Foundations of R Software
Prof. Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology, Kanpur

Lecture - 49
Graphics- Scatter Plots and Bar Plots

Hello friends, welcome to the course Foundations of R Software. Now, from this lecture we are going to talk about new aspect of this R software. This is about how to create the Graphics. You know that whenever we get the data and we want to dig some information from inside the data, then that can be done through the analytical tools like for example, you know mean etc. And, the second is this graphical view, yeah both of these methods they have their own advantages.

And, the idea is that we should actually use this analytical and graphical tools together, but anyway. The question here is now that how can you create the graphics in the R software? And, I can promise you in the beginning itself, the type of possibilities, the amount of options, the number of options what you have and the amount of control what you have in creating a graphic in the R software possibly that is excellent. The only thing is this when you want to control something then you have to work on the individual parameters.

It is something like this you know that when we want to cook some curry or some dish, there are some readymade spices which are available in the market. You just bring the spices and this is a combination of say 10, 12 different spices which are mixed in different quantities. You try to put it in the vegetable and it will give you some flavour. But, in case if you really want to change that flavour that is not possible from those spices or rather you have to mix those 10, 12 spices manually with your choice of a quantity.

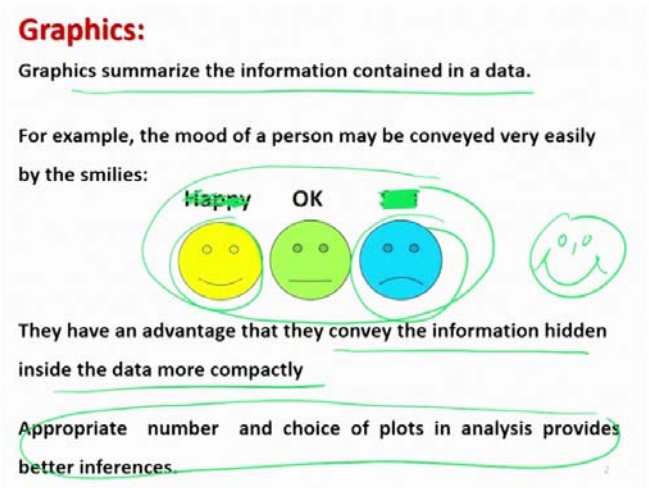
And, that will give you a different type of flavour and, but the biggest advantage will be then the taste of the curry is going to be under your full control. So, that is the story with the R software also and compared to those software's where you can click, click, click and get the graphics. So, now, in this R software there are many options and there are many packages which can be used for creating the beautiful graphics.

My objective here is to give you some idea that how you can begin and how you can think, how you can control the different parameters in the R software and, how you can manage these graphics in the R software. Surely, I cannot take care all possible type of graphics which are available here. So, I will try to initiate a thought process inside you that how to create the graphics and how to control different types of parameters.

And finally, the objective will be the type of graphics what you want, how you can give these values to create the graphic. And, definitely I am going to consider here only some collected common graphics. And, my idea is that I will first try to consider those graphics which you know like a bar plot, histogram, pie chart you know, you are doing it from your school days also.

So, that will give you some confidence, that ok how to create that graphics in the R software that you possibly have created in other software also. And, then you can learn that how you can manage those different parameters of that graphics to make it more informative and more beautiful. So, that is how we are going to proceed in learning the graphics. So, we begin our lecture and try to understand with a couple of examples.

(Refer Slide Time: 04:11)



Right ok. The first question comes here why should I use the graphics? So, we know that the graphic summarizes the information contained in the data. For example, if you really want to indicate that whether someone is happy or say normal ok or say sad, you can

very easily use these types of smileys nowadays. For example, if I do not even write here what is this here happy, but just looking at this face you can say that it is indicating that the person is happy.

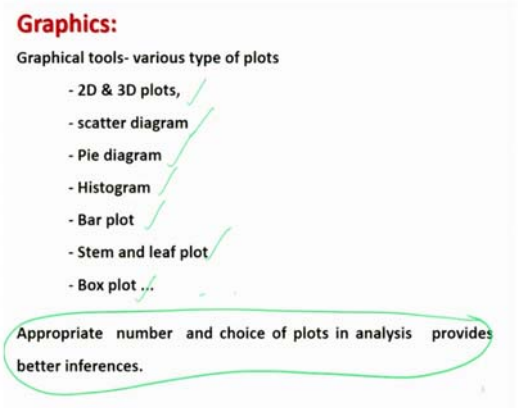
And, even if I try to remove here what is written over this sad, but still you can look at this face, this smileys and can say that ok the person is not happy. So, these are the graphics. So, graphics also conveys the similar information what is conveyed by the data and the advantage of this using the graphics is that they try to convey the information which is hidden inside the data more compactly, right.

If you want to simply explain that how to explain whether the person is happy, I am sure that you will take you have to write and explain very nicely and clearly. But, in the case of graphic you simply have to create a smiley like this one, that is all, right. If you want to make it here more happy, you can like even like this here.

So, that is the advantage. Now, this another question sometime I hear or I come to know from different people that people have a sort of thought process; that if you try to use more number of graphics or more number of plots.

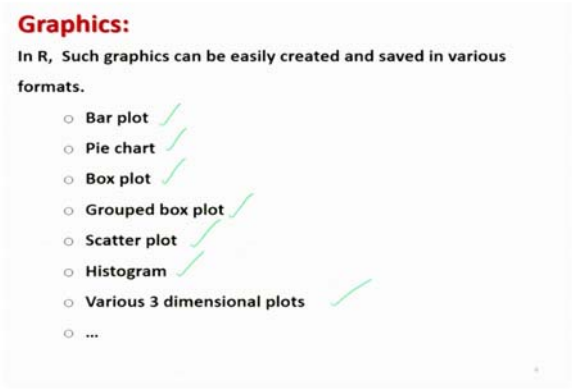
Then, the graphic analysis or the analysis becomes better and it provides better inferences. It is not correct, you have to simply use the appropriate number of graphic and that gives you the good quality of inferences out of that. It is just like as if doctor gives more number of medicine than required, it does not mean that the doctor is good. A good doctor is that one who gives only the required amount of medicines, right.

