Quantity Visualization

Part I: Visualizing Amounts

Introduction

- Importance of visualizing quantitative data across different categories.
- Common visualization techniques: bar plots, dot plots, heatmaps

Understanding Bar Plots

- Definition and purpose of bar plots in visualizing amounts
- Example: Highest grossing movies
- Vertical vs. horizontal bar plots
- Importance of arranging bars in meaningful order

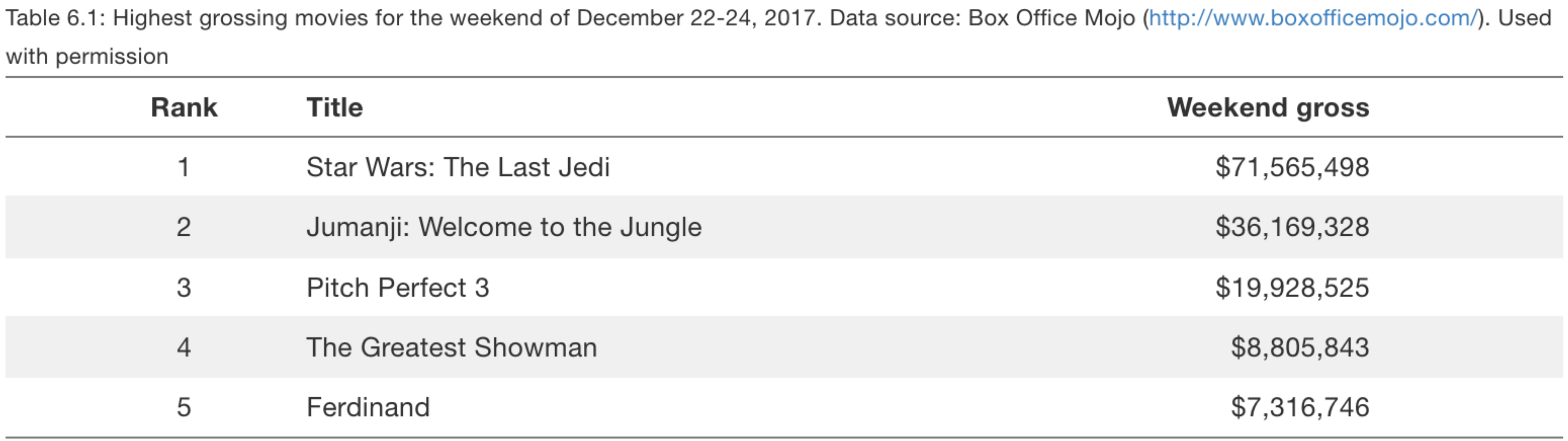
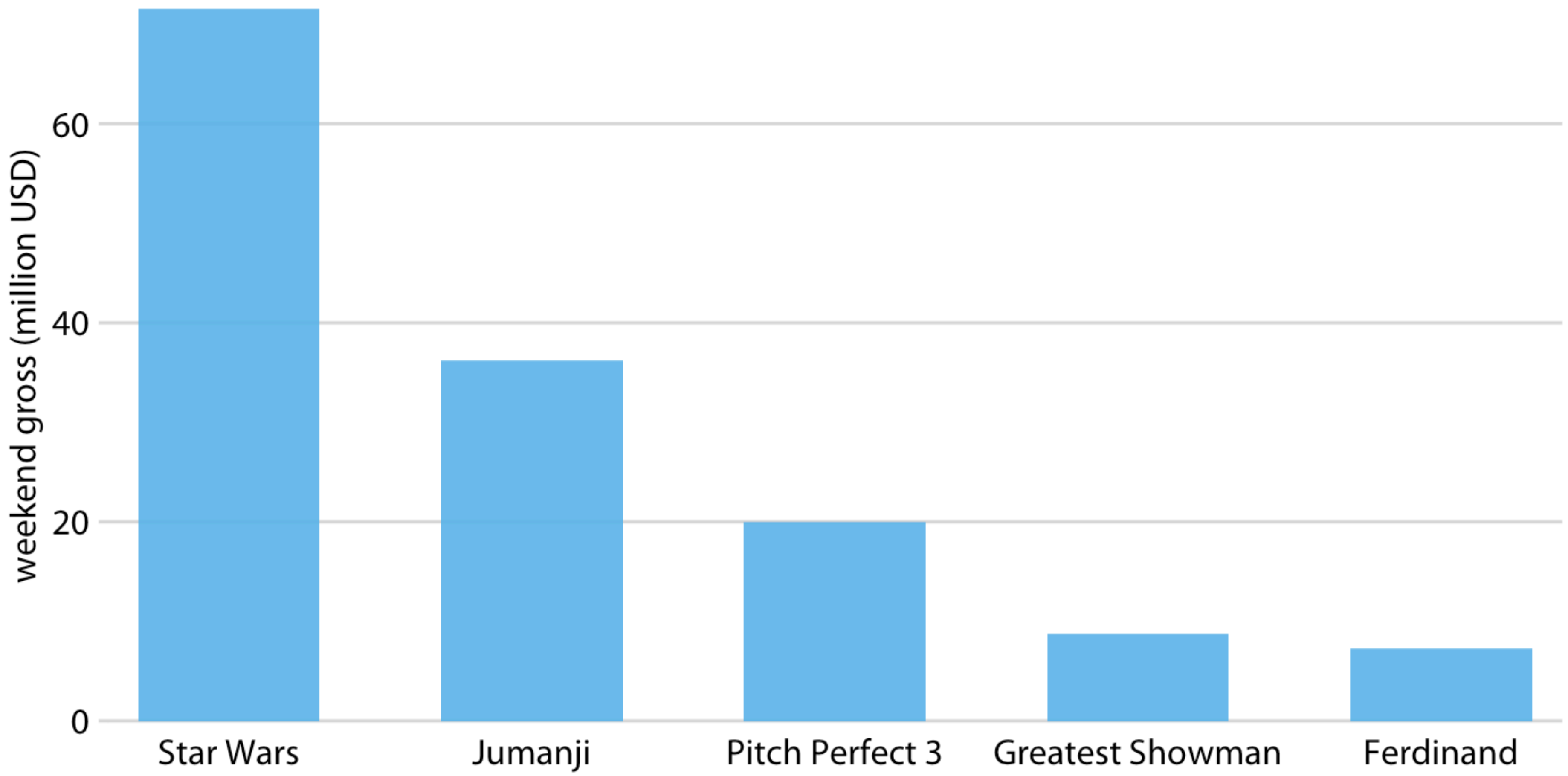


Table 6.1: Highest grossing movies for the weekend of December 22-24, 2017. Data source: Box Office Mojo (http://www.boxofficemojo.com/). Used with permission

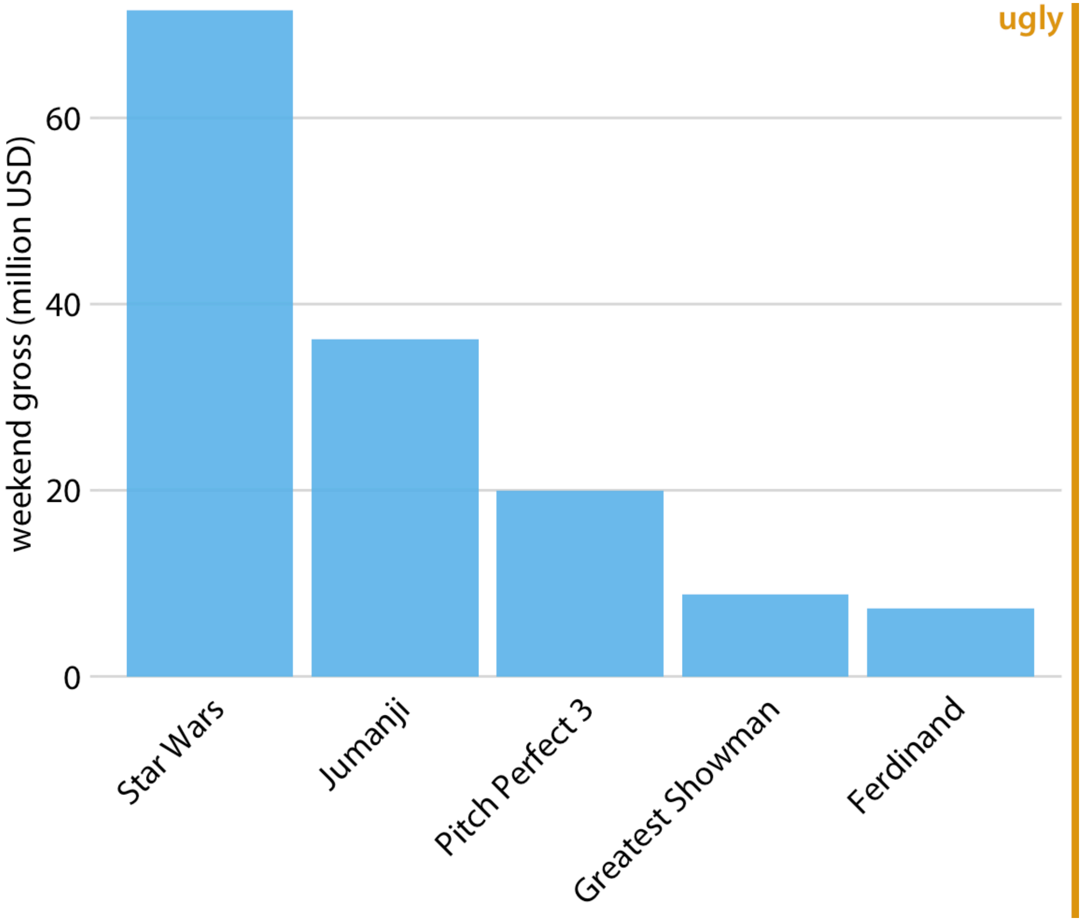
Rank	Title	Weekend gross
1	Star Wars: The Last Jedi	$71,565,498
2	Jumanji: Welcome to the Jungle	$36,169,328
3	Pitch Perfect 3	$19,928,525
4	The Greatest Showman	$8,805,843
5	Ferdinand	$7,316,746

Understanding Bar Plots (continued)

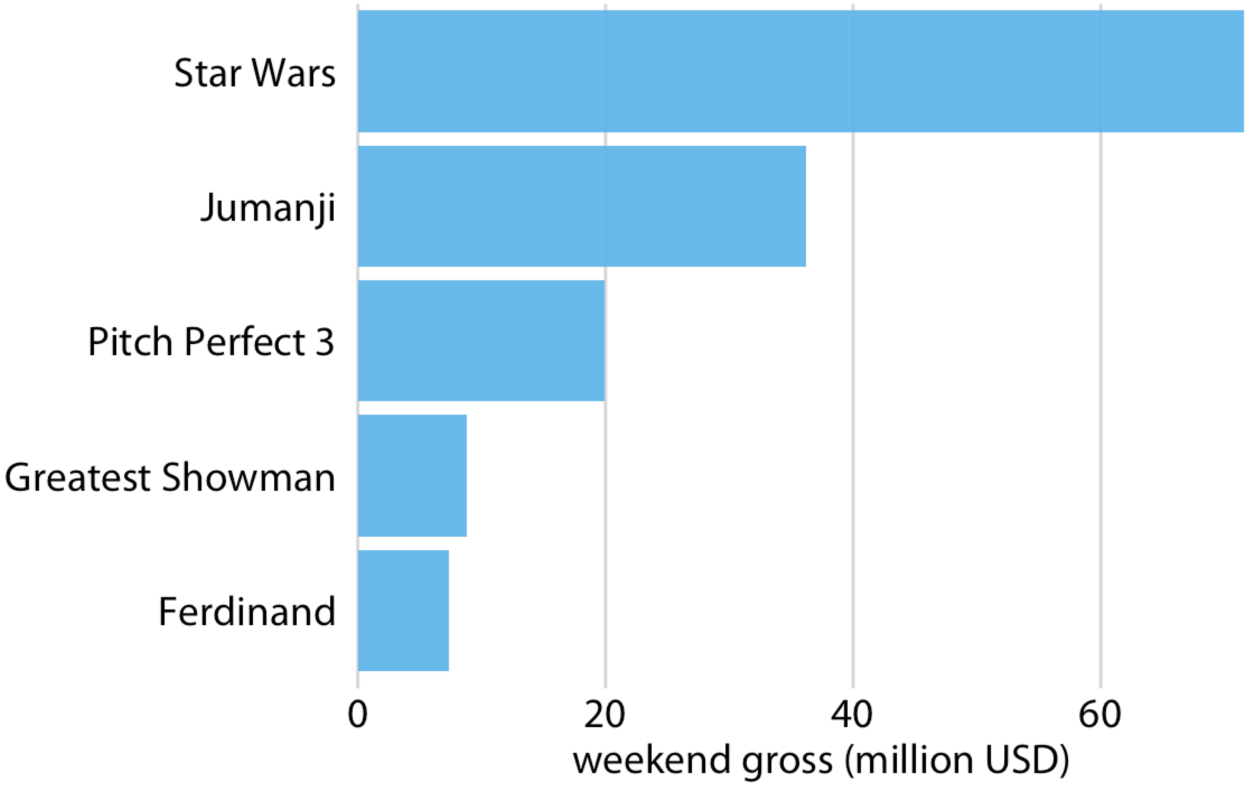


Highest grossing movies for the weekend of December 22-24, 2017, displayed as a bar plot

Understanding Bar Plots (continued)

