Essentials of Data Science with R Software - 2
Sampling Theory and Linear Regression Analysis
Prof. Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology, Kanpur

Introduction to R Software
Lecture - 08
Graphics and Plots

Hello, welcome to the course Essential of Data Science with R Software – 2, where we are dealing with the topics of Sampling Theory and Linear Regression Analysis and in this module we are going to talk about the topics of Introduction to R Software module. So, you may recall that up to now we have discussed different types of command, different types of functions, different types of mathematical operation related to scalar, matrices etc. and the data handling in R software.

Now, I am going to address the graphical part. In R there are different types of graphics and plots which are available and you can believe on me they are excellent. The only thing is this you have to learn, you have to practice and they can give you an excellent picture, excellent graphic, excellent plot. The only thing is this they are not click click click, but you have to just see what are the commands and what are the different parameters that have to be given.

But here in this lecture I am going to just discuss couple of graphics and plot which are needed for me in covering the further lectures. So, I will try to take here the first command which is here the plot command, but we, but before I try to go to the graphic part means how to create the graphics in R, let me address one thing more. Well, that is my experience that many time people ask me what is more important in this analysis in statistics or decision sciences whatever you want.

Whether graphics should be there or this analytical should be there, analytical means the formula like mean, median, standard deviation, correlation etc. etc. and graphical means some different types of graphics.

And people have a myth that if you try to put more number of graphics in your report, possibly the report will become better means I will simply declare before all of you this is a myth. Please do not do these things, means it is something like this if a patient goes to a doctor and the doctor gives him all sorts of medicines.

Do you think that will that be good or you try to choose the appropriate medicine which are effective for the disease? So, that is the same story with the graphics, every graphic has a peculiar property that it will try to demonstrate a particular type of hidden information. This is only you who has to think what type of information you want to dig out from the data because as I said data is deaf and dumb data cannot raise its hand and say I have this information, I have that information.

So, please think logically before you use a plot and try to see whether this is giving you relevant information or not, yes, I agree in the first shot it is very difficult. So, you have to try with different types of graphics and you have to see that which graphic is giving you what type of information and what is easy to understand and easy to communicate to others and what is more effective in digging out the information which is inside the data.

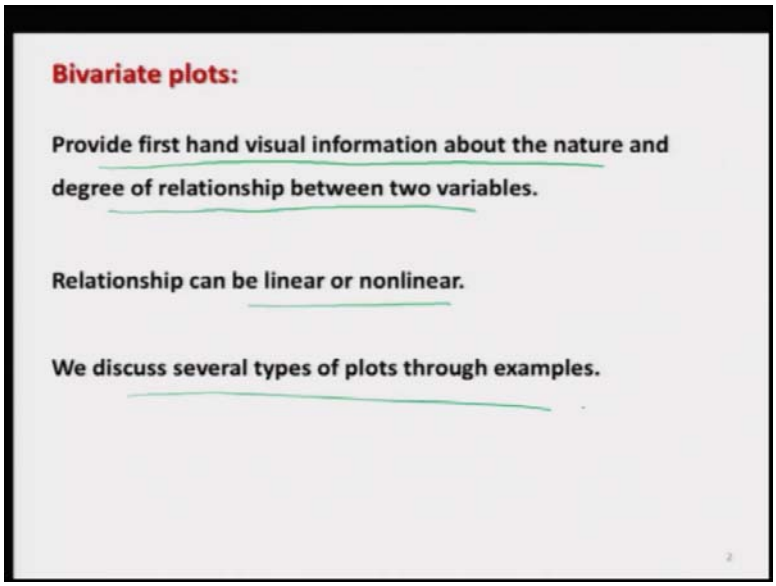
So, this is the moral of the story which you always have to keep in mind and that is my very sincere advice to you all, ok. So, this analytical tool and graphical tool both give us different types of information in different ways. So, my suggestion is this- both are important, the graphical will tell you the things in a shape of a graphi,c picture which are sometimes easy to understand and analytical tools will quantify the information.

By looking at the graphic possibly it might be difficult for you to say whether the average marks of this class is greater than the average marks of another class, but by finding out the arithmetic mean you can easily judge it.

There can always be a confusion on two similar type of graphic that whose variability is more, well you can see that whether the variability is smaller or more or bigger something like this, but unless and until you find out the standard deviation or variance you cannot quantify it, you cannot understand it. So, both are important and there is no ordering which is more important.

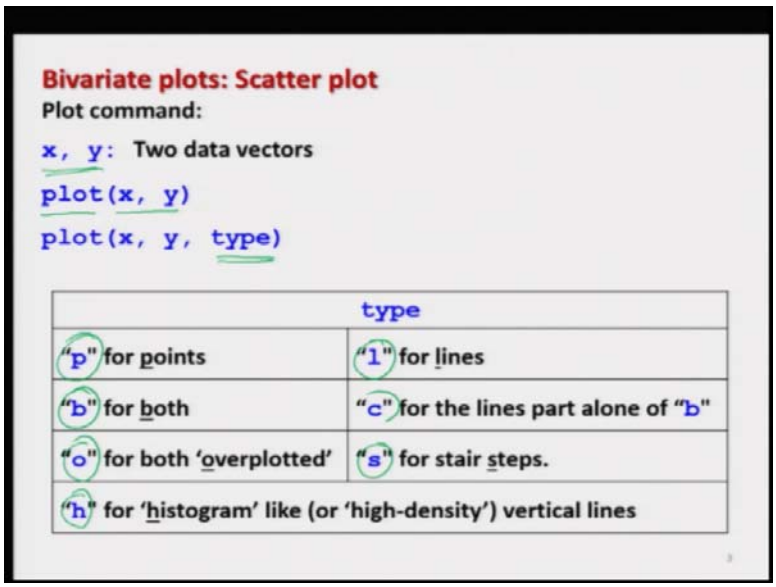
So, now continuing on this aspect that both are important, now I am going to address here the first plot which is bivariate plot. Bivariate plot means you are trying to create a plot between two variables, when you are trying to depict a relationship between the two variables that can be plotted, that can be created and that can be linear or non-linear, but we do not know unless and until you plot the data and data tells you whether the relationship between them, between the two variables is linear or non-linear.

(Refer Slide Time: 05:53)



So, I will say simply that the bivariate plots give us the first hand visual information about the nature and degree of relationship between the two variables.

(Refer Slide Time: 06:06)



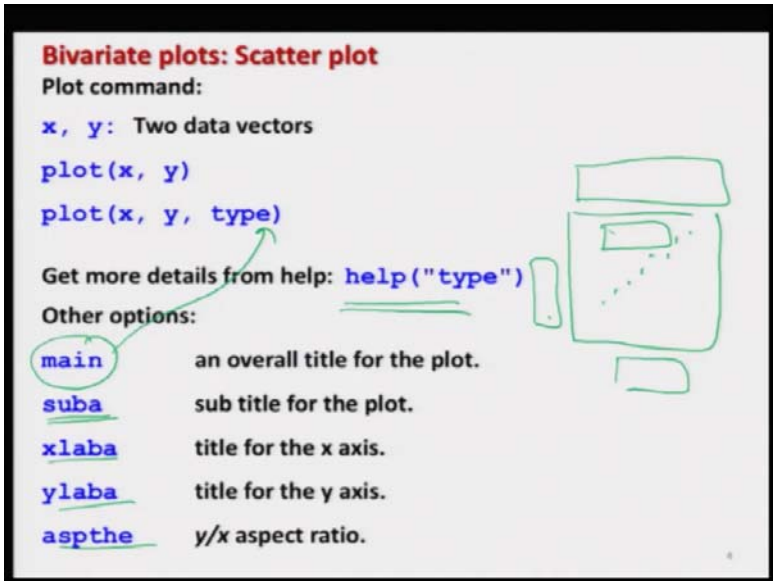
This relationship can be linear or non-linear and we will now take up different examples to see that how these graphs can be created. So, first let me consider here the command plot p l o t this is the command which is used to create a bivariate plot or a scatter diagram.

So, if I have two variables or two data vectors x and y of the same length then this plot command can be used plot x comma y inside the parenthesis. And when you are trying to create this plot

there is an option to write a type and type will give you different types of plot. For example, if you use type equal to p, this will give you points, if you say use type equal to l it will give you lines.

