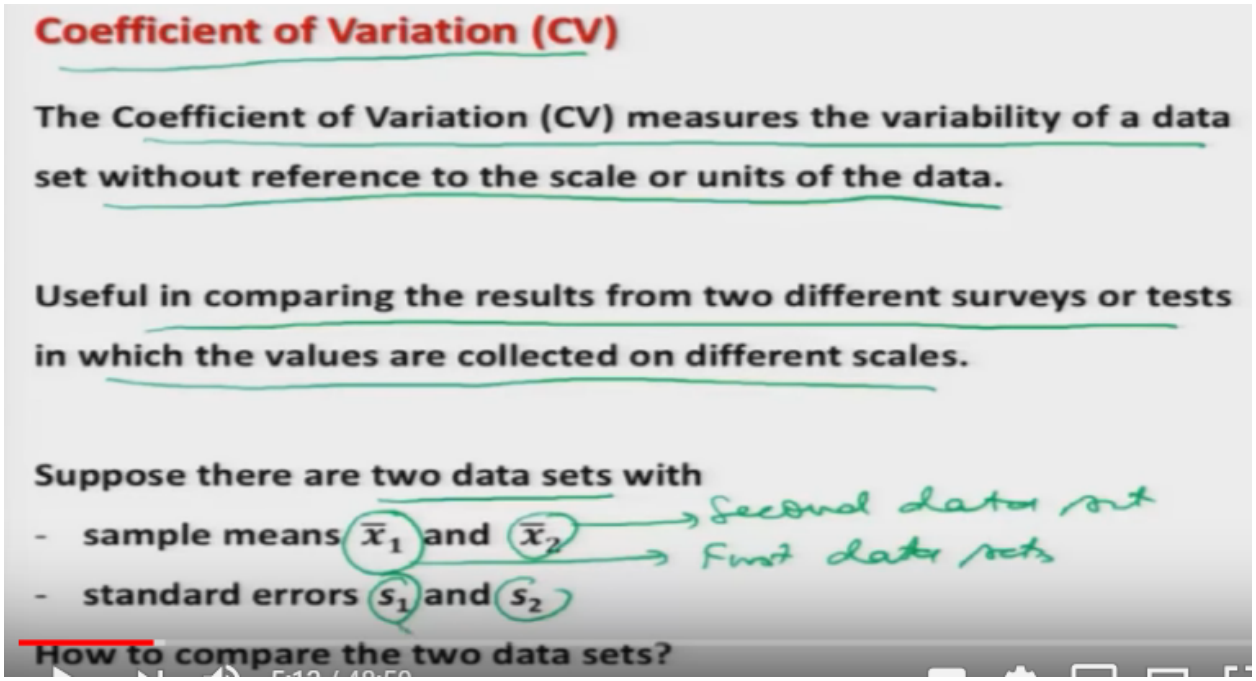
Lecture – 21
Variation Data – Coefficient of Variation and Boxplots

Welcome to the next lecture on the course descriptive statistics with R software. Now, you may recall that, up to now what we have done? We have considered two aspects of data, one is the central tendency of the data, and another is variation in data. And, both these aspects they are very important part of the information which is contained inside the data. Now, I am coming to another aspect, suppose we have data set and we want to know the variation in the data set that should also depend on the measure of central tendency. What does this mean? Up to now we have taken the aspects of central tendency and measure of variation separately, one by one. Now, I would like to have a measure which can inform me the information contained inside the data, base on arithmetic mean and variance. This will help us in getting an idea about the variability in the data in various types of situations. We are using either the mean or the variance may not really be advisable, and may not really give us the correct information. So in this lecture, we are going to first discuss about a tool or stat scale tool to measure the variation, this is called

as coefficient of variation. And, after this I will try to consider one quantitative measure, and say another graphical measure based on the R software, to have combined information on various aspects.

Refer Slide Time: (02:38)

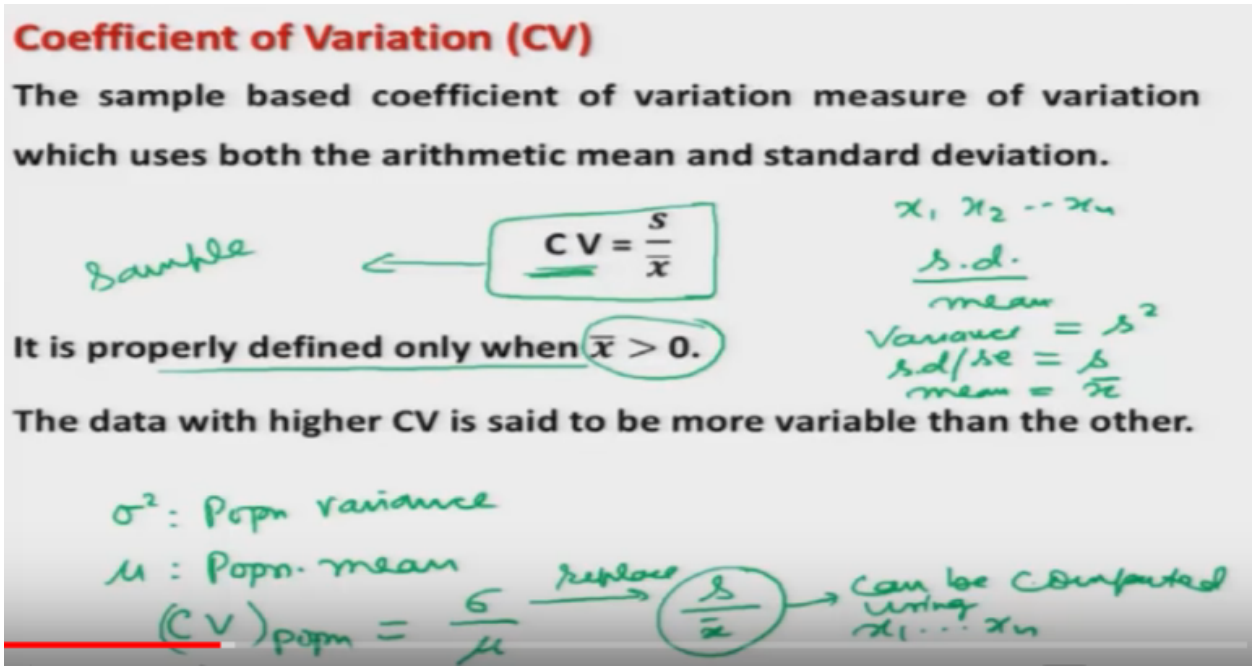


So, let us start our discussion with the coefficient of variation. The, coefficient of variation measures the variability of a data set without reference to the scale or units of the data. What does this mean? Suppose, I have got two different data sets, one is measured in say centimeters and says another is measured in say meters. Incase if you simply try to combine the mean and variances of the two data set and if you try to compare that, it might be little bit difficult. Similarly, in case if you want to compare the house rents say in India, and say house rent in US, the house rents in India they are given in Indian rupees, and they are with respect to the salaries that we get here. Similarly, if you go to US, the house rents are going to be in terms of US dollars, and they are also with respect to the salary structure in US, and the salary structures in US and India, they are very different. S, sometimes you have heard that people simply try to multiply the dollar by the exchange rate and they try to say that, Okay. I am earning this much or I'm spending this much, or I am paying this much of house rent. So, how to handle these types of situation, that can be done using the concept of coefficient of variation, Right.

So, this coefficient of variation will be useful to us, when we try to compare the results from two different surveys or they are obtained from two different tests and they are obtained on different scales. For example, if I say that I have got here two data sets, and we try to find out the arithmetic mean and standard errors of the two data sets. So, the sample mean of the first data set is obtained here they say x1,

first data sets mean and X by 2 is the arithmetic mean of second data sets. And, similarly I try to find out the standard errors s1 and s2, so a standard error s1 corresponds to the first data set, and standard error has to correspond to the second data set. Now, there are two aspects mean and standard errors or central tendency or variation. Now, how to compare the two data sets, that is the question that we are going to entertain here.

Refer Slide Time: (05:30)

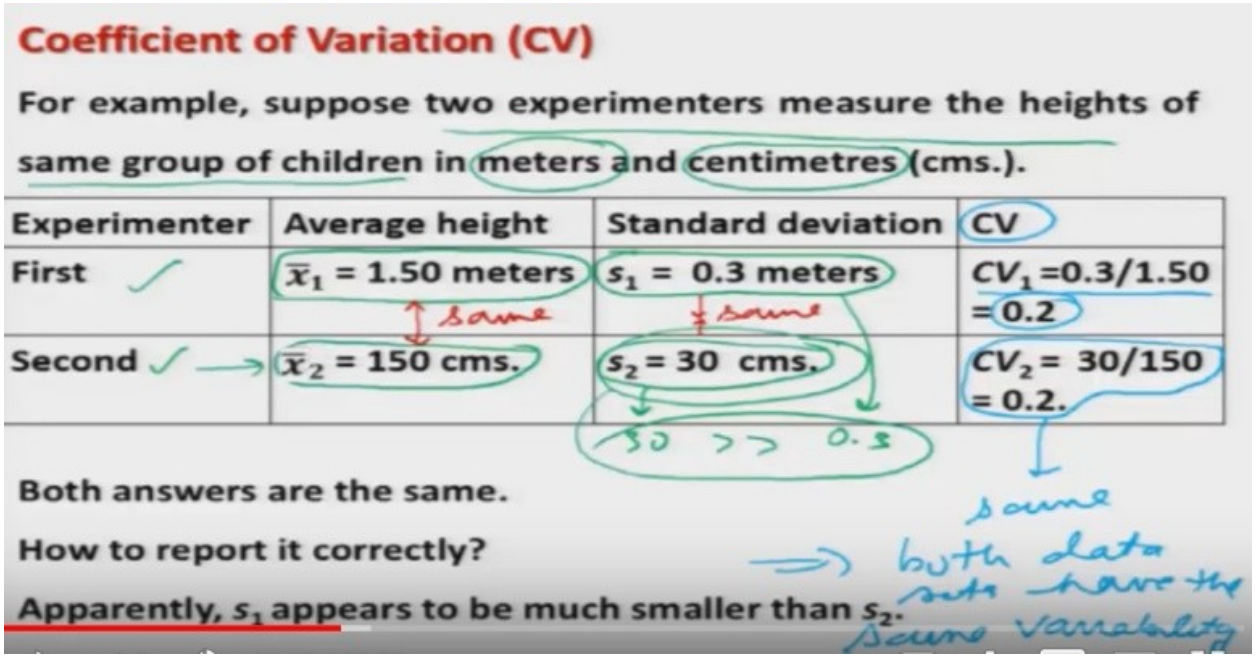


Now, the definition of coefficient of variation, as we had discussed in the earlier lecture in case of variance. That the variance can be for the entire population, which is usually unknown or the variance can be for the sample, that is called as sample variance, which is computed. Similarly, in the case of coefficient of variation, we have two versions, one for the population and one for the sample. But, here when we are trying to discuss the tools of descriptive statistics, we want to compute everything on the basis of given sample of data. So, I am going to discuss here the sample version of the coefficient of variation. Please keep this thing in mind. Okay? So, once I have got the data say x1, x2,…., xn, this can be either for the group data or say and group data or whatever you want. Then the coefficient of variation is defined as standard upon mean. So, if you remember that we had denoted the variance, sample variance by s square. So, the standard deviation or in simple language we call it standard error, whatever you want to call. I am trying to consider both as a similar meaning without any loss of generality. This is denoted by say here s. And, sample mean we already have denoted by x bar. So, this coefficient of variation briefly denoted as CV, this is defined as s upon x bar. Now, if you try to see a standard deviation will always be positive. So, this coefficient of variation is properly defined only when the mean is positive or X bar is

greater than zero. See here this definition of CV; this is based on the sample. Now, if you want to understand, what is the population counterpart? Then, if I say that Sigma square is the population variance. And, MU is the population mean, then CV of this population will be defined here as a Sigma upon MU. But, since we do not know the value of a Sigma or MU, so we try to replace it by s for Sigma and X bar for MU. So, this gives us the sample based definition of the coefficient of variation, which can be computed using the data x1, x2,…, xn. Right?

Now, the next question is how to take that decision? Because, coefficient of variation is also a measure of the variation in the data. So, if I have two data sets, then how we are going to measure it. And, now you can see here, that if I try to take here two data sets. In which, suppose the arithmetic mean of one data set is greater than the arithmetic mean of data set 2. And, suppose the standard deviation of first data set is smaller than the standard deviation of, of data set 2. So, what is happening? In one data set mean is larger, but the standard deviation is less, and in other data set just opposite, is happening. In that case, which of the data you have to choose? That cannot be answer directly by using the mean or standard deviation. So, in these situations the coefficient of variation helps us and we say simply try to find out the coefficient of variation of both that data sets. And, the higher value of coefficient of variation is going to indicate that the variability is higher. So, the data with higher CV is said to be more variable than the other.

Refer Slide Time: (10:14)

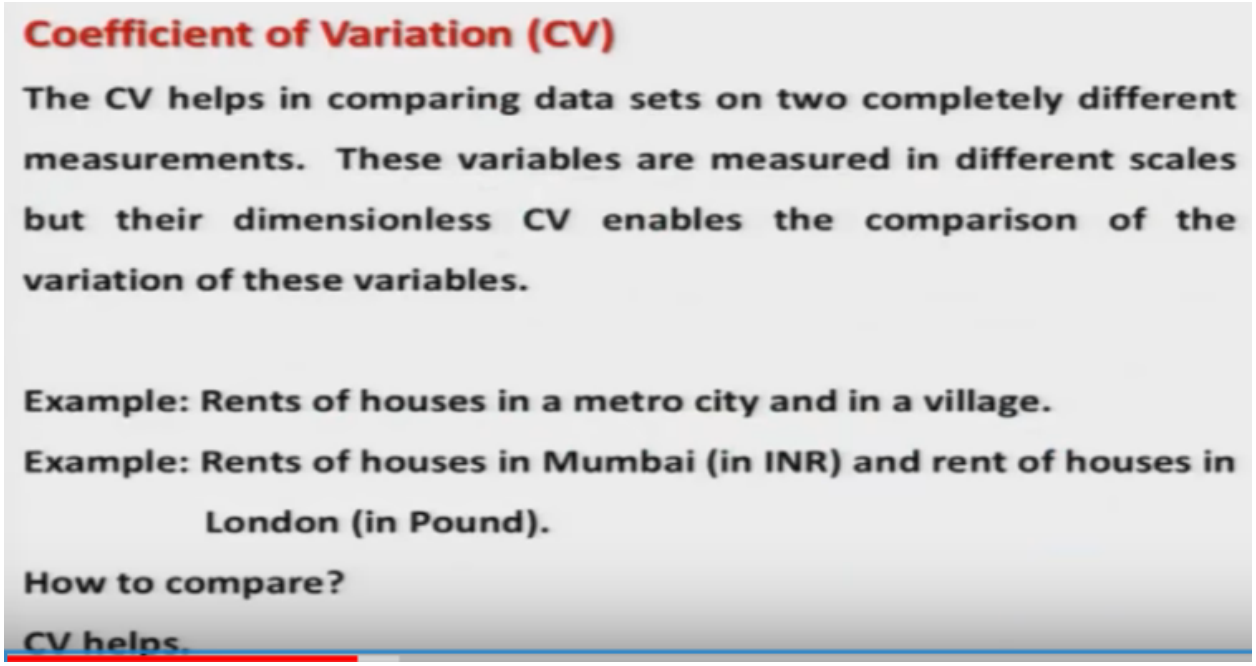


So, that is again similar to the interpretation of variance, higher the value of variance that means more variability, has the value of CV that means there is more variability. Just to explain you in more detail,

had we taken a simple example? Where two experimenters have collected the data on the heights of same group of children, one experimenter has taken the observations in meters and others experimenter has taken the observations in centimeters. And, they have found the average height and standard deviation of the two data sets. So, you can see here, I have tabulated the information, this is the first experimenter, second experimenter. And, first experimenter has found the average height to be 1.50 meters, and a standard deviation to be 0.3 meters. And, similarly the second experimenter he has found the average height to be 150 centimeters, and the standard deviation to be 30 centimeters. Now, this is the usual tendency to compare the standard deviations. So, you can see here, here the value of a standard deviation is 30. Whereas here this value is here 0.3 in the first case. So, obviously in the first look it appears that 30 is much much greater than 0.3. And, this indicates that the variability in the second data set is very very high. But, this conclusion is actually wrong. Because, if you try to see both sets of measurements, they have been taken on different scales, but they have got the same value. The heights say x1 bar and say here x bar 2 they have the value 1.50 meters and 150 centimeters, which are the same. And, similes and similarly standard deviation, they are point 3 meters and 30 centimeters they are also the same.

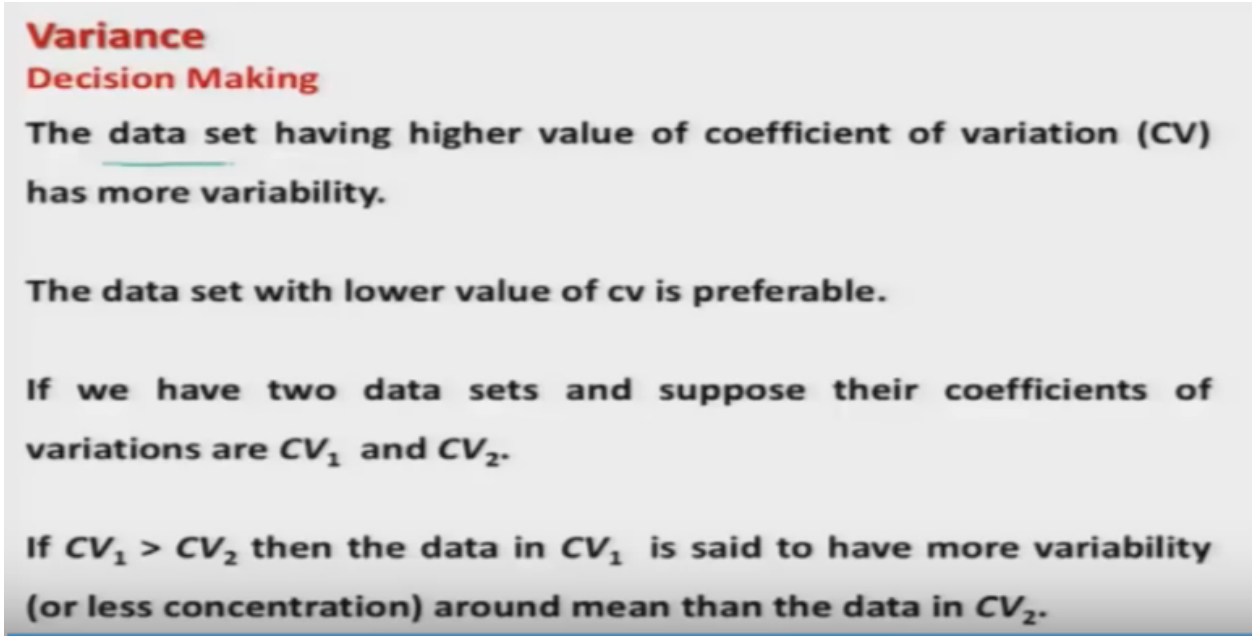
So, now in this case how to report it or how to identify it, how to know it, that is the question. So, in this case, this coefficient of variation comes to our help and risks you. So, if I try to find out the value of coefficient of variation in both the cases, then in the first case the coefficient of variation comes out to be standard deviation divided by mean, which is 0.3 divided by 1.5, and this comes out to be see here 0.2. And, in the second case also the coefficient of variation comes out to be 30 upon 150, which is equal to 0.2. So, you can see here, that both the values are same. And, this is indicating that both data sets have the same variability. And, this was not possible by looking only at the values of mean and standard deviation.

