

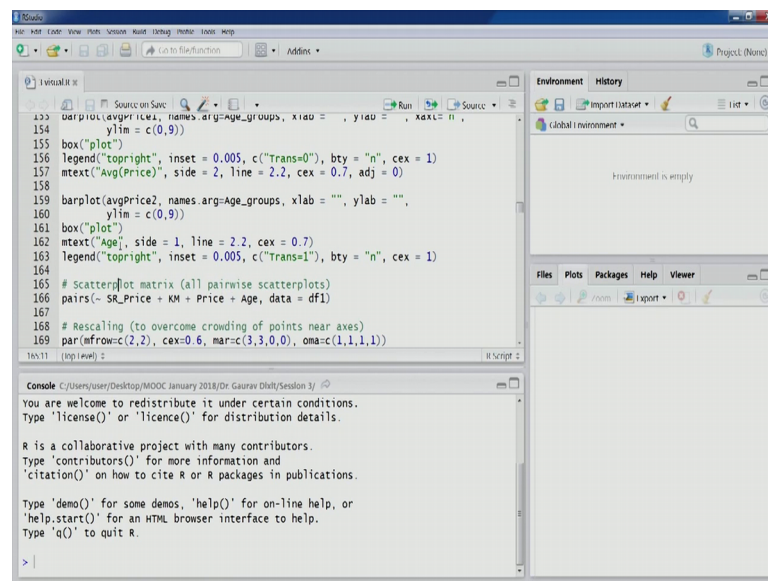
Business Analytics & Data Mining Modeling Using R
Dr. Gaurav Dixit
Department of Management Studies
Indian Institute of Technology, Roorkee

Lecture – 10
Visualization Techniques- Part IV
Multiple Panel Plotting

Welcome to the course Business Analytics and Data Mining Modelling Using R. So, in the previous lecture, we were discussing visualization techniques and we were in particular we were discussing a multiple panels. So, let us go back and restart our discussion from the same points let us go back to our studio.

So, last time.

(Refer Slide Time: 00:43)



```
153 barplot(avgPrice2, names.arg=Age_groups, xlab = "", ylab = "",  
154       ylim = c(0,9))  
155 box("plot")  
156 legend("topright", inset = 0.005, c("Trans=0"), bty = "n", cex = 1)  
157 mtext("Avg(Price)", side = 2, line = 2.2, cex = 0.7, adj = 0)  
158  
159 barplot(avgPrice2, names.arg=Age_groups, xlab = "", ylab = "",  
160       ylim = c(0,9))  
161 box("plot")  
162 mtext("Age", side = 1, line = 2.2, cex = 0.7)  
163 legend("topright", inset = 0.005, c("Trans=1"), bty = "n", cex = 1)  
164  
165 # Scatterplot matrix (all pairwise scatterplots)  
166 pairs(~ SR_Price + KM + Price + Age, data = df1)  
167  
168 # Rescaling (to overcome crowding of points near axes)  
169 par(mfrow=c(2,2), cex=0.6, mar=c(3,3,0,0), oma=c(1,1,1,1))
```

Console

C:\Users\user\Desktop\MOOC January 2018\Dr Gaurav Dixit\Session 3\>
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

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

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

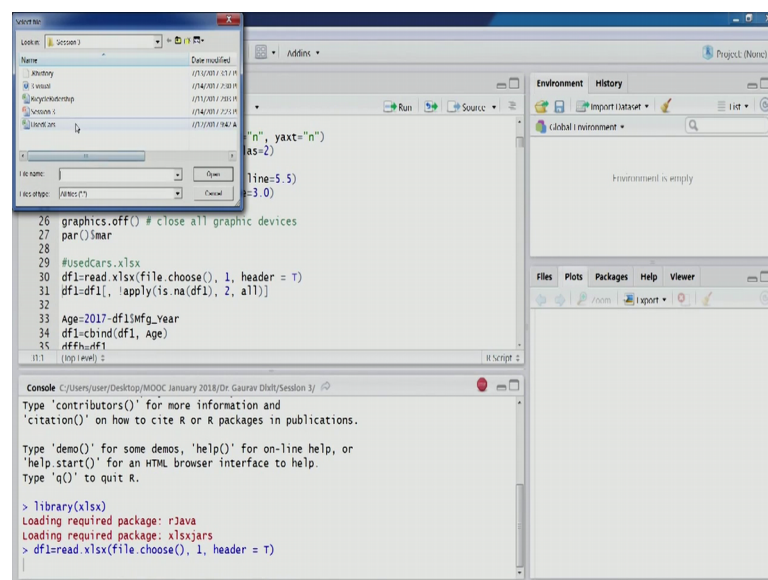
So, at the end of the lecture we were we were trying to cover separate panel for each groups. So, that I think we were able to complete and now today in this lecture let us move to scatter plot matrix. So, scatter plot matrix can be really useful in situations where you have a many numerical variables and you are trying to understand in different relationship between different pairs of a variables that could that is going to be useful for prediction task and supervised learning supervised learning prediction and classification task in supervised learning methods.

And in and in case of unsupervised learning methods it could be useful in understanding the information overlapped between 2 variables. So, if we get to see in one go and the relationship between different variables our visual perception can be much better specially in some situations.

So, the data set that we are going to use is the same one the users curve data set. So, again because we are starting a fresh so we need to import this particular data set again. So, let us do this let us reload this library. So, let us import the this this particular data set.

You can see the data set has been imported 79 observations of 11 variables that is visible in the environment section.

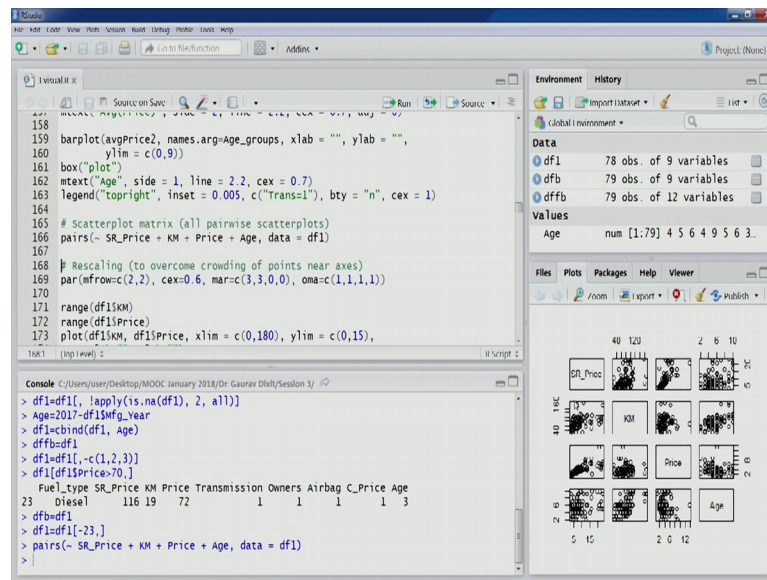
(Refer Slide Time: 02:18)



Now, let us also compute the age variables some of the things that let us also take full backup of this particular data frame that we are going to require later on the in this lecture. And first 3 observation as we understood in the previous lectures might not be were not important for some of the initial visualization techniques that we discussed once this is done.

In the previous sessions, we also identified one observation that we wanted to get rid of.