And if you give here type equal to b that mean it will give you both the points as well as lines and if you say here use a command c. possibly this will give you here line part alone of b and if you give here o this will give you over plot means lines and end points together and if you use here s it will give you here a stair a steps types of plot and if you use here the option type equal to s then possibly it will give you histogram like as 'high density' vertical line.

(Refer Slide Time: 07:27)

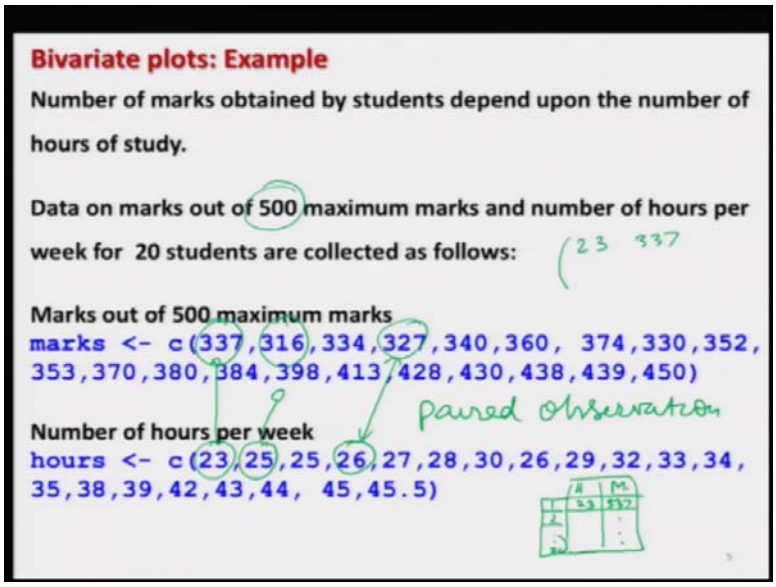


So, let me try to take an example over here and then I will try to show you, but before that whenever you are trying to create a graphic, the graphic may be will be looking like this, something like this, whatever it is and you always want to have a title of the graphic something on the x axis, something on the y axis, sometime the main title of the graphics all those things are possible.

I am not going to discuss them here in detail, but you can look into the help and this will give you help plot helpline they will give you all information. For example, if you use here the command main just over here means of the type inside the parenthesis main and then in and if you try to write down the title which you want on the plot can the output will have the title on the graphic also.

Similarly, if you want to have a subtitle of the plot that can be used by command suba, the title on the x axis, title on the y axis as aspect ratio etcetera everything can be controlled, you can control all sorts of parameter in the graphics but you have to first study it, read it and then understand it.

(Refer Slide Time: 08:36)

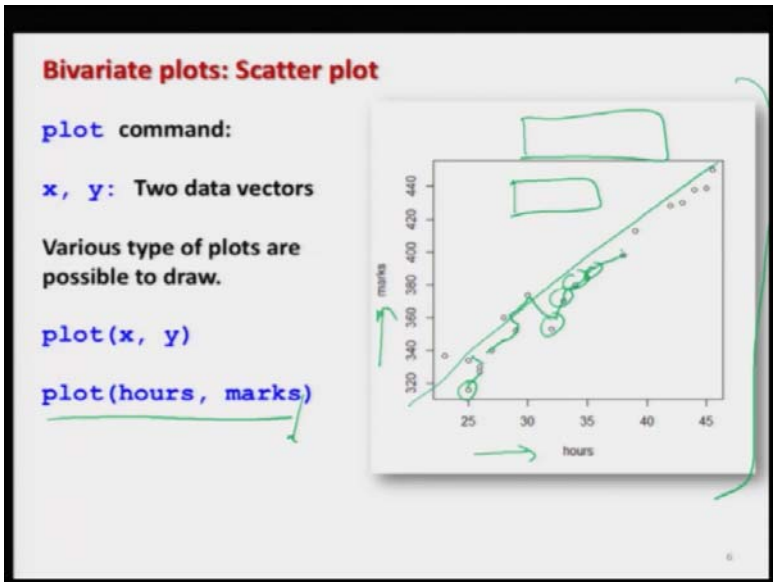


I will be here extremely simple to explain all these things. So, let me try to take here one simple example this example is related to 20 students, 20 students were given some examination and they have got the marks out of 500 and now some students have got lower mark, some students have got higher mark. So, it has been inquired by them that how many number of hours per week they studied. So, and then we have got here the paired observations, for example, if a student says 23 hours, he has got 337 marks.

And you can see so this and this here they are the paired observation, right. So, if one is x then and another is y then this is x 1 y 1 similarly this observation that if the student has studied 25 hours and he has got 316 marks, then similarly here student has studied 26 hours and he has got 327 marks.

So, these are paired observations, right. I can write them in the form of a table here like also like this that here number of hours and then here number of marks and then here a student number 1, 2 up to here 20. And then student has studied 23 hours. So, it. So, the student has got 337 marks and so on. So, this is my data.

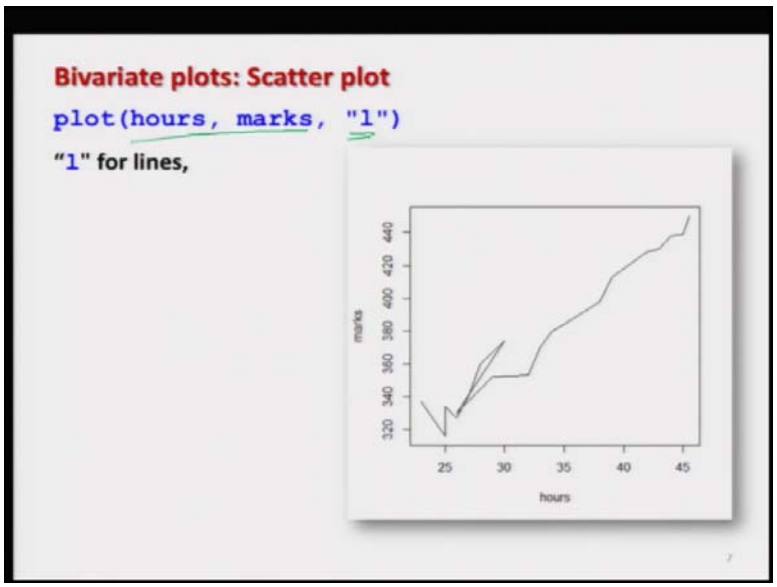
(Refer Slide Time: 10:12)



Now, in case if I simply use here the command here plot x y or now my variable is hours and marks. So, if I use plot hours marks I get here a graphic like this one, you can see here that all the data values they are just like a point over here. So, they are trying to give me an idea that there is a sort of linear trend here, right.

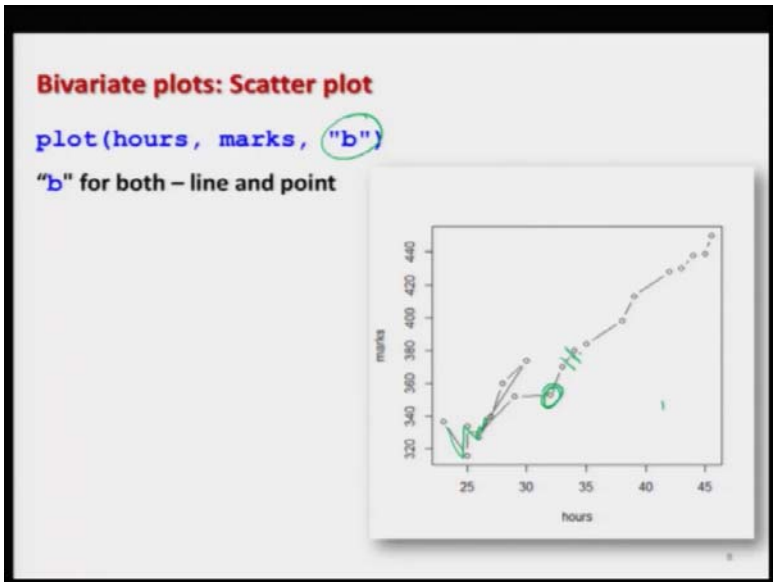
So, this will give us a lot of information to me now if you want to change here the title on the x axis, title on the y axis, some sub title, main title of the graph that you can use over here with different types of option, ok.

