HUST

ĐẠI HỌC BÁCH KHOA HÀ NỘI

HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

ONE LOVE. ONE FUTURE.

Visualization for table, multi-dimentional data

- Last lectures
  - Visual model and visual encoding
  - Graphical perception (visual decoding)
- Today lecture
  - Visualization for table, multi-dimentional data

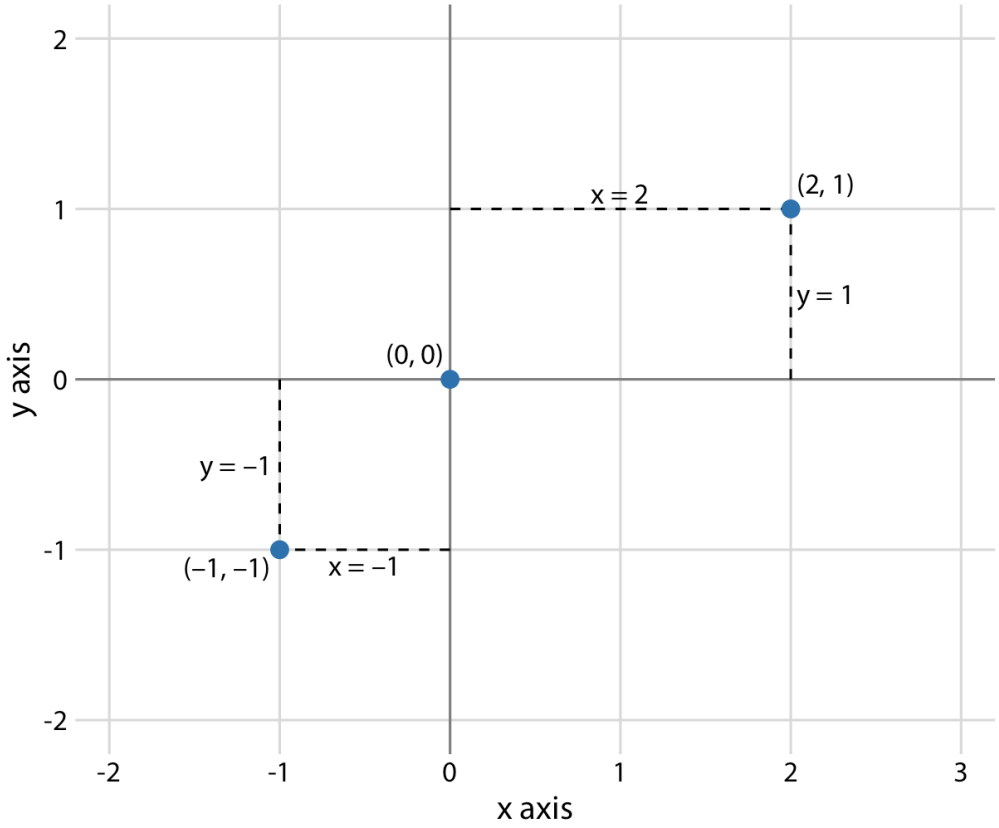
Outline

- Coordinate systems and axes
- Color scales
- Visualizing amounts
- Visualizing distributions
- Visualizing many distributions at once
- Next lesson
  - Visualizing proportions
  - Visualizing associations
  - Visualizing trends
  - Visualizing uncertainty

Coordinate systems and axes

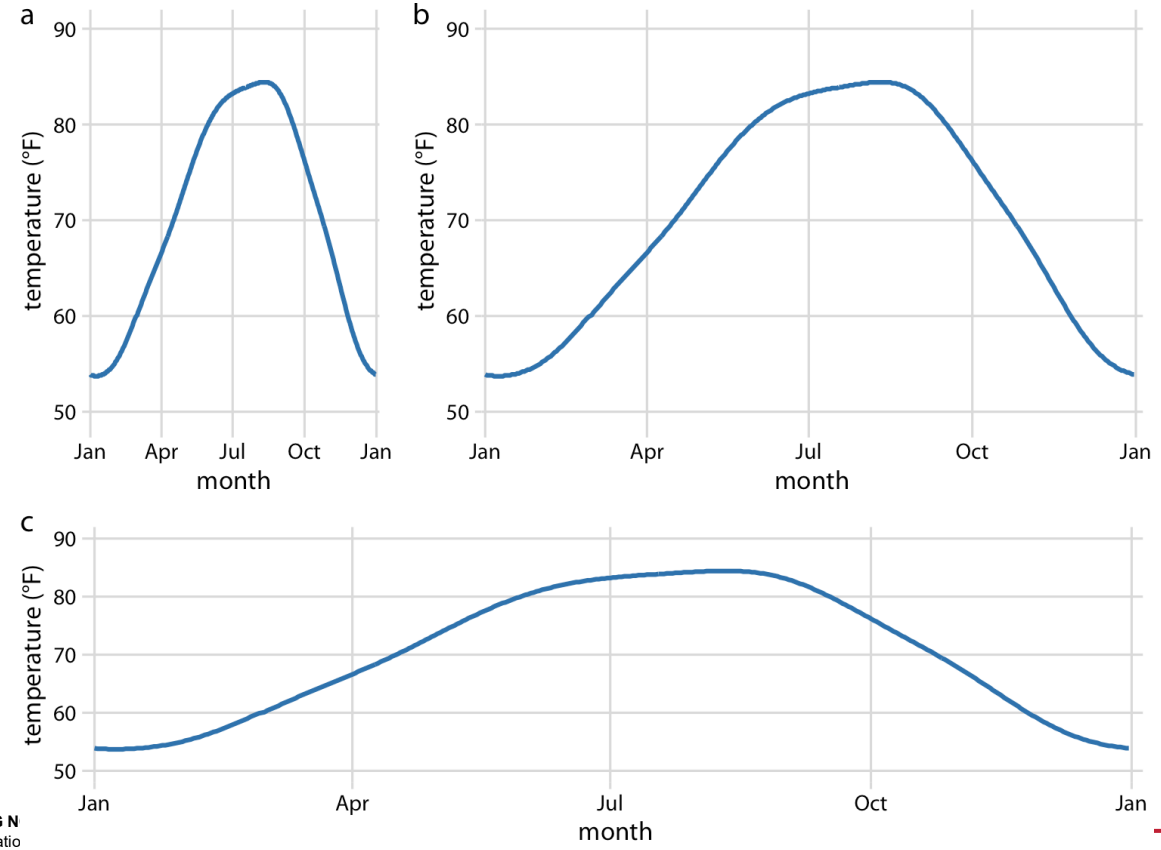
Cartesian Coordinates

- 2D Cartesian coordinate system
  - Each location is uniquely specified by an x and a y value.
  - The x and y axes run orthogonally to each other
  - Data values are placed in an even spacing along both axes



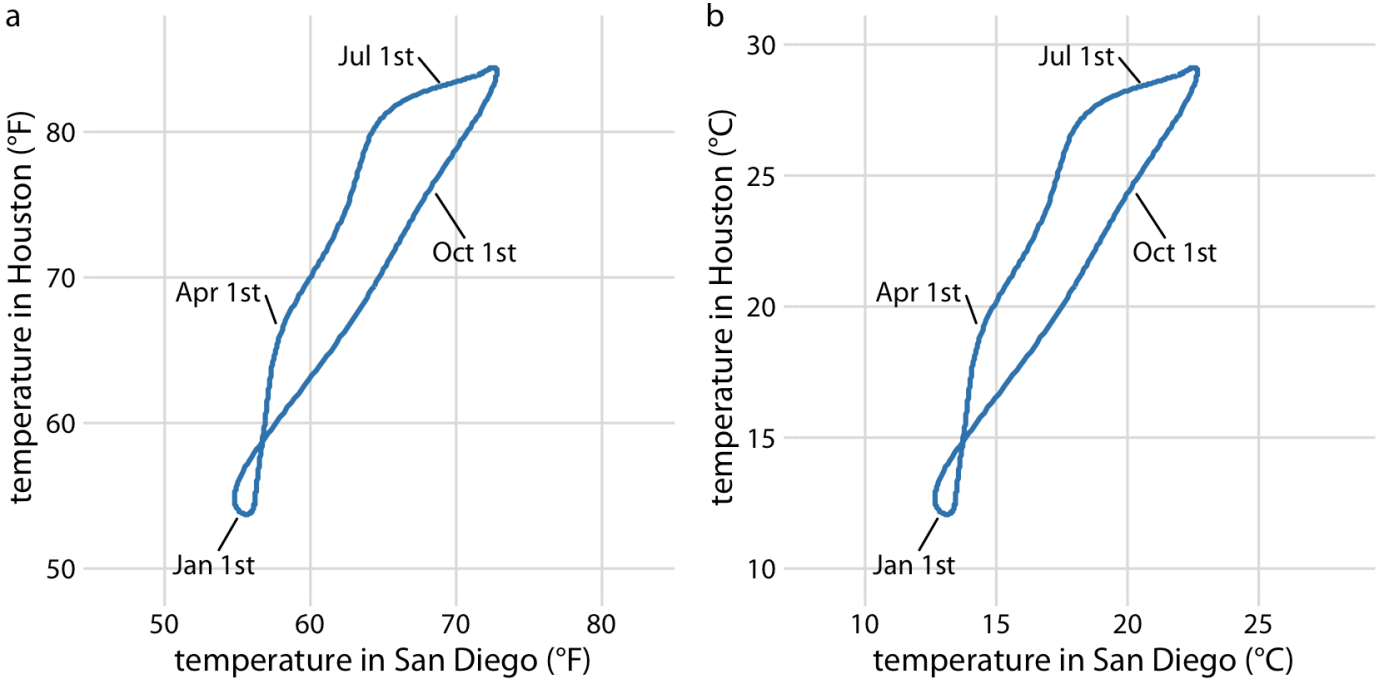
Example: Daily temperature for Houston

- The same figure in different aspect ratios.
- All three parts are valid visualizations of the temperature data.



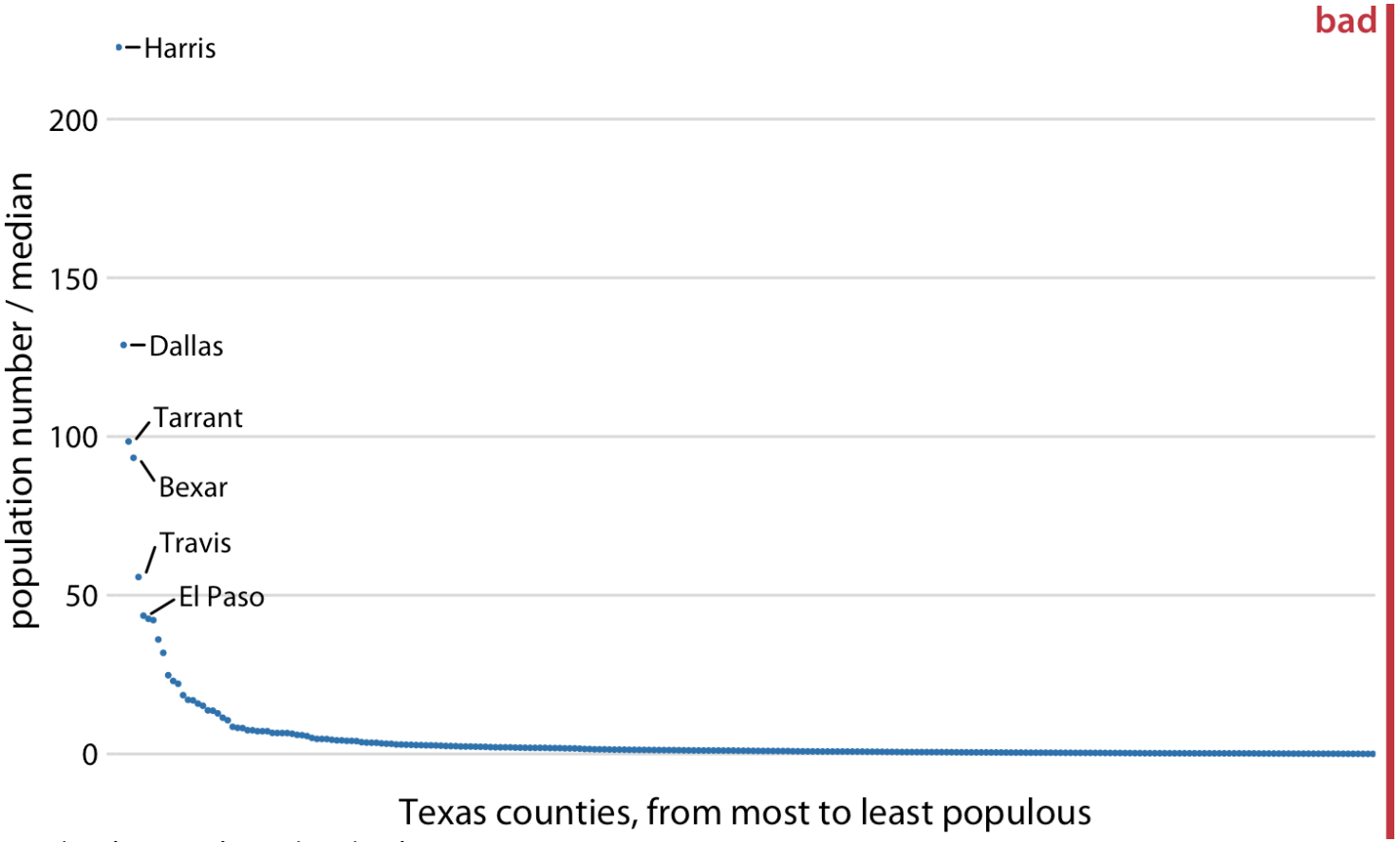
Example: Daily temperature for Houston

- If the x and y axes are measured in the same units, then the grid spacings for the two axes should be equal.



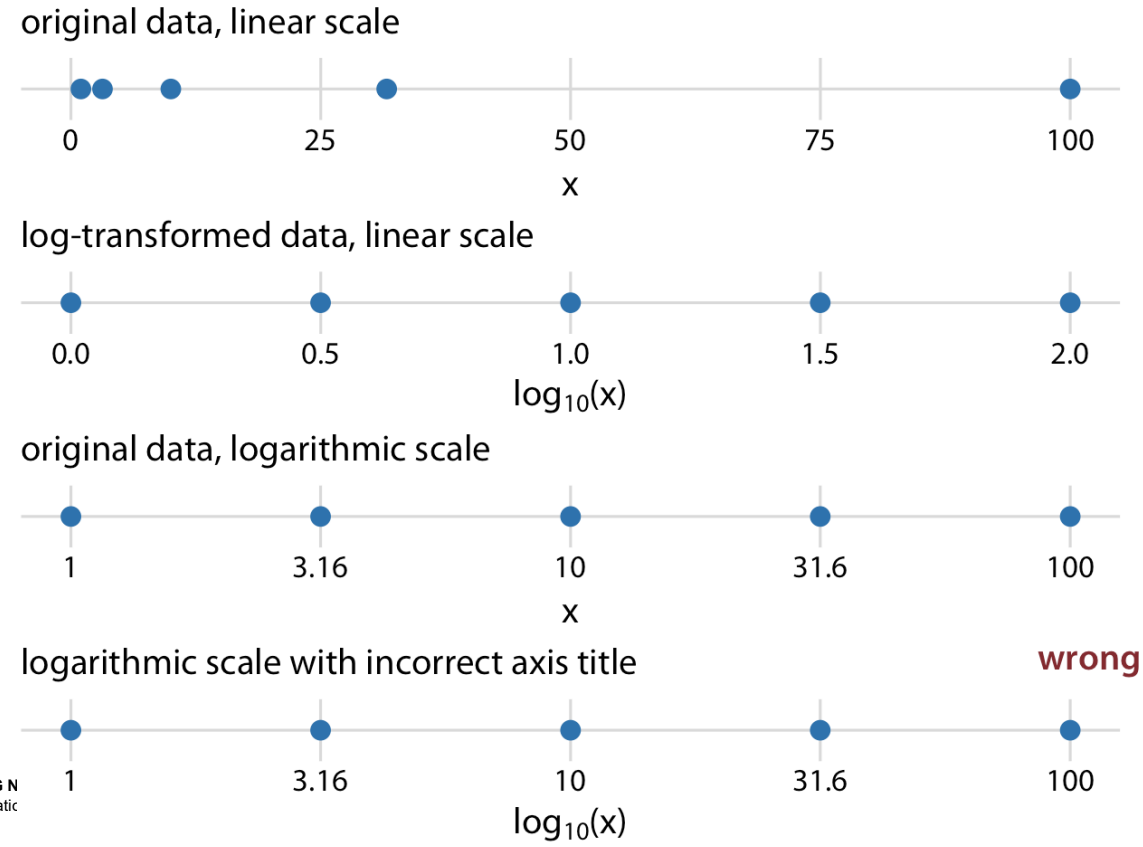
Example: Population numbers of Texas counties relative to their median value

- By displaying a ratio on a linear scale, we have overemphasized ratios > 1 and have obscured ratios < 1.
- Generally, ratios should not be displayed on a linear scale.



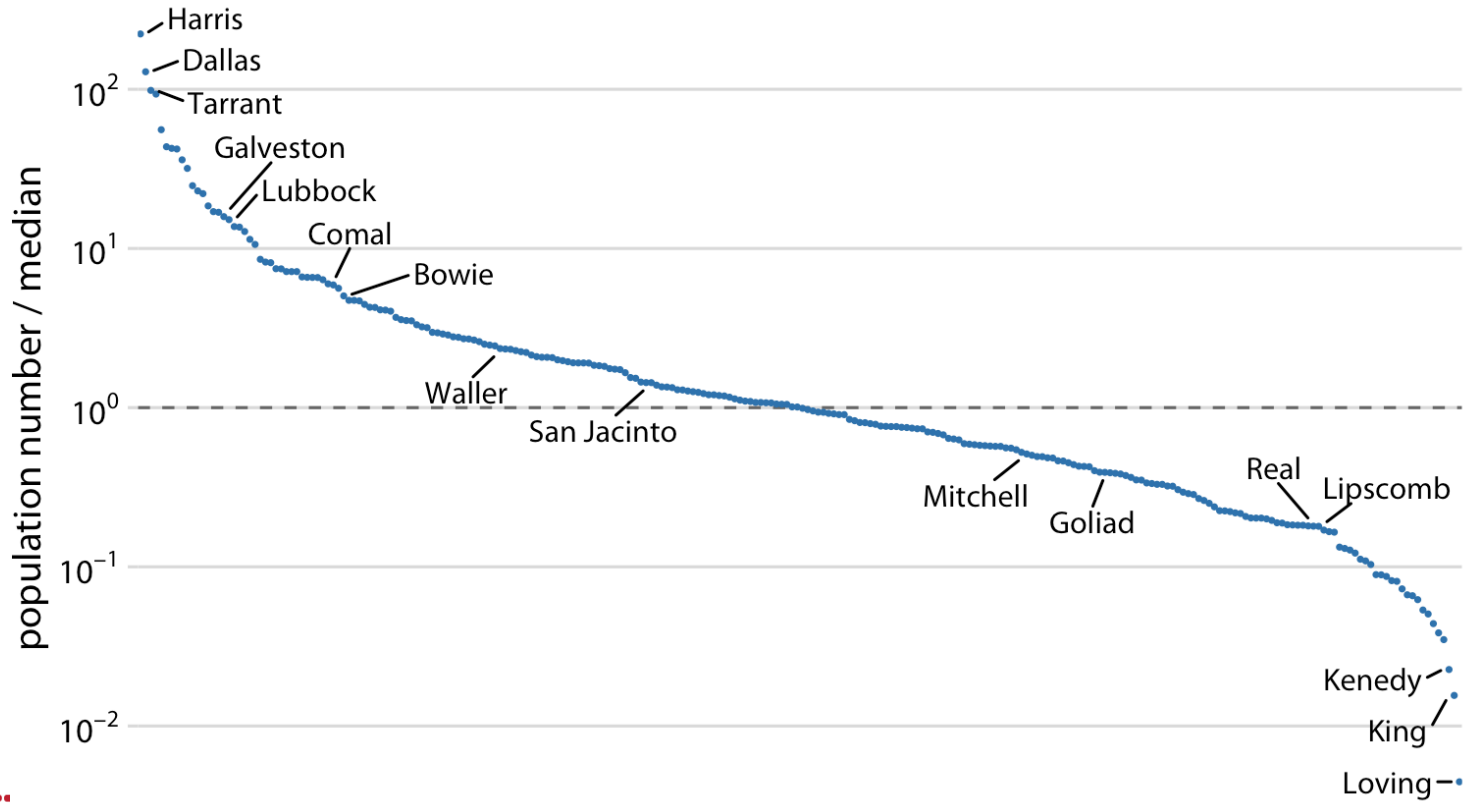
Nonlinear axes

- In a Cartesian coordinate system, the position scales is linear.
- In a nonlinear scale, even spacing in the visualization corresponds to uneven spacing in data units.

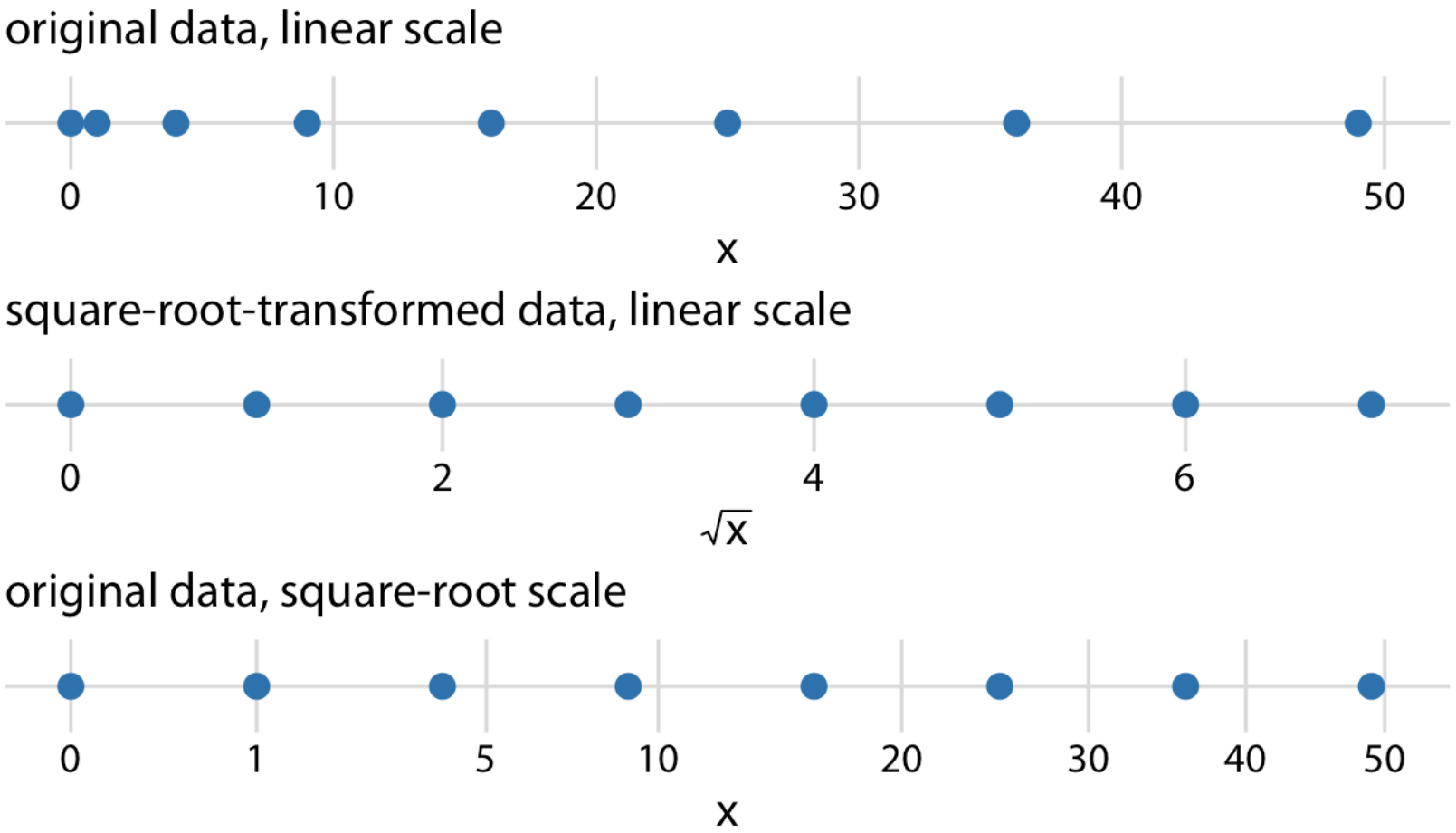


Example: Population numbers of Texas counties relative to their median value

- The dashed line indicates a ratio of 1, corresponding to a county with median population number. The most populous counties have approximately 100 times more inhabitants than the median county, and the least populous counties have approximately 100 times fewer inhabitants than the median county.

