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

Lecture - 41
Data Frames

Hello friends, welcome to the course Foundations of R Software. Now, from this lecture we are going to begin with a new topic, this is about Data Frame. So, one natural question that crops up is that, what is this data frame yeah ok. For a while we forget this question and let me ask you a very simple question, have you ever worked in a spreadsheet? Now, you will ask me what is a spreadsheet. For example: I believe that most of us have worked in a software Microsoft Excel.

And usually, we try to create a file in common language that is called as an excel file, right. If you try to recall how this excel file looks like, if you try to see there are some values and these values are arranged in rows and columns and you can see there that the values need not always be numerical values, they can be numerical values, character, names, string, etc.

And if you try to see with this excel file you can do many types of operations, you can do mathematical operations on those columns where you can do some such operations. You can sort, you can order, you can do different types of queries.

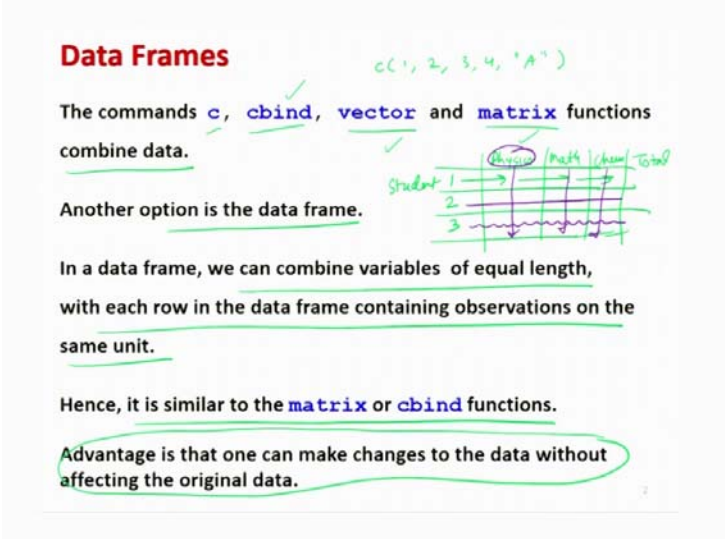
For example, if I say suppose I have some students and then I have entered their marks like physics, chemistry, maths and then their total, then if you want to know for example, that how many students are there who have got say more than 70 percent marks in mathematics and say less than 30 percent marks in chemistry.

This type of operation these types of queries, these types of questions you can generate in this file and you can get an answer, right. So, whatever we are trying to do in the spreadsheet which is the general name, generic name or in a simple words an excel file, the same operations can be done in the R software also, but on what. So, whatever is the spreadsheet or an excel file, that is the same thing which is called as data frame in the R software.

So, data frame is a format in which we try to arrange the data in the rows and columns on which we can do similar type of operations what we try to do in the spreadsheet or an excel file, right. So, first let us try to understand these concepts and then I will try to show you different operations and for this I will try to continue with this topic of data frame today in the lecture and then in the next couple of lectures.

So, I will try to take up these topics one by one and I will request you that whatever type of data file in MS Excel or any other equivalent software you have handled, try to look at it and try to recall that what type of operations you were doing over there. And here in this lecture and in the forthcoming lectures I will try to take up some common operations what you usually do. So, let us begin our lecture and try to understand what is this data frame.

(Refer Slide Time: 03:51)



So now, the first question comes here that whenever you are trying to handle with the data sets, then there are different type of data sets and you are interested in combining the data. So, in order to combine the data, you already have learned couple of topics like a c, cbind, vector matrix etc. If you try to recall you use this concept of data vector, where you would like to write the values as 1, 2, 3, 4 and then possibly here A and so on. So, this is also going to combine that with data.

Similarly, you had cbind, rbind in which you used to combine the values similarly you had vectors matrix etc. So, in all of them there was always some conditions; for example,

if you want to do the matrix manipulation then inside the matrix all the values have to be numerical values. So, similar to these different functions which are used for combining the data, another option is data frame, right. So, in this data frame what we try to do? We try to combine the variables of equal length, with each row in the data frame containing observation on the same unit.

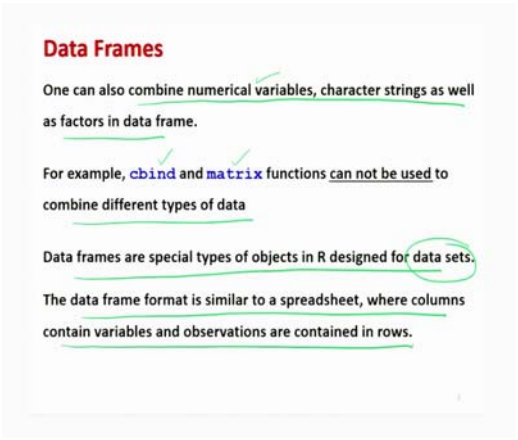
So, what does this mean? If you try to see here suppose I try to write down here the names of three subjects here, say physics, see here mathematics and here chemistry and then after that we try to find out the total of these marks.

