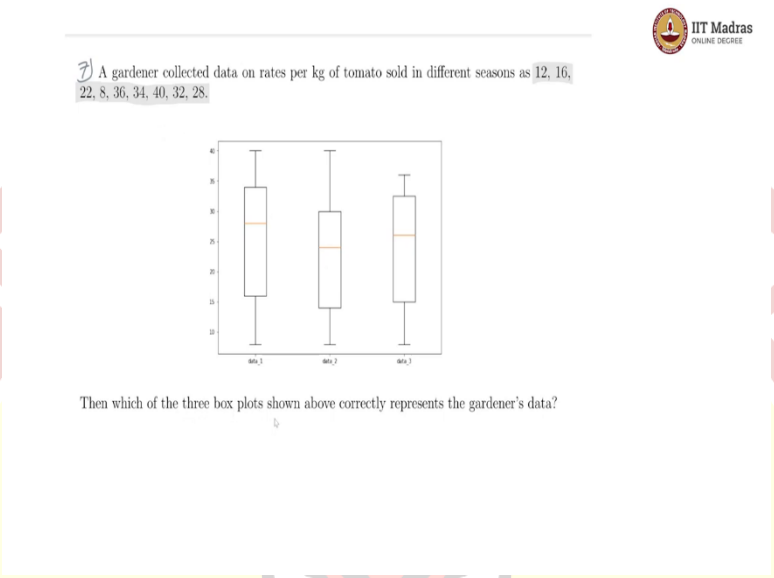
IIT Madras

ONLINE DEGREE

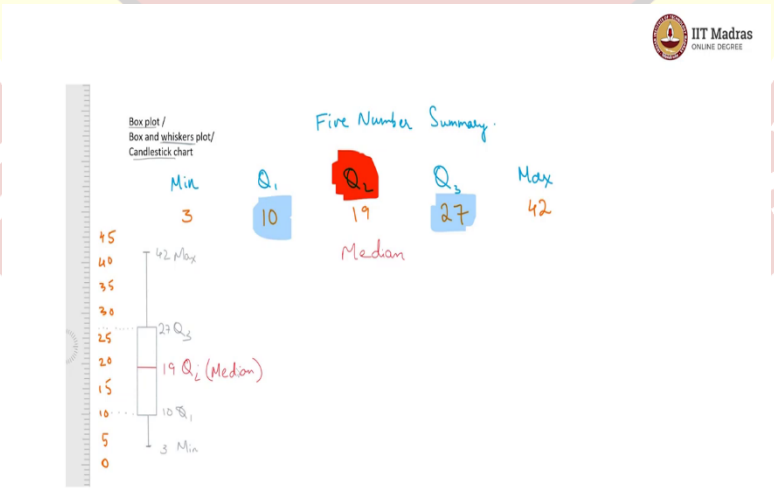
Statistics for data Science 1
Professor. Usha Mohan
Department of Management Studies
Indian Institute of Technology, Madras
Week 3- Box plot tutorial

(Refer Slide Time: 00:14)



Seventh question, as a gardener who collected data on the rates per kg of tomato, sold in different seasons. So, these are the data points which are 12, 16, 22, 8, 36, 34, 40 and 28. Then the question which of these 3 box plots correctly represents the gardener's data.

(Refer Slide Time: 00:50)



So, for this we have to first understand what a box plot is, so the box plot is called number of things, so it is called a box plot or box and whiskers plot or also candle sticks chart, these are

the various names given to the plot, and it is essentially a chart a plot to represent the five number summary. So, in the five-number summary what do we do we have something which is the minimum of the data, then we have the first quartile, then we have the second quartile and then we have the third quartile and finally we have the maximum of that data.

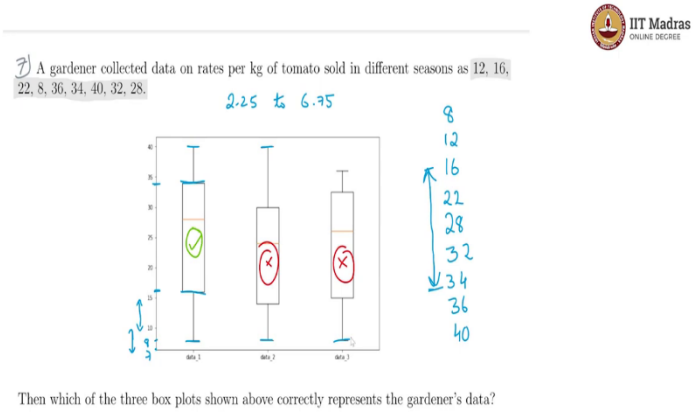
Let us imagine some data where the minimum is let us say 3. And the maximum is let us say 42 and the Q_1 is occurring at suppose 10, Q_2 is occurring at suppose 19, Q_3 is occurring at suppose 27. So, this is how the data is, we have found the five numbers summary of this data suppose. Now, for the box plot we first establish this markings vertically like this letters in this ruler, let us consider this is 0, this is 5, 10, 15, 20, 25, 30, 35, 40 and 45.

So, this is how the ruler is 0 is here and 45 is here, so what we do in a box plot is we draw a box a rectangular box from the Q_3 to Q_1 so Q_1 to Q_3 we draw a rectangular box which looks like this. So, here we can see that the upper part of the box is Q_3 and the lower part of the box is Q_1 10 to 27. So, this is basically nothing but the interquartile range Q_1 to Q_3 is the interquartile range and we are basically showing a box to represent the interquartile range.

So, this is why we call it the box plot and the whiskers part is that along with this box we also draw a vertical line from the Q_3 to the max here which is 42 and here we mark it of like this. And we do the same thing on the other end from Q_1 to the minimum which is 3, so here this is to 3 so this is our minimum and this is 42 which is our maximum. So, in this way the min is shown the max is shown the Q_1 and Q_3 are shown and what is left is the Q_2 which is also incidentally the median.

So, to indicate the median what we do is we draw this little line that is going through the box. So, this is our 19 which is Q_2 which is also the median because Q_2 is exactly 50 percentile which is a median of the data. And this plot is what is called your box plot or box and whiskers plot or sometimes even the candle sticks plot because it looks like a candle stick. Now, given this introduction, let us go to the question and see how to solve it.

(Refer Slide Time: 05:00)



So, since it is box plot we need to first rearrange our data as a ascending order we need to arrange the data so we will have 8 first 8 goes first, 12 appears to be next, 16 as after that, 22 comes after that, 28 is here and we have 32, 34, 36 and 40 so overall the 9 observations. So, first of all which box plot has a range of 8 to 40, so this box plot does not seem to be starting from 0 because this appears to be 5 units this length and this is even lessees on 5 units.

So, maybe this value is probably 7 and then 8 is likely to come about here, so all of these box plots seem to match that the upper limit the other side of the range is 40 for these two. So, either of these two could be our box plot and this one is definitely wrong. Now, let us look at the interquartile range, so we have 9 points we have seen. So, the interquartile range will come from if we did it exactly by as 25 percent, we would get from 2.25 to 6.75 and that means from the third value to the seventh value which is this.

So, 16 into 34 should be shown in our interquartile range. And that is happening for this box plot so this is about at 16 and this is at 34, so this is wrong and this is our correct box plot.