Multivariate Procedures with R

Prof. Shalabh

Department of Mathematics and Statistics

IIT Kanpur

Week – 04

Lecture – 16

Bar Diagram, Subdivided and Multiple Bar Diagrams

Hello friends, welcome to the course Multivariate Procedure with R. So, you can recall that up to now we were discussing basically the descriptive statistics, various tools which are used in the topic of descriptive statistics which helps us in taking on the intrinsic behaviour of the data or the hidden properties inside the data outside. And the way we worked that we started with the univariate case and gradually we extended it to a multivariate case. And in the last lecture that was a combination of descriptive statistics as well as graphics. So, now from this lecture and in the next couple of lectures we are going to understand how to create various types of graphics. Well, my approach will be the same that I will try to first introduce the graphics in the univariate and then I will try to show you that how they can be extended to a multivariate data.

Beside those things there are some very popular commonly used graphics which I would like to take here so that you are aware and in case if you have forgotten at least you can quickly revise them. Well, the main reason behind taking those simple graphics which most probably you know is that I want to illustrate that how different types of options can be added or different types of features can be added, how the graphics can be made more beautiful in the R software. And you will see that when we are trying to create the graphics you have the full control in most of the software where you try to put the data and press click, click, click and you get some graphics. There may, there you may not have the full control of the graphics but inside the R the advantage is that you have a full control of all the parameters.

The only condition is that you have to study it from the help of that basic command. But once you do it then it is always advantageous. For example, I can give you a very simple example to illustrate it. Suppose I have 20 companies and I want to plot suppose multiple bar diagrams stating the sale, expenditure, etc. for different years for all the 20 companies.

Now if I ask all the 20 companies to give me the graphic, they will create the graphic in their own way. Somebody may show the expenses by green bar, green color bar, somebody may show in a red color bar, third company may show in the yellow color bar, etc. So, now when you try to study them, when you try to compare them, it is not very convenient. But on the other hand, and well their ranges on the X and Y axis may be different. They may give different type of nomenclature also.

But now suppose if you can create a template just by writing the R command that how the graphics have to be created and if you circulate it to all the 20 companies, they simply have to enter the data and they will create the graphic, they will send it to us. Now it is very easy for me to understand because I know green color is indicating the expenses and red color is indicating the sales. So, that is how our life will become very easy and possibly we will be able to take quick decisions. The next issue, why should we consider the graphic? Well, we have done the analytical tools and now we are going for the graphics. Before moving forward, I can share with you, when you look at the numerical values then you get a precise idea that the mean is here and the median is this much.

But if you try to look at the graphics, it may not be possible to get the 100% accurate information by looking at the graphic. For example, when you look at the box plot, the different horizontal bars for the first quartile, second quartile, third quartile they are giving you a fair idea. But unless and until you take a scale and brought it to the Y axis, you will not be able to get the exact value. So, a combination of this analytical tools that is the numerical value and graphical is very helpful. And graphics are very easy to understand.

For example, if you see, if I take the same example which I explained earlier that if you want to express that someone is happy, we can always make a smiley like this. And if you want to express that you are unhappy, you can make a smiley with this type of face. And just by looking at those two smileys you can very well understand. Same is done with the graphics in the R software and in statistics. By looking at these graphics you can get lots of information yourself.

But the only thing is this, you need to do some practice. That is just like by looking at a X-ray or a ECG or a EEG, different doctors get different types of explanation. But definitely we always say if the experience of the doctor is more, possibly the doctor is going to give the proper and correct diagnosis. And the diagnosis is X-ray report is here for my chest and what is happening inside my chest? I do not know, doctor has not seen but by looking at the X-ray, the doctor can find out what is happening inside my chest. That is the beauty of the graphics.

