Ethics of Visualization and Storytelling with Data

Table of contents

Storytelling with Data

Storytelling with data

After data analysis, how should we communicate our findings?

- Effective communication often involves **telling a story**.
- The story should reflect what you believe to be true based on a rigorous analysis of the data
  - The story should reflect not just your conclusions but all assumptions and your process.
- Prepare visualizations that support your story and make a clear point.

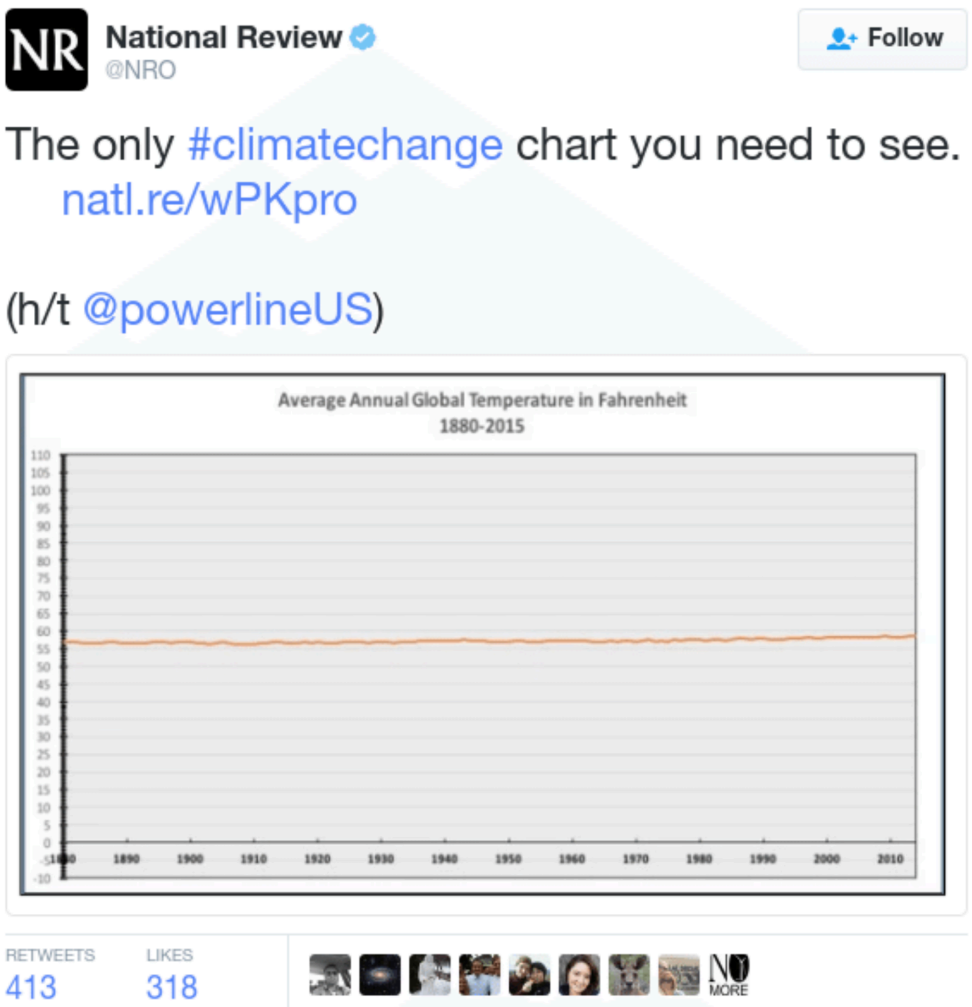
What is a story?

> A story is a collection of observations, facts, or events presented in a specific order such that they create an **emotional reaction**. - Clause Wilke

💬 Possible Discussion Questions 💬

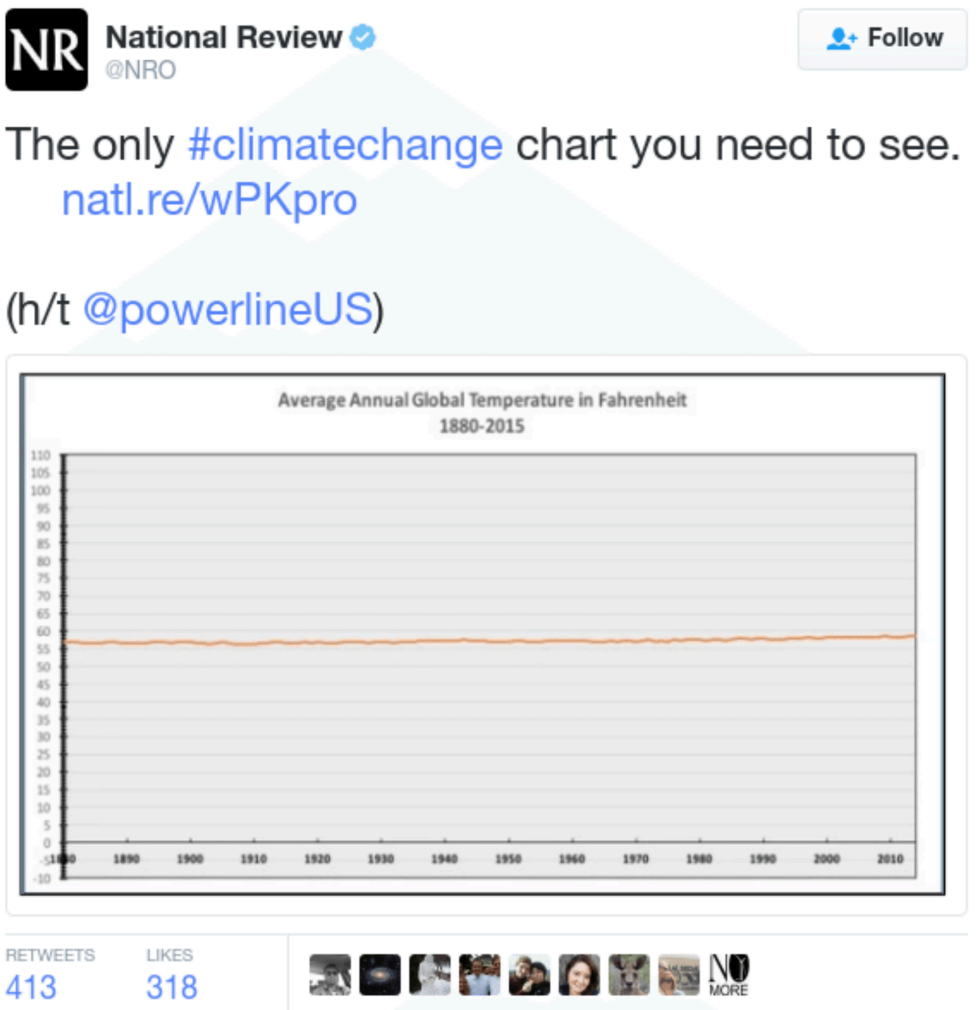
- The previous quote is somewhat provocative in the context of storytelling with data.
  - Why should we seek to create an **emotional reaction**? Shouldn't the goal be honest and clear reporting of evidence?
  - A story can create an **emotional reaction** but be misleading!
  - Discuss the phrase "Lies, damn lies and statistics!"
- We like to say "let the data speak for itself". Is that really possible? What would that require?

A Story About Climate Change



In 2015, the National Review, a conservative magazine, tweeted a chart that displayed the global temperature in Fahrenheit from 1800 to 2015. Noting that it was “the only chart you need to see”.

A Story About Climate Change

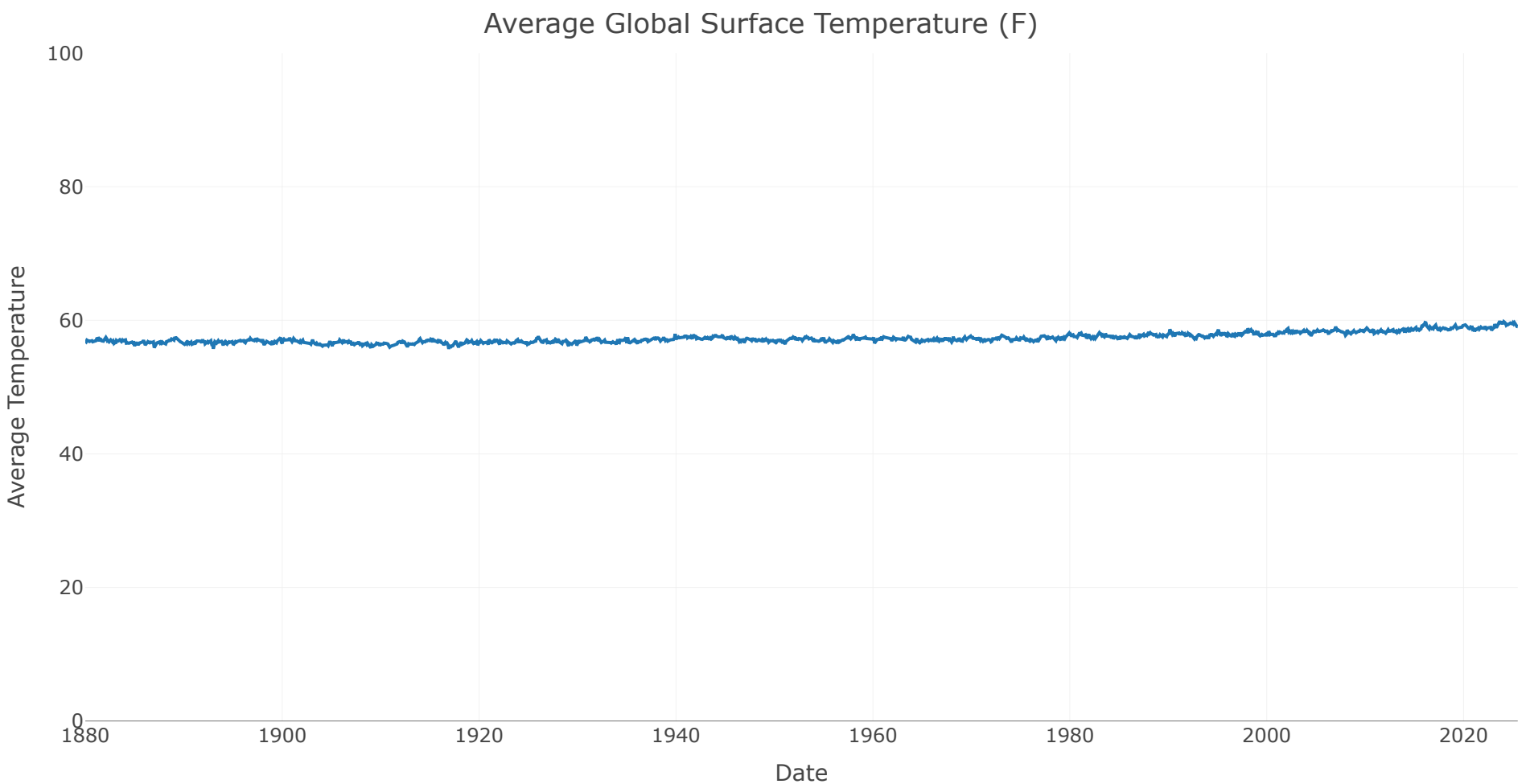


- Does this visualization make a point in service of a larger story?
- Is the point made by this graphic fair or misleading?