(Refer Slide Time: 10:57)



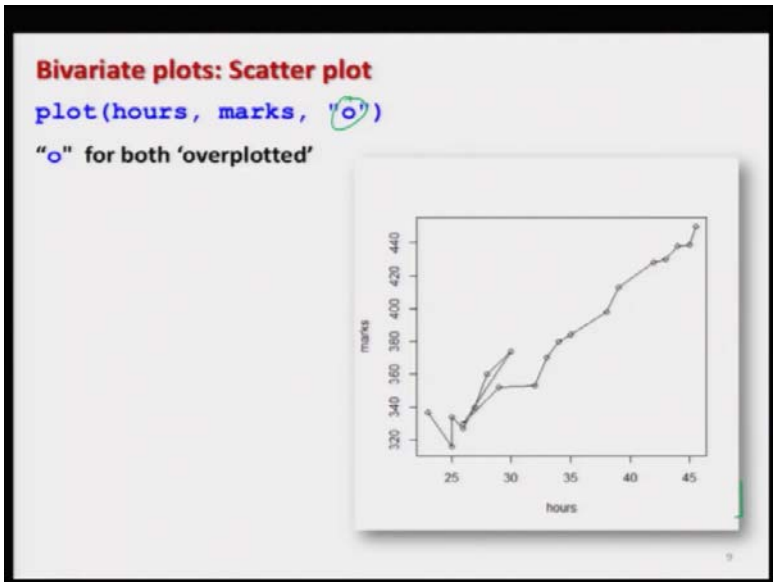
Now, in case if I use here that type equal to here “l” for example, l mean say a line. So, I take the same data set hours and marks, but I use here a command. So, you can see here the same command which was here given as point now all those lines are joined actually and so on right. So, you can see here that all the points are joined by lines, but the points are not appearing.

(Refer Slide Time: 11:25)



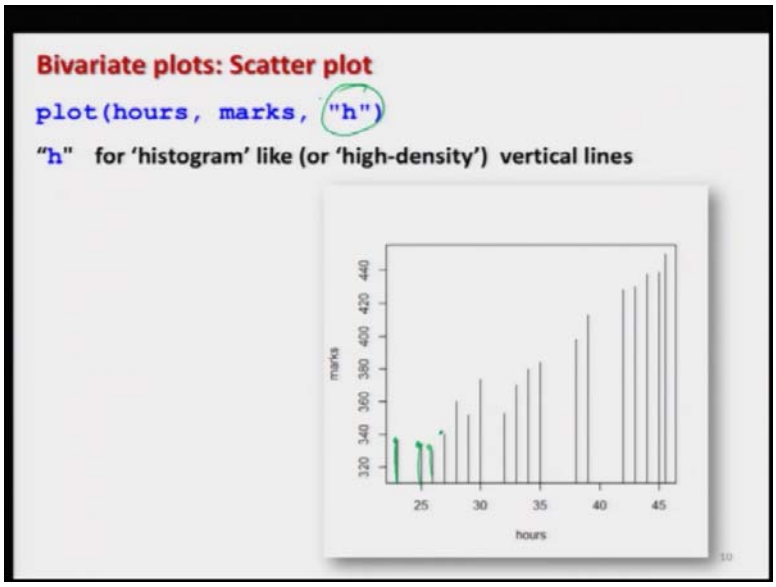
But if you want here lines and points both then I have an option here type is equal to “b” and you can see here that this points as well as lines both are given here in this graphic.

(Refer Slide Time: 11:40)



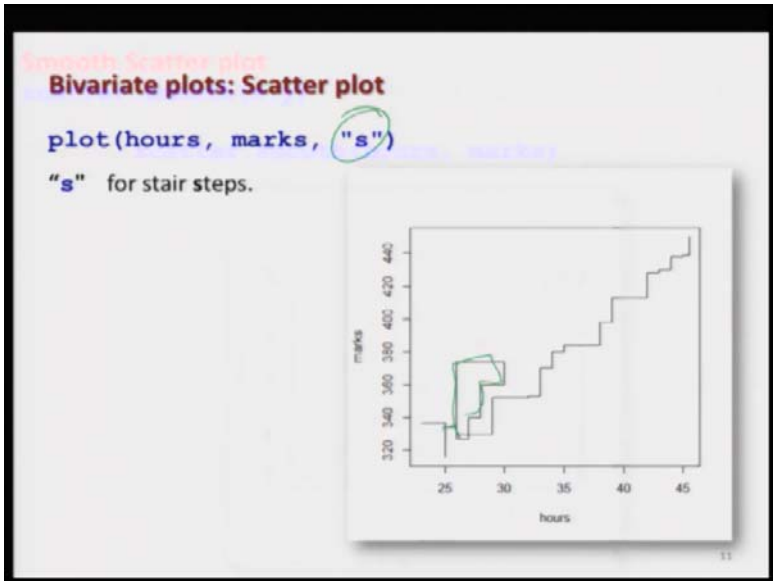
So, and if you right and, but if you try to see here, here the points are points and lines are difference there is no continuity. Now, in case if you want the continuity then you have to use the option type is equal to “o”; that means, over plotted. So, you can see here all the lines are crossing the points also and lines are overlapping the points also, right.

(Refer Slide Time: 12:07)



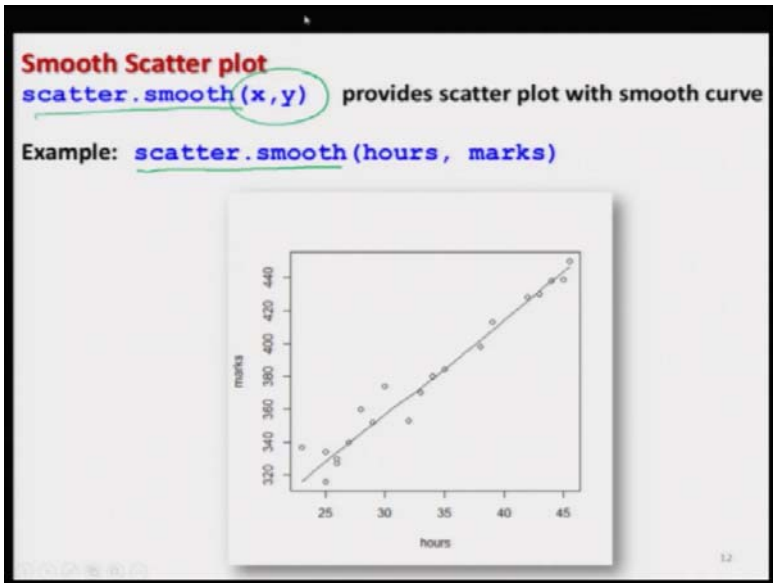
And similarly, if you want to if your data point is here, here, here and if you want to simply create it out of sort of high density histogram vertical lines type of graphic then you can use the option here type equal to “h” and you will get here this type of graphics.

(Refer Slide Time: 12:24)



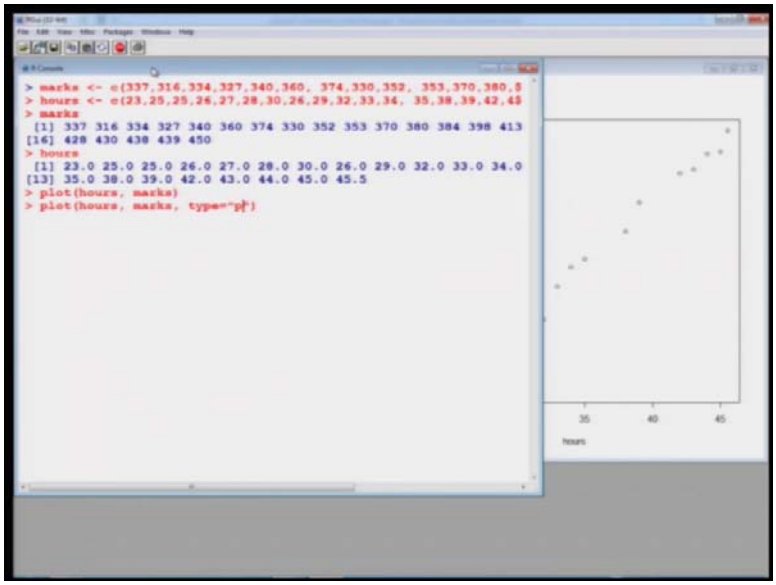
Similarly, if you want to have an option like a stair type of graphics. So, for that you have to use the command type equal to "s" and you can see here all the data has been in the form of a stair steps.

