Practitioners Course in Descriptive, Predictive and Prescriptive Analytics

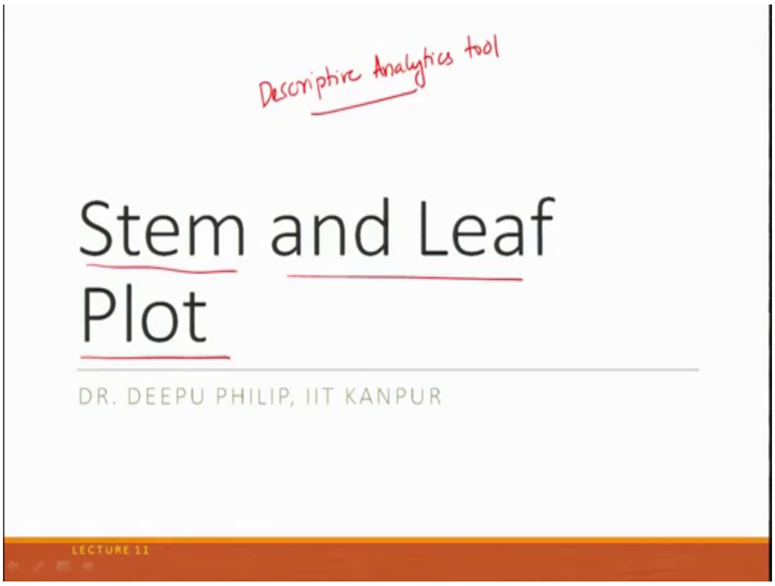
Prof. Deepu Philip
Dr. Amandeep Singh Oberoi
Department of Industrial and Management Engineering
Indian Institute of Technology, Kanpur
National Institute of Technology, Jalandhar

Lecture – 11
Stem and Leaf Plot

Good evening, I welcome all of you guys to the yet another lecture of the applied analytics course; which is for the descriptive, prescriptive and predictive, analytics and we are running this course from a practitioners approach.

So, the title of the course is practitioners approach in this case and we have been seeing different aspects of analytics why is it important why is the having a question in mind or a hypothesis in mind is important or not and we already saw that we require 3 things; important one is to identify the details of the independent variable dependent variable and appropriate descriptive statistics or appropriate the statistic test needs to be identified to find out the answer to the question the decision question in hand and we started looking into how to describe the data and we already saw some few techniques out of this and one of them we saw was earlier was frequency distribution where we were using it for ah handling large set of datas.

(Refer Slide Time: 01:23)



Descriptive Analytics tool

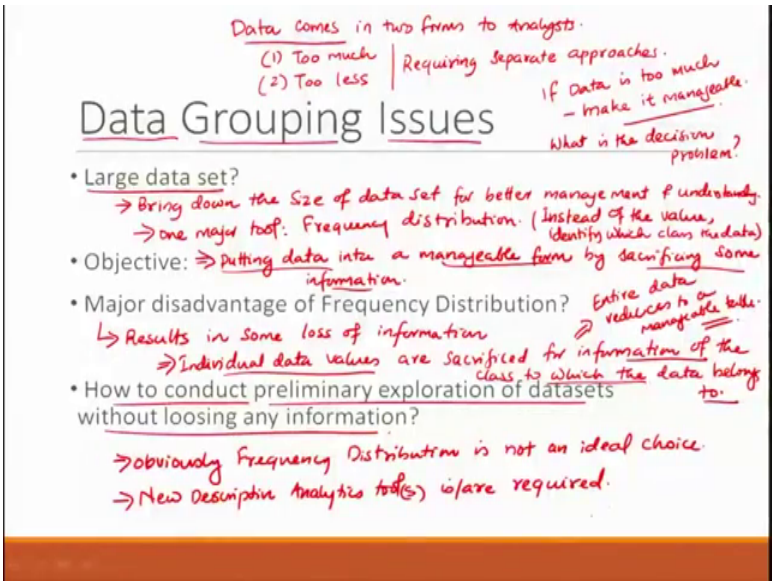
Stem and Leaf Plot

DR. DEEPU PHILIP, IIT KANPUR

LECTURE 11

So, today, we are going to see another new tool called stem and leaf plot and this is another descriptive analytics tool ok. So, we are going to study a new descriptive analytics tool called stem and leaf plot.