So now, you have here student number 1, student number 2, student number 3 and you try to write down the marks in physics, mathematics and chemistry for the student number 1 here, then you try to take the marks of student number 2 in physics, maths and chemistry here. And for student number 3, you try to write down the marks in physics, maths and chemistry here and then we try to find out their total.

So, this is what I am trying to say that number 1 the data is arranged in the rows and columns and every variable which is here physics that will have same number of observation in the rows. Because if you have say three students so there are going to be 3 marks in physics, 3 marks in mathematics and 3 marks in chemistry, right. So, this is equivalent or this is similar to the matrix or cbind function, but the only thing is this in the matrix if you want to do mathematical operation all the numerical values have to be of the same type, right, say numerical, right.

But in data frame you can have different types of variable, which can be numeric as well as character, I will try to show you, right. And one big advantage of using the concept of data frame is that, without making the change in the data file you can make the changes in the data.

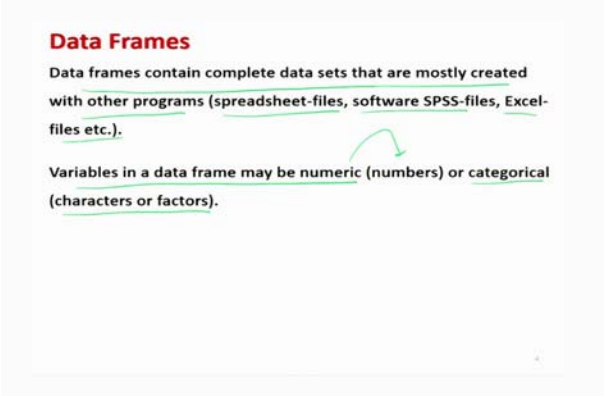
(Refer Slide Time: 06:29)



So, you will not change your original data file and in case if you want to use only a part of the data you can create such a data set without affecting the original data file, right. And another very big advantage is that, in case of data frame you can combine various types of objects, like as numerical values, character string as well as factors under the setup of keta frame.

For example: if you try to see the commands which you have used earlier, you have learnt earlier cbind and matrix they cannot be used to combine different types of data values, right. So, this data frame is also a special type of R object which is especially designed to handle the data sets, right. So, data from format is similar to spreadsheet where the columns contain variable and observations are contained in rows like that, right.

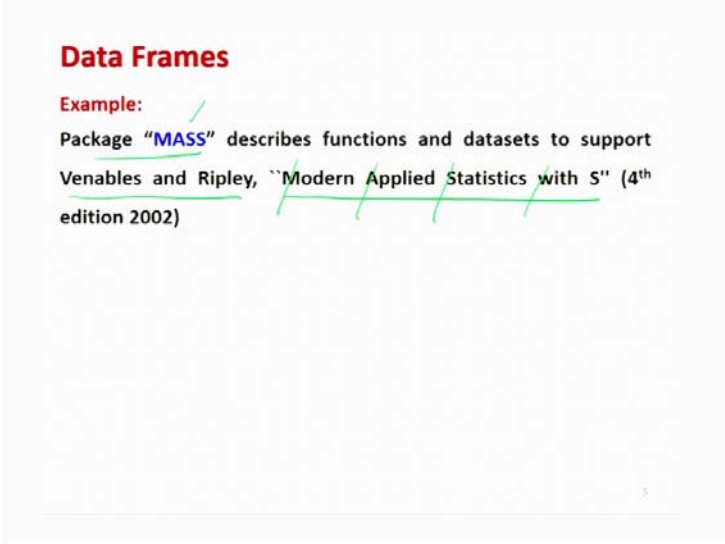
(Refer Slide Time: 07:05)



And data frame contain, they contain complete data that are mostly created with other programs like as a spreadsheet-file, software like SPSS-files, Excel-files etc.

And the variables in the data frame they can be numeric or they can be categorical which have the values. For example, numeric variables have the values in numbers whereas, the categorical variable has the values and characters or say factors and so on.

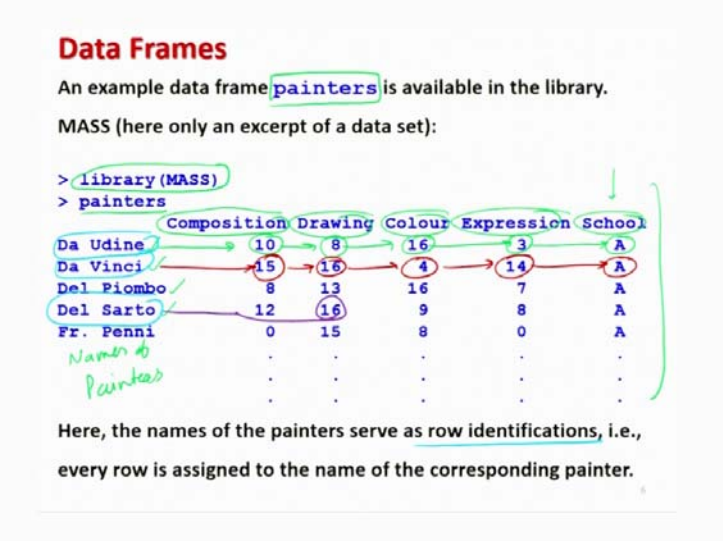
(Refer Slide Time: 07:29)



