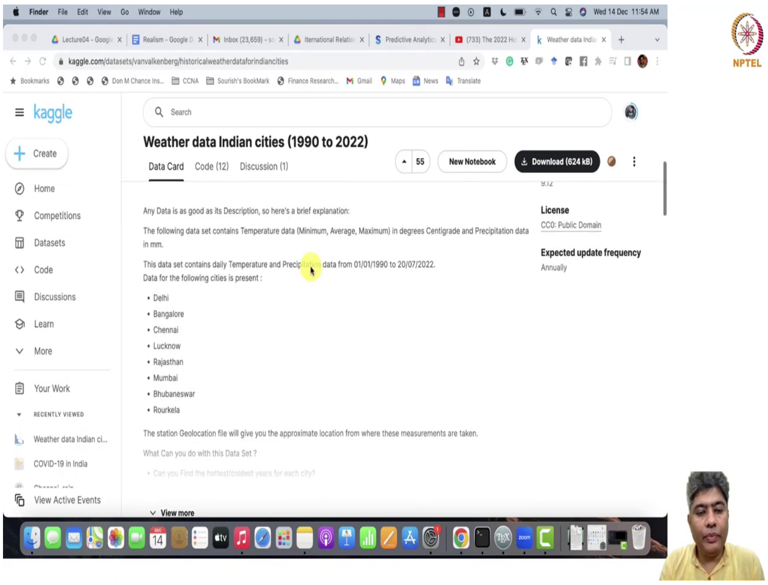
**Predictive Analytics - Regression and Classification**
**Prof. Sourish Das**
**Department of Mathematics**
**Chennai Mathematical Institute**

**Lecture - 16**
**Hands-on with R Part-4**

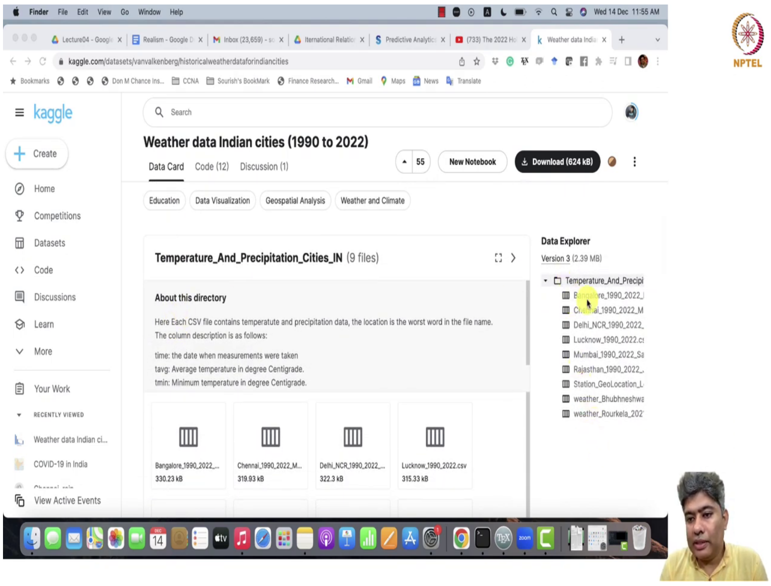
Welcome back to the part C of lecture 4. This part we are going to discuss the hands on, we are going to do some hands on of fitting linear models with feature engineering etcetera.

(Refer Slide Time: 00:36)



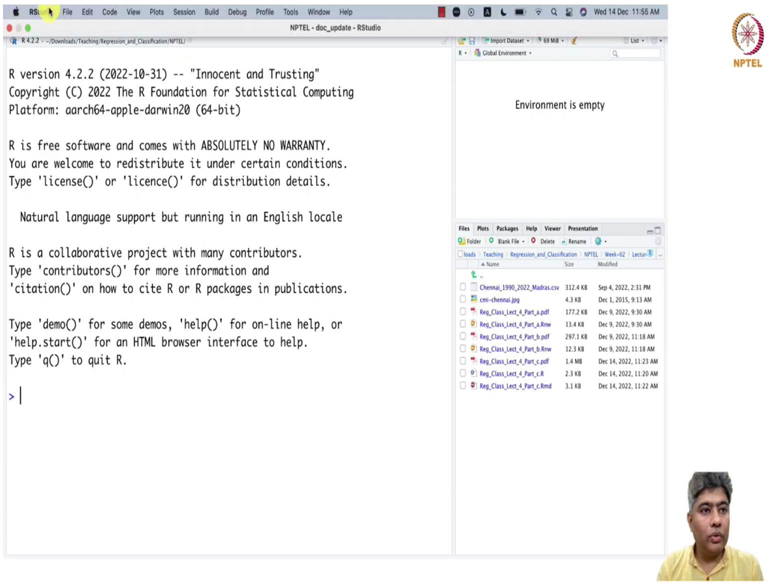

So, for this we are going to use the weather data of Indian cities from the Kaggle. I am going to share this data, this website this thing this page with you guys, as well as if you I am going to share the data. So, we are going to use the data for Chennai, you can try the same thing with other cities as well.

(Refer Slide Time: 00:55)

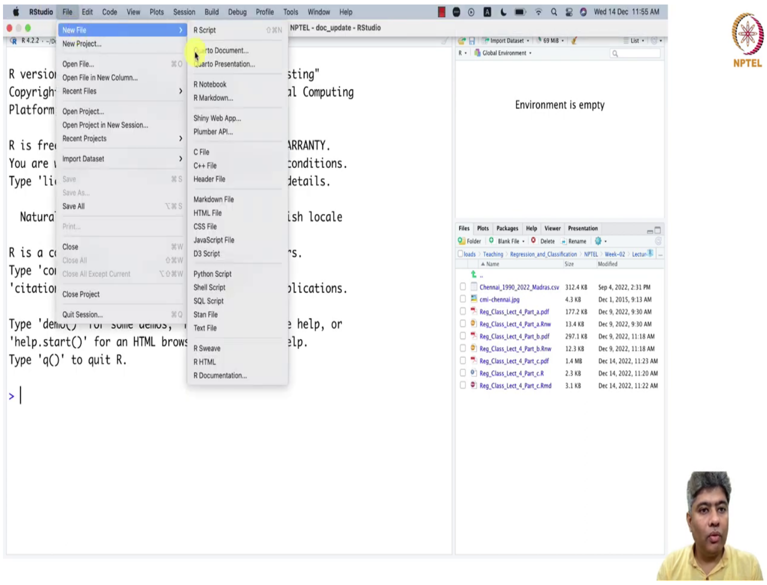



So, here is all the data sets are available. So, we are going to use the Chennai data set. The data set is available from 1990 to 2022, ok.

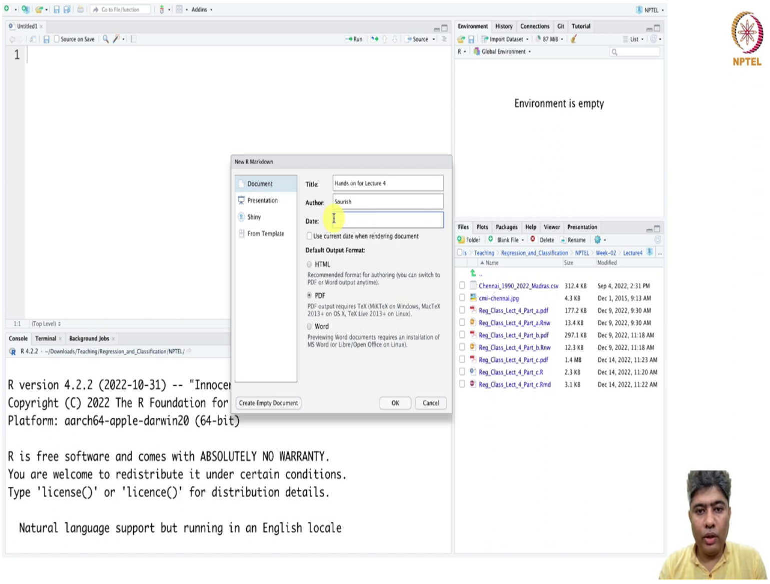
(Refer Slide Time: 01:15)



(Refer Slide Time: 01:17)

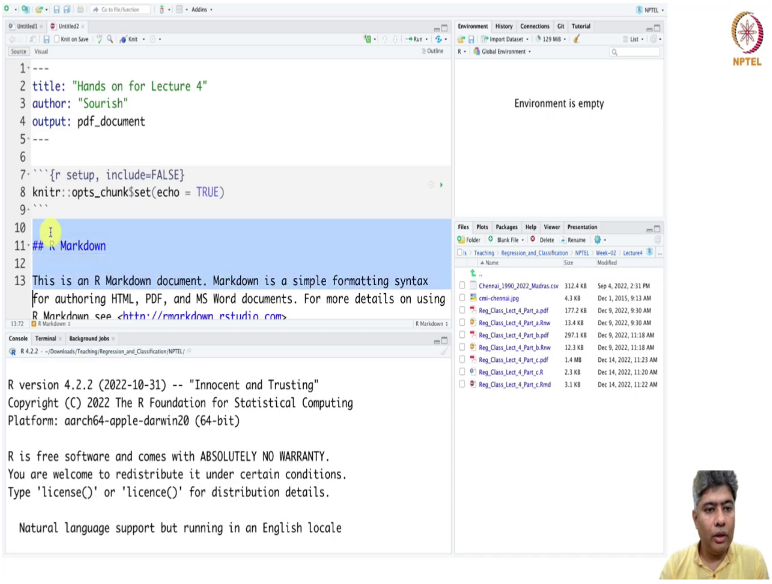


(Refer Slide Time: 01:22)



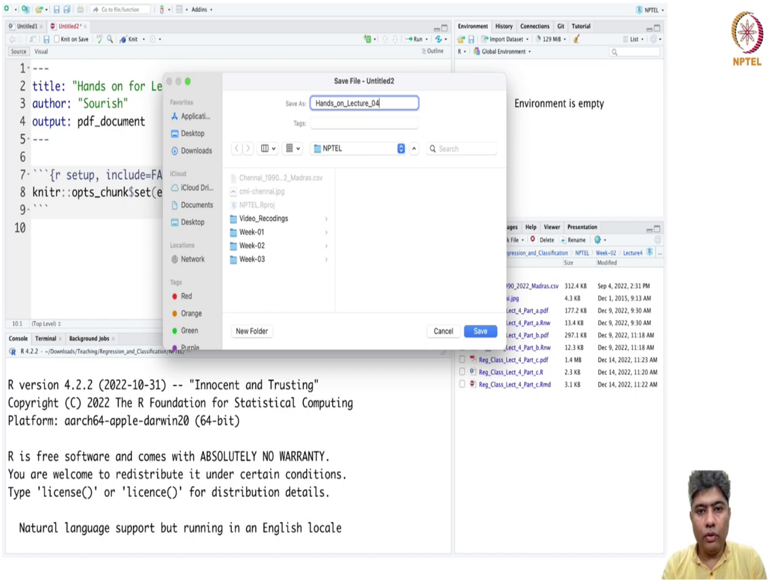

So, what we will do? We will first start R studio, ok. And in the R studio, first thing we will open is R script and an R markdown file, in the file, R markdown file.

