

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

**Lecture – 32**

**Data Frames**

Welcome to the next lecture on the course Introduction to R Software. You may kindly recall that in the earlier lecture we discussed about the data frame and in particular we talked about that how to extract the information on a particular variable from a data frame. And we had used the syntax like name of the data set and then name of the variables and they are joined together with the dollar sign and this is how we used to define the variable on which if we want to extract the information from the data frame. There are some other options to extract information on a particular variable from the data set without using the dollar sign.

So, we will try to start this lecture with this topic and you will try to do some more commands and we are going to use here the same data set that we have used in the last 2 lectures. And so can I try to have a quick revision of that data set this data set was the painters which was containing the information on different types of painters over here and longest.

(Refer Slide Time: 01:20)

**Data Frames**

An example data frame `painters` is available in the library MASS (here only an excerpt of a data set):

```
> 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
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.

Here, the names of the painters serve as row identifications, i.e., every row is assigned to the name of the corresponding painter.

2

And then it has information on several variable composition and drawing color expression school and these variables are your here the numeric variables and whether this is you here the categorical variable.

So, you can see here this is taking the values as A B C D and so on and these are the some numerical values and this is the screenshot of that data set that we had used earlier also.

(Refer Slide Time: 01:50)

### Data Frames

```

> 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
.	.	.	.	.	.
.	.	.	.	.	.
Rubens	18	13	17	17	G
Teniers	15	12	13	6	G
Van Dyck	15	10	17	13	G
Bourdon	10	8	8	4	H
Le Brun	16	16	8	16	H

3

So, now, let us try to come to our discussion.

(Refer Slide Time: 01:55)

### Data Frames

Attaching a data frame

With a command attach() over the data frame, the variables can be referenced directly by name.

It can address the names of a data frame directly, without the prefix dollar sign operator, e.g. painters\$.

**Example**

```
> attach(painters)
```

Variable names are

- Composition,
- Drawing, ✓
- Colour, ✓
- Expression, ✓
- School ✓

4

Let us start with the aspect that how to attach a data frame. There is a function what is called as `attach` and this function is used to something called `attach`. The syntax here is we try to write down `attach` and inside the arguments we try to write down the name of data set. Well, what really happens with `attach` that we will try to see.

The advantage of using this command is that then the variables can be referenced directly by their name and this will allow us to address the names of data frame directly without using the dollar sign operator. For example, earlier in order to extract the information on a particular variable we used to write down this name of data set then name or variable and we try to attach it and we try to join it with a dollar sign. So, here we are going to understand another option where I can address a data frame or a particular variable without using the dollar sign.

So, first we try to write here the command say `attach` `painters`, `painters` is the name of the data set you may recall that this painter data set has 5 variables `composition` `drawing` `color` `expression` and `school`. So, now, in case if you want to have information on any of this variable what we have to do that I can directly call these variables in the function.

(Refer Slide Time: 04:06)

The screenshot shows an R console window with the following content:

```
Data Frames
> summary(School) # Character variable
  A  B  C  D  E  F  G  H
10  6  6 10  7  4  7  4

> attach(painters)
> summary(School)
  A  B  C  D  E  F  G  H
10  6  6 10  7  4  7  4

painters$Composition
> summary(Composition) # Numeric variable
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.00   8.25   12.50   11.56  15.00   18.00

> summary(Composition)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.00   8.25   12.50   11.56  15.00   18.00
```

