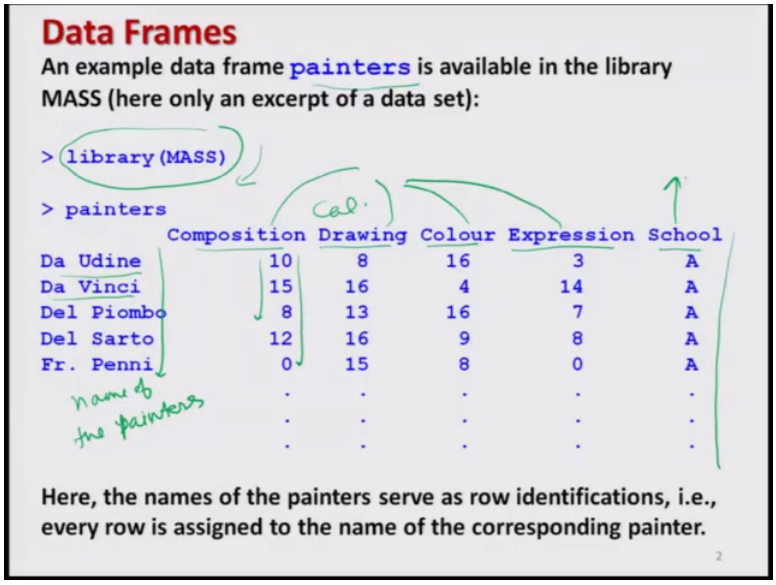
Introduction to R Software
Prof. Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology, Kanpur

Lecture - 31
Data Frames

Welcome to the lecture on Introduction to R Software. You may kindly recall that in the last lecture, we started a discussion on the topic of Data Frames. And we had considered some functions. Now in this lecture also, we will continue with similar type of topics and we will try to consider some more functions and their implementation in data frames. So, let us try to recall that in the earlier lecture, I had used a data frame or a data set that was called as here painters and this was contained in the library MASS.

(Refer Slide Time: 00:54)

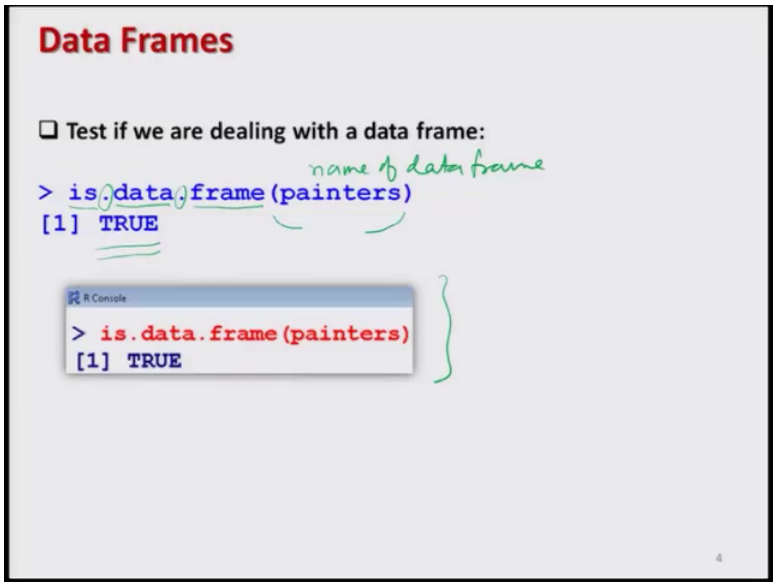


So, again I would request you that you please load your library with this here MASS and this data was actually like this; this was a long data set I had shown you in the last lecture. So, again I would try to use the same data set. Just for your remembrance in this data set, in the first column we had the name of the painters. We have some data on their composition, then drawing, and then color expression school.

So, all these variables they are arranged in see here columns. And the data on those variables is arranged in the rows. And out of these many variables, first row will give you the details of the first printer; second row will give us the details on the second

printer and so on. And the variables, composition, drawing, color, expression, they are some numeric variable and we have say this another variable here a school is a factor variable. So, they were certain factored in terms of a, b, c, d, e, f, g, h, right.

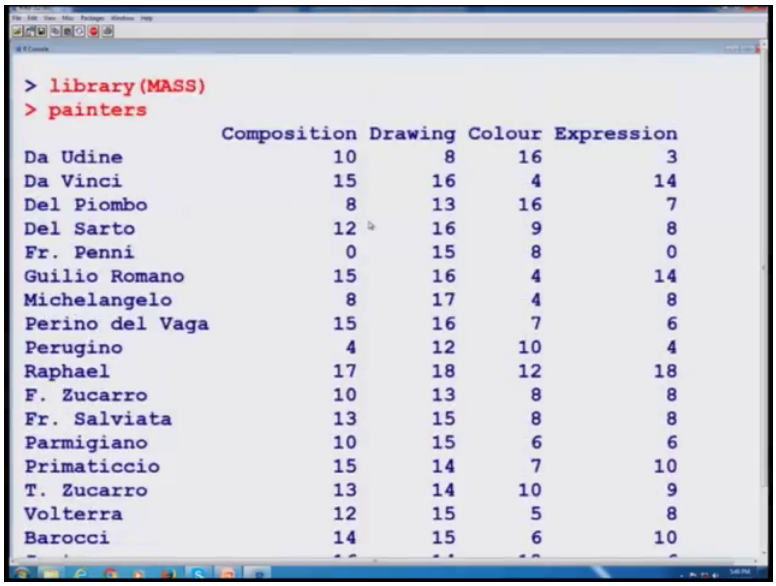
(Refer Slide Time: 02:17)



So, now we continue with this thing. So, that was the data set that we use last time. So, now, whenever you get a data set, you need to check whether it is really a data frame or something else; or that can be a matrix or something else. So, in order to test whether a given data set corresponds to the structure of data frame or not, we try to use the command here as usual we have done it earlier also another format. Here, I would use here is data frame and these three words are joined by the full stops as a delimiter.