(Refer Slide Time: 03:07)

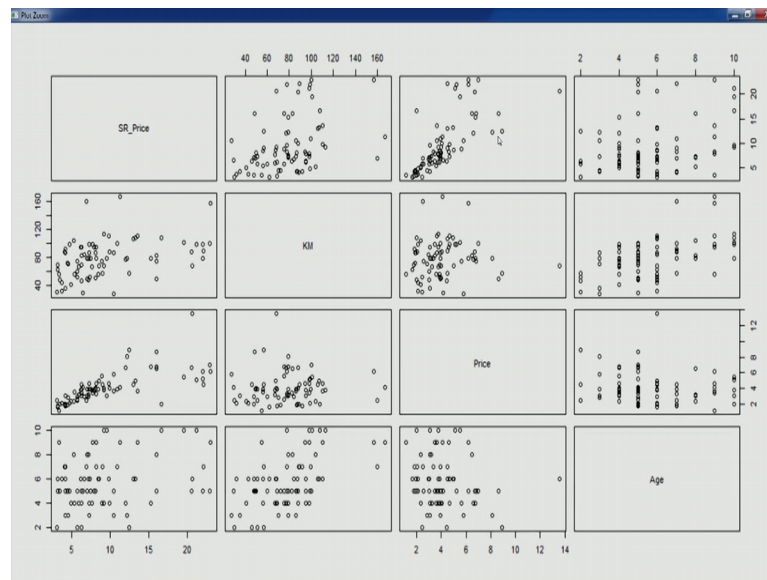


So, let us do the same today's well backup in eliminating the observation, now we can now we are ready to start with scatter plot matrix. So, let us go back to a scatter plot matrix yes. So, as discussed is scatter plot matrix can be useful to understand the relationships and information overlap.

So, now you can see we have a selected a 4 key a numerical variables continuous variable for scatter plot matrix as we understood in the previous lecture, that the for scatter plots both the a variables a variables that are going to be on x axis and y axis are supposed to be numerical variable. So, therefore, you can see S R Price kilometre and Price and age these 4 numerical variables can be used to create a to generate a scatter plot matrix. So, let us execute this code.

You can see the graphic here has been converted created let us zoom in.

(Refer Slide Time: 04:21)



Now, here you can see the you can see in the diagonal in the in the in the diagonal rectangles right in the diagonal boxes, you would see the minimum the variables that have been used to create this particular scatter plot matrix S R Price K M Price and Age you can the function that we have used to a generate this particular matrix is pairs. So, in this pairs function we have to pass on this formula in the formula you have to mention the name of the variables you know which are going to be used to create this scatter plot and then the data.

So, let us go back. So, for example, if you are interest interested in this particular this particular graphic then the y axis is going to be S R Price which is in the same row S R Price and the x axis is going to be represented by K M which is in the same column. So, x axis is K M and the y axis is S R Price. Similarly if you are interested in this particular graph then we would see that age is age variable, which is in the same row is going to be represented on the y axis and S R Price is going to represent the x axis while it is in the same column. So, that is how you can understand which variables are there in x axis and y axis.

Now, you can look at different plots and you can try to understand the relationship between them for example, this particular plot you can see this is between S R Price and price. So, you can see a linear kind of relationship is visible there see more majority of the points if you pass a line through the majority of points it is going to be a linear line

all right it is going to be a linear line. So, therefore, the relationship now you can understand from the variable variables itself that S R Price and Price both are based on prices therefore, there is going there is supposed to be a linear relationship. So, that is very visible in the data itself.

Now, if you if you are interested in kilometre and versus Price you can see this particular plot. So, K M and Price you can see most of the points they are clubbed here in this particular group. So, there does not seem to be much difference of you know you know much difference of Price on kilometre, we can also look at this particular graph in this case Price is on a y axis and K M is on x axis because Price is ever outcome variable of interest this is this particular plot is of more interest to us. So, here you can see the you know K M could be represented by this for particular data could be represented by horizontal line that actually you know signifying that there is not much of influence of K M on in determining Price similarly there are different plots and different kinds of relationship can be seen over there.

Now, if we if for example, if we are interested in some few other plots for example, Price and let say Price and age and this particular graph you would see that because the age is being represented by few numbers only. So, therefore, you would see for particular is cars of different prices are depicted here similarly also different. So, it is look like a bar chart bar chart kind of plot.

So, for different age group for different age numbers cars of different prices are being shown there by different point's data points. So, this particular these this particular scatter plot this kind of a scatter plot matrix can really be useful in terms of finding many relationship understanding many relationship, it can help us in finding a new variables it can also helps us in understanding the interaction terms if required it can also help us in grouping some of the categories it can also help us in you know sub setting the model. So, running a model on a on a subset of the full data set. So, those kind of things can actually be identified using these plots.

Now, let us move to our a next point let us go back to the slide.

(Refer Slide Time: 09:14)

The slide is titled "VISUALIZATION TECHNIQUES" in a bold, dark blue font. Below the title, there are two main bullet points. The first is "Multidimensional Visualization", which has several sub-bullets: "Multiple panels", "Color", "Size and shape", "Animation", "Aggregation, rescaling, and Interactivity", and "Main idea is to help build visual perception to support the subsequent analysis". The second main bullet point is "Open RStudio". At the bottom of the slide, there are logos for "IIT ROORKEE" and "NPTEL ONLINE CERTIFICATION COURSE", along with the slide number "11".

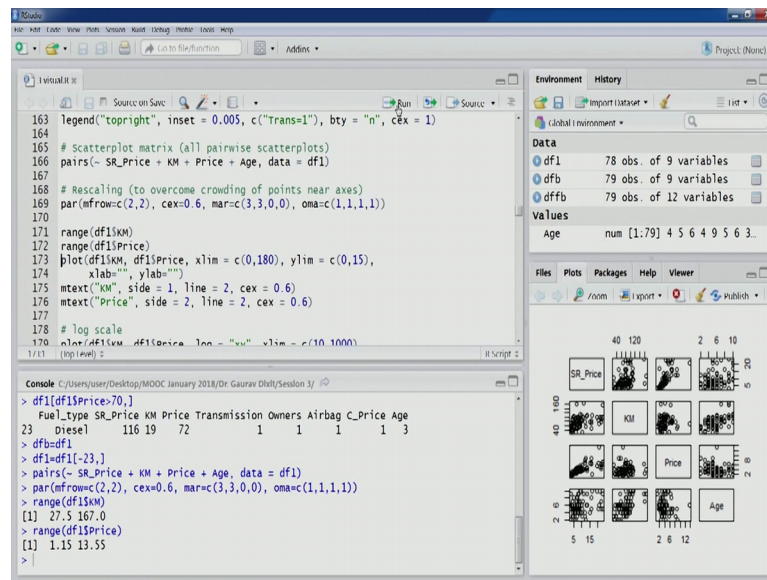
- **Visualization Techniques**
 - Multiple panels
 - Color
 - Size and shape
 - Animation
 - Aggregation, rescaling, and Interactivity
 - Main idea is to help build visual perception to support the subsequent analysis
- Open RStudio

So, next we are going to discuss is discuss these operations aggregation rescaling and interactivity. So, these operations can sometimes be really useful for the same things that we have been talking about. So, let us start with rescaling. So, let us go back to r studio again. So, rescaling can be really helpful we if there are crowding of crowding of points near axis in a near axis whether it is x axis or y axis if there are many points which are crowded near those axis near those axis.

So, therefore, we can do rescaling of x axis and y axis and get a better look of the data. So, how let see through an example. So, in this so now we are going to create 4 plots 4 back to back plots. So, therefore, we are trying to we are trying to divide our plotting area plotting region into 2 rows and 2 columns. So, 4 plots are going to be created and we have appropriately we have changed other settings like margin outer margin and this size of font and size of different text and numbers.

So, let us run this. So, first particular rescaling that first particular example is between a for is scatter plot between kilometre and price. So, let us again have a look at the range.

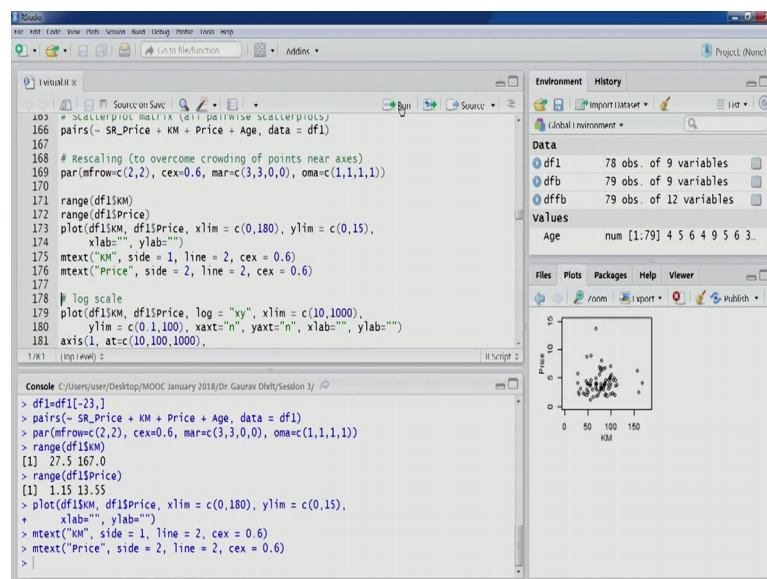
(Refer Slide Time: 10:55)



You can as you already know that range can be use to a specify the x axis x limit and the y limit in the plot function. So, you can see the appropriately the limits have been specified.

