**What is statistics?**

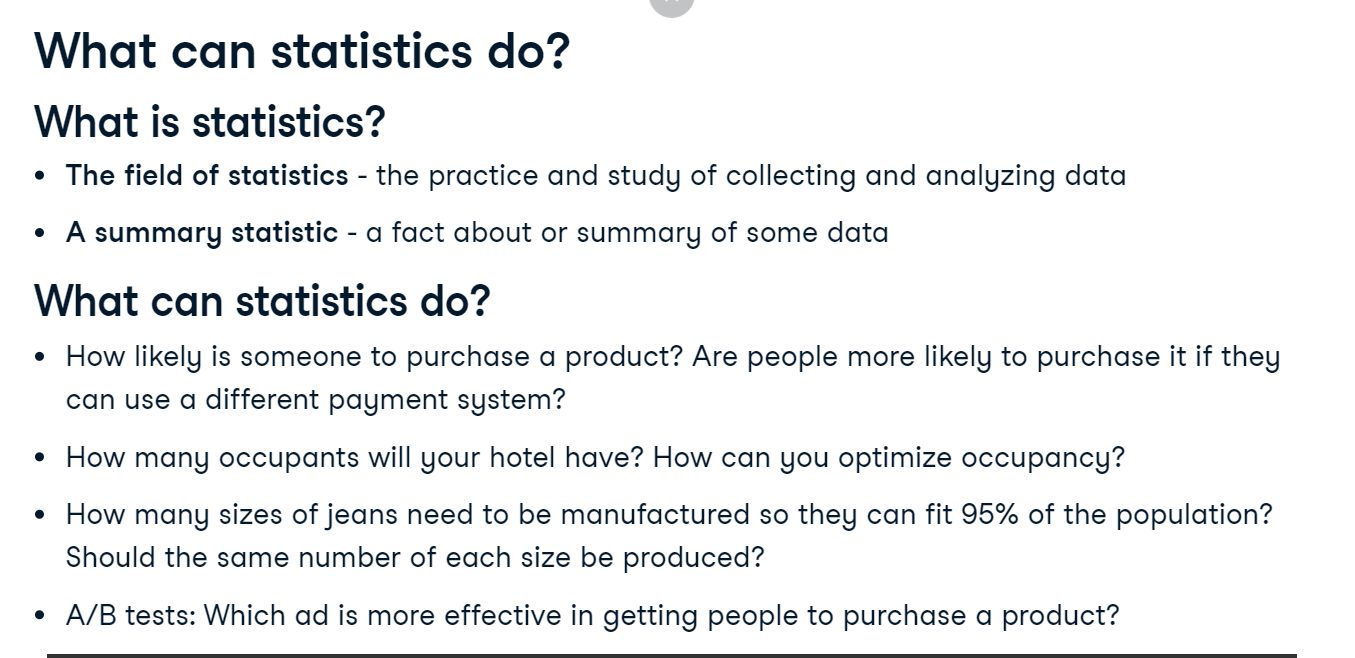
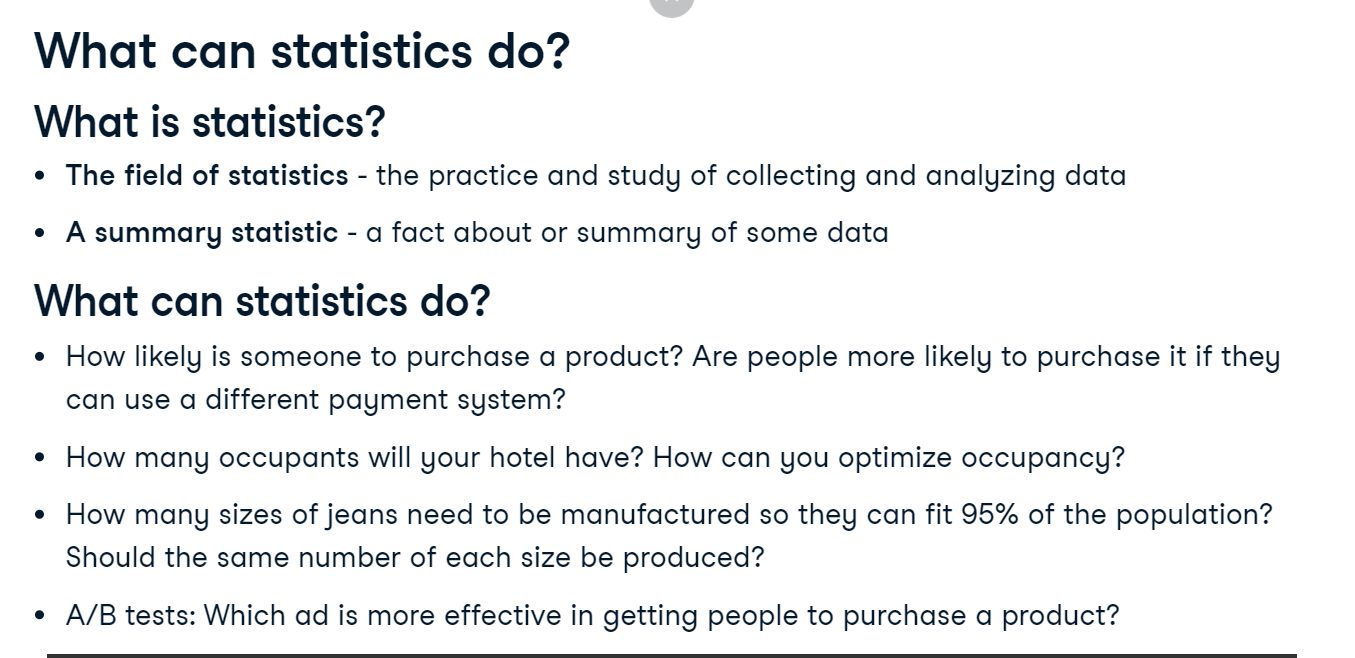
Hi and welcome to the course. My name is Maggie, and I'll be your host as we dive in to the world of statistics.

**What is statistics?**

So what is statistics anyway? We can talk about the field of statistics, which is the practice and study of collecting and analyzing data. We can also talk about a summary statistic, which is a fact about or summary of some data, like an average or a count.

**What can statistics do?**

A more important question, however, is what can statistics do? With the power of statistics, we can answer tons of different questions like: How likely is someone to purchase a product? Are people more likely to purchase it if they can use a different payment system? How many occupants will your hotel have? How can you optimize occupancy? How many sizes of jeans need to be manufactured so they can fit 95% of the population? Should the same number of each size be produced? A question like, Which ad is more effective in getting people to purchase a product? can be answered with A/B testing.

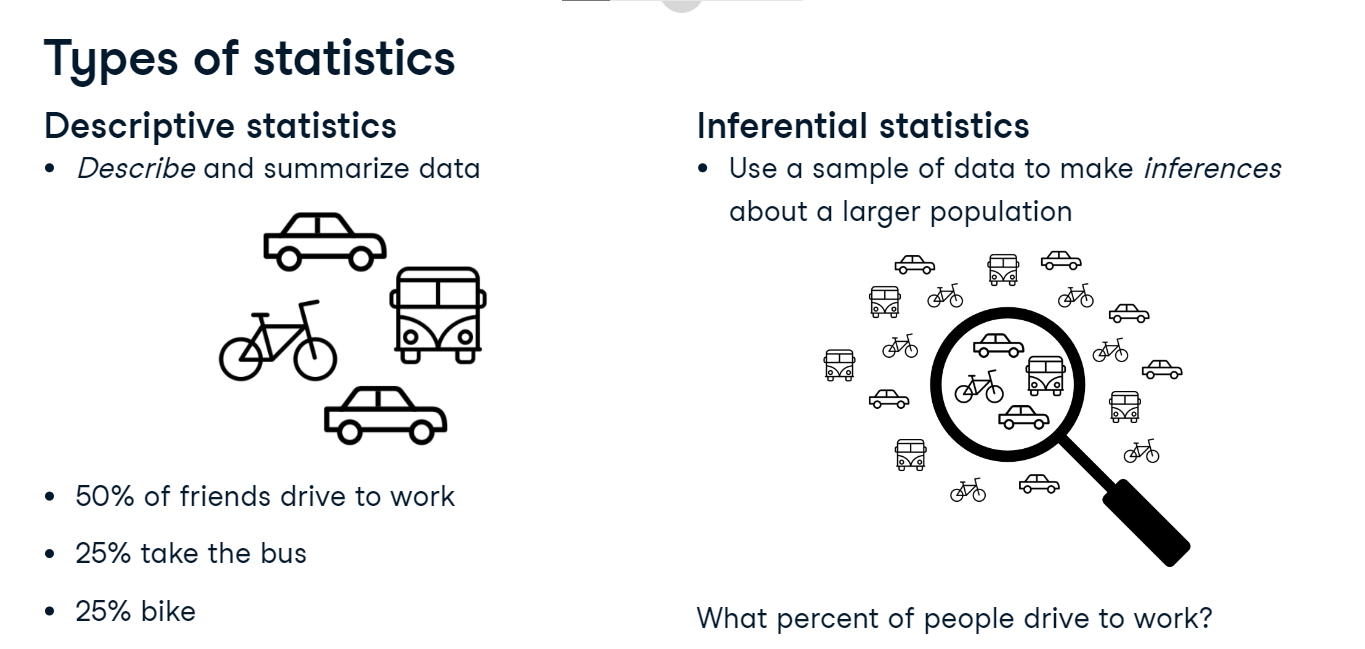


**What can't statistics do?**

While statistics can answer a lot of questions, it's important to note that statistics can't answer every question. If we want to know why the TV series Game of Thrones is so popular, we could ask everyone why they like it, but they may lie or leave out reasons. We can see if series with more violent scenes attract more viewers, but even if they do, we can't know if the violence in Game of Thrones is the reason for its popularity, or if other factors are driving its popularity and it just happens to be violent.

