Lecture 11 - Principles of figure design

Outline

- The principle of proportional ink
- Handling overlapping points
- Common pitfalls of color use
- Redundant coding
- Multi-panel figures
- Titles, captions, and tables
- Balance the data and the context
- Use larger axis labels
- Avoid line drawings
- Don’t go 3D

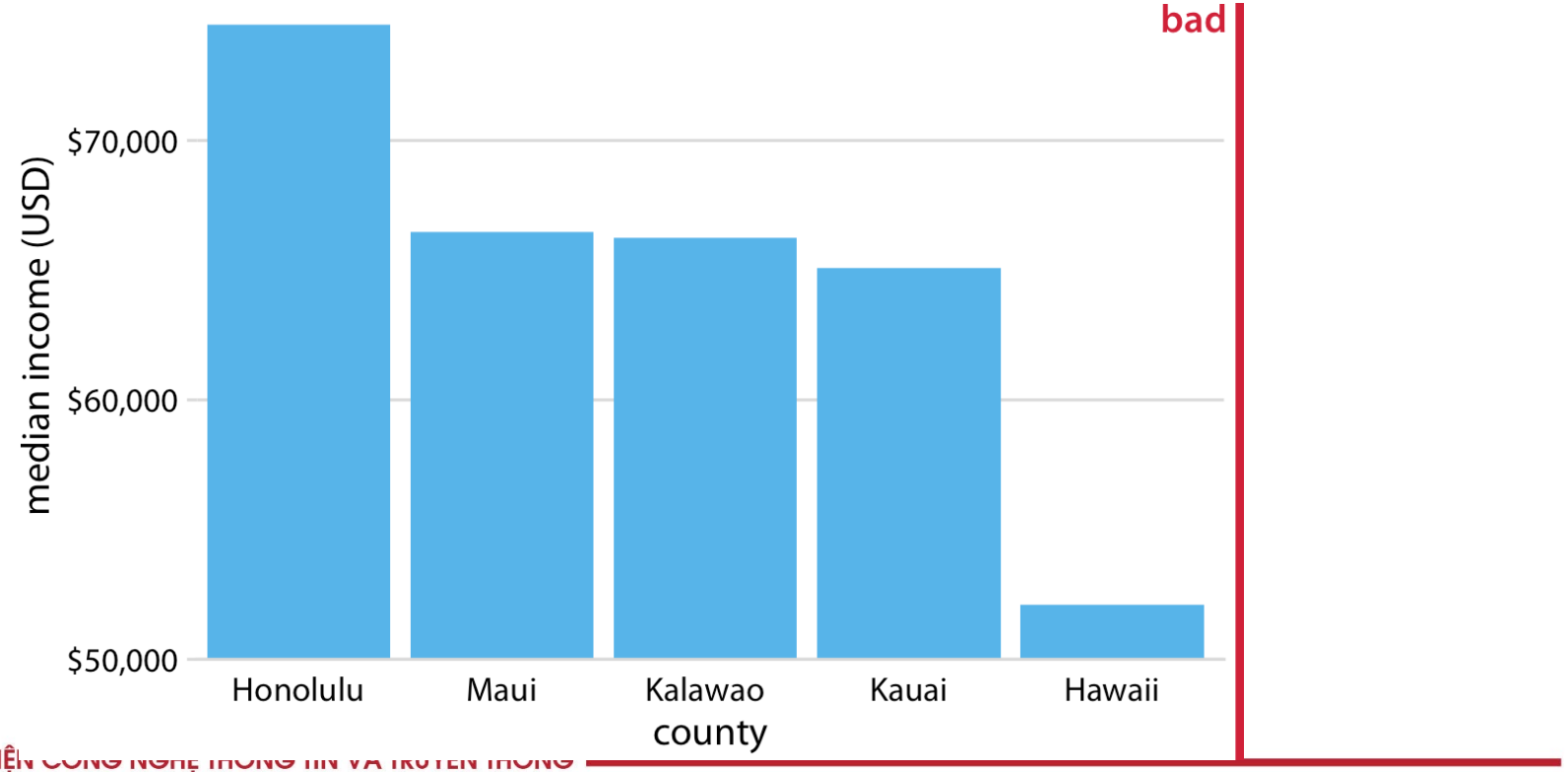
The principle of proportional ink

The principle of proportional ink

- When a shaded region is used to represent a numerical value, the area of that shaded region should be directly proportional to the corresponding value [Bergstrom and West 2016].
  - Make sure that there is no inconsistency

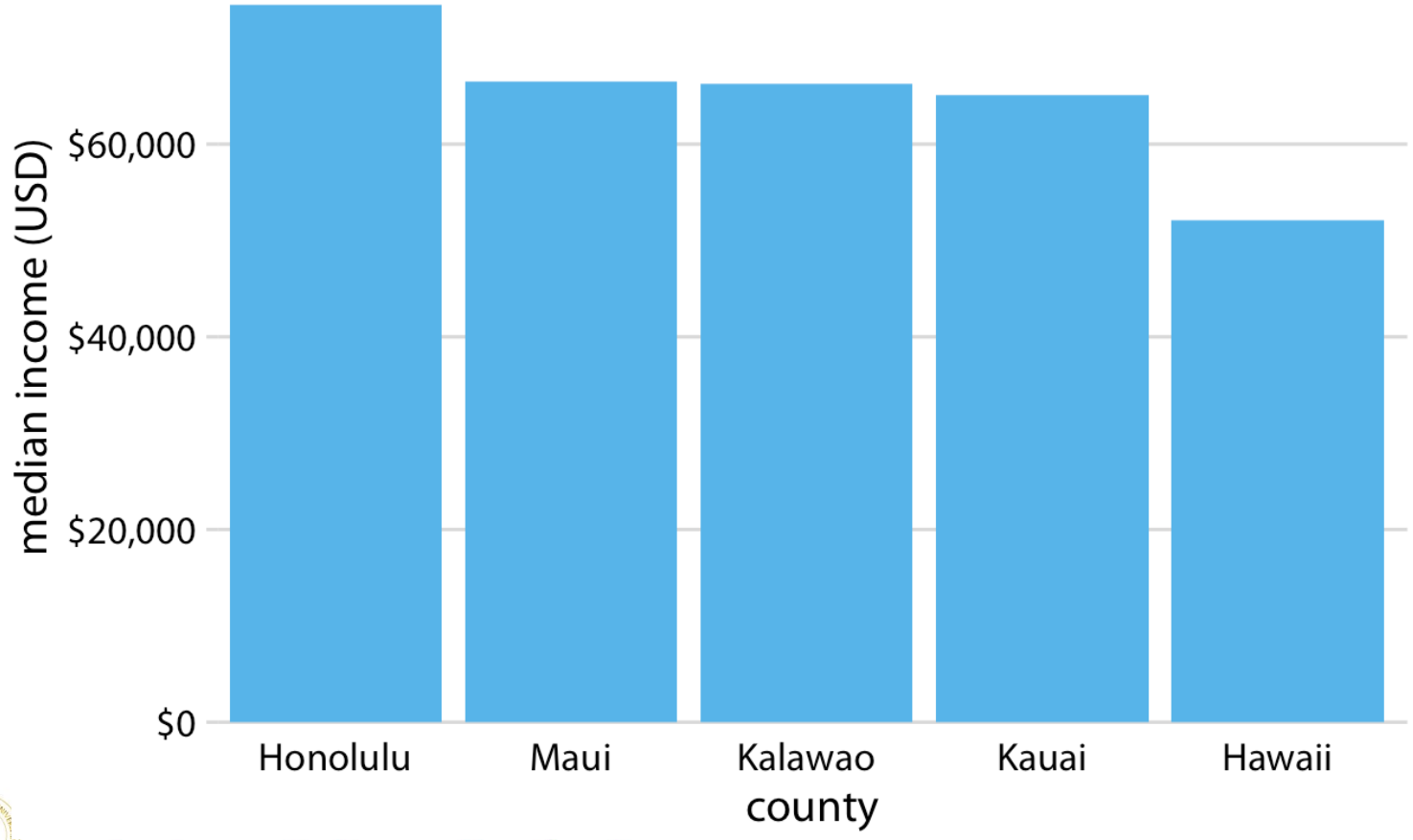
Visualizations along linear axes

- Median income in the five counties of the state of Hawaii
- The y-axis scale starts at $50,000 instead of $0
- The bar heights are not proportional to the values shown
- The income differential between the county of Hawaii and the other four counties appears much bigger than it actually is



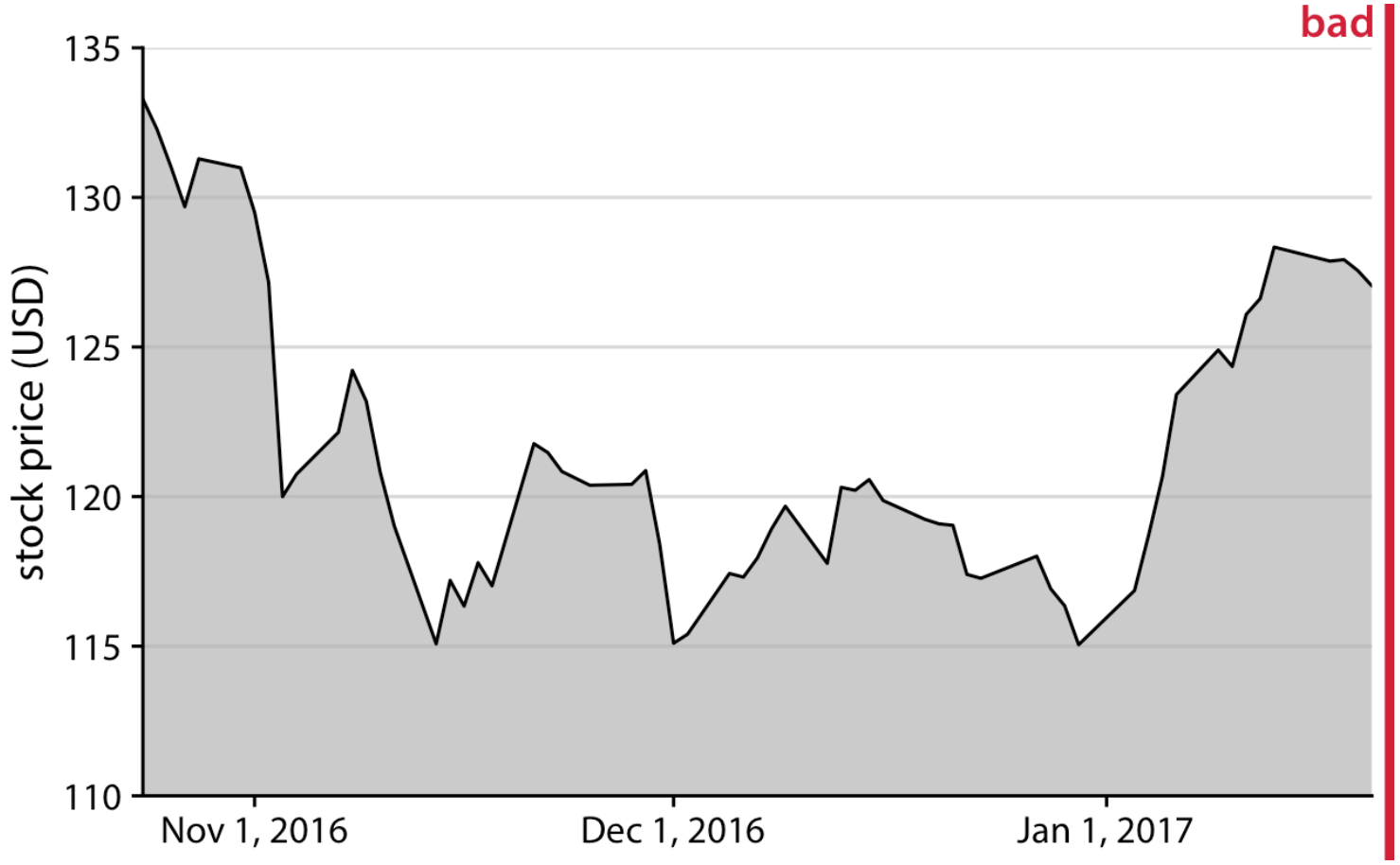
Example: Median income in the five counties of the state of Hawaii

- The y-axis scale starts at $0 and therefore the relative magnitudes of the median incomes in the five counties are accurately shown.
- Bars on a linear scale should always start at 0.



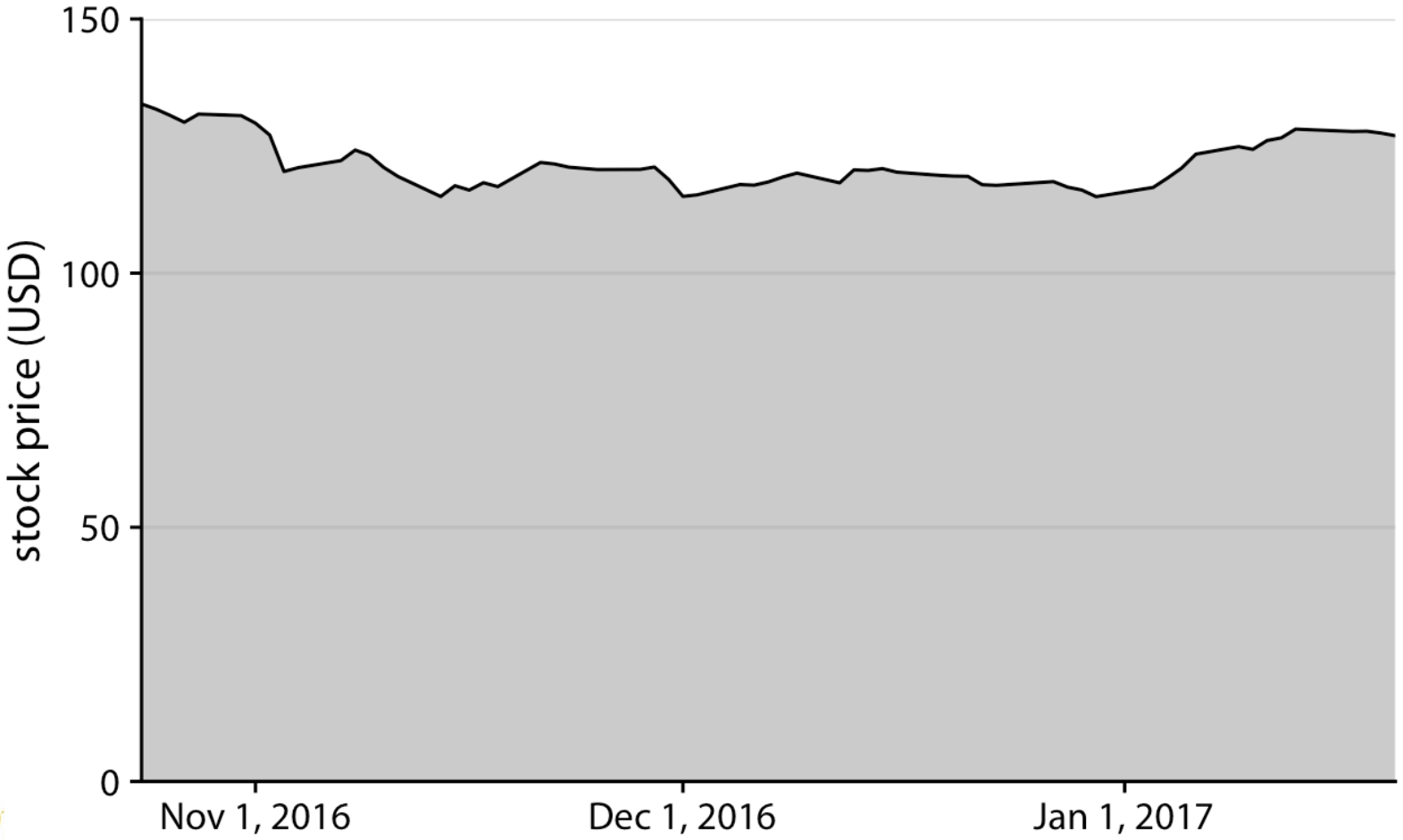
Example: Stock price of Facebook (FB) from Oct. 22, 2016, to Jan. 21, 2017

- The FB stock price collapsed around Nov. 1, 2016?



Example: Stock price of Facebook (FB) from Oct. 22, 2016, to Jan. 21, 2017

- More accurately relays the magnitude of the FB price drop around Nov. 1, 2016

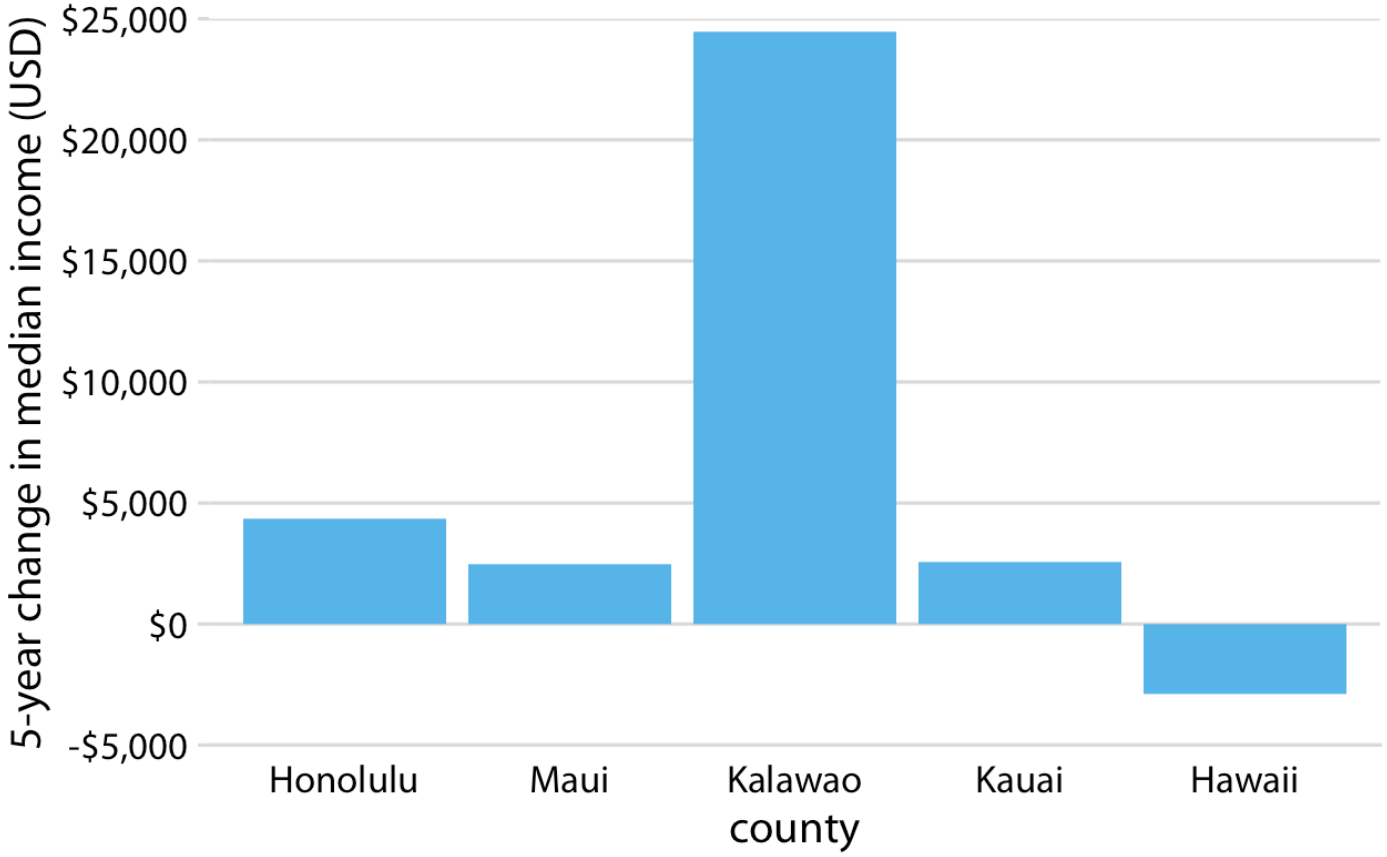


Question

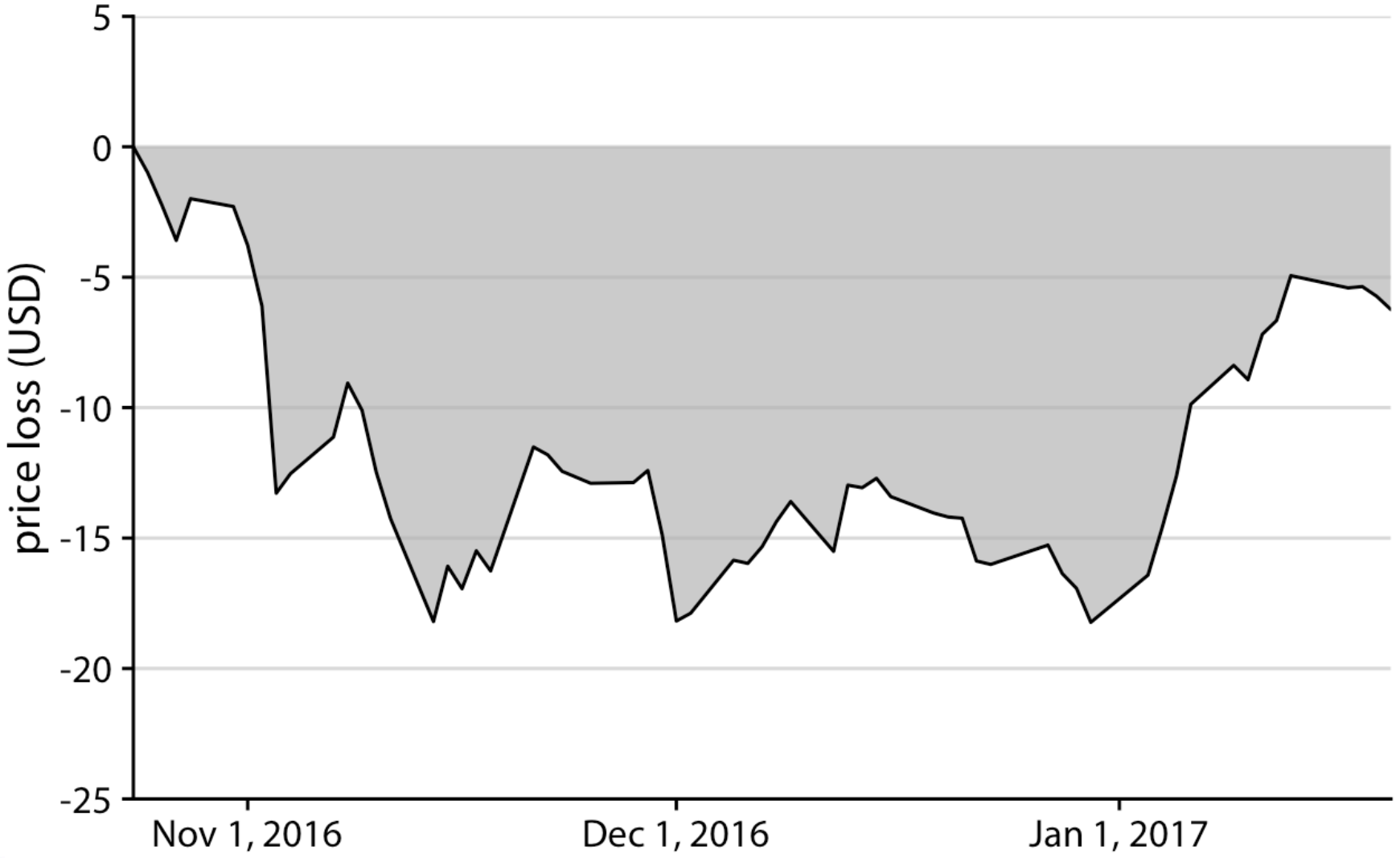
- Bars and shaded areas are not useful to represent small changes over time or differences between conditions, since we always must draw the whole bar or area starting from 0?

Example: Change in median income in Hawaiian counties from 2010 to 2015

- It is perfectly valid to use bars or shaded areas to show differences between conditions, if we make it explicit which differences we are showing.

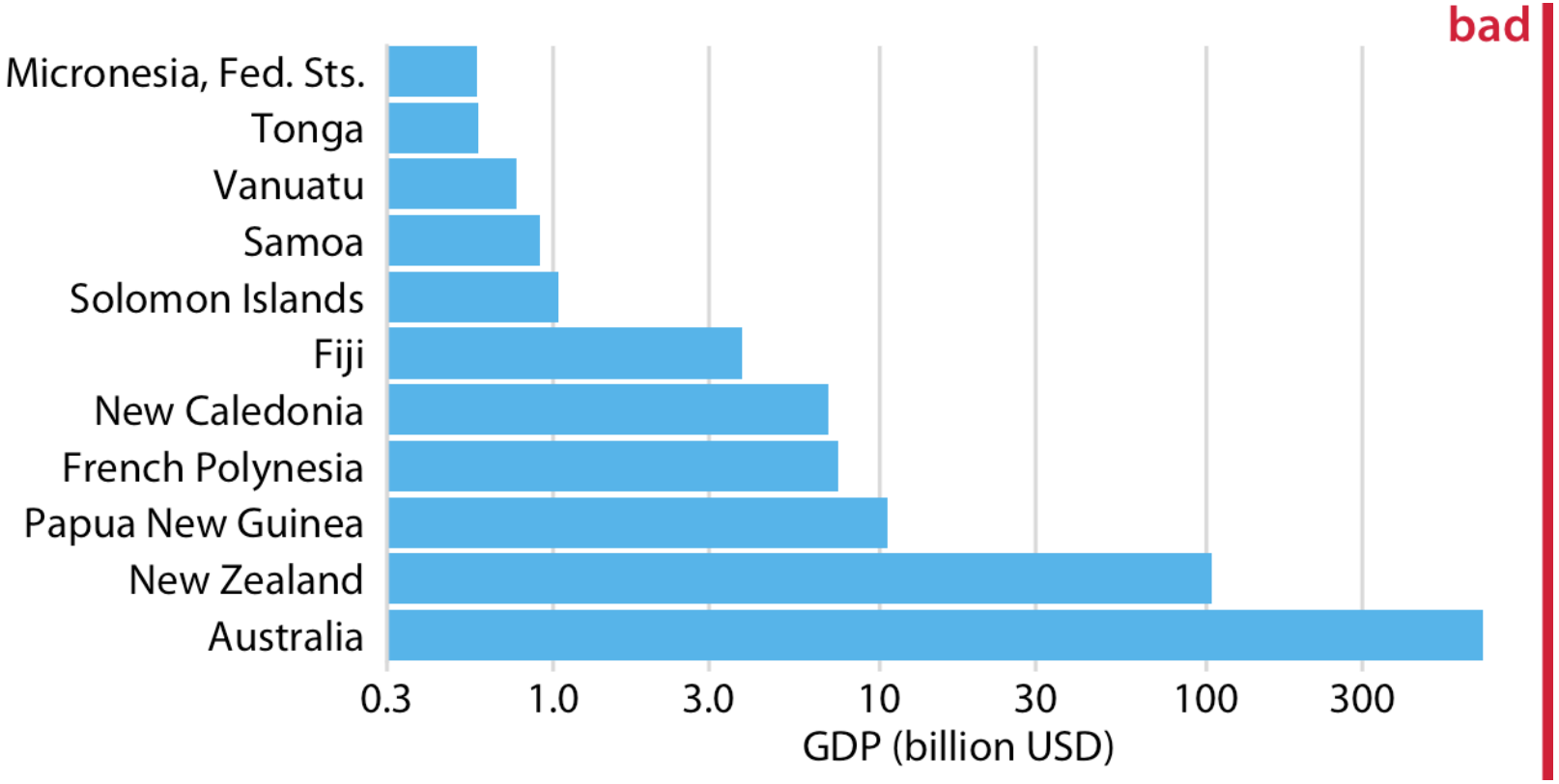


Example: Decline in Facebook (FB) stock price relative to the price of Oct. 22, 2016



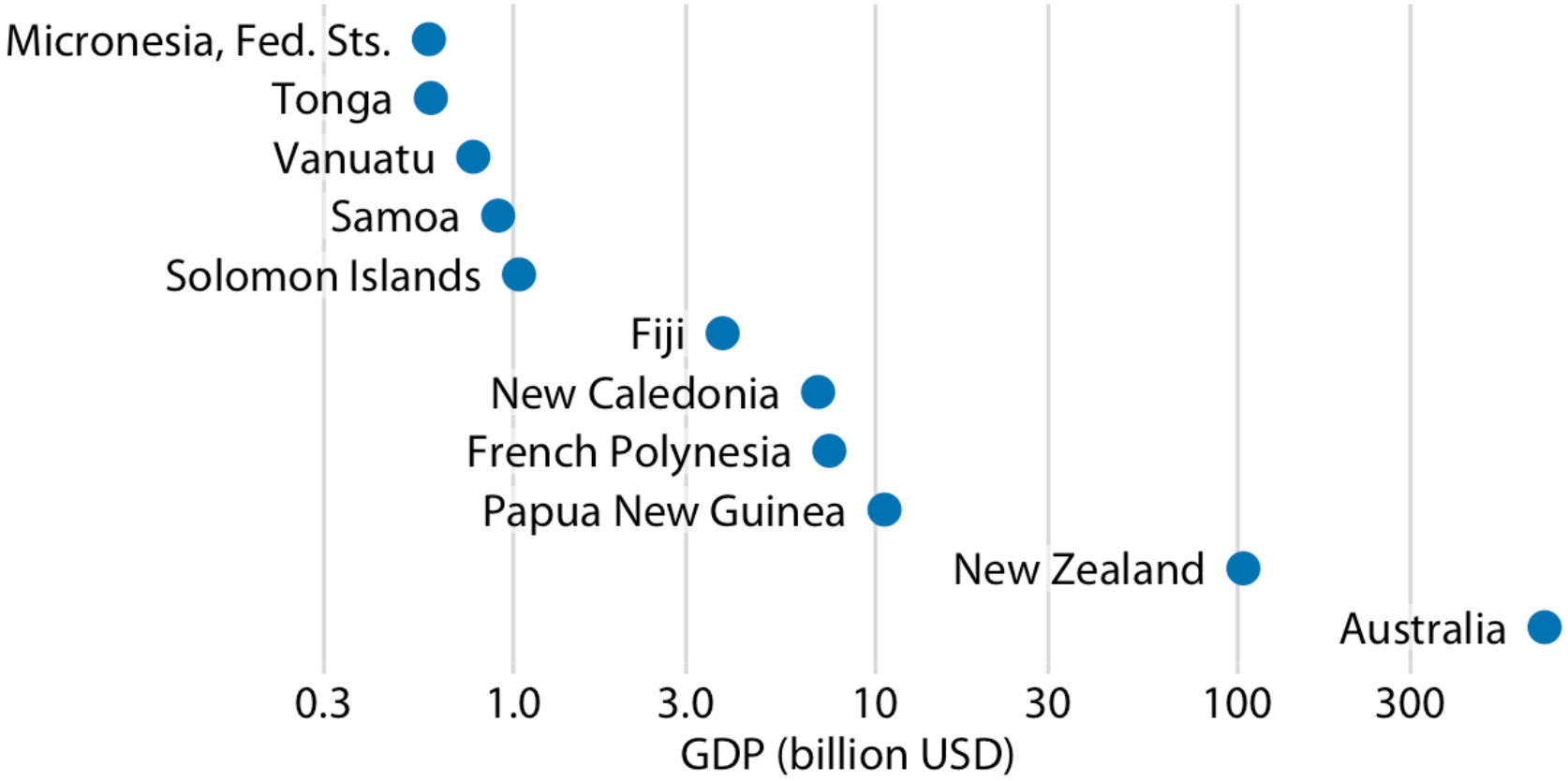
Visualizations along logarithmic axes

- For a logarithmic scale, data values are not linearly spaced along the axis



Example: GDP in 2007 of countries in Oceania

- Simply place a dot at the appropriate location along the scale for each country's GDP and avoid the issue of bar lengths altogether.



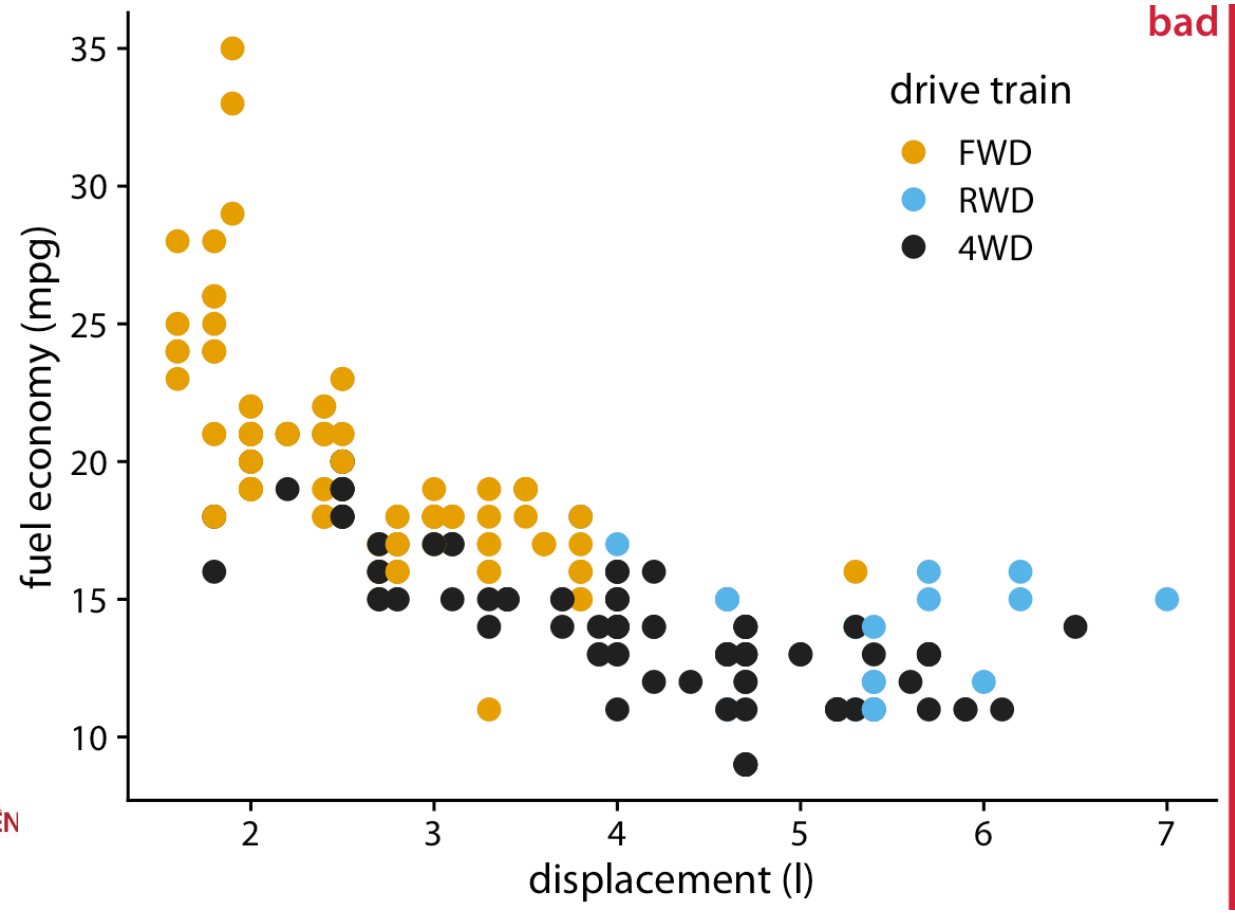
Handling overlapping points

Scenarios

- In the visualization of large or very large datasets, simple x–y scatterplots do not work very well because many points lie on top of each other and partially or fully overlap.
- In small datasets if data values were recorded with low precision or rounded, multiple observations have the same numeric values.

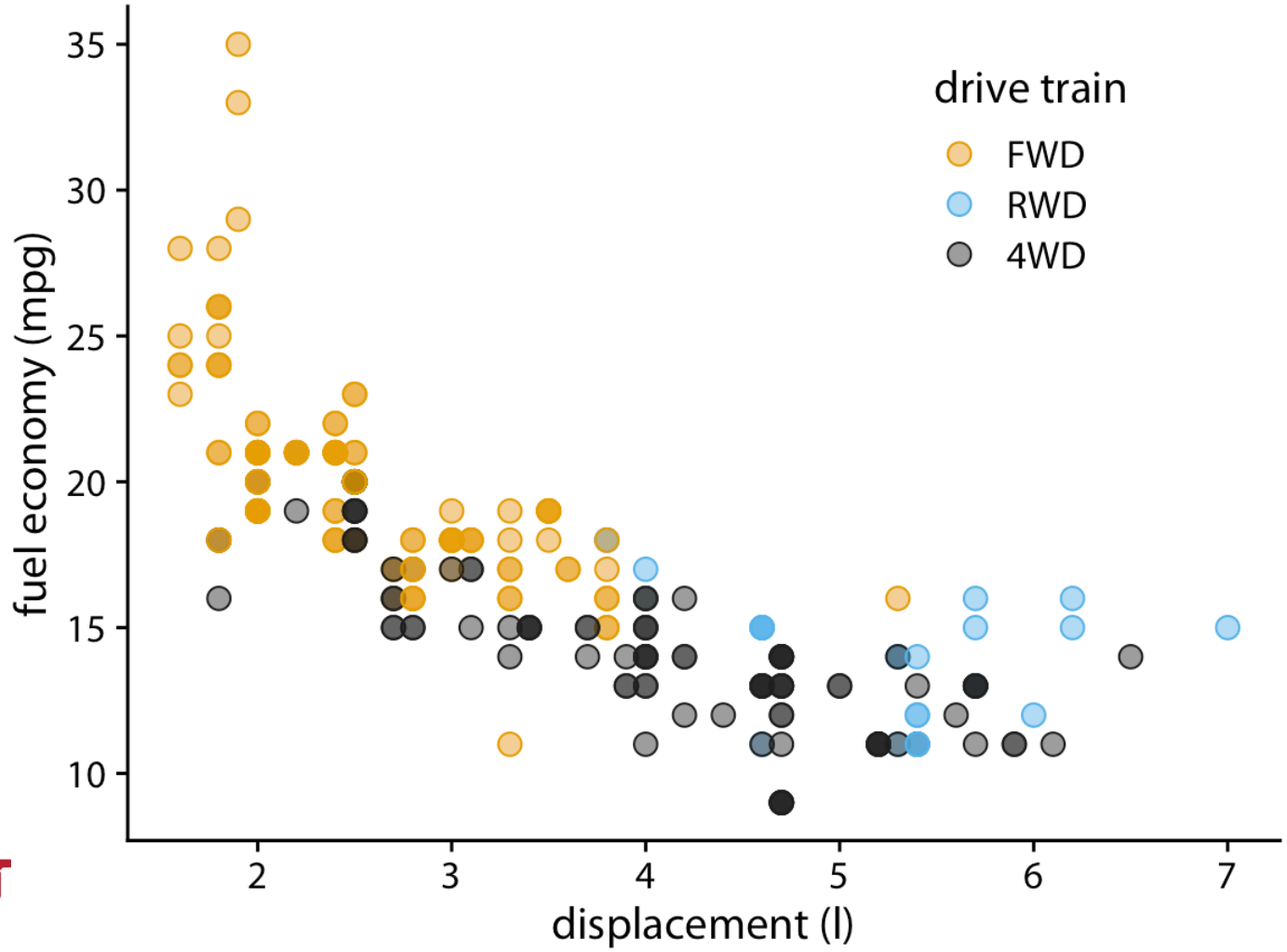
Partial transparency and jittering

- City fuel economy versus engine displacement, for popular cars released between 1999 and 2008
  - Each point represents one car.
  - Many car models have identical values.