(Refer Slide Time: 01:49)

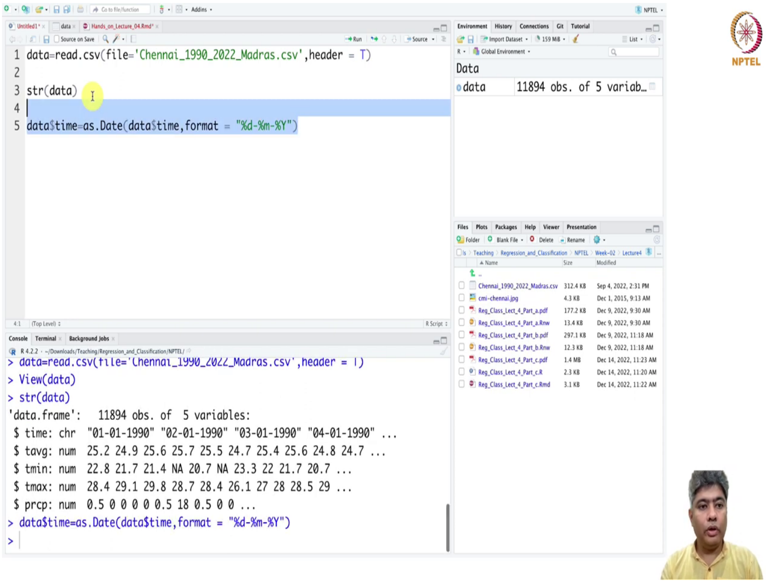



In the R markdown file, we will first write that hands on for lecture 4. We will choose PDF; we do not need, ok.

(Refer Slide Time: 02:02)

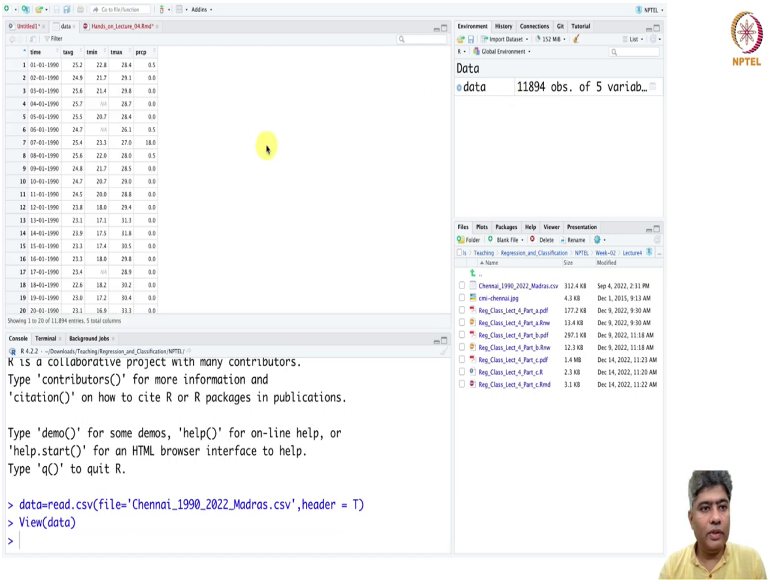


(Refer Slide Time: 02:17)




And then, what we will do, we will just save them as a hands on for lecture 4. And the first thing we have to do is we have to read the data. So, data equal to read dot csv file equal to Chennai, 1990 to 2022, Madras dot csv header equals to true.

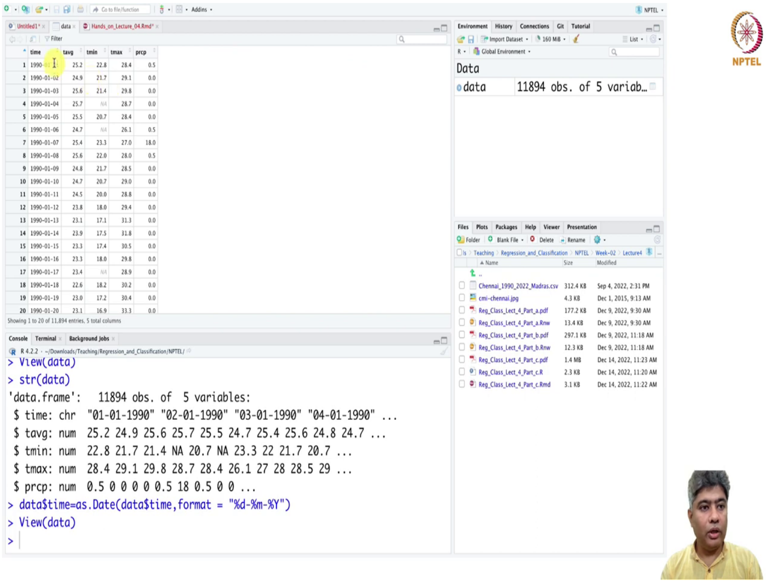
(Refer Slide Time: 02:44)



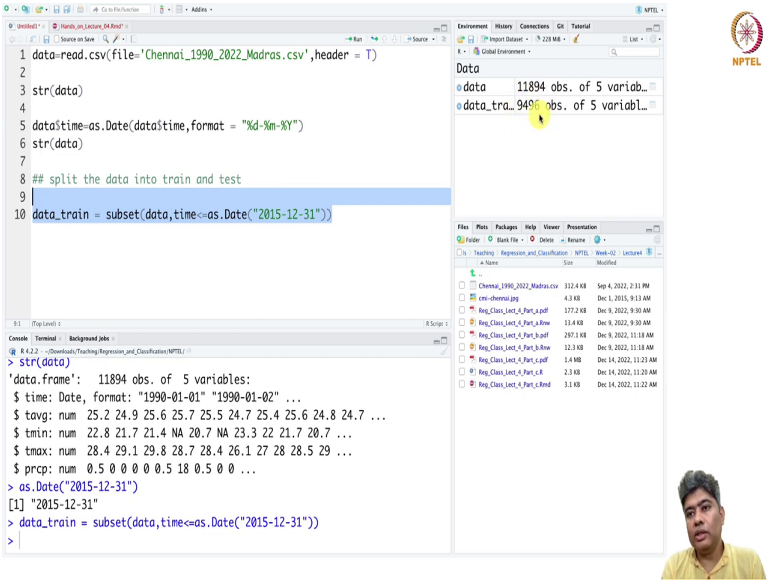

So, here is the data that we have seen. So, this is the first day of the data. So, if you look into the date, date, it is like day first, then month, then year. And we look into the structure of the data, it is the structure of the data is it is a data frame with 11894 observation and 5 variables.

And time is stored as character. So, we have to do it properly as a date fine, as a date column, we have to store it as a date column, data dollar time format should be percentage, date first, then percentage m, then percentage year.

(Refer Slide Time: 03:44)



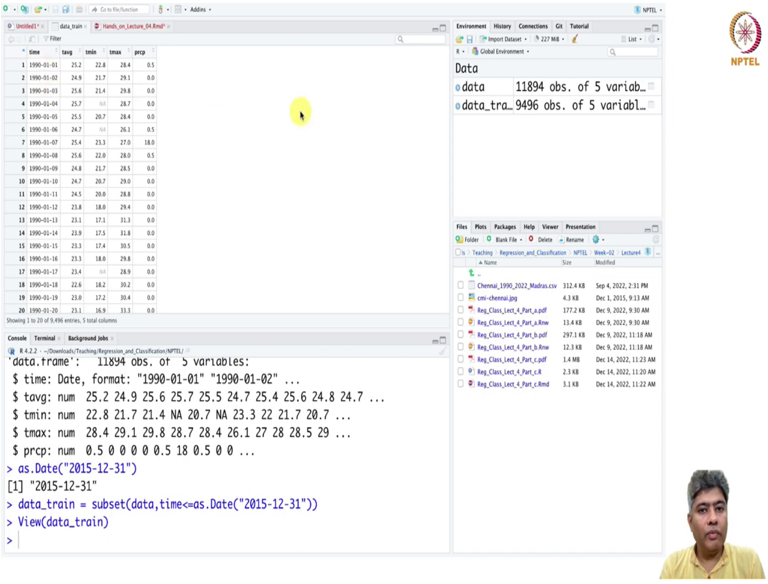
(Refer Slide Time: 03:50)



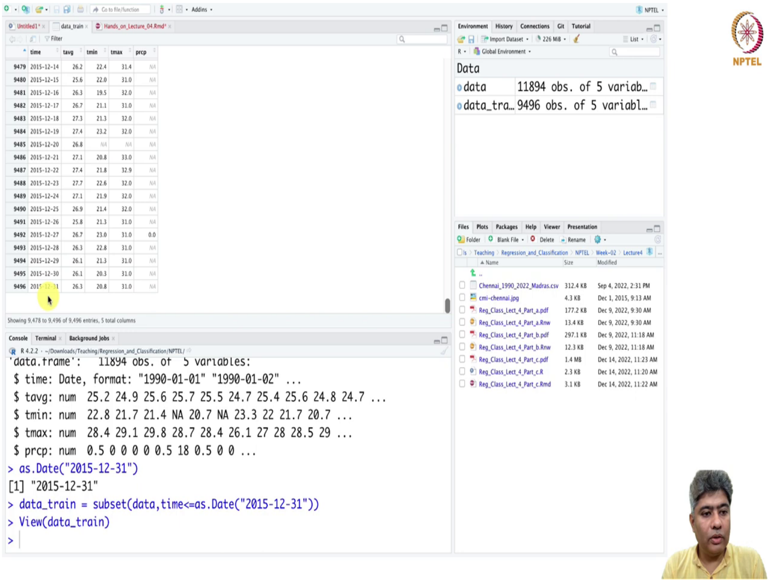

Now, if I do that and if we just check, it is read as this way. And if we know structure of data is (Refer Time: 03:55) this. Now, you can see the time is read as date with format in this format. So, now, it is a date format, not character format. So, the first thing we will see is we will split the data into train and test, split the data into train and test. So, data train equals to subset, data, time as less than equal to as dot Date.

Now, we have to give it in this format. So, date, month and day; year, month and day. So, the first thing will be 2015, then month will be 12, and day will be 31. I hope this is fine and then let us try. So, now, in the train data set you can see 2000 90; 9496 observations are there.

