In real estate, there is a famous saying that the most important thing is location, location, location.

In this recitation, we will be looking at regression trees, and applying them to data related to house prices and locations.

Boston is the capital of the state of Massachusetts, USA.

It was first settled in 1630, and in the greater Boston area there are about 5 million people.

The area features some of the highest population densities in America.

Here is a shot of Boston from above.

In the middle of the picture, we have the Charles River.

I'm talking to you from my office at MIT.

My office is here.

This is MIT here.

MIT lies in the city of Cambridge, which is north of the river, and south over the river there is Boston City, itself.

In this recitation, we will be talking about Boston in a sense of the greater Boston area.

However, if we look at the housing in Boston right now, we can see that it is very dense.

Over the greater Boston area, the nature of the housing varies widely.

This data comes from a paper, "Hedonic Housing Prices and the Demand for Clean Air," which has been cited more than 1,000 times.

This paper was written on a relationship between house prices and clean air in the late 1970s by David Harrison of Harvard and Daniel Rubinfeld of the University of Michigan.

The data set is widely used to evaluate algorithms of a nature we discussed in this class.

Now, in the lecture, we will mostly discuss classification trees with the output as a factor or a category.

Trees can also be used for regression tasks.

The output at each leaf of a tree is no longer a category, but a number.

Just like classification trees, regression trees can capture nonlinearities that linear regression can't.

So what does that mean?

Well, with classification trees we report the average outcome at each leaf of our tree.

For example, if the outcome is true 15 times, and false 5 times, the value at that leaf of a tree would be 15/(15+5)=0.75.

Now, if we use the default threshold of 0.5, we would say the value at this leaf is true.

With regression trees, we now have continuous variables.

So instead of-- we report the average of the values at that leaf.

So suppose we had the values 3, 4, and 5 at one of the leaves of our trees.

Well, we just take the average of these numbers, which is 4, and that is what we report.

That might be a bit confusing so let's look at a picture.

Here is some fake data that I made up in R.

We see x on the x-axis and y on the y-axis.

y is our variable we are trying to predict using x.

So if we fit a linear regression to this data set, we obtain the following line.

As you can see, linear regression does not do very well on this data set.

However, we can notice that the data lies in three different groups.

If we draw these lines here, we see x is either less than 10, between 10 and 20, or greater then 20, and there is very different behavior in each group.

Regression trees can fit that kind of thing exactly.

So the splits would be x is less than or equal to 10, take the average of those values.

x is between 10 and 20, take the average of those values.

x is between 20 and 30, take the average of those values.

We see that regression trees can fit some kinds of data very well that linear regression completely fails on.

Of course, in reality nothing is ever so nice and simple, but it gives us some idea why we might be interested in regression trees.

So in this recitation, we will explore the data set with the aid of trees.

We will compare linear regression with regression trees.

We will discuss what the cp parameter means that we brought up when we did cross-validation in the lecture, and we will apply cross-validation to regression trees.