In the previous video, we got a feel for how regression trees can do things linear regression cannot.

But what really matters at the end of the day is whether it can predict things better than linear regression.

And so let's try that right now.

We're going to try to predict house prices using all the variables we have available to us.

So we'll load the caTools library.

That will help us do a split on the data.

We'll set the seed so our results are reproducible.

And we'll say our split will be on the Boston house prices and we'll split it 70% training, 30% test.

So our training data is a subset of the Boston data where the split is TRUE.

And the testing data is the subset of the Boston data where the split is FALSE.

OK, first of all, let's make a linear regression model, nice and easy.

It's a linear model and the variables are latitude, longitude, crime, zoning, industry, whether it's on the Charles River or not, air pollution, rooms, age, distance, another form of distance, tax rates, and the pupil-teacher ratio.

The data is training data.

OK, let's see what our linear regression looks like.

So we see that the latitude and longitude are not significant for a linear regression, which is perhaps not surprising because linear regression didn't seem to be able to take advantage of them.

Crime is very important.

The residential zoning might be important.

Whether it's on the Charles River or not is a useful factor.

Air pollution does seem to matter-- the coefficient is negative, as you'd expect.

The average number of rooms is significant.

The age is somewhat important.

Distance to centers of employment (DIS), is very important.

Distance to highways and tax is somewhat important, and the pupil-teacher ratio is also very significant.

Some of these might be correlated, so we can't put too much stock in necessarily interpreting them directly, but it's interesting.

The adjusted R squared is 0.65, which is pretty good.

So because it's kind of hard to compare out of sample accuracy for regression, we need to think of how we're going to do that.

With classification, we just say, this method got X% correct and this method got Y% correct.

Well, since we're doing continuous variables, let's calculate the sum of squared error, which we discussed in the original linear regression video.

So let's say the linear regression's predictions are predict(linreg, newdata=test) and the linear regression sum of squared errors is simply the sum of the predicted values versus the actual values squared.

So let's see what that number is-- 3,037.008.

OK, so you know what we're interested to see now is, can we beat this using regression trees?

So let's build a tree.

The tree rpart command again.

Actually to save myself from typing it all up again, I'm going to go back to the regression command and just change "lm" to "rpart" and change "linreg" to "tree"-- much easier.

All right.

So we've built our tree-- let's have a look at it using the "prp" command from "rpart.plot." And here we go.

So again, latitude and longitude aren't really important as far as the tree's concerned.

The rooms aren't the most important split.

Pollution appears in there twice, so it's, in some sense, nonlinear on the amount of pollution-- if it's greater than a

certain amount or less than a certain amount, it does different things.

Crime is in there, age is in there.

Room appears three times, actually-- sorry.

That's interesting.

So it's very nonlinear on the number of rooms.

Things that were important for the linear regression that don't appear in ours include pupil-teacher ratio.

The DIS variable doesn't appear in our regression tree at all, either.

So they're definitely doing different things, but how do they compare?

So we'll predict, again, from a tree.

"tree.pred" is the prediction of the tree on the new data.

And the tree sum of squared errors is the sum of the tree's predictions versus what they really should be.

And then the moment of truth-- 4,328.

So, simply put, regression trees are not as good as linear regression for this problem.

What this says to us, given what we saw with the latitude and longitude, is that latitude and longitude are nowhere near as useful for predicting, apparently, as these other variables are.

That's just the way it goes, I guess.

It's always nice when a new method does better, but there's no guarantee that's going to happen.

We need a special structure to really be useful.

Let's stop here with the R and go back to the slides and discuss how CP works and then we'll apply cross validation to our tree.

And we'll see if maybe we can improve in our results.