Descriptive Statistics with R Software
Prof. Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology, Kanpur

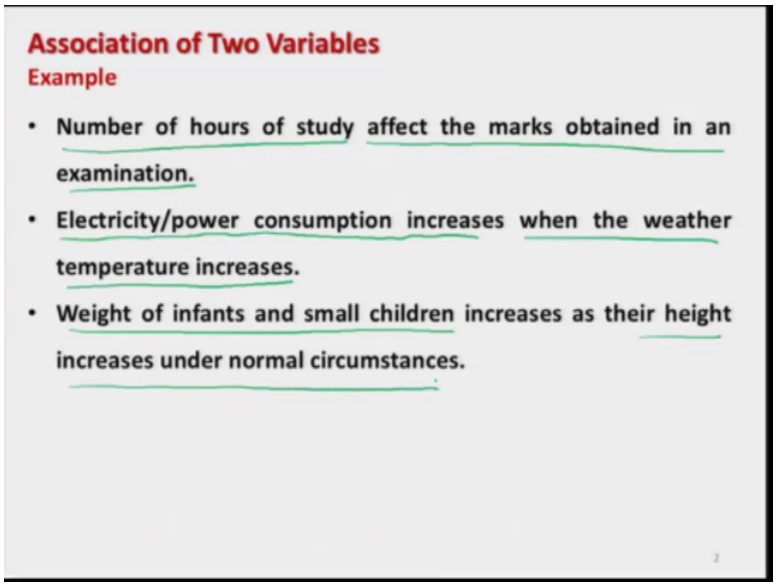
Lecture – 25
Associate of Variables - Univariate and Bivariate Scatter Plots

Welcome to the course on Descriptive Statistics with R Software. You may recall that in the earlier lectures up to now, we have considered the descriptive tools which are used for a univariate setup. Univariate setup means there was only one variable; and we try to visit the measures of central tendency measures of variations etcetera only on one variable. Now what will happen when we have more than one variable?

For example, there can be two variables which are interrelated. So, the question comes that how to know whether the two variables are interrelated or not and in case if they are interrelated how to measure their degree of association. So now, from this lecture we are going to attempt to study the descriptive tools which are used for more than one variables. Then there will be two types of tools: one graphical tools and say another are analytical tools like as quantitative tools.

So, in this lecture we are going to study on say Univariate and Bivariate Scatter Plots which is the graphical procedure.

(Refer Slide Time: 01:42)



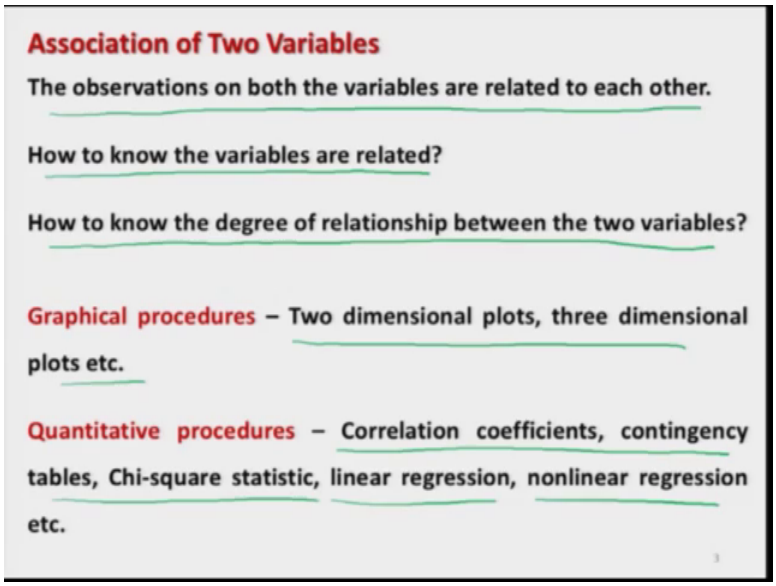
So, now I will get it is a start about discussion here and let me first take few examples. Now for example, we know from our experience that if a student studies for more number of hours usually he will get more marks in the examination. So now, if I try to take these two as variables, the numbers of hours of study and the marks obtained in the examination, then from experience we know that both these variables are associated and they are related.

But you can think that, if you have got a data set on two variables how would you know on the basis of given values that whether the two variables are related or not. And in case if they are related how to show it graphically and how to quantify the degree of dissociation? So, the first example which I have just taken is that, that the number of hours of study they affect the marks obtained in an examination. Similarly we also know that when the weather temperature increases for example, during summer, then we use more electrical appliances like a cooler air conditioner and so on; so the electricity or say power consumption increases.

So, I can say that the two variables power consumption and weather temperature they are related and their tendency is that as the temperature increases the power consumption also increases. Similarly in say another example we know that, weights of infants and small children they increase as the heights of those children increases under normal circumstances right. So, now, my question is this from this data set, from this type of information we have use our experience and based on that we are trying to conclude this things. But my question is this how to do it statistically, how to do it mathematically and more so over how to show it graphically and how to quantify it?

So, now I will be considering the association between two variables. So, I will consider two variables and I will first I to show you that what are the different types of plots available.

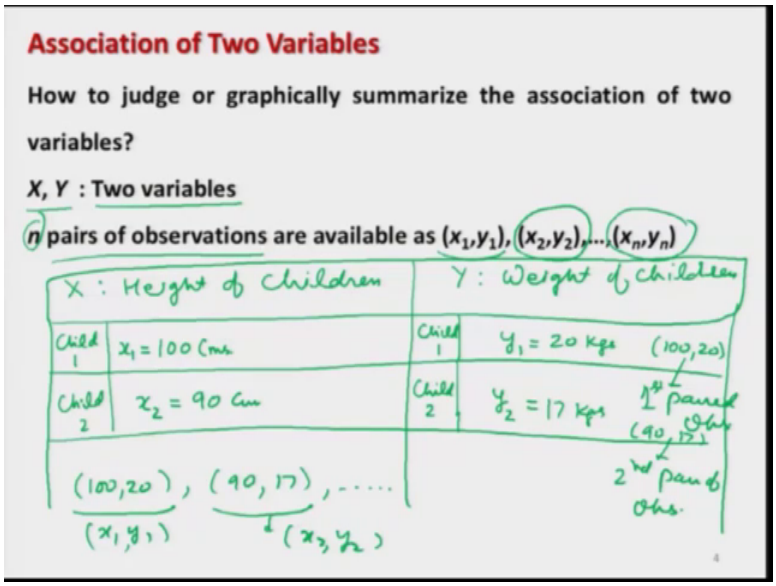
(Refer Slide Time: 04:19)



So, now my question is this that I have got the observations on two variables and both the variables are assumed to be related to each other. So, first question comes how to know that the variables are really related; and if they are related how to know what is the degree of relationship between the two variables. So, there are various graphical procedures like a two dimensional plot, three dimensional plots and so on.

And there are some quantitative procedures also like a correlation coefficients, contingency, tables, chi square statistics, linear regression, non-linear a regression and so on. So, we will try to study these tools one by one.