So, now before we move forward let me try to request you that you please upload the library MA double S in your software, that is inbuilt actually. So, we are going to use this package here MASS, MA double S, right.

So, if you remember in the beginning we had talked about a built in package MASS and this MASS was the, the name MASS is coming from the name of the book "Modern Applied Statistics with S". So, its MASS, actually that was the book which was written to explain S plus software and that was written by Professors Venables and Ripley. So, whatever the data sets were used in that books, they are combined in this package here MASS, right.

(Refer Slide Time: 08:09)



So, now I will try to tell you here many operation through examples which I personally believe that it is easier to learn, right. So, in this MASS package there is a data frame whose name is painters; painters is all in lower case alphabet, this is available in the library.

There are many data set we which are available in this package and actually it contains all the data sets which have been used in the book, right. And then yeah and the second thing what you have to keep in mind, that here I am trying to give you some idea that how the data is going to look like, sometime the data is big.

So, it may not be possible for me to give you here the entire screenshot or the entire detail or sometime even the screen is also going to the couple of screens actually. So, whatever I am doing here, you please try to keep on doing on your computer also for a better understanding, right. So, what I do here that first of all I try to load the package here MASS, right and then in order to access this data frame painters, I simply type here painters on the R console and you will get here this type of data set.

So, actually I can just briefly explain you before moving forward that what is this. So, this data set is containing the information about some painters. And for example, this Da Udine, Da Vinci, Del Piombo, Del Sarto etc. they are the names of different painters, right. And their paintings were analyzed by some experts and then they have tried to give

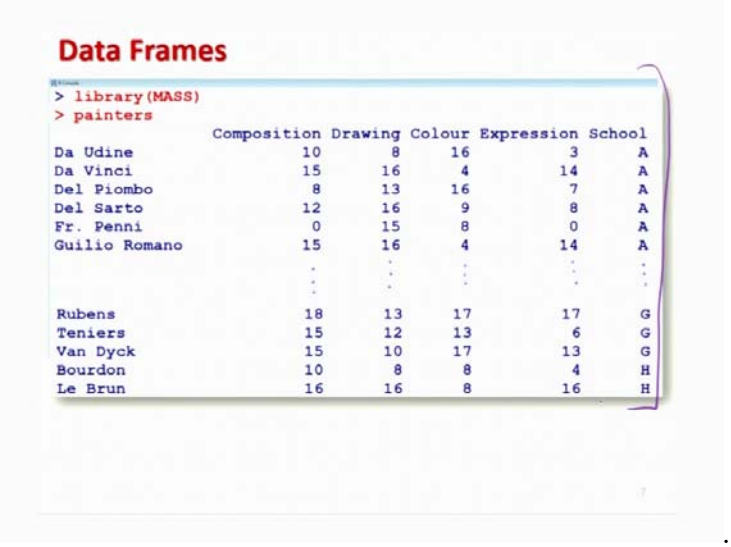
some values to those things and those variables on which the data has been collected on their work is like the composition, drawing, color, expression, school, etc.

For example: their school is something like there are different schools under which people try to learn this art, like as painting or music etc. So, this is that school, right. So, if you try to see here this Da Udine has a value of 10 units in the composition, then 8 value for the drawing, then 16 for the color, 3 for the expressions and the painter is coming from a school A.

And similarly, if you try to see about Da Vinci, the value of the variable composition is 15, drawing is 16, color is 4, expression is 14 and the school here is A. Yeah, I would like to very honestly accept that yeah, I do not know this much information about the drawing and they say paintings etc. So, kindly excuse me if I try to misinterpret, I will try my best not to do it, but yeah please do not expect that I know each and everything about the painting. So, that is my very honest confession, right.

