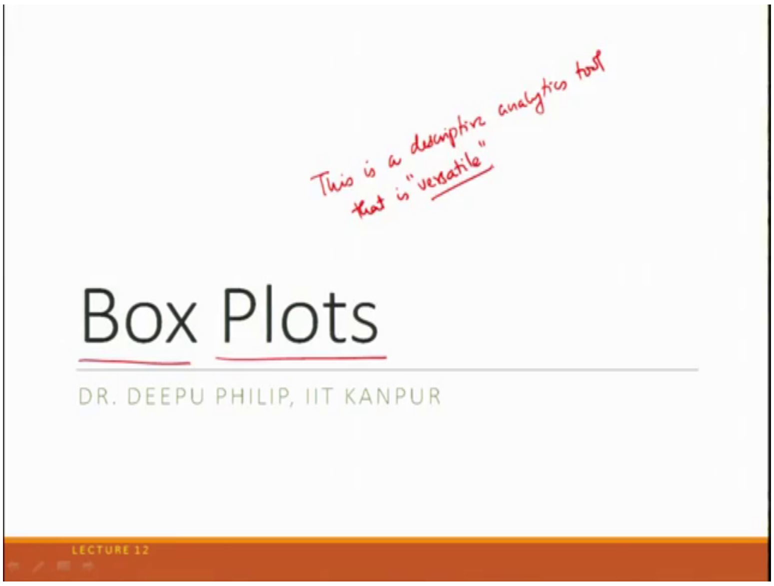
Practitioners Course in Descriptive, Predictive and Prescriptive Analytics
Prof. Deepu Philip
Dr. Amandeep Singh Oberoi
Department of Industrial & Management Engineering
Indian Institute of Technology, Kanpur
National Institute of Technology, Jalandhar

Lecture – 12
Box Plots

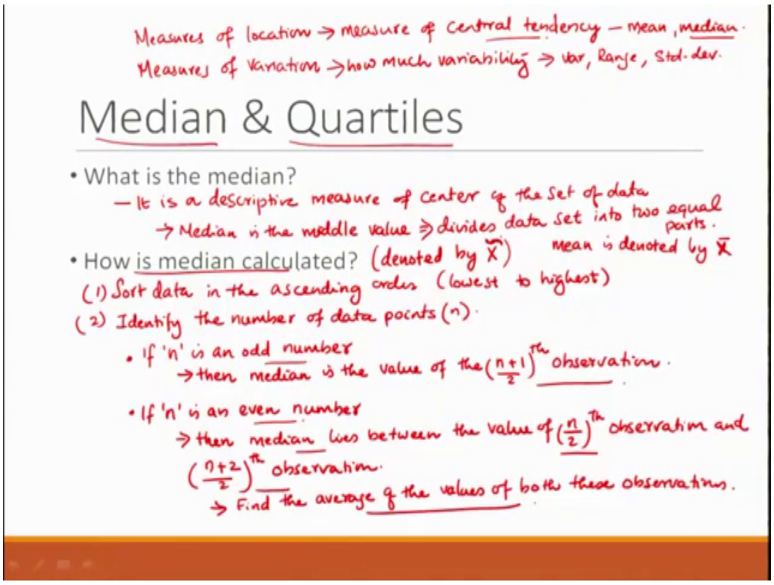
Good evening, welcome to get another lecture of the applied analytics course where we are looking at a practitioners approach to descriptive prescriptive and predictive analytics and we already seen the philosophy behind analytics, how is it different from data mining and the certain aspects of how do we look at the data describing the data and all this kind of stuff and today, we are going to deal with a new tool which is known as the box plots and box plots is a tool this is again.

(Refer Slide Time: 00:46)



This is a descriptive analytics tool that is versatile means it can be used for many other things and we will see the versatility of this tool in the lectures after this.

