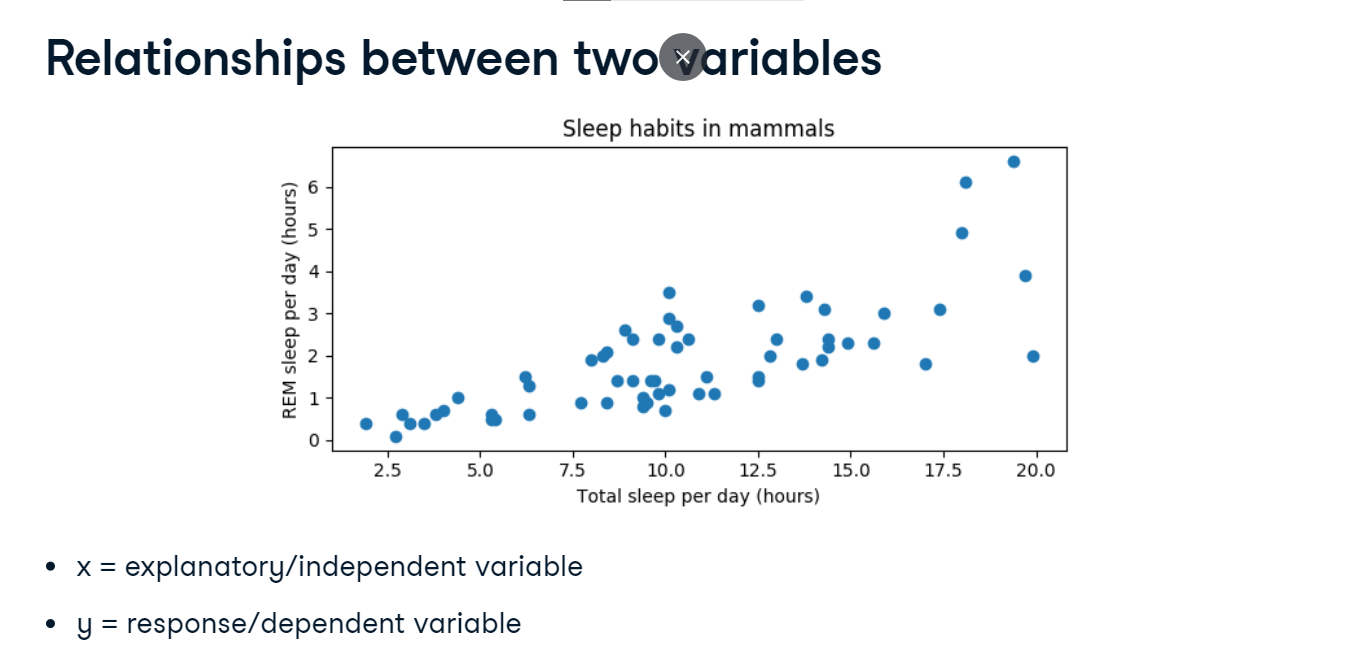
**Correlation**

Welcome to the final chapter of the course, where we'll talk about correlation and experimental design.

**Relationships between two variables**

Before we dive in, let's talk about relationships between numeric variables. We can visualize these kinds of relationships with scatter plots - in this scatterplot, we can see the relationship between the total amount of sleep mammals get and the amount of REM sleep they get. The variable on the x-axis is called the explanatory or independent variable, and the variable on the y-axis is called the response or dependent variable.