A Story About Climate Change

- The data are from NASA's Goddard Institute for Space Studies (GISS) analysis of global surface temperature change. (Hansen et al. 2010)
- The world's temperature is initially recorded in degrees Celsius.
- The GISS analysis combines available sea surface temperature records with meteorological station measurements of temperature.
- Then compared to the years 1951–1980 as the base period to determine change.
- Additional adjustments to the average global surface temperature are made, such as accounting for heterogeneity to minimize local non-climatic effects.
- The Washington Post addressed the tweet by identifying that the National Reviewer had scaled the plot to "hide the actual change in temperature."
- The National Review failed to address how they treated the data to get the Fahrenheit values.

Interactive Visualizations



Manipulate the plot by zooming in and out and evaluate temperature at time intervals.

Instructor Note

- We start with default ylimits to match the tweet, but show how the story changes if we zoom in.
- Use the “autoscale” button at the top to zoom in or zoom in manually. Double clicking on the plot resets to the default.
- Why is it maybe not a good idea to start the y-axis at 0 in this example?

💬 Possible Discussion Questions 💬

Interpretation:

- Is a 1 degree change a lot? Based on day to day experience it seems like not, but there is more to the story.
- Additional plots or results could drive home the story about impact of a 1 degree change. Hypothetical plots: how much does a 1 degree increase in global temperature impact the prevalence of wildfires? How much does a 1 degree increase impact the abundance of important species in the ecosystem?
- Domain expertise needed to contextualize findings!
- GISS show “temperature anomalies” by subtracting the mean temperature over some baseline period (1951-1980). Why might this be more desirable?

💬 Possible Discussion Questions 💬

What are some pros and cons of interactive visualizations?

- Pro: gives users the ability to explore the data and discover new patterns
- Con: lose control of the narrative, easier for others to distort findings
  - Figures should be designed to tell a story and have a clear purpose. Simply sharing an interactive visualization might make communication harder.
- Should a static visualization start the y-axis at 0 in this plot? Why or why not?

Poorly constructed vs unethical visualizations

- There is a difference between an **ugly** visualization and a **bad** visualization.
  - Ugly visualizations are aesthetically unappealing but can still effectively convey information.
  - Bad visualizations are misleading or confusing because of the way they are constructed.

Poorly constructed vs unethical visualizations

- There is also a fine line between poorly constructed visualizations and unethical presentations.
- Data scientists should remain honest, provide clarity and avoid violations of openness or objectivity.
- Ensure that your audience has enough information to make an informed decision that is not unduly influenced by the presenter's values and biases. (Schroeder 2022)

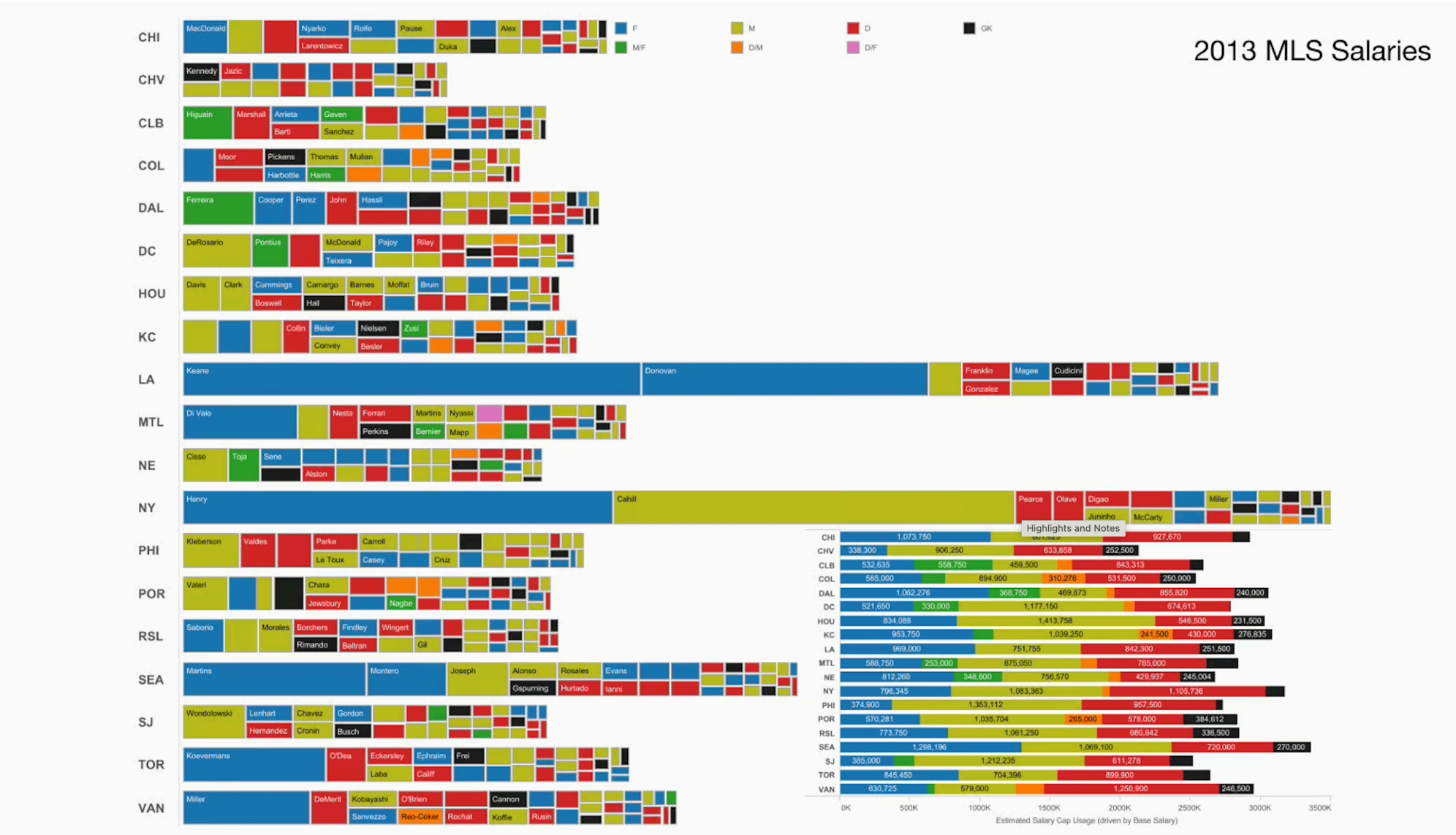
Features of Poorly Constructed Plots

The use of 3D Graphs

Excluding Information

- Excluding information that may impact a reader's understanding includes:
  - x- and y-axis labels,
  - using a variable scale of measurement (e.g. kg., log-transformation, etc.),
  - clear title or figure labels.
- **Deliberately** excluding data or information that might influence the conclusions would be unethical.

Including too Much Information



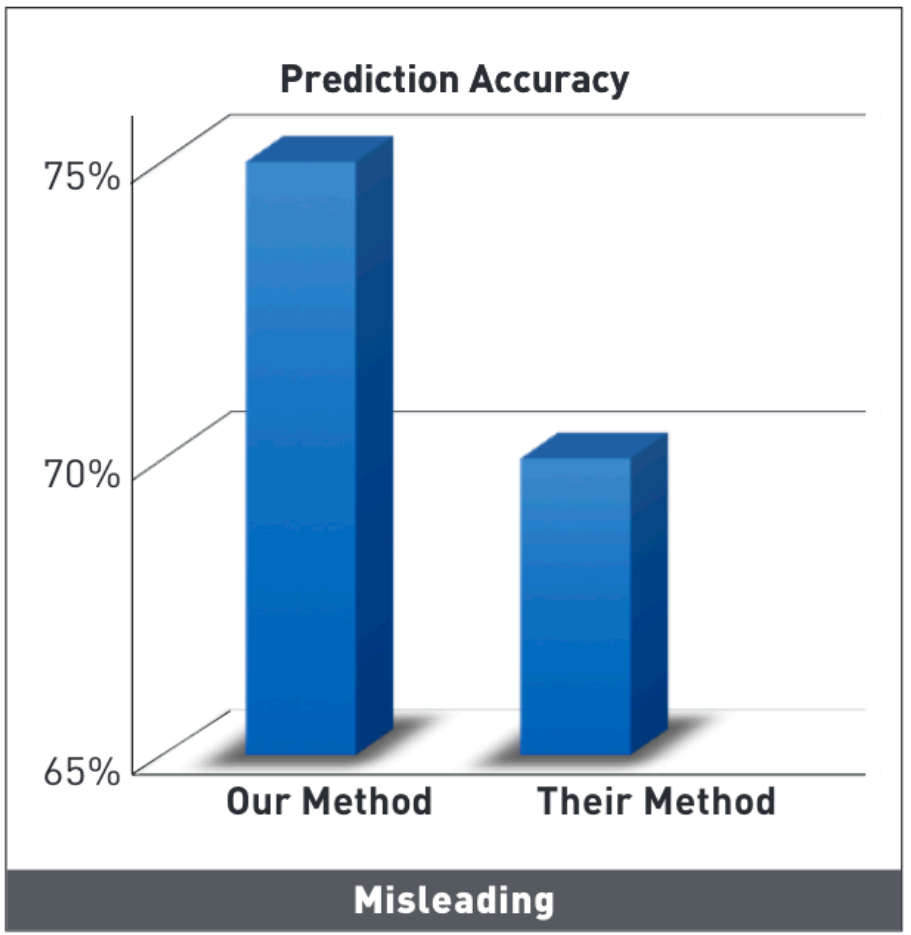
Including too Much Information

- Including **too much** information can have the same effect as leaving out information.
  - Leaves the audience with more questions and confusion.
  - General rule:
    - Provide enough information so that your audience can make an informed decision.
  - Exclude unnecessary text (subtitles, long title, jargon).
  - Address disproportionate number of numerical values (i.e., too many values on the y-axis).

