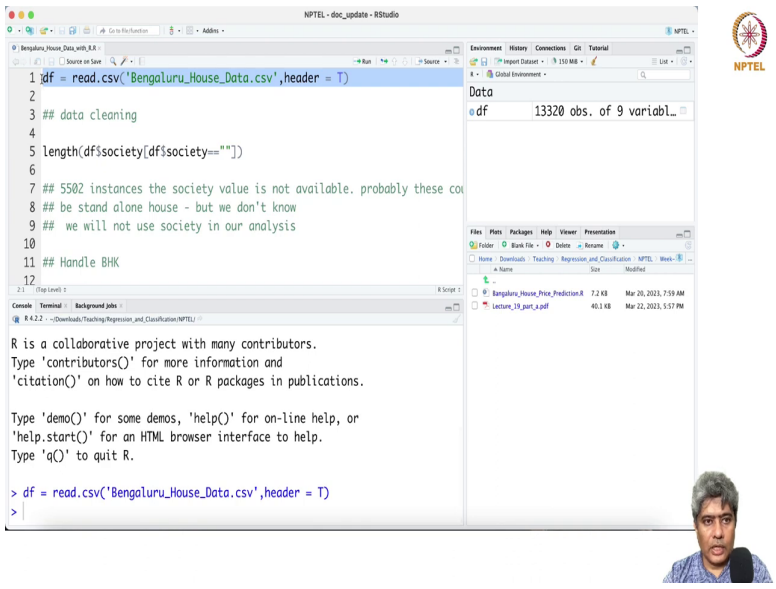
Predictive Analytics - Regression and Classification
Prof. Sourish Das
Department of Mathematics
Chennai Mathematical Institute

Lecture - 57
Hands on with R: Some Correction with Bangalore House Price Data Prediction

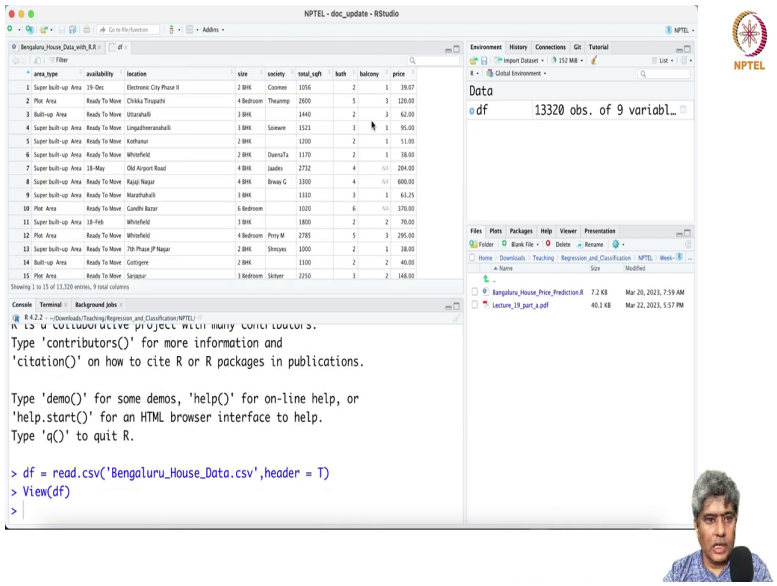
Hello all. In this video, we are going to do continue with Bangalore House Price Prediction Hands on Project. In this last week, I found there was some mistake later. I found some mistakes and today I am going to discuss those mistakes and how can we fix those mistakes.

(Refer Slide Time: 00:54)



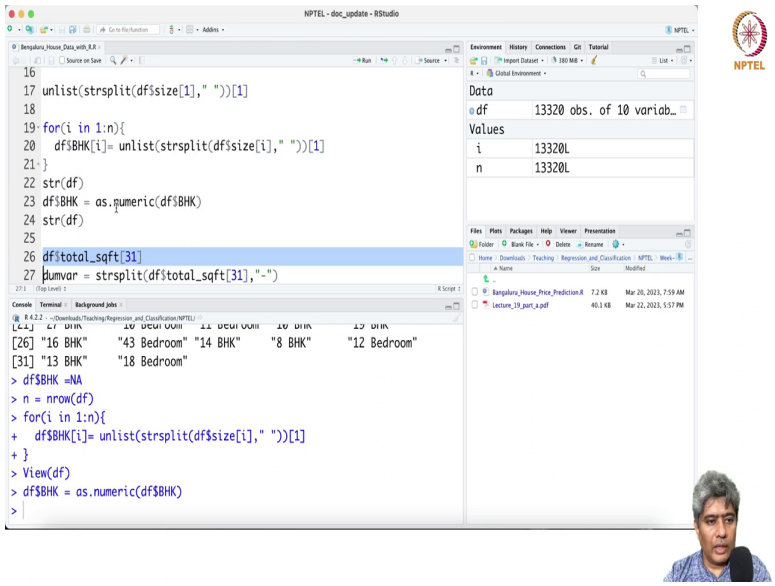
So, first let me show, let me just open the R, ok. So, let me. So, let me quickly go through the dataset that we were working on. So, here is the, we were reading in this line, we were reading the Bangalore House price dataset.

(Refer Slide Time: 01:22)



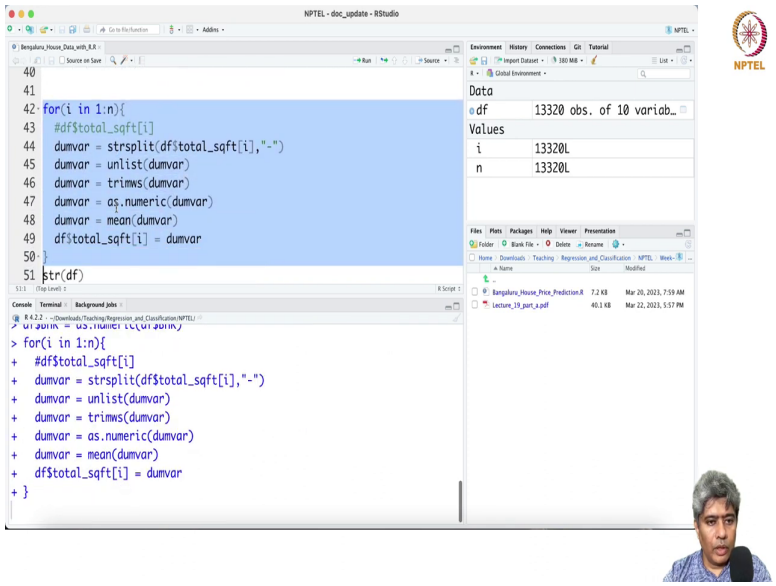
So, here is the dataset and so area type availability location, you have already seen this dataset.

(Refer Slide Time: 01:33)



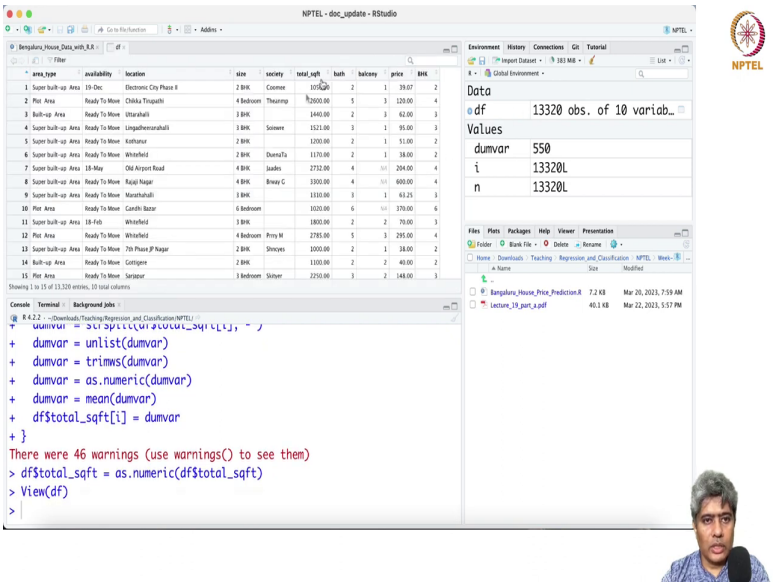
So, first we handled the BHK. So, from we there was a quote-unquote different ways of writing BHK. So, we created a column with variable name BHK. Initially it was a character variable, then we made it to numeric and then we handled the total square feet.

(Refer Slide Time: 02:00)



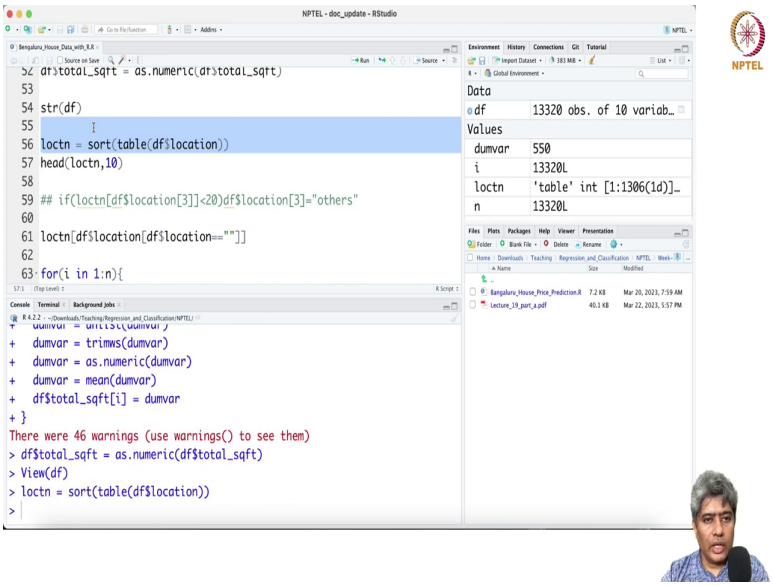
We come we compared the total square feet and then finally, total square feet was compare converted into numeric.

(Refer Slide Time: 02:08)

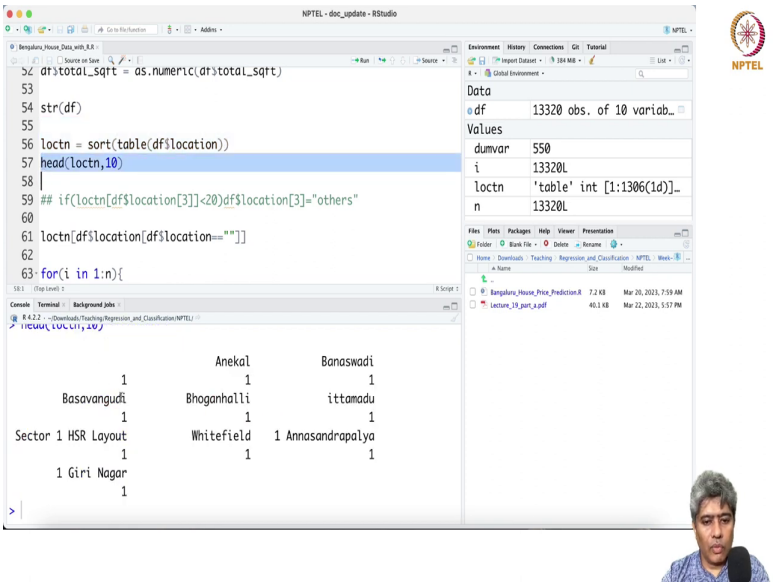


So, here is the total square feet and BHK, which were converted into from character to numeric.

(Refer Slide Time: 02:18)

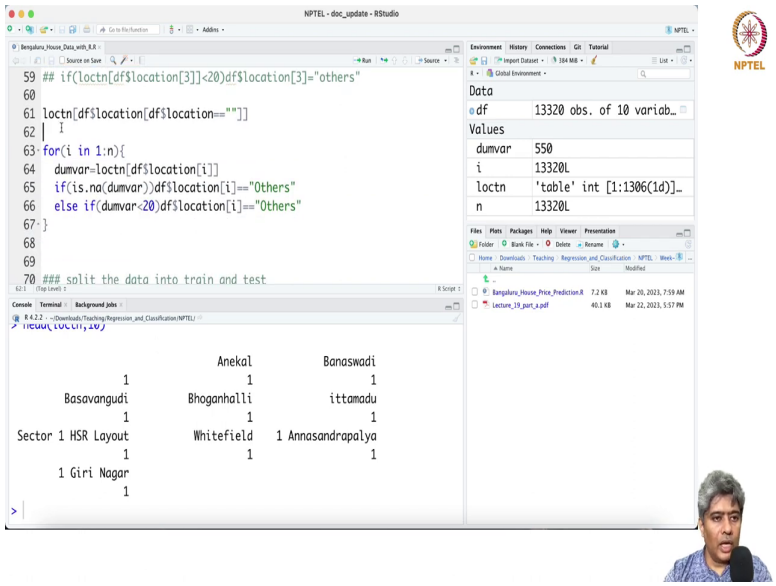


(Refer Slide Time: 02:27)



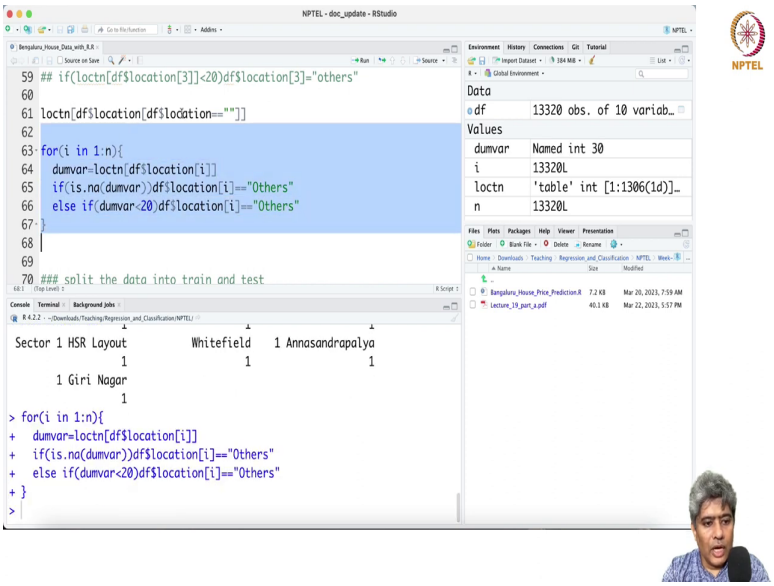
Now, then we took the locations and there were about quite a few 1300 locations that we found and out of that quite a few locations we have only one instances.

