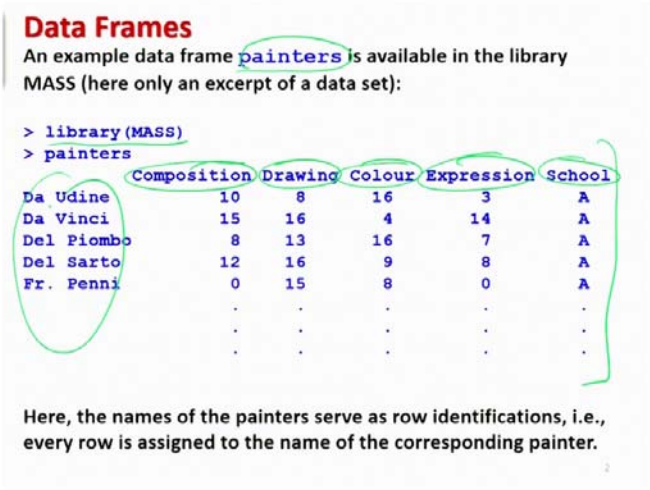
Foundations of R Software
Prof. Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology, Kanpur

Lecture - 43
Data Frames: Some More Operation

Hello friend, welcome to the course Foundations of R Software. And, you can recall that in the last two last couple of lectures, we had talked about the Data frames, right. So, now, in this lecture also we are going to continue with the same topic. And as we had learnt some basic operations in the data frame to extract different types of information in the last lecture.

So, similarly in this lecture also, we will try to learn some more concepts some more commands to do the common operations in the data frame. And for that we are going to once again use the package mass in which there is a data frame called painters. So, now, in this lecture, we are going to learn about some more operations. So, why not to take those examples and try to understand how to execute them? So, let us begin our lecture, right.

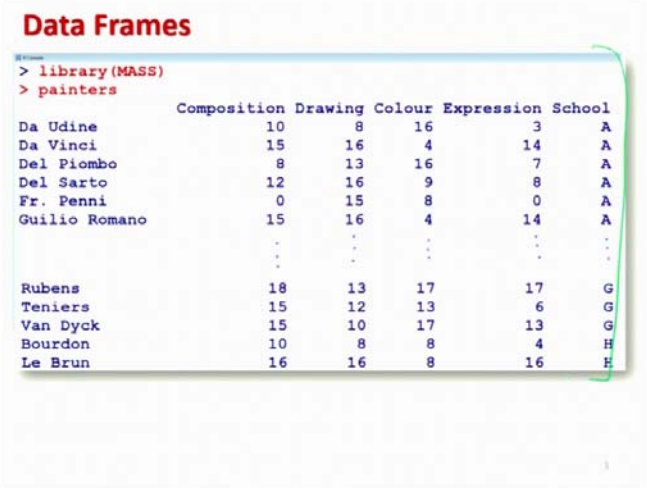
(Refer Slide Time: 01:13)



So, if you try to recall we were using the data set painters. So, for that you please upload your library by using the command library mass and then you can see this was the data

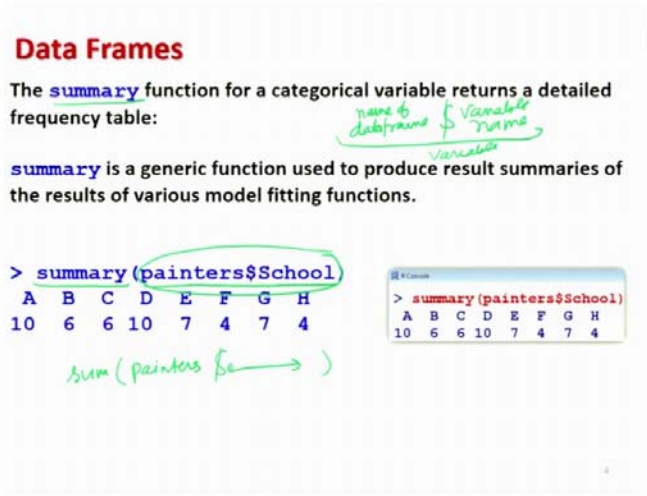
frame painters which has five variables; composition, drawing, color, expression, schools. And then in the rows there are names of different painters.

(Refer Slide Time: 01:37)



So, it looks like this, right. So, now you are familiar with this data set.

(Refer Slide Time: 01:41)



So, now one operation which we learnt in the last lecture was that if you want to extract a particular variable from a particular data frame then how to do it? So, for that, we had learnt that a you would try to use the command that you would use the name of the data

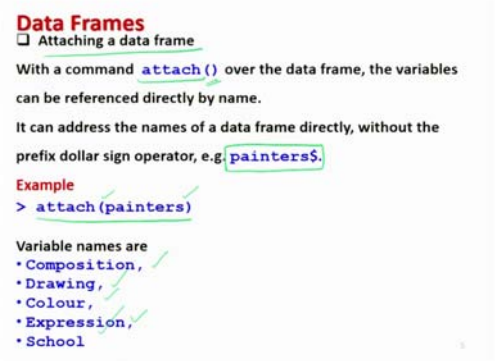
frame and then dollar command and then variable name. And then the advantage of using this was that now this works just like single variable and after that you can do any types of operation.

For example, if you want to find out the mean or sum or anything of those numerical values stored in a variable in a data frame then you can do it directly here, right. So, for example, we had used the command summary. And summary will give us like a mean first quartile, second quartile, third quartile, etc. for the quantitative variables. And then it will try to give you a sort of frequency distribution for the categorical variable right factor variable.

So, now, suppose if I try to take an example that how are you going to use this command over a variable then it is like this you will try to write down here the summary. And then you will try to write down here the painters dollar school, right. Now, similarly if you want to find out the sum of any value, so you will try to write down here the painters and then whatever is the name of the variable that you will try to write down here.

So, now, in this process if you try to see what essentially do you want you want to know the summary or the mean of a particular variable. And in order to access that variable you have to first write down the name of the data frame and then the name of the variable and both have to be joined by a dollar sign. So, basically in order to access a particular variable, first you have to write down the name of the data frame. So, now there is an alternative command in the R software to do such operations in which you do not need to write again and again the name of the data frame, right.

(Refer Slide Time: 04:11)

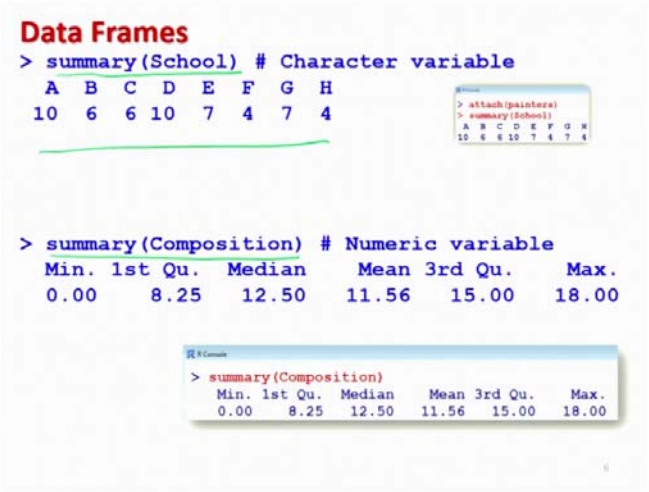


So, for that we try to use the concept of attach and we try to attach the data frame once you try to attach a data frame. Then all the variables whatever are contained in that data frame they can be accessed directly by their names, right. So, if you try to write down here the command like a t t a c h which is the function for attaching, then all in lower case here a double t a c h. And then inside the parentheses you try to write down the name of the data frame.

Then after that if you try to execute any command then you will not need to write down the name of the data frame followed by the dollar sign, right. So, for example, means once you try to attach the data frame here this painters just by writing a t t a c h all in lower case alphabets and the inside the parenthesis painters. Then in order to access any variable you need not to use this painters and dollar you can directly access them by their names for example, column names or the names of the variables in the painters were are composition drawing color expression school.

