Introduction to R Software

Swayam Prabha

Lecture 24

More Operations on Data Frames

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

Consider the data frame `painters` 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

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

```
> 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

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
```

R Console

```
> 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
```

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
> _
```

Data Frames

- Uninteresting columns can be eliminated.

```
> 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
```

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

Data Frames

R Console

```
> 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
>
```

Data Frames

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

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

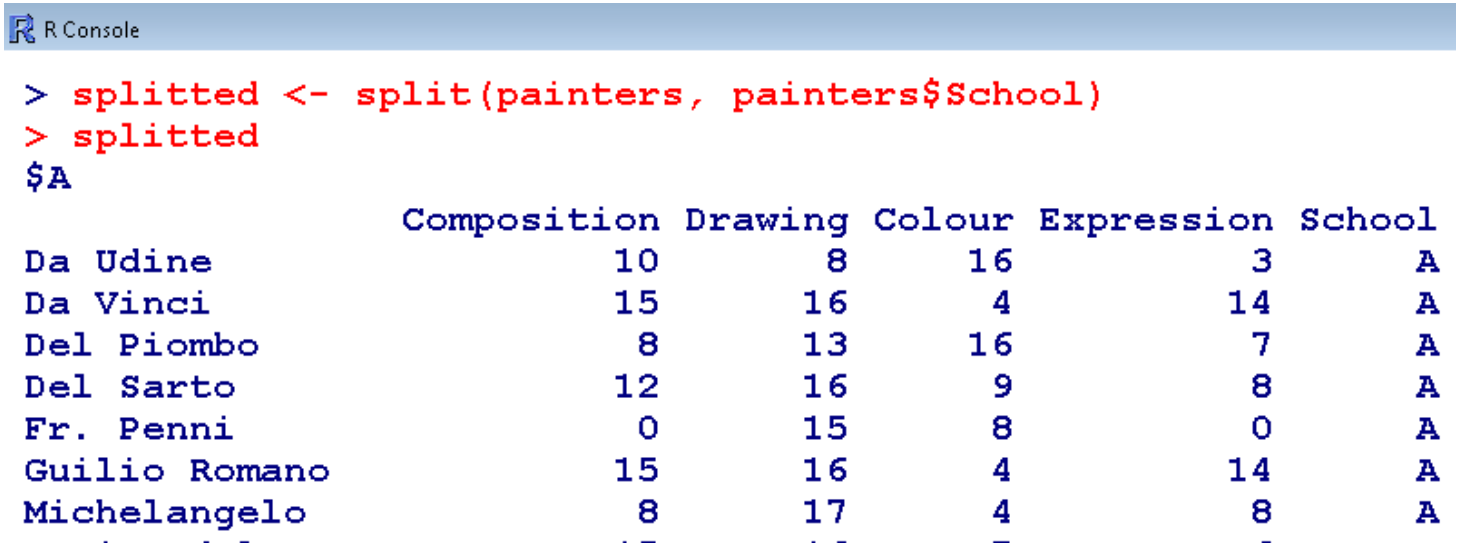
```
> splitted <- split(painters, painters$School)
```

Data Frames

```
> 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…



Data Frames

$B

	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

Contd…

R Console

```
> splitted $B
```

	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

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…

R Console

```
> 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

Data Frames

Contd…

```
. . .
$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
```

R Console

```
> splitted $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
```

Remark: If the data set is not attached, we have to use

`painters$School`.

Data Frames

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

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

Data Frames: Combining

- There are three main techniques :

`cbind()` – combining the columns of two data frames side-by-side.

`merge()` – joining two data frames using a common column.

`rbind()` – stacking two data frames on top of each other, appending one to the other.

Data Frames: Combining

- The command `cbind` horizontally merges two data frames side by side.

Example: Create two data frames as follows:

```
df1=data.frame(state=c("UP", "MP", "AP", "JK"),
  popnsize=c(1000,2000,3000,4000))

df2=data.frame(state=c("UP", "MP", "AP", "JK"),
  samplesize=c(100,200,300,400),
  surveycompleted=c("Yes", "No", "Yes", "No"))
```

Data Frames: Combining

```
> df1
  state popnsize
1    UP     1000
2    MP     2000
3    AP     3000
4    JK     4000
```

```
> df2
  state samplesize surveycompleted
1    UP        100             Yes
2    MP        200              No
3    AP        300             Yes
4    JK        400              No
```

Data Frames: Combining

```
> cbind(df1,df2)
  state popnsize state samplesize surveycompleted
1    UP     1000    UP        100             Yes
2    MP     2000    MP        200              No
3    AP     3000    AP        300             Yes
4    JK     4000    JK        400              No
```

Data Frames: Combining

R Console

