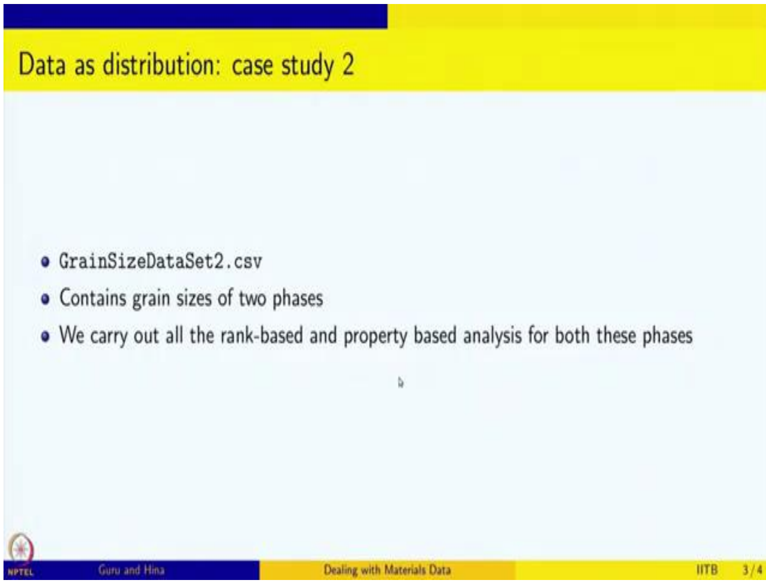
Dealing with Material Data: Collection, Analysis and Interpretation
Professor. M. P Gururajan
Professor Hina A Gokhale
Department of Metallurgical Engineering and Material Science
Indian Institute of Technology Bombay
Lecture 24
Case Study: Grain size in two phase steel

Welcome to dealing with materials data, we are looking at the collection analysis and interpretation of data from the material science and engineering and we have done two modules, we have done an introduction to R module and this module is meant for doing descriptive data analysis using R and we have already done analysis on two sets of data, one is on the conductivity of electromagnetic, tough pitch copper and we also did an analysis on grain size and we found that the conductivity measurements and the grain size data are of two different types.

The case of conductivity measurement, the repeated measurements just gave errors about some mean value, because those are measurement errors or random errors or uncertainties associated with the experiment. But, on the other hand, when we measured something like a grain size, it has naturally a distribution. Not all grains are of the same size, they follow a distribution and the distribution is not normal or Gaussian or anything like that it is slightly more complicated.

So, it makes sense to represent the data in this case, not just with the mean and standard deviation, like we did in the case of conductivity, but also by plotting something like a histogram plot to indicate how the data looks like.

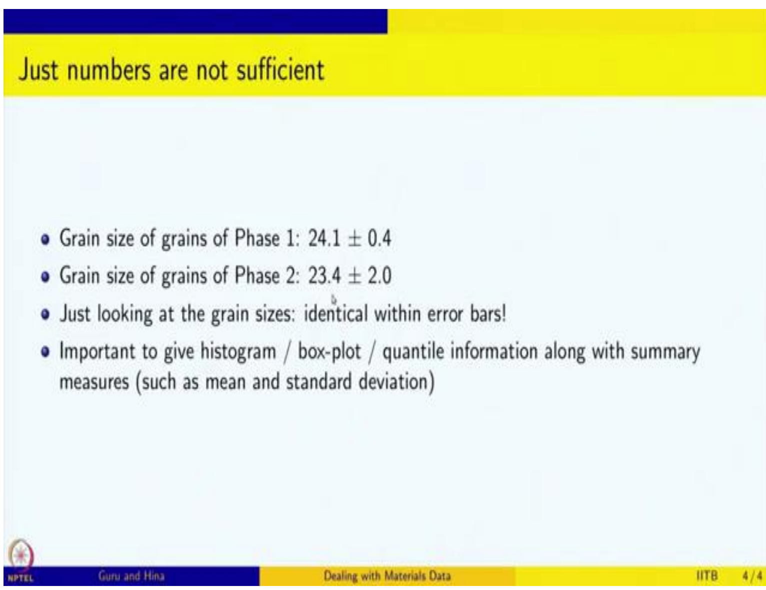
(Refer Slide Time: 01:49)



So let us now continue with the second data set. It is also data set which deals with distributions and this is the grain size data set to dot CSV and this is slightly complicated data set because it contains grain size of two phases and we are going to carry out all the rank based and property based analysis for both these phases and in all the cases we are going to do it, one next to the other.

So, we can have information about grain sizes of these two phases. But, we will also have a comparison between the grain sizes of phase one and phase two in terms of their rank based and property based summaries. So, that is what we are going to do.

(Refer Slide Time: 02:32)

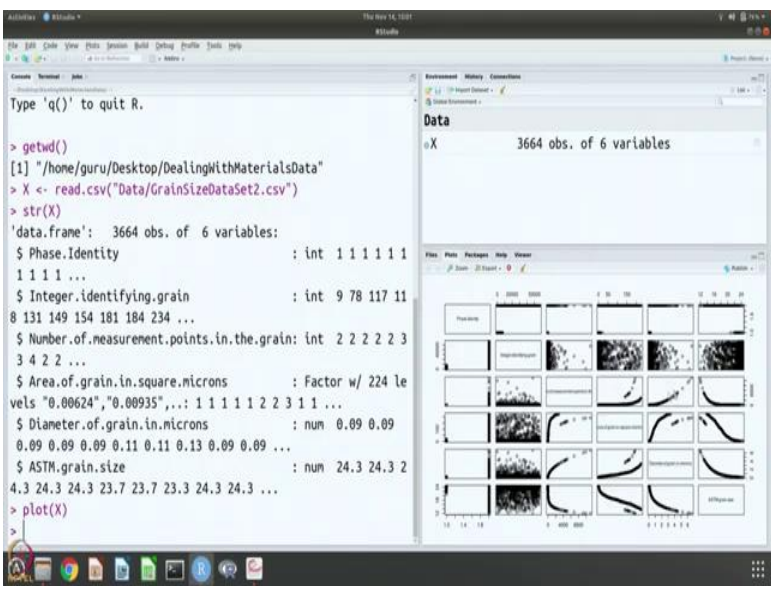


One of the things that you have to carry from this session is that if you just looked at the grain size, likely reported conductivity just by reporting the mean and standard deviation. You will see that the grain size of grains are phase one is 24.1, plus or minus point 4 and grain size of grains of phase 2 is 23.4, plus or minus 2.

Sometimes students make a mistake of thinking that these two grain sizes are different and grain size of grains of phase two is smaller than one that is not true. Because you also have to take into account the fact that there is an uncertainty, when we say 23.4 plus or minus 2, it means that the number could be anywhere between 21.4 to 25.4 and so the 21.4 to 25.4 actually covers 24 also, when we say 24.1 plus or minus point 4 that means, it is 23.7 to 24.5.

So, within the error bars, all that you can say is that these two phases have the same grain size. However, if you look at the histogram as we will do or look at the quantile information, you will see that the mean and standard deviation are not the complete picture. These two grain size distributions are very different even though they might end up giving you more or less the same grain size and that is the part that has to come through the session. So, that is what we are going to see and we are going to understand.