Refer Slide Time: (13:24)



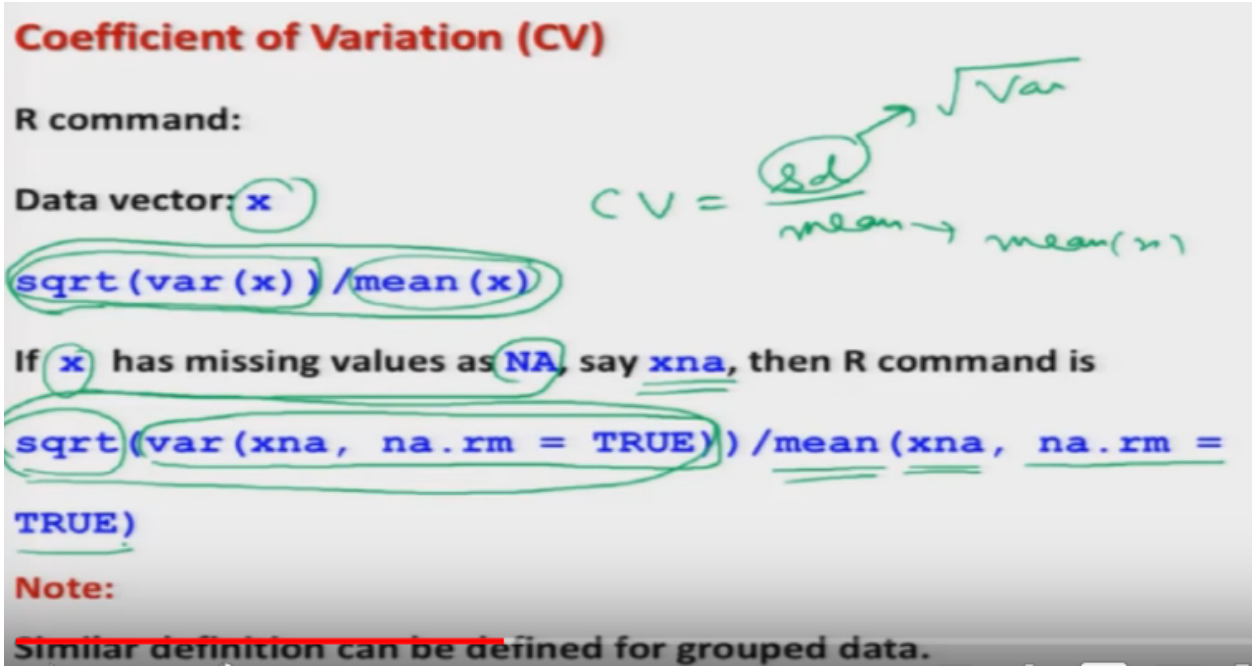
So, similarly this coefficient of variation also helps us in comparing the data sets on two completely different measurements. These measures, these observations can be obtained on different scales. But, the advantage what coefficient of variation has, that coefficient of variation is dimensionless. So, this helps us in the comparison of the variation in two data sets. For example, India if I take an example of rents of houses in a metro city and in a village. We know that the rents in, in a metro city in India they are very high, where is the rents in a village that are also very, very low. And, similarly if I try to take say another example the rents of houses in Mumbai, they will be in Indian rupees and the rent of houses in London, that will be in pounds. Now, how to compare when the data has been obtained in the same unit, as in the first case when we are trying to find out the rents in India in a metro city and a village and when the data has been obtained in different units. As in the case of, a range of houses in Mumbai and London, so how to compare? Then again this in this case this coefficient of variation helps us.

Refer Slide Time: (14:57)



Now, that question is how to use this concept of coefficient of variation in making a decision? So, the data set having higher value of coefficient of variation is thought to have more variability. So, definitely when we have lower value of coefficient of variation this is a variable. For example, in case if I have two data sets and suppose we have computed their coefficient s of variation as CV1 and CV2. Then, if CV1 is greater than CV2, then we consider or we say that the data in CV1 has more variability or say less concentration around the mean value than the data in CV2 or in the second data sets. And, similarly in case if I have the opposite that CV1 is a smaller than CV2 that means the data and CV1 has a smaller variability than the data in CV2, Right?

Refer Slide Time: (15:53)

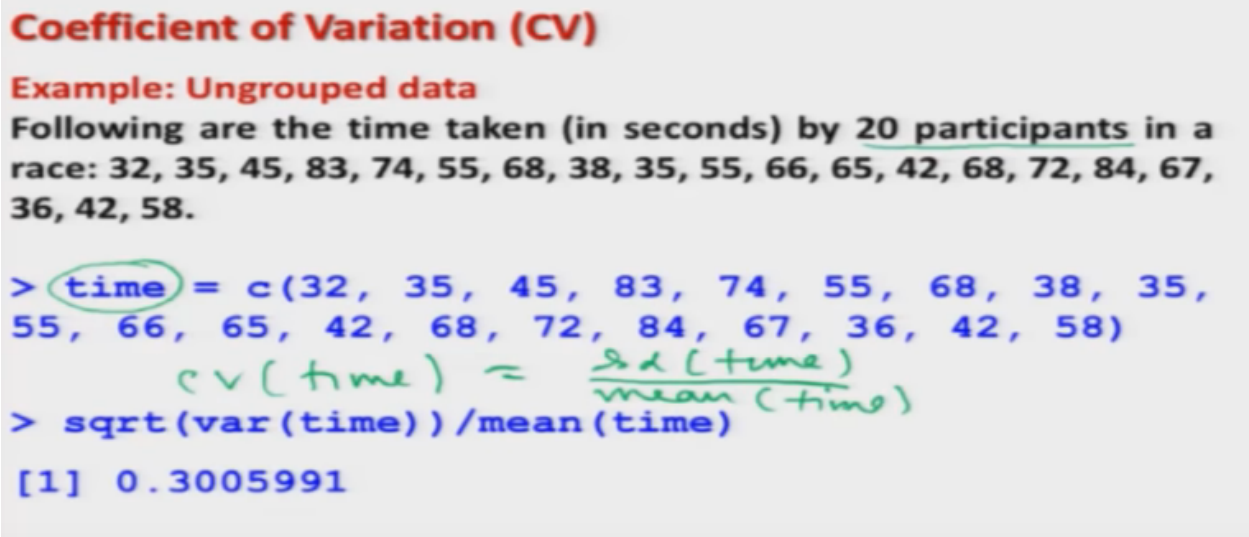


Now, the next aspect is how to compute it on the R software. So, as such there is no built-in command inside the R software to compute the coefficient of variation. But, computing the coefficient of variation is very simple and straight forward. This is only the function of standard deviation and arithmetic mean. And, we already have learnt how to compute the standard deviation and how to compute the arithmetic mean. So, just by using the same commands, we can always compute the coefficient of variation. So, this is how we are going to compute the coefficient of variation. Okay. So, if I say that we have got a data vector x then the coefficient of variation is going to be defined as like this. What is this? Coefficient of variation is simply here standard deviation upon mean. So, a standard deviation was nothing, but the square root of variance and mean was by the function here mean of x. So, this is what I'm trying to do, first I'm trying to find out the square root of the variant that is that will give me standard deviation divided by mean of x. And, this will give me the value of coefficient of variation.

Now, if I ask you that how would you compute the coefficient of variation in case the data is missing, then it is very simple. How? We already have learnt how to compute the standard deviation when the data is missing, and we also have learnt how to compute the arithmetic mean when the data is missing. So, you simply have to use the same syntaxes, same functions and you have to write down the syntax for computing the coefficient of variation. So, if you recall that what we had done earlier. Suppose my data vector x has got some missing values which are denoted by NA, and I'm denoting this data vector as xna. So, now I would try to find out the variance of this x and a by using the command inside the argument na dot rm is equal to true. So, this will help us in finding out the variance when the data vector has missing values and then we try to find out the square root, which will give me the standard deviation. So, this

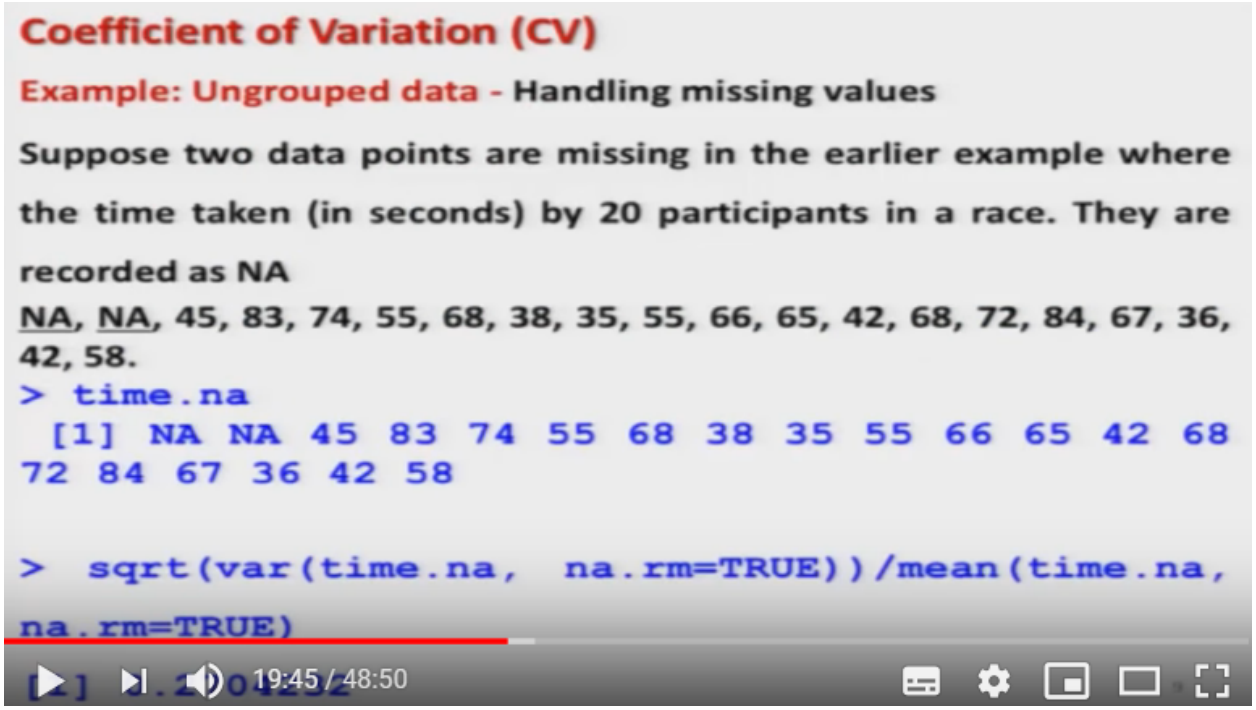
function will give me the standard deviation in the presence of missing data. And, similarly the mean function on the data vector xna with the argument na dot rm is equal to true, will give me the value of the mean when the data is missing. And, by using this command, we can always find out the value of the coefficient of variation. And, now using the same command you can also write down the syntax and command for computing the coefficient of variation in case of grouped data, that is not so difficult, Right?

Refer Slide Time: (19:03)



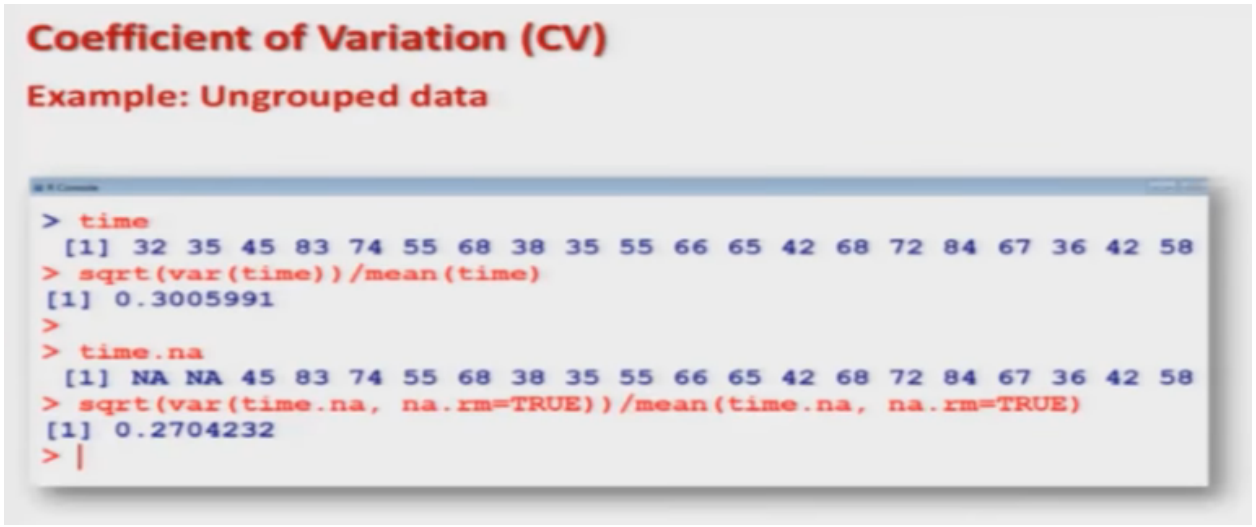
Now, I will try to take a small example to show you that how to measure it. Suppose, I have collected the data on 20 participants, in the time taken in a race and this data has been recorded inside a variable time. So, this is the same example that I have used earlier, now incase if I want to find out the coefficient of variation of time, so CV of time that is simply here standard deviation of time divided by mean of time. So, I am using here this syntax and this is giving me this value 0.3 and so on.

Refer Slide Time: (19:45)



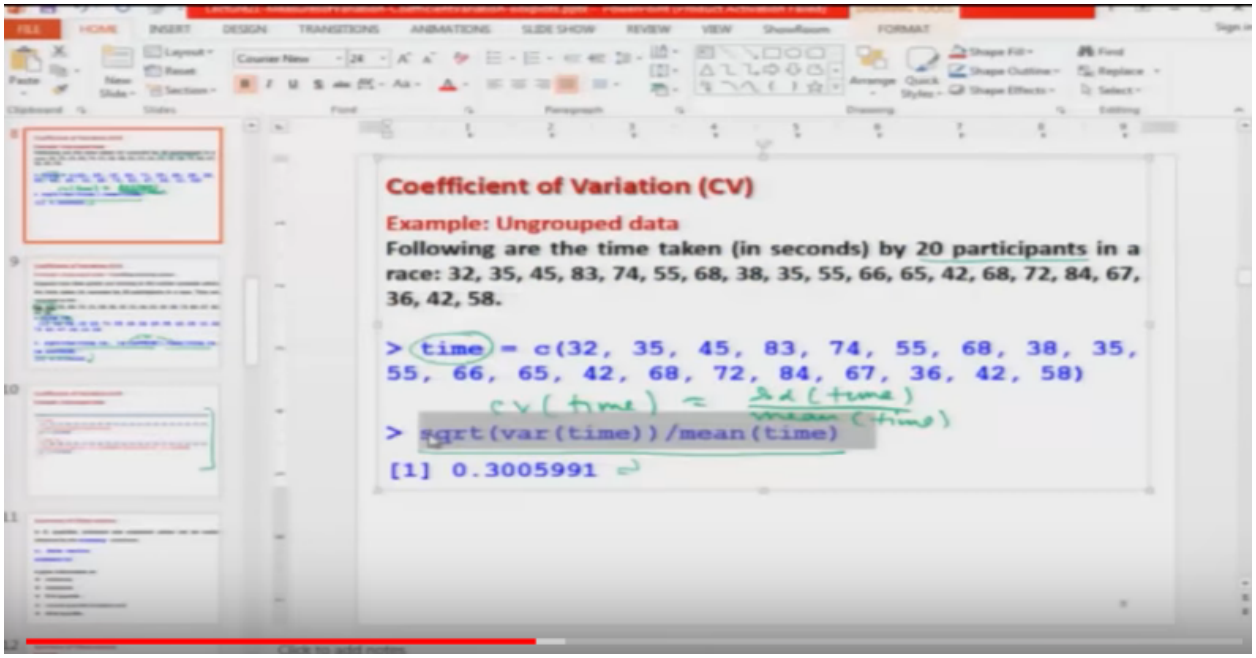
And, suppose if the data is missing, then in that case I have the same data in which I have replaced first two values by na, and this data has been stored inside a new data vector time dot na. And, I try to use the same command to compute the standard deviation in the presence of missing value and the command for mean, in the presence of missing values and I try to take the ratio and this will give me the coefficient of variation. And, this value comes out to be here 0.27.

