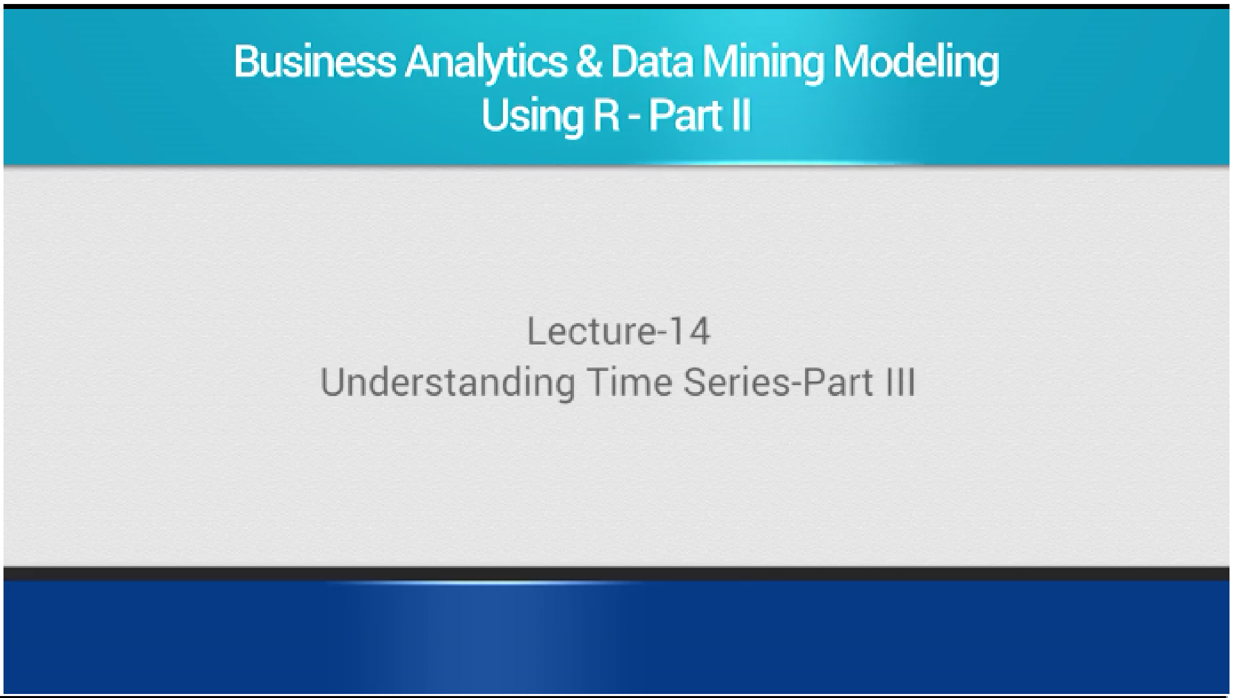
INDIAN INSTITUTE OF TECHNOLOGY ROORKEE
NPTEL
NPTEL ONLINE CERTIFICATION COURSE

Business Analytics & Data Mining Modeling Using R – Part II

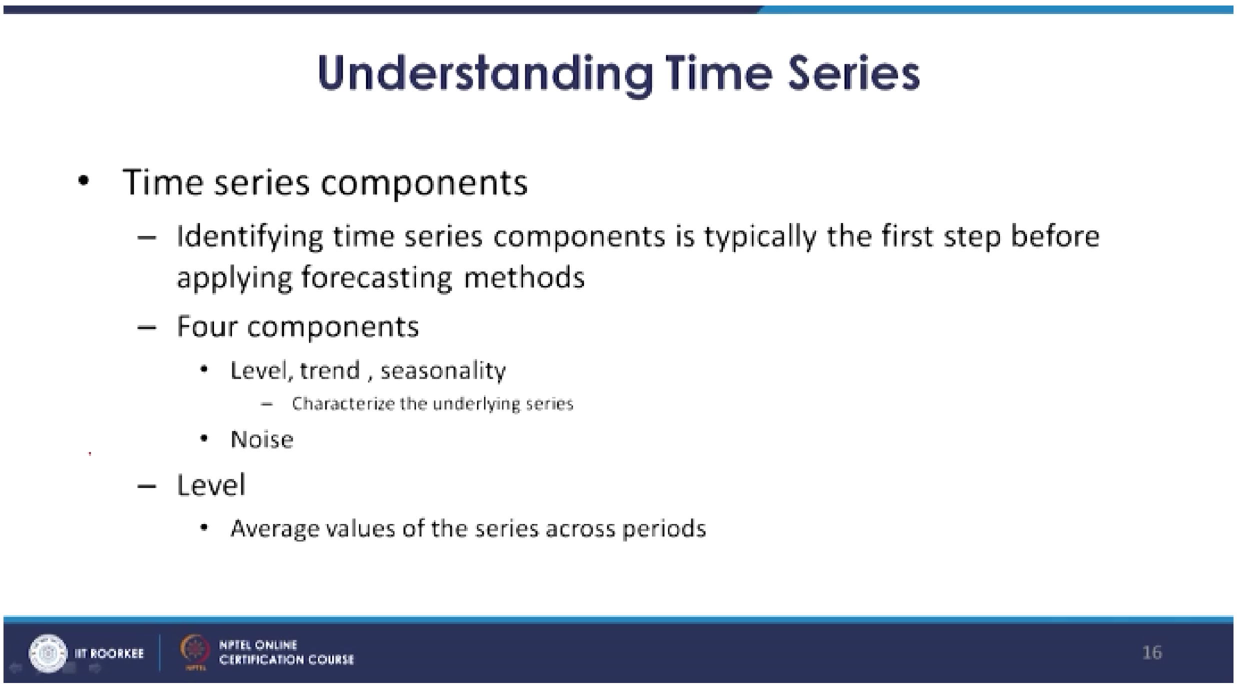
Lecture-14

Understanding Time Series– Part III

With
Dr. Gaurav Dixit
Department of Management Studies
Indian Institute of Technology Roorkee



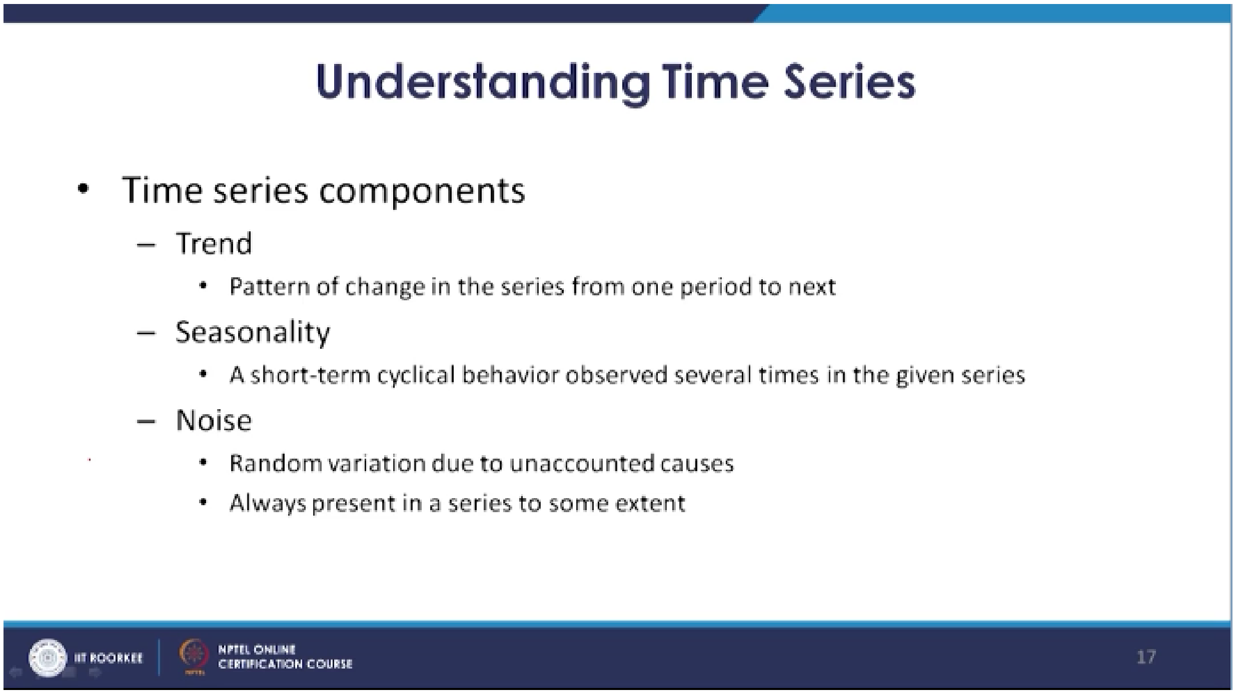
Welcome to the course Business Analytics and Data Mining Modeling Using R – Part 2, so in previous few lectures we started our discussion on our last module that is time series forecasting, we also started discussing the first topic in this particular module that is the understanding time series, so we have in previous few lectures we have been discussed a bit about time series.



