Data Frame

Data Frames

Un **Data Frame** es una tabla de doble entrada, formada por variables en las columnas y observaciones de esas variables en las filas, de manera que cada fila contiene los valores de las variables para un mismo caso o un mismo individuo.

- **Data()**: para abrir una ventana con la lista de los objetos de datos a los que tenemos acceso en la sesión actual de `R`(los que lleva la instalación básica de `R` y los que aportan los paquetes que tengamos cargados).
  - Si entramos **data(package = .packages(all.available = TRUE))** obtendremos la lista de todos los objetos de datos a los que tenemos acceso, incluyendo los de los paquetes que tengamos instalados, pero que no estén cargados en la sesión actual.

Obtiendo información del data frame

- **head(DataFrame, n)**: para mostrar las n primeras filas del data frame. Por defecto se muestran las 6 primeras filas.
- **tail(DataFrame, n)**: para mostrar las n últimas filas del data frame. Por defecto se muestran las 6 últimas filas.
- **str(DataFrame)**: para conocer la estructura global de un data frame.
- **names(DataFrame)**: para producir un vector con los nombres de las columnas.

Data Frame de Iris

Añadir una base de datos de iris

```
df <- iris #Cargando los datos de Iris a la variable df
head(df, 5)#Leer df y sólo muestra los 5 primeros datos
```

```
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
```

```
tail(df, 5)#Mostrar los 5 últimos datos de df
```

```
    Sepal.Length Sepal.Width Petal.Length Petal.Width   Species
146          6.7         3.0          5.2         2.3 virginica
147          6.3         2.5          5.0         1.9 virginica
148          6.5         3.0          5.2         2.0 virginica
149          6.2         3.4          5.4         2.3 virginica
150          5.9         3.0          5.1         1.8 virginica
```

Para consultar la información básica del Data Frame **df**, es decir, sólo muestra el nombre de las columnas de la tabla.

```r
names(df)#Para ver los nombres de las variables o columnas del Data Frame
```

```
[1] "Sepal.Length" "Sepal.Width"  "Petal.Length" "Petal.Width"  "Species"
```

Para ver su estructura:

```r
str(df)#Para ver la estructura del Data Frame
```

```
'data.frame':   150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

Como se puede observar, la estructura tiene 3 vectores numéricos y un factor de 3 niveles.

Estructura y filtrado de data frame

- **rownames(DataFrame)**: para producir un vector con los identificadores de las filas.
  - Rentiende siempre que estos identificadores son palabras, aunque sean números, de ahí que los imprima entre comillas.
- **colnames(DataFrame)**: para producir un vector con los identificadores de las columnas.
- **dimnames(DataFrame)**: para producir una lista formada por dos vectores(el de los identificadores de las filas y el de los nombres de las columnas).
- **nrow(DataFrame)**: para consultar el número de filas de una data frame.
- **ncol(DataFrame)**: para consultar el número de columnas de una data frame.
- **dim(DataFrame)**: para producir un vector con el número de filas y el de columnas de una data frame.
- **DataFrame.$nombre_variable**: para obtener una columnas concreta de un data frame
  - El resultado será un vector o un factor, según cómo esté definida la columna dentro del data frame.
  - Las variables de un data frame son internas, no están definidas en el entorno global de trabajo de R.

Ejemplos:

```r
#Nomes de la estructura de las filas
rownames(df)
```

```
  [1] "1"   "2"   "3"   "4"   "5"   "6"   "7"   "8"   "9"   "10"  "11"  "12"
 [13] "13"  "14"  "15"  "16"  "17"  "18"  "19"  "20"  "21"  "22"  "23"  "24"
 [25] "25"  "26"  "27"  "28"  "29"  "30"  "31"  "32"  "33"  "34"  "35"  "36"
 [37] "37"  "38"  "39"  "40"  "41"  "42"  "43"  "44"  "45"  "46"  "47"  "48"
 [49] "49"  "50"  "51"  "52"  "53"  "54"  "55"  "56"  "57"  "58"  "59"  "60"
 [61] "61"  "62"  "63"  "64"  "65"  "66"  "67"  "68"  "69"  "70"  "71"  "72"
 [73] "73"  "74"  "75"  "76"  "77"  "78"  "79"  "80"  "81"  "82"  "83"  "84"
 [85] "85"  "86"  "87"  "88"  "89"  "90"  "91"  "92"  "93"  "94"  "95"  "96"
 [97] "97"  "98"  "99"  "100" "101" "102" "103" "104" "105" "106" "107" "108"
[109] "109" "110" "111" "112" "113" "114" "115" "116" "117" "118" "119" "120"
[121] "121" "122" "123" "124" "125" "126" "127" "128" "129" "130" "131" "132"
[133] "133" "134" "135" "136" "137" "138" "139" "140" "141" "142" "143" "144"
[145] "145" "146" "147" "148" "149" "150"
```

Los datos que aparece el nombre de los identificadores de las filas. Esos identificadores son lo que aparece a izquierda del data frame

