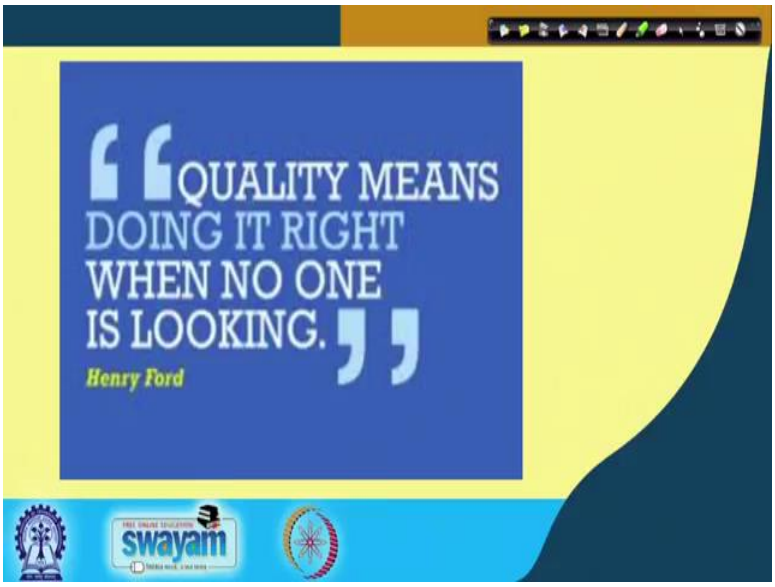
Six Sigma
Prof. Jitesh J Thakkar
Department of Industrial and Systems Engineering
Indian Institute of Technology, Kharagpur

Lecture – 19
Data Collection and Summarization (Part 2)

Hello friends, I welcome you to the part two of Data Collection and Summarization lecture number 19. We have already initiated our discussion on the topic of data collection and summarization in lecture 18 and if you recall then, we had seen couple of concepts like sampling population, categories of data and others.

(Refer Slide Time: 00:45)



Before we go into the details of data collection and summarization part 2; let us appreciate one very good quote on quality given by Henry Ford, a well known entrepreneur and a passionate person for building the quality into the product. So, Henry Ford says that quality means doing it right when no one is looking. There is lot of hidden meaning if you think it peacefully that I want to do something accurately, something in an appropriate manner. And it is not important that whether there is a supervisor or manager. So, when this kind of passion, this kind of attitude you cultivate in the people then they really aspire for achieving the quality in the organization. So, this quote has lot of meaning.

(Refer Slide Time: 01:55)

We had seen the variables and measurement scales in lecture 18, data collection method population and sampling. And we have also say seen the relative advantages and disadvantages of the various sampling methods like random sampling, systematic sampling, stratified sampling and cluster sampling.

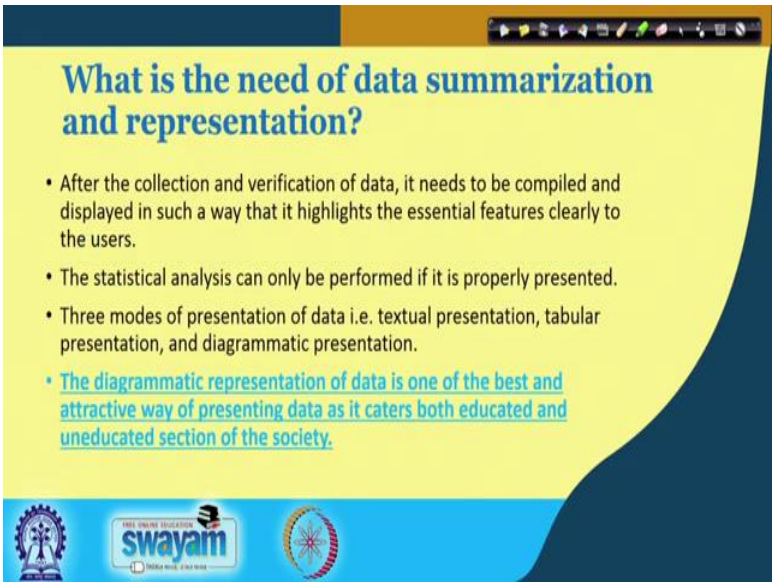
(Refer Slide Time: 02:21)

Now, this particular lecture part 2 on data collection and summarization, we will try to first understand: what is the need of data representation and summarization. Subsequently, this lecture will help you to appreciate various graphical methods; like

histograms and bar chart, frequency distribution, stem and leaf plots, box and whisker plots and Pareto diagram.

So, these are very well known and important methods for representing the data for summarizing the data, so that you can really investigate the phenomena in greater detail and reveal the causes or root causes behind the particular problem. So, data summarization and representation, what is the need?

(Refer Slide Time: 03:23)

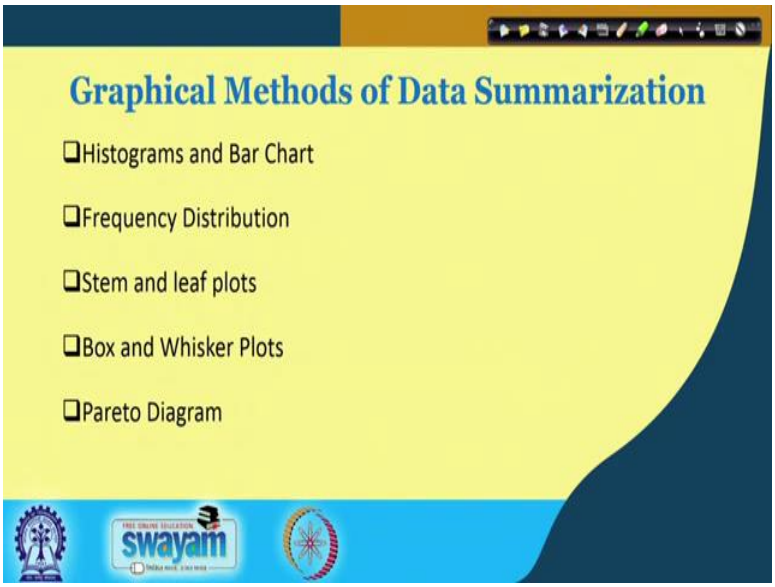


So, I would say that after you have collected the data and verified the data, you need to compile it. If you do not present in properly, then you will simply get lost. For example, I took a sample, random sample and I just collected the data on height of the people in eastern zone of India and I have all the data available. Now what to do? So, if you just have the data, it does not solve the purpose unless it is properly summarized and represented.

So, you not only gain the insights, but you can conduct the statistical analysis which is one step ahead, if your data is properly presented. So now, when you have to present the data, there are various formats you can present it in a textual form. You can use the tabular presentation or you can have diagrammatic presentation. Out of this the diagrammatic presentation of the data, is one of the best and attractive way of presenting data as it caters both educated and uneducated section of the society.

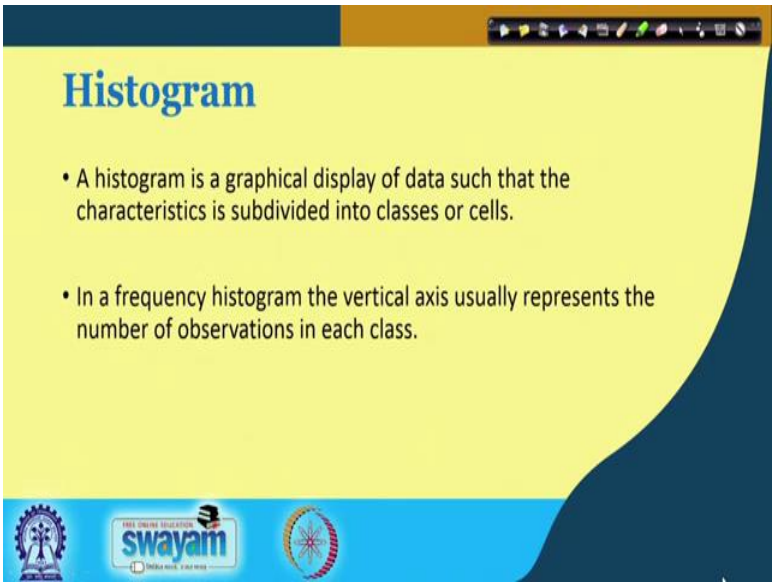
So, when something is presented through bar chart histogram, people can easily see, and without much botheration, they can understand that what is the problem or what exactly this particular situation is describing, this particular graph is trying to convey.

(Refer Slide Time: 05:10)



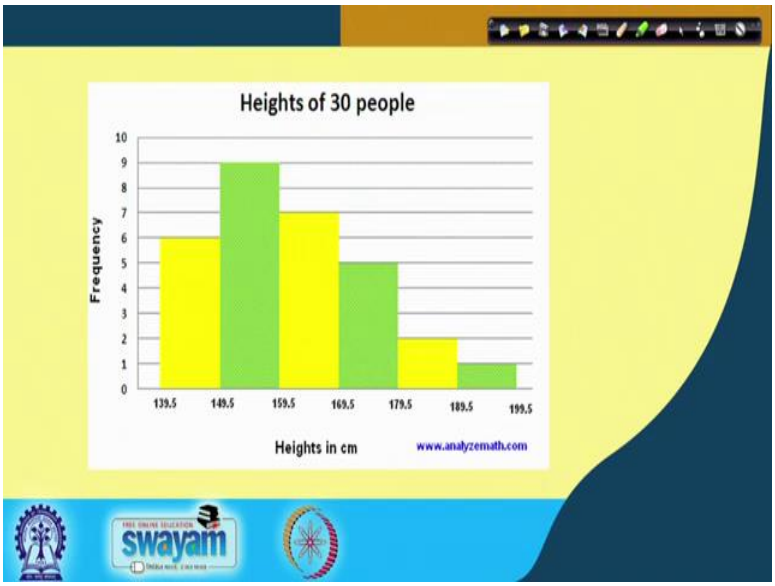
So, with this understanding, we will try to appreciate the various graphical methods of data summarization, histogram and bar chart, frequency distribution, stem and leaf plots box and whisker Plots and Pareto diagram. Let us begin with Histogram and Bar chart.

