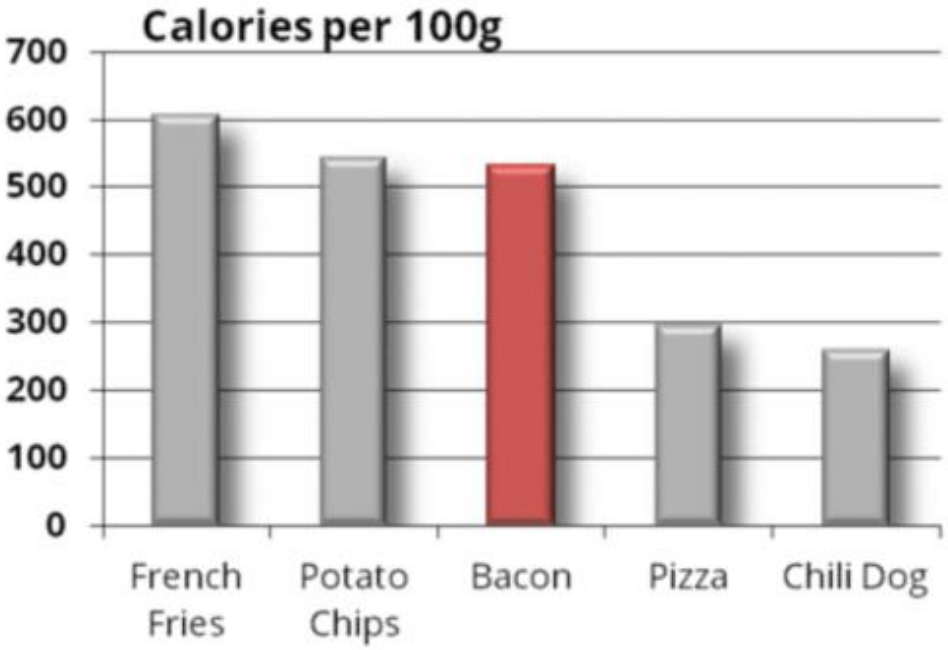




https://www.darkhorseanalytics.com/blo

Exercise to maximize the data-ink ratio

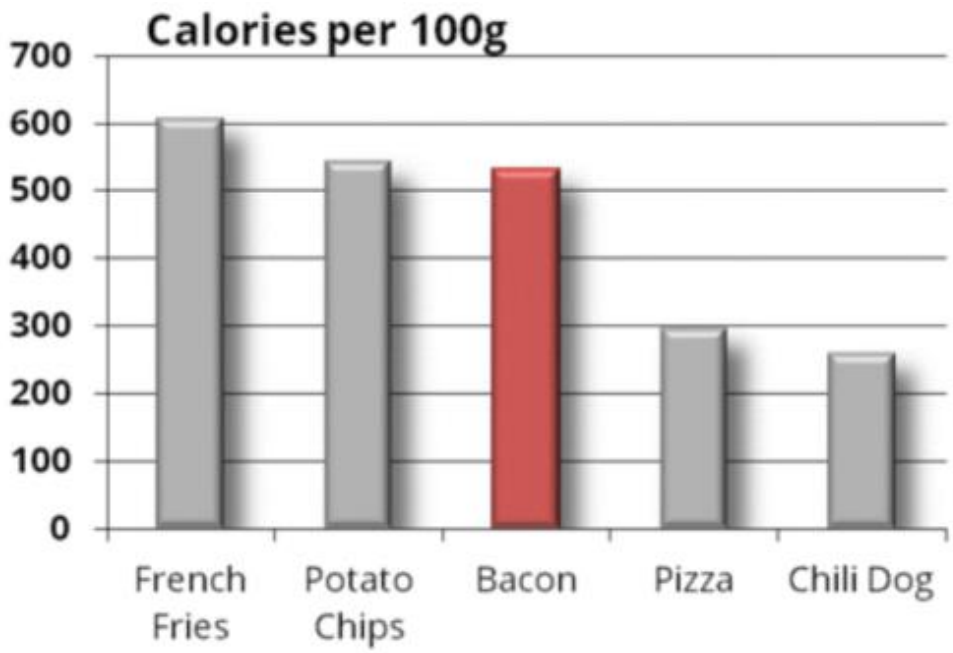


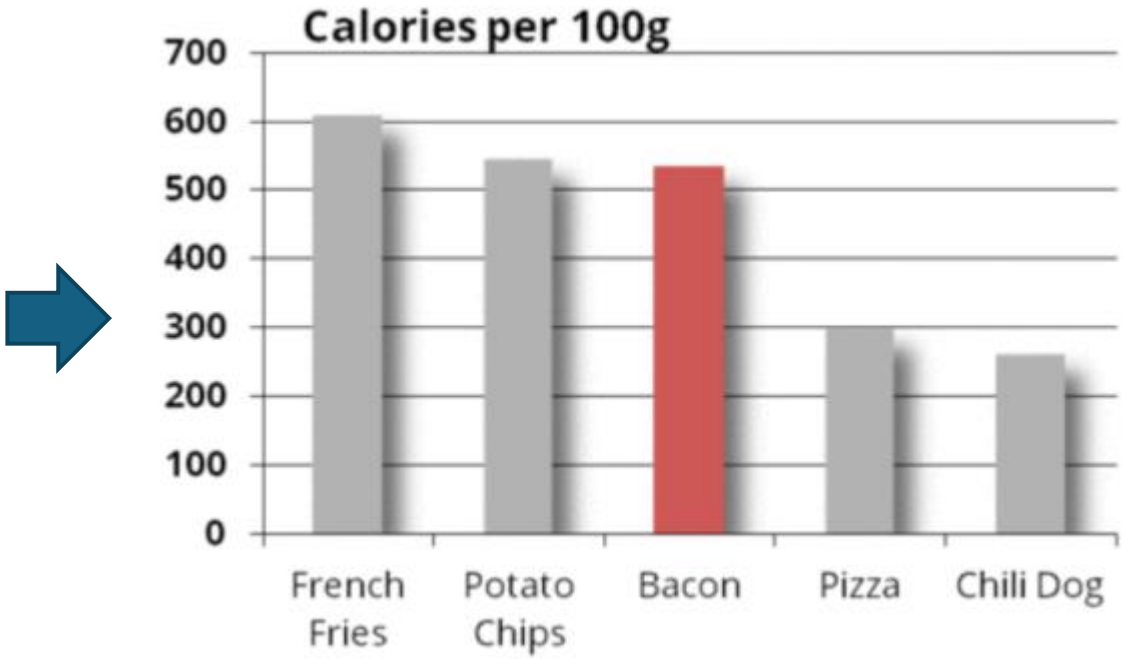


Colors use more data-ink

https://www.darkhorseanalytics.com/blo

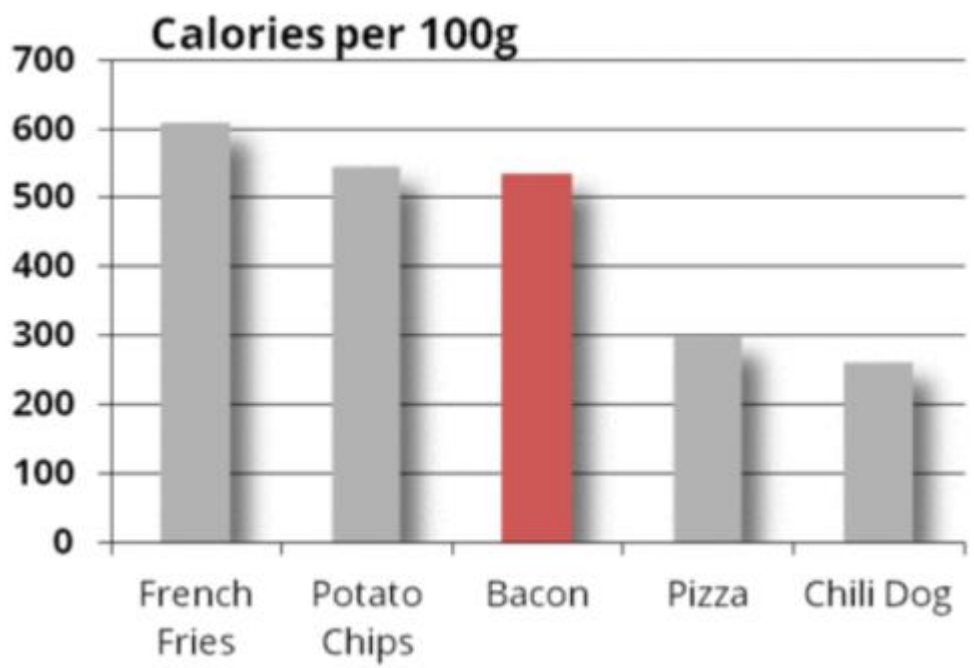
Exercise to maximize the data-ink ratio

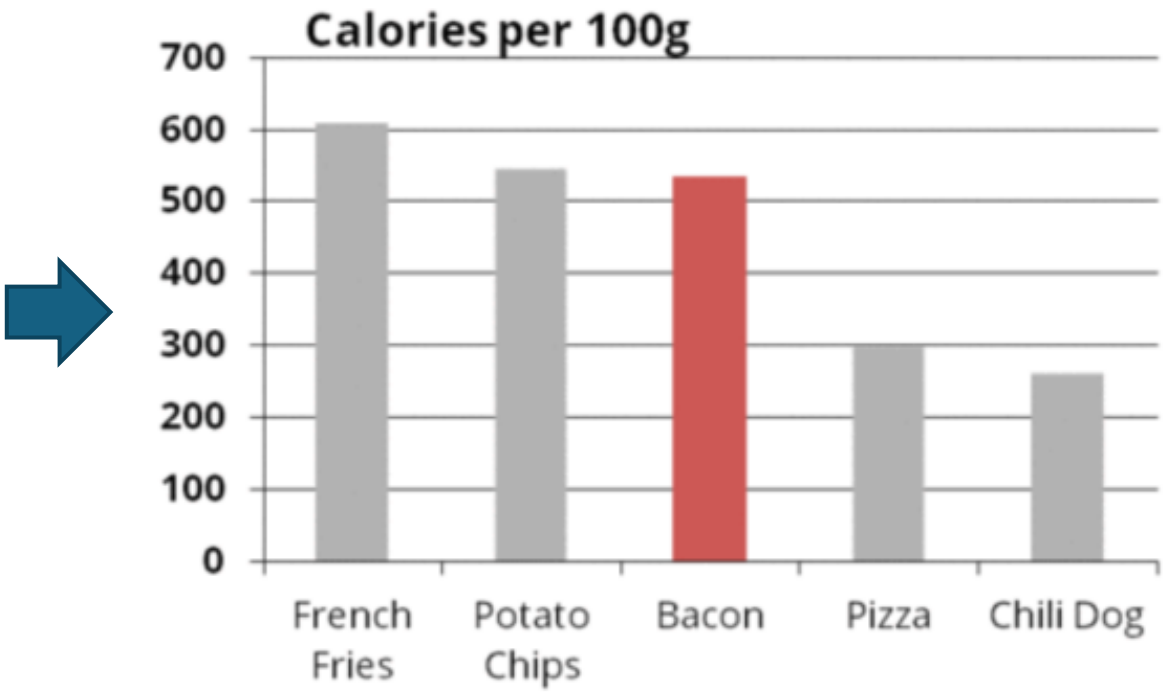




https://www.darkhorseanalytics.com/blo

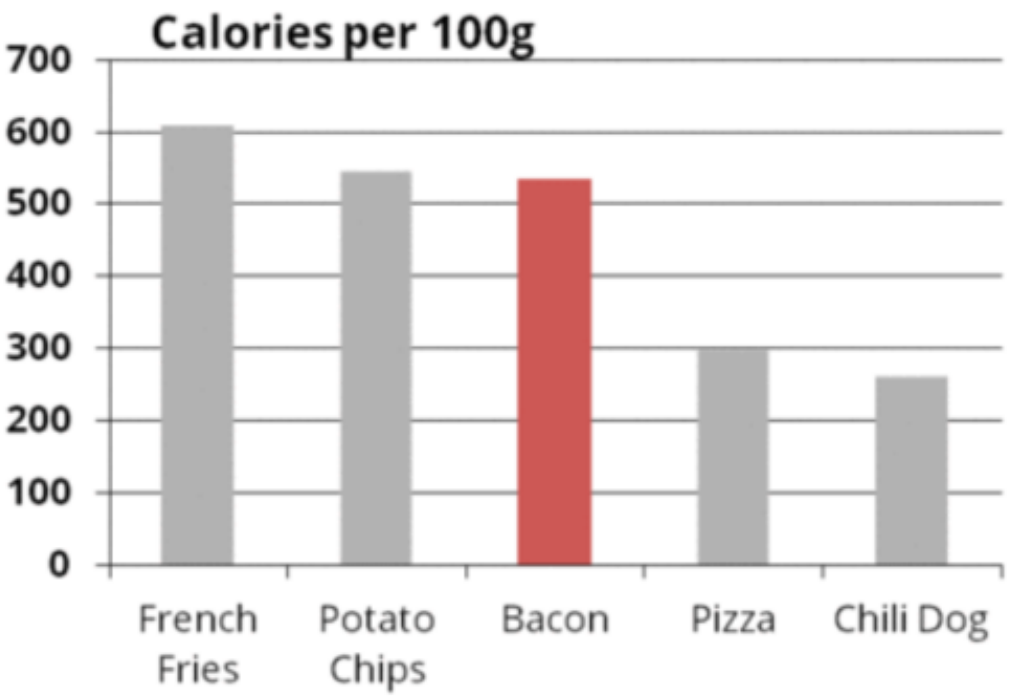
Exercise to maximize the data-ink ratio

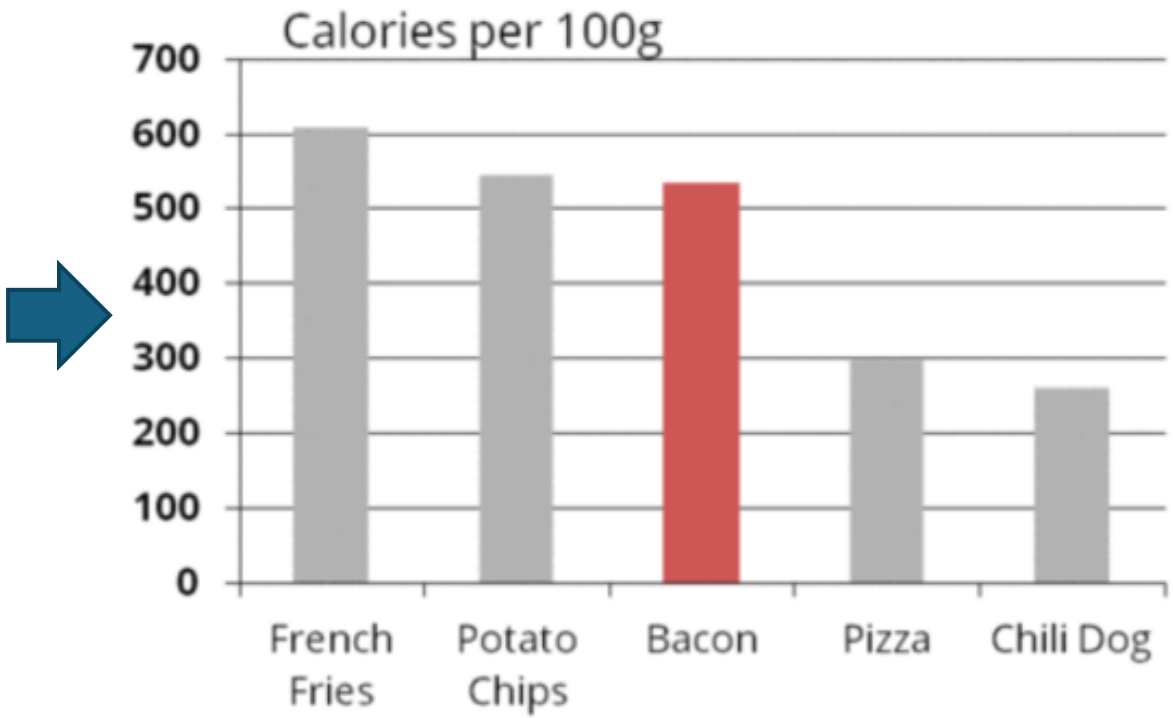




https://www.darkhorseanalytics.com/blo

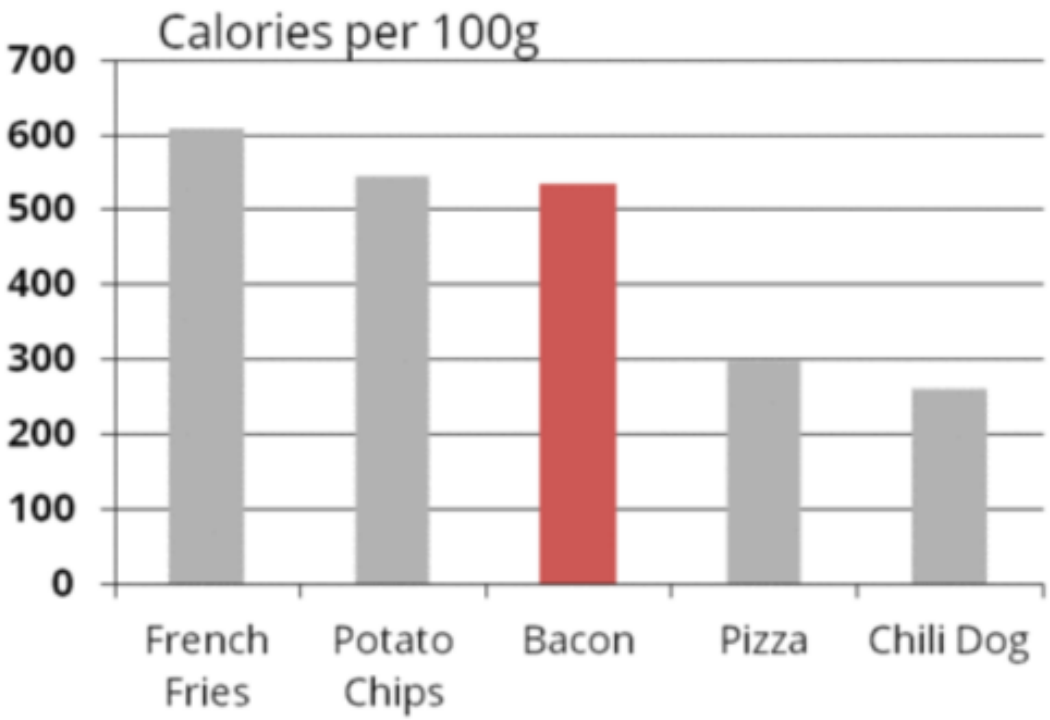
Exercise to maximize the data-ink ratio

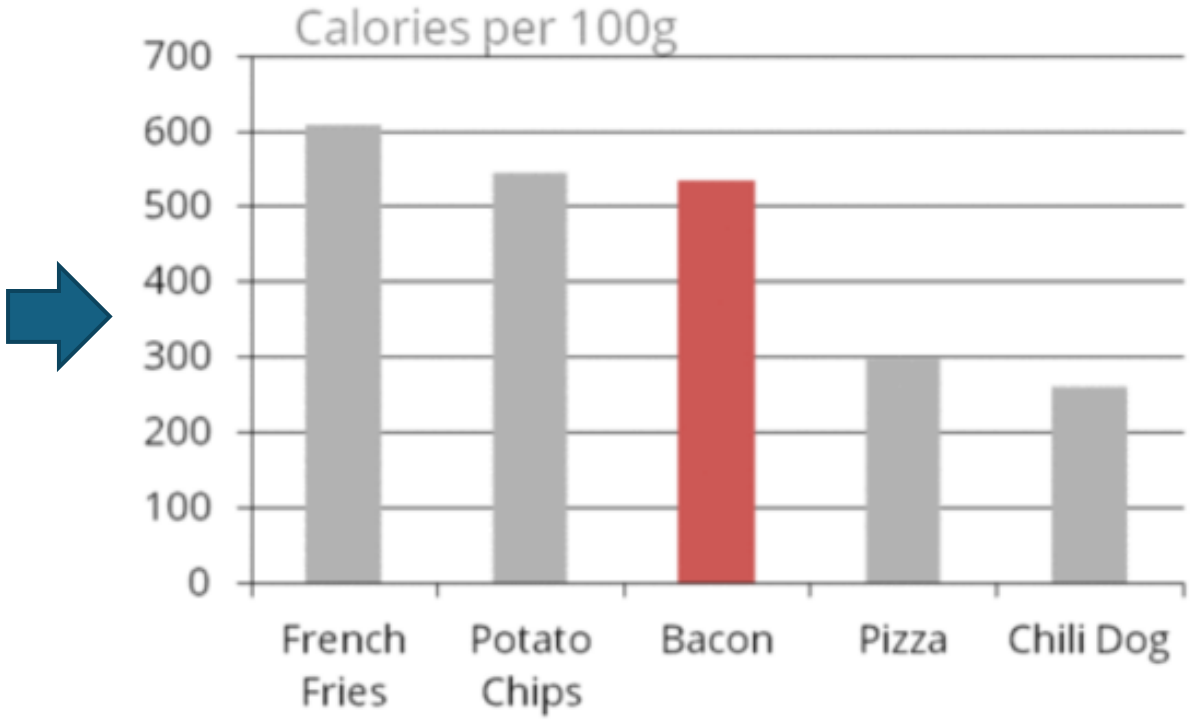




https://www.darkhorseanalytics.com/blo

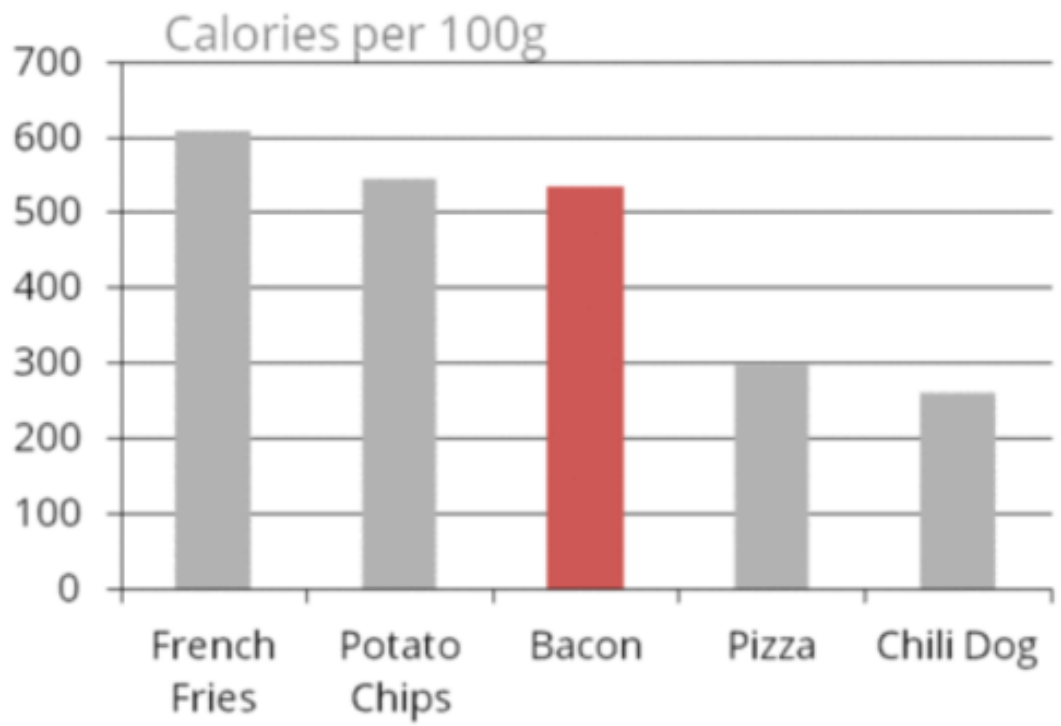
Exercise to maximize the data-ink ratio

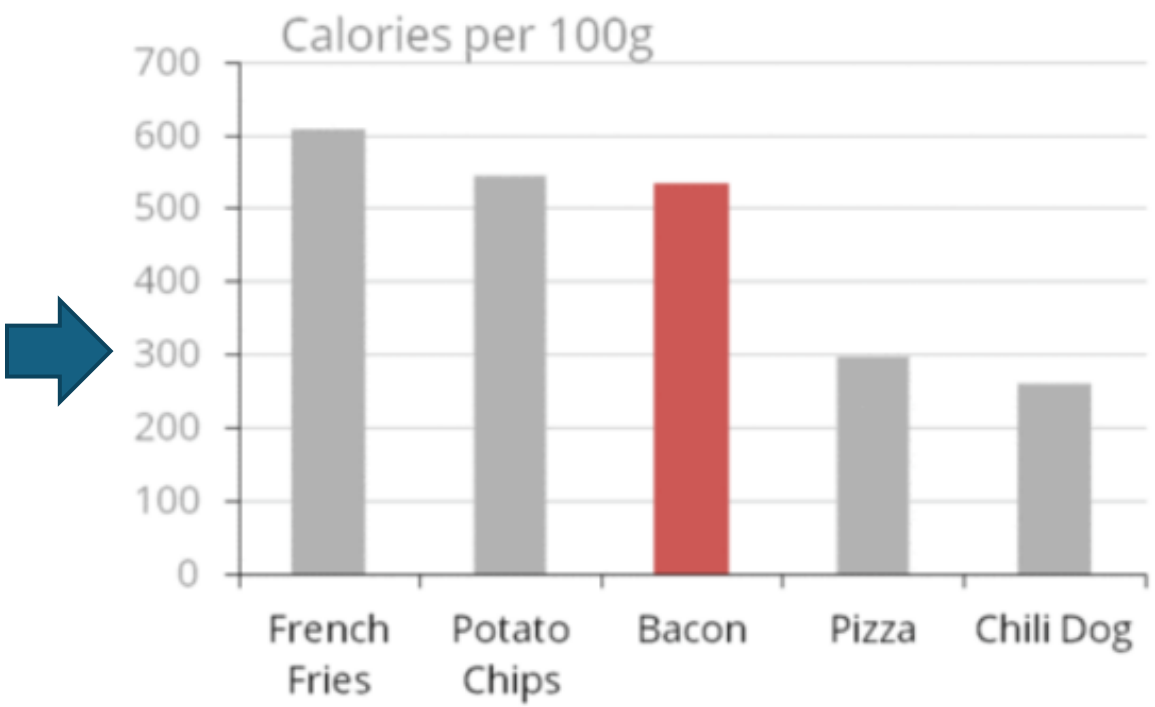




https://www.darkhorseanalytics.com/blo

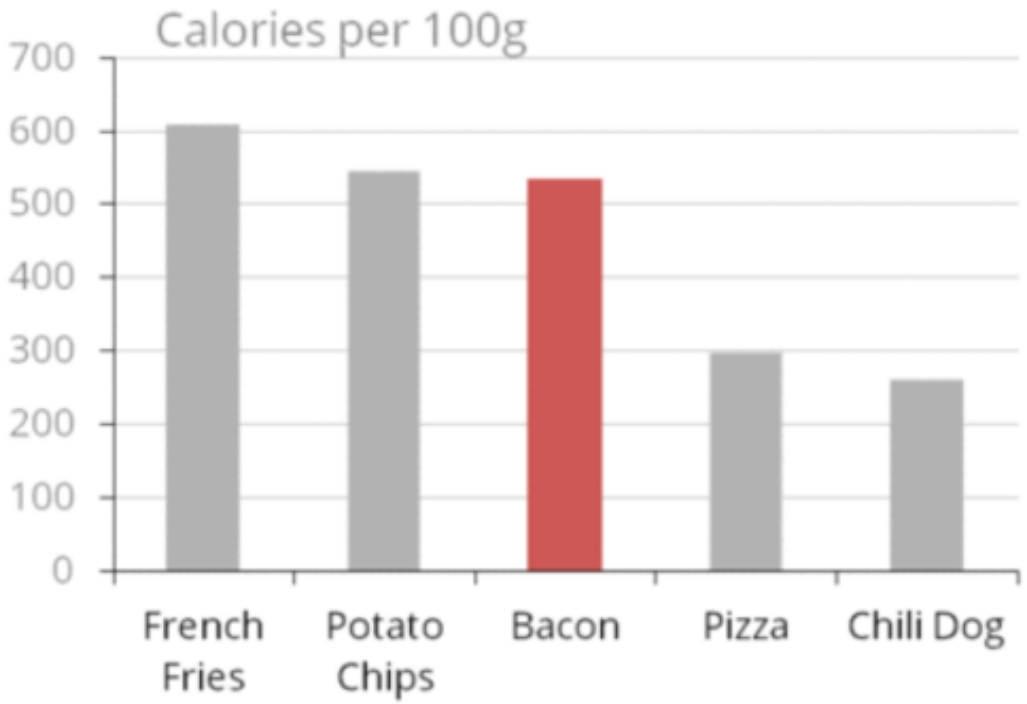
Exercise to maximize the data-ink ratio

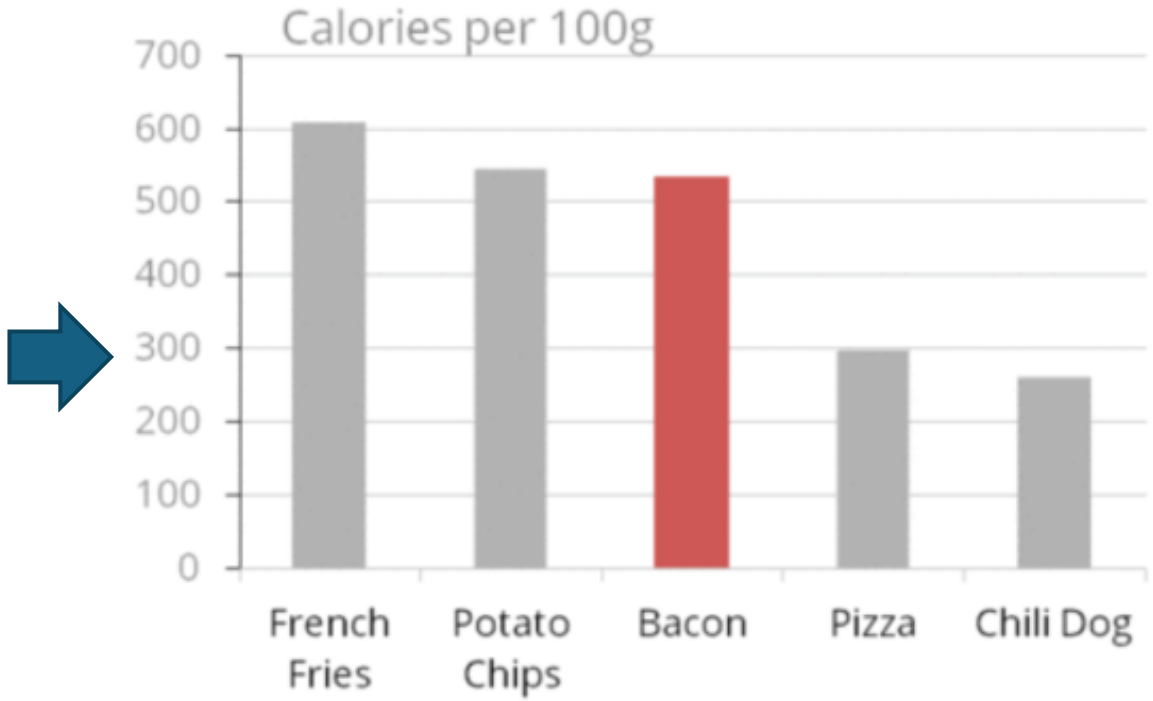