So, the command is dot data dot frame and inside the arguments, we have to write down the name of data set or name of data frame. And once you try to test it, it will give you the answer in terms of true or false, which are the logical values. And you can see here we have tested here and this painter's data set comes out to be a true value; that means, yes the data set painter is a data frame. Now we try to first check this statement whether this is a correct thing or not on the R console.

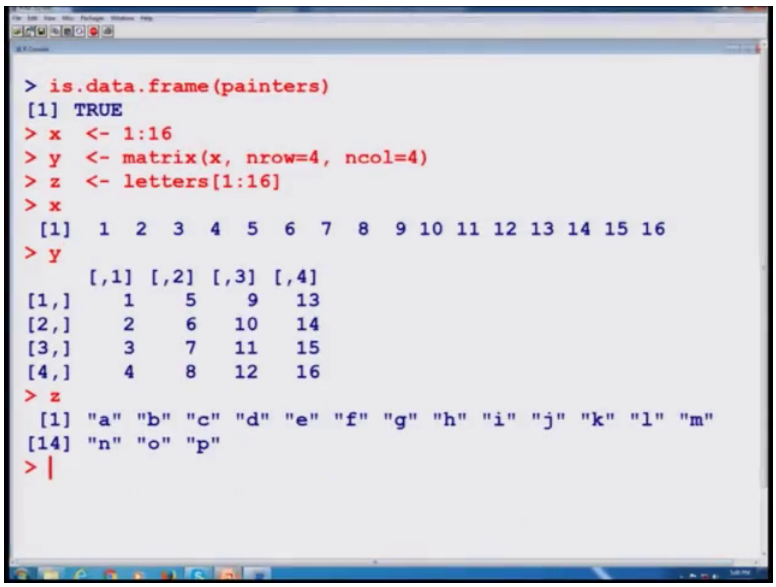
(Refer Slide Time: 03:43)



```
> library(MASS)
> painters
                Composition Drawing Colour Expression
Da Udine                 10       8     16          3
Da Vinci                 15      16      4         14
Del Piombo                8      13     16          7
Del Sarto                12      16      9          8
Fr. Penni                 0      15      8          0
Guilio Romano            15      16      4         14
Michelangelo              8      17      4          8
Perino del Vaga          15      16      7          6
Perugino                  4      12     10          4
Raphael                  17      18     12         18
F. Zucarro               10      13      8          8
Fr. Salviata             13      15      8          8
Parmigiano               10      15      6          6
Primaticcio              15      14      7         10
T. Zucarro               13      14     10          9
Volterra                 12      15      5          8
Barocci                  14      15      6         10
```

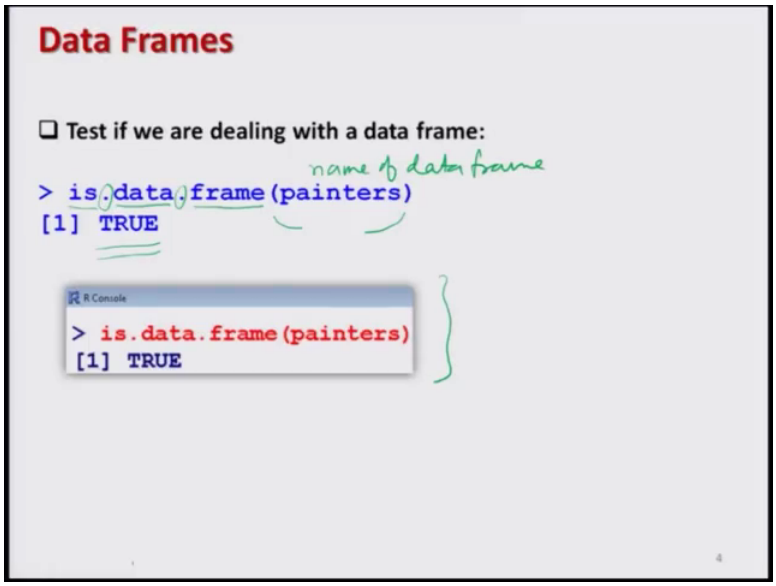
So, what we try to do here, that we come to R console. First, we try to load the library here MASS. And then after that the painters data set will come like this yet that you can see over here that is a huge data set anyway. So, after this I try to write down this command, whether this painters data set is what is it a data frame or not the answer comes out to be here true right.