So, looking at the graphic you can find out what is happening inside the data. The more you practice, the more you experience, you will become a better judge or better statistician to take out the hidden information from the data. So, with this objective we begin our lecture and in this lecture, we are going to take a very simple bar diagram, subdivided bar diagram and multiple bar diagram. Bar diagram I know, you are well aware of. But I am taking it so that I can explain you that what are the different features and I will try to extend it to a subdivided and multiple bar diagram.

In order to explain the graphics for the multivariate data set, I need to explain you the rules for creating the graphic in a univariate data set. So, let us begin our lecture. So, now we try to first understand that why graphics are required. So, as I said that graphics summarizes the information contained in the data. For example, these three types of pictures, what do you understand? Just by looking at this face structure like this, this or this, you are able to diagnose whether the person is happy, okay or sad.

So, the mood of a person can be very easily conveyed using these types of smileys. And they have an advantage that they convey the information hidden inside the data more compactly. And I would like to have your attention here that whenever we are trying to talk about these graphics, many times there is a misconception that if you try to create more types of graphics, then the analysis is better, right? But it is not correct. It is something like if somebody goes to a doctor for a disease, do you think that the doctor which is giving the greater number of medicines is better doctor than the doctor who is giving a smaller number of medicines? No. The number of medicines has no importance but the more important part is what is the appropriate medicine.

The same concept here is in statistics that appropriate number of choices, choice of plots in the analysis provides the better inferences. You have to know that what is the problem and what type of plot is going to diagnose it. What type of plot is going to bring the hidden information inside the data outside so that we can visualize it, right? And there are various types of graphical tools. There is two-dimensional plot, three-dimensional plot, scatter diagram, pie diagram, histogram, bar plot, stem leaf plot, box plot, wildland plot, star plot, can now possess many long lists and in R all of them can be created very easily and they can be saved in various types of formulas like jpeg or pdf format, postscript format. They can be copied and pasted into different type of application.

For example, bar chart, pie chart, box plot, group by box plot, scatter plot, co-plot, histogram, normal qq plot. They are used in different situations to diagnose different types of problems. For example, if you talk about the normal qq plot, normal qq plot is used to judge whether the sample data is coming from a normal distribution or not. Similarly, you have seen that box plot and group box plot, they are trying to compare different types of statistical properties of different data sets and they try to help you in their comparison. So, my advice to you all will be please try to understand that why the graphic is to be created, what is the objective of a graphic and then try to match it with the objective of the study and try to choose the appropriate graphic and do not follow the advice that a greater number of graphics are going to give you a better result.

My suggestion is that such situations are very confusing. So, try to see how many graphics are needed, what type of graphics are needed and how they have to be presented. For example, whether you need a bar plot or you need a multiple bar plot or you need a subdivided bar plot. So, if you try to use the correct graphic at the correct place possibly you will get the correct information. So, let me try to first give you some idea about the bar diagrams.

Bar diagrams they are very simple. You are studying them right from an early stage of your schooling. So, these bar diagrams if you recall they are like this type of bars and so on. So, this bar diagrams visualizes the relative or absolute frequency of observed value of a variable. What is the frequency? The frequency is the number of times a particular data is occurring and there can be relative frequency, there can be absolute frequency.

Absolute frequency means the number of times and relative frequency is the number of times divided by the total number of times. Well, this concept if you want to understand you can look into the course Desperately Statistic with R software where I have explained this concept in more detail but here, I want to simply use them. And the bar diagram consists of one bar for each category. So, if there are suppose four shops one in east, one in west, one in north and one in south so these are four categories and so there can be four possible bar. And the characteristics in the bar they are determined by the height of the bar.

So, the height of each bar is determined either by the absolute frequency or the relative frequency of the respective category which is shown on the y axis. For example, in this one if you try to see this is the height of the bar so either that can be here that can be absolute frequency or this can be relative frequency. You know that relative frequency always lies between 0 and 1. And the more important part you have to keep in mind that width of the bar is not important and it can be arbitrary. For example, if you have a bar diagram of this size and see here this size this width has no meaning.

