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

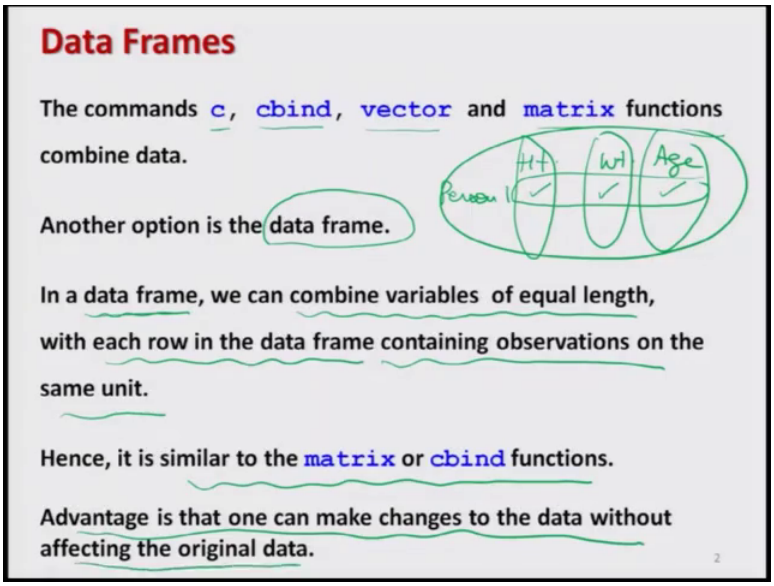
Lecture - 30
Data Frames

Welcome to the lecture on the course Introduction to R Software. Now in this lecture we are going to take up another topic, which is about Data Frames. What are data frames data frames are a special type of R object in which we can store the data, we can call the data, and we can do different types of mathematical manipulations over that data set. So, whatever operation we can do on a spreadsheet, how to get them in R software that is the objective of this lecture, right.

So, can I try to start with the lecture and first we try to understand what are we really going to do. If I ask you what is a data set, you can imagine that data set is a collection of data on say different variables for example; I can collect the data on say height weight age of say 20 different persons and can arrange them in rows and column. So, that is a data set. Now when this data set is collected individually or say separately or say variable by variable so; that means, I can have 3 sets of data one on height another on say this weight and say third on age.

So, these are 3 different (Refer Time: 01:41) sets and when we want to combine them together we have learned that there are different types of options in R by which I can combine for example, the 3 data sets, into one data set right. And that would be called as a data frame. So, first question comes how to combine this 3 variables together or the data on these 3 variables together.

(Refer Slide Time: 02:16)



So, we have learned in R that earlier we have learned about the command c, cbind vector or say matrix. These are the popular functions that we have learned which help us in combining the data for example, matrix can combine the data into rows and columns. Similar to these functions we have another option which is called as data frame.

Now, in this data frame what happens that we have some information on different types of data sets. And these data sets are containing the variables and each of the variable is containing then information or individual unit. So, you can see here that I can arrange the variables is has a column one variable number 2 column 2 variable number 3 column 3.

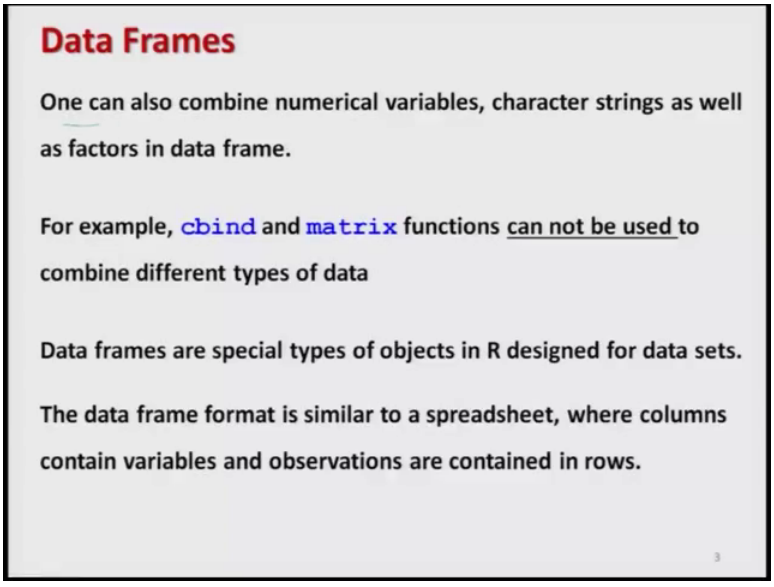
And then, now I can arrange the data in say rows say person number one then person number 2 then person number 3. So, whenever we are trying to consider a data frame than in the data frame we can combine variables of equal length with each row in the data frame containing observations on the same unit. What does this mean for example, my first column is height second column is weight and third column is here age something like this over here height weight and see here age; now, person number one height of person number one weight of person number one and age of person number one.