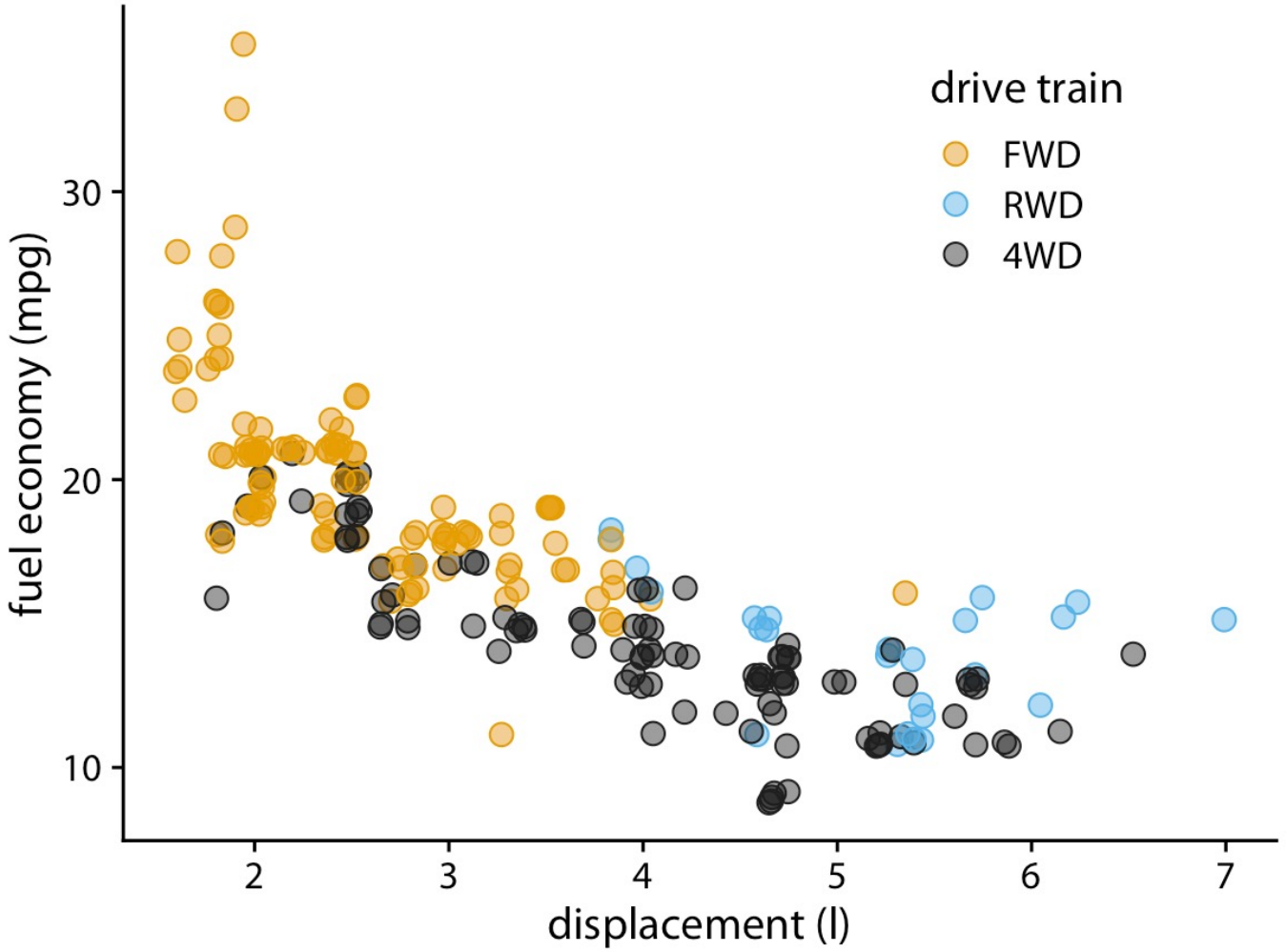
Example: City fuel economy versus engine displacement

- Points have been made partially transparent
- Points that lie on top of other points can now be identified by their darker shade



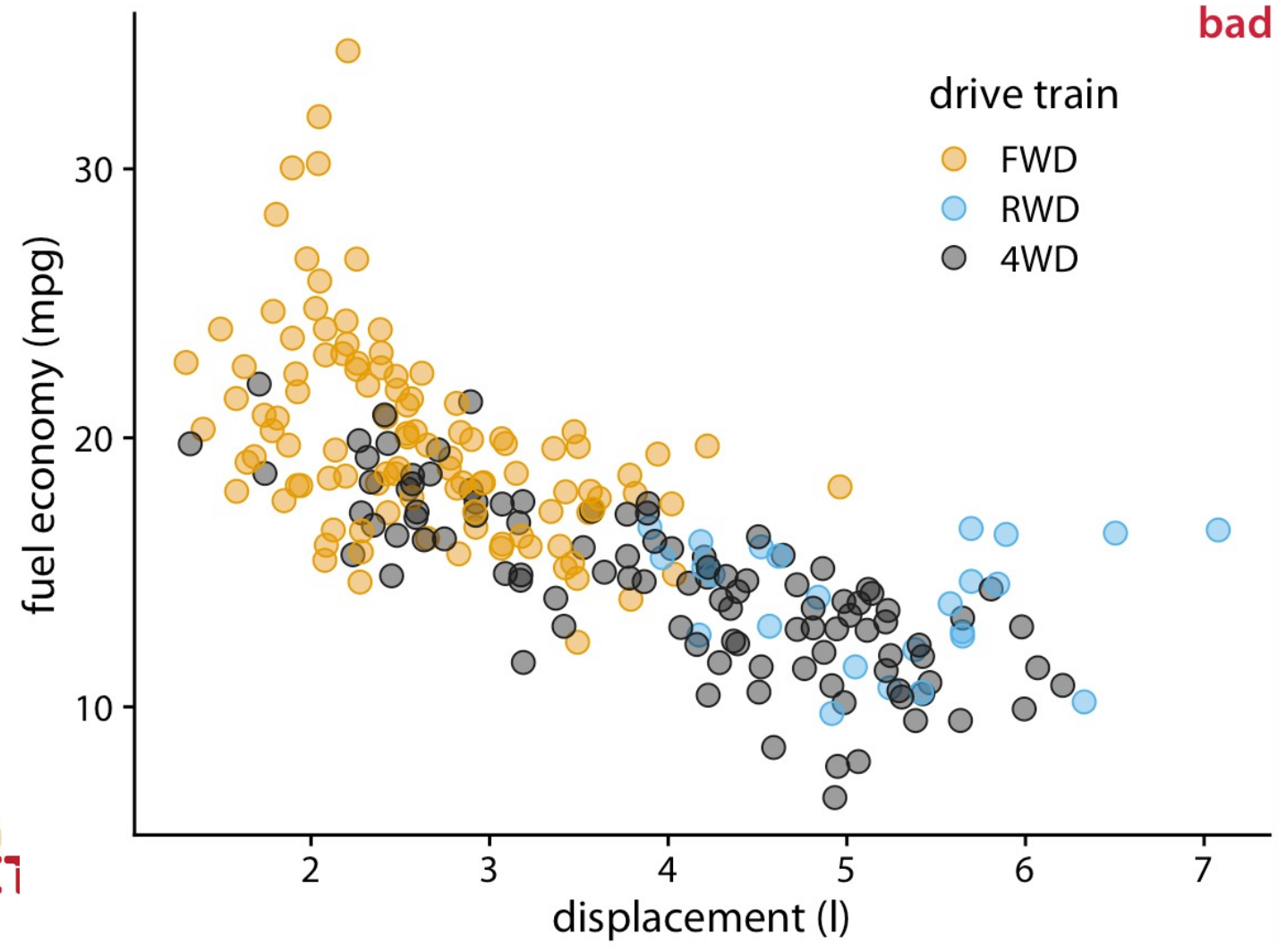
Example: City fuel economy versus engine displacement

- By adding a small amount of jitter to each point, we can increase the visibility of the overplotted points without substantially distorting the message of the plot.



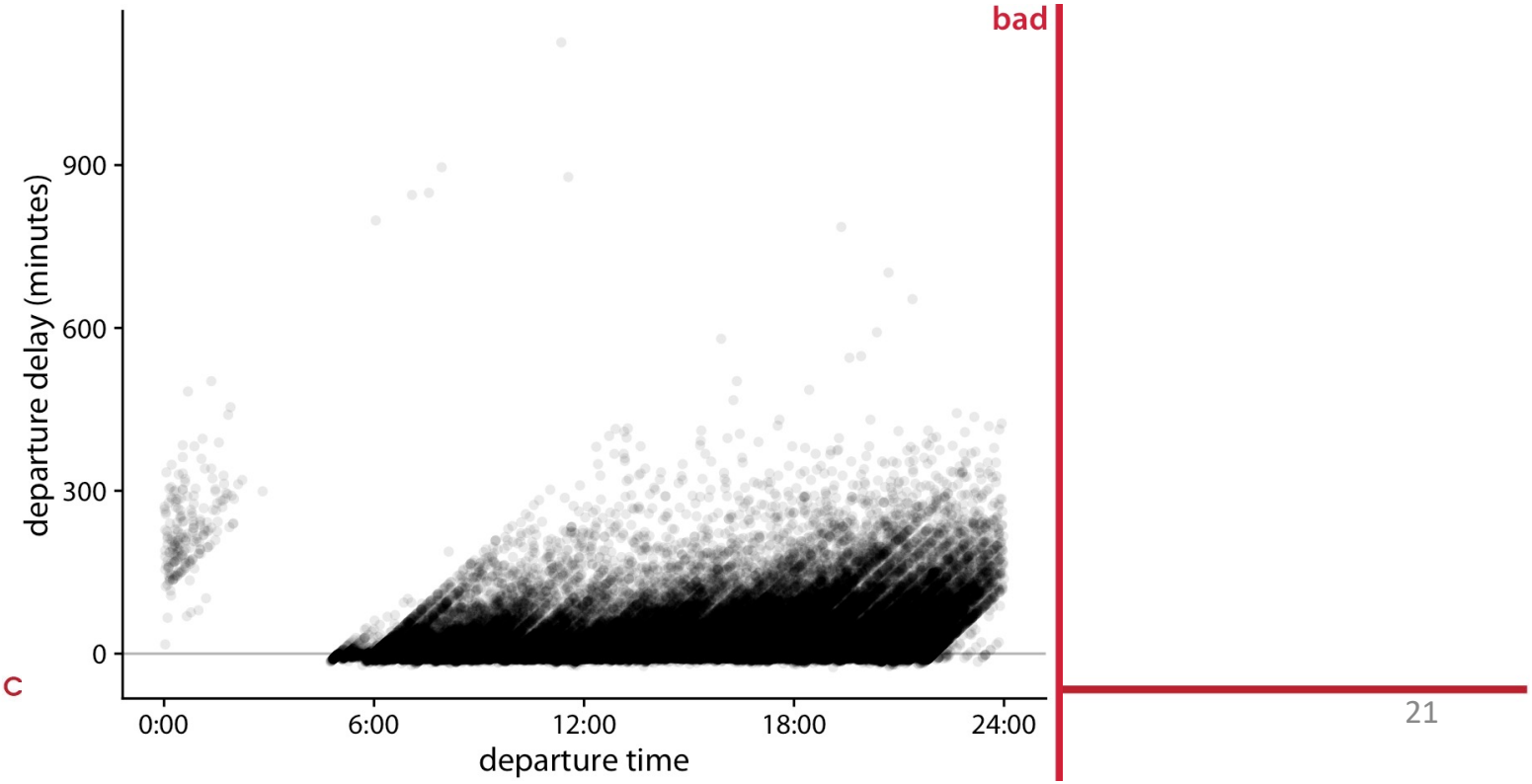
Adding too much jitter

- The visualization does not accurately reflect the underlying dataset



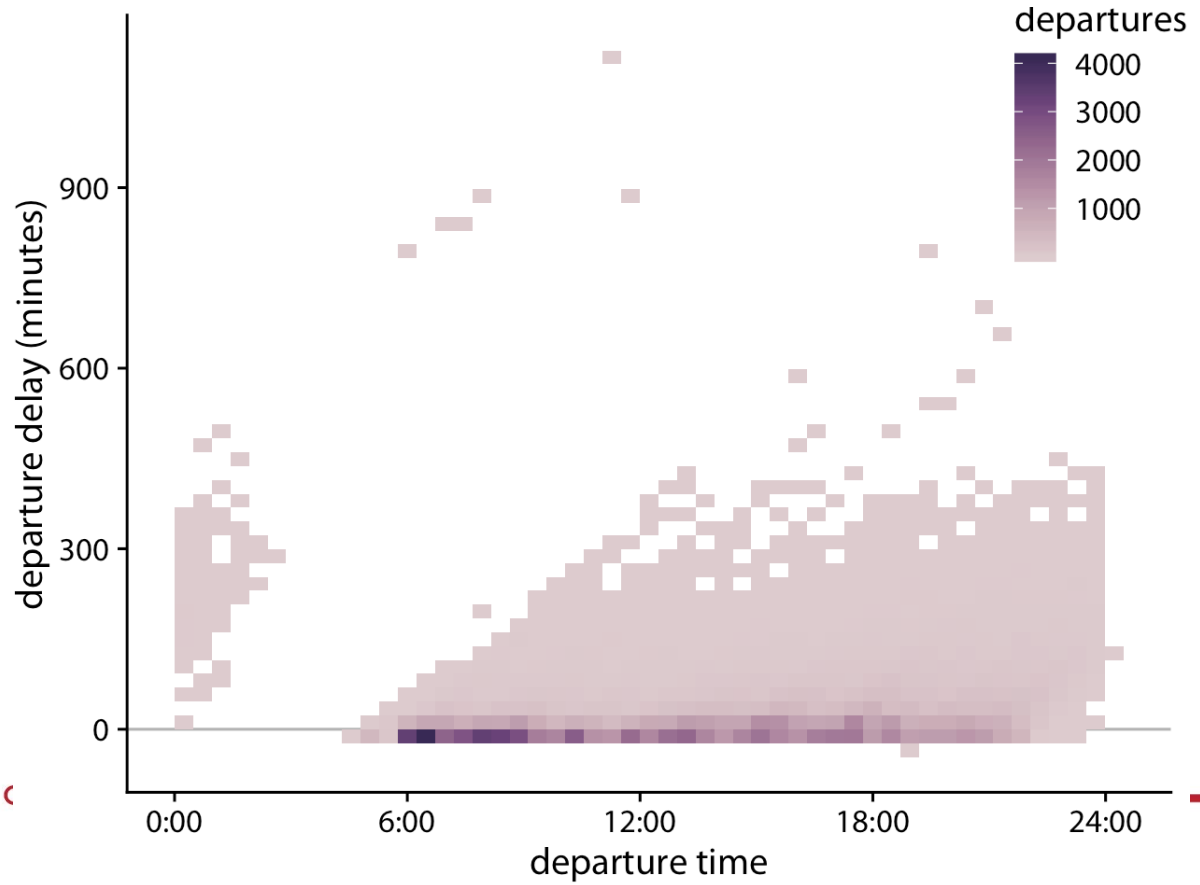
Example: Departure delay in minutes versus flight departure time

- When the number of individual points gets very large, partial transparency (with or without jittering) will not be sufficient
  - Areas with high point density will appear as uniform blobs of dark color,
  - Areas with low point density are barely visible



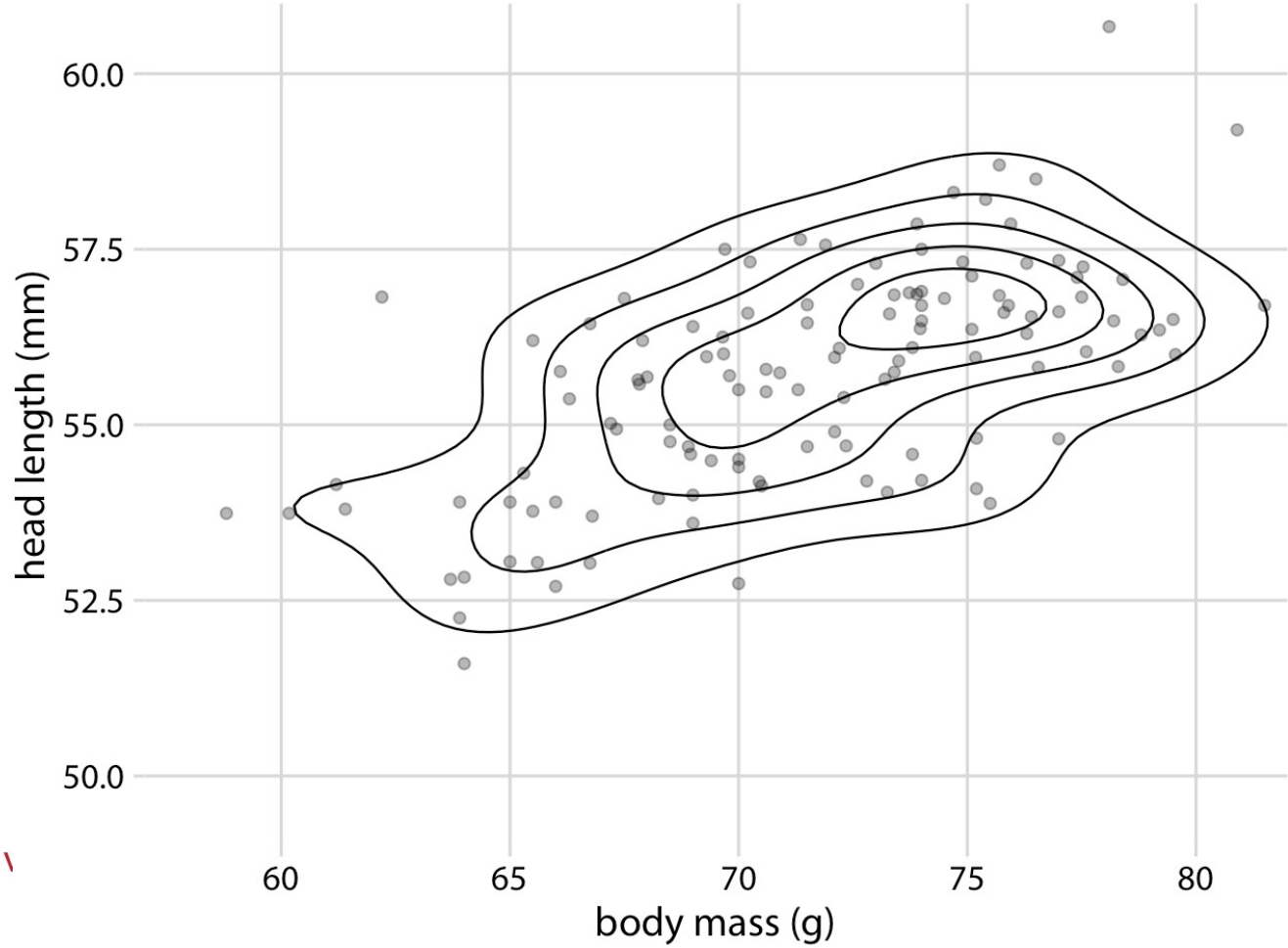
2D histogram

- Instead of plotting individual points, we bin the data in two dimensions
  - Subdivide the entire x–y plane into small rectangles, count how many observations fall into each one, and then color the rectangles by those counts.



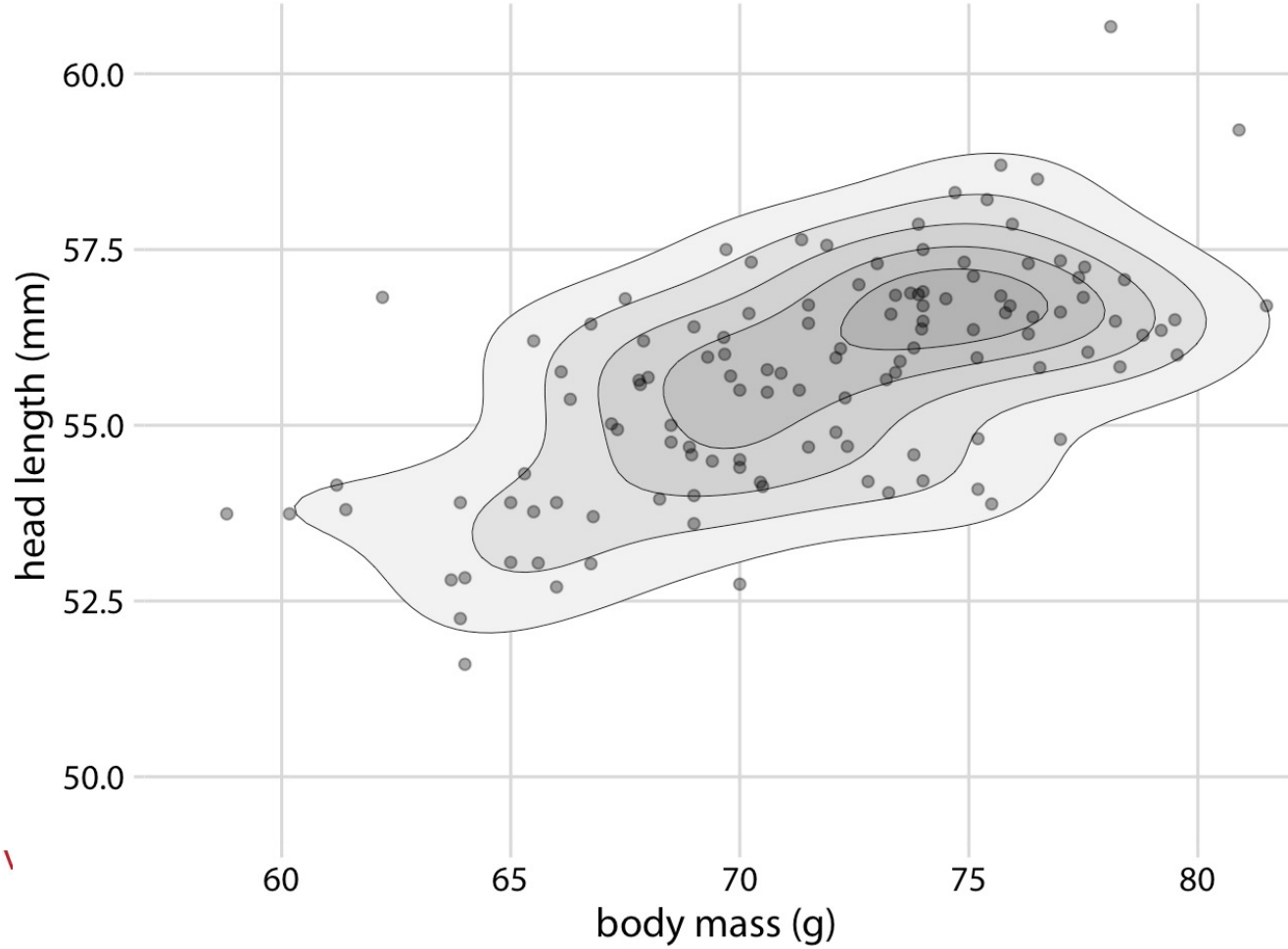
Example: Head length versus body mass for 123 blue jays

- Estimate the point density across the plot area.
- The lines indicate regions of similar point density.
- The point density increases toward the center of the plot.



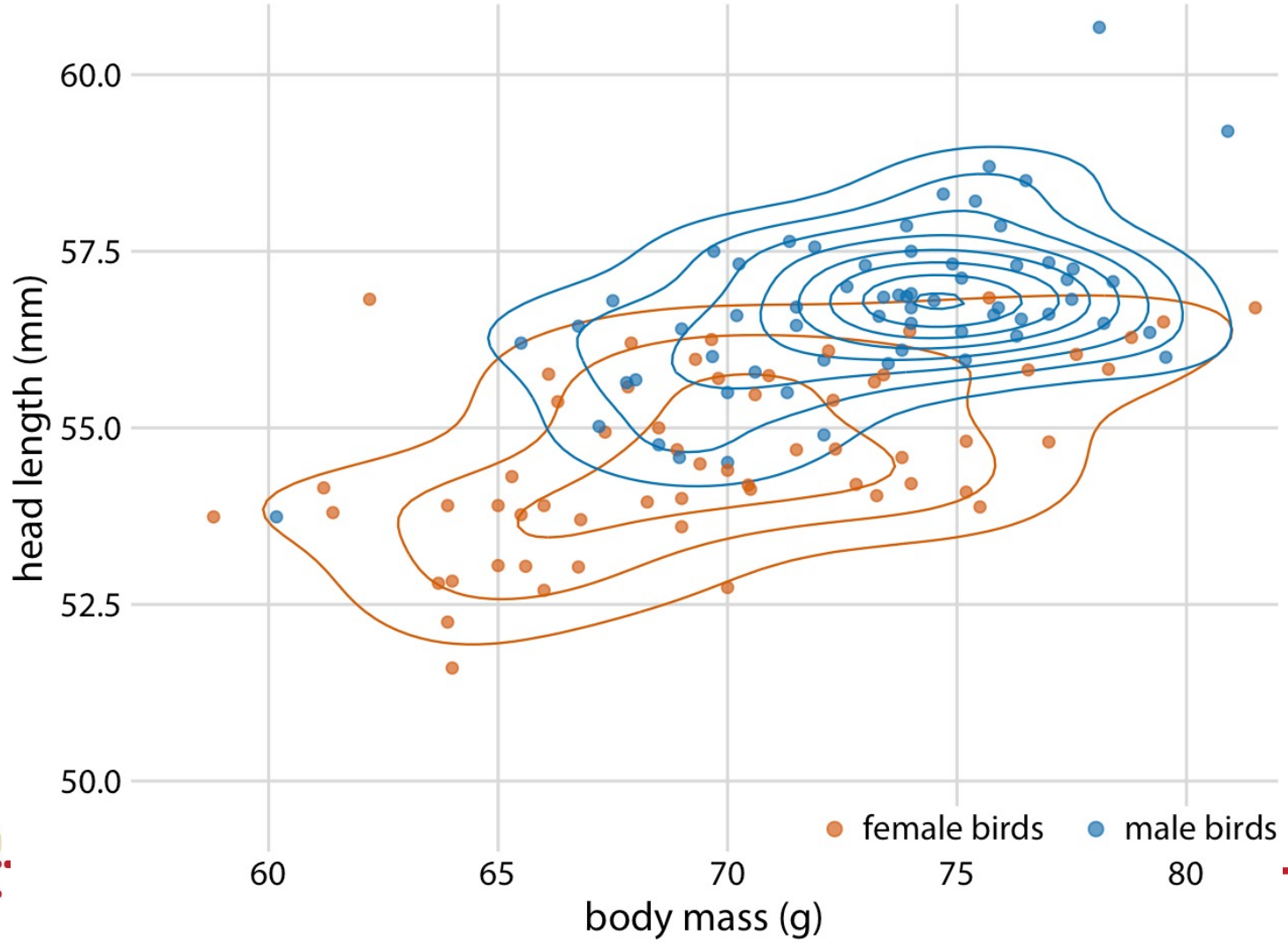
Example: Head length versus body mass for 123 blue jays

- The areas enclosed by the contour lines are shaded with increasingly darker shades of gray.
- This shading creates a stronger visual impression of increasing point density toward the center of the point cloud.

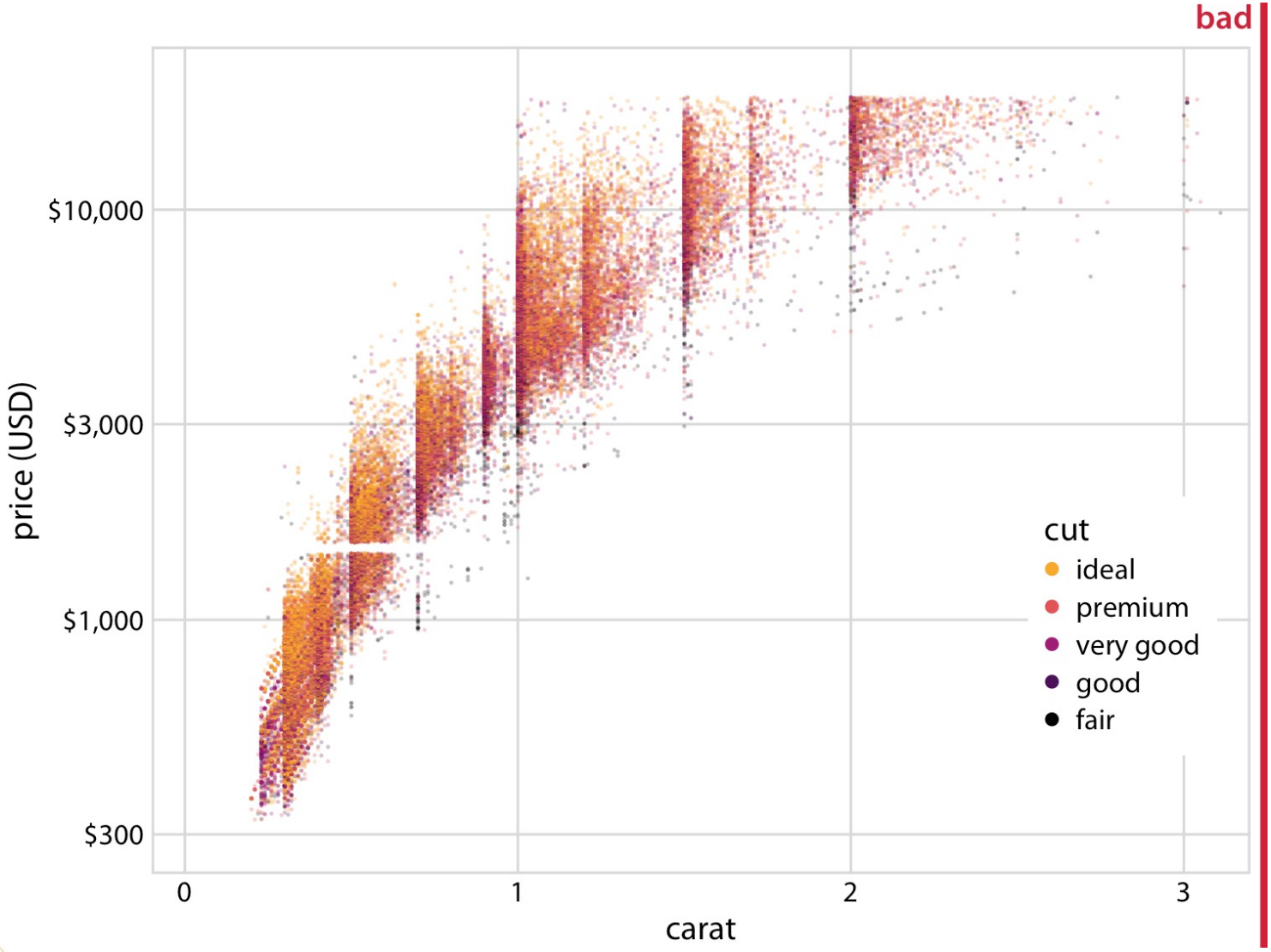


Example: Head length versus body mass for 123 blue jays

- Indicate the birds' sex by color when drawing contour lines.

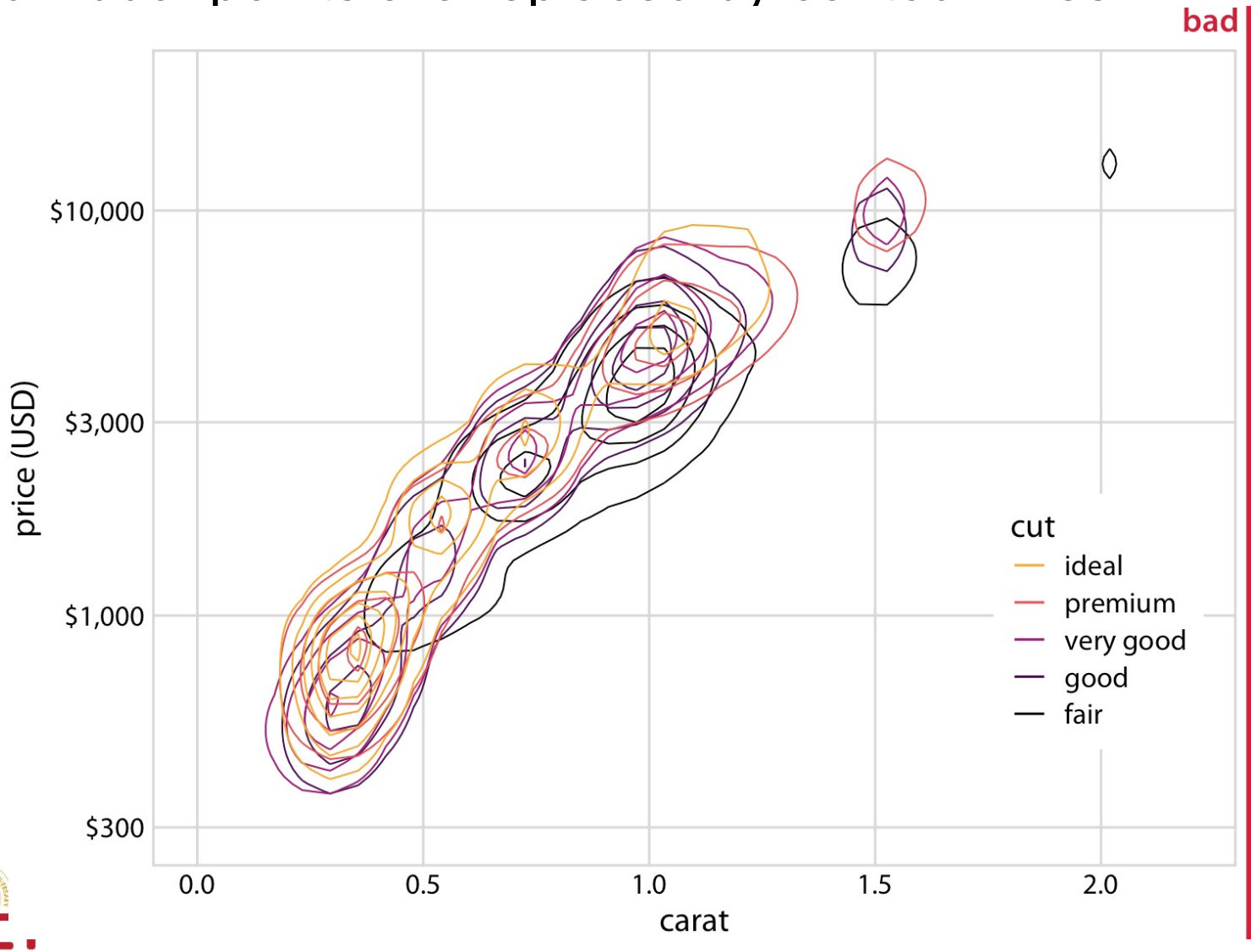


Example: Price of diamonds versus their carat value



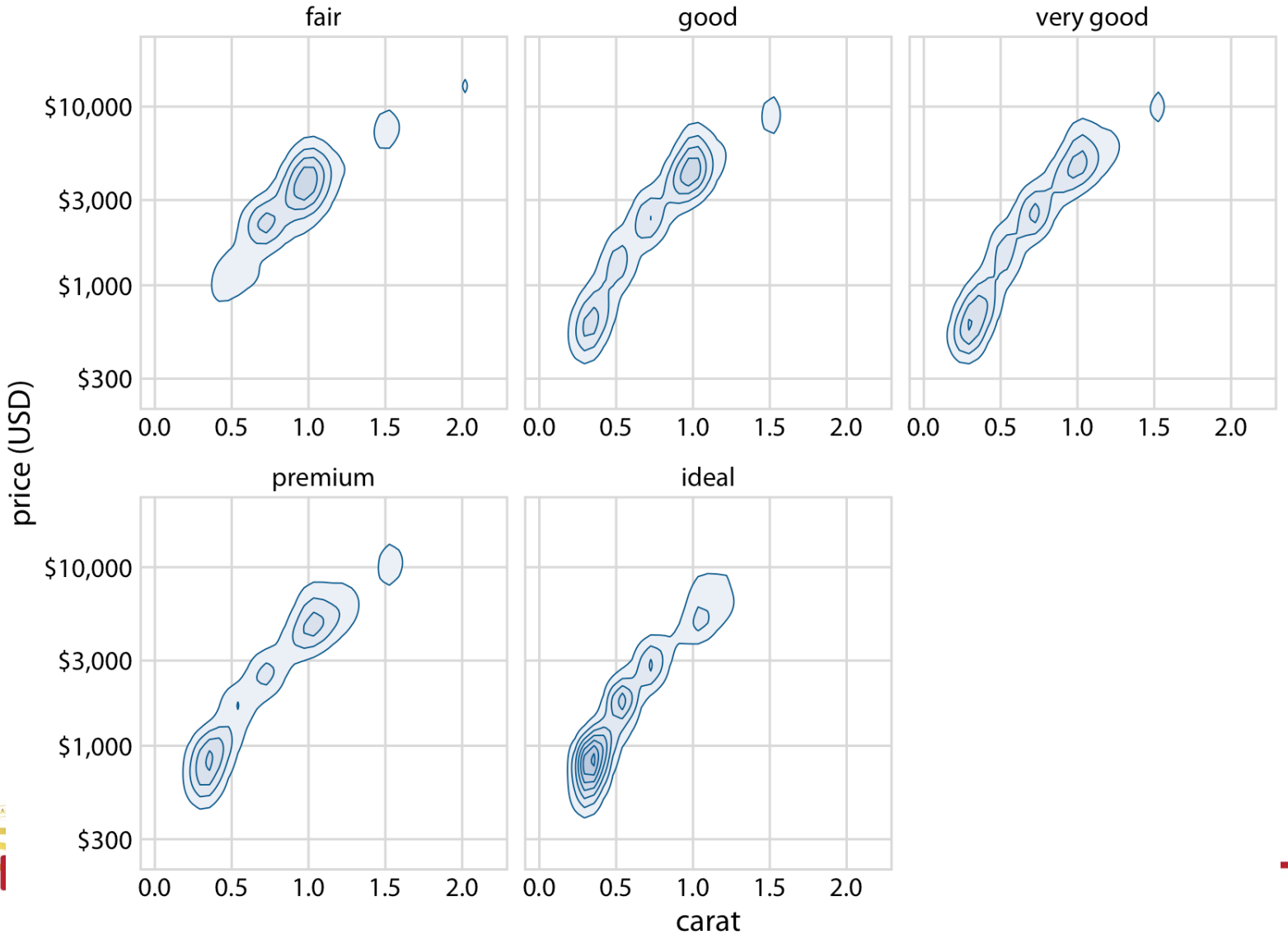
Example: Price of diamonds versus their carat value

- Individual points are replaced by contour lines



Example: Price of diamonds versus their carat value

- Draw the contour lines for each cut quality in its own plot panel.

