Ejemplos dplyr

Salvador Carrillo Fuentes

Abril de 2019

Load data

```
library(readr)

brc <- read.csv("breast-cancer.data", header=FALSE)
names <- read.csv("breast-cancer.names1.csv", header=FALSE)

names(brc) <- as.character(names$V1)
names(brc)
```

```
##  [1] "Class"       "age"         "menopause"   "tumor_size"  "inv_nodes"
##  [6] "node_caps"   "deg_malig"   "breast"      "breast_quad" "irradiat"
```

Funciones principales de `dpylr`:

```
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
```

5 funciones b?sicas:

- filtrar filas del *dataset* `filter`
- ordenar *dataset* seg?n ciertas variables : `arrange` - `top_n`
- seleccionar columnas: `select`
- crear nuevas columnas: `mutate`
- sumarizaci?n del *dataset*: `summarise`

Funcionamiento:

- 1 arg: dataset
- resto args: dice qu? hacer
- nombre columnas nunca con comillas
- resultado es un `data.frame`

Filter

```r
# knitr::kable(filter(brc, deg_malig == 3),  longtable = TRUE)
library(knitr)
kable(head(filter(brc, deg_malig == 3)),  longtable = TRUE)
```

Class	age	menopause	tumor_size	inv_nodes	node_caps	deg_malig	breast	breast_quad	irradiat
no-recurrence-events	30-39	premeno	30-34	0-2	no	3	left	left_low	no
no-recurrence-events	40-49	premeno	0-4	0-2	no	3	left	central	no
no-recurrence-events	50-59	ge40	25-29	0-2	no	3	left	right_up	no
no-recurrence-events	40-49	premeno	30-34	0-2	no	3	left	left_up	no
no-recurrence-events	50-59	premeno	30-34	0-2	no	3	left	left_low	no
no-recurrence-events	60-69	ge40	30-34	0-2	no	3	left	left_low	no

```r
kable(head(filter(brc, deg_malig == 3, irradiat == "no")),  longtable = TRUE)
```

Class	age	menopause	tumor_size	inv_nodes	node_caps	deg_malig	breast	breast_quad	irradiat
no-recurrence-events	30-39	premeno	30-34	0-2	no	3	left	left_low	no
no-recurrence-events	40-49	premeno	0-4	0-2	no	3	left	central	no
no-recurrence-events	50-59	ge40	25-29	0-2	no	3	left	right_up	no
no-recurrence-events	40-49	premeno	30-34	0-2	no	3	left	left_up	no
no-recurrence-events	50-59	premeno	30-34	0-2	no	3	left	left_low	no
no-recurrence-events	60-69	ge40	30-34	0-2	no	3	left	left_low	no

```r
kable(head(filter(brc, deg_malig == 3 & irradiat == "no" | breast == "left")),  longtable = TRUE)
```

Class	age	menopause	tumor_size	inv_nodes	node_caps	deg_malig	breast	breast_quad	irradiat
no-recurrence-events	30-39	premeno	30-34	0-2	no	3	left	left_low	no
no-recurrence-events	40-49	premeno	20-24	0-2	no	2	left	left_low	no
no-recurrence-events	60-69	ge40	15-19	0-2	no	2	left	left_low	no
no-recurrence-events	50-59	premeno	25-29	0-2	no	2	left	left_low	no
no-recurrence-events	60-69	ge40	20-24	0-2	no	1	left	left_low	no
no-recurrence-events	40-49	premeno	50-54	0-2	no	2	left	left_low	no

Ordenar *dataset*

```
brc1 <- filter(brc, deg_malig == 3 & irradiat == "no" | breast == "left")

kable(head(arrange(brc1, age, tumor_size)),  longtable = TRUE)
```

Class	age	menopause	tumor_size	inv_nodes	node_caps	deg_malig	breast	breast_quad	irradiat
no-recurrence-events	30-39	premeno	10-14	0-2	no	2	left	right_low	no
no-recurrence-events	30-39	premeno	15-19	0-2	no	1	left	left_low	no
no-recurrence-events	30-39	lt40	15-19	0-2	no	3	right	left_up	no
no-recurrence-events	30-39	premeno	15-19	0-2	no	1	left	left_low	no
recurrence-events	30-39	premeno	15-19	6-8	yes	3	left	left_low	yes
no-recurrence-events	30-39	premeno	20-24	0-2	no	2	left	right_low	no

```
kable(head(arrange(brc1, desc(age), tumor_size)),  longtable = TRUE)
```