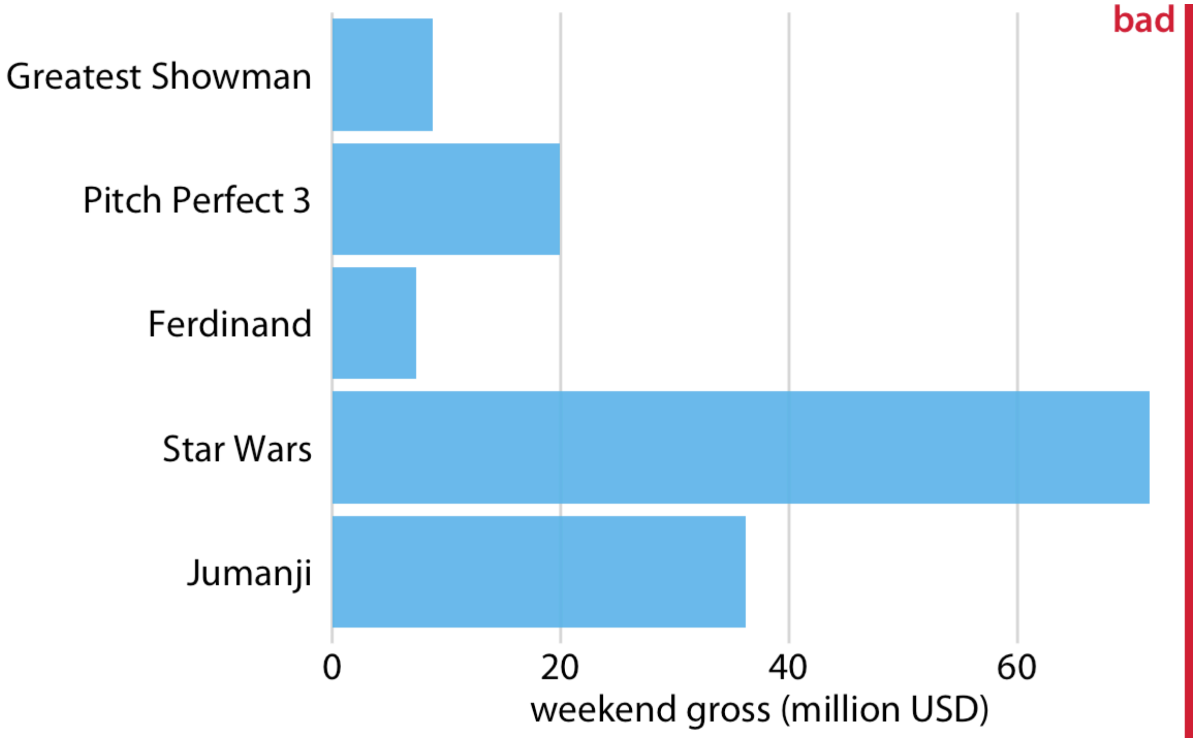


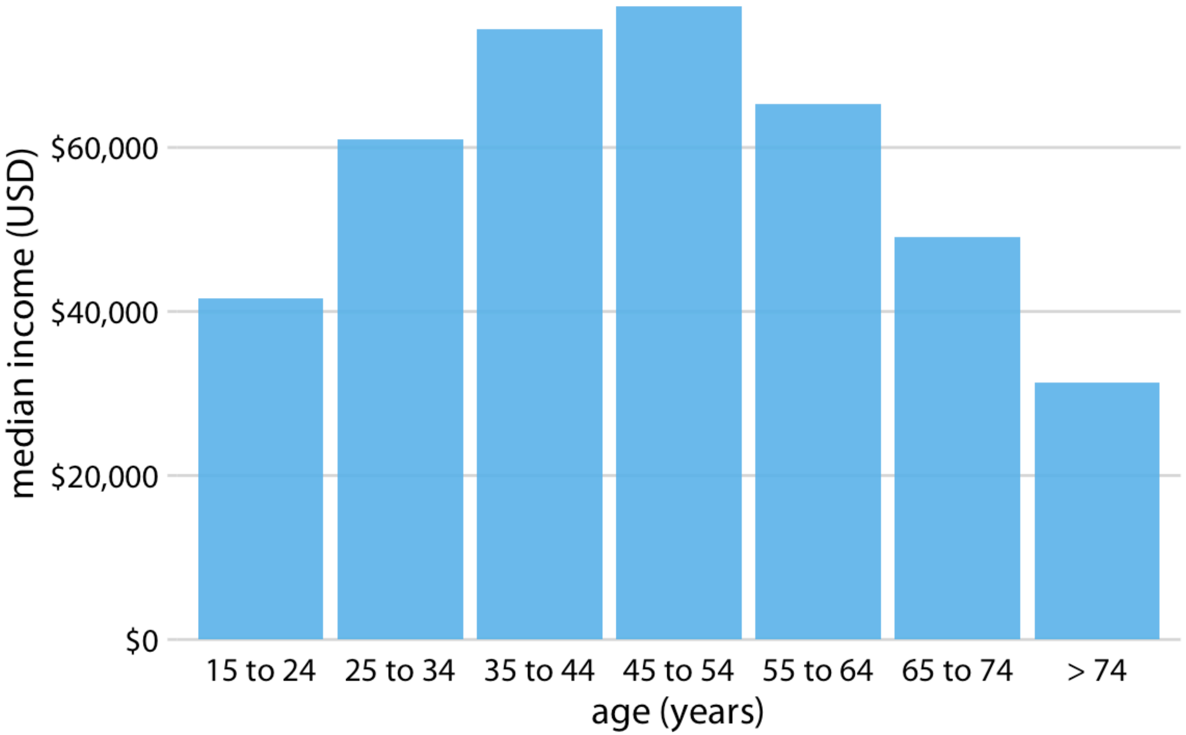
Rotated axis tick labels tend to be difficult to read and require awkward space use underneath the plot.



Highest grossing movies for the weekend of December 22-24, 2017, displayed as a horizontal bar plot

Understanding Bar Plots (continued)

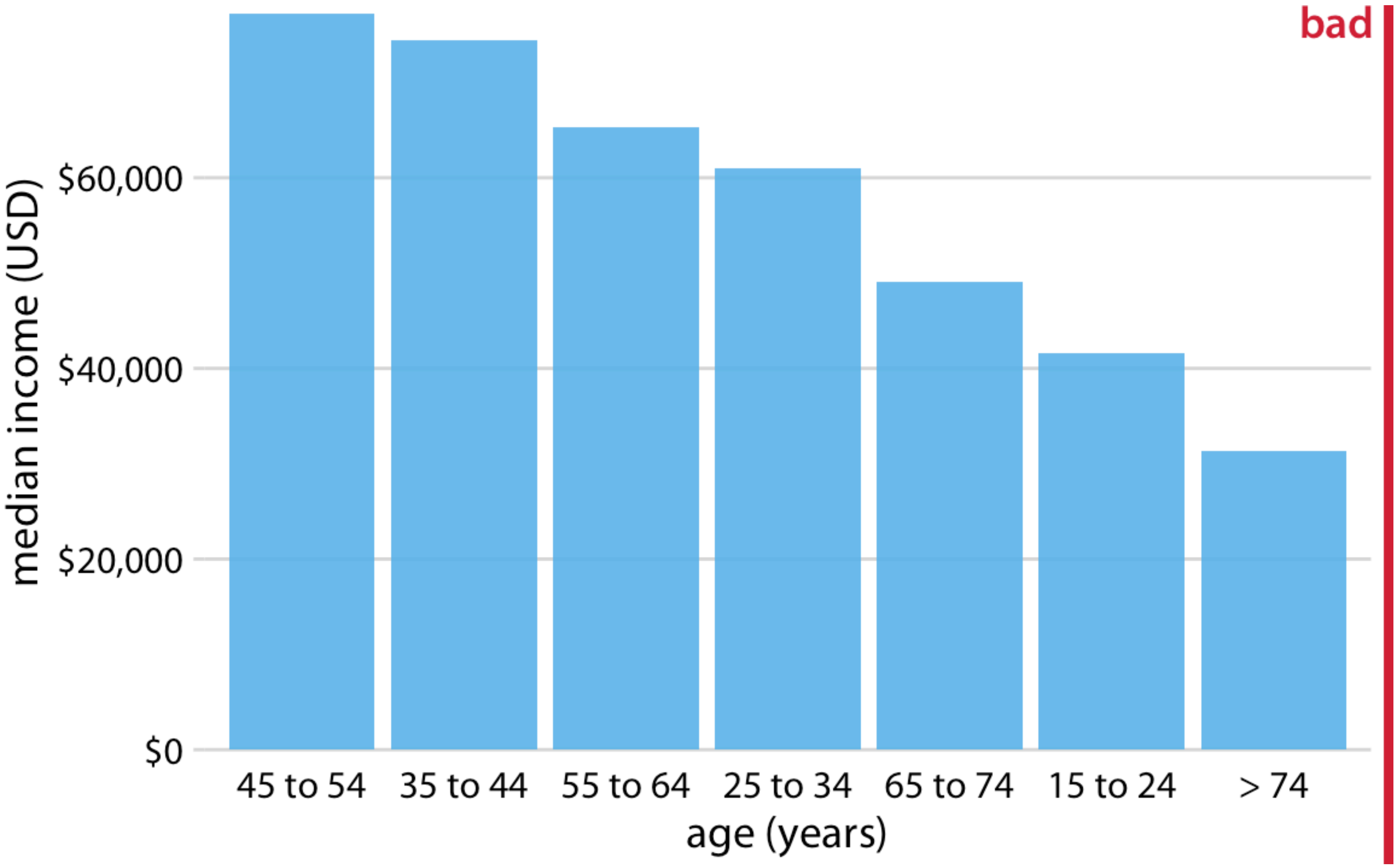




This arrangement of bars is arbitrary, it doesn't serve a meaningful purpose, and it makes the resulting figure much less intuitive

2016 median U.S. annual household income versus age group. The 45–54 year age group has the highest median income.

Understanding Bar Plots (continued)

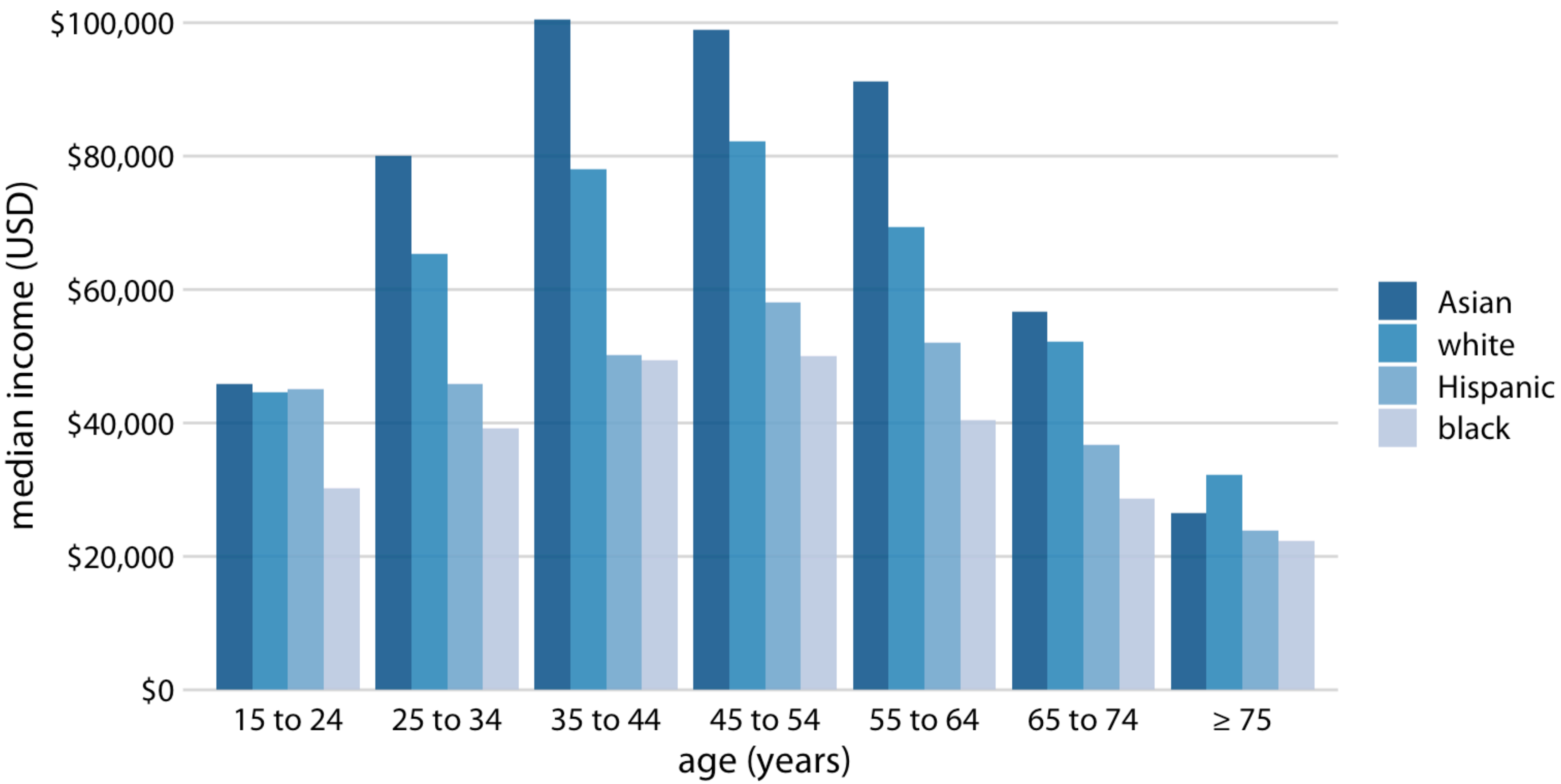


2016 median U.S. annual household income versus age group, sorted by income. While this order of bars looks visually appealing, the order of the age groups is now confusing.

Grouped and Stacked Bars

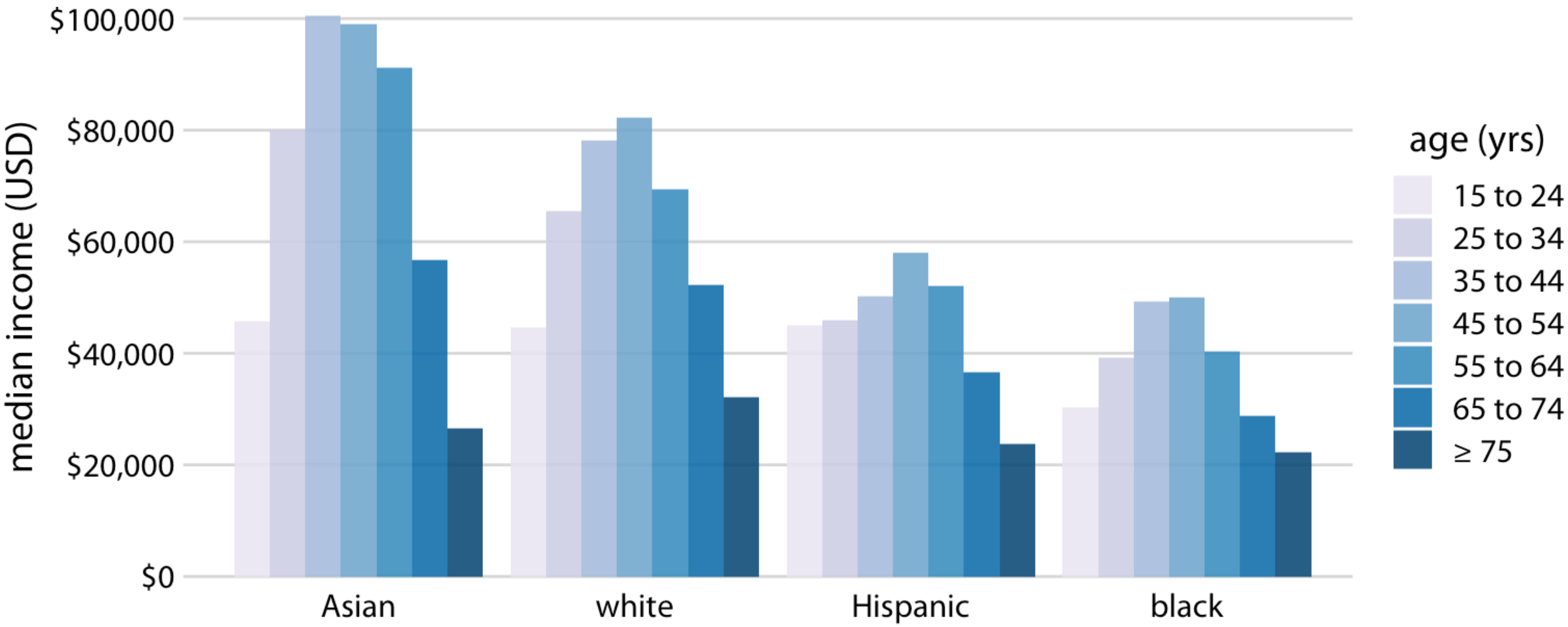
- Visualizing two categorical variables simultaneously
- Example: Median U.S. annual household income versus age group and race.
- Comparison between grouped and separate bar plots
- When to use stacked bars

Grouped and Stacked Bars (continued)



