MITOCW | MIT15_071S17_Session_1.4.06_300k

In this video, we will see how we can add a new variable to our data frame.

Suppose that we want to add a variable to our USDA data frame that takes a value 1 if the food has higher sodium than average, and 0 if the food has lower sodium than average.

Let's do this step by step.

To check if the first food in the dataset has a higher amount of sodium compared to the average, we can simply ask R to dig up the first value in the Sodium vector, using the square brackets and the index 1.

And then compare it using the greater-than sign to the mean of the Sodium vector, and then do not forget to remove the non-available entries.

And we obtain TRUE.

How about the 50th food?

Well, let's go back using the Up arrow, and simply change the index 1 to 50, and now we get FALSE.

This means that the first food has higher sodium content than average, and the 50th food has lower sodium content than average.

Now, we can write the same command, but on all the vector Sodium.

Let's use the Up arrow, and delete the square brackets with the index 50.

But we know we have 7,000 foods, and we really don't want to output 7,000 values right now.

So how about instead, we just save the output to a vector, and we're going to call it HighSodium.

And now let's look at the structure of the HighSodium vector.

And then we see that the HighSodium vector indeed has all these values-- TRUE and FALSE-- which are called logicals.

So basically the type of the HighSodium vector is logical.

But remember, we said we wanted values 1's and 0's.

So instead of TRUE, we want 1.

And instead of FALSE, we want a value of 0.

Well, to do this, we need to change the data type of HighSodium to numeric, and we can do this using the as.numeric function.

So let's use the Up arrow twice, and then enclose this logical expression by the as.numeric function.

So as.numeric, and now look up the structure of HighSodium, and now we see that we turned it into a numerical vector with values 0's and 1's.

Now, this vector, HighSodium, is not associated with the USDA data frame.

How can we add a variable, HighSodium, to our data frame?

Well, simply we need to use the dollar notation.

So let's go back twice to the command where we created the HighSodium vector, and then simply right now, instead of just calling it HighSodium, we associate it with the USDA data frame using the dollar notation.

Now, pressing Enter, and going and checking the structure of the USDA data frame, we see that we just added the HighSodium variable that was not present before, and it's a numerical variable with values 1's and 0's.

Now we can do the same, and add the variables HighProtein, HighCarbs, HighFat, similarly to our data frame.

Well, let's do this quickly using the Up arrow, and then let's go and replace Sodium now by Protein.

So again, here Sodium is replaced by Protein.

And then we're going to call this new variable HighProtein.

And do the same with TotalFat.

So instead of Protein, we're going to have TotalFat, and then replace it here again, and the variable name is going to be HighFat.

And finally, Carbohydrates-- so here is the vector of Carbohydrates, and this is, too, getting the Carbohydrates vector.

And finally, this last variable that we want to add is called HighCarbs.

And now looking at the structure of the USDA data frame, we see that we successfully added these three new variables, which are high HighProtein, HighFat, and HighCarbs, in addition to the HighSodium variable that we added previously.

So how can we now find relationships between these variables, and also the original variables that we had in the USDA data frame?

Well, we're going to be using the table and the tapply functions in our next video.