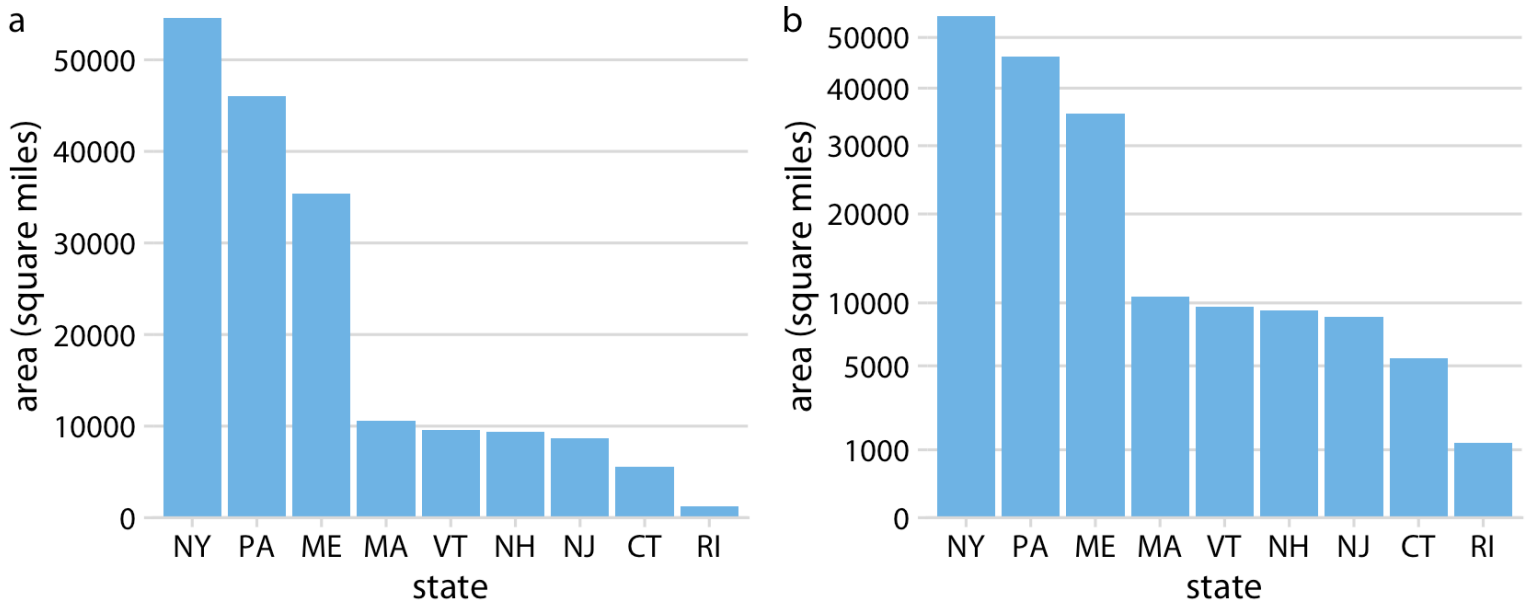


Square-root scales



Example: Areas of northeastern US states

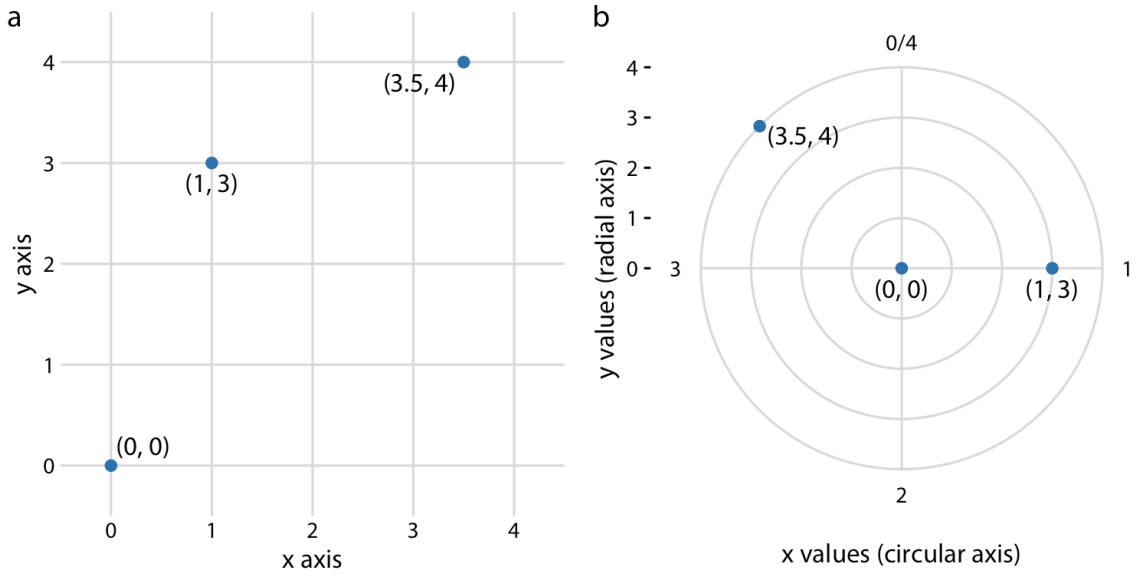
- The square-root scale is the natural scale for data that comes in squares



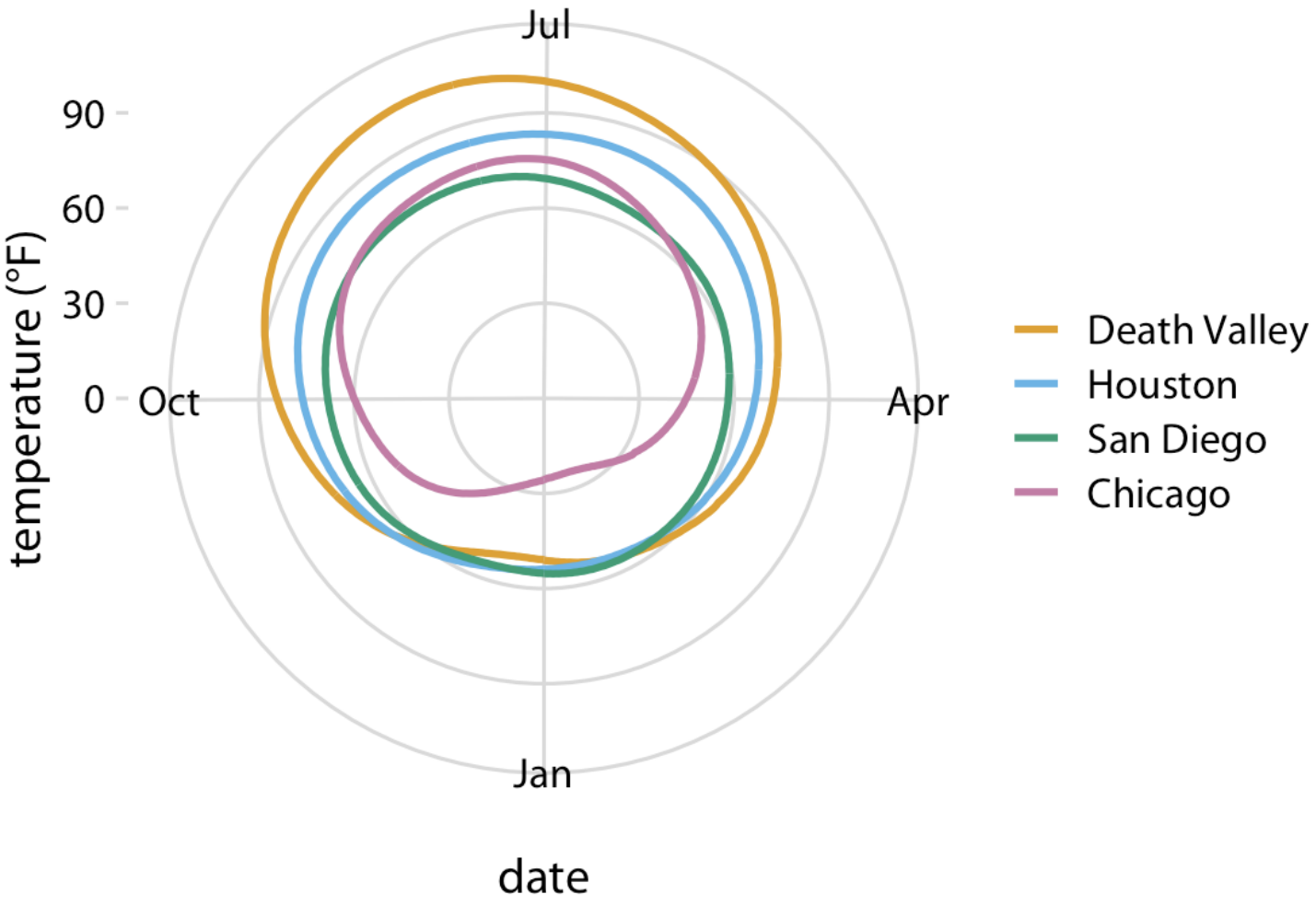
(a) Areas shown on a linear scale. (b) Areas shown on a square-root scale

Coordinate systems with curved axes

- In the polar coordinate system, we specify positions via an angle and a radial distance from the origin.
  - The angle axis is circular
- Useful for data of a periodic nature, such that data values at one end of the scale can be logically joined to data values at the other end.



Example: Daily temperature normals for four selected locations in the US



Color scales

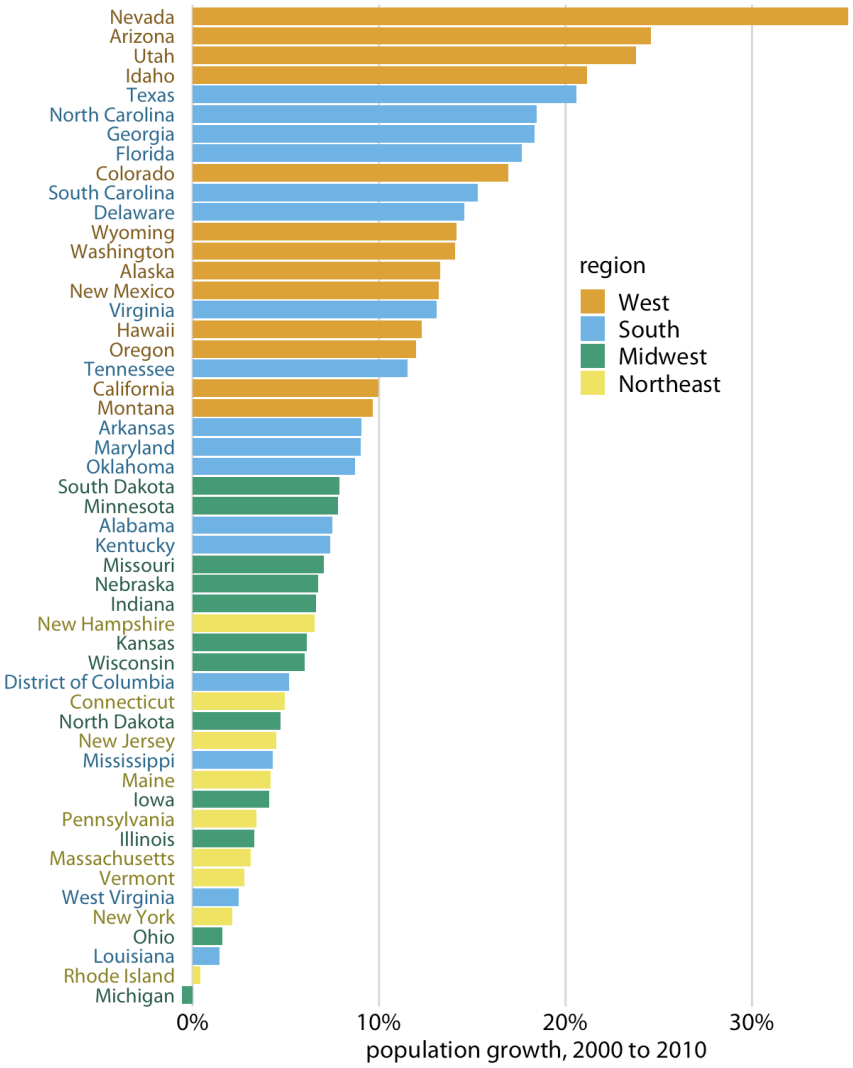
Use cases

- Three fundamental use cases for color in data visualizations
  - Distinguish groups of data from each other
  - Represent data values
  - Highlight

Color as a tool to distinguish

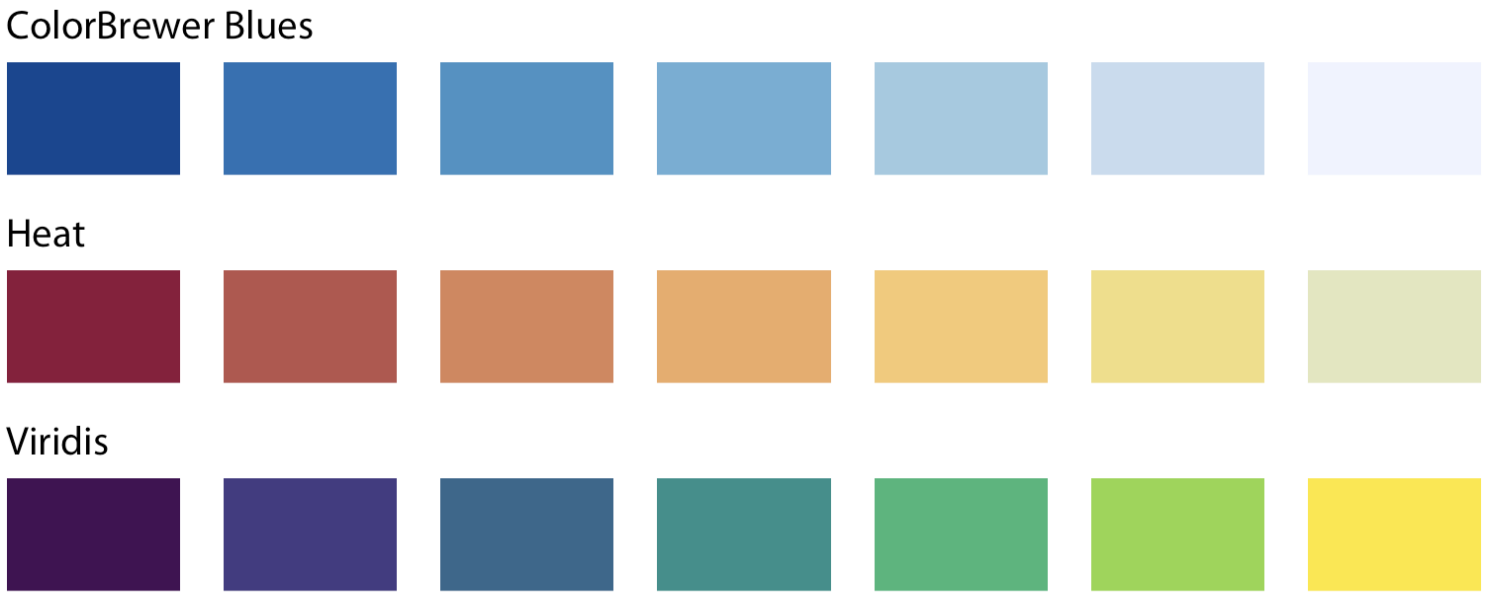
- A qualitative color scale:
  - A finite set of specific colors that are chosen to look clearly distinct from each other while also being equivalent to each other
- Distinguish discrete items or groups that do not have an intrinsic order, such as different countries on a map or different manufacturers of a certain product.

Example: Population growth in the US from 2000 to 2010

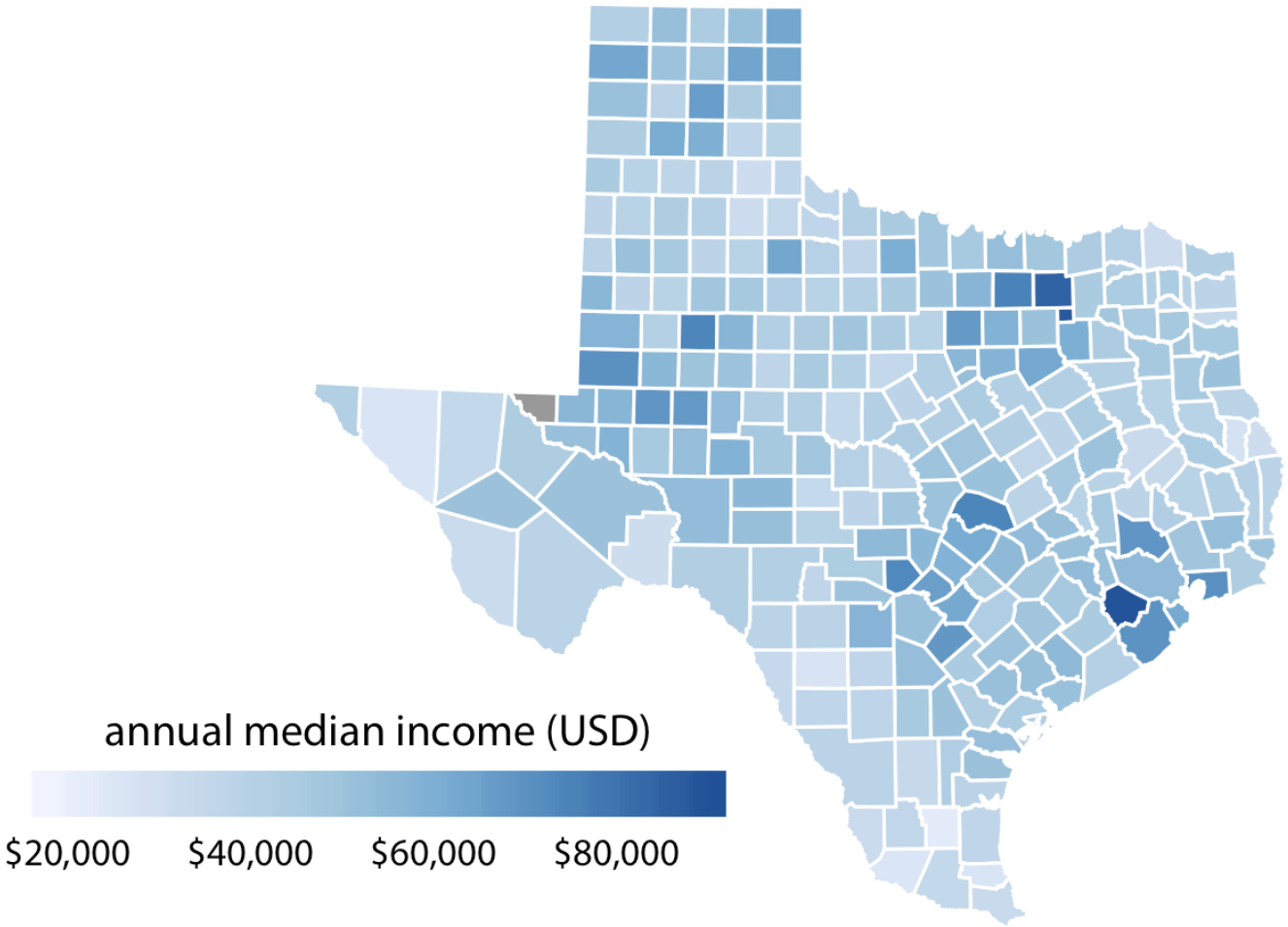


Color to represent data values

- A sequential color scale:
  - Represent quantitative data values, such as income, temperature, or speed
  - Clearly indicate which values are larger or smaller than which other ones, and how distant two specific values are from each other



Example: Median annual income in Texas counties

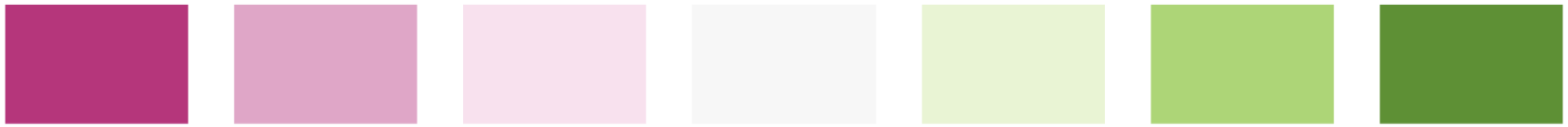


Diverging color scales

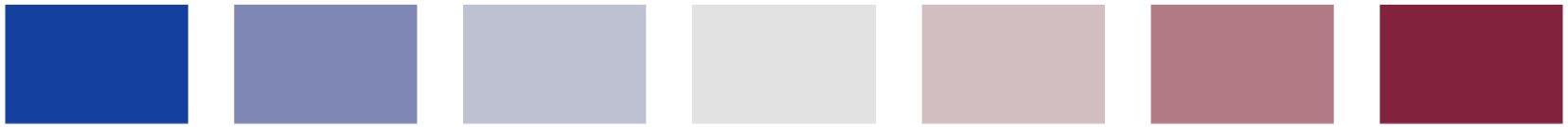
- Visualize the deviation of data values in one of two directions relative to a neutral midpoint.
- Two sequential scales stitched together at a common midpoint color.