Refer Slide Time: (20:19)



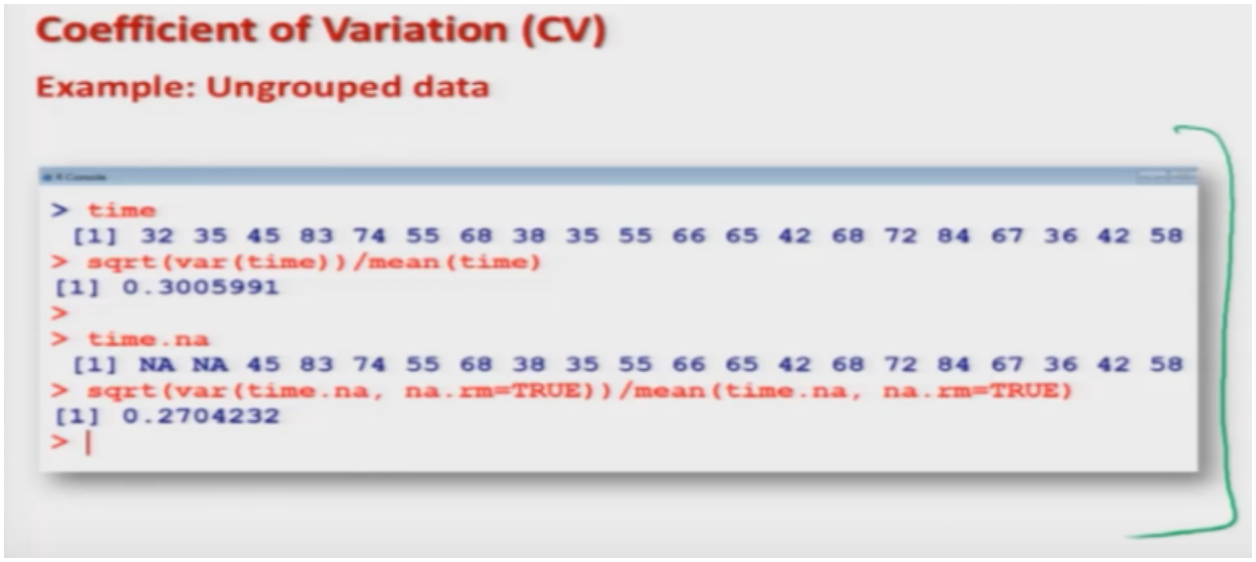
And, you can see here this is the screenshot of the same thing what I have done. Now, I will try to show you on the R console also.

Video Start Time: (20:27)



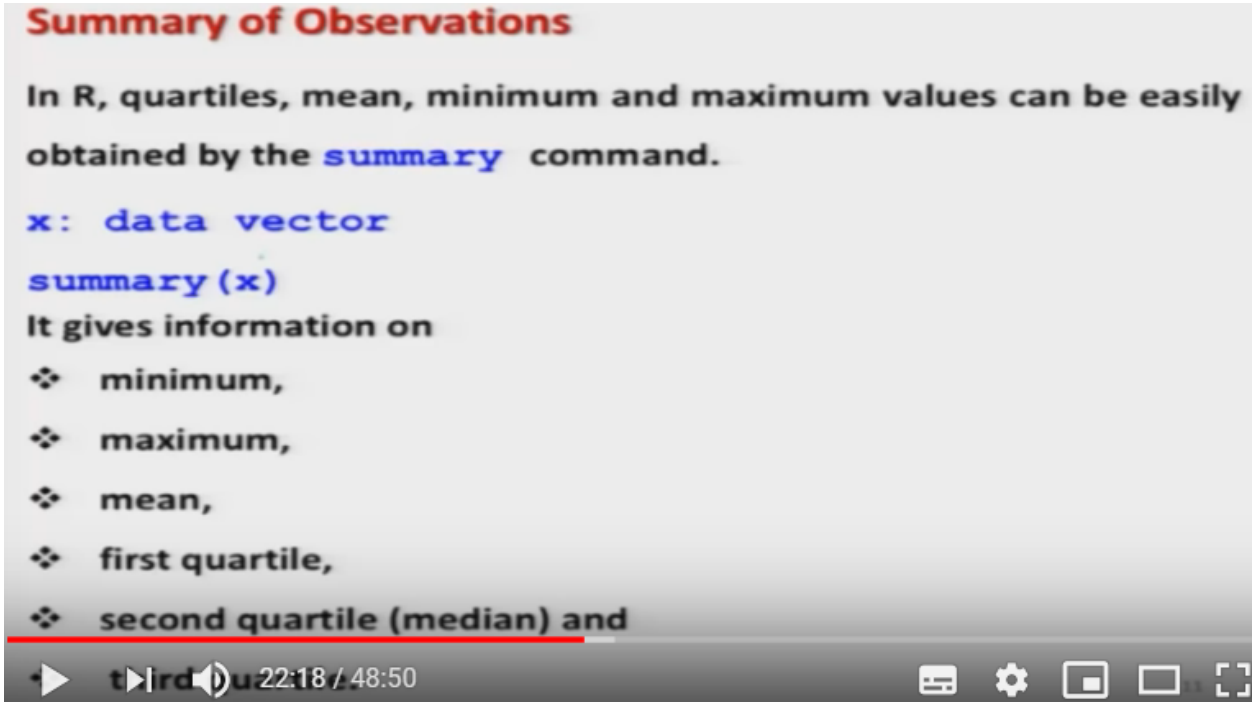
I'm trying to take the command and this data has already been copied, which is the same data set had a time. And, if I try to find out the coefficient of variation, this is giving me this thing, Right. And, similarly incase if you want to find out in the presence of missing values. Then, I already have stored the data time dot na, you can see here and if I try to use the same command on this data set, you will get here the same value. And, the same value has been reported here in this slide, this is the screenshot.

Video End Time: (21:08)



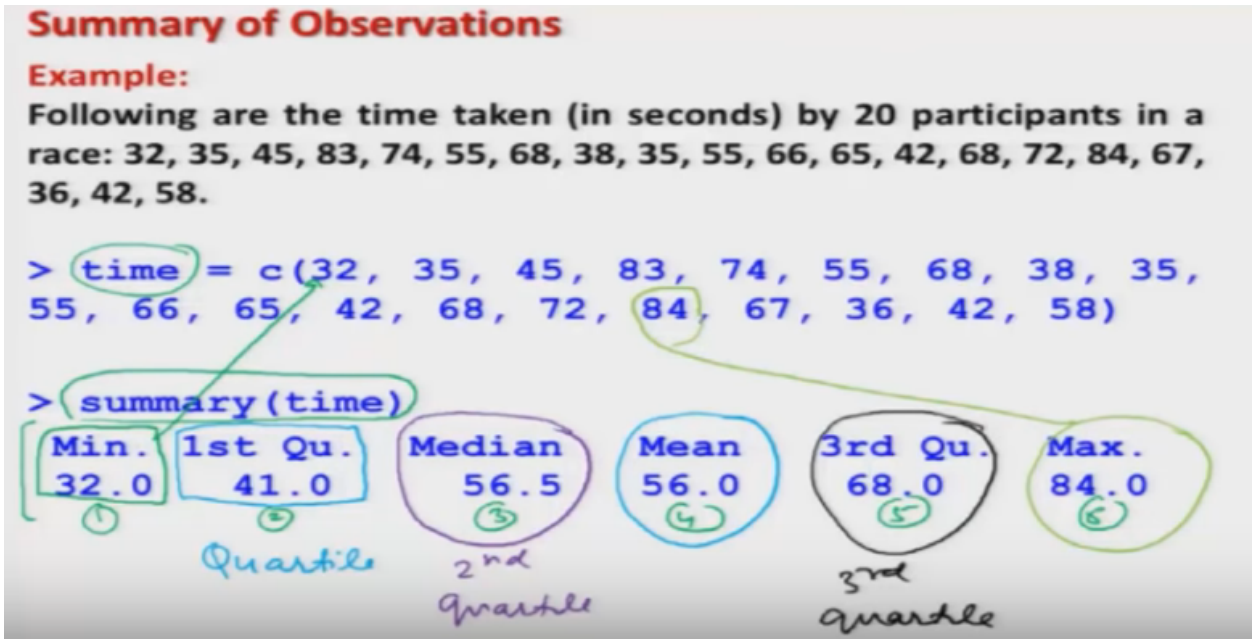
Now, I have completed different types of measures of variation that we had aimed. Now, I'm trying to address another aspect, when we get the data, then data has all sorts of feature. And, we would like to have all sorts of feature like measure of central tendency partitioning values variation and so on. Up to now what we have done, we have taken all these aspects one by one. Means, how to compute maximum among the data values, minimum range, quartiles and so on. In R software there is a command, by which you can compute all these values like as minimum value maximum value and different types of coal tiles in the single shot.

Refer Slide Time: (22:18)



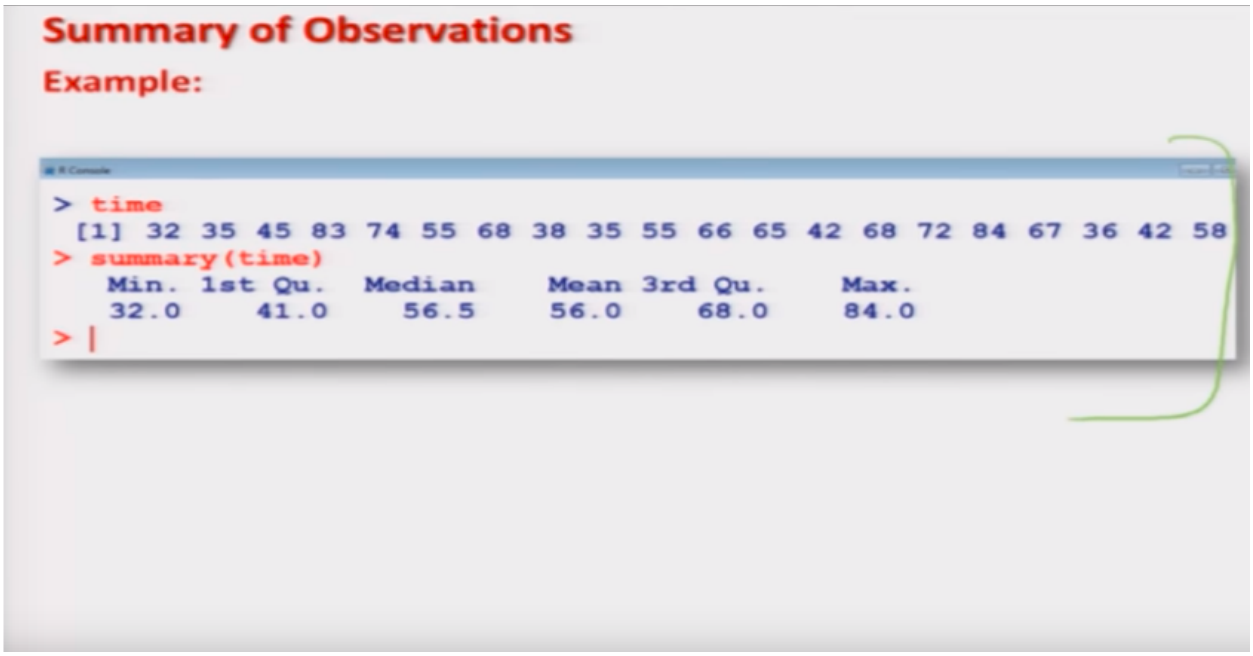
So, I'm trying to discuss now this command, which is a summary command. So, in R, there is a command, summary (s u mm a r y) . And, this summary commands provides us a comprehensive information on different types of quartiles, mean, minimum and maximum values of the data sets. So, if my data vector is denoted by x, then we use the command (s u mm a r y) summary and inside the arguments here x. And, if you try to use this, then the outcome of this command will give us an information on the minimum values, minimum value, maximum value, arithmetic mean, first quartile, second quartile this is median and the third quartile of the data set which is contained inside the data vector x.

Refer Slide Time: (23:13)



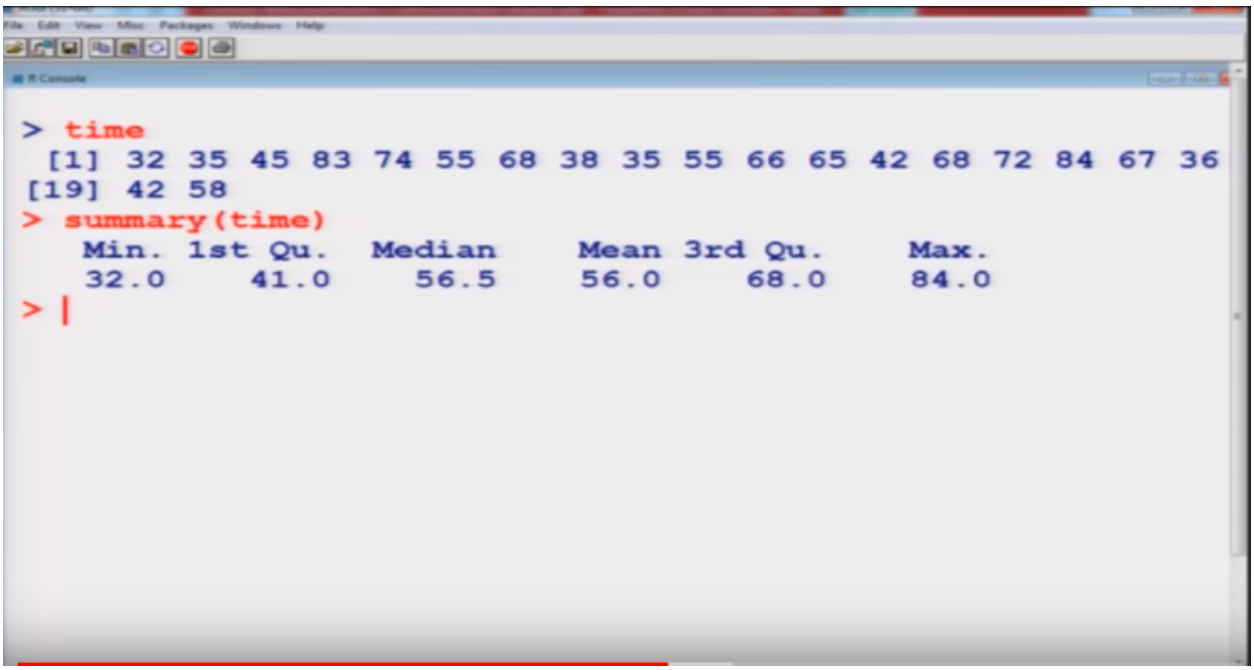
Now, let us try to take an example to understand it. And, now I take the same example, that, that I took earlier where we have the data on 20 participants in the time taken in a race and the data has been stored in a variable here time. So, now when I say summary time, then on the execution we get here an outcome like this one. So, you can see here there are 6 values, first value here, second value here, third value here, fourth value, fifth and sixth. So, now if you try to understand what are they trying to give us. The first value here is giving us the minimum value. Minimum of the values contained in that time data vector, which is 32, and we can see here this is here the 32. Similarly, the second value giving us the value of the first quartile, remember this is quartile not the quantile, Right. So, you know how to compute the first, second or third quartile. So, but here in this case you need not to compute it separately, but it will give you inside the same outcome. Similarly, the third one is the median value, which is the second quartile. Third value is here the mean that is the arithmetic mean. Similarly, fourth value is the value of the third quartile, and the fourth and the last one the sixth one is the maximum value of this observations, which is here 84, you can see here this is the value. So, by using this summary command you can get all this common information in a very comprehensive way and that is the advantage here.

Refer Slide Time: (25:30)



And, here you can see here this is here the, the screenshot. I will try to show you on the R console also so if you try to come here.

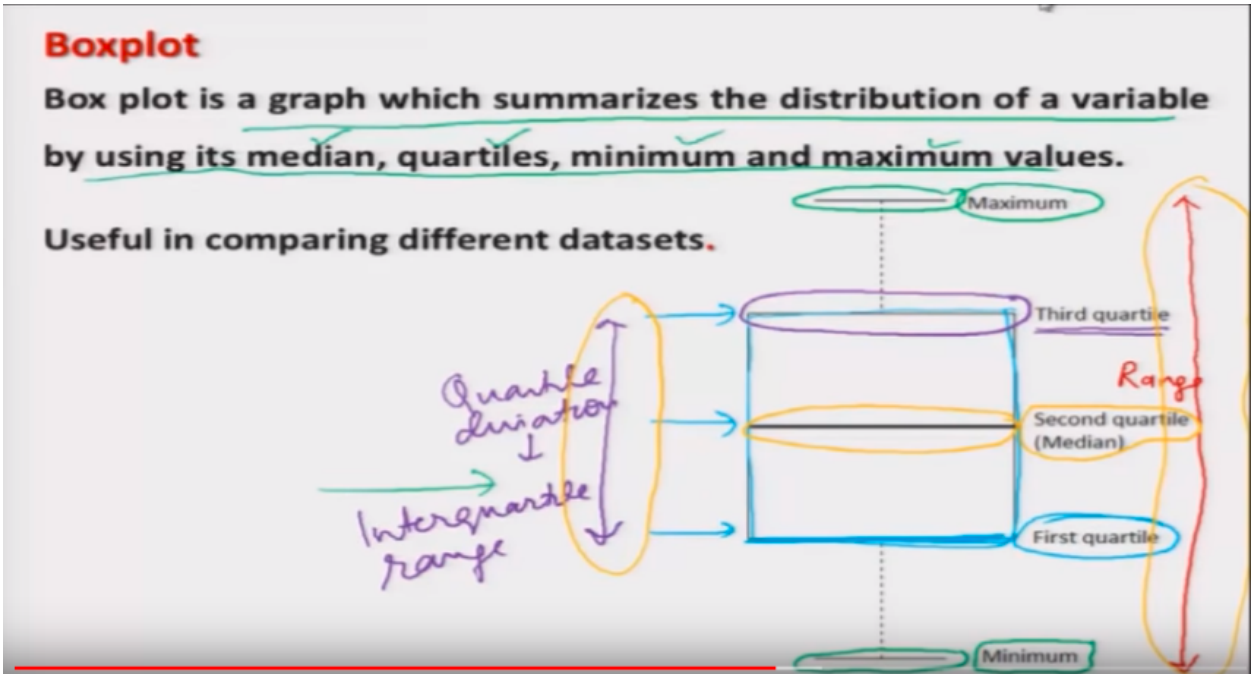
Refer Slide Time: (25:54)



If you say see here that data on time here is already stored there, and if I try to write down here summary of your time, we get here this value, Okay. But it cut us now come back to our slides and we try to discuss now another aspect. So, now you can see that, this summary command is trying to give you different type

of information minimum, maximum, quartiles, means in a comprehensive way and when we started the discussion on descriptive statistics I had told you that there are two types of tools, one are quantitative tools and other are graphical tools. So, now the next aspect is this whatever the information has been provided by the summary command can it be represented in a graphical way now, and now, What will be the advantage of this thing? Suppose if you have two data set or even more than two data sets and if you want to compare all the characteristics at the same time, you can use the summary command on all the data vectors and also you can create the graphics. So, graphics will give you a visual comparison of the information contained inside the different data sets, and in order to do so there is a graphic which is called as box plot. So, now I try to discuss what is a box plot, and how to construct it inside the R software, okay.

