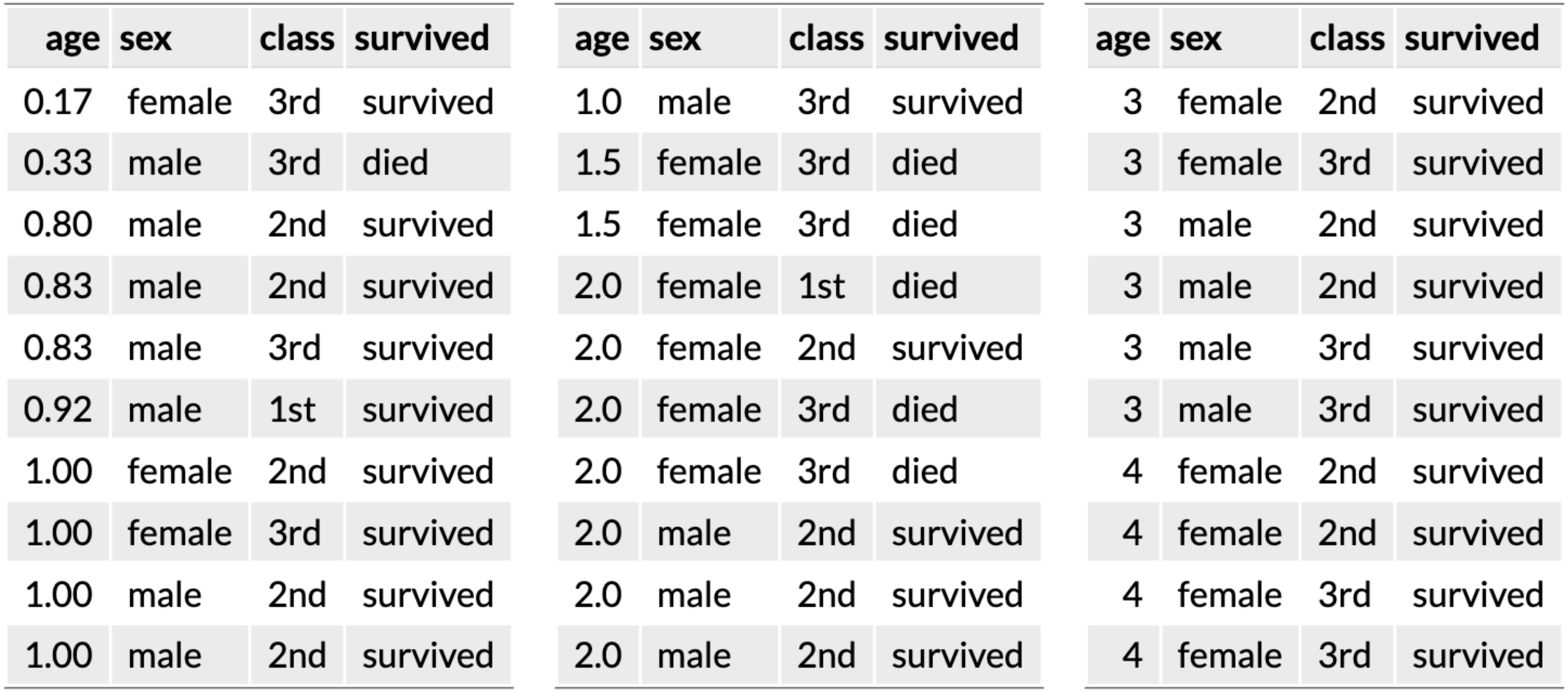
CSCI 491: Data Visualization

8- Visualizing Distributions

MONTANA STATE UNIVERSITY

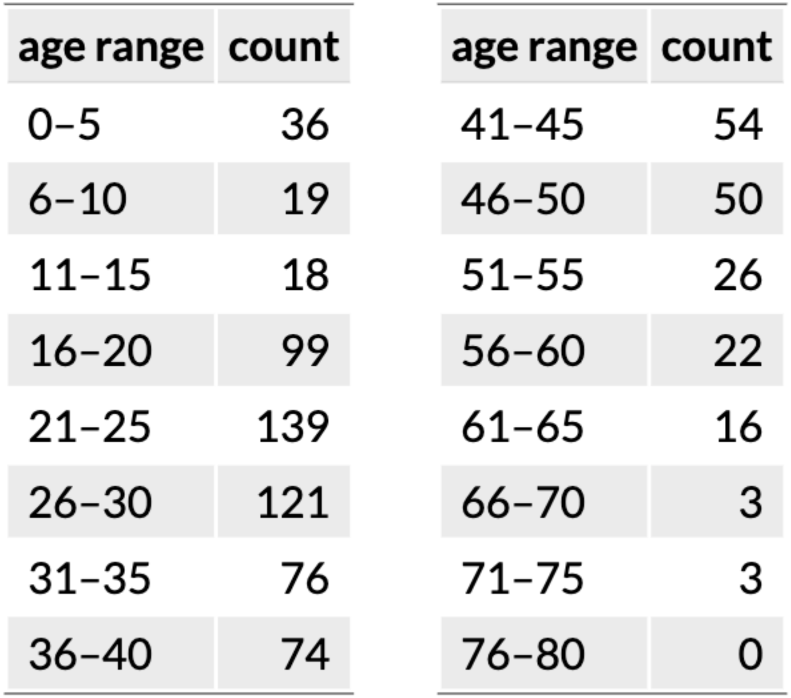
Who was on Titanic?



age	sex	class	survived
0.17	female	3rd	survived
0.33	male	3rd	died
0.80	male	2nd	survived
0.83	male	2nd	survived
0.83	male	3rd	survived
0.92	male	1st	survived
1.00	female	2nd	survived
1.00	female	3rd	survived
1.00	male	2nd	survived
1.00	male	2nd	survived
1.0	male	3rd	survived
1.5	female	3rd	died
1.5	female	3rd	died
2.0	female	1st	died
2.0	female	2nd	survived
2.0	female	3rd	died
2.0	female	3rd	died
2.0	male	2nd	survived
2.0	male	2nd	survived
2.0	male	2nd	survived
3	female	2nd	survived
3	female	3rd	survived
3	male	2nd	survived
3	male	2nd	survived
3	male	3rd	survived
3	male	3rd	survived
4	female	2nd	survived
4	female	2nd	survived
4	female	3rd	survived
4	female	3rd	survived

Who was on Titanic?

- We can obtain a sense of the age distribution among the passengers by grouping all passengers into bins with comparable ages and then counting the number of passengers in each bin Numbers of passengers with known age on the Titanic.

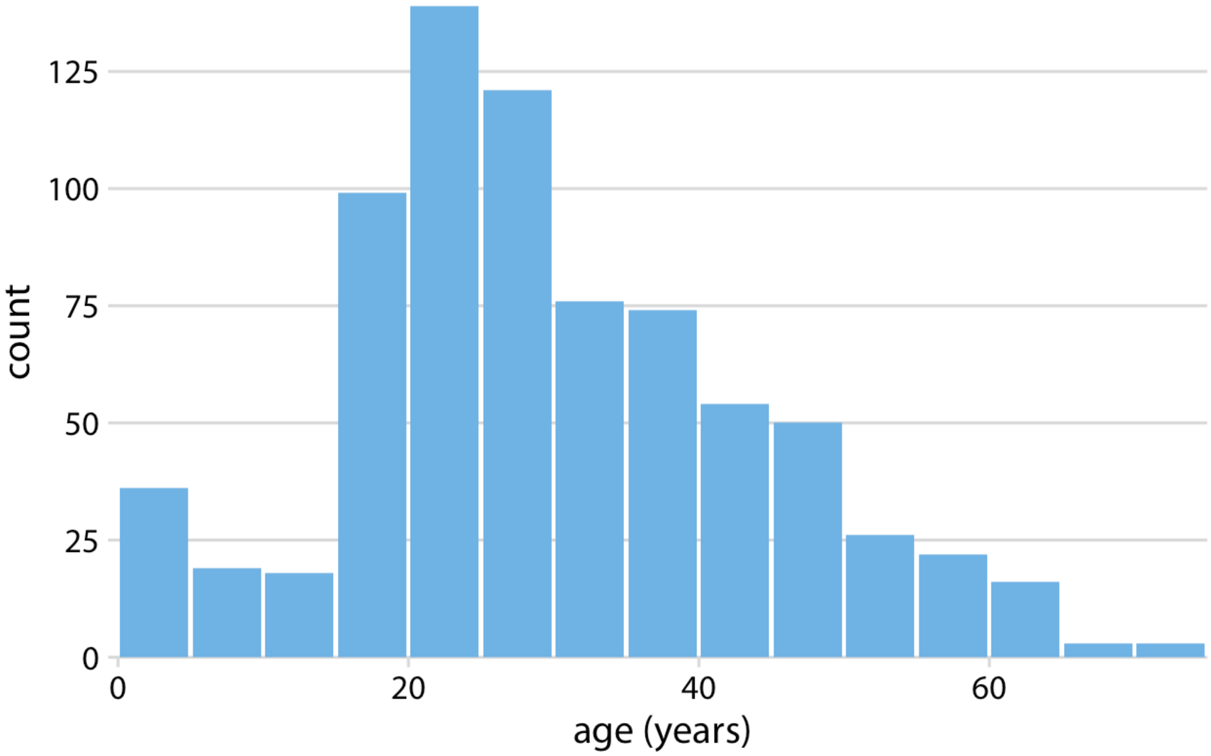


age range	count
0–5	36
6–10	19
11–15	18
16–20	99
21–25	139
26–30	121
31–35	76
36–40	74

age range	count
41–45	54
46–50	50
51–55	26
56–60	22
61–65	16
66–70	3
71–75	3
76–80	0

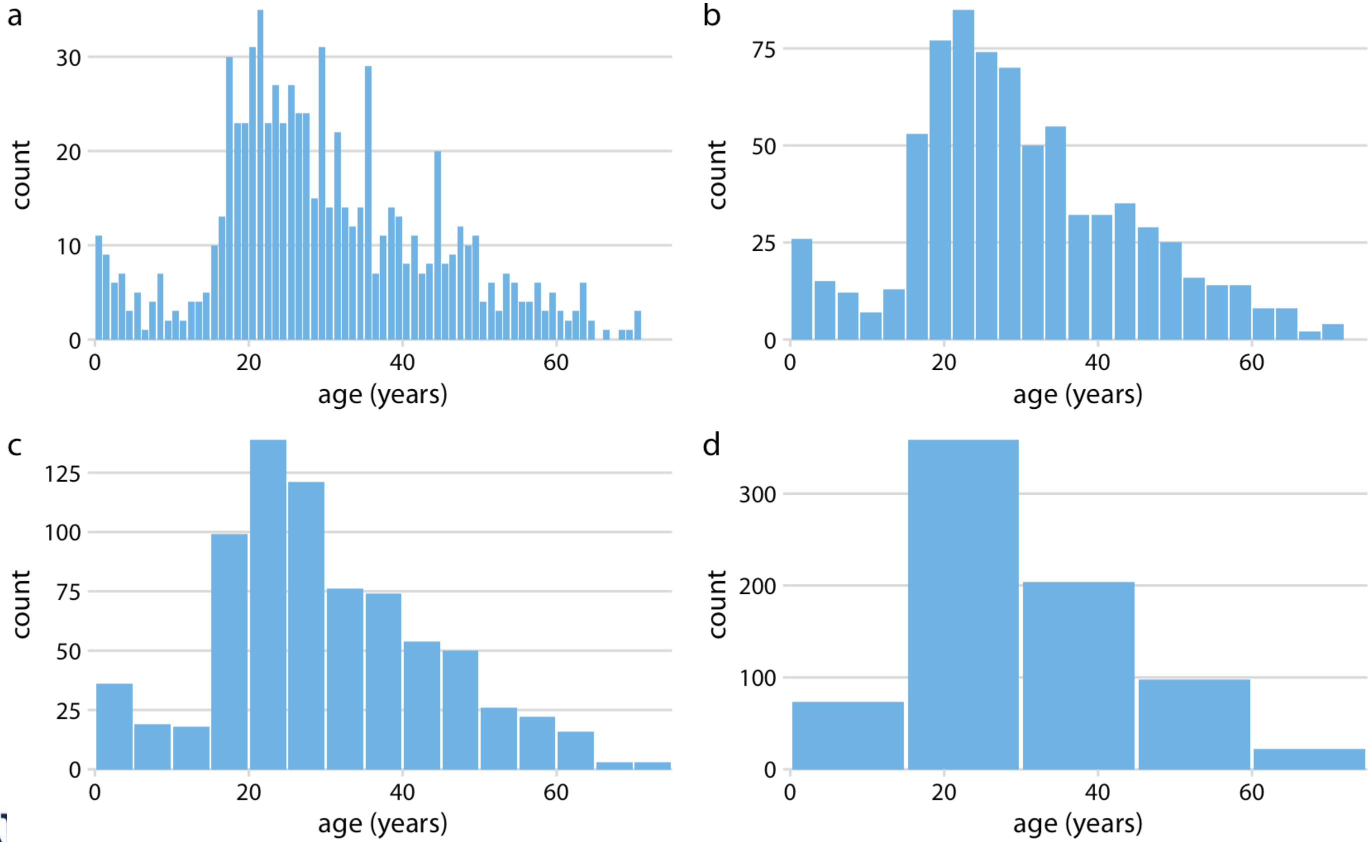
Histogram

- Note that all bins must have the same width for visualization to be a valid histogram



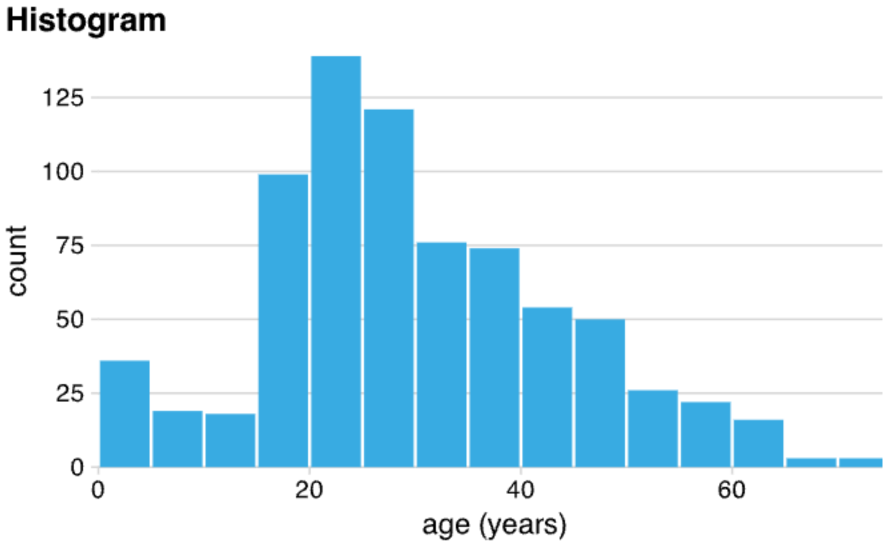
Bin Width Matters!

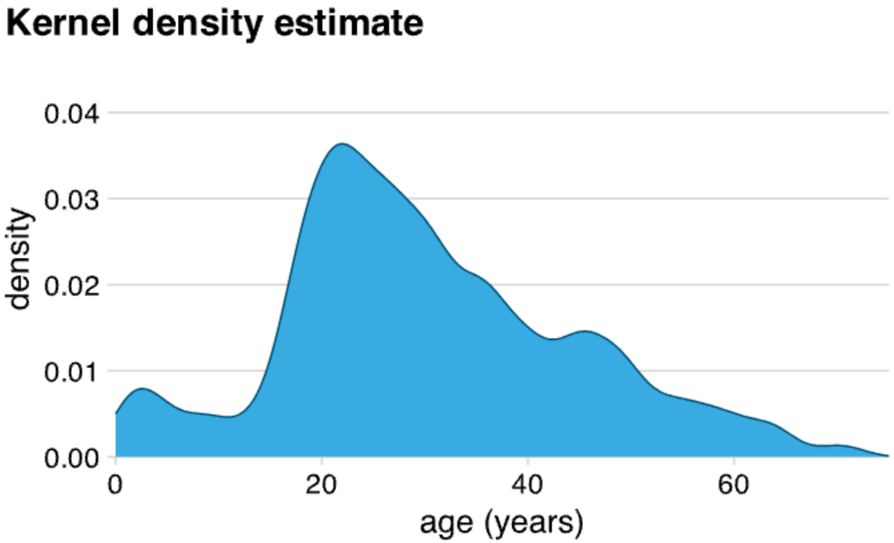
- Because histograms are generated by binning the data, their exact visual appearance depends on the choice of the bin width.



Density Plots

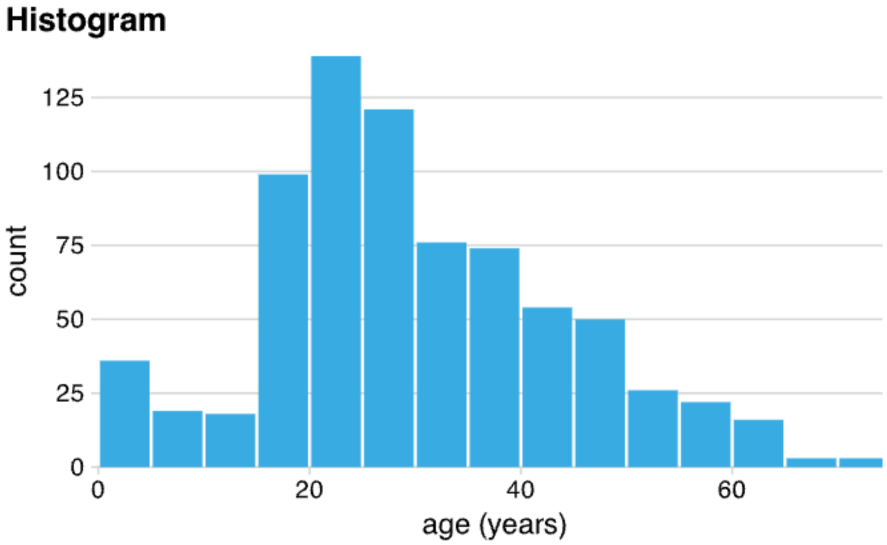
- These days we see we see them increasingly being replaced by density plots.
- In density plots we attempt to visualize the underlying probability distribution of the data by drawing an appropriate continuous curve.
- Kernel density estimation is the most commonly used method for this estimation procedure.

