

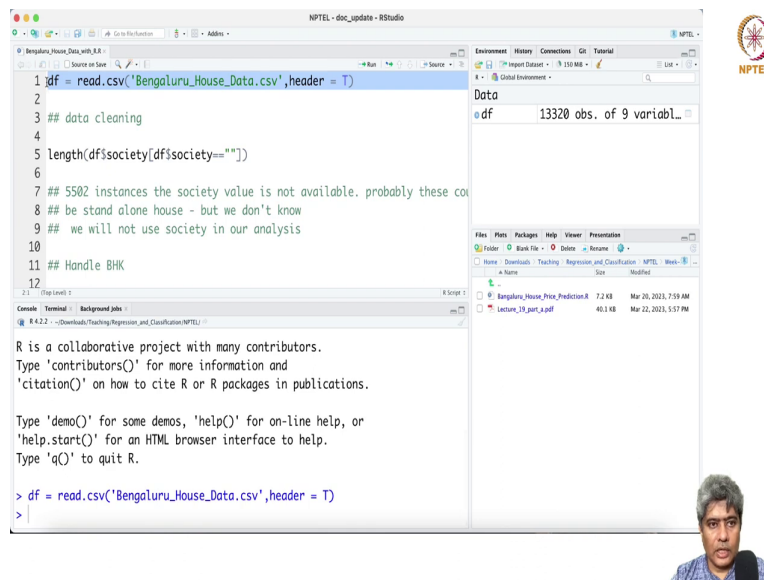
Predictive Analytics - Regression and Classification
Prof. Sourish Das
Department of Mathematics
Chennai Mathematical Institute

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)



The screenshot shows the RStudio interface with the following code in the editor:

```
1 df = read.csv('Bangaluru_House_Data.csv', header = T)
2
3 ## data cleaning
4
5 length(df$society[df$society==""])
6
7 ## 5502 instances the society value is not available. probably these co
8 ## be stand alone house - but we don't know
9 ## we will not use society in our analysis
10
11 ## Handle BHK
12
```

The console output shows the R help text for the 'read.csv' function:

```
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> df = read.csv('Bangaluru_House_Data.csv', header = T)
>
```

The Data Viewer on the right shows the structure of the data frame 'df':

Variable	Count
df	13320 obs. of 9 variabl...

The Files pane shows the following files:

Name	Size	Modified
Bangaluru_House_Price_Prediction.R	7.2 KB	Mar 20, 2023, 7:19 AM
Lecture_57_part_1.pdf	49.1 KB	Mar 22, 2023, 5:57 PM

The NPTEL logo is visible in the top right corner of the RStudio window.

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)

The screenshot shows the RStudio interface. The main window displays a data frame with 15 rows and 9 columns. The columns are: area_type, availability, location, size, society, total_sqft, bath, balcony, and price. The data includes various house listings with details like location (e.g., Electronic City Phase II, Chikka Tingathi), size (e.g., 2 BHK, 4 Bedroom), total_sqft (e.g., 1056, 2600), bath (e.g., 2, 5), balcony (e.g., 2, 3), and price (e.g., 39.07, 120.00).

The terminal window shows the following R code and output:

```
> df = read.csv('Bengaluru_House_Data.csv', header = T)
> View(df)
>
```

The terminal also displays the following text:

```

Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

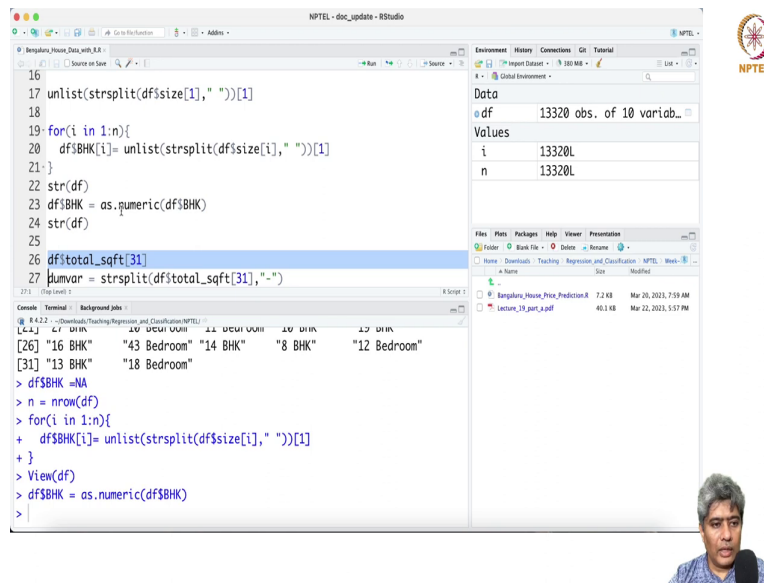
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

```

The RStudio interface also shows a file explorer on the right with a folder named 'Bengaluru_House_Price_Prediction.R' and a file named 'Lecture_19_part_1.pdf'.

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

(Refer Slide Time: 01:33)



The screenshot shows the RStudio interface with the following code in the script editor:

```
16
17 unlist(strsplit(df$size[1], " ")[1])
18
19 for(i in 1:n){
20   df$BHK[i]= unlist(strsplit(df$size[i], " ")[1])
21 }
22 str(df)
23 df$BHK = as.numeric(df$BHK)
24 str(df)
25
26 df$total_sqft[31]
27 plumar = strsplit(df$total_sqft[31], "-")
```


The console output shows the following steps and results:

```
> df$BHK =NA
> n = nrow(df)
> for(i in 1:n){
+   df$BHK[i]= unlist(strsplit(df$size[i], " ")[1])
+ }
> View(df)
> df$BHK = as.numeric(df$BHK)
>
```

The console also displays the following data for rows 26 and 31:

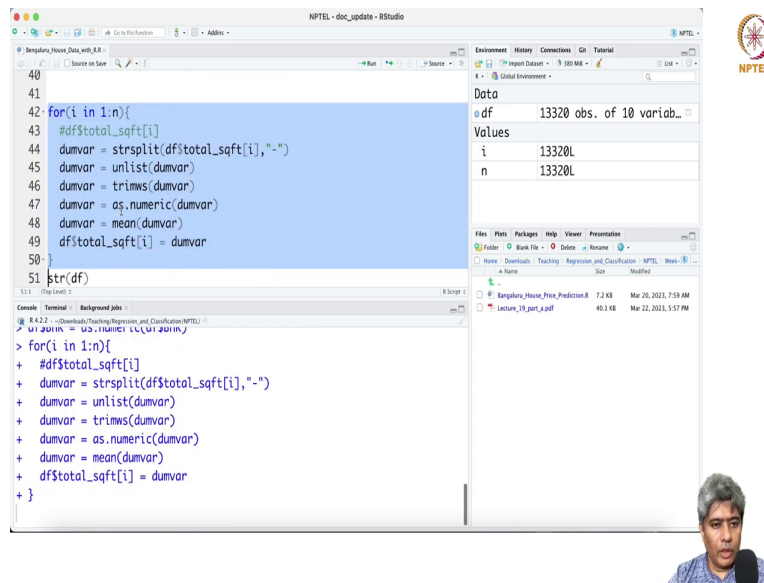
```
[26] "16 BHK"      "43 Bedroom" "14 BHK"      "8 BHK"      "12 Bedroom"
[31] "13 BHK"      "18 Bedroom"
```

The Environment pane on the right shows the data frame 'df' with 13320 observations and 10 variables. The 'Values' section shows 'i' as 13320L and 'n' as 13320L. The Files pane shows the current project files, including 'Bangalore_House_Price_Prediction.R' and 'Lecture_19_part_1.pdf'.



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)



The screenshot displays the RStudio interface. The main editor window contains the following R code:

```
40  
41  
42 for(i in 1:n){  
43   #df$total_sqft[i]  
44   dumvar = strsplit(df$total_sqft[i], "-")  
45   dumvar = unlist(dumvar)  
46   dumvar = trimws(dumvar)  
47   dumvar = as.numeric(dumvar)  
48   dumvar = mean(dumvar)  
49   df$total_sqft[i] = dumvar  
50 }  
51 str(df)
```


The console window shows the execution of the code:

```
51:1 str(df)  
> str(df)  
> for(i in 1:n){  
+ #df$total_sqft[i]  
+ dumvar = strsplit(df$total_sqft[i], "-")  
+ dumvar = unlist(dumvar)  
+ dumvar = trimws(dumvar)  
+ dumvar = as.numeric(dumvar)  
+ dumvar = mean(dumvar)  
+ df$total_sqft[i] = dumvar  
+ }  
[1] 13320L  
[1] 13320L
```

The Environment pane on the right shows a data frame 'df' with 13320 observations and 10 variables. The 'Values' section shows the first two rows of the data frame:

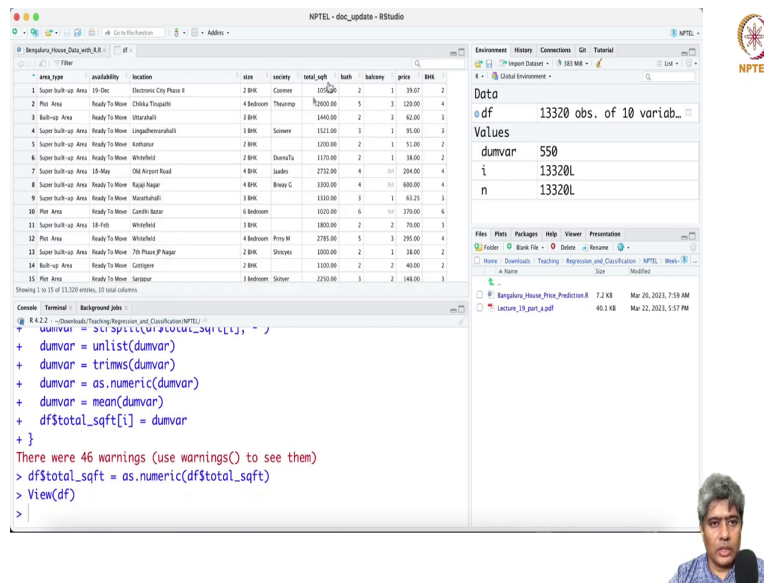
Variable	Value
i	13320L
n	13320L

The Files pane shows a folder structure with files like 'Bangalore_House_Price_Prediction.R' and 'Lecture_19_part_1.pdf'.



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

(Refer Slide Time: 02:08)



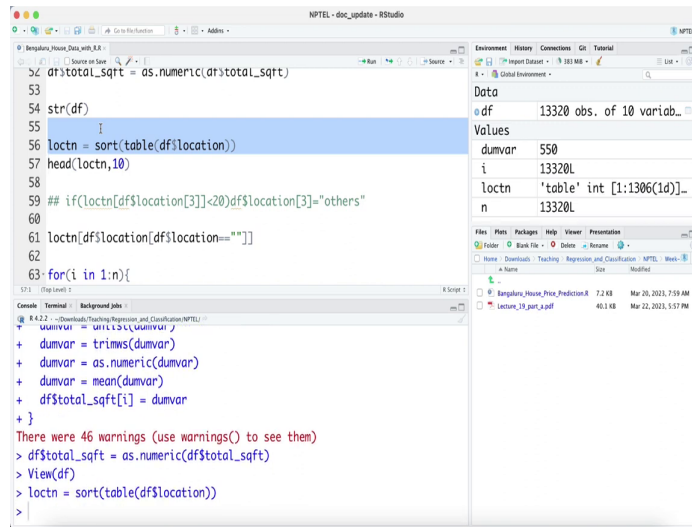
The screenshot displays the RStudio interface with the following components:

- Environment Pane:** Shows the loaded data frame 'df' with 13320 observations and 10 variables.
- Console:** Contains the following R code and output:

```
df$total_sqft = strtoi(df$total_sqft, -)
+ dumvar = unlist(dumvar)
+ dumvar = trimws(dumvar)
+ dumvar = as.numeric(dumvar)
+ dumvar = mean(dumvar)
+ df$total_sqft[i] = dumvar
+ }
There were 46 warnings (use warnings() to see them)
> df$total_sqft = as.numeric(df$total_sqft)
> View(df)
>
```
- Table:** A data frame with columns: area_type, availability, location, size, society, total_sqft, bath, balcony, price, BHK. It lists 15 different property listings.
- Environment:** Lists files like 'Bangalore_House_Price_Prediction.R' and 'Lecture_19_part_1.pdf'.

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

(Refer Slide Time: 02:18)



The screenshot shows the RStudio interface with the following content:

```
df$total_sqft = as.numeric(df$total_sqft)
53
54 str(df)
55
56 loctn = sort(table(df$location))
57 head(loctn,10)
58
59 ## if(loctn[df$location[3]]<20)df$location[3]="others"
60
61 loctn[df$location[df$location==""]]
62
63 for(i in 1:n){
```

Console output:

```
R 4.2.2 -- Command: Training Regression and Classification NPTEL
> dumvar = list(dumvar)
+ dumvar = trimws(dumvar)
+ dumvar = as.numeric(dumvar)
+ dumvar = mean(dumvar)
+ df$total_sqft[i] = dumvar
+ }
There were 46 warnings (use warnings() to see them)
> df$total_sqft = as.numeric(df$total_sqft)
> View(df)
> loctn = sort(table(df$location))
>
```

Environment pane:

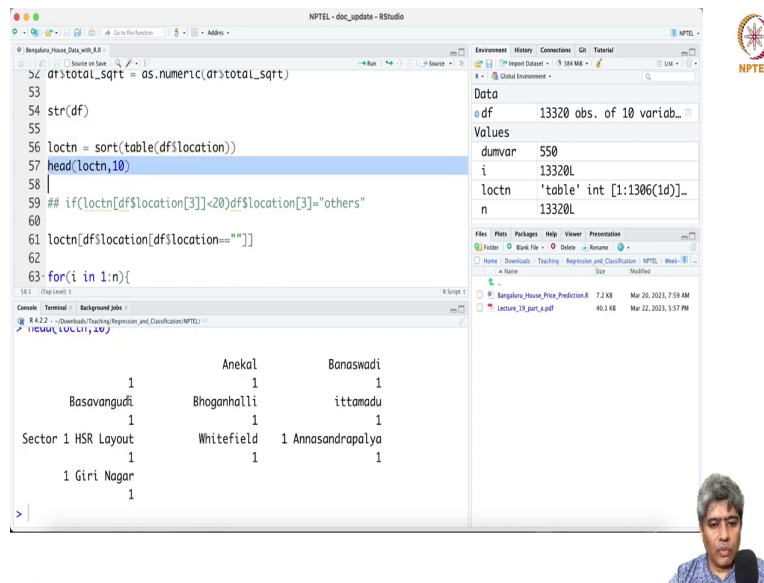
Data	
df	13320 obs. of 10 variab...
Values	
dumvar	550
i	13320L
loctn	'table' int [1:1306(1d)]...
n	13320L

Files pane:

Name	Type	Modified
Bangluru_House_Price_Prediction.R	7.2 KB	Mar 20, 2023, 7:59 AM
Lecture_19_part_1.pdf	40.1 KB	Mar 22, 2023, 5:57 PM



(Refer Slide Time: 02:27)

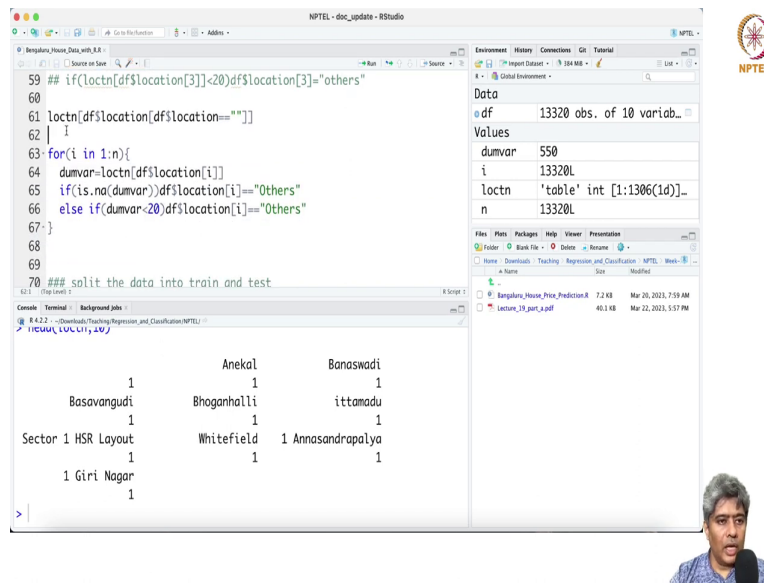


```
df$total_sqrt = as.numeric(df$total_sqrt)
53
54 str(df)
55
56 loctn = sort(table(df$location))
57 head(loctn,10)
58
59 # if(loctn[df$location[3]]<20)df$location[3]="others"
60
61 loctn[df$location[df$location==""]]
62
63 for(i in 1:n){
```

	Anekal	Banaswadi
1	1	1
Basavangudi	Bhoganhalli	ittamadu
1	1	1
Sector 1 HSR Layout	Whitefield	1 Annasandrapalya
1	1	1
1 Giri Nagar		
1		

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)



```
59 ## if(loctn[df$location[3]]<20)df$location[3]="Others"
60
61 loctn[df$location[df$location==""]]
62 |
63 for(i in 1:n){
64   dumvar=loctn[df$location[i]]
65   if(is.na(dumvar))df$location[i]="Others"
66   else if(dumvar<20)df$location[i]="Others"
67 }
68
69
70 ### split the data into train and test
```

	Anekal	Banaswadi
1	1	1
Basavangudi	Bhoganhalli	ittamadu
1	1	1
Sector 1 HSR Layout	Whitefield	1 Annasandrapalya
1	1	1
1 Giri Nagar		
1		

NPTEL logo



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)

```
NPTEL - doc_update - RStudio  
Bangalore_House_Data_with_X3.R  
59 ## if(loctn[df$location[3]]<20)df$location[3]="others"  
60  
61 loctn[df$location[df$location==""]]   
62  
63 for(i in 1:n){  
64   dumvar=loctn[df$location[i]]  
65   if(is.na(dumvar))df$location[i]="Others"  
66   else if(dumvar<20)df$location[i]="Others"  
67 }  
68  
69  
70 ### split the data into train and test  
68.1 (Stop Level) :  
Sector 1 HSR Layout      Whitefield  1 Annasandrapalya  
      1  
      1 Giri Nagar  
      1  
> for(i in 1:n){  
+   dumvar=loctn[df$location[i]]  
+   if(is.na(dumvar))df$location[i]="Others"  
+   else if(dumvar<20)df$location[i]="Others"  
+ }  
>
```



(Refer Slide Time: 03:15)

The screenshot shows the RStudio interface with the following components:

- Environment Pane:** Shows a data frame named 'df' with 13320 observations and 10 variables.
- Console:** Displays the following R code and its output:

```
1 1 1 1
1 Giri Nagar
1
> for(i in 1:n){
+   dumvar=loctn[df$location[i]]
+   if(is.na(dumvar))df$location[i]=="Others"
+   else if(dumvar<20)df$location[i]=="Others"
+ }
> View(df)
```
- Table:** A data table with columns: area_type, availability, location, size, society, total_sqft, bath, balcony, price, BHK. It lists 15 different property listings.
- Environment:** Lists files like 'Bangalore_House_Price_Prediction.R' and 'Lecture_19_part_1.pdf'.

