02-Carga de Datos

Juan Gabriel

Contents

Tipos de ficheros

Losa tipos de ficheros más usados son:

- CSV
- XML
- JSON

Ficheros CSV

Para leer un csv tenemos la función *read.csv*.

```
data <- read.csv("../data/tema1/auto-mpg.csv", header = TRUE, sep = ",")

names(data)#devuelve el nombre de las columnas
```

```
## [1] "No"           "mpg"          "cylinders"    "displacement"
## [5] "horsepower"   "weight"       "acceleration" "model_year"
## [9] "car_name"
```

```
str(data)#estructura de los datos
```

```
## 'data.frame':    398 obs. of  9 variables:
##  $ No          : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ mpg         : num  28 19 36 28 21 23 15.5 32.9 16 13 ...
##  $ cylinders   : int  4 3 4 4 6 4 8 4 6 8 ...
##  $ displacement: num  140 70 107 97 199 115 304 119 250 318 ...
##  $ horsepower  : int  90 97 75 92 90 95 120 100 105 150 ...
##  $ weight      : int  2264 2330 2205 2288 2648 2694 3962 2615 3897 3755 ...
##  $ acceleration: num  15.5 13.5 14.5 17 15 15 13.9 14.8 18.5 14 ...
##  $ model_year  : int  71 72 82 72 70 75 76 81 75 76 ...
##  $ car_name    : Factor w/ 305 levels "amc ambassador brougham",..: 66 184 165 86 8 18 11 79 42 112
```

```
knitr::kable(summary(data)) #resumen de estadísticos básicos
```

No	mpg	cylinders	displacement	horsepower	weight	acceleration	mod
Min. : 1.0	Min. : 9.00	Min. :3.000	Min. : 68.0	Min. : 46.0	Min. :1613	Min. : 8.00	Min
1st Qu.:100.2	1st Qu.:17.50	1st Qu.:4.000	1st Qu.:104.2	1st Qu.: 76.0	1st Qu.:2224	1st Qu.:13.82	1st
Median :199.5	Median :23.00	Median :4.000	Median :148.5	Median : 92.0	Median :2804	Median :15.50	Medi
Mean :199.5	Mean :23.51	Mean :5.455	Mean :193.4	Mean :104.1	Mean :2970	Mean :15.57	Mea
3rd Qu.:298.8	3rd Qu.:29.00	3rd Qu.:8.000	3rd Qu.:262.0	3rd Qu.:125.0	3rd Qu.:3608	3rd Qu.:17.18	3rd

No	mpg	cylinders	displacement	horsepower	weight	acceleration	mod
Max. :398.0	Max. :46.60	Max. :8.000	Max. :455.0	Max. :230.0	Max. :5140	Max. :24.80	Max
NA	NA	NA	NA	NA	NA	NA	

Si no hay cabecera:

```
data <- read.csv("../data/tema1/auto-mpg-noheader.csv", header = FALSE, sep =",",
                 col.names = c("Numero", "MillasPorGalon", "Cilindrada",
                               "Desplazamiento", "Caballos", "Peso",
                               "Aceleracion", "Año", "Modelo"))

xtable(head(data, 5)) # Muestra los 5 primeros por pantalla
```

```
## % latex table generated in R 3.5.1 by xtable 1.8-3 package
## % Tue Sep 25 12:31:01 2018
## \begin{table}[ht]
## \centering
## \begin{tabular}{rrrrrrrrrl}
##   \hline
##  & Numero & MillasPorGalon & Cilindrada & Desplazamiento & Caballos & Peso & Aceleracion & Año & Mode
##   \hline
## 1 &   1 & 28.00 &   4 & 140.00 &  90 & 2264 & 15.50 &  71 & chevrolet vega 2300 \\
##   2 &   2 & 19.00 &   3 & 70.00 &  97 & 2330 & 13.50 &  72 & mazda rx2 coupe \\
##   3 &   3 & 36.00 &   4 & 107.00 &  75 & 2205 & 14.50 &  82 & honda accord \\
##   4 &   4 & 28.00 &   4 & 97.00 &  92 & 2288 & 17.00 &  72 & datsun 510 (sw) \\
##   5 &   5 & 21.00 &   6 & 199.00 &  90 & 2648 & 15.00 &  70 & amc gremlin \\
##    \hline
## \end{tabular}
## \end{table}
```

```
xtable(tail(data, 5)) # muestra los 5 últimos por pantalla
```

```
## % latex table generated in R 3.5.1 by xtable 1.8-3 package
## % Tue Sep 25 12:31:01 2018
## \begin{table}[ht]
## \centering
## \begin{tabular}{rrrrrrrrrl}
##   \hline
##  & Numero & MillasPorGalon & Cilindrada & Desplazamiento & Caballos & Peso & Aceleracion & Año & Mode
##   \hline
## 394 & 394 & 16.50 &   6 & 168.00 & 120 & 3820 & 16.70 &  76 & mercedes-benz 280s \\
##   395 & 395 & 34.50 &   4 & 105.00 &  70 & 2150 & 14.90 &  79 & plymouth horizon tc3 \\
##   396 & 396 & 38.10 &   4 & 89.00 &  60 & 1968 & 18.80 &  80 & toyota corolla tercel \\
##   397 & 397 & 30.50 &   4 & 98.00 &  63 & 2051 & 17.00 &  77 & chevrolet chevette \\
##   398 & 398 & 19.00 &   6 & 232.00 & 100 & 2634 & 13.00 &  71 & amc gremlin \\
##    \hline
## \end{tabular}
## \end{table}
```

