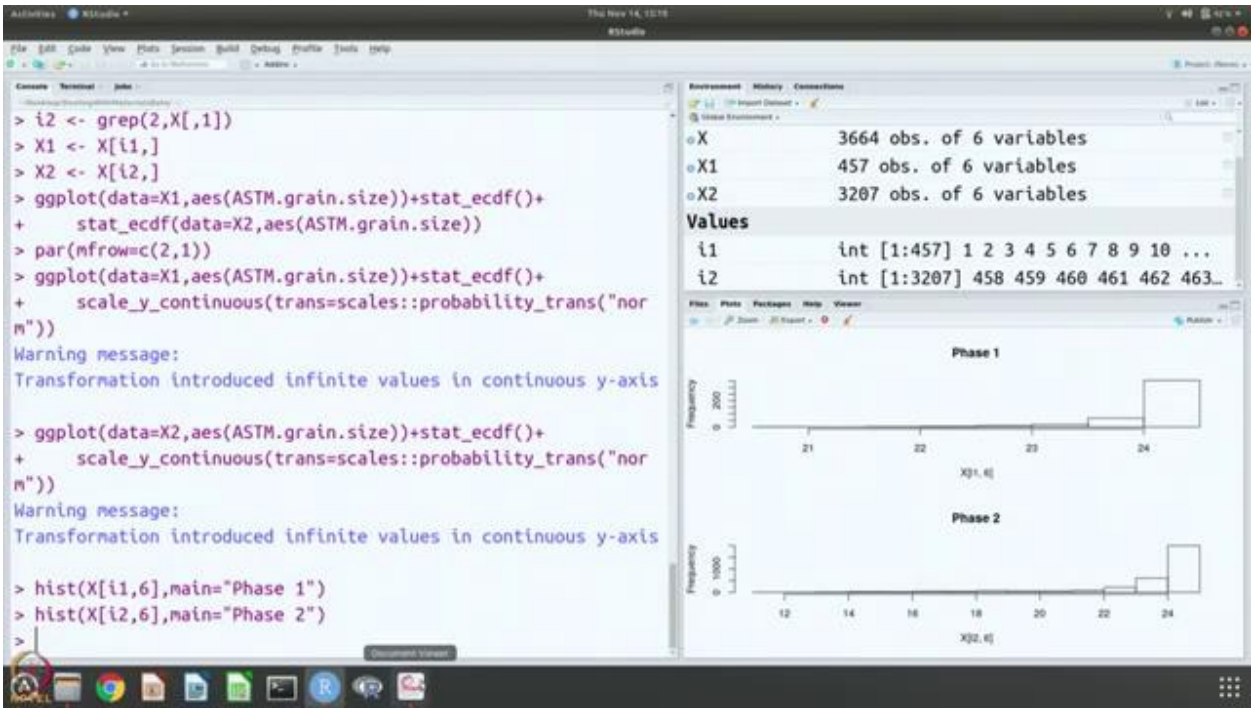
Dealing with Material Data: Collection, Analysis and Interpretation
Professor. M. P Gururajan
Professor Hina A Gokhale
Department of Metallurgical Engineering and Material Science
Indian Institute of Technology Bombay
Lecture 25
Grain size in a two phase steel: Descriptive statistics

We are looking at the data set two which consists of grain size data from two different phases and we are doing the rank based properties and their representation.

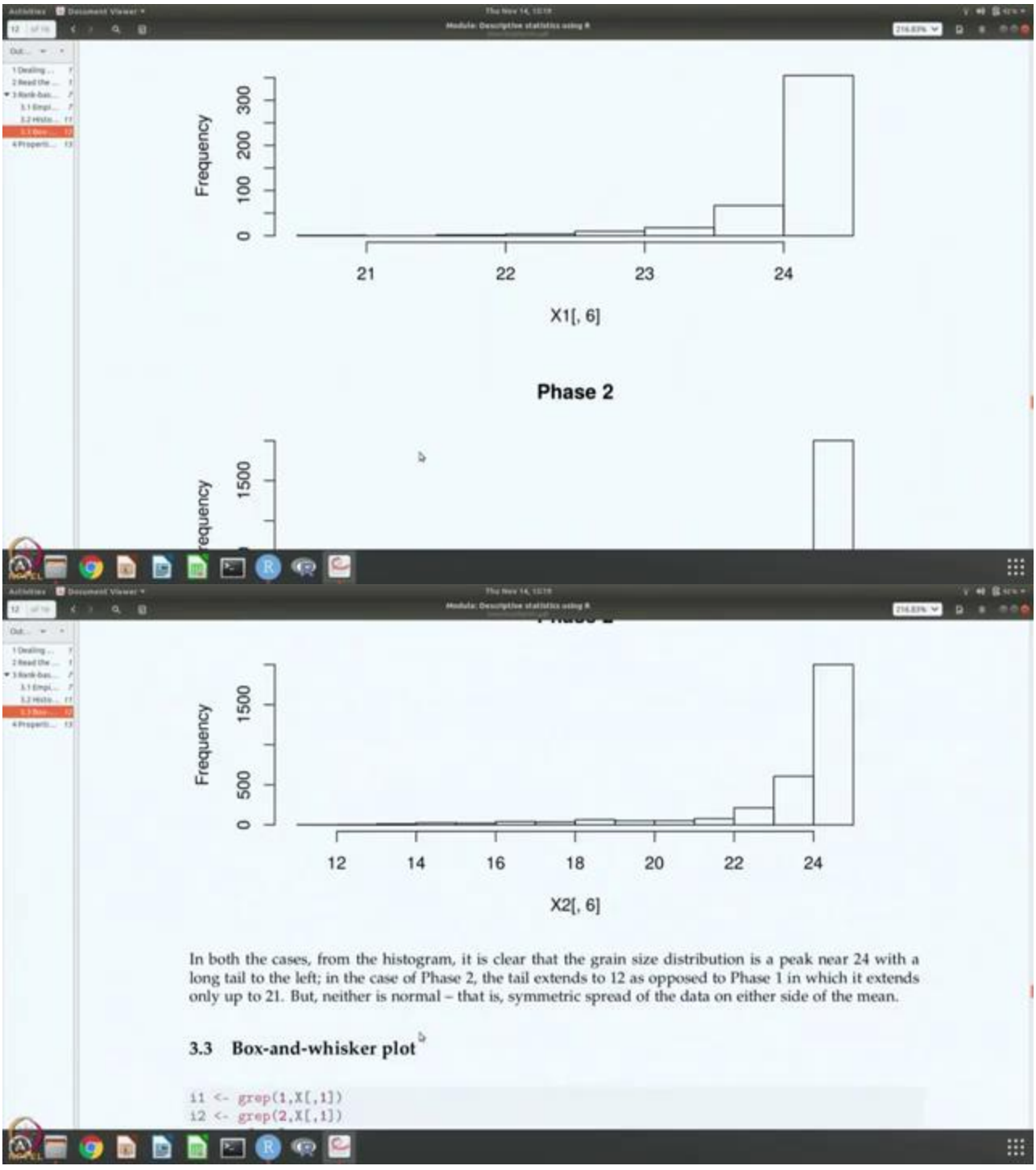
(Refer Slide Time: 00:29)