(Refer Slide Time: 02:32)

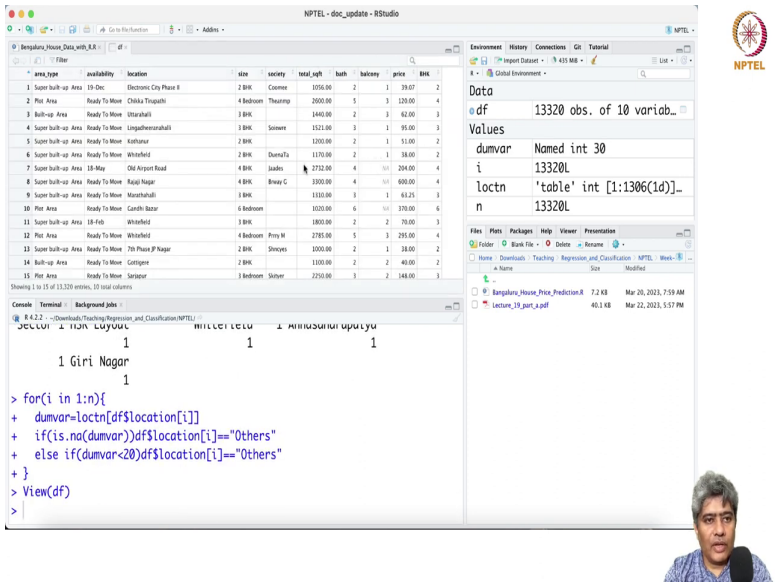


So, what we did first thing that we ran this and we say that ok, if dummy variable is less than 20, then it is others. But what I made a mistake that I was made a double equal to sign. In the double equal to sign actually it will not do anything ok, and it will just check whether it is true or false and it will go ahead.

(Refer Slide Time: 03:08)

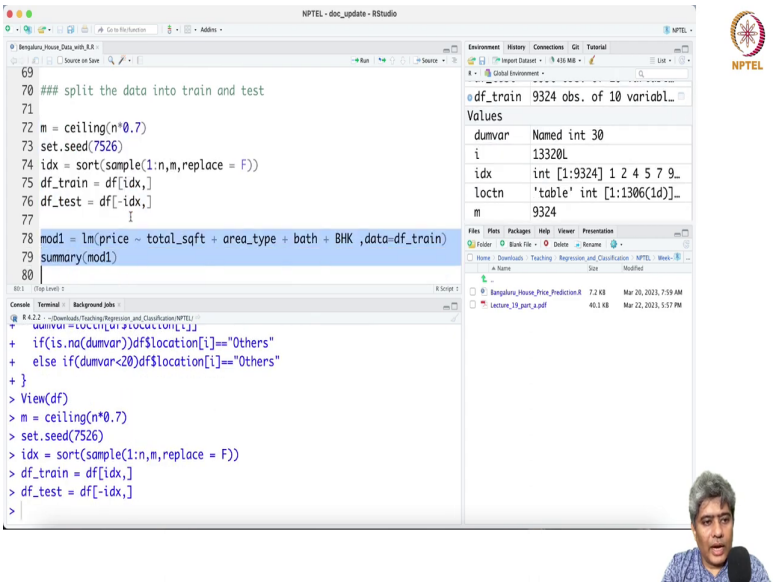


(Refer Slide Time: 03:15)

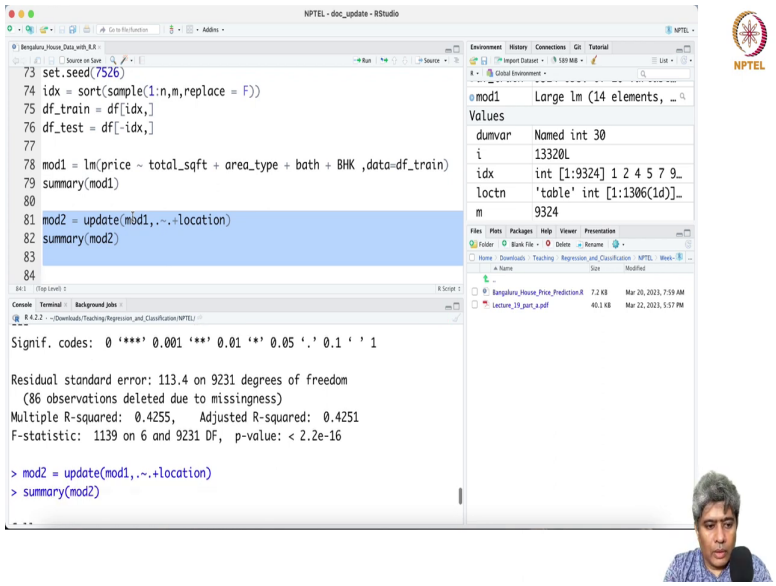


So, basically, we should have put others here. And first if you do not run it, if you do not run this, if you run this, you will not get any others effectively here, none of them are others. There is no others locations. So, this so, basically what I was doing that there are locations, which has instances less than 20. If it is less than 20, you put it into the others location and rest of them will be as usual.

(Refer Slide Time: 03:33)

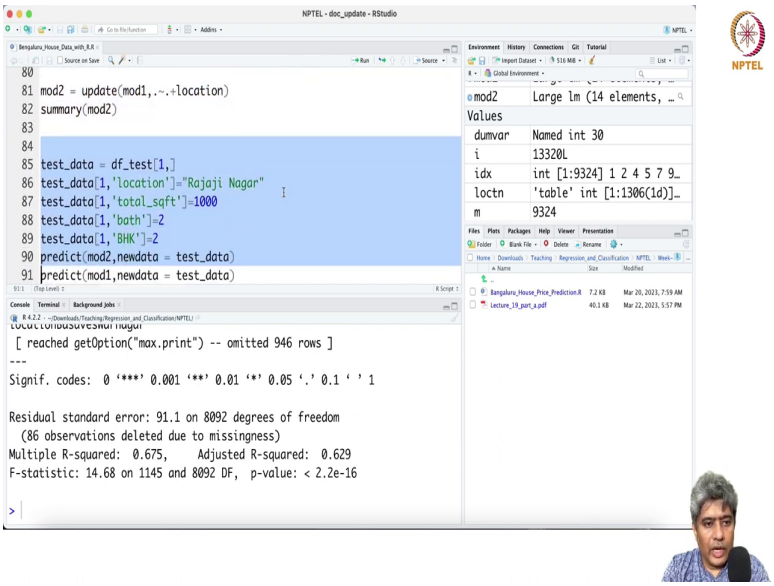


(Refer Slide Time: 03:44)

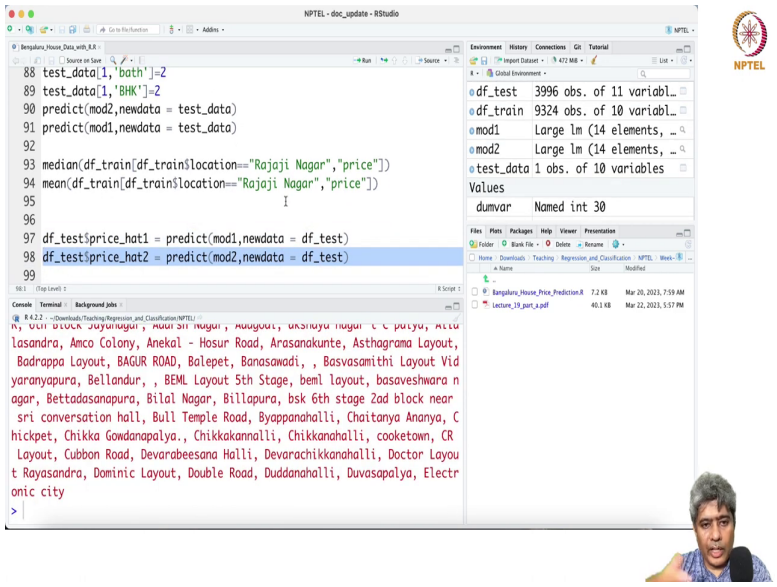


So, but it was sort of an ad-hoc correction. So, and then when we did run this, what we found that. So, it was not doing anything. We found that location was the first model had a 42 percent accuracy and the second model had about 62 percent accuracy, ok. Yeah, about 62 percent in sample accuracy.

(Refer Slide Time: 03:58)

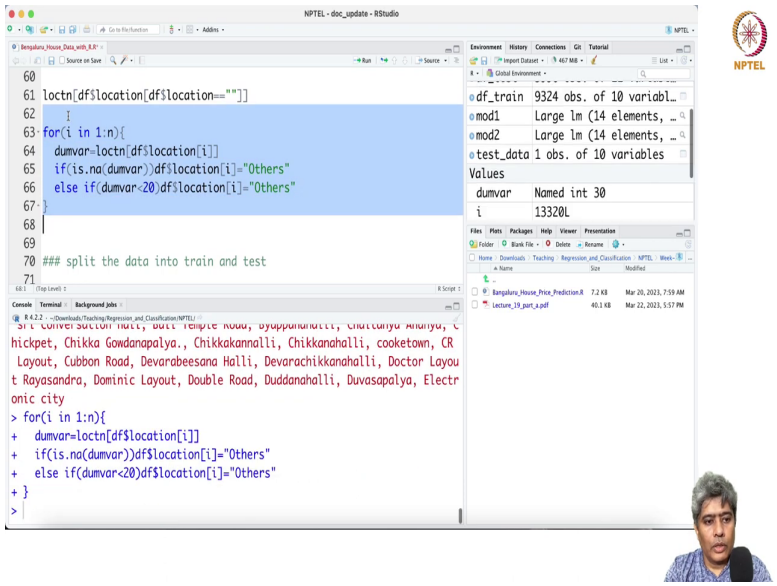


(Refer Slide Time: 04:09)



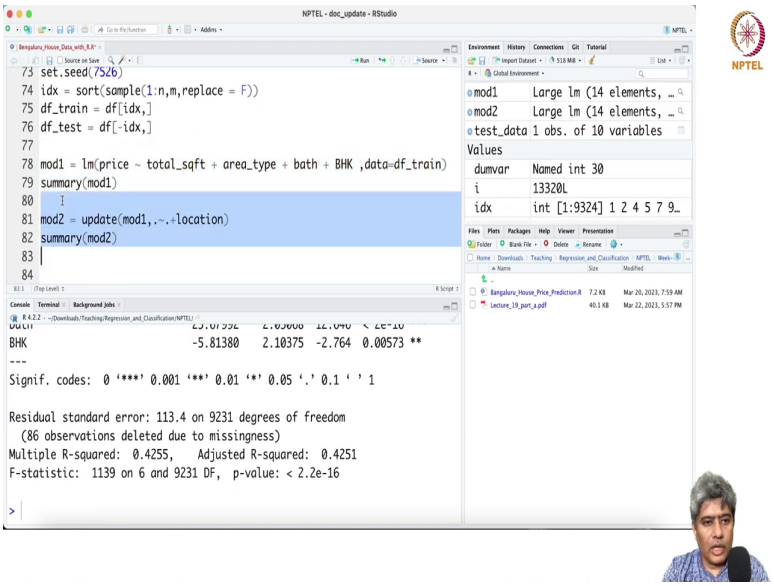
And then we did some test of the values and for Rajaji Nagar and all. And then we did a overall try to do the prediction. The prediction was failing. One reason the prediction was failing that there were locations, which were which came in the test data set, but they were not those locations were not there in the training data set. As a result, those the model was not trained for those locations, ok. And that is why it was throwing the error. And one reason was it was not correctly coded here.