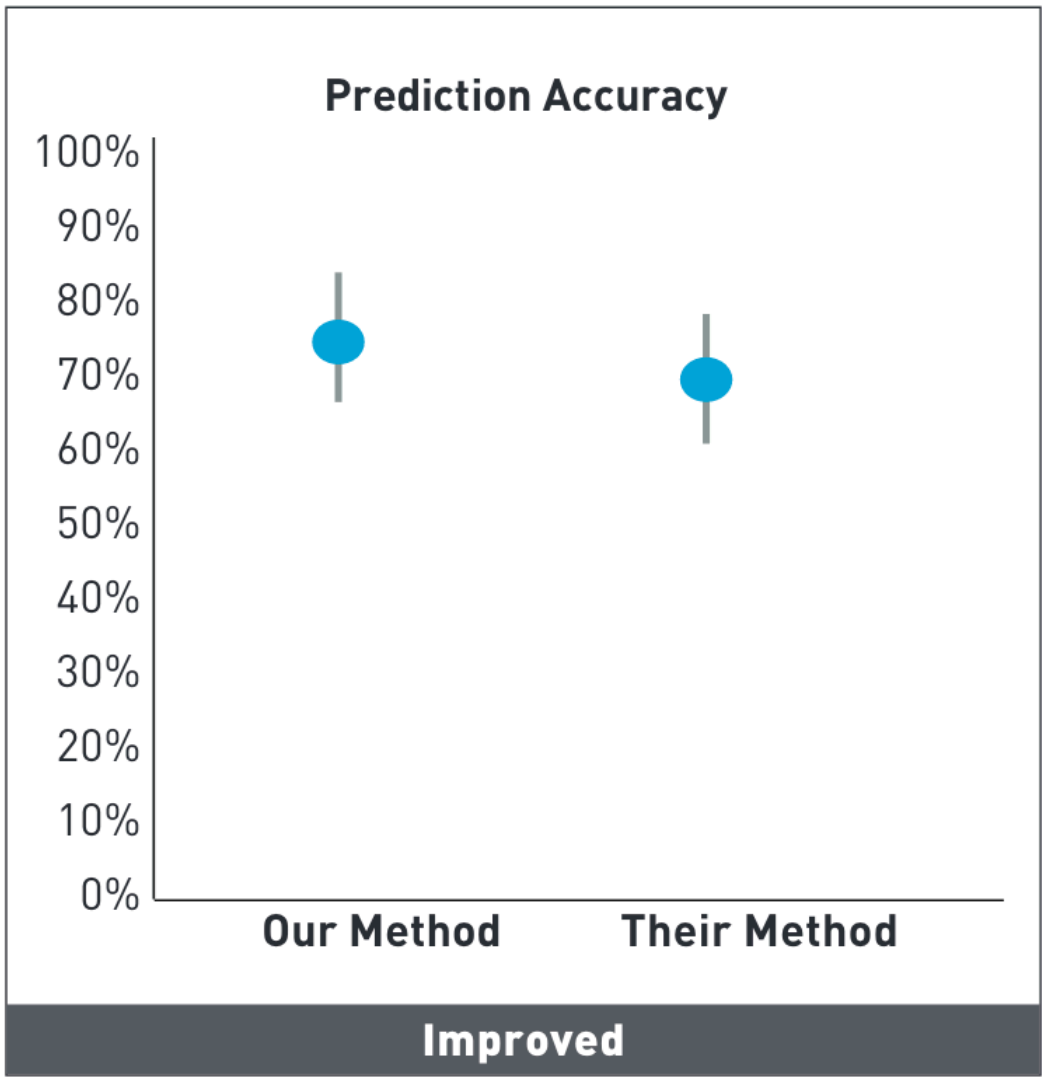
Distorted axis limits

- Some functions by default fit axis ranges to natural data scales.
- Non-zero y-axes may cause small differences to appear much larger than truth.
- When dealing with normalized data ensure that the variables are placed on the same **scale**.

Distorted axis limits



Improved axis limits



Visual Communication

> Excellence in statistical graphics consists of complex ideas communicated with clarity, precision, and efficiency.”
>
> – (Tufte 2001)

Instructor Note

Edward Tufte is a Yale professor who is well known for his studies on the characteristics of data visualizations and how such characteristics impact readers. His foundational book ‘’The Visual Display of Quantitative Information’’ (1983) reviews several principles to creating informative data graphics.

How does Tufte’s quote on communicating with clarity, precision and efficiency relate to some of the previous discussions around storytelling with data and creating an emotional reaction?

Tufte’s graphic principles to creating useful figures:

Accessibility

Accessibility & Color

- Human perceptions of color vary. Upwards of 5-8% of men are colorblind! (less common for women)
- Colorblindness and complete blindness present accessibility concerns related to visual presentation of information. (Johnson 2024)
- (Petroff 2024) used a machine learning technique to identify color sequences which considered color blindness types, color lightness, color sequence preferences, and aesthetic preferences.

Accessibility & Visualizations

Table 1. Final Color Sequences for Six, Eight, and Ten Colors

Six Colors					Eight Colors				
	R	G	B	min ΔE_{cvd}		R	G	B	min ΔE_{cvd}
blue	87	144	252	100.0	blue	24	69	251	100.0
orange	248	156	32	57.1	orange	255	94	2	66.9
red	228	37	54	21.3	red	201	31	22	18.2
purple	150	74	139	21.3	purple	200	73	169	18.1
gray	156	156	161	21.3	gray	173	173	125	18.1
purple	122	33	221	20.5	light blue	134	200	221	18.1
					blue	87	141	255	18.1
					gray	101	99	100	18.1

Ten Colors				
	R	G	B	min ΔE_{cvd}
blue	63	144	218	100.0
orange	255	169	14	56.8
red	189	31	1	33.4
gray	148	164	162	22.3
purple	131	45	182	18.3
brown	169	107	89	16.4
orange	231	99	0	16.3
tan	185	172	112	16.1
gray	113	117	129	16.1
light blue	146	218	221	16.1

Note: The sequences are ordered from top to bottom, and sRGB values $[0, 255]$ are given. The names shown are the most-probable names based on the color-naming model described in Appendix D. The minimum perceptual distances, min ΔE_{cvd}, take into account color-vision-deficiency simulations.

Accessibility & Color

- Surrounding colors influence the perception of other colors:
  - Choose colors that are easier to distinguish among.
- Color *association*:
  - Avoid color schemes that engage in stereotypes (e.g., black associated with unsafe and white associated with approachable).
  - Keep color schemes neutral
  - Consider how different mediums (e.g., computer screen, printouts, software) used to display the graphic may influence color.

Instructor Note

Accessibility is the practice of ensuring that all individuals have equal access to resources, programs, and opportunities, regardless of their abilities or backgrounds.

In more recent years, the study of colors in statistical visualizations has burgeoned since visualizations rely on color for many different reasons (e.g. distinction of classes, degrees,effects). Findings from those studies highlight several key themes that researchers need to consider while creating visualizations.

For instance, an individual's degree of color blindness may influence the perception of a graph or figure. One in every 12 men (8%) and one in every 200 women (0.5%) in the U.S. are born with inherited color defective vision of some type and degree (Johnson, 2024). The most common color-blindness types are red-green and blue-purple. Any color that contains the affected primary color is *distorted*.

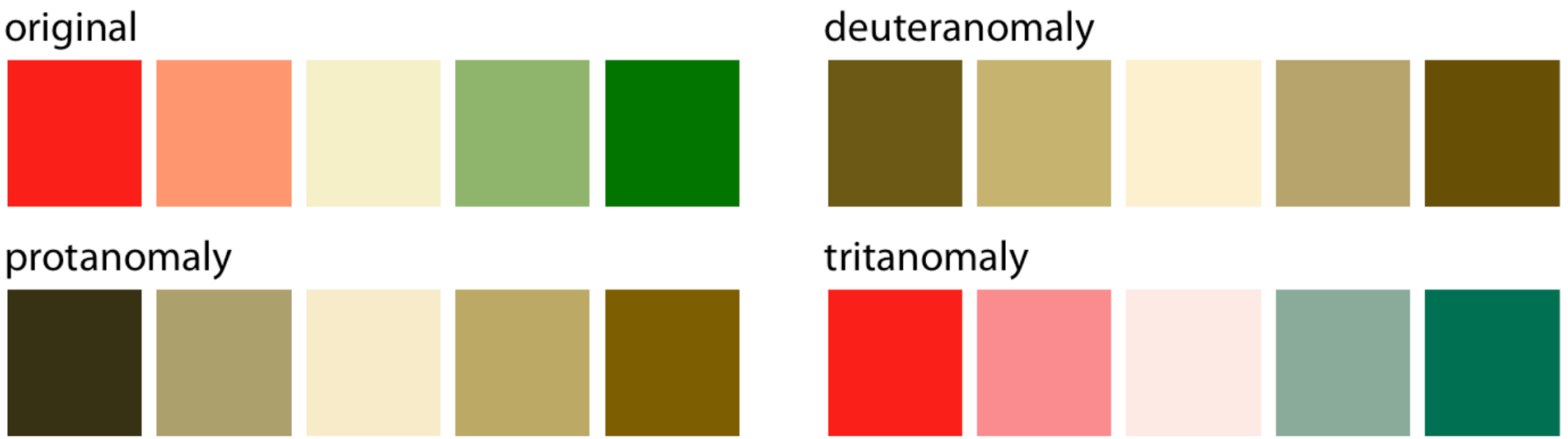
Instructor Note

Using neural networks, Petroff (2024), was able to identify sequence of colors that may improve the accessibility of plots. Several variables were used as predictors/features: color blindness types, color lightness, preferred color sequence, and aesthetic preferences. Three color sequences were identified. All started with the color blue.

Different mediums play a considerable role in the perception of graphs, figures, and tables. Colors may be more saturated on a computer than a piece of paper. Different software may use different varying default colors in visual aides. A data scientist needs to consider their audience.

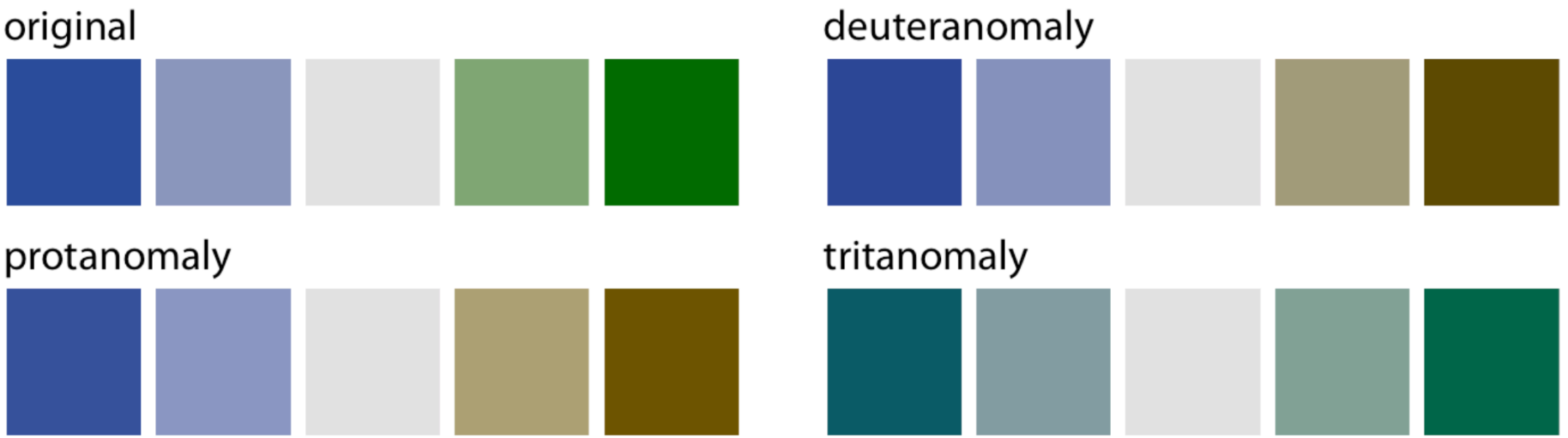
Deuteranomaly, protanomaly, and tritanomaly are three types of color blindness. Deuteranomaly and protanomaly are red-green color blindness conditions that cause difficulty distinguishing between reds and greens, with deuteranomaly making green appear redder and protanomaly making red appear dimmer and more green. Tritanomaly is a less common blue-yellow color blindness.

Red-Green Colorblindness



Red-green color-vision deficiency is the most common

Blue-Yellow Colorblindness



Blue-yellow color-vision deficiency is rarer but does occur; it is more common in older people

Accessible Color Palettes

Accessibility

- Researchers have used machine learning to automate data extraction from graphs and figures and consequentially provide the natural language descriptions to support visually disabled users. (Shahira and Lijiya 2021)
- There are **R** packages that assist visually impaired researchers with developing figures and graphs. (J. R. Godfrey 2018)
  - `colorBlindness` package — view your plots under different color-deficiency simulations
  - `BrailleR` package—an R add-on specifically targeted at blind users—to extract detailed information, such as data ranges, grammatical constructs, etc. to generate speech strings that are embedded into the image annotation. (A. J. R. Godfrey et al., n.d.)
  - `gridSVG` — exports graphics using the grid package to SVG, which can be read by a screen reader for the visually impaired. (Murrell and Potter, n.d.)