So in the last part of previous lecture we started our discussion on time series components, so let's start from there itself, so we talked about the time series components and typically we said as mentioned the slide as well that this is probably the first step that one has to perform, before we can move ahead and apply more formal techniques, so identifying time series components is typically the first step before applying forecasting methods.

So as we discussed there are four components for a time series, so three components we have grouped together here, the level, trend and seasonality, because these are the three components which actually characterize the underlying series and the remaining part is the noise, so noise is the fourth component.

So then we started discussing each of these four components one by one, so let's again discuss the first component that is level, so as we discussed in the previous lecture level is average values of the series across periods, so this gives us an indication about the, whether on an average level, overall average level, how the values are being taken by the time series, so that's sense that indication we get from the level of the time series.

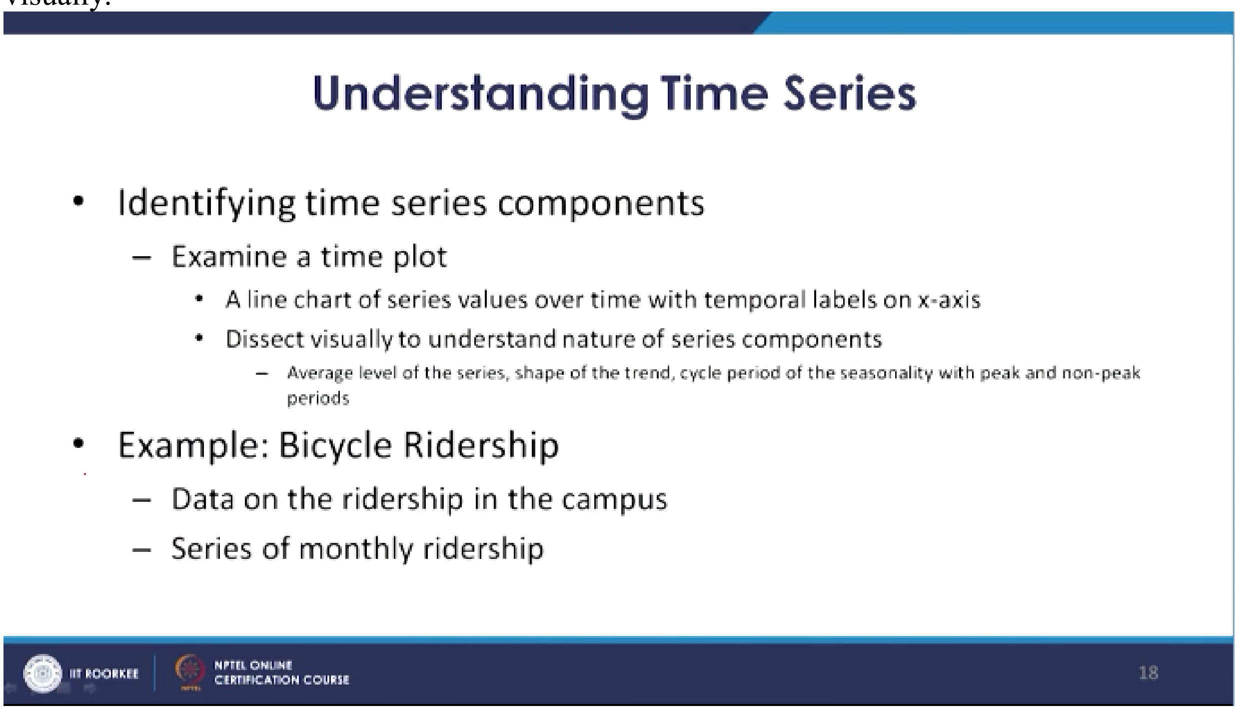


If we look at the second component trend, this is more important so this particular trend tells us about how pattern, how values are changing in series from one period to the next, so pattern, so capturing the pattern of that change, how the values are changing from one period to next that is defined as trend, so as you can see in the slide, so trend is pattern of change in the series from one period to next, this is more important part of time series and as we have discussed in previous lecture, two types of approaches where one is model driven, the another one is data driven so there if we are looking for you know, if we are specifically focusing on trend component, how we should, what kind of technique, what kind of method we should be using for modeling, so lot of that is going to depend on this particular component, trend component, so we'll discuss this in more detail in this and coming lectures.

The third component which is also very important is seasonality, so what is seasonality? So as you can see in the slide that seasonality can be defined as a short term cyclical behavior observed several times in the given series, so typically what you know swings that we see in a particular series, so if those swings are you know short term you know they keep on reoccurring in a short-term manner, in a cyclical manner, several times within the you know specified, series within the overall time period of the series that is there, so that is defined as seasonality, so seasonality is a short term cyclical behavior observed several times in the given series, so trend and seasonality these are the two most important components of time series, so whenever we are looking to model a time series for forecasting purposes, it is these two components trend and seasonality which we specifically focused on.

So this discussion will happen later in the coming lectures as well, the last component of time series is noise, so as we have done in our previous course as well, the noise is defined in the you know similar fashion, so noise here is also random variation due to unaccounted causes, so anything that we cannot understand it becomes part of the noise term, so random variation due to unaccounted causes, we call it, we define it as a random variation, because there is any pattern, if there is any structural left in the series then we are supposed to capture that, so whatever is left over you know so that should not obviously have any trend or you know any structural format or anything, so therefore the remaining part should be the random you know

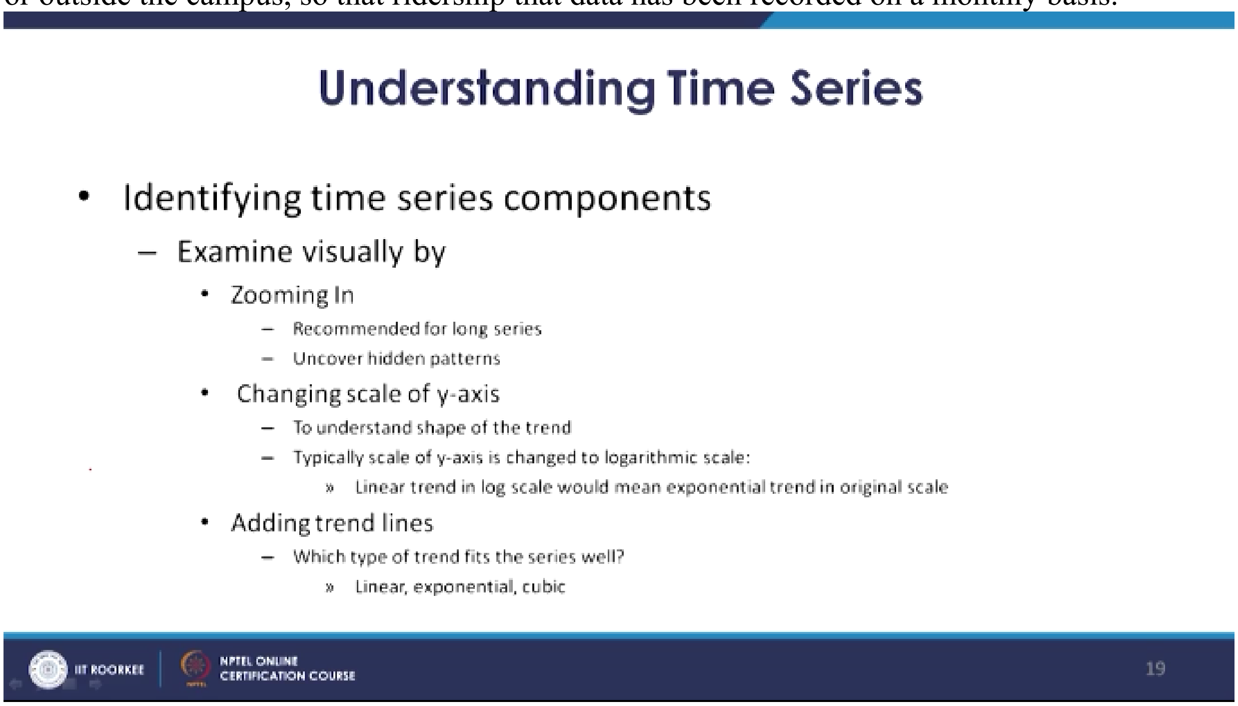
variation part which is why noise is defined in this fashion, so noise is random variation due to unaccounted causes. So always present in a series to some extent so there is always going to be you know some noise that is going to be associated or present in a series.
So these are four components, so first one is level, the average values, the overall average values that are taken by the series, then the second one is the trend which tells about the you know change, rate of change, pattern of change in the series, and the third one is about the seasonality if there are any cyclical behavior that is present in the series, then the fourth one is the noise, the unaccounted part, so let's move forward.
So once we understand that these are the four important components of a series and specifically and you know understanding or dissecting trend and seasonality components is more important for our modeling purposes for forecasting purposes, so how do we go ahead and identify these time components, so there are various ways to do this, so first we'll look at some of the visual techniques you know visual inspection that could be used to identify these components, and you know these components, how whatever we have learned by applying these visual techniques, visual methods, how they can be incorporated in our formal modeling process, so we'll see that later, so let's start with how some of these you know time series components can be inspected visually.



