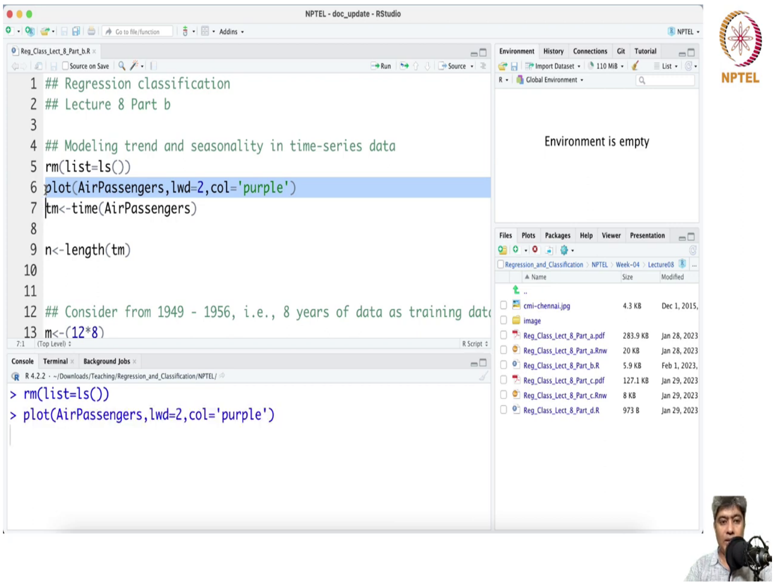
Predictive Analytics - Regression and Classification
Prof. Sourish Das
Department of Mathematics
Chennai Mathematical Institute

Lecture - 27
Hands on with R Part - 6

Hello all, welcome back to the part b of lecture series 8. In the previous video we discussed how to use simple linear regression model technique to model time series data in that case we consider two objectives; one was to a model a forecast a long trend or long term forecast using modeling the trained and seasonality and then short term forecast using autoregressive model.

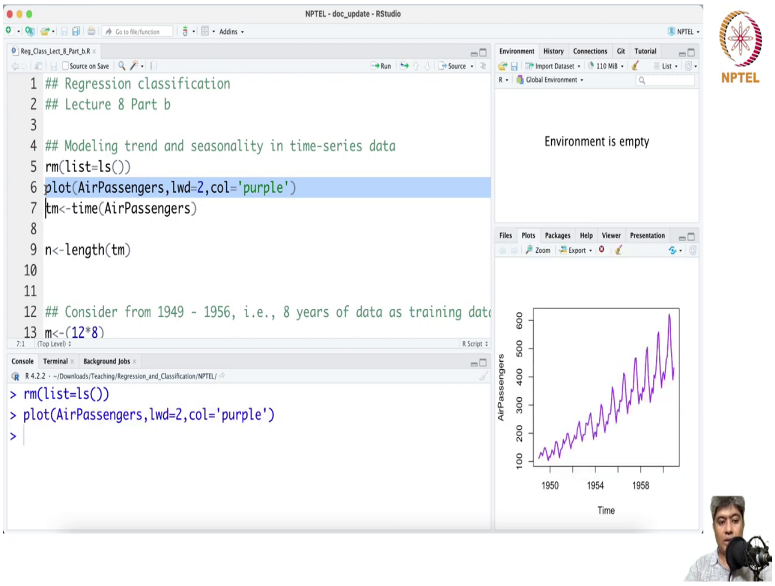
In this video we are going to see how those models were actually fitted with the air passenger data.

(Refer Slide Time: 01:09)

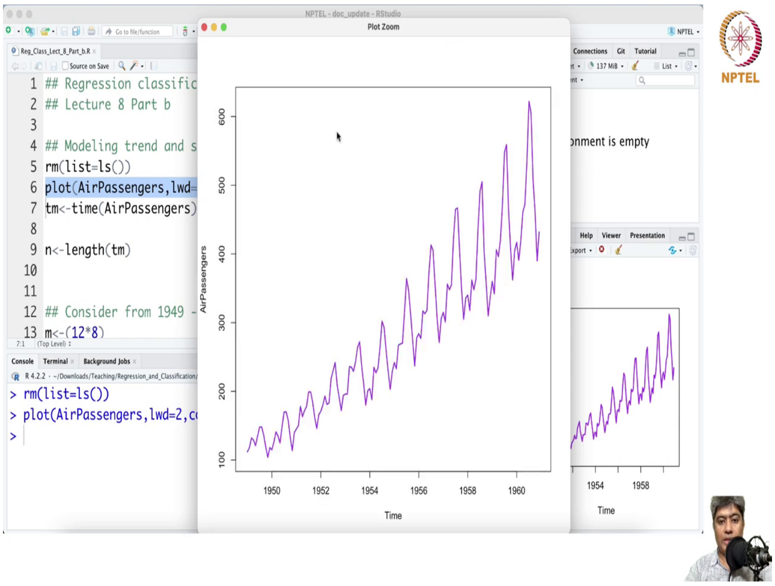


So, let me go and start the r and I will share this code on the SWAYAM portal in the NPTEL portal. So, first line here is it is just remove the environment any variable are there. So, it will just remove the environment clean the environment. If you just run air passengers data it is available in the base data set.

(Refer Slide Time: 01:32)

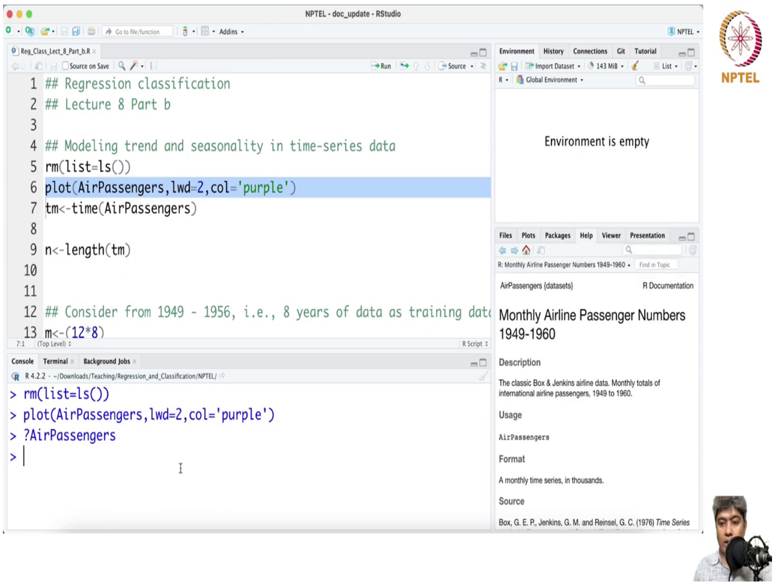


(Refer Slide Time: 01:37)



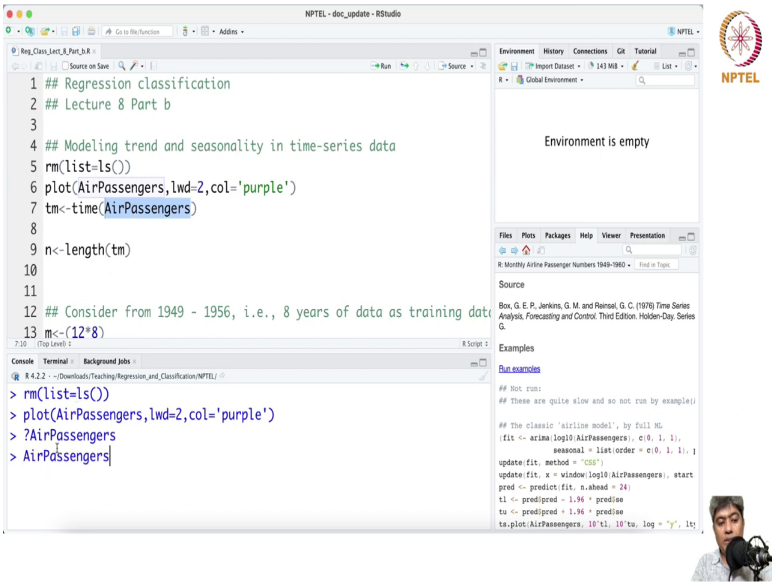
Then it will just plot the air passenger data set the data set is over the time period from 9 early 1950s to 1960s and here is the air passenger values.

(Refer Slide Time: 01:51)



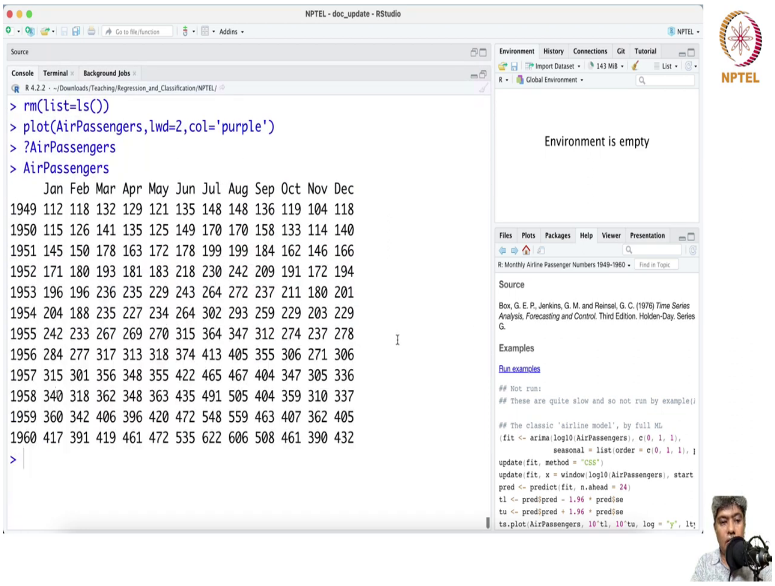
In fact, if you just do question mark air passengers here it is talking about the classic box gains airline data set, monthly total of international air passengers between 1949 and 1960.

(Refer Slide Time: 02:13)



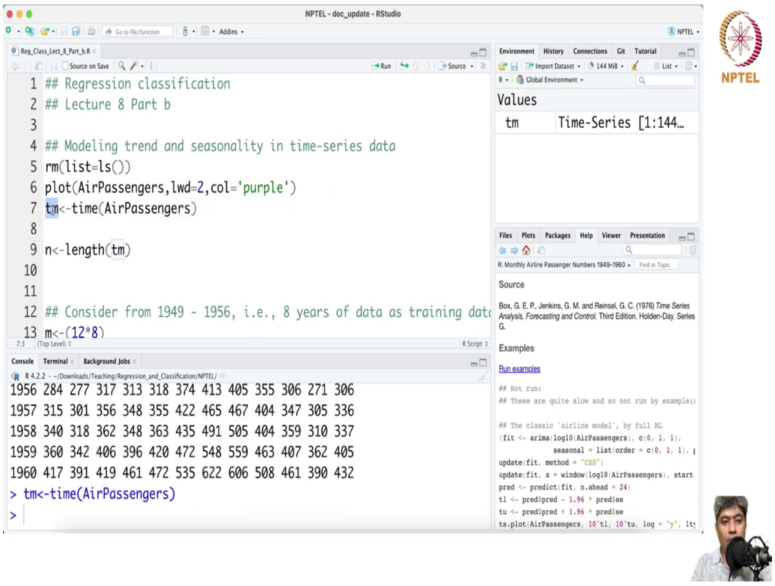
So, here is the data and if you just run this in fact, this data just copy and paste it and just run it.

(Refer Slide Time: 02:21)



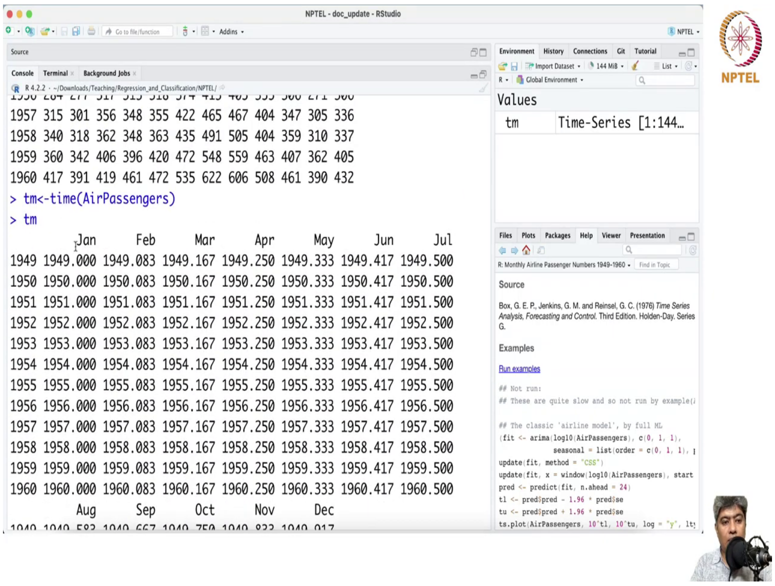
You can see the entire data set its a small data set where you can see the entire data set Jan, February, March, April, May up to December and then if each row sort of or a particular year. So, its a nice data set which you can see and now we are going to model it.

(Refer Slide Time: 02:46)



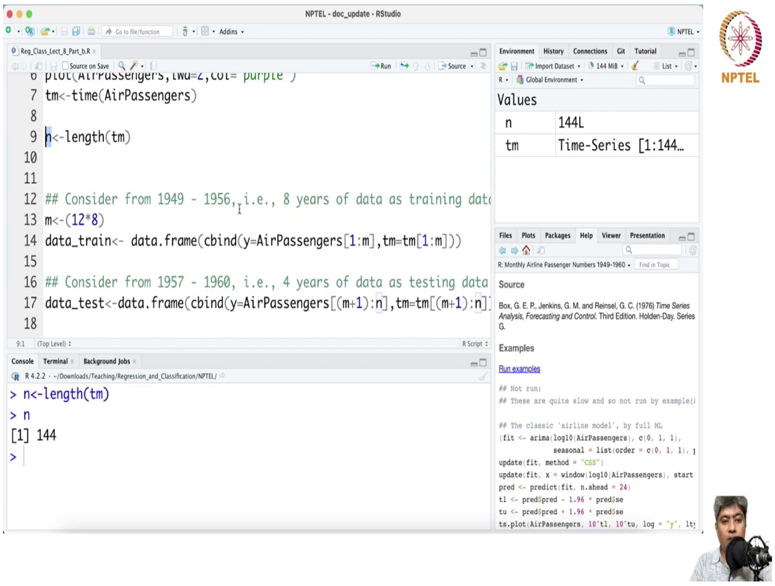
Now from here I am going to extract the time.