(Refer Slide Time: 04:26)

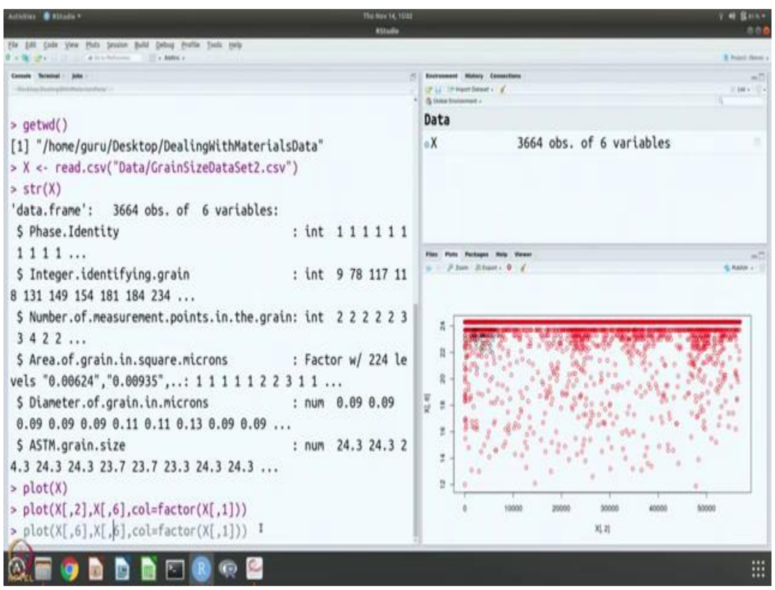


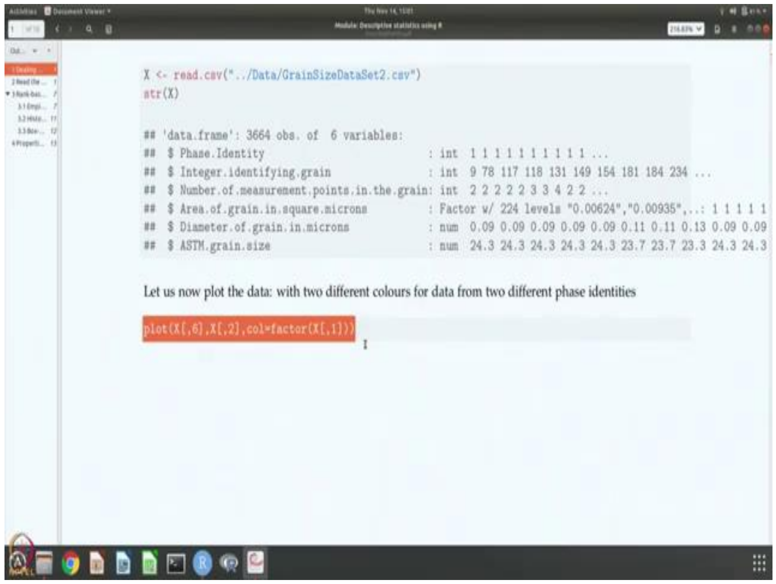
So, let us do as usual we want to open R and so we have to check the version, we have to make sure that we are in the right directory and so we are all set to now start importing the data and in this case, the data is in CSV format. So, read dot CSV and it is from data directory and grain size two is the data set two. As you can see, there are 3664 observations and there are 6 variables 6 variables because in addition to the 5 variables that you saw in the other case, here the grain at the phase identity is also included before grain identity.

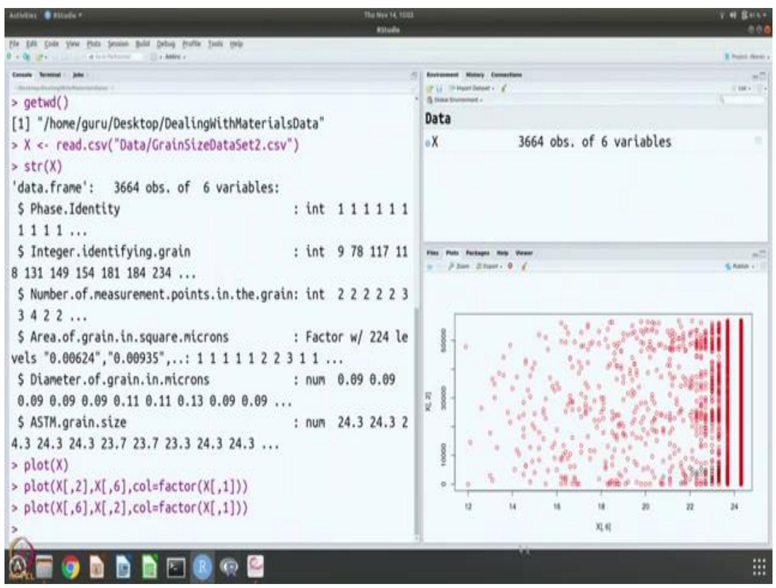
Of course, you can get more information by looking at the structure of this object x. It is a data frame, there are 300 and, 3,664 observations, there are 6 variables, the variables are phase identity, integer, identifying grade number of measurement points, the area of grain diameter of grain and ASTM grain size, like we did earlier, we are going to be worried only about integers identifying the grain and ASTM grain size. Of course, for the two phases, phase identity is the, is this is a 2 phase microstructure. So, there are 2 phase identities 1 and 2 and we are going to be working with these two.

First thing of course is to plot we can always try plot x so it is a 6 by 6. So, you should have 36 boxes 3, 4, 5, 6 and 3, 6. So 36 boxes are there and so against phase identity against integer identifying grain etc, all the parameters are plotted and so you can see how the data looks. So, this is the first step and this does not distinguish between the grain identities.

(Refer Slide Time: 06:58)





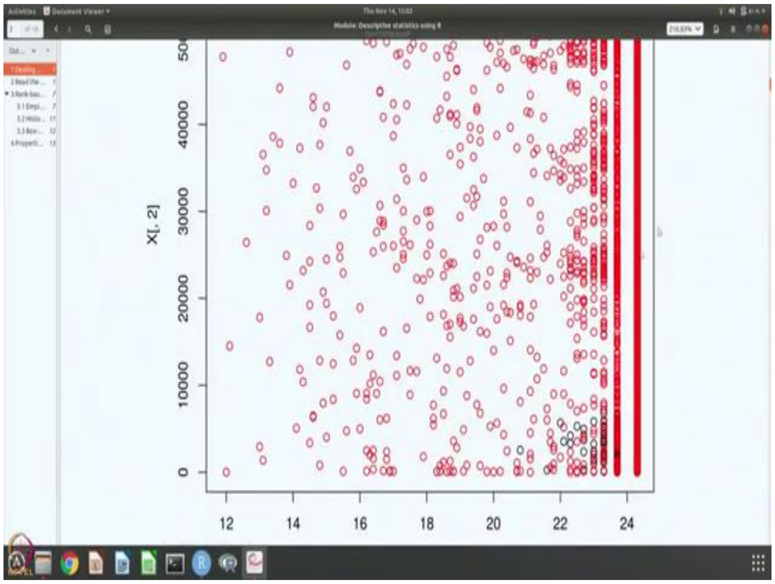


So, let us do the plotting of, what is this command? So, we want to plot the grain identity versus grain size and we want to color them according to the phase identity. So, phase identity, one should get one color and phase identity two should get another color.

So, this is how the plot looks. So, these are the sizes and these are the grain identities, you can also switch them, which is how it was the original one and you will get this I am showing this plot because this is closer to what you would see in the case of a dot chart.

