Tema 5 : Data frames

Alonso Pizarro Lagunas

25/10/2021

Data frames

Un data frame es una tabla de doble entrada en donde cada variable formará parate de una columna y cada fila una observación para cada variable de un individuo.

```
df = iris
head(df,5)
```

```
  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)
```

```
    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
```

```
names(df)
```

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

```
str(df)
```

```
'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 ...
```

Del data frame *Orange* obtenemos

```
df2 = Orange

names(df2)
```

```
[1] "Tree"          "age"           "circumference"
```

```
rownames(df2)
```

```
 [1] "1"  "2"  "3"  "4"  "5"  "6"  "7"  "8"  "9"  "10" "11" "12" "13" "14" "15"
```

```
[16] "16" "17" "18" "19" "20" "21" "22" "23" "24" "25" "26" "27" "28" "29" "30"
[31] "31" "32" "33" "34" "35"
```

```
dimnames(df2)
```

```
[[1]]
 [1] "1"  "2"  "3"  "4"  "5"  "6"  "7"  "8"  "9"  "10" "11" "12" "13" "14" "15"
[16] "16" "17" "18" "19" "20" "21" "22" "23" "24" "25" "26" "27" "28" "29" "30"
[31] "31" "32" "33" "34" "35"

[[2]]
[1] "Tree"          "age"           "circumference"
```

```
str(df2)
```

```
Classes 'nfnGroupedData', 'nfGroupedData', 'groupedData' and 'data.frame':  35 obs. of  3 variables:
 $ Tree         : Ord.factor w/ 5 levels "3"<"1"<"5"<"2"<..: 2 2 2 2 2 2 2 4 4 4 ...
 $ age          : num  118 484 664 1004 1231 ...
 $ circumference: num  30 58 87 115 120 142 145 33 69 111 ...
 - attr(*, "formula")=Class 'formula'  language circumference ~ age | Tree
  .. ..- attr(*, ".Environment")=<environment: R_EmptyEnv>
 - attr(*, "labels")=List of 2
  ..$ x: chr "Time since December 31, 1968"
  ..$ y: chr "Trunk circumference"
 - attr(*, "units")=List of 2
  ..$ x: chr "(days)"
  ..$ y: chr "(mm)"
```

```
head(df2,4)
```

```
  Tree  age circumference
1    1  118            30
2    1  484            58
3    1  664            87
4    1 1004           115
```

```
tail(df2,4)
```

```
   Tree  age circumference
32    5 1004           125
33    5 1231           142
34    5 1372           174
35    5 1582           177
```

```
df2$Tree[1:10]
```

```
 [1] 1 1 1 1 1 1 1 2 2 2
Levels: 3 < 1 < 5 < 2 < 4
```

Acceso al data frame

Iris

```
df = iris

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
```

```
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
```

```
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
```

Orange

```
dataOrange = Orange

dataOrange[c(10:12),]
```

```
   Tree  age circumference
10    2  664           111
11    2 1004           156
12    2 1231           172
```

```
dataOrange[c(2:17), c(1,3)]
```

```
   Tree circumference
2     1            58
3     1            87
4     1           115
5     1           120
6     1           142
7     1           145
8     2            33
9     2            69
10    2           111
11    2           156
12    2           172
13    2           203
```

```
14     2           203
15     3            30
16     3            51
17     3            75
```

```
dataOrange[2,3]
```

```
[1] 58
```

```
knitr::kable(dataOrange[dataOrange$circumference <= 50,], caption = "titulo 1")
```

Table 1: titulo 1

	Tree	age	circumference
1	1	118	30
8	2	118	33
15	3	118	30
22	4	118	32
29	5	118	30
30	5	484	49

Carga de ficheros local

```
df = read.table("Curso de R basic/data/olive.txt", header = TRUE,
                col.names = c("breed", "sale_price", "shoulder",
                              "fat_free", "percent_ff", "frame_scale",
                              "back_fat", "sale_height", "sale_weight"),
                sep = "", dec = ".")
knitr::kable(head(df,10), caption = "titulo 2")
```