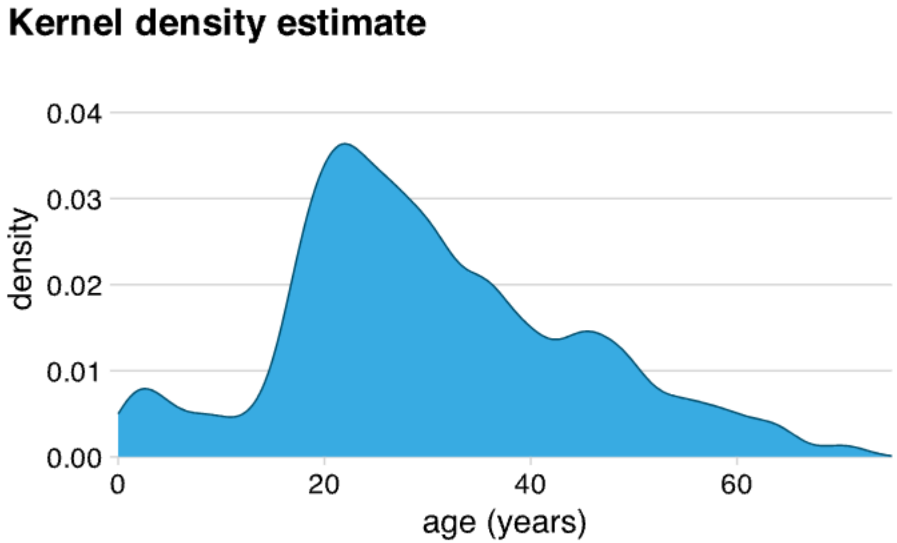




Density Plots

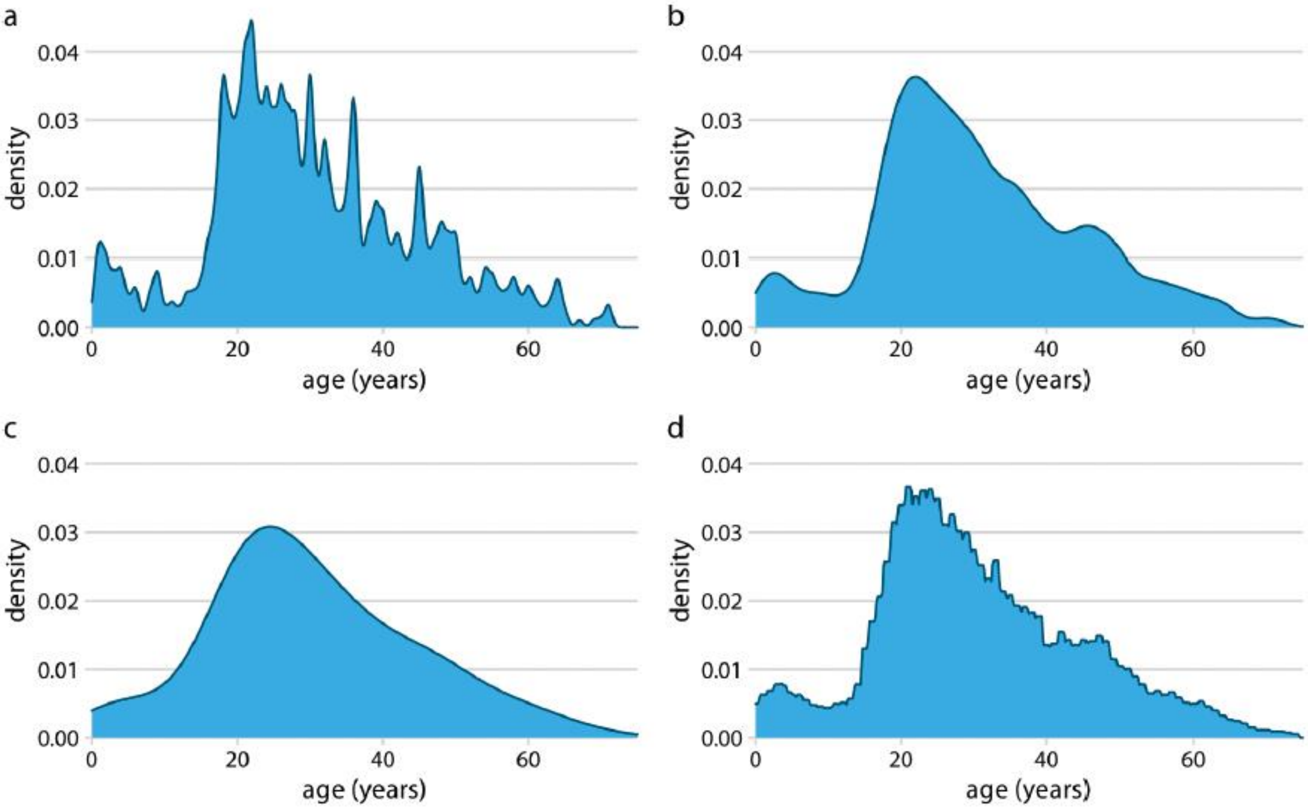
- In kernel density estimation, we draw a continuous curve (the kernel) with a small width (controlled by a parameter called bandwidth) at the location of each data point, and then we add up all these curves to obtain the final density estimate.
- The most widely used kernel is a Gaussian kernel, but there are many other choices.





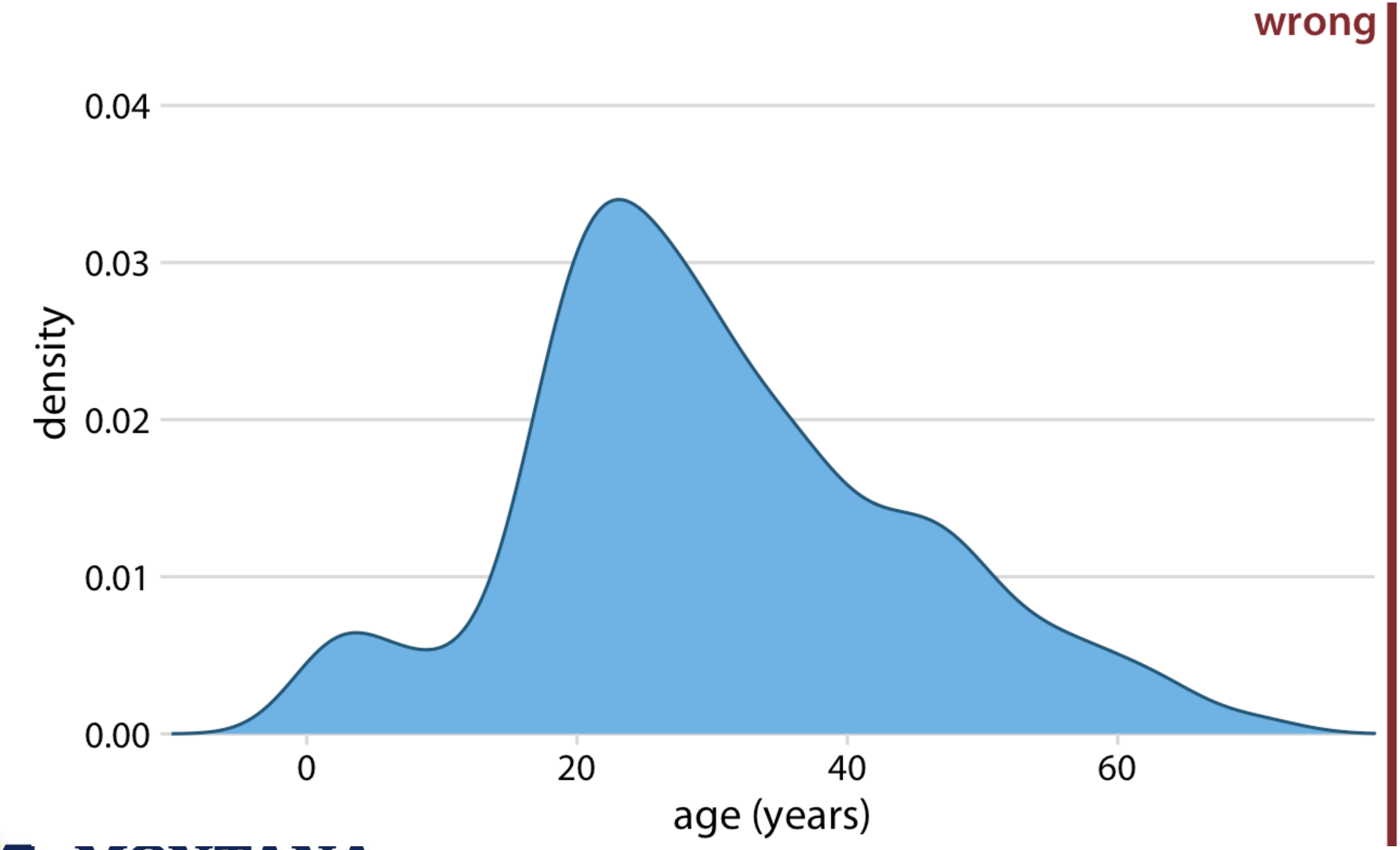
Kernel Choice and Bandwidth Matter

- The bandwidth parameter behaves similarly to the bin width in histograms.



(a) Gaussian kernel, bandwidth = 0.5; (b) Gaussian kernel, bandwidth = 2; (c) Gaussian kernel, bandwidth = 5; (d) rectangular kernel, bandwidth = 2. Data source: Encyclopedia Titanica.

Density Plot Pitfall



Density Plot Pitfall

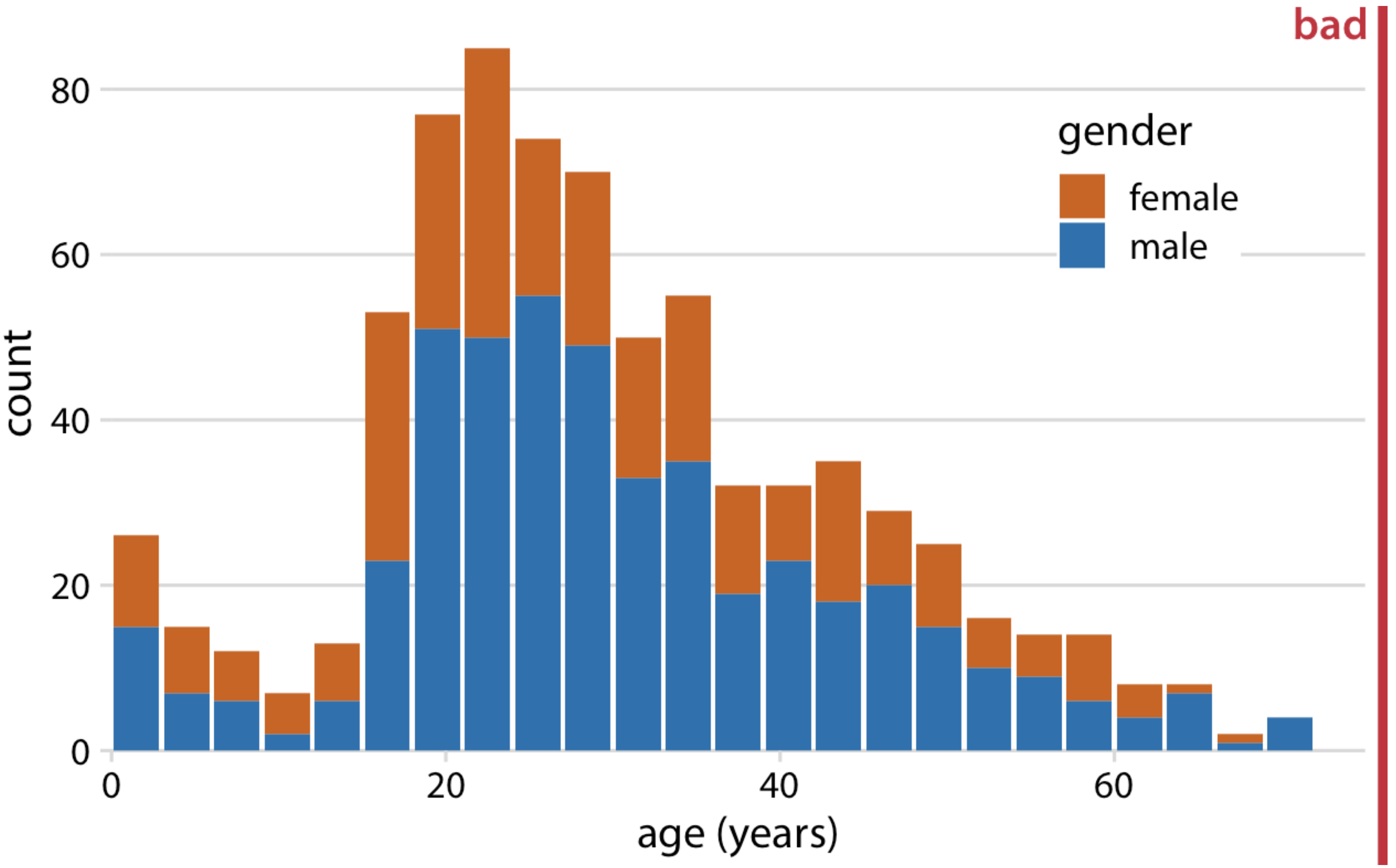
- DPs have a tendency to produce the appearance of data where none exists, in particular in the tails. As a consequence, careless use of density estimates can easily lead to figures that make nonsensical statements.
- Histogram or Density Plots?

Visualizing Multiple Distributions at the Same Time

- In many scenarios we have multiple distributions we would like to visualize simultaneously.
- How the ages of Titanic passengers are distributed between men and women.
- One commonly employed visualization strategy in this case is a stacked histogram

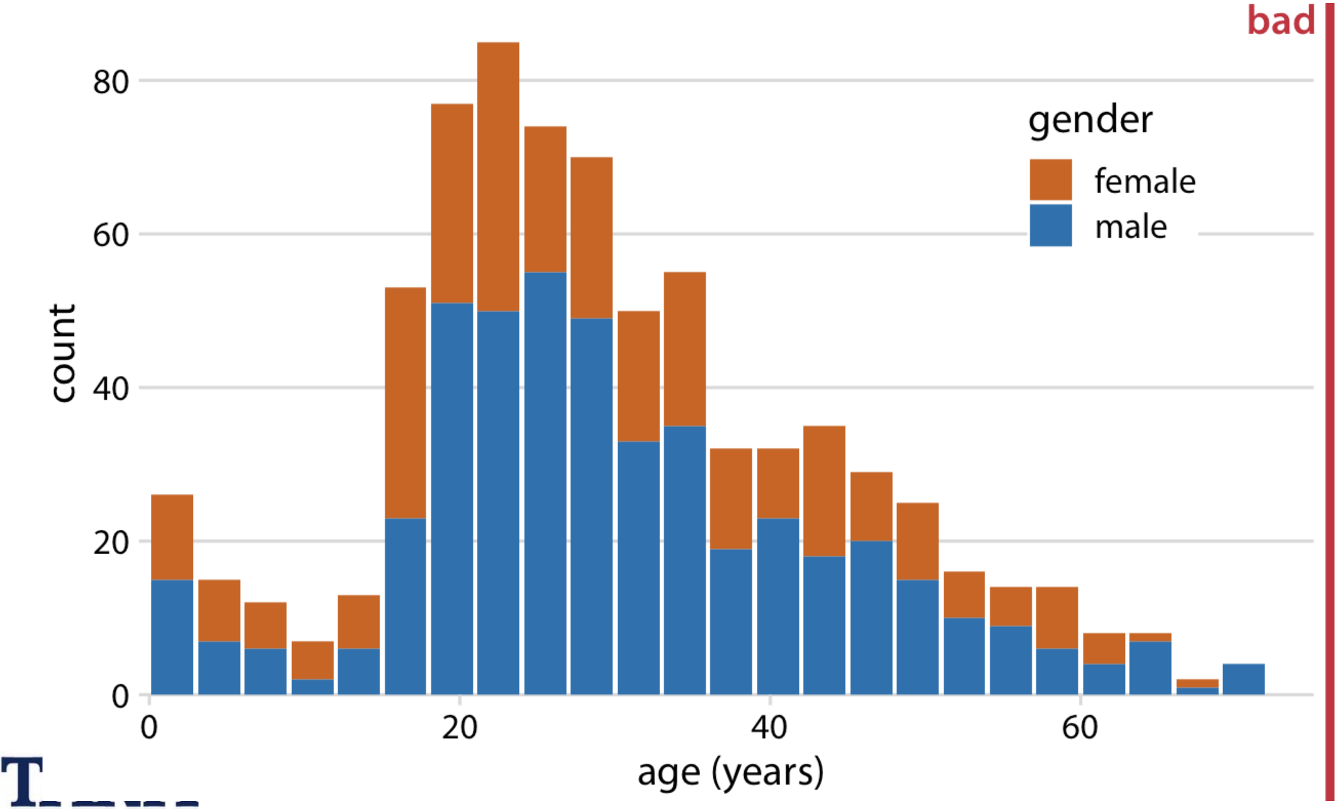
Stacked Histograms

- What is the problem with this plot?



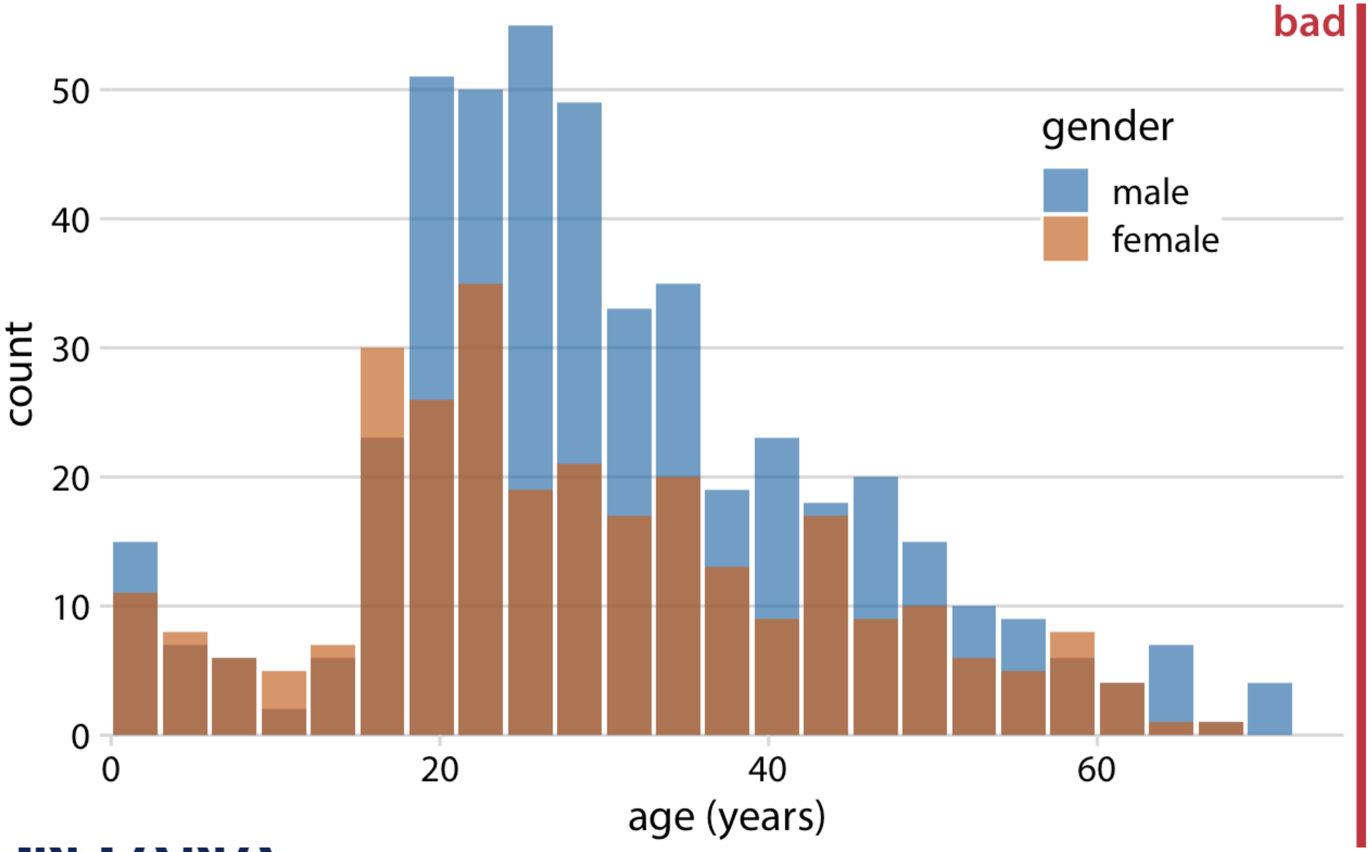
Stacked Histograms

- Stacked histograms are easily confused with overlapping
- histograms. The heights of the bars cannot easily be compared to each other.



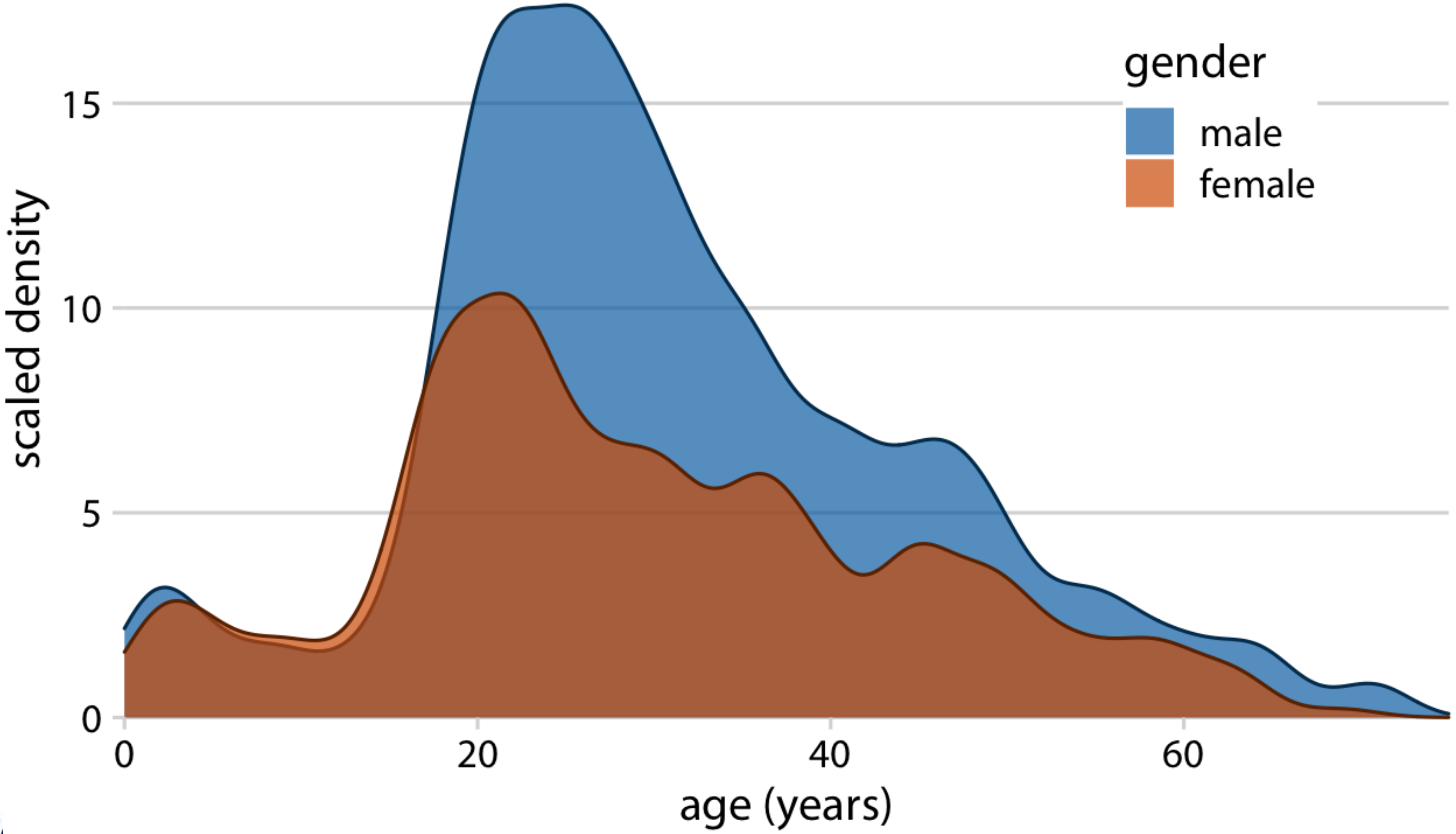
Overlapped Histograms

- How about this one?



