Engineering Econometrics
Prof. Rudra P. Pradhan
Vinod Gupta School of Management
Indian Institute of Technology, Kharagpur

Lecture-10
Exploring Data on Spreadsheets (Contd.)

Hello everybody, this is Rudra Pradhan here welcome to engineering econometrics and today we will continue with the Exploring Data on Spreadsheets. And in the last couple of lectures we have already highlighted the use of data and that to how to enter the data how to process the data with the help of a particular chapter called as a excel spreadsheet. In fact we have discuss in details the operations the kind of you know features like you know we have the 20 you know functions, we have a statistical functions, then we have a data analysis pack, then x we have a excel start.

So; that means, typically in the last couple of lectures we specifically give trace on a on the use of data with the help of you know excel spreadsheet. So; that means, a in the process of you know econometric engineering or in the process of you know that the data analysis, you should know how to you know enter the data then we know get a kind of you know structure.

So, that you know we can do the analysis in a better way and we can apply particular model in a much attractive way. So, that is how we should know details about the data and how to record the data, how to process the data how to structure the data. So, you know in totals we should know in details about the data recording, data processing, data structuring in all these are you know the essential in the process of you known data analysis and the use of the engineering econometrics, you solve some of the problems as per the particular you know organizational requirement of the kind of you know sector requirement.

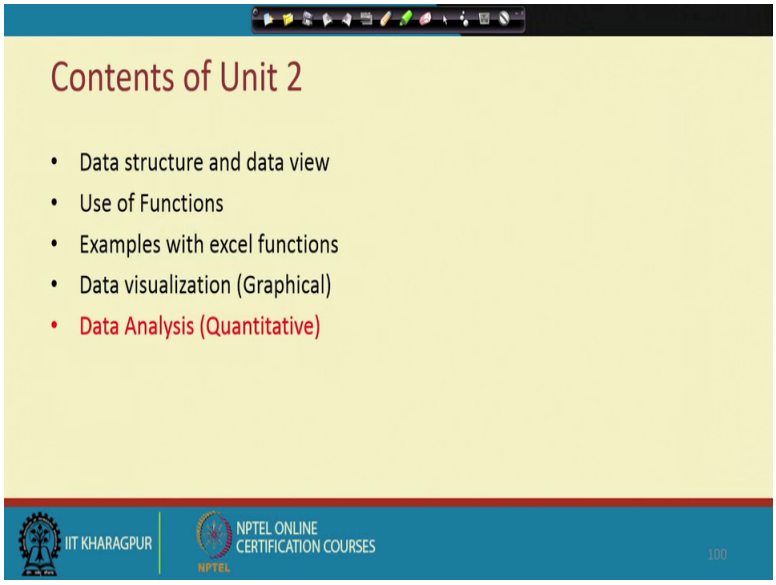
So, in this in this lecture we specifically gives stress on the again the you specific use of you know excel spreadsheet and that to the advance use of data and sometimes you know we have a first data, and by the help of first of data with the simple you know the processing, will get some kind of you know inference. And that may not be very attractive inference, but somehow the a these inferences can help you lot to choose a

particular model and at a you know that may be the best model a as per the particular you know engineering problems requirement.

So, typically in last lectures, will have gone through the component (Refer Time: 03:01) of data visualization, where we have discuss details about the plottings and that too we have gone through various types of you know graphs that to know the nature of data, the kind of you know structure of the data and we like to know the kind of you know functionality, all these things are you know we can clear highlighted.

And what I have mentioned the same data can be visualized through various you know you know diagrams or you know graphs starting with the line a you know the line charts, then you know bar diagrams and something that and then a you know we will get means at the end of the day we will get a kind of you know structure that you know, yes that I g like this and this is how we did do or we will go for the kind of you know analysis. So, in this in this lecture will go little bit more about this you know that is the typically on quantitative aspects.

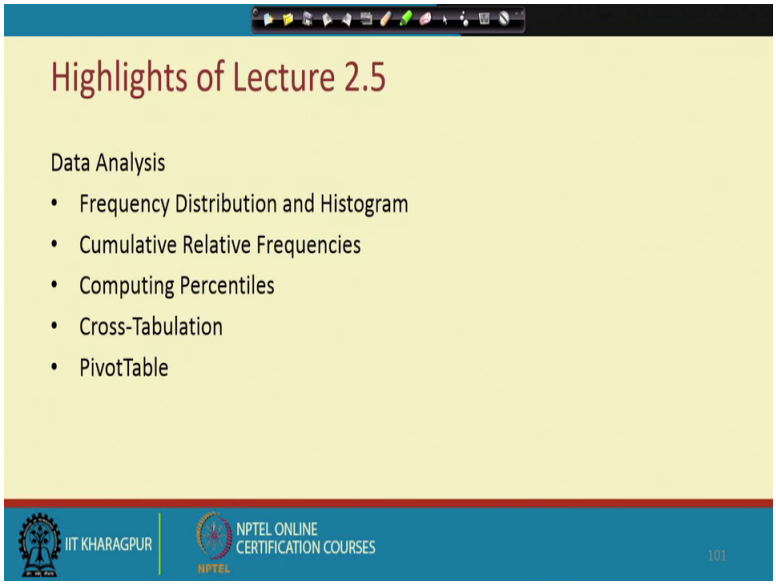
(Refer Slide Time: 03:54)



So; that means, technically what we have done in the last lecture, just you know the actual data and we just observing they are the kind of you know lecture or the kind of you know actual (Refer Time: 04:13) observations; just you plot these data and you can get the our viewer for instance whether the data is uniformly distributed. So, or there is a kind of no a linearity structure or there is a kind of non non-linearity structure. So, once

you actually plot the data you can get in details and with the help of these details, we can go for you know advance modeling.

(Refer Slide Time: 04:38)



So, in this case you know we like to know something more about to the data processing. In fact in the next unit we will be got details about the discrete econometrics where we will go through some of the basic statistics and little bit you know inferential statistics to know how the advance econometrics modeling can be process and that too can be applied, to solve some of the engineering problems.

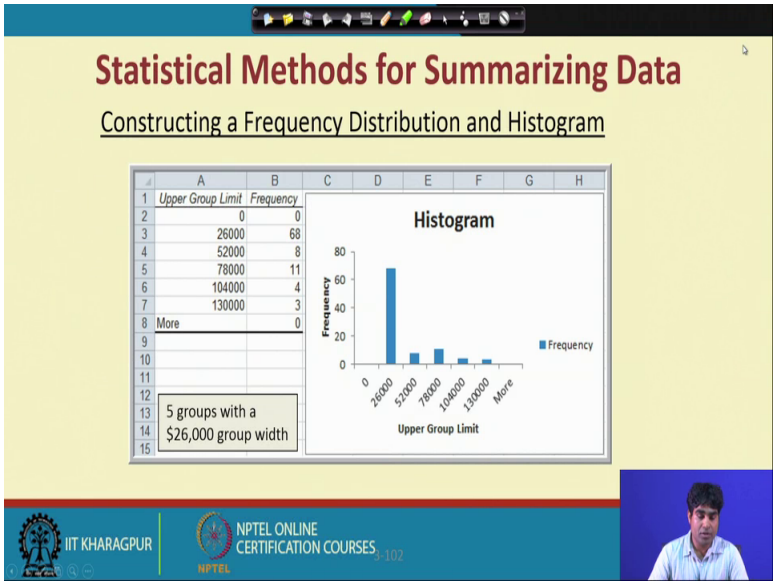
So, we like to know you know some of the you know data analysis requirements; that means, you know when we have a data some you know basic idea you must have, for instance the first hand data is called as a primary data and then if you will you somebody will use the primary data in a another kind of in a research or some other way of you know use, where it is called as you know secondary data or you know secondary kind of in a research.

