Business Analytics & Data Mining Modeling Using R

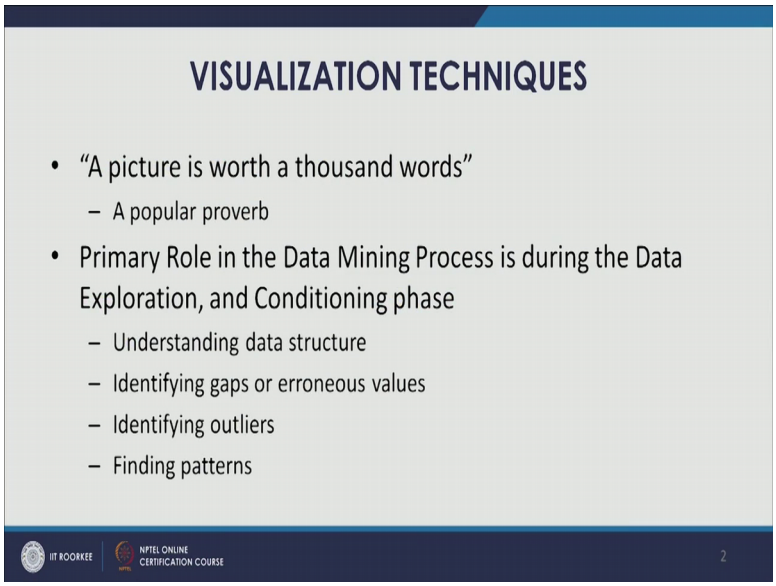
Dr. Gaurav Dixit
Department of Management Studies
Indian Institute of Technology, Roorkee

Lecture – 07
Visualization Techniques- Part I

Welcome to the course business analytics and data mining modelling using R. So, we completed our first module that was about general overview of data mining. Now, we are moving into our second module that is about data exploration and conditioning data preparation so those topics. Now, first lecture is going to be on visualization techniques, so let say start. Now, you might have a come across this particular popular proverb that a picture is worth a 1000 words.

Now, this is also a well known fact that visual processing of human a brain is much higher than numerical or mathematical processing. So, that is the main underlying importance of visualization techniques in any modelling process including a data mining modelling statistical modelling. So, therefore, if we as humans or as domain knowledge expert or as analyst or data scientist if we get to see, have you look at the data or to understand some graphs, to see some graphs, to see some plots.

(Refer Slide Time: 01:23)



So, we are able to exploit our domain knowledge, our expertise in a much improved fashion. So, that been the bases. So, we are going to start our discussion on visualization techniques.

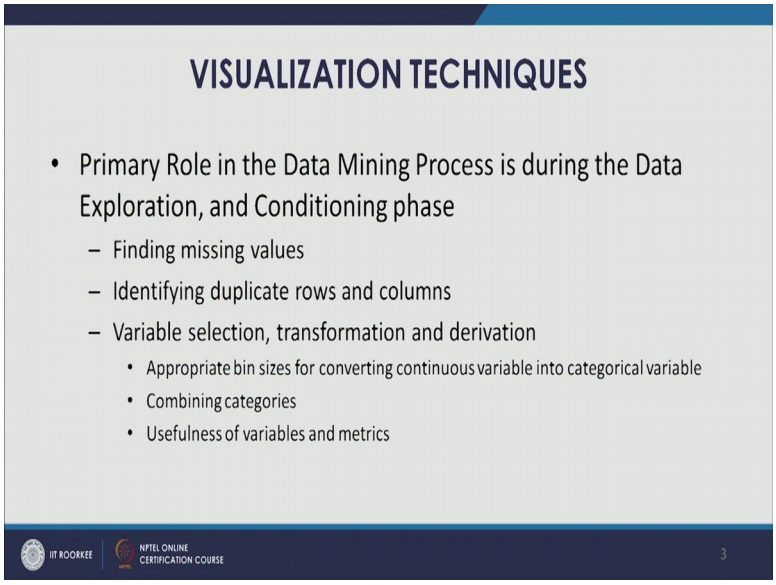
So, generally in visualization techniques they have primary role in the data mining process during this data exploration and conditioning phase, different phases we have already talked about in the previous lecture; data mining process specifically. Now, this primary role of visualization technique in this phase, data exploration and conditioning phase can be summed up with these points. So, we try to understand the structure of the data that is available. So, that is a 1 goal that is done using visualization techniques.

The second 1 being the identifying gaps or awareness values, so sometimes there might be few rows or could be duplicate few rows could be you know some of the values might be missing or some of the values might look out of place or they might look awareness. So, therefore, those gaps, some shelves might not have any values at all. So, identifying those gaps would also be part of visualization techniques.

Identifying outlier, so some of the values would be far away from the mean or median values, where the majority of the major chunk of the values are align. So, some of some of the values are might be far away. So, identifying those values whether they are valid point or whether they are awareness values that also need to be determined, so that we can move ahead for further analysis.

Now, finding patterns, as we said that visual processing of human brain is much better, so therefore, if we get to see the data, the plots, graph. So, we can easily find some we, can easily see some patterns which can in turn be helpful in terms of identifying appropriate data mining techniques or statistical techniques and then use them for our modelling process. So, these are some of the roles, where visualization techniques can be useful.

(Refer Slide Time: 04:00)



Now, building on those points, few other things could be missing values that we already talked about, identifying duplicate row and columns. So, that is also important sometimes some rows and columns could be duplicate. So, we would like to avoid that, because many statistical techniques might have this you know assumption that cases should be independent, in that case you know duplicate rows could be a problems similarly duplicate columns would also be problem in some statistical modelling techniques, where multicollinearity could be an issue.

