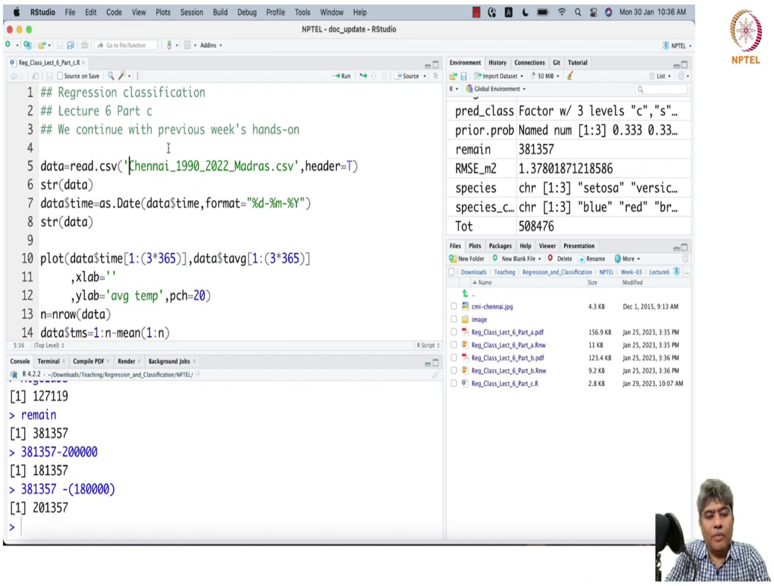
**Predictive Analytics - Regression and Classification**
**Prof. Sourish Das**
**Department of Mathematics**
**Chennai Mathematical Institute**

**Lecture - 22**
**Hands on with R Part - 5**

Welcome back to last part of lecture 6. And in this part, we will do some Hands on with R. And let me just you know try to share with this stuff.

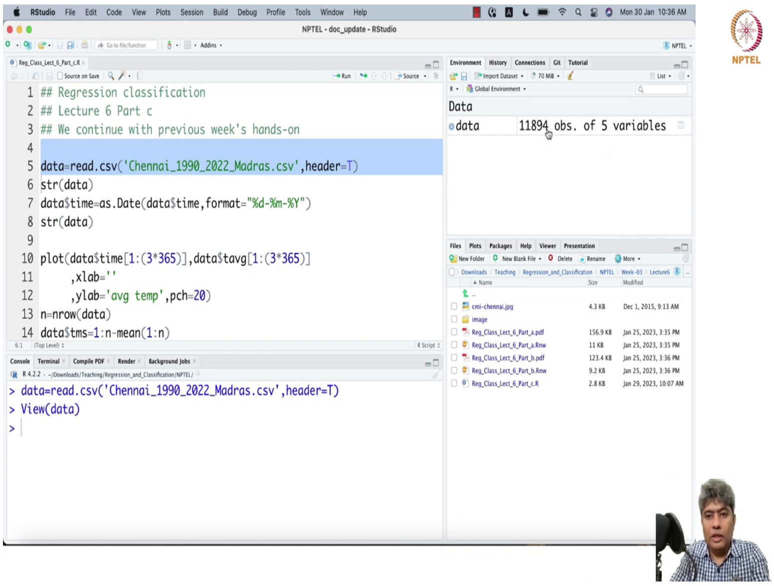
(Refer Slide Time: 00:33)



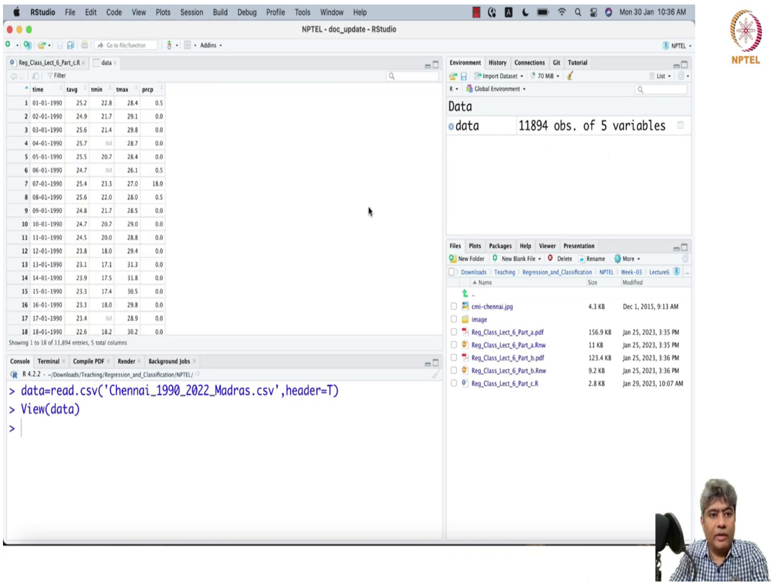

So, this is and this is the quote that I already have shared with you in the lecture material part that we are using Chennai, 1990 we have done some work with the data, this data set. And in this data set, what we have done is we have we tried to fit simple sine cosine kind of I mean some regression model with sine cosine Fourier engine term or engineer term.

So, let us start with this. We have this data. Say, let me just first clean the environment, that will help you.

(Refer Slide Time: 01:30)

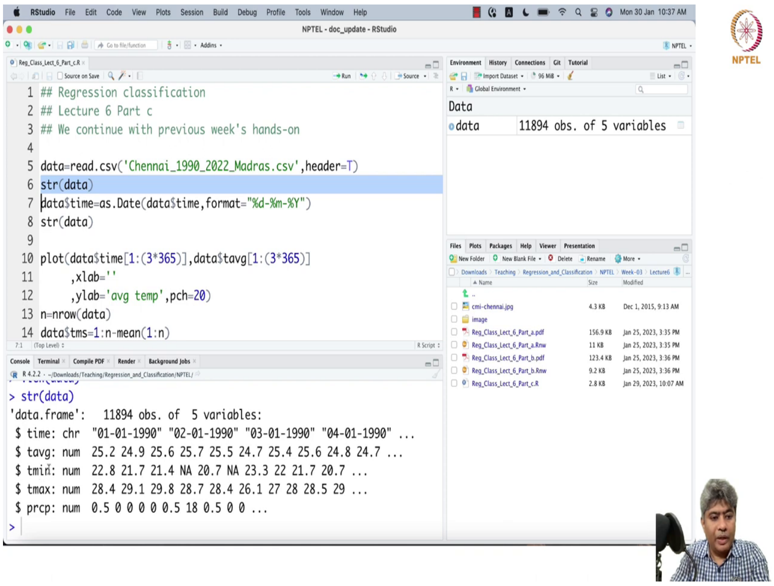


(Refer Slide Time: 01:33)



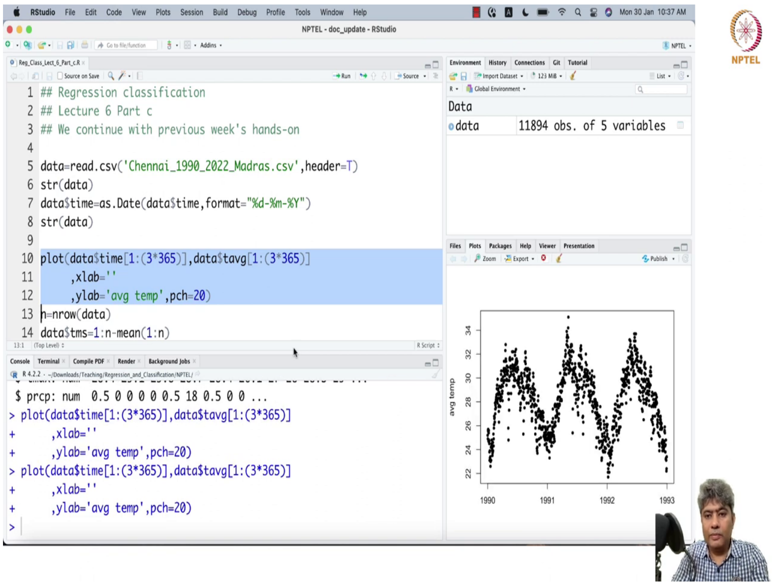

So, first, so this is the data set that we have. We have seen this before, last week.

(Refer Slide Time: 01:41)



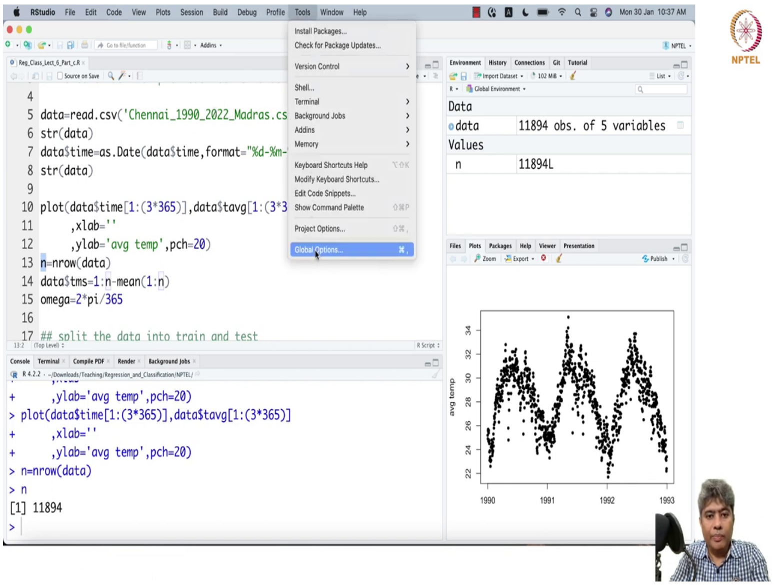

Now, the structure of the data set is everything is numeric but time was taken as character. So, we convert it into time and now you have this as a date format.