(Refer Slide Time: 01:11)



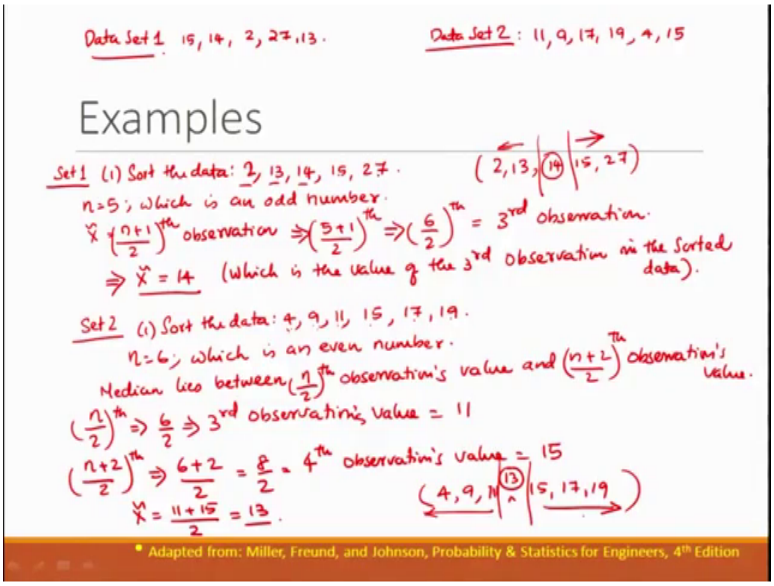
So, let us talk about the some of the fundamental concepts before starting with the box plots and first let us talk about median and quartiles. So, we all know that there is multiple ways to measure the. So, there is measures of location and measures of variation we already discussed this in earlier, but location is sometimes like for example, is measure of central tendency central tendency which means; where is the centre of the data located ok. So, couple example will be like mean median they are all that examples they are all part of the central tendency and measures of variation, we know that how much variability in the data how much variability in the data and some example is variance range standard deviation etcetera ok.

So, one such case of the measure of central tendency is the median and let us say what is a median and median it is descriptive measure is a descriptive measure of centre of the set of the data you are try to measure the centre of the set of the data or its also as said central tendency, but it is in a way median is the middle value which implies, it divides data set data set into two equal parts ok, ah, now the obvious question is how is median calculated, the we have not seen this in the class will quickly see how it is done the first thing to do it is sort.

So, how is the mean median calculated sort data in the ascending order? Ascending order means lowest to highest then second step identify the number of data points the number of data points usually a denoted by lowercase n small n if n is an odd number what we are saying here is if the n is a odd number then median is the value median is the value of the n plus 1 by 2; 2th observation if you find how many data points are there in the data set and then the value of n is odd; if it is an odd number is not divisible by 2, then the median is the value of the n plus 1 by 2th observation if n is an even number it can only be order an even see if it is an even number then what do we do then median lies between the value of n b 2th observation and n plus 2 by 2th observation ok, then find the average of the values of both these observations.

So, if it is the if n is odd is an odd number you find the n plus 1 by 2th observation and the value of that will give you the median if n is an even number then the median will lie between n plus 2th observation and n plus 2 by 2th observation and the average of these two values whatever the value set you get that is what is called as the median. So, the median is denoted by ok, it is also denoted by x tilde, tilde is the squiggly thing, mean on the other hand mean is denoted by denoted by x bar the line on the top ok, this is the tilde this thing the curvy or. So, let us look into an example quick example. So, let us take data set one.

(Refer Slide Time: 07:20)



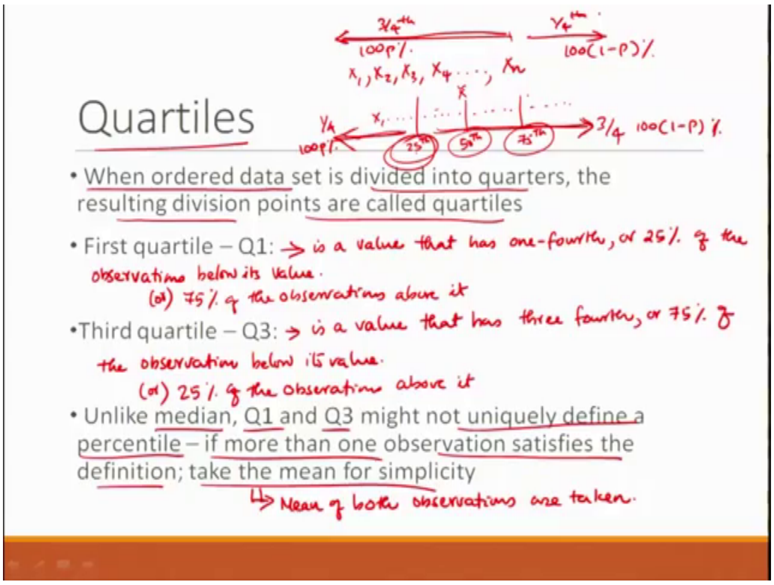
The first data set let us talk about it as 15, 14, 2, 27 and 13, and the data set 2, let us talk about second data set as 11, 9, 17, 19, 4 and 15, ok. So, let us start with the data set one. So, set one the first step is step one sort the data. So, the sorting will give us it will be giving as two, then it will give us 13, then it is 14, then 15 and 27, fine. So, the value of n equal to 1, 2, 3, 4, 5 equal to 5 which is an odd number odd number ok.