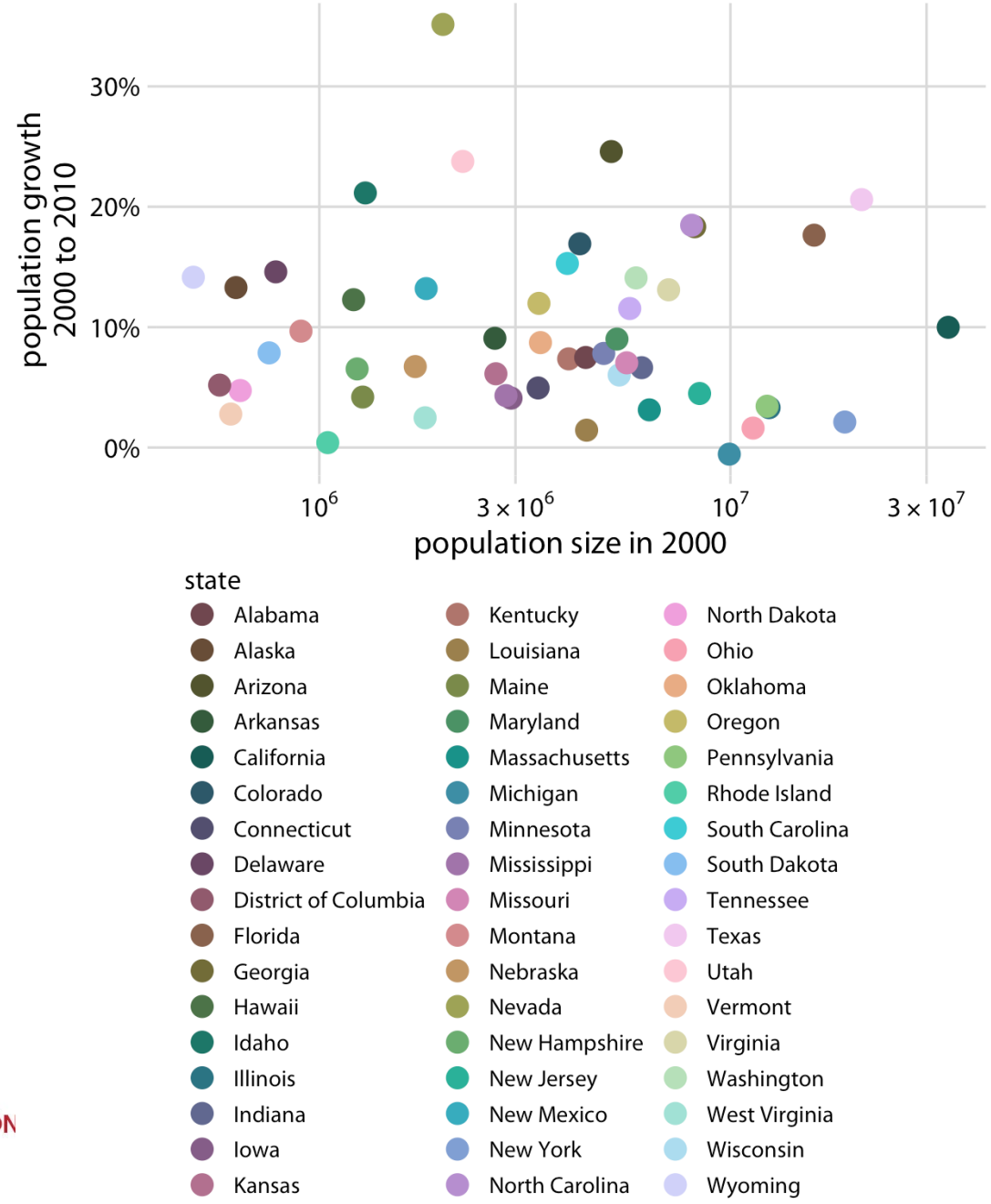


Common pitfalls of color use

Encoding too much or irrelevant information

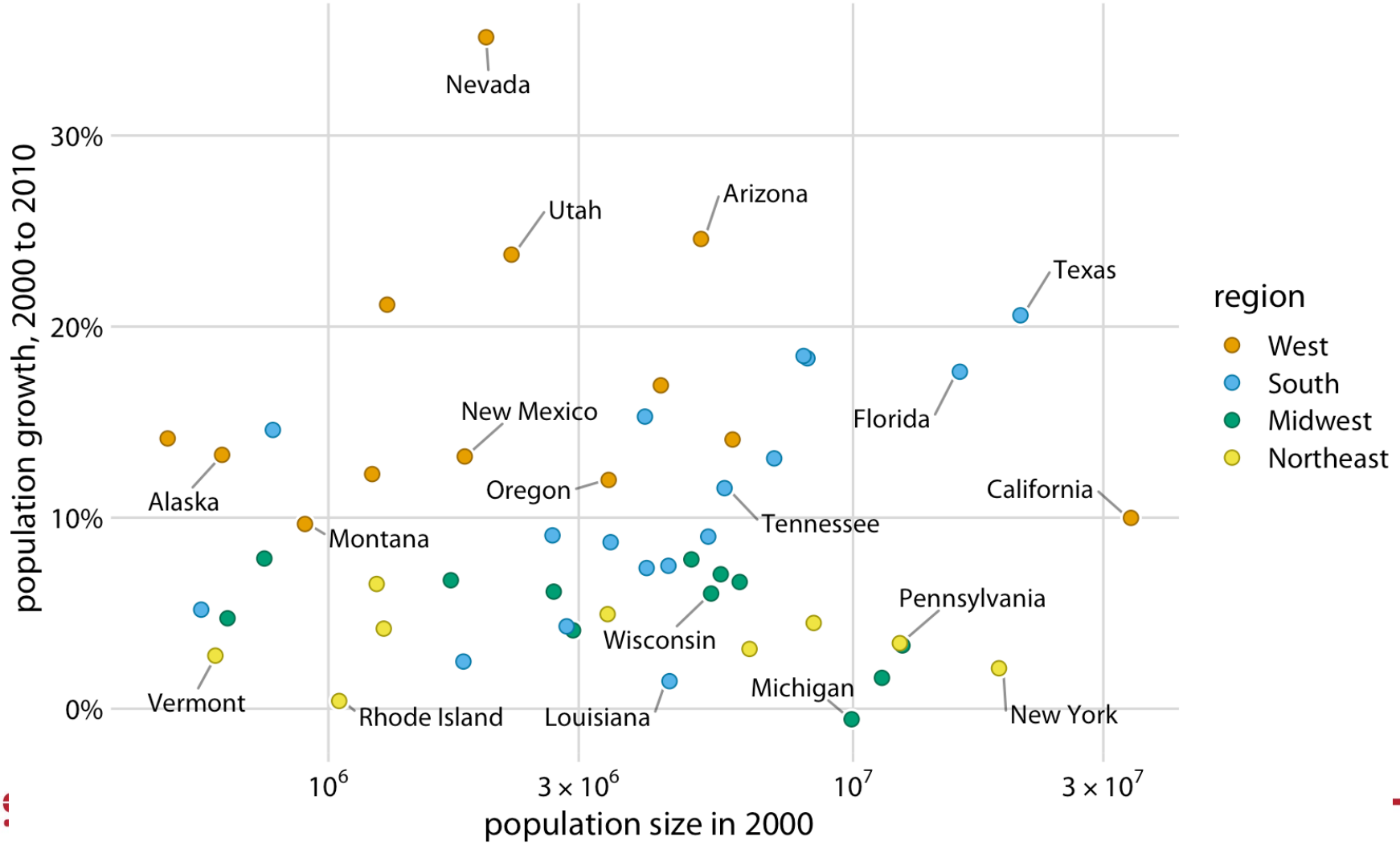
- Population growth versus population size for all 50 US states and the District of Columbia.
  - Qualitative color scales work best when there are three to five different categories that need to be colored

bad



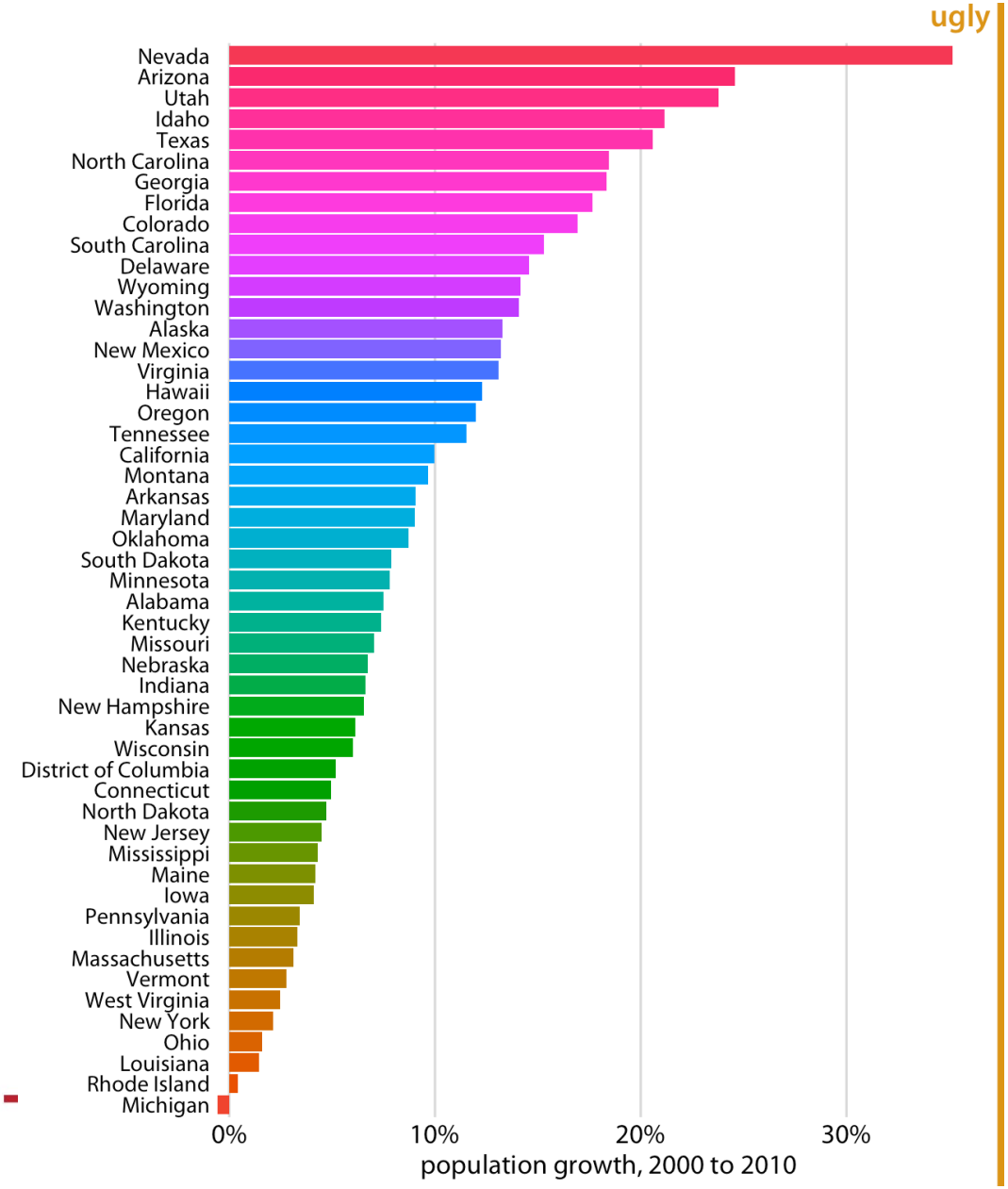
Example: Population growth from 2000 to 2010 versus population size in 2000

- Use color only to indicate the geographic region of each state
- Identify individual states by direct labeling.



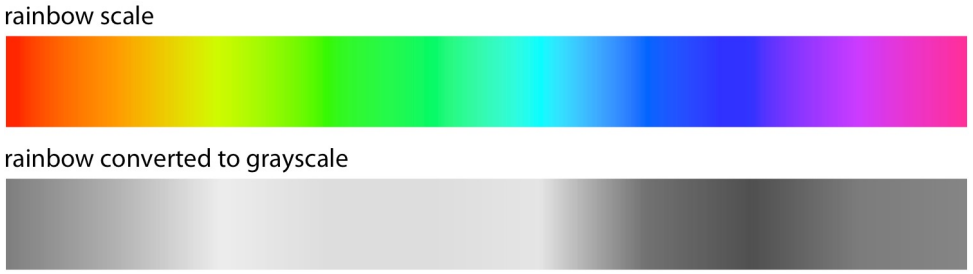
Example: Population growth in the US from 2000 to 2010

- Coloring for the sake of coloring, without having a clear purpose for the colors

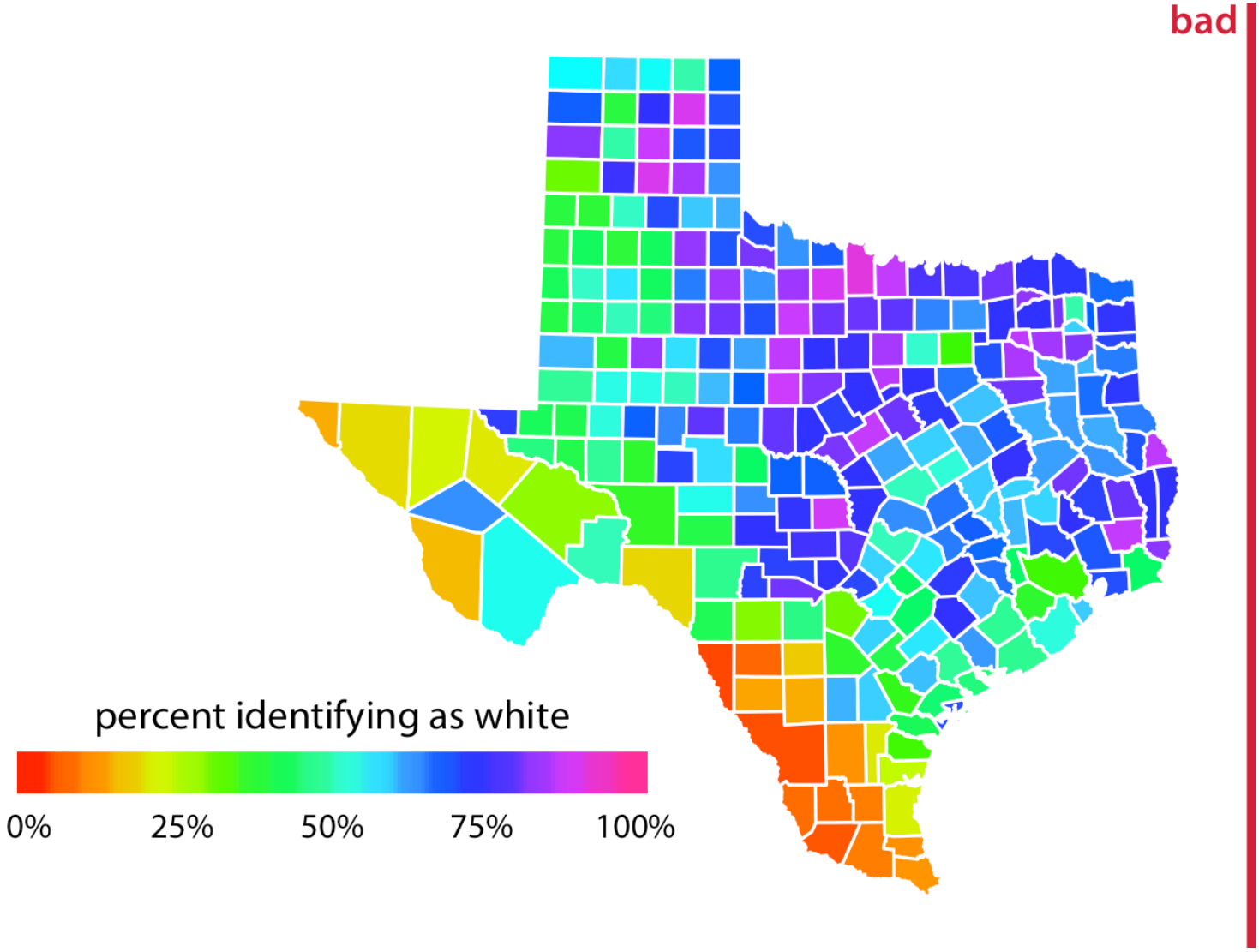


Using nonmonotonic color scales to encode data values

- Sequential color scales that can represent data values
  - The colors need to clearly indicate which data values are larger or smaller than which other ones, and the differences between colors need to visualize the corresponding differences between data values.
- The rainbow color scale is highly nonmonotonic
  - Tends to obscure data features and/or highlight arbitrary aspects of the data.
  - The colors are overly saturated.



Example: Percentage of people identifying as white in Texas counties



Not designing for Color-Vision deficiency

- Some of the readers may have some form of color-vision deficiency (CVD i.e., are colorblind)
  - Red–green color-vision deficiency.
  - Blue and green (blue–yellow color-vision deficiency).
- Sequential scales will generally not cause any problems for people with CVD
  - It presents a continuous gradient from dark to light colors.

original

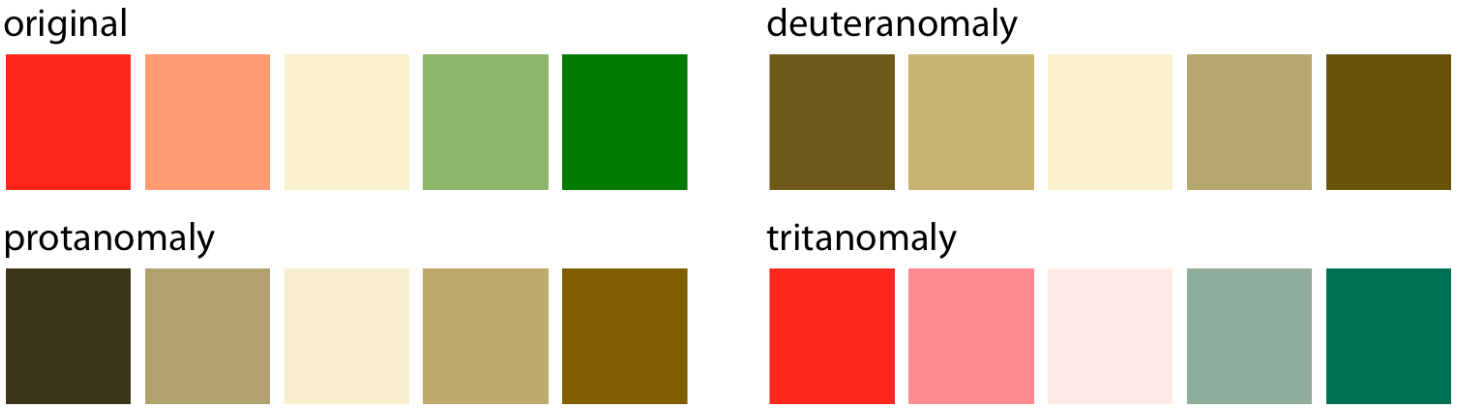
deuteranomaly

protanomaly

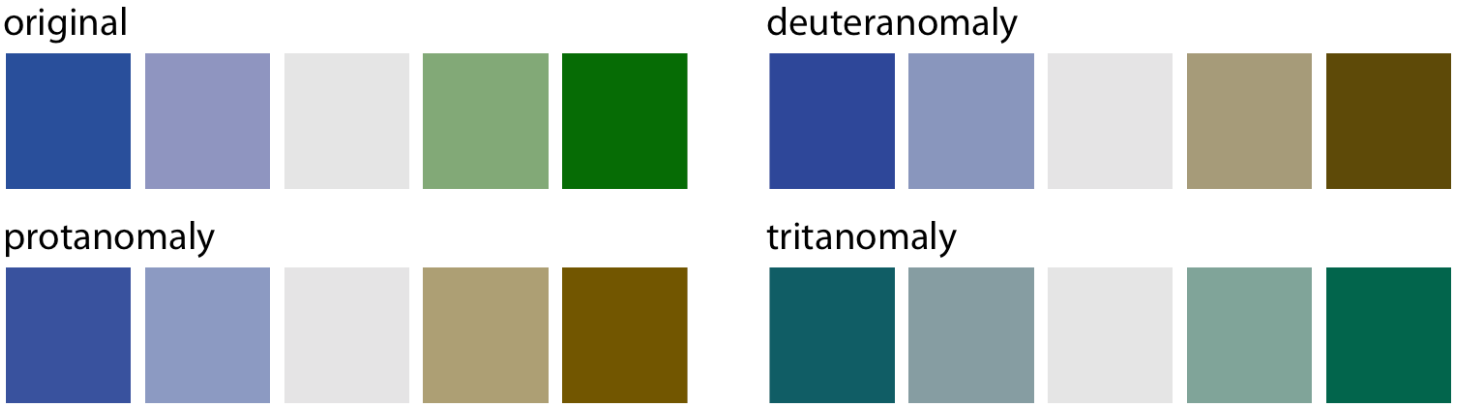
tritanomaly

Diverging color scales

- A red–green contrast becomes indistinguishable under red–green CVD (deuteranomaly or protanomaly)

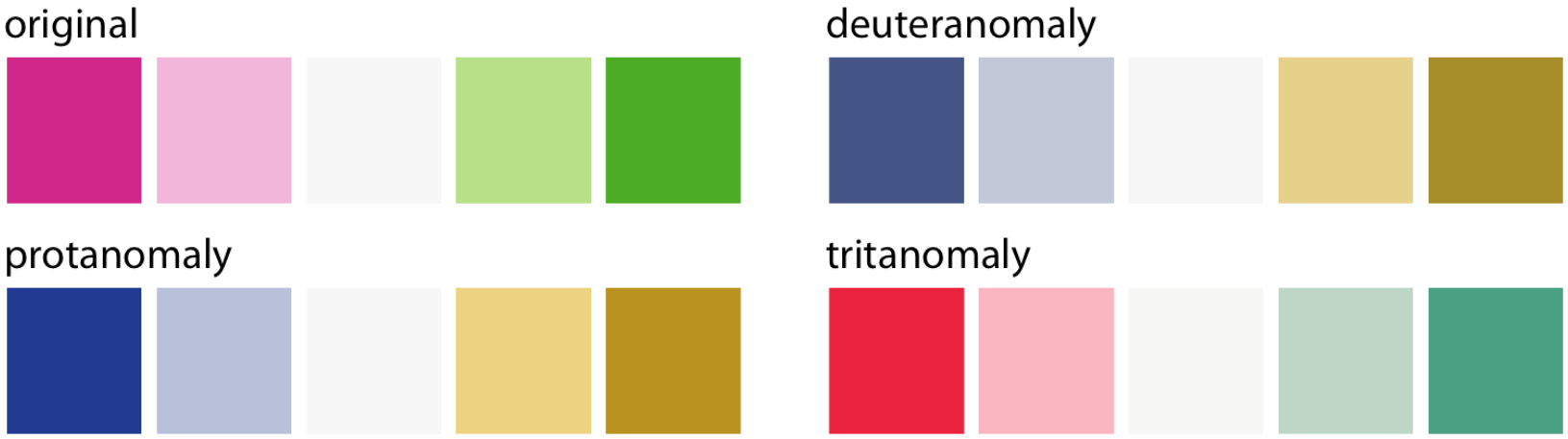


- A blue–green contrast becomes indistinguishable under blue–yellow CVD (tritanomaly).



Diverging color scales

- The ColorBrewer PiYG (pink to yellow-green) scale looks like a red–green contrast to people with regular color vision but works for people with all forms of color-vision deficiency.



Qualitative color scales

- Different colors need to be distinguishable from each other under all forms of CVD.
- Qualitative color palette for all color-vision deficiencies.

- Colorblind-friendly color scale [Okabe and Ito 2008].

Name	Hex code	Hue	C, M, Y, K (%)	R, G, B (0–255)	R, G, B (%)
Orange	#E69F00	41°	0, 50, 100, 0	230, 159, 0	90, 60, 0
Sky blue	#56B4E9	202°	80, 0, 0, 0	86, 180, 233	35, 70, 90
Bluish green	#009E73	164°	97, 0, 75, 0	0, 158, 115	0, 60, 50
Yellow	#F0E442	56°	10, 5, 90, 0	240, 228, 66	95, 90, 25
Blue	#0072B2	202°	100, 50, 0, 0	0, 114, 178	0, 45, 70
Vermilion	#D55E00	27°	0, 80, 100, 0	213, 94, 0	80, 40, 0
Reddish purple	#CC79A7	326°	10, 70, 0, 0	204, 121, 167	80, 60, 70
Black	#000000	N/A	0, 0, 0, 100	0, 0, 0	0, 0, 0

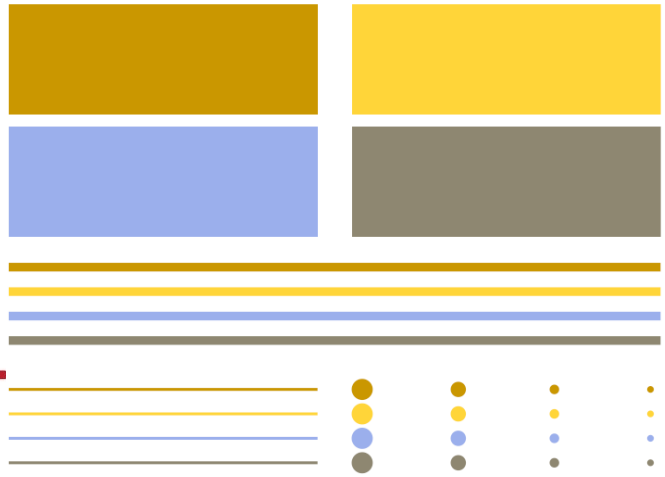
Colored elements become difficult to distinguish at small sizes

- This problem becomes exacerbated in the CVD simulations.

original

deuteranomaly

protanomaly



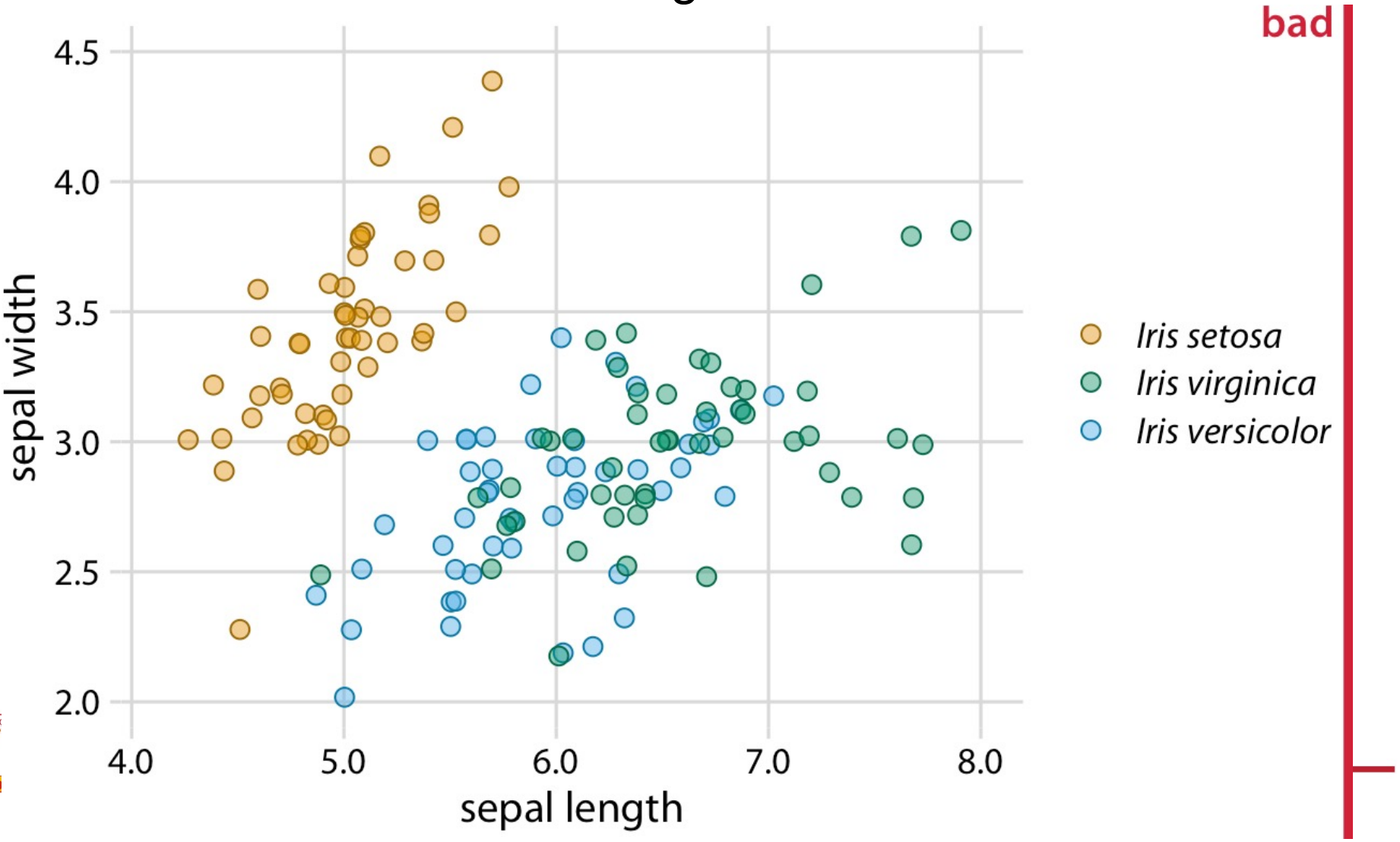
tritanomaly

Redundant coding

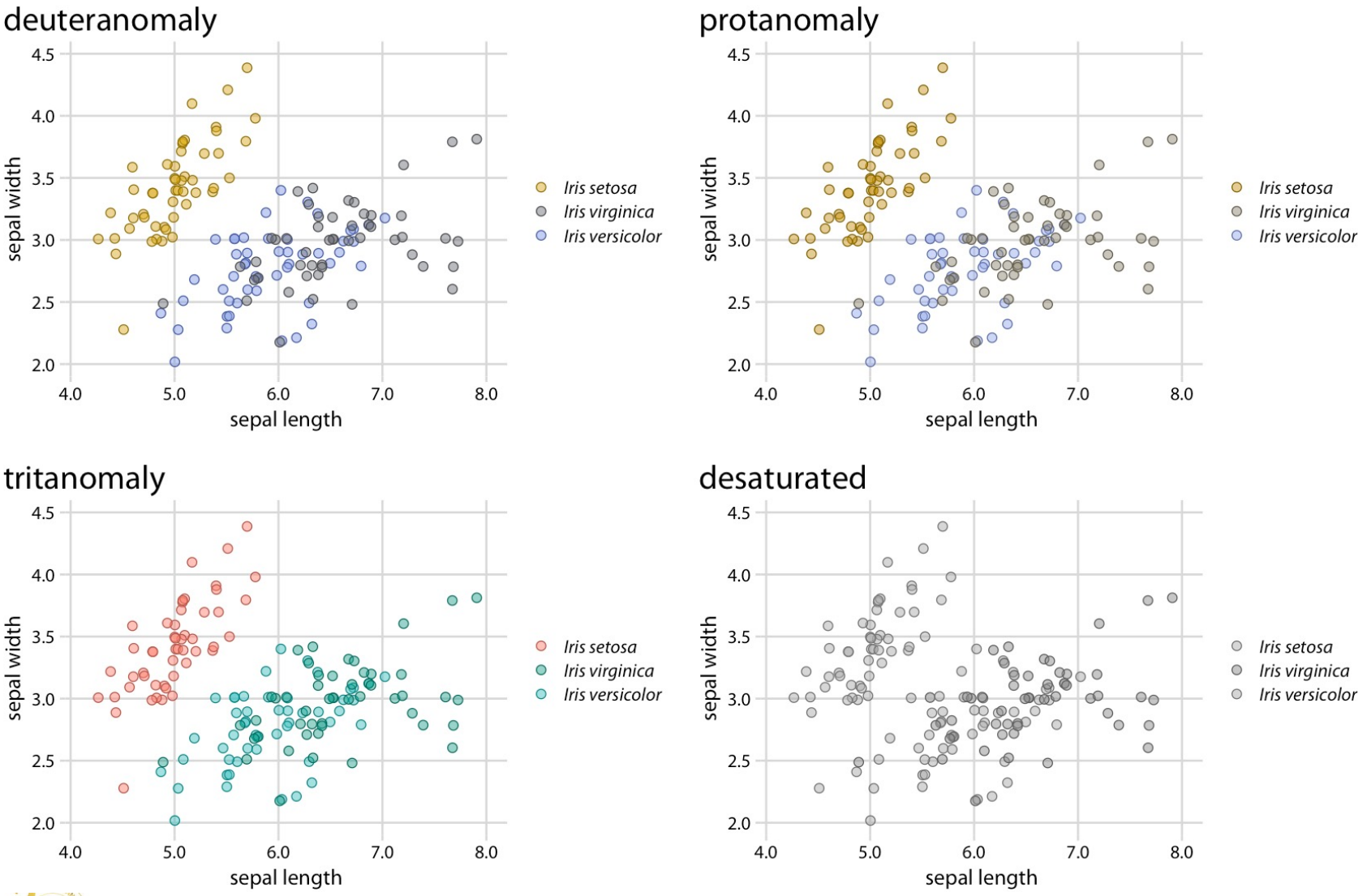
Use color to enhance the visual appearance of the figure without relying entirely on color to convey key information.

Example: Sepal width versus sepal length for three different

- The virginica points in green and the versicolor points in blue are difficult to distinguish from each other.

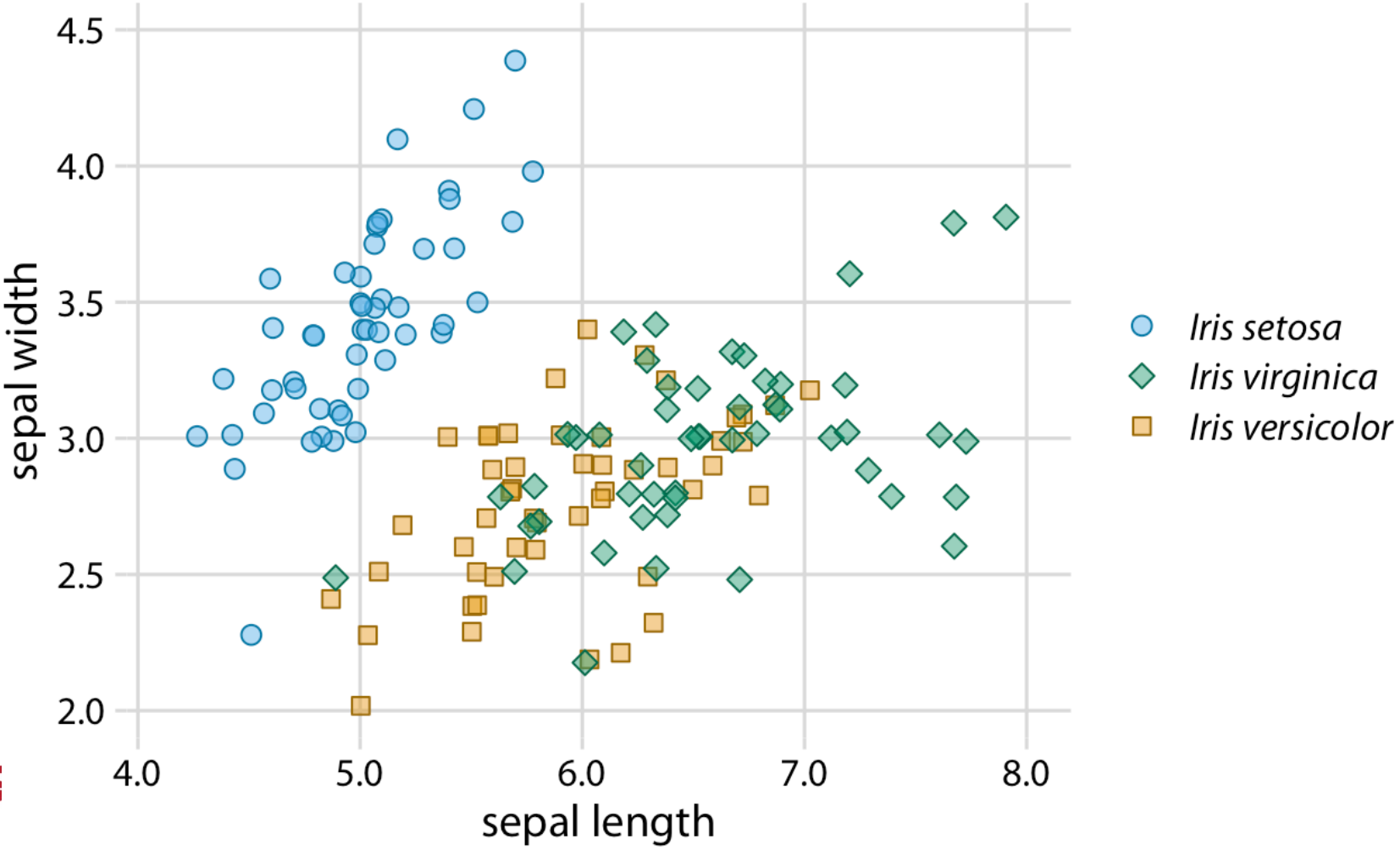


Color-vision deficiency simulation



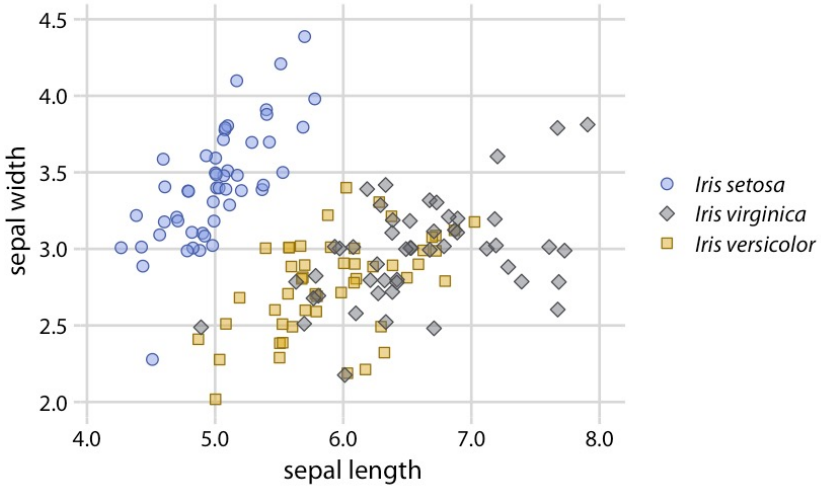
Example: Sepal width versus sepal length for three different Iris species

- Swap the colors used for Iris setosa and Iris versicolor, so that the blue is no longer directly next to the green.
- Use three different symbol shapes, so that the points all look different.