(Refer Slide Time: 05:02)



So, now first let me try to describe the setup that what are the variables and how the observations have been obtained; and now we are interested in creating the graphs. So, I simply assume here suppose there are two variables and these two variables are denoted by capital X and capital Y and small n number of pairs of observations are available on these two variables.

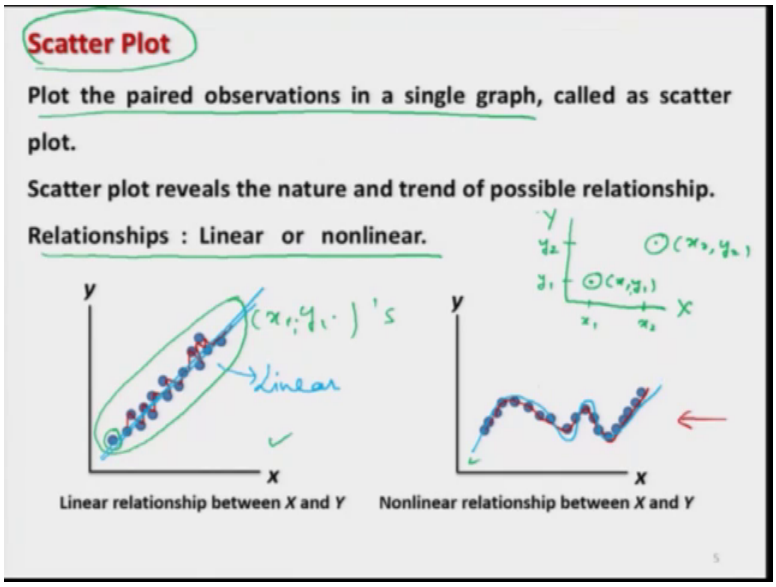
And these observations are they denoted at x 1, y 1 which are occurring together x 2, y 2, which are occurring together and lastly x n y n which are occurring together. What does this mean? Suppose I take here X variable as say here height of children and I try to take here Y to be the say here, weight of children and now I try to collect the observation on these two variables like this. Suppose I take a child and I try to measure his height and suppose this heights comes out to be 100 centimeters.

So, for this child number one the height is coming out to be 100 centimeter. So, I try to denote it say here x 1 is equal to 100 centimeter. And, then I try to find out the weight of the same child, child number 1 and support this weight comes out to be see here 20 kilograms. So, I try to denote this first value as y 1 equal to 20. So now, this x 1 y 1 which is here see here, 100 and 20 this is my first pair observation. And similarly if I try to take second child say child number 2; and if I try to measure the height of this child suppose this comes out be 90 centimeters. So, this is going to denote the height of the

second child height is denoted by capital X. So, I can denote the height of the second child by x 2; and similarly I try to find out the weight of this child.

So, weight is given by Y and its value is given by small y, so I try to write down the y 2 which is indicating the weight of second child and suppose we observe that, this weight is suppose 17 kgs. So now, this x 2 y 2 which is equal to here 90 comma 17 this (Refer Time: 08:21) my second pair of observation and so on, we can say collect more number of observation and this observation will be given as 100 comma 20, 90 comma 17 and so on. So, this will indicate that this is the value of x 1 y 1 and this is the value of x 2 y 2. So, as soon as I write that there are n pairs of observations that mean, these are some numerical values which we have obtained by experimenting the data or by observing the data in any phenomena.

(Refer Slide Time: 08:58)



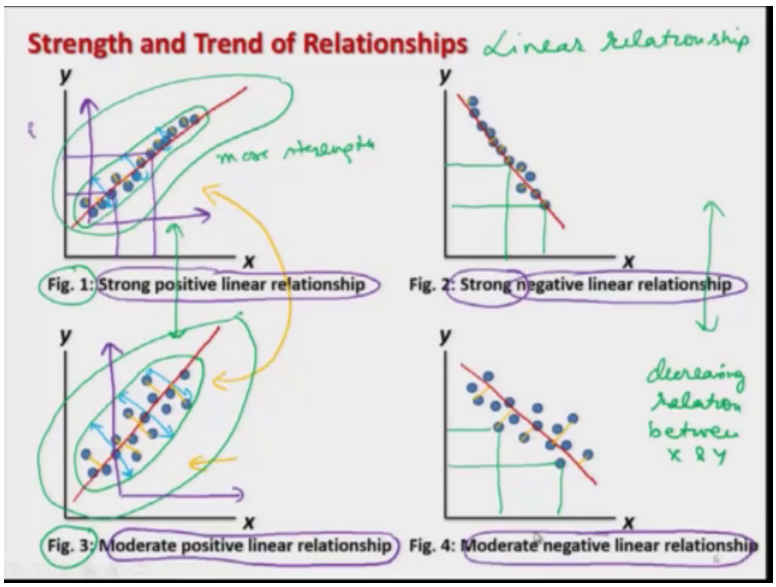
Now, after doing this my first question will be I would like to know, whether there is any relationship between, the two variables or not and that I would like to judge on the basis of given set of data. In order to do this thing we have a plot which is called as a scatter plot and in scatter plot what we do that we try to plot the paired observation in a single graph. How? For example, I have here two variables X and here Y. So, I will try to take the value of here x 1 and here y 1 and I will try this point over here. Similarly, I try to take the value of here x 2 and the value of here y 2 and I try to plot it here.

So, this point is x the point x 1 y 1 and this point is here x 2 y 2, so this is called a scatter plot. Now, this scatter plot can reveal different types of nature and trends of possible relationship. So, for example, we can broadly divided the nature of relationship to be linear or non-linear. So, here you can see that I have just plotted I scatter plot here and here. So, you can see here these circles here they are trying to indicate the value of some observation right. So, all these circles they are essentially trying to plot the values of x is and y is.

But if you try to look at the pattern in this graph, what you see that you can see here that there is a sort of trend mean the trend is that that as the value of X is increasing the value of Y is also increasing and this is happening more. So, over in a linear fashion there is a sort of linear relationship. So, I can say here that by looking at the scattered diagram I can conclude that the relationship in this case is linear; that means, there exist a linear relationship between X and Y.

Similarly, if you try to look into this figures second figure, you can see here that once again these points are plotted here, but you can see here that this points are something like they are not actually linear they are not showing that there is a really a linear trend, but the trend is something like this. So, this is indicating that the relationship between X and Y is non-linear right.

(Refer Slide Time: 11:47)



After this once we have a decided that the relationship is linear or say non-linear then how to see whether the strength of the relationship is say more or less. So, we are going to now consider here only the linear relationship. And, similar type of conclusion will also be there for the non-linear relationship, but I will not consider it here in this lecture. So, if you try to see here in these two graphics, figure number 1 and figure number 3 on the left hand side.

In case if you try to see here in the figure number 1 these points are concentrated inside this band; and in figure number 3 the points are concentrated in this band. And now if you try to observe the width of this band; and compare it with the band width of figure number 1 and you can see here that in case of figure number 3, the observations are scattered more than in the case of figure number 1, but in both the cases you can see that the trend is almost the linear which I am denoting with the red colour. In both the cases you can see here that the trend is nearly linear.