(Refer Slide Time: 12:40)



So, now, let me try to take this data here on the R console and I try to show you that how this graphics are working, ok.

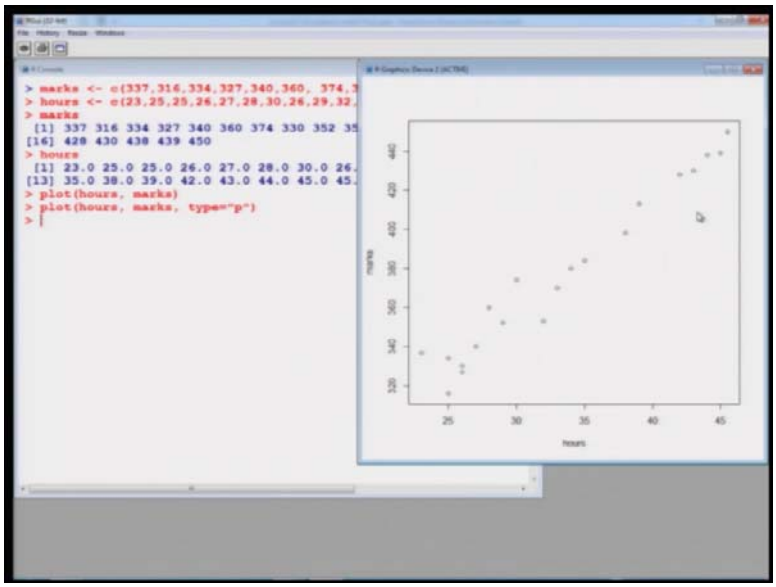
(Refer Slide Time: 12:55)



So, first I try to create my data set here- marks. So, I can create here marks, right and then I try to take it here hours. So, you can see here these are the marks which I have entered and these

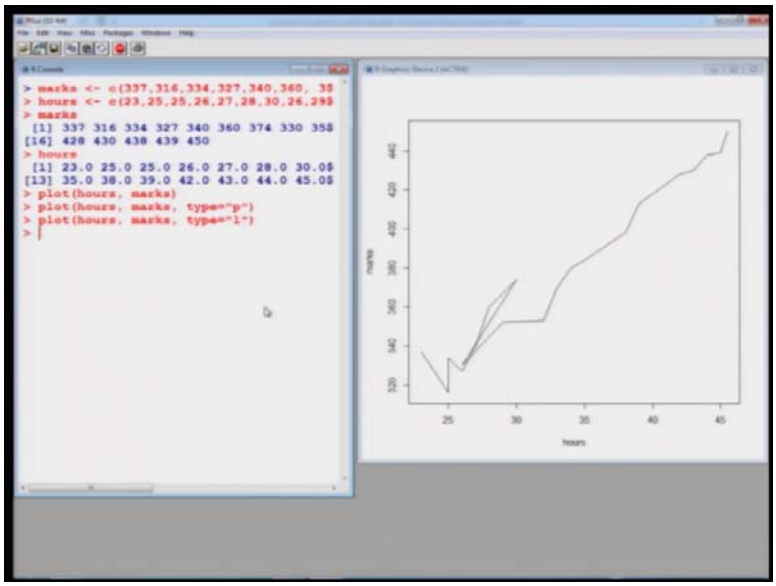
are hour the, these are the hours which I have entered and now if I say here the same command plot x, y. So, I come here and if I try to type here plot hours, marks and after that I will try to take here I will try to use here the same command here type, ok.

(Refer Slide Time: 13:39)



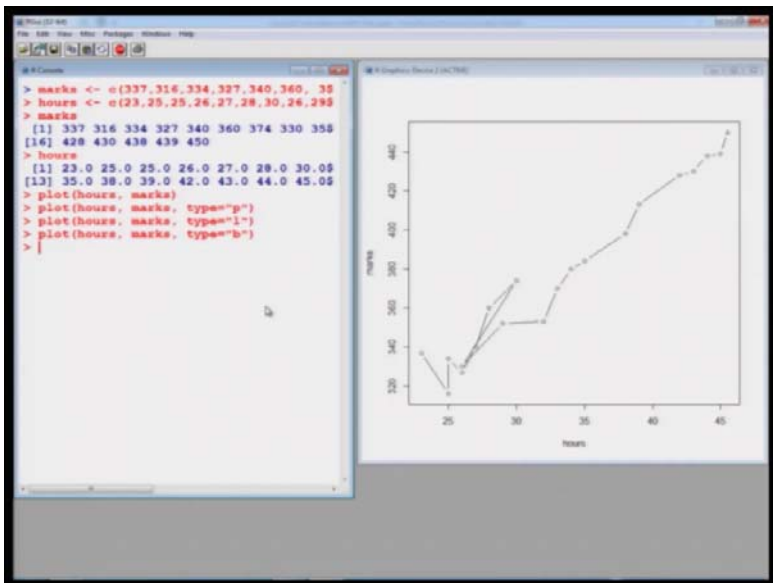
So, let me try to see here as soon as I see you see I see here a graph like this one, right and if I try to make it here if I try to put here type is equal to say here “l”; that means, line or if I say here first I may get a point. So, you can see here that is the same thing so, the default here is actually point, right. Let me make this window small so that you can see this graphic change in the graphics very clearly, ok.

(Refer Slide Time: 14:17)



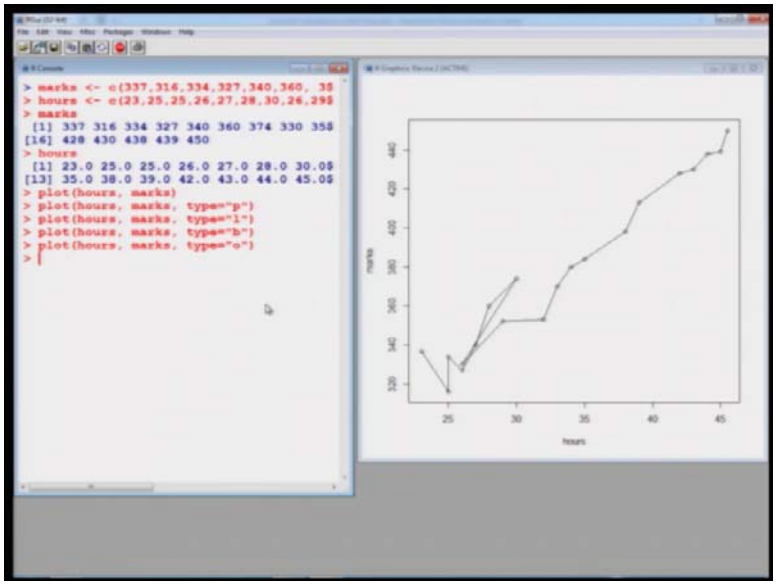
Now, I try to change here my plot I make it here “l” that means line. So, you can see here now the graph is changed to line.

(Refer Slide Time: 14:27)



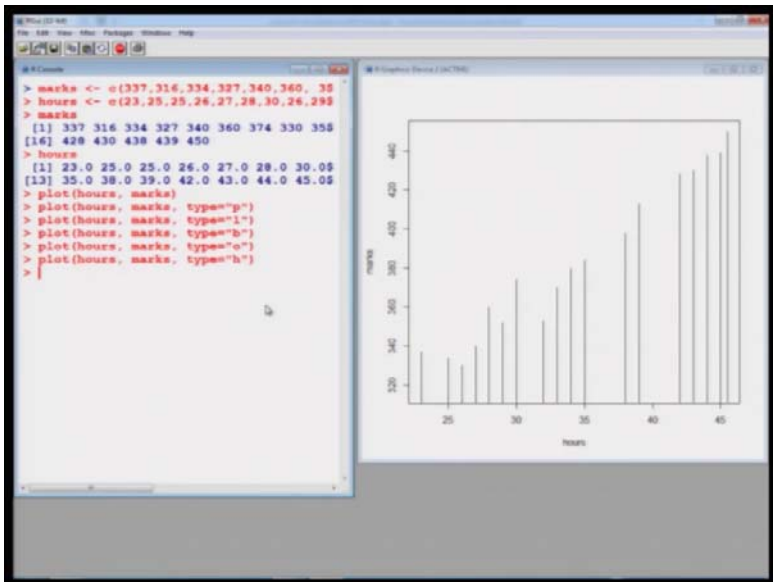
Now, in case if I trying to make it a point as well as here line I use the here option “b” both. So, you can see here the same graphic is changed to both; that means, point and line, but the points and line are not over plotted.