(Refer Slide Time: 02:49)



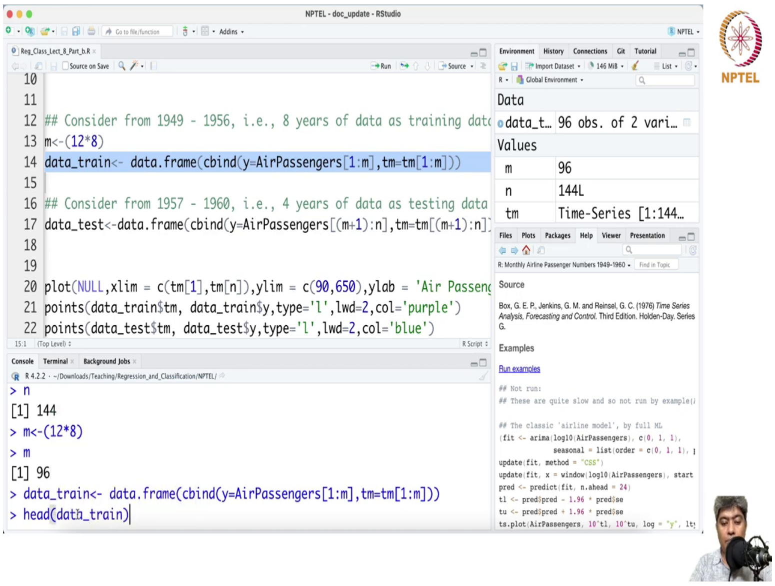
So, tm if I just plot the tm, you can see that Jan, Feb, March every month the time is being converted into a numerical value.

(Refer Slide Time: 03:08)



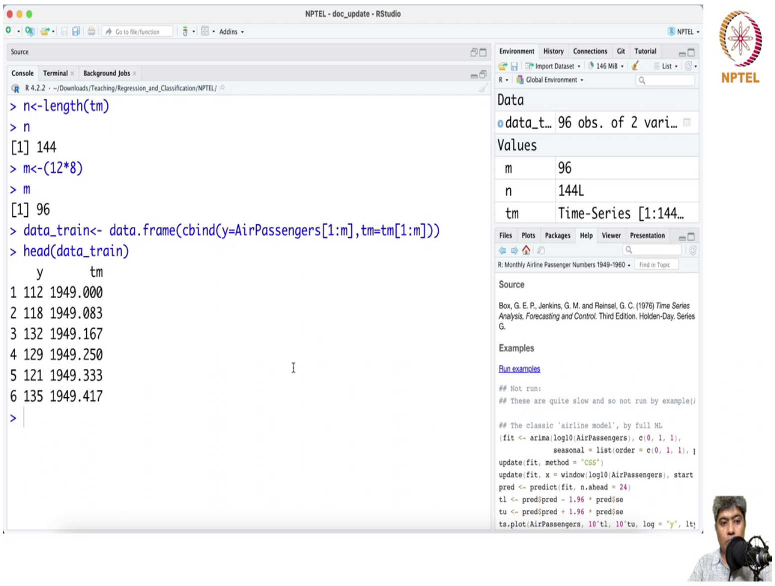
So, we call it tm and n is the length of the tm. So, n is so, about 144 values are there. So, we consider from 1949 to 1956 that is 8 years of data as training data. So, 12 into 8 every month we have 12 months of data over 8 years.

(Refer Slide Time: 03:35)



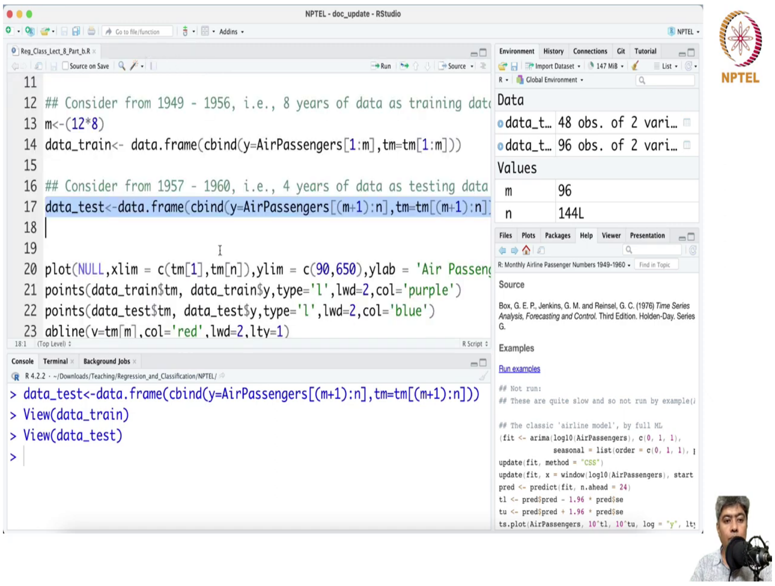
So, we use we are going to use 96 data points out of the; out of the 144 data points as training data. So, I am defining my training data as data frame where y equal to AirPassengers from 1 is to m and tm stands for time. So, if I just now write head of data sorry header data_train.

(Refer Slide Time: 04:09)



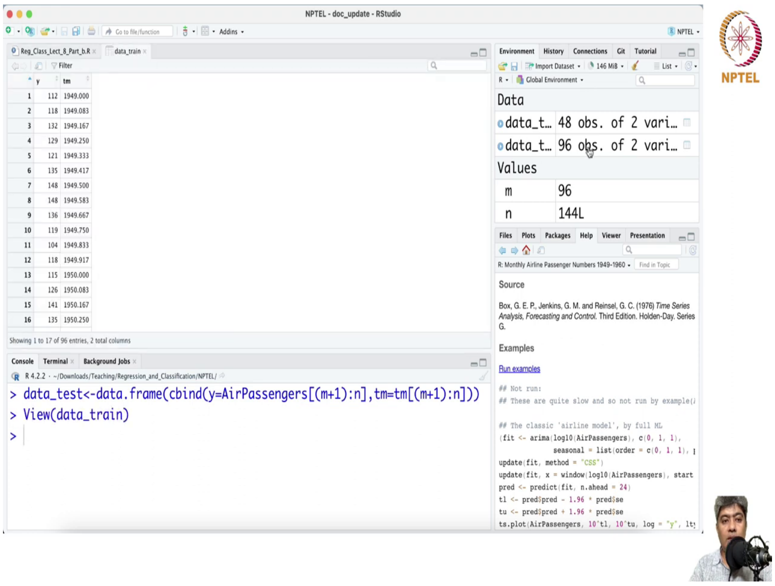
So, you can see that this is the data on the first few days of data, few months of data ok.

(Refer Slide Time: 04:20)



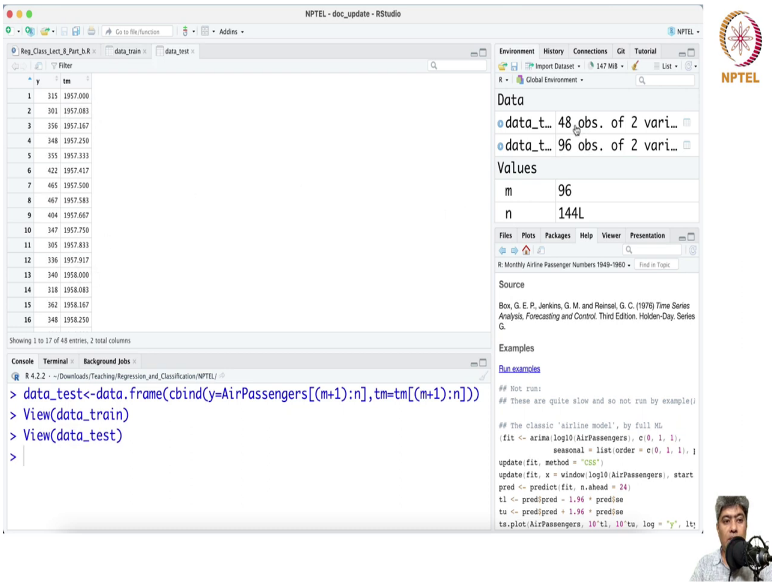
Now, we are going to consider from 1957 to 1960 that is 57,58, 59 and 60 4 years of data as testing data. So, about one two third data we are going to use for training and one-third data we are going to use it for testing ok. So, here is the we are running it as a testing.

(Refer Slide Time: 04:44)

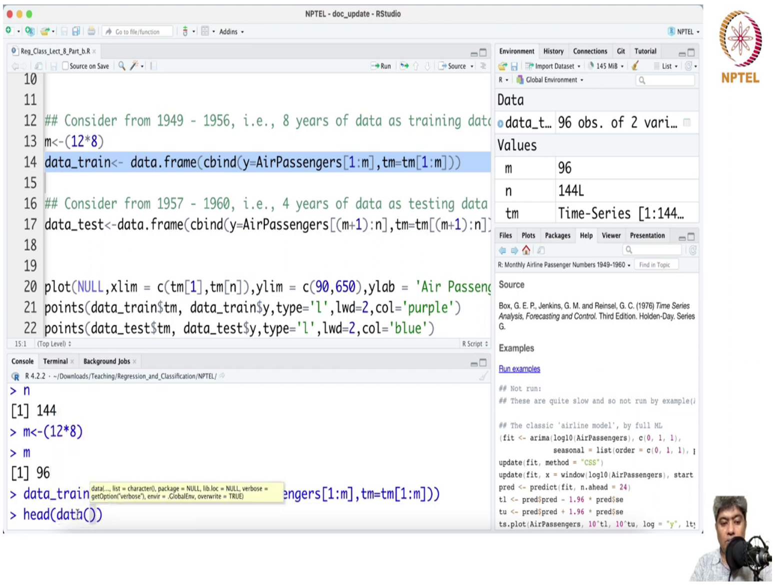


So, here is the test data set it is a trained data set and this is the test data set ok.

(Refer Slide Time: 04:51)



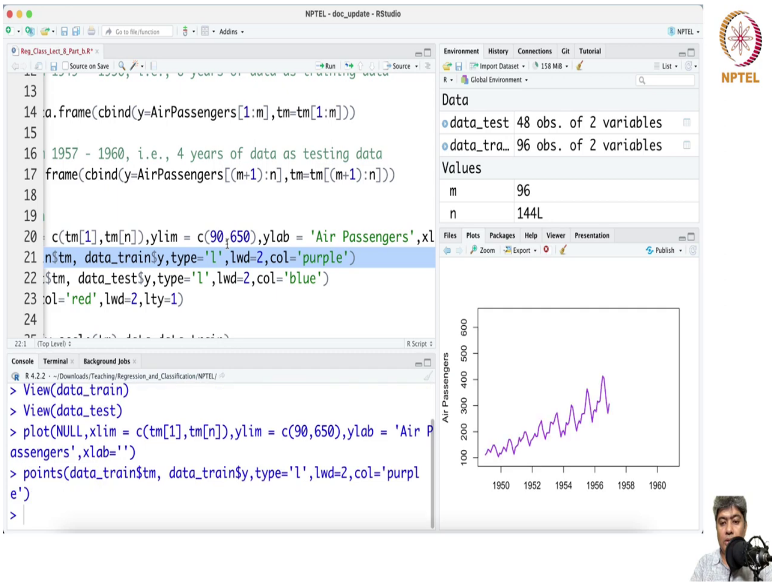
(Refer Slide Time: 04:56)



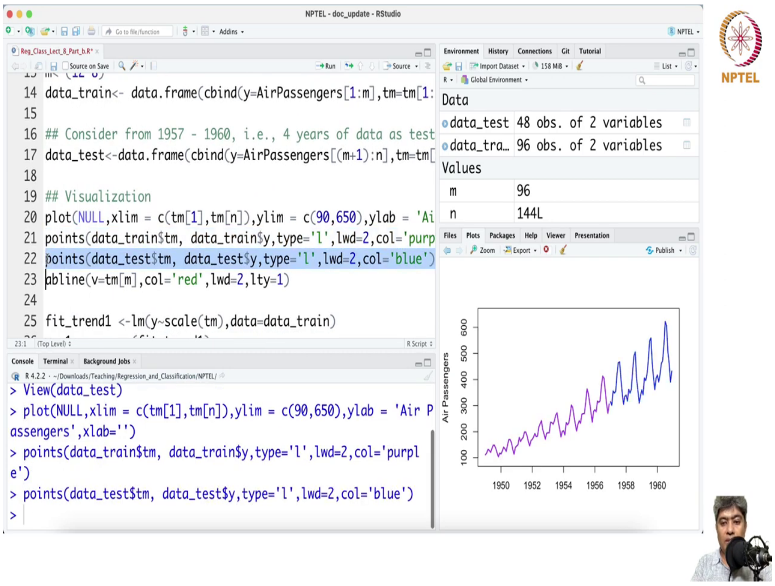
Now, this piece of line is going to create the just that frame of the plot I am going to create some visualization let us as our discussion you know that visualization is sometimes helps to understand the it helps us to understand the data and what kind of model can be used.

So, here I am just creating a framework just on the exact same thing do not plot anything just x axis will have the times and y axis will have the dates. Then from the time train you plot the data with purple color.

(Refer Slide Time: 05:46)

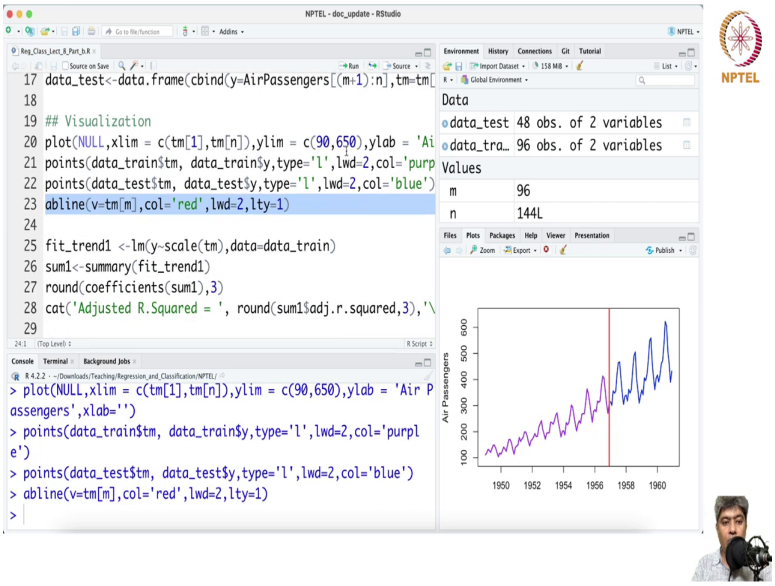


(Refer Slide Time: 05:51)



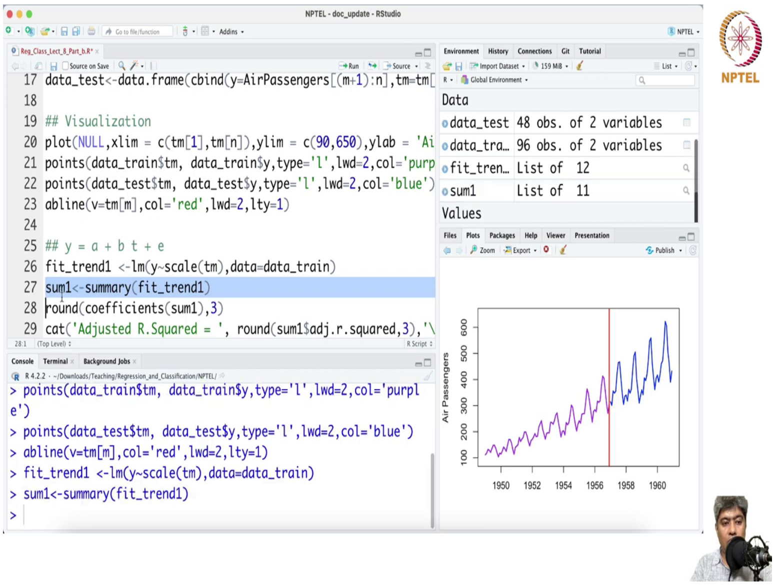
And then from the test data set you plot the data with the blue color.

