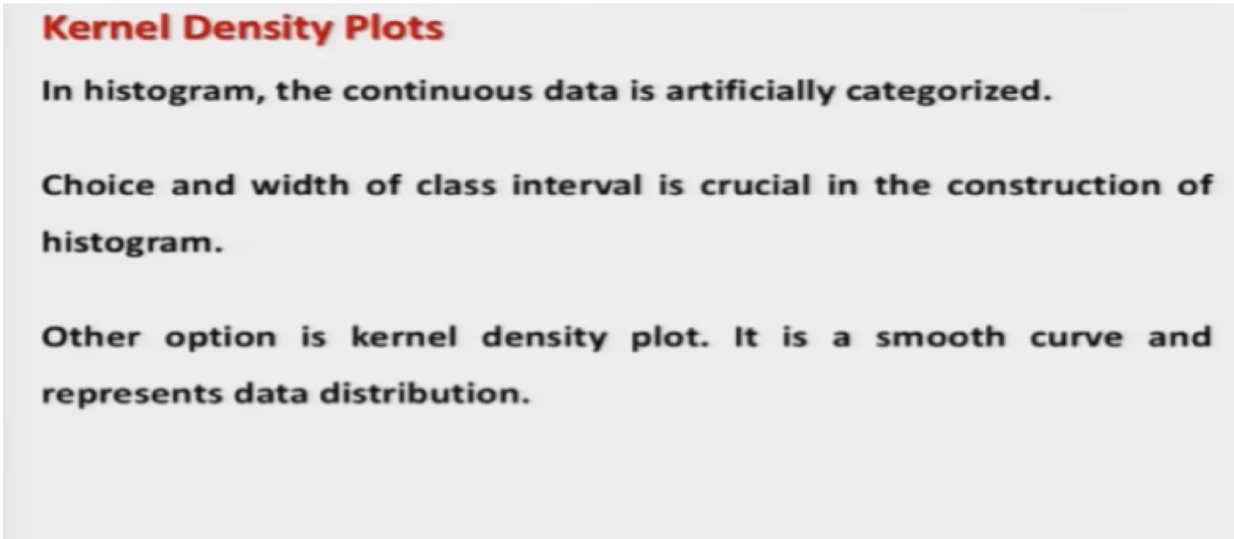
## Lecture – 13

## Graphics and Plots - Kernel Density and Stem-Leaf Plots

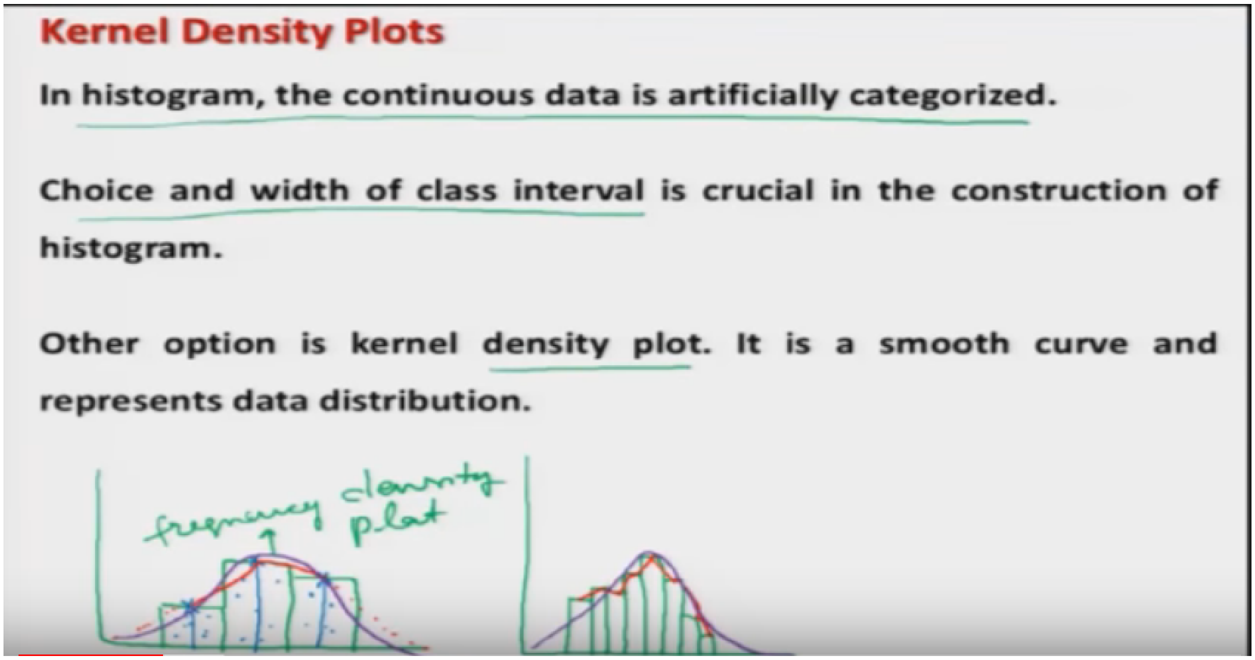
Welcome to the, lecture on the descriptive statistics with R software. You may recall that, in the last lecture, we had a discussion on the construction of histogram. Now, in this lecture we will Continue on the same idea and we will try to discuss, kernel density plot. And after that, I will take up the stem-and-leaf plot.

Refer slide time: (0:41)




Now, let me start our discussion, in case if you recall, what do you do in case of construction of histogram? First you have a frequency, distribution in which you have class intervals and then you try to plot the bars, where the lower and upper limit of the bars, are indicating the lower and upper limit of the class interval. And in case if your data becomes quite large, then what will happen? Ideally you assume that, whatever are the observations which are contained inside the bar, they are assumed to be concentrated at the midpoint of the class interval. So, in case if your, number of data point become large, then obviously, you will have to create more number of bars. And if you are in case, if I say in very simple word, in case if you want to make your, histogram more precise, then you need to make more bars. So, what will really happen, that we try to first understand and then based on that, how to represent the data, that we try to see, actually this can be done, through the concept of kernel density plot.

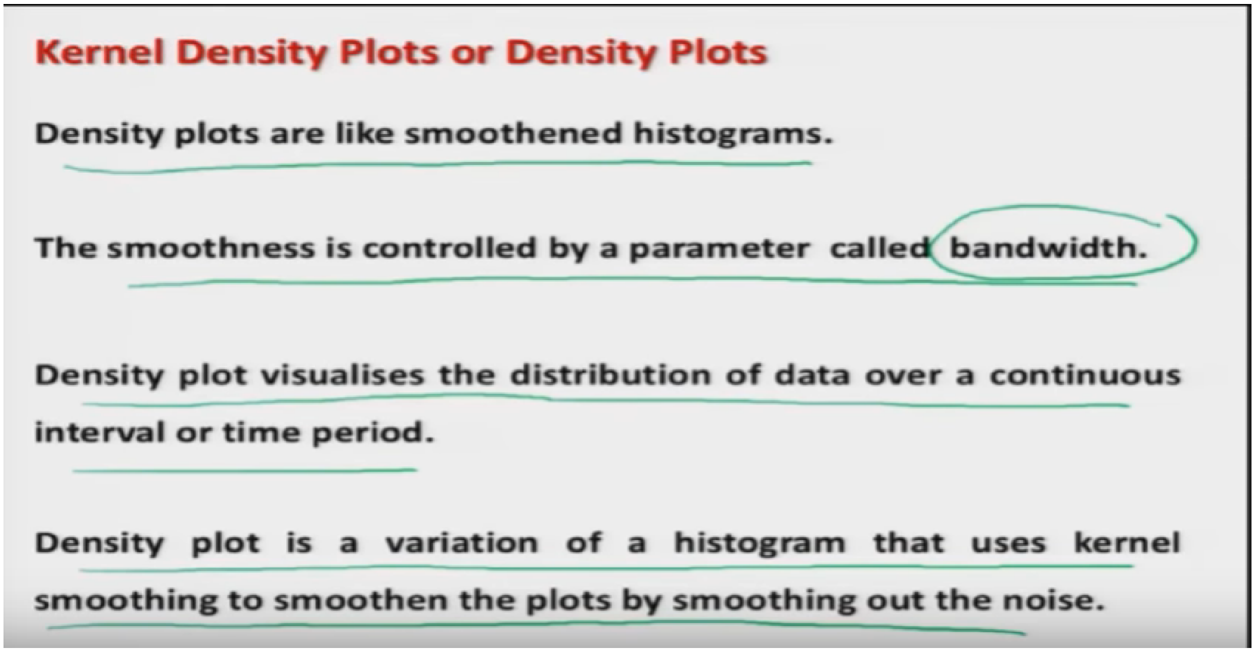
Refer slide time: (2:05)




So, first you try to see here, what happens? In case of histogram, the continuous data is artificially categorized, in different class interval. The choice and width of class interval is, very crucial in the construction of histogram. for example, you may recall, that in case if you have our data, the histogram may look like this, Or other alternative is this, that in case if you try to make the, width of the, bass could be smaller, then it may look like this and so on. So, if you try to see, that in case if you want to make your histogram, more precise or if the number of the data points, which are becoming very very large, you need to create here, more number of bars. And what ideally, you assume here, is that whatever is the data concentrated inside the bar. That is concentrated at the midpoint of the interval. Right?? So, what we can do? That we can, join these points, like this, like this.

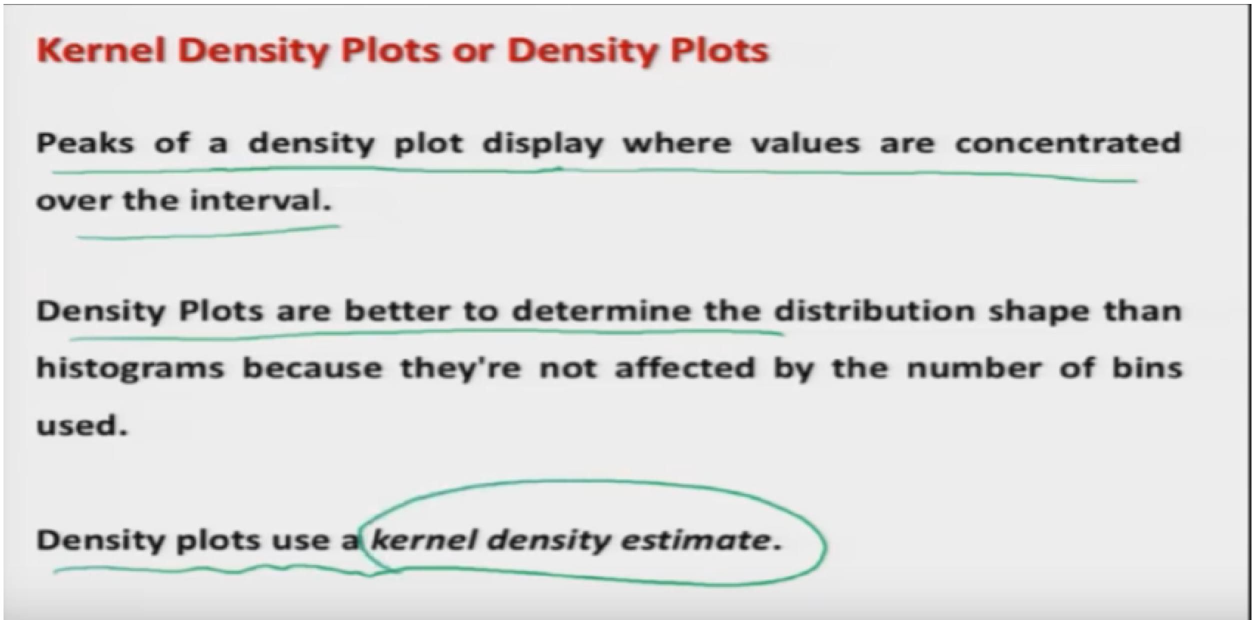
And similarly, in case if you want to do it here, this will look like this, they are the straight lines and so on. and then yah means, these points can be extended, so that they finish, on the x-axis, but now, in case if this number of bars are increasing, then another option is this, that I can join the midpoints of the bars, by a smooth curve, like this, you see I will put my pen here and then I will try to make, a small curve like this, or something like this and this, curve has been drawn in such a way, such that, it is passing through with, most of the midpoints of the class interval, this curve is called as, 'Frequency Density Plot', or in simple word this is called as, 'Density Plot'. Now, this is the concept. Now, our next issue is this, when we really want to implement it, for computation. Then how to get it done. and in order to construct a, such type of plots, we try to take the help of, kernel functions and based on that, we try to create or construct the, kernel density plots.