https://www.darkhorseanalytics.com/blo

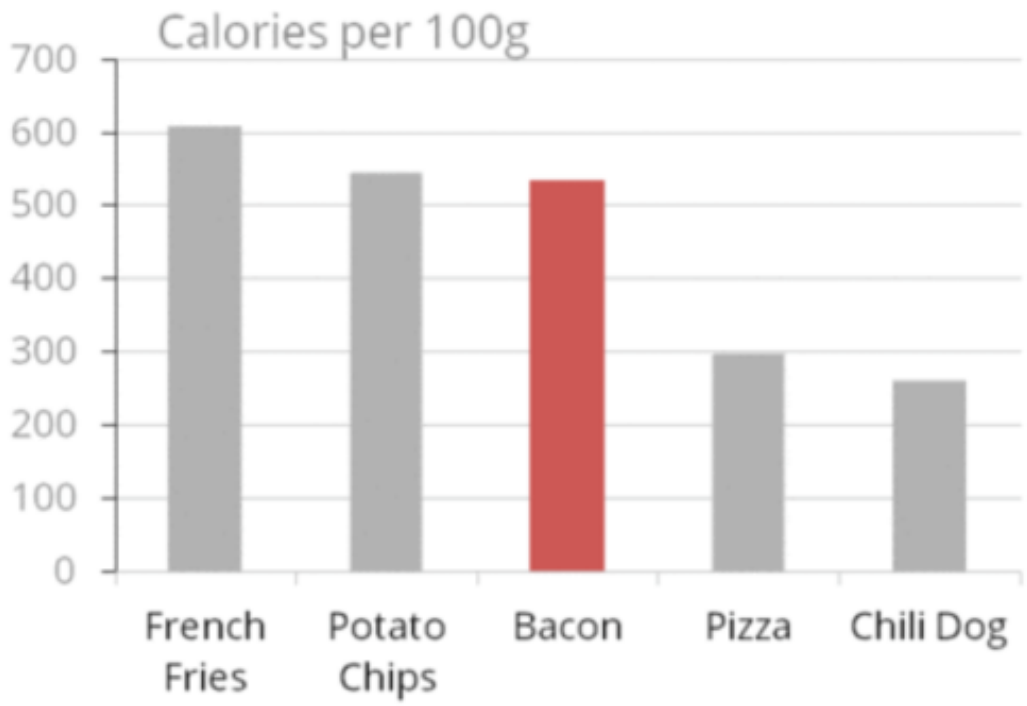
Exercise to maximize the data-ink ratio

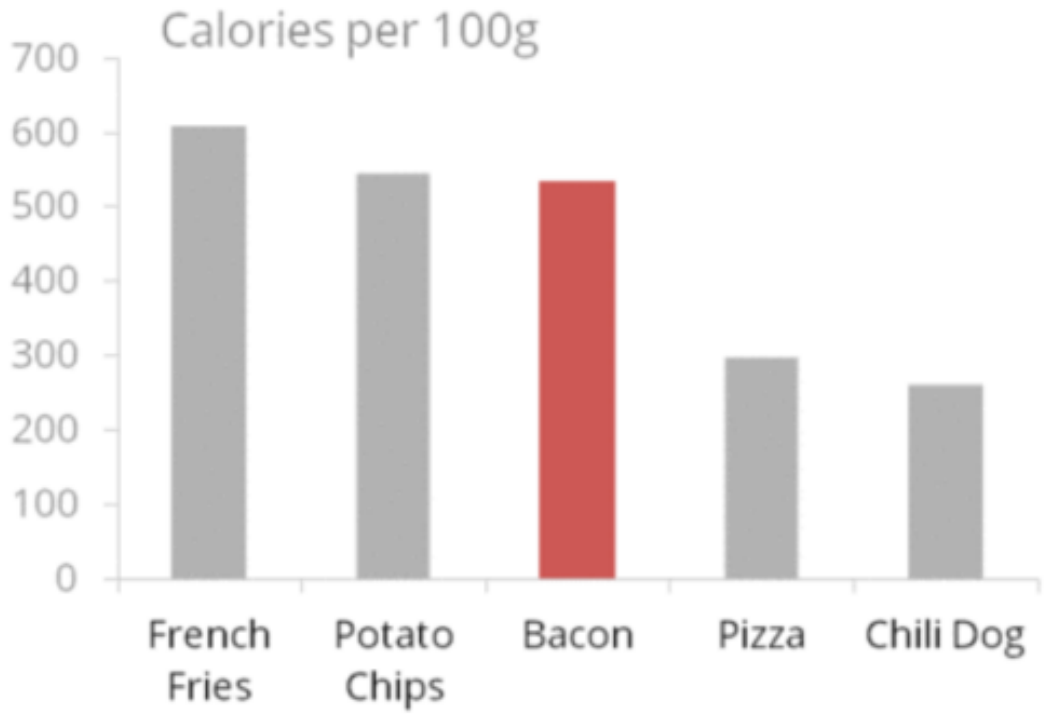




https://www.darkhorseanalytics.com/blo

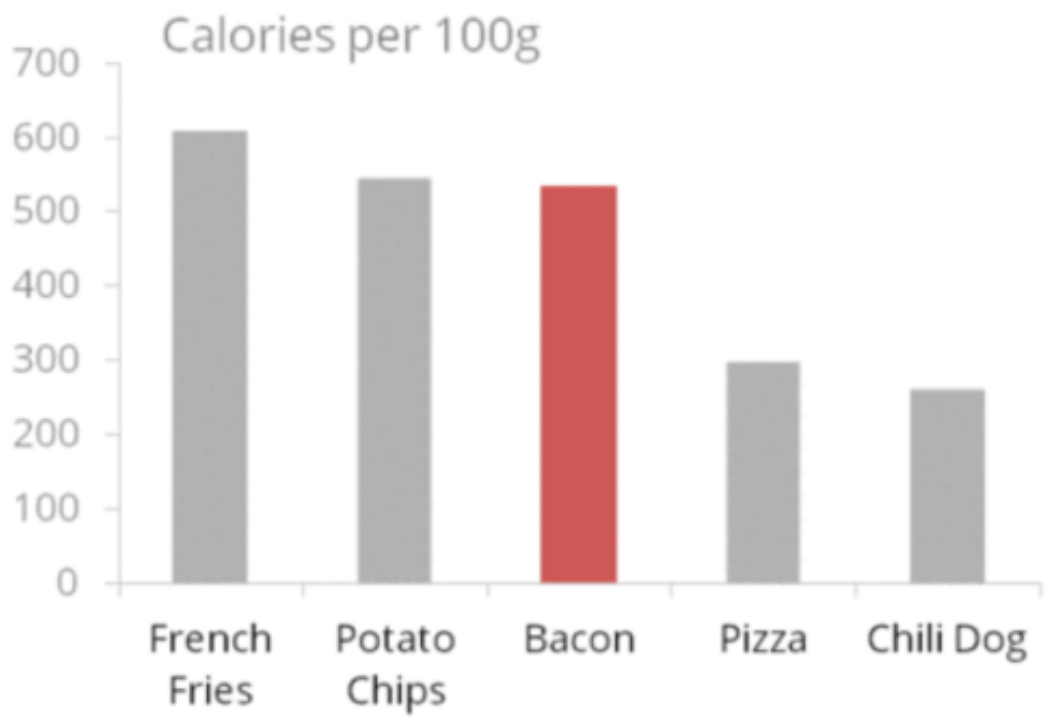
Exercise to maximize the data-ink ratio

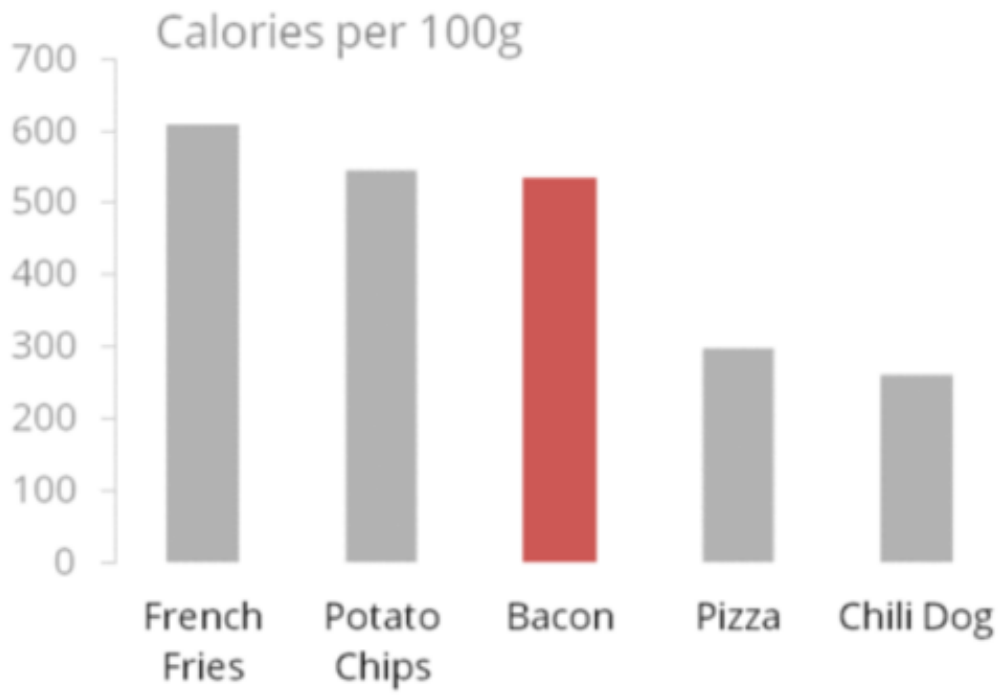




https://www.darkhorseanalytics.com/blo

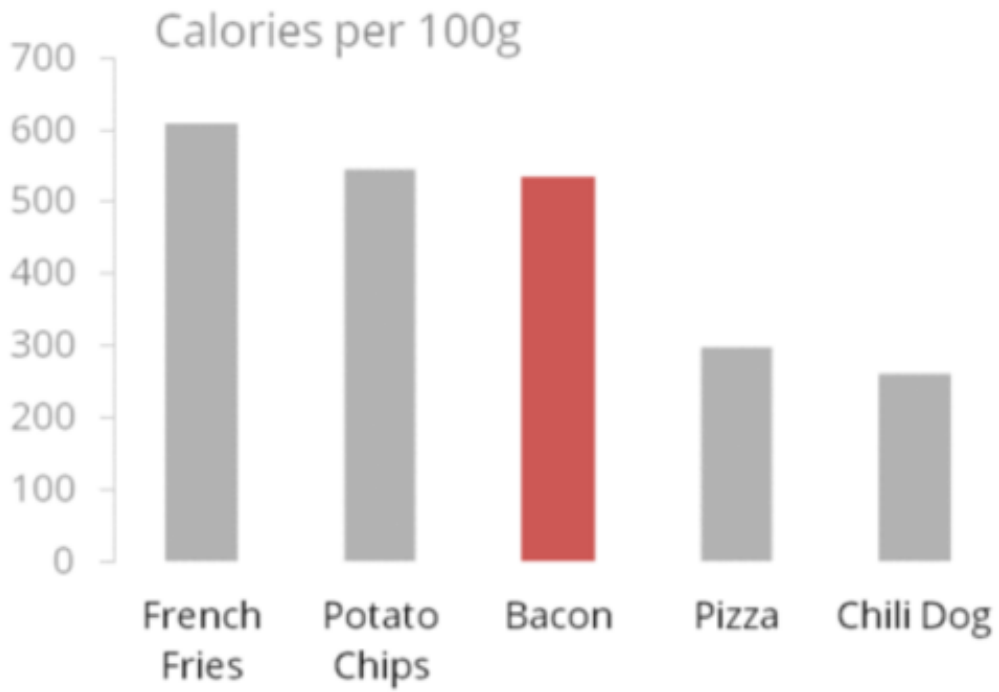
Exercise to maximize the data-ink ratio

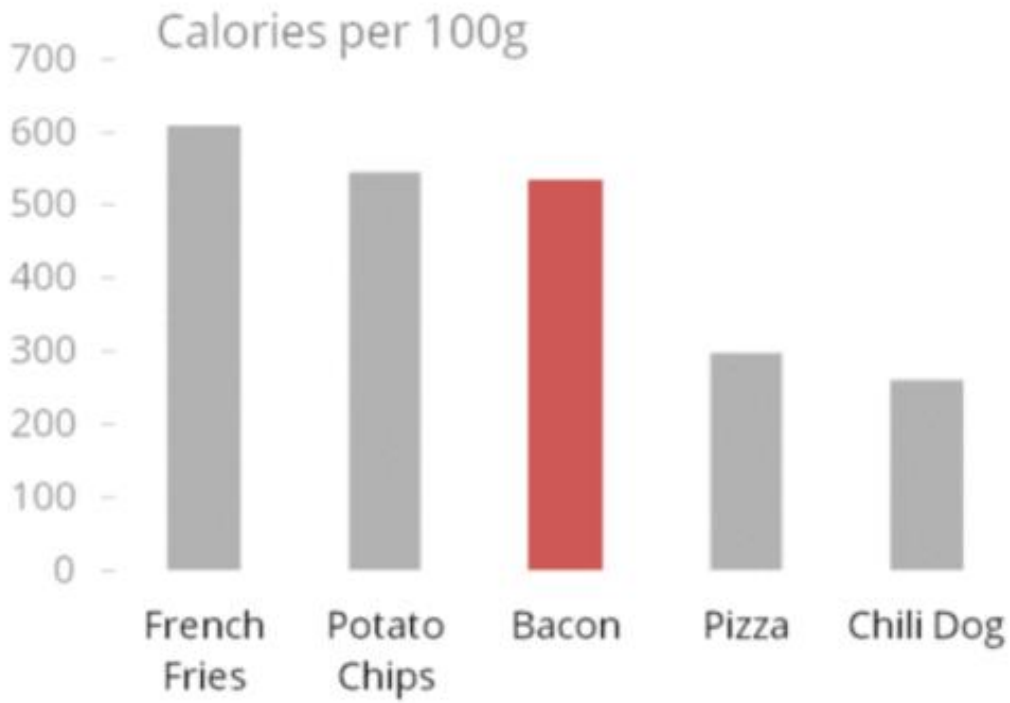




https://www.darkhorseanalytics.com/blo

Exercise to maximize the data-ink ratio



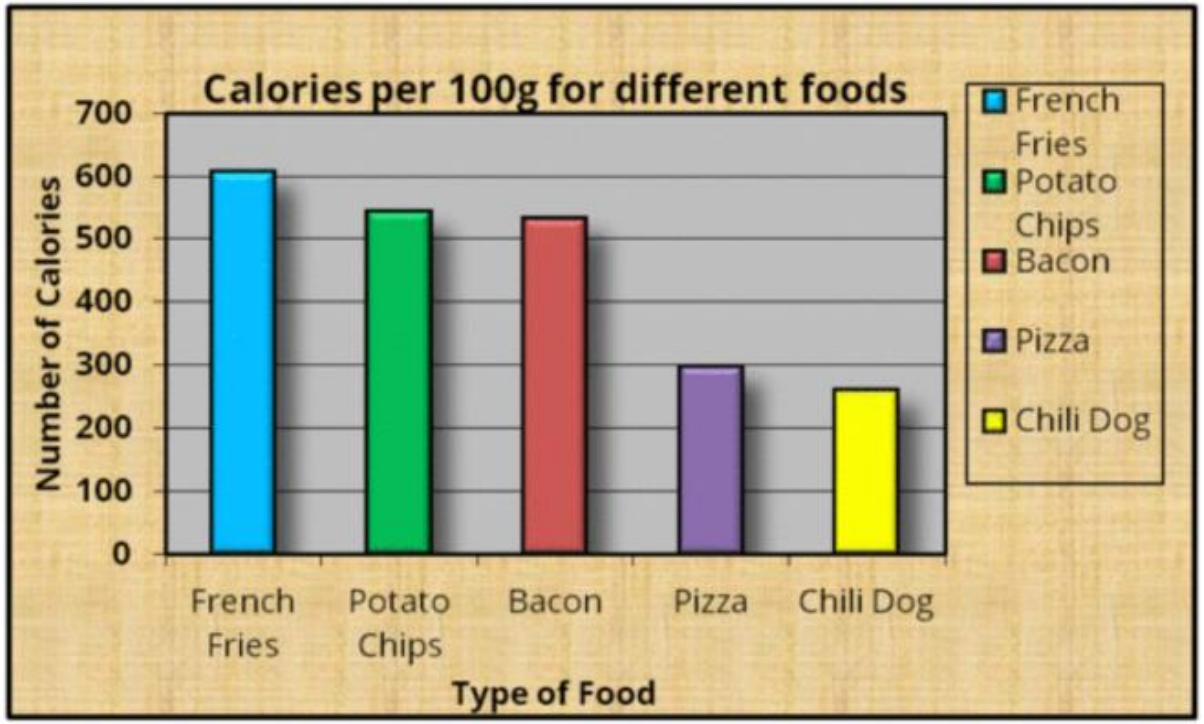


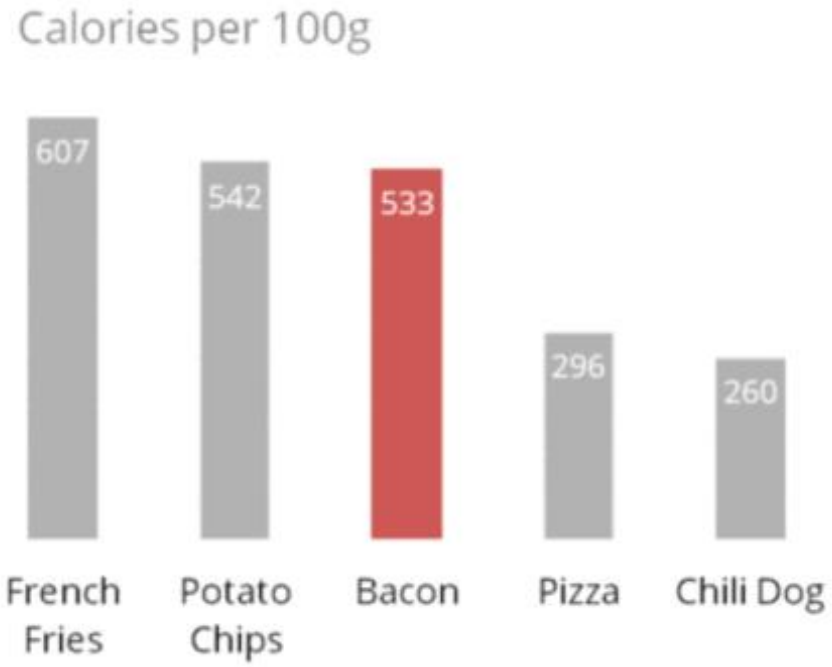
https://www.darkhorseanalytics.com/blo

Exercise to maximize the data-ink ratio

https://www.darkhorseanalytics.com/blo

Good visualization keeps the data-ink ratio high without losing information

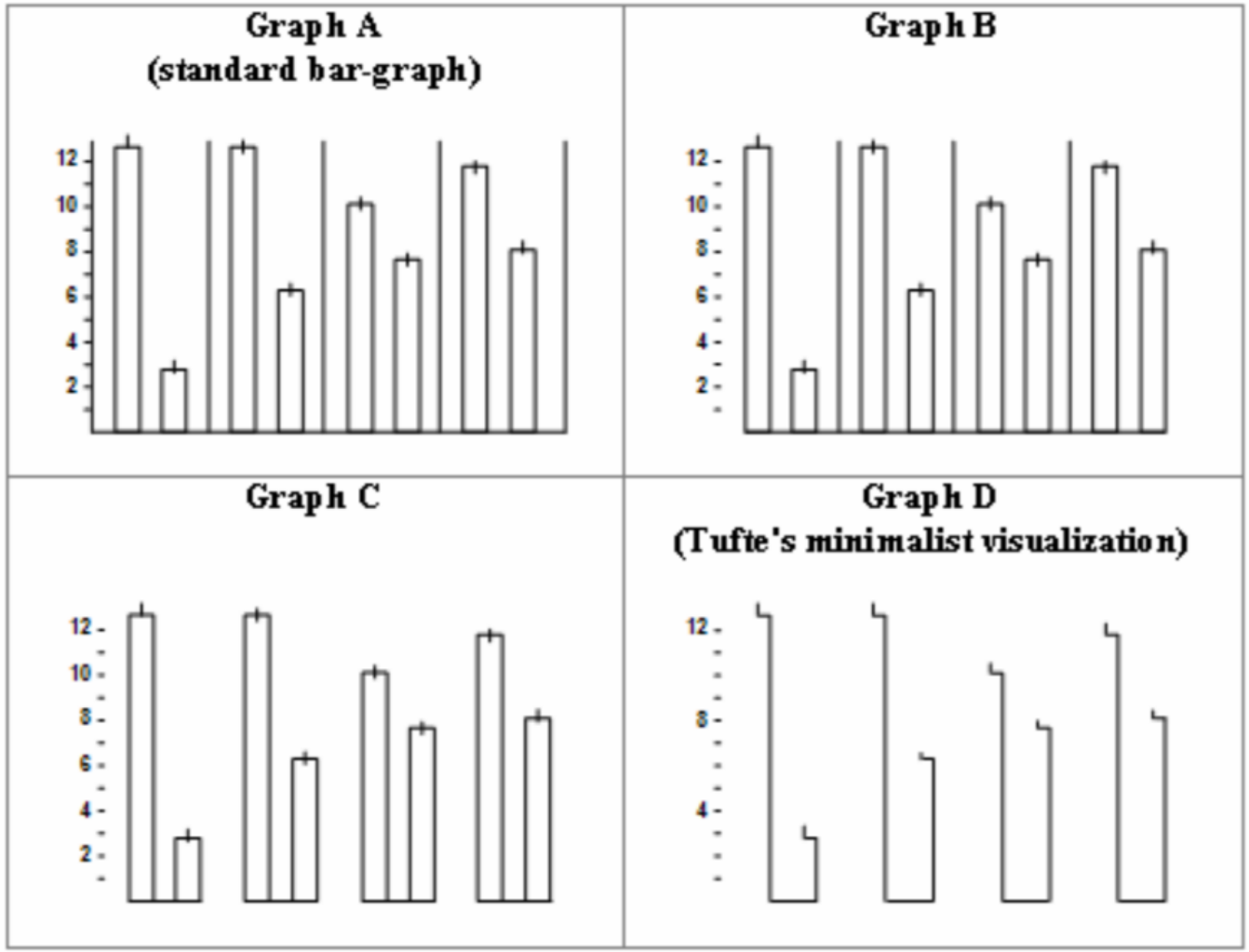




- Did we lose any information?
- “Everything should be made as simple as possible, but no simpler” - Einstein

Let’s not go crazy with maximizing data-ink ratio!

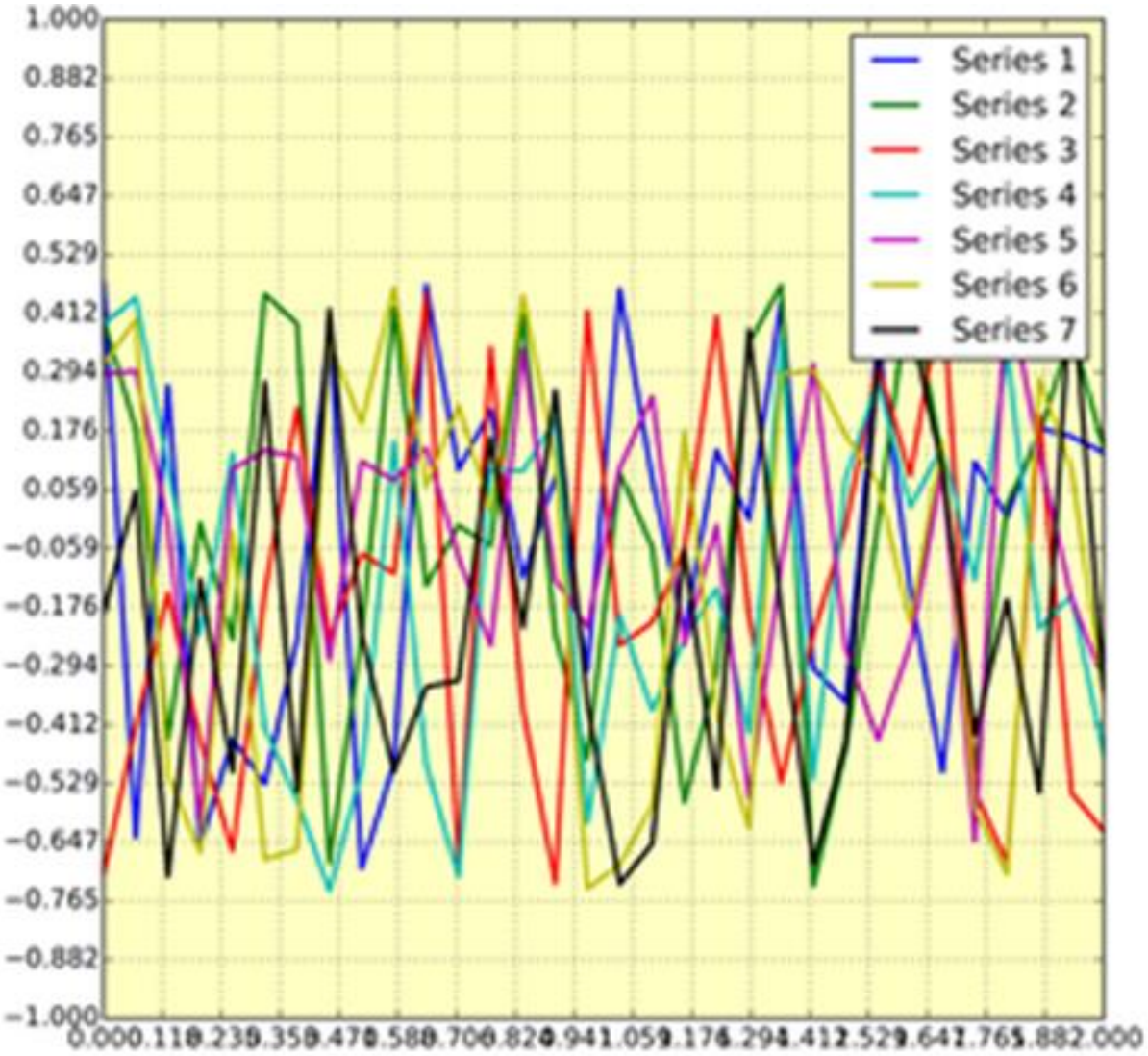