Refer Slide Time: (27:24)

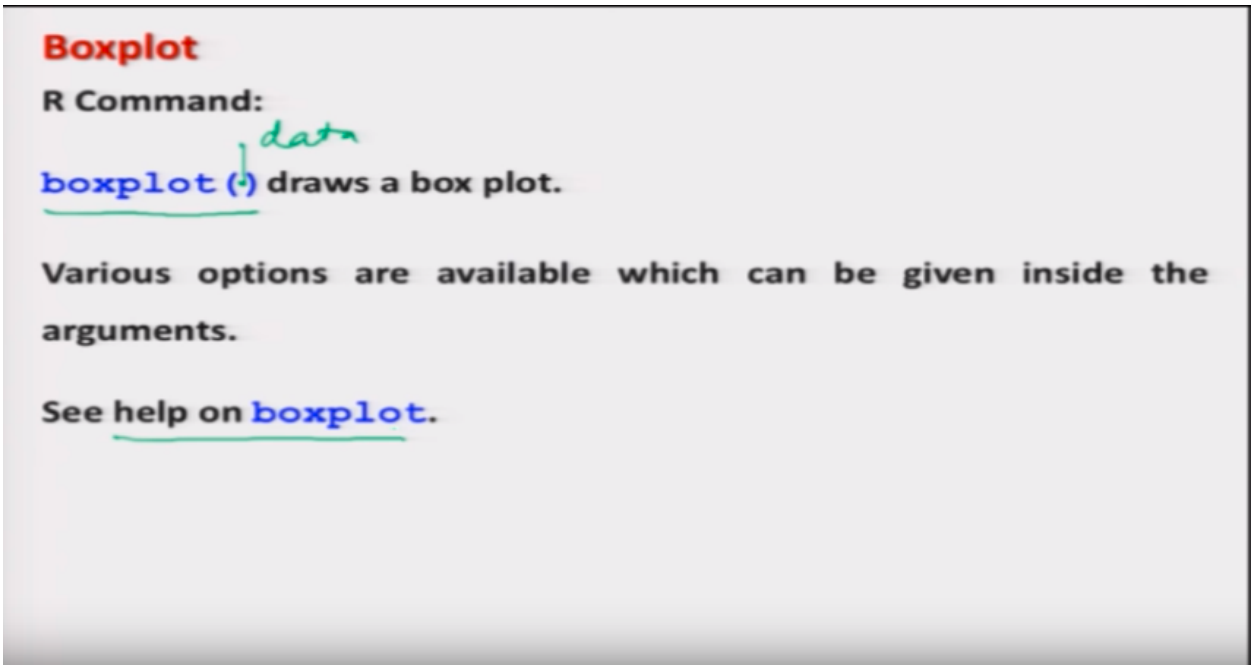


So, this box plot is actually a graph. We summarize the distribution of the variable by using different types of information like as median, quartiles, minimum, maximum. I remember this will not give you the value of the mean. So, this box plot looks like actually this. How you can see here the graph, you can see here there are two lines here one in the bottom please try to observe the pen. One in the bottom and now other in the top in green color you can see here these two values are giving the minimum value of the data set and the upper value is giving the maximum value of the data set.

Now, in case if you try to find out the difference between the minimum and maximum value, What you will get? you will get here the value of range, and similarly if you try to look on these three lines which I

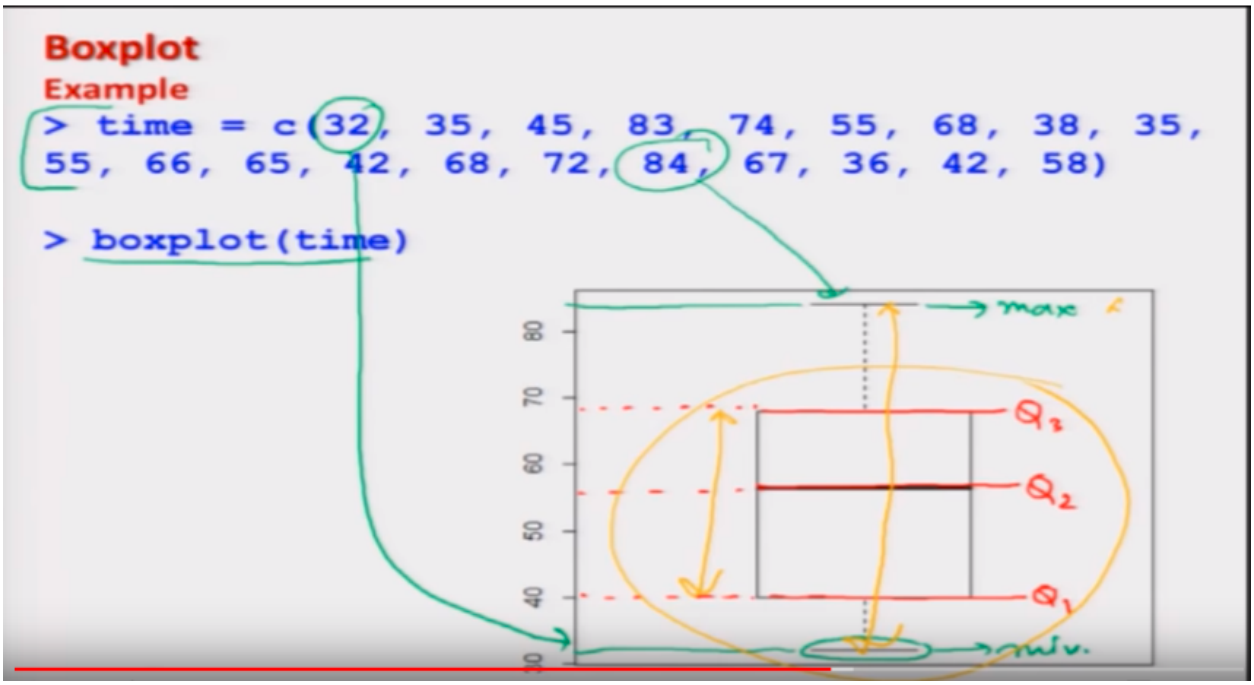
am indicating here first second and third. So, if you try to see this is here as sort of here a box and the lower edge of the box, which is here. That is giving you the information on first quartile. Now, I will between the color of the pen, so you can see it clearly. Similarly, the upper edge, which is here, this is giving us the information on third quartile. So, if you try to see if you try to find out the difference between the two, don't you think that this will give you an idea of the quartile deviation and also in some sense, it will give you the information on interquartile range. These two measures we had discussed as the measure of variation. Now, finally if you try to look in the line in the middle of the box. This line is going to give you information on the median, which is the second quartile Q_2 . So, you can see that inside this box there are several measures which are combined and just by looking at this difference, and this difference you can also compare the variation by looking at the middle values you can compare the median and so on.

Refer Slide Time: (30:12)



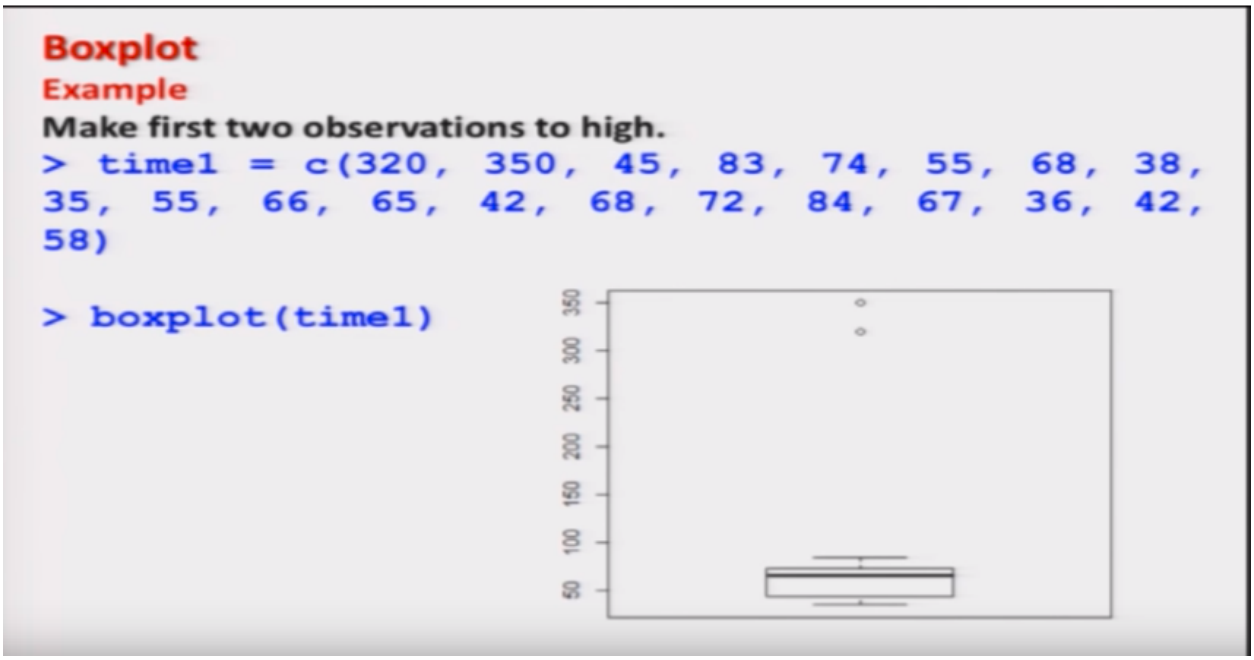
So, let's try to first see the applications through the software inside the R software there is a command here box plot and inside the arguments you have to give the data vector for which you want to create a box plot and there are several arguments, which are several options available in the plotting of box plot. I would request you to go to the help menu and try to see what are the different arguments, and what are their uses in creating the box plot, you can make the legion's, names, colors etcetera, shapes, etcetera.

Refer Slide Time: (30:45)



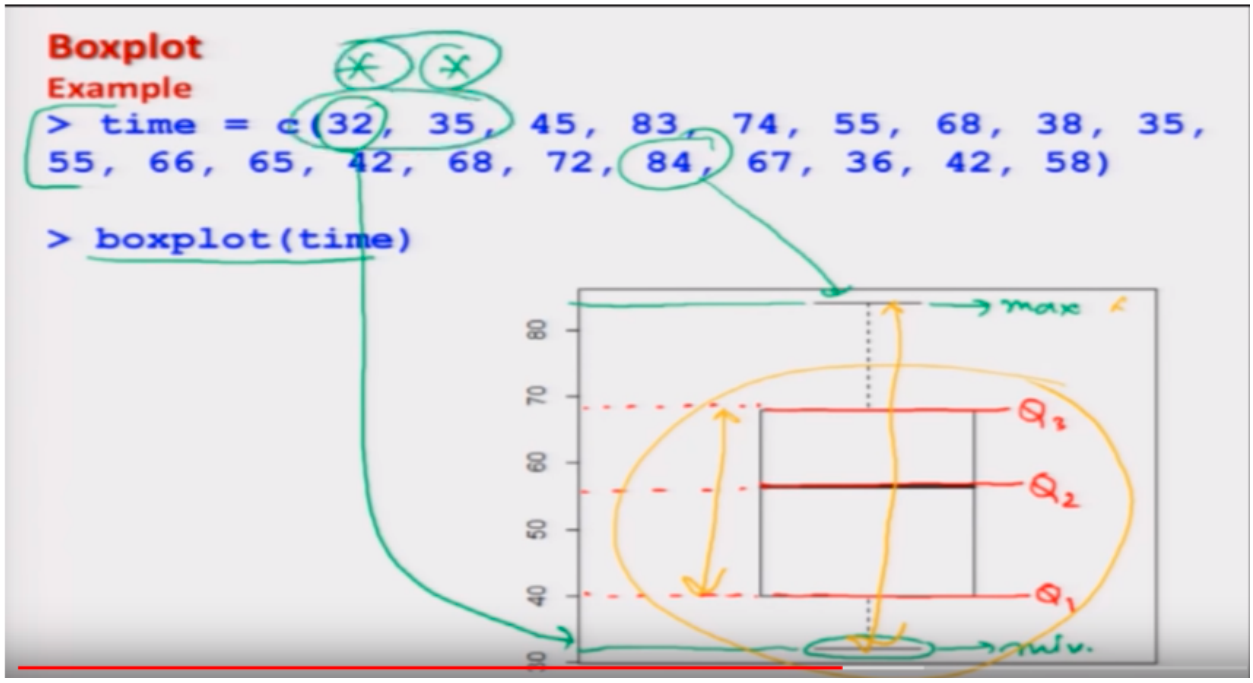
So, let me take here the same example which I considered earlier so this is the same data set on time and now, I have created here the box plot on the set time so you can see here this upper value this is going to give me the maximum value of this time, and which is here 84. You can see here this is somewhere here and similarly the bottom line this is going to give me the minimum value of this data set which is here 32 somewhere here you can see here this is the data set. Now, these three-values bottom value, second value, and third value. These are trying to give me the information on first quartile, middle in the second quartile and upper edge in the third quartile. So, you can see here these values are somewhere here, so you can compare it with the values that you have just obtained in the summary command and by looking at the difference of these two edges here and these two edges here you can have the idea on the variation in the data in terms of range and quartile division or say interquartile range. Now, you can see here, why this plot is called as box plot? you can see here that all the information is contained inside this box so that was possibly the reason that it was called as box plot.

Refer Slide Time: (32:29)



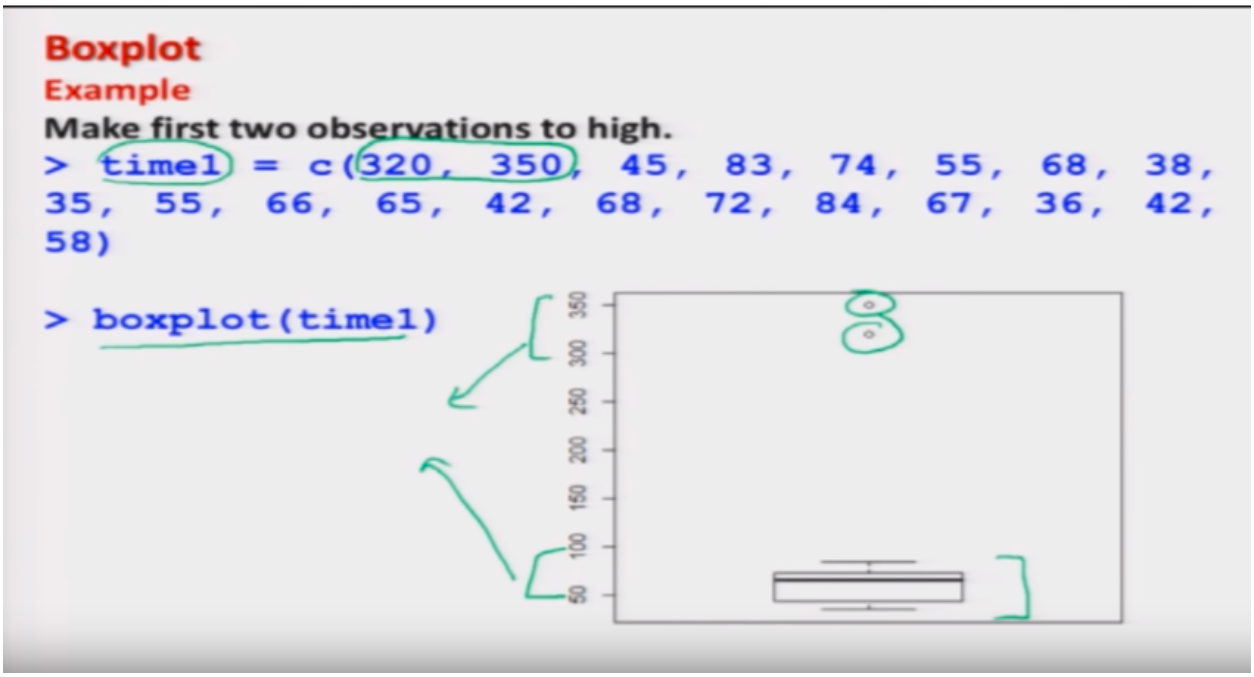
Now, what is the use of this box plot, how it is going to help us?

Refer Slide Time: (32:38)



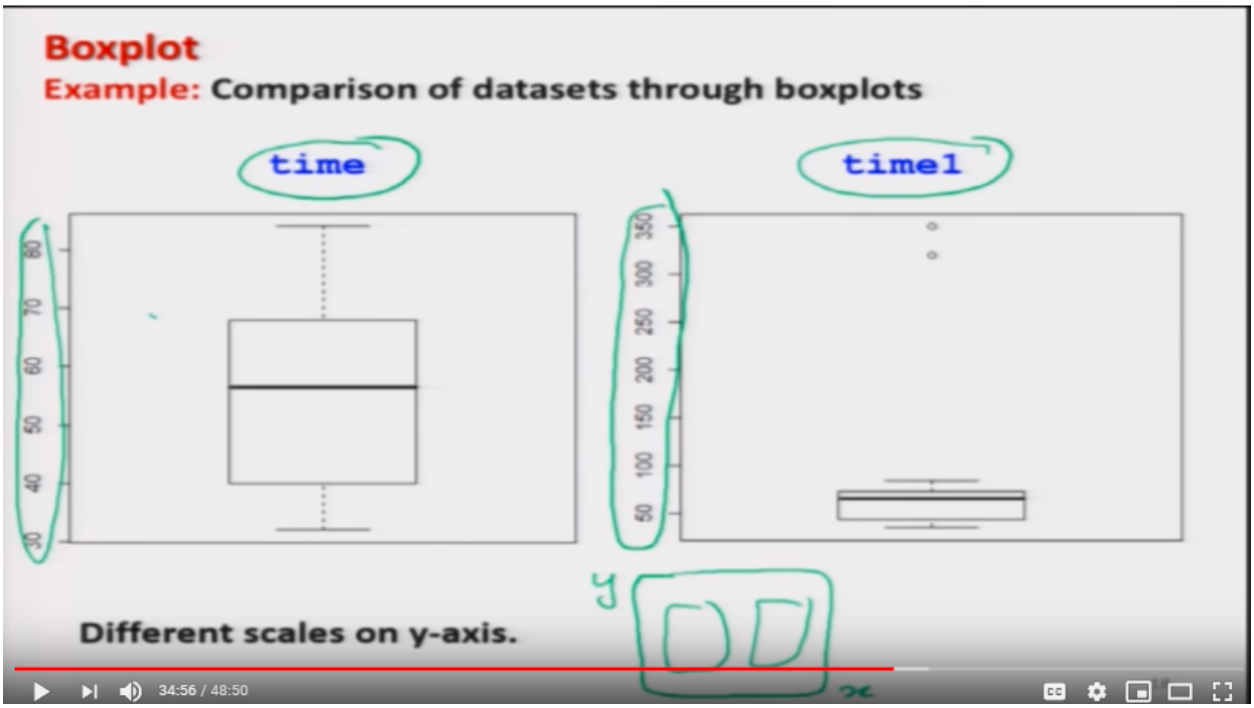
So, now what I have done here first you please try to notice, what are the first two values which I am indicating here, these are 32 and 35.