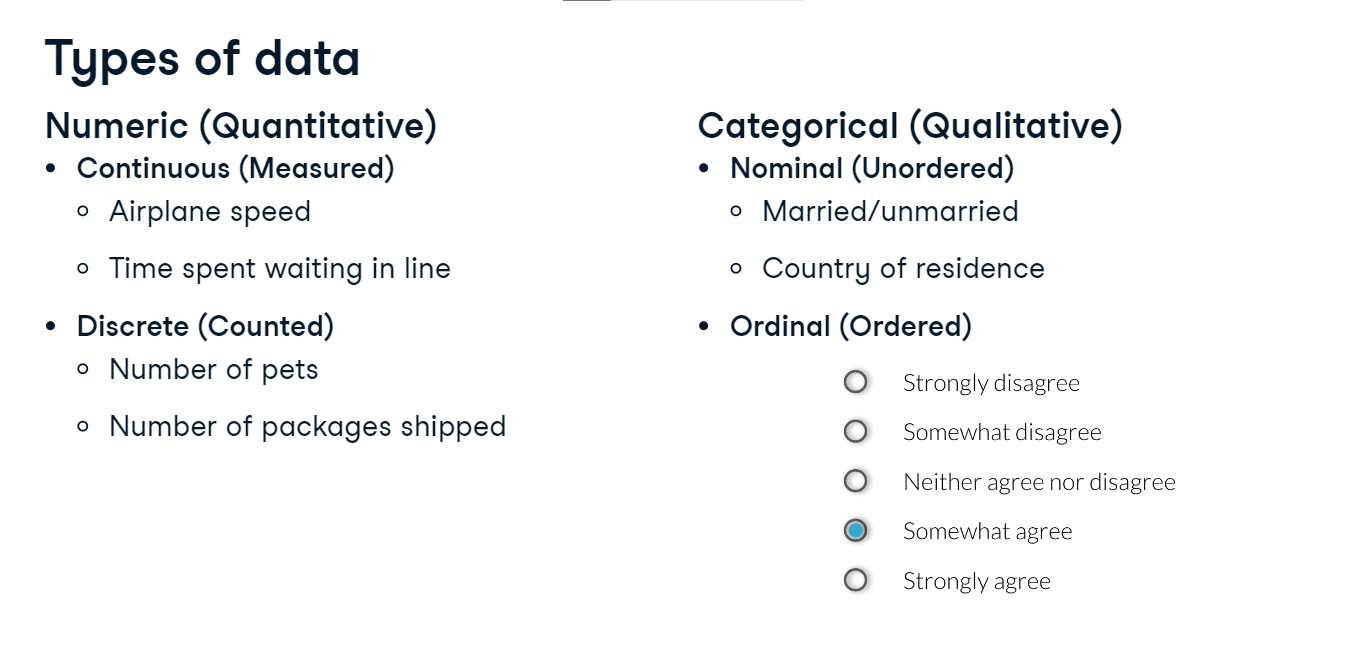
**Types of statistics**

There are 2 main branches of statistics: descriptive statistics and inferential statistics. Descriptive statistics focuses on describing and summarizing the data at hand. After asking four friends how they get to work, we can see that 50% of them drive to work, 25% ride the bus, and 25% bike. These are examples of descriptive statistics. Inferential statistics uses the data at hand, which is called sample data, to make inferences about a larger population. We could use inferential statistics to figure out what percent of people drive to work based on our sample data.

****

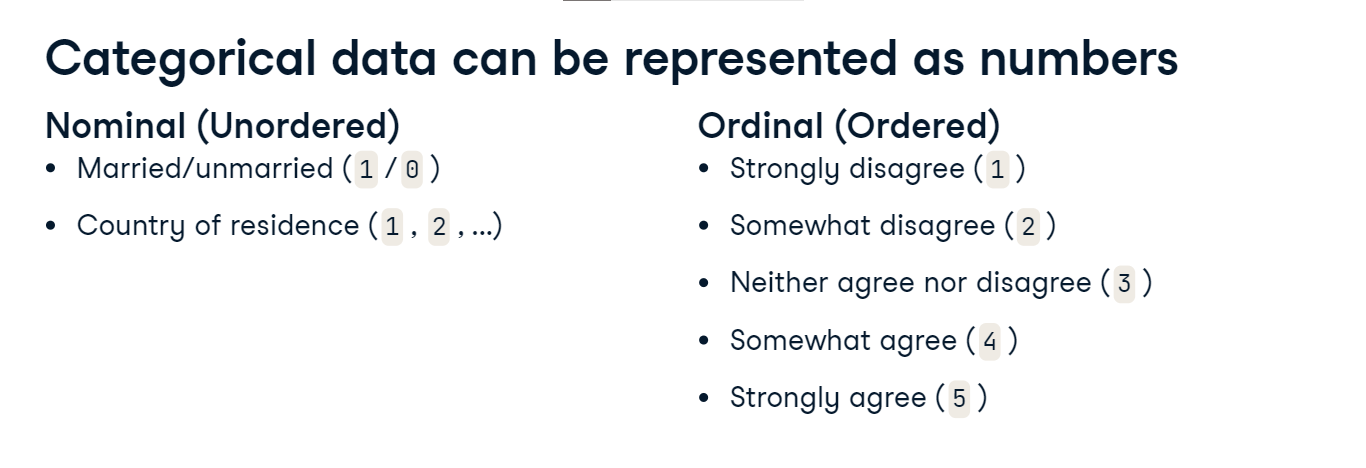
**Types of data**

There are two main types of data. Numeric, or quantitative data is made up of numeric values. Categorical, or qualitative data is made up of values that belong to distinct groups. It's important to note that these aren't the only two types of data that exist - there are others too, but we'll be focusing on these two. Numeric data can be further separated into continuous and discrete data. Continuous numeric data is often quantities that can be measured, like speed or time. Discrete numeric data is usually count data, like number of pets or number of packages shipped. Categorical data can be nominal or ordinal. Nominal categorical data is made up of categories with no inherent ordering, like marriage status or country of residence. Ordinal categorical data has an inherent order, like a survey question where you need to indicate the degree to which you agree with a statement.



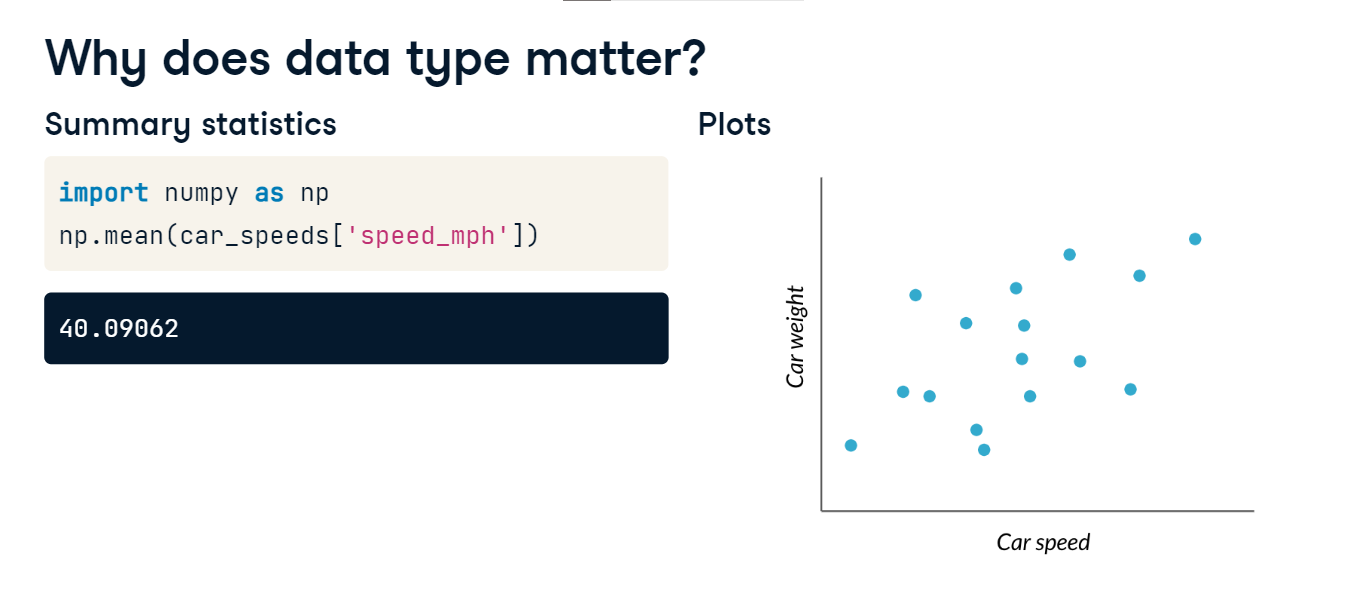
**Categorical data can be represented as numbers**

Sometimes, categorical variables are represented using numbers. Married and unmarried can be represented using 1 and 0, or an agreement scale could be represented with numbers 1 through 5. However, it's important to note that this doesn't necessarily make them numeric variables.



