# Filter y subset de data frames

Téllez Gerardo Rubén

19/6/2021

## Castring

- **as.character(columna o vector)**: Convierte los datos a palabras.
- **as.integer(columna o vector)**: Convierte los datos a números enteros.
- **as.numeric(columna o vector)**: Transformar todos los datos de un objeto a números reales.
- **as.factor(columna o vector)**; Convierte los datos a factores

```
vino = read.table("wine.data", sep = ",", header = FALSE, col.names = c("Clase de distribución","Alcohol

head(vino[vino$Ácido.malico > 4, ])
```

```
##     Clase.de.distribución Alcohol Ácido.malico Ceniza Alcalinidad.de.la.ceniza
## 46                      1   14.21         4.04   2.44                    18.9
## 123                     2   12.42         4.43   2.73                    26.5
## 124                     2   13.05         5.80   2.13                    21.5
## 125                     2   11.87         4.31   2.39                    21.0
## 130                     2   12.04         4.30   2.38                    22.0
## 137                     3   12.25         4.72   2.54                    21.0
##     Magnesio Fenoles.totales Flavonoides Fenoles.no.flavonoides
## 46       111            2.85        2.65                  0.30
## 123      102            2.20        2.13                  0.43
## 124       86            2.62        2.65                  0.30
## 125       82            2.86        3.03                  0.21
## 130       80            2.10        1.75                  0.42
## 137       89            1.38        0.47                  0.53
##     Proantocianidinas Intensidad.del.color Matiz OD280.OD315.de.vinos.diluídos
## 46               1.25                 5.24  0.87                         3.33
## 123              1.71                 2.08  0.92                         3.12
## 124              2.01                 2.60  0.73                         3.10
## 125              2.91                 2.80  0.75                         3.64
## 130              1.35                 2.60  0.79                         2.57
## 137              0.80                 3.85  0.75                         1.27
##     Prolina
## 46     1080
## 123     365
## 124     380
## 125     380
## 130     580
## 137     720
```

- **droplevels(DF)**: para borrar los niveles sobrantes de todos los factores, ya que las columnas que son factores heredan en los sub-dataframes todos los niveles del factor original, aunque no estén en el trozo seleccionado.

## Formas de filtrado de tydiverse o dplyr

- **select(DF, parámetros)**: para especificar qué se quiere extraer de un dataframe
  - **strarts_with("x")**: extrae las columnas cuyo nombre empiece con la palabra "x"
  - **ends_with("x")**: extrae las variables cuyo nombre termine con la palabra "x"
  - **contains("x")**: extrae las vaiables cuyo nombre contenga la palabra "x"

```
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
```

```
head(select(vino, starts_with("Ácido")))
```

```
##   Ácido.malico
## 1         1.71
## 2         1.78
## 3         2.36
## 4         1.95
## 5         2.59
## 6         1.76
```

```
head(select(vino, ends_with("ceniza")))
```

```
##   Ceniza Alcalinidad.de.la.ceniza
## 1   2.43                    15.6
## 2   2.14                    11.2
## 3   2.67                    18.6
## 4   2.50                    16.8
## 5   2.87                    21.0
## 6   2.45                    15.2
```

```
head(select(vino, contains("flavonoides")))
```

```
##   Flavonoides Fenoles.no.flavonoides
## 1        3.06                   0.28
## 2        2.76                   0.26
## 3        3.24                   0.30
## 4        3.49                   0.24
## 5        2.69                   0.39
## 6        3.39                   0.34
```

## Subset natural

- **subset(DF, condición, select = columnas)**: para extraer del data frame las filas que cumplen con la condición y las columnas especificadas
  - Si se quiere todas las filas, no se especifica condición
  - Si se quiere todas las columnas no se especifica select
  - Las variables en la condición se especifican con su nombre, sin añadir antes el nombre del data frame

```
# Selecciona los vinos con GL < 13 y las columnas 1 a 3
sta = subset(vino, Alcohol > 13, select = 1:3)
head(sta)
```

```
##   Clase.de.distribución Alcohol Ácido.malico
## 1                     1   14.23         1.71
## 2                     1   13.20         1.78
## 3                     1   13.16         2.36
## 4                     1   14.37         1.95
## 5                     1   13.24         2.59
## 6                     1   14.20         1.76
```

```
# Selecciona vinos con GL > 13, y las columnas de select
st = subset(vino, Alcohol > 13, select = c(Alcohol, Ácido.malico))
head(st)
```

```
##   Alcohol Ácido.malico
## 1   14.23         1.71
## 2   13.20         1.78
## 3   13.16         2.36
## 4   14.37         1.95
## 5   13.24         2.59
## 6   14.20         1.76
```