Manejo de datos con R

Oscar Perpiñán Lamigueiro
http://oscarperpinan.github.io

Manejo de datos
con R

Oscar Perpiñán
Lamigueiro
http://
oscarperpinan.
github.io

Fuentes de datos

- The R Datasets Package
- Enlaces en Bibsonomy
- …

Manejo de datos
con R

Oscar Perpiñán
Lamigueiro
http://
oscarperpinan.
github.io

setwd, getwd, dir

Manejo de datos
con R

Oscar Perpiñán
Lamigueiro
http://
oscarperpinan.
github.io

```
getwd()
old <- setwd("~/github/intro")
dir()
dir(pattern='.R')
dir('data')
```

read.table

Manejo de datos
con R

Oscar Perpiñán
Lamigueiro
http://
oscarperpinan.
github.io

Fuentes de datos

Lectura de datos

Datos agregados

Cambio de
formato

```
CO2 <- read.table('data/CO2_GNI_BM.csv', header=TRUE,
    sep=',')
head(CO2)
```

```
   Country.Name Country.Code                            Indicator.Name
1      Finland          FIN           CO2 emissions (kg per PPP $ of GDP)
2      Finland          FIN       CO2 emissions (metric tons per capita)
3      Finland          FIN           GNI, PPP (current international $)
4      Finland          FIN GNI per capita, PPP (current international $)
5       France          FRA           CO2 emissions (kg per PPP $ of GDP)
6       France          FRA       CO2 emissions (metric tons per capita)
    Indicator.Code      X2000       X2001       X2002       X2003
1 EN.ATM.CO2E.PP.GD 3.923481e-01 4.099378e-01 4.265803e-01 4.785172e-01
2    EN.ATM.CO2E.PC 1.007322e+01 1.087588e+01 1.174433e+01 1.321467e+01
3 NY.GNP.MKTP.PP.CD 1.318800e+11 1.374500e+11 1.434180e+11 1.428710e+11
4 NY.GNP.PCAP.PP.CD 2.548000e+04 2.649000e+04 2.758000e+04 2.741000e+04
5 EN.ATM.CO2E.PP.GD 2.384221e-01 2.370408e-01 2.231432e-01 2.287341e-01
6    EN.ATM.CO2E.PC 6.016236e+00 6.303892e+00 6.171683e+00 6.236447e+00
       X2004       X2005       X2006       X2007       X2008       X2009
1 4.289469e-01 3.389595e-01 3.786006e-01 3.341890e-01 2.792975e-01           NA
2 1.280953e+01 1.040875e+01 1.254696e+01 1.208669e+01 1.063578e+01           NA
3 1.573070e+11 1.618390e+11 1.761810e+11 1.913600e+11 2.033820e+11 1.93598e+11
4 3.009000e+04 3.085000e+04 3.345000e+04 3.618000e+04 3.828000e+04 3.62600e+04
5 2.212615e-01 2.105801e-01 1.918882e-01 1.770519e-01 1.720425e-01           NA
6 6.232062e+00 6.219341e+00 6.026120e+00 5.864109e+00 5.873142e+00           NA
     X2010 X2011
1       NA     NA
2       NA     NA
3 1.98866e+11     NA
4 3.70700e+04     NA
5       NA     NA
6       NA     NA
```

