Dealing with Materials Data: Collection, Analysis and Interpretation

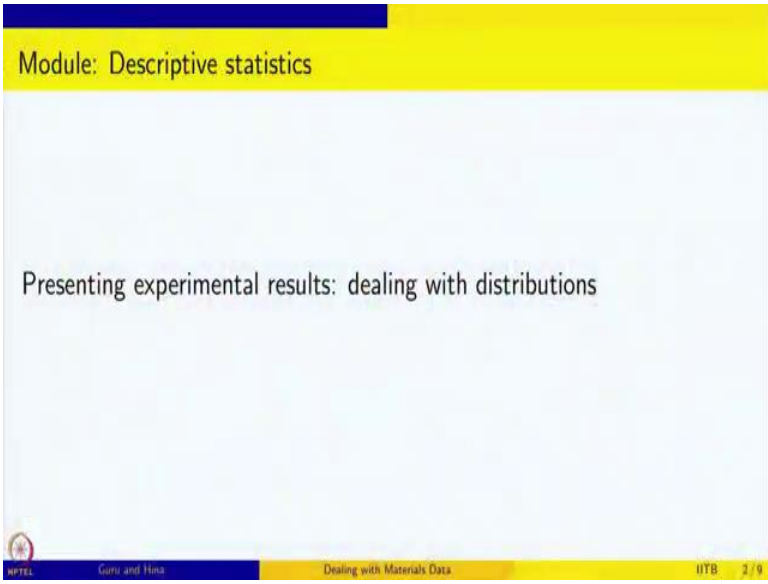
Professor M. P. Gururanjan
Professor Hina A Gokhale
Department of Metallurgical Engineering and Materials Science
Indian Institute of Technology, Bombay

Lecture No 22

Dealing with distributions: Grain size data

Welcome to this course on Dealing with Materials Data, we are going to look at the collection analysis and interpretation of materials data. We have already done one module on introduction to R, and we are learning now how to use R to do descriptive statistics.

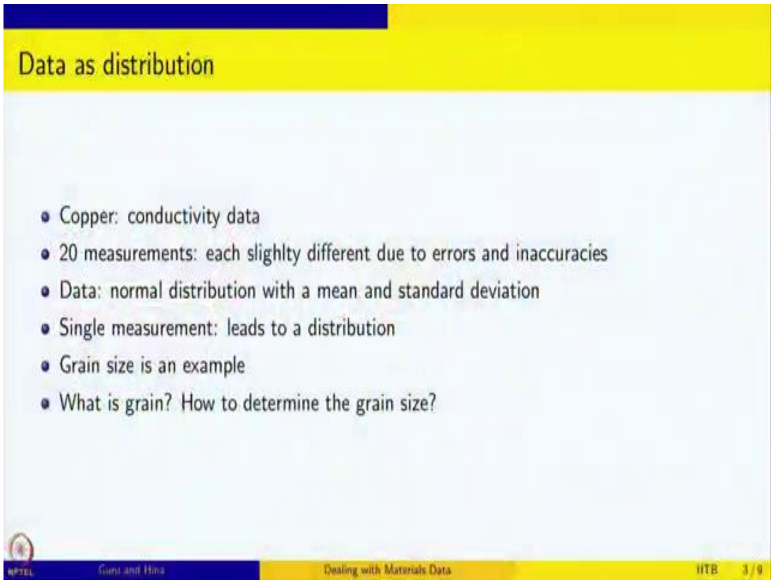
(Refer Slide Time: 0:35)



And in this we are going to look at how to present experimental results and we have already seen how to present experimental results taking the conductivity of ETP copper as an example. So, there were 20 measurements, and we presented those measurements in many different ways and we have also found, those are rank based reports like histograms and dot charts and things like that.

And then we also made the summary of the data. We have also prepared summary based reports like mean and the standard deviation and variance and quantities like that, quantiles and quantities like that. So, in this session, we are going to look at another very common data that you would see in material science and engineering. Which has a slightly different character compared to the previous data that we looked at, namely the conductivity of ETP copper.

(Refer Slide Time: 1:32)

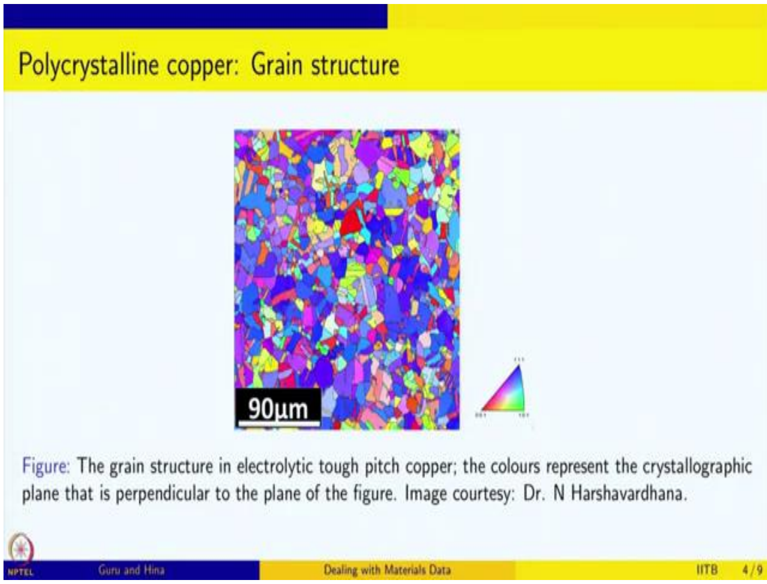


So, let us take a look at that. And when we looked at copper conductivity data, there were about 20 measurements, and each was slightly different due to errors and inaccuracies and the data itself was a normal distribution and the mean was the value that we reported as the conductivity and standard deviation said how much spread is there about this mean in the data.

When you do experiments, if you repeat the experiments, what are the values and how much away are they from the mean. So, this, these are the two quantities that completely describe the data. So, we reported the conductivity itself as mean 101.3 plus or minus standard deviation. It was either point 1 or 0.1 percent in this case both the relative error and absolute terror happened to be same numerically, but you can report it in either ways.

However, sometimes a single measurement leads to a distribution of values. Grain size is an example. So, we will look at what is a grain and how do we determine grain size? So, that is what we will look at now.

(Refer Slide Time: 2:46)



This shows the example of polycrystalline copper and the colors represent different grains. So, this is basically the grain structure in polycrystalline copper. This is the same electronic tough pitch copper on which the conductivity measurements have been made by Doctor Harshvardhana. This is actually taken from his thesis.

And there is a color triangle here and this triangle tells you that if you see a grain which is colored red for example, that means that perpendicular to the plane of this screen, this grain has {001} family of planes the normal is (001). Similarly, if you see any blue colored the grain that means, that that grain is oriented in such a way that the in the plane of this screen perpendicular to that is the normal of (111).

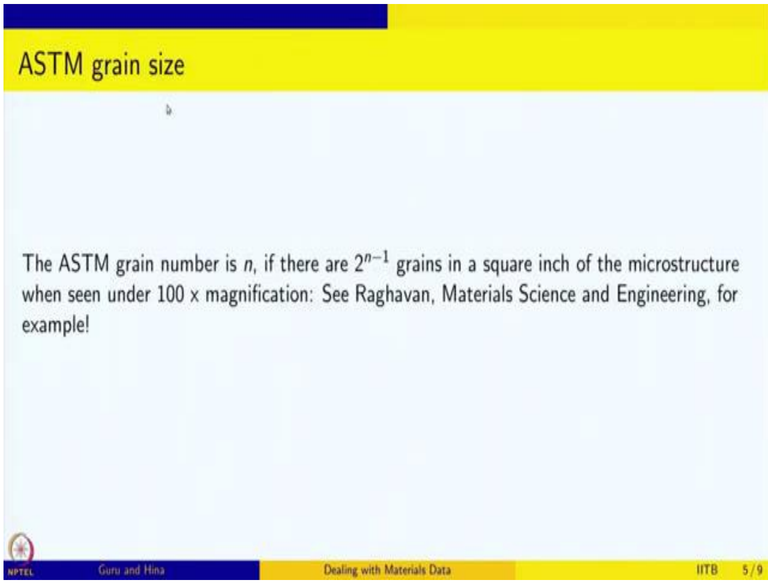
So, this is basically 111 plane and the normal leaves the normal to the plane and so, that is what is colored as blue, so, near about these, then the color slightly changes. So, anything that is bluish basically means that the normal perpendicular to the plane of this figure is given by that family of planes and so this represents the grain structure. And this is for 101 family we use the green. So, it is you can see that it is mostly blue and some red and some green, yellow and things like that.

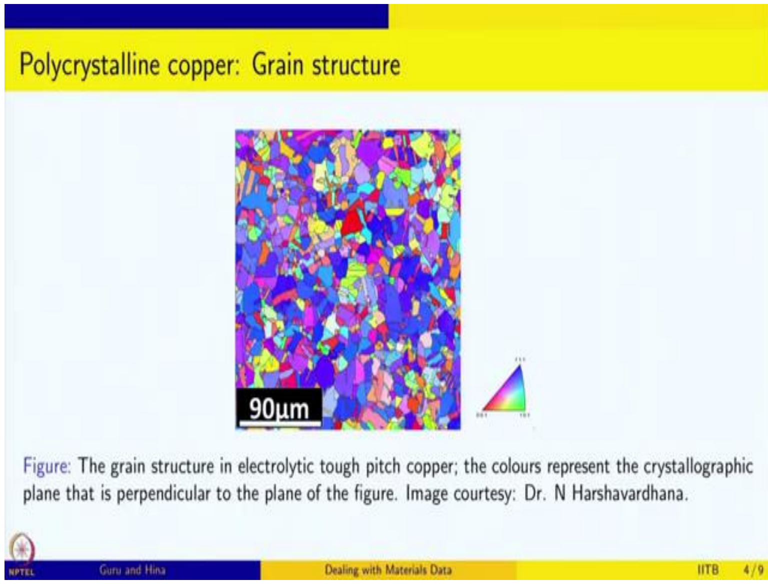
So, this is the grain structure. Now, in order to say what is the size of grains in a material, so you can see this is one single measurement, it is one single micrograph, and it gives you the grain's shape, size and distribution and if you look at the different grains, they all have different sizes. For example, there are some intermediate sized ones, there are big ones, and there are very small ones.

For example, this is very small. So, you can see that there is a variety of grain sizes that are coming out of one single measurement and this is very, very common. So, most of the materials are polycrystalline metallic materials and alloys are polycrystalline and one typically measures the grain size and as you can see, it is not sufficient to give a number and a standard deviation as in the case of conductivity, because in the case of conductivity most of the values were lying slightly away from the mean and that deviation was because of random errors or uncertainties in our measurement, but that is not the case here, the grain size itself is distributed.

And so we need to give this information so sometimes just giving the mean and standard deviation might not be sufficient or might not represent the true nature of what you are measuring or observing.

