# Descarga de datos de EDUCAbase

Ilustración: Alumnado matriculado por enseñanza

Marzo 2024

Método: Basado en la información disponible en la web

1. Extrae el nombre de los archivos del código fuente html
2. Genera url para cada una de las bases de datos en formato .csv
3. Lee y guarda los datos directamente de la web

```
library(xml2)
library(rvest)
library(stringr)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v purrr     1.0.1
## v forcats   1.0.0     v readr     2.1.4
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter()         masks stats::filter()
## x readr::guess_encoding() masks rvest::guess_encoding()
## x dplyr::lag()            masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

## Extrae nombre de bases de datos de la fuente html

```
html_source <- "https://estadisticas.educacion.gob.es/EducaDynPx/educabase/index.htm?type=pcaxis&path=/
pg <- read_html(html_source)
href <- html_attr(html_nodes(pg, "a"), "href")
href <- href[!is.na(href)]
href <- href[nchar(href)>1]

pattern <- ".*&file= *(.*?) *.px&"
file <- unique(str_match(href, pattern)[,2])
file <- file[!is.na(file)]
file <- paste0(file, ".csv")
```

## Genera url de bases de datos en .csv

```
base_url <- "https://estadisticas.educacion.gob.es/EducaJaxiPx/files/_px/es/csv_bdsc/no-universitaria/a
suf_url <- "_bdsc?nocab=1"
myurl <- paste0(base_url, file, suf_url)
```

## Guarda en directorio y lee los archivos .csv (opcional)

```
#dir <- "/home/eldani/MEGA/Work/Projects/Ongoing/mefd_stat/datos"
#dir_full <- file.path(dir, file)
#lapply(seq_along(file), function(x) download.file(myurl = myurl[[x]], destfile = dir_full[[x]]))
#df <- lapply(dir_full, function(x) read.csv2(myurl(x)))
```

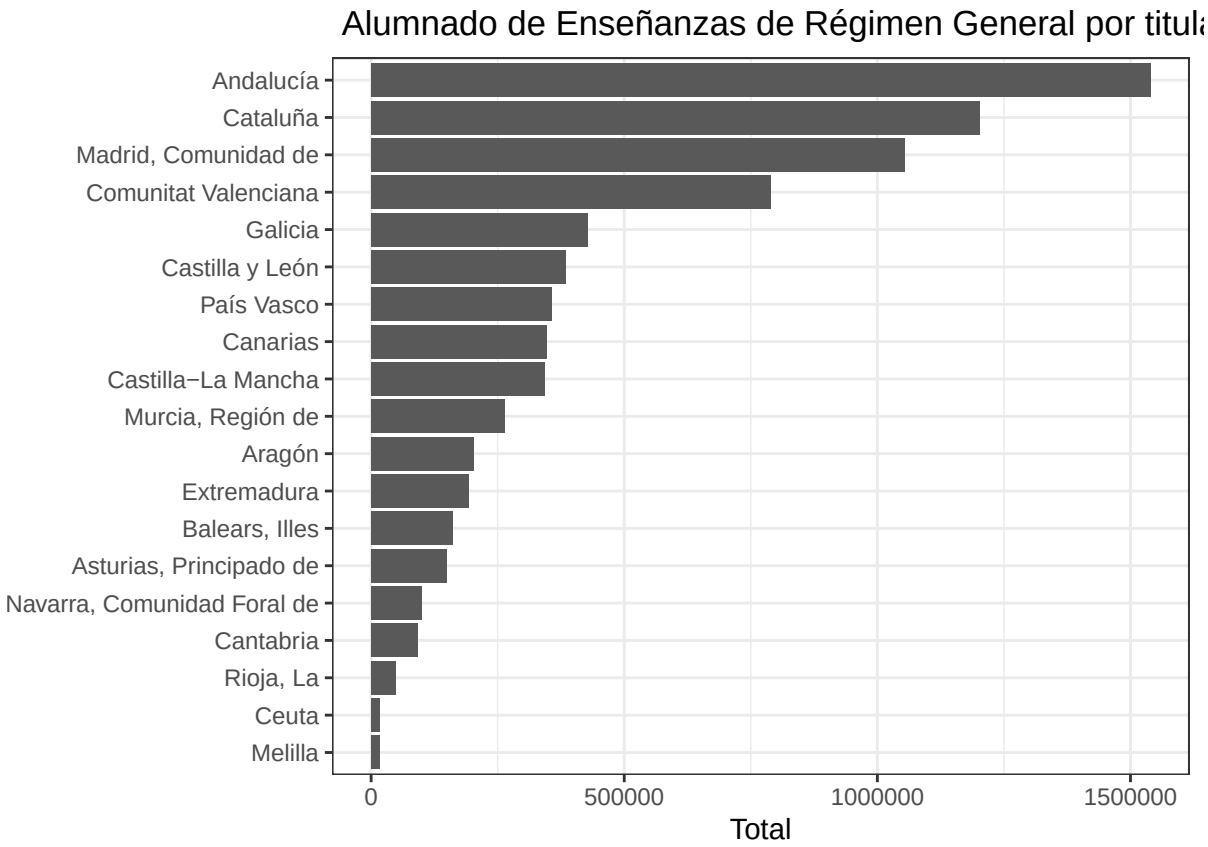
## Lee los archivos .csv directo de la web

```
alumnado <- lapply(myurl, function(x) read.csv2(url(x)))
names(alumnado) <- file
```

## Visualización de datos seleccionados

```
df <- alumnado[["alumnado_1_01.csv"]] %>%
      filter(Titularidad.del.centro == "TODOS LOS CENTROS") %>%
      filter(Comunidad.autónoma != "TOTAL") %>%
      mutate(Total = as.numeric(gsub('\\.', '', Total))) %>% # remover puntos
      group_by(Comunidad.autónoma) %>%
      summarise(Total = mean(Total, na.rm = TRUE))

ggplot(df, aes(x = Total, y = reorder(Comunidad.autónoma, Total))) +
  geom_bar(stat = "identity") +
  ylab("") +
  theme_bw() +
  ggtitle(" Alumnado de Enseñanzas de Régimen General por titularidad del centro, comunidad autónoma")
```


Alumnado de Enseñanzas de Régimen General por titula
Andalucía
Cataluña
Madrid, Comunidad de
Comunitat Valenciana
Galicia
Castilla y León
País Vasco
Canarias
Castilla−La Mancha
Murcia, Región de
Aragón
Extremadura
Balears, Illes
Asturias, Principado de
Navarra, Comunidad Foral de
Cantabria
Rioja, La
Ceuta
Melilla
0
500000
1000000
1500000
Total