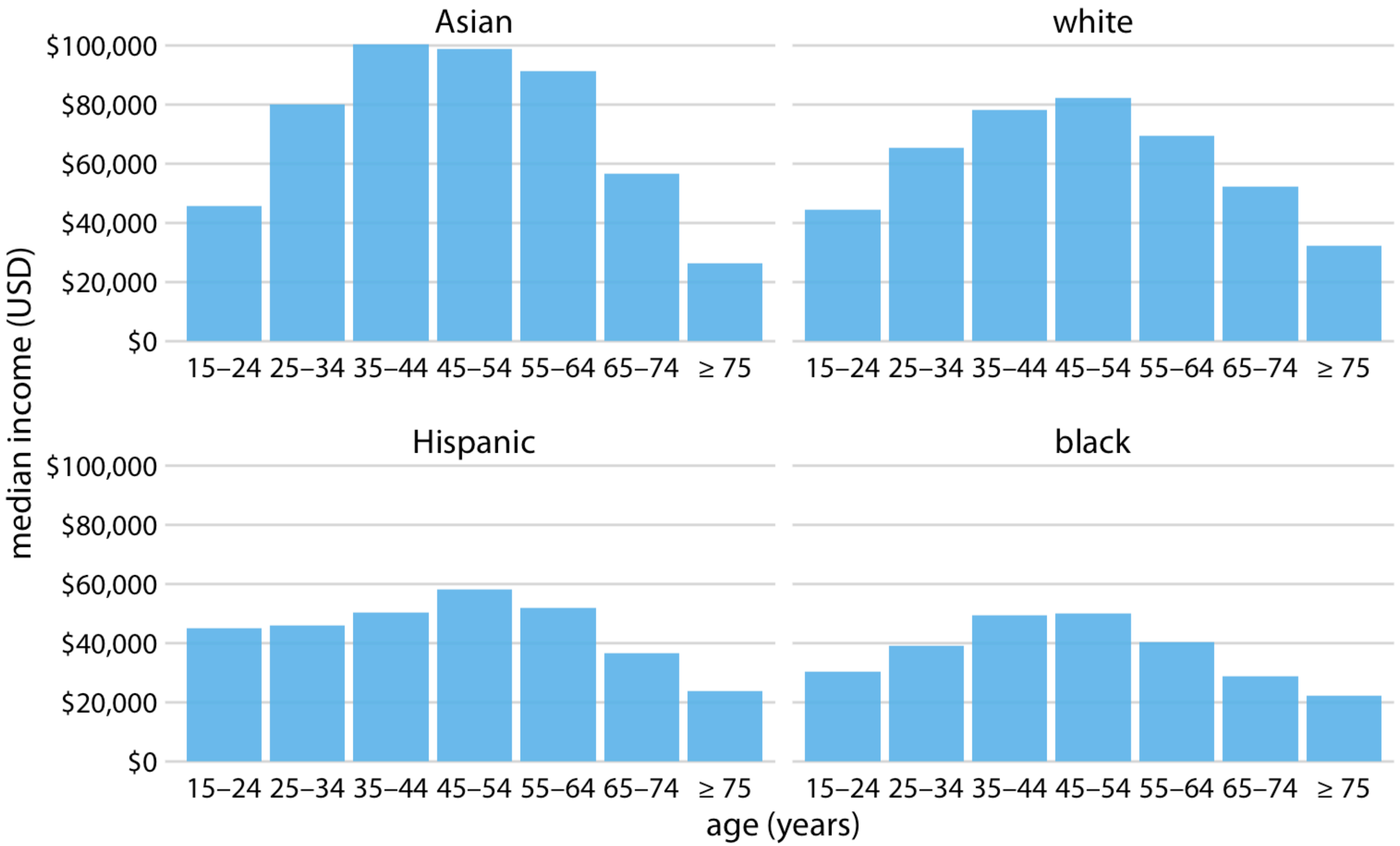
2016 median U.S. annual household income versus age group and race. Age groups are shown along the *x* axis, and for each age group there are four bars, corresponding to the median income of Asian, white, Hispanic, and black people, respectively.

Grouped and Stacked Bars (continued)



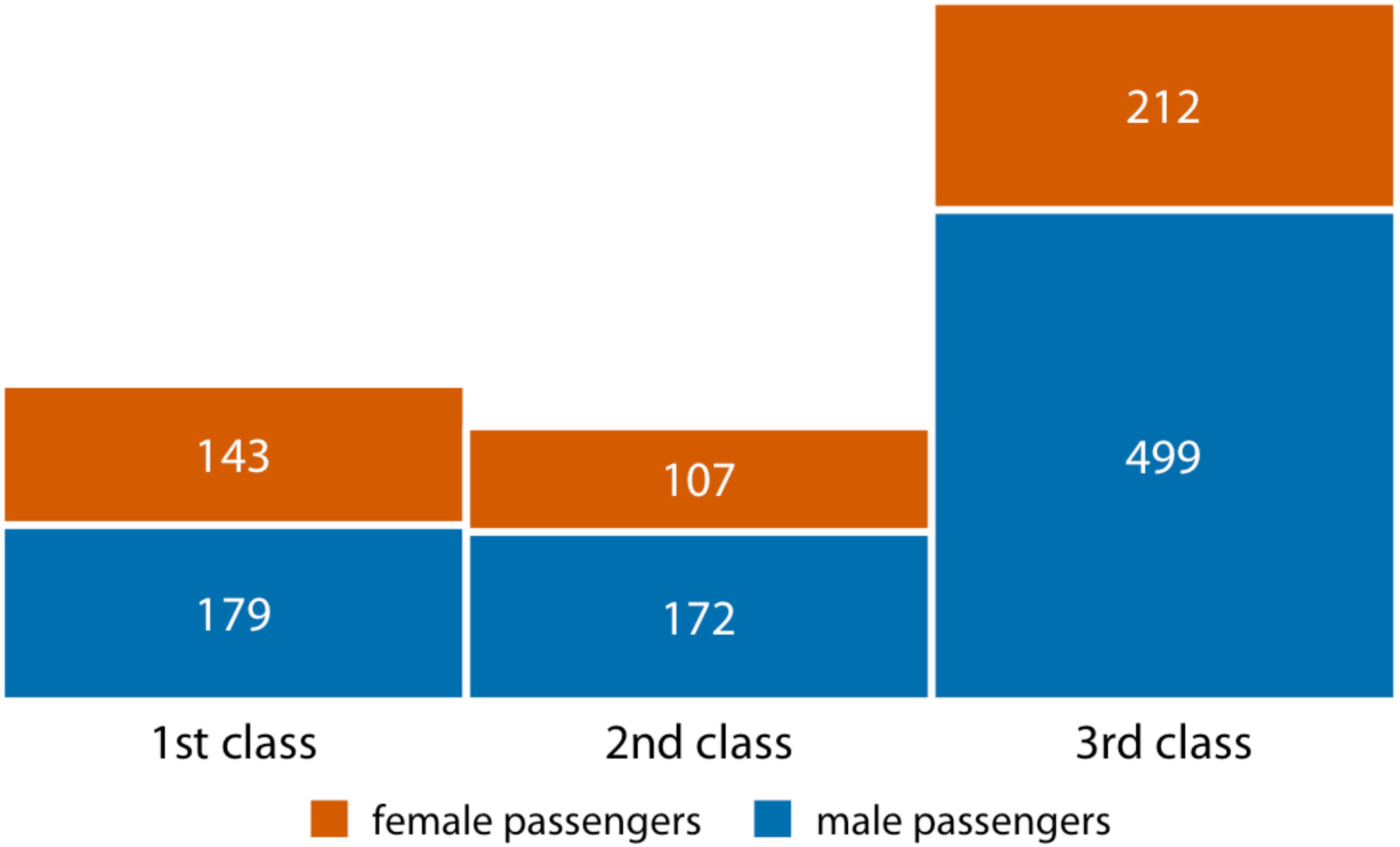
Now race is shown along the *x* axis, and for each race we show seven bars according to the seven age groups.

Grouped and Stacked Bars (continued)



Instead of displaying this data as a grouped bar plot, now the data is shown as four separate regular bar plots. This choice has the advantage that we don't need to encode either categorical variable by bar color.

Grouped and Stacked Bars (continued)

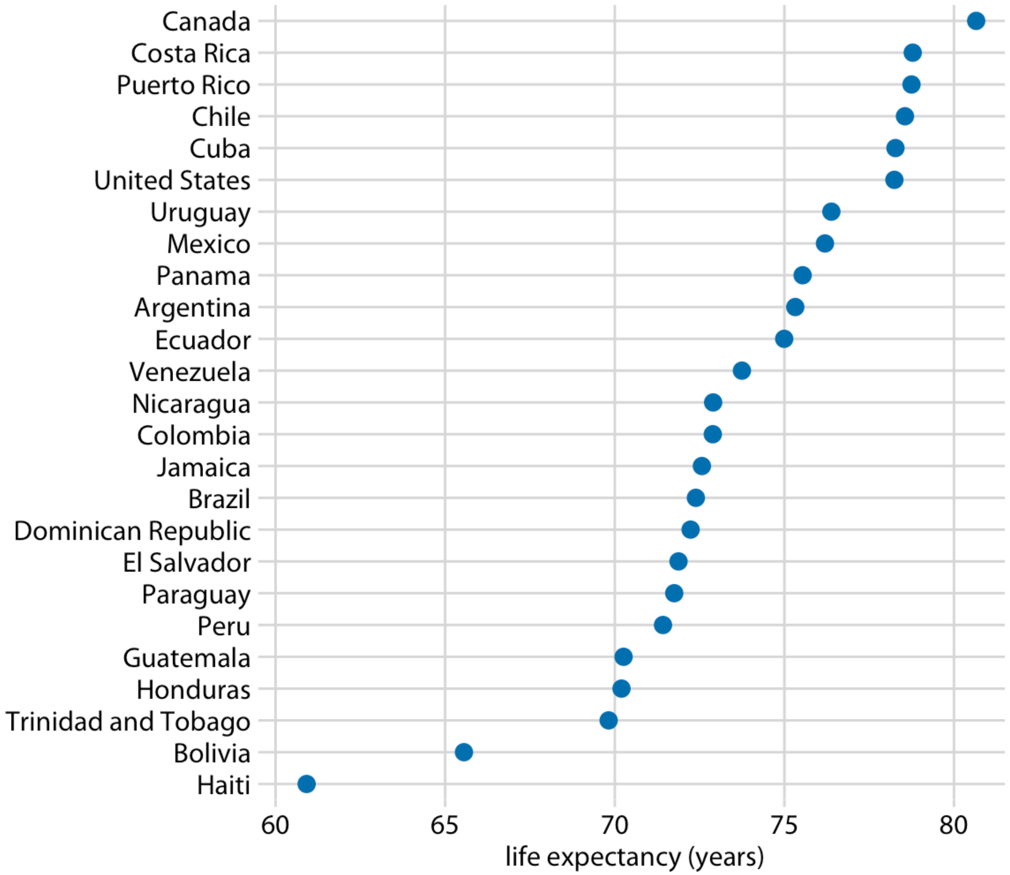


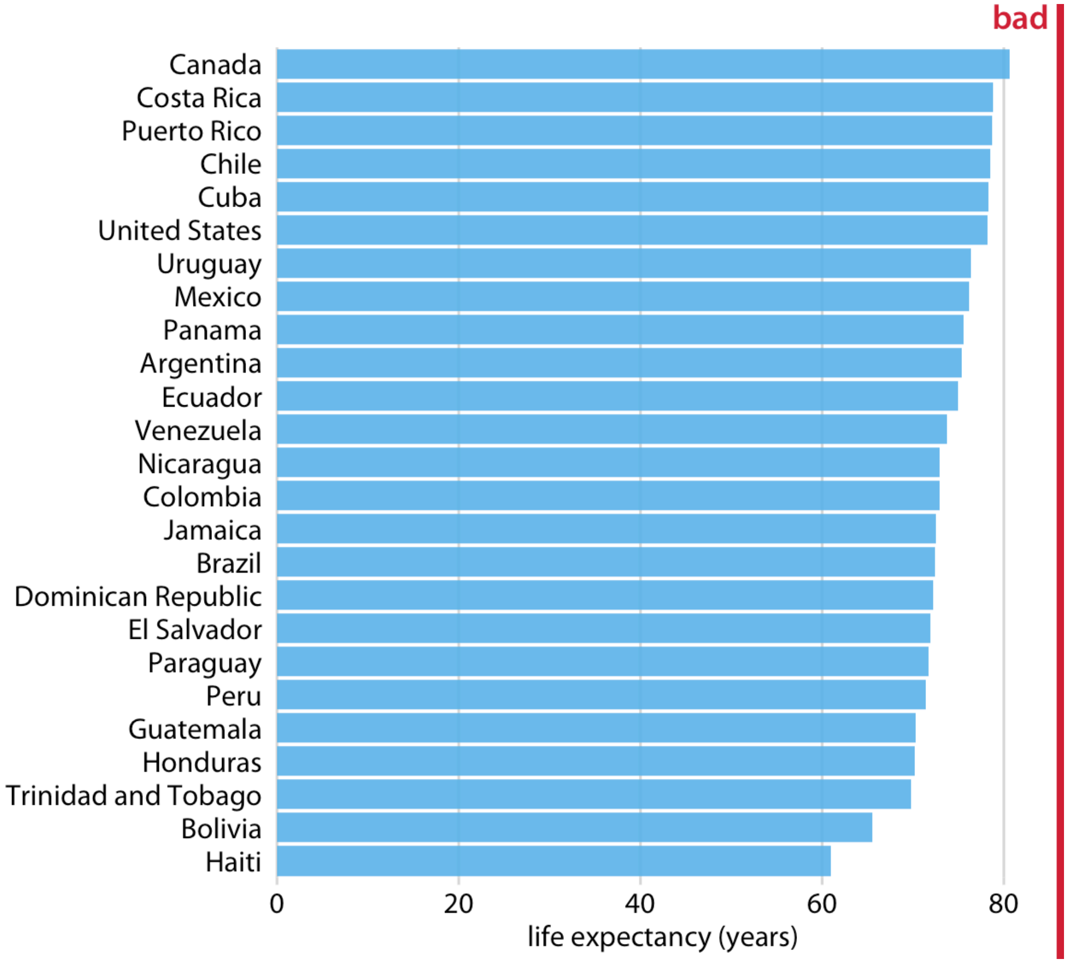
Numbers of female and male passengers on the Titanic traveling in 1st, 2nd, and 3rd class.

Dot Plots and Heatmaps

- Alternatives to Bar Plots: Dot plots and Heatmaps
- Example: Life expectancies of countries in the Americas
- Comparison between bars, dots, and heatmaps
- Importance of data ordering in dot plots and heatmaps

Dot Plots (continued)

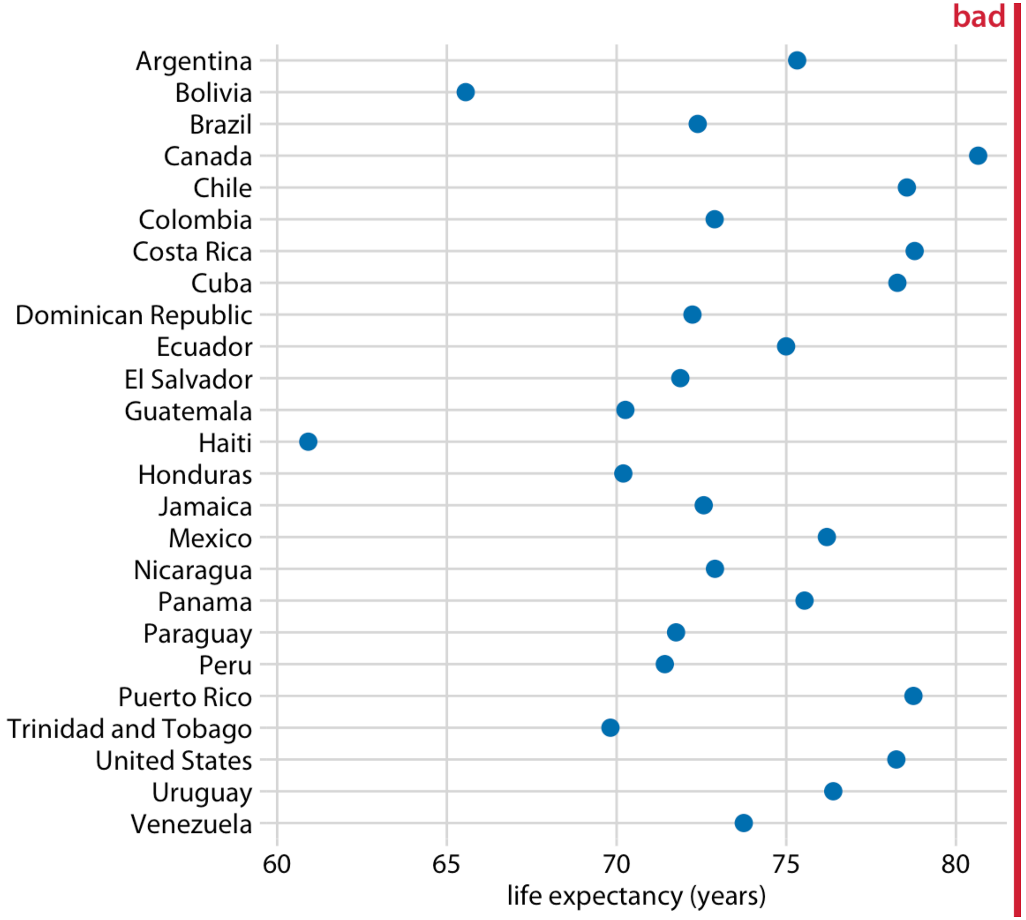




Life expectancies of countries in the Americas, for the year 2007.