So, therefore, so these, is specific things, these are specific term that I just talked about, will discuss them in more detail when we you know come up some statistical technique like regression and logistic regression. So, another important role played by visualization technique is about variable selection, transformation derivation. So, sometimes when we apply some of the visualization techniques on datasets, we are able to identify some of the variable which could be useful for the data mining task, some of the variable which could actually be transformed to suit our goal in a much in improved a fashion, much better fashion, derivation will also get some ideas, some directions about new variable derivations.

So, all these kind of things are possible through visualization techniques. Some examples are given here, for example, appropriate bin sizes for converting continuous variable into categorical variable that is something, when we look at the data, when we look at some of the graphs that we are going to cover in this lecture. So, we will get some idea what

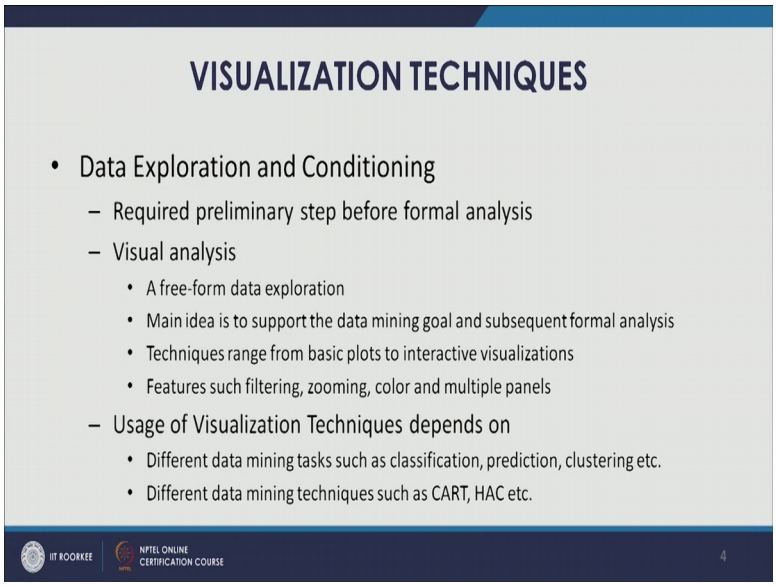
should be the bin sizes for a continuous variable for it to be converted into a categorical variable.

Combining categories, sometimes you know some categorical variable might be having many categories which might not be which all of them might not be useful for the specific task at end. So, sometimes it might be required or it might be mandatory by the data at some of the groups can be combined, so some of the categories could be reduced and you should be able to keep only the meaningful groups, meaningful categories for our appropriate task mainly classification. Another important role could be usefulness of variables and metrics. So, while we are exploring the data using visualization techniques, we will also be able to understand, which variables are important and which metrics are going to be used for performance evaluation etcetera.

Now, this particular phase data exploration and conditioning phase this is considered to be a required frame e ester before formal analysis and we say formal analysis we actually mean the data mining techniques and statistical technique like regression tress artificial neural network discriminate analysis. So, before we go ahead with those formal analysis of these techniques, this a particular step is mandatory; kind of mandatory where we apply you know some of the visualization techniques on data and do some preliminary processing, preliminary analysis.

Now, visual analysis let us understand visual analysis, role of visual analysis a bit more. So, it could be considered a free form data exploration. So, when we talk about regression analysis that is very structured kind of analysis that we do, but when we talk about visual analysis there is mainly we are exploring so and that to in a free form. We try many plots, many graphs that we are going to cover later in the lecture and which try to learn something about the data, which is going to help us in our further analysis.

(Refer Slide Time: 07:23)



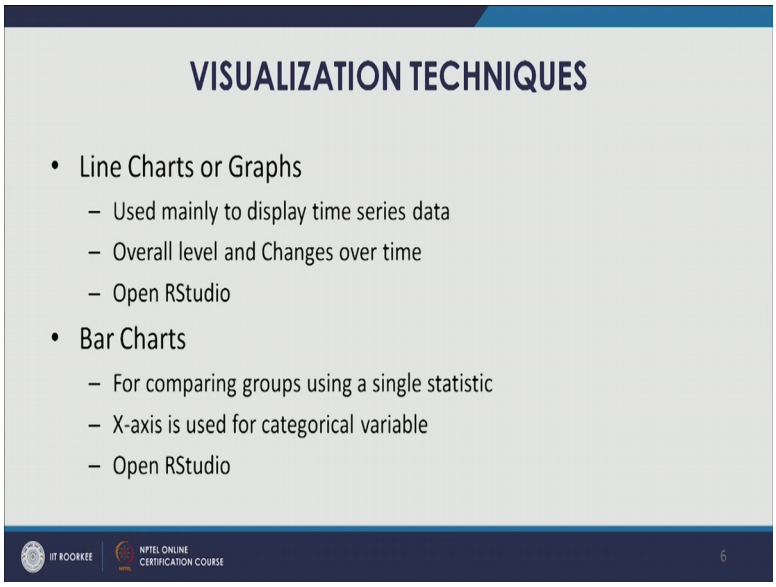
Now, as mentioned in the second point, that main idea is to support the data mining goal and subsequent formal analysis that is going to take place. Now, techniques in visual analysis they range from basic plots, that will cover you know line graphs, bar plots, a scatter plots. So these are, be, will cover from these some of these basic plots to interactive visualizations. So, interactive visualization they will a cover the multivariate nature of the data sets, will later discuss that it is generally the kind of modelling that is required is generally multivariate in nature.