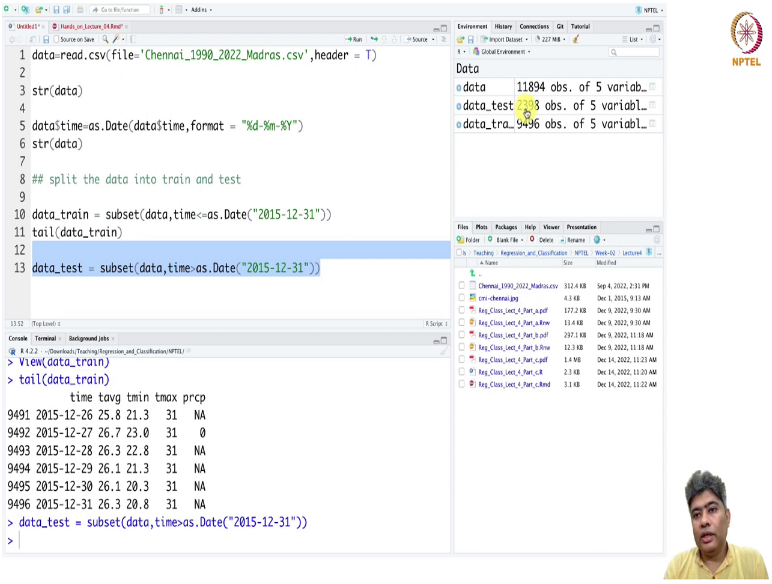
(Refer Slide Time: 05:30)



(Refer Slide Time: 05:34)

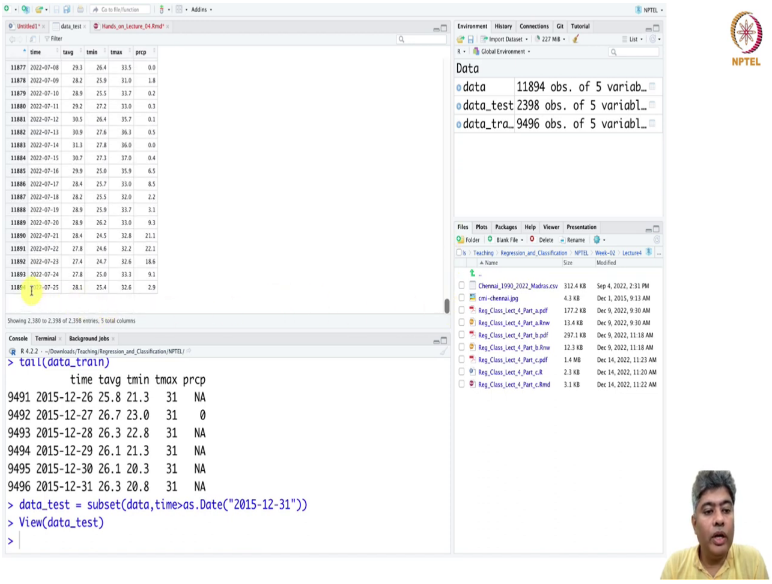


(Refer Slide Time: 05:40)



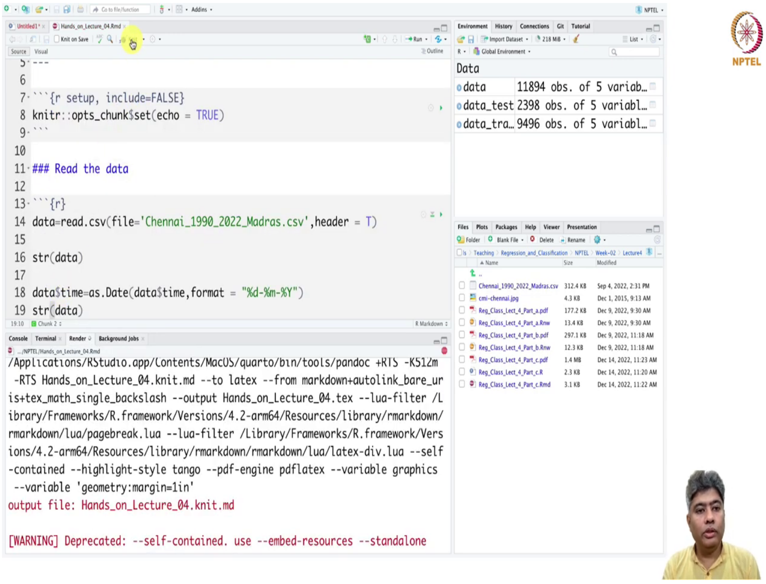

So, it is starting from 1990 and last day is 31st December, 2012. So, also if I just say train tail of the data train. So, up to the last day is 31st December 2015, 31st December 2015. This is the train data. So, same way we can define the test data.

(Refer Slide Time: 06:14)



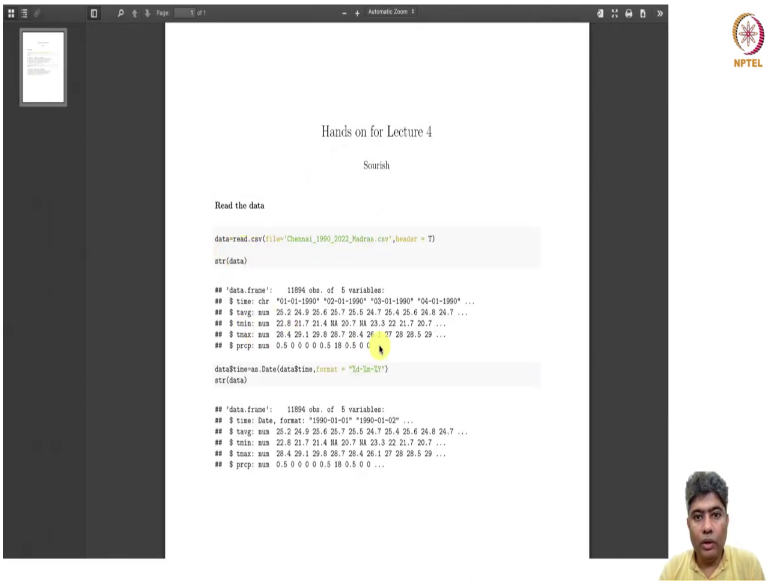

So, we will just take the same thing, instead of less than equal to we will make it simply greater than. If we do that test data has 2398 and the first day is 1st January 2016, is the first day of the data. And the last day of the data is 25th July 2022. So, this is what we are going to do now. What we will do? We will put it in the knitr.

(Refer Slide Time: 06:40)



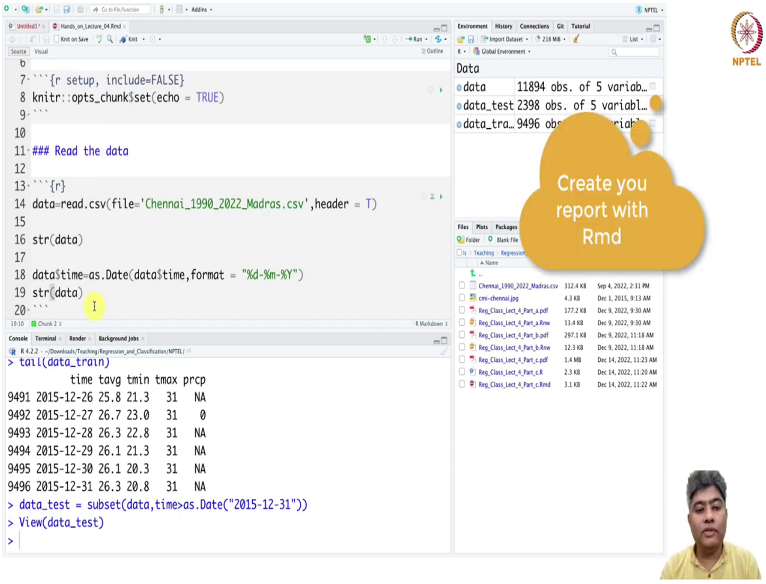

So, first thing is we and we will keep going read the data. So, up to this we can just put it in the read the data part, so here, and then if we just run it.

(Refer Slide Time: 07:10)

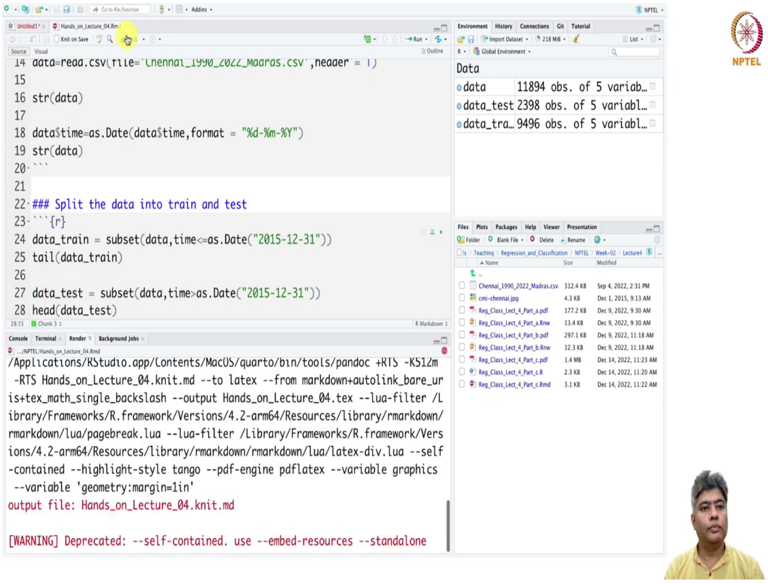



So, here in the R markdown, nicely it is come in a PDF; you can create the report of your analysis alongside whatever you are doing.

(Refer Slide Time: 07:21)

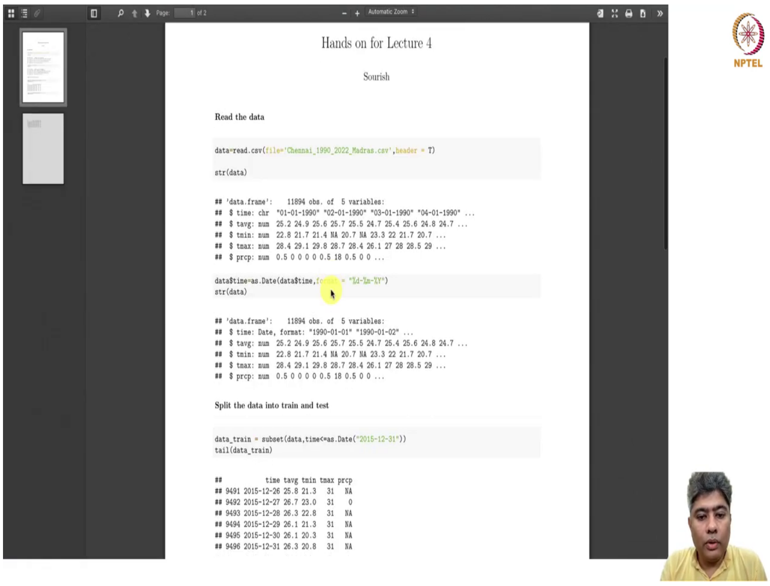


(Refer Slide Time: 07:29)

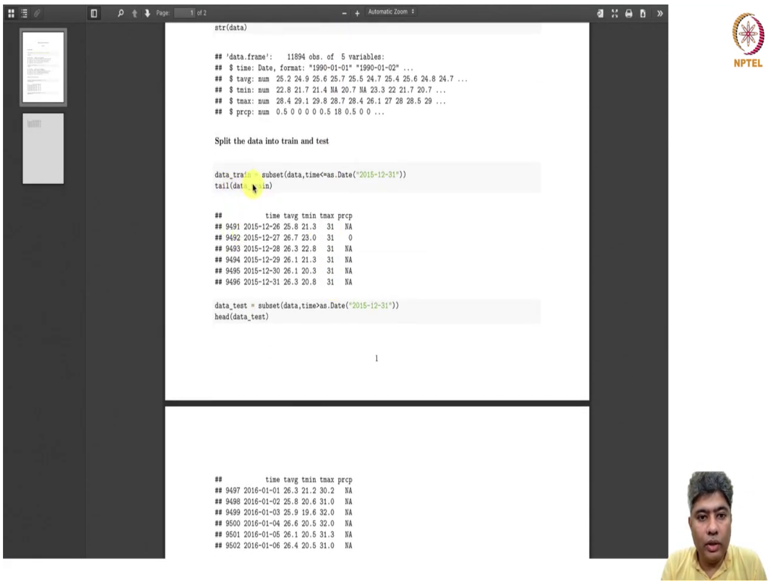



So, this is very important. When you are doing analysis you create the report as well whatever you are doing. So, this is a very handy thing in R. So, weak the data and then split the data into train and test R. So, ok we can just, we can just write it in this way and head data test.