(Refer Slide Time: 01:38)



Data comes in two forms to Analysts.
(1) Too much | Requiring Separate approaches.
(2) Too less

If Data is too much
– make it manageable.
What is the decision problem?

Data Grouping Issues

- Large data set?
  - → Bring down the Size of data set for better management & understanding.
  - → one major tool: Frequency distribution. (Instead of the value, identify which class the data)
- Objective: ⇒ Putting data into a manageable form by sacrificing some information.
- Major disadvantage of Frequency Distribution?
  - ↳ Results in some loss of information
  - ⇒ Individual data values are sacrificed for information of the class to which the data belongs to.
  - ⇗ Entire data reduces to a manageable table.
- How to conduct preliminary exploration of datasets without loosing any information?
  - ⇒ obviously Frequency Distribution is not an ideal choice.
  - → New Description Analytics tool(s) is/are required.

So, as we said earlier the when you have large set of data remember, in the earlier discussion, we said data comes in two forms to analysts. The people who are doing data analysis or data analytics, it comes in two forms number one too much or number two too less. So, and each one of these scenarios require and requiring separate approaches ok, in this case, we started seeing that when you have a large data set. So, today what we are talking going to talk about these data grouping issues which are predominantly pertaining to the large data sets.

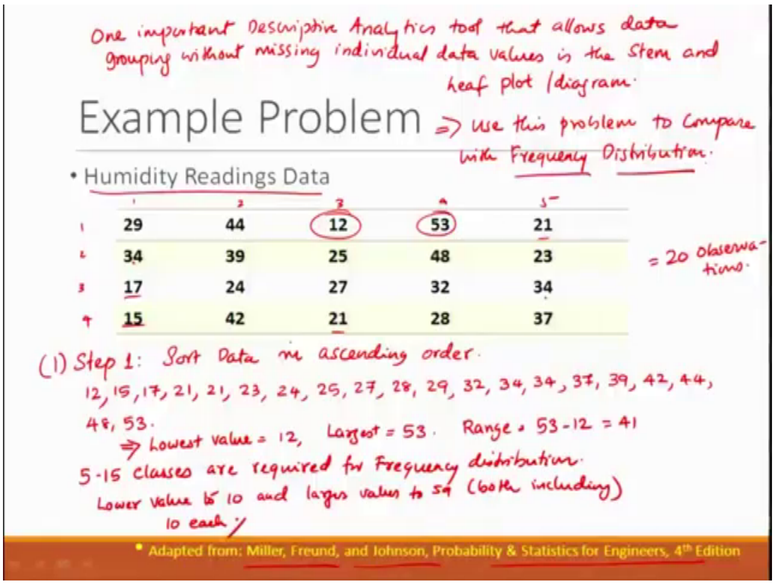
So, when the data set is large or if data is too much make it manageable that is the fundamental idea behind it how do we make it manageable one of the way to do this is ok. So, what we are doing is when you have a large data set ok, bring down the size of data set for better management and understanding, see we are still trying to figure out what the research question is we are try still trying to identify, what is the decision problem or for that we are trying to get a feeling of the data and when you have too much of data, then you have to bring in the data side out size out for manageability and the main reason of why do you bring it down through manageability the objective behind it is putting data into manageable form by sacrificing some information or as we said earlier one major tool is frequency distribution tool of handling large data set is frequency distribution frequency distribution.

So, where what we are doing is we are identifying not the individual values here instead of the value we are looking at identify which class it belongs to which class the data belongs to ok. So, that is what we are looking in the frequency distribution the idea of this is that. So, that you can put the data into a manageable form and by in doing that you sacrifice some information; so what is the major disadvantage? The disadvantage of frequency distribution as I said earlier is it results in some loss of information what is the loss of information individual data values individual data values are sacrificed for information of the class to which the data belongs to.