(Refer Slide Time: 04:01)



```
> is.data.frame(painters)
[1] TRUE
> x  <- 1:16
> y  <- matrix(x, nrow=4, ncol=4)
> z  <- letters[1:16]
> x
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16
> y
     [,1] [,2] [,3] [,4]
[1,]    1    5    9   13
[2,]    2    6   10   14
[3,]    3    7   11   15
[4,]    4    8   12   16
> z
 [1] "a" "b" "c" "d" "e" "f" "g" "h" "i" "j" "k" "l" "m"
[14] "n" "o" "p"
>
```

(Refer Slide Time: 04:13)

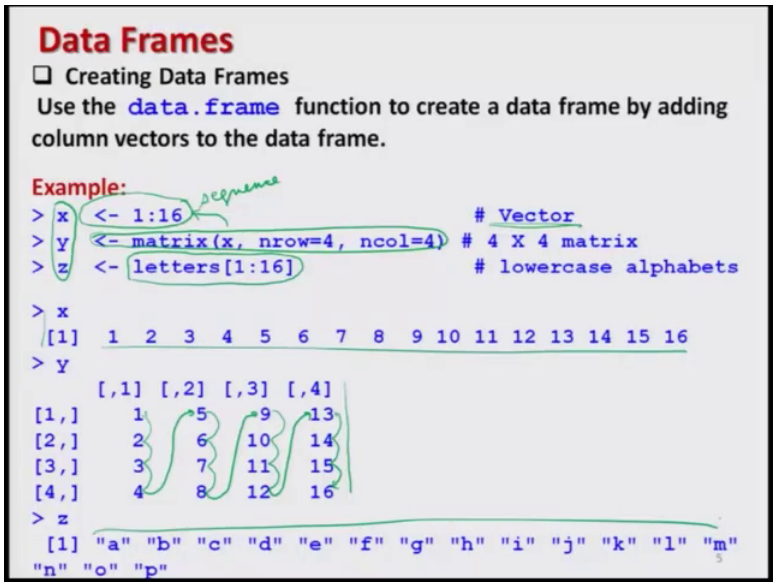


So, now I am confident that, the data set which I am going to use in my examples, it is a data frame. So, after this we come back to a slides and the next question comes, how to construct a data frame? We are using data frame, we are testing a data set to know whether it is a data frame or not, but in case if I can understand; how a data frame is created, possibly I will have a more, deeper inside into this concept. So, why not to create a small data frame, and try to see what happens.

So, in order to create a data frame, first we have to recall the basic concept that we had discussed in the earlier lectures, that in data frame we can combine different types of variables. So, what I try to do here that I try to generate three different types of variables: one number, numeric variable, another matrix and say third is alphabet. And then I try to combine them together and see what happens with that data frame or how to combine them in the format of a data frame.

So, I will generate here three variables: one numeric, one matrix, and one character, using the alphabets. And then I will try to combine these three variables together and I would like to combine them in the format of a data frame. Then let us try to see what happens.

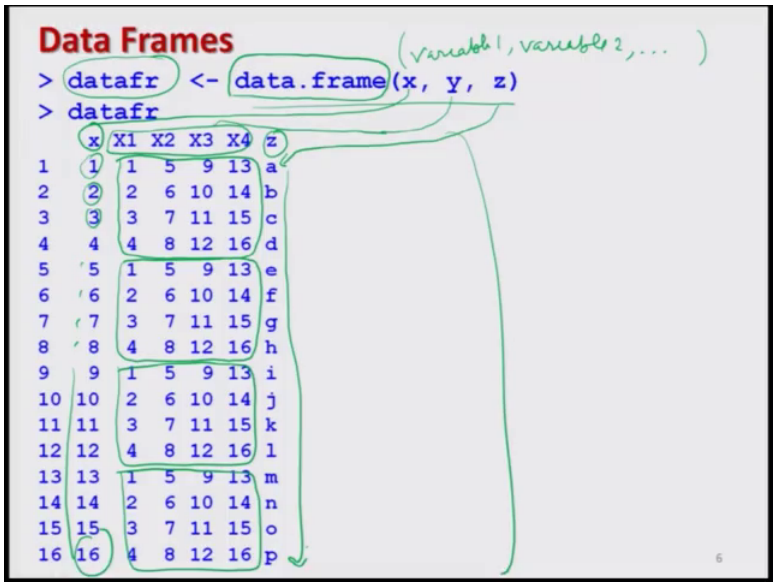
(Refer Slide Time: 05:56)



So, we have taken here three variables: x, y and z. First variable is a sequence: this is here sequence. Sequence of numbers from 1 to 16, and this is a vector. Similarly another variable here y this is here a matrix, this is a 4 by 4 matrix which is constructed by the 16 elements in x, and then the third variables z contains a sequence of 16 alphabets in lowercase, by using the command letters from 1 to 16. So, this will give me the first 16 alphabets in lowercase. So, the outcome of x, y, z looks like this; x this is the sequence of 1 to 16 number, this matrix if you try to see this is something like this: 1, 2, 3, 4 and then 5, 6, 7, 8 then 9, 10, 11, 12 and then here 13, 14, 15, 16 and then here z this is a sequence of alphabets from a to p.

Now, the next step is this I want to combine it. And this is the place where comes the difference. In order to combine the data frames, we have to use here a command data dot frame.

(Refer Slide Time: 07:26)



So, I use here a command here data dot frame, and inside the arguments I try to write down all the variables say: variable 1, variable 2 and so on separated by comma. That is what I have to do and this is the syntax of the function data dot frame. That will create a data frame of x, y and z.