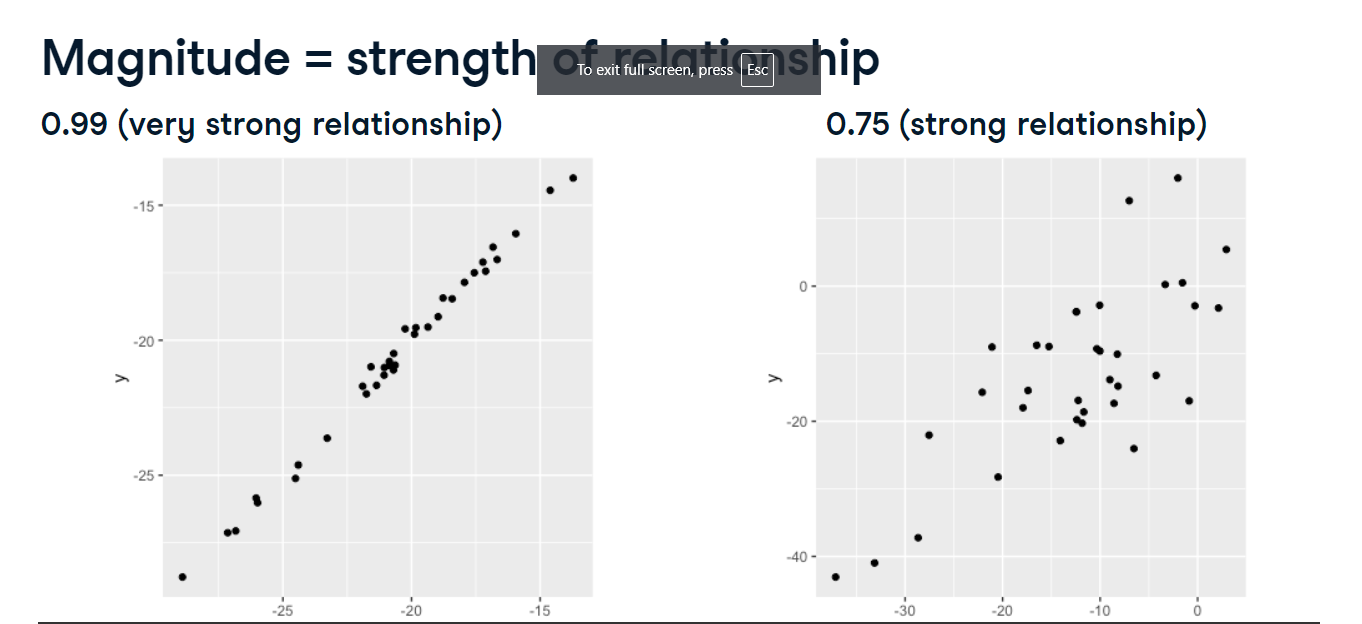
**Correlation coefficient**

We can also examine relationships between two numeric variables using a number called the correlation coefficient. This is a number between -1 and 1, where the magnitude corresponds to the strength of the relationship between the variables, and the sign, positive or negative, corresponds to the direction of the relationship.

**Magnitude = strength of relationship**

Here's a scatterplot of 2 variables, x and y, that have a correlation coefficient of 0-point-99. Since the data points are closely clustered around a line, we can describe this as a near-perfect or very strong relationship. If we know what x is, we'll have a pretty good idea of what the value of y could be.

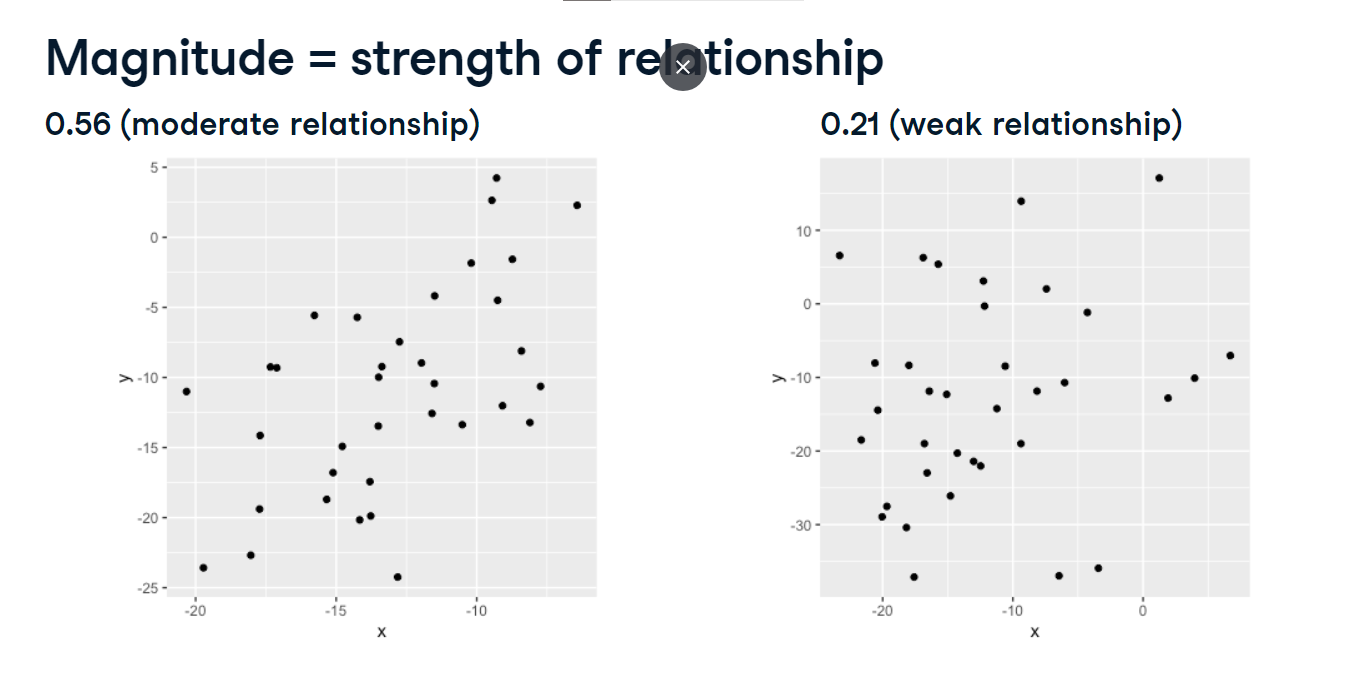
Here, x and y have a correlation coefficient of 0-point-75, and the data points are a bit more spread out.



