

# Manejo de datos con R

Oscar Perpiñán Lamigueiro

<http://oscarperpinan.github.io>

Lectura de datos

Indexado

Cambio de  
formato

Datos agregados

Unión de  
`data.frame`

Lectura de datos

Indexado

Cambio de formato

Datos agregados

Unión de `data.frame`

# setwd, getwd, dir

En setwd hay que especificar el directorio que contiene el repositorio.

```
getwd()  
old <- setwd("~/github/intro")  
dir()
```

```
dir(pattern='.R')
```

```
[1] "ClasesMetodos.R"      "datos.R"              "estadistica.R"  
[4] "factorDateCharacter.R" "Funciones.R"          "graficos.R"  
[7] "intro.R"              "raster.R"             "zoo.R"
```

```
dir('data')
```

```
[1] "aranjuez.csv"          "aranjuez.RData"       "CO2_GNI_BM.csv"  
[4] "El.Arenosillo.txt"    "InformeDatos.zip"    "NREL-Hawaii.csv"  
[7] "radiacion_datos.csv"  "SIAR.csv"            "SI Smm2008-CMSAF.zip"
```

Manejo de datos  
con R

Oscar Perpiñán  
Lamigueiro  
[http://  
oscarperpinan.  
github.io](http://oscarperpinan.github.io)

Lectura de datos

Indexado

Cambio de  
formato

Datos agregados

Unión de  
data.frame

# Lectura de datos con read.table

Manejo de datos  
con R

Oscar Perpiñán  
Lamigueiro  
[http://  
oscarperpinan.  
github.io](http://oscarperpinan.github.io)

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

	X	TempAvg	TempMax	TempMin	HumidAvg	HumidMax	WindAvg	WindMax	Rain
1	2004-01-01	4.044	10.71	-1.969	88.3	95.9	0.746	3.528	0
2	2004-01-02	5.777	11.52	1.247	83.3	98.5	1.078	6.880	0
3	2004-01-03	5.850	13.32	0.377	75.0	94.4	0.979	6.576	0
4	2004-01-04	4.408	15.59	-2.576	82.0	97.0	0.633	3.704	0
5	2004-01-05	3.081	14.58	-2.974	83.2	97.0	0.389	2.244	0
6	2004-01-06	2.304	11.83	-3.379	84.5	96.5	0.436	2.136	0

  

	Radiation	ET
1	5.490	0.5352688
2	6.537	0.7710499
3	8.810	0.8361229
4	9.790	0.6861381
5	10.300	0.5152422
6	9.940	0.4886631

Lectura de datos

Indexado

Cambio de  
formato

Datos agregados

Unión de  
data.frame

# Lectura de datos con read.csv

Manejo de datos  
con R

Oscar Perpiñán  
Lamigueiro  
[http://  
oscarperpinan.  
github.io](http://oscarperpinan.github.io)

```
aranjuez <- read.csv('data/aranjuez.csv')  
names(aranjuez)[1] <- 'Date'
```

```
class(aranjuez)
```

```
[1] "data.frame"
```

```
names(aranjuez)
```

```
[1] "Date"      "TempAvg"   "TempMax"   "TempMin"   "HumidAvg"   "HumidMax"  
[7] "WindAvg"   "WindMax"   "Rain"       "Radiation" "ET"
```

Lectura de datos

Indexado

Cambio de  
formato

Datos agregados

Unión de  
data.frame

# Inspeccionamos el resultado

```
head(aranjuez)
```

	Date	TempAvg	TempMax	TempMin	HumidAvg	HumidMax	WindAvg	WindMax	Rain
1	2004-01-01	4.044	10.71	-1.969	88.3	95.9	0.746	3.528	0
2	2004-01-02	5.777	11.52	1.247	83.3	98.5	1.078	6.880	0
3	2004-01-03	5.850	13.32	0.377	75.0	94.4	0.979	6.576	0
4	2004-01-04	4.408	15.59	-2.576	82.0	97.0	0.633	3.704	0
5	2004-01-05	3.081	14.58	-2.974	83.2	97.0	0.389	2.244	0
6	2004-01-06	2.304	11.83	-3.379	84.5	96.5	0.436	2.136	0
	Radiation	ET							
1	5.490	0.5352688							
2	6.537	0.7710499							
3	8.810	0.8361229							
4	9.790	0.6861381							
5	10.300	0.5152422							
6	9.940	0.4886631							

```
tail(aranjuez)
```