CARTO Earth

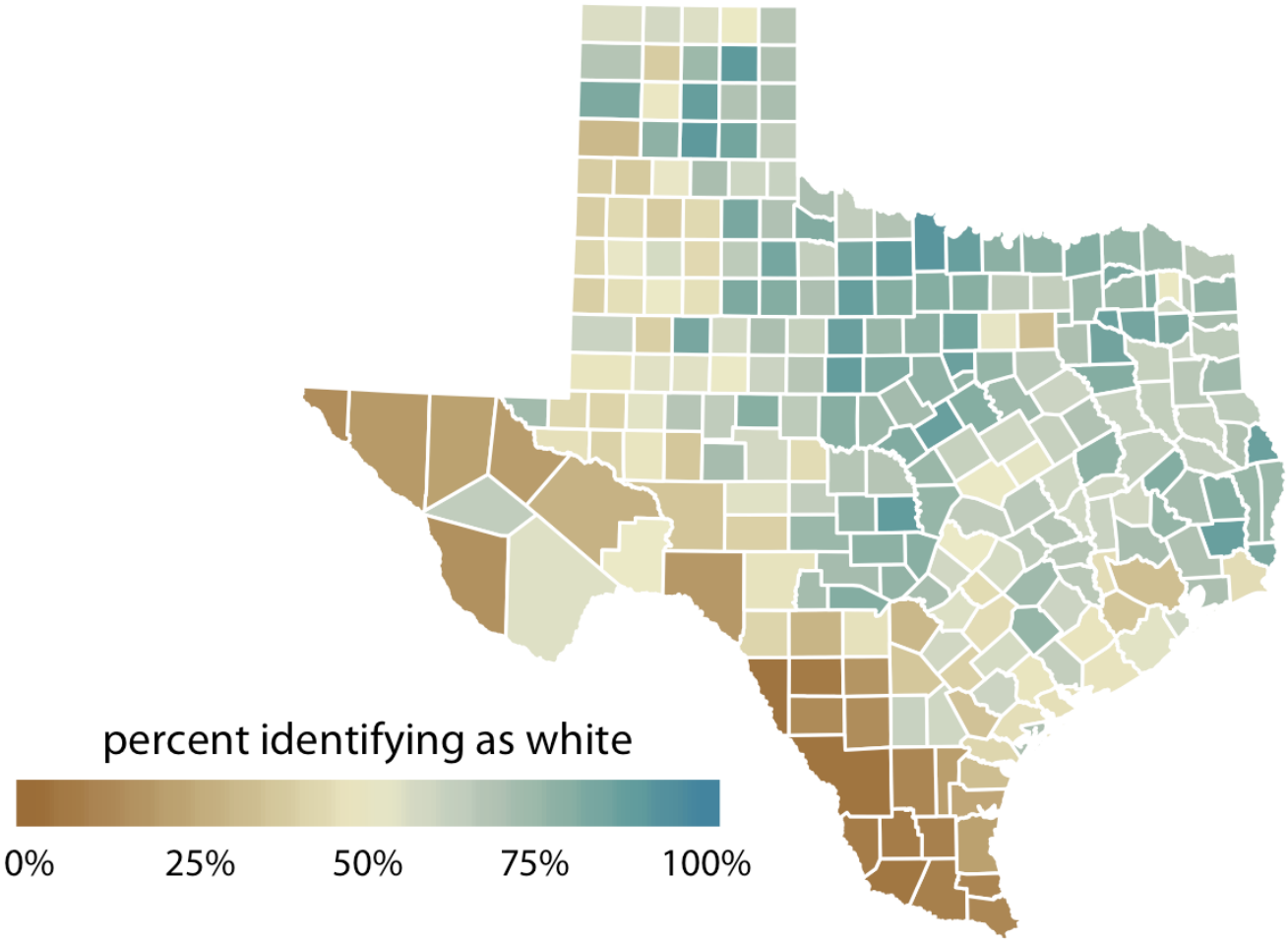
ColorBrewer PiYG



Blue-Red

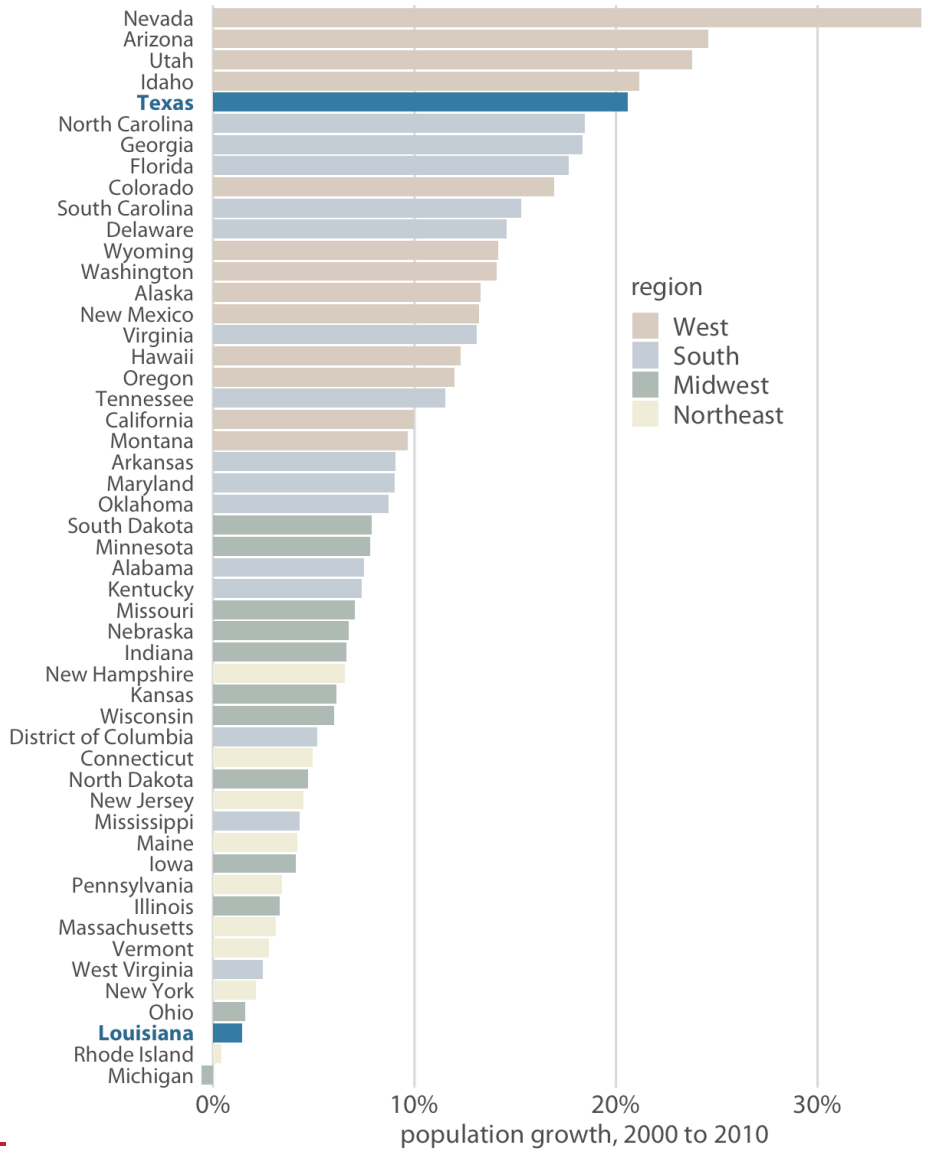


Example: Percentage of people identifying as white in Texas counties



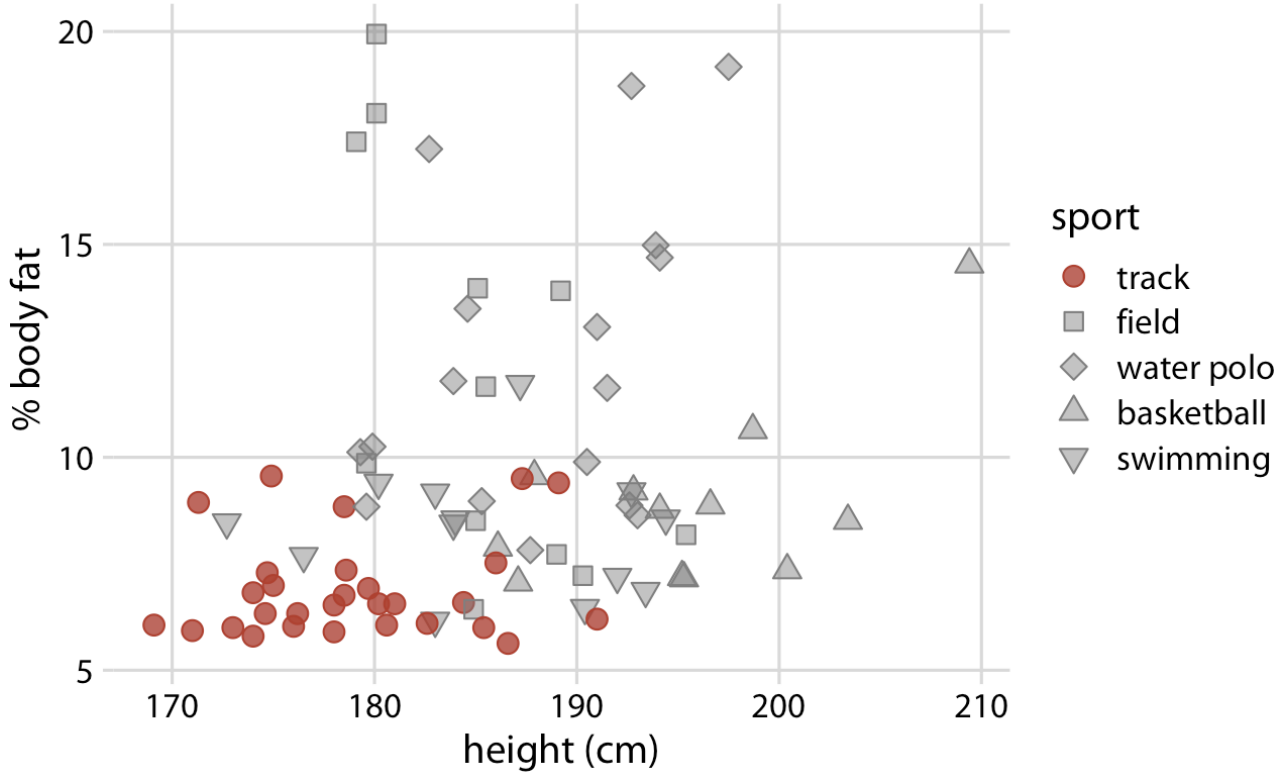
Color as a tool to highlight

- Specific categories or values in the dataset that carry key information about the story we want to tell.
- Color these figure elements in a color or set of colors that vividly stand out against the rest of the figure.



Example

- Track athletes are among the shortest and leanest of male professional athletes participating in popular sports.

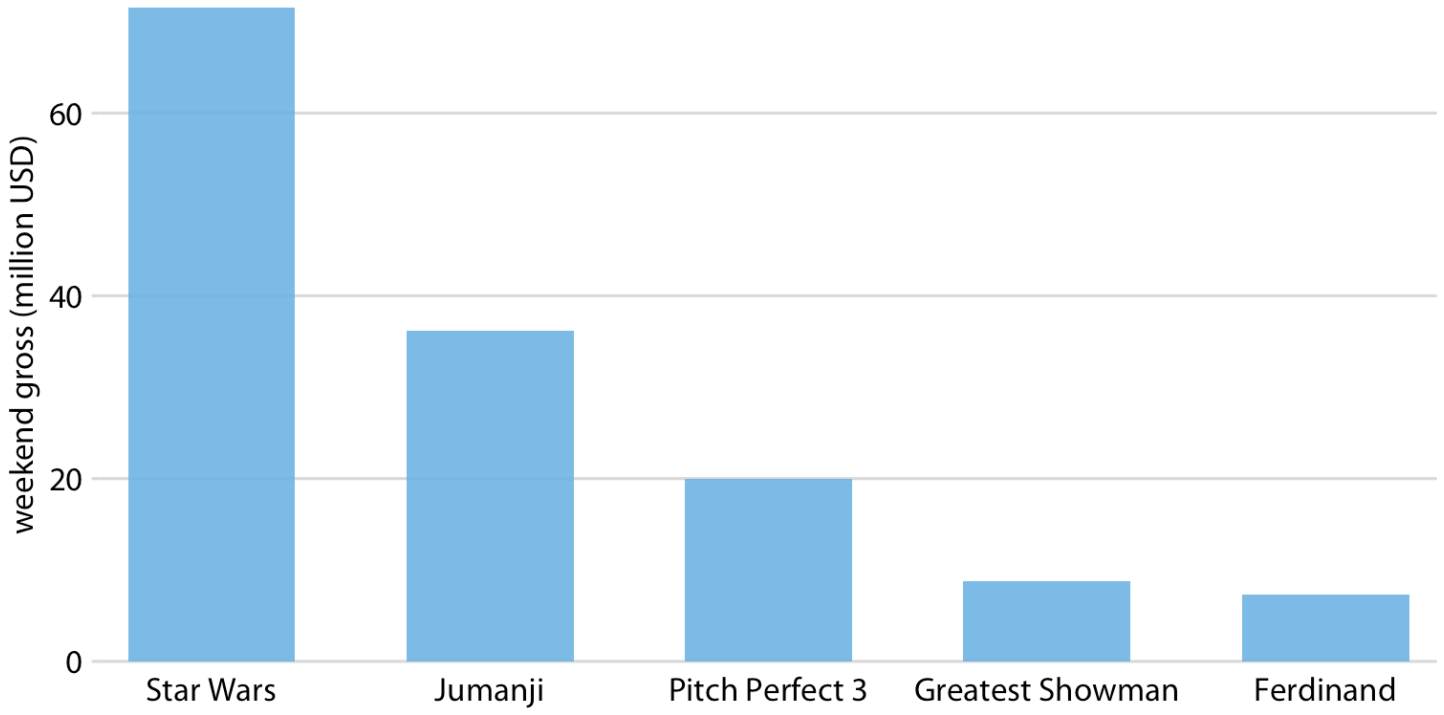


Visualizing amounts

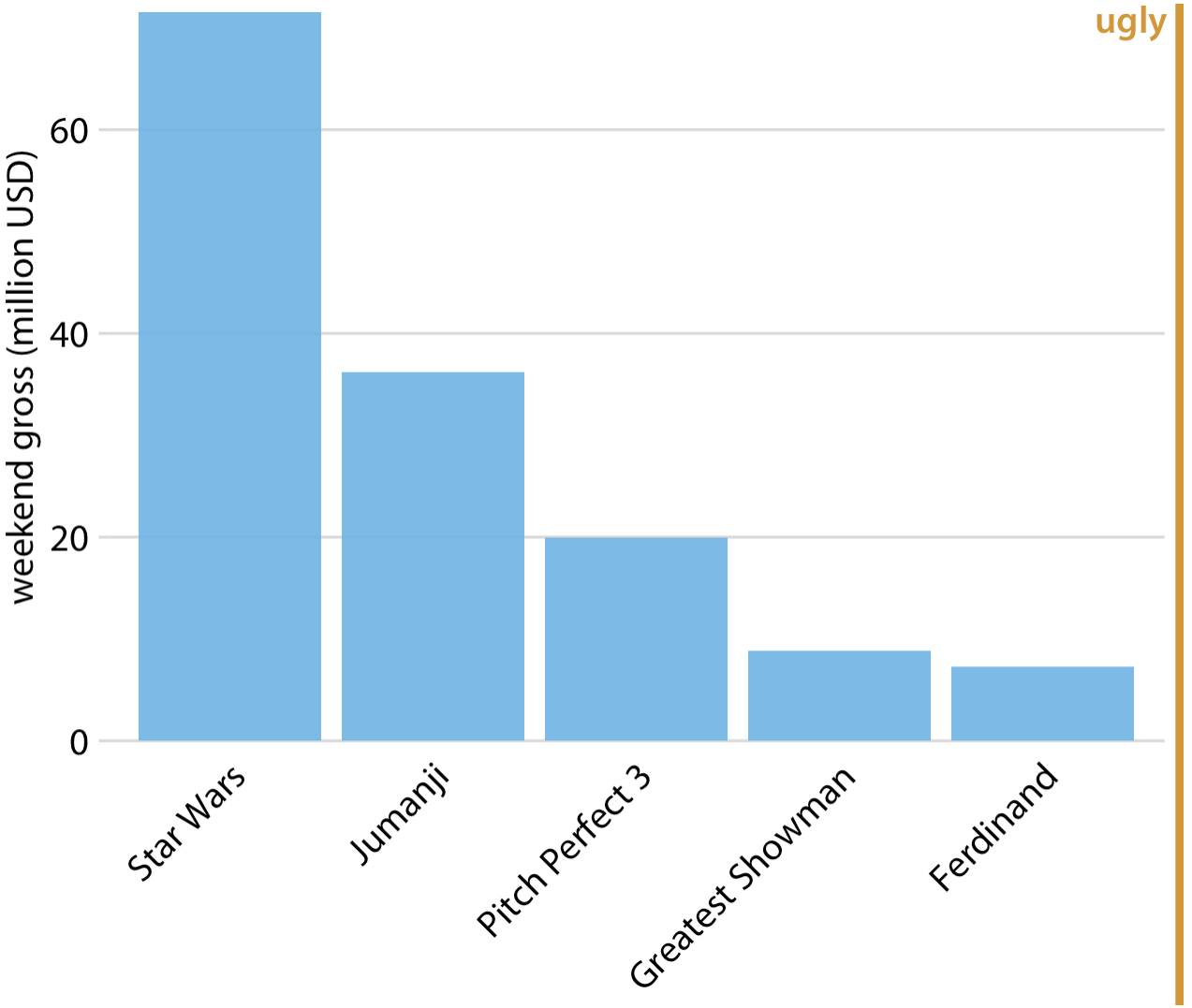
Scenario

- Visualize the magnitude of some set of numbers
  - a set of categories (e.g., brands of cars, cities, or sports) and a quantitative value for each category.

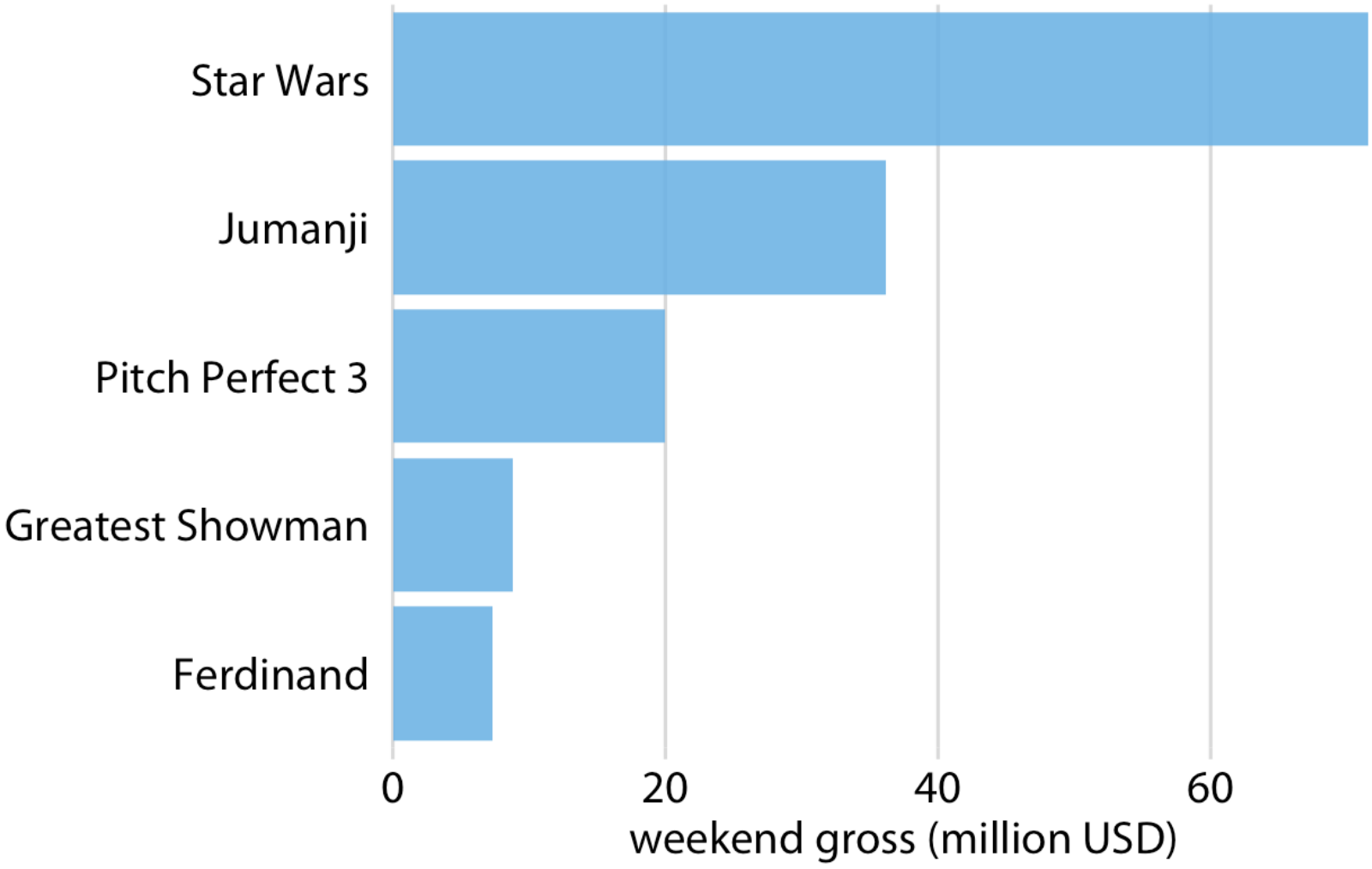
Example: Highest-grossing movies for the weekend of December 22–24, 2017



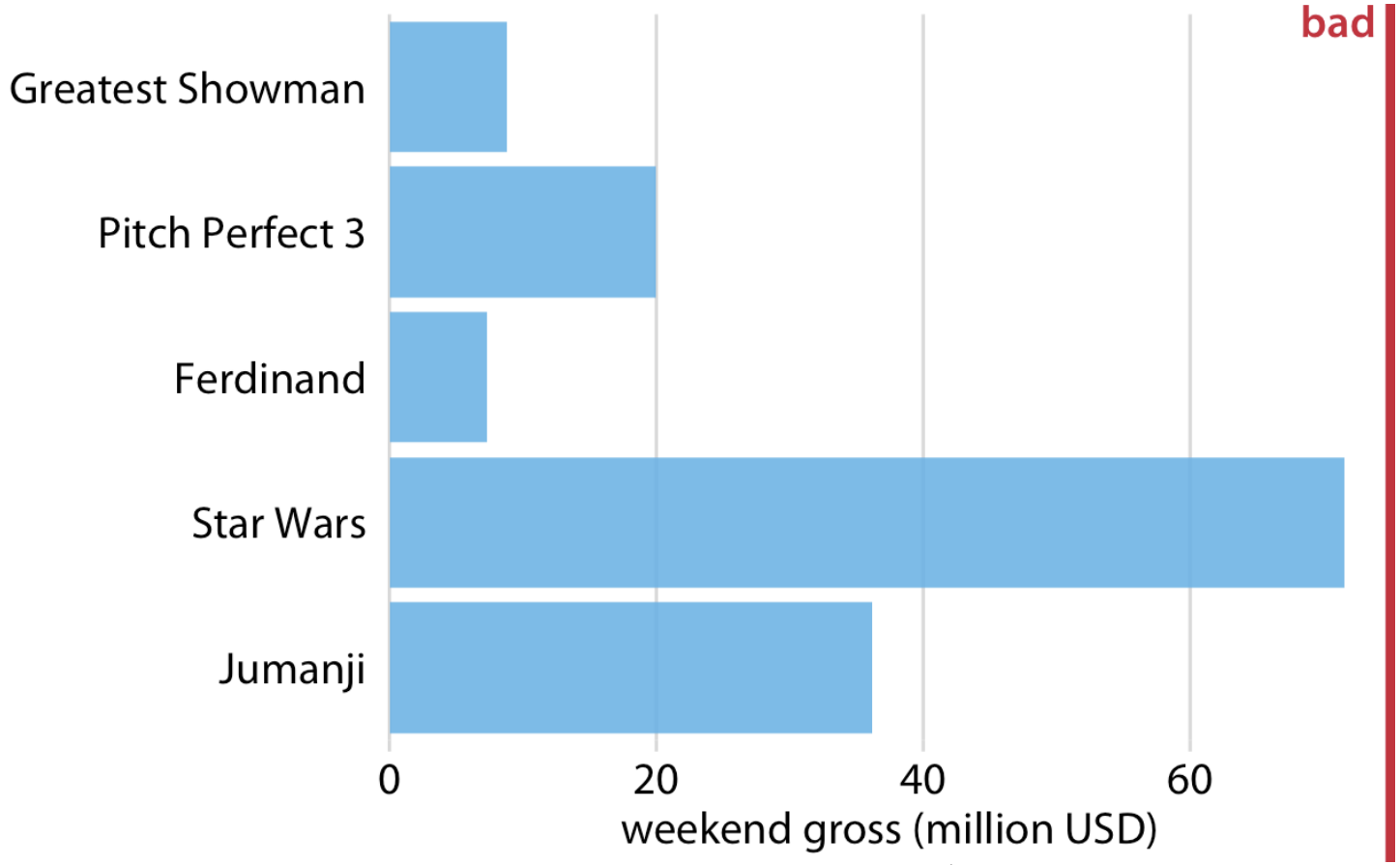
A bar plot with rotated axis tick labels