So, let us continue and what next what we want to do is to do a histogram plot of this data for both the phases. So, this is phase 1. So, it should say phase 1 and we want to do a histogram plot of phase two data from grains of phase two.

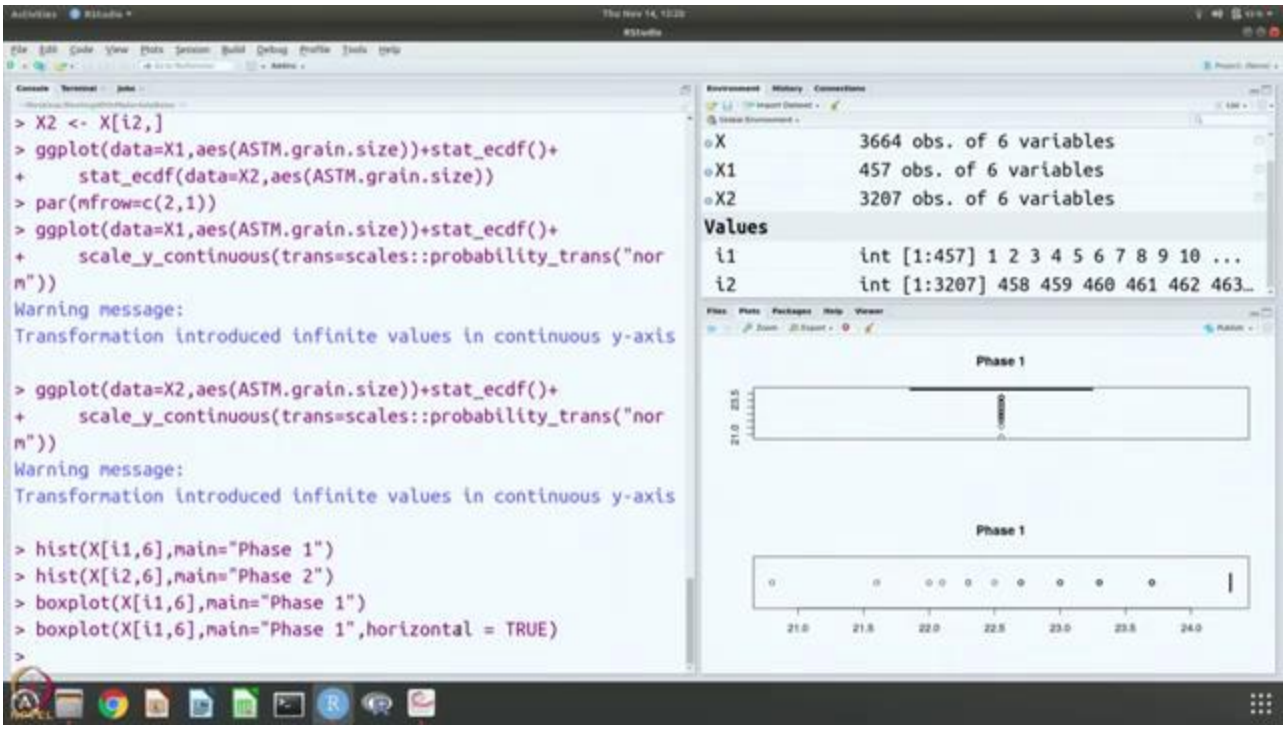
So, let us do that and then call it as. So, this is the same information we have seen it several times now. So, both peak somewhere around 24 and both have a tail only to the left and this is a much longer tail and a relatively probably fatter tail as compared to this.