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.

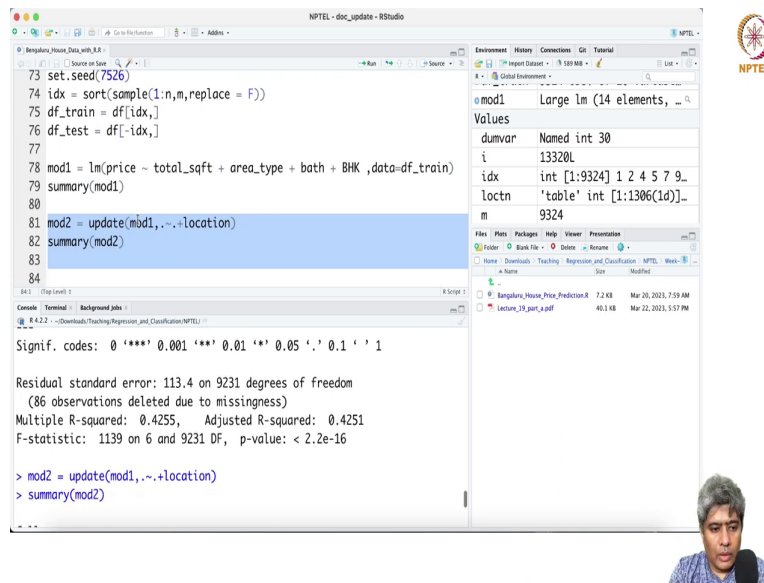
(Refer Slide Time: 03:33)

```
69
70 ## split the data into train and test
71
72 m = ceiling(n*0.7)
73 set.seed(7526)
74 idx = sort(sample(1:n,m,replace = F))
75 df_train = df[idx,]
76 df_test = df[-idx,]
77
78 mod1 = lm(price ~ total_sqft + area_type + bath + BHK ,data=df_train)
79 summary(mod1)
80
81
```

```
## 4.4.22 - Overview Training Regression and Classification (NPTEL)
> dumvar = table(df$location)
+ if(is.na(dumvar))df$location[i]="Others"
+ else if(dumvar<20)df$location[i]="Others"
+ }
> View(df)
> m = ceiling(n*0.7)
> set.seed(7526)
> idx = sort(sample(1:n,m,replace = F))
> df_train = df[idx,]
> df_test = df[-idx,]
>
```



(Refer Slide Time: 03:44)



The screenshot shows the RStudio interface with the following code in the editor:

```
73 set.seed(7526)
74 idx = sort(sample(1:n,m,replace = F))
75 df_train = df[idx,]
76 df_test = df[-idx,]
77
78 mod1 = lm(price ~ total_sqft + area_type + bath + BHK ,data=df_train)
79 summary(mod1)
80
81 mod2 = update(mod1,~. +location)
82 summary(mod2)
83
84
```

The console output shows the following summary statistics for the model:

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 113.4 on 9231 degrees of freedom
(86 observations deleted due to missingness)
Multiple R-squared:  0.4255,    Adjusted R-squared:  0.4251
F-statistic: 1139 on 6 and 9231 DF,  p-value: < 2.2e-16
```

The environment pane shows the following variables:

Variable	Value
mod1	Large lm (14 elements, ...)
dumvar	Named int 30
i	13320L
idx	int [1:9324] 1 2 4 5 7 9...
loctn	'table' int [1:1306(1d)]...
m	9324

The file explorer shows the following files:

Name	Type	Modified
Bangalore_House_Price_Prediction.R	7.2 KB	Mar 20, 2023, 7:59 AM
Lecture_19_part_1.pdf	40.1 KB	Mar 22, 2023, 5:57 PM

A small video thumbnail of a man is visible in the bottom right corner of the RStudio window.

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)

The screenshot shows the RStudio interface with the following content:

```
80  
81 mod2 = update(mod1, ~. + location)  
82 summary(mod2)  
83  
84  
85 test_data = df_test[1,]  
86 test_data[1, 'location'] = "Rajaji Nagar"  
87 test_data[1, 'total_sqft'] = 1000  
88 test_data[1, 'bath'] = 2  
89 test_data[1, 'BHK'] = 2  
90 predict(mod2, newdata = test_data)  
91 predict(mod1, newdata = test_data)
```

Environment: mod2 Large lm (14 elements, ...)

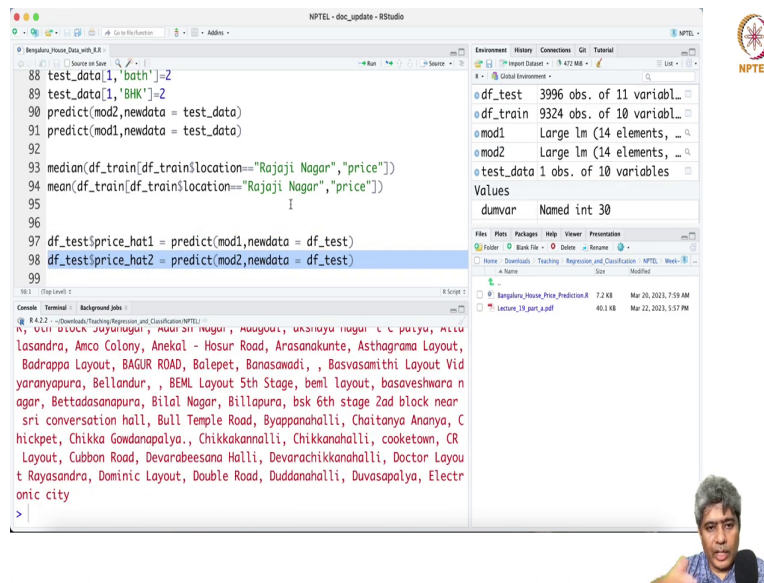
Values	
dumvar	Named int 30
i	13320L
idx	int [1:9324] 1 2 4 5 7 9_
loctn	'table' int [1:1306(1d)]_
m	9324

Console:

```
[ reached getOption("max.print") -- omitted 946 rows ]  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 91.1 on 8092 degrees of freedom  
(86 observations deleted due to missingness)  
Multiple R-squared:  0.675,    Adjusted R-squared:  0.629  
F-statistic: 14.68 on 1145 and 8092 DF,  p-value: < 2.2e-16  
  
>
```



(Refer Slide Time: 04:09)



The screenshot shows an RStudio window titled 'NPTEL - doc_update - RStudio'. The script editor contains the following R code:

```
88 test_data[1,'bath']=2
89 test_data[1,'BHK']=2
90 predict(mod2,newdata = test_data)
91 predict(mod1,newdata = test_data)
92
93 median(df_train[df_train$location=="Rajaji Nagar","price"])
94 mean(df_train[df_train$location=="Rajaji Nagar","price"])
95
96
97 df_test$price_hat1 = predict(mod1,newdata = df_test)
98 df_test$price_hat2 = predict(mod2,newdata = df_test)
99
```

The Environment pane on the right shows the following objects:

- df_test: 3996 obs. of 11 variables
- df_train: 9324 obs. of 10 variables
- mod1: Large lm (14 elements)
- mod2: Large lm (14 elements)
- test_data: 1 obs. of 10 variables
- Values: dumvar Named int 30

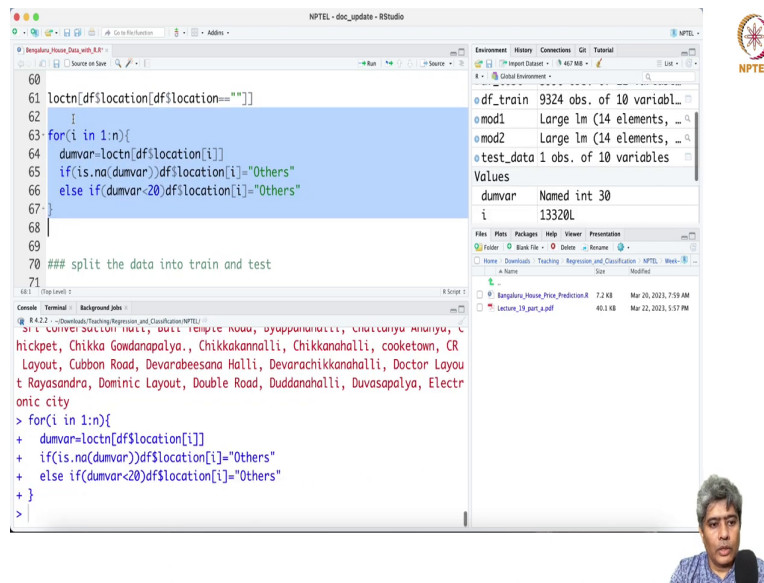
The Console pane shows the following output:

```
R 4.2.2 -- Download Training Regression and Classification NPTEL
Rajaji Nagar, Hosur Road, Anasankunte, Asthagrama Layout,
Lasandra, Amco Colony, Anekal - Hosur Road, Anasankunte, Asthagrama Layout,
Badrappa Layout, BAGUR ROAD, Balepet, Banasawadi, , Basvasamithi Layout Vid
yaranyapura, Bellandur, , BEML Layout 5th Stage, beml layout, basaveshwara n
agar, Bettadasanapura, Bilal Nagar, Billapura, bsk 6th stage 2ad block near
sri conversation hall, Bull Temple Road, Byappanahalli, Chaitanya Anyana, C
hickpet, Chikka Gowdanapalya., Chikkakannalli, Chikkanahalli, cooaketown, CR
Layout, Cubbon Road, Devarabeesana Halli, Devarachikkanahalli, Doctor Layou
t Rayasandra, Dominic Layout, Double Road, Duddanahalli, Duvasapalya, Electr
onic city
>
```

The NPTEL logo is visible in the top right corner of the RStudio window.

And then we did some test of the values and for Rajaji Nagar and all. And then we did an 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)



The screenshot shows the RStudio interface with the following components:

- Source Editor:** Contains R code for cleaning the 'location' variable in a dataset named 'df'.

```
60 loctn[df$location[df$location==""]]
61
62 for(i in 1:n){
63   dumvar=loctn[df$location[i]]
64   if(is.na(dumvar))df$location[i]="Others"
65   else if(dumvar<20)df$location[i]="Others"
66 }
67
68
69
70 ## split the data into train and test
71
```
- Environment:** Shows the current environment with objects: 'df_train' (9324 obs. of 10 variables), 'mod1' (Large lm (14 elements)), 'mod2' (Large lm (14 elements)), and 'test_data' (1 obs. of 10 variables). The 'Values' section shows 'dumvar' as a Named integer vector of length 30, with the value '13320L'.
- Console:** Shows the execution of the code from the source editor, with the output of the 'for' loop visible.
- Files:** Shows the project files, including 'Bangalore_House_Price_Prediction.R' and 'Lecture_19_part_1.pdf'.
- NPTEL Logo:** Located in the top right corner of the RStudio window.
- Video Feed:** A small video feed of the presenter is visible in the bottom right corner.

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)

```
NPTEL - doc_update - RStudio
Bangalore_House_Data_with_BHK.R
73 set.seed(7526)
74 idx = sort(sample(1:n,m,replace = F))
75 df_train = df[idx,]
76 df_test = df[-idx,]
77
78 mod1 = lm(price ~ total_sqft + area_type + bath + BHK ,data=df_train)
79 summary(mod1)
80
81 mod2 = update(mod1, ~. + location)
82 summary(mod2)
83
84
```

```
Console Terminal Background Jobs
R 4.2.2 --Downloads/Teaching/Regression_and_Classification/NPTEL
DUULH
BHK
-5.81380 2.10375 -2.764 0.00573 **
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 113.4 on 9231 degrees of freedom
(86 observations deleted due to missingness)
Multiple R-squared: 0.4255, Adjusted R-squared: 0.4251
F-statistic: 1139 on 6 and 9231 DF, p-value: < 2.2e-16
```

Environment History Connections GR Tutorial
Project Status 9.312 MB
Global Environment

mod1 Large lm (14 elements, ...
mod2 Large lm (14 elements, ...
test_data 1 obs. of 10 variables

Values
dumvar Named int 30
i 13320L
idx int [1:9324] 1 2 4 5 7 9...

Files Plots Packages Help Viewer Presentations
Folder Bank File Delete Rename
Home Downloads Teaching Regression_and_Classification NPTEL Week-6
Name Type Modified
Bangalore_House_Price_Prediction.R 7.2 KB Mar 20, 2022, 7:59 AM
Lecture_19_part_1.pdf 40.3 KB Mar 22, 2022, 5:57 PM



(Refer Slide Time: 04:53)

The screenshot shows an RStudio session with the following code and output:

```
73 set.seed(7526)
74 idx = sort(sample(1:n,m,replace = F))
75 df_train = df[idx,]
76 df_test = df[-idx,]
77
78 mod1 = lm(price ~ total_sqft + area_type + bath + BHK ,data=df_train)
79 summary(mod1)
80
81 mod2 = update(mod1, ~. + location)
82 summary(mod2)
83
84
```

The console output for `summary(mod1)` is:

```
LocationYeshwanthpur      0.29638
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 110.5 on 9083 degrees of freedom
(86 observations deleted due to missingness)
Multiple R-squared:  0.4637,    Adjusted R-squared:  0.4546
F-statistic:  51 on 154 and 9083 DF,  p-value: < 2.2e-16
```


The environment pane shows:

- `mod1`: Large lm (14 elements)
- `mod2`: Large lm (14 elements)
- `test_data`: 1 obs. of 10 variables

The console also shows the values for `df_test`:

```
Values
dumvar Named int 30
i       13320L
idx     int [1:9324] 1 2 4 5 7 9...
```

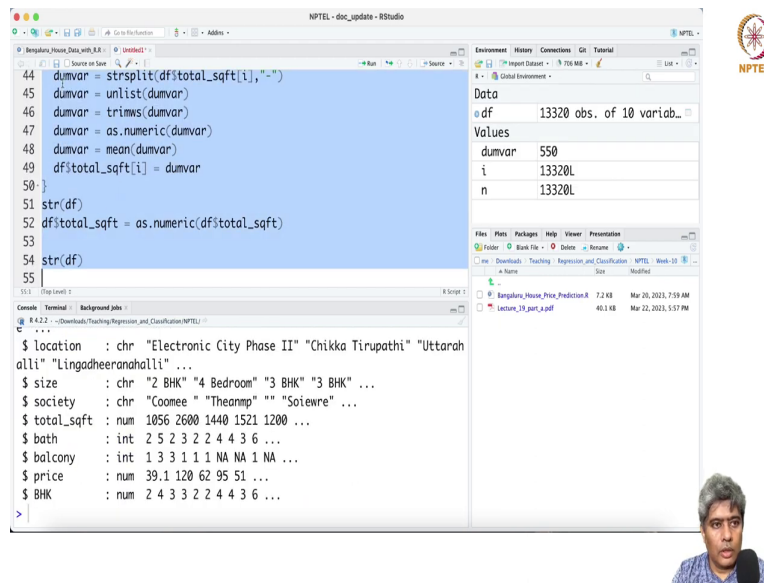
The file explorer shows a folder named "Regression_and_Classification/NPTEL/Week-8" containing files like "Bangalore_House_Price_Prediction.R" and "Lecture_19_part_1.pdf".



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)





The screenshot shows the RStudio interface with the following content:

```
44 dumvar = strsplit(df$total_sqft[i], "-")
45 dumvar = unlist(dumvar)
46 dumvar = trimws(dumvar)
47 dumvar = as.numeric(dumvar)
48 dumvar = mean(dumvar)
49 df$total_sqft[i] = dumvar
50 }
51 str(df)
52 df$total_sqft = as.numeric(df$total_sqft)
53
54 str(df)
55
```

The console output shows the structure of the data frame:

```
$ location : chr "Electronic City Phase II" "Chikka Tirupathi" "Uttarahalli" "Lingadheeranahalli" ...
$ size : chr "2 BHK" "4 Bedroom" "3 BHK" "3 BHK" ...
$ society : chr "Coomee" "Theanmp" "" "Soiewre" ...
$ total_sqft : num 1056 2600 1440 1521 1200 ...
$ bath : int 2 5 2 3 2 2 4 4 3 6 ...
$ balcony : int 1 3 3 1 1 1 NA NA 1 NA ...
$ price : num 39.1 120 62.95 51 ...
$ BHK : num 2 4 3 3 2 2 4 4 3 6 ...
```

The Environment pane shows the data frame 'df' with 13320 observations and 10 variables. The 'Values' pane shows the values for 'dumvar' (550), 'i' (13320L), and 'n' (13320L).



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)

The screenshot shows the RStudio interface with the following R code in the editor:

```
106     ,data=df_train)
107 summary(mod3)
108 exp(predict(mod3,newdata = test_data))
109
110 ##-----
111 ## Random stratified sampling strategy
112
113 location_rms=unique(df$location)
114 df_train=df[1,]
115 df_test = df[1,]
116 set.seed(7526)
117
```

The console output shows the following variable types:

```
$ location : chr "Electronic City Phase II" "Chikka Tirupathi" "Uttarahalli" "Lingadheeranahalli" ...
$ size     : chr "2 BHK" "4 Bedroom" "3 BHK" "3 BHK" ...
$ society  : chr "Coomee" "Theanmp" "" "Soiewre" ...
$ total_sqft : num 1056 2600 1440 1521 1200 ...
$ bath     : int 2 5 2 3 2 2 4 4 3 6 ...
$ balcony  : int 1 3 3 1 1 1 NA NA 1 NA ...
$ price    : num 39.1 120 62.95 51 ...
$ BHK     : num 2 4 3 3 2 2 4 4 3 6 ...
```

The Environment pane on the right shows the following data:

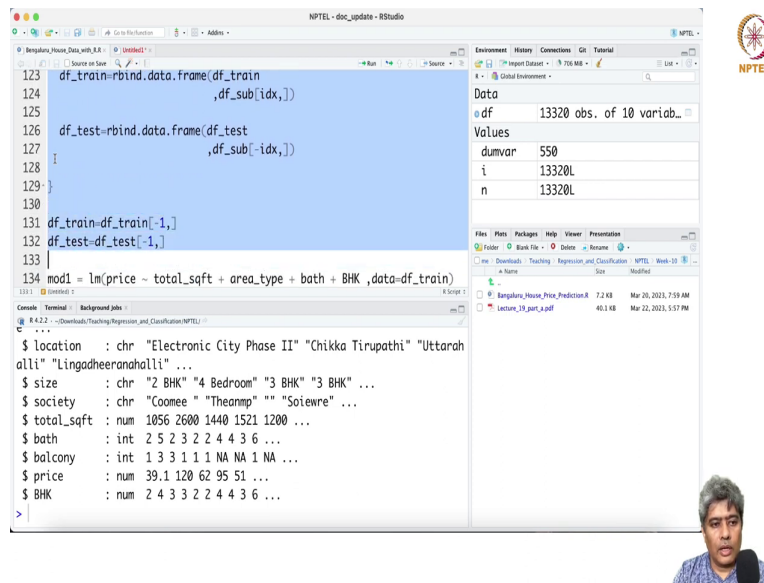
Variable	Value
df	13320 obs. of 10 variab...
dumvar	550
i	13320L
n	13320L

The Files pane shows the following files:

Name	Size	Modified
Bangalore_House_Price_Prediction.R	7.2 KB	Mar 20, 2023, 7:59 AM
Lecture_19_part_1.pdf	40.1 KB	Mar 22, 2023, 5:57 PM

A small video inset in the bottom right corner shows a man with grey hair and a blue shirt, gesturing with his right hand.

(Refer Slide Time: 06:55)



The screenshot shows the RStudio interface. The script editor contains the following R code:

```
123 df_train=rbind.data.frame(df_train
124                           ,df_sub[idx,])
125
126 df_test=rbind.data.frame(df_test
127                           ,df_sub[-idx,])
128
129 }
130
131 df_train=df_train[-1,]
132 df_test=df_test[-1,]
133
134 mod1 = lm(price ~ total_sqft + area_type + bath + BHK ,data=df_train)
```

The console shows the output of the `df_train` data frame:

```
$ location : chr "Electronic City Phase II" "Chikka Tirupathi" "Uttarah
allii" "Lingadheeranahalli" ...
$ size : chr "2 BHK" "4 Bedroom" "3 BHK" "3 BHK" ...
$ society : chr "Coomee" "Theanmp" "" "Soiewre" ...
$ total_sqft : num 1056 2600 1440 1521 1200 ...
$ bath : int 2 5 2 3 2 2 4 4 3 6 ...
$ balcony : int 1 3 3 1 1 1 NA NA 1 NA ...
$ price : num 39.1 120 62.95 51 ...
$ BHK : num 2 4 3 3 2 2 4 4 3 6 ...
```

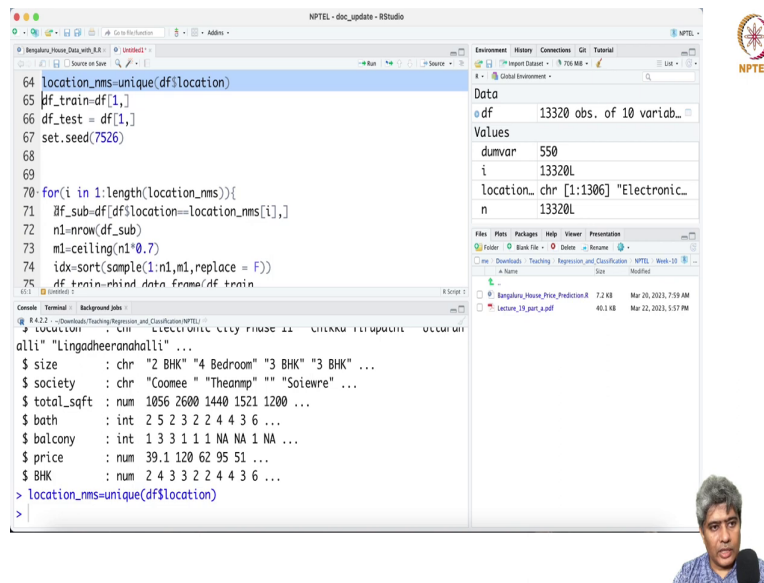
The Data Viewer on the right shows the structure of the `df` data frame:

Variable	Value
df	13320 obs. of 10 variab...
dumvar	550
i	13320L
n	13320L

The NPTEL logo is visible in the top right corner of the RStudio window.

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)



The screenshot shows an RStudio session with the following code in the editor:

```
64 location_rms=unique(df$location)
65 df_train=df[1,]
66 df_test = df[1,]
67 set.seed(7526)
68
69
70 for(i in 1:length(location_rms)){
71   df_sub=df[df$location==location_rms[i],]
72   n1=nrow(df_sub)
73   m1=ceiling(n1*0.7)
74   idx=sort(sample(1:n1,m1,replace = F))
75   df_train=rbind(data.frame(df_train
```

The console output shows the structure of the data frame:

```
$ size      : chr "2 BHK" "4 Bedroom" "3 BHK" "3 BHK" ...
$ society   : chr "Coomee" "Theanp" "Soiewre" ...
$ total_sqft: num 1056 2600 1440 1521 1200 ...
$ bath      : int 2 5 2 3 2 2 4 4 3 6 ...
$ balcony   : int 1 3 3 1 1 1 NA NA 1 NA ...
$ price     : num 39.1 120 62.95 51 ...
$ BHK       : num 2 4 3 3 2 2 4 4 3 6 ...
```

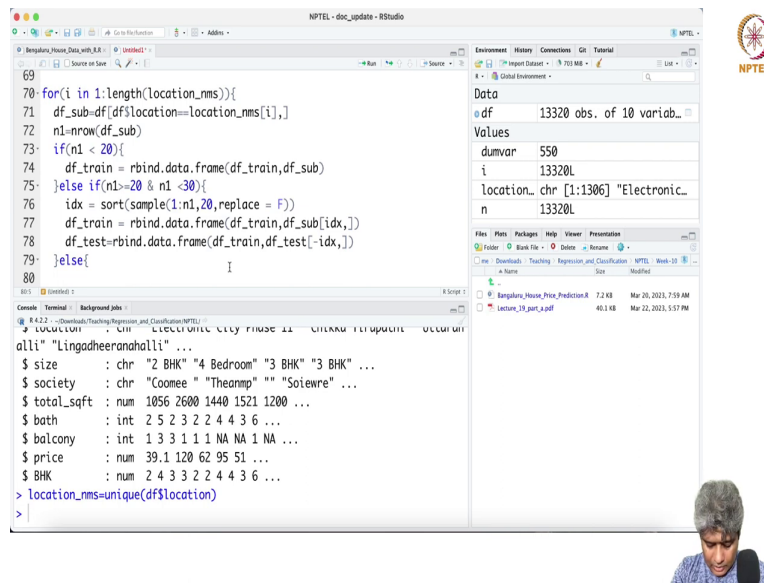
The Data Inspector on the right shows the following summary for the 'df' data frame:

Variable	Value
df	13320 obs. of 10 variab...
dumvar	550
i	13320L
location_chr [1:1306]	"Electronic_..."
n	13320L

The NPTEL logo is visible in the top right corner of the RStudio window.

So, let me it is I hope this is location names, yeah. So, it is for the i th 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)



The screenshot shows an RStudio window with the following R code in the editor:

```
69  
70 for(i in 1:length(location_nms)){  
71   df_sub=df[df$location==location_nms[i],]  
72   n1=nrow(df_sub)  
73   if(n1 < 20){  
74     df_train = rbind.data.frame(df_train,df_sub)  
75   }else if(n1==20 & n1 <30){  
76     idx = sort(sample(1:n1,20,replace = F))  
77     df_train = rbind.data.frame(df_train,df_sub[idx,])  
78     df_test=rbind.data.frame(df_train,df_test[-idx,])  
79   }else{  
80     I
```

The console output shows the summary of a data frame:

```
$ size      : chr "2 BHK" "4 Bedroom" "3 BHK" "3 BHK" ...  
$ society   : chr "Coomee" "Theanmp" " " "Soiewre" ...  
$ total_sqft : num 1056 2600 1440 1521 1200 ...  
$ bath      : int 2 5 2 3 2 2 4 4 3 6 ...  
$ balcony   : int 1 3 3 1 1 1 NA NA 1 NA ...  
$ price     : num 39.1 120 62.95 51 ...  
$ BHK       : num 2 4 3 3 2 2 4 4 3 6 ...  
> location_nms=unique(df$location)  
>
```

The Environment pane on the right shows the data frame 'df' with 13320 observations and 10 variables. The 'Values' pane shows the following values:

Variable	Value
dumvar	550
i	13320L
location_chr [1:1306]	"Electronic_..."
n	13320L

The Files pane shows the following files:

File Name	Size	Modified
Bangalore_House_Price_Prediction.R	7.2 KB	Mar 20, 2023, 7:59 AM
Lecture_19_part_1.pdf	40.1 KB	Mar 22, 2023, 5:57 PM

The NPTEL logo is visible in the top right corner of the RStudio window.

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.

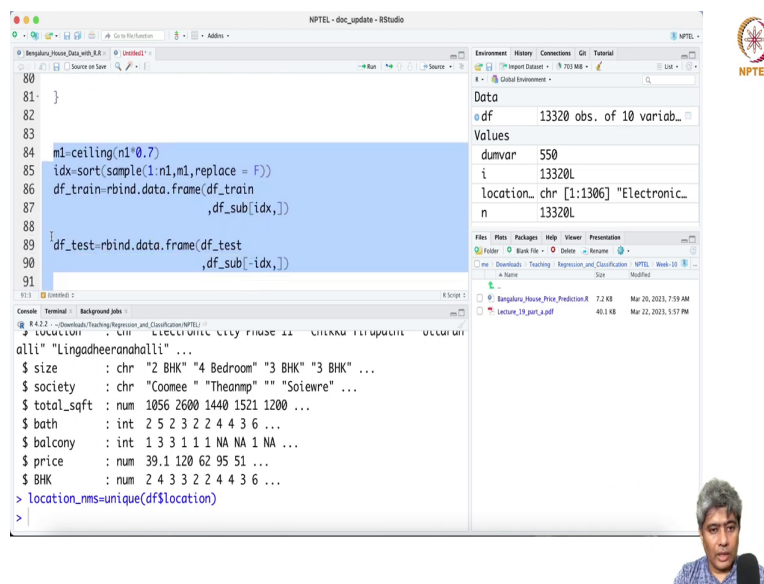
So, df_{train} equal to $rbind$ dot $data$ dot $frame$ df_{train} comma df_{sub} . All the instances goes to the training data set. I am not going to I will not have enough data to even test that location, if going forward more, but at least my model will be trained. And if in future, if somebody wants 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)



The screenshot shows an RStudio window with the following code in the editor:

```
80  
81: }  
82  
83  
84 m1=ceiling(n1*0.7)  
85 idx=sort(sample(1:n1,m1,replace = F))  
86 df_train=rbind.data.frame(df_train  
87 ,df_sub[idx,])  
88  
89 df_test=rbind.data.frame(df_test  
90 ,df_sub[-idx,])  
91  
92
```

The console output shows the following summary:

```
$ size      : chr "2 BHK" "4 Bedroom" "3 BHK" "3 BHK" ...  
$ society   : chr "Coomee" "Theanmp" "" "Soiewre" ...  
$ total_sqft : num 1056 2600 1440 1521 1200 ...  
$ bath      : int 2 5 2 3 2 2 4 4 3 6 ...  
$ balcony   : int 1 3 3 1 1 1 NA NA 1 NA ...  
$ price     : num 39.1 120 62 95 51 ...  
$ BHK       : num 2 4 3 3 2 2 4 4 3 6 ...  
> location_nms=unique(df$location)  
>
```

The Environment pane on the right shows the 'Data' environment with a data frame 'df' containing 13320 observations and 10 variables. The 'Values' pane shows the following values:

Variable	Value
dumvar	550
i	13320L
location_chr [1:1306]	"Electronic_"
n	13320L

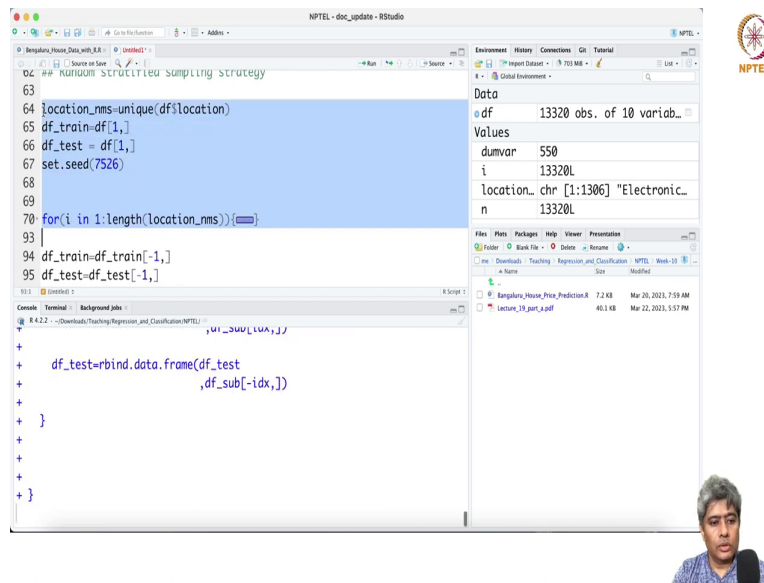
The Files pane shows the following files:

File Name	Size	Modified
Bangalore_House_Price_Prediction.R	7.2 KB	Mar 20, 2023, 7:59 AM
Lecture_10_part_2.pdf	40.1 KB	Mar 22, 2023, 5:57 PM

The NPTEL logo is visible in the top right corner of the RStudio window.

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)



The screenshot shows the RStudio interface with the following components:

- Source Editor:** Contains R code for random stratified sampling. Lines 64-66 define `location_rms`, `df_train`, and `df_test`. Lines 67-69 set a seed and loop through `location_rms`. Lines 94-95 show the final `df_train` and `df_test` assignments.
- Environment:** Shows a data object with 13320 observations and 10 variables.
- Data Preview:** Displays the first few rows of the data frame.
- Files Panel:** Lists files in the current project, including `Bangalore_House_Price_Prediction.R` and `Lecture_19_part_1.pdf`.
- Console:** Shows the execution output of the code, including the `df_test` data frame structure.

Variable	Value
df	13320 obs. of 10 variab...
Values	
dumvar	550
i	13320L
location_ chr [1:1306]	"Electronic_
n	13320L

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

(Refer Slide Time: 11:29)

The screenshot displays the RStudio interface. The main editor window contains the following R code:


```
128 }
129 }
130
131 df_train=df_train[-1,]
132 df_test=df_test[-1,]
133
134 mod1 = lm(price ~ total_sqft + area_type + bath + BHK ,data=df_train)
135 summary(mod1)
136
137
138 mod2 = update(mod1, ~. + Location)
139 summary(mod2)
```

The Environment pane on the right shows the following variables:

Variable	Value
df_train	10778 obs. of 10 variab...
dumvar	550
i	1306L
idx	int [1:20] 1 2 3 4 5 6 7_
location_chr	[1:1306] "Electronic_
m1	26

The Console window shows the execution of the code, including the output of the `summary(mod1)` function. The output is partially visible, showing the model's coefficients and statistics.

The Files pane on the right shows the current project files, including `Bangalore_House_Price_Prediction.R` and `Lecture_19_part_1.pdf`.



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)

The screenshot shows an RStudio window with the following code in the editor:

```
67 set.seed(7526)
68
69
70 for(i in 1:length(location_rms)){}
93
94 df_train=df_train[-1,]
95 df_test=df_test[-1,]
96
97 I
98 mod1 = lm(price ~ total_sqft + area_type + bath + BHK ,data=df_train)
99 summary(mod1)
100
```

The console displays the following summary output for the linear model:

```
Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.258e+01 3.964e+00 -10.743 < 2e-16
total_sqft 5.406e-02 9.635e-04 56.109 < 2e-16
area_typeCarpet Area 2.770e+00 1.477e+01 0.188 0.851
area_typePlot Area 6.759e+01 4.191e+00 16.127 < 2e-16
area_typeSuper built-up Area 4.428e-01 3.100e+00 0.143 0.886
bath 3.897e+01 2.000e+00 19.488 < 2e-16
BHK -1.551e+01 2.079e+00 -7.461 9.23e-14

(Intercept) ***
total_sqft ***
```

The environment pane on the right shows the following values for the model object:

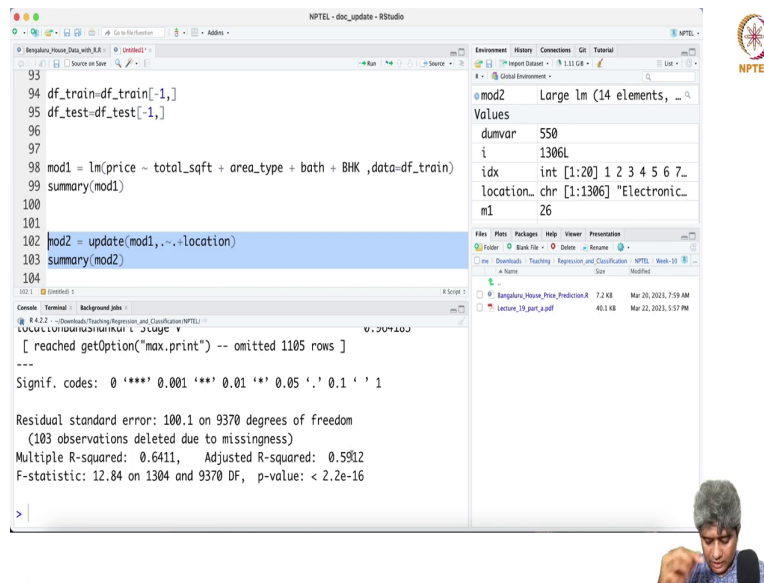
Variable	Value
dumvar	550
i	1306L
idx	int [1:20] 1 2 3 4 5 6 7_
location_	chr [1:1306] "Electronic_
m1	26

The file explorer on the right shows the following files:

File Name	Size	Modified
Bangluru_House_Price_Prediction.R	7.2 KB	Mar 20, 2022, 7:59 AM
Lecture_19_part_a.pdf	40.1 KB	Mar 22, 2022, 5:57 PM



(Refer Slide Time: 12:07)



The screenshot shows the RStudio interface with the following code in the script editor:

```
93  
94 df_train=df_train[-1,]  
95 df_test=df_test[-1,]  
96  
97  
98 mod1 = lm(price ~ total_sqft + area_type + bath + BHK ,data=df_train)  
99 summary(mod1)  
100  
101  
102 mod2 = update(mod1,~.~location)  
103 summary(mod2)  
104
```

The console output shows the summary for model 2:

```
---  
[ reached getOption("max.print") -- omitted 1105 rows ]  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 100.1 on 9370 degrees of freedom  
(103 observations deleted due to missingness)  
Multiple R-squared:  0.6411,    Adjusted R-squared:  0.5912  
F-statistic: 12.84 on 1304 and 9370 DF,  p-value: < 2.2e-16  
  
>
```

The Environment pane on the right shows the following values for the model object:

Variable	Value
dumvar	550
i	1306L
idx	int [1:20] 1 2 3 4 5 6 7_
location_	chr [1:1306] "Electronic_
m1	26

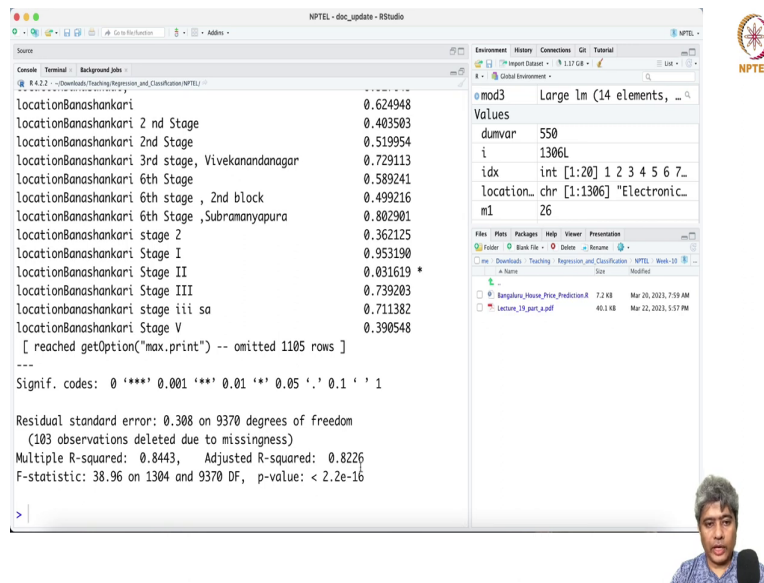
The Files pane shows the following files:

File Name	Size	Modified
Bangalore_House_Price_Prediction.R	7.2 KB	Mar 20, 2023, 7:59 AM
Lecture_19_part_1.pdf	40.1 KB	Mar 22, 2023, 5:57 PM

The NPTEL logo is visible in the top right corner of the RStudio window.

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)



The screenshot shows the RStudio interface with the following content:


```
Source
Console Terminal Background Jobs
R 4.2.2 --(Downloads/Teaching/Regression_and_Classification/NPTEL)
locationBanashankari 0.624948
locationBanashankari 2nd Stage 0.403503
locationBanashankari 2nd Stage 0.519954
locationBanashankari 3rd stage, Vivekanandanagar 0.729113
locationBanashankari 6th Stage 0.589241
locationBanashankari 6th stage , 2nd block 0.499216
locationBanashankari 6th Stage ,Subramanyapura 0.802901
locationBanashankari stage 2 0.362125
locationBanashankari Stage I 0.953190
locationBanashankari Stage II 0.031619 *
locationBanashankari Stage III 0.739203
locationBanashankari stage iii sa 0.711382
locationBanashankari Stage V 0.390548
[ reached getOption("max.print") -- omitted 1105 rows ]
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.308 on 9370 degrees of freedom
(103 observations deleted due to missingness)
Multiple R-squared: 0.8443, Adjusted R-squared: 0.8226
F-statistic: 38.96 on 1304 and 9370 DF, p-value: < 2.2e-16

>
```

Environment History Connections GUI Tutorial
Global Environment
mod3 Large lm (14 elements, ...)
Values
dumvar 550
i 1306L
idx int [1:20] 1 2 3 4 5 6 7_
location_ chr [1:1306] "Electronic_
m1 26

Files: Files Packages Help Viewer Presentation
Folder Bank File Delete Rename
Downloads Teaching Regression_and_Classification NPTEL Week-10
Name Size Modified
Bangalore_House_Price_Prediction.R 7.2 KB Mar 20, 2023, 7:59 AM
Lecture_10_part_1.pdf 40.1 KB Mar 22, 2023, 5:57 PM



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)

The screenshot shows an RStudio window with the following code in the editor:

```
102 mod2 = update(mod1, ~. + location)
103 summary(mod2)
104
105
106 mod3 = lm(log(price) ~ log(total_sqft) + area_type + log(bath) + log(BHFI
107 + location
108 , data=df_train)
109 summary(mod3)
110
111 df_test$price_hat1 = predict(mod1, newdata = df_test)
112
113
```

The console output shows the following statistics:

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.308 on 9370 degrees of freedom
(103 observations deleted due to missingness)
Multiple R-squared:  0.8443,    Adjusted R-squared:  0.8226
F-statistic: 38.96 on 1304 and 9370 DF,  p-value: < 2.2e-16

> df_test$price_hat1 = predict(mod1, newdata = df_test)
>
```

The Environment pane on the right shows the following objects:

- df_test: 297618 obs. of 11 variables
- df_train: 10778 obs. of 10 variables
- mod1: Large lm (14 elements)
- mod2: Large lm (14 elements)
- mod3: Large lm (14 elements)
- Values: dumvar = 550

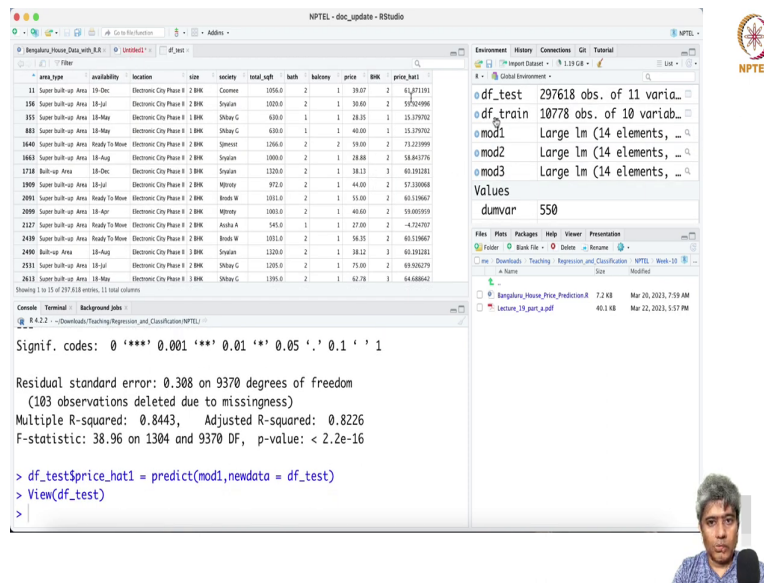
The Files pane shows the following files:

- Bangalore_House_Price_Prediction.R (7.2 KB, Mar 26, 2023, 7:59 AM)
- Lecture_19_part_1.pdf (40.1 KB, Mar 22, 2023, 5:57 PM)

A small portrait of a man is visible in the bottom right corner of the RStudio window.

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)



The screenshot shows an RStudio interface with the following components:

- Environment Pane:** Shows variables `df_test` (297618 obs. of 11 variab.), `df_train` (10778 obs. of 10 variab.), `mod1` (Large lm (14 elements, ...)), `mod2` (Large lm (14 elements, ...)), and `mod3` (Large lm (14 elements, ...)). A variable `dumvar` has a value of 550.
- Console:** Displays the following output:

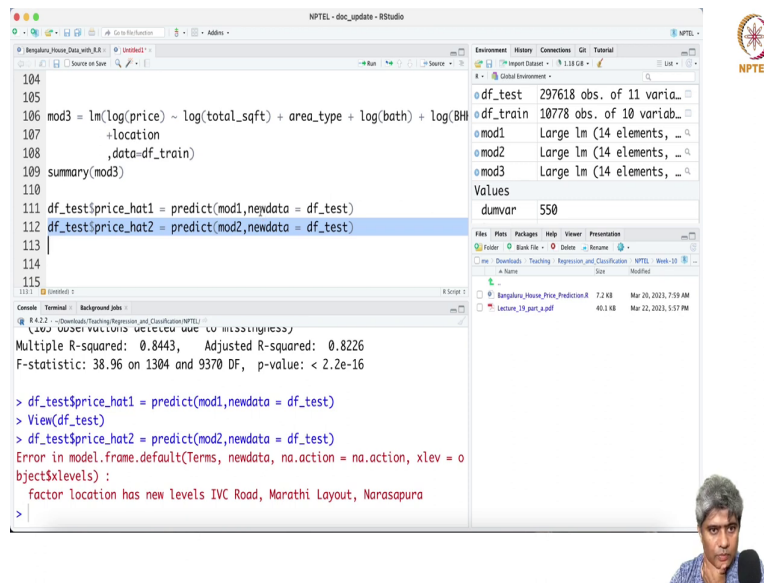
```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.308 on 9370 degrees of freedom
(103 observations deleted due to missingness)
Multiple R-squared:  0.8443,    Adjusted R-squared:  0.8226 
F-statistic: 38.96 on 1304 and 9370 DF,  p-value: < 2.2e-16

> df_test$price_hat1 = predict(mod1, newdata = df_test)
> View(df_test)
>
```
- Table:** A data table with columns: `id`, `area_type`, `availability`, `location`, `size`, `society`, `total_sqft`, `bath`, `bhkone`, `price`, `bhk`, and `price_hat1`. It lists various properties with their details.
- Terminal:** Shows the command `R 4.2.2 - Downloads/Teaching/Regression_and_Classification/NPTEL/`.



(Refer Slide Time: 13:26)



The screenshot displays the RStudio interface with the following content:

```
104
105
106 mod3 = lm(log(price) ~ log(total_sqft) + area_type + log(bath) + log(BH
107 +location
108 ,data=df_train)
109 summary(mod3)
110
111 df_test$price_hat1 = predict(mod1,newdata = df_test)
112 df_test$price_hat2 = predict(mod2,newdata = df_test)
113 |
114
115
```

Environment pane:

- df_test: 297618 obs. of 11 variables
- df_train: 10778 obs. of 10 variables
- mod1: Large lm (14 elements)
- mod2: Large lm (14 elements)
- mod3: Large lm (14 elements)

Values pane:

dumvar: 550

Console:

```
R 4.2.2 -- Command-Line Training Regression and Classification (NPTEL)
(loading variables deleted due to missingness)
Multiple R-squared: 0.8443, Adjusted R-squared: 0.8226
F-statistic: 38.96 on 1304 and 9370 DF, p-value: < 2.2e-16

> df_test$price_hat1 = predict(mod1,newdata = df_test)
> View(df_test)
> df_test$price_hat2 = predict(mod2,newdata = df_test)
Error in model.frame.default(Terms, newdata, na.action = na.action, xlev = o
bject$levels) :
  factor location has new levels IVC Road, Marathi Layout, Narasapura
>
```

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)

The screenshot shows an RStudio window titled 'NPTEL - doc_update - RStudio'. The script editor contains the following R code:

```
for(i in 1:length(location_rms)){
71 df_sub=df[df$location==location_rms[i],]
72 n1=nrow(df_sub)
73 if(n1 < 20){
74   df_train = rbind.data.frame(df_train,df_sub)
75 }else if(n1>=20 & n1 <30){
76   idx = sort(sample(1:n1,20,replace = F))
77   df_train = rbind.data.frame(df_train,df_sub[idx,])
78   df_test=rbind.data.frame(df_test,df_sub[-idx,])
79 }else{
80   m1=ceiling(n1*0.7)
81   idx=sort(sample(1:n1,m1,replace = F))
```

The Environment pane on the right shows the following objects:

- df_test: 297618 obs. of 11 variables
- df_train: 10778 obs. of 10 variables
- mod1: Large lm (14 elements)
- mod2: Large lm (14 elements)
- mod3: Large lm (14 elements)
- Values: dumvar = 550

The Console pane shows the following output and error:

```
Multiple R-squared: 0.8443, Adjusted R-squared: 0.8226
F-statistic: 38.96 on 1304 and 9370 DF, p-value: < 2.2e-16

> df_test$price_hat1 = predict(mod1,newdata = df_test)
> View(df_test)
> df_test$price_hat2 = predict(mod2,newdata = df_test)
Error in model.frame.default(Terms, newdata, na.action = na.action, xlev = o
bject$xlevels) :
  factor location has new levels IVC Road, Marathi Layout, Narasapura
```

A small video inset in the bottom right corner shows the presenter, a man with grey hair, looking at the camera.

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)

The screenshot shows the RStudio interface with the following content:

```
105
106 mod3 = lm(log(price) ~ log(total_sqft) + area_type + log(bath) + log(BH
107 +location
108 ,data=df_train)
109 summary(mod3)
110
111 df_test$price_hat1 = predict(mod1,newdata = df_test)
112 df_test$price_hat2 = predict(mod2,newdata = df_test)
113
114
115
```

Environment: R 4.2.2 (64-bit) on x86_64-pc-linux-gnu

Global Environment

df_test 2542 obs. of 12 variabl...
df_train 10778 obs. of 10 variabl...
mod1 Large lm (14 elements, ...
mod2 Large lm (14 elements, ...
mod3 Large lm (14 elements, ...

Values

dumvar 550

Files: Files Packages Help Viewer Presentations

Folder Bank File Delete Rename
Downloads Teaching Regression and Classification NPTEL Week-10
Name Size Modified

Bangalore_House_Price_Prediction.R 7.2 KB Mar 20, 2023, 7:59 AM
Lecture_10_part_1.pdf 40.1 KB Mar 22, 2023, 5:57 PM

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.308 on 9370 degrees of freedom
(103 observations deleted due to missingness)
Multiple R-squared: 0.8443, Adjusted R-squared: 0.8226
F-statistic: 38.96 on 1304 and 9370 DF, p-value: < 2.2e-16

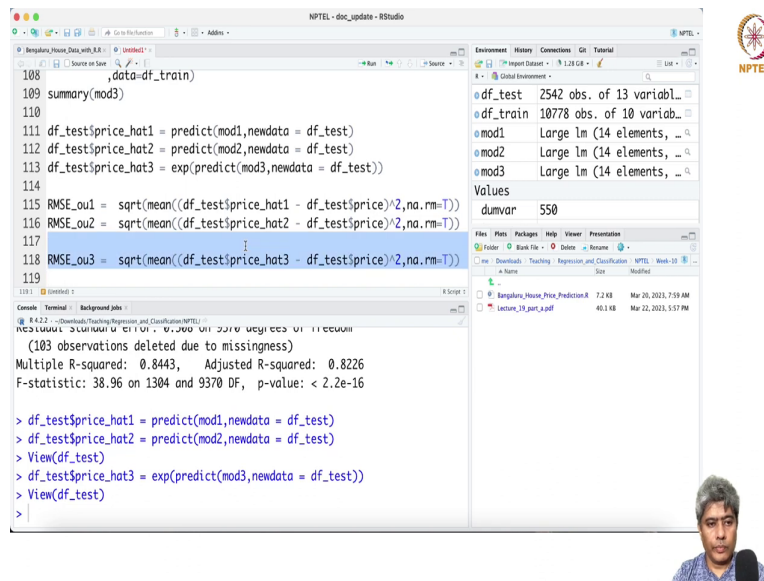
```
> df_test$price_hat1 = predict(mod1,newdata = df_test)
> df_test$price_hat2 = predict(mod2,newdata = df_test)
>
```

NPTEL

Small video inset of a man in the bottom right corner.