(Refer Slide Time: 5:45)



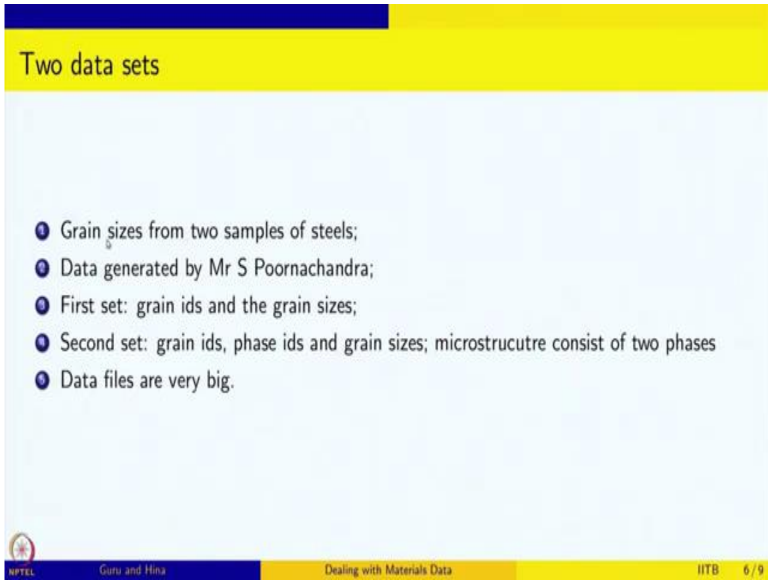


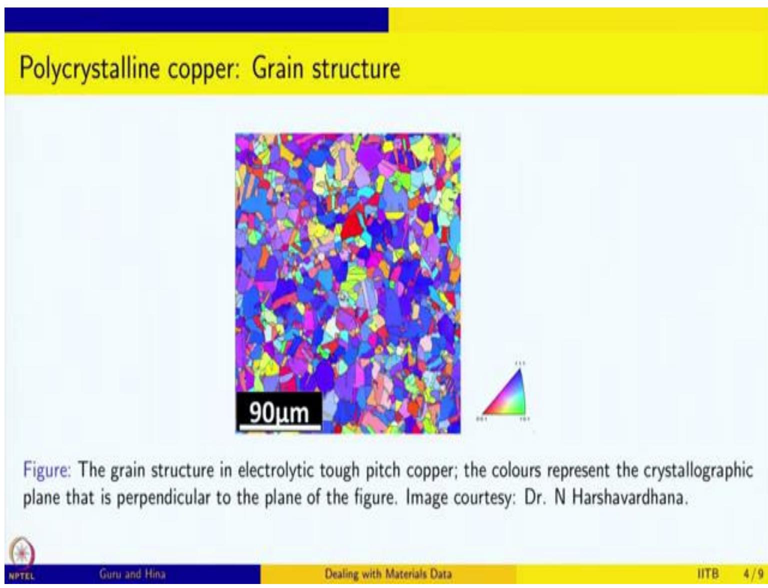
So, there is this concept called ASTM grain size and it is basically a number indicating what is the grain size in a material and it is defined as follows. So, you take the microstructure under 100 magnification and in that 100 times magnification microstructure, you take 1 square inch, count the number of grains and ASTM grain number is the n if there are 2 to the power n minus 1 grains in that square.

So, we basically take a micrograph and we make sure that the magnification is 100x, and then we take 1 square inch of that and count the number of grains and based on that, then we give a number and this is called ASTM grain number for that microstructure or for that material.

And this is described in detail in Raghavan’s book on “Material Science and Engineering” for example. So, so, we have this grain structure and we have the grain size measured by the ASTM grain number.

(Refer Slide Time: 6:49)





And I am going to show two data sets. Both datasets give grain sizes in some steels, and these are two different samples of steels. And this data is generated by Mr. S Poornachandra, who is a PhD student at IIT Bombay and he has given me this data set and the first set has grain ids and grain sizes, I will show you the data set itself, we will open it in Libre Office and see and the second set is slightly more involved because the second set is for a steel which consists of two phases.

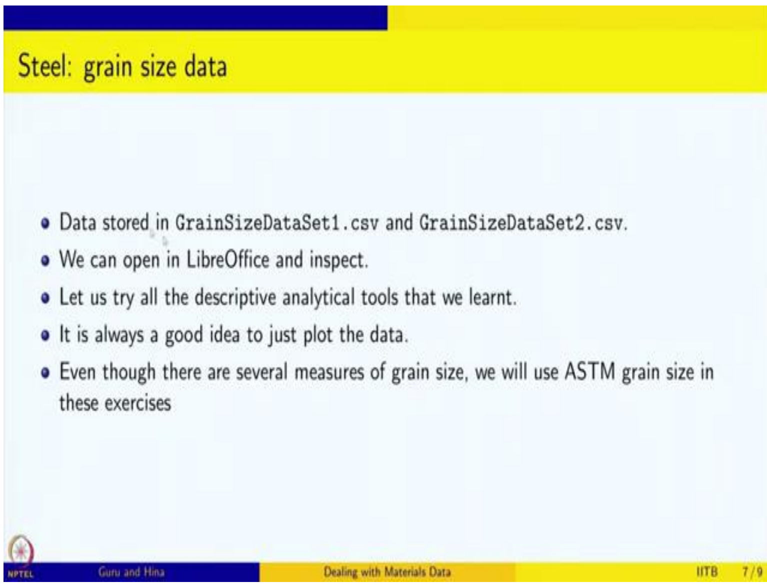
So, in addition to grain id you also have a phase id. So, it either say that this is grain of phase 1 or grain of phase 2 and then it will give the size of that grain, ok. That is because the micro structure consists of two phases.

On the other hand, if you look at copper for example, it is a single phase everything is copper and then we are getting the grains sizes. But, sometimes it can happen that there is more than 1 phase and this is true for most of the alloys that are used in engineering application. Hardly any of them are single phase materials. So, they will always have more than 1 phase.

And the second set is given to, to deal with such scenarios. So, you have in addition two grain ids and grain sizes, also the phase ids. These data files are very big. As you will see, it is no longer practical to enter these numbers by hand. Fortunately for us, these are data files that are generated from the computer.

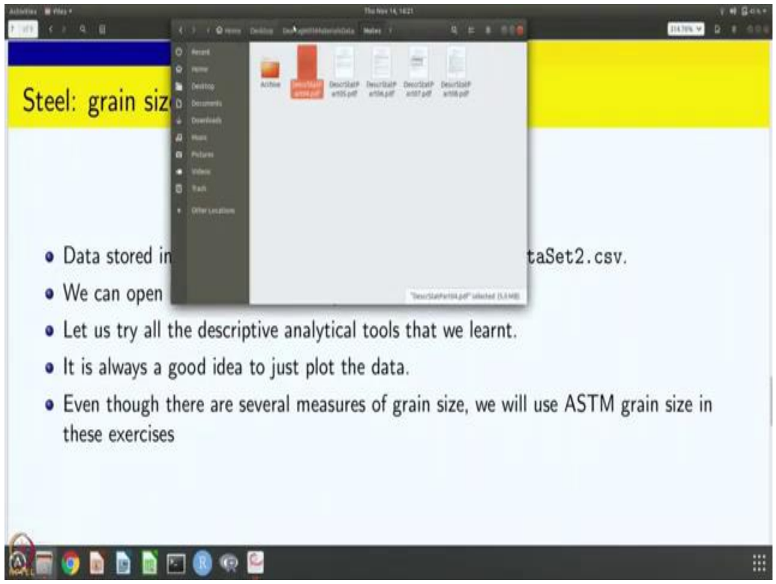
So, you can save them in the CSV format, which is what Mr. S Poornachandra has done and given the data files to us for our study, and we are going to load this CSV data file and we are going to do the analysis on that.

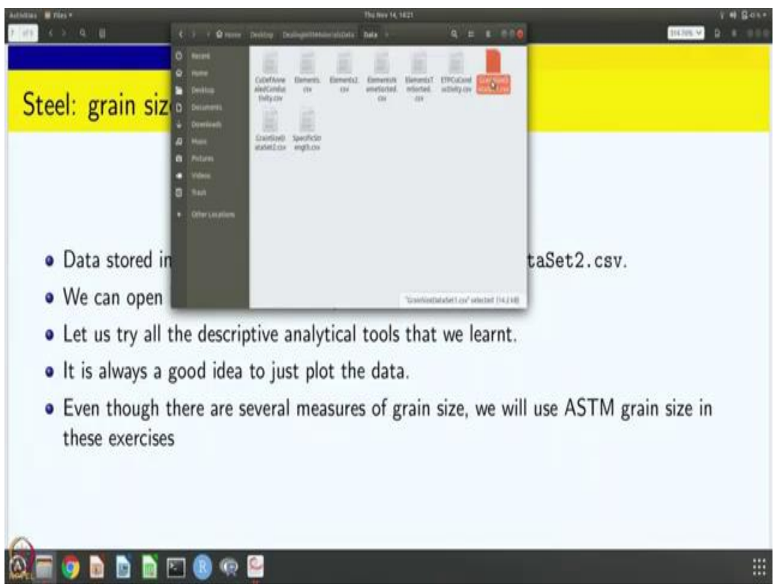
(Refer Slide Time: 8:45)



So, the two data sets, one is called grainsizedataset1.csv and the other one is called grainsizedataset2.csv. So, as you might have noticed, we want to give names as much as possible which are intuitive and easy to understand and clear to follow. So, we are willing to open these files in Libre Office and inspect so let us do that.