So, earlier, if you wanted to access the data on the variable school or expression you would like write this painters dollar school or painters dollar expression, right.

(Refer Slide Time: 05:45)

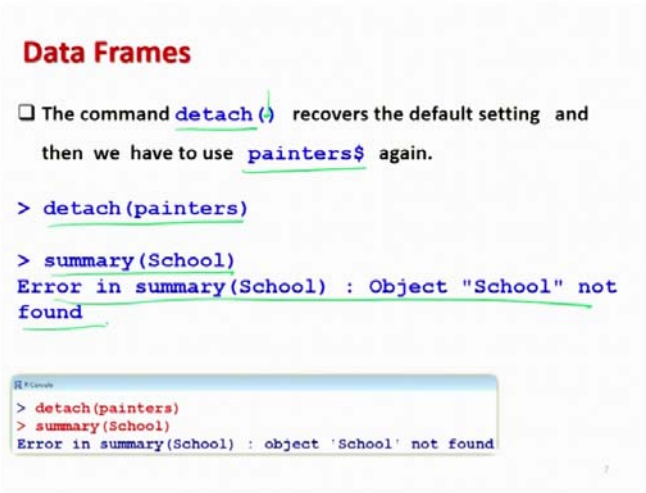


But after this, you can see here that without writing the painters dollar, you can if you want to find out the same command that summary of a school, you will get here this option. And similarly, if you try to write down here summary of the variable composition

then you need not to write down the painters dollar. So, actually when you are trying to work with a single database for a very long time then actually these type of commands are really useful, right.

So, you can see here although I am not saying that you cannot do to writing the name of the data frame and the dollar. So, and there is no difference between the two also, but it is just more convenient, right.

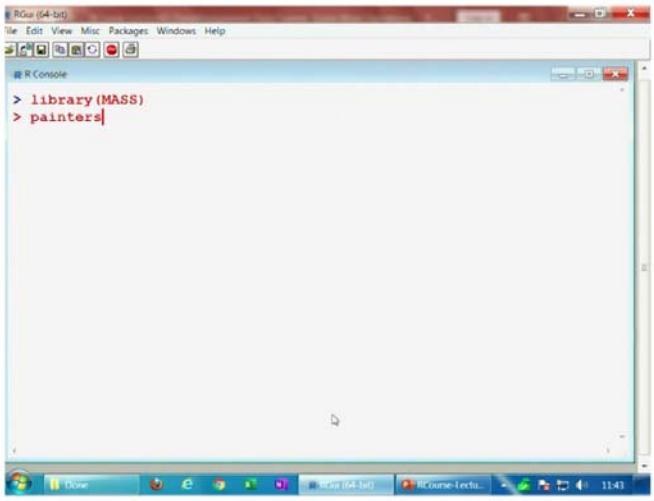
(Refer Slide Time: 06:31)



And once you have done the job means once you have work. So, as you have use the command attach then you also have to use the command detach d e t a c h, right. So, once you use here the command here d e t a c h and inside the parentheses you have to write down the name of the data frame. Then, it will go back to earlier and after that if you want to call a variable then you have to use once again like painters dollar and the variable name, right.

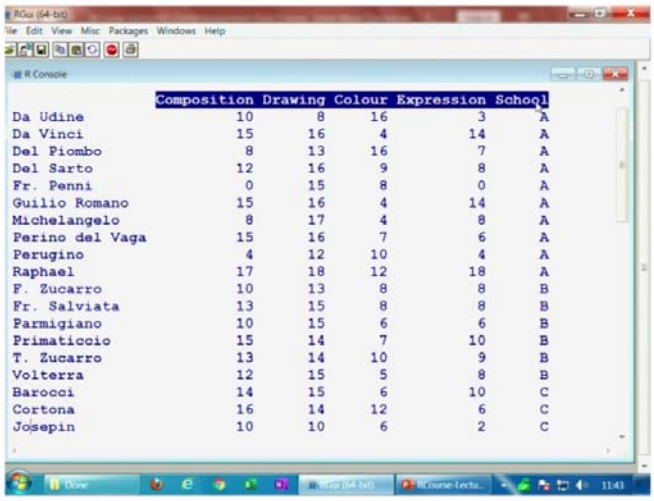
For example, if I try to show you here suppose I execute here detach painters and then I try to now give you here once again summary school it says that error in summary objective school not found whereas, just before that you are trying to find out this summary without any problem. So, this is the way by which we can write down these things. So, let me try to first show you this operation on the R console. So, that you become more confident and then I will try to show you here right, ok.

(Refer Slide Time: 07:41)



So, now I first try to upload here this package here MASS. So, this is here. Now, this and then you can see here this is your here data frame painter.

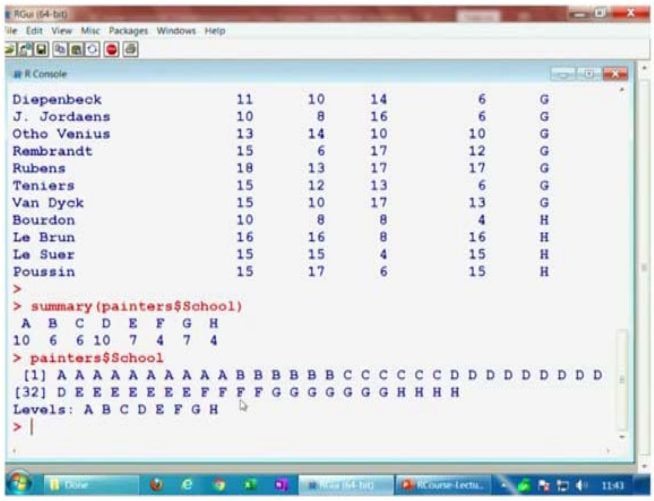
(Refer Slide Time: 07:51)



	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

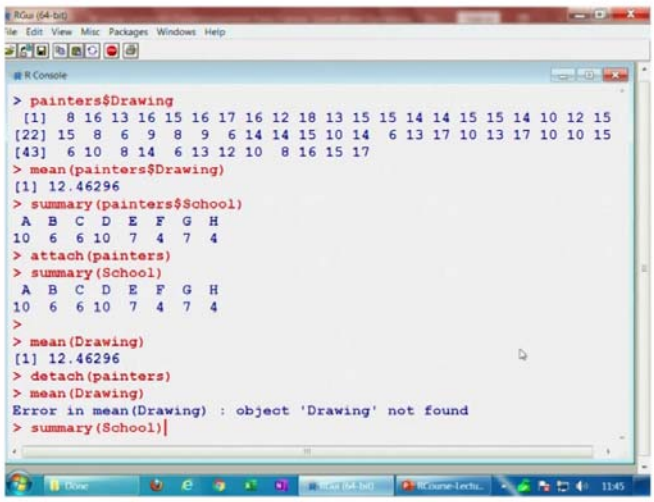
So, you have here different types of variable here composition drawing etc. and so on, right.

(Refer Slide Time: 08:01)



So, now in case if you try to execute here, this command here say here summary; summary of here the painters, painters and after that you try to here take here school. So, you can see here this we are come like this, right. And, if you try to see the data in this painters dot school here is like this, right. So, similarly let me try to take here one more example.