(Refer Slide Time: 05:56)



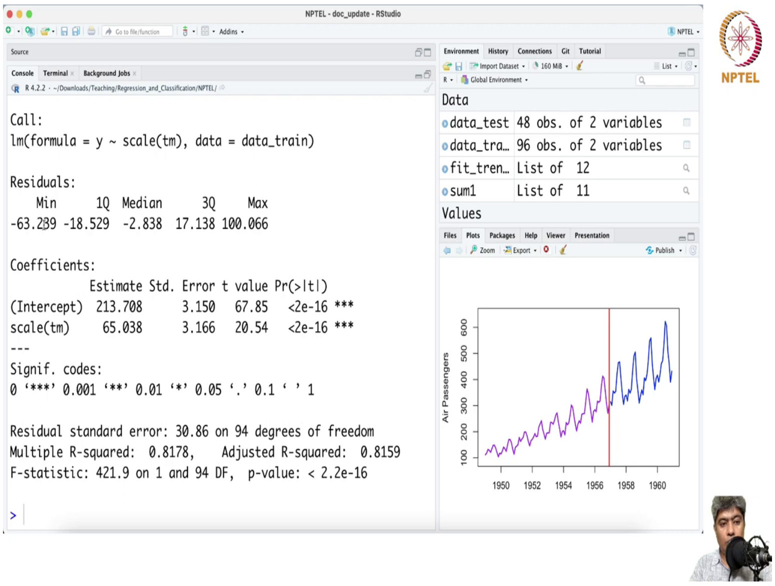
And this is the red line that we draw. So, this part of the data we are going to use as a train data and this part of the data we are going to use as a test data.

(Refer Slide Time: 06:14)



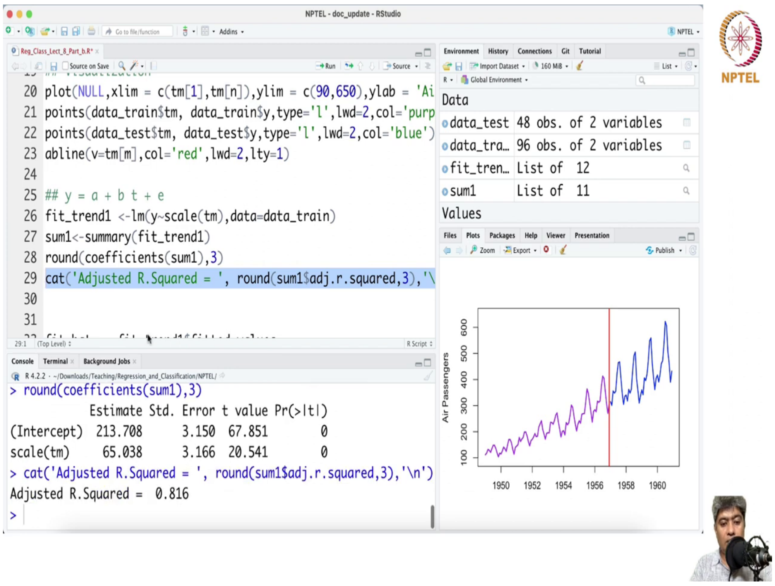
So, the first model we are going to fit is y equal to a plus b time plus error ok, now we just fit the data summary.

(Refer Slide Time: 06:31)



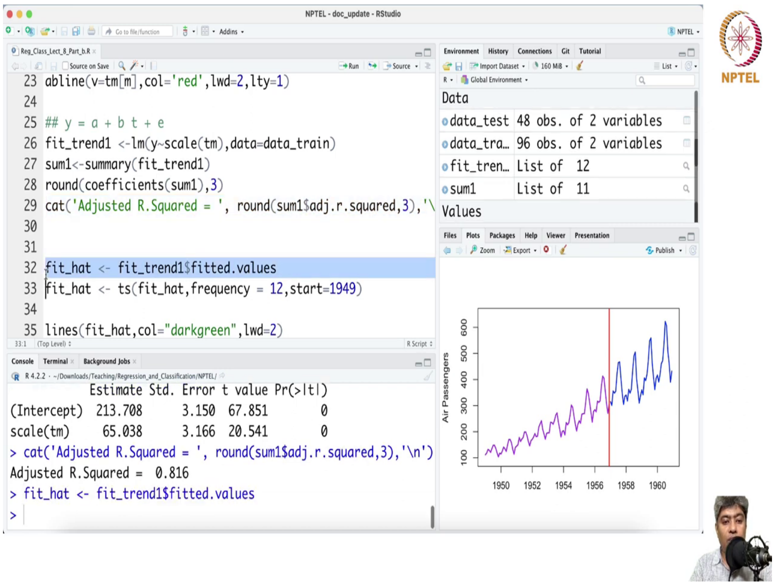
And this is the summary. So, this is what we have seen it in our presentation also.

(Refer Slide Time: 06:41)



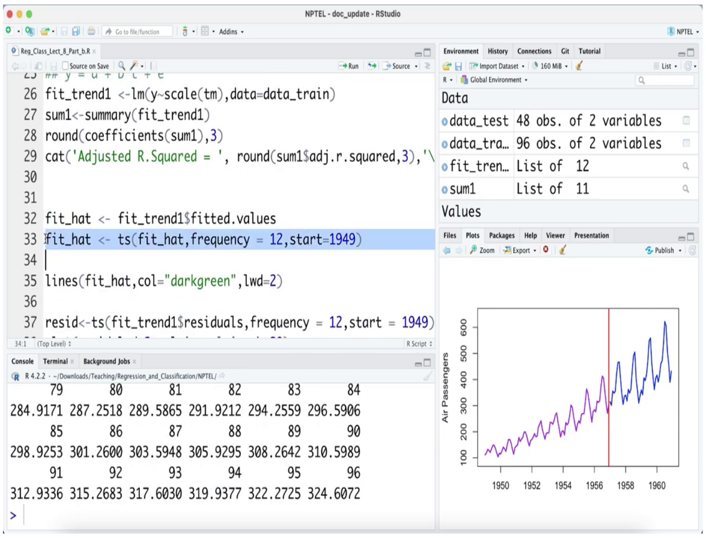
If we just so, this is what we have seen and adjusted r square is 81.6 percent so; that means, 81.6 percent of the variability in the air passenger data can be simply explained by the simple linear trend.

(Refer Slide Time: 07:05)

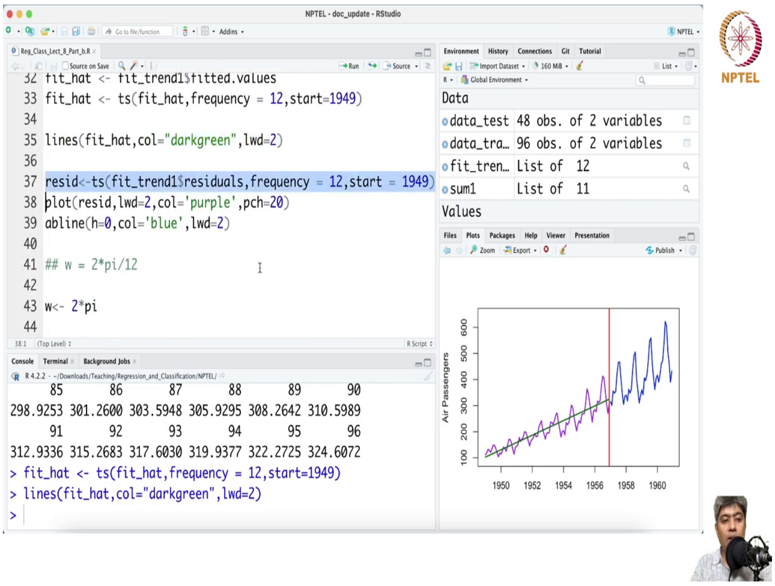


Now from that I can compute the fitted values.

(Refer Slide Time: 07:06)

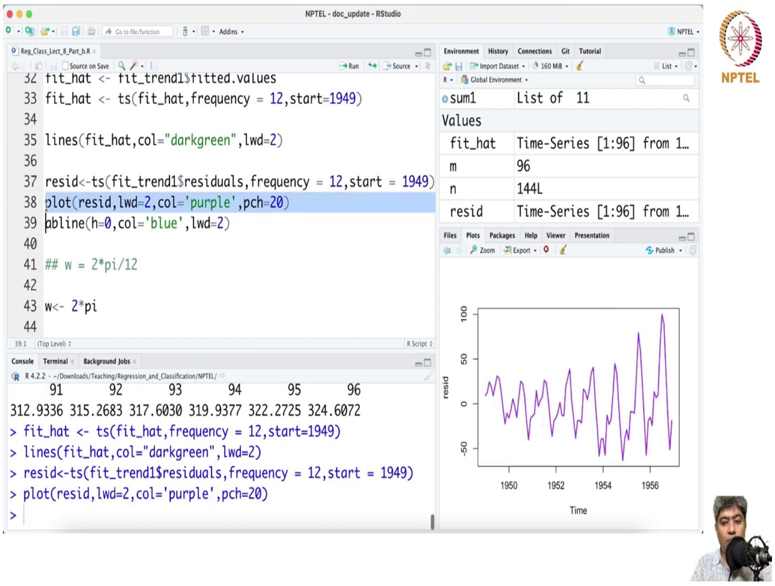


(Refer Slide Time: 07:09)



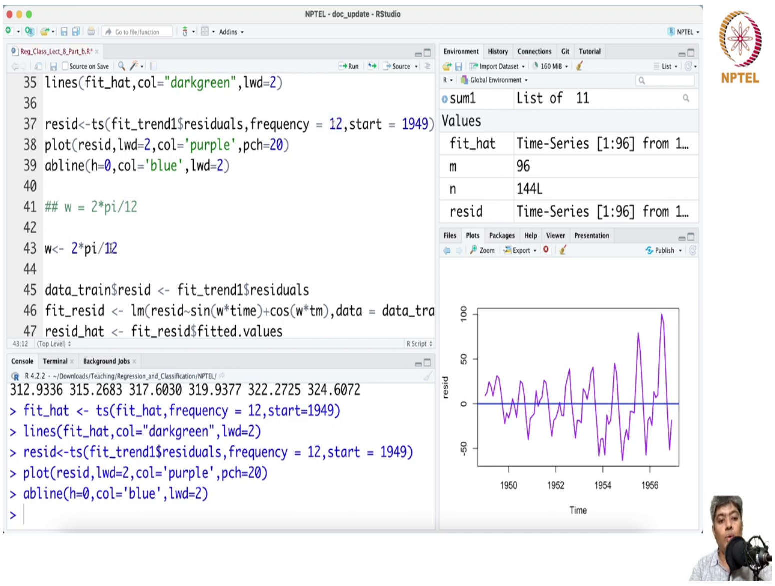
And I just define it as a time series and plot the line. So, this is the fitted values and from there we from the fit trend.

(Refer Slide Time: 07:24)



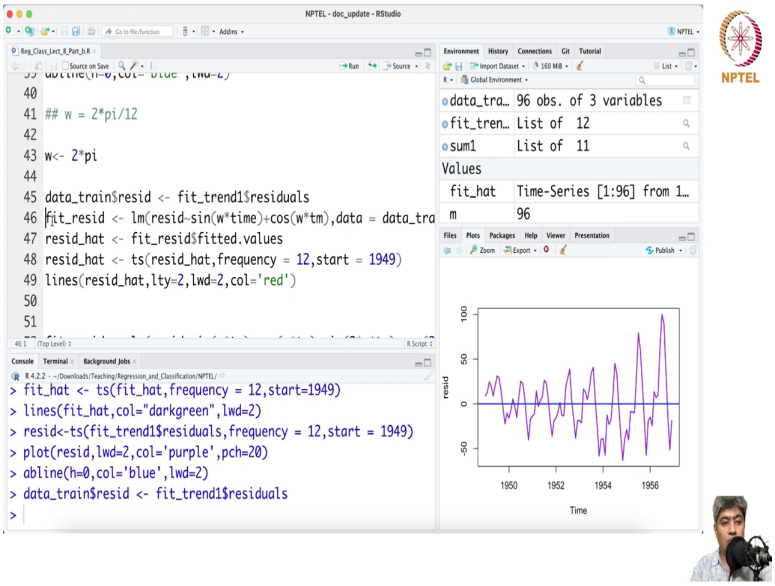
We just extract the residuals as frequencies as time series data and plot them and this is these are the residuals you can see this is the residuals.

(Refer Slide Time: 07:37)



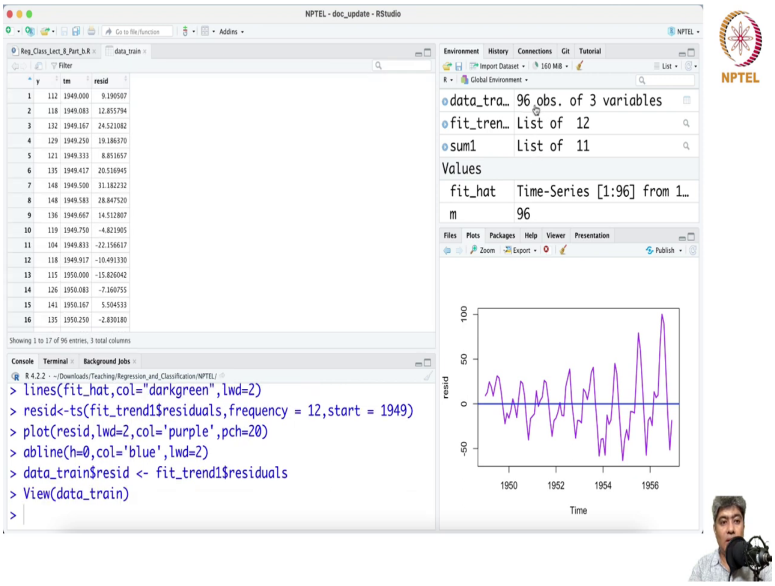
And at we plot a 0 at a at 0 the sum will be negative sum will be positive and we can see that residuals have the seasonality. Now we now we plot 2 pi omega by 2 pi now we do not need to define it as a 12 because in the we have defined it as a time series data and here frequency is already defined as 12. So, that is why I do not need to divide it by 12.

