

# [0010] DEFAD: Representación y tabulación de datos

## Manipulación de datos

Elvira Ferre Jaén, [elvira@um.es](mailto:elvira@um.es)  
Antonio Maurandi, [amaurandi@um.es](mailto:amaurandi@um.es)

Universidad de Murcia

Marzo 2018



# Introducción



# Instalación y carga

## ■ Instalación

```
install.packages( "dplyr" )  
install.packages( "tidyr" )  
# install.packages( "tidyverse" )
```

## ■ Carga de las librerías

```
library( dplyr )  
library( tidyr )  
# library( tidyverse )
```

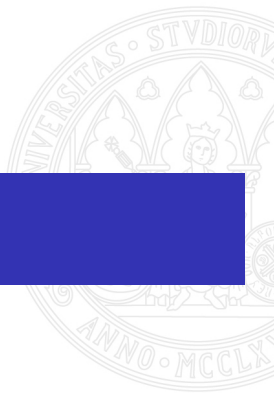
# Base de datos de ejemplo: mtcars

```
head( mtcars )
```

##	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
## Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
## Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4	4
## Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1	4	1
## Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
## Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3	2
## Valiant	18.1	6	225	105	2.76	3.460	20.22	1	0	3	1



dplyr



## select

```
head( select( mtcars, mpg, cyl, disp, hp, drat ) )
```

##	mpg	cyl	disp	hp	drat
## Mazda RX4	21.0	6	160	110	3.90
## Mazda RX4 Wag	21.0	6	160	110	3.90
## Datsun 710	22.8	4	108	93	3.85
## Hornet 4 Drive	21.4	6	258	110	3.08
## Hornet Sportabout	18.7	8	360	175	3.15
## Valiant	18.1	6	225	105	2.76



## select

```
mtcars %>% select( mpg:disp )%>%  
  head
```

##		mpg	cyl	disp
##	Mazda RX4	21.0	6	160
##	Mazda RX4 Wag	21.0	6	160
##	Datsun 710	22.8	4	108
##	Hornet 4 Drive	21.4	6	258
##	Hornet Sportabout	18.7	8	360
##	Valiant	18.1	6	225



## select

```
select( mtcars, mpg:drat )
select( mtcars, -cyl, -disp )
select( mtcars, starts_with( "d" ) )
select( mtcars, ends_with( "p" ) )
select( mtcars, contains( "a" ) )
# ?select_helpers
```



## filter

```
mtcars %>% filter( mpg > 20, disp > 80 )
```

##	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
## 1	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
## 2	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
## 3	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
## 4	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
## 5	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
## 6	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
## 7	21.5	4	120.1	97	3.70	2.465	20.01	1	0	3	1
## 8	26.0	4	120.3	91	4.43	2.140	16.70	0	1	5	2
## 9	30.4	4	95.1	113	3.77	1.513	16.90	1	1	5	2
## 10	21.4	4	121.0	109	4.11	2.780	18.60	1	1	4	2

## filter

```
filter( mtcars, mpg > 20, disp > 80, gear == 4 )
```

##	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
## 1	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
## 2	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
## 3	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
## 4	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
## 5	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
## 6	21.4	4	121.0	109	4.11	2.780	18.60	1	1	4	2

## filter

```
filter(mtcars, mpg > 20, disp > 80 | vs == 1, gear == 4 )
```

##	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
## 1	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
## 2	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
## 3	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
## 4	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
## 5	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
## 6	32.4	4	78.7	66	4.08	2.200	19.47	1	1	4	1
## 7	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
## 8	33.9	4	71.1	65	4.22	1.835	19.90	1	1	4	1
## 9	27.3	4	79.0	66	4.08	1.935	18.90	1	1	4	1
## 10	21.4	4	121.0	109	4.11	2.780	18.60	1	1	4	2

## summarise

```
summarise( mtcars, mean( disp ), sd( disp ) )
```

```
##   mean(disp) sd(disp)  
## 1    230.7219 123.9387
```