Refer Slide Time: (32:48)



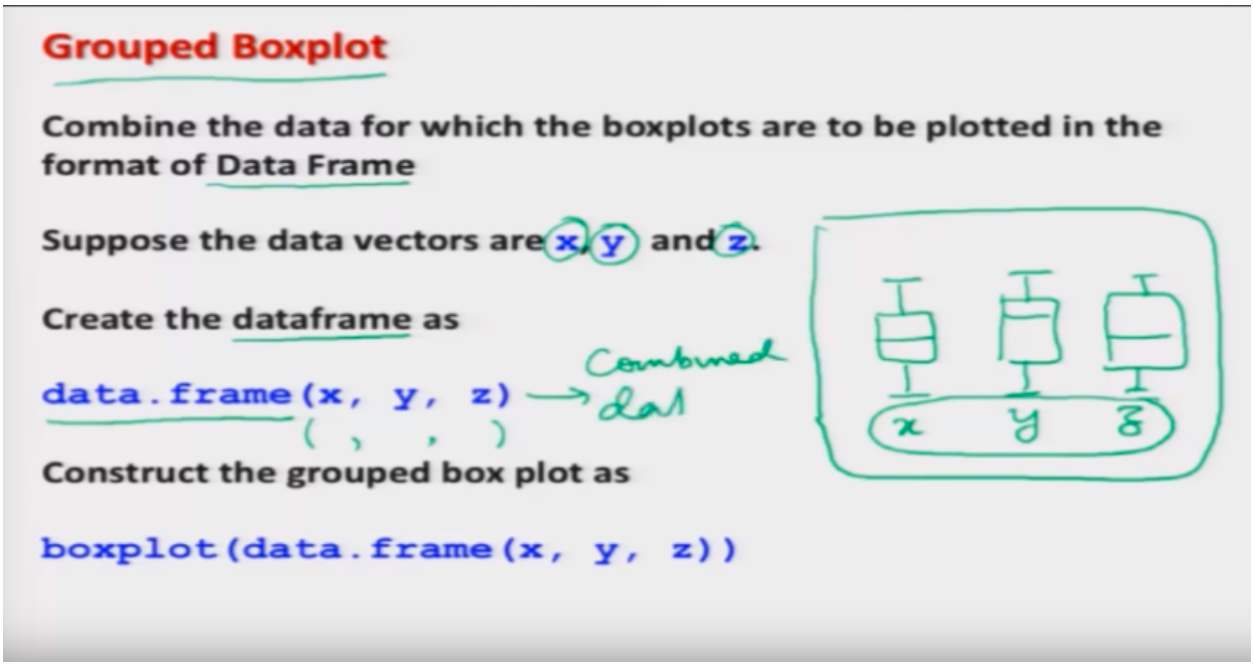
Now, what I do in the same data set, I try to make it 320 and 350 very high value and I give this data a new name time1, and then I try to create the box plot the time1. You can see here that this box plot is very, very different than what you have obtained in the earlier case. Not only this, it is also showing you that there are two observations which are possibly the extreme observation. So, by looking at this graph this is giving you information well when you are trying to analyze that data please try to take a look at the values, which are somewhere between 300 and 350. These values are unusual because they are very, very far away from the remaining data, all our data is lying between 50 and 100 here, where these two values are lying between 300 and 350. But my objective was that I want to compare it. So, I try to artificially make these two plots side-by-side.

Refer Slide Time: (34:01)



So, I have simply copied and pasted these two graphs manually and you can see here that the first graph is for the data set time, and the second graph is for the data set time1. but now you can see still we are not very comfortable. Why? because the range on the y-axis in the both data set, they are different. So, they are not really comparable well they are comparable, but you need to put a harder work in the comparison. So, we would like to make a graph in such a way where we have only this type of boundary on say x and y axis and this plot should be inside the same boundary, so that they are comparable. So, in order to do so, what we have to do? This I am going to now discuss to demonstrate.

Refer Slide Time: (34:59)



We have a graphic, which is called as grouped box plot. This graphic will combine different types of datasets and it will create a box plots inside the format of a boxplot using the data inside the format what is called as data frame. What is this data frame and why this is needed? Suppose we have here three data vectors, which I am indicating by x, y and z. So, what we want here is the following that we want here a graphic, which is enclosed inside this rectangle where, there are three box plots like this and they are indicating the box plots for the three data vectors x, y, z. So, in order to create this combined box plot first we need to combine that data, and in order to combine the data, we have a concept of data frame, and we do as follows that we use the command here data dot frame and inside the arguments we try to give the names of the data vector. Which are separated by commas and this will give us a data set in the framework of the concept of data frame and this data will be a combined data set.

Now, at this stage question crops of that, what is a data frame? well data frame is a method to combine different types of data sets in R. Well it is not really possible for me to explain this concept in detail here but these concepts have been explained in the lectures on the course introduction to R software, if you wish you can have a look on those lectures or you can look into the help menus of the R software to have an idea about the concept of data frame, Right. But in this lecture, I will be using only this command just to combine the data sets, so I am not going into that detail. I have given you the command that how to use it. If you want some advanced feature advanced knowledge above this concept, I would request you to look into the books and help menus, Okay.

Refer Slide Time: (37:47)

Grouped Boxplot

Combine the data for which the boxplots are to be plotted in the format of Data Frame

Suppose the data vectors are x, y and z.

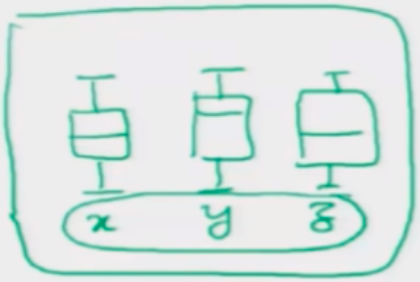
Create the dataframe as

```
data.frame(x, y, z)
```

Combined data set

Construct the grouped box plot as

```
boxplot(data.frame(x, y, z))
```



So, now my objective is very simple, I would try to create a box plot for this data set which has been combined using the concept of data frame.

Refer Slide Time: (37:56)

Grouped Boxplot

Example

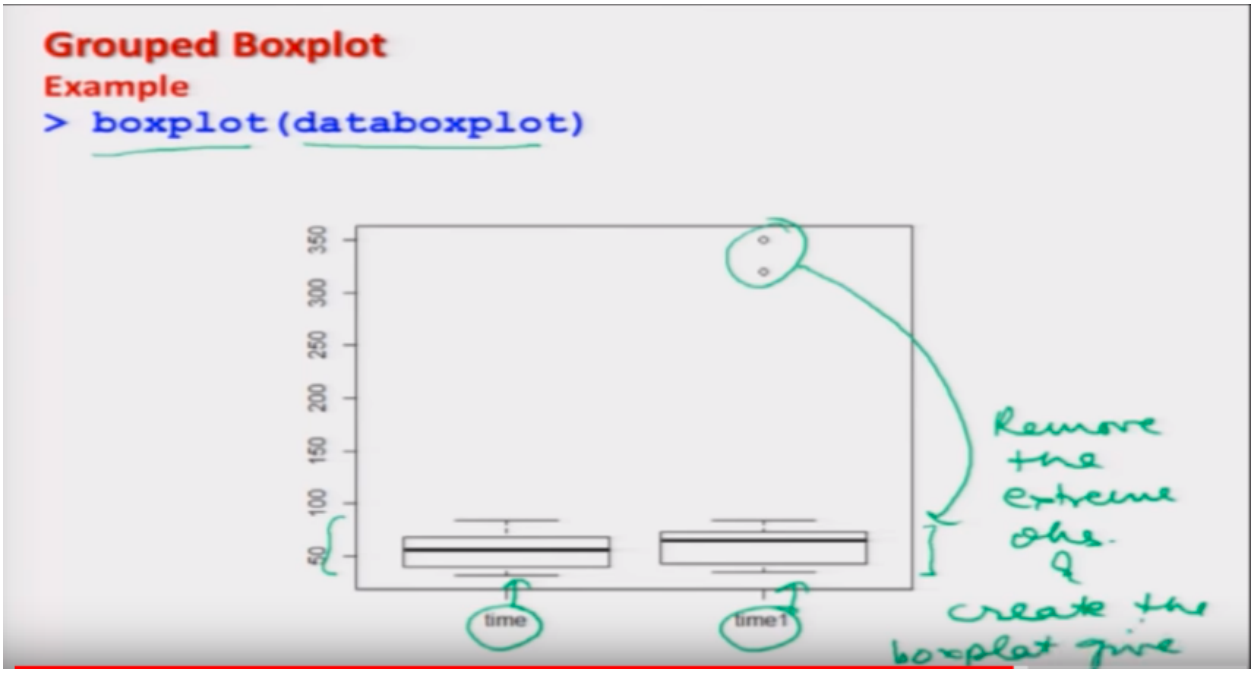
```
> time = c(32, 35, 45, 83, 74, 55, 68, 38, 35,
55, 66, 65, 42, 68, 72, 84, 67, 36, 42, 58)

> time1 = c(320, 350, 45, 83, 74, 55, 68, 38,
35, 55, 66, 65, 42, 68, 72, 84, 67, 36, 42, 58)
```

Create the data frame as follows:

```
> databoxplot = data.frame(time, time1)
```

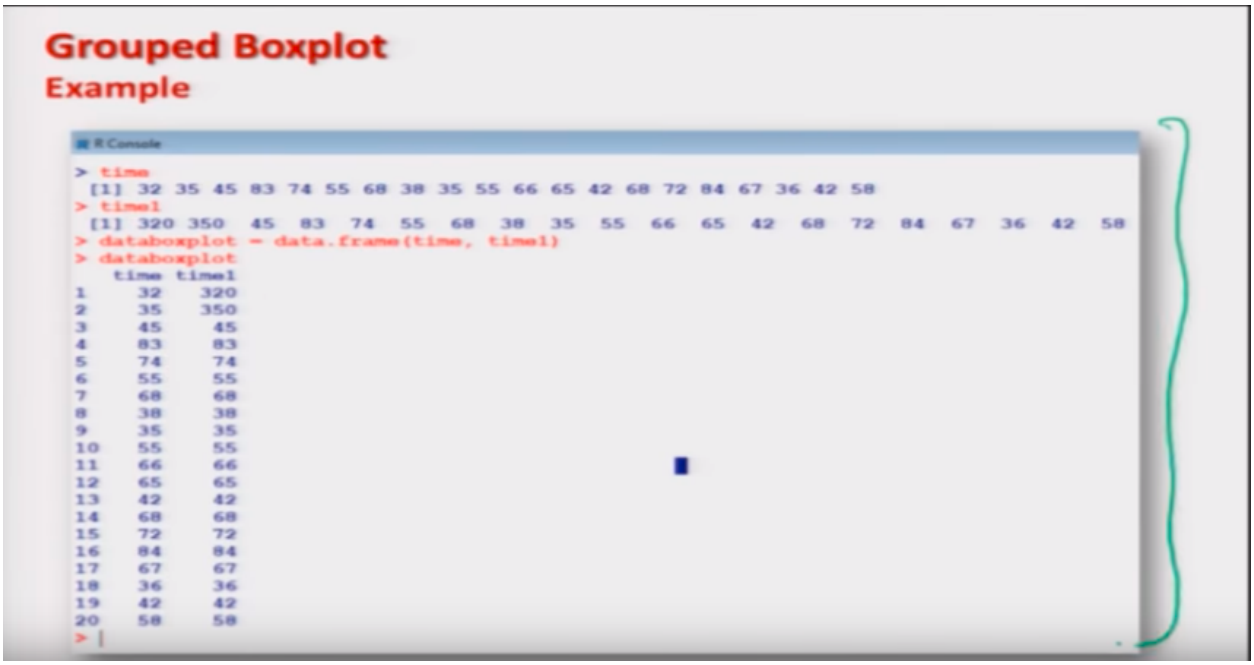
So, now what I do here is the following, I try to take the same data set time and here time ,1 and I try to combine it and create a data frame using the command data dot frame, and inside the argument time separated by comma and then time 1, and I store this data as a data box plot.

Refer Slide Time: (37:56)



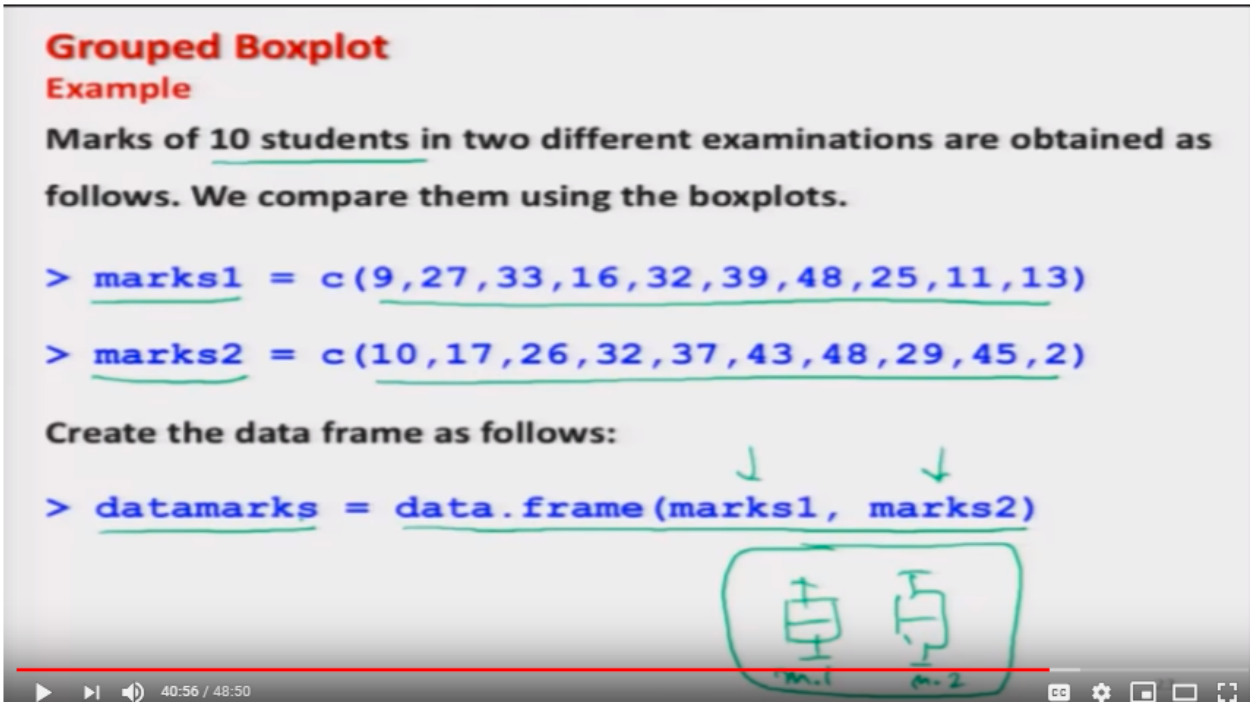
And, after this I use the command box plot and, and give the data inside the argument as data box plot. We, which has been obtained through data frame. Now you can see here that both the graphics have been combined together, Right. This is the box plot for the time, and this is the box plot for the time ,1 and here you can see that that is indicating the presence of two extreme observations, Right. So, in this case you can see that we have combined the two box plots, but they are not really informative because, the ranges or both the box plots are very different, Right. This is here and this is indicating that there are two extreme observations, so this is giving a different box plot that if, what we wanted for the ideal thing is that, after looking at this picture first try to remove the extreme observations, extreme observation and create the box plot again, and you will see that this will give you more information, and I would like to illustrate the same thing with a different example.

Refer Slide Time: (39:39)



So, this is here the screenshot of the creation of Dakota frame of time and time1 data.

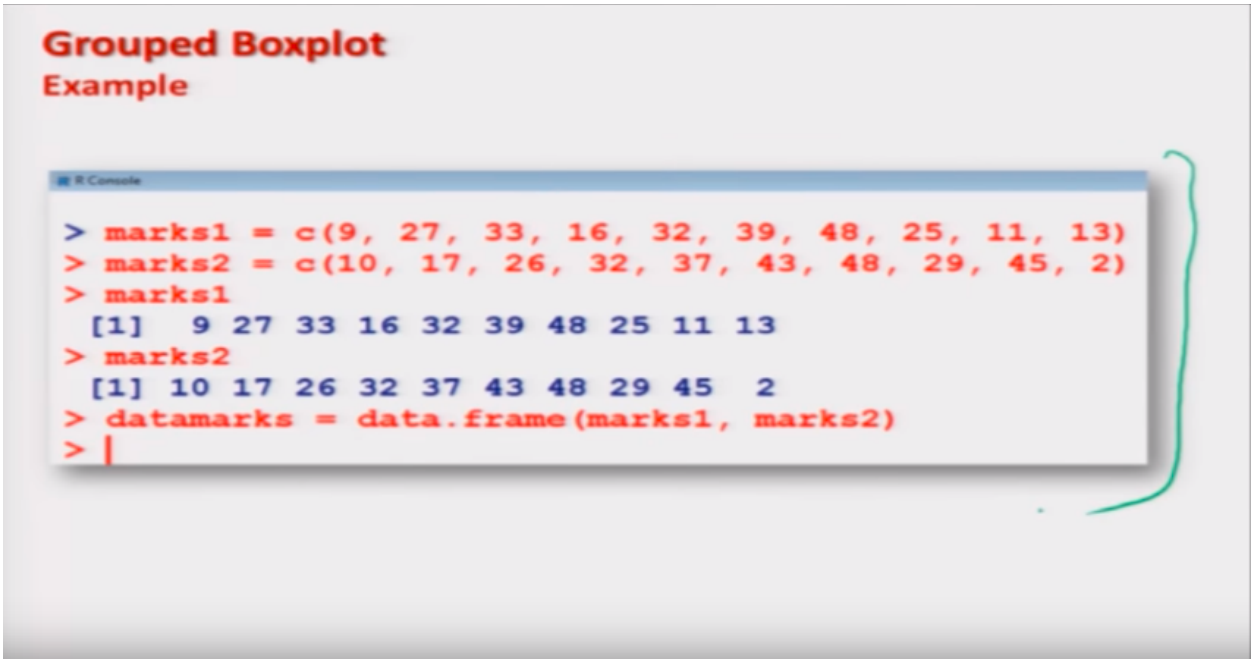
Refer Slide Time: (39:47)



Now, I will take another example, to show you the utility of box plots. Suppose the max of ten students in two different examinations are obtained and we would like to compare the marks using the concept of box plots. So, the marks in the first examination they are stored here inside the data vector named as marks1.