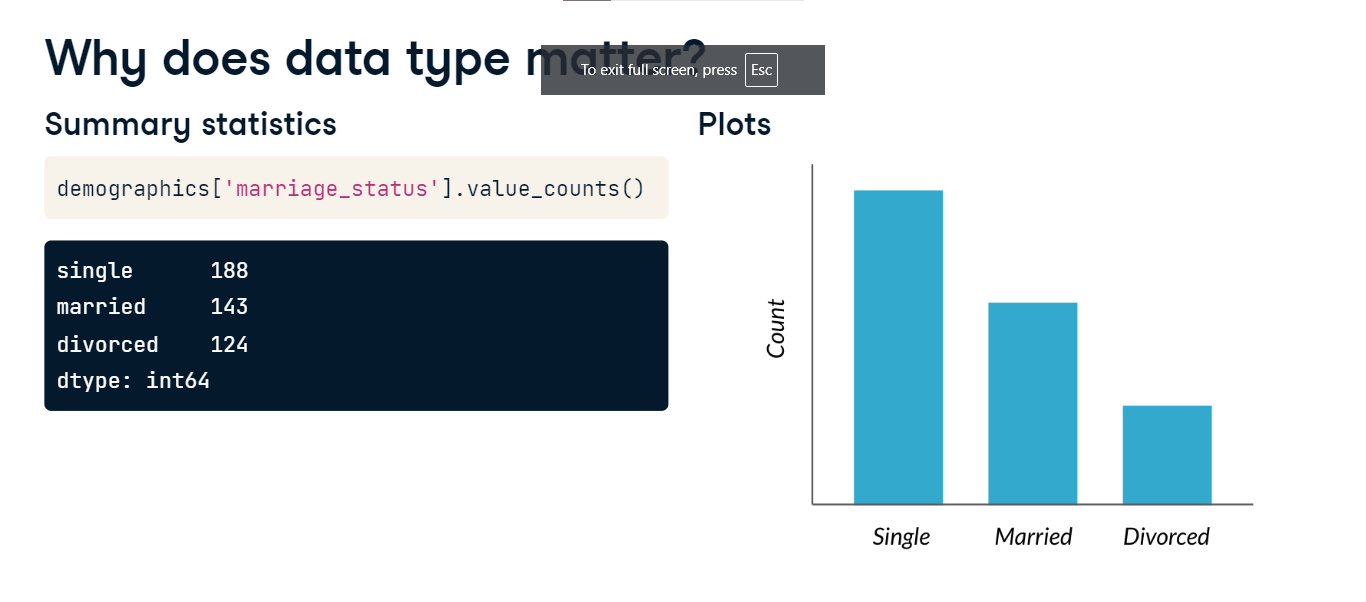
**Why does data type matter?**

Being able to identify data types is important since the type of data you're working with will dictate what kinds of summary statistics and visualizations make sense for your data, so this is an important skill to master. For numerical data, we can use summary statistics like mean, and plots like scatter plots, but these don't make a ton of sense for categorical data.



**Why does data type matter?**

Similarly, things like counts and barplots don't make much sense for numeric data.



**Measures of center**

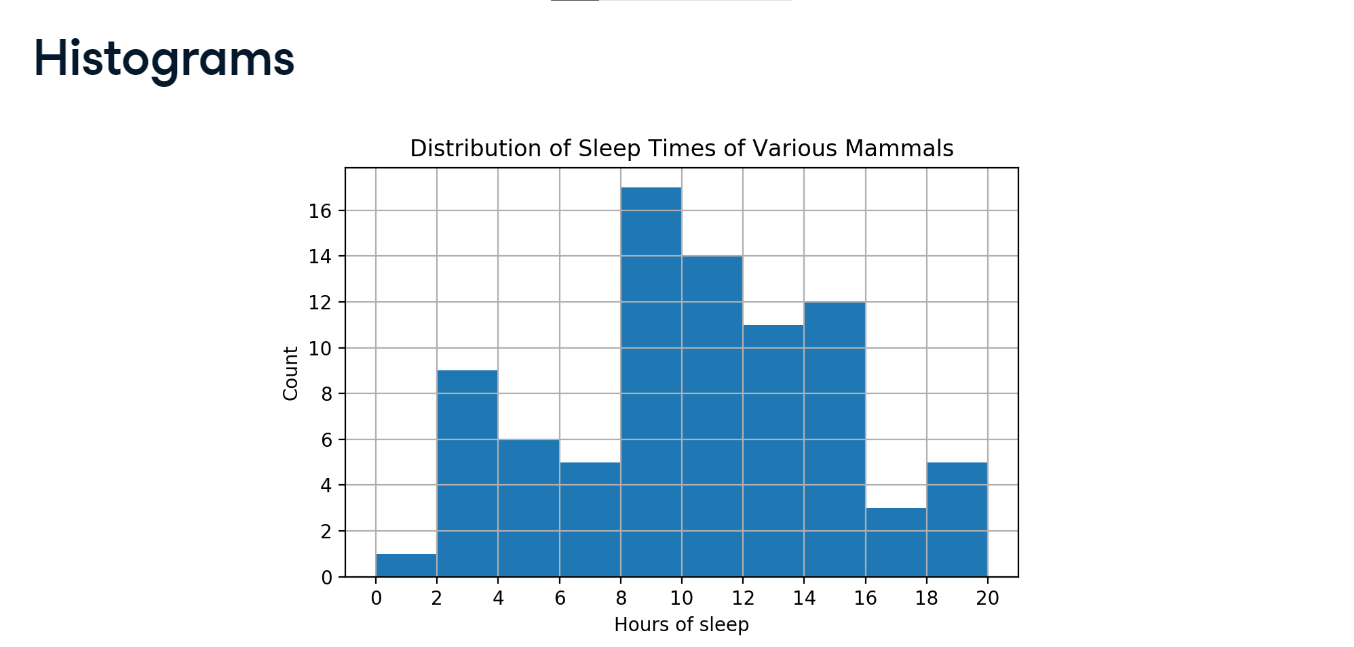
In this lesson, we'll begin to discuss summary statistics, some of which you may already be familiar with, like mean and median.

**Mammal sleep data**

In this video, we'll look at data about different mammals' sleep habits.

**Histograms**

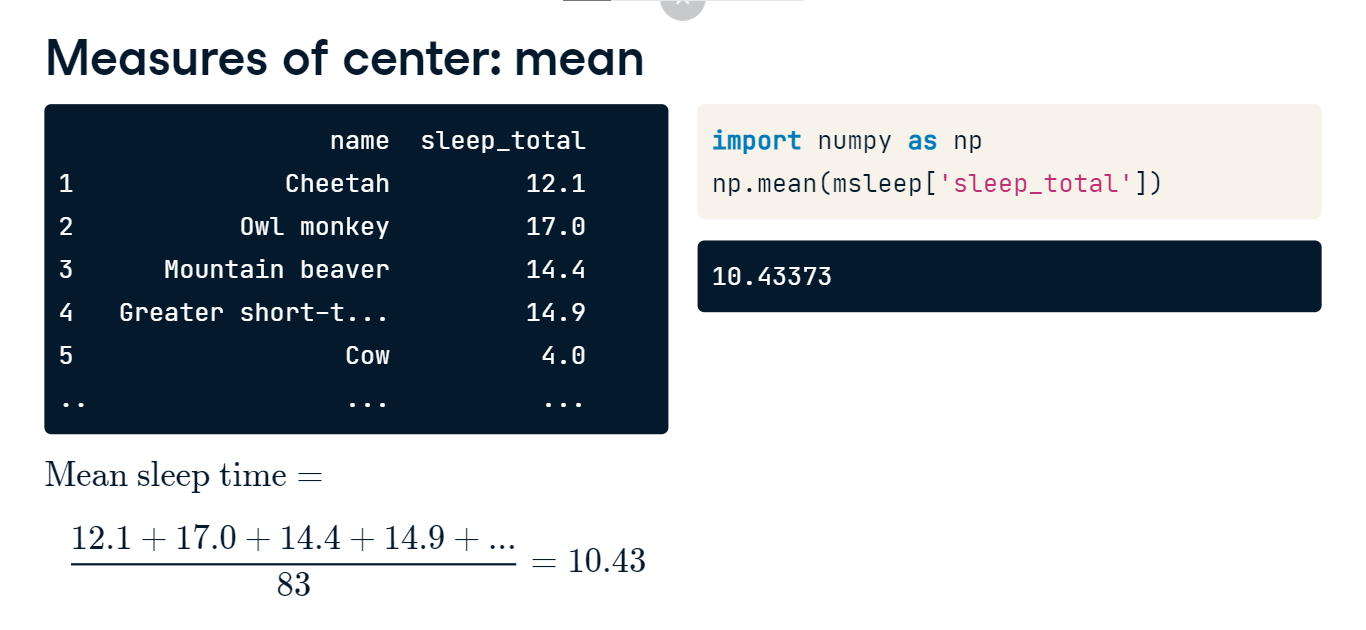
Before we dive in, let's remind ourselves how histograms work. A histogram takes a bunch of data points and separates them into bins, or ranges of values. Here, there's a bin for 0 to 2 hours, 2 to 4 hours, and so on. The heights of the bars represent the number of data points that fall into that bin, so there's one mammal in the dataset that sleeps between 0 to 2 hours, and nine mammals that sleep two to four hours. Histograms are a great way to visually summarize the data, but we can use numerical summary statistics to summarize even further.



