Spanish TDT movies report

Guillermo Diaz

2022-09-19

Índice

INTRODUCTION

Link

This line of code is quite magical. Change any opts_chunk here and it will change the entire document. You can set any option you want: `echo`, `include`, `warning` or `messages`.

```
knitr::opts_chunk$set(
    echo = TRUE,
    message = FALSE,
    warning = FALSE
)
```

LOAD AND CLEAN DATA

The data set contains 144 instances and 9 variables. From now on, we won't be using the variable `Description`, which included a brief synopsis of each film. Let's see a few examples of the data.

```
library(knitr)
df_no_desc <- df %>%
  select(1:8)
```

```
df_no_desc %>%
  head(6) %>%
  knitr::kable()
```

date_time	channel	sp_title	original_title	year	genre	country	length
2022-09-13 00:19:00	Paramount Network	Vanilla Sky	Vanilla Sky	2001	Drama	NA	NA
2022-09-13 00:42:00	Neox	Ruslan: la venganza del asesino	Driven to Kill	2009	Acción	NA	NA
2022-09-13 01:10:00	laSexta	La mujer del pastor	The Pastor's Wife	2011	Drama	NA	NA
2022-09-13 13:10:00	La 2	El sonido de un tambor	Cimarron: The Sound of a Drum	1968	Western	NA	NA
2022-09-13 16:05:00	TRECE	Comando secreto	The Secret War of Harry Frigg	1968	Comedia	NA	NA
2022-09-13 16:20:00	La 1	La cuchara de Elli	Tessa Hennig - Elli gibt den Löffel ab	2012	Drama	NA	NA

The variables names are formatted to work with them in R, not to be shown in a document. We can clean them.

```
df_clean <- df_no_desc %>%
  dplyr::rename(
    Date = date_time,
    Channel = channel,
    "Spanish title" = sp_title,
    "Original title" = original_title,
    Year = year,
    Genre = genre,
    Country = country,
    Length = length
  )

df_clean %>%
  head() %>%
  knitr::kable()
```

Date	Channel	Spanish title	Original title	Year	Genre	Country	Length
2022-09-13 00:19:00	Paramount Network	Vanilla Sky	Vanilla Sky	2001	Drama	NA	NA
2022-09-13 00:42:00	Neox	Ruslan: la venganza del asesino	Driven to Kill	2009	Acción	NA	NA
2022-09-13 01:10:00	laSexta	La mujer del pastor	The Pastor's Wife	2011	Drama	NA	NA
2022-09-13 13:10:00	La 2	El sonido de un tambor	Cimarron: The Sound of a Drum	1968	Western	NA	NA
2022-09-13 16:05:00	TRECE	Comando secreto	The Secret War of Harry Frigg	1968	Comedia	NA	NA

Date	Channel	Spanish title	Original title	Year	Genre	Country	Length
2022-09-13 16:20:00	La 1	La cuchara de Elli	Tessa Hennig - Elli gibt den Löffel ab	2012	Drama	NA	NA

This is so much better. I want to see some examples with `Country` and `Length` data.

```
df_clean %>%
  drop_na() %>%
  head() %>%
  knitr::kable()
```

Date	Channel	Spanish title	Original title	Year	Genre	Country	Length
2022-09-18 00:26:00	Neox	Tenemos que hablar	Tenemos que hablar	2016	Comedia	España	91 min
2022-09-18 00:35:00	Antena 3	Suplantación de identidad	The Cheating Pact	2013	Suspense	Estados Unidos	85 min
2022-09-18 00:53:00	Cuatro	Colonia V	The Colony	2013	Ciencia ficción	Canadá	95 min
2022-09-18 01:15:00	La 1	Amor, ladrón, diamantes	Liebe, Diebe, Diamanten	2015	Drama	Alemania	90 min
2022-09-18 01:30:00	TRECE	Sol naciente	Rising Sun	1993	Suspense	Estados Unidos	129 min
2022-09-18 01:45:00	Paramount Network	Shame	Shame	2011	Drama	Reino Unido	101 min

I don't like to see the unit in the `Length` column. I noticed to that film genres are in spanish. Let's clean both columns.