- Maximum data-ink ratio is not an absolute rule
- E.g., Bar charts should have bars standing from 0.
- The more important thing is balance



Inbar et al. Proceedings of the ECCE 2007 Conference, 28-31 August 2007, London,

Good visualization avoids clutter

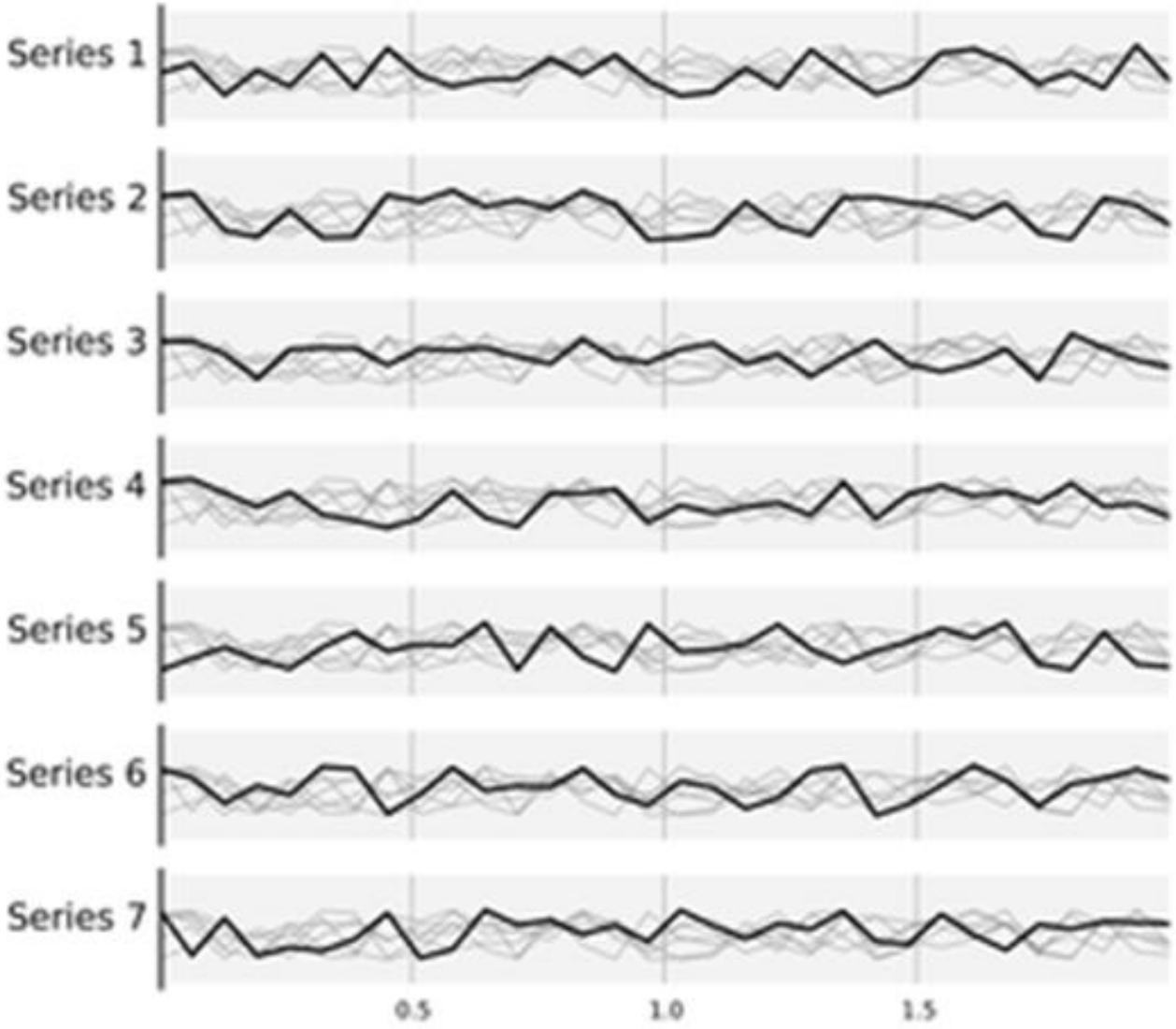
- All the unnecessary visual elements that
  - Do not improve the message
  - Add confusion
- E.g.
  - Too many lines of different colors
  - The legend overlaps the graphics content
  - Too many labels in the legend
  - Gratuitously colored background
  - Useless grid lines not helpful to understand
  - Gratuitously given x-axis and y-axis ticks



Rougier et al. Ten Simple Rules for Better Figures, PLoS Comp. Bio. 20

Good visualization avoids clutter

- How does it improve the criticisms?
  - Too many lines of different colors
  - The legend overlaps the graphics content
  - Too many labels in the legend
  - Gratuitously colored background
  - Useless grid lines
  - Gratuitously given x-axis and y-axis ticks
- What graphics design components does this lack?