It is only here actually the height of the bar which is making a difference. Well do not try to get confused with histogram. Histogram has a different concept but now at the moment I am talking of the only bar diagram. Yeah, you can see that when all the bars are joined together and then it looks like a histogram but histogram is basically for the continuous data and various bar diagrams are created only for discrete data. Discrete data means categorical data.

So, when I try to explain you the histogram then I will explain you that there the width is important but in bar diagram width is not important. In order to create the bar diagram in R software we have the command here barplot and here this x is the data vector and later on I will show you it can be a matrix, it can be a data frame and it will try to give you different types of univariate bar plot or say multivariate plots. Then here is the width. For example, if you want to control the width of the bars then you can choose here different values then here there is a space etc. and you will see that there are many other options. So, the best way to understand what is the meaning for example is of width is to create the same bar plot for the same data set just by changing the width.

Try to take width equal to 1 and try to take width equal to say 5 and try to see the difference in the two graphics that will give you an idea that how the value of width is affecting the width of the bars. So, in case if you are interested in creating the bar plot with absolute frequency then simple command is barplot and table of x because you see this bar plot is created for the tabulated data means all the raw data is converted into different groups and then or say different classes and then the frequencies of those groups or classes are plotted. And in case if you want to plot the bar plot with relative frequency then relative frequency is divided by the total number of observations right. So, in the frequency of the data divided by total number of observations. So, if you want to create the bar plot then you have to simply consider here table of x divided by length of x.

So, length of x is going to give you the total number of observations in the data vector x and then whatever will be plotted in the bar plot that will be the relative frequency right. So, relative frequency always lie between 0 and 1. So, the maximum value of the height will always be 1 and it is more useful when the variation in the categories value is very high. So, for example if I try to say take suppose in terms of absolute frequency one data is here like this another data is here like this somebody may have a class with the frequencies like this. So, in such situation it is better to make here a plot with relative frequency such that the upper limit here is 1 and this may look like this actually.

So, this may look here like this. So, now just to give you an idea that there are so many options which are available in the bar plot. Well, I am just trying to give you here this screen just to make you understand. I am not going to give you all the details but what I said earlier that when you are trying to plot the graphics in R then you have lots of control in your hand and this is an example. Suppose if you want to create a bar plot, they can be created with vertical bars or say horizontal bars right. The basic command here is bar plot x where this data is going to be here in say this different formats data vector matrix or other format.

So, I will try to explain you then you have here width that you can control then you have a space names which is going to give different names to different bars. Regan test if you want to write something over there right and then you have here density you have here angle you have a color mean who can give different colors to different bars you can give different types of borders. This is the main title of the graph then you have something over here and then the value on the x axis the value on the y axis the limit of the values on the x axis the axis name etc. You can see here you have to write down this statement or this command only once after understanding it but now it will give you the graphic which you want to control with your own hand right. And if you want to understand those things you simply have to write down here type the help within double quotes bar plot and try to look into the help menu you will get all the information for example what height is going to indicate how width is going to indicate space names dot or etc etc etc.

Our only thing is written here in must detail the only constraint is that you have to read it that's all right. So, now let me try to take here an example and try to show you that how such plots can be created. So, I am trying to take here an example here I would like to create three bars right. So, I am considering here that there are three sales persons in a shop and they are indicated by the numbers 1 2 and 3. So, whenever a customers enter into the shop one of the sales person attend them and then the number of persons which the three sales persons have attended in the first handed out of the first handed customer is recorded and this is here like this for example one that means the customer was attended by the sales person one then again one then two that means the third customer is attended by the sales person two and so on.

So, I have just stored this data as a sales per inside this data vector right. Now if I try to create here a bar plot here like this bar plot table sales per so you can see here that they are there are here three bars which are trying to indicate that the sales person number one that person has attended almost say 28 29 customers similarly the sales person number two has attended about 42 43 customers and sales person number three has nearly attended 28 29 percent. So, by looking at this graphic I can say that the sales number one sales person number one and three they have attended almost the same number of customers where the sales person number two is more efficient and this sales person has handled the maximum number of customers. So, this type of hidden information you can obtain just by looking at this graphic right. Now in case if you try to see here this values are 0 10 20 30 40 and one of the problems here is that if you try to see this value here is on the see here this upper limit this is going outside the range although you can control it but this is the default setting.