Para leer desde la red

```
data <- read.csv("https://frogames.es/course-contents/r/intro/tema1/WHO.csv")

#head(data)
```

```
#summary(data)

#str(data)
```

Cargar datos desde XML

Para cargar XML hace falta la librería 'XML'.

```
library(XML)

url <- "../data/tema1/cd_catalog.xml"

xmlDoc <- xmlParse(url)

rootNode <- xmlRoot(xmlDoc)

data <- xmlSApply(rootNode, function(x) xmlSApply(x, xmlValue))

df <- data.frame(t(data), row.names = NULL)

xtable(df, caption = "Datos desde un XML")
```

% latex table generated in R 3.5.1 by xtable 1.8-3 package % Tue Sep 25 12:31:03 2018

	TITLE	ARTIST	COUNTRY	COMPANY	PRICE	YEAR
1	Empire Burlesque	Bob Dylan	USA	Columbia	10.90	1985
2	Hide your heart	Bonnie Tyler	UK	CBS Records	9.90	1988
3	Greatest Hits	Dolly Parton	USA	RCA	9.90	1982
4	Still got the blues	Gary Moore	UK	Virgin records	10.20	1990
5	Eros	Eros Ramazzotti	EU	BMG	9.90	1997
6	One night only	Bee Gees	UK	Polydor	10.90	1998
7	Sylvias Mother	Dr.Hook	UK	CBS	8.10	1973
8	Maggie May	Rod Stewart	UK	Pickwick	8.50	1990
9	Romanza	Andrea Bocelli	EU	Polydor	10.80	1996
10	When a man loves a woman	Percy Sledge	USA	Atlantic	8.70	1987
11	Black angel	Savage Rose	EU	Mega	10.90	1995
12	1999 Grammy Nominees	Many	USA	Grammy	10.20	1999
13	For the good times	Kenny Rogers	UK	Mucik Master	8.70	1995
14	Big Willie style	Will Smith	USA	Columbia	9.90	1997
15	Tupelo Honey	Van Morrison	UK	Polydor	8.20	1971
16	Soulsville	Jorn Hoel	Norway	WEA	7.90	1996
17	The very best of	Cat Stevens	UK	Island	8.90	1990
18	Stop	Sam Brown	UK	A and M	8.90	1988
19	Bridge of Spies	T'Pau	UK	Siren	7.90	1987
20	Private Dancer	Tina Turner	UK	Capitol	8.90	1983
21	Midt om natten	Kim Larsen	EU	Medley	7.80	1983
22	Pavarotti Gala Concert	Luciano Pavarotti	UK	DECCA	9.90	1991
23	The dock of the bay	Otis Redding	USA	Stax Records	7.90	1968
24	Picture book	Simply Red	EU	Elektra	7.20	1985
25	Red	The Communards	UK	London	7.80	1987
26	Unchain my heart	Joe Cocker	USA	EMI	8.20	1987

Table 2: Datos desde un XML

Leer desde HTML

Para leer de HTML

```
url <- "../data/tema1/WorldPopulation-wiki.htm"

tables <- readHTMLTable(url)

world.most.populous <- tables[[6]]

xtable(world.most.populous, caption = "Lugares más poblados")
```

% latex table generated in R 3.5.1 by xtable 1.8-3 package % Tue Sep 25 12:31:03 2018

	Rank	Country / Territory	Population	Date	Approx. % of world population	Source
1	1	China[note 4]	1,385,310,000	September 9, 2017	18.3%	[91]
2	2	India	1,321,010,000	September 9, 2017	17.5%	[92]
3	3	United States	325,732,000	September 9, 2017	4.31%	[93]
4	4	Indonesia	261,600,000	October 31, 2016	3.46%	[94]
5	5	Pakistan	208,848,000	September 9, 2017	2.76%	[95]
6	6	Brazil	207,985,000	September 9, 2017	2.75%	[96]
7	7	Nigeria	188,500,000	October 31, 2016	2.49%	[97]
8	8	Bangladesh	163,106,000	September 9, 2017	2.16%	[98]
9	9	Russia	146,773,226	June 1, 2017	1.94%	[99]
10	10	Japan	126,750,000	July 1, 2017	1.68%	[100]

Table 3: Lugares más poblados