(Refer Slide Time: 08:12)

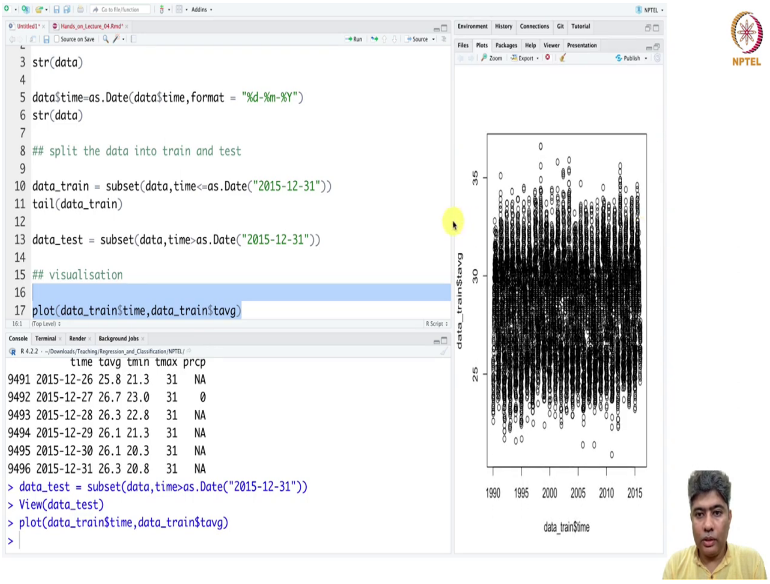


(Refer Slide Time: 08:15)



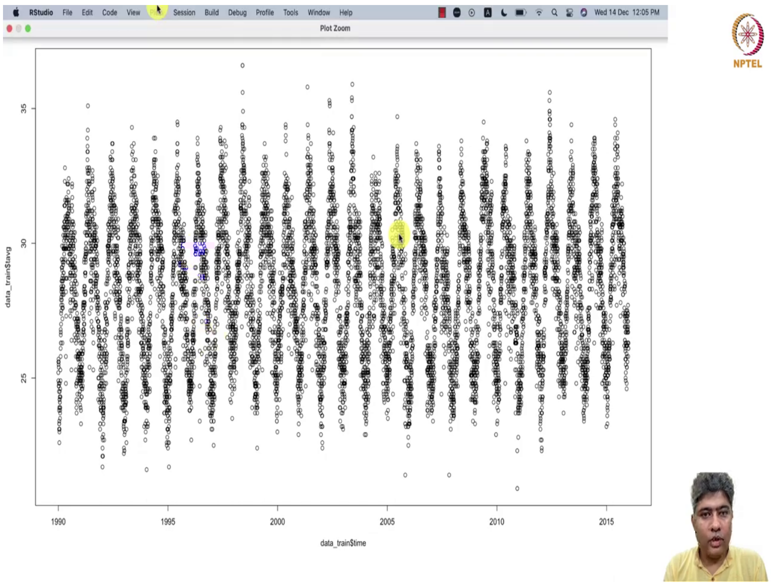

So, here is the splitting the data. We have trained data, we have defined, and here is the test data. Now, what we will do? We will do some visualization. Let us go there and in the main R script, let us do some visualisation, ok.

(Refer Slide Time: 08:30)



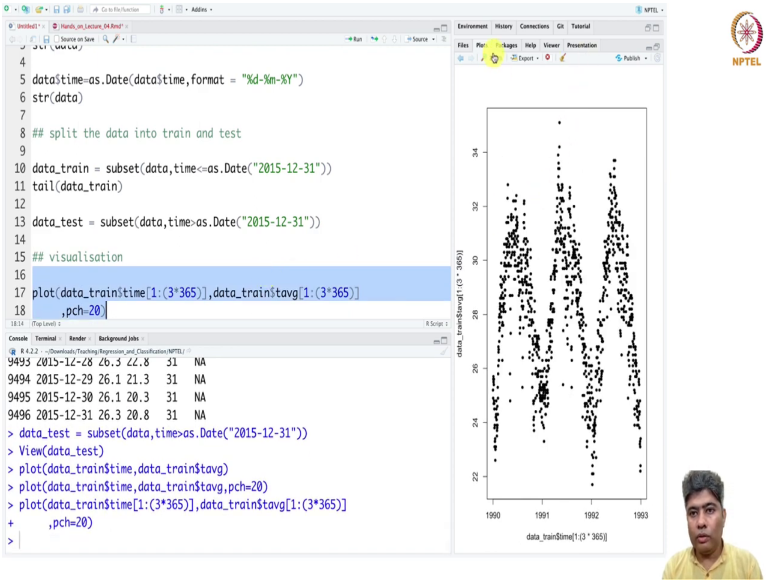

In the visualization, what we will do? We will take the from the train data, we will take the time on the x axis and data train as the average time, we will plot them.

(Refer Slide Time: 09:02)



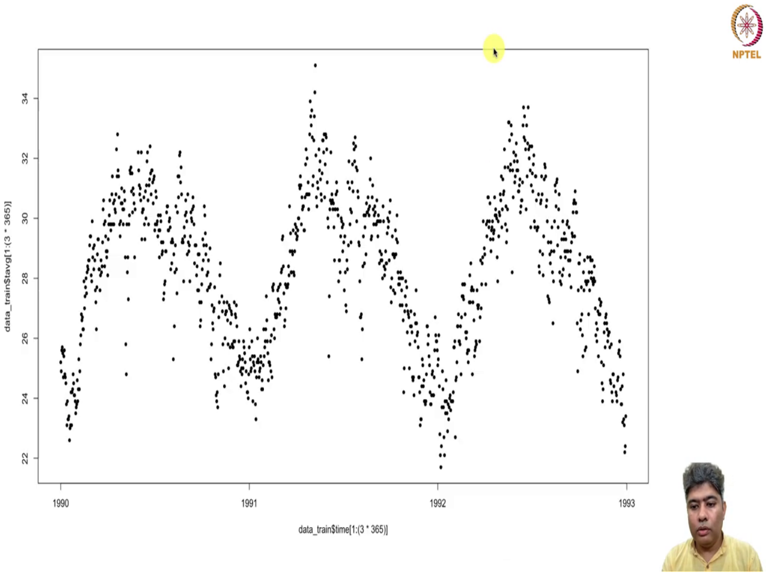
So, you can see. So, this is how it looks like somewhat sinusoidal. So, what we will do? We will put a pch equal to 20, alright, yeah. It looks like somewhat sinusoidal. Every annual behavior is sort of there.

(Refer Slide Time: 09:25)



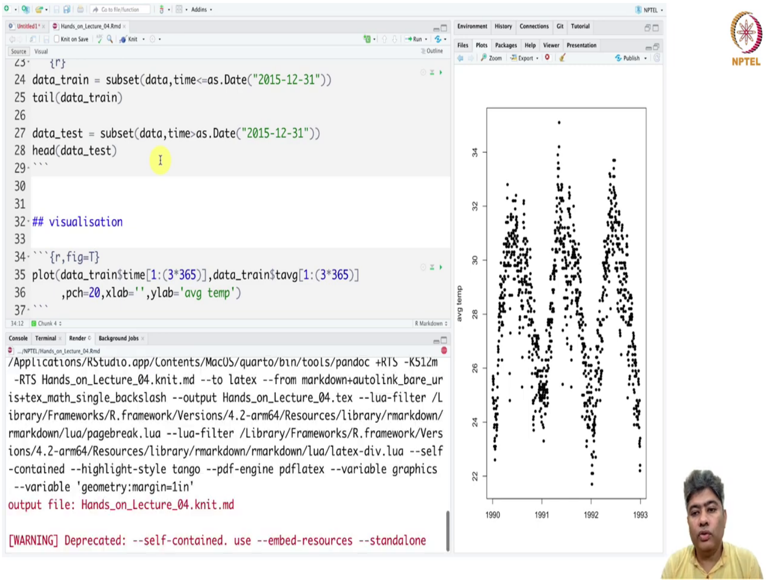

So, in order to just make sure, if it is truly that, we can (Refer Time: 09:28) is 2 into maybe 3 years of data we can plot, 3 into 365, yeah.

(Refer Slide Time: 09:44)




So, this is from 1990-91 is one cycle, 91 to 92 is another cycle, and 92 to 93 is another, 3 cycles, 3 years of data, and it is a sinusoidal behavior as expected.

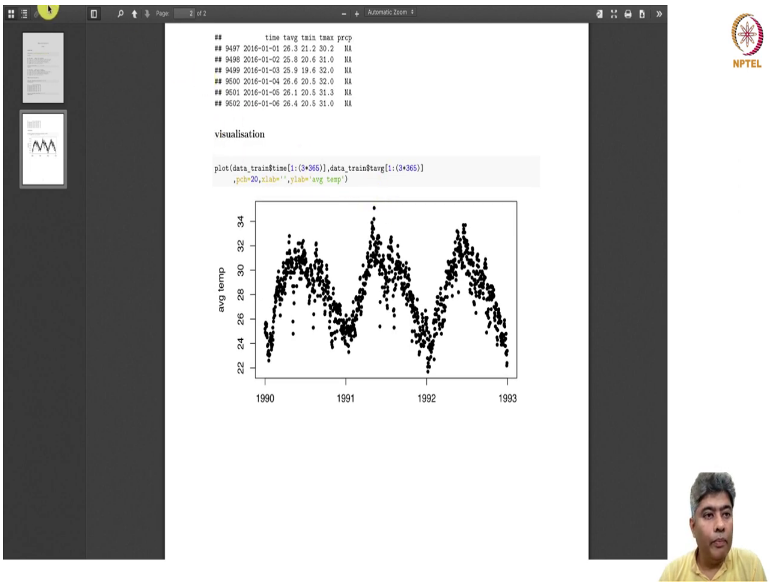
(Refer Slide Time: 10:14)




Now, what we have is if you look into the plot the y lab has is bit, scriptic and x lab is also bit scriptic. What we can do? We can in xlab, we do not need to give anything some sort of self-explanatory, in the ylab in the y label, the label of y needs to be given is a average temperature, ok. So, this is average temperature.