(Refer Slide Time: 05:28)



So, histogram we all know and you must have plotted many times. So, typically it is a graphical display of data such that the characteristic is sub divided into classes or cells. In a frequency histogram typically you have vertical axis which represents the number of observations in each particular class. So, we will see just some examples also.

(Refer Slide Time: 05:55)



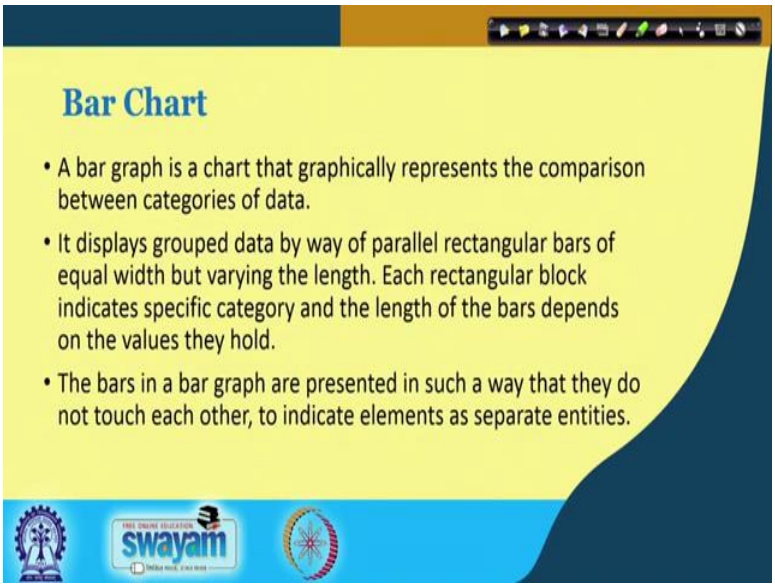
Let us see this is the data I have about heights of 30 people. And on vertical axis as I mentioned, you can see the frequency and x axis, you have the heights in centimeter. In a logical way, I will tell you how it is. I have created the intervals and this interval basically will capture the number of people falling into this range. So, here my typically range is 139.5 to 149.5 then 149.5 to 159.5 and so on. So, an increment of 10 is set. Now, when you have to classify this 30 people for their height, you can say that, ok 6 people they are falling in the interval of 139.5 to 149.5 height in centimeter and similar way. So, this graphical display immediately will tell you that fine what is the interval in which maximum number of people they fall.

So, this can have a lot of value. Suppose, you are manufacturing the garments, suppose you want to design the door or many other things, as you can understand; this kind of distribution, this kind of data can tell you that what would be an appropriate dimension for your design.

(Refer Slide Time: 07:34)

So, now the question comes; is there a difference between histogram and bar chart? So, many a times you get confused that, fine both they display some kind of columns and they are same, but there is some difference.

(Refer Slide Time: 07:53)

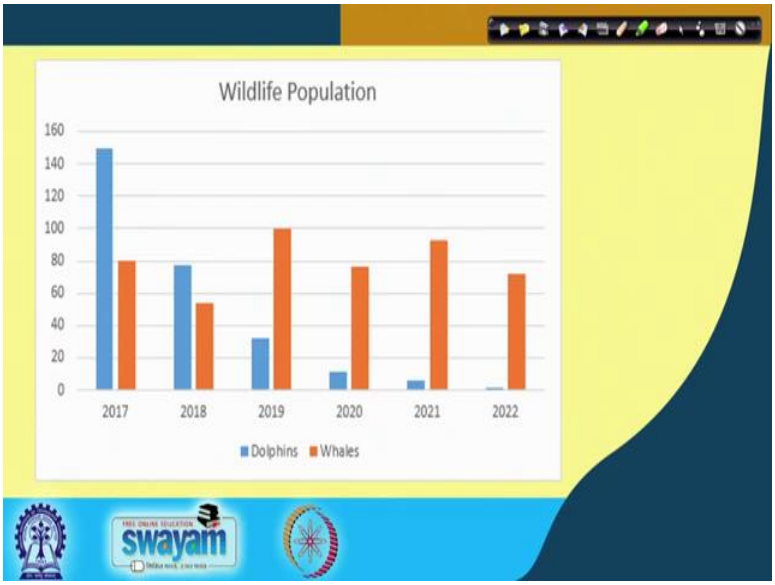


So, when you say bar chart, it is typically a chart that graphically represents the comparison between categories of data. So, if my purpose is to compare the categories of data, then I will make use of bar chart. But, if my purpose is just to identify the

frequency falling into a particular range then the appropriate representation is the histogram.

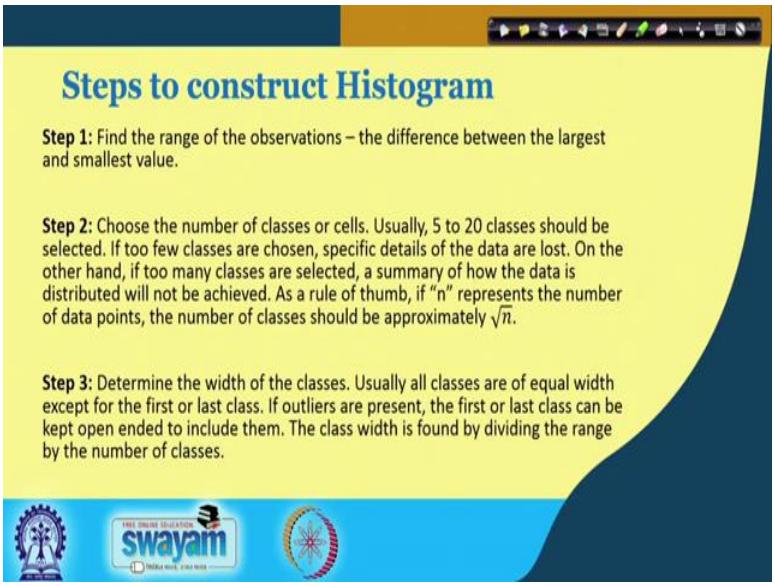
So, it displays group data by way of parallel rectangular bars of equal width, but varying length. And each rectangular block indicates specific category and the length of the bars depends on the values they hold. So, the bars in the bar graph are presented in such a way that they do not touch each other. You have seen in case of histogram that they touch each other and there is the continuous, say formation of the class. In this case, it is not important or it should not touch each other.

(Refer Slide Time: 09:03)



So, if we see the example here there is a comparison of wild life population and I am considering dolphins and whales. So, I just want to compare and also may be, also I want to project, then you will say in 2017 (Refer Time: 09:22) blue color means dolphins and orange is my whales. So, this is the population, may be in thousands and same way 2018, 2019, 2020 and so on. So, this display just helps me to compare two different wildlife populations. And the representation is through rectangular bars placed side by side represented with different color.

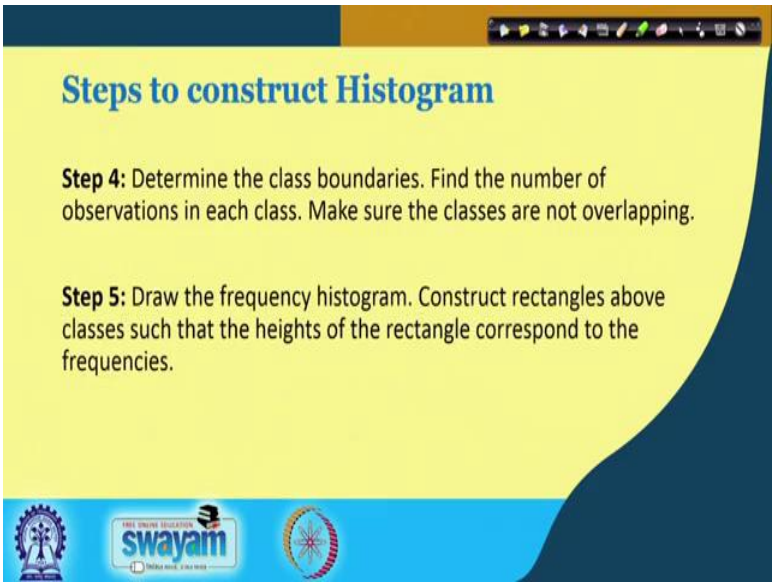
(Refer Slide Time: 09:54)



Now, if we see the steps involved typically in histogram, I will go line by line and help you to appreciate. Step 1: Find the range of the observation. So, typically this is the difference between the largest and the smallest value. Step 2: Choose the number of class or cells. Usually, it should be 5 to 20 classes. If too few class you choose, then specific details of the data are lost. And on the other side, if it is too many classes then a summary of how the data is distributed typically that is our interest will not be properly captured or achieved.

So, as a rule of thumb, if n represents the number of data points, suppose you have collected 50 data, 100 data, then the number of classes should be approximately square root of n. Now depending upon your convenience, you can little bit manipulate plus and minus, but usually it should be $\sqrt{n}$. Step 3: Determine the width of the classes so, you have to decide the interval. And usually, we consider that all columns or classes are of equal width except for the first or last class, because you have to begin and terminate at a particular value. If some outliers are present the first and last class can kept open ended to include them. And the class width is found by dividing the range by the number of classes.