So, now what is happening when you when you get the first hand data the data may be in a (Refer Time: 05:42). So, we need actually proper structure before we go for you know processing the data as per the requirement. For instance the first hand reporting and then the kind of you know processing. So, let us say it is a kind of you know a aspect or something kind of you know height factors in a particular in organization. So, randomly you can choose and then you will be find you know there are may be 100 peoples and

then you will find there are different types of you know peoples, having you know different levels of age, difference levels of height and there may be chance that you know a particular you know numbers may be repeated frequently.

So, that is why you know we know some of the basic statistic, through which all these details can be clustered and you know put in a kind of you know structure so, that you know it can be easily you know you know process and the anybody can understand very quickly. So, that is how in this lectures will know something little bit about the distributions, frequency distributions, histogram then cumulative frequencies, then percentile and some of the basic statistics and a the kind of you know per structure collage you know cross tabulations, pivot tables and a you know you know more of the basic statistics a something like that. And a you will see here this is a kind of you know a structure and a here is what is happening.

(Refer Slide Time: 07:07)



So, a there is a actually you know observations and the actual observations are you know in a different steps so; that means, here the range is the something called as you know 0, 26 26000 then 52000 a 78000 then 1,04,000 1,30,000 something like that.

So that means, you know you just you know enter the data and then you will find you know the particular data can be clustered, in such a way that you know it will be in a kind of you know proper structure for instance you see here all together. So, we have actually 68 observations 8 observations 11 observation 4 observation, so the these are all

actually different numbers with respect to a particular you know group. So that means, technically let us assume that this is a income group.

So, there are people a there are you know figures 0 and the there are you know in that particular organization you will not find any people whose income is actually 0. Then there are 26000 if you will having a income 26000 there are 68 numbers and then there are people you know 18 numbers they have the income 52000; similarly there are 11 peoples whose income is 78000 and there are 7 peoples who are you know above 1 lakh.

So; that means, now you just you know a I will tell you the process of the you know recording and the kind of you know structuring through a e excel spreadsheet. Just you know since lets you know here the total observation is you say 80, so now you just a randomly ask you know 1 to 80. So, what is your you know income levels then you know you just 1 by 1 you know enter the data, then finally you wants to enter the data you will see you know 0 at the 0. So, no people are there in the 26000 then there are you know 68 peoples.

So, now you just enter 1, by 1 then finally you can process it and that they are particular process can be you know through excel sheet and so the entered instead of writing so many in the data points, so you have to just write you know 6 numbers and then followed by a corresponding frequencies. So that means, the enter spread sheet now you know (Refer Time: 09:36) means clustered into two different you know columns, where in the first column we have we have 6 observations and followed by second clause we have a 6 observations and the first column represent the different income levels and second column represents the frequency against these income in all levels.

So; that means just you know just look here in that particular figure, so the first hand collections and now the we reporting the of these first hand information in a kind of you know systematic format and that is a that is very much possible. And that is elements it can be very easily processed through excel spread sheet only and again you can also plot these a figures and in order to know details about this kind of you know processing.

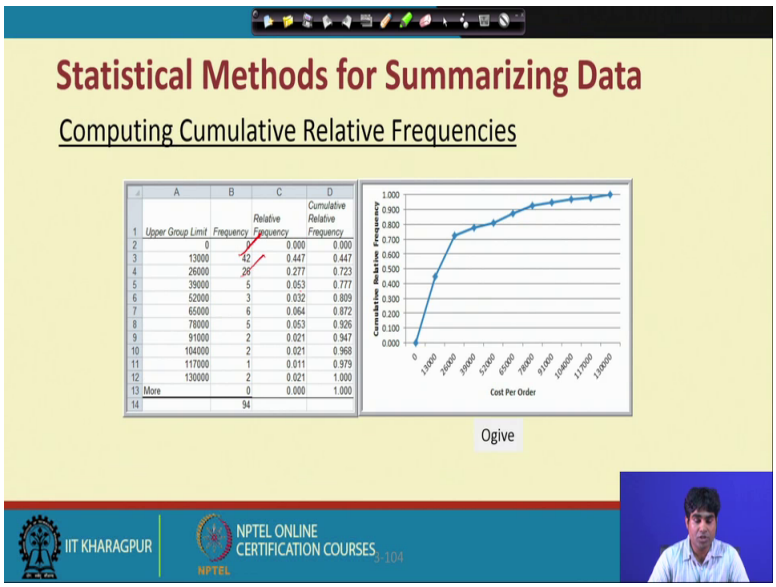
So, now likewise the that means 26000, so the a you know frequency is 68 that is highest and then a it is 52, so it is actually third highest then this is second highest then this is forth highest then this is 5th highest and there is no other figures. So that means, it is a

very easy to you know record and you know observe the particular you know a pairs in a much better way you can say so that means, technically.

So, again so with help of excel spread sheet we can no details and then it can be also summarized a particulars you know structural you know frame works. So, that you know it can be easily understood and anybody can also understand even elevens can a also understand the kind of you know structure the kind of you know reporting and the kind of you know inference. Similarly, so we can have actually another kind of you know structure here so at the different levels; that means, we are just increasing the a income levels and then again you are recording the particular you know frequency levels then you can also visualize it.

So, the visualization we will give you which particular group I means income group is having high frequency and which particular income group is having you know a low frequency. So that means, it is a typical actually a different kind of you know a structure through which you can observe the data and you know get some kind of you know inference; infinity is not the kind of you know in depth inference which you can do through a I know econometric modeling. But this is somehow you know basic inference we can draw from the a primary data or the first hand or you can say the first hand data.

(Refer Slide Time: 12:14)



So, a obviously, the particular data can be further structure, against you know income levels frequency then relative frequency and then a you know called as cumulative

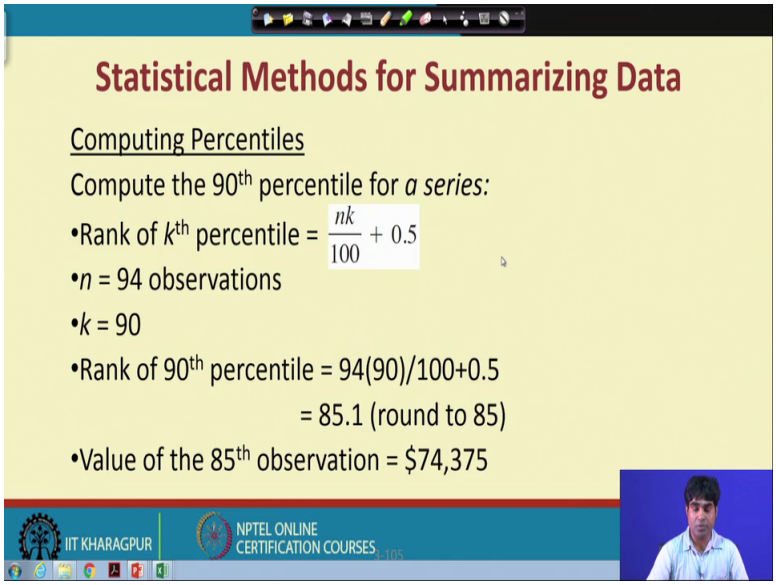
frequency. So that means, a relative frequency means actually so frequency divided by you know total and that will come in to the relative frequency again.