(Refer Slide Time: 04:33)

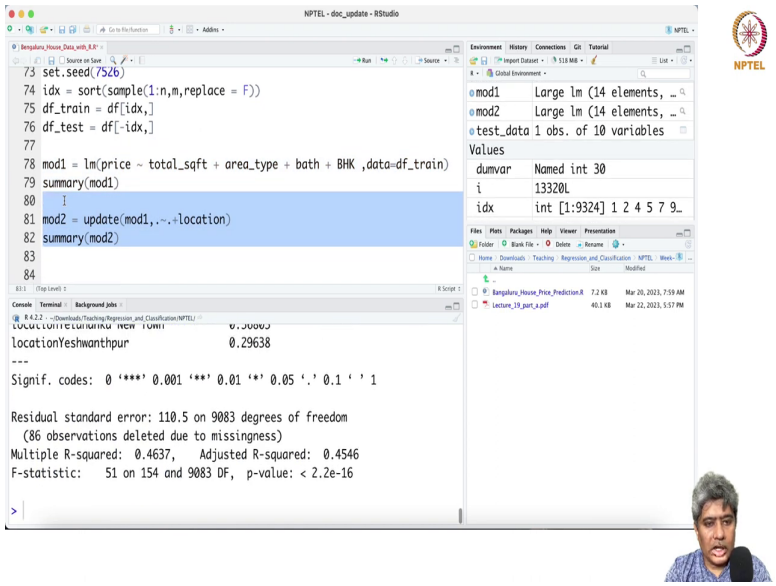


So, if you now, if I do it now, now it will be doing correctly coding. Now, if you do that and run this, ok.

(Refer Slide Time: 04:50)



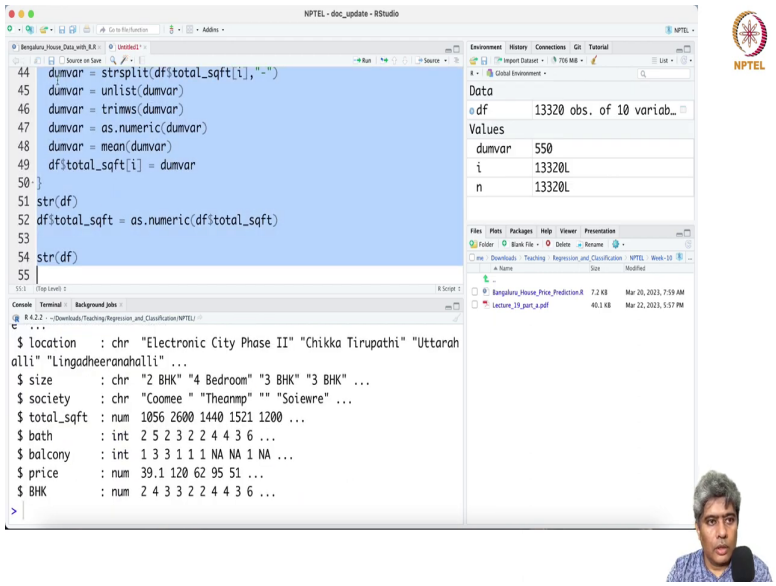
(Refer Slide Time: 04:53)



And now, you see 42 percent and if I add location, then after adding location, you are getting only 45 percent. So that means, if you if I really correctly do the others, then it will reduce the my accuracy of the model. And this adding other locations just because I do not have enough data, I will put it in club them into others is not a very good idea, correct.

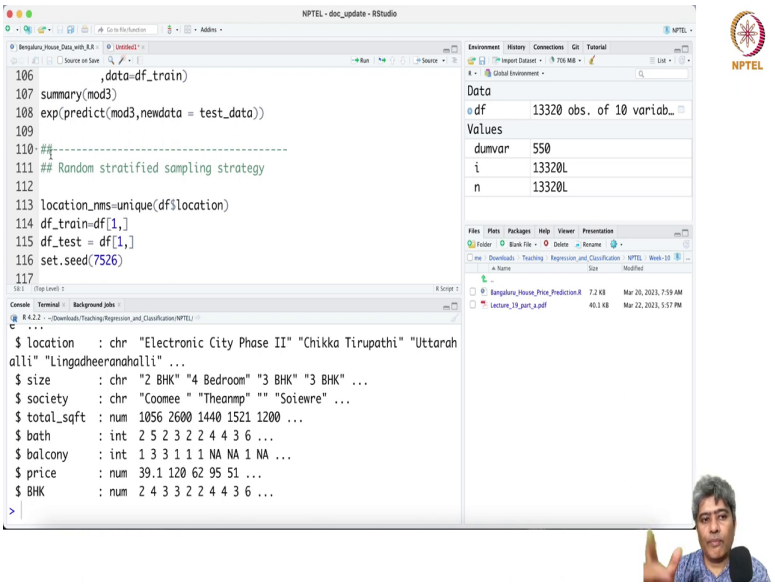
Because it is actually reducing the model accuracy to even in the in sample accuracy it is reducing. So, I do not think we should go ahead with this correction. So, we have to do a correction here. So, let me start a new script here. And let me just put, you know, let me just take up to this. So, what I will do here, let me just. So, I was up to location, right. Let me just take up to location. Let me just clean the entire thing, up to I will just take this up to this, ok.

(Refer Slide Time: 06:11)

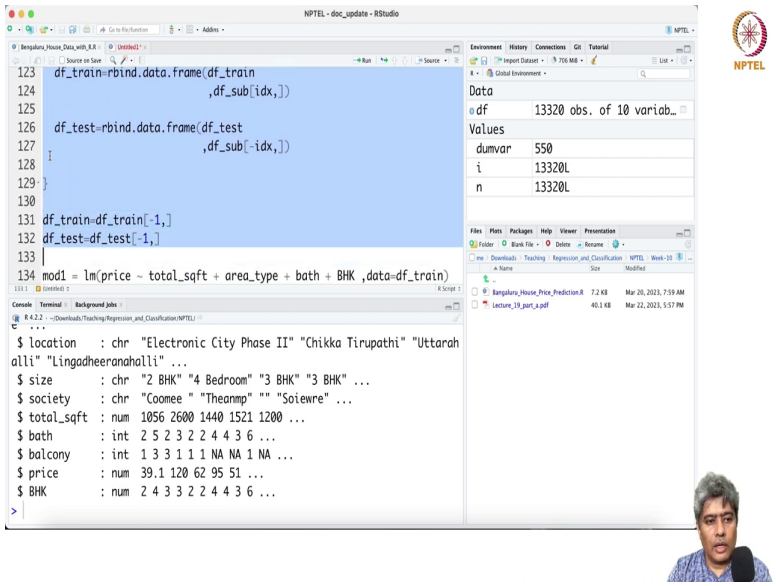


And then here is the location up to total square feet and then here is the location, correct. So, so then what I will do, I will just go and fix, I will just take this random stratified sampling strategy, but I will do in a effect in a slightly different way.

(Refer Slide Time: 06:40)

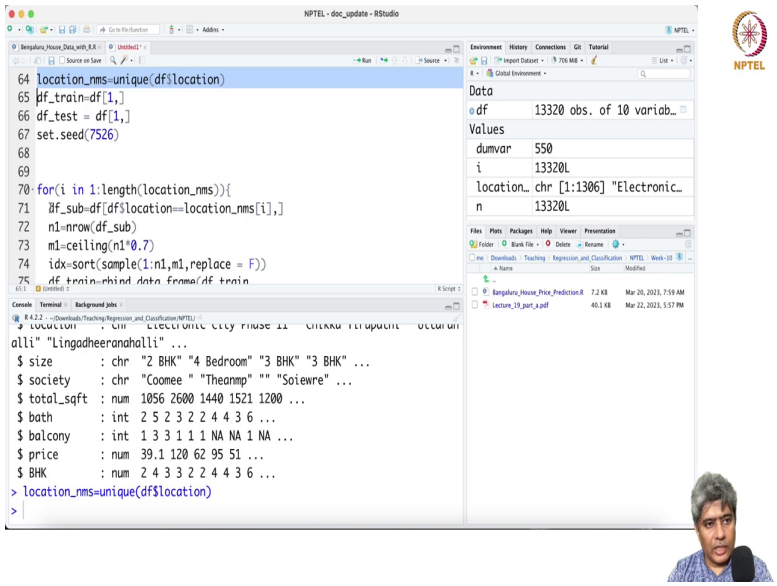


(Refer Slide Time: 06:55)



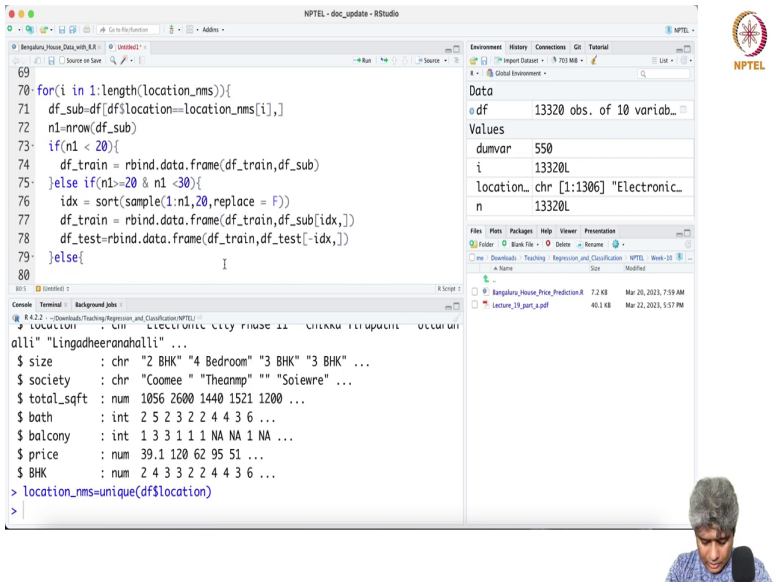
So, I will tell you this how we will do that. Let me just copy this part, ok. Let me just copy this part and you will understand what I am doing here. So, first thing I am going to do is what this piece of code is doing. It is taking.

(Refer Slide Time: 07:13)



So, let me it is I hope this is location names, yeah. So, it is for the ith location, it is just creating a subset, sub data set df sub. And then it is seeing what is the number of samples in that df sub.

(Refer Slide Time: 07:38)



Now, here I am going to do something interesting. If n1 is less than 20, if n1 is less than 20, that means, I have total number of instances for a location is less than 20, I am going to put all the locations for that, all the instances for that location into the training data set, ok.

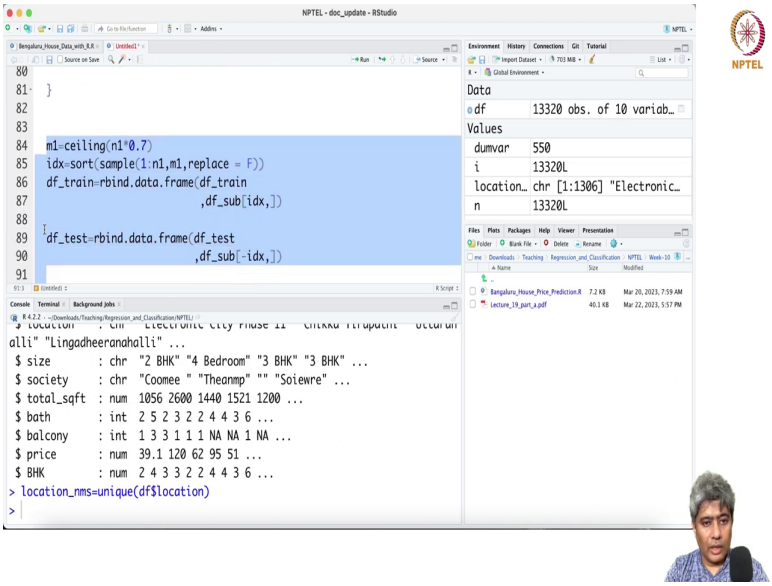
So, df train equal to rbind dot data dot frame df train comma df sub. All the instances goes to the training data set. I am not going to I will not have enough data to even test that location, if going forward more, but at least my model will be trained. And if in future, if somebody wants to come up with a test location test additional points, we can do the test or at least model will have the capacity to give you a for a test point. What will be the predicted value? It will be able to give you that.

Else if now here is an interesting thing if what happens is n1 is less than equal to 20 and n1 is less than 30. So, between if it is the values between 20 and 30, if it is value between 20 and

30, then what you do is you randomly draw 20 samples. So, minimum 20 sample you need boss. So, sort sample 1 is to n1.

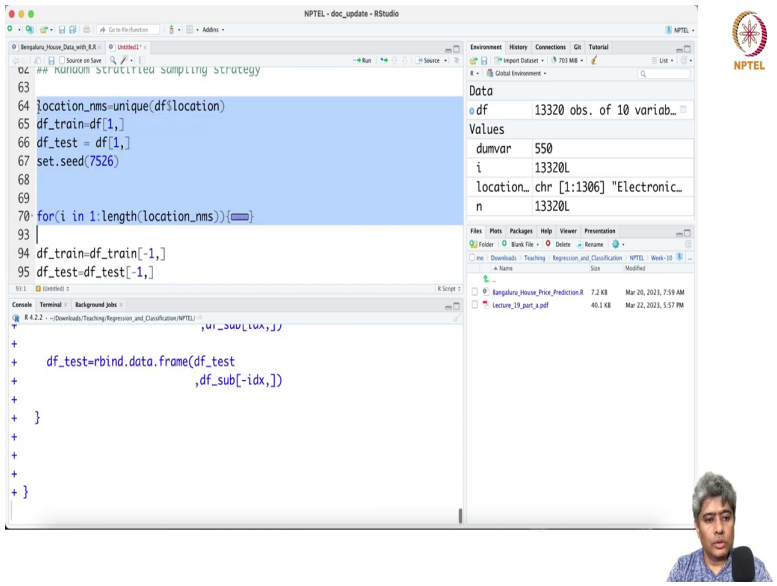
So, the if there is a 28 instances, you randomly draw 20 of them replace equal to false and put them in train, df train to df train and idx. So, those 20 samples will go into the training and rest of the samples will go to the if there are 28 instances so, 8 sample will go to the df test. So, df test minus idx, ok. And else it will just behave like a normal like 70 percent will go to the df test and, yeah.