Refer slide time: (5:09)



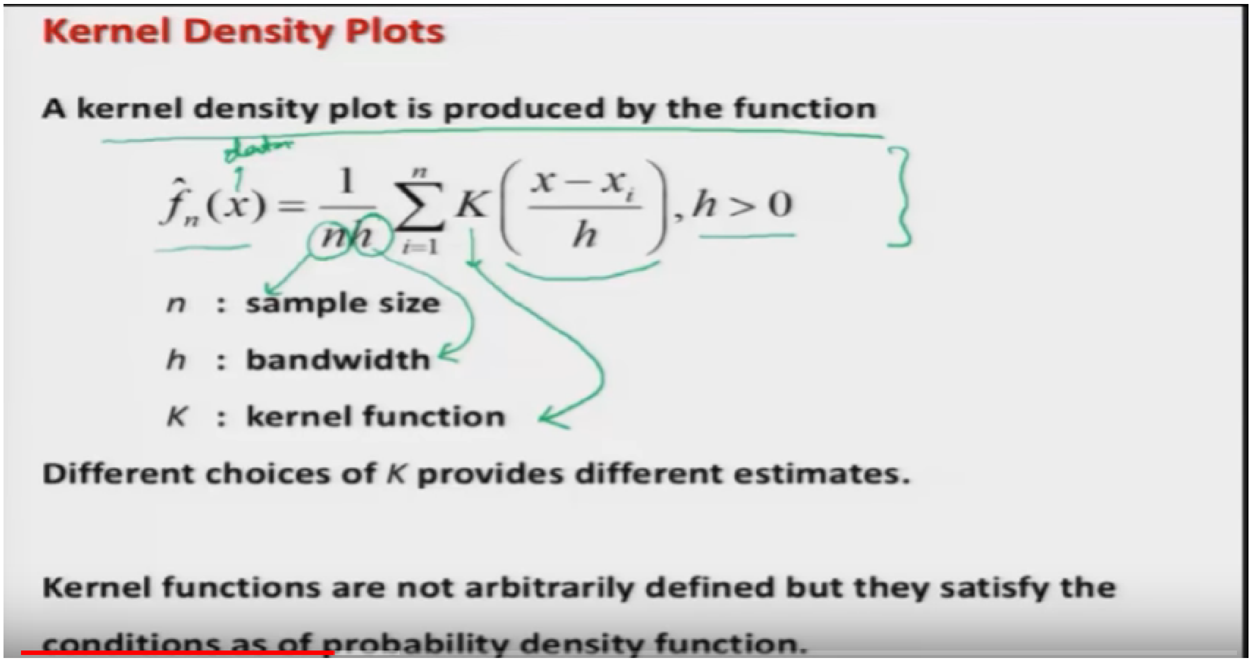

These density plots are like, smoothened histograms. Smooth and histograms, means, we are trying to create the histograms and then, we are trying to, join the midpoints of the bar, by a smooth curve. This is really, what we mean. and now, this is smoothness, this depends on the number of bars, in case if you have large number of bars, the curve is going to be more smooth and as, the number of bar becomes, larger and larger and they tend towards infinity, this number tends to infinity, then this curve will become a, perfectly smooth curve, this is the basic idea. So, in order to plot it, this smoothness is controlled by a parameter, which is called as, 'Bandwidth'. And the density plots, helps us in visualizing the distribution of the data, over a continuous interval or a time period. And in case, if I want to explain this, candle density plot, in very simple words, then I would say, that this is only a variation of a histogram and the histogram is being constructed, by using the concept of kernels. Right? So, density plus is simply a variation of histogram, that uses, candle is smoothing, to smoothen the plots, by smoothing out the noise, this noise is coming, because of the, class intervals. Right?

Refer slide time: (6:40)




And whenever, we are trying to create such a kernel density plot, the peaks of the density plot, they display where values are concentrated over the interval. That is the value of the frequency. And the advantage of kernel density plots, over the histogram is that, the shape, of the density plot is not affected by the number of bins, number of bins mean, number of bars. Where as in case of histogram, the shape of the histogram is, determined by the width of the bins, number of the bins and so on. So, that is why, density plot play up, better role than the histogram, when you have large theta. And these density plots are, constructed using the concept of kernel density estimate. So, let us try to first understand, what is this kernel density estimates?

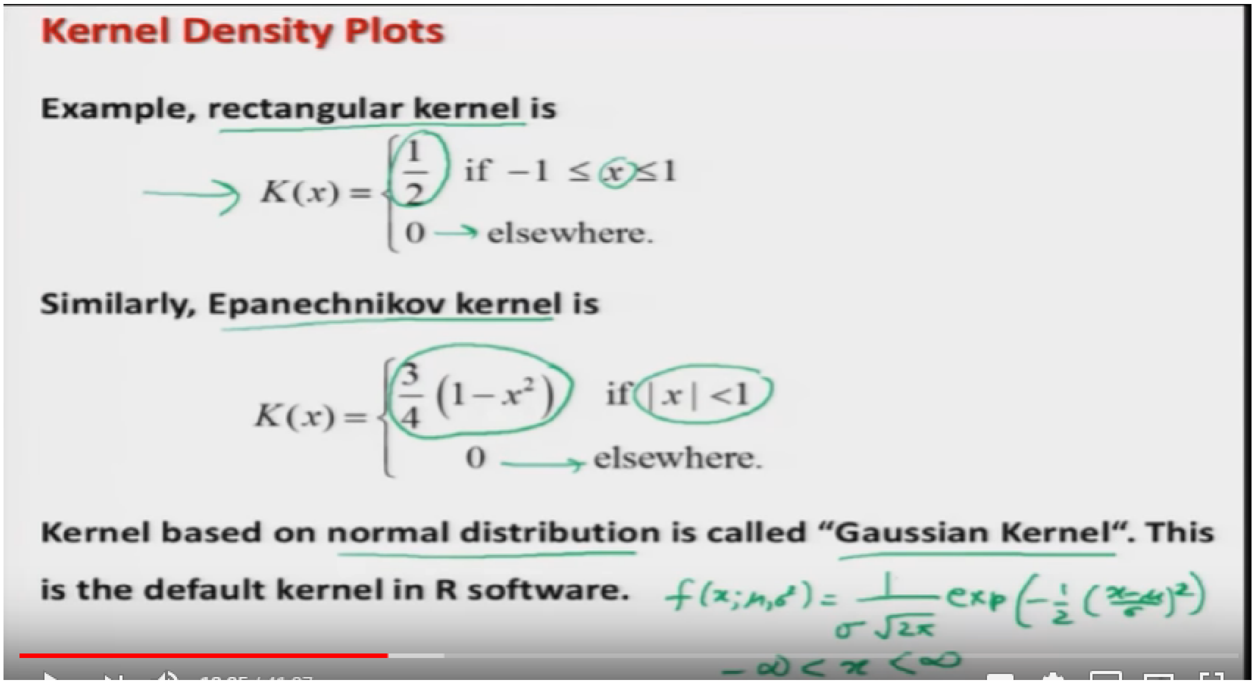
Refer slide time: (7:45)




What are this kernel density functions? So, a kernel density plot is essentially, defined by your, kernel density function. and how this, kernel density function is defined, you can see here, I am writing here, a mathematical function, which is here F hat and X and X here, is a data or the variable on which, we are trying to create the data or we try to collect the data, this is given by 1 upon n H, summation over a function. And the form of this function is going to be determined by the function K. and inside the argument X minus X I upon H, so that X is a function of X minus X I upon H and we are H is some positive value. Here, this is small and this is denoting the sample size and this is small H, this is trying to indicate the bandwidth, this is the parameter that is going to control the, smoothness of a kernel density function. And this K here, this is called heresy, say kernel functions. This kernel density plot, is not arbitrarily defined, this function also has, some properties, which have to be satisfied, in case if a function has to be treated like as a kernel density function and these properties are, similar to the properties or probability, density function.

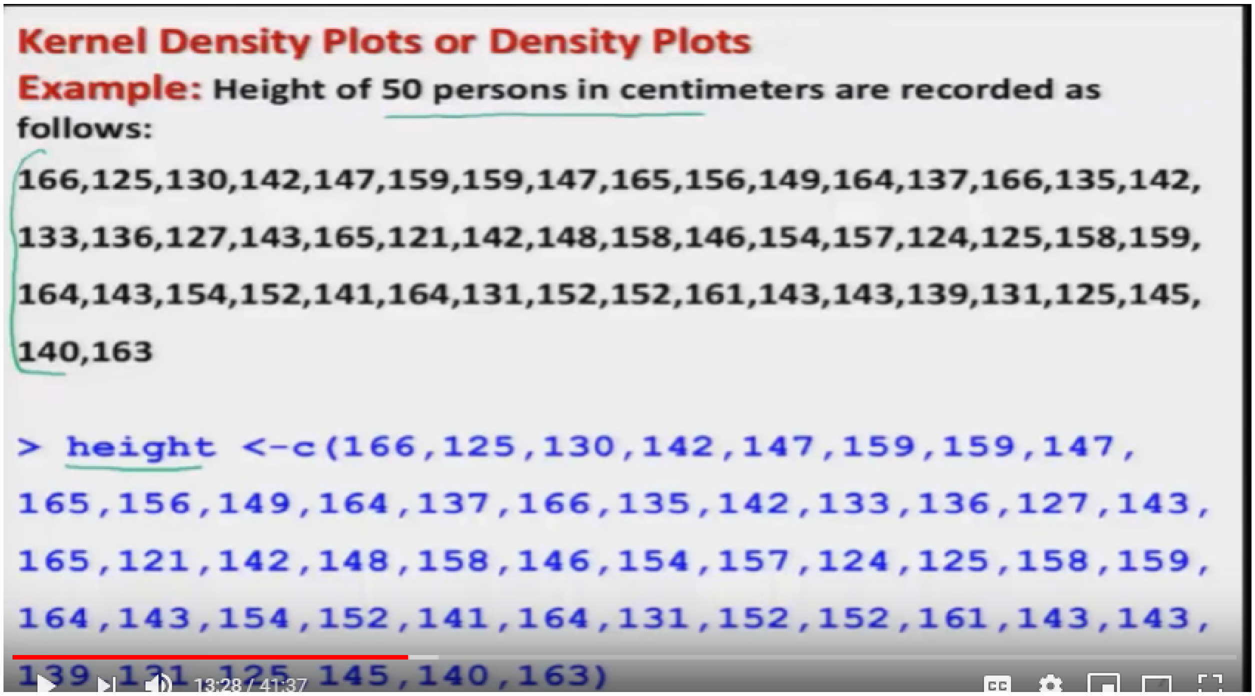
Well those who do not have a background stresses, I can tell them, that in, statistics we have probability density function, for our continuous random variable and these are the function which have certain properties and these function helps, in determining the probabilities, of events. For example, you have heard normal density function, gamma density function, chi-square functions, T distribution, chi-square distribution and F, F distribution and so on. So here, I'm not going into that detail, but I just wanted to give you, some idea. And in this case of a kernel function, in case if I try to choose, different types of K, that is the kernel function. We will get different estimates and constitutively, we will get different types of plots. Because, now I can say briefly, that this kernel density plots, are going to be constructed on the values of the candle function, that we are going to obtain, on the basis of given set of data. So we have data, we try to estimate the values of, kernel density functions and then, we try to plot them. So, obviously in case, if you try to change the form of the kernel function, your estimate may also change. But, definitely those kernel choices are taken in such a way, such that, there is not much difference among different kernels. And the aim is to provide a function, which is more efficiently presenting the truth, frequency distribution of the data. Okay?  So, just to give you an idea, what I am going to use here? And what choices are available in our,