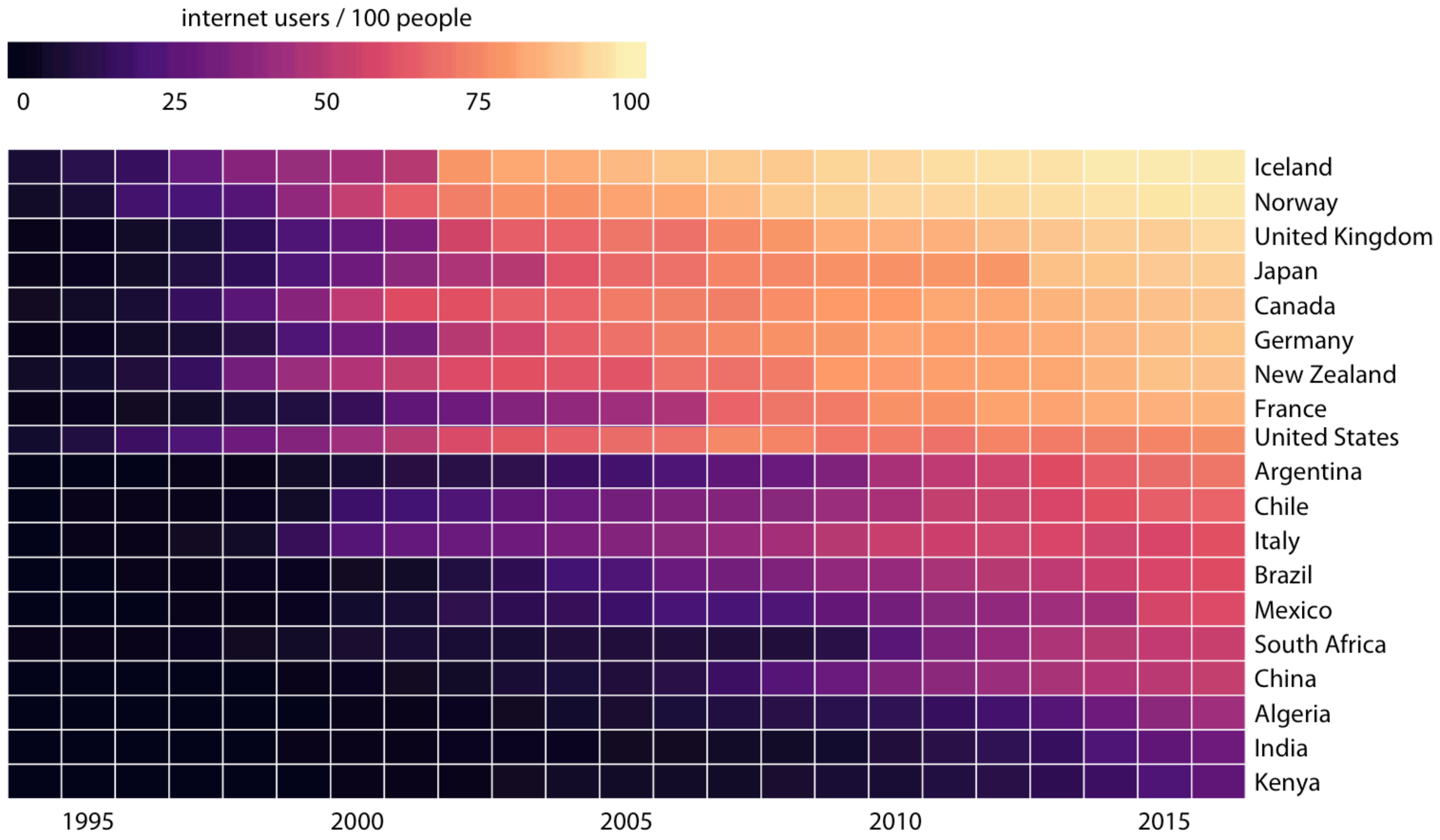
Dataset is not suitable visualization with bars. Bars are too long and draws attention away from differences in life expectancy among the different countries.

Dot Plots (continued)



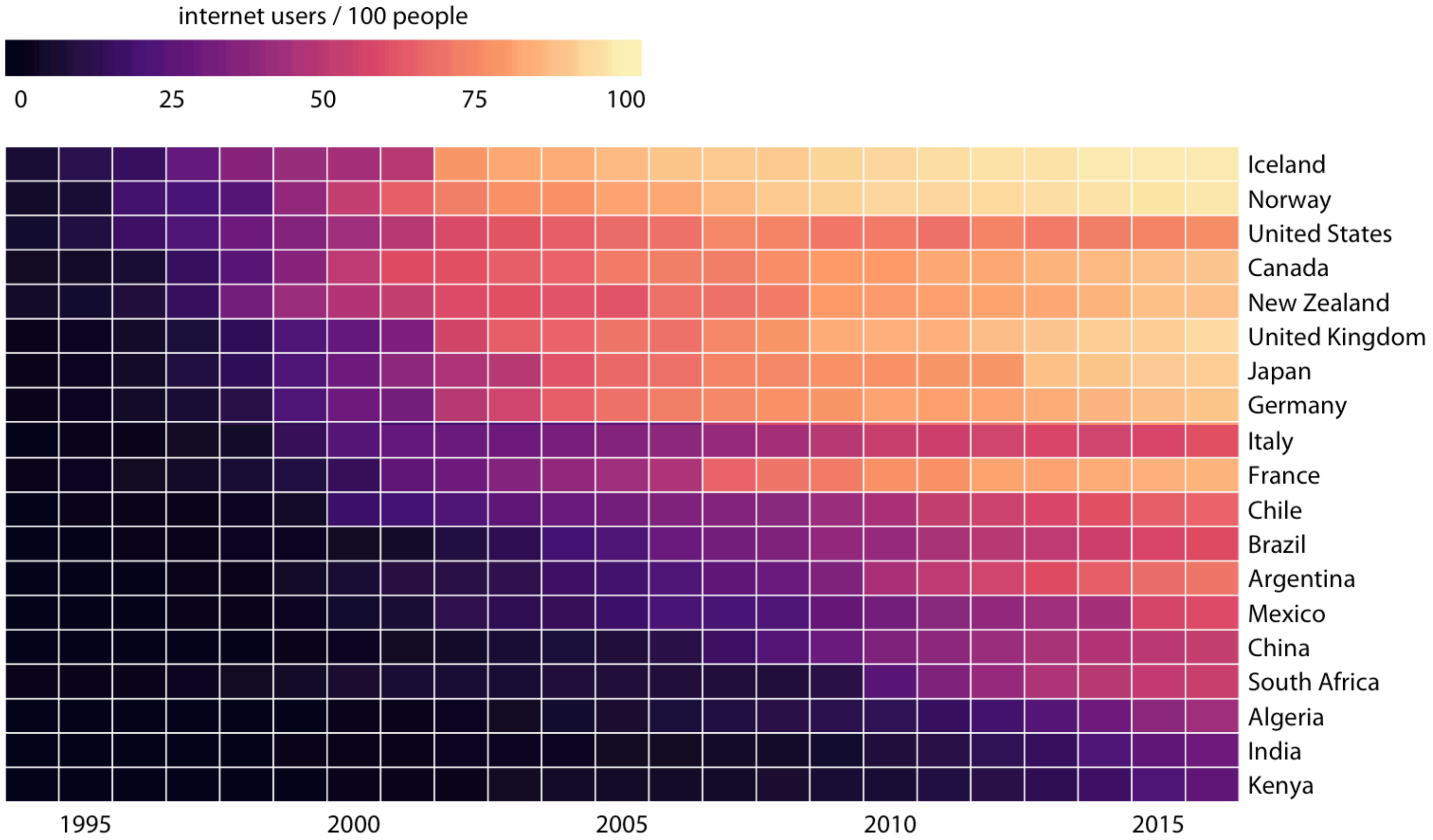
Countries are ordered alphabetically, which causes dots to form a disordered cloud of points. making the figure difficult to read.

Heatmaps (continued)



Internet adoption over time, for select countries. Color represents the percent of internet users for the respective country and year. Countries were ordered by percent internet users in 2016.

Heatmaps (continued)



Internet adoption over time, for select countries. Countries were ordered by the year in which their internet usage first exceeded 20%.

Practical Considerations

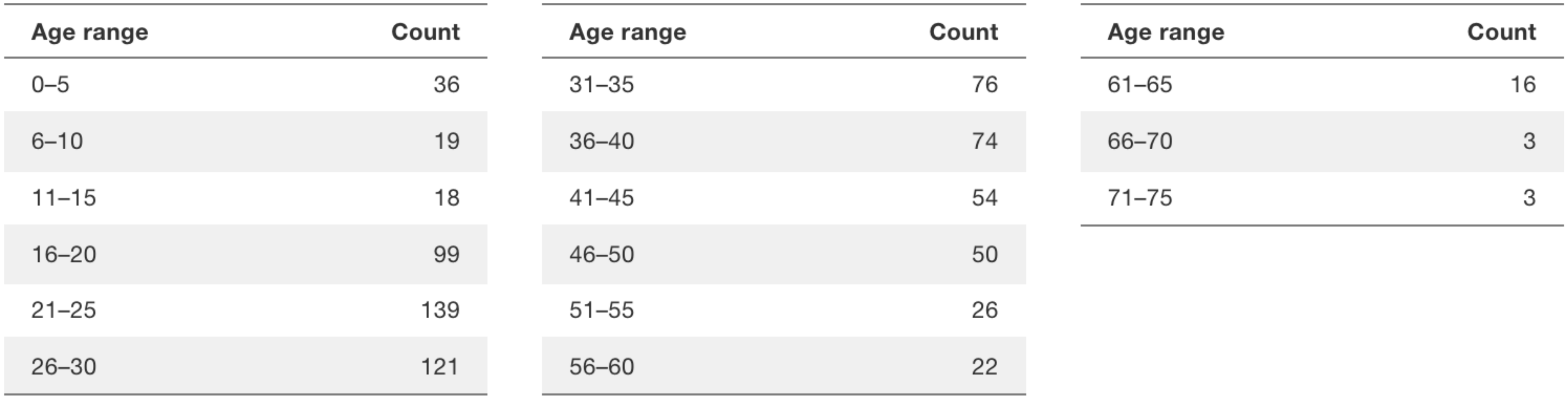
- Importance of clear labeling and axis orientation
- Considerations for choosing the appropriate visualization technique
- Example: Internet adoption over time for select countries
- Choosing the right visualization for the story you want to convey

Histograms and Density Plots

Part II

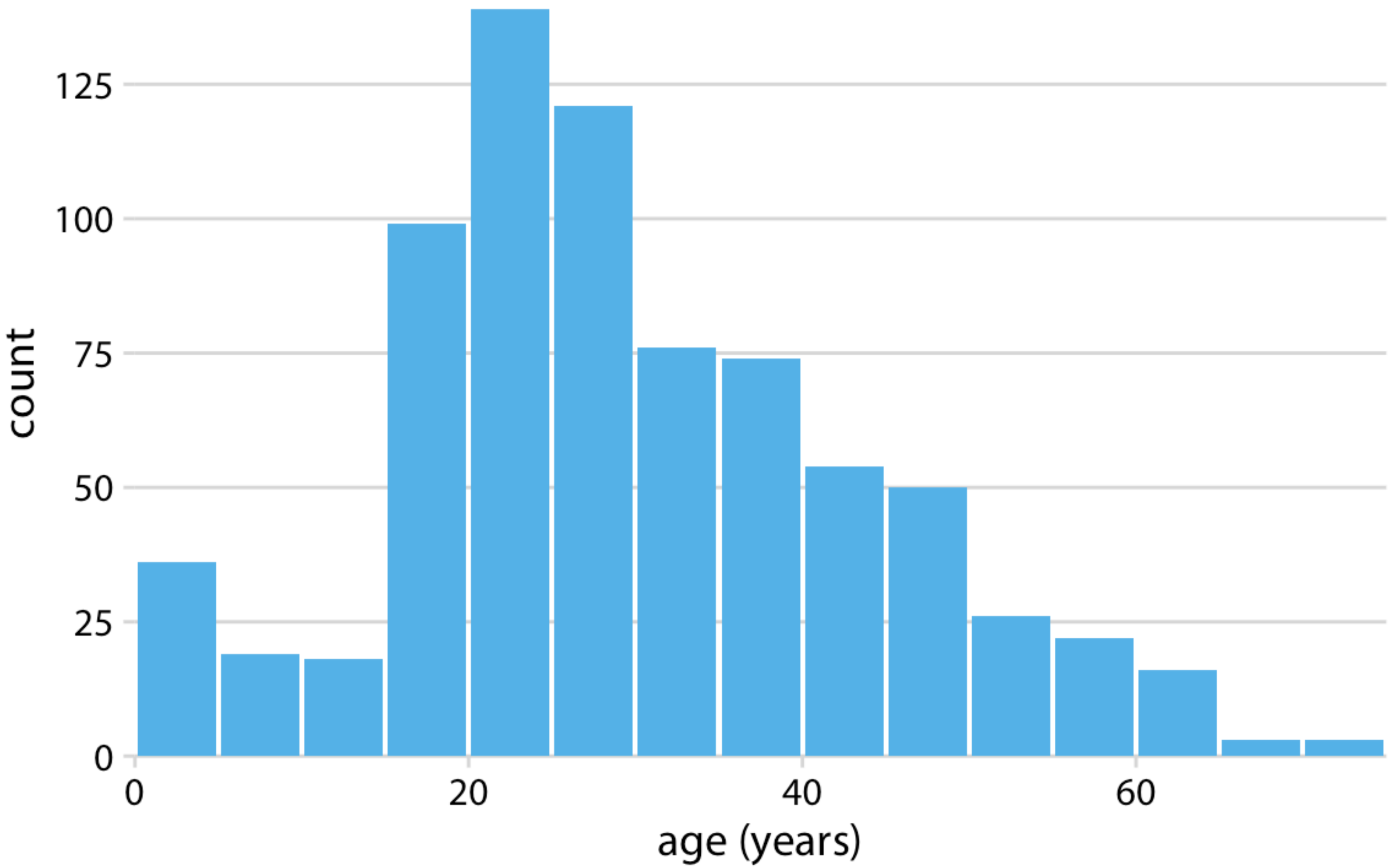
Visualizing a Single Distribution

- Histogram visualization (e.g. age distribution among Titanic passengers)