(Refer Slide Time: 08:12)

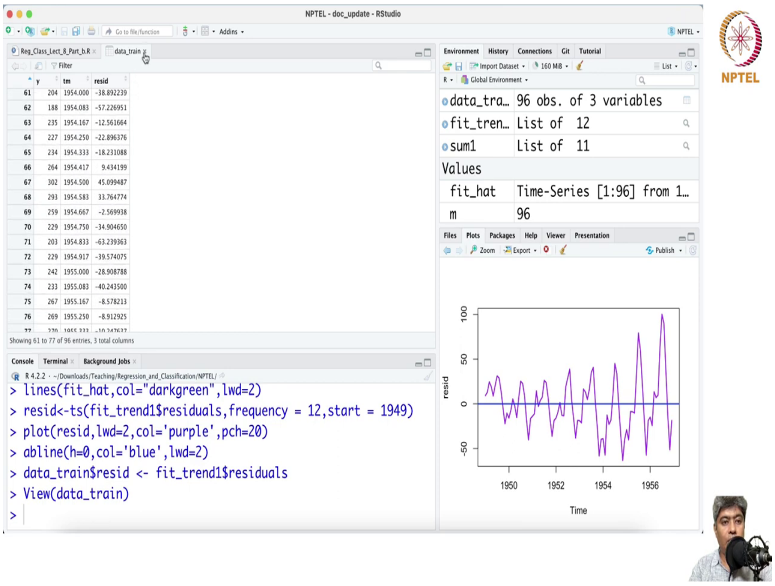


So, because data it is already being defined as a frequency of 12.

(Refer Slide Time: 08:25)

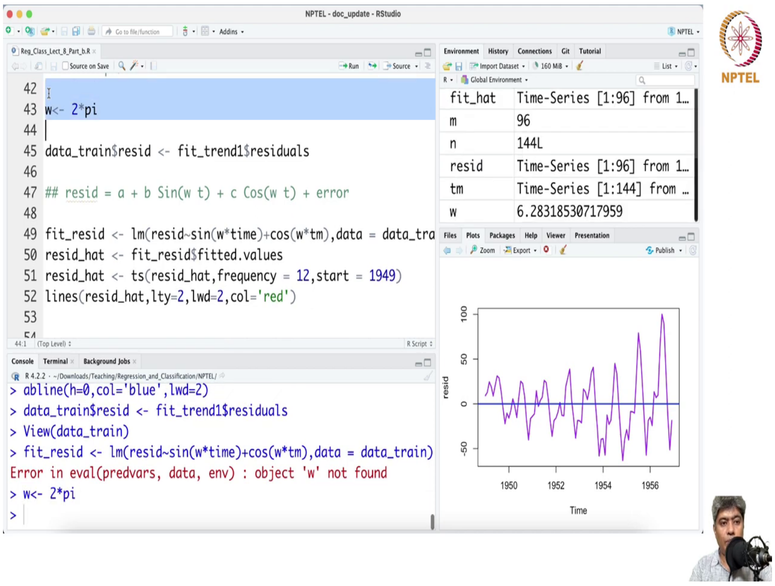


(Refer Slide Time: 08:31)



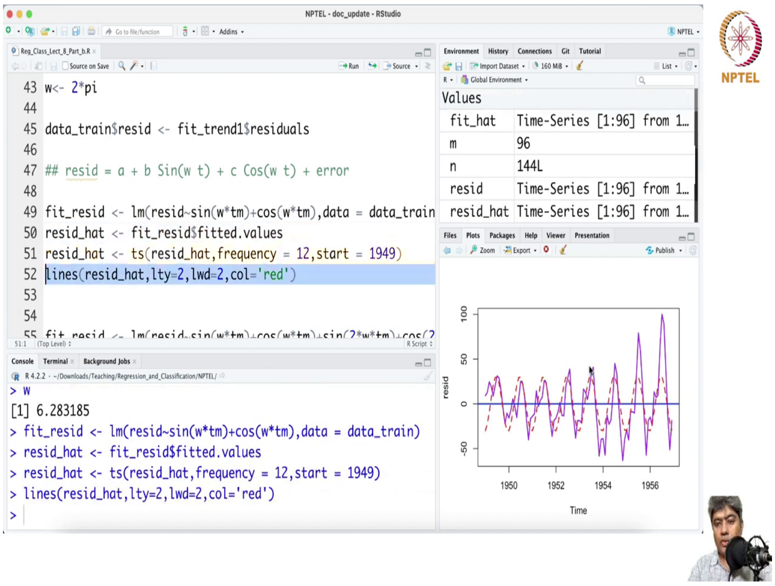
So, data set is residuals and now in the data set train you see I have extracted the residuals ok extracted the residuals and fit the residuals as a function of sin omega time plus cos omega time ok.

(Refer Slide Time: 08:45)

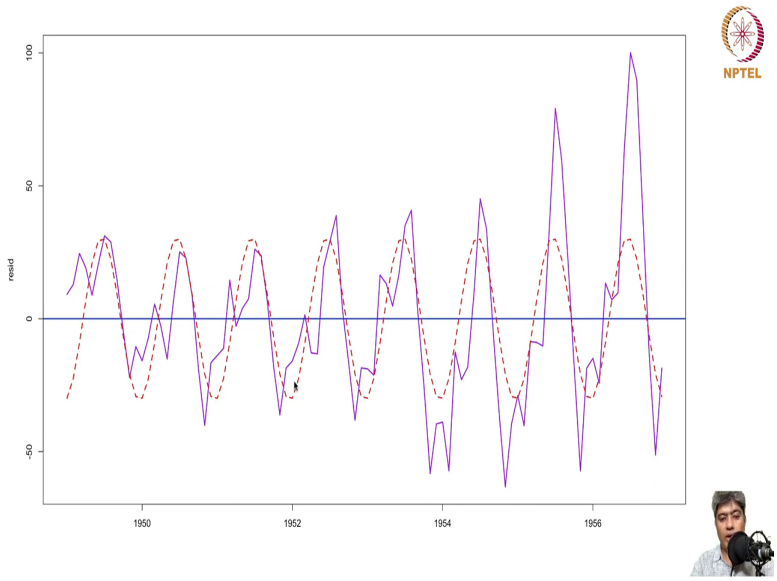


Now, this is the model what we want to fit residual as a function of a plus b times Sin omega t plus c times cos omega t plus error this is the model we want to fit. So, we fit this model sorry we need to run this guy yeah.

(Refer Slide Time: 09:19)

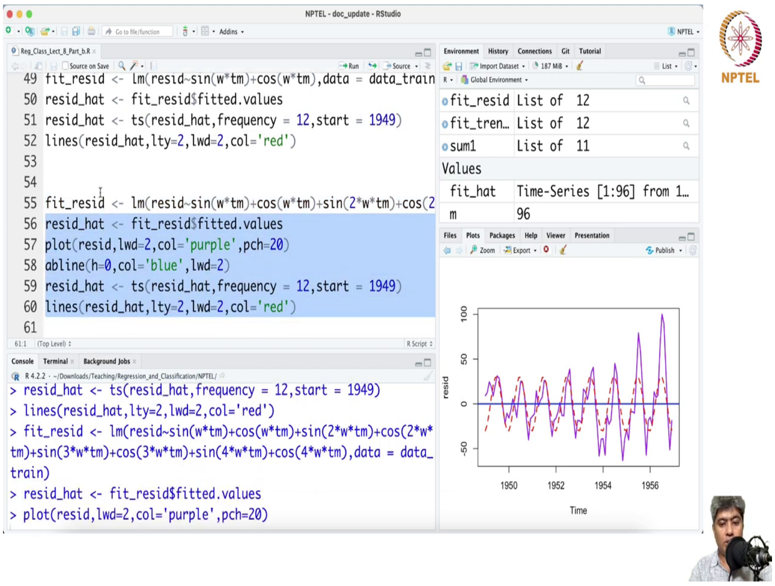


(Refer Slide Time: 09:27)



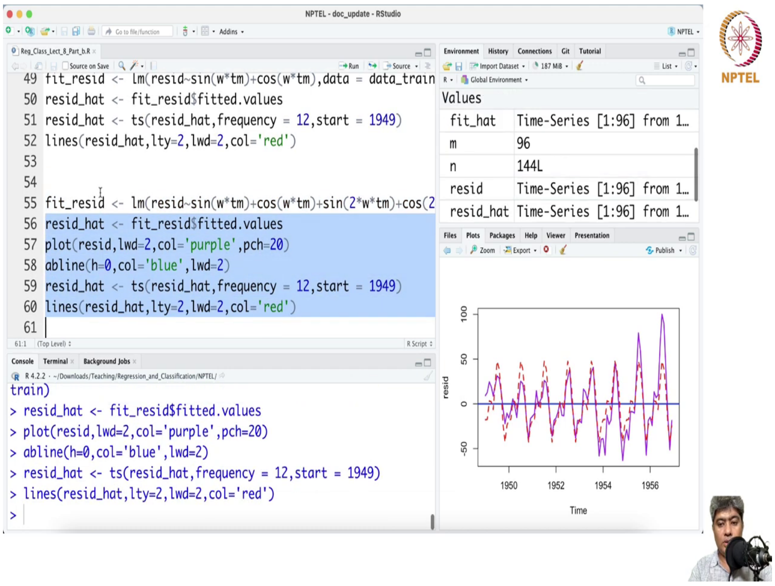
Now you can see that let me just enlarge this thing. So, now, you can see that this sin cosine is trying to capture the seasonality annual seasonality of the data, but it misses the like you know highs and lows of the data.

(Refer Slide Time: 09:51)

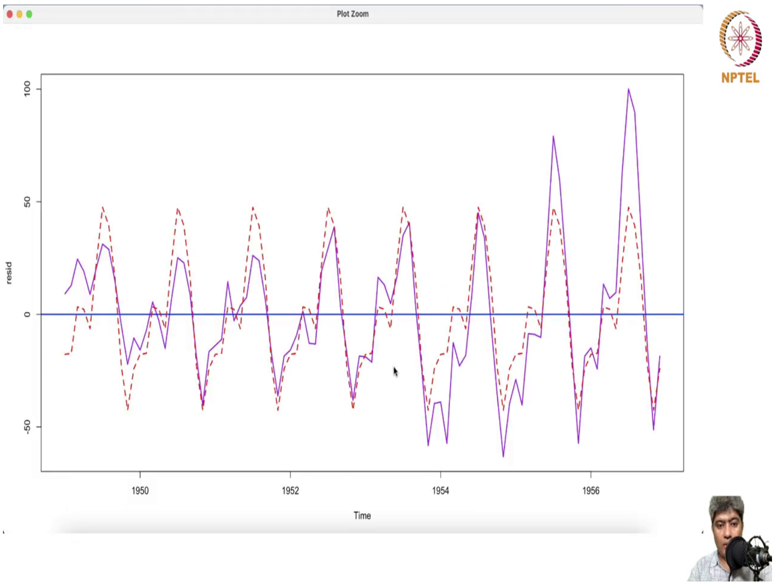


Now, it for then we fit a second order residual Fourier transform.

(Refer Slide Time: 09:57)

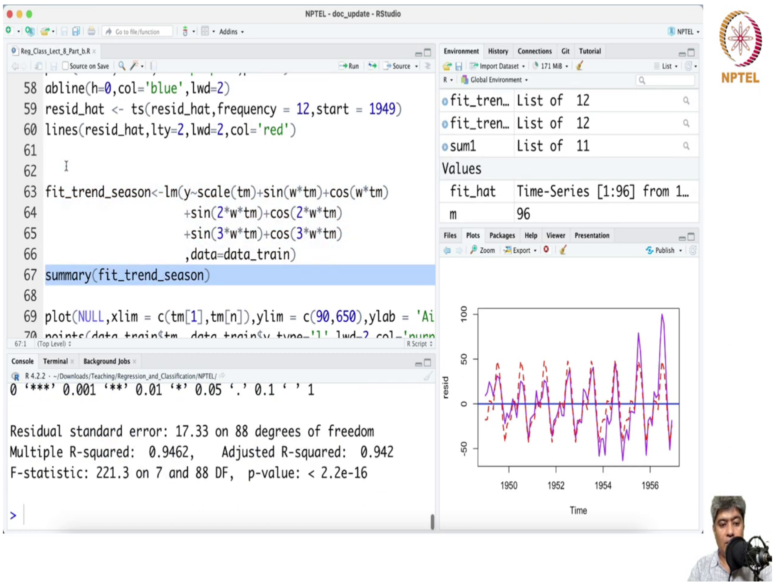


(Refer Slide Time: 09:59)



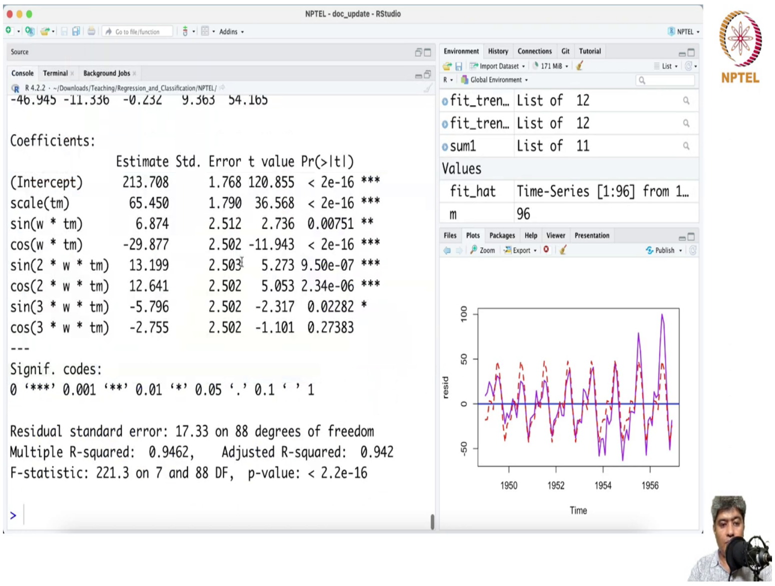
And we plot this and now we can see that this red curve is actually trying to capture the this this local seasonality this local seasonality is trying to be captured by the you know the second term Fourier.

(Refer Slide Time: 10:17)



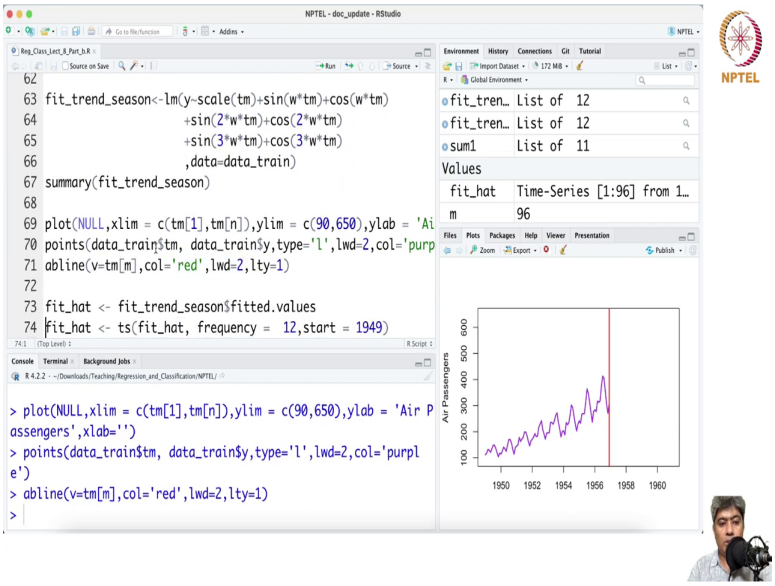
So, now if we fit a trend and seasonality together what we are getting is.