So, therefore, some of the advance plots or interactive visualization can be really helpful for formal analysis. Now, the usage of visualization techniques that also depends on the kind of pass that we have, so some of the visualization techniques, some of the charts and plots would be more suitable for classification, some others would be more suitable for prediction, some others would be more suitable for clustering. So, therefore, it is the data mining task that will also drive, the way, the kind of visualization techniques that we are going to apply.

Now, different data mining techniques is also, such as CART and HAC that is hierarchical agglomerative clustering. So, CART is classification and regression tree modelling. So, some of these data mining techniques also have their own specific, I know visualization techniques, their own charts and graphs. So, that is also important to understand here. That we are not going to apply, everything that we learnt on you know

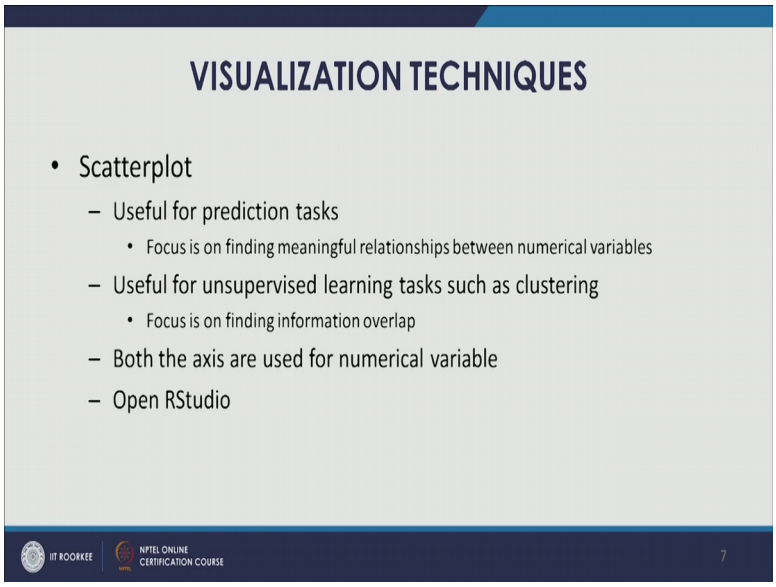
every technique that we are going to follow in subsequent formula analysis, but it is going to be task specific as well, classification, protection or clustering and also it sometimes it is going to be specific to the a particular technique.

(Refer Slide Time: 10:05)

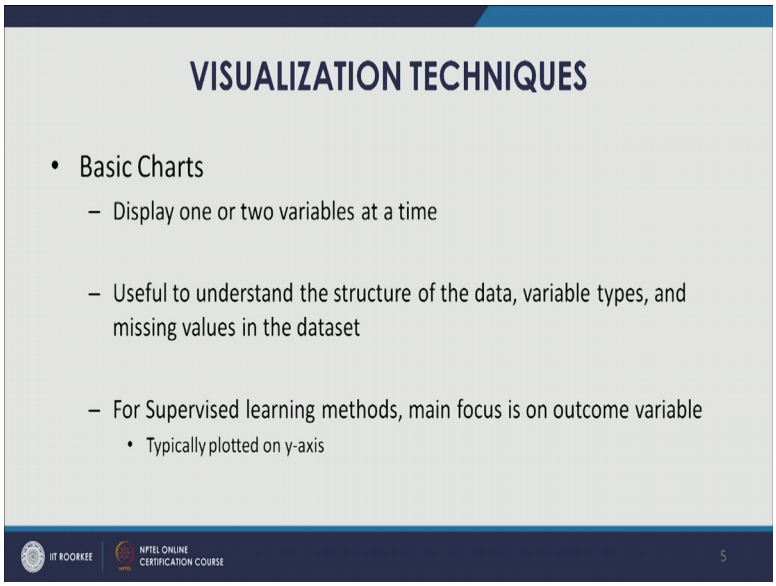


Now, let us start our discussion with the basic charts. So, as I said 3 important charts or graphs we are going to discuss, first 1 being line charts graphs, second 1 bar charts, third 1 being scatter plot. So, let us have basic discussion on charts.

(Refer Slide Time: 10:13)



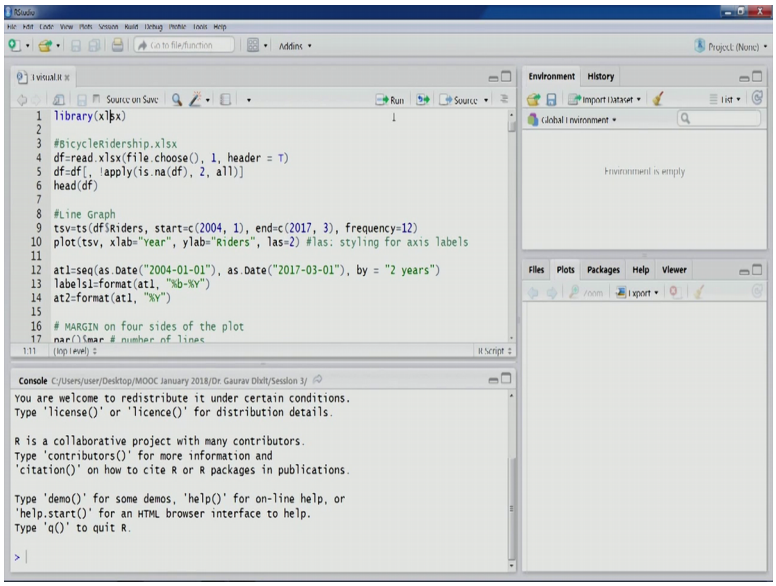
(Refer Slide Time: 10:17)



So, generally these basics are display 1 or 2 variables at a time. So, generally they are going to be 2 dimensional graphics and then there are going to be you know 2 variables we are going to pick and 1 is going to be 1 axis and other 1 is on y axis. So, generally 1 or 2 variables at a time and the a main idea being, to understand the structure of the data, variable types and missing values in the data set. So, generally these are the points where these basic charts are going to be useful.