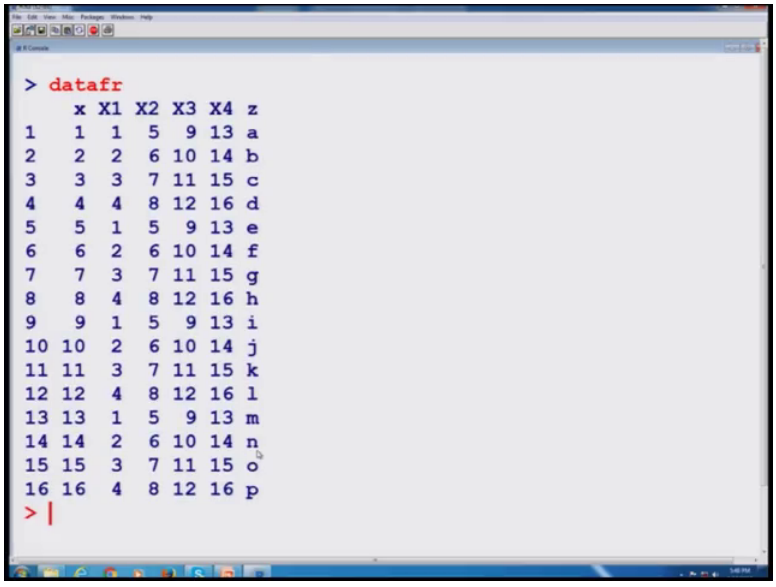
So, once you try to enter it, you can see here now this data frame or the outcome of this data frame is stored in a variable say datafr that is a short form of data frame just for the sake of understanding. And this comes out to be in this format. So, now, first we try to understand what is really happening if you try to see we have here three variables: x, y and z. This here x is coming over here x and this was a sequence from 1 to 16. So, you can see here this is here 1, this is 2, 3, three and up to here this is coming here 16.

So, that is my first variable that was the sequence from 1 to 16. Now my y variable was this was here a matrix. So, now, there is no y here you can see well, this is my here y and you can see here, this is the matrix 4 by 4 matrix which we have defined using the 16 elements. And then this is repeated here. Like this you can see here and third variable which we have combined is z; which is a sequence of here alphabets from 1 to 16 and this is the variable. You can see here a, b, c, d up to here p. So, now, you can see this is the structure of the data set that we have got, but this is a data frame. This has all the characteristics of data frame which is defined in the R package. So, let us try to first do it

over the R console and try to see what really happens. So, I try to define here say here x, y and z say x then here y and then here z.

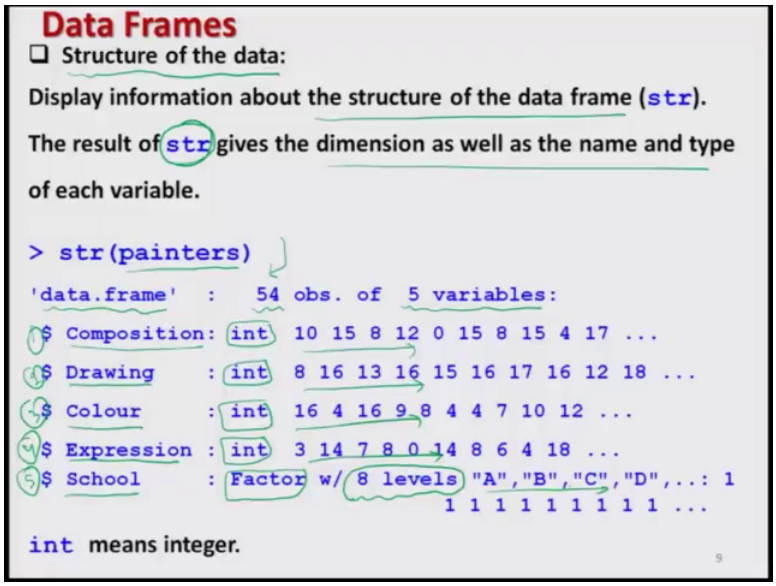
So, you can see here this is my x, this is my here y and this is my here z. And now I try to use here this command to create the data frame. And this comes out to be here like this. So, now, I have created a data frame.

(Refer Slide Time: 11:04)



Let us try to see how it looks like. You can see here this is like this; this is the same outcome that I have copied and pasted on the slide. So, now, let us come back to a slides and here are the screenshots the screenshot of the outcome of x, y, z variables and here is the outcome of the data frame. So, you can see here these are the same thing that we have operated over the R console.

(Refer Slide Time: 11:38)



Now, after this once we have understood how a data frame is created? We try to explore some other aspect. The next aspect is that we would like to know, what is the structure of the data? Structure of the data; that means, in a data frame there can be a numeric, variable, character or string or different types of things can be combined, so whenever we have got a huge data set, we cannot look into the entire data set manually, but we would like to use some command to know, what is the structure of data set? What types of variables are there? How many variables are there? How many observations are there and so on?