(Refer Slide Time: 01:58)



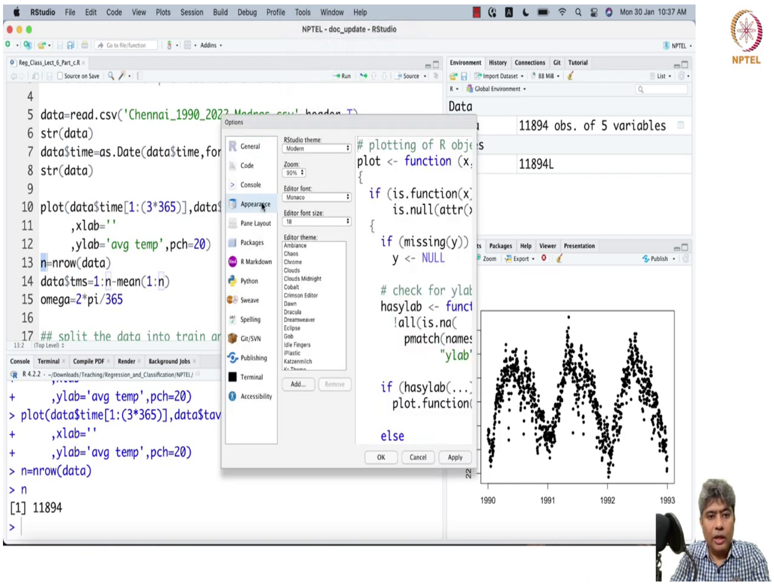
Then, now we plot this. Now, it is the data set. We have 118895 data points.

(Refer Slide Time: 02:11)

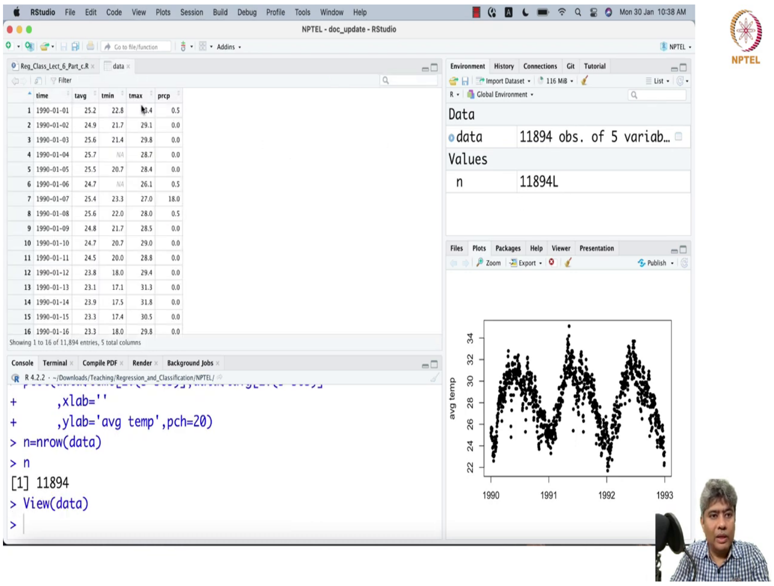



Let me increase the font size a bit. You go to tools, you go to global options, then appearance (Refer Time: 02:25) maybe zoom 200 percent and apply and ok, right.

(Refer Slide Time: 02:21)

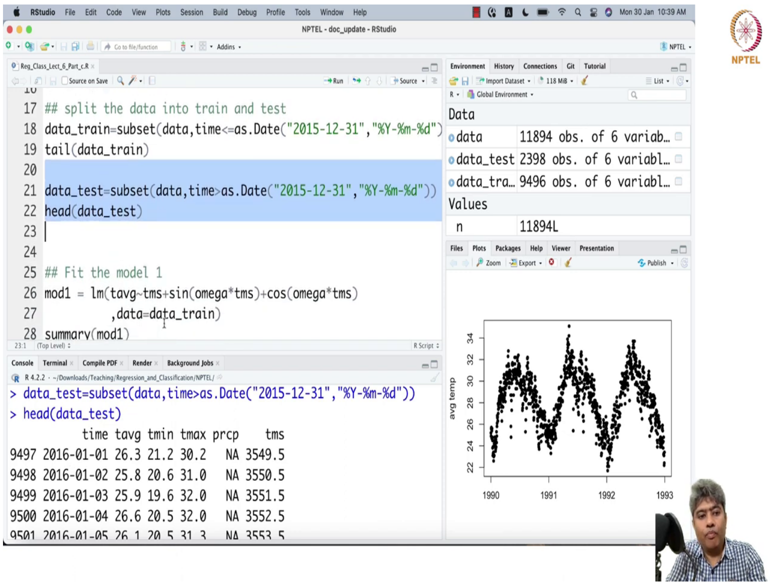


(Refer Slide Time: 02:42)




And then what we are doing, we are creating another data set another column. So, there is a time t, average temperature, minimum temperature, maximum temperature and precipitation, right.

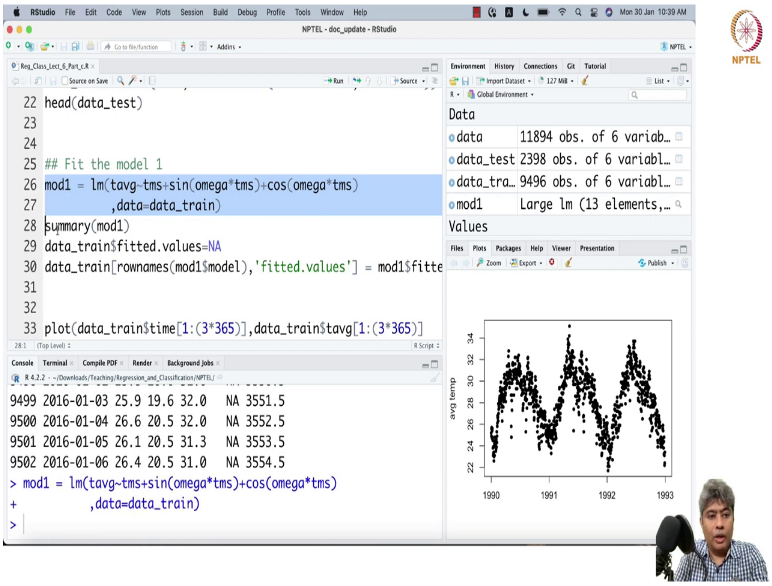
(Refer Slide Time: 03:14)



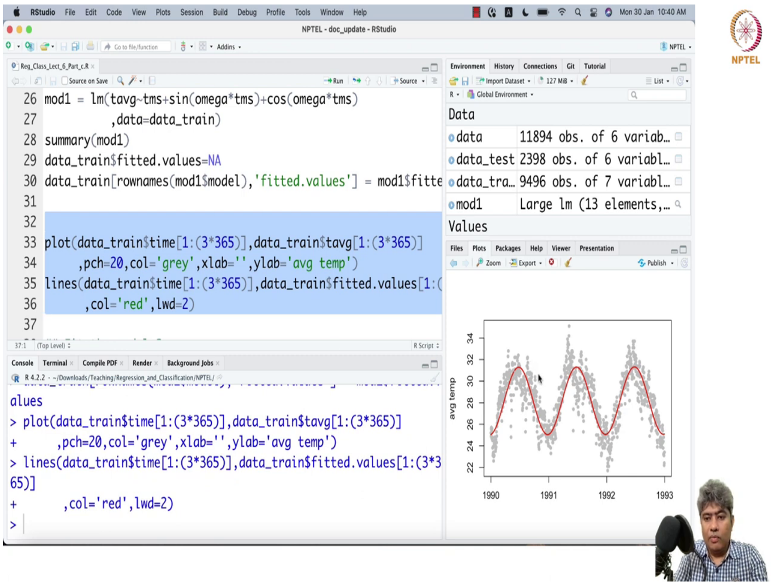

Now, I am creating a new column omega, ok. Now, after creating column, you can see that what we have done. I have subtract the average time; just location, I just shift the location and then I split the data into train and test. So, any data before 31st December 2015, I have used it as train data.

And if you just put a tail, so you can see the last value is of the train data is 31st December 2015. And any data before after 31st December 2015 is test data. So, I am just plotting the first few rows. So, from 1st January 2016, we are going to use it as test data. So, then we fit first model, ok.