So, here we are sacrificing the individual data values and instead we are getting the information on which class the data value belongs to. So, the individual data value is no longer there inside the membership of a particular class is what we are getting and. So, that the whole data reduces to a manageable tabular format. So, the idea is that entire data reduces to a manageable table that is what the frequency distribution is all about. So, now, if you want to do if you want to conduct preliminary exploration of data set without losing any information then; obviously, frequency distribution is not an ideal choice.

So, hence new descriptive analytics tool or tools is or are required and one such tool that we are going to see, today is called the stem and leaf plot or stem and leaf diagram which allows us to study the data large dataset without losing too much of information.

(Refer Slide Time: 07:50)



One important Descriptive Analytics tool that allows data grouping without missing individual data values is the Stem and leaf plot /diagram.

Example Problem

⇒ use this problem to Compare with Frequency Distribution.

- Humidity Readings Data

	1	2	3	4	5
1	29	44	12	53	21
2	34	39	25	48	23
3	17	24	27	32	34
4	15	42	21	28	37

= 20 observations.

(1) Step 1: Sort Data in ascending order.

12, 15, 17, 21, 21, 23, 24, 25, 27, 28, 29, 32, 34, 34, 37, 39, 42, 44, 48, 53.

⇒ Lowest value = 12, Largest = 53. Range = 53-12 = 41

5-15 classes are required for Frequency distribution.

Lower value is 10 and larger value to 54 (both including)

10 each /

- Adapted from: Miller, Freund, and Johnson, Probability & Statistics for Engineers, 4th Edition

So, one way to study this stem and leaf plot. So, one important descriptive analytics tool that allows data grouping without missing individual data values is the stem and leaf plot or stem and leaf diagram you can call it either way and what he said we will first do an example and with that example we will compare it between we. So, what we will do is use this problem to compare with frequency distribution.

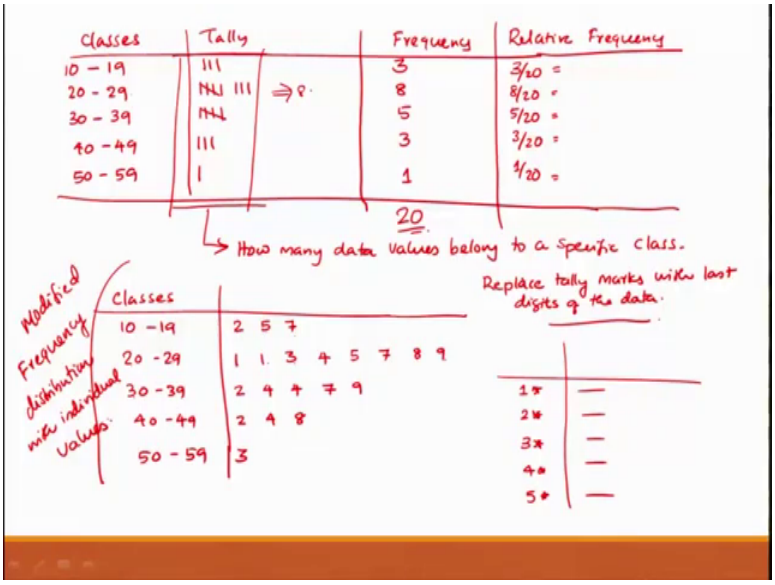
So, we compare with the frequency distribution how the stem and leaf plot is done and using that then we will see how which one is better stuff like that this example is adapted from the Miller Freund and Johnson probability and statistics for engineers and scientists and fourth edition and this data there is you can see that there is 1, 2, 3, 4 and 5; 5 columns and 1, 2, 3, 4, 4 rows. So, you have somewhere close to 20 observations here observations or 20 data values are available here, and this data is about the humidity readings they rounded to the nearest integer and this data we will use it to first plot of frequency distribution and then we will see how frequency distribution can be converted to stem and leaf plot in which the missing observation can be retrieved.