(Refer Slide Time: 05:58)



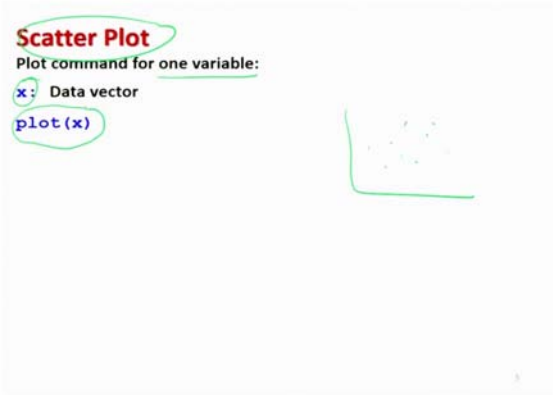
So, now in R software or in statistics in general there are various types of graphics which are available; 2-dimensional plot, 3-dimensional plot, scatter diagram, pie diagram, histogram, bar plot, stem leaf program plot, box plot, Venn plot, many many things. What we have to always keep in mind that appropriate number and choice of plot in the analysis provides better inferences, that is our key word in this entire analysis.

(Refer Slide Time: 06:23)



So, now the question here is this in R these type of graphics can be created? Actually there is a long list, but some of the popular graphics which you know which you already have learnt in your past are like bar plot, pie chart, box plot, group box plot, scatter plot, histogram, various types of 3-dimensional plots etcetera and that is a very long list I will say again.

(Refer Slide Time: 08:42)

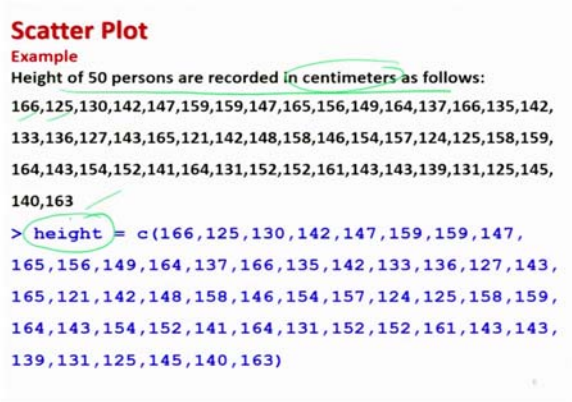


But anyway, we are trying to learn here how these things can be done and I will try to take some collected these commands and I will try to show you how to create the graphics. Yes, once you create the graphics, after that you have to analyze it and then you have to dig out different types of information which I am not doing it here. But, my main objective is that I want to show you that how the graphics can be created ok.

So, suppose we have 2 variable, 3 variable based on that you can have univariate graphics, bivariate graphics, tri and 3 dimensional graphics etcetera. But, at this moment we are in the beginning I am simply assuming here that we have here only 1 variable on, which we have collected the data and the data is stored in this x.

So, if you want to simply make a scatter plot, scatter plot is something like this here on the x and y axis. You simply try to plot the data, right. So, this can be obtained by the command plot p l o t, right and inside the parentheses you write here x.

(Refer Slide Time: 07:36)

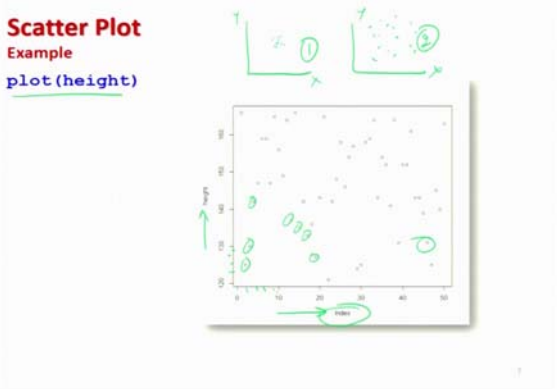


And, after that there are many options which I would request you once again to look into the help menu and try to understand them. Some of the options, some of the popular common options I will try to show you here. So, let me try to take an example here and with these examples I will try to create different types of graphics and I will try to use different options to show you that how options can change the look of the graphics.

So, suppose I have a data here where I have collected the data on the heights of 50 persons and the data is recorded in centimeters. So, their heights are like 166 centimeter,

125 centimeters and so on and this data is stored in a variable here height in this data vector, right.

(Refer Slide Time: 08:16)



So, now if you try to use here this height so, you can see here means corresponding to the index it will try to plot these values on the x y axis. You can see here these are the data values which are obtained here, yeah. This scatter diagram gives out information for example, if you have got two scatter sets and on the same scale of x and y, if one scatter plot comes out to be here like this and another plot comes out to be here like this.

Then, if this is plot number 1, this is the plot number 2 then it indicates that variability in the data in the plot number 1 is less than the variability in the plot number 2. So, different types of such statistical outcomes are obtained in the first step through this plot command. So, this is how you can create the plot command. You can use the plot command to create a scatter plot, right.

Now, you can have different types of options that you can see here, here it is trying to take the index. And, then here height then you can change this name, you can change this ticks on the x axis, y axis, you can change the color of these dots and you can manage many. So, what I am going to do here that I will try to take here a couple of examples and through those examples I will try to show you that how you can control these aspects.

And, in this lecture I am going to explain you in quite detail. And, then from the next lecture the same commands are going to be repeated, then you will not need to repeat them again, right.