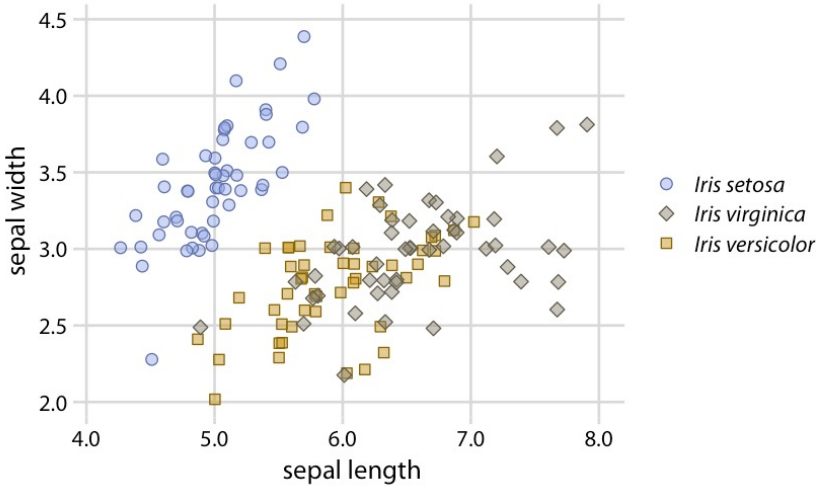


Example: Color-vision deficiency simulation

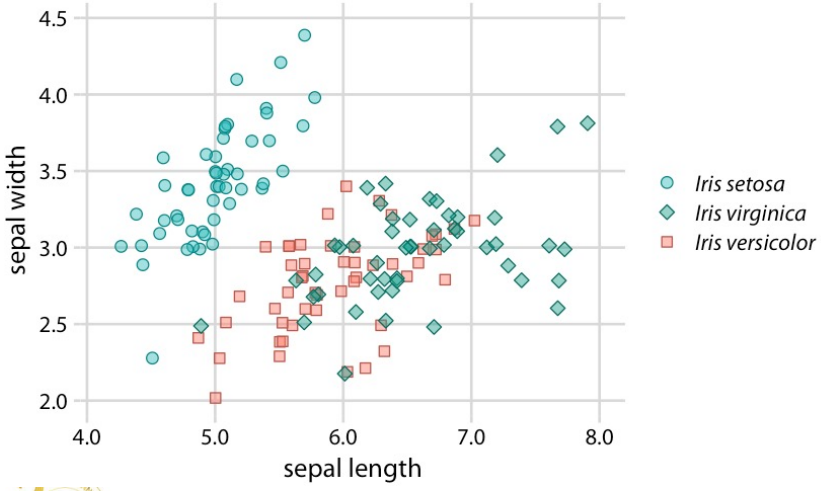
deuteranomaly



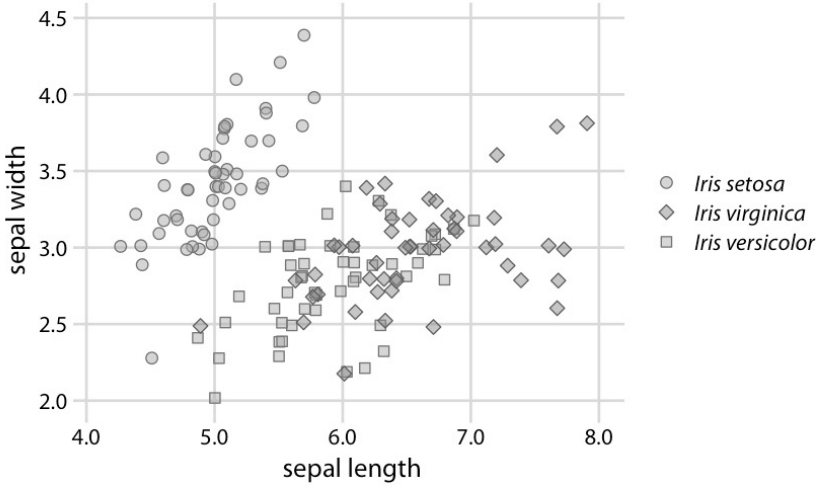
protanomaly



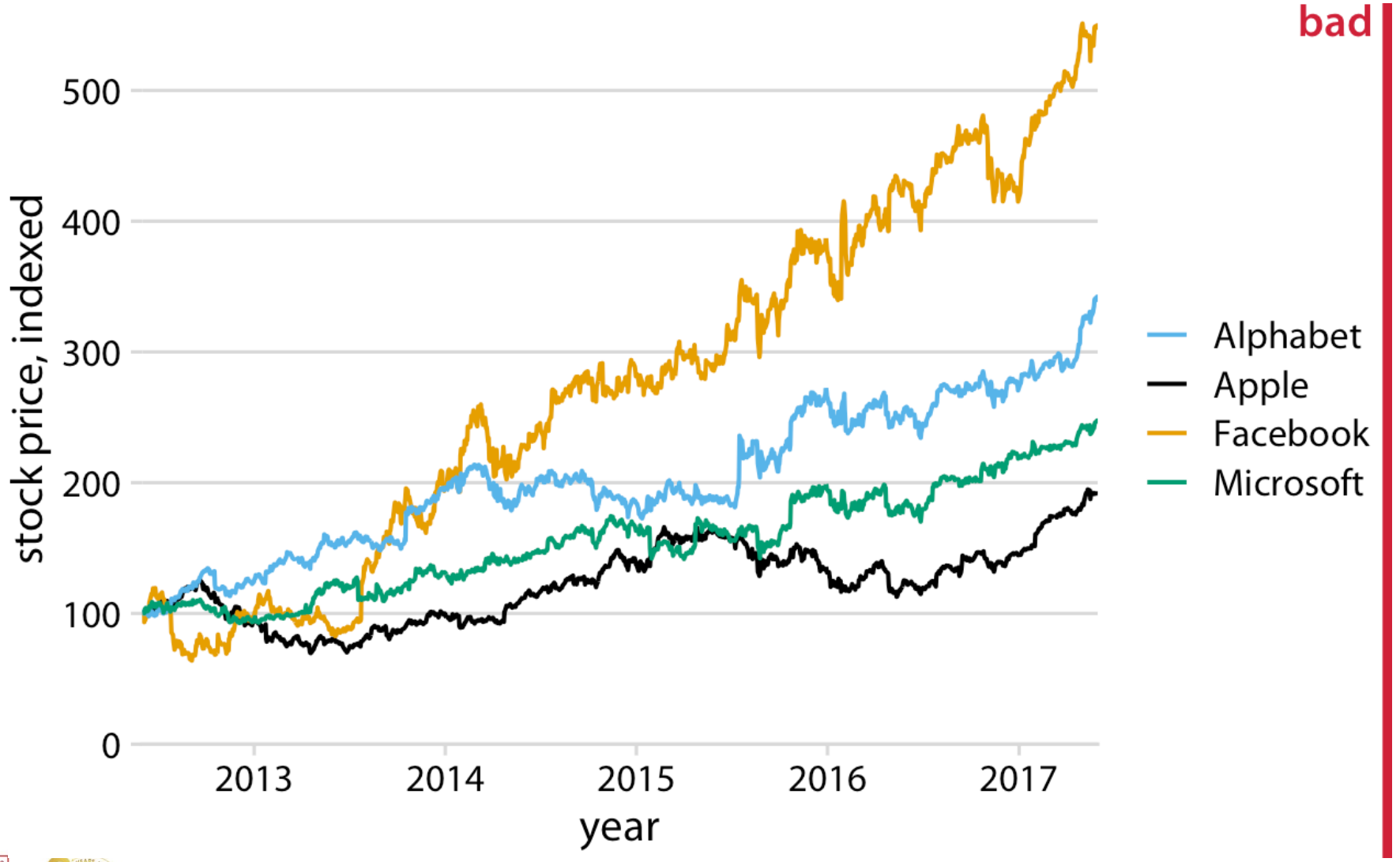
tritanomaly



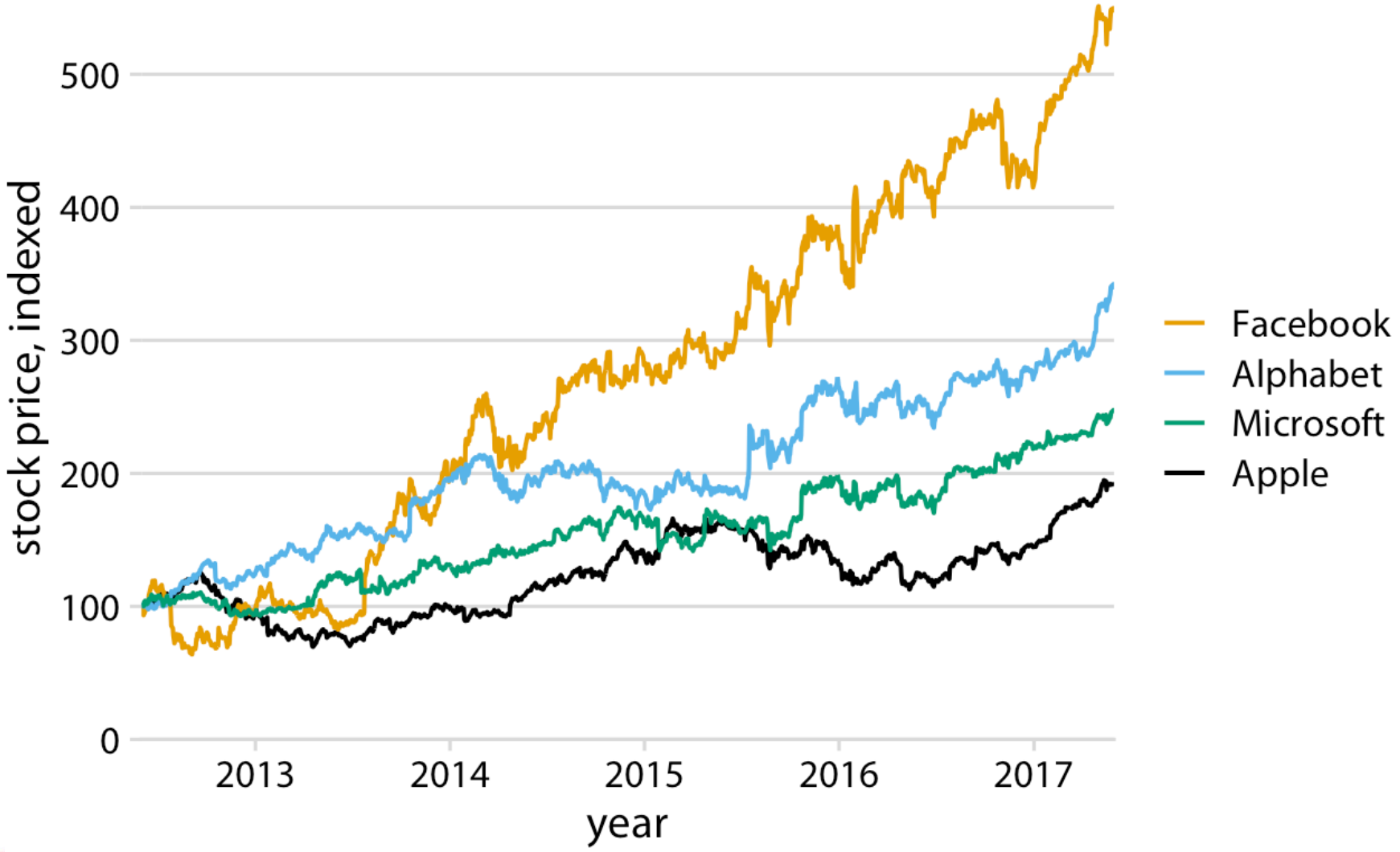
desaturated



Example: Stock price over time for four major tech companies

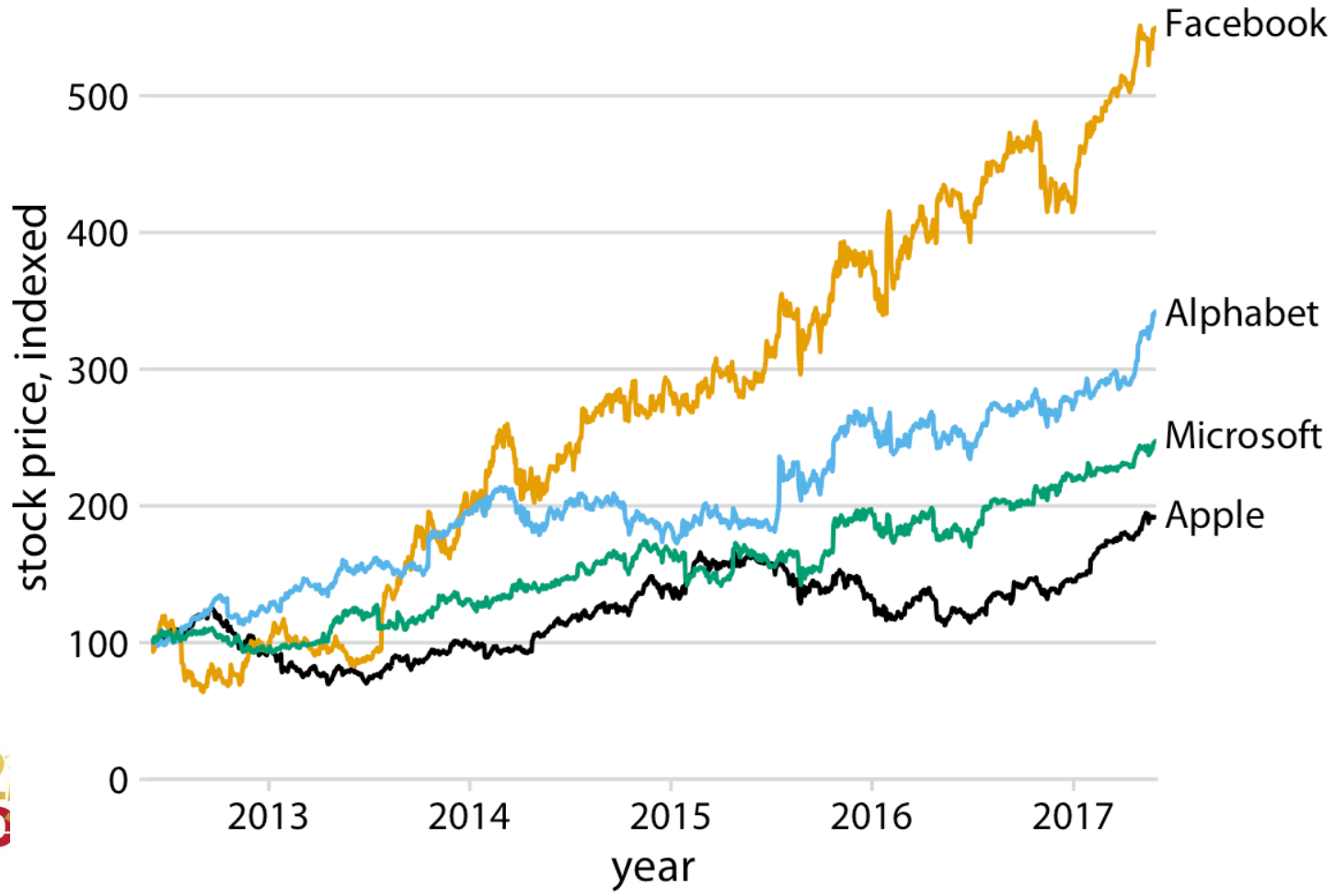


Example: Stock price over time for four major tech companies

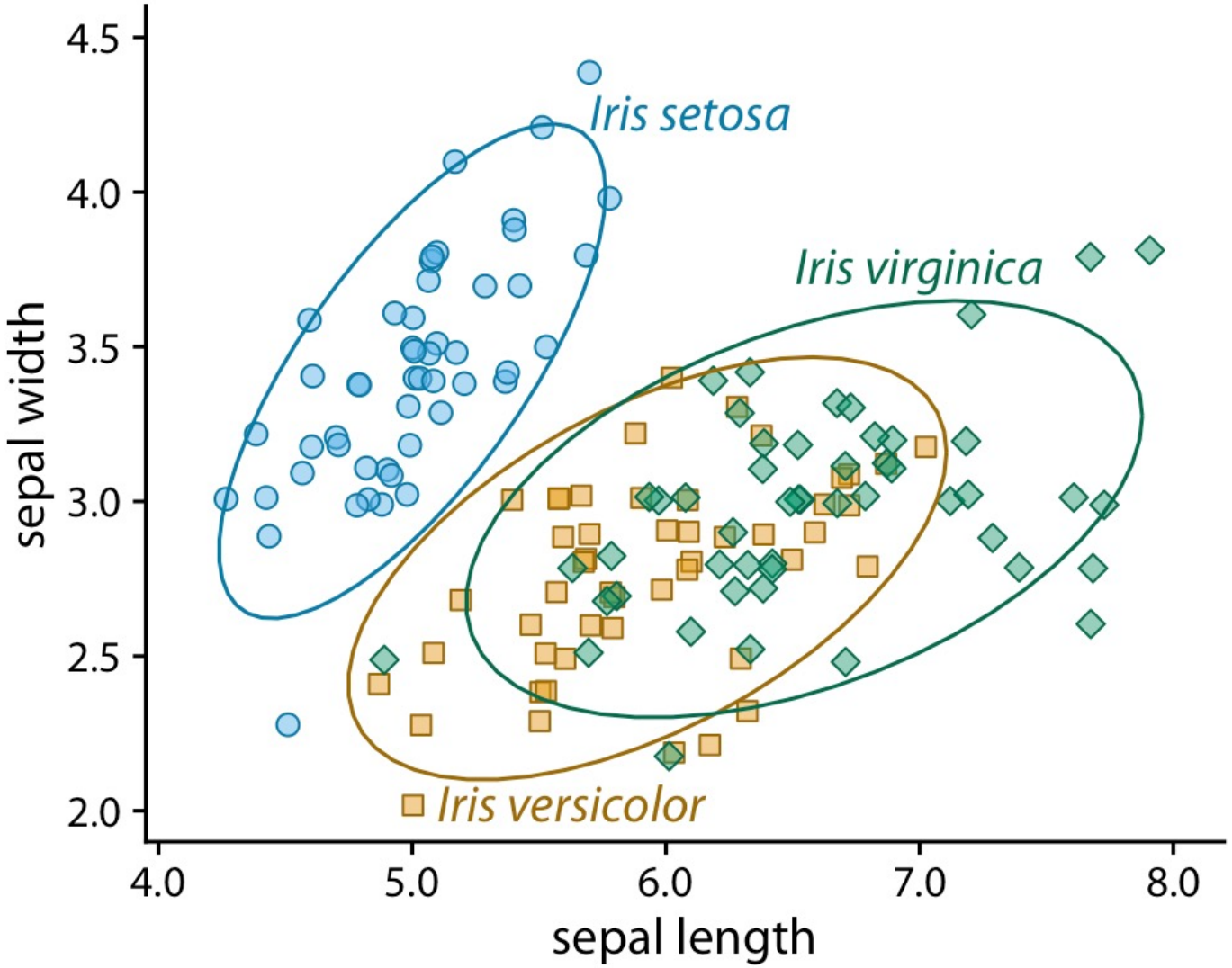


Designing figures without legends

- Whenever possible, design your figures so they don't need a separate legend.
- In direct labeling, we incorporate appropriate text labels or other visual elements that serve as guideposts to the rest of the figure.

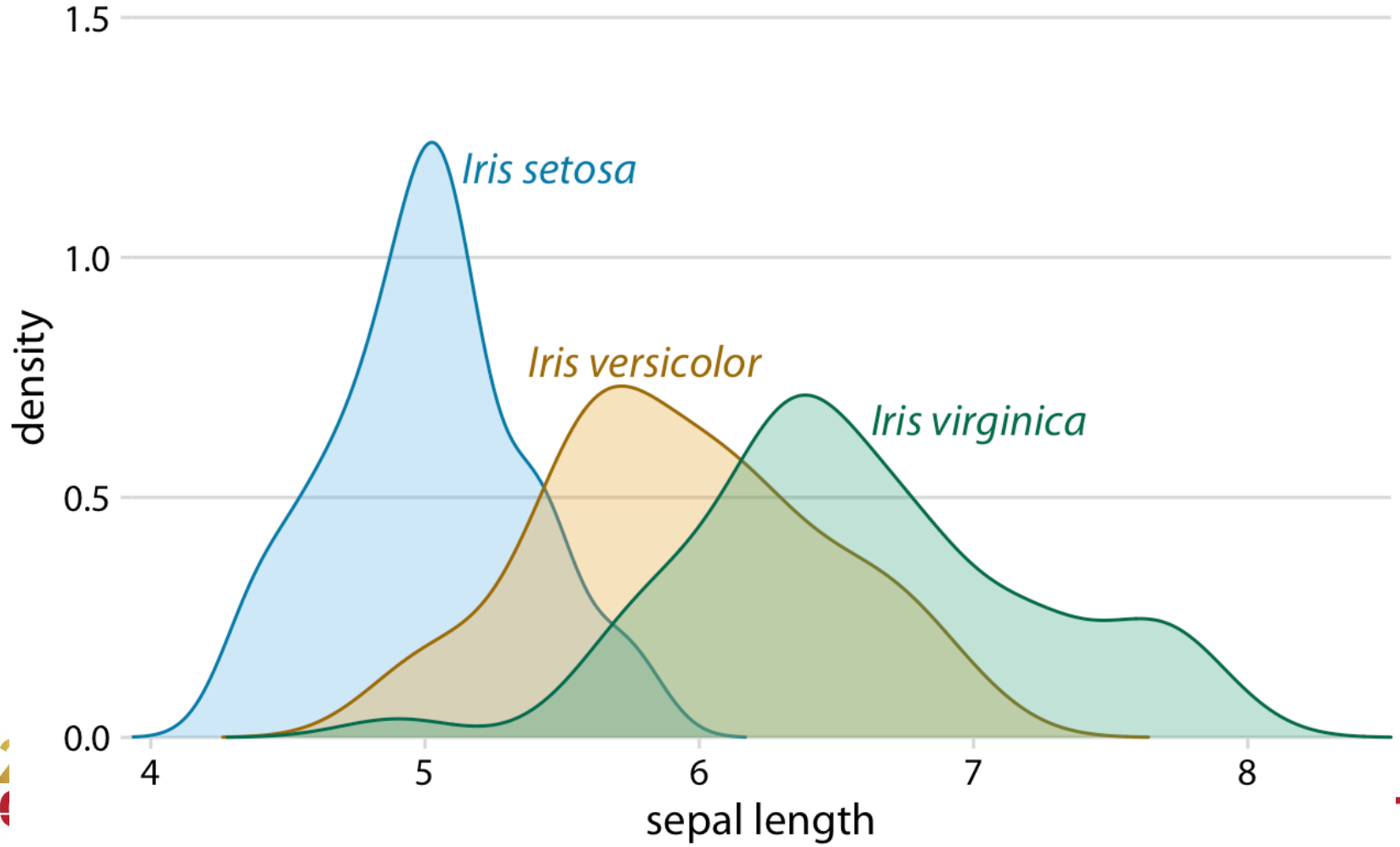


Example: Sepal width versus sepal length for three different

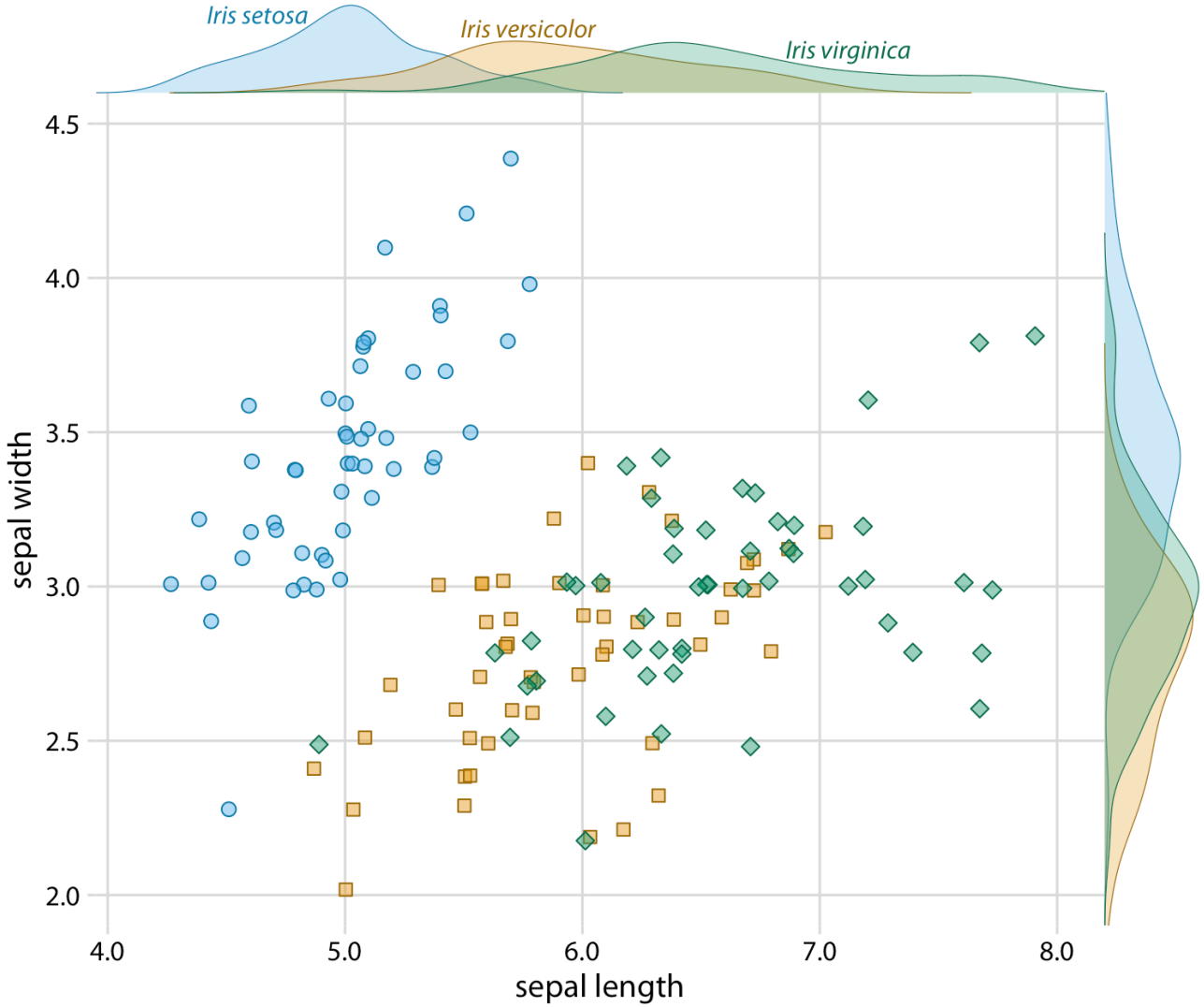


Example: Density estimates of the sepal lengths of three different Iris species

- Each density estimate is directly labeled with the respective species name.

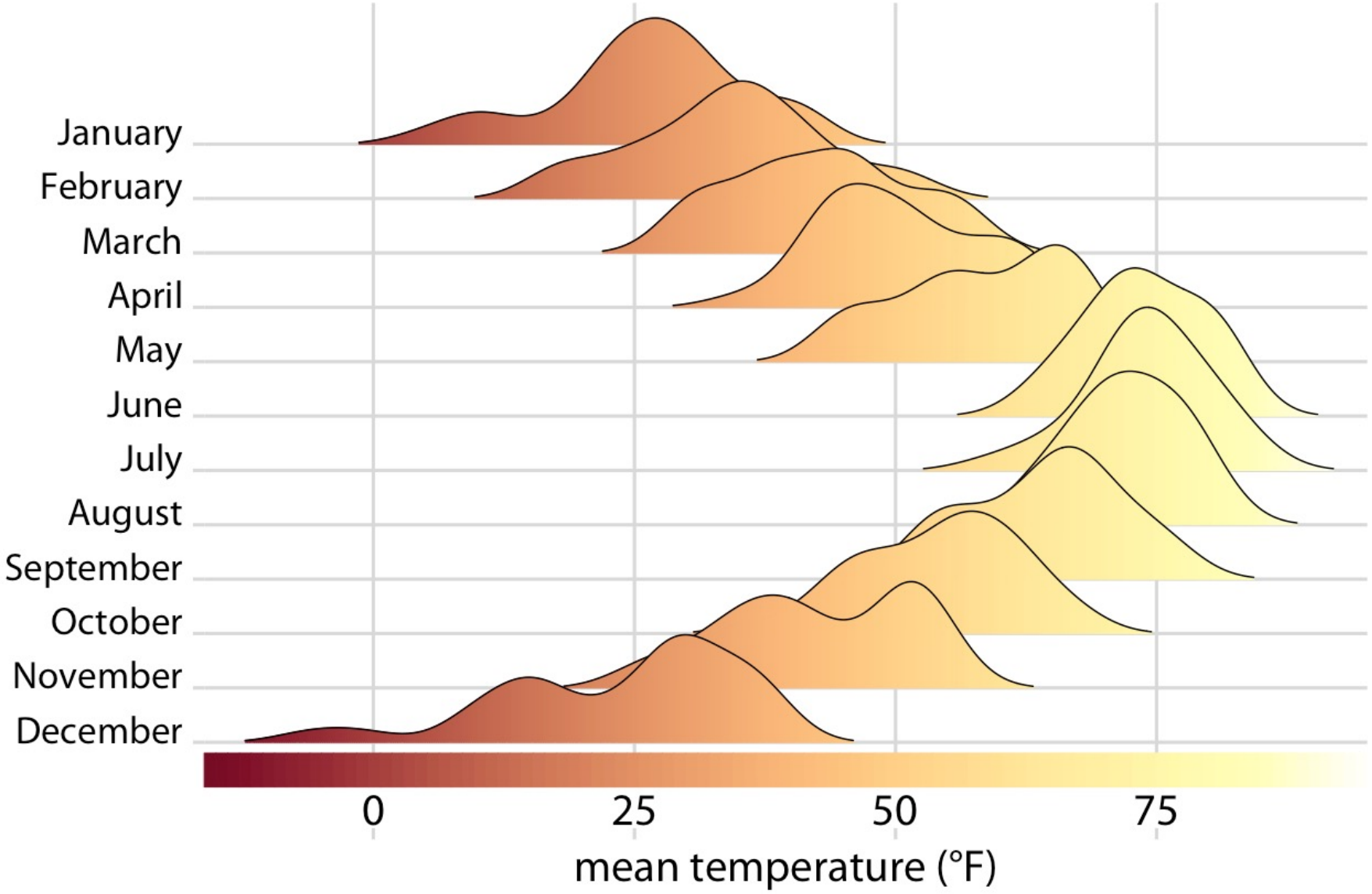


Density plots into the margins of a scatterplot



Example: Temperatures in Lincoln, NE, in 2016

- Integrate the color legend into the x axis



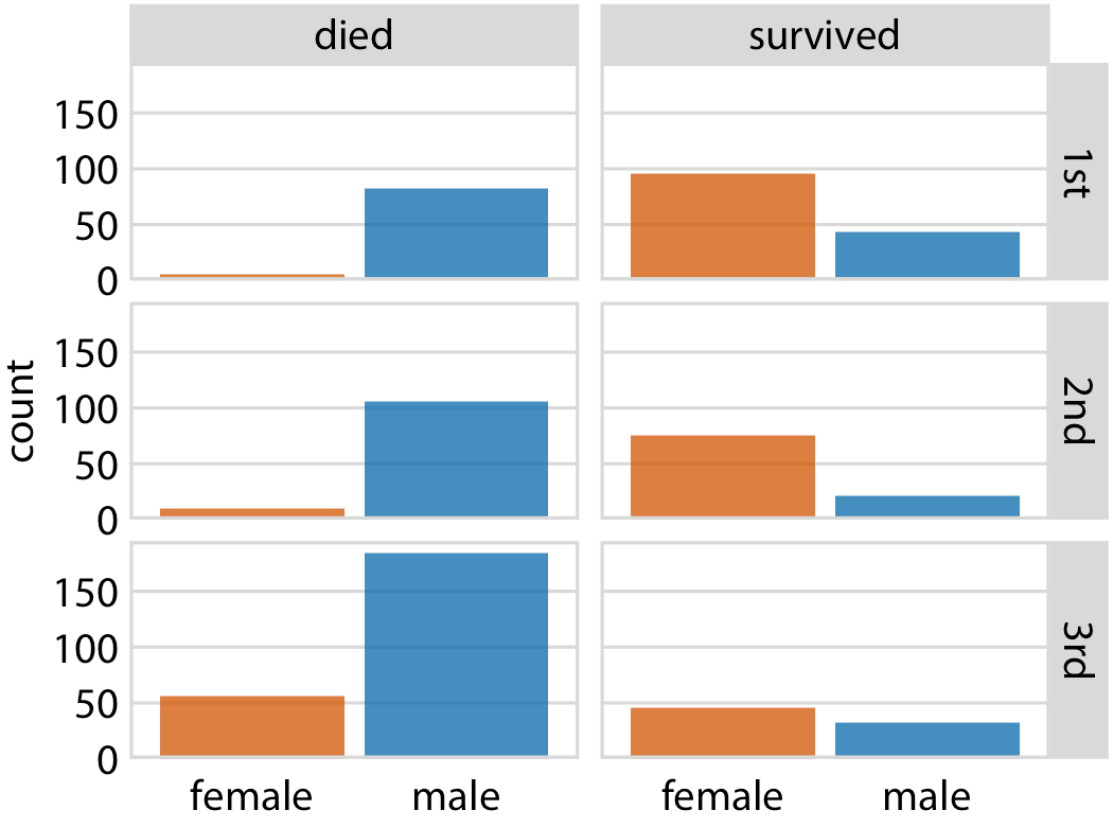
Multi-panel figures

Scenarios

- When datasets become large and complex, they often contain much more information than can reasonably be shown in a single figure panel.
- It can be helpful to create multi-panel figures.
- Compound figures
  - Separate figure panels assembled in an arbitrary arrangement (which may or may not be grid-based).
  - Show entirely different visualizations, or possibly even different datasets.

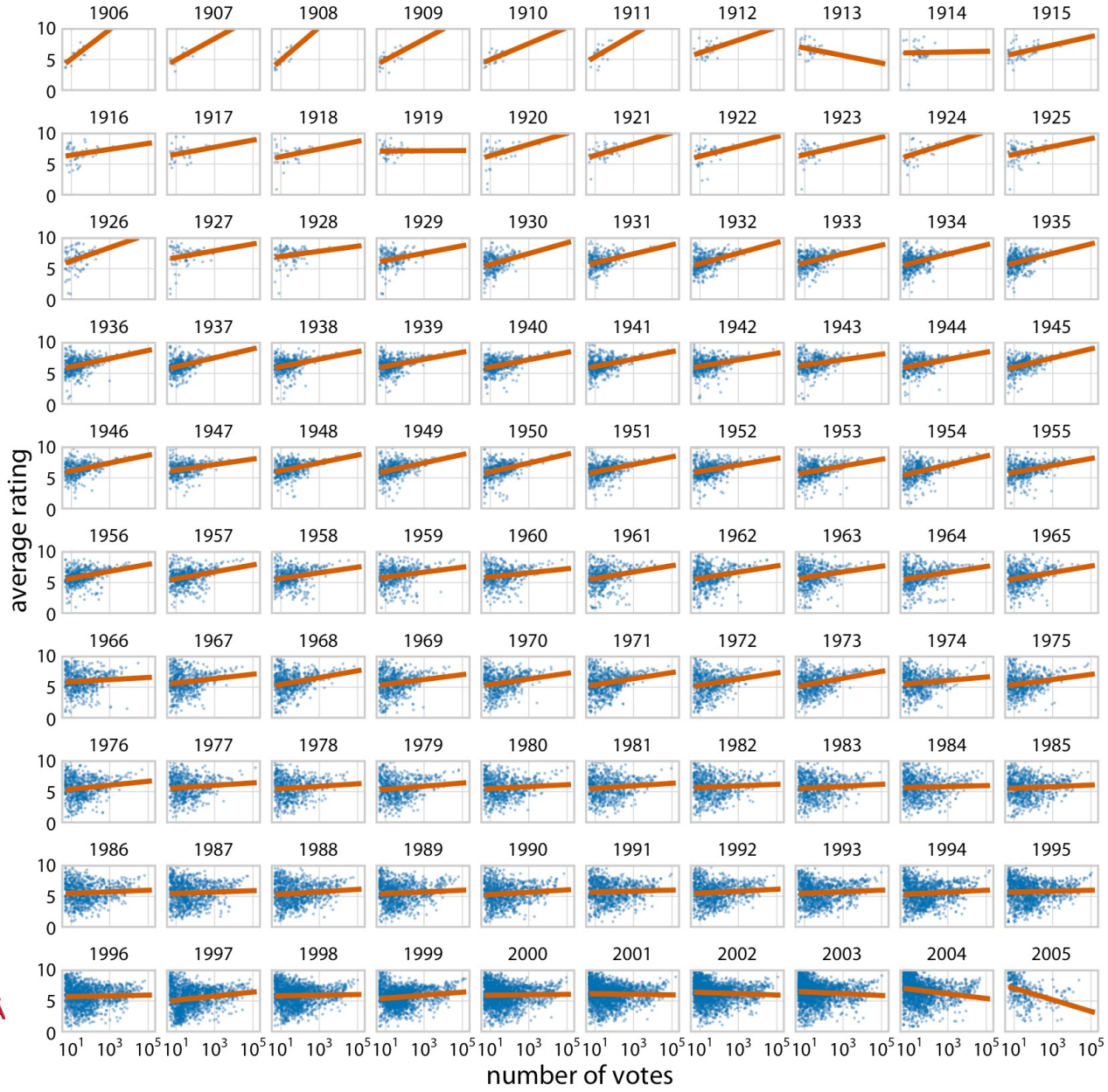
Small multiples

- Small multiples
  - Plots consisting of multiple panels arranged in a regular grid.
  - Each panel shows a different subset of the data, but all panels use the same type of visualization.
- Sometimes referred to as “faceting”.
- Example
  - Breakdown of passengers on the Titanic by gender, survival, and class in which they traveled (1st, 2nd, or 3rd).

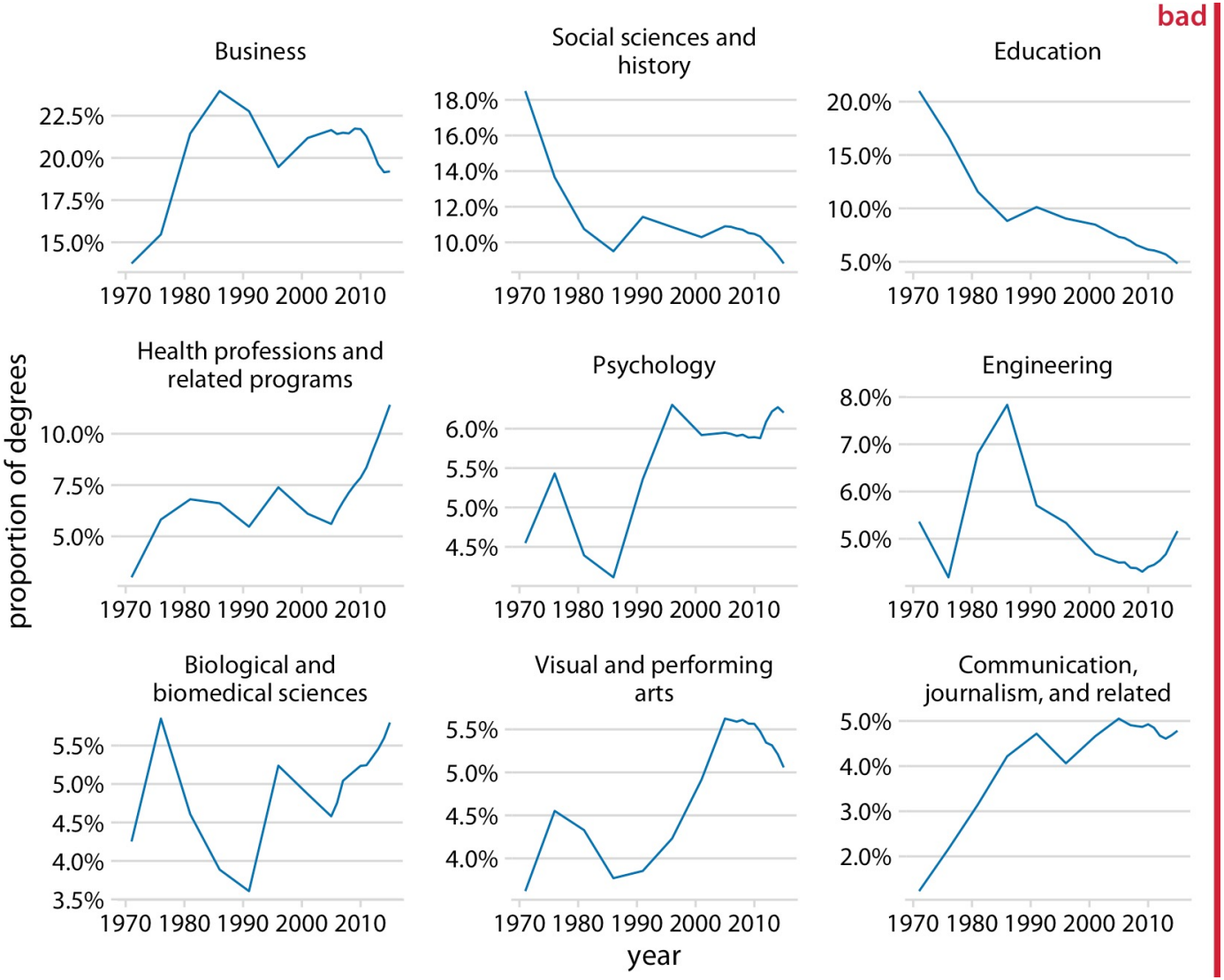


Example: Average movie rankings versus number of votes

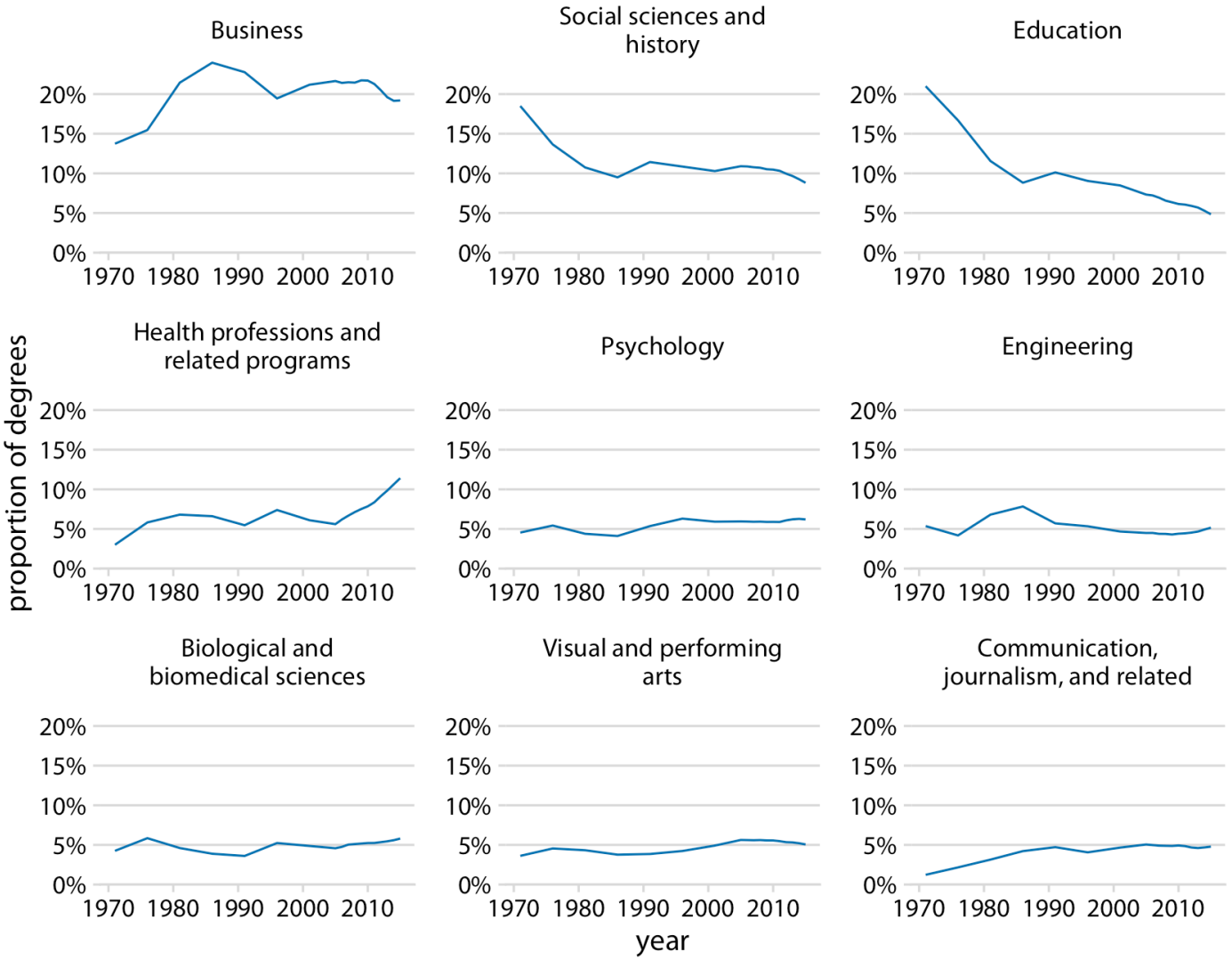
- It is important that each panel uses the same axis ranges and scaling