(Refer Slide Time: 11:01)



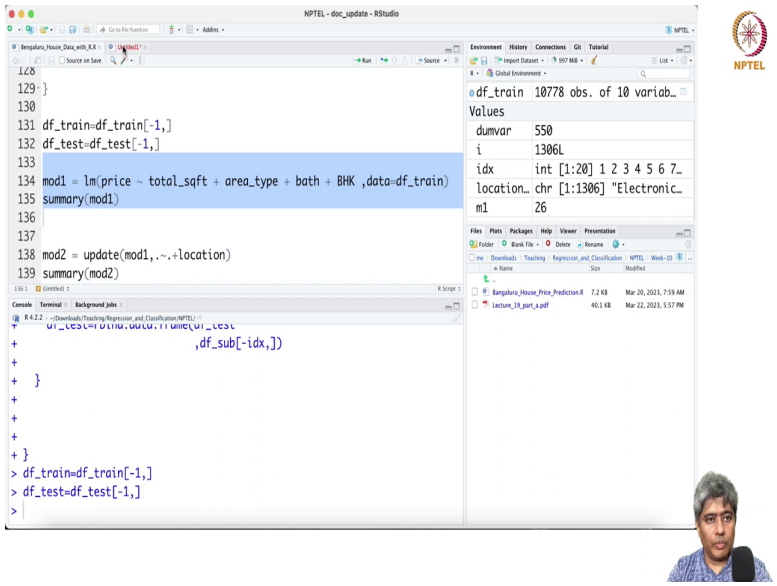
So, here I already have written. So, 70 percent will go to the train and rest of the samples will go to the test. So, this is what I am suggesting here.

(Refer Slide Time: 11:17)



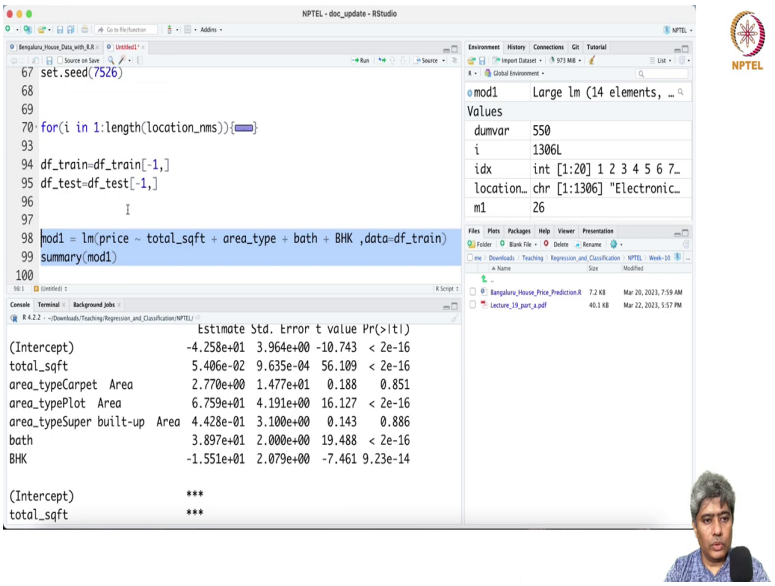
So, let me just run it through, ok. And then I am going to drop the first few rows.

(Refer Slide Time: 11:29)

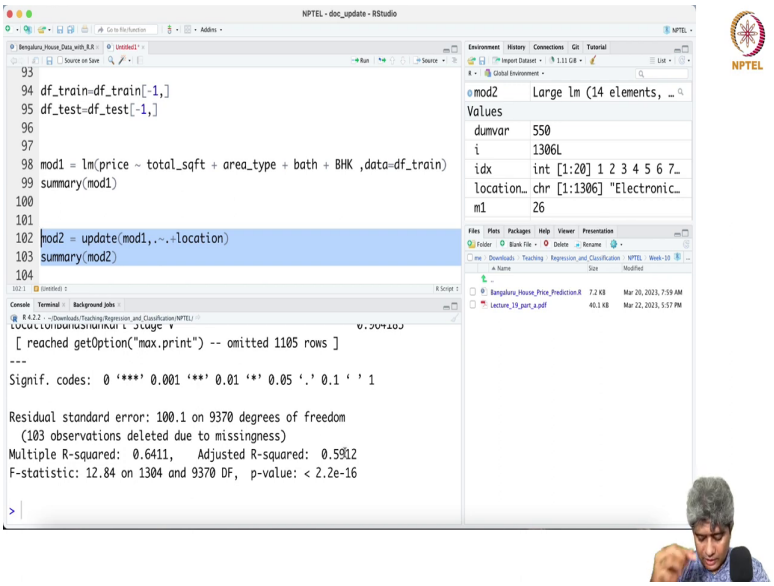


Now, I am going to call the first model. I am going to run the first model that we have had, ok. So, let us run the first model, copy the first model. And that gives us 39 percent accuracy, ok. That is fine. We will live with that.

(Refer Slide Time: 11:59)

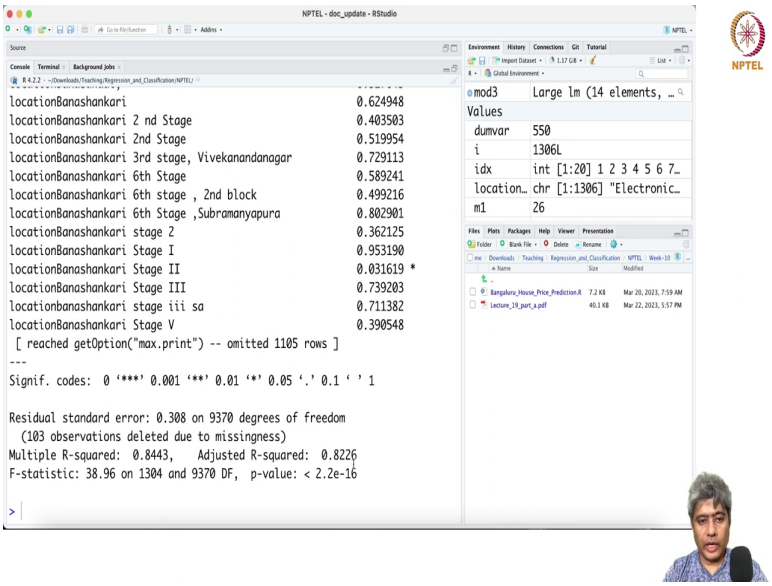


(Refer Slide Time: 12:07)



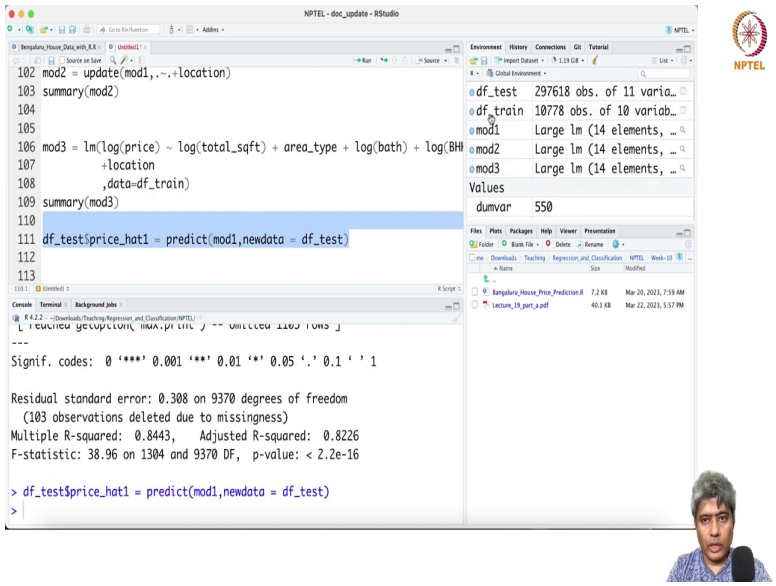
And if you add model 2 so, that gives us 59 percent accuracy, ok. And finally, if we just do model 3 and summary 3.

(Refer Slide Time: 12:32)



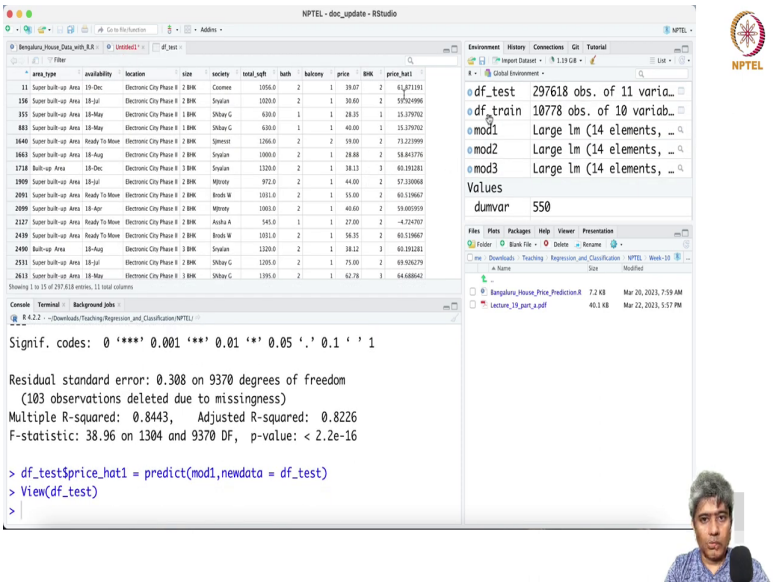
Now, we are getting almost 82 percent accuracy. So, we are reaching to the prior level if we have a correct accurate correctly done modeling.

(Refer Slide Time: 12:44)

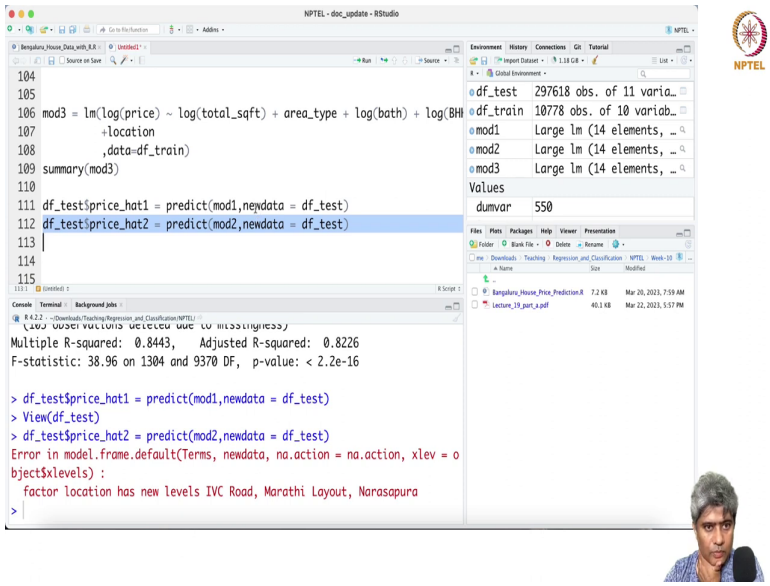


Now, I am going to do the prediction. So, df test equal to sorry, dollar sign price hat1 from the first model mod1 newdata equal to df test, ok.

(Refer Slide Time: 13:14)

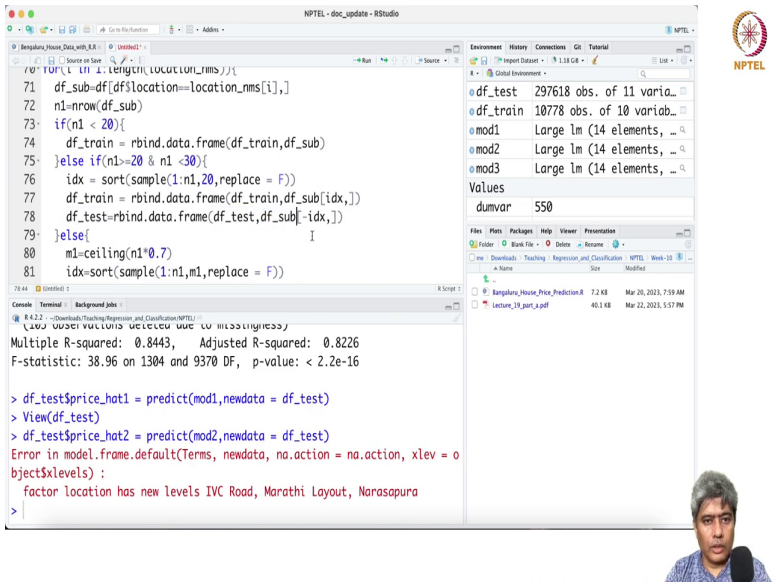


(Refer Slide Time: 13:26)



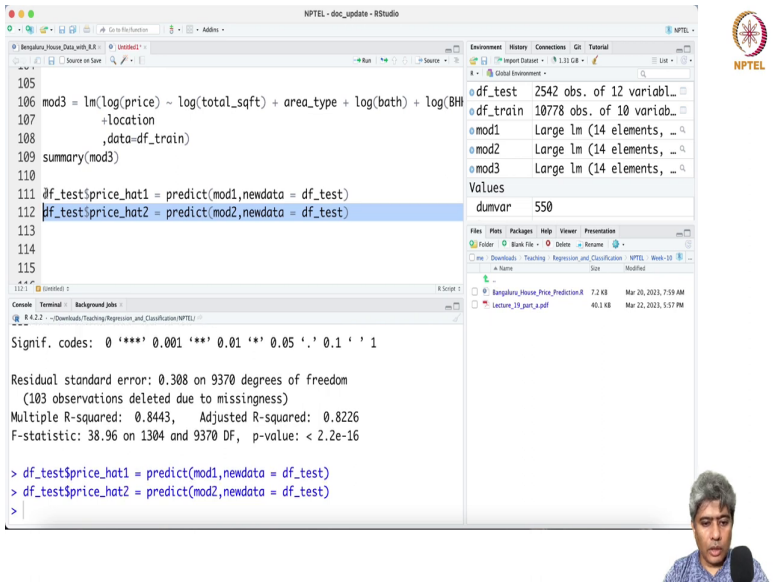
So, if I have the so, this is the first prediction, ok. So, this is the first prediction. And then we will have the second prediction from the second model. I think I have there must be some issue. Let me check the code. If I am not sure if I have done it correctly something wrong must be, ok. df train is (Refer Time: 13:56) sub oh, ok.