(Refer Slide Time: 08:35)



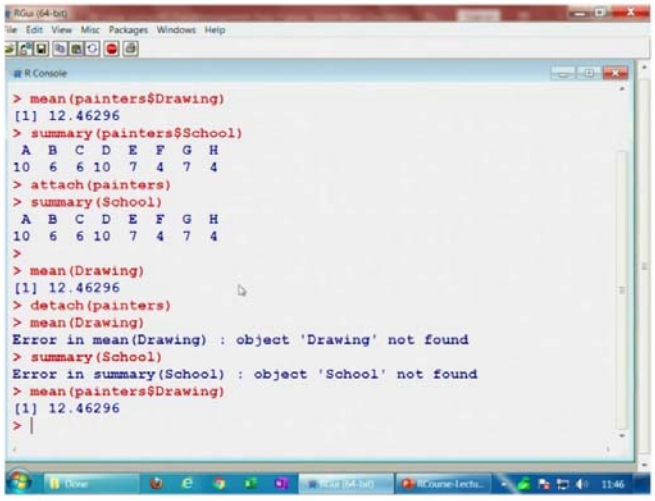
So, you can see here the data in the drawing is here like this, right. And, suppose I want to find out the arithmetic mean of this data. So, I simply try to write down here mean, but

then I have to write down here painters dollar and then drawing, right. So, it will give you here this value right now I try to use here the command attach right and you can see here what happens ok. Let me try to just, repeat this option here also so, that you can compare it very easily, right. So, now, I try to use here attach and then painters, right.

Now, I try to find out simply here the summary of this painters, dollar, school just by writing here summary and inside parenthesis I will write simply the name of the variable. We can see here you get here the same outcome this outcome is the same as this outcome and now in case if you want to find out here the mean of this drawing. So, I can write down here mean and simply here drawing, right. So, you can see here this is the same value which we had obtained here. So, that is the advantage right, ok.

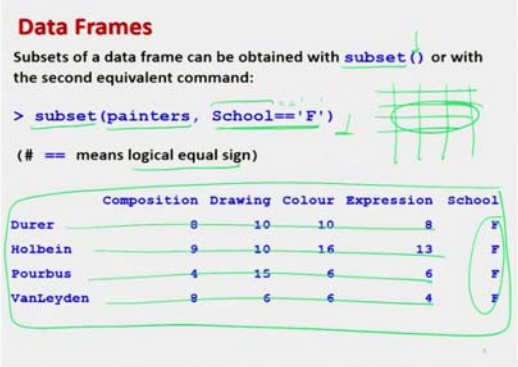
So, now let me try to give you idea about some more operation yeah, but let me to also try to show you the about detach. So, now, I try to use here detach painters. And now you will see that in the same screen, now if you try to find out here mean or drawing it says there is error. And if you want to find out the summary of a school it says error, right.

(Refer Slide Time: 10:19)



But, just before that it was working and if you try to find out here the mean by writing the painters dollar school or summary you can find out easily here. So, that is what is happening, right.

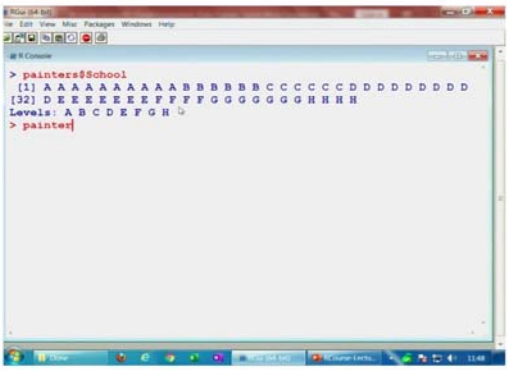
(Refer Slide Time: 10:43)



So, now, let me try to show you here one more operation that is about how will you try to extract a subset from the data frame, right. For example, if I try to say here I have here different datas arranged in rows and columns like this. And suppose you want to just select some particular data set out of this whole data frame, right. So, subset of the data frame can be obtained by using the command here subset s u b s e t, right. And, inside the parenthesis you have to write down the name of the data frame and then you have to write down the condition that what type of data frame do you want, right.

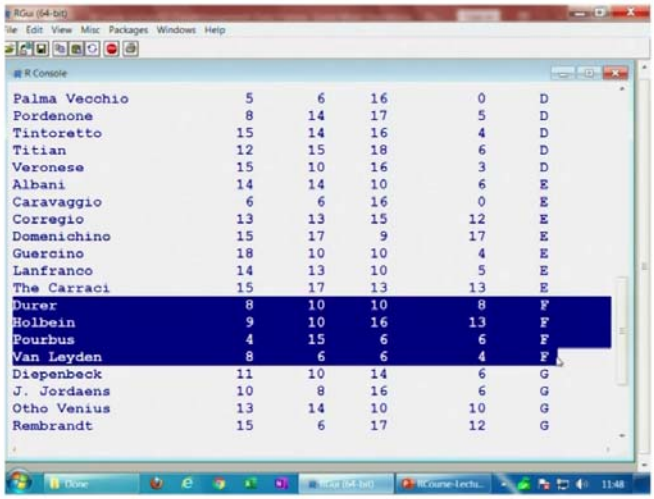
For example, you have seen that when you were trying to see the data on the school the school had different category like A, B, C, D, E, F etc., right. So, if you want to have the subset of those painters who are coming from the school F, then how to get it done? Ok. So, let me try to first show you that how this data will look like on the R software. So, that you are you can understand it better, right.

(Refer Slide Time: 12:05)



So, if I try to write down here this is here painters, dollar, say school, right. So, you can see here that there are here four such painters who are coming from the school whose name is or categorized by F.

(Refer Slide Time: 12:27)



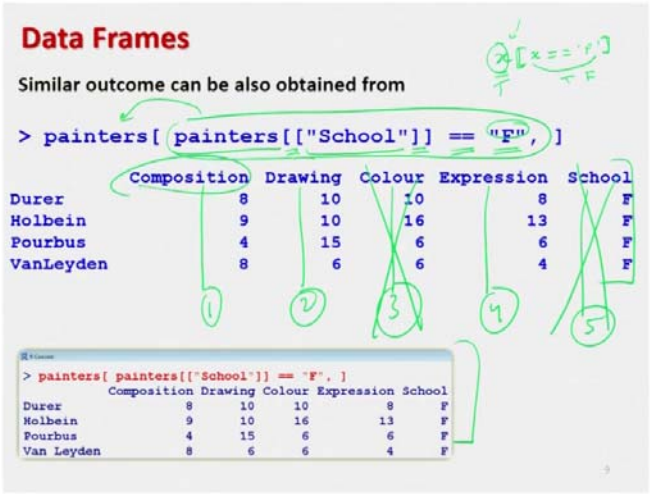
And if you try to see here if you try to see here, these are the four painters you can see here whose schools are here F in the last column like here, here, here and here. And out of this bigger data set you simply want to extract this only part of the data set where the school is equal to F. Now, you have to use your earlier knowledge right the first question comes here that how are you going to indicate that schools are going to be equal to only this F. So, F is a factor variable.

So, this F is a character. So, I try to write down here the name of the variable is school. And then double equality sign double equality sign you know this is the logical equal and then within course I try to write down here the name of the character that is here capital F. So, I try to use here the command subset and then I try to write down here the name of the data frame painters. And after that, I have to write down the condition this condition can be anything for example, if you want to know that how many painters are there whose drawing values are say more than 10, etcetera.

So, this type of data set you want to extract from the bigger data frame painters. So, once you try to execute it you will now see here this type of outcome, right. So, you can see

here that we have here these four schools and this is the same data set which you saw on the R console also, right.

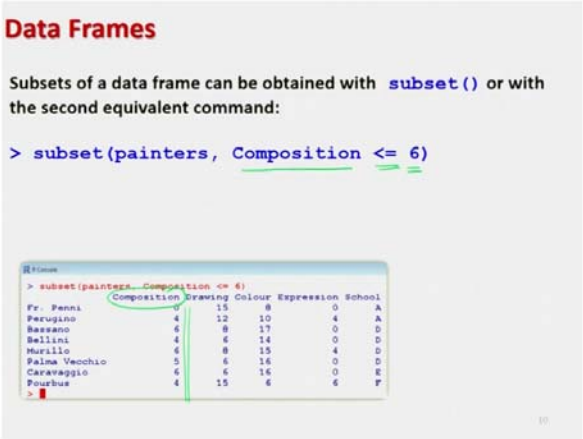
(Refer Slide Time: 14:01)



And similarly there is a another outcome if you try to recall you had done a command like something like x, x is equal to say here F type of thing. So, the outcome of this variable x in which the values are equal to here F means they will they will be obtained in terms of true and false. And then you try to write down here x and inside the square bracket, if you try to write this condition then whatever are the values in the x for which the value is coming out to be true they are also reported.