But what is happening in this trend, in this figure number 1 you can see here that those points are very close to the line. All the scatter point they are line very close to the line in red colour. Similarly in the case of figure 3 here, if you try to see the points are lying quite far away from the line you can see here in the orange line I am trying to denote the deviations. And when I try to compare these deviations of observation from the trend line or the red colour line in figure number 3 and figure number 1, I can say here that the strength of the linear relationship in figure number 1 is more.

Why? Because in figure number 1 here the points are lying more close to the line and in figure number 3 the points are quite away from the line and comparison to figure number 1. Now, there is another thing which you have to observe, in figure number 1 and say figure number 3 now try to observe my line in purple colour; you can see here as the values of X are increasing the values of Y's are also increasing. And, the same thing is happening in figure number 3 also as the values of X are increasing the values of Y s are also increasing. You can see here in figure number 1 this is my here X this is my here Y and now if I try to take another X here this is my here another Y right.

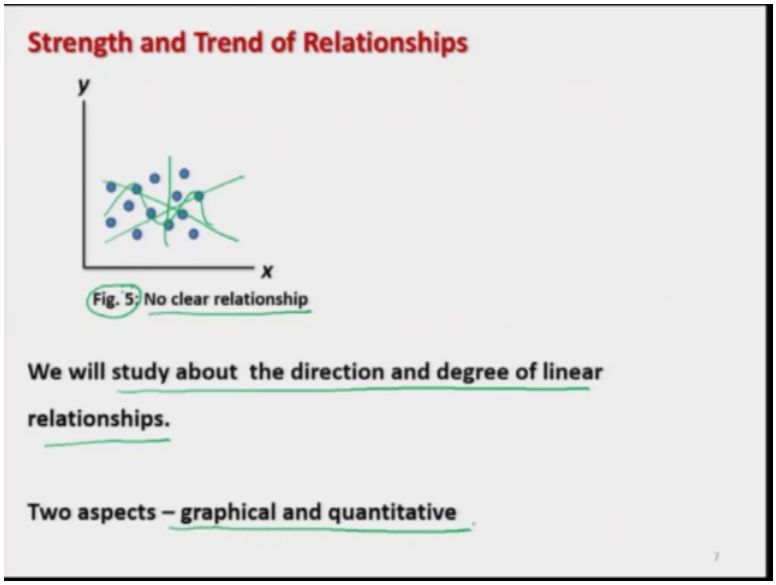
So, this is indicating that the relationship is increasing or we call is positive. So, this is what I mean which I have written in the title strong positive linear relationship; that means, the relationship is linear I am this is the strongly positive in comparison to the

relationship in figure number 3, where I am saying that the relationship is positive, but it is moderate positive and it is linear. Similarly, now incase if you try to observe in figure number 2 and figure number 4, here you can see here in this case as the values of X are increasing the values of Y's are decreasing this is happening in figure number 2 and the same thing is happening in figure number 4.

So, in this case I can say that the relationship is decreasing relation between say here X n Y n both the cases figure 2 and figure 4. Now, in case if you try to create here are line that is called trend line you can see here this is like this and in the figure number 4 this is like this now, if you try to analyze what are the deviations of individual observation from this line. So, if you first observe in figure number 4 these points are lying quite away from the trend line you can see here in comparison to the figure number 2 because in figure number 2. Now, if you observe these points are very very close to the line in comparison to figure number 4.

So, I can say now here that in figure number 2 the relationship between X and Y is quite strong and since it is decreasing relationship. So, we call it as a negative linear relationship, because the relationship is linear; and similarly in the case of figure 4 I will call it as a moderate negative linear relationship.

(Refer Slide Time: 17:44)

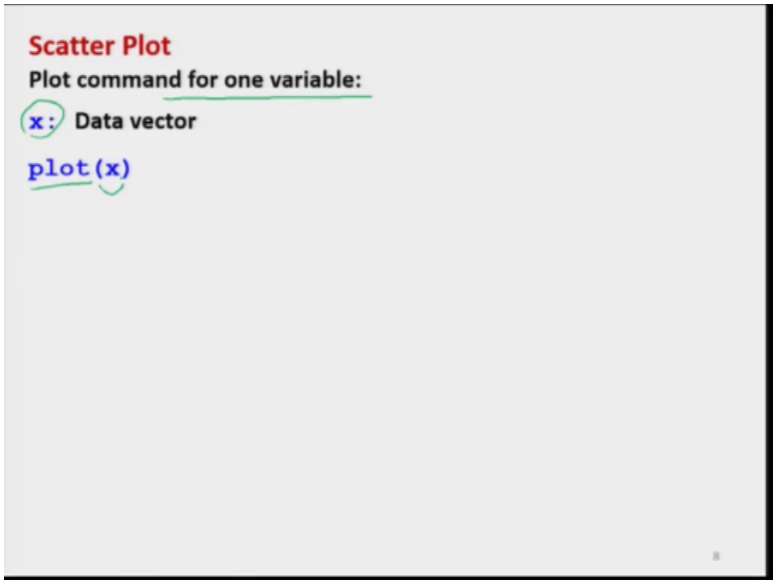


And, similarly incase if you try to plot X and Y and if you get no clear relationship for example, it is happening here in figure number 5. For example, you cannot say here

whether there is an increasing trend or decreasing trend or say varies the trend something like this.

So, in this case by looking at the scatter plot, I can see here there is no clear relationship and even we do not know whether it is linear or non-linear or even this is positive or negative. So, now, in this lecture we are going to study the aspect of linear relationship. So, we will assume that whatever things we are going to do in those cases the relationship between X and Y is linear and there will be two expects say graphical and say quantitative right.

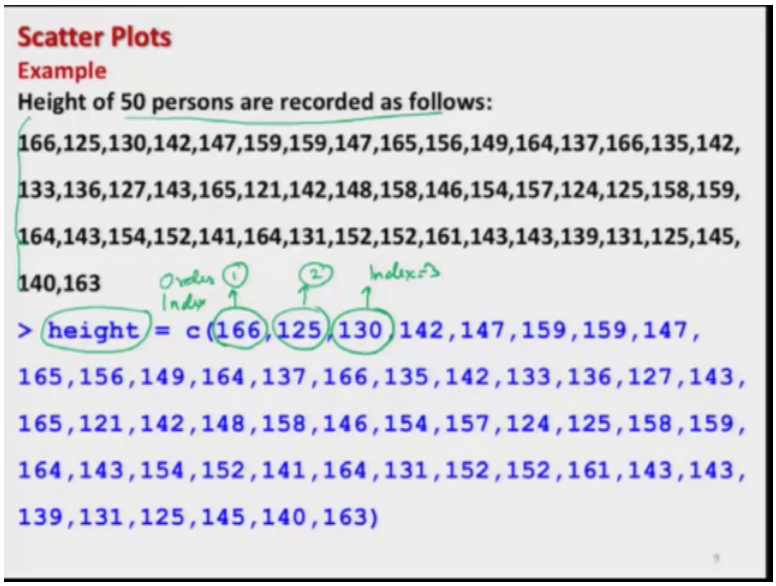
(Refer Slide Time: 18:37)



So, now what I will do that, I will try to take some examples and I will show you that what are the commands on the R software; and I will also show you that how to execute them and how to understand the outcome. First I am going to discuss here the command which is plot command, one thing which I would like to make clear here that this plot command can be used to create the scatter diagram in univariate as well as bivariate set up I had not covered it when I covered the univariate graphics.