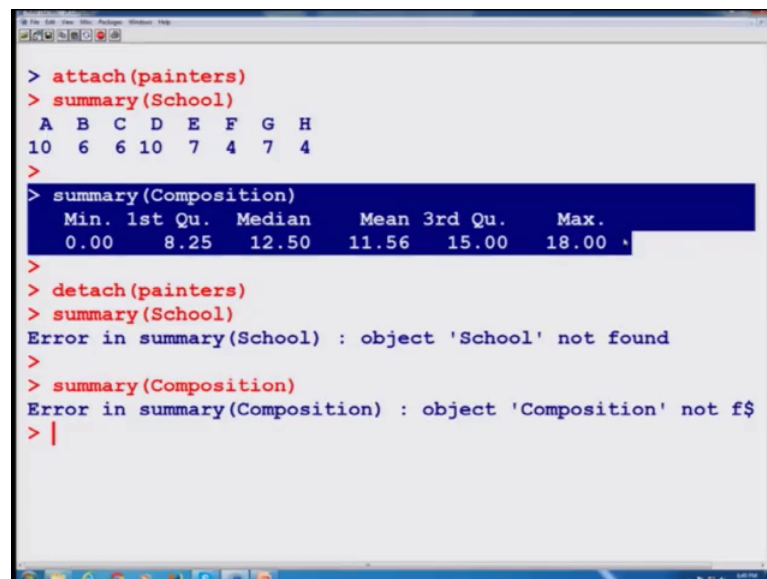
Handwritten annotations in green include: `summary(painters$School)` above the first summary, `painters$Composition` above the second summary, and a bracket on the right side grouping the `attach` and `summary(School)` commands.

For example, you may recall that earlier we had used the function `summary` and suppose I want to have some information about the variable `school` then I will write the name of the variable then the name of the data set and then I will try to join them by a dollar sign. And this is how I used to find out the summary of the variable `school`. But now you can

see here once you have attached the data set painters I can use here the summary command directly over the school without using the dollar sign and this give us this option and the screenshot is given over here you can see. So, this example I am taking using a character variable and similarly I would try to take another example which is a numeric variable and for example, composition.

So, if you try to see here now I am not writing here painter dollar composition something like painters, dollar, composition, but I am directly using here the variable name composition. So, I am now trying to operate the summary function directly over this variable and I get the same information over the minimum value first quartile median mean third quartile maximum and this is the screenshot over here. So, why not to do it over the R console and try to see what do we obtain over here. So, first I try to do here say attach painters so now, here done.

(Refer Slide Time: 06:02)



```
> attach(painters)
> summary(School)
 A B C D E F G H
10 6 6 10 7 4 7 4
>
> summary(Composition)
  Min. 1st Qu.  Median    Mean 3rd Qu.   Max.
 0.00   8.25   12.50   11.56   15.00   18.00
>
> detach(painters)
> summary(School)
Error in summary(School) : object 'School' not found
>
> summary(Composition)
Error in summary(Composition) : object 'Composition' not found
> |
```

And now I would like to use my command here say summary of the school and you can see here that this gives here say my outcome and similarly if I try to use here another summary function on the composition variable it gives us this thing. So, you can see here now I am not doing the same operation which I did earlier.

(Refer Slide Time: 06:39)

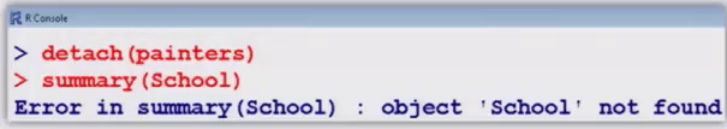
### Data Frames

❑ The command `detach()` recovers the default setting and then we have to use `painters$` again.

```
> detach(painters)
```

*(name of data set)*

```
> summary(School)
Error in summary(School) : Object "School" not found
```



The screenshot shows the R Console with the following text:

```
> detach(painters)
> summary(School)
Error in summary(School) : object 'School' not found
```

6

Now we come back to a slides and try to continue. Well, we have now learned how to attach a data set after we are done we would like to remove this function; that means, this is done by the function `detach` and once I use this function then we go back to the previous one and this function `detach` recovers the default settings and once I have detach the data set then I have to use the `painters` dollar as I did earlier.

So, in order to detach a data set which was attached earlier we try to use the function `detach` and inside the argument we have to write the name of data set and then it is done. And once you use this command then if you try to find out the summary of your school it will say that there is some problem this and this variable is not found. And here is the screenshot of the outcome, but we would like to do it over the R console also. So, you can see here that I have attached here, the data set `painters` earlier now I am trying to remove it by using the function `detach` `detach` `painters` and it is now done.

