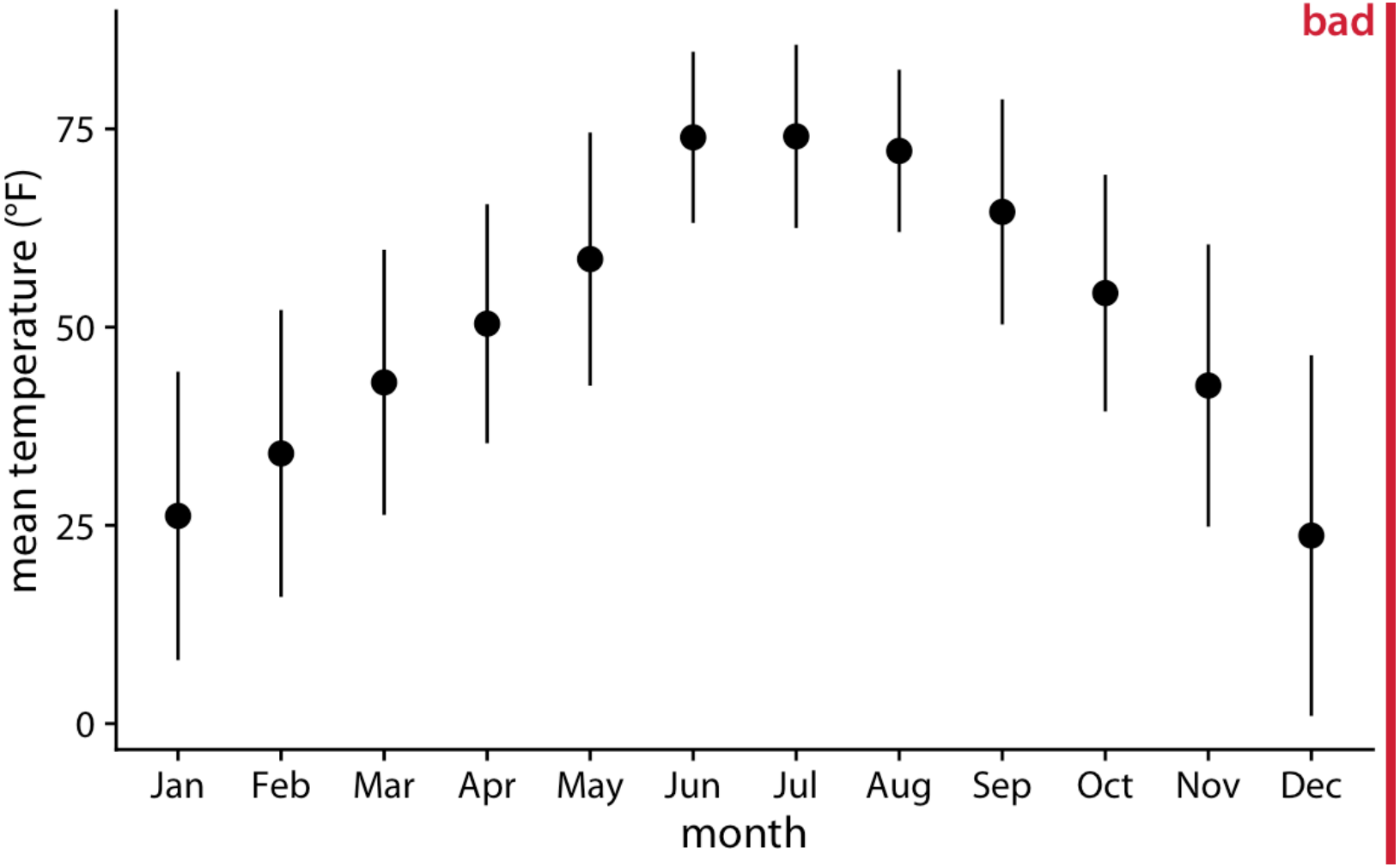
Dispersion Visualization

Visualizing many distributions at once

Scenario: Weather Data

- Limitations of previous visualization methods
- Example scenario: Visualizing temperature variations across different months (showing twelve temperature distributions at once)
- Definition of response variable and grouping variables

Response and Grouping Variables

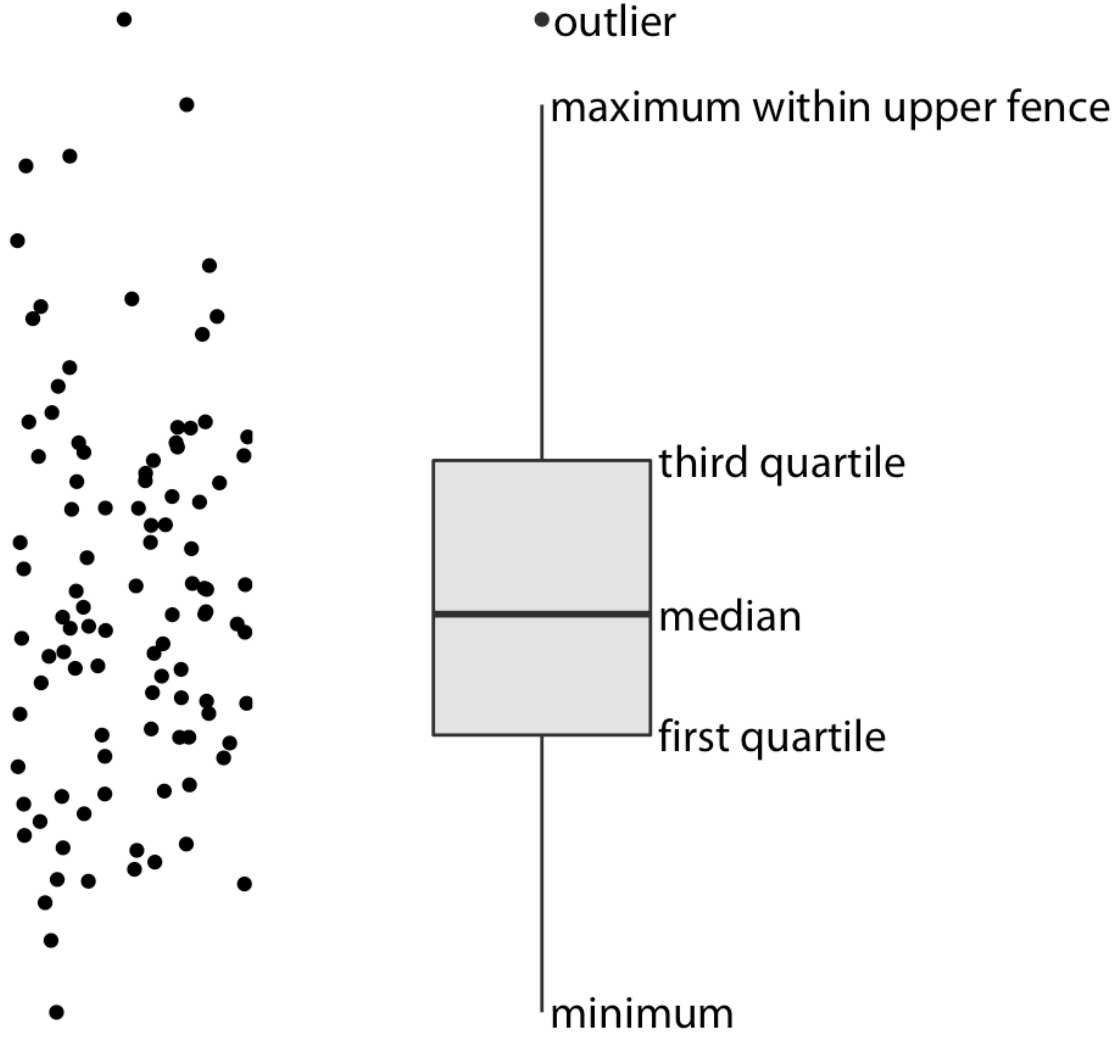


Mean daily temperatures in Lincoln, Nebraska in 2016 are represented by points indicating average daily mean temperatures for each month, with error bars depicting twice the standard deviation, criticized for misusing error bars to display population variability.

Visualizing Distributions Along Vertical Axis

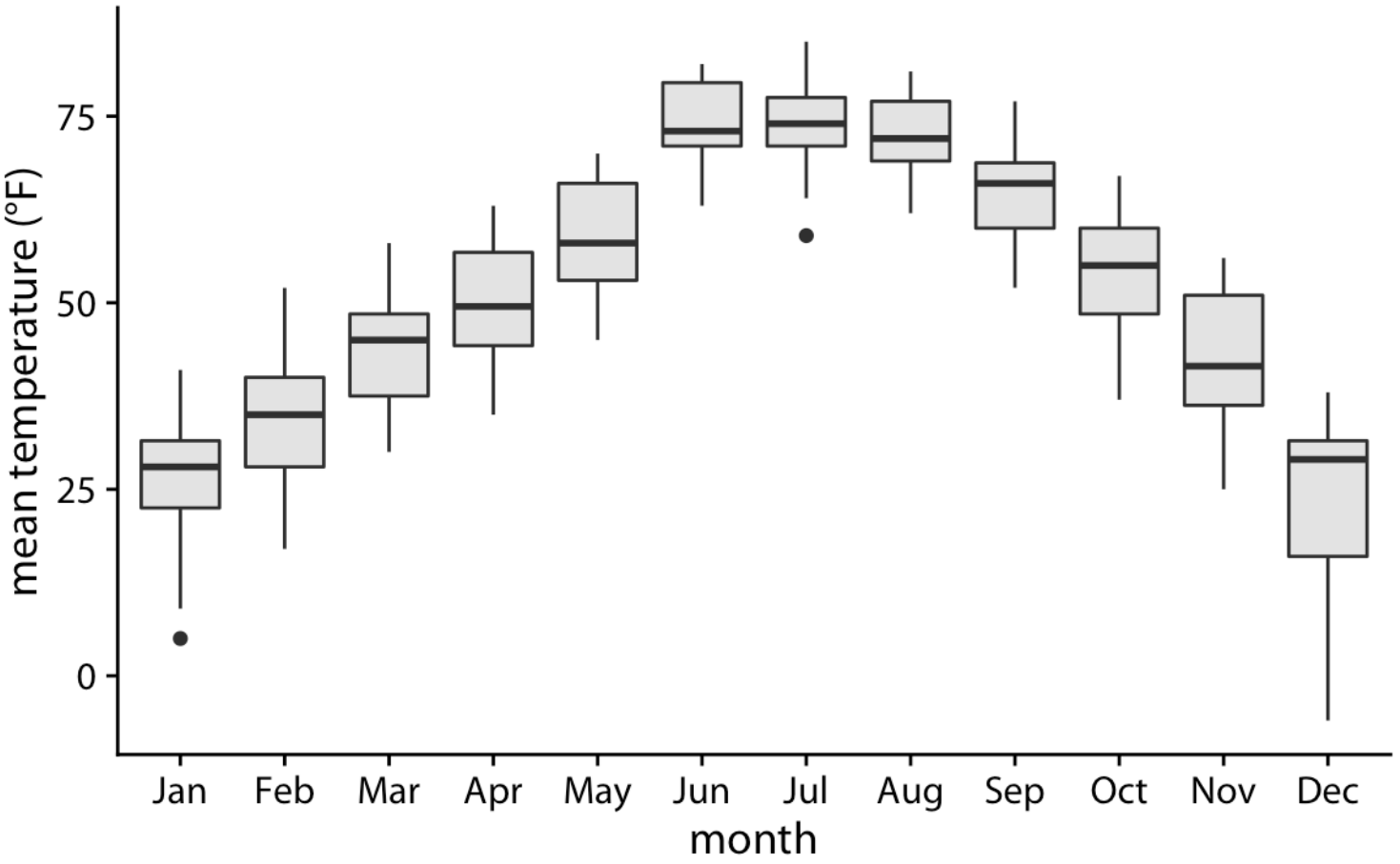
- Approaches to visualize distributions along the vertical axis
- Shortcomings of mean/median with error bars
- Introduction to boxplots as a solution

Anatomy of Box Plot



Anatomy of a boxplot: Represents median with a line, encloses middle 50% of data in a box, and extends whiskers to data range or 1.5 times the box height; outliers shown as individual dots.