	Date	TempAvg	TempMax	TempMin	HumidAvg	HumidMax	WindAvg	WindMax	Rain
2893	2011-12-26	3.366	13.88	-3.397	81.5	100	0.556	3.263	0.000
2894	2011-12-27	2.222	13.33	-4.005	87.0	100	0.369	1.842	0.000
2895	2011-12-28	1.810	12.33	-4.682	85.0	100	0.540	3.401	0.203
2896	2011-12-29	2.512	11.92	-4.682	77.2	100	0.546	4.420	0.203
2897	2011-12-30	1.006	11.05	-5.822	79.7	100	0.446	2.832	0.000
2898	2011-12-31	2.263	12.67	-3.938	80.3	100	0.270	1.950	0.000
	Radiation	ET							
2893	9.44	0.5358751							
2894	9.52	0.4386931							
2895	9.59	0.5183545							
2896	9.72	0.5428373							
2897	9.74	0.4614953							
2898	8.11	0.4246270							

Manejo de datos  
con R

Oscar Perpiñán  
Lamigueiro  
[http://  
oscarperpinan.  
github.io](http://oscarperpinan.github.io)

Lectura de datos

Indexado

Cambio de  
formato

Datos agregados

Unión de  
`data.frame`

# Inspeccionamos el resultado

```
summary(aranjuez)
```

Date	TempAvg	TempMax	TempMin
2004-01-01: 1	Min. : -5.309	Min. : -2.362	Min. : -12.980
2004-01-02: 1	1st Qu.: 7.692	1st Qu.: 14.530	1st Qu.: 1.515
2004-01-03: 1	Median : 13.810	Median : 21.670	Median : 7.170
2004-01-04: 1	Mean : 14.405	Mean : 22.531	Mean : 6.888
2004-01-05: 1	3rd Qu.: 21.615	3rd Qu.: 30.875	3rd Qu.: 12.590
2004-01-06: 1	Max. : 30.680	Max. : 41.910	Max. : 22.710
(Other) : 2892			NA's : 4

  

HumidAvg	HumidMax	WindAvg	WindMax
Min. : 19.89	Min. : 35.88	Min. : 0.251	Min. : 0.000
1st Qu.: 47.04	1st Qu.: 81.60	1st Qu.: 0.667	1st Qu.: 3.783
Median : 62.58	Median : 90.90	Median : 0.920	Median : 5.027
Mean : 62.16	Mean : 87.22	Mean : 1.174	Mean : 5.208
3rd Qu.: 77.38	3rd Qu.: 94.90	3rd Qu.: 1.431	3rd Qu.: 6.537
Max. : 100.00	Max. : 100.00	Max. : 8.260	Max. : 10.000
	NA's : 13	NA's : 8	NA's : 128

  

Rain	Radiation	ET
Min. : 0.000	Min. : 0.277	Min. : 0.000
1st Qu.: 0.000	1st Qu.: 9.370	1st Qu.: 1.168
Median : 0.000	Median : 16.660	Median : 2.758
Mean : 1.094	Mean : 16.742	Mean : 3.091
3rd Qu.: 0.200	3rd Qu.: 24.650	3rd Qu.: 4.926
Max. : 49.730	Max. : 32.740	Max. : 8.564
NA's : 4	NA's : 13	NA's : 18

Manejo de datos  
con R

Oscar Perpiñán  
Lamigueiro  
[http://  
oscarperpinan.  
github.io](http://oscarperpinan.github.io)

Lectura de datos

Indexado

Cambio de  
formato

Datos agregados

Unión de  
data.frame

# Valores ausentes

- ▶ NA está definido como logical

```
class(NA)
```

```
[1] "logical"
```

- ▶ Operar con NA siempre produce un NA

```
1 + NA
```

```
[1] NA
```

- ▶ Esto es un «problema» al usar funciones

```
mean(aranjuez$Radiation)
```

```
[1] NA
```

```
mean(aranjuez$Radiation, na.rm = TRUE)
```

```
[1] 16.74176
```

Manejo de datos  
con R

Oscar Perpiñán  
Lamigueiro  
[http://  
oscarperpinan.  
github.io](http://oscarperpinan.github.io)