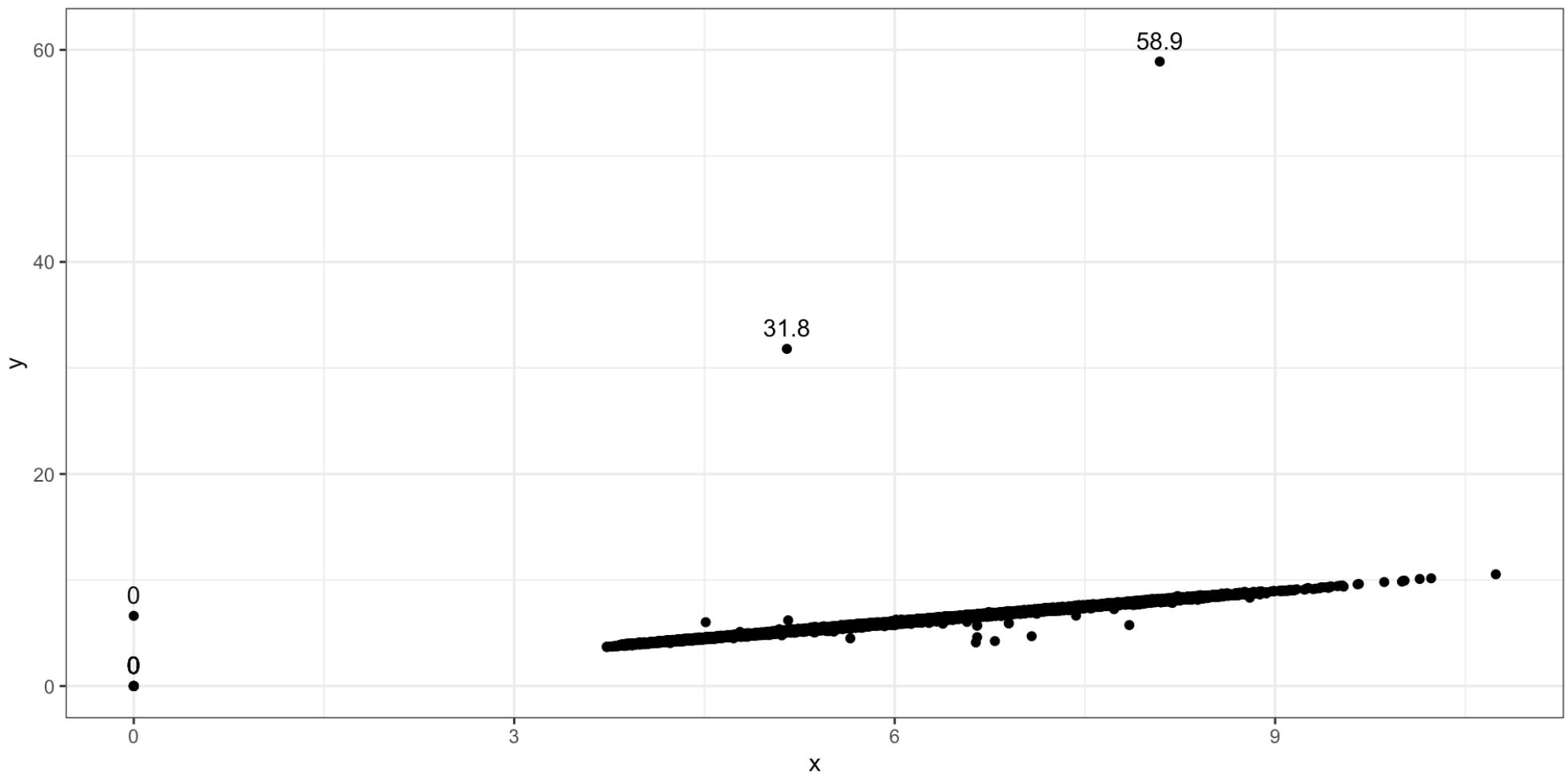
Instructor Note

Can have students simulate colorblindness on visualizations that students previously generated (e.g. using `colorBlindness`) or experiment with simulators online (e.g. https://www.color-blindness.com/coblis-color-blindness-simulator/)

Outliers and anomalies

Diamonds Data

Length (x) and width (y) of diamonds listed by online jewelers.



Instructor Note

The diamond data set contains prices and other attributes of over 50,000 diamonds. Most diamonds are “square-ish” so x (length) and y (width) tend to be very similar.

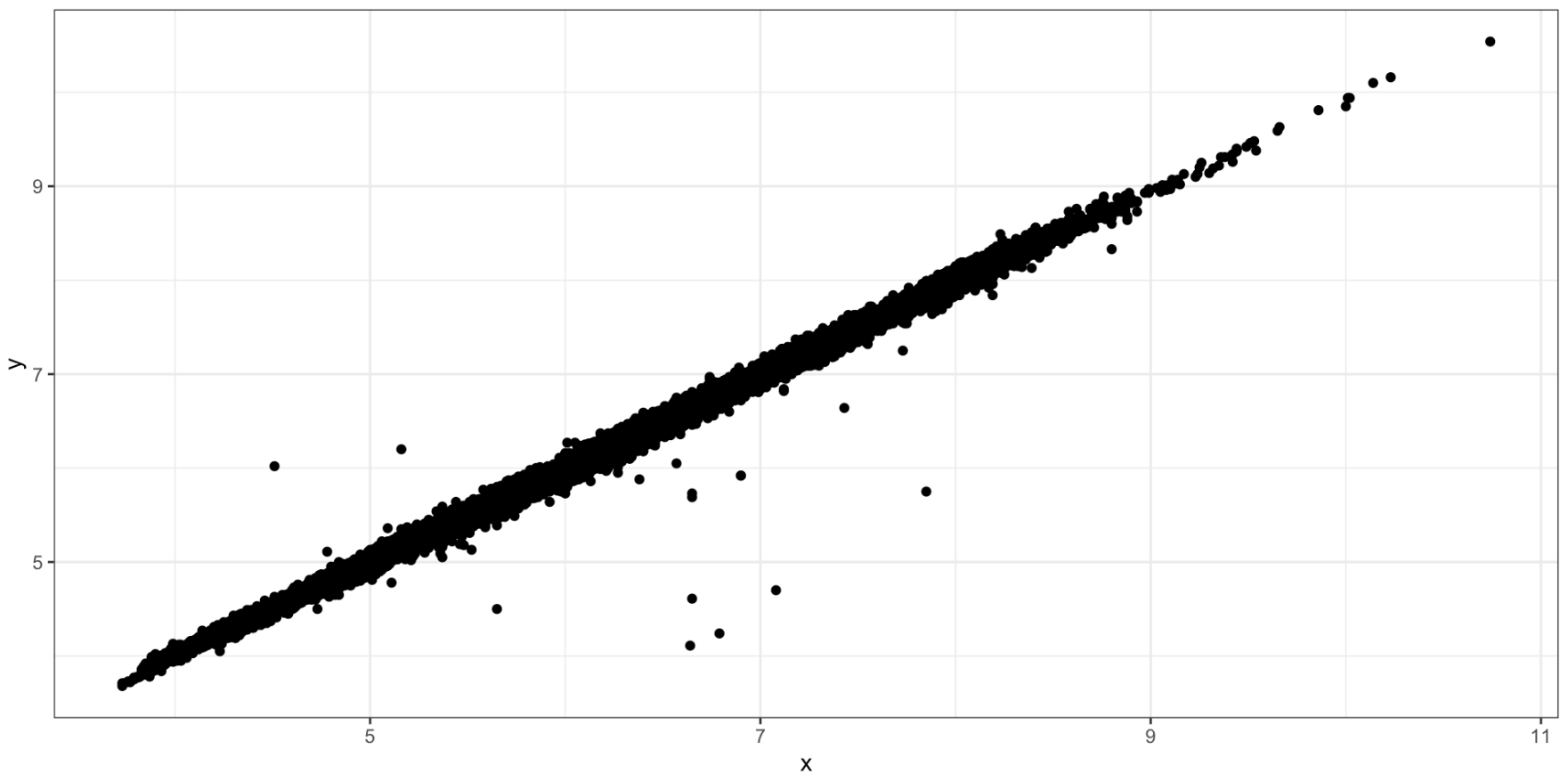
Mistakes could have been introduced by coding for missing values (e.g., 0 or 999) and simple data entry errors

💬 Possible Discussion Questions 💬

- How do those extreme values impact the perceived relationship between length and width of diamonds after being removed?
- Would you drop the outliers and all their additional observed features? When is this appropriate?

Outliers and anomalies

Zooming in and excluding outliers from the plot reveals the fundamental relationship more clearly.



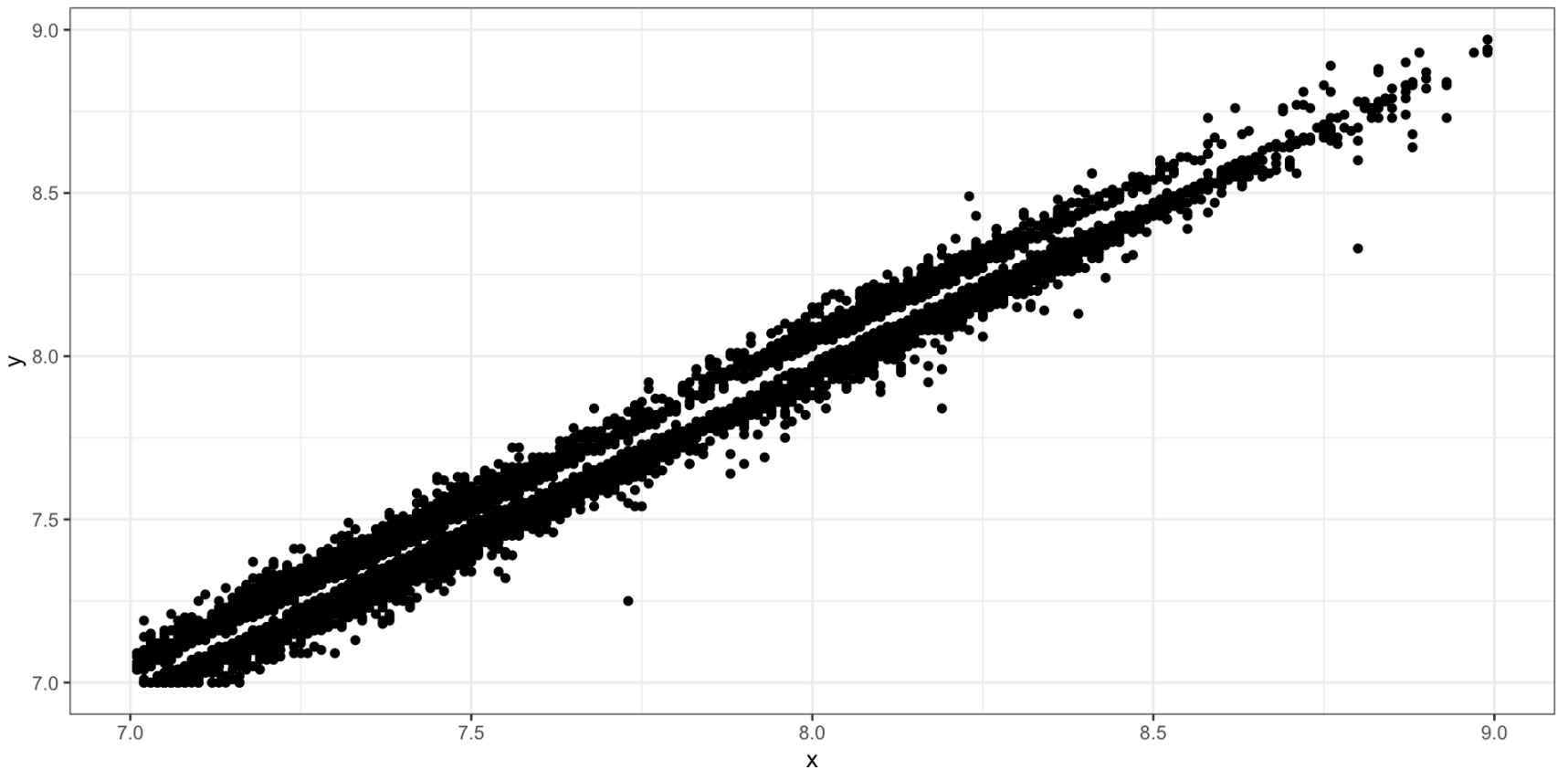
Outliers and Anomalies

- During data cleaning and data exploration, evaluate the data for “unusual values” using visual aides or summary statistics.
- Units with extreme high, low, or otherwise unusual values are classified as outliers.