Anatomy of Box Plot (continued)

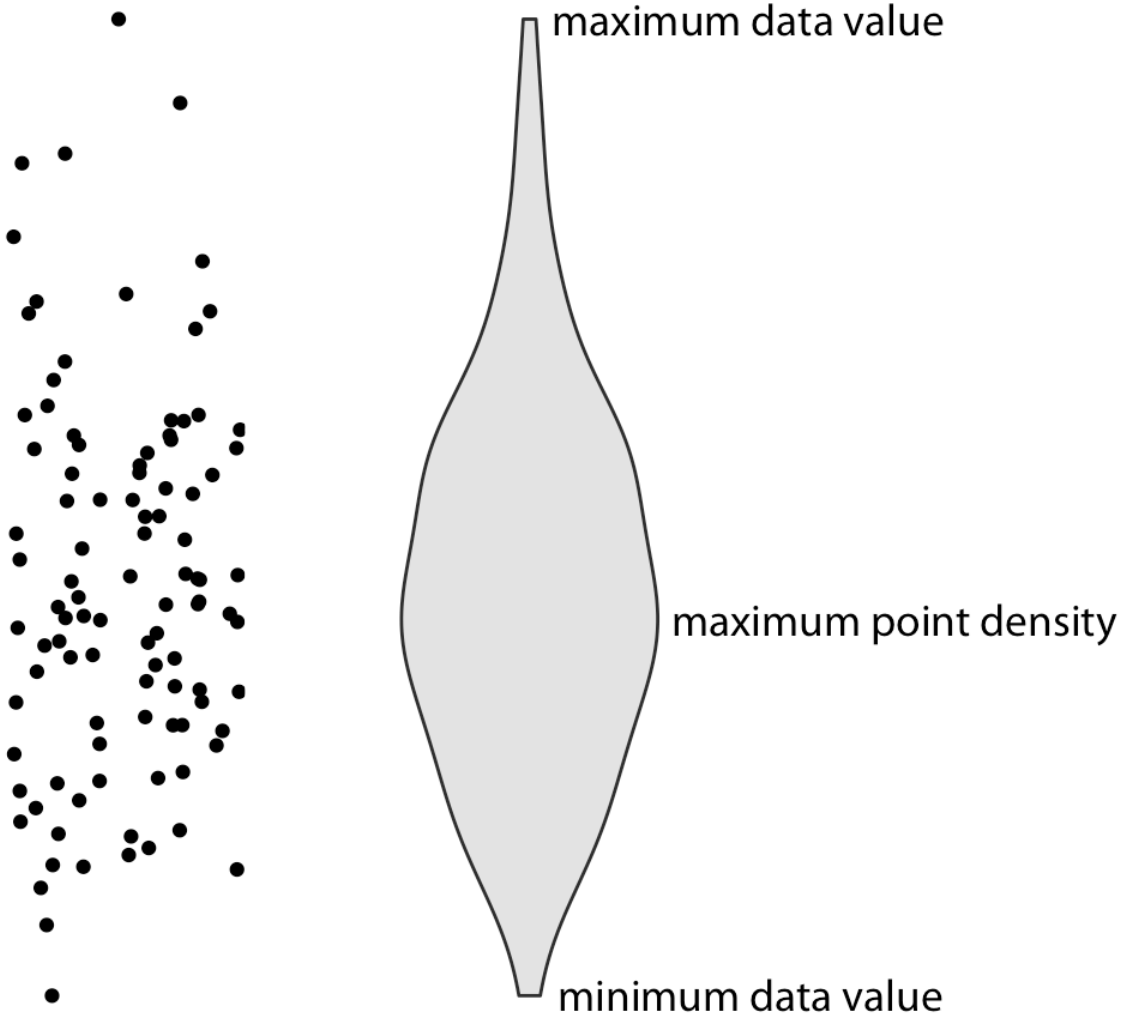


Mean daily temperatures in Lincoln, Nebraska, visualized as boxplots

Boxplots vs. Violin Plots

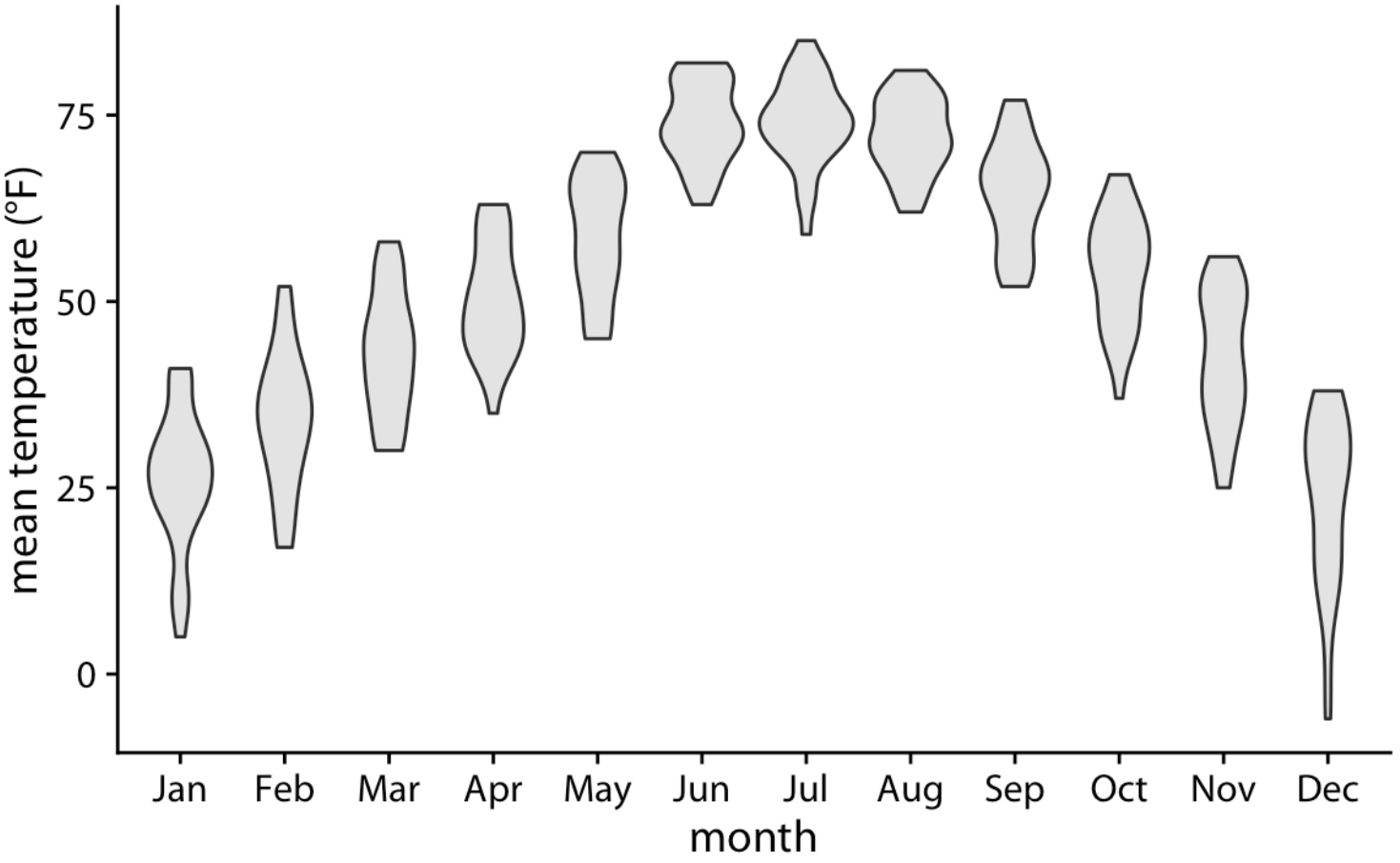
- Comparison between boxplots and violin plots
- Benefits of using violin plots for nuanced data representation
- Anatomy of a violin plot
- Advantages and considerations when using violin plots

Anatomy of a Violin Plot



Anatomy of a violin plot:
Represents density estimate with symmetric violins mirroring data values, where width indicates point density at each value.

