In this video, we'll create a heat map on a map of the United States.

We'll be using the data set murders.csv, which is data provided by the FBI giving the total number of murders in the United States by state.

Let's start by reading in our data set.

We'll call it murders, and we'll use the read.csv function to read in the data file murders.csv.

Let's take a look at the structure of this data using the str function.

We have 51 observations for the 50 states plus Washington, DC, and six different variables: the name of the state, the population, the population density, the number of murders, the number of murders that used guns, and the rate of gun ownership.

A map of the United States is included in R. Let's load the map and call it statesMap.

We can do so using the map_data function, where the only argument is "state" in quotes.

Let's see what this looks like by typing in str(statesMap).

This is just a data frame summarizing how to draw the United States.

To plot the map, we'll use the polygons geometry of ggplot.

So type ggplot, and then in parentheses, our data frame is statesMap, and then our aesthetic is x = long, the longitude variable in statesMap, y = lat, the latitude variable, and then group = group.

This is the variable defining how to draw the United States into groups by state.

Then close both parentheses here, and we'll add geom_polygon where our arguments here will be fill="white"-- we'll just fill all states in white-- and color="black" to outline the states in black.

Now in your R graphics window, you should see a map of the United States.

Before we can plot our data on this map, we need to make sure that the state names are the same in the murders data frame and in the statesMap data frame.

In the murders data frame, our state names are in the State variable, and they start with a capital letter.

But in the statesMap data frame, our state names are in the region variable, and they're all lowercase.

So let's create a new variable called region in our murders data frame to match the state name variable in the statesMap data frame.

So we'll add to our murders data frame the variable region, which will be equal to the lowercase version-- using the tolower function that we used in the text analytics lectures-- and the argument will be murders$State.

This will just convert the State variable to all lowercase letters and store it as a new variable called region.

Now we can join the statesMap data frame with the murders data frame by using the merge function, which matches rows of a data frame based on a shared identifier.

We just defined the variable region, which exists in both data frames.

So we'll call our new data frame murderMap, and we'll use the merge function, where the first argument is our first data frame, statesMap, the second argument is our second data frame, murders, and the third argument is by="region".

This is the identifier to use to merge the rows.

Let's take a look at the data frame we just created using the str function.

We have the same number of observations here that we had in the statesMap data frame, but now we have both the variables from the statesMap data frame and the variables from the murders data frame, which were matched up based on the region variable.

So now, let's plot the number of murders on our map of the United States.

We'll again use the ggplot function, but this time, our data frame is murderMap, and in our aesthetic we want to again say x=long, y=lat, and group=group, but we'll add one more argument this time, which is fill=Murders so that the states will be colored according to the Murders variable.

Then we need to add the polygon geometry where the only argument here will be color="black" to outline the states in black, like before.

And lastly, we'll add scale_fill_gradient where the arguments here, we'll put low="black" and high="red" to make our color scheme range from black to red, and then guide="legend" to make sure we get a legend on our plot.

If you hit Enter and look at your graphics window now, you should see that each of the states is colored by the number of murders in that state.

States with a larger number of murders are more red.

So it looks like California and Texas have the largest number of murders.

But is that just because they're the most populous states?

Let's create a map of the population of each state to check.

So back in the R Console, hit the Up arrow, and then, instead of fill=Murders, we want to put fill=Population to color each state according to the Population variable.

If you look at the graphics window, we have a population map here which looks exactly the same as our murders map.

So we need to plot the murder rate instead of the number of murders to make sure we're not just plotting a population map.

So in our R Console, let's create a new variable for the murder rate.

So in our murderMap data frame, we'll create the MurderRate variable, which is equal to murderMap$Murders-- the number of murders-- divided by murderMap$Population times 100,000.

So we've created a new variable that's the number of murders per 100,000 population.

Now let's redo our plot with the fill equal to MurderRate.

So hit the Up arrow twice to get back to the plotting command, and instead of fill=Population, this time we'll put fill=MurderRate.

If you look at your graphics window now, you should see that the plot is surprisingly maroon-looking.

There aren't really any red states.

Why?

It turns out that Washington, DC is an outlier with a very high murder rate, but it's such a small region on the map that we can't even see it.

So let's redo our plot, removing any observations with murder rates above 10, which we know will only exclude Washington, DC.

Keep in mind that when interpreting and explaining the resulting plot, you should always note what you did to create it: removed Washington, DC from the data.

So in your R Console, hit the Up arrow again, and this time, after guide="legend", we'll type limits=c(0,10) and hit Enter.

Now if you look back at your graphics window, you can see a range of colors on the map.

In this video, we saw how we can make a heat map on a map of the United States, which is very useful for organizations like the World Health Organization or government entities who want to show data to the public organized by state or country.

In the next video, we'll conclude by discussing the analytics edge of predictive policing.