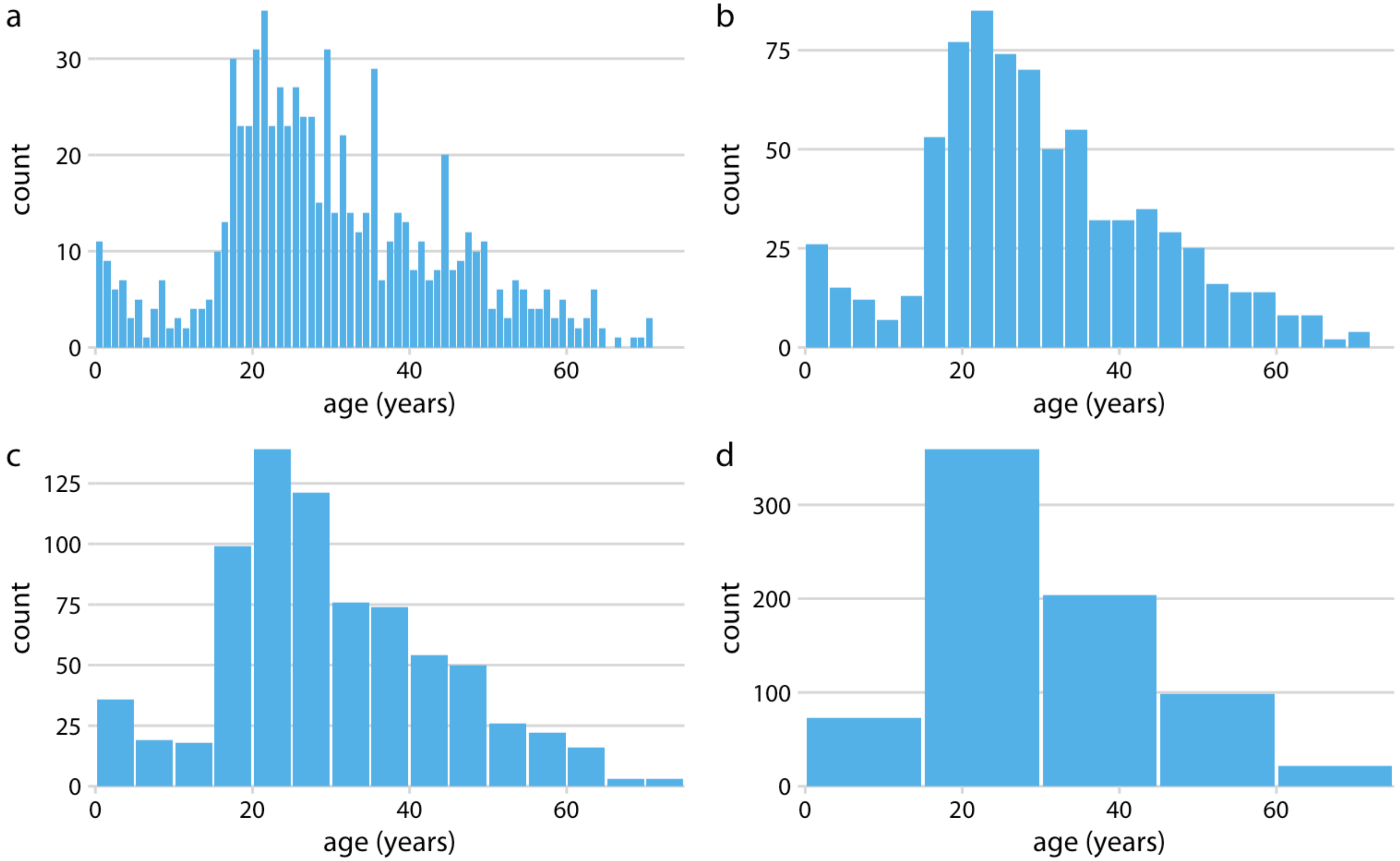
Age range	Count
0–5	36
6–10	19
11–15	18
16–20	99
21–25	139
26–30	121
31–35	76
36–40	74
41–45	54
46–50	50
51–55	26
56–60	22
61–65	16
66–70	3
71–75	3

Single Distribution Visualization (continued)



Histogram of the ages of Titanic passengers

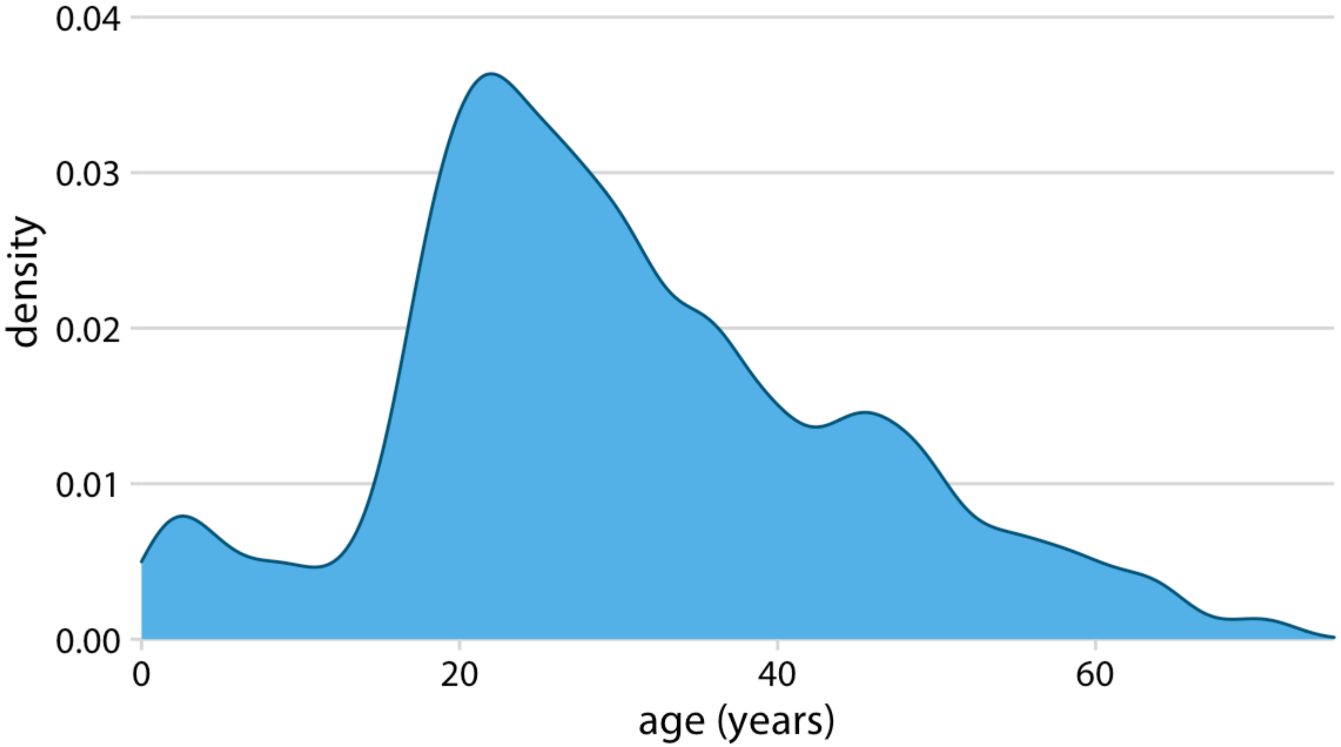
Single Distribution Visualization (continued)



Histograms depend on the chosen bin width. Here, the same age distribution of Titanic passengers is shown with four different bin widths: (a) one year; (b) three years; (c) five years; (d) fifteen years.

Density Plots

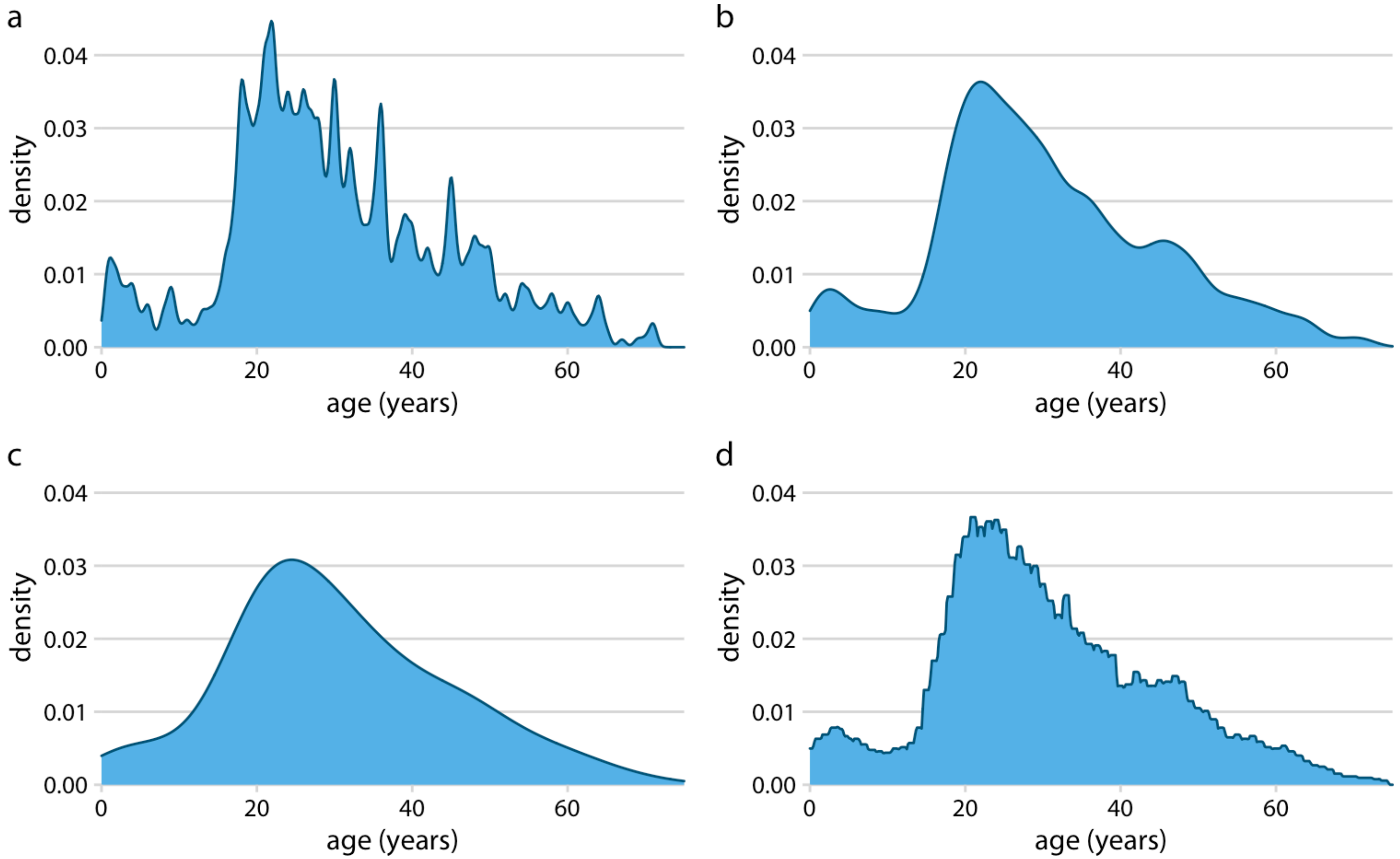
- Kernel density estimation for probability distribution



Kernel density estimate of age distribution of Titanic passengers. Curve height scaled such that the area under the curve equals one. The density estimate was performed with a Gaussian kernel and a bandwidth of 2.

Density Plots (continued)

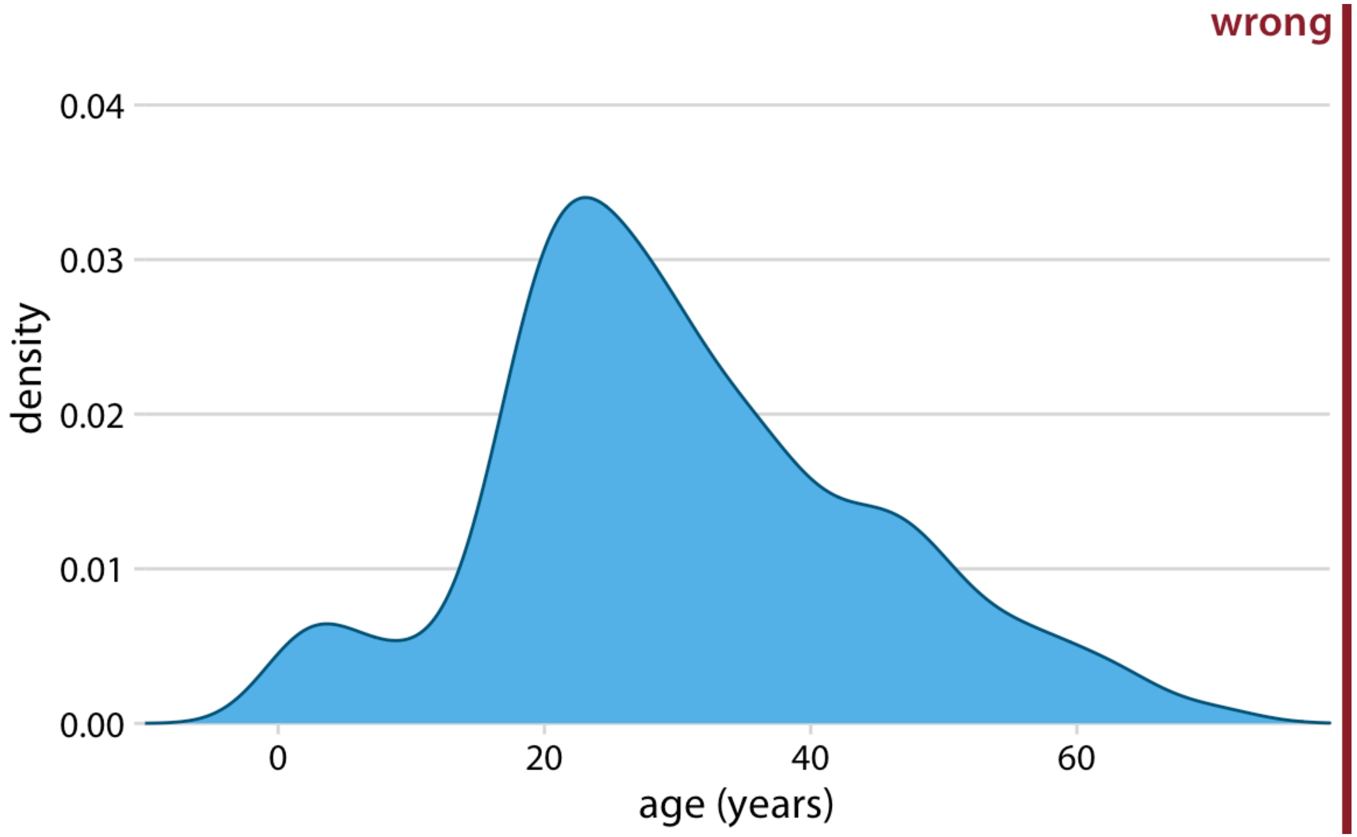
- Impact of bandwidth and kernel choice



Kernel density estimates depend on the chosen kernel and bandwidth. Here, the same age distribution of Titanic passengers is shown for four different combinations of these parameters: (a) Gaussian kernel, bandwidth = 0.5; (b) Gaussian kernel, bandwidth = 2; (c) Gaussian kernel, bandwidth = 5; (d) Rectangular kernel, bandwidth = 2.