**How long do mammals in this dataset typically sleep?**

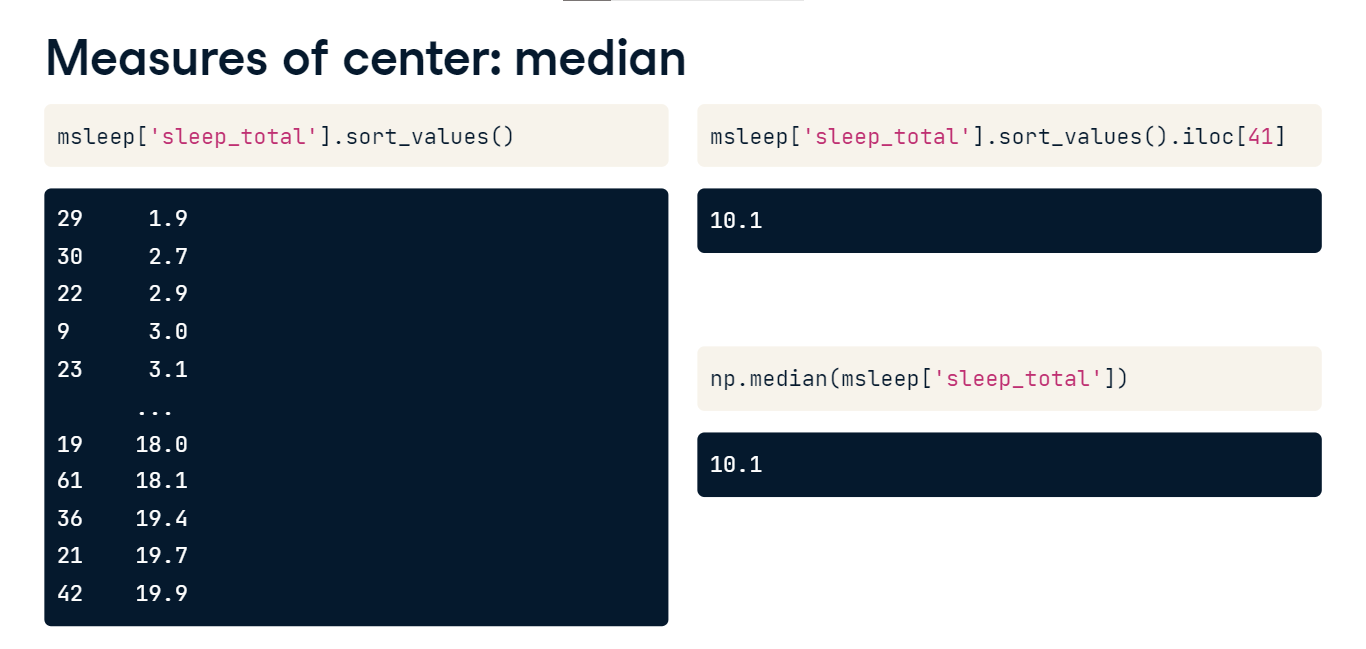
One way we could summarize the data is by answering the question, How long do mammals in this dataset typically sleep? To answer this, we need to figure out what the "typical" or "center" value of the data is. We'll discuss three different definitions, or measures, of center: mean, median, and mode.

**Measures of center: mean**

The mean, often called the average, is one of the most common ways of summarizing data. To calculate mean, we add up all the numbers of interest and divide by the total number of data points, which is 83 here. This gives us 10-point-43 hours of sleep. In Python, we can use numpy's mean function, passing it the variable of interest. 

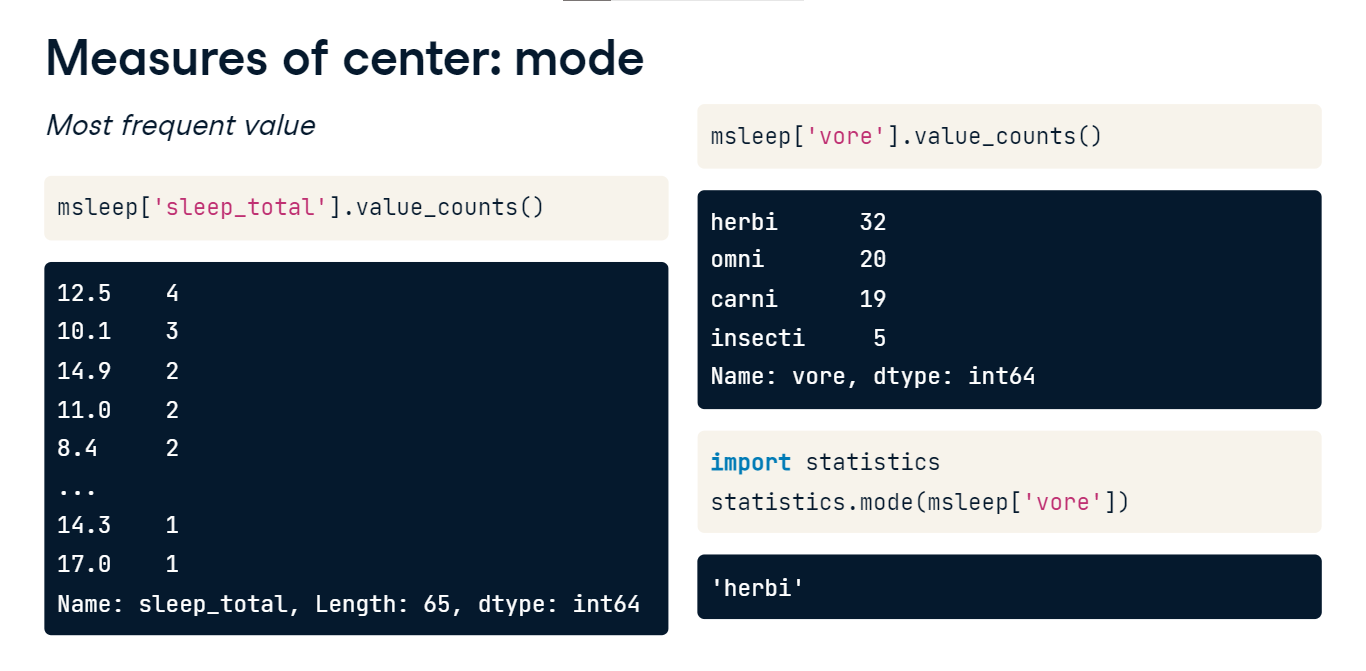
**Measures of center: median**

Another measure of center is the median. The median is the value where 50% of the data is lower than it, and 50% of the data is higher. We can calculate this by sorting all the data points and taking the middle one, which would be index 41 in this case. This gives us a median of 10-point-1 hours of sleep. In Python, we can use np-dot-median to do the calculations for us.



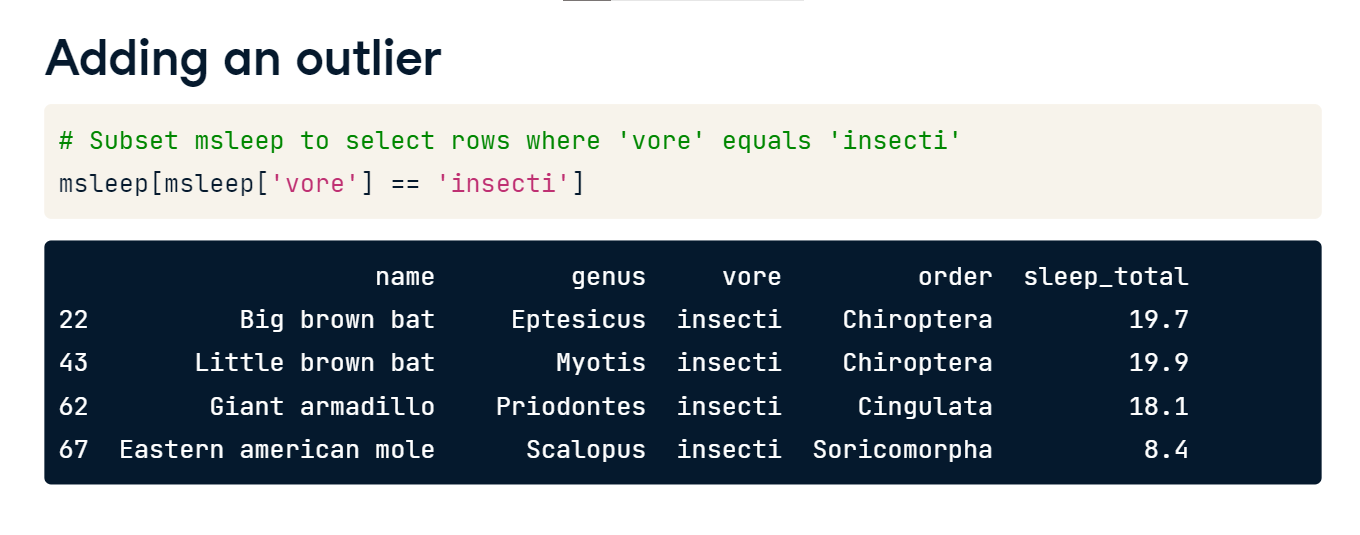
**Measures of center: mode**

The mode is the most frequent value in the data. If we count how many occurrences there are of each sleep\_total and sort in descending order, there are 4 mammals that sleep for 12.5 hours, so this is the mode. The mode of the vore variable, which indicates the animal's diet, is herbivore. We can also find the mode using the mode function from the statistics module. Mode is often used for categorical variables, since categorical variables can be unordered and often don't have an inherent numerical representation.



**Adding an outlier**

Now that we have lots of ways to measure center, how do we know which one to use? Let's look at an example. Here, we have all of the insectivores in the dataset.



**Adding an outlier**

We get a mean sleep time of 16-point-5 hours and a median sleep time of 18-point-9 hours.