So, if will add one after another level of frequency then will get you know cumulative frequency. So, again you can check the you know check the data with the respect to you know cumulative frequency, like you know engineering economics you will find you know a you know cause factor individual cause factors and total cause factors, so it is like that only. So, a every point of every stage of you know production you adjust adding the you know African syllabus and then you will get something called as you know cumulative frequency.

So, then you know every point you know you find out the midpoints and the you join the midpoints then you will get a particular you know curve and that particular curve is called as a Ogive. So, this is a graphical plotting of the particular you know distributions and we like to know how we these particular you know distribution.

Of course there are couple of you know distinguish statistics are they are which can be you know better explain the structure of these distributions which you can discuss in the next unit. But in the mean times if you have a data you can process it and structure it as per the particular you know requirement and you know you can say the better understanding of the you know primary structure or you know primary data or you know firsthand data.

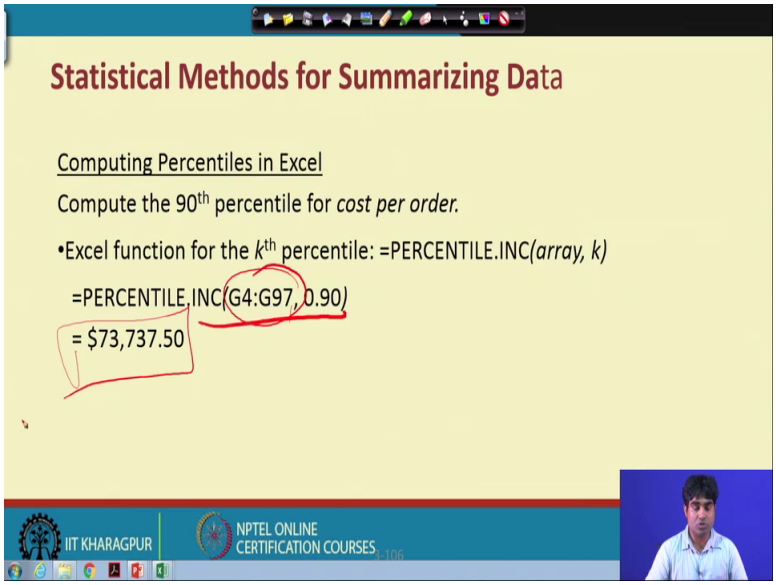
(Refer Slide Time: 13:53)



And a some of the a some of the structure a you know you can also follow by calculating percentile calculating, you know you know say average then calculating medians. In fact, in this case you know you can just check how much you know percentiles within this particular you know observations, so particular you know groups.

So, it is actually means that means technically excel has these options. In fact, this is your formula through which you can actually a calculate the percentile, but excel can help you against to do this kind of you know calculations and a then report the figures as per you had you know a requirement. So, this is have the kind of you know basic structure through which you can understand the data and predictica kind of you know environment.

(Refer Slide Time: 14:50)

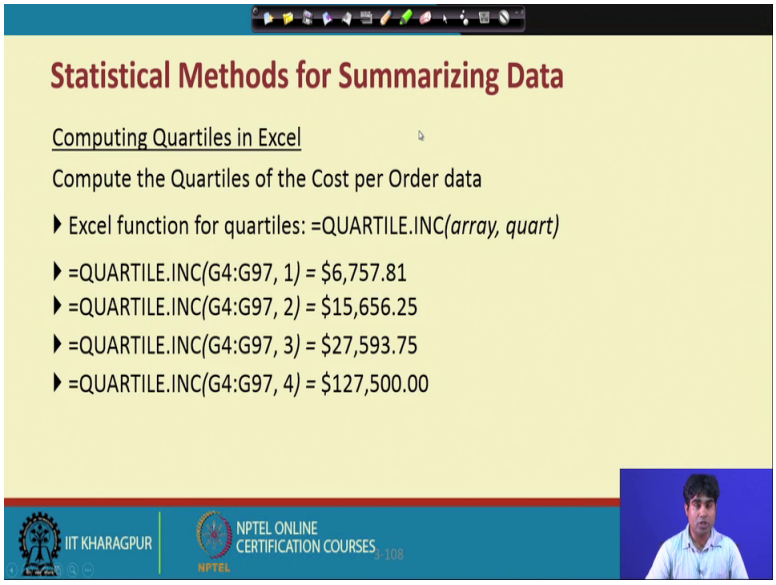


And here you see I mean see in the particular you know series how to repot you know percentile, just you go to the excel you know spread sheet and a you know just below of that particular series you just put it equal to and then as per the percentile requirement you know just indicate this series like here. I means this is this is have the particular you know series a where you can find out the kind of you know requirement ok, so this is what the requirement. So, that means, technically in this case.

So, we have series G4 to G97 which we have discussed in the last lecture and then you just in ask for you know percentile 90 percentile then by the 40 you can get to know the numbers ok. So, which you can you know better you can you can you know see in the

case of you know excel spread sheet. So, I will show you this one with the help of in a spread sheet and again there is a quartiles you know first quartiles second quartiles third quartiles, so this will give you different you know steps of the distributions.

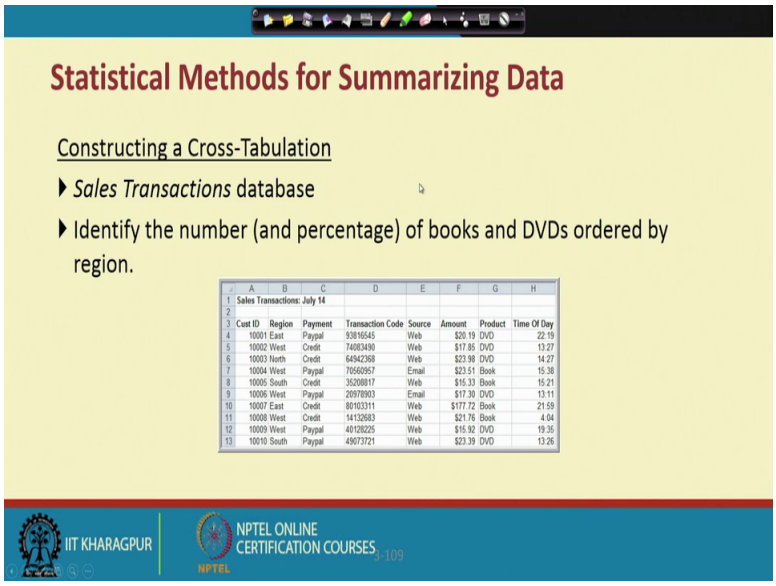
(Refer Slide Time: 15:58)



So, there means you know on the idea been this particular you know you know discussion is that, so we like to torture the first hand data you know heavily in order to get better inference and then with the help of you know the inference we can use some kind of you know advance econometrics tools. So that means, you know typically one of the a basic requirement of econometric modeling is you know understand the data perfectly it is not the kind of you know collection of the data or recording of the data.