read.csv, read.csv2

Manejo de datos
con R

Oscar Perpiñán
Lamigueiro
http://
oscarperpinan.
github.io

Fuentes de datos

Lectura de datos

Datos agregados

Cambio de
formato

- ▶ `read.csv` y `read.csv2` son como `read.table` con valores por defecto para encabezado y separadores

```
CO2 <- read.csv('data/CO2_GNI_BM.csv')
```

```
names(CO2)
head(CO2)
tail(CO2)
```

```
summary(CO2)
```

```
 Country.Name  Country.Code
Brazil : 4    BRA   : 4
China  : 4    CHN   : 4
Finland: 4    DEU   : 4
France : 4    ESP   : 4
Germany: 4    FIN   : 4
Greece : 4    FRA   : 4
(Other):16    (Other):16
                                    Indicator.Name              Indicator.Code
CO2 emissions (kg per PPP $ of GDP)           :10    EN.ATM.CO2E.PC   :10
CO2 emissions (metric tons per capita)        :10    EN.ATM.CO2E.PP.GD:10
GNI per capita, PPP (current international $):10     NY.GNP.MKTP.PP.CD:10
GNI, PPP (current international $)             :10    NY.GNP.PCAP.PP.CD:10


    X2000              X2001              X2002
Min.   :0.000e+00  Min.   :0.000e+00  Min.   :0.000e+00
1st Qu.:1.000e+00  1st Qu.:1.000e+00  1st Qu.:1.000e+00
```

table, `data.frame` de ejemplo

Manejo de datos
con R

Oscar Perpiñán
Lamigueiro
http://
oscarperpinan.
github.io

Fuentes de datos

Lectura de datos

Datos agregados

Cambio de
formato

```
chromosome <- gl(3, 10, labels = c('A', 'B', 'C'))
probeset <- gl(3, 10, labels = c('X', 'Y', 'Z'))
ensg <- gl(3, 10, labels = c('E1', 'E2', 'E3'))
symbol <- gl(3, 10, labels = c('S1', 'S2', 'S3'))
XXA_00 <- rnorm(30)
XXA_36 <- rnorm(30)
XXB_00 <- rnorm(30)
```

```
chromo <- data.frame(chromosome, probeset, ensg,
    symbol,
    XXA_00, XXA_36, XXB_00)
head(chromo)
```

```
  chromosome probeset ensg symbol     XXA_00      XXA_36     XXB_00
1          A        X   E1     S1 -0.9055000  0.39414564 -0.6600499
2          A        X   E1     S1 -0.7349505  0.50905638 -1.2176382
3          A        X   E1     S1 -0.3175336 -0.78622185 -2.0760214
4          A        X   E1     S1 -0.6429121 -0.05785827  1.1762978
5          A        X   E1     S1  0.3393823 -0.22097023 -0.5870772
6          A        X   E1     S1 -0.8632702 -0.13718955  0.3643206
```

table

Manejo de datos
con R

Oscar Perpiñán
Lamigueiro
http://
oscarperpinan.
github.io

Fuentes de datos

Lectura de datos

Datos agregados

Cambio de
formato

```r
table(chromo$chromosome, chromo$XXA_00 > 0)
```

```
    FALSE TRUE
A       6    4
B       5    5
C       4    6
```

```r
table(chromo$probeset, chromo$XXA_00 > -1 & chromo$
    XXA_00 < 1)
```

```
    FALSE TRUE
X       1    9
Y       6    4
Z       4    6
```

xtabs

Manejo de datos
con R

Oscar Perpiñán
Lamigueiro
http://
oscarperpinan.
github.io

Fuentes de datos

Lectura de datos

Datos agregados

Cambio de
formato

```
xtabs(XXA_00 > 1 ~ chromosome + probeset,
      data=chromo)
```

```
          probeset
chromosome X Y Z
         A 1 0 0
         B 0 4 0
         C 0 0 2
```

tapply

Manejo de datos
con R

Oscar Perpiñán
Lamigueiro
http://
oscarperpinan.
github.io

Fuentes de datos

Lectura de datos

Datos agregados

Cambio de
formato

```
tapply(CO2$X2000, CO2$Indicator.Name,
       FUN=mean)
```

```
            CO2 emissions (kg per PPP $ of GDP)
                                   4.777875e-01
       CO2 emissions (metric tons per capita)
                                   7.580861e+00
GNI per capita, PPP (current international $)
                                   1.981000e+04
             GNI, PPP (current international $)
                                   2.078196e+12
```