So, if you want to use that type of command also. So, there is an alternative command that you try to write down here this painters and within the double quotes you try to write down the name of the variable that is here is school and it is logically equal to F. And F is going to be inside the double quotes and then you try to write that what are the values under this condition which have the value true and inside the data vector inside the data frame painters, right. So, it will also give you the same value here, right. So, here also you will get the same outcome.

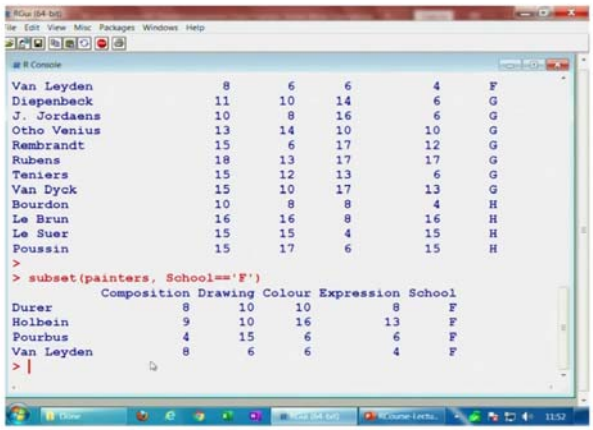
(Refer Slide Time: 15:15)



So, and similarly if you want to have a subset of those painters whose composition values are less than or equal to 6. So, you can see here this is your here composition variable and all those painters whose composition values are smaller than or equal to 6 they will be reported here. So, you can see here now you can means extract different type of data sets.

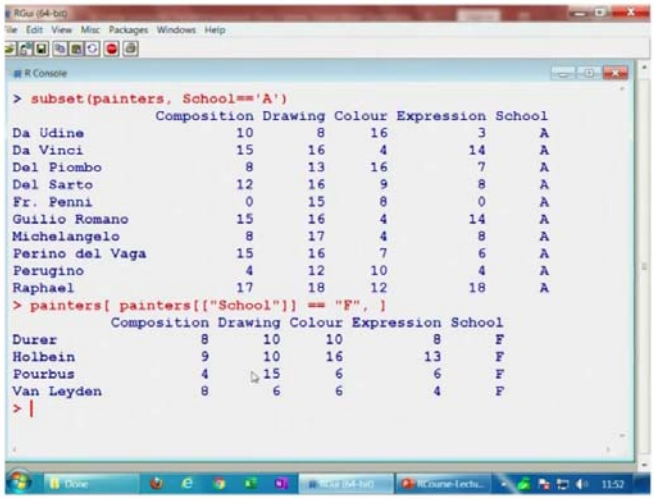
So, why not to first means execute these operations on the R console, right. So, firstly, let me try to use here this command here subset and let us try to see how it operates. So, you can see here with the subset command we are going to extract this value of data set Durer, Holbein, Pourbus and Van Leyden, right.

(Refer Slide Time: 16:07)



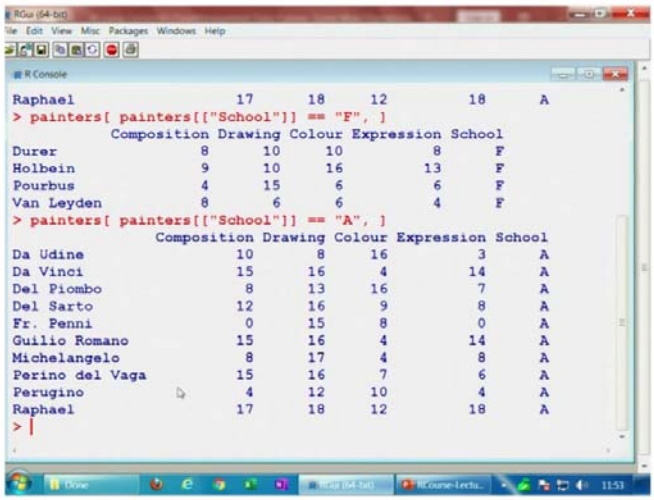
So, if you try to see here I am simply writing here subset and then painters, school is equal to F. And I have got a same data set Durer, Holbein Pourbus and Van Leyden, right.

(Refer Slide Time: 16:25)



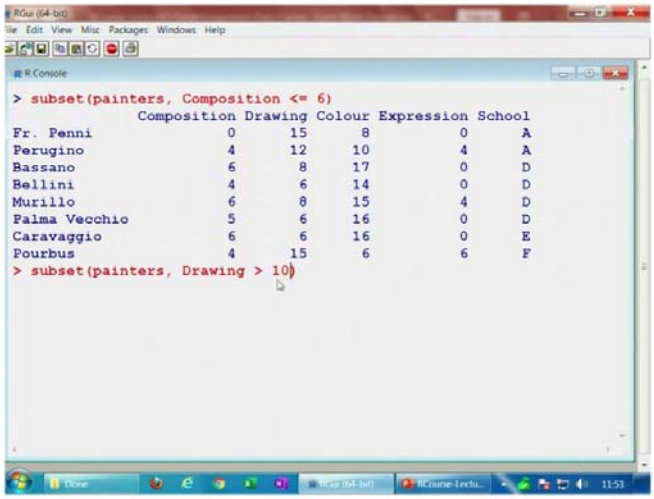
Similarly, if you want to find out here those painters whose school is here A, you can write down here like this, right. And if you want to see the other alternative which I shown you here if you try to use it here you here get the same outcome.

(Refer Slide Time: 16:39)



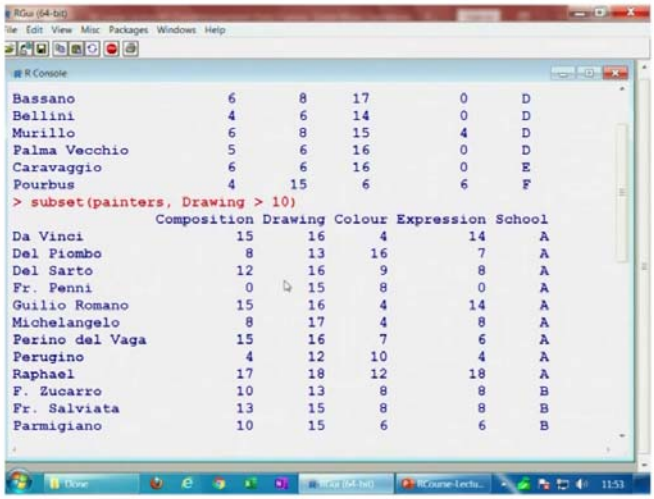
And similarly, if you want to take it here the school here to be A, you get here the same outcome. So, these are different ways. So, you know that when you are trying to work in the programming in order to do the same thing there are different ways actually.

(Refer Slide Time: 16:59)



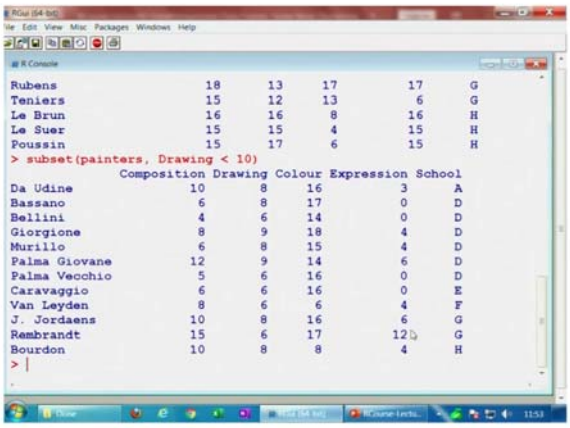
And similarly, if you want to have the subset of that data set in which the composition has a value which is less than or equal to 6, then you can do like this. And similarly, if you want to have it on the on here drawing is more than 10. So, you can change this command here, and I can write down here drawing is greater than suppose here 10.