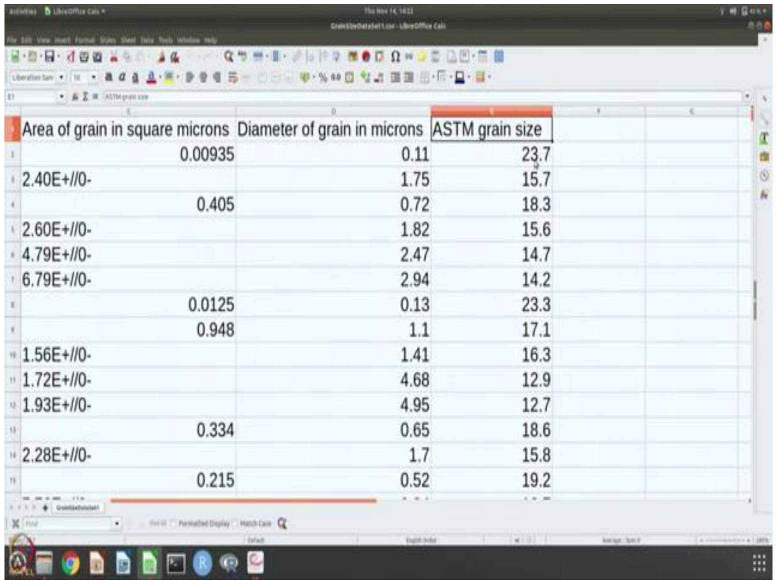
(Refer Slide Time: 9:11)


Steel: grain siz
Data stored i
taSet2.csv.
We can open
Let us try all the descriptive analytical tools that we learnt.
It is always a good idea to just plot the data.
Even though there are several measures of grain size, we will use ASTM grain size in these exercises


Steel: grain siz
Data stored i
taSet2.csv.
We can open
Let us try all the descriptive analytical tools that we learnt.
It is always a good idea to just plot the data.
Even though there are several measures of grain size, we will use ASTM grain size in these exercises

Integer identifying grain	Number of measurement points in the grain	Area of grain in square microns	Diameter
1	3	0.00935	
2	769	2.40E+//0-	
3	130	0.405	
4	833	2.60E+//0-	
5	1535	4.79E+//0-	
6	2178	6.79E+//0-	
7	4	0.0125	
8	304	0.948	
9	500	1.56E+//0-	
10	5520	1.72E+//0-	
11	6182	1.93E+//0-	
12	107	0.334	
13	730	2.28E+//0-	
14	69	0.215	

Number of measurement points in the grain	Area of grain in square microns	Diameter of grain in microns	AST
3	0.00935	0.11	
769	2.40E+//0-	1.75	
130	0.405	0.72	
833	2.60E+//0-	1.82	
1535	4.79E+//0-	2.47	
2178	6.79E+//0-	2.94	
4	0.0125	0.13	
304	0.948	1.1	
500	1.56E+//0-	1.41	
5520	1.72E+//0-	4.68	
6182	1.93E+//0-	4.95	
107	0.334	0.65	
730	2.28E+//0-	1.7	
69	0.215	0.52	



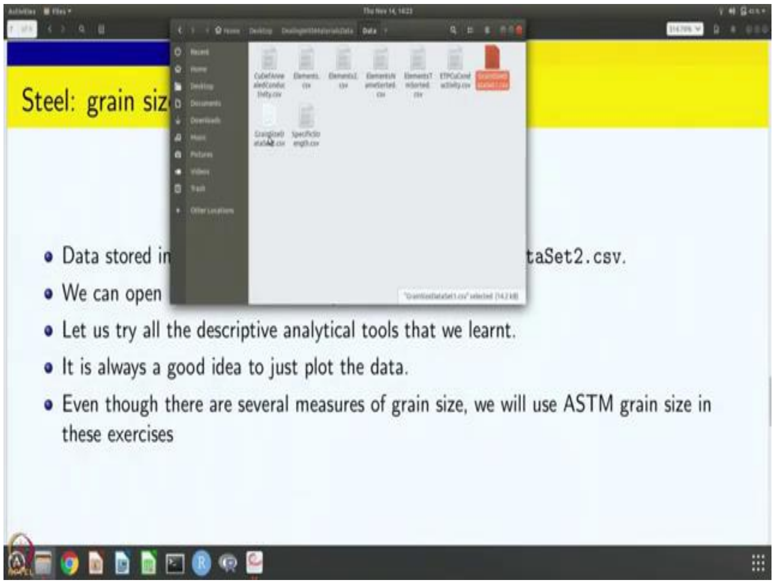
Area of grain in square microns	Diameter of grain in microns	ASTM grain size
0.00935	0.11	23.7
2.40E+//0-	1.75	15.7
0.405	0.72	18.3
2.60E+//0-	1.82	15.6
4.79E+//0-	2.47	14.7
6.79E+//0-	2.94	14.2
0.0125	0.13	23.3
0.948	1.1	17.1
1.56E+//0-	1.41	16.3
1.72E+//0-	4.68	12.9
1.93E+//0-	4.95	12.7
0.334	0.65	18.6
2.28E+//0-	1.7	15.8
0.215	0.52	19.2

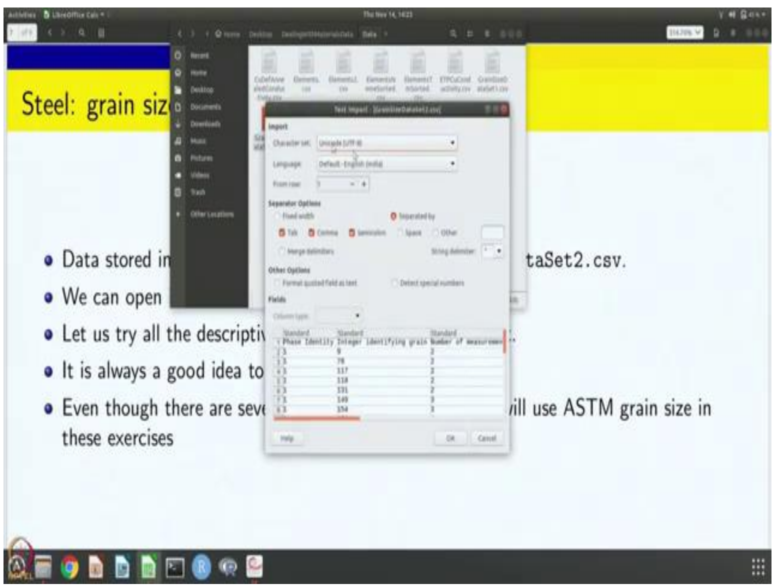
So, let me go to the data and let us open grainsizedataset1.csv, and so there is this column integer identifying grain like 1, 2, 3, 4 etc. and these are the number of measurement points in the grain and the area of grain in square microns is given. So, this is also a measure of the grain size, you can give the area of the grain in, in square microns and then you can also give let us go here, the diameter of the grain in micron.

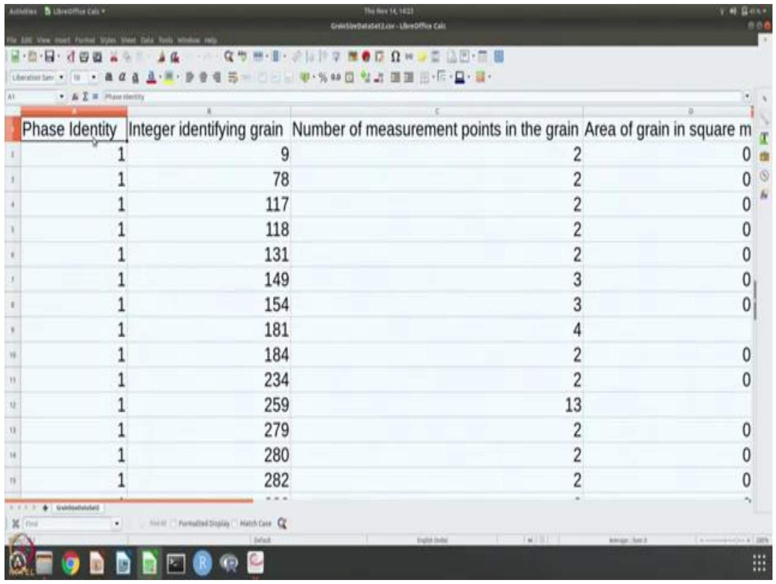
So, obviously, as you are seen the grains are not circular or spherical, but you can get the equivalent circle of this area, what is going to be the diameter of such a circle. So, it is possible to give an equal and diameter for grains, that is one way of defining an equivalent diameter, but that need not be the only way, but this is again another measure this is, this is an area measure, this is a length to measure.

So, we also have the ASTM grain size, which is like a number so, you can also give a number of measures. So, there are three different measures of grain sizes that we see here and one is the area measure, the other one is some length measure, the other one is a number. And for our analysis, we are going to use the ASTM grain size with the integer identifying grain. So, this is one single data set and it already gives you large number of grain sizes and so we are interested in looking at the distribution. So, the data itself is distribution.

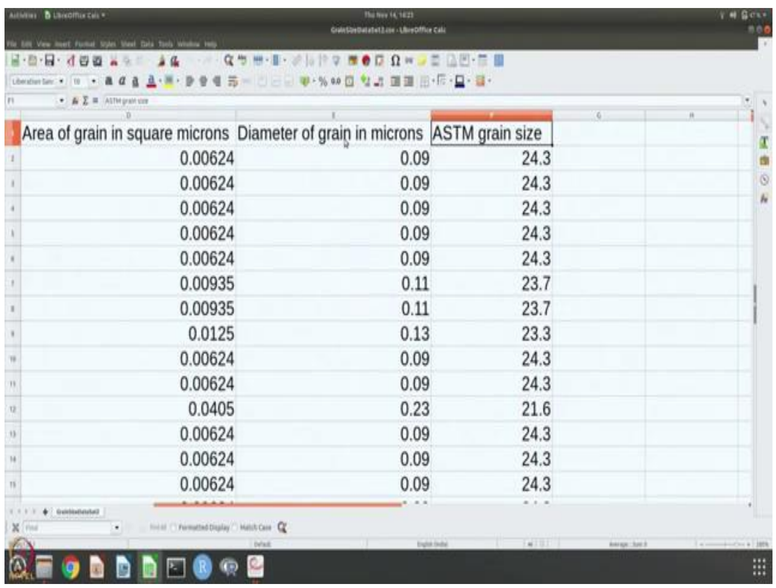
(Refer Slide Time: 10:55)







Phase Identity	Integer identifying grain	Number of measurement points in the grain	Area of grain in square m
1	9	2	0
1	78	2	0
1	117	2	0
1	118	2	0
1	131	2	0
1	149	3	0
1	154	3	0
1	181	4	
1	184	2	0
1	234	2	0
1	259	13	
1	279	2	0
1	280	2	0
1	282	2	0