Now if I try to find out here the summary school which I have obtained here earlier you can see this is the function which I use and it has a this type of outcome please try to concentrate on the highlighted part earlier it was giving this outcome, but now once I use it is giving me that there is array in this function. And similarly if I try to use over the composition this is again giving us that there is array in the summary composition this and this form variable is not found whereas, earlier when the data was a attach I was

getting this outcome. So, this is how you can attach and detach a data set while doing any analysis.

(Refer Slide Time: 09:04)

**Data Frames**

Subsets of a data frame can be obtained with `subset()`.

Example: `> subset(painters, School=='F')`

(# == means logical equal sign)

	Composition	Drawing	Colour	Expression	School
Durer	8	10	10	8	F
Holbein	9	10	16	13	F
Pourbus	4	15	6	6	F
VanLeyden	8	6	6	4	F

Now, we discuss next topic. Suppose we have an objective to extract a subset of data from the data frame earlier we had discussed how to extract a data value using the matrix notations but now, I want to extract more than one values which are constituting a subset of the data frame. So, the question is how to get it done.

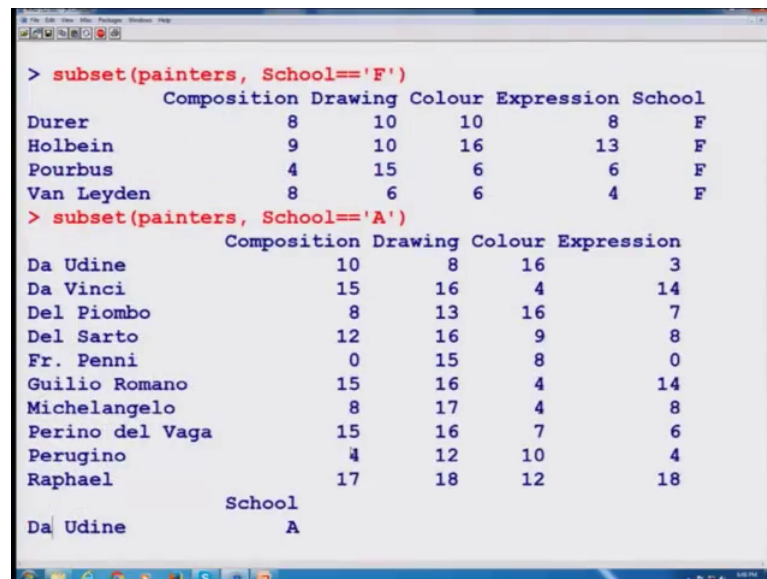
So, in order to extract a subset the syntax is here subset and inside the argument we write the name of data set and other options, so why not to take an example and try to understand that how it is being done. So, we are again going to use the data set painters. So, I try to write down here subset then I am trying to write down here the name of data set and then I am writing here the options which I want to use. And suppose I want to extract a subset of the data set from this data frame painters where the school is actually F you can see here I have used here the double equality sign between the logical equal sign exactly equal to.

So, now I am asking are to provide me those data values from the entire data set for which the value of the variable school is F that is what I am trying to write down here. And this F is written you can see here inside the double quotes that is the standard notation once I try to do it here I get an outcome of this type. So, you can see here now from the entire data frame the subset command is giving me that there are 1 2 3 4

observations whose school is equal to F and corresponding to which this is giving me this entire data set. So, this is a subset of the entire data set contained in the data frame painters right.

So, why not to do it toward the R console and try to see what happens.

(Refer Slide Time: 11:45)



```
> subset(painters, School=='F')
  Composition Drawing Colour Expression School
Durer          8      10      10          8      F
Holbein         9      10      16         13      F
Pourbus         4      15       6          6      F
Van Leyden      8       6       6          4      F

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

  School
Da Udine  A
```

You can see here we get this type of outcome and similarly if you want to have what is the data set for which the school is equal to here you can get here all this information you can see here. Similarly for school set here see here D suppose I want to have this information I get here like this there these many painters have gone to school D.