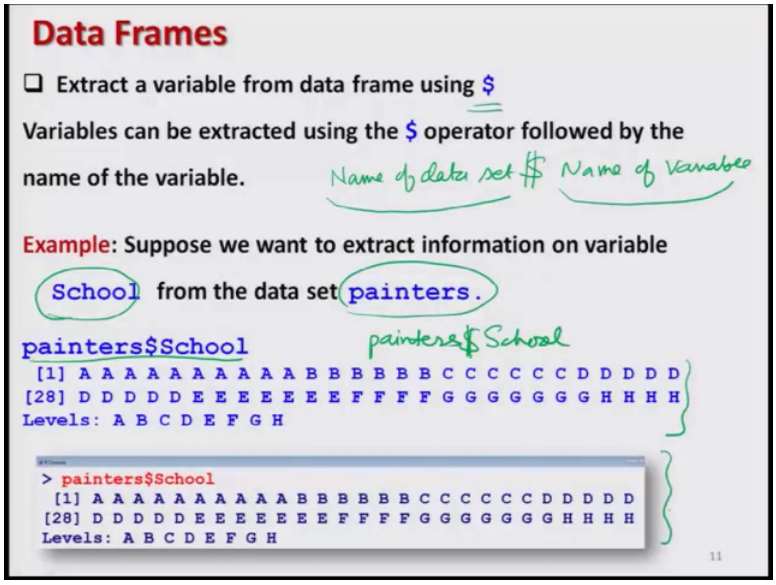
So, all such information can be obtained, using the function str. This function displays information about the structure of the data frame, and it give us the dimension as well as name and type of the each of the variable. For example, here we are dealing with the data set painters. So, you can see here, as soon as I type here str inside the argument the painters which is the usual syntax: str inside the bracket try to write down the name of the data set. As soon as I enter here we see this tells us that this is a data frame, there are 54 observations and there are 5 variables. What are those 5 variables? Number one this is composition, number two this is drawing, number three this is color, number four this is expression and the last fifth variable is say school.

Now it is telling us about the nature of the variable. So, this composition is say integer, drawing is integer, color is integer, expression is integer and school is a factor variable.

And here briefly it is trying to show us some sample of the values, in all the integer values. For the factor variable it is trying to show that there are 8 levels, and if you remember the levels were a, b, c, d, e, f, g, h.

So, it is trying to show us a sample of these things right. So, in case if you want to get any information about the structure of the data frame, please try to use the command str and here is the say here the screenshot of the outcome, but definitely we will try to use it ourselves. So, str so you can see here this comes out to be like this. So, now, we come back to our slides.

(Refer Slide Time: 14:43)

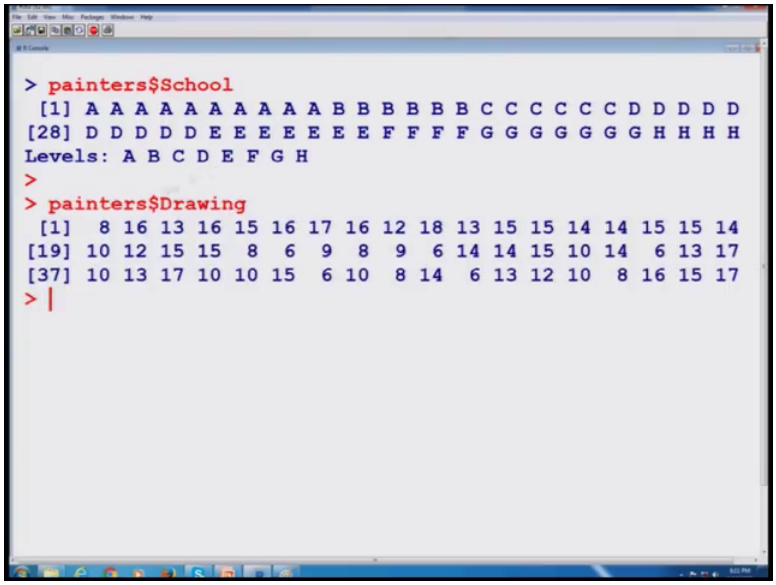


Next, we try to consider another option. Suppose we are interested in extracting some information about a particular variable from the data frame. You may remember that a data frame consists of different types of variable and we also have done this topic earlier in the lecture, but just for the sake of completeness I would try to repeat it here quickly.

So, in order to extract some information about a particular variable, we try to use the symbol, dollar. And the rule is very simple. First you write the name of data set, and then write name of variable and then join them together by your dollar sign, that is; that simple syntax by which I can get the information on particular variable from the data set. So, why not to do this in the example that we have taken earlier in the data set painters?

