Ejercicios Dataframes

David Criado Ramón

10/11/2019

Introducción a R (4)

1. Acceso y selección de secciones de un data frame

La sintaxis general para acceder a un data frame es my_frame[rows, columns]. Vamos a trabajar con un ejemplo que viene por defecto con R USArrests. Este data frame contiene la información por cada estado Americano de las tasas de criminales (por 100.000 habitantes). Los datos de las columnas se refieren a asesinatos, violaciones y porcentaje de la población que vive en áreas urbanas. Los datos son de 1973. Contesta a las siguientes preguntas sobre los datos:

- Las dimensiones del dataframe.

```
dim(USArrests)
```

```
## [1] 50  4
```

El dataframe es bidimensional con 50 muestras (filas) y 4 variables (columnas).

- La longitud del dataframe (filas o columnas).

```
length(USArrests)
```

```
## [1] 4
```

La longitud devuelve el número de columnas.

- Número de columnas.

```
ncol(USArrests)
```

```
## [1] 4
```

- ¿Cómo calcularías el número de filas?

```
nrow(USArrests)
```

```
## [1] 50
```

- Obtén el nombre de las filas y las columnas para este data frame.

```
rownames(USArrests)
```

```
##  [1] "Alabama"        "Alaska"         "Arizona"        "Arkansas"
##  [5] "California"     "Colorado"       "Connecticut"    "Delaware"
##  [9] "Florida"        "Georgia"        "Hawaii"         "Idaho"
## [13] "Illinois"       "Indiana"        "Iowa"           "Kansas"
## [17] "Kentucky"       "Louisiana"      "Maine"          "Maryland"
## [21] "Massachusetts"  "Michigan"       "Minnesota"      "Mississippi"
## [25] "Missouri"       "Montana"        "Nebraska"       "Nevada"
## [29] "New Hampshire"  "New Jersey"     "New Mexico"     "New York"
## [33] "North Carolina" "North Dakota"   "Ohio"           "Oklahoma"
## [37] "Oregon"         "Pennsylvania"   "Rhode Island"   "South Carolina"
## [41] "South Dakota"   "Tennessee"      "Texas"          "Utah"
```

```
## [45] "Vermont"        "Virginia"       "Washington"     "West Virginia"
## [49] "Wisconsin"      "Wyoming"
```

```
colnames(USArrests)
```

```
## [1] "Murder"   "Assault"  "UrbanPop" "Rape"
```

Échale un vistazo a los datos, por ejemplo, a las 6 primeras filas.

```
head(USArrests, 6)
```

```
##            Murder Assault UrbanPop Rape
## Alabama      13.2     236       58 21.2
## Alaska       10.0     263       48 44.5
## Arizona       8.1     294       80 31.0
## Arkansas      8.8     190       50 19.5
## California    9.0     276       91 40.6
## Colorado      7.9     204       78 38.7
```

Ordena de forma decreciente las filas de nuestro data frame según el porcentaje de población en área urbana. Para ello investiga la función order() y sus parámetros.

```
x <- USArrests
x <- x[order(x$UrbanPop, decreasing=T),]
head(x, )
```

```
##               Murder Assault UrbanPop Rape
## California       9.0     276       91 40.6
## New Jersey       7.4     159       89 18.8
## Rhode Island     3.4     174       87  8.3
## New York        11.1     254       86 26.1
## Massachusetts    4.4     149       85 16.3
## Hawaii           5.3      46       83 20.2
```

¿Podrías añadir un segundo criterio de orden?, ¿cómo?

```
y <- USArrests
y <- y[order(y$UrbanPop, y$Murder, decreasing=T),]
head(y, 10)
```

```
##               Murder Assault UrbanPop Rape
## California       9.0     276       91 40.6
## New Jersey       7.4     159       89 18.8
## Rhode Island     3.4     174       87  8.3
## New York        11.1     254       86 26.1
## Massachusetts    4.4     149       85 16.3
## Illinois        10.4     249       83 24.0
## Hawaii           5.3      46       83 20.2
## Nevada          12.2     252       81 46.0
## Florida         15.4     335       80 31.9
## Texas           12.7     201       80 25.5
```

Sí, podemos realizarlo indicando las columnas del dataframe en orden de preferencia para ordenar.

Muestra por pantalla la columna con los datos de asesinato.

```
USArrests$Murder
```

```
##  [1] 13.2 10.0  8.1  8.8  9.0  7.9  3.3  5.9 15.4 17.4  5.3  2.6 10.4  7.2
## [15]  2.2  6.0  9.7 15.4  2.1 11.3  4.4 12.1  2.7 16.1  9.0  6.0  4.3 12.2
## [29]  2.1  7.4 11.4 11.1 13.0  0.8  7.3  6.6  4.9  6.3  3.4 14.4  3.8 13.2
## [43] 12.7  3.2  2.2  8.5  4.0  5.7  2.6  6.8
```

Muestra las tasas de asesinato para el segundo, tercer y cuarto estado.

```
USArrests$Murder[2:4]
```

```
## [1] 10.0  8.1  8.8
```

Muestra las primeras cinco filas de todas las columnas.

```
USArrests[1:5, ]
```

```
##            Murder Assault UrbanPop Rape
## Alabama      13.2     236       58 21.2
## Alaska       10.0     263       48 44.5
## Arizona       8.1     294       80 31.0
## Arkansas      8.8     190       50 19.5
## California    9.0     276       91 40.6
```

Muestra todas las filas para las dos primeras columnas.

```
USArrests[,1:2]
```

```
##                Murder Assault
## Alabama          13.2     236
## Alaska           10.0     263
## Arizona           8.1     294
## Arkansas          8.8     190
## California        9.0     276
## Colorado          7.9     204
## Connecticut       3.3     110
## Delaware          5.9     238
## Florida          15.4     335
## Georgia          17.4     211
## Hawaii            5.3      46
## Idaho             2.6     120
## Illinois         10.4     249
## Indiana           7.2     113
## Iowa              2.2      56
## Kansas            6.0     115
## Kentucky          9.7     109
## Louisiana        15.4     249
## Maine             2.1      83
## Maryland         11.3     300
## Massachusetts     4.4     149
## Michigan         12.1     255
## Minnesota         2.7      72
## Mississippi      16.1     259
```

```
## Missouri          9.0     178
## Montana           6.0     109
## Nebraska          4.3     102
## Nevada           12.2     252
## New Hampshire     2.1      57
## New Jersey        7.4     159
## New Mexico       11.4     285
## New York         11.1     254
## North Carolina   13.0     337
## North Dakota      0.8      45
## Ohio              7.3     120
## Oklahoma          6.6     151
## Oregon            4.9     159
## Pennsylvania      6.3     106
## Rhode Island      3.4     174
## South Carolina   14.4     279
## South Dakota      3.8      86
## Tennessee        13.2     188
## Texas            12.7     201
## Utah              3.2     120
## Vermont           2.2      48
## Virginia          8.5     156
## Washington        4.0     145
## West Virginia     5.7      81
## Wisconsin         2.6      53
## Wyoming           6.8     161
```