So, the first step as we said earlier step one sort data in ascending order now ok. So, the sorting will give us is if you look at this the lowest value in this whole regard is that is a 70 that is a 12. So, the lowest value is a 12. So, the data will start from 12, then we are 12, then there is a after that then we have a 15, then there is a 17. So, 15, 17, then we have 21; 21. So, 21, 21, then we have 23, then there is 24, then there is 25, 27, 28, 29, 32, 34, 34, two 34s, right this one and this one, 34 and 37, 39, then 42, 44, then 48 and 53. So, 53 is the largest value right, I do not see there is a nearer larger value than this.

So, this implies the lowest value equals 12 largest equals 53 and range equals 53 minus 12. So, that will be 41, right, yes. So, now, the question is you can look at multiple ways, we saw how to make classes and all those kind of things, but to make the life easy I am just going to make the classes vary from. So, it is about 41. So, if I have 5 classes, I can make it 2, I think I can make classes in different ways. So, what I will do is that we have a minimum of 5 to 15 classes is what we studied yesterday, in the earlier in the frequency distribution classes are required for frequency distribution ok. So, I can do something like make it as a 5 classes of 10 and then I can start from 10, I can go all the where to 60.

So, I will start the lower value and revise the lower value to 10 and larger value largest value to 59 includes something like this ok. So, then I can have a class size of like 10 each ok. So, let us see, how it can be done. So, let me make a frequency distribution out of this for 5 classes.

(Refer Slide Time: 13:26)



Classes	Tally		Frequency	Relative Frequency
10 – 19	|||		3	3/20 =
20 - 29	𝍸 |||	⇒ 8.	8	8/20 =
30 - 39	𝍸		5	5/20 =
40 - 49	|||		3	3/20 =
50 - 59	|		1	1/20 =
			20	

How many data values belong to a specific class.

Modified Frequency distribution with individual values.

Classes	
10 - 19	2 5 7
20 - 29	1 1. 3 4 5 7 8 9
30 - 39	2 4 4 7 9
40 - 49	2 4 8
50 - 59	3

Replace tally marks with last digits of the data.

1*	—
2*	—
3*	—
4*	—
5*	—

So, the frequency distribution will be classes and you have the tally and the frequency oops sorry somewhat here. So, the first class I will start from 10 the lowest value and I will go all the way to 19, both 10 and 19 included because all the values are single digit values then the next well start from 20 and go all the way to 29, then 30 all the way to 39, 40 all the way to 49 and 50, all the way to 59 ok. So, in which I should get all the values completed as part of this and I can also say that in the if you look at the previous data, you can see that between the first class which is 10 and 19 I have 3 values 12, 15 and 17.

So, in a frequency distribution I will be doing 1, 2, 3. So, that will be the 3, then 22, 29 is my next class. So, then I go back to the previous one. So, 22, 29. So, this is 1, 2, 3, 4, 5, 6, 7, 8, ok, there are 8 of them. So, it starts with 20. So, 21 to 29, I will 29 is also included because that is how the class is being defined. So, if you look at in this I will have 8 of them. So, it is 1, 2, 3, 4, 5, 6, 7, 8, then the next class is 30 to 39. So, I can go and do it as 30 to 39 will be 1, 2, 3, 4, 5. So, 5 of them. So, it will be 1, 2, 3, 4, 5 next is 42, 49. So, I will go as 1, 2, 3, 3 of them.