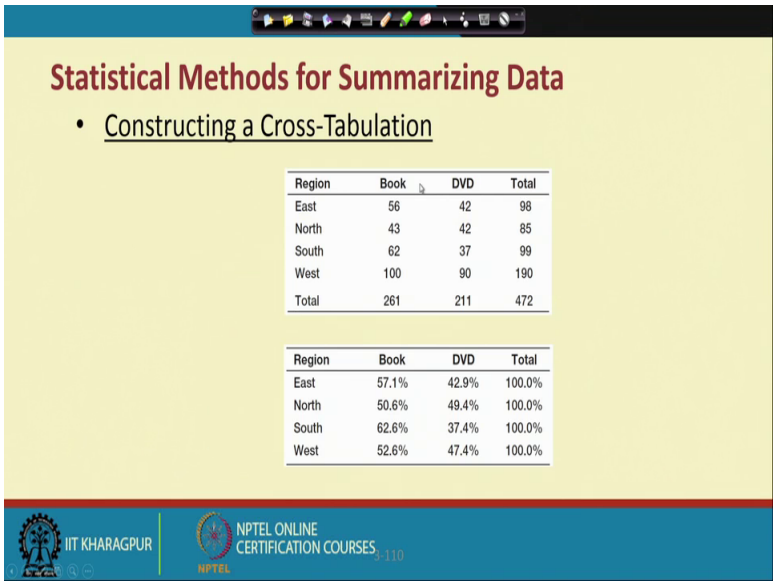
But you have to o understand the data very perfectly that is why you need actually , very much you know what we can called as you know torture heavily you have to torture the data. So, that you can get you know better inference and that will help in lot to go for you know advance econometric modeling.

(Refer Slide Time: 16:57)



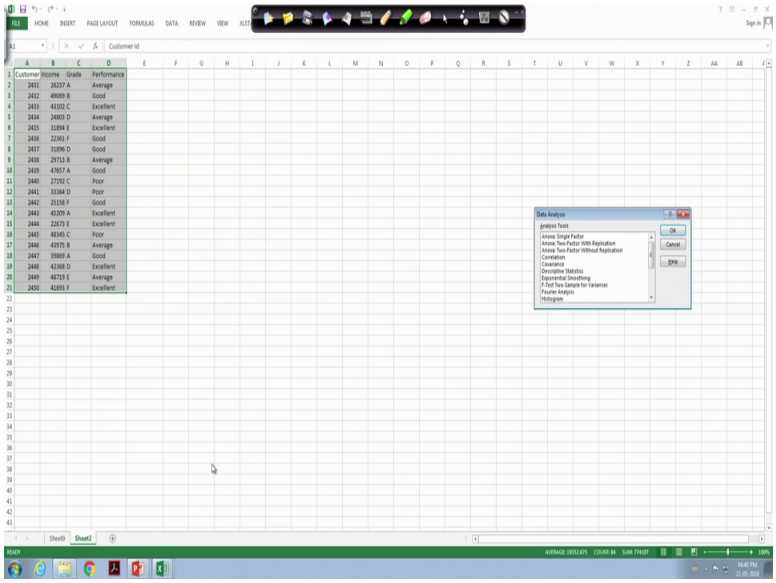
So, likewise you have a different you know a structure through which you can actually a you know observed with that data and predict this situation as per the particular requirement.

(Refer Slide Time: 17:06)



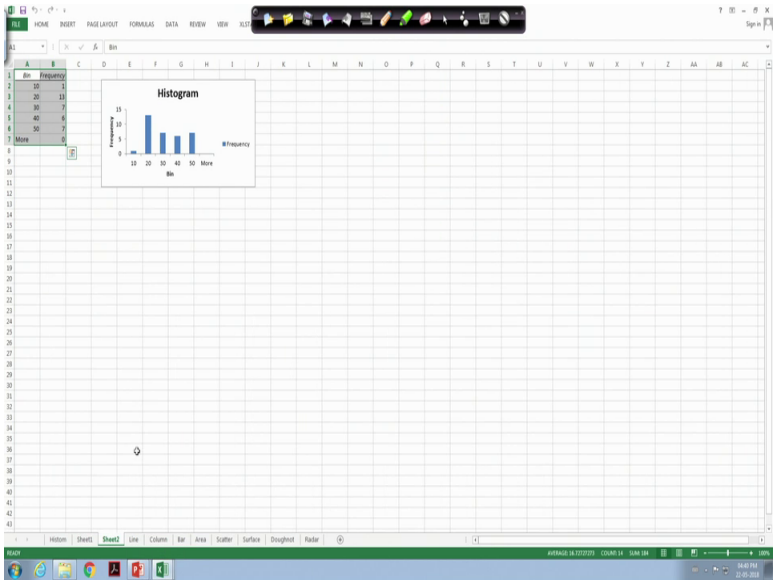
So, let me give you a kind of you know hint, a how in the you know help of you know excel spread sheet.

(Refer Slide Time: 17:15)



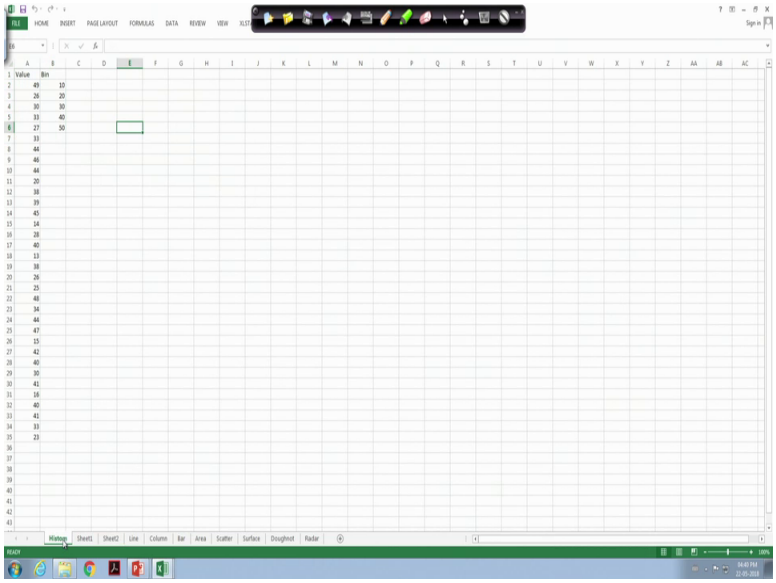
You can actually explain the particular you know requirement and as see here is so this is have this is have the particular you know case ok.

(Refer Slide Time: 17:35)



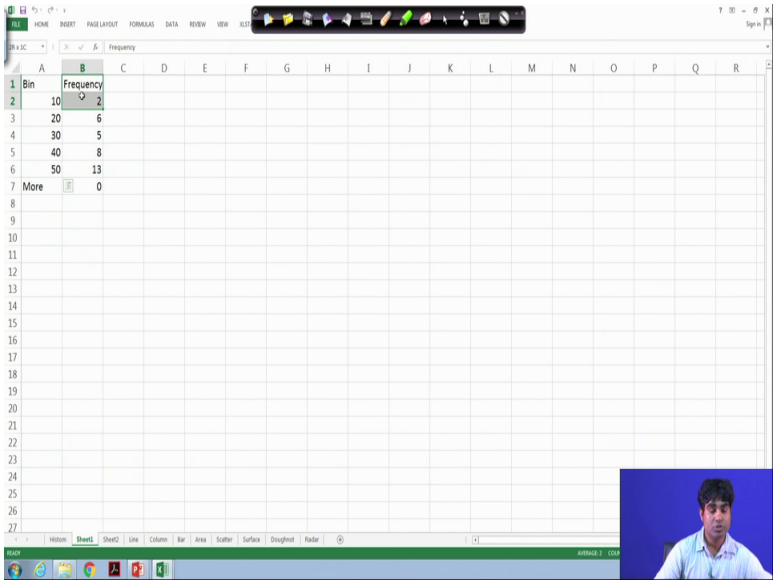
So, let us let us see here.