Muestra todas las filas de las columnas 1 y 3.

```
USArrests[,c(1,3)]
```

```
##                Murder UrbanPop
## Alabama          13.2       58
## Alaska           10.0       48
## Arizona           8.1       80
## Arkansas          8.8       50
## California        9.0       91
## Colorado          7.9       78
## Connecticut       3.3       77
## Delaware          5.9       72
## Florida          15.4       80
## Georgia          17.4       60
## Hawaii            5.3       83
## Idaho             2.6       54
## Illinois         10.4       83
## Indiana           7.2       65
## Iowa              2.2       57
## Kansas            6.0       66
## Kentucky          9.7       52
## Louisiana        15.4       66
## Maine             2.1       51
## Maryland         11.3       67
## Massachusetts     4.4       85
## Michigan         12.1       74
```

```
## Minnesota         2.7       66
## Mississippi      16.1       44
## Missouri          9.0       70
## Montana           6.0       53
## Nebraska          4.3       62
## Nevada           12.2       81
## New Hampshire     2.1       56
## New Jersey        7.4       89
## New Mexico       11.4       70
## New York         11.1       86
## North Carolina   13.0       45
## North Dakota      0.8       44
## Ohio              7.3       75
## Oklahoma          6.6       68
## Oregon            4.9       67
## Pennsylvania      6.3       72
## Rhode Island      3.4       87
## South Carolina   14.4       48
## South Dakota      3.8       45
## Tennessee        13.2       59
## Texas            12.7       80
## Utah              3.2       80
## Vermont           2.2       32
## Virginia          8.5       63
## Washington        4.0       73
## West Virginia     5.7       39
## Wisconsin         2.6       66
## Wyoming           6.8       60
```

Muestra sólo las primeras cinco filas de las columnas 1 y 2.

```
USArrests[1:5,1:2]
```

```
##            Murder Assault
## Alabama      13.2     236
## Alaska       10.0     263
## Arizona       8.1     294
## Arkansas      8.8     190
## California    9.0     276
```

Extrae las filas para el índice Murder

```
USArrests$Murder
```

```
##  [1] 13.2 10.0  8.1  8.8  9.0  7.9  3.3  5.9 15.4 17.4  5.3  2.6 10.4  7.2
## [15]  2.2  6.0  9.7 15.4  2.1 11.3  4.4 12.1  2.7 16.1  9.0  6.0  4.3 12.2
## [29]  2.1  7.4 11.4 11.1 13.0  0.8  7.3  6.6  4.9  6.3  3.4 14.4  3.8 13.2
## [43] 12.7  3.2  2.2  8.5  4.0  5.7  2.6  6.8
```

Vamos con expresiones más complicadas.

¿Qué estado tiene la menor tasa de asesinatos?, ¿qué línea contiene esa información?, obtén esa información.

```
rownames(USArrests)[which.min(USArrests$Murder)]
```

```
## [1] "North Dakota"
```

¿QUé estados tienen una tasa inferior al 4 %?, obtén esa información.

```
rownames(USArrests)[USArrests$Murder < 4]
```

```
##  [1] "Connecticut"   "Idaho"          "Iowa"          "Maine"
##  [5] "Minnesota"     "New Hampshire" "North Dakota"  "Rhode Island"
##  [9] "South Dakota"  "Utah"           "Vermont"       "Wisconsin"
```

¿Qué estados están en el cuartil superior (75) en los que a población en zonas urbanas se refiere?

```
rownames(USArrests)[USArrests$UrbanPop > 75]
```

```
##  [1] "Arizona"       "California"    "Colorado"      "Connecticut"
##  [5] "Florida"       "Hawaii"        "Illinois"      "Massachusetts"
##  [9] "Nevada"        "New Jersey"    "New York"      "Rhode Island"
## [13] "Texas"         "Utah"
```

Carga el set de datos co2 y realiza las siguientes acciones:

- a) Ordena alfabéticamente los datos en función de la variable Plant. Recuerda que Plant es un factor. Imprime el resultado por pantalla para comprobarlo.

```
CO2[order(as.character(CO2$Plant)),]
```

```
##    Plant        Type  Treatment conc uptake
## 64   Mc1 Mississippi    chilled   95   10.5
## 65   Mc1 Mississippi    chilled  175   14.9
## 66   Mc1 Mississippi    chilled  250   18.1
## 67   Mc1 Mississippi    chilled  350   18.9
## 68   Mc1 Mississippi    chilled  500   19.5
## 69   Mc1 Mississippi    chilled  675   22.2
## 70   Mc1 Mississippi    chilled 1000   21.9
## 71   Mc2 Mississippi    chilled   95    7.7
## 72   Mc2 Mississippi    chilled  175   11.4
## 73   Mc2 Mississippi    chilled  250   12.3
## 74   Mc2 Mississippi    chilled  350   13.0
## 75   Mc2 Mississippi    chilled  500   12.5
## 76   Mc2 Mississippi    chilled  675   13.7
## 77   Mc2 Mississippi    chilled 1000   14.4
## 78   Mc3 Mississippi    chilled   95   10.6
## 79   Mc3 Mississippi    chilled  175   18.0
## 80   Mc3 Mississippi    chilled  250   17.9
## 81   Mc3 Mississippi    chilled  350   17.9
## 82   Mc3 Mississippi    chilled  500   17.9
## 83   Mc3 Mississippi    chilled  675   18.9
## 84   Mc3 Mississippi    chilled 1000   19.9
```

```
## 43   Mn1 Mississippi nonchilled   95   10.6
## 44   Mn1 Mississippi nonchilled  175   19.2
## 45   Mn1 Mississippi nonchilled  250   26.2
## 46   Mn1 Mississippi nonchilled  350   30.0
## 47   Mn1 Mississippi nonchilled  500   30.9
## 48   Mn1 Mississippi nonchilled  675   32.4
## 49   Mn1 Mississippi nonchilled 1000   35.5
## 50   Mn2 Mississippi nonchilled   95   12.0
## 51   Mn2 Mississippi nonchilled  175   22.0
## 52   Mn2 Mississippi nonchilled  250   30.6
## 53   Mn2 Mississippi nonchilled  350   31.8
## 54   Mn2 Mississippi nonchilled  500   32.4
## 55   Mn2 Mississippi nonchilled  675   31.1
## 56   Mn2 Mississippi nonchilled 1000   31.5
## 57   Mn3 Mississippi nonchilled   95   11.3
## 58   Mn3 Mississippi nonchilled  175   19.4
## 59   Mn3 Mississippi nonchilled  250   25.8
## 60   Mn3 Mississippi nonchilled  350   27.9
## 61   Mn3 Mississippi nonchilled  500   28.5
## 62   Mn3 Mississippi nonchilled  675   28.1
## 63   Mn3 Mississippi nonchilled 1000   27.8
## 22   Qc1      Quebec    chilled   95   14.2
## 23   Qc1      Quebec    chilled  175   24.1
## 24   Qc1      Quebec    chilled  250   30.3
## 25   Qc1      Quebec    chilled  350   34.6
## 26   Qc1      Quebec    chilled  500   32.5
## 27   Qc1      Quebec    chilled  675   35.4
## 28   Qc1      Quebec    chilled 1000   38.7
## 29   Qc2      Quebec    chilled   95    9.3
## 30   Qc2      Quebec    chilled  175   27.3
## 31   Qc2      Quebec    chilled  250   35.0
## 32   Qc2      Quebec    chilled  350   38.8
## 33   Qc2      Quebec    chilled  500   38.6
## 34   Qc2      Quebec    chilled  675   37.5
## 35   Qc2      Quebec    chilled 1000   42.4
## 36   Qc3      Quebec    chilled   95   15.1
## 37   Qc3      Quebec    chilled  175   21.0
## 38   Qc3      Quebec    chilled  250   38.1
## 39   Qc3      Quebec    chilled  350   34.0
## 40   Qc3      Quebec    chilled  500   38.9
## 41   Qc3      Quebec    chilled  675   39.6
## 42   Qc3      Quebec    chilled 1000   41.4
## 1    Qn1      Quebec nonchilled   95   16.0
## 2    Qn1      Quebec nonchilled  175   30.4
## 3    Qn1      Quebec nonchilled  250   34.8
## 4    Qn1      Quebec nonchilled  350   37.2
## 5    Qn1      Quebec nonchilled  500   35.3
## 6    Qn1      Quebec nonchilled  675   39.2
## 7    Qn1      Quebec nonchilled 1000   39.7
## 8    Qn2      Quebec nonchilled   95   13.6
## 9    Qn2      Quebec nonchilled  175   27.3
## 10   Qn2      Quebec nonchilled  250   37.1
## 11   Qn2      Quebec nonchilled  350   41.8
## 12   Qn2      Quebec nonchilled  500   40.6
```

```
## 13   Qn2      Quebec nonchilled  675   41.4
## 14   Qn2      Quebec nonchilled 1000   44.3
## 15   Qn3      Quebec nonchilled   95   16.2
## 16   Qn3      Quebec nonchilled  175   32.4
## 17   Qn3      Quebec nonchilled  250   40.3
## 18   Qn3      Quebec nonchilled  350   42.1
## 19   Qn3      Quebec nonchilled  500   42.9
## 20   Qn3      Quebec nonchilled  675   43.9
## 21   Qn3      Quebec nonchilled 1000   45.5
```