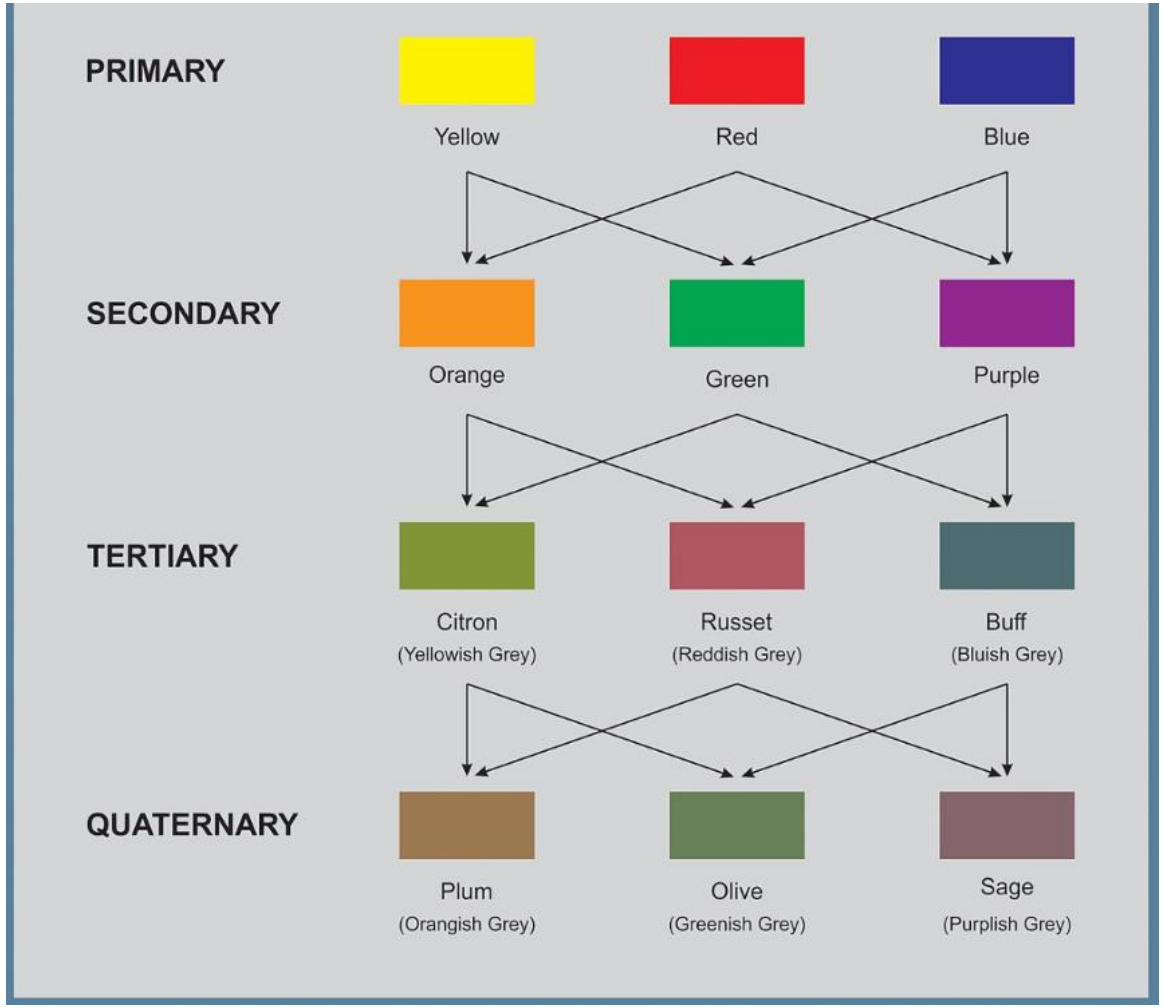
Rougier et al. Ten Simple Rules for Better Figures, PLoS Comp. Bio. 20

How to suit colors to purpose

If multiple data objects in the foreground,

1. Use complementary colors for contrast
2. Use analogous colors to accompany
3. Use triadic colors to compare 3 groups
4. Use split-complementary colors to highlight two groups

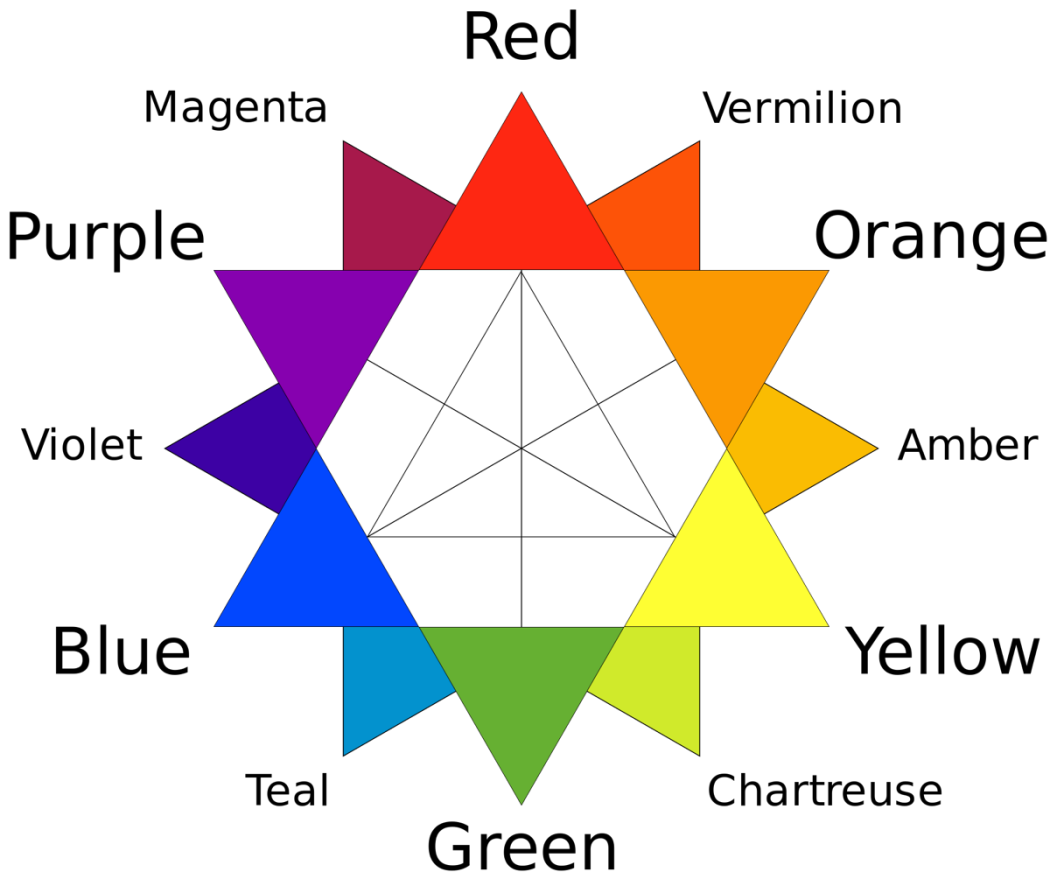
Degree of colors



https://en.wikipedia.org/wiki/Color_theory/

- Primary and secondary colors for most direct effect
- Tertiary and quaternary for indirect effect

Degree of colors



- Primary and secondary colors for most direct effect
- Tertiary and quaternary for indirect effect

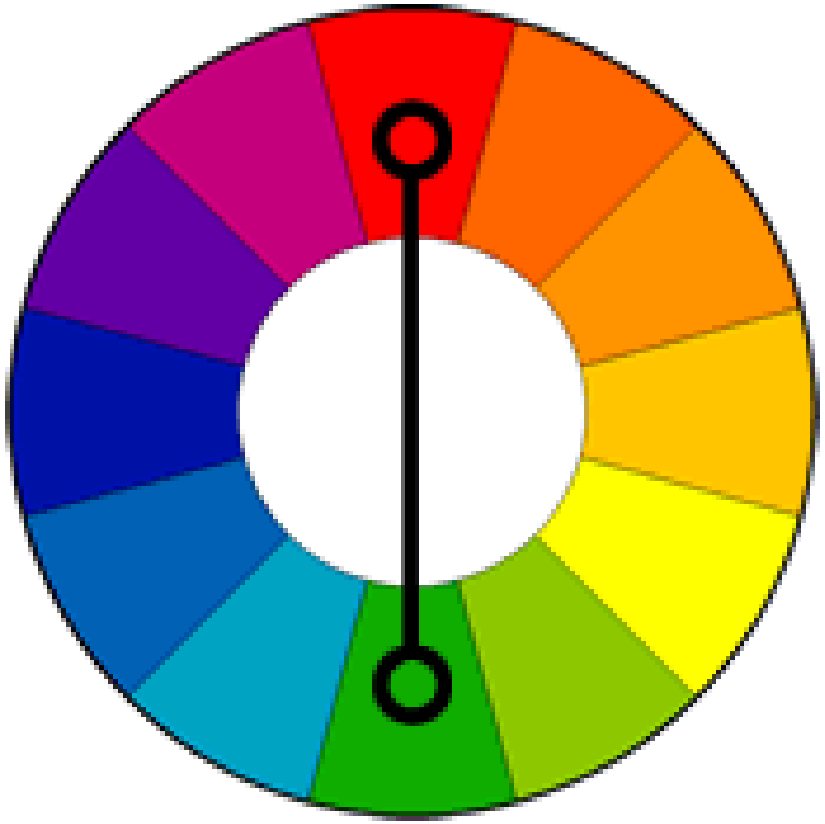
←A reference for color picking

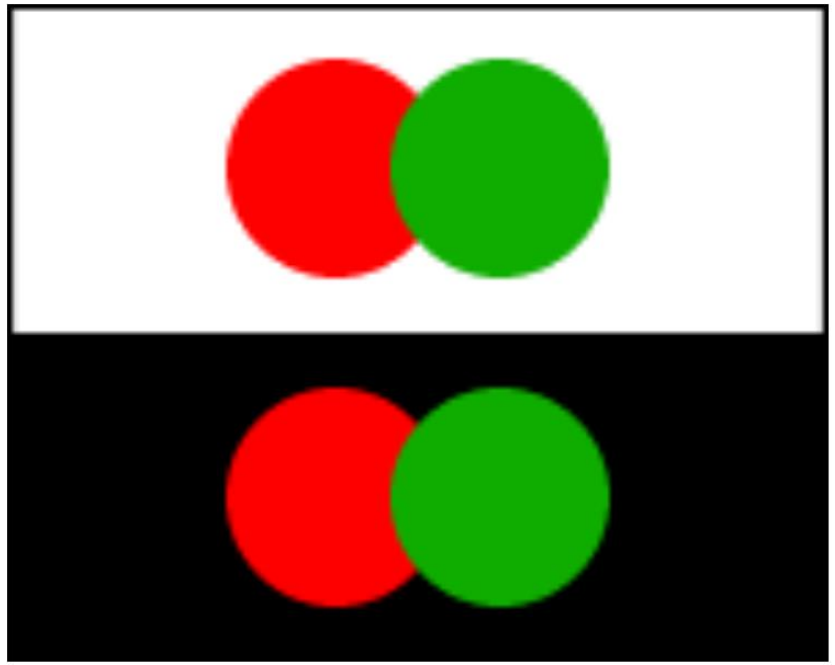
https://en.wikipedia.org/wiki/Color_theory/

Use complementary colors to contrast

Colors oppose to each other

- Create a vibrant look
- Tricky to use in large doses
- In scientific figures with two groups to contrast (e.g. good and bad genes)



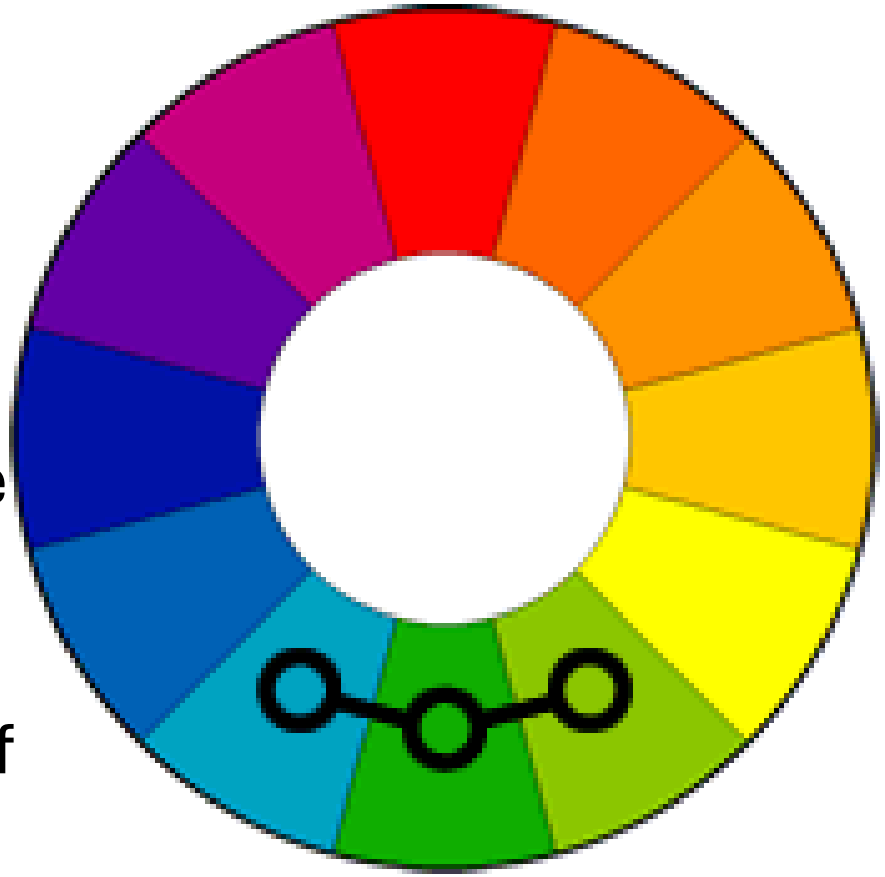


https://www.tigercolor.com/color-lab/color-theory/color-

Use analogous colors to emphasize similarity

Colors next to each other

- Create a serene look
- Often found in nature
- One to dominate, a second to support, the third as an accent
- In scientific figures with multiple groups of similar attributes

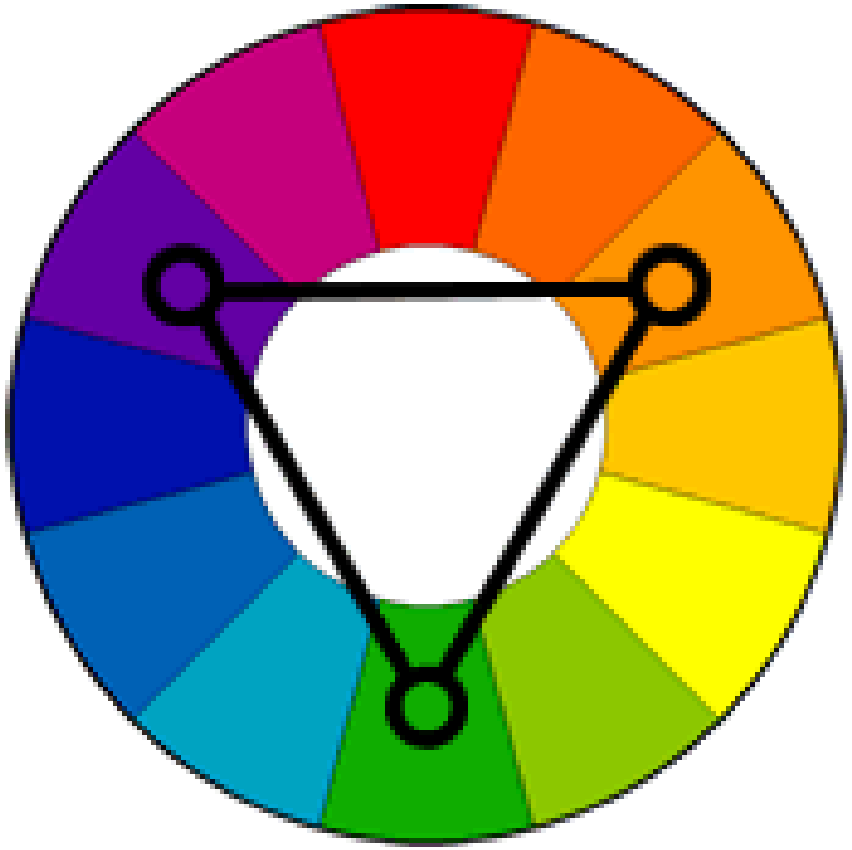


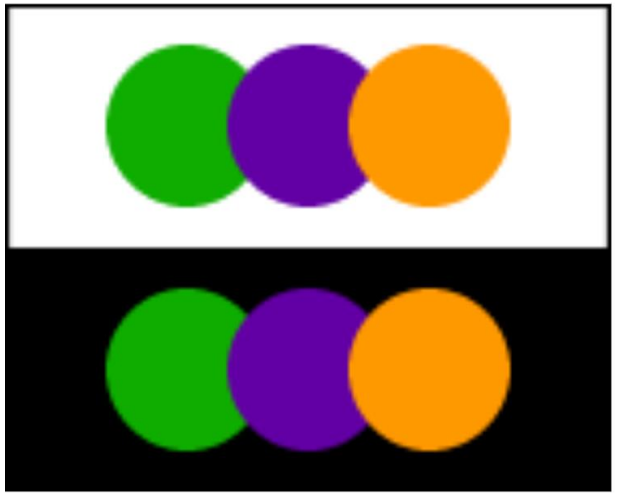
https://www.tigercolor.com/color-lab/color-theory/color-

Use triadic colors to compare 3 groups

Colors evenly spaced

- Create a vibrant look
- One to dominate, the other two for accent
- In scientific figures with multiple groups to compare of different attributes



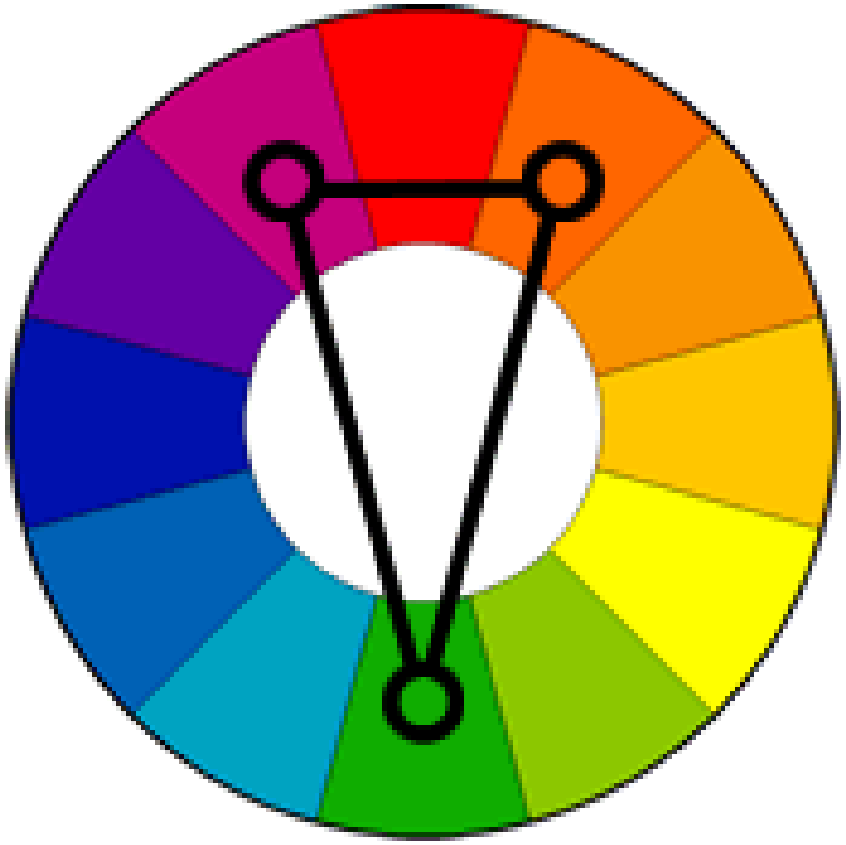


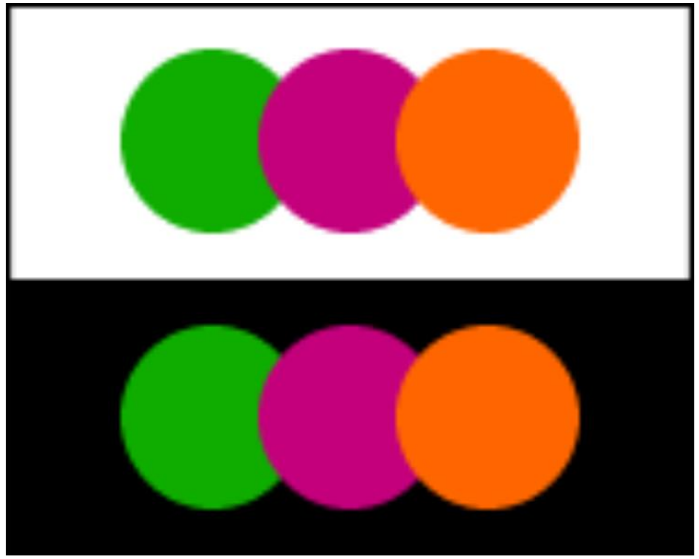
https://www.tigercolor.com/color-lab/color-theory/color-

Use split-complementary colors to highlight two groups

A variation of the complementary color scheme

- The base color + two complementary colors
- Strong visual contrast with less tension
- In scientific figures with multiple groups when comparing one to the other two





https://www.tigercolor.com/color-lab/color-theory/color-

How to match background and foreground colors

Background color matters for clarity and style

e.g.