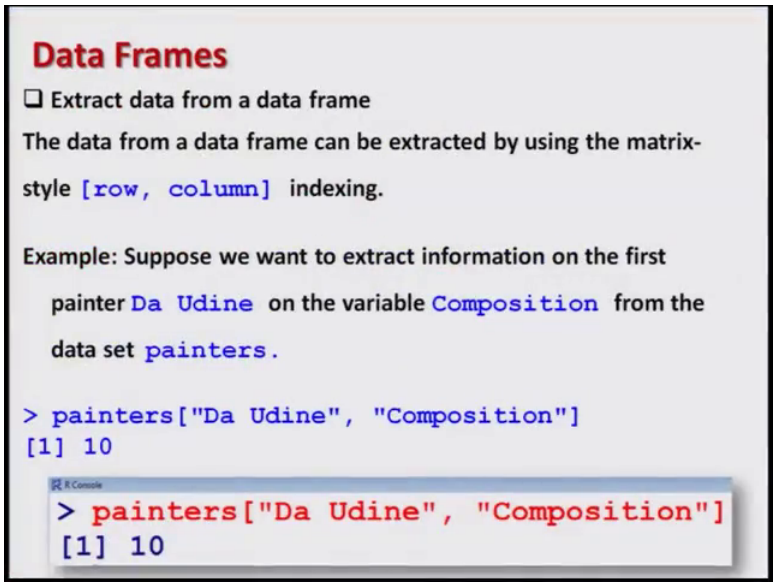
So, if you remember in the painters you had here several variables, composition, drawings, color, expression and school. And suppose I want to extract information about the variable say school from this data set called as painters. So, what I have to write name of the data set, painters, name of the variable say school, and then I try to join it by a dollar sign and that is what I am trying to do it here and as soon as I do it here you can see that I get this type of outcome over here, and this is the screenshot of the outcome, but let us try to do it over the R console for the sake of understanding.

(Refer Slide Time: 16:53)



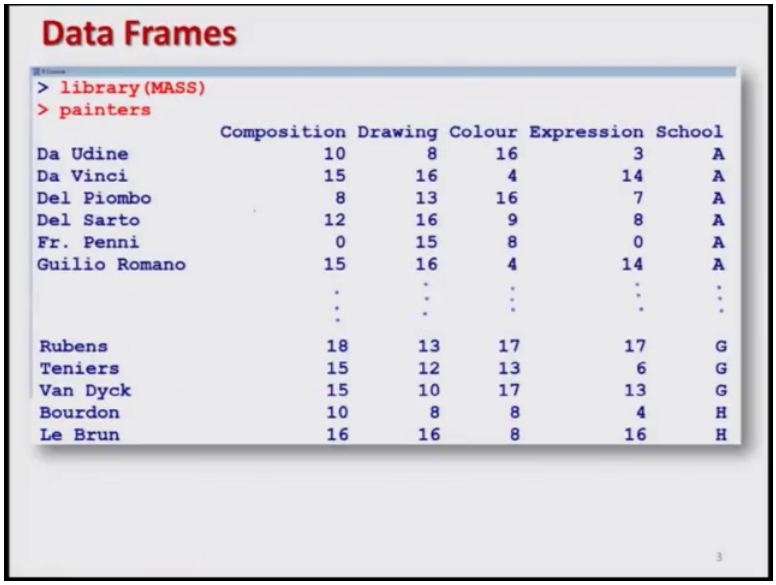
So, now suppose if I want to have some information on say another variable say here drawing. So, what I have to do here, that I have to write down here the variable name drawing, in the data set painters and joined by this dollar sign. And as soon as I enter here you can see I get the information here on the variable drawing contained in these data set painters. So, let us come back to our slides and consider another aspect.

(Refer Slide Time: 17:24)



Now, suppose my question is how to extract some data from a data frame? In order to extract a particular data from a data frame, we try to use the matrix style combination; that means, I try to use here something like square bracket and here I try to specify the name of a row and column, from which I need to extract the information. Let us try understand this through an example, but before tell if I can start to see what we want to do? So, if we try to see here in the earlier slide, I had given you the information about this data frame. If you try to see here we have here a painter Da Udine and this is the name on the first row.

(Refer Slide Time: 18:21)

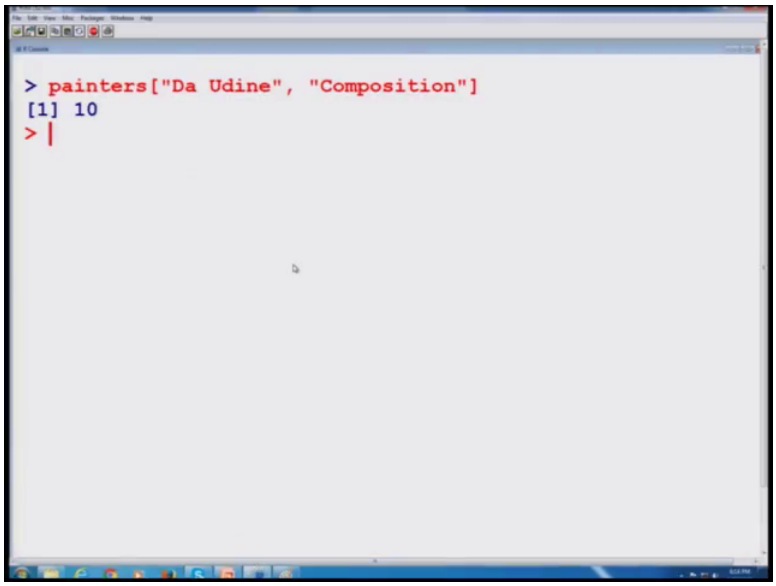


And suppose I want to have information on the composition; that means, I want to extract this value this data.