Pitfalls of Density Plots

- Example of nonsensical data prediction



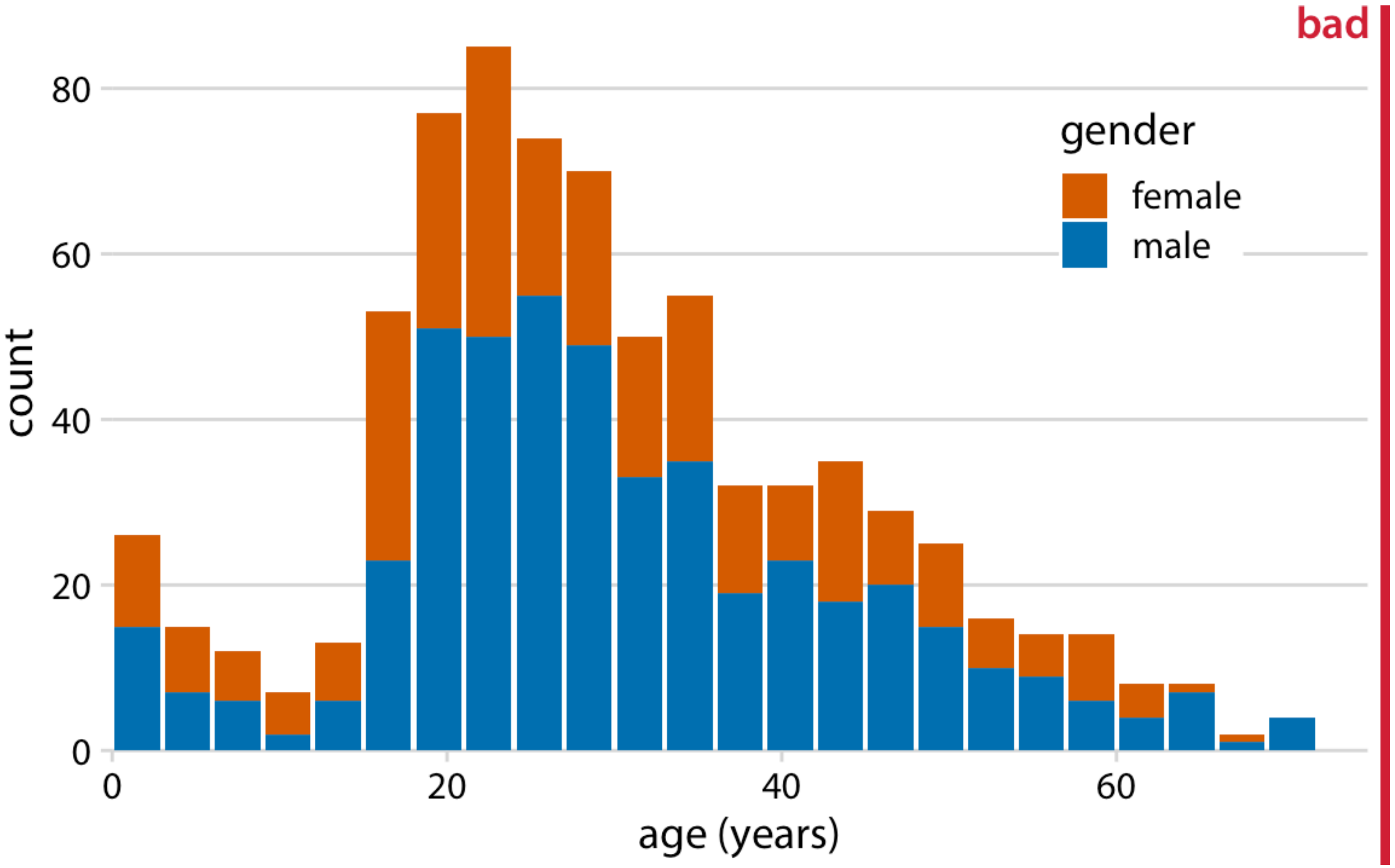
Kernel density estimates can extend the tails of the distribution into areas where no data exist and no data are even possible. Here, the density estimate has been allowed to extend into the negative age range.

Histogram vs. Density Plot

- Considerations for choosing between histogram and density plot
- Alternatives like empirical cumulative density functions
- Exploration based on specific data and goals

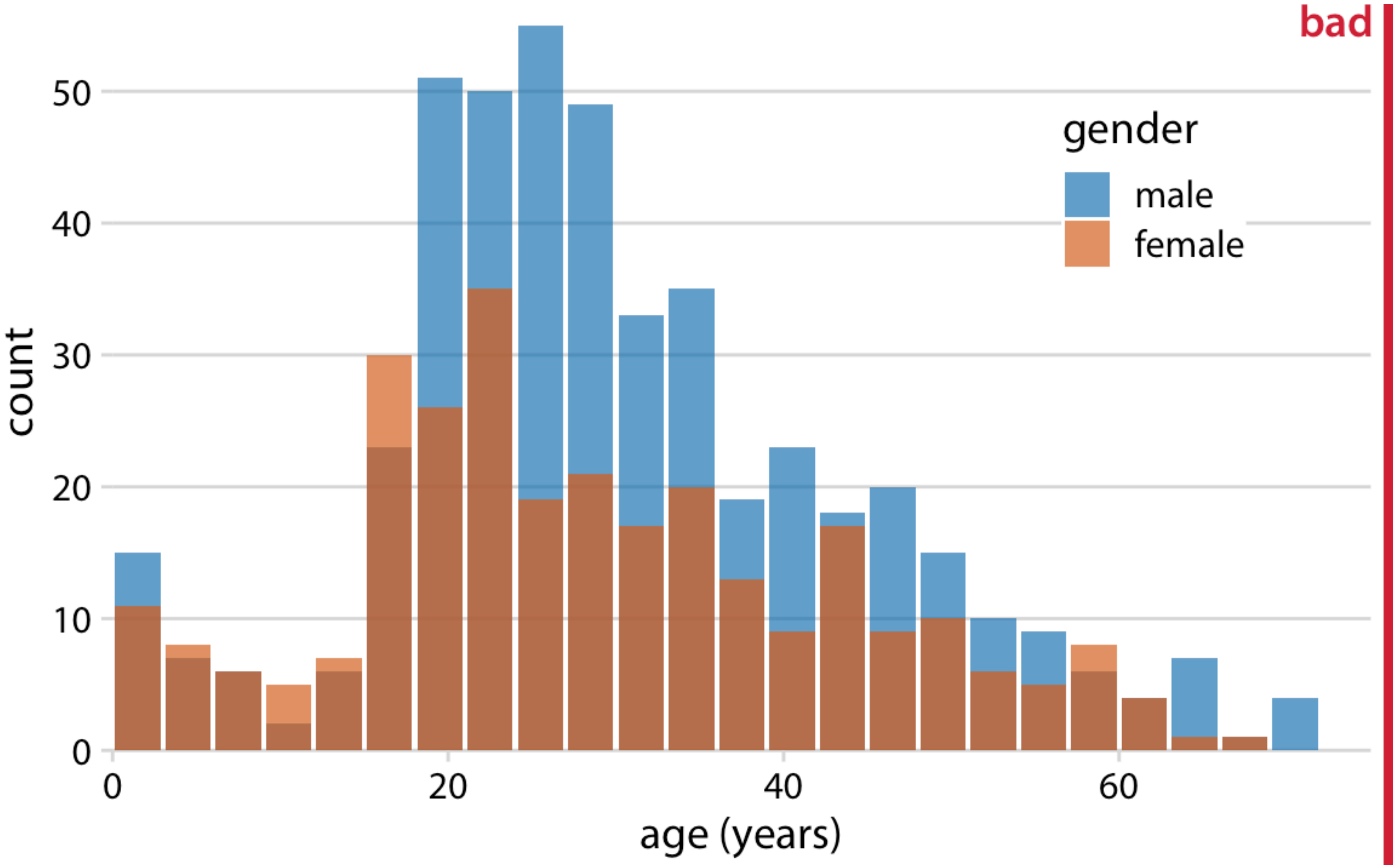
Visualizing Multiple Distributions

- Example: Comparing ages of Titanic passengers by gender



Histogram of the ages of Titanic passengers stratified by gender. This figure has been labeled as “bad” because stacked histograms are easily confused with overlapping histograms, and the heights of the bars representing female passengers cannot easily be compared to each other.

Visualizing Multiple Distributions (continued)

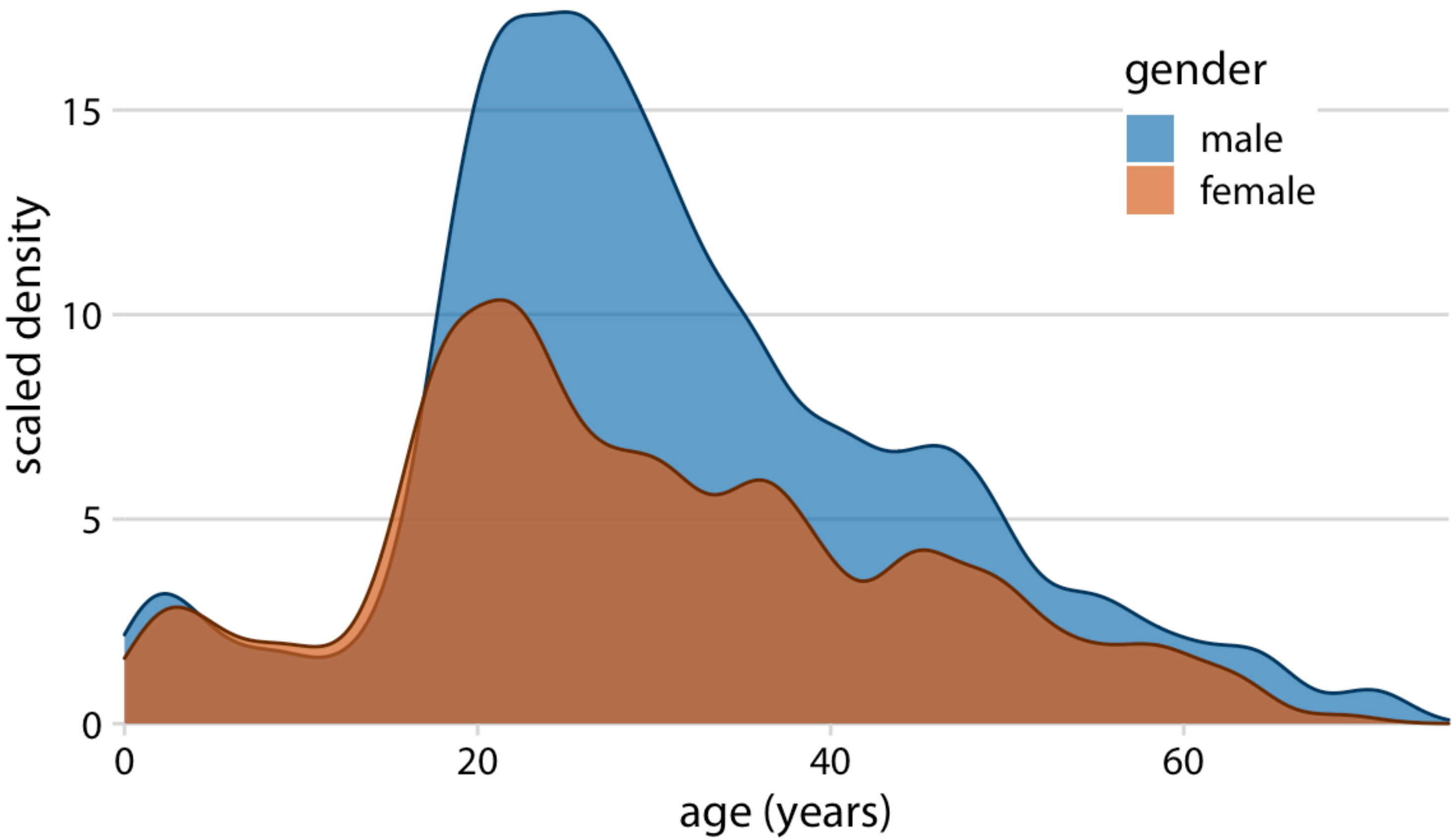


Age distributions of male and female Titanic passengers, shown as two overlapping histograms. This figure has been labeled as “bad” because there is no clear visual indication that all blue bars start at a count of 0.

Better Approaches

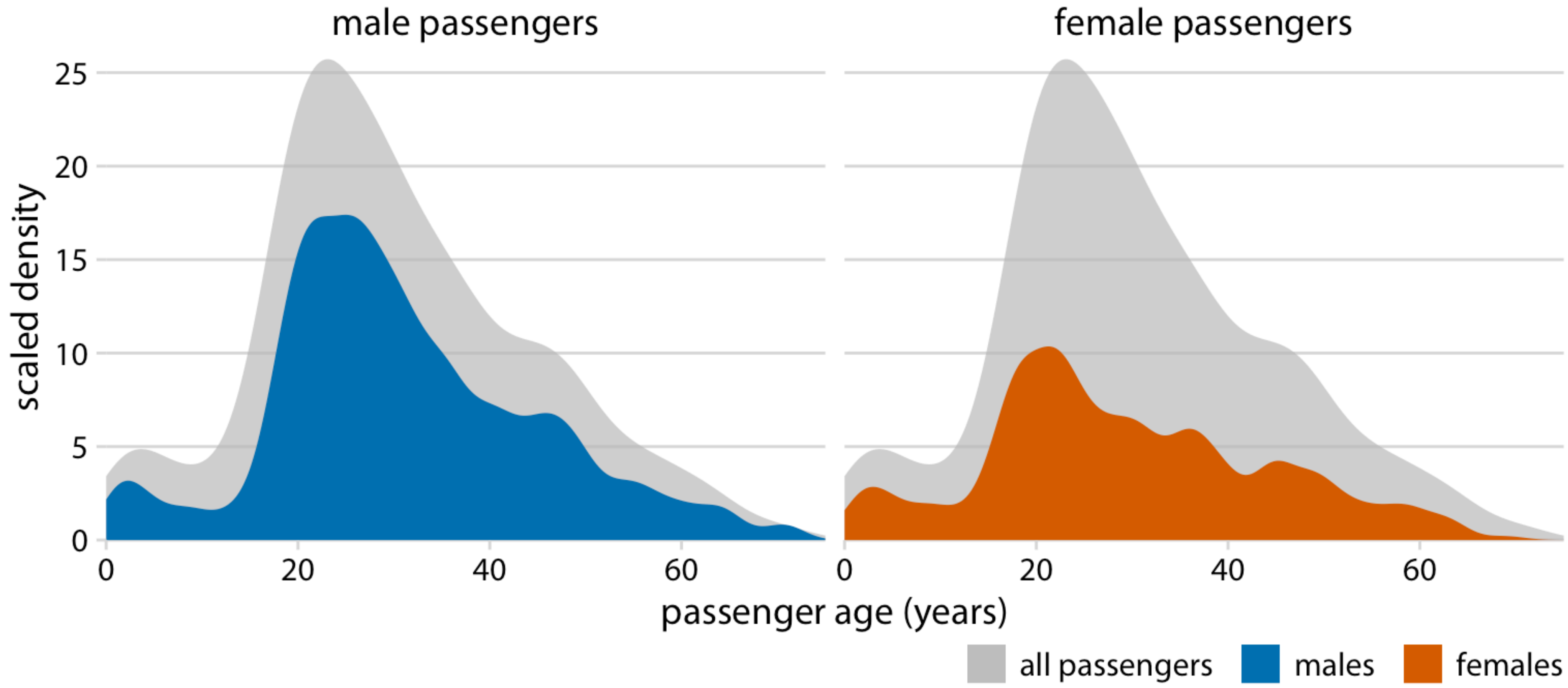
- Density estimates of ages by gender
- Age distributions shown as a proportion of the total
- Age pyramid visualization

Better Approaches (continued)



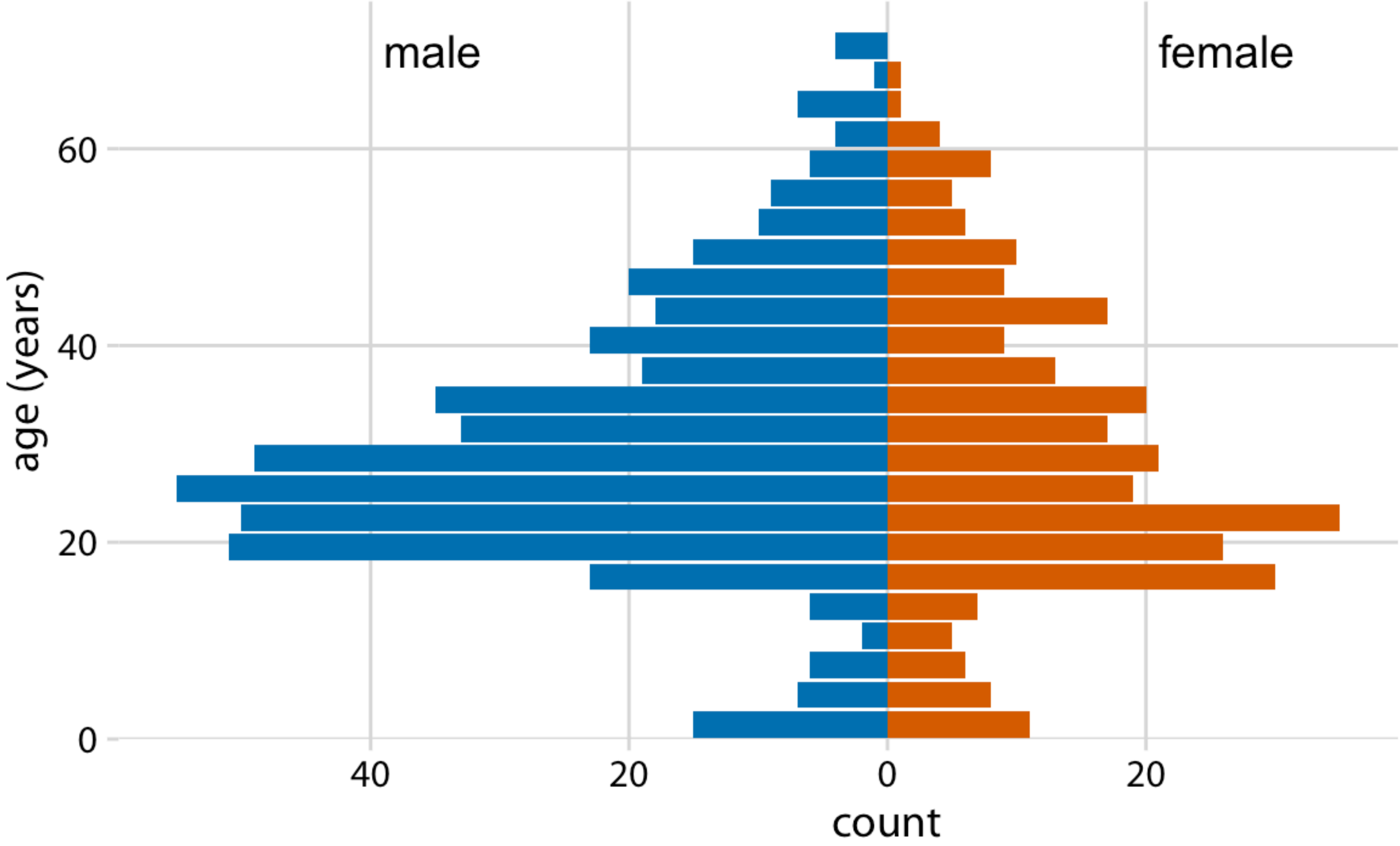
Density estimates of the ages of male and female Titanic passengers. To highlight that there were more male than female passengers, the density curves were scaled such that the area under each curve corresponds to the total number of male and female passengers with known age (468 and 288, respectively).