(Refer Slide Time: 17:37)



So, this is we start with you know the simply let us say histograms. So, these are all kind of you know observations and that will followed by a frequency distributions.

(Refer Slide Time: 17:44)



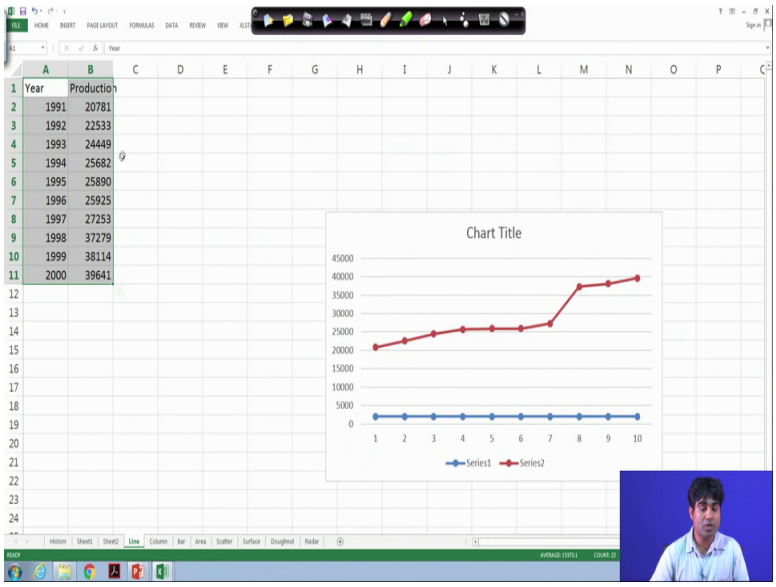
So, also the observation is like this, so a you know the group follows 10 20 30 40 and a 50 and more than that and then we have a frequency here.

So, just you just what will you do you can just in highlight these ones and then you go for you know or something like that you know histograms. So, you refine these pictures

are you know visibles clearly ok. So, let me give you hint like this see here at the histograms will be a appears here.

So, now a you know this an things are you know plotted here so that means it is the first hand information's. So, in fact the particular recording can you know give some kind of you know inference, but now these particular recording can be visualize you know more in the form of a graph which can give you better kind of you know inference again.

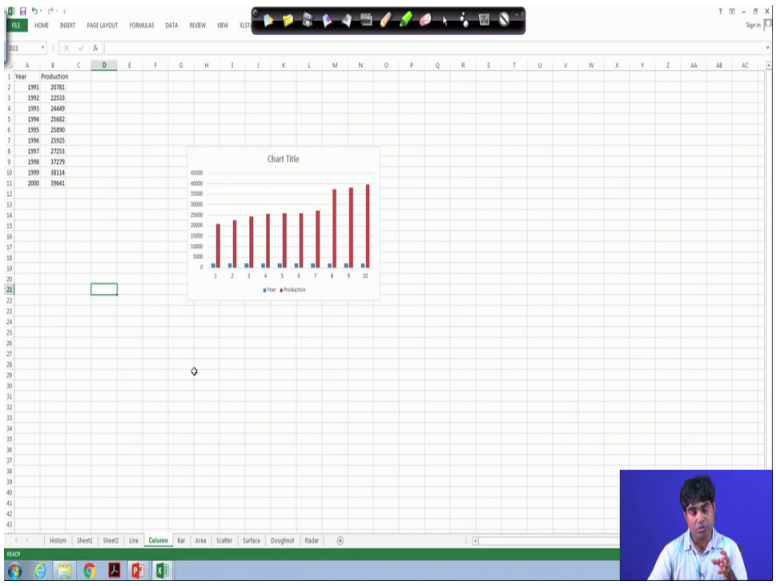
(Refer Slide Time: 18:47)



You know which we have a line diagram let say this is early productions and then we will find how this percents are you know plotted with you know different you know structure. So that means, technically so this is a better kind of you know inference through which you can actually you know understand the data as per the particular you know a requirement ok, so this is what the a data.

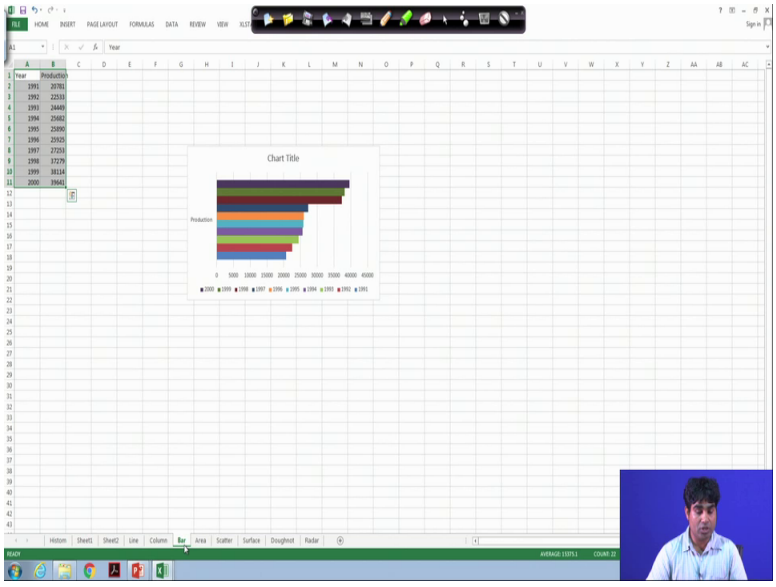
So, again so the plotting will be a showing this particular plotting is showing actually it is kind of line diagram and then again is this an things you can transfer to a column diagram.

(Refer Slide Time: 19:25)



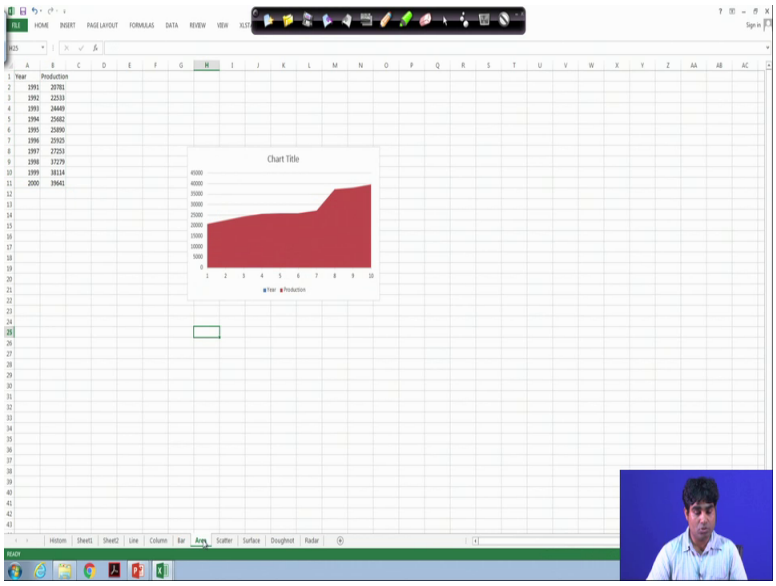
So, means a so it is a kind of you know different kind oh of you know approach all together because, some information's we are visualizing in a different angle all together which we have discussed in the last lecture. But with the you know practically you can take the examples like this and then you can actually, so the kind of you know situation.