- b) Ordena los datos en función del incremento de la variable uptake y el orden alfabético de la planta (en ese orden).

```
CO2[order(CO2$uptake, as.character(CO2$Plant)),]
```

```
##    Plant        Type  Treatment conc uptake
## 71   Mc2 Mississippi    chilled   95    7.7
## 29   Qc2      Quebec    chilled   95    9.3
## 64   Mc1 Mississippi    chilled   95   10.5
## 78   Mc3 Mississippi    chilled   95   10.6
## 43   Mn1 Mississippi nonchilled   95   10.6
## 57   Mn3 Mississippi nonchilled   95   11.3
## 72   Mc2 Mississippi    chilled  175   11.4
## 50   Mn2 Mississippi nonchilled   95   12.0
## 73   Mc2 Mississippi    chilled  250   12.3
## 75   Mc2 Mississippi    chilled  500   12.5
## 74   Mc2 Mississippi    chilled  350   13.0
## 8    Qn2      Quebec nonchilled   95   13.6
## 76   Mc2 Mississippi    chilled  675   13.7
## 22   Qc1      Quebec    chilled   95   14.2
## 77   Mc2 Mississippi    chilled 1000   14.4
## 65   Mc1 Mississippi    chilled  175   14.9
## 36   Qc3      Quebec    chilled   95   15.1
## 1    Qn1      Quebec nonchilled   95   16.0
## 15   Qn3      Quebec nonchilled   95   16.2
## 80   Mc3 Mississippi    chilled  250   17.9
## 81   Mc3 Mississippi    chilled  350   17.9
## 82   Mc3 Mississippi    chilled  500   17.9
## 79   Mc3 Mississippi    chilled  175   18.0
## 66   Mc1 Mississippi    chilled  250   18.1
## 67   Mc1 Mississippi    chilled  350   18.9
## 83   Mc3 Mississippi    chilled  675   18.9
## 44   Mn1 Mississippi nonchilled  175   19.2
## 58   Mn3 Mississippi nonchilled  175   19.4
## 68   Mc1 Mississippi    chilled  500   19.5
## 84   Mc3 Mississippi    chilled 1000   19.9
## 37   Qc3      Quebec    chilled  175   21.0
## 70   Mc1 Mississippi    chilled 1000   21.9
## 51   Mn2 Mississippi nonchilled  175   22.0
## 69   Mc1 Mississippi    chilled  675   22.2
## 23   Qc1      Quebec    chilled  175   24.1
## 59   Mn3 Mississippi nonchilled  250   25.8
## 45   Mn1 Mississippi nonchilled  250   26.2
## 30   Qc2      Quebec    chilled  175   27.3
## 9    Qn2      Quebec nonchilled  175   27.3
```

```
## 63   Mn3 Mississippi nonchilled 1000   27.8
## 60   Mn3 Mississippi nonchilled  350   27.9
## 62   Mn3 Mississippi nonchilled  675   28.1
## 61   Mn3 Mississippi nonchilled  500   28.5
## 46   Mn1 Mississippi nonchilled  350   30.0
## 24   Qc1      Quebec    chilled  250   30.3
## 2    Qn1      Quebec nonchilled  175   30.4
## 52   Mn2 Mississippi nonchilled  250   30.6
## 47   Mn1 Mississippi nonchilled  500   30.9
## 55   Mn2 Mississippi nonchilled  675   31.1
## 56   Mn2 Mississippi nonchilled 1000   31.5
## 53   Mn2 Mississippi nonchilled  350   31.8
## 48   Mn1 Mississippi nonchilled  675   32.4
## 54   Mn2 Mississippi nonchilled  500   32.4
## 16   Qn3      Quebec nonchilled  175   32.4
## 26   Qc1      Quebec    chilled  500   32.5
## 39   Qc3      Quebec    chilled  350   34.0
## 25   Qc1      Quebec    chilled  350   34.6
## 3    Qn1      Quebec nonchilled  250   34.8
## 31   Qc2      Quebec    chilled  250   35.0
## 5    Qn1      Quebec nonchilled  500   35.3
## 27   Qc1      Quebec    chilled  675   35.4
## 49   Mn1 Mississippi nonchilled 1000   35.5
## 10   Qn2      Quebec nonchilled  250   37.1
## 4    Qn1      Quebec nonchilled  350   37.2
## 34   Qc2      Quebec    chilled  675   37.5
## 38   Qc3      Quebec    chilled  250   38.1
## 33   Qc2      Quebec    chilled  500   38.6
## 28   Qc1      Quebec    chilled 1000   38.7
## 32   Qc2      Quebec    chilled  350   38.8
## 40   Qc3      Quebec    chilled  500   38.9
## 6    Qn1      Quebec nonchilled  675   39.2
## 41   Qc3      Quebec    chilled  675   39.6
## 7    Qn1      Quebec nonchilled 1000   39.7
## 17   Qn3      Quebec nonchilled  250   40.3
## 12   Qn2      Quebec nonchilled  500   40.6
## 42   Qc3      Quebec    chilled 1000   41.4
## 13   Qn2      Quebec nonchilled  675   41.4
## 11   Qn2      Quebec nonchilled  350   41.8
## 18   Qn3      Quebec nonchilled  350   42.1
## 35   Qc2      Quebec    chilled 1000   42.4
## 19   Qn3      Quebec nonchilled  500   42.9
## 20   Qn3      Quebec nonchilled  675   43.9
## 14   Qn2      Quebec nonchilled 1000   44.3
## 21   Qn3      Quebec nonchilled 1000   45.5
```