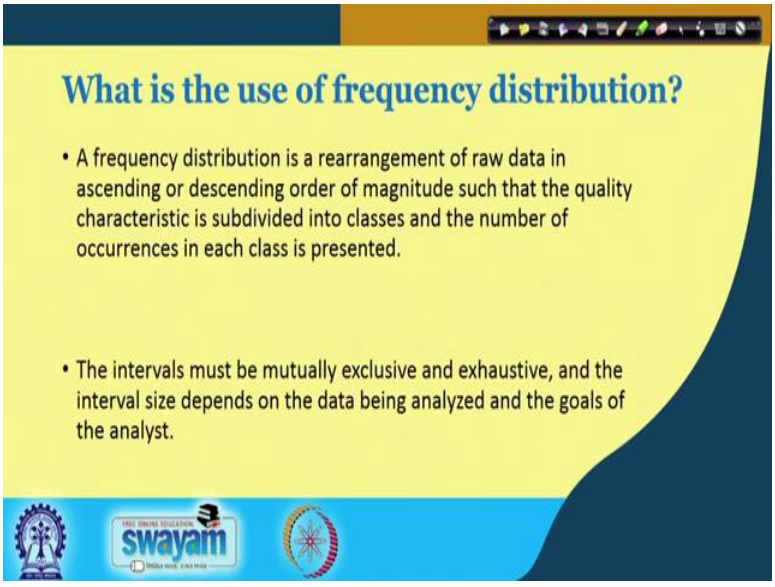
(Refer Slide Time: 11:35)



Step 4: Determine the class boundaries. So, find the number of observations in each class. Make sure the classes are not overlapping. And finally, step 5: Draw the frequency histogram. Construct rectangles above classes such that the heights of the rectangle corresponding to the frequencies.

So, that is the top part your rectangle and typically this will tell you that, what is the frequency, how many are falling in a particular class. The frequency distribution is another way of representing the data.

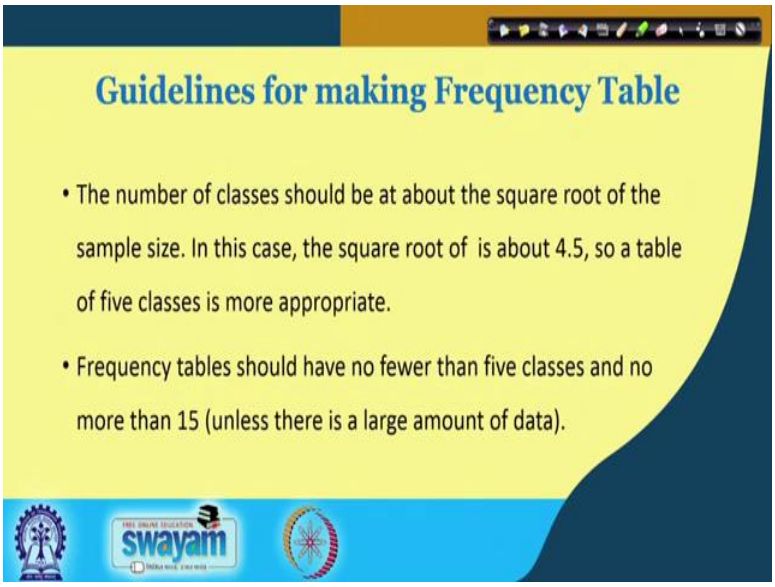
(Refer Slide Time: 12:14)



So, typically it is a rearrangement of the data. Many a times I am interested to see observe or analyze a particular phenomenon happening and which particular class this is happening more number of times then I will make use of frequency distribution. And it rearranges the data in ascending or descending order of magnitude such that the quality characteristic sub divided into classes and the number of occurrences in each class is presented.

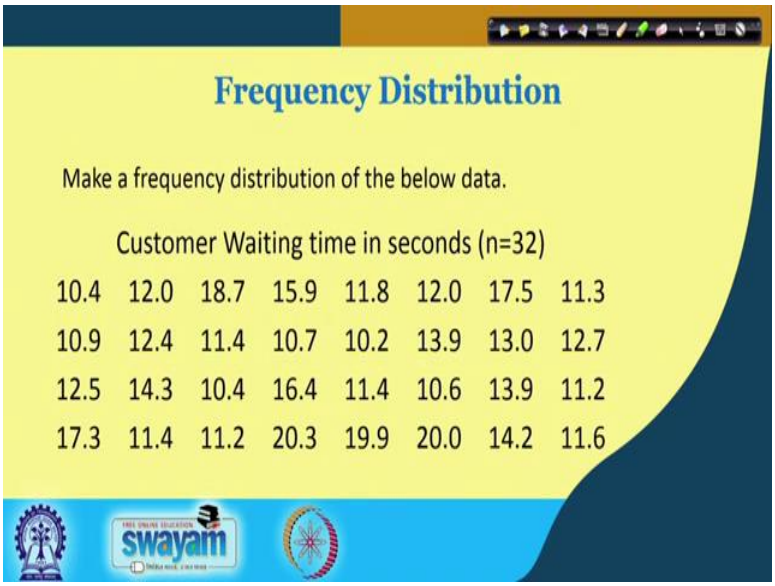
So, here the interval must be mutually exclusive and exhaustive, there should not be overlapping. And the interval size depends on the data being analyzed and the goals of the analyzed. So, not necessary that you will have equal size interval, but it depends upon the data being analyzed and the goal of the analyst, what exactly here wants to investigate.

(Refer Slide Time: 13:18)



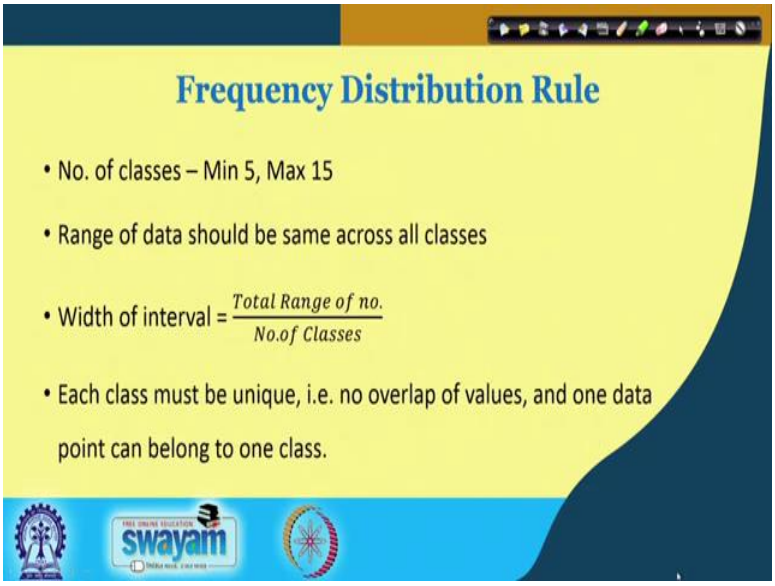
There are some guidelines for making frequency table, the number of classes should be at about square root of the sample size. In this case, the square root is about 4.5. So, a table of five classes is more appropriate. And frequency table should have no fewer than five classes and no more than 15 unless there is a large amount of data.

(Refer Slide Time: 13:48)



So, if you just see this particular example, I am just talking about. So, you have total 32 data set and if you see the number of classes square root of the sample size. So, if we just see the square root of 32 then maybe you can say that, let it be somewhere around say five classes or something like that.

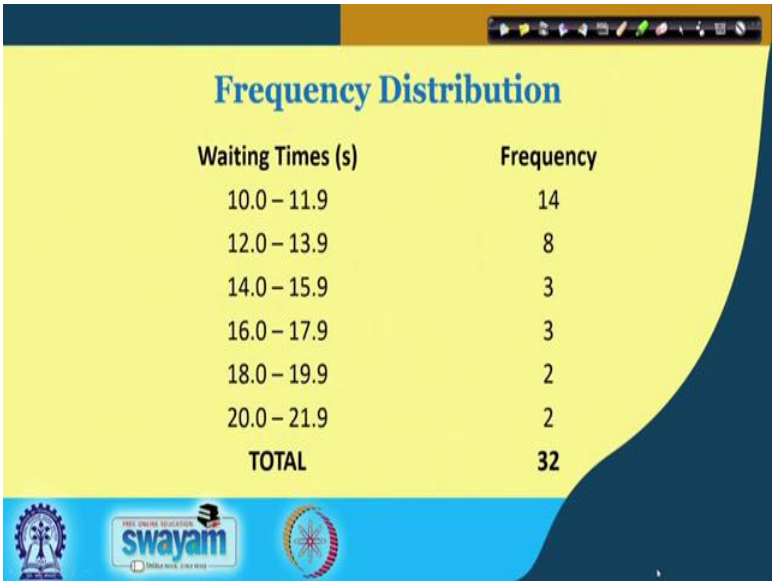
(Refer Slide Time: 14:20)



So, here number of classes, you will have minimum 5, maximum 15. And range of data should be same across all classes. And $Width\ of\ interval\ =\ \frac{Total\ Range\ of\ number}{No.\ of\ Classes}$.

So, each class must be unique and no overlap of values and one data point can belong to one class only.

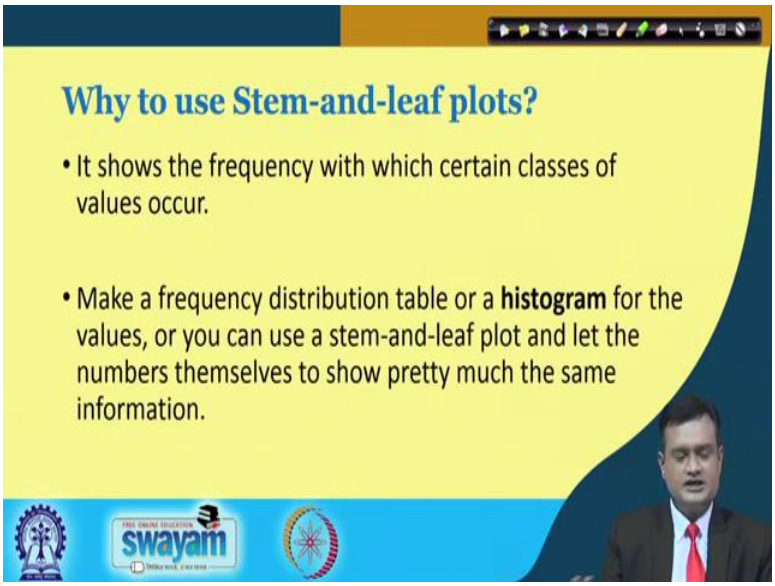
(Refer Slide Time: 14:46)



So, we have here divided the data of waiting time in 1, 2, 3, 4, 5, 6 classes. And you can see that 10 to 11.9, the frequency is 14, 12 to 13.9, the frequency is 8, 14 to 15.9, the frequency is 3. And what exactly, I can reveal that here if the waiting time is 10 to 11, then the number of people or customer falling in this particular range basically are 14. If the waiting time is 12 to 13.9, than the number of people falling in this range or the number of time, you can represent it in that way also basically 8 and so on.