Horizontal bar plot

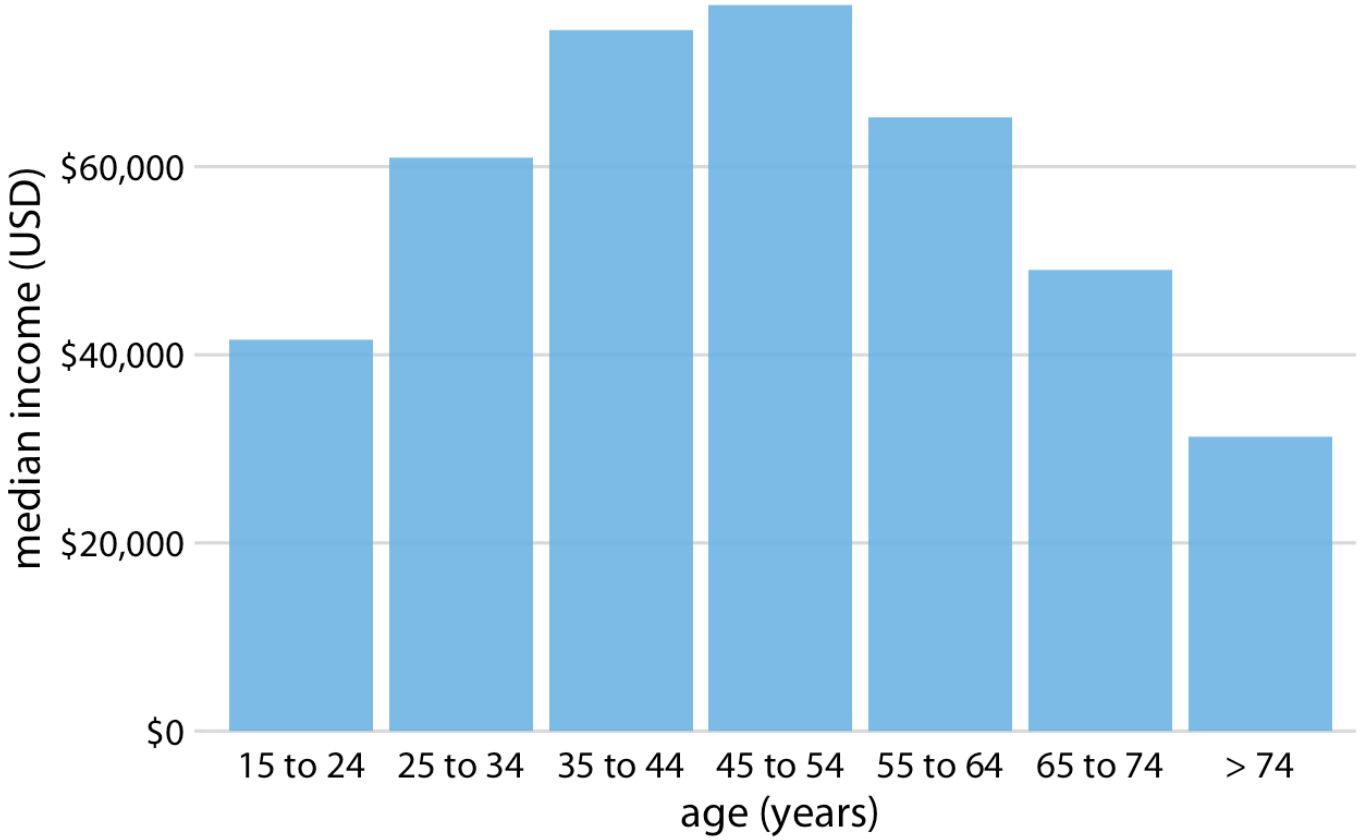


Example: Highest-grossing movies for the weekend of December 22–24, 2017



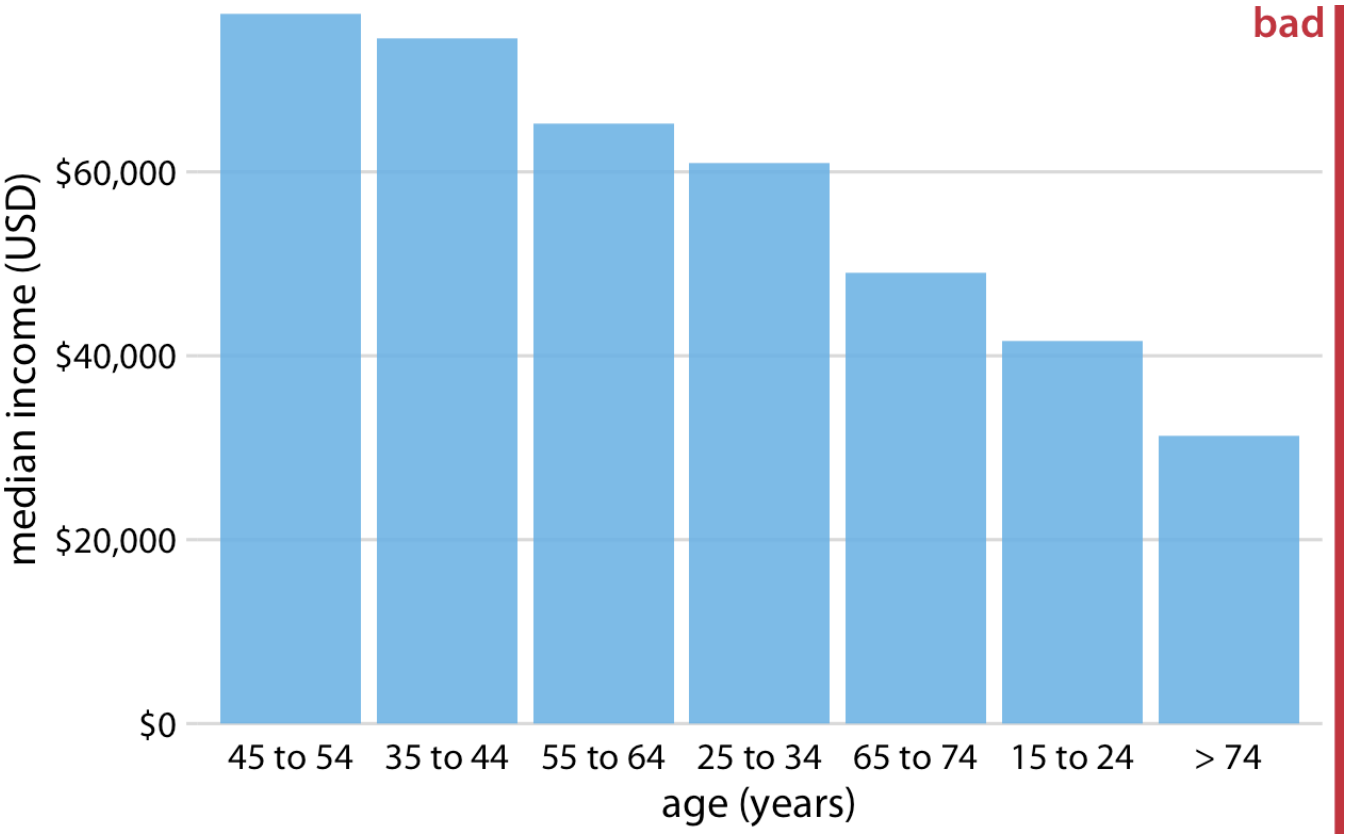
- This arrangement of bars is arbitrary, doesn't serve a meaningful purpose, and makes the resulting figure much less intuitive.

Example: 2016 median US annual household income versus age group



- Only rearrange bars when there is no natural ordering to the categories the bars represent.

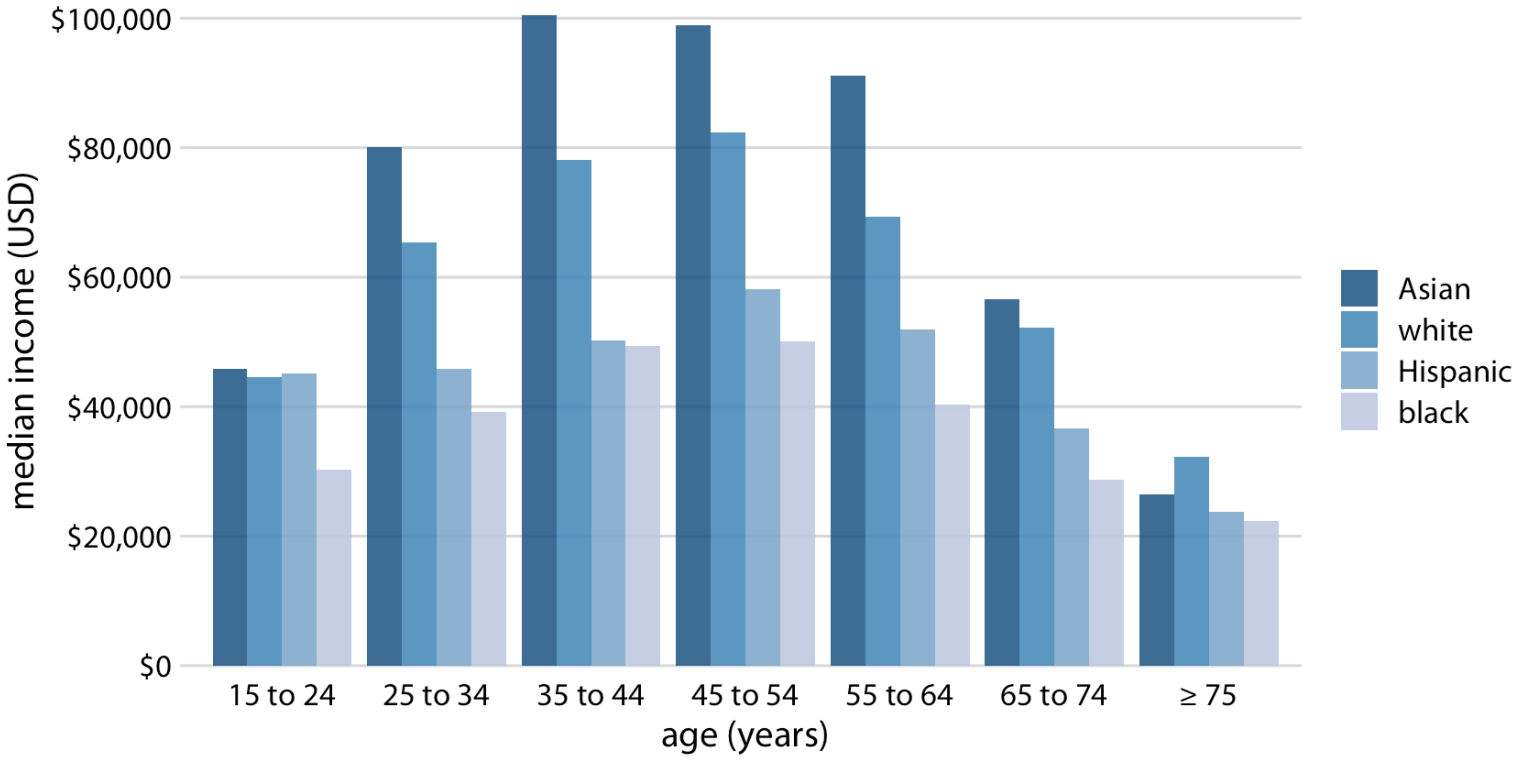
Example: 2016 median US annual household income versus age group



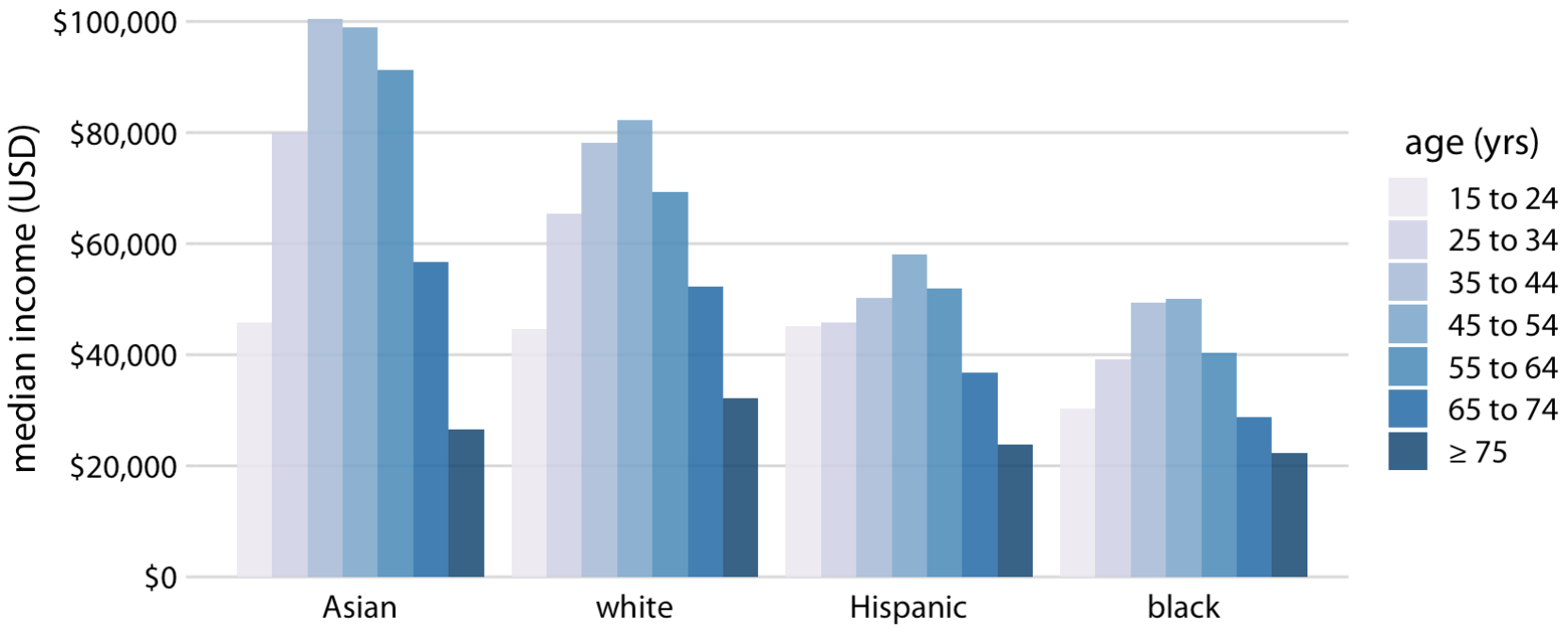
- While this order of bars looks visually appealing, the order of the age groups is now confusing

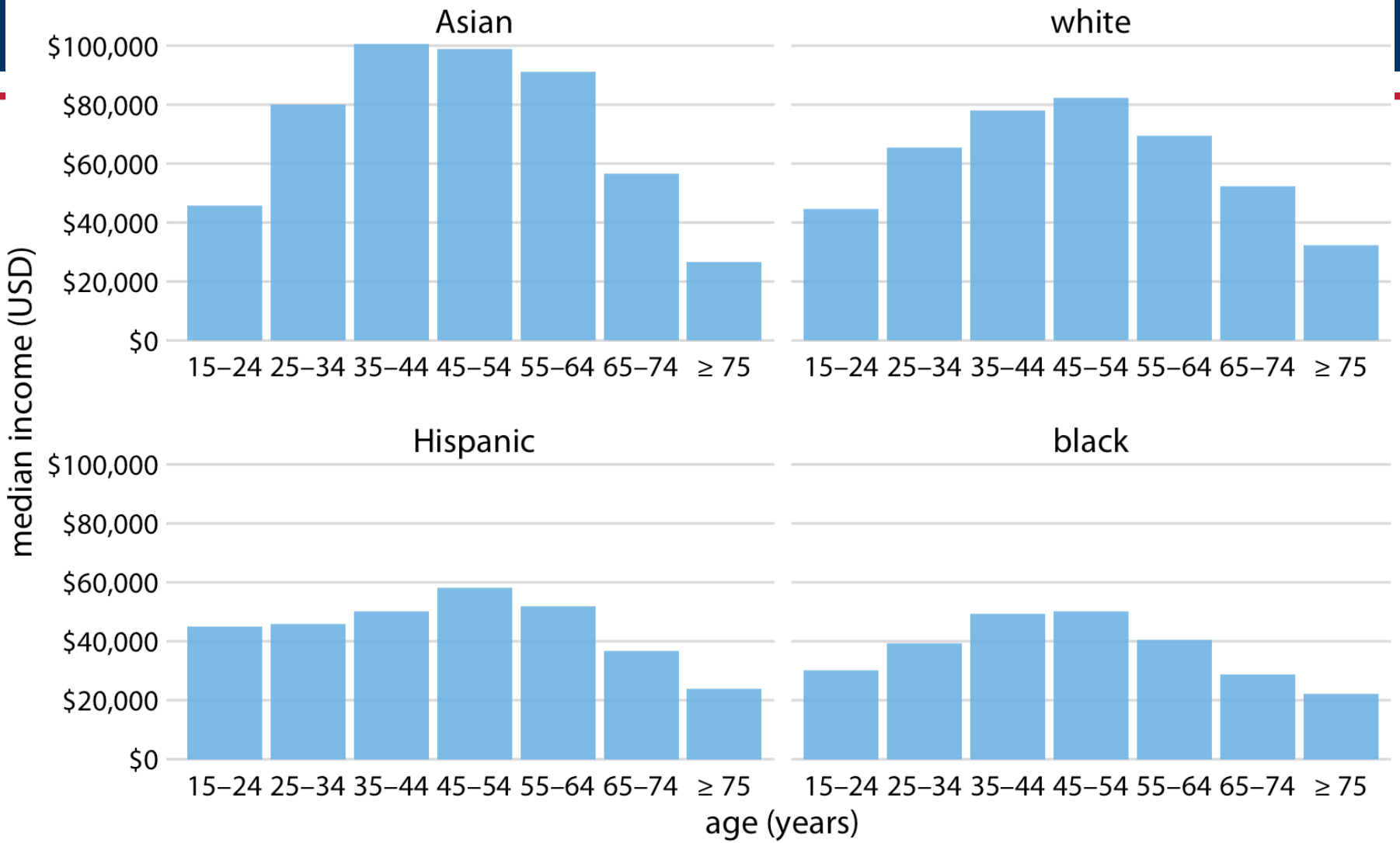
Grouped and Stacked Bars

- Frequently, we are interested in > two categorical variables at the same time.
  - Grouped bar plot (groups of bars side-by-side)



Example: 2016 median US annual household income versus age group and race



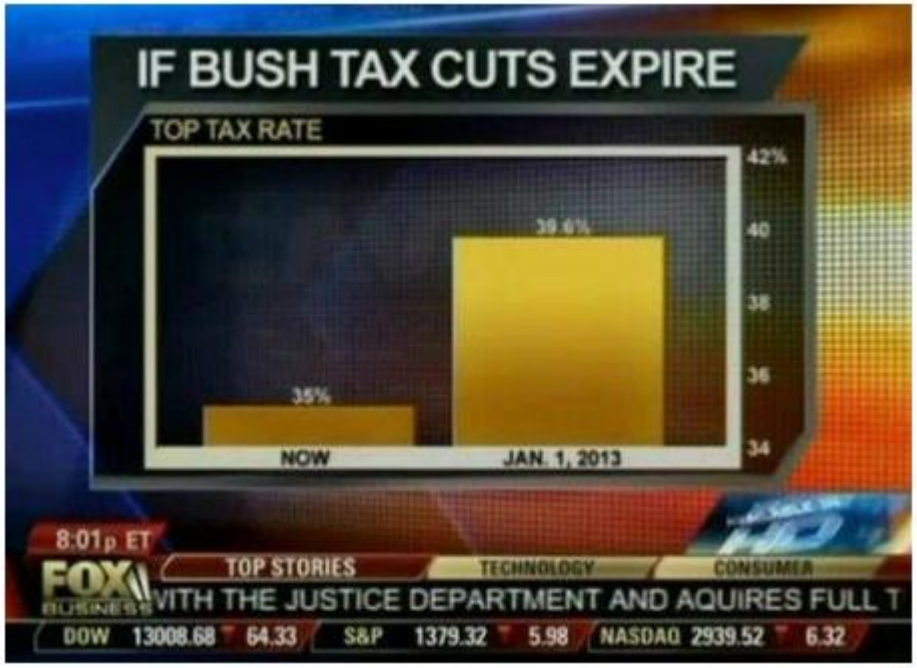


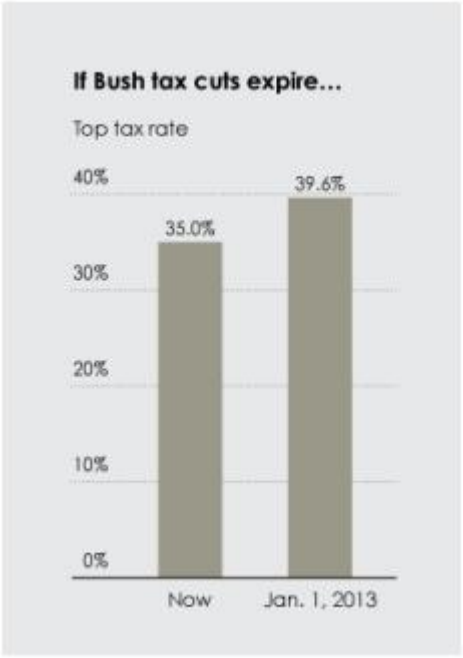
- Show the data as four separate regular bar plots. This choice has the advantage that we don't need to encode either categorical variable by bar color

Stack bars on top of each other

- Useful when the sum of the amounts represented by the individual stacked bars is a meaningful amount.
  - Numbers of female and male passengers on the Titanic traveling in 1st, 2nd, and 3rd class