So, the rule is x tilde is the n plus 1 by 2th observation which implies 5 plus 1 by 2th which implies 6 by 2th, it implies the third observation ok. So, which is the third observation third observation the sorted order is 1, 2, 3. So, 14 is the third observation this implies x tilde the median is 14, ok, the value of median is 14 which is the value of the third observation in the sorted data remember that the data has to be sorted no matter what fine. So, now, similar let us look at the set two step again first one sort the data. So, sorting gives us we start with the sorting it is 4, 9, 11, 15, 17 and 19 ok. So, the value of n equal to 1, 2, 3, 4, 5, 6, n equal to 6 which is an even number ok; this is a number that is divisible by 2. So, even number then what do we do. So, then median lies between n by 2th observation value observations value n 10 plus 2th observation value observations value ok

So, what is the n by 2th observation, n by 2th implies, it is 6 by 2 implies 3. So, it is a third observation, value which is 1, 2, 3, 11, right, ok, now observation value then n plus 2 by 2th observation implies 6 plus 2 by 2 implies 8 by 2 or the fourth observations value and this value both come from the sorted order that is 1, 2, 3, 4 as 15 ok. So, now, median x tilde is equal to the midpoint of this, ok, 11 plus 15 divided by 2 which is equal to 13, ok. So, the value 13 which is between these ok; so if you look at it here you can see that it is 2, 13 is one set and the other one is 15 and 27 and 14 becomes the median.

So, it divides the data into two equal parts if you look think at there is two value and two values here, similarly, if you look at this case you can see that the values 4, 9, 11, then you have 15, 17 and 19 and you have 13 in between which is your median which divides data into two equal parts you can see that the 11 up to 4 to 11 is one part, this is the other ok. So, this is the simplest way how the data if the median is calculated ok. Now once you know how to calculate the median we can do many more fancy things out of this.

(Refer Slide Time: 13:01)

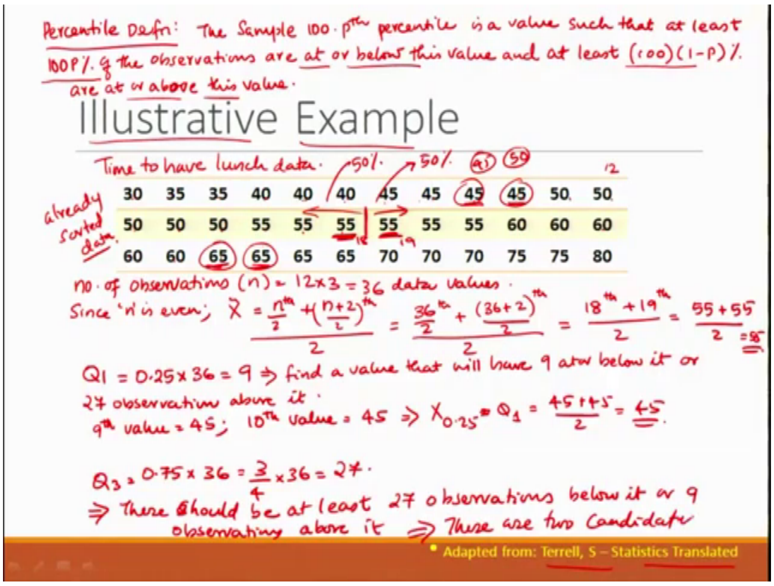


So, first thing to do let us say is that let us talk about the concepts of quartile quartiles. So, when the data is ordered when the ordered data is divided into quartiles or quarters then the resulting division points are called as quartiles ok. So, if you think about data has x 1, x 2, x 3, x 4, etcetera up to x n and if you divide them into quarters ok, you think about it as the first one x 1 to some number will give you the first quarter, then from here to median x tilde is second quarter then the third one is the data and then the last one above this.

