In this video, we'll see how to color our points by region and how to add a linear regression line to our plot.

Here we have our plot from the last video, where we've colored the points dark red.

Now, let's color the points by region instead.

This time, we want to add a color option to our aesthetic, since we're assigning a variable in our data set to the colors.

To do this, we can type ggplot, and then first give the name of our data, like before, WHO, and then in our aesthetic, we again specify that x = GNI and y = FertilityRate.

But then we want to add the option color = Region, which will color the points by the Region variable.

And then we just want to add the geom_point function and hit Enter.

Now, in our plot, we should see that each point is colored corresponding to the region that country belongs in.

So the countries in Africa are colored red, the countries in the Americas are colored gold, the countries in the Eastern Mediterranean are colored green, etc.

This really helps us see something that we didn't see before.

The points from the different regions are really located in different areas on the plot.

Let's now instead color the points according to the country's life expectancy.

To do this, we just have to hit the up arrow to get back to our ggplot line, and then delete Region and type LifeExpectancy.

Now, we should see that each point is colored according to the life expectancy in that country.

Notice that before, we were coloring by a factor variable, Region.

So we had exactly seven different colors corresponding to the seven different regions.

Here, we're coloring by LifeExpectancy instead, which is a numerical variable, so we get a gradient of colors, like this.

Lighter blue corresponds to a higher life expectancy, and darker blue corresponds to a lower life expectancy.

Let's take a look at a different plot now.

Suppose we were interested in seeing whether the fertility rate of a country was a good predictor of the percentage of the population under 15.

Intuitively, we would expect these variables to be highly correlated.

But before trying any statistical models, let's explore our data with a plot.

So now, let's use the ggplot function on the WHO data again, but we're going to specify in our aesthetic that the x variable should be FertilityRate, and the y variable should be the variable, Under15.

Again, we want to add geom_point, since we want a scatterplot.

This is really interesting.

It looks like the variables are certainly correlated, but as the fertility rate increases, the variable, Under15 starts increasing less.

So this doesn't really look like a linear relationship.

But we suspect that a log transformation of FertilityRate will be better.

Let's give it a shot.

So go ahead and scroll up in your R console to the previous line, and instead of x = FertilityRate, we want x = log(FertilityRate).

And hit Enter.

Now this looks like a linear relationship.

Let's try building in a simple linear regression model to predict the percentage of the population under 15, using the log of the fertility rate.

So let's call our model, model, and use the lm function to predict Under15 using as an independent variable log(FertilityRate).

And our data set will be WHO.

Let's look at the summary of our model.

It looks like the log of FertilityRate is indeed a great predictor of Under15.

The variable is highly significant, and our R-squared is 0.9391.

Visualization was a great way for us to realize that the log transformation would be better.

If we instead had just used the FertilityRate, the R-squared would have been 0.87.

That's a pretty significant decrease in R-squared.

So now, let's add this regression line to our plot.

This is pretty easy in ggplot.

We just have to add another layer.

So use the up arrow in your R console to get back to the plotting line, and then add stat_smooth(method = "lm"), and hit Enter.

Now, you should see a blue line going through the data.

This is our regression line.

By default, ggplot will draw a 95% confidence interval shaded around the line.

We can change this by specifying options within the statistics layer.

So go ahead and scroll up in the R console, and after method = "lm", type level = 0.99, and hit Enter.

This will give a 99% confidence interval.

We could instead take away the confidence interval altogether by deleting level = 0.99 and typing se = FALSE.

Now, we just have the regression line in blue.

We could also change the color of the regression line by typing as an option, color = "orange".

Now, we have an orange linear regression line.

As we've seen in this lecture, scatterplots are great for exploring data.

However, there are many other ways to represent data visually, such as box plots, line charts, histograms, heat maps, and geographic maps.

In some cases, it may be better to choose one of these other ways of visualizing your data.

Luckily, ggplot makes it easy to go from one type of visualization to another, simply by adding the appropriate layer to the plot.

We'll learn more about other types of visualizations and how to create them in the next lecture.

So what is the edge of visualizations?

The WHO data that we used here is used by citizens, policymakers, and organizations around the world.

Visualizing the data facilitates the understanding of global health trends at a glance.

By using ggplot in R, we're able to visualize data for exploration, modeling, and sharing analytics results.