(Refer Slide Time: 14:42)



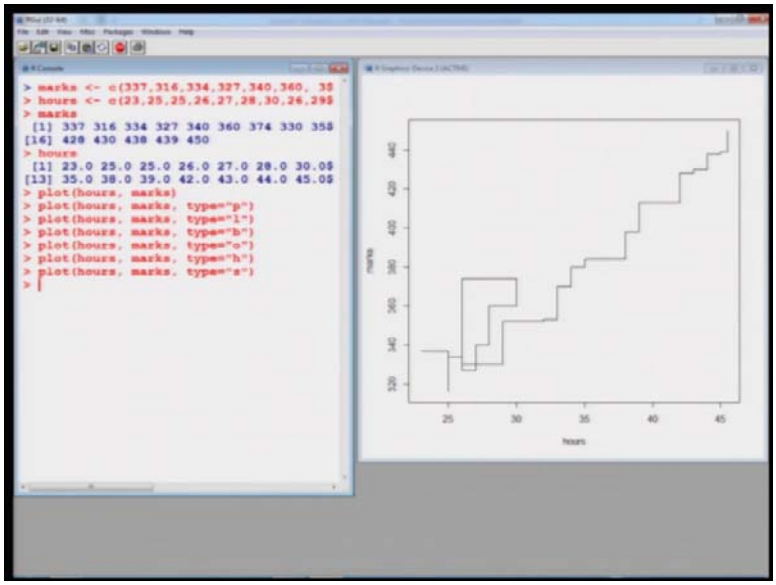
So, if I want to have an over plotted graph, I have to use the command type equal to “o” and now you can see here the same graph is changed to a graph where lines and points they are overlapping, right.

(Refer Slide Time: 15:00)



Similarly, if you want to have here high density lines like as histogram type of plot, you can see here just by changing type equal to “h” you will get this type of graphic.

(Refer Slide Time: 15:12)



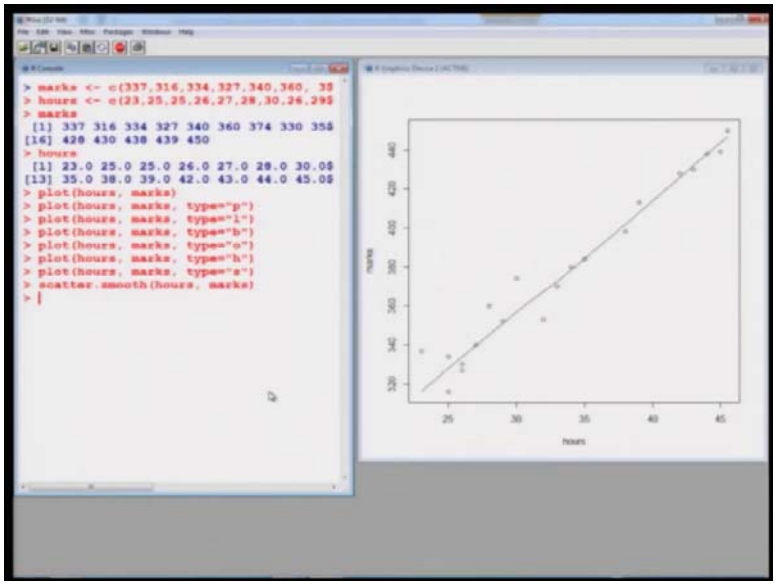
And if you want to have type stairsteps use type equal to “s” and you will get here this type of graphic, right, ok. Now, I try to introduce here one more type of graphic because actually you can see here that when you are trying to plot hours and marks, you can see that that there is a very clear indication that the relationship between hours and marks is approximately linear.

So, now you would like to see what type of curve will be there which can approximate the relationship between marks and hours in a decent way, in a reasonable way, for that we have different types of tools in statistics, say principle of least square regression analysis and there are many things which provide us the graphic on the basis of given set of data.

So, now, I am not going into that detail at this moment, but my objective is that if you want to have a plot where you want to have points as well as you want to have the sort of a smooth curve also which is depicting the trend in the data set then we have a command here scatter dot smooth.

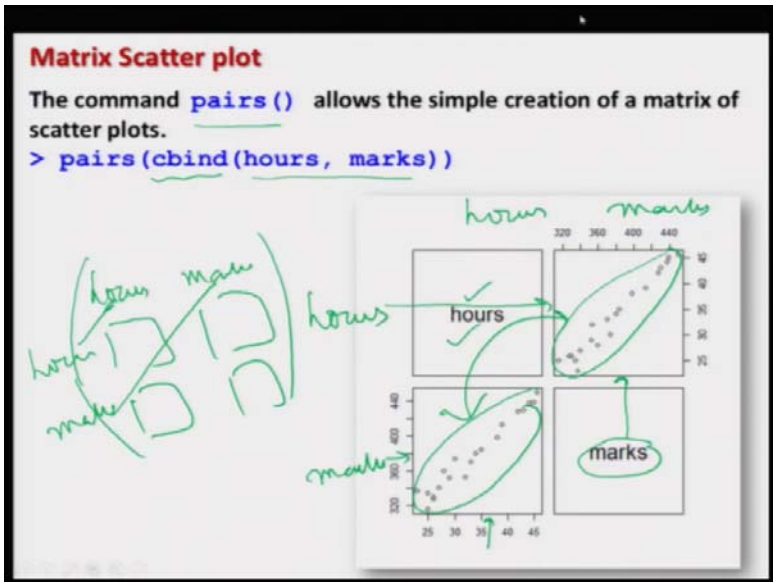
So, the command is scatter dot smooth and inside the parenthesis you have to give the name of the data vector. So, I have used here the data vector hours and marks. So, I will try to use here the command scatter smooth and I will try to show you here how do I get it.

(Refer Slide Time: 16:55)



So, I already have the data. So, you can see here, this is the curve which I am getting here right.

(Refer Slide Time: 17:13)



Now, I try to introduce you another type of curve, whenever you are trying to deal with multivariate, what do you mean by multivariate, usually there are more than one variables which are inter related and they are affecting the process. So, you would always be interested in finding out that whether there is a correlation among the different variables or not.

So, suppose if I have three variables say x, y and z, then there can be correlation between x and y, x and z and y and z and if I have four variables say a b c d then there will be correlation between a b, a c, a d correlation between b c, b d correlation between c d.

So, now, if you try to plot these curves one by one separately then that is not very convenient to do this correlation analysis. But on the other hand, if you have a facility where all those graphics can be created in a single frame, then the understanding becomes more easier.

The data will give you more information which you can analyze more easily and that is more informative, as you can imagine that you have say 5 or 6 curves means lying down side by side and you have to look on each of the curve one by one, but rather there is one graphic on which you can see all the curves at a single time.

So, for that we have got a matrix plot, the concept is matrix plot matrix is something like the graphics will appear in the form of a matrix and all the graphics will appear together. So, in order to create the matrix plot, we have a command here pairs. So, let us try to understand it.

So, the command here is p a i r s pairs. So, this allows us to create the matrix plot, but in order to create the matrix plot, first I have to bind the data. So, for that there is a command here cbind. So, hours and marks they itself they themselves are data vector, but I need to combine them that is the requirement of this command pair. So, I use here the command cbind and then hours and plot.

And if you try to plot it here first I would try to explain you how the things are coming. So, on the x axis this is here say here hours and then here marks and then here hours and then here mark. So, you can see here this is a sort of 2 by 2 matrix where you have here hours, marks and see here hours and here marks.