Table 2: titulo 2

breed	sale_price	shoulder	fat_free	percent_ff	frame_scale	back_fat	sale_height	sale_weight
1	1075	75	226	7823	672	36	60	29
1	1088	73	224	7709	781	31	61	29
1	911	54	246	8113	549	31	63	29
1	966	57	240	7952	619	50	78	35
1	1051	67	259	7771	672	50	80	46
1	911	49	268	7924	678	51	70	44
1	922	66	264	7990	618	49	56	29
1	1100	61	235	7728	734	39	64	35
1	1082	60	239	7745	709	46	83	33
1	1037	55	213	7944	633	26	52	30

Carga desde URL

```
options(width = 120)
df2 = read.table("https://people.sc.fsu.edu/~jburkardt/data/csv/freshman_kgs.csv",
                 header = TRUE, sep = ",", col.names = c("sex", "Weight_Sep",
                                                         "Weight_Apr", "BMI_Sep",
                                                         "BMI_Apr"), dec = ".")
```

```
knitr::kable(head(df2, 5), caption = "titulo 3")
```

Table 3: titulo 3

sex	Weight_Sep	Weight_Apr	BMI_Sep	BMI_Apr
M	72	59	22.02	18.14
M	97	86	19.70	17.44
M	74	69	24.09	22.43
M	93	88	26.97	25.57
F	68	64	21.51	20.10

```
names(df2)
```

```
## [1] "sex"        "Weight_Sep" "Weight_Apr" "BMI_Sep"    "BMI_Apr"
```

```
str(df2)
```

```
## 'data.frame':    67 obs. of  5 variables:
##  $ sex       : chr  "M" "M" "M" "M" ...
##  $ Weight_Sep: int  72 97 74 93 68 59 64 56 70 58 ...
##  $ Weight_Apr: int  59 86 69 88 64 55 60 53 68 56 ...
##  $ BMI_Sep   : num  22 19.7 24.1 27 21.5 ...
##  $ BMI_Apr   : num  18.1 17.4 22.4 25.6 20.1 ...
```

Factores en un data frame

```
df3 = read.table("https://people.sc.fsu.edu/~jburkardt/data/csv/cities.csv",
                 header = TRUE, sep = ",")

str(df3)
```

```
'data.frame':   128 obs. of  10 variables:
 $ LatD : int  41 42 46 42 43 36 49 39 34 39 ...
 $ LatM : int  5 52 35 16 37 5 52 11 14 45 ...
 $ LatS : int  59 48 59 12 48 59 48 23 24 0 ...
 $ NS   : chr  " N" " N" " N" " N" ...
 $ LonD : int  80 97 120 71 89 80 97 78 77 75 ...
 $ LonM : int  39 23 30 48 46 15 9 9 55 33 ...
 $ LonS : int  0 23 36 0 11 0 0 36 11 0 ...
 $ EW   : chr  " W" " W" " W" " W" ...
 $ City : chr  " Youngstown" " Yankton" " Yakima" " Worcester" ...
 $ State: chr  " OH" " SD" " WA" " MA" ...
```

```
head(df3,18)
```

```
   LatD LatM LatS NS LonD LonM LonS EW            City State
1    41    5   59  N   80   39    0  W      Youngstown    OH
2    42   52   48  N   97   23   23  W         Yankton    SD
3    46   35   59  N  120   30   36  W          Yakima    WA
4    42   16   12  N   71   48    0  W       Worcester    MA
5    43   37   48  N   89   46   11  W Wisconsin Dells    WI
6    36    5   59  N   80   15    0  W   Winston-Salem    NC
7    49   52   48  N   97    9    0  W        Winnipeg    MB
```

```
8    39   11   23  N   78    9   36  W       Winchester    VA
9    34   14   24  N   77   55   11  W       Wilmington    NC
10   39   45    0  N   75   33    0  W       Wilmington    DE
11   48    9    0  N  103   37   12  W        Williston    ND
12   41   15    0  N   77    0    0  W     Williamsport    PA
13   37   40   48  N   82   16   47  W       Williamson    WV
14   33   54    0  N   98   29   23  W    Wichita Falls    TX
15   37   41   23  N   97   20   23  W          Wichita    KS
16   40    4   11  N   80   43   12  W         Wheeling    WV
17   26   43   11  N   80    3    0  W  West Palm Beach    FL
18   47   25   11  N  120   19   11  W        Wenatchee    WA
```