Better Approaches (continued)



The colored areas show the density estimates of the ages of male and female passengers, respectively, and the gray areas show the overall passenger age distribution.

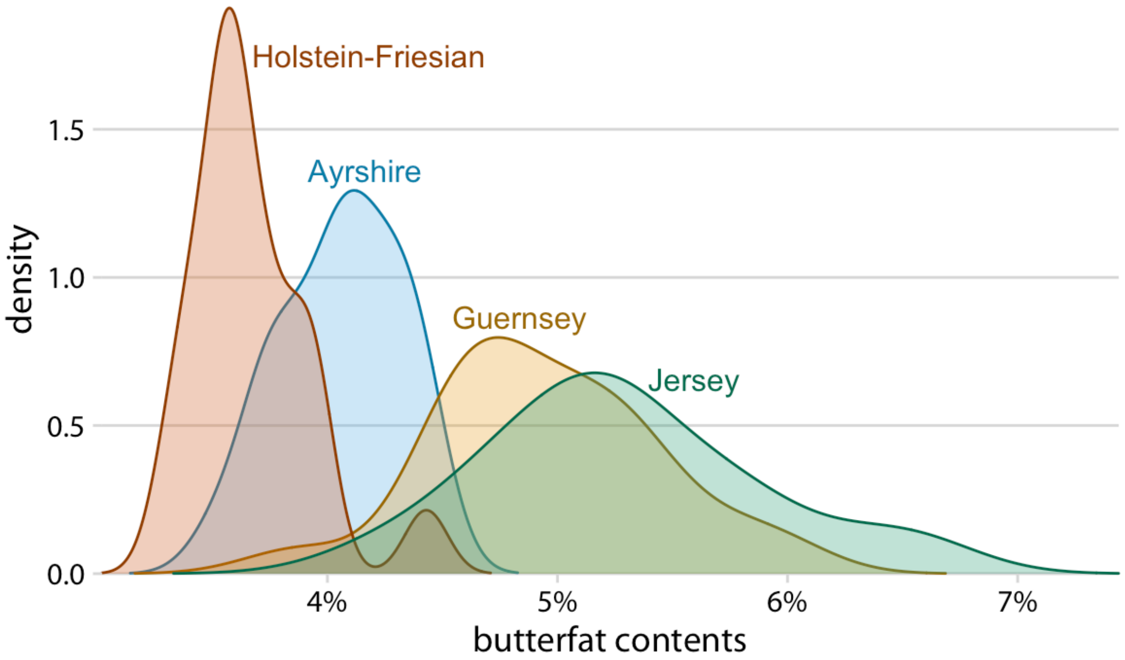
Better Approaches (continued)



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

Handling Multiple Distributions

- Challenges of visualizing multiple distributions
- Recommendation for using density plots
- Example: Density estimates of butterfat percentage in cattle breeds



Density estimates of the butterfat percentage in the milk of four cattle breeds. Data Source: Canadian Record of Performance for Purebred Dairy Cattle

Empirical cumulative distribution functions and q-q plots

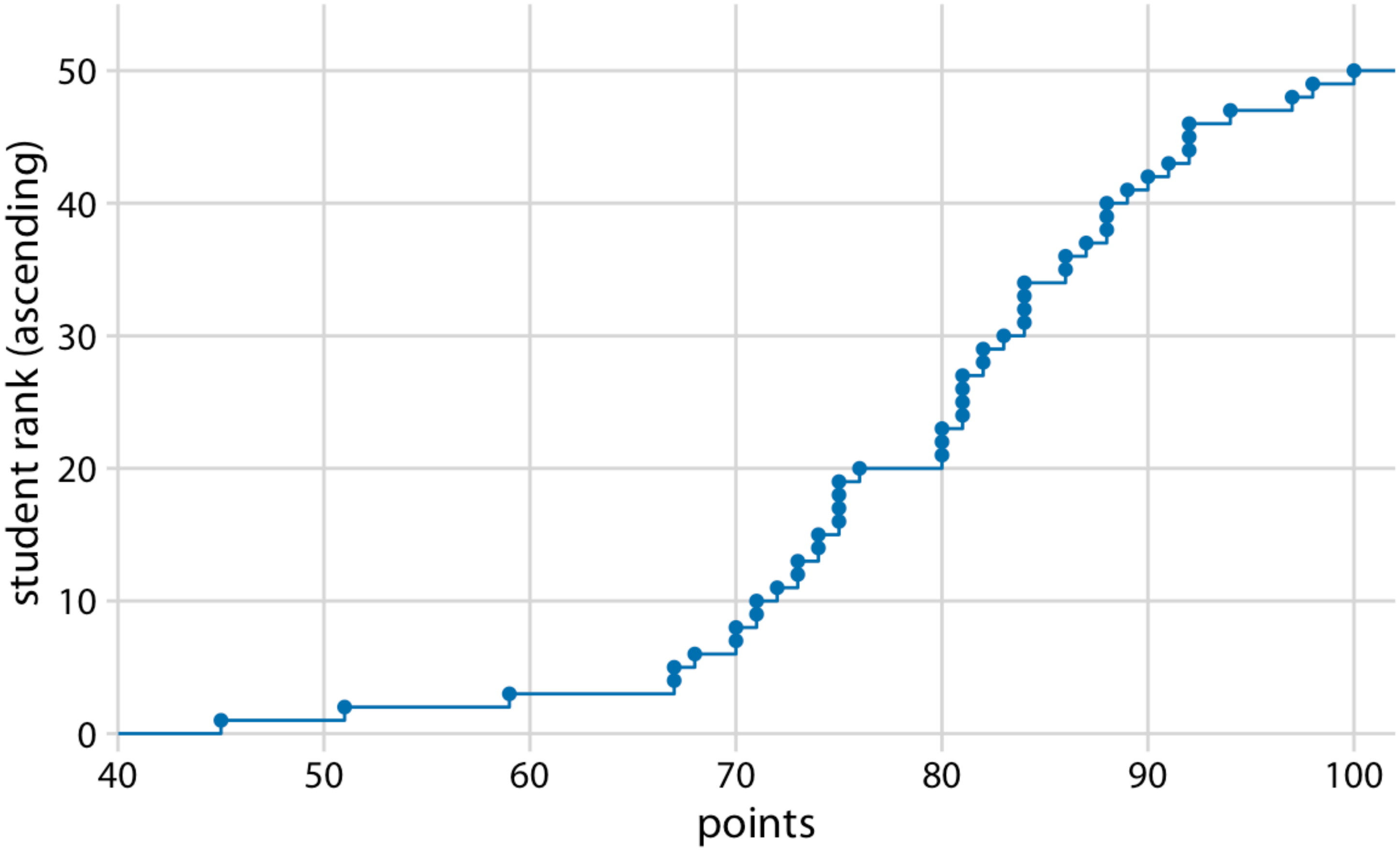
Part III

Limitations of Histograms and Density Plots

- Recap the limitations of histograms and density plots
- Need for techniques that require no arbitrary parameter choices
- Alternative visualization techniques to histograms and density plots: ECDFs and Q-Q Plots

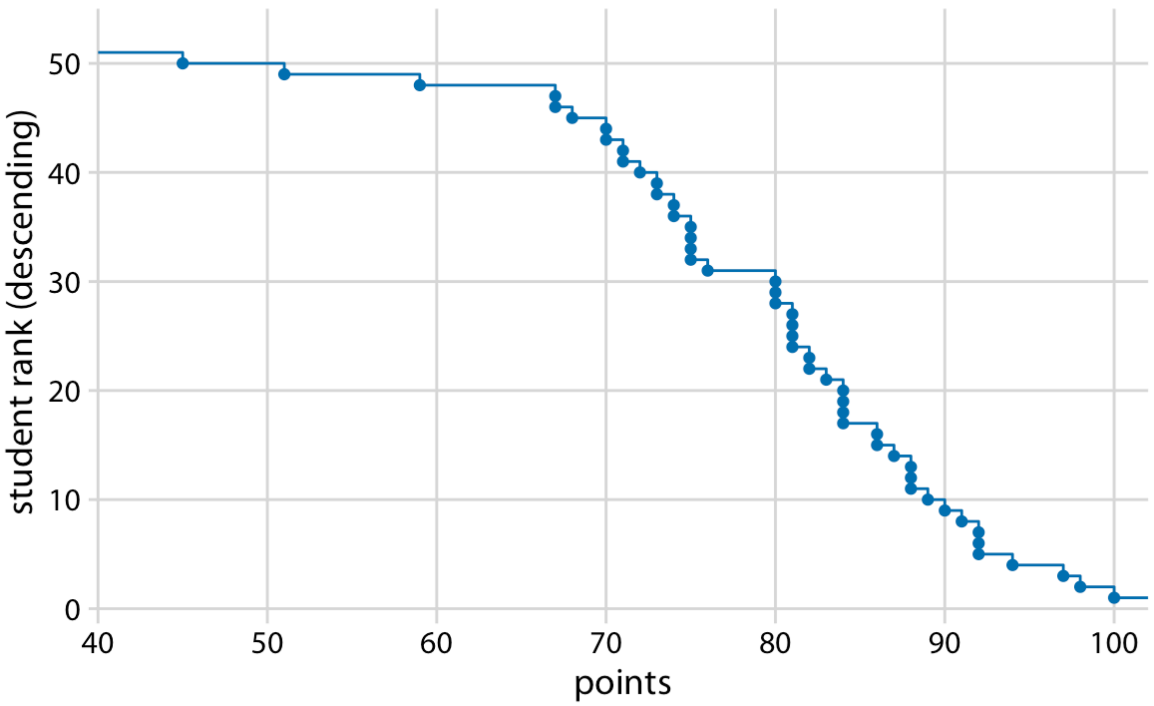
Empirical Cumulative Distribution Functions (ECDFs)

- Concept of using a hypothetical example of student grades

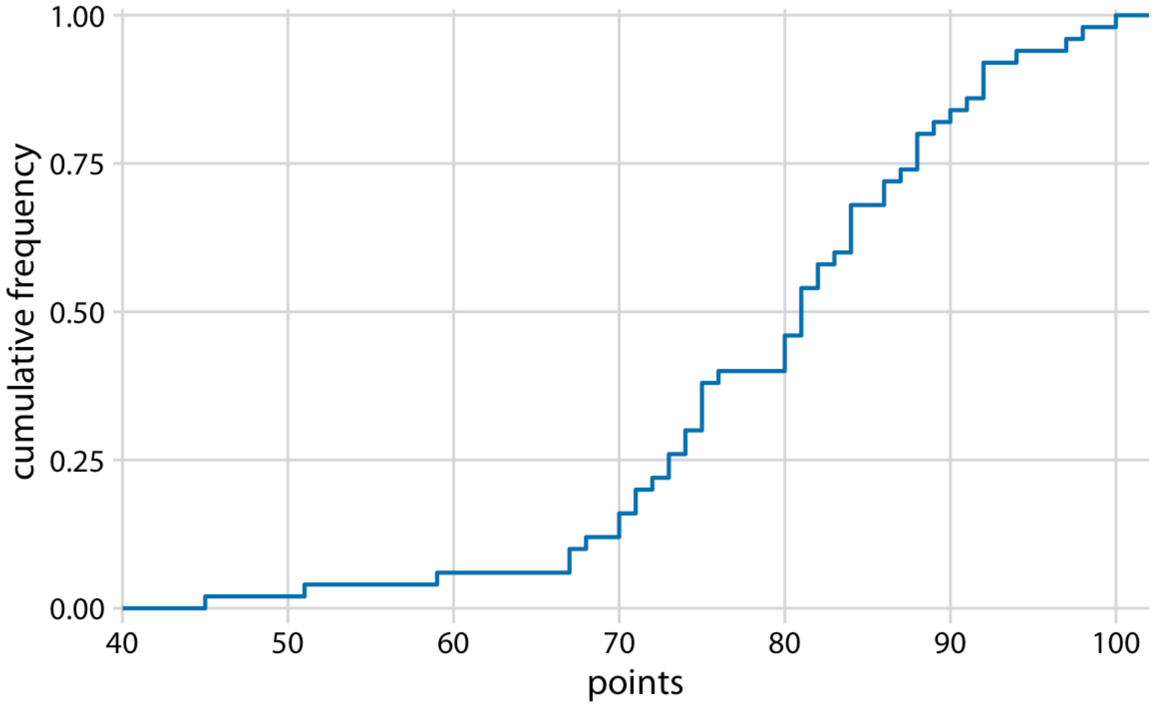


Empirical cumulative distribution function of student grades for a hypothetical class of 50 students

Normalized ECDFs

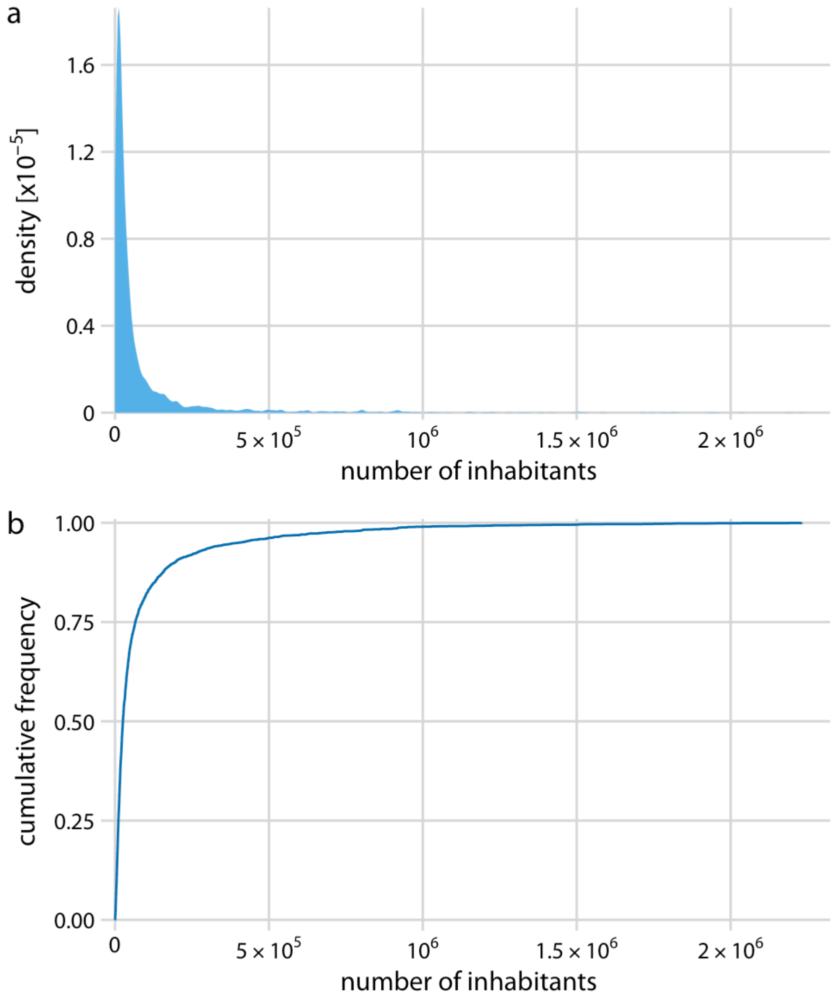


Distribution of student grades plotted as a descending ecdf



Ecdf of student grades. The student ranks have been normalized to the total number of students, such that the *y* values plotted correspond to the fraction of students in the class with at most that many points.