So, if you divide this into quarter this is the middle point the fiftieth percentage this is 25th percentile and this is the 75 percentile ok. So, these kind of points, these percentiles point are typically what was called as the quartiles. So, the first one we going to study is the first quartiles or what we call as the Q 1 the first quartile, Q 1 is a value that has that has one-fourth or 25 percentage of the observations of the observations below, its value right. Similarly, if you talk about it Q 3, it is a value that has three-fourth or 75 percentage of the observations below its value and other way to think about these or 25 percentage of the observations above it ok.

Here you can think about it as or 75 percentage of the observations above it. So, as I said here when you think about the 25th percentile which is this side of it will contain the one- fourth of the data and this will be the three-fourth of the data to think about the 75th percentile this will be the three-fourth of the data and this will be the one-fourth of the data ok. So, unlike median which is exactly a middle point Q 1 and Q 3 might not uniquely define a percentile; that means, it may not be in exactly in the middle of two observations because then some point of time more than one observations can satisfy the definition then if that happens when we take the mean of both the observations here the mean of both observation are taken. So, let us talk about this general definition of percentiles ok. So, we will talk about in the next slide.

(Refer Slide Time: 16:49)



So, let us say before we get into this example let us talk about the definition of the percentile definition how do we define this the definition of the percentile the sample 100 time pth percentile is a value such that at least at least 100 times P percentage 100 P percentage of the observations of the observations are at or below are at or below this value and at least below this value and at least and at least 100 times one minus P percentage are at or above this value. So, if you look at this the definition says the 100 pth percentile of a value is a value such that at least 100 P of the observations 100 P percentage of the observations are at or below this value and at least 100 times 1 minus P percentage of observations are at or above this value. So, if you think about the previous case is early, I was talking about the 25th percentile means 25 percentage of the observations are below at or below this value or 75 percentages.

So, this is the 100 P percentage this is the 100 times one minus P percentage that is above the value, similarly this in our case is the 100 P percentage and this is the 100 times 1 minus P percentage in the case of the 75th percentile here is the three-fourth and the one-fourth ok. So, let us do an example and you should doing an example you should be helps us to understand how things are this is our time to have lunch data time to have lunch data and the credit of this data goes to Terral; Steven Terral's statistic translated book.

So, we will use this data do demonstrate or illustrate this examples. So, first is number of observations that is n which is 12 time 3; 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 and there are 3. So, the 12 columns and 3 rows there are 36 data values data values and this is anyway in sorted ok, this is already sorted data. So, the first step is already given to you sorted the 36 values. So, now, since n is even since n is even x tilde is the mean of n by 2th and n plus 2 by 2th observations average which is equal to the n is 36. So, is the 36 by 2 2th observations value and 36 plus 2 by 2th observations values average which is 36 by 1 is 18; 18th and 19th observations average is what we need to take.

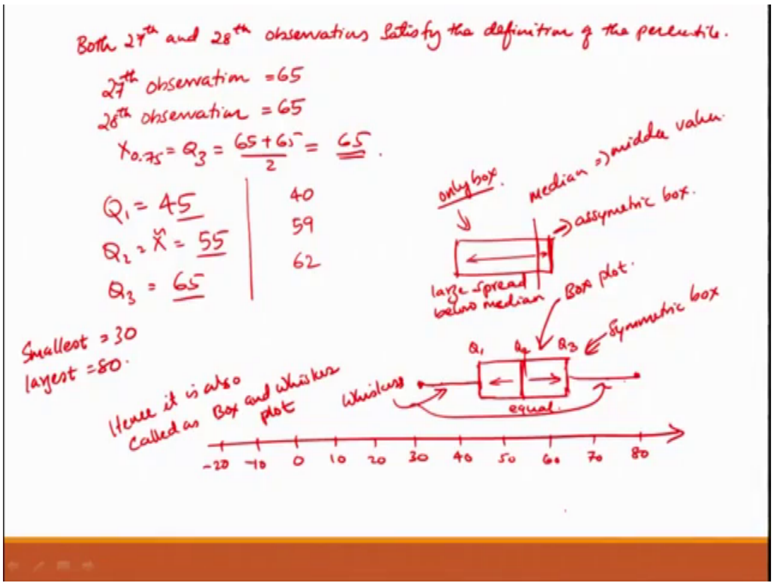
So, that is the first row is 12, then 13, 14, 15, 16, 17, 18. So, this is one 18th observation this is also the other one the 19th observation. So, the median can be calculated as 55 plus 55 by 2 which is equal to 55 ok. So, the median comes to be 55 and both the values are identical. So, the median will also be 55. So, what we are saying here is the value here somewhere here exactly will split the data the number of data points below of this is the 50 percentage of the data and the values about this is the other 50 percentage of the data that is why it is the median.