Can you read this easily?

How about this one, then?

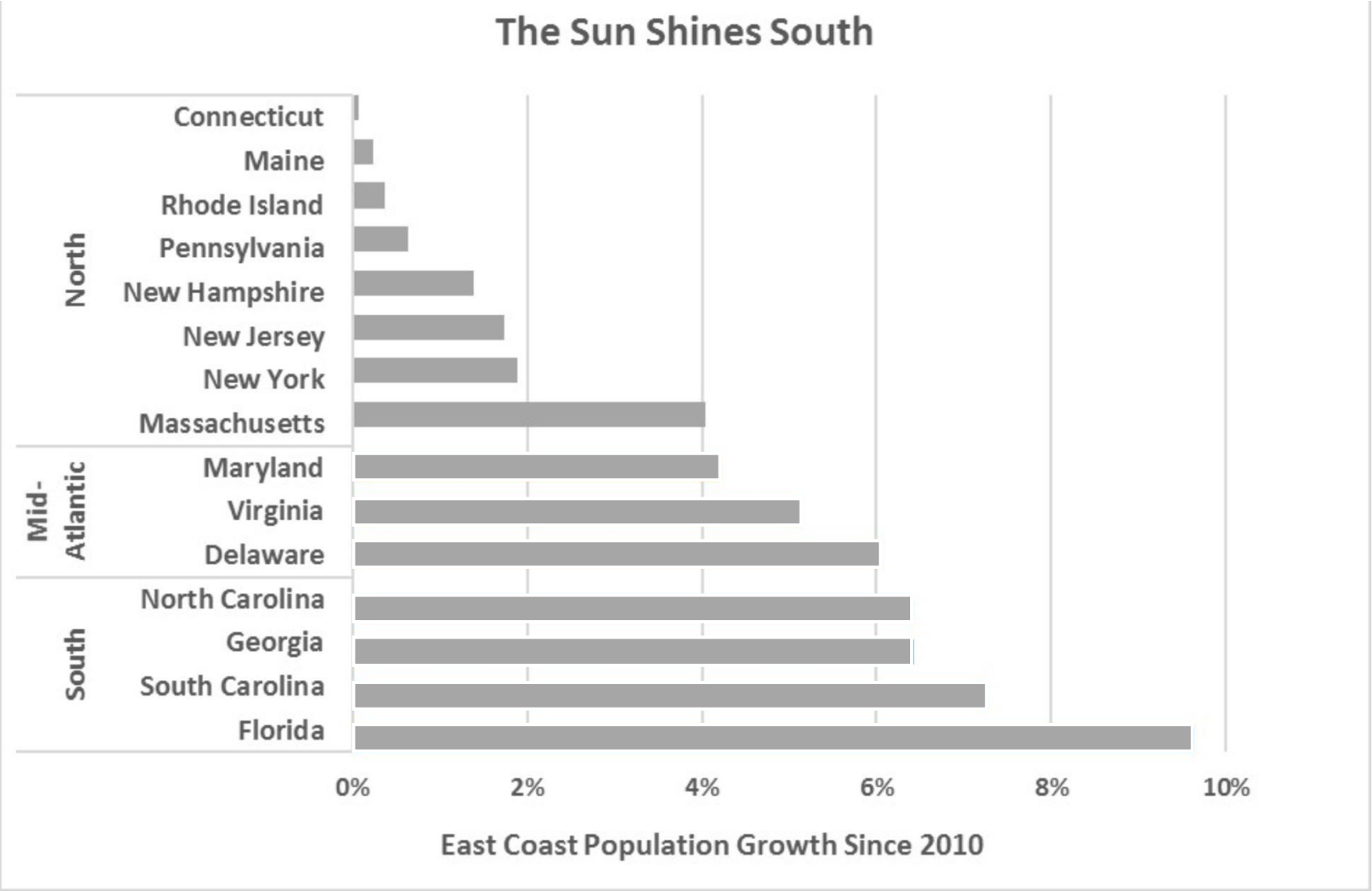
You can do better than this.

How to use colors for scientific figures

1. Colors to categorize
2. Colors to indicate ordinal values
3. Colors to indicate values and directionalities
4. Colors to highlight
5. Colorblind-friendly visualizations

Colors to categorize

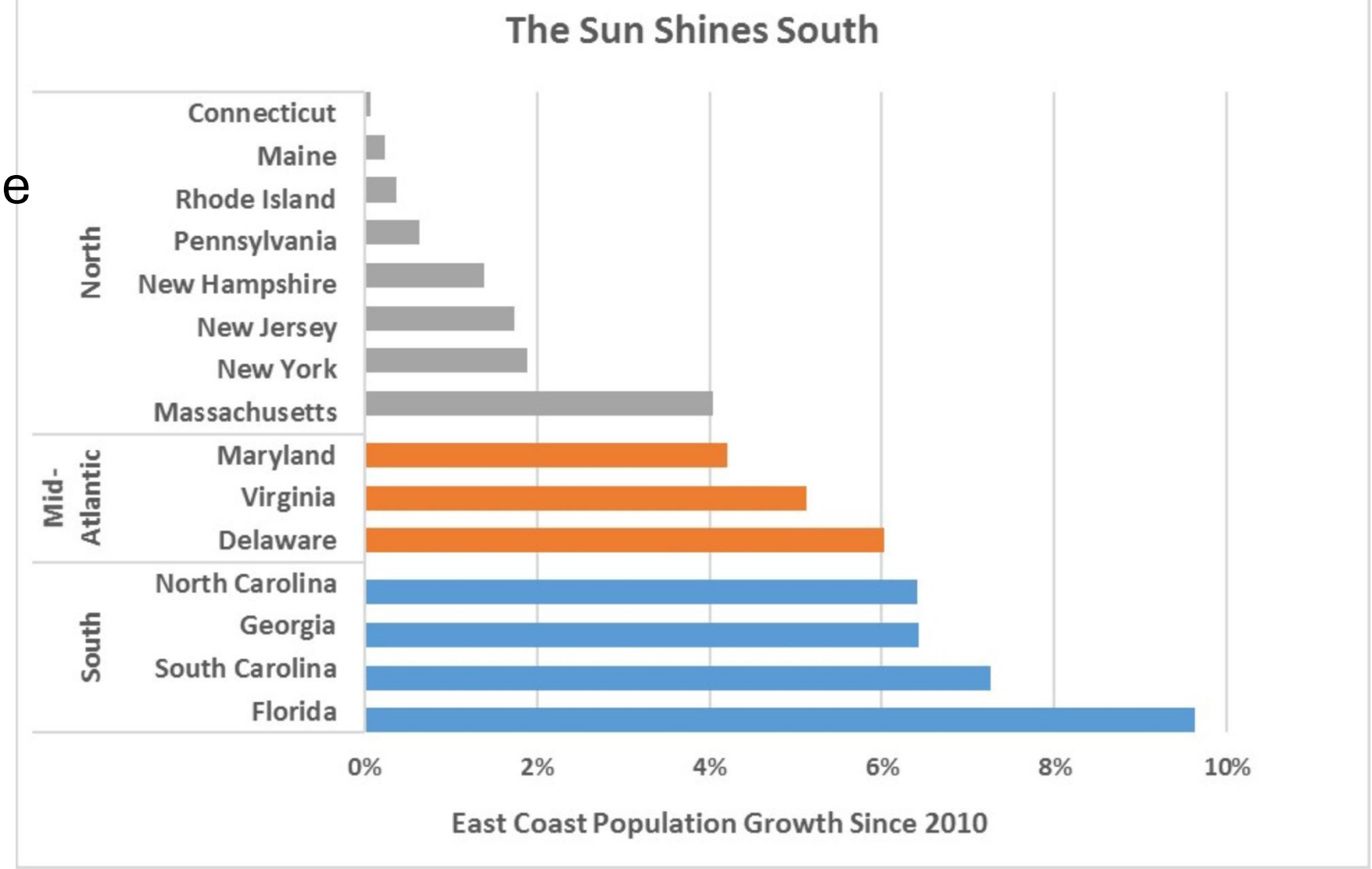
- Growth trend shared by region
- However, the presenter's focus is not clear



Adapted from Decennial Census and 2016 State Population Estimates

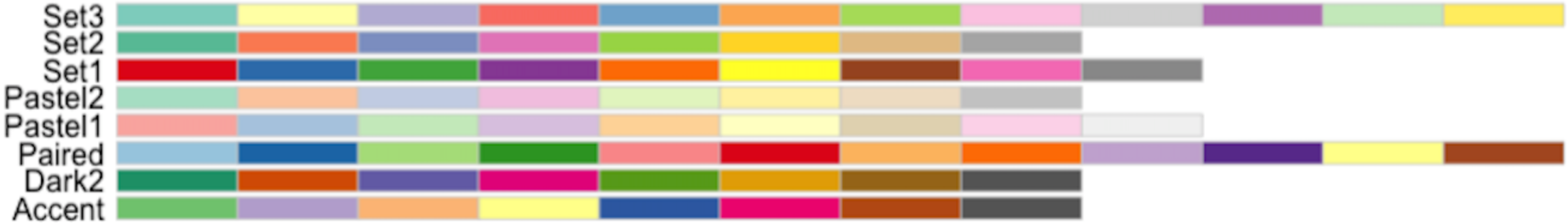
Colors to categorize

- Coloring helps the categorization with the focus on Mid-Atlantic and South



Decennial Census and 2016 State Population Estimates

Palette to categorize by color



https://python-graph-gallery.com/python-colors/

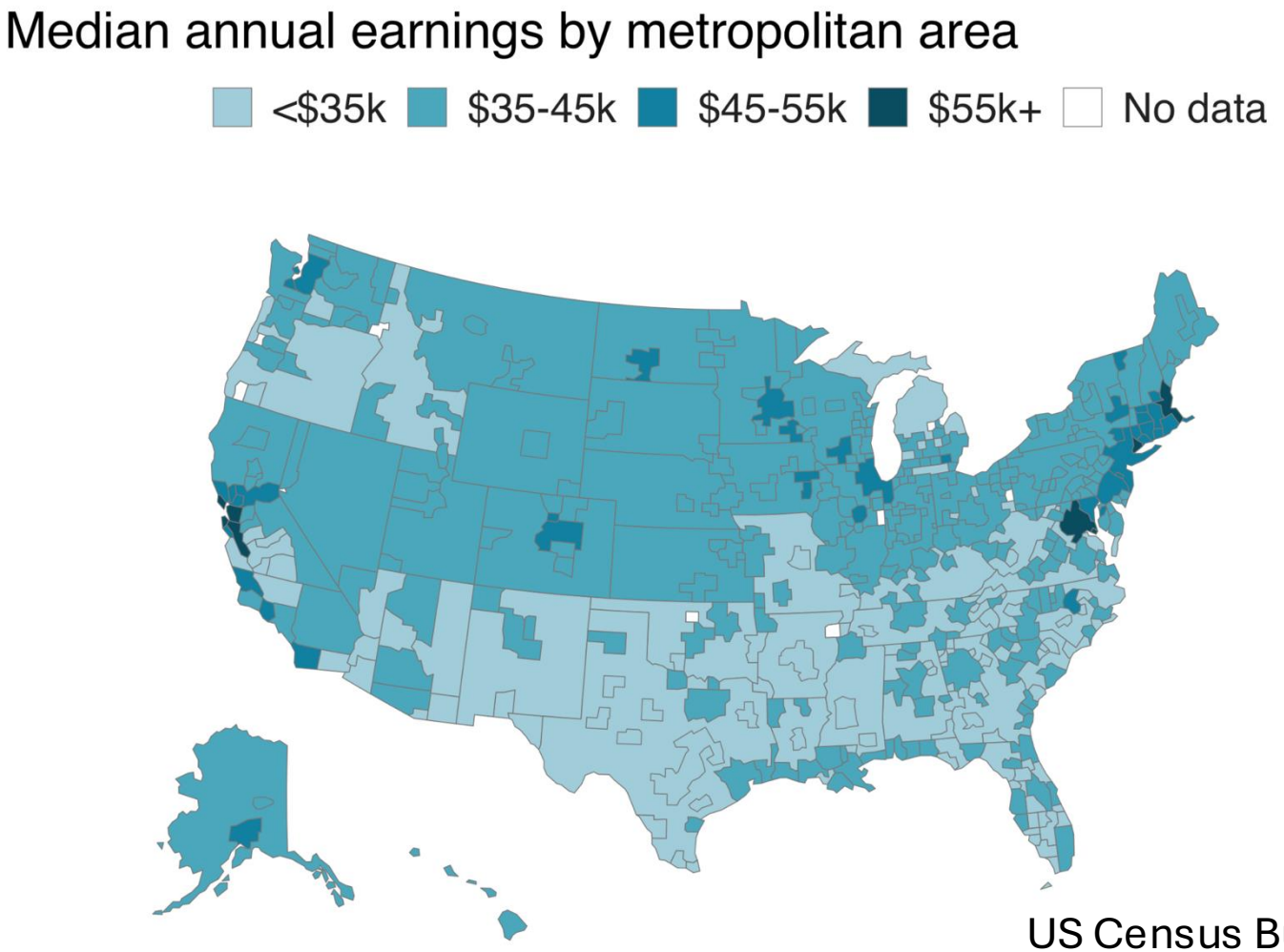
Categorical palette in matplotlib for categorical values

- That have no ordinal relationship
- Using more than 3 may lead to confusion

Colors to indicate ordinal values

Colors need to indicate

- Which values are higher or lower
- How distant two values from each other.
- The color scale should vary uniformly

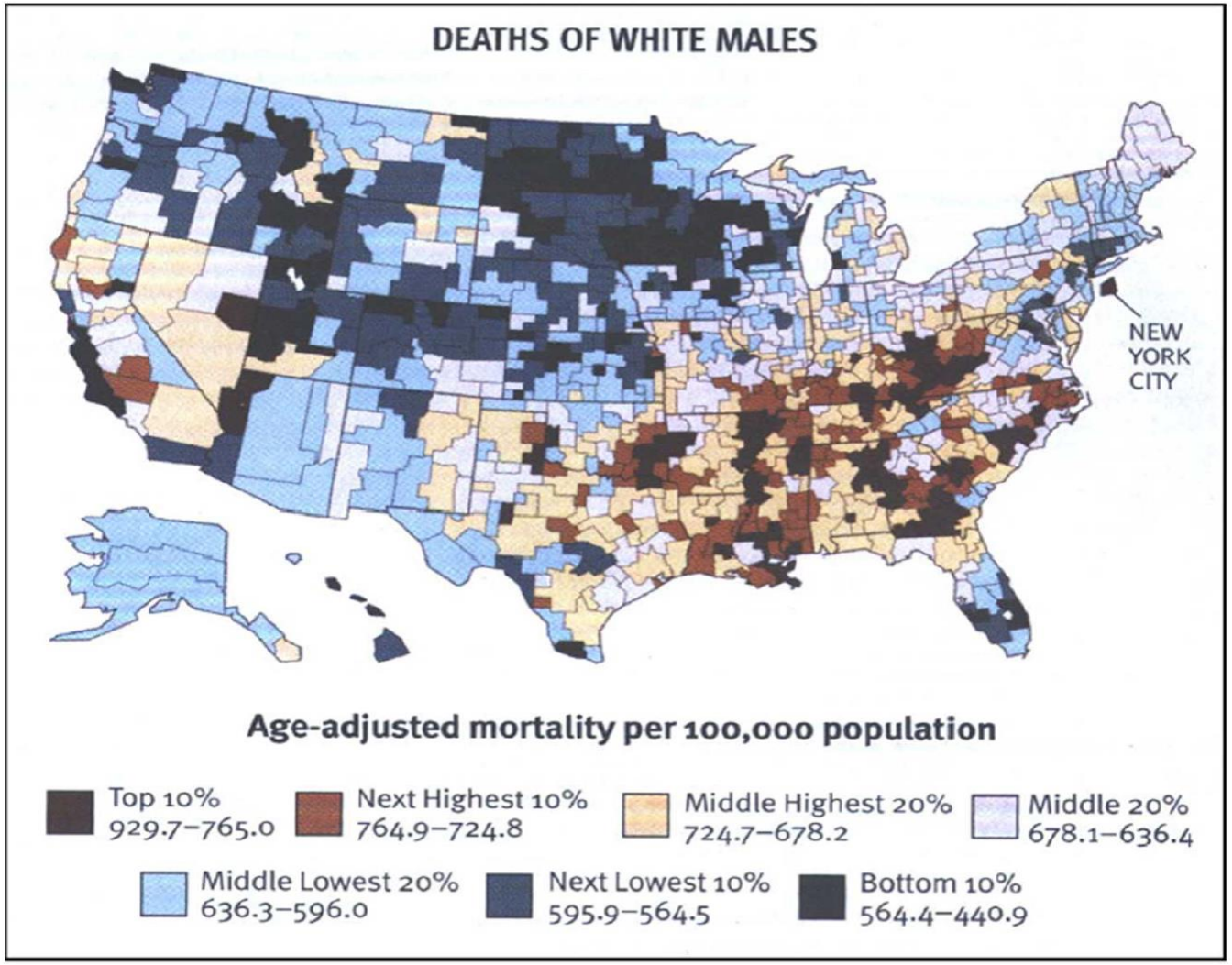


US Census Bureau

Colors to indicate ordinal values

Colors need to indicate

- Which values are higher or lower
- How distant two values from each other.
- The color scale *should* vary uniformly



http://www2.cs.uh.edu/~gnawali/courses/cosc6397-f13/intro-visualization.p

Colors to indicate ordinal values

Good, in that we can see

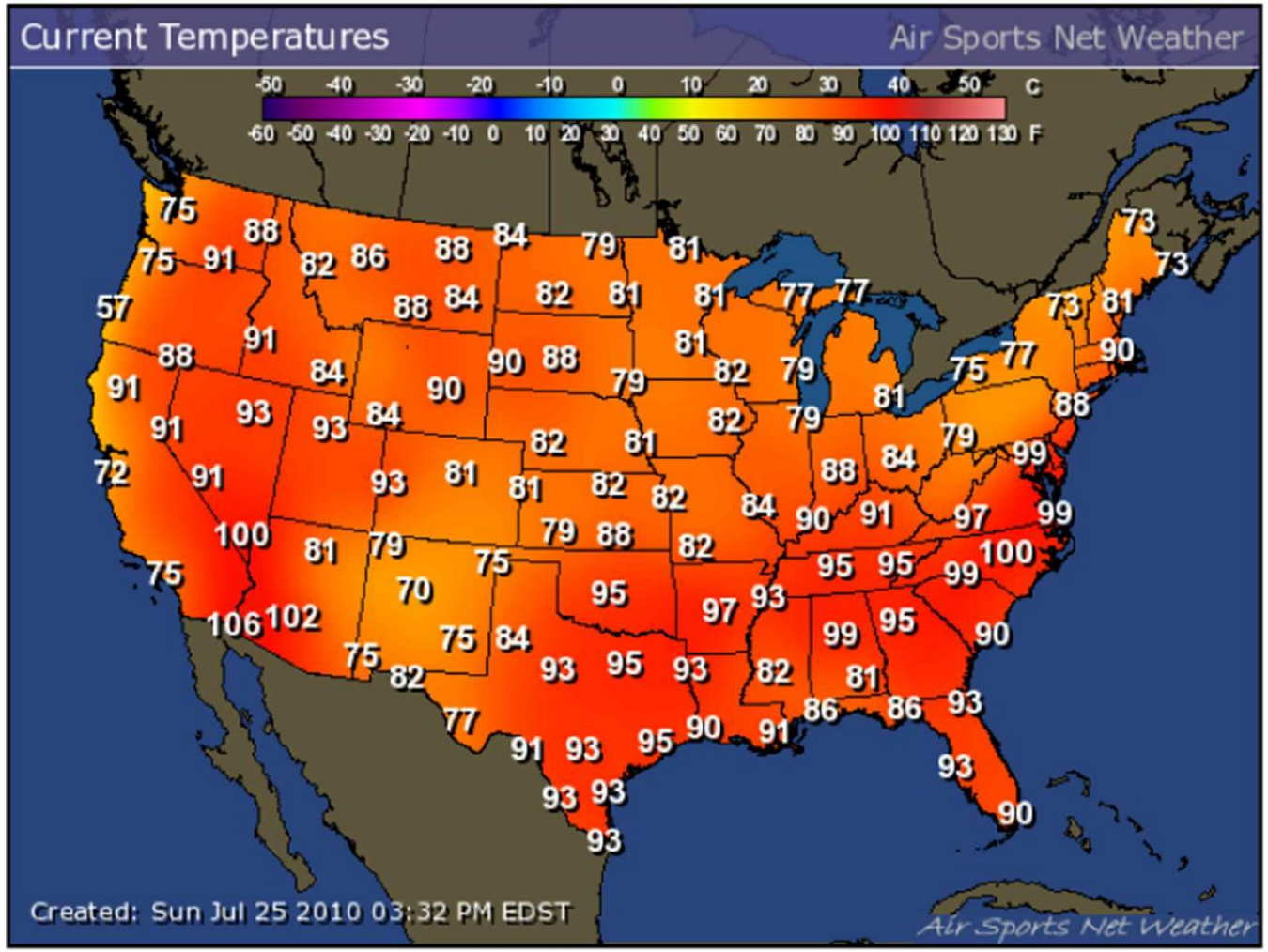
- which values are larger or smaller
- how distant two values from each other (kind of)