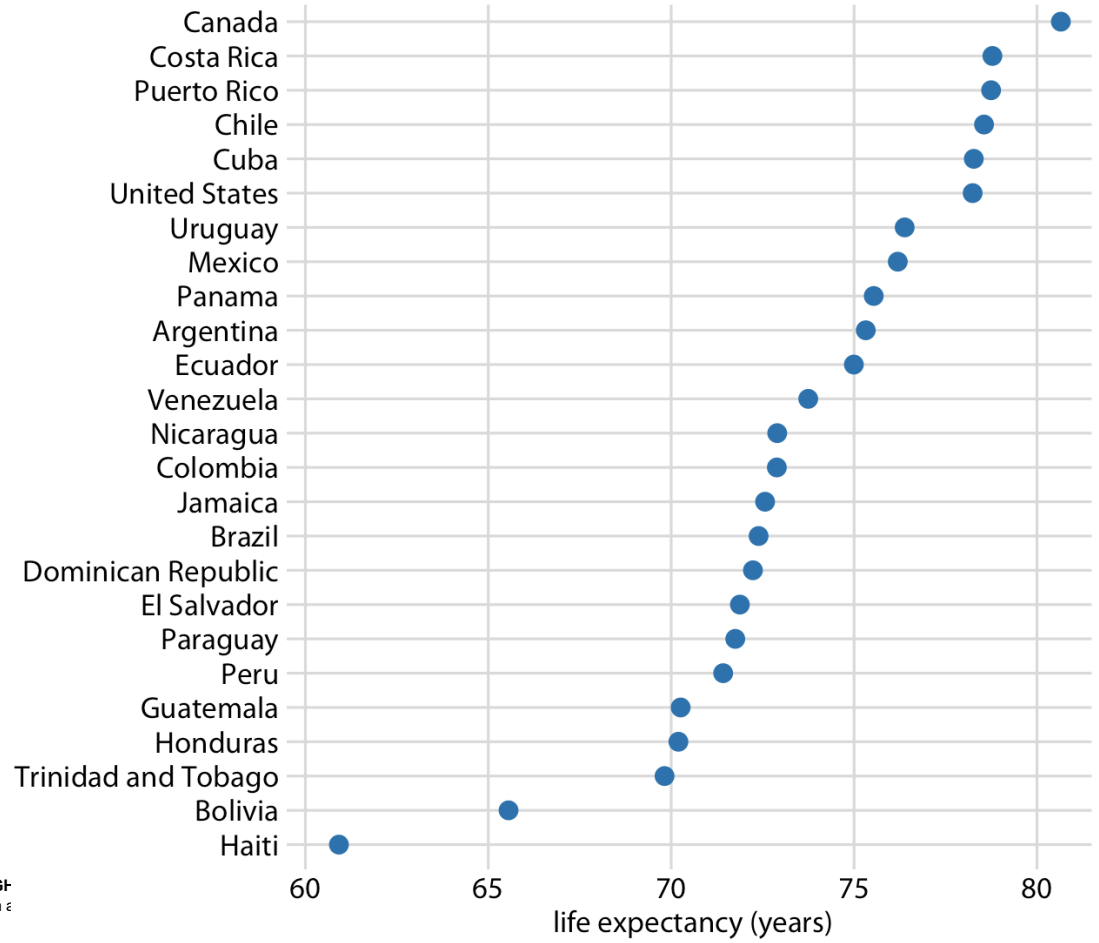
Bar chart: Baseline



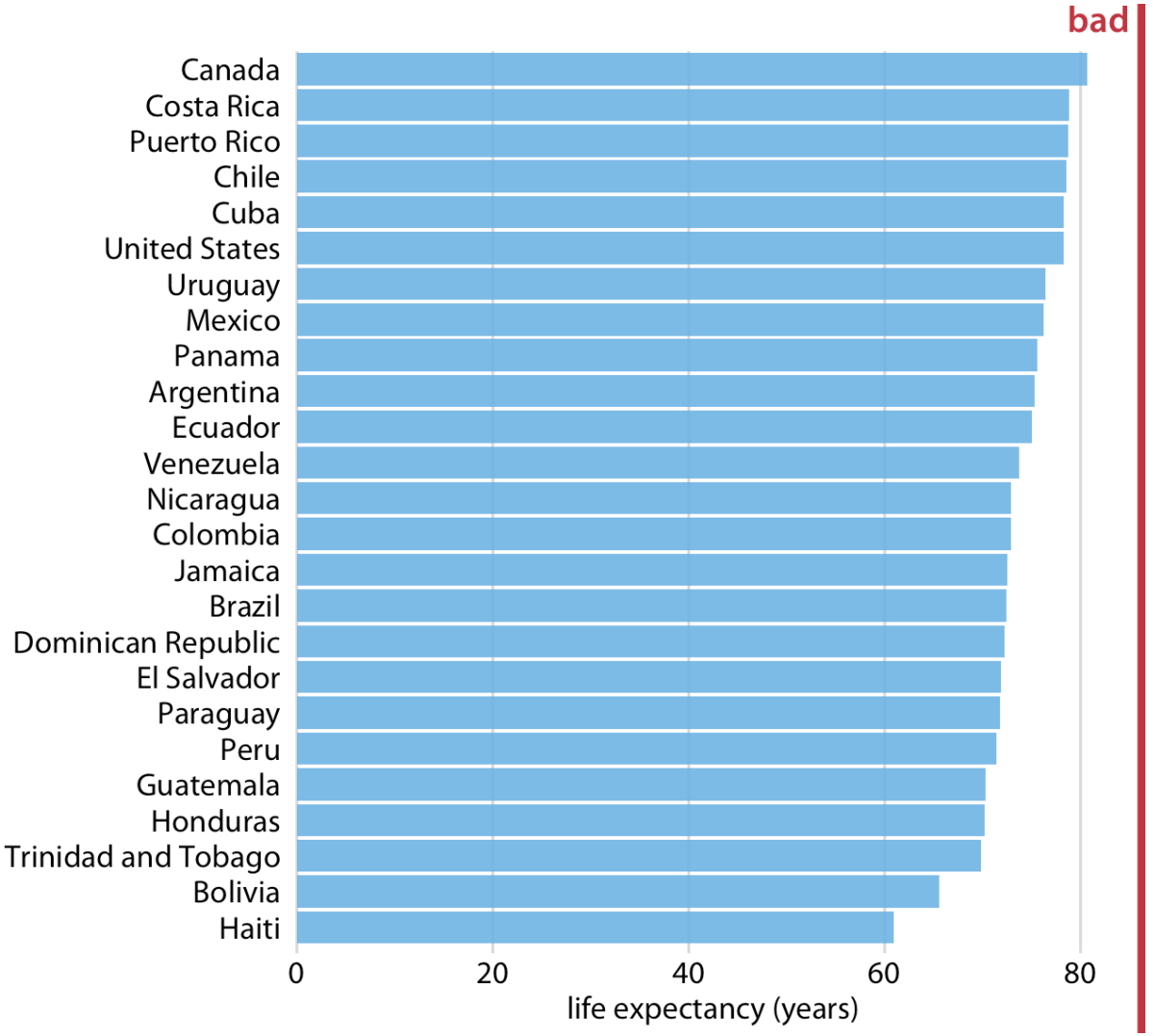


Dot Plots

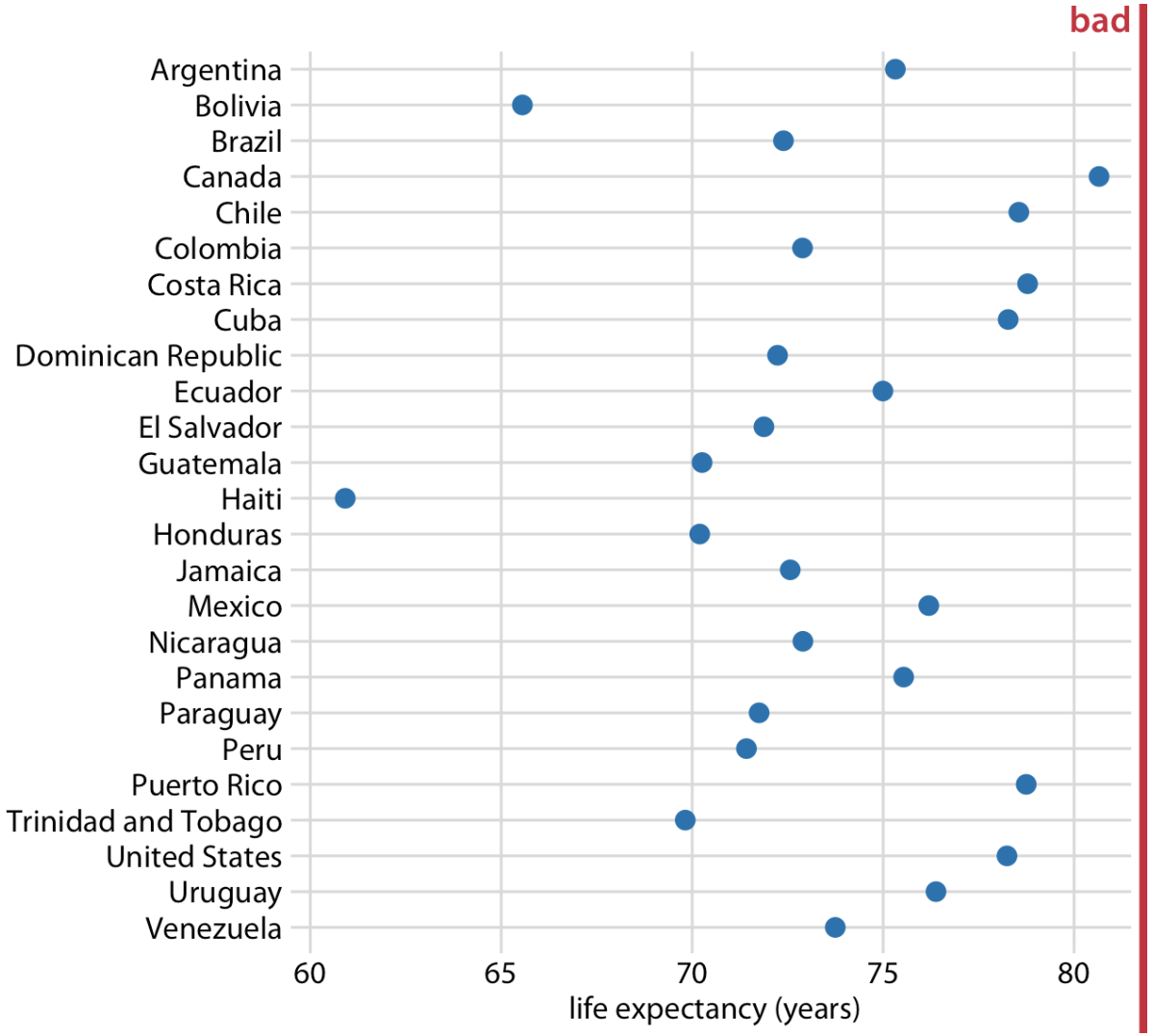
- Example: Life expectancies of countries in the Americas, for the year 2007



Example: Life expectancies of countries in the Americas, for the year 2007

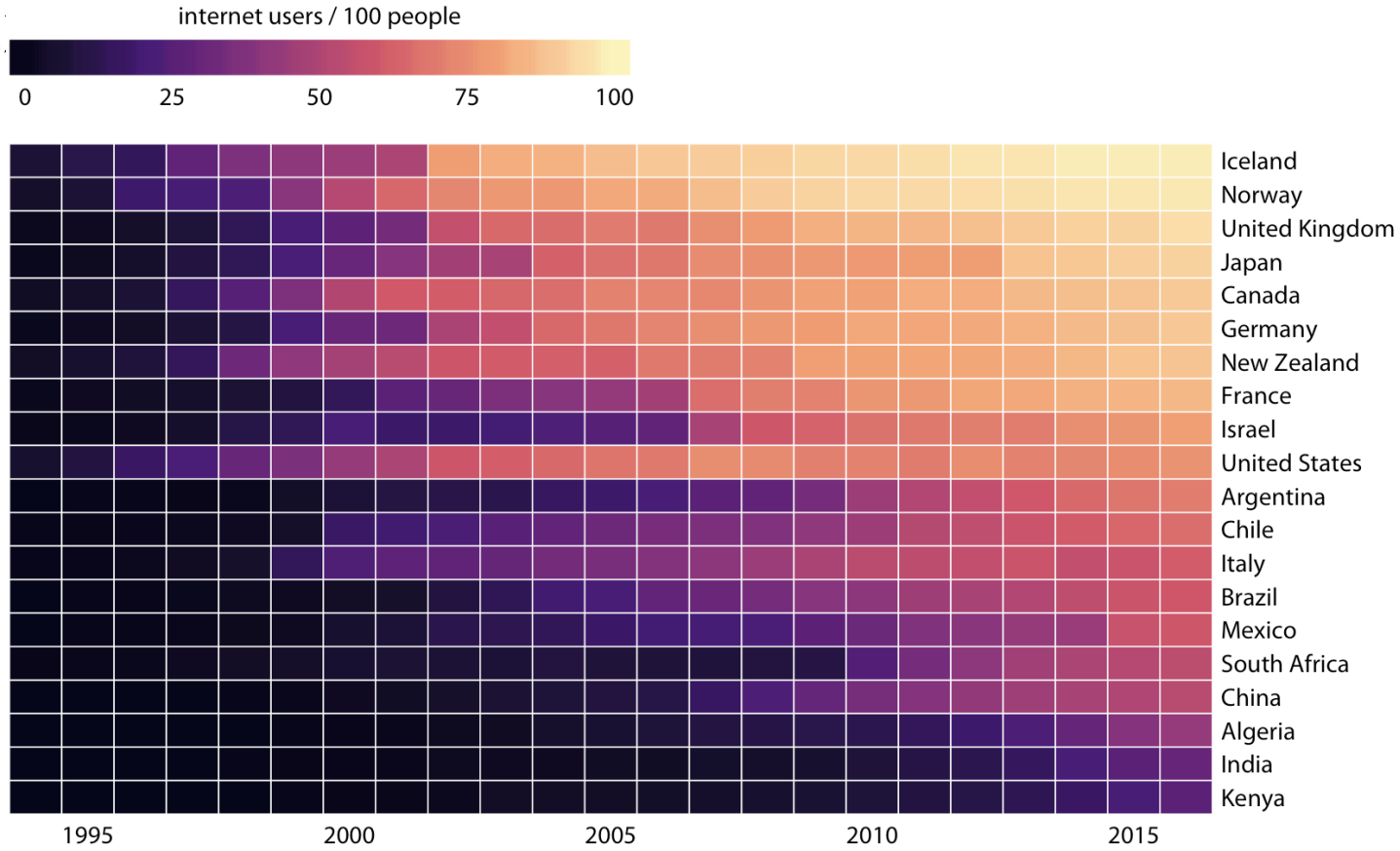


Pay attention to the ordering of the data values



Heatmap: map data values onto colors

- Harder to determine the exact data values shown.
- E



Visualizing distributions

Histograms and Density Plots

Visualizing a single distribution

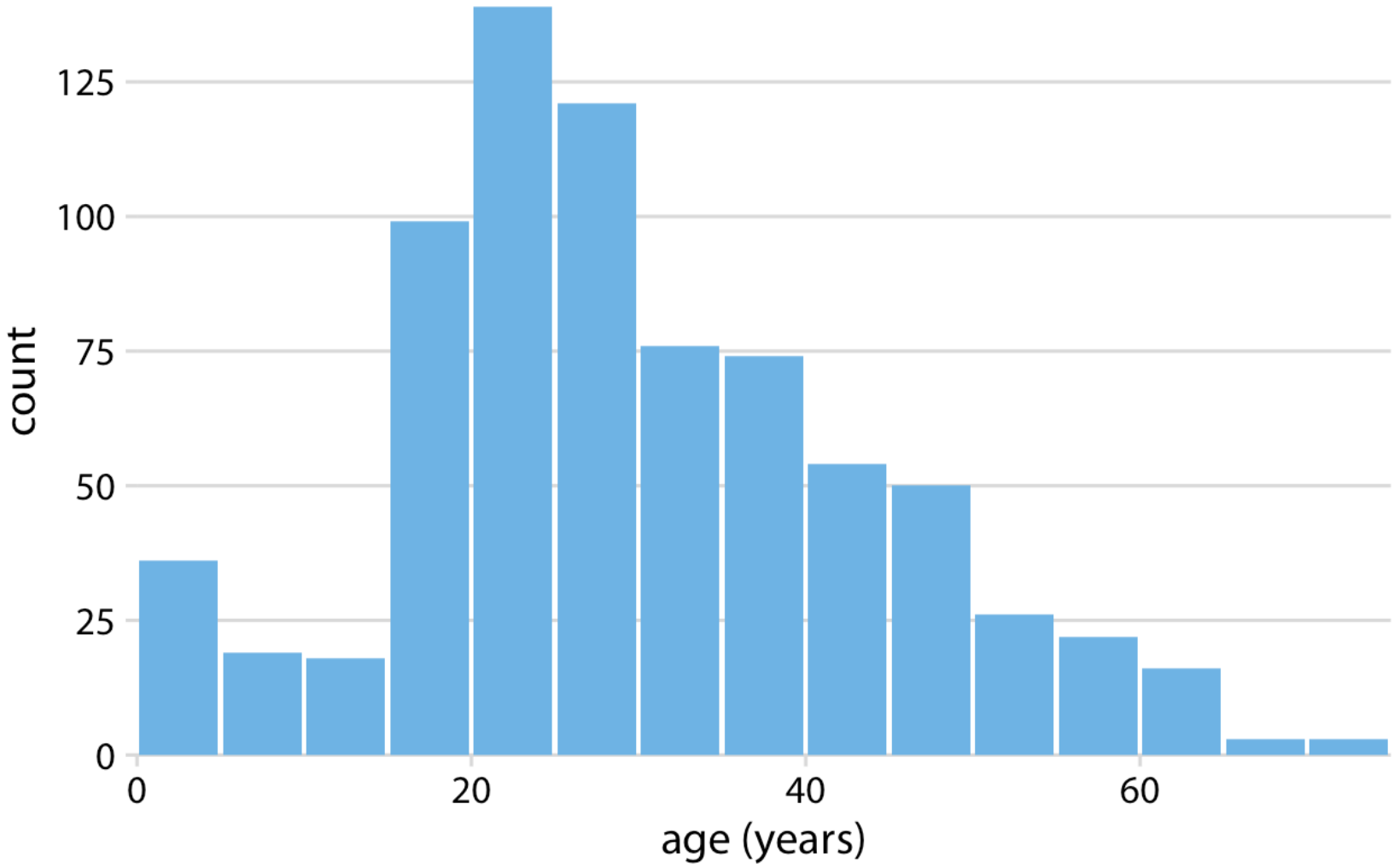
- To understand how a particular variable is distributed in a dataset.

Age range	Count
0–5	36
6–10	19
11–15	18
16–20	99
21–25	139
26–30	121

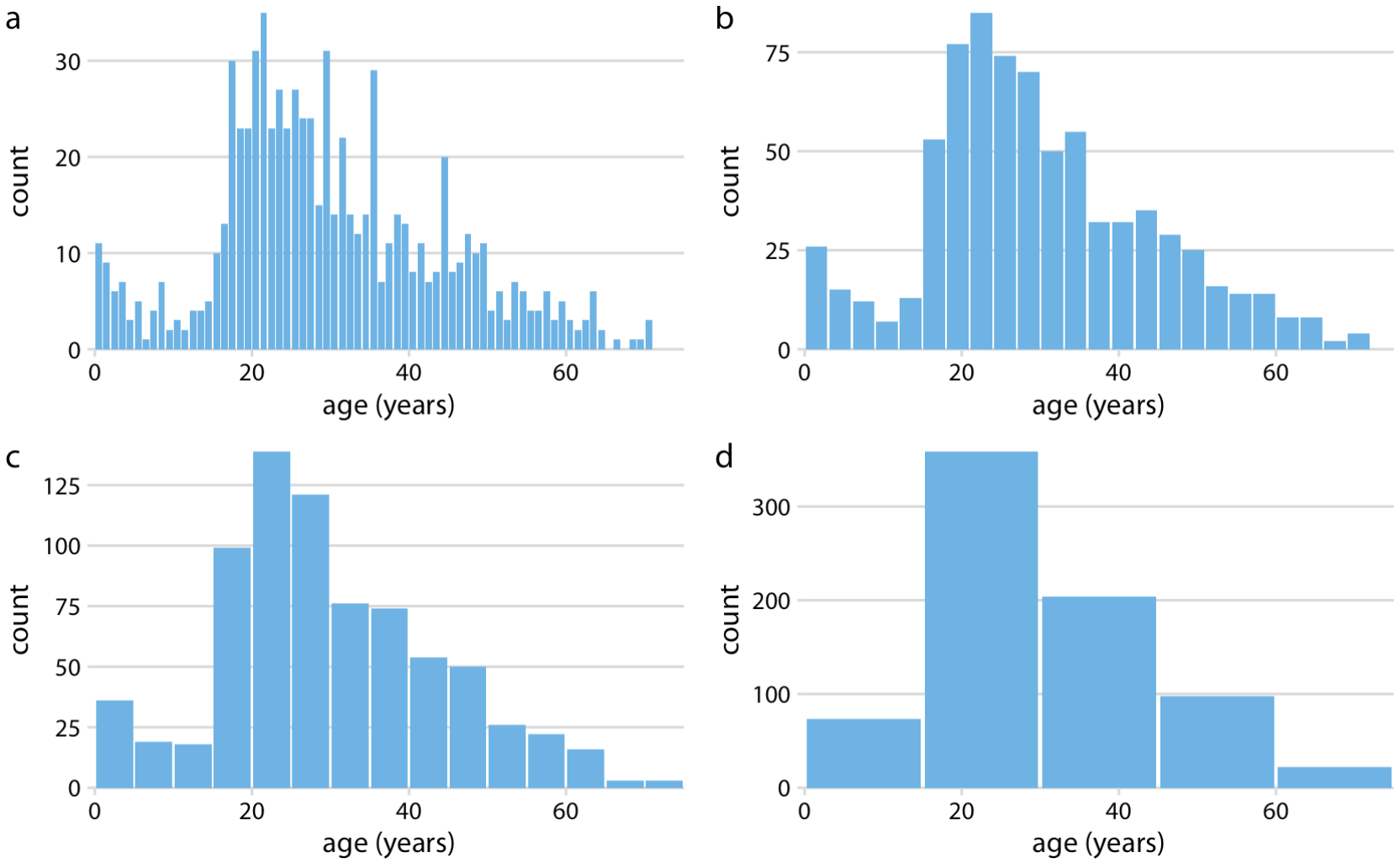
Age range	Count
31–35	76
36–40	74
41–45	54
46–50	50
51–55	26
56–60	22