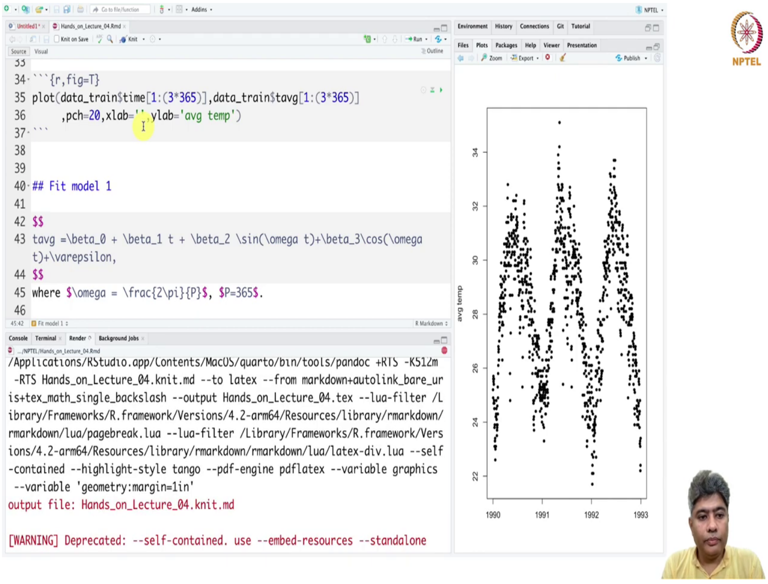
So, we will try to fit some kind of sinusoidal model in the for the average temperature. So, we can put it in R and in R markdown, we have to just say R. And here you have to say figure, fig equal true. So, that it will produce the figure as you wish.

(Refer Slide Time: 11:09)




So, now you can see it produces the figure as expected. If you do not give this figure equal to true, then it will not in (Refer Time: 11:23) reproduce the figure, ok.

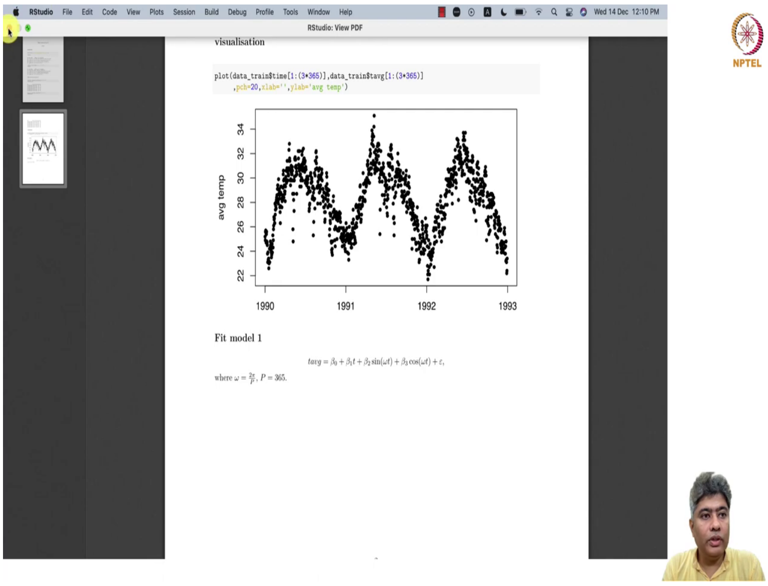
(Refer Slide Time: 11:35)



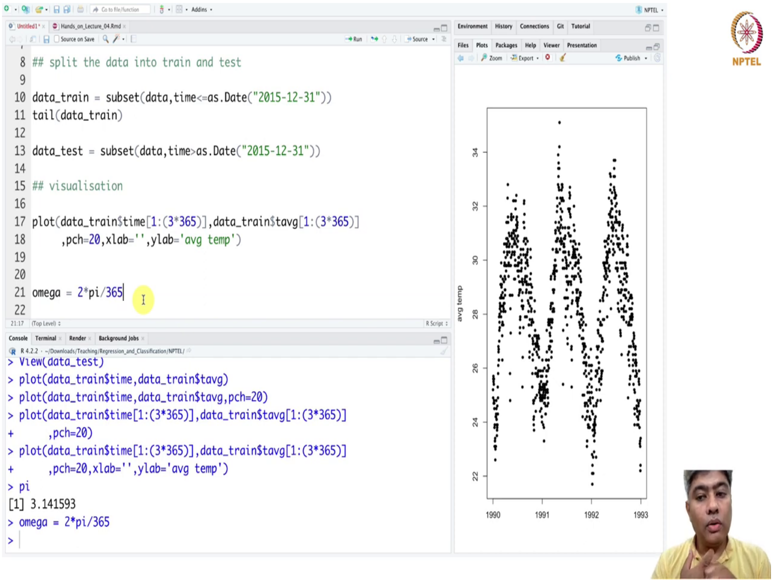

Next, we want to fit a model. So, what model we want to fit? Let us first fit a model, fit model, first model that we want to fit is some kind of sinusoidal. So, we can use in R markdown t average equals to beta naught plus beta 1 equal to prime. Now, that will create if there is any trend is there. And then, a beta 2 times sin maybe omega t plus beta 3 times cos omega t plus some error, ok.

This is the model that we want to fit. But we have to at the same time we have to say something about the omega. So, omega, where omega is equal to 2 fraction of 2 pi of p naught. This p is 365, this p is our daily data is 365, ok. So, this is the model that we want to fit, ok.

(Refer Slide Time: 13:56)

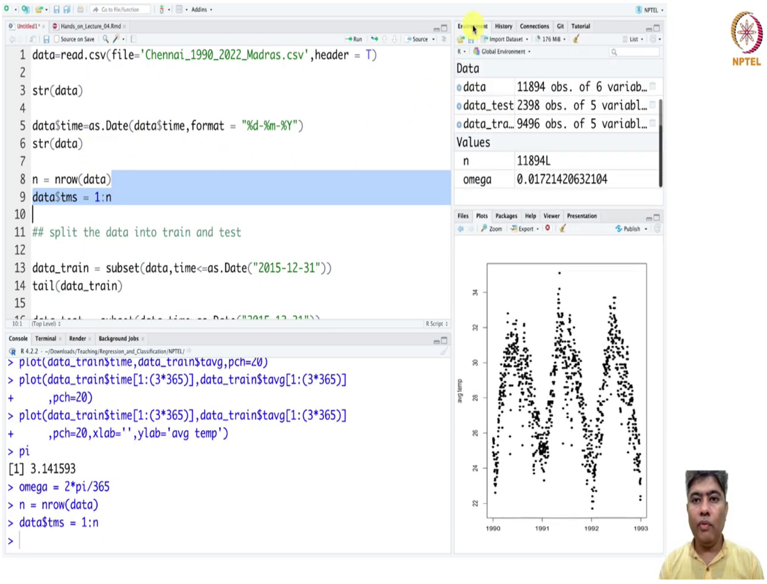


(Refer Slide Time: 14:04)



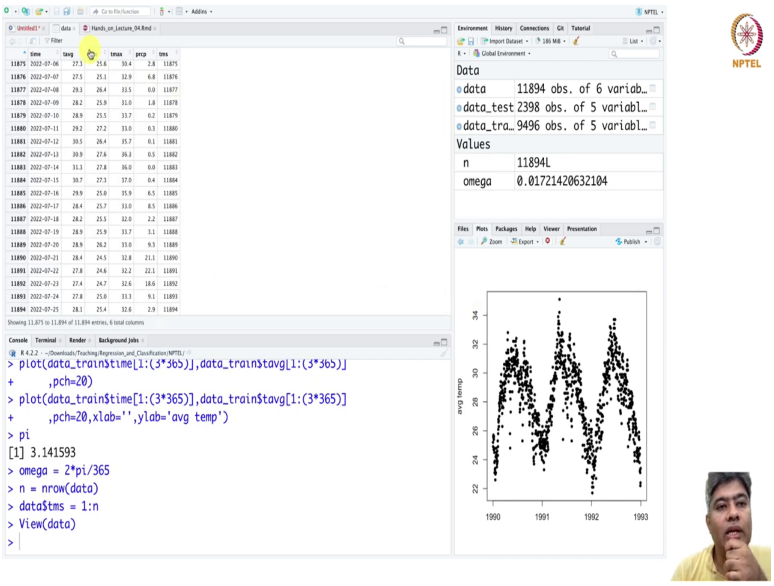

So, this is the model we want to fit. So, first thing we have to do is we have to first thing is omega equal to 2 pi by 365, pi itself is defined as 3.141. So, omega would be like this. Now, we have to have, if I go there I have to have a time to be defined. So, what I will do is here and this time component, we will require both in train and test data set.

(Refer Slide Time: 14:54)



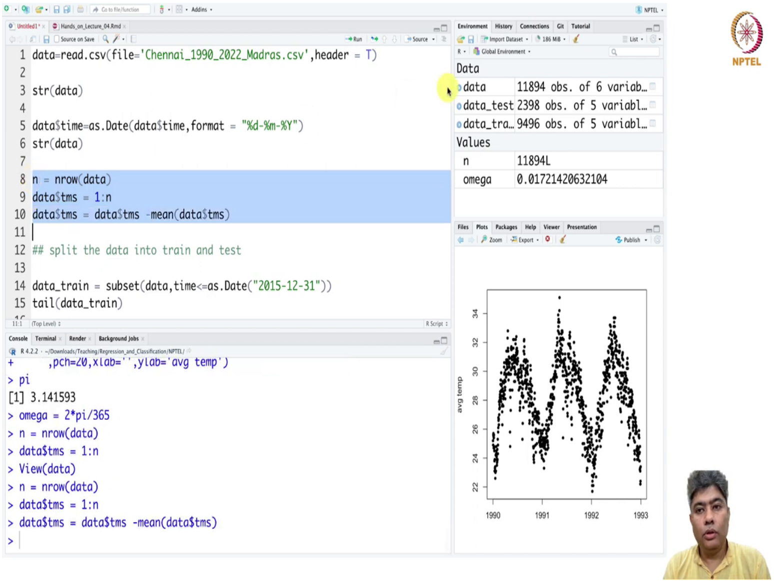
So, what I am going to do is here before the training and test we are going to, I am going to tms, a new variable which is 1 is to n, where n is n row of data. Now, the problem will be if you take a look into the data tms, it is starting from 1 and it is going into 11894.

