Chapter 9

Fundamentals of Data Visualization

April 28, 2023

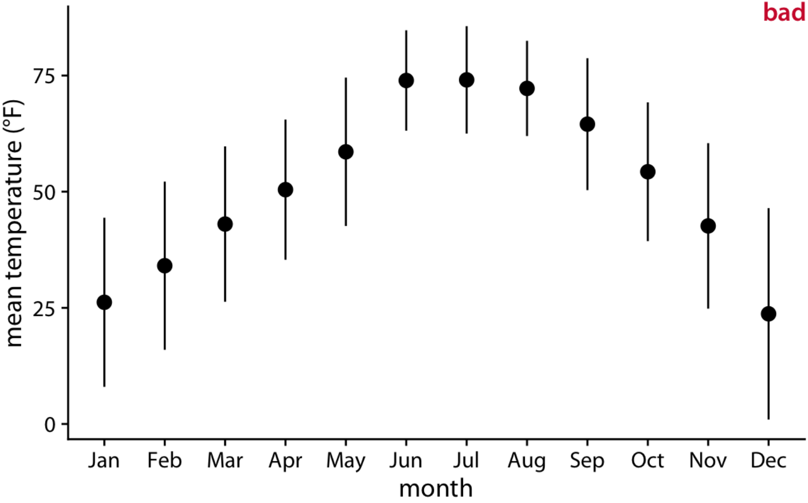
Visualizing many distributions at once

- We may want to visualize how temperature varies across different months while also showing the distribution of observed temperatures within each month.
- This scenario requires showing twelve temperature distributions at once, one for each month.
- Viable approaches include boxplots, violin plots, and ridgeline plots.

Visualizing many distributions at once

- The **response variable** is the variable whose distributions we want to show.
- The **grouping variables** define subsets of the data with distinct distributions of the response variable.
- For example, for temperature distributions across months, the response variable is the temperature and the grouping variable is the month.
- All techniques discussed in this chapter draw the response variable along one axis and the grouping variables along the other.

Points and error bars



bad

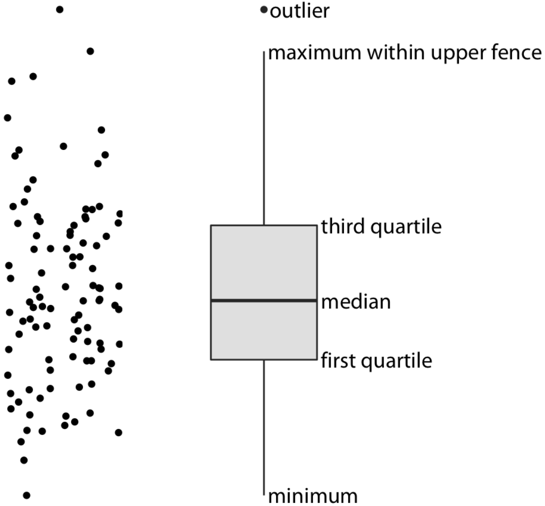
What’s wrong with points and error bars

- First, by representing each distribution by only one point and two error bars, we are losing a lot of information about the data.
- Second, it is not immediately obvious what the points represent, even though most readers would likely guess that they represent either the mean or the median.
- Third, it is definitely not obvious what the error bars represent.
- Do they represent the standard deviation of the data, the standard error of the mean, a 95% confidence interval, or something else altogether?
- There is no commonly accepted standard.

What’s wrong with points and error bars

- In the figure, they represent twice the standard deviation of the daily mean temperatures, meant to indicate the range that contains approximately 95% of the data.
- However, error bars are more commonly employed to visualize the standard error (or twice the standard error for a 95% confidence interval), and it is easy for readers to confuse the standard error with the standard deviation.
- The standard error quantifies how accurate our estimate of the mean is, whereas the standard deviation estimates how much spread there is in the data around the mean.
- It is possible for a dataset to have both a very small standard error of the mean and a very large standard deviation.
- Fourth, symmetric error bars are misleading if there is any skew in the data, which is the case here and almost always for real-world datasets.

