# Dataset Gapminder

## Importa los datos con readr

Recuerda guardar los archivos csv en tu directorio de trabajo para poder cargarlos a R sin tener que especificar su ruta, con las siguientes órdenes:

```
library(readr)

gapminder <- read_delim("Data/gapminder_1800.csv",  ";", escape_double = FALSE, trim_ws = TRUE)
```

```
## Rows: 46995 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ";"
## chr (2): geo, country
## dbl (3): year, income, gdp
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
hiv <- read_delim("Data/hiv.csv",  ";", escape_double = FALSE, trim_ws = TRUE)
```

```
## Rows: 275 Columns: 34
## -- Column specification --------------------------------------------------------
## Delimiter: ";"
## chr  (1): country
## dbl (31): 1979, 1980, 1981, 1982, 1983, 1984, 1985, 1986, 1987, 1990, 1991, ...
## lgl  (2): 1988, 1989
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

## Produce un conjunto de datos ordenado con las herramientas tidyverse

El conjunto de datos debe contener información desde el año 1991 en adelante, y queremos terminar con columnas country (factor), year (numérica) y prevalence (numérica). Es importante especificar el tipo correcto de cada variable porque nos permitirá operar con ellas sin problemas más adelante.

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v purrr     1.0.1
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     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 dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
prev <- hiv %>%
  gather(key="year", value="prevalence", -country) %>% #gater() transforma los datos de formato ancho a
  drop_na() %>% #eliminar los valores nulos
  mutate(year = as.numeric(year),
           prevalence = as.numeric(prevalence),
           country = as.character(country)) %>%  # funcion mutate()para crear o modificar el tipo de v
  filter(year > 1990) #funcion filter() que permite filtrar los casos
```

## Une ambas bases de datos

```
gap_hiv <- gapminder %>%
  select(country, year, income) %>%  # la funcion select() permite seleccionar las columnas de interés
  inner_join(prev, by=join_by(country, year)) # inner_join() permite unir ambas bases de datos
```

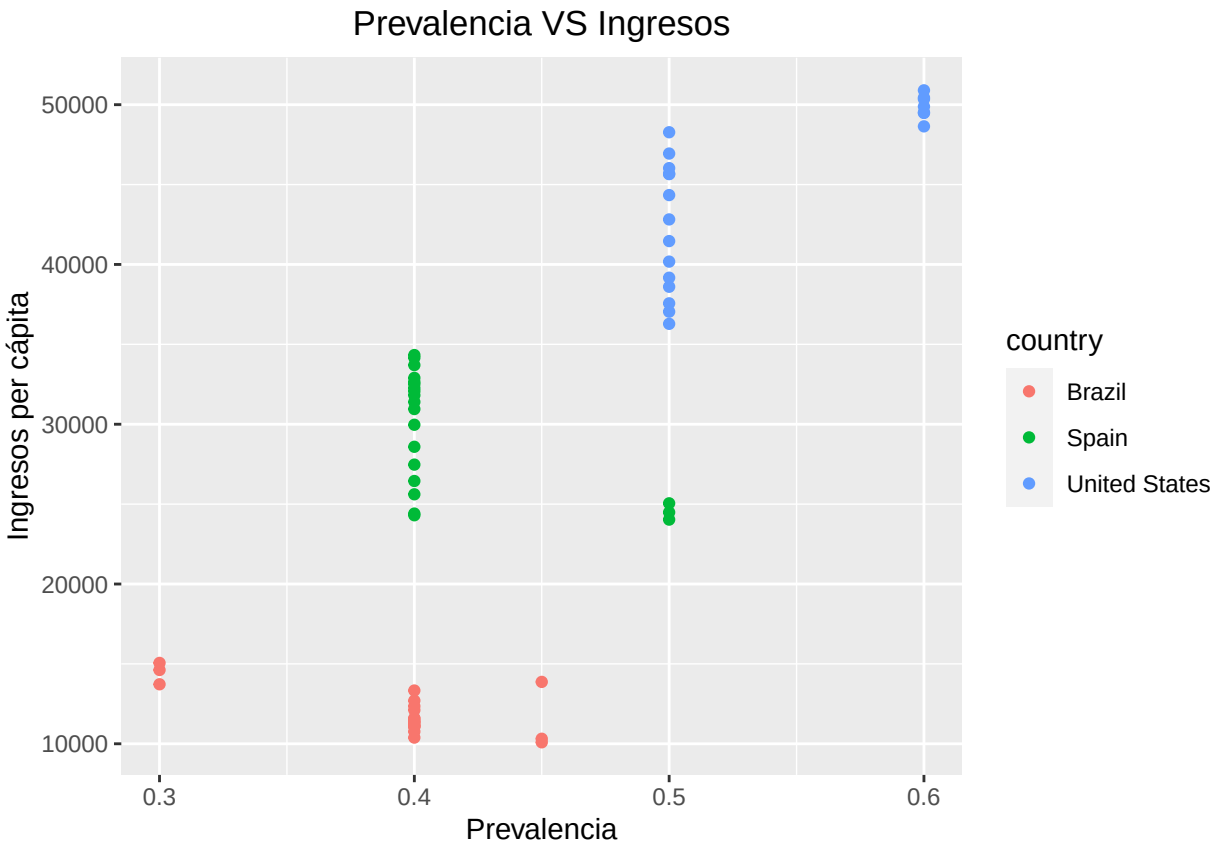
## Realiza los gráficos para responder a las preguntas formuladas

Usa diferentes colores para los 3 países: Brazil, Spain y United States.