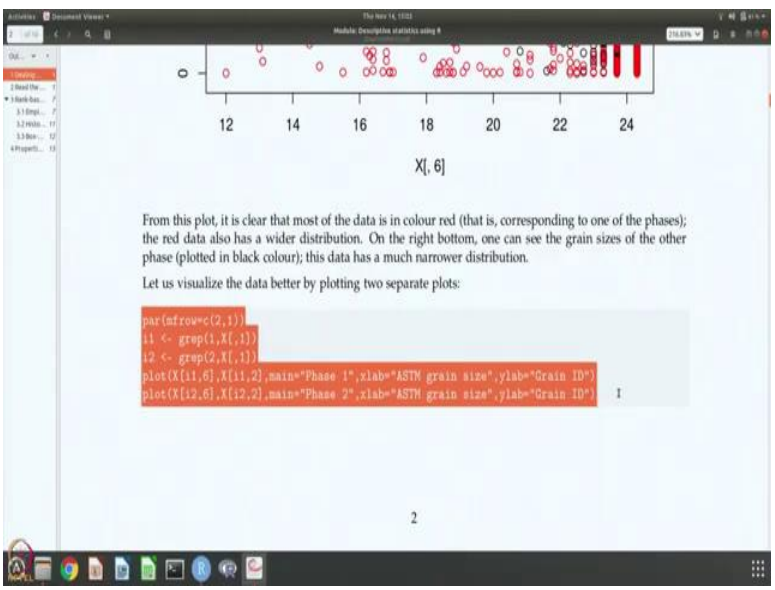
But it is very difficult here we know to distinguish between these black points and red points. However, in dot chart, they will be separated. So, dot chart is always a nice way of visualizing the data instead of plain scatterplot that you can make. Of course, you can also play with the plain scatterplot yourself and come up with commands which will separate these data. But there is an easier way by using the existing libraries.

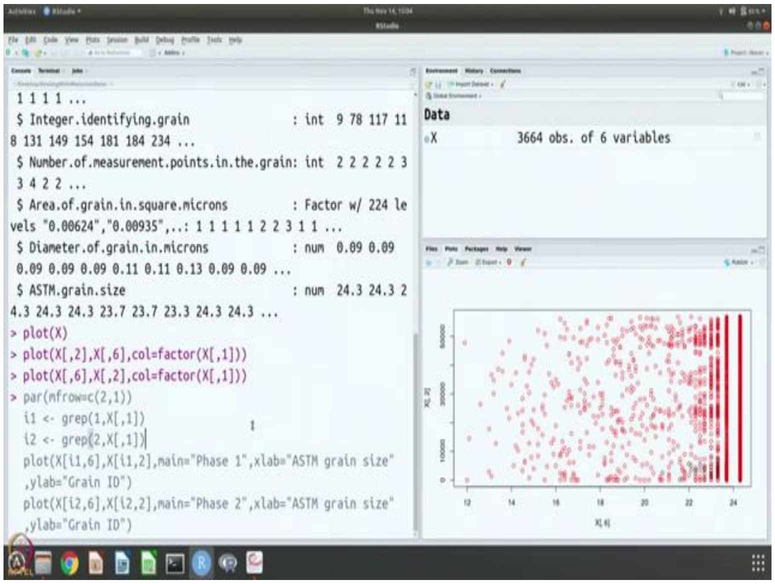
(Refer Slide Time: 08:12)



So, that is what and here it is much clearer. So, there are lots of red points and fewer black points and the black points are all clustered here and the red points are spread all over. So, black points are between 20 and 24. Whereas red points are between 12 and 24 and odd. So, this is an important point. So, we will come back to it.

(Refer Slide Time: 08:38)

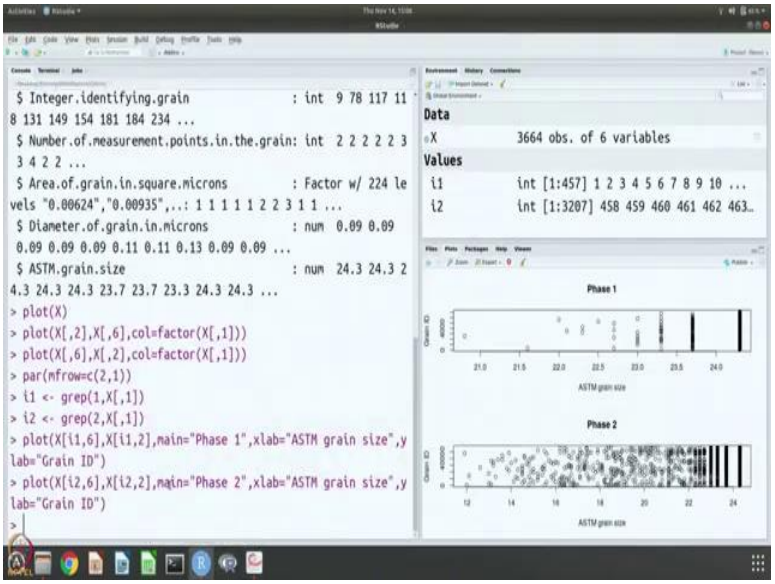




So, let us do the other thing. Let us separate out the data. So, this tells R that we are going to make two plots, and one on top of another and those two plots we are going to plot by using phase one phase two. So, the grep command, so is going to get all the line numbers or rows, which have data of phase one and this grep command is going to get the row numbers of all the data points which has data for phase two, that is what I1 and I2 do. So, if I plot x I16 versus xI12.

So, this will be only data that is corresponding to phase one, because we have separated that those line numbers and I2 is for those data rows which have data about phase 2. So, that is what this is and of course we are going to label them as phase 1 and phase 2 and x label is ASTM grain size y label is grain ID.

(Refer Slide Time: 09:53)

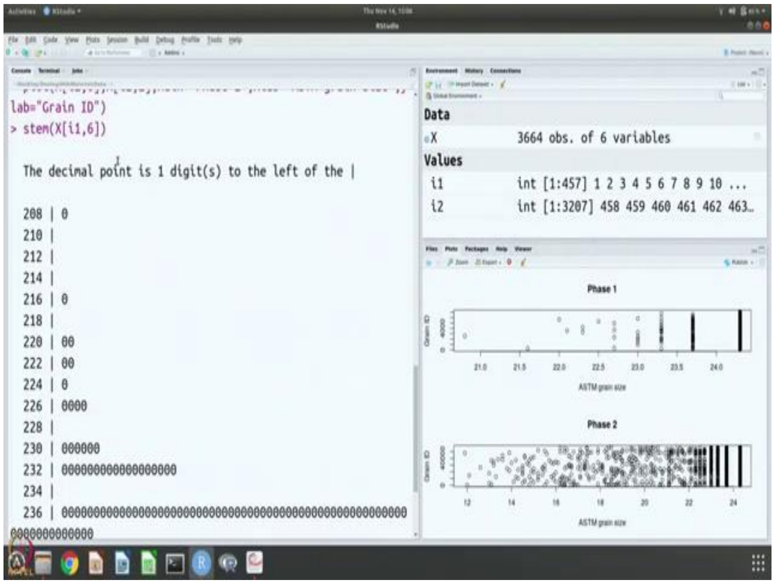


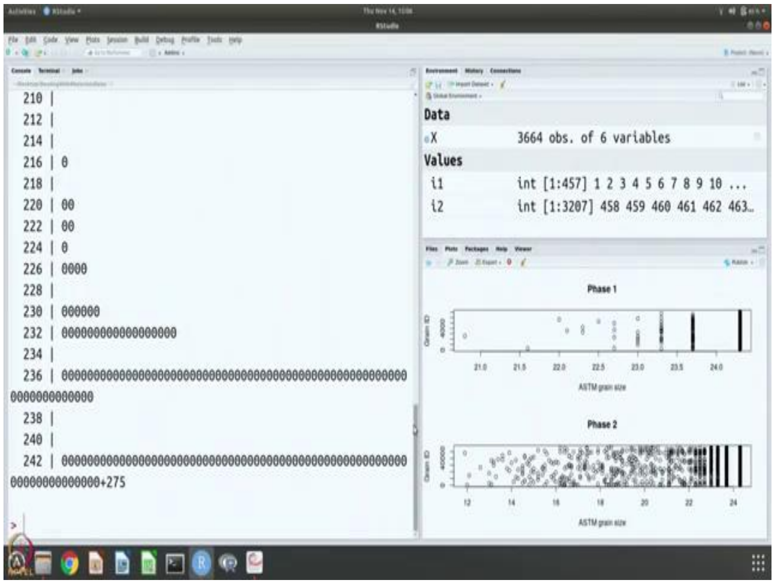
So, let us do this plot and you can see that phase 1 has grain sizes, ranging somewhere between 20 and 24 and odd and phase 2 has between 12 and 24. So, this is about to 3, this is about 12 the spread is 4 times as much, like I said we later we are going to see that both of them are willing to give you a grain size which were somewhere which is somewhere about 24 both are going to show you the same grain size and of course, this is going to show more of spread it will show a spread of two, as compared to spread of 0.4 here.