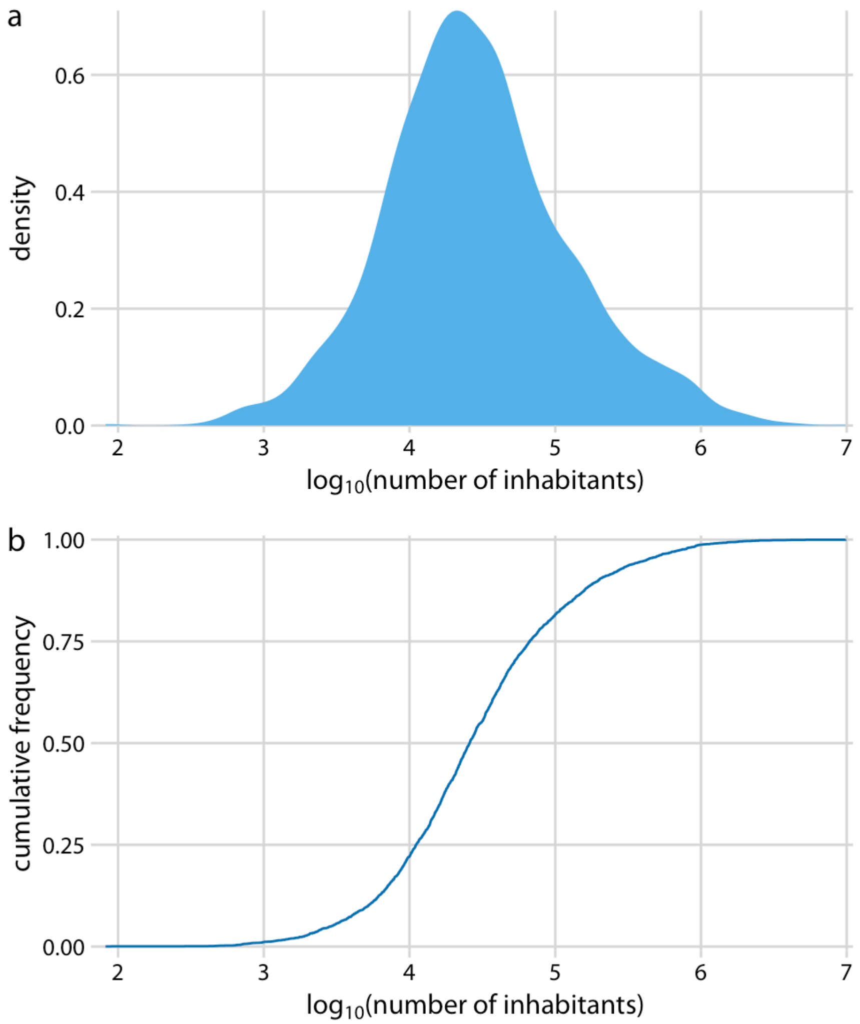
Handling Highly Skewed Distributions



- Challenge of visualizing highly skewed distributions
- Example: population counts in US counties

Distribution of the number of inhabitants in US counties, according to the 2010 US Census. (a) Density plot. (b) Empirical cumulative distribution function

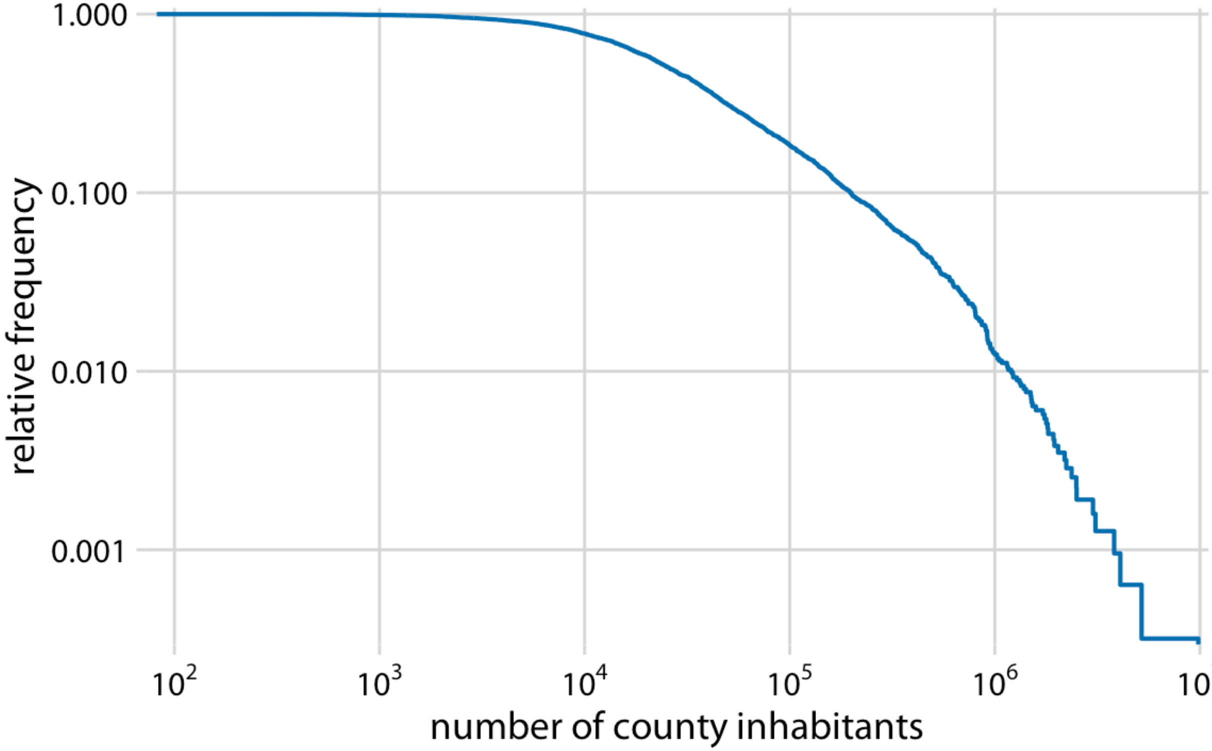
Log-Transformation



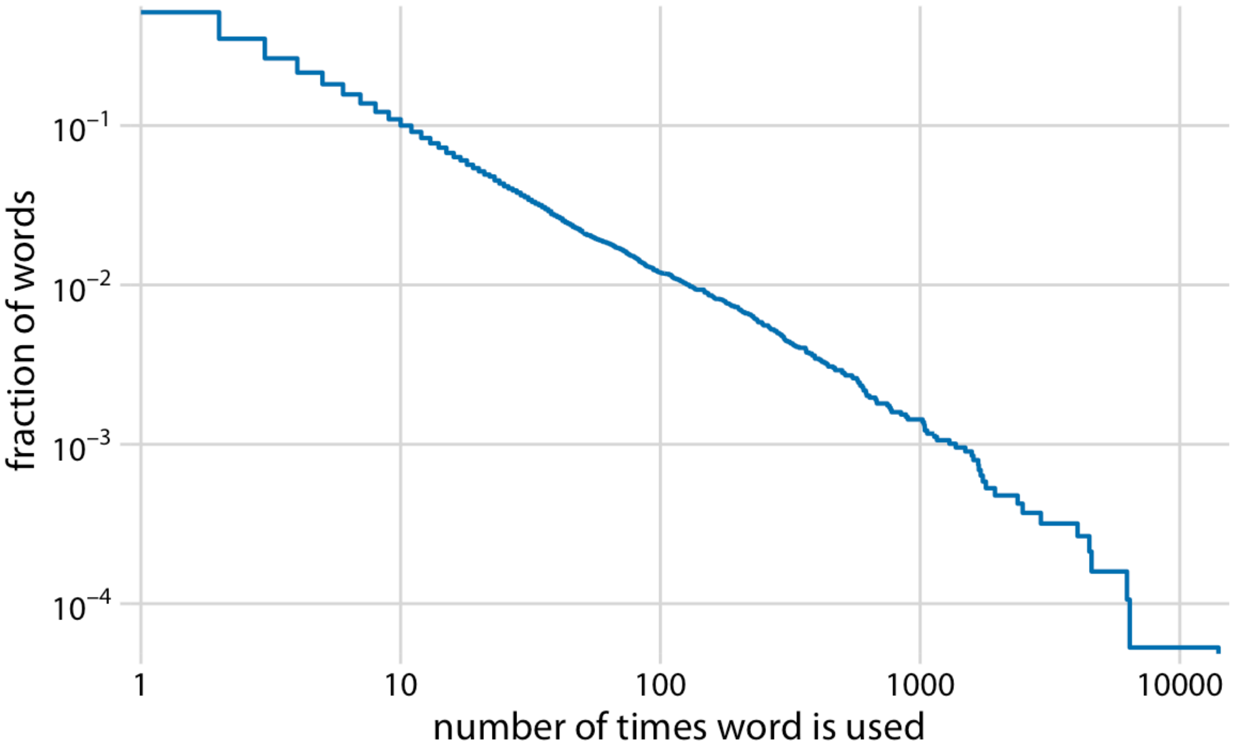
- Log-transformation for skewed distributions.
- Log-transformed data enables better visualization

Distribution of the logarithm of the number of inhabitants in US counties. (a) Density plot. (b) Empirical cumulative distribution function

Power Law Distributions



Relative frequency of counties with at least that many inhabitants versus the number of county inhabitants

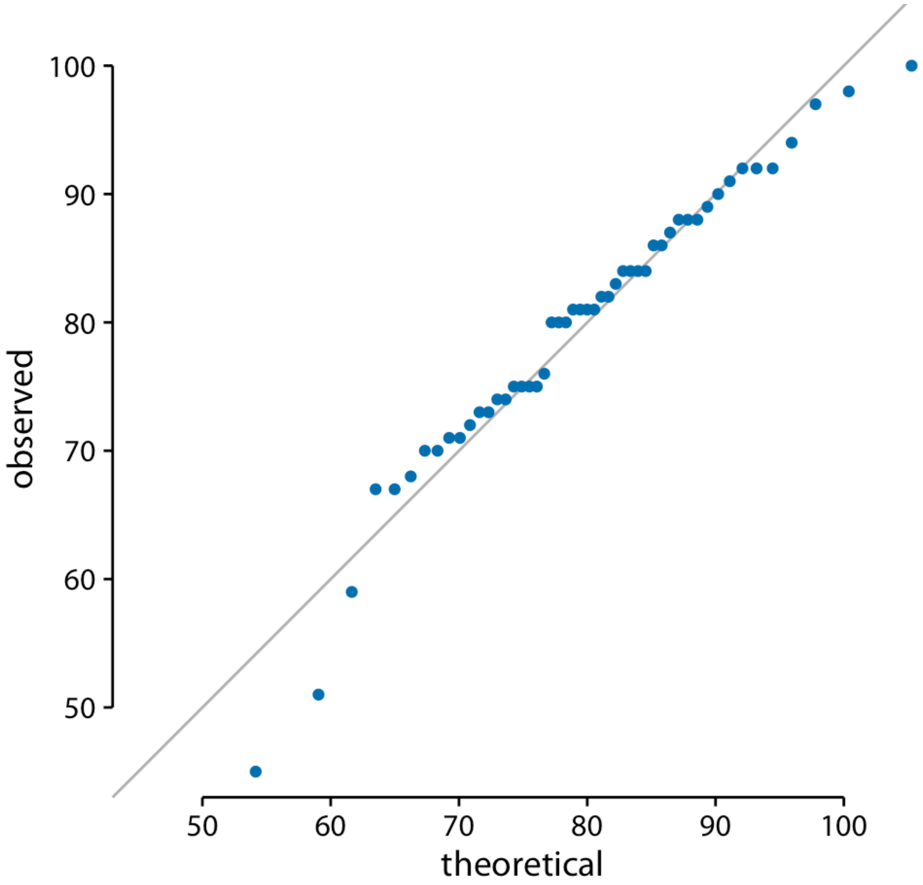


Relative frequency of words that occur at least that many times in the novel versus the number of times words used

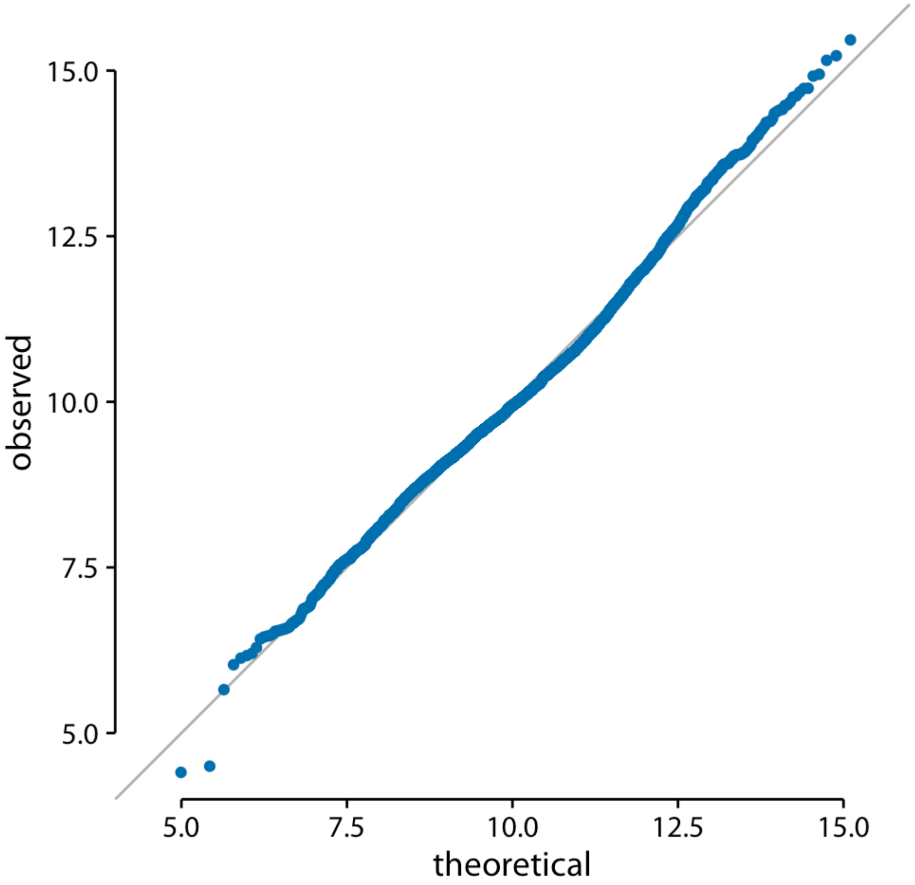
Quantile-Quantile (Q-Q) Plots

- Q-Q plots as a visualization technique
- Comparing observed data to theoretical distributions
- Example: Q-Q plot for student grades distribution

Q-Q Plots (continued)



q-q plot of student grades



q-q plot of the logarithm of the number of inhabitants in US counties