So the most important you know graphic, the technique, visualization technique that can be applied is you know time plot, so we should create a time plot and by examine the time plot of a particular series we can look for, inspect for manually observe what are the you know component and how they are there, so first point is always going to be examining a time plot, so what is a time plot? So it is a line chart of series values overtime with temporal levels on X axis, so in previous course on business analytics and data mining modeling, so where we have, where we have had extensive discussion on visualization technique, we have also taken you know I think the same dataset, the same example on bicycle ridership and we have you know analyzed looked at the time plots and other things, so same thing, similar things we are going to repeat here in our discussion time series forecasting as well, so time plot is a line chart of series values overtime with temporal levels on X axis, on X axis we have time index and on Y axis we have the series values, so that line chart is essentially the time plot.

So this time plots we are supposed to dissect you know, visually this time plot to understand nature of series components, so we look to identify average level of the series, so once we have a look at the time plot we'll be able to see the you know visually also we'll able to identify which value on the Y axis is typically in a way defining the level of the series, we can also see the shape of trend whether it is you know linear or you know exponential or polynomial so that kind of thing we can also observe using the time plot of series.

Then cycle period of the seasonality, so if there are any cyclical behavior that is present in the series that should also be clearly visible in the time plot, we would be able to you know more often they are not, we would be able to identify the, you know cycle period and if there are you know and what are the peak you know period, what are the non-peak periods, so in that sense so some of these details we would be able to observe from time plot, so what we are going to do now we'll discuss this example dataset bicycle ridership, so this particular dataset, data is on the ridership, ridership in the campus so this is about the bicycle ridership, so the number of bicycles that are being used in the campus for you know travelling within the campus, so within or outside the campus, so that ridership that data has been recorded on a monthly basis.



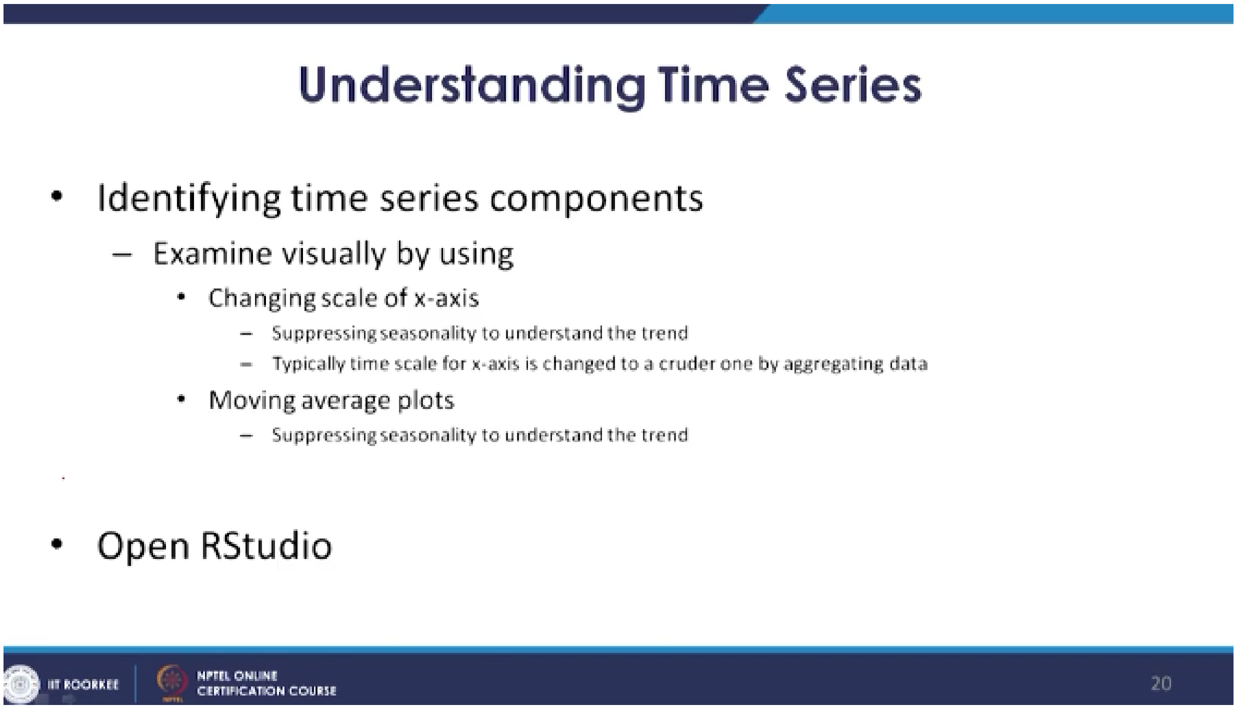
So as you can see here next point series of monthly ridership, so for each month number of riders that have been there, so that is recorded and this particular you know series, number of riders per month and based on that the series has been created, so we are going to analyze this series to find out the you know time series components, so before we go ahead and do that in our R environments let's discuss a few more points about identifying this time series components, so one thing that we talked about is time plot, so visually inspecting, visually analyzing time plot can help us in identifying time series components, other ways are also listed here, so if we can examine time series visually by zooming in, so sometimes the period that is involved in a time series could be quite long, so in that sense when we create a time plot so the values can get mixed up and if there are you know any you know for example if there is any cyclical behavior shown by the series that would not be clearly visible, so there are you know some components which would not be you know very clearly you know inspect it using a time plot just because the specified period is quite long, so zooming in is one way we can you know handle a long series, so as you can see zooming in recommended for long series so it can uncover hidden patterns, so typically when we talk about hidden patterns it could be understanding the cyclical behavior and therefore the seasonality component of the time series.

If we look at other ways, so the second one is also there, the changing scale of Y axis, so as we talked about that on Y axis we have the values related to time series variable, for example the dataset that we just discussed bicycle ridership, so on Y axis we are going to have the number or riders, and on X axis the time index is going to indicate the month 1, 2 and 3, 4 and so on, so that would essentially be Jan, Feb, March, so in that fashion, so for each of those calendar year months what has been the you know ridership, what has been the number of riders in the campus, so that we can, that we typically can observe in the time plot.

Now changing a scale of Y axis can help us understand the shape of trend, so as we talked about what is a trend, so a trend essentially we try to capture the pattern of change in the series from one period to next, so how changing scale of Y axis is going to help, so let's look at the other point, so typically scale of Y axis is changed to logarithmic scale, so what would happen if we do that? So for example if linear trend is visible, so if we create the you know, if we change the Y axis scale to you know logarithmic scale then what is going to happen is if there is any you know linear trend that is visible in the log scale that would been exponential trend in the original scale, so sometimes in the original scale if the exponential trend is not quite clearly visible due to the nature of values taken by the series, so if we transform the scale on Y axis let's make it logarithmic scale and if somewhat linear trend is visible then that can give us the confirmation that the original series, in the original series the trend must have been near to exponential, so in that fashion changing a scale of Y axis can give us you know understanding about the trend, that pattern of change in the series from one period to next.