Example: Trends in bachelor's degrees conferred by US institutions of higher learning

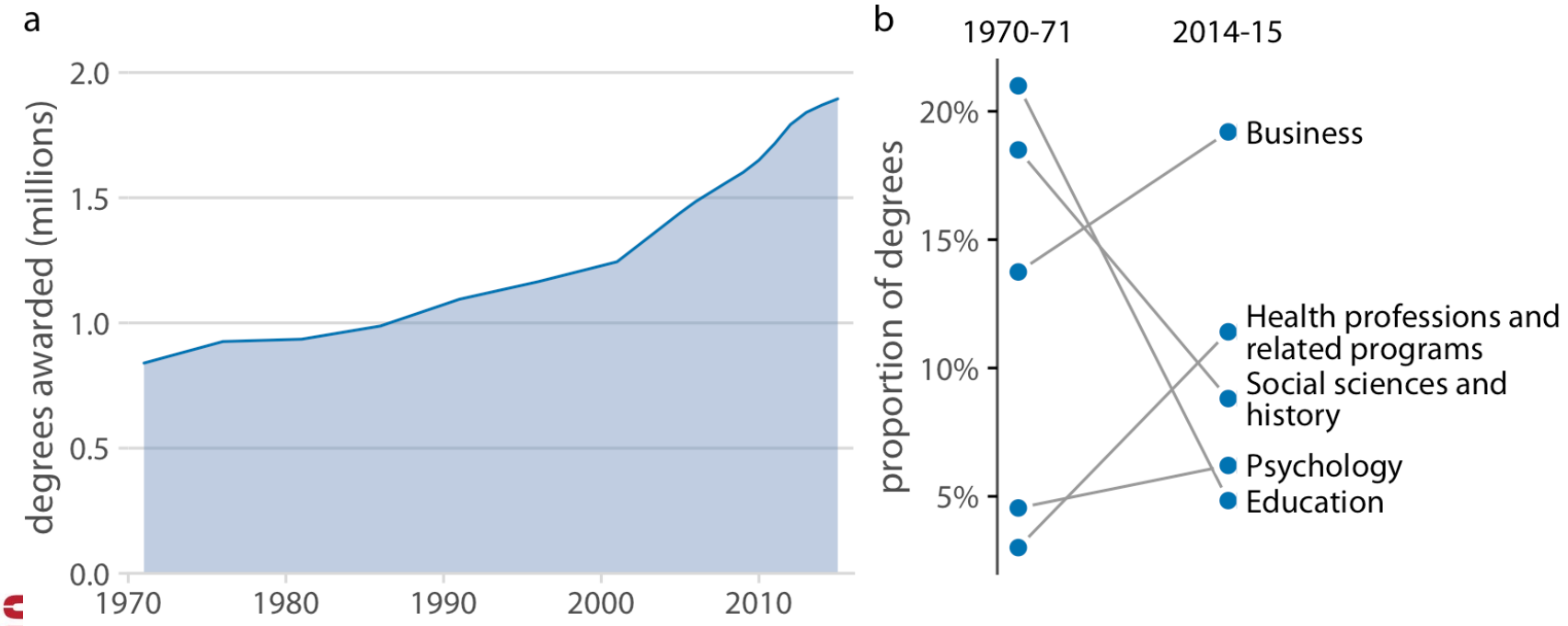


Example: Trends in bachelor's degrees conferred by US institutions of higher learning



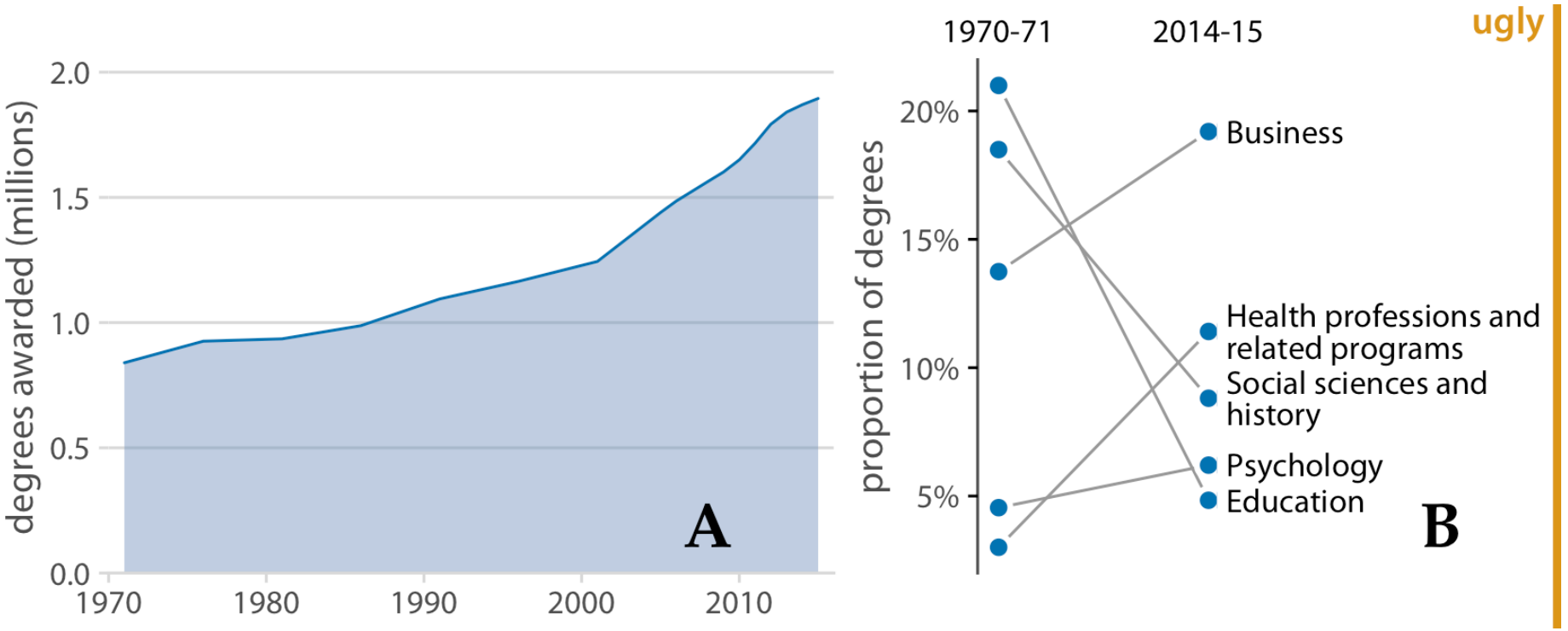
Compound figures

- Combine several independent panels into a figure that conveys one overarching point.
- Individual panels of the compound figure are labeled alphabetically
- Example
  - Trends in bachelor's degrees conferred by US institutions of higher learning



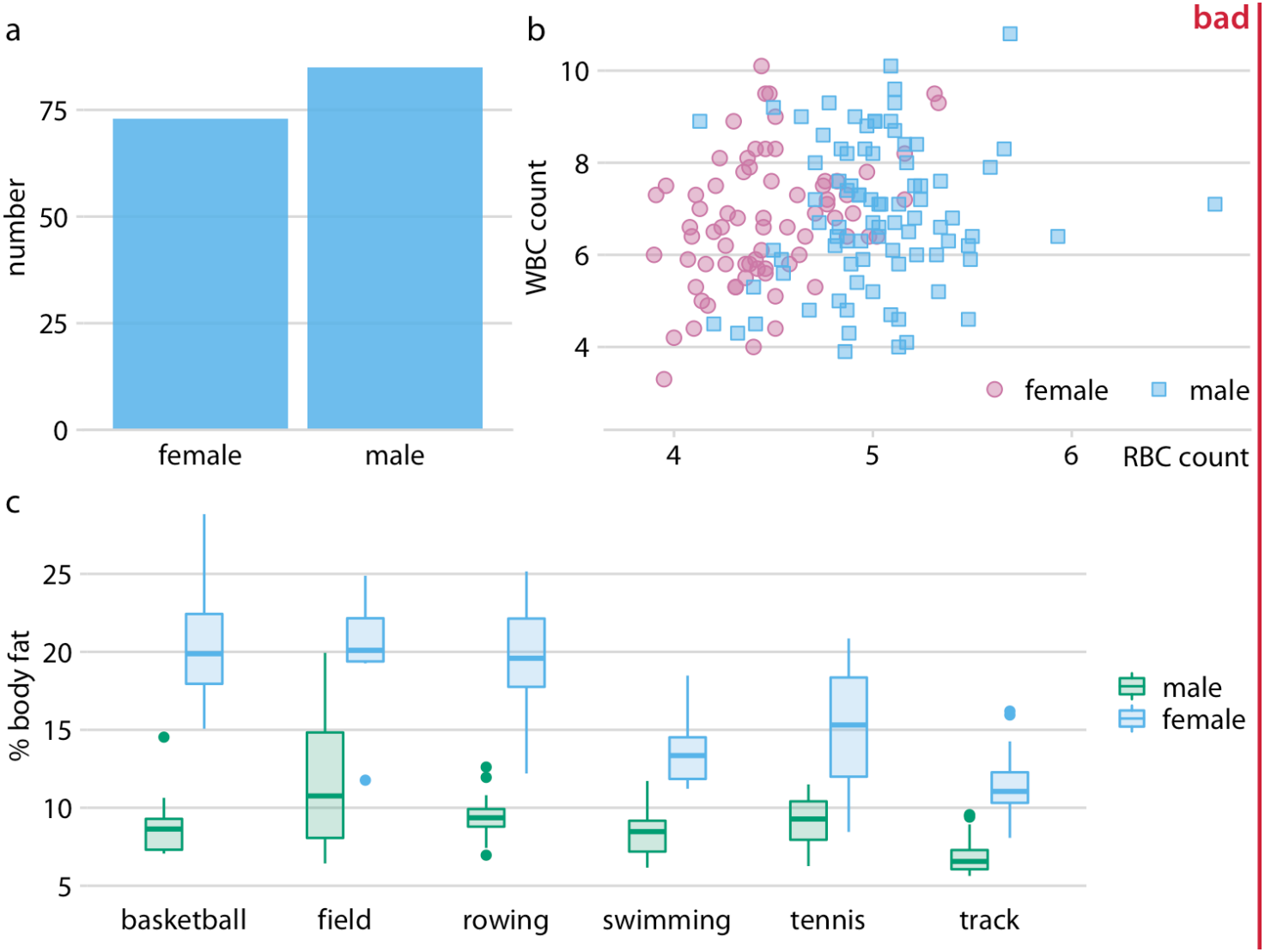
Example: Trends in bachelor's degrees conferred by US institutions of higher learning

- Poor labeling
  - The panel labels are too large and thick, they are in the wrong font, and they are placed in an awkward location

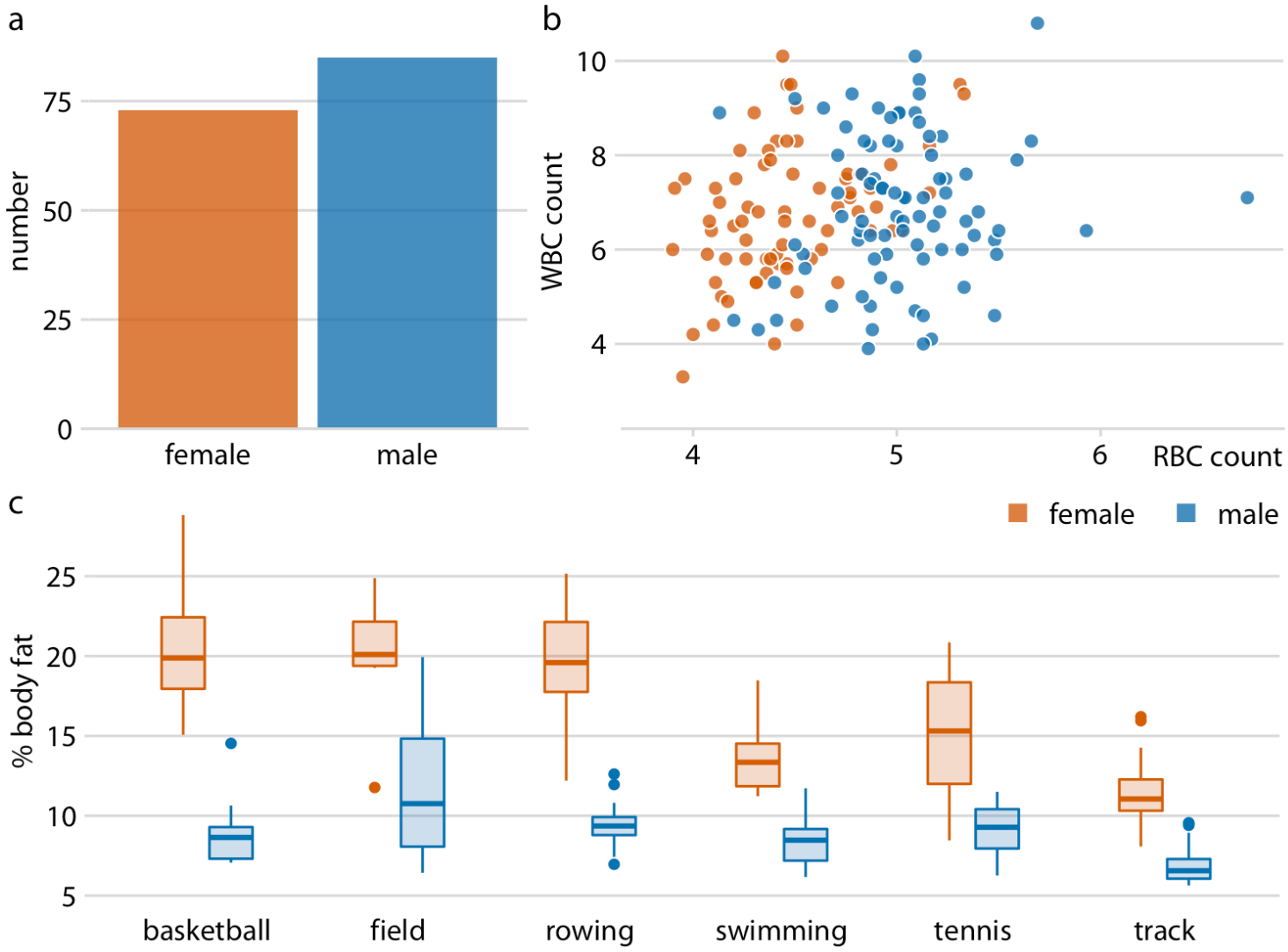


Example: Physiology and body composition of male and female athletes

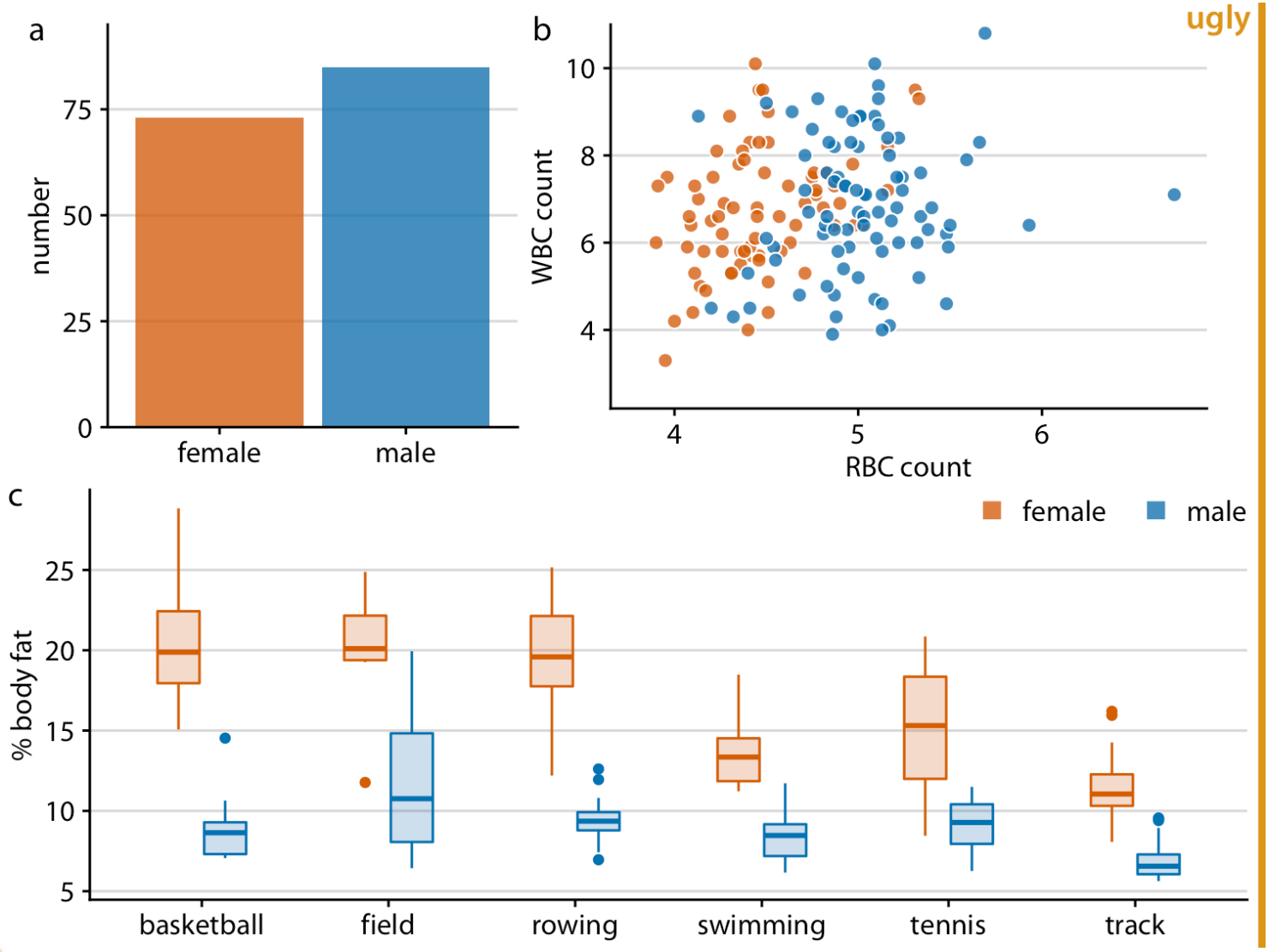
- (a) shows the number of men and women in the dataset.
- (b) shows the counts of red and white blood cells for men and women.
- (c) shows the body fat percentages of men and women, broken down by sport.



Example: Physiology and body composition of male and female athletes



Example: Physiology and body composition of male and female athletes



Titles, captions, and tables

Data is placed into context by accompanying titles, captions, and other annotations

Figure titles and captions

- The job of the title is to accurately convey to the reader what the figure is about, what point it makes.

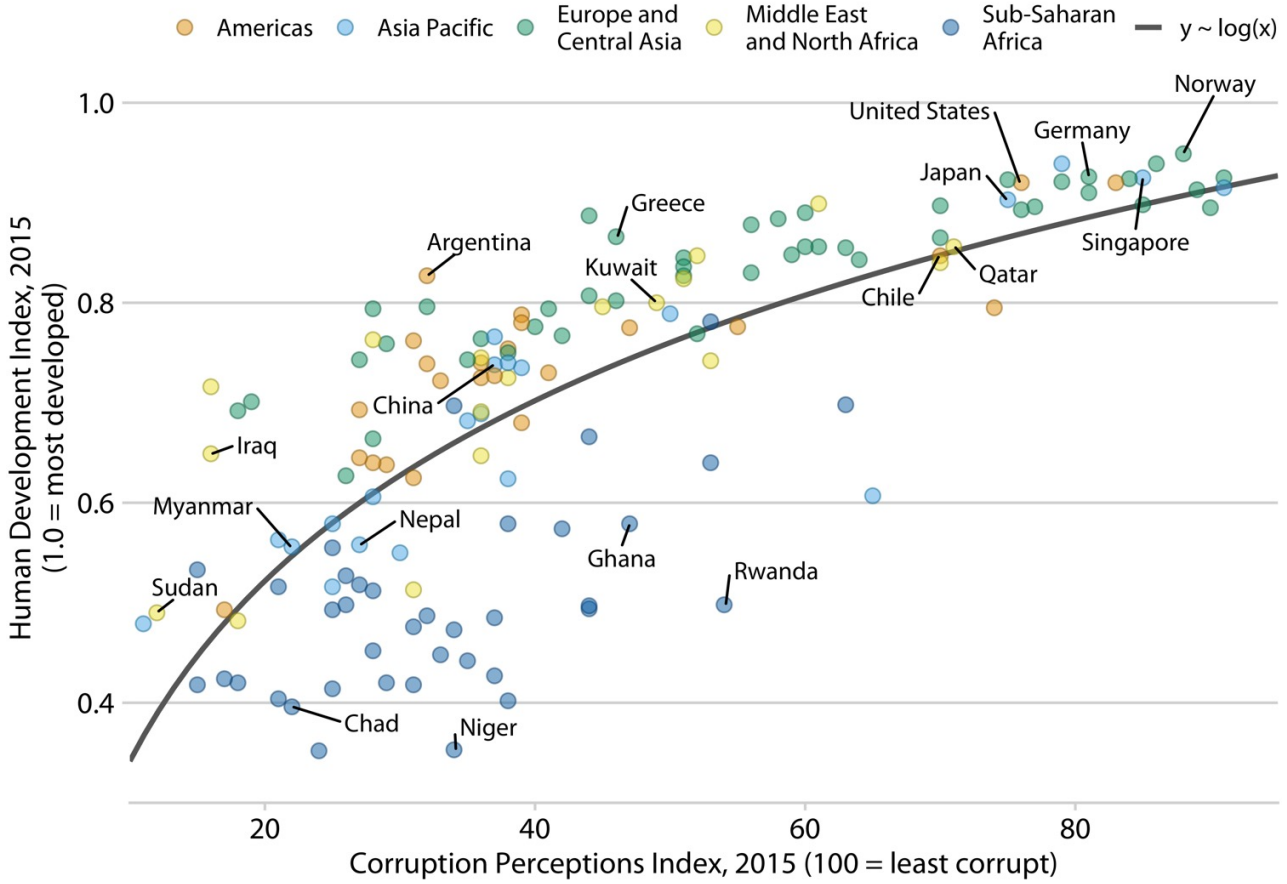
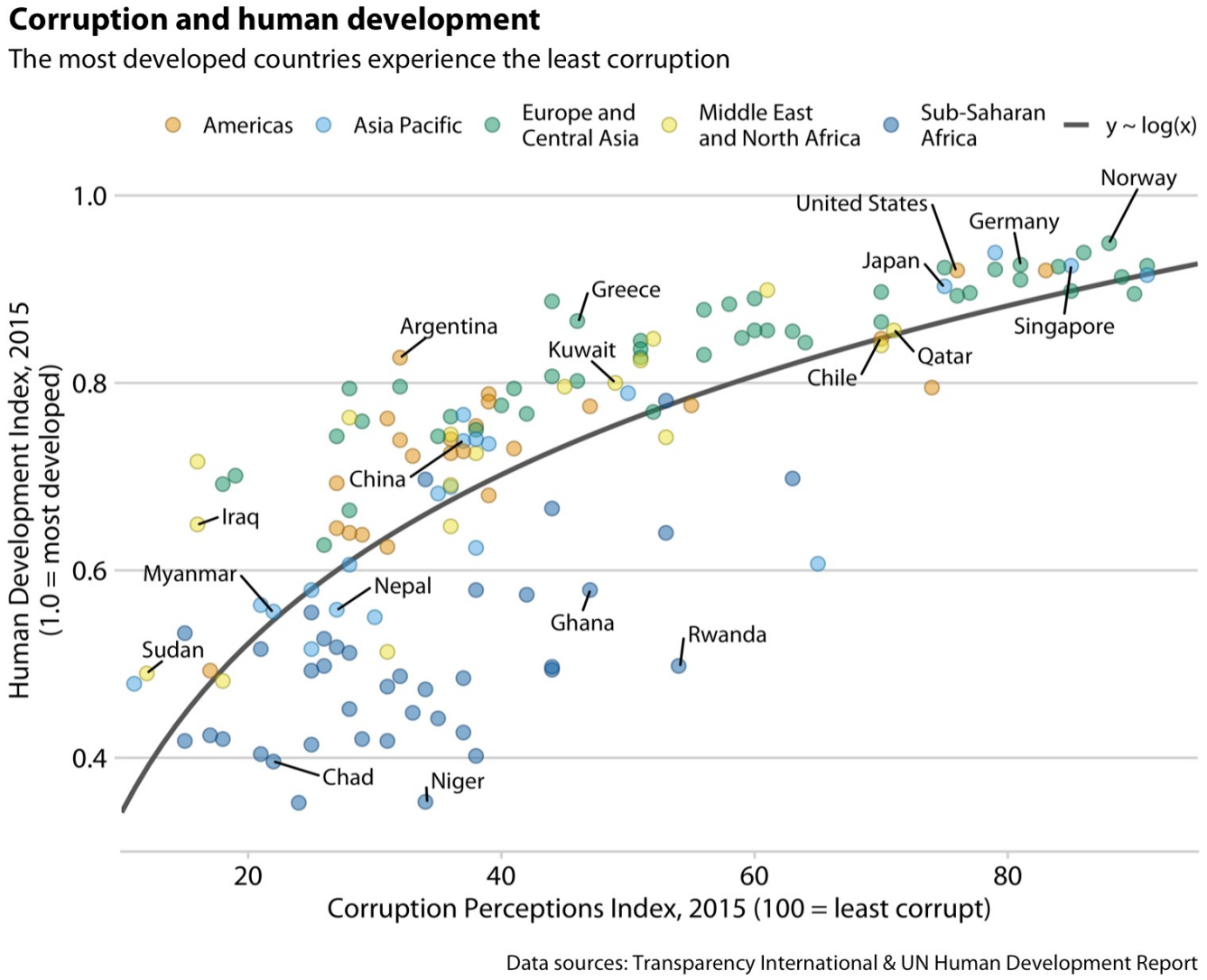
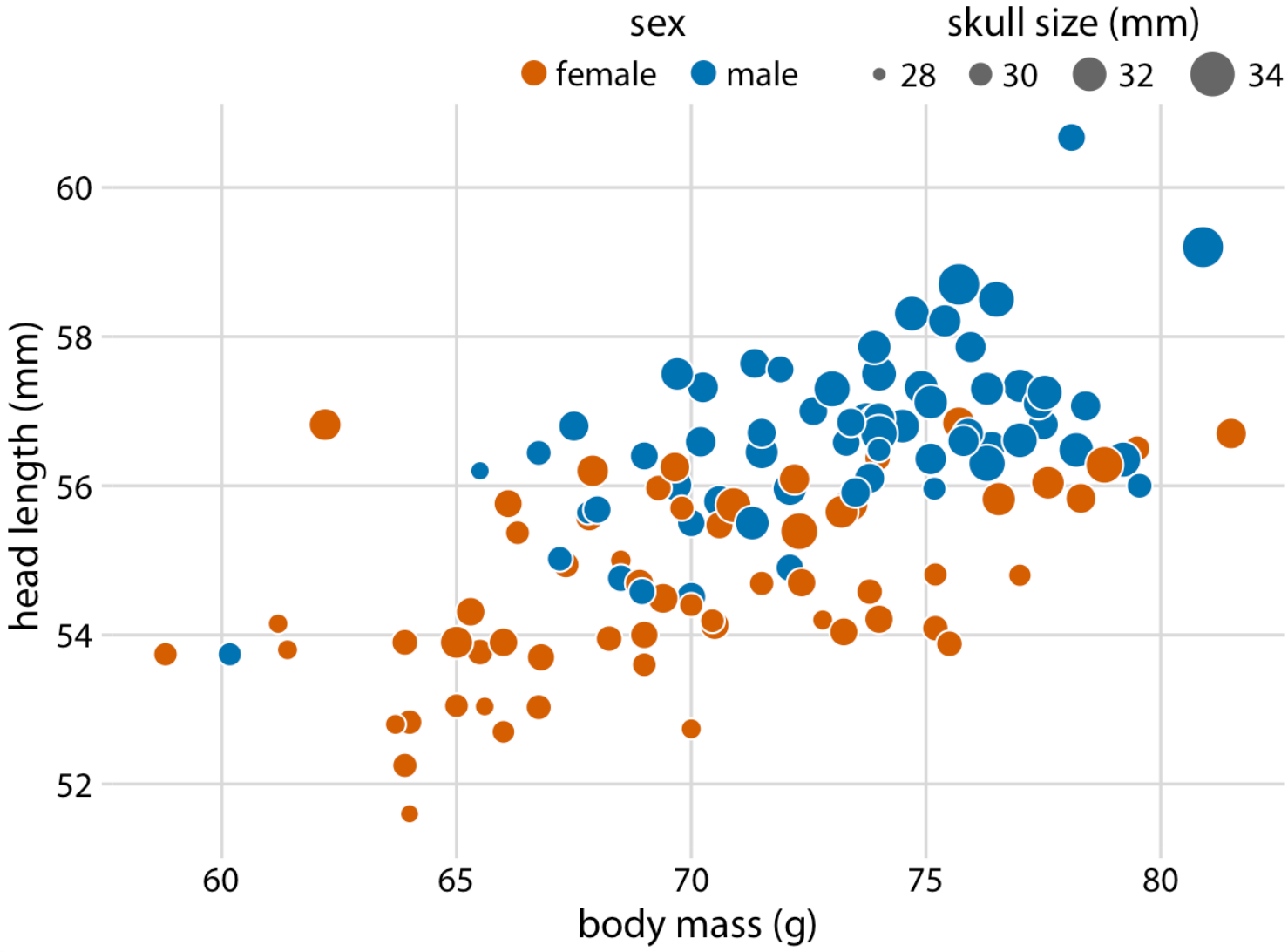


Figure 22-1. Corruption and human development: the most developed countries experience the least corruption. Original figure concept: [The Economist online 2011]. Data sources: Transparency International & UN Human Development Report.

The title, subtitle, and data source statements have been incorporated into the figure

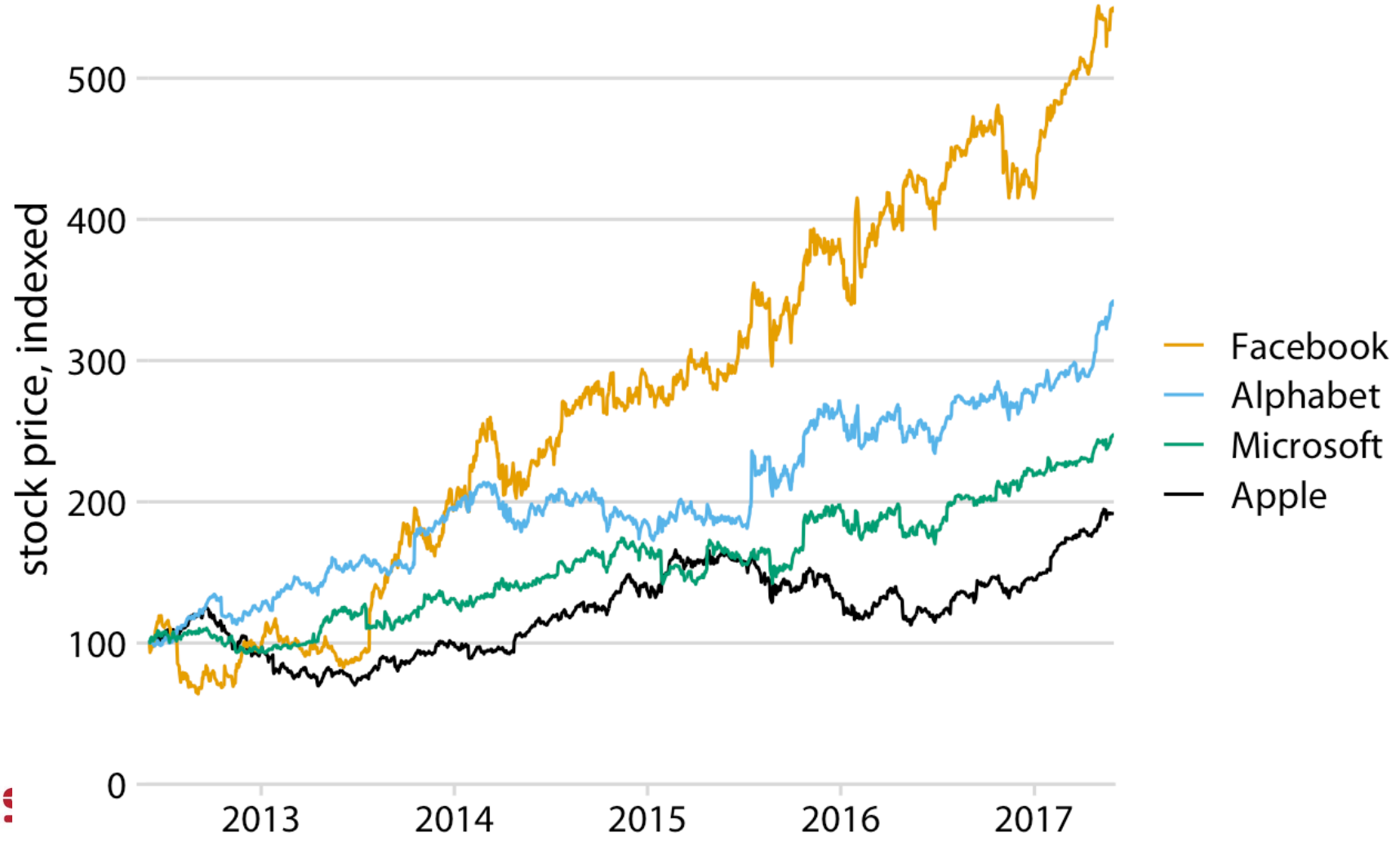


Example: Head length versus body mass for 123 blue jays



Example: Stock price over time for four major tech companies

- Axis or legend titles can be omitted, namely when the labels themselves are fully explanatory.



Example: Stock price over time for four major tech companies

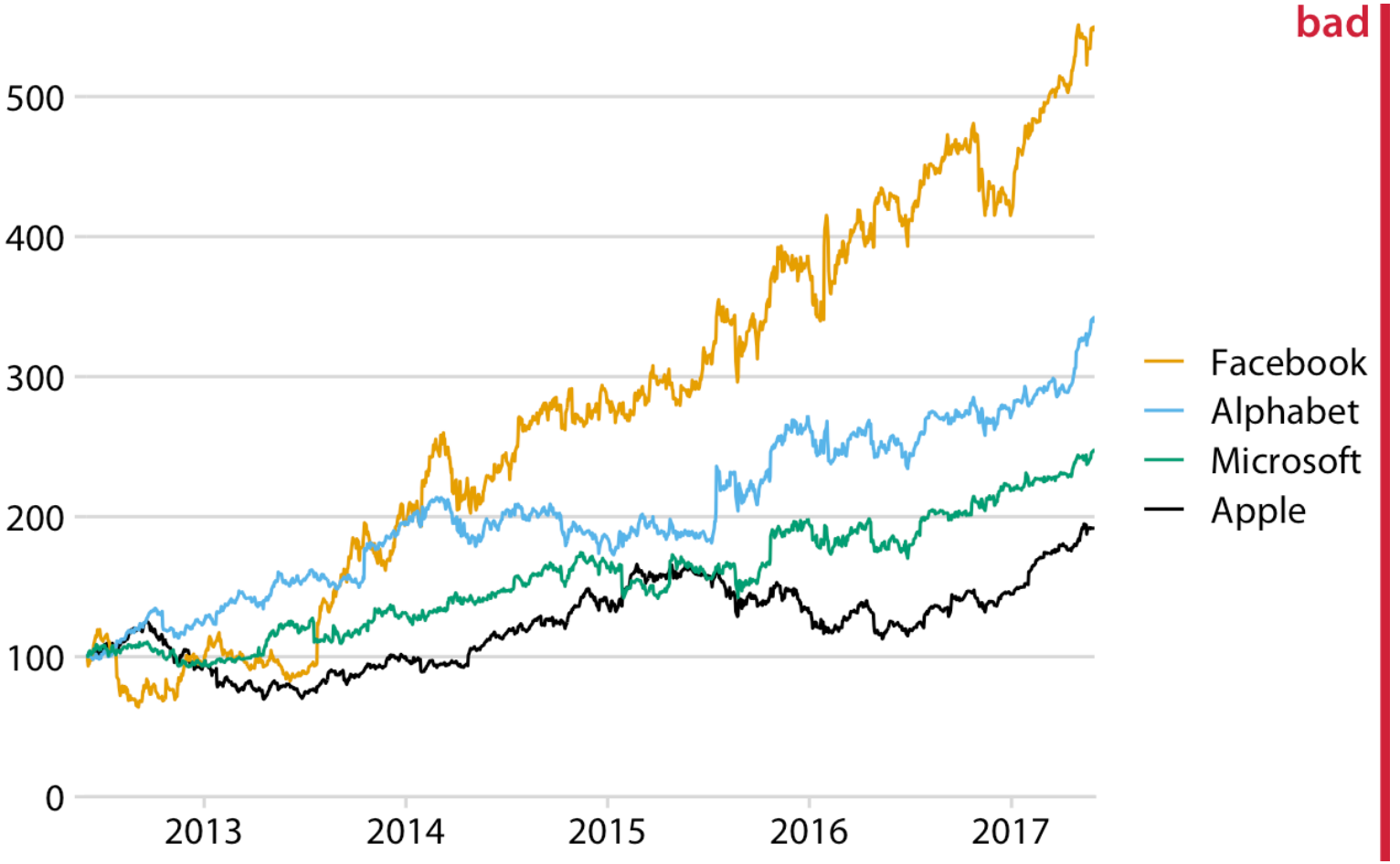
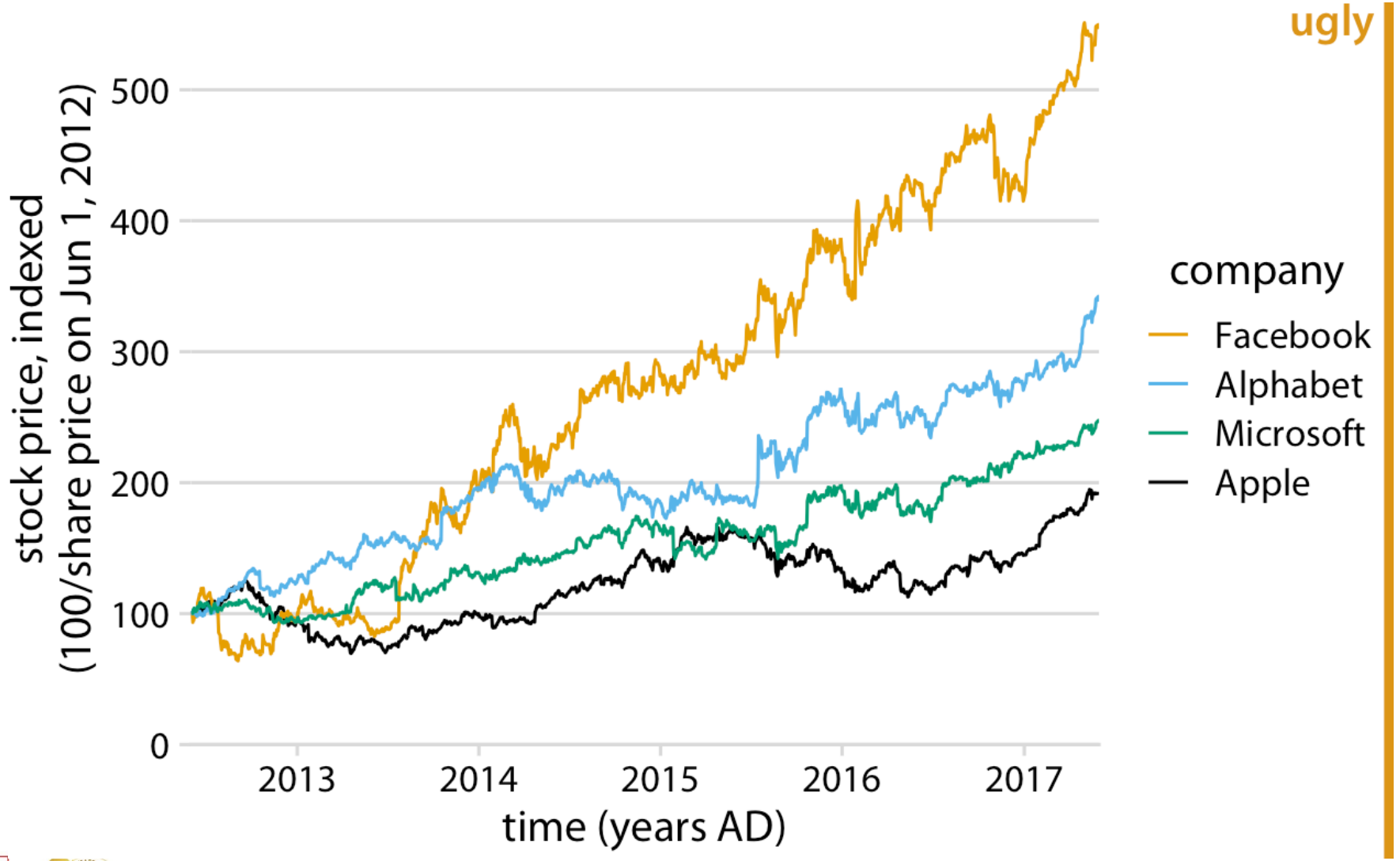


Figure 100: Stock price over time for four major tech companies. The stock price for each company has been normalized to equal 100 in June 2012

Example: Stock price over time for four major tech companies



Examples of poorly and appropriately formatted tables

a

ugly

Rank	Title	Amount
1	*Star Wars: The Last Jedi*	$71,565,498
2	*Jumanji: Welcome to the Jungle*	$36,169,328
3	*Pitch Perfect 3*	$19,928,525
4	*The Greatest Showman*	$8,805,843
5	*Ferdinand*	$7,316,746

b

ugly

Rank	Title	Amount
1	*Star Wars: The Last Jedi*	$71,565,498
2	*Jumanji: Welcome to the Jungle*	$36,169,328
3	*Pitch Perfect 3*	$19,928,525
4	*The Greatest Showman*	$8,805,843
5	*Ferdinand*	$7,316,746

c

Rank	Title	Amount
1	*Star Wars: The Last Jedi*	$71,565,498
2	*Jumanji: Welcome to the Jungle*	$36,169,328
3	*Pitch Perfect 3*	$19,928,525
4	*The Greatest Showman*	$8,805,843
5	*Ferdinand*	$7,316,746

d

Rank	Title	Amount
1	*Star Wars: The Last Jedi*	$71,565,498
2	*Jumanji: Welcome to the Jungle*	$36,169,328
3	*Pitch Perfect 3*	$19,928,525
4	*The Greatest Showman*	$8,805,843
5	*Ferdinand*	$7,316,746

Tables

- Key rules for table layouts
  1. Do not use vertical lines.
  2. Do not use horizontal lines between data rows. (Horizontal lines as a separator between the title row and the first data row or as a frame for the entire table are fine.)
  3. Text columns should be left aligned.
  4. Number columns should be right aligned and should use the same number of decimal digits throughout.
  5. Columns containing single characters should be centered.
  6. The header fields should be aligned with their data; i.e., the heading for a text column will be left aligned and the heading for a number column will be right aligned.