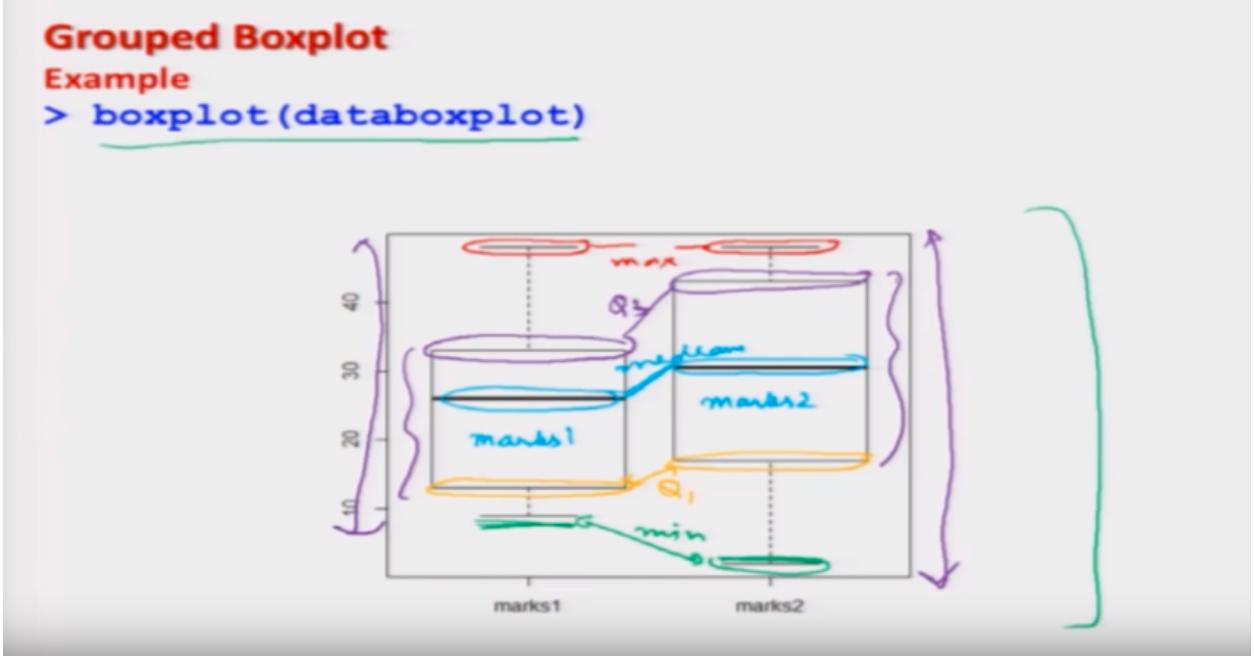
and the marks of those ten students in second examination they are stored here, and they are stored in a the data vector marks2. So, What I want here is the following? I want here a graphic like this one, where there are two box plots one indicating the marks1 and another box plot indicating the marks2. So, in order to do, so first I try to create here a data frame using the command data dot frame and inside the arguments, I try to give the data vector, which I would like to combine. And, this data suppose I am trying to store as say data marks, Right, Okay.

Refer Slide Time: (40:59)



Now this is the screenshot of this operation and now I try to create the box plot of this data box plot data which has been obtained in the framework of a data frame like this box plot of data box plot.

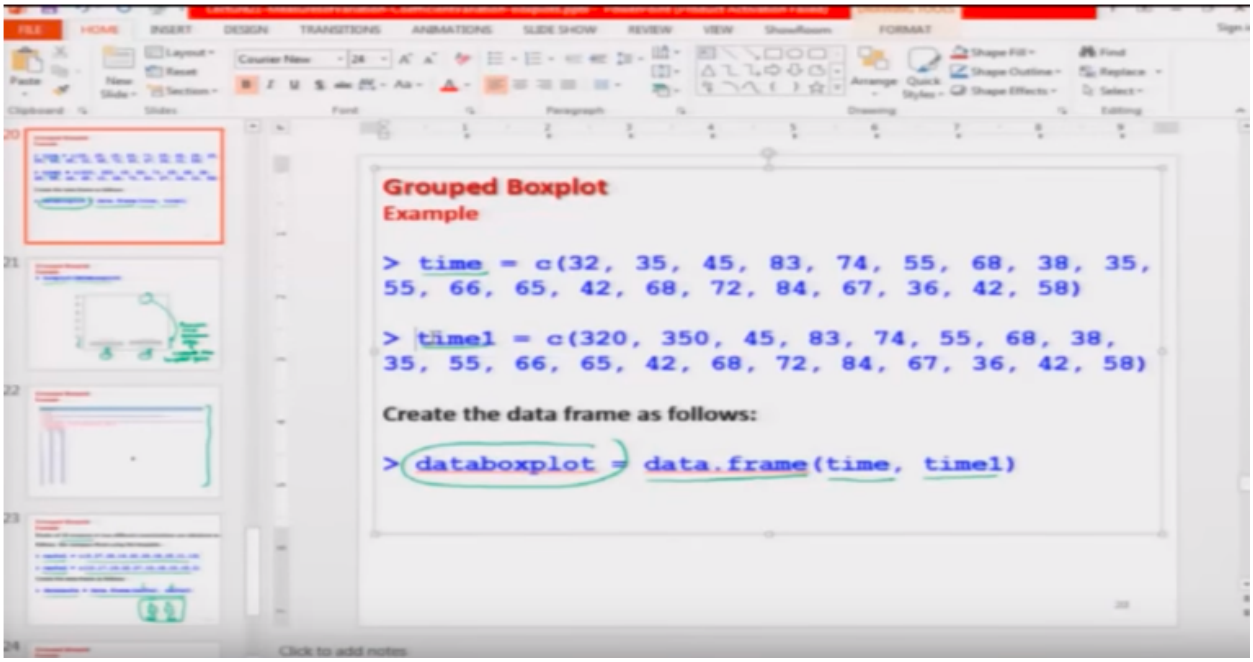
Refer Slide Time: (41:03)



Now, you can see here now here you are very nice and clear picture. By looking at these two values these green color lines, you can easily compare that which data set or which of the group of students has got the lower marks or the minimum marks. Similarly, if you try to look at the red pink which I am highlighting here by comparing this to you can have an idea about the maximum marks obtained by the students. Similarly, if you try to look in the middle value these are trying to give you the idea of the medians. So, by looking at these values you can see here, that the median in the marks two case is higher than the median marks in the case of marks1.

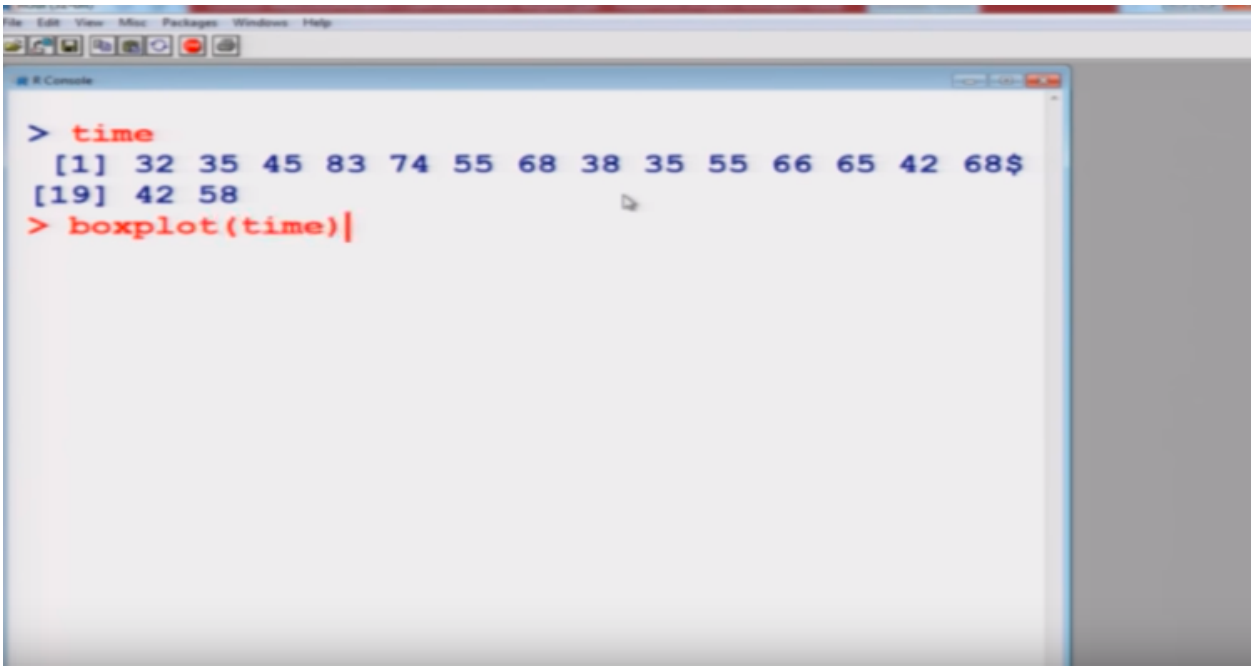
Similarly, in case if you try to compare this orange color part this will give you the idea of the first quartile and similarly if you try to compare the third quartiles highlighted by this violet lines you can compare it and you can see here that the third quartile Q_3. That Q_3 in the case of marks2 has higher value then in the case of marks1 and you can see here the range in the marks1 and the range in the marks2. So, you can see here that very clearly that the range of the marks2 is higher than the range of the data in marks1, and similarly you can also compare the quartile deviation and interquartile ranges, so you can see here this is how we can obtain it, now I would try to show you the construction of the box plots and the group box plots on the R software, Right.

Refer Slide Time: (43:14)



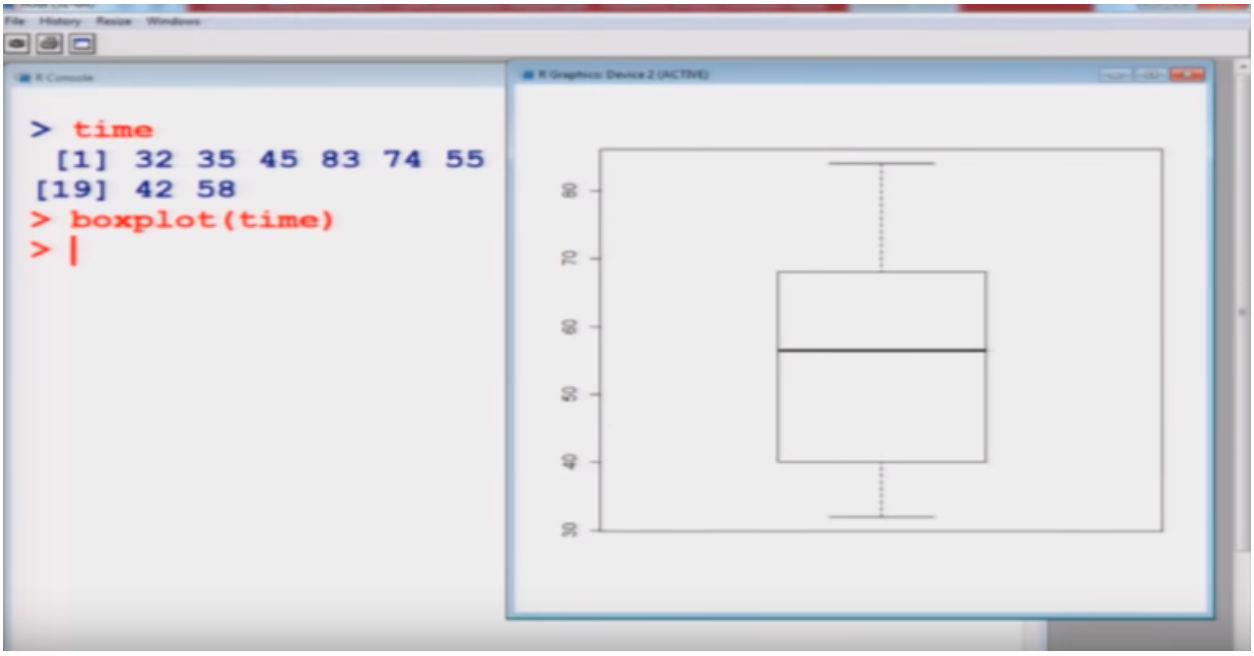
So, you can see here we already have the data on time.

Refer Slide Time: (43:19)



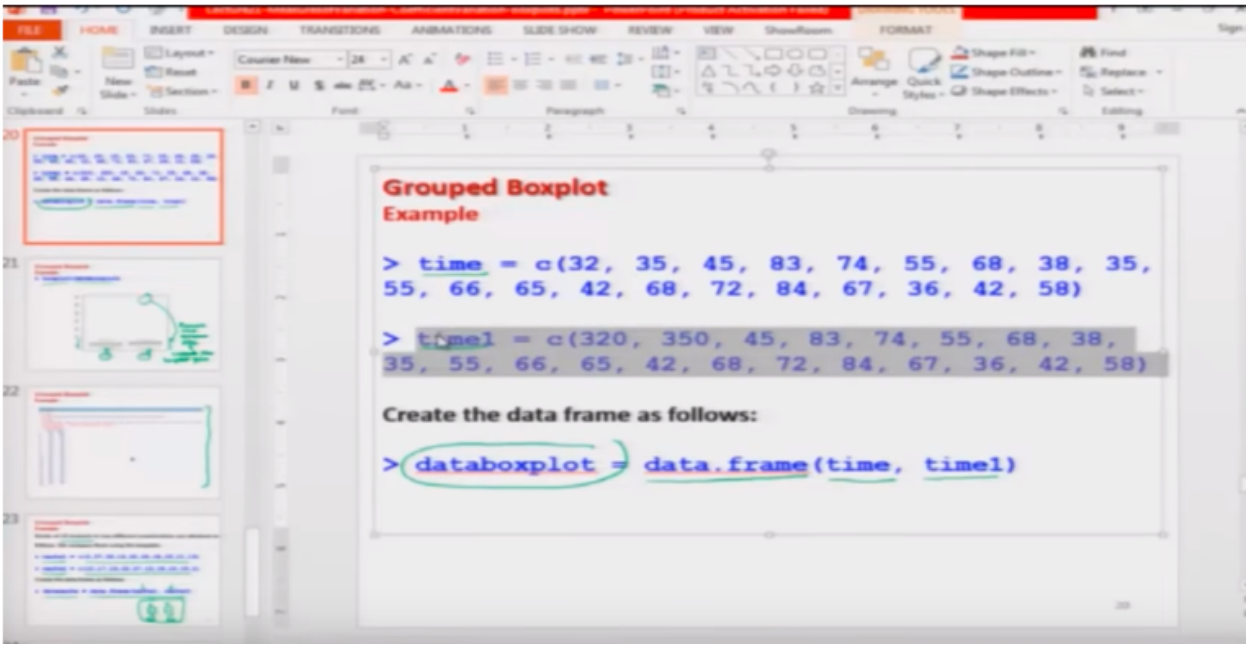
So, if I try to create here the box plot it will look like box plot of say head time.

Refer Slide Time: (43:29)



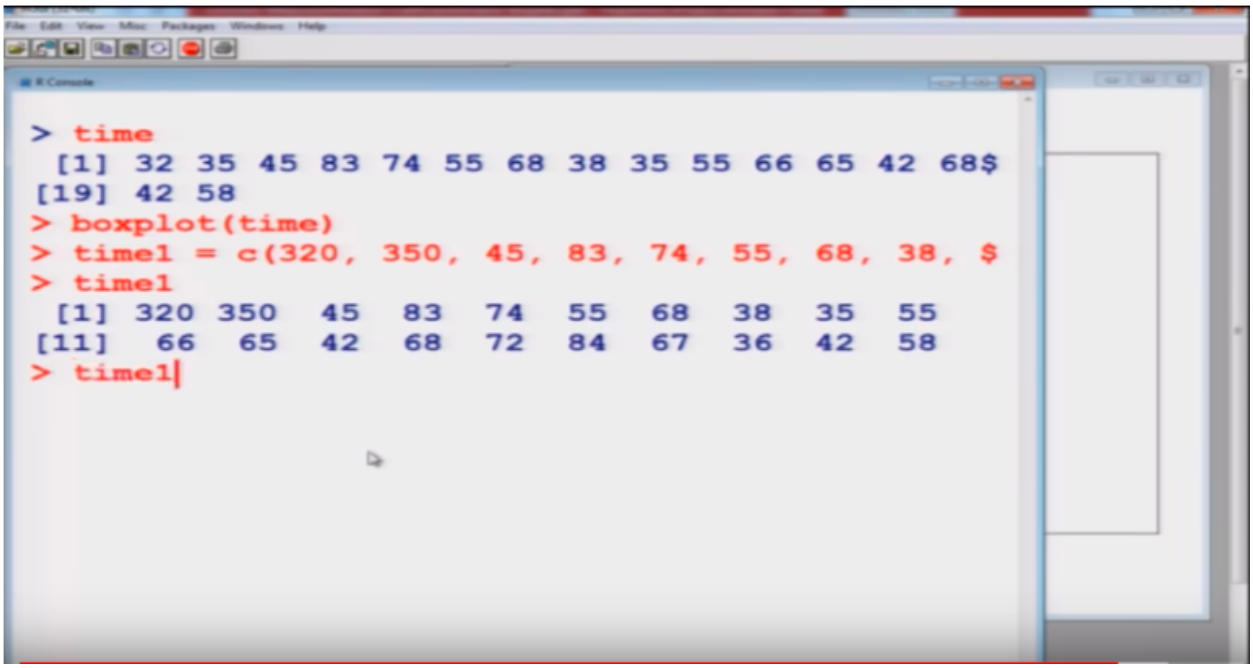
And you can see here this comes out to be like this.