(Refer Slide Time: 14:14)



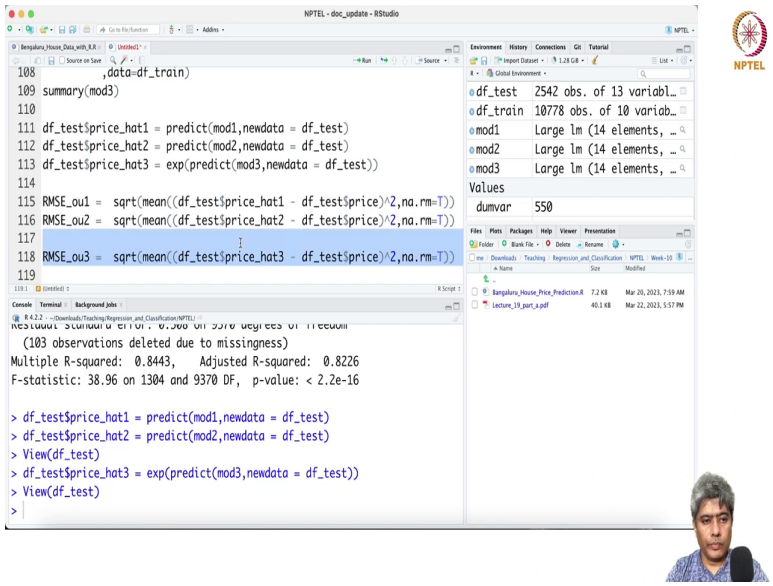
So, here is a mistake I can see controls need see df test. And it should be df sub, ok. So, this is the mistake that I made and then df test sub, yeah. Rest of the thing is fine. So, let me just run it once more. I hope let me just run it once more. So, these are the typical mistake I often make. But I hope mistake is divine. Some people say mistake is divine ok, 39 percent and let us see the model 2 ok 59 percent and let us run the model 3, 82 percent.

(Refer Slide Time: 15:02)



So, now it is fine. Now, you see there is no issue. So, we can run this df test, yeah, 62, 1131, 8 50, alright.

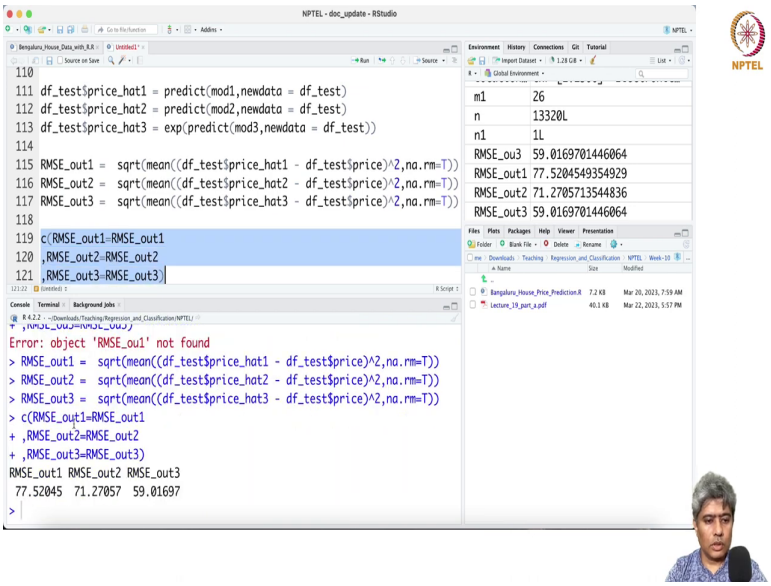
(Refer Slide Time: 15:20)



And finally, we just put 3, hat3 exp and model 3. If features run over that, I am sure test will have the third values which are looks like reasonably, ok, alright. So, we have some test values. And now what I am going to do is I am going to calculate the Root Mean Square Error RMSE out1 is equal to, alright. So, df test price 1 minus df test dollar price square.

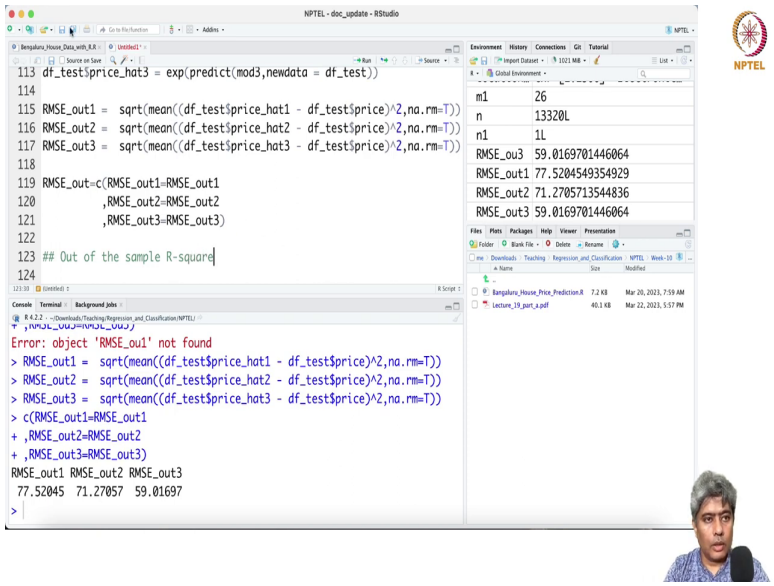
Take the mean. You have to take na dot rm equals to True because there could be some cases where you may get, na's if the all values are not available. Will see out2 will be based on price at 2 and see out3 will be based on price hat3, ok.

(Refer Slide Time: 17:09)



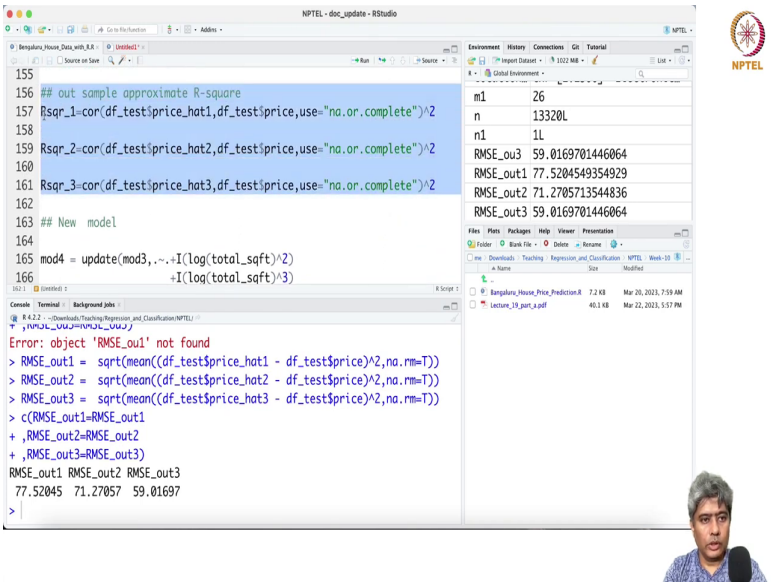
So, alright so, if I just say RMSE out equal to out1. 2 equal to out2 and 3 equal to 3 c. So, out, out, out, out, out. So, let us run this guy. And now so, this is the root mean square error 77, 71 and 59.

(Refer Slide Time: 18:11)



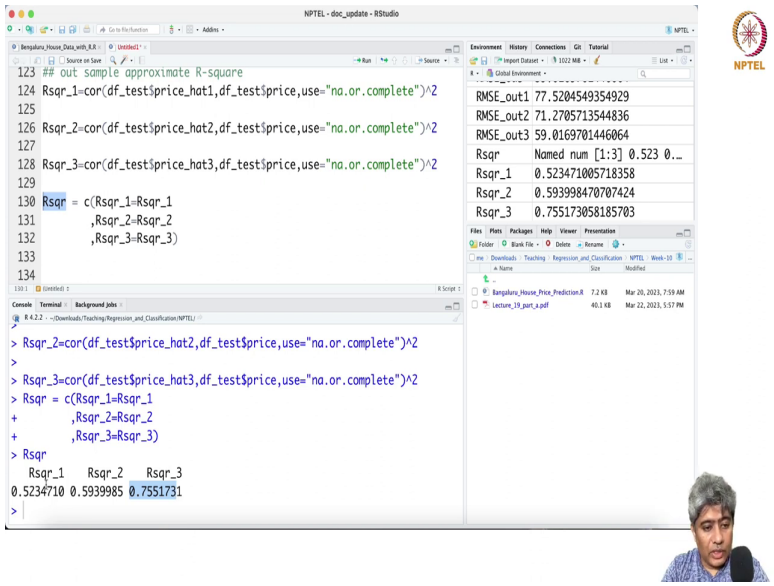
So, this is RMSE out sample RMSE, ok. So, let us, ok. We will do the next is we compute Out of the sample R-square out of the sample out of the sample R-square. I think what we can do. We can just take these questions. Take these things probably, yeah.

(Refer Slide Time: 18:54)



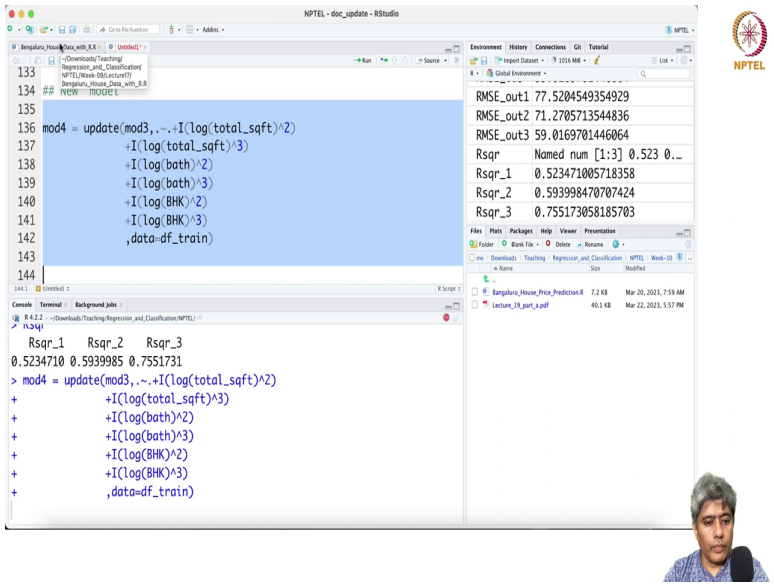
Out of the sample and then if we just compute this or see and I think R square 1, 2, 3. These are out of the sample R square.

(Refer Slide Time: 19:12)



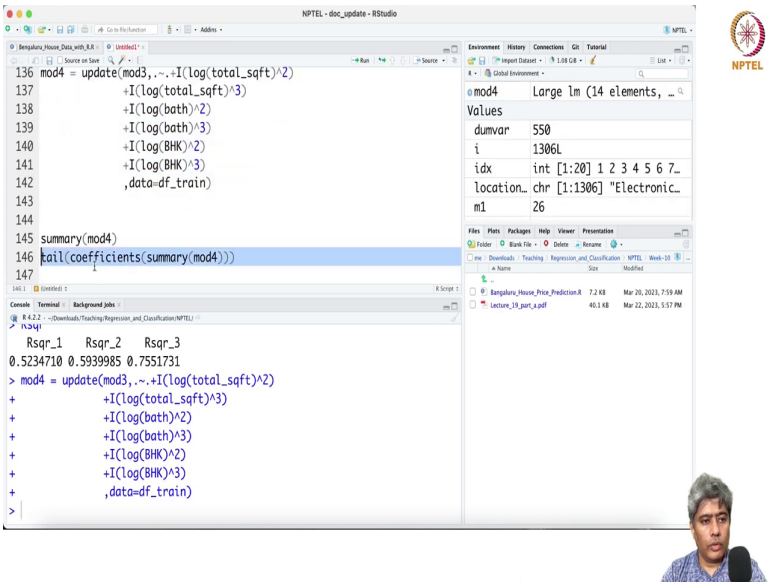
And 3 equal to 3 R square. So, out of the sample R square for the third model is near 75 percent where the in sample was near 82 percent. So, there is not much over fitting is happening, which is very good actually. Very decently looks like there is not much over fitting happening.

