Introduction to R Software

Swayam Prabha

Lecture 23

Data Frames: Creation and Operations

Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

Slides can be downloaded from

http://home.iitk.ac.in/~shalab/sp

Data Frames

❑ Creating Data Frames

Use the `data.frame` function to create a data frame by adding column vectors to the data frame.

Example:

```
> x  <- 101:116                          # Vector
> y  <- matrix(x, nrow=4, ncol=4)        # 4 X 4 matrix
> z  <- letters[1:16]                    # lowercase alphabets
> x
[1] 101 102 103 104 105 106 107 108 109 110 111 112 113
14] 114 115 116
> y
     [,1] [,2] [,3] [,4]
[1,]  101  105  109  113
[2,]  102  106  110  114
[3,]  103  107  111  115
[4,]  104  108  112  116
> z
[1] "a" "b" "c" "d" "e" "f" "g" "h" "i" "j" "k"  "l" "m"
"n" "o" "p"
```

Data Frames

R Console

```
> x  <- 101:116
> y  <- matrix(x, nrow=4, ncol=4)
> z  <- letters[1:16]
>
> x
 [1] 101 102 103 104 105 106 107 108 109 110 111 112 113
[14] 114 115 116
>
> y
     [,1] [,2] [,3] [,4]
[1,]  101  105  109  113
[2,]  102  106  110  114
[3,]  103  107  111  115
[4,]  104  108  112  116
>
> z
 [1] "a" "b" "c" "d" "e" "f" "g" "h" "i" "j" "k" "l" "m"
[14] "n" "o" "p"
>
```

Data Frames

```
> datafr  <- data.frame(x, y, z)
> datafr
    x   X1  X2  X3  X4 z
1  101 101 105 109 113 a
2  102 102 106 110 114 b
3  103 103 107 111 115 c
4  104 104 108 112 116 d
5  105 101 105 109 113 e
6  106 102 106 110 114 f
7  107 103 107 111 115 g
8  108 104 108 112 116 h
9  109 101 105 109 113 i
10 110 102 106 110 114 j
11 111 103 107 111 115 k
12 112 104 108 112 116 l
13 113 101 105 109 113 m
14 114 102 106 110 114 n
15 115 103 107 111 115 o
16 116 104 108 112 116 p
```

Data Frames

R Console

```
> datafr  <- data.frame(x, y, z)
> datafr
     x  X1  X2  X3  X4 z
1  101 101 105 109 113 a
2  102 102 106 110 114 b
3  103 103 107 111 115 c
4  104 104 108 112 116 d
5  105 101 105 109 113 e
6  106 102 106 110 114 f
7  107 103 107 111 115 g
8  108 104 108 112 116 h
9  109 101 105 109 113 i
10 110 102 106 110 114 j
11 111 103 107 111 115 k
12 112 104 108 112 116 l
13 113 101 105 109 113 m
14 114 102 106 110 114 n
15 115 103 107 111 115 o
16 116 104 108 112 116 p
>
```

Data Frames

Consider the data frame `painters` which 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.

Data Frames

R Console

```
> 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
```

Data Frames

- **Structure of the data:**

Display information about the structure of the data frame (`str`).

The result of `str` gives the dimension as well as the name and type of each variable.

```
> str(painters)
'data.frame'  :   54 obs. of  5 variables:
 $ Composition: int  10 15 8 12 0 15 8 15 4 17 ...
 $ Drawing    : int  8 16 13 16 15 16 17 16 12 18 ...
 $ Colour     : int  16 4 16 9 8 4 4 7 10 12 ...
 $ Expression : int  3 14 7 8 0 14 8 6 4 18 ...
 $ School     : Factor w/ 8 levels "A","B","C","D",..: 1
                                  1 1 1 1 1 1 1 1 1 ...
```

`int` means integer.

Data Frames

R Console

```
> str(painters)
'data.frame':   54 obs. of  5 variables:
 $ Composition: int  10 15 8 12 0 15 8 15 4 17 ...
 $ Drawing    : int  8 16 13 16 15 16 17 16 12 18 ...
 $ Colour     : int  16 4 16 9 8 4 4 7 10 12 ...
 $ Expression : int  3 14 7 8 0 14 8 6 4 18 ...
 $ School     : Factor w/ 8 levels "A","B","C","D",..: 1 1 1 1 1 1 1 1 1 1 ...
```