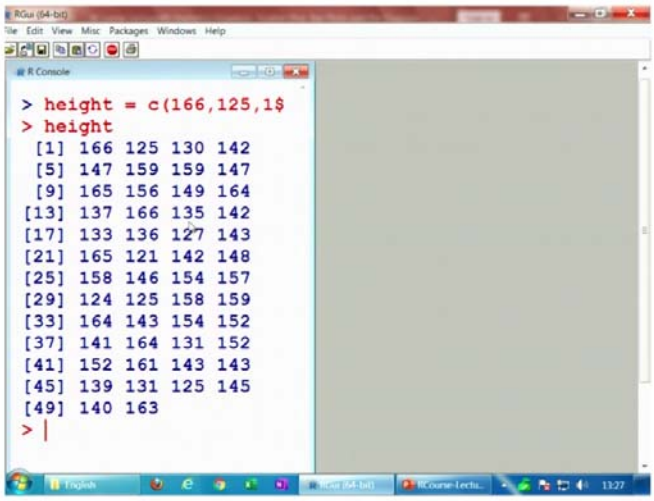
(Refer Slide Time: 09:37)

So, now for example, in this case if you try to see, if you simply want to change the colors of these dots. You can see here earlier these dots are in default black color, you can see here, but now in this plot they are in red color. So, how to get it done? Simply, have to use the plot command with the height and now you have to give it an option col and then you have to write r e d red within the double quotes yeah.

There is a way in which you can specify these colors in the R software, right. So, for example, r e d means red and that is going to change all the dots into a red color you can see here.

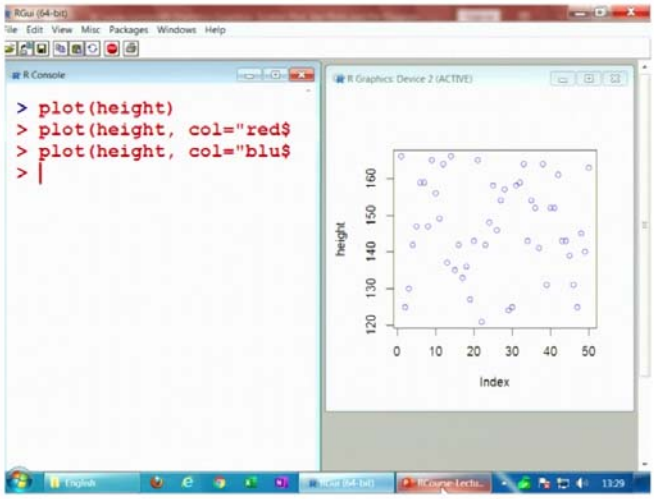
So, the graph is the same, but this color is changed. So, this is how actually can add here more options here and they will try to change the view of the graphics. You can add here that the title, somewhere on the main title. You can indicate here what is indicating by different legends etcetera etc, but anyway let us try to learn them with one by one. But, before I move forward let me try to show you here that on the R console that how it will look like and how I am going to manage it.

(Refer Slide Time: 10:40)



So, that we are trying to understand the outcome. So, you can see here I already have reduced the width of my window. But, if you try to see here this is my here data, the data is the same and now if I try to create here plot here.

(Refer Slide Time: 10:53)



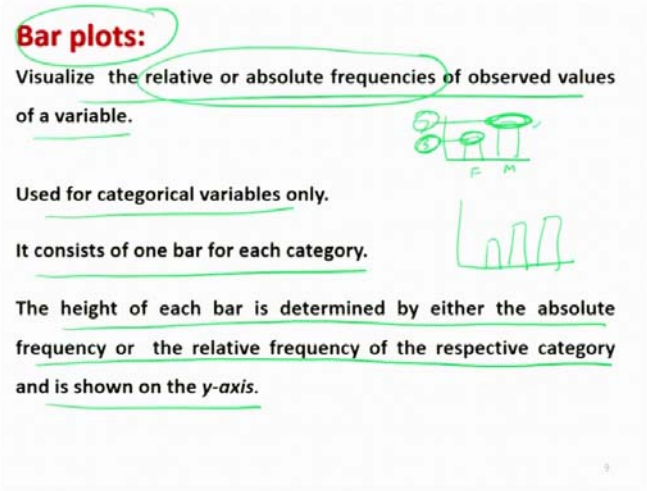
So, I will try to write down here plot and then the name of the variable, say here height. Now, you can see here there is another window which is opened here and in which you can see here this is here the graphic. So, if you want to means store this graphic, you can export it, you can there is an option here resize and etcetera etc. And, the most simple

option to save this graphic is this you can, right click on your mouse and you can see here there are different options copy as meta file, copy as bitmap and print etc. etc., right.

So, these things you can do very easily. So, now, you can see here this is the way I am going to do. On the left window, I will try to write down on this window I will try to write my commands and the outcomes are going to be shown on this window here on the, right hand side, right. Now, if you try to see here, if I try to add here the command here color see by using the command col and then if I write down here red; you can see here what happens.

Now, if you try to see here on the in this graph what happened as soon as I try to execute it command. As soon as I say enter, you can see here this is the color becomes here red. And, similarly if you want to make it here, suppose here blue you can see here you simply have to makes you l is equal to blue and you can see here on the, right hand side this is changing. So, that is the way I am going to represent all the graphics over here, right.

(Refer Slide Time: 12:16)



So, let me try to take here one more very interesting example and through which I will try to explain many many things, right. So, we are now going to consider about bar plots. So, the first question comes here, what are these bar plots? So, these bar plots are used to

visualize the relative or absolute frequency of observed values of a variable, right and this is used in the categorical variables.

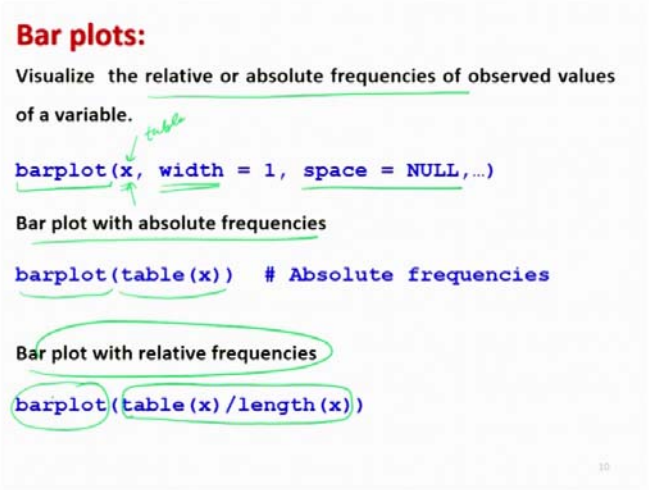
Now, tell me what is this relative frequency or absolute frequency? That was the reason I had explained you these concept in the last lecture; so, now, you know that how to compute the relative frequency, absolute frequency, what is the meaning of this. So, it will not be difficult for you to understand what the bar plots are trying to indicate. And, there will be a one to one correspondence between an analytical tool and a graphical tool, right.