¿How many unique values there are in each variable?

```
library(purrr)
df %>% map_dbl(
  n_distinct
)
```

```
##      date_time        channel       sp_title original_title           year
##            131             10            137            137             55
##          genre        country         length    description
##             16              9             44            137
```

So, if there are 144 instances, why do we only have 137 movies? I guess some of them were broadcasted more than once. These are the ones:

```
df_clean %>%
  group_by(`Spanish title`) %>%
  summarise(emisiones = n()) %>%
  filter(emisiones > 1)
```

```
## # A tibble: 7 x 2
```

```
##   `Spanish title`                     emisiones
##   <chr>                                   <int>
## 1 ¡Viven!                                     2
## 2 Cómo entrenar a tu dragón 2                 2
## 3 Diario de Greg 3: Días de perros            2
## 4 Grace Kelly: Los millones perdidos          2
## 5 Indiana Jones y la última cruzada           2
## 6 Querido fotogramas                          2
## 7 Se llamaba Grace Kelly                      2
```

MAPPING THE DATA

First, we have to prepare the data. Country names are in spanish, let's translate them.

```
df_map <- df_clean %>%
  group_by(`Country`) %>%
  summarise(Movies = n())

df_map <- df_map %>%
  mutate(
    Country = case_when(
      Country == "Alemania" ~ "Germany",
      Country == "Canadá" ~ "Canada",
      Country == "España" ~ "Spain",
      Country == "Estados Unidos" ~ "United States",
      Country == "Italia" ~ "Italy",
      Country == "Reino Unido" ~ "England"
    )
  )
df_map
```

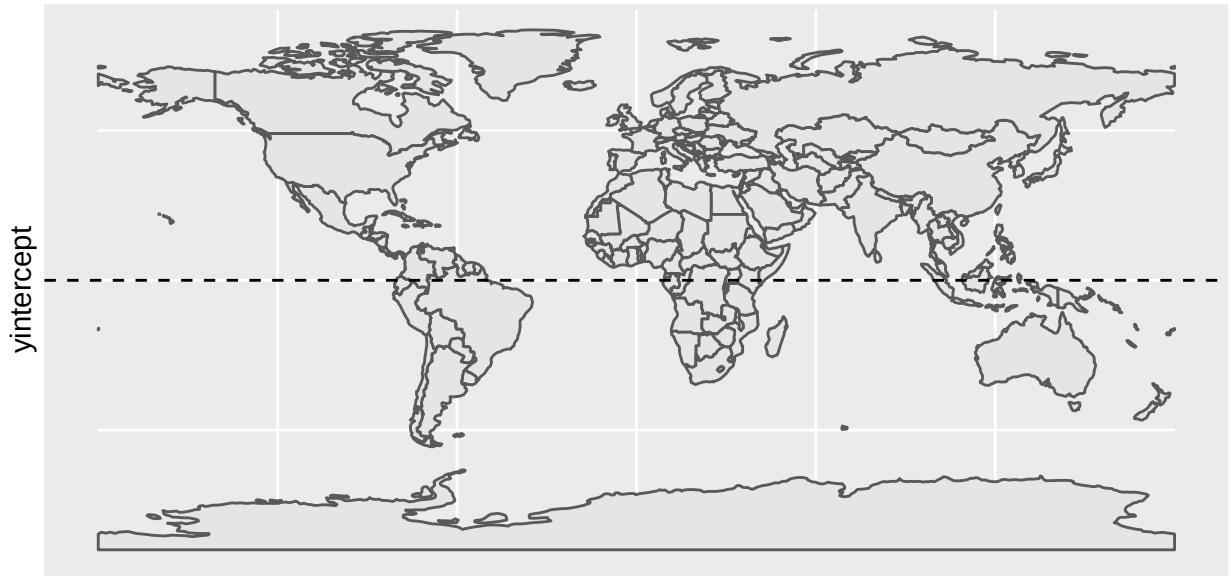
```
## # A tibble: 9 x 2
##   Country       Movies
##   <chr>          <int>
## 1 Germany            3
## 2 Canada             1
## 3 <NA>               1
## 4 Spain              4
## 5 United States     26
## 6 <NA>               1
## 7 Italy              1
## 8 England            2
## 9 <NA>             105
```