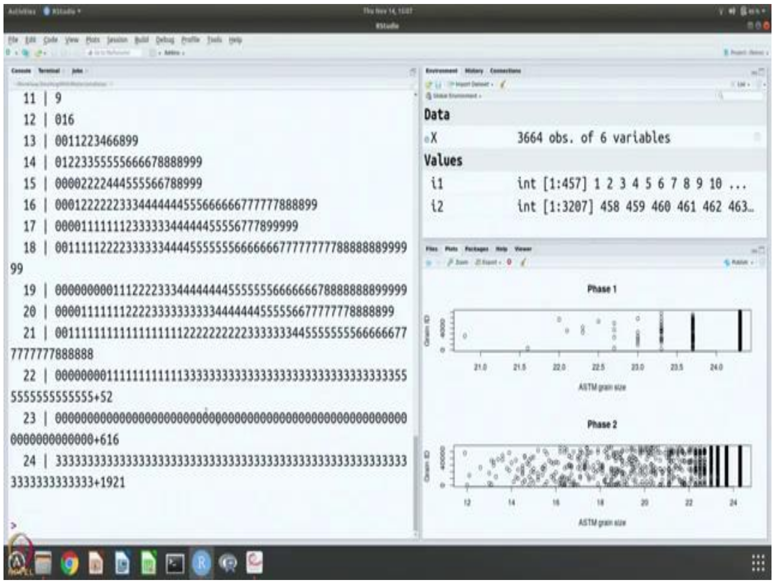
So, 5 times however, looking at this data now, it is very clear that the phase one and phase two grain sizes are completely different in terms of their distribution, even though the overall grace properties like me might be the same.

So, that is the point behind making this broad. Like I said the, you do not have to make all this two plot separate, dot chart will do it automatically for us. So we are going to look at that. But before we do the dot chart.

(Refer Slide Time: 11:15)





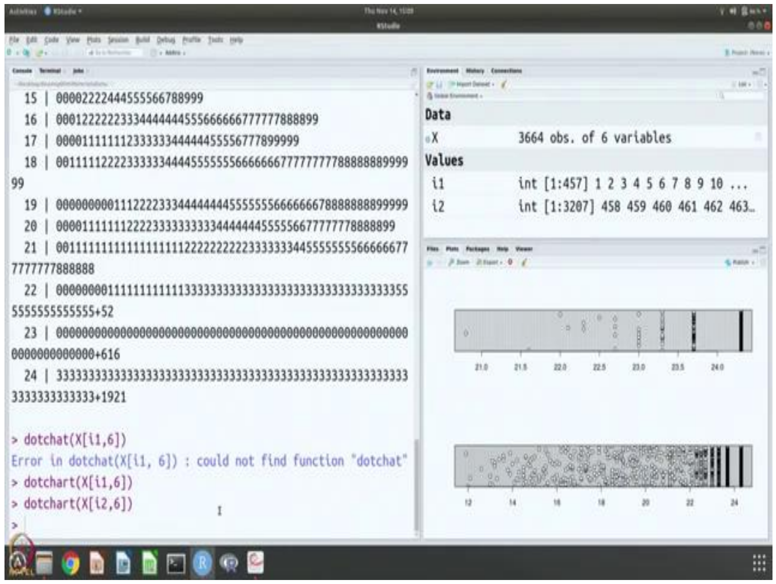


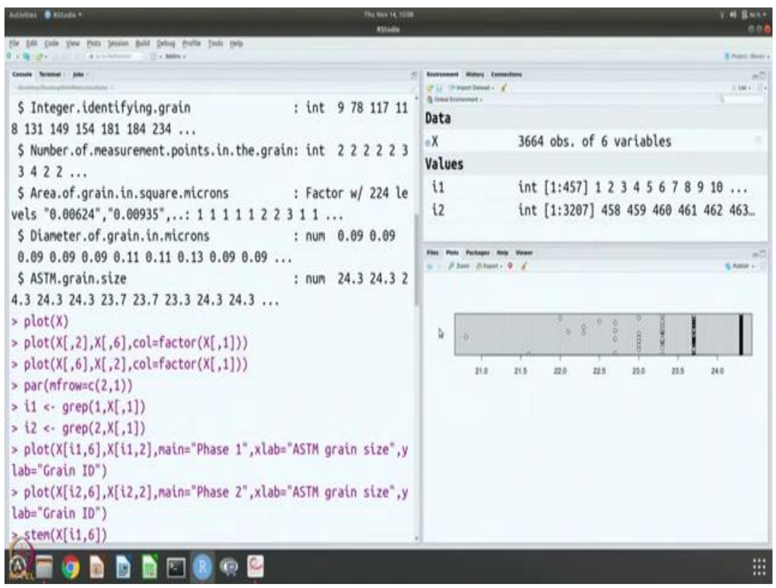
Let us do the stem plots and again, we are going to do two stem plots. First stem plot is for identity one, you can see. So, the decimal point is one digit to the left of the pipe symbol. So, it is 20.8 and 21.0, 21.2, etc, 21.6, and 22.0, 22.0, 22.0. So, these are the data points and you can see that it has a tail and then it slowly peaks and the peak is here. There are 275 data with this 24.2 number and that is why the average falls somewhere about 24.2 because the rest of the numbers are small compared to this value and the data goes like this and then it just speaks right.

So, it has a long tail on one side, but there is nothing on this side there is no tail at all I mean this is a peak and from the peak on 1 side, you see that the data has a tail. So we can now do this for 2 stem plot because it is much more skewed. So, you see these data points 11, 12, 13, etc and then you see that there are 52 more data points 616 more data points, thousand 924 more data points.

So, this is also the peak and from there, it just goes this way. So stem and leaf plots are nice to know about the structure of the data how it looks like, which will become apparent when we do the histogram plot, which we are going to do.