So, what happens, that you have seen that when you are trying to find out the absolute frequencies or the relative frequencies, there were some categories. And, you were trying to find out what is the absolute or relative frequencies in those categories using the command table. So, now, those frequencies are going to be plotted with respect to the categories, right. So, one category will have one bar each. For example, if you try to recall we had taken the one example where we have the variable here as a gender, say female and say male.

So, their frequency was here 3 and 7; so, that will be indicated like this. Similarly, we had taken one more example of direction, where you had 3 direction of a restaurant and then we had created the frequency table in terms of absolute and relative frequency. So, now, we are going to plot those frequencies against those categories under this bar plots. And, the rule here in the bar plot is that the height of the each bar is determined by the either the absolute frequency or the relative frequencies of the respective categories.

And that is indicated on the y axis; that means, if the height of the bar is more; that means, the frequency is more. It is like for example you can see here in this example, the frequency here is 7 and in the first case the frequency here is 3, but that is indicated by the height of these bars, right ok.

(Refer Slide Time: 14:19)



So, now the next question comes here, how to create such bar plots in the R software? So, we have here both the option that we can use either the relative frequencies or the absolute frequencies of the observed values of a variable and then we have a command here bar plot b a r p l o t. All in lower case alphabets, after that we write down the data and now you have to be very careful, this data is going to be in the tabular format.

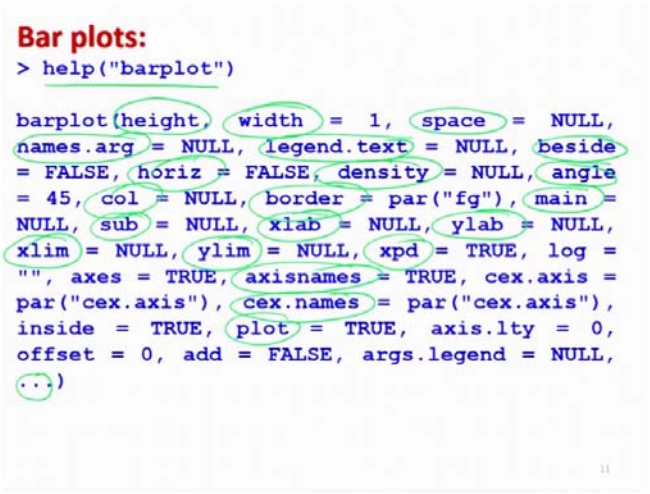
So, first you have to create the frequencies because this bar plot is going to plot the frequencies not the data. So, first you have to obtain the frequencies and those frequencies are going to be used as an input here x. Then, after this there are many commands actually, first option here is width. So, that you can control the width of the bar, then space means you can control the spacing between two bars, right and so on. So, I will try to take up these some common operations through these options one by one.

So, right so, the first question is how to create a bar plot with the absolute frequencies? So, for that you simply have to use the command bar plot and inside the parentheses if your data is given in x, then you have to write down here table x. So, table x is the command which gives you the frequency or the absolute frequencies of the data in x, right.

And, similarly if you want to a create the bar plot with the relative frequencies, you know that you simply have to revise the absolute frequencies by their length. So, you

simply have to use here the same command which you use to find out the relative frequency which was table x divided by length of x and then you have to use here the bar plot, right. So, this is how you are going to do it ok.

(Refer Slide Time: 15:55)

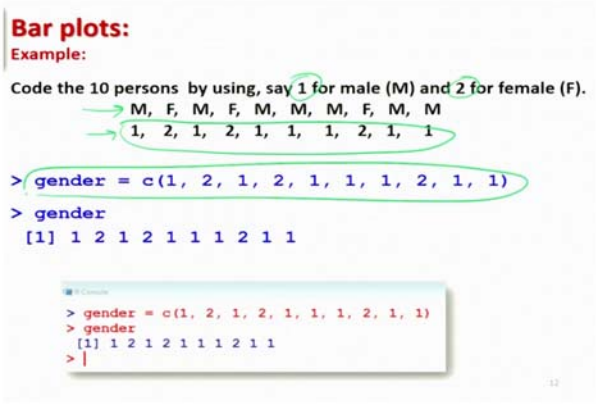


So, now if you miss my very sincere request to you all is that, without failing please try to look into the help of this barplot. You can see here that how many things I can show you here which are possible and at the end I also have written here 3 dot; that means, continued. You can give here the data, you can control the width, you can control the space.

You can give the names, you can control the legends, you can control the beside, horizontal, density, angle, color, border, main title and then the sub title, the level on the x axis, level on the y axis, the limits on the x axis, limits on the y axis, xpd, axisnames, cex names, plot etcetera etcetera so, called you can see. So, now, if I try to explain you each and everything in this lecture possibly this will become the longest lecture.

So, I will request you that you please try to go through it at least once. I am not asking you to keep in mind, but at least you must know what are the different possibilities. So, that whenever you are trying to create the graphics and whenever you want to compare the graphic with any other graphic, you can know how you can create the same graphic in the R software.