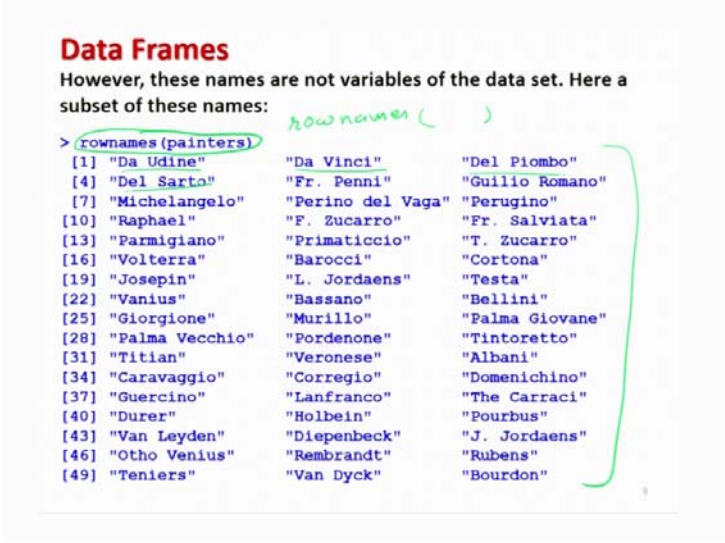
So, you can see here that in this case whatever are the names here, this one this one etc. they are working as a row identification. Means if somebody ask me, what is the value of the drawing for the painter Del Sarto. So, I will go here and I will try to find out here this value, right. So, in this case you can see here that every row has got this type of information, right.

(Refer Slide Time: 11:19)



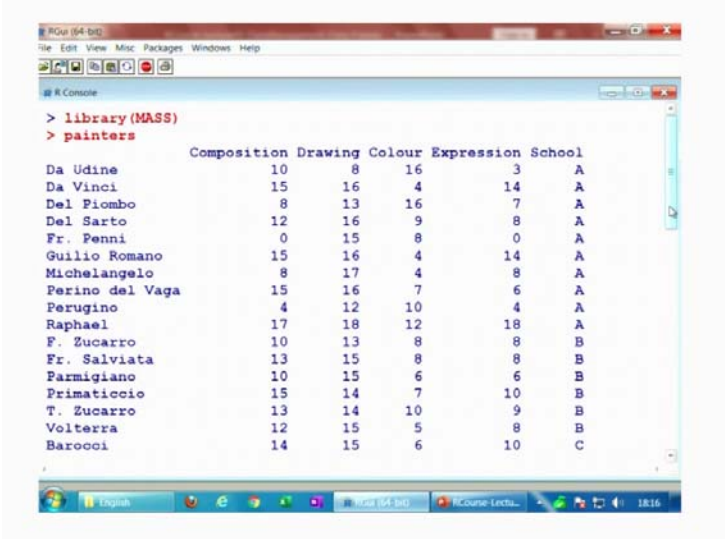
So, now you can see here if you try to see on the R console actually, this will look like this.

(Refer Slide Time: 11:25)



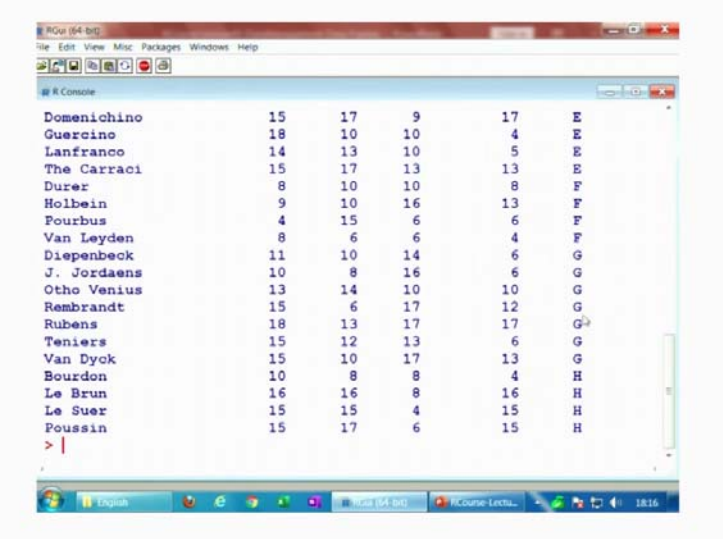
So, before moving forward let me try to show you these things on the R console.

(Refer Slide Time: 11:29)



So, if I try to see here, first I try to make a library mask and it is here like this, right. And if I try to show you here the painters it is here like this, right. One thing you have to keep in mind, that I have decreased the phone size so that I can show you most of the operation on a single screen, right.

(Refer Slide Time: 11:53)

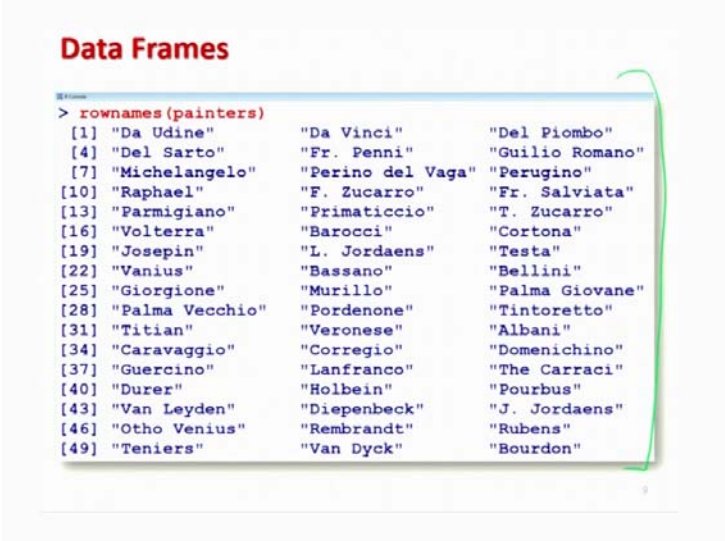


So, what I will request you that as I am moving forward you also try to do the same operations on your computer, right.