tapply

Manejo de datos
con R

Oscar Perpiñán
Lamiguero
http://
oscarperpinan.
github.io

Fuentes de datos

Lectura de datos

Datos agregados

Cambio de
formato

```
tapply(CO2$X2000, CO2[,c("Indicator.Name", "Country.
    Name")],
        FUN=mean)
```

```
                                                Country.Name
Indicator.Name                                        Brazil        China
  CO2 emissions (kg per PPP $ of GDP)           2.699746e-01 1.140619e+00
  CO2 emissions (metric tons per capita)        1.892645e+00 2.696862e+00
  GNI per capita, PPP (current international $)  6.820000e+03 2.340000e+03
  GNI, PPP (current international $)             1.188790e+12 2.948850e+12
                                                Country.Name
Indicator.Name                                       Finland       France
  CO2 emissions (kg per PPP $ of GDP)           3.923481e-01 2.384221e-01
  CO2 emissions (metric tons per capita)        1.007322e+01 6.016236e+00
  GNI per capita, PPP (current international $)  2.548000e+04 2.566000e+04
  GNI, PPP (current international $)             1.318800e+11 1.558990e+12
                                                Country.Name
Indicator.Name                                       Germany       Greece
  CO2 emissions (kg per PPP $ of GDP)           3.929031e-01 4.598579e-01
  CO2 emissions (metric tons per capita)        1.012147e+01 8.391709e+00
  GNI per capita, PPP (current international $)  2.549000e+04 1.832000e+04
  GNI, PPP (current international $)             2.095450e+12 2.000130e+11
                                                Country.Name
Indicator.Name                                         India       Norway
  CO2 emissions (kg per PPP $ of GDP)           7.448517e-01 2.391275e-01
  CO2 emissions (metric tons per capita)        1.125975e+00 8.641315e+00
  GNI per capita, PPP (current international $)  1.500000e+03 3.565000e+04
  GNI, PPP (current international $)             1.575930e+12 1.601000e+11
                                                Country.Name
Indicator.Name                                         Spain United States
  CO2 emissions (kg per PPP $ of GDP)           3.428950e-01  5.568755e-01
  CO2 emissions (metric tons per capita)        7.312922e+00  1.953626e+01
  GNI per capita, PPP (current international $)  2.115000e+04  3.569000e+04
```

aggregate

```
aggregate(X2000 ~ Indicator.Name,
  data=CO2, FUN=mean)
```

```
                     Indicator.Name         X2000
1       CO2 emissions (kg per PPP $ of GDP) 4.777875e-01
2       CO2 emissions (metric tons per capita) 7.580861e+00
3 GNI per capita, PPP (current international $) 1.981000e+04
4         GNI, PPP (current international $) 2.078196e+12
```

```
aggregate(cbind(X2000, X2001) ~ Indicator.Name,
  data=CO2, FUN=mean)
```

```
                     Indicator.Name         X2000         X2001
1       CO2 emissions (kg per PPP $ of GDP) 4.777875e-01 4.591328e-01
2       CO2 emissions (metric tons per capita) 7.580861e+00 7.725765e+00
3 GNI per capita, PPP (current international $) 1.981000e+04 2.066300e+04
4         GNI, PPP (current international $) 2.078196e+12 2.182390e+12
```

```
aggregate(X2000 ~ Indicator.Name + Country.Name,
  data=CO2, FUN=mean)
```

```
                      Indicator.Name Country.Name        X2000
1       CO2 emissions (kg per PPP $ of GDP)       Brazil 2.699746e-01
2       CO2 emissions (metric tons per capita)       Brazil 1.892645e+00
3 GNI per capita, PPP (current international $)       Brazil 6.820000e+03
4         GNI, PPP (current international $)       Brazil 1.188790e+12
5       CO2 emissions (kg per PPP $ of GDP)        China 1.140619e+00
6       CO2 emissions (metric tons per capita)        China 2.696862e+00
7 GNI per capita, PPP (current international $)        China 2.340000e+03
8         GNI, PPP (current international $)        China 2.948850e+12
9       CO2 emissions (kg per PPP $ of GDP)      Finland 3.923481e-01
```