So, if you try to make everything on the scale of 0 1 then it is a better option to plot or create the bar plot with respect to relative frequency. So, this is obtained as bar plot then it is here the relative frequency yes that means the frequency of a group divided by the total frequency which is in the case of discrete data it is length of the data vector. So, if you try to see here this is the bar plot if you try to see this structure and the earlier structure the structures are the same the only difference is on the scale right. So, it is more advantage if you try to make here the scale try to define the scale to be here 0 and 1 and if you have suppose 20 companies and if you ask them to give the bar plot of their monthly expenses and if you ask them to create only the relative frequencies then by looking at the height of the bar you can very easily find out at which company is spending more or less and so on right. And now in case if you want to provide different type of information to this graph for example if you want to put here title here like as customers attended by the salesperson suppose you want to give it then you have to use here the option here main is equal to customers attended by salesperson and this will appear here remaining part bar plot table salesperson will remain as the same right.

And now similarly if you want to incorporate more information then you can see here that this part is the same as earlier and now, I have added here see here names dot arg which is here say sp1 sp2 sp3 for example I want to give the information on the plot that which bar corresponds to which salesperson so this is here sp1 sp2 and sp3. And now I want to indicate here that what is here sp1 sp2 sp3 so I want to write down here salespersons and inside the parenthesis sp. So, you can see here this can be obtained by writing here xlab is equal to within double quotes salespersons and within parenthesis sp right. And similarly, if I want to write down title on the y axis suppose you can see here the number of customers so this can be obtained by ylab equal to number of customers in double quotes. So, you can see here that gradually I have introduced more say options here to make the graphics more informative right.

Suppose if I want to change the colors of this graph also so this part you can see here this is the same as earlier and now only this part col is equal to red green orange in double quotes inside the data vector has been written and it is changing the color and if you try to see the order in which these colors are mentioned in the command the bars are colored in the same order red is at the first position so it is at the first position green is at the second position it is at the second position orange is at the third position it is at the third position right.

So, now if you try to see here this is about the bar diagram and my objective is not to teach you here about the bar diagram because I am sure you know about it but my objective was to give you some option so that you can understand and get convinced that different types of options are possible to add which will give us the good graphics. For example if you try to see here suppose I want to make here bar plot now I am going to make a mistake here I try to write down here sales per and I do not write here the command table you can see here it is giving here like this but this is what you do not want you want to create the bar plot for each category but it has created the bar plot for each value which you do not want and if you did that is why if you try to see the command here is bar plot along with table right. As soon as you give here this command here table sales per this will come out to be here as say these three graphics these three bars and if you want to make it here with respect to here the relative frequency we can see here on the y axis the range is going up to 40 and yeah well you can also control it by xlim and ylim on the x and y axis but here it is taking the default value but if you want to now plot it with respect to relative frequency you can see here now this range is changed to 0 to 0.4 right.

And similarly if you want to create if you want to put here the main title then if you just in the same command if you try to add main is equal to this you can see here now this title has appeared here and similarly if you want to add here the information about the sales person and the titles on the x and y axis this if you try to see this is now here the same command I have copied and pasted here you can see here this is here sp1, sp2, sp3 and now if you want to change the color of the graphic also you can just I can copy and paste the same command here you can see here this is here like this and if you try to change the name of the color suppose I make it here blue you can see here it is now here blue. So, well in our software in order to give different types of colors there are these commands are available and you can find them in the help menu also so that you understand that how these good graphics can be created right. So, let me try to create clear the screen for our next operation. Okay now I come to an idea the next idea which is again the bar plot but it is subdivided or component bar diagram right. Why it is used? In subdivided or component bar diagram the total magnitude of the variables is divided into different or say various parts.

