Manejo de datos con R

Oscar Perpiñán Lamigueiro

25 de Enero de 2013

Contenidos

Fuentes de datos

- The R Datasets Package
- Enlaces en Bibsonomy

Contenidos

setwd, getwd, dir

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

read.table

- Con un fichero local

```
download.file('http://oscarperpinan.github.com/
    spacetime-vis/data/CO2_GNI_BM.csv',
            destfile='data/CO2_GNI_BM.csv')

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

- Directamente de un enlace URL

```
CO2 <- read.table('http://oscarperpinan.github.com/
    spacetime-vis/data/CO2_GNI_BM.csv',
              header=TRUE, sep=',')
head(CO2)
```

read.csv, read.csv2

- `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')
```

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

Contenidos

table

```
chromo <- data.frame(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))
```

```
table(chromo$chromosome, df$XXA_00 > 0)
table(chromo$probeset, df$XXA_00 > -1 & df$XXA_00 <
    1)
```

xtabs

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

tapply

```
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

```
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)

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

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

aggregate

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

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

aggregate(cbind(X2000, X2001) ~
        Indicator.Name + Country.Name,
        subset=(Country.Name %in% c('United␣States',
            'China')),
              data=CO2, FUN=mean)
```

aggregate

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

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

Contenidos

stack

- 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))
```

- Pasamos de formato `wide` a `long`

```
stack(CO2China)
```

reshape: wide a long

- Primer intento

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

- Añadimos argumentos

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

reshape: long a wide

- Primero escogemos las columnas de interés

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

- Ahora cambiamos formato

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

- Y ponemos nombres al gusto

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

head(CO2wide)
```