Violin Plots for Temperature Data

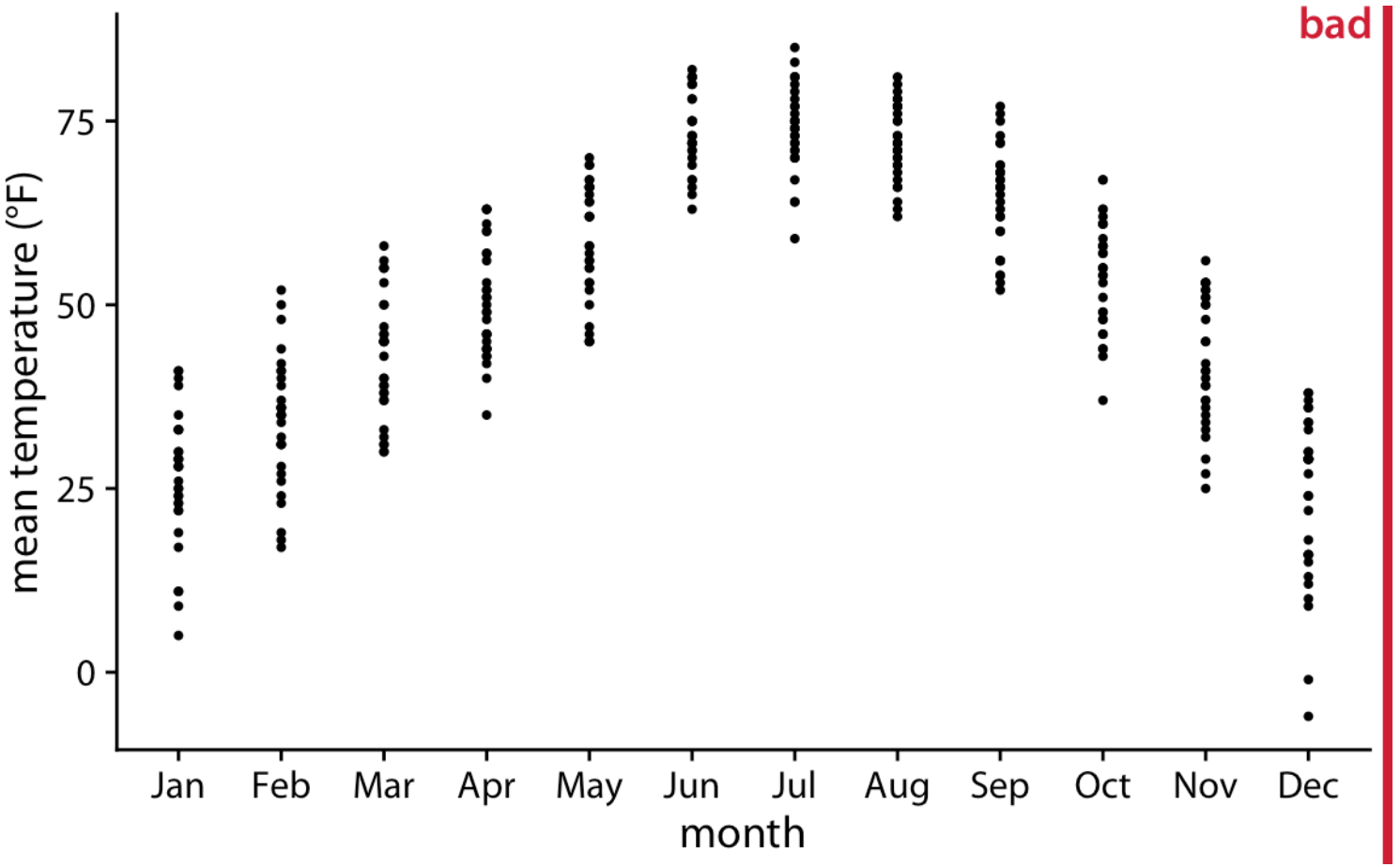


Mean daily temperatures in Lincoln, Nebraska, visualized as violin plots

Strip Charts and Sina Plots

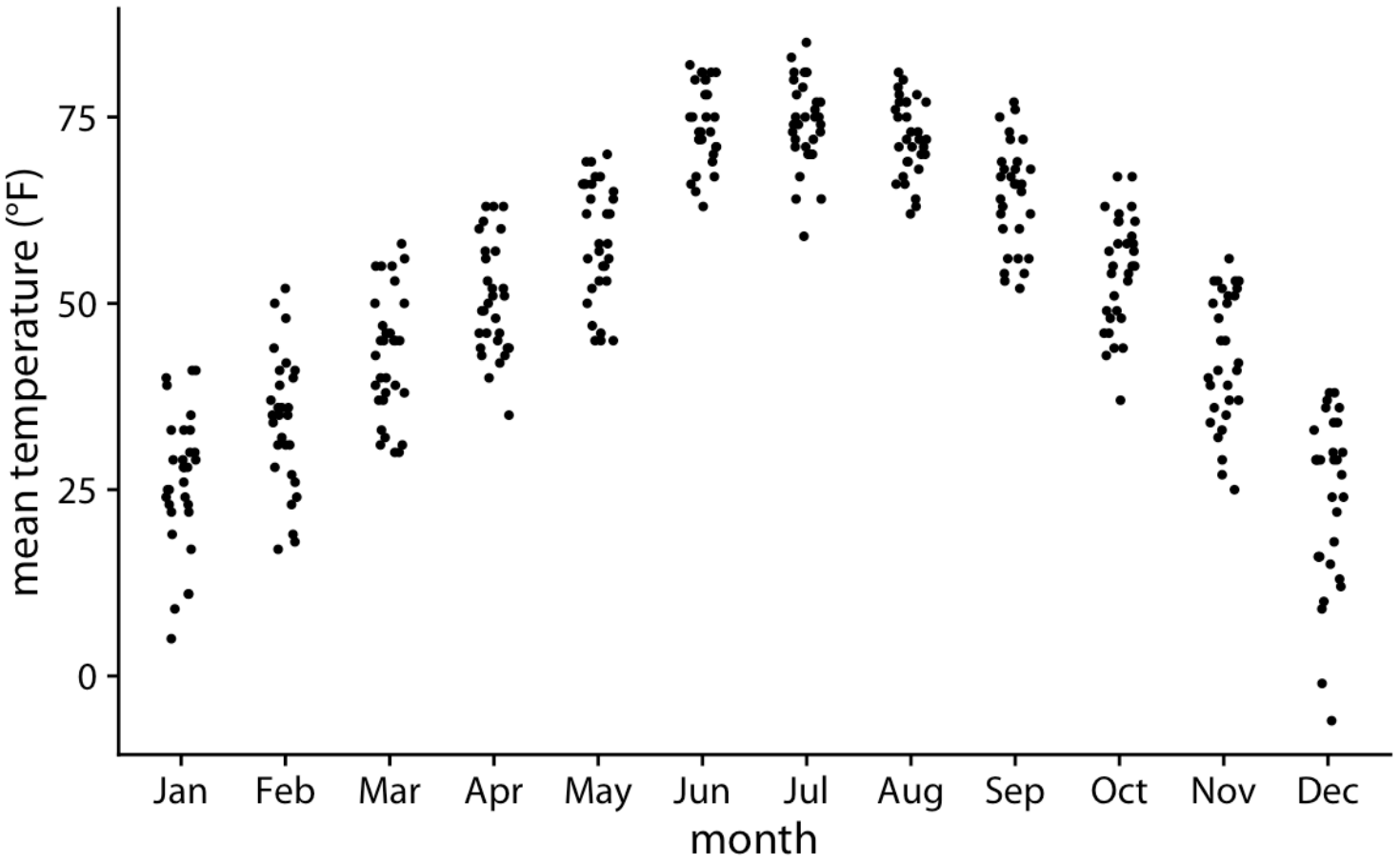
- Introduction to strip charts for visualizing individual data points.
- Limitations of strip charts with overplotting.
- Sina plots as a hybrid approach.
- Demonstration of temperature data visualized as strip charts and sina plots.

Strip Charts



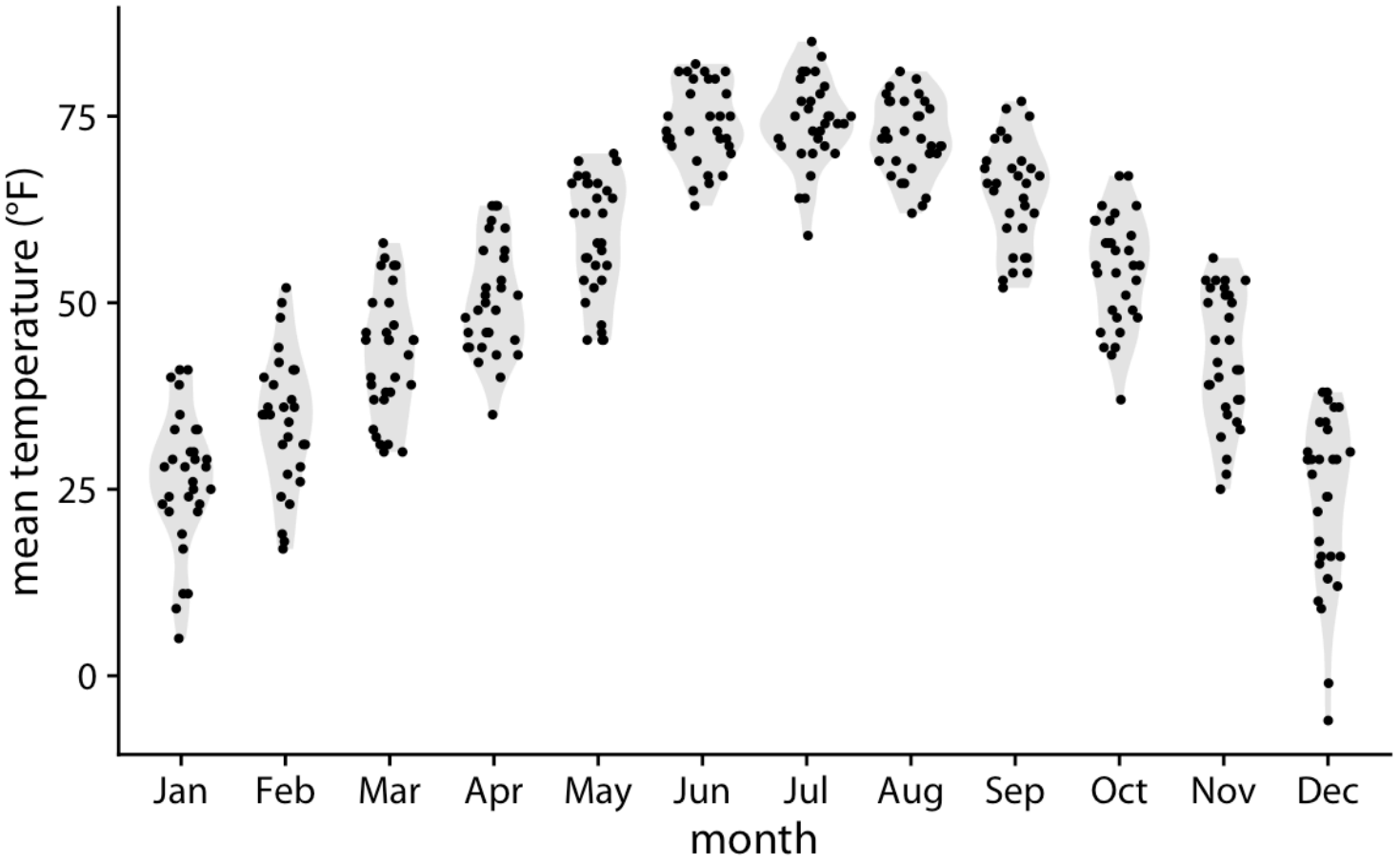
Mean daily temperatures in Lincoln, Nebraska, visualized as strip chart, labeled "bad" due to overlapping points hindering identification of common temperatures.

Strip Charts (continued)



Mean daily temperatures in Lincoln, Nebraska, visualized as strip chart. The points have been jittered along the *x* axis to better show the density of points at each temperature value.