(Refer Slide Time: 10:31)



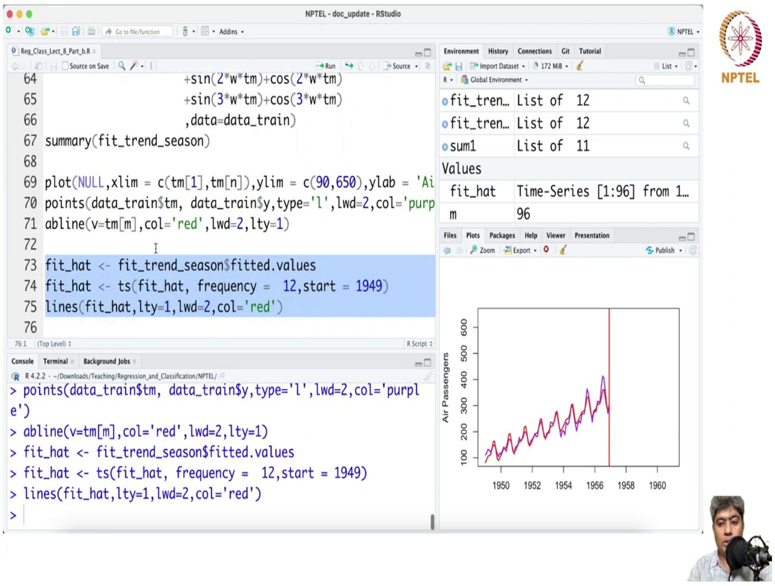
And then almost so, that you know the time that you know the sin cosine everything is kind of extremely significant and adjusted R square is 0.942. So, 94 percent of the x variability get explained by the model.

(Refer Slide Time: 10:52)



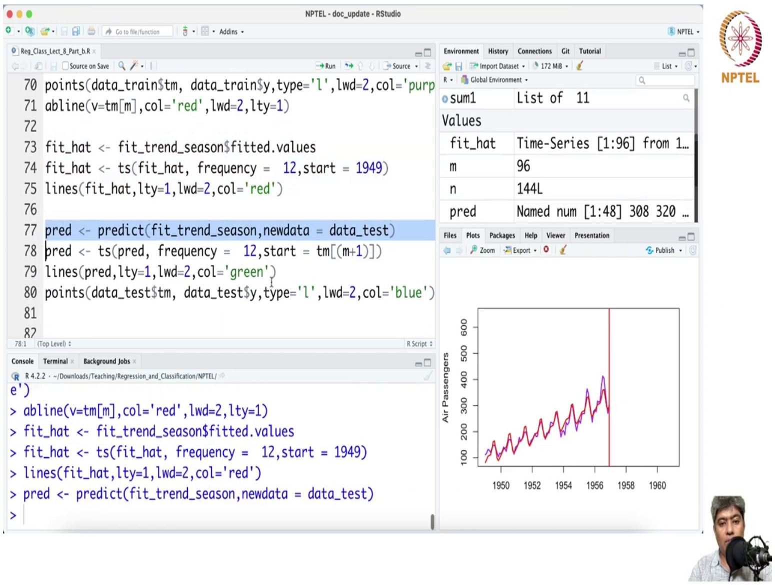
And then we plot the model and then fit hat.

(Refer Slide Time: 10:59)



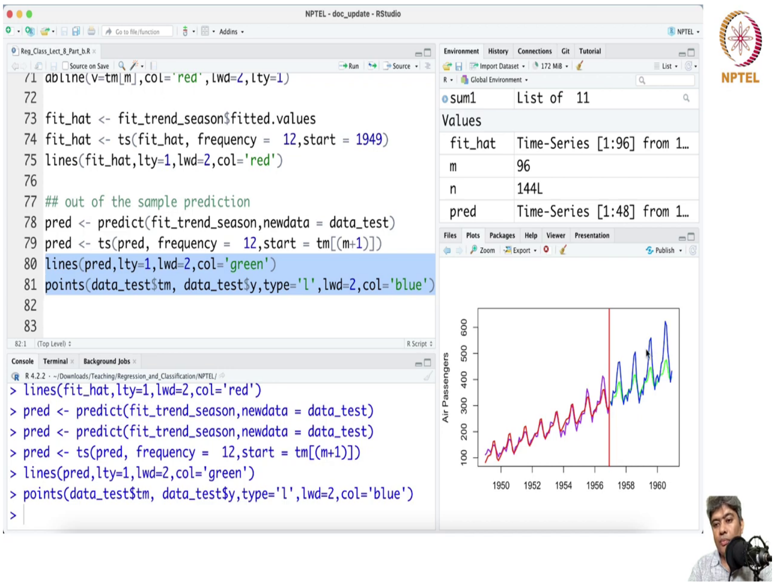
So, we can in the in sample the fitting is quite good.

(Refer Slide Time: 11:04)



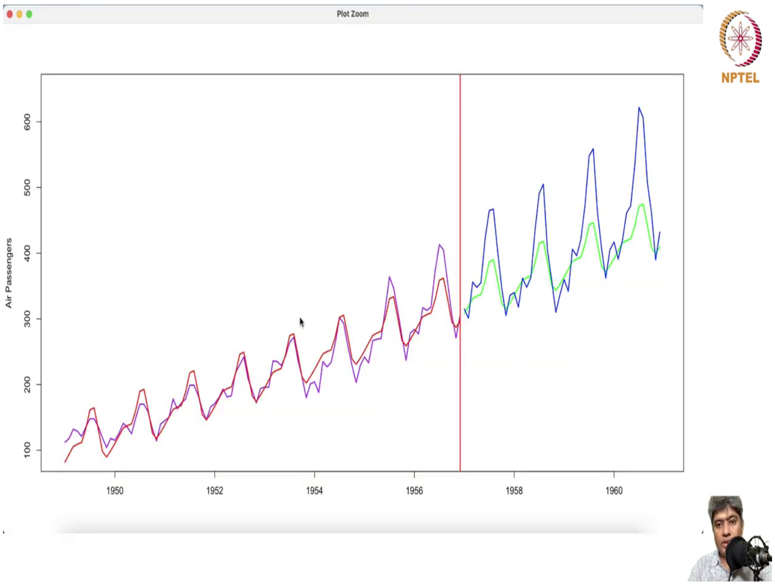
And let us do the out of the sample prediction this is out of the sample prediction: out of the sample prediction.

(Refer Slide Time: 11:10)



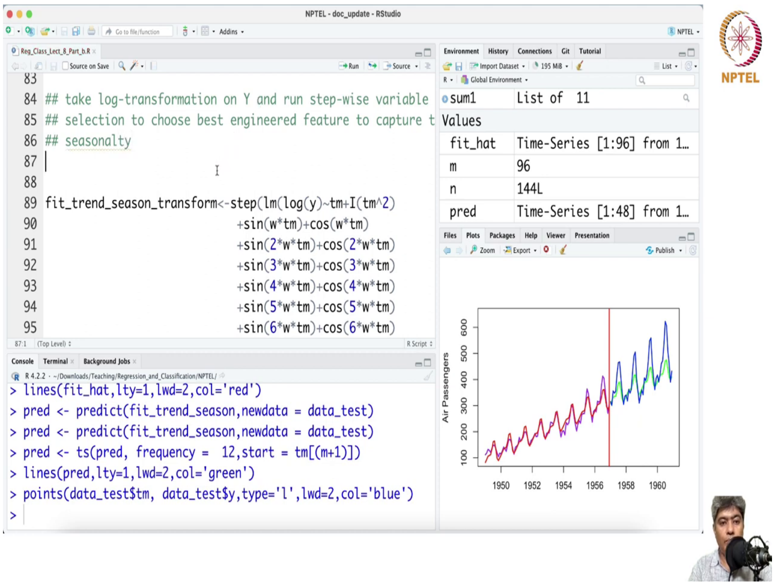
So, we do that and then we can see that you know its definitely missing the if you just let me just.

(Refer Slide Time: 11:36)



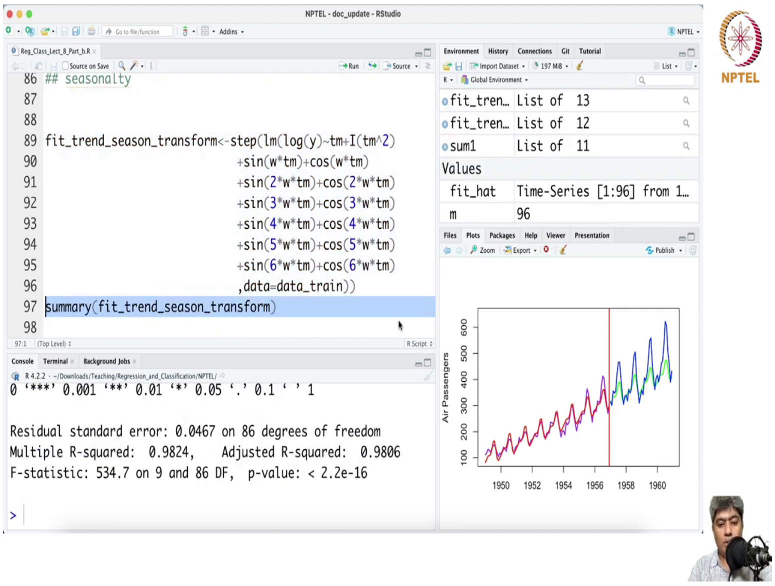
So, this reds and this greens are out of the sample forecast and clearly it is missing the highs of the you know max the peak seasons.

(Refer Slide Time: 11:53)

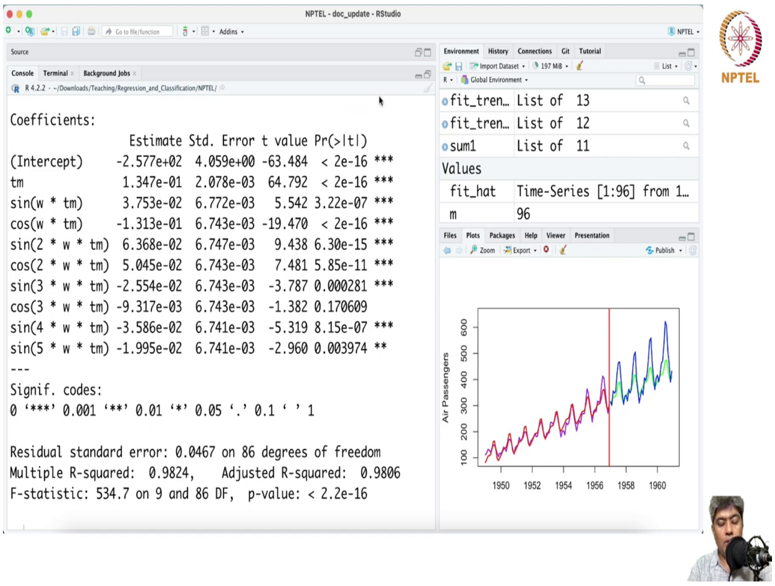


So, we take the log transformation we trend and seasonality with log transformation and then we on that we run a step by selection transformation and run step wise variable selection to choose best engineered feature to capture the seasonality dt ok. So, let me just run this. So, you can see sin cos now all the cos terms of fourth and fifth order cos terms are being dropped ok.

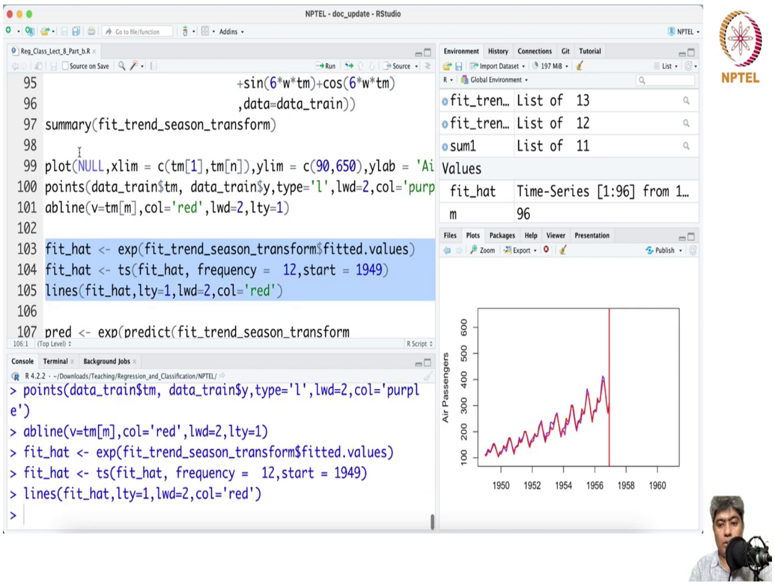
(Refer Slide Time: 12:42)



(Refer Slide Time: 12:47)

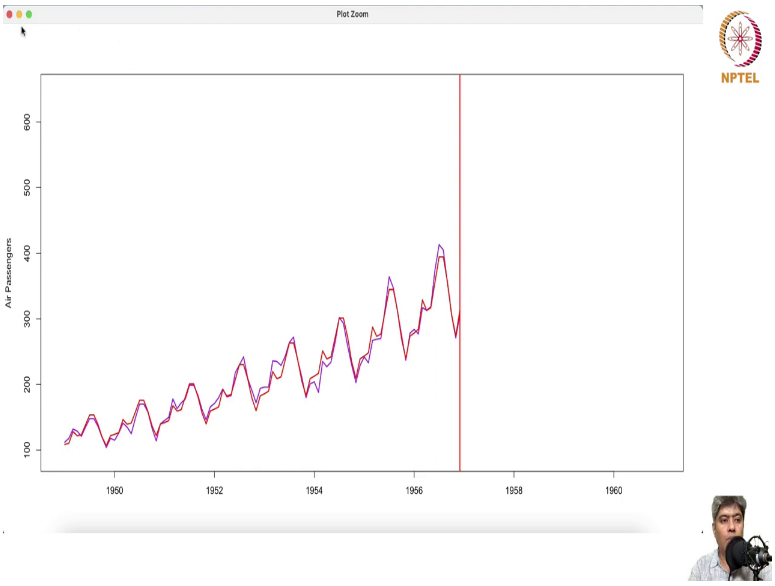


(Refer Slide Time: 12:59)



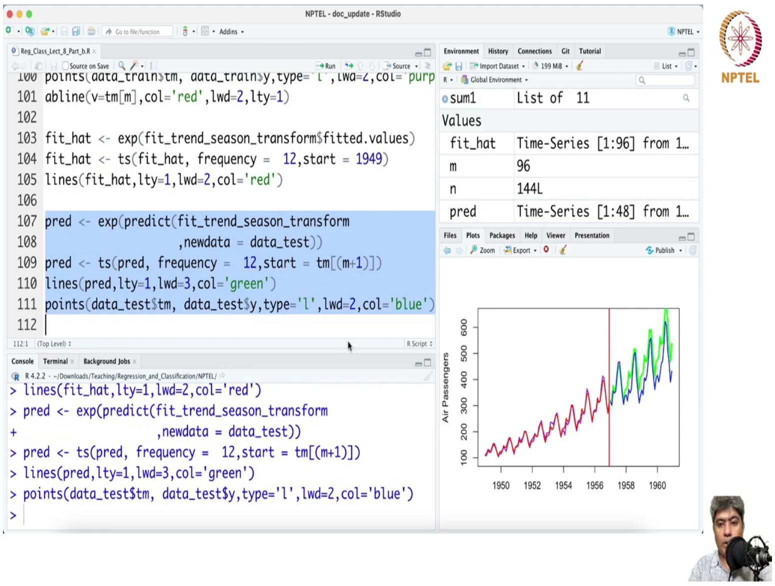
So, and sin 6 and cos 6 also being dropped. So, now, if we draw the plot.