Now, I try to take up here different types of operation what you can do in a usual spreadsheet. So, it is going to be like that, that you are simply getting the file from some external sources and you do not know what are the contents of the file. So, you would like to find a different type of information by using the R functions from this file or from this data frame. For example, in case if you want to find what are the row names in the data frame painter, right, you have seen that in the names of the painters are given and they are the identification marks.

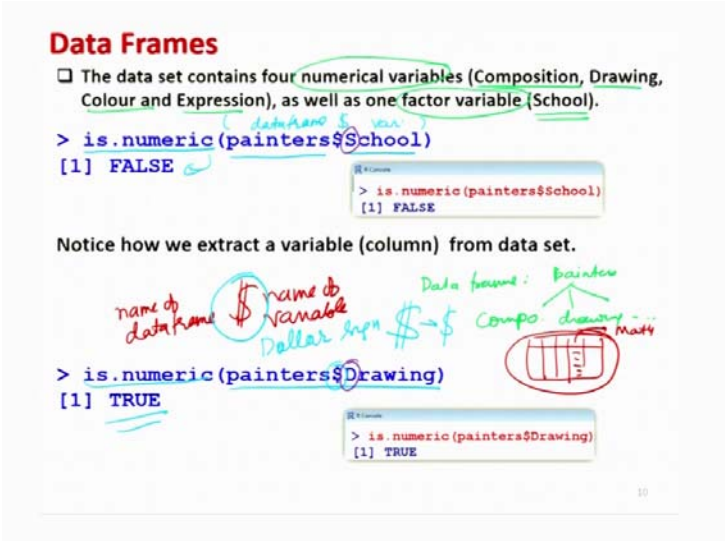
So, for that you have a command here row names, row like as names and then within the parenthesis you try to write down the name of the data frame. So, you can see here you will get here this type of outcome. So, the first name is Da Udine, Da Vinci, Del Piambo then Del Sarto etc. So, there are so many names and that is why if you try to see here, I have given here that it is continued and that is what I meant that when said that it is not possible for me to show you all the values here, right anyway.

(Refer Slide Time: 13:01)



So, this is the screenshot of the same operation. So, now it is clear that if you want to find out the names of the row in any data frame, you simply have to use the command rownames ok.

(Refer Slide Time: 13:13)



Now, the first question comes here, there are different types of variables here, right, you can see here actually this composition its some there are some numbers whereas, in the school there are some alphabet. But; obviously, you have to just imagine that this file is not for you means that is inside some storage and you cannot just open it and can

determine that which of the variables are going to be or what is the structure of the file, what is the behavior of the file, what is the type of the variable, etc.

So, suppose I want to know that what is the type of the variable. For example, you can see here, that here as I shown you here these variables here composition, drawing, color, expression you have here the numerical data, but for the school you have got here data like A, B, C D etc. So, there are here four variables say composition, drawing, color and expression, which have got some numerical values and there is here one variable here school, which is F factor variable, right. But as such I do not know.

So, I would like to first understand how to get it done. So, now, I am going to tell you two things, one thing you already know and the second thing is I will be explaining you in detail in the forthcoming lecture, but here I am going to use it. Now, if you try to see here, you have here a data frame, whose name is painters and this has got several variables, like as composition, then drawing, etc. So, now, if you want to access only the data of a particular variable, then how to get it done.

So, there are different ways, but one simple option is that you simply try to write down the name of data frame and then put here a sign and then write down here the name of variable. So, what will happen? That the data corresponding to this particular variable in this data frame that will be attached and then you can do different types of operations over that data value. It is something like this for example, a spreadsheet where you have this type of column and suppose this column contains the marks in maths.

And you have some the marks here and you want to access only the marks in the mathematics. So, how to get it done?