So, this can help me to understand that what is the overall waiting time in my particular say company, my organization, service counter whatever may be the case. And based on that, I can take the corrective measures to improve upon my service time or reducing the waiting time. Another representation which is very much interesting and will reveal something additional compared to histogram and your frequency distribution. This representation is called stem and leaf plot.

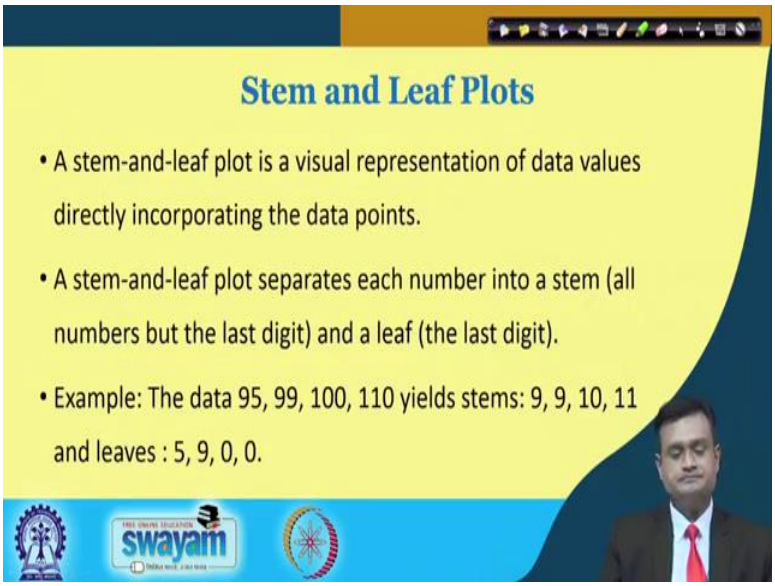
(Refer Slide Time: 16:21)



So, let us see what is the uniqueness of this. So, typically it shows the frequency with which certain classes of value occur. So, you have the waiting data, time data, you have let us say service time data. Now I want to study that how this particular data in terms of frequency with respect to various classes they occur.

So, you make a frequency distribution table or histogram for the values or you can use stem and leaf plot. And let the number themselves should show pretty much the same information. So, you can use multiple say, representation method and you should be able to reveal same phenomena, same observation. When you plot it in terms of histogram or frequency distribution or may be in terms of stem and leaf plot.

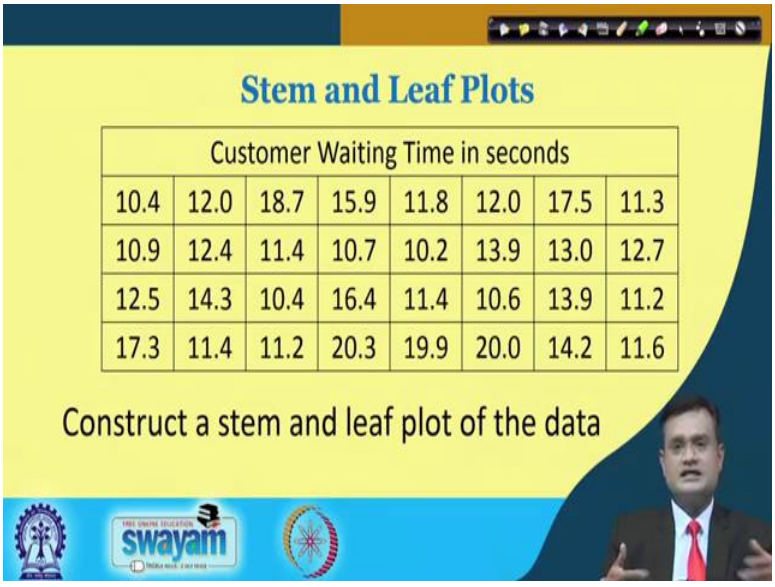
(Refer Slide Time: 17:19)



So, typically stem and leaf plot is visual representation of data values directly incorporating the data points. And it separates each number in to stem all numbers, but the last digit and the leaf the last digit. So, you have the data and these partitions, the data, particular data in to two stem and leaf stem is you will say all number, but not the last digit and leaf means the last digit.

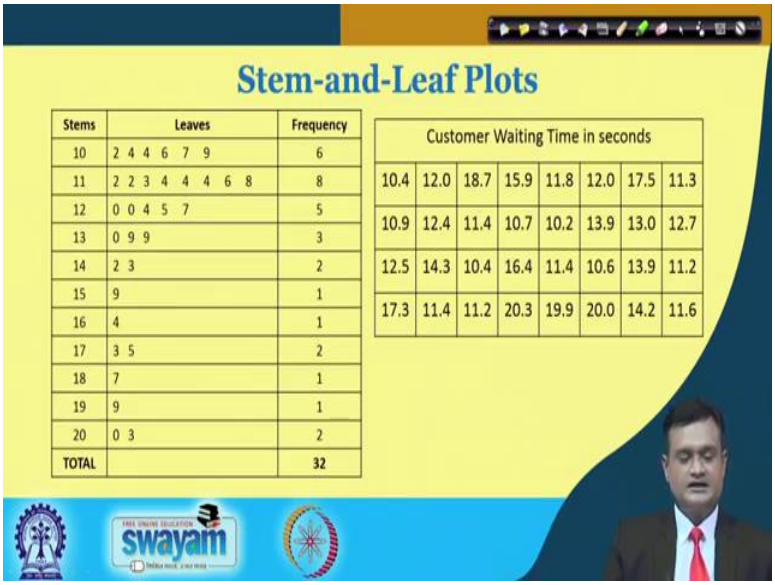
So, for example, you have data 95, 99,100 110, this typically will yield stems like 9, 9, 10 and 11. So, last digit I am just putting under leaves and this would be 5, 9, 0, 0.

(Refer Slide Time: 18:17)



We will see one more example where decimal values are included and this will again help us to appreciate that how we can reveal the data pattern, data prevailing in the particular set and what kind of analysis or conclusions we can make? So, for example, you have again customer waiting time in seconds and we are supposed to construct stem and leaf plot for this data set.

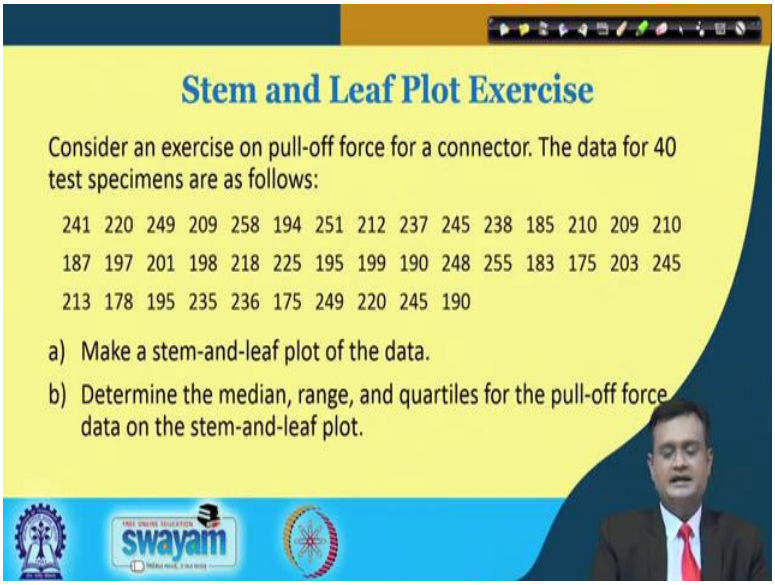
(Refer Slide Time: 18:48)



So, here you can see that your data is like 10.4, 12.0, 18.7, 15.9 and so, on. So, there are decimal values. What I am doing here, I have identified the first two digit or the number before decimal point. And I can say my stems are 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, and 20. Now, if there is a data which has 10.4, 10.2 something like that then leaf part I will put in the leaves column. So, 2, 4, 4, 6, 7, 9 first row you can say that you have 6 such reading which begins with 10 and has the value 2, 4, 4, 6, 7, 9 after decimal.

Similar way 11, 12, 13 and so on and this is what exactly we have done. So, you have the tabulated data, this is what you have, this is the tabulated data. And now you want to this is your tabulated data and you want to convert this into say frequency distribution through stem and leaf representation.