(Refer Slide Time: 04:06)

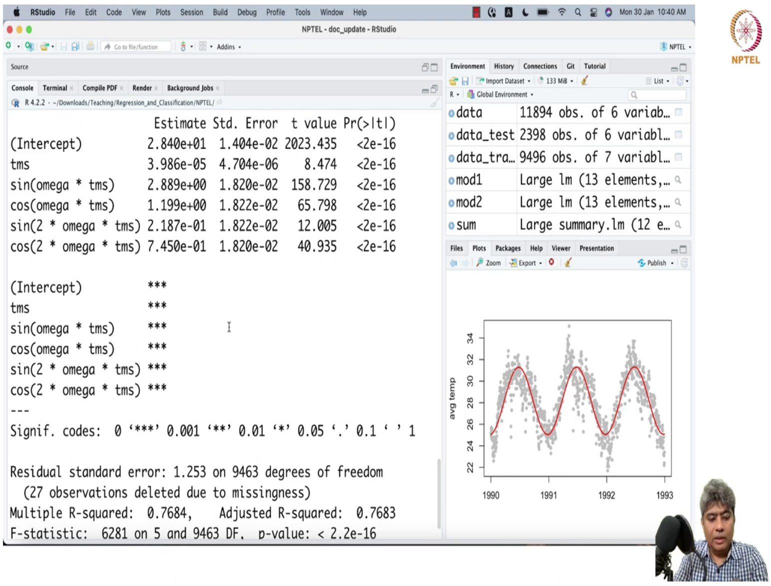


(Refer Slide Time: 04:30)



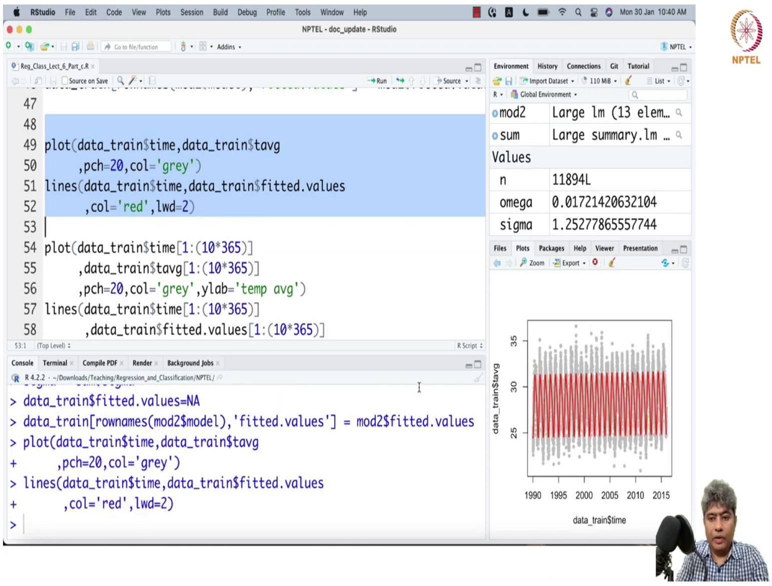

And this is the first model that we fit it. You can see with one that we did last week this model we fitted last week. Then, we created took the fitted model, fitted values and plot the fitted values through the average temperature.

(Refer Slide Time: 04:46)

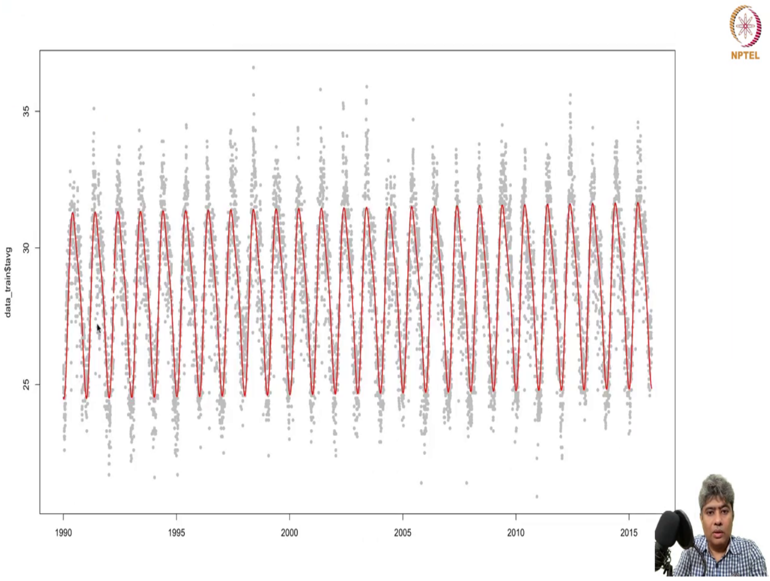



Then, we fit the second model, ok. And what we are seeing that what we are seen the second model with sin 2 omega type t plus cos 2 omega t, they are all significant and name various fitted model, fit the model.

(Refer Slide Time: 05:17)

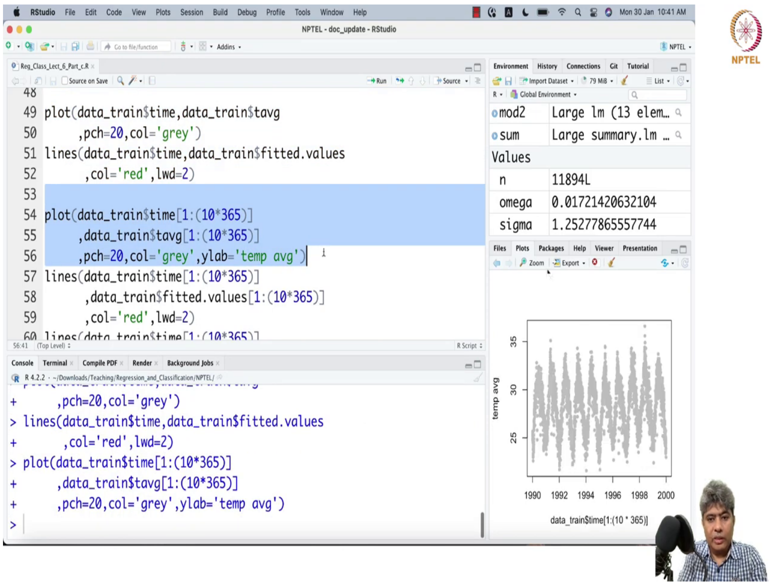


(Refer Slide Time: 05:24)

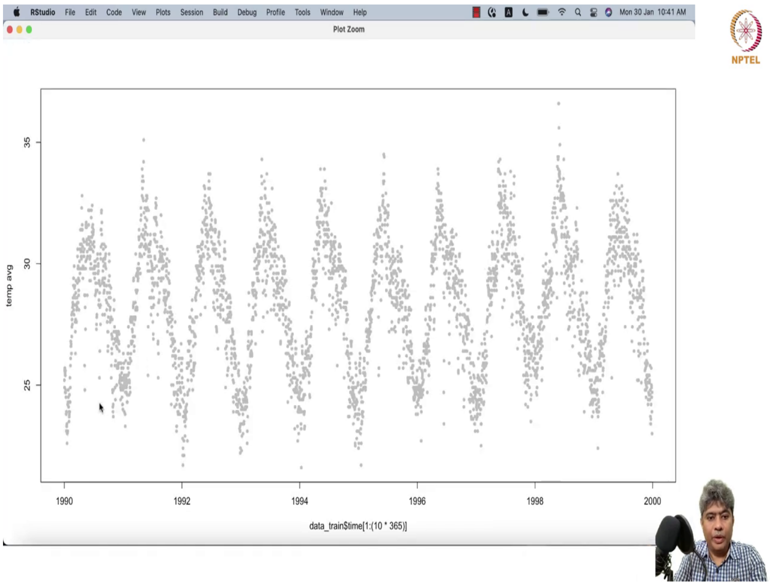



Then, so this is sort of a let me just know. So, this is you can; so, every year we have some kind of sine cosine behavior. This is kind of expected. At the same time, it is kind of going up. We can see there is an increasing trend, model is picking up a increasing trend.

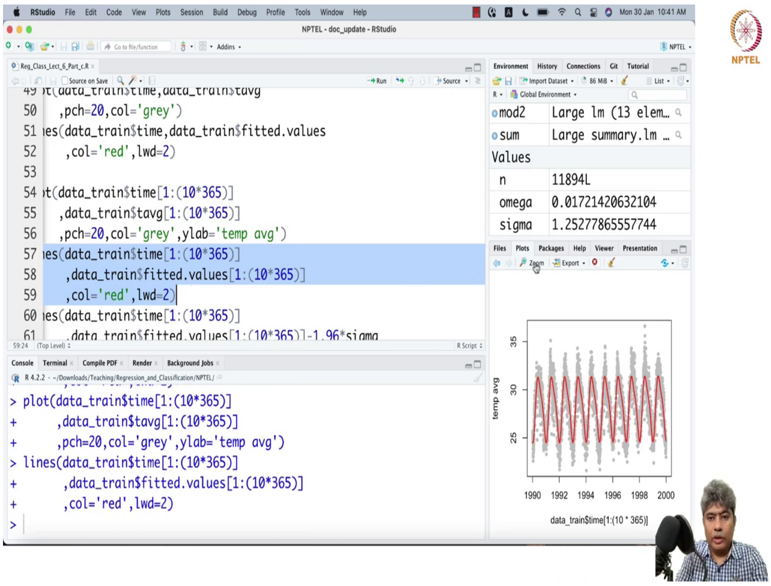
(Refer Slide Time: 05:50)



(Refer Slide Time: 05:53)