So, now it is fine. Now, you see there is no issue. So, we can run this df test, yeah, 62, 1131, 850, alright.

(Refer Slide Time: 15:20)



```
108 data=df_train)
109 summary(mod3)
110
111 df_test$price_hat1 = predict(mod1,newdata = df_test)
112 df_test$price_hat2 = predict(mod2,newdata = df_test)
113 df_test$price_hat3 = exp(predict(mod3,newdata = df_test))
114
115 RMSE_ou1 = sqrt(mean((df_test$price_hat1 - df_test$price)^2,na.rm=T))
116 RMSE_ou2 = sqrt(mean((df_test$price_hat2 - df_test$price)^2,na.rm=T))
117
118 RMSE_ou3 = sqrt(mean((df_test$price_hat3 - df_test$price)^2,na.rm=T))
119
```

Console Terminal Background Jobs

```
R 4.2.2 - Download: Training Regression and Classification NPTEL
(103 observations deleted due to missingness)
Multiple R-squared: 0.8443, Adjusted R-squared: 0.8226
F-statistic: 38.96 on 1304 and 9370 DF, p-value: < 2.2e-16

> df_test$price_hat1 = predict(mod1,newdata = df_test)
> df_test$price_hat2 = predict(mod2,newdata = df_test)
> View(df_test)
> df_test$price_hat3 = exp(predict(mod3,newdata = df_test))
> View(df_test)
>
```

Environment History Connections GUI Tutorial

df_test 2542 obs. of 13 variabl...
df_train 10778 obs. of 10 variabl...
mod1 Large lm (14 elements, ...
mod2 Large lm (14 elements, ...
mod3 Large lm (14 elements, ...


Values

dumvar 550

Files Packages Help Viewer Presentations

Folder Bank File Delete Rename
Downloads Teaching Regression and Classification NPTEL Week-10
Name Size Modified

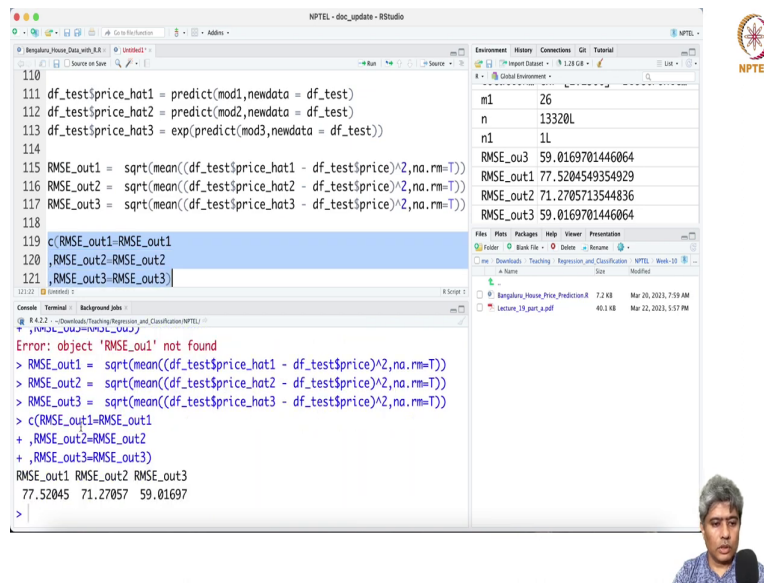
Bangalore_House_Price_Prediction.R 7.2 KB Mar 20, 2023, 7:59 AM
Lecture_10_part_1.pdf 40.1 KB Mar 22, 2023, 5:57 PM



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)



The screenshot shows an RStudio window with the following R code in the editor:

```
110  
111 df_test$price_hat1 = predict(mod1,newdata = df_test)  
112 df_test$price_hat2 = predict(mod2,newdata = df_test)  
113 df_test$price_hat3 = exp(predict(mod3,newdata = df_test))  
114  
115 RMSE_out1 = sqrt(mean((df_test$price_hat1 - df_test$price)^2,na.rm=T))  
116 RMSE_out2 = sqrt(mean((df_test$price_hat2 - df_test$price)^2,na.rm=T))  
117 RMSE_out3 = sqrt(mean((df_test$price_hat3 - df_test$price)^2,na.rm=T))  
118  
119 c(RMSE_out1-RMSE_out1  
120 ,RMSE_out2-RMSE_out2  
121 ,RMSE_out3-RMSE_out3)
```

The console shows the execution of the code and the resulting RMSE values:

```
Error: object 'RMSE_out1' not found  
> RMSE_out1 = sqrt(mean((df_test$price_hat1 - df_test$price)^2,na.rm=T))  
> RMSE_out2 = sqrt(mean((df_test$price_hat2 - df_test$price)^2,na.rm=T))  
> RMSE_out3 = sqrt(mean((df_test$price_hat3 - df_test$price)^2,na.rm=T))  
> c(RMSE_out1-RMSE_out1  
+ ,RMSE_out2-RMSE_out2  
+ ,RMSE_out3-RMSE_out3)  
RMSE_out1 RMSE_out2 RMSE_out3  
77.52045 71.27057 59.01697  
>
```

The Environment pane on the right shows the following variables:

Variable	Value
m1	26
n	13320L
n1	1L
RMSE_out3	59.0169701446064
RMSE_out1	77.5204549354929
RMSE_out2	71.2705713544836
RMSE_out3	59.0169701446064

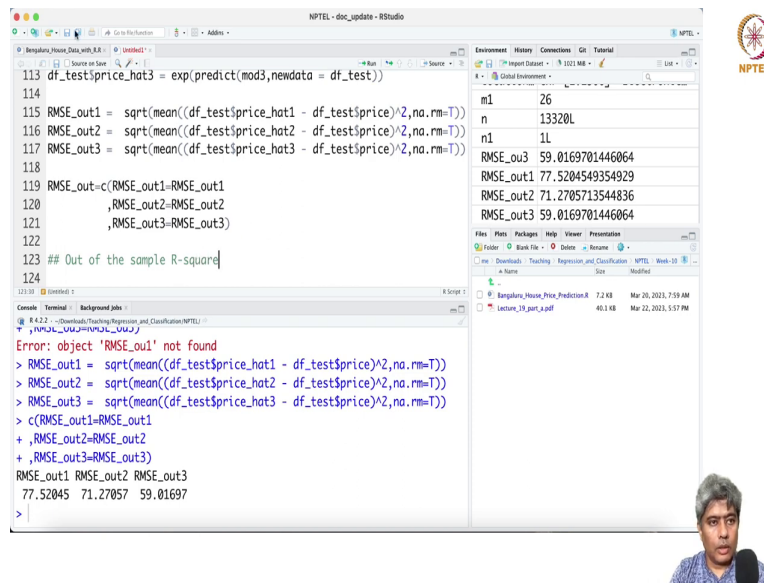
The Files pane shows the following files:

File Name	Size	Modified
Bangalore_House_Price_Prediction.R	7.2 KB	Mar 20, 2023, 7:59 AM
Lecture_19_part_1.pdf	40.1 KB	Mar 22, 2023, 5:57 PM

The NPTEL logo is visible in the top right corner of the RStudio window.

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)



The screenshot shows an RStudio window with the following code in the editor:

```
113 df_test$price_hat3 = exp(predict(mod3,newdata = df_test))
114
115 RMSE_out1 = sqrt(mean((df_test$price_hat1 - df_test$price)^2,na.rm=T))
116 RMSE_out2 = sqrt(mean((df_test$price_hat2 - df_test$price)^2,na.rm=T))
117 RMSE_out3 = sqrt(mean((df_test$price_hat3 - df_test$price)^2,na.rm=T))
118
119 RMSE_out=c(RMSE_out1-RMSE_out1
120           ,RMSE_out2-RMSE_out2
121           ,RMSE_out3-RMSE_out3)
122
123 ## Out of the sample R-square
124
```

The console shows the following error and subsequent commands:

```
Error: object 'RMSE_out1' not found
> RMSE_out1 = sqrt(mean((df_test$price_hat1 - df_test$price)^2,na.rm=T))
> RMSE_out2 = sqrt(mean((df_test$price_hat2 - df_test$price)^2,na.rm=T))
> RMSE_out3 = sqrt(mean((df_test$price_hat3 - df_test$price)^2,na.rm=T))
> c(RMSE_out1-RMSE_out1
+ ,RMSE_out2-RMSE_out2
+ ,RMSE_out3-RMSE_out3)
RMSE_out1 RMSE_out2 RMSE_out3
77.52045 71.27057 59.01697
```

The Environment pane on the right shows the following variables:

m1	26
n	13320L
n1	1L
RMSE_out3	59.0169701446064
RMSE_out1	77.5204549354929
RMSE_out2	71.2705713544836
RMSE_out3	59.0169701446064

The Files pane shows the following files:

Bangalore_House_Price_Prediction.R	7.2 KB	Mar 20, 2023, 7:59 AM
Lecture_19_part_1.pdf	40.1 KB	Mar 22, 2023, 5:57 PM

The NPTEL logo is visible in the top right corner of the RStudio window.

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.

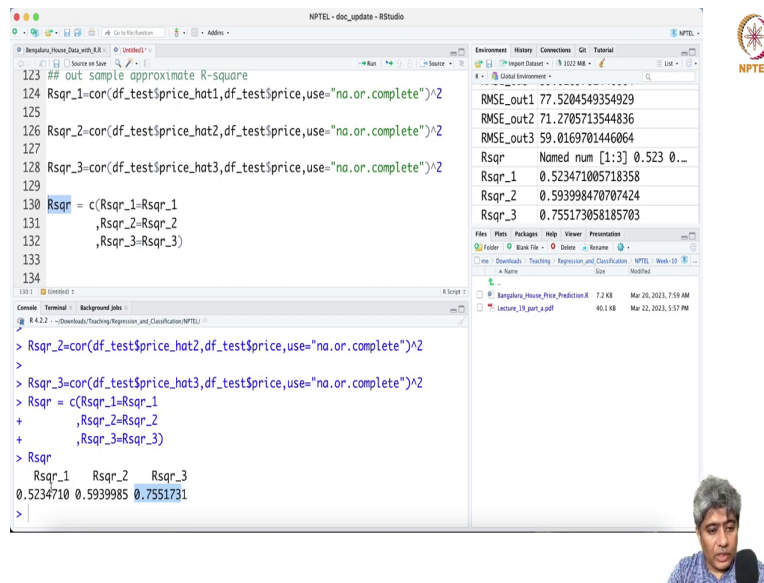
(Refer Slide Time: 18:54)

```
155
156 ## out sample approximate R-square
157 Rsqr_1=cor(df_test$price_hat1,df_test$price,use="na.or.complete")^2
158
159 Rsqr_2=cor(df_test$price_hat2,df_test$price,use="na.or.complete")^2
160
161 Rsqr_3=cor(df_test$price_hat3,df_test$price,use="na.or.complete")^2
162
163 ## New model
164
165 mod4 = update(mod3,.-,I(log(total_sqft)^2)
166               +I(log(total_sqft)^3)
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
```



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)



The screenshot displays an RStudio session with the following code and output:

```
123 ## out sample approximate R-square
124 Rsqr_1=cor(df_test$price_hat1,df_test$price,use="na.or.complete")^2
125
126 Rsqr_2=cor(df_test$price_hat2,df_test$price,use="na.or.complete")^2
127
128 Rsqr_3=cor(df_test$price_hat3,df_test$price,use="na.or.complete")^2
129
130 Rsqr = c(Rsqr_1=Rsqr_1
131           ,Rsqr_2=Rsqr_2
132           ,Rsqr_3=Rsqr_3)
133
134
```

Environment pane output:

RMSE_out1	77.5204549354929
RMSE_out2	71.2705713544836
RMSE_out3	59.0169701446064
Rsqr	Named num [1:3] 0.523 0...
Rsqr_1	0.523471005718358
Rsqr_2	0.593998470707424
Rsqr_3	0.755173058185703

Console output:

```
> Rsqr_2=cor(df_test$price_hat2,df_test$price,use="na.or.complete")^2
>
> Rsqr_3=cor(df_test$price_hat3,df_test$price,use="na.or.complete")^2
> Rsqr = c(Rsqr_1=Rsqr_1
+         ,Rsqr_2=Rsqr_2
+         ,Rsqr_3=Rsqr_3)
> Rsqr
  Rsqr_1  Rsqr_2  Rsqr_3
0.5234710 0.5939985 0.7551731
>
```

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)

The screenshot shows the RStudio interface with the following components:

- Source Editor:** Contains R code for model selection. Lines 136-144 are highlighted in blue. The code defines a new model 'mod4' by adding terms to 'mod3':

```
mod4 = update(mod3, ~., +I(log(total_sqft)^2)
+I(log(total_sqft)^3)
+I(log(bath)^2)
+I(log(bath)^3)
+I(log(BHK)^2)
+I(log(BHK)^3)
,data=df_train)
```
- Environment:** Displays the results of the model selection process:

RMSE_out	Rsqr
RMSE_out1	77.5204549354929
RMSE_out2	71.2705713544836
RMSE_out3	59.0169701446064
Rsqr	Named num [1:3] 0.523 0...
Rsqr_1	0.523471005718358
Rsqr_2	0.593998470707424
Rsqr_3	0.755173058185703
- Console:** Shows the execution of the R code and the resulting Rsqr values:

```
Rsqr_1 Rsqr_2 Rsqr_3
0.5234710 0.5939985 0.7551731
> mod4 = update(mod3, ~., +I(log(total_sqft)^2)
+I(log(total_sqft)^3)
+I(log(bath)^2)
+I(log(bath)^3)
+I(log(BHK)^2)
+I(log(BHK)^3)
,data=df_train)
```
- Files:** Lists files in the current project, including 'Bangalore_House_Price_Prediction.R' and 'Lecture_19_part_1.pdf'.
- Background Jobs:** Shows the status of background jobs.
- Terminal:** Displays the R version number: 'R 4.2.2'.

A small video inset of the presenter is visible in the bottom right corner of the RStudio window.

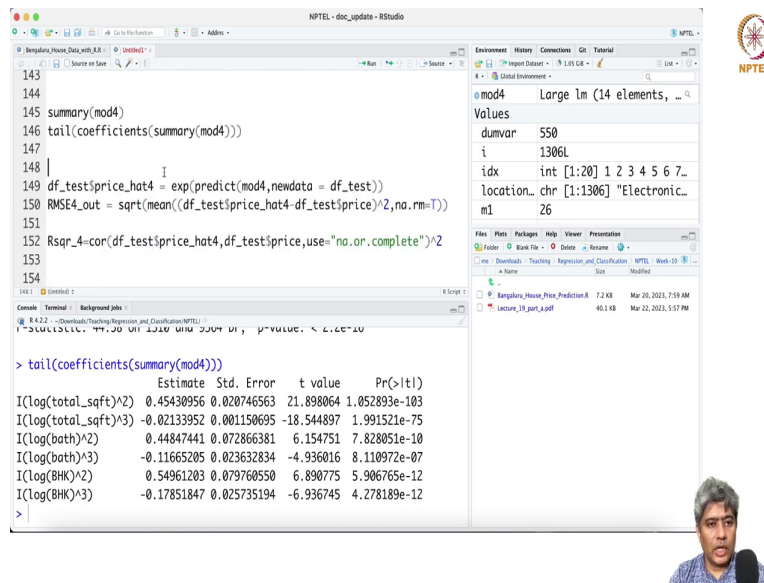
Now, we can choose some new models here with all the, you know, all the engineered feature. Let us run that.

(Refer Slide Time: 20:34)

```
NPTEL - doc_update - RStudio  
Bangluru_House_Data_with_X3.R | Untitled1 |  
136 mod4 = update(mod3, ~., -I(log(total_sqft)^2)  
137 +I(log(total_sqft)^3)  
138 +I(log(bath)^2)  
139 +I(log(bath)^3)  
140 +I(log(BHK)^2)  
141 +I(log(BHK)^3)  
142 ,data=df_train)  
143  
144  
145 summary(mod4)  
146 tail(coefficients(summary(mod4)))  
147  
Console | Terminal | Background Jobs | R Script |  
R 4.2.2 | Downloads/Teaching/Regression_and_Classification/NPTEL |  
Rsqr_1 Rsqr_2 Rsqr_3  
0.5234710 0.5939985 0.7551731  
> mod4 = update(mod3, ~., -I(log(total_sqft)^2)  
+ +I(log(total_sqft)^3)  
+ +I(log(bath)^2)  
+ +I(log(bath)^3)  
+ +I(log(BHK)^2)  
+ +I(log(BHK)^3)  
+ ,data=df_train)  
>
```



(Refer Slide Time: 20:40)



```
143
144
145 summary(mod4)
146 tail(coefficients(summary(mod4)))
147
148 |
149 df_test$price_hat4 = exp(predict(mod4,newdata = df_test))
150 RMSE4_out = sqrt(mean((df_test$price_hat4-df_test$price)^2,na.rm=T))
151
152 Rsqr_4=cor(df_test$price_hat4,df_test$price,use="na.or.complete")^2
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
```

Environment: R 4.2.2 (64-bit) on Windows x86_64-mingw-x86_64

mod4 Large lm (14 elements, ...)