Balance the data and the context

Scenarios

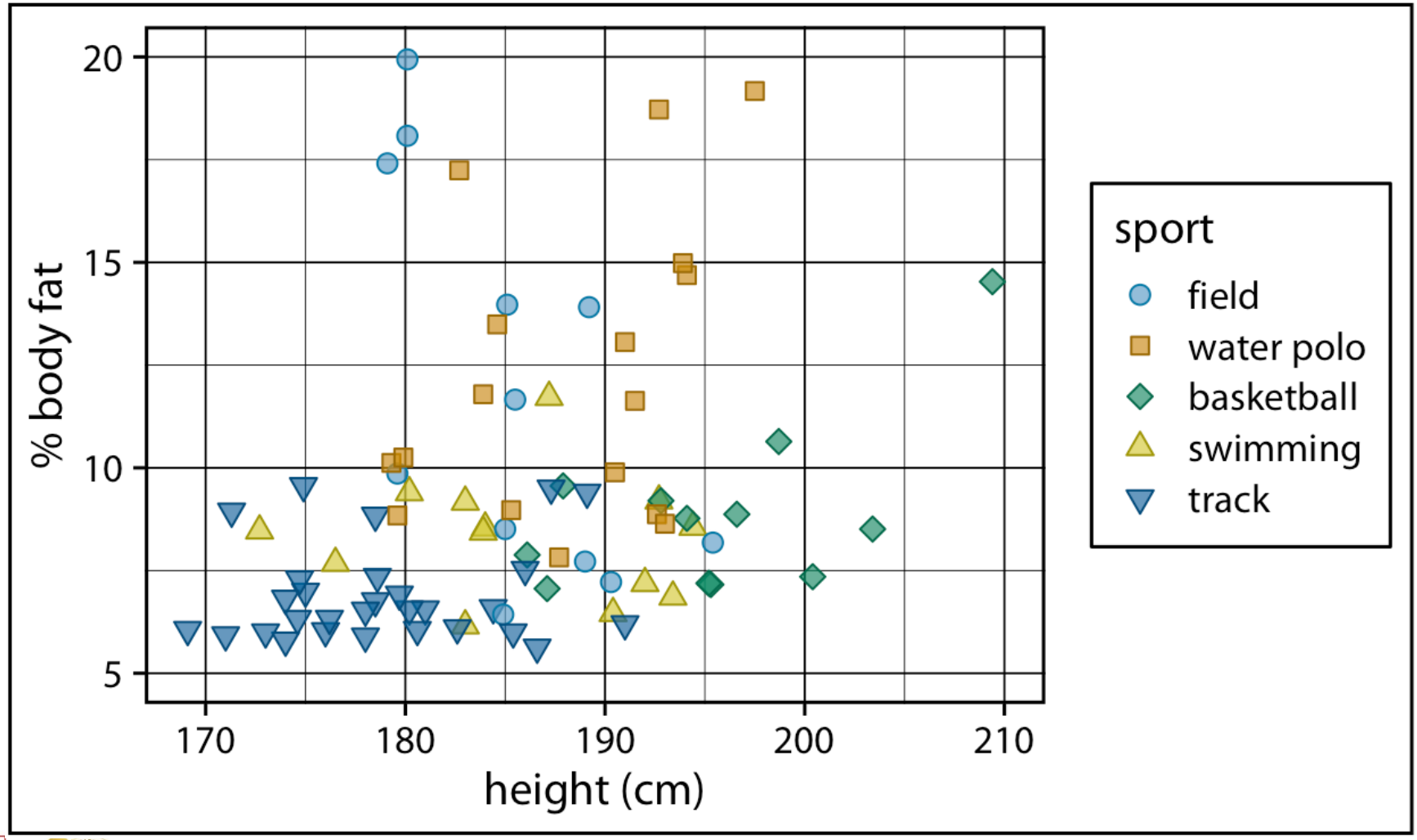
- The graphical elements in any visualization can be broadly subdivided into elements that represent data and elements that do not.
- Example
  - Points in a scatterplot, the bars in a histogram or bar plot, or the shaded areas in a heatmap.
  - Plot axes, axis ticks and labels, axis titles, legends, and plot annotations.

Providing the appropriate amount of context

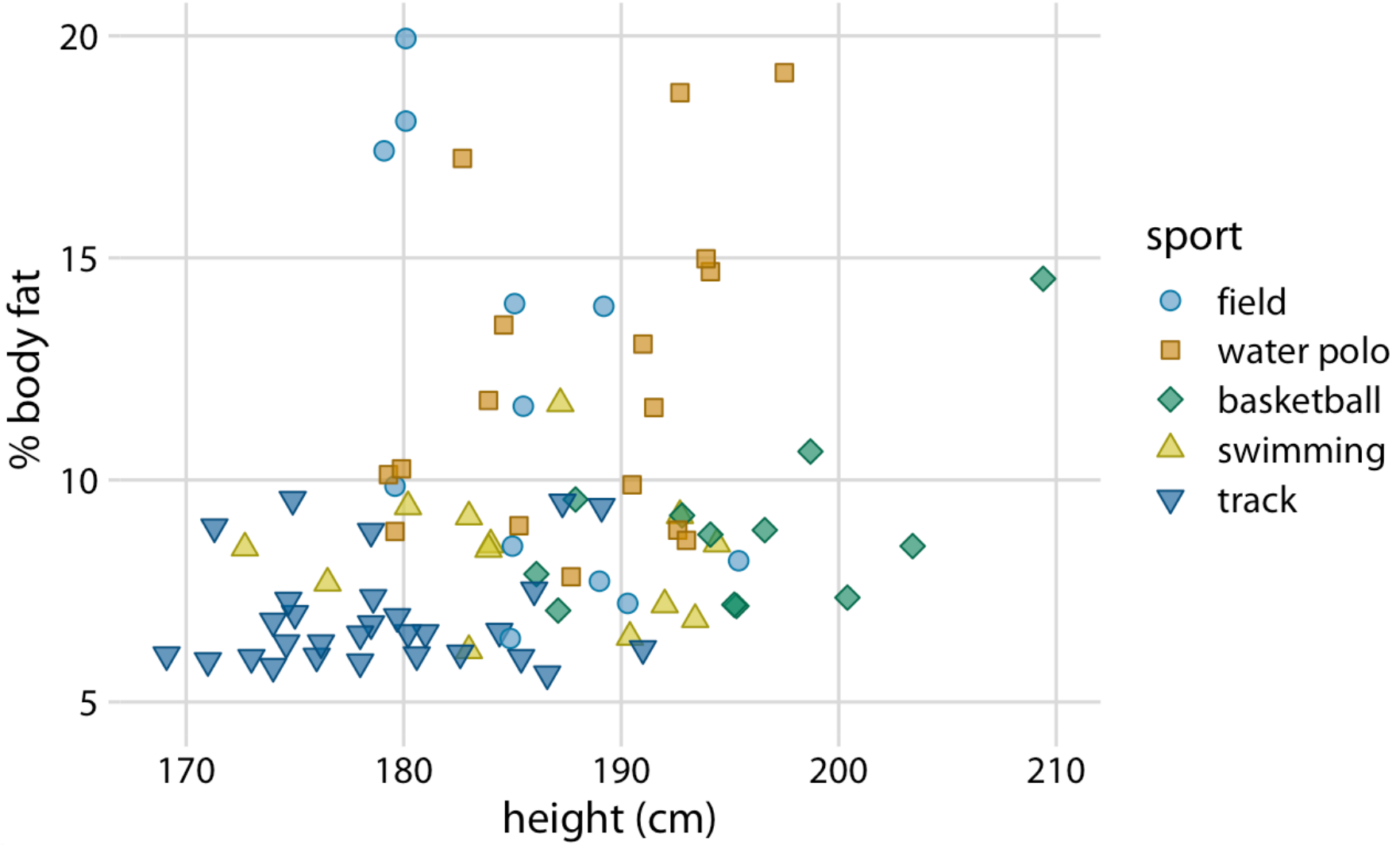
- “Data–ink ratio” is defined as the “proportion of a graphic’s ink devoted to the non-redundant display of data information.” (Edward Tufte)
- Maximize the data–ink ratio, within reason.

Example: Percent body fat versus height in professional male Australian athletes

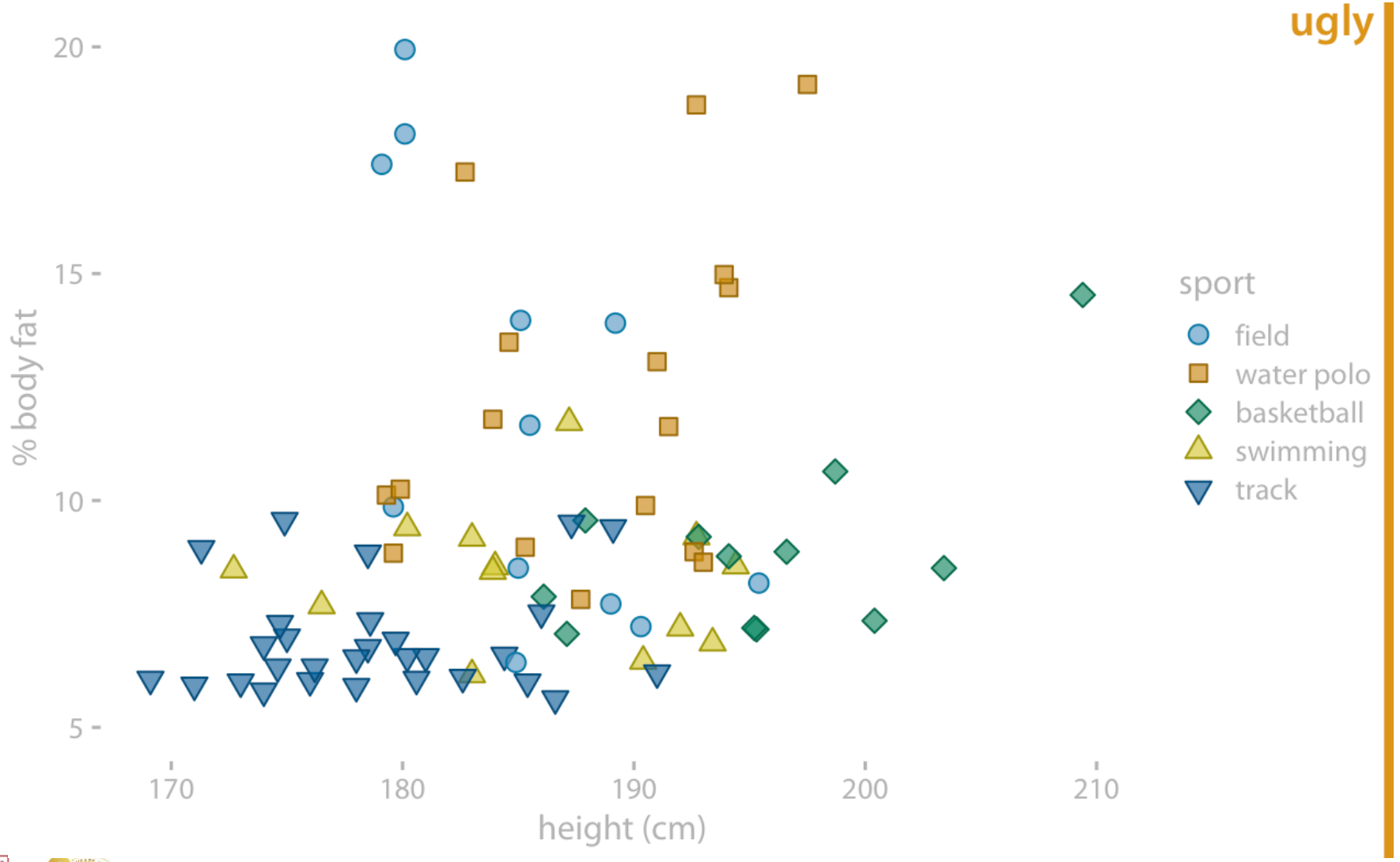
ugly



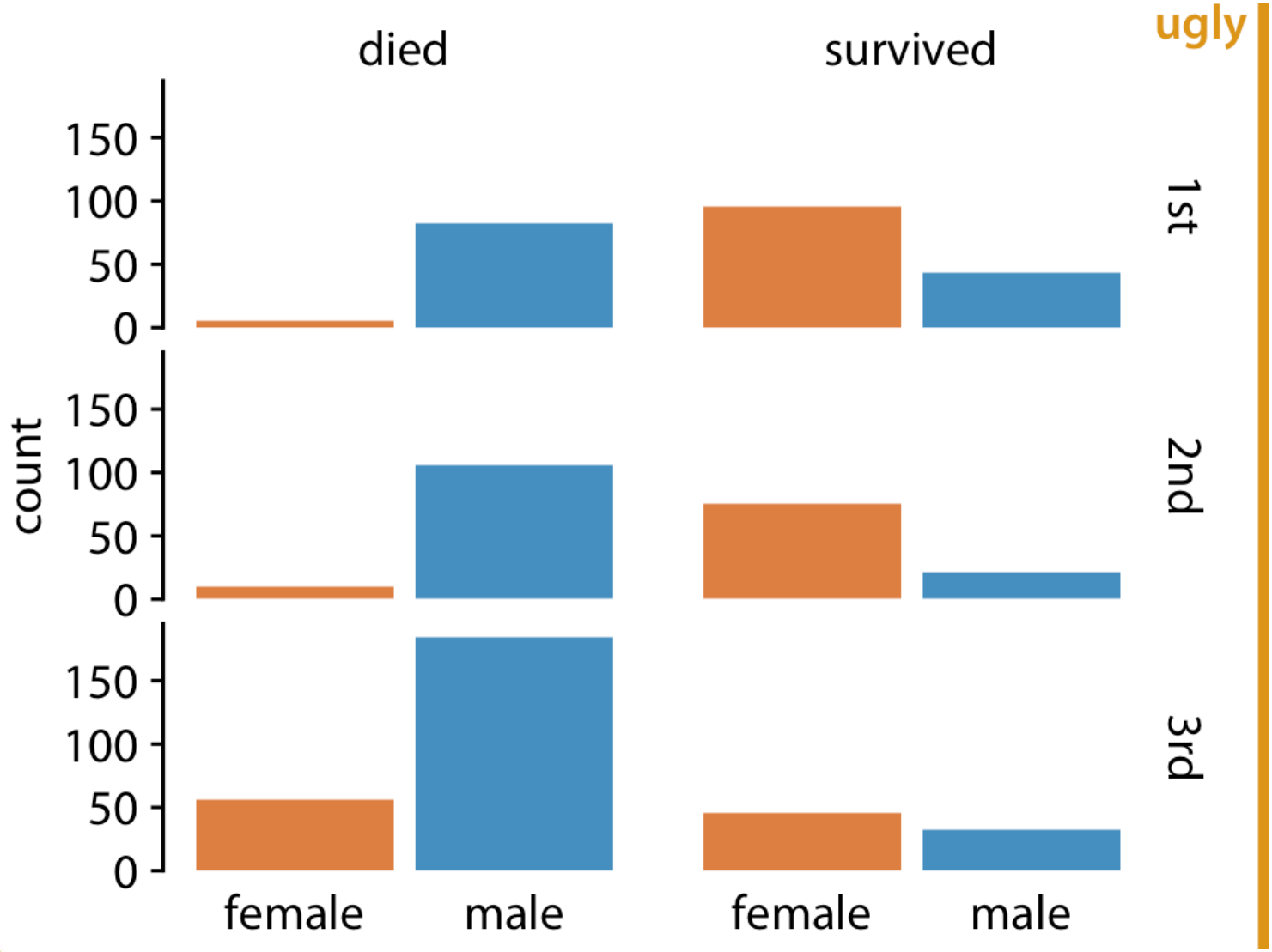
Example: Percent body fat versus height in professional male Australian athletes



Example: Percent body fat versus height in professional male Australian athletes



Example: Survival of passengers on the Titanic, broken down by gender and class

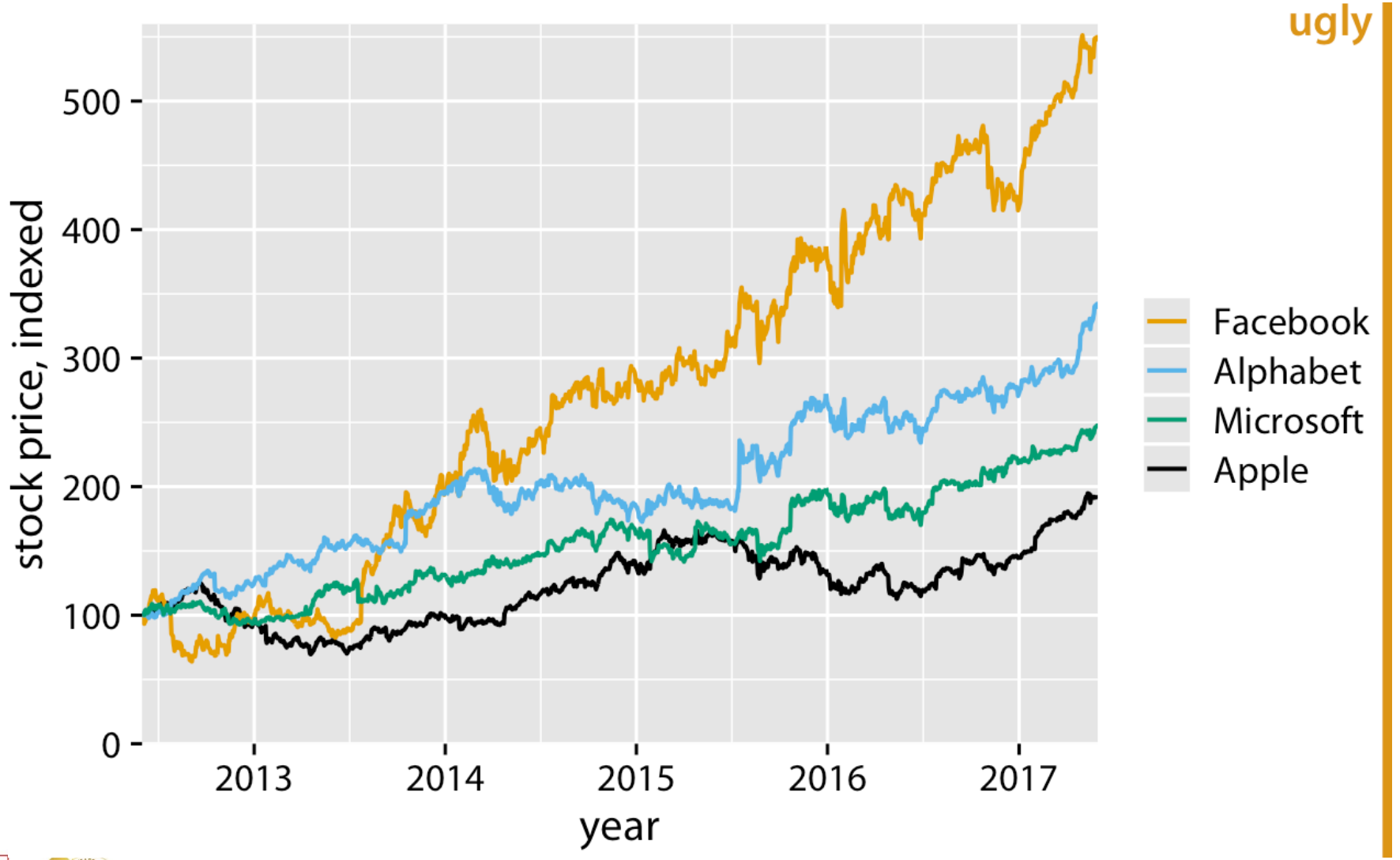


Example: Survival of passengers on the Titanic, broken down by gender and class

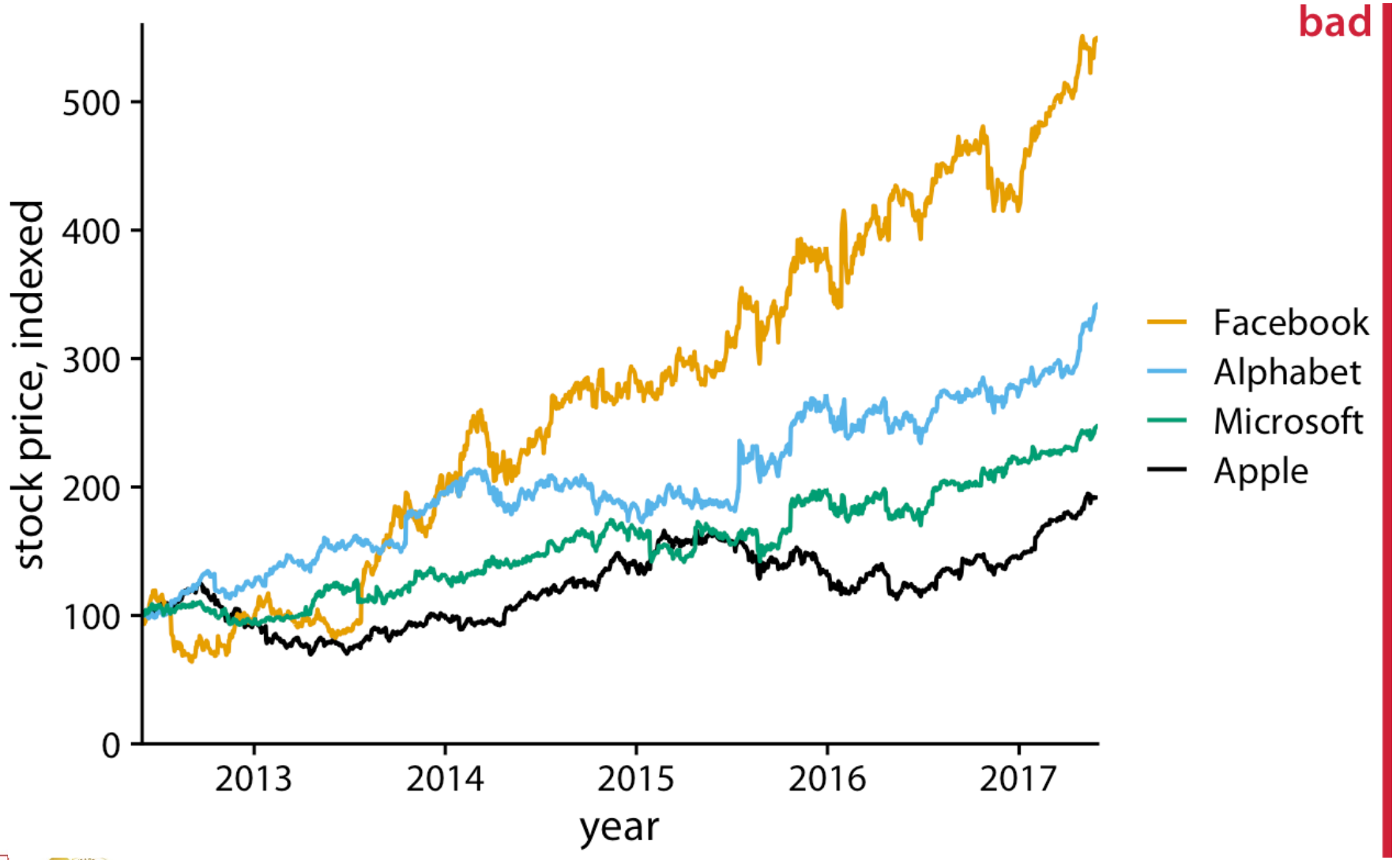
Background grids

- Grid lines in the background of a plot can help the reader discern specific data values and compare values in one part of a plot to values in another part.
- Grid lines can add visual noise, when they are prominent or densely spaced.

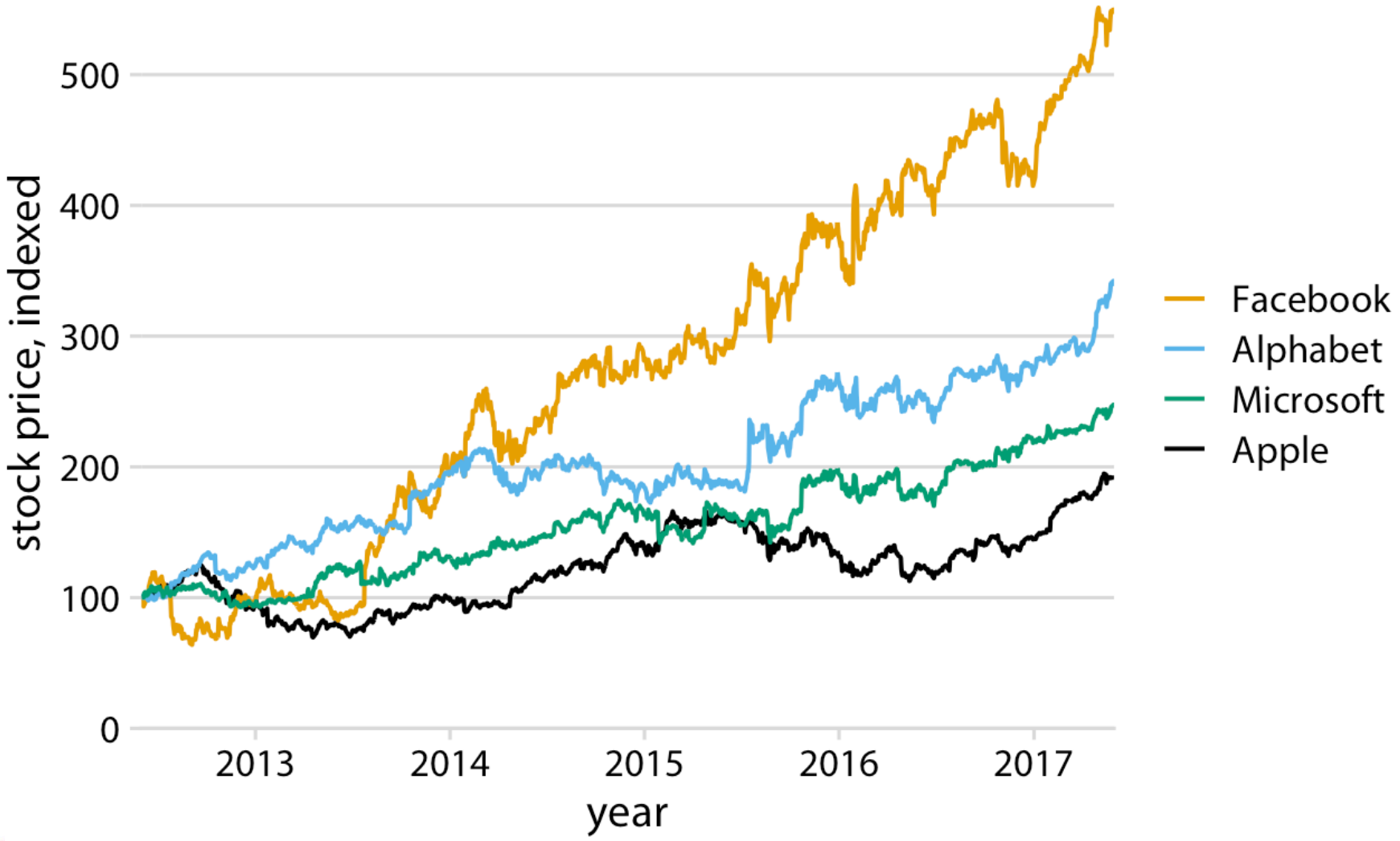
Example: Stock price over time for four major tech companies



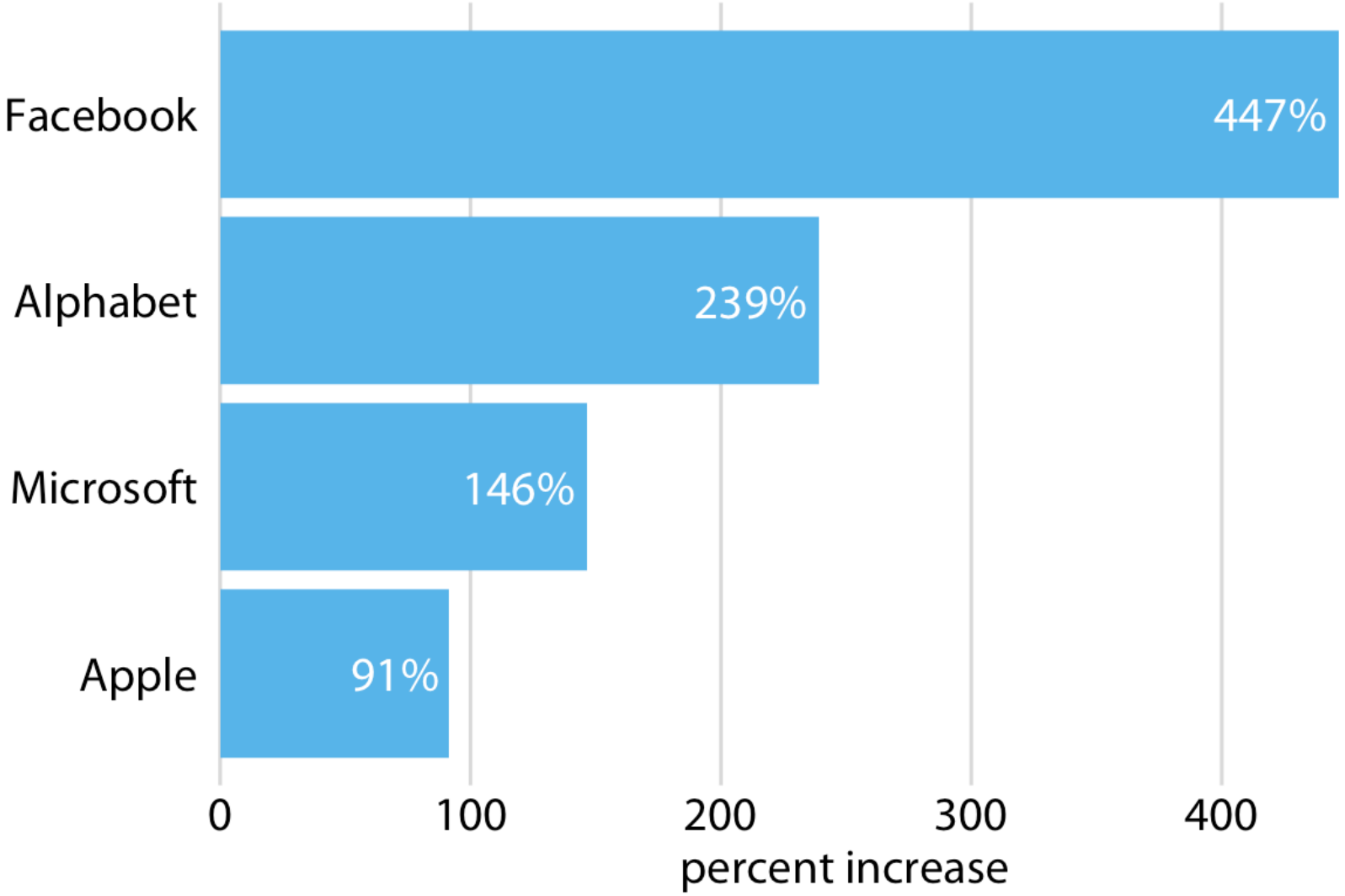
Example: Indexed stock price over time for four major tech companies



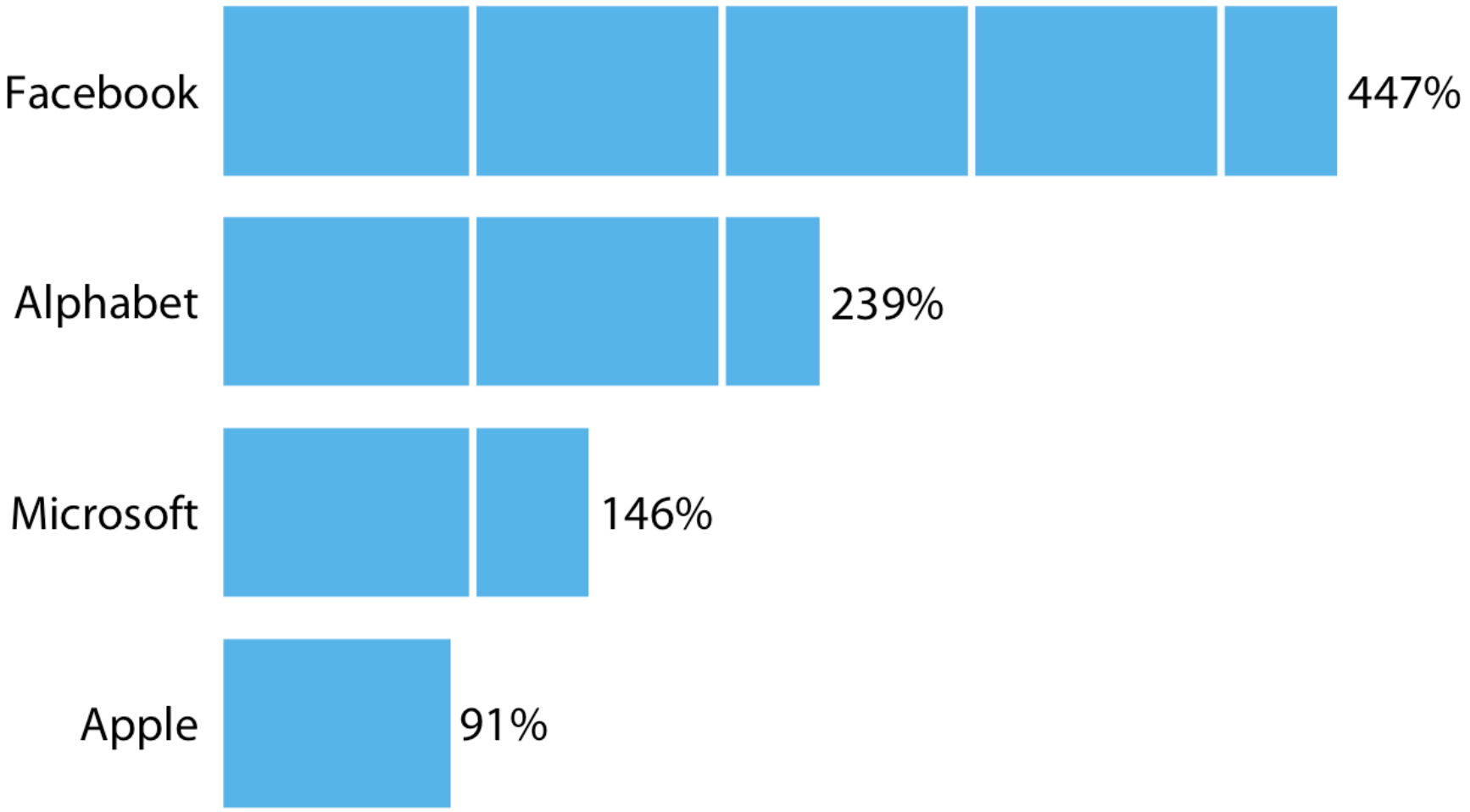
Example: Indexed stock price over time for four major tech companies