**Magnitude = strength of relationship**

In this plot, x and y have a correlation of 0-point-56 and are therefore moderately correlated.

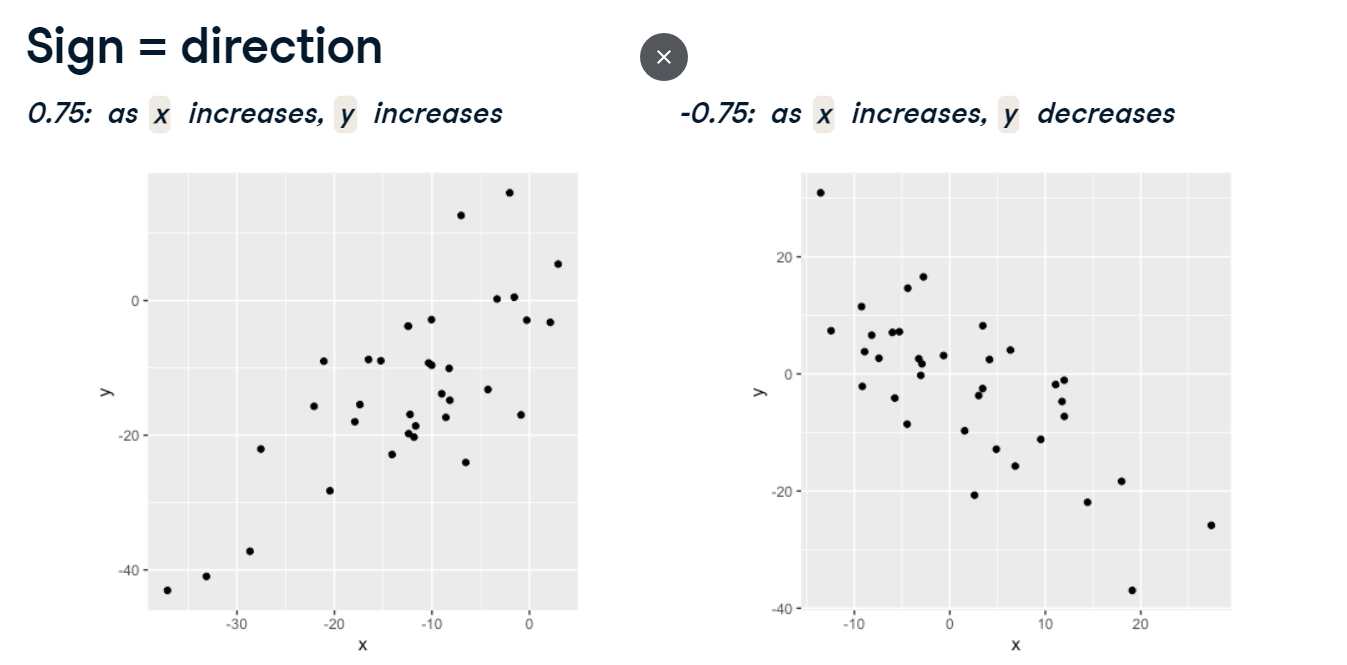
A correlation coefficient around 0-point-2 would be considered a weak relationship.

** Magnitude = strength of relationship**

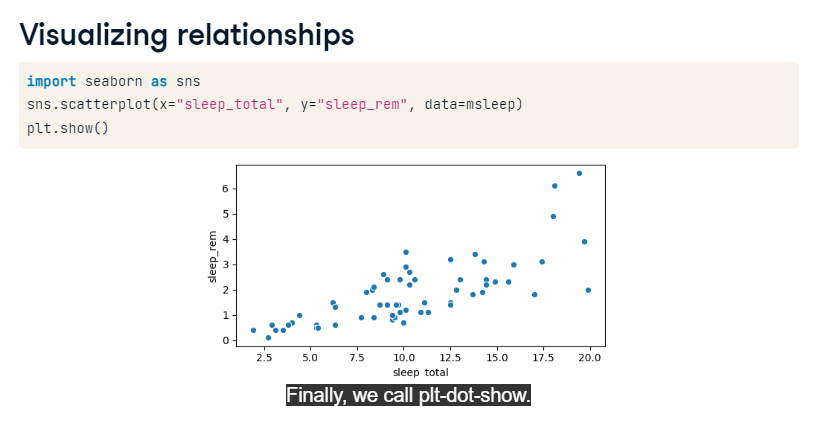
When the correlation coefficient is close to 0, x and y have no relationship and the scatterplot looks completely random. This means that knowing the value of x doesn't tell us anything about the value of y.

**Sign = direction**

The sign of the correlation coefficient corresponds to the direction of the relationship. A positive correlation coefficient indicates that as x increases, y also increases. A negative correlation coefficient indicates that as x increases, y decreases.