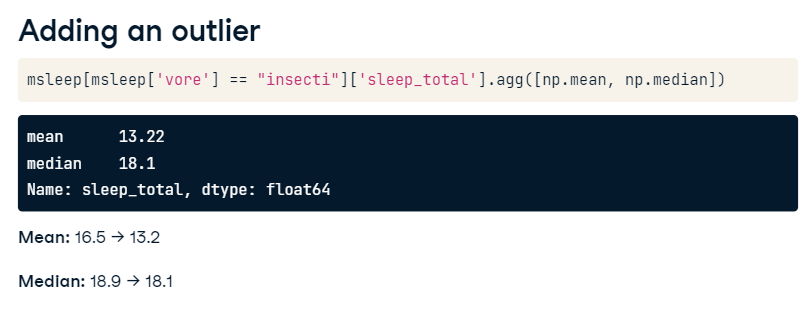


Now let's say we've discovered a new mystery insectivore that never sleeps.



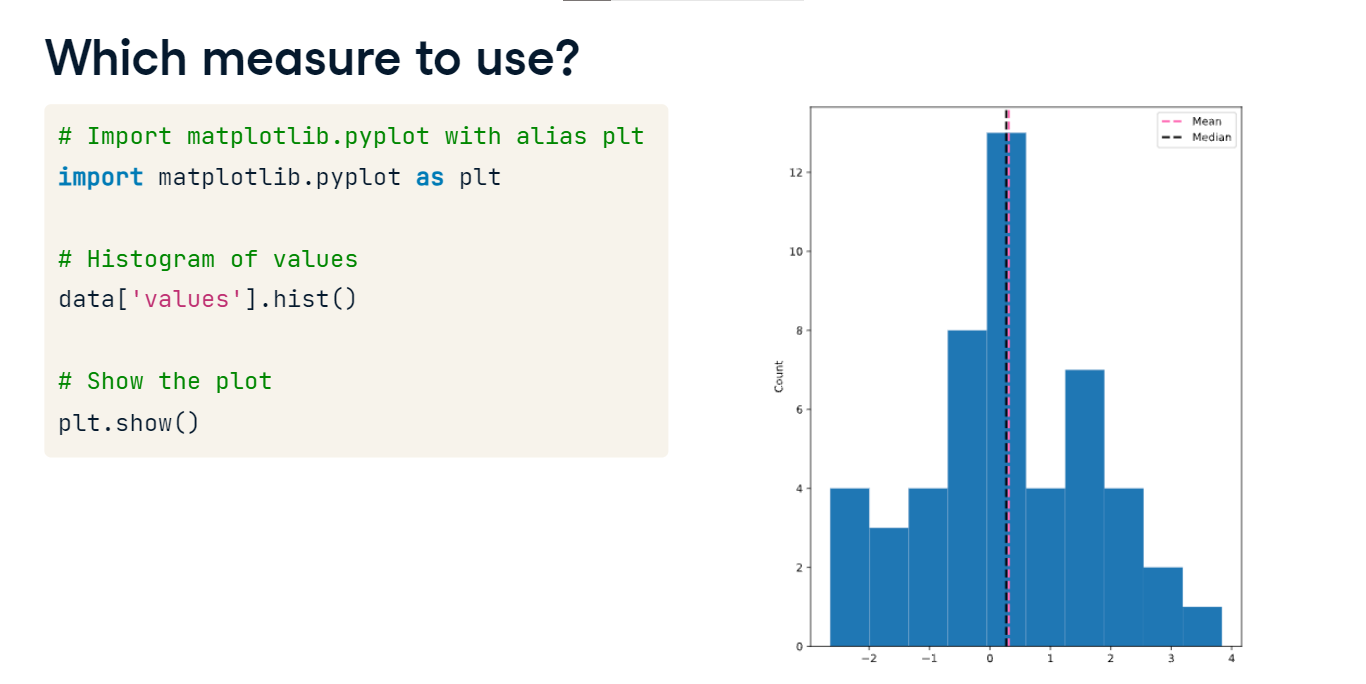
**Adding an outlier**

If we take the mean and median again, we get different results. The mean went down by more than 3 hours, while the median changed by less than an hour. This is because the mean is much more sensitive to extreme values than the median.



**Which measure to use?**

Since the mean is more sensitive to extreme values, it works better for symmetrical data like this. Notice that the mean, in black, and median, in red, are quite close.

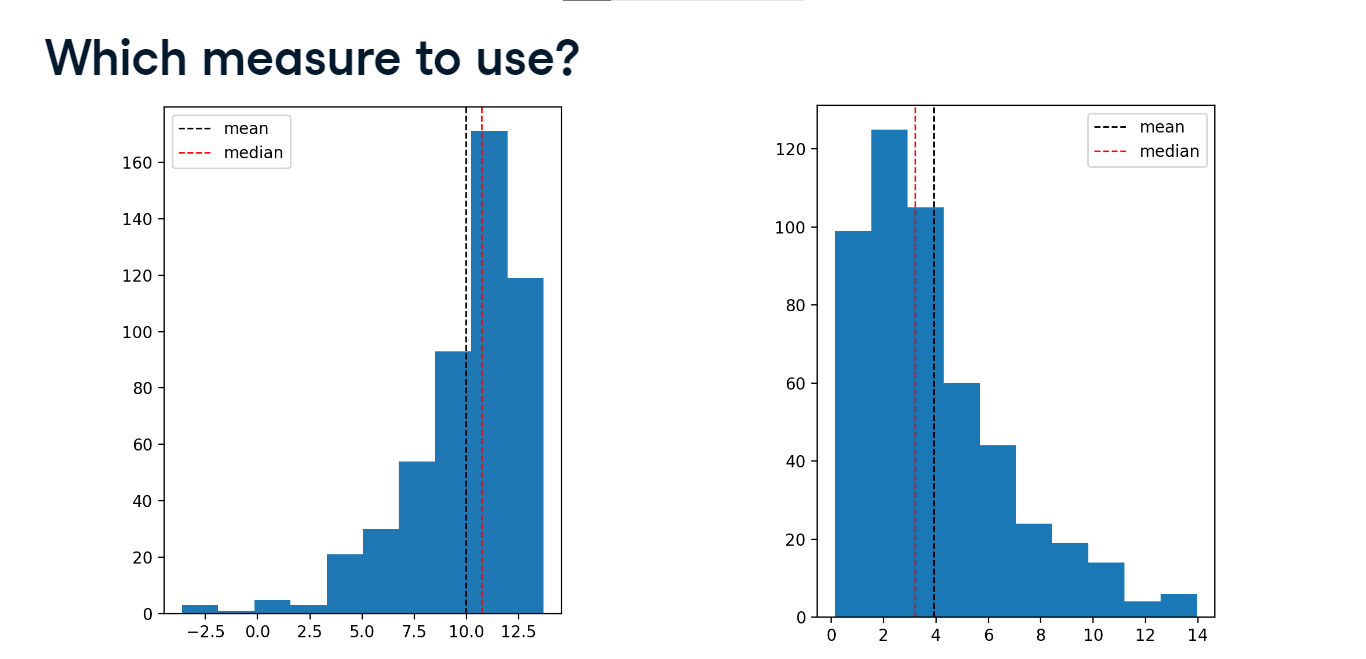


**Skew**

However, if the data is skewed, meaning it's not symmetrical, like this, median is usually better to use. In this histogram, the data is piled up on the right, with a tail on the left. Data that looks like this is called left-skewed data. When data is piled up on the left with a tail on the right, it's right-skewed.

**Which measure to use?**

When data is skewed, the mean and median are different. The mean is pulled in the direction of the skew, so it's lower than the median on the left-skewed data, and higher than the median on the right-skewed data. Because the mean is pulled around by the extreme values, it's better to use the median since it's less affected by outliers.

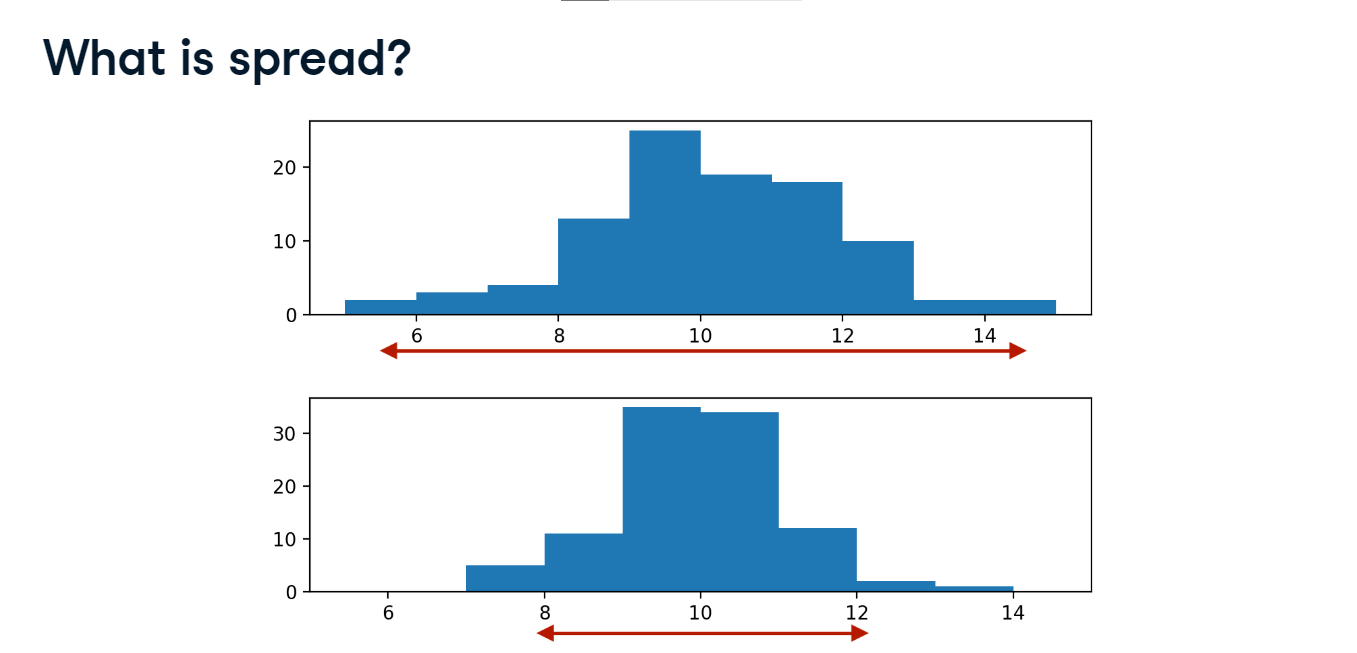


## Measures of spread

In this lesson, we'll talk about another set of summary statistics: measures of spread.