Example: Percent increase in stock price from June 2012 to June 2017

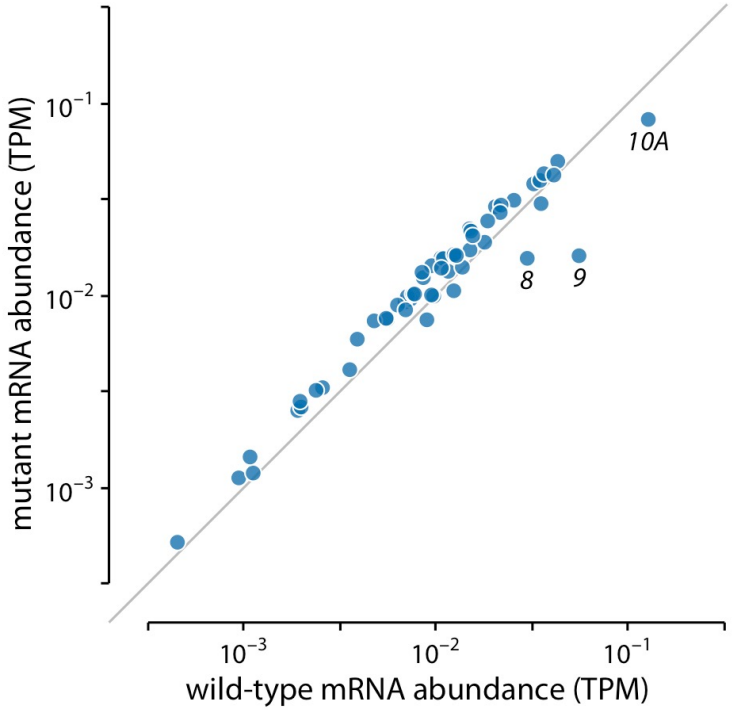


Example: Percent increase in stock price from June 2012 to June 2017

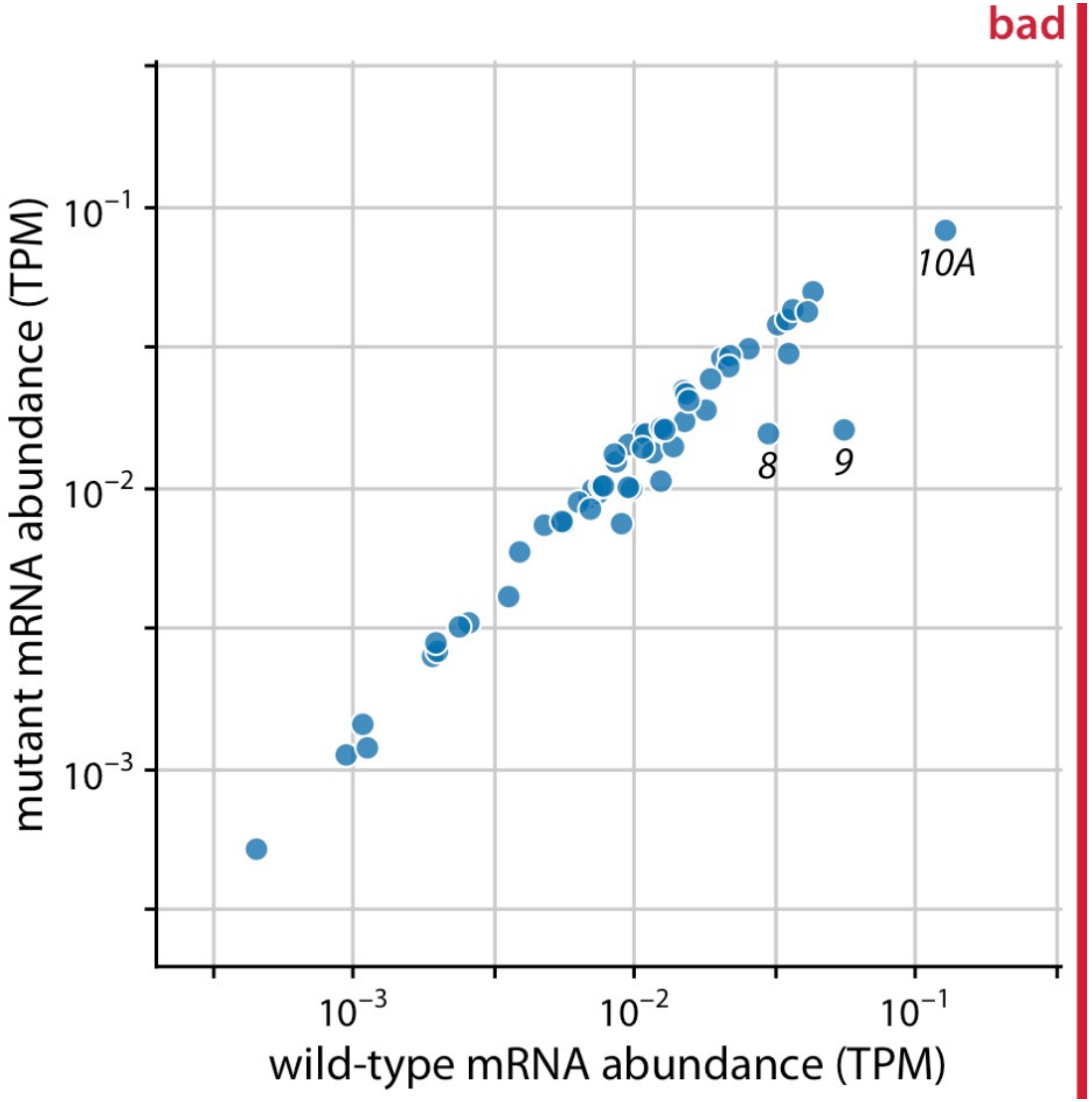


Paired data

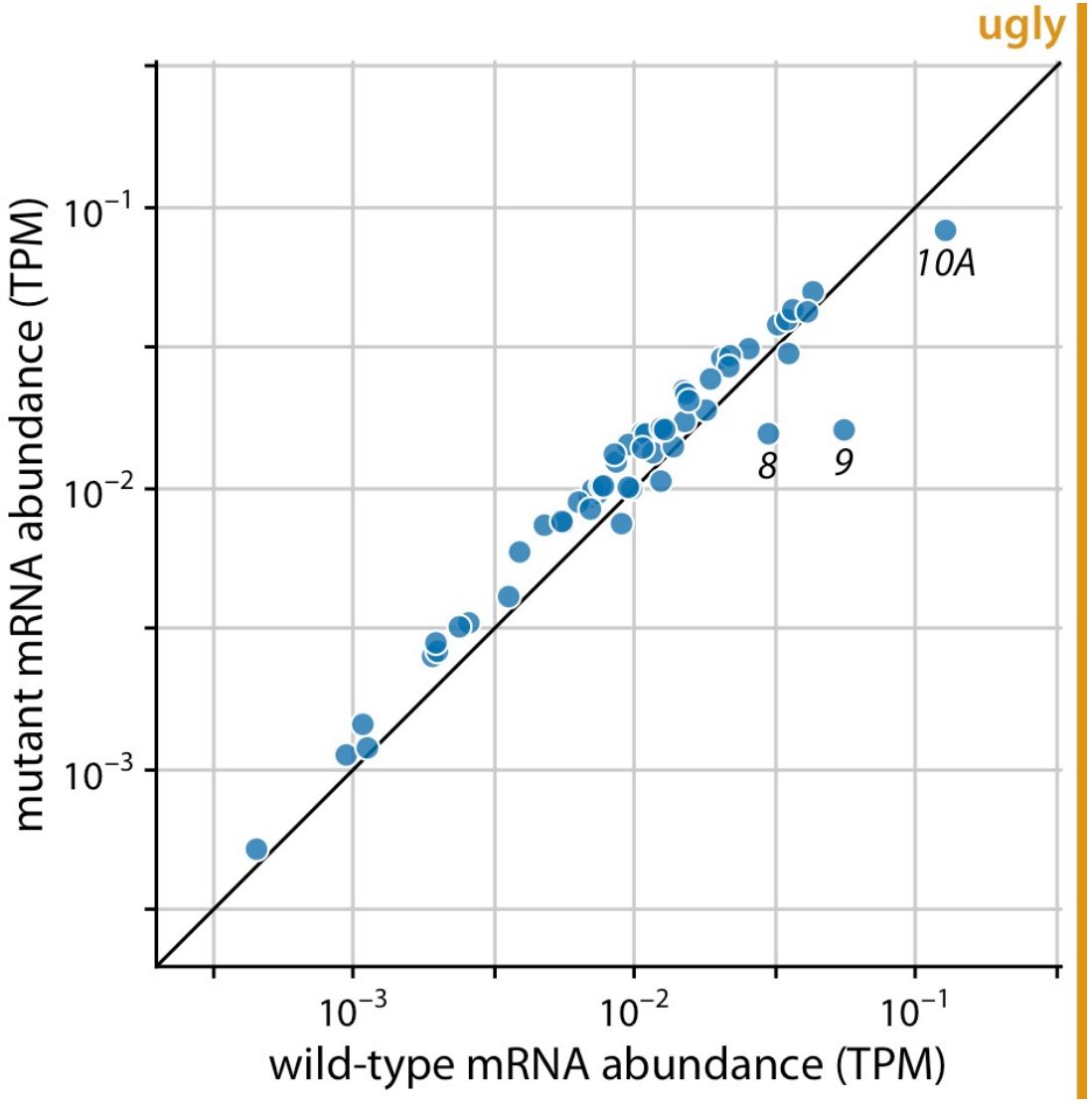
- Where the relevant comparison is the x = y line, such as in scatterplots of paired data.
  - Draw a diagonal line rather than a grid.
- Example
  - Gene expression levels in a mutant bacteriophage T7 relative to wild type.



Example: Gene expression levels in a mutant bacteriophage T7 relative to wild type

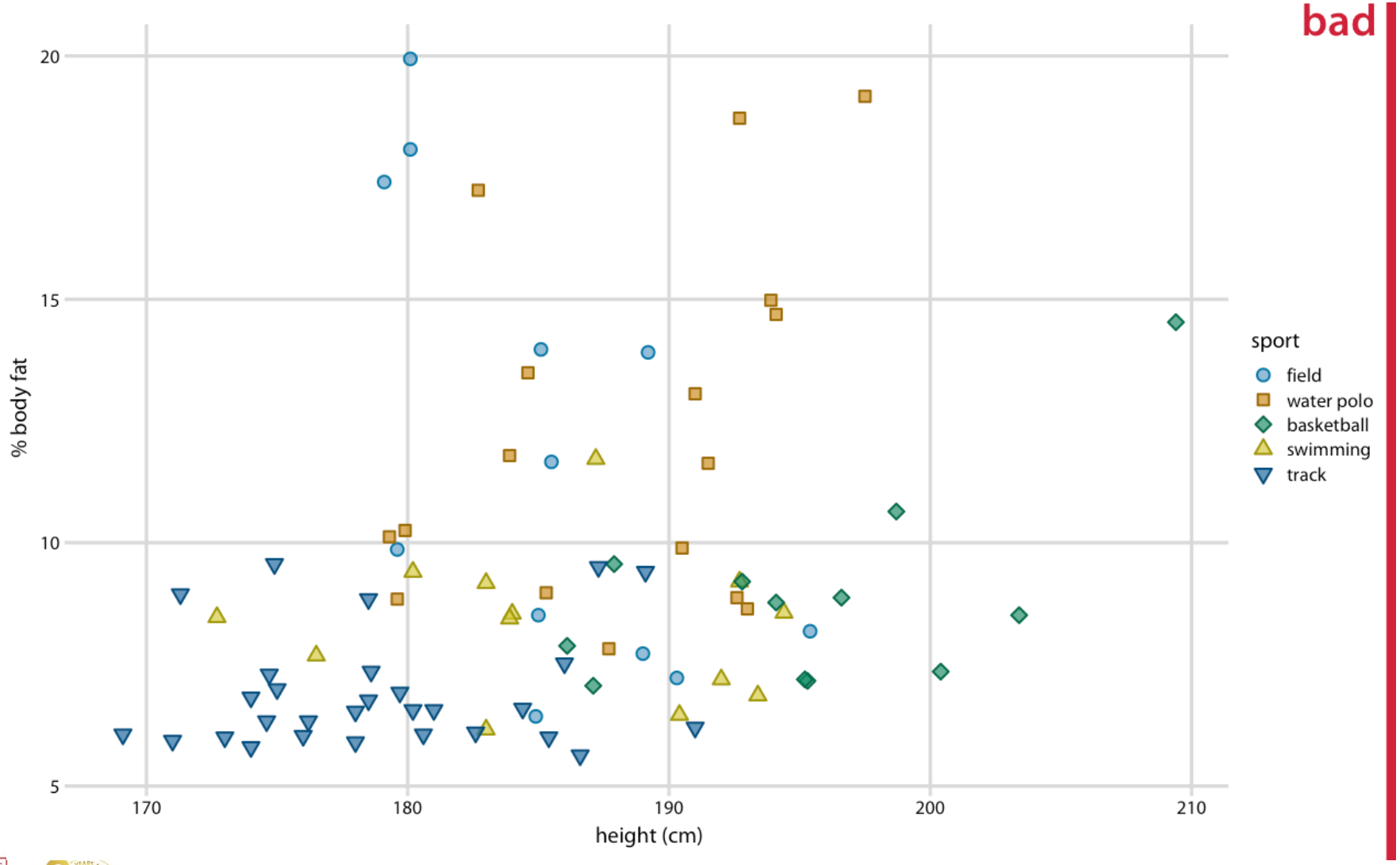


Example: Gene expression levels in a mutant bacteriophage T7 relative to wild type

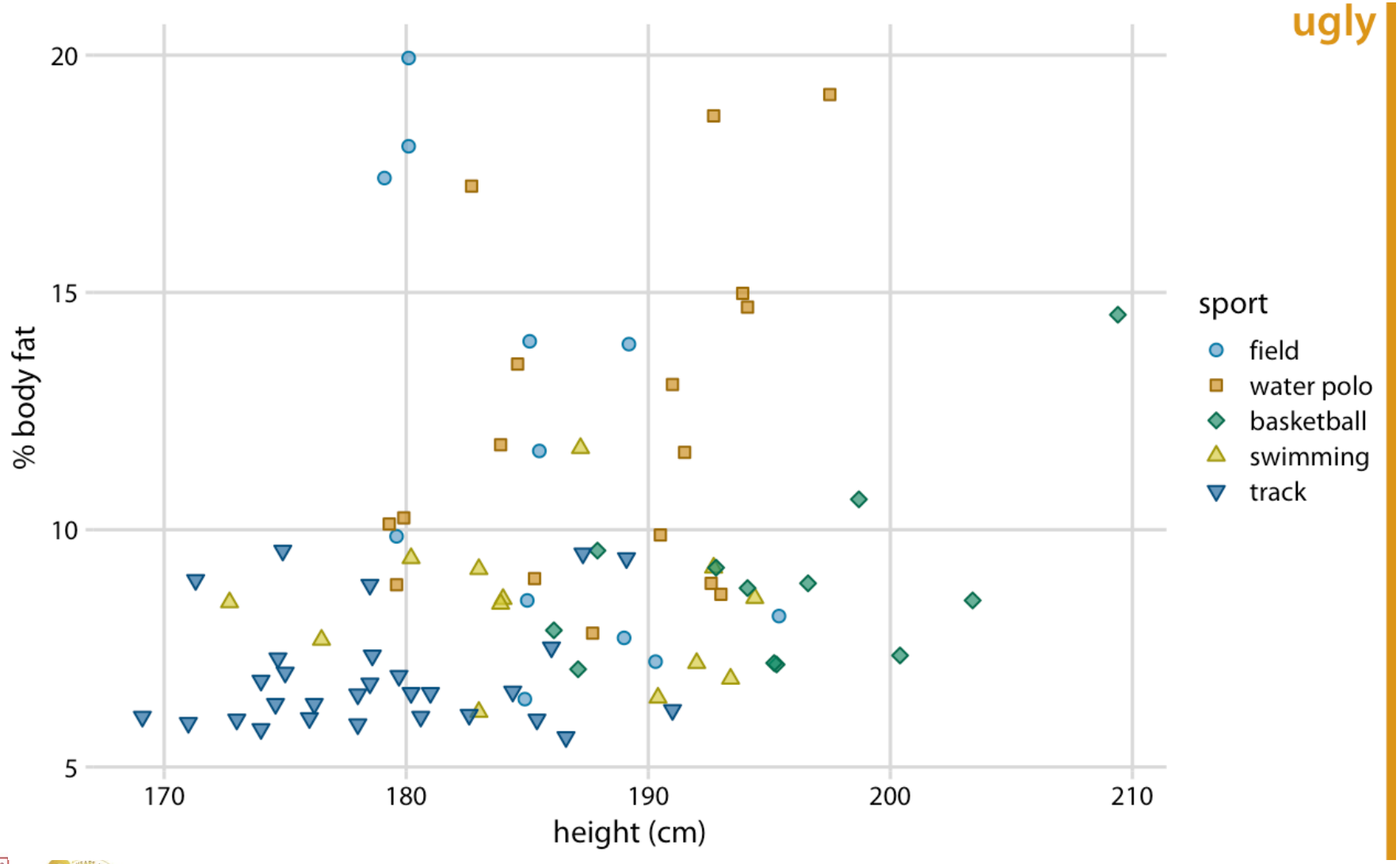


Use larger axis labels

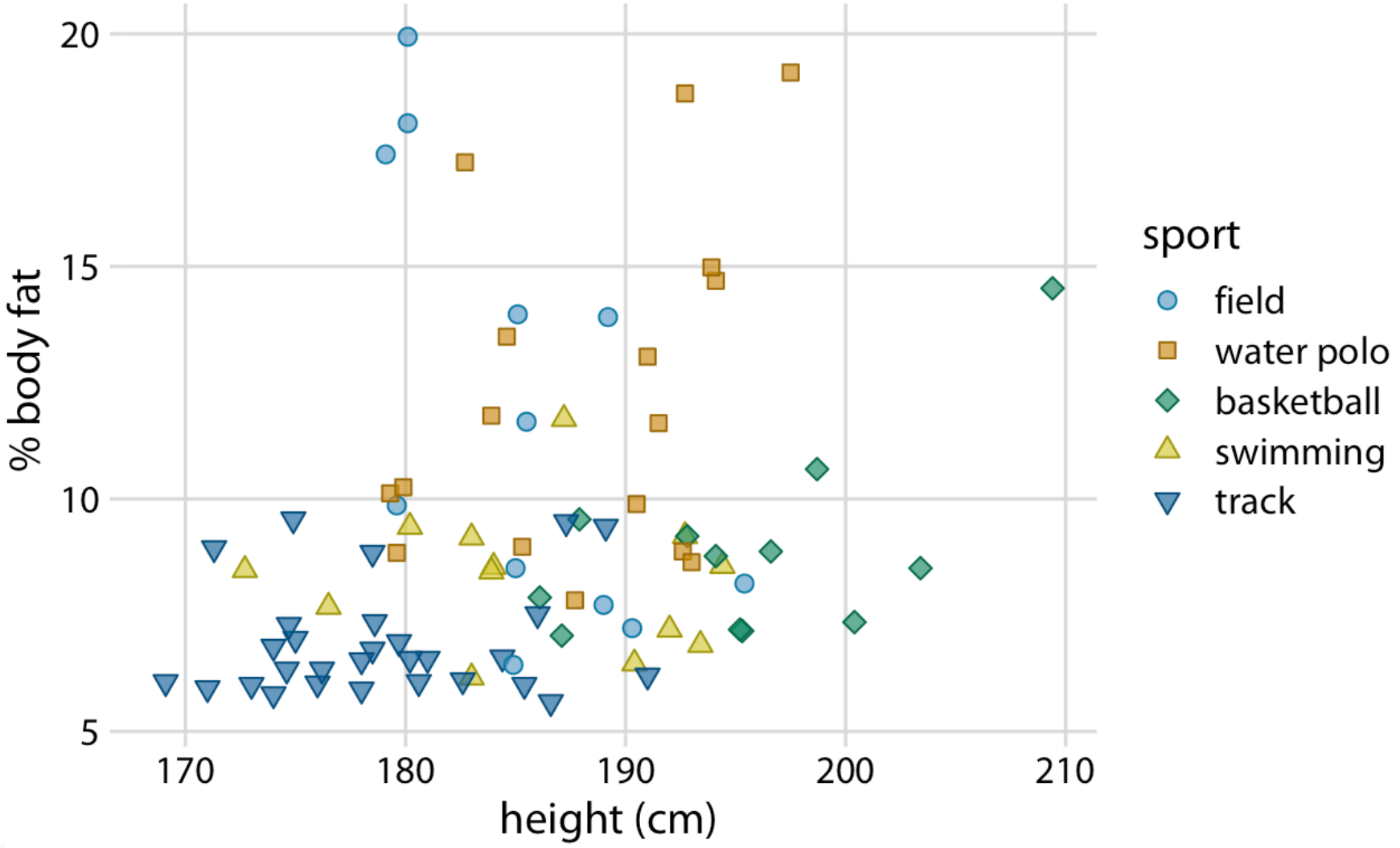
Example: Percent body fat versus height in professional male Australian athletes



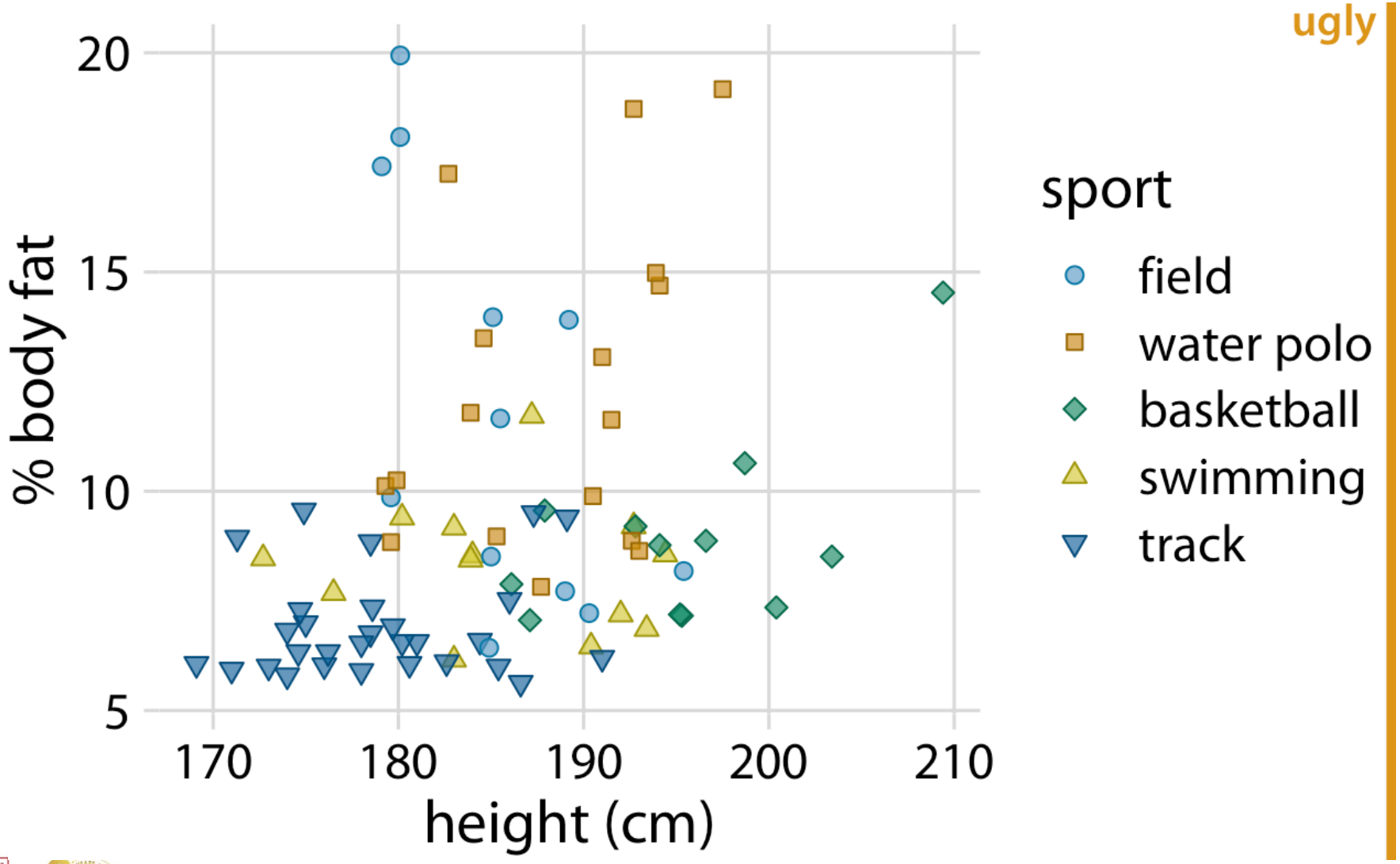
Example: Percent body fat versus height in professional male Australian athletes



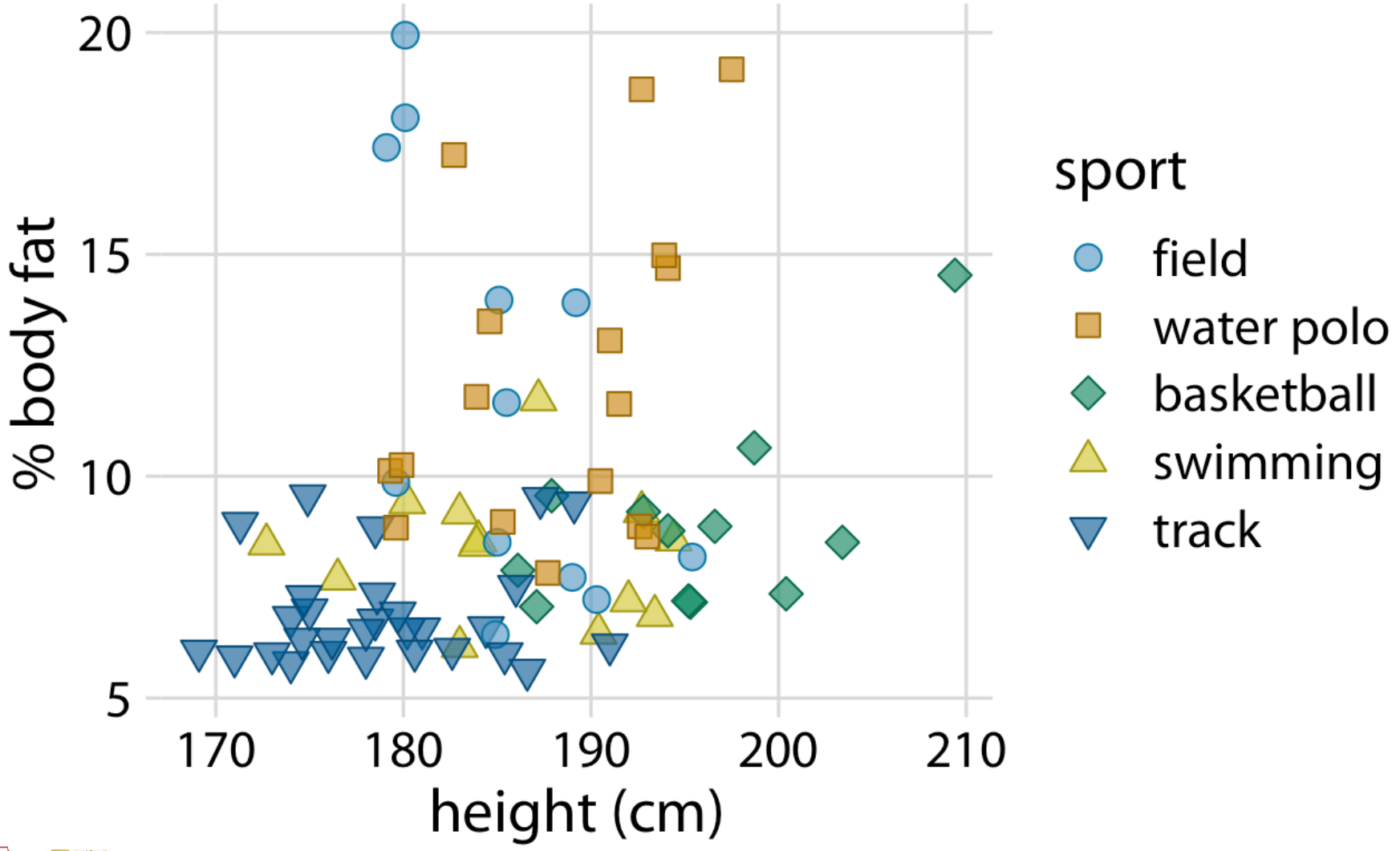
Example: Percent body fat versus height in professional male Australian athletes



Example: Percent body fat versus height in professional male Australian athletes



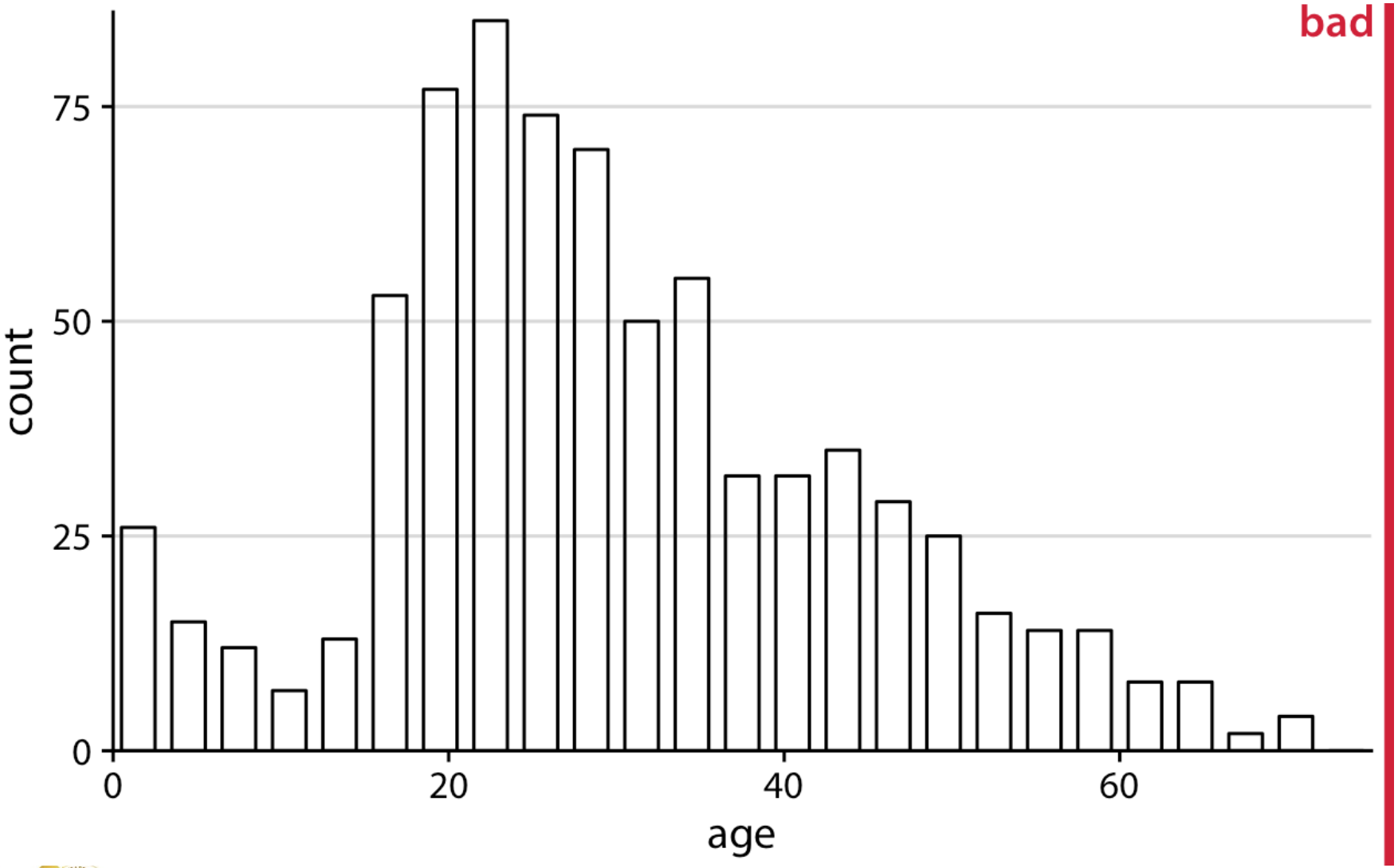
Example: Percent body fat versus height in professional male Australian athletes



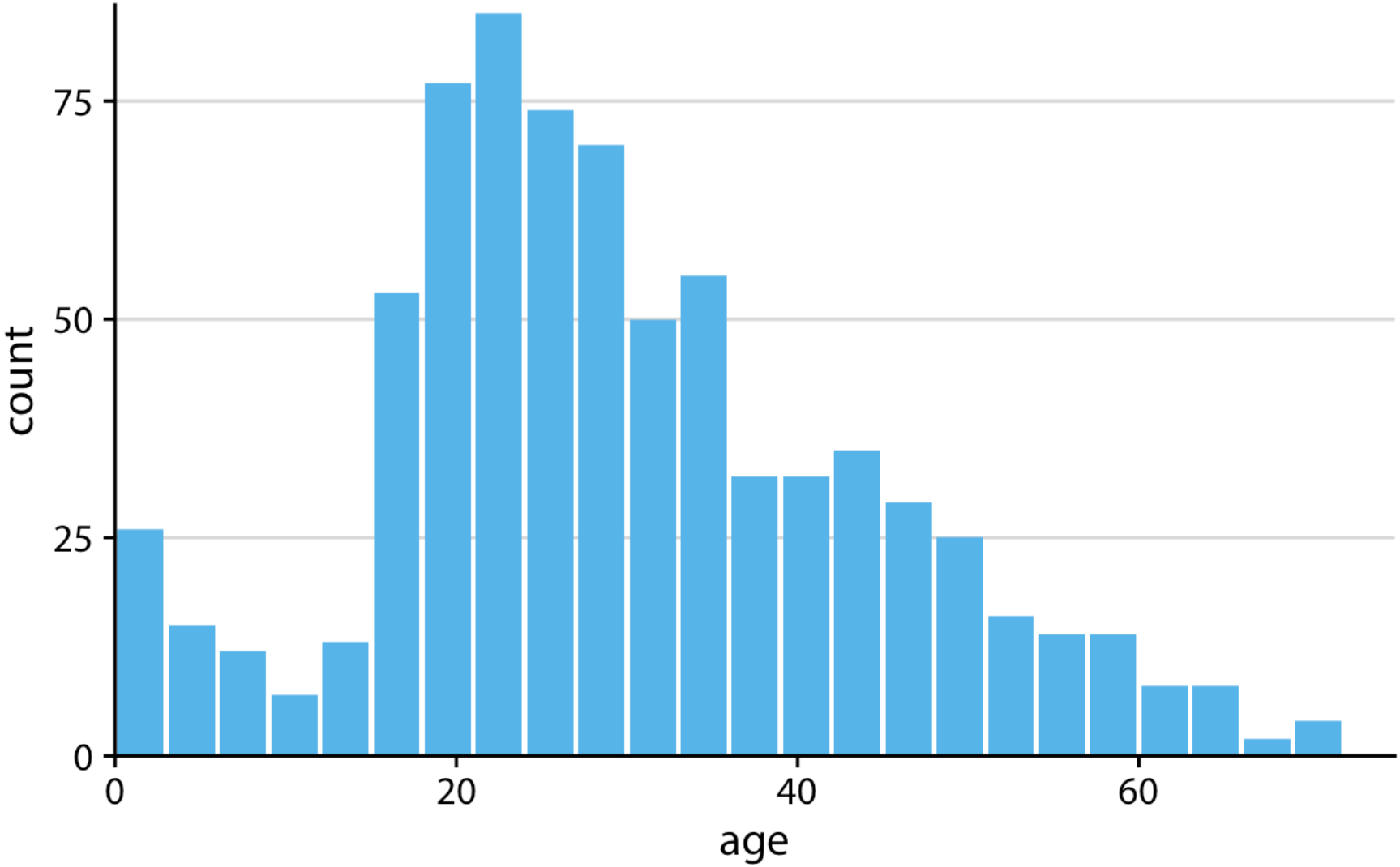
Avoid line drawings

Whenever possible, visualize your data with solid, colored shapes rather than with lines that outline those shapes.