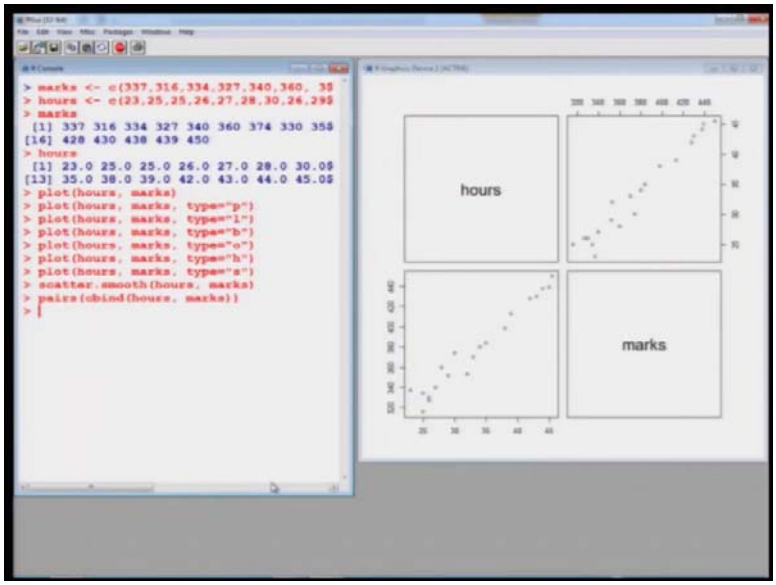
So, now, you will have here four graphics over here. So, four graphics will be here between hours and hours, marks and marks and so on. So, now, so, you can see here this is the graphic between hours and hours which is not interesting that will be a straight line.

And similarly, here this is a graphic between marks and marks. So, this is also not interesting, but this is here a graphic between hours and marks hours is on the y axis and marks is on the x axis. And similarly, this graphic is between marks on the y axis and hours on the x axis. So, you can see actually this graphic and this graphic they are just means transpose of each other if you try to rotate it possibly they will surely they will give you the same thing, right.

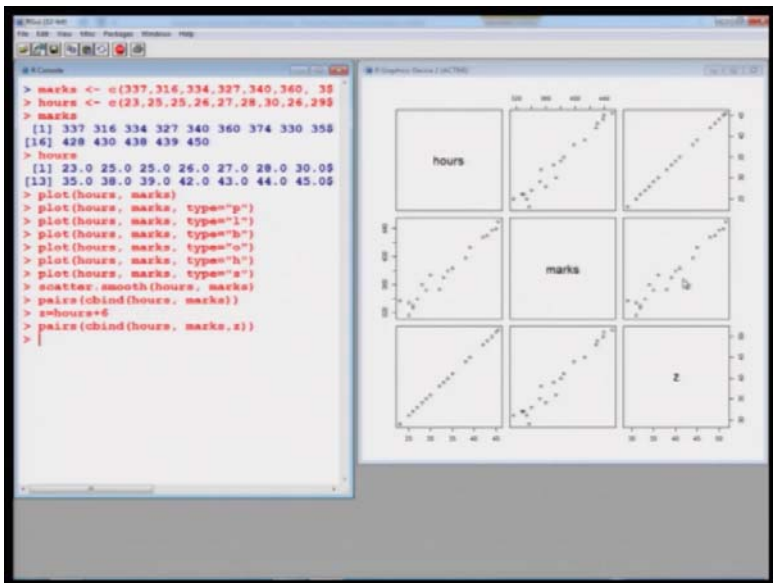
So, but the idea here is that you can see two graphics together on a single plot and this is the idea of this here matrix scatter plot, right. So, let me try to first show you here this matrix

scatter plot and then I will try to give you one more plot. So, if I try to take this command over here and if I execute it on the R console, you can see here, this is the graphic.

(Refer Slide Time: 22:02)



(Refer Slide Time: 22:10)



And suppose if I create here another here say here some data set say hours say here say plus here 6, right and if I try to make a well that is the only hypothetical situation you can see here now there will be three curves. And you can see here that this hours and z they are exactly the same thing there is only a difference of 6 units.

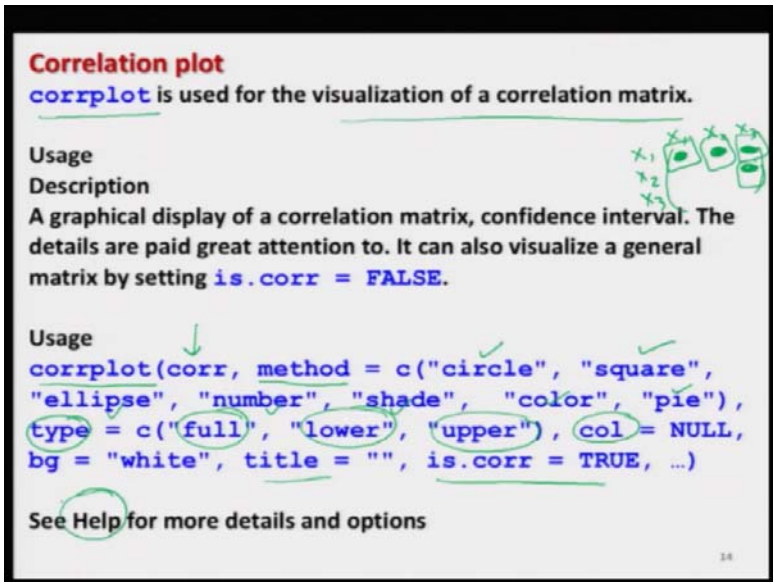
So, this curve here just observe my this cursor this curve is exact line because the correlation coefficient will be exactly equal to 1 and in case if you try to increase the number of variables over there you will get more graphics on a single panel.

Now I try to explain you another type of plot which is correlation plot you know that the correlation coefficient measures the degree of linear relationship between the two variables. And the value of correlation coefficient lies between minus 1 and plus 1, minus sign is indicating the direction of the relationship that if one variable increases another decreases and the positive sign of the correlation coefficient indicates that if one variable increases then the second variable also increases and whatever is the magnitude of this correlation coefficient that indicates the strength of the degree of linear relationship.

For example, if I say correlation coefficient is close to 1; that means, all the points are lying very close to a straight line, but if the value of the correlation coefficient becomes less than the scatteredness in the data around the line becomes more. So, the magnitude of this correlation coefficient will always be between 0 and 1 and the direction will be denoted by plus or minus sign.

So, another concept came that if I can indicate the degree of linear relationship by different types of shades of color or different types of a structure, possibly they will also give the same information about the degree of linear relationship among different variable, but that will be more easy to understand with this objective this correlation plot was created and there is a special package in R which is needed to plot this type of plot. So, first let me try to explain you that how this correlation plot is created and how you are going to interpret it, ok.

(Refer Slide Time: 25:03)



So, in order to create the correlation plot we have a command here c o r r p l o t that is correlation plot and it is used to visualize the correlation matrix. So, correlation matrix is a very simple thing if you say if I have three variables x 1, x 2, x 3 and if I try to write them in the form of a matrix then here this would be the correlation coefficient between first variable first variable, this will be the correlation coefficient between first variable and second variable.

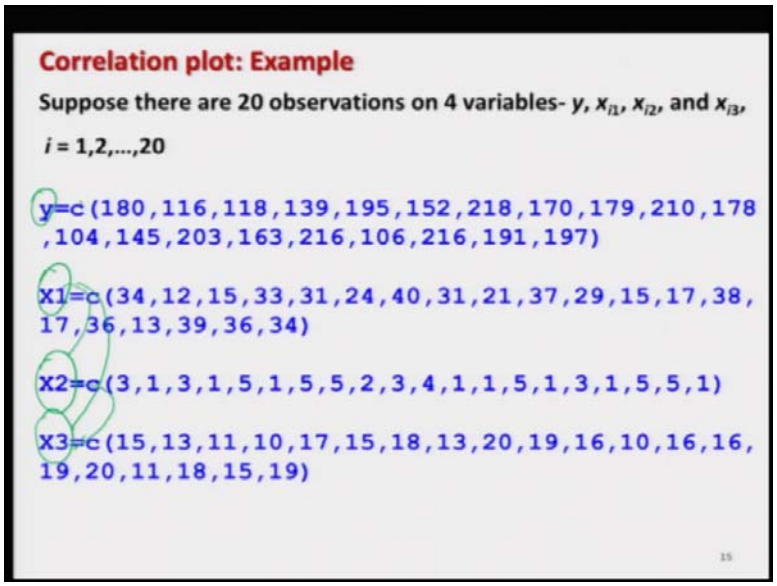
This will be here the correlation coefficient between first and third variable, this will be here the correlation coefficient between second and third variable and so on. So, this matrix essentially tries to indicate in a graphic way. So, all these values they will become a graph, it is something like this, right.

