MITOCW | MIT15_071S17_Session_4.4.04_300k

So, we saw in the previous video that the house prices were distributed over the area in an interesting way, certainly not the kind of linear way.

And we wouldn't necessarily expect linear regression to do very well at predicting house price, just given latitude and longitude.

We can kind of develop an intuition more by plotting the relationship between latitude and house prices-- which doesn't look very linear-- or the longitude and the house prices, which also looks pretty nonlinear.

So, we'll try fitting it in a linear regression anyway.

So, let's call it latlonlm.

And we'll use the LM command, linear model, to predict house prices based on latitude and longitude using the Boston data set.

If we take a look at our linear regression, we see that r squared is around 0.1, which is not great.

The latitude is not significant, which means the north-south differences aren't going to be really used at all.

Longitude is significant, and it's negative.

Which we can interpret as, as we go towards the oceans-- we go towards the east-- house prices decrease linearly.

So this all seems kind of unlikely, but let's work with it.

So let's see how this linear regression model looks on a plot.

So let's plot the census tracts again.

OK.

Now, remember before, we had-- from the previous video-- we plotted the above-median house prices.

So we're going to do that one more time.

Median was 21.2.

We had-- the color was red.

And we used solid dots.

Ha.

Oops.

See what I did there?

I used the plot command, instead of the points command, and it plotted just the new points.

I meant to plot the original points and use the points command to plot it on top of the existing plot.

OK.

So that's more like it.

So now we have the median values with the above median value census tracts.

So, OK, we want to see, now, the question we're going to ask, and then plot, is what does a linear regression model think is above median.

So we could just do this pretty easily.

We have latlonlm$fitted.values and this is what the linear regression model predicts for each of the 506 census tracts.

So we'll plot these on top.

Boston$LON-- take all the census tracts, such that the latlonlm's fitted values are above the median.

Take the latitudes, too.

And I'm going to make them blue, but let's pause for a moment and think.

If we use the dots again, we'll cover up the red dots and cover up some of the black dots.

What we won't be able to see is where the red dots and the blue dots match up.

You know, we're interested in seeing how the linear regression matches up with the truth.

So it'd be ideal if we could plot the linear regression blue dots on top of the red dots, in some way that we can still see the red dots.

It turns out that you can actually pass in characters to this PCH option.

So since we're talking about money, let's plot dollar signs instead of points.

And there you have it.

So, the linear regression model has plotted a dollar sign for every time it thinks the census tract is above median value.

And you can see that, indeed, it's almost as-- you can see the sharp line that the linear regression defines.

And how it's pretty much vertical, because remember before, the latitude variable was not very significant in the regression.

So that's interesting and pretty wrong.

One thing that really stands out is how it says Boston is mostly above median.

Even knowing-- we saw it right from the start-- there's a big non-red spot, right in the middle of Boston, where the house prices were below the median.

So the linear regression model isn't really doing a good job.

And it's completely ignored everything to the right side of the picture.