Refer slide time: (11:06)



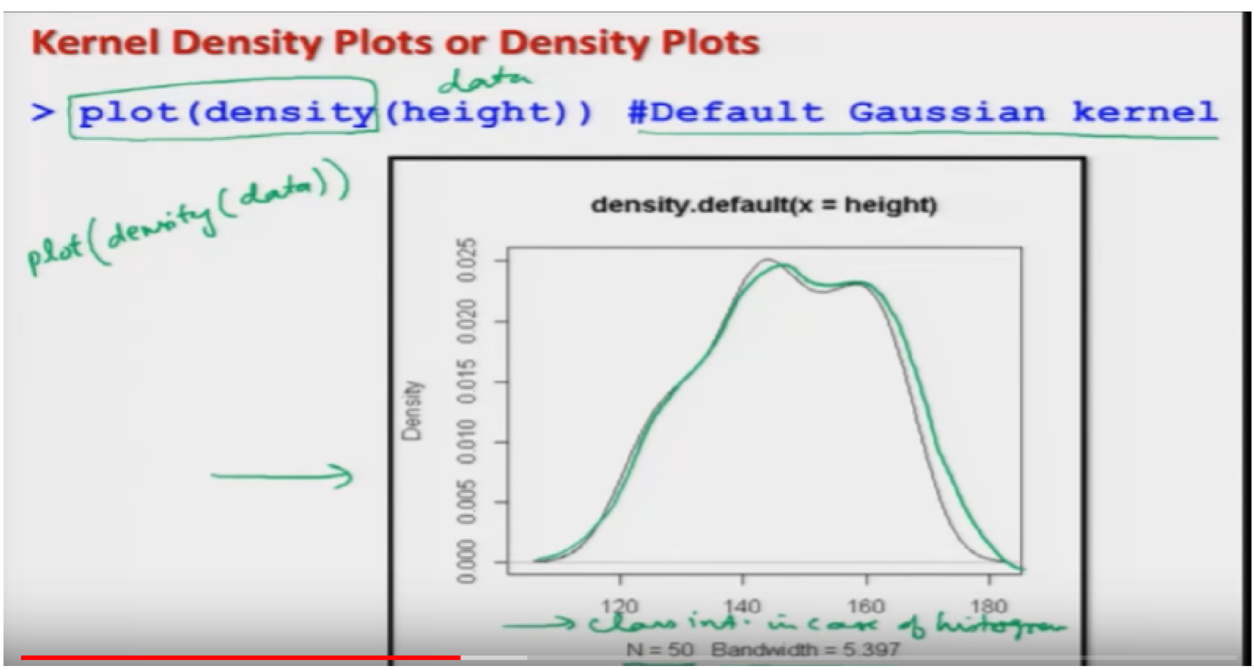

We are going to discuss here, three choices of kernel function. One function here, is a rectangular kernel and second is, a Panasonic oh food kernel and third is, normal distribution kernel or this is called as, 'Gaussian kernel'. Right? What is this rectangular kernel? this rectangular kernel is defined, as a function, which takes value 1 upon 2, in case if the value of x is lying between minus 1 and plus 1, else this is taking the value 0. Similarly if Panasonic of kernel, this takes the value, 3 upon 4 into 1 minus X square, if absolute value of X is smaller than 1 and 0 in all other places. similarly the Gaussian kernel, this is based on the normal distribution, we know, that the probability density function of normal distribution, with parameters mu and Sigma I Square, this is given by 1 upon Sigma root 2 pi, exponential of minus 1 upon 2 X minus mu upon Sigma whole square. And here, this X lies between say, minus infinity and plus infinity. Right? And actually, when we are trying to construct such, a kernel density plots in R software, then in case if you don't give any choice, then this Gaussian kernel is the default choice. But, I will try to show you that, how the density plot looks, when we try to change the, type of kernel or the choice of kernel. So, could we take here, the same example, that I had taken in the case of histogram and I will try to construct the, candle consider plot, in the R software and I will try to compare them with histogram also. So now, let us try to consider an example, in which

Refer slide time: (13:12)



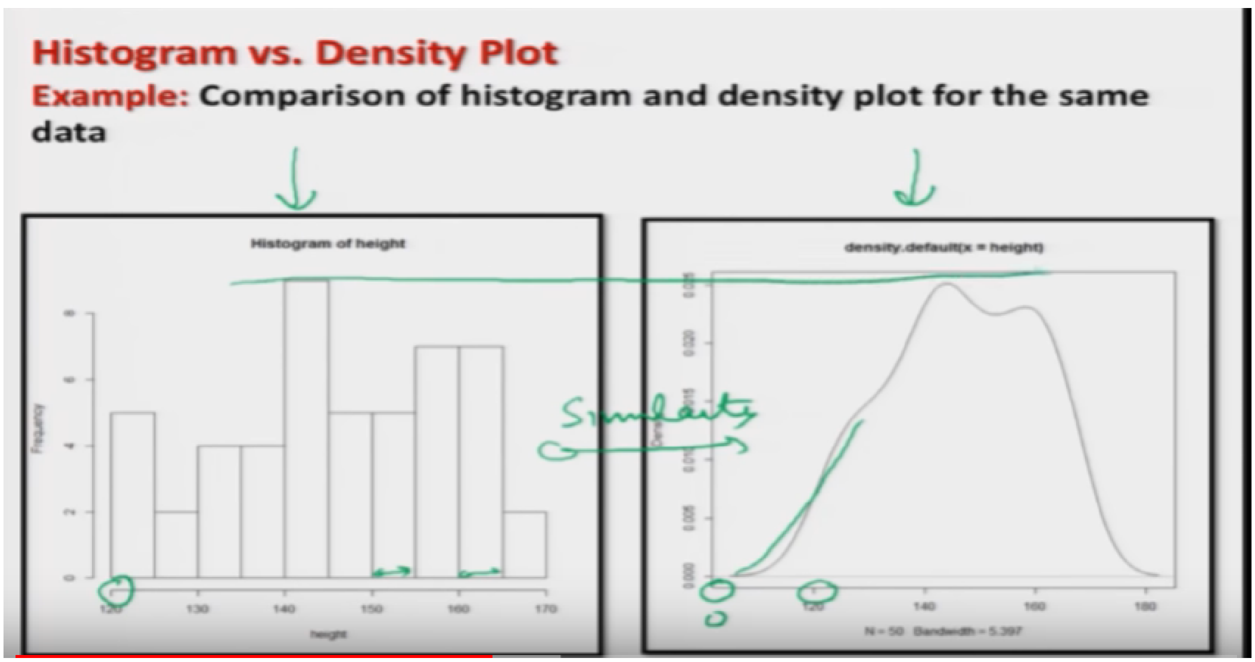

the heights of 50 percents are recorded in, 20 meters. And we would like to create a, density plot for this data. So, this data has been stored in a variable, here height, here like this.

Refer slide time: (13:30)



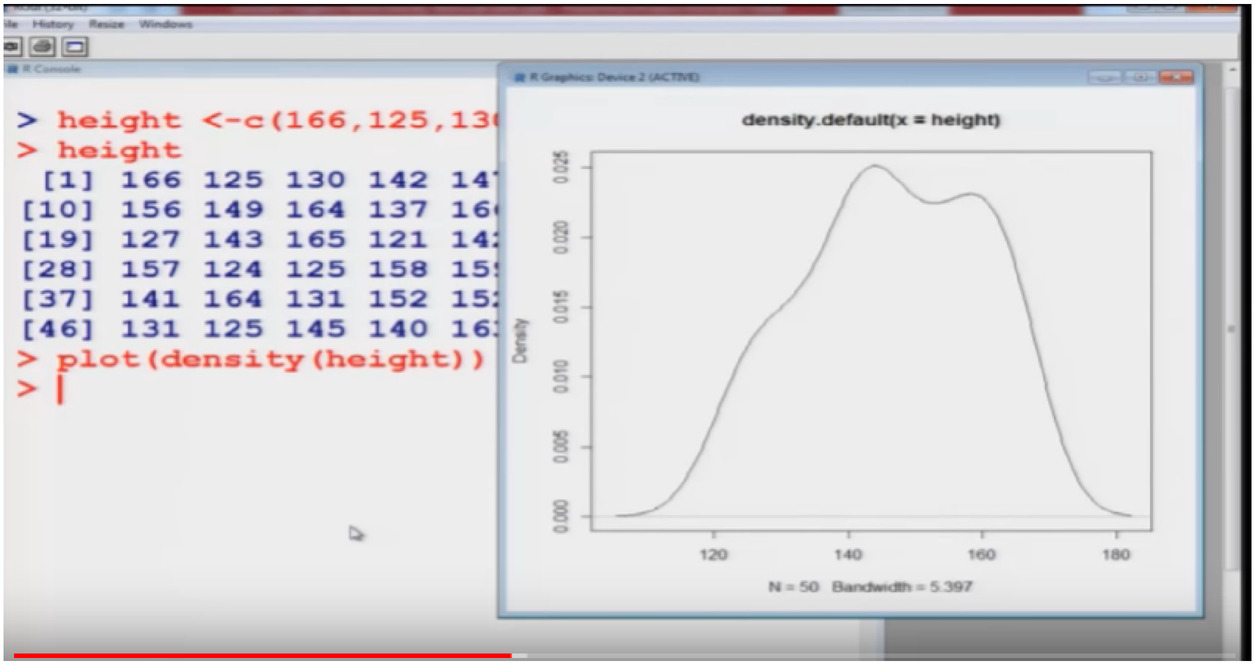
And after this, in case if you want to plot this, kernel density plot. Then the command in R software is, plot density. And so, I would try to write here plot, inside the argument density and then, inside the arguments, I have to give the Data. So, this is the command here, for plotting the density and remember, this is plot, inside the argument you have to write, density and then, once again inside the argument, you have to give the, data. So here, the data is given by, variable here height and in case if you, execute, it you will get here, this type of graph. So, you can see here, here are the number of observations, here 15. The bandwidth here, is controlled by the Gaussian kernel, which is the, default kernel, when we are not specifying anything and you can see here, this curve, looks like this, this is a smooth curve, hunting this. So, this is called a, 'Kernel Density Plots'. And here, these are the values, of something like a class intervals, in case of, case of, histogram. And this type of a smooth curve helps us, in getting an idea, about the distribution of the, numerical values and these types of curve, are actually more useful, when you have large number of data that will give you a much better information, than the histogram. Okay?