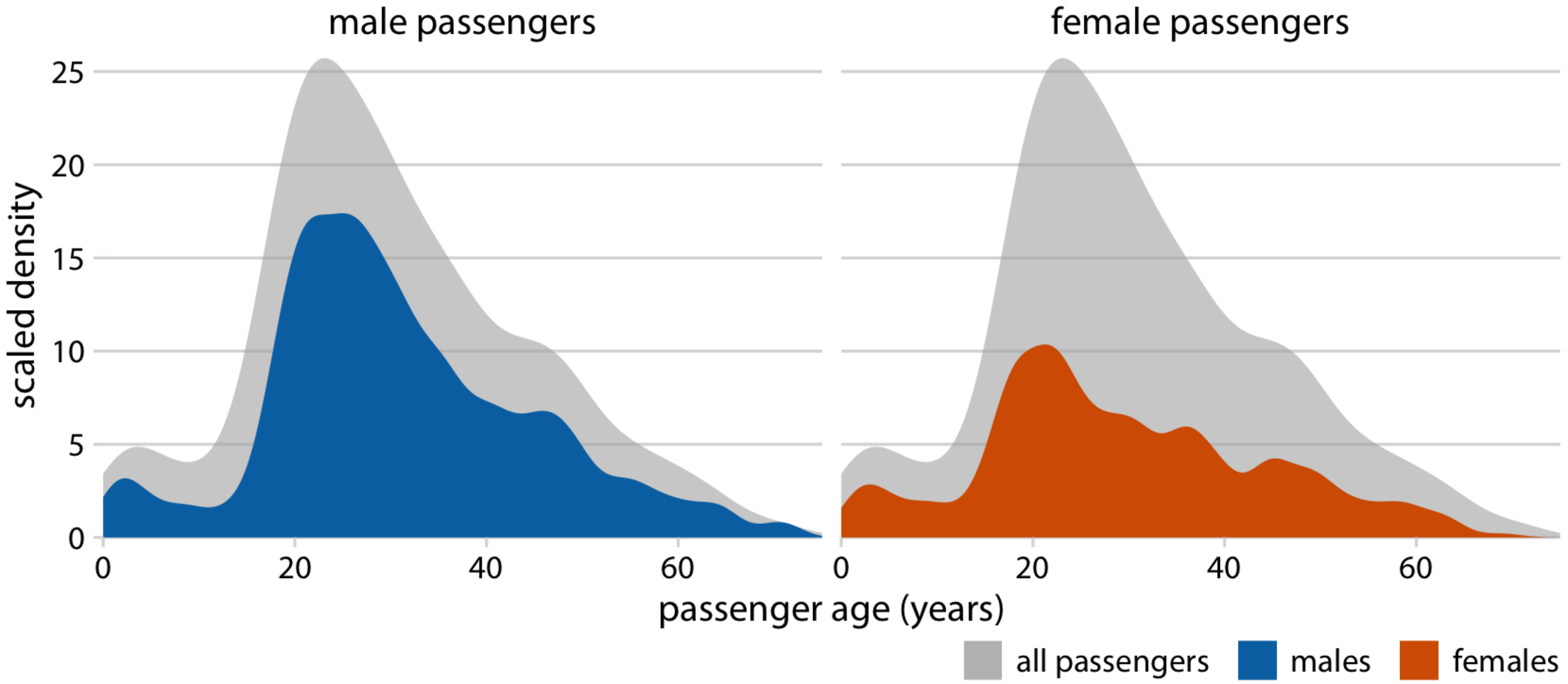
Overlapped Density Plots

- How about this one?

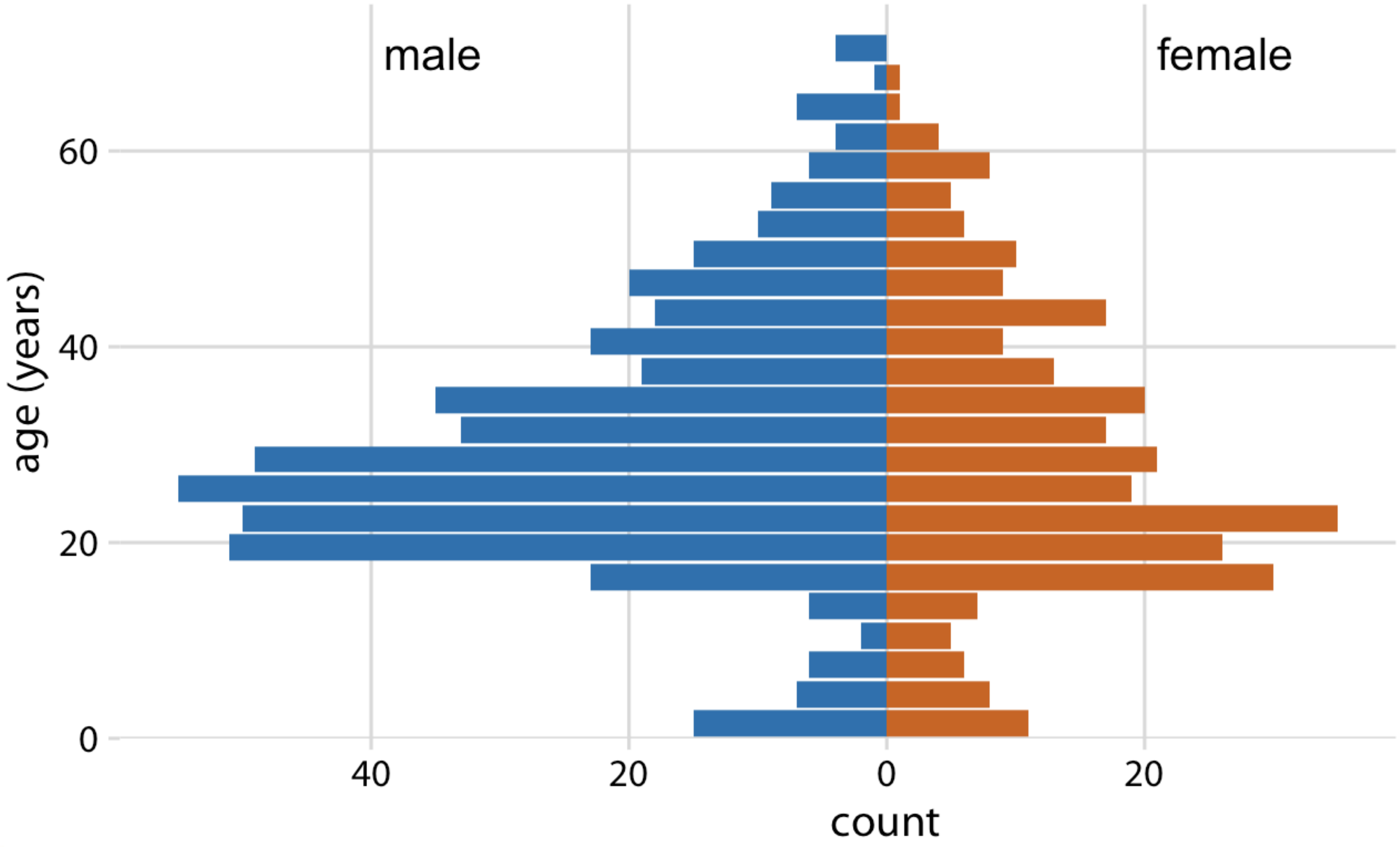


Overlapped Density Plots

- Is it better now?

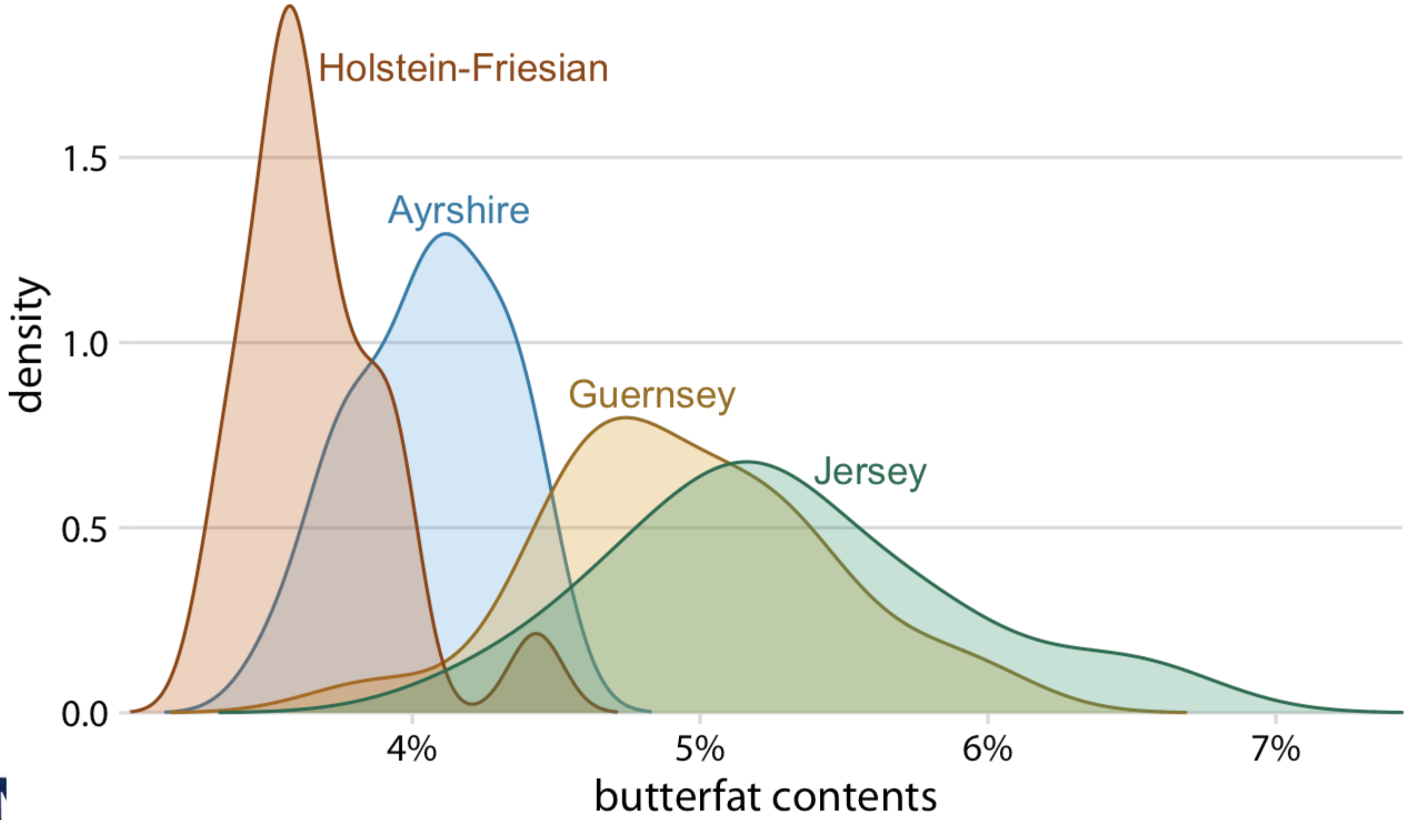


Age Pyramid

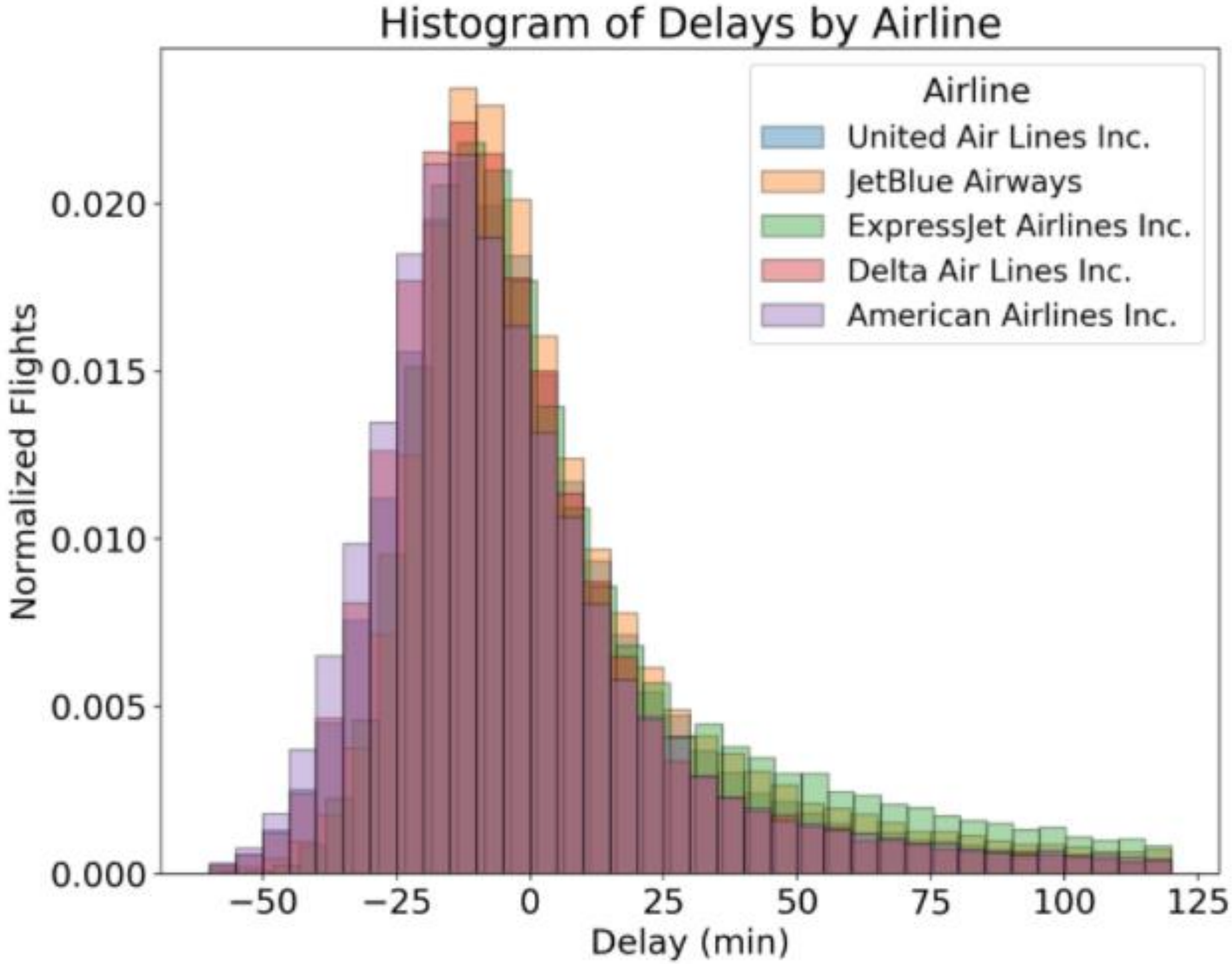


Multiple Overlapped Density plots

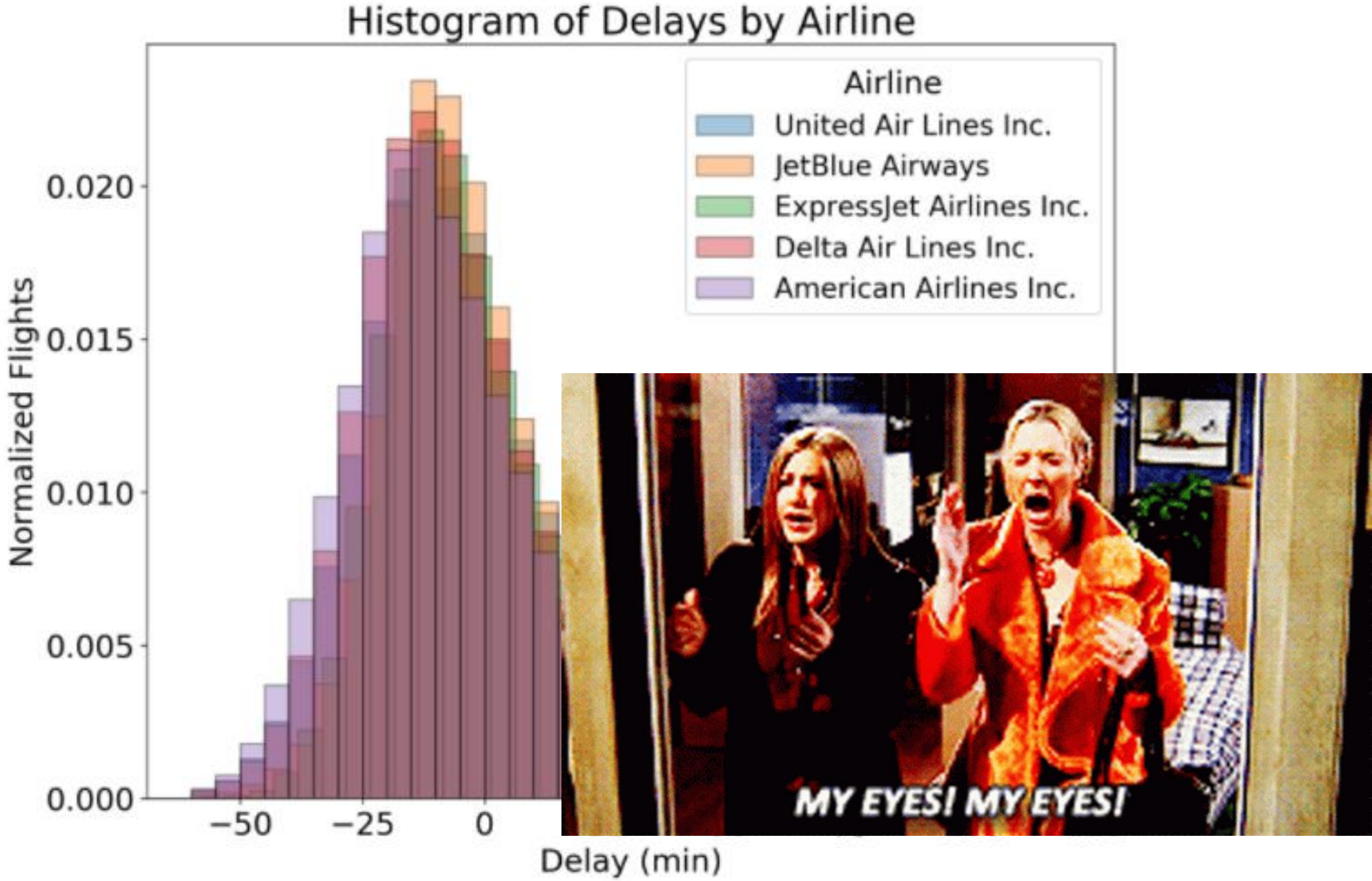
- To visualize several distributions at once, kernel density plots will generally work better than histograms.