## summarise

```
mtcars %>% summarise( media = mean( disp ),  
                      sd = sd( disp ) )
```

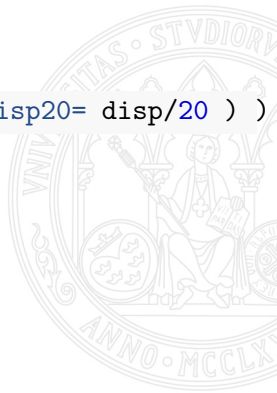
```
##      media      sd  
## 1 230.7219 123.9387
```



## mutate

```
head( mutate( select( mtcars, mpg:disp ), disp20= disp/20 ) )
```

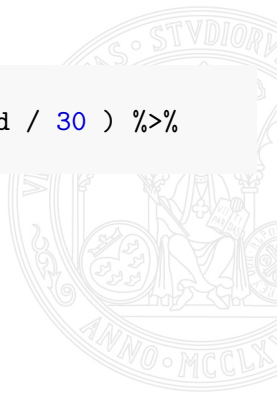
```
##      mpg cyl  disp  disp20
## 1  21.0   6   160    8.00
## 2  21.0   6   160    8.00
## 3  22.8   4   108    5.40
## 4  21.4   6   258   12.90
## 5  18.7   8   360   18.00
## 6  18.1   6   225   11.25
```



## mutate

```
mtcars %>% select( mpg:disp ) %>%  
  mutate( md = mpg * disp, div = md / 30 ) %>%  
  head
```

```
##      mpg cyl disp      md      div  
## 1 21.0   6  160 3360.0 112.00  
## 2 21.0   6  160 3360.0 112.00  
## 3 22.8   4  108 2462.4  82.08  
## 4 21.4   6  258 5521.2 184.04  
## 5 18.7   8  360 6732.0 224.40  
## 6 18.1   6  225 4072.5 135.75
```



## group\_by

```
group_by( mtcars, gear ) %>%
  select( -carb ) %>%
  head()
```

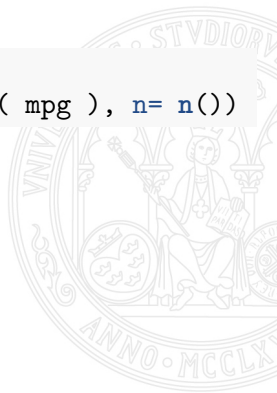
```
## # A tibble: 6 x 10
## # Groups:   gear [2]
##   mpg   cyl  disp    hp  drat    wt   qsec    vs  am  gear
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1  21.0     6   160   110  3.90  2.620  16.46     0     1     4
## 2  21.0     6   160   110  3.90  2.875  17.02     0     1     4
## 3  22.8     4   108    93  3.85  2.320  18.61     1     1     4
## 4  21.4     6   258   110  3.08  3.215  19.44     1     0     3
## 5  18.7     8   360   175  3.15  3.440  17.02     0     0     3
## 6  18.1     6   225   105  2.76  3.460  20.22     1     0     3
```



## group\_by

```
mtcars %>% group_by(carb) %>%
  summarise( media= mean( mpg ), suma= sum( mpg ), n= n())
```

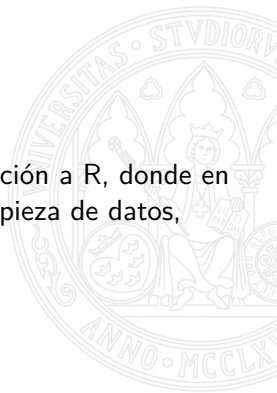
```
## # A tibble: 6 x 4
##   carb   media  suma    n
##   <dbl> <dbl> <dbl> <int>
## 1     1 25.34286 177.4     7
## 2     2 22.40000 224.0    10
## 3     3 16.30000  48.9     3
## 4     4 15.79000 157.9    10
## 5     6 19.70000  19.7     1
## 6     8 15.00000  15.0     1
```



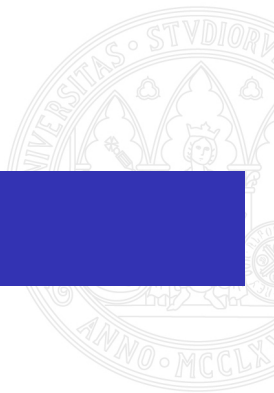
# Más ejemplos

Os dejo aquí un pequeño curso en español de introducción a R, donde en particular explica cómo manipular datos con *dplyr*, limpieza de datos, gráficos, etc .