(Refer Slide Time: 13:12)



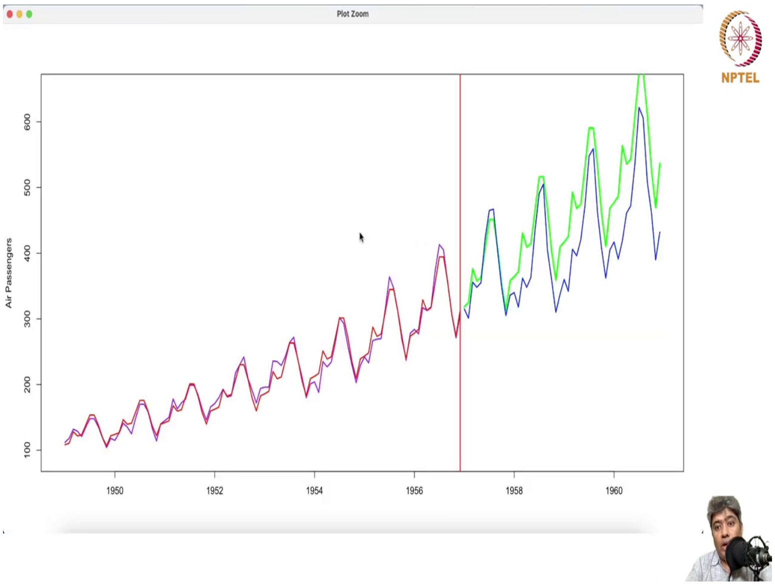
So, this is in sample fitting which is pretty much picking up the entire model enter data nicely doing.

(Refer Slide Time: 13:18)



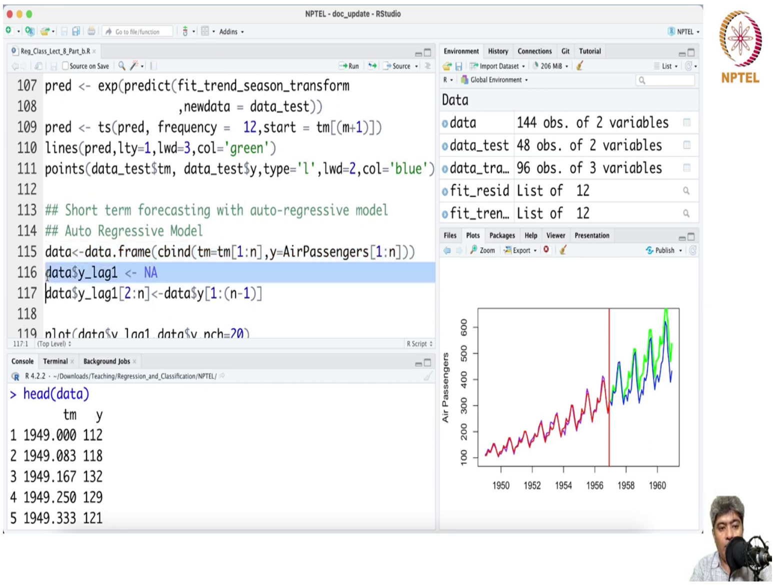
And this is out of the sample.

(Refer Slide Time: 13:24)



Forecasting the green is out of the sample forecasting it is slightly actually over estimating now.

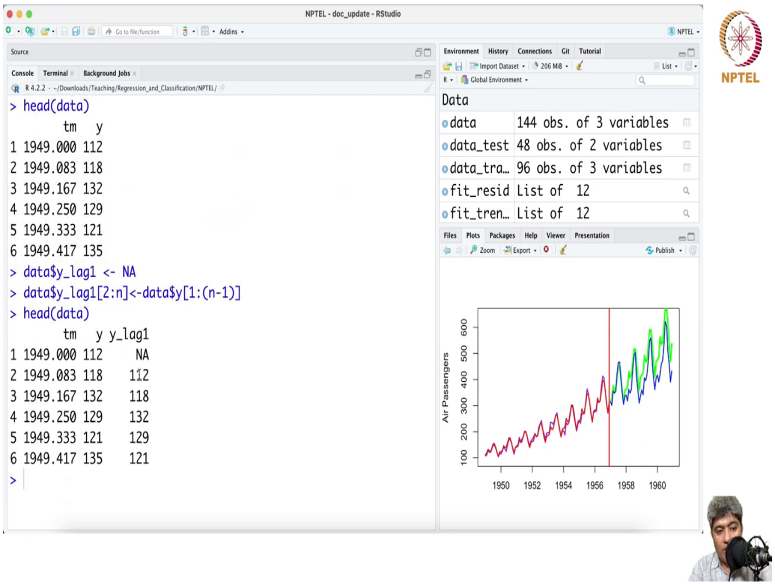
(Refer Slide Time: 13:31)



So, anyway so, this was the modeling long term modeling remember that actually why I am calling it long term modeling I am doing the entire forecast just standing here. Entire green forecast is being done while you standing in the 1957 remember that. So, it is I am doing the entire 4 year forecast while standing in the 1957 and it is doing reasonably good.

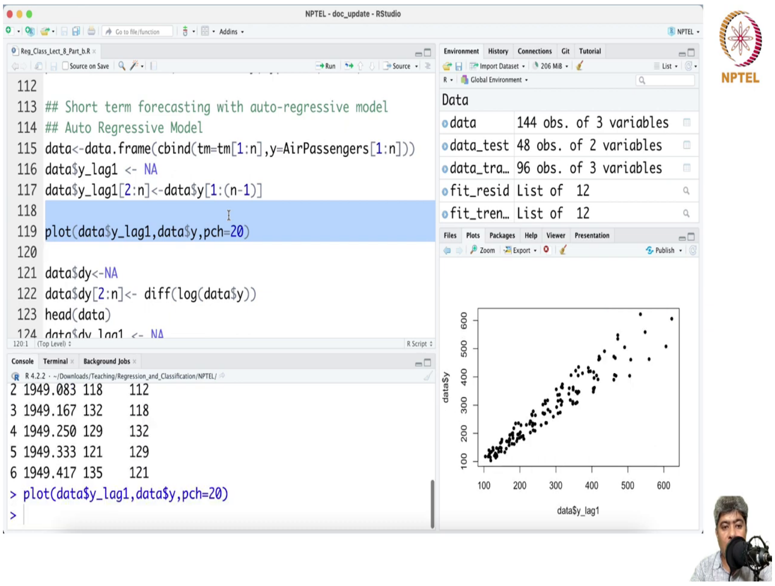
You can see it is yes of course, farther away the error will be higher, but at least the 1st year and 2nd year it has done p t decent job and it has picked up the nature of the trend and seasonality to a great extent. Now, we are going to do the short term forecasting short term forecasting with auto-regressive model ok. So, here is the data simple basic data that we have. So, basic time and the y ok and then we are taking the lag ones just created the lag ones.

(Refer Slide Time: 15:00)



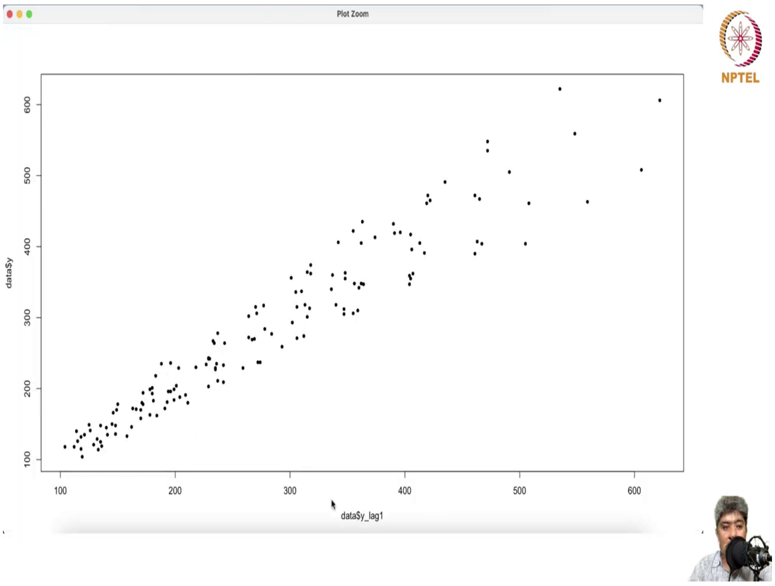
So, why so, 112 is being just placed here. So, if it is 112 on the February of 1949 the lag is 112 it has been kept here. If 132 is the value for the March in 112 18 is the lag value and that is being kept here similarly 132 is being kept here. So, that is how you create the lag variable ok.

(Refer Slide Time: 15:35)



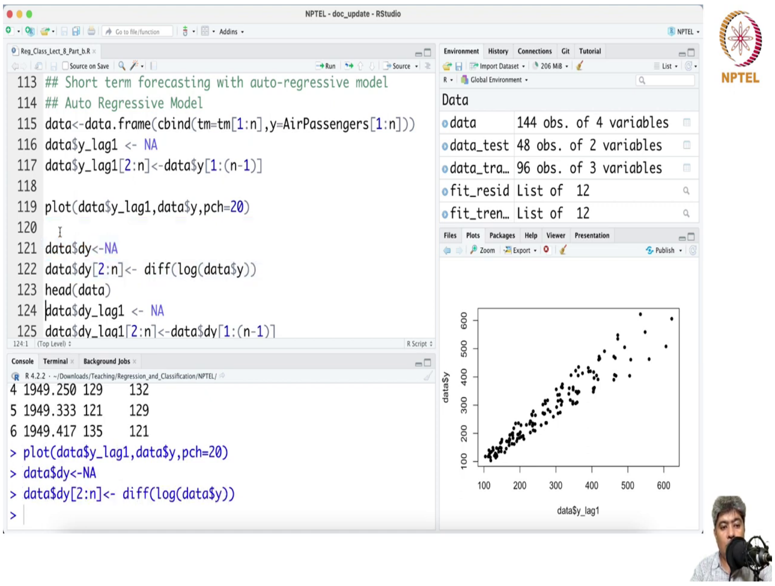
Once you create the lag variable now you plot the lag variable versus the actual variable.

(Refer Slide Time: 15:43)



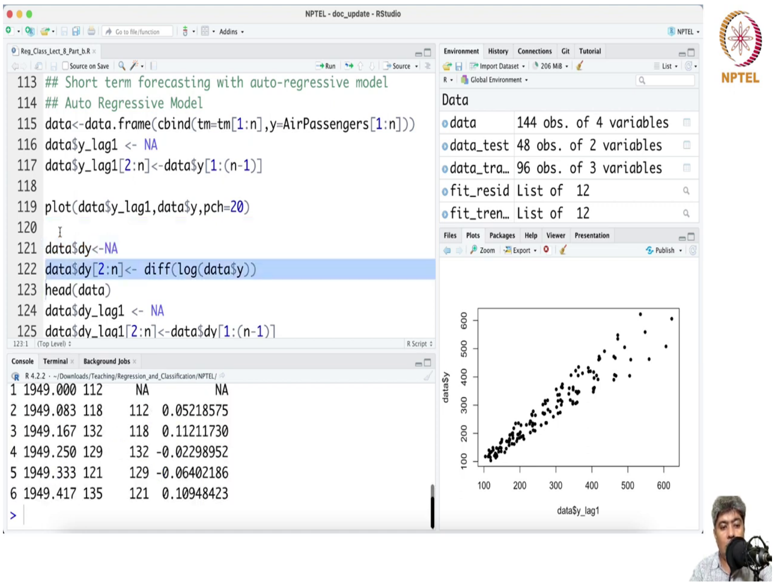
So, now, we can see this is the lag variable versus this is the original variable and there ok.

(Refer Slide Time: 15:50)



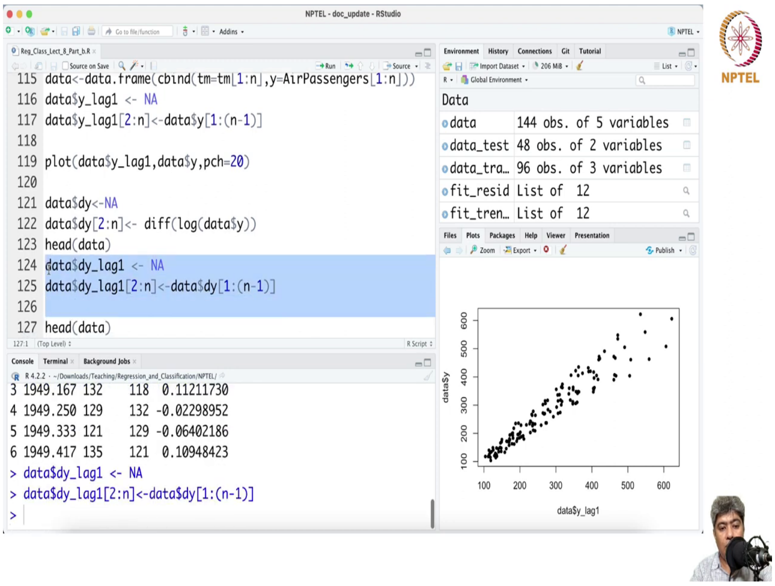
Now I am going to create a log difference of the values.