(Refer Slide Time: 13:14)

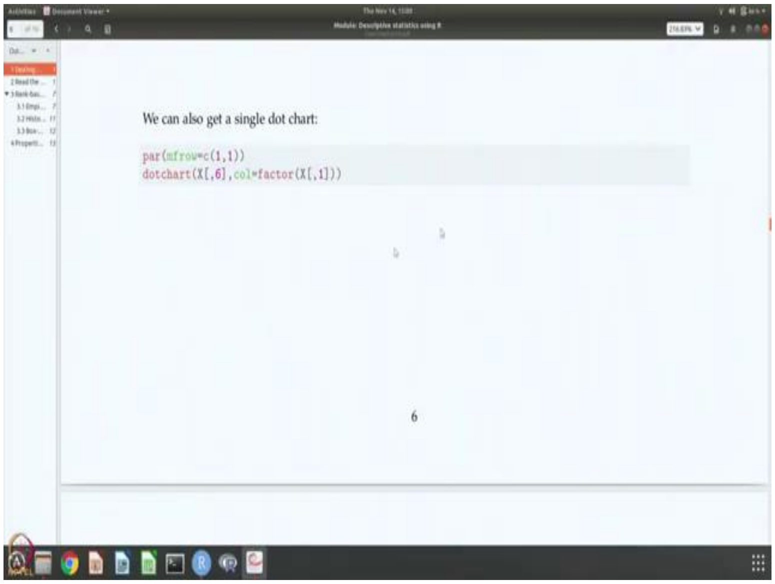


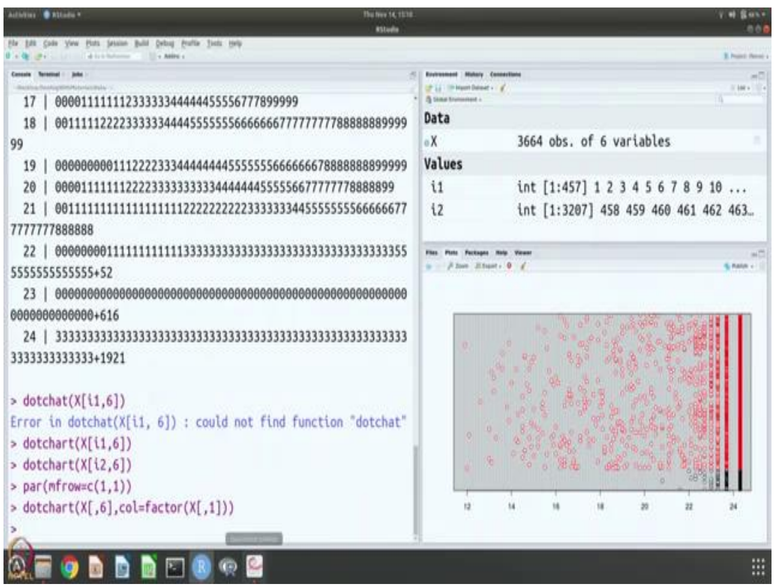


So, before that we are going to do the dot chart. Let us do the dot chart. So, we had x. Notice that. So, let us do the dot chart, notice that because we previously said that you have to do these two plots, two rows of plots, it continues.

If you want to change it, you have to give again, give the command to make sure that it starts plotting single, but in this case, we want to see the two plots for phase 1 and phase two. So let us continue. So, this is the second chart, so you can see that this is the same as the two plots we made previously and so dot chart again gives you this information.

(Refer Slide Time: 14:32)

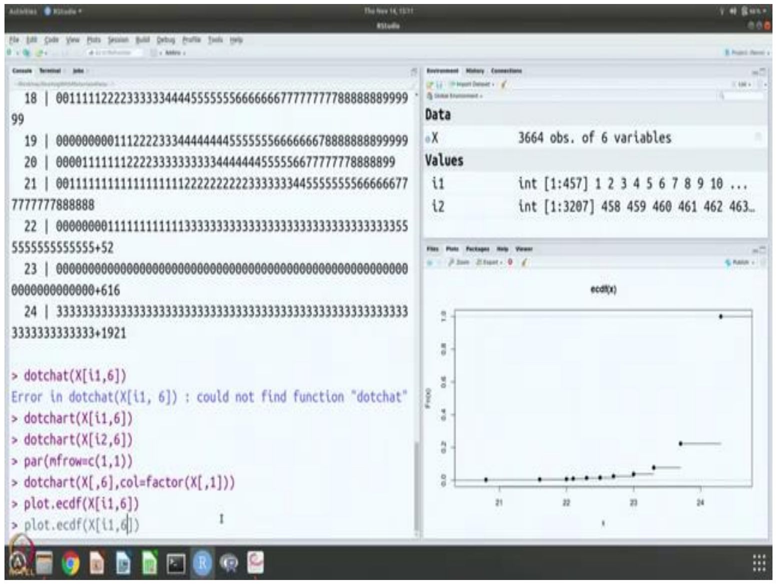


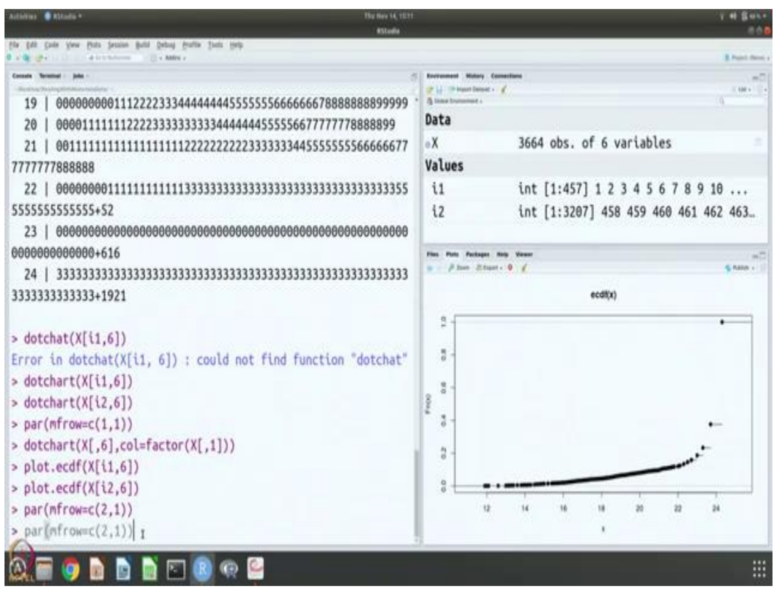


Let us make a single dot chart. So how do we do? We say, and then we do a chart of x and then we color. We color by the next factor X 1. So, we are going to say dot chart of the 6 column which is the sizes and color by factor of 1, which means phase 1 and phase 2 should be plotted with different colors and here is a plot.

Now you can see that dot chart automatically separates out the black points from red points previously, just our scatterplot actually overlap them, but dot chart automatically separate them out and puts them. So you do not have to make two different plots, just by calling dot chart with factor. Now you can look at how the data looks like. So, which is very good, which is a nice way of looking at the data. Let us now move to the rank based properties.

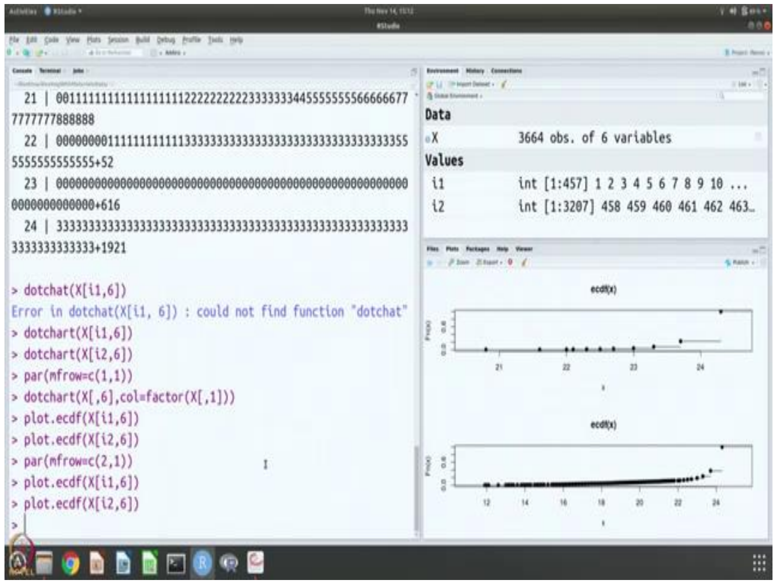
(Refer Slide Time: 16:17)

