Let's see how regression trees do.

We'll first load the rpart library and also load the rpart plotting library.

We build a regression tree in the same way we would build a classification tree, using the rpart command.

We predict MEDV as a function of latitude and longitude, using the Boston dataset.

If we now plot the tree using the prp command, which is to find an rpart.plot, we can see it makes a lot of splits and is a little bit hard to interpret.

But the important thing is look at the leaves.

In a classification tree, the leaves would be the classification we assign that these splits would apply to.

But in regression trees, we instead predict a number.

That number is the average of the median house prices in that bucket or leaf.

So let's see what that means in practice.

So we'll plot again the latitude-- the points.

And we'll again plot the points with above median prices.

I just scrolled up from my command history to do that.

Now we want to predict what the tree thinks is above median, just like we did with linear regression.

So we'll say the fitted values we can get from using the predict command on the tree we just built.

And we can do another points command, just like we did before.

The fitted values are greater than 21.2 The color is blue.

And the character is a dollar sign.

Now we see that we've done a much better job than linear regression was able to do.

We've correctly left the low value area in Boston and below out, and we've correctly managed to classify some of those points in the bottom right and top right.

We're still making mistakes, but we're able to make a nonlinear prediction on latitude and longitude.

So that's interesting, but the tree was very complicated.

So maybe it's drastically overfitting.

Can we get most of this effect with a much simpler tree?

We can.

We would just change the minbucket size.

So let's build a new tree using the rpart command again: MEDV as a function of LAT and LON, the data=boston.

But this time we'll say the minbucket size must be 50.

We'll use the other way of plotting trees, plot, and we'll add text to the text command.

And we see we have far fewer splits, and it's far more interpretable.

The first split says if the longitude is greater than or equal to negative 71.07-- so if you're on the right side of the picture.

So the left-hand branch is on the left-hand side of the picture and the right-hand-- So the left-hand side of the tree corresponds to the right-hand side of the map.

And the right side of the tree corresponds to the left side of the map.

That's a little bit of a mouthful.

Let's see what it means visually.

So we'll remember these values, and we'll plot the longitude and latitude again.

So here's our map.

OK.

So the first split was on longitude, and it was negative 71.07.

So there's a very handy command, "abline," which can put horizontal or vertical lines easily.

So we're going to put a vertical line, so v, and we wanted to plot it at negative 71.07.

OK.

So that's that first split from the tree.

It corresponds to being on either the left or right-hand side of this tree.

We'll plot the-- what we want to do is, we'll focus on one area.

We'll focus on the lowest price prediction, which is in the bottom left corner of the tree, right down the bottom left after all those splits.

So that's where we want to get to.

So let's plot again the points.

Plot a vertical line.

The next split down towards that bottom left corner was a horizontal line at 42.21.

So I put that in.

That's interesting.

So that line corresponds pretty much to where the Charles River was from before.

The final split you need to get to that bottom left corner I was pointing out is 42.17.

It was above this line.

And now that's interesting.

If we look at the right side of the middle of the three rectangles on the right side, that is the bucket we were predicting.

And it corresponds to that rectangle, those areas.

That's the South Boston low price area we saw before.

So maybe we can make that more clear by plotting, now, the high value prices.

So let's go back up to where we plotted all the red dots and overlay it.

So this makes it even more clear.

We've correctly shown how the regression tree carves out that rectangle in the bottom of Boston and says that is a low value area.

So that's actually very interesting.

It's shown us something that regression trees can do that we would never expect linear regression to be able to do.

So the question we're going to answer in the next video is given that regression trees can do these fancy things with latitude and longitude, is it actually going to help us to be able to build predictive models, predicting house prices?

Well, we'll have to see.