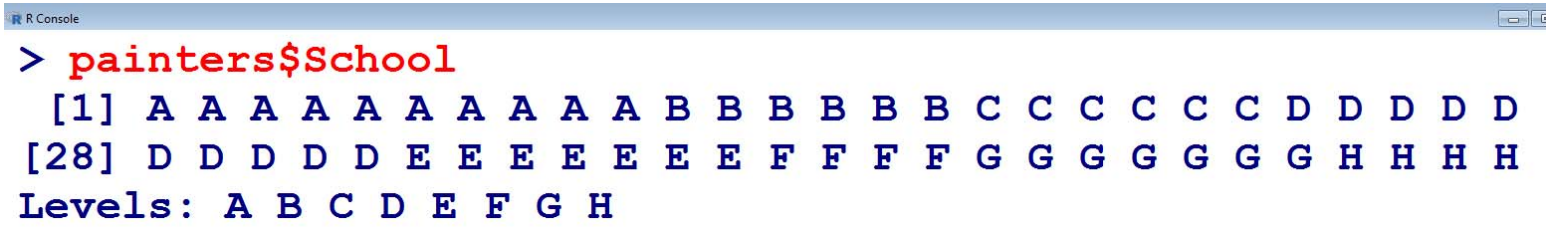
Data Frames

- Extract a variable from data frame using **$**

Variables can be extracted using the **$** operator followed by the name of the variable.

Example: Suppose we want to extract information on variable **School** from the data set **painters.**

```
painters$School
 [1] A A A A A A A A A A B B B B B B C C C C C C D D D D D
[28] D D D D D E E E E E E E F F F F G G G G G G G H H H H
Levels: A B C D E F G H
```



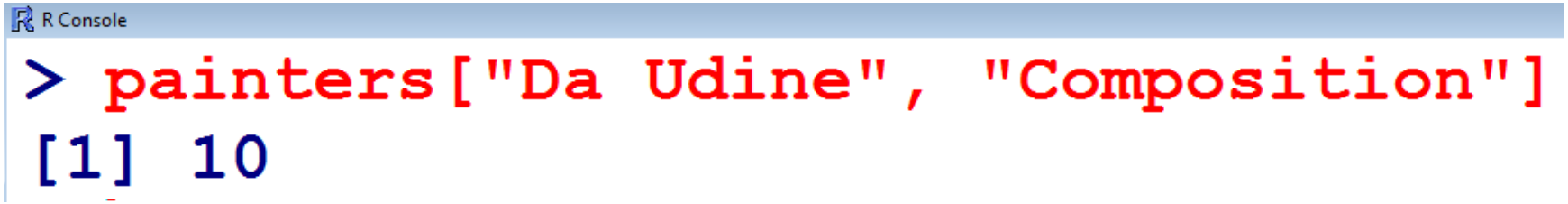
Data Frames

❑ **Extract data from a data frame**

The data from a data frame can be extracted by using the matrix-style `[row, column]` indexing.

Example: Suppose we want to extract information on the first painter `Da Udine` on the variable `Composition` from the data set `painters`.

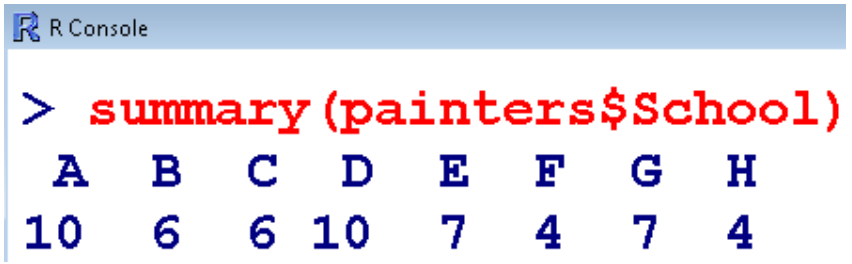
```
> painters["Da Udine", "Composition"]
[1] 10
```



Data Frames

The `summary` function for a categorical variable returns a detailed frequency table:

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



We will learn later:

`summary` is a generic function used to produce result summaries of the results of various model fitting functions.

Data Frames

Using the `summary` function, we can get a quick overview of descriptive measures for each variable: (*We will learn later*).

```
> summary(painters)
  Composition       Drawing         Colour       Expression        School
 Min.   : 0.00   Min.   : 6.00   Min.   : 0.00   Min.   : 0.000   A      :10
 1st Qu.: 8.25   1st Qu.:10.00   1st Qu.: 7.25   1st Qu.: 4.000   D      :10
 Median :12.50   Median :13.50   Median :10.00   Median : 6.000   E      : 7
 Mean   :11.56   Mean   :12.46   Mean   :10.94   Mean   : 7.667   G      : 7
 3rd Qu.:15.00   3rd Qu.:15.00   3rd Qu.:16.00   3rd Qu.:11.500   B      : 6
 Max.   :18.00   Max.   :18.00   Max.   :18.00   Max.   :18.000   C      : 6
                                                                  (Other): 8
```

The categories F and H, each present 4 times in the variable "`School`", are summed under the category `Other` as `8` with the corresponding frequency. i.e., only the 6 most frequent values are displayed.