(Refer Slide Time: 15:31)

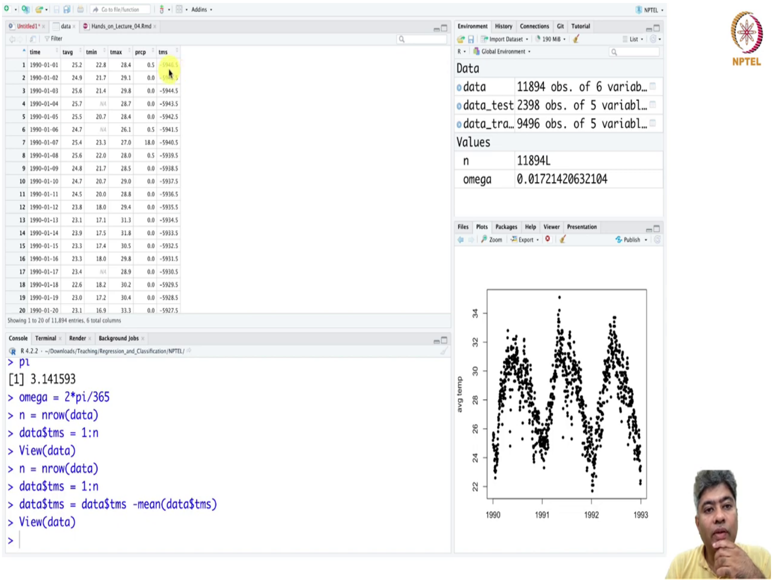



So, it is better, if we just enter the time, it will just handling the data will be better, the coefficient will be little bit meaningful.

(Refer Slide Time: 15:48)

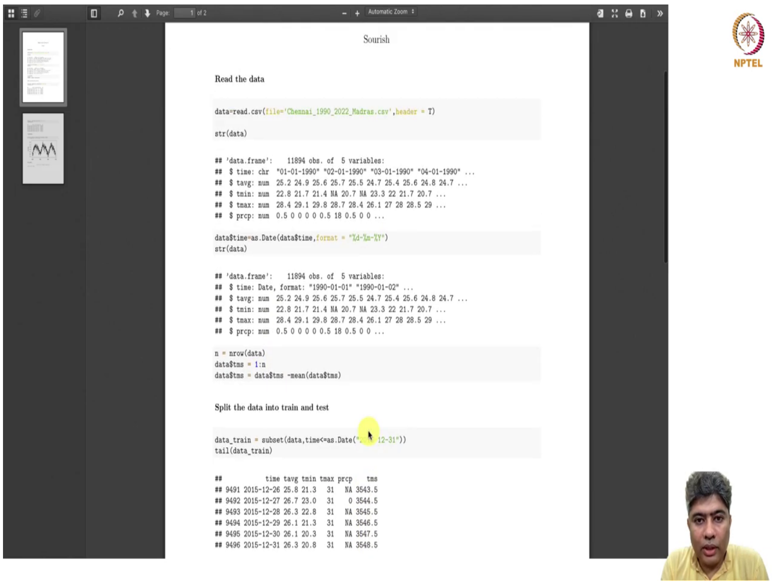


(Refer Slide Time: 16:05)



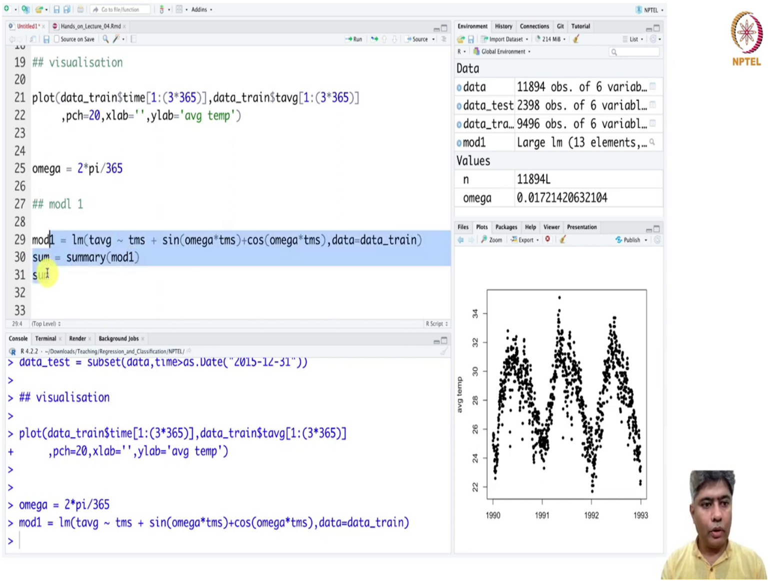

So, what we will do? We just subtract the mean of these guys, ok. So, if I do that, then it is starting from minus 5946.5 and it is going ending at 5946.5. So, it is just shifting the horizon, shifting the time from you know centering at 0. Now, if I just run the whole thing, it will be no problem.

(Refer Slide Time: 16:50)



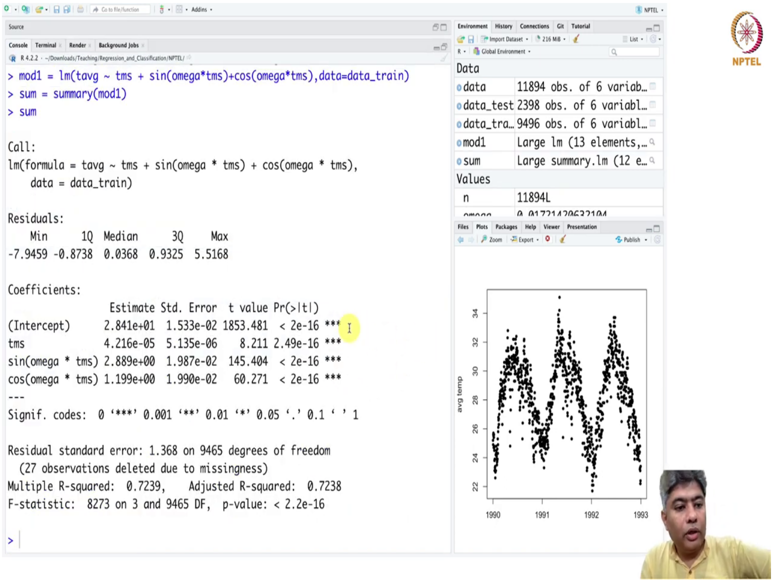

So, we have to add this part in the markdown file, in the beginning. And now if I just run it, you can see the tms has taken care of, alright.

(Refer Slide Time: 17:02)




Next, what we will do? We will fit the model, first model. Model point mod 1 tms plus sin omega (Refer Time: 17:25). Now, we have to provide data equal to data train. And if we just put summary.

(Refer Slide Time: 18:11)

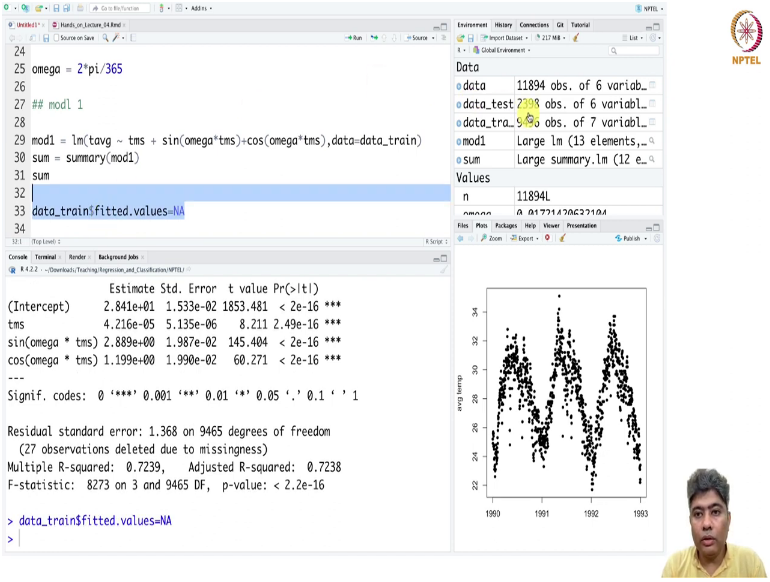



Now, this is the first model to fit. Now, what are the; let us spend understanding what; spend some time to understand what is going on here The first is intercept of the model that is beta naught and then second one is the this is the coefficient corresponding to t. So, if we look into the markdown file, so beta naught plus beta 1 t. So, this beta naught beta 1 times t is, this is the standard error, this is the t value and the p value is really small.

So, that means, with increasing time the average temperature in Chennai is increasing. This is sin omega t which has a significant effect, this is cos omega t, the this is the coefficient of the; so, beta 3, this is beta 2, and this is beta 3 and both of them are significant. Of course, sin cosine will have a significant effect, but the most important thing is trained is significant and it is positive, it is very small, it is very small point after 50421, so it is increasing at a very slow rate.

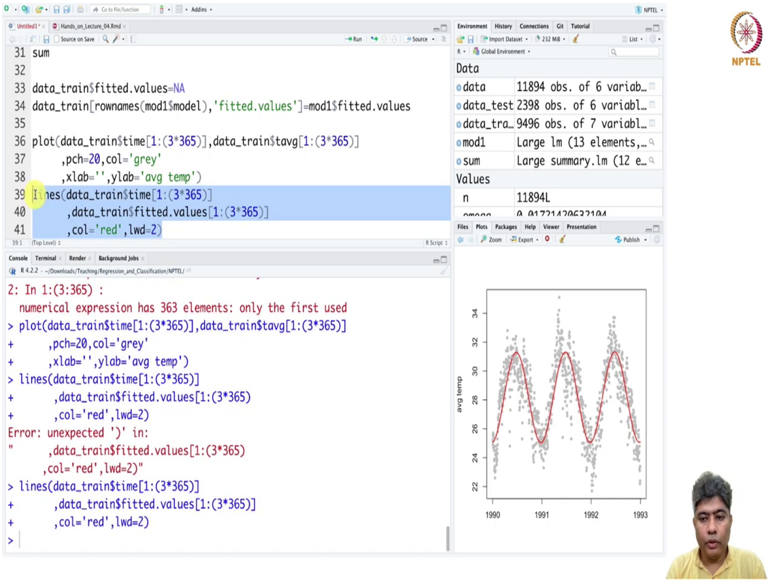
But we cannot ignore it. It is small enough, it is significant enough and it is saying that it is increasing over time.

(Refer Slide Time: 20:06)




Now, what we are going to do is we are going to plot the model, we have to see just how the fitting is working or not. So, before plot, what we will do? We will in the data train, what we will do we will just take the fitted values equal to NA. We are creating a new column with all the NAs.