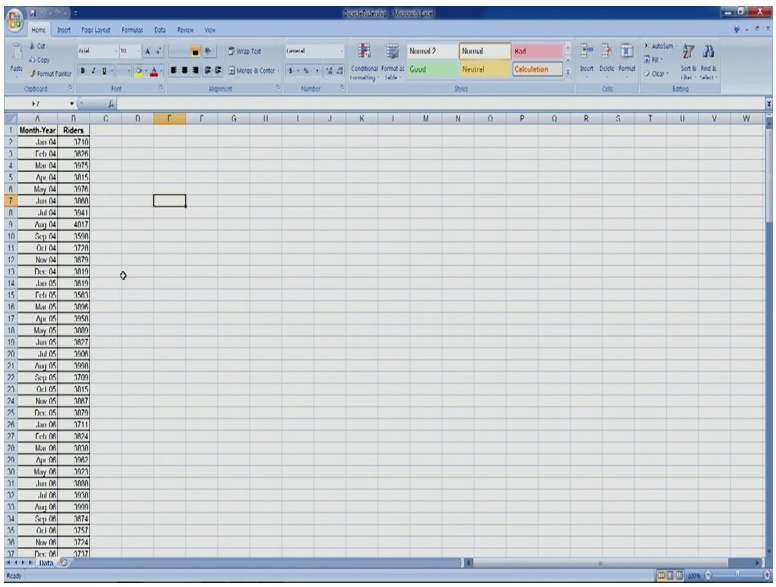
So, the basic charts for supervised learning methods, generally main focus is on outcome variable. So, that is typically plotted on a y axis for and supervised learning, so basic chart you can also be used for unsupervised learning methods as well, so will see that through examples using r, so let us a move to our next discussion on line charts.

(Refer Slide Time: 11:38)



So, line charts the main used mainly to display time series data. So, we try to a see the overall level and the changes that happen in the data overtime. So, let us understand, let us learn line charts through an example, let us open r studio. So, let us load this, a particular library xlsx. So, this is the data set that we are going to use bicycle leadership, let us understand this particular data set.

(Refer Slide Time: 11:59)

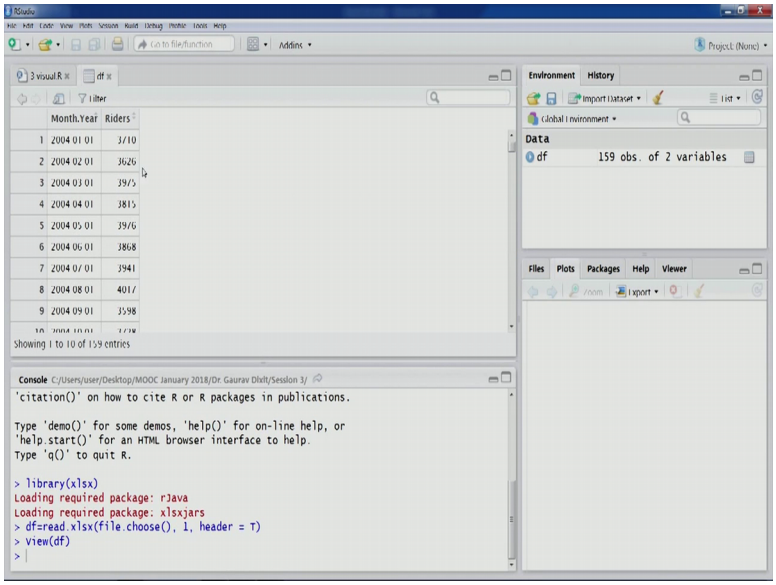


So, if you look at the actual data on this excel file, you would see that the data starts from January 2004 and it goes up to March a 2017 having a 159 data points and the second

column, the second variable is on riders which is actually the number of individuals riding bicycles. So, this is a mainly to reflect the a bicycle leadership in the I I T Roorkee campus, but this being mainly hypothetical data.

So, this data have created; hypothetical data have created for this demonstration purpose. So, a let us load this particular, let us import this particular data set has we have been doing in previous lectures. You can see in the environment section, the data set has been imported, you would see that there are 159 observation and 2 variables if you want to see the data here in the r environment or studio environment you can see here, month year is the first variable and then the riders, so the data because, this being time series data, so data is mainly on the riders, so it displays the number of riders in a particular month.

(Refer Slide Time: 13:20)



Now, this second line, the second line of code is actually about, if there are any as we talked about in the previous lecture if there are any deleted columns in excel files they would actually be picked up in the r environment. So, therefore, we want to get rid of those columns. So, this particular line this particular apply function is going to help in that. Now, let us look at the, let us have a look at the first 6 observation you can see for different months, Jan Feb March and a different months, we can see the number of riders that are there.