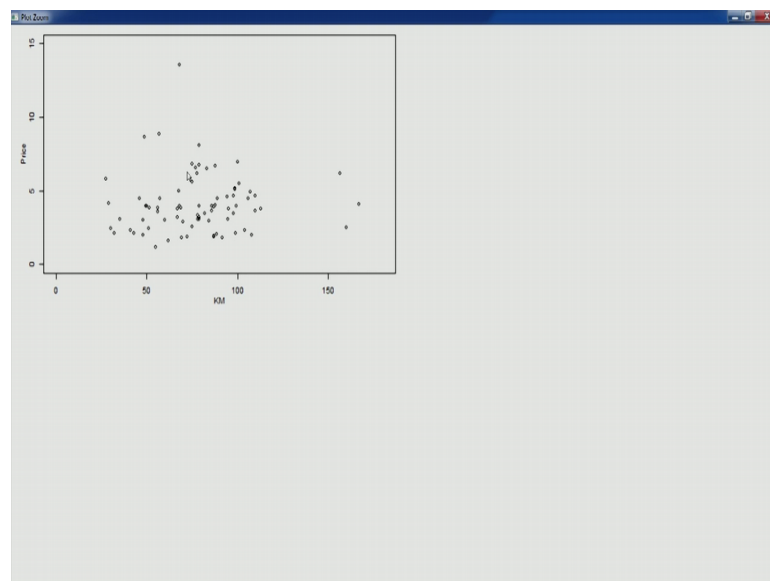
Now, let us run this particular plot you would see because as I said we wanted to you know we wanted to generate 4 plots in the same you know plot area.

(Refer Slide Time: 11:12)



So, therefore, you can see one fourth of the one fourth of the area has been taken by the first plot. Now let us create the axis labels we can see Price verses kilometre and you can also see the points. Now if we want to zoom into this particular plot though we have talked about this plot many times that there is not much influence of K M on price, but if you want to have a much closer look if you want to have much closer look then scaling can be really useful.

(Refer Slide Time: 11:42)



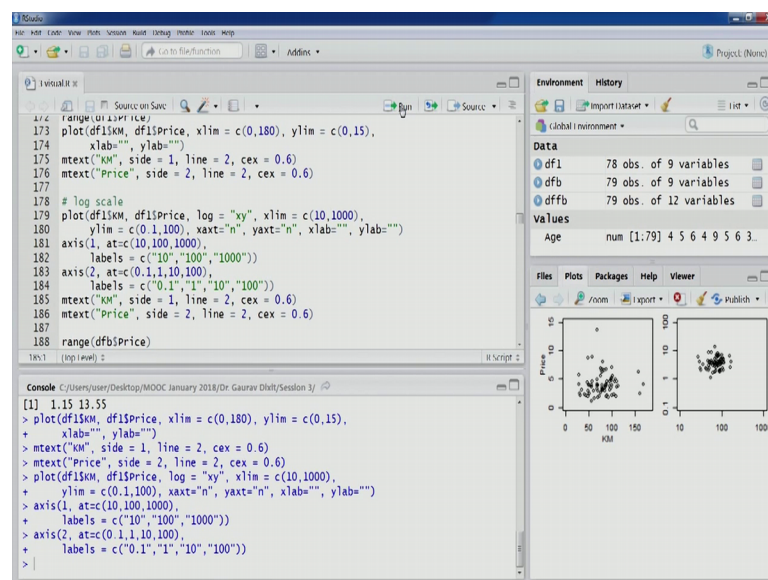
How we will see. So, let us change this scale of x axis and y axis into log scale. So, when we talk about this log scaling we are essentially changing the spacing of points on x axis and y axis. So, points are not going to be equally spaced in x axis and y axis they are going to follow the logarithmic log base is you know scaling and is spacing would be accordingly changed. So, how that can be done so in the plot function in r there is this argument log which can be actually which can be used to do the perform different kinds of scaling for example, if you if you just want to change the scaling for x axis. So, we can say log we can assign log as x and then if you if you just want to scale y axis then y can be a log can be assigned the y axis if you want to change both the axis then that that is the case that we are doing here right now. So, we have to say x y. So, this is between K M and price.

Now, let us talk about the limits now limits as you as in the previous plot we had used 0 to 180 for x axis. Now in case of log in case of log you would see that this is 10 to 10 10

to 1000 reason being that is the reason being that 180 is more than 100 and the next you know spacing point appropriately spacing point in a log is scale is going to be 1000. So, it is going to be like 1 10 100 and 1000 or in the other direction 1.1.0.01 in that sense.

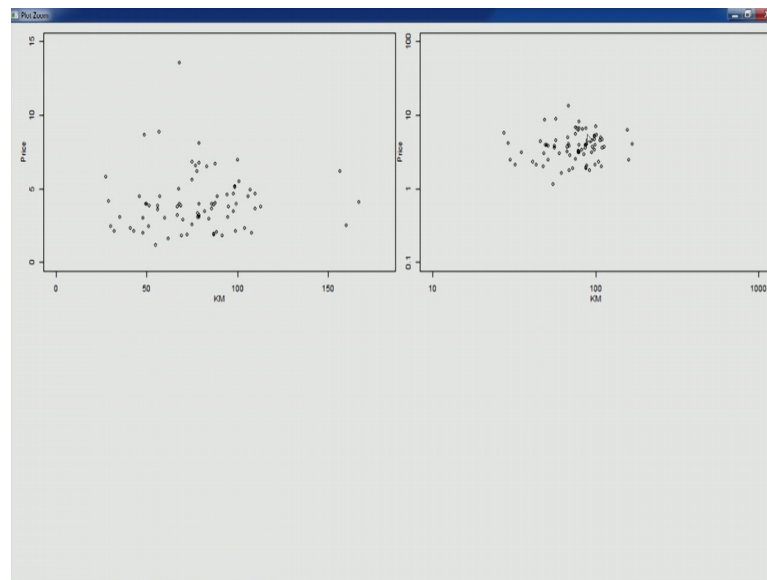
So, therefore, we have to make sure that the all the values are within the range. So, appropriate limit for the for the values to lie in the in the in the plot region this could be appropriate 10 to 1000 for 0 and 180. Similarly 0 to 15 we can have 0.1 2 100. Let us execute this particular line you can see now the visibility of all these points is much clear and this is mainly because of the spacing change in a scale and they are and thereby change in spacing of points in x axis and y axis.

(Refer Slide Time: 14:31)



Now, we are trying to recreate both the axis x axis and y axis and we are also trying to re label these axis now let us zoom in.