Then let's move forward the next thing that we can do to visually inspect the time series is, adding trend line so which type of trend fits the series well, so we can add in the time plot itself so we can add some of the trend line, for example typically we expect the trends to be linear exponential or cubic, so therefore we can add these trend lines on top of the time plot and we can visually you know analyze whether those trend lines are fitting the time series values or fitting the data or not, so that would also give us a sense in terms of the kind of trend that is there in the series, so adding trend lines can help us in identifying the trend of the series.

Let's move forward, so there are few more points which can also be used to find out you know to understand, to identify different time series components, so the first one here in this particular slide is changing the scale of X axis, just like we talked about changing the scale of Y axis and how, you know, if in the original series exponential trend is not, you know cannot be clearly confirmed and changing the you know scale of Y axis to logarithmic and if we see that close about linear trends, so therefore that confirmation can be done.



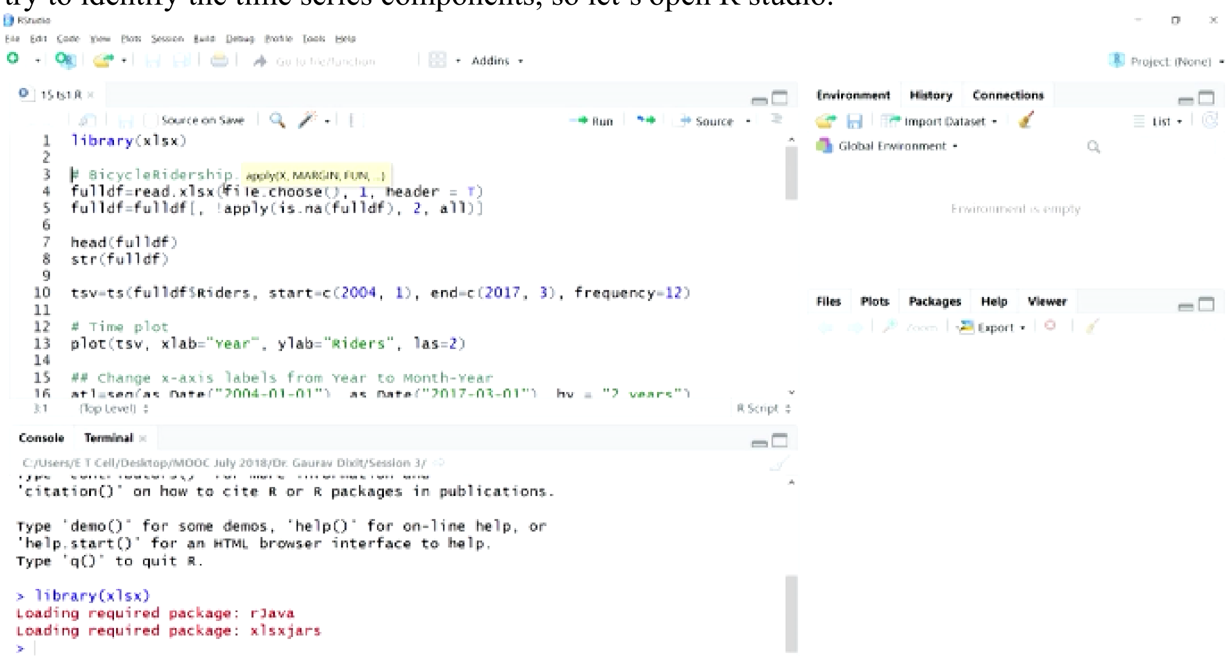
Similarly we can also change the scale of X axis, so typically this is done so that do suppressed seasonality, so seasonality can be suppressed by changing the scale of X axis and that would actually help us in understanding the trend component, so suppressing seasonality to understand the trend, if we change the scale of X axis, so typically time scale for X axis is changed to a cruder one by aggregating data, so as we have talked about in the previous lecture also that when we are discussing the level component specifically that if we aggregate data to a course of level then some of the sources of noise might cancel out, might average out, therefore the forecasting part of the series might increase and all that, so that was the discussion.

Similarly, if we change the scale of X axis so essentially that would mean aggregating data, so if we change this scale to a cruder one, then of course some of the noises will cancel out and typically when the noises cancel out we are, you know some of the cyclical behavior, so as you might have seen before that if you create a plot you know for a shorter time period you would essentially see you know some sort of cycles in those plots, in those time plots, however if you take longer horizon so those values average out and you see a clear trend, so we are essentially talking about that scenario, so if we change the scale of X axis for example if we take the India's GDP number and if we create a plot for GDP numbers for last decade from 2005 to 2015 or 2007 to 2017 that kind of you know series we plot we would you know see ups and downs in the GDP numbers, however if we take a plot from, for last 3 decades and we would see a linear trend there, right where the GDP has been continuously growing, so once we change the you know scale of the series, once we increase the length, the some of the things, some of these sources of noise cancel out, so essentially that also suppresses the seasonality you know component of the series, and that would actually help us in understanding in a visualizing the trend component of the series, so this is one thing that can be done.

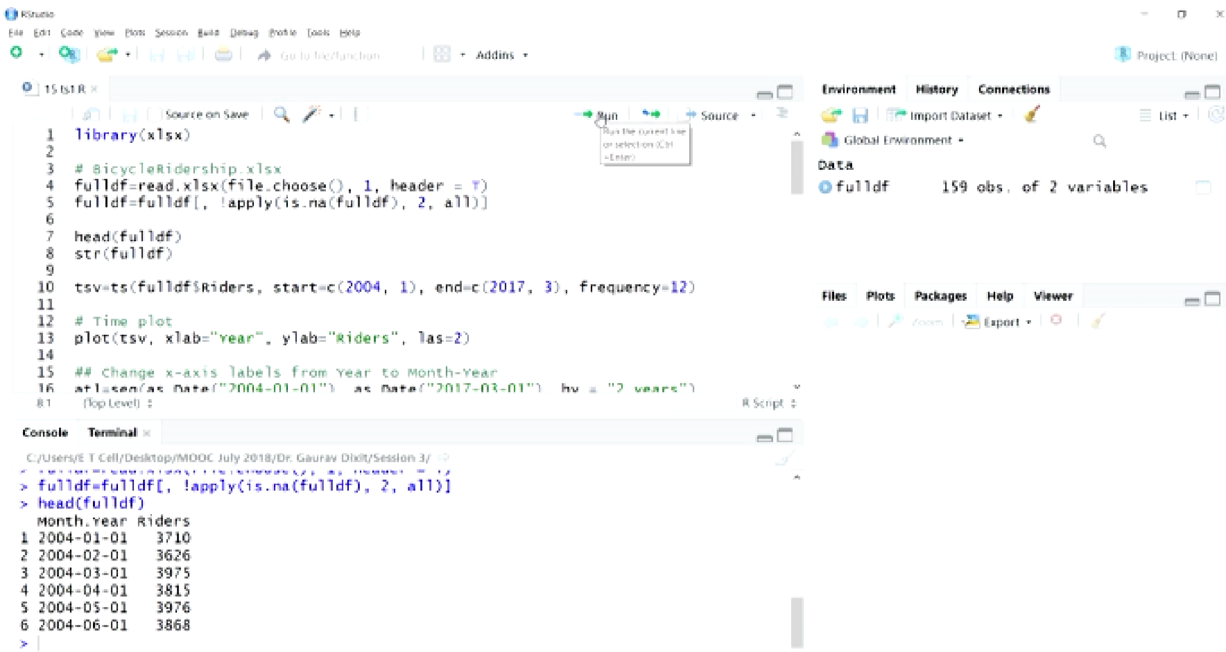
So next point is about moving average plots, so this particular point is actually does the same thing, so if we create moving average plots, so that would also suppress seasonality and help us in understanding the trend component, so this particular point is more related to what we are going to cover, a later in this module that is about smoothing methods, so the moving average plots there essentially part of that particular technique, smoothing methods so we'll discuss more of, more on this point in those lectures.

So what we are going to do is, till now we have discussed a number of points about how to identify time series components, so we started with time plot and how time plot can be examined to find out different components of the series, we discussed this dataset by cycle ridership and then some other points which could be used, for example zooming in specifically for a long series, then the next one that we discussed is changing a scale of Y axis, so that would also help us in terms of understanding the you know trend of the series, then the another thing adding trend lines so that would also help us in understanding trend of the series, so if we look at the points that we are discussing, we are mainly focusing on the trend component, so the trend component is the most important component in modeling series or forecasting a time series.