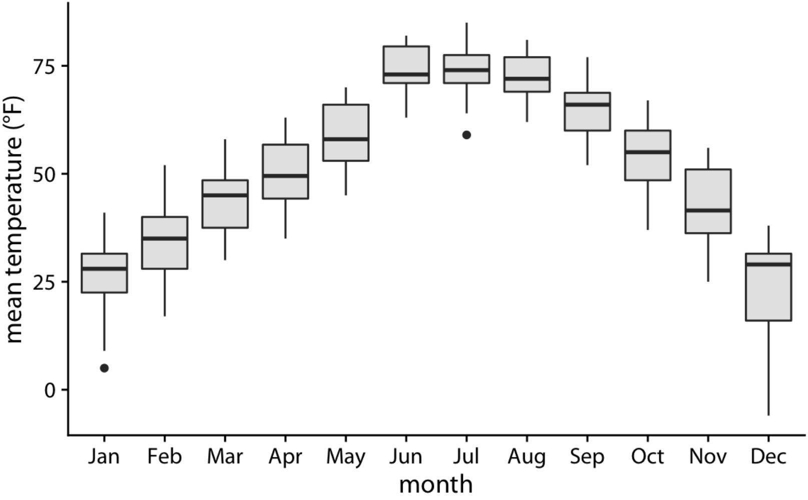
Boxplots


outlier
maximum within upper fence
third quartile
median
first quartile
minimum

Boxplots

- Shown are a cloud of points (left) and the corresponding boxplot (right).
- Only the y values of the points are visualized in the boxplot.
- The line in the middle of the boxplot represents the median, and the box encloses the middle 50% of the data.
- The top and bottom whiskers extend either to the maximum and minimum of the data or to the maximum or minimum that falls within 1.5 times the height of the box, whichever yields the shorter whisker.
- The distances of 1.5 times the height of the box in either direction are called the upper and the lower fences.
- Individual data points that fall beyond the fences are referred to as outliers and are usually showns as individual dots.

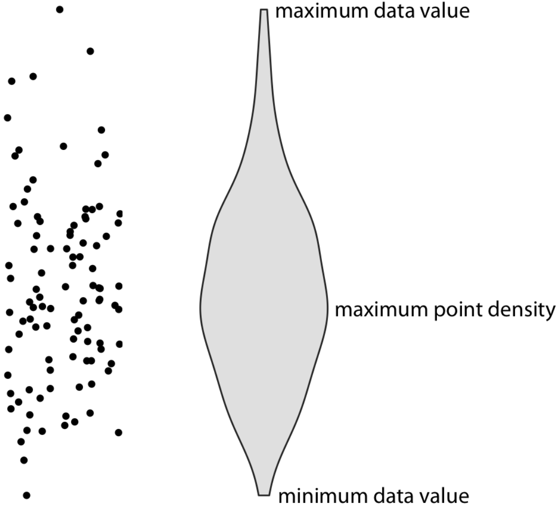
Boxplots work well plotted next to each other



Boxplots

- Boxplots were invented by the statistician John Tukey in the early 1970s, and they quickly gained popularity because they were highly informative while being easy to draw by hand.
- Most data visualizations were drawn by hand at that time.
- However, with modern computing and visualization capabilities, we are not limited to what is easily drawn by hand.

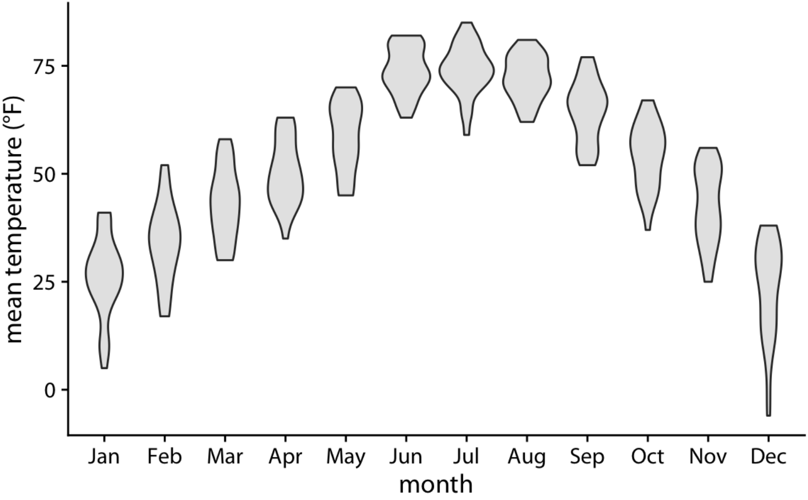
Violin plots



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.
- Violins are therefore symmetric.
- Violins begin and end at the minimum and maximum data values, respectively.
- The thickest part of the violin corresponds to the highest point density in the dataset.
- 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.