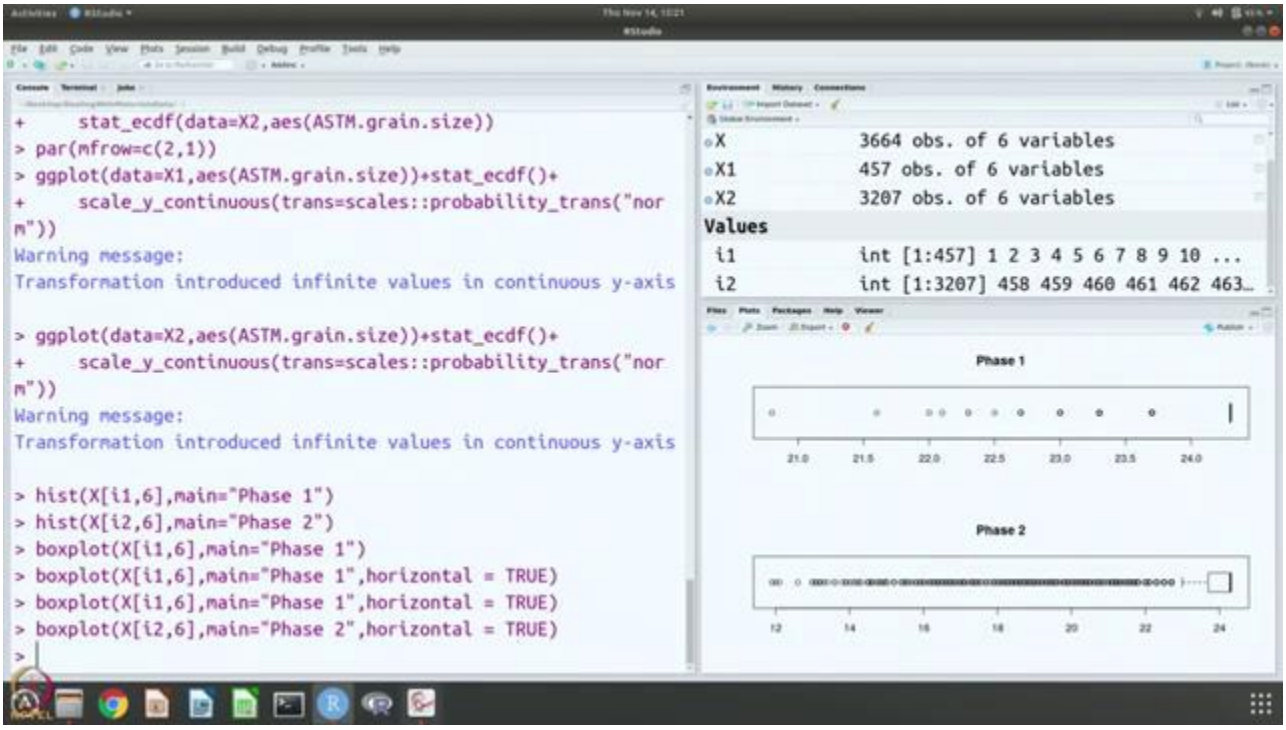
(Refer Slide Time: 01:34)



So, this is what we have been noticing and so you can clearly see here in these pictures how it looks like. So, you can do the box and whisker plot also.

(Refer Slide Time: 1:51)