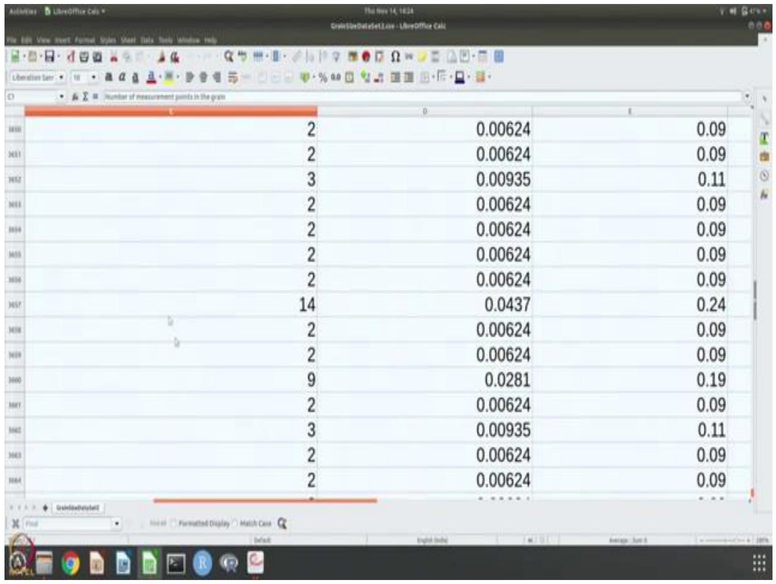


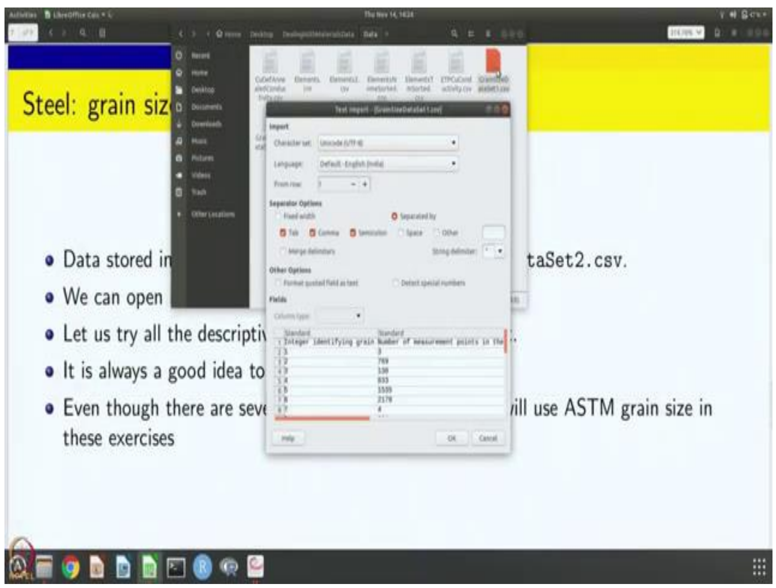
Area of grain in square microns	Diameter of grain in microns	ASTM grain size
0.00624	0.09	24.3
0.00624	0.09	24.3
0.00624	0.09	24.3
0.00624	0.09	24.3
0.00624	0.09	24.3
0.00935	0.11	23.7
0.00935	0.11	23.7
0.0125	0.13	23.3
0.00624	0.09	24.3
0.00624	0.09	24.3
0.0405	0.23	21.6
0.00624	0.09	24.3
0.00624	0.09	24.3
0.00624	0.09	24.3

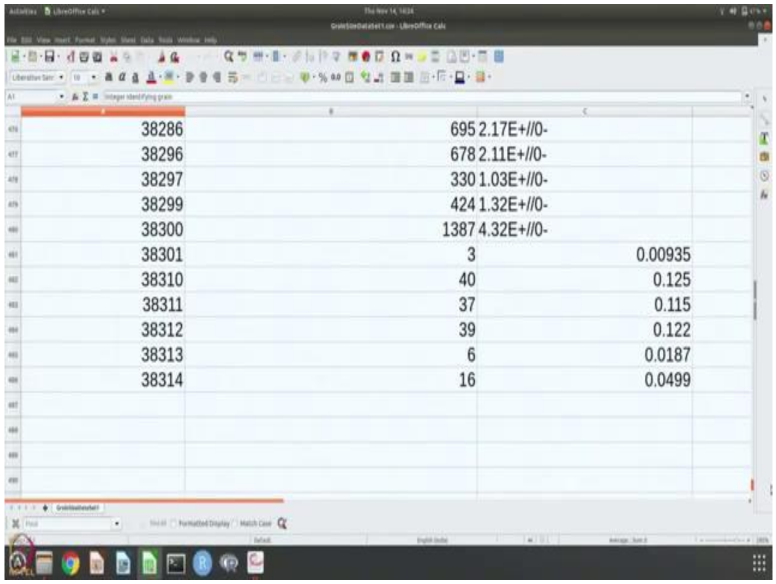
If you look at the second data set, it is very similar to the first 1, except that now there is an extra column which is called phase identity. So, it again after giving the phase identity, then it gives integer, which identifies the grain and the number of measurement points in the grain and the area of the grain and diameter of the grain and the ASTM number.

By the way, the number of measurement points in the grain should also be proportional to the size of the grain because if you are taking measurements at periodic distances, then if you have larger area, you will have more measurements. So, this is also at some level another measure of the size of the grain.

(Refer Slide Time: 11:49)



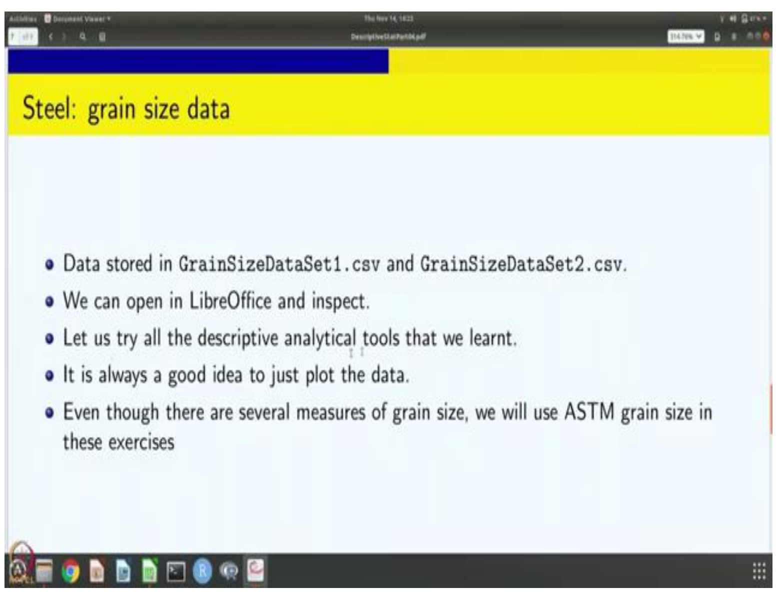




But as you can see, these data files are too big. So, for example, the grain size too, if you go down and you can see somewhere of the order of 3600 data points are there. And similarly, the first a data set that we had that grain size 1, it is not that big, but it is still reasonably big and so I think this has about 480 or 500 data points. So, so we have about 486 data points.

So, obviously, generating such a data file by putting data, data by hand into R is not practical and it is also not meaningful because manual entry can introduce its own errors. So, this data comes from the computer and it is stored as csv so that we can import this data into R and start working with it.

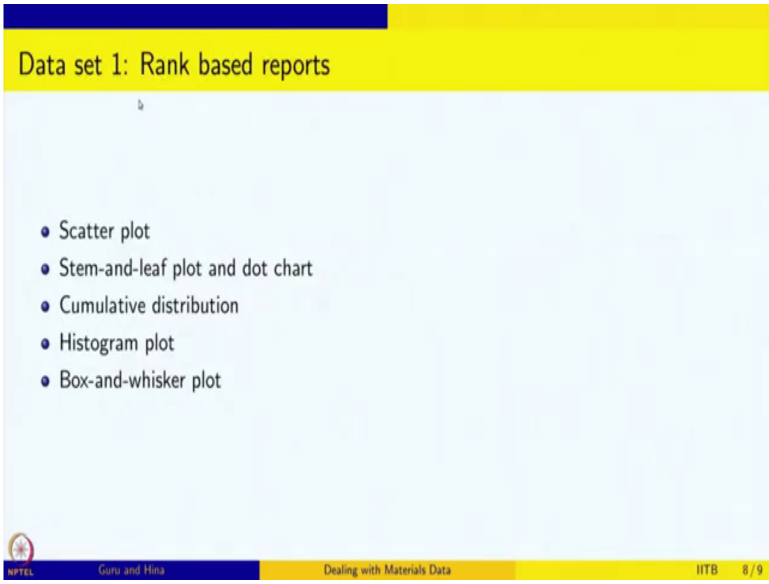
(Refer Slide Time: 12:47)



So, we are going to try all the descriptive analytical tools that we learned while dealing with connectivity. In looking at this grain size data for both the sets. It is always a good idea to just plot the data to have an idea of what the data looks like, of course, you can open in Libre Office and inspect, but that is still very cursory and you can try to get a overall picture of the data just by trying to plot this data. So, we are going to do that also and we are going to mostly use ASTM grain size for our exercises.

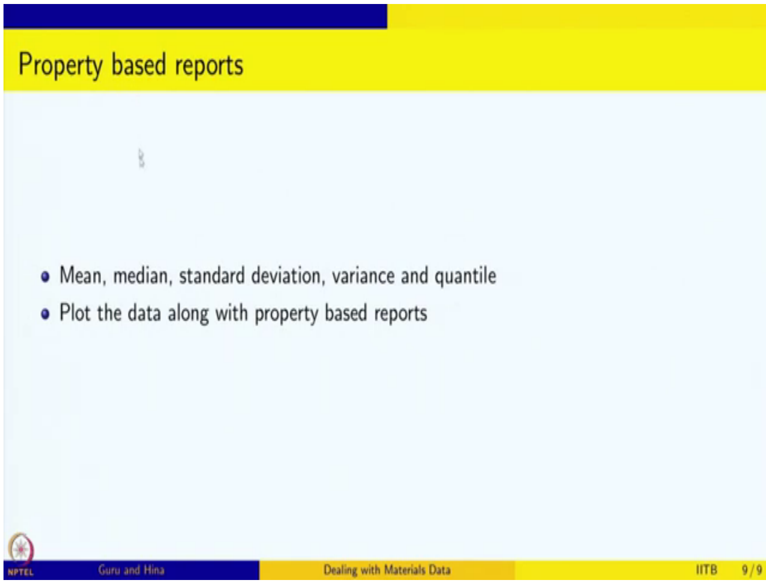
So, any of the measure of grain size can be used, but we are going to stick to ASTM grain size for this, for this session, ok. So, let us go.

(Refer Slide Time: 13:40)



For the data set 1, let us do this rank based reports. We have learnt about several ranks based reports, scatter plot, stem-and-leaf plot, dot chart, cumulative distribution, histogram plot and box-and-whisker plot. So, we are able to do all these rank based reports.