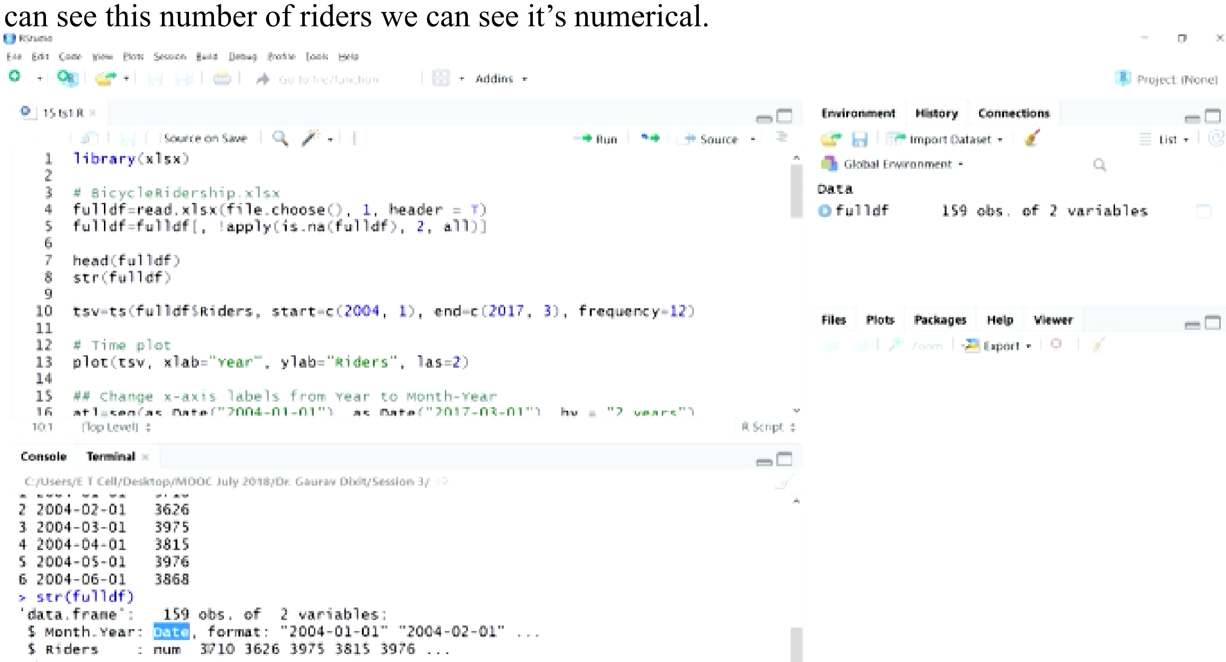
Then few more points we discussed about changing a scale of X axis and how that can suppress seasonality and help us in understanding the trend component, similarly moving average plots will also enable us in doing the same thing, so what we'll do is we'll open R studio and go through this dataset that we talked about bicycle ridership and apply some of these points and try to identify the time series components, so let's open R studio.



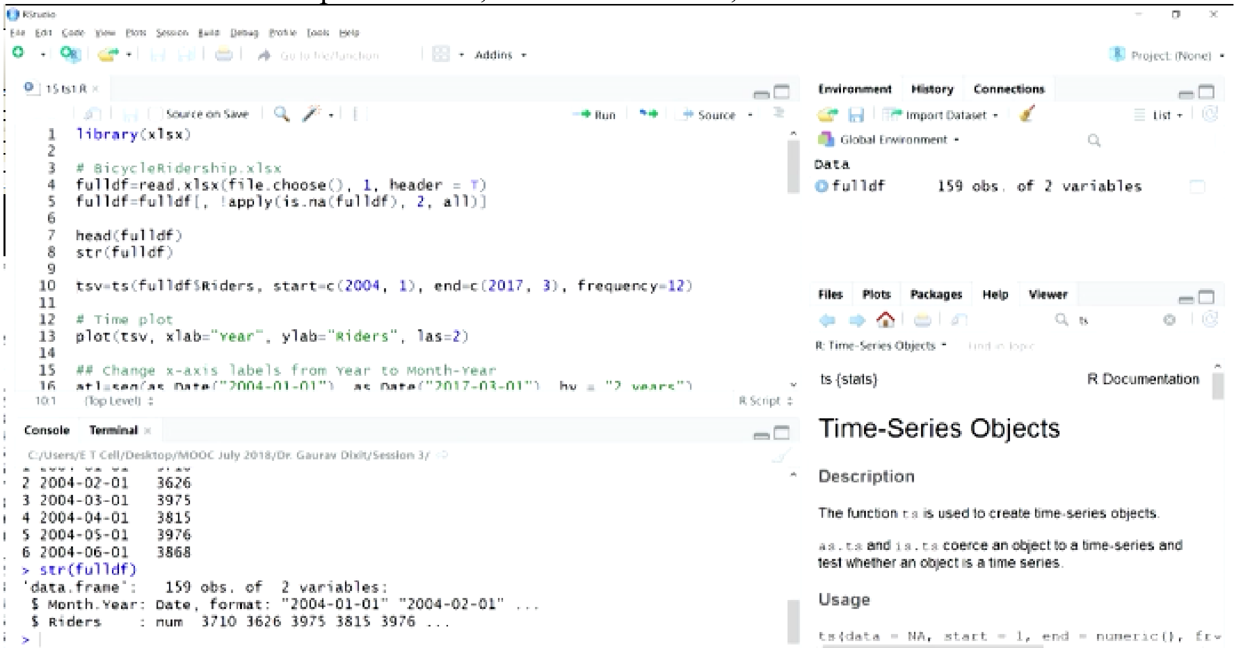
So let's load this library xlsx, so this is the dataset bicycleridership.xlsx is the file that we would like to import into R environment, so let's run this code. So this is the file, you would see that in environment section the file has been loaded and we can see full DF is the data frame, we have 159 observations of two variables, let's run this part of code also to eliminate any deleted columns.



Now we'll have, let's have a look at the first 6 values, so you can see the first variable is month, year which is essentially capturing the time specific information, the second variable is the main variable the time series variable which is capturing the riders, the number of riders and you know for different months of year, you can see in the first variable month, year, you can see the only variation is in different rows is the month value is changing from 1, 2, 3, 4 onwards, so these are first 6 observations and let's look at the structure of this particular time series data, so now in this we can see data frame 159 observations, 2 variables, month, year we can see it has been identified as a data you know date type of you know data structure in R environment, we can see this number of riders we can see it's numerical.

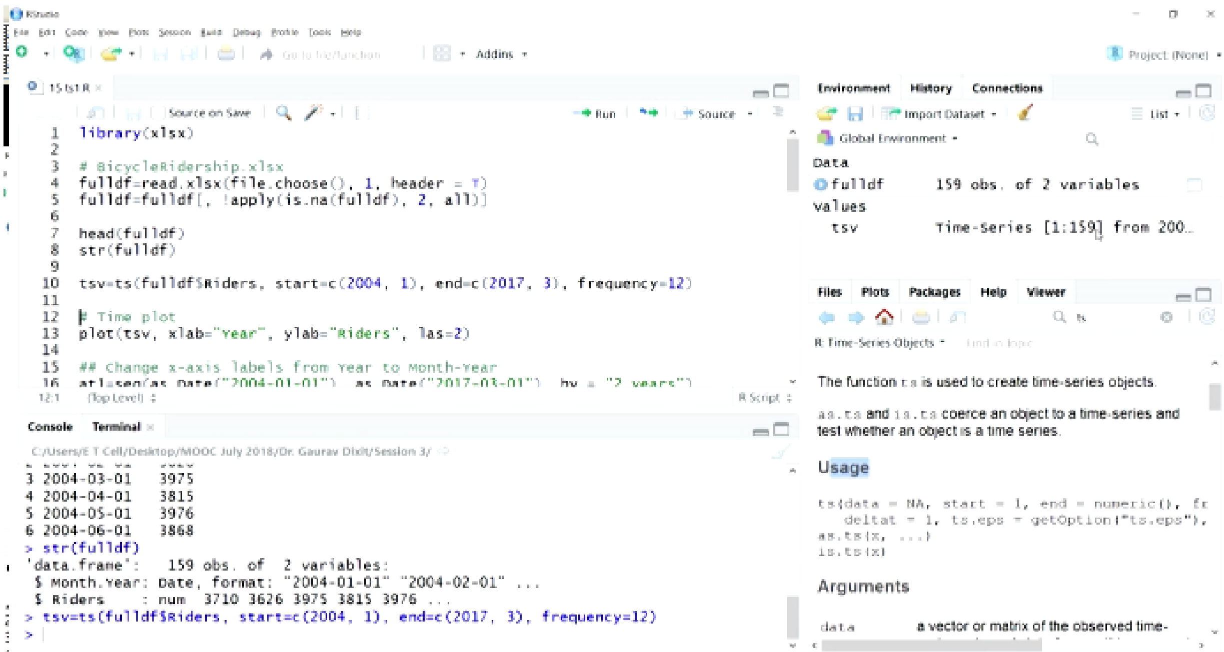


Now to create before we can go ahead and analyze the time series and use it for modeling purpose, we need to create a time series object, so for this we typically use this TS function, so if we are interested in more, if you are interested in finding more about this particular function so we can look at the help section TS, so we can find out, so we can see here time series