So, before generating a line graph, we need to create a time series vector here. So, t s is the command, if you understood in understanding more about a particular function in r

you can do so using help section t s, you can see t s is the function mainly for time series objects and you would get a detailed uses a t s different arguments, you can get detailed help here in the help section.

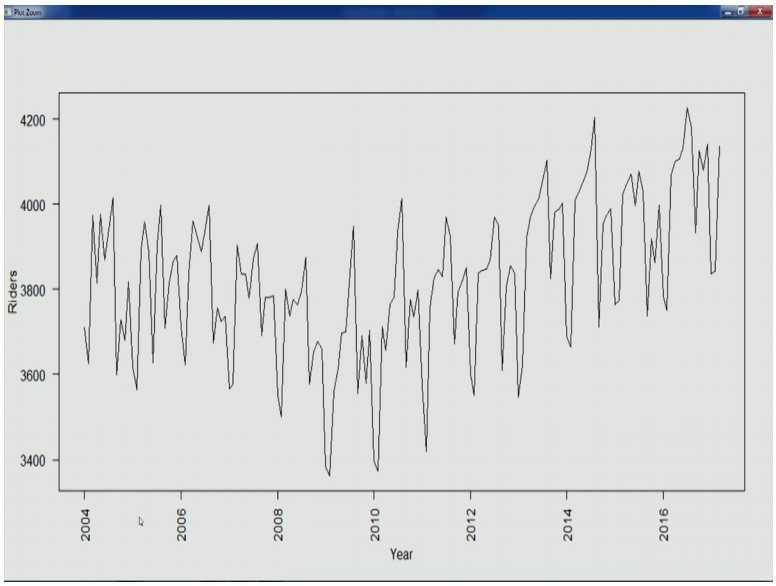
Now, let us go back to our code. So, here you can see that in the t s function the first argument is actually the data. So, we are passing on this argument data frame and is specifically this riders variable, so that is the first argument. So, we want to create a time series object out of number of riders for month for every month. So, you can see the starting of a date is from 2004 and 1 is for first month of the year January and then ending of this time series data is 2017 and then March that is third month and the frequency is mentioned 12, that is be mainly because the it is the monthly.

So, therefore, in a year we have 12 data points. So, therefore, frequency has been mentioned as 12. So, let us create this time series vector, this has been created, you can see the same in our environment the t s v has been created and time series is a number of values 1 2, 159. Now, if you want to plot this particular time series, so that is going to be our, a line graph. So, you can see again plot is the command that we used previously as well. So, in this case we have just 1 variable t s v and the this is going to be this particular rider riders is going to be on y axis as you can see that y label has been given and the x axis is mainly going to be used for time scale.

So, in this particular code, the time is scale would v determined by the function plot by default, the default settings could be applied for time scale and you would see then another argument l a s is there, that is for styling for axis labels. So, how the axis labels are going to be displayed? More information on this particular argument, you can find in help. You can type plot in the help section and you will get more information on different arguments, some of the arguments would be available in the par command that is for parameters; different parameters for graphical settings. So, you would see l a s somewhere there, you can see here. So, in the parameter p a r, par command, so you can see different styling of axis label that is displayed over there. So, you can see I have mentioned l a s is equal 2, so that means, I want my axis to be always perpendicular to the axis.

So, the labelling of points would is always going to be perpendicular to the axis. So, let us see how it is going to be displayed.

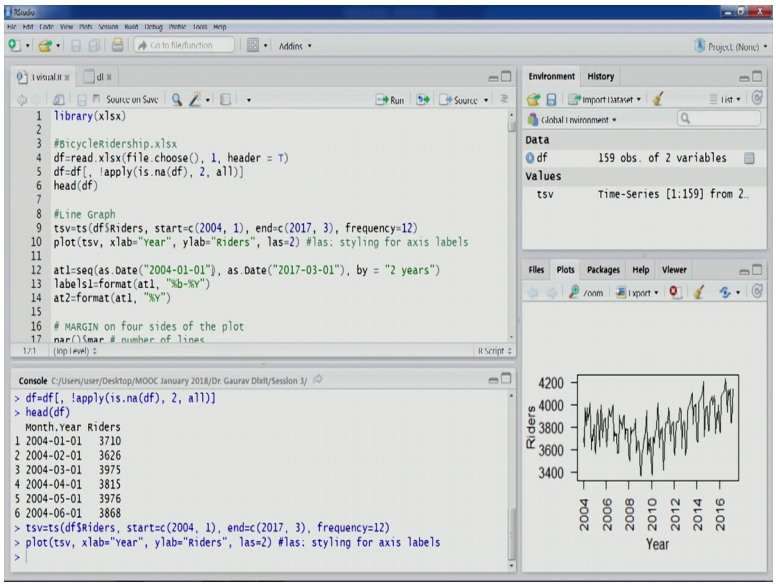
(Refer Slide Time: 18:30)



So, this is the plot, so you would see in the x axis years are being displayed and you would see different tick marks for different years have been picked up by the a plotting function, that is plot here, this case and the riders was defined by us, so that was the part of the time series vector that we created and this particular. Now, you would also see that labels for these axis are perpendicular to the axis itself, so the way they are here.

Now, you can also look at the plot as well, would see this is a line graph, so all the points for different points in time, riderships are there and these points have been connected and that represent a line graph. So, this can help us understand the main label of the data. So, you can understand the level of values that are being taken here. So, a 1 particular line which can pass through these points, somewhere in between can be considered as the main level of this particular graphic and then you can also see the changes over time. So, it is look this particular graph looks polynomials, so the changes looks a polynomial in nature.