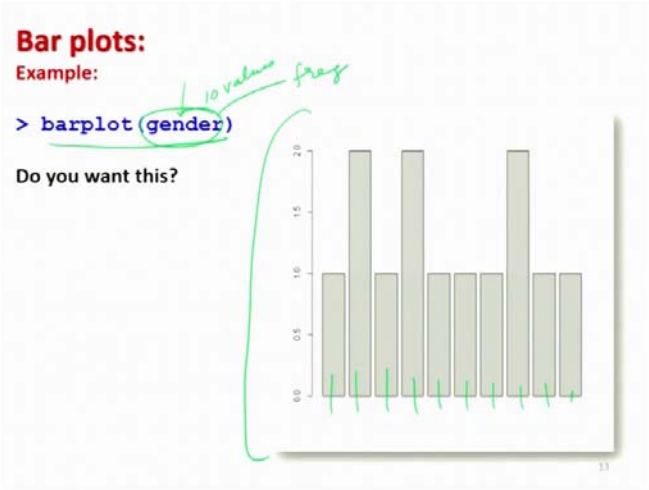
(Refer Slide Time: 17:07)



Now, I try to take here a very simple example that I took in the last lecture and I will try to create the bar plots on the same data set. And, I will try my best to keep the examples as minimum as possible so, that you can feel a connection between the analytical and graphical tool.

So, you can recall that we had considered the data of 10 people and we had coded them as male and female. The male category was indicated by 1 and female category was indicated by 2. And, we had the data on male and female like the gender of the persons and then it was coded here like this as gender and this variable was stored in gender.

(Refer Slide Time: 17:44)

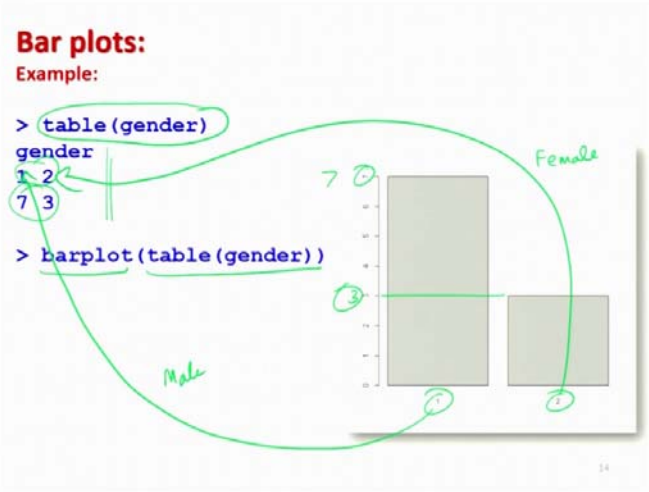


Now, I try to create here a bar plot. I simply try to give it here bar plot gender and I get here this thing. Well, my question to you here is that do you want this? Think about it. There are only here two categories here, male and female. And, if you try to recall that was the same outcome of the frequency table also, then why it is trying to give you here 1 2 3 4 5 6 7 8 9 10, these are 10 bars and you have just said here that each bar indicates only one category.

So, what is this happening? Well, you have made a mistake, sorry rather I will say I have made a mistake here, that we have created the bar plot on the gender which is which are 10 values and you have to create the bar plot on the frequency. So, if you try to replace this gender by the table gender, then it will give you the correct value. Well, I have done it intentionally because I know that in many many software the command is simply, that you try to give the variable name inside the parenthesis.

So, frequency table is created automatically, but here it is not happening in the R software. So, that is why I have taken this example to just indicate you that I have made this mistake, but you please do not make this mistake.

(Refer Slide Time: 18:55)

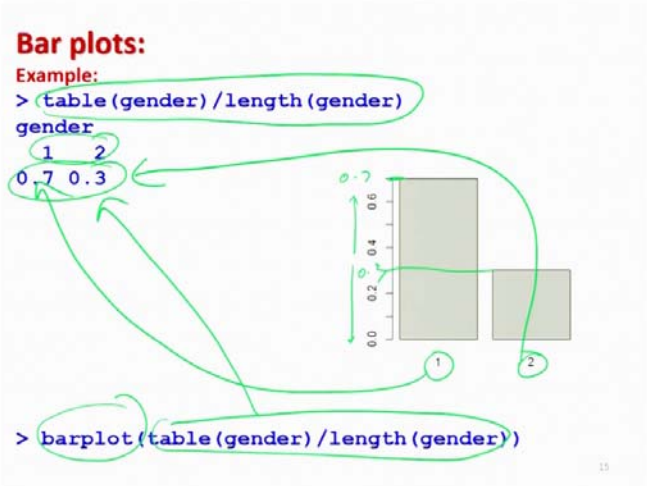


So, now if you try to see here, if you try to operate the command table over gender then the gender data is like this. Two categories 1 and 2 and the values are here 7 and 3. Now, if you try to use the command bar plot on the table gender, it will give you like this. You

can see here exactly this is here 7 and this is exactly here 3. So, this is category 1, this is category 2 which is coming from here category 1 and category 2 from here like this.

Category 1 here is male and this is indicating the male category and this is indicating here the female categories, right. So, this is the one of the very basic bar plot and after that you will have many options to change the name, change the colors and put a title on the graph, on the axis etcetera; so, which I will try to show you very soon, right.

(Refer Slide Time: 19:44)

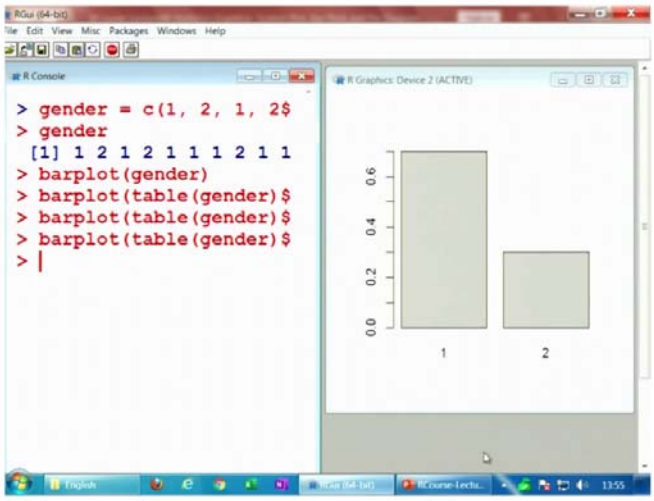


Now, in case if you want to create this bar plot with respect to the relative frequency. So, here you can see you have plotted it with respect to the absolute frequency. Now, you want to make it with respect to the relative frequency. So, the command here is very simple. First you have to find out the relative frequency, for that you have to use the command, table gender divided by length of gender, that you already did in the last lecture.