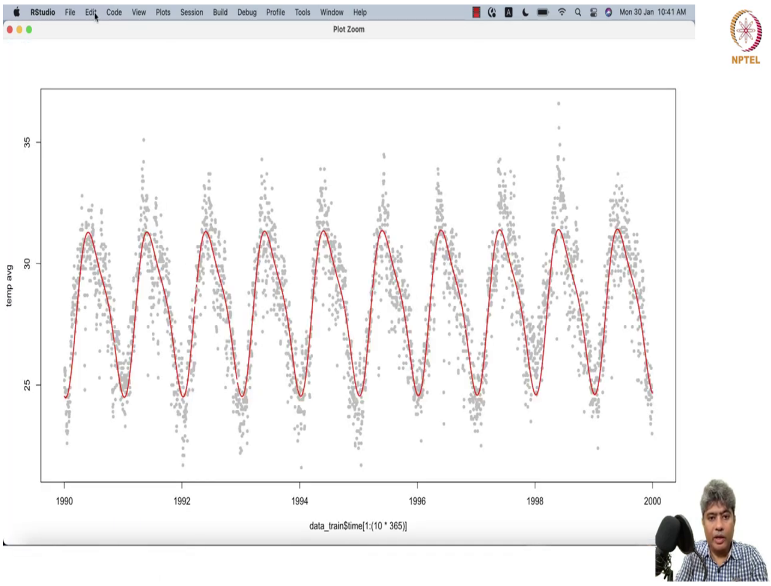


(Refer Slide Time: 06:04)

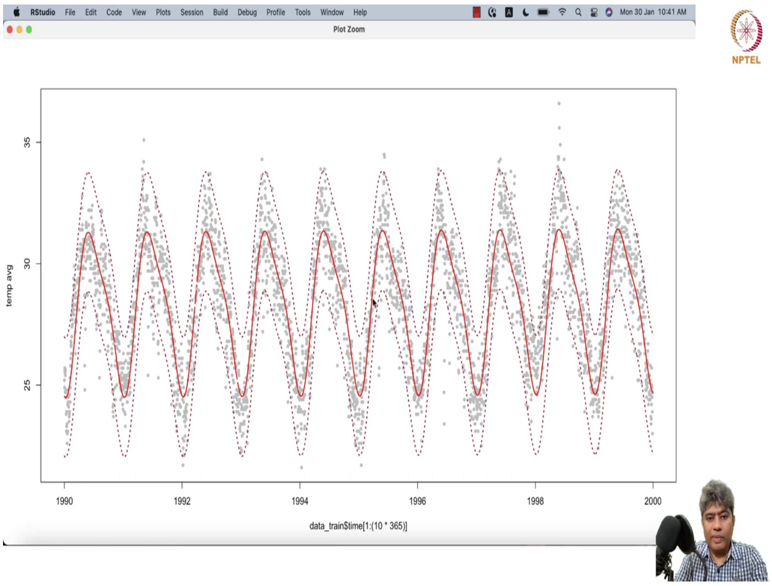


And we just here we just plotted the first 10 years from 1990 to 2000. And so, here is the first 10 years average expected behavior.

(Refer Slide Time: 06:06)

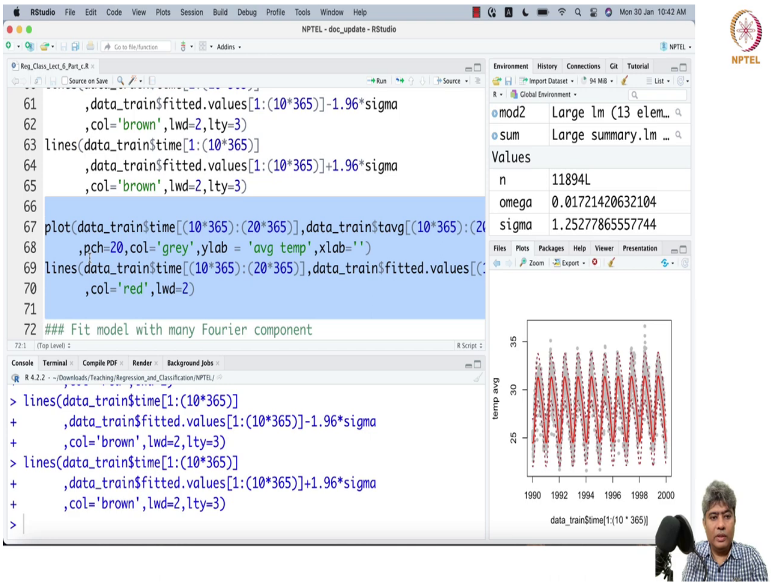


(Refer Slide Time: 06:17)



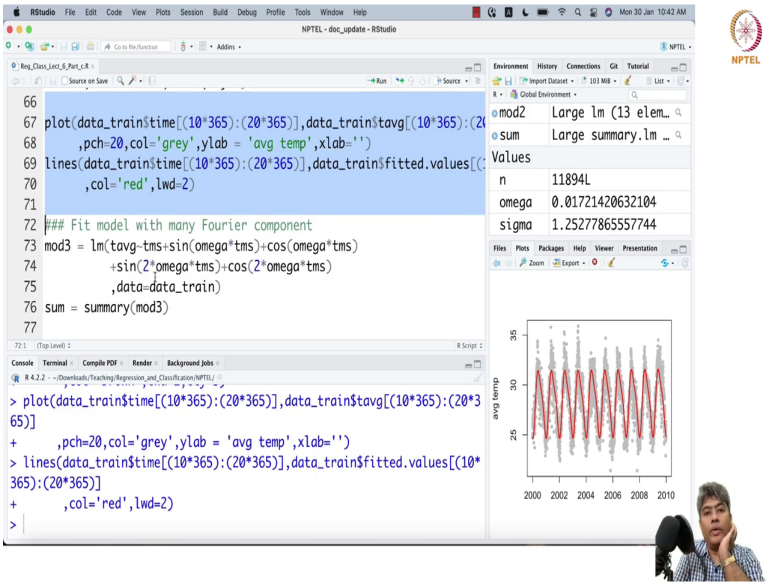

And then, if we just we have a 95 percent confidence band as well, ok.

(Refer Slide Time: 06:29)



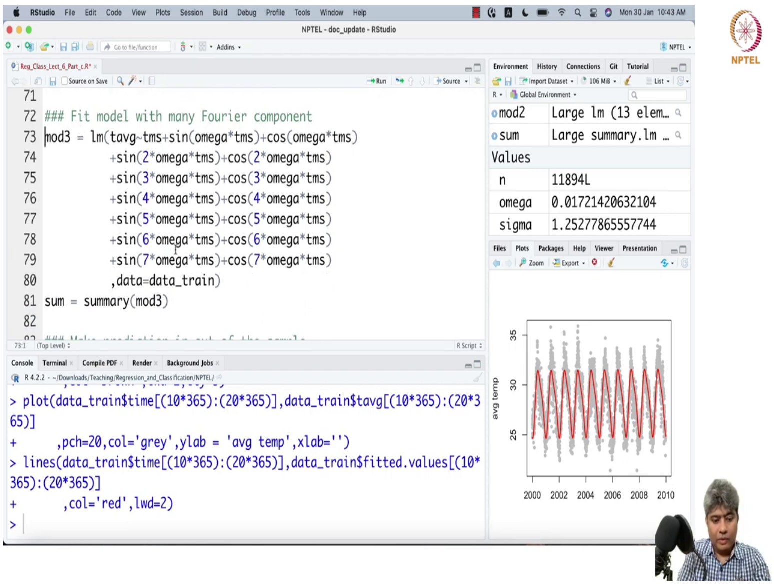

Now, and this is the from 2000 to 2010, ok.

(Refer Slide Time: 06:37)



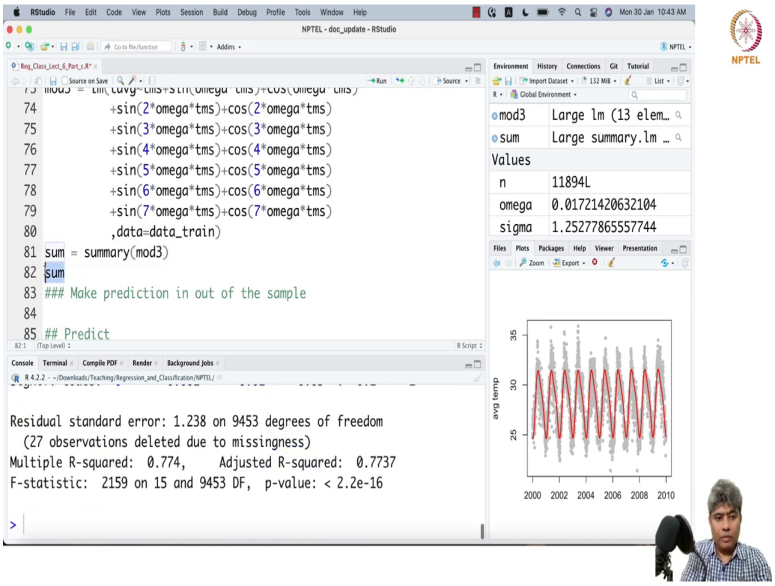

Now, we are going to add why we are only going to stop at 2 omega and cos 2 omega. We can in add as many cases as we want. Say may be up to 7 omega or 8 omega, ok 4, 4.