Refer Slide Time: (43:34)



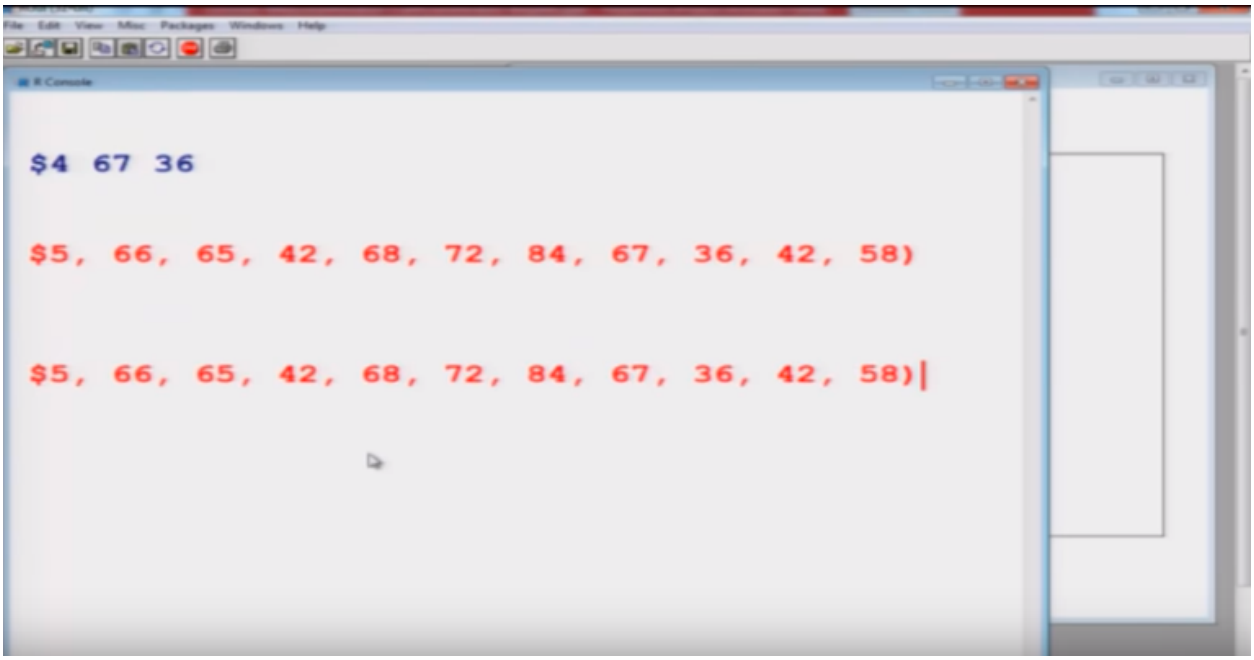
Now, similarly if you try to take another data set time1.

Refer Slide Time: (43:48)



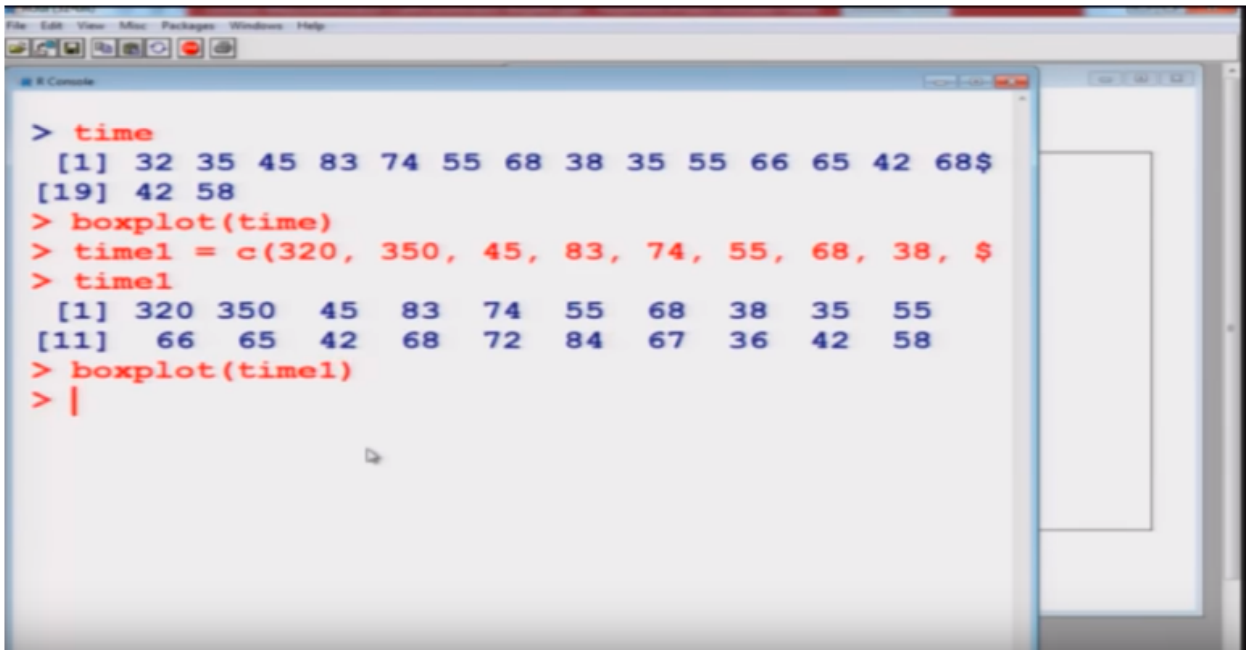
Where we have increased the values of two observation first two observation.

Refer Slide Time: (43:50)



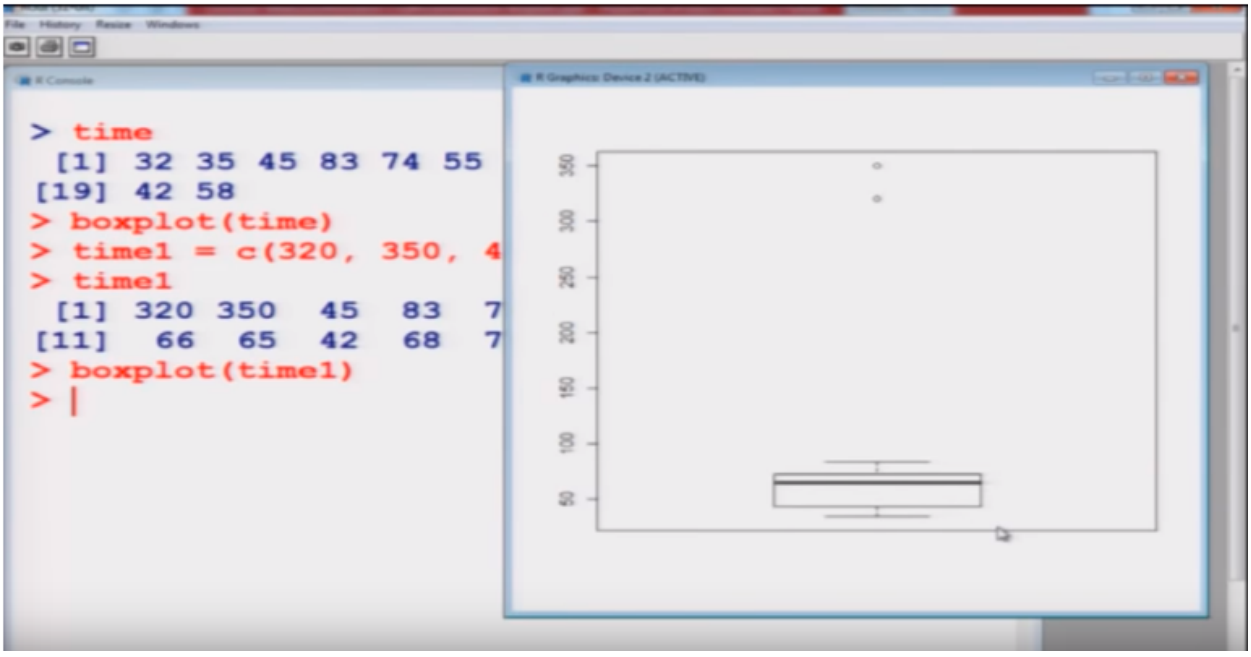
Then in this case this data set is here time1 and you can see here that there are two extreme values.

Refer Slide Time: (43:52)



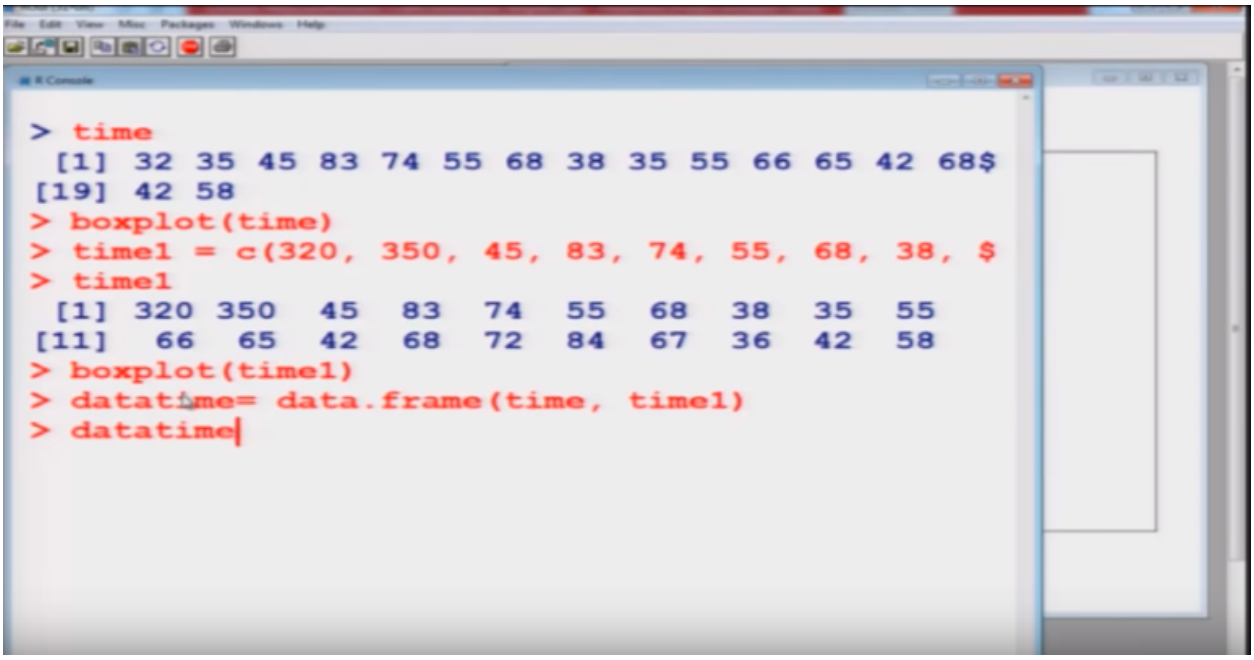
And, if you try to create here the box plot of the same.

Refer Slide Time: (43:54)



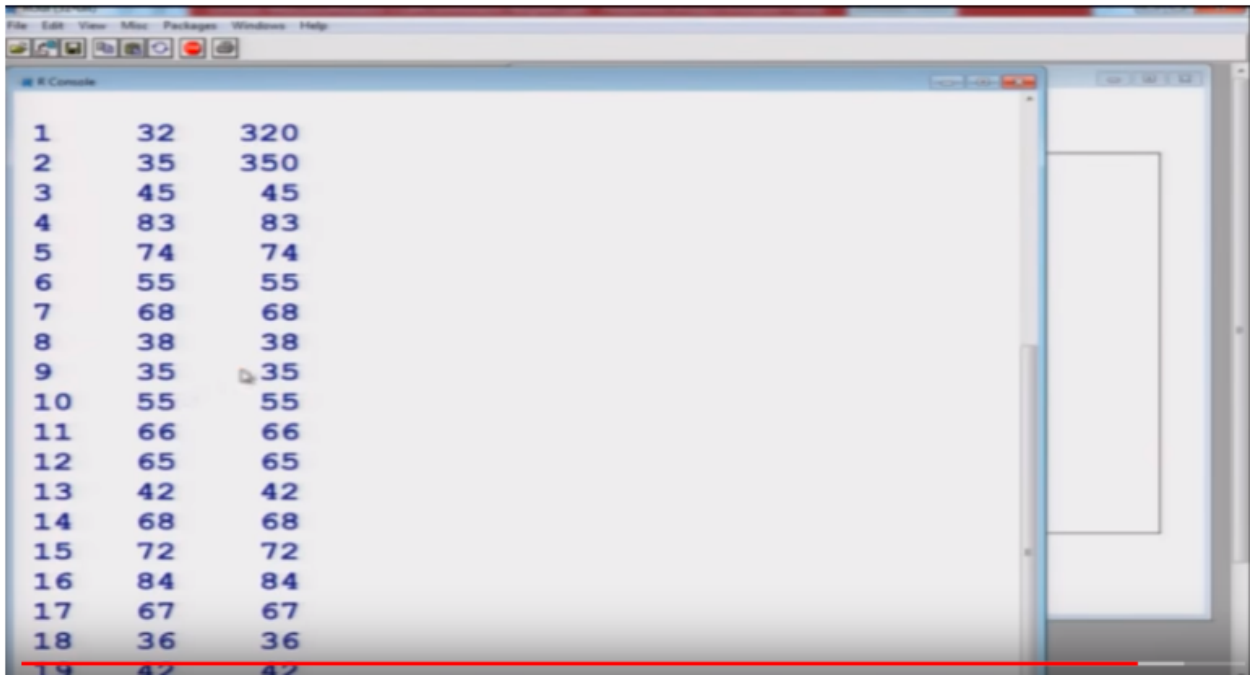
This comes out to be here like this which we had reproduced in the slides, Right.

Refer Slide Time: (44:03)



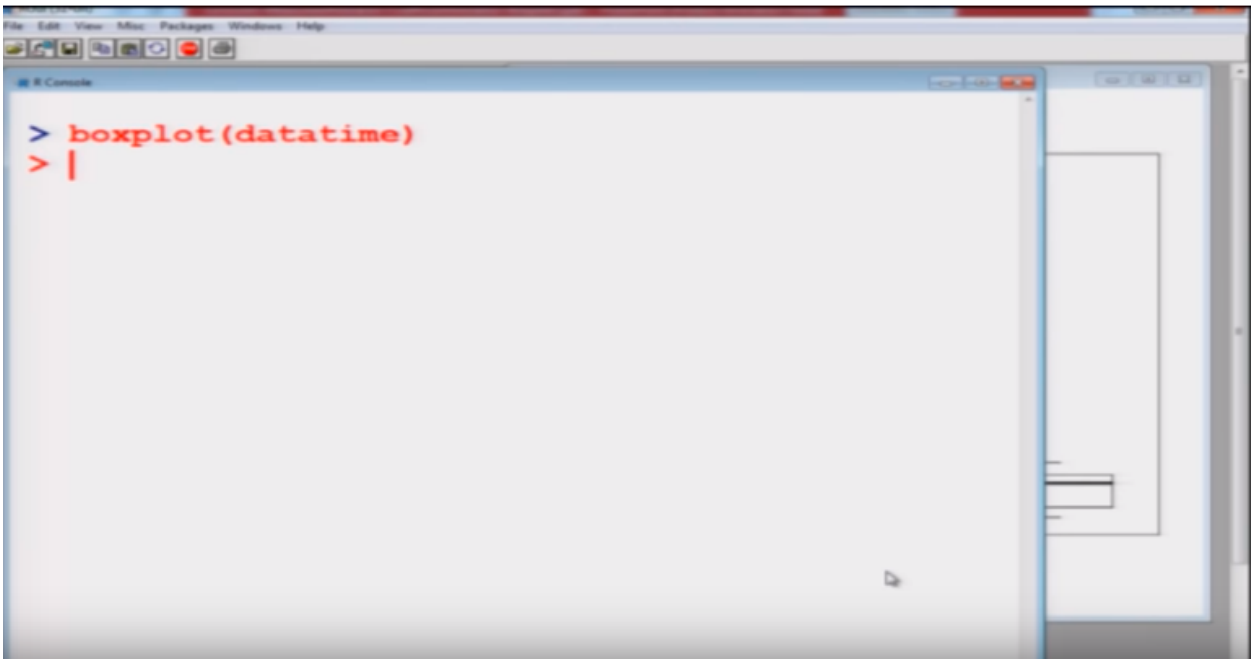
And, in case if you try to combine here the data here see here, data the time is equal to Data dot frame and inside the bracket, time separated by comma time1 argument closed.

Refer Slide Time: (44:32)



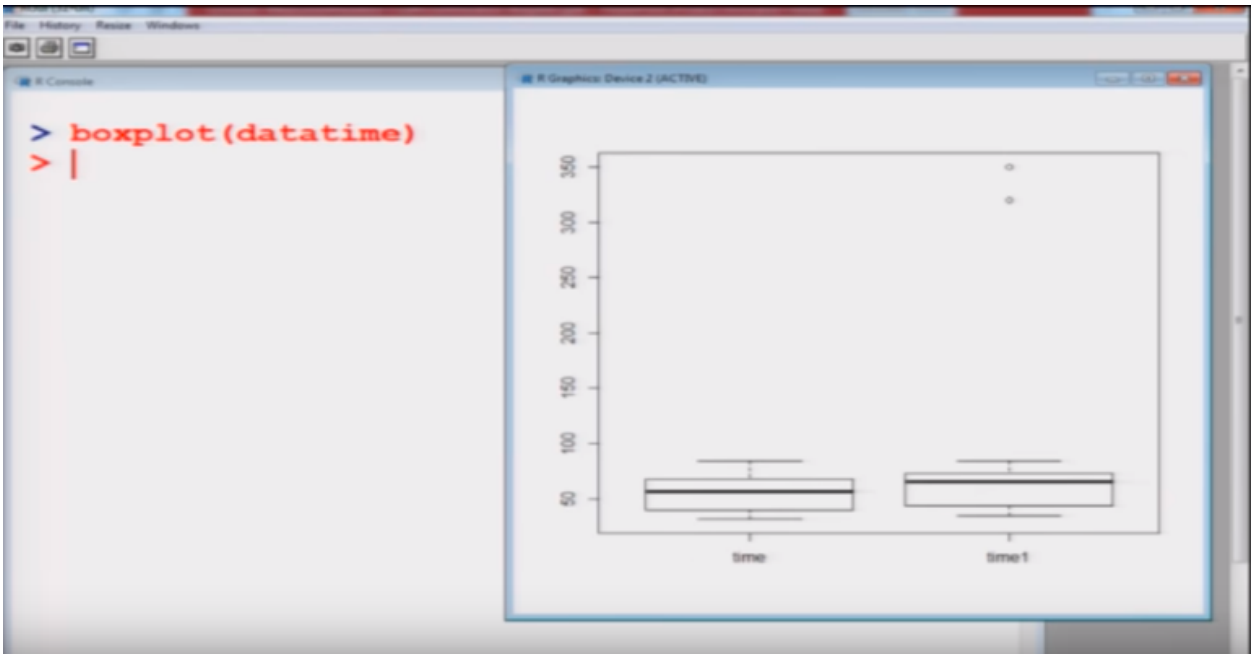
You will get here at the, the combined data on time and time1 in the data frame mode and you can see here this is the data which I have obtained here.