So, now, similarly plus do Q 1 Q 1 is the 25th percentile. So, which is 0.25 times 36 which is 9 which is one-fourth of the 36, 9 implies find a value that will have 9 at or below it or 27 observations because there are 36 of the remaining 26 observations above it ok, in this case, if you can think about it you can think about it as 1, 2, 3, 4, 5, 6, 7, 8, 9, this is the 9th observation, you can think about it and it will have 9 of them and above this will have 27 of them or if you take this 10th observations also you can find that from starting from here it will 9 below and it will also have the 27 above it at 10 above ok.

So, these two points will actually work in our case. So, then the 9th observation is 9th value is equal to 45, then similarly 10th value is also 45. So, x 0.2 or the Q 1, first quarter is equal to 40 plus 40 by 2 which will be 45 in software packages typically software packages like are an old it will if these values let us say instead of 45 and this one this was 45 and 50 if that was the case, then it would actually extrapolated between it correctly, but in our case we will just take the average of that if that is that was the case similarly let us calculate Q 3 the Q 3 is the 70-50 percentile 0.75 times 36 or 3 by 4 times 36 which gives us the value of 27 this implies that there should be at least 27 observations below it or 9 observations above it ok.

So, if you look at it the cardiac are this is 12 another 12, 24, 25, 26, 27. So, this is one point where the number of observations below this are 27 including this or it could think about here this 28th observation on which including this one you have 9 observations above it. So, we have two candidates which implies there are two candidates. So, in that if you think about it what we are going to do.

(Refer Slide Time: 25:01)

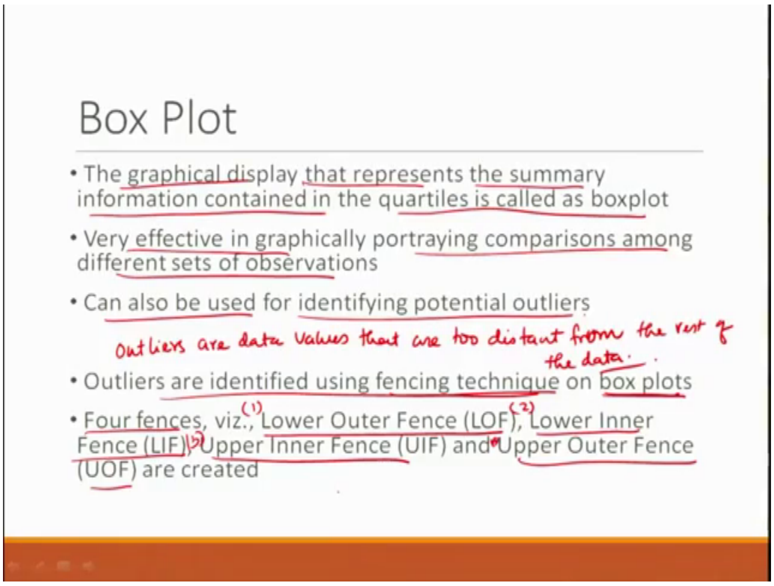


Now, is both 27th and 28th observations satisfy the definition of the percentile. So, what we do the 27th observation is as I said is 65 ok, 27th observation is the value is 65, 28th observation the value of the same is also 65 ok. So, the Q 3 x 0.75 the 17th percentile value of Q 3 is 65 plus 65 by 2 if find the average of it which is 65 again right.

So, we have now the Q 1 we calculated it as from the previous slide we calculate the value of Q 1 as 45 Q 2 which is your tilde the median where is calculated as we calculate the median as 55 and Q 3, the 75th percentile is calculated as 65. So, we have 3 of these values that we calculated now before getting into this let us take a scenario where we make this as a graph ok. So, let us take the x axis like long numbers scale it starts from here as 0, 10, 20, 30, 40, 50, 60, 70, 80 like this the number goes and they said you have minus 10 minus 20 like this and if we find out what is the Q 1 in this number scale Q 1 is 45.

