Before we jump into R, let's understand the data.

Each entry of this data set corresponds to a census tract, a statistical division of the area that is used by researchers to break down towns and cities.

As a result, there will usually be multiple census tracts per town.

LON and LAT are the longitude and latitude of the center of the census tract.

MEDV is the median value of owner-occupied homes, measured in thousands of dollars.

CRIM is the per capita crime rate.

ZN is related to how much of the land is zoned for large residential properties.

INDUS is the proportion of the area used for industry.

CHAS is 1 if a census tract is next to the Charles River, which I drew before.

NOX is the concentration of nitrous oxides in the air, a measure of air pollution.

RM is the average number of rooms per dwelling.

AGE is the proportion of owner-occupied units built before 1940.

DIS is a measure of how far the tract is from centers of employment in Boston.

RAD is a measure of closeness to important highways.

TAX is the property tax per $10,000 of value.

And PTRATIO is the pupil to teacher ratio by town.

Let's switch over to R now.

So let's begin to analyze our data set with R. First of all, we'll load the data set into the Boston variable.

If we look at the structure of the Boston data set, we can see all the variables we talked about before.

There are 506 observations corresponding to 506 census tracts in the Greater Boston area.

We are interested in building a model initially of how prices vary by location across a region.

So let's first see how the points are laid out.

Using the plot commands, we can plot the latitude and longitude of each of our census tracts.

This picture might be a little bit meaningless to you if you're not familiar with the Massachusetts-Boston area, but I can tell you that the dense central core of points corresponds to Boston city, Cambridge city, and other close urban cities.

Still, let's try and relate it back to that picture we saw in the first video, where I showed you the river and where MIT was.

So we want to show all the points that lie along the Charles River in a different color.

We have a variable, CHAS, that tells us if a point is on the Charles River or not.

So to put points on an already-existing plot, we can use the points command, which looks very similar to the plot command, except it operates in a plot that already exists.

So let's plot just the points where the Charles River variable is set to one.

Up to now it looks pretty much like the plot command, but here's where it's about to get interesting.

We can pass a color, such as blue, to plot all these points in blue.

And this would plot blue hollow circles on top of the black hollow circles.

Which would look all right, but I think I'd much prefer to have solid blue dots.

To control how the points are plotted, we use a PCH option, which you can read about more in the help documentation for the points command.

But I'm going to use PCH 19, which is a solid version of the dots we already have on our plot.

So by running this command, you see we have some blue dots in our plot now.

These are the census tracts that lie along the Charles River.

But maybe it's still a little bit confusing, and you'd like to know where MIT is in this picture.

So we can do that too.

I looked up which census tract MIT is in, and it's census tract 3531.

So let's plot that.

We add another point, the longitude of MIT, which is in tract 3531, and the latitude of MIT, which is in census tract 3531.

I'm going to plot this one in red, so we can tell it apart from the other Charles River dots.

And again, I'm going to use a solid dot to do it.

Can you see it on the little picture?

This little red dot, right in the middle.

That's exactly what we were looking at from the picture in Video One What other things can we do?

Well, this data set was originally constructed to investigate questions about how air pollution affects prices.

So the air pollution variable is this NOX variable.

Let's have a look at a distribution of NOX.

boston$NOX.

So we see that the minimum value is 0.385, the maximum value is 0.87 and the median and the mean are about 0.53, 0.55.

So let's just use the value of 0.55, it's kind of in the middle.

And we'll look at just the census tracts that have above-average pollution.

So we'll use the points command again to plot just those points.

So, points, the latitude--no the longitude first.

So we want the census tracts with NOX levels greater than or equal to 0.55.

We want the latitude of those same census tracks.

Again, only if the NOX is greater than 0.55.

And I guess a suitable color for nasty pollution would be a bright green.

And again, we'll use the solid dots.

So you can see it is pretty much the same as the other commands.

Wow okay.

So those are all the points have got above-average pollution.

Looks like my office is right in the middle.

Now it kind of makes sense, though, because that's the dense urban core of Boston.

If you think of anywhere where pollution would be, you'd think it'd be where the most cars and the most people are.

So that's kind of interesting.

Now, before we do anything more, we should probably look at how prices vary over the area as well.

So let's make a new plot.

This one's got a few too many things on it.

So we'll just plot again the longitude and the latitude for all census tracts.

That kind of resets our plot.

If we look at the distribution of the housing prices (Boston MEDV), we see that the minimum price (remember units are thousands of dollars, so the median value of owner-occupied homes is in thousands of dollars) is around five, the maximum is around 50.

So let's plot again only the above-average price points.

So we'll go: points(boston$LON[boston$MEDV>=21.2].

We can also plot the latitude: boston$LATboston$LAT[boston$MEDV>=21.2].

We'll reuse that red color we used for MIT.

And one more time, with we'll do the solid dots.

So what we see now are all the census tracts with above-average housing prices.

As you can see, it's definitely not simple.

There's census tracts of above-average and below-average mixed in between each other.

But there are some patterns.

For example, look at that dense black bit in the middle.

That corresponds to most of the city of Boston, especially the southern parts of the city.

Also, on the Cambridge side of the river, there's a big chunk there of dots that are black, that are not red, that are also presumably below average.

So there's definitely some structure to it, but it's certainly not simple in relation to latitude and longitude at least.

We will explore this more in the next video.