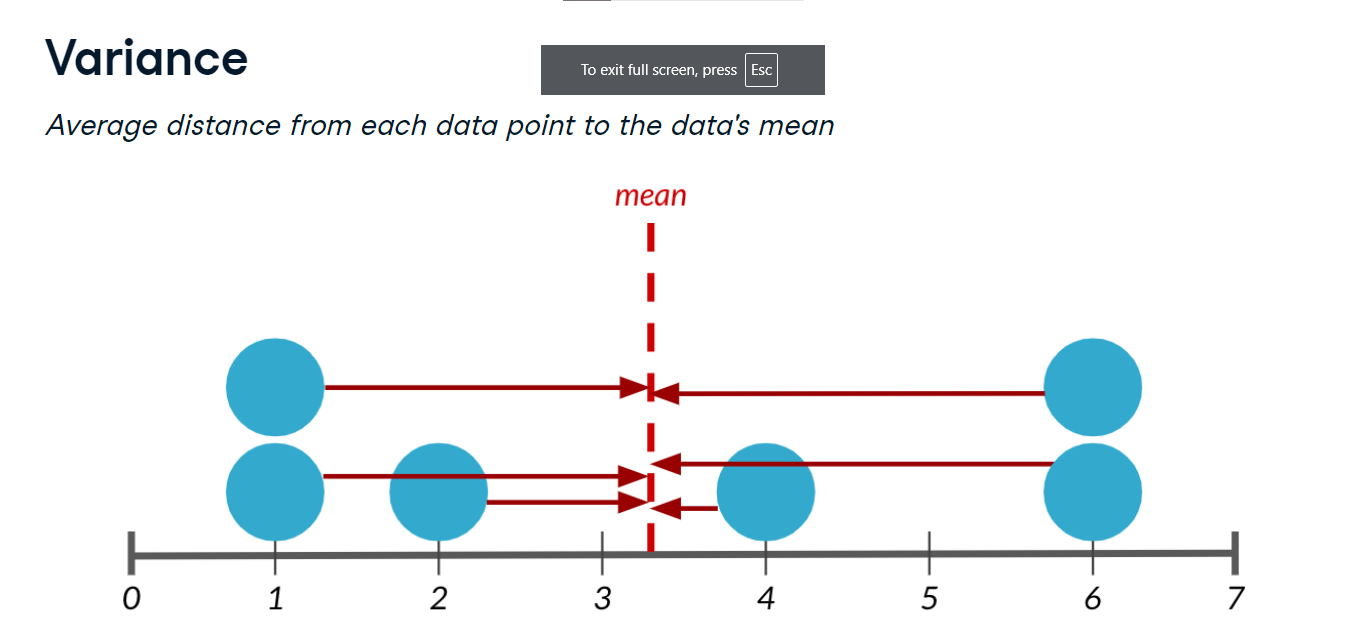
**What is spread?**

Spread is just what it sounds like - it describes how spread apart or close together the data points are. Just like measures of center, there are a few different measures of spread.



**Variance**

The first measure, variance, measures the average distance from each data point to the data's mean.

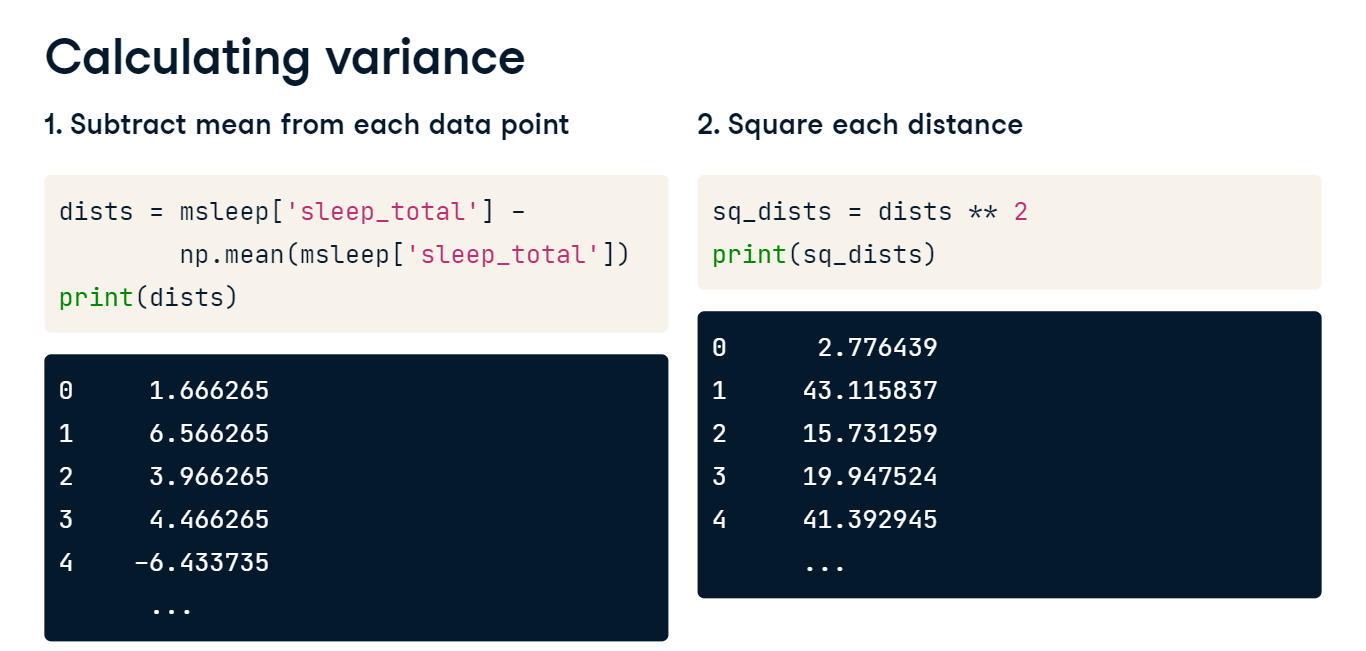


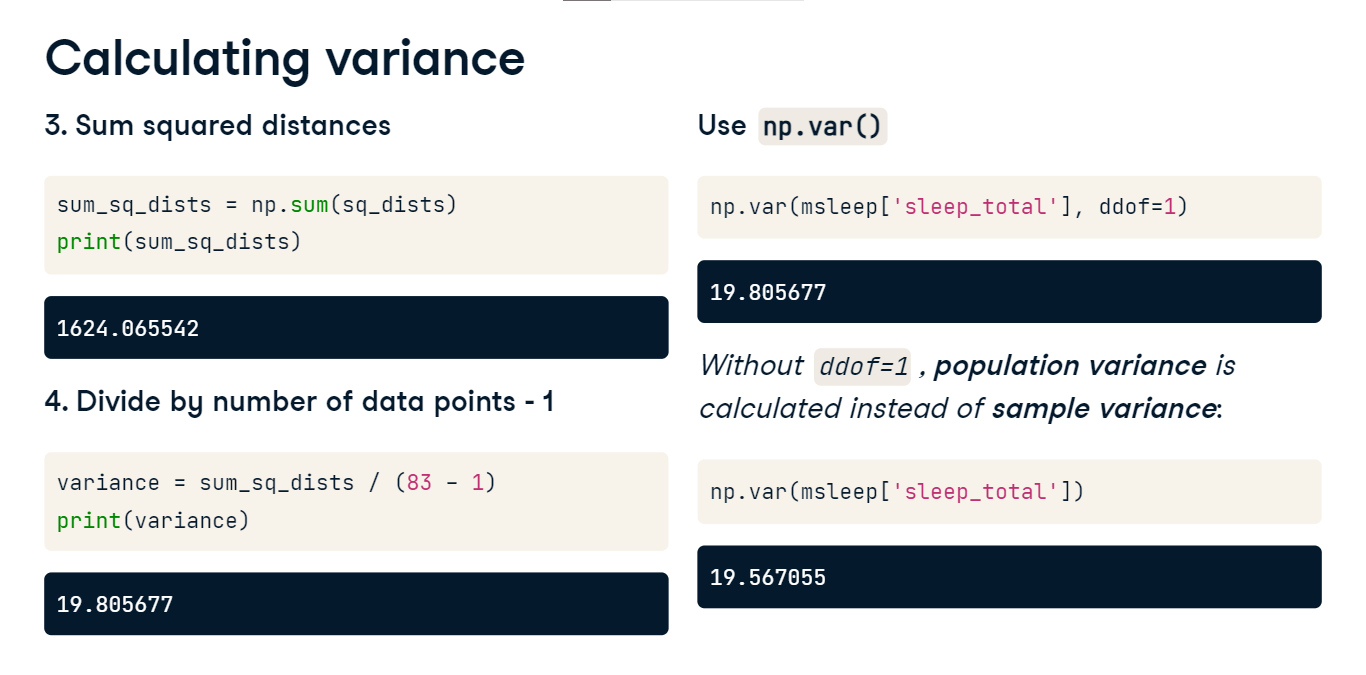
**Calculating variance**

To calculate the variance, we start by calculating the distance between each point and the mean, so we get one number for every data point. We then square each distance

