

Manejo avanzado

Combinación y reestructuración de bases de datos

ISCIII

Combinación y reestructuración

- ❶ Combinación de bases de datos (`rbind`)
- ❷ Estratificación de base de datos (`split`)
- ❸ Cruzando bases de datos (`merge`)
- ❹ Reestructuración de bases de datos (`melt` y `dcast`)

Combinando data.frames

Mismas variables pero distintos registros

```
require(data.table) # simplifica la manipulación de data.frame

DF1=data.table(id="Luke Skywalker", altura=172, color.ojos="azul")
DF2=data.table(id="Darth Vader", altura=202, color.ojos="amarillo")
DF3=data.table(id="Leia Organa", altura=150, color.ojos="marrón")

rbind(DF1,DF2,DF3)
```

```
##           id altura color.ojos
## 1: Luke Skywalker   172      azul
## 2:   Darth Vader   202 amarillo
## 3:   Leia Organa   150  marrón
```

Ejercicio

Juntar las dos bases siguientes

```
datos=data.table(datos) #convierte datos en un objeto data.table  
datos.hombres=subset(datos, sexo=="Hombre", select=c(sexo,peso))  
datos.mujeres=subset(datos, sexo=="Mujer", select=c(sexo,peso))
```

```
str(datos.hombres)
```

```
## Classes 'data.table' and 'data.frame':  100 obs. of  2 variables:  
## $ sexo: chr  "Hombre" "Hombre" "Hombre" "Hombre" ...  
## $ peso: num  79.2 80.8 80.8 79.8 80.7 ...
```

```
str(datos.mujeres)
```

```
## Classes 'data.table' and 'data.frame':  100 obs. of  2 variables:  
## $ sexo: chr  "Mujer" "Mujer" "Mujer" "Mujer" ...  
## $ peso: num  59.6 60 61.3 60.7 60.1 ...
```

Estratificando data.frames

Operación inversa a la combinación

```
temp=subset(datos,select=c(sexo,estado.civil,peso))  
estratos=split(temp, by="sexo")
```

```
str(estratos)
```

```
## List of 2  
## $ Mujer :Classes 'data.table' and 'data.frame': 100 obs. of 3 variable  
## ..$ sexo : chr [1:100] "Mujer" "Mujer" "Mujer" "Mujer" ...  
## ..$ estado.civil: chr [1:100] "Casado" "Soltero" "Soltero" "Divorciado"  
## ..$ peso : num [1:100] 59.6 60 61.3 60.7 60.1 ...  
## $ Hombre:Classes 'data.table' and 'data.frame': 100 obs. of 3 variable  
## ..$ sexo : chr [1:100] "Hombre" "Hombre" "Hombre" "Hombre" ...  
## ..$ estado.civil: chr [1:100] "Casado" "Divorciado" "Divorciado" "Casa  
## ..$ peso : num [1:100] 79.2 80.8 80.8 79.8 80.7 ...
```

Estratificando data.frames

Estratificación de acuerdo a más de una variable

```
temp=subset(datos,select=c(sexo,estado.civil,peso))  
estratos=split(temp, by=c("sexo","estado.civil") )
```

```
str(estratos[1:2])
```

```
## List of 2  
## $ Mujer.Casado :Classes 'data.table' and 'data.frame': 20 obs. of 3 v  
## ..$ sexo : chr [1:20] "Mujer" "Mujer" "Mujer" "Mujer" ...  
## ..$ estado.civil: chr [1:20] "Casado" "Casado" "Casado" "Casado" ...  
## ..$ peso : num [1:20] 59.6 60.3 58 58.4 60.6 ...  
## $ Mujer.Soltero:Classes 'data.table' and 'data.frame': 44 obs. of 3 v  
## ..$ sexo : chr [1:44] "Mujer" "Mujer" "Mujer" "Mujer" ...  
## ..$ estado.civil: chr [1:44] "Soltero" "Soltero" "Soltero" "Soltero" .  
## ..$ peso : num [1:44] 60 61.3 60.4 57.9 60.4 ...
```

Cruzando data.frames

Mismos registros pero distintas variables

```
DF1=data.table(id=c("C-3P0", "R2-D2", "Chewbacca"), altura=c(167, 96, 228))  
DF2=data.table(id=c("C-3P0", "R2-D2"), peso=c(75, 32))  
  
merge(DF1, DF2, by="id")
```

```
##           id altura peso  
## 1: C-3P0      167    75  
## 2: R2-D2       96    32
```

Cruzando data.frames

Mismos registros pero distintas variables

```
DF1=data.table(id=c("C-3P0", "R2-D2", "Chewbacca"), altura=c(167, 96, 228))
DF2=data.table(id=c("C-3P0", "R2-D2"), peso=c(75, 32))

merge(DF1, DF2, by="id", all.x=TRUE)
```

```
##           id altura peso
## 1:      C-3P0   167   75
## 2: Chewbacca   228  NA
## 3:      R2-D2    96   32
```