Manejo de datos
con R

Oscar Perpiñán
Lamigueiro
http://
oscarperpinan.
github.io

Fuentes de datos

Lectura de datos

Datos agregados

Cambio de
formato

aggregate

```
aggregate(cbind(X2000, X2001) ~
  Indicator.Name + Country.Name,
  data=CO2, FUN=mean)
```

	Indicator.Name	Country.Name	X2000
1	CO2 emissions (kg per PPP $ of GDP)	Brazil	2.699746e-01
2	CO2 emissions (metric tons per capita)	Brazil	1.892645e+00
3	GNI per capita, PPP (current international $)	Brazil	6.820000e+03
4	GNI, PPP (current international $)	Brazil	1.188790e+12
5	CO2 emissions (kg per PPP $ of GDP)	China	1.140619e+00
6	CO2 emissions (metric tons per capita)	China	2.696862e+00
7	GNI per capita, PPP (current international $)	China	2.340000e+03
8	GNI, PPP (current international $)	China	2.948850e+12
9	CO2 emissions (kg per PPP $ of GDP)	Finland	3.923481e-01
10	CO2 emissions (metric tons per capita)	Finland	1.007322e+01
11	GNI per capita, PPP (current international $)	Finland	2.548000e+04
12	GNI, PPP (current international $)	Finland	1.318800e+11
13	CO2 emissions (kg per PPP $ of GDP)	France	2.384221e-01
14	CO2 emissions (metric tons per capita)	France	6.016236e+00
15	GNI per capita, PPP (current international $)	France	2.566000e+04
16	GNI, PPP (current international $)	France	1.558990e+12
17	CO2 emissions (kg per PPP $ of GDP)	Germany	3.929031e-01
18	CO2 emissions (metric tons per capita)	Germany	1.012147e+01
19	GNI per capita, PPP (current international $)	Germany	2.549000e+04
20	GNI, PPP (current international $)	Germany	2.095450e+12
21	CO2 emissions (kg per PPP $ of GDP)	Greece	4.598579e-01
22	CO2 emissions (metric tons per capita)	Greece	8.391709e+00
23	GNI per capita, PPP (current international $)	Greece	1.832000e+04
24	GNI, PPP (current international $)	Greece	2.000130e+11
25	CO2 emissions (kg per PPP $ of GDP)	India	7.448517e-01
26	CO2 emissions (metric tons per capita)	India	1.125975e+00
27	GNI per capita, PPP (current international $)	India	1.500000e+03
28	GNI, PPP (current international $)	India	1.575930e+12

Manejo de datos con R

Oscar Perpiñán Lamigueiro

http://oscarperpinan.github.io

Fuentes de datos

Lectura de datos

Datos agregados

Cambio de formato

aggregate

Manejo de datos con R

Oscar Perpiñán Lamigueiro
http://
oscarperpinan.
github.io

Fuentes de datos

Lectura de datos

Datos agregados

Cambio de formato

```
aggregate(cbind(XXA_00, XXA_36, XXB_00) ~
  ensg + chromosome + symbol,
  data = chromo, FUN = mean)
```

```
  ensg chromosome symbol      XXA_00      XXA_36      XXB_00
1  E1          A     S1 -0.04095755 -0.2906532 -0.54415660
2  E2          B     S2  0.32532301 -0.2300200 -0.05928762
3  E3          C     S3  0.17182752 -0.2650071 -0.10832663
```

```
aggregate(cbind(XXA_00, XXA_36, XXB_00) ~ ensg ,
  data = chromo, FUN = mean)
```

```
  ensg      XXA_00      XXA_36      XXB_00
1  E1 -0.04095755 -0.2906532 -0.54415660
2  E2  0.32532301 -0.2300200 -0.05928762
3  E3  0.17182752 -0.2650071 -0.10832663
```