```
head(df, 10)
```

```
   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1           5.1         3.5          1.4         0.2  setosa
2           4.9         3.0          1.4         0.2  setosa
3           4.7         3.2          1.3         0.2  setosa
4           4.6         3.1          1.5         0.2  setosa
5           5.0         3.6          1.4         0.2  setosa
6           5.4         3.9          1.7         0.4  setosa
7           4.6         3.4          1.4         0.3  setosa
8           5.0         3.4          1.5         0.2  setosa
9           4.4         2.9          1.4         0.2  setosa
10          4.9         3.1          1.5         0.1  setosa
```

```
#Nombres de la estructura de las columnas
colnames(df)
```

```
[1] "Sepal.Length" "Sepal.Width"  "Petal.Length" "Petal.Width"  "Species"
```

Los datos que aparece el nombre de los identificadores de las columnas. Esos identificadores son lo que aparece a la parte de arriba del data frame

Para ver que los datos de nombres de fila o columna siempre devuelve un vector, basta pedir la información completa como en el ejemplo a continuación:

```
dimnames(df)
```

```
[[1]]
  [1] "1"   "2"   "3"   "4"   "5"   "6"   "7"   "8"   "9"   "10"  "11"  "12"
 [13] "13"  "14"  "15"  "16"  "17"  "18"  "19"  "20"  "21"  "22"  "23"  "24"
 [25] "25"  "26"  "27"  "28"  "29"  "30"  "31"  "32"  "33"  "34"  "35"  "36"
 [37] "37"  "38"  "39"  "40"  "41"  "42"  "43"  "44"  "45"  "46"  "47"  "48"
 [49] "49"  "50"  "51"  "52"  "53"  "54"  "55"  "56"  "57"  "58"  "59"  "60"
 [61] "61"  "62"  "63"  "64"  "65"  "66"  "67"  "68"  "69"  "70"  "71"  "72"
 [73] "73"  "74"  "75"  "76"  "77"  "78"  "79"  "80"  "81"  "82"  "83"  "84"
 [85] "85"  "86"  "87"  "88"  "89"  "90"  "91"  "92"  "93"  "94"  "95"  "96"
 [97] "97"  "98"  "99"  "100" "101" "102" "103" "104" "105" "106" "107" "108"
[109] "109" "110" "111" "112" "113" "114" "115" "116" "117" "118" "119" "120"
[121] "121" "122" "123" "124" "125" "126" "127" "128" "129" "130" "131" "132"
[133] "133" "134" "135" "136" "137" "138" "139" "140" "141" "142" "143" "144"
[145] "145" "146" "147" "148" "149" "150"

[[2]]
[1] "Sepal.Length" "Sepal.Width"  "Petal.Length" "Petal.Width"  "Species"
```

Como se aprecia, es un vector bidimensional (matriz), donde una dimensión son los identificadores de las filas y la otra dimensión son los identificadores de las columnas.

Para saber el número de filas:

```
nrow(df)
```

```
[1] 150
```

Para saber el número de columnas:

```r
nrow(df)
```

```
[1] 150
```

Para saber el número de filas y columnas en una única consulata:

```r
dim(df)
```

```
[1] 150   5
```

Como se observa, devuelve un vector con dos posiciones, donde la posición cero indica la longitud de las filas y en la posición uno indica la longitud de columnas.

Para poder acceder a una componente en concreto del data frame. El resultado puede ser un vector o un factor, eso va depender del tipo de dato que esté almacenado.