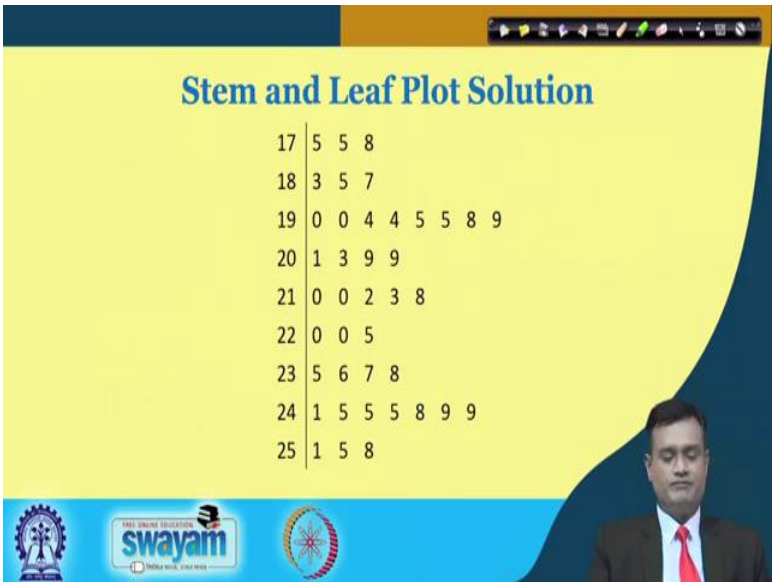
(Refer Slide Time: 20:23)



So, if we see then what happens here. If you just see the data, you will at least get an idea that how your data is distributed. And whether it follows a kind of normal distribution, the data is skewed or it is skewed on the right side, left side, positively skewed, negatively skewed such kind of observation, you can just make through this simple kind of representation called Stem- and- Leaf Plots.

Now, you can consider this exercise on pull off force and this is done for a particular connector. So, there is a pull off force and the data is collected for 40 specimens and you can see the numbers to 41, 220 and so, on. So, first we want to have a make a stem and leaf plot for this data and determine the median range and quartiles for the pull off force which is the phenomena under investigation on the stem and leaf plot.

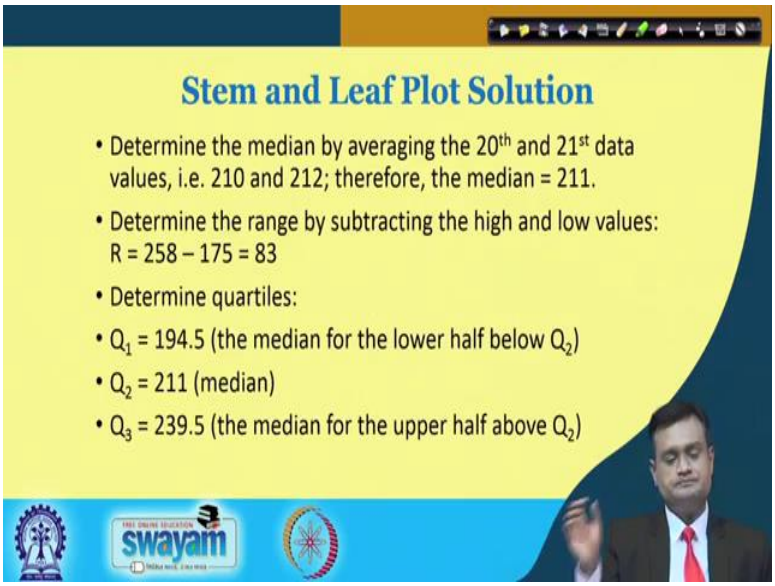
(Refer Slide Time: 21:35)



So, just see when I arrange the data then this particular representation, you will see that you have the values 175 means 17 stem is there, 183 means 18 stem is there. Similar way 190 means 19 stem is there. When I just rearrange the data as per my stem and leaf procedure then I can see that 17 there are three data point,18 there are three data point,19 there are 1, 2, 3, 4, 5, 6, 7 total eight data point, 20 four data point and so on.

So, I can say that yes maximum data is following in this particular range or the particular number is having the maximum number of data set. And, I can just get an overview or idea that how my data set is distributed.

(Refer Slide Time: 22:33)

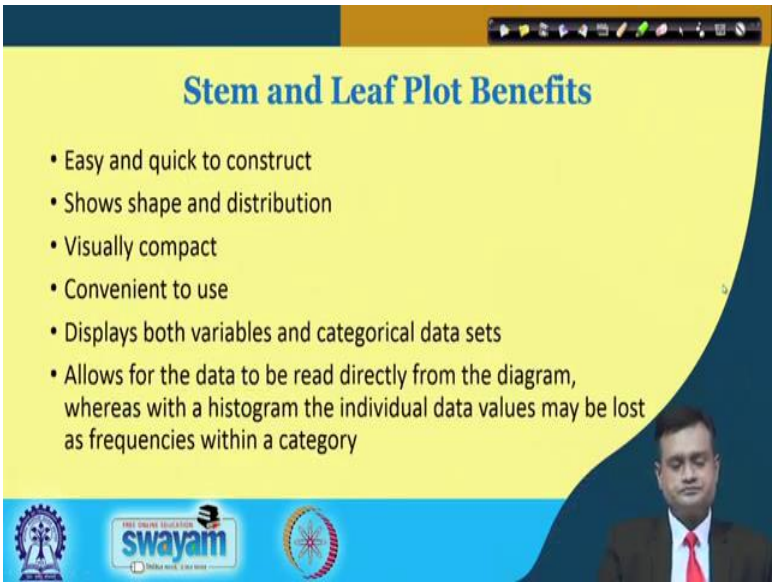


So, you can do little bit more analysis by going into detail and what you can do determine the median. So, you all know what is mode, median and mean. So, median is basically the middle value and here I you have two values, because there is no single middle value. So, you can take the average of 20th and 21st data; so, 210, 212 and your median is 211.

Now, you determine the range. So, you have the maximum value 258, minimum value 175 so, range is 83. Now I am interested in determining the quartiles so, that I can understand that quartile means 25 percent quarter. How much data is really falling in a particular quartile? So, Q_1 is 194.5, the median of the lower half of the Q_2, Q_2 is basically your median value that you have determined 211 and Q_3 is 239.5 the median of the upper half above Q_2.

So, please understand Q_1, Q_3 basically they are below or lower and upper half and whatever data set you have in the lower and upper half below Q_2 above Q_2. You consider that data set and find out the median value of that data set which is basically your Q_1 and Q_3 respectively.

(Refer Slide Time: 24:12)

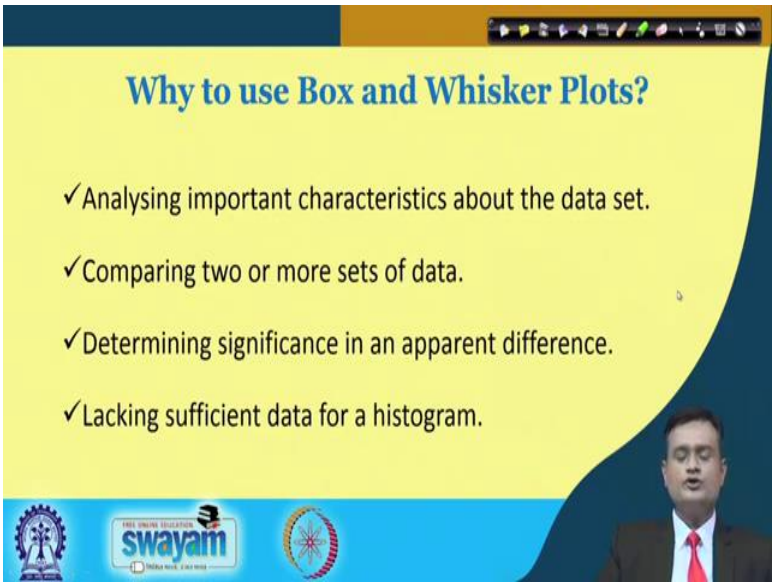


So, once this is done, you can very well understand how your data is distributed. And you can say that it is easy and quick to construct not much hustle is involved shows shape and distribution, as I mention how your data is distributed. Visually compact not much say representation is required, it is visually compact. So, you can grasp the entire information, situation with a very little say observation. And understand the problem convenient to use displays both variables as well as categorical data set, this is the unique feature and allows for the data to be read directly from the diagram whereas, with histogram the individual data value may be lost as frequency within a category.

So, here there is to and fro that even when you look at the stem and leaf representation, you can also read the data value which is not possible when you develop the histogram or frequency distribution.

The another representation which is also very well known is box and whisker plot. Let us try to appreciate what additional information; this data representation method can give us.

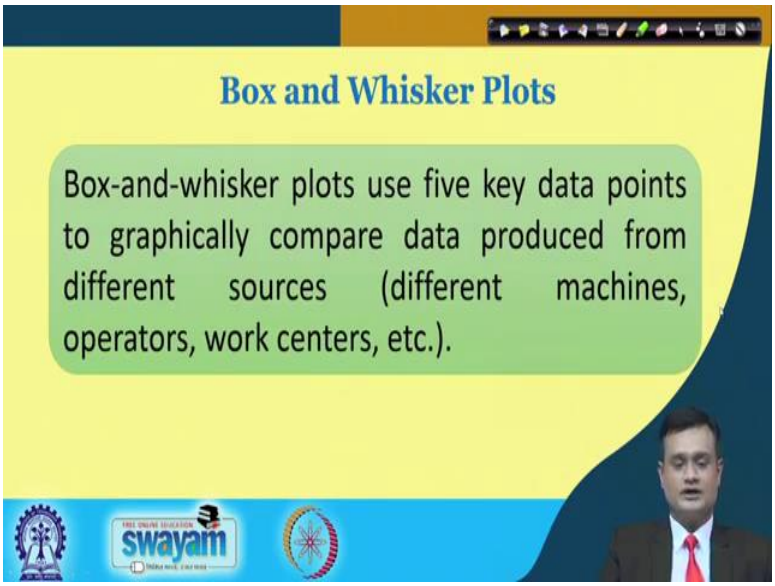
(Refer Slide Time: 25:37)



So, why to use box and whisker plot? When already, we have histogram, bar chart frequency distributions, stem and leaf plot. So, typical it helps us to analyze the important characteristic about the data set. And many a times, I am not interested only in one data set, I am also interested in comparing two data set. So, suppose, let us say I have a vehicle and it is giving me some mileage per liter, now I want to compare this vehicle with the other vehicle. So, you can or efficiency of one machine with respect to other machine.