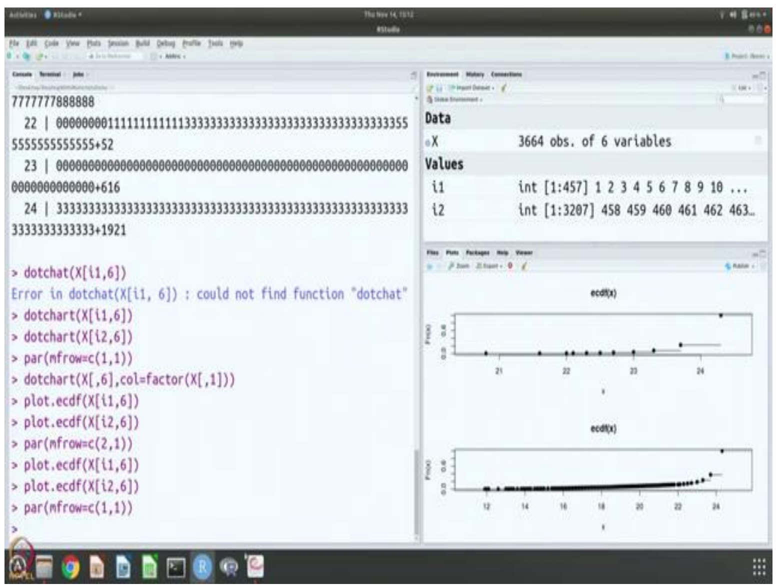




So, we will do the cumulative distribution function of course. So, you can say ecdf, plot dot ecdf of X, 1, 6. So, this is the cumulative distribution function and you can see it just goes like that, because we know that there is a peak somewhere here and it has a tail, and this is true for two also. It is also but it is a much longer and fatter tail as compared to the other one right.

(Refer Slide Time: 16:48)

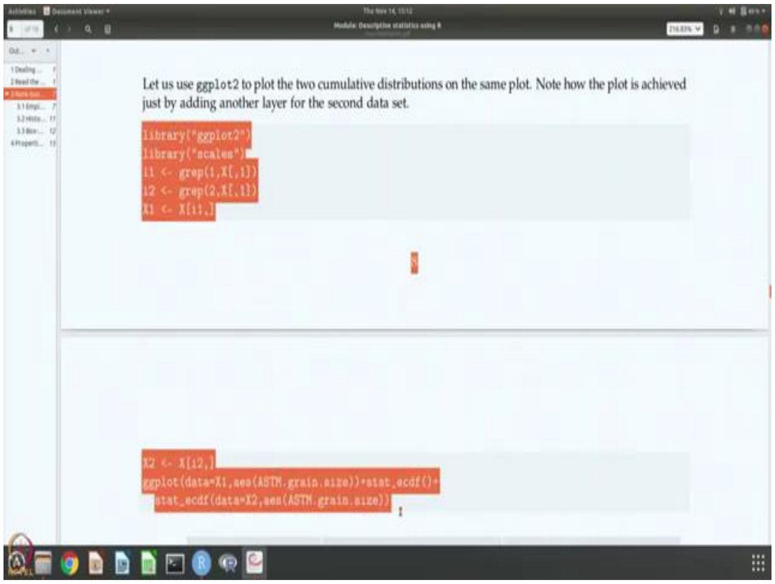


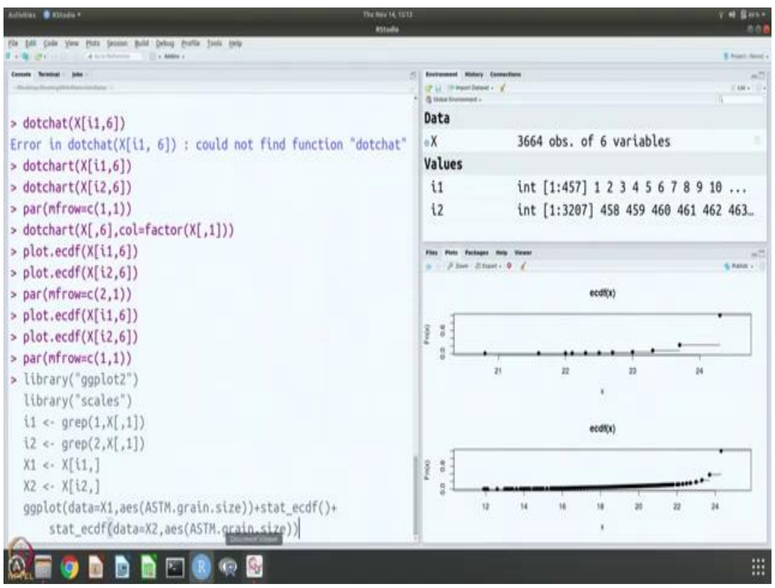


So, if you actually do this two rows then you will immediately see because see both of them go from 0 to 1, and in this case you have so both are skewed. Both have long tails and the tail is only on one side. That is what I mean by skew. But, but the tail is much longer here and because they are on the same scale, you can also see that this tail is much fatter than this.

So, so these information about these skewedness and how much information is there in detail, but peak at this value somewhere around this value is where they peak. But, their tails have different distributions, different characteristics. So, let us go back to single plot.

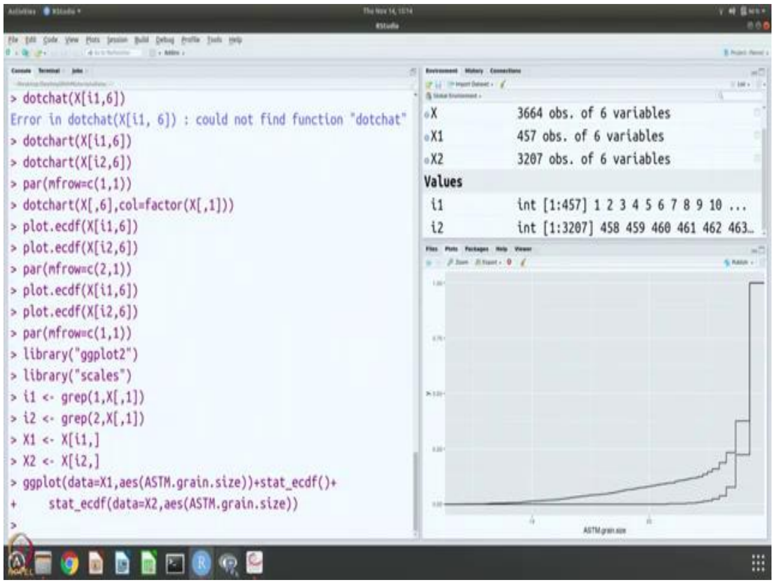
(Refer Slide Time: 17:38)





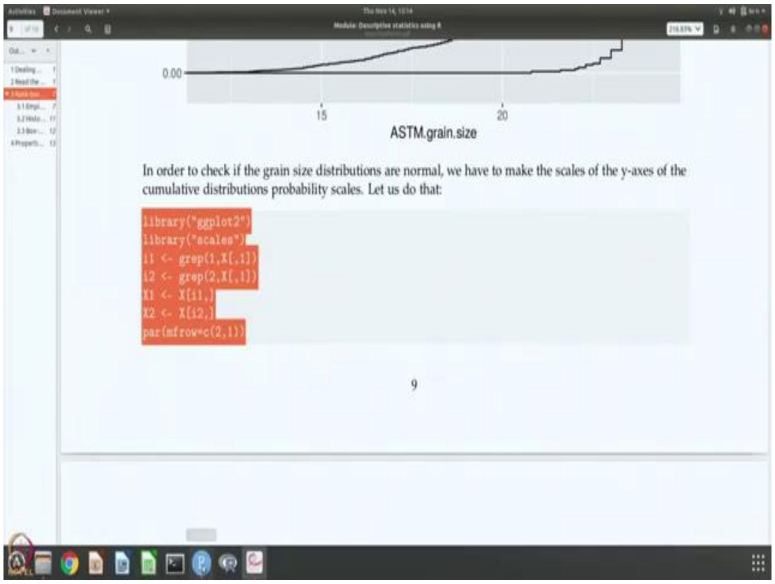
So, we have looked at the empirical cumulative distribution function and we will do. So, can we do that using ggplot? Let us do that also. So, we use the library ggplot and library scales. The reason why we are doing this is because we want to see whether the data will show any normal behavior. Obviously, it is not well to show because it is peaking at one end and it is highly skewed. Normal Distribution should have nice symmetric tails on either side, which is not the case. So, we are obviously not expecting, but if you see some kind of this kind of skew, what kind of plots do you get, for the probability scale

