R para Ciencia de Datos: Hoja de Referencia

data.table

Aprende R para Ciencia de Datos en www.datademia.es

Datademia

data.table

data.table es un paquete de R que proporciona una versión de alto rendimiento del marco de datos de base R con sintaxis y mejoras de funciones para facilitar su uso, comodidad y velocidad de programación.

Carga el paquete:

```
> library(data.table)
```

Crea un data.table DT

```
DT <- data.table(V1=c(1L,2L),
                 V2=LETTERS[1:3],
                 V3=round(rnorm(4),4),
                 V4=1:12)
```

Forma general: DT[i, j, by]

"En DT, selecciona filas usando **i**, luego calcula **j** y agrupa por **by**"

Selecciona filas usando i

```
> DT[3:5,]                    Selecciona de la 3ra a la 5ta fila
> DT[3:5]                     Selecciona de la 3ra a la 5ta fila
> DT[V2=="A"]                 Selecciona filas que tienen el valor A en la columna V2
> DT[V2 %in% c("A","C")]      Selecciona filas que tengan el valor A o C en la columna V2
```

Manipula columnas usando j

```
> DT[,V2]                                  Devuelve V2 como un vector
[1] "A" "B" "C" "A" "B" "C" ...
> DT[,.(V2,V3)]                            Devuelve V2 y V3 como un data.table
> DT[,sum(V1)]                             Devuelve la suma de todos los elementos de
[1] 18                                     V1 en un vector
> DT[,.(sum(V1),sd(V3))]                   Devuelve la suma de todos los elementos de
   V1        V2                            V1 y la desviación estándar de V3 en una
1: 18 0.4546055                            tabla de datos.

>DT[,.(Aggregate=sum(V1),                  Igual que el anterior, con nuevos nombres.
       Sd.V3=sd(V3))]
  Aggregate     Sd.V3
1:        18 0.4546055

> DT[,.(V1,Sd.V3=sd(V3))]                  Seleccione la columna V2 y calcula la
                                           desviación estándar de V3, que devuelve un
                                           valor único y se recicla
> DT[,.(print(V2), plot(V3), NULL)]        Imprime la columna V2 y traza V3
```

Usando j por grupo

```
> DT[,.(V4.Sum=sum(V4)),by=V1]             Calcula la suma de V4 para cada grupo en V1
   V1 V4.Sum
1:  1     36
2:  2     42
>DT[,.(V4.Sum=sum(V4)),                    Calcula la suma de V4 para cada grupo en V1
       by=.(V1,V2)]                        y V2
>DT[,.(V4.Sum=sum(V4)),                    Calcula la suma de V4 para cada grupo en el
       by=sign(V1-1)]                      signo (V1-1)
  sign V4.Sum
1:    0     36
2:    1     42
> DT[,.(V4.Sum=sum(V4)),                   Igual que el anterior, con un nuevo nombre
       by=.(V1.01=sign(V1-1))]             para la variable por la que está agrupando
> DT[1:5,.(V4.Sum=sum(V4)), by=V1]         Calcula la suma de V4 para cada grupo en V1
                                           después de seleccionar las primeras 5 filas
> DT[,.N,by=V1]                            Cuenta el número de filas para cada grupo en V1
```

Añadiendo/Actualizando Columnas por referencia en j usando :=

```
> DT[,V1:=round(exp(V1),2)]                V1 se actualiza por lo que está después: =
> DT                                       Devuelve el resultado llamando a DT
 V1 V2 V3 V4
 1: 2.72 A -0.1107 1
 2: 7.39 B -0.1427 2
 3: 2.72 C -1.8893 3
 4: 7.39 A -0.3571 4
 ...
> DT[,c("V1","V2"):=list(round(exp(  Las columnas V1 y V2 se actualizan con lo que
V1),2),LETTERS[4:6])]                      está después de :=
> DT[,':='(V1=round(exp(V1),2),            Alternativa a la anterior. Con [],
 V2=LETTERS[4:6])][]                       imprime el resultado en la pantalla
          V1  V2      V3  V4
 1:    15.18   D -0.1107   1
 2: 1619.71   E -0.1427   2
 3:    15.18   F -1.8893   3
 4: 1619.71   D -0.3571   4
> DT[,V1:=NULL]                            Elimina V1
> DT[,c("V1","V2"):=NULL]                  Elimina las columnas V1 y V2
> Cols.chosen=c("A","B")
> DT[,Cols.Chosen:=NULL]                   Elimina la columna con el nombre Cols.chosen
                                           Eliminar las columnas especificadas en la
> DT[,(Cols.Chosen):=NULL]                 variable Cols.chosen
```