Refer slide time: (15:17)



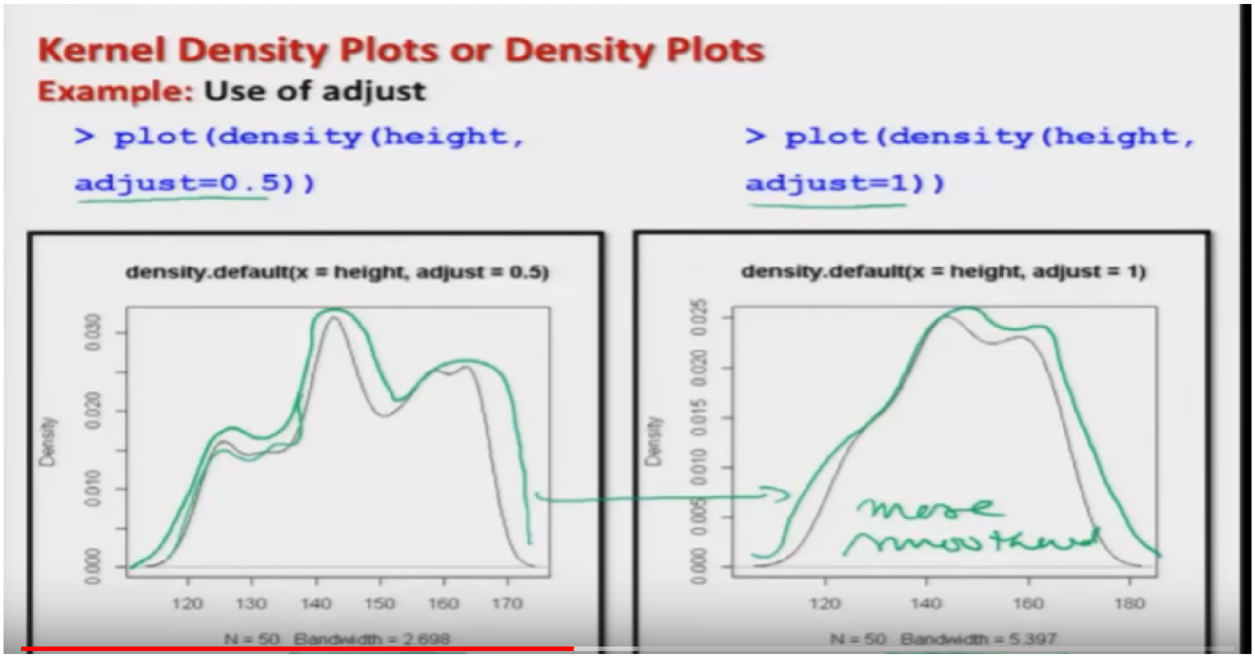

Now, in case if I try to plot the histogram and here, this density plot side-by-side. Then it will help you, in comparing the two. So you can see here, I have plotted the histogram, of the same data and here, is the Gaussian kernel plot. Now, you can see here, that how the things are happening over here. So, one thing what you have to keep in mind here, that the histogram is, is starting from hundred twenty, whereas, kernel density plaus, is starting with zero. and then, what is really happening, that the width of these bars, is now made a smaller and smaller, and that is controlled by, the choice of the kernel and then, all this frequency ,they have been joined together, to give this type of curve. So, you can see the, similarity between that two graphs. For example, you can see here, the maximum frequency is here, at the same place. But, this is giving us, the more information. But, before going further, let me try to, plot this data, on the R console and I try to show you here, so that you are more confident.

Refer slide time: (16:33)



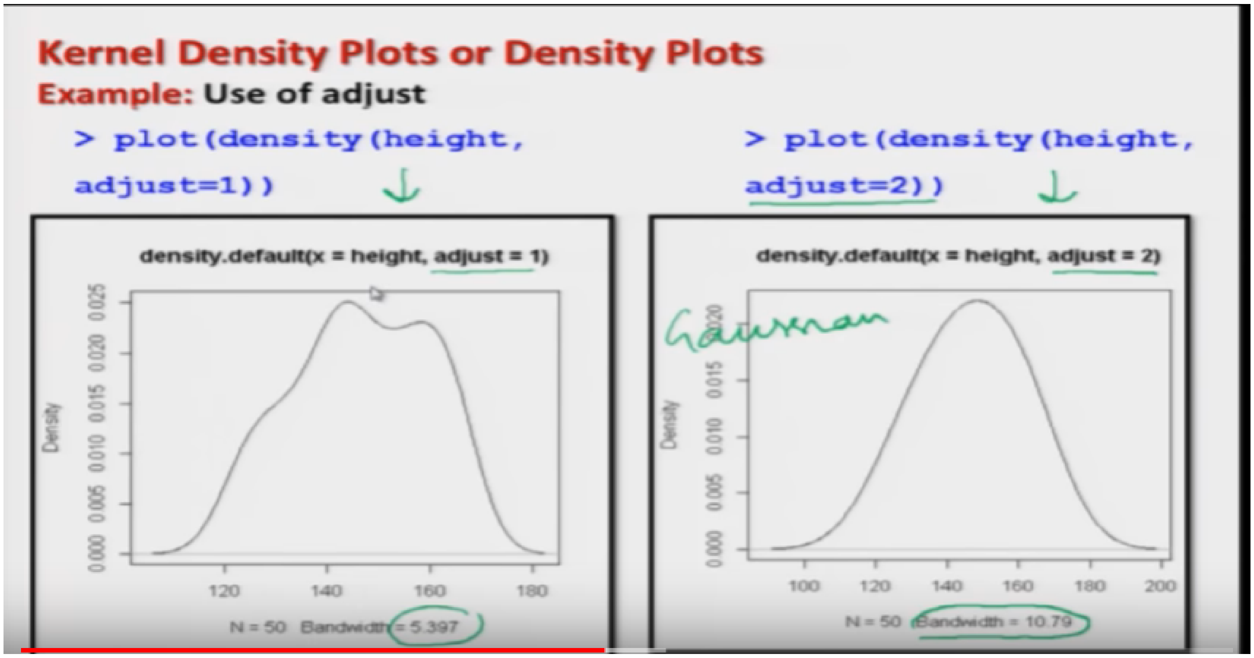

So, now let me copy this, data on the R console. and you can see here, this is the data on the height and now, I use this plot, density command and if I, execute it on the R console, you can see here, you get, this type of kernel density plot. Based on the, choice of Gaussian kernel. Right? And now, I will show you, one more aspect,

Refer slide time: (17:09)



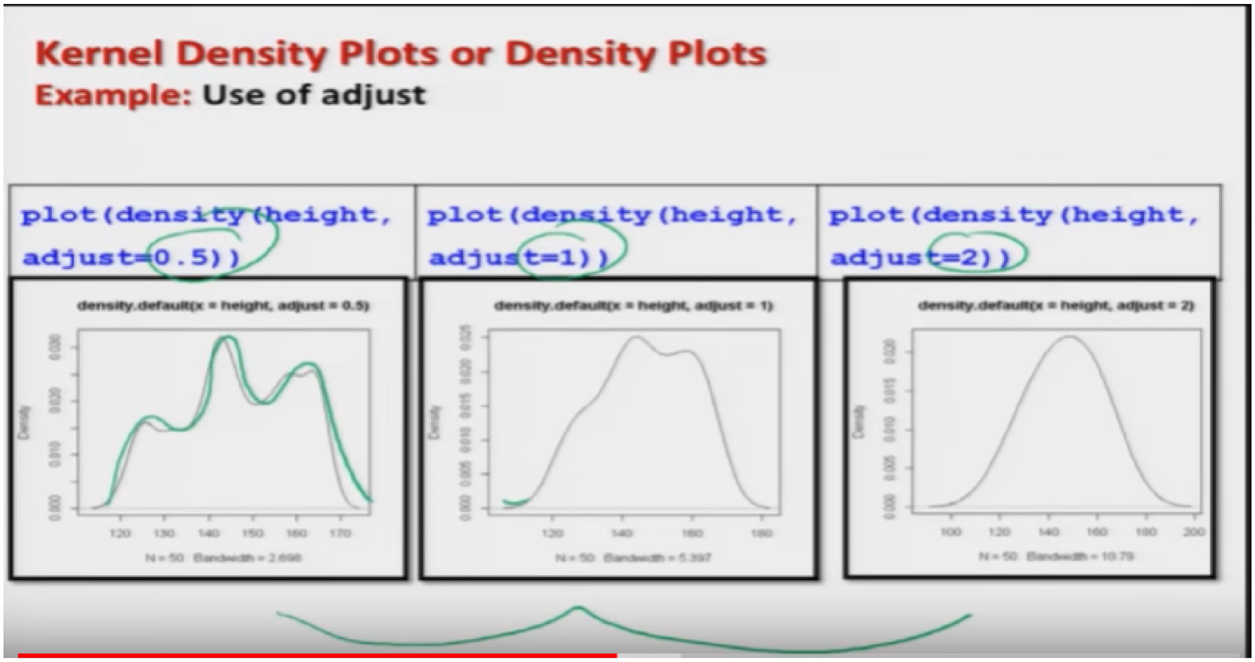

You can adjust, the plots this kernel density plots, using a command, adjust. this adjust is going to be a parameter, that takes some numerical value and if you, try to see here, the number of bins or the structure of the curve, that is going to be, controlled by this, adjust parameter, for the sake of understanding and explanation, I am trying to take here, two possible values of adjust at 0.5 and here, adjust equal to 1 and then I am trying to plot the same date, using the, same command., so, you can see here, the structure of this data set and structure of this data set. So, you can see here, the data sets are the same in both situation and the same Gaussian kernel density function has been used to construct, these densities. But, their structure is, different. this is giving you more variation and we're here, the second one this is, more is smooth and, and you can see here, the in, the first case, the value of bandwidth chosen is, two point six nine eight and in the second case ,that chosen bandwidth is, five point three nine seven. Which is corresponding to a just equal to one? Right? Now, I would say, that a question comes, what should be the proper value of edges? so for that, I will say, try to create a histogram, try to look into your experimental data and try to choose certain values of edges and try to plot your curve and whatever, curve is representing, the situation in a more, authentic way, more honestly, try to choose that value of a just.