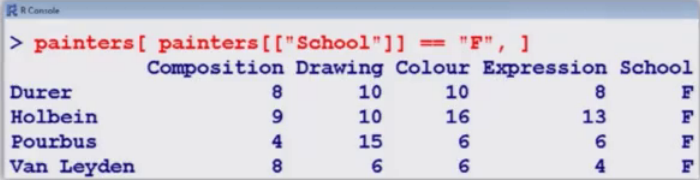
(Refer Slide Time: 12:13)

### Data Frames

Similar outcome can be also obtained from

```
> painters[ painters[["School"]] == "F", ]
```

	Composition	Drawing	Colour	Expression	School
Durer	8	10	10	8	F
Holbein	9	10	16	13	F
Pourbus	4	15	6	6	F
VanLeyden	8	6	6	4	F



```
> painters[ painters[["School"]] == "F", ]
```

	Composition	Drawing	Colour	Expression	School
Durer	8	10	10	8	F
Holbein	9	10	16	13	F
Pourbus	4	15	6	6	F
Van Leyden	8	6	6	4	F

So, let us now come back to our slides and now we are going to discuss another aspect. The thing whatever we have done using here the subset command can also be done in the same way that you did earlier you might recall that you had used the symbol of double brackets and this was like this here using double brackets over here. So, if you try to see here I am trying to write down here the name of the variable school inside the double quotes and then I am trying to say this variable is from the data set called painters and then I am trying to say from this part a variable name school from a data set called painters. And now I am trying to say please use the data set painters and inside this data set try to find out the values of variable school under this painters which are equal to F. Be careful F does not mean false here F is one of the categories of the school the schools are denoted by A B C D E F and so on right.

So, now if you try to do it here you get here this outcome. That is the same outcome you can see here which you have got earlier you can see here.



(Refer Slide Time: 13:51)

### Data Frames

Subsets of a data frame can be obtained with `subset()`.

Example: *name of data set*  
> `subset(painters, School=='F')` == ==

(# == means logical equal sign)

	Composition	Drawing	Colour	Expression	School
Durer	8	10	10	8	F
Holbein	9	10	16	13	F
Pourbus	4	15	6	6	F
VanLeyden	8	6	6	4	F

7

This is the same outcome that you got here and this here is the screenshot you can try it right.

(Refer Slide Time: 14:05)

### Data Frames

Subsets of a data frame can be obtained with `subset()` or with the second equivalent command:

> `subset(painters, Composition <= 6)`

```
R Console
> subset(painters, Composition <= 6)
  Composition Drawing Colour Expression School
Fr. Penni    0      15      8           0      A
Perugino     4      12     10           4      A
Bassano      6       8     17           0      D
Bellini      4       6     14           0      D
Murillo      6       8     15           4      D
Palma Vecchio 5       6     16           0      D
Caravaggio   6       6     16           0      E
Pourbus      4      15      6           6      F
```

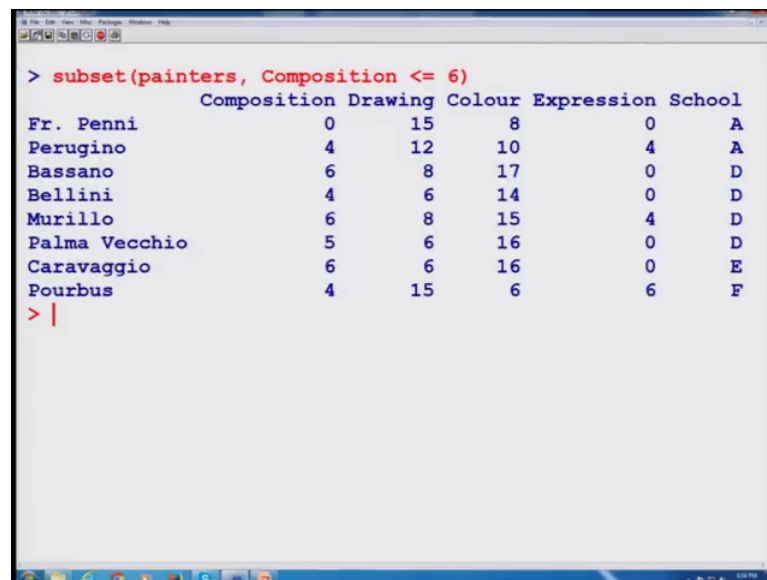
9

Now, I try to do some more calculations using this subset functions in this case we have found a subset where the school takes value F. Now suppose my objective is to extract a subset of data from the data set or the data frame painters where the variable composition is less than or equal to 6 this is the logical operator of less than or equal to something like this right, if you remember. And as soon as I do it I get an outcome of this type here you

can see here that it is trying to give us the name and other details of these painters for which the value of the composition is smaller than 6, you can see here none of the values is more than 6 and corresponding to these observations it is trying to give me all other information over here.