Class	age	menopause	tumor_size	inv_nodes	node_caps	deg_malig	breast	breast_quad	irradiat
no-recurrence-events	70-79	ge40	0-4	0-2	no	1	left	right_low	no
no-recurrence-events	70-79	ge40	10-14	0-2	no	2	left	central	no
recurrence-events	70-79	ge40	15-19	9-11	?	1	left	left_low	yes
no-recurrence-events	70-79	ge40	20-24	0-2	no	3	left	left_up	no
no-recurrence-events	60-69	lt40	10-14	0-2	no	1	left	right_up	no

Class	age	menopause	tumor_size	inv_nodes	node_caps	deg_malig	breast	breast_quad	irradiat
no-recurrence-events	60-69	ge40	10-14	0-2	no	1	left	left_low	no

Ejercicio:

Extraer los 10 pacientes con menopausia con mayor tama?o del tumor y menor n?mero de nodos invasores ordenados por edad:

```r
brc2 <- filter(brc, menopause == "premeno")

# head(arrange(brc2, age, desc(tumor_size), inv_nodes), n=10)
kable((arrange(brc2, age, desc(tumor_size), inv_nodes))[1:10, ], longtable = TRUE)
```

Class	age	menopause	tumor_size	inv_nodes	node_caps	deg_malig	breast	breast_quad	irradiat
no-recurrence-events	20-29	premeno	35-39	0-2	no	2	right	right_up	no
no-recurrence-events	30-39	premeno	5-9	0-2	no	2	left	right_low	no
no-recurrence-events	30-39	premeno	40-44	0-2	no	2	right	right_up	no
no-recurrence-events	30-39	premeno	40-44	0-2	no	2	left	left_low	yes
recurrence-events	30-39	premeno	40-44	0-2	no	1	left	left_up	no
no-recurrence-events	30-39	premeno	40-44	3-5	no	3	right	right_up	yes
recurrence-events	30-39	premeno	35-39	0-2	no	3	left	left_low	no
recurrence-events	30-39	premeno	35-39	0-2	no	3	left	left_low	no
recurrence-events	30-39	premeno	35-39	9-11	yes	3	left	left_low	no
no-recurrence-events	30-39	premeno	30-34	0-2	no	3	left	left_low	no

top_n()

```r
brc1 <- filter(brc, age =="40-49")

top_n(brc1, 1)
```

```
## Selecting by irradiat
```

```
##                    Class   age menopause tumor_size inv_nodes node_caps
## 1  no-recurrence-events 40-49   premeno      35-39      9-11       yes
## 2  no-recurrence-events 40-49   premeno      35-39      9-11       yes
## 3  no-recurrence-events 40-49   premeno      40-44       3-5       yes
## 4  no-recurrence-events 40-49   premeno        5-9       0-2         no
## 5  no-recurrence-events 40-49   premeno      45-49       0-2         no
## 6  no-recurrence-events 40-49   premeno      25-29       0-2          ?
## 7  no-recurrence-events 40-49   premeno      25-29       0-2         no
## 8  no-recurrence-events 40-49      ge40      40-44     15-17        yes
## 9  no-recurrence-events 40-49      ge40      30-34       0-2         no
## 10 no-recurrence-events 40-49   premeno      30-34       0-2         no
## 11 no-recurrence-events 40-49   premeno      20-24       0-2         no
## 12 no-recurrence-events 40-49   premeno      35-39       0-2        yes
## 13 no-recurrence-events 40-49   premeno      35-39       0-2        yes
## 14 no-recurrence-events 40-49   premeno      25-29       0-2         no
## 15 no-recurrence-events 40-49   premeno      20-24       6-8         no
## 16 no-recurrence-events 40-49   premeno      15-19     12-14         no
## 17 no-recurrence-events 40-49   premeno      25-29       0-2         no
## 18 no-recurrence-events 40-49   premeno      10-14       0-2         no
## 19    recurrence-events 40-49      ge40      20-24       3-5         no
## 20    recurrence-events 40-49   premeno      20-24       3-5        yes
## 21    recurrence-events 40-49   premeno      30-34     12-14        yes
## 22    recurrence-events 40-49   premeno      50-54       0-2         no
## 23    recurrence-events 40-49   premeno      30-34       0-2         no
## 24    recurrence-events 40-49   premeno      20-24       3-5        yes
## 25    recurrence-events 40-49      ge40      25-29     12-14        yes
## 26    recurrence-events 40-49   premeno      25-29       0-2         no
##    deg_malig breast breast_quad irradiat
## 1          2  right     left_up      yes
## 2          2  right    right_up      yes
## 3          3  right     left_up      yes
## 4          1   left    left_low      yes
## 5          2   left    left_low      yes
## 6          2   left   right_low      yes
## 7          3  right     left_up      yes
## 8          2  right     left_up      yes
## 9          2   left     left_up      yes
## 10         2  right    right_up      yes
## 11         3  right    left_low      yes
## 12         3  right     left_up      yes
## 13         3  right    left_low      yes
## 14         1  right    left_low      yes
## 15         2  right    left_low      yes
## 16         3  right   right_low      yes
## 17         2   left     left_up      yes
## 18         2   left    left_low      yes
## 19         3  right    left_low      yes
## 20         2  right    right_up      yes
## 21         3   left     left_up      yes
## 22         2  right    left_low      yes
## 23         1   left    left_low      yes
## 24         2   left    left_low      yes
## 25         3   left   right_low      yes
## 26         2   left    left_low      yes
```

```r
head(top_n(brc1, -1))
```

```
## Selecting by irradiat
##                  Class   age menopause tumor_size inv_nodes node_caps deg_malig
## 1 no-recurrence-events 40-49   premeno      20-24       0-2        no         2
## 2 no-recurrence-events 40-49   premeno      20-24       0-2        no         2
## 3 no-recurrence-events 40-49   premeno        0-4       0-2        no         2
## 4 no-recurrence-events 40-49   premeno      50-54       0-2        no         2
## 5 no-recurrence-events 40-49   premeno      20-24       0-2        no         2
## 6 no-recurrence-events 40-49   premeno        0-4       0-2        no         3
##   breast breast_quad irradiat
## 1  right    right_up       no
## 2   left    left_low       no
## 3  right   right_low       no
## 4   left    left_low       no
## 5  right     left_up       no
## 6   left     central       no
```