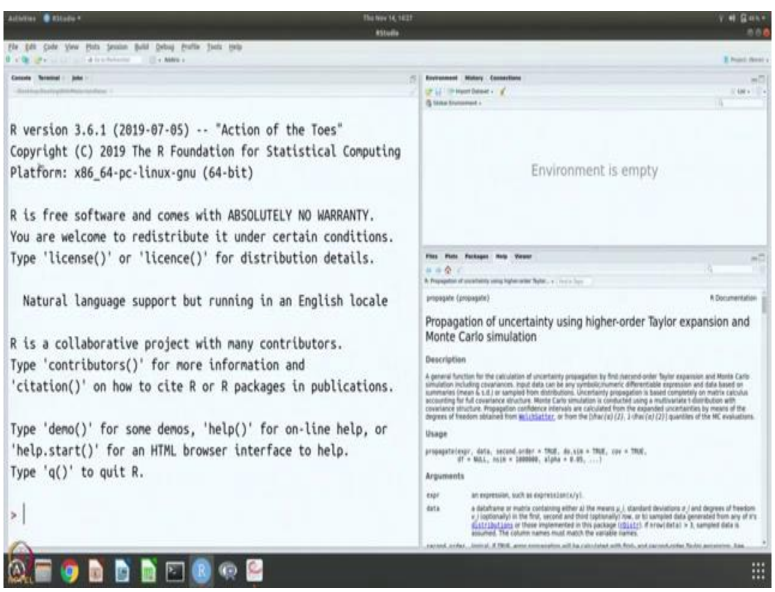
(Refer Slide Time: 13:59)

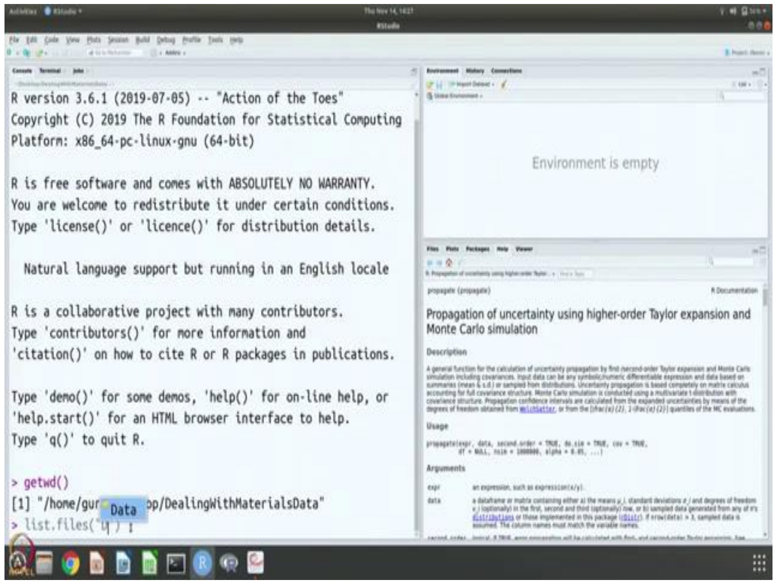


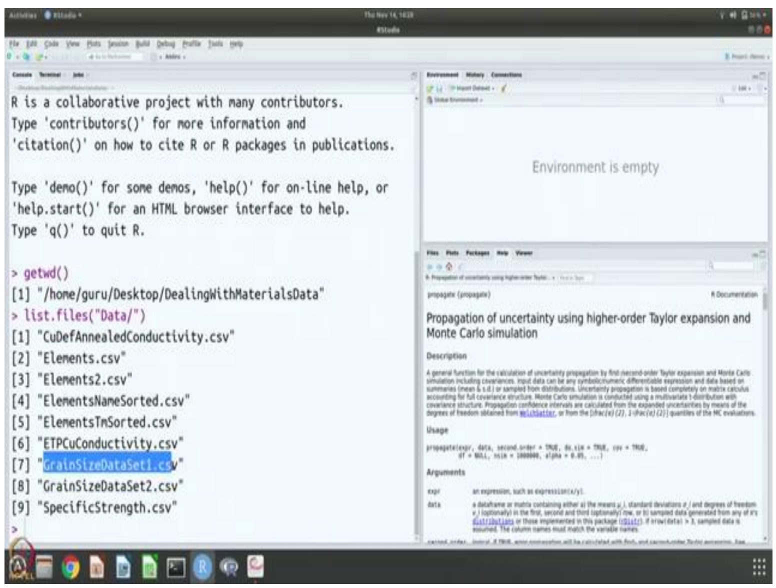
And for the data set 1 we are also going to do the property based reports. Those are mean, median, standard deviation, variance in quantile. And of course, finally, we are going to plot the data and we are going to indicate this property based values on the plot to have a better understanding of the data.

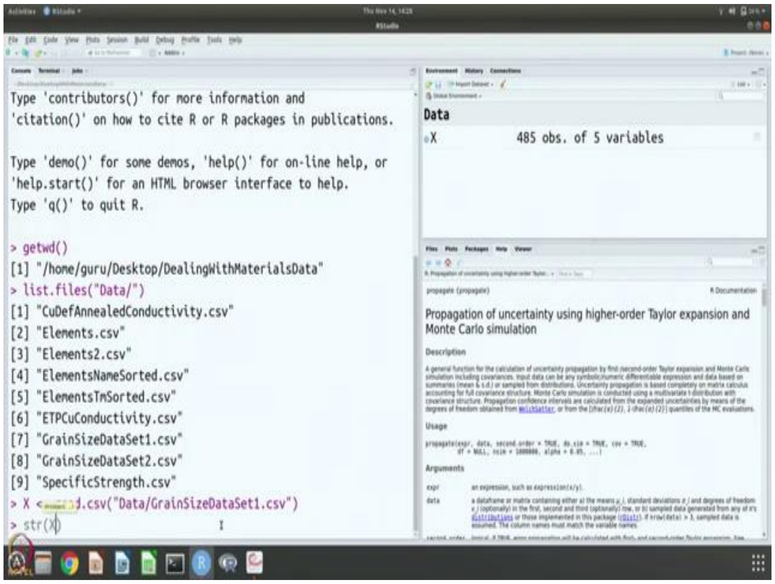
So, that is what we are going to do in this session. So, for data set 1 and we will come back to data set 2 in the next one, ok.

(Refer Slide Time: 14:28)




R version 3.6.1 (2019-07-05) -- "Action of the Toes"
Copyright (C) 2019 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
Natural language support but running in an English locale
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
> getwd()
[1] "/home/gur
p/DealingWithMaterialsData"
Data
> list.files("
Environment is empty
Propagation of uncertainty using higher-order Taylor expansion and Monte Carlo simulation
Description
Usage
Arguments


R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
> getwd()
[1] "/home/guru/Desktop/DealingWithMaterialsData"
> list.files("Data/")
[1] "CuDefAnnealedConductivity.csv"
[2] "Elements.csv"
[3] "Elements2.csv"
[4] "ElementsNameSorted.csv"
[5] "ElementsTmSorted.csv"
[6] "ETPCuConductivity.csv"
[7] "GrainSizeDataSet1.csv"
[8] "GrainSizeDataSet2.csv"
[9] "SpecificStrength.csv"
>
Environment is empty
Propagation of uncertainty using higher-order Taylor expansion and Monte Carlo simulation
Description
Usage
Arguments



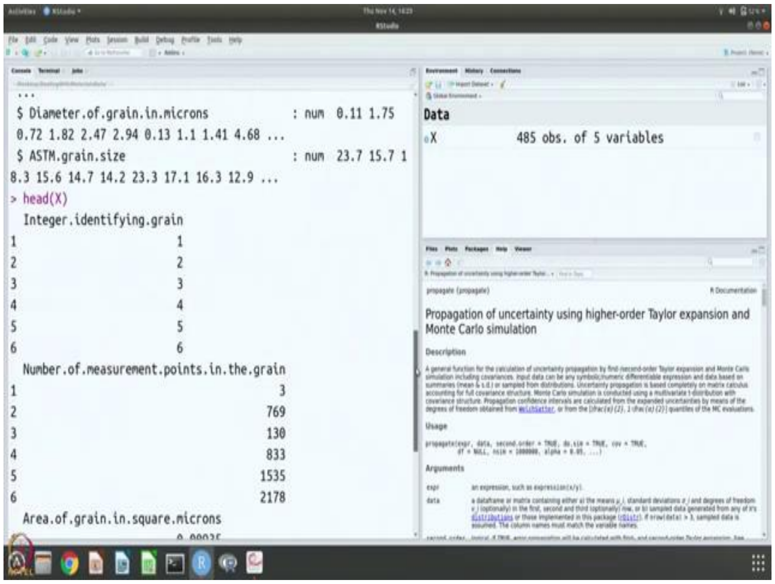
So, as usual to do the data analysis, we have to open R we have to look at the R version it is 3.6.1 "Action of the Toes' and we have to find out which directory we are that is by using get working directory. So, we are in the dealing with materials data directory. So, we are going to load the data and so to load the data we need to know which are the data files so let us just look at the files in the data directory.

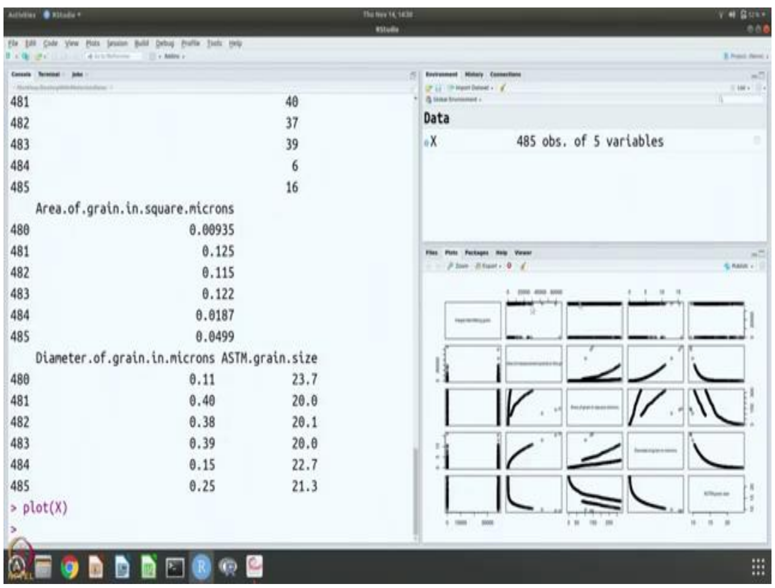
So, if I say data, so there are all these files and we are interested in grainsizedataset1.csv and grainsizedataset2.csv. So, first we are going to deal with grainsizedataset1.csv. So, let us do that. So, let us load the data into the variable x. and importing is done by using read csv because it is a csv file and we are going to say from the data directory, and it is grainsizedataset1dotcsv.

So, let us read it, ok. So, as you can see immediately R tells you that there are 485 observations and 5 variables which we have already opened in Libre Office and saw so let us get some more information on the X so it is a data frame. It has 485 observations and 5 variables and those 5 variables are listed here, integer identifying grain, number of measurement points in the green, area of grain in square microns, diameter of grain in microns and ASTM grain size.

And you can see that integer is for example int, it goes as 1 2 3 4 etc. and the number of measurement points is 3 769 130 8 etc. area of grain has some 321 levels 0.00624, 0.00935, etc. and diameter of the grain again is a number it is 0.11 1.75 etc. and as you can see 0.11 and 1.75. So, number of measurement points is 3 and 769. So, that is also sort of consistent with what one would expect and the ASTM grain size is given 23.7, 15.7 etc., ok.