So, this is arranged in a row and variables are arranged in a columns. So, this altogether can be called as a data frame. So, if you try to see this operation is very much similar to matrix or say c bind function. Then what is the difference between the 2. Why should I

use the structure of data frame rather than using the matrix or say c bind function? The different is the following, one of the advantage with the data frame is that one can make changes to the data without affecting the original data, means I can make different types of manipulation over the data frame and the original data set will not be altered. That will remain as such that is the first advantage.

And second advantage is that that in our data frame one can combine the numerical values character string as well as factors.

(Refer Slide Time: 04:53)



So, different types of these variables can be combined together. For example, you have seen in any spreadsheet for example, the first column is name that is a string say mister x misses y mister z and so on.

Then this can be their height weight age which are some numbers. Then the fourth column can be gender say male or female. So, that is a factor. So, all these things are possible in data frame all different types of data sets can be combined under the structure of data frame. And whereas, if you try to compare the data frame with the matrix or say c bind function, then this possibility is not available in c bind and matrix function c bind. And matrix function cannot be used to combine different types of data.

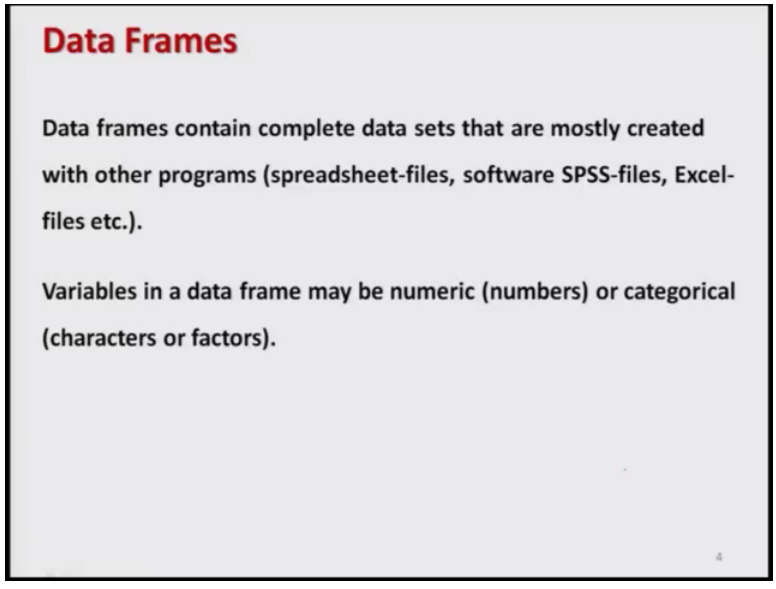
That is another advantage of using the setup of data frame and data frames are nothing, but they are simply some special types of objects in R and these objects R designed

especially for dealing with data sets. And the format of our data frame as I said earlier is very much similar to a spreadsheet. Where the columns contain variable and observations are contained in rows, first column second column third column columns are always vertical and rows are always horizontal.

So, when I say a column one that can say contain data on say on height second column can contain the data on weight third column can contain the data on say age. So, first column second column third column they are the they are containing the variable names. Now when I go for the rows I try to take the person number one and record it is age weight and send other information and I put it in the row number one. Then I take the person number 2 and then I record his age weight height etcetera and I put them in the row number 2. So, the data is arranged row wise that is horizontal.

So, that is the main thing what you have to keep in mind that what really happens in a spreadsheet that is the usual structure and when I talk of data frame.

(Refer Slide Time: 07:34)

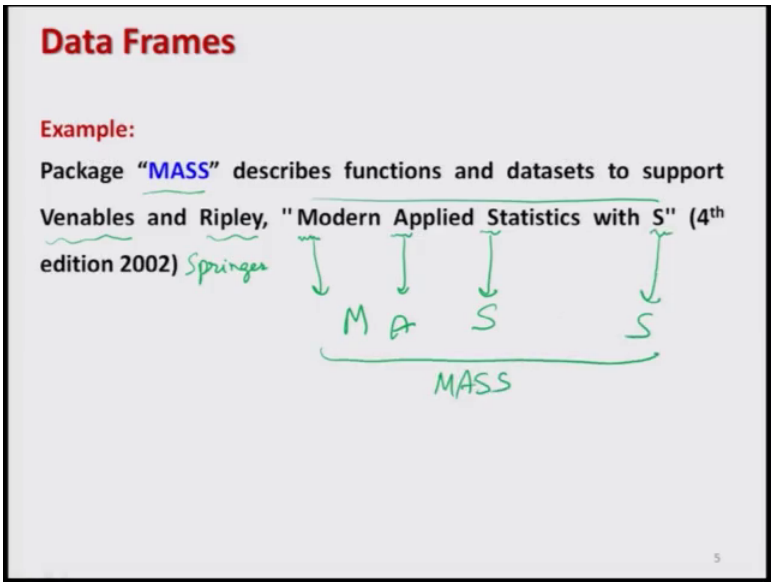


Data frame not only considers the data which is collected in the say (Refer Time: 07:41) spreadsheet like as for example, I took a software say MS excel, that is not always necessary, but another advantage of using data frame is that data frame can consider the data from other types of files also. For example, there is a statistical software what is called as SPSS right.