- c) Ordena de nuevo los datos en función del increment de la variable uptake y el orden alfabético reverso de la planta (en ese orden).

```
CO2[c(order(CO2$uptake), order(as.character(CO2$Plant), decreasing=T)),]
```

```
##      Plant        Type  Treatment conc uptake
## 71     Mc2 Mississippi    chilled   95    7.7
## 29     Qc2      Quebec    chilled   95    9.3
## 64     Mc1 Mississippi    chilled   95   10.5
```

```
## 43     Mn1 Mississippi nonchilled   95   10.6
## 78     Mc3 Mississippi    chilled   95   10.6
## 57     Mn3 Mississippi nonchilled   95   11.3
## 72     Mc2 Mississippi    chilled  175   11.4
## 50     Mn2 Mississippi nonchilled   95   12.0
## 73     Mc2 Mississippi    chilled  250   12.3
## 75     Mc2 Mississippi    chilled  500   12.5
## 74     Mc2 Mississippi    chilled  350   13.0
## 8      Qn2      Quebec nonchilled   95   13.6
## 76     Mc2 Mississippi    chilled  675   13.7
## 22     Qc1      Quebec    chilled   95   14.2
## 77     Mc2 Mississippi    chilled 1000   14.4
## 65     Mc1 Mississippi    chilled  175   14.9
## 36     Qc3      Quebec    chilled   95   15.1
## 1      Qn1      Quebec nonchilled   95   16.0
## 15     Qn3      Quebec nonchilled   95   16.2
## 80     Mc3 Mississippi    chilled  250   17.9
## 81     Mc3 Mississippi    chilled  350   17.9
## 82     Mc3 Mississippi    chilled  500   17.9
## 79     Mc3 Mississippi    chilled  175   18.0
## 66     Mc1 Mississippi    chilled  250   18.1
## 67     Mc1 Mississippi    chilled  350   18.9
## 83     Mc3 Mississippi    chilled  675   18.9
## 44     Mn1 Mississippi nonchilled  175   19.2
## 58     Mn3 Mississippi nonchilled  175   19.4
## 68     Mc1 Mississippi    chilled  500   19.5
## 84     Mc3 Mississippi    chilled 1000   19.9
## 37     Qc3      Quebec    chilled  175   21.0
## 70     Mc1 Mississippi    chilled 1000   21.9
## 51     Mn2 Mississippi nonchilled  175   22.0
## 69     Mc1 Mississippi    chilled  675   22.2
## 23     Qc1      Quebec    chilled  175   24.1
## 59     Mn3 Mississippi nonchilled  250   25.8
## 45     Mn1 Mississippi nonchilled  250   26.2
## 9      Qn2      Quebec nonchilled  175   27.3
## 30     Qc2      Quebec    chilled  175   27.3
## 63     Mn3 Mississippi nonchilled 1000   27.8
## 60     Mn3 Mississippi nonchilled  350   27.9
## 62     Mn3 Mississippi nonchilled  675   28.1
## 61     Mn3 Mississippi nonchilled  500   28.5
## 46     Mn1 Mississippi nonchilled  350   30.0
## 24     Qc1      Quebec    chilled  250   30.3
## 2      Qn1      Quebec nonchilled  175   30.4
## 52     Mn2 Mississippi nonchilled  250   30.6
## 47     Mn1 Mississippi nonchilled  500   30.9
## 55     Mn2 Mississippi nonchilled  675   31.1
## 56     Mn2 Mississippi nonchilled 1000   31.5
## 53     Mn2 Mississippi nonchilled  350   31.8
## 16     Qn3      Quebec nonchilled  175   32.4
## 48     Mn1 Mississippi nonchilled  675   32.4
## 54     Mn2 Mississippi nonchilled  500   32.4
## 26     Qc1      Quebec    chilled  500   32.5
## 39     Qc3      Quebec    chilled  350   34.0
## 25     Qc1      Quebec    chilled  350   34.6
```

```
## 3      Qn1      Quebec nonchilled  250   34.8
## 31     Qc2      Quebec    chilled  250   35.0
## 5      Qn1      Quebec nonchilled  500   35.3
## 27     Qc1      Quebec    chilled  675   35.4
## 49     Mn1 Mississippi nonchilled 1000   35.5
## 10     Qn2      Quebec nonchilled  250   37.1
## 4      Qn1      Quebec nonchilled  350   37.2
## 34     Qc2      Quebec    chilled  675   37.5
## 38     Qc3      Quebec    chilled  250   38.1
## 33     Qc2      Quebec    chilled  500   38.6
## 28     Qc1      Quebec    chilled 1000   38.7
## 32     Qc2      Quebec    chilled  350   38.8
## 40     Qc3      Quebec    chilled  500   38.9
## 6      Qn1      Quebec nonchilled  675   39.2
## 41     Qc3      Quebec    chilled  675   39.6
## 7      Qn1      Quebec nonchilled 1000   39.7
## 17     Qn3      Quebec nonchilled  250   40.3
## 12     Qn2      Quebec nonchilled  500   40.6
## 13     Qn2      Quebec nonchilled  675   41.4
## 42     Qc3      Quebec    chilled 1000   41.4
## 11     Qn2      Quebec nonchilled  350   41.8
## 18     Qn3      Quebec nonchilled  350   42.1
## 35     Qc2      Quebec    chilled 1000   42.4
## 19     Qn3      Quebec nonchilled  500   42.9
## 20     Qn3      Quebec nonchilled  675   43.9
## 14     Qn2      Quebec nonchilled 1000   44.3
## 21     Qn3      Quebec nonchilled 1000   45.5
## 15.1   Qn3      Quebec nonchilled   95   16.2
## 16.1   Qn3      Quebec nonchilled  175   32.4
## 17.1   Qn3      Quebec nonchilled  250   40.3
## 18.1   Qn3      Quebec nonchilled  350   42.1
## 19.1   Qn3      Quebec nonchilled  500   42.9
## 20.1   Qn3      Quebec nonchilled  675   43.9
## 21.1   Qn3      Quebec nonchilled 1000   45.5
## 8.1    Qn2      Quebec nonchilled   95   13.6
## 9.1    Qn2      Quebec nonchilled  175   27.3
## 10.1   Qn2      Quebec nonchilled  250   37.1
## 11.1   Qn2      Quebec nonchilled  350   41.8
## 12.1   Qn2      Quebec nonchilled  500   40.6
## 13.1   Qn2      Quebec nonchilled  675   41.4
## 14.1   Qn2      Quebec nonchilled 1000   44.3
## 1.1    Qn1      Quebec nonchilled   95   16.0
## 2.1    Qn1      Quebec nonchilled  175   30.4
## 3.1    Qn1      Quebec nonchilled  250   34.8
## 4.1    Qn1      Quebec nonchilled  350   37.2
## 5.1    Qn1      Quebec nonchilled  500   35.3
## 6.1    Qn1      Quebec nonchilled  675   39.2
## 7.1    Qn1      Quebec nonchilled 1000   39.7
## 36.1   Qc3      Quebec    chilled   95   15.1
## 37.1   Qc3      Quebec    chilled  175   21.0
## 38.1   Qc3      Quebec    chilled  250   38.1
## 39.1   Qc3      Quebec    chilled  350   34.0
## 40.1   Qc3      Quebec    chilled  500   38.9
## 41.1   Qc3      Quebec    chilled  675   39.6
```

```
## 42.1   Qc3      Quebec    chilled 1000   41.4
## 29.1   Qc2      Quebec    chilled   95    9.3
## 30.1   Qc2      Quebec    chilled  175   27.3
## 31.1   Qc2      Quebec    chilled  250   35.0
## 32.1   Qc2      Quebec    chilled  350   38.8
## 33.1   Qc2      Quebec    chilled  500   38.6
## 34.1   Qc2      Quebec    chilled  675   37.5
## 35.1   Qc2      Quebec    chilled 1000   42.4
## 22.1   Qc1      Quebec    chilled   95   14.2
## 23.1   Qc1      Quebec    chilled  175   24.1
## 24.1   Qc1      Quebec    chilled  250   30.3
## 25.1   Qc1      Quebec    chilled  350   34.6
## 26.1   Qc1      Quebec    chilled  500   32.5
## 27.1   Qc1      Quebec    chilled  675   35.4
## 28.1   Qc1      Quebec    chilled 1000   38.7
## 57.1   Mn3 Mississippi nonchilled   95   11.3
## 58.1   Mn3 Mississippi nonchilled  175   19.4
## 59.1   Mn3 Mississippi nonchilled  250   25.8
## 60.1   Mn3 Mississippi nonchilled  350   27.9
## 61.1   Mn3 Mississippi nonchilled  500   28.5
## 62.1   Mn3 Mississippi nonchilled  675   28.1
## 63.1   Mn3 Mississippi nonchilled 1000   27.8
## 50.1   Mn2 Mississippi nonchilled   95   12.0
## 51.1   Mn2 Mississippi nonchilled  175   22.0
## 52.1   Mn2 Mississippi nonchilled  250   30.6
## 53.1   Mn2 Mississippi nonchilled  350   31.8
## 54.1   Mn2 Mississippi nonchilled  500   32.4
## 55.1   Mn2 Mississippi nonchilled  675   31.1
## 56.1   Mn2 Mississippi nonchilled 1000   31.5
## 43.1   Mn1 Mississippi nonchilled   95   10.6
## 44.1   Mn1 Mississippi nonchilled  175   19.2
## 45.1   Mn1 Mississippi nonchilled  250   26.2
## 46.1   Mn1 Mississippi nonchilled  350   30.0
## 47.1   Mn1 Mississippi nonchilled  500   30.9
## 48.1   Mn1 Mississippi nonchilled  675   32.4
## 49.1   Mn1 Mississippi nonchilled 1000   35.5
## 78.1   Mc3 Mississippi    chilled   95   10.6
## 79.1   Mc3 Mississippi    chilled  175   18.0
## 80.1   Mc3 Mississippi    chilled  250   17.9
## 81.1   Mc3 Mississippi    chilled  350   17.9
## 82.1   Mc3 Mississippi    chilled  500   17.9
## 83.1   Mc3 Mississippi    chilled  675   18.9
## 84.1   Mc3 Mississippi    chilled 1000   19.9
## 71.1   Mc2 Mississippi    chilled   95    7.7
## 72.1   Mc2 Mississippi    chilled  175   11.4
## 73.1   Mc2 Mississippi    chilled  250   12.3
## 74.1   Mc2 Mississippi    chilled  350   13.0
## 75.1   Mc2 Mississippi    chilled  500   12.5
## 76.1   Mc2 Mississippi    chilled  675   13.7
## 77.1   Mc2 Mississippi    chilled 1000   14.4
## 64.1   Mc1 Mississippi    chilled   95   10.5
## 65.1   Mc1 Mississippi    chilled  175   14.9
## 66.1   Mc1 Mississippi    chilled  250   18.1
## 67.1   Mc1 Mississippi    chilled  350   18.9
```

```
## 68.1   Mc1 Mississippi    chilled  500   19.5
## 69.1   Mc1 Mississippi    chilled  675   22.2
## 70.1   Mc1 Mississippi    chilled 1000   21.9
```