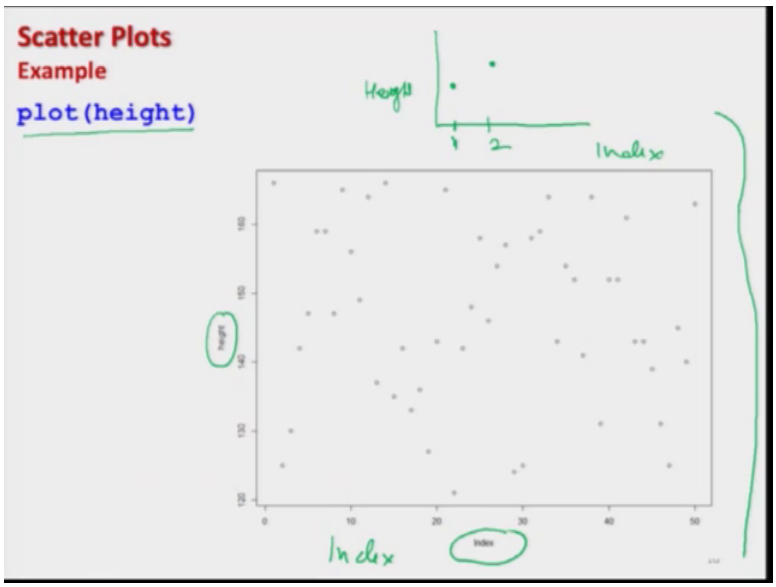
Because I knew that I am going to cover the topic on plot. So, why not to cover it together right? So, increase if you have only one variable univariate case then the data on that variable is stored in a data vector x and then the R command is plot p l o t and inside the argument you have to write the data vector x ok.

(Refer Slide Time: 19:49)



Now, I try to take simple example and I try to show you here that how the graph will look like. So, I have once again taken the same example which I considered in the earlier slide that, we have collected the data on height of 50 persons and which has been recorded here and this data has been stored inside a variable whose name has been given as height like this.

(Refer Slide Time: 20:16)



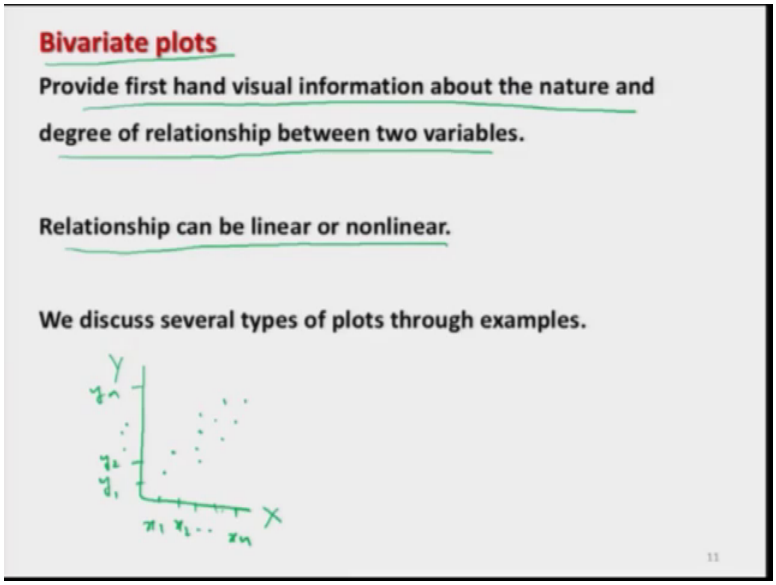
Now, I would like to plot this data. So, I give the command here plot and inside the argument height; and you will see here you will get here this type of outcome on the arc

console. And my objective is not to show you here is the figure, but I want to show you that how to interpret it. You can see here on the x axis this is giving me the index what is the index. For example, if you try to see in this data set what is my first observation? This is my first observation. So, this is my here second observation.

So, this index is trying to give the order in which the data has been given. So, this was observation will have index 1, second observation will have index 2, third observation one 130 will have index equal to 3; and in this plot what they have done? They have taken index here they try to take the value of index number one and whatever is the value here of the x data for example, here is this is height they will plot it here they will try to take the index number two they will plot the data here somewhere we were it lies.

So, here you can see here on the y axis we have here the height. So, this is only I scatter diagram of one variable. So, from here actually you can have the information on say central tendency or say dispersion if the data is more consternated all it is more concentrated around a particular value. So, all the types of measures of central tendency and dispersions can be viewed from this type of graph ok.

(Refer Slide Time: 22:10)

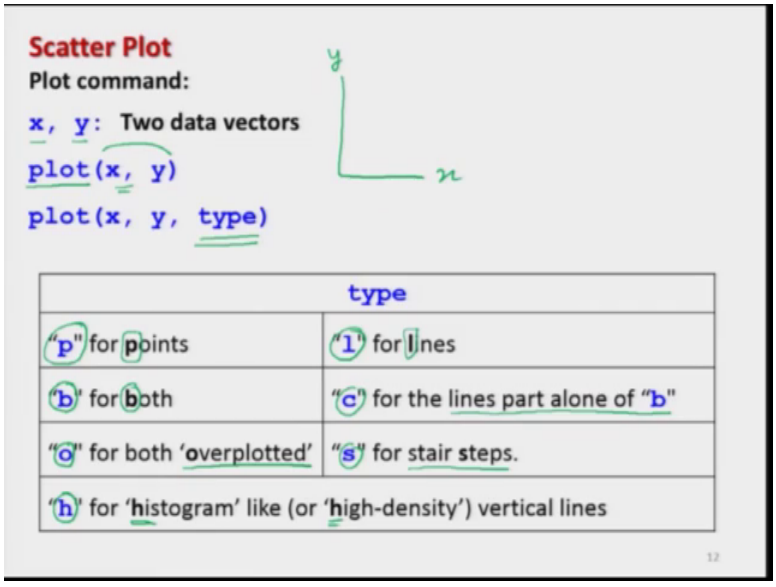


Now, I come to bivariate plots. So, bivariate plot mean there are two variables and these are the plot which gives us. The first and visual information about the nature and degree of relationship between the two variables; whether the two variables are related or not and if they are related whether their relationship is linear or say non-linear. So, all this

type of information can be used by these bivariate plots. So, in bivariate plots what we try to do we will take two variable here X which will be plotted on X axis and say values of Y variable which will plotted on Y axis and then whatever are the values of say here X 1 X 2 and so on, here X n and similarly on the y axis all the values y 1 y 2.

See here y n they will be plotted here like this and they will try to show you the trend and degree of relationship. So, we will try to take some examples and we will try to see that what are the commands and how they have to be executed.