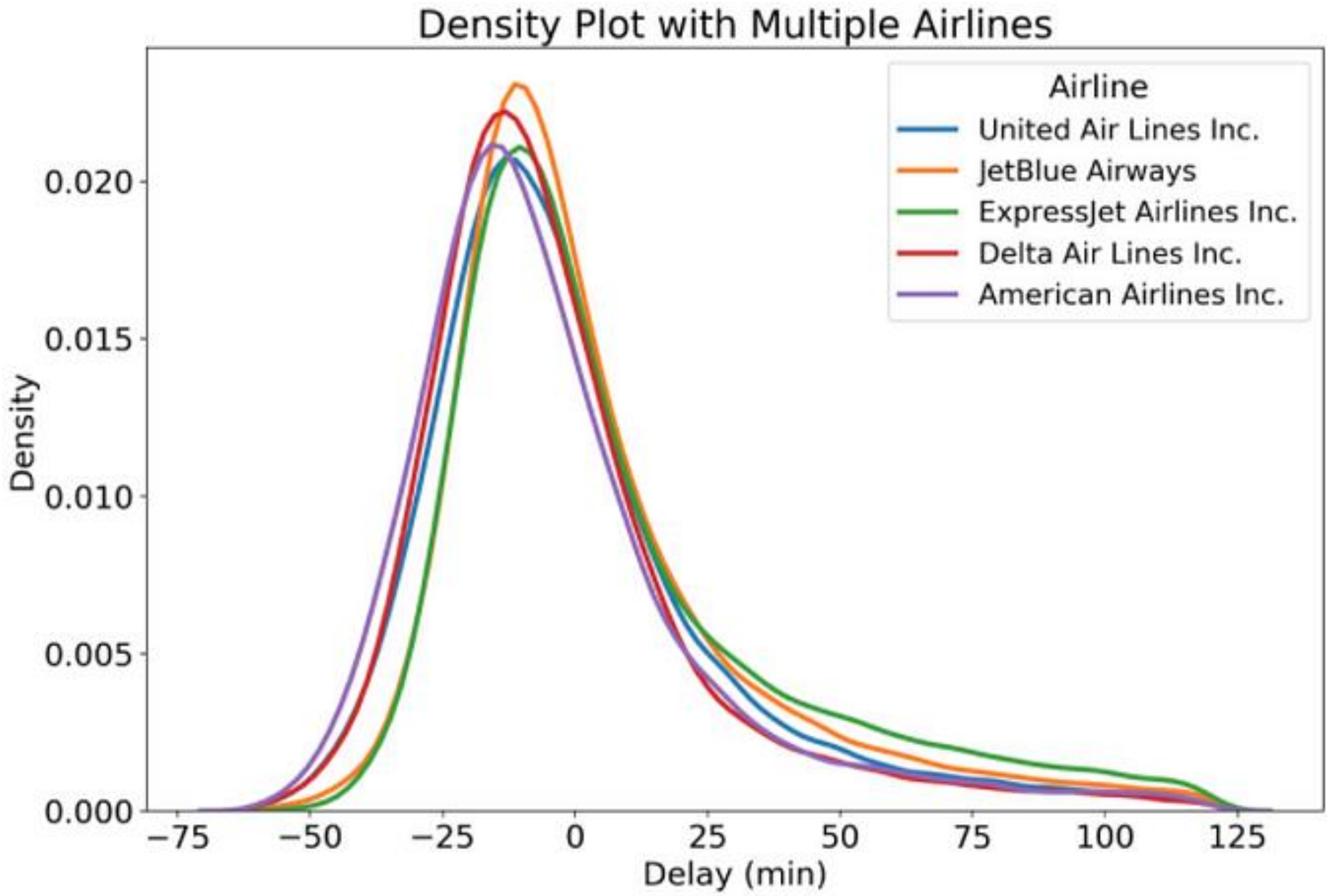
What do you think of this plot?



What do you think of this plot?



How about this one?



Recap: Pros and Cons of Histogram and KDEs

- They are intuitive and visually appealing
- The resulting figure depends on parameters the user has to choose, such as the bin width for histograms and the bandwidth for density plots.
- As a result, both have to be considered as an interpretation of the data rather than a direct visualization of the data itself.

Methods to calculate the number of bins of a histogram

- Number of bins= $1+3.3\log_{10}(n)$

n: number of observations

- Sturge's rule: No of bins = $1+\log_{2}(n)$

Alternatives

- As an alternative to using histograms or density plots, we could simply show all the data points individually like strip charts and boxplots.
- However, this approach becomes less effective for very large datasets.
- To solve this problem, statisticians have invented empirical cumulative distribution functions (ECDFs) and quantile-quantile (q-q) plots.
- These types of visualizations require no arbitrary parameter choices
- They are a little less intuitive than a histogram, or a density plot
- They are often found in highly technical publications.

CDF of a Random Variable

- The cumulative distribution function (cdf) gives the probability that the random variable X is less than or equal to x and is usually denoted F(x). The cumulative distribution function of a random variable X is the function given by:

$$F(x) = \mathrm{P}[X \leq x].$$

- CDF of a Continuous Random Variable:

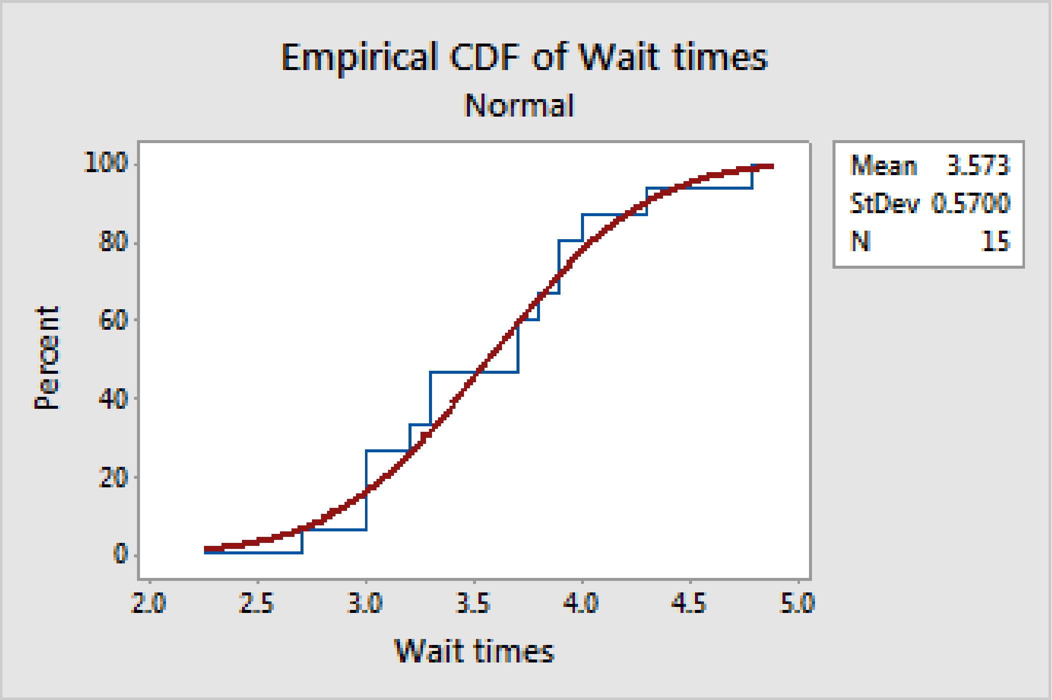
$$F(x) = \int_{-\infty}^{x} f(t)\mathrm{d}t.$$

- CDF of a discreet Random Variable:

$$F_X(x) = \mathrm{P}(X \leq x) = \sum_{x_i \leq x} \mathrm{P}(X = x_i) = \sum_{x_i \leq x} p(x_i).$$

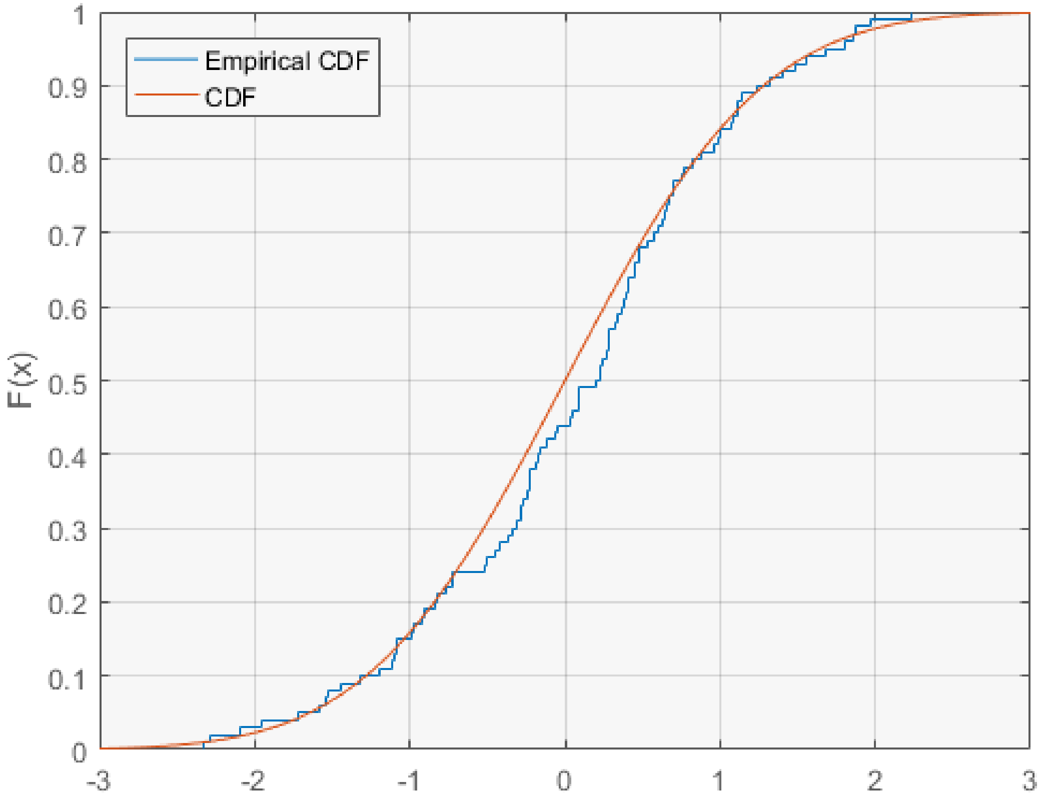
Empirical CDF

- An Empirical CDF is an estimator of the Cumulative Distribution Function.

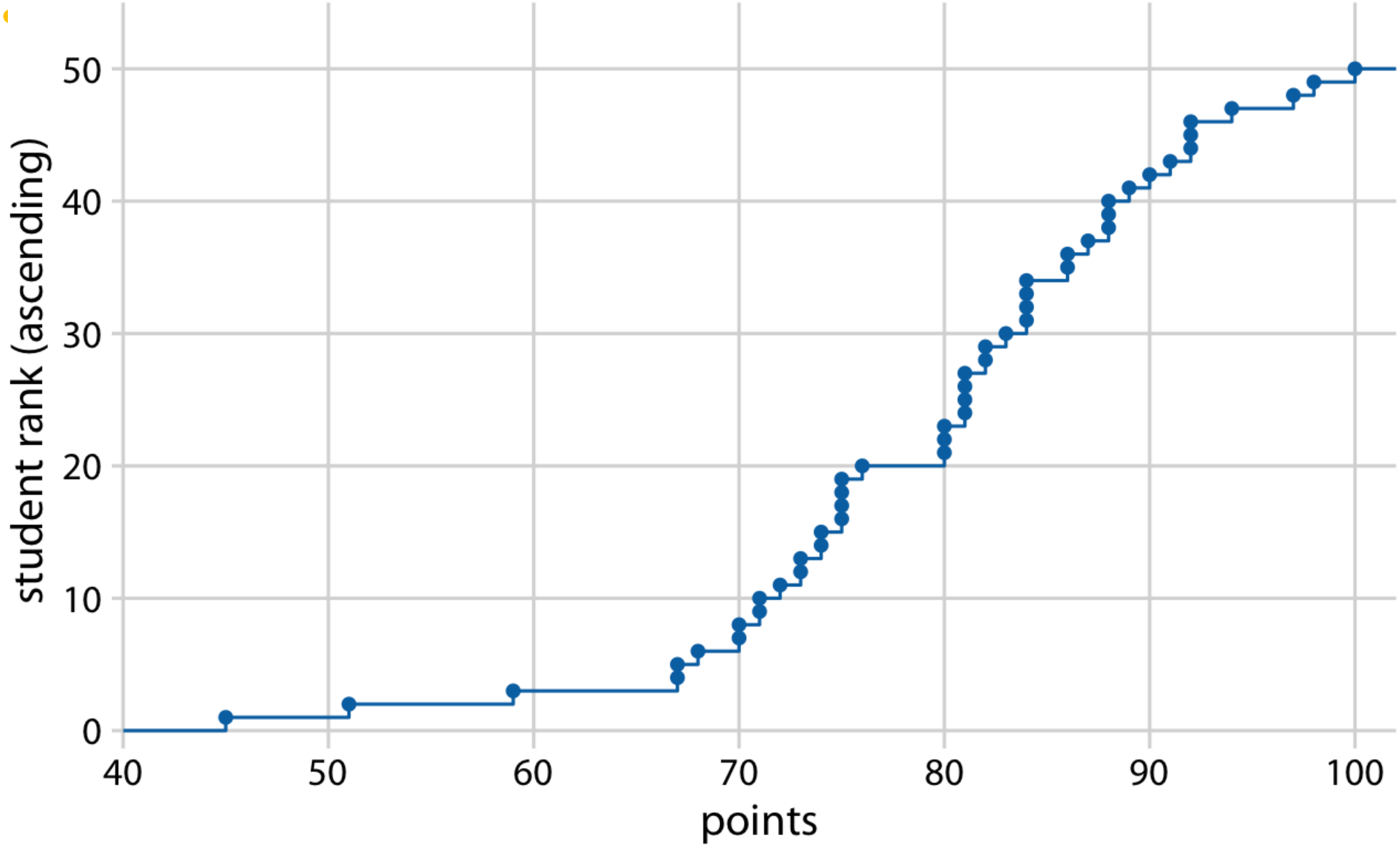


ECDF vs. CDF

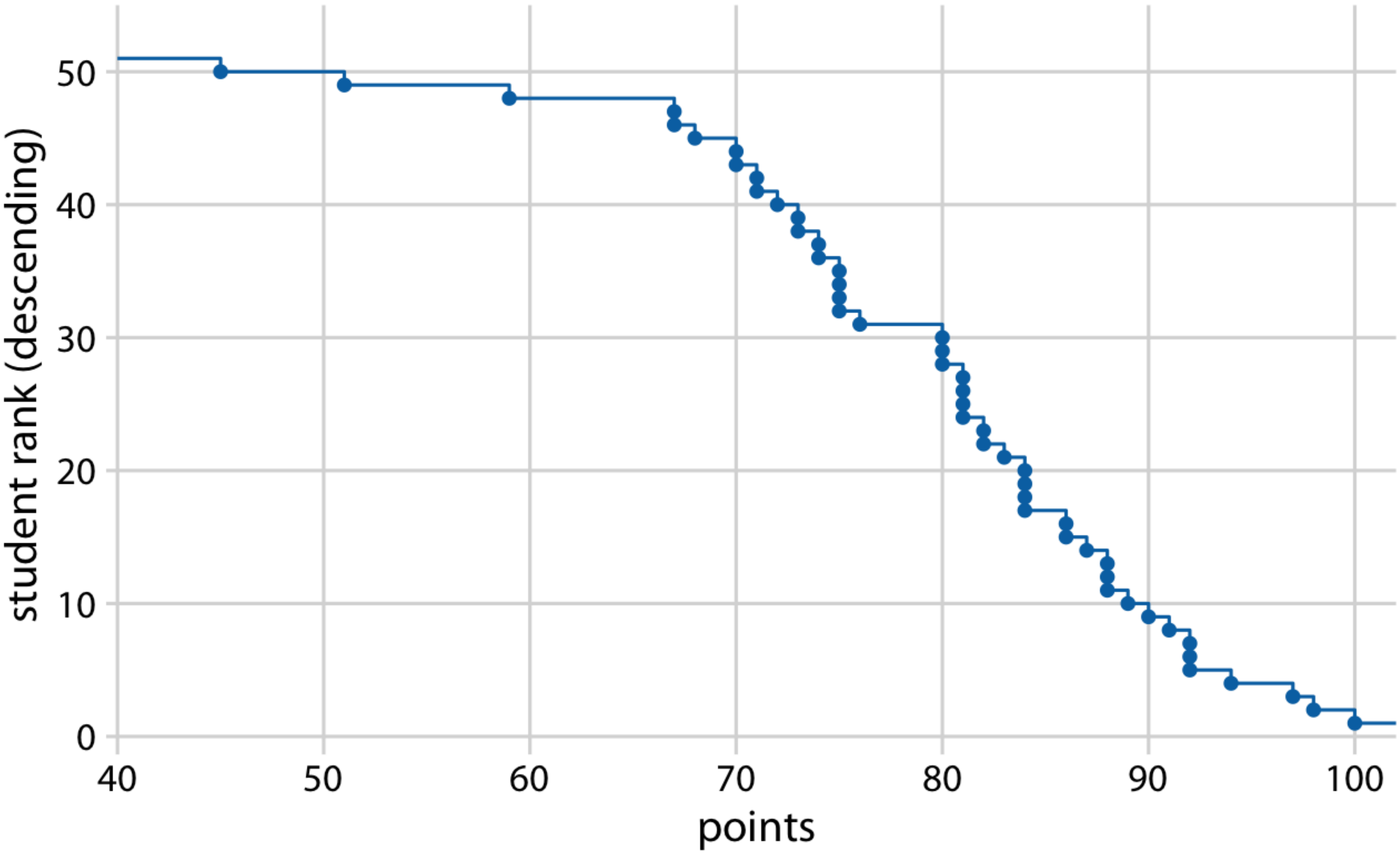
- The empirical CDF is built from an actual data set (Discreet)
- The CDF is a theoretical construct - it is what you would see if you could take infinitely many samples (continuous)



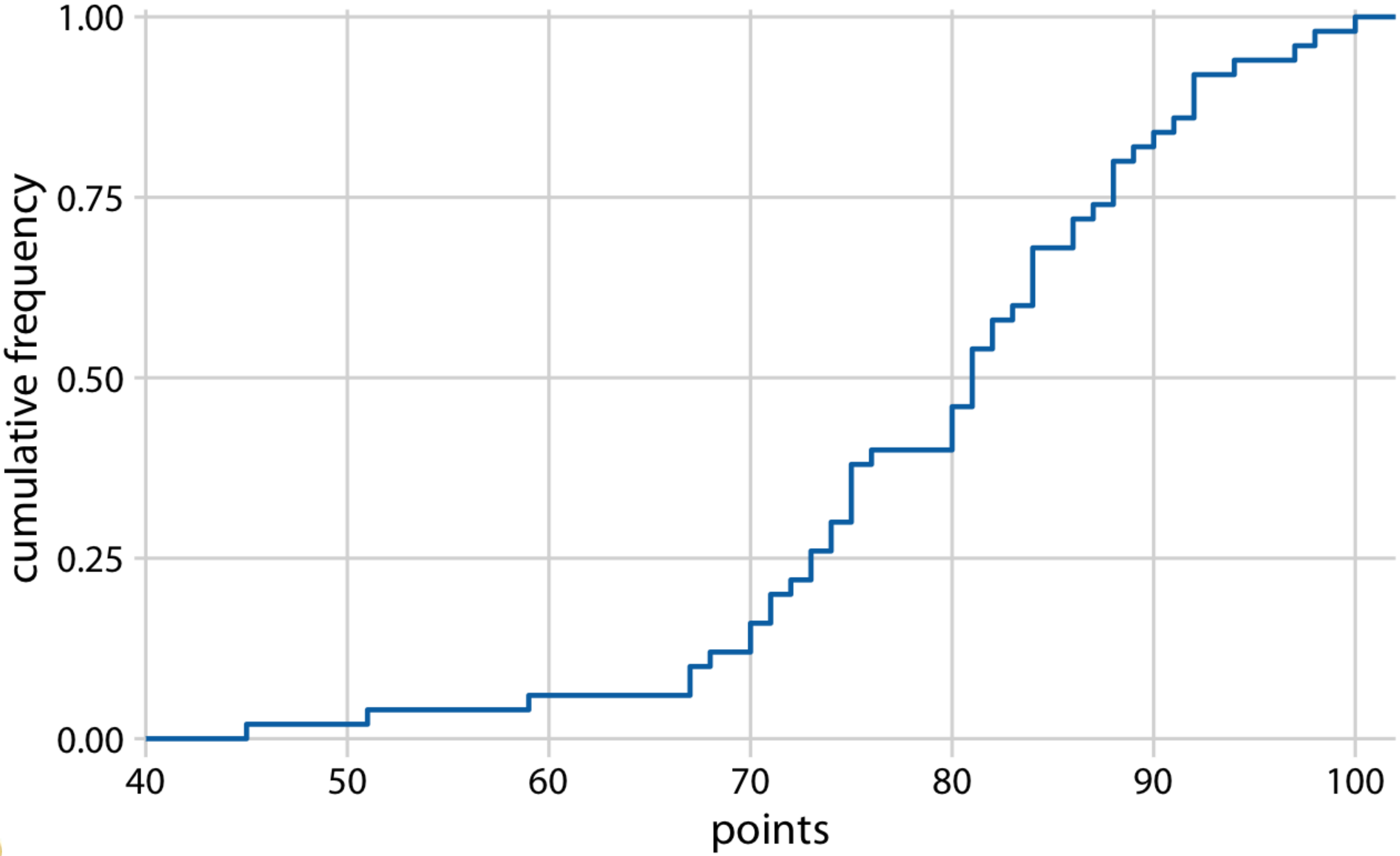
Empirical Cumulative Distribution Functions



Descending ECDF (Exceedance curve)



ECDF with Normalized Y scale



In-class Problem

iClicker

- An insurance company collected the data of all claimed incidents and company payments (assessed damage) for the last two years.
- The company is interested in learning the percentage of incidents in which damages is more than certain values. What kind of plot is more appropriate for this goal?