Outliers

- Best to pre-specify your criteria for excluding outliers prior to analysis.
- Show results with and without outliers removed. Clearly state:
  - How many observations were excluded
  - The method used for detection
  - How results differ between analyses

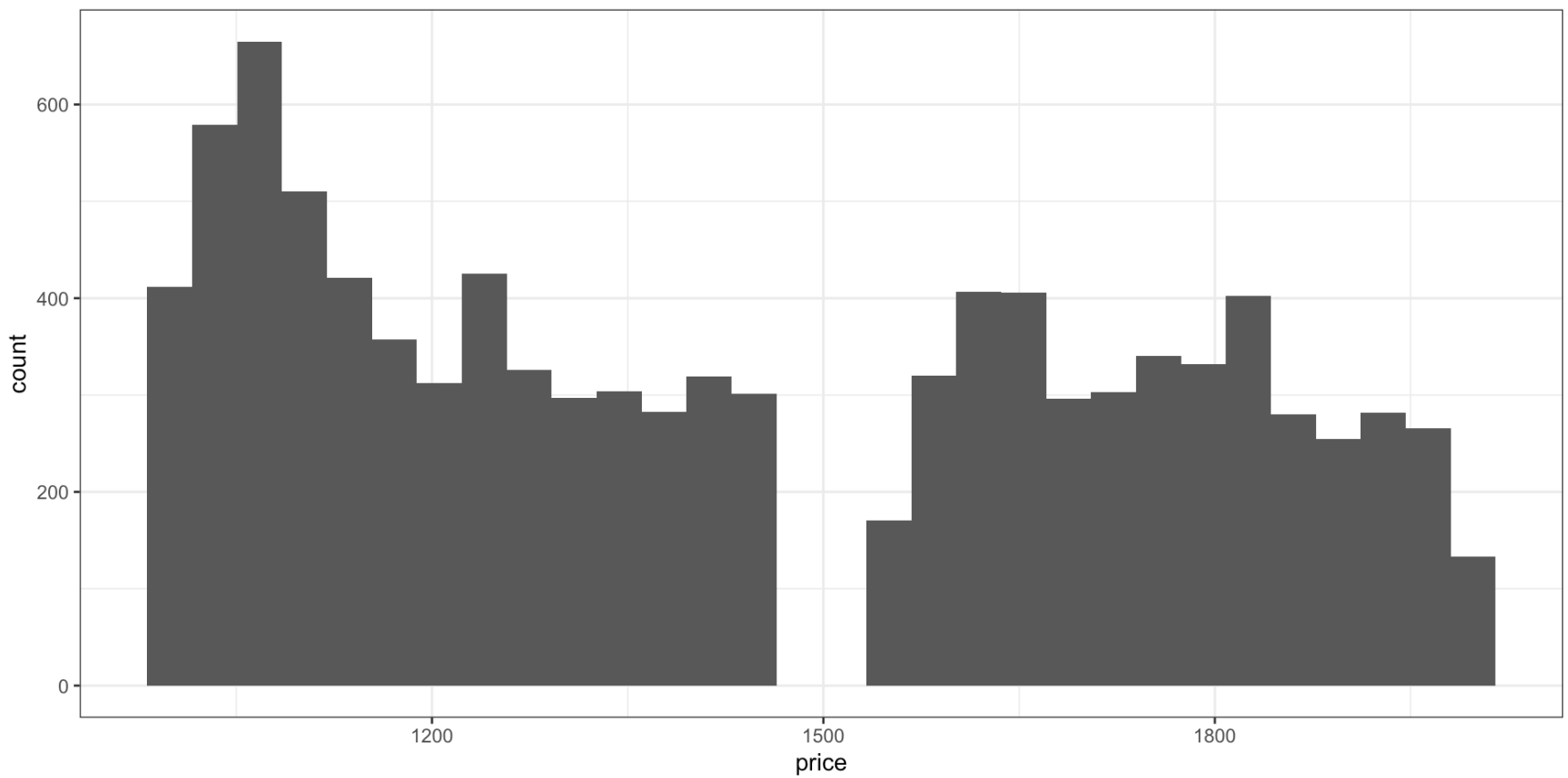
Anomalies



Zooming in further on $7 < x < 9$ and $7 < y < 9$ reveals a band where there are no observations.

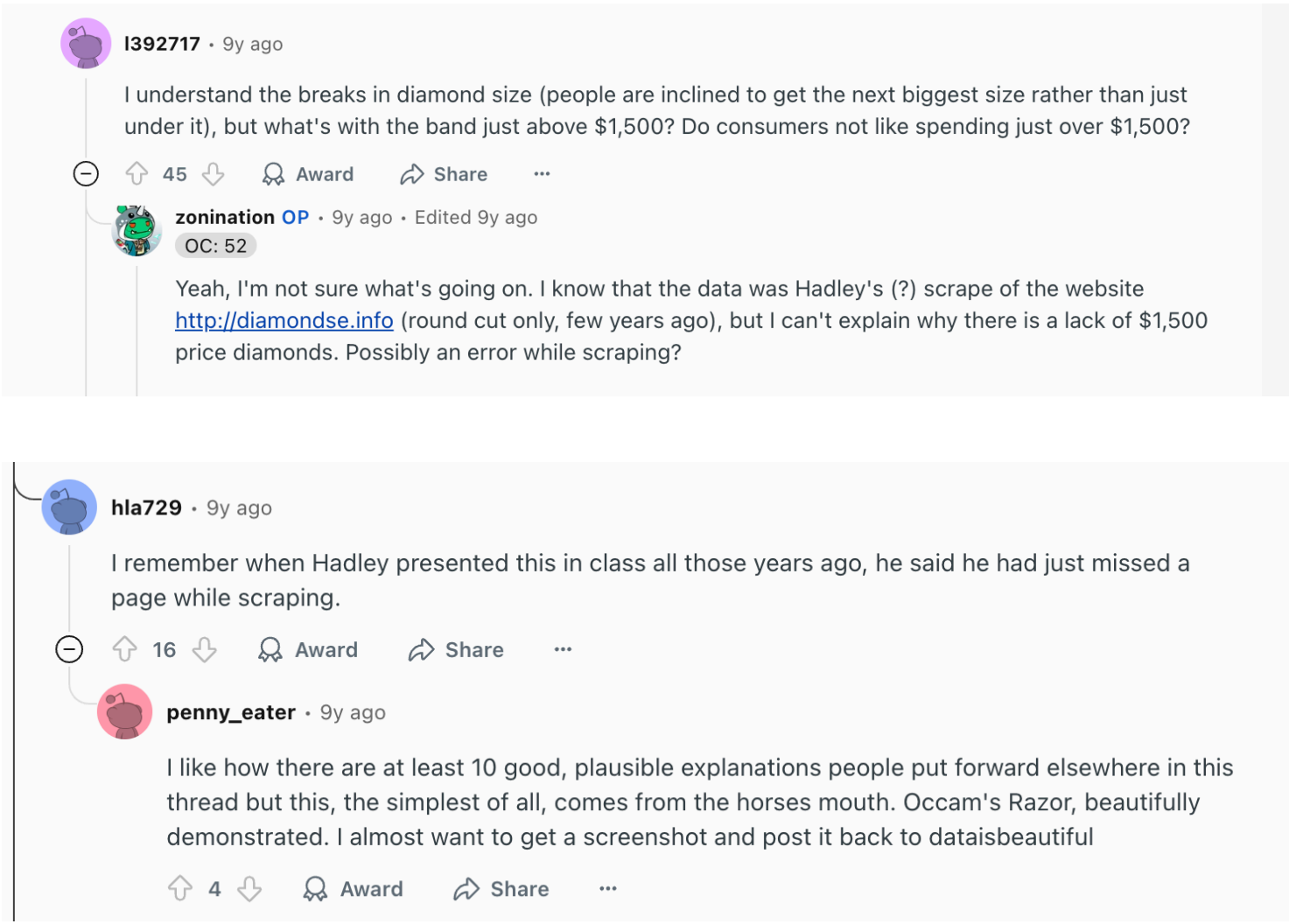
Anomalies

Looking at a different feature, the price of the diamond, shows a lack of diamonds priced at $1500. Why?



Anomalies

Buried in a Reddit thread:



l392717 • 9y ago

I understand the breaks in diamond size (people are inclined to get the next biggest size rather than just under it), but what's with the band just above $1,500? Do consumers not like spending just over $1,500?

45 Award Share ...

zonination OP • 9y ago • Edited 9y ago

OC: 52

Yeah, I'm not sure what's going on. I know that the data was Hadley's (?) scrape of the website http://diamondse.info (round cut only, few years ago), but I can't explain why there is a lack of $1,500 price diamonds. Possibly an error while scraping?

hla729 • 9y ago

I remember when Hadley presented this in class all those years ago, he said he had just missed a page while scraping.

16 Award Share ...

penny_eater • 9y ago

I like how there are at least 10 good, plausible explanations people put forward elsewhere in this thread but this, the simplest of all, comes from the horses mouth. Occam's Razor, beautifully demonstrated. I almost want to get a screenshot and post it back to dataisbeautiful

4 Award Share ...

💬 Possible Discussion Questions 💬

- What role does the *data collector* have in ensuring the quality, representativeness, and completeness of the data?
- What role does the *data analyst* have in identifying any anomalies or data collection errors?
- May consider returning to a discussion on stories/explanations and Occam's Razor. It's easy to come up with compelling stories about the data, but often it's something as simple as a mistake in data collection.

Ethical Principles from the OECD

Instructor Note

The following four principles out of eight privacy principles established by the Organization for Economics Co-operation and Development (OECD) especially relate to data visualizations. The principle is first described followed by suggestions to comply with it while creating visualizations. Consider to ask students if they can think of any other practices that will help them comply with these principles.

Data Quality

- Principle: Be sure that the data are accurate, complete and kept up-to-date.
- Poor data will create poor graphics (Tufte 2001), which may confuse readers and lead to a wrong conclusion.
- Evaluate the data for errors, missing values, and extreme values prior to graphing.