(Refer Slide Time: 17:19)



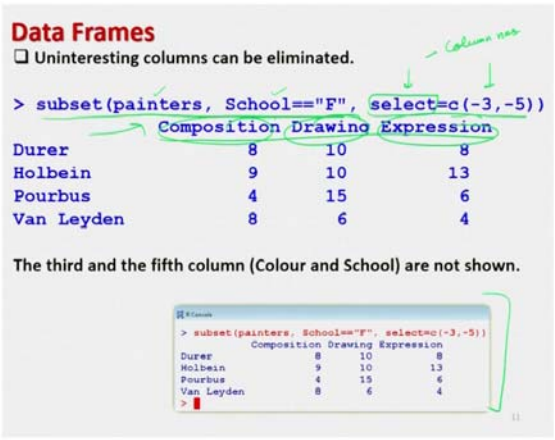
And you can see here you get here. All these data set right means every drawing is more than 10, right.

(Refer Slide Time: 17:25)



And similarly, if you want to have here all those painters whose drawing value is smaller than ten less than 10. So, it is here like this. So, you can see here that you can make here different types of such operations, right. So, now, let me come back to our slides and let us try to understand one more command ok.

(Refer Slide Time: 17:49)



Now, I try to explain you here one more operation in the data frame suppose you have got a data frame where there are several columns and you want to create here a subset in which you want to ignore or you want to eliminate some of the columns for that we have

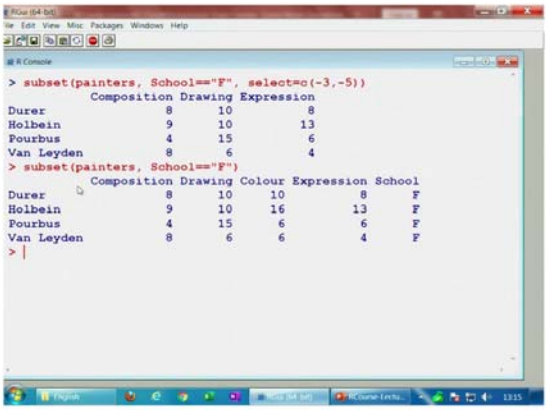
a command here select s e l e c t. So, for example, if you want to have a subset from this painters whose school is equal to here F, but now earlier you had shown that this has here all the values here you can see here like this composition, drawing, color, expression and school.

So, here you can see here this composition is here column number 1, drawing here is column number 2, color here is column number 3, expression here is column number 4, and the school here is column number 5, right. So, what I try to do here? Suppose I want to remove or eliminate the column number 3 and column number 5. So, under this option select s e l e c t all in lower case alphabet equal to I will try to write down the column numbers and then I will use here the minus sign.

So, as soon as I use here the minus sign that will inform the R ok, that these columns have to be eliminated from the subset, right. So, for example, if I try to take here -3 and -5, so I try to use here the data vector. So, what will happen in my outcome? This third and fifth column are going to be eliminated and you will get here this type of outcome.

So, you can see here that in your means earlier this 3 and here 5. They are going to be eliminated. So, you have only here composition drawing and expression and if you try to see here, here you have got here composition drawing and here expression and this is here the screenshot. So, let me try to show you this operation on the R console also.

(Refer Slide Time: 20:05)

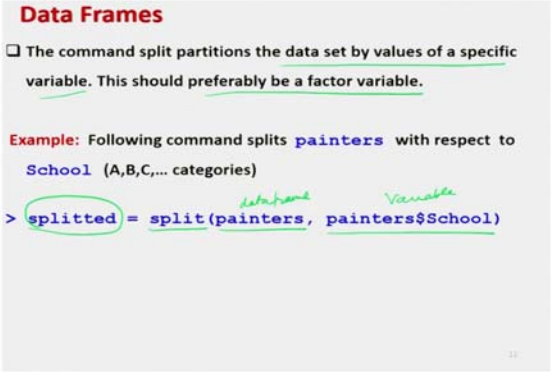


So, if you try to see here I try to take here like this. So, you can see here that here the third and fifth column are eliminated. And if you want to see here what are those

columns you can simply see here there are composition, drawing, color, expression, school. And then we have here only composition, drawing and expression and the column number 3 for the color and column number 5 for the school they are eliminated here.

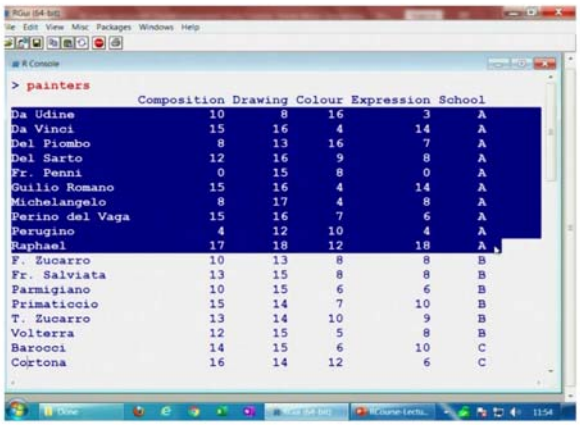
And yeah this data can be stored in a new variable and then you can do different types of operation on that data frame, we can see here that you can make here different types of such operations, right. So, now, let me come back to our slides and let us try to understand one more command.

(Refer Slide Time: 20:55)



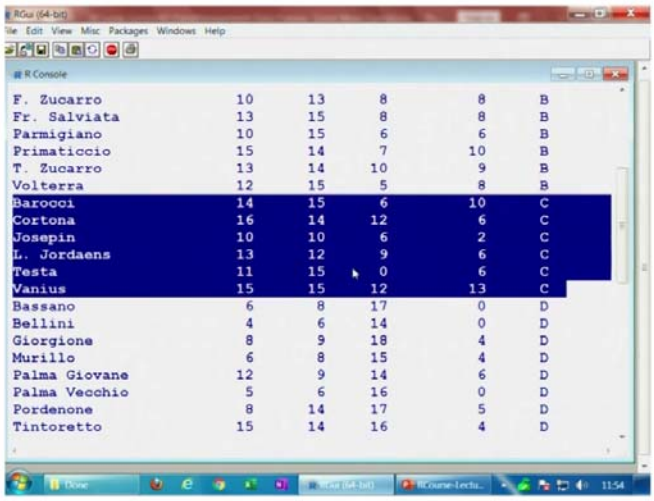
So, now if you try to see suppose let me try to take the same example here and yeah first let me try to show it on the R console itself so, that you can understand it better, right.

(Refer Slide Time: 21:07)



So, if you try to see here in these data set painters you can see here this is here is school. And these many values they are corresponding to when the school is A, and then the there are these many values which are highlighted on the screen which are corresponding to school B.

(Refer Slide Time: 21:29)

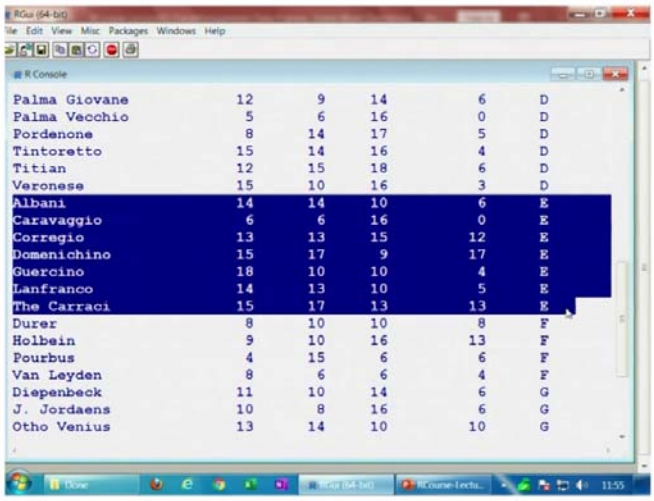


And then there are values which are corresponding to school C.