**Calculating variance**

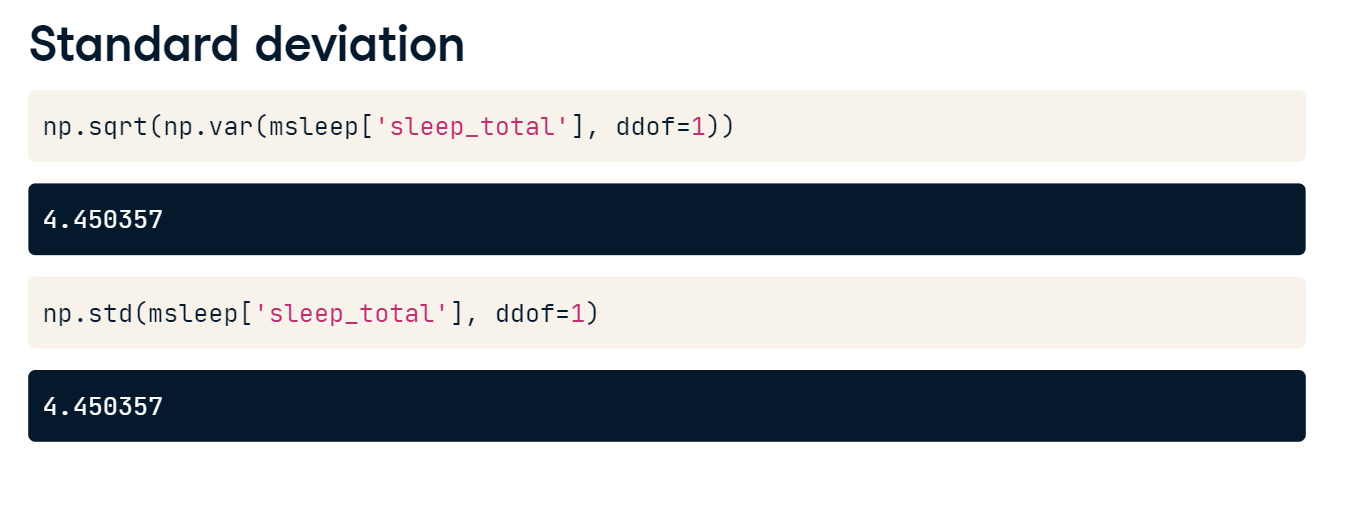
and then add them all together. Finally, we divide the sum of squared distances by the number of data points minus 1, giving us the variance. The higher the variance, the more spread out the data is. It's important to note that the units of variance are squared, so in this case, it's 19-point-8 hours squared. We can calculate the variance in one step using np-dot-var, setting the ddof argument to 1. If we don't specify ddof equals 1, a slightly different formula is used to calculate variance that should only be used on a full population, not a sample.



****

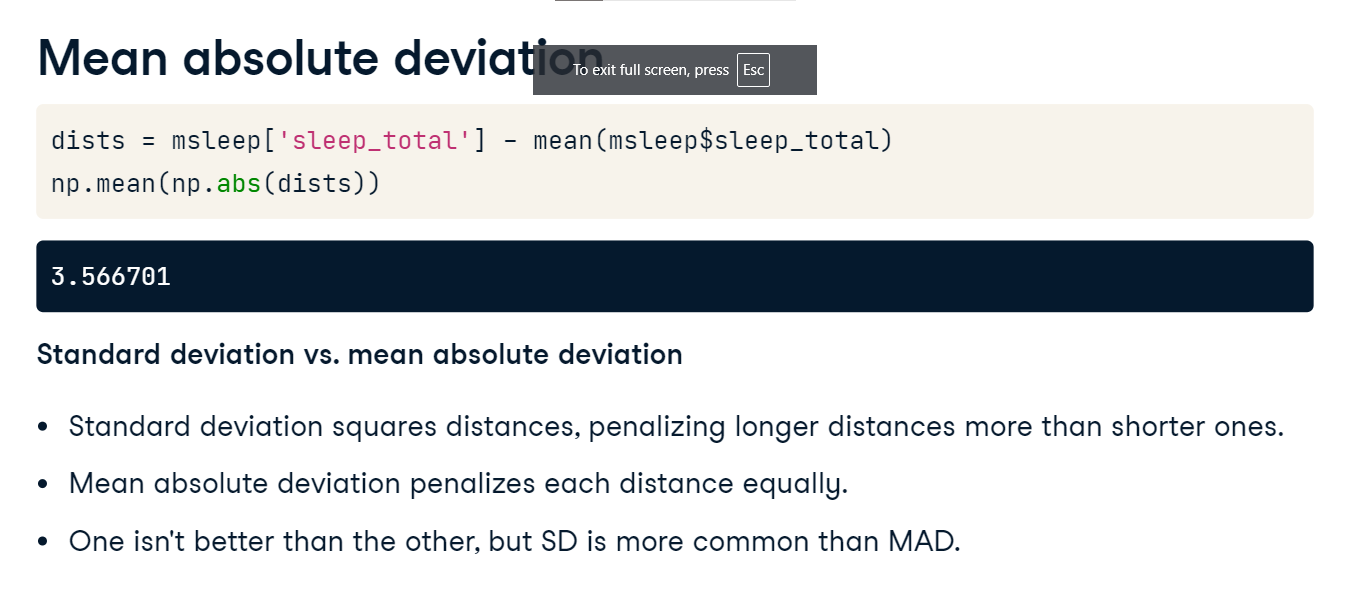
**Standard deviation**

The standard deviation is another measure of spread, calculated by taking the square root of the variance. It can be calculated using np-dot-std. Just like np-dot-var, we need to set ddof to 1. The nice thing about standard deviation is that the units are usually easier to understand since they're not squared. It's easier to wrap your head around 4 and a half hours than 19-point-8 hours squared.



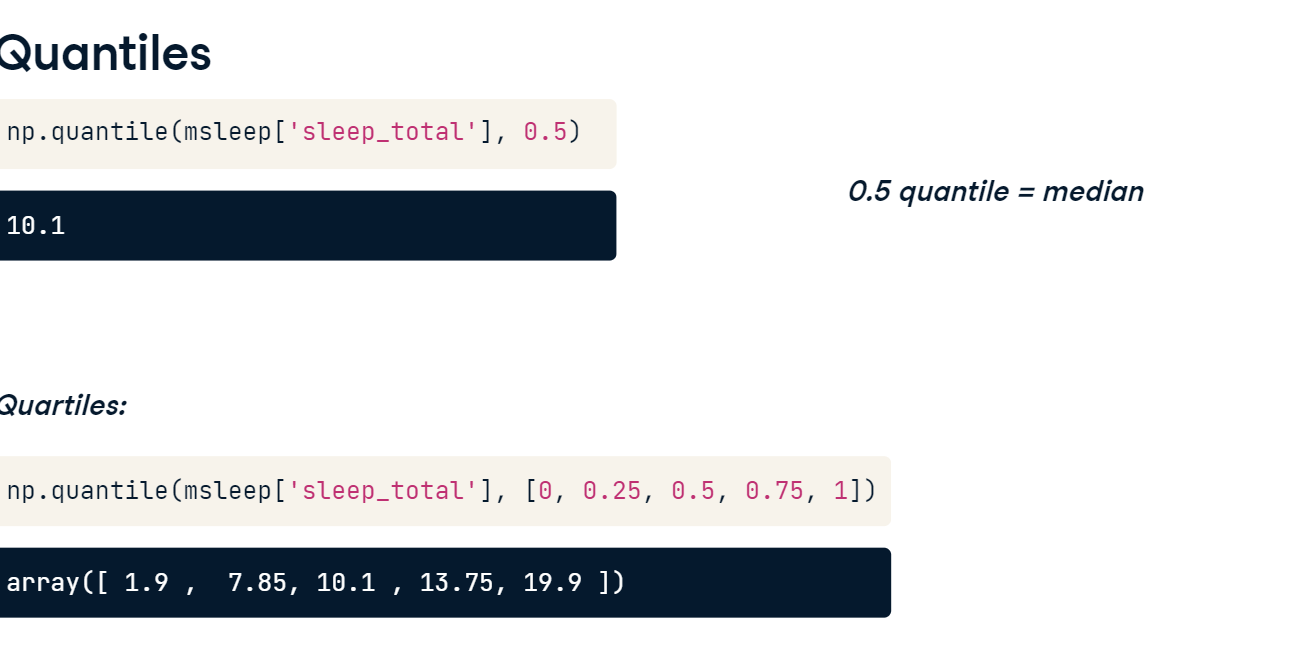
**Mean absolute deviation**

Mean absolute deviation takes the absolute value of the distances to the mean, and then takes the mean of those differences. While this is similar to standard deviation, it's not exactly the same. Standard deviation squares distances, so longer distances are penalized more than shorter ones, while mean absolute deviation penalizes each distance equally. One isn't better than the other, but SD is more common than MAD.