Values

dumvar	550
i	1306L
idx	int [1:20] 1 2 3 4 5 6 7_
location_chr	[1:1306] "Electronic_
m1	26

```
> tail(coefficients(summary(mod4)))
              Estimate Std. Error t value Pr(>|t|)
I(log(total_sqft)^2)  0.45430956 0.020746563  21.898064 1.052893e-103
I(log(total_sqft)^3) -0.02133952 0.001150695 -18.544897 1.991521e-75
I(log(bath)^2)       0.44847441 0.072866381   6.154751 7.828051e-10
I(log(bath)^3)      -0.11665205 0.023632834  -4.936016 8.110972e-07
I(log(BHK)^2)       0.54961203 0.079760550   6.890775 5.906765e-12
I(log(BHK)^3)      -0.17851847 0.025735194  -6.936745 4.278189e-12
```

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)

```
180
181 RMSE_out=c(RMSE1_out-RMSE1_out
182           ,RMSE2_out-RMSE2_out
183           ,RMSE3_out-RMSE3_out
184           ,RMSE4_out-RMSE4_out)
185
186 Out_Rsqr = c(Rsqr_1-Rsqr_1
187             ,Rsqr_2-Rsqr_2
188             ,Rsqr_3-Rsqr_3
189             ,Rsqr_4-Rsqr_4)
190
191 # Tree Refression
192
193 console Terminal Background Jobs
194 @ 8.8.22 - Download Training Regression and Classification (NPTEL)
195 I(Log(total_sqft)^3) -0.02133952 0.001150695 -18.544897 1.991521e-75
196 I(Log(bath)^2) 0.44847441 0.072866381 6.154751 7.828051e-10
197 I(Log(bath)^3) -0.11665205 0.023632834 -4.936016 8.110972e-07
198 I(Log(BHK)^2) 0.54961203 0.079760550 6.890775 5.906765e-12
199 I(Log(BHK)^3) -0.17851847 0.025735194 -6.936745 4.278189e-12
200 > df_test$price_hat4 = exp(predict(mod4,newdata = df_test))
201 > RMSE4_out = sqrt(mean((df_test$price_hat4-df_test$price)^2,na.rm=T))
202 >
203 > Rsqr_4=cor(df_test$price_hat4,df_test$price,use="na.or.complete")^2
204 >
```



(Refer Slide Time: 21:21)

The screenshot shows an RStudio window titled "NPTEL - doc_update - RStudio". The editor contains the following R code:

```
149 RMSE4_out = sqrt(mean((df_test$price_hat4-df_test$price)^2,na.rm=T))
150
151 Rsqr_4=cor(df_test$price_hat4,df_test$price,use="na.or.complete")^2
152
153 RMSE_out=c(RMSE1_out-RMSE1_out
154           ,RMSE2_out-RMSE2_out
155           ,RMSE3_out-RMSE3_out
156           ,RMSE4_out-RMSE4_out)
157
158 Out_Rsqr = c(Rsqr_1-Rsqr_1
159            ,Rsqr_2-Rsqr_2
160            ,Rsqr_3-Rsqr_3)
```

The console shows the following output and error messages:

```
> Rsqr_4=cor(df_test$price_hat4,df_test$price,use="na.or.complete")^2
> RMSE_out=c(RMSE1_out-RMSE1_out
+           ,RMSE2_out-RMSE2_out
+           ,RMSE3_out-RMSE3_out
+           ,RMSE4_out-RMSE4_out)
Error: object 'RMSE1_out' not found
> RMSE_out
Error: object 'RMSE_out' not found
>
```

The Environment pane on the right shows the following objects:

- df_test: 2542 obs. of 14 variabl...
- df_train: 10778 obs. of 10 variabl...
- mod1: Large lm (14 elements, ...)
- mod2: Large lm (14 elements, ...)
- mod3: Large lm (14 elements, ...)
- mod4: Large lm (14 elements, ...)

The Files pane shows the following files:

- Bangluru_House_Price_Prediction.R: 7.2 KB, Mar 20, 2023, 7:59 AM
- Lecture_19_part_1.pdf: 40.1 KB, Mar 22, 2023, 5:57 PM

The NPTEL logo is visible in the top right corner of the RStudio window.

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)

```
NPTEL - doc_update - RStudio
Bangalore_House_Data_with_XGB | Untitled1 |
Source of Save | Run | Source
147
148 df_test$price_hat4 = exp(predict(mod4,newdata = df_test))
149 RMSE4_out = sqrt(mean((df_test$price_hat4-df_test$price)^2,na.rm=T))
150
151 Rsqr_4<-cor(df_test$price_hat4,df_test$price,use="na.or.complete")^2
152
153 RMSE_out=c(RMSE_out
154             ,RMSE4_out-RMSE4_out)
155
156 Out_Rsqr = c(Rsqr_1-Rsqr_1
157              ,Rsqr_2-Rsqr_2
158              ,Rsqr_3-Rsqr_3)
159
Console | Terminal | Background Jobs
R 4.2.2 - Downloaded Teaching Regression and Classification NPTEL
> RMSE_out=c(RMSE2_out-RMSE2_out
+           ,RMSE3_out-RMSE3_out
+           ,RMSE4_out-RMSE4_out)
Error: object 'RMSE1_out' not found
> RMSE_out
Error: object 'RMSE_out' not found
> RMSE_out=c(RMSE_out
+           ,RMSE4_out-RMSE4_out)
Error: object 'RMSE_out' not found
>
Environment | History | Connections | GUI | Tutorial
NPTEL
Global Environment
df_test 2542 obs. of 14 variabl...
df_train 10778 obs. of 10 variab...
mod1 Large lm (14 elements, ...
mod2 Large lm (14 elements, ...
mod3 Large lm (14 elements, ...
mod4 Large lm (14 elements, ...
Values
Files | Plots | Packages | Help | Viewer | Presentations
Folder | Blank File | Delete | Rename
Download | Teaching Regression and Classification NPTEL | Week-10
Name Size Modified
Bangalore_House_Price_Prediction.R 7.2 KB Mar 20, 2023, 7:59 AM
Lecture_10_part_1.pdf 40.1 KB Mar 22, 2023, 5:57 PM
```



(Refer Slide Time: 21:55)

The screenshot shows an RStudio window with the following R code in the editor:

```
117 RMSE_out3 = sqrt(mean((df_test$price_hat3 - df_test$price)^2, na.rm=T))
118
119 RMSE_out=c(RMSE_out1=RMSE_out1
120             ,RMSE_out2=RMSE_out2
121             ,RMSE_out3=RMSE_out3)
122
123 ## out sample approximate R-square
124 Rsqr_1=cor(df_test$price_hat1,df_test$price,use="na.or.complete")^2
125
126 Rsqr_2=cor(df_test$price_hat2,df_test$price,use="na.or.complete")^2
127
```

The console shows the following error messages:

```
Error: object 'RMSE1_out' not found
> RMSE_out
Error: object 'RMSE_out' not found
> RMSE_out=c(RMSE_out
+             ,RMSE4_out=RMSE4_out)
Error: object 'RMSE_out' not found
> RMSE_out=c(RMSE_out1=RMSE_out1
+             ,RMSE_out2=RMSE_out2
+             ,RMSE_out3=RMSE_out3)
>
```

The environment pane on the right shows the following objects:

idx	int [1:20]	1 2 3 4 5 6 7...
location_chr	[1:1306]	"Electronic_...
m1		26
n		13320L
n1		1L
RMSE_out3		59.0169701446064
RMSE_out	Named num [1:3]	77.5 71...

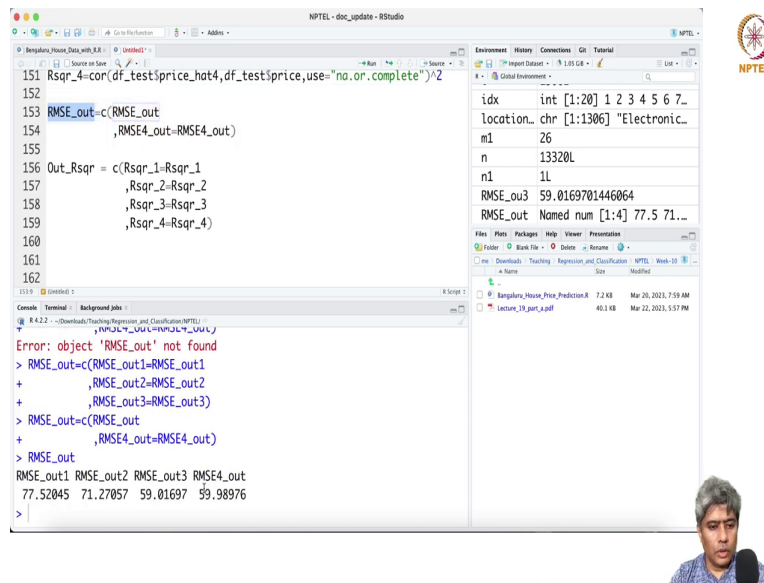
The file pane on the right shows the following files:

Bangluru_House_Price_Prediction.R	7.2 KB	Mar 20, 2023, 7:59 AM
Lecture_19_part_1.pdf	40.1 KB	Mar 22, 2023, 5:57 PM

The NPTEL logo is visible in the top right corner of the RStudio window.

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)



The screenshot shows an RStudio session with the following code and output:

```
151 Rsqr_4=cor(df_testiprice_hat4,df_testiprice,use="na.or.complete")^2
152
153 RMSE_out=c(RMSE_out
154             ,RMSE4_out=RMSE4_out)
155
156 Out_Rsqr = c(Rsqr_1=Rsqr_1
157              ,Rsqr_2=Rsqr_2
158              ,Rsqr_3=Rsqr_3
159              ,Rsqr_4=Rsqr_4)
160
161
162
```

The console shows an error message and the resulting RMSE values:

```
Error: object 'RMSE_out' not found
> RMSE_out=c(RMSE_out1=RMSE_out1
+            ,RMSE_out2=RMSE_out2
+            ,RMSE_out3=RMSE_out3)
> RMSE_out=c(RMSE_out
+            ,RMSE4_out=RMSE4_out)
> RMSE_out
RMSE_out1 RMSE_out2 RMSE_out3 RMSE4_out
77.52045  71.27057  59.01697  59.98976
```

The Environment pane shows the following variables:

Variable	Class	Value
idx	int	[1:20] 1 2 3 4 5 6 7...
location_chr	chr	[1:1306] "Electronic_...
m1	num	26
n	num	13320L
n1	int	1L
RMSE_out3	num	59.0169701446064
RMSE_out	Named num	[1:4] 77.5 71...

The Files pane shows the following files:

File Name	Size	Modified
Bangluru_House_Price_Prediction.R	7.2 KB	Mar 20, 2023, 7:59 AM
Lecture_19_part_1.pdf	40.1 KB	Mar 22, 2023, 5:57 PM

The NPTEL logo is visible in the top right corner of the RStudio window.

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)

The screenshot shows an RStudio session. The script editor contains the following R code:

```
149 RMSE4_out = sqrt(mean((df_test$price_hat4-df_test$price)^2,na.rm=T))
150
151 Rsqr_4=cor(df_test$price_hat4,df_test$price,use="na.or.complete")^2
152
153 RMSE_out=c(RMSE_out
154           ,RMSE4_out=RMSE4_out)
155
156 Out_Rsqr = c(Rsqr_1=Rsqr_1
157             ,Rsqr_2=Rsqr_2
158             ,Rsqr_3=Rsqr_3
159             ,Rsqr_4=Rsqr_4)
160
```


The console shows the output of the code:

```
RMSE_out1 RMSE_out2 RMSE_out3 RMSE4_out
77.52045 71.27057 59.01697 59.98976
> Out_Rsqr = c(Rsqr_1=Rsqr_1
+             ,Rsqr_2=Rsqr_2
+             ,Rsqr_3=Rsqr_3
+             ,Rsqr_4=Rsqr_4)
> Out_Rsqr
Rsqr_1 Rsqr_2 Rsqr_3 Rsqr_4
0.5234710 0.5939985 0.7551731 0.7330323
>
```

The Environment window shows a data frame with the following structure:

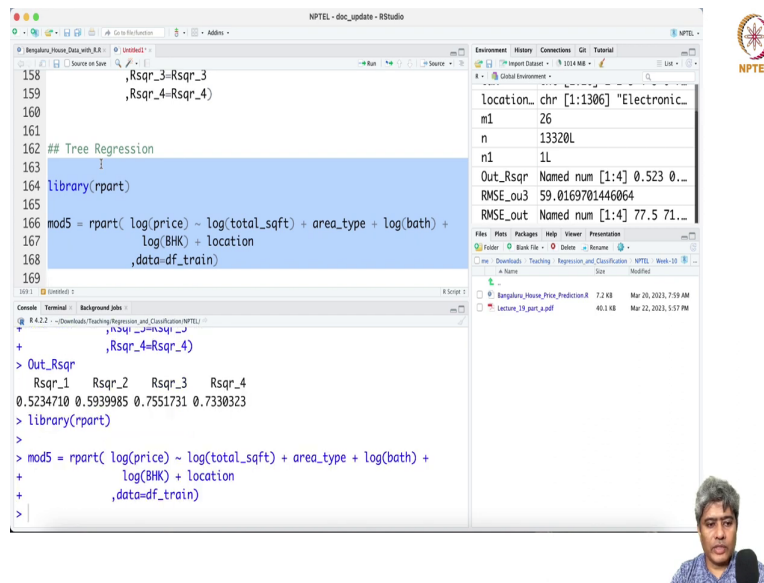
location_	chr [1:1306]	"Electronic_
m1	26	
n	13320L	
n1	1L	
Out_Rsqr	Named num [1:4]	0.523 0...
RMSE_out3	59.0169701446064	
RMSE_out	Named num [1:4]	77.5 71...

The Files window shows a folder structure: Teaching > Regression and Classification > NPTEL > Week-10.



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)



The screenshot shows an RStudio session with the following code in the editor:

```
158     ,Rsqr_3=Rsqr_3
159     ,Rsqr_4=Rsqr_4)
160
161
162 ## Tree Regression
163 |
164 library(rpart)
165
166 mod5 = rpart( log(price) ~ log(total_sqft) + area_type + log(bath) +
167              log(BHK) + location
168              ,data=df_train)
169
```

The console output shows the results of the rpart function:

```
> Out_Rsqr
  Rsqr_1  Rsqr_2  Rsqr_3  Rsqr_4
0.5234710 0.5939985 0.7551731 0.7330323
> library(rpart)
>
> mod5 = rpart( log(price) ~ log(total_sqft) + area_type + log(bath) +
+              log(BHK) + location
+              ,data=df_train)
>
```

The Environment pane on the right shows the following objects:

Object	Class	Value
location_chr	chr [1:1306]	"Electronic_..."
m1	chr	26
n	num	13320L
n1	num	1L
Out_Rsqr	Named num [1:4]	0.523 0...
RMSE_ou3	num	59.0169701446064
RMSE_out	Named num [1:4]	77.5 71...

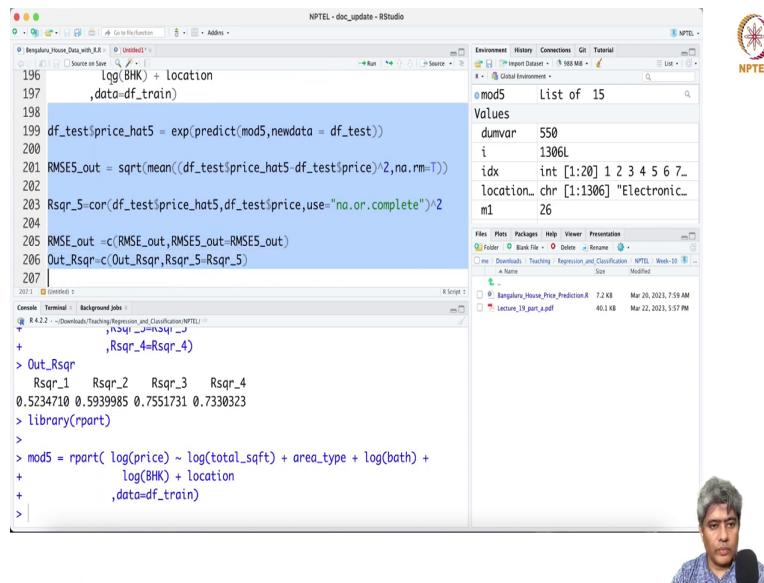
The Files pane shows the following files:

File	Size	Modified
Bangalore_House_Price_Prediction.R	7.2 KB	Mar 20, 2023, 7:59 AM
Lecture_19_part_a.pdf	40.1 KB	Mar 22, 2023, 5:57 PM

The NPTEL logo is visible in the top right corner of the RStudio window.