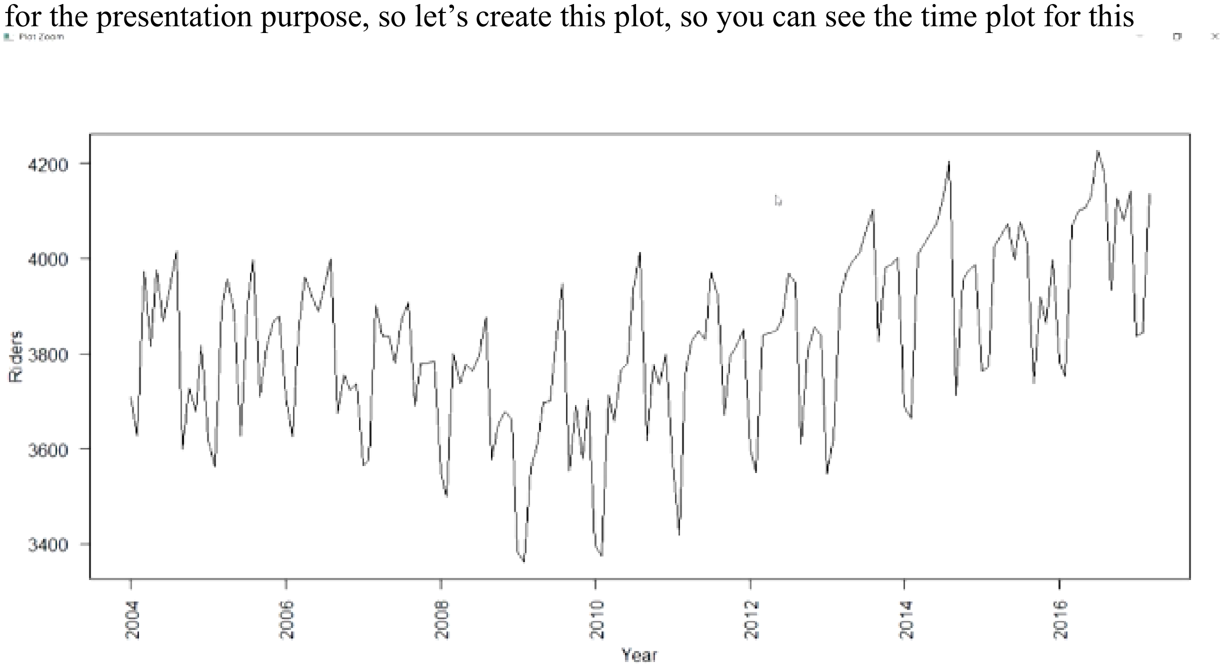
objects, the function TS is used to create time series objects, more detail about this function and different arguments that could be used, you can find out on these usage and arguments detail, so let's go back to our code, so here we can see I'm calling TS function and in the first argument itself I'm using this data frame that I have full DF and the variable riders.

Now this time series variable that is the number of riders you know that is going to be used to create this time series object, so in the second argument I'm also specifying the start and end time of the series, so and then I'm also specifying the frequency, so as you would see if the you know frequency is being specified then I typically don't need to specify the third, let's say third argument that is end of the series, however we can do you know, maybe we can you know use all three function because these are the, these are the arguments that typically we would be using while creating this time series objects in R studio environment, so start of the series is this we can see here 2004 January, so that is the start and we have data up to 2017 March, so let's create this object, so you would see a time series vector has been created, you can see here in

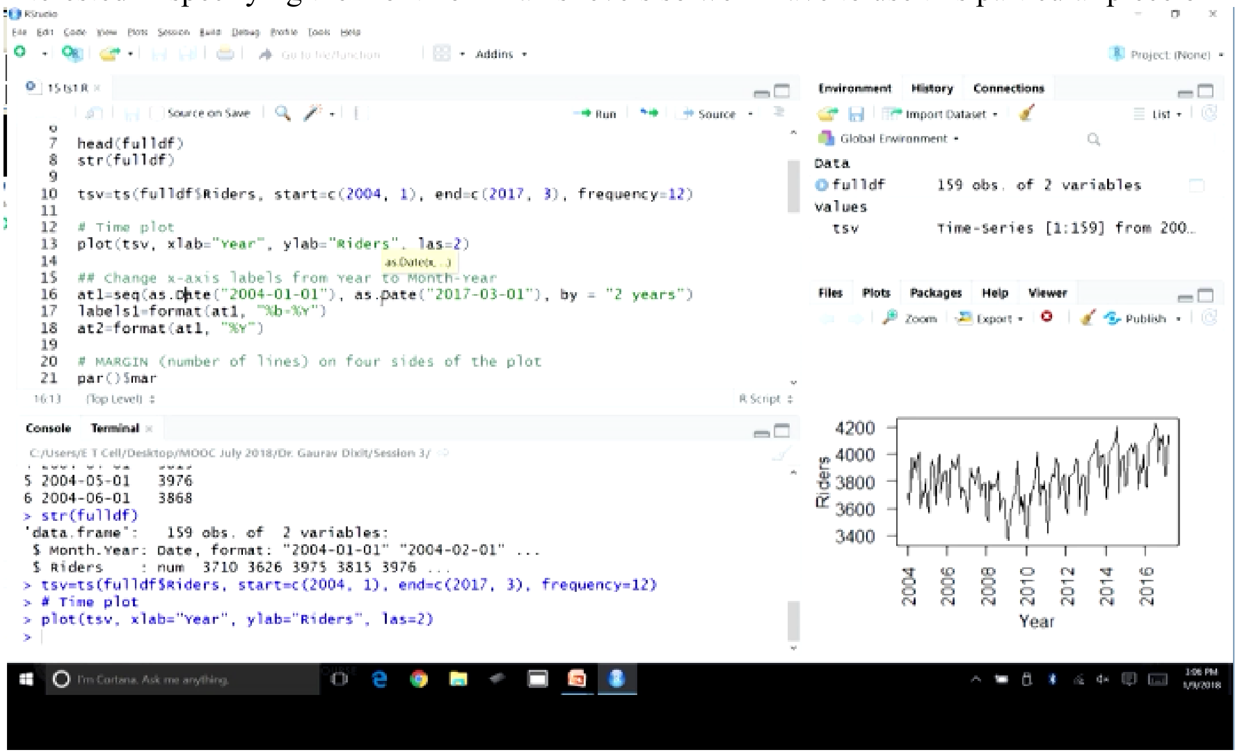


the environment section time series 1 to 59 and so these are the, this time series object has been created using the function.