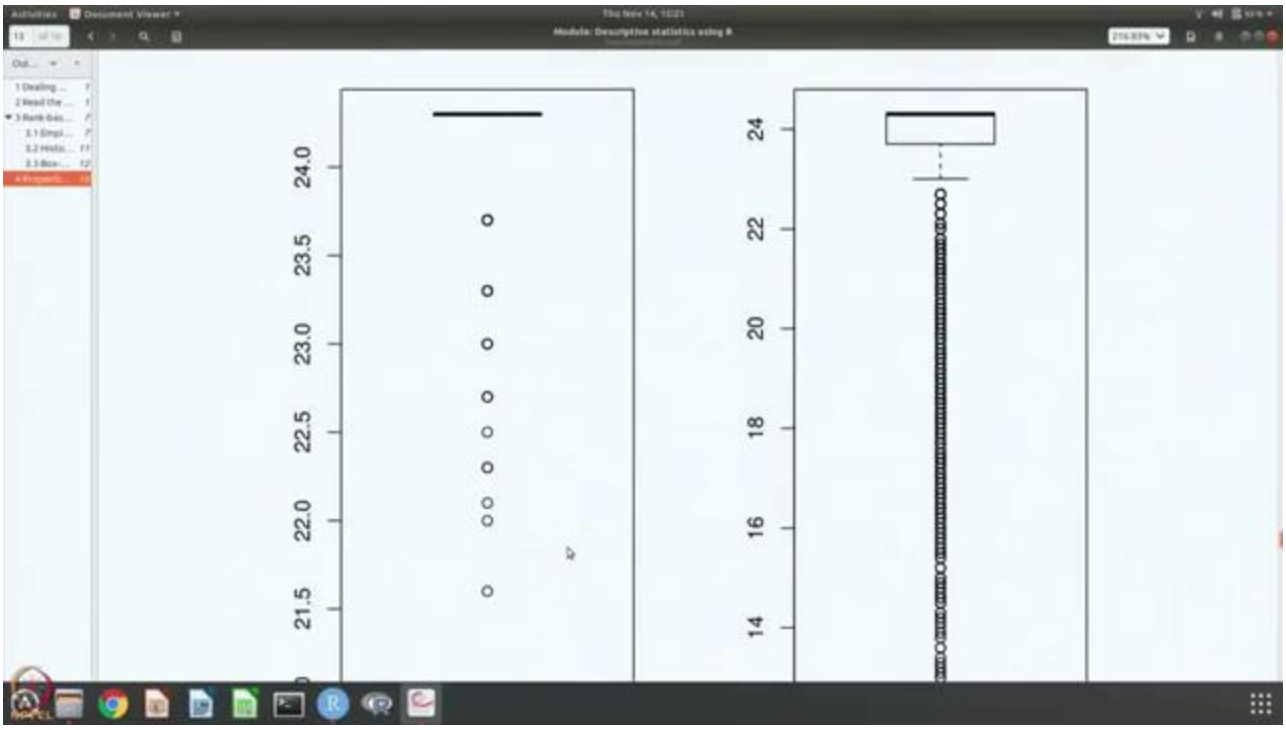


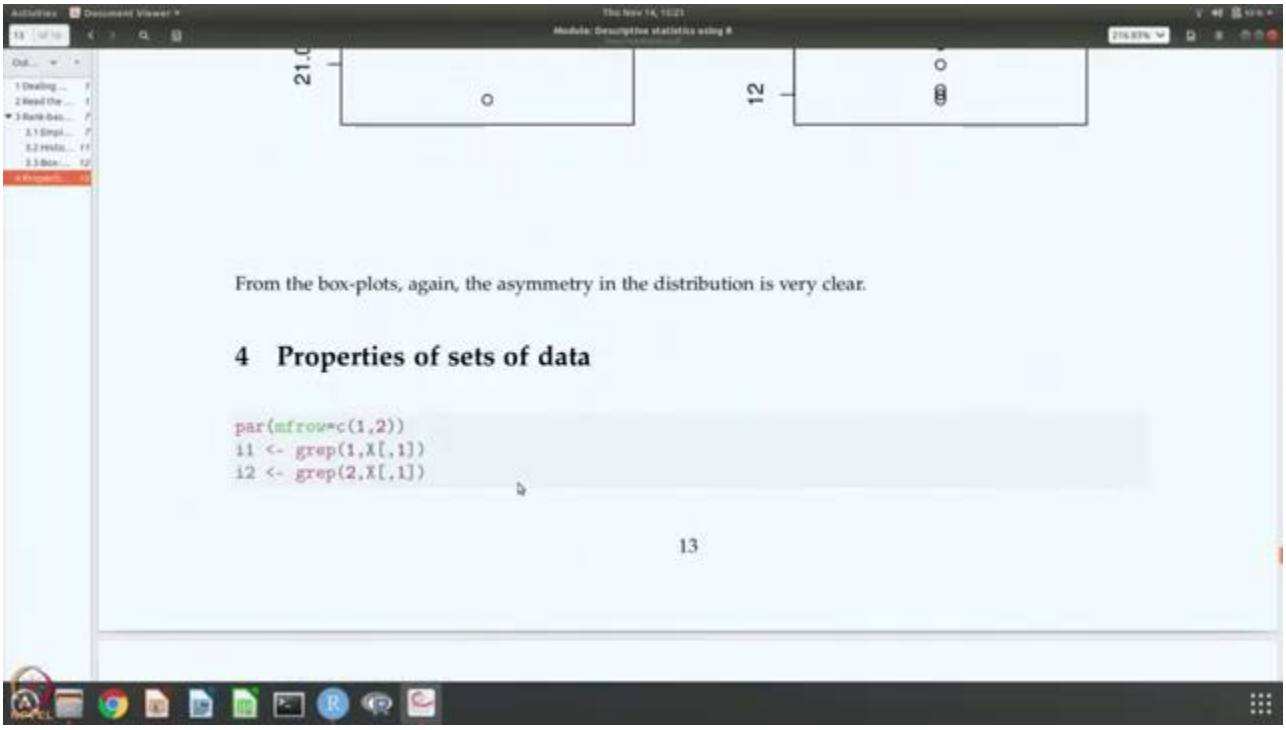
So, we just need to change and let us, let us make it horizontal, true I have to do this again and I have to do it for phase two. You can clearly see that these data sets are really, really different from what we have been seeing.

So, this point is generally the median value and you can see that in both cases, the data point that occurs maximum number of times is actually the last value and then you have tail and here the tail is much longer and here there are lots of outliers.

You know, this is like the box is supposed to be second and third quartile and beyond. This is supposed to be fourth quarter. There are no data points that is because there is no tail on that side at all and on this side you have one but lots of data points are lying outside of that and here you know everything all these points are outside. So, the, the second and third quantiles, quantiles and there is no fourth quantile everything is sort of collapsed here.

(Refer Slide Time: 03:26)