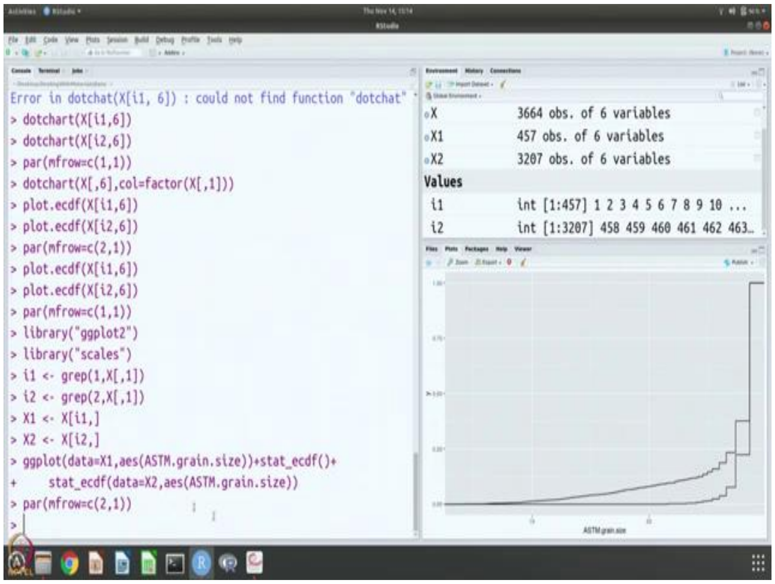
(Refer Slide Time: 18:56)



So, let us take a look at it, let us. So, I have plotted the both this on the same plot. So, usually ggplot he has to say data is x1 and you brought the grain size and that is ecdf and then you have to plot the second data. Again the ASTM grain size and you have to plot the empirical cumulative distribution function. So that is what we have done and they are on the same plot, and then what do we do?

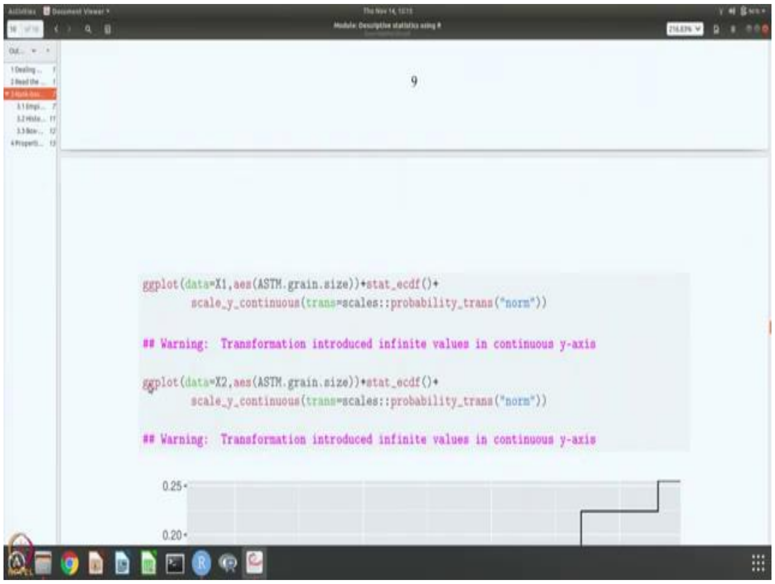
(Refer Slide Time: 19:36)

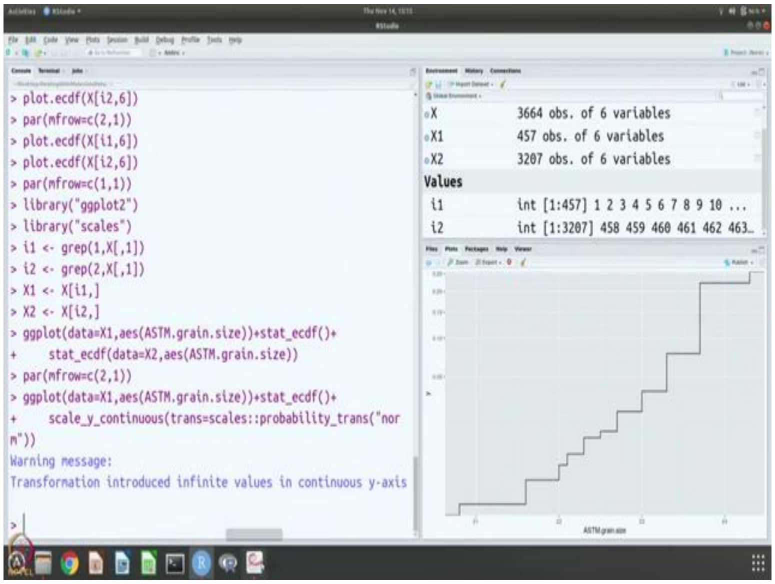


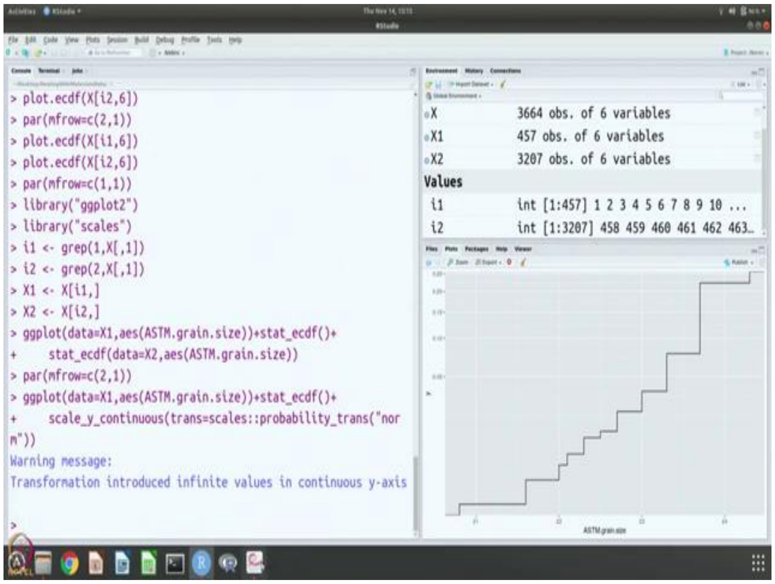


Obviously, we are going to change the scale the chain the scale these are not needed we already have this information. We also have this information. So, this is also not needed. So, we are going to make two plots,

(Refer Slide Time: 20:04)