So, there are many many these options which are available over here, but I will try to introduce here some basic options. For example, the command here is c o r r p l o t and then you have to give here the data which you want to plot.

Then there is another option here which is method that you want to indicate the correlation by a circle, by a square, by an ellipse, by number, by shade, by color or by pie chart and so on.

And then there is another option here type that will indicate whether you want the full correlation matrix or lower triangular matrix or say upper triangular matrix and then you will have an option to fill different types of colors and title etc. etc. So, that I would suggest you that you please go to the help of this correlation plot and try to look over there.

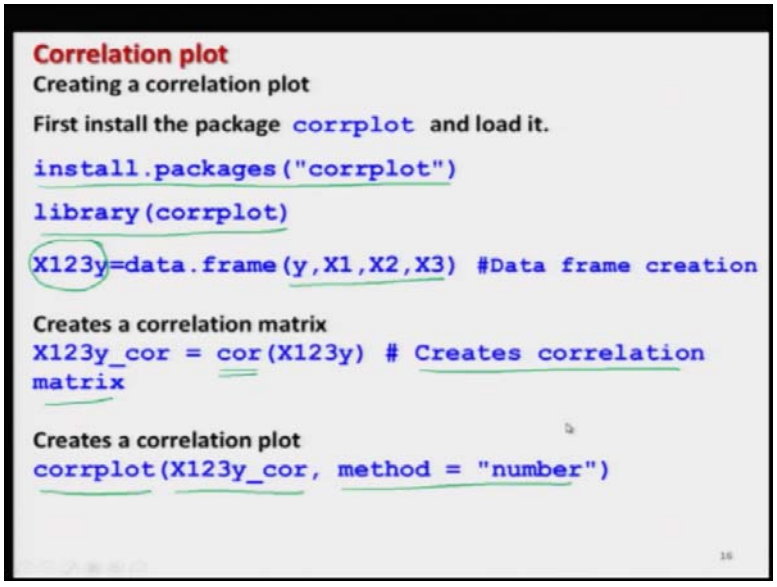
(Refer Slide Time: 26:46)



And I will try to take here a simple example to show you that how the things are happening. Suppose I have just taken say four variables say which are indicated by here y, x 1, x 2 and x 3, right. So, and there are 20 observations and actually they are all paired observations, right.

So, means first observation with first observation of all the variables and so on and I simply want to create here a correlation plot which are going to indicate all type of correlation, correlation coefficient between y and x 1, y and x 2, y and x 3, correlation between x 1 and x 2, x 1 and x 3, correlation between x 2 and x 3 and so on.

(Refer Slide Time: 27:37)

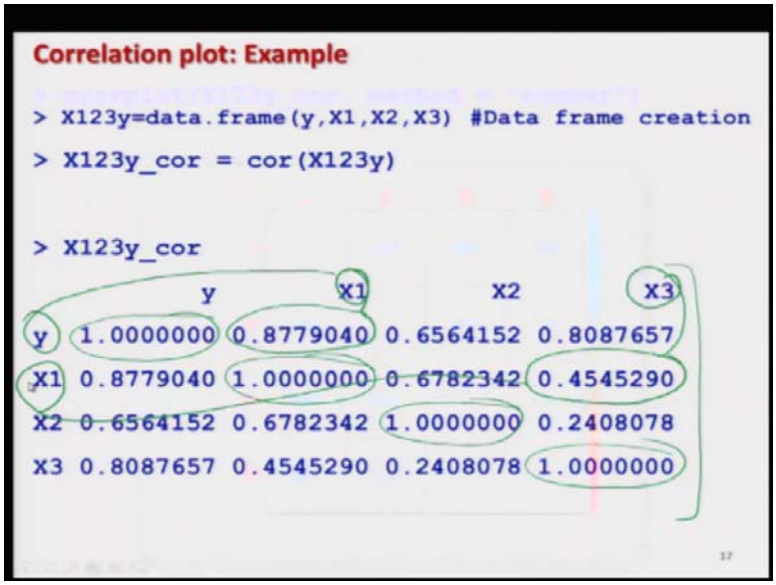


So, for that I have to install a first package because this correlation plot is not available in the base package. So, in order to do it I use the command here install dot packages and I and inside the double quotes. I write the name of the package and if I do it on the R console this will install the "corrplot" package on the R Software on my computer and then I load it by the by using the command library so that is the first step which we have to do.

Now, this corrplot requires the data to be in the format of data frame. So, I try to create here a data frame of the four variables y, x 1, x 2, x 3 using the same command that we just discussed and I store them in a new variable name x 1 2 3 y. So, that is indicating that x 1, x 2, x 3 and y together, right and then I try to create a correlation matrix.

So, once I try to take the data structure in the framework of a data frame then and if I try to find out the correlation coefficient of a data frame, this will create a correlation matrix right. Now, this correlation matrix is plotted under the correlation plot. So, I try to use here the command "corrplot" and if I use here the correlation matrix and method, I am saying here "number" later on I will try to change it and will show you different thing.

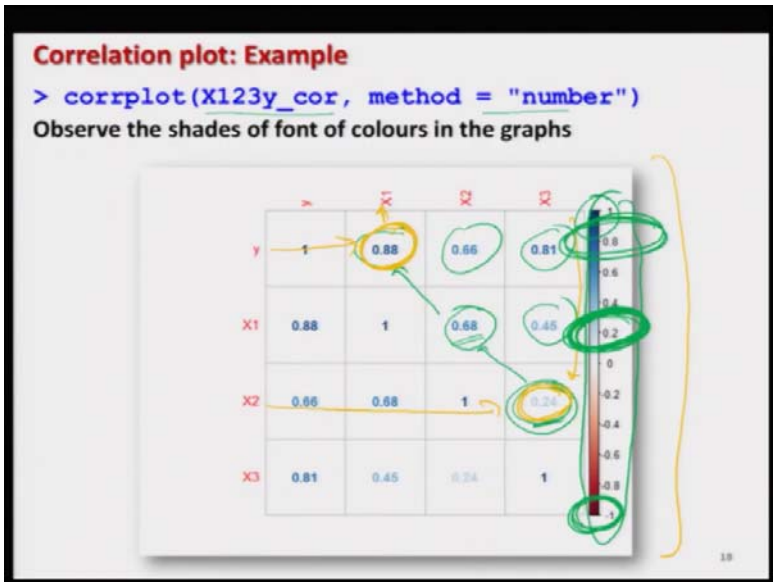
(Refer Slide Time: 29:10)



So, once I try to do it here you can see here the outcome will look like this that x123 correlation matrix will look like this, you can see here that all the elements on the diagonal matrix they are the 1, they are indicating the correlation coefficient between y and y between y and x1, sorry, x1 and x1, x2 and x2 and so on.

And off diagonal elements they are trying to indicate for example, this value is trying to indicate the correlation coefficient between this here x1 and here y and similarly if I try to take here this value over here. So, this is trying to indicate the correlation coefficient between x3 and here this here x1 and so on. So, this is my correlation matrix.

(Refer Slide Time: 29:54)



And if I try to plot it using the command here corrplot, this correlation matrix and method is equal to “number” you can see here you will get this type of picture. Now, first have a look what is this trying to indicate.

You can see here that this color is extremely light and if you move in this direction this color of this number 0.68, this is becoming darker than the number 0.24 and if you move more further, this number here is 0.88 and it is the it is darker than the number 0.68 and 0.24 and similar is the shade of different number over here this and this.

So, one thing what you have to keep in mind here that different shades are indicating the degree of linear relationship. If the value here is 0.24 and if here is the value here is 0.88. So, that is indicating that the degree of linear relationship indicated by 0.88 that is between x1 and y is much more stronger than the correlation coefficient between say here x2 and here x3.