(Refer Slide Time: 23:16)



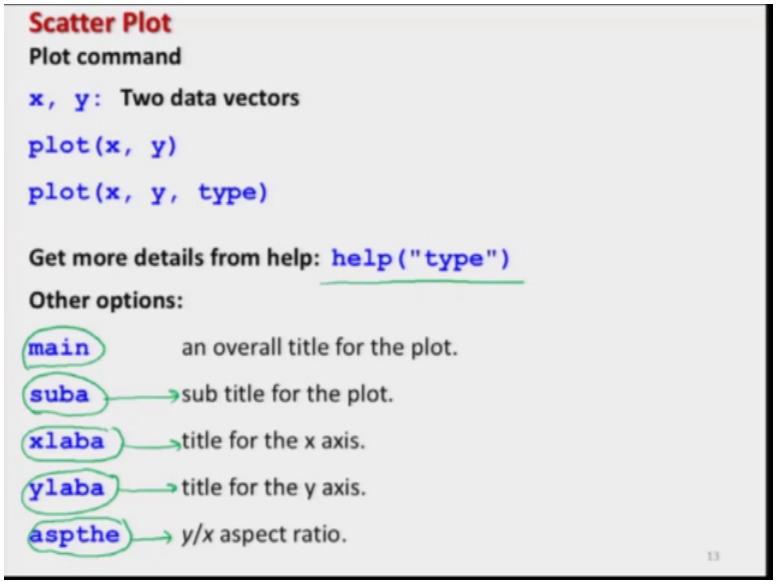
So, now in order to create a bivariate plot, suppose we have two variable which I am decorating by small x and small y in which the data of these two variables have been stored. Now in case if I want to make a plot like on x axis I will have the value of x variable and on y axis, I will have the value of the y variable this can be executed by the command plot. But now inside the arguments I have to give the variables x and y separated by comma. So, this is how we have to do now there is another option available in this.

Command which is here type just by giving various values to type we can create different types of graphics. For example, if I said type equal to p then it will give me only the point which is the default value and if I say here p is equal to l this is going to give me the lines. So, you can see here I have made this p and here l to be bold which is indicating that what is the meaning so you can easily remember also. And, if you want

this point a line to both to be present, then I have to use the type equal to b, which is coming from both. And similarly if you try to use that type equal to c then I will get the lines which are only part of the b because then we have the points and lines both.

And, similarly if I try to use here the command o then I will get a over plotted means the point and line are both are over plotted inside the graph. Similarly, if I try to take the type equal to s then I will get a graph which will look like a stair steps. And finally, if I choose type equal to h, then this will give me a lines which will look like as if I have created the histograms of this data or there will be some high density vertical lines.

(Refer Slide Time: 25:38)

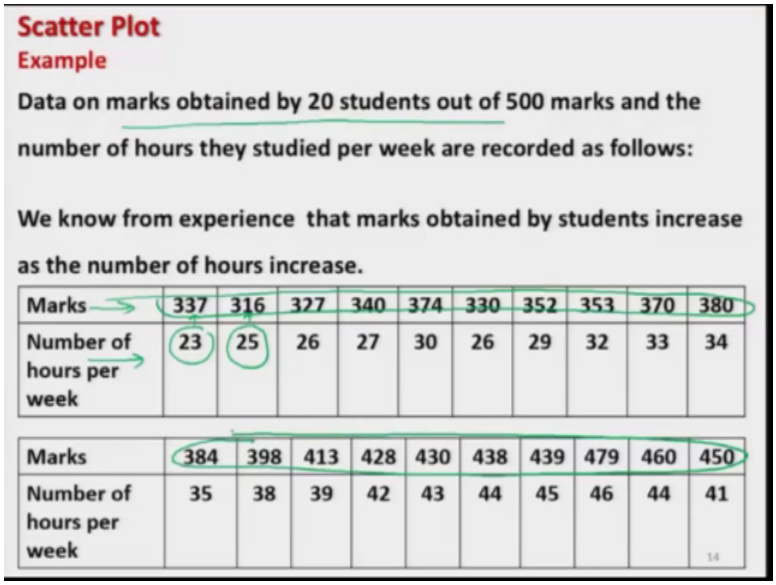


So, now let me try to take an example and I try to show you that how these different types of graphical look like, but before that there are some more options which are available in the plot command. For example, in case if you want to give a title then you have to use the main inside the argument and you have to just give whatever you want to give as a title of the plot. Similarly, if you want to give us subtitle of the plot, then you have to use the command suba and if you want to give the title on the x axis then the command here is x laba.

And similarly for y axis title the command is y lab a, and if you want to maintain the aspect ratio that mean how the graph will look like should it be more discharge on in the x direction or in the y direction; then this is given by the option aspthe. And, if you give a numerical value to this thing this will maintain the aspect ratio and definitely if you want

to have more information I will suggest you that, please try to look into the help of this type command or say help on plot command right.

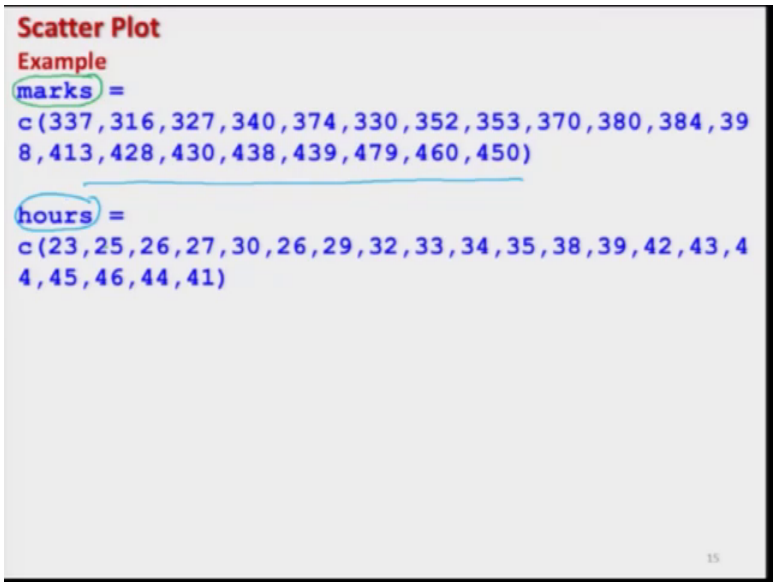
(Refer Slide Time: 26:40)



Now, I will take a simple example to demonstrate how to use this command and how the graphs will look like. As we have discussed that we know that number of ours of study of a students this affects the marks obtained examination. So, the number of hours of study and the marks obtained in the examination they are related, but this is from my experience. So, we have collected the data on 20 students that with that how many hours every weekday have studied and finally, how many marks they have obtained in the examination out of 500.