So, now if somebody has got a data file which is an output or from an SPSS that can be used by using the concept of data frame. Similarly, the output of other software also can also be called in R using the setup of data frame. That is another advantage of using the setup of data frame in R all right. And means finally, you have to just keep in mind that the variables in the data frame may be numerics which are taking some numbers as their observations or they can be also categorical variables; that means, which are containing the observation in the form of some characters or say factors. So, you can see that data frame is a very general setup because can combine different types of information on different types of variable and then after that we can play with the data.

So, in order to understand about this data frame, what are we going to do? We are going to use a package which is in built inside the R software.

(Refer Slide Time: 09:23)



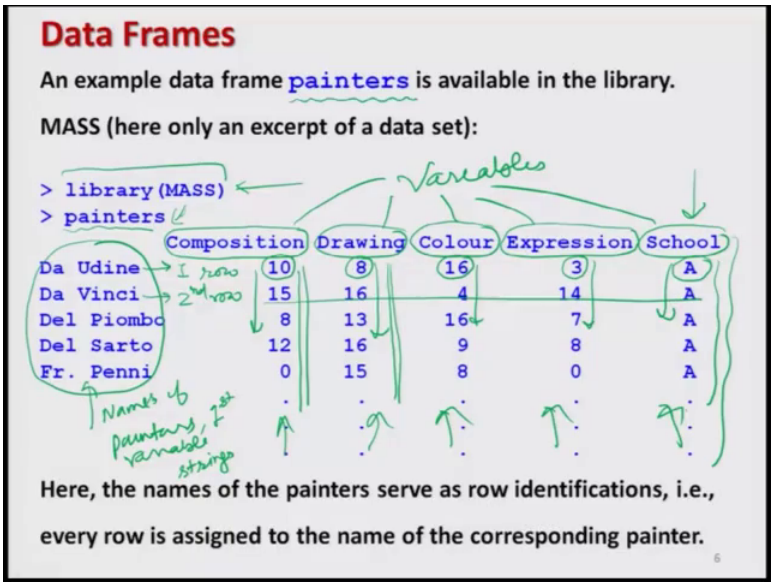
This package is called here as say MASS; capital M capital A capital S capital S. Actually this package contains the functions and data sets which were used in a book that was written by professors venables and Ripley and the title of the book is modern applied statistics with s, and in it is fourth edition published in 2002. You may recall that in the initial lecture when I introduce R, I have told you that R was basically developed on the lines of another language that is called as S language.

And in order to understand that s language venables and Ripley, they had written a book modern applied statistics with S and it is fourth edition and this was actually published

by springer right. And this was a quite popular book and so in R, we are also going to use the same data set that was provided with that book and named MASS.

Here it is a short form of say modern first letter of modern m, then first letter of say applied say here a first letter of s statistics say here s and with this here s. So, this is creating here the package MASS right. And this MASS is containing different types of variable different types of data set. And here in order to understand the functions utilities and different types of operation with the a data frame, we are going to use one of the data set which is contained in this package MASS right. So, one example data frame that is available in this package MASS is called as painters.

(Refer Slide Time: 11:22)



That is available in the library of MASS right. This painter is actually collecting the data on the work of various painters right. For example, there are some famous painters in this world like as Da Udine, Da Vinci and so on.

So, what this data set is trying to do is that it has collected the observation on their painting work. For example, it has collected the data on their composition, drawing, color, expression and school. The school does not mean our primary school or say or say colleges, there are some schools of arts right. So, they are also getting the information that how much is the composition what is the nature of drawing color expression and say and from which of the school that painter is coming right. So, you can see here that the

observations on different variables are collected and these variables are here composition drawing color expression and say school.

So, first we try to understand; how does this data set actually look like. So, in order to understand what we are going to do that first of all we have to load the library, if you remember in the initial lecture I have shown you that whenever you want to use anything specific which is not included in the base package of R that has to be uploaded or that has to be loaded say specifically.

And in order to do so we are going to note the library MASS and which is loaded by like this we try to write library and inside the arguments, we try to write down the package that is MASS. And after this once this package is uploaded, and then I try to see what is happening in the data set painters for that. I simply try to write down painters and I enter and I get here this type of information.

Well, I will try to show you on the R console also, but before that let us try to see what type of outcome, we are going to get you can see here that here these are the name of the painters. And here this is information on say composition here is some data, on say their drawing here, you can see there is some data on say color in this column the data is on say another variable expression. And here there is another variable in which they are trying to give the information on say school their schools are categorized say a school a school b school c and so on.