(Refer Slide Time: 19:57)



Now, you want to improve this particular graph, so that is also possible. So, will try to, so if you want to improve this particular plot. So, first thing that you need to do is to create a sequence that can be used in creating our labels and tick marks in the axis is specifically the x axis that is for time. So, what I am trying to do here is I mean creating a sequence of, I am creating this particular vector using this sequence function.

So, sequence function will create values in this case a date sequence starting from 2004 up to 2017 and the difference, so these are going to be equally space points. So, difference being 2 years, so let us create it, then another function that could be used for formatting of a vector, so format is the function and for example, the sequence that I have just created in at 1 in this particular variable, so we can format this. So, format command would actually allow you the particular information that you want to retain in your specific format, in your customize format. So, in this case I am using percentage b and percentage y, which are mainly for month and year.

So, I am living out the date, day information and I am keeping the month and year information using this particular format function. So, let us execute this particular code and you would see label size being created. Now, in the environment section you would also see that labels a Jan 2004 and then Jan 2006, Jan 2008, so these kind of labels have been created. Now, if we want only the year part, then again we can use the format

function and we can extract only the year related information from this at 1 vector, so that is also possible, so let us do that.

Now, another important aspect of creating plots in r is, margin, margin the plots. So, there is this command par, p a r par parameter for parameter that we can actually use. So, there is another, so there is, this a variable that is available in a par command that is for margin m a r. So, this will actually tell us the different margin in a 4 sides of the plots. So, these 4 slide 1 with these 4 slides, 1 being the bottom, then you have left side, then you have top side and then the right side. So, all these 4 side what is going to be the margin that we can actually defined using this particular function, par.

So, let us execute this line. So, you would see the default setting for margin 5.1. So, these numbers actually represent the number of lines, so 5.1 number of lines, 4.1, 2.1, this is the margin that is by default that is there, when we create a plot. So, we want to change this particular because we want to change the axis and you would see a lot of spaces being taken the way we are a labelling the axis, which being the perpendicular to the axis. So, lot of space is required in this case, so therefore, lot of margin is required.

So, we need to change the margin. So, therefore, you would see that, the first margin is actually for the bottom side. So, we see it is the highest 1 8. So, we want more margin here and then we want 4 in the left side, then we want 4 in the in the top side and just 2 on the right side and then there is another point 1 addition to this. So, once this particular margin has been created, now we can go ahead and start getting our new plot. This new plot is on the same a time series vector that we have created, but the graphic is going to look slightly different much better.

So, in this case we want to create a new plot and we want to create a new axis. So, you would have to use these parameters in the plot command x a x t and y a x t. So, in this case I have assigned n value, that means the x axis and y axis, would not be plotted in the graph. So, they would disappear and labels also have kept them as null. So, there are not going to be any labels just the graph, the line graph is going to be displayed without any x axis or y axis or any labels for those axis, you would see that a box and a graph within that particular box is displayed over here.

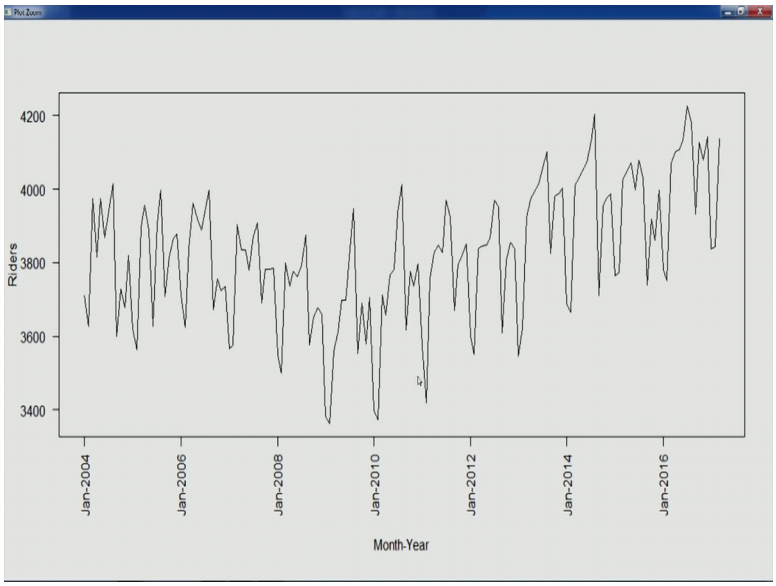
Now, we will have to create this axis. So, axis is the function that is available in r, which could be used to create this function. So, x is 1, 1 means 1 is for x axis and then the next

line is for axis 2 the 2 is for y axis. So these, this axis function can actually be used to create x axis and y axis. Now, you would see that at this is the tick mark location we are using at 2, so at 2 is going to be used. Labels you would see, labels 1 that, so these labels are, so at these tick marks at these points, the labels that have that are there in labels 1 they are going to be displayed and styling of these label is going to be again l a s is 2. So, therefore, it is going to be perpendicular to the axis.

So, let us do this and you would see an x axis has been created which is slightly different from the previous plot. In the previous plot, only the year was displayed, now you can see on the month and year is has been displayed Jan 2004. So, why we are doing this, main reason being the way we have, the kind of data that we have in the data, it is monthly data. So, we have ridership data month wise. So, therefore, it is more appropriate for us to create a, generate a plot which also shows a month not just the year. So, month and year that information is being defected in the plot now, now similarly we can create the y axis.