Let me try to explain you here with an example suppose there are three shops shop 1, shop 2 and shop 3 and the number of customers which are arriving in these shops on day 1, day 2, day 3, day 4 they are recorded here like this and you want to get here a data I can show you like this one right. So, means what do you want here? You want to have here like this where there is a bar here for say shop 1 there is a bar here for say shop 2 and there is a bar here for shop 3 and the cumulative number of customers which are coming here on day 1, day 2, day 3, day 4 this should be plotted here by subdividing the bar space or the height of the bar. So, in order to create such subdivided or component bar diagram the basic command remain the same bar plot but the way we are entering the data that changes and we have to input the data in a matrix format right. So, if you try to see this is here the data and we want to create here a matrix like this one, this one, this one where the rows are going to indicate the days and columns are going to indicate the shop numbers. So, now you know that how to enter this data and how to create such a matrix.

So, that I will not repeat but anyway I have given here the basic command a matrix number of rows are 4 number of columns are 3 and the data has to be entered by row so I have written it in this order like this. If I try to show you here the data I have written is this and this and this and this here because by row. So, now you can understand that why I had taken the topic of matrix in the beginning that because that is going to be used in different places. So, and then I try to store this data as CUST in this matrix format and this is my here this matrix right. Now I try to use the same command here bar plot and the variable name is in the matrix format right and this will create a subdivided or component bar diagram with columns of the matrix as a bar right means if I try to see here this columns will become the 3 bars like this one and the individual values inside the bars they will be partitioned and this section inside the bars will indicate the values in the cumulative form right.

How I will try to show you here. For example, if you try to give the command here bar plot CUST inside the parenthesis in the R console then you will get here this type of bar diagram which is a subdivided bar diagram but how to interpret it and what values are it going to give. So, if you try to see here, I will try to use different colors of pens so that you can follow me. So, this is here a 0.2 and this here is this part black part.

This value here is somewhere here which is actually here 2. What is this 2? This is this value right. Now on day 1 and day 2 together if you try to see how many customers have come. So, if I try to say here day 1 and day 2 together there are 26 plus 2 customers so which is equal to here 28. So, now this 28 is plotted here on the second bar. So, this is always indicating here the value 2 plus 26 which is equal to 28.

Now about the third part here this one in red color what is this indicating? This is trying to indicate how many customers in the first 3 days that is day 1, 2 and 3. If you try to see how many are this is 2 plus 26 plus 42 and this value here is 70 which is plotted here 70. And after that what is happening how many total numbers of customers have visited the shop 1 in the first 4 days. So, this is the sum of here all 2, 26, 42 and here 30 which is here like this and this here is plotted here like this.

So, this how the subdivided bar diagram is created. Now the same thing is happening in the other two bars also which are indicating the shop number 2 and shop number 3. So, this is how they are created. So, now if you want to make it more beautiful you want to give them different colors you can use the options names.arg, xlab, ylab, color etc. Now one thing very important for you to understand and observe these commands more or less they remain the same for all the graphics.

That you will see in general I am not saying that there is 100 percent guarantee but more or less they will not change. For example, if you want to give the color usually you will see the command will be col. So, now if you try to see with the gust you are giving the same data and if you want to have this here names shop 1, shop 2, shop 3 which is given here by this one names.arg, shop 1, shop 2, shop 3 and then you have here shops on the x axis which is given here by xlab, then you have here days on the y axis which is here given by here days here ylab and then if you want to give here the colors then you can see here that the colors is here red, green, orange, brown.

So, you can see here this is the color ordering. So, now this is now you can create and you can insert more information also such that they look more beautiful more informative. So, let me try to give you this example or try to show you this example on the R software so that you get confident. Try to see this is here I have written the matrix. So, let me try to say here this is here the name of the matrix, see here x.