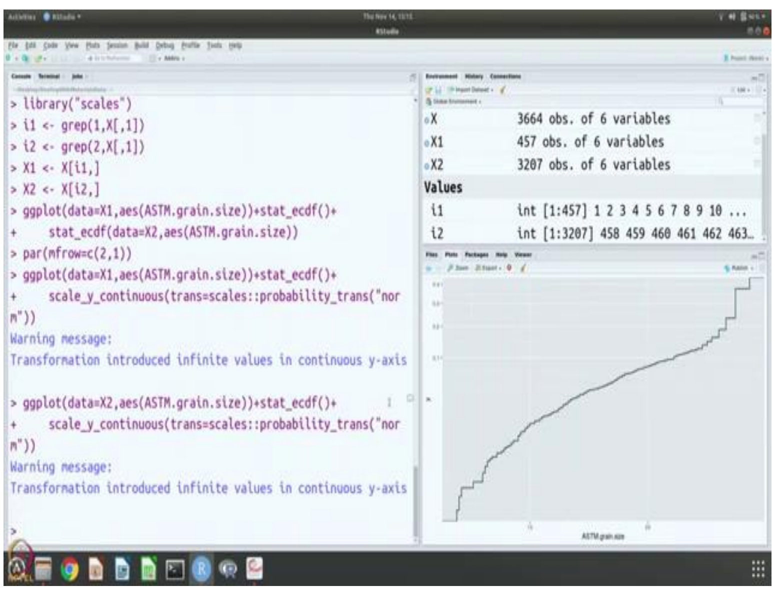






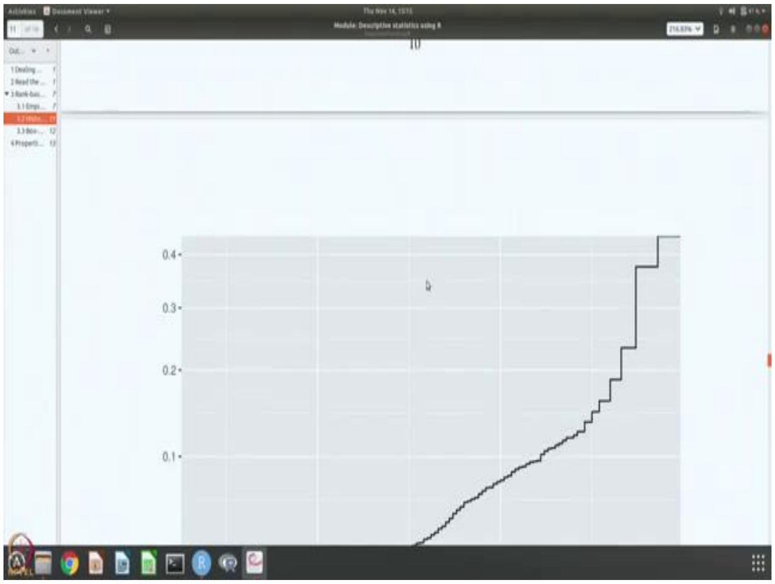
And what are those plots going to be? We are going to do the probability distribution. The scale, we are going to change it to probability scale normal probability scale. Obviously, we do not expect it to be normal.

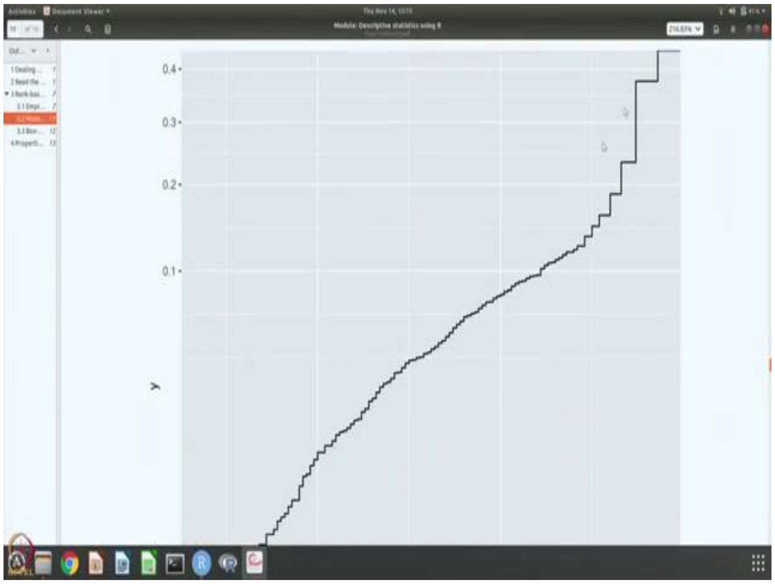
(Refer Slide Time: 20:29)

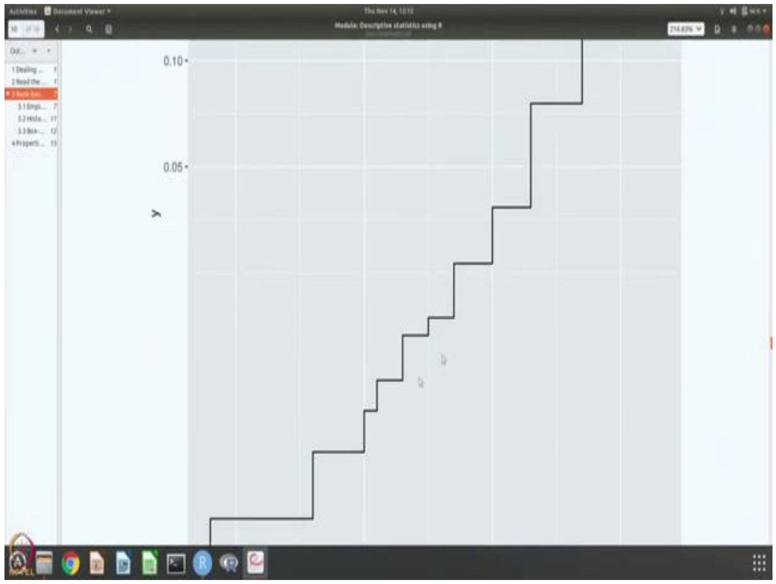


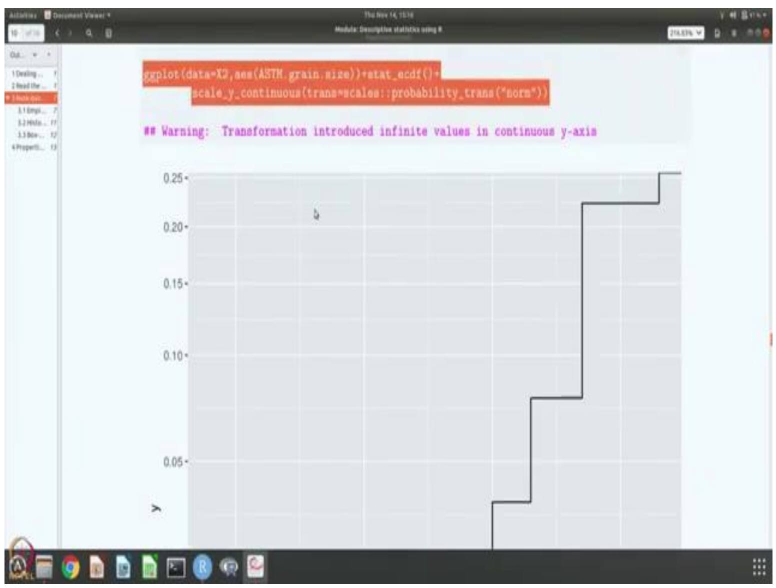
So, we do not expect this curves to turn out to be a straight line. We are just confirming and so you can see, it is not clear here that these two figures are one on top of each other, but you will see it here.

(Refer Slide Time: 20:51)









So, in both the cases, so this is also not a straight line, sort of some amount of deviation. And here it is very clearly seen that it is not a straight line at all and you can see that this does not go up to it is only point 4 here and it is only point 25 here, the scale is not going up to 1, as you would have seen in other cases.

That is because the data is not symmetric about the mean. It is only 1 side that you have the tail. So, that is what is seen in this also. So we will come back and we will do more of the other analysis, histograms and box plots, etc for the same data set, thank you.