(Refer Slide Time: 16:00)



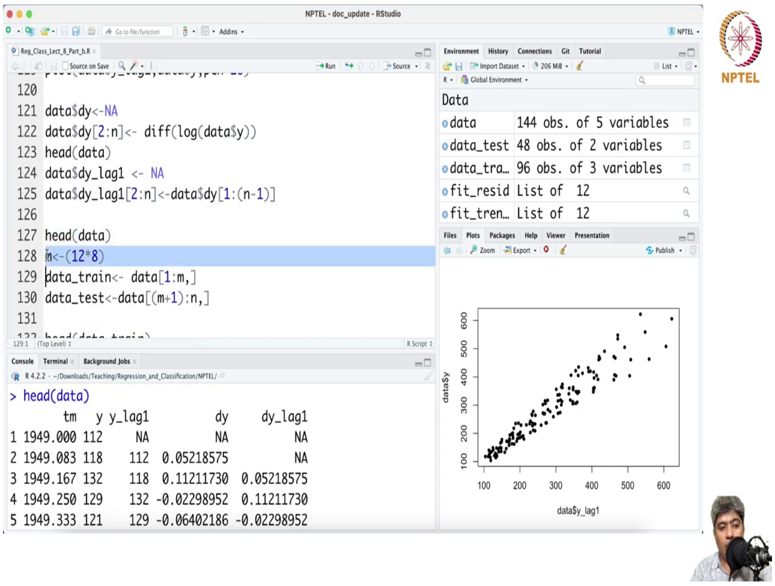
And then. So, if you now you see these are the delta y log differences.

(Refer Slide Time: 16:07)



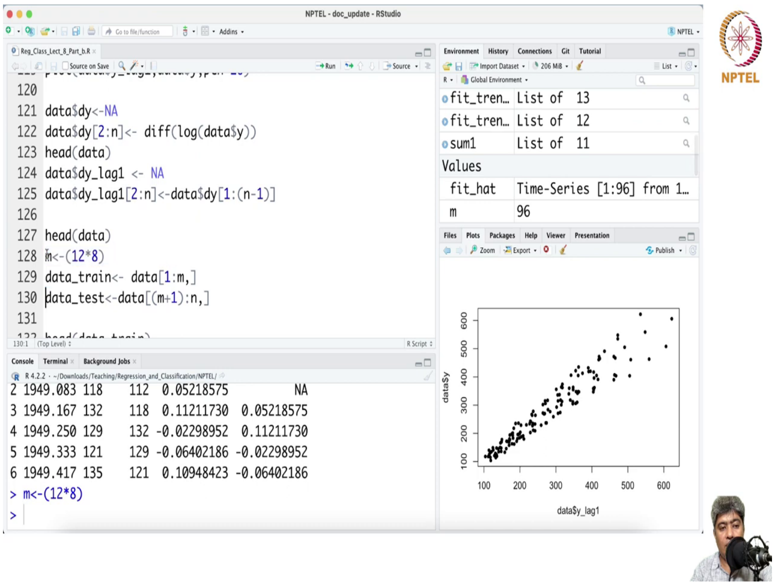
And then lag of the log differences.

(Refer Slide Time: 16:09)

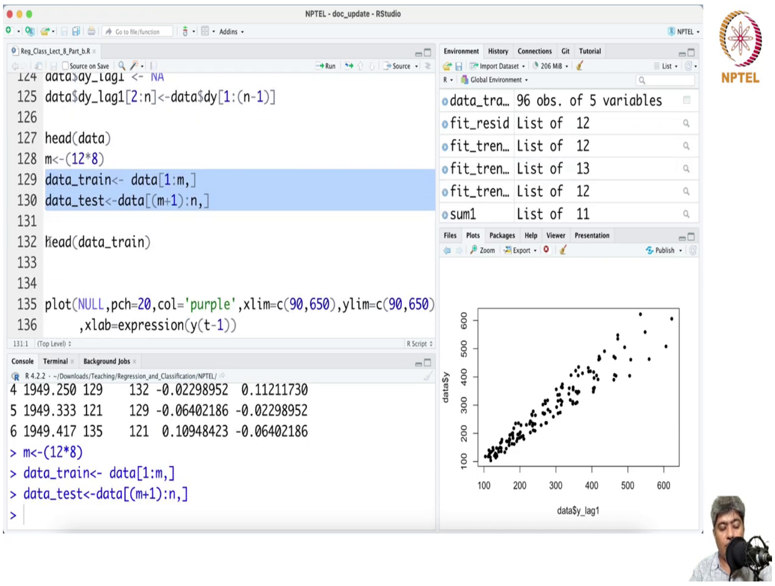


I have created lag of the log differences.

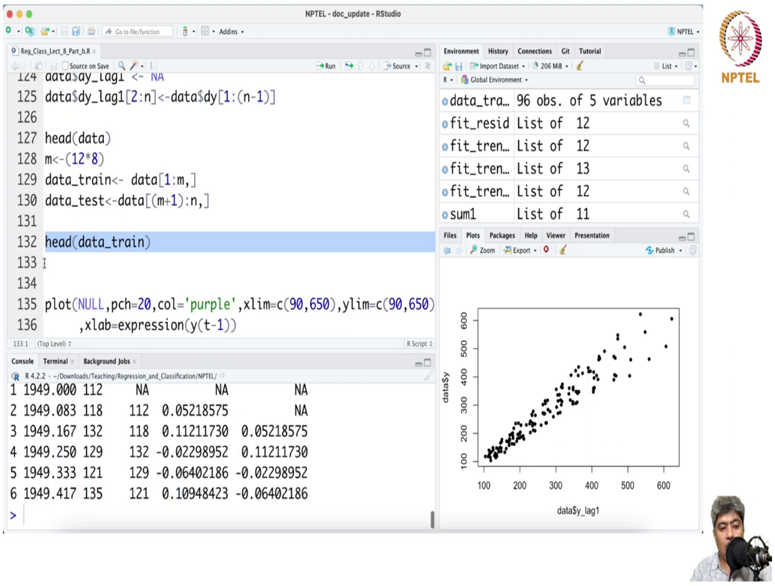
(Refer Slide Time: 16:15)



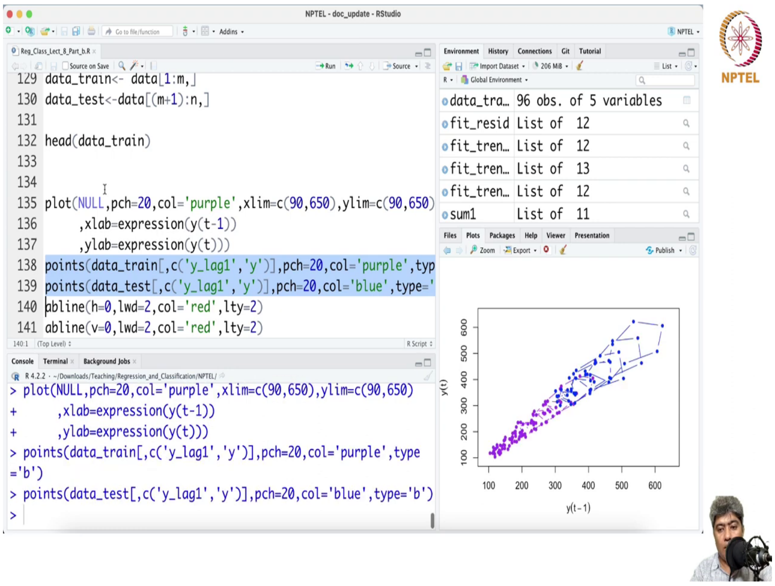
(Refer Slide Time: 16:18)



(Refer Slide Time: 16:21)

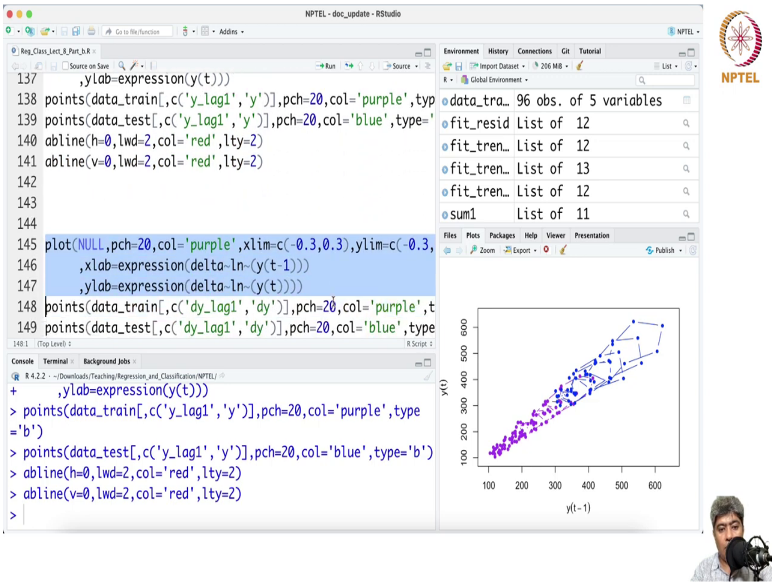


(Refer Slide Time: 16:27)



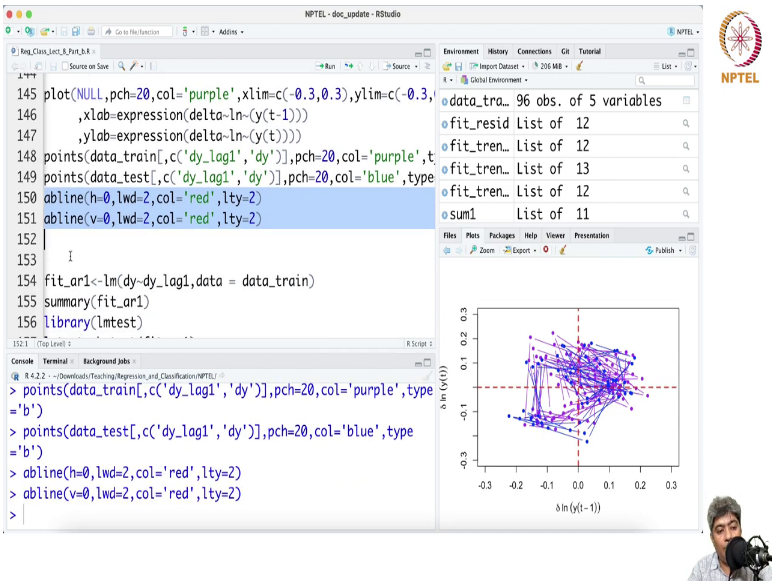
Again I am going to take the first 8 years of data as training data and next 4 years of data as the 4 years of data as the test data now I am going to plot this y t versus yt minus 1.

(Refer Slide Time: 16:40)



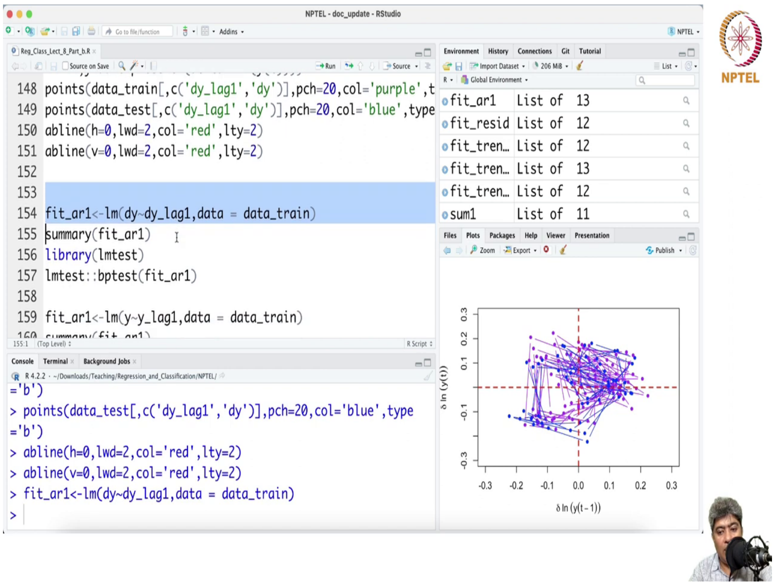
So, you can see this you know sort of a points ok.

(Refer Slide Time: 16:46)



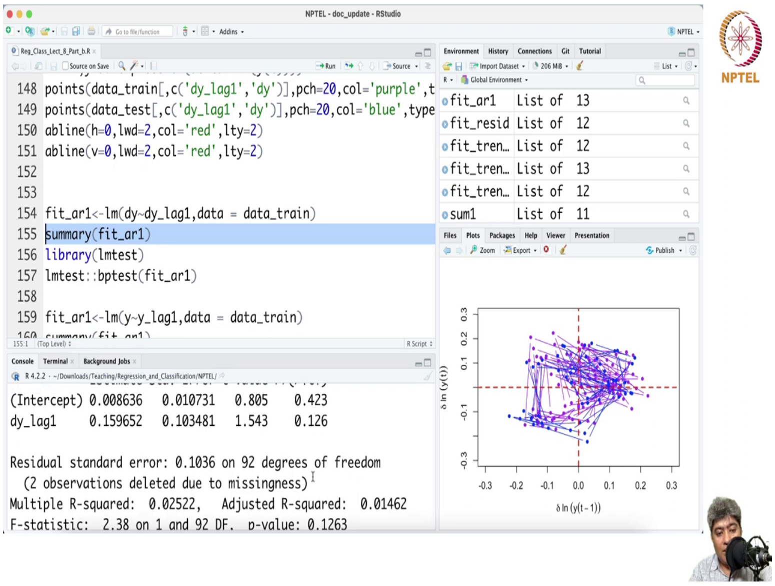
And then also this plot you can see the this is a delta log y t minus 1 versus delta log y t and you can see that there is a sort of a hole in the data right and this whole represent the seasonality ok. So, question is there a second hole there or is it just why this case is so tense. So, this is something the topological data analysis people do consider why it is happening why the way it is happening this topology what is the topology of the data alright.

(Refer Slide Time: 17:30)



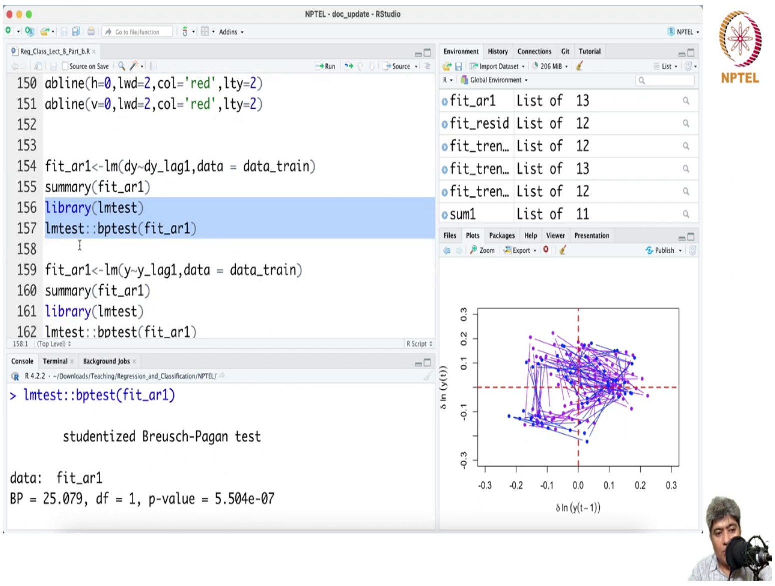
So, now, I am going to fit this model in delta y as a function of delta log y lag of y.