(Refer Slide Time: 20:04)

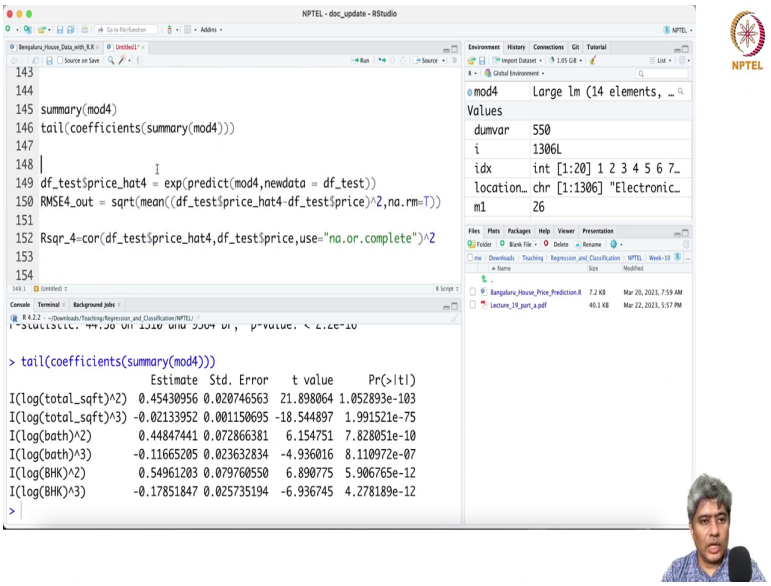


Now, we can choose some new models here with all the, you know, all the engineered feature. Let us run that.

(Refer Slide Time: 20:34)

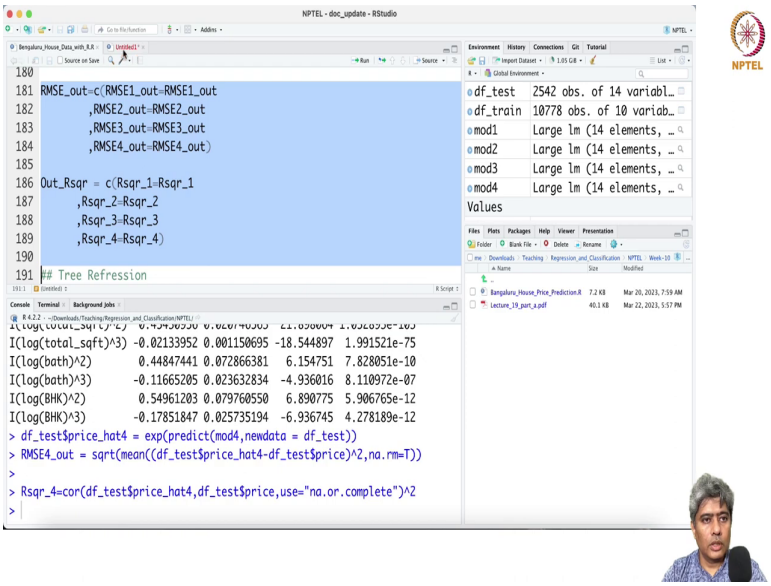


(Refer Slide Time: 20:40)

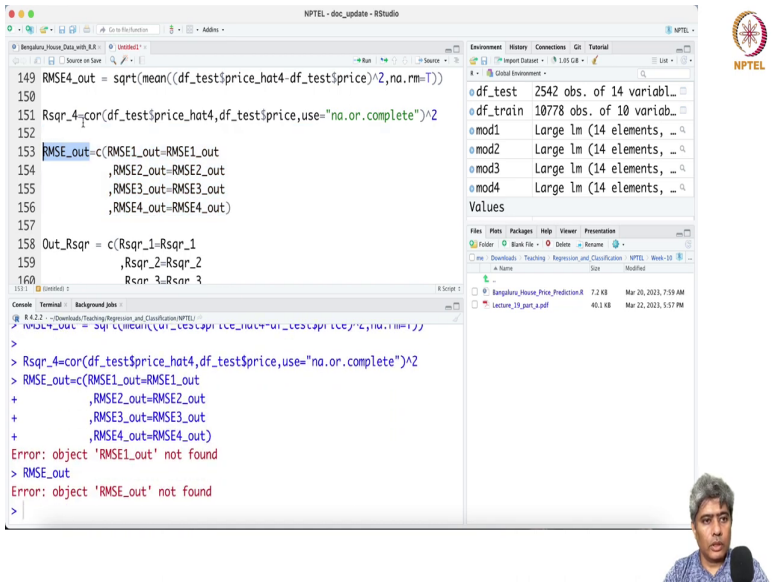


And if you run this so, these are the all if effectively total square feet square cube and bathroom, BHK they all have a quadratic and cubic kind of effect. So, we can just compute both these values, those out of the sample RMSE and R square for both.

(Refer Slide Time: 21:10)

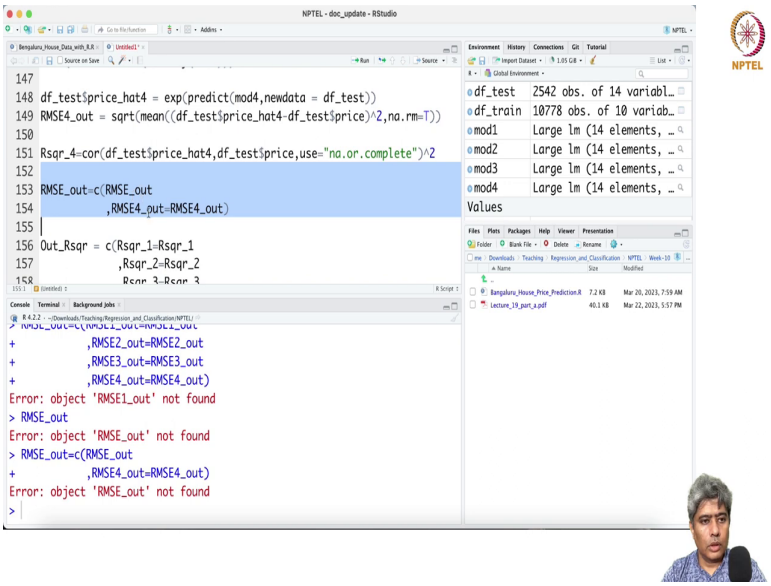


(Refer Slide Time: 21:21)

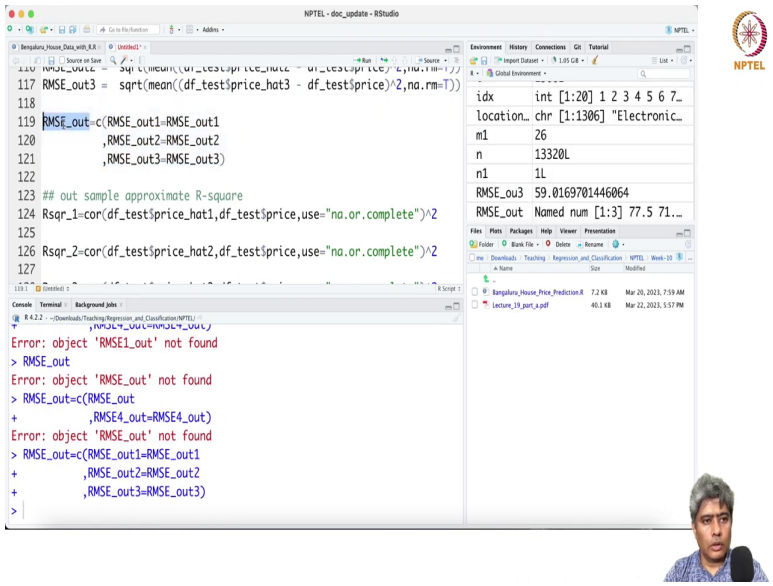


Both model 4 and then we just compute this 2. So, if I just run this out. So, this is my, oh, ok.

(Refer Slide Time: 21:37)

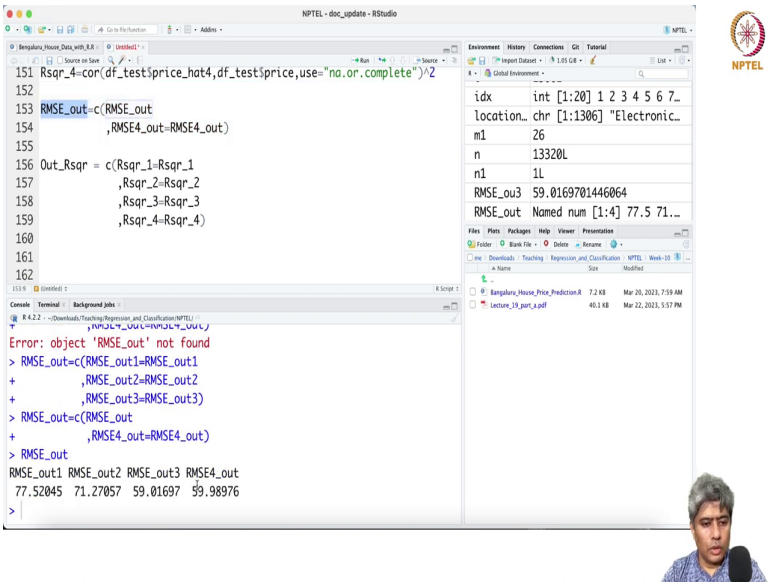


(Refer Slide Time: 21:55)



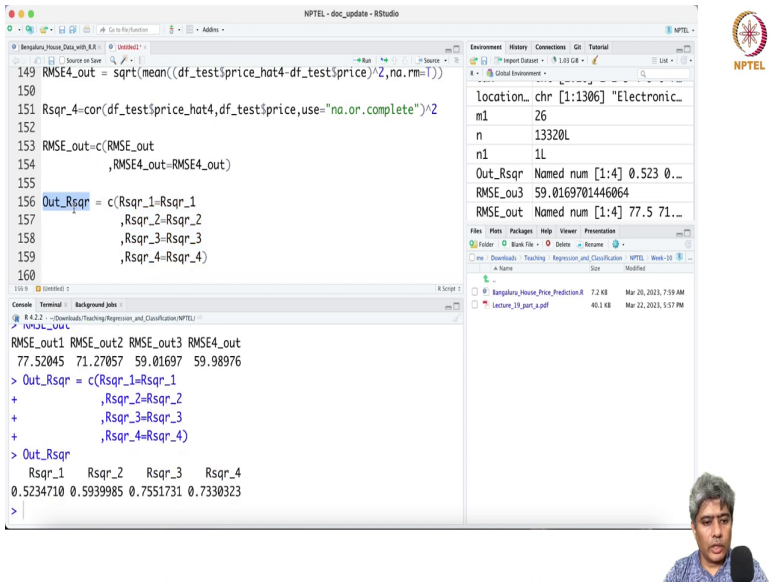
So, this is, I think we had a RMSE out. I think here I have a RMSE out equal to RMSE out comma this. Just a minute. I think I have this RMSE out (Refer Time: 21:58) RMSE out.

(Refer Slide Time: 22:04)



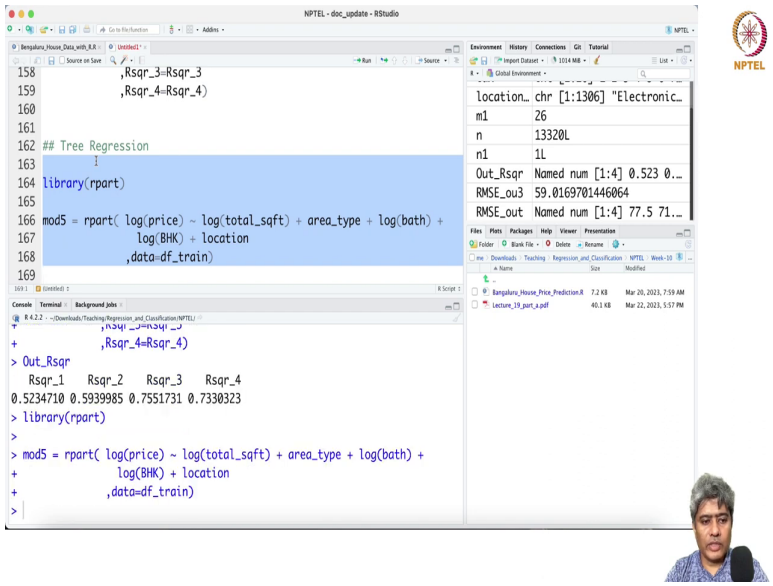
So, I can have this so, 59.98 and 59.01. So, looks like 3rd model is better than the 4th model.