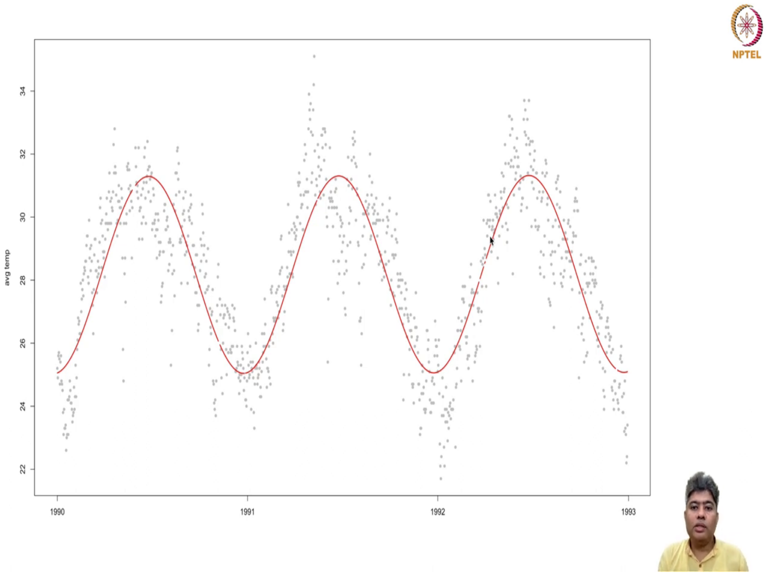
(Refer Slide Time: 20:41)




And then, data train sorry rownames and then model (Refer Time: 20:55) model. So, we have taken the, from the model we have taken the fitted values and we put it in the train data set. We extracted those values and; so, now, we can plot the train dollar, train data dollar, t average pch equal to 20, color equal to grey, x labels we can keep it empty, y lab equal to average temperature.

First, we are taking these, well we can take the first few years to just understand what is going on. First 3 years and then lines, we can just take this comma; instead of t average I want to fitted values, because color equal to say may be red, line width equal to 2. Now, we just, I have not close this here, yeah.

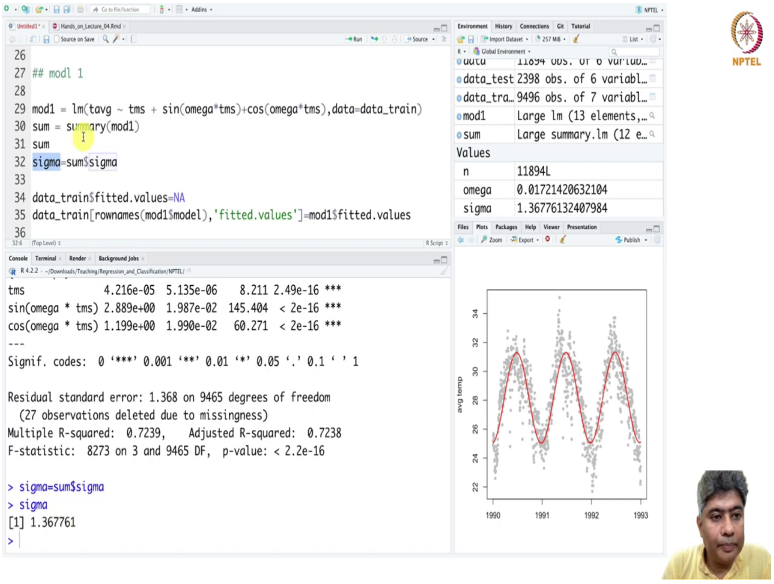
(Refer Slide Time: 23:48)




So, now if I just zoom it, you can see that the fitted model is doing very nicely between going through the sinusoidal way, going through the middle of the points. It is not doing pretty bad, but there are quite a fewer from far away, quite few from bit down. So, this is maybe June, July where lot of rain happens. So, that time, during the rain the temperature drops. So, average temperature drops around that time.

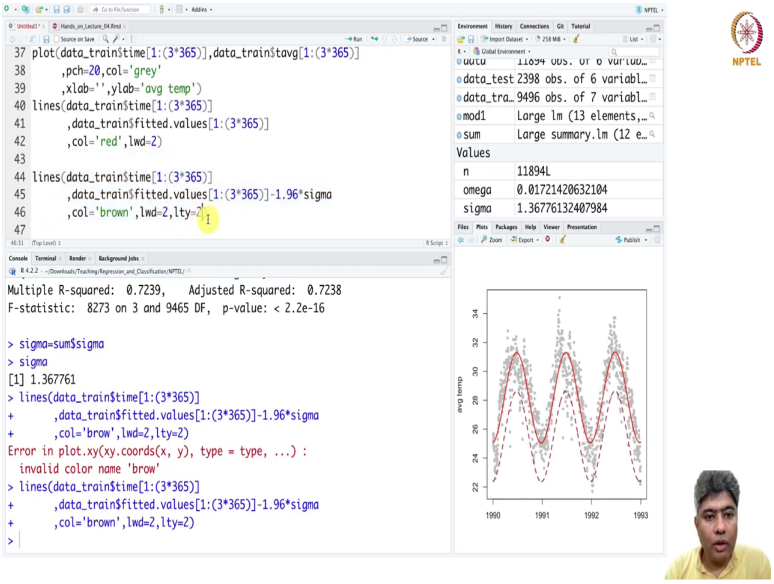
Now, what we can do? We can take the from the summary; if you look into the summary this residual standard error is 1.38 which is actually the sigma.

(Refer Slide Time: 24:37)



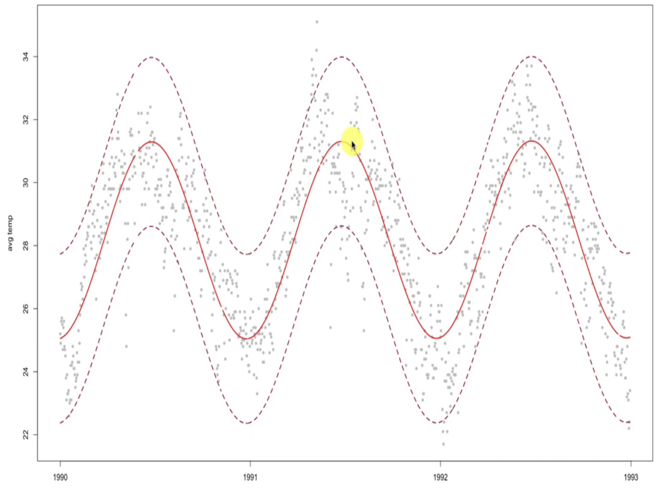

So, we can take the sigma and summary we can calculate the sigma. So, this is our residual standard error 1.367.

(Refer Slide Time: 25:03)



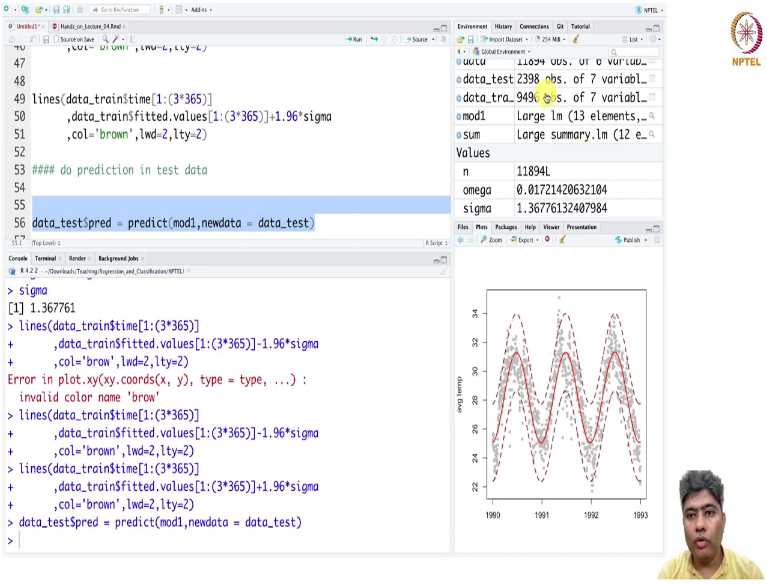

Now, what we can do? We can create a band 0.96 times sigma, if color maybe some different color, maybe brown and we take lty equal to 2. I have to take brown, yeah.

(Refer Slide Time: 25:35)



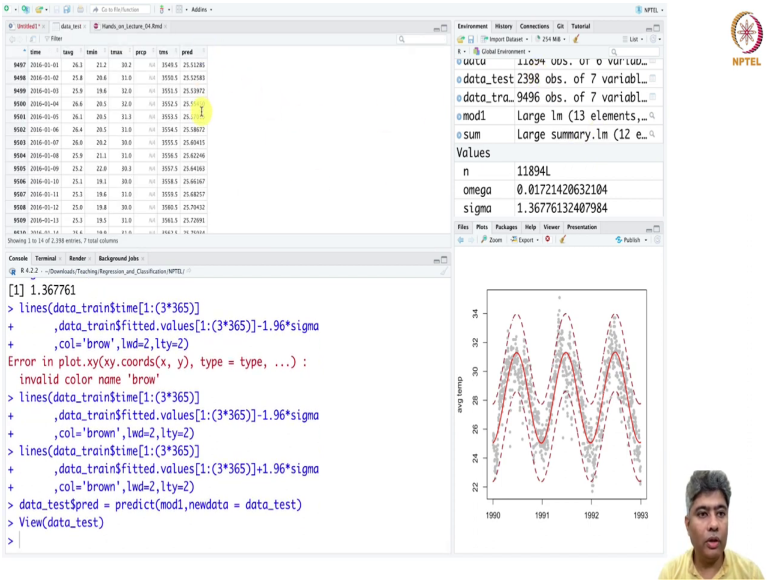

And similarly, we have this is lower bound, this gives us a lower bound you can see and similarly we can create a upper bound as well. So, now, you can see there is an upper bound, ok. And interestingly, you know looks like within the bound, all data points are nicely fitting in. So, it is a pretty decent model, it is a pretty decent model. Now, what we will do? We can just plot this. We can now fit, do the prediction in the, this fitting is inside the training data.

(Refer Slide Time: 26:17)

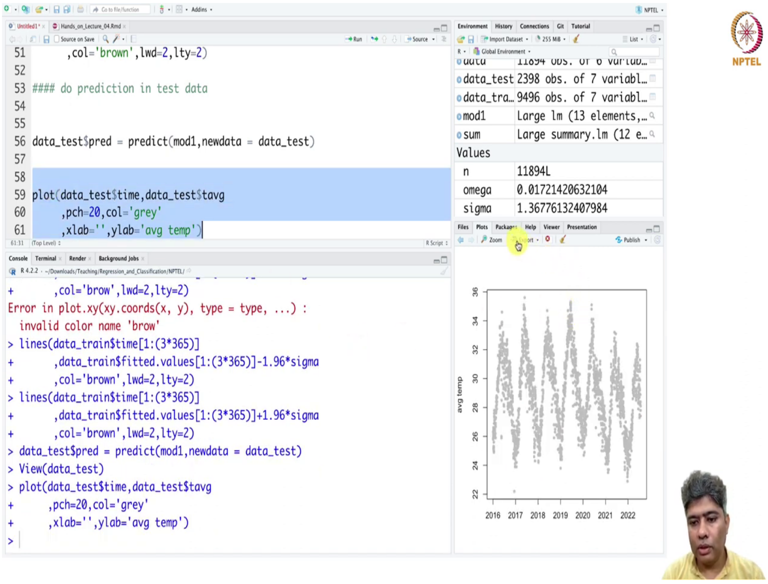



So, what we need to do is we need to look into the how it is doing in the test data. So, do prediction, do prediction in test data. So, what we will do? In data test dollar red equal to predict mod 1, new data equal to data test.