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)



The screenshot displays the RStudio interface with the following code in the script editor:

```
196 log(BHK) + location
197 ,data=df_train)
198
199 df_test$price_hat5 = exp(predict(mod5,newdata = df_test))
200
201 RMSE5_out = sqrt(mean((df_test$price_hat5-df_test$price)^2,na.rm=T))
202
203 Rsqr_5=cor(df_test$price_hat5,df_test$price,use="na.or.complete")^2
204
205 RMSE_out =c(RMSE_out,RMSE5_out-RMSE5_out)
206 Out_Rsqr=c(Out_Rsqr,Rsqr_5-Rsqr_5)
207
```

The console output shows the results of the Rsqr_4 and Rsqr_5 calculations:

```
> Out_Rsqr
  Rsqr_1  Rsqr_2  Rsqr_3  Rsqr_4
0.5234710 0.5939985 0.7551731 0.7330323
> library(rpart)
>
> mod5 = rpart( log(price) ~ log(total_sqft) + area_type + log(bath) +
+               log(BHK) + location
+               ,data=df_train)
```

The Environment pane on the right shows the 'mod5' object as a List of 15 values:

Value	
dumvar	550
i	1306L
idx	int [1:20] 1 2 3 4 5 6 7_
location_chr	[1:1306] "Electronic_
m1	26

The Files pane shows the project files, including 'Bangalore_House_Price_Prediction.R' and 'Lecture_19_part_1.pdf'.

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.

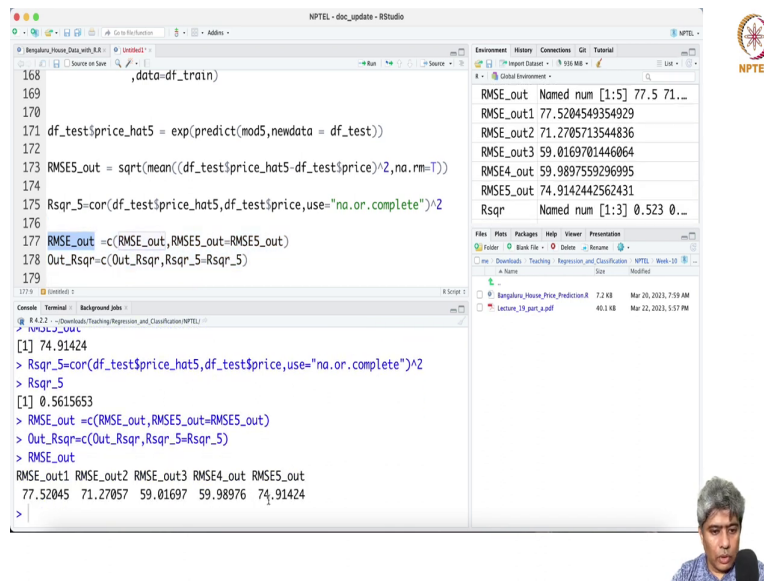
(Refer Slide Time: 23:22)

```
168     ,data=df_train)
169
170
171 df_test$price_hat5 = exp(predict(mod5,newdata = df_test))
172
173 RMSE5_out = sqrt(mean((df_test$price_hat5-df_test$price)^2,na.rm=T))
174 | I
175 Rsqr_5=cor(df_test$price_hat5,df_test$price,use="na.or.complete")^2
176
177 RMSE_out =c(RMSE_out,RMSE5_out-RMSE5_out)
178 Out_Rsqr=c(Out_Rsqr, Rsqr_5=Rsqr_5)
179
```

```
>
> mod5 = rpart( log(price) ~ log(total_sqft) + area_type + log(bath) +
+             log(BHK) + location
+             ,data=df_train)
> df_test$price_hat5 = exp(predict(mod5,newdata = df_test))
>
> RMSE5_out = sqrt(mean((df_test$price_hat5-df_test$price)^2,na.rm=T))
> RMSE5_out
[1] 74.91424
>
```



(Refer Slide Time: 23:34)





```
168     ,data=df_train)
169
170
171 df_test$price_hat5 = exp(predict(mod5,newdata = df_test))
172
173 RMSE5_out = sqrt(mean((df_test$price_hat5-df_test$price)^2,na.rm=T))
174
175 Rsqr_5=cor(df_test$price_hat5,df_test$price,use="na.or.complete")^2
176
177 RMSE_out =c(RMSE_out,RMSE5_out-RMSE5_out)
178 Out_Rsqr=c(Out_Rsqr,Rsqr_5-Rsqr_5)
179
```

```
Console Terminal Background Jobs
R 4.2.2 ...Downloads/Teaching/Regression_and_Classification/NPTEL/
> RMSE_out
[1] 74.91424
> Rsqr_5=cor(df_test$price_hat5,df_test$price,use="na.or.complete")^2
> Rsqr_5
[1] 0.5615653
> RMSE_out =c(RMSE_out,RMSE5_out-RMSE5_out)
> Out_Rsqr=c(Out_Rsqr,Rsqr_5-Rsqr_5)
> RMSE_out
RMSE_out1 RMSE_out2 RMSE_out3 RMSE4_out RMSE5_out
77.52045 71.27057 59.01697 59.98976 74.91424
>
```

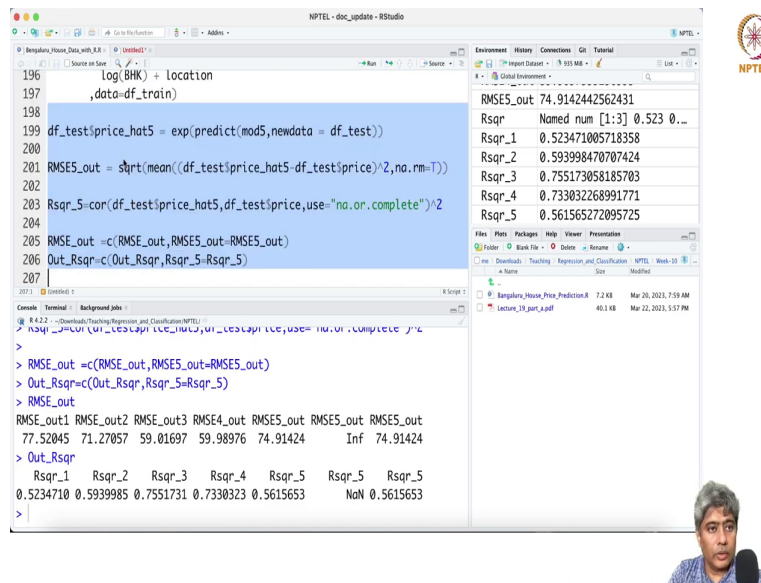
RMSE_out	Named num [1:5]
RMSE_out1	77.5204549354929
RMSE_out2	71.2705713544836
RMSE_out3	59.0169701446064
RMSE4_out	59.989759296995
RMSE5_out	74.914242562431

Rsqr	Named num [1:3]
Rsqr_1	0.523 0...



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)



The screenshot shows an RStudio session with the following code and output:

```
196 log(BHK) ~ location
197 ,data=df_train)
198
199 df_test$price_hat5 = exp(predict(mod5,newdata = df_test))
200
201 RMSE5_out = sqrt(mean((df_test$price_hat5-df_test$price)^2,na.rm=T))
202
203 Rsqr_5=cor(df_test$price_hat5,df_test$price,use="na.or.complete")^2
204
205 RMSE_out =c(RMSE_out,RMSE5_out-RMSE5_out)
206 Out_Rsqr=c(Out_Rsqr,Rsqr_5-Rsqr_5)
207
```

The console output shows the results of the calculations:


```
> RMSE_out =c(RMSE_out,RMSE5_out-RMSE5_out)
> Out_Rsqr=c(Out_Rsqr,Rsqr_5-Rsqr_5)
> RMSE_out
RMSE_out1 RMSE_out2 RMSE_out3 RMSE4_out RMSE5_out RMSE5_out RMSE5_out
77.52045 71.27057 59.01697 59.98976 74.91424 Inf 74.91424
> Out_Rsqr
Rsqr_1 Rsqr_2 Rsqr_3 Rsqr_4 Rsqr_5 Rsqr_5 Rsqr_5
0.5234710 0.5939985 0.7551731 0.7330323 0.5615653 NaN 0.5615653
>
```

The Environment pane shows the following variables:

RMSE5_out	74.9142442562431
Rsqr	Named num [1:3] 0.523 0...
Rsqr_1	0.523471005718358
Rsqr_2	0.593998470707424
Rsqr_3	0.755173058185703
Rsqr_4	0.733032268991771
Rsqr_5	0.561565272095725

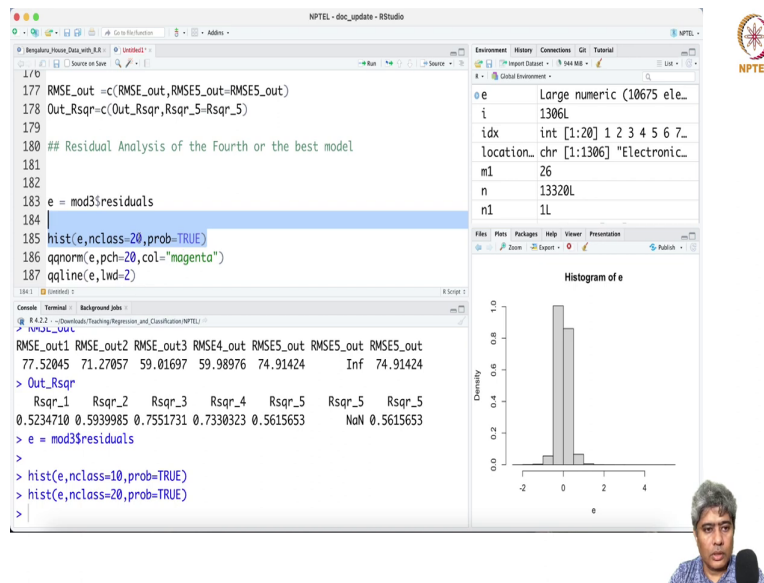
The Files pane shows the following files:

File Name	Size	Modified
Bangalore_House_Price_Prediction.R	7.2 KB	Mar 20, 2023, 7:59 AM
Lecture_19_part_1.pdf	40.1 KB	Mar 22, 2023, 5:57 PM



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

(Refer Slide Time: 24:26)



The image shows an RStudio session with the following code in the script editor:

```
177 RMSE_out =c(RMSE_out,RMSE5_out-RMSE5_out)
178 Out_Rsq=c(Out_Rsq,Rsq_5-Rsq_5)
179
180 ## Residual Analysis of the Fourth or the best model
181
182
183 e = mod3$residuals
184
185 hist(e,nclass=20,prob=TRUE)
186 qqnorm(e,pch=20,col="magenta")
187 qqline(e,lwd=2)
```



The console output shows the following data:

```
RMSE_out1 RMSE_out2 RMSE_out3 RMSE4_out RMSE5_out RMSE5_out
77.52045 71.27057 59.01697 59.98976 74.91424 Inf 74.91424
> Out_Rsq
  Rsqr_1  Rsqr_2  Rsqr_3  Rsqr_4  Rsqr_5  Rsqr_5  Rsqr_5
0.5234710 0.5939985 0.7551731 0.7330323 0.5615653  NaN  0.5615653
> e = mod3$residuals
> hist(e,nclass=10,prob=TRUE)
> hist(e,nclass=20,prob=TRUE)
>
```

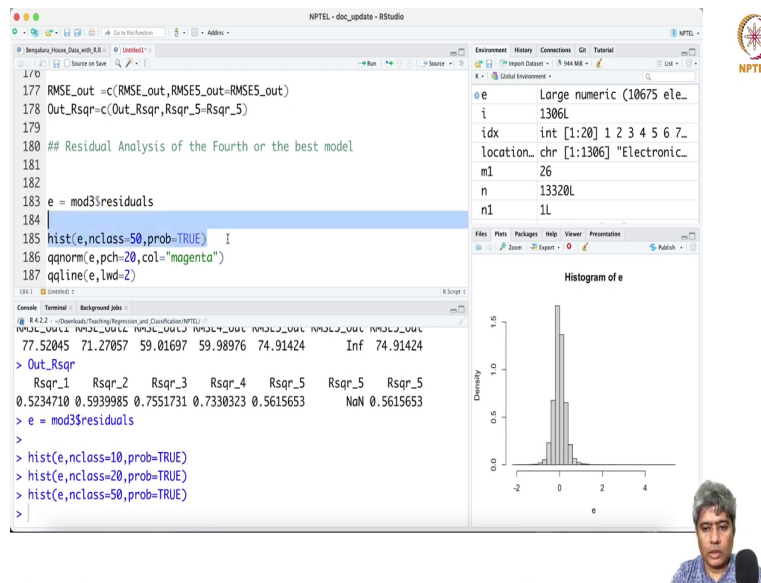
The environment pane shows the following variables:

Variable	Class	Length
e	Large numeric	10675 elements
i	integer	1306L
idx	integer	[1:20] 1 2 3 4 5 6 7...
location	character	[1:1306] "Electronic_...
m1	numeric	26
n	numeric	13320L
n1	integer	1L

A histogram titled "Histogram of e" is displayed, showing the density of residuals. The x-axis is labeled "e" and ranges from -2 to 4. The y-axis is labeled "Density" and ranges from 0.0 to 1.0. The histogram shows a distribution centered around 0, with a peak density of approximately 0.9.



(Refer Slide Time: 24:39)



The screenshot shows the RStudio interface with the following code in the script editor:

```
177 RMSE_out =c(RMSE_out,RMSE5_out-RMSE5_out)
178 Out_Rsq=c(Out_Rsq, Rsqr_5-Rsqr_5)
179
180 ## Residual Analysis of the Fourth or the best model
181
182
183 e = mod3$residuals
184
185 hist(e,nclass=50,prob=TRUE) I
186 qqnorm(e,pch=20,col="magenta")
187 qqline(e,lwd=2)
```

The console output shows the following results:

```
77.52045 71.27057 59.01697 59.98976 74.91424 Inf 74.91424
> Out_Rsq
  Rsqr_1  Rsqr_2  Rsqr_3  Rsqr_4  Rsqr_5  Rsqr_5  Rsqr_5
0.5234710 0.5939985 0.7551731 0.7330323 0.5615653 NaN 0.5615653
> e = mod3$residuals
>
> hist(e,nclass=10,prob=TRUE)
> hist(e,nclass=20,prob=TRUE)
> hist(e,nclass=50,prob=TRUE)
>
```

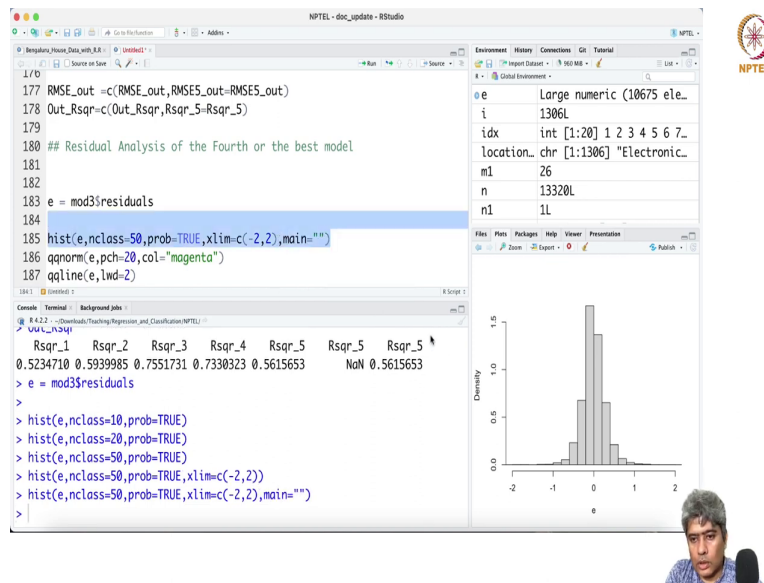
The environment pane on the right shows the following variables:

Variable	Class	Length
e	Large numeric	10675 elements
i	integer	1306L
idx	integer	[1:20] 1 2 3 4 5 6 7...
location_chr	character	[1:1306] "Electronic_...
m1	numeric	26
n	numeric	13320L
n1	integer	1L

A histogram titled "Histogram of e" is displayed, showing the density of residuals. The x-axis is labeled "e" and ranges from -2 to 4. The y-axis is labeled "Density" and ranges from 0.0 to 1.5. The histogram shows a distribution centered around 0, with a peak density of approximately 1.4.

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)



The screenshot shows the RStudio interface with the following R code in the editor:

```
177 RMSE_out =c(RMSE_out,RMSE5_out-RMSE5_out)
178 Out_Rsq=c(Out_Rsq, Rsqr_5-Rsqr_5)
179
180 ## Residual Analysis of the Fourth or the best model
181
182
183 e = mod3$residuals
184
185 hist(e,nclass=50,prob=TRUE,xlim=c(-2,2),main="")
186 qqnorm(e,pch=20,col="magenta")
187 qqline(e,lwd=2)
```

The console output shows the following data:

```
Rsqr_1 Rsqr_2 Rsqr_3 Rsqr_4 Rsqr_5 Rsqr_5 Rsqr_5
0.5234710 0.5939985 0.7551731 0.7330323 0.5615653 NaN 0.5615653
```

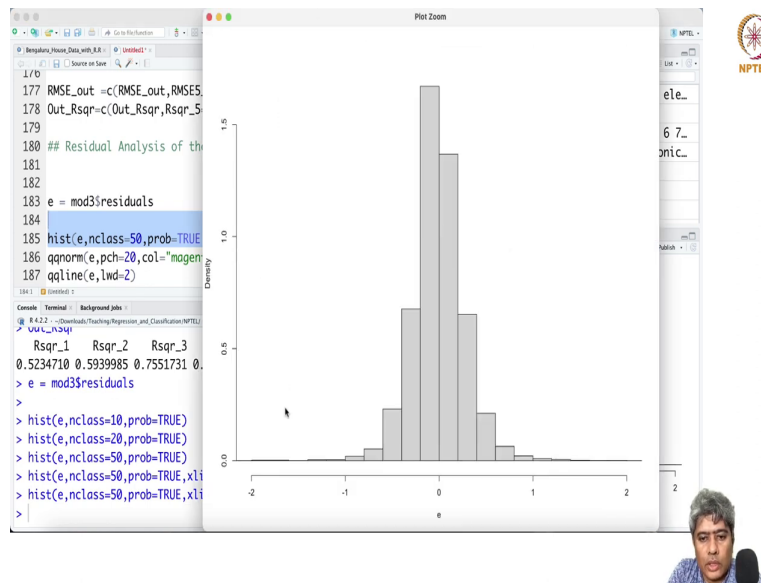
The histogram shows the density of residuals (e) on the x-axis (ranging from -2 to 2) and density on the y-axis (ranging from 0.0 to 1.5). The distribution is centered around 0.