Seleccionar columnas

- `starts_with()`
- `ends_with()`

```r
head(select(brc, age, tumor_size))
```

```
##     age tumor_size
## 1 30-39      30-34
## 2 40-49      20-24
## 3 40-49      20-24
## 4 60-69      15-19
## 5 40-49        0-4
## 6 60-69      15-19
```

```r
kable(head(select(brc, breast, breast_quad, everything())))
```

breast	breast_quad	Class	age	menopause	tumor_size	inv_nodes	node_caps	deg_malig	irradiat
left	left_low	no-recurrence-events	30-39	premeno	30-34	0-2	no	3	no
right	right_up	no-recurrence-events	40-49	premeno	20-24	0-2	no	2	no
left	left_low	no-recurrence-events	40-49	premeno	20-24	0-2	no	2	no
right	left_up	no-recurrence-events	60-69	ge40	15-19	0-2	no	2	no
right	right_low	no-recurrence-events	40-49	premeno	0-4	0-2	no	2	no
left	left_low	no-recurrence-events	60-69	ge40	15-19	0-2	no	2	no

```r
head(select(brc, Class:inv_nodes)) # desde:hasta
```

```
##                  Class   age menopause tumor_size inv_nodes
## 1 no-recurrence-events 30-39   premeno      30-34       0-2
## 2 no-recurrence-events 40-49   premeno      20-24       0-2
## 3 no-recurrence-events 40-49   premeno      20-24       0-2
## 4 no-recurrence-events 60-69      ge40      15-19       0-2
## 5 no-recurrence-events 40-49   premeno        0-4       0-2
## 6 no-recurrence-events 60-69      ge40      15-19       0-2
```

```r
head(select(brc, contains("_")))
```

```
##   tumor_size inv_nodes node_caps deg_malig breast_quad
## 1      30-34       0-2        no         3    left_low
## 2      20-24       0-2        no         2    right_up
## 3      20-24       0-2        no         2    left_low
## 4      15-19       0-2        no         2     left_up
## 5        0-4       0-2        no         2   right_low
## 6      15-19       0-2        no         2    left_low
```

```r
head(select(brc, starts_with("a")))
```

```
##     age
## 1 30-39
## 2 40-49
## 3 40-49
## 4 60-69
## 5 40-49
## 6 60-69
```

```r
br1 <- select(brc, breast, breast_quad, everything())
br2 <- select(br1, -c(Class, irradiat));
head(br1)
```

```
##   breast breast_quad                Class   age menopause tumor_size inv_nodes
## 1   left    left_low no-recurrence-events 30-39   premeno      30-34       0-2
## 2  right    right_up no-recurrence-events 40-49   premeno      20-24       0-2
## 3   left    left_low no-recurrence-events 40-49   premeno      20-24       0-2
## 4  right     left_up no-recurrence-events 60-69      ge40      15-19       0-2
## 5  right   right_low no-recurrence-events 40-49   premeno        0-4       0-2
## 6   left    left_low no-recurrence-events 60-69      ge40      15-19       0-2
##   node_caps deg_malig irradiat
## 1        no         3       no
## 2        no         2       no
## 3        no         2       no
## 4        no         2       no
## 5        no         2       no
## 6        no         2       no
```

```r
head(br2)
```

```
##   breast breast_quad   age menopause tumor_size inv_nodes node_caps deg_malig
## 1   left    left_low 30-39   premeno      30-34       0-2        no         3
## 2  right    right_up 40-49   premeno      20-24       0-2        no         2
## 3   left    left_low 40-49   premeno      20-24       0-2        no         2
## 4  right     left_up 60-69      ge40      15-19       0-2        no         2
## 5  right   right_low 40-49   premeno        0-4       0-2        no         2
## 6   left    left_low 60-69      ge40      15-19       0-2        no         2
```