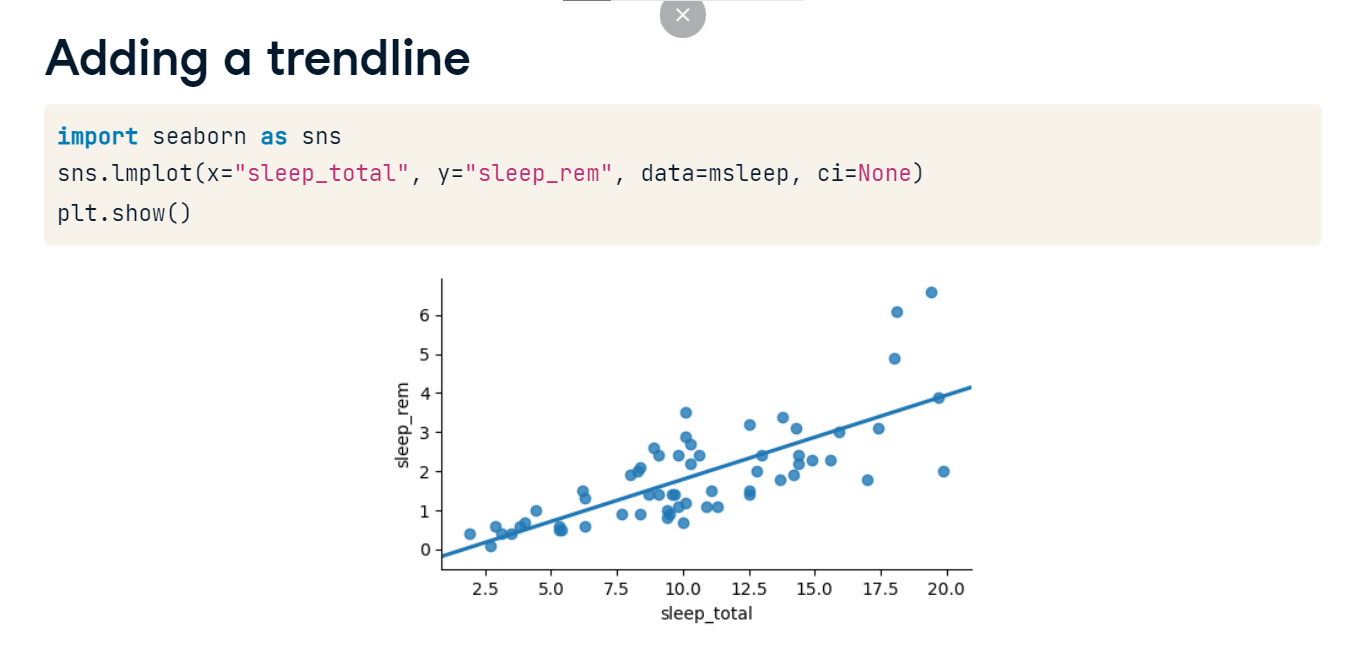


**Visualizing relationships**

To visualize relationships between two variables, we can use a scatterplot. We'll use seaborn, which is a plotting package built on top of matplotlib. We import seaborn as sns, which is the alias commonly used for seaborn. We create a scatterplot using sns-dot-scatterplot, passing it the name of the variable for the x-axis, the name of the variable for the y-axis, as well as the msleep DataFrame to the data argument. Finally, we call plt-dot-show. 

**Adding a trendline**

We can add a linear trendline to the scatterplot using seaborn's lmplot() function. It takes the same arguments as sns-dot-scatterplot, but we'll set ci to None so that there aren't any confidence interval margins around the line. Trendlines like this can be helpful to more easily see a relationship between two variables.