stack

Manejo de datos
con R

Oscar Perpiñán
Lamigueiro
http://
oscarperpinan.
github.io

Fuentes de datos

Lectura de datos

Datos agregados

Cambio de
formato

- ▶ Primero escogemos un subconjunto

```
CO2China <- subset(CO2,
  subset=(Country.Name=='China' &
  Indicator.Name=='CO2 emissions (kg per PPP $ of
    GDP)'),
  select=-c(Country.Name, Country.Code,
    Indicator.Name, Indicator.Code))
head(CO2China)
```

```
      X2000    X2001    X2002    X2003    X2004    X2005    X2006     X2007
29 1.140619 1.054772 1.007715 1.098485 1.133811 1.079371 1.027606 0.9255433
      X2008 X2009 X2010 X2011
29 0.8556903    NA    NA    NA
```

stack

Manejo de datos
con R

Oscar Perpiñán
Lamigueiro
http://
oscarperpinan.
github.io

- ▶ Pasamos de formato `wide` a `long`

`stack(CO2China)`

```
      values    ind
1   1.1406188 X2000
2   1.0547715 X2001
3   1.0077152 X2002
4   1.0984850 X2003
5   1.1338112 X2004
6   1.0793710 X2005
7   1.0276060 X2006
8   0.9255433 X2007
9   0.8556903 X2008
10        NA X2009
11        NA X2010
12        NA X2011
```

reshape: wide a long

Manejo de datos
con R

Oscar Perpiñán
Lamigueiro
http://
oscarperpinan.
github.io

Fuentes de datos

Lectura de datos

Datos agregados

Cambio de
formato

▶ Primer intento

```
CO2long <- reshape(CO2,
    varying=list(names(CO2)[5:16]),
    direction='long')
head(CO2long)
```

```
    Country.Name Country.Code                              Indicator.Name
1.1      Finland          FIN         CO2 emissions (kg per PPP $ of GDP)
2.1      Finland          FIN      CO2 emissions (metric tons per capita)
3.1      Finland          FIN               GNI, PPP (current international $)
4.1      Finland          FIN GNI per capita, PPP (current international $)
5.1       France          FRA         CO2 emissions (kg per PPP $ of GDP)
6.1       France          FRA      CO2 emissions (metric tons per capita)
        Indicator.Code time      X2000 id
1.1 EN.ATM.CO2E.PP.GD    1 3.923481e-01  1
2.1    EN.ATM.CO2E.PC    1 1.007322e+01  2
3.1 NY.GNP.MKTP.PP.CD    1 1.318800e+11  3
4.1 NY.GNP.PCAP.PP.CD    1 2.548000e+04  4
5.1 EN.ATM.CO2E.PP.GD    1 2.384221e-01  5
6.1    EN.ATM.CO2E.PC    1 6.016236e+00  6
```

reshape: wide a long

Manejo de datos
con R

Oscar Perpiñán
Lamigueiro
http://
oscarperpinan.
github.io

Fuentes de datos

Lectura de datos

Datos agregados

Cambio de
formato

▶ Añadimos argumentos

```
CO2long <- reshape(CO2,
   varying=list(names(CO2)[5:16]),
   timevar='Year', v.names='Value',
   times=2000:2011,
   direction='long')
head(CO2long)
```

```
       Country.Name Country.Code                        Indicator.Name
1.2000      Finland          FIN        CO2 emissions (kg per PPP $ of GDP)
2.2000      Finland          FIN      CO2 emissions (metric tons per capita)
3.2000      Finland          FIN             GNI, PPP (current international $)
4.2000      Finland          FIN GNI per capita, PPP (current international $)
5.2000       France          FRA        CO2 emissions (kg per PPP $ of GDP)
6.2000       France          FRA      CO2 emissions (metric tons per capita)
       Indicator.Code Year        Value id
1.2000 EN.ATM.CO2E.PP.GD 2000 3.923481e-01  1
2.2000   EN.ATM.CO2E.PC 2000 1.007322e+01  2
3.2000 NY.GNP.MKTP.PP.CD 2000 1.318800e+11  3
4.2000 NY.GNP.PCAP.PP.CD 2000 2.548000e+04  4
5.2000 EN.ATM.CO2E.PP.GD 2000 2.384221e-01  5
6.2000   EN.ATM.CO2E.PC 2000 6.016236e+00  6
```