```
> df1
  state popnsize
1    UP     1000
2    MP     2000
3    AP     3000
4    JK     4000
> df2
  state samplesize surveycompleted
1    UP        100             Yes
2    MP        200              No
3    AP        300             Yes
4    JK        400              No
> cbind(df1,df2)
  state popnsize state samplesize surveycompleted
1    UP     1000    UP        100             Yes
2    MP     2000    MP        200              No
3    AP     3000    AP        300             Yes
4    JK     4000    JK        400              No
>
```

Data Frames: Merging

- The command `merge` horizontally merges two data frames by common columns or row names.

Example: Create two data frames as follows:

```
df1=data.frame(state=c("UP", "MP", "AP", "JK"),
  popnsize=c(1000,2000,3000,4000))

df2=data.frame(state=c("UP", "MP", "AP", "JK"),
  samplesize=c(100,200,300,400),
  surveycompleted=c("Yes", "No", "Yes", "No"))
```

Variable "`state`" is common between the two data frames and we want to merge the two data frames with respect to `state`.

Data Frames: Merging

- The command `merge` horizontally merges two data frames by common columns or row names.

Usage : `merge(x, y, ...)`

Arguments :

`x, y` : data frames, or objects to be coerced to one.

`by, by.x, by.y` : specifications of the columns used for merging.

`sort` : logical.

Data Frames: Merging

Arguments :

`sort` **: logical.**

`suffixes` **: a character vector of length 2 specifying the suffixes to be used for making unique the names of columns in the result which are not used for merging (appearing in by etc).**

`no.dups` **: logical indicating that suffixes are appended in more cases to avoid duplicated column names in the result.**

Data Frames: Merging

```
> df1
  state popnsize
1    UP     1000
2    MP     2000
3    AP     3000
4    JK     4000

> df2
  state samplesize surveycompleted
1    UP        100             Yes
2    MP        200              No
3    AP        300             Yes
4    JK        400              No

> merge(df1,df2,by="state")
  state popnsize samplesize surveycompleted
1    AP     3000        300             Yes
2    JK     4000        400              No
3    MP     2000        200              No
4    UP     1000        100             Yes
```

Data Frames: Merging

R Console

```
> df1
  state popnsize
1    UP     1000
2    MP     2000
3    AP     3000
4    JK     4000
>
> df2
  state samplesize surveycompleted
1    UP        100              Yes
2    MP        200               No
3    AP        300              Yes
4    JK        400               No
> merge(df1,df2,by="state")
  state popnsize samplesize surveycompleted
1    AP     3000        300              Yes
2    JK     4000        400               No
3    MP     2000        200               No
4    UP     1000        100              Yes
>
```

Data Frames: Combining vertically

- The command `rbind` stacks two data frames on top of each other, appending one to the other

Example: Create two data frames as follows:

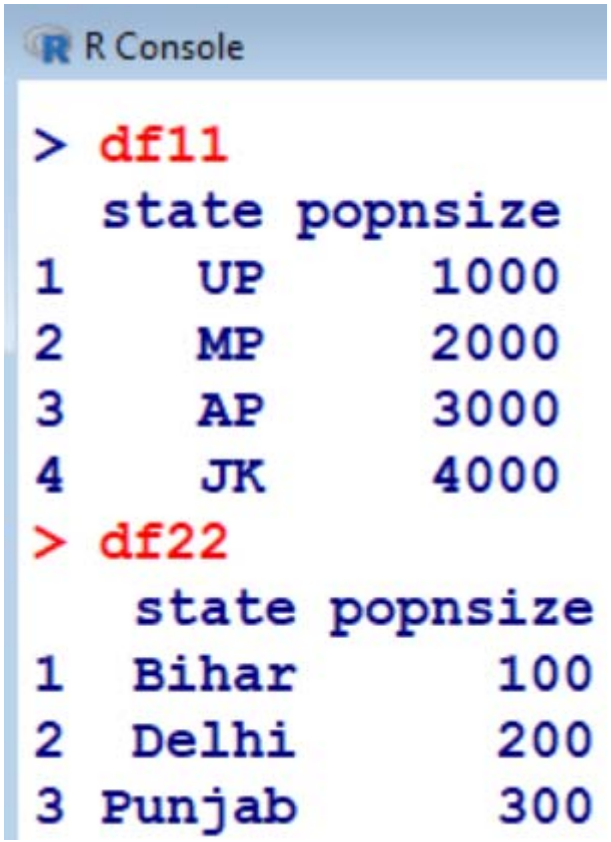
```
df11=data.frame(state=c("UP", "MP", "AP",
  "JK"), popnsize=c(1000,2000,3000,4000))

df22=data.frame(state=c("Bihar", "Delhi",
  "Punjab"), popnsize =c(100,200,300))
```

Data Frames: Combining vertically

```
> df11
  state popnsize
1    UP      1000
2    MP      2000
3    AP      3000
4    JK      4000

> df22
   state popnsize
1  Bihar      100
2  Delhi      200
3  Punjab      300
```



Data Frames: Combining vertically

```
> rbind(df11,df22)
   state popnsize
1     UP     1000
2     MP     2000
3     AP     3000
4     JK     4000
5  Bihar      100
6  Delhi      200
7 Punjab      300
```