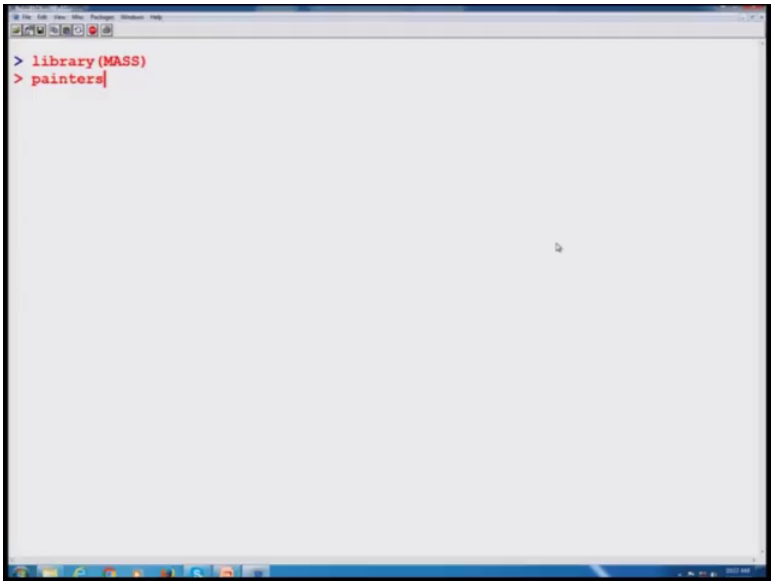
So, you can see here what is now happening first thing these are here variables. And the data on say any individual painter for example, the first painted the Udine that is contained in the first row; that means, that composition for the first printer is the value or the observation on the drawing of first painter is 8, the observation on the variable color for the first painter is 16. The data on the variable expression for the first painter is 3 and the first printer belongs to a, school a school whose category is a this is for the second row and this contains the data for the second painter and so on right.

So, you can notice here one thing that this is here the first variable. This painters name this is my first variable. And you can see here that these are strings. These are the names whereas this data here 10, 15, 8, 12 or in the third column on the drawing, this is 8 16 13 and so on. These are some numeric values and the data on the say variable here say

school this is in the form of same alphabets a b c d and so on, which are denoting a factor right.

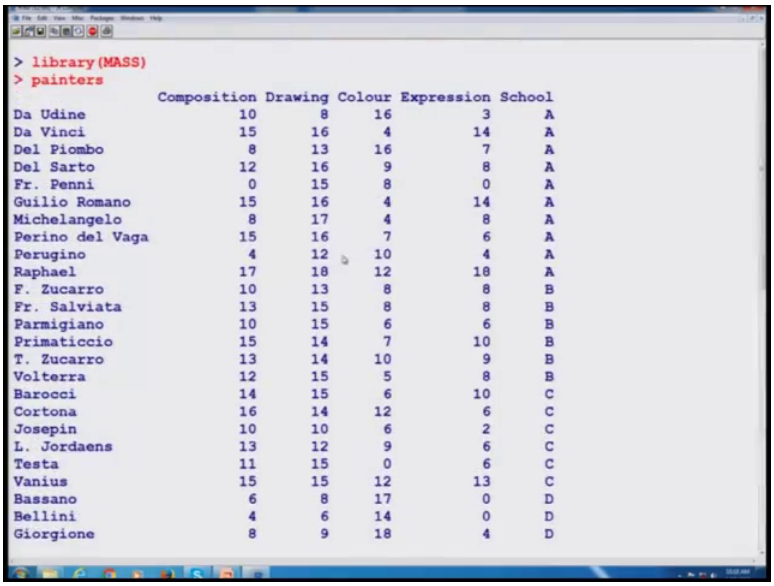
So, you can see here by these here columns we are trying to identify different types of variable and in the rows we are trying to give the data for individual this painter right. Let us try to do it over the R console and see how it looks like and what really happens. So, first I try to say load my package MASS using the library function, and then you can see there is no error.

(Refer Slide Time: 16:12)



So, I believe that this data has been loaded, now I try to see means I can also view the content of the data and if you try to recall in the initial lectures we have learned that how to see the content of a package and everything, so that you can refer to those lectures and can see how to get it done. Here my objective is that as I told you that, I want to use a particular data set and whose name is painters.

(Refer Slide Time: 16:50)



```
> library(MASS)
> painters
               Composition Drawing Colour Expression School
Da Udine                10       8     16          3      A
Da Vinci                15      16      4         14      A
Del Piombo               8      13     16          7      A
Del Sarto               12      16      9          8      A
Fr. Penni                0      15      8          0      A
Guilio Romano           15      16      4         14      A
Michelangelo             8      17      4          8      A
Perino del Vaga         15      16      7          6      A
Perugino                 4      12     10          4      A
Raphael                 17      18     12         18      A
F. Zucarro              10      13      8          8      B
Fr. Salviata            13      15      8          8      B
Parmigiano              10      15      6          6      B
Primaticcio             15      14      7         10      B
T. Zucarro              13      14     10          9      B
Volterra                12      15      5          8      B
Barocci                 14      15      6         10      C
Cortona                 16      14     12          6      C
Josepin                 10      10      6          2      C
L. Jordaens             13      12      9          6      C
Testa                   11      15      0          6      C
Vanius                  15      15     12         13      C
Bassano                  6       8     17          0      D
Bellini                  4       6     14          0      D
Giorgione                8       9     18          4      D
```