So, now we have extracted a subset from the painters data set under that the composition variables takes value less than or equal to 6. So, let us try to do it over the R console and see what we obtain.

(Refer Slide Time: 15:31)



```
> subset(painters, Composition <= 6)
      Composition Drawing Colour Expression School
Fr. Penni         0     15      8          0      A
Perugino          4     12     10          4      A
Bassano           6      8     17          0      D
Bellini           4      6     14          0      D
Murillo           6      8     15          4      D
Palma Vecchio    5      6     16          0      D
Caravaggio        6      6     16          0      E
Pourbus           4     15      6          6      F
> |
```

You can see here that is the same outcome that we have obtained.

(Refer Slide Time: 15:39)

### Data Frames

Uninteresting columns can be eliminated.

```
> subset painters, School=="F", select=c(-3,-5))
```

	1	2	3	4	5
	Composition	Drawing	Expression	Colour	School
Durer	8	10	8		
Holbein	9	10	13		
Pourbus	4	15	6		
Van Leyden	8	6	4		

The third and the fifth column (Colour and School) are not shown.

10

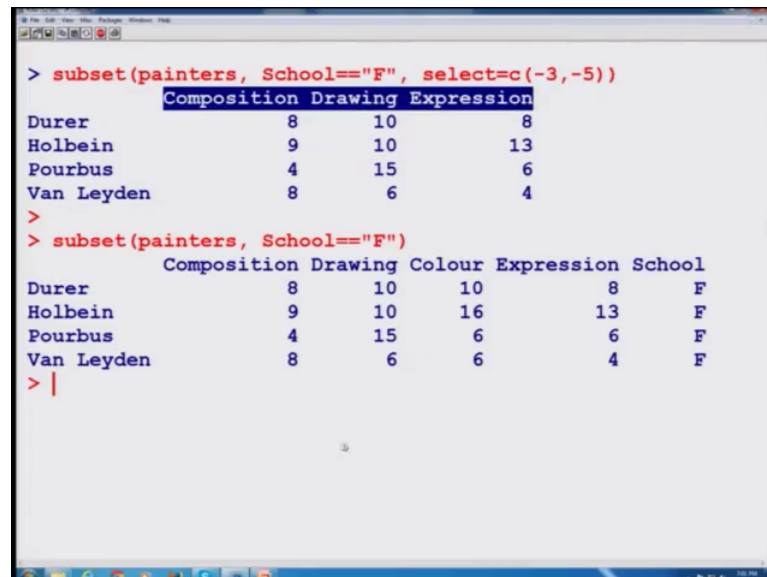
Now, take another example. So, now, suppose I am not interested in say 2 columns say third column and fifth column and further what I want is the following. I want to have a subset of the data from the data frame painters for those painters who have gone to school F. So, this part is going to give me the data set for which the school is equal to F and this is the same data set you can see here we had obtained here first we try to have a look. So, that you can compare the outcome this is here this thing, this is the data set we have obtained here and you can see here first column here this composition, second column is drawing, third is color, fourth is here expression and fifth is here say school and suppose I am not interested in the third and fifth column and I want to remove it. So, that is what I am trying to do here.

So, I am now trying to say here select this is another option what third and fifth column and I want to remove it. So, I am using here a negative sign and since there are 2 columns so I am trying to combine them the with the command c. So, I am so now, finally, I am trying to say please try to find a subset of the data from the data frame painters for which the school is equal to F and select 2 columns third and fifth columns and since I have used here negative signs. So, that is indicating that please remove them.