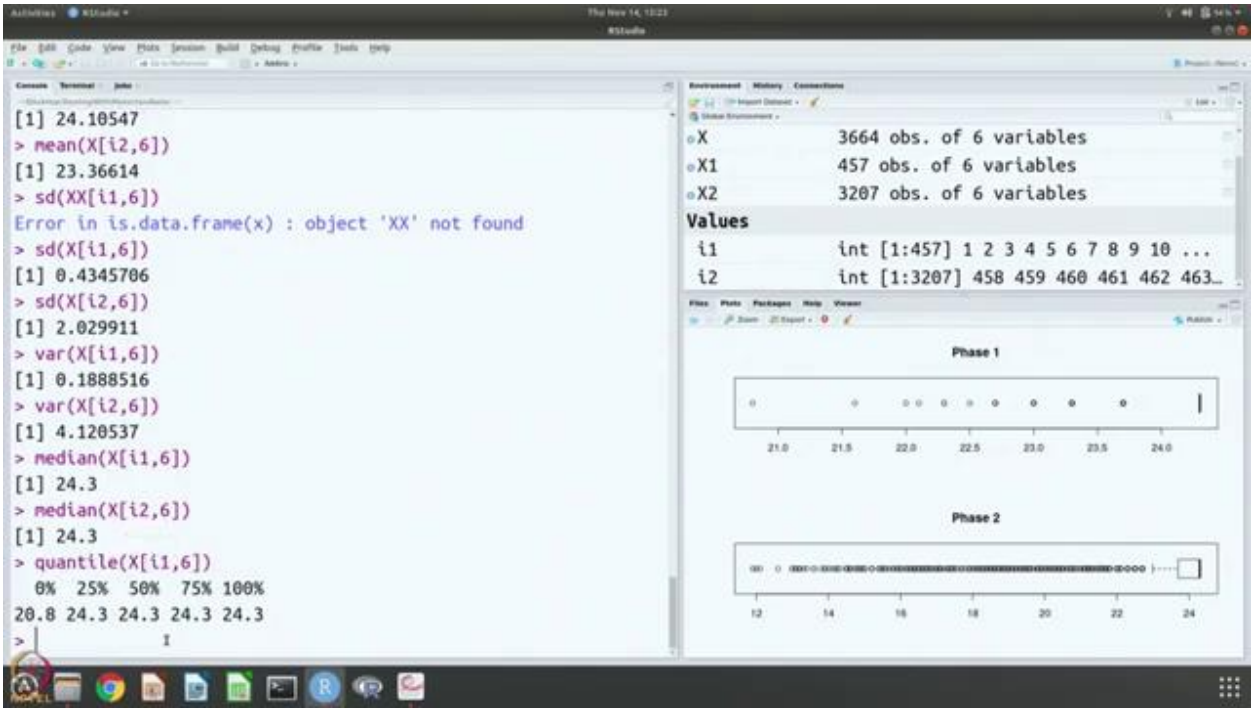




From the box-plots, again, the asymmetry in the distribution is very clear.

4 Properties of sets of data

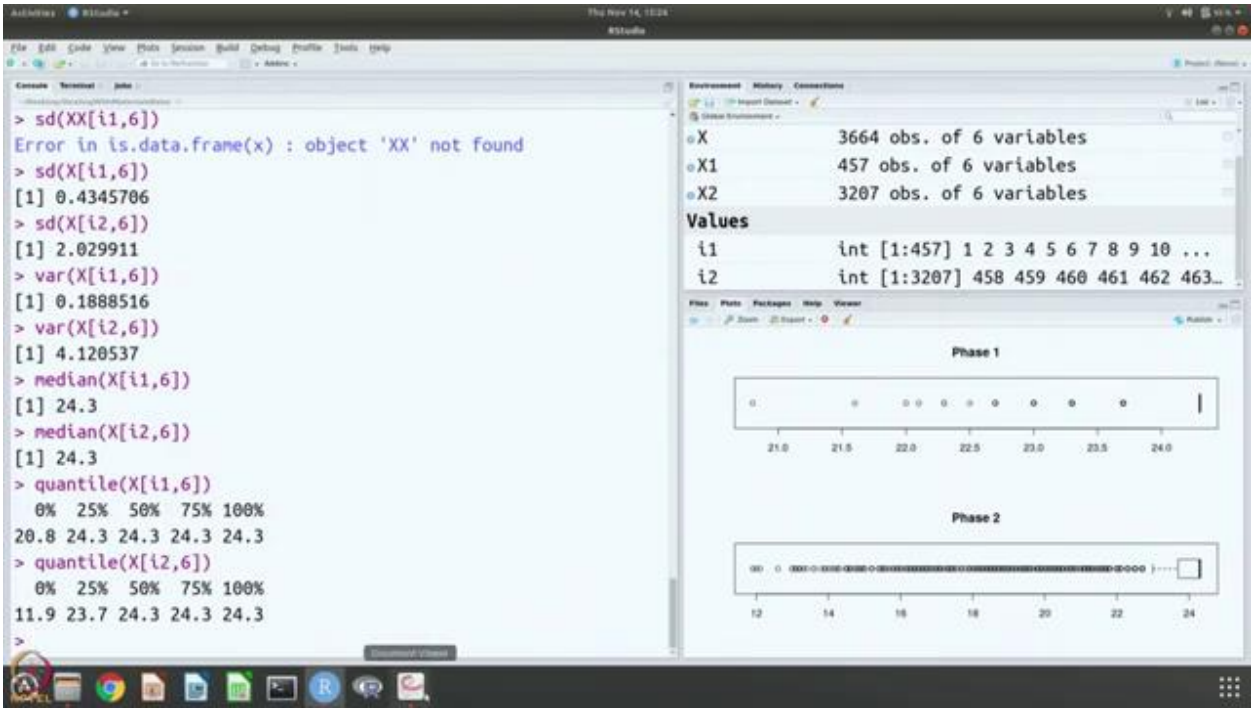
```
par(mfrow=c(1,2))
i1 <- grep(1,X[,1])
i2 <- grep(2,X[,1])
```



So, this boxplot also gives you more information about how the data looks. So, now let us try to get the properties of these data sets. So, let us get the mean of the two. So, mean is 24.1 and 23.4. Now the standard deviation will tell you this values we have already seen, and so you can see that for 1, it is 24.1 plus or minus point 4, and for 2 it is 23.4 plus or minus 2. So, it is not surprising because most of the values are here and the average turns out to be here.