(Refer Slide Time: 14:50)



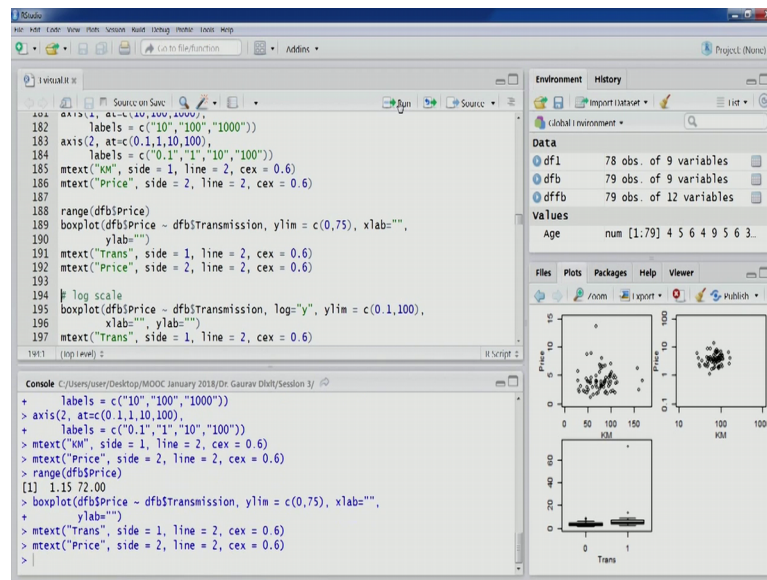
So, now you can compare these 2 plots. So, here the that a that horizontal line that can that can represent this particular these particular data points is not that much usually, you know Perceval in in plot 1, but in plot 2 pretty much you can see that this seems to be a that kind of relationship there is not much this is a horizontal line. So, therefore, there is not a much influence of K M on price.

So, this is much better visible in this log scale you can see the points this 10 100 and a 1000 you can see the most of the points they were in this range right around 100 20 around from 25 to 100 20. You can see the points are still lying in the in the same range, but because of the change in the in this is scale and therefore, is spacing of points the visibility of the of these data points have changed and therefore, we can more easily perceive the relationships.

So, let us go back and now we are going to create a box plot and try to understand how the scaling rescaling can actually be really it can be helpful in case of a box plot. So, this is this in this for this example we are using this data frame which we had taken backup. So, in this in why we are taking this particular data frame we had eliminated the one particular out lier point in the data frame 1 d f 1 now we wanted back so that the importance of rescaling could be emphasized in much better manner.

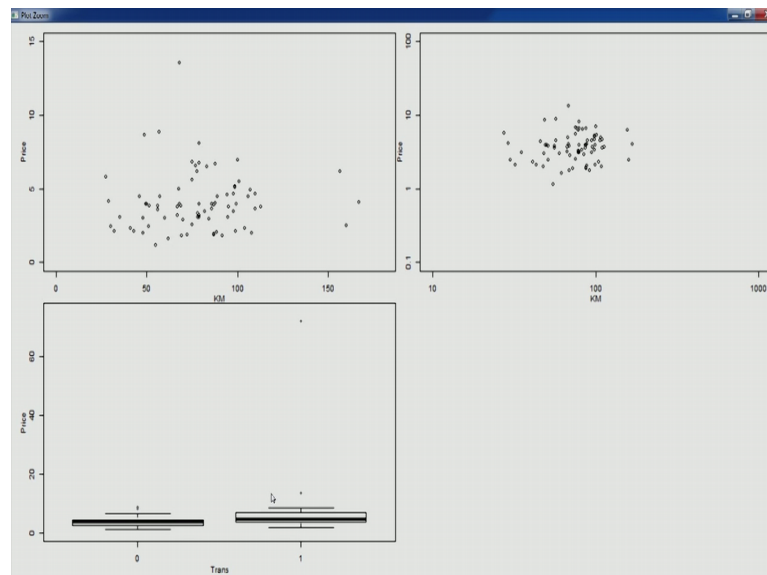
We will see that.

(Refer Slide Time: 16:45)



So, range of this you can see this point is back 72 now we have to change our limits you can see that. So, this particular box plot is between Price and transmission transmission being on the x axis being the this being the categorical variable. So, let us plot this this is the plot let us label the axis now let us zoom in let us have a look at the plot.

(Refer Slide Time: 17:14)

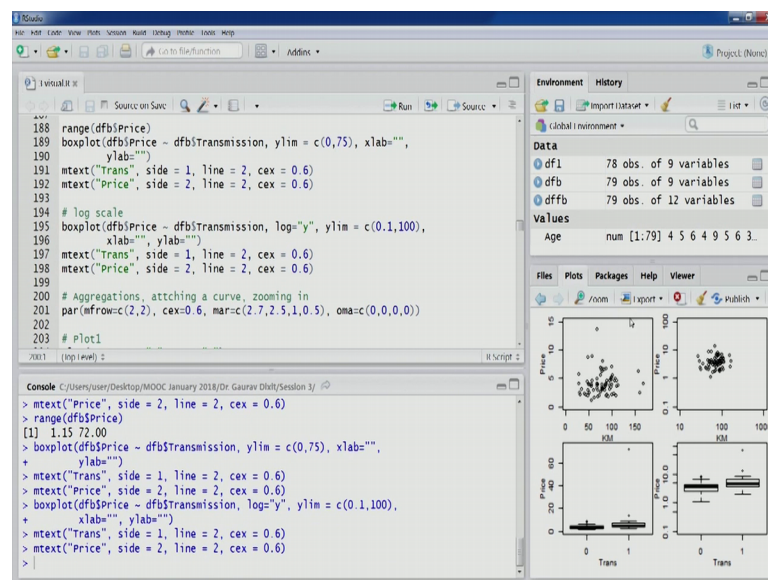


Now, you would see the plot is in this particular case because of this particular observation you can see one observation lying here, because of this and this observation

in the automatic transmission category because of this observation most of the whole box has crowded into x axis.

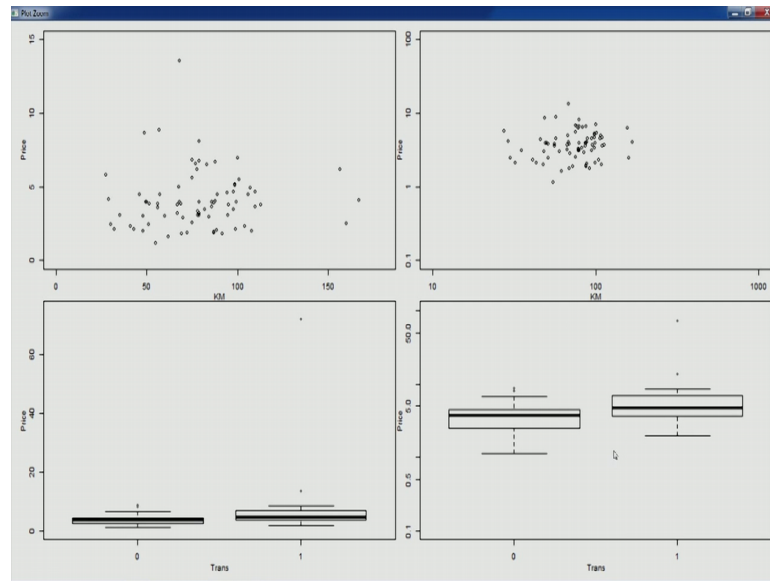
So, therefore, comparison of these 2 boxes is becoming a very very difficult. So, therefore, rescaling can really be helpful in this particular case. So, what we are going to do is we will use the log scale.

(Refer Slide Time: 17:50)



See the log y has been selected because now only the y is the a numerical variable. So, therefore, only there rescaling is rescaling is required and y limit have been changed appropriately. So, that it covers all the data points you can see 0 75 is very well very well within this particular range and let us execute this code you would see a plot has been created let us label the axis.

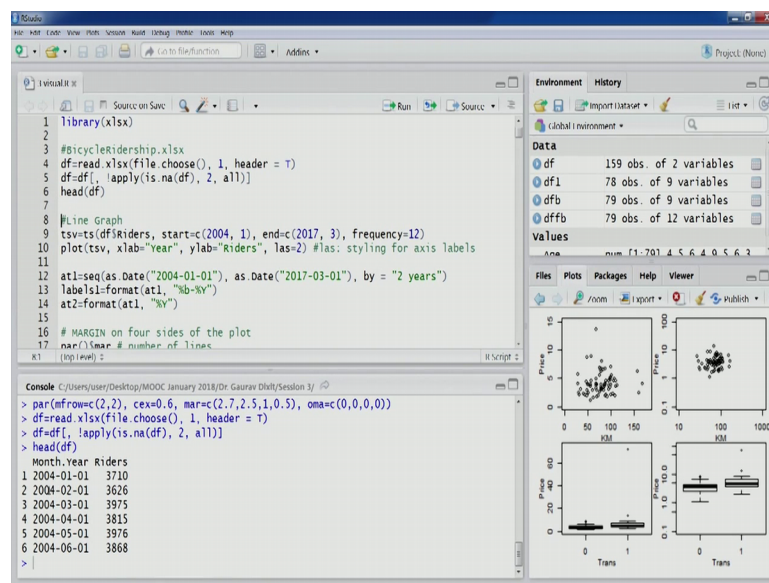
(Refer Slide Time: 18:20)



Now, you see the boxes are looking much better and the comparison could be easily performed. Now the point which was outlier now you can see the spacing between this particular point and the a box plot the main box plot has changed to a great extent and this is the result of scaling now comparison could be easily done.