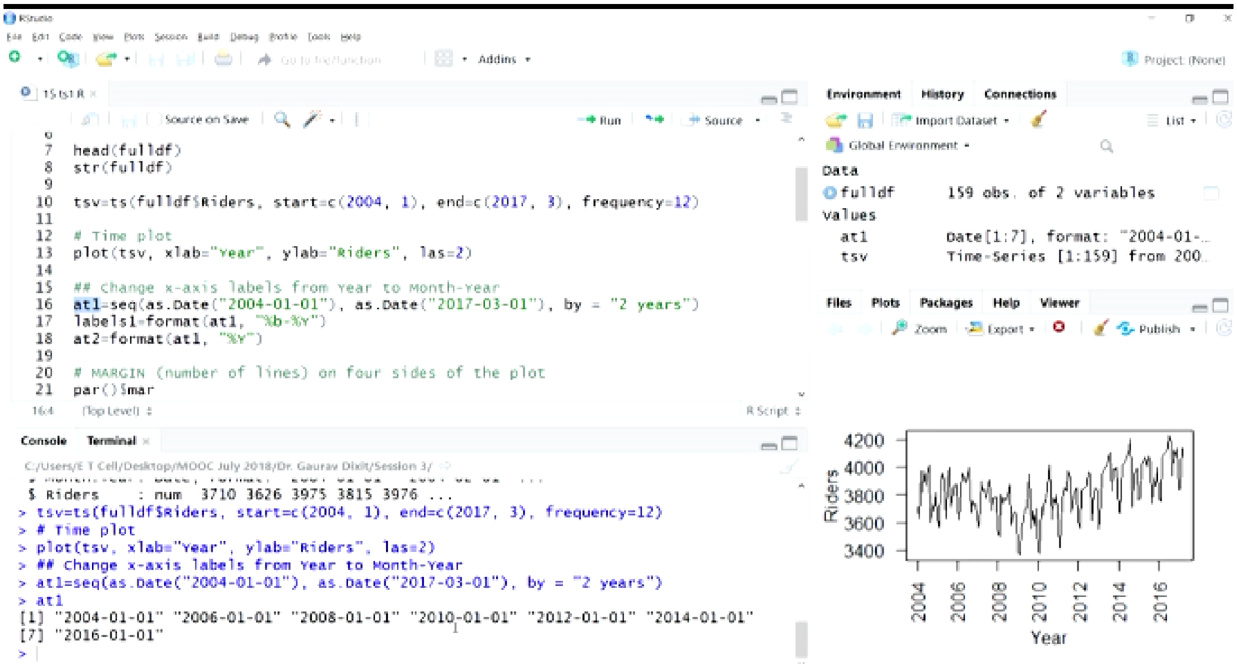
Now to create the time plot we need to call this function plot first argument is the time series object that we have just created TSV, the time series vector then the levels you can see right now year has been mentioned for X axis and riders has been mentioned for Y axis, however the level for X axis should can also be changed to month year because the data that we have in the series is monthly data, so monthly ridership data that we have. And then LAS is also mentioned for the presentation purpose, so let's create this plot, so you can see the time plot for this



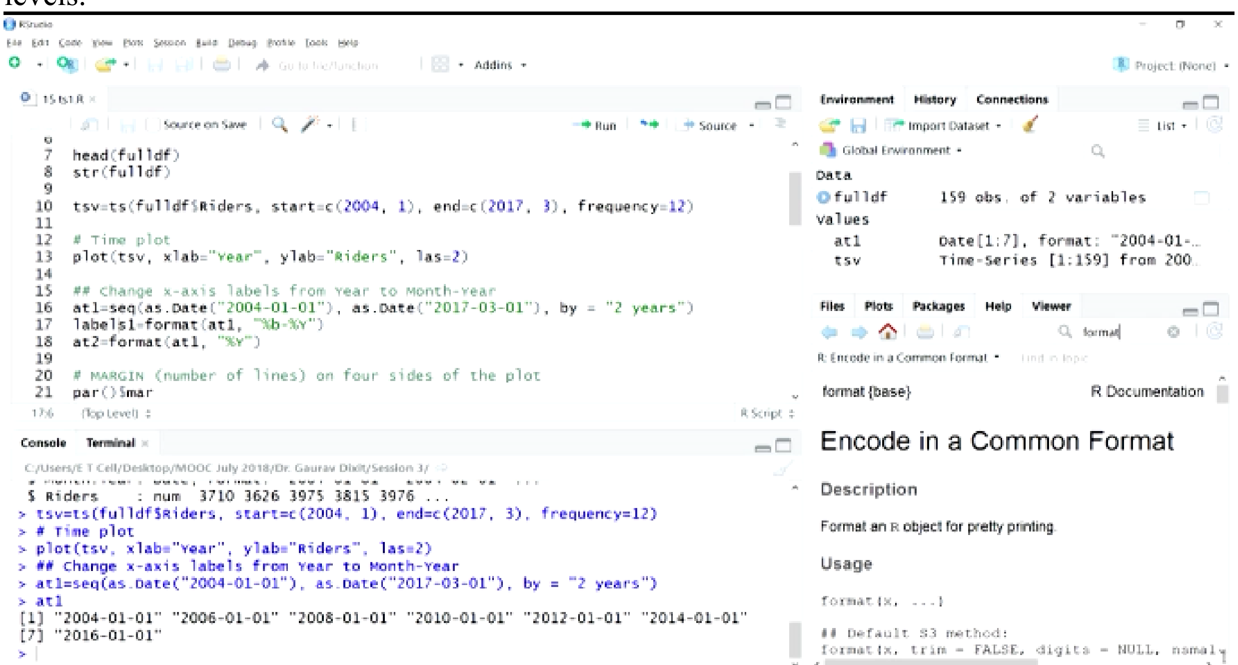
particular series here it has been created, so you can see we have data starting from 2004 up to 2017 as we saw in the code the end of the series, so you can ridership of this particular you know time series has been changing month by month, so we can see here or this has been changing, so from this looking at this particular plot, this particular time plot itself we should be able to identify few of the components of the time series, but before we go ahead and analyze this time plot we can do certain transformation in the plot itself to reflect more detailed information, so for example we would like to change the levels on X axis, so for that you know because if we want to specify the month as well, because this is monthly data so if we are interested in specifying the month on X axis levels so we'll have to use this particular piece of



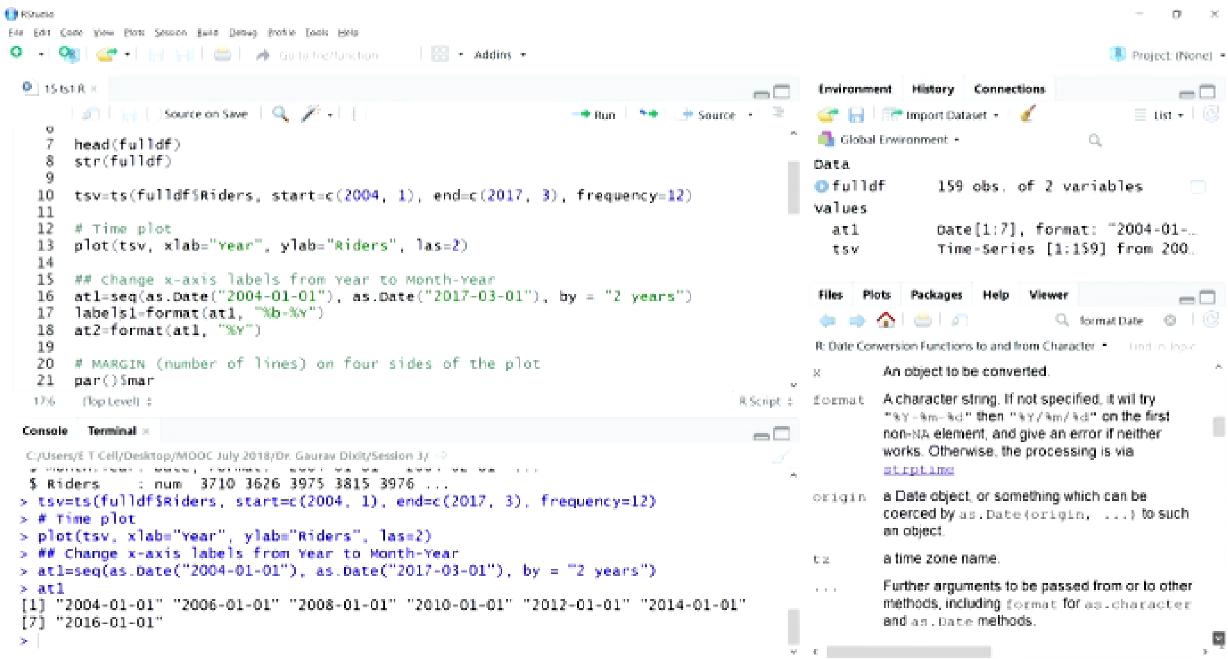
code, here you can see first time creating a sequence, so you can see changing you know the command here, change X axis levels from year to month here, so here first I need to create the sequence as.date and the starting date I have mentioned here, then the ending, end date of the series I've mentioned also, and then by two years, so because those are the tick marks that I might want here in the plot, so you know by you know 2 years I would like to have this sequence and month name to be depicted on the plot, so let's run this, we'll have the sequence.
Now once this has been created you can see at 1 this is the date object having 7 values here as you can see, so from that what we can do is we can format this particular date object and get the levels so from this date information we are interested in finding the, we are interested in extracting the information of month and year, so in that sense if you are interested in understanding this function a bit more how we are doing it, so at 1 as I said the sequence 2004-01-01 to up to 2017-03-01 having two years interval has been created, so more if you are interested in looking at these values we can run this and you can look at the values at 1 you can