(Refer Slide Time: 22:18)



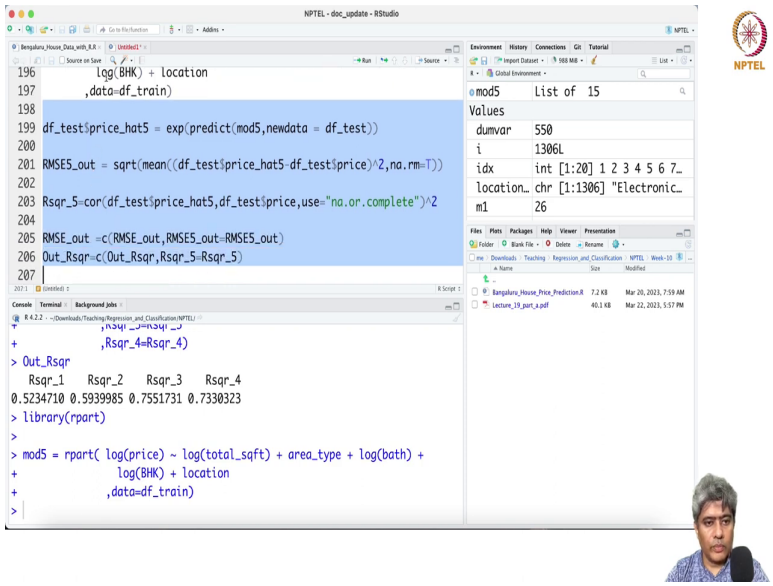
And let me out of the sample R square, let us look into this. Yeah, 3rd model looks like better than the 4th model. So, looks like third model is better. So, now we can have try the Tree Regression also, Tree Regression also.

(Refer Slide Time: 22:50)



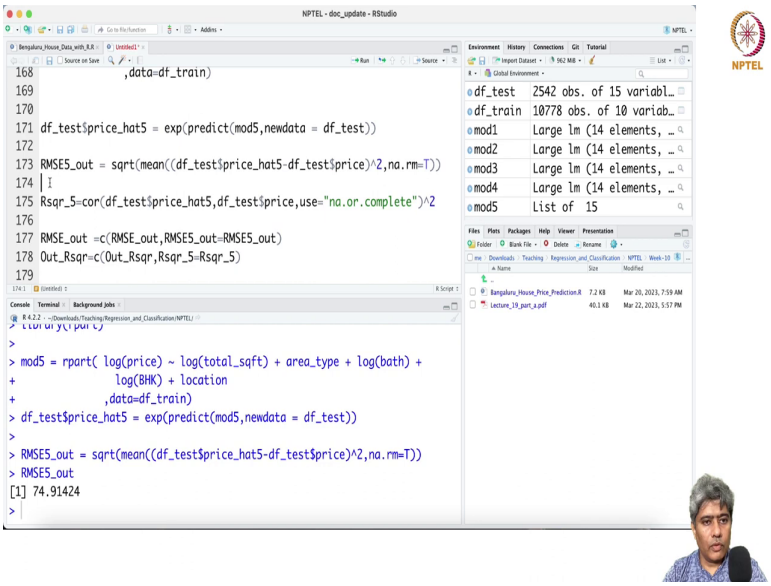
So, we can try Tree Regression as well here. I am not using again, random forest because it will take lot of time. We can try that, but I was trying that it was taking long time. So, I am not, I am refraining myself doing that, ok.

(Refer Slide Time: 23:12)

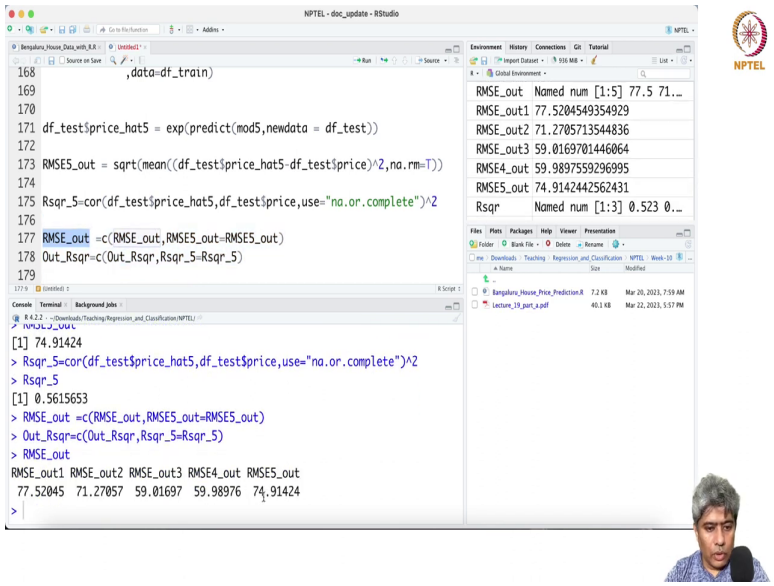


So, it is not like I cannot do that. It just need to, you have to wait long time for some reason my system it was taking just too much time.

(Refer Slide Time: 23:22)

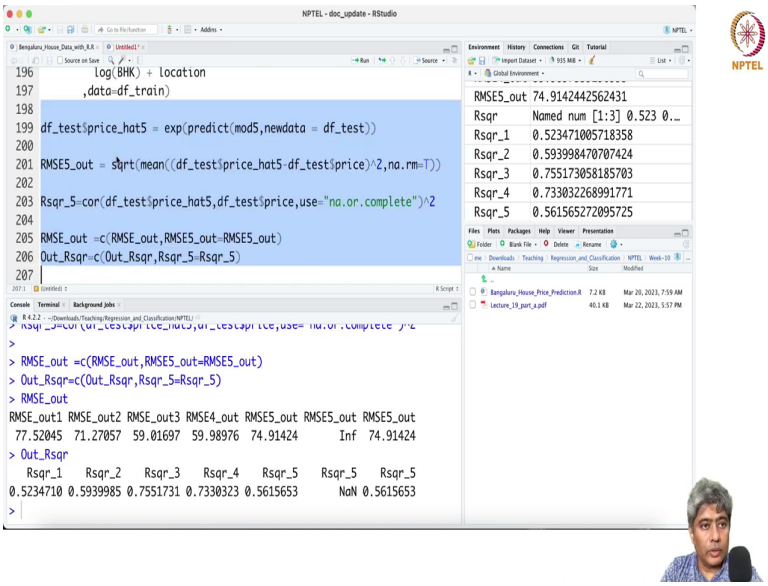


(Refer Slide Time: 23:34)



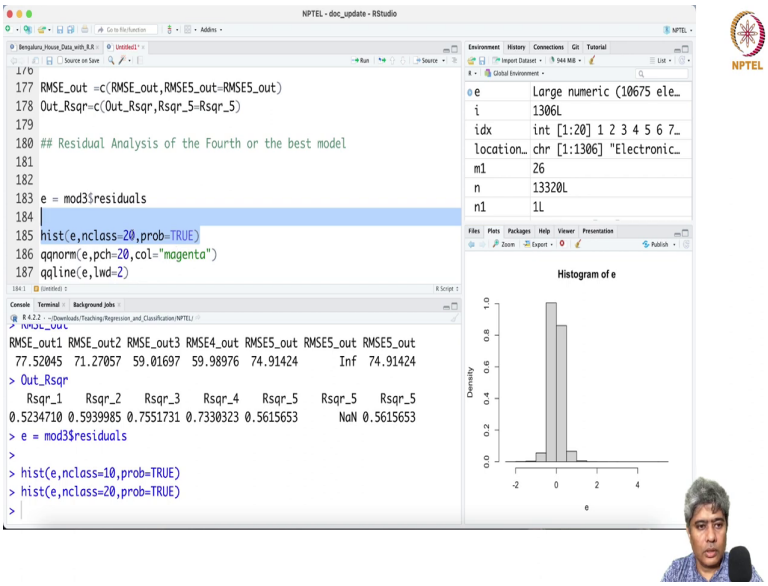
So, I am not trying that in this case, ok. So, RMSE out 74 and out of the sample R square is 56 if I just (Refer Time: 23:29) them. So, ok, so, Tree Regression is giving us 75 percent, it is not as good as the, you know, model 3 or model 4 after log, even after log transformation. We have taken log price, log total square feet and all those things. So, even there it is not giving decision trees, not giving tree regression is not giving that well.

(Refer Slide Time: 23:58)

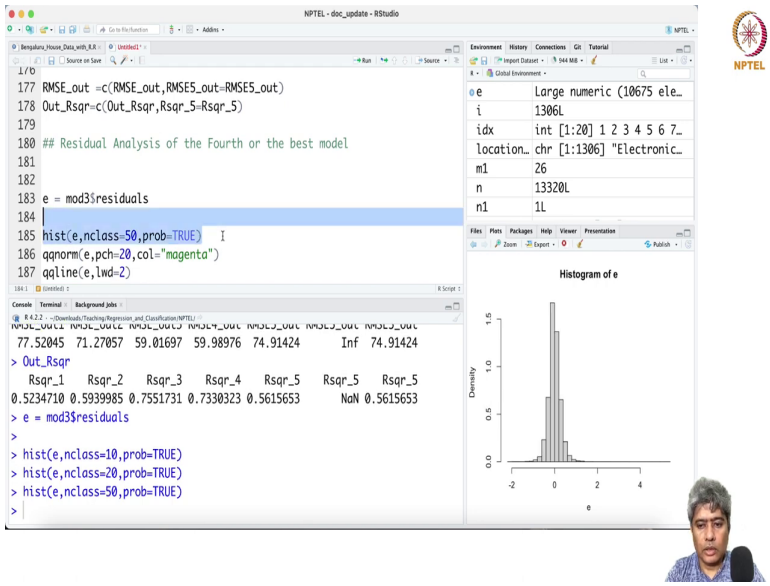


So, you have to be very careful so, alright and some residual analysis for the best model.

(Refer Slide Time: 24:26)

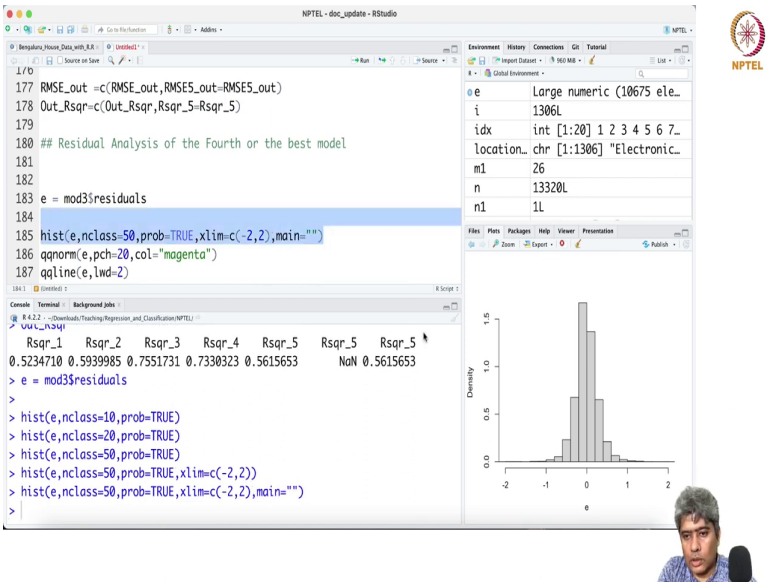


(Refer Slide Time: 24:39)



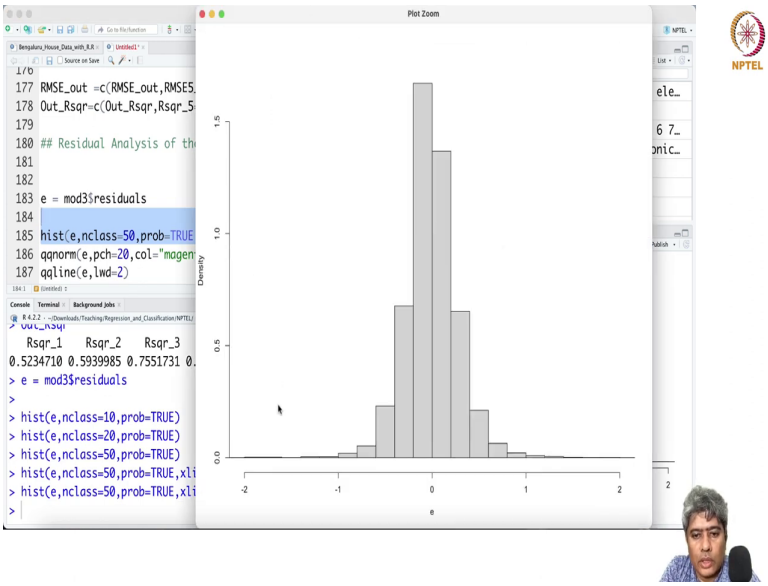
So, I have we have to do some residual analysis that if we run the residual analysis now, whatever that. So, residual analysis is from the 3rd model, we got from the best model, we got and the 3rd model. So, effectively if we just do 20, let us see yeah, or maybe 50 might be it will be better, yeah.