So, this could be the benefit of rescaling in some situations where crowding happens. So, next discussion point is on aggregations attaching a curve and zooming in. So, we will go through some of the examples and we will see how aggregations and curve how we can append a curve or add a curve to the existing plot and how we can zoom in and how it can actually help us in doing some of the visual analysis task.

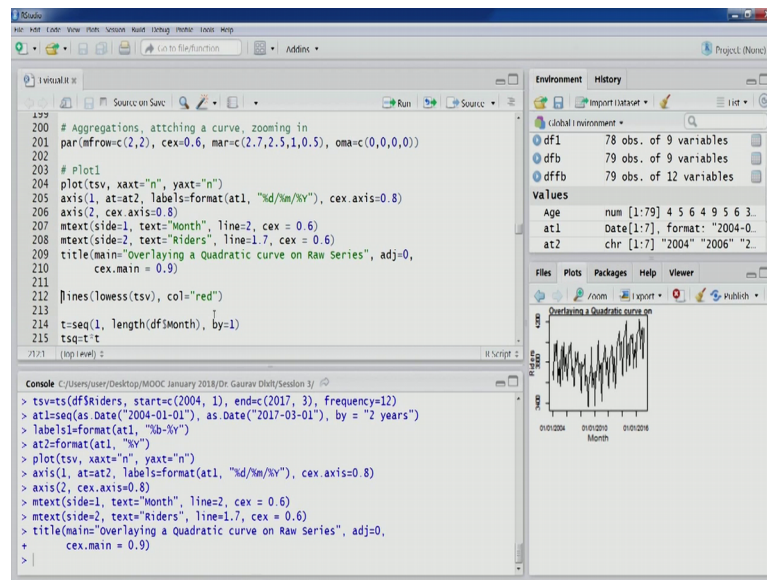
(Refer Slide Time: 19:14)



So, let us. So, again we are going to create 4 back to back plot. So, the similar parameter the `par` function has been specified and called appropriately you can see 2 and 2 rows 2 columns and a margins have also been specified. So, that we are able to use the plotting region effectively. Now our first plot is now for the in this particular case we are going to use the time series data that riders data that, we had earlier used. So, let us import that particular data set bicycle ridership dot xlsx you can this particular data set is the time series data the first variable being from month and year the time scale related information and the number of riders for every month and this covering years from 2004 to 2017.

Let us also create this time series vector that we are going to require later on let us also create these variables at 1 labels at 2.

(Refer Slide Time: 20:52)

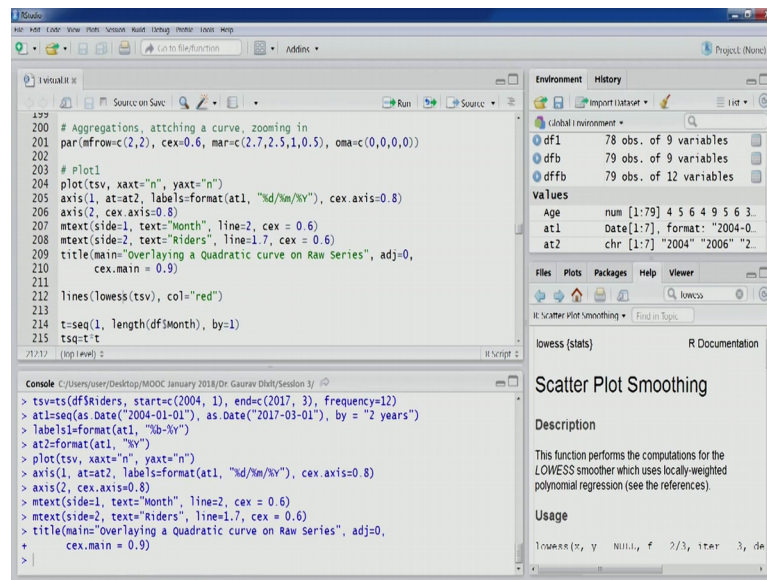


Now, let us go back to aggregations yes now once this we have created the time series vector now let us plot. So, this plotting we have done before you can see the time series has been plotted one fourth of the plotting region has been covered. So, because we are going to plot 4 back to back graphs. So, let us recreate the axis. So, x axis you can see aggregation the code that we are using that is you can understand the kind of labelling that we are doing right, now you can see a day month and year and the same is depicted in the plot. So, labels have been changed let us recreate the axis y axis this is mainly being done to accommodate to have the access and also to be able to accommodate 4 4 graphs in the plotting region.

Now, let us label the both the axis month and riders we are also going to provide a appropriate title, because we are in this particular in this particular example we are going to add a append or attach a curve on raw series. So, raw series is displayed now this is the title you can see a title has been displayed there now lines is one function that can actually be used to create a to create a curve.

So, therefore, if you are interested in finding more information on lines you can go into the help section you can see add connected line segments to a plot this is a generic function.

(Refer Slide Time: 22:47)



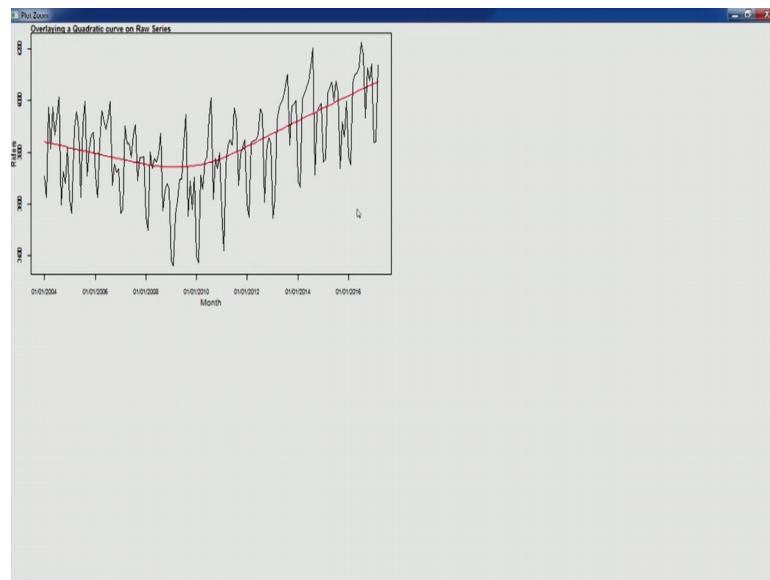
```
200 # Aggregations, attaching a curve, zooming in
201 par(mfrow=c(2,2), cex=0.6, mar=c(2.7,2.5,1,0.5), oma=c(0,0,0,0))
202
203 # Plot1
204 plot(tsv, xaxt="n", yaxt="n")
205 axis(1, at=at2, labels=format(at1, "%d/%m/%Y"), cex.axis=0.8)
206 axis(2, cex.axis=0.8)
207 mtext(side=1, text="Month", line=2, cex = 0.6)
208 mtext(side=2, text="Riders", line=1.7, cex = 0.6)
209 title(main="Overlaying a Quadratic curve on Raw Series", adj=0,
210       cex.main = 0.9)
211
212 lines(lowess(tsv), col="red")
213
214 t=seq(1, length(df$Month), by=1)
215 tsq=t^2
```

The screenshot shows the RStudio interface. The main editor displays R code for plotting a time series. The code includes data aggregation, plotting, and adding a LOWESS smoothing line. The console shows the execution of the code, including the creation of the time series object and the plotting commands. The environment pane on the right shows the variables in the workspace, including 'df1', 'dfb', 'dfdb', 'Age', 'at1', and 'at2'. The 'lowess (stats)' package documentation is also visible on the right.