mutate():

```r
br1 <- mutate(brc, distancia = 4 - deg_malig);
head(br1)
```

```
##                  Class   age menopause tumor_size inv_nodes node_caps deg_malig
## 1 no-recurrence-events 30-39   premeno      30-34       0-2        no         3
## 2 no-recurrence-events 40-49   premeno      20-24       0-2        no         2
## 3 no-recurrence-events 40-49   premeno      20-24       0-2        no         2
## 4 no-recurrence-events 60-69      ge40      15-19       0-2        no         2
## 5 no-recurrence-events 40-49   premeno        0-4       0-2        no         2
## 6 no-recurrence-events 60-69      ge40      15-19       0-2        no         2
##   breast breast_quad irradiat distancia
## 1   left    left_low       no         1
## 2  right    right_up       no         2
## 3   left    left_low       no         2
## 4  right     left_up       no         2
## 5  right   right_low       no         2
## 6   left    left_low       no         2
```

rename():

```r
br2 <- rename(br1, dist = distancia);
head(br2)
```

```
##                  Class   age menopause tumor_size inv_nodes node_caps deg_malig
## 1 no-recurrence-events 30-39   premeno      30-34       0-2        no         3
## 2 no-recurrence-events 40-49   premeno      20-24       0-2        no         2
## 3 no-recurrence-events 40-49   premeno      20-24       0-2        no         2
## 4 no-recurrence-events 60-69      ge40      15-19       0-2        no         2
## 5 no-recurrence-events 40-49   premeno        0-4       0-2        no         2
## 6 no-recurrence-events 60-69      ge40      15-19       0-2        no         2
##   breast breast_quad irradiat dist
## 1   left    left_low       no    1
## 2  right    right_up       no    2
## 3   left    left_low       no    2
## 4  right     left_up       no    2
## 5  right   right_low       no    2
## 6   left    left_low       no    2
```

transmute():

```
head(transmute(brc2, test = deg_malig / 61.0237))
```

```
##         test
## 1 0.04916123
## 2 0.03277415
## 3 0.03277415
## 4 0.03277415
## 5 0.03277415
## 6 0.03277415
```

summarize():

Realiza un resumen de una estad?stica y guardarlos en un *dataframe*:

```
summarise(brc, a = mean(brc$deg_malig))
```

```
##          a
## 1 2.048951
```

```
summarise(brc, a = median(brc$deg_malig))
```

```
##   a
## 1 2
```

Agregaci?n de datos

Agrega datos de varias columnas:

```
data("mtcars")
aggdata <-aggregate(mtcars, by=list(mtcars$cyl, mtcars$vs), FUN=mean, na.rm=TRUE)
print(aggdata)
```

```
##   Group.1 Group.2      mpg cyl   disp       hp     drat       wt     qsec vs
## 1       4       0 26.00000   4 120.30  91.0000 4.430000 2.140000 16.70000  0
## 2       6       0 20.56667   6 155.00 131.6667 3.806667 2.755000 16.32667  0
## 3       8       0 15.10000   8 353.10 209.2143 3.229286 3.999214 16.77214  0
## 4       4       1 26.73000   4 103.62  81.8000 4.035000 2.300300 19.38100  1
## 5       6       1 19.12500   6 204.55 115.2500 3.420000 3.388750 19.21500  1
##          am     gear     carb
## 1 1.0000000 5.000000 2.000000
## 2 1.0000000 4.333333 4.666667
## 3 0.1428571 3.285714 3.500000
## 4 0.7000000 4.000000 1.500000
## 5 0.0000000 3.500000 2.500000
```