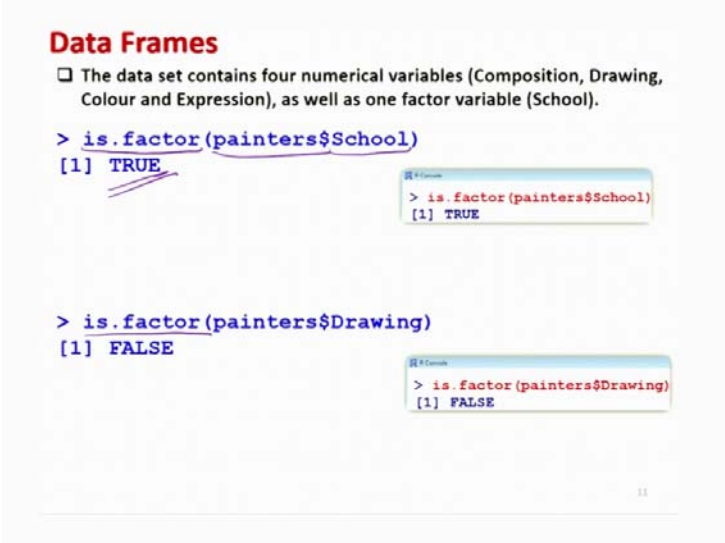
So, for that you try to write down the name of the data frame in which all these values are stored and then try to write down the name of this variable and both of these, the name of the data frame and the name of the variable they are joined together with the dollar sign, right, this is the do double lar, dollar sign. It is like this S and then here two vertical signs this is available on your keyboard, right and then keyboard it is written like here like this.

So, suppose I want to see whether the variable whose name is school in this data frame painters is it a numeric? So, I try to write down here our command is dot numeric and

within the parenthesis, I try to write down the name of the data frame and then here the variable name and then I try to attach it with the dollar sign. So, you can see here this comes out to be here for, right. And similarly if you want to know whether the data on the drawing is numeric or not. So, you try to write down here painters, color, drawing and then is dot numeric and it will come out to be true yeah.

Very important point what you have to keep in mind that whatever is the exact name of the variable, that you have to give here. For example, you can see here that in the variable here is school, this capital S and in the word here drawing this capital D, they are in the capital letters, right. So, that is what you have to keep in mind and similarly if you try to take here the variable here is school. Now, because you have seen here that this school is coming out to be non numeric. So, now, you would like to check is it a factor.

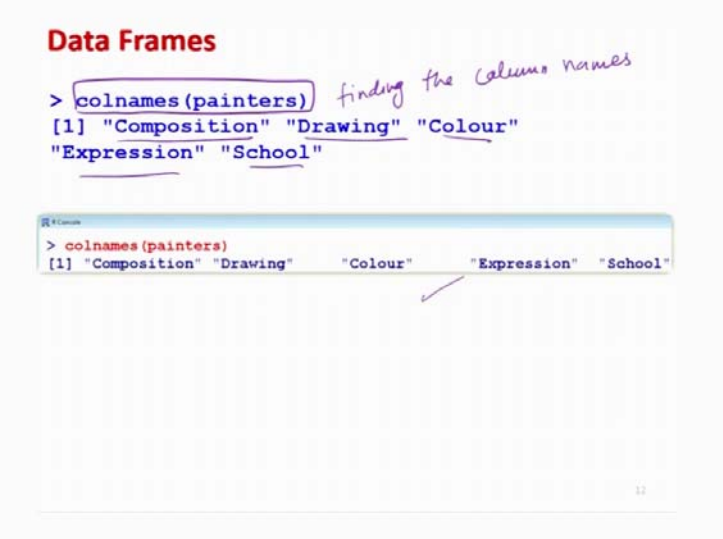
(Refer Slide Time: 16:59)



So, if you try to write down is dot factor and within parentheses you write painters dollar school it will come out to be true.

And similarly, if you try to see here you have seen here is numeric painters drawing, this is here true, but now if you want to know is dot factor this painters dollar drawing. That means, you want to know whether the variable under drawing and it data values is it a factor answer comes out to be here false, right.

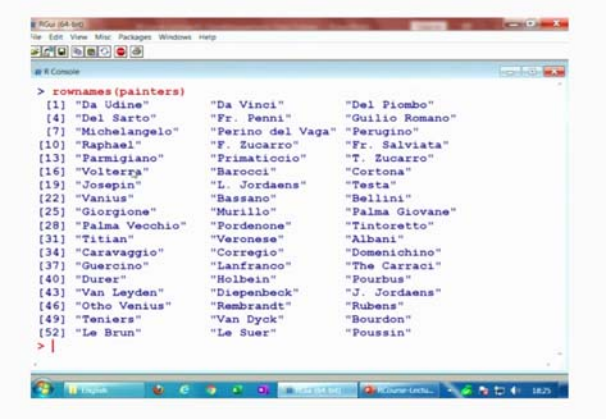
(Refer Slide Time: 17:25)



So, let me try to show you these things on the R console, but before that as you have learnt about the row names. So, similarly we have here one more function here for finding the column names, right.