So, we particularly we did not want much changes in a y axis. So, therefore, it has been displayed as is. Now, if you want to a label the axis, so this is another command m text that can actually be used to label the axis, x axis and y axis. So, again in m text also first argument is going to be you have to select the axis. So, side is equal to 1 mean, being the x axis if you want to a change y axis and it has to be side h 2. Then you can mention text argument is there, where you can mention the level of the axis and then a line argument is there which will actually tell you that, how many number of lines the labelling would be created after how many number of lines below the axis or from the axis. So, let us execute this particular code and you would see that month year have has been created, similarly for y axis you would see riders has been created.

(Refer Slide Time: 28:00)



In the zoomed version of this graph, you can see the same month, year and rider. So, this is much better plot then the previous 1, where we just had the year information on the time scale, now we have month and year information on the time scale.

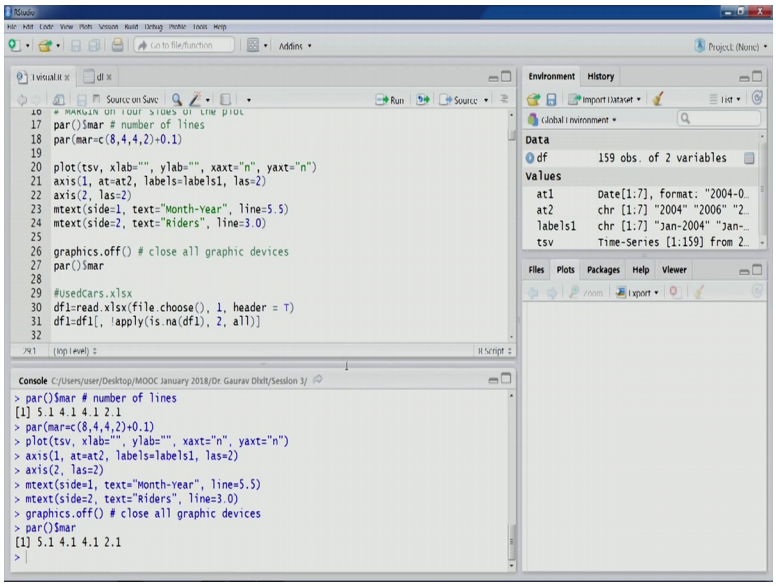
Now, next function that we can learn is about graphics dot off. So, this particular function if you call this function, so all the plots would actually disappear from your r studio environment and there is an another command def dot off, if you run that particular command then only the current plot the default plot that is being displayed only that put, we deleted or erased. You can achieve the same effect using these 2 points, these 2 tabs here in the plot section; you can see these 2 tabs, so that can be achieved.

So, we want to get rid of all these plots, so let us run this, you would see everything would actually disappear. Now, you can again check the margin, once this all the devices are closed, the par setting, this par function and many settings related to this particular function would actually be reset. So, this we can check for the margin because we change the margin you can see, that the default numbers are again having set for margins.

Now, this was our discussion on line charts and line graphs, more discussion on a line graphs and how we can use it further would be covered in time series forecasting time series, where we are going to learn more about line chart and how it could actually be used before the formal analysis and time series forecasting can actually happen, so will

learn more in those lectures. So, let us move to our a next basic chart that is bar chart. So, the main youthfulness of this particular bar chart is for comparing group using a single statistic, so will see, how that is done. Now, generally in x axis is used for categorical variable. So, generally x axis is going to be reserved for categorical variable and we try to understand more on that particular variable using y axis. So, let us do this through an example.

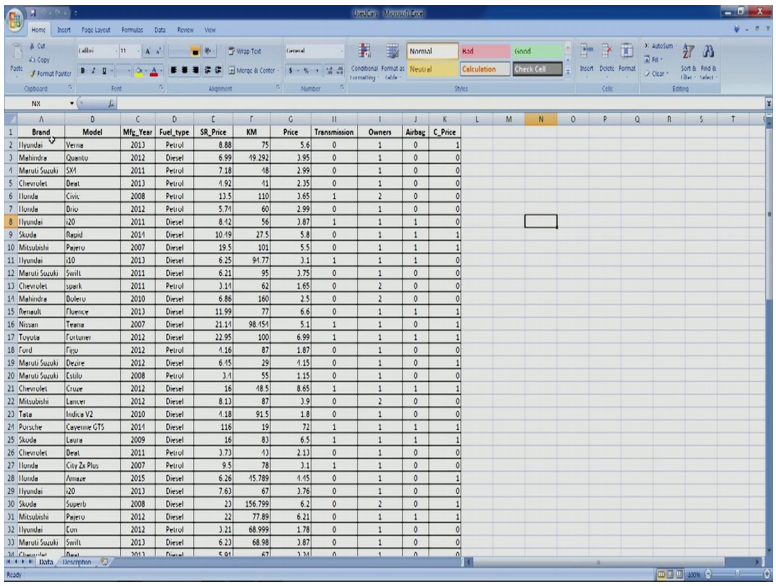
(Refer Slide Time: 30:25)



So, to understand bar chart and other charts we are going to use this particular data set that we are already familiar with, used cars data set. So, let us load this particular file, so you can see this particular data set has been loaded here, we have 79 observation and 11 variables. So, let us run this a command, so that we are able to get rid of the deleted columns, so there were no deleted columns.