Sina Plot

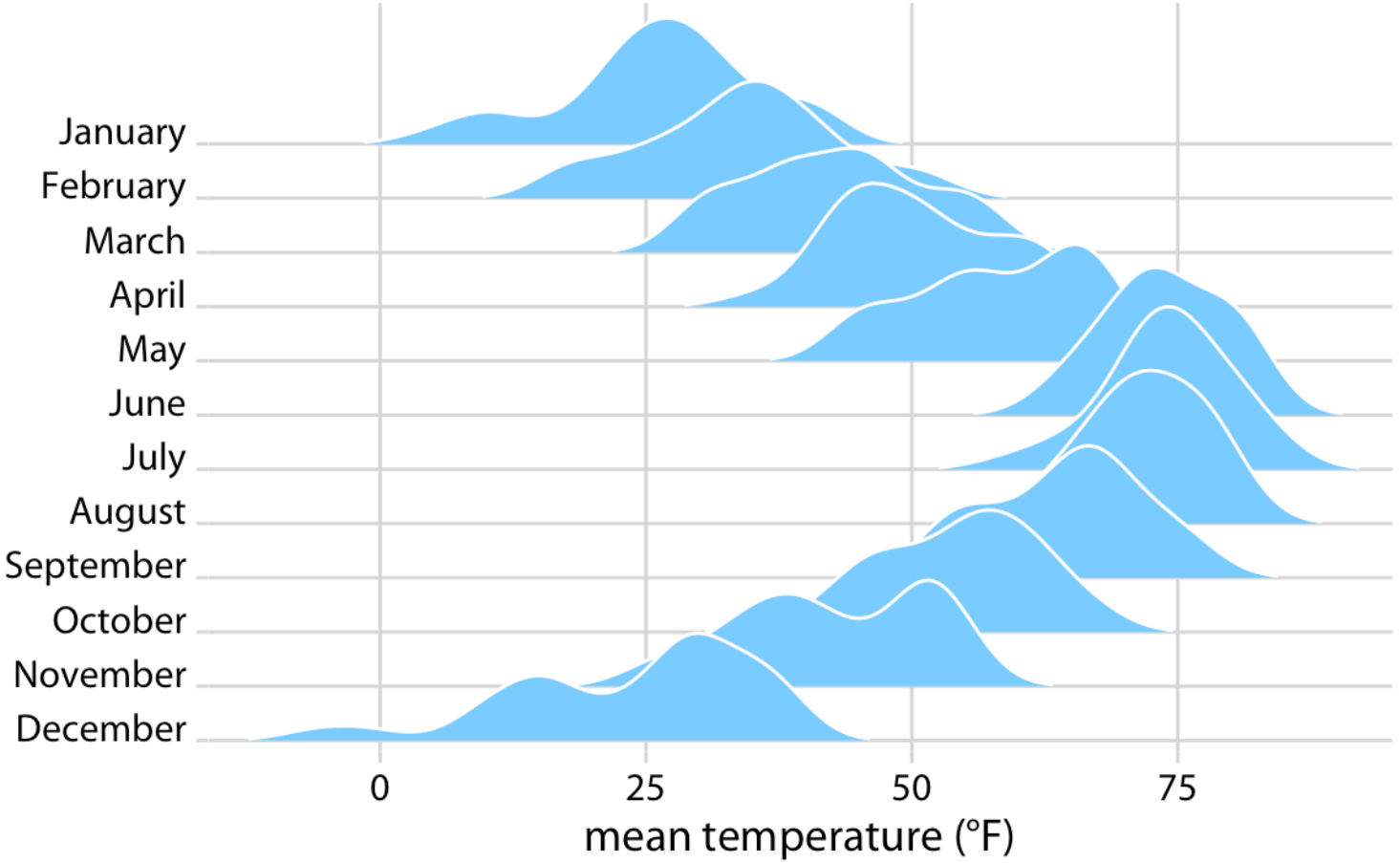


Mean daily temperatures in Lincoln, Nebraska, visualized as a sina plot (combination of individual points and violins). The points have been jittered along the *x* axis in proportion to the point density at the respective temperature.

Visualizing Distributions Along Horizontal Axis

- Introduction to visualizing distributions along the horizontal axis
- Explanation of ridgeline plots as a method for staggered distribution visualization
- Anatomy of a ridgeline plot

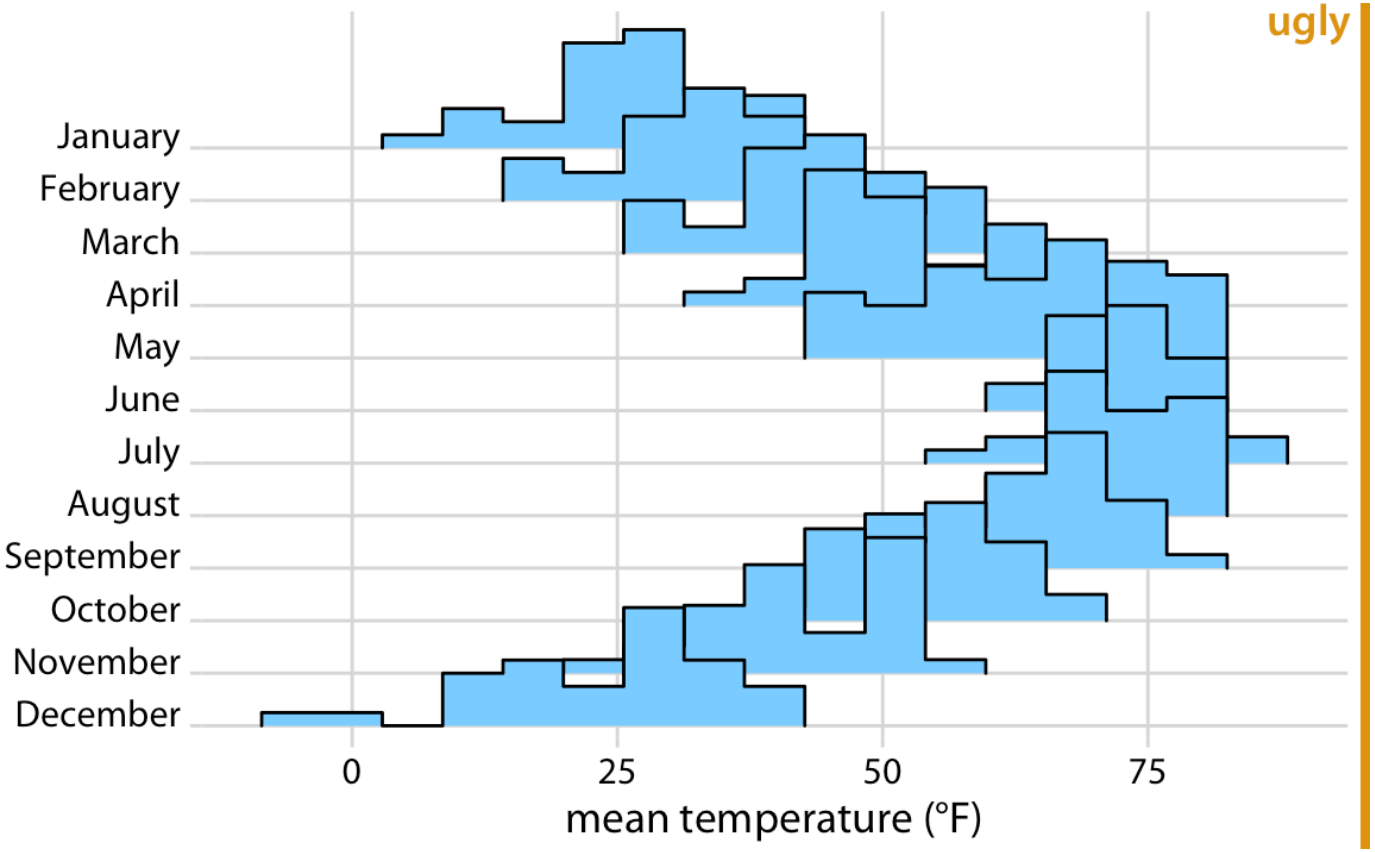
Ridgeline Plot



Temperatures in Lincoln, Nebraska, in 2016, visualized as a ridgeline plot. For each month, we show the distribution of daily mean temperatures measured in Fahrenheit

Ridgeline Plots for Temperature Data

- Comparing ridgeline plots and histogram-based ridgeline plots

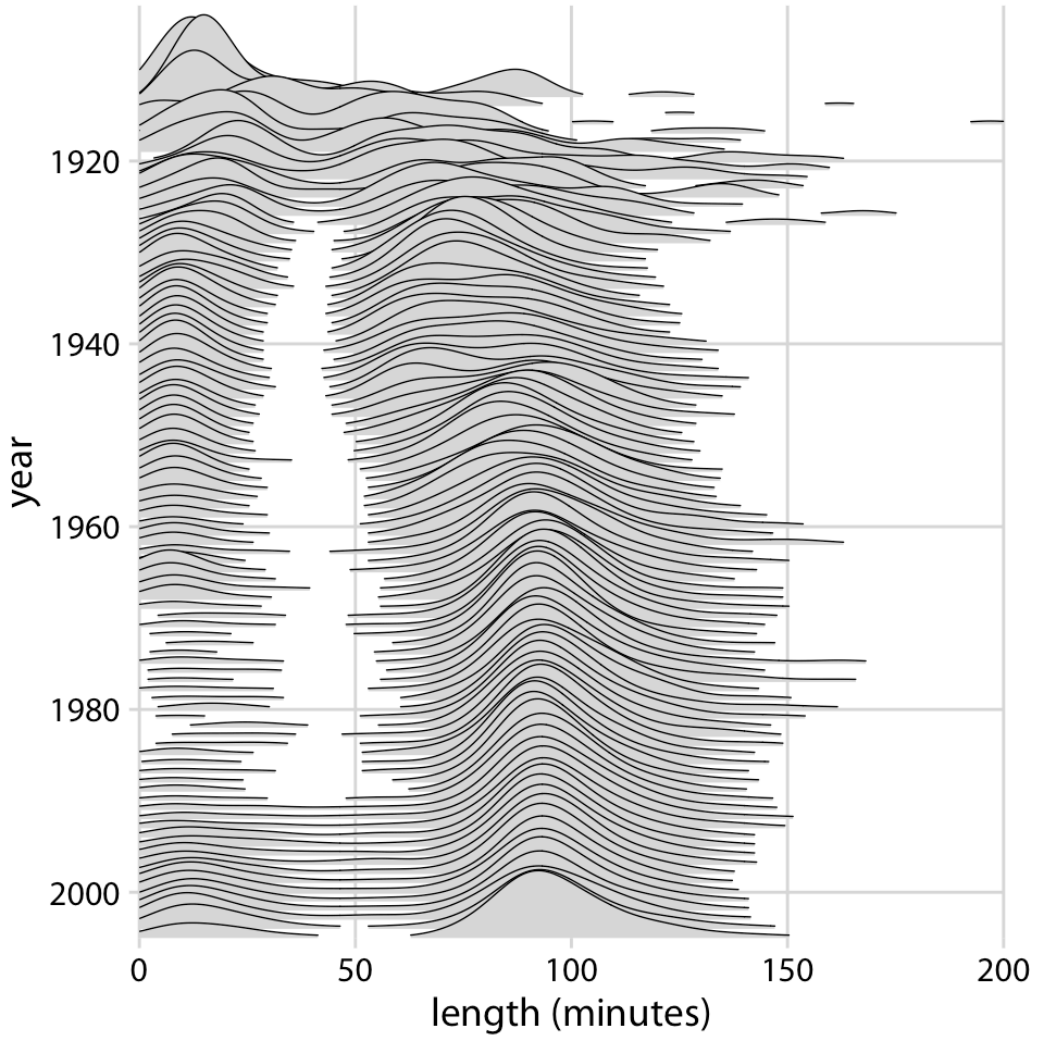


Temperatures in Lincoln, Nebraska, in 2016, visualized as a ridgeline plot of histograms. The individual histograms don't separate well visually, and the overall figure is quite busy and confusing.

