Let's go ahead and start R. The first thing you'll see in the R console is the version of R you are using, and other basic information related to licensing, citations, and demos.

To clear the console, you can simply go to Edit and select Clear Console.

We'll start by reading in our dataset USDA.csv, which contains all foods in the USDA database in 100-gram amounts.

You should have already downloaded the dataset to your computer.

To be able to read the dataset in R, we first need to navigate to the directory in our computer, where the data file, USDA.csv, is saved.

To do so, if you are on a Mac, go to the Misc menu, and select Change Working Directory.

If you are on a PC, go to the File menu, select Change Directory, and then navigate to the folder where you saved the csv file.

Then press OK.

Nothing has happened in R until now, except changing the working directory.

To double-check that we are in the right working directory, we can type getwd, which stands for get working directory, and then R gives us the path to the folder we just selected.

Now we should be ready to read in our dataset.

We will use the function read.csv, since the dataset was given to us in a csv format.

And let's save the output to a data frame, and call it USDA.

And this is equal to read.csv, and this takes, as an input, the name of the csv file, which is USDA.csv, and don't forget the quotation marks around the name of the csv file.

Pressing Enter, R now read the information from the dataset, and saved it to the data frame, USDA.

Now it's time to learn about our data.

We can use the structure function, or str in R, and give it the input USDA.

This gives us the following information.

We have 7,058 observations, or foods in our dataset, along with 16 different variables.

The first variable gives a unique identification number for each of the foods, starting with the number 1,001.

The second variable gives a text description of each of the foods.

The third variable is the amount of calories in 100 grams of these foods, and it's given to us in kilocalories.

Then we also have information about the protein, total fat, carbohydrate, saturated fat, and sugar levels in grams, as well as the sodium, cholesterol, calcium, iron, potassium, and vitamin C levels, in milligrams.

And finally, the amount of vitamin E and vitamin D in what is known as international units, and this is a standard measurement for drugs and vitamins.

Now to obtain high-level statistical information about our dataset, we can use the summary function, and give it, as an input, the USDA data frame.

The summary function gives us information such as the minimum, the maximum, and the mean values across all 7,058 foods for each of the 16 different variables.

For instance, the maximum amount of cholesterol is 3,100 milligrams, whereas the mean is only 41.55 milligrams.

We also have information about the number of non-available entries.

For instance, we have 1,910 foods that are missing entries for their sugar levels.

Now, scrolling through this information, a startling observation is the maximum level of sodium, which is 38,758 milligrams.

This number is huge, given that the daily recommended maximum is only 2,300 milligrams.

Let's investigate this variable further in our next video.