Refer Slide Time: (44:37)



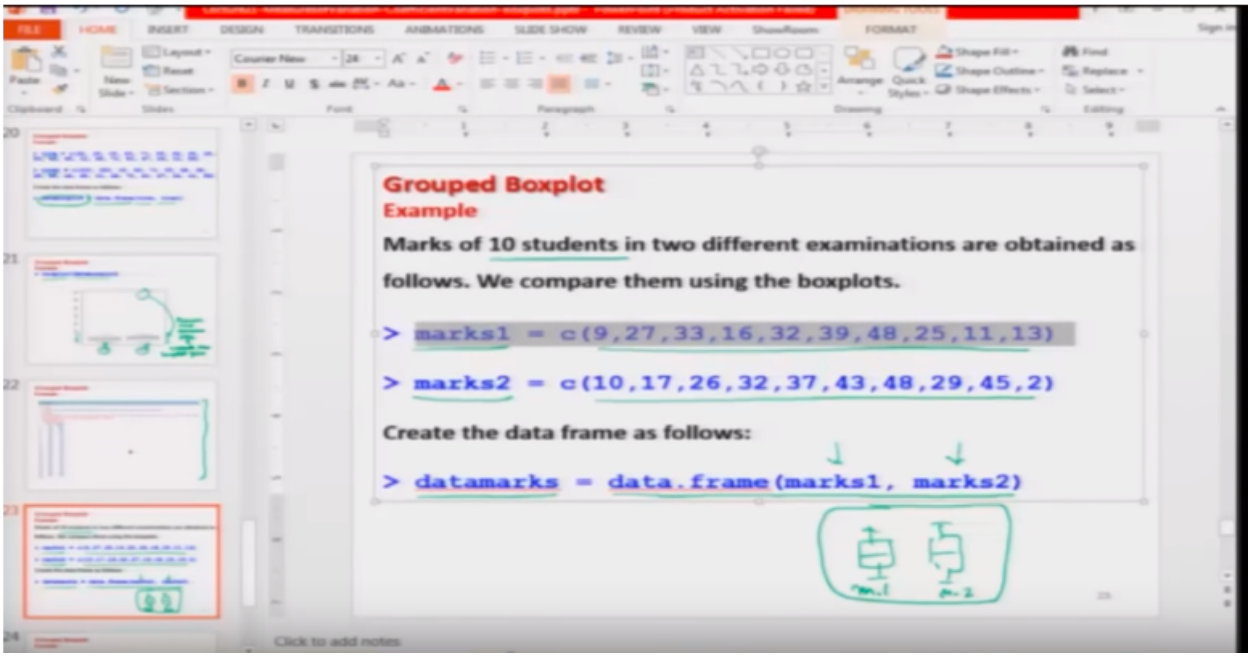
Now, I will simply try to make sure the box plot of the same data set.

Refer Slide Time: (44:43)



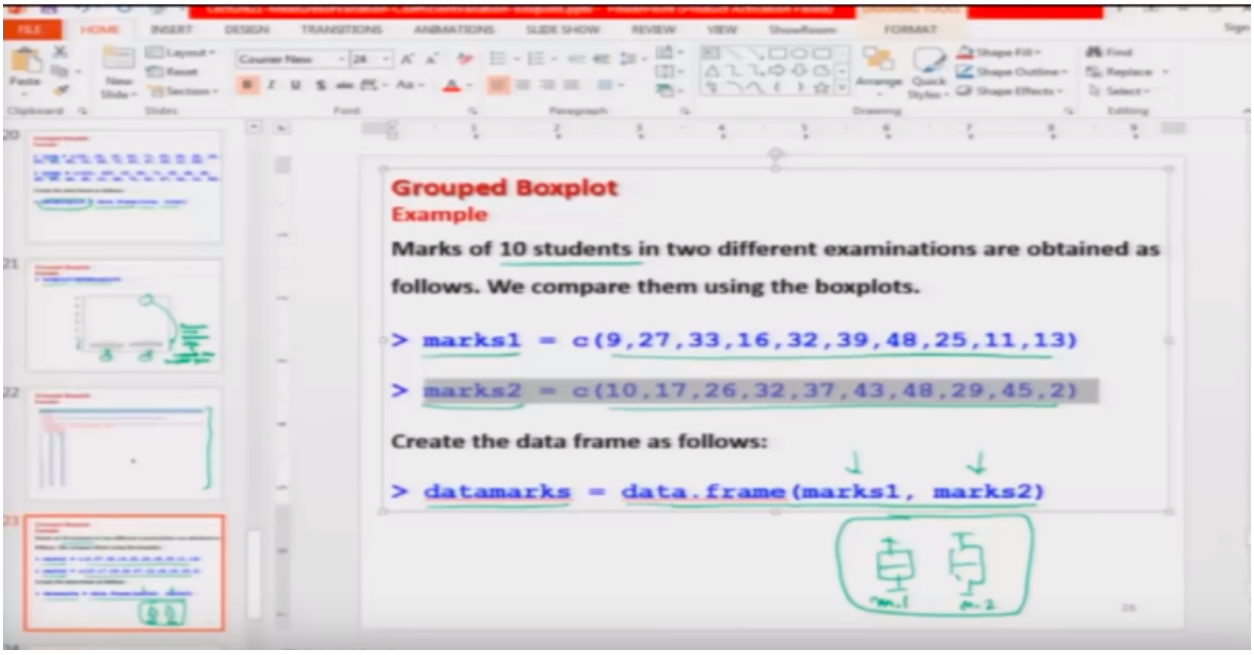
So, you can see here if I try to create the box plot this comes out to be here like this which we had to reproduce whenever slides.

Refer Slide Time: (44:53)



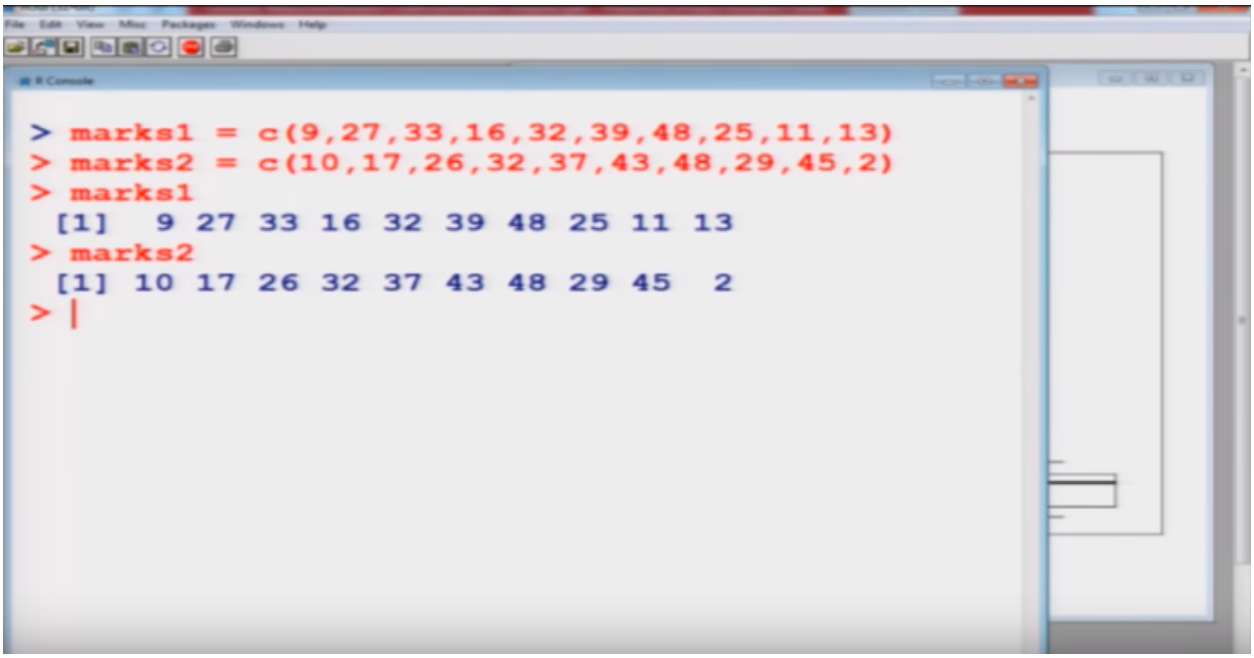
And, similarly if I try to take here the data on marks.

Refer Slide Time: (45:04)



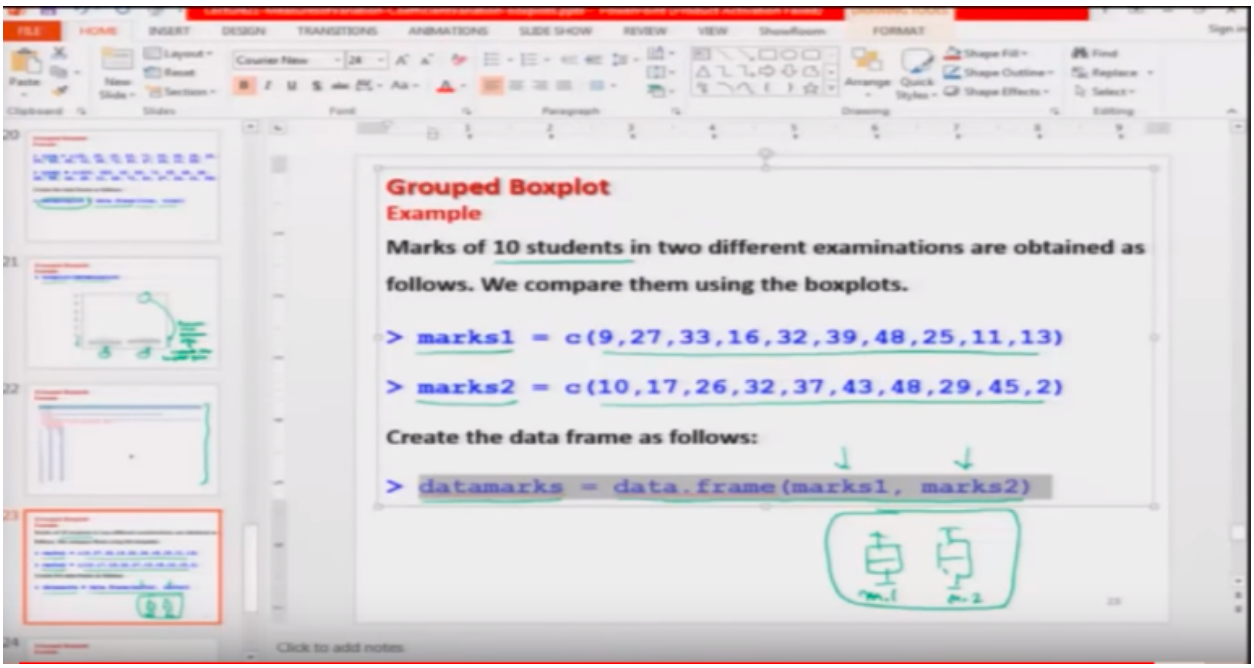
So, data on marks here is like this. and data on here mark here is like this.

Refer Slide Time: (45:00)



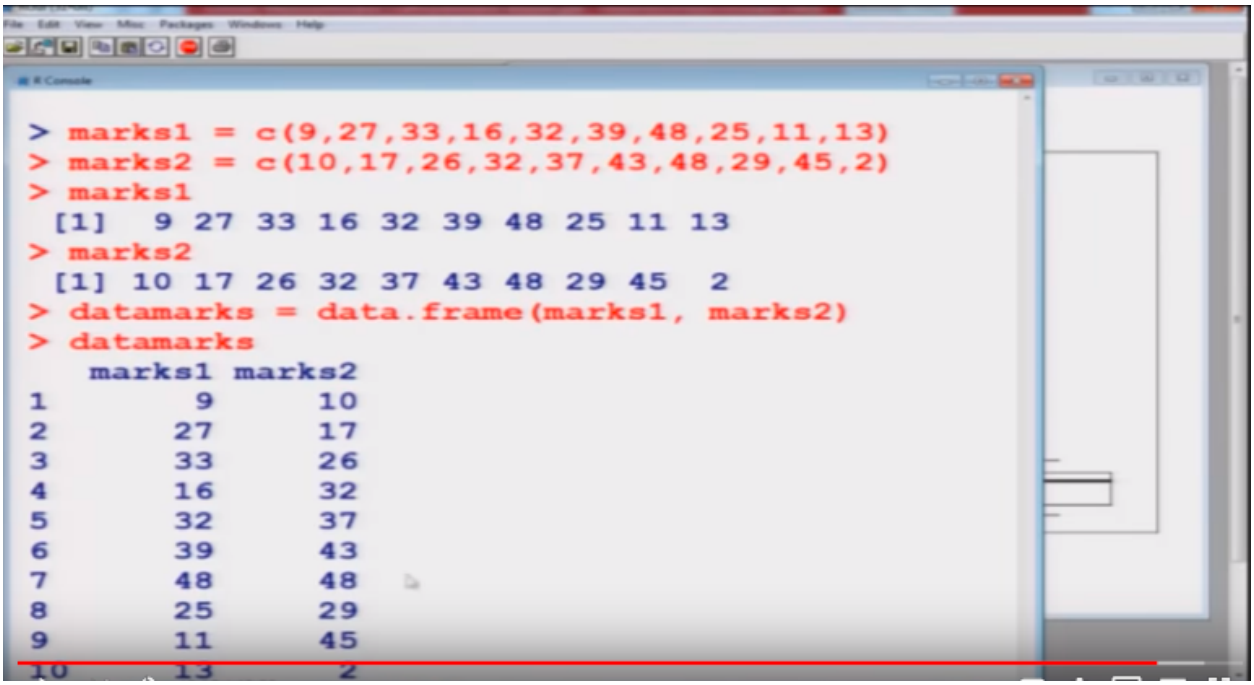
You can see here, this is the data on marks1 and marks2, Right.

Refer Slide Time: (45:14)



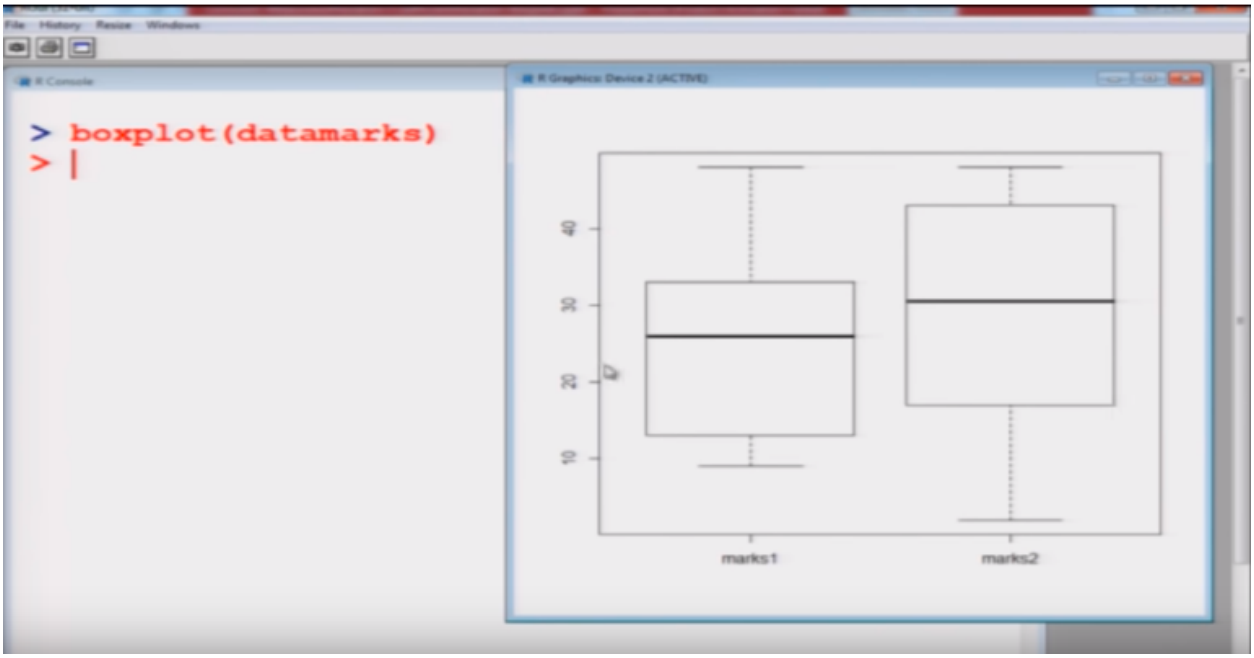
And, I would like to combine this data into a data frame.

Refer Slide Time: (45:19)



So, I try to use as the data frame come on and you can see here this data has been combined see here click at a frame like this, Right.

Refer Slide Time: (45:29)



So, I would like to now create a box plot of the same thing two box plot of the data marks you can see here no this looks like this so you can compare it and can have some fruitful conclusions, Okay.

So, now I would stop here in this lecture and I would also complete the discussion on the topics of different measures of variation. So, we have discussed different types of measures of variation and every measure will give you a different information, and a different numerical value your experience in dealing with the data sets and using these measures will give you more insight into the aspect that how to interpret, how to say whether the variability is low or variability is higher, this is always a relative term. But, remember one thing from the statistics point of view in case if the data has very high variability then most of the usual statistical tool will not really work well. They will give you some information, but that information may be misleading. So, it is very important to use the appropriate tool to bring the information out from the data regarding its inherent variability, different samples taken by different people from the same population they may have different variation, so if you try to use different types of tools ideally all the tools should give you the same information, but they will have different numerical values. So, I would request you that please try to look into the books, try to understand the concept of a variation and the data, try to look the different drawbacks different advantages of all these tools all the tools cannot be applied in all the situations, and more importantly how to compute them on the R software, this is what you have to learn. So, I would request you that you take some datasets and try to employ all the tools whatever you have done up to now. Different measures of central tendency, different measures of variation and try to see how these values are trying to provide you different pieces of information. So, you practice, and we will see you in the next lecture with a new topic. Till then, good bye!