So, you have you will have to provide the coordinates x and y. Now in this particular case we are providing the those coordinates using low lowest function the lowest function let us find out. So, lowest function is actually a smoother scatter plot smoothing function. So, it has actually takes the same x y points.

So, these points have are being taken from the time series vectors and the lowest is applying some computations related to our lowest smoother and we are going to get a smooth line added to the plot. The colour of the line has been this red we are going to use red colour for this particular line. So, let us execute this code you would see a line red curve has been added you can see that.

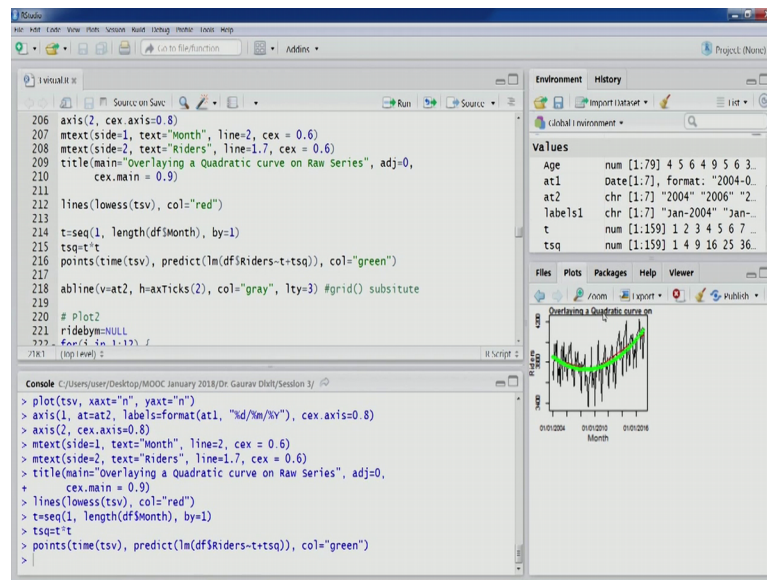
(Refer Slide Time: 23:43)



So, this particular curve has been created using the x and using the coordinated from the time series vectors if we want a more a curve, which is more representing of the data points you know we should be able to approximate what this particular line graph is looking like and then the representative curve we should be able to try and plot here. So, for example, this particular a line graph this particular time series is looking to follow a polynomial curve. So, probably a quadratic a curve can actually be over laid on this. So, let us do that. So, we will add a quadratic curve on this.

So, let us create this. So, t we need to create a because t will become a predictor for this particular quadratic curve. So, let us create t . So, you would see that t has been created in the you can see same in the in environment section. So, it is nothing, but series of numbers 1 2 3 4 depending on the number of points that are there and because this is being a quadratic let us create the t square. So, this has been completed now you can see that we are using a point's function which also does the similar kind of thing it plots points on a on depending on the coordinates.

(Refer Slide Time: 25:15)

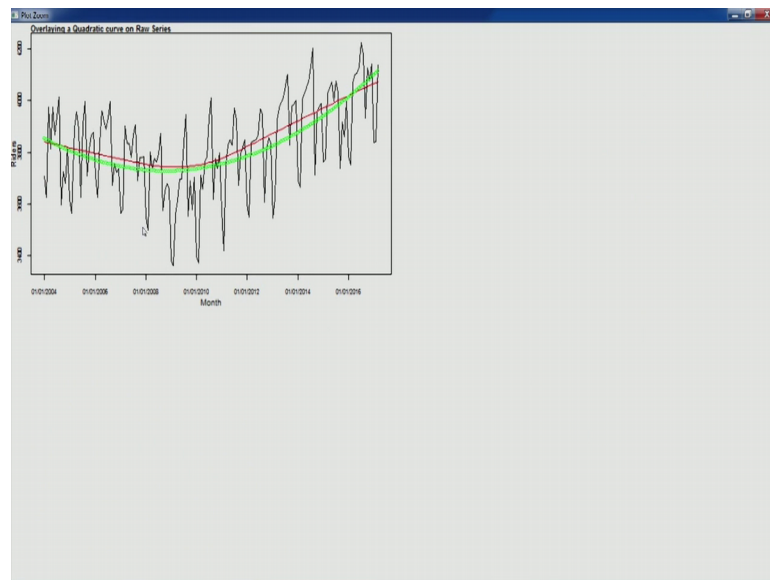


So, coordinates are being passed using the time function and they are using time function we are extracting these coordinates from the t s v time series vector and then predict. So, time is coming from t s v and we are using predict function that we have discussed before. So, we are using linear modelling. So, in these linear modelling 2 predictors 1 is a t and another 1 is t square they are have been model into riders.

So, this particular prediction they will actually give us the y coordinates and the time function that is extracting time from the t s v a vector and once both these points are there we are plotting them using the points function, colour is green for this particular curve.

So, let us execute this line you would see a green has been a green curve has been added.

(Refer Slide Time: 26:15)



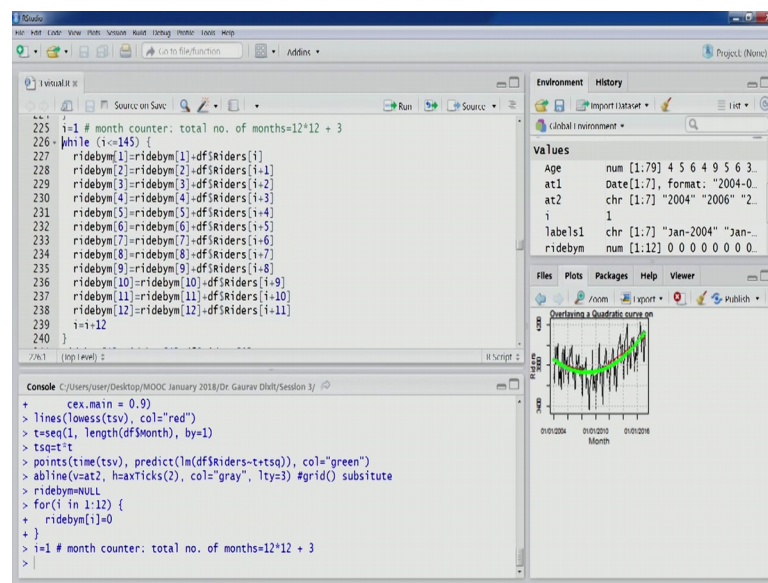
You can see. So, this seems to be much this a particular curve seems to be a representing the line graph in a much better fashion, but you can see the you can see the connected points see points have been plotted it is not the line previous one was the line. So, lines is the function that is used to create a line and the points with the function that is used to a plot points, but the because these being see the points we get a we get a sense of a curve being added now this is also following a polynomial a quadratic curve.

So, now let us also add the grid. So, a b line is the function which can be used to create a vertical and horizontal lines a x text function can be used to a get the default text that are generated in the plot that are that are generated using the plot function. So, this text can be extracted using extract function and we can have horizontal lines from coming out of those text and this has been generated at 2 that we used that we computed before has been used.

So, therefore, you can see that this grid lines these dot lines have been created appropriately. So, that we get a. So, that we are able to get a better look better look of the graph and we want to compare some of the values. So, we can easily understand them now let us come to our plot number 2. So, this plot number 2 is monthly average.

So, we are trying to take monthly average over the years.