(Refer Slide Time: 17:39)



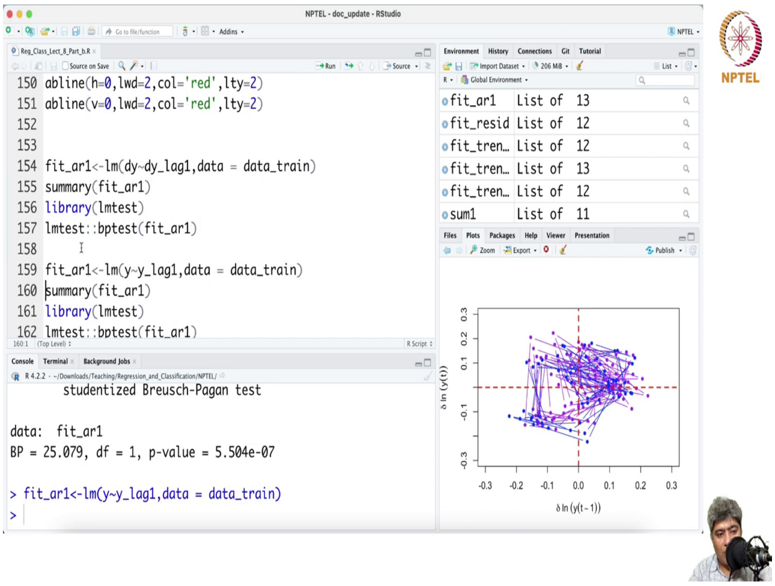
Then summary interestingly you see they are not very really not very significant none of them are really adjusted us, but not very high.

(Refer Slide Time: 17:53)



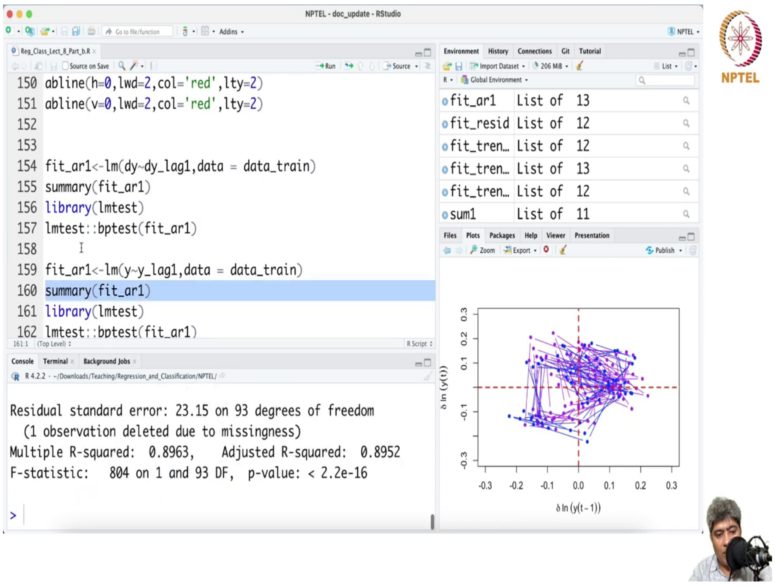
So, if I just run the bptest Breusch-Pagan test they are not also homogeneous. So, probably I will not use this model at delta low delta y level.

(Refer Slide Time: 18:07)

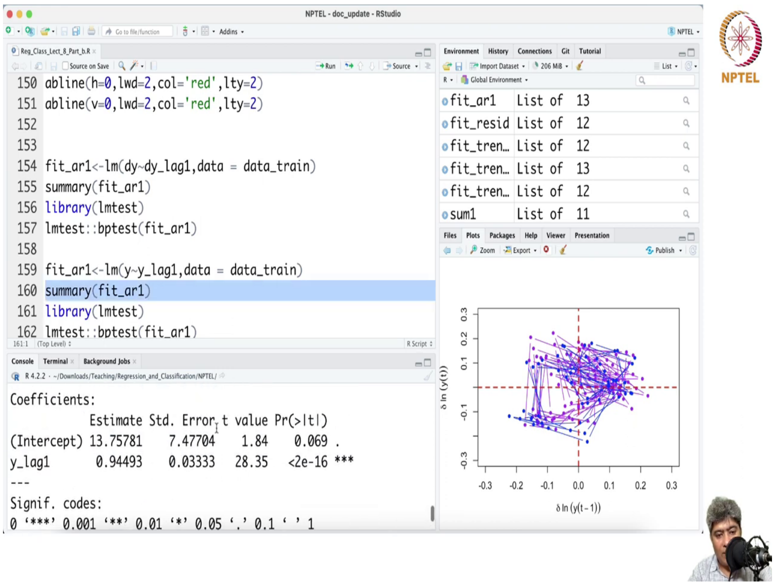


So, if you just model fit this model x simply y versus log y as a function of lag y auto simple auto regressive model.

(Refer Slide Time: 18:15)

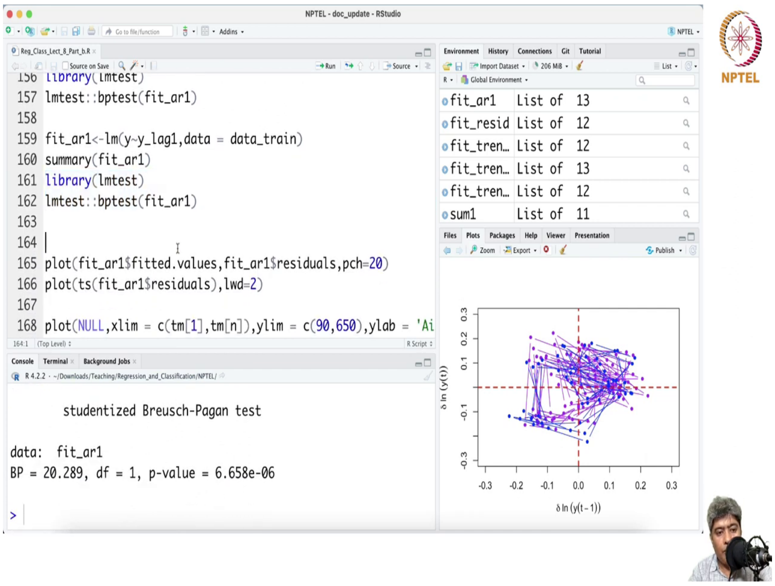


(Refer Slide Time: 18:18)



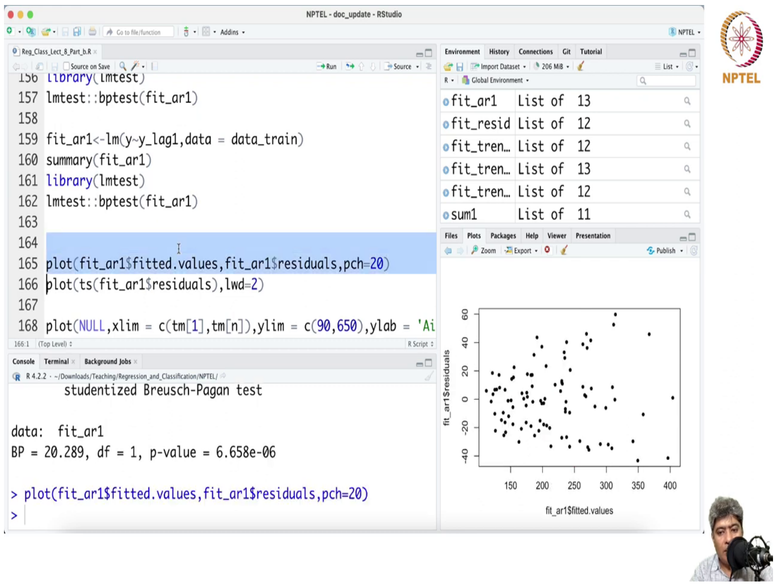
So, at least immediately what we are seeing that it is very strong model with Adjusted R-square 0.89 and we run a Breusch-Pagan test.

(Refer Slide Time: 18:27)



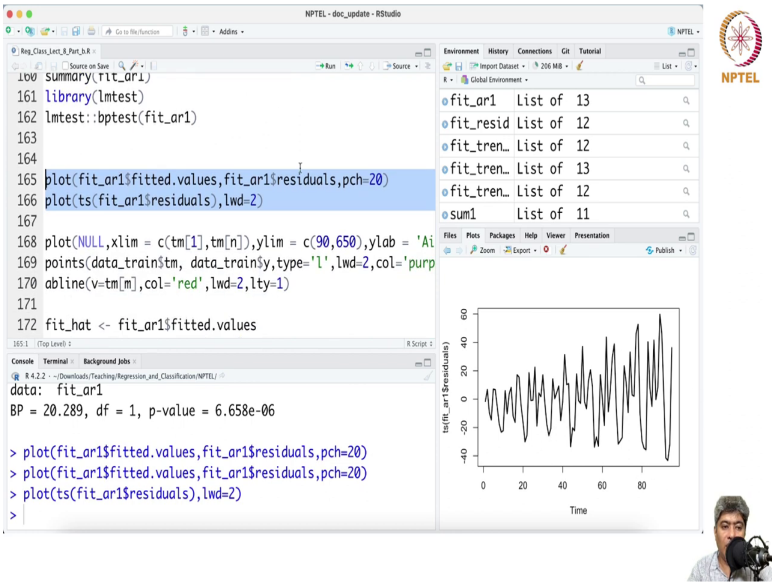
The Breusch-Pagan test still rejects the you know homogeneity.

(Refer Slide Time: 18:36)



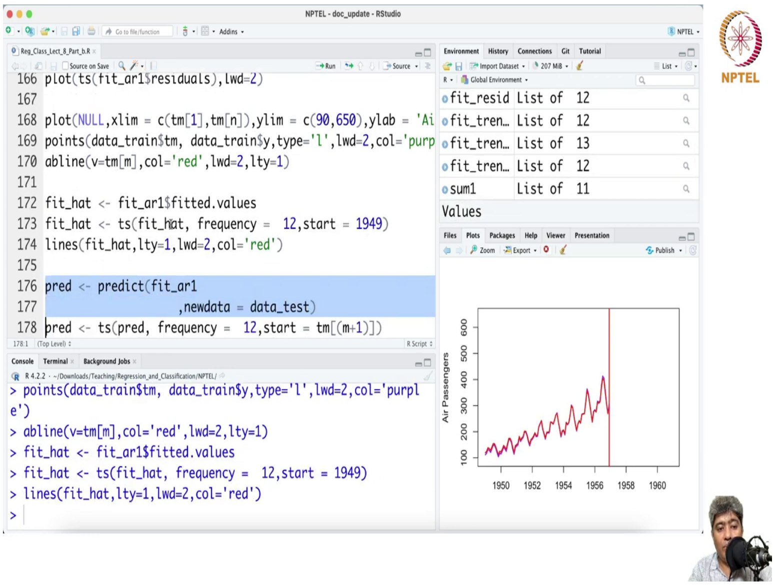
And if we fit this you can clearly see at the beginning it was much more dense, but as it increases the residual increases with that its slight a fan kind of you know fan kind of behavior. So, clearly its not homogeneous.

(Refer Slide Time: 18:55)



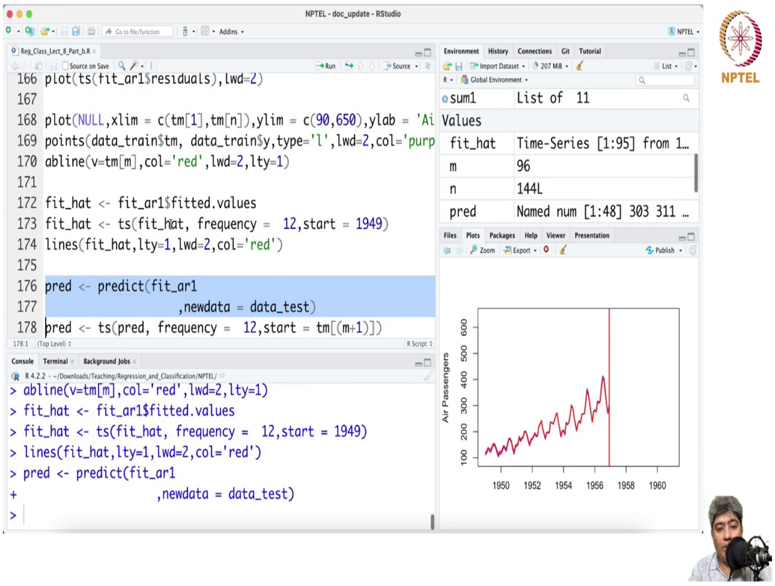
And a line plot also sort of in initially it is the residuals these are the residuals much less, but the residuals much higher at the end points.

(Refer Slide Time: 19:12)

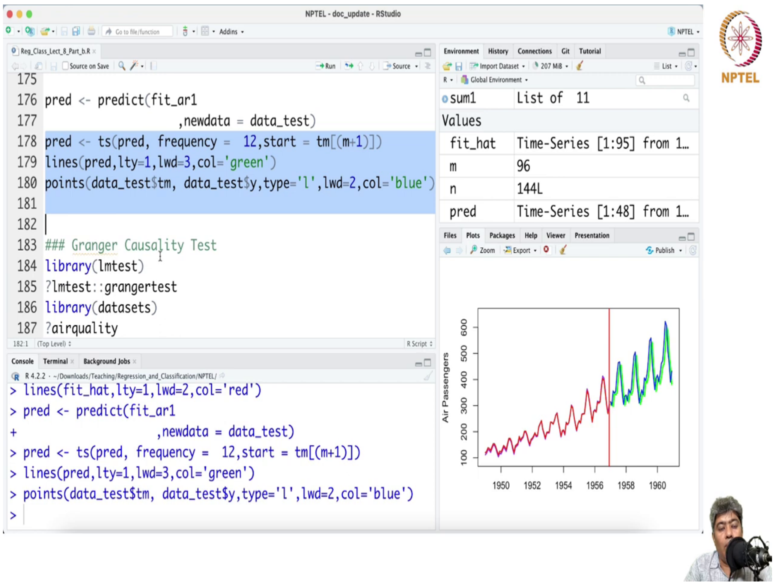


So, if we just fit this guys, but still if you fit this guys that they are doing pretty good job in terms of prediction and in terms of out of the sample prediction.

(Refer Slide Time: 19:17)



(Refer Slide Time: 19:19)



So, we will stop here we will stop here and next video we will move to the Granger Causality test and what is Granger causality and how it is going to you know play a big role in the causal inference and how simple linear regression model that we discussed in this; discussing in this can be used to develop some Granger causal model and Granger can be used to in in defining Granger causal inference.

Thank you very much, see you in the next video, see you soon, take care bye.