(Refer Slide Time: 06:47)

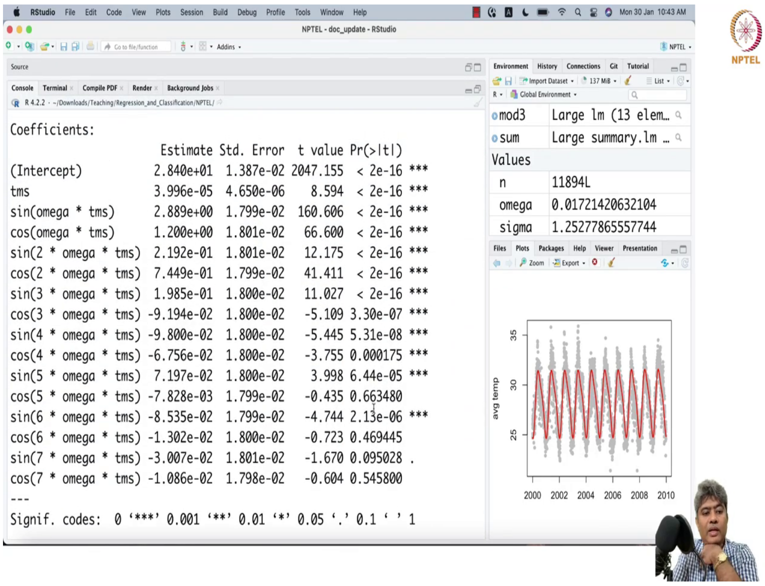



So, I am adding as many engineered feature as possible and I am not worried about over fitting much because I have too many datas, almost 11000 data points I have. So, I am not worried, ok. So, let me just try 7 Fourier terms.

(Refer Slide Time: 07:44)

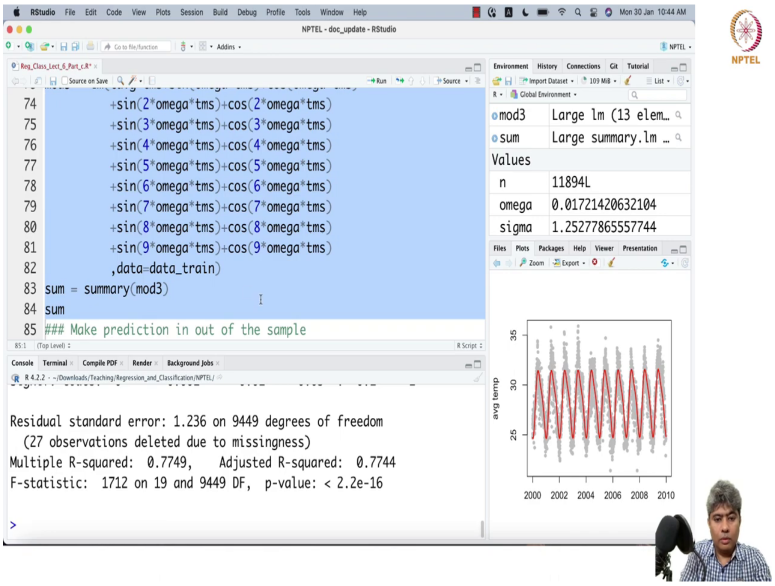


(Refer Slide Time: 07:51)

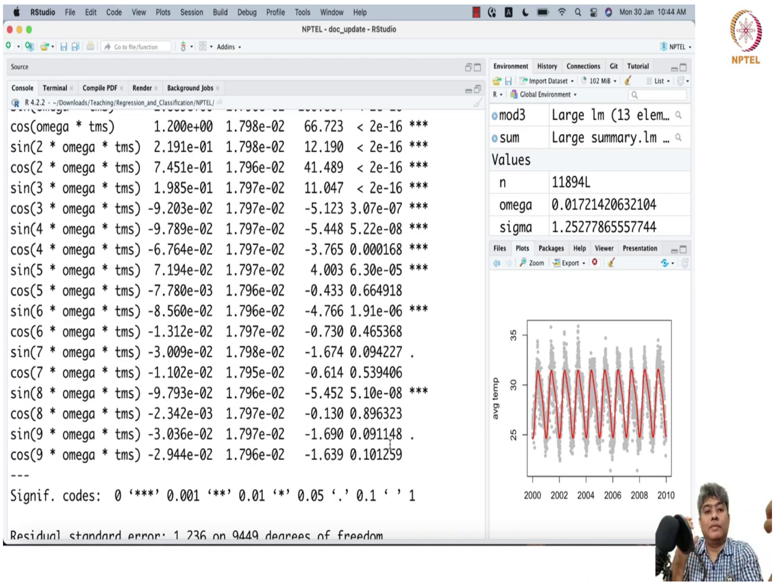



And then, let us see, we can see that there are some Fourier terms like cos 5 omega is not significant. Actually, in fact, after 3 omega onwards the significance has dropped to an extent though sin 6 omega does have a effect, but cos 6 omega 7 omega they do not have a effect. You can even try few more, if you want know never know. If in case there are some higher things 9, 9. So, this is yeah.

(Refer Slide Time: 08:20)

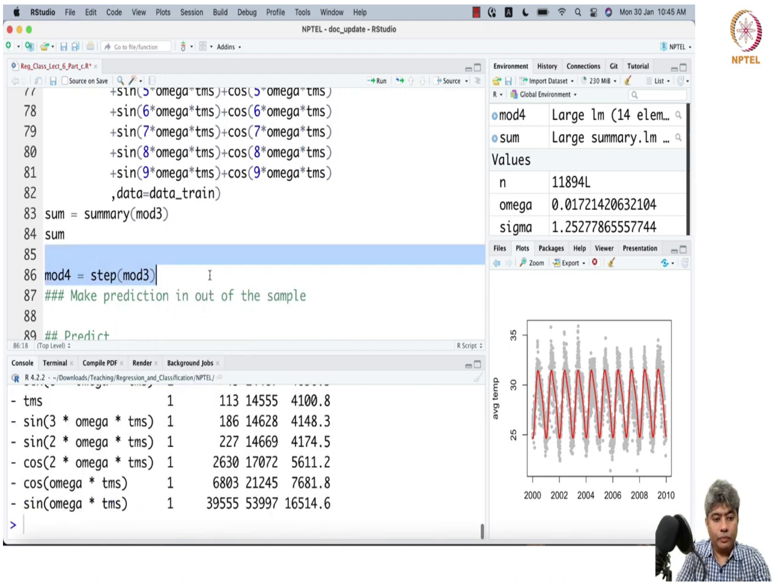


(Refer Slide Time: 08:47)




So, yeah from 5, 6, 7, 8, may have, but 9; 8, 9 they do not have much effect. So, what we will do? We will is; we do not want to keep the keep growing the model size, ok. And if we do that we will see that there might be eventually over fitting which is happening.

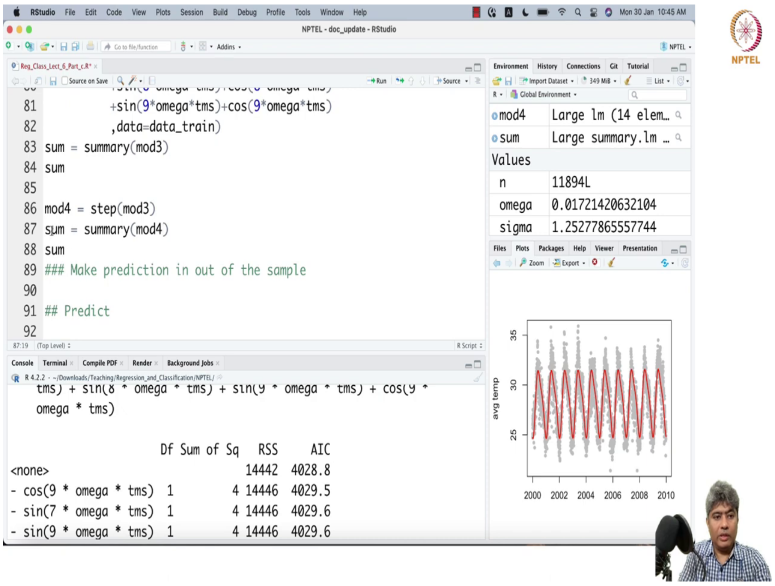
(Refer Slide Time: 09:16)



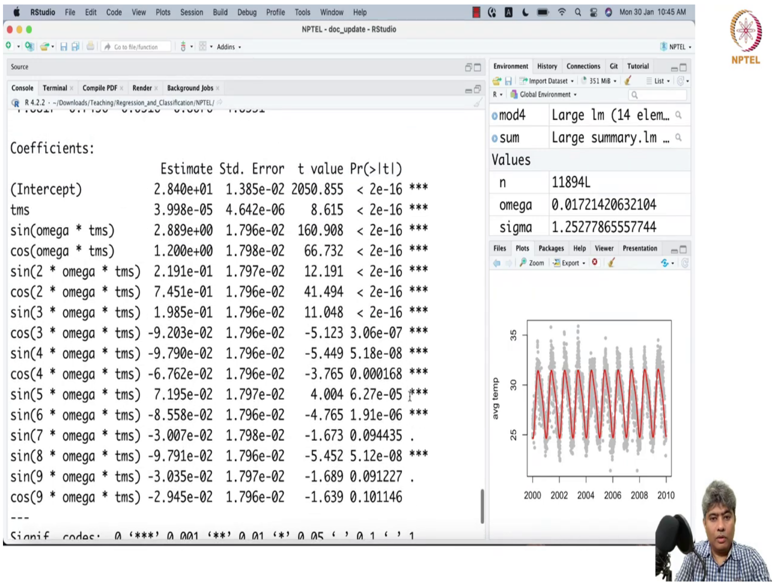

So, what we will do? We will keep this model, then the same time we will fit another model for model 4, model 4 and we will apply step wise variable selection on the model 3. So, if we just, you see it just fitted these models, so these models were fitted, ok. So, yeah here, from here.

So, this is the model that was fitted with this AIC, and then this is the model that was fitted, and eventually, it settled down to a model and let us see what was that.