Lectura de datos

Indexado

Cambio de  
formato

Datos agregados

Unión de  
`data.frame`



# Valores ausentes

Manejo de datos  
con R

Oscar Perpiñán  
Lamigueiro  
[http://  
oscarperpinan.  
github.io](http://oscarperpinan.github.io)

Las funciones `is.na` y `anyNA` los identifican

```
anyNA(aranjuez)
```

```
[1] TRUE
```

```
which(is.na(aranjuez$Radiation))
```

```
[1] 1861 1867 1873 1896 1897 1908 1923 2153 2413 2587 2600 2603 2684
```

```
sum(is.na(aranjuez$Radiation))
```

```
[1] 13
```

Lectura de datos

Indexado

Cambio de  
formato

Datos agregados

Unión de  
`data.frame`

Lectura de datos

**Indexado**

Cambio de formato

Datos agregados

Unión de `data.frame`

# Indexado con []

## ► Filas

```
aranjuez[1:5,]
```

	Date	TempAvg	TempMax	TempMin	HumidAvg	HumidMax	WindAvg	WindMax	Rain
1	2004-01-01	4.044	10.71	-1.969	88.3	95.9	0.746	3.528	0
2	2004-01-02	5.777	11.52	1.247	83.3	98.5	1.078	6.880	0
3	2004-01-03	5.850	13.32	0.377	75.0	94.4	0.979	6.576	0
4	2004-01-04	4.408	15.59	-2.576	82.0	97.0	0.633	3.704	0
5	2004-01-05	3.081	14.58	-2.974	83.2	97.0	0.389	2.244	0
	Radiation	ET							
1	5.490	0.5352688							
2	6.537	0.7710499							
3	8.810	0.8361229							
4	9.790	0.6861381							
5	10.300	0.5152422							

## ► Filas y Columnas

```
aranjuez[10:14, 1:5]
```

	Date	TempAvg	TempMax	TempMin	HumidAvg
10	2004-01-10	10.85	16.59	5.676	84.9
11	2004-01-11	7.59	9.23	4.806	95.4
12	2004-01-12	7.41	10.24	5.200	93.1
13	2004-01-13	8.35	11.38	4.137	91.3
14	2004-01-14	8.74	13.32	2.857	86.9

Manejo de datos  
con R

Oscar Perpiñán  
Lamigueiro  
[http://  
oscarperpinan.  
github.io](http://oscarperpinan.github.io)

Lectura de datos

Indexado

Cambio de  
formato

Datos agregados

Unión de  
`data.frame`

# Indexado con []

Manejo de datos  
con R

Oscar Perpiñán  
Lamigueiro  
[http://  
oscarperpinan.  
github.io](http://oscarperpinan.github.io)

## ► Condición basada en los datos

```
idx <- with(aranjuez, Radiation > 20 & TempAvg < 10)

head(aranjuez[idx, ])
```

	Date	TempAvg	TempMax	TempMin	HumidAvg	HumidMax	WindAvg	WindMax	Rain
82	2004-03-22	9.78	16.12	4.340	51.65	87.9	1.526	7.660	0
83	2004-03-23	8.50	15.52	-0.290	50.10	83.3	1.533	6.027	0
85	2004-03-25	7.47	14.58	1.584	49.66	76.6	1.138	5.939	0
100	2004-04-09	8.83	15.52	2.056	47.50	70.8	1.547	6.125	0
101	2004-04-10	7.04	13.85	-0.155	54.45	85.8	1.448	6.958	0
102	2004-04-11	7.50	15.19	-1.699	54.98	91.0	1.126	7.590	0
	Radiation	ET							
82	21.92	3.075785							
83	20.62	2.881419							
85	22.44	2.849603							
100	25.45	3.566452							
101	21.07	2.943239							
102	20.99	2.905479							

Lectura de datos

Indexado

Cambio de  
formato

Datos agregados

Unión de  
data.frame

# subset

```
subset(aranjuez,  
  subset = (Radiation > 20 & TempAvg < 10),  
  select = c(Radiation, TempAvg,  
    TempMax, TempMin))
```