(Refer Slide Time: 24:43)

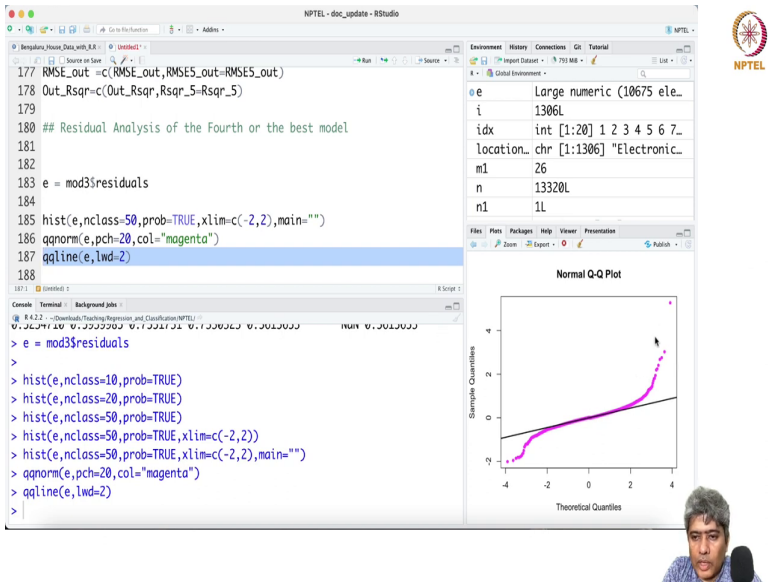


So, it looks like xlim equal to minus 2, 2. So, it is main equal to this.

(Refer Slide Time: 25:05)

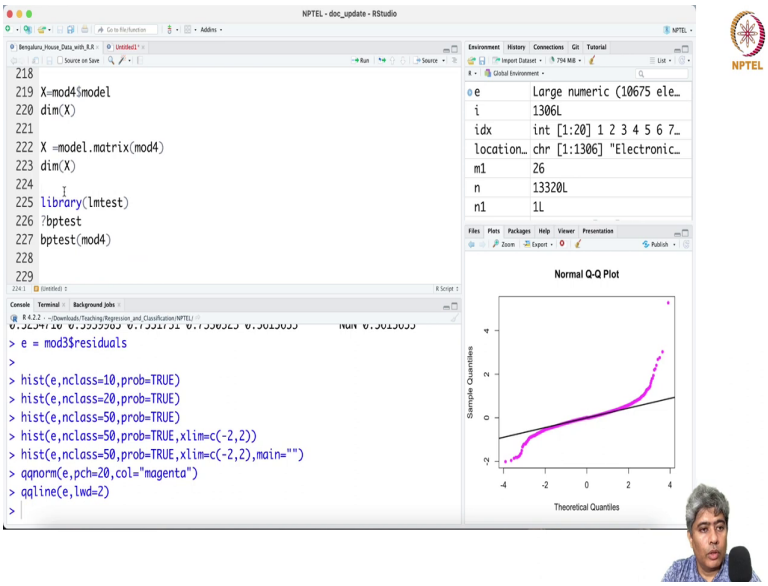


(Refer Slide Time: 25:16)

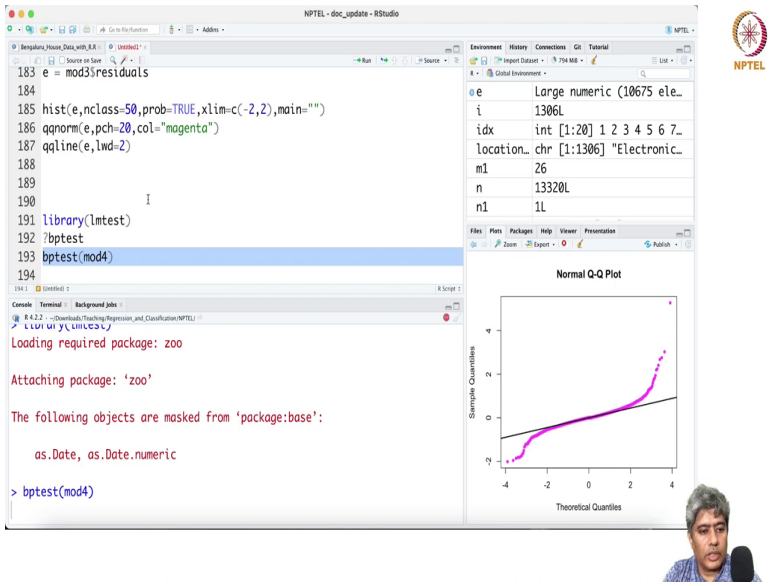


And if I just zoom it, we can see that it is somewhat behaves like a normal or bell shaped. But if we just do the qqnorm, we can see that it is not really, it has both side is quite heavy tail distribution in the middle, it is doing fine. That means there are quite a few cases where you are having very high under estimation or overestimation.

(Refer Slide Time: 25:46)

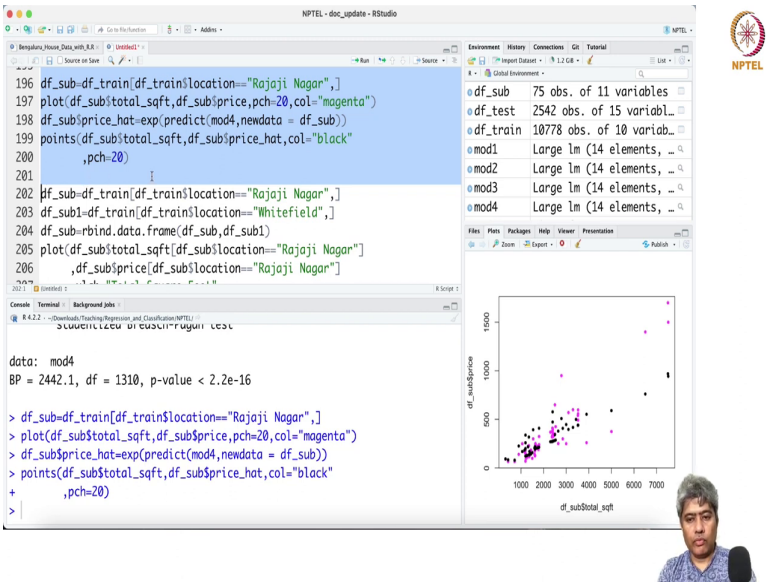


(Refer Slide Time: 25:59)

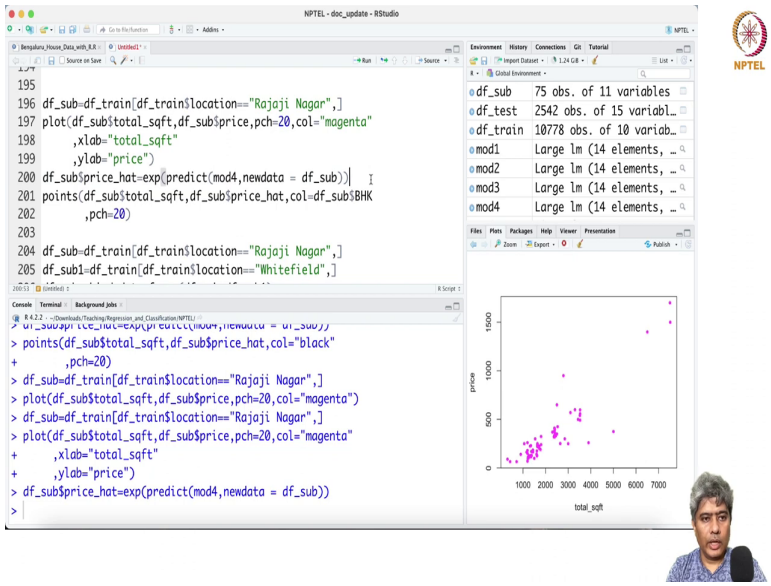


And I was also trying to do the you know, let me just all those bptest, Breusch Pagan test. I did those Breusch Pagan test tool and turns out Breusch Pagan test is was also not very good, yeah. And if I just do few more analysis for say, Rajaji Nagar, ok.

(Refer Slide Time: 26:50)

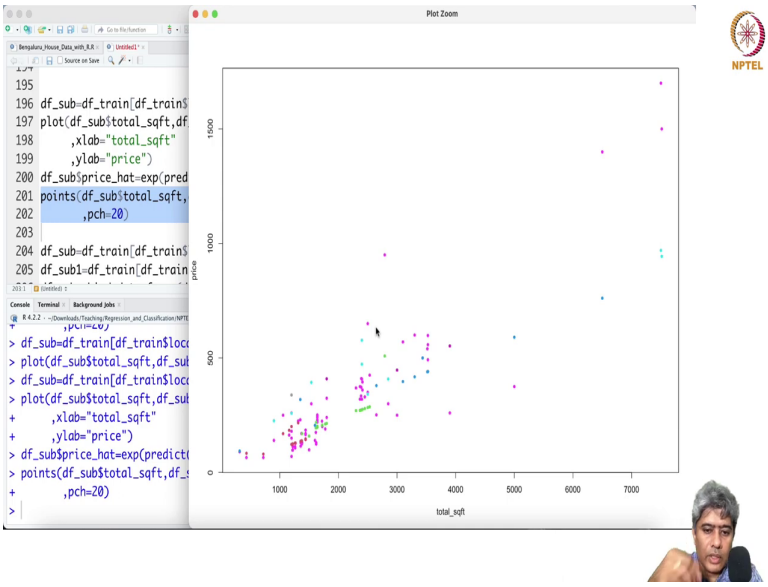


(Refer Slide Time: 27:10)



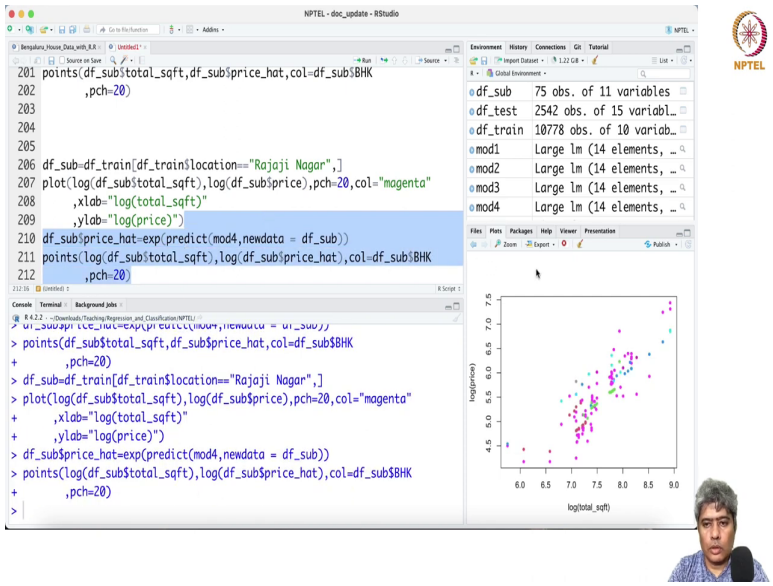
So, one thing I was thinking of that, instead of in the color, instead of giving black, I was what we can do, we can just say df sub dollar, whatever the BHK that we had. So, if we just run this in this way. So, these are the plots that we have. So, on the x axis, x label, we have the total square feet and on the y axis, y label will be price. So, if we just run it. So, this is total square feet versus price and then we have the price hat and if we just put a points.

(Refer Slide Time: 28:00)



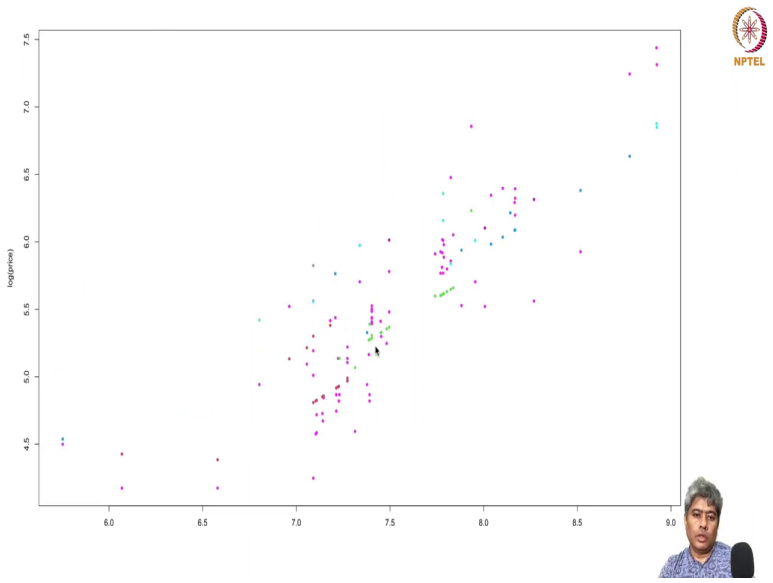
And now if he just say, you see this certain prices, different prices, you just have different value. Now, so, this is an interesting phenomena that we are seeing and then I think what we have is what we are trying, we can try to plot in log scale as well. So, probably that will be bit easier.

(Refer Slide Time: 28:39)



So, if we just put log scale and we can just say log of total square feet and log of price and then total square feet and price hat.

(Refer Slide Time: 29:24)



So, let me just zoom it. So, there are some rate cases, rate at the 2 BHK green, I think 3 BHK and blue at the 4 BHK and then. So, as people left from 2, 3, 4, then it is kind of went on different, different level. So, price for number of same number of the exact square feet, if you have more BHK, I think your price premium will go up; you have to pay a premium of the price for the number of BHK because functionality of the home goes up.

So, with this, I will stop here. I think now most of the issues of these data analysis is fixed. I am going to share this data, this code correct code on the Swayam portal of the NPTEL. And in the next video, we will begin with a new data analysis with a new real life data.

Thank you very much. See you in the next video.