Fundamentals of Data Visualization
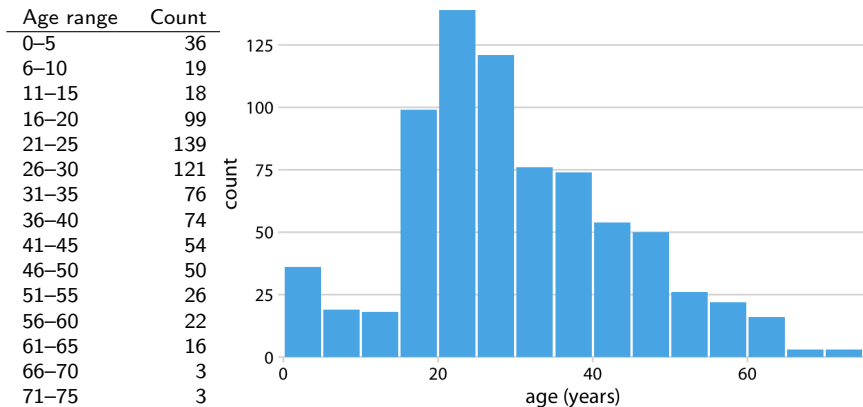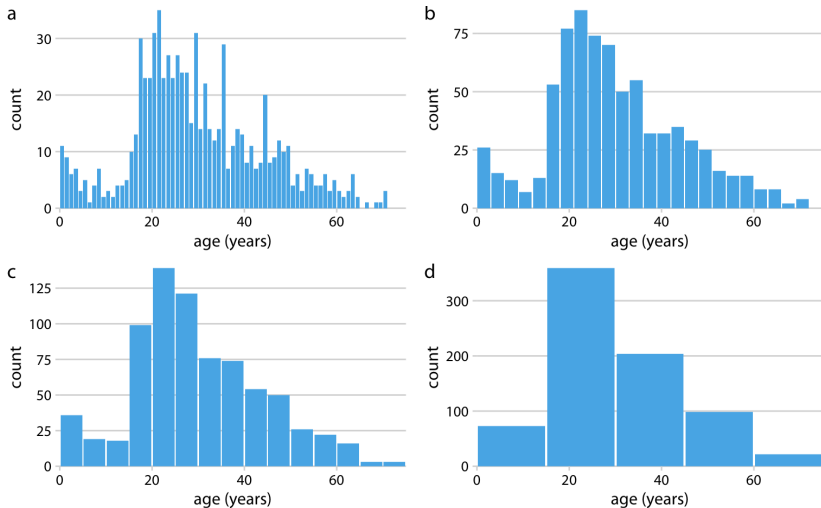
Chapter 7

April 27, 2023

Visualizing Distributions

- We often want to know how a variable is distributed.
- For example, we might want to know how many passengers of what ages there were on the Titanic, i.e., how many children, young adults, middle-aged people, seniors, and so on.
- This is called the age **distribution**.
- A standard visualization is either a histogram or a density plot.

A single distribution

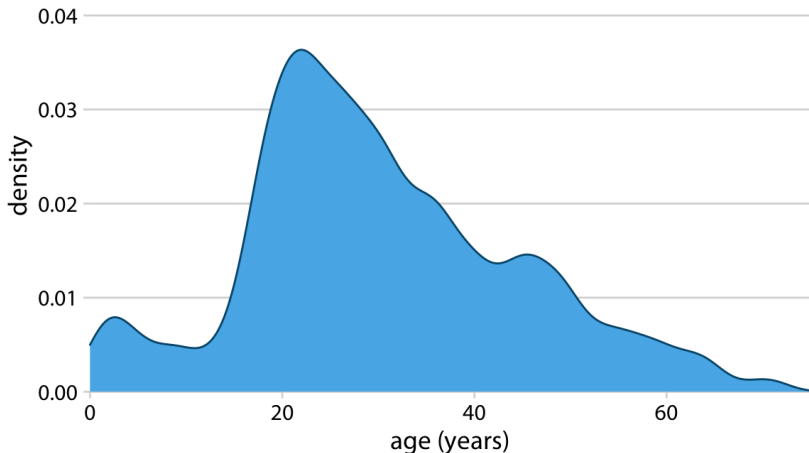| Age range	Count
0–5	36
6–10	19
11–15	18
16–20	99
21–25	139
26–30	121
31–35	76
36–40	74
41–45	54
46–50	50
51–55	26
56–60	22
61–65	16
66–70	3
71–75	3

Bin width is crucial

- When making a histogram, always explore multiple bin widths.