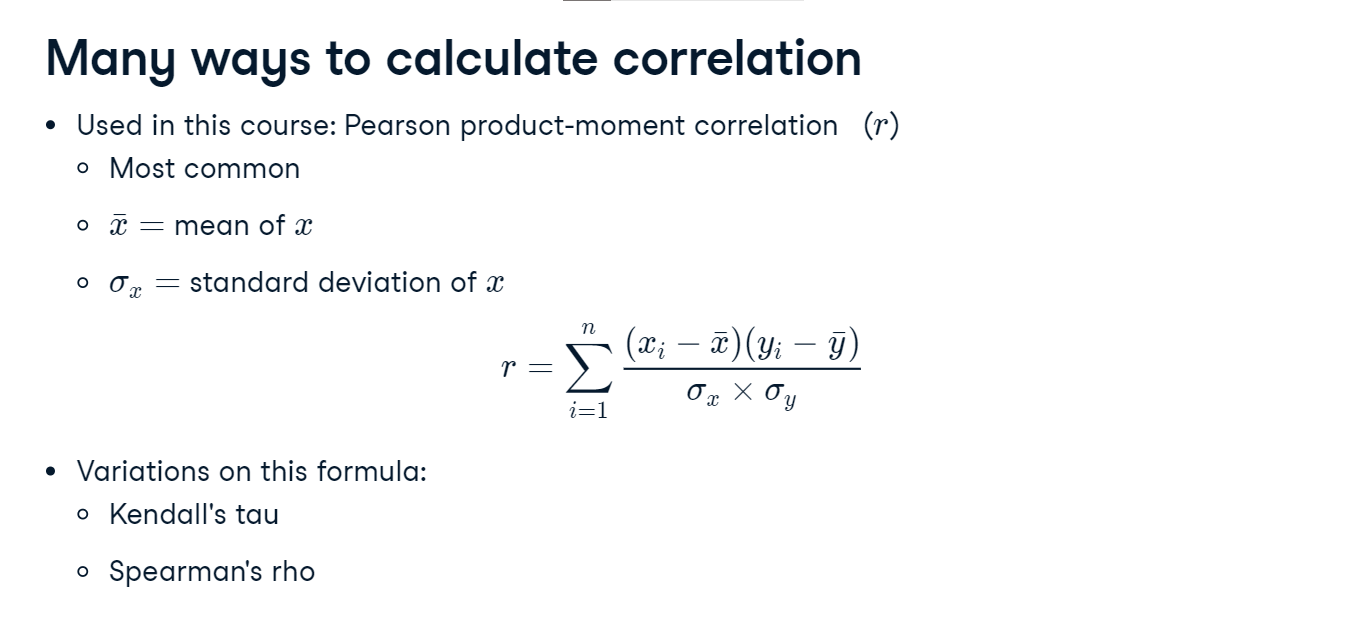
**Computing correlation**

To calculate the correlation coefficient between two Series, we can use the dot-corr method. If we want the correlation between the sleep\_total and sleep\_rem columns of msleep, we can take the sleep\_total column and call dot-corr on it, passing in the other Series we're interested in. Note that it doesn't matter which Series the method is invoked on and which is passed in since the correlation between x and y is the same thing as the correlation between y and x.



**Many ways to calculate correlation**

There's more than one way to calculate correlation, but the method we've been using in this video is called the Pearson product-moment correlation, which is also written as r. This is the most commonly used measure of correlation. Mathematically, it's calculated using this formula where x and y bar are the means of x and y, and sigma x and sigma y are the standard deviations of x and y. The formula itself isn't important to memorize, but know that there are variations of this formula that measure correlation a bit differently, such as Kendall's tau and Spearman's rho, but those are beyond the scope of this course.



**Correlation caveats**

00:00 - 00:06

While correlation is a useful way to quantify relationships, there are some caveats.

**2. Non-linear relationships**

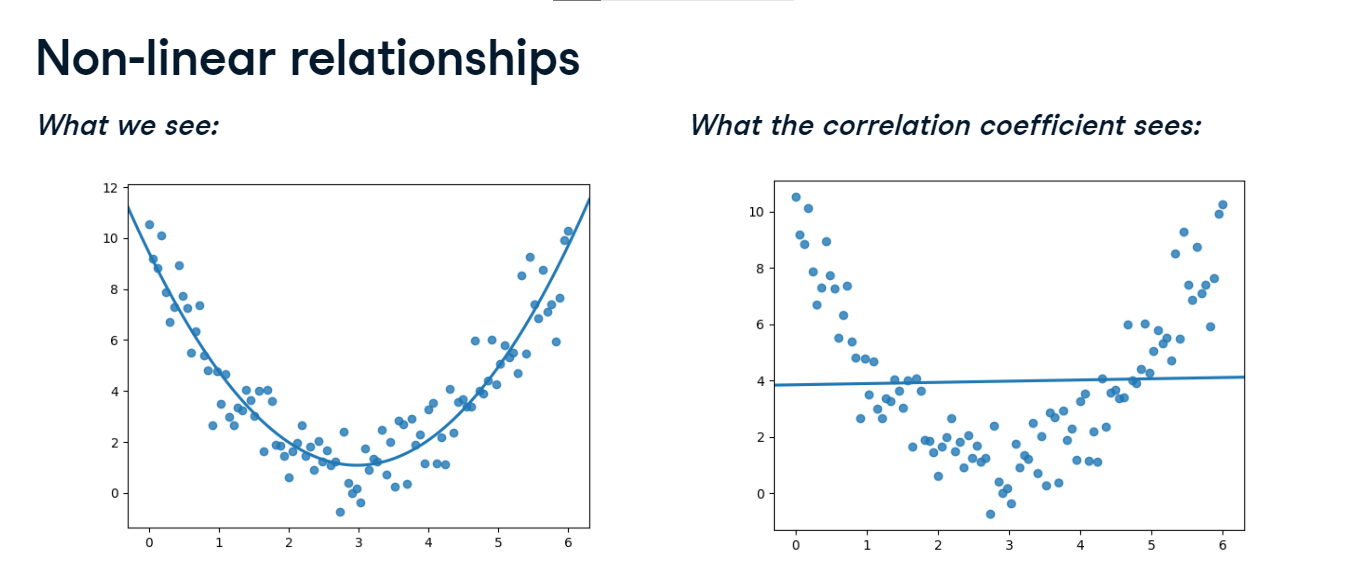
00:06 - 00:16

Consider this data. There is clearly a relationship between x and y, but when we calculate the correlation, we get 0-point-18.