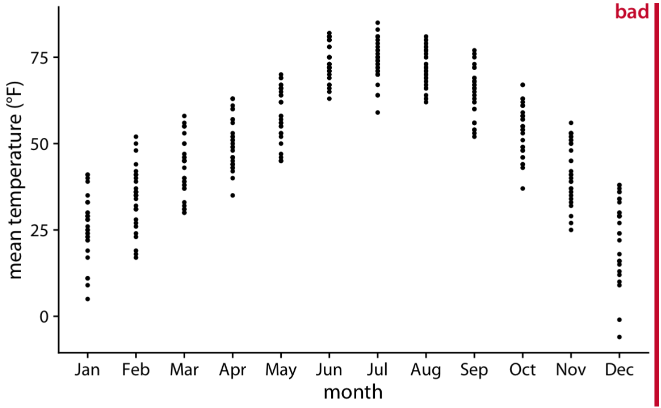
Violin plots



Violin plots

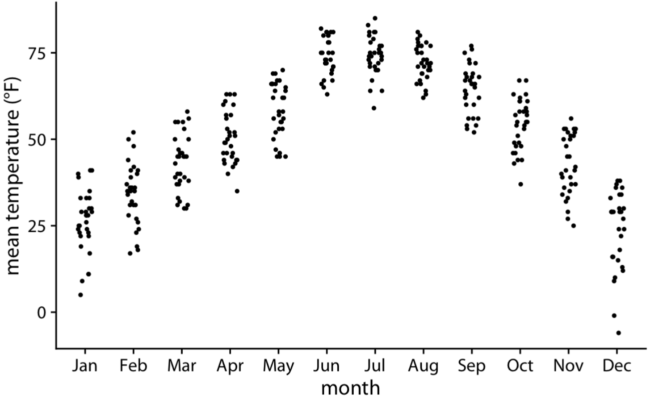
- We can see several bimodal distributions.
- But:
- violin plots, like density estimates, show data where there is none.

Showing the data itself



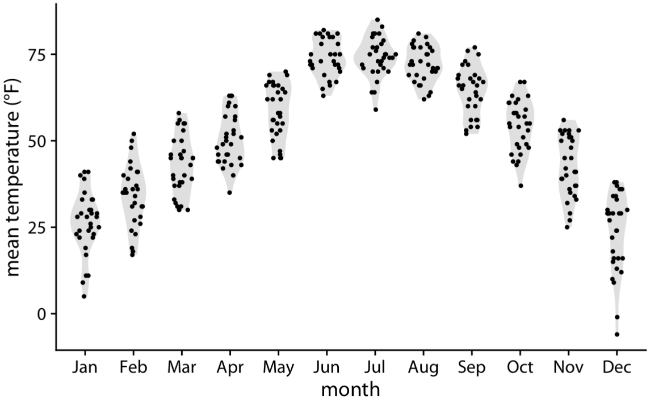
- Many points plotted on top of each other.

Jittered data



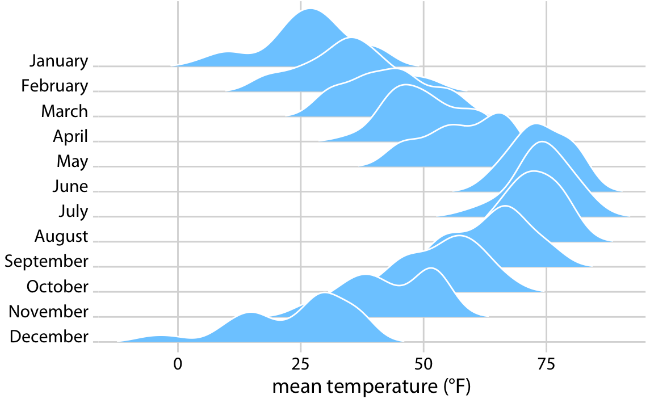
- Add random noise in the x direction.
- Whenever the dataset is too sparse to justify the violin visualization, plotting the raw data as individual points will be possible.