And, now you simply have to use the bar plot command here over this outcome. This outcome is that there are two categories and their relative frequencies are here like 0.7 and 0.3. And, now you will see here that it is giving you here the values which are like 0, 0.2 and so on, because the relative frequencies they are the proportions. So, they will always lie between 0 and 1 and if you try to see it is here is 0.7.

So, for the category 1, the value here is 0.7 and for category 2 the value here is 0.3 and category 1 is given here and category 2 is given here. So, now, you can see here this is one of the very basic fundamental plot. But I will try to show you more options in the next example, but before that let us try to first create this example on the R console.

(Refer Slide Time: 20:51)

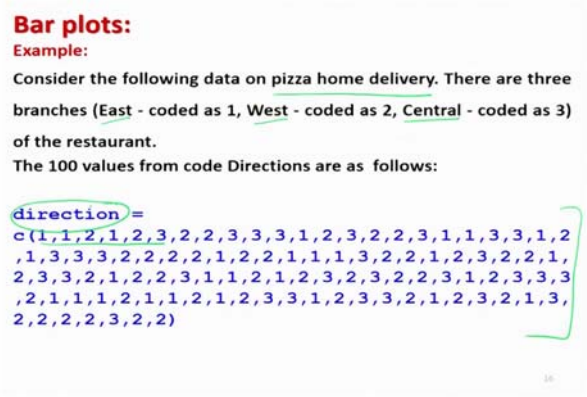


So, right let me clear the screen and I try to create here my here data which is here gender. And, now if you try to create here the bar plot over this bar plot, you can see now I am going to make a mistake here. This will come like this, right. So, because this is the individual values 1 2 1 2 and so on, that is the thing which you do not need.

So, what you have to do here? You have to give the input the data in terms of the table, table command. And, as soon as you enter here, you get this graph over here on the, right hand side which has only 2 bars. This first bar it is indicating here 1 and second bar is indicating here 2 and first bar is indicating the male and the bar number 2 is indicating the females.

And, now if you try to create here this here bar plot with respect to the relative frequency here; so, you can see here it is becoming here like this and you can see here on the this y axis, this absolute frequencies are now converted into relative frequencies. And, you are getting here the same bar plot, but now with respect to the relative frequency, right.

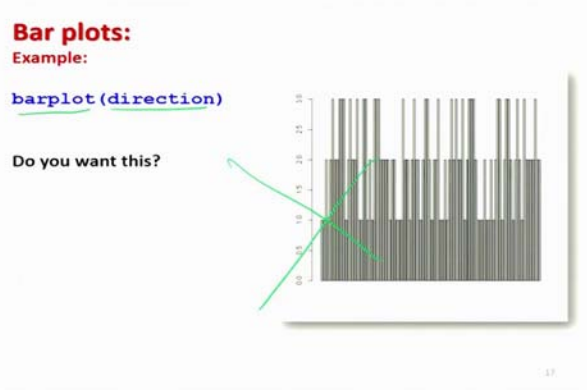
(Refer Slide Time: 21:53)



Now, I try to take here one more example and through this example I will try to show you a couple of details related to the options. And, I will try to show you how you can control the graphics in the in the way you want. So, I try to consider here the same data which I considered in the last lecture, that there is a restaurant who has got three branches in the city in the east, in the west and in the central part of the city.

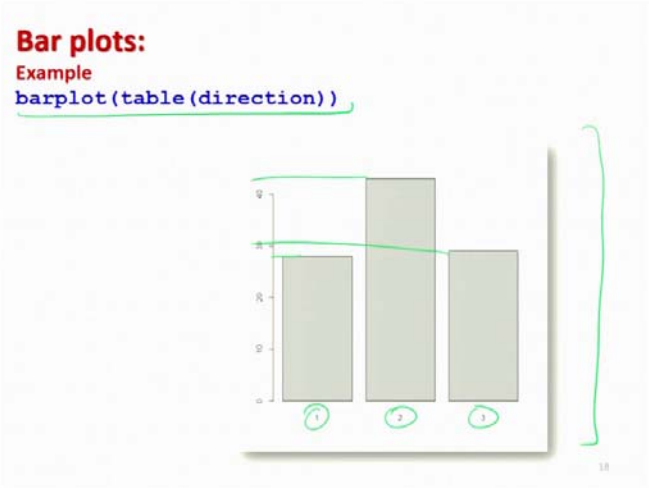
And, this restaurant is delivering the pizza at home. So, now, that they are getting the orders at center places and then they are trying to contact the branch which is the closest to the place of delivery. And, this is the data on the 100 values of delivery that which branch has delivered the food, right. So, this is like 1, 1, 2, 1, 2 and so on. So, I have created this data vector here direction.

(Refer Slide Time: 22:43)



And that is the same example which I did in the last lecture, where we had created the frequency table and we had obtained the absolute and relative frequencies. So, now, I want to create the bar plot. So, I try to make here the first here a mistake and I use here the command bar plot direction. So, you will see these are the 100 values, but this is the thing which you do not want. Now, you know why? Because you want to give the data in terms of here table in the form of frequencies.

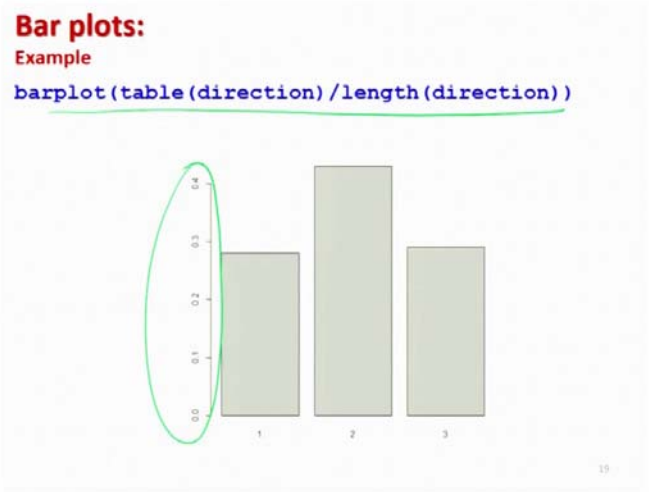
(Refer Slide Time: 23:03)



So, I try to write down here barplot tabled direction and now you can see here you will get here a similar bar plot what you got in the last example and so, there are 3 directions; 1, 2 and 3 and here are these frequencies. Now, in this example I am going to explain you that how you can add different features.