(Refer Slide Time: 16:59)



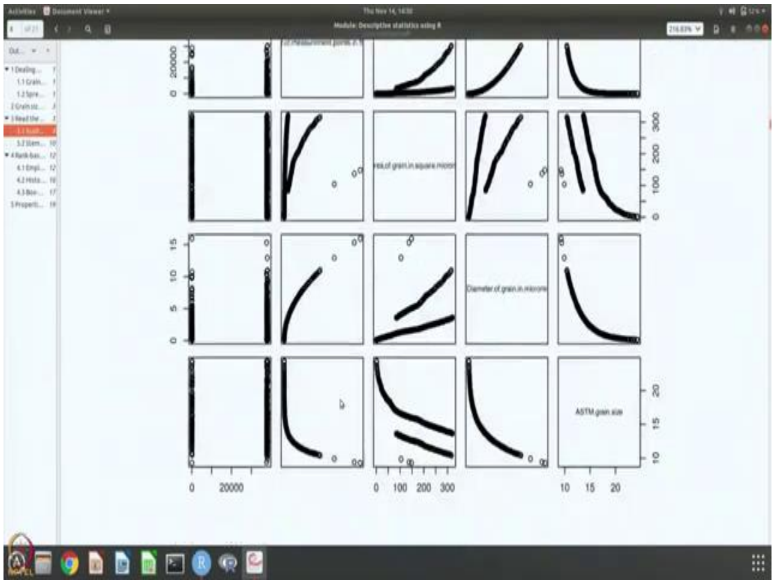


So, the easiest thing to do is to just you can you know without opening even in Libre Office so you can say head x for example. So, it will give you the first few lines, 5, 6 lines. So, you can do and you can also do the similar command till to look at the last few lines. So, this is another way of taking a look at the data, but this is not the complete data.

So, one of the easiest ways to get the complete data is to plot x, like I told you last time, when you say just plot x for a data frame it makes a table of plots. So, it takes each of these variables, there are 5 variables and it plots each against all the other variables. So, 5 into 5 there are 25 boxes that you can see and.

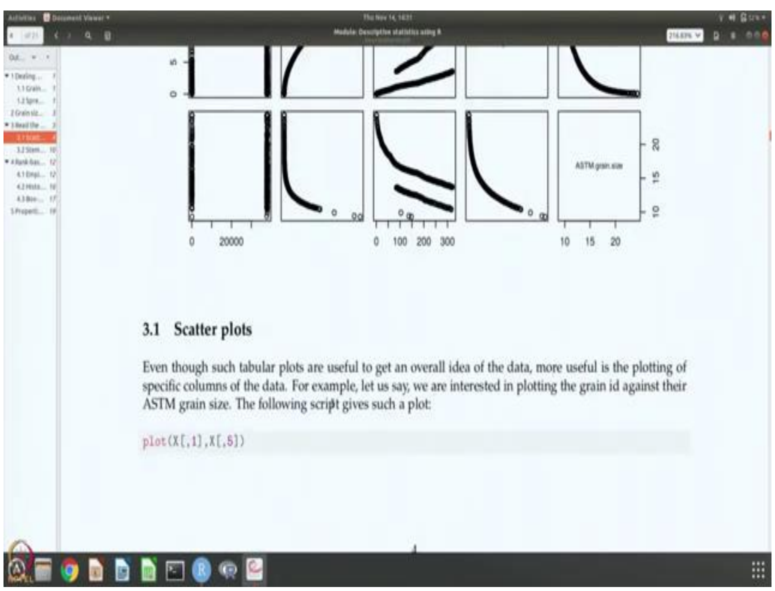
So, you see 4 4 4 4 4. So, 20 plots are there and our interest is with the ASTM grain size. In fact, we are interested in looking at the integer identifying grain and the ASTM grain size, ok. So, that is what we are going to plot and, and see. So, let us do that.

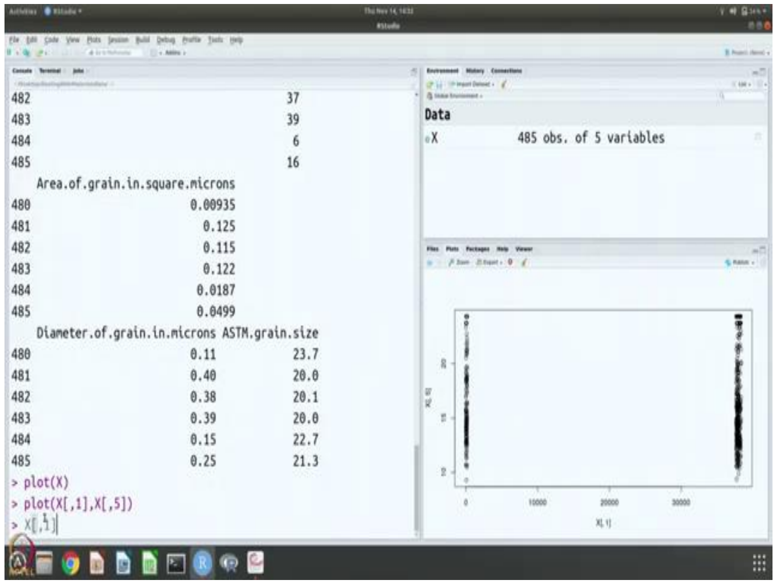
(Refer Slide Time: 18:08)

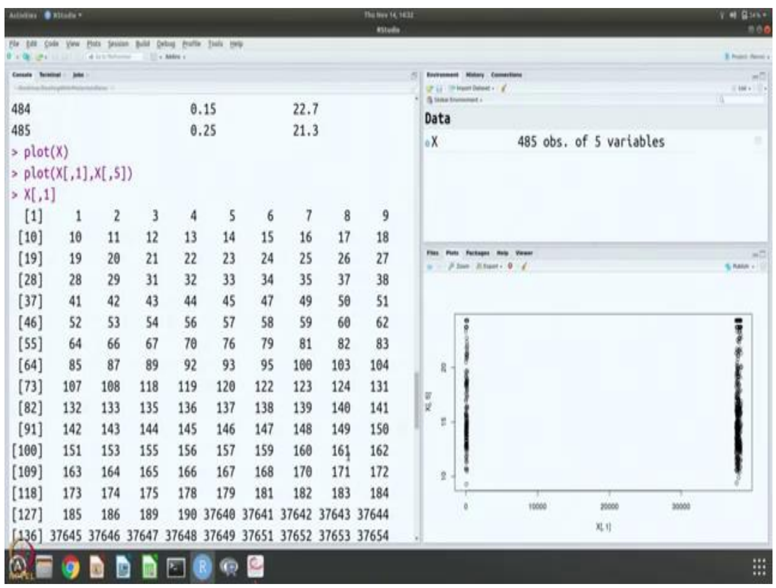


So, let us go and look at so, so this is how the figure looks. So, this is a slightly bigger picture. So, you can clearly see what is there along the diagonals, and how the data plot looks like.

(Refer Slide Time: 18:22)





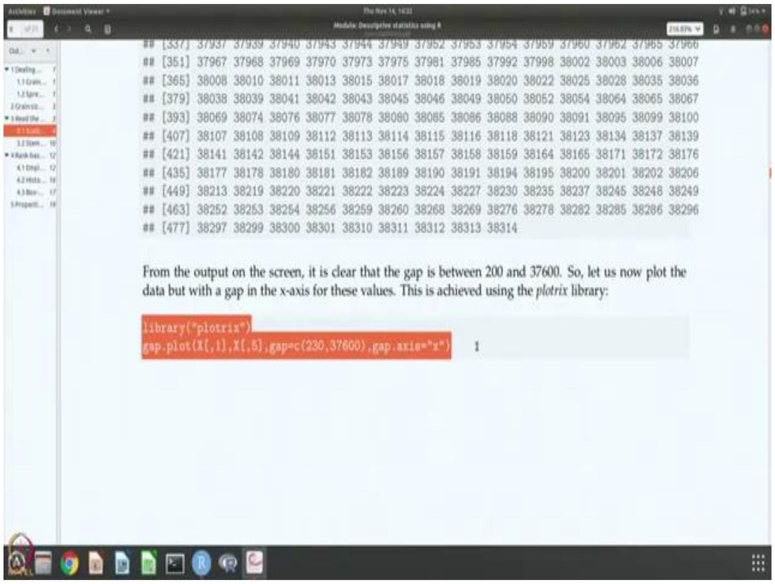


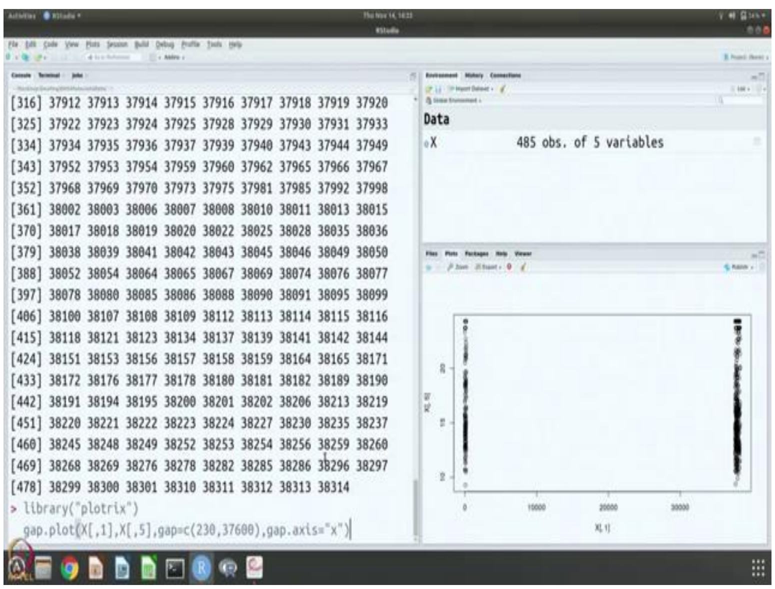
So, the first thing that we want to do is to make us scatterplot. We want to say that, ok, integer identifying grain against the ASTM grain size. So, let us just plot that quantity and see. When I plot that quantity, you see that there are lots of data points, quite close to 0 here and lots of data points are somewhere about 37,000 or 38,000. We know that there are 485 observations, and all of them are clustered in two places, and rest of it in the middle is empty.