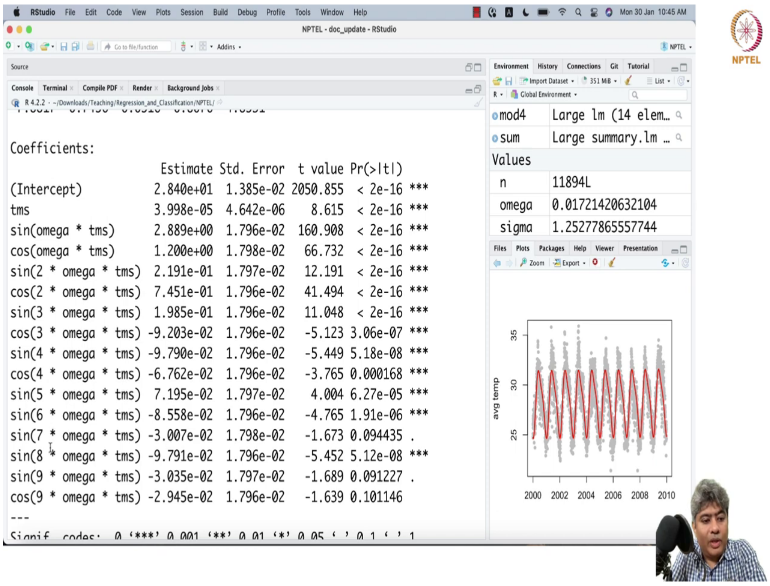
(Refer Slide Time: 10:11)



(Refer Slide Time: 10:18)

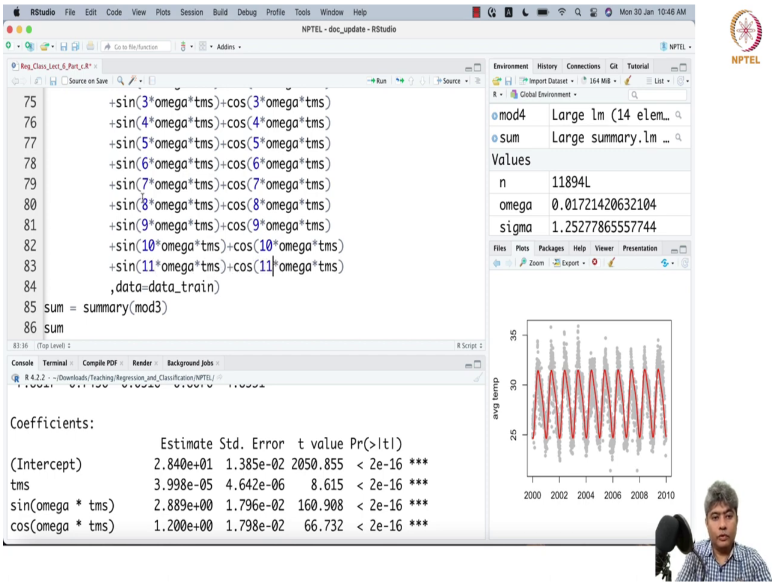


(Refer Slide Time: 10:24)



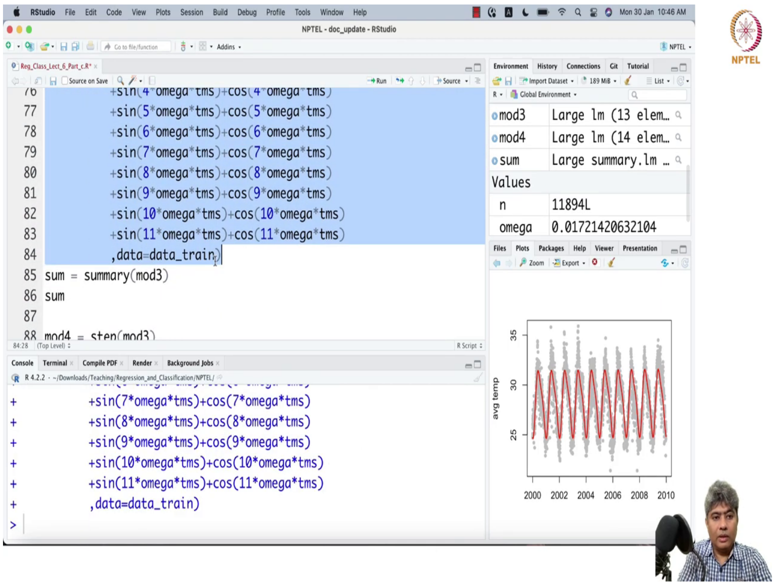
So, I can just copy and paste it here, and I taking some summary. So, it has drop the cos terms and keep only the sine terms that has effect in the model though it is it did not throw away the cos 9.

(Refer Slide Time: 10:41)

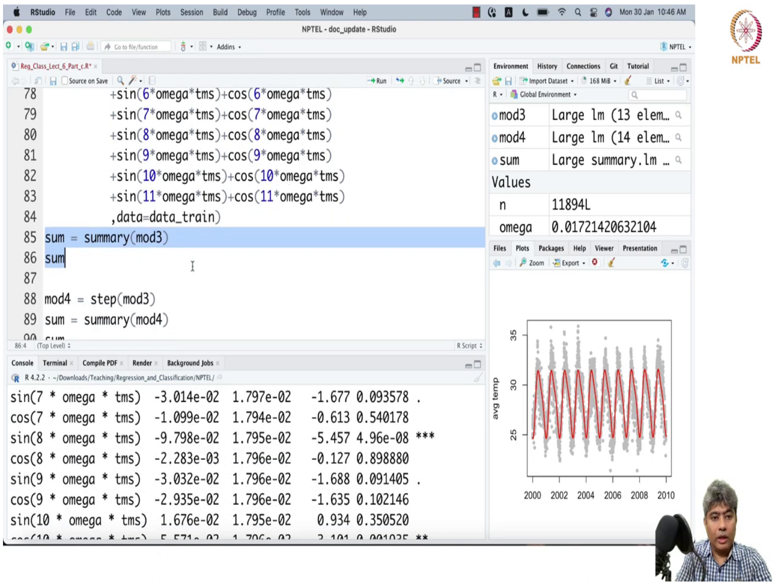



So, we can actually; that means, we can add few more terms.

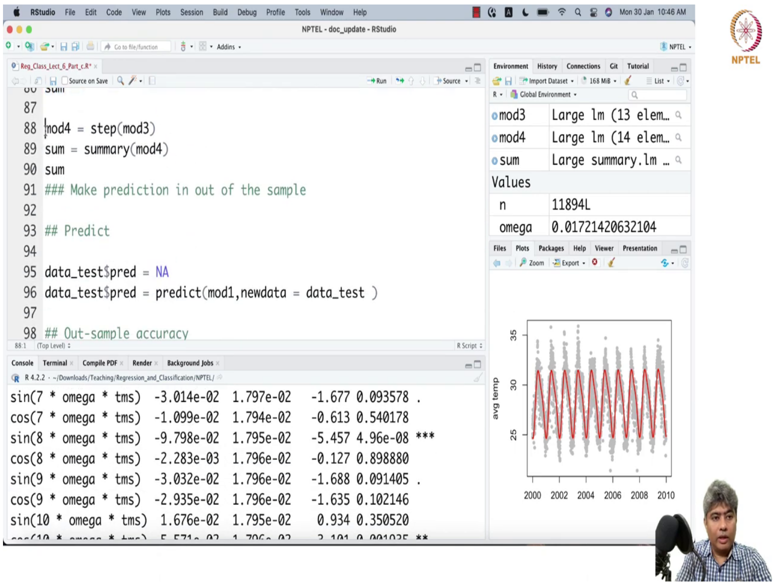
(Refer Slide Time: 11:00)



(Refer Slide Time: 11:09)

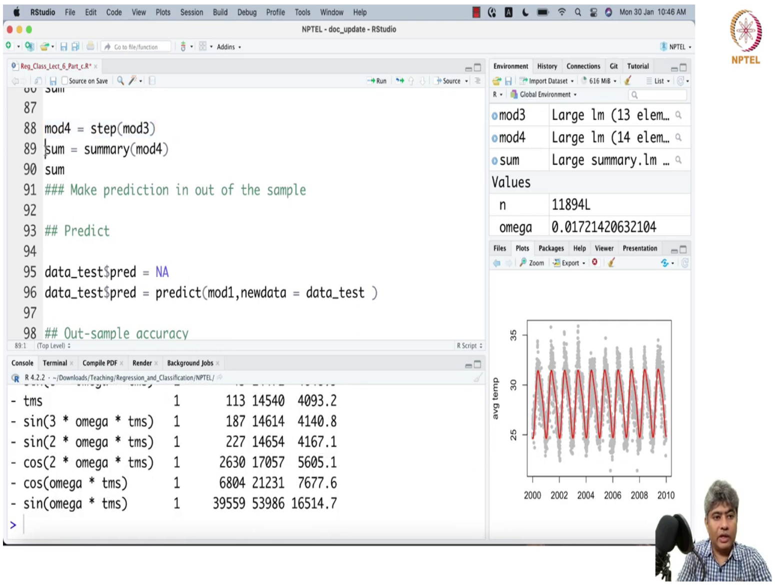


(Refer Slide Time: 11:10)



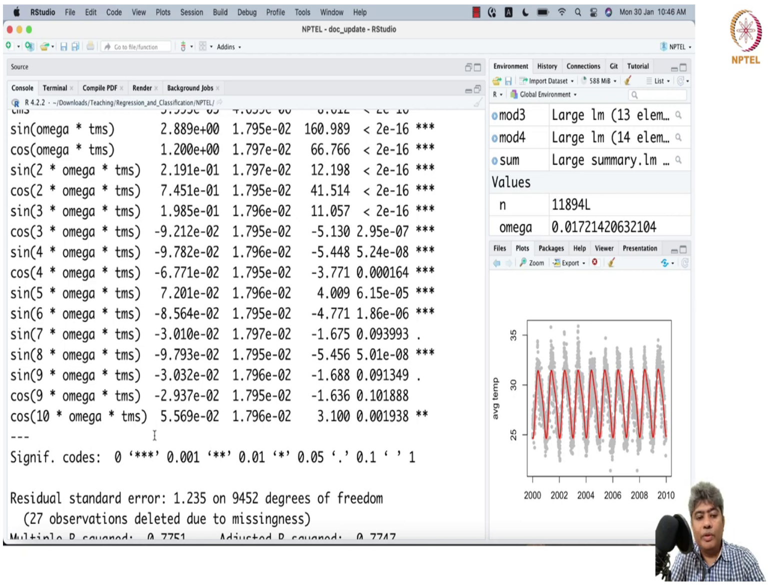
And let me try this and so yeah looks good, right.