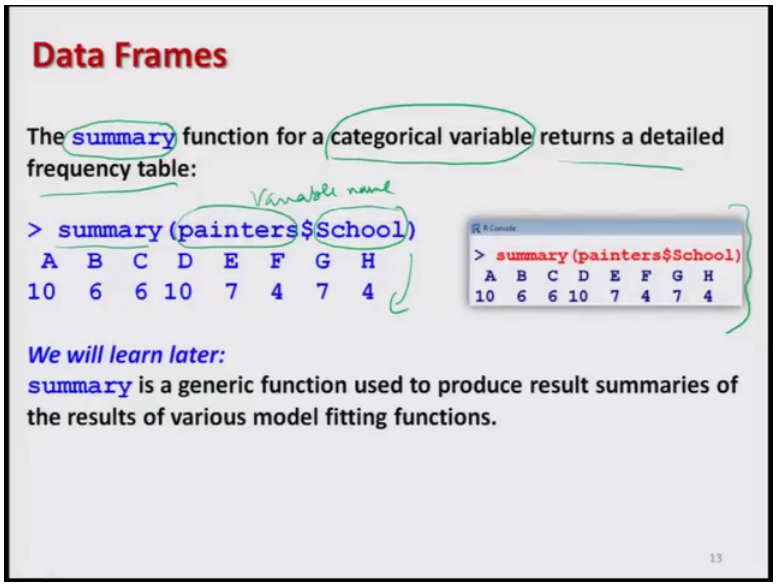
So, I can consider as if this information is given to us in the form of a matrix. And then I try to write down the command here as like this over here, that I will write here the name on the row, but notice one thing I am trying to write it inside the double quotes and then I try to write down the name of the variable for which I need this information. And this is again written inside the double quotes.

So, as soon as you try to enter here you get here the value therefore; the same value which I shown you earlier. And this is the screenshot here. So, why not to do it over the R console and try to see, what do we get here? You can see here you are getting the same thing. And similarly, if you want to extract any particular information you can just do it.

(Refer Slide Time: 19:41)



(Refer Slide Time: 19:50)

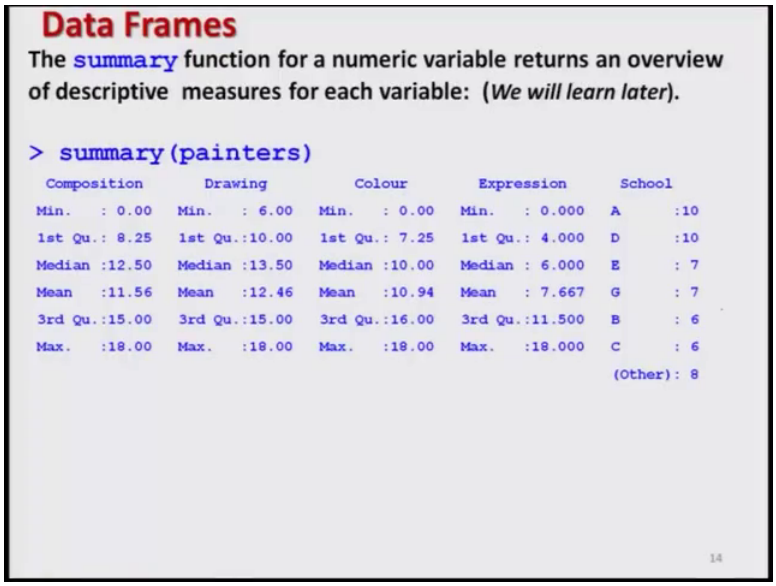


So, now we come back to a slide; and try to see what more can be done with this data frame. Here, I am just going to take some example just to show you that these things are simply possible on data frame. Well, we have not done these functions up to now, but definitely; we will try to do them in more detail in the forthcoming lectures.

So, here I would try to simply show you an illustration, using the function summary. There is a function summary that we will do after some time, but here I can just briefly tell you that summary is a function. When this function is used over a categorical variable, then it returns a detail frequency table. What is frequency table that we will try to understand although; I am sure that you must have done it in class 10 or so. Suppose, I want to create here a frequency table, for the categorical variable say school from the data set painters.

So, I try to use this as variable name and I try to use here function summary inside the arguments, I try to write down the name of variable and I get here this time of outcomes. You can see that this is here the screenshot you can try it yourself we will do it in more details when we come to the summary function right. So, you can see here that inside a data frame I can extract any particular variable and I can use different types of function.

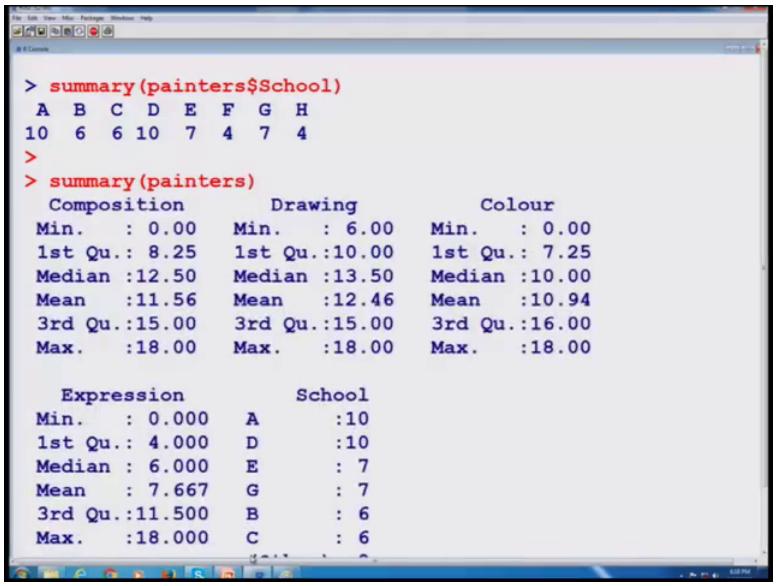
(Refer Slide Time: 21:32)



Similarly, as a next example; suppose, if I am trying to use the summary function over the entire data set painters, then the outcome will come out to be like this and here what I am trying to show you that when I am trying to use the data set painter then some of the variables are quantitative, they are numeric variable, and some are qualitative; that is categorical variable and for them you can see here for the quantitative variable it is trying to give us a minimum value first quartile, median, mean, third quartile, maximum value and so on.