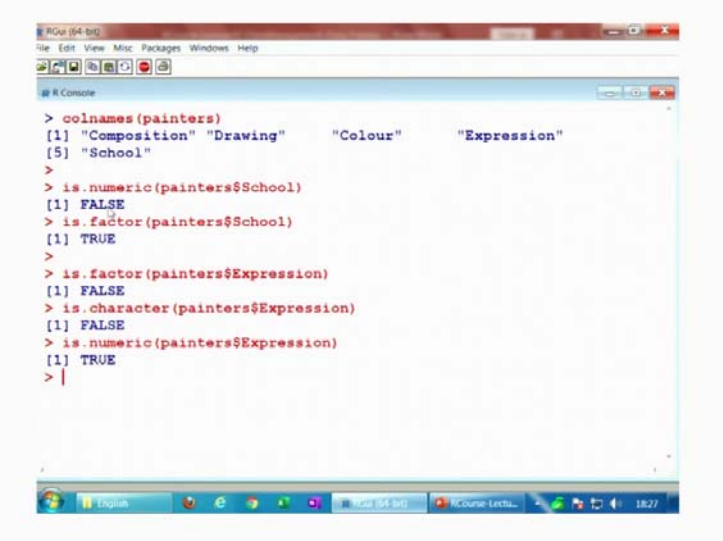
And if you try to operate it you see here this is here composition, drawing, color, expression, school, it comes out to be here like this. So, let me try to show you these operations on the R console and then we try to obtain. You can see here it is here like this. So, you can see here these are the names of the columns and these are the names of here row, right.

(Refer Slide Time: 17:59)



So, what I try to do here that, I simply try to find out the row names, row names of painters it is here like this you can see here, right.

(Refer Slide Time: 18:07)



And similarly, if I try to find out here the column name. So, I have to take here colnames all in lower case alphabets and you can see here it is here like this, right. Now, here I would like to explain you the simple way by which you will not make the mistake when you are trying to address a particular variable. Suppose I want to know is dot numeric now what? Suppose I want to know about this school. So, now, the question here is there is always a possibility that you may make a mistake in writing the school, beginning the lower case alphabet.

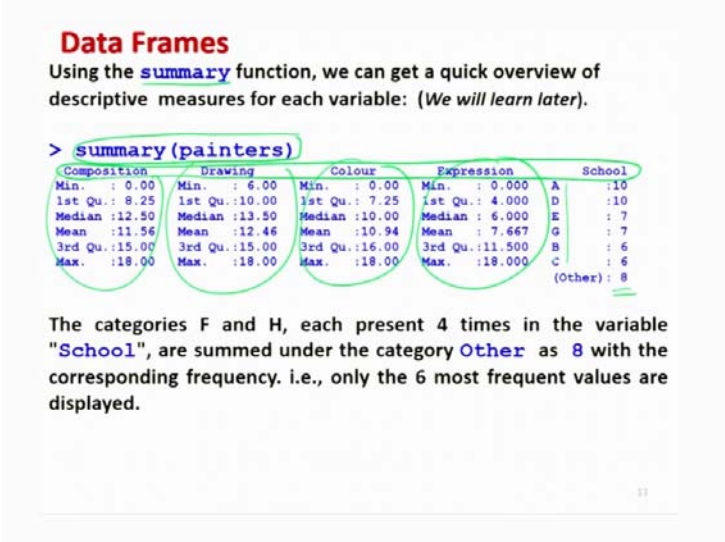
So, the best option is that which I suggest is that you try to copy this name of this data frame so that there is no mistake in making a mistake, then you put here the dollar sign and then whatever is the name of the variable exactly you just try to copy it, leaving the double quotes and you try to write down here. So, this will ensure that you are not making any mistake in typing or reproducing the name of the variable, right.

So, you can see here this comes out to be here false, right. And similarly, if you see here is a factor variable, you can see here it will come out to be here true, right. And similarly, if you try to look for this expression here the same thing. So, if I try to see here is this a factor variable, answer comes out to be here false, right. But if I try to see here is it

character false, then I would like to know is it numeric, is dot numeric and the same expression and this comes out to be here true.

So, similarly you can make here different types of such operations and which will give you more insight about the structure of the data frame, right. So, and yeah if you so now, you have seen here these operations.

(Refer Slide Time: 19:51)



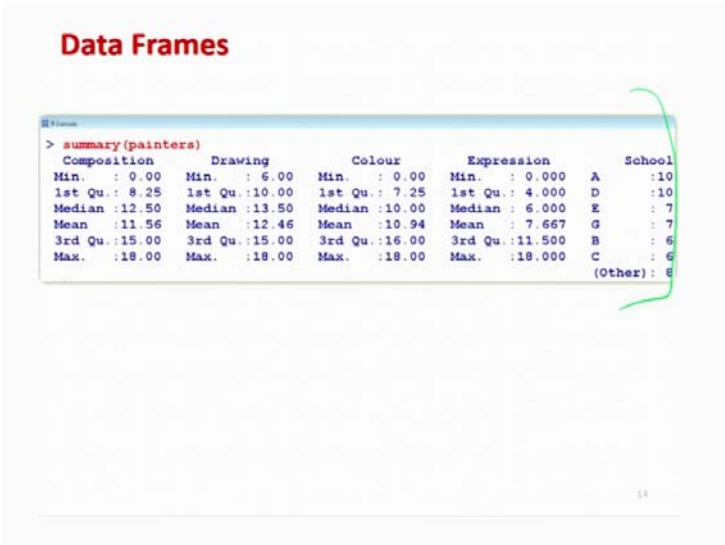
Now, I want to give you here just one example that how the data operations on this data frames is going to help us. For example, we are going to learn about this concept at a later stage and we have a here function summary s u m m a r y. So, this summary function gives us the information about or a quick overview about the minimum value, maximum value, 1st quartile, 2nd quartile, mean, 3rd quartile.