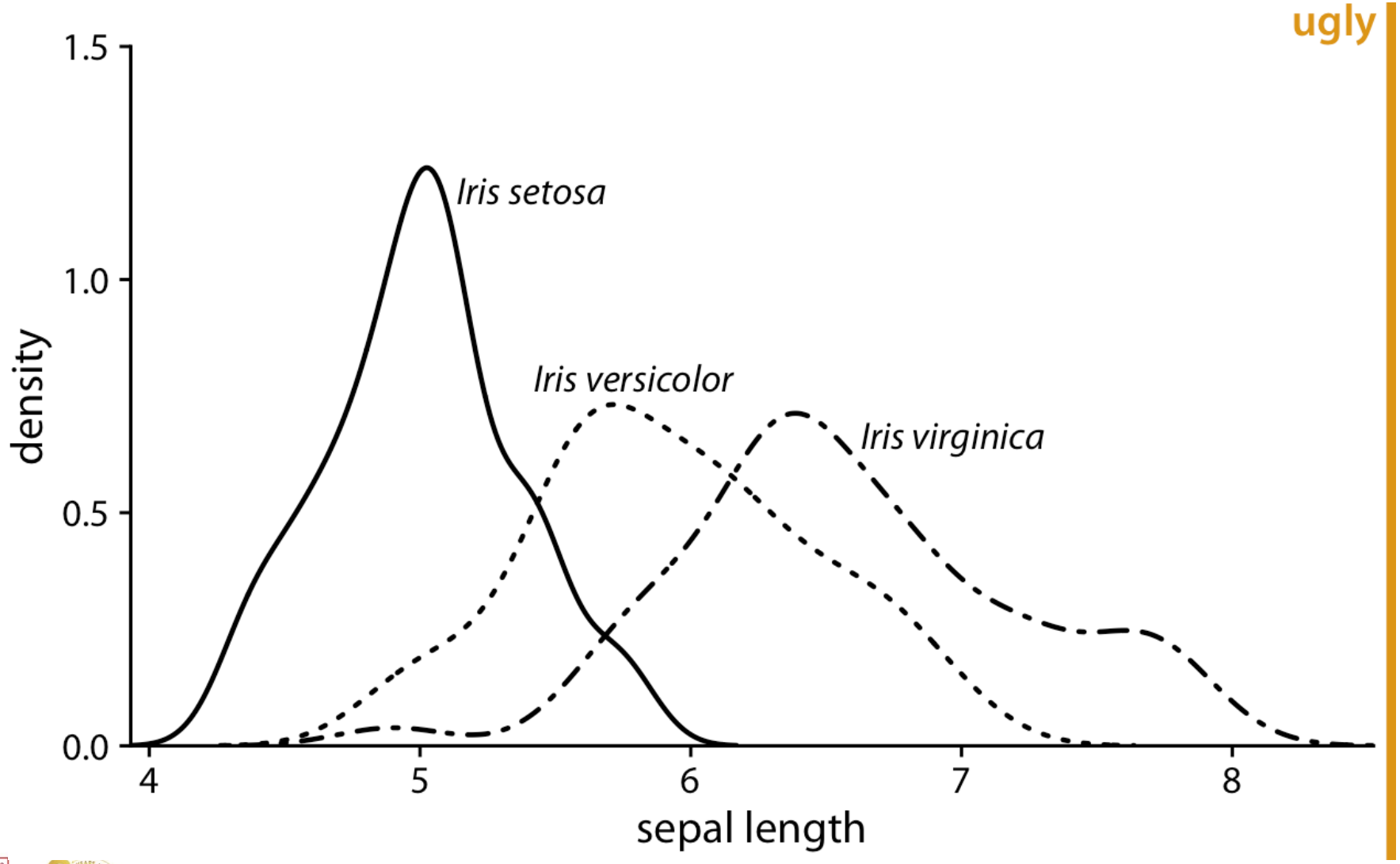
Example: Histogram of the ages of Titanic passengers, drawn with empty bars



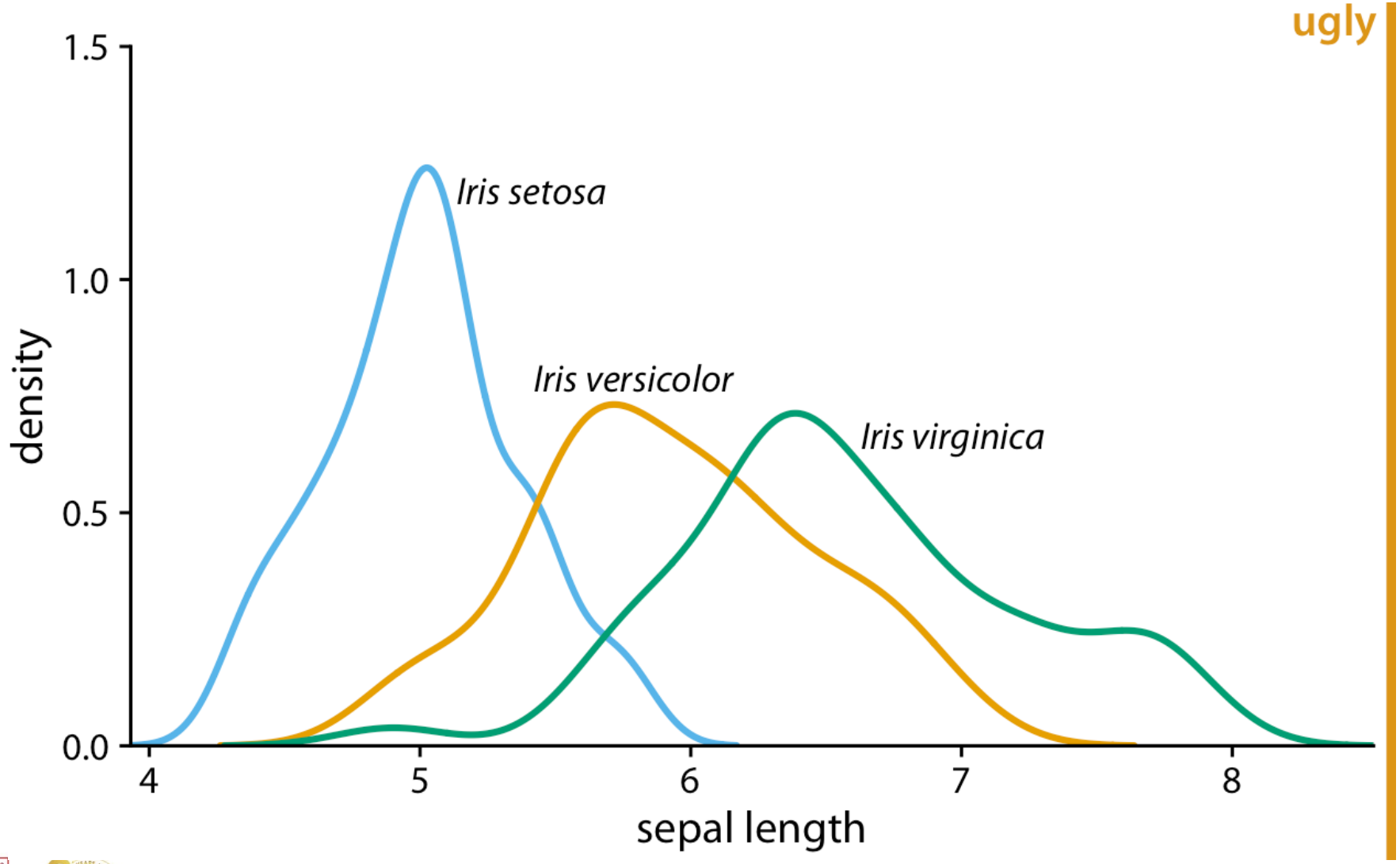
Example: Histogram of the ages of Titanic passengers



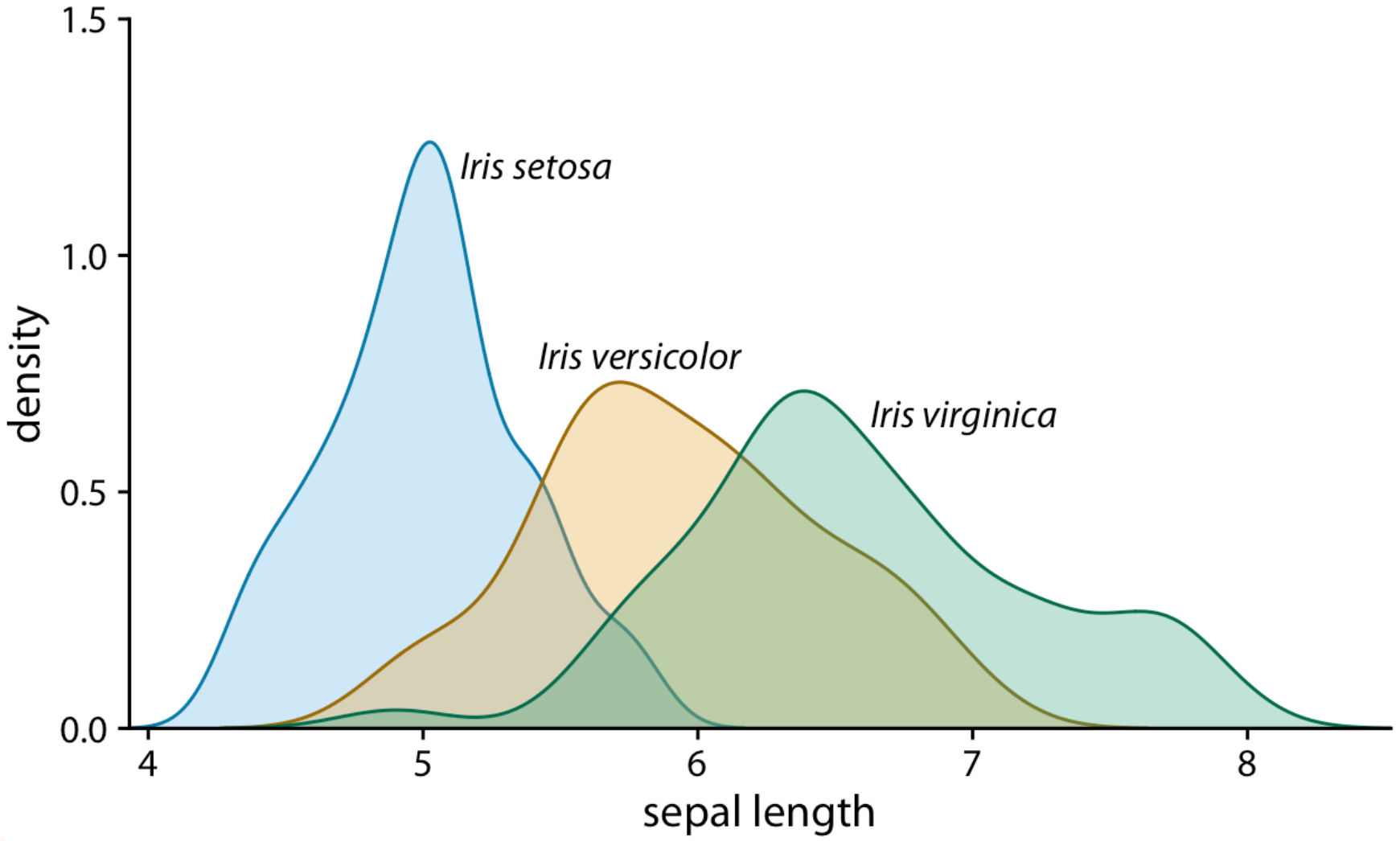
Example: Density estimates of the sepal lengths of three different Iris species



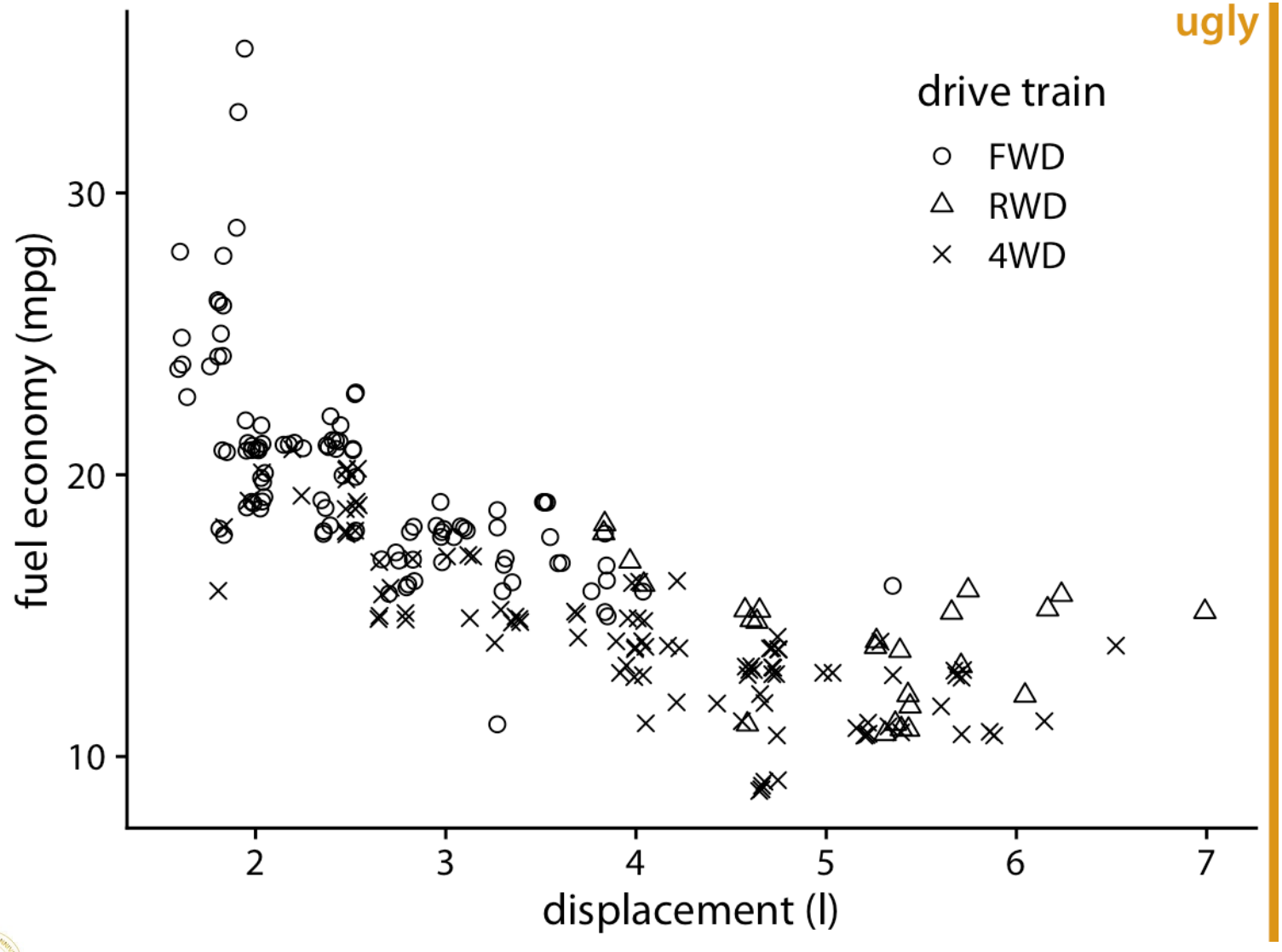
Example: Density estimates of the sepal lengths of three different Iris species



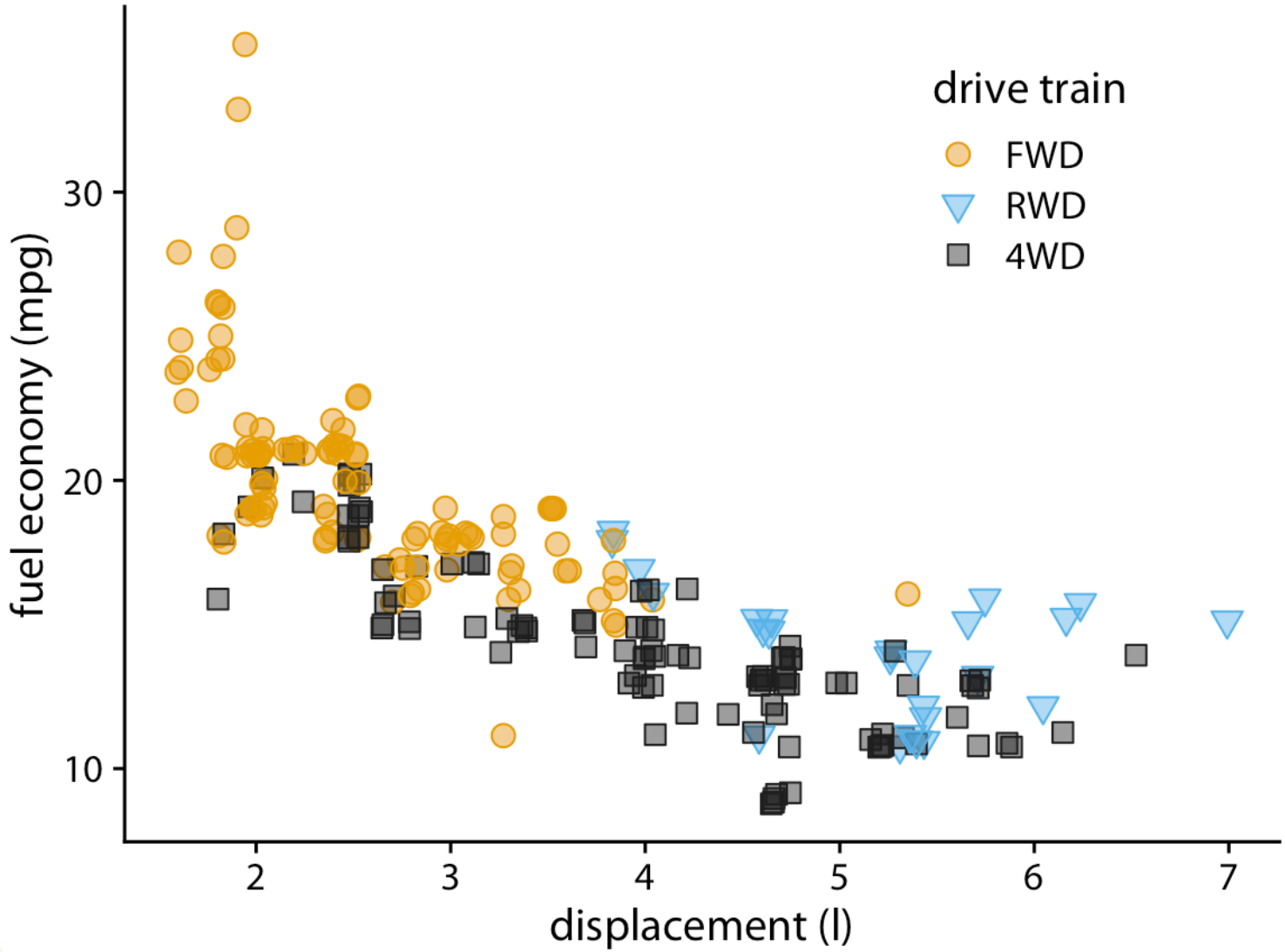
Example: Density estimates of the sepal lengths of three different Iris species



Example: City fuel economy versus engine displacement



Example: City fuel economy versus engine displacement



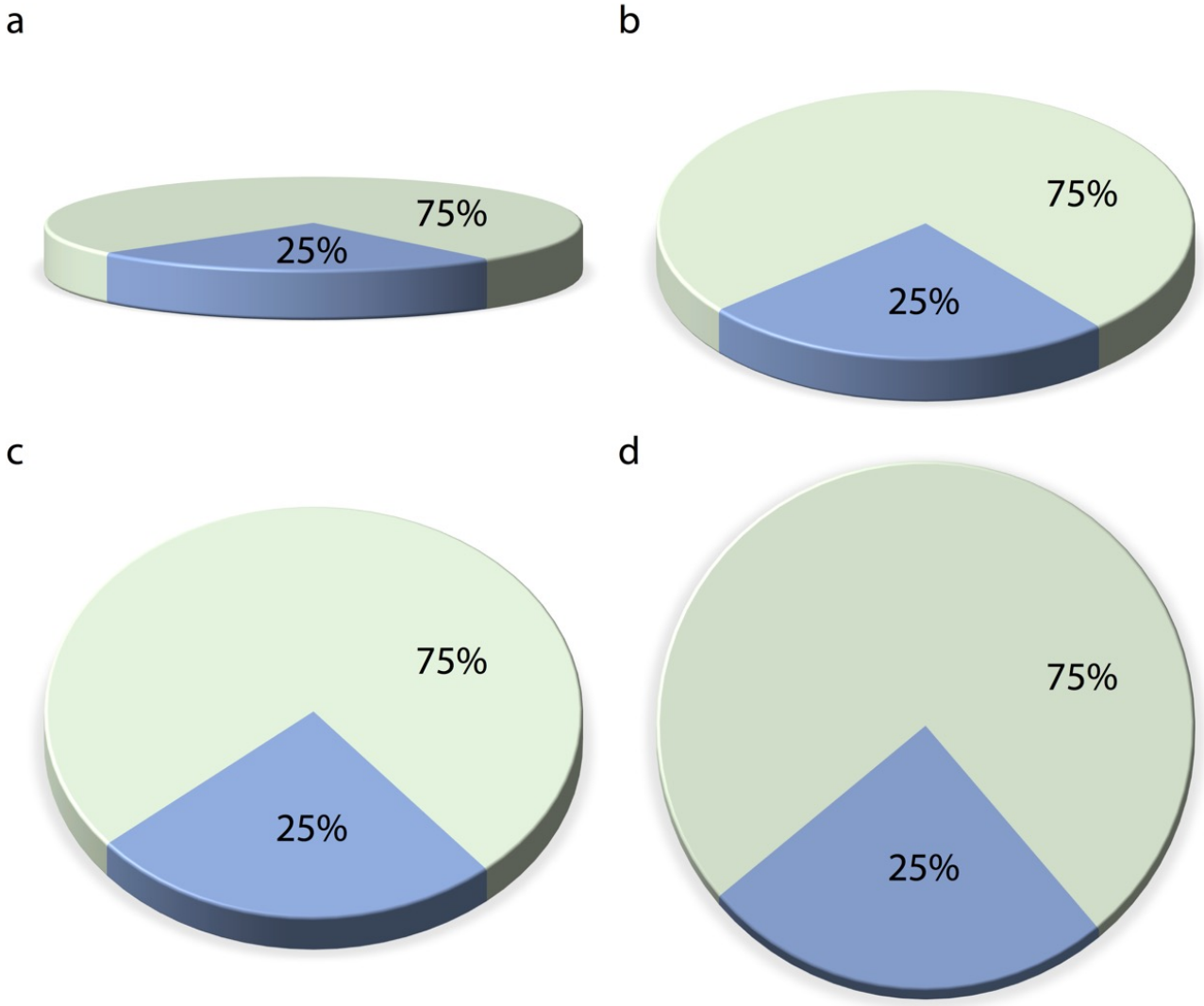
Don’t go 3D

Almost always inappropriately used.

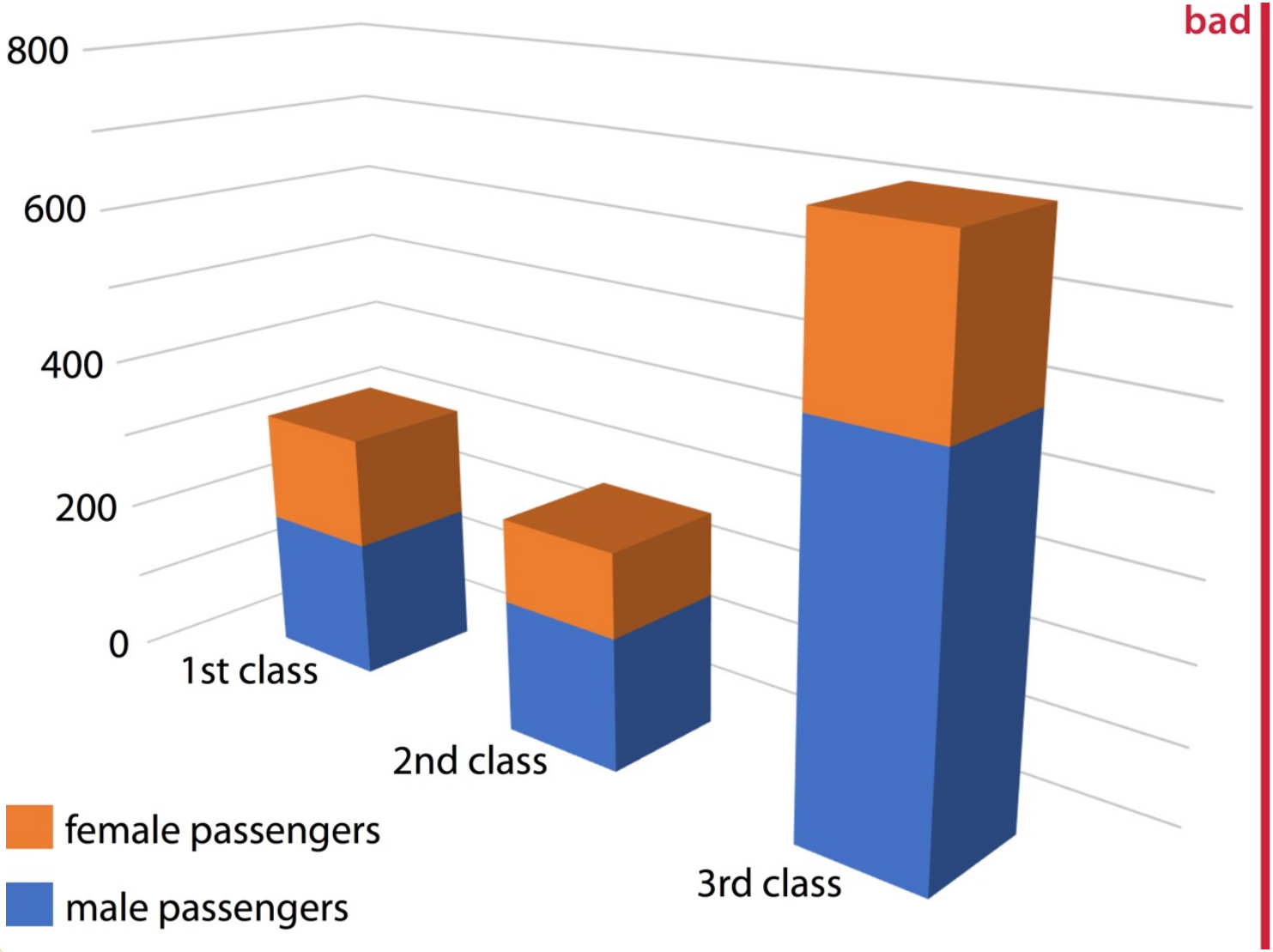
Avoid gratuitous 3D

- The third dimension does not convey any actual data.
- 3D is used simply to decorate and adorn the plot.

Example: The same 3D pie chart shown from four different angles

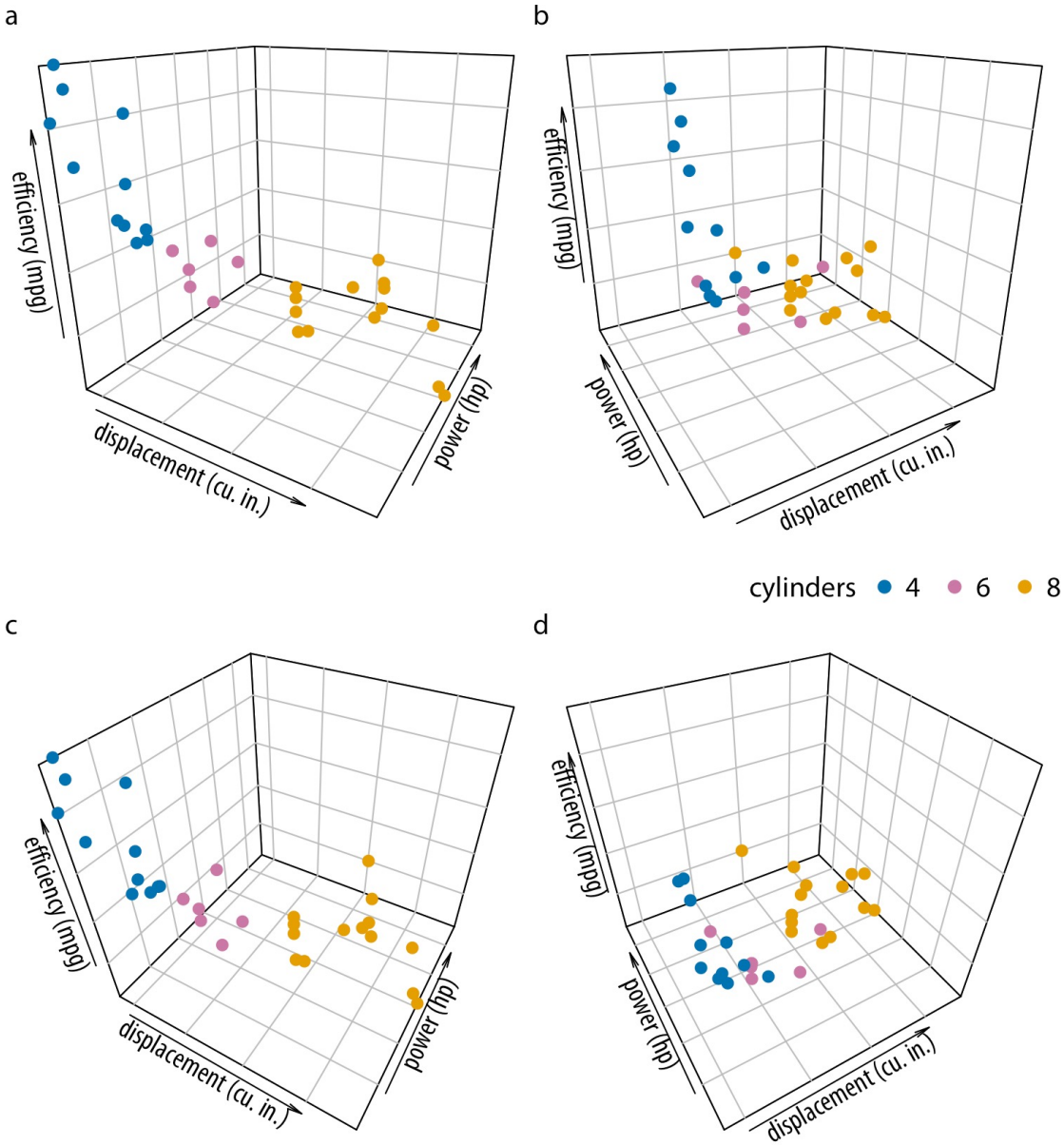


Example: Numbers of female and male passengers on the Titanic traveling in 1st, 2nd, and 3rd class

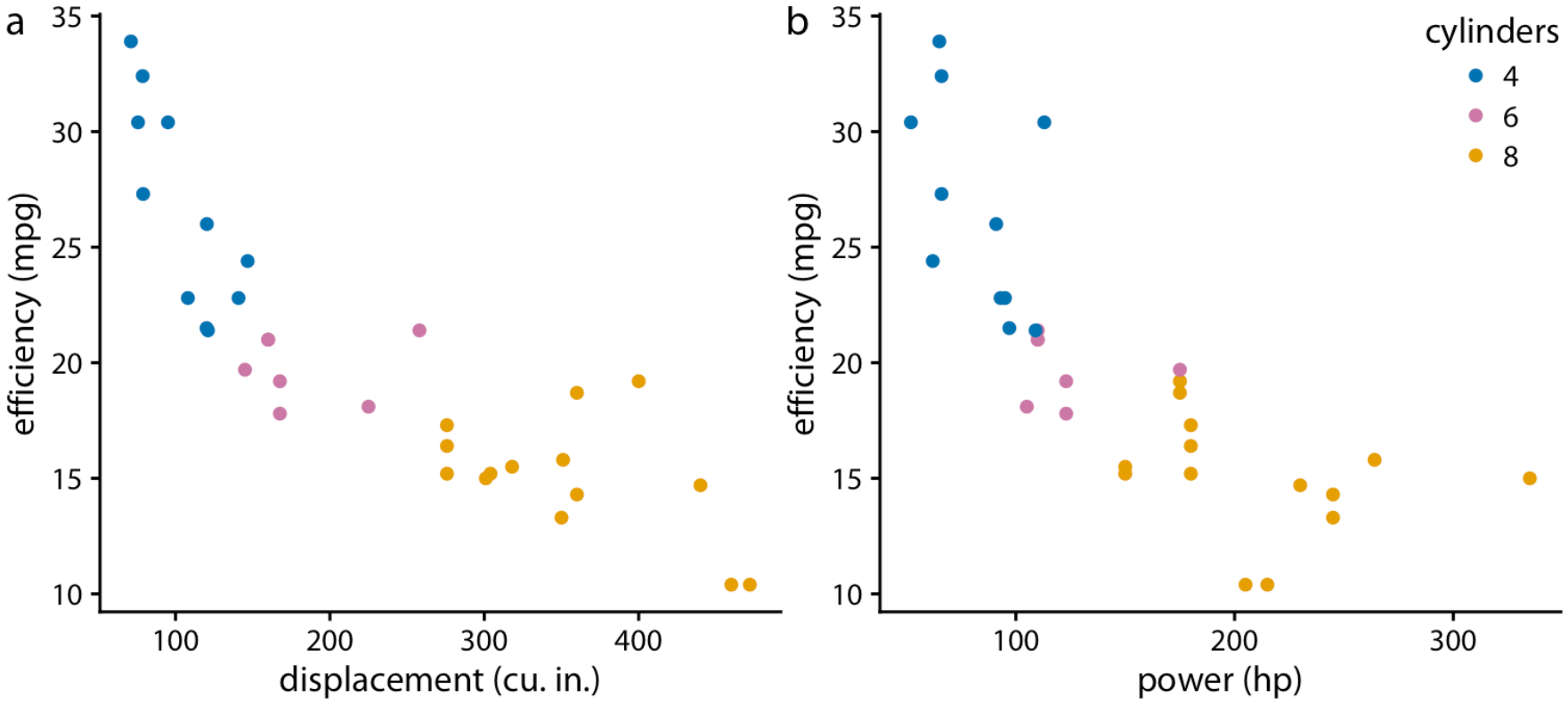


Avoid 3D position scales

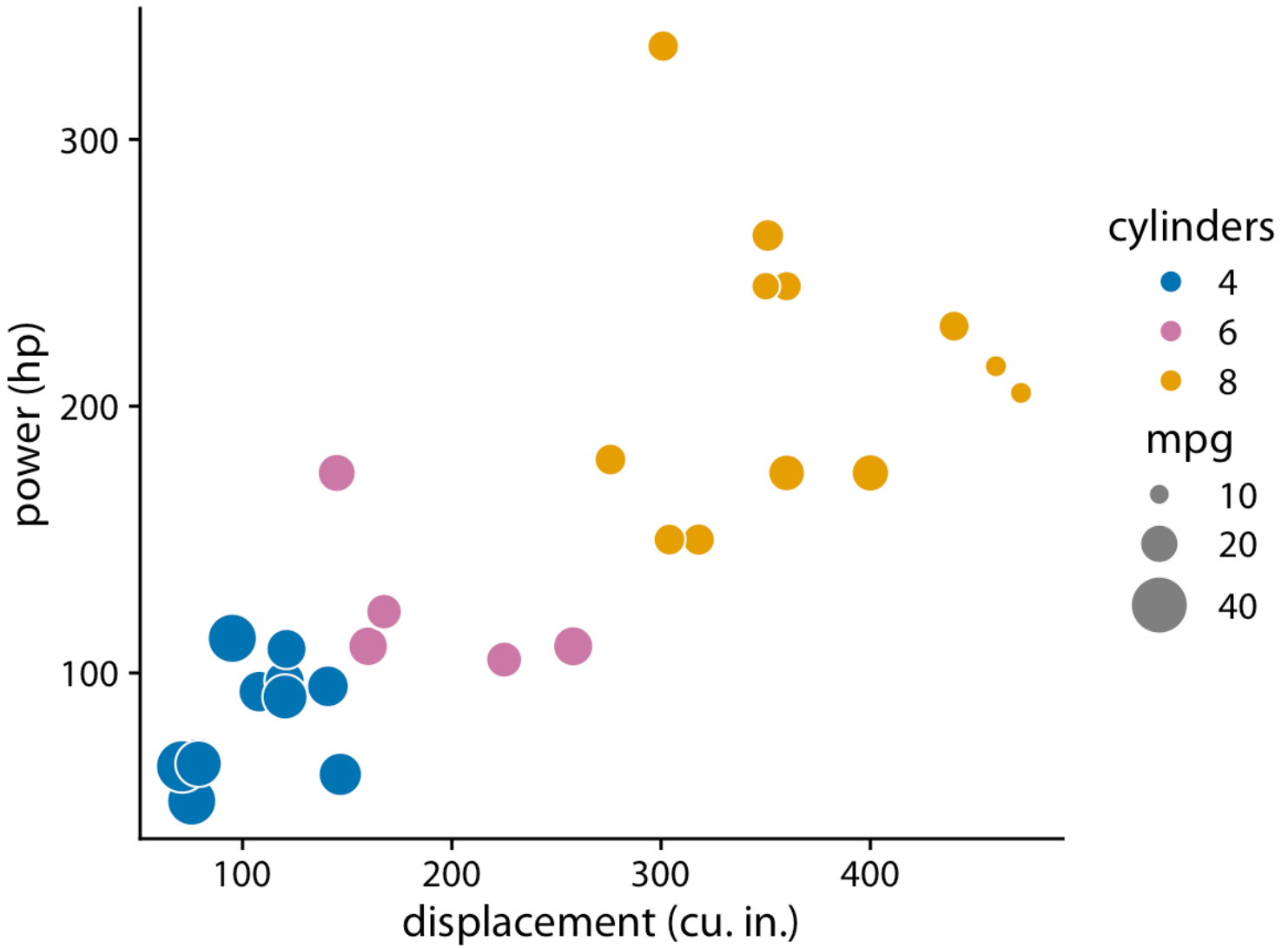
- The use of the third dimension serves an actual purpose.
- Nevertheless, the resulting plots are frequently difficult to interpret.
- Example: Fuel efficiency versus displacement and power for 32 cars



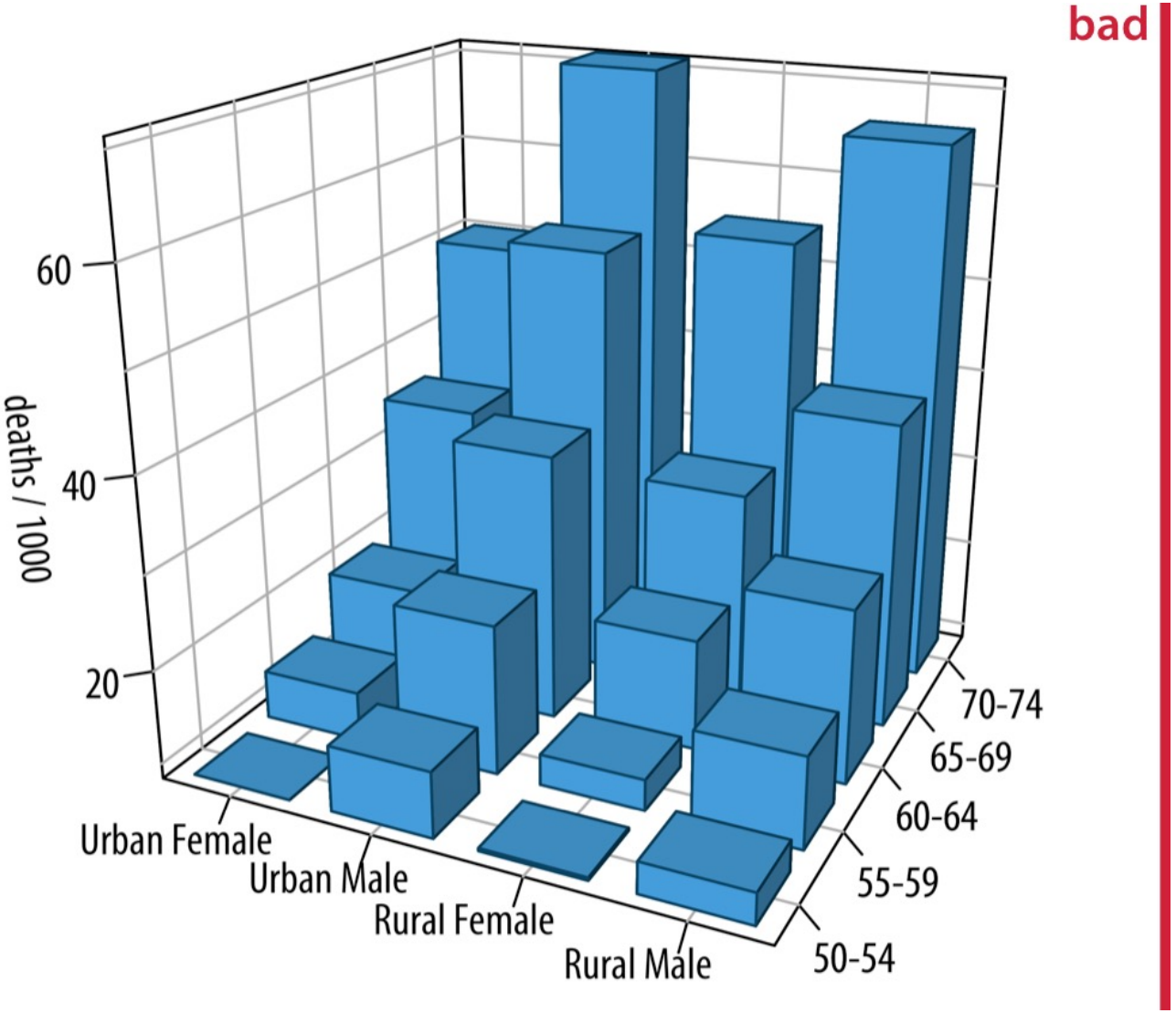
Example: Fuel efficiency versus displacement (a) and power (b)



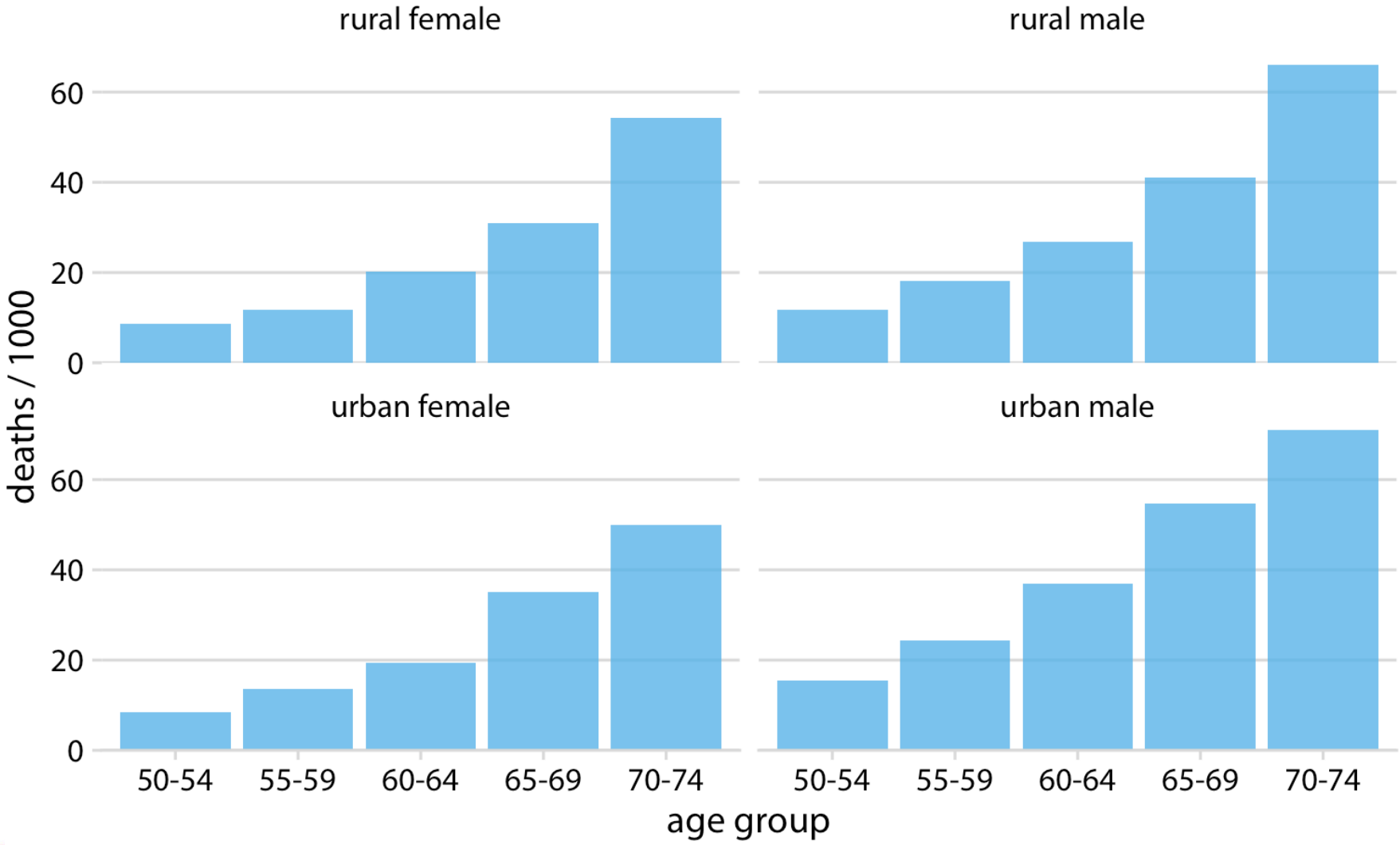
Example: Power versus displacement for 32 cars, with fuel efficiency represented by dot size



Example: Mortality rates in Virginia in 1940, visualized as a 3D bar plot



Example: Mortality rates in Virginia in 1940, visualized as a small multiples plot



Appropriate use of 3D visualizations

- The visualization is interactive and can be rotated by the viewer.
- If the visualization isn't interactive, showing it slowly rotating, will allow the viewer to discern where in 3D space different graphical elements reside.
- Use 3D visualizations when we want to show actual 3D objects and/or data mapped onto them.

Example: Relief of the Island of Corsica in the Mediterranean Sea

Example: Patterns of evolutionary variation in a protein