Use Limitation

- Principle: Personal data should not be disclosed, made available or otherwise used for purposes other than those specified in accordance.
- Ensure that graphics do not contain personal participant information or that given information cannot be directly related to participants.

Openness Principle

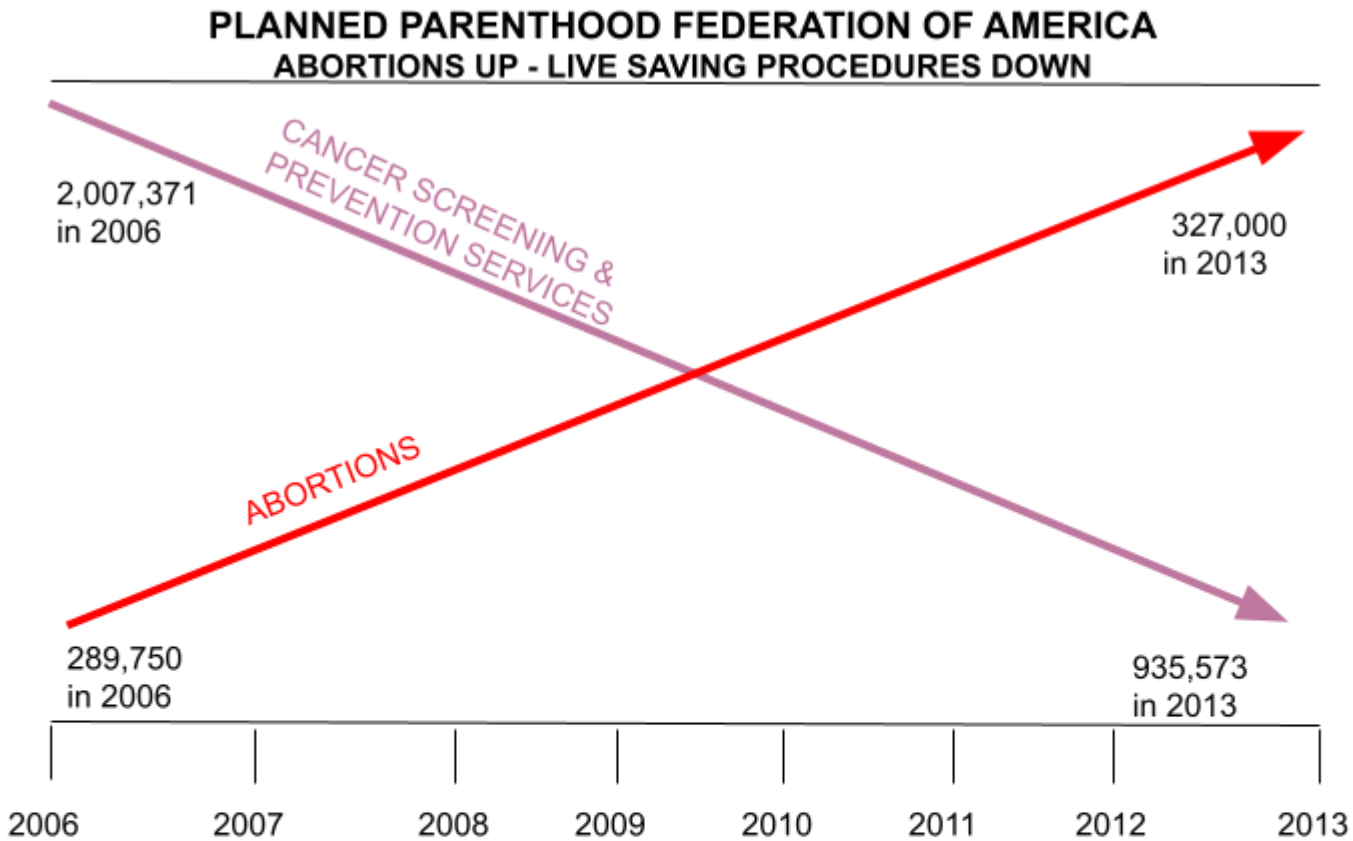
- Principle: There should be a general policy of openness about developments, practices and policies with respect to personal data.
- Remain transparent about the data used in figures.
- Be transparent about how your background may have influenced the graphic.
- Describe the procedures used to create graphics, including the transformation of data and the removing of outliers.
- Do not manipulate graphics that change the patterns of the data.

Accountability principles

- Principle: A data controller should be accountable for complying with measures which give effect to the principles stated above.
- If the data in a figure are wrong ensure that your superior knows and the graphic is corrected.
- Do not let others influence the procedures used to present data or findings.
- Ensure that you use the proper graphic to display your data.

Comparability

How is this figure misleading?



Source: American United for Life

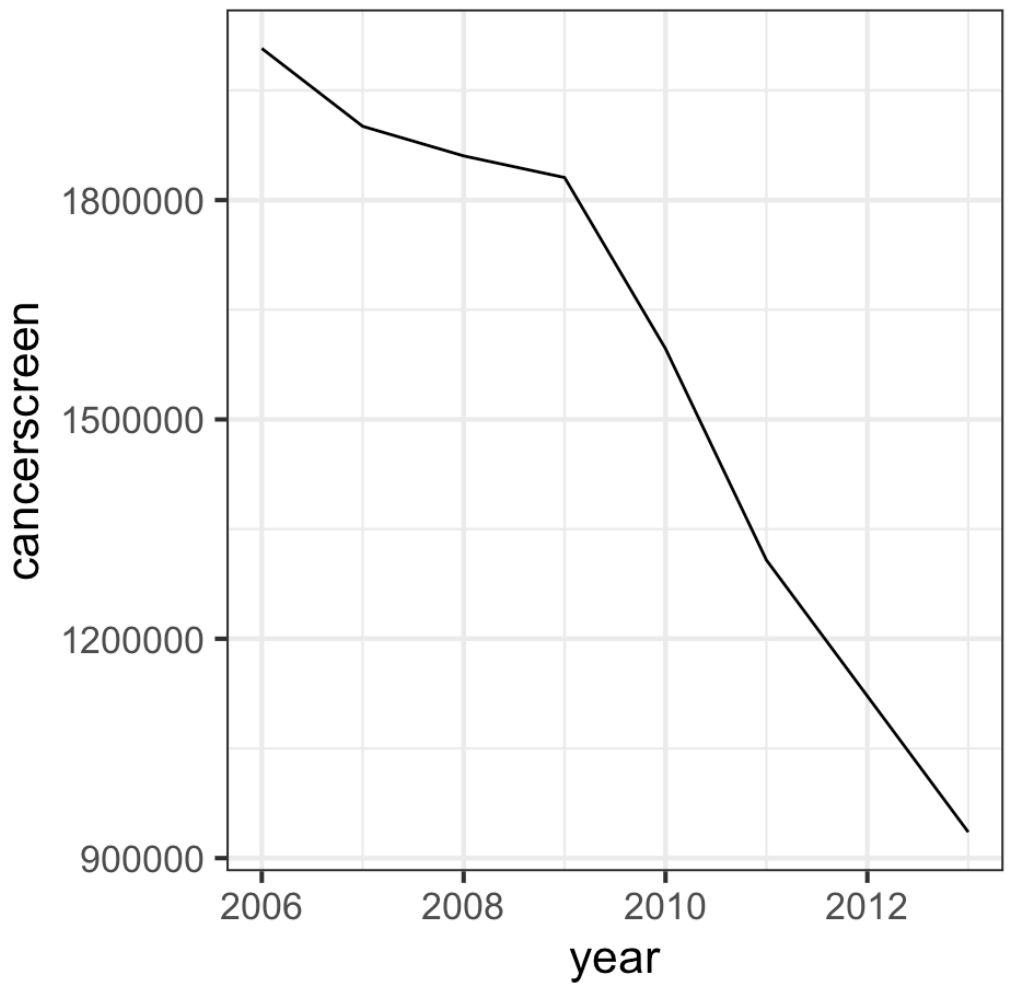
Instructor Note

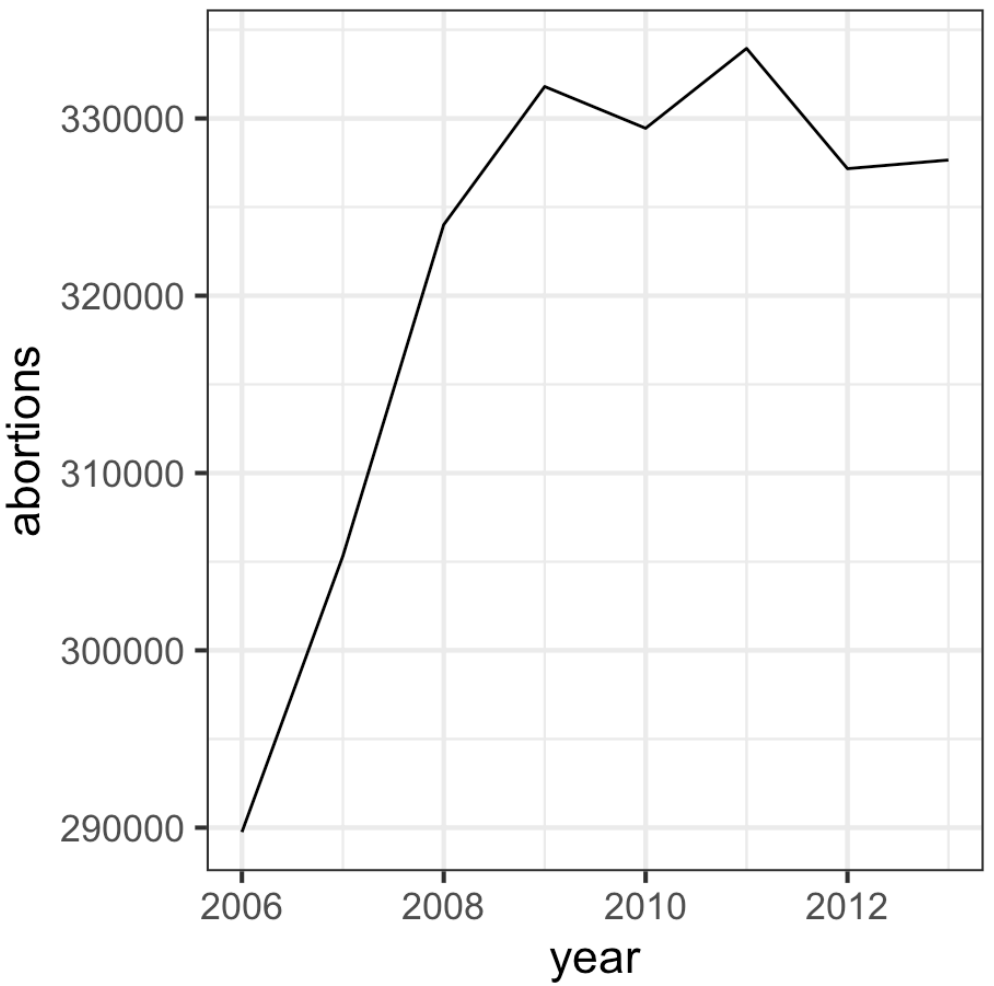
Context about the preceding chart: the infographic was presented in 2015 by Republican congressman Jason Chaffetz, the Chairman of the House Committee on Oversight and Government Reform. This chart was presented to Cecile Richards, the head of Planned Parenthood, during her testimony at public hearing. The chart suggested that Planned Parenthood performs more abortions than cancer screenings and prevention services.
Source: Committee of Oversight and Accountability Democrats

Key points to note:

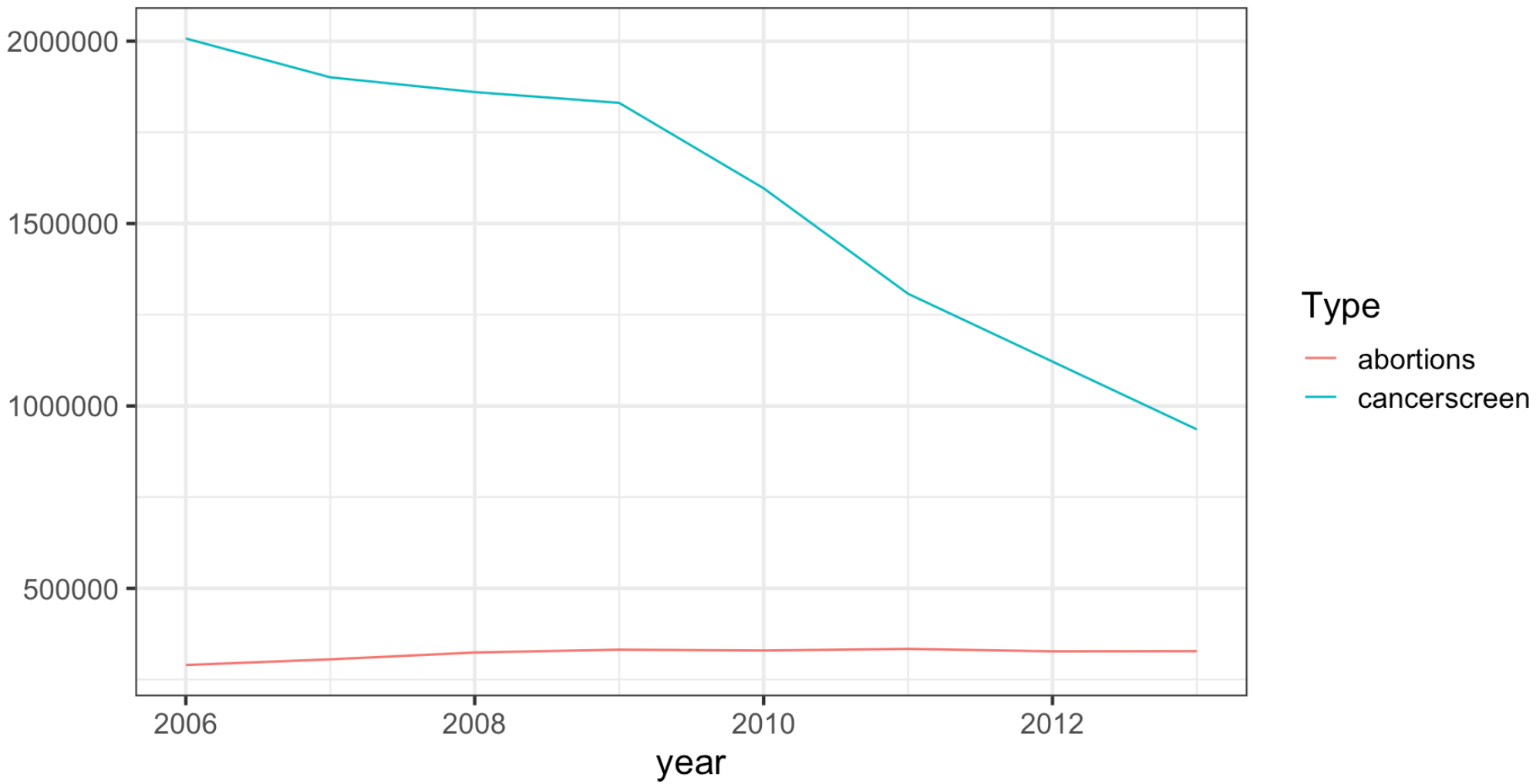
- The total number of cancer screenings and abortions vary dramatically
- Usually best to start the y-axis at 0
- The arrows don't correspond to real data (this is an infographic, not a visualization of real data).
- Can ask students to recreate the visualization, with the data presented as separate lines on the same plot or as two side by side plots and discuss how presenting the data in different formats can be used to tell different stories. Examples follow.

Side by side plot





On the same plot



💬 Possible Discussion Questions 💬

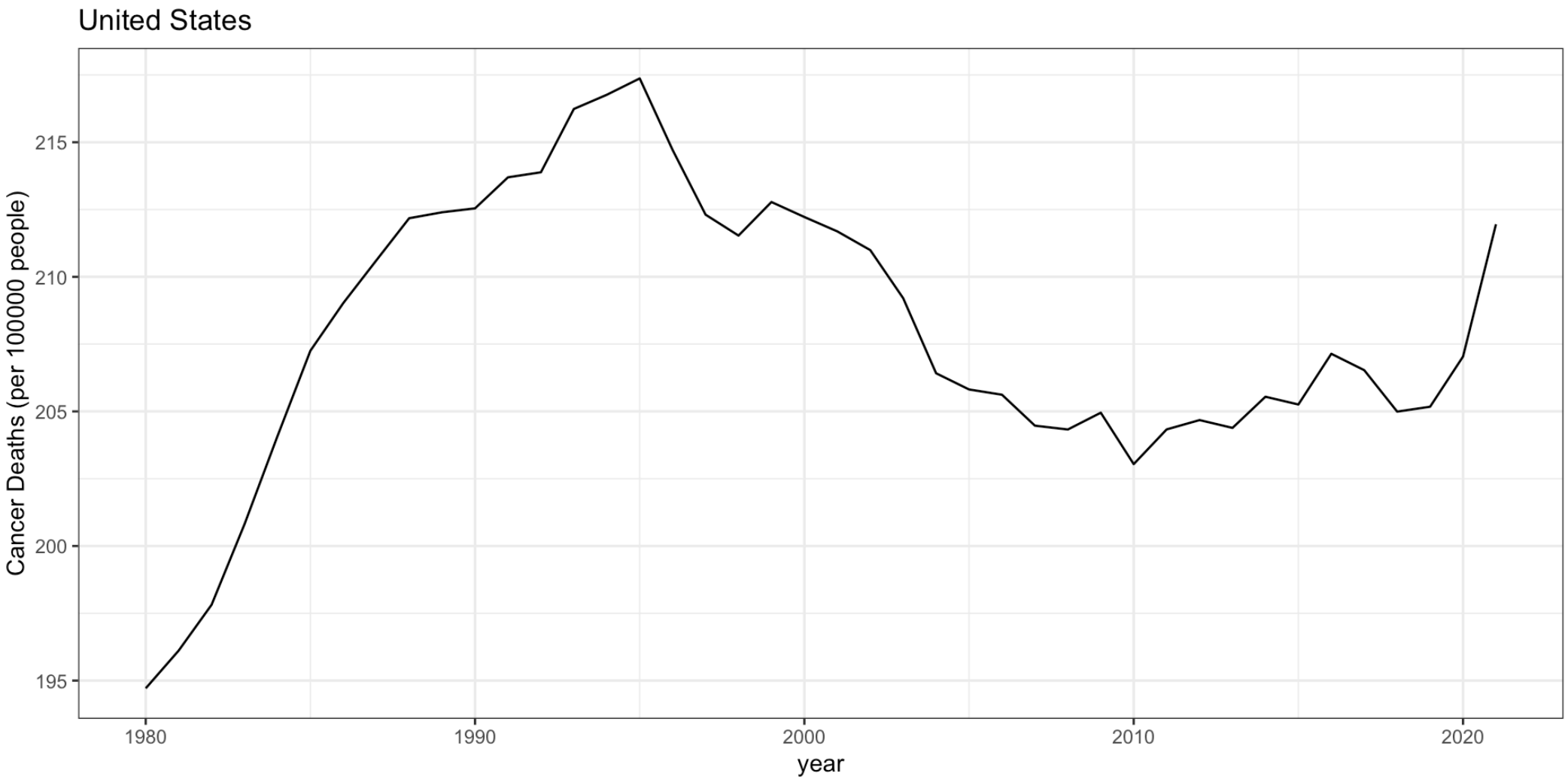
What's the best way to compare data across multiple groups?

- Same plot? Different plot?
- Absolute vs relative? Depends on the goal.
  - Cancer screenings did indeed drop by almost a factor of 2 over the 7 year span, likely due to changes in access to healthcare.
  - Cancer screenings still represent far more of the services provided
- What is the best way to accurately and ethically convey the meaning of this data?

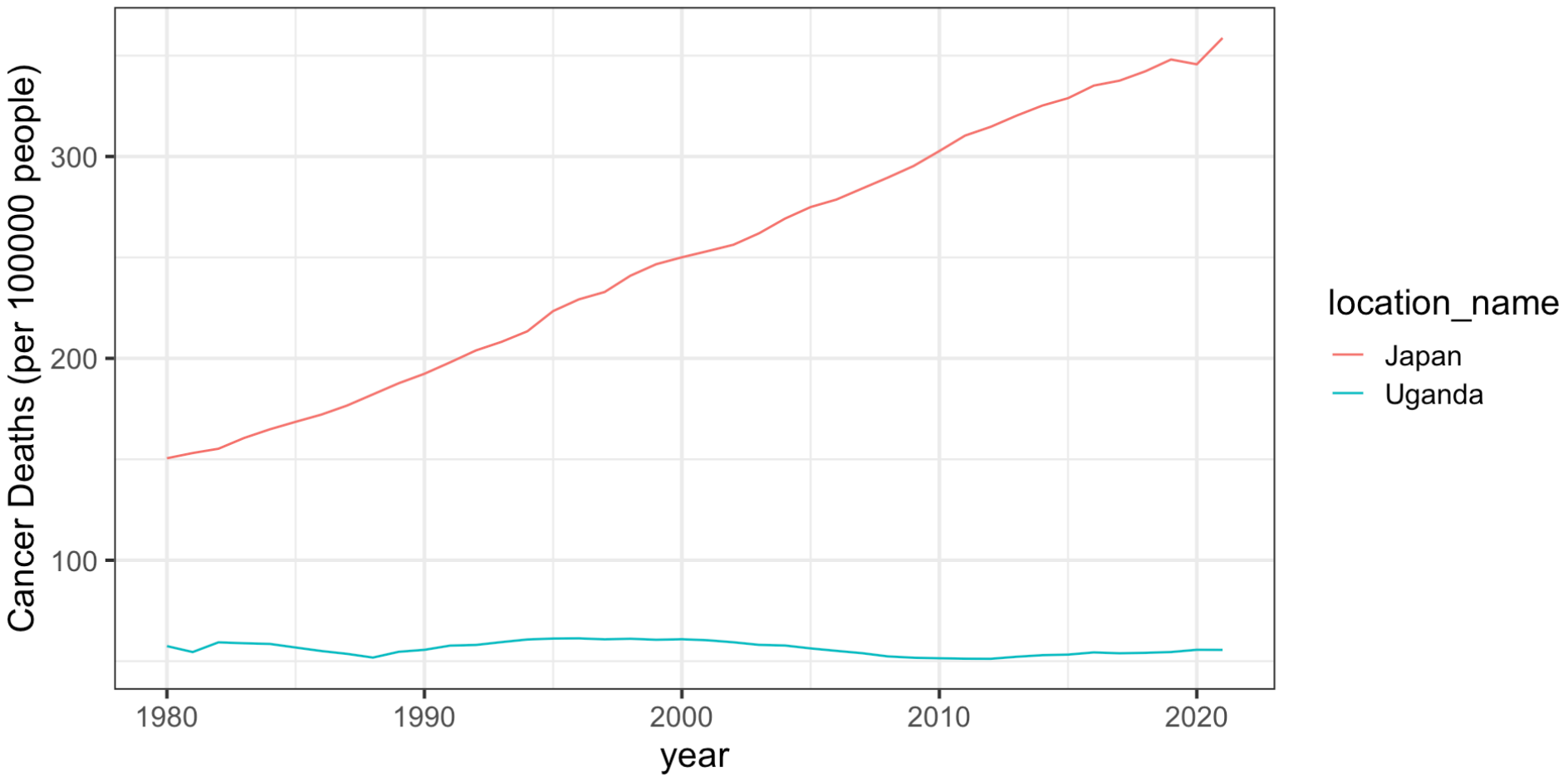
Case study: cancer deaths

- Since 1980, have cancer deaths in the United States been increasing, decreasing or remaining roughly constant?
- Before you look at data, hypothesize about possible trends in cancer deaths and suggest reasons for those trends.