Now, you get here the outcome you can see here you are getting away here only one composition variable 2 drawing variable and 3 expression variable. The original 2 variables which were the third and fifth variable here you can see the color and say

school they are no more here. So, you have removed those 2 columns from this subset right and this is here the screenshot, but we would also like to do it on the R console to see what happens.

(Refer Slide Time: 19:01)



```
> subset painters, School=="F", select=c(-3,-5)
  Composition Drawing Expression
Durer           8      10         8
Holbein         9      10        13
Pourbus         4      15         6
Van Leyden      8       6         4
>
> subset painters, School=="F"
  Composition Drawing Colour Expression School
Durer           8      10        10         8      F
Holbein         9      10        16        13      F
Pourbus         4      15         6         6      F
Van Leyden      8       6         6         4      F
> |
```

So, let us try to use this command over the R console and see what happens over here. So, you can see here the third and fifth column are removed just for the sake of understanding I try to get the subset without removing the third and fifth column you can see here this is the data set. So, here you can see that third column here color is and the fifth column here the school is missing from this list. So, this is how you can remove some columns which are indicating the data on certain variables in R.

(Refer Slide Time: 19:46)

**Data Frames**

□ The command `split` partitions the data set by values of a specific variable. This should preferably be a factor variable.

*split ( )*

**Example:** Following command splits `painters` with respect to `School` (A,B,C,... categories)

*factor variable*

```
> splitted <- split(painters, painters$School)
```

*split ( name of data set, name of factor variable )*

12

So, now we come back to our slides and now we cut a straight to consider another aspect. Suppose you have got a data frame the data frame will consist different types of data values and different types of variable and suppose you want to split the data set according to some variable, usually we try to choose a factor variable by which I can partition the data set or split the data set.

So, the question which we are going to answer here is that how to split a data set in a data frame. So, for that we have here a function which is called here as a split function and the syntax is very simple split and inside the arguments you have to write the syntax. So, let us try to take an example and try to understand its implications and its implementations. Suppose I am going to use here the data set painters and one of the factor variable in this data set is school this is my here factor variable. So, now, what is my interest you may recall that in the data set painters there were 8 categories of a school A B C D E F G H and it was something like school A, school B, school C up to school H. Now I want to split the data set according to the school say at a there should be split, at B there should be a split at C there should be a split and so on, that means, I want to get a sort of subset of data set which are splitted according to the factor variable see here is school.

So, now how to write this command the syntax is very simple try to write down the function split name of data set and then write down the name of variable or other more

specifically name of factor variable by which we want to split the data and this is separated by comma. So, here in this example I would like to use the data frame painters and I would like to split the data set with respect to school. So, this is how we try to define the variable name school right. So, this is painters dollar school and I try to store this value inside another say variable here say splitted that that is going to be a long outcome that will go through with different slides. So, first we try to understand the outcome and then I will try to show you over the R console.

(Refer Slide Time: 22:52)

```
> splitted
$A
  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

Contd..
> splitted <- split(painters, painters$School)
> splitted
$A
  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
```

The outcome what we get it is something like this; this is not the complete one that will go to several slides.

So, first split is occurring here with respect to the category A and you can see here that all the data values corresponding to category A they are appearing here. And this is the screenshot also I am going to show you here and this actually continues that is what I am writing here is continued.

(Refer Slide Time: 23:23)

### Data Frames

ⓈC

	Composition	Drawing	Colour	Expression	School
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

Contd...

```
> splitted ⓈC
```

	Composition	Drawing	Colour	Expression	School
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