	Radiation	TempAvg	TempMax	TempMin
82	21.92	9.780	16.12	4.340
83	20.62	8.500	15.52	-0.290
85	22.44	7.470	14.58	1.584
100	25.45	8.830	15.52	2.056
101	21.07	7.040	13.85	-0.155
102	20.99	7.500	15.19	-1.699
104	25.76	9.420	17.47	0.115
461	24.29	7.460	14.66	-0.081
462	25.25	7.930	17.35	-1.686
463	24.56	9.800	19.08	-1.484
1146	20.08	7.170	18.20	-3.746
1157	20.90	4.378	12.03	-6.353
1159	21.87	7.920	18.54	-2.941
1160	20.35	7.830	16.49	-2.807
1521	21.54	8.100	19.29	-4.075
2244	20.49	6.121	15.15	-0.940
2245	21.02	5.989	16.94	-3.208
2246	20.22	9.020	19.74	-2.068
2261	23.00	9.500	14.96	3.662
2262	20.40	9.910	14.70	4.668
2263	24.09	9.440	16.89	0.794
2265	23.64	9.680	16.35	2.938
2295	22.46	8.730	13.84	1.740

Manejo de datos  
con R

Oscar Perpiñán  
Lamigueiro  
[http://  
oscarperpinan.  
github.io](http://oscarperpinan.github.io)

Lectura de datos

Indexado

Cambio de  
formato

Datos agregados

Unión de  
`data.frame`

Lectura de datos

Indexado

Cambio de formato

Datos agregados

Unión de `data.frame`

# Forma simple con stack

```
aranjuezWide <- aranjuez[, c('Radiation',  
                             'TempAvg', 'TempMax',  
                             'WindAvg', 'WindMax')]
```

► Pasamos de formato wide a long

```
aranjuezLong <- stack(aranjuezWide)
```

```
head(aranjuezLong)
```

	values	ind
1	5.490	Radiation
2	6.537	Radiation
3	8.810	Radiation
4	9.790	Radiation
5	10.300	Radiation
6	9.940	Radiation

```
summary(aranjuezLong)
```

	values	ind
Min.	:-5.309	Radiation:2898
1st Qu.:	3.158	TempAvg :2898
Median :	8.720	TempMax :2898
Mean :	12.074	WindAvg :2898
3rd Qu.:	19.970	WindMax :2898
Max.	:41.910	
NA's	:149	

Manejo de datos  
con R

Oscar Perpiñán  
Lamigueiro  
[http://  
oscarperpinan.  
github.io](http://oscarperpinan.github.io)

Lectura de datos

Indexado

Cambio de  
formato

Datos agregados

Unión de  
data.frame

# Más flexible con reshape2

Manejo de datos  
con R

Oscar Perpiñán  
Lamigueiro  
[http://  
oscarperpinan.  
github.io](http://oscarperpinan.github.io)

Lectura de datos

Indexado

Cambio de  
formato

Datos agregados

Unión de  
`data.frame`

- ▶ `reshape2` es un paquete que puede facilitar la transformación de `data.frame` y matrices.

```
library(reshape2)
```



# melt para cambiar de *wide* a *long*

Manejo de datos  
con R

Oscar Perpiñán  
Lamigueiro  
[http://  
oscarperpinan.  
github.io](http://oscarperpinan.github.io)

```
aranjuezLong2 <- melt(aranjuez, id.vars = 'Date',  
                      variable.name = 'Variable',  
                      value.name = 'Value')
```

```
head(aranjuezLong2)
```

	Date	Variable	Value
1	2004-01-01	TempAvg	4.044
2	2004-01-02	TempAvg	5.777
3	2004-01-03	TempAvg	5.850
4	2004-01-04	TempAvg	4.408
5	2004-01-05	TempAvg	3.081
6	2004-01-06	TempAvg	2.304

Lectura de datos

Indexado

Cambio de  
formato

Datos agregados

Unión de  
data.frame

# dcast para cambiar de *long* a *wide*

Manejo de datos  
con R

Oscar Perpiñán  
Lamigueiro  
[http://  
oscarperpinan.  
github.io](http://oscarperpinan.github.io)

```
aranjuezWide2 <- dcast(aranjuezLong2,  
                        Variable ~ Date)  
head(aranjuezWide2[, 1:10])
```