Case study: cancer deaths

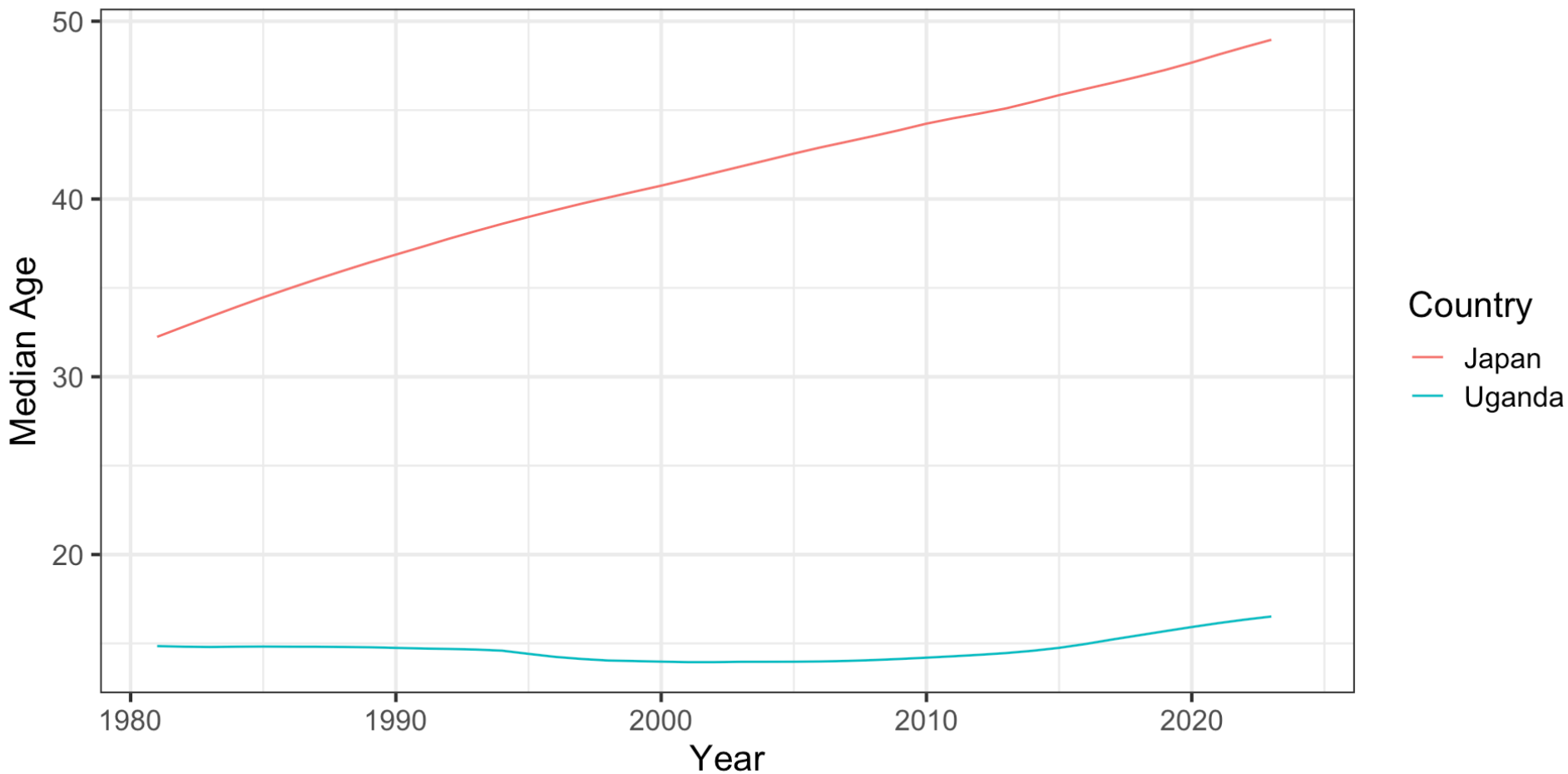


What story does the data tell?



This visualization is valid, but perhaps misleading. Why?

Median age over time

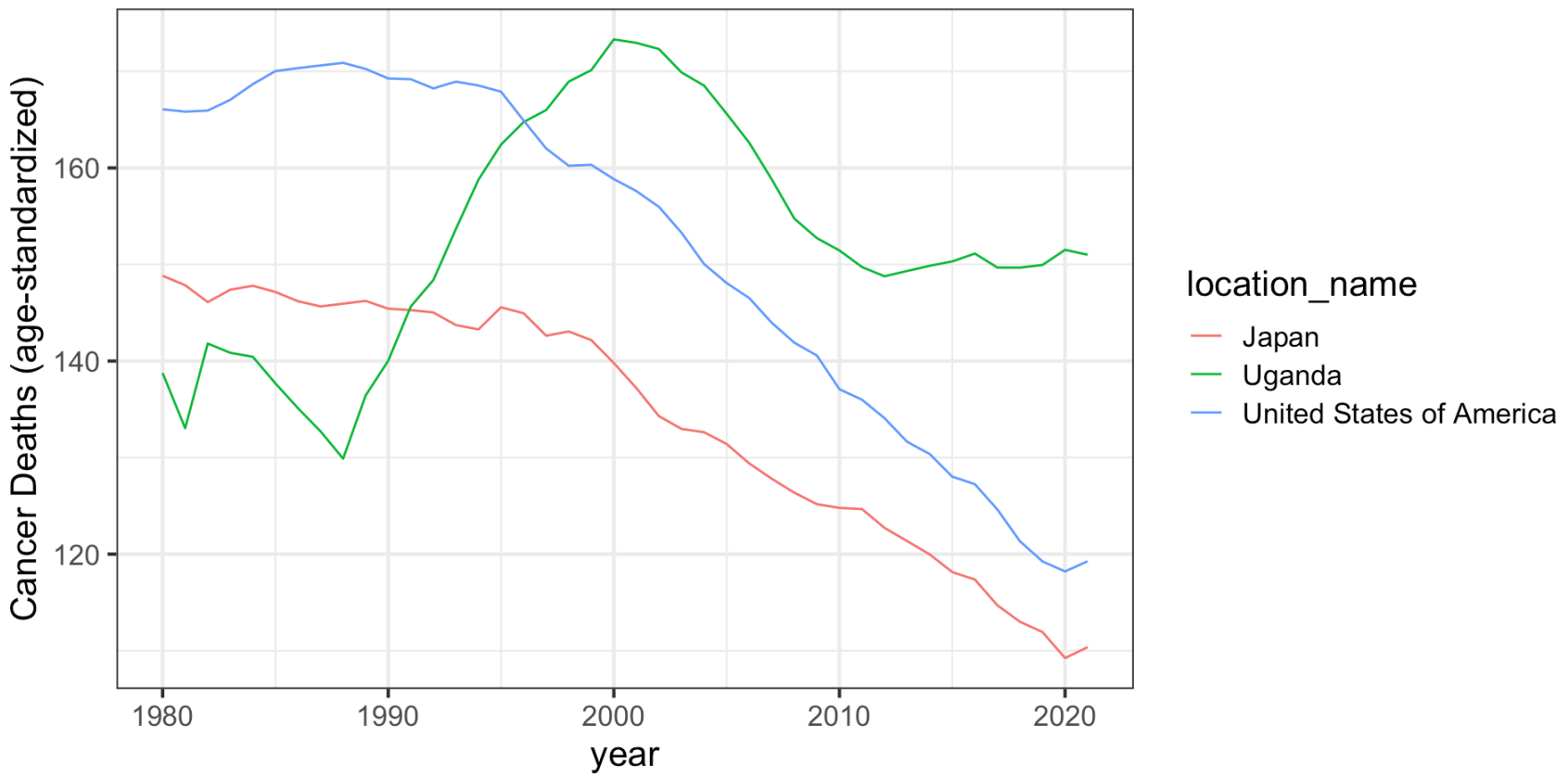


Source: ourworldindata.org

Instructor Note

- Japan is a country that is rapidly aging
- Uganda is a relatively young country with a lower life expectancy
- Age is one of the most important predictors of cancer prevalence. Fewer people live to old age to get cancer in Uganda.

Age-Standardized Cancer Deaths



See https://ourworldindata.org/age-standardization for more discussion.

Comparability

- Always consider relevant reference or comparison groups
- Ask: is **standardization** needed for fair comparison? Are there lurking variables?
- Highlights that expertise is often needed to generate the most appropriate visualization
- Visualizations can be misleading without the relevant context

💬 Possible Discussion Questions 💬

Some useful discussion questions about responsibility and power dynamics:

- How do power dynamics change what stories get told with data and how that data is presented?
- Is the creator of a visualization responsible for how that visualization is interpreted and used?
- What makes somebody qualified to visualize data?

References

Godfrey, A. Jonathan R., Debra Warren, James Thompson, Paul Murrell, Timothy Bilton, and Volker Sorge. n.d. “BrailleR: Improved Access for Blind Users.” https://doi.org/10.32614/CRAN.package.BrailleR.

Godfrey, Jonathan R. 2018. “An Accessible Interaction Model for Data Visualisation in Statistics.” *ICCHP*.

Hansen, J., R. Ruedy, M. Sato, and K. Lo. 2010. “GLOBAL SURFACE TEMPERATURE CHANGE.” *Reviews of Geophysics* 48 (4): RG4004. https://doi.org/10.1029/2010RG000345.

Johnson, Donald D. 2024. “Color Adaptation for Color Deficient Learners.”

Murrell, Paul, and Simon Potter. n.d. “gridSVG: Export ’Grid’ Graphics as SVG.” https://doi.org/10.32614/CRAN.package.gridSVG.

Petroff, Matthew A. 2024. “Accessible Color Sequences for Data Visualization.” http://arxiv.org/abs/2107.02270.

Schroeder, S. Andrew. 2022. “An Ethical Framework for Presenting Scientific Results to Policy-Makers.” *Kennedy Institute of Ethics Journal* 32 (1): 33–67. https://doi.org/10.1353/ken.2022.0002.

Shahira, K. C., and A. Lijiya. 2021. “Towards Assisting the Visually Impaired: A Review on Techniques for Decoding the Visual Data From Chart Images.” *IEEE Access* 9: 52926–43. https://doi.org/10.1109/ACCESS.2021.3069205.

Tufte, Edward R. 2001. *The Visual Display of Quantitative Information*. 2nd ed. Cheshire, Connecticut: Graphic Press.