(Refer Slide Time: 19:47)



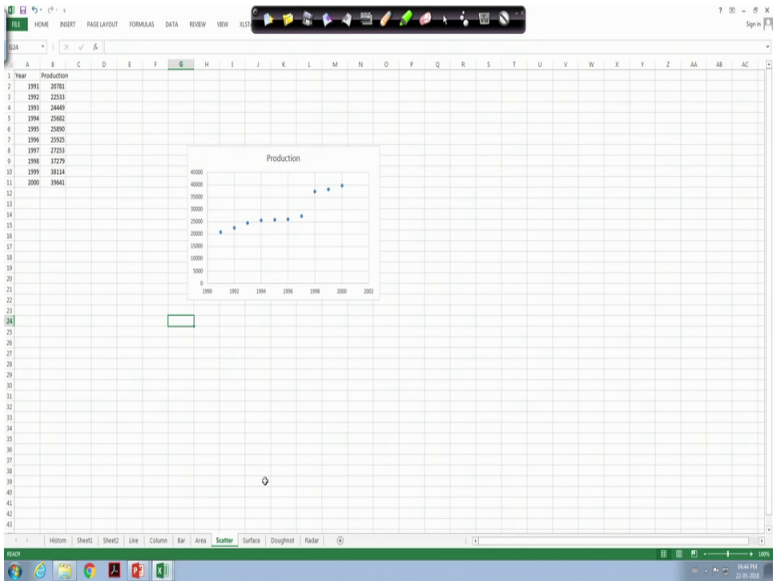
Similarly, this same data you can observe through bar diagrams and again.

(Refer Slide Time: 19:52)



You can also observed through area diagrams and again.

(Refer Slide Time: 19:55)

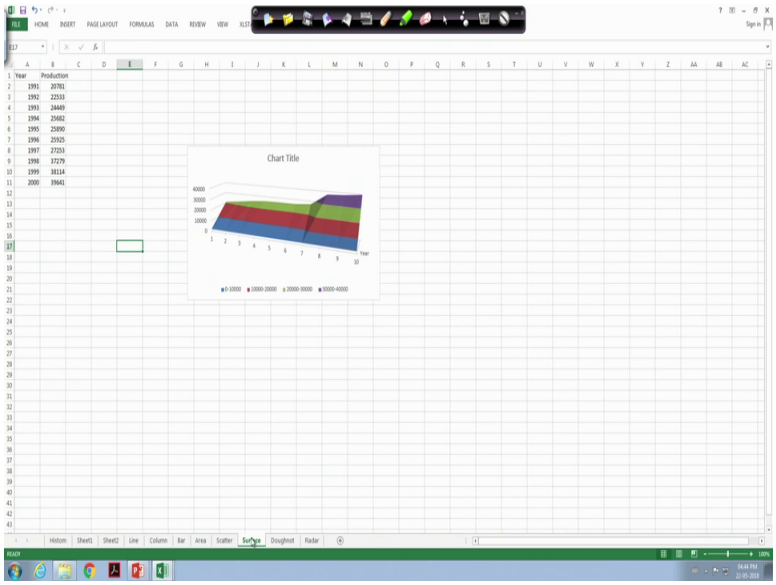


We can observe through you know what we can call as you know scattered as scattered plotting, that means, technically what I am saying that you know this an information. So, we are not getting any kind of you know extra inference pro for these processing or by using different you know graphs and diagrams. But what is the main idea here that you know we should not different a kinds of you know plotting to visualize the data in order to understand the data perfectly and by using different plotting it will give you some kind

of you know robustness check or you know in depth kind of you know understanding or you know in depth kind of you know visualization.

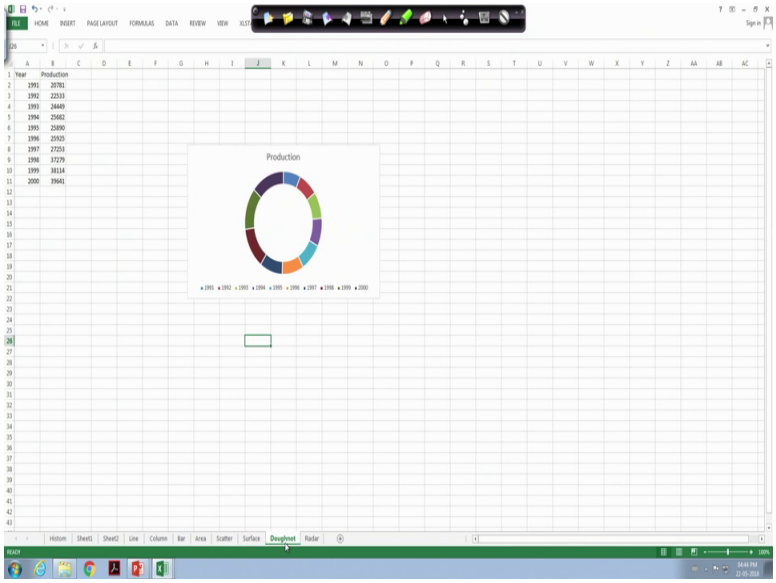
So, that you know there is there should not be any kind of you know error for you know any kind of you know advance data modeling or advance you know econometric modeling. So, that is why you have to torture the data heavily a in the first instance, so that you know later stage when we will go for advance econometric modeling you not find any difficulty.

(Refer Slide Time: 21:04)



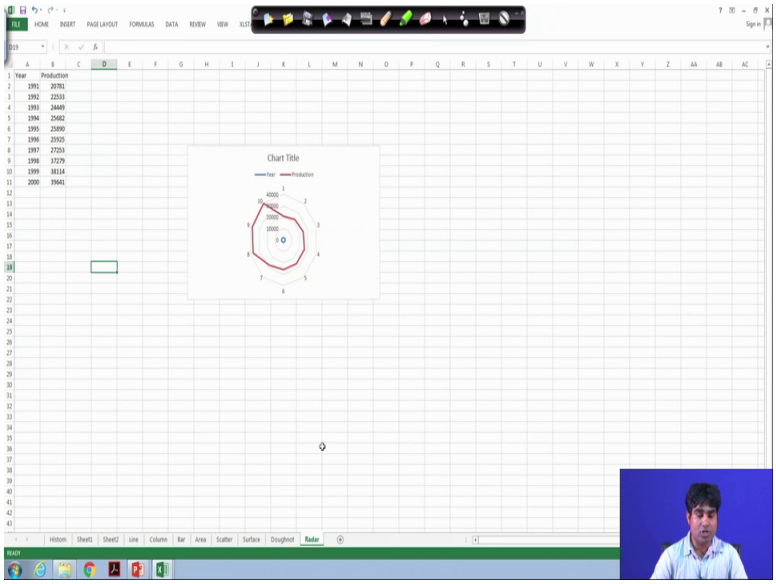
So, you have to torture the data then a you have to get then inference as per the particular new a requirement, you send data against in the form of a chart we can be visualized and can so as per the particular requirement and similarly a dot net a dot net kind of you know plotting.

(Refer Slide Time: 21:20)



So that means we have a different levels of you know plotting to understand the data a to get some kind of you know inference a as per the particular requirement similarly.

(Refer Slide Time: 21:31)



We have also discussed (Refer Time: 21:35) rather the adder kind of you know diagrams, so that means see here so if you start with here actually, then this is a simply you know year wise production data from 1991 to 2000 and we are just checking at the you know production behaviors with the help of you know different charts and diagram and we are

just visualizing to know what is a happening a you know in a in this kind of you know you know that I you know reporting.

So, same things you know same data will a line chart column chart bar diagrams area diagrams scatter plotting surface diagram dot net radar. So that means, see here a same data as you know and the same data points and same number of years, just we are you know watching the behavior of the data or behavior of the productions with respective you know the different time periods and we with respect different you know plotting.