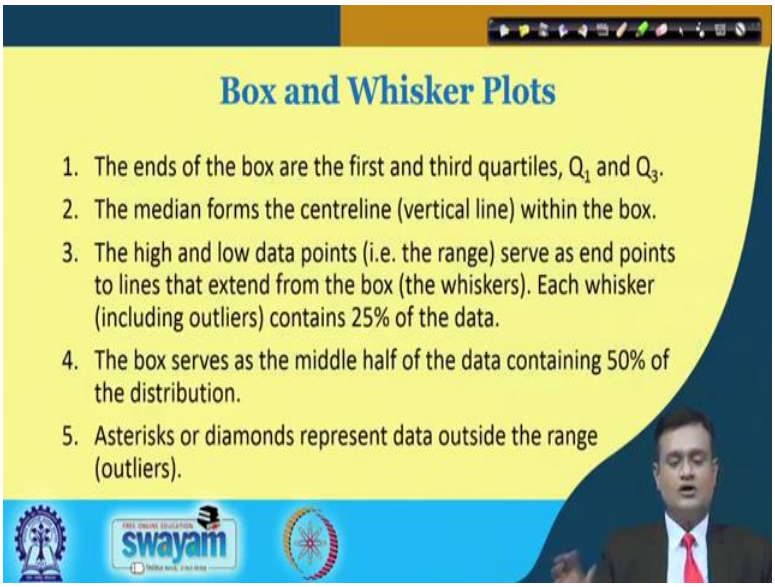
The productivity of one worker with respect to other worker or group of worker in such situation this kind of representation can help us to investigate deeper compared to the previous one. So, determining significance in an apparent difference and lacking sufficient data for a histogram.

(Refer Slide Time: 26:42)



So, box and whisker plot use five key data points to graphically compare data produced from different sources, different machines as I mentioned operators, work centers, you have an interest in comparing the two different units or maybe it may be three or more.

(Refer Slide Time: 27:01)

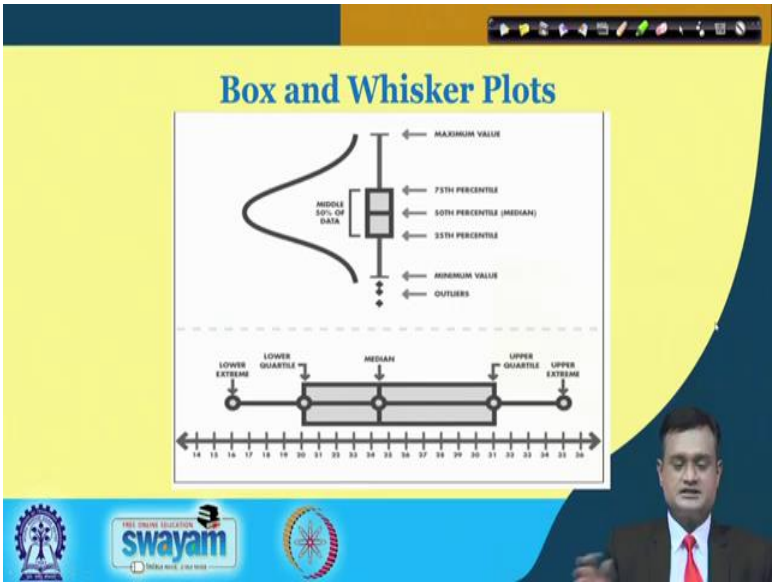


So, box and whisker plot there are certain features, let us try to appreciate. The ends of the box are the first and third quartiles, we have already seen the concept of quartile, quarter 25 percent. So, the ends of the box are the first and third quartile Q_1 and Q_3.

Second the median forms the center line, vertical line within the box. So, that is the median and typically, we call it as Q_2.

The high and low data points; that means, the range serves as end points to the line that extend from the box. The whiskers typically here the boxes are called whiskers. Each whisker including outlier contains 25 percent of the data point. Then the box serves us the middle half of the data containing 50 percent of the distribution. And asterisks or diamonds represent data outside the range and typically call it as outliers.

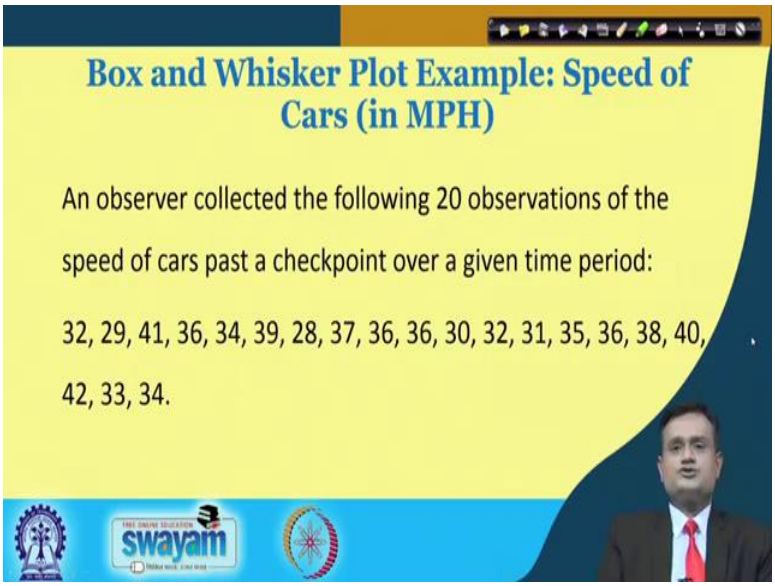
(Refer Slide Time: 28:11)



So, we will see the example that will make you more comfortable with this kind of data representation. Just see this example and what I have done here; you can see that if I use the pen then you can just focus on this. So, this is my 50 percent middle of the data. And what I say that the median value this is the median value, this is the median value is 50 percentile, 25 percent this side that is 75 percentile and 25th percentile.

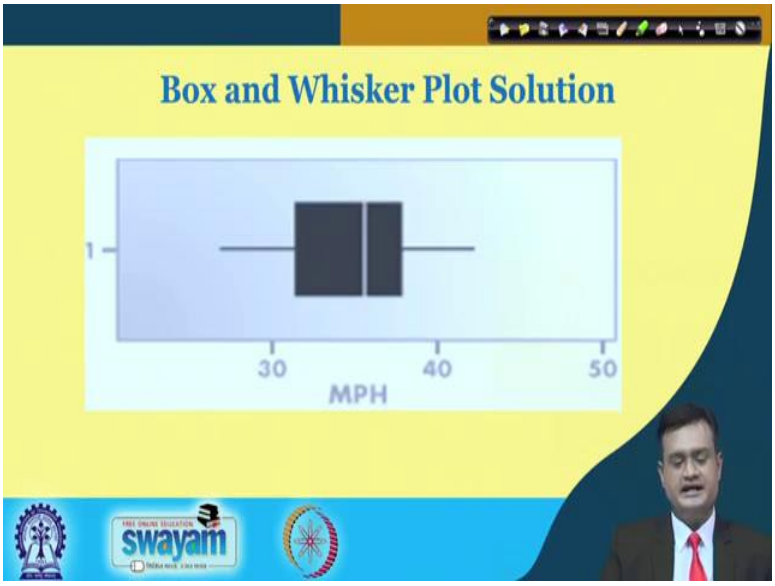
I am putting here minimum value and also I am putting here the end point. I am putting here maximum value with the end point and typically if it is distributed exactly like this then I would say that it follows the normal distribution. Now, I have outliers typically you can see here and this outliers I am just representing beyond the end points either at the maximum side or the minimum side. And if you see this particular representation which is shown here, this one then again there is lower quartile, there is upper quartile, then there is median and you can have lower extreme and the upper extreme.

(Refer Slide Time: 29:58)



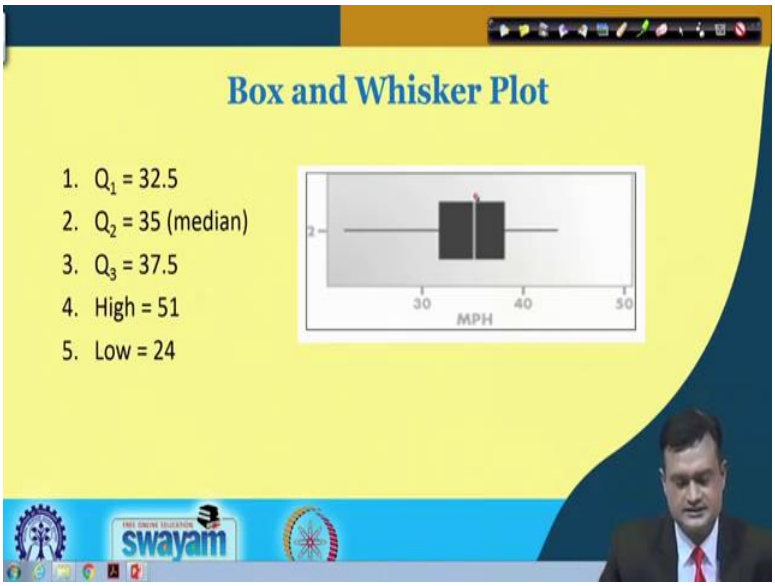
So, this representation really helps us to appreciate the distribution of the data set. Now, let us try to undertake one exercise that can help us better. An observer collected some 20 observations of the speed of cars past a check point over a given time period 32, 29, 41, 36 and so on.

