

# Manejo avanzado

## Filtro y cálculo en un data.frame

ISCIH

# Filtro y cálculo

- ❶ Filtrando una base de datos (`subset`)
- ❷ Cálculo de columnas (`:=`)
- ❸ Cálculo agrupado (`by`)
- ❹ Elaboración de tablas (`dcast`)

## Un simple ejemplo

```
DF = data.frame(  
  id = c("Luke Skywalker", "Darth Vader", "Leia Organa", "C-3P0", "R2-D2"),  
  peso = c(77, 136, 49, 75, 32),  
  altura = c(172, 202, 150, 167, 96),  
  especie = c("Humana", "Humana", "Humana", "Droid", "Droid"))
```

DF

```
##           id peso altura especie  
## 1 Luke Skywalker   77   172  Humana  
## 2   Darth Vader  136   202  Humana  
## 3   Leia Organa   49   150  Humana  
## 4         C-3P0   75   167   Droid  
## 5         R2-D2   32    96   Droid
```

## Filtrando la base de datos

```
subset(DF, peso < 50) # personajes que pesan menos de 50 kg
```

```
##           id peso altura especie
## 3 Leia Organa   49   150  Humana
## 5      R2-D2   32    96   Droid
```

```
# Nombre y peso de personajes cuyo nombre termina en 'er'
```

```
subset(DF, grepl("er$",id), select = c(id, peso) )
```

```
##           id peso
## 1 Luke Skywalker   77
## 2      Darth Vader 136
```

## Ejercicio

A partir de la siguiente base de datos

```
## 'data.frame':    200 obs. of  9 variables:
## $ ID             : num  137 174 200 23 39 90 40 115 72 27 ...
## $ edad           : num  37 85 29 13 49 12 85 31 39 70 ...
## $ sexo           : chr   "Mujer" "Mujer" "Hombre" "Hombre" ...
## $ estado.civil   : chr   "Casado" "Soltero" "Casado" "Divorciado" ...
## $ nivel.estudios : chr   "Bajo" "Alto" "Bajo" "Alto" ...
## $ peso           : num  59.6 60 79.2 80.8 80.8 ...
## $ altura         : num  151 149 169 171 171 ...
## $ fumador        : chr   "No" "No" "No" "Si" ...
## $ diabetes       : chr   "No" "Si" "Si" "Si" ...
```

extraer el peso, la altura y el habito con el tabaco de los hombres divorciados.

## Cálculo sobre columnas

```
require(data.table) # simplifica la manipulación de data.frame
DT = data.table(DF) # convierte DF en un objeto data.table

DT[,imc:=peso/(altura/100)^2] # cálculo del índice de masa corporal
DT
```

##		id	peso	altura	especie	imc
## 1:	Luke Skywalker	77	172	Humana	26.02758	
## 2:	Darth Vader	136	202	Humana	33.33007	
## 3:	Leia Organa	49	150	Humana	21.77778	
## 4:	C-3P0	75	167	Droid	26.89232	
## 5:	R2-D2	32	96	Droid	34.72222	

## Cálculo agrupado

```
# IMC respecto a una referencia grupal  
DT[,exceso.imc := imc / mean(imc) , by=especie]  
DT
```

##	id	peso	altura	especie	imc	exceso.imc
## 1:	Luke Skywalker	77	172	Humana	26.02758	0.9623755
## 2:	Darth Vader	136	202	Humana	33.33007	1.2323864
## 3:	Leia Organa	49	150	Humana	21.77778	0.8052381
## 4:	C-3PO	75	167	Droid	26.89232	0.8729213
## 5:	R2-D2	32	96	Droid	34.72222	1.1270787

## Ejercicio

Partiendo de la base datos, realizar el mismo cálculo que en el ejemplo anterior, pero usando como referencia el IMC medio por sexo, estado civil y nivel de estudios.

##	ID	edad	sexo	estado.civil	nivel.estudios	exceso.imc
##	1:	137	37	Mujer	Casado	Bajo 0.9816565
##	2:	174	85	Mujer	Soltero	Alto 1.0024880
##	3:	200	29	Hombre	Casado	Bajo 0.9937224
##	4:	23	13	Hombre	Divorciado	Alto 0.9916774
##	5:	39	49	Hombre	Divorciado	Bajo 1.0045125
##	---					
##	196:	42	72	Mujer	Divorciado	Bajo 1.0111110
##	197:	145	50	Hombre	Divorciado	Medio 0.9852689
##	198:	57	41	Hombre	Soltero	Medio 1.0105011
##	199:	169	44	Mujer	Soltero	Bajo 1.0174449
##	200:	143	54	Mujer	Divorciado	Medio 1.0079203



## Agregando filas

```
DT[,.(media = mean(imc), N = .N), by=especie]
```

```
##      especie      media N
## 1:   Humana 27.04514 3
## 2:    Droid 30.80727 2
```

```
setorder(datos,sexo,estado.civil) # ordena por sexo y estado civil
datos[,.(prev.fumador = mean(fumador=="Si") ),by=.(sexo,estado.civil)]
```

```
##      sexo estado.civil prev.fumador
## 1: Hombre      Casado      0.6666667
## 2: Hombre  Divorciado      0.4102564
## 3: Hombre      Soltero      0.3870968
## 4:  Mujer      Casado      0.3500000
## 5:  Mujer  Divorciado      0.4166667
## 6:  Mujer      Soltero      0.5000000
```

## Creando tablas

```
porcentaje<-function(x) percent(mean(x=="Si")) # require(scales)
dcast(datos,estado.civil ~ sexo, fun=porcentaje, value.var ="fumador")
```

```
##      estado.civil Hombre Mujer
## 1:      Casado      67%   35%
## 2:   Divorciado      41%   42%
## 3:      Soltero      39%   50%
```

```
dcast(datos, estado.civil ~ ., fun=porcentaje,
      value.var = c("fumador", "diabetes", "cancer.prostata"),
      subset=(sexo=="Hombre"))
```

```
##      estado.civil fumador diabetes cancer.prostata
## 1:      Casado      67%      43%          80%
## 2:   Divorciado      41%      44%          67%
## 3:      Soltero      39%      48%          84%
```

- 1 Elaborar una tabla de contingencia donde viene reflejada la relación entre consumo de tabaco y cáncer de próstata en hombres.

##	fumador	Casos	No casos
## 1:	No	42	10
## 2:	Si	34	14

- 2 Calcular el tiempo de supervivencia mediano (en días) en casos de cáncer de mama (mujeres) por estado civil y nivel de estudios.

##	estado.civil	Alto	Bajo	Medio
## 1:	Casado	57.0 days	26.0 days	62.5 days
## 2:	Divorciado	17.0 days	46.5 days	34.0 days
## 3:	Soltero	68.5 days	60.0 days	42.0 days