So, at the end of the day it will give you better inference as per the particular requirement and which will help you lot for you know I know getting some kind of you know better you know insights in the process of you know using converting modeling for engineering problem.

So, now a for a you know for a better understanding again, so we have to gone through different level of you know structuring. So, one such mechanism is called as you know cross tabulations. So, for a examples so here is there is a regional regionalize information's and then that is with respect to three different products book DVD book DVD and a something can be of you know CD.

But here is in this particular you know plotting so we are just showing 2 items a that is actually book and DVD and that to with respect of for different you know region of you know let us say particular you know country something kind of you know state, a then you know we just a check the you know size and then we visualize the how these kind of you know (Refer Time: 23:46) about. So that means, see here is in the case of you know cross tables, so it is the actual observations and this is actually region wise. So, the total books sales let us have up to you know total 261 and then again DVD sales 211 and this is actually together a total sale sales that is you know 261 plus 211 that is coming 472.

Again so in the eastern region, so we have a actually sale of you know book 56 and sale of book you know 42, so again it will appearing 98 single do not the region there will be sale a books and there is a sale of you know DVD and again it will coming 85. Similarly, for south region and similarly for you know west region; so that means, technically this is actually total selling across the region and across the you know sales and now in order to understand the better inference about these particular you know the data.

So, we will transfer the data in a percentile which is possible in the case of you know cross tabulation and that means technically a so the eastern region, so the books selling is you know 57 1 per 57 0.1 percent in northern region is 50.6 percent and southern region 62.6 percent and western region 52.6 percent. So that means, technically it is very easy to calculate and excel can help you lot to do this transformations so that means technically.

So, it is simply actually 56 by 261 divide mean divided by 261 multiplied by100. So, the 56 number will transparent to this percentage, similarly if you are doing this one then it can just you know scroll it. So, you will get you know other percentage for you know other regions and similarly for DVD you can also transfer the actual number in 2 percentage and that means technically what is the duty of this particular you know structure you know the spread sheet is that.

So, you report the data first hand data then you transfer into another form which can give you better visualization and some kind of you know better inference and that that itself will give you some kind of you know you know insights1 some kind of you know better inference as per the particular you know engineering a in a engineering problems requirement. So, that is why so knowing the excel spread sheet and knowing some kind of you know a subtier for the process of you know data for the analysis of data as per the requirement.

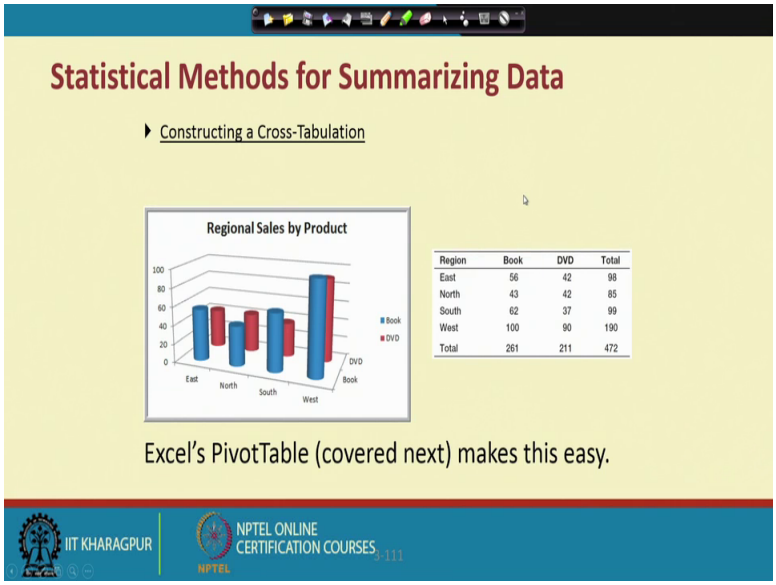
So, it is always you know suggestive that you know you must be a acquitted with these particular systems and this it is not that you should not know the a particular you know subtiers you should not be operations the kind of you know features, the kind of you know items availability, like you know graphical plotting we just we just means we have just gone through it. So, where you know we the same data we are plotting in different you know charts and diagrams.

So, you will find you know different look all together and it gives a different kind of you know you know understanding and the kind of you know it is better kind of you know understanding with a kind of you know different looks. So, it is completely what we use the term as you called as you know robustness check.

So, it will give you complete visualization and complete clarity a without any loop holes. So, now we are observing or you know torturing the data in different angles by a different tools and you know different you know structures all together. So, like wise so

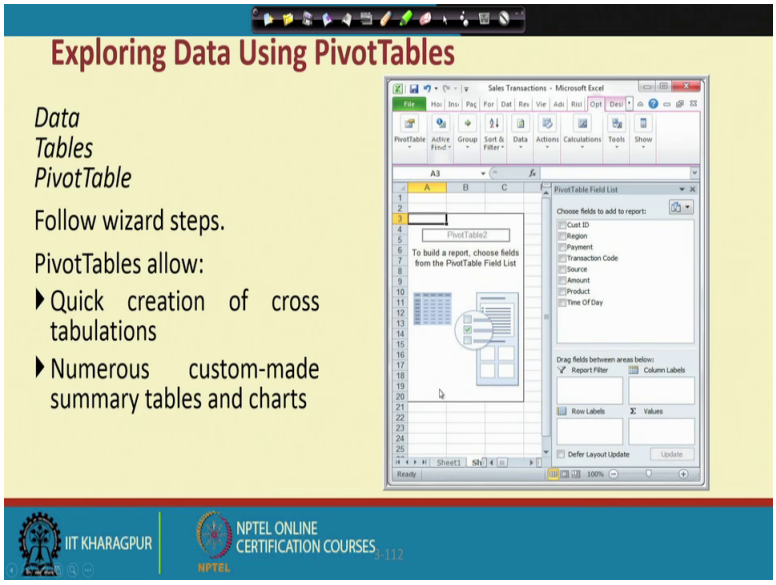
there are you know many different ways you can actually understand the data and then you can actually observe the data as per the particular you know requirements.

(Refer Slide Time: 27:39)



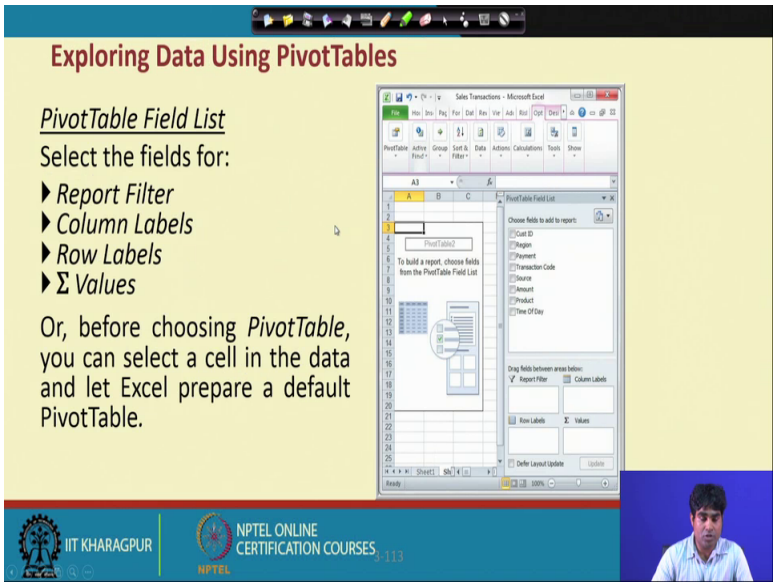
Similarly, this an things we are know that plotting here so see here is the actual data and that will be a you know plotted. So, that you know you can you know visualize much better way. So, here actually it is a in quantity figures, so whenever you know graphically you plot then you can know better you know means we here judgment will be much better then compare to this you know just reporting the figure in the excel sheet.