Scalability of Ridgeline Plots

- Scalability of ridgeline plots for large datasets
- Example of movie length distributions over time visualized using ridgeline plots
- Benefits of ridgeline plots for trend analysis

Ridgeline plot scalability

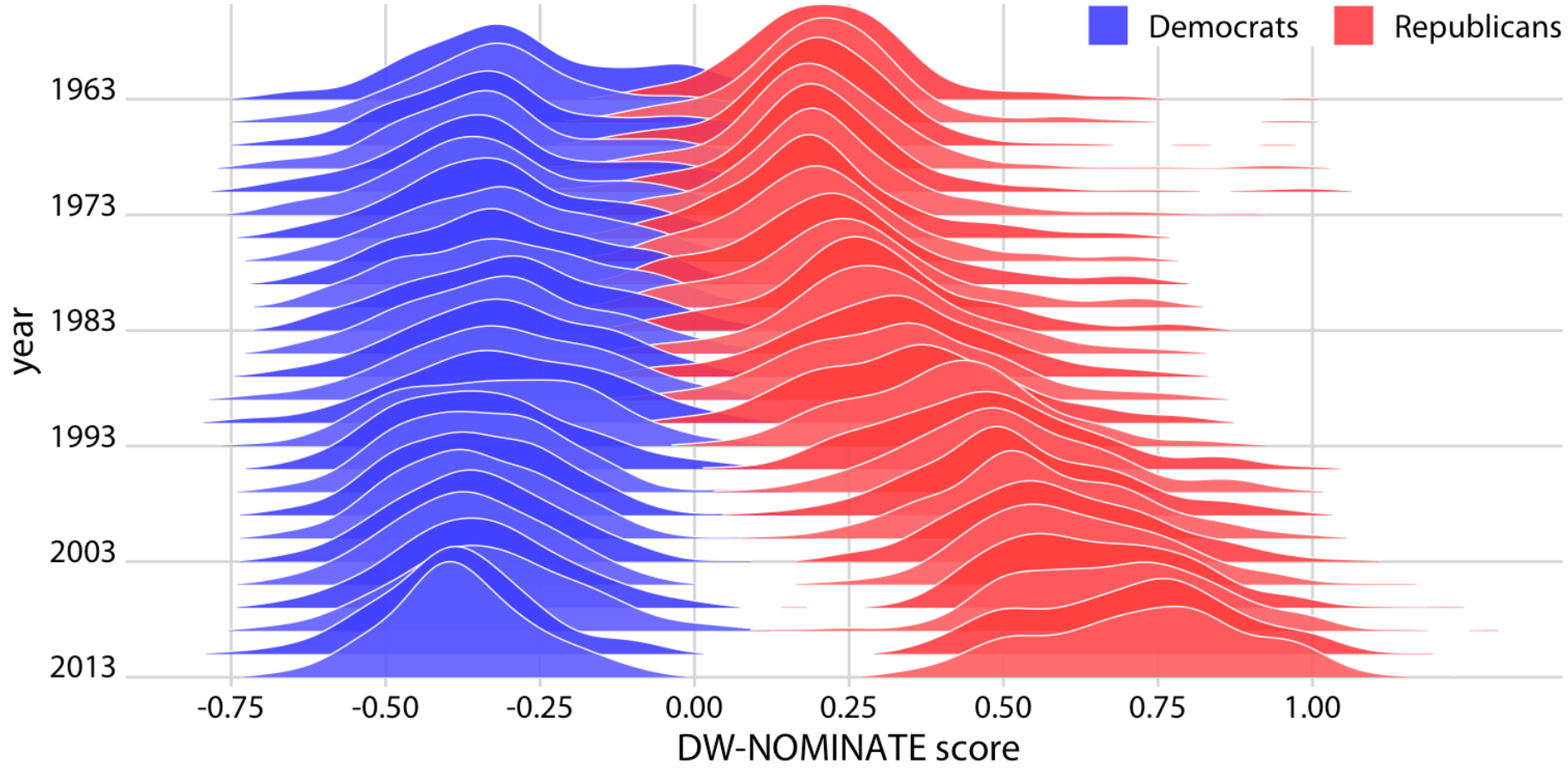


Evolution of movie lengths over time. Since the 1960s, the majority of all movies are approximately 90 minutes long.

Comparative Ridgeline Plots

- Application of ridgeline plots for comparing trends over time
- Example of voting patterns in the U.S. House of Representatives visualized using ridgeline plots

Ridgeline Plots for Voting Patterns Over Time



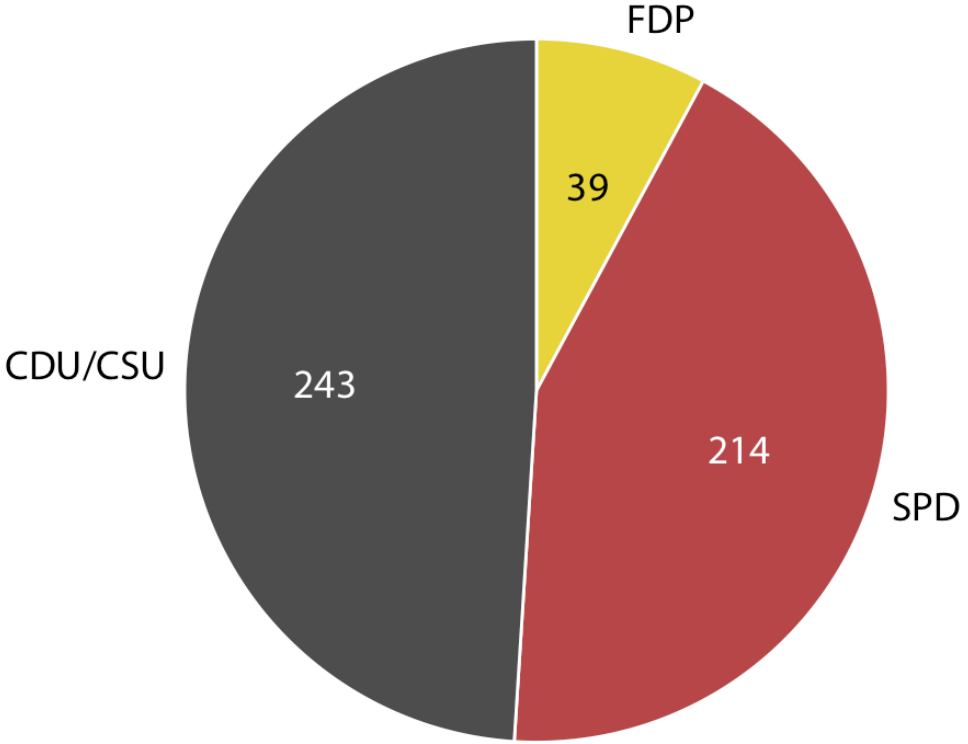
Voting patterns in the U.S. House of Representatives, showing DW-NOMINATE score distributions for each Congress from 1963 to 2013 for Democrats and Republicans, highlighting increasing polarization over time.

Dispersion Visualization

Part II

Pie Charts

- Concept of visualizing proportions
- Examples: Men and women in a group, political party percentages
- When to use: Simple fractions, small datasets.

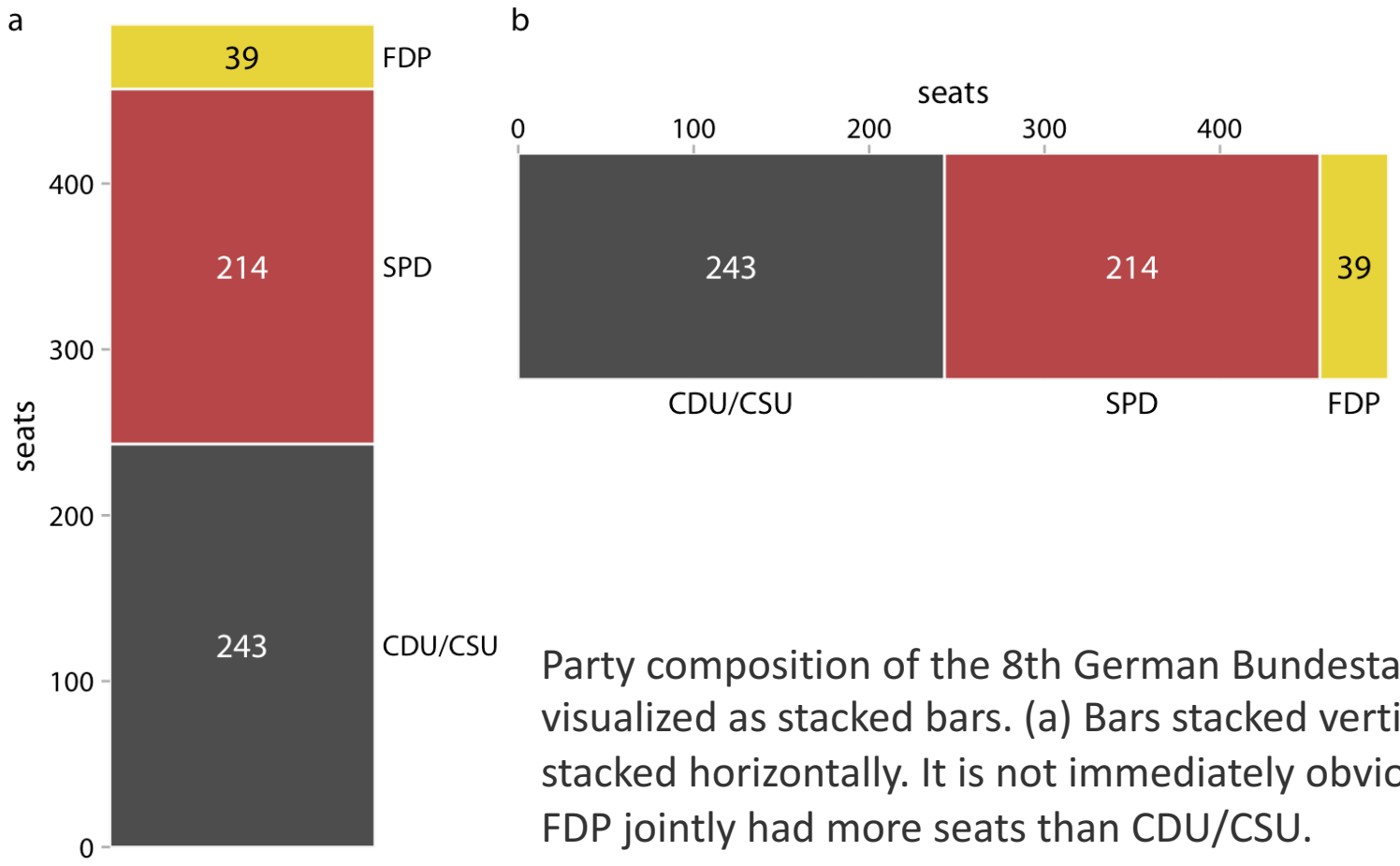


Visualization of the party composition of the 8th German Bundestag from 1976 to 1980 as a pie chart, illustrating the ruling coalition's small majority of SPD and FDP over the opposition CDU/CSU.