Age range	Count
61–65	16
66–70	3
71–75	3

Example: Histogram of the ages of Titanic passengers



Histograms depend on the chosen bin width

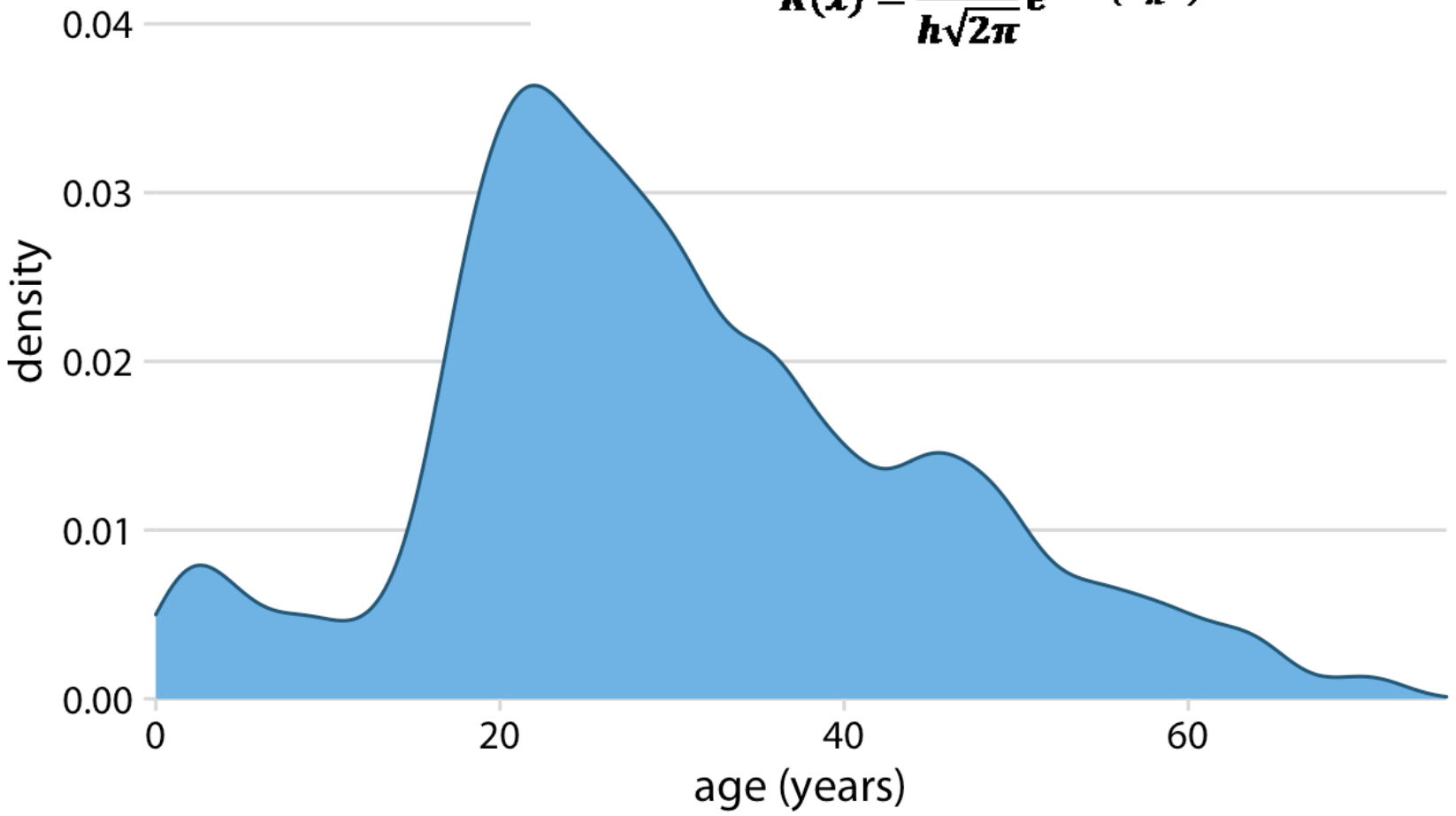


Density plots

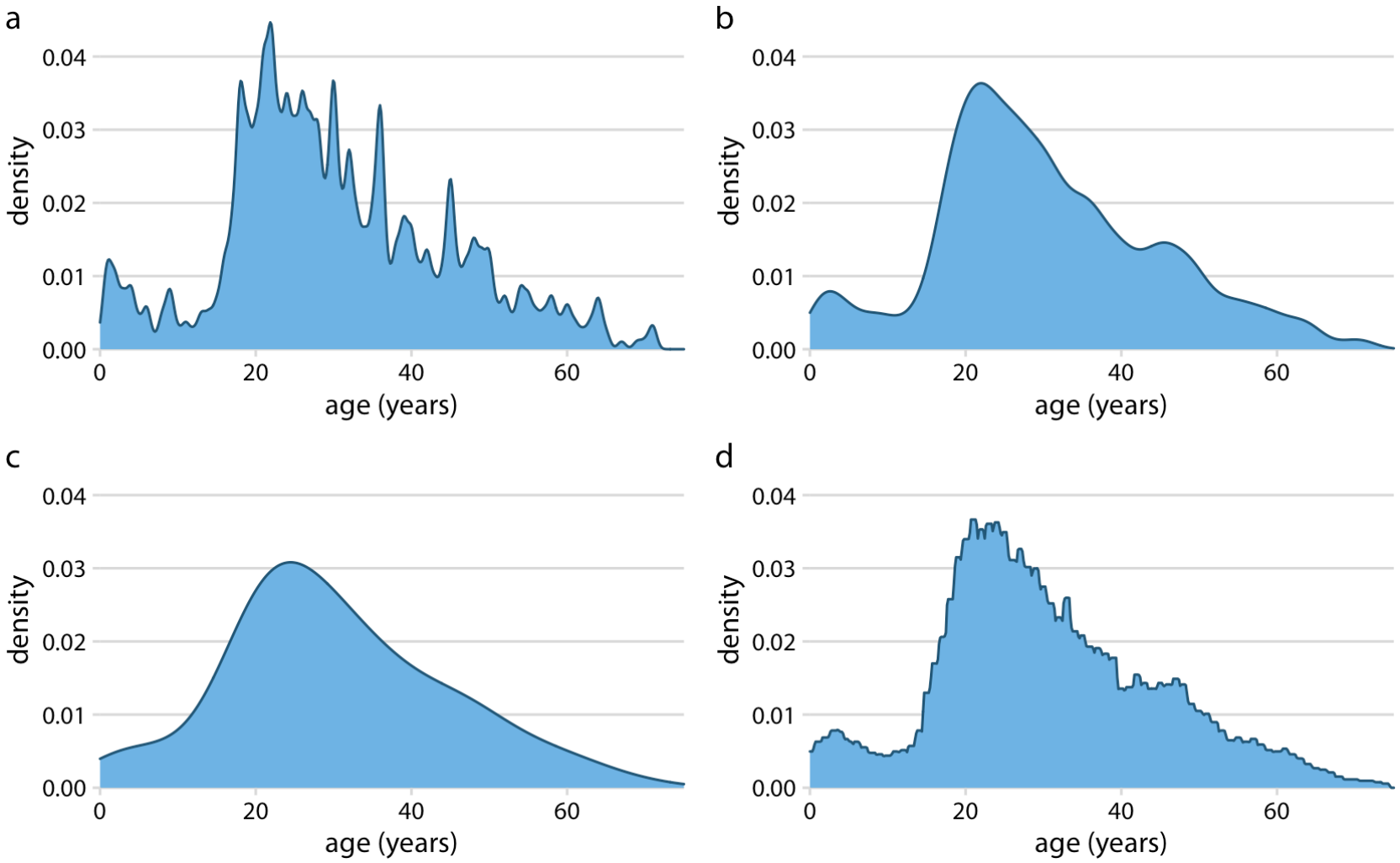
- Attempt to visualize the underlying probability distribution of the data by drawing an appropriate continuous curve
  - Estimated from the data
  - The most used method is called kernel density estimation

Example: Kernel density estimate of the age distribution of passengers on the Titanic

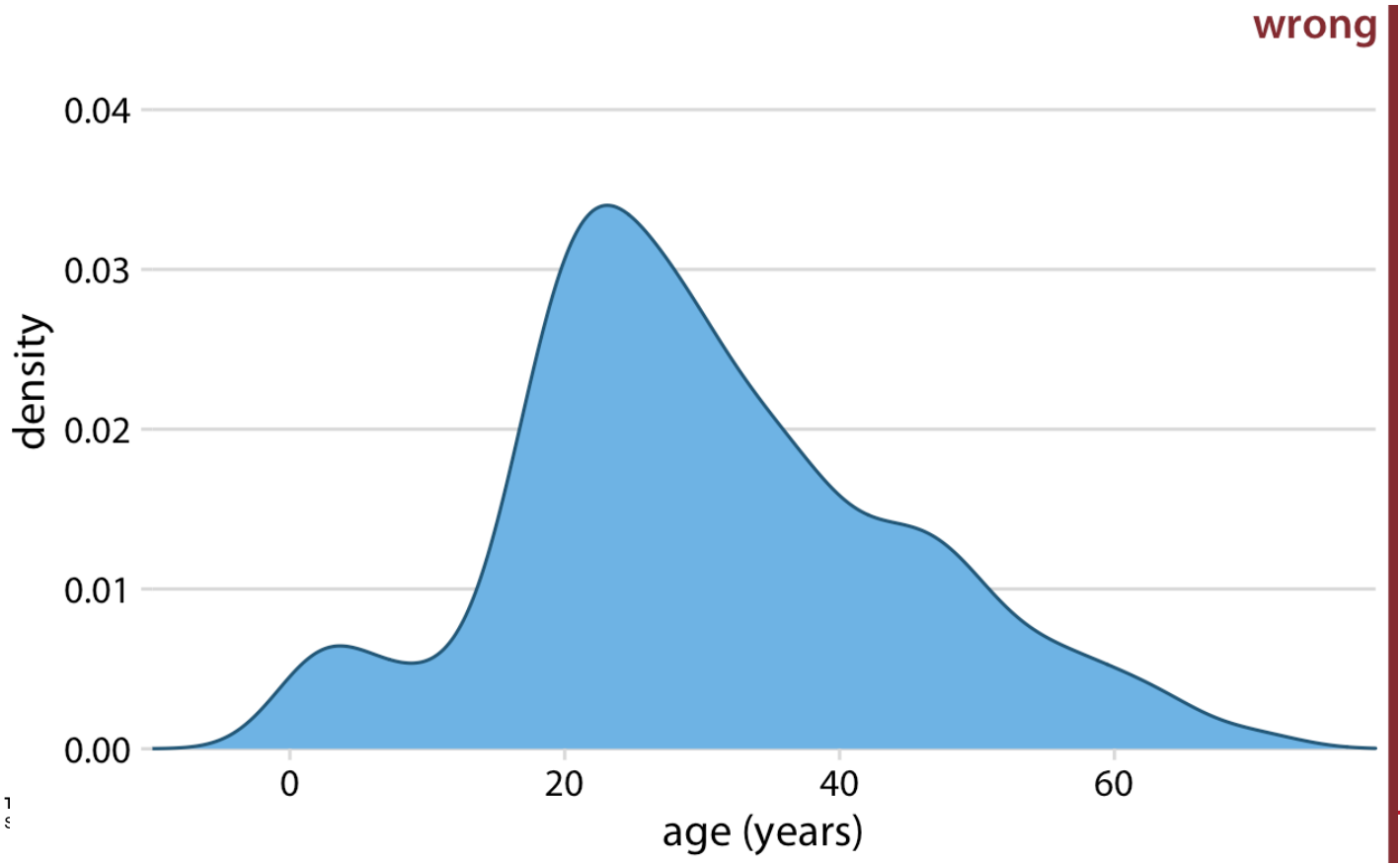
$$K(x) = \frac{1}{h\sqrt{2\pi}} e^{-0.5\left(\frac{x-xi}{h}\right)^2}$$



Kernel density estimates depend on the chosen kernel and bandwidth

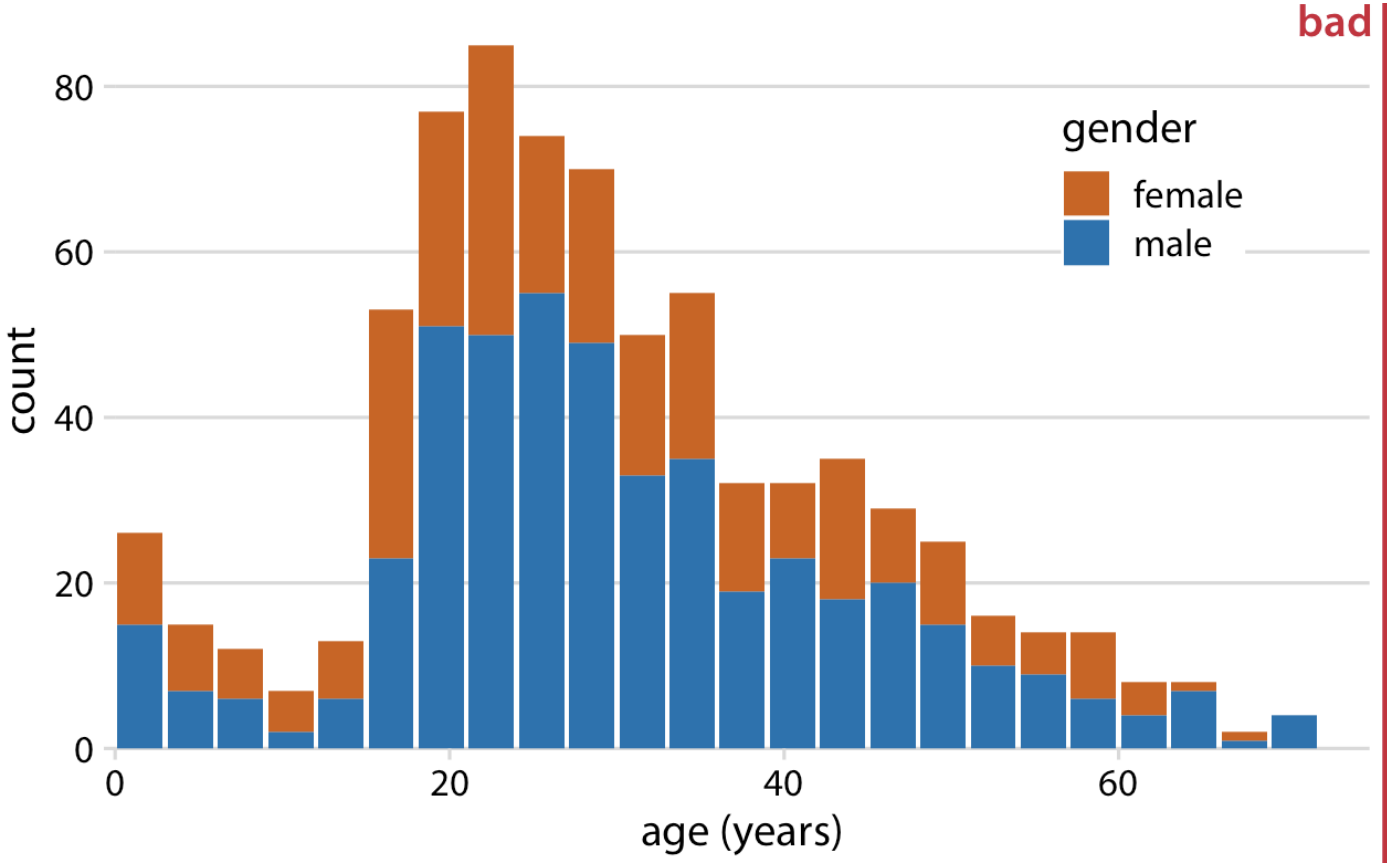


- Kernel density estimates can extend the tails of the distribution into areas where no data exists, and no data is even possible.



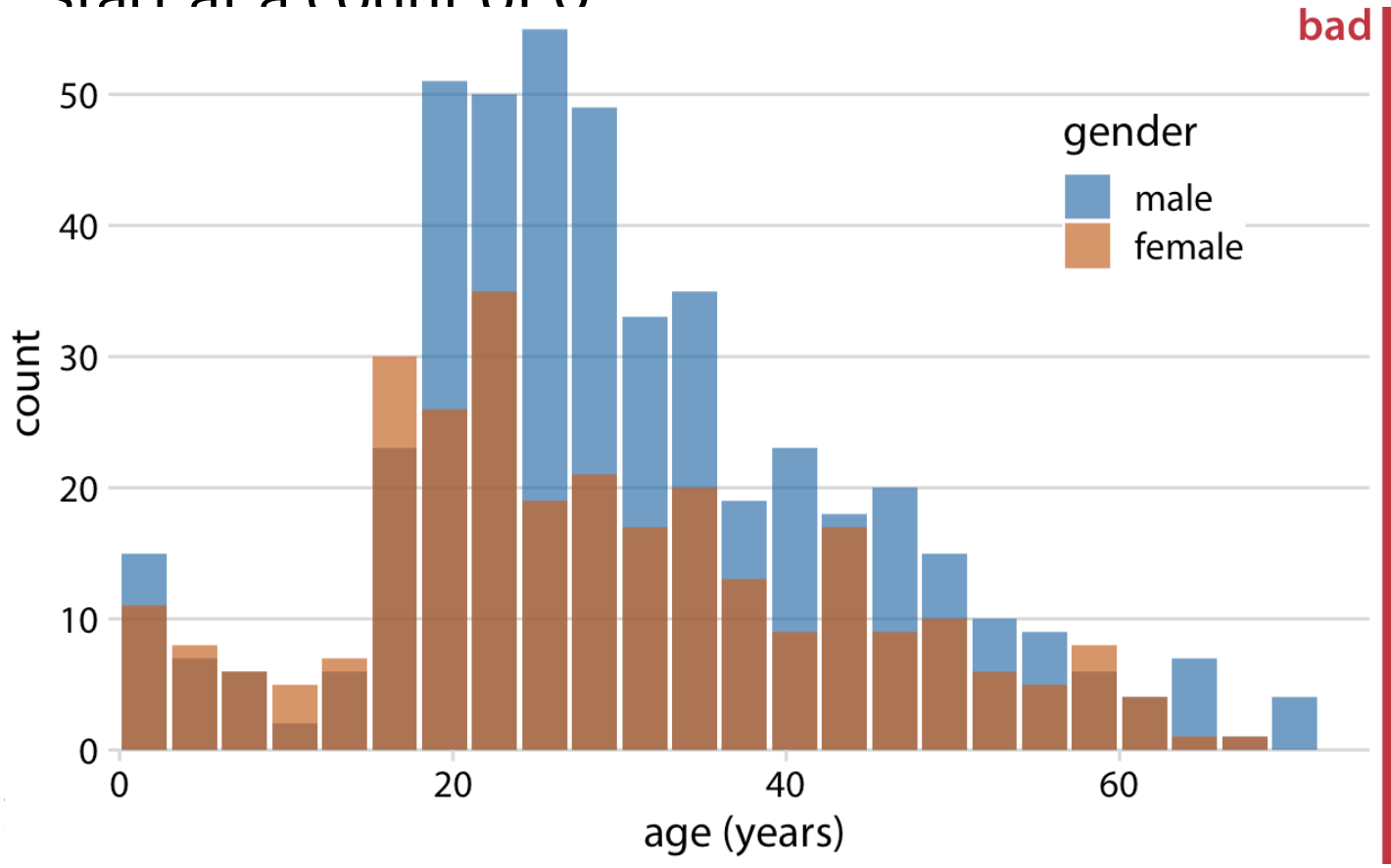
Visualizing multiple distributions at the same time

- Histogram of the ages of Titanic passengers stratified by gender
  - The heights of the bars representing female passengers cannot easily be compared to each other



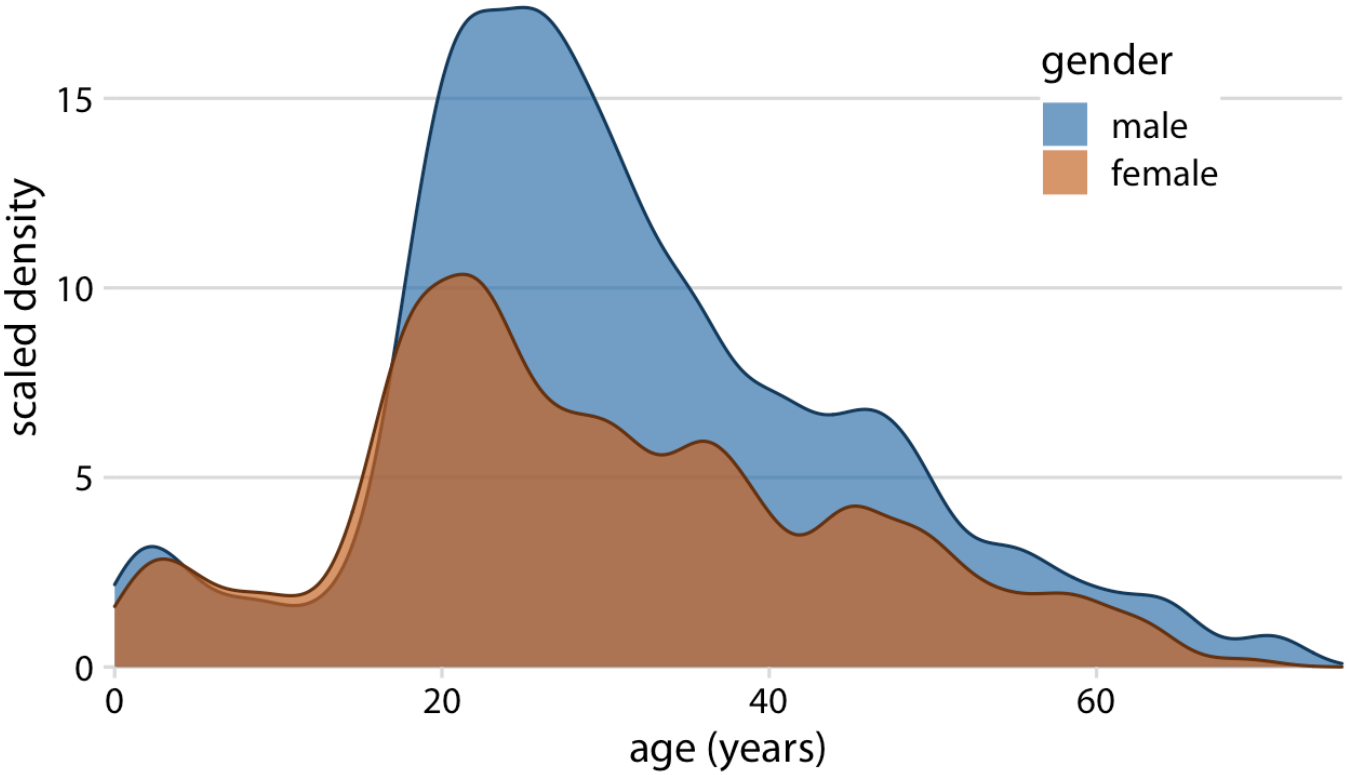
Example: Age distributions of male and female Titanic passengers

- Having all bars start at zero and making the bars partially transparent
  - There is no clear visual indication that all blue bars start at a count of 0