Also good also because

- of color psychology
- it gives exact numbers

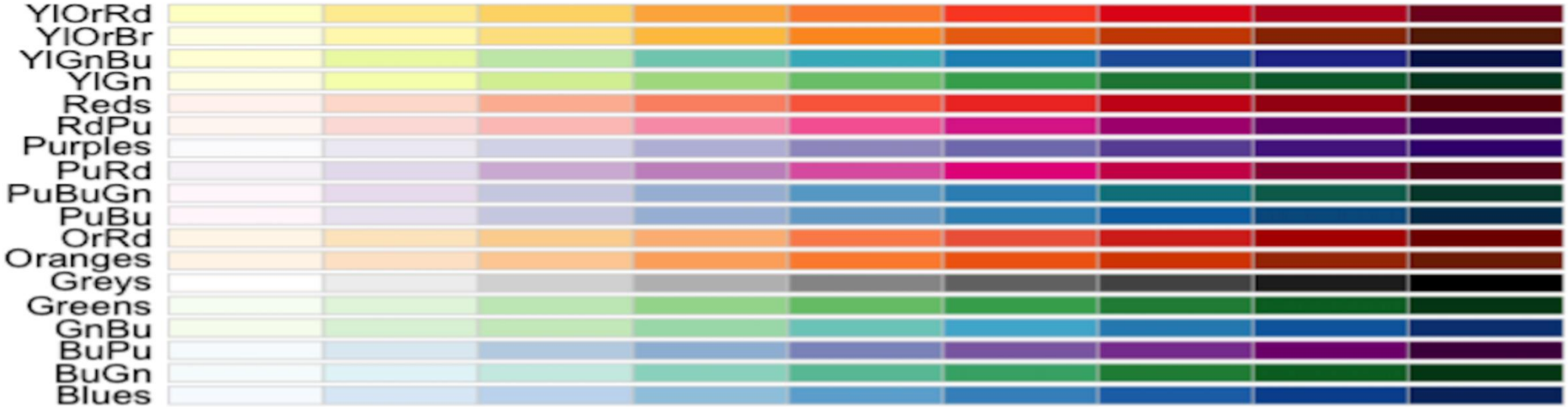
But it would be better if

- the color scale were more distinguishable
- the background were gray



http://www2.cs.uh.edu/~gnawali/courses/cosc6397-f13/intro-visualization.p

Palettes to indicate ordinal values



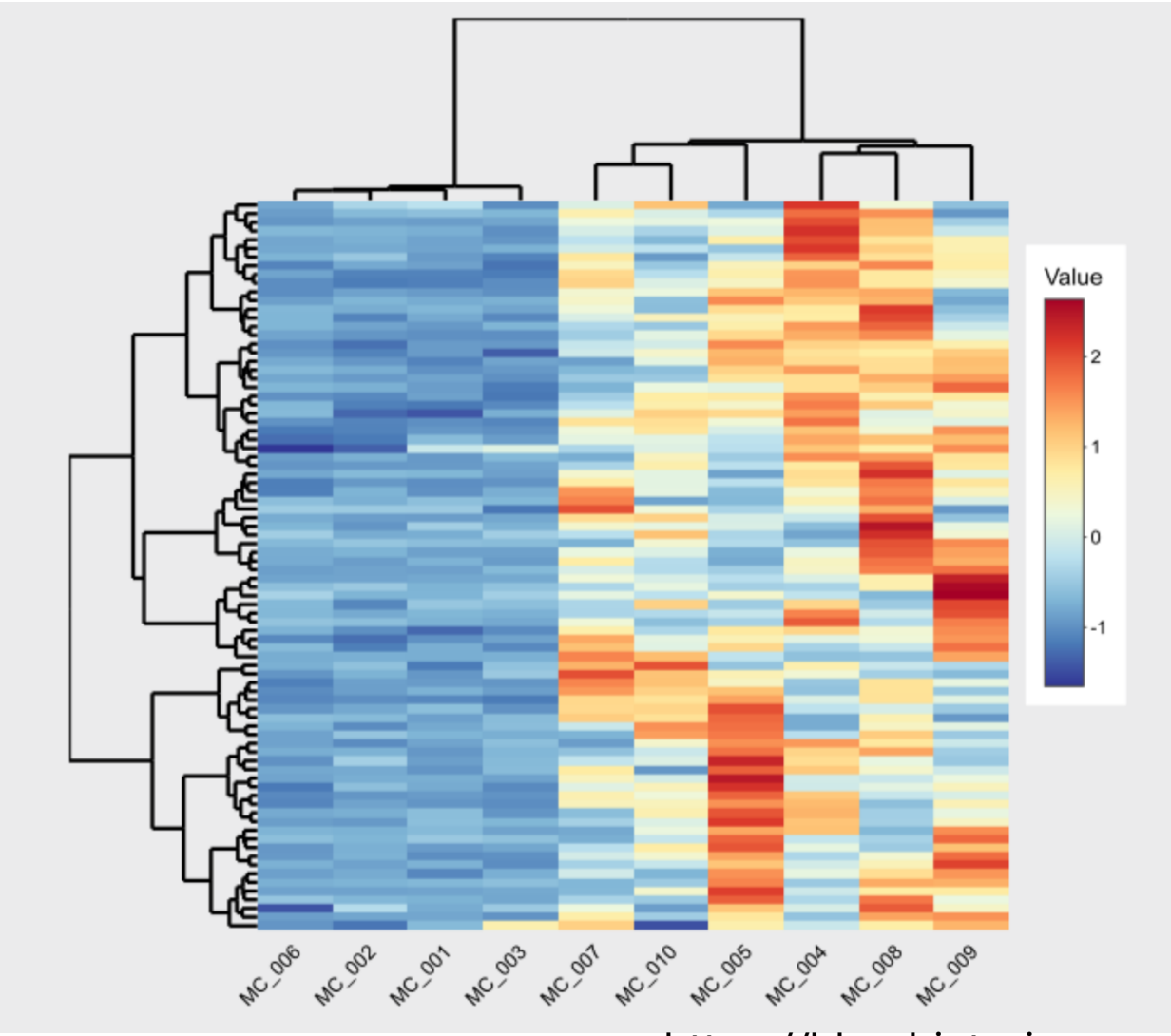
https://python-graph-gallery.com/python-colors/

<u>Sequential</u> palette in matplotlib for ordinal values

- Ordinal relationship
- Spread evenly across

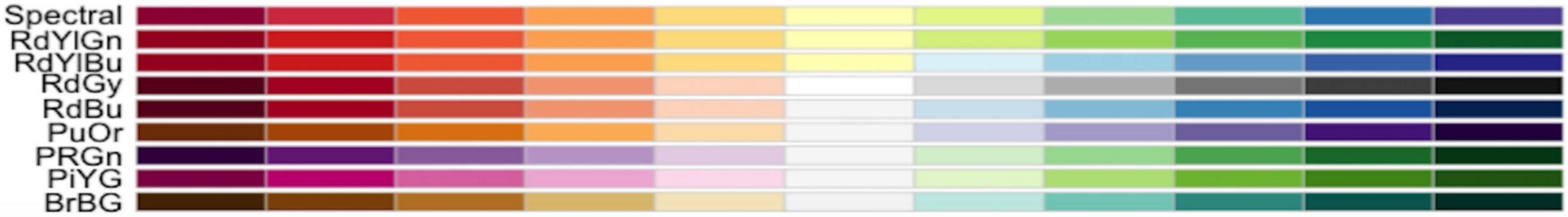
Colors to indicate values and directionalities

Hypothetical example showing gene expression change (case vs. control)



https://blog.bioturing.com/2018/09/24/heatmap-color-scal

Palettes to represent values and directionalities



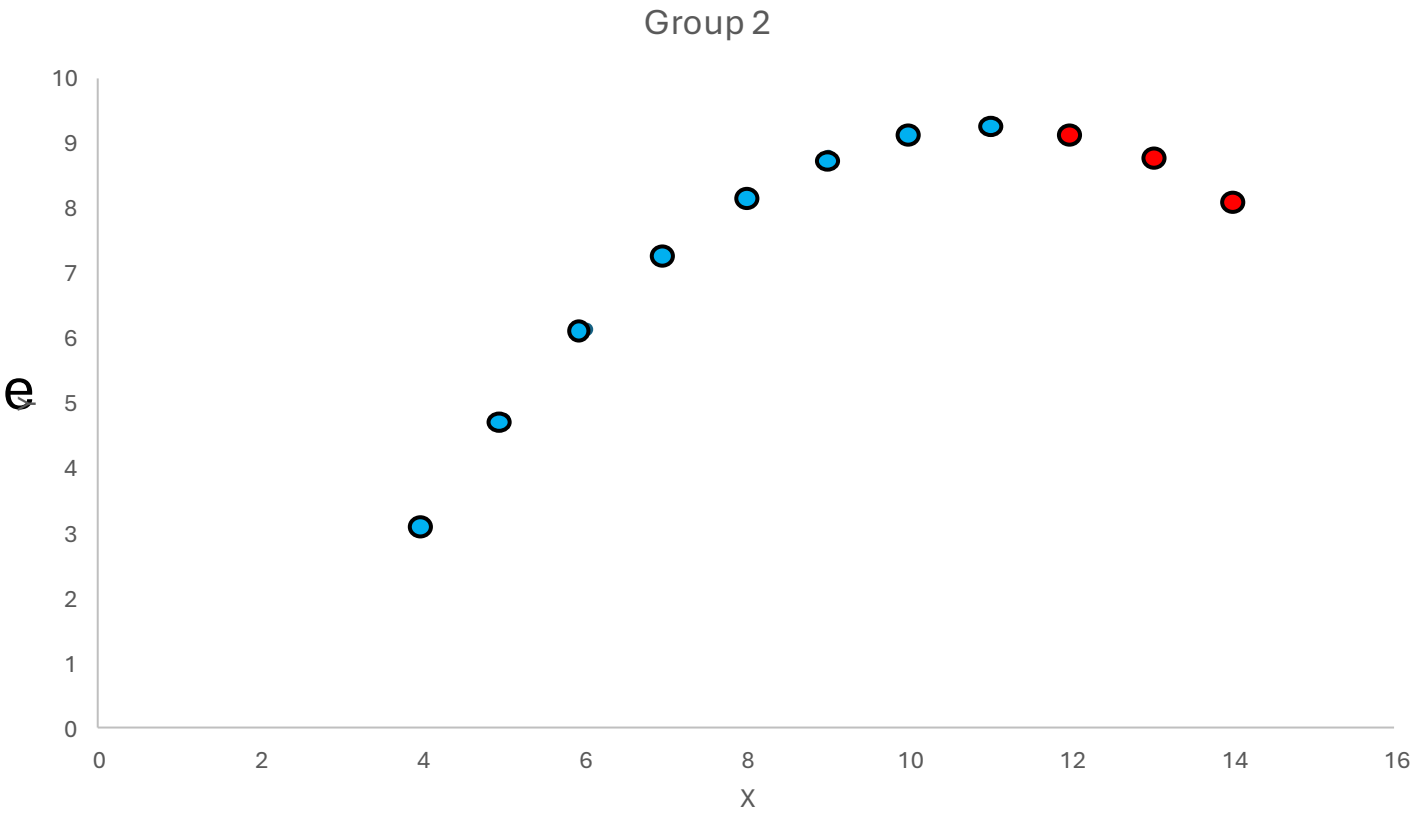
https://python-graph-gallery.com/python-colors/

<u>Diverging</u> palette in matplotlib for ordinal values.

The middle/reference value should map to the “middle”/”neutral” color.

Colors to highlight

- We wish to highlight the 3 rightmost points
- Red is OK for doing that
- Red is *not* OK when it is put with blue
- The colors *compete for attention*

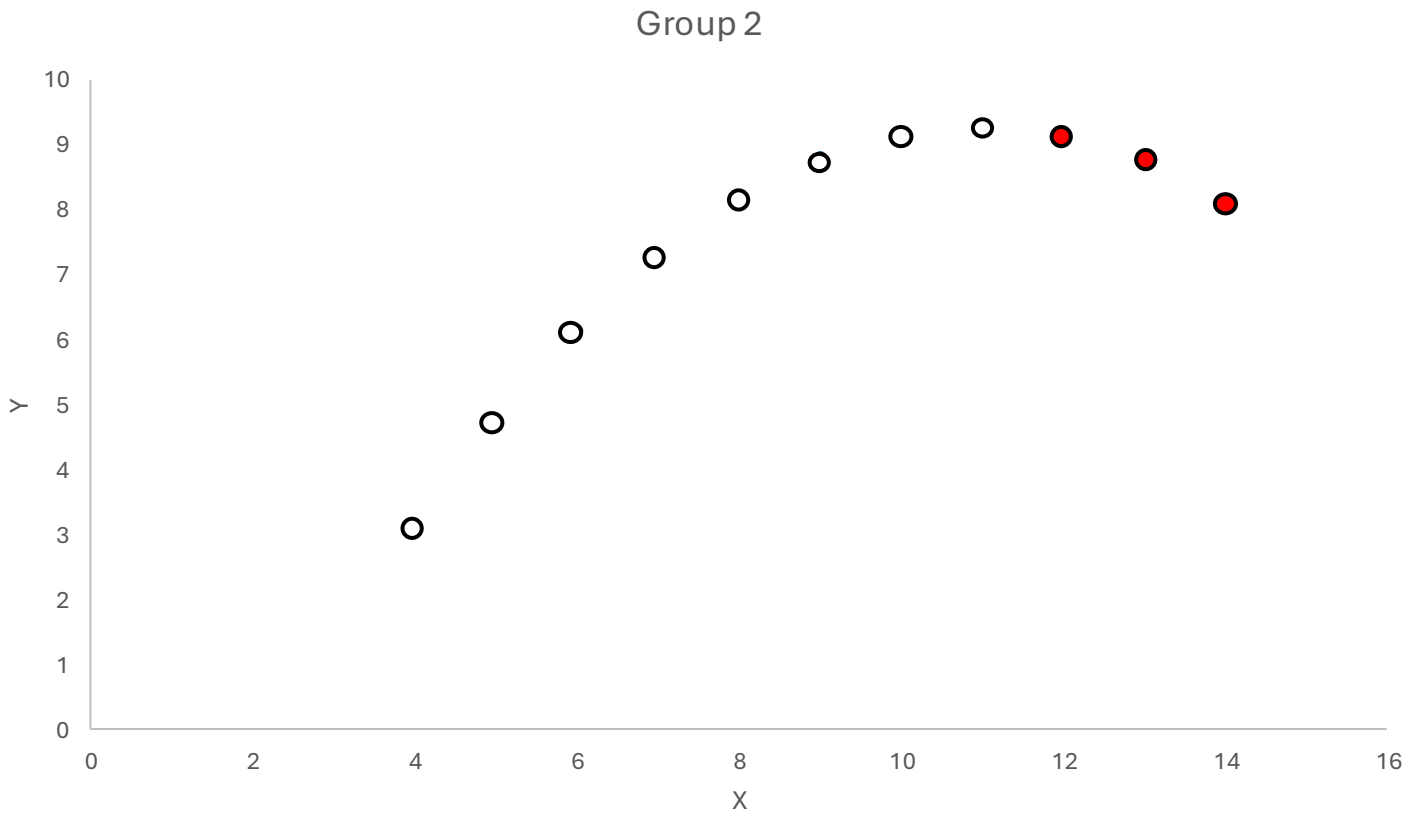


Colors to highlight

This is better.

Clearly focuses attention on the highlighted points, without any distraction due to competing colors.

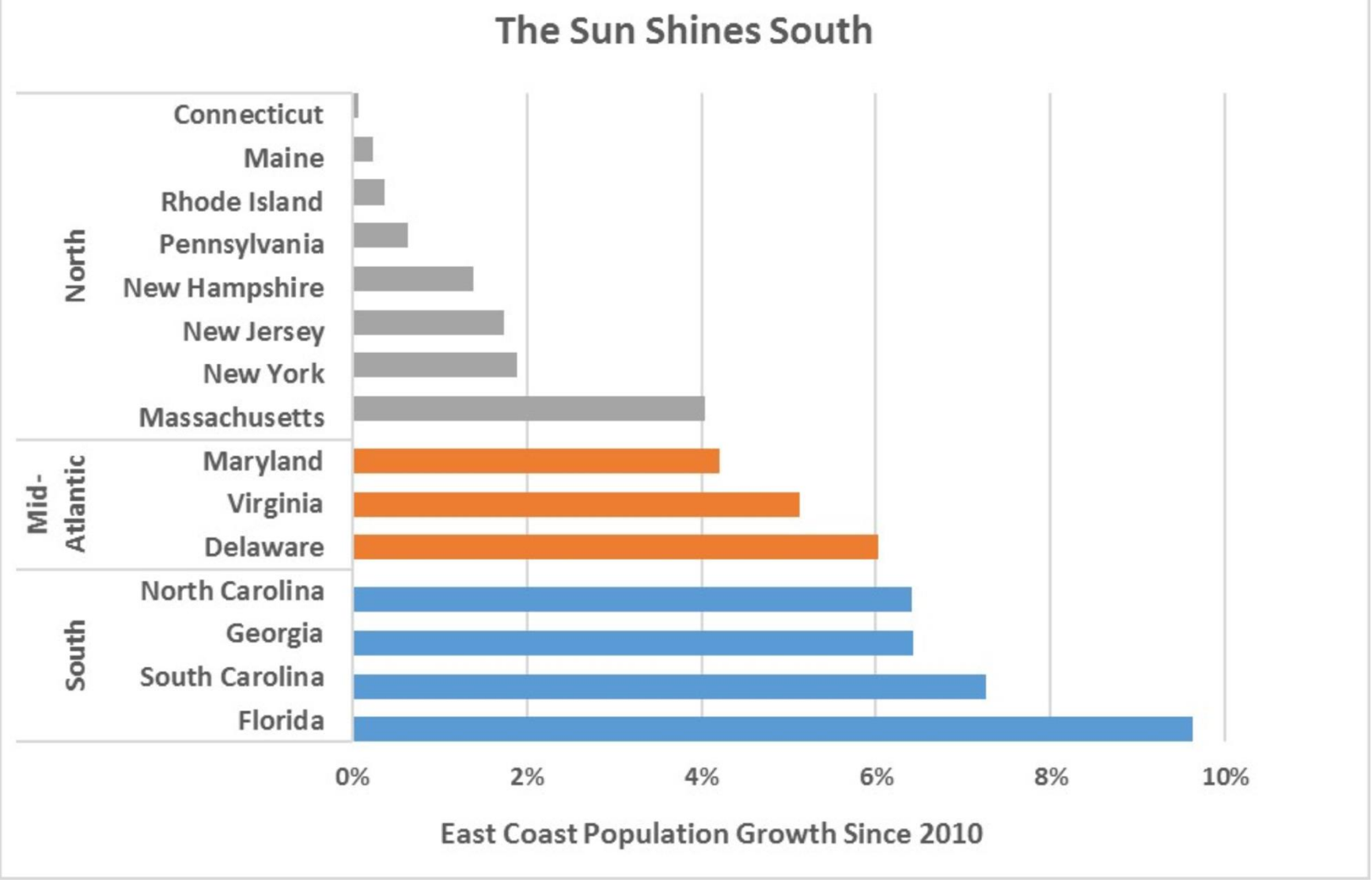
You need to set a clear focus <u>before</u> coloring!