Stacked Bars

- Example: Market share of five hypothetical companies, 2015–2017.
- Advantages and limitations of stacked bars.
- When to use: Comparison over time, multiple conditions.

Stacked Bars (continued)

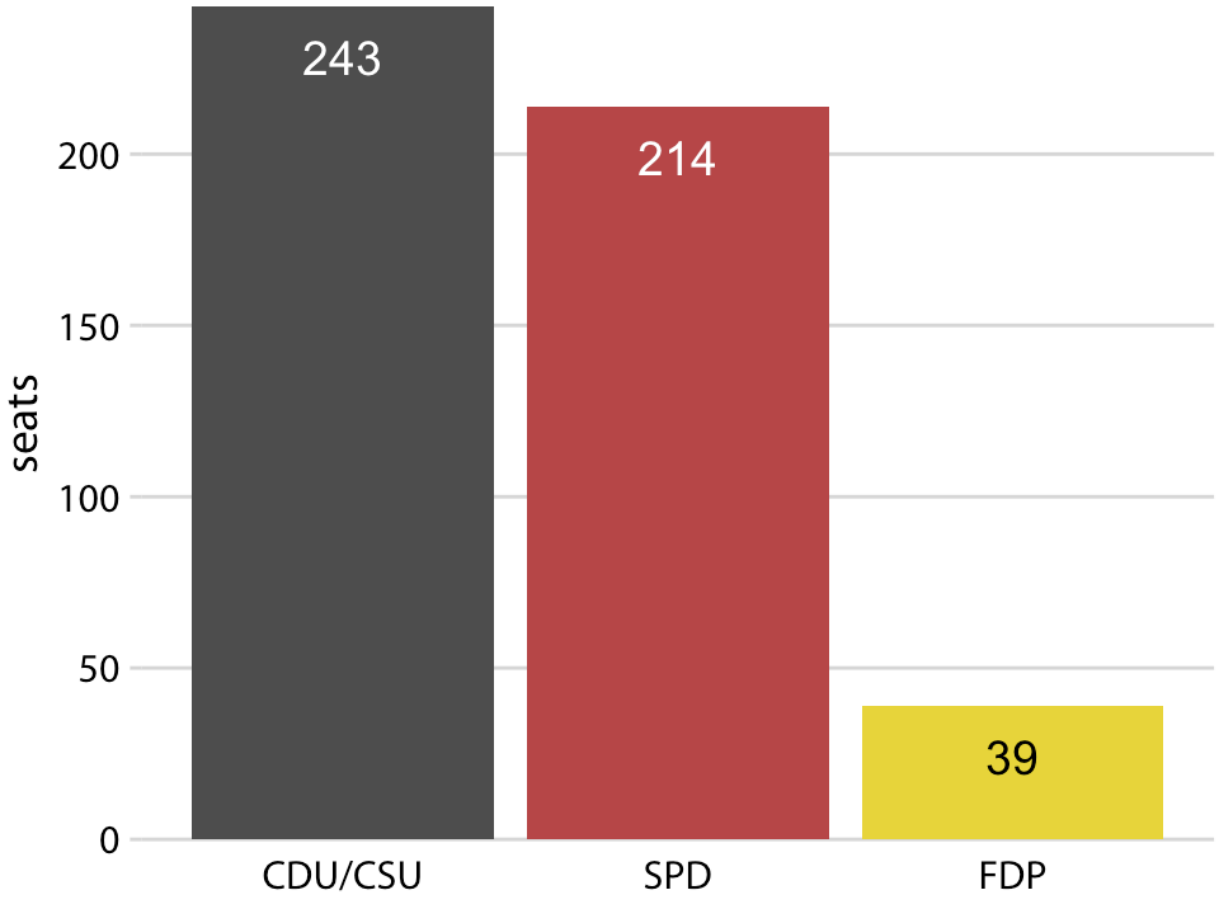


Party composition of the 8th German Bundestag, 1976–1980, visualized as stacked bars. (a) Bars stacked vertically. (b) Bars stacked horizontally. It is not immediately obvious that SPD and FDP jointly had more seats than CDU/CSU.

Side-by-Side Bars

- Example: Market share of five hypothetical companies, 2015–2017.
- Advantages and limitations of side-by-side bars.
- When to use: Direct comparison of individual fractions.

Side-by-Side Bars (continued)



Party composition of the 8th German Bundestag, 1976–1980, visualized as side-by-side bars. It is not immediately obvious that SPD and FDP jointly had more seats than CDU/CSU.

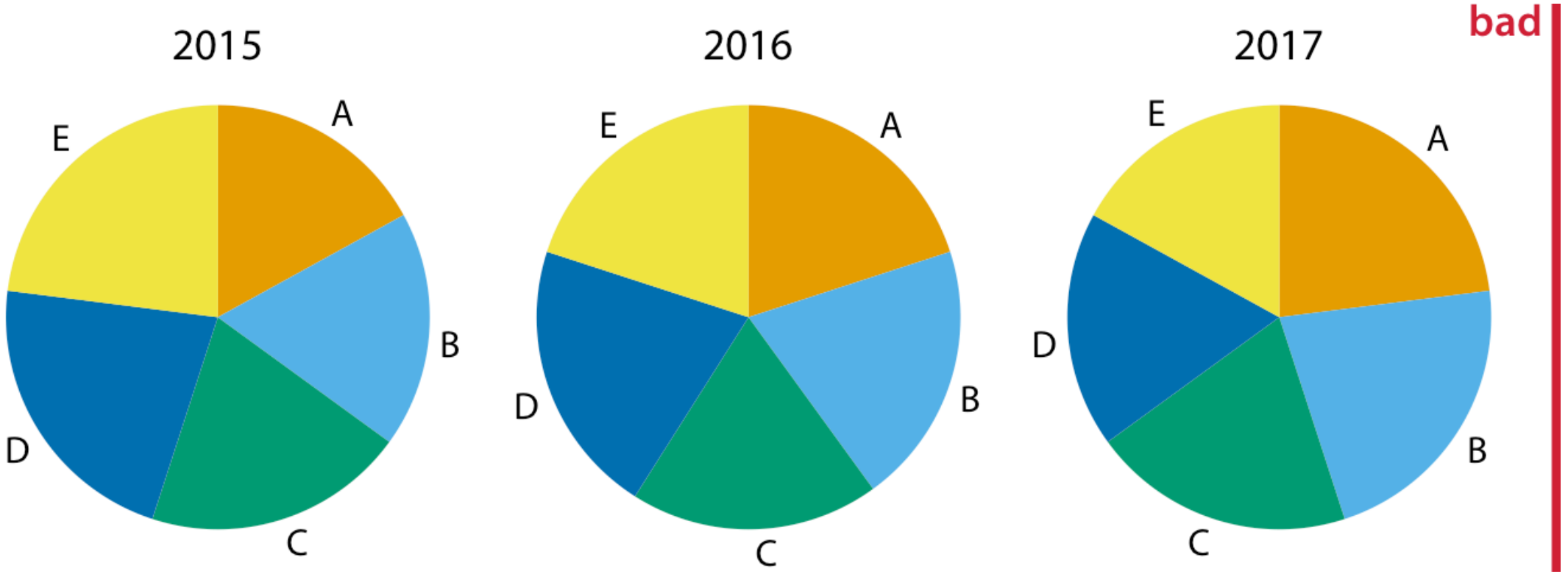
Pros and cons to visualizing proportions

	Pie chart	Stacked bars	Side-by-side bars
Clearly visualizes the data as proportions of a whole	✔	✔	✖
Allows easy visual comparison of the relative proportions	✖	✖	✔
Visually emphasizes simple fractions, such as 1/2, 1/3, 1/4	✔	✖	✖
Looks visually appealing even for very small datasets	✔	✖	✔
Works well when the whole is broken into many pieces	✖	✖	✔
Works well for the visualization of many sets of proportions or time series of proportions	✖	✔	✖

Stacked Densities

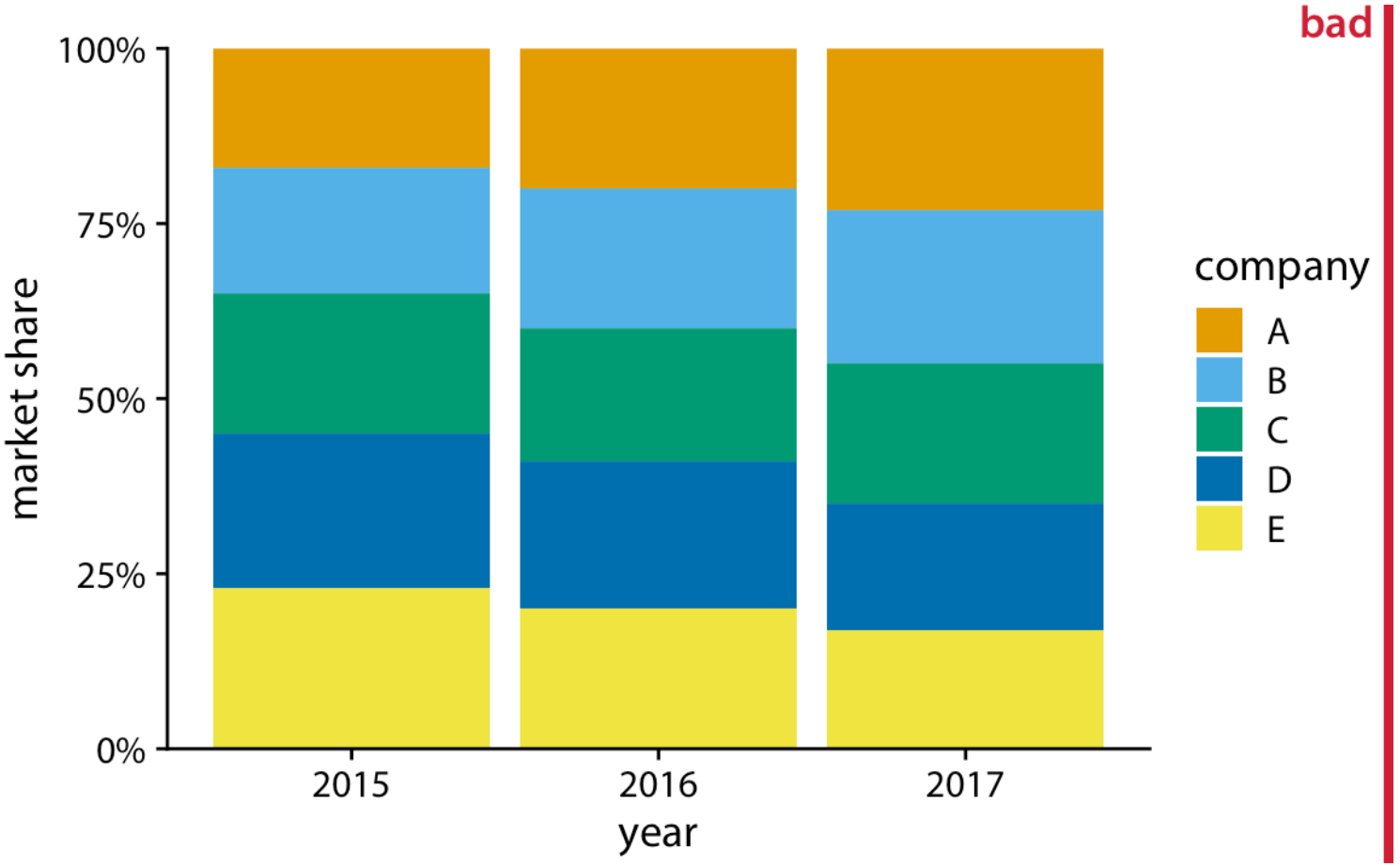
- Example: Health status by age.
- Advantages and limitations of stacked densities.
- When to use: Continuous variables, trends over time.