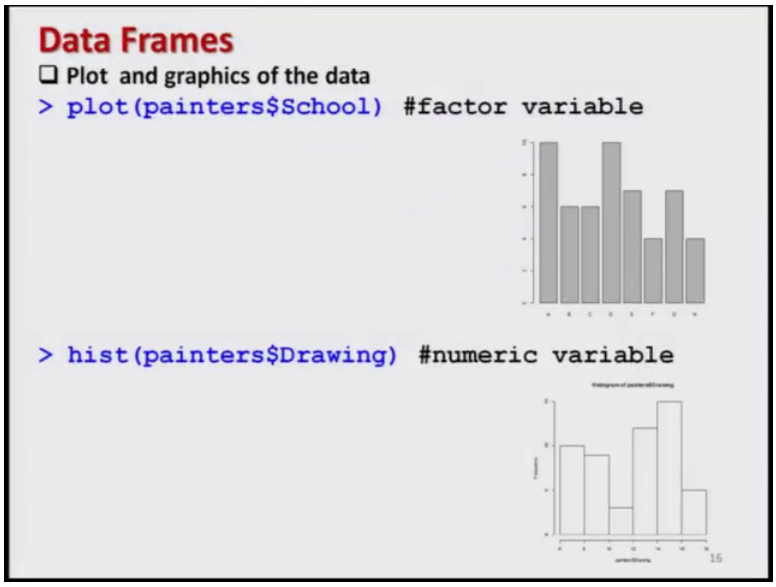
And this is true for all the quantitative variables and for the categorical variable here school it is trying to give us the frequency right and this others is a the frequency of the two categories F and here H which are here 4 and 4, that you can see yourself in the data set record just try to do this thing for the R console and try to see what happens. So, you can see here, I am getting here over this thing.

(Refer Slide Time: 22:45)



And then, I will try to write down the summary of painters data set. You can see here you get the same thing right. This is the screen shot of the same outcome. And similarly, I would try to give you another illustration.

(Refer Slide Time: 23:03)

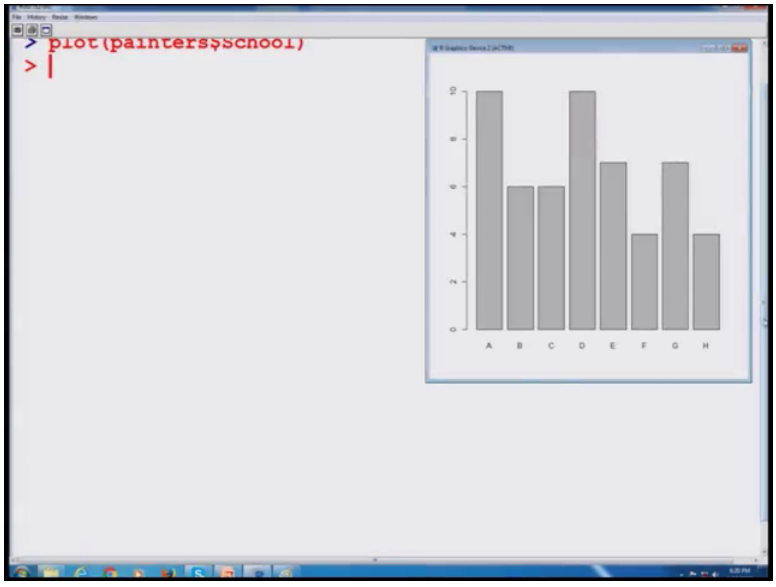


For example, in case if you want to do some graphics on these variables, that it is also possible. Although; we have not discussed these functions of to not that we do later on, but my objective once again I will say is here simply to show you that different types of functions can be operated over the R object data frame.

So, here I am trying to use here a variable school and which is a factor variable, and then in the second example I am trying to use another variable here drawing, which is a numeric variable and if you want to do some graphic with the factor variable one possible option is to use the function plot, and similarly for the numeric variable one option is to create histogram.

So, you can see here that here we are getting these types of graphics and we can also create these graphics over the R console to see the outcome.

(Refer Slide Time: 24:00)



So, you can see here that this comes out to be the graphic. And similarly, if you try to go for another say histogram, so here you can see here; and this is also possible here. This is here and this is the same screenshot on this one. So, by these two examples; I am not trying to give you the details of the functions like summary or say plot or histogram, but my modest objective is to show you, that different types of function can be used over the data frame also and they give us some suitable outcome.

So now, I would like to stop here and I would request you to have a quick look on the different types of examples from your book, and try to attempt some questions. And in the next lecture I will try to take some more aspects on the data frame, till then goodbye.