Introduction to R Software

Data Frames

Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology Kanpur

Data Frames

The commands `c`, `cbind`, `vector` and `matrix` functions combine data.

Another option is the data frame.

In a data frame, we can combine variables of equal length, with each row in the data frame containing observations on the same unit.

Hence, it is similar to the `matrix` or `cbind` functions.

Advantage is that one can make changes to the data without affecting the original data.

Data Frames

One can also combine numerical variables, character strings as well as factors in data frame.

For example, `cbind` and `matrix` functions <u>can not be used</u> to combine different types of data

Data frames are special types of objects in R designed for data sets.

The data frame format is similar to a spreadsheet, where columns contain variables and observations are contained in rows.

Data Frames

Data frames contain complete data sets that are mostly created with other programs (spreadsheet-files, software SPSS-files, Excel-files etc.).

Variables in a data frame may be numeric (numbers) or categorical (characters or factors).

Data Frames

Example:

Package "MASS" describes functions and datasets to support Venables and Ripley, ``Modern Applied Statistics with S'' (4th edition 2002)

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.

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

However, these names are not variables of the data set. Here a subset of these names:

```
> rownames(painters)
 [1] "Da Udine"        "Da Vinci"        "Del Piombo"
 [4] "Del Sarto"       "Fr. Penni"       "Guilio Romano"
 [7] "Michelangelo"    "Perino del Vaga" "Perugino"
[10] "Raphael"         "F. Zucarro"      "Fr. Salviata"
[13] "Parmigiano"      "Primaticcio"     "T. Zucarro"
[16] "Volterra"        "Barocci"         "Cortona"
[19] "Josepin"         "L. Jordaens"     "Testa"
[22] "Vanius"          "Bassano"         "Bellini"
[25] "Giorgione"       "Murillo"         "Palma Giovane"
[28] "Palma Vecchio"   "Pordenone"       "Tintoretto"
[31] "Titian"          "Veronese"        "Albani"
[34] "Caravaggio"      "Corregio"        "Domenichino"
[37] "Guercino"        "Lanfranco"       "The Carraci"
[40] "Durer"           "Holbein"         "Pourbus"
[43] "Van Leyden"      "Diepenbeck"      "J. Jordaens"
[46] "Otho Venius"     "Rembrandt"       "Rubens"
[49] "Teniers"         "Van Dyck"        "Bourdon"
```

Data Frames

R Console

```
> rownames(painters)
 [1] "Da Udine"        "Da Vinci"        "Del Piombo"
 [4] "Del Sarto"       "Fr. Penni"       "Guilio Romano"
 [7] "Michelangelo"    "Perino del Vaga" "Perugino"
[10] "Raphael"         "F. Zucarro"      "Fr. Salviata"
[13] "Parmigiano"      "Primaticcio"     "T. Zucarro"
[16] "Volterra"        "Barocci"         "Cortona"
[19] "Josepin"         "L. Jordaens"     "Testa"
[22] "Vanius"          "Bassano"         "Bellini"
[25] "Giorgione"       "Murillo"         "Palma Giovane"
[28] "Palma Vecchio"   "Pordenone"       "Tintoretto"
[31] "Titian"          "Veronese"        "Albani"
[34] "Caravaggio"      "Corregio"        "Domenichino"
[37] "Guercino"        "Lanfranco"       "The Carraci"
[40] "Durer"           "Holbein"         "Pourbus"
[43] "Van Leyden"      "Diepenbeck"      "J. Jordaens"
[46] "Otho Venius"     "Rembrandt"       "Rubens"
[49] "Teniers"         "Van Dyck"        "Bourdon"
```

Data Frames

- The data set contains four numerical variables (Composition, Drawing, Colour and Expression), as well as one factor variable (School).

```
> is.numeric(painters$School)
[1] FALSE
```

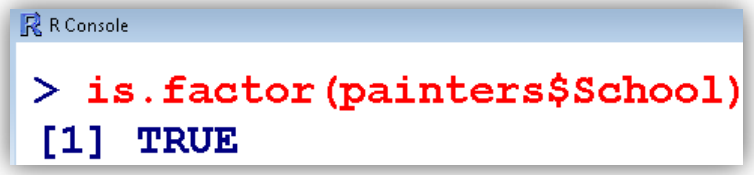
Notice how we extract a variable (column) from data set.

```
> is.numeric(painters$Drawing)
[1] TRUE
```

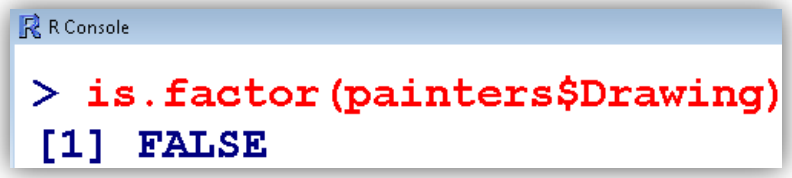
Data Frames

- The data set contains four numerical variables (Composition, Drawing, Colour and Expression), as well as one factor variable (School).

```
> is.factor(painters$School)
[1] TRUE
```




```
> is.factor(painters$Drawing)
[1] FALSE
```



Data Frames

```
> colnames(painters)
[1] "Composition" "Drawing" "Colour"
"Expression" "School"
```

R Console

```
> colnames(painters)
[1] "Composition" "Drawing"     "Colour"      "Expression"  "School"
```

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
```