(Refer Slide Time: 27:59)

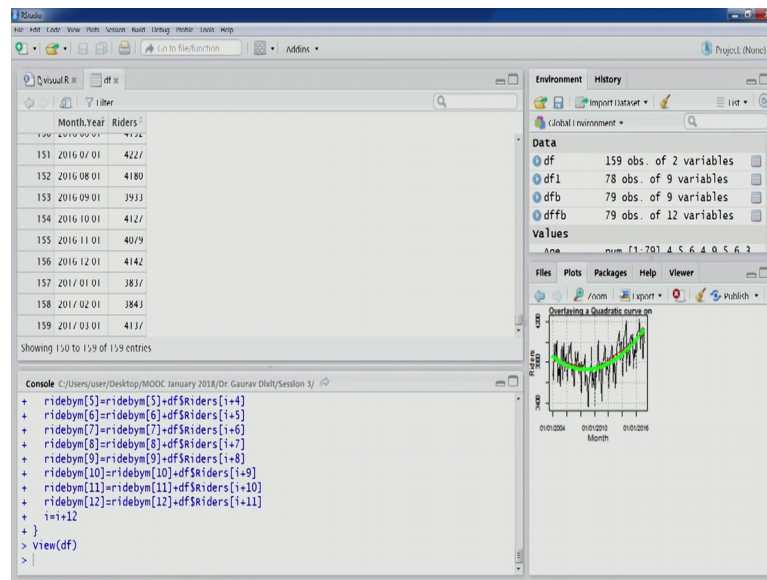


So, we have 13 years in total. So, for different months we want to we want to compute averages of average number riders for each month Jan Feb March a similarly for all 12 months and therefore, from those and then we want to plot them in a line graph. So, we will we would be required to compute those averages. So, let us do these so riders by month. So, this is the variable and 12 months are there. So, let us a first initialize all these 12 values 12 variables this is our counter.

So, we have total because we because we have 13 months and the and the 2017 we had only 3 months Jan Feb and March. So, therefore, we have this 144 plus 3 140 a 7 months in total. So, this being our counter I variable being our counter you can see in this while loop we are trying to accumulate all the riders' month wise.

So, a ride by m one is representing the month of Jan then 2 is representing the month of Feb similarly the last 1 a 12 ride by m 12 and the braces and the brackets is representing the a month of December and we are trying to accumulate all these numbers and then later on we are going to average them. So, that we get the average average a numbers average number of riders by month and later on we are going to plot them. So, let us run this now there are 3 more months in the year 2017 that we saw in the data before.

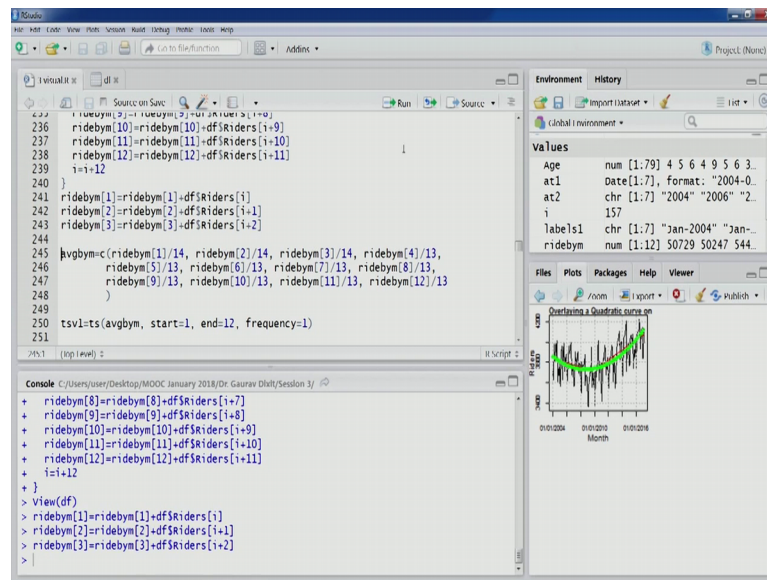
(Refer Slide Time: 29:52)



If you want to have a relook of the data you can do that d f is the a data frame can see 2004 starting from 2004 we have month for number of riders for every month and in the last month.

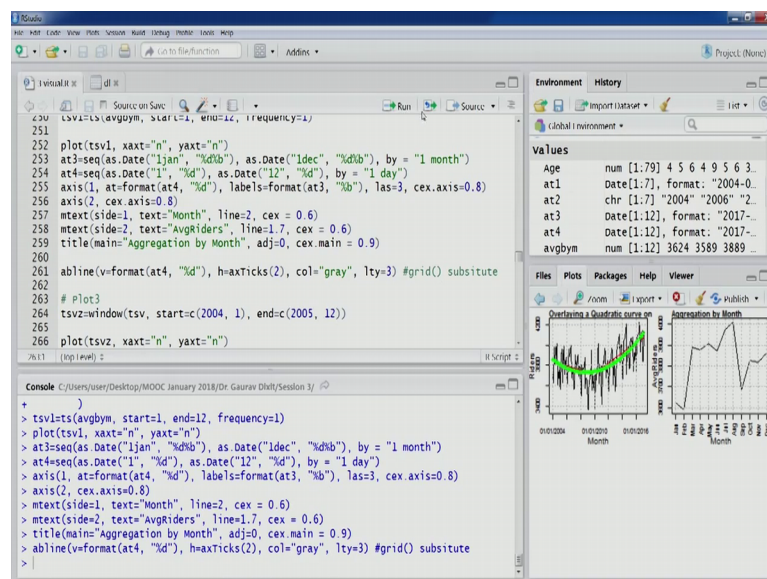
Let us these to the last year you can see we have a the rate on 3 only 3 months. So, the code that we discussed here this these particular 3 lines or adding those 3 numbers as well for these months Jan Feb and March. So, let us execute these lines now once we have all these numbers let us take average. So, you as you can see all these numbers are being divided by the number of months. So, Jan is counted 14 times and then the others are there these 3 months are counted 14 times and then rest of them have been counted for 13 times because there was 13 years of data.

(Refer Slide Time: 31:01)



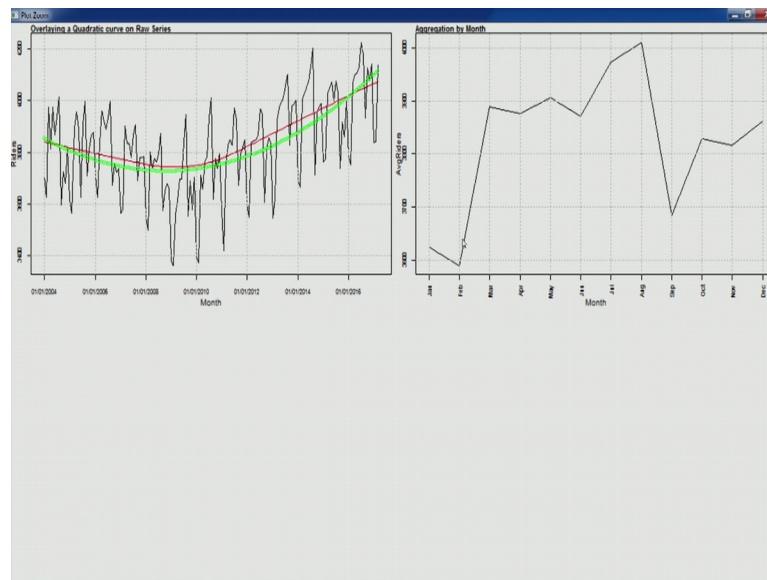
So, let us take average now let us create another time series vector this is now going to we using this particular data average number of riders by month. So, let us create this particular time series vector now let us plot it you can see this data.

(Refer Slide Time: 31:23)



Now, let us recreate the axis. So, we need to create labels. So, x axis y axis and the labelling for both the axis and then title followed by a grid now let us look at the graph this is our graph.