Refer slide time: (19:13)



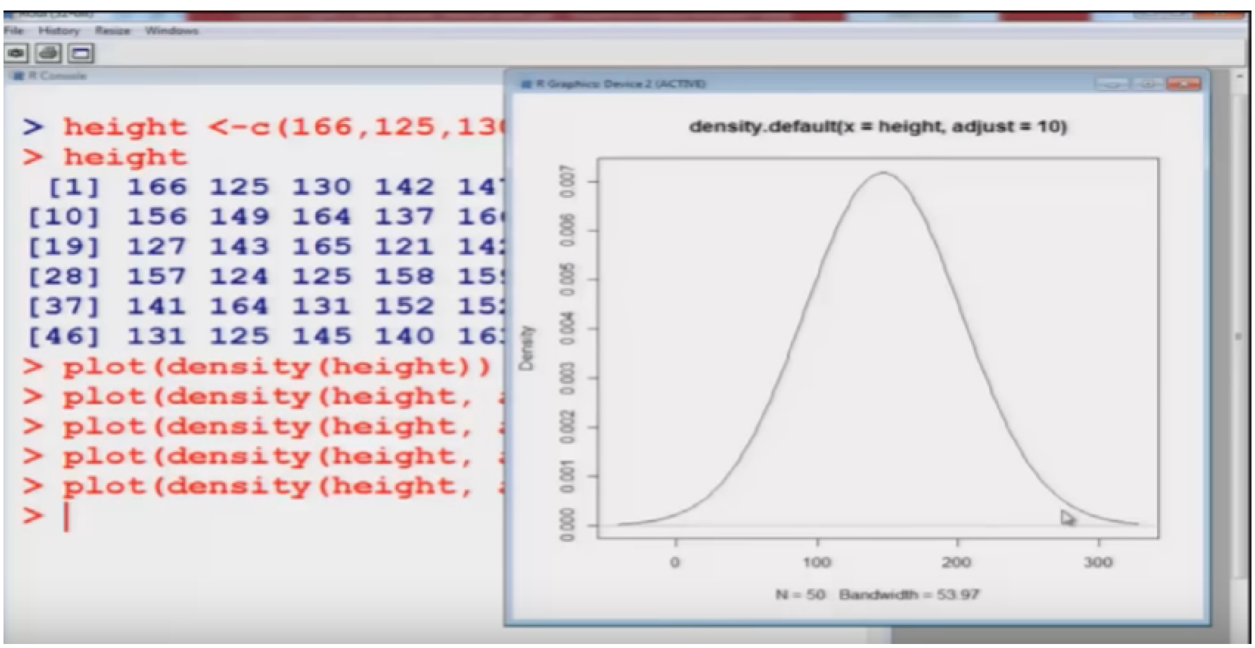

and similarly, in case if you want to increase the value of edges, for example, in the next slide I am trying to, create the same density plot, using the adjust equal to one and two. And you will see that, as I try to increase the value of edges parameter, the smoothness increases more. For example, you can see here, this first graph, this is the same data with a just equal to 1 and the second graph, this is with a just equal to 2. You can see here, with a just equal to 2 ,this is more like a Gaussian curve or normal curve. Because, here the bandwidth has, increased from five point three nine seven to, ten point seven nine, nearly double. Right? So, this is how you play with your adjust parameter.

Refer slide time: (20:06)



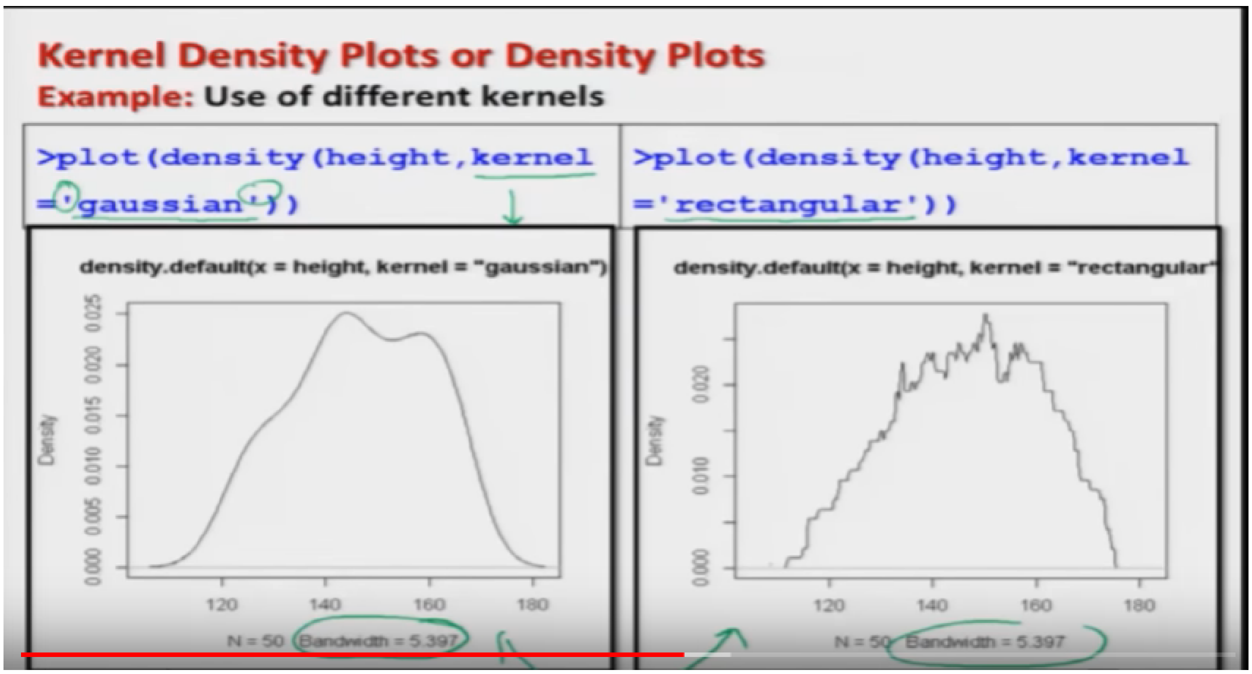

And now, in this slide, I am simply trying to give you, all the three slides or say all the three graphics, with a just equal to 0.51 and 1, two together. So, that you can make a better inter-comparison, you can see, that as the value of edges is increasing, this degree of smoothness, here is increasing. So, this is how you can play with, this data. So, first let me show you, this execution on the R console.

Refer slide time: (20:43)



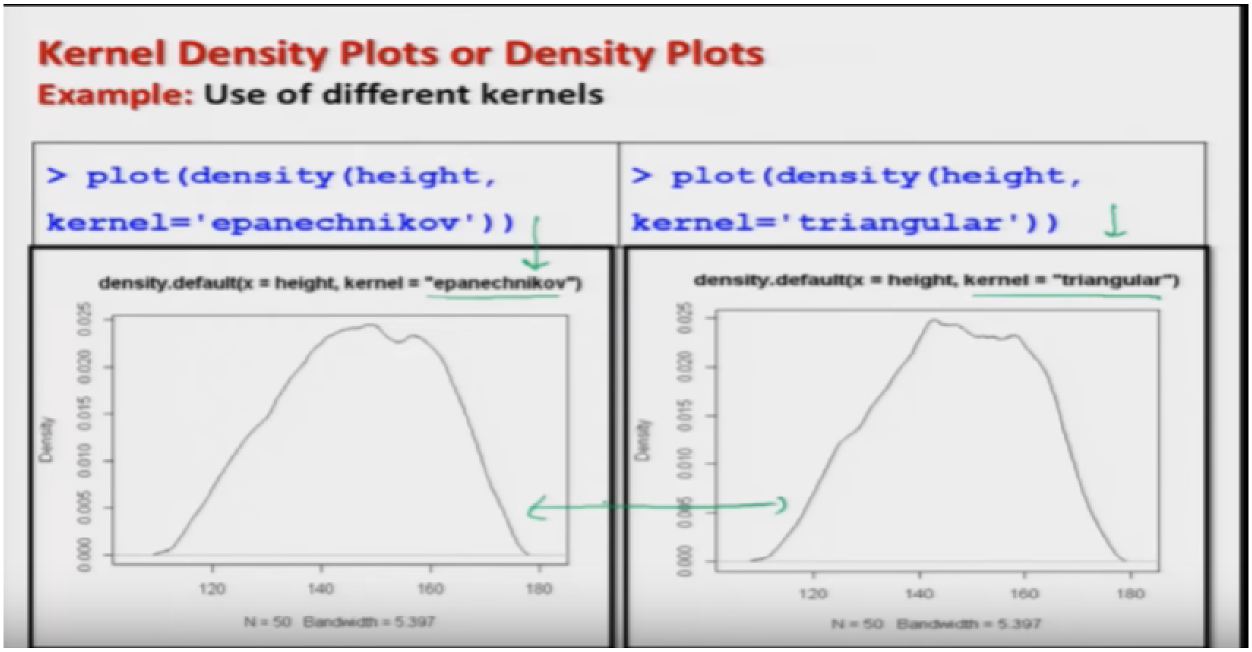
So, you can see here. Now, I'm trying to use here, first the density plot, with a just equal to 0.5 and then click on I try to take the value here 1, 1 and then, I try to take the value here 2, you can see here, this becoming more normal and suppose just for the sake of illustration, I try to take the value to be here 10, you can see here, this is now, more towards the normal curve and so on. Right? Ok. Now, let me try to address, this another feature of kernel concerti plot.

Refer slide time: (21:30)



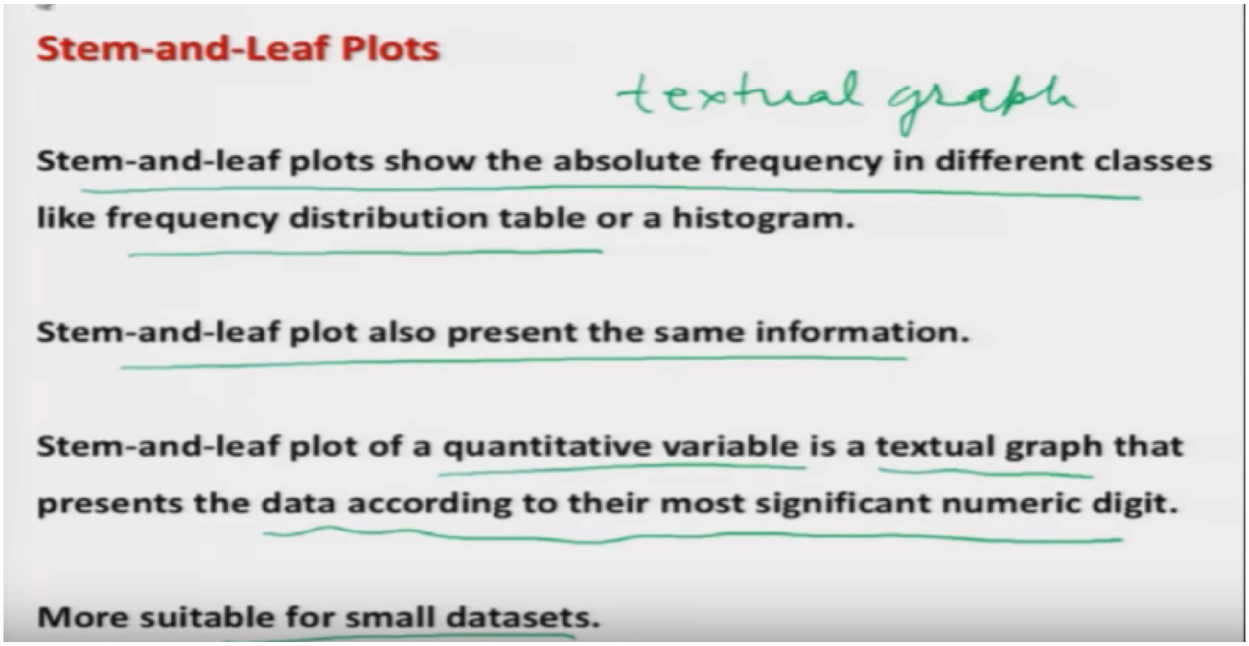

That in case if I try to use, a different kernel function. So, you can see here, on the upper left hand side, I'm trying to create a deacon city plot, using the Gaussian kernel, which is the default kernel and for that, now I'm giving the choice, as kernel KERNAL equal to, inside the single quotes, Gaussian GAUSSION. and similarly, I am trying to take here, another kernel here, that is a rectangular kernel and if you try to compare here, both these graphs, you can see here, the structure is almost the same, means, even here, if you try to plot the graph, this will look, something like this, on the similar lines. But, the structures are different. So, now you are the one, who has to decide that, which of the kennel is going to, give you the more representation of the data. And you can notice that, the bandwidth in both the cases that is here the same, five point three nine seven. Right? And if you want to increase the, bandwidth and the choice of Cardinal, both together, then you try to use the parameter edges, as well as, the kernel choice.