**3. Non-linear relationships**

00:16 - 00:30

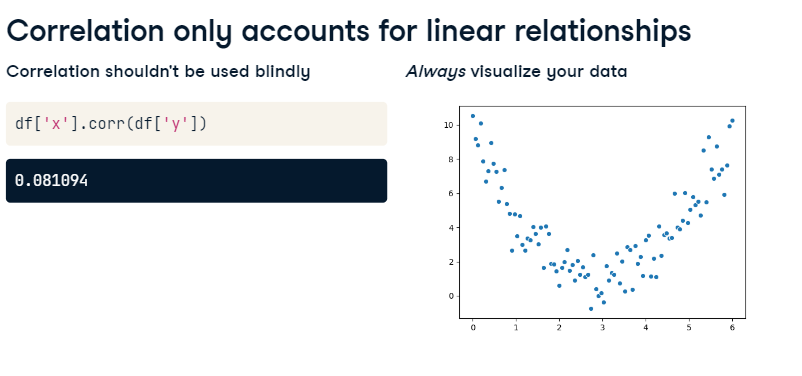
This is because the relationship between the two variables is a quadratic relationship, not a linear relationship. The correlation coefficient measures the strength of linear relationships, and linear relationships only.



**4. Correlation only accounts for linear relationships**

00:30 - 00:38

Just like any summary statistic, correlation shouldn't be used blindly, and you should always visualize your data when possible.



**5. Mammal sleep data**

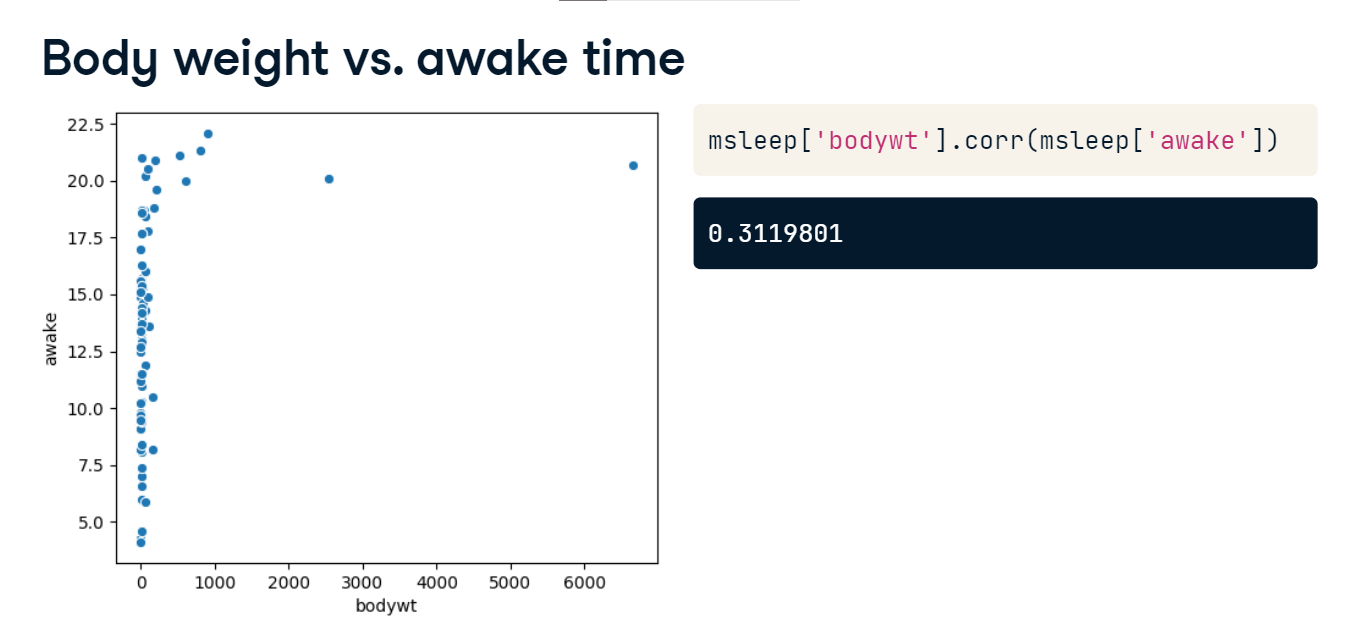
00:38 - 00:41

Let's return to the mammal sleep data.