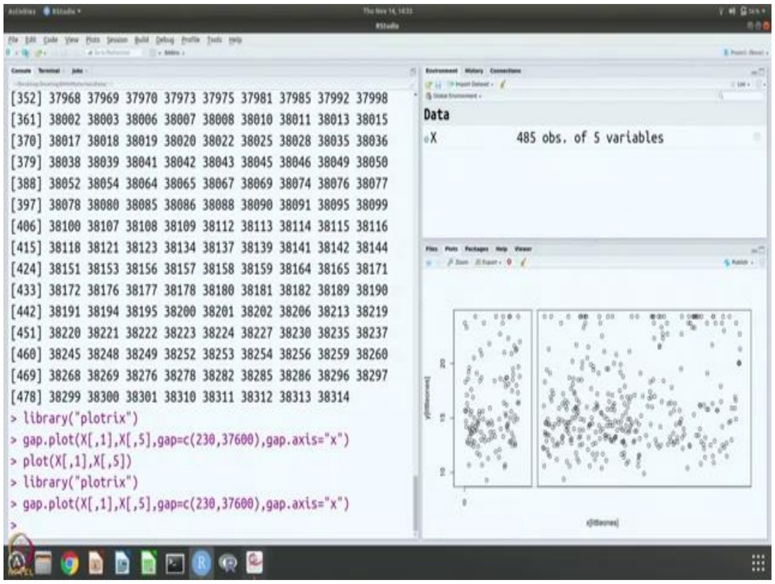
So, this picture is really not very helpful for me to understand how the data looks. So, I want to understand why and because everything is clustered around the, the first variable near about 0 or near about 30,000. Let us just look at what this integer identifying grains looks like. So, if you do that, of course, you can see that the numbers initially start as 1 2 3 4 intuitively, that makes sense and somewhere about 190 suddenly there is a jump to 37,640. And that is why after 190 you do not see anything before 37,600 and (())(19:44).

So, which means that it is not meaningful plot to plot data like this. So, for the scatterplot let us try to remove this gap in the figure and trying to make a scatterplot.

(Refer Slide Time: 20:05)





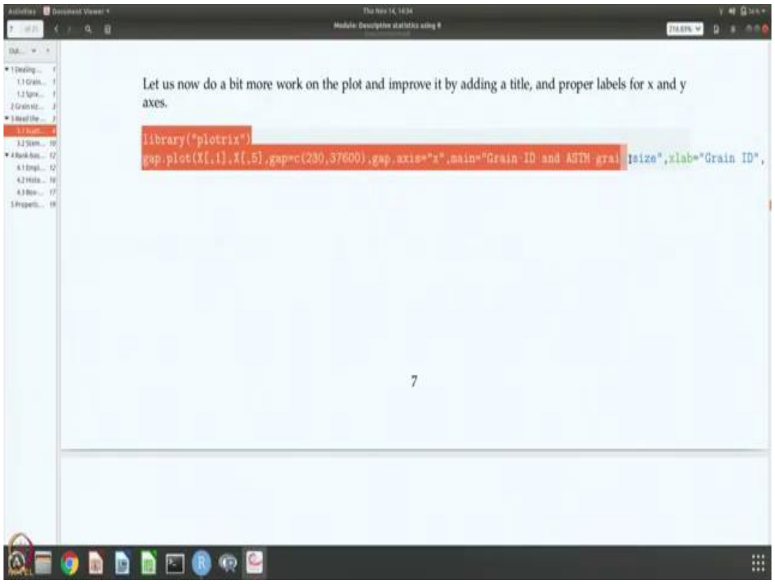


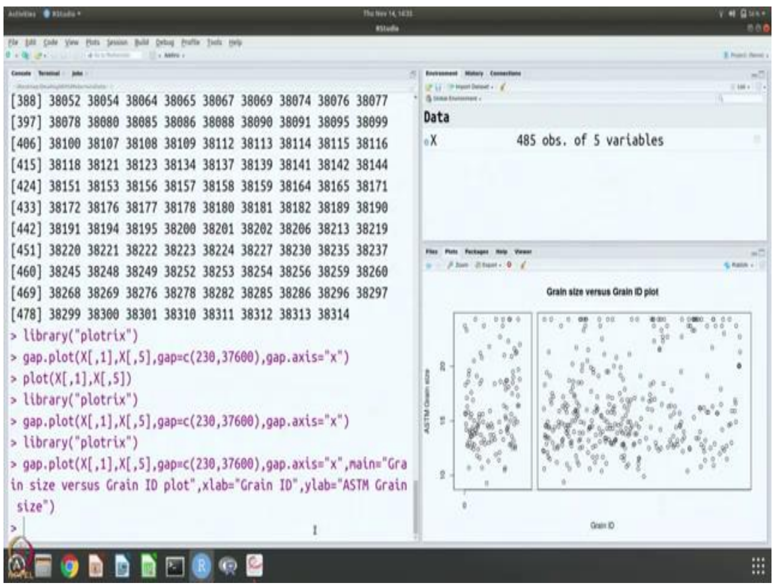
For doing that, we have to use a library and that library is known as plottricks, ok. So, let us take these two commands and let us put it in our. So library, I am going to use the plottricks library and plottricks library allows you to plot a gap plot. So, x1 and x5, that is what we are plotting the first column versus the fifth column.

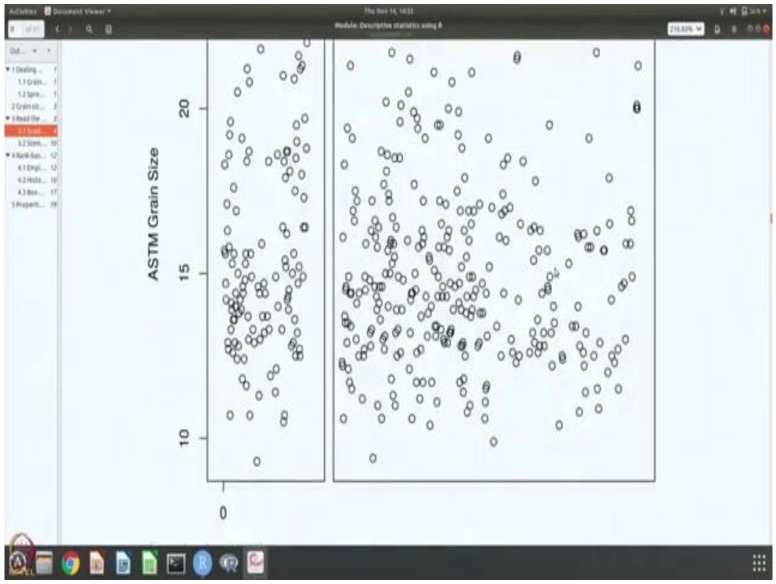
But there is a gap and the gap is I am telling that ok the gap is between to the 230 and 37,600. So, those data points will be left out the gap axis is x, because that is the x axis and I want to leave out in x these points, and for the rest we are going to plot. So, if you do that, of course, now the data is easier to visualize.

But there is a difference. You know, when we plotted see the x was labeled 1000; 0, 10,000 20,000 30,000 etc but when you do the gap plot the tick marks have disappeared so we have to get them back.

(Refer Slide Time: 21:25)



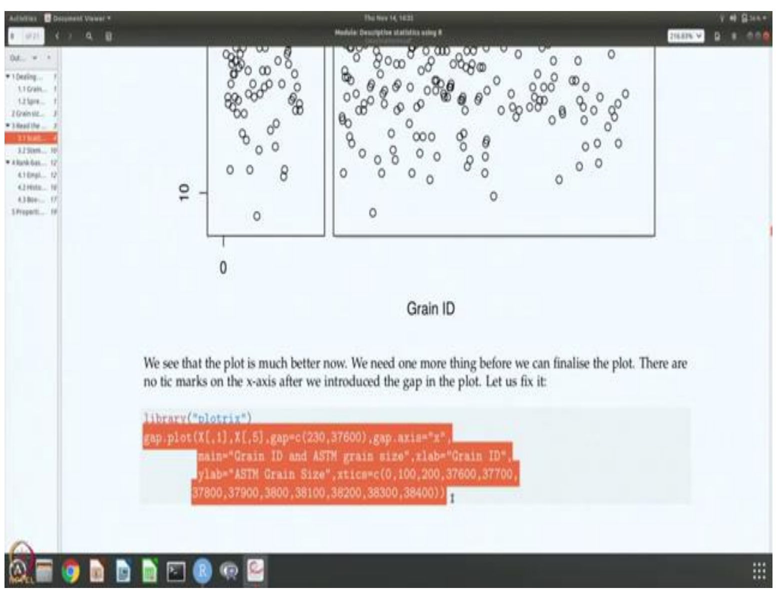


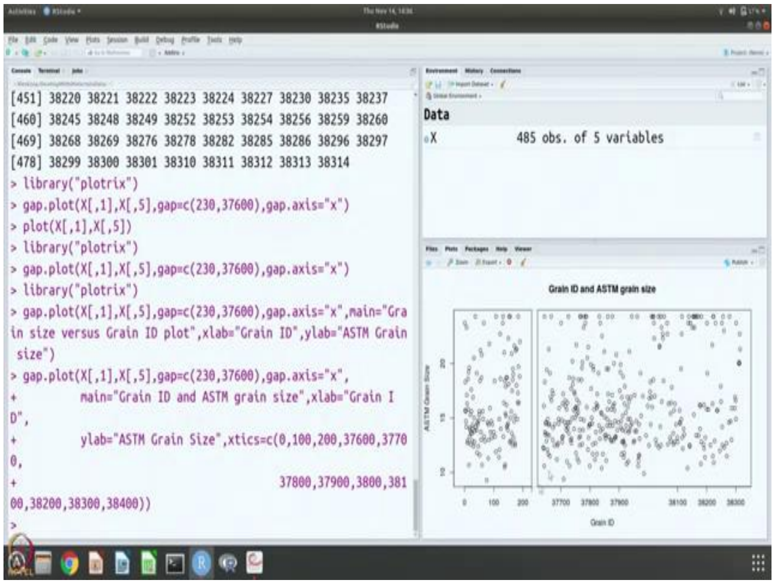


So, we will do that there is a way to get the tick marks, ok. There is also a way to get the other information. So, ok. So, so, let us go back and do this. So, of course you can get the title, this is the so the, the labels also says the x little ones, y little ones, etc.

So, let us change, let us say, let us call this plot as grain size versus grain ID lot and x label should be grain ID and y label should be ASTM grain size. Let us do that. So, we have grain size versus grain ID plot, ASTM grain size versus grain ID so that is what is given and that is what this command is and you can see grain ID versus.