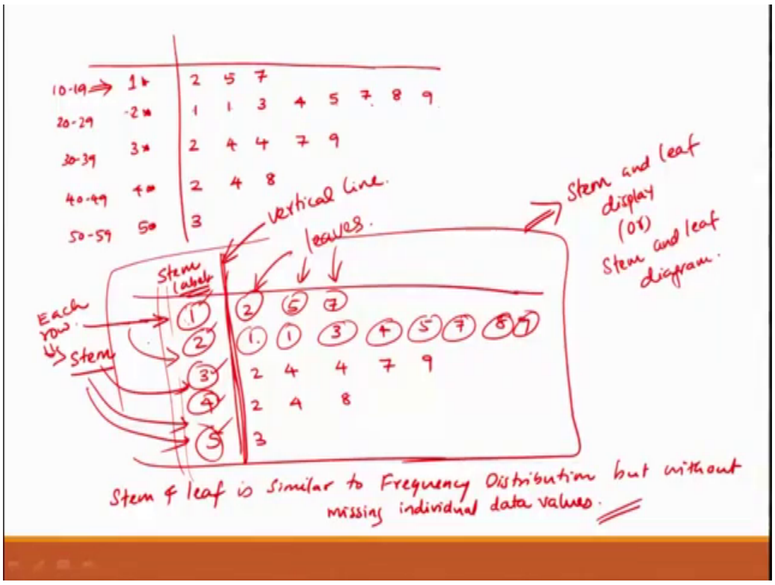
So, I will combine mark 3 here and the 52, 59, there is only one we already saw that 58th value. So, that me wrote it and the frequency will come to be this is 3, this is 8, this is 5, this is 3 and this is 1, and the total number of frequency will come to 5 plus 3; 8, 16, 19, 20. So, there are total 20 observations and then we know how to do the relative frequency and the relative frequency can be calculated by 3 out of 28 out of 25 out of 23 out of 21 out of 20 like this, and this will give you the individual relative frequencies and then we were using histogram to plot this. Now in this process in doing this is the area where we identify how many data values belong to a specific class and that is what this tally identifies for you.

So, in this process you are losing this information now assume that instead of doing this I revised this diagram in such a way that lets look at a scenario where I am having the same classes 10 to 19, then 22, 29, 32, 39, 42, 49 and 52, 59 let us say I have a classes like this and instead of the tally instead of the tally I have 3 values belonging to this in the 10 to 90, if I write only the last digits. So, remember the replace tally marks with last digit digits of the data if I do that then I can write it as this portion 10 to 90, if you go back, I can see that I have the values of 12, 15 and 70. So, the digits last digit is 2, 5 and 7. So, then I can write at as here as 2, 5 and 7 ok. So, this is the last 3 digits. So, instead of the 3 tally these 3 tally, we are replacing it with the last 3 digits of the data value. So, now, let us look at 20 to 29, if you look into that then we can find out that lets go back and take a look into it we have values of 21, 21, 23, 24, 25, 27, 28 and 29.

So, you have digits 1, 1, 3, 4, 5 and 7; 7, 8 and 9 ok. So, using that let us try to write that digits right here which is one 21 the next one is 21, then we have 23, then we have 24, then 25, 27, 28 and 29, ok, if the. So, there is 1, 2, 3, 4, 5, 6, 7, 8 these are the same 8 values right here, but the last digits of it now if you look at the 32, 39, class we have what we call as the 32, 34, 34, 37, 39. So, if you have to 4, 4, 7 and 9 are the last digits, ok, if we do that then we have the last 5 digits will be 2, 4 and 4 and 7 and 9 ok, similarly 42, 49, we go back we will see that 42, 44 and 48, 3 values last digits 2, 4 and 8 are available here.

So, we will write here the last as 2 of 4 and 8 and then we go back and take a look into the last data and we find that in 50 the last digit is 53. So, 3 is the last digit. So, we come back and write the last digit as 3 ok. So, this type of a display where the tally marks are replaced with the last digits of the data such a display ok. Now we can say that these are the digits. So, now, we can say 10 to 19 within this the last digits are 2, 5 and 7. So, the value that is important to us is the tenth position which is the one value. So, you can easily replace this instead of this 10 or 19, you can write it by saying that fine I will change the diagram instead of the classes I will write it as the first 10 to 19, I can write it as one star whatever it is then 20 to 29, I will write it as 2 star the 30 to 39, I will write 3 star and 40 to 49, I will write it as 4 star and 50 to 59, I will write it as 5 star like this and then I can write the digits right here I can do that part. So, let us see how I can translate transform this current modified version. So, this we can say that modify modified frequency distribution with individual values if that is the case, then let us see how we can do in that particular format that we were talking about. So, I will end up doing it this particular fashion we can have the 1 star which is the 2 star, 3 star, 4 star and 5 star.