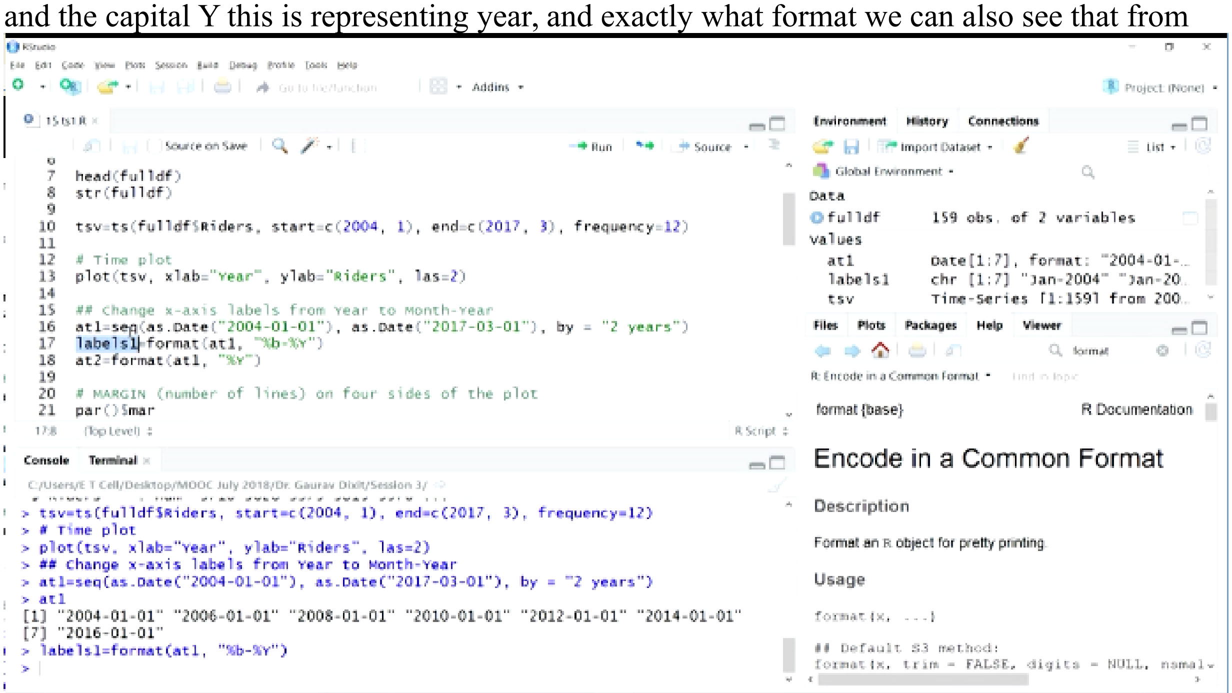
see 2004-01-01,2006-01-01, 2008, so this is the gap is 2 years, and we can see this particular sequence 7 values are here, now we are going to format this because we want to create X axis levels.



Now we can look at the, if you are interested in more detail information on this particular you know function you can look at here, format, so you'll see that format this particular function we can look at this particular function allows us understanding more about, more about the, understanding more about how we are going to use, how we are going to format these levels so



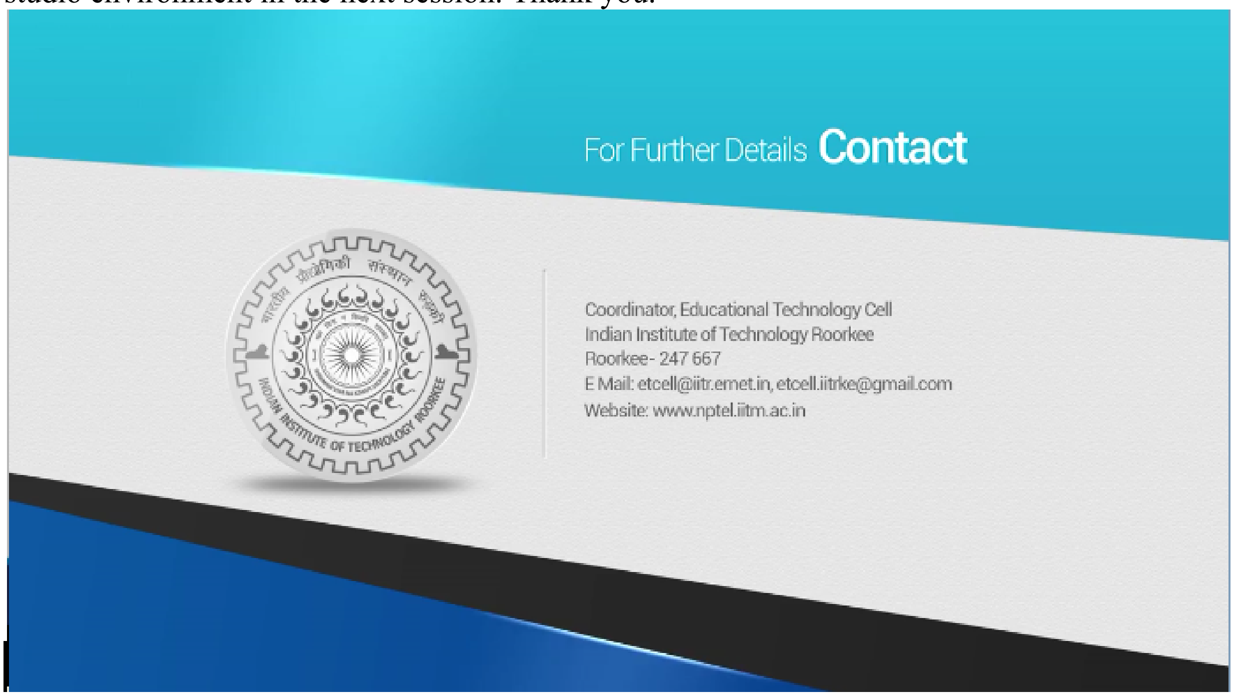
format.date is the specific function that we might be interested in, so you can see here, format X and other arguments, so here we can find out different levels that could be used, so here we will see that in this, let's go back to format, so we are, I just want to show you this, these formatting option that I have picked here, I have picked from this function, so I think the details would be here down in this particular function, so you can find out the detail of these formats, so what these different formats represent actually, so in this we'll see that either in the format or on the format.date you would see that this percentage B and percentage Y, so B is representing month, and the capital Y this is representing year, and exactly what format we can also see that from



the output, so what I'll do I'll run this particular part of code and we'll see this, so labels I have just created this particular variable you can see, this is character vector and you can see in this format we'll get, so a smaller b that actually represented the month in this form, and then we had Y which represented year, so now we have the labels in month and year format, now we can use this.

Now to identify the tick marks also, so if we go back to the plotting so we need to identify these tick marks where the labeling are to be plotted and the labels name, so what we have just created in the labels one, we have created the labels for us, now we need, we also need to compute these tick marks where these labels are to be plotted, so at 2 will actually allow us that, so let's run this.

Now we'll stop at this point and we'll continue our discussion on time series forecasting and specifically this part of the code where we are looking to analyze time series components in R studio environment in the next session. Thank you.



For Further Details Contact
Coordinator Educational Technology Cell
Indian Institute of Technology Roorkee
Roorkee – 247 667
E Mail:-etcell@iitr.ernet.in, iitrke@gmail.com
Website: www.nptel.iitm.ac.in

Acknowledgement
Prof. Ajit Kumar Chaturvedi
Director, IIT Roorkee
NPTEL Coordinator
IIT Roorkee
Prof. B. K Gandhi
Subject Expert
Dr. Gaurav Dixit
Department of Management Studies
IIT Roorkee
Produced by

Mohan Raj.S
Graphics
Binoy V.P
Web Team
Dr. NibeditaBisoyi
Neetesh Kumar
Jitender Kumar
Vivek Kumar
Dharamveer Singh
Gaurav Kumar
An educational Technology cell
IIT Roorkee Production