```r
#Acceder todos los datos de variable de nombre Sepal
df$Sepal.Length
```

```
  [1] 5.1 4.9 4.7 4.6 5.0 5.4 4.6 5.0 4.4 4.9 5.4 4.8 4.8 4.3 5.8 5.7 5.4 5.1
 [19] 5.7 5.1 5.4 5.1 4.6 5.1 4.8 5.0 5.0 5.2 5.2 4.7 4.8 5.4 5.2 5.5 4.9 5.0
 [37] 5.5 4.9 4.4 5.1 5.0 4.5 4.4 5.0 5.1 4.8 5.1 4.6 5.3 5.0 7.0 6.4 6.9 5.5
 [55] 6.5 5.7 6.3 4.9 6.6 5.2 5.0 5.9 6.0 6.1 5.6 6.7 5.6 5.8 6.2 5.6 5.9 6.1
 [73] 6.3 6.1 6.4 6.6 6.8 6.7 6.0 5.7 5.5 5.5 5.8 6.0 5.4 6.0 6.7 6.3 5.6 5.5
 [91] 5.5 6.1 5.8 5.0 5.6 5.7 5.7 6.2 5.1 5.7 6.3 5.8 7.1 6.3 6.5 7.6 4.9 7.3
[109] 6.7 7.2 6.5 6.4 6.8 5.7 5.8 6.4 6.5 7.7 7.7 6.0 6.9 5.6 7.7 6.3 6.7 7.2
[127] 6.2 6.1 6.4 7.2 7.4 7.9 6.4 6.3 6.1 7.7 6.3 6.4 6.0 6.9 6.7 6.9 5.8 6.8
[145] 6.7 6.7 6.3 6.5 6.2 5.9
```

Se puede acceder solamente a una cantidad especificada

```r
df$Sepal.Length[1:10]
```

```
##  [1] 5.1 4.9 4.7 4.6 5.0 5.4 4.6 5.0 4.4 4.9
```

Sub-data frames

- **DataFrame[n, m]**: para extraer "trozos" del data frame por filas y columnas(funciona exactamente igual que en matrices) donde n y m pueden definirse como:
  - intervalos
  - condiciones
  - números naturales
  - no poner nada
  - Si sólo queremos definir la subtabla quédandonos con algunas variables, basta aplicar el nombre del data frame al vector de variables.
  - Estas construcciones se pueden usar tambíen para reordenar las filas o columnas.

De esa forma se puede acceder a los datos del data frame de forma personalizada.

Ejemplos:

```r
#Datos de las 10 primeras filas
df[1:10, ]
```

```
   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1           5.1         3.5          1.4         0.2  setosa
2           4.9         3.0          1.4         0.2  setosa
```

```
3           4.7         3.2          1.3         0.2  setosa
4           4.6         3.1          1.5         0.2  setosa
5           5.0         3.6          1.4         0.2  setosa
6           5.4         3.9          1.7         0.4  setosa
7           4.6         3.4          1.4         0.3  setosa
8           5.0         3.4          1.5         0.2  setosa
9           4.4         2.9          1.4         0.2  setosa
10          4.9         3.1          1.5         0.1  setosa
```