A. Ascending ECDF

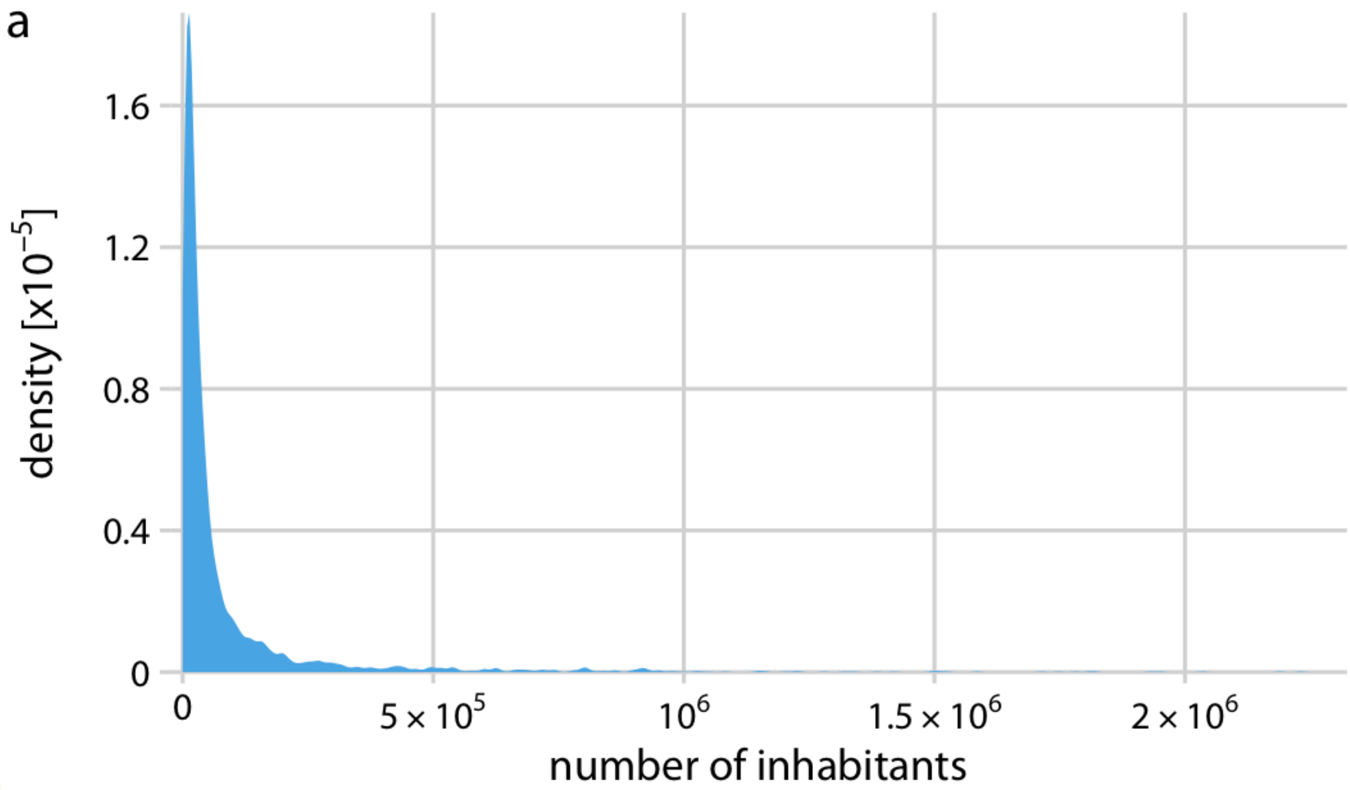
B. Descending ECDF

C. Histogram

D. Kernel Density Estimate

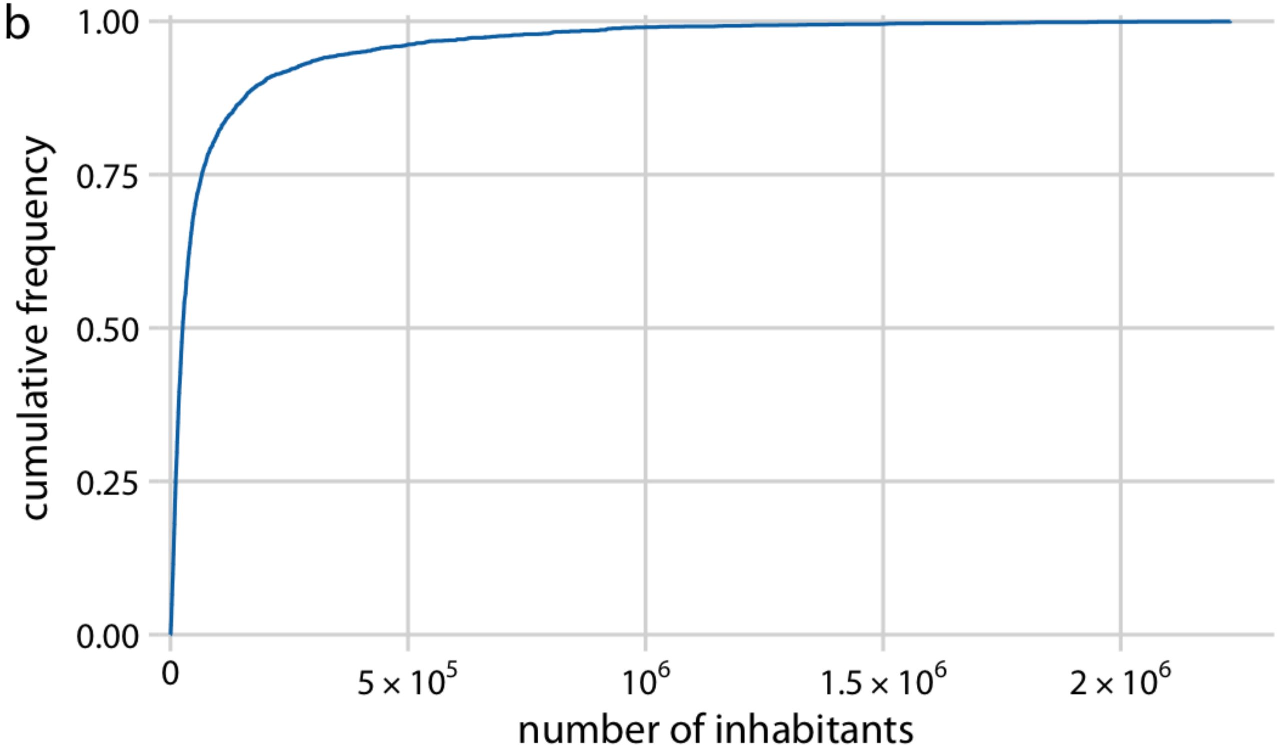
Highly Skewed Distributions

- The number of people living in different US counties according to the 2010 US Census according to the 2010 US Census



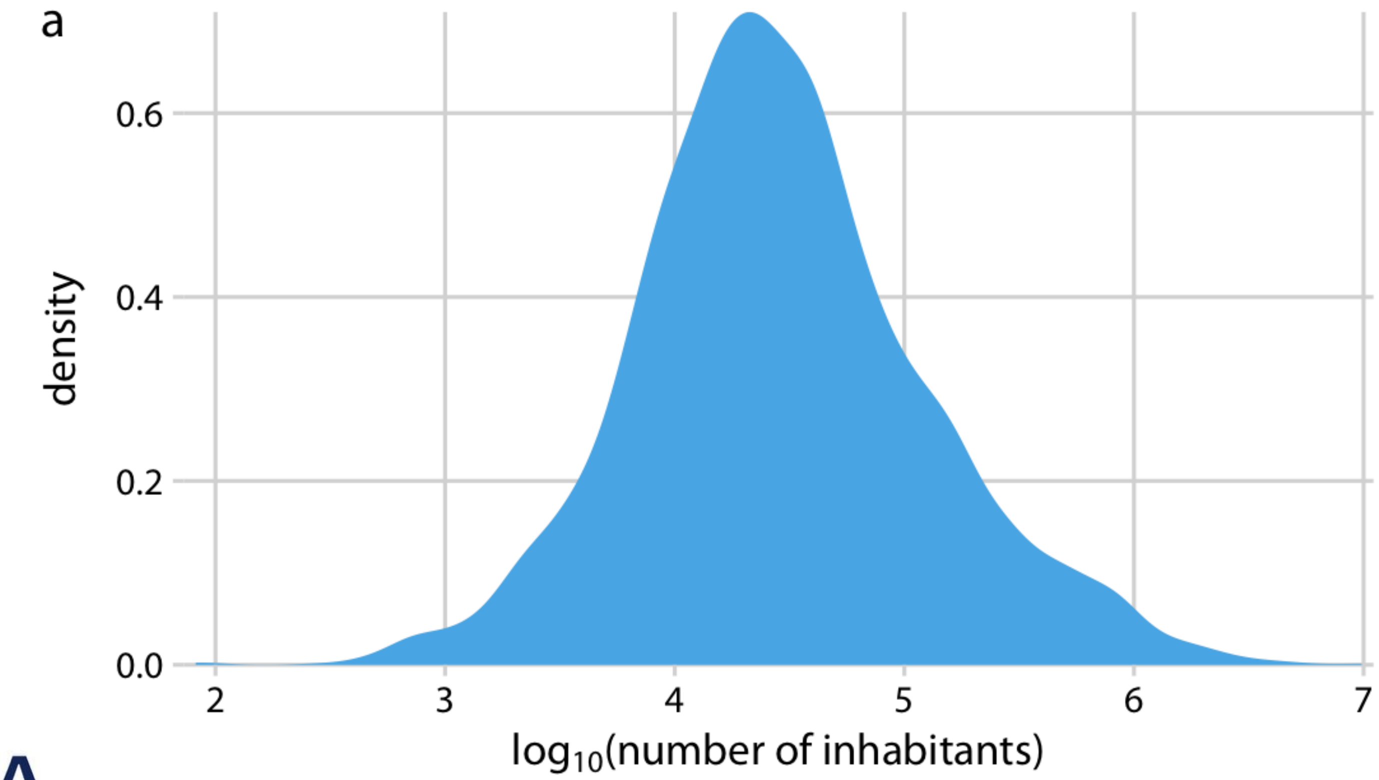
Highly Skewed Distributions

- This distribution has a very long tail to the right.
- If we try to visualize the distribution of population counts as either a density plot or an ECDF, we obtain figures that are essentially useless

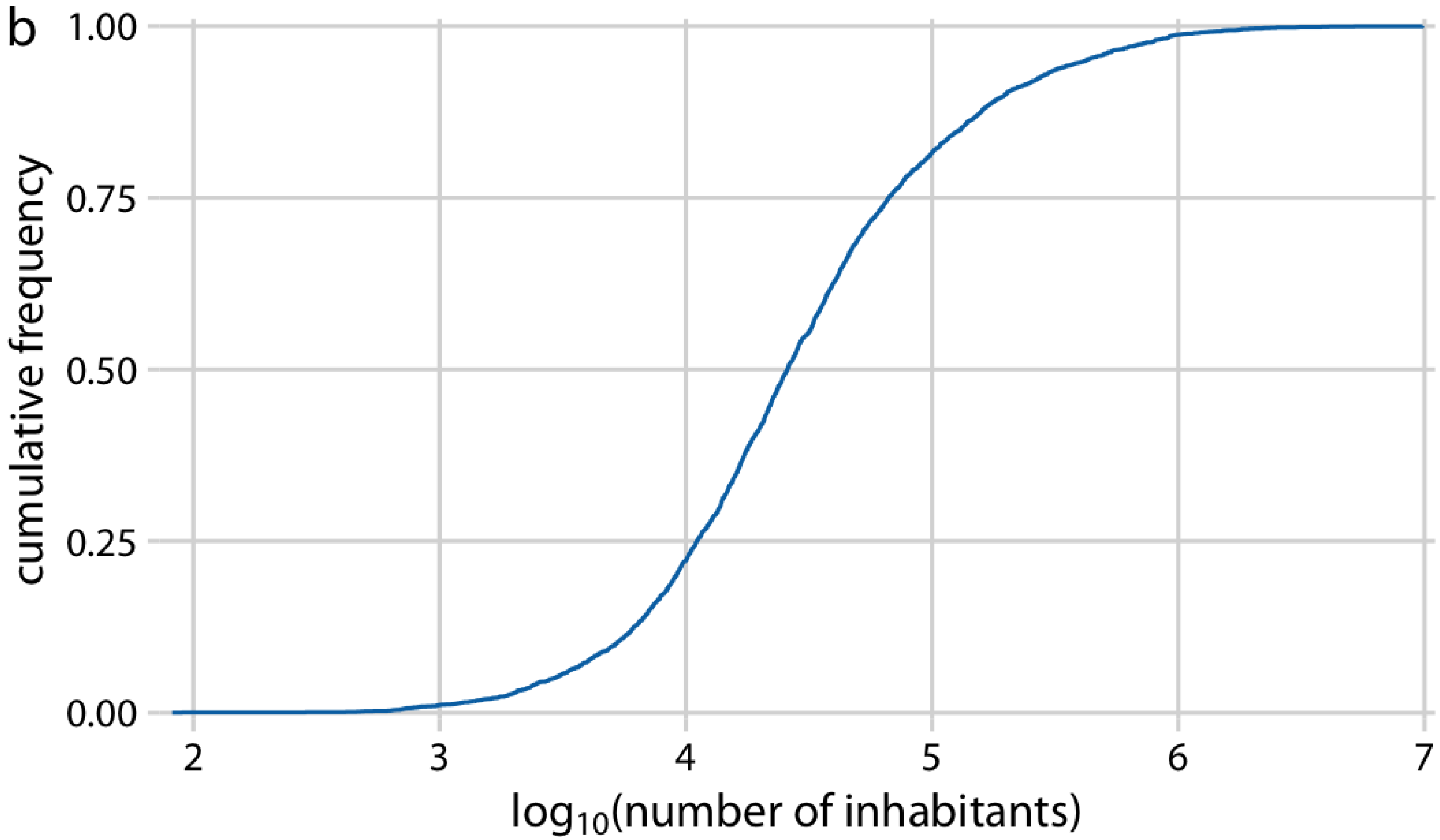


Log-transformed Density Plot

- This transformation works here because the distribution of population numbers in counties is a nearly perfect log-normal distribution

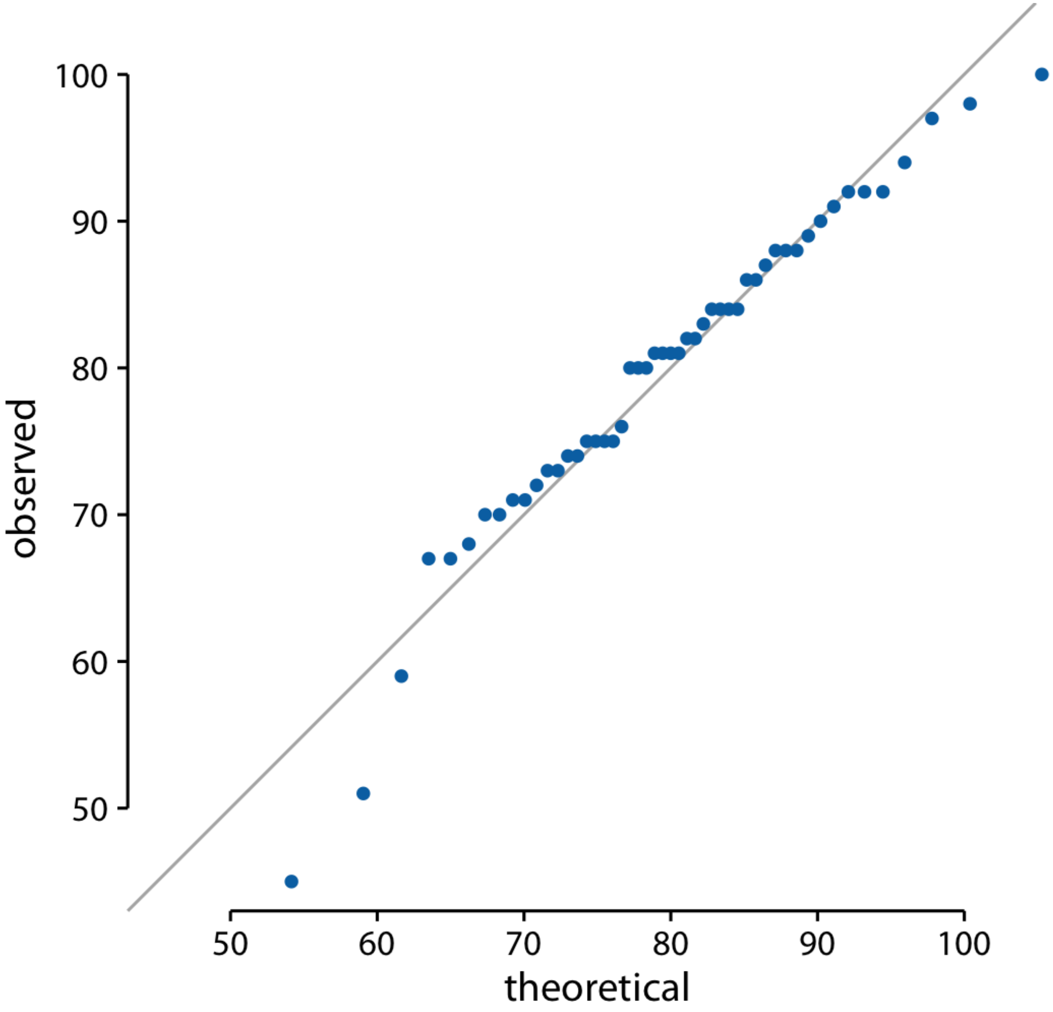


Log-transformed ECDF



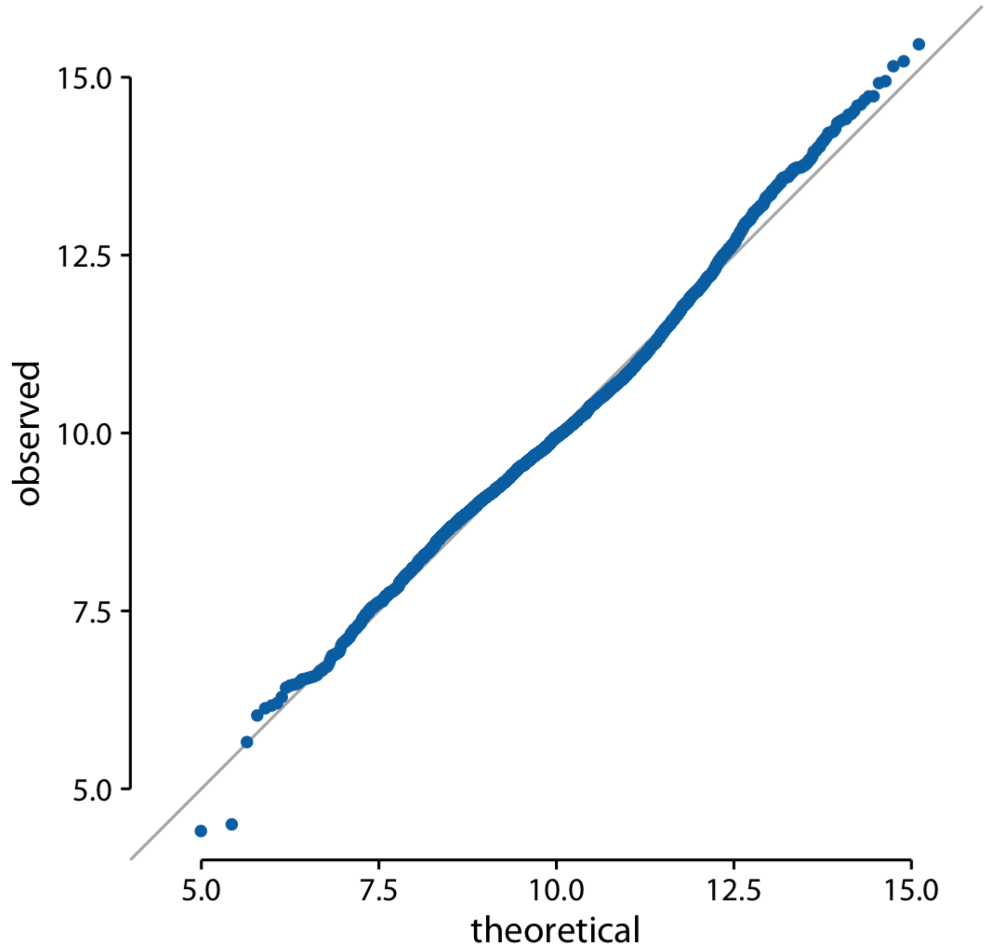
Quantile-Quantile plot

- Quantile-quantile (q-q) plots are useful visualizations when we want to determine to what extent the observed data points do or do not follow a given distribution.