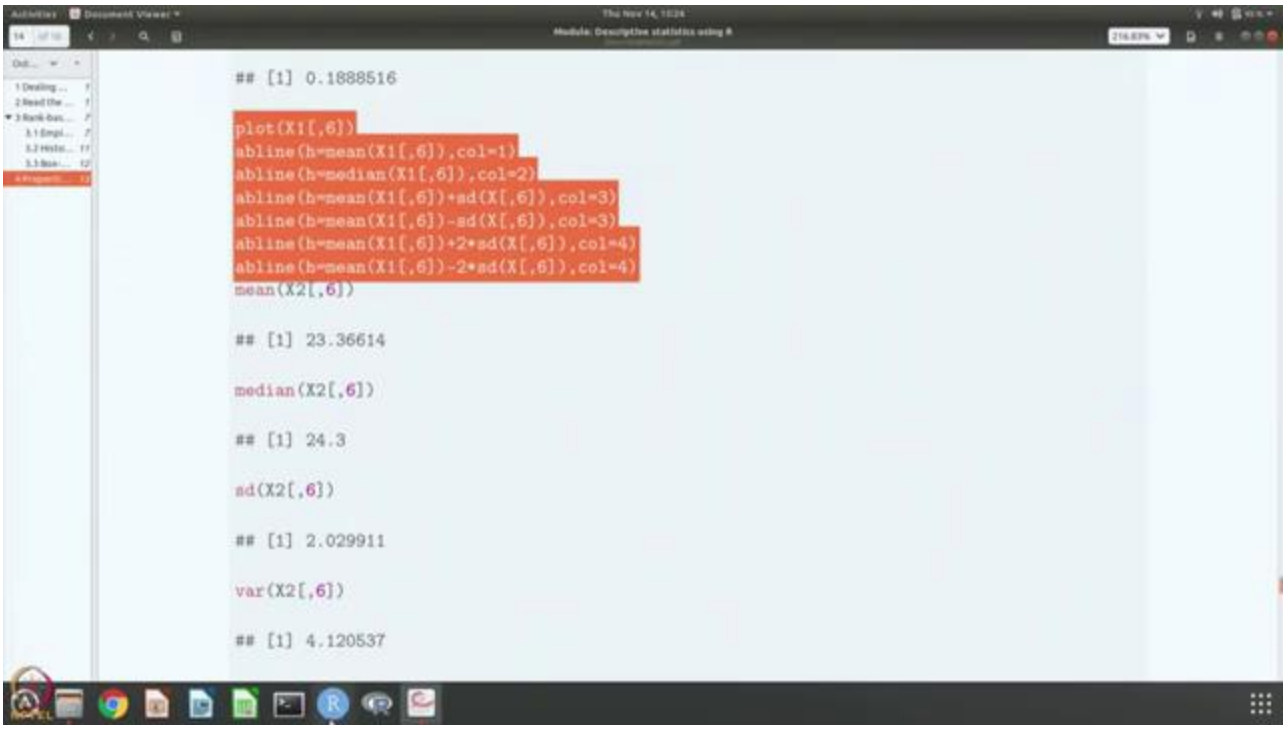
Of course, the standard deviation is much larger as expected because the spread is much higher. So, it is 5 times the standard deviation and you variants will be correspondingly different because the variance is just square of this standard deviation. So, it is 2 times 4 and it is 0.4, 0.16. So, this is just square and so mean, and then you can get median values. We already know from the boxplots and so in both cases the median value is 24.3 and that is what you see here both cases just use your 24.3.

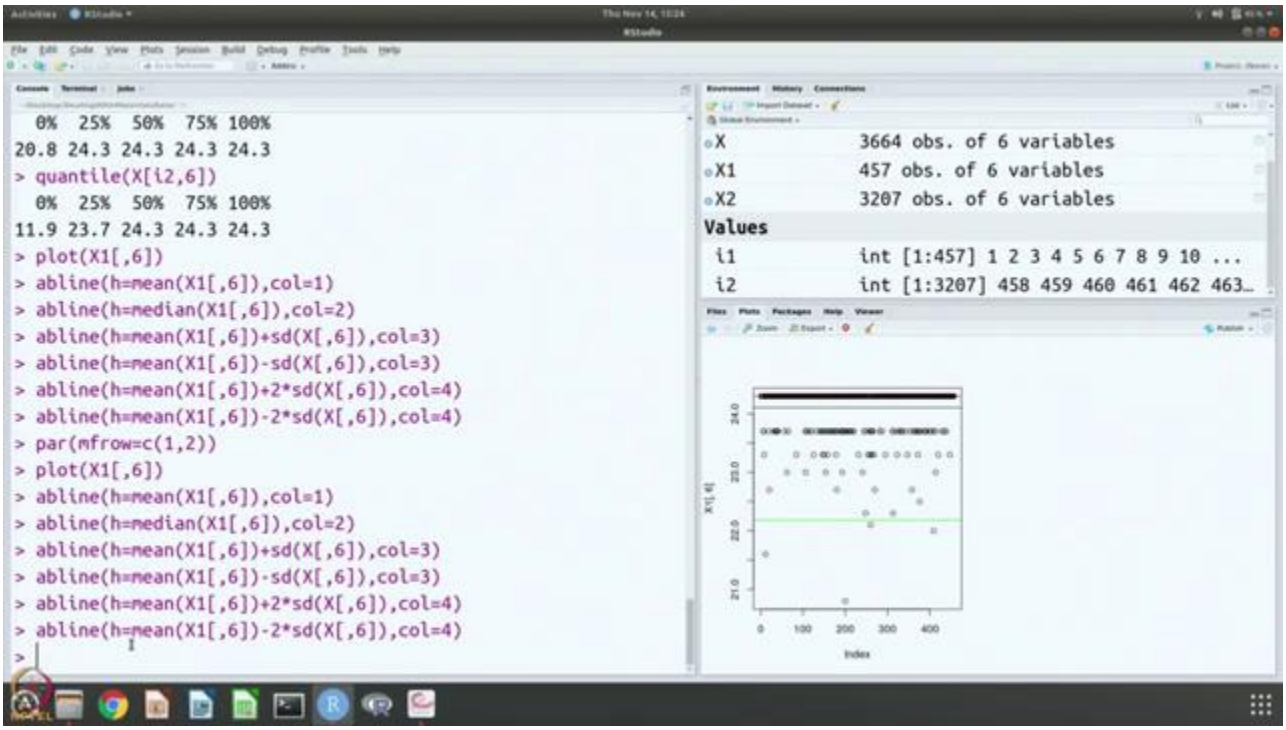
(Refer Slide Time: 05:35)



So, quantiles again the information is already there from the boxplot splitting them in terms of numbers and you can see that the spread is here from 11 to 24.3 and here it is, and you can see 50 percent, 75 percent hundred percent everything falls in this 24.3 and so, also here and there is some slight distribution here. I mean in fact here 25 percent onwards everything here only, the first quantiles is slightly different value here only from the 50 percent. So, that is the third, fourth quantile they are all on the same value.