(Refer Slide Time: 30:19)

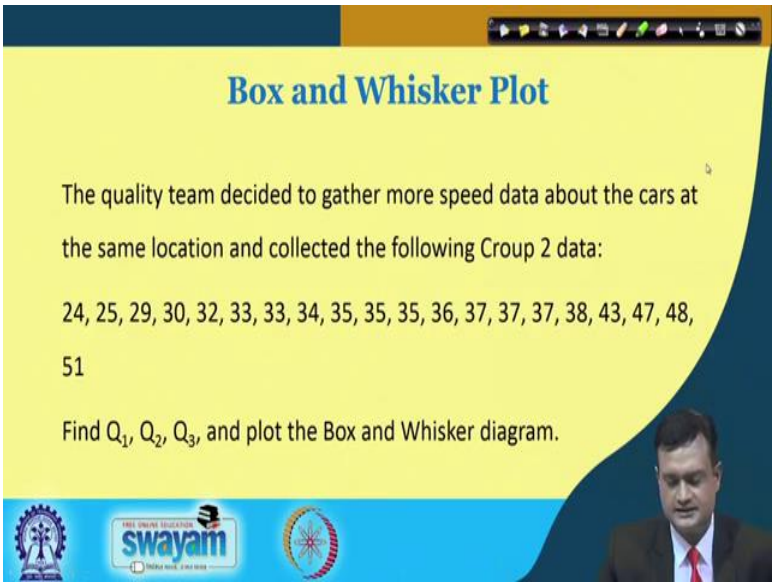


Now, MPH means miles per hour that is the speed data. And I am just putting it as I explained finding the mid and then putting the quartile wise data, I can see that my median which is shown here is somewhere here.

(Refer Slide Time: 30:38)



(Refer Slide Time: 30:46)

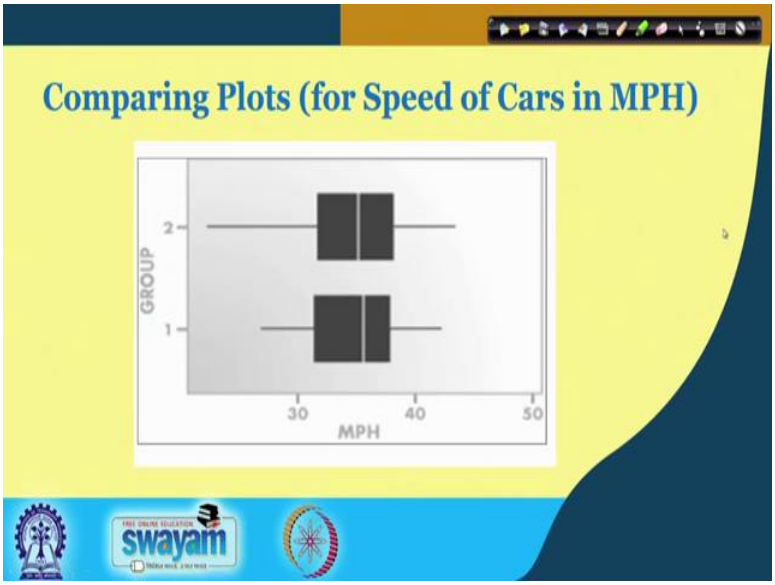


This is the data set and you can see here that Q_1 is 32.5. Then Q_2 is 35 that is the median, Q_3 is 37.5. So, I have put the median value here and 32.5, this is the Q_1, I have put here first quartile and Q_3 that is 37.5, I have put here high is 51 and low is 24. So, I have just displayed this data set through a box and whisker plot.

Now, the question is what to interpret. So, when I am not comparing it with any other just this will tell me that whether my data is normally distributed, it means exactly 25 percent, 25 percent on either side of the median. And the middle part is covering 50

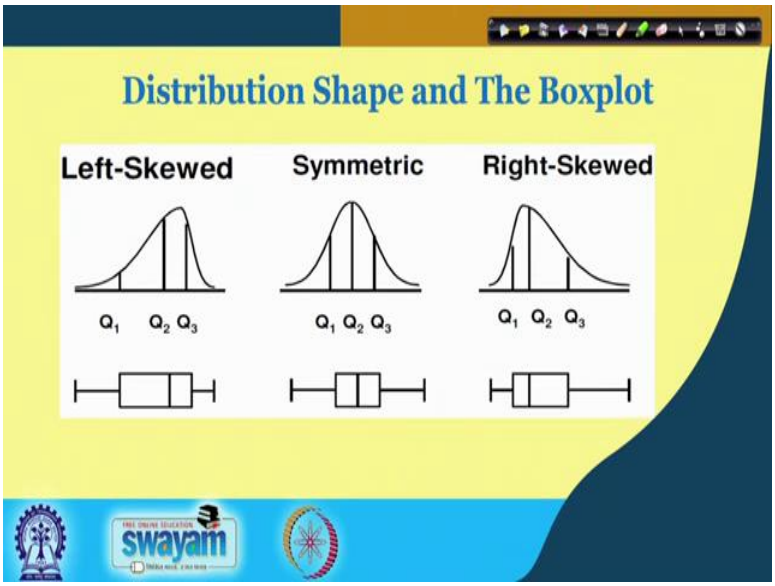
percent and also it helps me to understand that which side there is more data accumulation and what are the outliers.

(Refer Slide Time: 32:04)



So, now if we go little bit further then I can compare two plots may be of two different vehicles in terms of miles per hour. And you can see that clearly this will immediately help me to understand that the median values are different. And in the first plot more or less, say equal distribution is there on either side of the median, but in the second plot you will see that miles per hour on the left side, it is more means more data is there on the left side whereas, the right side the number of data point is less.

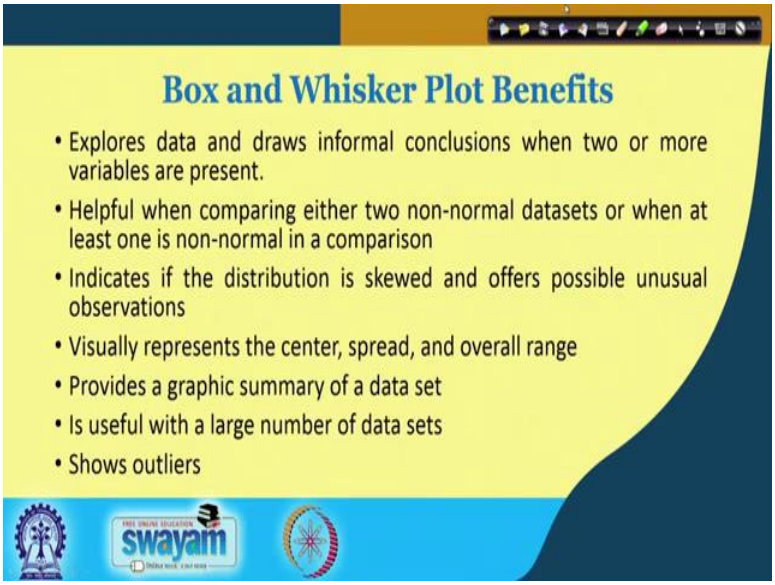
(Refer Slide Time: 32:51)



So, this helps me to compare the performance of two cars in terms of miles per hour by using the box and whisker plot. And also you can easily rectify that my distribution with respect to quartile as I mentioned. If it is exactly like this, then it is symmetric Q_1, Q_2 and Q_3, you have 50 percent, 25 percent, 25 percent on either side. And it is symmetric, but if you have the distribution like this then it is shifted on the left side which is called right skewed.

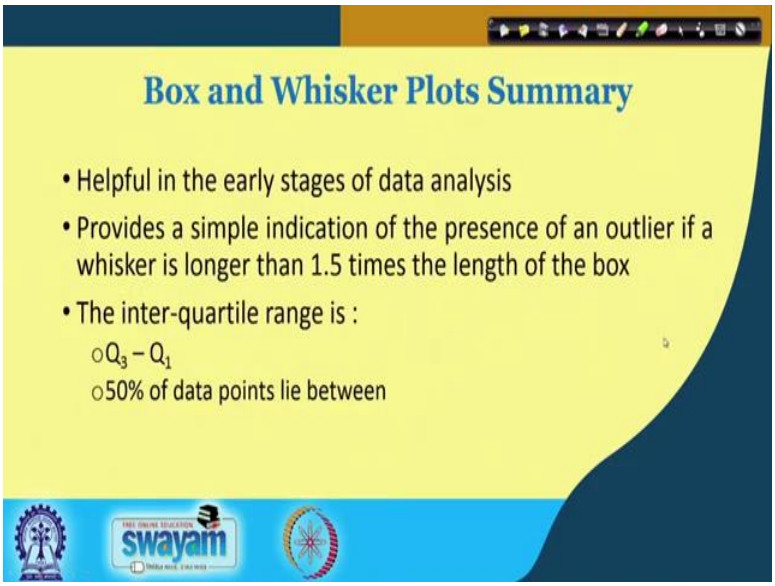
So, just it is opposite here it is shifted on the right side. So, it is called left skewed. And this helps me to understand how my data is distributed. We will subsequently see that what exactly it means, when I say how my data is distributed and what its importance in the statistical analysis, we will do it in the coming time.

(Refer Slide Time: 33:49)

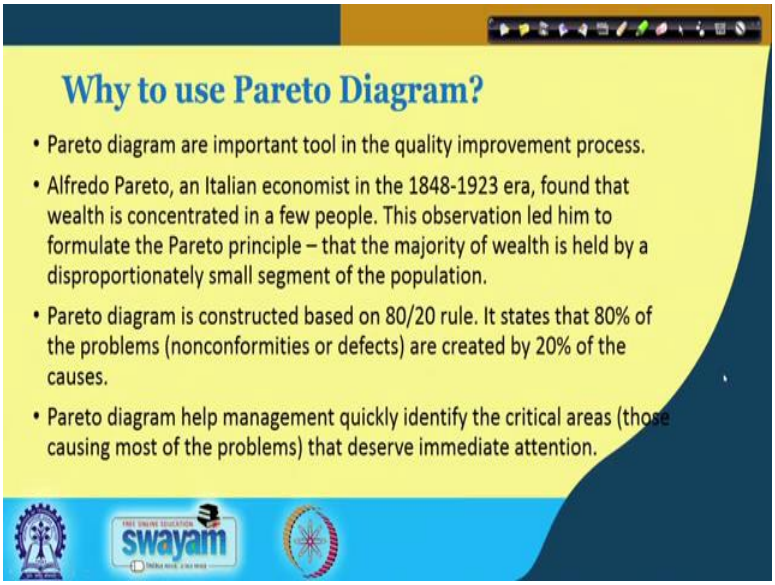


So, box and whisker plot typically, if we summarize the benefits, explores data and draws informal conclusions, when two or more variables are present helpful when comparing either two non-normal data sets or when at least one is non-normal in a comparison. And, typically indicates that if the distribution is skewed and offer possible say, unusual observation and it shows also the outliers. So, these are some of the benefits of my box and whisker plot. And you can have the interquartile range and so on.