- Tutorial



# The pipes



# Tuberías

La sintaxis *'the pipe'*, de *tuberías*, permite expresar de forma clara una secuencia de múltiples operaciones.

```
iris %>% select( starts_with( "sep" ) ) %>%
  filter( Sepal.Length > 5 ) %>%
  summary
```

##	Sepal.Length	Sepal.Width
##	Min. :5.100	Min. :2.200
##	1st Qu.:5.600	1st Qu.:2.800
##	Median :6.100	Median :3.000
##	Mean :6.130	Mean :3.048
##	3rd Qu.:6.575	3rd Qu.:3.300
##	Max. :7.900	Max. :4.400

# Tuberías

El operador `%>%` coge el *output* ('la salida') de una sentencia de código y la convierte en el *input* ('el argumento') de una nueva sentencia.

```
iris %>% select(contains( "." ) ) %>%
  apply( 2, summary, digits = 3 )
```

##	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
## Min.	4.30	2.00	1.00	0.1
## 1st Qu.	5.10	2.80	1.60	0.3
## Median	5.80	3.00	4.35	1.3
## Mean	5.84	3.06	3.76	1.2
## 3rd Qu.	6.40	3.30	5.10	1.8
## Max.	7.90	4.40	6.90	2.5

Al describir el código ponemos pensar en él como un “ENTONCES”.



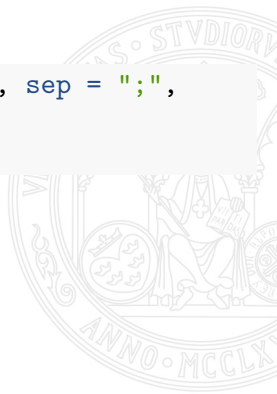
tidyr



# Datos de ejemplo: trat

```
trat <- read.table( "files/tratamiento.csv", sep = ";",  
                   head = TRUE )  
  
head( trat )
```

##	id	genero	raza	m0	m1	m3
## 1	1	1	3	35	25	16
## 2	2	2	1	37	23	12
## 3	3	2	1	36	22	14
## 4	4	1	2	34	21	13
## 5	5	2	3	60	43	22
## 6	6	1	3	54	46	26



## gather

```
trat2 <- trat %>%  
  gather( "mes", "faltaRac", m0, m1, m3 )  
  
head( trat2 )
```

```
##   id genero raza mes faltaRac  
## 1  1      1    3  m0        35  
## 2  2      2    1  m0        37  
## 3  3      2    1  m0        36  
## 4  4      1    2  m0        34  
## 5  5      2    3  m0        60  
## 6  6      1    3  m0        54
```



# spread

```
trat3 <- spread(trat2, mes, faltaRac )  
head( trat3 )
```

```
##      id genero raza m0 m1 m3  
## 1  1      1      3 35 25 16  
## 2  2      2      1 37 23 12  
## 3  3      2      1 36 22 14  
## 4  4      1      2 34 21 13  
## 5  5      2      3 60 43 22  
## 6  6      1      3 54 46 26
```



## gather vs spread

```
head( trat ); head( trat3 )
```

```
##   id genero raza m0 m1 m3
## 1  1      1    3 35 25 16
## 2  2      2    1 37 23 12
## 3  3      2    1 36 22 14
## 4  4      1    2 34 21 13
## 5  5      2    3 60 43 22
## 6  6      1    3 54 46 26
```

```
##   id genero raza m0 m1 m3
## 1  1      1    3 35 25 16
## 2  2      2    1 37 23 12
## 3  3      2    1 36 22 14
## 4  4      1    2 34 21 13
## 5  5      2    3 60 43 22
## 6  6      1    3 54 46 26
```



# Enlaces y referencias



# Enlaces y referencias

- Enlace a la página de CRAN de dplyr
- Data Wrangling Cheatsheet
- Introduction to dplyr
- Tidy data, Hadley Wickham
- The tidy tools manifesto
- Enlace a la página de CRAN de tidyr