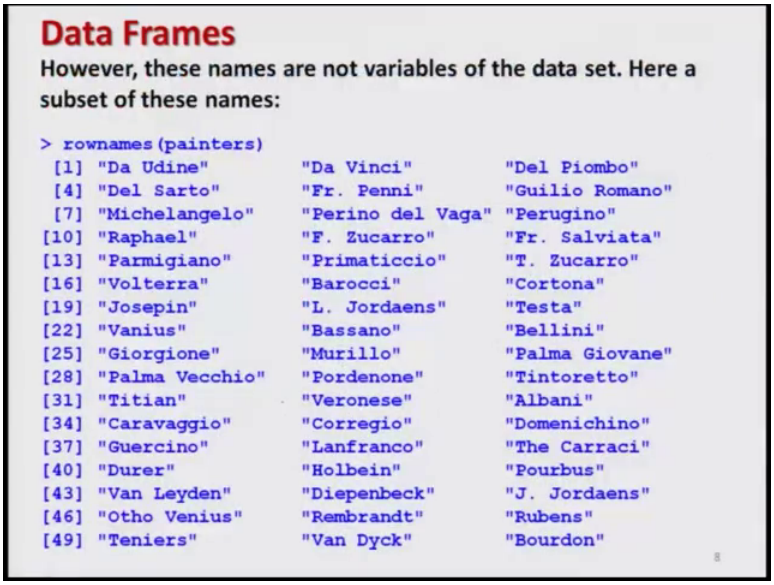
So, now I try to see what is contained in the data set painters. So, you can see here this is quite long, but if you try to see here gradually, I will try to explain you can see here that this is a quite long data set where the first variable is here, say actually the first variable is here the names of the painters, second is the composition, third is drawing, fourth is here color, fifth here is expressions, and last variable. Here is say school and you can see here because of the name of the first painter and whose values for composition is 10 8 for drawing 16 for color 3 for expression and a for school.

And you can see here that there are different types of schools also here a, here it is b please try to concentrate where I am highlighting then b, then there is under school here c and then school here d. And here you can also see in the first column that there are say the names of different painters and if I try to scroll it you can see here that this data is something like this. So, there are. So, many school d school e school f school g and school h right. So, these are the different categories which are the factors and so on right.

So, you can see this is how the data on painters look like, now we have different types of jobs that I can operate. So, what now I will do I will try to take different types of tasks which usually someone likes to do over a data set and we will try to see how we can do it. Now in the next slide I have given a snapshot of this data set. And you can see here that I have not given you the complete snapshot, but here is something like a dot because

this was a pretty long data set. And it was not possible to take the complete screen shot right otherwise you cannot see anything, but I have given you the complete idea.

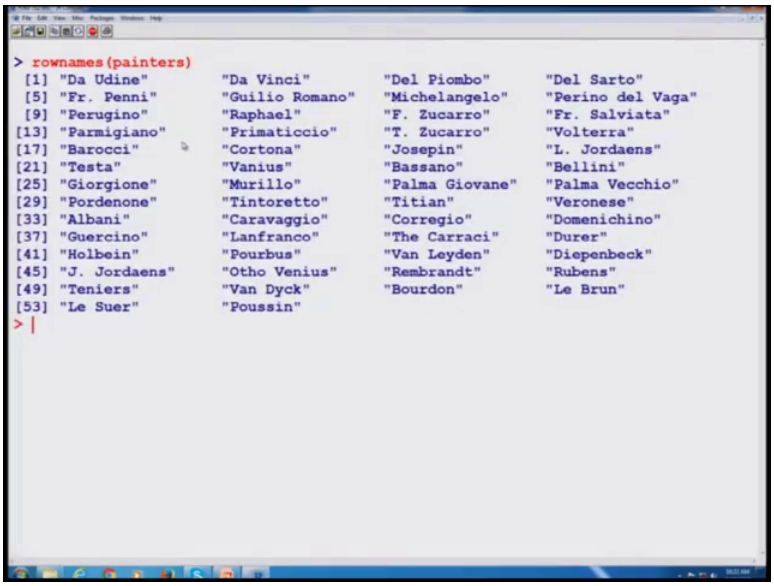
(Refer Slide Time: 18:51)



Now, suppose I want to know the first task or the first question is this that what are the row names or what are the names given in the rows right. Now you can see this data set can be a very huge and with some with say eyes you cannot see the entire data set in a single shot. So, I would like to extract the information. So, if you try to recall in the data set what was the row name the first column that was containing the name of different painters.

