In this video, we'll do some basic data analysis using our WHO data.

To access a variable in a data frame, you always have to link it to the data frame it belongs to with the dollar sign.

To see this, let's first try typing Under15 and hitting Enter.

R responds with an error.

That's because R won't recognize this variable name since it doesn't know to look in the data frame WHO.

Now type WHO$Under15 and hit Enter.

This outputs the Under15 vector of the data frame WHO.

We can compute some statistics about this variable, such as the mean, using the mean function and then in parentheses typing WHO$Under15.

Close the parentheses and hit Enter.

This tells us that the average percentage of the population under 15 is 28.7.

We can also compute the standard deviation using the sd function.

So type sd and then in parentheses WHO$Under15.

Close the parentheses and hit Enter.

This tells us that the standard deviation of the percentage of the population under 15 is 10.5.

We can also get the statistical summary of just one variable using the summary function like we did before for the whole data frame.

To do this, we can type summary, and then in parentheses WHO$Under15.

Close the parentheses and hit Enter.

This gives the minimum value, the first quartile, the median value, the mean, the third quartile, and the maximum value of the variable Under15.

The first quartile is the value for which 25% of the data is less than that value, and the third quartile is the value for which 75% of the data is less than that value.

This output tells us that there's a country with only 13% of the population under 15.

Let's see which country it is using the which.min function.

So we can type which.min and then in parentheses our variable WHO$Under15.

Close the parentheses and hit Enter.

This returns the number 86, which is the row number of the observation with the minimum value of Under15.

To see which country is in row 86, we can type WHO$Country, for the country name, and then in square brackets 86, and hit Enter.

So Japan is the country with the minimum percentage of the population under 15.

Now let's see which country has the maximum percentage of the population under 15.

We can do this with the which.max function.

So type which.max and then in parentheses WHO$Under15.

Close the parentheses and hit Enter.

This tells us that the 124th observation has the maximum value of the variable Under15.

We can look up the country of the 124th observation by typing WHO$Country and then in square brackets 124.

Close the square brackets and hit Enter.

So Niger is the country with the maximum percentage of the population under 15.

Let's now create a scatter plot of GNI versus fertility rate.

You can do this using the plot function.

So type plot and then in parentheses WHO$GNI, the variable we want on our x-axis, comma, and then WHO$FertilityRate, the variable we want on our y-axis.

Close the parentheses and hit Enter.

A scatter plot should appear.

Income, or GNI, is on the x-axis, and fertility rate is on the y-axis.

Each point in the scatter plot is a country.

We can see that most countries here either have a low GNI or a high GNI but a low fertility rate.

However, there are a few countries for which both the GNI and the fertility rate are high.

Let's investigate.

We'll use the subset function to identify the countries with a GNI greater than 10,000, and a fertility rate greater than 2.5.

So go back to your R console and then type Outliers-- this is what we'll call our subset-- equals subset, and then in parentheses WHO comma GNI greater than 10,000 and FertilityRate greater than 2.5.

Close the parentheses and hit Enter.

When we used subset before, we only had one condition to define which observations to keep in the subset.

Here we have two conditions, separated by the and symbol.

This means that both conditions must be true for all observations in the subset.

We can see how many rows of data are in our subset by using the nrow function.

So type nrow for number of rows, and in parentheses the name of our subset, Outliers.

This tells us that there are seven countries for which the GNI is greater than 10,000 and the fertility rate is greater than 2.5.

Now let's output just the country names, GNI, and fertility rate of these seven countries to investigate further.

There's an easy way of doing this, and we'll use this technique several times in this class when we just want to extract a few variables from a data set.

So type the name of our data set, Outliers, and then in square brackets we'll make a vector of the names of the variables we want to output.

So c, and then in parentheses, "Country" for the country name, comma, "GNI" comma, and then "FertilityRate".

Close the parentheses, close the square brackets, and hit Enter.

This shows us the values of these three variables for the seven observations of outliers.

We can see that one of the seven countries is Equatorial Guinea, a country that is very rich per capita due to oil production, but the wealth is distributed very unevenly.

In the next video, we'll see how to create different types of plots in R and then build some summary tables.