So, it is somewhere here. So, this is your Q 1 ok. So, this is Q 1 45 and the Q 2 the at median is 55. So, will draw another line here 55 Q 2 and the third one is Q 3 which is 65. So, we will draw another line here which is Q 3 and if we connect all of them together with the rectangle this is called as the box or the box plot. So, plot is the graphical representation of the quartile information we will see the definitions later ok.

(Refer Slide Time: 27:57)



And so, if we go back to the slide the graphical display that represents the summary information contained in the quarter is called as the box plot ok.

So, when we talk about this whatever summery information contained in the quartiles how are they is clearly defined by boxplot assume that in a scale like this you had instead of this assume that the values happened to be 40 and 59 and 62. So, then the box would be something like 40 here, 59 will be closed to 60 and other one is 62 which will be close here and you will have a box like this ok. So, here you can see that it is a symmetric box plot because it is symmetric about the median this is equidistant equal here this is not symmetric it is asymmetric ok.

So, this is an asymmetric box. So, it tells you how the data is distributed with respect to the median and this is your median right median this is our median or the middle value here you can see that the data has a Q a large spread below medium ok. So, this is an important information or the spread above the median is also small. So, that kind of a thing and; obviously, the second part is it is a very effective in graphically portraying comparisons among different sets of observations.

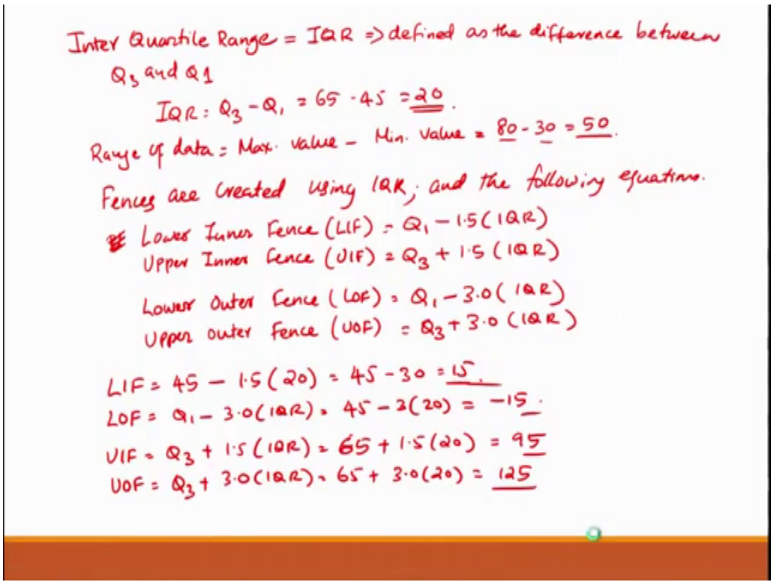
So, if you assume that these two are different box plot than we can easily compare them because we know that the Q 3 of the first box plot is lesser than the Q 3 of this and stuff like that and the Q 1 also less the median is that spread or these patterns are clearly demonstratable using this box plot then box plot can also be used for identifying potential outliers. So, before completing this before we getting into the outlier business let us point out which are the smallest values of box plot of the smallest and the largest values.

So, if we go back to the data the smallest value is 30 and the largest value is 80. So we come here this smallest is 30 largest is 80. So, if we take here is 30 you draw point and here is 80 you draw a point and you them with the box this is typically called as the whiskers ok. So, hence this is also called as hence, it is also called as called as box and whisker plot when you connect these smallest and the largest values with then it becomes the box and whisker plot. So, this is this is only box this is box and whisker this one is box and whisker you can see that the bottom one is the box and whisker plot ok.