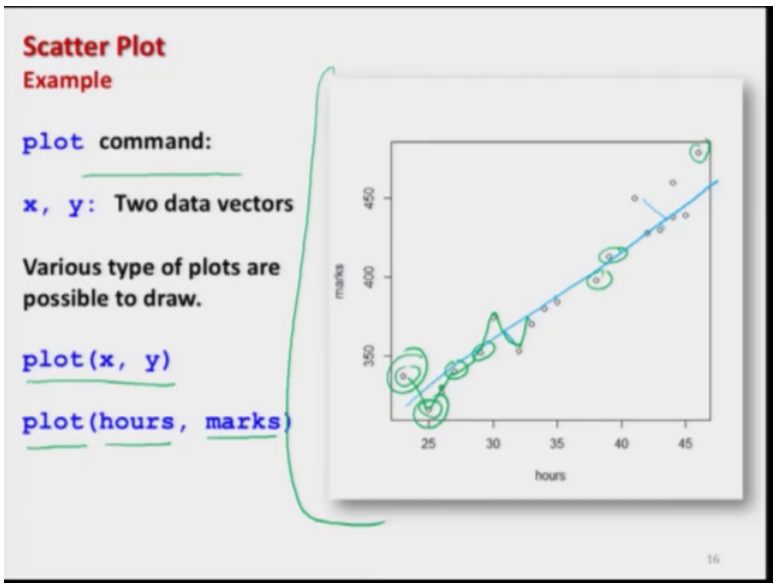
And those marks and the number of hours of these 20 students have been recorded as follows that for example, if a student has studied 23 hours per week then he has got 337 marks; if a student has studied 25 hours every week then he has got 316 marks and so on. So, the first row is denoting the marks and second row denoting the number of hours per week which a student has studied and this data for 20 student has been obtained here.

(Refer Slide Time: 27:53)



So, now what I do here first let me try to compile all the data into data vectors. So, I try to create a data vector here marks in which I tried to store all the observation which are given in the first row here, this and here this and which is here. And, now in the second case I have taken variable here hours in which I have stored the data which is given in the second row here and I have created here two data vectors, marks and hours as I will be using this example again and again. So now, I have explain you the genesis of this data set.

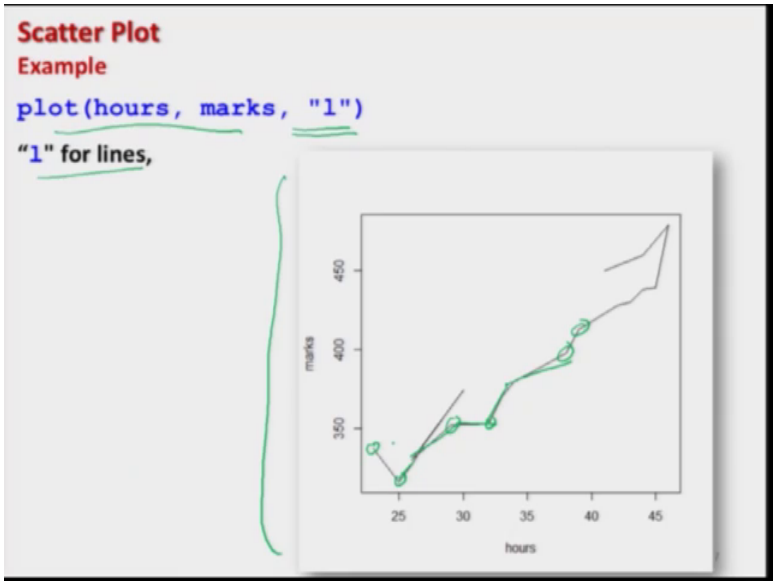
(Refer Slide Time: 28:44)



Now, I would try to simply make a plot of these two data vectors hours and marks. So, I will be using here the plot commands for the bivariate plots which was plot xy. So, now I will say here plot inside the arguments I will try to give that two data vectors hours and marks separated by comma and this will give me here the graphic like this one. So, you can see here that these are the data points and by looking at these data points you can see here the that, there is going to be a sort of linear trend; trend means that most of the points are going to lie nearly on a straight line.

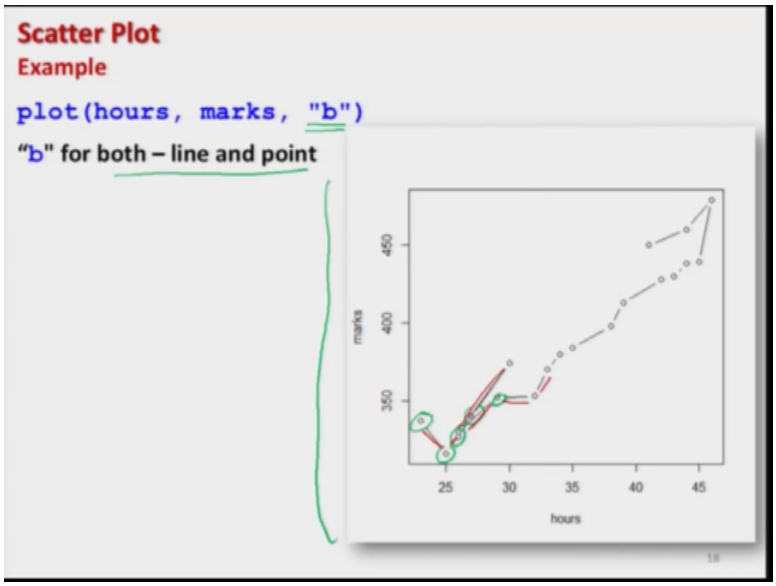
Now, in case these points are line close to the line or away from line, that will give some idea about the extent of the relationship between marks obtained in the examination and number of hours studied right.

(Refer Slide Time: 29:46)



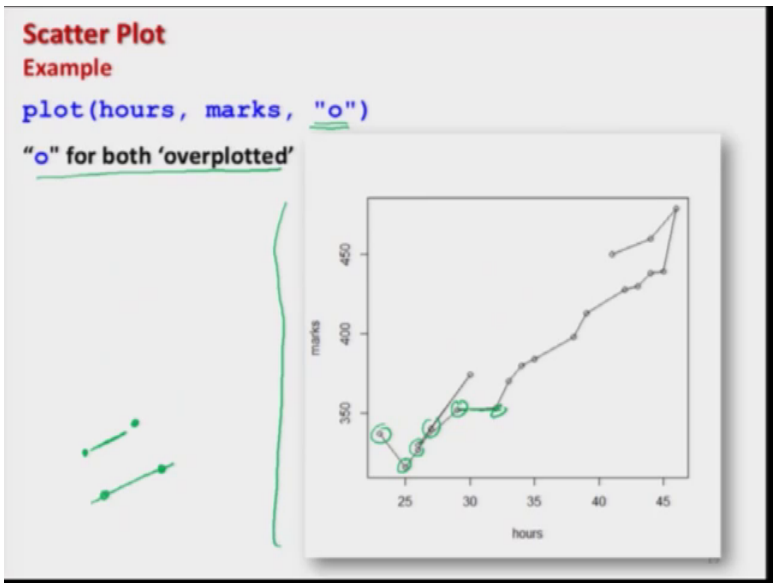
Now, in case if I want to change this type then now I am going to use here a type which is here l if you remember this l was used as the type to create the graph with lines. So, now you can see here once I try to do it here I get here a graph like this one in which all those data points like this one here and they are inter connected by such lines and so on. And if you wish you can compare it from this curve, if you try to make it here this, this, this, this and so on connected them by line that will give you a line plot.