(Refer Slide Time: 34:45)



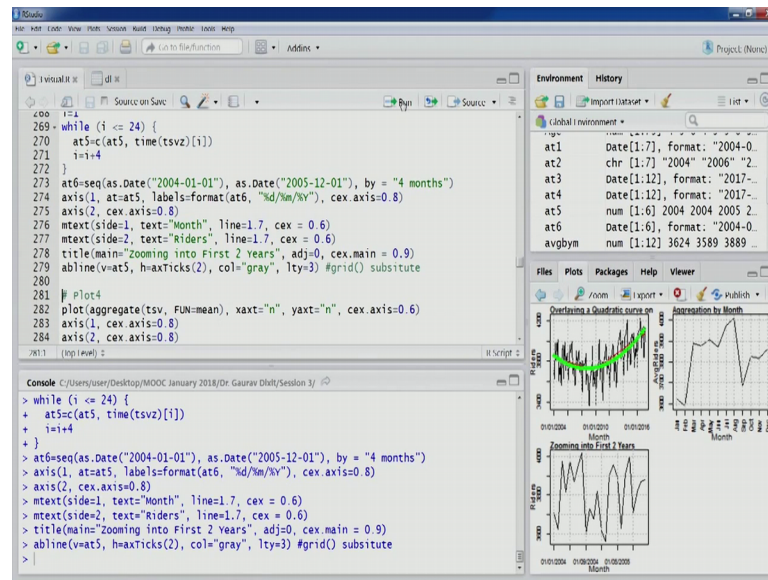
These is representing this is aggregation by month and monthly by for every month average number of riders over the years are being represented. So, you can see that if you look at this particular graph we can easily understand that in the month of July and august the ridership's the numbers of riders are at higher so on in an average sense. So, in an average sense in the month of July and August the numbers of riders are on the higher side and we if you look at the other months right. So, in the month of let say June there is a dip in the ridership and then there is a lowest numbers are in the month of Jan and Feb.

So, a number of riders are on the low on the lower side in the Jan and Feb because this particular data is reflecting the number of you know bicycle ridership in the in the I t roorkee campus you can you can see initially when the in the month of July and august this particular semester the environment the environment is much conducive of a bicycle ridership, but in in the month of Jan and Feb this cold weather. So, therefore, environment is not that much conducive there therefore, the ridership is also affected.

So, the same thing can is very well captured in this particular aggregation. So, let us move to our next plot this next plot is about so zooming in. So, we are going to zoom into zoom in zoom into particular year. So, first 2 years of data so we are going to have a look at the first 2 years of data. So, this can also be in a way you know we can subset the data and then plot it. So, that can also be done or otherwise we can zoom in using this particular function. So, window is the function in our that can be used to subset a time

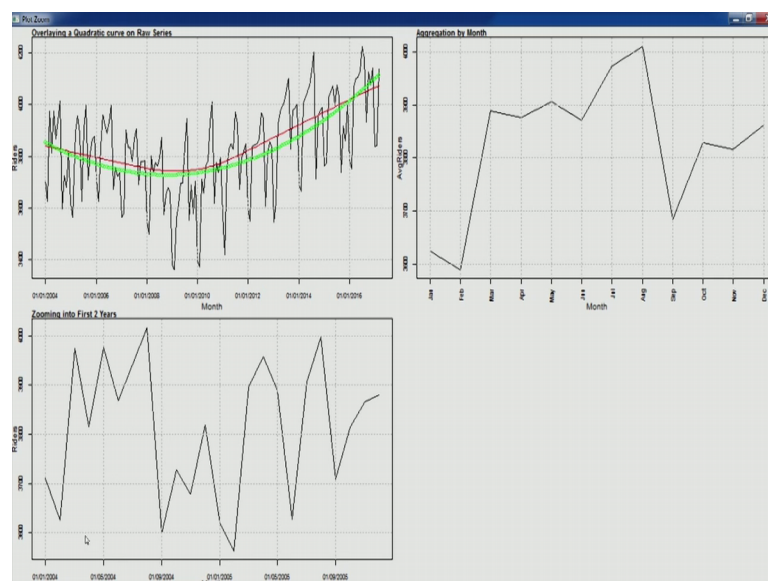
series. So, we can create a new time series a subset time series from the existing one. So, t_{sv} is the existing time series and we can create a subset using the start and end arguments. So, 2004 to 2005 data we are going to create a subset of.

(Refer Slide Time: 34:17)



So, let us execute this code let us plot you can see a plot has been created let us recreate the tick marks and axis let us create the labels in title you mean into first 2 years and also grid let us go back to the zoomed graph.

(Refer Slide Time: 34:44)

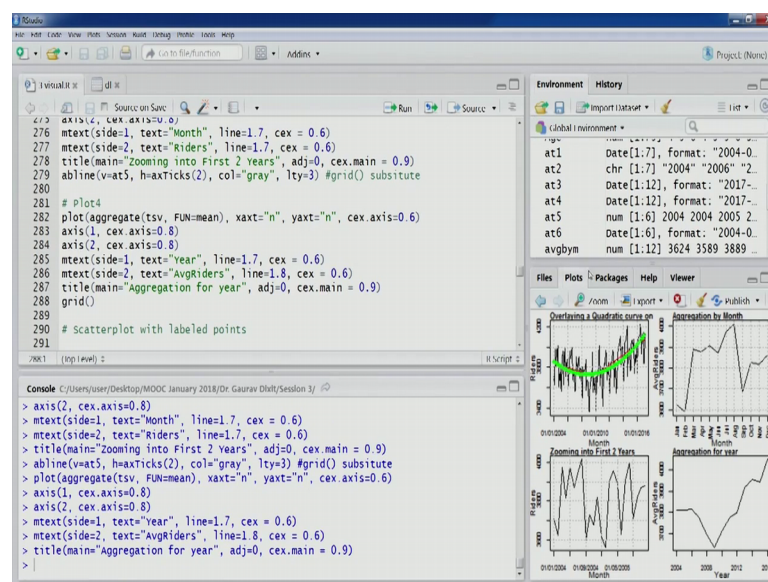


You can see is this is the these are the this particular line graph is representing first 2 years and you can have a look at the data.

Now, let us move into the a next plot this fourth plot is actually aggregation by year. So, if we are interested in looking in in looking into the you know a global sense we are trying to have a look at the data in a global sense what is the global pattern. So, for example, when we aggregated data using months so, that was mainly that can be called you know to understand the seasonality if there is any seasonality present in the data more what these terms actually mean more discussion would be when we come to time series forecasting, but right now we can understand that when we aggregate year wise we can have a look at the global things when we aggregate by month we can have a look at the full graph the r series gives a global trend look aggregation by month gives us a you know seasonality a look year on year also give us a some other patterns some global patterns.

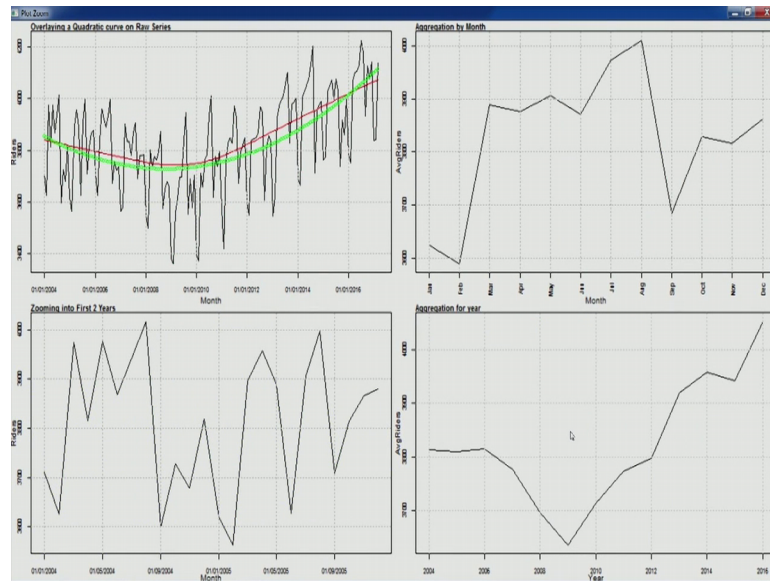
So, let us do that. So, we are going to aggregate this particular. So, aggregate is the function we are aggregating using mean using this particular time series. So, let us plot.

(Refer Slide Time: 36:21)



So, for every year these are the riders and this is the graph let us recreate the axis title and grid.

(Refer Slide Time: 36:39)



Now, let us have a look now you can see starting from 2004 to 2016 the data is the we can see the line graph and average rider average number of riders are depicted for each year and we can see overall year on year there was a dipper on 2009 and after that there the number of riders have been increasing year on year. So, we get the overall sense over the years what has been happening. So, global pattern can be seen very easily in this aggregation. So, we will stop here and in the next lecture we will start restart our discussion with on scatter plot with labelled points.

Thank you.