Now, with this with this you can also say that the box plot is also a good tool for identifying potential outliers. So, outliers are. So, one we will see what outliers are later very clear definition, but outliers are data values that are too distant from the rest of the data that is what we call as the outliers and box plot is a good technique box plot can be enhanced if you add the fencing technique on to the box plot then you can use it to identify the outliers in the data for the outliers in the data there are 4 types of fences in this number one is our lower outer fence typically called as a LOF number two is the lower inner fence typically called as a LIF the third one is called as the UIF upper inner fence and the fourth one is called as the upper outer fence called as the UOF. So, we create all these fence on to the box plot and for using that we try to identify the outliers

So, let us see how to make the fences ok. So, the first thing before making the fences;

(Refer Slide Time: 32:56)



We need to understand this deifier concept called inter quartile range ok. So, remember in our previous example, we had the Q 1 as 45 and Q 3 as 65. So, using that we can identify what is called as a topic called inter quartile range typically known as IQR defined as defined as the difference the difference between Q 3 and Q 1. So, IQR in our case is equal to Q 3 minus Q 1 which is pretty much 65 minus 45 which is equal to 20. So, this is called as inter quartile range. So, what is the range of the data range of data is equal to max value minus minimum value which is 80 minus 30 which is 50. So, the IQR is the range between the Q 3 and Q 1 whereas, the data range is between the min and the max value.

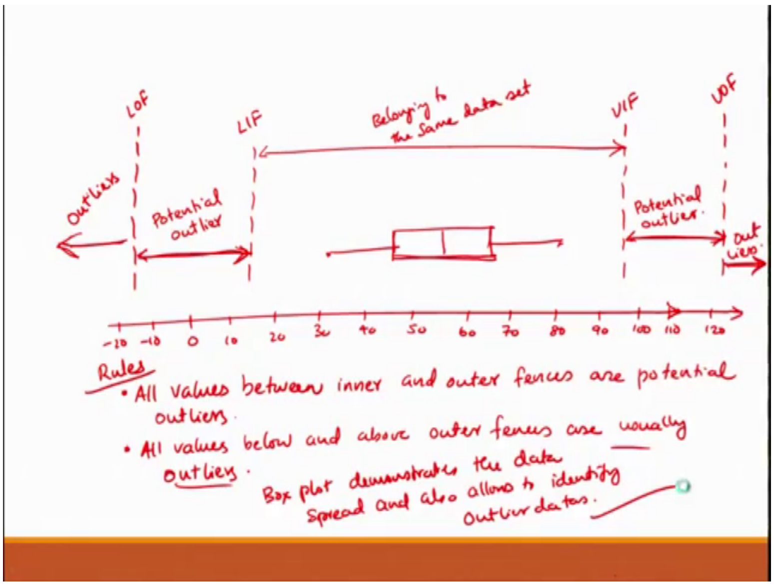
So, the fences are created using IQR equations IQR and the following equations what are those equations upper let us start with the lower fences to make our life easy we will start with the lower inner fence or LIF this is given by Q 1 the quartile the first quartile 1 value Q 1 minus 1.5 times IQR that is your lower inner fence. So, upper inner fence will then be symmetric on the other side UIF will be do not be Q 1 it will be Q 3 plus 1.5 times IQR that will your upper inner fence and the. Now let us talk about the outer fences lower outer fence the LOF is equal to Q 1 minus 3 times IQR and upper outer fence outer fence that is u o f is equal to Q 3 plus 3 times IQR ok.

So, if you visualize this in this case you already know; what is the values of Q 1 and Q 3? So, in our case we calculate it we can get it the LIF is equal to Q 1 the value of Q 1 if

45 and that is minus 1.5 times IQR we calculate the value as 20. So, that is 45 minus 45 minus that will come to 20; 20 half the thirties. So, that will be equal to you get the value of 15 ok, then similarly upper inner fence, let us talk about the lower outer fence first lower outer fence will be again Q 1 minus 3 times IQR which is equal to 45 minus 3 times 20 which will give us the value as minus 15 ok.

Similarly, the upper inner fence is equal to Q 3 plus 1.5 times IQR which is equal to for it will be what to 65 plus 1.5 times 20 which will give us 95 and the upper outer fence is equal to Q 3 plus 3 times IQR which is equal to 65 plus 3 times 20 which will give you one 25. So, these are the fence values that we get using these fences we can embellish the box plot ok. So, we will see how to do that. So, as we did earlier.