Density plots

- Curve is estimated from the data using **kernel density estimation**.
- We add up all values in a small width (the *bandwidth*).
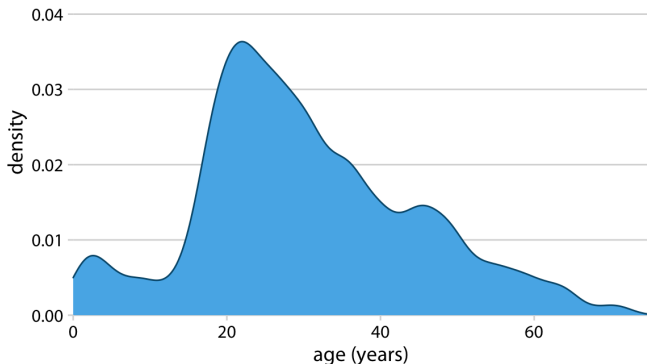- Usually weighted by a Gaussian kernel.

Density curves depend upon kernel and bandwidth
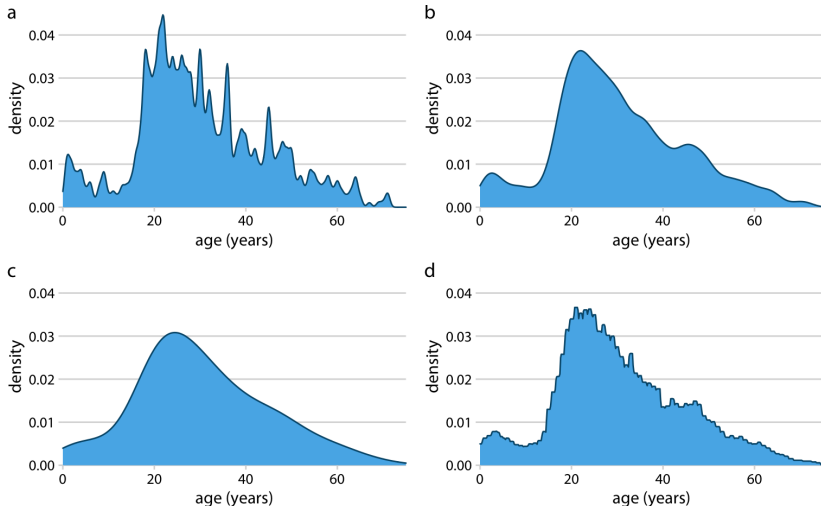
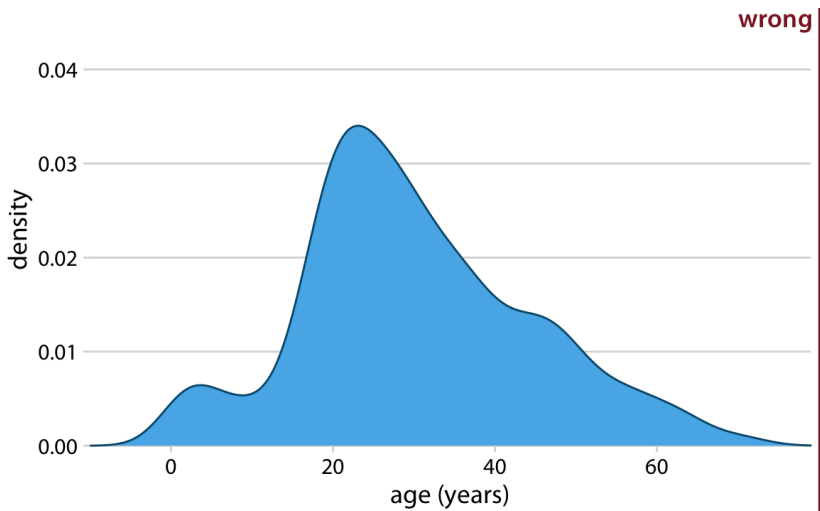(a) Gaussian kernel, bandwidth = 0.5; (b) Gaussian kernel, bandwidth = 2;
(c) Gaussian kernel, bandwidth = 5; (d) Rectangular kernel, bandwidth = 2