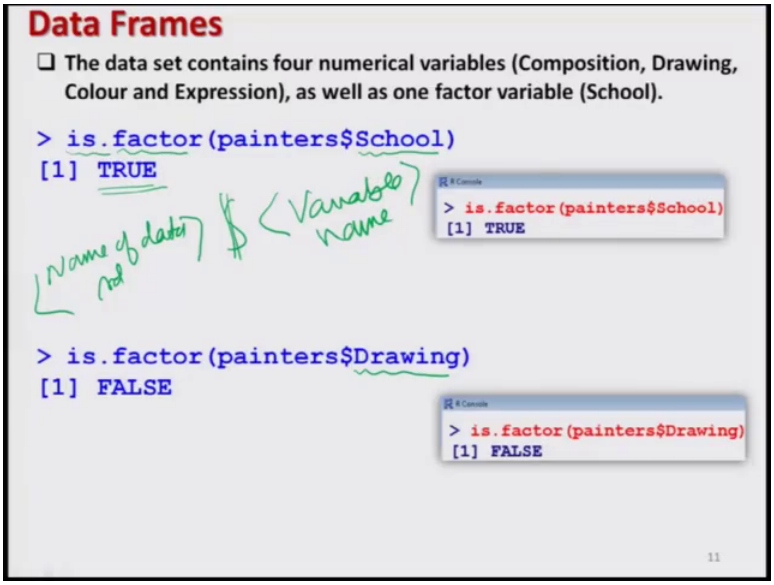
So, the value on the first row is the name of the first painter, the value on the second row is the name of second painter the value of the third row is the name of the third painter and so on. So, suppose I want to see what are the different row names in order to see the row names we have a command here R o w n a m e s simple row names. And inside the argument you try to write down this data sets name that is painters. And you can see here you get this type of thing. And let us try to do it means over the R console itself you can see here this is the first row name, this is the second row name. And now I am trying to extract this information from here right.

(Refer Slide Time: 20:16)



So, I come here, and I try to find out what are the row names and you can see here. So, it is trying to give us the same information which is contained in the rows right. Now let us try to come back to a slide. Now here I have given the screenshot of the same outcome. So, you can try it yourself also and the output will look like this.

(Refer Slide Time: 20:42)



Now, I will take another say options or say another type of task which can be done over a data frame (Refer Time: 20:52). We can see here that in this data set we have seen that there are 4 numerical variables or 4 variables which are taking the values in numerical.

These variables are composition drawing color and say expression. And there is one factor variable also that was the school which is taking the values as a b c d e f g h right. Now when there is a huge data set you cannot see it yourself so, I would like to know that; what is the nature of a variable. So, the first task you have to now understand is how to extract the information on particular variable from the data frame, the data frame has more than one variables and you want to have information only on one particular variable.

o, suppose I start with the variable say school and I do not know whether the information on school is of numeric type or is a factor, I do not know. So, I am trying to is first extract the information from the data set painters on the variable say school. So, please try to see how I am doing it in order to define the variable school from the data set painters first I try to write down here the name of data set. This is here painters and then I write here a sign dollar sign. And then I try to write down here the name of the variable. And now in case if I try to do like this painter's dollar school that will fetch the information on the variable school contained in the data set called as painters. Now I want to know whether this variable school is a numeric variable or not.

So, we have learnt in the earlier lectures, that how to judge this thing I can use here the function is dot numeric and then inside the argument I can give the variables name. So, the variable name here is my here painter's dollar school. And please be careful there is a difference between a small letters and say capital letter, if you try to write down say school all in small letters then this is different from this school where the first letter is the uppercase alphabet.

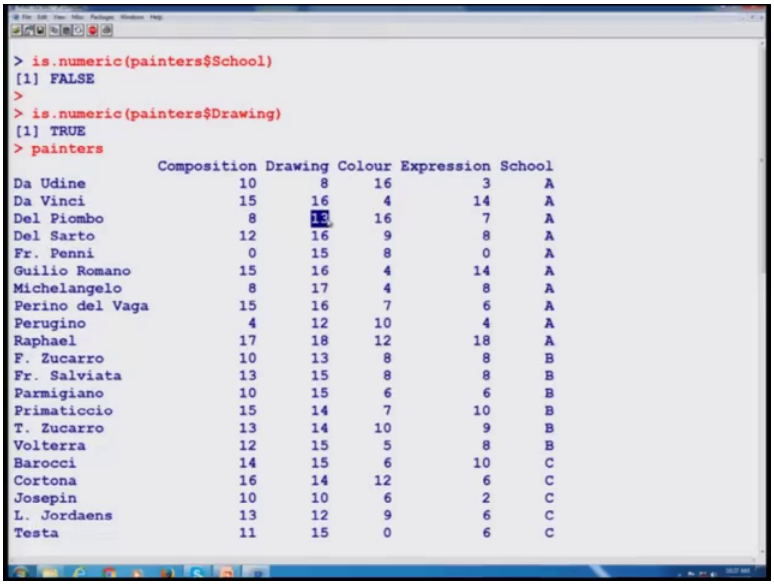
So, be careful with lower case and upper case when you are trying to call our variable all right. Now if you try to see here the answer comes out to be here false. And that is correct also you have got the data on a school essay ABCD. So, this is not a numeric and here is their screenshot of the same thing right and suppose I want to know what about say another variable drawing, what it is nature. So, again my objective is to extract the variable drawing from the data set painter.

So, I need two informations: one is the name of the data set and second is the name of the variable. And again I try to follow the same rules that first I write the name of the data set. Then separated by dollar sign and then the name of the variable which is

occurring inside the data set and suppose I want to know is it also a numeric you can see that the observation on drawings are numbers.