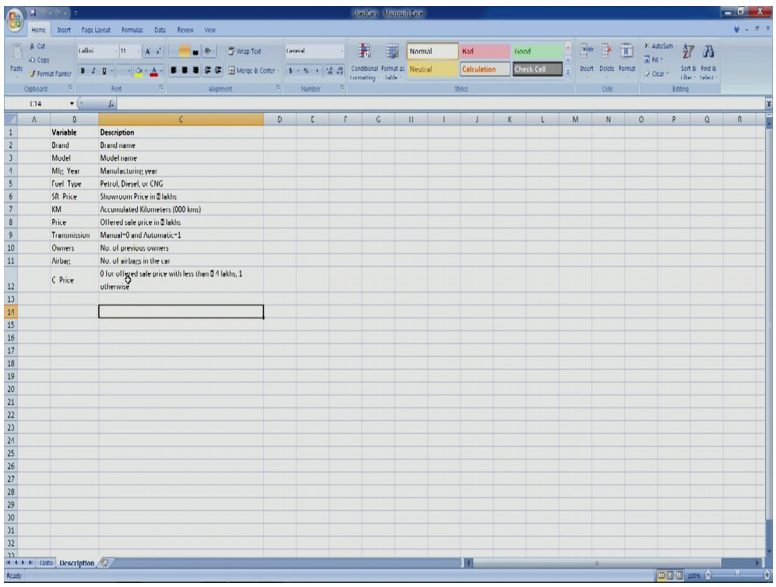
So, let us look at the data as well, so let us go back to the original excel file. So, this data set, we already familiar, but let us again have a look.

(Refer Slide Time: 31:21)



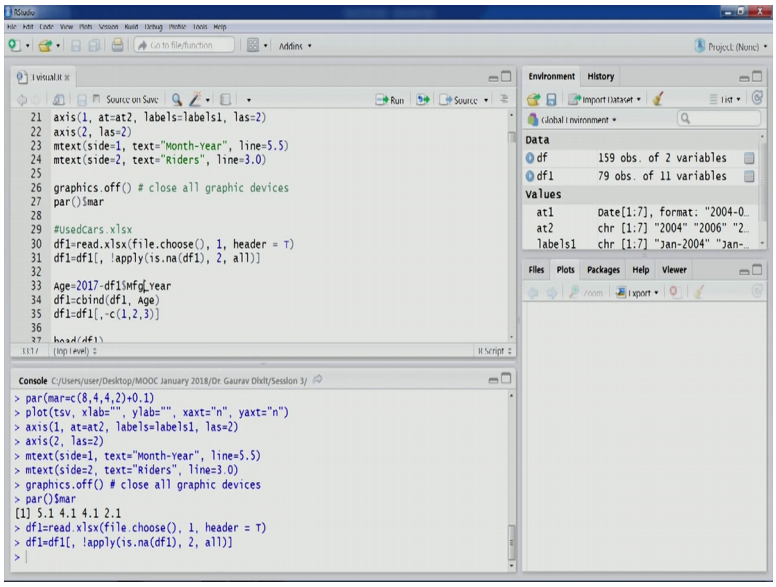
So, that data set is about used cars, so we have this information, through we have information on our used car like brand, model, manufacturing year, fuel type, showroom price, kilometres accumulated, the offered price and the transmission whether it is manual or automatic, then the number of owners and then the number of air bags and then this another variable c underscore price, which has been created manually.

(Refer Slide Time: 31:49)



So, all this particular variable is, if is 0, if offered sales price is less than 4 lakhs and 1 otherwise, so these are the variables and this particular data sets. So, from that data set we can actually see that there is the 1 variable there was the manufacturing year.

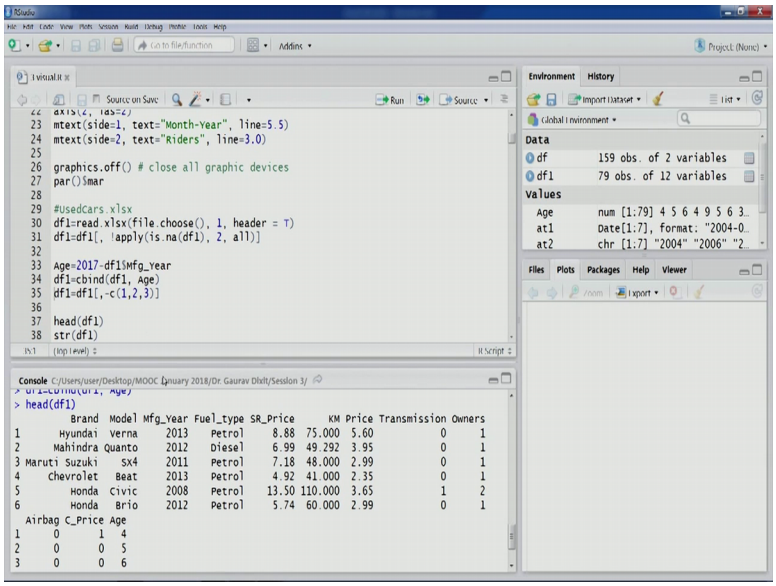
(Refer Slide Time: 32:15)



So, from the manufacturing year and if the current year is 2017, we can actually calculate the age of the vehicle. So, that we can do using this particular code, we can subtract manufacturing year from 2017 and we can create age and once this is done we can use c mind command to add this particular column in the data frame.

Now, in the data set, that we have already seen let us have a look again. We might not be interested, you would see that age has been added there, you can see age there and you would also see that some of the variables would like brand, model and manufacturing year they might not be required now, so will get rid of them.

(Refer Slide Time: 33:04)



Now, these are the variables that we are interested in. So, let us stop here and will continue our discussion in the next part.

Thank you.