Best of both worlds: sina plots



- Jittering is proportional to the density.
- Sina plot drawn on top of the violins to show relation.

Ridgeline plot

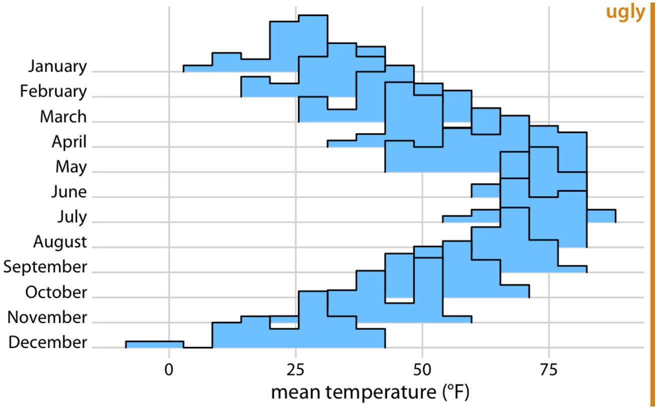


- Used to show trends in distributions over time.
- A violin plot on its side, but frequently easier to understand.
- Bimodal November is easier to see than in the violin plot.

Ridgeline plot

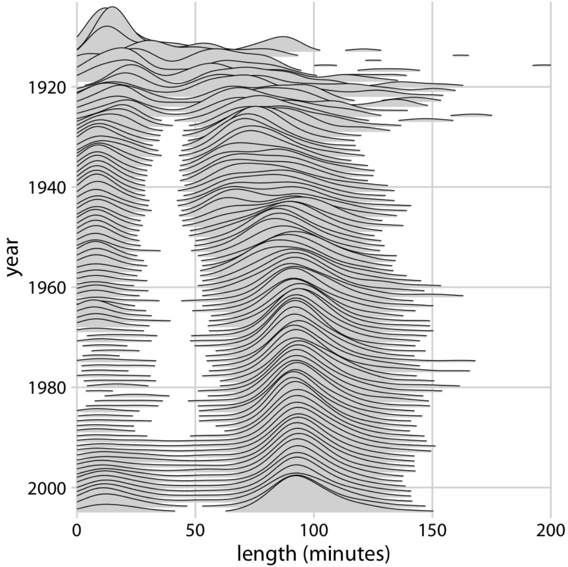
- Because the x axis shows the response variable and the y axis shows the grouping variable, there is no separate axis for the density estimates in a ridgeline plot.
- Density estimates are shown alongside the grouping variable.
- This is no different from the violin plot.
- The purpose of the plot is to allow for easy comparison of density shapes and relative heights across groups.

Histograms don’t work as well

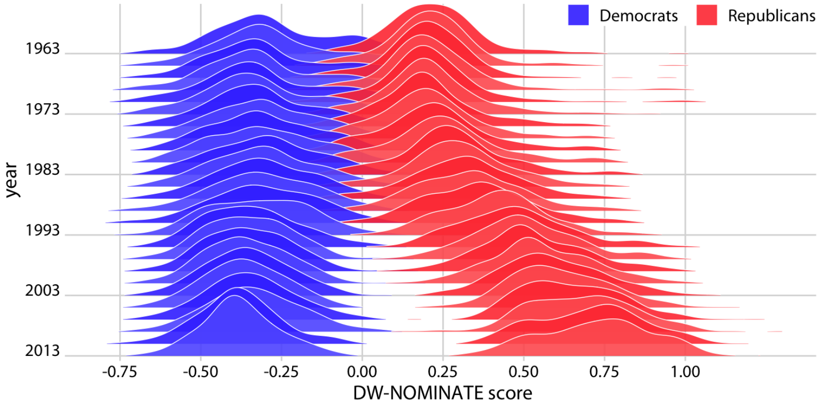


- Because the vertical lines in these ridgeline histograms appear always at the exact same x values, the bars from different histograms align with each other in confusing ways.

Ridgeline plots scale to many distributions



Contrast two distributions over time


Democrats
Republicans
year
DW-NOMINATE score