Density curves

- Usually scaled so the area sums to one.
- For the Titanic data, ages range from 0 to 75.
- The average height should be about $1/75 \approx 0.013$

Density curves produce data where there is none

Histograms vs. Density plots

- Some people are vehemently against density plots and believe that they are arbitrary and misleading.
- Others realize that histograms can be just as arbitrary and misleading.
- Density estimates have an inherent advantage over histograms as soon as we want to visualize more than one distribution at a time.
- Cumulative density and q-q plots plot distributions without the arbitrary choices, but are difficult to interpret.

# Stacked bar charts for multiple distributions
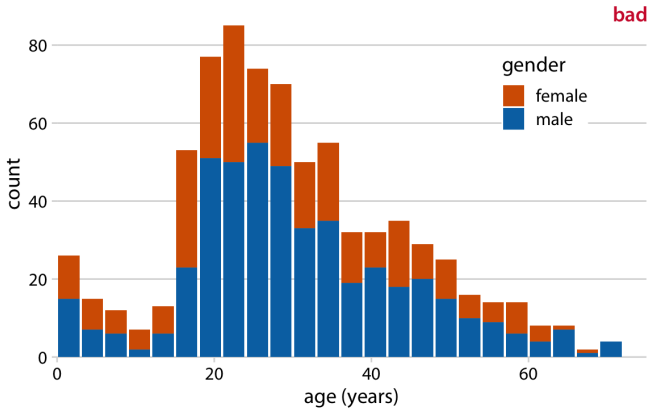
- Very difficult to interpret. Where is the zero for women?
- Were men and women passengers generally of the same age, or was there an age difference between the genders?
- The bars for women cannot be compared; they do not have a common zero.

All bars start at zero, but are transparent
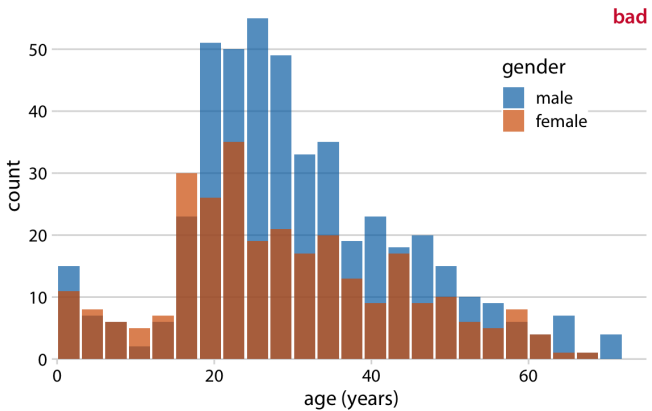
- Now it appears that there are actually three different groups, not just two.
- We're still not entirely sure where each bar starts and ends.
- A semi-transparent bar drawn on top of another tends to not look like a semi-transparent bar but instead like a bar drawn in a different color.