Case for side-by-side bars



Market share of five hypothetical companies for 2015–2017, visualized as pie charts, faces challenges in comparing shares within years and tracking changes across years.

Case for side-by-side bars (continued)

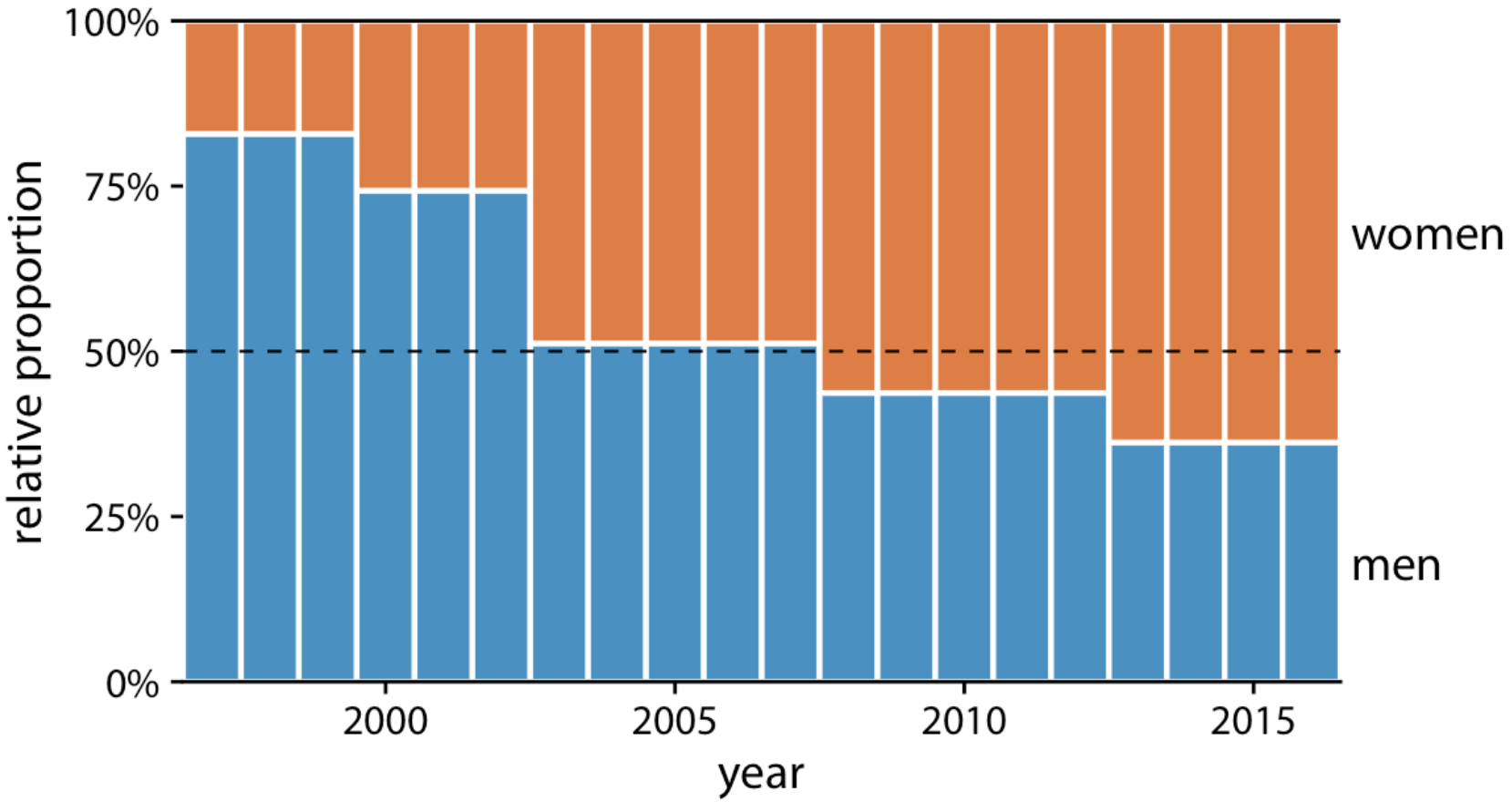


Market share of five companies for 2015–2017, shown as stacked bars, with difficulties in comparing relative shares within years and tracking changes across years for certain companies due to bar location shifts.

Case for side-by-side bars (continued)

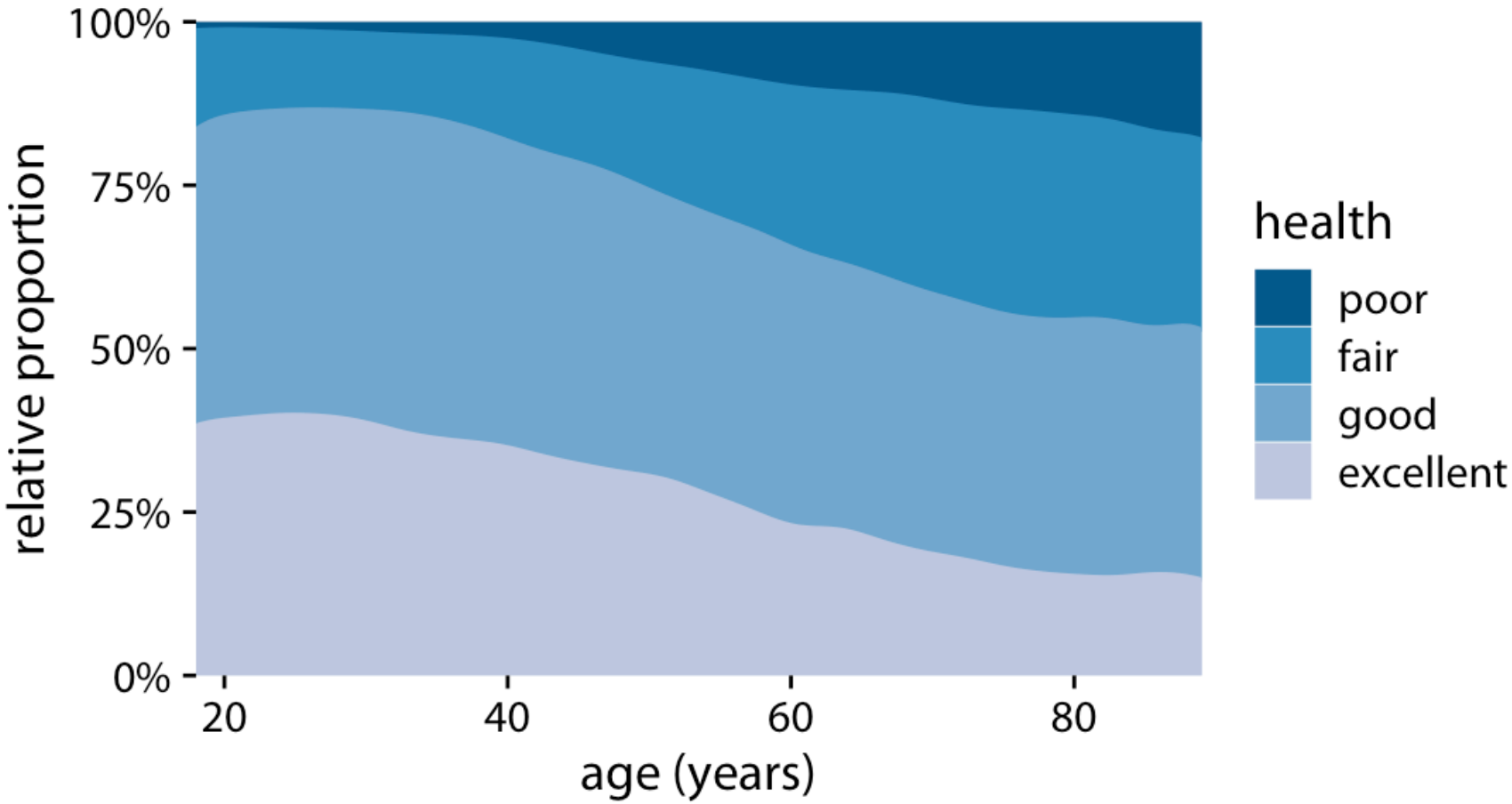
Market share of five hypothetical companies for the years 2015–2017, visualized as side-by-side bars.

Case for stacked bars



Change in the gender composition of the Rwandan parliament over time, 1997 to 2016.