So, this is obviously a numeric variable and the answer comes out to be here true and these is the screenshot of this thing, but let us try to do it on the R console and see what really happens.

(Refer Slide Time: 25:15)



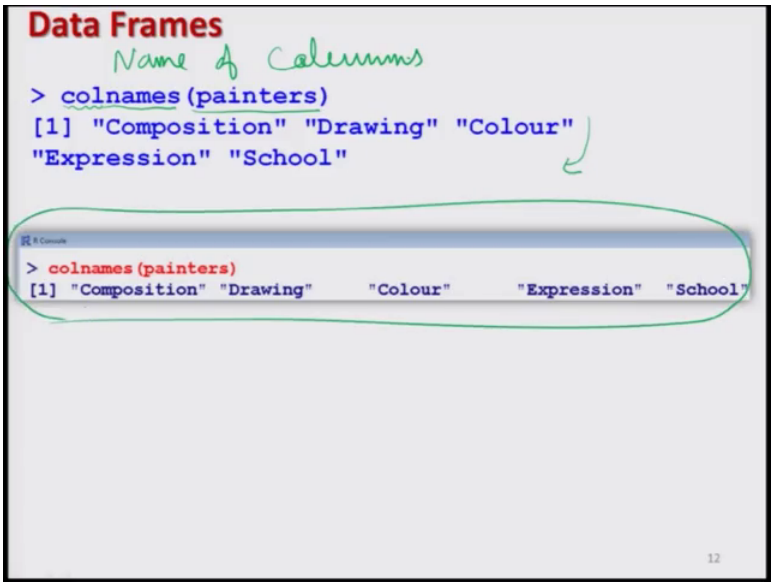
So, you can see here this is coming out to be here false. And similarly say this is another here drawing this is coming out to be here something here true, right. And you can see here once again that what about the variables is school and drawing, you see here school here is alphabets this is not numeric and on the observation on say here drawing they are here 8 16 and so on these are some numeric values right.

So, the same information has been obtained by these commands. And similarly in case if you want to go for say other option for example, I would like to check whether the school variable contained in the painters is it a factor. So, I can use the command here is dot factor and inside the bracket inside the arguments I use the same rule, name of the data set separated by dollar and the name of the variable that is what you have to remember right. I can write down here say again name of data set then followed by a dollar sign and then followed by the variable name. That is what you have to always keep in mind to fetch a particular variable. And the answer comes out to be here true yes the school is a b c d e f. So, that is a factor and now we also have seen that another

variable here drawing this is getting the values in terms of some number. Now I want to check here whether the variable drawing is a factor variable or not.

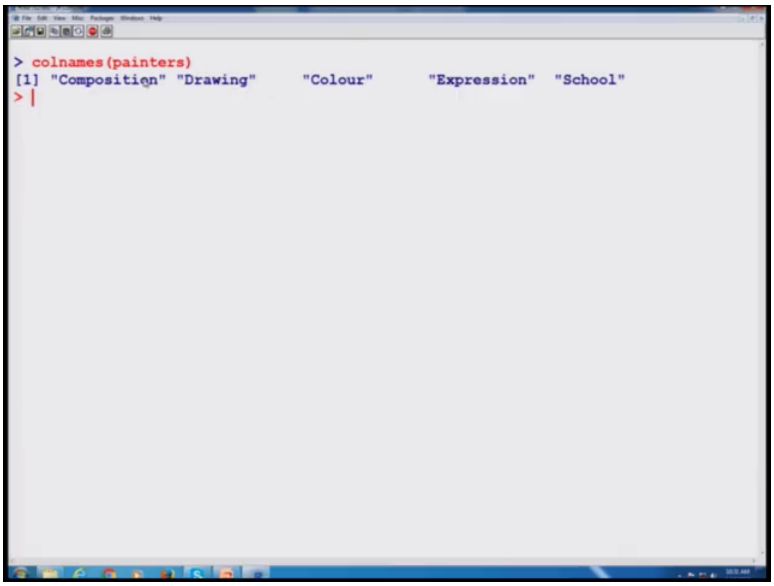
So, again I try to write down the name of the data painters then dollar and then variable name drawing and I try to see here is dot factor, answer comes out to be here false. So, why I am trying to take different example, I am simply trying to show you that if you have a huge data set then without looking into the data set you can extract all this information just by using different types of commands. Now means earlier I had taken the names of the rules means I try to extract the name of the rows from this data set. Now suppose I want to extract the name of the columns. So, how to get it done?