Exportando datos a ficheros

Particularmente usando `write.table(df, file = "")`

```
write.table(df3, file ="Curso de R basic/data/ciudades.txt",
                                                    dec = ".")

df4 = read.table("Curso de R basic/data/ciudades.txt", header = TRUE, dec =".")
head(df4)
```

```
  LatD LatM LatS NS LonD LonM LonS EW             City State
1   41    5   59  N   80   39    0  W       Youngstown    OH
2   42   52   48  N   97   23   23  W          Yankton    SD
3   46   35   59  N  120   30   36  W           Yakima    WA
4   42   16   12  N   71   48    0  W       Worcester    MA
5   43   37   48  N   89   46   11  W  Wisconsin Dells    WI
6   36    5   59  N   80   15    0  W    Winston-Salem    NC
```

Construyendo data frames

Ejemplo 1

```
algebra = c(1,2,0,5,4,6,7,5,5,8)
analysis = c(3,3,2,7,9,5,6,8,5,6)
statistics = c(4,5,4,8,8,9,6,7,9,10)
grades = data.frame(Alg = algebra, An = analysis, Stat = statistics)
str(grades)
```

```
'data.frame':   10 obs. of  3 variables:
 $ Alg : num  1 2 0 5 4 6 7 5 5 8
 $ An  : num  3 3 2 7 9 5 6 8 5 6
 $ Stat: num  4 5 4 8 8 9 6 7 9 10
```

```
calculus = c(5,4,6,2,1,0,7,8,9,6)
grades2 = cbind(grades, calculus)

grades2
```

```
   Alg An Stat calculus
1    1  3    4        5
2    2  3    5        4
3    0  2    4        6
4    5  7    8        2
```

```
5    4  9    8        1
6    6  5    9        0
7    7  6    6        7
8    5  8    7        8
9    5  5    9        9
10   8  6   10        6
```

Ejemplo 2

```
gender = c("H", "M", "M", "M", "H")
age = c(23,45,29,30,18)
family = c(2,3,4,2,5)
df5 = data.frame(genero = gender, edad = age, familia = family, stringsAsFactors = TRUE)

df5
```

```
  genero edad familia
1      H   23       2
2      M   45       3
3      M   29       4
4      M   30       2
5      H   18       5
```

```
row.names(df5) = c("P1","P2","P3","P4","P5")


str(df5)
```

```
'data.frame':   5 obs. of  3 variables:
 $ genero : Factor w/ 2 levels "H","M": 1 2 2 2 1
 $ edad   : num  23 45 29 30 18
 $ familia: num  2 3 4 2 5
```

```
#fix(df5) # fix para editar el data frame al igual que se hacía con vectores

dimnames(df5) = list(
  c("nombre1", "nombre2", "nombre3", "nombre4", "nombre5"),
  c("sexo", "edad", "integrantes")
)

df5 = rbind(df5, c("H",30,1)) # añadir datos de columna
df5
```

```
        sexo edad integrantes
nombre1    H   23           2
nombre2    M   45           3
nombre3    M   29           4
nombre4    M   30           2
nombre5    H   18           5
6          H   30           1
```

```
df5$Ingresos = c(10000, 12000,12000, 13500, 11500, 13000) # otra manera de crear una #columna
df5
```

```
        sexo edad integrantes Ingresos
nombre1    H   23           2    10000
nombre2    M   45           3    12000
nombre3    M   29           4    12000
```

```
nombre4    M   30          2    13500
nombre5    H   18          5    11500
6          H   30          1    13000
```

Cambiando los tipos de datos

Esto recibe el nombre de casting en programación

- as.character
- as.integer
- as.numeric