Data Frames

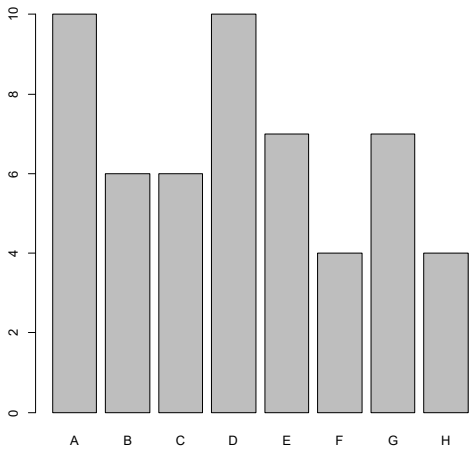
R Console

```
> summary(painters)
  Composition       Drawing          Colour        Expression     School
 Min.   : 0.00   Min.   : 6.00   Min.   : 0.00   Min.   : 0.000   A      :10
 1st Qu.: 8.25   1st Qu.:10.00   1st Qu.: 7.25   1st Qu.: 4.000   D      :10
 Median :12.50   Median :13.50   Median :10.00   Median : 6.000   E      : 7
 Mean   :11.56   Mean   :12.46   Mean   :10.94   Mean   : 7.667   G      : 7
 3rd Qu.:15.00   3rd Qu.:15.00   3rd Qu.:16.00   3rd Qu.:11.500   B      : 6
 Max.   :18.00   Max.   :18.00   Max.   :18.00   Max.   :18.000   C      : 6
                                                                  (Other): 8
```

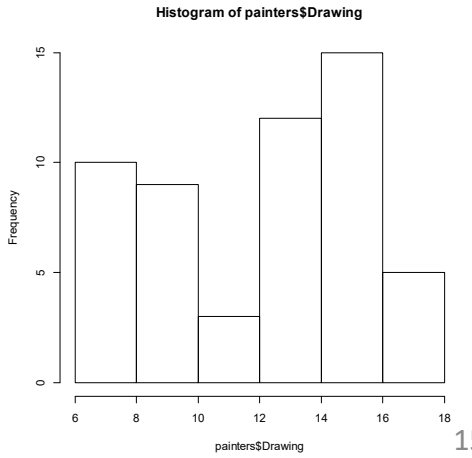
Data Frames

- Plot and graphics of the data

```
> plot(painters$School) #factor variable
```




```
> hist(painters$Drawing) #numeric variable
```



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$`.

Data Frames

- Attaching a data frame

Example

```
> attach(painters)
```

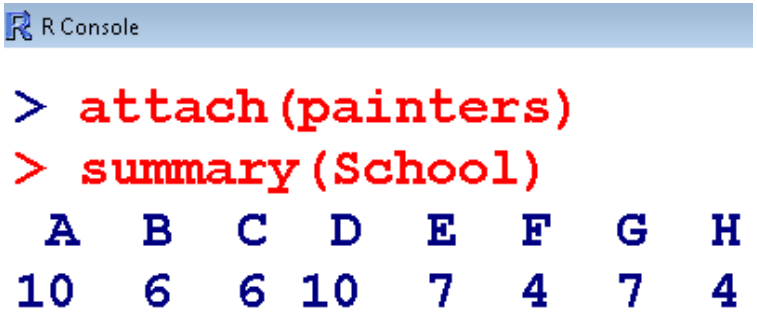
Variable names are

- `Composition,`
- `Drawing,`
- `Colour,`
- `Expression,`
- `School`

Data Frames

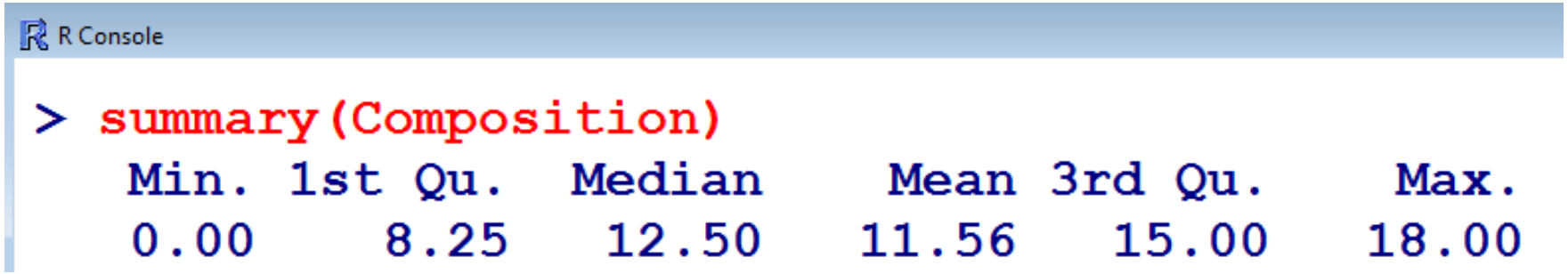
```
> summary(School) # Character variable
```

```
  A   B   C   D   E   F   G   H
 10   6   6 10   7   4   7   4
```



Data Frames

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



Data Frames

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

```
> detach(painters)

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