Refer slide time: (22:46)



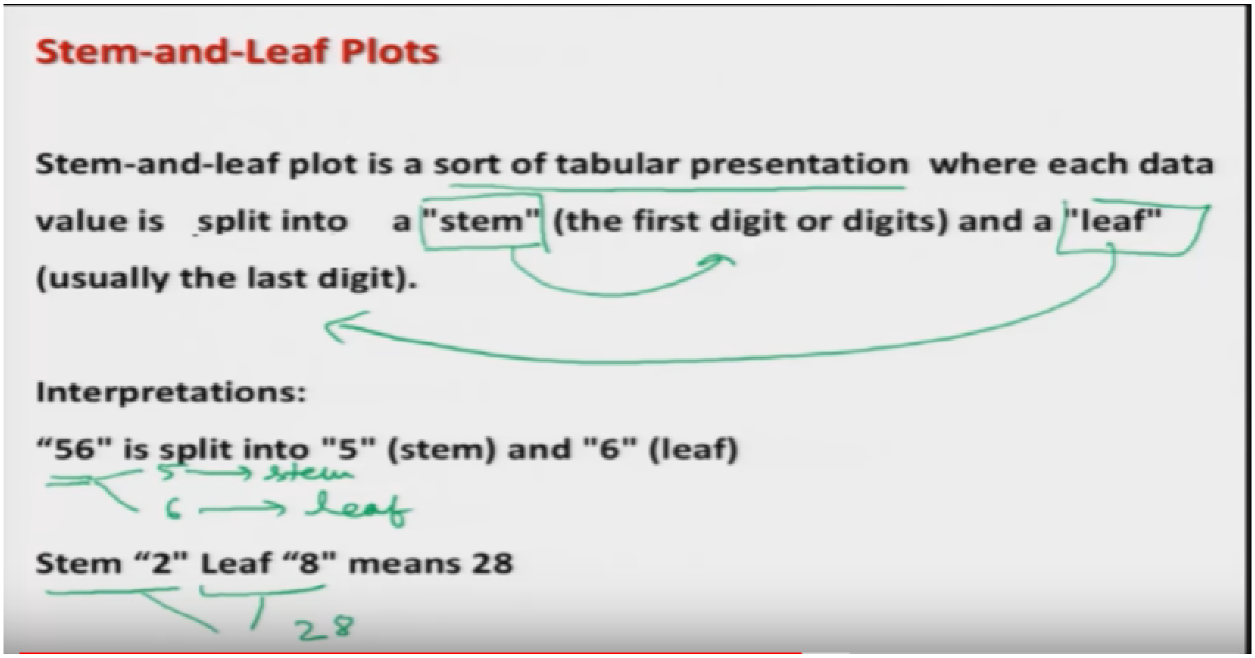

and similarly in case if you try to choose other, kernels like as here, in this first graphic, I'm trying to use the, kernel if EPANECHNIKOV and in this one, I'm trying to use a triangular kernel and so on. So, you can see here, the structure is more liked similar, but then, they are not exactly the same. So once again, I will try to show you, all this graphics on the R console, to make you more confident and then, I will move to another graph. so you can see here, here I'm trying to make the, Gaussian kernel, then I'm trying to make the, density plot based on, rectangular kernel, which was reproduced in the slides and similarly if I try to take, a pence Nikko kernel, then it comes out to be, like this and if I try to take a triangular kernel, then again this curve changes, you can see here.  Right? So, this kernel density plot, will help you in getting a smooth density plots and they are more or less, like similarly to the, histograms also.

Refer slide time: (24:06)



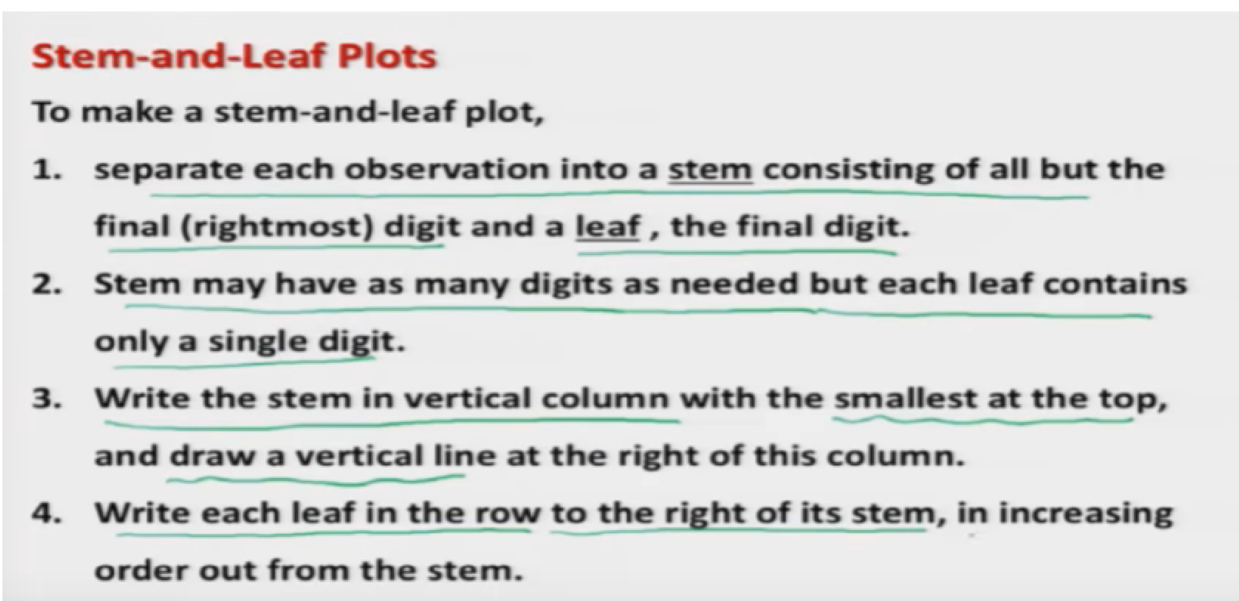

Now, we come to next topic, that is stem and leaf plots. These are once again, another way of representing the data. And in this case, the absolute frequencies in different classes, they are represented in a different way, for example, we  represented the frequency or the absolute frequency, using the bar diagrams, in case of discrete data and histogram, in the case of continuous data. Right? So, in this, case we have ,a data set and the graphic is, going to be a sort of combination of the, graph and numbers. Graph or text, so that is why, this is also called as a, 'Textual Graphed'. Means, it is trying to use the, text as well as graph. So, this estimate leaf plots, show the absolute frequency, in different classes in the frequency table or in a histogram. And they also, present that same information, the only difference is that, this is based on a quantitative variable and these are the textual graph. And the presentation in this graphic is based on the, data according to their most significant numeric digits. And this type of graphic, is actually more suitable, for a small data set. So, I would advise you, in case of your large data sets, try to go for histogram or say, density plots.

Refer slide time: (25:44)



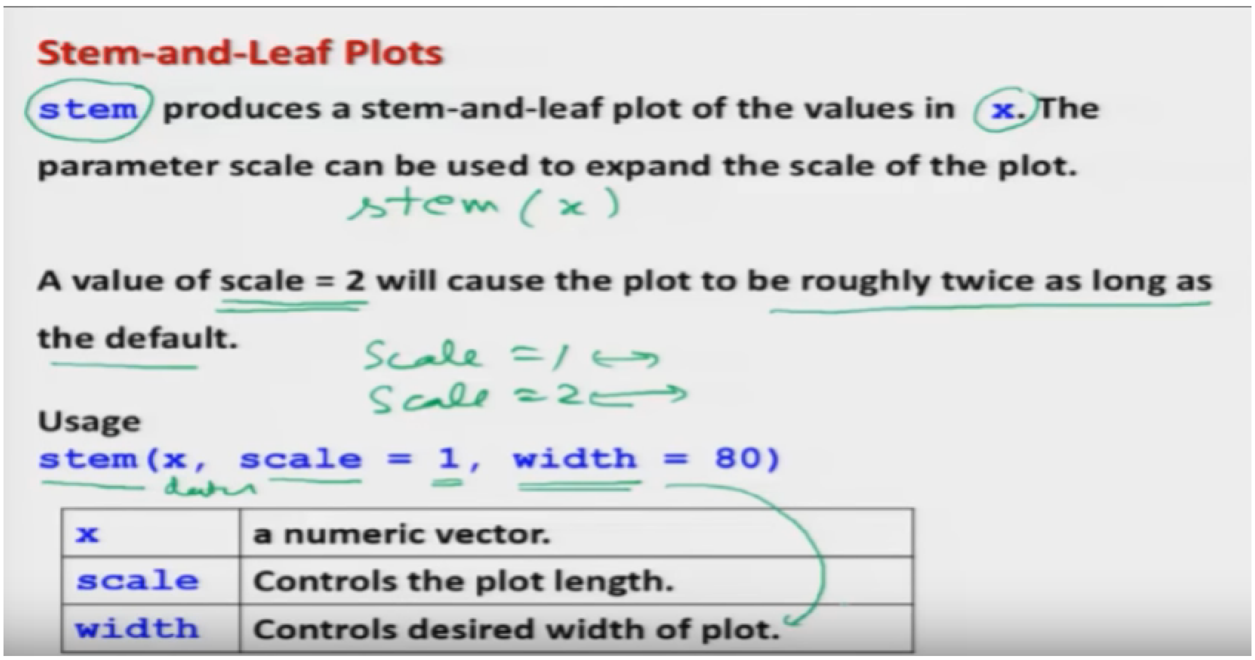

And this is stem and leaf plot ,this, this is a sort of actually, tabular presentation. Where each of the data value, is going to be splitter, into two parts. One is called, 'Stem', and another is called, 'leaf'. Usually in a stem leaf plot, the stem is representing the first digit or the first digits of the data. Or leaf is going to usually, represent the last credit of the data. For example, in case if I have a data, 56. Then 56 is going to be splitter into, two parts, five and here six. This type is called a stem and this six is called here leaf. Similarly in case if I say that, I have a number, whose stem is to and leaf is eight, this means, that the number is 28. So, this is how we try to, represent the interpretations stem and leaf.