Quantile-Quantile plot

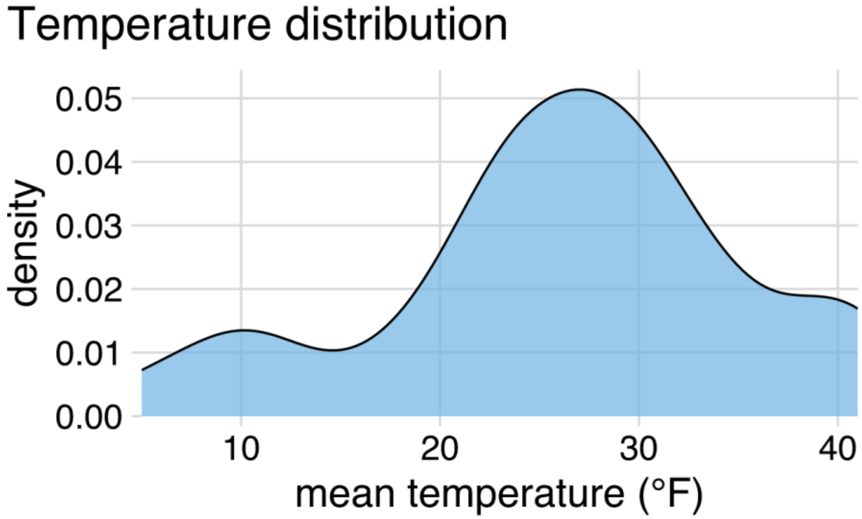
- q-q plot of the logarithm of the number of inhabitants in US counties



Visualizing Many Distributions at Once

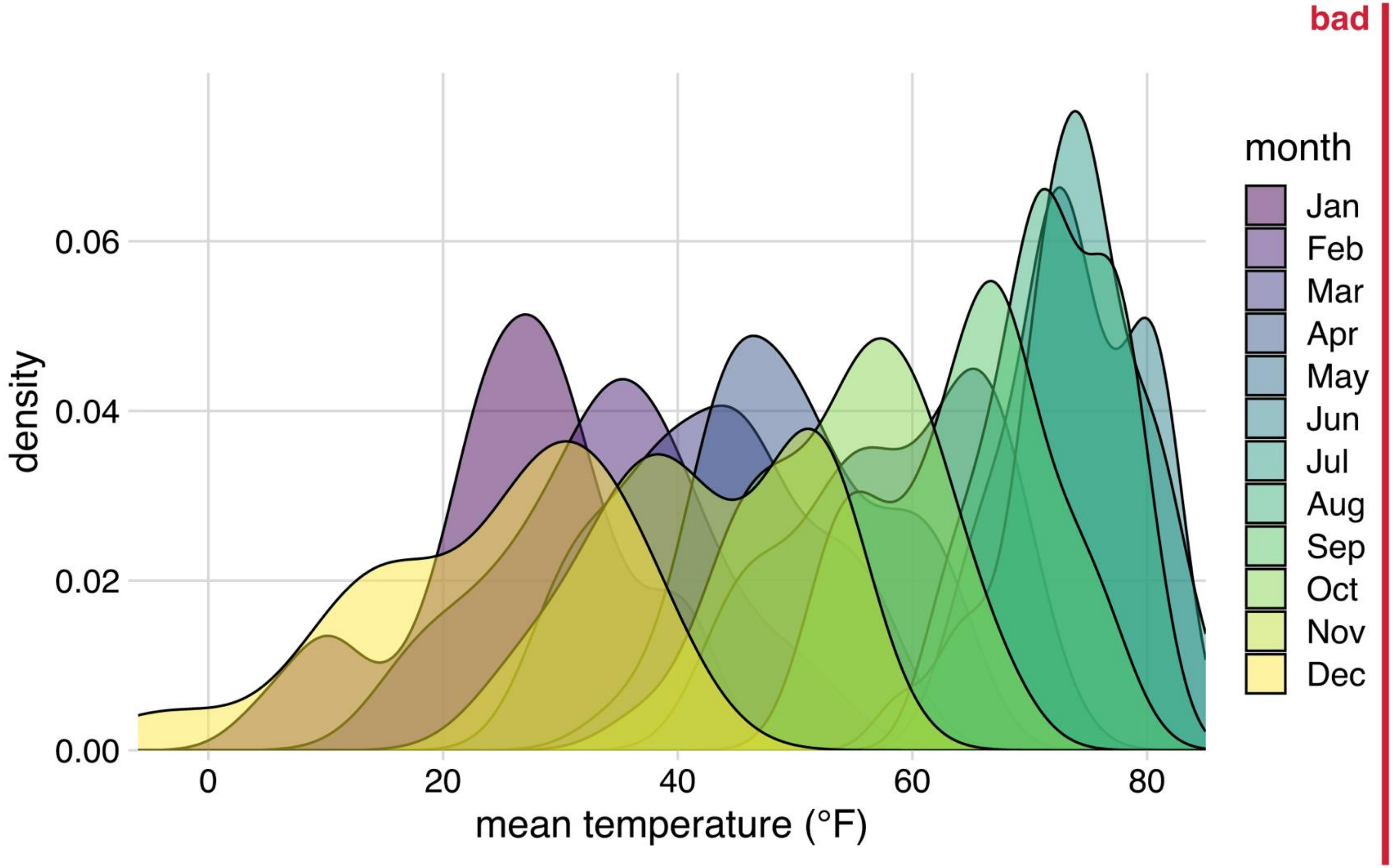
Mean temperatures in Lincoln, NE, in January 2016:

date	mean temp
2016-01-01	24
2016-01-02	23
2016-01-03	23
2016-01-04	17
2016-01-05	29
2016-01-06	33
2016-01-07	30
2016-01-08	25
2016-01-09	9
2016-01-10	11

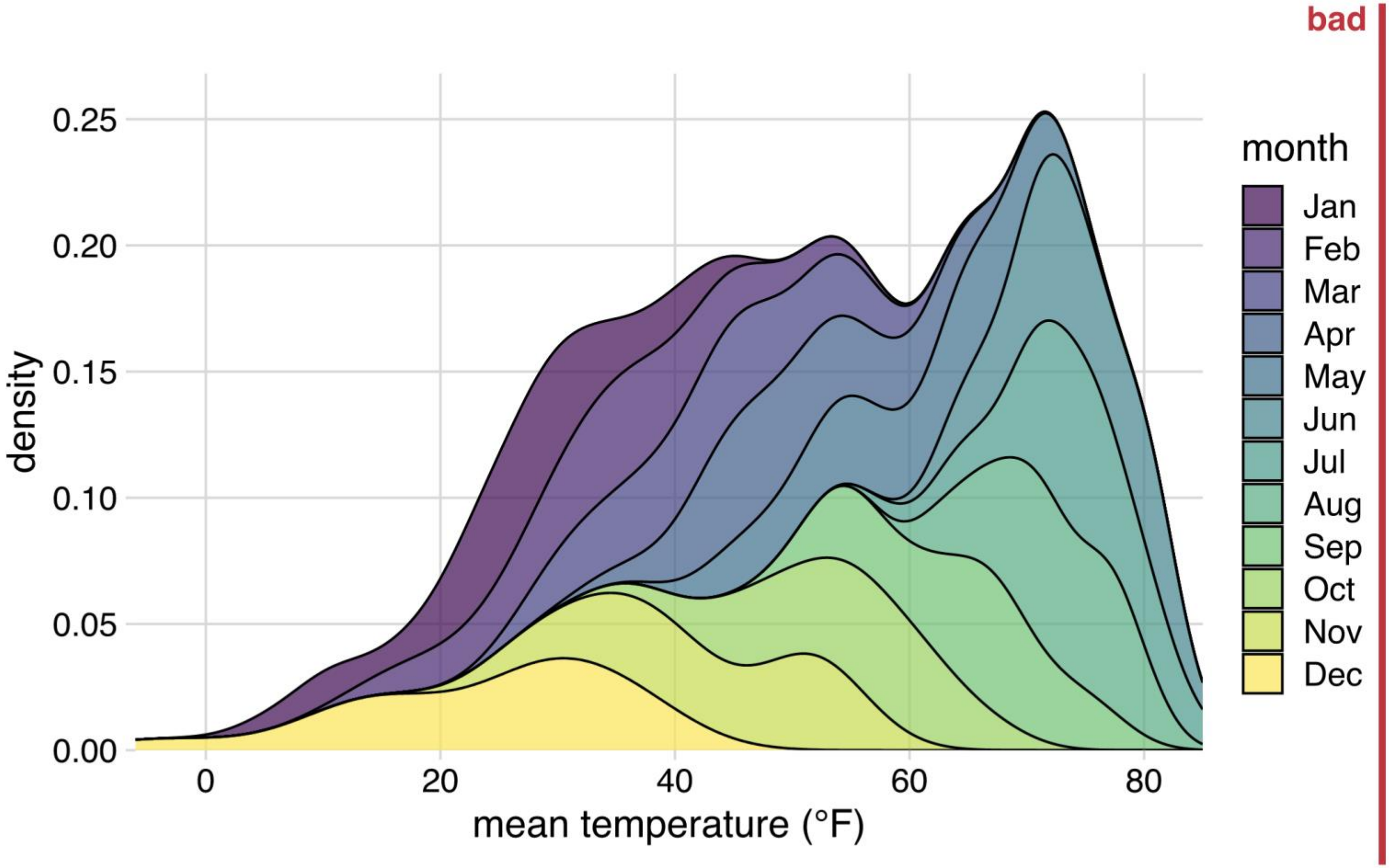


- How can we compare distributions across months?