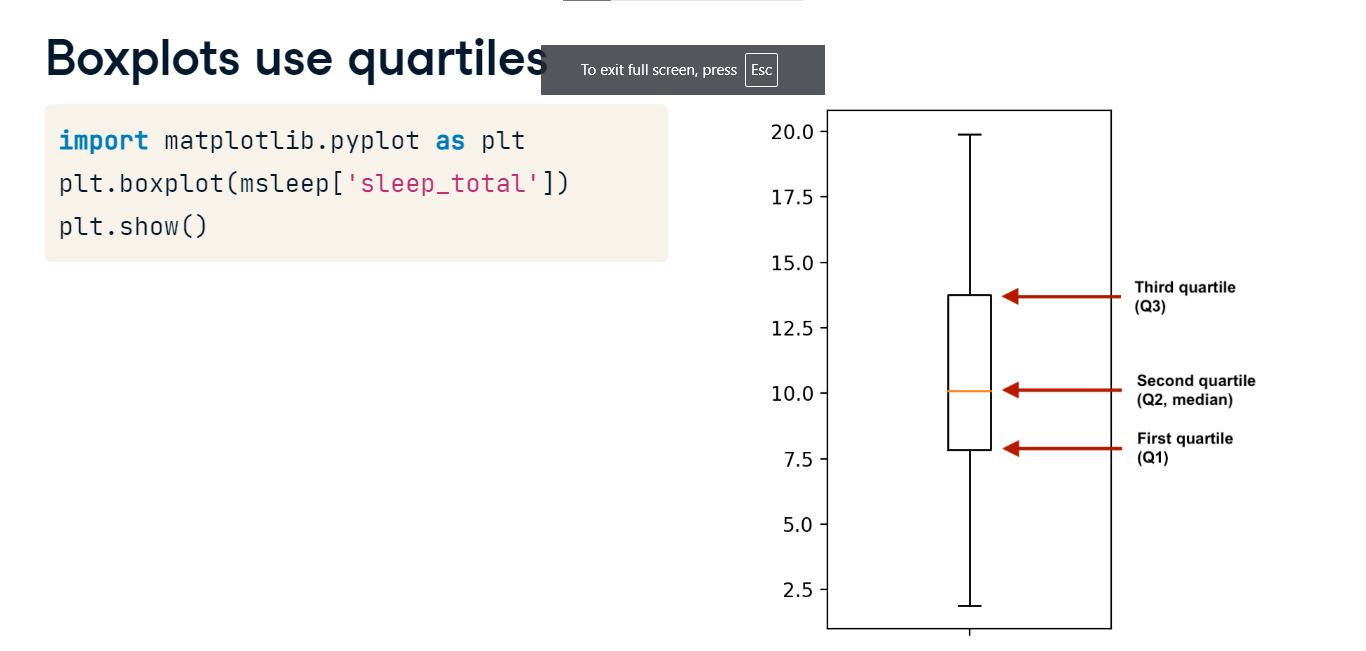
**Quantiles**

Before we discuss the next measure of spread, let's quickly talk about quantiles. Quantiles, also called percentiles, split up the data into some number of equal parts. Here, we call np-dot-quantile, passing in the column of interest, followed by point-5. This gives us 10-point-1 hours, so 50% of mammals in the dataset sleep less than 10-point-1 hours a day, and the other 50% sleep more than 10-point-1 hours, so this is exactly the same as the median. We can also pass in a list of numbers to get multiple quantiles at once. Here, we split the data into 4 equal parts. These are also called quartiles. This means that 25% of the data is between 1-point-9 and 7-point-85, another 25% is between 7-point-85 and 10-point-10, and so on.



**Boxplots use quartiles**

The boxes in box plots represent quartiles. The bottom of the box is the first quartile, and the top of the box is the third quartile. The middle line is the second quartile, or the median.



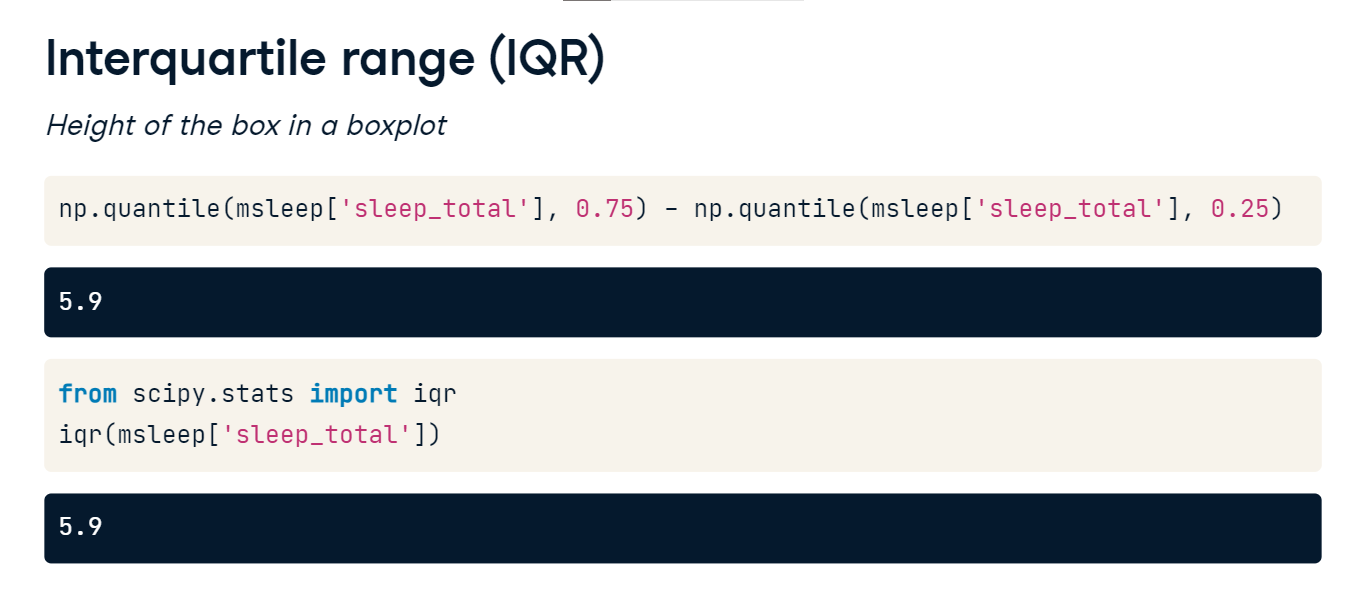
**Quantiles using np.linspace()**

Here, we split the data in five equal pieces, but we can also use np-dot-linspace as a shortcut, which takes in the starting number, the stopping number, and the number intervals. We can compute the same quantiles using np-dot-linspace starting at zero, stopping at one, splitting into 5 different intervals.



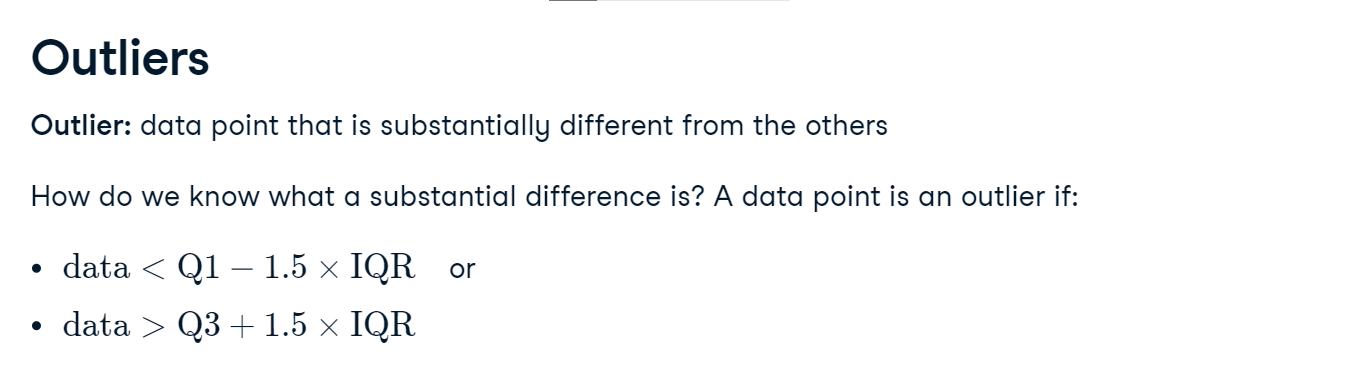
**Interquartile range (IQR)**

The interquartile range, or IQR, is another measure of spread. It's the distance between the 25th and 75th percentile, which is also the height of the box in a boxplot. We can calculate it using the quantile function, or using the iqr function from scipy-dot-stats to get 5-point-9 hours.



**Outliers**

Outliers are data points that are substantially different from the others. But how do we know what a substantial difference is? A rule that's often used is that any data point less than the first quartile minus 1-point-5 times the IQR is an outlier, as well as any point greater than the third quartile plus 1-point-5 times the IQR.



**Finding outliers**

To find outliers, we'll start by calculating the IQR of the mammals' body weights. We can then calculate the lower and upper thresholds following the formulas from the previous slide. We can now subset the DataFrame to find mammals whose body weight is below or above the thresholds. There are eleven body weight outliers in this dataset, including the cow and the Asian elephant.



**All in one go**

Many of the summary statistics we've covered so far can all be calculated in just one line of code using the dot-describe method, so it's convenient to use when you want to get a general sense of your data.