(Refer Slide Time: 11:16)



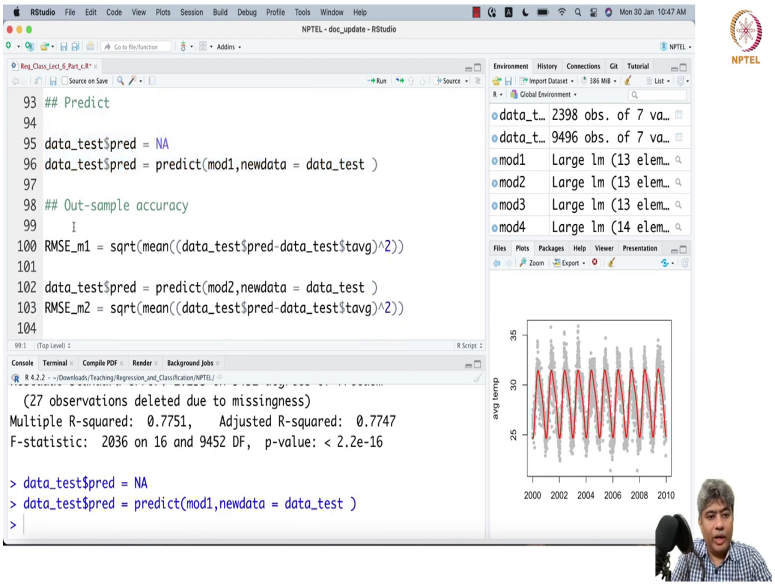

And then let me just run this stepwise forward section, and see just run and let me just write.

(Refer Slide Time: 11:21)



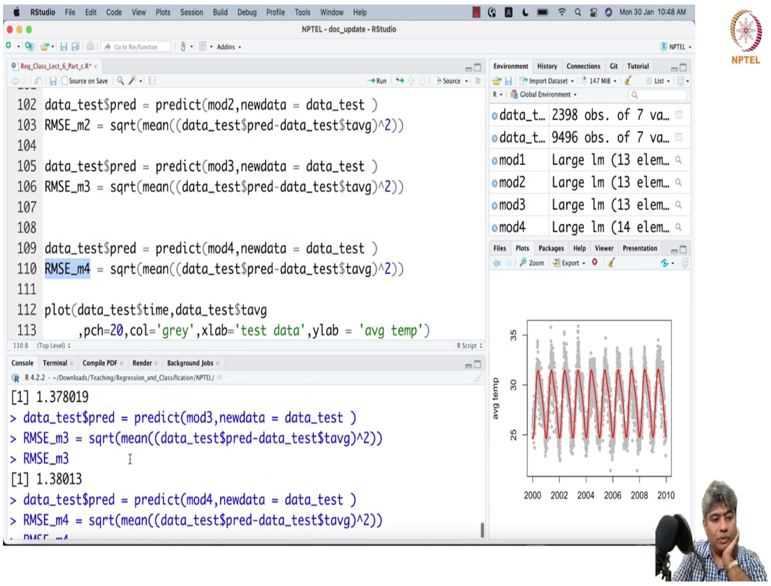

So, now it has completely drop the 11th Fourier term and it stopped at 10th Fourier term. So, we can just stop at here. So, we can just. Now, what we will do? We will do the prediction. So, first we will do prediction and we will calculate the out of the sample accuracy.

(Refer Slide Time: 11:46)




So, in the test, we are creating prediction and here is the out of the sample accuracy and then we will do. So, what was the RMSE? So, 1.5 c for the first model. Then, if we do prediction for the second model and I may see would be 1.37.

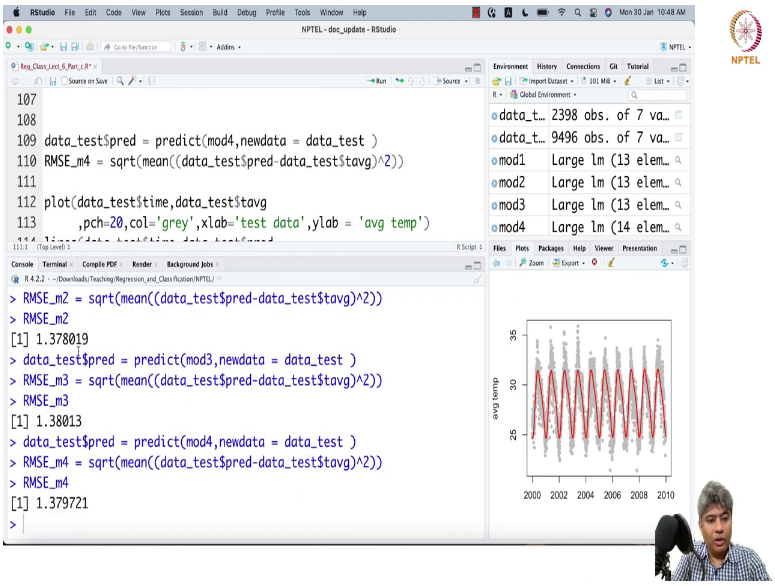
(Refer Slide Time: 12:15)




Now, if we do the same thing for the third model, let us done that and the RMSE 3 is 1.38. Now, you can see that RMSE has gone up because we have added too many models. So, that we can see too many Fourier term, too many engineered term. And as a result, you can see that there are some over fitting tendencies being picked up, and then, we have the step wise selected, for selective model.

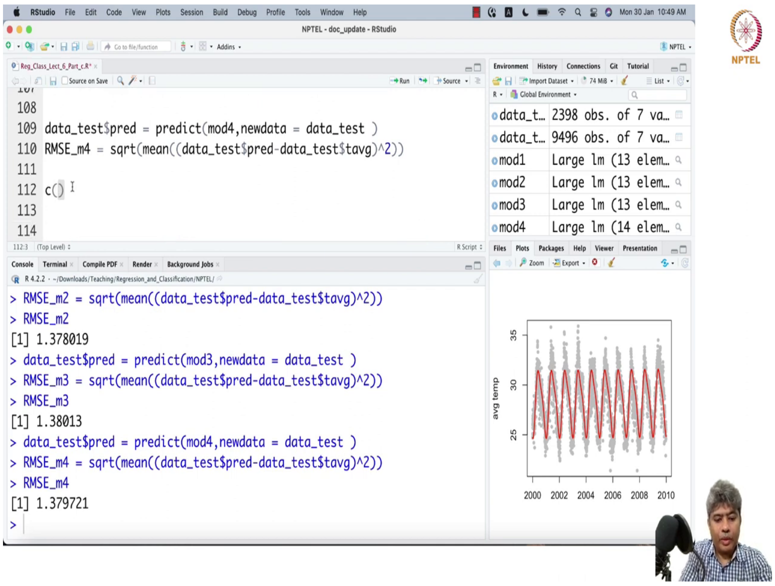
So, from there if we calculate the RMSE out of the sample RMSE is 1.379. So, this is where it is stopping.

(Refer Slide Time: 13:18)



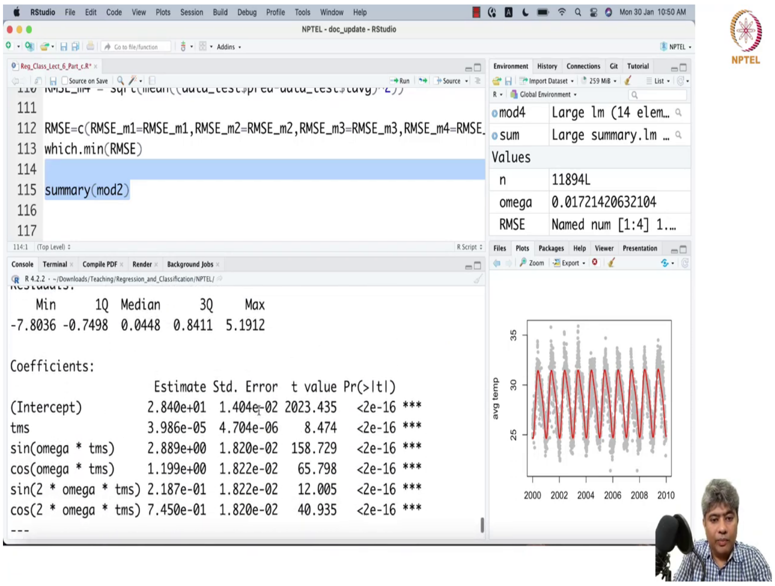

So, either we can go for 4th model or we can go for because end of the day the simpler 1.378. The simpler say second model with 2 sine cosine Fourier transform has a lower RMSE than a bigger more complex model, even if after doing a stepwise selection.