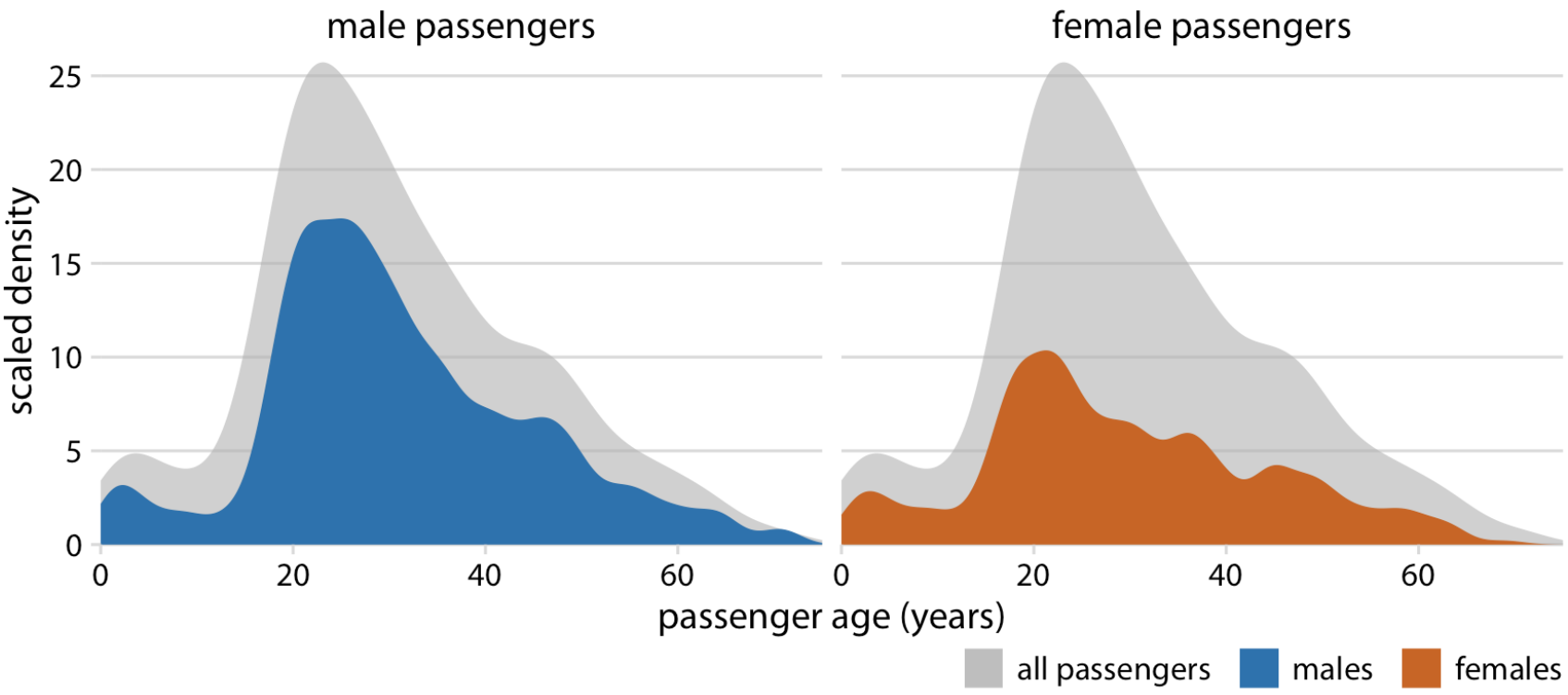
Overlapping density plots

- Continuous density lines help the eye keep the distributions separate



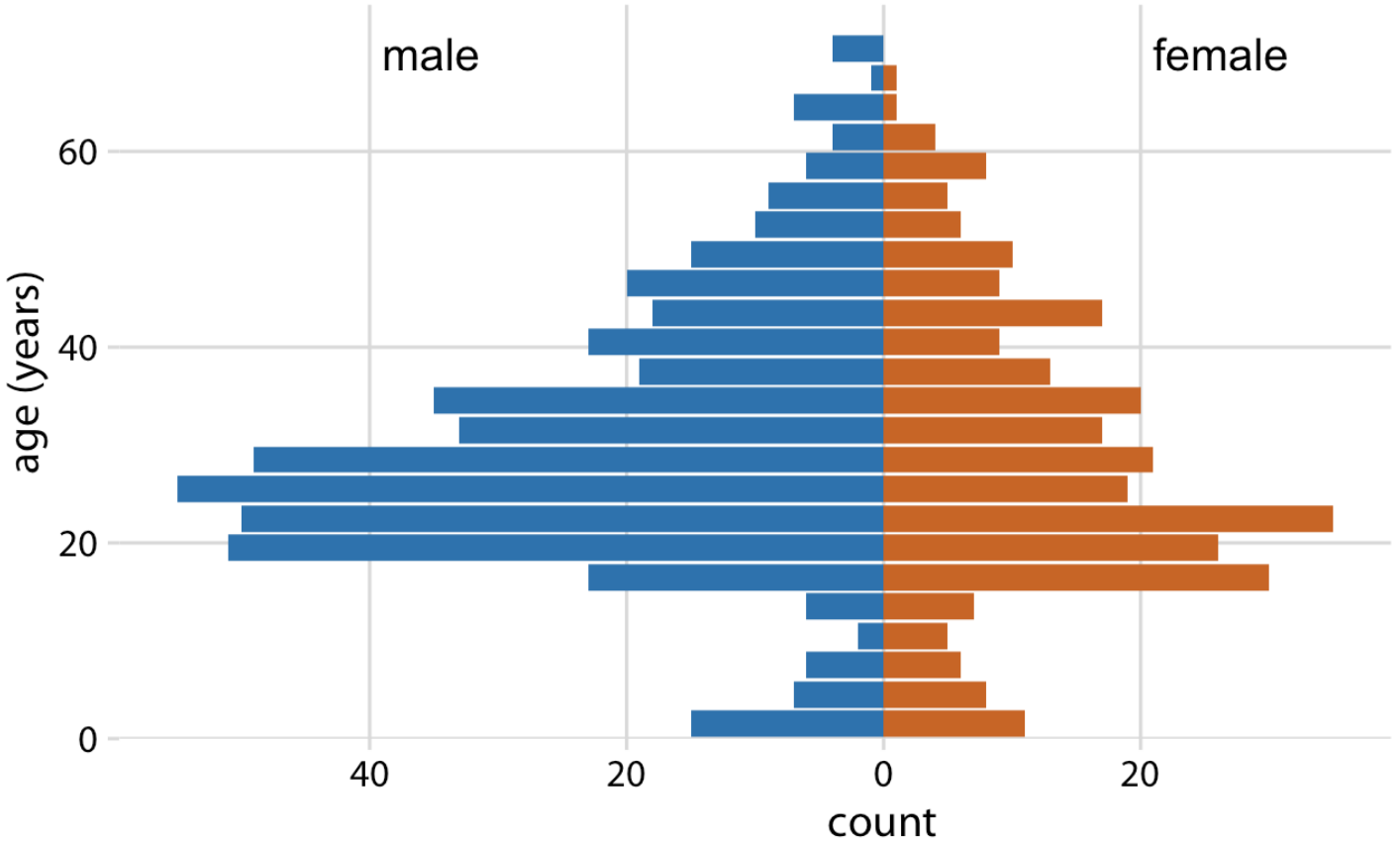
Example: Age distributions of male and female

- Show the age distributions separately, each as a proportion of the overall age distribution



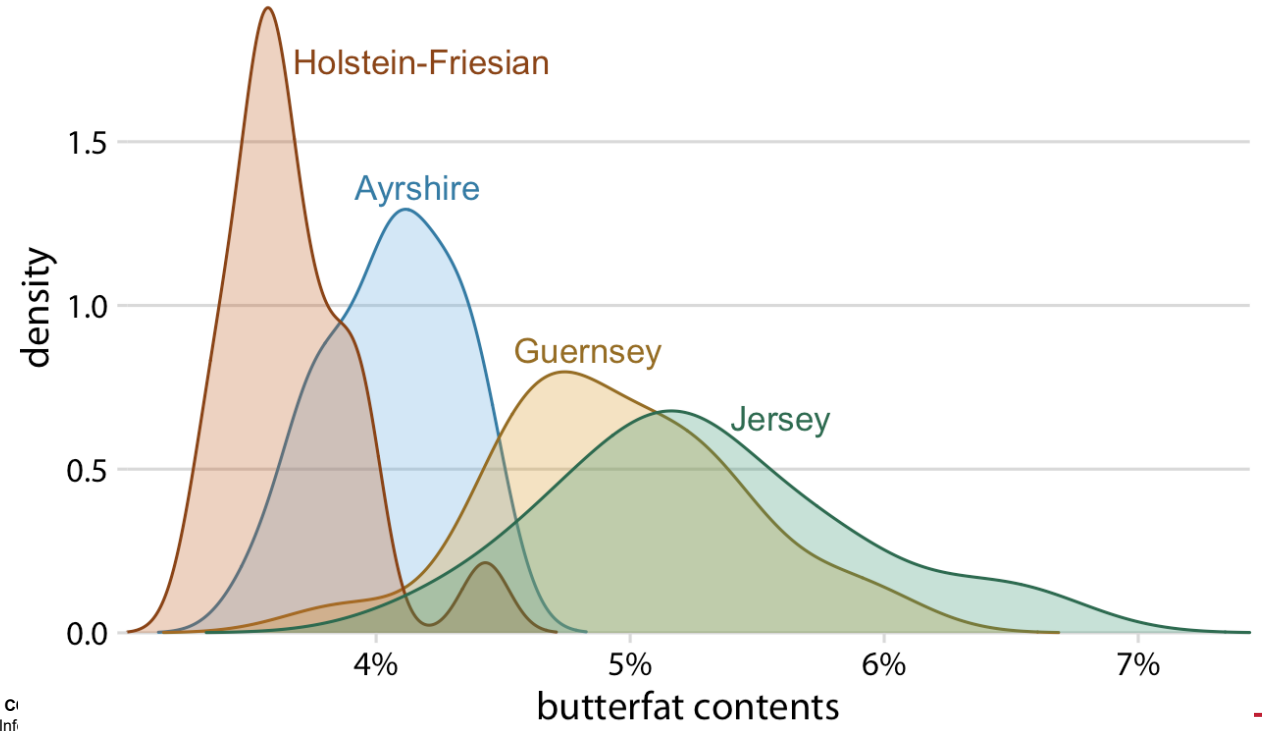
Separate histograms, rotated by 90 degrees

- The age distributions of male and female Titanic passengers visualized as an age pyramid



Density plots for multiple distributions

- Work well if the distributions are somewhat distinct and contiguous
- Example: Density estimates of the butterfat percentage in the milk of four cattle breeds



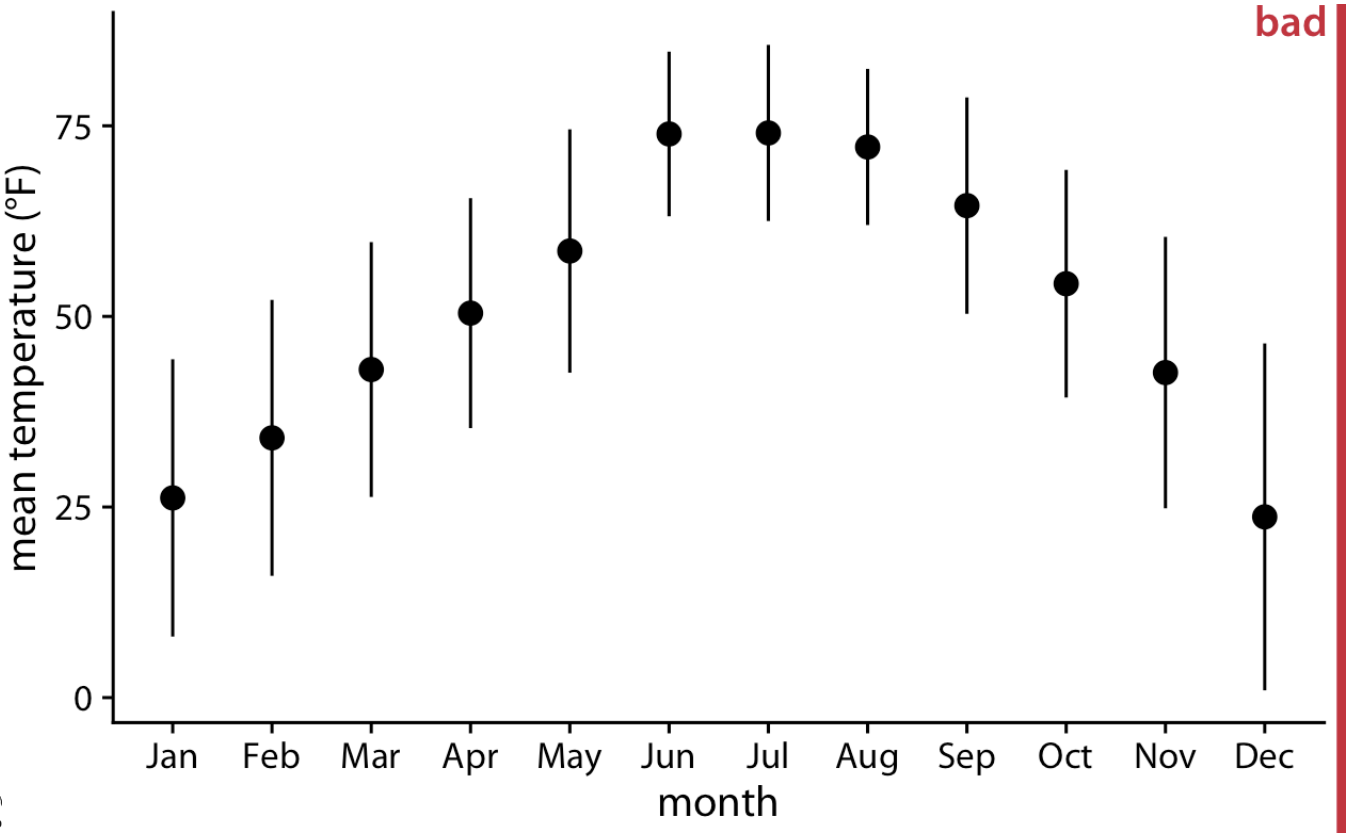
Visualizing many distributions at once

Scenarios

- Visualize multiple distributions at the same time.
  - For example, visualize how temperature varies across different months while also showing the distribution of observed temperatures within each month.

Visualizing distributions along the vertical axis

- Show their mean or median as points,
- Show some indication of the variation around the mean or median shown by error bars.

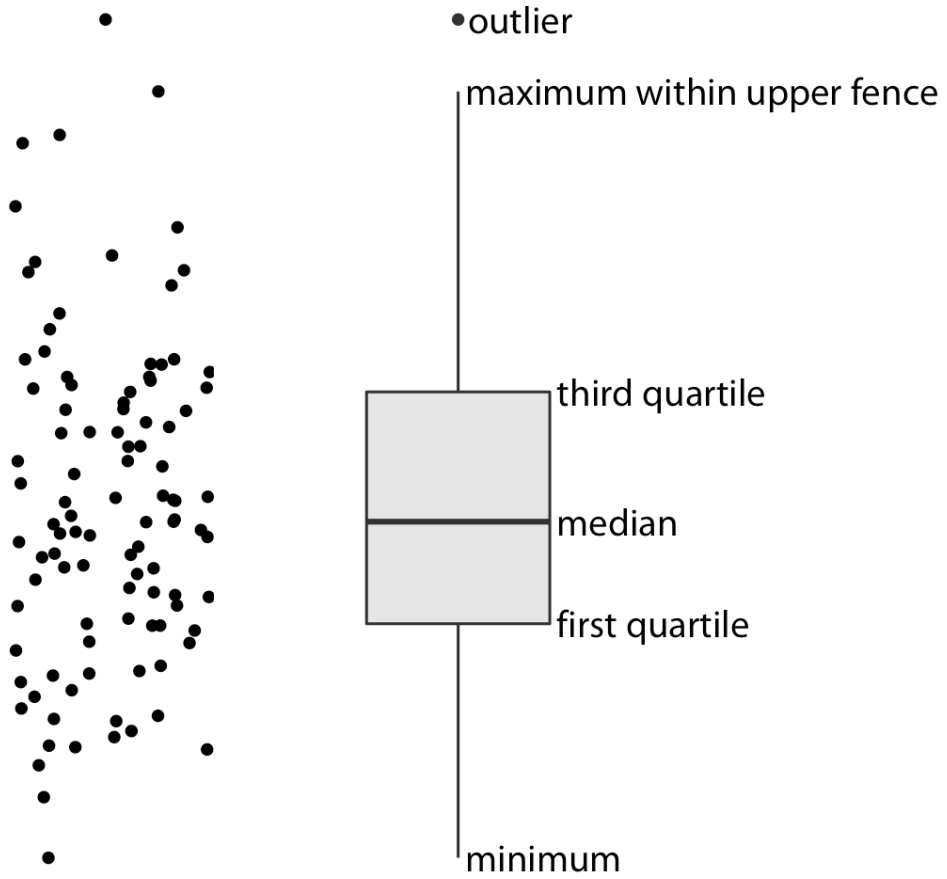


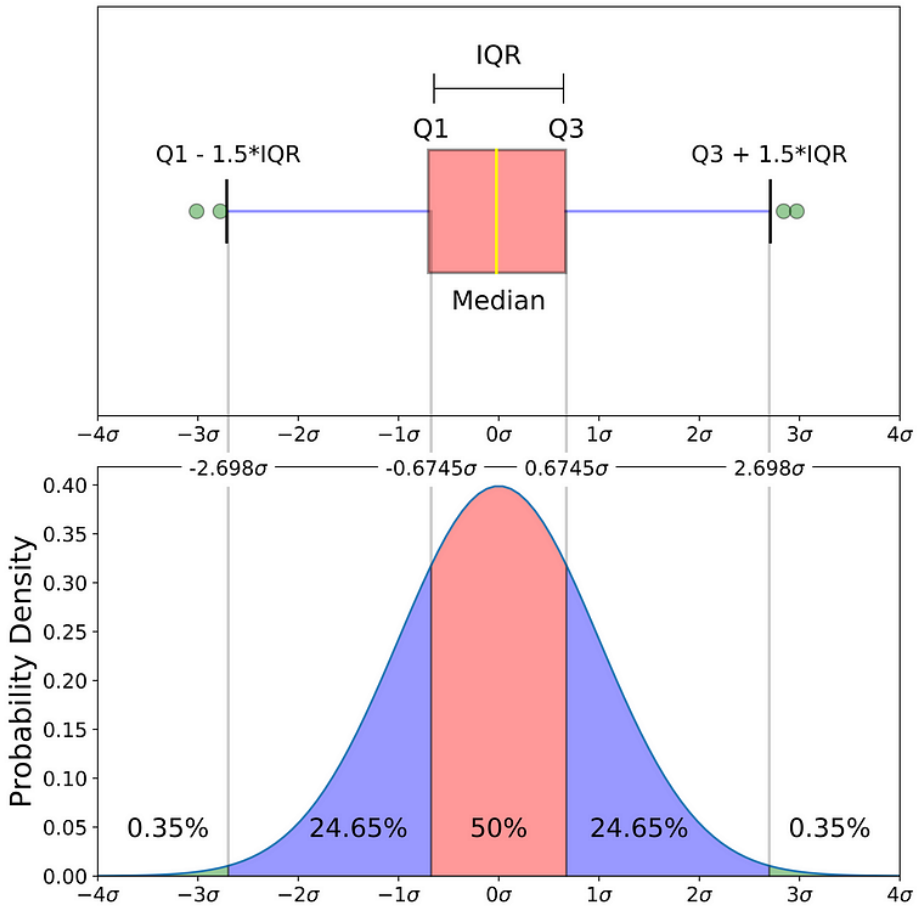
Problems of the approach

- Loss a lot of information about the data
- Is not immediately obvious what the points represent
- Is not obvious what the error bars represent

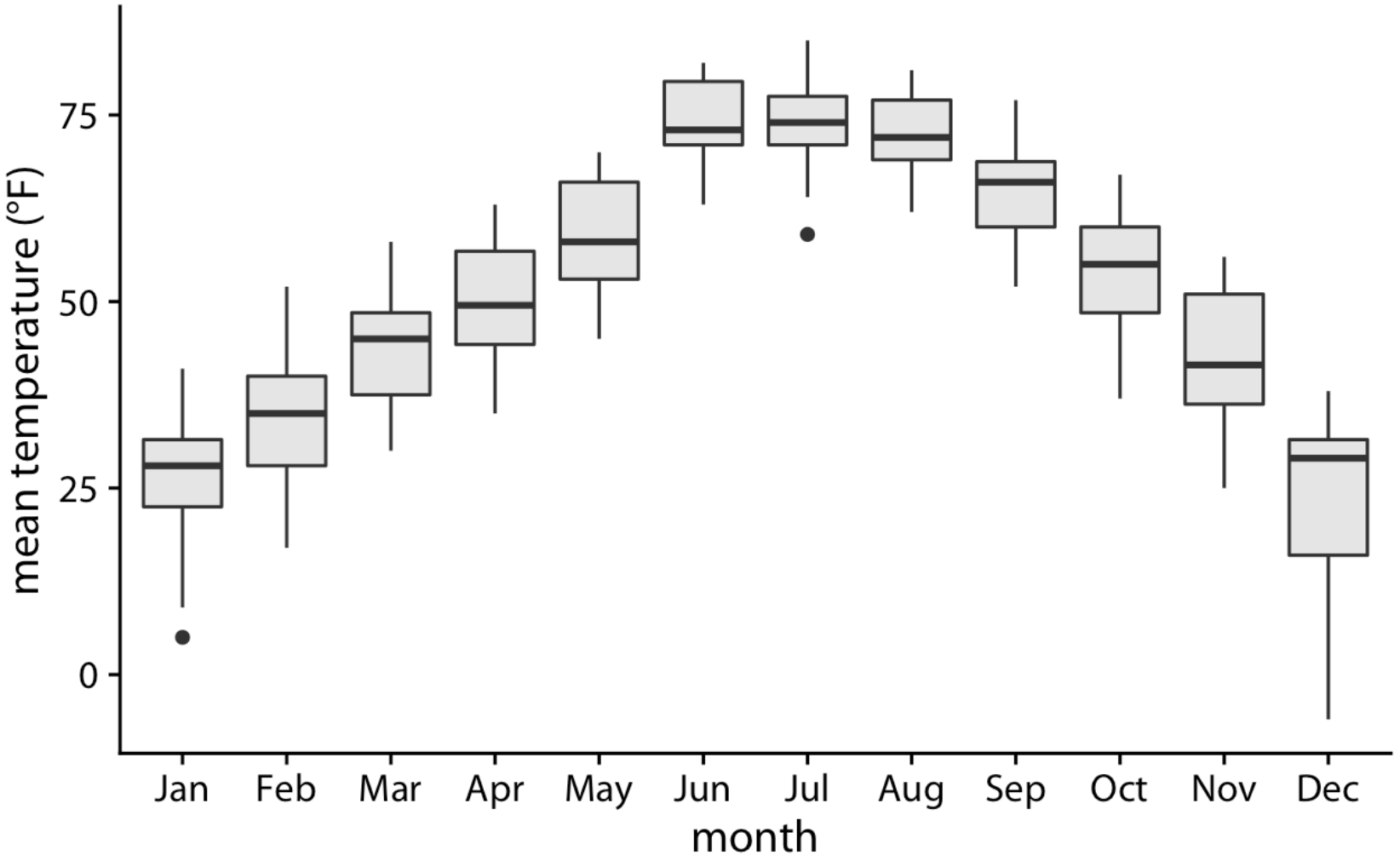
Boxplot

- Invented by the statistician John Tukey in 1970s



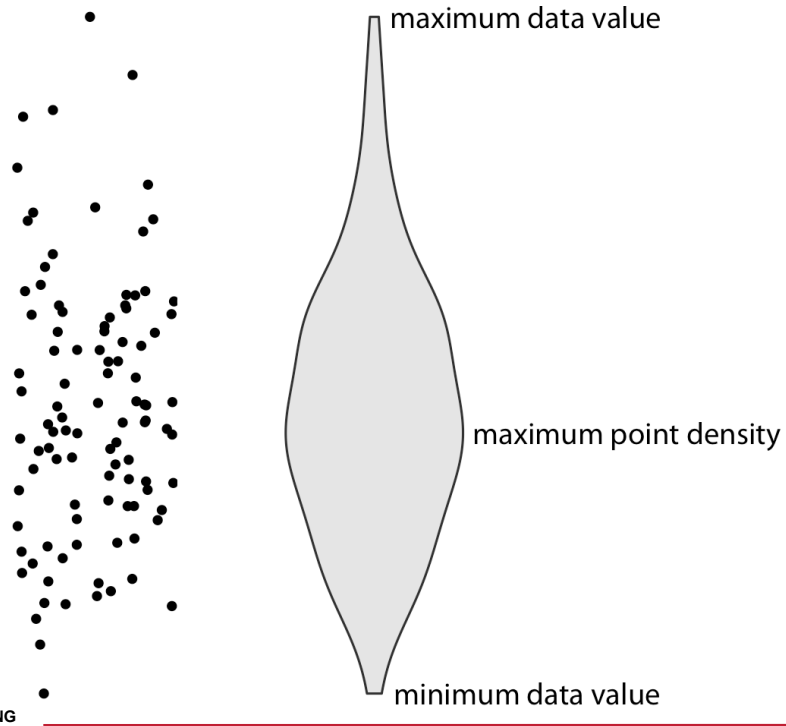


Example: Mean daily temperatures in Lincoln, NE, visualized as boxplots

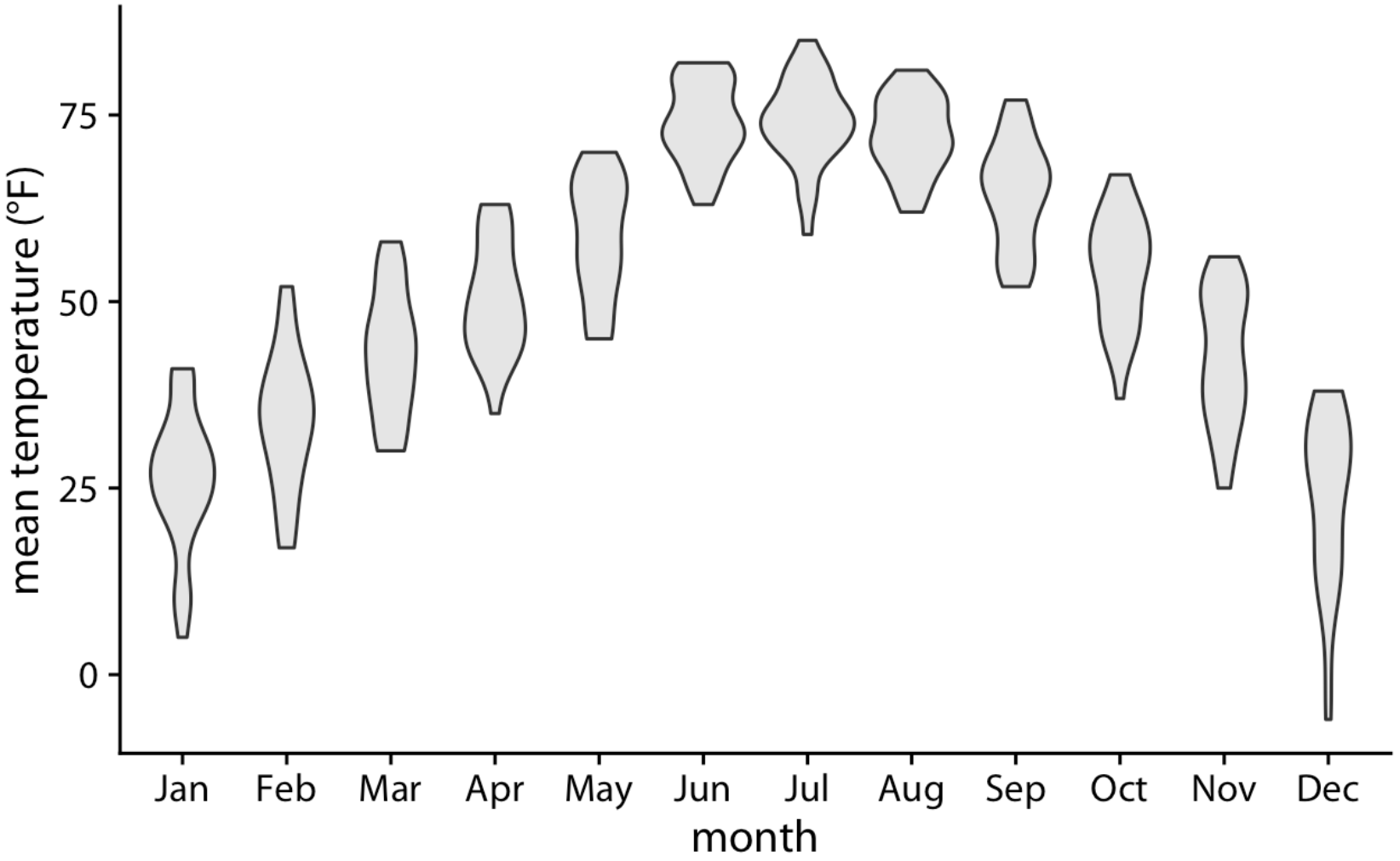


Violin plots

- Only the y values of the points are visualized in the violin plot.
- The width of the violin at a given y value represents the point density at that y value.
- Technically, a violin plot is a density estimate rotated by 90 degrees and then mirrored.