(Refer Slide Time: 37:38)



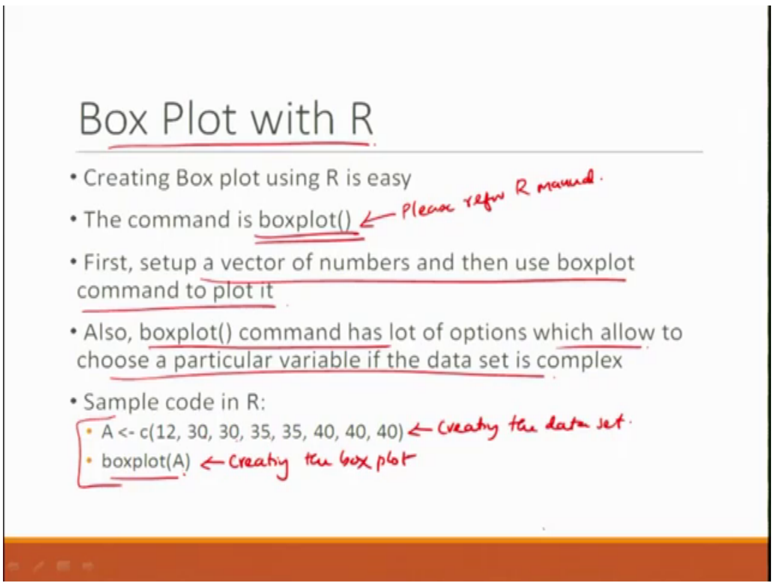
We will draw the line again just to remind the how to draw the box plot once again 0 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, let us assume it this way and here we have minus 10 minus 20 like this and the starting value of the fence was 45. So, we start with the Q 1 45, Q 2 as 55 and Q 3 as 65, all the 3 values drawn like this and the minimum value was 30 maximum value was 80, they are connected to the box which heats the box and whisker plot this is how you draw the box plot.

Now the question is the fences. So, the first value of the fence is at 15. So, 10, 15. So, this is where the lower inner fence will come into picture and the upper inner fence was at the value of as we seen earlier is was the value of 95. So, we come to 90 and 95 this is where the upper inner fence will come into picture. Now we see; what are the lower outer fences? So, lower outer fence was minus 15. So, if you go to the diagram we will fine minus 10 minus 15 will be write here.

So, it will be the lower outer fence and similarly if you look into it we find out that the upper outer fence is at the value 125. So, we come here at 120, 125 and this is our upper outer fence. So, the trickier is that this fencers we can find out the outliers the idea is this all values between the rule the rules are this all values between inner and outer fences are potential outliers ok. So, this area this is a potential outlier ok, this is another potential outlier then all values below and above outer fences are usually outliers ok.

So, this area and this area these are outliers. So, the values between these inner fences are supposed to be belonging to the same data set. So, as long as the values are lying between this inner fences we can call it as they are kind of behaving like the other data points, but the values between the inner and outer fences typically become what you call as the potential outliers and the values outside the outer fences below the outer fence and above the upper outer fence becomes outliers designating outliers. So, box plot demonstrates the data spread and also allows to identify outlier datas ok. So, this is one good advantage of the box plot then; obviously, if you are doing box plot then one question is been have a large data set it is very hard to do box plots with your by hand.

(Refer Slide Time: 42:16)



So, you would require software and in this case the particular software that we are creating here is the using as we said this class we will using R box plot they are extremely easy to do with R compare to excel and the command that you need to learn to do it is the box plot function. So, it is basically box plot. So, what you need to do is first you need to set up a vector of numbers and then use box plot command to plot the same.

So, first you set up the number and then you use the command and the command also has lot of options which you can learn and find out how which allows you to make quite lot of different complicated type of box plots, but the simple code is like this you give a data setting this is the creating the data set ok. So, you create a vector a with it is a set of 12, 30, 30, 30, 35, these kind of values and then you ask into here is what you are doing creating the box plot ok. So, when you we use this box plot command and feed the data value the R software will then do the box plot for you and you can art the fences and all other aspect the box plot by manipulating this box plot command.

So, to find out the details of this please refer R manual ok, but to make a box plot in R with the large data set is very simple is just two lines of quote with that we come to the conclusion of the box plot descriptive analytic tool today. And in the next session we will now start with the some aspects of a, what is data warehousing what is business intelligence what is data mart and that kind of concepts and then we will start with a normal distributions after which we will move towards hypothesis testing which is the fundamental corners throne of data analytics thank you for your patience listening and will we will see you soon.

Thank you.