(Refer Slide Time: 28:05)



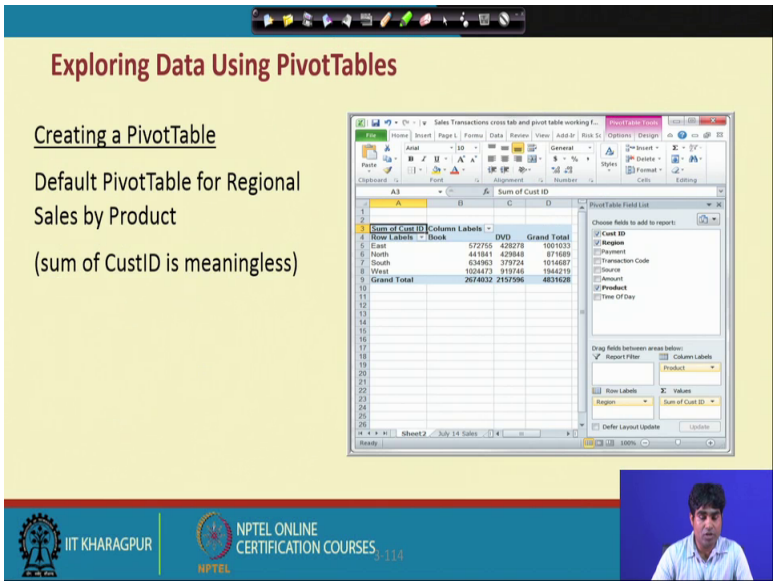
So, similarly so means in the excel sheet we have actually absents called as you know pivot table cross tabulation and some of the a statistical tools which can help you lot to you know a understand the data and to visualize the data in much better way. And you can use some of the basic tools like you know pivot table cross tabulation and they that itself will give you some kind of you know better inference, in order to understand the data or as per the particular you know you know problems requirement.

(Refer Slide Time: 28:41)



So, likewise you have a different you know you know structural together in the excel spread sheet.

(Refer Slide Time: 28:47)



So, my suggestion is that you know you must be very you know very acquainted with the excel spread sheet, the how to how to you know a operate the excel sheet and that to the use of the data and the kind of you know data analysis. In fact, the data analysis pack and excel start is a little bit you know complex and they very complicated sometimes depending upon the particular you know technically requirement and the kind of you know problem requirement.

So, you know so again the suggestion is that you know it is not that only have to know the excel spread sheet of the kind of you know the data entry data processing data visualization data analysis in a basic a at the basic levels. And at the same times you should know the advance operations like you know the at data analysis pack and the kind of you know kind of you know called as you know excel start pack something like that, for instance if you got to this you know excel spread sheet again. So, you will find so there are you know information here.

So, for instance you just you just go through to the excel spread sheet and here will find this is actually data and here do will you see there is a option called as you know the data in the top of the menu and you just click the data, will you find extremely right in the top of the corner again there is a structure called as you know data analysis you just click here.

So, there you will find you know you will find there is a actually variety of you know we you know it means variety of sums are there and here all kinds of you know statistical tools are they are starting with you knows as simple one you know complex one. Likes you know simple dispute statistic co variance co relations ANOVA then smoothing techniques you then you know obedience test a testing procedures histogram moving average a the (Refer Time: 30:56) percentile regression sampling all these you know tools already their actually.

So, what you can do so you just you know enter the data first hand data and then with the help of all these you know statistical tools or the analytical tools. So, you have to actually understand the data in a much better way and tried to get some kind of you know inference as per the particular requirement and there is a high chance that you know some of the engineering problems can be solved by using these simple technique, but if the problem is very complex and it cannot be you know understand very easily. So, then in that constant we have to go for you know advance technique and then operate the problem as per the particular you know requirement.

So, a so the suggestion is that you know be acquainted with the a you know excel spread sheet, how to enter the data how to understand the data how to visualize the data a how to get some kind of you know basic inference as per the advance modeling requirement or the kind of you know engineering problems requirement. Once you know all these details then you know it will give you some kind of you know better you know kind of you know environment, where here you know problem can be solved in a much better way and as per the particular you know requirement.

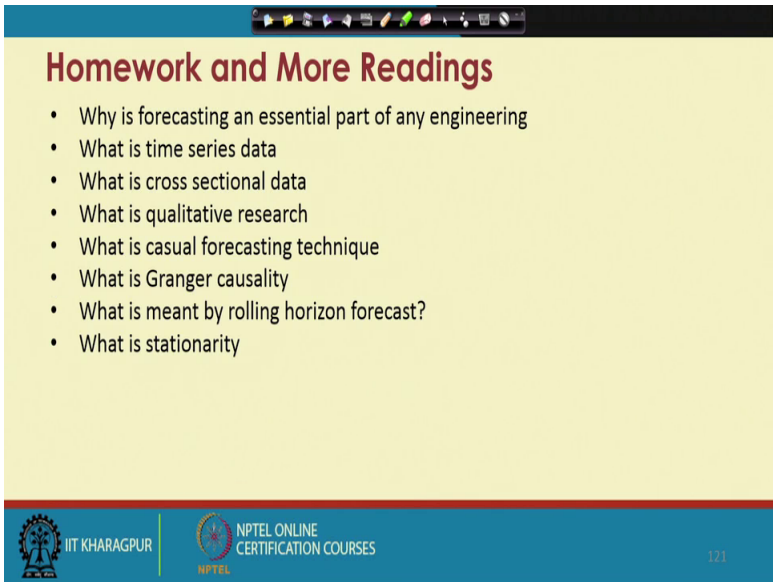
So, with this we will close this unit and in the a in the next unit will we start with you know basic a kind of you know say basic econometrics that that is what we called as you know discrete econometric, where will start with you know all these basic statistic then will obtain you know we will solve some of the problems by the help of these you know data analysis pack.

So, now a in this particular lecture I what we have done here that you know we just go through all these you know duty of the excels you know excel spread sheets and then we get to know what are the operations what are the functionality and how to operate how to visualize and how to get some kind of you know basic inference. So, now that itself will

take you in some in a advance level, so that where you know we can use some of the you know inferential statistics or (Refer Time: 33:13) statistics to solve some of the problems as per the particular you know engineering econometrics requirement.

So that means, technically so a you have to very careful about these you know data and the use of you know spread sheet as per the engineering econometrics requirement. So, a likewise you have a different kind of you know structural together and.

(Refer Slide Time: 33:40)



So, my suggestion is that you know you must be a you know against you know acquainted with the excel spread sheet and before you know close this lectures a like to you know suggested that you know you should go through some of the component like you know understand what is the forecasting structure and the data structures, like time series data cross sectional data qualitative research forecasting with in a different techniques.

Like you know smoothing techniques moving average and some of the advance techniques like you know regressive causality a, so stationarity all these things are they are send with the help of the actual excel spread sheet. And the kind of you know some of the statistical structure you can be easily acquainted with the systems and then we will tried to get some kind of you know inference as per the particular engineering econometrics requirement these will it stop here.

Thank you very much have a nice day.