**6. Body weight vs. awake time**

00:41 - 01:00

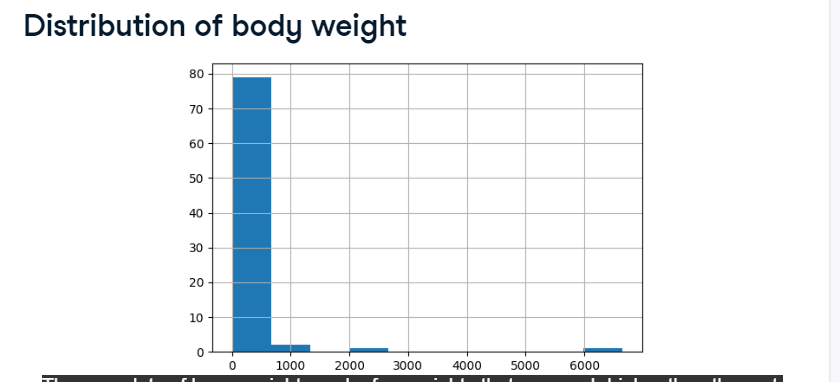
Here's a scatterplot of each mammal's body weight versus the time they spend awake each day. The relationship between these variables is definitely not a linear one. The correlation between body weight and awake time is only about 0-point-3, which is a weak linear relationship.



**7. Distribution of body weight**

01:00 - 01:12

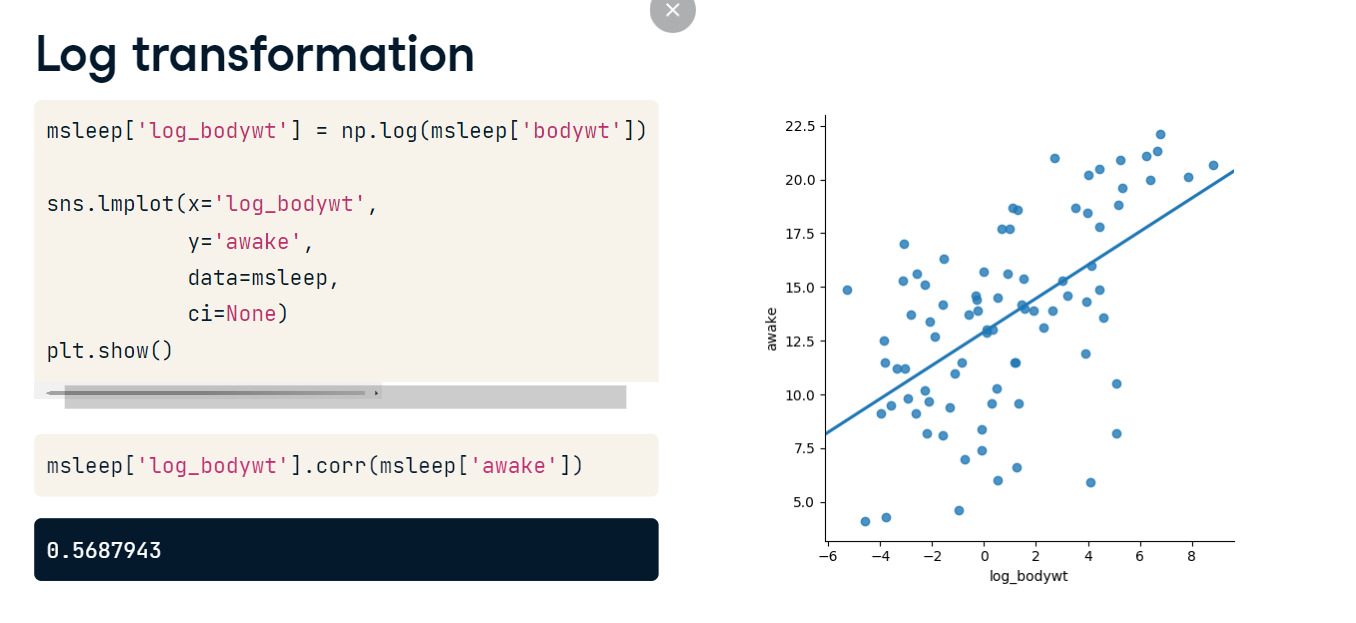
If we take a closer look at the distribution for bodywt, it's highly skewed. There are lots of lower weights and a few weights that are much higher than the rest.



**8. Log transformation**

01:12 - 01:44

When data is highly skewed like this, we can apply a log transformation. We'll create a new column called log\_bodywt which holds the log of each body weight. We can do this using np-dot-log. If we plot the log of bodyweight versus awake time, the relationship looks much more linear than the one between regular bodyweight and awake time. The correlation between the log of bodyweight and awake time is about 0-point-57, which is much higher than the 0-point-3 we had before.