Refer slide time: (26:45)



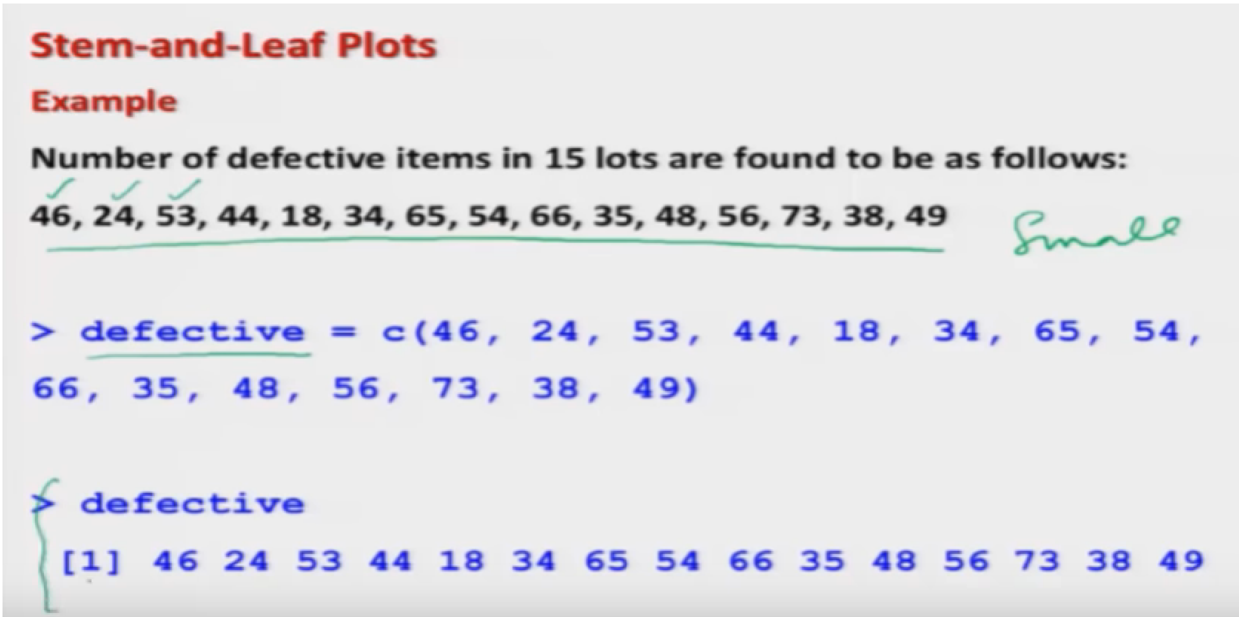
In order to create, the stem-and-leaf plots, what we have to do? First we need to, separate each observation, into a stem consisting of all, but the final, which is the rightmost digit and a leaf which is the final digit. Actually this stem may have, as many as possible digits, as needed but each leaf contains only a, single digit, this is the key point. and then, we try to write down, the values of the stem in, vertical column, in cutaway such that the smallest value is on the top and then we draw a, vertical line and after this, we write the value of the leaf, in each row, to the right of the, the stem. And this is again done, in increasing order.

Refer slide time: (27:39)



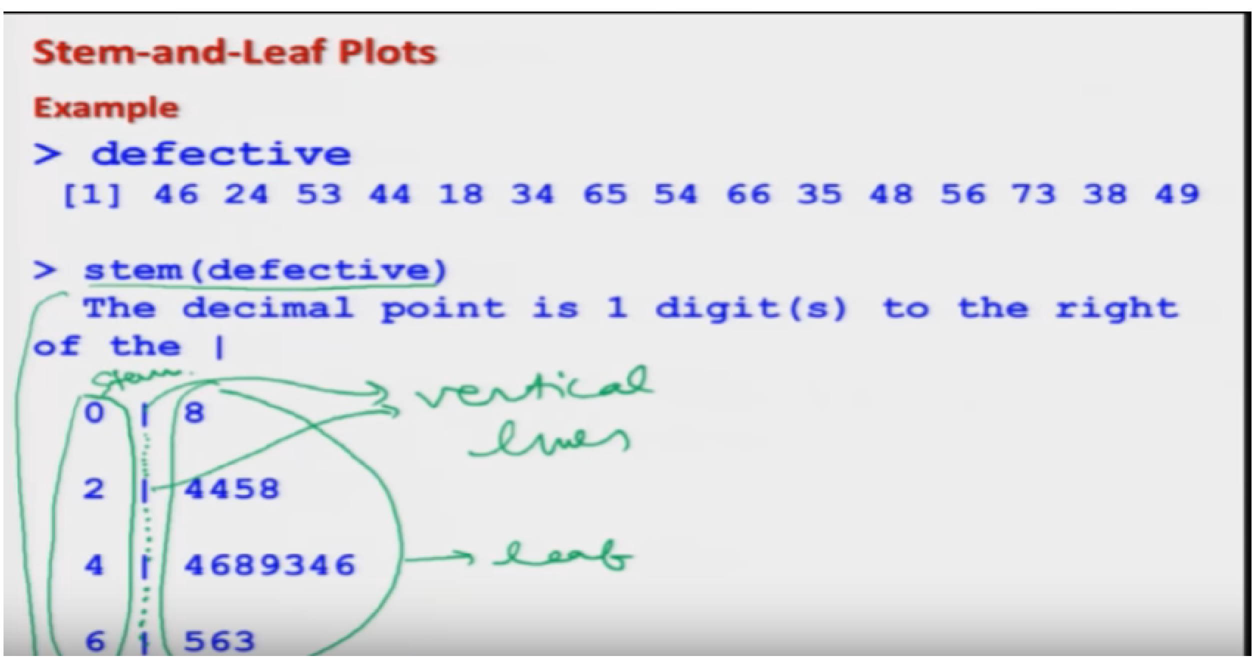

So, first let me, give you here, the R command and can I will try to show you, how the stem and leaf plot, look like which will make you understand better. So, in R we have a command STEM, this command, produces the stem and leaf plot, of a variable here, of the some values here in data X .so, we need to write here, say stem inside the argument, see here, X. and then, the size of the stem, that means, how it has to go, this is controlled by a parameter, what we call it as say, ‘Scale’. So, in case if I try to choose here, two values of scale. Scale equal to 1 and is scale equal to here 2. then the interpretation will be, that when I am trying to choose the value 2 here, then this will give me, a stem and leaf plot, which is roughly that ,twice as long, as of the default value, which is scale equal to 1. So, in this case, when we want to constrict the stem and leaf plot, the R command is, stem and inside the argument, I have to give the data and then, I have to give the scale value, this can be number and this control the length of the plot and similarly, there is another parameter here width, which controls the width of the plot. So, scale controls the length of the plot and which controls the width of the plot. And the default value is here, usually 80 is taken.

Refer slide time: (29:18)



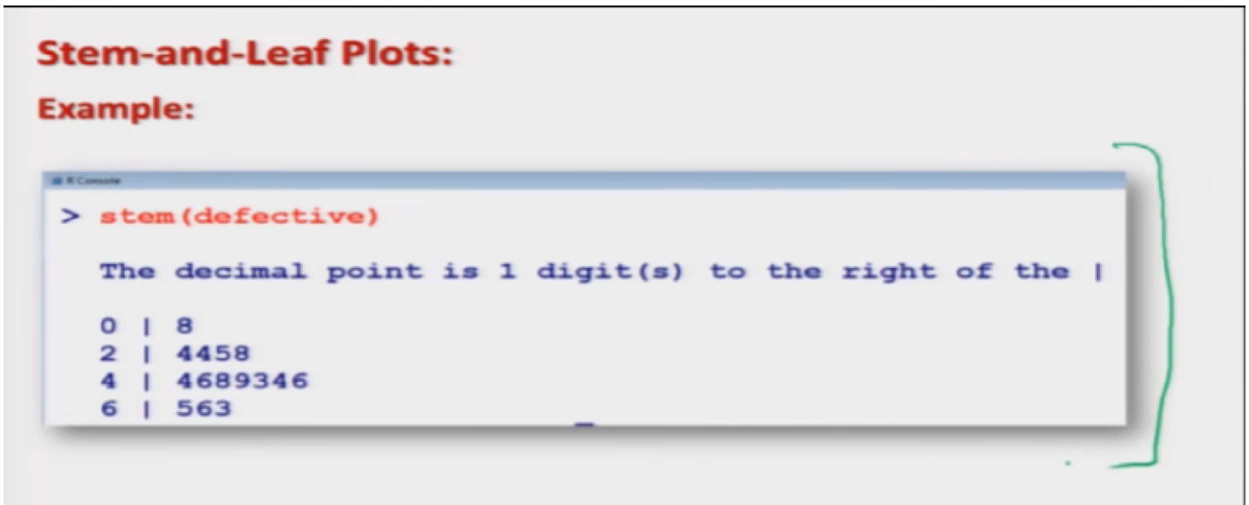

So, let me try to take an example, over here and then we try to understand it. Suppose there are, 15 a loss of an item and the, number of defective items in those Lots are found, to be as follows, like this. So there are, 46 defective number of pieces, 24 number of defective pieces ,53 numbered or defective pieces and so on. And yeah, that's a small cricketer said, only consisting of 15 values, so and we would like to construct the stem-and-leaf plot, for this data set, so I try to put all this data inside a, variable 2 defective, so this will look like this and then, I will try to execute, the arc command over here.

Refer slide time: (30:03)



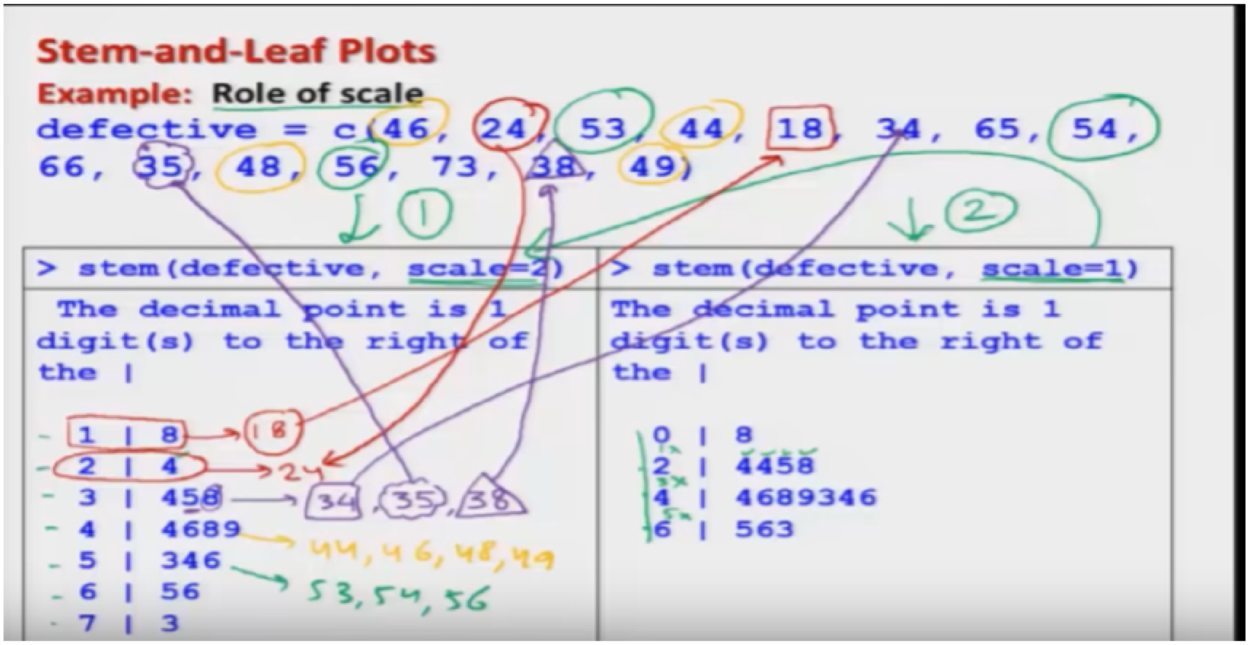
So, as soon as, I say here, stem defective. Then this the outcome will look like this, which I will show you, on the our screen also, on the R console, also. So, you can see here, that this part here is, the stem. And this part here is, the leaf. And these are the, vertical lines, which I was discussing. So, you can make this, vertical lines to be something like this, so that there is a partition over here, so that stem and leaf, they are separated. So, first question comes that, what it is trying to interpret? The interpretation, of this graphic is easy to understand, in case if I try to add here, the scale parameter. So, first I would try to show you here,

Refer slide time: (31:04)




The outcome, the screenshot of the outcome here and first, I will try to give you, the explanation and then I will show you, how to plot it R console.