(Refer Slide Time: 34:23)



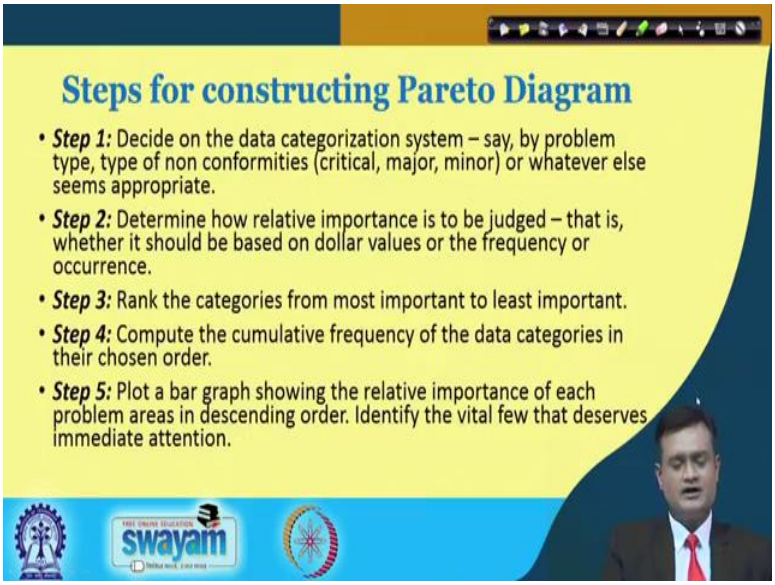
(Refer Slide Time: 34:29)



Now finally, let us try to see the most important representation that is Pareto diagram. And, why to use Pareto diagram; these particular representation is derived in the era of 1848 to 1923. And there was a renowned Italian economist named Alfredo Pareto what he observed that majority of the wealth is held by disproportionately small segment of the population. So, 80 percent of the wealth is possessed by 20 percent of the people and only 20 percent of the wealth is available for the 80 percent of the people.

So, based on this he thought that same applies to many other domains that 20 percent basically creates the maximum problem, 80 percent. And 80 percent of the issues, they only contribute to the 20 percent of the problem. So, typically if you can figure out the critical areas 20 percent creating maximum problem. Then you can apply the selective control put more resources on those problematic area and manage your situation well. So, thus Pareto diagram there are various simple steps to follow.

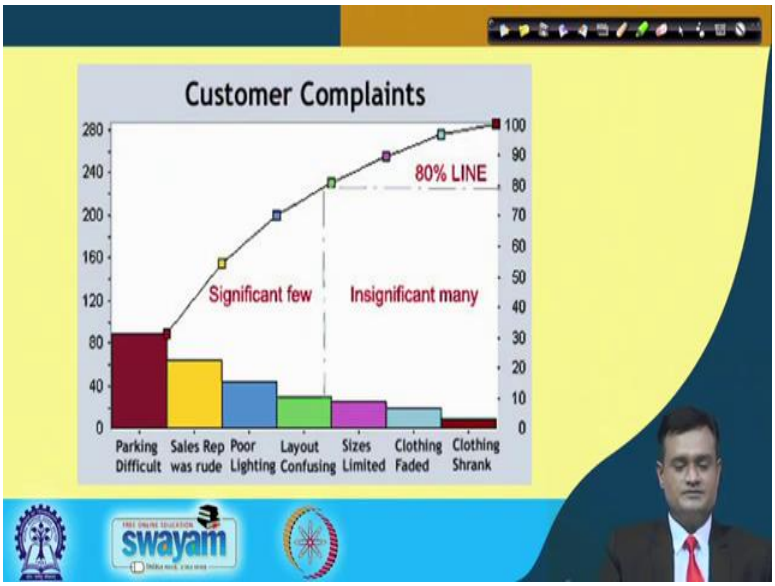
(Refer Slide Time: 35:52)



Decide on the data categorization system may be say by problem type, type of conformities, critical major minor or whatever else seems appropriate. Step 2: Determine how relative importance is to be judged that is whether it should be based on dollar values, frequency or occurrence. Step 3: You rank the categories from most important to least important.

Step 4: compute the cumulative frequency of the data category in their chosen order. And 5: Plot a bar graph typically showing the relative importance of each problem area in descending order. Identify the vital few that, deserves immediate attention means 20 percent of the problem which causes 80 percent problems or 20 percent issues which causes 80 percent of the problems.

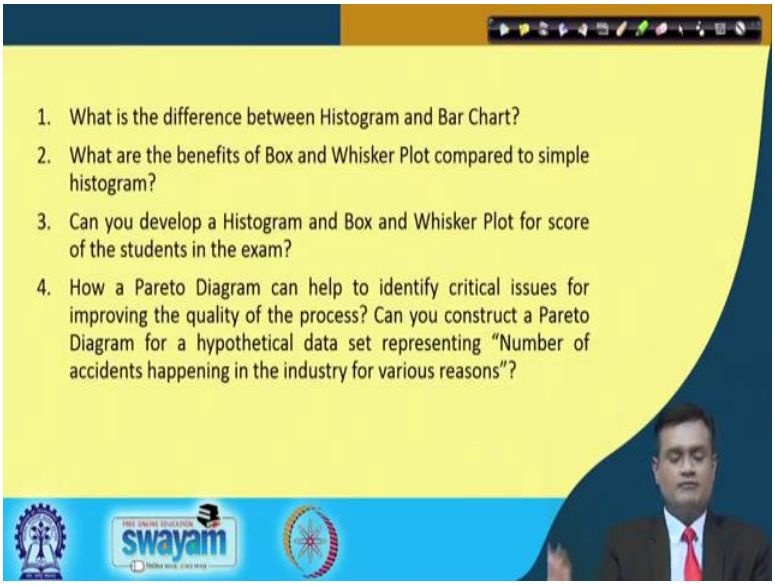
(Refer Slide Time: 36:53)



Just see the typical customer complaint, a restaurant is interested in identifying the critical causes behind the customer complaint and could be like parking difficulty, sales representative, poor lightning, lay out confusing, size is limited clothing fad faded. So, many be it is an example of garment shop clothing (Refer Time: 37:22) and so on. Now, they have put it in a descending manner and when they plotted the cumulative frequency, you can see that this typical 80 percent line, it cuts it here.

And you can say that, if you really address the problems on this side, then these problems will mainly contribute to the 80 percent of the customer complaint. And handling this problem will drastically help you to reduce the customer complaint which is 80 percent in this particular example. So, parking difficulty, sales representative behavior, poor lighting situation, lay out handle, this 4 instead of all and 80 percent of the customer complaint can be addressed.

(Refer Slide Time: 38:29)

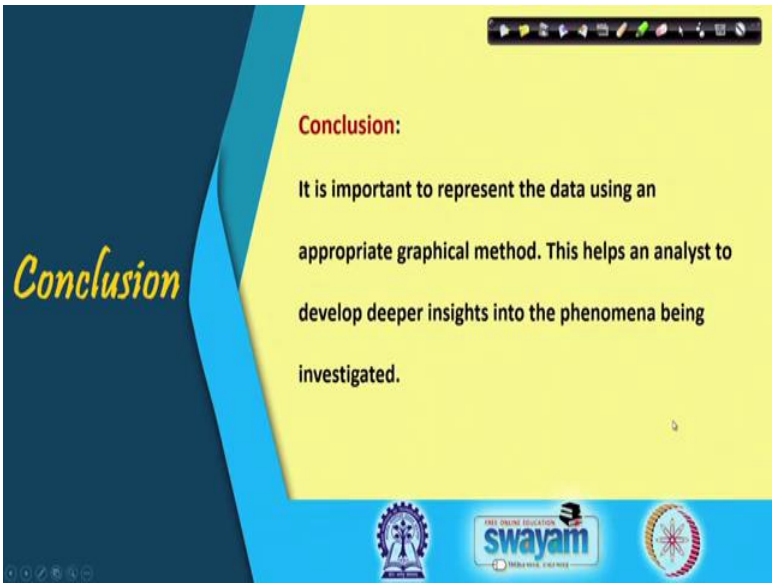


So, with this understanding, we can say that Pareto diagram is a very useful tool in analyzing the situation, just to summarize and help you to introspect, I am floating couple of questions. What is the difference between histogram and bar chart? What are the benefits of box and whisker plot compared to the simple histogram? Can you develop histogram and box and whisker plot for score of the students in the exam? And typically how a Pareto Diagram can help to identify the critical issues for improving the quality of the process? And can you construct a Pareto diagram for a hypothetical data set representing number of accidents happening in the industry for various reasons?

(Refer Slide Time: 39:13)

So, try to address this questions that will help you to have a quick recap of all the concepts we have covered. And with this you can go through couple of references, I have mentioned for better and detailed understanding.

(Refer Slide Time: 39:20)



We can conclude that it is important to represent the data using an appropriate method, graphical method and this helps an analyst to develop deeper insights into the phenomena being investigated.

So, thank you very much for your interest in learning the data summarization, collection, representation. We have discussed this topic in two part, part 1 lecture 18 and part 2 Lecture 19. Please, revise it solve couple of examples go through couple of examples available on the net or the suggested books and try to strengthen your understanding. We will meet with new topic in the next class till that time keep revising, enjoy, be with me.