Indices y claves

```
> setkey(DT,V2)                            Establece una clave en V2
> DT["A"]                                  Devuelve todas las filas donde la columna clave
   V1 V2      V3  V4                       (establecida a V2) tiene el valor A
 1: 1  A -0.2392   1
 2: 2  A -1.6148   4
> DT[c("A","C")]                           Devuelve todas las filas donde la columna clave
                                           (V2) tiene el valor A o C
> DT["A",mult="first"]                     Devuelve la primera fila de todas las filas que
                                           coinciden con el valor A en la columna clave V2
> DT["A",mult="last"]                      Devuelve la última fila de todas las filas que
                                           coinciden con el valor A en la columna clave V2
> DT[c("A","D")]                           Devuelve todas las filas donde la columna clave
   V1 V2      V3  V4                       V2 tiene el valor A o D
 1: 1  A -0.2392   1
 2: 2  A -1.6148   4
 3: NA D      NA  NA
> DT[c("A","D"),nomatch=0]                 Devuelve todas las filas donde la columna clave
 V1 V2 V3 V4                               V2 tiene el valor A o D
 1: 1 A -0.2392 1
 2: 2 A -1.6148 4
> DT[c("A","C"),sum(V4)]                   Devuelve la suma total de V4, para las filas de la
                                           columna clave V2 que tengan valores A o C
> DT[c("A","C"),sum(V4),by=.EACHI]         Devuelve la suma de la columna V4 para las filas
  V2 V1                                    de V2 que tienen el valor A, y otra suma para las
 1: A 22                                   filas de V2 que tienen el valor C
 2: C 30

> setkey(DT,V1,V2)                         Ordenar por V1 y luego por V2 dentro de cada
                                           grupo de V1 (invisible)
> DT[.(2,"C")]                             Seleccione filas que tengan el valor 2 para la
  V1 V2      V3 V4                         primera clave (V1) y el valor C para la segunda
 1: 2  C  0.3262  6                        clave (V2)
 2: 2  C -1.6148 12
> DT[.(2,c("A","C"))]                      Seleccione filas que tengan el valor 2 para la
   V1 V2      V3 V4                        primera clave (V1) y dentro de
 1: 2  A -1.6148  4                        esas filas el valor A o C para la segunda clave
 2: 2  A  0.3262 10                        (V2)
 3: 2  C  0.3262  6
```

Familia set

```
> setnames(DT,"V2","Rating")               Cambia el nombre V2 a Rating (invisible)
> setnames(DT,                             Cambia 2 nombres de columnas (invisible)
           c("V2","V3"),
           c("V2.rating","V3.DC"))
> setcolorder(DT,
          c("V2","V1","V4","V3"))          Cambia el orden  de las columnas usando el
                                           vector escogido (invisible)
```

Operaciones avanzadas de data table

```
> DT[.N-1]                                 Devuelve la penúltima fila del DT
> DT[,.N]                                  Devuelve el número de filas
> DT[,.(V2,V3)]                            Devuelve V2 y V3 como data.table
> DT[,list(V2,V3)]                         Devuelve V2 y V3 como data.table
> DT[,mean(V3),by=.(V1,V2)]                Devuelve el resultado de j, agrupado por todas
 V1 V2         V3                          las combinaciones posibles de los grupos
 1:  1  A  0.4053                          especificados en by
 2:  1  B  0.4053
 3:  1  C  0.4053
 4:  2  A -0.6443
 5:  2  B -0.6443
 6:  2  C -0.6443
```

.SD & .SDcols

```
> DT[,print(.SD),by=V2]                    Mira lo que contiene .SD
> DT[,.SD[c(1,.N)],by=V2]                  Selecciona la primera y la última fila
                                           agrupadas por V2
> DT[,lapply(.SD,sum),by=V2]               Calcula la suma de columnas en .SD
                                           agrupadas por V2
> DT[,lapply(.SD,sum),by=V2,               Calcula la suma de V3 y V4 en .SD agrupados
.SDcols=c("V3","V4")]                      por V2
   V2      V3 V4
 1: A -0.478 22
 2: B -0.478 26                            Calcula la suma de V3 y V4 en .SD agrupados
> DT[,lapply(.SD,sum),by=V2,               por V2
.SDcols=paste0("V",3:4)]
```

Encadenamiento

```
> DT <- DT[,.(V4.Sum=sum(V4)),             Calcula la suma de V4, agrupado por V1
           by =V1]
   V1 V4.Sum
1: 1     36
2: 2     42

> DT[V4.Sum>40]                            Selecciona el grupo del cual la suma es> 40
> DT[,.(V4.Sum=sum(V4)),                   Seleccione el grupo del cual la suma es> 40
       by=V1][V4.Sum>40]                   (encadenamiento)
   V1 V4.Sum
1: 2     42
> DT[,.(V4.Sum=sum(V4)),                   Calcula la suma de V4, agrupado por V1,
       by=V1][order(-V1)]                  ordenado por V1
   V1 V4.Sum
1: 2     42
2: 1     36
```

Datademia

Aprende R para Ciencia de Datos en www.datademia.es