(Refer Slide Time: 21:35)

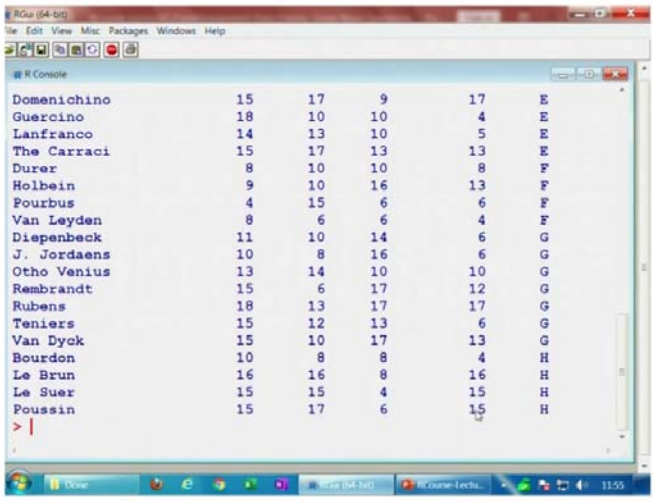
Then there are values which are corresponding to school D.

(Refer Slide Time: 21:39)



And similarly there are values which are corresponding to school E.

(Refer Slide Time: 21:45)

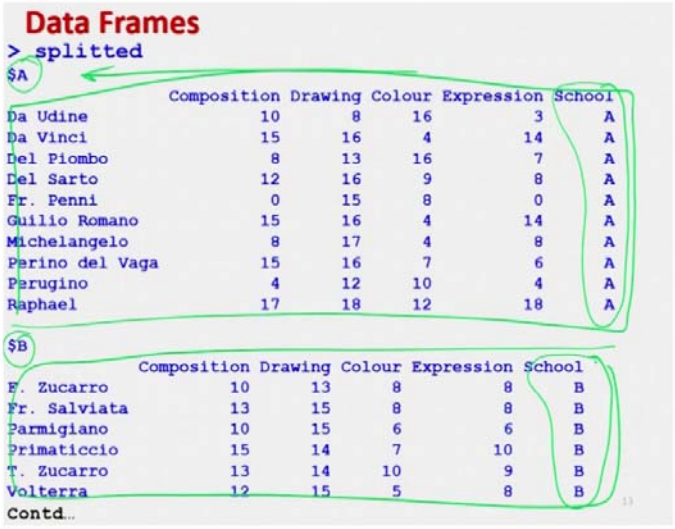


Then F and then here G. And then can here H finally, right. So, you want to split the data at all these point wherever the school is changing or you can actually split at any other value also depending on the variable. So, now how to split the data is the question that we are going to answer. So, in order to split the data set by a value that is corresponding to any specific variable that can be done by the command split and we always prefer to have this splitting preferably for a factor variable.

So, for example, I have shown you for this school this is the factor variable and the values are like A, B, C, D, E, F, G, H, right. So, let us try to see how it will look like because this will continue to a couple of screen.

So, that is why I have used here the commands here is split and then whatever is the outcome I have stored here in this variable is splitted. So, I try to command here say split. And then I try to write down here the name of the data frame say painters and then comma. And then I try to write down here the name of the variable by which I want to partition it, ok.

(Refer Slide Time: 23:15)



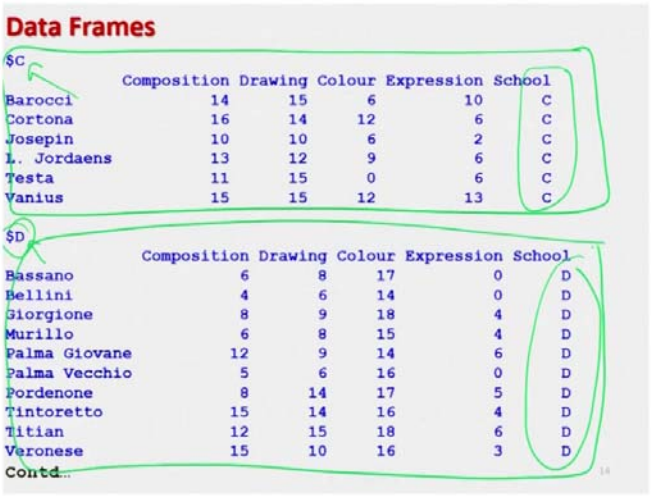
	Composition	Drawing	Colour	Expression	School
Da Udine	10	8	16	3	A
Da Vinci	15	16	4	14	A
Del Piombo	8	13	16	7	A
Del Sarto	12	16	9	8	A
Fr. Penni	0	15	8	0	A
Guilio Romano	15	16	4	14	A
Michelangelo	8	17	4	8	A
Perino del Vaga	15	16	7	6	A
Perugino	4	12	10	4	A
Raphael	17	18	12	18	A

	Composition	Drawing	Colour	Expression	School
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

So, now if you try to execute this command you get here this type of outcome that is going to continue in the couple of screens. So, I have simply copied it here, and I will try to show it on the R console also. So, the value of the split will be the first value here will be dollar A. So, it is trying to indicate that this is the first variable by which the partitioning has been done. So, you can see here in this case all the schools have got the value A. So, that is indicated by this here dollar A.

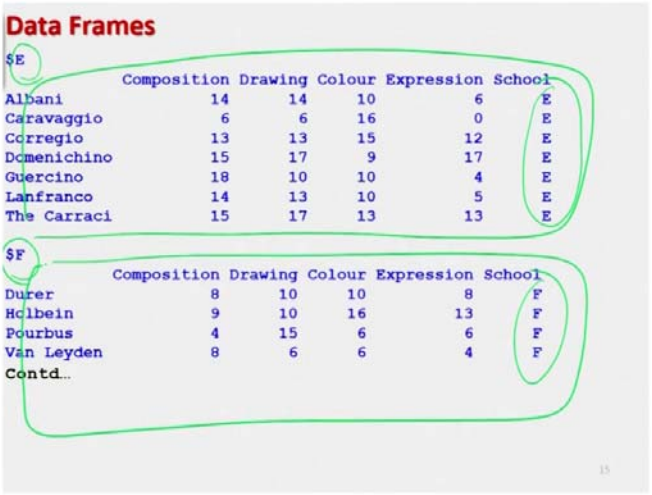
And then entire data set which is corresponding to this school A, this is stored here and after that this is here is say dollar B and whatever is the data corresponding to school B, this is stored here like this.

(Refer Slide Time: 24:01)



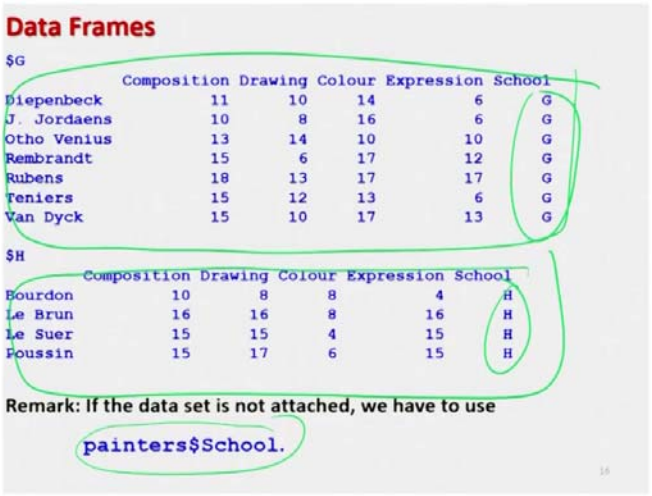
And similarly, if you try to continue, this is here is school C. And this data corresponds to school C. And this is indicated here by here dollar C. And similarly, we have here dollar D. And then school Ds are here and the entire data which is here which is corresponding to school D, right.