Case for stacked densities

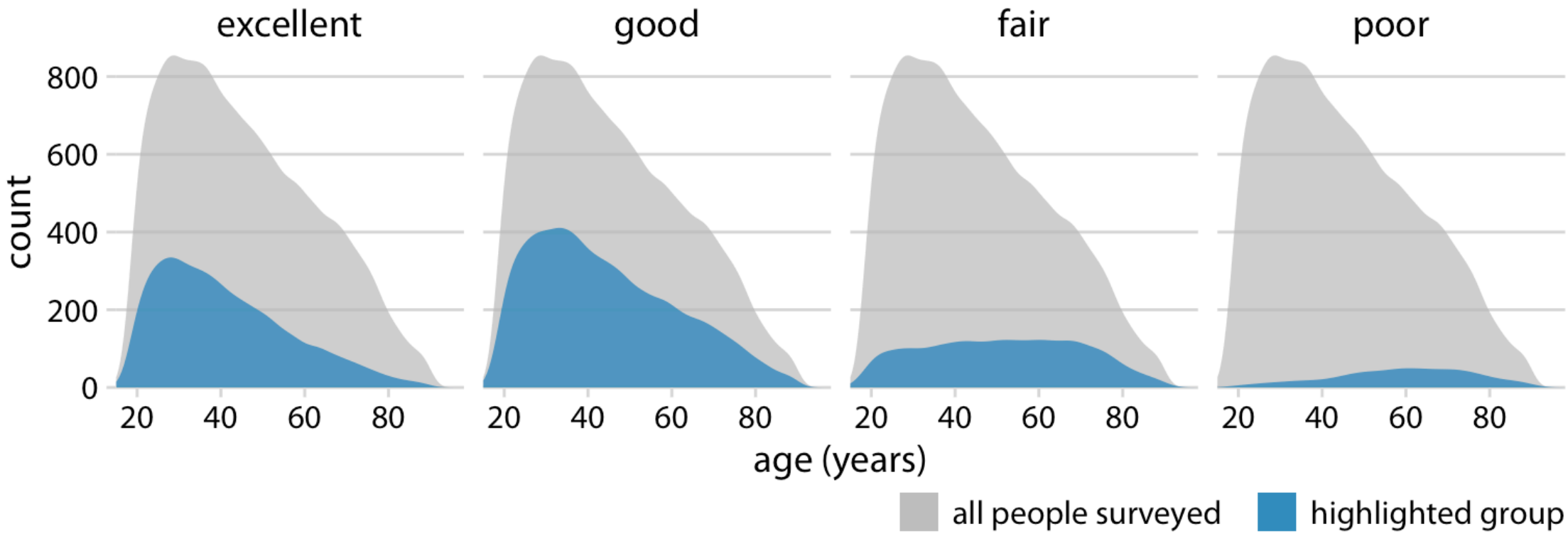


Health status by age, as reported by the general social survey

Visualizing Proportions Separately

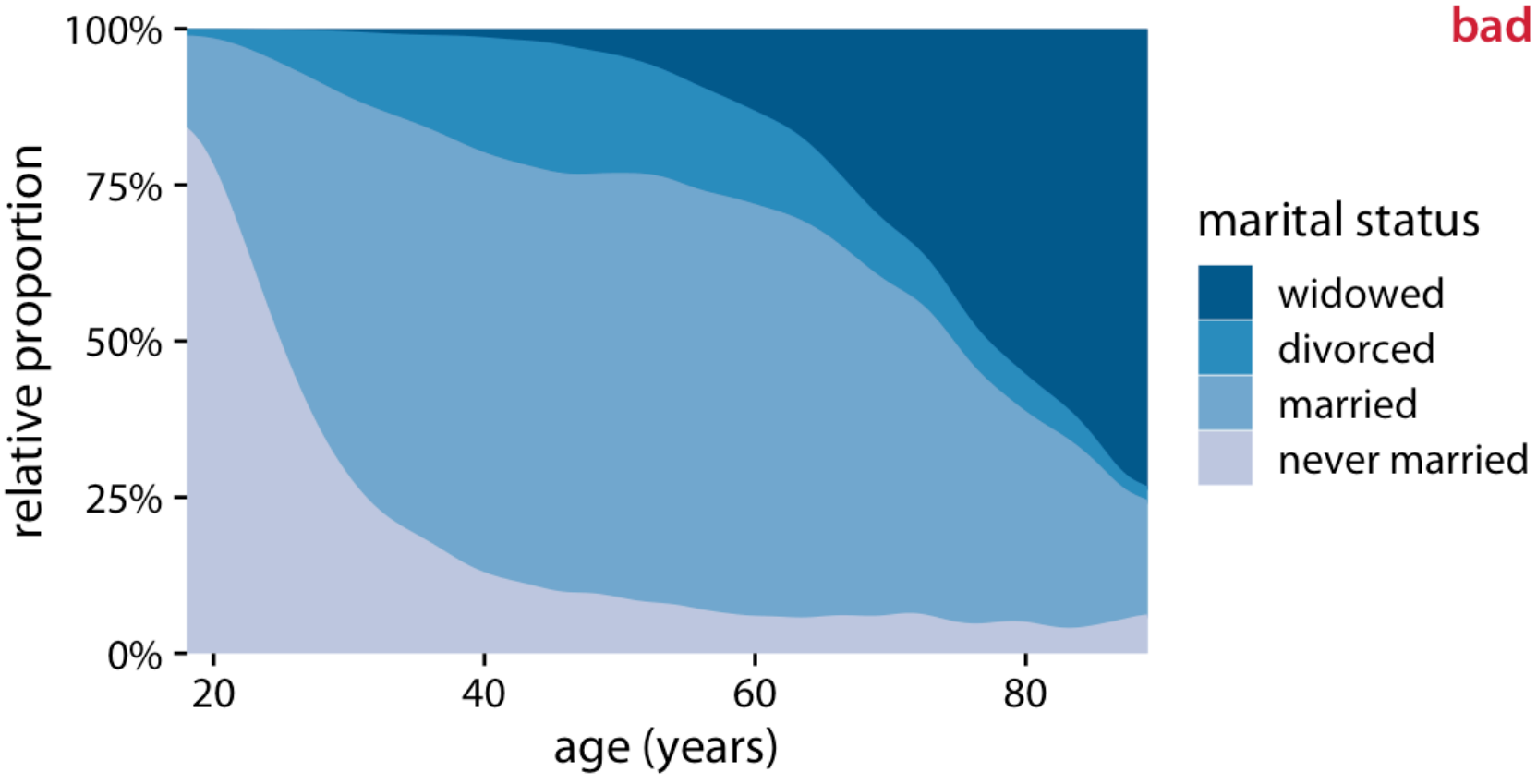
- Example: Health status by age, shown as proportion of the total.
- Advantages of separating proportions.
- Example: Marital status by age, shown as proportion of the total.

Visualizing proportions as parts of the total



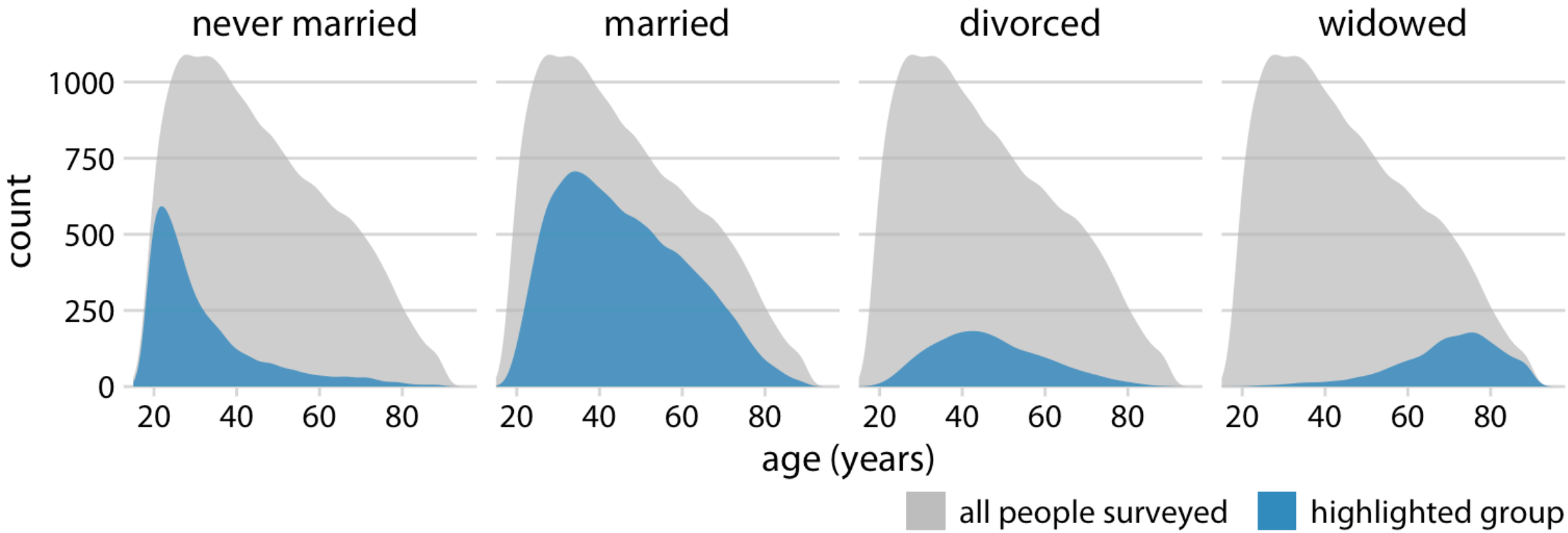
Health status by age, proportions of the total survey population, with colored areas indicating density estimates of ages for each health status and gray areas showing the overall age distribution.

Visualizing proportions as parts of the total
(continued)



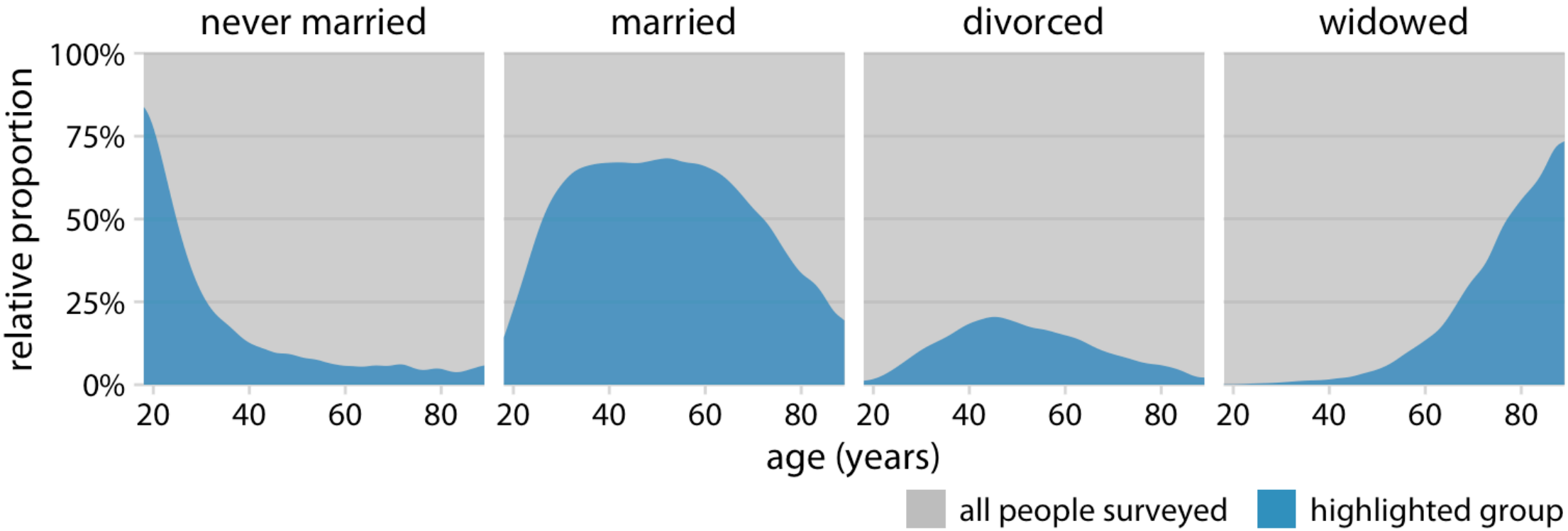
Marital status by age, labeled as "bad" due to significant distortion caused by varying frequencies of never married and widowed individuals across age groups.

Visualizing proportions as parts of the total
(continued)



Marital status by age, with colored areas representing density estimates of ages for each marital status and gray areas indicating overall age distribution.

Visualizing proportions as parts of the total
(continued)



Marital status by age, with blue areas representing the percentage of people at each age with the respective status, and gray areas representing all other marital statuses.