So, if you try to see here if you try to use here this summary painters s u m m r y and then within parenthesis you try to write down the name of the data frame.

Then all information about all the variables will be available here in a single command. For example, this is for here composition this is for here drawing, this is for here color, this is for here expression and so on, right. And yeah one thing you have to here that these are the categories here A, D, E, G, B, C and others are here actually A. So, there are more categories, but only so it is showing here as say other.

So, at this moment you do not have to think about these thing, but the main thing is I want to show you that, if you want to have the information or any statistical or mathematical operation over all the variables, yeah if that is applicable. Then you can use the concept of data frame and that will help you in giving you a quick answer, right.

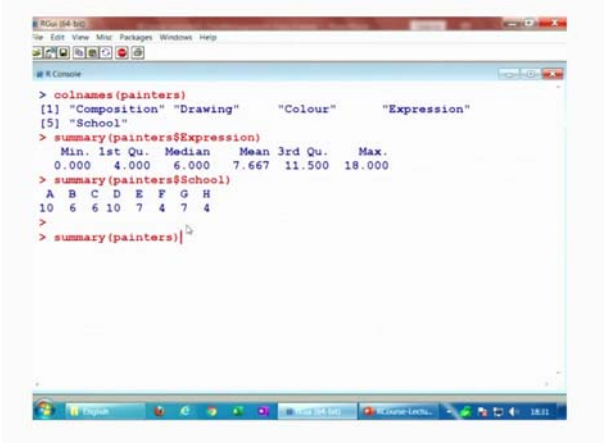
(Refer Slide Time: 21:15)



So, this is here the screenshot of the same thing and yeah if you want to do it for the individual variables also you can also do it.

So, let me try to show you here it on the R console. So, you can see here that you have all these variables by the names here painters dot expression etc.

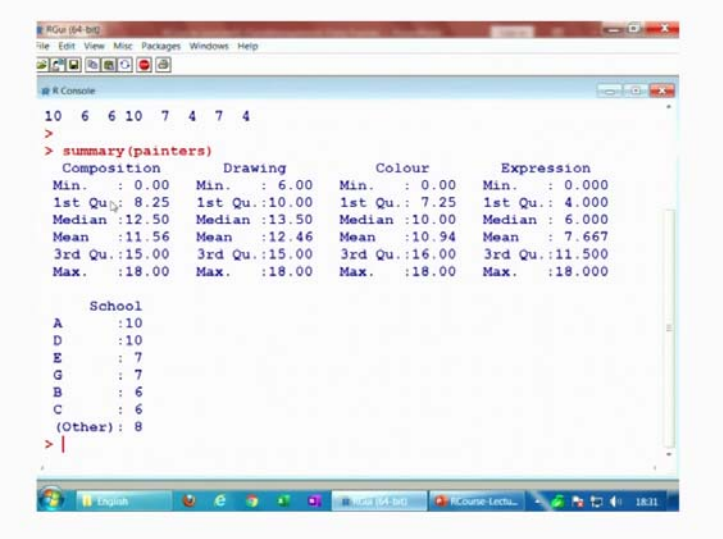
(Refer Slide Time: 21:35)



So, if I try to see here what are your here columns of here painters, right. So, that will help us. So, now, if you want to make the this summary command for this expression. So, I have to write down here painters, dollar, expression and if I say here enter it will give me the first value.

Similarly, if I want to use it for here see here school. So, one by one I can do it like this, but in case if I try to do it over the whole data frame and I simply write down here summary and then within parenthesis the name of the data frame, you can see here that it will give you all the values like this here, right.

(Refer Slide Time: 22:11)



So, this is what I meant ok. So, I think let me stop in this lecture and my idea in this lecture was simply to introduce you with the concept of data frame. So, that you can settle it in your mind and try to do this very small operation. And in the next lecture I will try to give you some more commands here which are going to really help you in making different types of operation in the R software.

So, you try to have a look on the concept of data frame and try to practice with this very simple command that how can you be comfortable when we are trying to meet in the next lecture. And believe me once you are comfortable in the next lecture, I am going to repeat this command I am going to use this command. So, you try to practice and I will see in the next lecture. Till then goodbye.