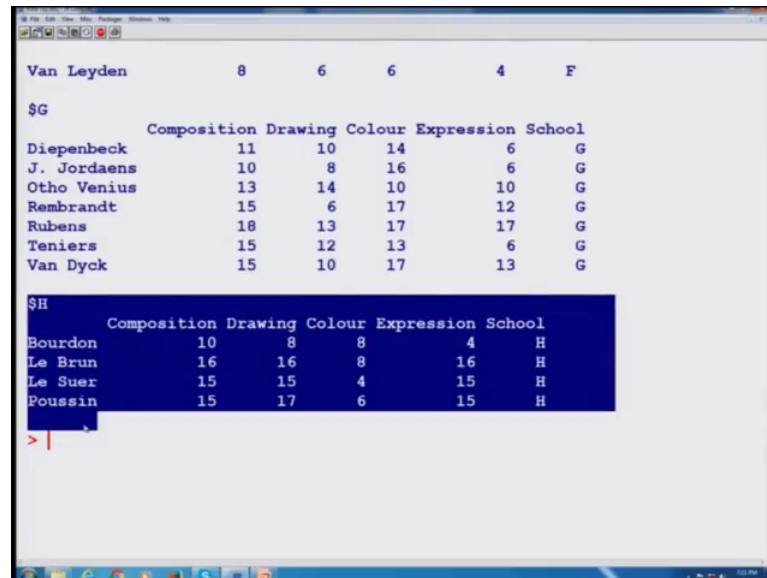
15

Then we get all the data values corresponding to the vector variable here B something like here it again continues and this is the screenshot of the outcome which is showing me here all the factor values here be this continuous further. And then it splits up the data with the factor variable here C you can see over here this further continues and this is here the screenshot where all the values are splitted with respect to here C.

Similarly, here this continues further and finally, it comes to the last factor here H and you can see here there are 4 values corresponding to factor variable having value H and this is the screenshot of this thing right. Before I go to show you the outcome on the R console I would like to make a remark that in case if the data set is not attached we have to use this painters dollar a school to provide the variable name.

So, now we try to do the same thing over here and we try to get this syntax here. So, now, you can see here that the values which are stored in a splitted are like this.

(Refer Slide Time: 24:50)



```
Van Leyden      8      6      6      4      F

$G
      Composition Drawing Colour Expression School
Diepenbeck     11     10     14         6      G
J. Jordaens    10      8     16         6      G
Otho Venius    13     14     10        10      G
Rembrandt     15      6     17        12      G
Rubens        18     13     17        17      G
Teniers       15     12     13         6      G
Van Dyck      15     10     17        13      G

$H
      Composition Drawing Colour Expression School
Bourdon        10      8      8         4      H
Le Brun       16     16      8        16      H
Le Suer       15     15      4        15      H
Poussin       15     17      6        15      H
> |
```

So, here you can see the outcome is going through several screen. Now here is the first split with respect to here factor A that you can see going further with some more values. Now then there is another split with respect to here factor B and then we have here another split with respect to factor C, then we have here another splitter is with respect to here factor D and similarly we have here one more split with respect to factor E and similarly we have here under split with respect to factor F, then we have another split with respect to factor G and finally, we have the last split with respect to H.

So, if you try to see what is really happening that this data is arranged with respect to the school variable A B C D E F G H and then wherever is the intersection of A and B there it occurs I split then again wherever is the intersection of B and C there against occurs a split and the data is splitted with respect to the factor variables A B C D E F G and H.



(Refer Slide Time: 25:49)



**Data Frames**

The objects `splitted$A` to `splitted$H` are themselves data frames:

```
> is.data.frame(splitted$A)
[1] TRUE
```

*R Console*

```
> is.data.frame(splitted$A)
[1] TRUE
```

17

So, now, let us come back to our slides and here you can see these are the splitted commands that we did earlier right. And one thing what you have to notice here that is very important that after you are splitting a data frame several splits occur depending on the factor variable these splitted data sets are itself data frames. So, when we are trying to split a data frame the splits are also the data frame that you have to keep in mind and this can be checked by using the command `is.data.frame`.

For example, here suppose I try to say here is split it with respect to factor A and another variable is split dollar H with respect to factor H and I am trying to see all the variables all the splits which have occurred with respect to A B C D E F G H they are itself the data frames. And in order to check it I will simply try to write `is.data.frame` and inside this I will try to write down here the variable and this comes out to here true; that means, it is really a data frame and this is here the screenshot of the outcome right.

So, now, with this description I would like to stop with the details on data frame and I have given you some details about the topic on data frames, now I would request you once again to do some practice take some question try to solve them and we will see in that next lecture, till then good bye.