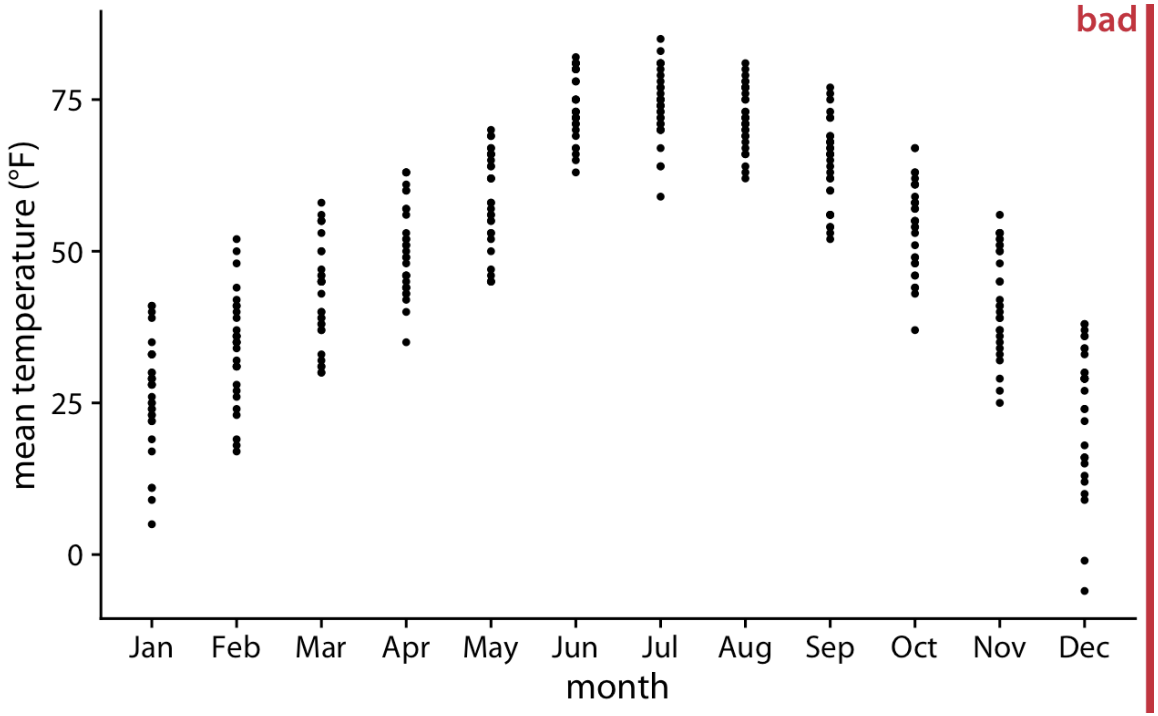


Example: Mean daily temperatures in Lincoln, NE, visualized as violin plots



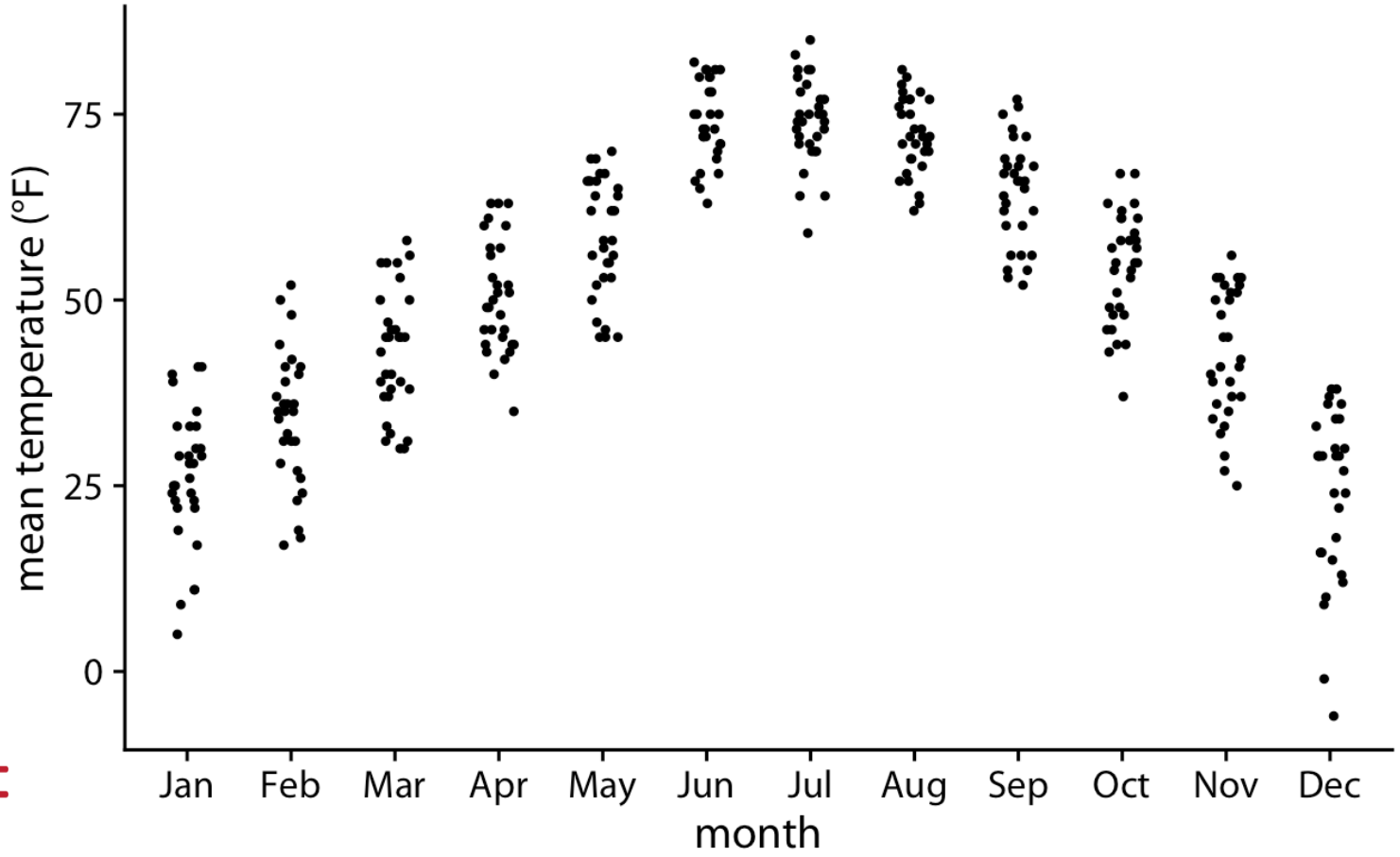
Strip charts

- Displays numerical data along a single strip.
- A good alternative to boxplots when the sample sizes are small so that you can see the individual data points.



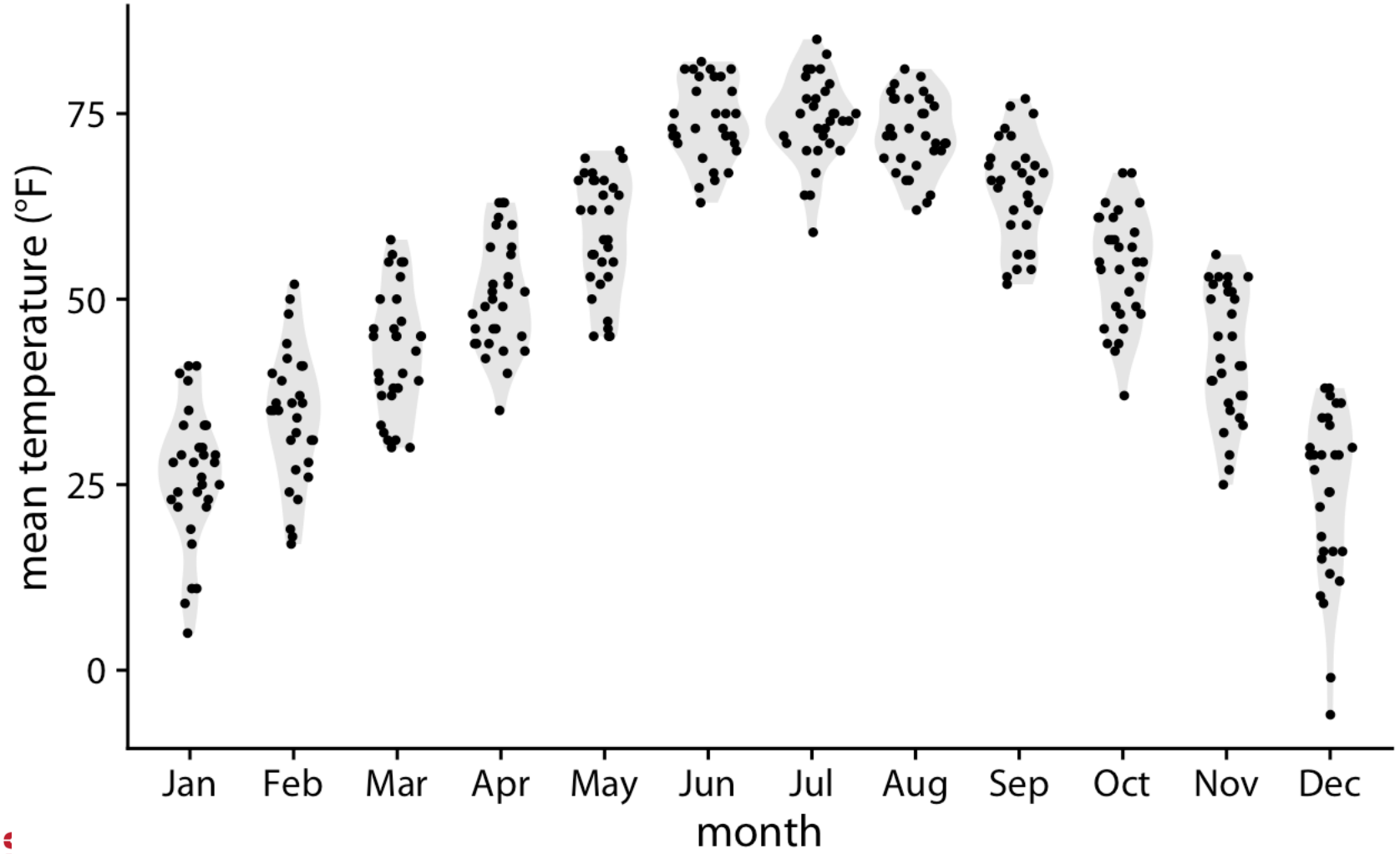
Example: Mean daily temperatures in Lincoln, NE, visualized as strip charts

- Spread out the points somewhat along the x axis, by adding some random noise in the x dimension (jittering)



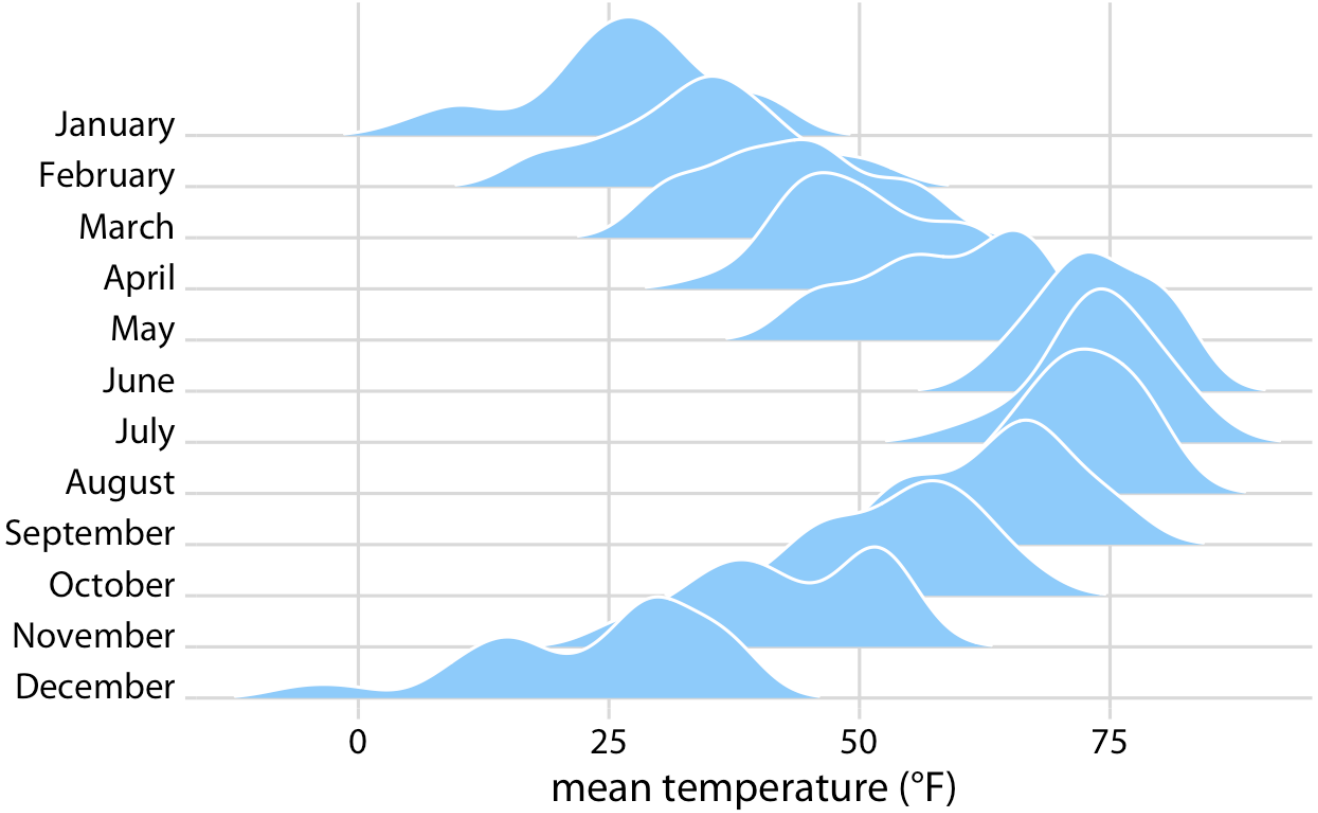
Example: Mean daily temperatures in Lincoln, NE, visualized as sina plots

- A combination of individual points and violins.



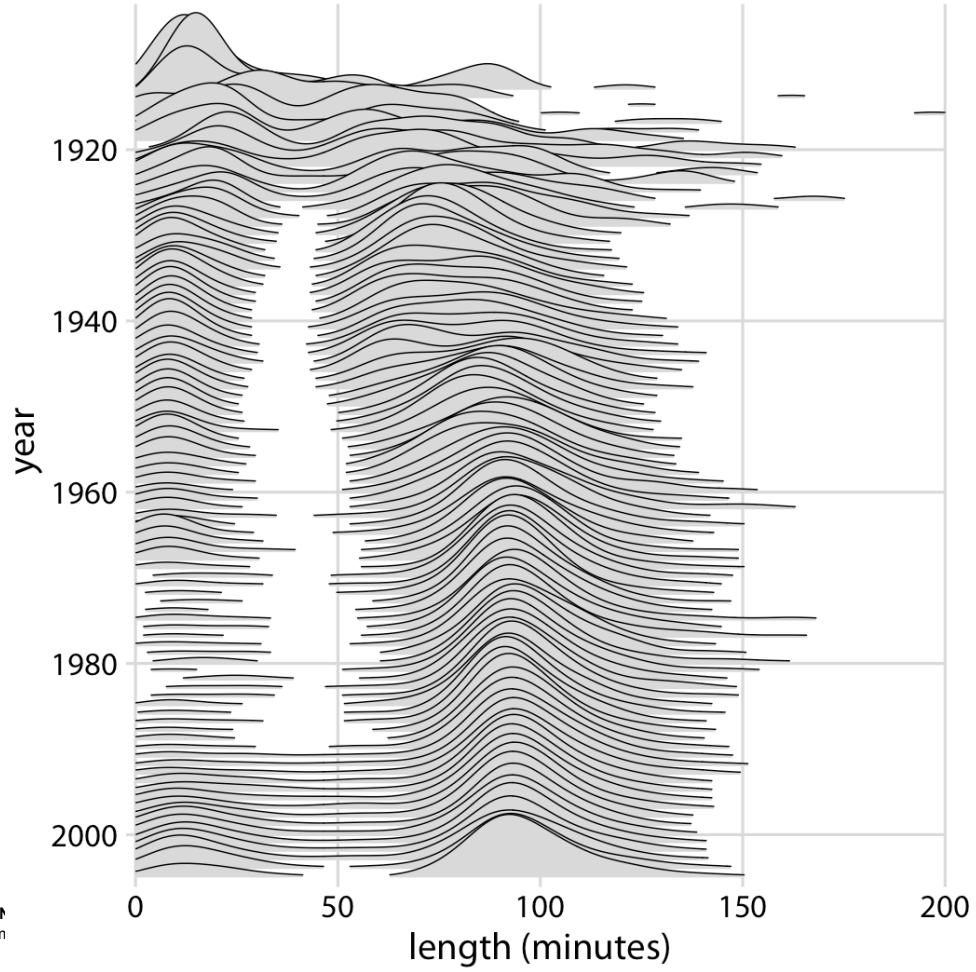
Visualizing distributions along the horizontal axis

- Ridgeline plots
  - These plots look like mountain ridgelines.
  - Standard ridgeline plot uses density estimates.
  - The purpose of the plot is not to show specific density values but instead to allow for easy comparison of density shapes and relative heights across groups.



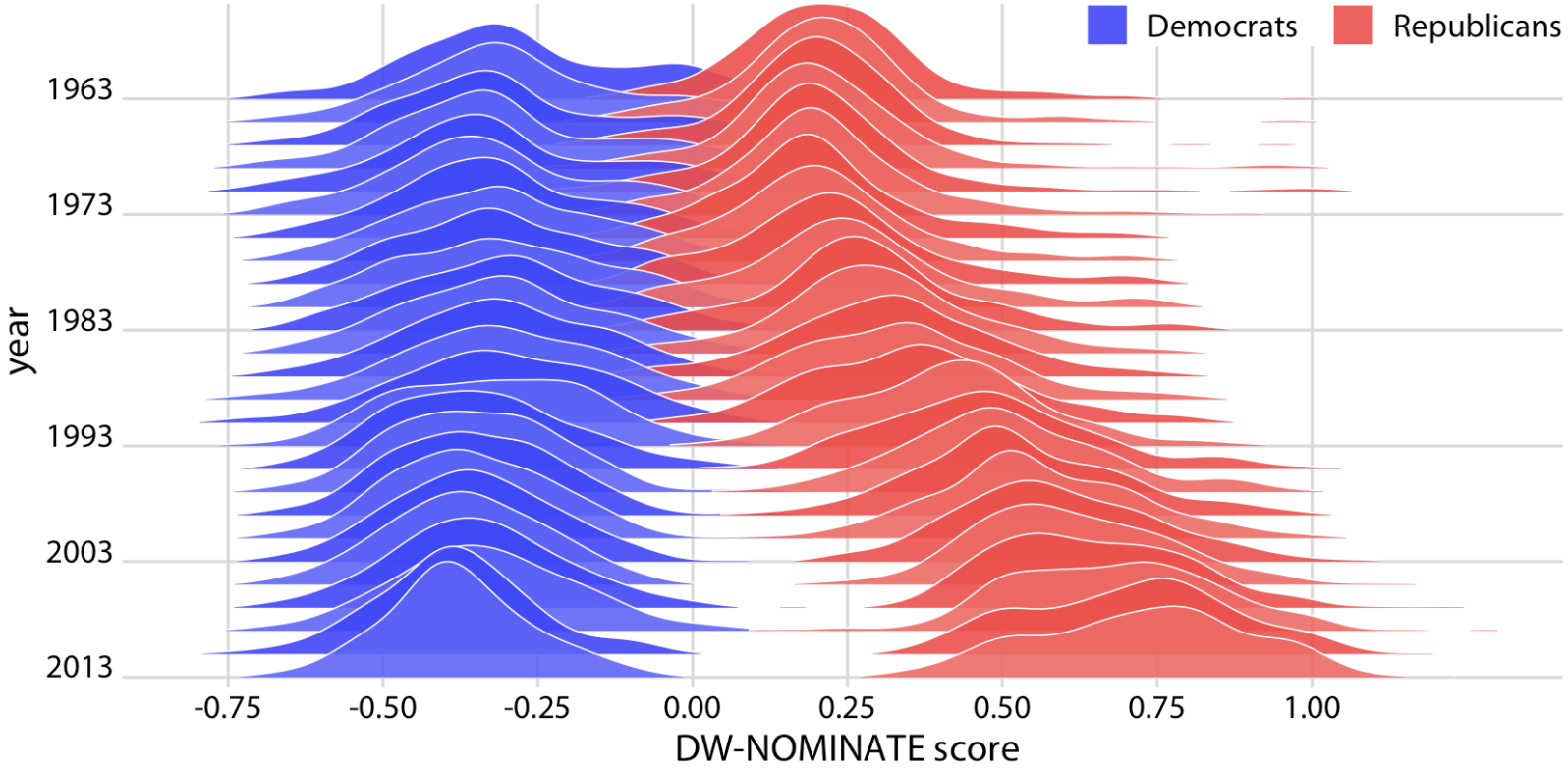
Example: Evolution of movie lengths over time

- Ridgeline plots scale to very large numbers of distributions



Example: Voting patterns in the US House of Representatives

- Ridgeline plot is used to compare two trends over time.
- Voting patterns have become increasingly



HUST

ĐẠI HỌC BÁCH KHOA HÀ NỘI

HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

ONE LOVE. ONE FUTURE.