(Refer Slide Time: 30:33)



Similarly, if you try to choose here the option for typical to b; b means both line and point. So, you can see here this is the combination of the first two graph that we have obtained here those points are here; all those points are here and they are connected by these lines like this, this, this, this, this and so on. So, this is how you can obtain this type of plot.

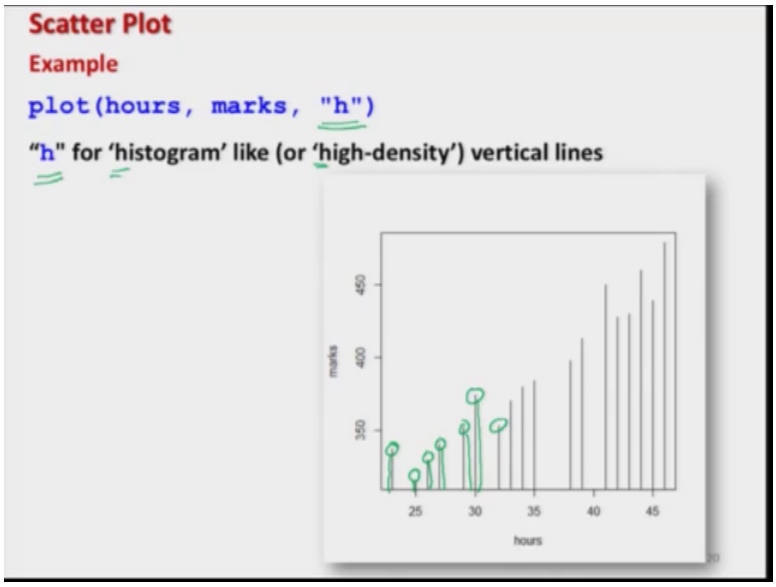
(Refer Slide Time: 31:07)



Similarly, if I try to use the option of type equal to o, o means over plotted what you mean by over plotted? In the earlier graph if there are two points here like this there were

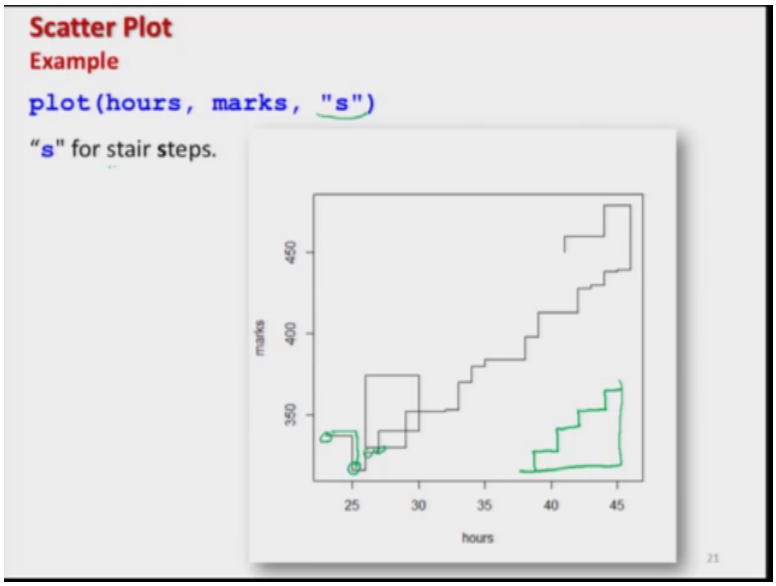
simply joined like this, but now I am saying that there are two points like this and they and the line will cross through like this one. So, you can see here these are the points and. So, on which I just joined from point to point. So, the line and those dots both are over plotted. So, this type of graphs will again give us a different type of information.

(Refer Slide Time: 31:49)



Now, in case if I use the type h. So, we have discussed this h means this is a sort of histogram or say high density vertical lines. So, we can see here we have this data here like this and this data has been join on the x axis like here this. So, this is another type of plot which can be obtained by using the type h.

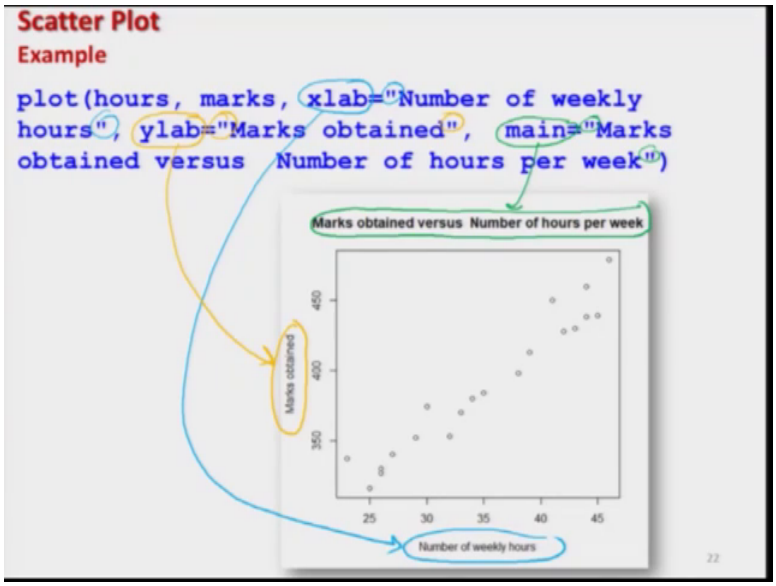
(Refer Slide Time: 32:16)



And similarly if you want to create a stairs steps type of plot then use that type equal to s; and if you see the plot, which we have obtained here these are my data points and so on, you can see here first the that what we are my data points in the first curve.

You can see here this point is here like this and this and the same thing is here like this the points are here and here and now they have joined by see stairs type of plot. And, similarly you can see here these points are going like a steps means you have seen that when we climb on the roof then there are stairs like this one. So, that is why this plot is called as a stair step type plot.

(Refer Slide Time: 33:01)

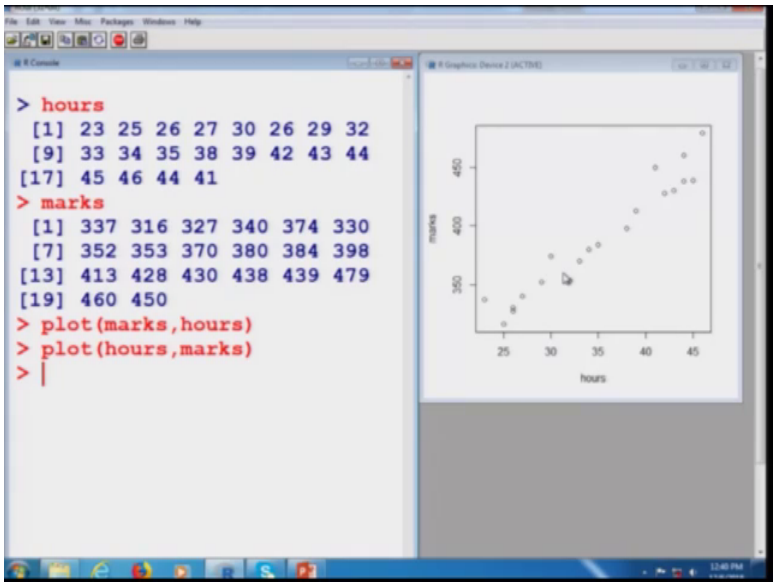


And, now in case if you want to make it more informative, suppose you want to add here the title of the plot titles on the x and y axis. So, you have to use different options and if you remember we have discuss these things in more detail when we have discussed the graphics in the case of univariate data. So, for example, now exactly on the same way suppose I want to give here a title like marks obtained versus number of hours per week.