(Refer Slide Time: 22:38)



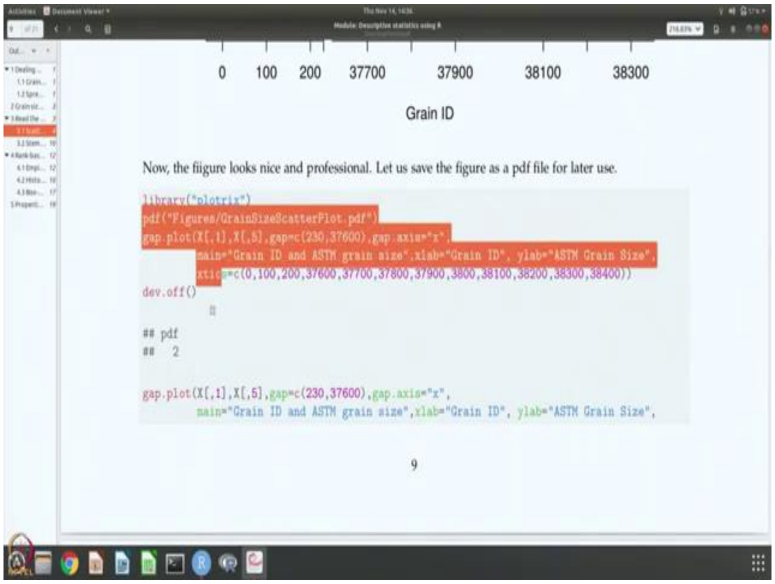


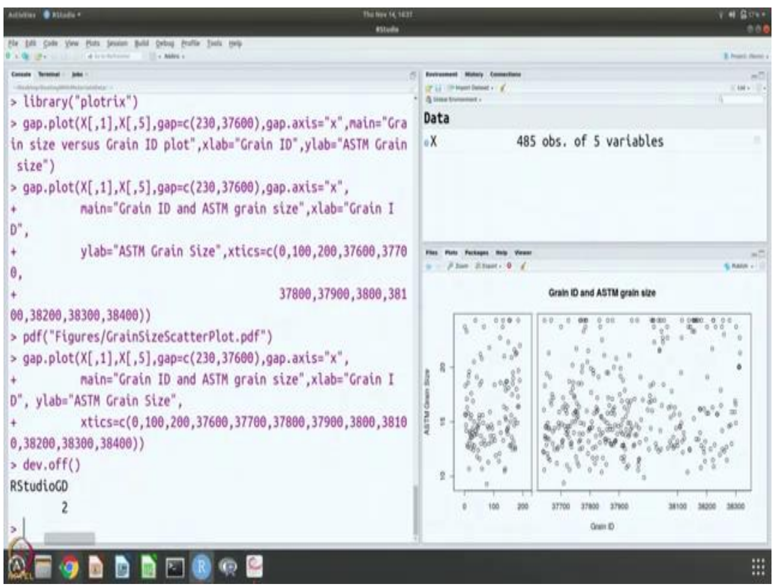
So, now let us also introduce the x tick marks, ok. So, I am going to cut paste this command, ok. So, let us look at this command again it says is gap plot, 1, which is the first column ID grain ID, 5, which is the fifth column which is the ASTM grain size and we are plotting and it is a gap plot.

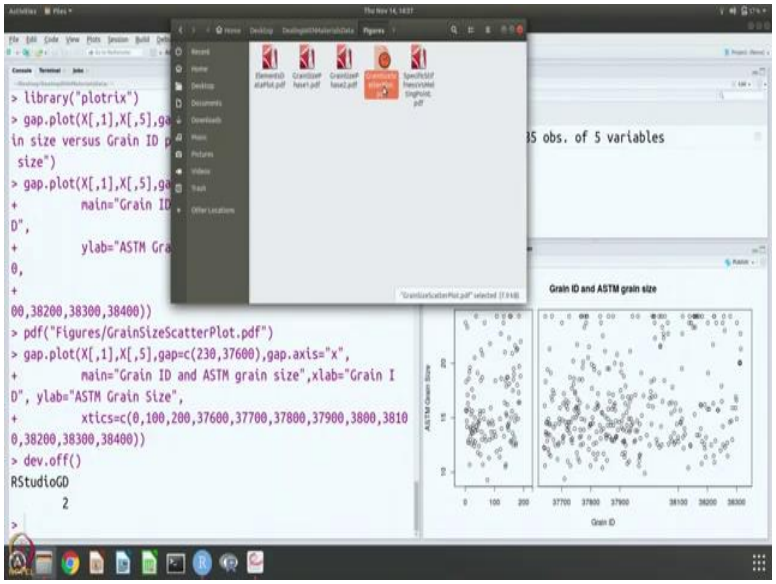
So, we are saying that there is a gap in x axis and the gap is from 230 to 37,600 and so, the, the plot is called grain ID and ASTM grain size plot x label is grain ID, y label is a same grain size, then we are saying introduce the x tick marks and the technique should go from 0, 100, 200 that is this part and then 37,600 onwards up to 38,400 because we can see that the data is up to 38 3 on 4. So, 38,400 should about cover the entire range.

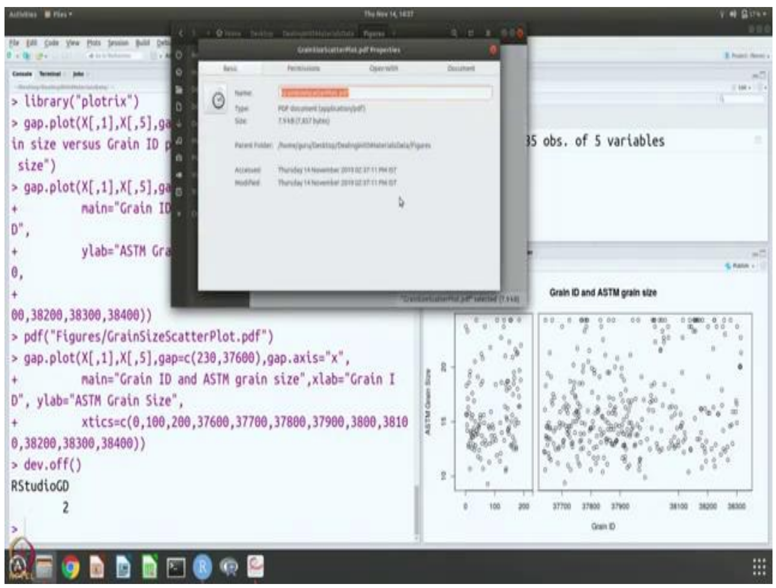
If we do that, of course, we have the complete data now plotted the, and, and figure looks very neat and professional now. So, you can save this figure, ok. So, let us do that.

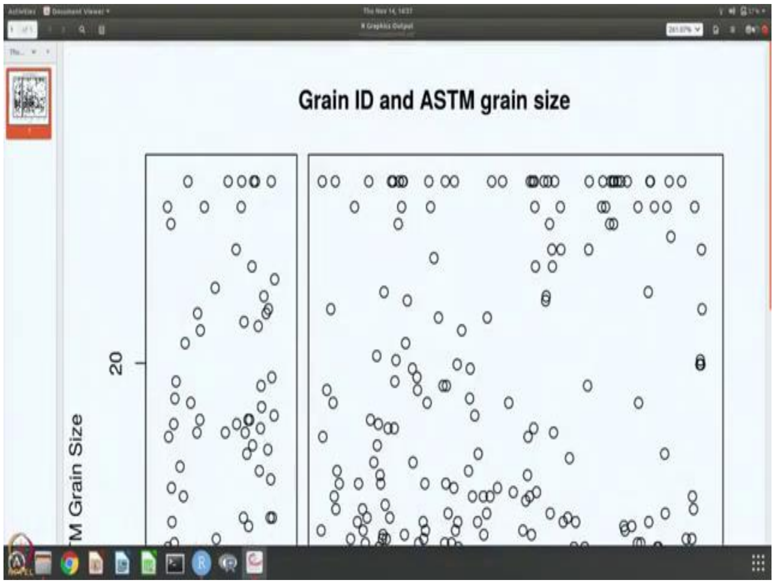
(Refer Slide Time: 23:53)






> library("plotrix")
> gap.plot(X[,1],X[,5],ga
in size versus Grain ID p
size")
> gap.plot(X[,1],X[,5],ga
+ main="Grain ID
D",
+ ylab="ASTM Gra
0,
+
00,38200,38300,38400))
> pdf("Figures/GrainSizeScatterPlot.pdf")
> gap.plot(X[,1],X[,5],gap=c(230,37600),gap.axis="x",
+ main="Grain ID and ASTM grain size",xlab="Grain I
D", ylab="ASTM Grain Size",
+ xtics=c(0,100,200,37600,37700,37800,37900,3800,3810
0,38200,38300,38400))
> dev.off()
RStudioGD
2
>
5 obs. of 5 variables
Grain ID and ASTM grain size
ASTM Grain Size
Grain ID

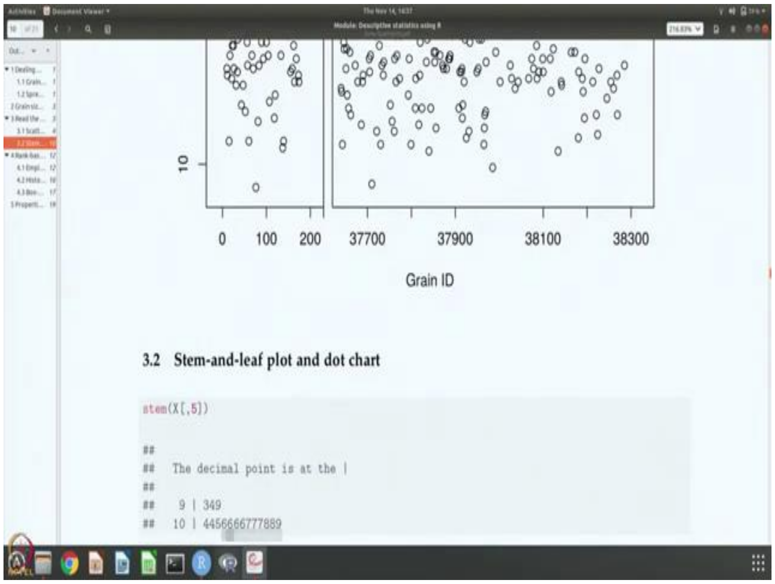

> library("plotrix")
> gap.plot(X[,1],X[,5],ga
in size versus Grain ID p
size")
> gap.plot(X[,1],X[,5],ga
+ main="Grain ID
D",
+ ylab="ASTM Gra
0,
+
00,38200,38300,38400))
> pdf("Figures/GrainSizeScatterPlot.pdf")
> gap.plot(X[,1],X[,5],gap=c(230,37600),gap.axis="x",
+ main="Grain ID and ASTM grain size",xlab="Grain I
D", ylab="ASTM Grain Size",
+ xtics=c(0,100,200,37600,37700,37800,37900,3800,3810
0,38200,38300,38400))
> dev.off()
RStudioGD
2
>
5 obs. of 5 variables
Grain ID and ASTM grain size
ASTM Grain Size
Grain ID



We have done it already. So, we will do once more. This is very common. So, we want to save it as a PDF in the figures directory, we want to call this as grain size scatterplot dot PDF and that is what the name of this file be and we are just going to give the plotting commands and device of to tell our to close this PDF file and come back to showing figures to you on the screen.

So, we do and there is a plot that is generated. So, we can go to the figures directory and see that there is a file that is generated this is a grain size scatter plotter dot PDF. So, you can look at the properties and you can see that it is just generated now, ok. So, this is the plot that is generated. Grain ID versus ASTM grain size, ok. So, what is the next step?

(Refer Slide Time: 25:07)



Stem and leaf plot and dot chart and other measures. So, we will do that next.