1. Statistical plotting

In the previous lesson, you saw how to create histograms that compare distributions of data. How can we make these comparisons more formal? Statistical plotting is a set of methods for using visualization to make comparisons. Here, we'll look at two of these techniques.

2. Adding error bars to bar charts

The first is the use of error bars in plots. These are additional markers on a plot or bar chart that tell us something about the distribution of the data. Histograms, that you have seen in the previous lesson, show the entire distribution. Error bars instead summarize the distribution of the data in one number, such as the standard deviation of the values. To demonstrate this, we'll use the data about heights of medalists in the 2016 Olympic Games. There are at least two different ways to display error bars. Here, we add the error bar as an argument to a bar chart. Each call to the ax-dot-bar method takes an x argument and a y argument. In this case, y is the mean of the "Height" column. The yerr key-word argument takes an additional number. In this case, the standard deviation of the "Height" column, and displays that as an additional vertical marker.

3. Error bars in a bar chart

Here is the plot. It is helpful because it summarizes the full distribution that you saw in the histograms in two numbers: the mean value, and the spread of values, quantified as the standard deviation.

4. Adding error bars to plots

We can also add error bars to a line plot. For example, let's look at the weather data that we used in the first chapter of this course. To plot this data with error bars, we will use the Axes errorbar method. Like the plot method, this method takes a sequence of x values, in this case, the "MONTH" column, and a sequence of y values, in this case, the column with the normal average monthly temperatures. In addition, a yerr key-word argument can take the column in the data that contains the standard deviations of the average monthly temperatures.

5. Error bars in plots

Similar to before, this adds vertical markers to the plot, which look like this.

6. Adding boxplots

The second statistical visualization technique we will look at is the boxplot, a visualization technique invented by John Tukey, arguably the first data scientist. It is implemented as a method of the Axes object. We can call it with a sequence of sequences. In this case, we create a list with the men's rowing "Height" column and the men's gymnastics "Height" column and pass that list to the method. Because the box-plot doesn't know the labels on each of the variables, we add that separately, labeling the y-axis as well. Finally, we show the figure, which looks

7. Interpreting boxplots

like this. This kind of plot shows us several landmarks in each distribution. The red line indicates the median height. The edges of the box portion at the center indicate the inter-quartile range of the data, between the 25th and the 75th percentiles. The whiskers at the ends of the thin bars indicate one and a half times the size of the inter-quartile range beyond the 75th and 25th percentiles. This should encompass roughly 99 percent of the distribution if the data is Gaussian or normal. Points that appear beyond the whiskers are outliers. That means that they have values larger or smaller than what you would expect for 99 percent of the data in a Gaussian or normal distribution. For example, there are three unusually short rowers in this sample, and one unusually high gymnast.