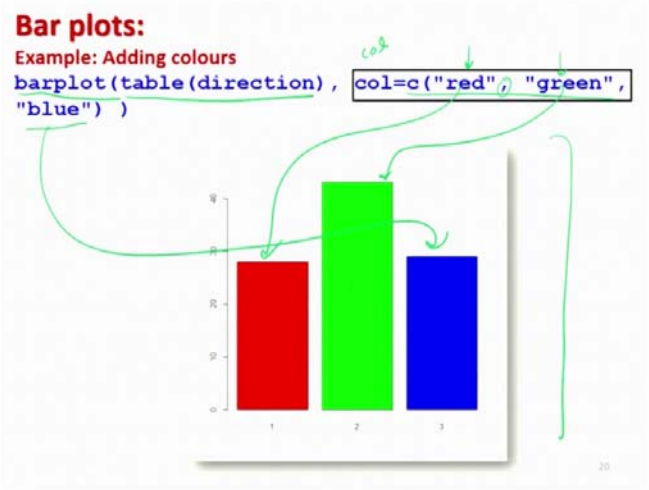
So, please try to be careful, I will try to explain you first with the help of this screen shot. So, I will try to make some changes in the command and I will try to explain you the changes with the screenshot. And, after that I will try to very quickly review that how these things are going to happen in the R console, right.

(Refer Slide Time: 23:44)



So, now in case if you want to have this bar plot with respect to the relative frequency, you can simply use the command here bar plot and then table divided by length and then you can see here now this values are in the form of relative frequency. So, anyway now you have learnt this type of bar plot very easily and this is the very basic fundamental default bar plot which is provided by the R software.

(Refer Slide Time: 24:07)



Now, suppose I have different types of needs and requirements. Suppose, I want to change the color of this bars, how to get it done? So, if you try to see here my command

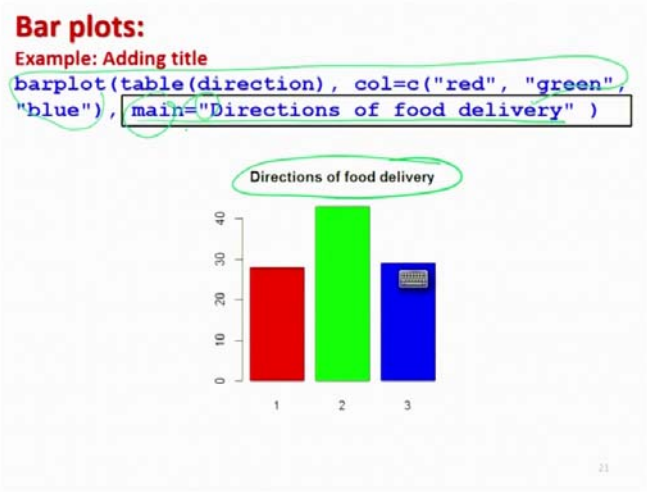
remains the same here bar plot and inside the parentheses this is stable direction, but now after this I have to add one option to do one job.

So, now I would like to inform you that if you want to add the colors in the bars, your command here is col and these colors are given in the form of a data vector. Suppose, I want to make the first bar to be red so, I will write down here within double quotes red then comma and then I want to make the second bar as green. So, I will write down here green g r double e n and then the third bar I want to make blue. So, I will write down here blue.

So, the one very important point which I want to inform you that whatever these options I am going to explain you here in detail, they will more or less will be valid in other commands also yeah. There may be some change, but usually in the common graphics these commands are going to work. So, you have to understand two things that if you want to make a change in the graphic, how you can do it and how you can give the values and what type of effect they are going to create on the graphics.

So, as soon as I add here this option, you can see here this is the option here. The first graph will becomes here red, second green and third here blue. So, that is a very simple option. Now, I will try to keep on adding different options to give a to add different features.

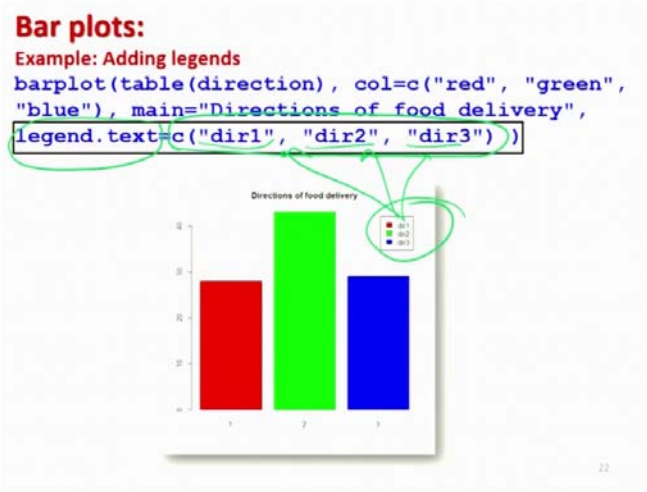
(Refer Slide Time: 25:33)



And, you have to observe that how I am adding and how they are affecting the outcome of the graphic. So, now, if you try to see I have just copied the earlier command up to this point here and then I have just added here one option which is here main, you can see here main here like this, right.

So, this main is going to add the title of the graph. Suppose, you want to add here directions of food delivery as one of the main title of the cor; you can see here and you can see here this is added here, right. So, the rule here is that if you want to add here the main title of the cor, you have to use the command here m a i n. And, then after that you have to give whatever you want to add inside the double quotes, you can see here like this ok, right.