Transparent density plots
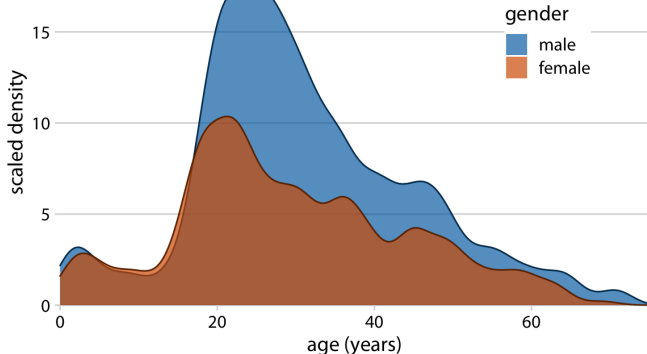
- Density plots generally do better with transparency, because the curves usually don't line up.
- This dataset is bad for transparency; the curves are similar up to about age 17.
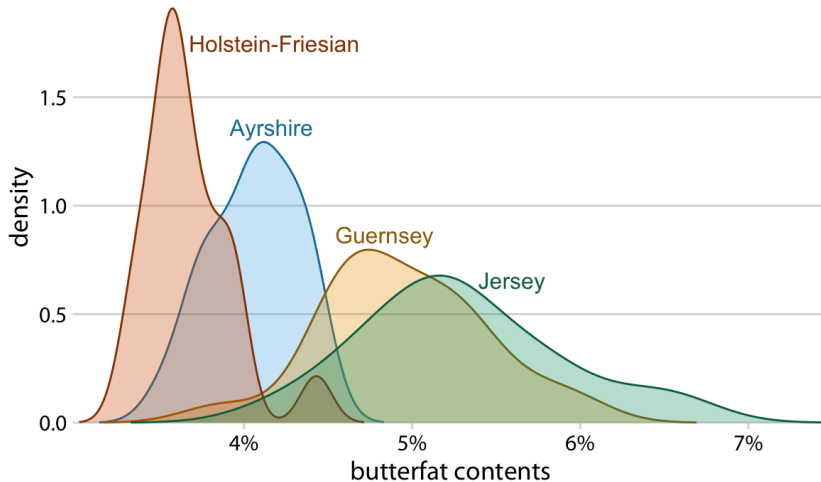
Separate density plots
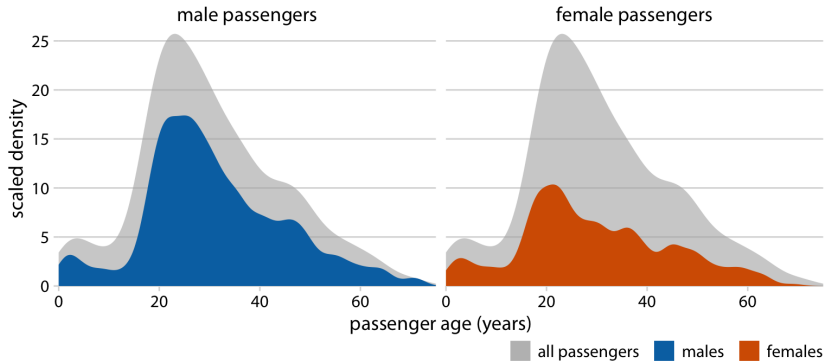
- Show each subpopulation in relation to the total.
- Easy to see there were many fewer women at around age 20.

Age pyramid popular with populations
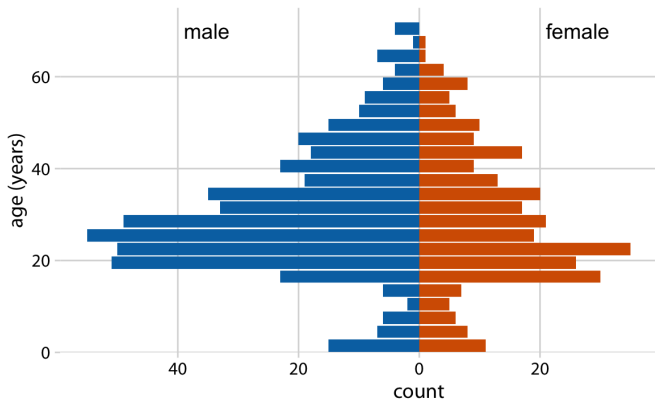
- Only works for two distributions

Density plots for multiple distributions work well

- Distributions must be somewhat distinct.
- Distributions should be contiguous.