(Refer Slide Time: 21:33)



10-19 →	1*	2 5 7
20-29	2*	1 1 3 4 5 7 8 9
30-39	3*	2 4 4 7 9
40-49	4*	2 4 8
50-59	5*	3

1	2 5 7
2	1 1 3 4 5 7 8 9
3	2 4 4 7 9
4	2 4 8
5	3

Stem & leaf is similar to Frequency Distribution but without missing individual data values.

So, which means this 1 star means this is the 10 to 19 this is the 20 to 29 this is the 30 to 39, 40 to 49 and 50 to 59. So, what we are saying here is that this is represented at the digitize is important to assist the one ok. So, when I write here it as basically 2, 5 and 7 then; that means, the values here are the thee values in this 10 to 19 class or the with the 10s digit as 1 we have 12, 15 and 17 ok.

Similarly, if you look at the previous case you will see that one; 1, 3, 4, 5, 7, 8, 9 is the next one ok. So, we write it the same way. So, it is one 1, 3, 4, 5, 7, 8, 9. So, this means these values are 21, 21, 23, 24, 25, 27, 28 and 29. So, you can see that you are not losing any of the information as part of this similarly 32, 39 we had 32, 34, 34, 37, and 39 that 5 values.

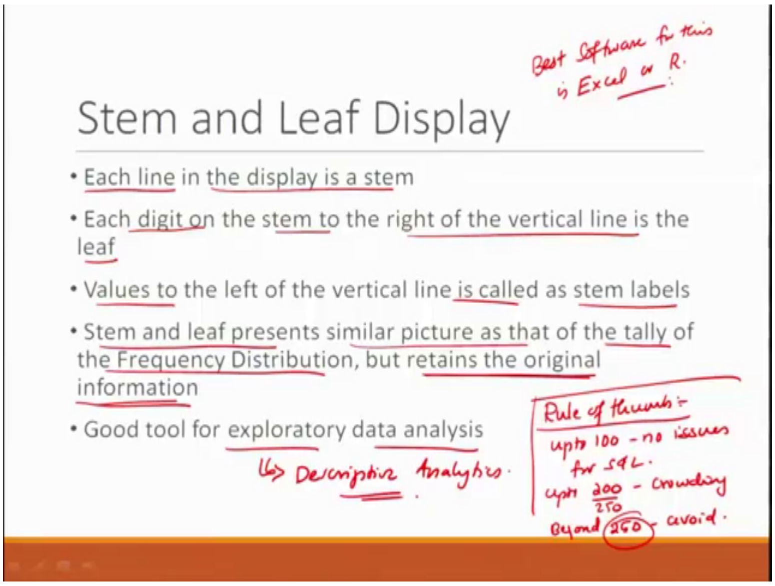
So, you can say that the units digit 2, 4, 4, 7, 9 gives you the last values and in the same way 40 is 40 to 44 and 48 and 50 is 53 ok. So, it gives you a similar appearance. So, here at the if you look at the previous diagram you can kind of see that ok, this is how the a data was behaving in this case and think about it is a rough case rough system and you can kind of say that you know a similar pattern is also observed here.

So, some amount of pattern without losing the significance of the system can be easily obtained or easily visualized in a system like this and you can also say that instead of using these stars you can replace the whole system with the help of without using any star you can say that 1, 2, 3 4 and 5 are the digits the 10 digits and you have your 2, 5 and 7; 1, 1, 3, 4, 5, 7, 8, 9 ok, then 2, 4, 4, 7, 9 and 2, 4, 8 and 3, this is what you can call as the this diagram that I just drew. Now this is called as the stem and leaf display or stem and leaf diagram ok.