Para este ejercicio vamos a usar el dataset state.x77. Asegúrate de que el objeto es un data frame, si no lo es fuerza su conversión.

```
class(state.x77)
```

```
## [1] "matrix"
```

```
df <- as.data.frame(state.x77)
class(df)
```

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

- Averigua cuántos estados tienen ingresos (Income) menores que 4300. Pista investiga la función subset().

```
rownames(subset(df, df$Income < 4300))
```

```
##  [1] "Alabama"        "Arkansas"       "Georgia"        "Idaho"
##  [5] "Kentucky"       "Louisiana"      "Maine"          "Mississippi"
##  [9] "Missouri"       "New Hampshire"  "New Mexico"     "North Carolina"
## [13] "Oklahoma"       "South Carolina" "South Dakota"   "Tennessee"
## [17] "Texas"          "Utah"           "Vermont"        "West Virginia"
```

- Averigua cuál es el estado con los ingresos más altos.

```
rownames(df)[which.max(df$Income)]
```

```
## [1] "Alaska"
```

- Crea un dataframe 2, df2, con los datasets existentes en R: state.abb, state.area, state.division, state.name, state.region. Las filas tienen que ser los nombres de los estados.

```
df2 <- cbind(state.abb, state.area, state.division, state.region)
df2 <- as.data.frame(df2, row.names=state.name)
```

- Elimina de todas las variables la palabra state. Busca alguna función para strings.

```
names(df2) <- gsub("state.", "", names(df2))
names(df2)
```

```
## [1] "abb"      "area"     "division" "region"
```

- Borra la variable div de df2. Estás borrando una única variable del dataframe.

```
df2$division <- NULL
```

- Añade por columnas el nuevo dataframe df2 al dataframe state.x77. Elimina las variables Life Exp, HS Grad, Frost, abb y area.

```
df <- cbind(df, df2)
df$`HS Grad` <- NULL
df$`Life Exp` <- NULL
df$Frost <- NULL
df$abb <- NULL
df$area <- NULL
```

- Añade una variable que categorice el nivel de formación (illiteracy) de manera que [0, 1) is low, [1,2) is some, [2, inf) is high. Hazlo de dos formas utilizando la función cut() y usando ifelse().

```
df$CutVersion <- cut(df$Illiteracy, breaks=c(0,1,2,Inf),
                     labels=c("low", "some", "high"))
df$CutVersion
```

```
##  [1] high some some some some low  some low  some some some low  low  low
## [15] low  low  some high low  low  some low  low  high low  low  low  low
## [29] low  some high some some low  low  some low  low  some high low  some
## [43] high low  low  some low  some low  low
## Levels: low some high
```

```
df$IfElseVersion <- factor(ifelse(df$Illiteracy < 1, "low",
                           ifelse(df$Illiteracy < 2, "some", "high")))
df$IfElseVersion
```

```
##  [1] high some some some some low  some low  some high some low  low  low
## [15] low  low  some high low  low  some low  low  high low  low  low  low
## [29] low  some high some some low  low  some low  some some high low  some
## [43] high low  low  some low  some low  low
## Levels: high low some
```

- Encuentra qué estado del oeste (west) tiene la formación más baja y los mayores ingresos. ¿Qué estado es?

Según la ayuda el factor state.region indica la región y el último valor 4 se corresponde al oeste (west)

```
x <- subset(df, df$region==4)
rownames(x)[which.min(x$Illiteracy)]
```

```
## [1] "Nevada"
```

```
rownames(x)[which.max(x$Income)]
```

```
## [1] "Alaska"
```