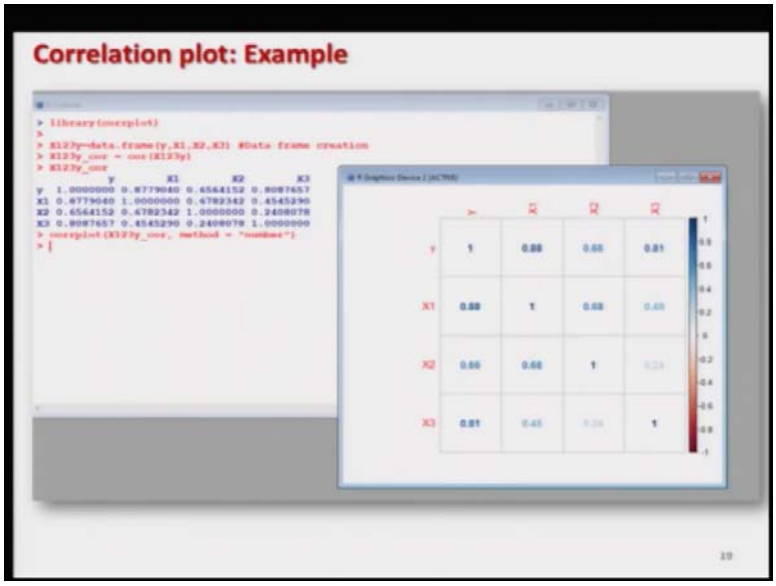
So, this number is indicated by the number which is printed and the shade of the number is indicating the degree of the linear relationship. So, you can see here by looking at this entire graph, you get a very clear cut picture about the behavior of the correlation matrix or the

behavior of different types of correlation coefficient among all the variables and that is the advantage of this thing.

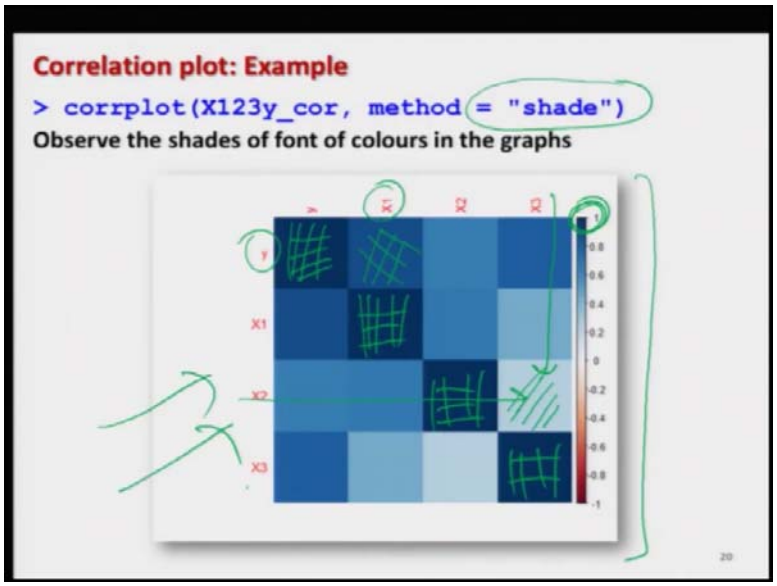
And here also you can see on this scale here, they are also trying to give the scale which is here minus 1 which is the means if a shade of the number is close to red that will indicate a minus, if the shade is close to blue that will indicate a positive relationship and then if the value is close to -0.8 you can see here this is darker and if the value is close to 0.2 you can see my here pen then it is lighter.

So, this is the role of correlation plot that by looking at this matrix you can get an idea about the magnitude as well as direction of different types of correlation coefficients among all type of pairs of variables, right.

(Refer Slide Time: 32:38)



(Refer Slide Time: 32:49)



And you can see here this is the screenshot this is just to assure you that it is obtainable in the R Software. Now I do one thing that in the same command I try to change here the option and I try to use here the method equal to "shade." So, now, you can see here now the number is not there and you get here a picture like this one a graphic like this one. So, where you can see that these are the different shades and this and once again as I explained the shades are trying to give the degree of linear relationship.

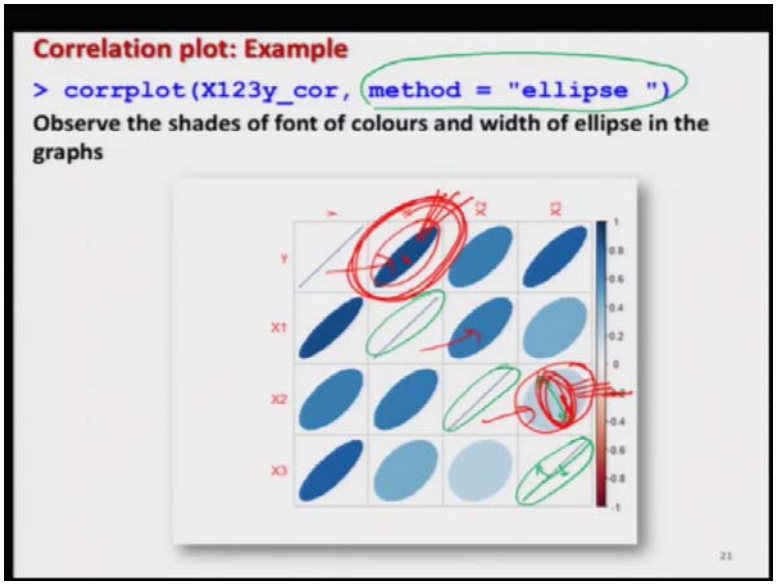
For example if you try to look over here this shade is lighter and this shade is more darker. So, means; obviously, I can say that the correlation coefficient between x 1 and y is stronger than the correlation coefficient between here say x 3 and here x 2 1.

On the other hand if you try to look in the first column like this one or the values on the diagonal elements they are the darkest, why they are the darkest because, they are the correlation coefficient between the same variable which is always equal to 1. So, this is indicated by here +1 sign.

So, you can see here from this graphic you are trying to get a different type of picture which is more compact and it gives you an idea that how the different variables are inter related with respect to the correlation coefficient or with respect to the linear relationship. Well, here in this option, there is no number whereas, in the earlier option you can see there were numbers.

So, that is the difference between the two, but I would suggest you that depending on your need, you have to choose an appropriate method and as I said putting all the graphics in the same report will indicate that possibly you have not understood which graphic has to be used and that is why you are trying to put everything together. So, that is not an intelligent move actually.

(Refer Slide Time: 34:45)



Now, I try to change this option over here and I try to use here another option here method. And I try to put it here method equal to "ellipse." So, now, you can see here those here a squares which over here in this graphic they are now converted into a sort of ellipse and you can see here that there are two aspect, there are two aspects, one is the color and another is the this width of the ellipse, you can see here in these columns the width of the ellipse is minimum because that is trying to show you that the degree of linear relation is perfectly one.

So, the relation correlation coefficient is 1. So, the relationship is perfectly one. Now if you try to move away from this line then this will indicate the degree of linear relationship is decreasing for example, here if you try to observe my pen in this color red, here if you try to compare here and here now this width is here less and this width here is more. So, this is indicating and then also the color of this thing here is lighter blue and color here is darker blue.

So, that is indicating two thing- by this width, it is indicating that the correlation coefficient is lower it means there is a low degree of correlation coefficient whereas, if you try to look over here then the width of the ellipse is less; that means, the correlation coefficient is high and so

the degree of linear relationship between the two corresponding variable is high. Now if you try to compare the shade of this ellipse then this ellipse with the lightest, then this is darker and this is the most darkest.

So, the shades are also indicating the degree of linear relationship means if the shade is light; that means, the degree of linear relationship is say weak. And if the shade is darker that mean the degree of linear relationship is stronger and that is high means close to 1 or something like that. So, I will request you to do it yourself. Up to now I have shown you all the execution, all the execution of commands, online on the R Software.

Now for this correlation plot you try to install the package, you load it using the library command and you have all the commands from the slides you try to execute them. Well, there can be one issue which I have encountered in my experience. That sometime when you try to install a package in R it gives you some strange messages, I do not know why?

Because I am not a computer scientist and I am not a very good computer programmer also. I just survive, but my experience is that if you try to do it in for example, in R studio you try to install the software and you load the software you can do it very easily.

So, my suggestion is that if you face any problem in R sometime there is a problem of version or anything do not lose your heart, do not get disappointed, but try in R studio I am sure that it will work. So, you practice it and I will see you in the next lecture till then good bye.