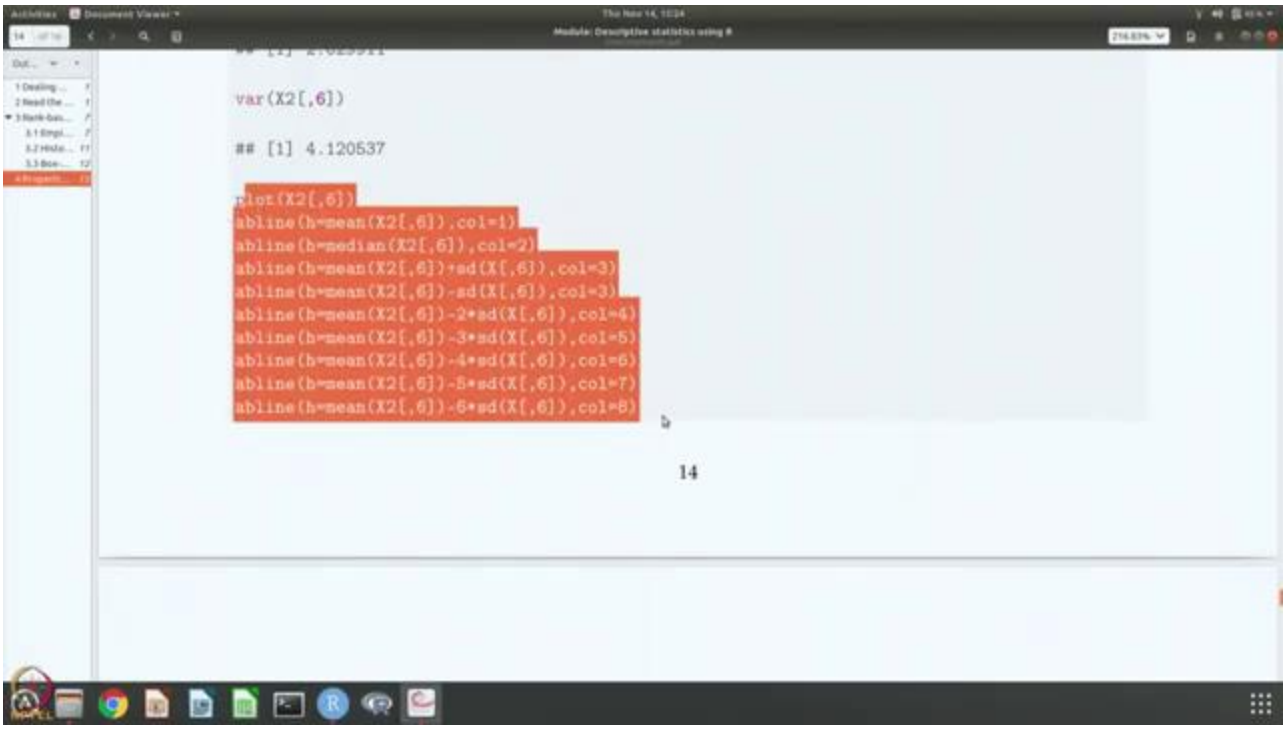
(Refer Slide Time: 06:13)

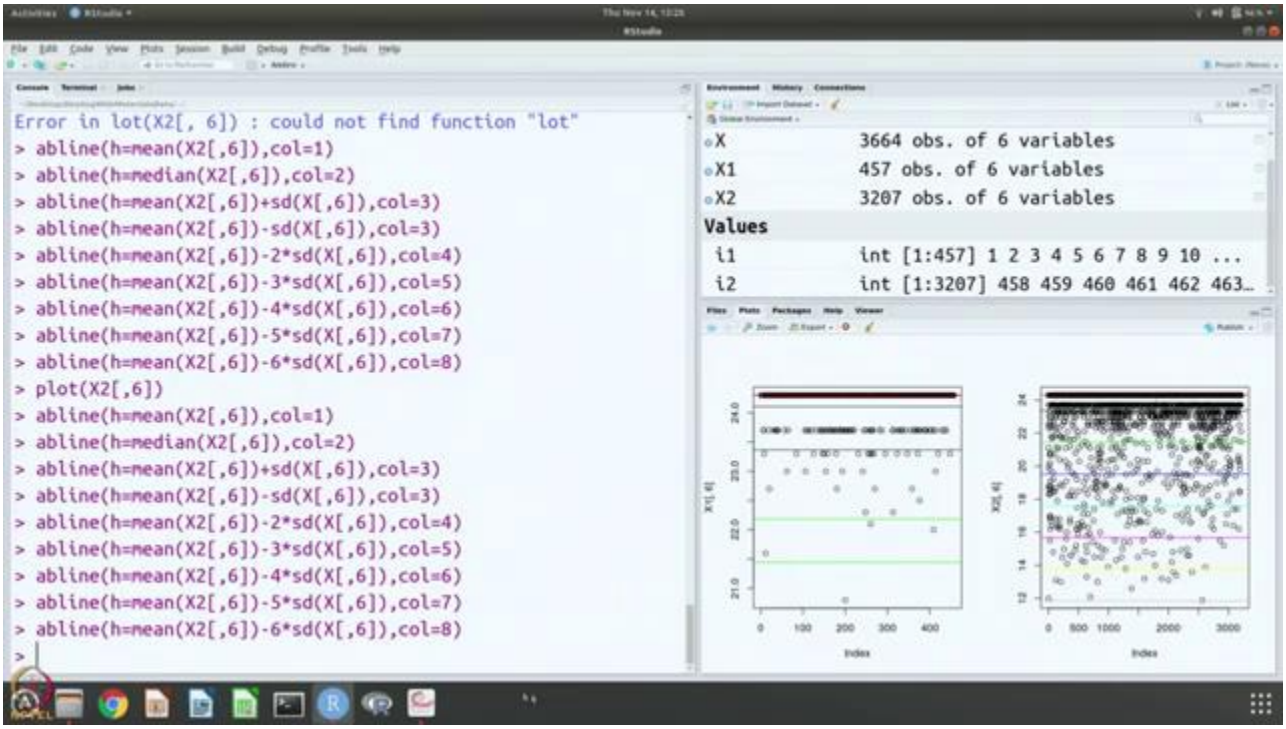




So, this is what we have seen also and as we did earlier of course, let us plot everything in one go. These plots are useful, so, I am going to change this a little bit, instead of.