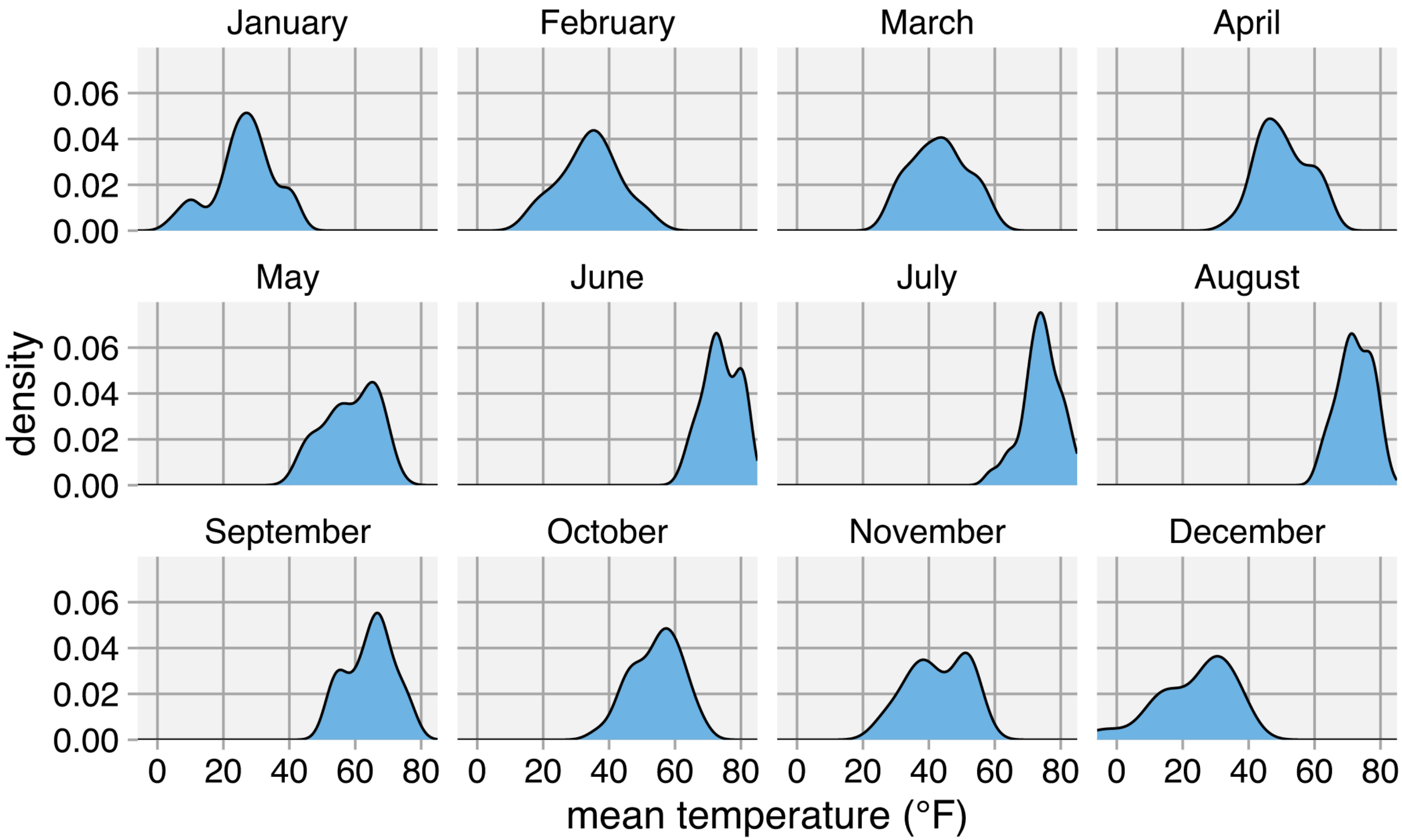
Many overlapping density plots



Stacked density plots

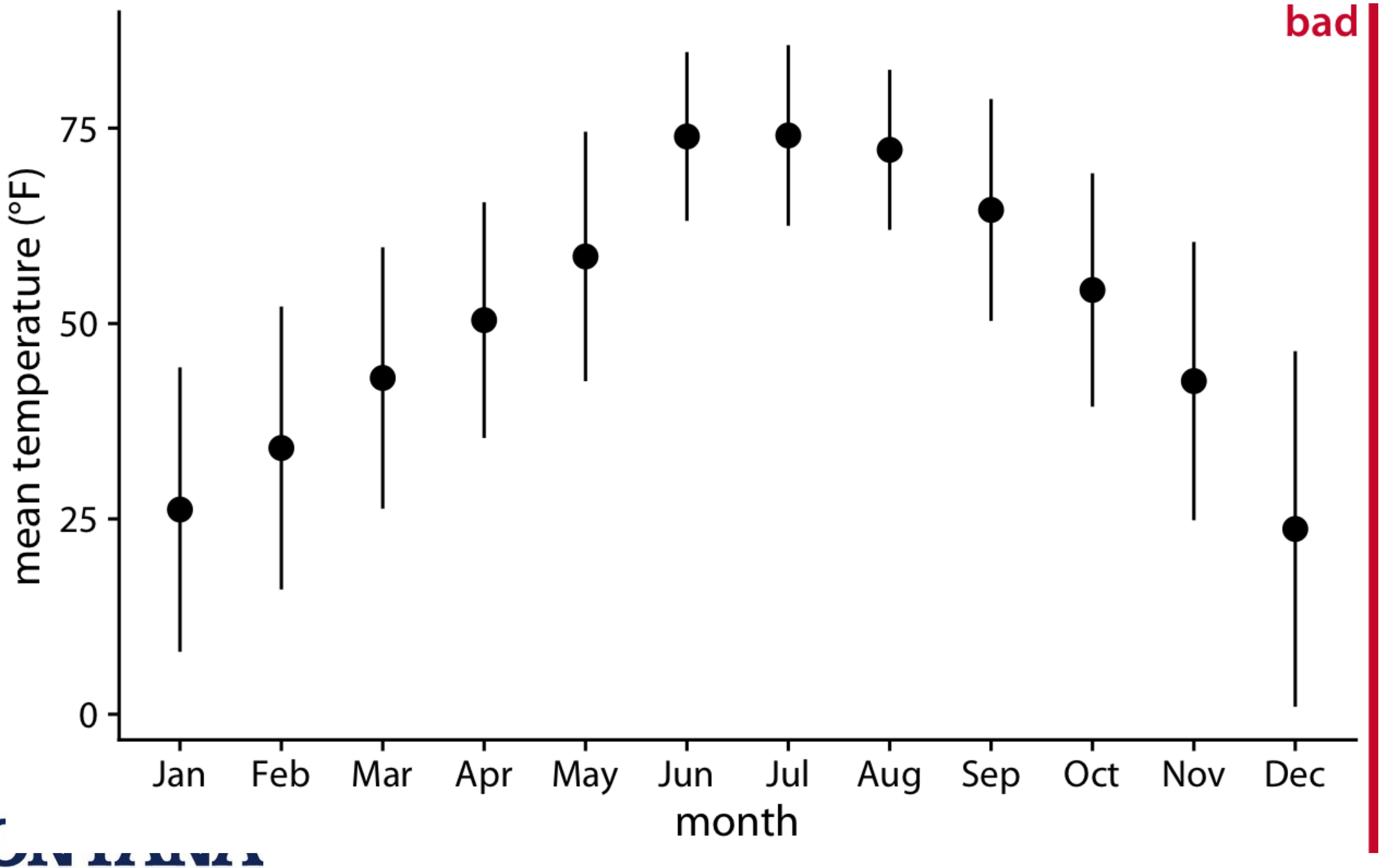


Small multiples

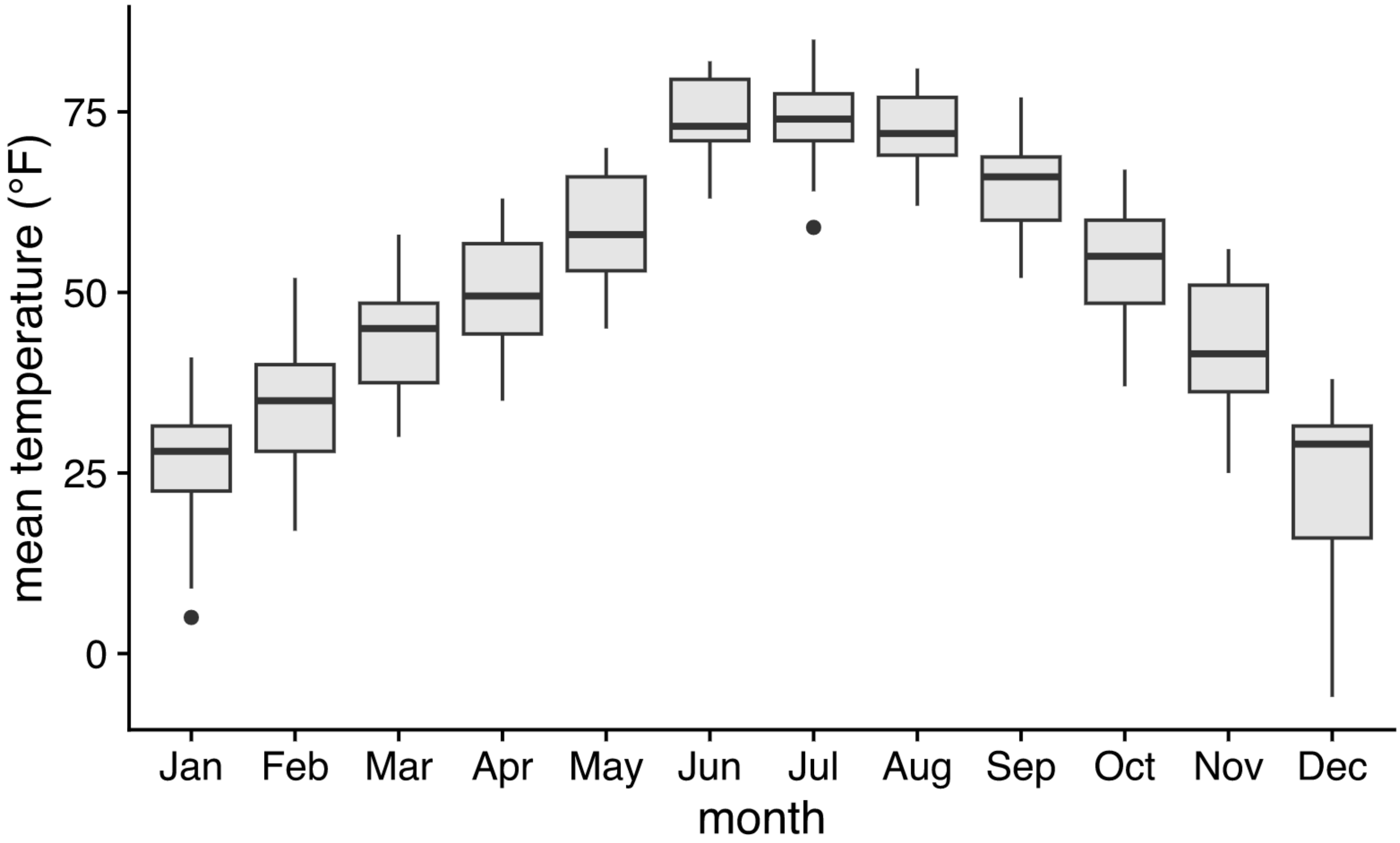


Visualizing Distributions Along the Vertical Axis

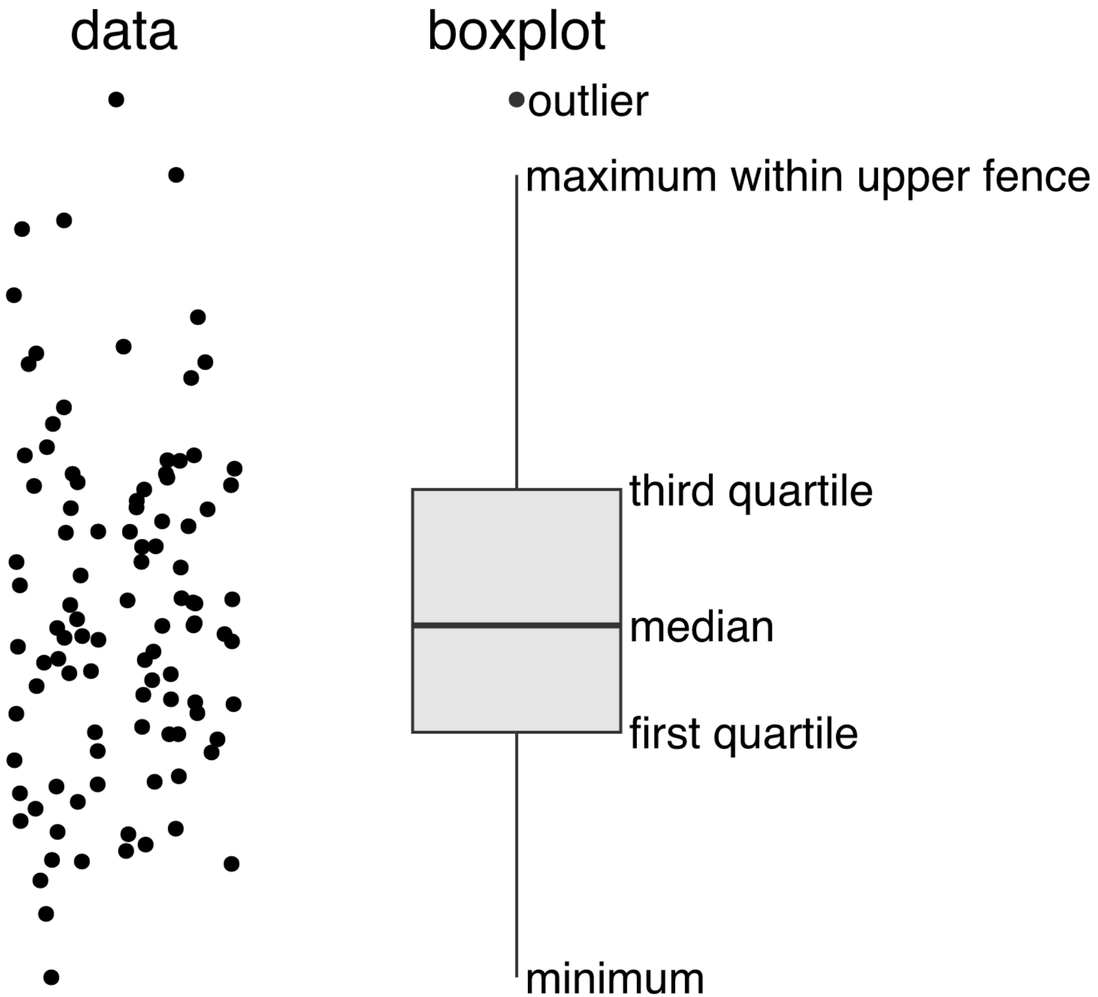
- a



Show values along y, conditions along x

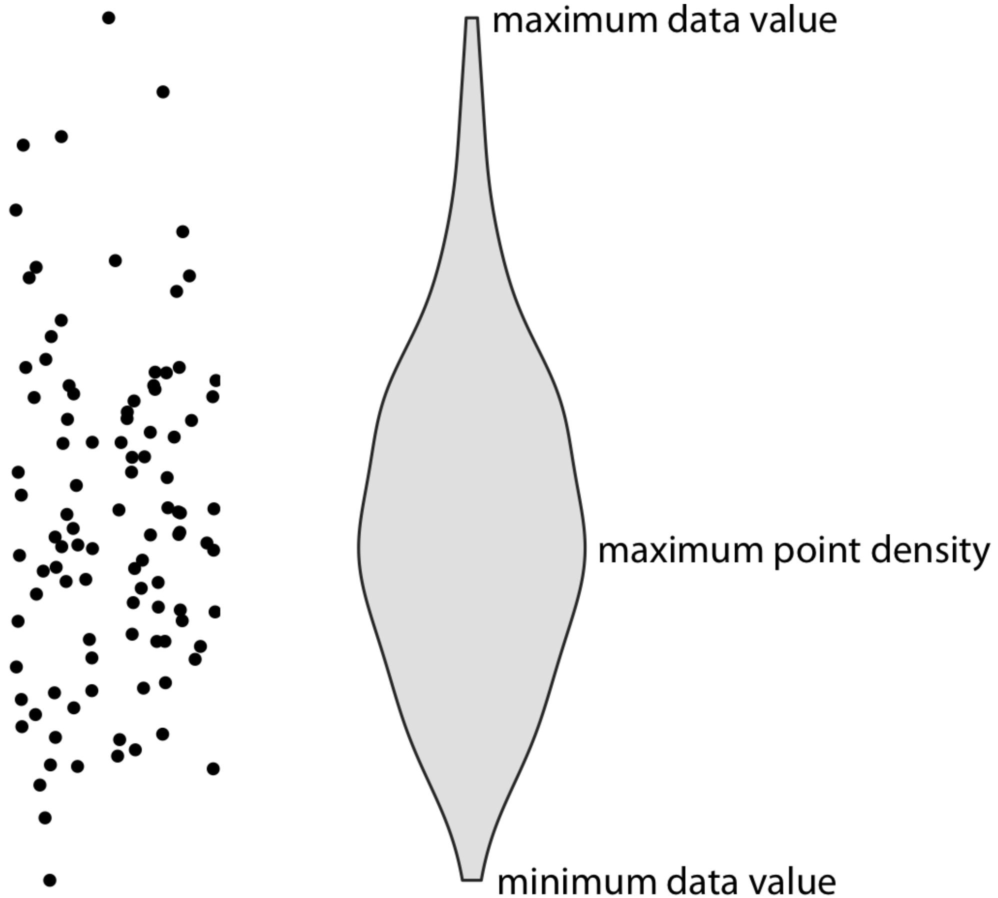


How to read a boxplot



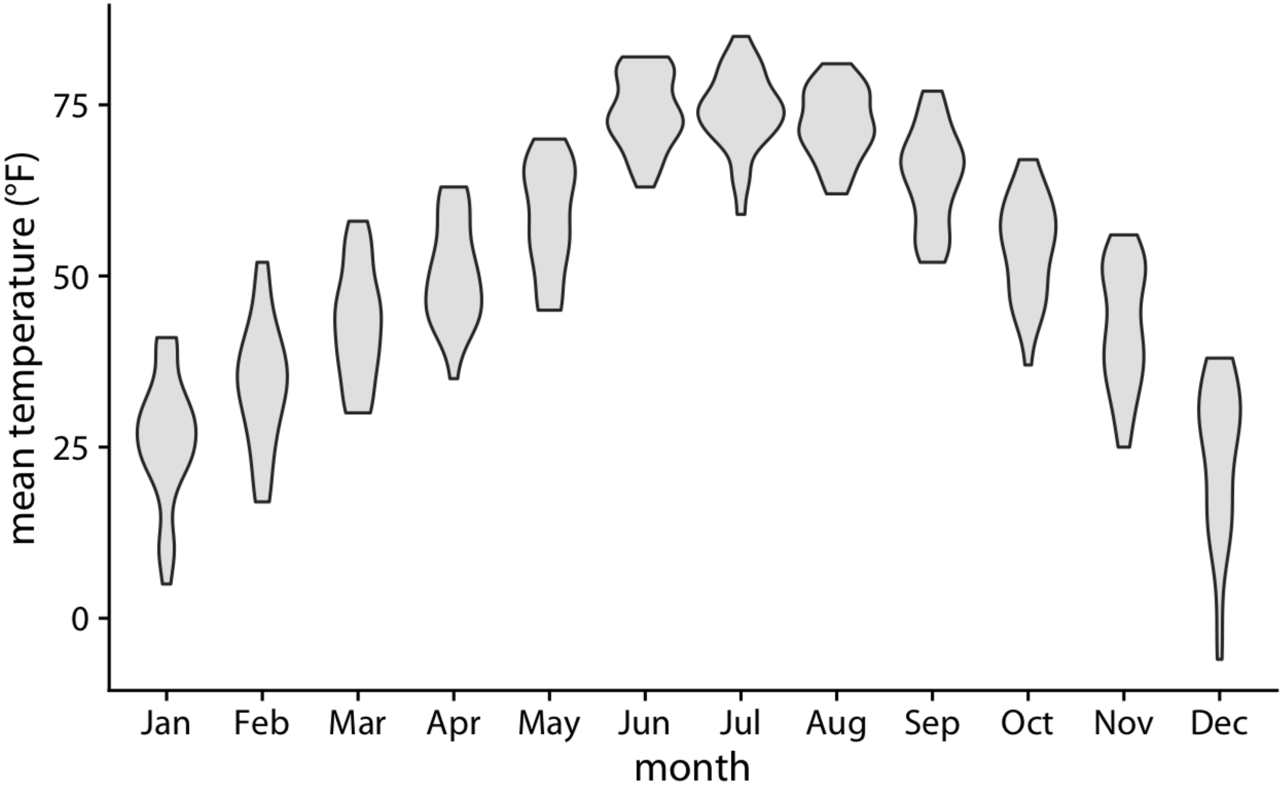
Violins plot

- If you like density plots, consider violins
- A violin plot is a density plot rotated 90 degrees and then mirrored.



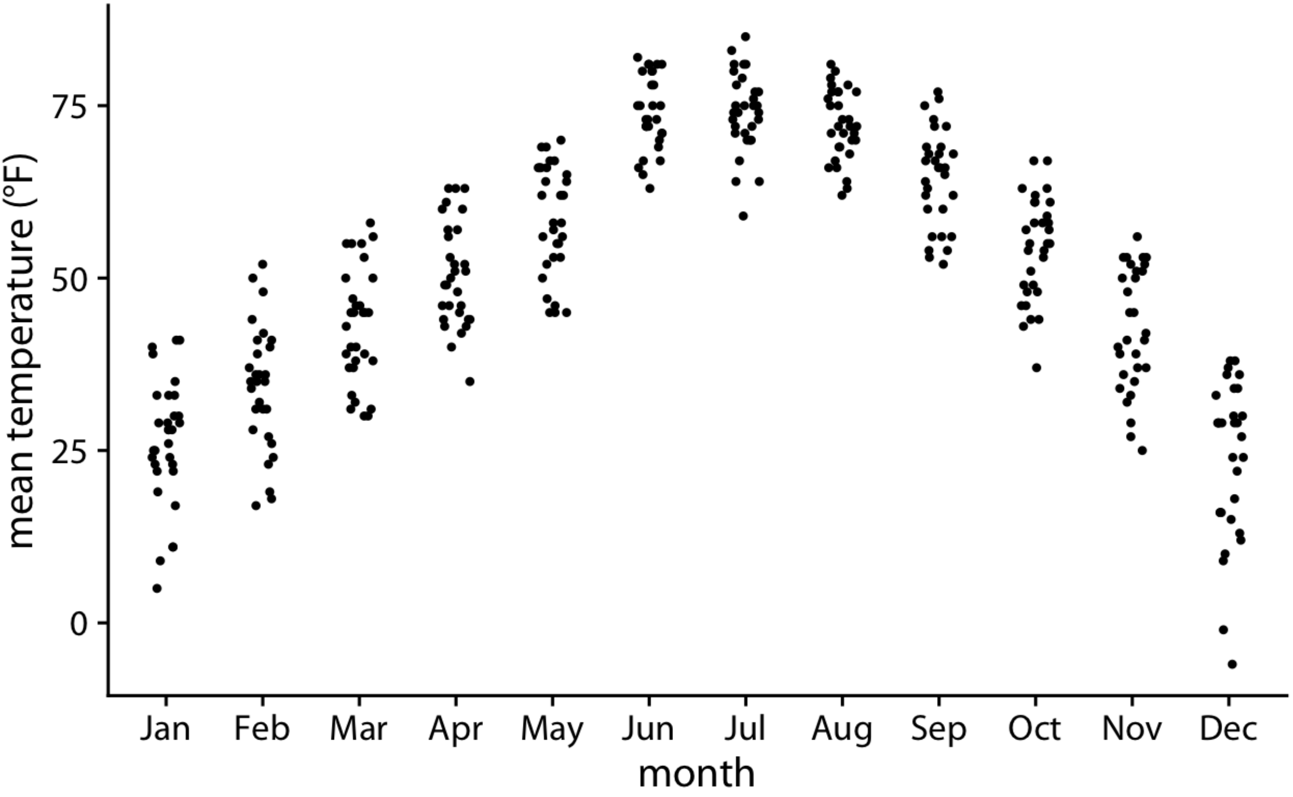
Violins plot

- Before using violins to visualize distributions, verify that you have sufficiently many data points in each group to justify showing the point densities as smooth lines.



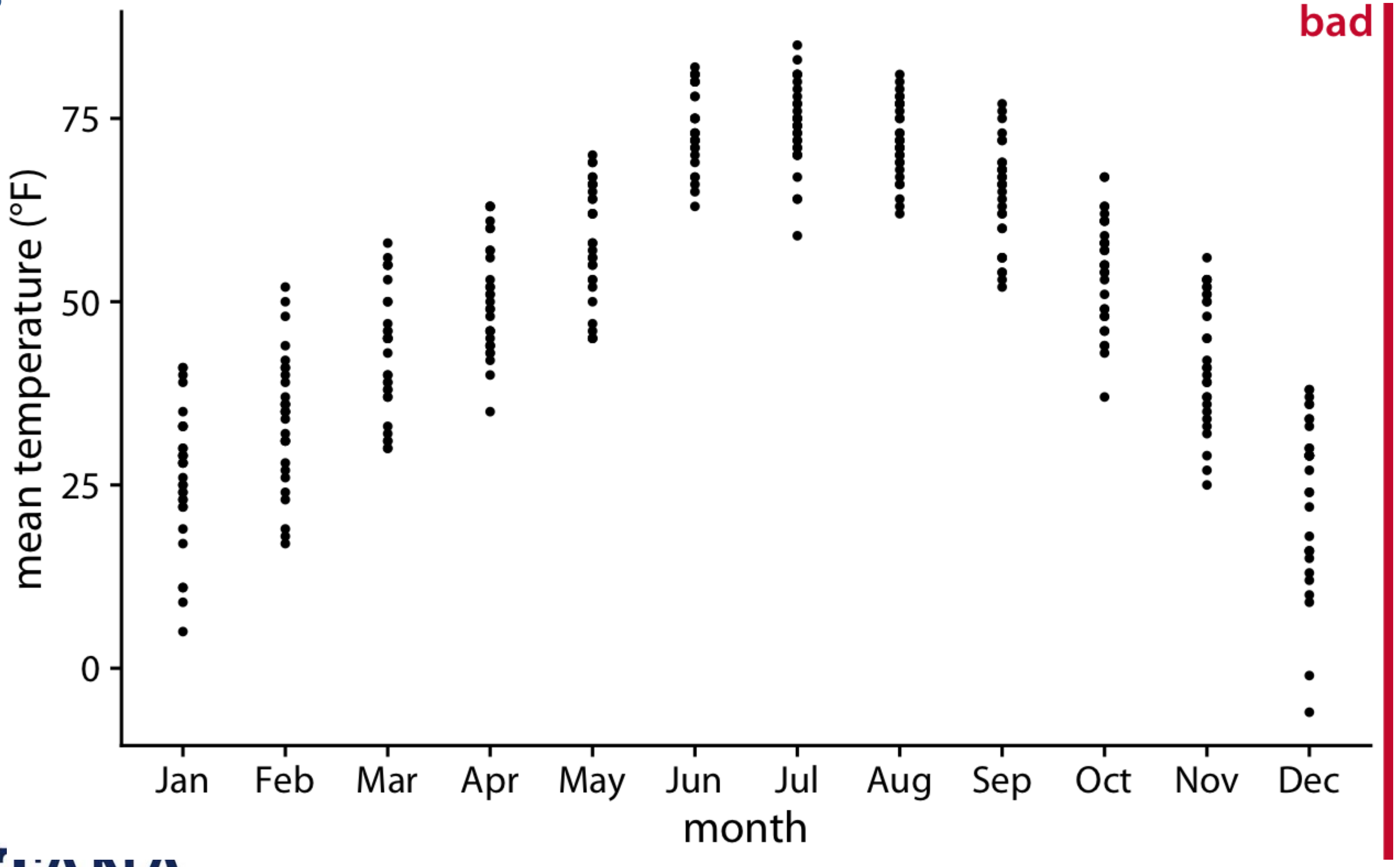
Strip charts

- For small datasets, you can also use a strip chart
- Advantage: Can see raw data points instead of abstract representation.