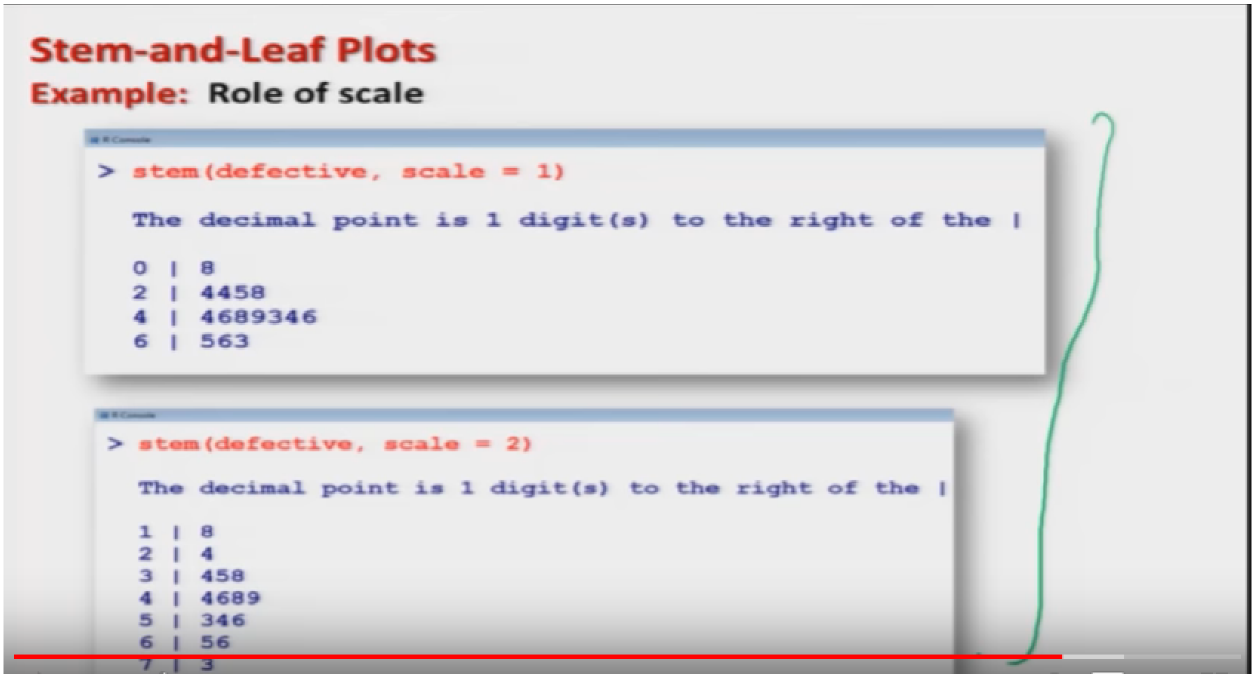
Refer slide time: (31:13)




So, now if you try to see here, first I'm trying to discuss here, the role of, parameter scale. So, you can see here, I have created here, to stem plots. One here is like this, number one and here number two. A number one, I am using the scale 2 and in number two, I'm using the scale value here one, that that I used earlier. You can see here, that there is a difference, that here, there is only here one, two, three four, four values here, whereas here, one, two, three, four, five, six, seven. Then in case if you come on the leaf part, corresponding to two, you can see here, there are four values. 4458, whereas in the case of scale equal to two, there is only here one value and there is no here three, three is not present here, five is not present here and similarly one is not present over here. So, when I am trying to increase the value of, the scale from one to here 2, then this is giving me a more clear, stem-and-leaf plot. So, now let us try to see, in the figure number one, what is the interpretation? Now, first you try to look in the data, I will change the color of my pen, so that you can clearly see it, in the data set, try to see there is here a value one eight. Now, and now try to look at, this thing.

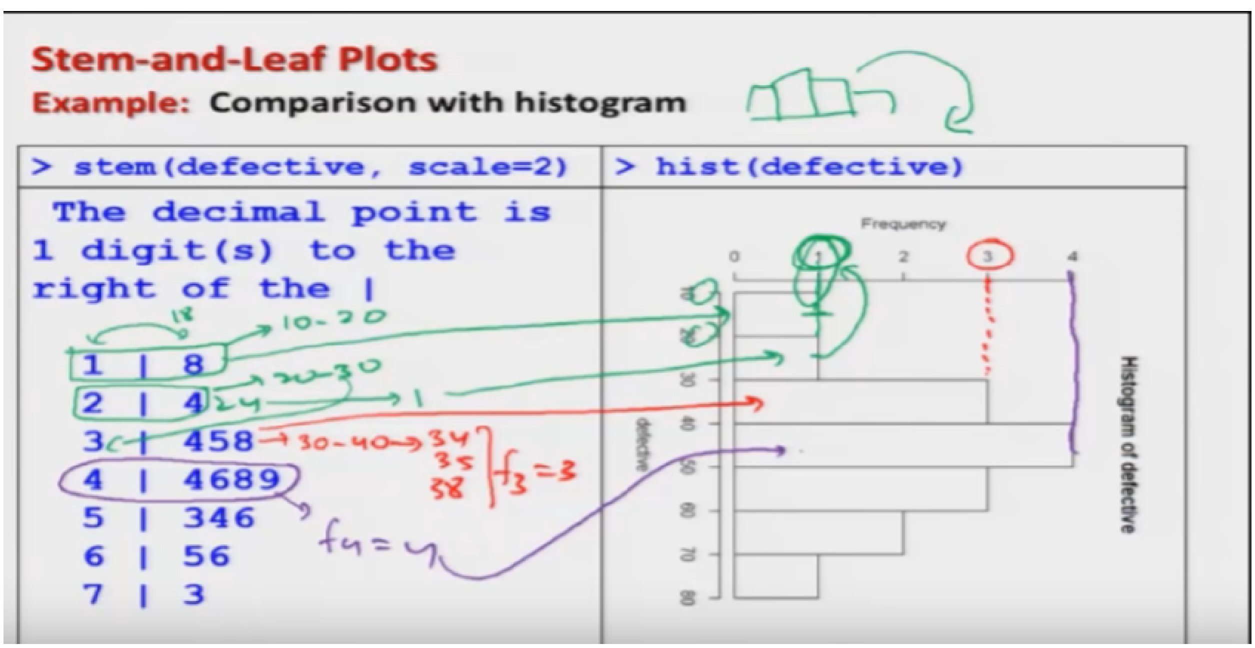
So, this is actually ,indicating one edge stem and eight as leaf and this is the same value, which is given here. And similarly if you, try to look at the second value here, 2,4. So this is actually here, 2,4. What is this indicating? This is also Indicating, a value inside our data and if you try to look on the data, this is the value here 24, this value is indicated here. So, these are the cases, where we have, single digits only. Now, I try to take care, that third column. So you can see here, this is indicating here, something like three as it stem and the interpretation is 4 dessert is 34. Now, next value here is five, so the interpretation goes, the stem here is three and the leaf here is five. so the value here is thirty-five, then the third value here is eight, so this becomes your stem three and value here eight, that is leaf is eight, so then number is 38. And you can see, that these numbers are present in your data, so if you try to see here 34. 34 is present in the data here like this. And if you try to see here, 35, 35 here is this data is present here and similarly if you try to see here 38, 38 is present here, in this data. And similarly if you try to Interpret, this here fourth line, this number is going to be 44, it is representing actually four values, 44, 46, 48 and 49. And you can see here ,that these values are also present, here in this data says 44, 46, 48 and 49. And similarly, if you will try to choose other values, you can also see here, that this is here 53, 54 and 56 and these are again given here, in the data set 53, 54 and here 56. so you can see here, that this stem and leaf plots, they are trying to represent, the sort of histogram in terms of the frequency, of the given data set, how?

Refer slide time: (35:29)




Let me try to show you here, well this is here the screenshot, of the two with a scale equal to 1 and is scale equal to 2.

Refer slide time: (35:36)



So, in order to compare it, I am trying to make here the histogram and then I have the histogram will look like this and so on, I have rotated so that it, becomes comparable with the stem-and-leaf plot. So, you can see here, in this data set this is here 1, 8. So this is indicating here, like this, so now, if you try to see this is here 1. So this is indicating the frequency, that the number of values between 10 and 20, is only here 1. So, the same thing is happening here, also we are trying to read here 1 ,8 as 18. so that is indicating, that and the next value here is 24, so that is indicating, that the class interval here is, 10 to 20 and say, 20 to 30 and so on. 30 is coming because, of here this 3. So, you can see here, that the number of values, in the interval, first interval is only here one, this is here 8, so this is indicating here, on the y axis here 1. Now, in the second case also, there is only one value here, which is 24, so again, this is indicating in this second bar, so again you can see here there is only here, one value which is indicating here. and similarly if I try to take here, the third one ,so third one has, the interval now, say 30 to 40 and then there are here three values 34, 35 and 38. So, the frequency here, f3 becomes here 3 and you can see here, that this interval is representing here, in this bar, where the frequency here is 3. And similarly in the fourth case, if you say observe means,  I can say simply here, that the frequency in this case is here 4, f4 here is 4 and this value is indicated over here and you can see here, that this frequency here is 4.

So now, you can see here, that the stem-and-leaf plot and the histogram, they are more or less comparable, the only difference is that, in case of a stem and leaf plot, it is also trying to give an idea, about the individual values, we're inside the histogram, the individual values are lost. So, I'm not going to compare to see their advantages or disadvantages, of these two graphics. But, I will say that depending on the need, you try to create a graphic, according to your need. Now, I will try to show you, this stem plot on the R console. So, firstly let me try to copy this data, so I can close this thing and then you can see here, this is the defective data here Right? And then I try to make it here, I try to create here, a stem. So you can see here, this is the same system, what was presented here? And then I try to play with the, scale part. So, I try to create here, the two actually this is what you have obtained here, this is actually here, the same width scale, with a scale equal to here one .so you can see here, that these two are the something. But, definitely in case if you try to add here ,say here, is scale equal to here 2. So, you get here, this thing. So, that is the same thing, what we have done here? So, I will now stop here and I will also try to close our discussion on different types of graphics, in one dimensional. There are some graphics, which are available for two-dimensional, when we have two variables. But, that I will discuss, when we are trying to deal with the data on, two variables, at this moment, first I will try to take up all the issues, when we have the data only on, one variable. So, now I will stop here, but definitely I am not saying that, that these are the only possible graphics available, there are many more graphics available and day by day, their new types of graphics are, also coming into picture, because of the use of software. But. I'm sure that this type of background. will surely help you and it will take out your, fear from your heart, that it is difficult to create graphics in R software, in comparison to a software, where you simply have to do some clicks. What you have to do? Just make some practice and if you practice, I'm sure that, you will be very very comfortable, in making more beautiful and more interactive graphics. So, you practice it and we will see you in the, next lecture with a, new topic. Till then, Good Bye.