So, we have discussed that this can be given by the command main. So, I tried to write here is the main is equal to the title inside the double quotes. And similarly if you want to give here title on the x-axis, suppose I want to give number of weekly hours. So, then in order to do this thing we have the command x lab. So, I tried to give whatever the title I want to give on x axis by writing x lab is equal to the title inside the double quotes and this gives me here an outcome like this.

Similarly, incase if you want to have a title on the y axis suppose I have want to give here marks obtained then for that we have the command y lab. So, I write y lab is equal to the title which I want to give inside the double quotes and this outcome is given over here. And, similarly there are some other options which I would say that please try to look into the help menu and try to see here. So now, I will try to show you that how these graphics using the plot command or constructed inside the R software.

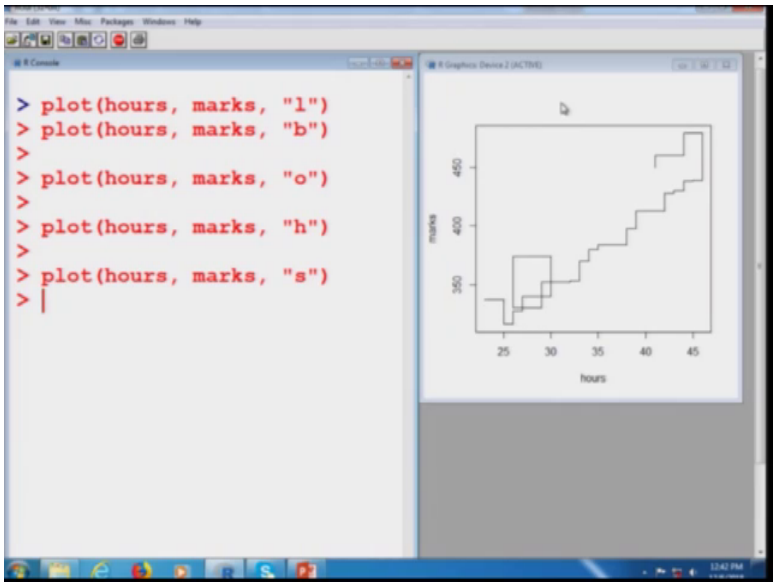
(Refer Slide Time: 34:48)



So, let us come to our software part. So, you can see here that I already have stored the data h o u r s hours and on the marks, we can see here like this, this is the same data that we have obtained in the example you can verify it. Now I am trying to see here plot between see here marks and see here hours and as soon as you enter it you can you get here this type of graphic right. So, you can see here this is the same graphic that we have obtained.

Now, in case if you simply try to change the order of the variable here you can see here at this moment marks are coming on x axis and hours are coming on y axis. Now, suppose I want to interchange with in place of mark I will give hours and in place of hours I will give the mark. So, I will type it plot hours and marks and now as soon as I enter you can see here that this graphic is changed you can see like this.

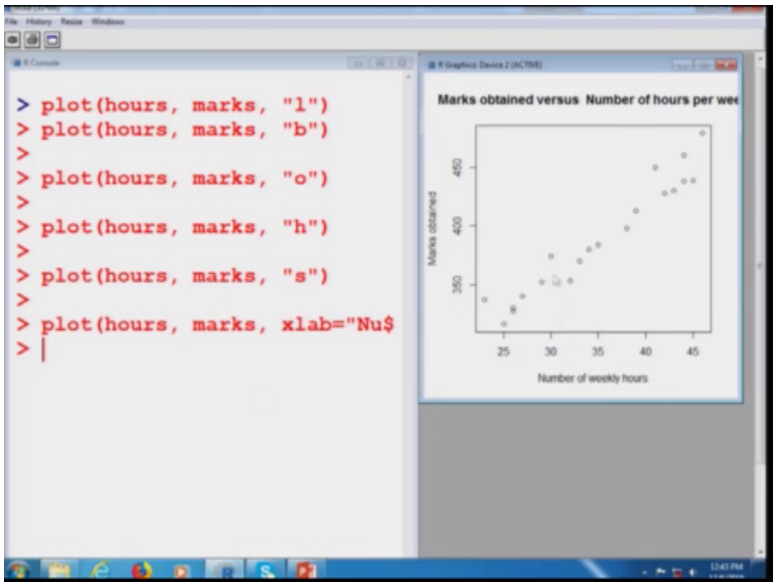
(Refer Slide Time: 36:02)



So, now I will continue with the same command and I will clear the screen, but I will try to show you that what are the different options which you can use using the type option. So, if you see here this was our first option was type equal to l by which I will get the lines. So now, you can see that as soon as a press enter this graphic will change you can see here this is now l and suppose if you want to have line and points both then I have to give the type equal to b, so you can you see here you get a different graphic. And, similarly if you want to use the over plotted option by choosing the type equal to o small o then as soon as I press enter I get this curve this is the same graphic which we have seen inside the slides.

And similarly if you want to have a histogram or say high density lines then using the type equal to h we get here this type of command we can see here. And finally, if you want to have still type graphic then I have to use the type equal to s and as soon as I do. So, I will get here an graphic like this one, so you have now seen that creating graphic is not difficult at all. Now, suppose if you want to add some more features like as there are some default features that whatever is the name of the variable that will be coming on the x and y axis, but suppose if you want to add titles on x axis title on y axis and main title then you have to use the same command. So, you see here on my slides that I had shown you here.

(Refer Slide Time: 38:02)



Now, I will try to copy this command and I will try to paste it on the R console. You can see here this is coming like this; and now you can see here as soon as I have press the enter this is graphics changed now you have the titles on x y axis as well as the other titles. And similarly in case if you want to change the colours this is also possible using the options, but definitely I have show you the basic operation you can make it as beautiful as possible as informative as possible whatever you want.

So, now I would stop in this lecture and in this lecture, I have given you an idea of the plot command. I am not saying at all that which of the command or which of the graphic or which of the type is the best it depends only on you people or the experimenter. The experimenter has to decide or you people have to decide what type of information you want from the graphic and which is the graphic which is more suitable to provide that information in the correct way? And, this also comes by practice and experience.

So, at this moment the objective is to learn how to create graphics and how to make them more informative more interactive using the R commands. So, you please take some data set try to experiment on it, try to give different types of option, try to create the plots and try to see what type of information they are going to provide try to take different types of types.

And, you practice and we will see you in the next lecture till then good bye.