(Refer Slide Time: 07:04)

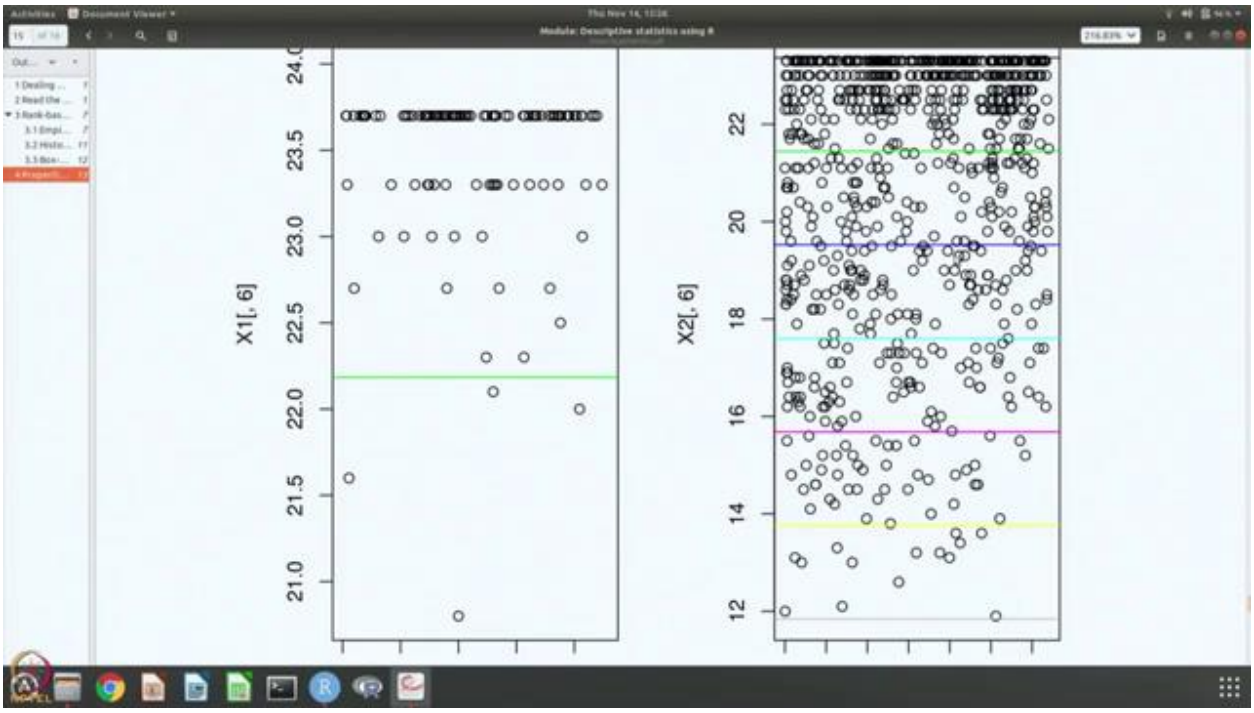




So, let us do this, for the second data set also second dataset you can see 1 standard deviation 2, 3, 4, 5, 6 etc on the 1 side because after the on the on the plus side there is nothing. So, let us do that. So, there is some problem with the command.

So, you can see that, we mark the mean and we mark the median and we mark the 1,2 standard deviation etc. But here we are, we have been marking 1, 2, 3, 4, 5, 6. So, up to 6 standard deviations here to go before you actually take all the data into, into account. So, this is a really large spread and that is what is shown in this figure also.

(Refer Slide Time: 08:12)



So, you can see that, so, somewhere here this read is median, black is mean and about the black, the first standard deviation is green, and second is blue. In this case, you do not even see the blue, you see the blue, and then this light blue and purple and yellow and I do not know what this color is.

But, so, there are several standard deviations from the mean on 1 side you have to go before you encompass all the data points. So, we and this information we have also seen in the case of quantile. So, to summarize, we have looked at grain size distribution, we have looked at 2 different cases 1 was very straightforward.

We just had the data for one phase, because it was a single phase material. In the other case, it was a two phase material and we pulled out data and we separated it into phase 1 and phase 2 and we carried out the same analysis, the there are three things that we did.one is just plotting, that is stem and leaf or dot chart, scatterplot.

Second one is to prepare rank based properties that is by doing some analysis like cumulative distribution, boxplot, histogram plot and things like that and the last one is to prepare summary based reports like mean, median, standard deviation, variance, quantiles, etc and of course, we tried to put the scatterplot along with the summary based reports to get a better handle on how the data looks like.

And it is very clear from this data that if you are having something like grain size, it is better not just to report mean and standard deviation, but also some information about the distribution and probably the best way to represent the distribution is by giving the histogram plot and so, that is also common histogram our cumulative distribution plot gives an idea about how the data looks like in this case, how the grain size spread is in these cases.

So, we will continue with more of the analysis and this example and the previous example is to show that sometimes in Materials Science and Engineering, you come across data sets, which are to be represented as distribution not just in terms of simple numbers. So, thank you.