**9. Other transformations**

01:44 - 02:13

In addition to the log transformation, there are lots of other transformations that can be used to make a relationship more linear, like taking the square root or reciprocal of a variable. The choice of transformation will depend on the data and how skewed it is. These can be applied in different combinations to x and y, for example, you could apply a log transformation to both x and y, or a square root transformation to x and a reciprocal transformation to y.



**10. Why use a transformation?**

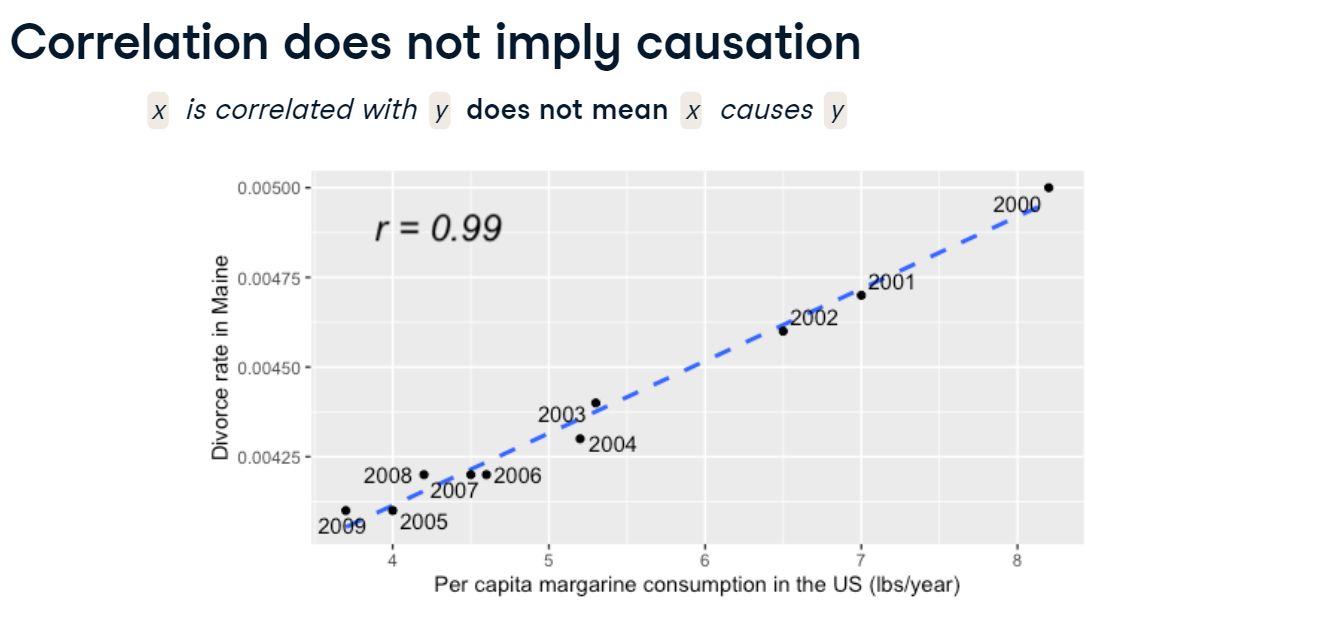
02:13 - 02:34

So why use a transformation? Certain statistical methods rely on variables having a linear relationship, like calculating a correlation coefficient. Linear regression is another statistical technique that requires variables to be related in a linear manner, which you can learn all about in this course.

**11. Correlation does not imply causation**

02:34 - 03:14

Let's talk about one more important caveat of correlation that you may have heard about before: correlation does not imply causation. This means that if x and y are correlated, x doesn't necessarily cause y. For example, here's a scatterplot of the per capita margarine consumption in the US each year and the divorce rate in the state of Maine. The correlation between these two variables is 0-point-99, which is nearly perfect. However, this doesn't mean that consuming more margarine will cause more divorces. This kind of correlation is often called a spurious correlation.



**12. Confounding**

03:14 - 03:34

A phenomenon called confounding can lead to spurious correlations. Let's say we want to know if drinking coffee causes lung cancer. Looking at the data, we find that coffee drinking and lung cancer are correlated, which may lead us to think that drinking more coffee will give you lung cancer.

**13. Confounding**

03:34 - 03:39

However, there is a third, hidden variable at play, which is smoking.

**14. Confounding**

03:39 - 03:43

Smoking is known to be associated with coffee consumption.

**15. Confounding**

03:43 - 03:48

It is also known that smoking causes lung cancer.

**16. Confounding**

03:48 - 04:35

In reality, it turns out that coffee does not cause lung cancer and is only associated with it, but it appeared causal due to the third variable, smoking. This third variable is called a confounder, or lurking variable. This means that the relationship of interest between coffee and lung cancer is a spurious correlation. Another example of this is the relationship between holidays and retail sales. While it might be that people buy more around holidays as a way of celebrating, it's hard to tell how much of the increased sales is due to holidays, and how much is due to the special deals and promotions that often run around holidays. Here, special deals confound the relationship between holidays and sales.