So, this diagram stem and leaf is similar to frequency distribution, but without missing individual data values it will be tricky for us to do the frequency I mean relative frequency and cumulative frequency in those kinds of things because you have to then again count these values and put that other thing it could be confusing. So, people already draw up to this much ok. So, these individual draws each rows each row this is called as the stem ok. So, the stems these are all stems the; this is stem they are all stems individual values here in each stem these are the leaves ok. So, if you think about this as a tree a trunk of a tree and then these are the branches or the leaves that are coming out of each of the stem then you can call these as the leaves of the diagram ok.

So, this kind of a display is what we call as the stem and leaf display.

(Refer Slide Time: 26:17)



Best Software for this is Excel or R.

Stem and Leaf Display

- Each line in the display is a stem
- Each digit on the stem to the right of the vertical line is the leaf
- Values to the left of the vertical line is called as stem labels
- Stem and leaf presents similar picture as that of the tally of the Frequency Distribution, but retains the original information
- Good tool for exploratory data analysis

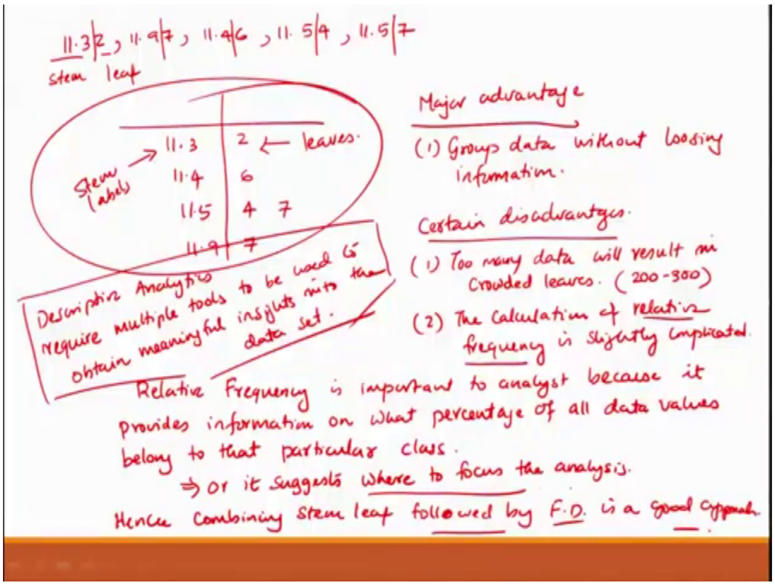
↳ Descriptive Analytics.

Rule of thumb :-
upto 100 - no issues for S&L.
upto 200 / 250 - crowding
Beyond 250 - avoid.

And each line in the display as we said earlier each line in the display is a stem ok. So, if we go back then we know that these are these temp each line in the display this is system now each digit on the stem to the right of the vertical line is the leaf each stem to the right of the vertical line is the leaf. So, each stem to the right of the vertical line this is the vertical line vertical line the right of it each one of these individual values is called as the leaves. So, a value to the left of the vertical line is also called as a stem levels ok. So, the right side is the leaves the left side is called as the stem labels. So, if you go back you will see that these ones these are the stems and each one of these are these individual values are the stem labels ok

So, one is the stem labor for the first row kind of thing. Now the stem and leaf what it does is it presents a similar picture as that of the tally of the frequency distribution, but retains the original information ok. So, it is a lot of similarity as like that of the frequency distribution, but it retains the original information that is the biggest difference of this and it is a good tool for exploratory data analysis, it allows you to analyze the data or this is a good to this is also called as descriptive analytics, we are able to describe the data without losing the plot ok, now there are few things that we need to think about so; obviously, people will ask a question.

(Refer Slide Time: 28:17)



Major advantage

(1) Groups data without loosing information.

Certain disadvantages.