The Environment pane shows the following variables:

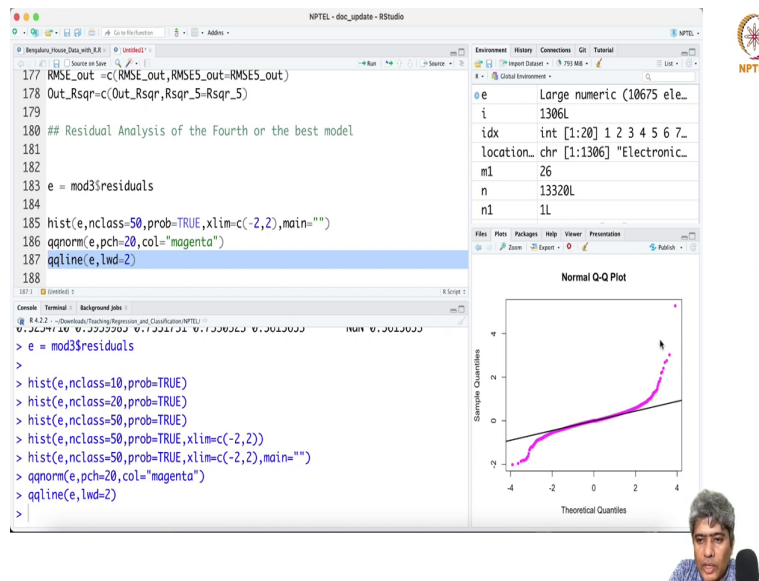
Variable	Class	Value
e	Large numeric	(10675 elements)
i	integer	1306L
idx	integer	[1:20] 1 2 3 4 5 6 7_
location_	character	[1:1306] "Electronic_
m1	integer	26
n	integer	13320L
n1	integer	1L

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

(Refer Slide Time: 25:05)

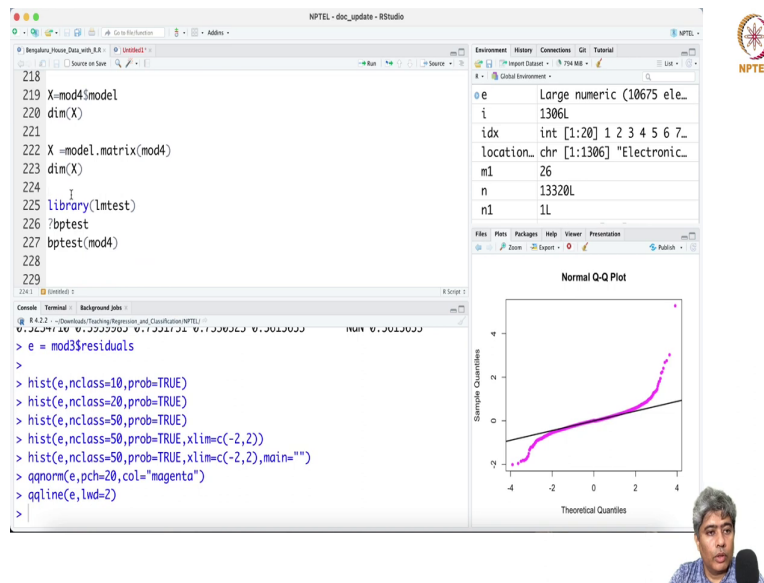


(Refer Slide Time: 25:16)



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)



The image displays the RStudio interface with the following components:

- Source Editor:** Contains R code for fitting a model and testing its residuals.

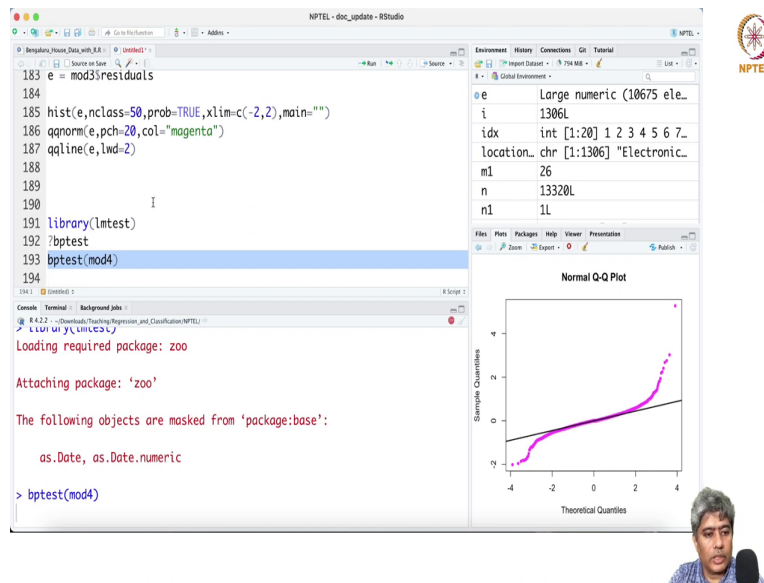
```
218  
219 X=mod4$model  
220 dim(X)  
221  
222 X =model.matrix(mod4)  
223 dim(X)  
224  
225 library(lmtest)  
226 ?bptest  
227 bptest(mod4)  
228  
229
```
- Environment:** Shows the objects created in the workspace:

e	Large numeric (10675 ele...
i	1306L
idx	int [1:20] 1 2 3 4 5 6 7_
location_	chr [1:1306] "Electronic_
m1	26
n	13320L
n1	1L
- Console:** Shows the execution of the following commands:

```
> e = mod3$residuals  
>  
> hist(e, nclass=10, prob=TRUE)  
> hist(e, nclass=20, prob=TRUE)  
> hist(e, nclass=50, prob=TRUE)  
> hist(e, nclass=50, prob=TRUE, xlim=c(-2,2))  
> hist(e, nclass=50, prob=TRUE, xlim=c(-2,2), main="")  
> qqnorm(e, pch=20, col="magenta")  
> qqline(e, lwd=2)  
>
```
- Plots:** A Normal Q-Q Plot is displayed, showing the sample quantiles of the residuals against the theoretical quantiles. The data points are plotted in magenta, and a black reference line is shown. The plot title is "Normal Q-Q Plot".



(Refer Slide Time: 25:59)



The screenshot displays the RStudio interface. The script editor on the left contains the following R code:



```
183 e = mod3$residuals
184
185 hist(e, nclass=50, prob=TRUE, xlim=c(-2,2), main="")
186 qqnorm(e, pch=20, col="magenta")
187 qqline(e, lwd=2)
188
189
190
191 library(lmtest)
192 ?bptest
193 bptest(mod4)
194
```

The console on the bottom left shows the execution of `library(lmtest)` and `?bptest`, indicating that the 'zoo' package is being attached and some objects are masked from the 'base' package.

The environment pane on the right shows the following variables:

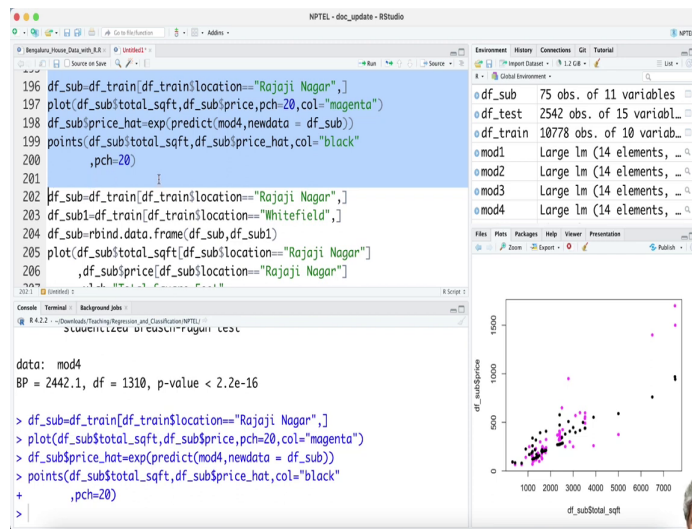
Variable	Value
e	Large numeric (10675 ele...)
i	1306L
idx	int [1:20] 1 2 3 4 5 6 7_
location_	chr [1:1306] "Electronic_
m1	26
n	13320L
n1	1L

The bottom right pane displays a Normal Q-Q Plot. The x-axis is labeled 'Theoretical Quantiles' and the y-axis is labeled 'Sample Quantiles'. The plot shows a magenta line representing the data points, which follows a black diagonal line, indicating that the residuals are approximately normally distributed.



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)



The screenshot shows the RStudio interface with the following R code in the editor:

```
196 df_sub=df_train[df_train$location=="Rajaji Nagar",]  
197 plot(df_sub$total_sqft,df_sub$price,pch=20,col="magenta")  
198 df_sub$price_hat=exp(predict(mod4,newdata = df_sub))  
199 points(df_sub$total_sqft,df_sub$price_hat,col="black"  
200 ,pch=20)  
201  
202 df_sub=df_train[df_train$location=="Rajaji Nagar",]  
203 df_sub1=df_train[df_train$location=="Whitefield",]  
204 df_sub=rbind.data.frame(df_sub,df_sub1)  
205 plot(df_sub$total_sqft[df_sub$location=="Rajaji Nagar"]  
206 ,df_sub$price[df_sub$location=="Rajaji Nagar"]  
207 )
```

The Environment pane on the right shows the following objects:

- df_sub: 75 obs. of 11 variables
- df_test: 2542 obs. of 15 variables
- df_train: 10778 obs. of 10 variables
- mod1: Large lm (14 elements)
- mod2: Large lm (14 elements)
- mod3: Large lm (14 elements)
- mod4: Large lm (14 elements)

The Console shows the following output:

```
data: mod4  
BP = 2442.1, df = 1310, p-value < 2.2e-16  
  
> df_sub=df_train[df_train$location=="Rajaji Nagar",]  
> plot(df_sub$total_sqft,df_sub$price,pch=20,col="magenta")  
> df_sub$price_hat=exp(predict(mod4,newdata = df_sub))  
> points(df_sub$total_sqft,df_sub$price_hat,col="black"  
+ ,pch=20)  
>
```

The plot shows a scatter plot of `df_sub$total_sqft` (x-axis, 0 to 7000) versus `df_sub$price` (y-axis, 0 to 15000). The data points are colored magenta and have a size of 20. A black line represents the predicted values from the model.



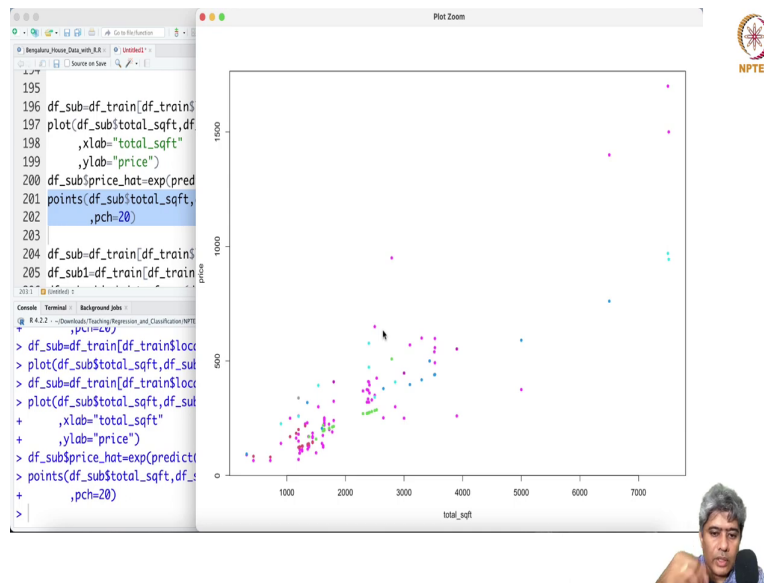
(Refer Slide Time: 27:10)

The screenshot shows the RStudio interface with the following components:

- Source Editor:** Contains R code for data manipulation and plotting. Lines 195-205 show filtering data for 'Rajaji Nagar' and 'Whitefield', plotting 'total_sqft' vs 'price' in magenta, and calculating predicted values using a model.
- Environment:** Lists objects: df_sub (75 obs. of 11 variables), df_test (2542 obs. of 15 variables), df_train (10778 obs. of 10 variables), and four linear models (mod1, mod2, mod3, mod4).
- Console:** Shows the execution of the code, including the plot command and the calculation of predicted values.
- Plots:** A scatter plot titled 'price' vs 'total_sqft' showing data points in magenta. The x-axis ranges from 0 to 7000, and the y-axis ranges from 0 to 15000.
- NPTEL Logo:** Located in the top right corner of the RStudio window.
- Speaker:** A small video feed of a person is visible in the bottom right corner of the RStudio window.

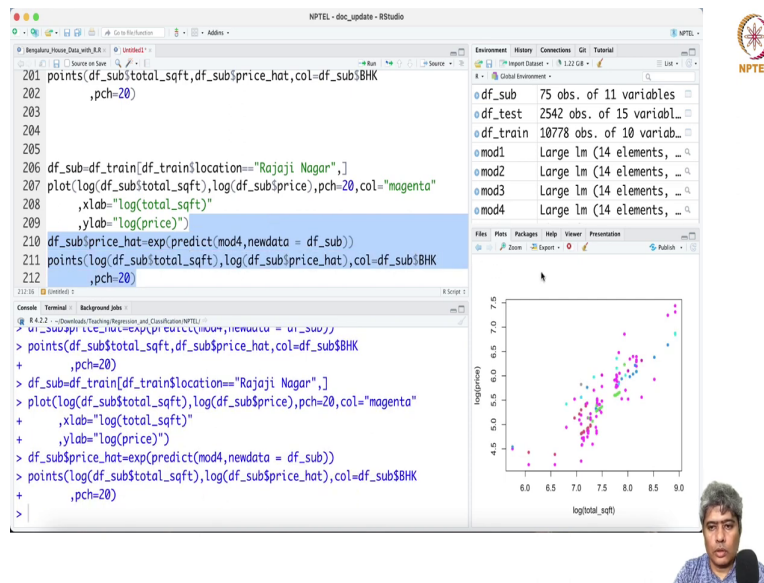
So, one thing I was thinking of that, instead of in the color, instead of giving black, I was what we can do, we can just say df sub dollar, whatever the BHK that we had. So, if we just run this in this way. So, these are the plots that we have. So, on the x axis, x label, we have the total square feet 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)



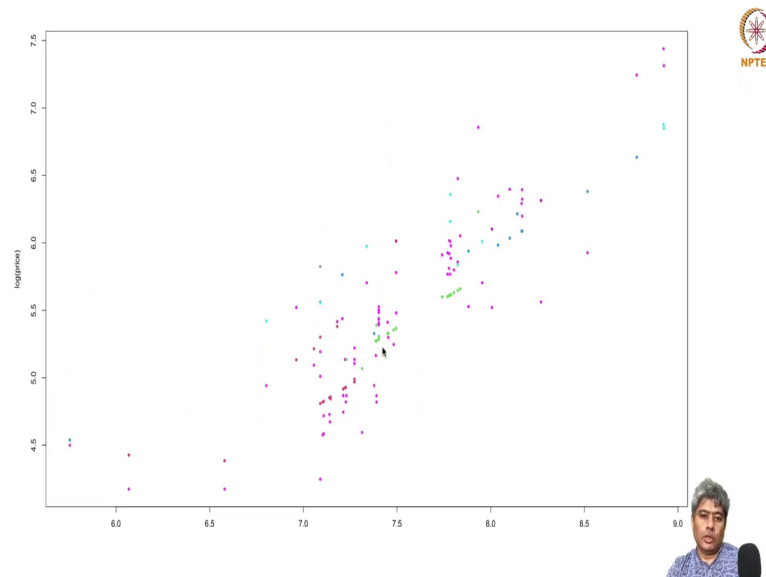
The screenshot displays the RStudio interface. The script editor on the left contains the following R code:

```
201 points(df_sub$total_sqft, df_sub$price_hat, col=df_sub$BHK
202 ,pch=20)
203
204
205
206 df_sub=df_train[df_train$location=="Rajaji Nagar",]
207 plot(log(df_sub$total_sqft), log(df_sub$price), pch=20, col="magenta"
208 , xlab="log(total_sqft)"
209 , ylab="log(price)")
210 df_sub$price_hat=exp(predict(mod4, newdata = df_sub))
211 points(log(df_sub$total_sqft), log(df_sub$price_hat), col=df_sub$BHK
212 , pch=20)
```

The console on the left shows the execution of this code. The Environment pane on the right lists several objects: `df_sub` (75 obs. of 11 variables), `df_test` (2542 obs. of 15 variables), `df_train` (10778 obs. of 10 variables), and four linear models (`mod1`, `mod2`, `mod3`, `mod4`), each with 14 elements. The plot window on the right shows a scatter plot with `log(total_sqft)` on the x-axis (ranging from 6.0 to 9.0) and `log(price)` on the y-axis (ranging from 4.5 to 7.5). The data points are colored by neighborhood (`BHK`) and shaped by model (`price_hat`).

So, if we just put log price 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.