Colors to highlight

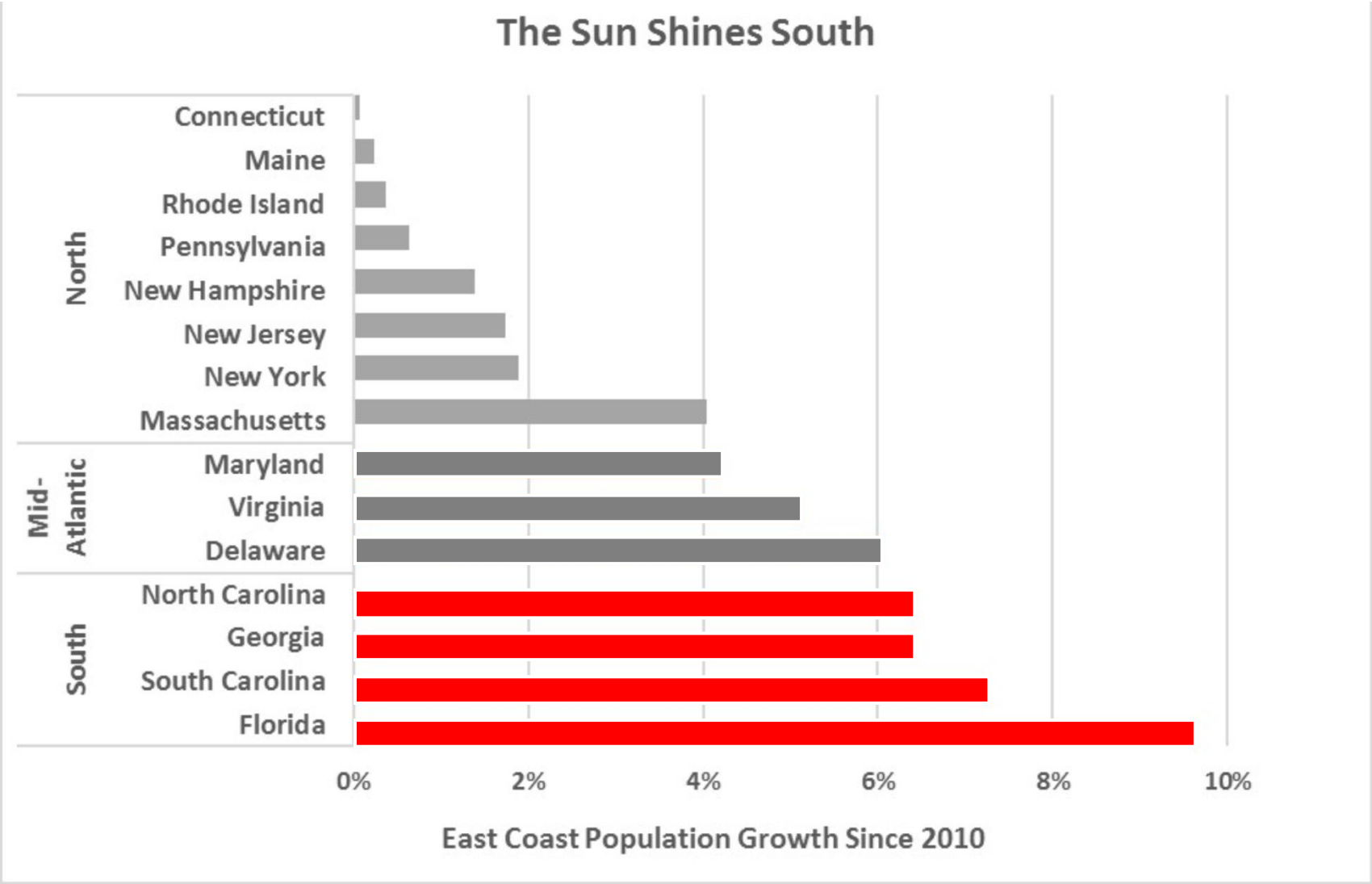
How does the perceived message/focus change with different highlights?

Compare this and the following two versions of the same figure.



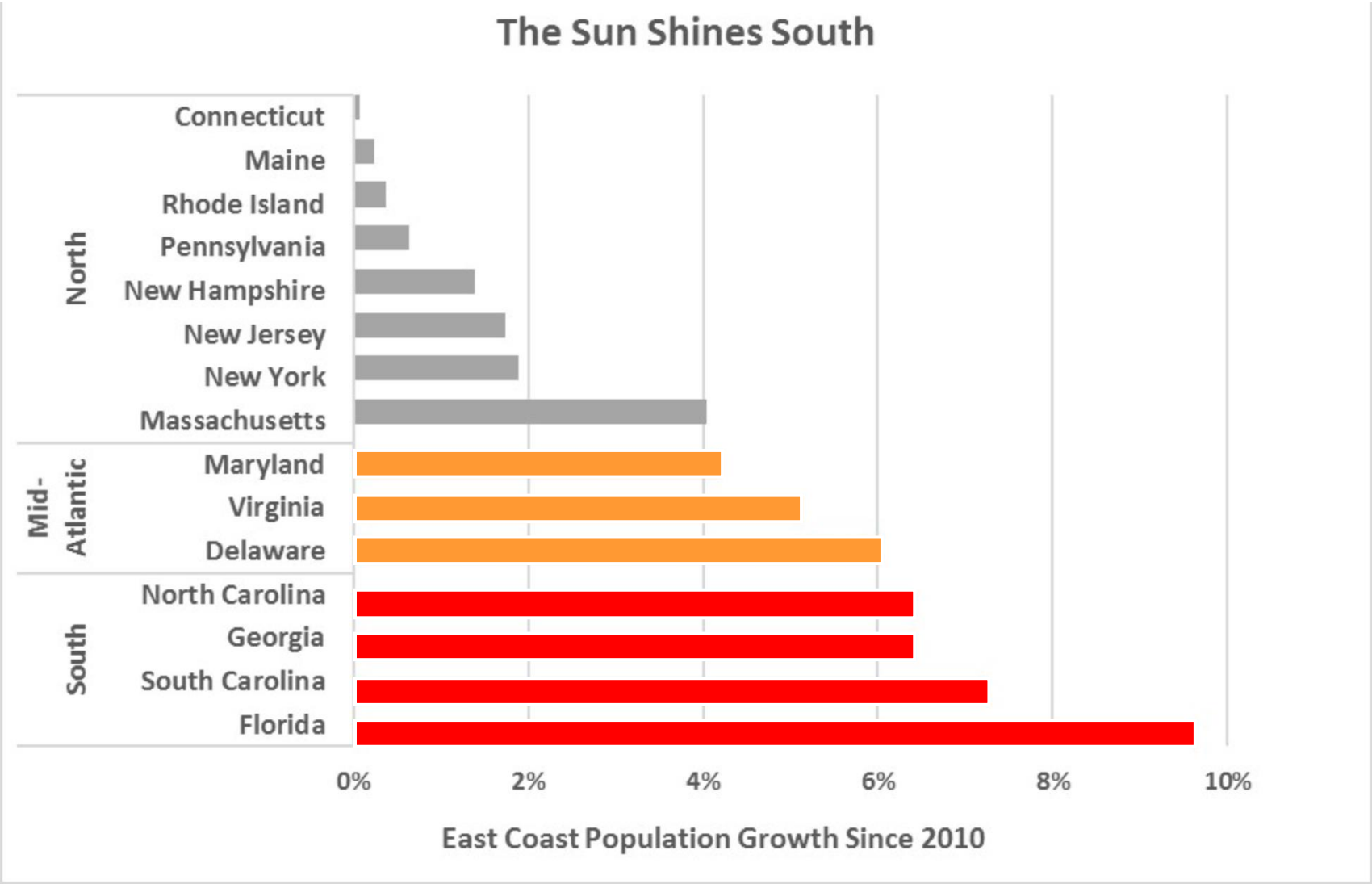
Decennial Census and 2016 State Population Estimates

Colors to highlight



Adapted from Decennial Census and 2016 State Population Estimates

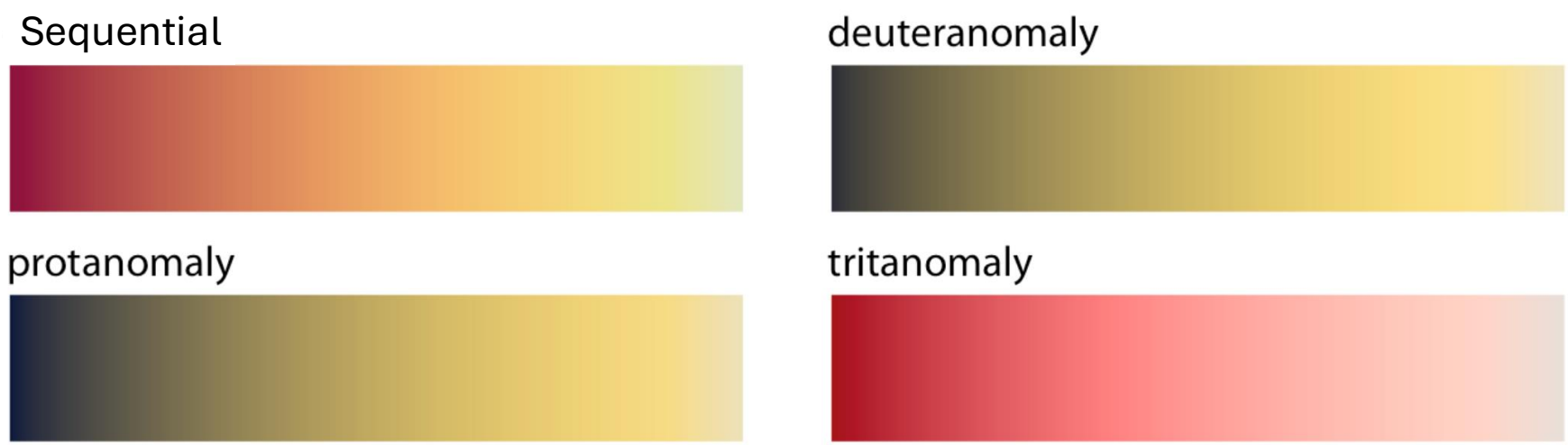
Colors to highlight



Adapted from Decennial Census and 2016 State Population Estimates

Colorblind-friendly visualizations

- ~ 8% of males and 0.5% of females suffer from some sort of color-vision deficiency
  - Less/not sensitive to green (deuteranomaly/deteranopia)
  - Less/not sensitive to red (protanomaly/protanopia)
  - Blue–yellow color-vision deficiency (tritanomaly/tritanopia)
- Deuteranomaly is the most common form whereas tritanomaly is relatively rare
- Sequential scales probably not a problem – gradient still clearly per



Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed. (see chapter

Colorblind-friendly visualizations

Diverging scales can be more difficult to make robust

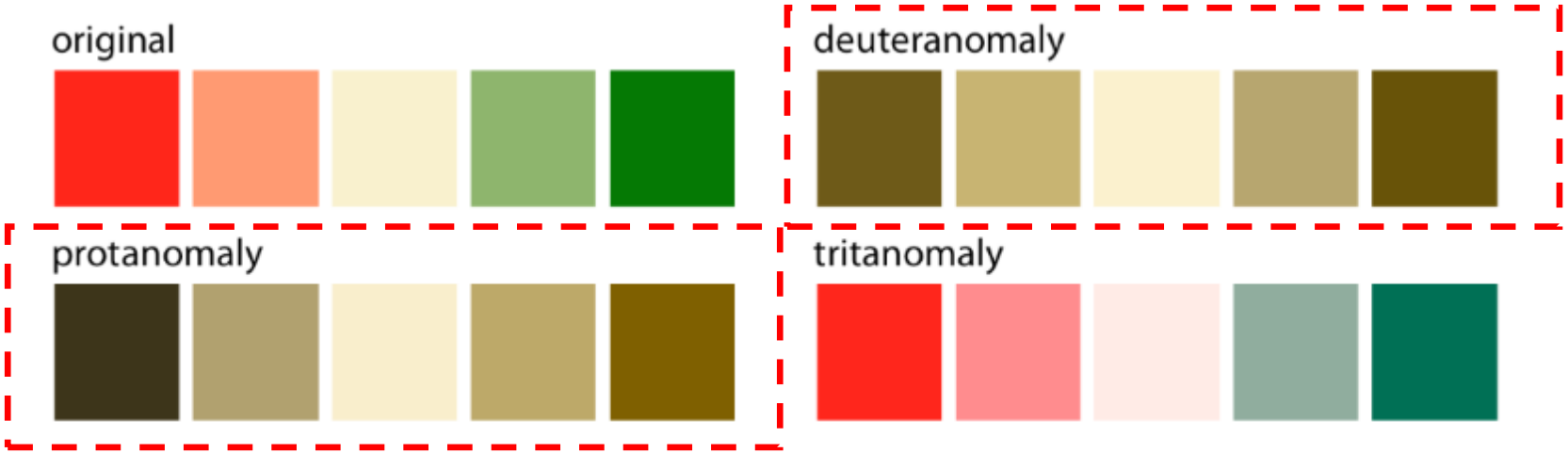


Figure 19.7: A red–green contrast becomes indistinguishable under red–green cvd (deuteranomaly or protanomaly).

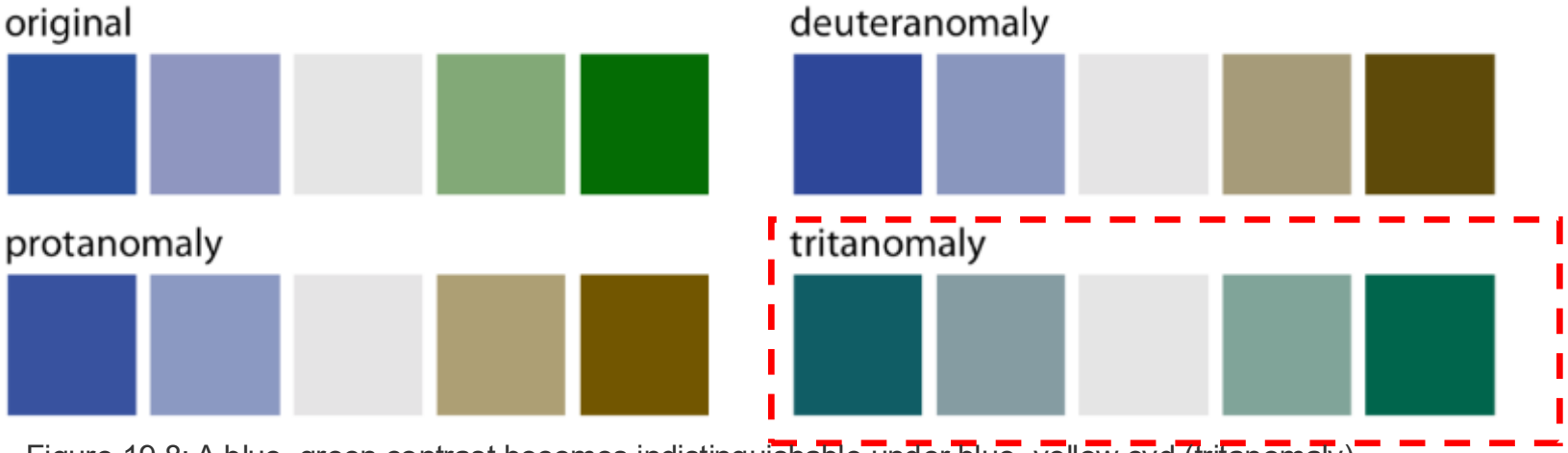
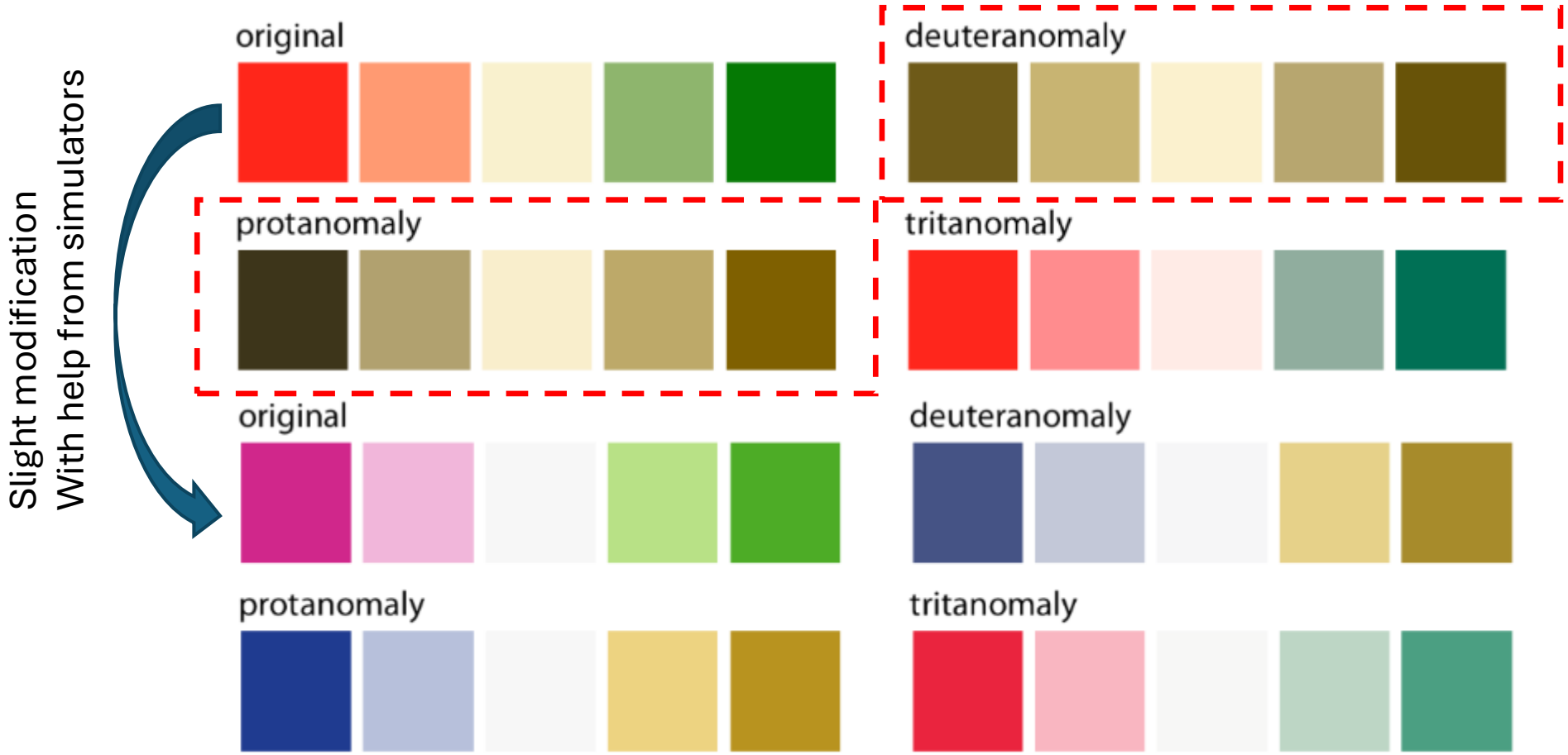


Figure 19.8: A blue–green contrast becomes indistinguishable under blue–yellow cvd (tritanomaly).

Colorblind-friendly visualization

- The “original” on the bottom is more robust as a diverging scale



- Categorical scales can also be problematic. Use a web-based simulator

https://davidmathlogic.com/colorblind/#%23D81B60-%231E88E5-%23FFC107-%23004D40

A wealth of colors to use in R or Python

- Which colors to use for effective communication?

https://python-graph-gallery.com/python-colors/

Color psychology?

Color can

- Evoke emotions
- Create impressions

e.g. red, blue, gray, black, green and brown

- Help strengthen your argument

The meaning of colors may be context-specific.