(Refer Slide Time: 13:48)



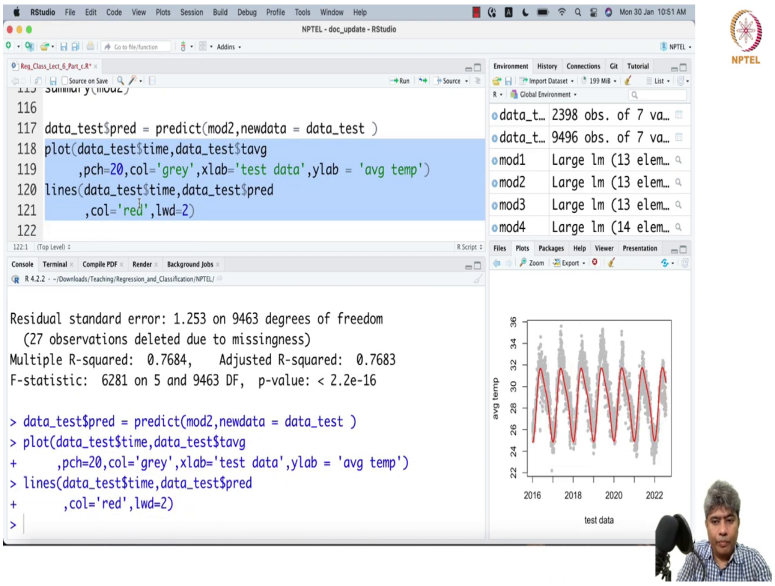

So, if I have to choose out of these 3, I will go for the second model. So, if I have to let me just you know let me just is equal to RMSE 1, RMSE 2 equal to RMSE 2, RMSE 3 equal to RMSE 3, RMSE 4 equal to RMSE 4, ok. So, clearly, if I just say RMSE is this and which dot min of RMSE; clearly the second one is the RMSE m2, second model has the maximum minimum RMSE.

(Refer Slide Time: 13:48)



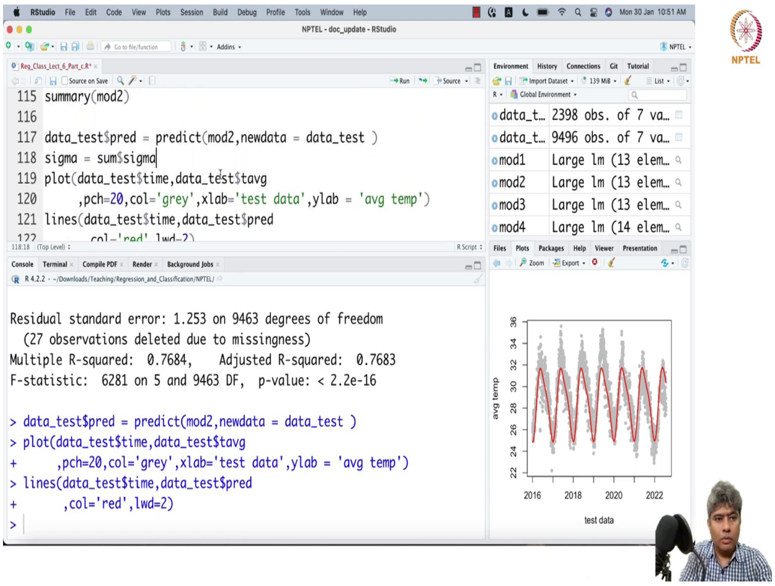

So, even if after doing the statewise selection, significant dimension reduction, the model is still complex and doing some bit of over fitting perhaps. So, we can do summary of mod 2.

(Refer Slide Time: 15:33)

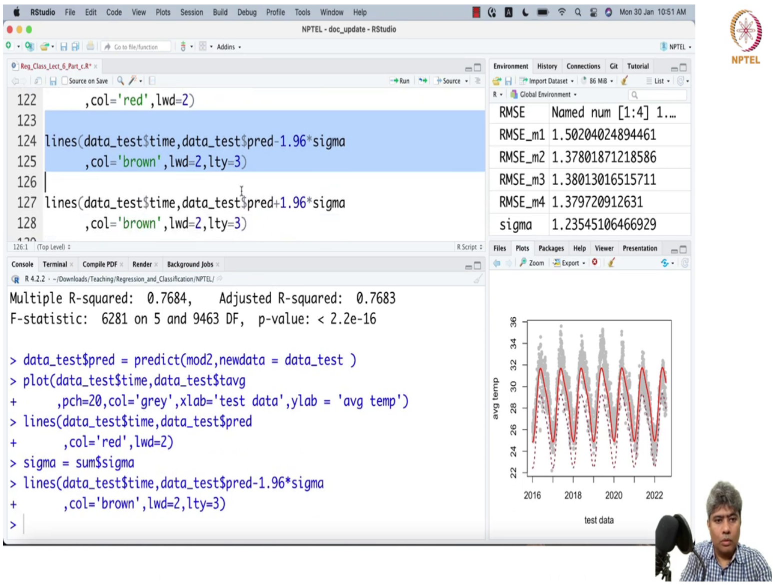



So, we can look into the parsimonious model. This is our parsimonious model. Perhaps, we just be happy with this. We just take this line and run once more. With this, we want this and sigma, only the sigma of this, correct sigma for this model. We have to just go up.

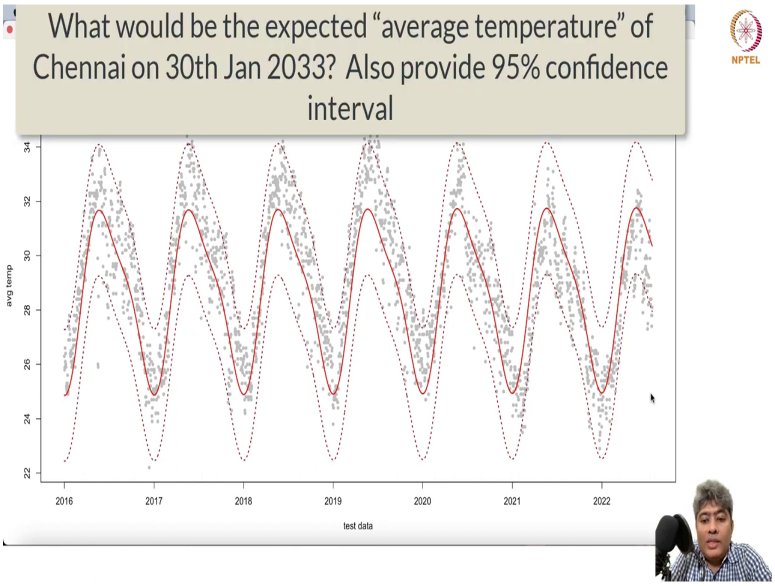
(Refer Slide Time: 16:09)



(Refer Slide Time: 16:19)

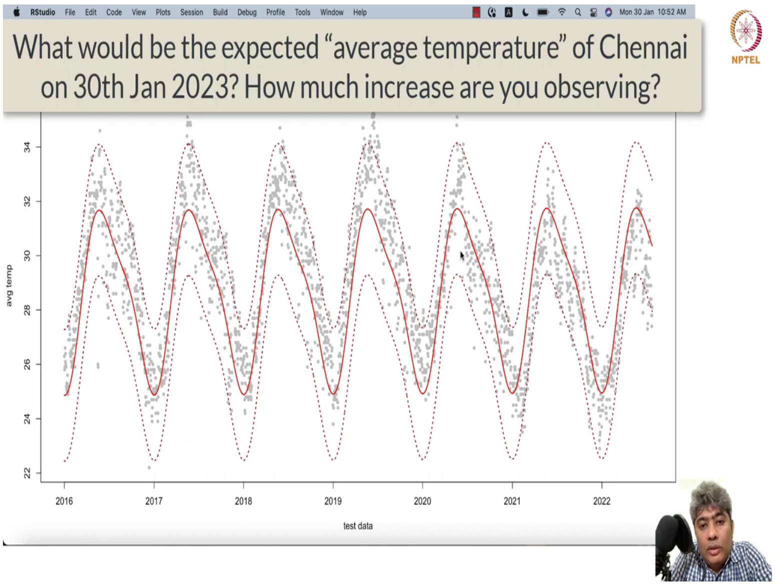


(Refer Slide Time: 16:26)




So, now if we just look into this, so this is our final out of the sample performance from 2016 to 2022. Now, here is one self-assessed assignment for you guys use this model and can you find what would be the temperature of Chennai, average expected temperature of Chennai in January in year 2033, that is 10 years from today. In January, say 30th January, what would be the expected temperature? Can you with this kind of confidence interval can you figure that out?

(Refer Slide Time: 17:25)




And if, what is the expected temperature and in the 30th January on 2023, and how much increase you are observing. Can you do this self assessed exercise and let me know that what is your findings? That will be really fun exercise. And I think that will be we all want to know that, the answer to that question.

Thank you very much. See you in the next lecture. Bye.