El estado del oeste que tiene formación más baja es Nevada y el que tiene más ingresos es Alaska

```
x
```

```
##            Population Income Illiteracy Murder   Area region CutVersion
## Alaska            365   6315        1.5   11.3 566432      4       some
## Arizona          2212   4530        1.8    7.8 113417      4       some
## California      21198   5114        1.1   10.3 156361      4       some
## Colorado         2541   4884        0.7    6.8 103766      4        low
## Hawaii            868   4963        1.9    6.2   6425      4       some
## Idaho             813   4119        0.6    5.3  82677      4        low
## Montana           746   4347        0.6    5.0 145587      4        low
## Nevada            590   5149        0.5   11.5 109889      4        low
## New Mexico       1144   3601        2.2    9.7 121412      4       high
## Oregon           2284   4660        0.6    4.2  96184      4        low
## Utah             1203   4022        0.6    4.5  82096      4        low
## Washington       3559   4864        0.6    4.3  66570      4        low
## Wyoming           376   4566        0.6    6.9  97203      4        low
##            IfElseVersion
## Alaska              some
## Arizona             some
## California          some
## Colorado             low
## Hawaii              some
## Idaho                low
```

```
## Montana              low
## Nevada               low
## New Mexico          high
## Oregon               low
## Utah                 low
## Washington           low
## Wyoming              low
```

Vamos a trabajar con la library(hfligths). Inspecciona el dataframe y familiarízate con las variables.

```
library("hflights")
library("dplyr")
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
df <- hflights
names(df)
```

```
##  [1] "Year"              "Month"             "DayofMonth"
##  [4] "DayOfWeek"         "DepTime"           "ArrTime"
##  [7] "UniqueCarrier"     "FlightNum"         "TailNum"
## [10] "ActualElapsedTime" "AirTime"           "ArrDelay"
## [13] "DepDelay"          "Origin"            "Dest"
## [16] "Distance"          "TaxiIn"            "TaxiOut"
## [19] "Cancelled"         "CancellationCode"  "Diverted"
```

- Busca todos los vuelos del 1 de enero. (Para que quepan en el PDF sólo vamos mostrar unos pocos resultados del dataframe. Quitar head para obtener todos los resultados)

```
df %>% filter(DayofMonth==1 & Month==1) %>% select(Month, DayofMonth, FlightNum) %>%
  head(., 20)
```

```
##    Month DayofMonth FlightNum
## 1      1          1       428
## 2      1          1       460
## 3      1          1      1121
## 4      1          1      1294
## 5      1          1      1700
## 6      1          1      1820
## 7      1          1      1994
## 8      1          1       731
## 9      1          1       620
## 10     1          1       622
## 11     1          1         1
## 12     1          1         5
## 13     1          1         6
## 14     1          1        33
```

```
## 15     1          1        35
## 16     1          1        47
## 17     1          1        52
## 18     1          1        59
## 19     1          1        60
## 20     1          1        62
```

- Busca los vuelos que están sólo operados por American Airlines (AA) o por United Airlines(UA)

```
df %>% filter(UniqueCarrier %in% c("AA", "UA")) %>% select(UniqueCarrier, FlightNum) %>%
  head(., 20)
```

```
##    UniqueCarrier FlightNum
## 1             AA       428
## 2             AA       428
## 3             AA       428
## 4             AA       428
## 5             AA       428
## 6             AA       428
## 7             AA       428
## 8             AA       428
## 9             AA       428
## 10            AA       428
## 11            AA       428
## 12            AA       428
## 13            AA       428
## 14            AA       428
## 15            AA       428
## 16            AA       428
## 17            AA       428
## 18            AA       428
## 19            AA       428
## 20            AA       428
```

- Crea un nuevo dataframe con las variables vuelo, hora de salida, hora de llegada y número de vuelo.

```
df2 <- df %>% select(FlightNum, DepTime, ArrTime)
```

- Selecciona la variable vuelo y aquellas que contengan la palabra “Taxi” o “Delay”

```
df %>% select(grep("*Taxi*|*Delay*", names(df))) %>% head(10)
```

```
##      ArrDelay DepDelay TaxiIn TaxiOut
## 5424      -10        0      7      13
## 5425       -9        1      6       9
## 5426       -8       -8      5      17
## 5427        3        3      9      22
## 5428       -3        5      9       9
## 5429       -7       -1      6      13
## 5430       -1       -1     12      15
## 5431      -16       -5      7      12
## 5432       44       43      8      22
## 5433       43       43      6      19
```

- Crea una tabla que contenga el Unique carrier y el retraso de salida sólo para aquellos vuelos con un retraso superior a una hora (60 minutos) ordenados de forma decreciente

```
df %>% select(UniqueCarrier, DepDelay) %>%
       filter(DepDelay > 60) %>%
       arrange(desc(DepDelay)) %>% head(20)
```

```
##    UniqueCarrier DepDelay
## 1             CO      981
## 2             AA      970
## 3             MQ      931
## 4             UA      869
## 5             MQ      814
## 6             MQ      803
## 7             CO      780
## 8             CO      758
## 9             DL      730
## 10            MQ      691
## 11            AA      677
## 12            AA      653
## 13            XE      628
## 14            UA      588
## 15            CO      576
## 16            UA      563
## 17            WN      548
## 18            UA      535
## 19            AA      525
## 20            MQ      520
```

- Crea una variable que se llame mph y que se calcula como la distancia/tiempo en el aire.

```
df <- df %>% mutate(mph = Distance/ActualElapsedTime)
head(df$mph, 10)
```

```
##  [1] 3.733333 3.733333 3.200000 3.200000 3.612903 3.500000 3.200000
##  [8] 3.796610 3.154930 3.200000
```

- Crea una nueva tabla organizada por destino y que para destino ponga la media de los retrasos en la llegada.

```
df %>% group_by(Dest) %>% summarize(Media = mean(ArrDelay, na.rm=T))
```

```
## # A tibble: 116 x 2
##    Dest   Media
##    <chr>  <dbl>
##  1 ABQ     7.23
##  2 AEX     5.84
##  3 AGS     4
##  4 AMA     6.84
##  5 ANC    26.1
##  6 ASE     6.79
##  7 ATL     8.23
##  8 AUS     7.45
##  9 AVL     9.97
## 10 BFL   -13.2
## # ... with 106 more rows
```