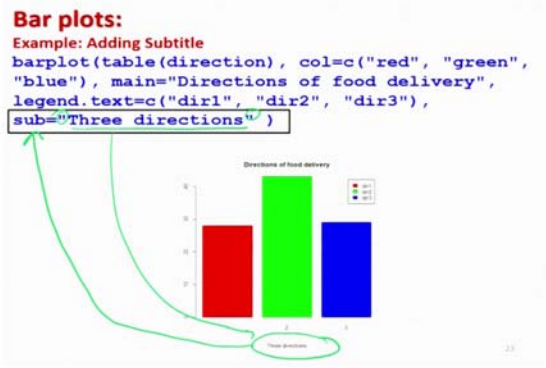
(Refer Slide Time: 26:16)



After this, suppose you want to add here some legends. For example you know, you do not know what is this different bars are indicating. For example, here you can see this red bar is indicating the direction number 1, green bar is indicating the direction number 2 and this blue bar is indicating the direction number 3. So, if you want to add this information here, you have this type of graphics here. You can see here where you are trying to write down dir1 again the red color, dir2 again the green color and dir3 again the blue color.

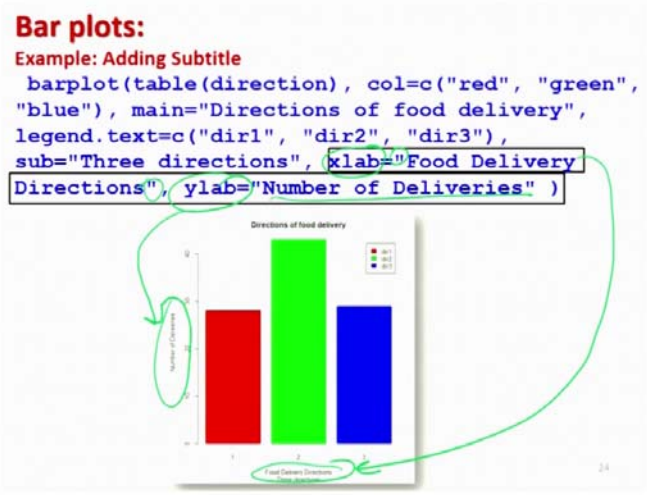
So, for that you have to add here an option legend l e g e n d dot t e x t and within this data vector you have to give in the same order the way you want to add the legend. And, I have given here within double quote dir1, dir2 and dir3 and same thing is now added here in the legend in the graphics. So, that is how you can see, you can very easily add the legends also without any problem, right.

(Refer Slide Time: 27:13)



Now, in case if you want to add here some subtitles on the of this graph. Suppose, I want to add here you can see here this I have written here now three directions. So, in order to do this thing, I am using here the command here sub s u b is equal to produced here. So, this is how you can see you can add here the subtitle now.

(Refer Slide Time: 27:31)



Similarly, if you want to add here some titles on the axis, then you have a command here xlab and ylab. We will try to look into the graphic on the x axis I have added here delivery direction, on the y axis I have added here number of deliveries. And, in order to do it I am using the command here xlab and then within double quotes I am trying to write down here Food Delivery Directions which is produced here.

And, then I am trying to write down here ylab is equal to Number of Deliveries, like as here and it is produced here. So, you can see here if you want to write something on the x and y axis as a title, then you can use the command here xlab x l a b and y l a b respectively, right. So, that is how we try to add over these different options in the graph and we try to work on this. So, now, I will try to just use my here this graph, earlier graph and I will try to show you that how these commands are working on the R software.

(Refer Slide Time: 28:26)

Like I will clear the screen and you can see here; now you have added here the colors. Then, in the next command you are trying to add the main title over here. So, now, you have to just observe as soon as I execute it what happens in the graphic. Now, you see this main title is going to be added here, directions of the food delivery and then I tried to use here this command for adding the legends.

So, you can see here that something will be added here as soon as I try to execute it. You can see here it is here, yeah you can adjust these things, but yeah this position can also be adjusted without any problem. And, then after that you try to add here subtitles. So, I try to simply copy and paste to avoid any mistake.

So, as soon as I try to execute it here, you can see here that this three directions here, that is added. And, similarly if you try to mark here something on the x and y axis here, you can write down here this here like this. And, see here number of deliveries is added on the y axis and food delivery direction that is added on the x direction.

And, you can see here this is the same command which I have shown you on my slides, right. And, after that there are many options and what I covered in this lecture here it is not the end of the lecture, not the end of the opportunities, but it is the beginning of the lecture and the beginning of the opportunities. I have taken here only very selected options which I believe that most of you know about them. So, I thought that you would always try to create a graphic which you know very well.

So, that is why I have taken here very common options which are required which are usually present in most of the graphics. So, that you get convinced that it is not difficult to create such graphics in the R software. Yes, the only thing is this for the first time you have to understand and then you have to add these different options in your command. But, if you try to recall and if you try to understand the advantage, you are sitting at one central office, you have created this command only once.

And, you are circulating it to all your offices all over the country, all over the world and from all the places you are getting the graphics which are using the same color, same title etcetera. So, now, the comparison of those graphics will become very simple in the long run. So, that is the sort of investment in the short run, but it has very long run advantages and besides this thing you can control each and every parameter in this graphic.

You can decide about the tick marks on the x axis, y axis, you can decide for the location of the legends etcetera etc. many things. So, now, this is your turn that please try to look into the help menu. Try to pick up one of the option and try to create a very simple graphic and try to keep on adding one option at a time. And, then try to see at the end

what are you getting, whatever is written there and whatever you are getting how they are matching to each other.

And, this will make you make you a better programmer and I am sure that you will learn how to create beautiful graphics. Now, in the next time I will try to continue some more graphics, but then I will not be going into this much detail. So, I would request that you please try to revise it and try to see which of the command is doing what, because similar type of commands I am going to use in the next lecture. So, you practice and I will see you in the next lecture.

Till then goodbye.