### 1. Gráfico prevalencia vs Ingresos per cápita.

Primero debemos elegir los 3 países con los que trabajaremos y luego realizar un gráfico de puntos donde mediante el color se indiquen los 3 países seleccionados.

```
gap_hiv %>%
  filter(country %in% c("Brazil","Spain","United States") ) %>% #El operador %in% verifica si el valor
  ggplot(aes(x = prevalence, y = income, colour = country)) + #grafica prevalencia vs income, por país
  geom_point() +
  xlab("Prevalencia") + ylab("Ingresos per cápita") + ggtitle("Prevalencia VS Ingresos") +
  theme(plot.title=element_text(hjust=0.5))
```

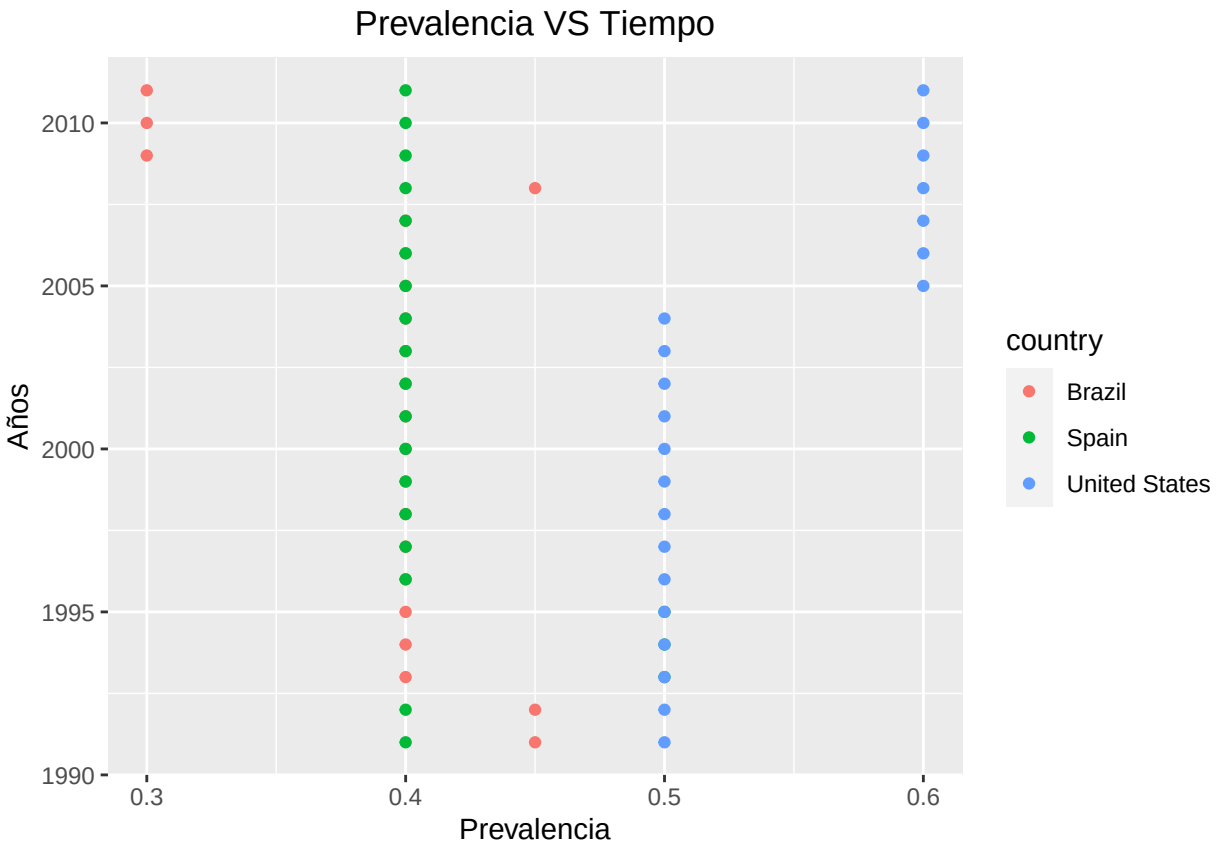



Parece ser que la relación entre la prevalencia del VIH y los ingresos per cápita es que a mayores ingresos, mayor prevalencia de esta enfermedad (en este caso sería en EEUU).

**2. Prevalencia vs Tiempo.**

El mismo procedimiento que antes pero en lugar de ingresos per cápita utilizaremos la variable tiempo.

```
gap_hiv %>%
  filter(country %in% c("Brazil","Spain","United States") ) %>%
  ggplot(aes(x = prevalence, y = year, colour = country)) + #grafica prevalencia vs tiempo, por país
  geom_point() +
  xlab("Prevalencia") + ylab("Años") + ggtitle("Prevalencia VS Tiempo") +
  theme(plot.title=element_text(hjust=0.5))
```




Desde 1990, el país que de nuevo ha tenido una mayor prevalencia de VIH, ha sido EEUU, con una prevalencia de entre 0,5 y 0,6 a lo largo de estos 30 años, siendo mayor a partir del 2005. En España la prevalencia ha sido estable a lo largo de todos los años, mientras que en Brasil la prevalencia ha sido discontinua a lo largo de los años, con un pico entre el 1990 y 1995, y otro en los últimos años.