(Refer Slide Time: 27:44)



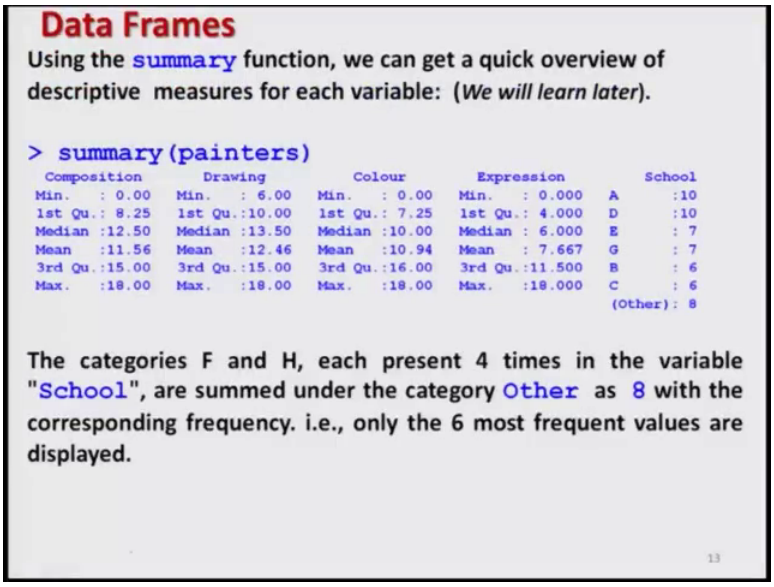
So, now my objective here is this I want to extract the name of columns for; that means, again the rule is very simple simply write c o l which is indicating column. And the names call names and inside the argument try to write down the name of the data set which is here painters and you will get here this type of outcome you can see here this is the outcome that will come in the R console, but let us try to see here that whether this is working or not. First you can see here that in this data set here painters, these are my columns composition drawing color expression and school.

(Refer Slide Time: 28:32)



And now when I want to have this column names, I do not know the data set I have not seen the data set it is giving us the same answer. So, this is the way I can extract the information on the names of the column.

(Refer Slide Time: 28:46)



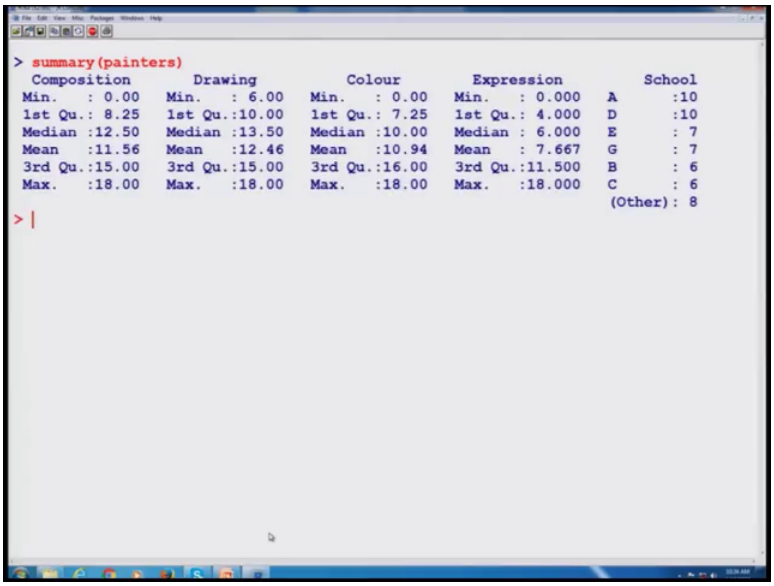
So, we again come back to our slides and I would try to show you some more applications on this data set. Well I am going to use here a function which is called here a summary function. Well, we have not done this summary function up to now, but we will take it in the forthcoming lectures, but here why objective is simply to show you that

different types of functions which we are going to learn for the numerical manipulation and say statistical computations in the forthcoming lectures. They will also be working on the data frame type of objects actually this summary function is a very simple function. And it gives us a quick overview of the descriptive measures of a variable for example, mean median mode quartiles minimum value maximum value, and so on; I believe that you all are familiar with these simple concepts.

So, for example, if I want to know the summary statistics then I can use the function here say here summary. And then I can use here the name of the data set inside the arguments and then I will get here this type of outcomes all right for example, if you try to understand it this is trying to take all the variables one by one composition, then drawing color expression and then school. And then in this first column it is trying to choose; what is the minimum value of the values in composition first quartile, median, mean, third quartile and maximum value. And this information is given over here for all the variables and similarly here there is something here like you see here other, what is this other is going if you remember there are categories a b c d e f g h right, but here it is not showing the categories f and h under the school and it is simply trying to show that this f has 4 values and h has 4 values which are taken together they are your here 8 values right.

So, this is simply trying to count or trying to measure the frequency of those variables right. So, now, the quickest is here the screenshot, but we also try to do it over the R console and we try to see what happens. So, you can see here we get the same outcome.

(Refer Slide Time: 31:10)



So, now we stop here in this lecture and I have tried my best to give you an overview of for say another R object, which is called as data frame and with this data frame. There are different types of functions commands and tasks which can be executed. So, instead of taking those commands function in theory, I have taken an example from a data frame that is the data set painters contained in the library MASS. And I am trying to take different type of tasks and I am trying to illustrate how they can be executed with this data set.

So, we will also continue with the same strategy in the next lecture. And we will again be using the same data set painters from the library MASS. And we will try to learn to execute some other types of this task. So you practice it, and I will see you in the next lecture, till then goodbye.