reshape: long a wide

Manejo de datos
con R

Oscar Perpiñán
Lamigueiro
http://
oscarperpinan.
github.io

Fuentes de datos

Lectura de datos

Datos agregados

Cambio de
formato

▶ Primero escogemos las columnas de interés

```
CO2subset <- CO2long[c("Country.Name",
      "Indicator.Name",
      "Year", "Value")]
head(CO2subset)
```

```
      Country.Name                              Indicator.Name Year
1.2000      Finland           CO2 emissions (kg per PPP $ of GDP) 2000
2.2000      Finland        CO2 emissions (metric tons per capita) 2000
3.2000      Finland                GNI, PPP (current international $) 2000
4.2000      Finland GNI per capita, PPP (current international $) 2000
5.2000       France           CO2 emissions (kg per PPP $ of GDP) 2000
6.2000       France        CO2 emissions (metric tons per capita) 2000
            Value
1.2000 3.923481e-01
2.2000 1.007322e+01
3.2000 1.318800e+11
4.2000 2.548000e+04
5.2000 2.384221e-01
6.2000 6.016236e+00
```

reshape: long a wide

▶ Ahora cambiamos formato

```
CO2wide <- reshape(CO2subset,
  idvar=c('Country.Name','Year'),
  timevar='Indicator.Name',
  direction='wide')
head(CO2wide)
```

```
        Country.Name Year Value.CO2 emissions (kg per PPP $ of GDP)
1.2000        Finland 2000                              0.3923481
5.2000         France 2000                              0.2384221
9.2000        Germany 2000                              0.3929031
13.2000        Greece 2000                              0.4598579
17.2000        Norway 2000                              0.2391275
21.2000         Spain 2000                              0.3428950
        Value.CO2 emissions (metric tons per capita)
1.2000                                     10.073216
5.2000                                      6.016236
9.2000                                     10.121466
13.2000                                     8.391709
17.2000                                     8.641315
21.2000                                     7.312922
        Value.GNI, PPP (current international $)
1.2000                              1.31880e+11
5.2000                              1.55899e+12
9.2000                              2.09545e+12
13.2000                             2.00013e+12
17.2000                             1.60100e+11
21.2000                             8.51462e+11
        Value.GNI per capita, PPP (current international $)
1.2000                                            25480
```

Manejo de datos
con R

Oscar Perpiñán
Lamigueiro
http://
oscarperpinan.
github.io

Fuentes de datos

Lectura de datos

Datos agregados

Cambio de
formato

reshape: `long` a `wide`

Manejo de datos
con R

Oscar Perpiñán
Lamigueiro
http://
oscarperpinan.
github.io

Fuentes de datos

Lectura de datos

Datos agregados

Cambio de
formato

▶ Y ponemos nombres al gusto

```
names(CO2wide)[3:6] <- c('CO2.PPP', 'CO2.capita',
'GNI.PPP', 'GNI.capita')

head(CO2wide)
```

```
        Country.Name Year    CO2.PPP CO2.capita     GNI.PPP GNI.capita
1.2000       Finland 2000 0.3923481  10.073216 1.31880e+11      25480
5.2000        France 2000 0.2384221   6.016236 1.55899e+12      25660
9.2000       Germany 2000 0.3929031  10.121466 2.09545e+12      25490
13.2000       Greece 2000 0.4598579   8.391709 2.00013e+11      18320
17.2000       Norway 2000 0.2391275   8.641315 1.60100e+11      35650
21.2000        Spain 2000 0.3428950   7.312922 8.51462e+11      21150
```