(Refer Slide Time: 24:19)



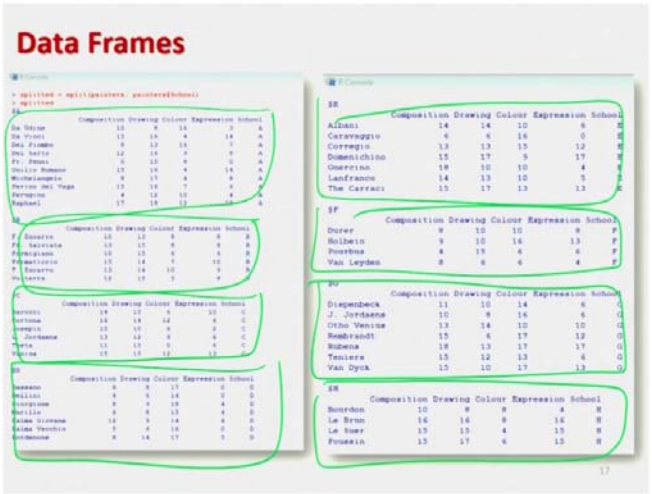
And similarly, we have here school E and here is school F. And then their data is indicated by dollar E and here dollar F. And this is the data which is here like this for E and this is the data for here F.

(Refer Slide Time: 24:35)



And similarly, for the G and H also this is here the data corresponding to G. And then corresponding to here H you can see here like this, right. So, in yeah in case if you attach you can simply use here the variable name and if you do not attach the data frame then you have to use the name of the variables using the name of the data frame and the variable, right.

(Refer Slide Time: 25:03)



And yeah you can see here this is how the screenshot will look like this is your here A this is your here B, this is your here C, this is your here D, this is your here E, this is your

here F, this is your here G, and this is your here H yeah definitely. If I have to compile all the things in a single screen the phone size is going to be (Refer Time: 25:27) smaller.

(Refer Slide Time: 25:27)

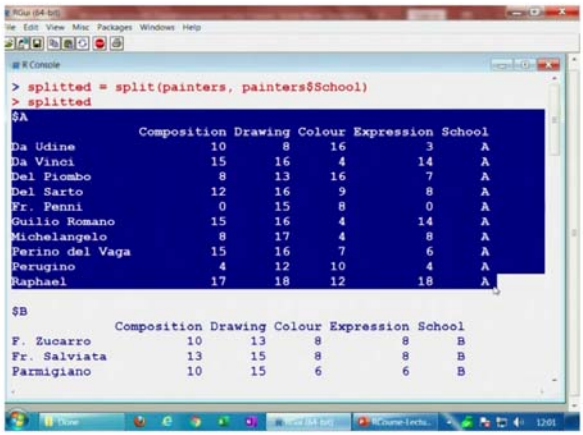
And then, yeah you have to just reduce your phone size. So, I will request you that you try to do the same operation yourself on your computer and try to see how the values are going to be, right. Now, the next question comes that whatever is the outcome of this splitted value number 1, how are you going to access only a subset based on this splitting.

So, if you simply try to write down here the name of the and followed by dollar sign and then here the name of the factor variable under which the data has been splitted; for example, if you want to have this data here which is for here A like this one here. So, you simply have to write down here this dollar A along with this splitted sign. So, if you try to see here you are I am simply trying to write down here splitted dollar A.

So, that will give me the data, and if you want to see whether the splitted value are they also data frames; for example, you have a splitted dollar a splitted dollar B, etcetera, etc. up to splitted dollar H, right. These are the partitions or these are the values which are obtained by splitting the data frame with respect to the variably school A to H, then are these partition value are data frame?

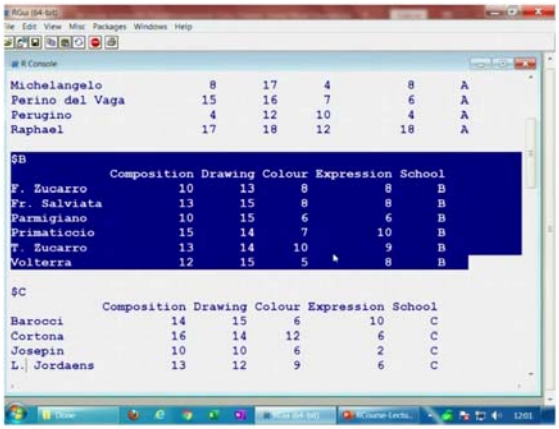
So, if you try to see here I use here the command is dot data frame and inside the parenthesis I splitted dollar A, and this gives me answer true so; that means, whatever splitting you have done right that is also a data frame, right. So, let me try to do this example on the R console and try to show you that how do you get here this type of outcome, ok.

(Refer Slide Time: 27:23)



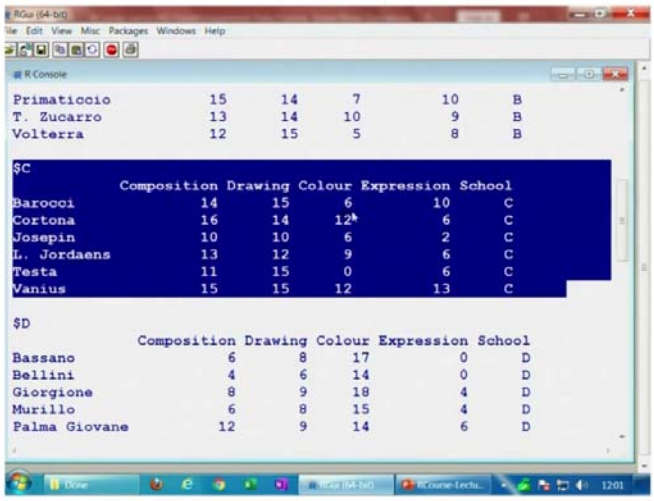
So, if you try to see here as soon as I try to execute this split the split command s p l i t all in lowercase alphabets with the name of the data frame and the name of the variable painters, dollar, school, then you can see here this is going to be like this. And yeah definitely it is going to continue in a couple of screen. So, you can see here this is here dollar A.

(Refer Slide Time: 27:47)



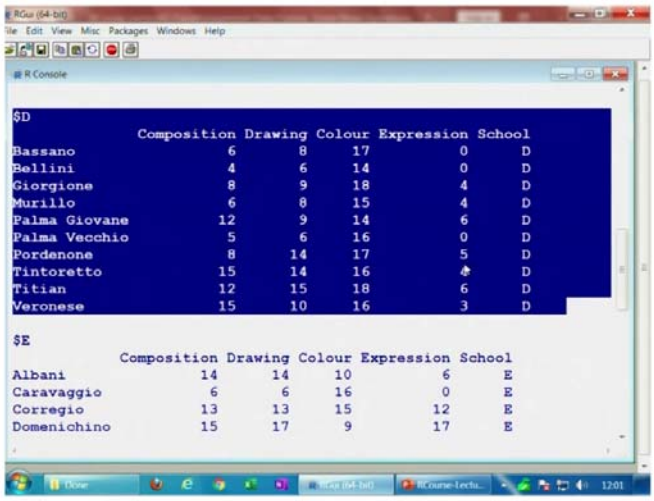
This is here dollar B.

(Refer Slide Time: 27:49)



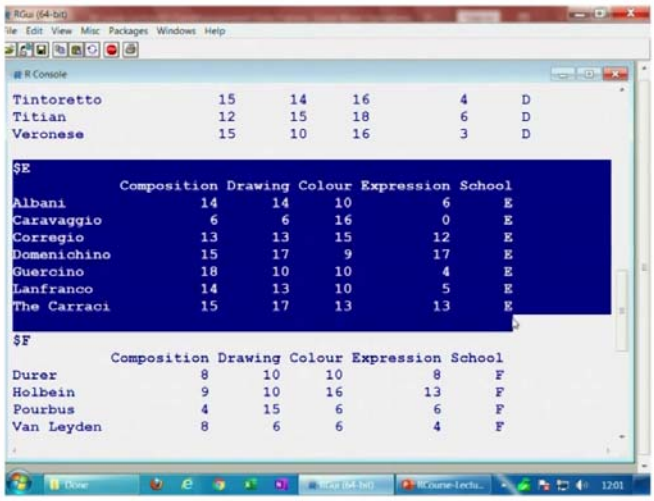
And then similarly you have a dollar C.