- Calcula para cada compañia el minimo y el máximo de sus retrasos en salidas y llegadas. Ayuda: usa las funciones adicionales como contains_ para cada compañia calcula que dos dias del año fueron los

que tuvieron mas retraso. Ten encuenta que el valor mas pequeño es el primero por defecto, asi que tendras que usar "desc" para poder hacer el ranking.

```
df %>% group_by(UniqueCarrier) %>%
  slice(which.max(DepDelay), which.min(DepDelay),
        which.max(ArrDelay), which.min(ArrDelay))
```

```
## # A tibble: 60 x 22
## # Groups:   UniqueCarrier [15]
##     Year Month DayofMonth DayOfWeek DepTime ArrTime UniqueCarrier FlightNum
##    <int> <int>      <int>     <int>   <int>   <int> <chr>             <int>
##  1  2011    12         12         1     650     808 AA                 1740
##  2  2011     2         13         7    2005    2109 AA                  653
##  3  2011    12         12         1     650     808 AA                 1740
##  4  2011     6         11         6    1753    2106 AA                 1294
##  5  2011     2         28         1    2117      13 AS                  731
##  6  2011     6         18         6    1825    2055 AS                  731
##  7  2011     2         28         1    2117      13 AS                  731
##  8  2011    12         10         6    1826    2039 AS                  731
##  9  2011    10         29         6    2015      17 B6                  622
## 10  2011     8         13         6     616    1103 B6                  620
## # ... with 50 more rows, and 14 more variables: TailNum <chr>,
## #   ActualElapsedTime <int>, AirTime <int>, ArrDelay <int>,
## #   DepDelay <int>, Origin <chr>, Dest <chr>, Distance <int>,
## #   TaxiIn <int>, TaxiOut <int>, Cancelled <int>, CancellationCode <chr>,
## #   Diverted <int>, mph <dbl>
```

Vamos a trabajar con otro dataframe. Descarga el fichero student.txt de la plataforma PRADO, almacena la información en una variable llamada "students". Ten en cuenta que los datos son tab-delimited y tienen un texto para cada columna. Comprueba que R ha leído correctamente el fichero imprimiendo el objeto en la pantalla.

```
students <- read.table("student.txt",header=T)
students
```

```
##    height shoesize gender population
## 1     181       44   male     kuopio
## 2     160       38 female     kuopio
## 3     174       42 female     kuopio
## 4     170       43   male     kuopio
## 5     172       43   male     kuopio
## 6     165       39 female     kuopio
## 7     161       38 female     kuopio
## 8     167       38 female    tampere
## 9     164       39 female    tampere
## 10    166       38 female    tampere
## 11    162       37 female    tampere
## 12    158       36 female    tampere
## 13    175       42   male    tampere
## 14    181       44   male    tampere
## 15    180       43   male    tampere
## 16    177       43   male    tampere
## 17    173       41   male    tampere
```

-Imprime solo los nombres de la columnas.

```
names(students)
```

```
## [1] "height"     "shoesize"   "gender"     "population"
```

-Llama a la columna height solo.

```
students$height
```

```
##  [1] 181 160 174 170 172 165 161 167 164 166 162 158 175 181 180 177 173
```

-¿Cuantas observaciones hay en cada grupo?. Utiliza la función table(). Este commando se puede utilizar para crear tablas cruzadas (cross-tabulations)

```
table(students$gender)
```

```
##
## female   male
##      9      8
```

```
table(students$population)
```

```
##
##  kuopio tampere
##       7      10
```

```
table(students$height, students$shoesize)
```

```
##
##       36 37 38 39 41 42 43 44
##   158  1  0  0  0  0  0  0  0
##   160  0  0  1  0  0  0  0  0
##   161  0  0  1  0  0  0  0  0
##   162  0  1  0  0  0  0  0  0
##   164  0  0  0  1  0  0  0  0
##   165  0  0  0  1  0  0  0  0
##   166  0  0  1  0  0  0  0  0
##   167  0  0  1  0  0  0  0  0
##   170  0  0  0  0  0  0  1  0
##   172  0  0  0  0  0  0  1  0
##   173  0  0  0  0  1  0  0  0
##   174  0  0  0  0  0  1  0  0
##   175  0  0  0  0  0  1  0  0
##   177  0  0  0  0  0  0  1  0
##   180  0  0  0  0  0  0  1  0
##   181  0  0  0  0  0  0  0  2
```

-Crea nuevas variables a partir de los datos que tenemos. Vamos a crear una variable nueva “sym” que contenga M si el genero es masculino y F si el genero es femenino. Busca en la ayuda información sobre la función ifelse(). Crea una segunda variable “colours” cuyo valor será “Blue” si el estudiante es de kuopio y “Red” si es de otro sitio.Con los datos anteriores de height y shoesize y las nuevas variables crea un nuevo data.frame que se llame students.new

```
students.new <- as.data.frame(students) %>%
    mutate(sym = ifelse(gender == "male", "M", "F")) %>%
    mutate(colours = ifelse(population == "kuopio", "Blue", "Red")) %>%
    select(height, shoesize, sym, colours)
head(students.new)
```

```
##   height shoesize sym colours
```

```
## 1    181       44   M    Blue
## 2    160       38   F    Blue
## 3    174       42   F    Blue
## 4    170       43   M    Blue
## 5    172       43   M    Blue
## 6    165       39   F    Blue
```

- Comprueba que la clase de students.new es un dataframe

```
class(students.new)
```

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

- Crea dos subsets a partir del dataset student. Dividelo dependiendo del sexo. Para ello primero comprueba que estudiantes son hombres (male). Pista: busca información sobre la función which.

```
student <- as.data.frame(students)
subset1 <- which(student$gender=="male")
```

-Basándote en esa selección dada por which() toma solo esas filas del dataset student para generar el subset student.male

```
student.male <- student[subset1,]
student.male
```

```
##    height shoesize gender population
## 1     181       44   male     kuopio
## 4     170       43   male     kuopio
## 5     172       43   male     kuopio
## 13    175       42   male    tampere
## 14    181       44   male    tampere
## 15    180       43   male    tampere
## 16    177       43   male    tampere
## 17    173       41   male    tampere
```

- Repite el procedimiento para seleccionar las estudiantes mujeres (females)

```
student.female = student[-subset1,]
student.female
```

```
##    height shoesize gender population
## 2     160       38 female     kuopio
## 3     174       42 female     kuopio
## 6     165       39 female     kuopio
## 7     161       38 female     kuopio
## 8     167       38 female    tampere
## 9     164       39 female    tampere
## 10    166       38 female    tampere
## 11    162       37 female    tampere
## 12    158       36 female    tampere
```

- Utiliza la function write.table() para guarder el contenido de student.new en un archivo.

```
write.table(students.new, "student_new.txt")
```