Lectura de datos

Indexado

Cambio de  
formato

Datos agregados

Unión de  
`data.frame`

```
Using Value as value column: use value.var to override.  
Variable 2004-01-01 2004-01-02 2004-01-03 2004-01-04 2004-01-05 2004-01-06  
1 TempAvg      4.044      5.777      5.850      4.408      3.081      2.304  
2 TempMax     10.710     11.520     13.320     15.590     14.580     11.830  
3 TempMin     -1.969      1.247      0.377     -2.576     -2.974     -3.379  
4 HumidAvg     88.300     83.300     75.000     82.000     83.200     84.500  
5 HumidMax     95.900     98.500     94.400     97.000     97.000     96.500  
6 WindAvg       0.746      1.078      0.979      0.633      0.389      0.436  
2004-01-07 2004-01-08 2004-01-09  
1      2.080      6.405     12.060  
2     11.500     13.380     15.330  
3     -3.109     -1.301      9.830  
4     87.000     88.600     86.800  
5     96.600     97.400     94.500  
6      0.449      1.188      2.737
```

Lectura de datos

Indexado

Cambio de formato

Datos agregados

Unión de `data.frame`

# aggregate

Manejo de datos  
con R

Oscar Perpiñán  
Lamigueiro  
[http://  
oscarperpinan.  
github.io](http://oscarperpinan.github.io)

Lectura de datos

Indexado

Cambio de  
formato

Datos agregados

Unión de  
`data.frame`

```
aranjuez$rainy <- aranjuez$Rain > 0
```

```
aggregate(Radiation ~ rainy, data = aranjuez,  
          FUN = mean)
```

```
      rainy Radiation  
1 FALSE    19.63325  
2  TRUE    10.26028
```

# Variable categórica con cut

```
aranjuez$tempClass <- cut(aranjuez$TempAvg, 5)

aggregate(Radiation ~ tempClass, data = aranjuez,
          FUN = mean)
```

```
      tempClass Radiation
1 (-5.34,1.89]  8.805389
2  (1.89,9.09]  9.014178
3  (9.09,16.3] 14.554177
4  (16.3,23.5] 21.912414
5  (23.5,30.7] 26.192742
```

```
aggregate(Radiation ~ tempClass + rainy,
          data = aranjuez, FUN = mean)
```

```
      tempClass rainy Radiation
1 (-5.34,1.89] FALSE  9.869134
2  (1.89,9.09] FALSE 10.718837
3  (9.09,16.3] FALSE 17.238283
4  (16.3,23.5] FALSE 23.238145
5  (23.5,30.7] FALSE 26.392665
6 (-5.34,1.89]  TRUE  6.822955
7  (1.89,9.09]  TRUE  7.063932
8  (9.09,16.3]  TRUE 11.091063
9  (16.3,23.5]  TRUE 15.802522
10 (23.5,30.7]  TRUE 22.545862
```

Manejo de datos  
con R

Oscar Perpiñán  
Lamigueiro  
[http://  
oscarperpinan.  
github.io](http://oscarperpinan.github.io)

Lectura de datos

Indexado

Cambio de  
formato

Datos agregados

Unión de  
data.frame

# Agregamos varias variables

```
aggregate(cbind(Radiation, TempAvg) ~ tempClass,  
          data = aranjuez, FUN = mean)
```

	tempClass	Radiation	TempAvg
1	(-5.34,1.89]	8.805389	0.3423095
2	(1.89,9.09]	9.014178	5.6663267
3	(9.09,16.3]	14.554177	12.5219084
4	(16.3,23.5]	21.912414	19.7486310
5	(23.5,30.7]	26.192742	26.0496953

```
aggregate(cbind(Radiation, TempAvg) ~ tempClass +  
          rainy,  
          data = aranjuez, FUN = mean)
```

	tempClass	rainy	Radiation	TempAvg
1	(-5.34,1.89]	FALSE	9.869134	0.3550122
2	(1.89,9.09]	FALSE	10.718837	5.6657481
3	(9.09,16.3]	FALSE	17.238283	12.6959488
4	(16.3,23.5]	FALSE	23.238145	19.9486604
5	(23.5,30.7]	FALSE	26.392665	26.0896408
6	(-5.34,1.89]	TRUE	6.822955	0.3186364
7	(1.89,9.09]	TRUE	7.063932	5.6669887
8	(9.09,16.3]	TRUE	11.091063	12.2973563
9	(16.3,23.5]	TRUE	15.802522	18.8267565
10	(23.5,30.7]	TRUE	22.545862	25.3210345