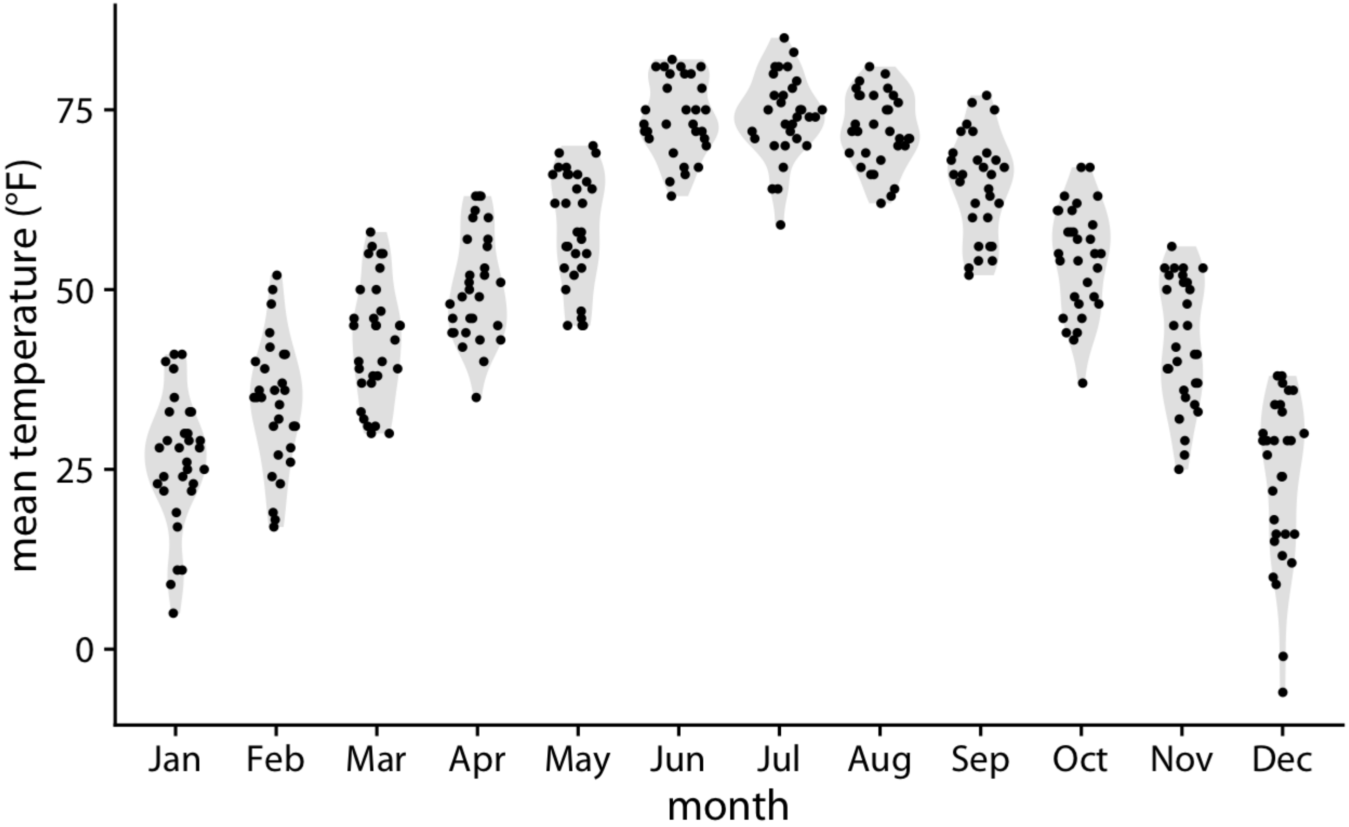
Strip charts

- Horizontal jittering may be necessary to avoid overlapping points



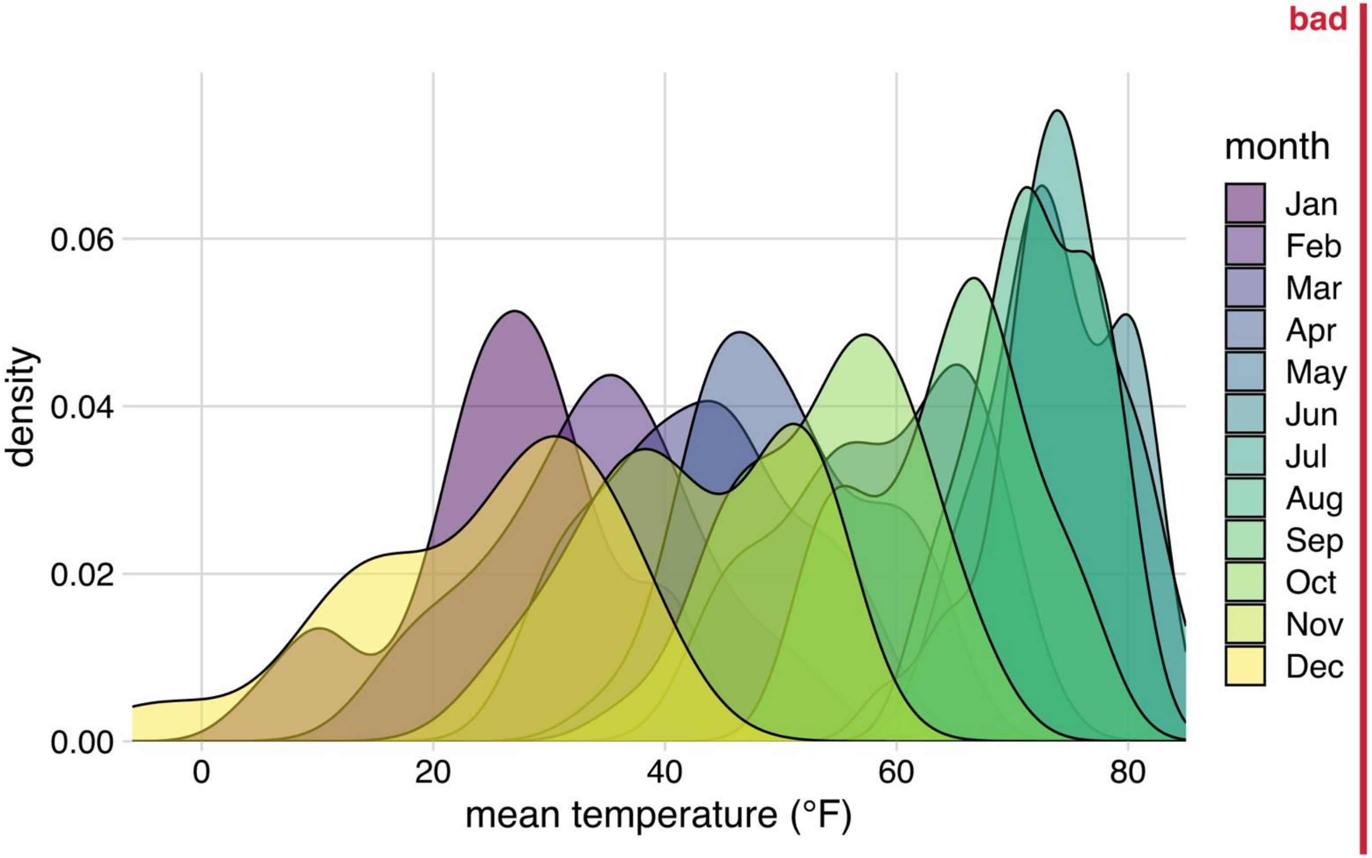
Sina plot

- We can also jitter points into violins
- Such plots are called Sina plots, to honor Sina Hadi Sohi.



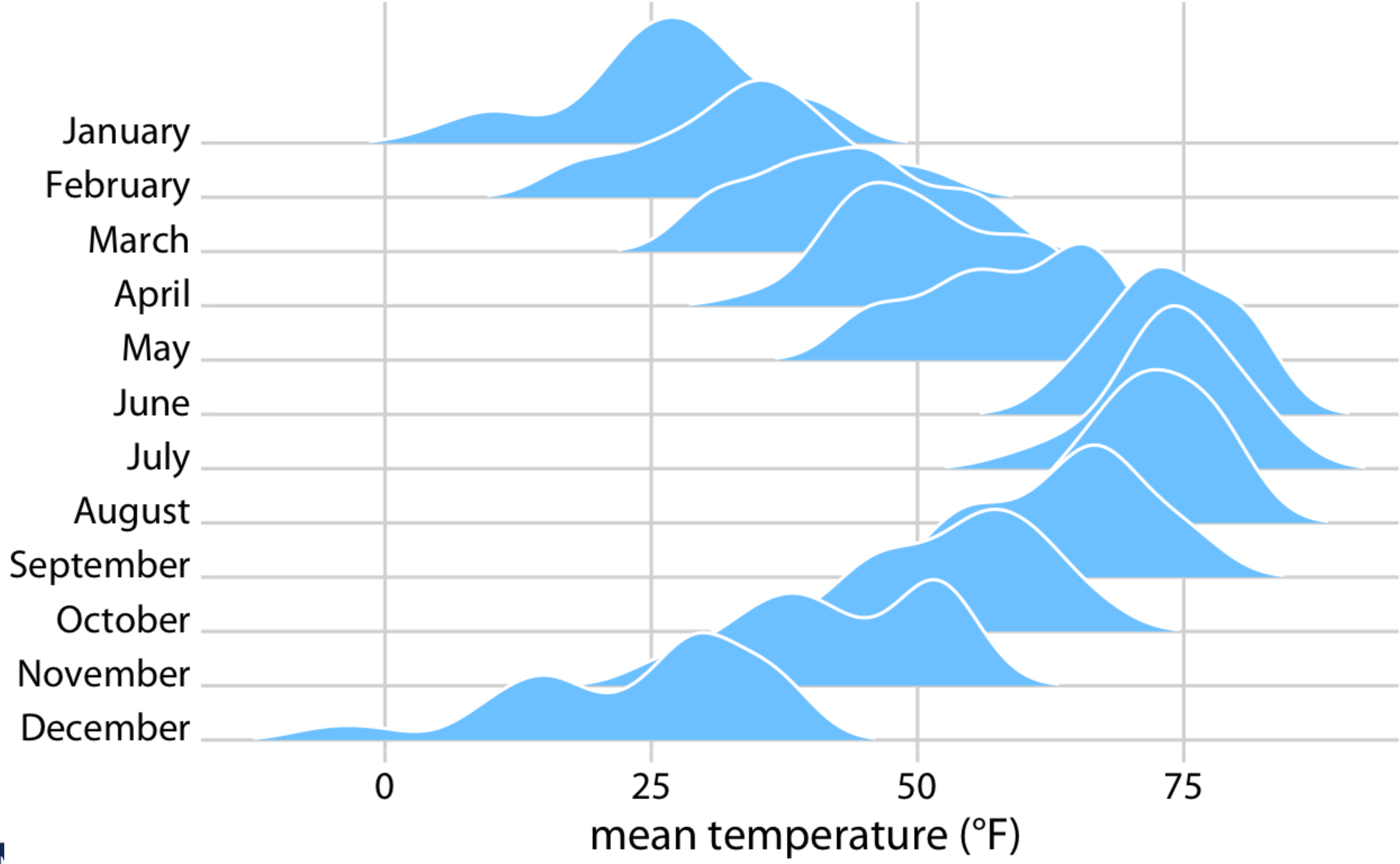
Visualizing Distributions Along the Horizontal Axis

- But maybe there's hope for overlapping density plots?
- How about we stagger the densities vertically?

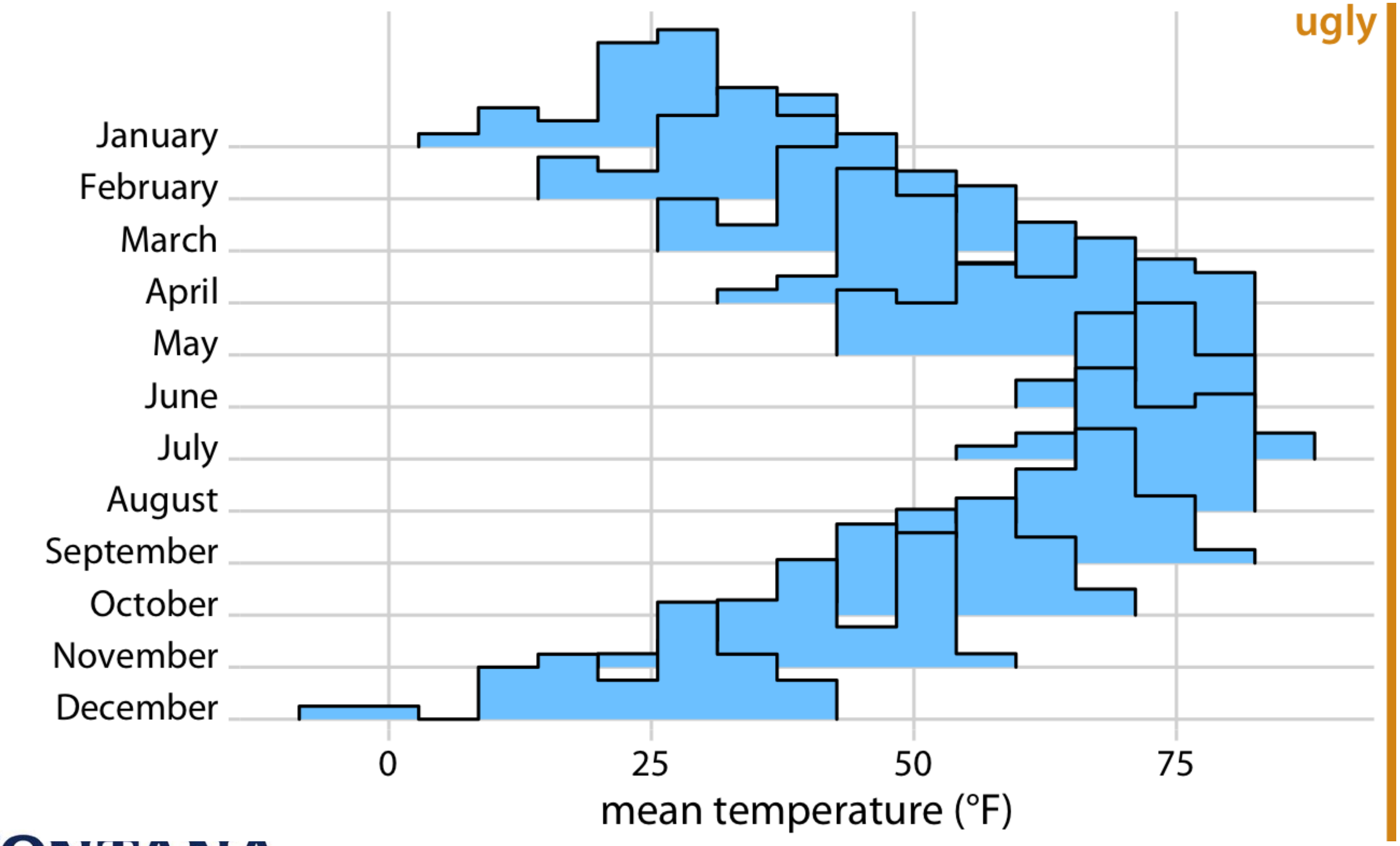


Ridgeline plot

- Vertically staggered density plots are called ridgelines
-

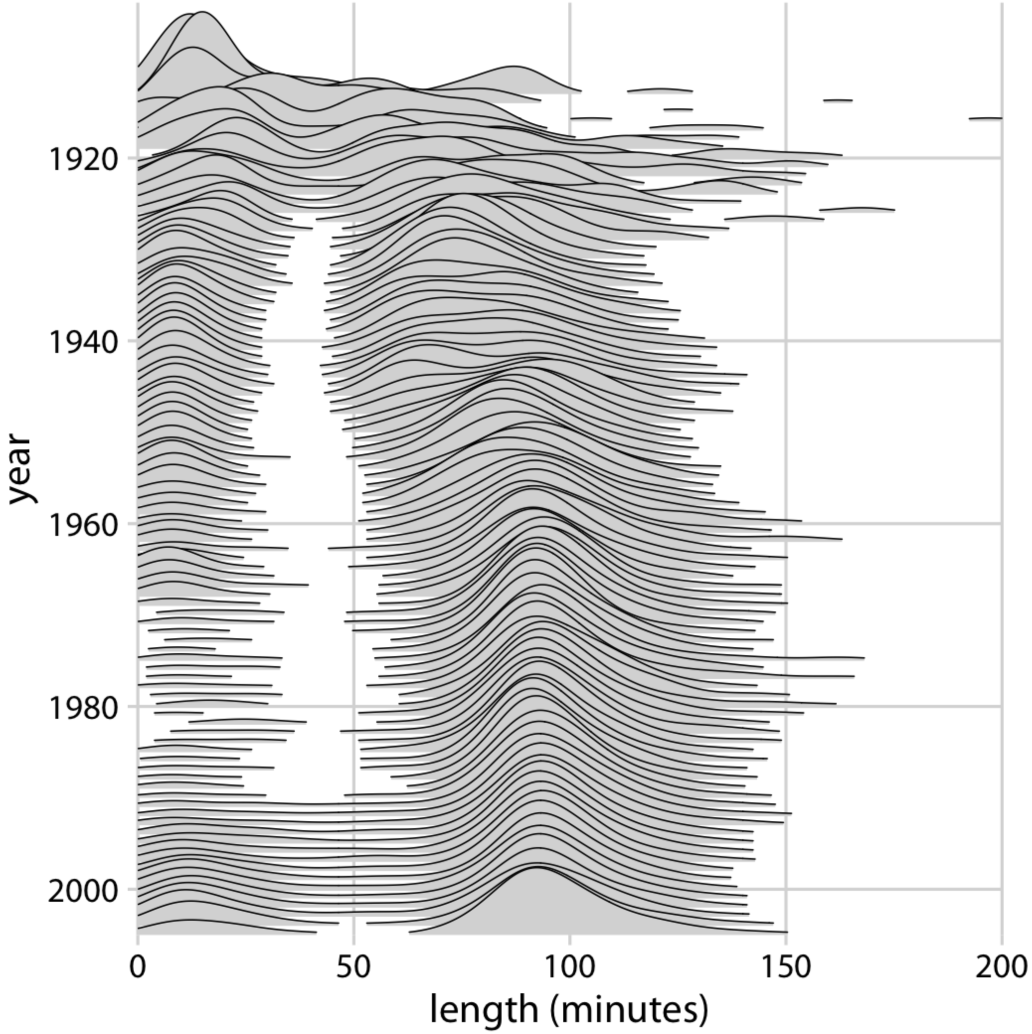


Ridgeline plot



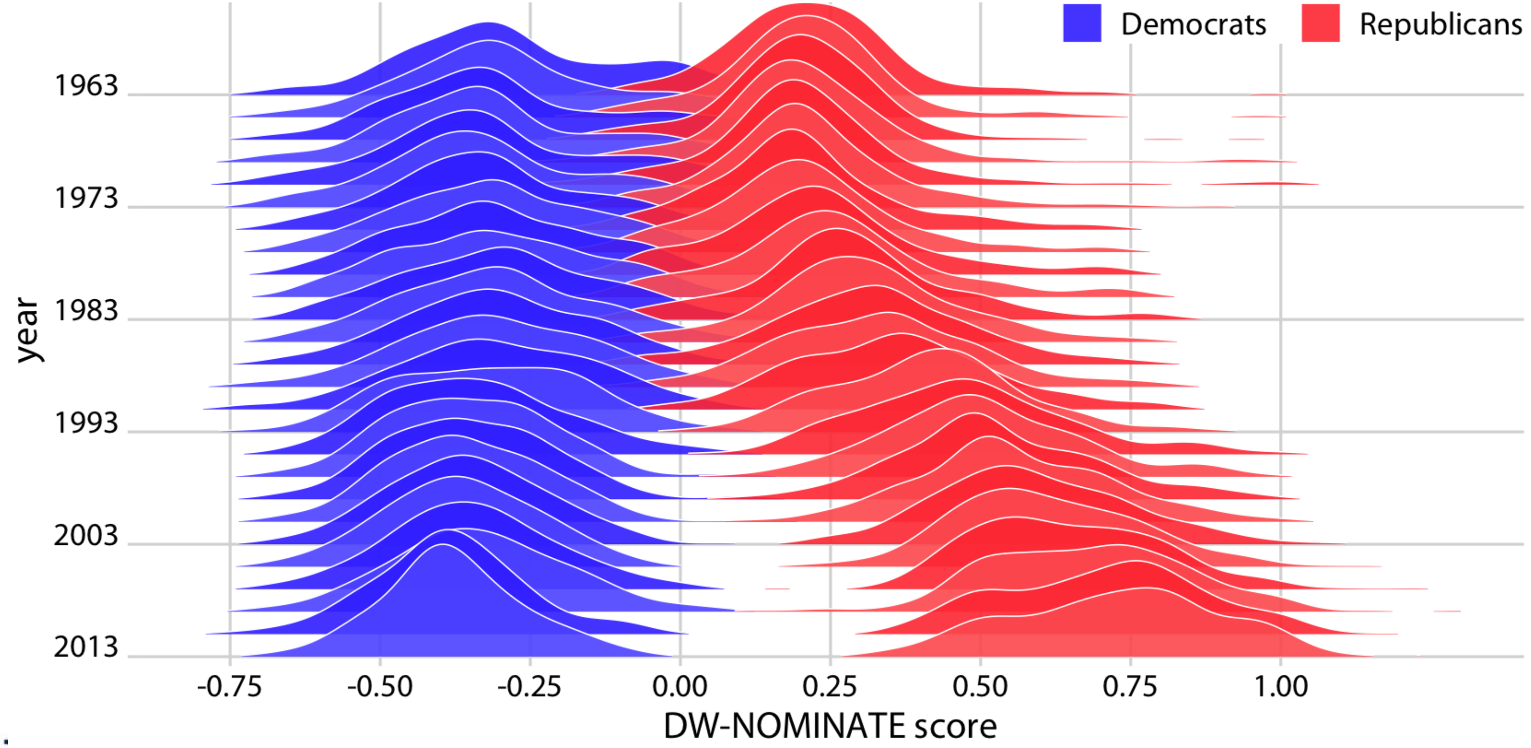
Ridgeline plot for Very Large Numbers of Distributions

- Evolution of movie lengths over time. Since the 1960s, the majority of all movies have been approximately 90 minutes long. Data source: Internet Movie Database (IMDB).
-



Ridgeline plot for Comparing Two Trends

- Voting patterns in the US House of Representatives have become increasingly polarized. DW-NOMINATE scores are frequently used to compare voting patterns of representatives between parties and over time.


🟦 Democrats
🟥 Republicans
year
DW-NOMINATE score