Ejercicio

Juntar las dos bases siguientes

```
socio.demo=subset(datos, select=ID:nivel.estudios)
basal=subset(datos, select=c(ID,peso:diabetes))
head(socio.demo, 3)
```

```
##      ID edad  sexo estado.civil nivel.estudios
## 1: 137   37 Mujer      Casado          Bajo
## 2: 174   85 Mujer      Soltero          Alto
## 3: 200   29 Hombre     Casado          Bajo
```

```
head(basal,3)
```

```
##      ID      peso  altura fumador diabetes
## 1: 137 59.58221 150.7163      No        No
## 2: 174 59.95427 149.2075      No        Si
## 3: 200 79.20674 168.9795      No        Si
```

Reestructuración de data.frames

Formato wide: una columna para cada variación

```
wide=subset(datos, sexo=="Mujer", select=c(ID:sexo,fdiag_cm,fdef))  
  
head(wide)
```

	ID	edad	sexo	fdiag_cm	fdef
## 1:	137	37	Mujer	1977-05-05	1977.05.14
## 2:	174	85	Mujer	<NA>	<NA>
## 3:	115	31	Mujer	1980-01-01	1980.03.24
## 4:	72	39	Mujer	1996-10-12	1997.01.08
## 5:	19	24	Mujer	2003-03-18	2003.04.02
## 6:	15	42	Mujer	1996-05-13	1996.07.23

Reestructuración de data.frames

Formato long: un registro para cada variación

```
long = melt(wide, id=1:3) # id : variables que se quedan fijas  
setkey(long,ID) # reordenando la base por ID
```

```
head(long)
```

```
##      ID edad  sexo variable      value  
## 1:   2   13 Mujer fdiag_cm 2013-04-16  
## 2:   2   13 Mujer      fdef 2013.05.16  
## 3:   4    9 Mujer fdiag_cm 1989-09-09  
## 4:   4    9 Mujer      fdef 1989.11.24  
## 5:   6   66 Mujer fdiag_cm      <NA>  
## 6:   6   66 Mujer      fdef      <NA>
```

Reestructuración de data.frames

Formato long: un registro para cada variación

```
long = melt(wide, measure=4:5) # measure : variables que varian  
setkey(long,ID) # reordenando la base por ID
```

```
head(long)
```

```
##      ID edad  sexo variable      value  
## 1:   2   13 Mujer fdiag_cm 2013-04-16  
## 2:   2   13 Mujer      fdef 2013.05.16  
## 3:   4    9 Mujer fdiag_cm 1989-09-09  
## 4:   4    9 Mujer      fdef 1989.11.24  
## 5:   6   66 Mujer fdiag_cm      <NA>  
## 6:   6   66 Mujer      fdef      <NA>
```

Reestructuración de data.frames

Volviendo al formato wide...

```
wide = dcast(long, ID + sexo + edad ~ variable)

head(wide)
```

	ID	sexo	edad	fdiag_cm	fdef
## 1:	2	Mujer	13	2013-04-16	2013.05.16
## 2:	4	Mujer	9	1989-09-09	1989.11.24
## 3:	6	Mujer	66	<NA>	<NA>
## 4:	9	Mujer	82	2009-05-21	2009.06.04
## 5:	10	Mujer	6	1984-03-24	1984.05.02
## 6:	14	Mujer	83	<NA>	<NA>

Ejercicio

Poner la siguiente base en formato long

VADeaths

	Rural Male	Rural Female	Urban Male	Urban Female
## 50-54	11.7	8.7	15.4	8.4
## 55-59	18.1	11.7	24.3	13.6
## 60-64	26.9	20.3	37.0	19.3
## 65-69	41.0	30.9	54.6	35.1
## 70-74	66.0	54.3	71.1	50.0

```
temp=data.table(VADeaths, keep.rownames=TRUE)
VADeaths.long = melt(VADeaths, id=1) # o bien melt(VADeaths, measure=1:4)
head(VADeaths.long )
```

	Var1	Var2	value
## 1	50-54	Rural Male	11.7
## 2	55-59	Rural Male	18.1
## 3	60-64	Rural Male	26.9
## 4	65-69	Rural Male	41.0
## 5	70-74	Rural Male	66.0
## 6	50-54	Rural Female	8.7

Ejercicio

... y volver al formato original a partir de

```
head(VADeaths.long)
```

```
##      Var1      Var2 value
## 1 50-54    Rural Male  11.7
## 2 55-59    Rural Male  18.1
## 3 60-64    Rural Male  26.9
## 4 65-69    Rural Male  41.0
## 5 70-74    Rural Male  66.0
## 6 50-54 Rural Female   8.7
```

```
dcast(VADeaths.long, Var1 ~ Var2, fun.var="value")
```

```
##      Var1 Rural Male Rural Female Urban Male Urban Female
## 1 50-54      11.7      8.7      15.4      8.4
## 2 55-59      18.1     11.7     24.3     13.6
## 3 60-64      26.9     20.3     37.0     19.3
## 4 65-69      41.0     30.9     54.6     35.1
## 5 70-74      66.0     54.3     71.1     50.0
```