Manejo de datos  
con R

Oscar Perpiñán  
Lamigueiro  
[http://  
oscarperpinan.  
github.io](http://oscarperpinan.github.io)

Lectura de datos

Indexado

Cambio de  
formato

Datos agregados

Unión de  
data.frame

# Agregamos a partir de un formato long

```
head(aranjuezLong2)
```

	Date	Variable	Value
1	2004-01-01	TempAvg	4.044
2	2004-01-02	TempAvg	5.777
3	2004-01-03	TempAvg	5.850
4	2004-01-04	TempAvg	4.408
5	2004-01-05	TempAvg	3.081
6	2004-01-06	TempAvg	2.304

```
aggregate(Value ~ Variable, data = aranjuezLong2,  
           FUN = mean)
```

	Variable	Value
1	TempAvg	14.404856
2	TempMax	22.531033
3	TempMin	6.887590
4	HumidAvg	62.160390
5	HumidMax	87.216887
6	WindAvg	1.173983
7	WindMax	5.208021
8	Rain	1.094255
9	Radiation	16.741759
10	ET	3.091191

Manejo de datos  
con R

Oscar Perpiñán  
Lamigueiro  
[http://  
oscarperpinan.  
github.io](http://oscarperpinan.github.io)

Lectura de datos

Indexado

Cambio de  
formato

Datos agregados

Unión de  
data.frame

Lectura de datos

Indexado

Cambio de formato

Datos agregados

Unión de `data.frame`



# Con merge

Manejo de datos  
con R

Oscar Perpiñán  
Lamigueiro  
[http://  
oscarperpinan.  
github.io](http://oscarperpinan.github.io)

- Primero construimos un `data.frame` de ejemplo

```
USStates <- as.data.frame(state.x77)
USStates$Name <- rownames(USStates)
rownames(USStates) <- NULL
```

- Lo partimos en estados «fríos» y estados «grandes»

```
coldStates <- USStates[USStates$Frost>150,  
                      c('Name', 'Frost')]  
largeStates <- USStates[USStates$Area>1e5,  
                      c('Name', 'Area')]
```

Lectura de datos

Indexado

Cambio de  
formato

Datos agregados

Unión de  
`data.frame`

# Con merge

Manejo de datos  
con R

Oscar Perpiñán  
Lamigueiro  
[http://  
oscarperpinan.  
github.io](http://oscarperpinan.github.io)

Lectura de datos

Indexado

Cambio de  
formato

Datos agregados

Unión de  
`data.frame`

- Unimos los dos conjuntos (estados «fríos» y «grandes»)

```
merge(coldStates, largeStates)
```

	Name	Frost	Area
1	Alaska	152	566432
2	Colorado	166	103766
3	Montana	155	145587
4	Nevada	188	109889

# merge usa match

- Estados grandes que también son fríos

```
idxLarge <- match(largeStates$Name,  
                  coldStates$Name,  
                  nomatch=0)  
idxLarge
```

```
[1] 1 0 0 2 5 6 0 0
```

```
coldStates[idxLarge,]
```

	Name	Frost
2	Alaska	152
6	Colorado	166
26	Montana	155
28	Nevada	188

Manejo de datos  
con R

Oscar Perpiñán  
Lamigueiro  
[http://  
oscarperpinan.  
github.io](http://oscarperpinan.github.io)

Lectura de datos

Indexado

Cambio de  
formato

Datos agregados

Unión de  
data.frame

# merge usa match

- Estados frios que también son grandes

```
idxCold <- match(coldStates$Name,  
                 largeStates$Name,  
                 nomatch=0)  
  
idxCold
```

```
[1] 1 4 0 0 5 6 0 0 0 0 0
```

```
largeStates[idxCold,]
```

```
      Name      Area  
2    Alaska 566432  
6   Colorado 103766  
26  Montana 145587  
28   Nevada 109889
```

Manejo de datos  
con R

Oscar Perpiñán  
Lamigueiro  
[http://  
oscarperpinan.  
github.io](http://oscarperpinan.github.io)

Lectura de datos

Indexado

Cambio de  
formato

Datos agregados

Unión de  
data.frame