(1) Too many data will result in Crowded leaves. (200-300)

(2) The Calculation of relative frequency is slightly complicated.

Descriptive Analytics require multiple tools to be used to obtain meaningful insights into the data set.

Relative Frequency is important to analyst because it provides information on what percentage of all data values belong to that particular class.

⇒ or it suggests where to focus the analysis.

Hence Combining Stem leaf followed by F.D. is a good approach.

If you have a data called 11.32, then 11.97 or something like this and 11.46 how would you make a stem and leaf kind of a thing; obviously, if you want you can think about it does this much being the; you know stem and the leaf.

So, this part is the stem this part becomes the leaf kind of a thing. So, then you can have a diagram in which you can say that it is 11.3, 11.4, 11.9 that kind of a thing in which 11.3 you will have two 11.4 you will have 6; 11.9, you will have 7 and if you get a new value called 11.54 or something like that then the question will be like it will be a then you will change you will basically saying that fine.

So, there will be you will split again 11.5 or 4 and then 11.9 of 7 and then you got an x value called 11.5 7. So, then that will get added right here 7 like this. So, in a system where you have a stem and leaf plot going on; so as I said earlier these are your stem labels and these are your leaves and the system where you are getting data continuously coming in you can use this mechanism to display your data without losing much of the result much of the information there are few ah.

So, the major advantage the major advantage is group data without losing information ok; however, there are some disadvantages to the system also certain disadvantages number one too many data will result in crowded leaves since you are putting the last digit values and you keep on putting it for a large set of values. So, typically this number is you know couple 100, 200 to 300 data values above which it starts looking really crowded.

Second thing is the calculation of relative frequency; frequency is slightly complicated the advantage of yd y uses relative frequency importance is because it also tells you relative frequency is important to analyst because it provides information on what percentage of all data values belong to that particular class or it suggests where to focus the analysis that is the most important aspect of the relative frequency and calculation of the relative frequency is not very straightforward when you have a diagram like this it is much more easier to do that in the case of a original frequency distribution.

But if you combine. So, hence combining stem leaf followed by frequency distribution is a good approach. So, first maybe do a stem and leaf diagram and after that you follow it up with a frequency distribution which in a way will allow you to analyze the data better. So, hence this brings us an important point descriptive analytics require multiple tools to be used to obtain meaningful insights into the data set. So, the is an important thing that you need to remember because in if you use just one tool you are pretty sure that it is no one single descriptive analytics tool that completely describes the data.

So, you will have to use multiple tools in Tandaman Dhorne cascading fashion to understand different nuances of the data. So, now, we have studied stem and leaf and we also see in frequency distribution and how they are connected to each other and how to even make a stem and leaf kind of a system and typically these things I will again recommend that you can use a Microsoft excel or something because the rule of thumb in this case is this rule of thumb for practitioners is up to 100, no issues up to 100 data points no issues for s and l stem and leaf up to 200 crowding 200 or 250 beyond 250 avoid ok.

So, up to 200 or 250 data points you can think about making stem and leaf, but beyond 20 50 data points avoid reusing stem and leaf because it will just defeat the purpose of the data will look. So, crowded that you would not be able to make any proper identification out of that data ok. So, with this what we will do today is we will conclude todays this short session on how to make the stem and leaf, but before concluding I will also say that the best software for this is excel only excel or r either one of them will do a good job of it, but lot of these things you can do it by yourself in excel and I would recommend you to study stem and leaf using excel as such ok.

And in the next class, what we will do is we will get into a little bit more complicated tools like box and whisker plot and how to do that using r and the r cod associated with it that will also be shown in the class and then from there we look into scatter diagram and then regression and those kind of things ok. So, we have already getting into different tools and analyzing this why I really hope that you guys are doing your homework and working on the problems and trying to solve this and bringing the skills to yourself because remember in analytics it is a skill that matters to you understanding of the technique and the skill how to use this technique is most important to you um. So, until the next class, it is thank you from my side and wish you all a good learning.

Thank you very much.