```r
#Datos de las 3 primeras columnas
df[, 1:3]
```

```
    Sepal.Length Sepal.Width Petal.Length
1            5.1         3.5          1.4
2            4.9         3.0          1.4
3            4.7         3.2          1.3
4            4.6         3.1          1.5
5            5.0         3.6          1.4
6            5.4         3.9          1.7
7            4.6         3.4          1.4
8            5.0         3.4          1.5
9            4.4         2.9          1.4
10           4.9         3.1          1.5
11           5.4         3.7          1.5
12           4.8         3.4          1.6
13           4.8         3.0          1.4
14           4.3         3.0          1.1
15           5.8         4.0          1.2
16           5.7         4.4          1.5
17           5.4         3.9          1.3
18           5.1         3.5          1.4
19           5.7         3.8          1.7
20           5.1         3.8          1.5
21           5.4         3.4          1.7
22           5.1         3.7          1.5
23           4.6         3.6          1.0
24           5.1         3.3          1.7
25           4.8         3.4          1.9
26           5.0         3.0          1.6
27           5.0         3.4          1.6
28           5.2         3.5          1.5
29           5.2         3.4          1.4
30           4.7         3.2          1.6
31           4.8         3.1          1.6
32           5.4         3.4          1.5
33           5.2         4.1          1.5
34           5.5         4.2          1.4
35           4.9         3.1          1.5
36           5.0         3.2          1.2
37           5.5         3.5          1.3
38           4.9         3.6          1.4
39           4.4         3.0          1.3
40           5.1         3.4          1.5
41           5.0         3.5          1.3
42           4.5         2.3          1.3
```

43	4.4	3.2	1.3
44	5.0	3.5	1.6
45	5.1	3.8	1.9
46	4.8	3.0	1.4
47	5.1	3.8	1.6
48	4.6	3.2	1.4
49	5.3	3.7	1.5
50	5.0	3.3	1.4
51	7.0	3.2	4.7
52	6.4	3.2	4.5
53	6.9	3.1	4.9
54	5.5	2.3	4.0
55	6.5	2.8	4.6
56	5.7	2.8	4.5
57	6.3	3.3	4.7
58	4.9	2.4	3.3
59	6.6	2.9	4.6
60	5.2	2.7	3.9
61	5.0	2.0	3.5
62	5.9	3.0	4.2
63	6.0	2.2	4.0
64	6.1	2.9	4.7
65	5.6	2.9	3.6
66	6.7	3.1	4.4
67	5.6	3.0	4.5
68	5.8	2.7	4.1
69	6.2	2.2	4.5
70	5.6	2.5	3.9
71	5.9	3.2	4.8
72	6.1	2.8	4.0
73	6.3	2.5	4.9
74	6.1	2.8	4.7
75	6.4	2.9	4.3
76	6.6	3.0	4.4
77	6.8	2.8	4.8
78	6.7	3.0	5.0
79	6.0	2.9	4.5
80	5.7	2.6	3.5
81	5.5	2.4	3.8
82	5.5	2.4	3.7
83	5.8	2.7	3.9
84	6.0	2.7	5.1
85	5.4	3.0	4.5
86	6.0	3.4	4.5
87	6.7	3.1	4.7
88	6.3	2.3	4.4
89	5.6	3.0	4.1
90	5.5	2.5	4.0
91	5.5	2.6	4.4
92	6.1	3.0	4.6
93	5.8	2.6	4.0
94	5.0	2.3	3.3
95	5.6	2.7	4.2
96	5.7	3.0	4.2

97	5.7	2.9	4.2
98	6.2	2.9	4.3
99	5.1	2.5	3.0
100	5.7	2.8	4.1
101	6.3	3.3	6.0
102	5.8	2.7	5.1
103	7.1	3.0	5.9
104	6.3	2.9	5.6
105	6.5	3.0	5.8
106	7.6	3.0	6.6
107	4.9	2.5	4.5
108	7.3	2.9	6.3
109	6.7	2.5	5.8
110	7.2	3.6	6.1
111	6.5	3.2	5.1
112	6.4	2.7	5.3
113	6.8	3.0	5.5
114	5.7	2.5	5.0
115	5.8	2.8	5.1
116	6.4	3.2	5.3
117	6.5	3.0	5.5
118	7.7	3.8	6.7
119	7.7	2.6	6.9
120	6.0	2.2	5.0
121	6.9	3.2	5.7
122	5.6	2.8	4.9
123	7.7	2.8	6.7
124	6.3	2.7	4.9
125	6.7	3.3	5.7
126	7.2	3.2	6.0
127	6.2	2.8	4.8
128	6.1	3.0	4.9
129	6.4	2.8	5.6
130	7.2	3.0	5.8
131	7.4	2.8	6.1
132	7.9	3.8	6.4
133	6.4	2.8	5.6
134	6.3	2.8	5.1
135	6.1	2.6	5.6
136	7.7	3.0	6.1
137	6.3	3.4	5.6
138	6.4	3.1	5.5
139	6.0	3.0	4.8
140	6.9	3.1	5.4
141	6.7	3.1	5.6
142	6.9	3.1	5.1
143	5.8	2.7	5.1
144	6.8	3.2	5.9
145	6.7	3.3	5.7
146	6.7	3.0	5.2
147	6.3	2.5	5.0
148	6.5	3.0	5.2
149	6.2	3.4	5.4
150	5.9	3.0	5.1

```r
#Datos de las 10 primeras filas y columnas de 2 a 4
df[1:10, 2:4]
```

```
   Sepal.Width Petal.Length Petal.Width
1          3.5          1.4         0.2
2          3.0          1.4         0.2
3          3.2          1.3         0.2
4          3.1          1.5         0.2
5          3.6          1.4         0.2
6          3.9          1.7         0.4
7          3.4          1.4         0.3
8          3.4          1.5         0.2
9          2.9          1.4         0.2
10         3.1          1.5         0.1
```

```r
#Datos cumpliendo una expresión booleana
df[df$Species == "setosa" & df$Sepal.Width > 4, ]
```

```
   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
16          5.7         4.4          1.5         0.4  setosa
33          5.2         4.1          1.5         0.1  setosa
34          5.5         4.2          1.4         0.2  setosa
```

El resultado anterior son los datos de las especies que se llama ***setosa*** con una ***anchura de pétalos*** mayor que cuatro. El resultado sigue siendo un data frame. Con eso se puede hacer consultas con filtros más elaborados, por ejemplo:

```r
#Datos cumpliendo una expresión booleana además de uun filtrado al final
df[df$Species == "setosa" & df$Sepal.Width > 4, ][c(1,3), c(2,5)]
```

```
   Sepal.Width Species
16         4.4  setosa
34         4.2  setosa
```