(Refer Slide Time: 27:51)



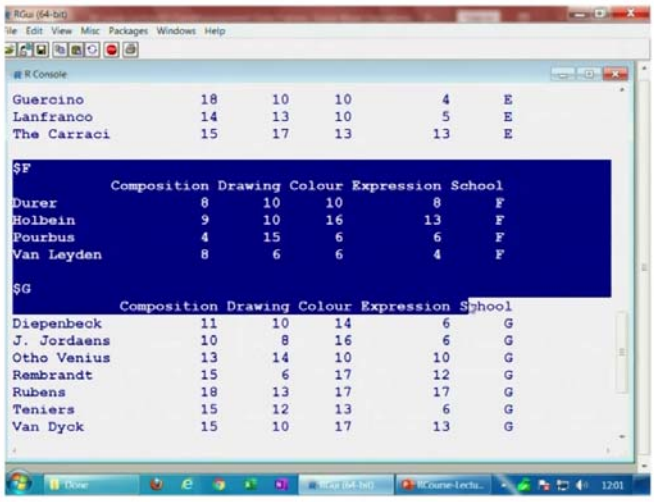
Then dollar D.

(Refer Slide Time: 27:53)



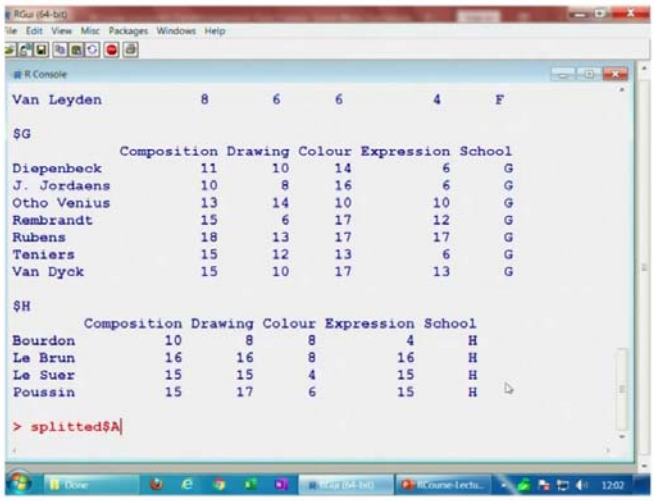
And then dollar E.

(Refer Slide Time: 27:57)



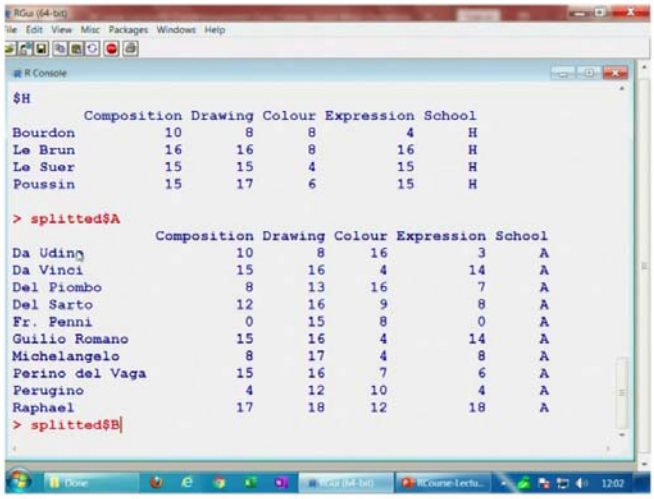
And then dollar F.

(Refer Slide Time: 28:01)



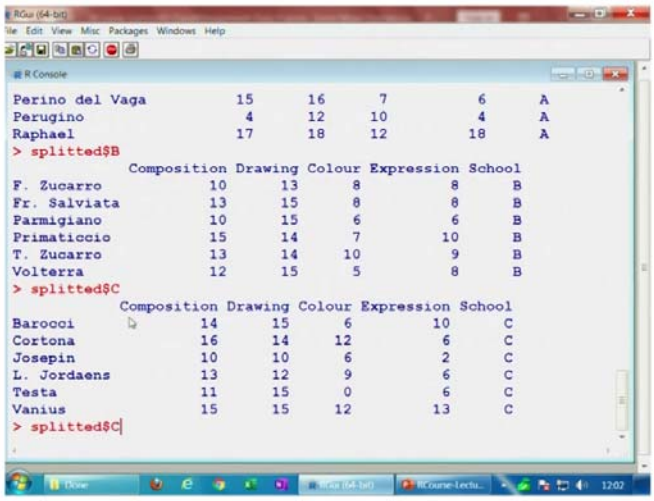
And then dollar G, and then here dollar H. So, all this dollar A, B, C, D, E, F, G, H they are going to indicate the sort of set of the data frame where school is equal to, B, C, D, E, F, G, H respectively, right.

(Refer Slide Time: 28:21)



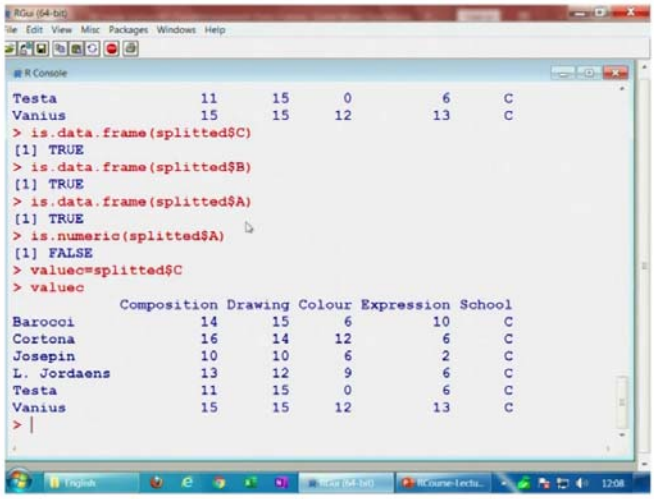
So, if you try to see here if you want to see what is the data under a. So, you simply have to write down here splitted A and you will get here this data. So, this is how you can access the different partition value splitted B.

(Refer Slide Time: 28:27)



So, this will give you here the data set under the splitted B. And similarly, if you try to write down here splitted dollar C which is the name of the variable under which you have stored these values. So, it will actually work like this.

(Refer Slide Time: 28:59)



And if you want to see here whether these values say either A, B or C whatever it is it a data frame. So, you try to write down here is dot data dot frame and then you will see here this is here true, right. So, then it is whether it is C or B or here a this is here true

right and if you try to see here is it a numeric is dot numeric for example, it will say here false.

So, you can now see here that whatever partitions you have obtained by splitting the data set with respect to the variable school they are also data frames. So, now, once they are the data frame then you can execute all the commands whatever you have learnt under the data frame on these data sets also and for that you can save them in different values. For example, if I say here like this suppose if I want to save here the values of say school C.

So, I can say here value say here c and now you can see here the value c here is like this and after that yeah means you can make different types of operation whatever you want on it right ok. So, now, we come to an end to this lecture and then you have seen that ok this was a also a pretty simple lecture. And I have tried my best to explain you that how the things are functioning when you are trying to deal with the data frame.

And this is the most important part in any programming that you have to understand that how the software is working and how it is giving you the output and what is the nature and behavior of the output. So, that will help you in writing the correct program for executing what you want. So, I would now request you once again that you please try to take a data frame or try to see or try to look into your older some excel file or spreadsheet and whatever operation you were trying to do there try to do it in the R software also.

Yes, there is a question that how can you read the data from the excel file or any other software in the in inside the R software, that I will try to tell you after say one more lecture on that data frame. So, you try to practice it and try to understand how these commands are working and I will see you in the next lecture, till then goodbye.