Science does not necessarily share all the impressions.

https://www.greenapplelane.com/blog/logo-design-color-

Same graphic with different colors creates different meanings

Store locations in red

→ predatory nature of company A

https://www.litigationinsights.com/best-colors-trial-graphic

Same graphic with different colors creates different meanings

The impression of the graphic shifted

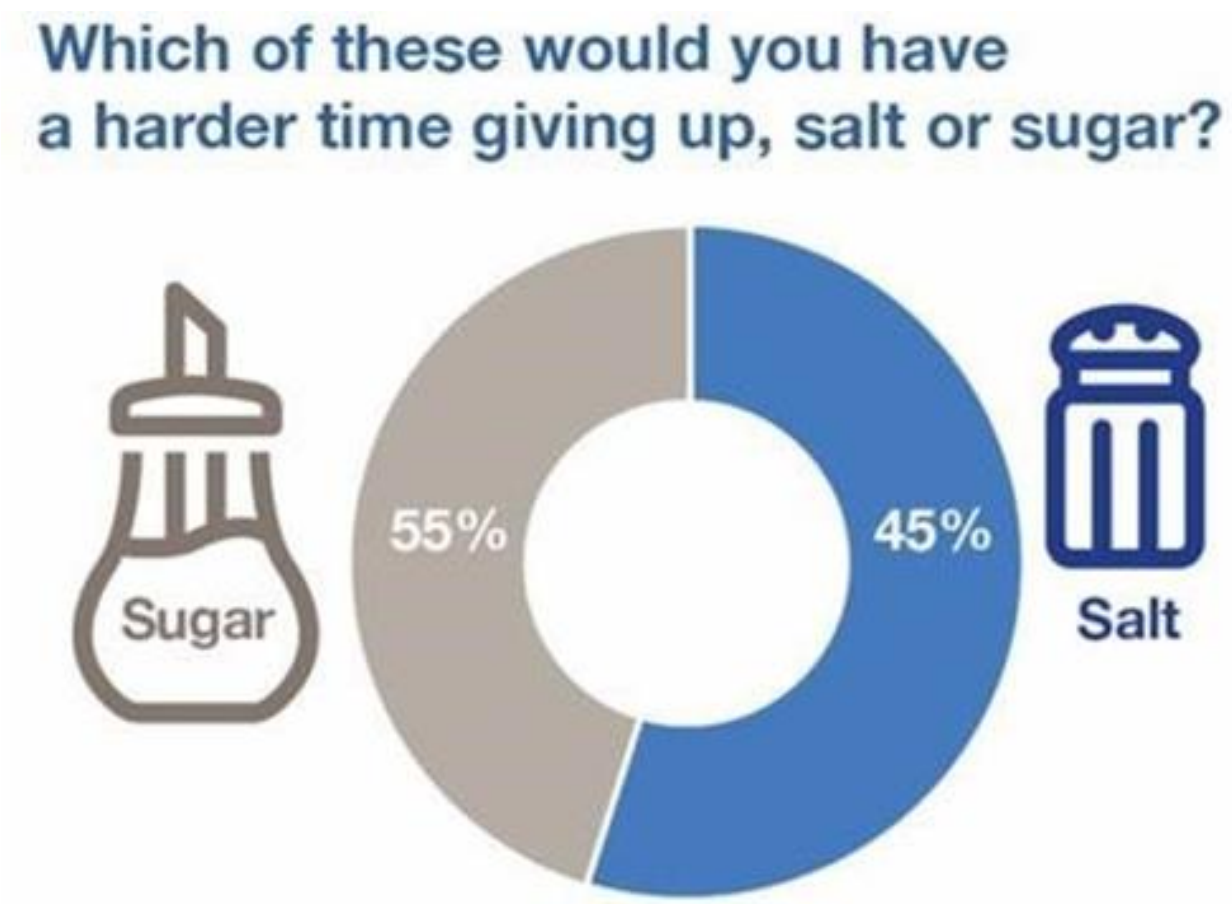
- Store locations in green

➡ company growth and expansion

https://www.litigationinsights.com/best-colors-trial-graphic

Color misuse makes counter-intuitive figures

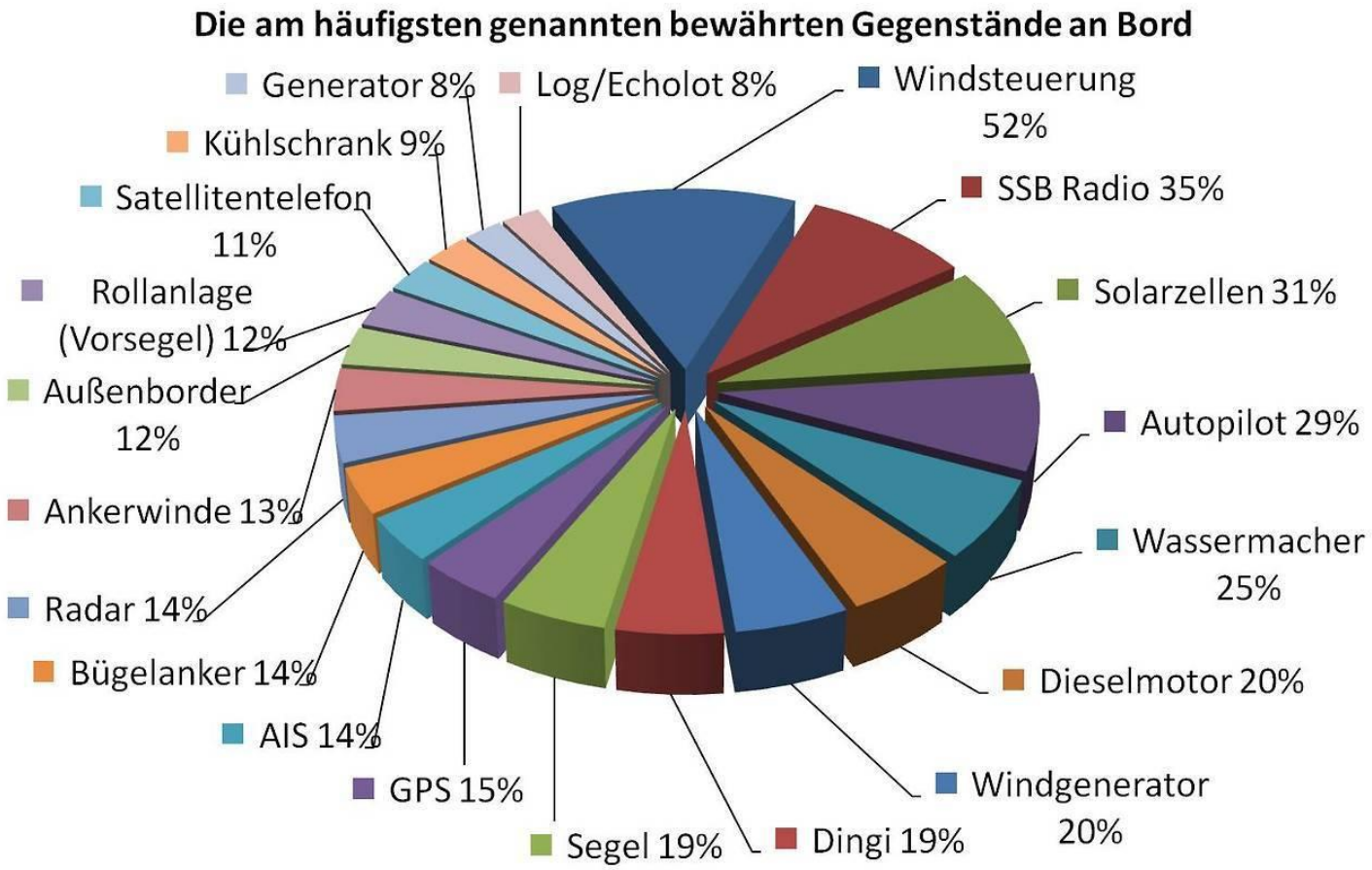
Good visualization reflects actual values



https://viz.wtf/post/169392205697/fewer-people-would-have-a-hard-time-giving-up

Good visualization follows assumptions

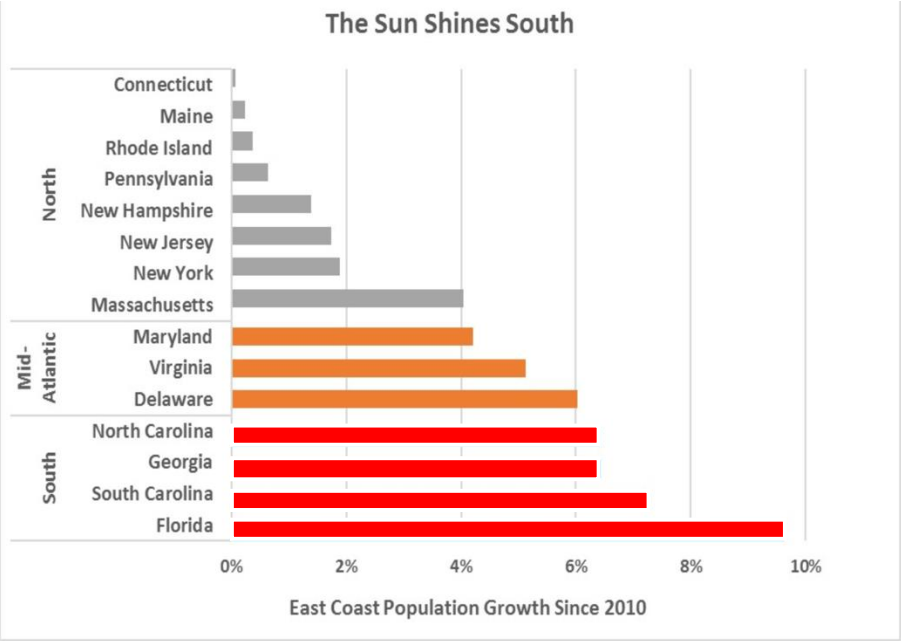
E.g., what is always assumed in a pie chart?

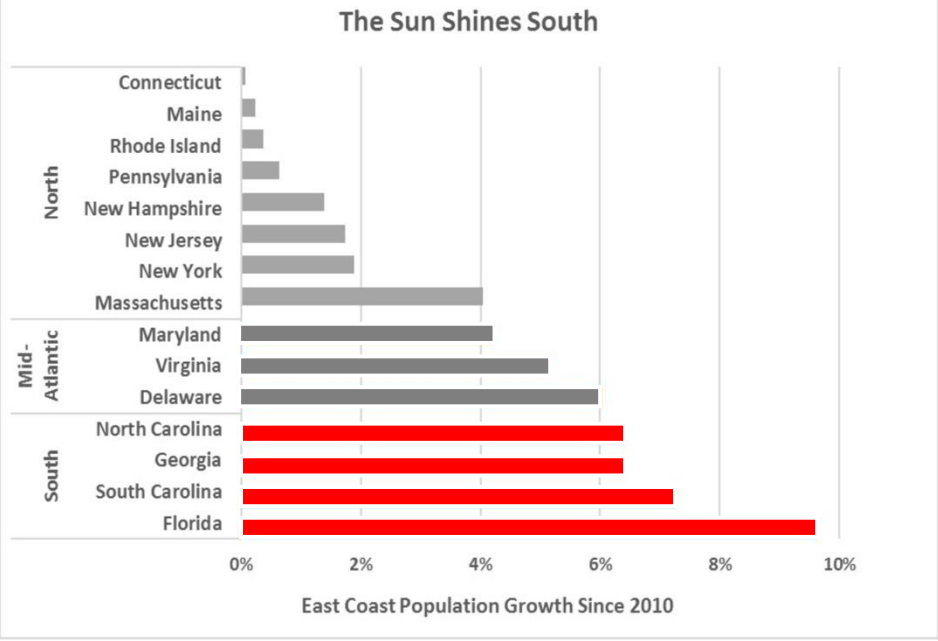


https://viz.wtf/post/162907112938/i-guess-its-best-to-show-all-the-data-and-

Good visualization highlights your message

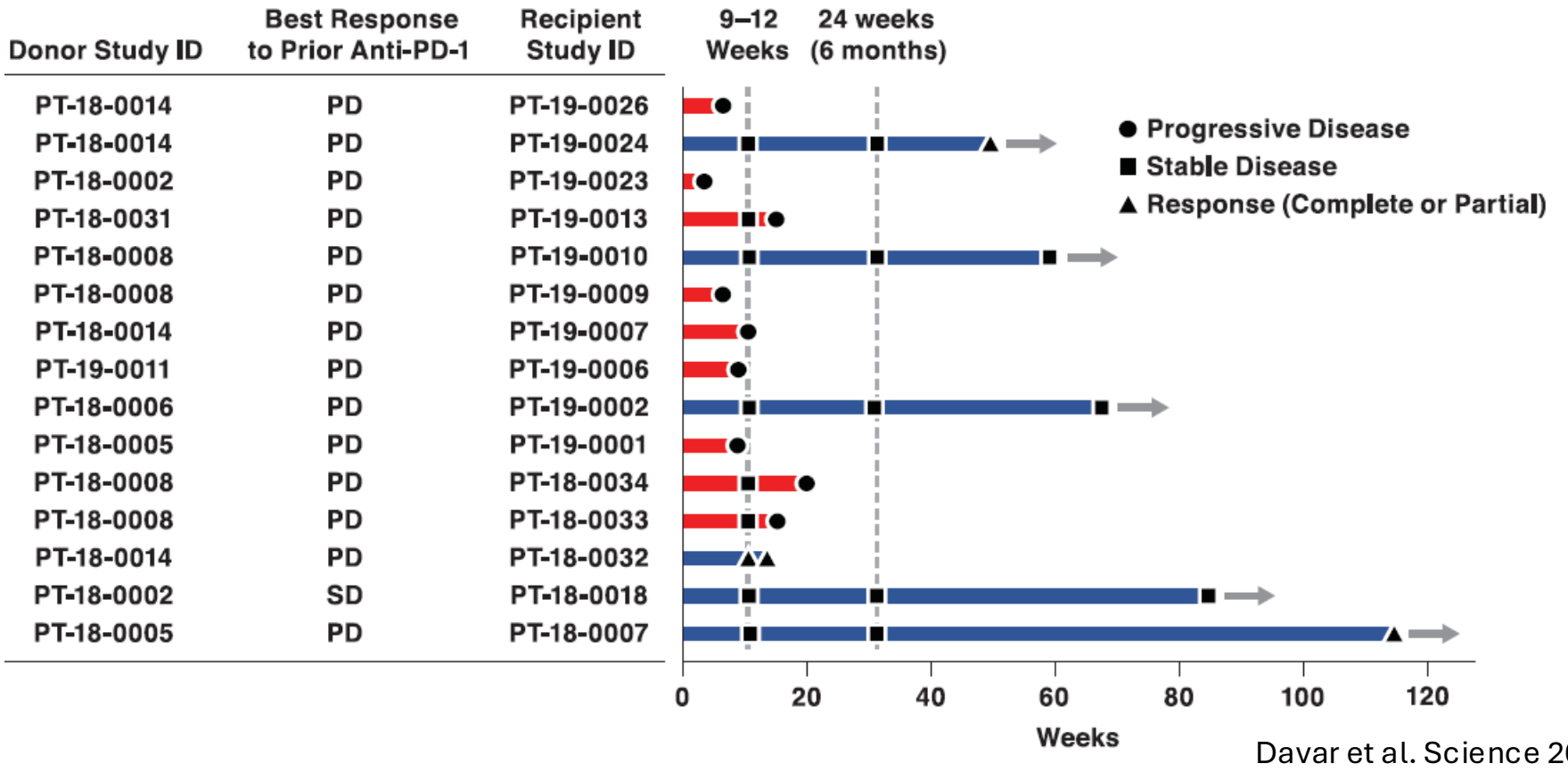
- Matters more in presentation
- Clearly identify the message first





Adapted from Decennial Census and 2016 State Population Estimates

Color helps focus our attention on the message



Davar et al. Science 2021

Is “Stable Disease” good or bad?

Do not trust defaults

Good visualization

Unlike in visual arts, there are clearly good and bad visualization in science.

Tufte in *The Visual Display of Quantitative Information* says:

> Excellence in statistical graphics consists of complex ideas communicated with clarity, precision, and efficiency. Graphical displays should
>
> - show the data
> - induce the viewer to think about the substance rather than about methodology, graphic design, the technology of graphic production, or something else
> - avoid distorting what the data have to say
> - present many numbers in a small space
> - make large data sets coherent
> - encourage the eye to compare different pieces of data
> - reveal the data at several levels of detail, from a broad overview to the fine structure
> - serve a reasonably clear purpose: description, exploration, tabulation, or decoration
> - be closely integrated with the statistical and verbal descriptions of a data set.

Suggestion: when you make or see a figure, always ask the purpose.

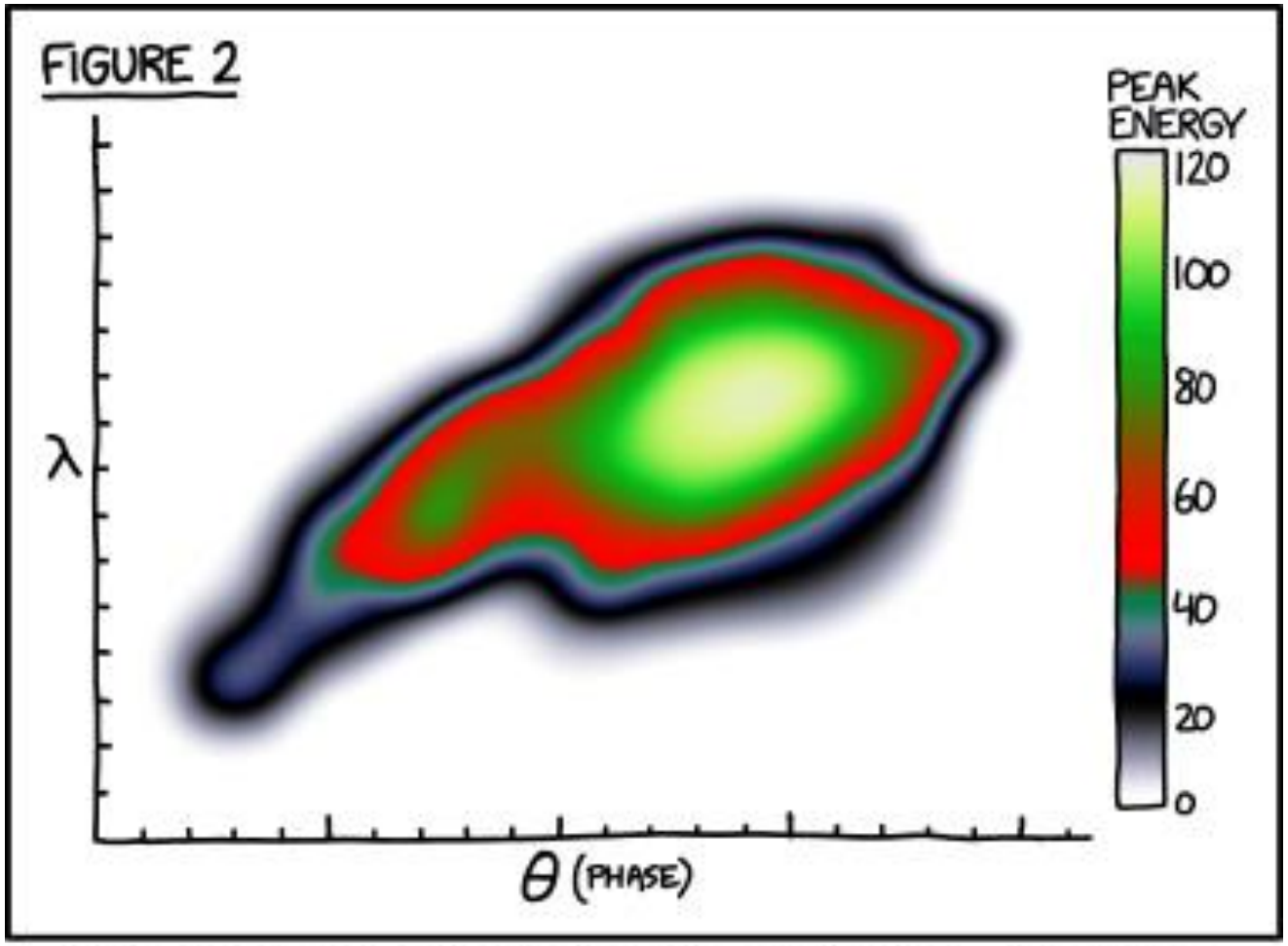
"Why am I making this figure?"

"In one sentence, what message am I trying to convey?"

"Why did the author bother including this figure?"

Visualization software/packages

- Biorender
- Cytoscape
- Python (Matplotlib)
- Inkscape, GIMP, Illustrator
- TikZ, PGF
- ImageMagick
- D3.js
- R/ggplot2
- Tableau



EVERY YEAR, DISGRUNTLED SCIENTISTS COMPETE FOR THE PAINBOW AWARD FOR WORST COLOR SCALE.

https://imgs.xkcd.com/comics/painbow_award.png