pipes

`x %>% f(y)` ≡ `f(x, y)`

Por ejemplo, `brc %>% filter(age=="40-49")` ≡ `filter(brc, age=="40-49")`

```
head(filter(brc, age=="40-49"))
```

```
##                  Class   age menopause tumor_size inv_nodes node_caps deg_malig
## 1 no-recurrence-events 40-49   premeno      20-24       0-2        no         2
## 2 no-recurrence-events 40-49   premeno      20-24       0-2        no         2
## 3 no-recurrence-events 40-49   premeno        0-4       0-2        no         2
## 4 no-recurrence-events 40-49   premeno      50-54       0-2        no         2
## 5 no-recurrence-events 40-49   premeno      20-24       0-2        no         2
## 6 no-recurrence-events 40-49   premeno        0-4       0-2        no         3
##   breast breast_quad irradiat
## 1  right    right_up       no
## 2   left    left_low       no
## 3  right   right_low       no
## 4   left    left_low       no
## 5  right     left_up       no
## 6   left     central       no
```

```
brc %>%
  filter(age=="40-49") %>%
  select(tumor_size) %>%
  top_n(5)
```

```
## Selecting by tumor_size
```

```
##   tumor_size
## 1      50-54
## 2      40-44
## 3      40-44
## 4      40-44
## 5        5-9
## 6      45-49
## 7      40-44
## 8      40-44
## 9      50-54
```

```
top_n(select(filter(brc, age=="40-49"), tumor_size), 5)
```

```
## Selecting by tumor_size
```

```
##   tumor_size
## 1      50-54
## 2      40-44
## 3      40-44
## 4      40-44
## 5        5-9
## 6      45-49
## 7      40-44
## 8      40-44
## 9      50-54
```

group_by():

```
brc %>% group_by(age)
```

```
## # A tibble: 286 x 10
## # Groups:   age [6]
##    Class age   menopause tumor_size inv_nodes node_caps deg_malig breast
##    <chr> <chr> <chr>     <chr>      <chr>     <chr>         <int> <chr>
##  1 no-r~ 30-39 premeno   30-34      0-2       no                3 left
##  2 no-r~ 40-49 premeno   20-24      0-2       no                2 right
##  3 no-r~ 40-49 premeno   20-24      0-2       no                2 left
##  4 no-r~ 60-69 ge40      15-19      0-2       no                2 right
##  5 no-r~ 40-49 premeno   0-4        0-2       no                2 right
##  6 no-r~ 60-69 ge40      15-19      0-2       no                2 left
##  7 no-r~ 50-59 premeno   25-29      0-2       no                2 left
##  8 no-r~ 60-69 ge40      20-24      0-2       no                1 left
##  9 no-r~ 40-49 premeno   50-54      0-2       no                2 left
## 10 no-r~ 40-49 premeno   20-24      0-2       no                2 right
## # ... with 276 more rows, and 2 more variables: breast_quad <chr>,
## #   irradiat <chr>
```

Podemos usar group_by() junto con summarise():

```
brc %>%
    group_by(age) %>%
    summarise(mean_deg_malig=mean(deg_malig))
```

```
## ‘summarise()‘ ungrouping output (override with ‘.groups‘ argument)
```

```
## # A tibble: 6 x 2
##   age   mean_deg_malig
##   <chr>          <dbl>
## 1 20-29           2   
## 2 30-39           2.14
## 3 40-49           2.07
## 4 50-59           2.07
## 5 60-69           1.98
## 6 70-79           1.5 
```

```
brc %>%
    group_by(age,tumor_size) %>%
    summarise(mean_deg_malig=mean(deg_malig),
              sd_deg_malig=sd(deg_malig))
```

```
## ‘summarise()‘ regrouping output by ’age’ (override with ‘.groups‘ argument)
```

```
## # A tibble: 46 x 4
## # Groups:   age [6]
##    age   tumor_size mean_deg_malig sd_deg_malig
##    <chr> <chr>               <dbl>        <dbl>
##  1 20-29 35-39                2          NA
##  2 30-39 0-4                  2           0
##  3 30-39 10-14                1.5         0.707
##  4 30-39 15-19                1.8         1.10
##  5 30-39 20-24                2.33        0.516
##  6 30-39 25-29                2.17        0.753
##  7 30-39 30-34                2.14        0.690
##  8 30-39 35-39                3           0
##  9 30-39 40-44                2           0.816
## 10 30-39 5-9                  2          NA
## # ... with 36 more rows
```