So, if you try to see here this is here the x matrix. Now if I try to see here bar plot senior x you can see here this is here like this. And if you want to create here beautiful graphic just by giving different types of options I can what I have shown you here you can see here that okay now I have taken here the name of the variable here as say x0cost. So, let me try to take it here x you can see here this is here like this. So, that is how you can create such graphics without any problem.

So, now I hope you have understood it. Now let me clear the window and let me try to give the last topic of this lecture. This is also a bar diagram but this is a multiple bar diagram. That means there are more than one bar, there are multiple bars and they are arranged in groups. So, its construction is like the subdivided bar diagram which we have just completed. I simply have to give only here one option beside equal to true. This beside equal to true will add the next bar diagram side by side. So, the basic command will remain the same here and then I have simply have to add here beside is equal to true BES IDE all in lower case is equal to logical variable true. So, I try to take here the same data set of this customer. That there are three shops, four deals and the number of customers in the campus they are here like this. And now if you try to see here that if I want to plot here the graphic in such a way I have to write down bar plot cost that is the data in matrix format and beside is equal to true.

Now if you try to see this type of data is coming. What is this thing? If you try to see this is your here shop number one, shop number two, shop number three and the first bar here day one, day two, day three, day four which is like a d1, d2, d2, d3, d4. Once again d1, d2, d3, d4 and if you try to see this height, this height is nothing but this height. And if I try to change the color, this second height, this is value here 26. And after that the third height, this height is day number three in shop number one. And after that the fourth height, this fourth height is this value 30 which is the shop number one on day four.

And similarly, others will go. So, now if you try to see the same data set is now trying to give you a different type of information. And in case if you want to make it more beautiful by giving different types of titles on the main graphics like as customers in shops, shop and shop one, shop two, shop three and the number of customers including if you want to give here d1, d2, d3 that can also be added. So, for that the same command what I have used earlier that is the same command. The only thing is this I have added here beside is equal to true.

And this is giving you here this type of graphic. So, you can see here it is not difficult at all. And let me try to give you this graphic on the R console and write. So, if you try to see here, if I try to see here bar plot see here X. If you try to see this X here is this matrix and then if I see here bar plot X, this will give me here a bar plot like this. But if I try to add here the command here beside is equal to true, then you can see here it is coming here like this.

And if I want to get here this thing, suppose I make it here X and if I try to come just copy and paste this command and if I try to give it here you can see here multiple bar diagram is also coming exactly in the same way as I shown in the figure. So, now we come to an end to this lecture. You can see how we started and how we concluded. We started with a univariate bar diagram and then we extended it to subdivided bar diagram and multiple bar diagram.

So, if you try to see within one variable there are different categories. It is a sort of multivariate setup and we want to have the information from the data on these categories also. So, for that the same concept of bar diagram is extended in different ways and as I said in the beginning that the data sets have lots of information and it depends on our own capability that how much we are successful in extracting the hidden information from the data and that is the job of a data scientist. So, from the same data set just three shops, four days you can get different types of information. So, it depends on the objective of the study that what type of information is required and based on that this is only you who is going to take the call what you want to plot.

Bar diagram or a subdivided bar diagram or a multiple bar diagram. Except you nobody can take this call and nobody will come to us to tell us that okay in this case you try to make this type of graphic and that is the role of a data scientist. The doctor always examines and then takes a decision that which medicine is to be given. Nobody gets the doctor that you have to give this medicine. So, that is why if you try to compare as a data scientist you have to understand this basic terminology, basic understanding of different types of graphics and then depending on the situation, depending on the condition you have to take the proper call that which of the graphic is going to give you the correct information.

And this will come once again I will say by experience and practice. So, you try to take some data sets and then try to see what type of information can be obtained. Try to work with this univariate bar diagram, multiple bar diagram, subdivided bar diagram and try to see how you can obtain different type of information from there. So, you try to practice it and I will see you in the next lecture till then goodbye. Thank you.