(Refer Slide Time: 27:05)

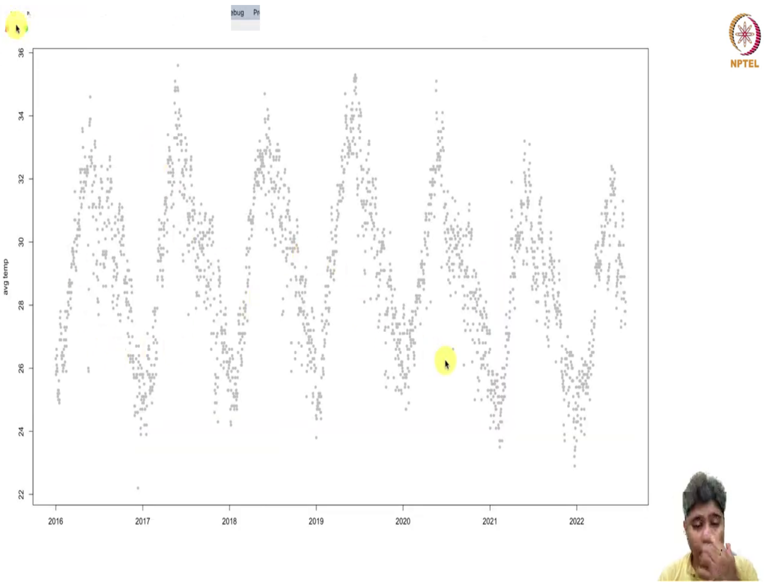


(Refer Slide Time: 27:19)



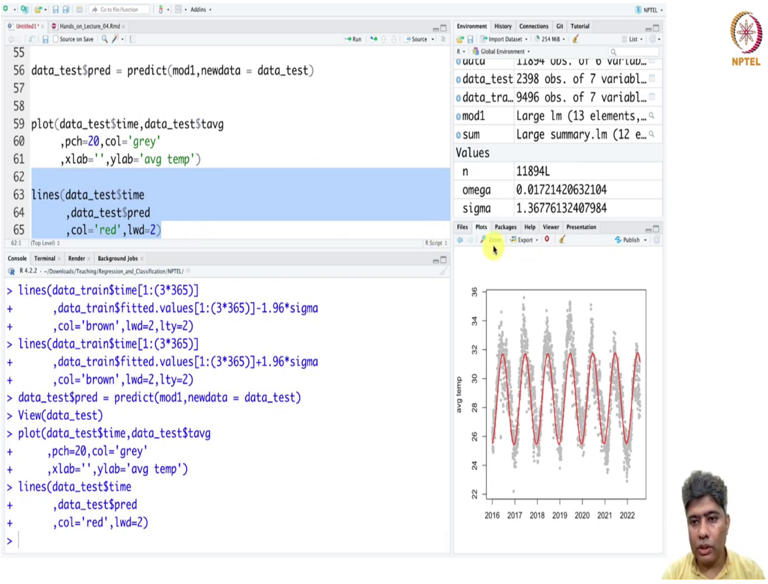

Now, if you go to the test data set, we got all the predicted values. And now similarly, we can do the plotting of first the test data set. And we do not need, we have only 5 years of data test data set, we do not need to keep it. And so, this is our test data set. Let us zoom it.

(Refer Slide Time: 27:40)

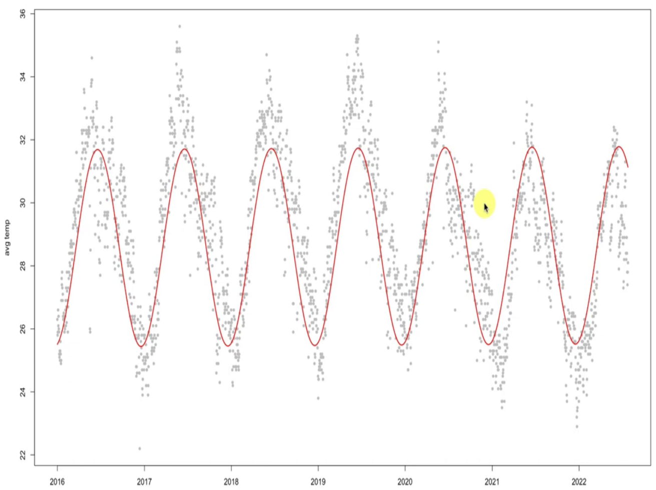



This is our original test data set. And now we can draw the lines here. So, this will be just, we have to follow it here in this way.

(Refer Slide Time: 28:01)

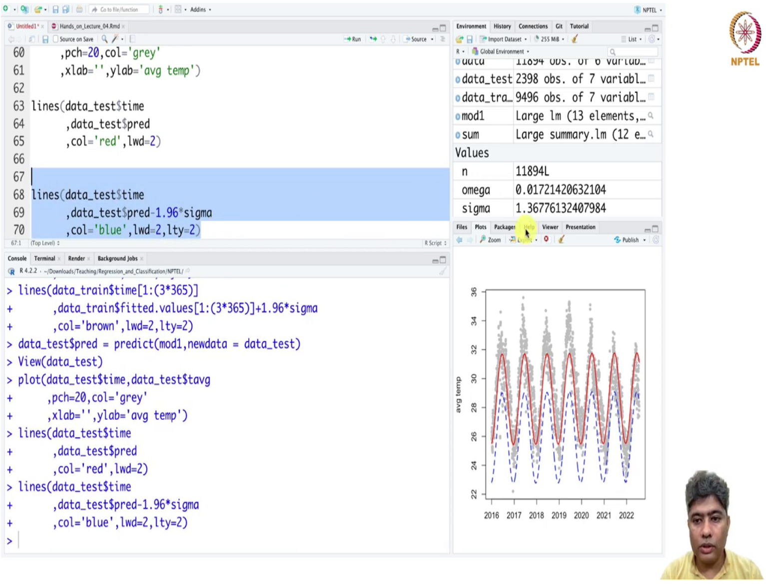


(Refer Slide Time: 28:16)

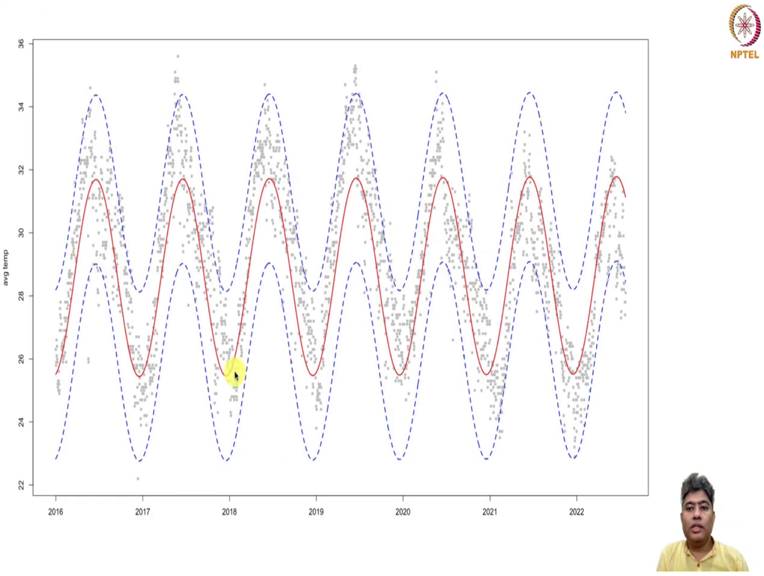



And instead of fitted values we have is pred, we do not need to keep it. So, looks like it has done ok, it has done ok, but we should do a prediction band. And let us see if it is inside the prediction band or not.

(Refer Slide Time: 28:35)



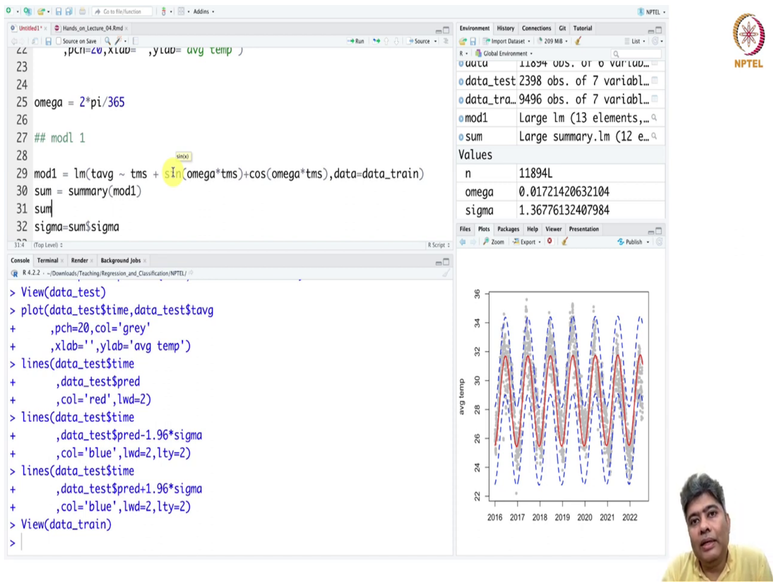
(Refer Slide Time: 28:54)




So, we have to just take this minus 1.96 times sigma. We would like to use some different color maybe blue with lty equal to 2. So, this is our lower band, this is our lower band and we have to do a upper band as well. So, all I have to do is here immediate plus and let us draw the train. So, in the test data set, most of our predictions are within the 95 percent predictive interval. So, we can say that this is a simple model with some sin cosine feature, and then it has done pretty good.

Now, little bit interesting point I would like to make here. So, if you see our data set, right our data set is in two-dimension effectively from a train point of view, we have effectively time and the average time, the dates, and the average temperature.

(Refer Slide Time: 30:02)




So, we have only two variable, average temperature and the time. From there we have created these, these two variables sin and cosine are our engineered feature, ok. These are our engineered feature, sin omega t and cos omega t are engineered feature. So, with the engineered feature we fit a nice model and which is doing pretty reasonable in terms of predicting the temperature.

So, we will stop here. And we will continue in on it, you can always ask me one question that why we are stopping at only two sin cosine Fourier engineered feature, Fourier feature. We can always go for higher order Fourier feature, like sin 2 omega t cos 2 omega t sin, 3 omega t cos 3 omega t.

In the next lectures, we will do those hands on with the higher dimension. But remember that more feature we will put, it will add more complexity in our model, our model will become more high dimensional. So, the this has some advantage and disadvantage. We will talk about those model, increasing model complexity in the next lecture.

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