Then, we plot a map

```
library(ggplot2)
library(sf)
library(rnaturalearth)

world <- ne_countries(scale = "small", returnclass = "sf")
```

```
world %>%
  ggplot() +
  geom_sf() +
  geom_hline(yintercept = 0, linetype = "dashed")
```



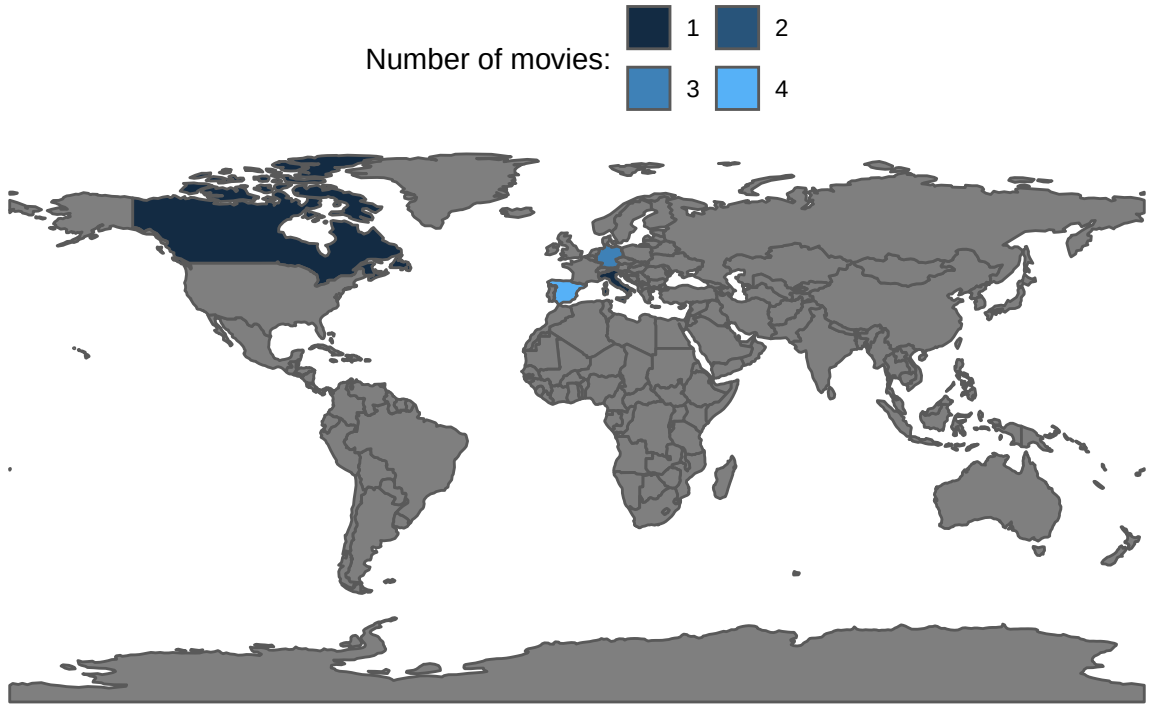
The latter was an empty world map. We are goint to fill it with our data. We create a suitable data frame.

```
world <- world %>%
  dplyr::rename("Country" = "sovereignt")

prueba <- left_join(world, df_map)
```

Let's see how it looks filled.

```
library(ggplot2)
library(sf)
prueba %>%
  ggplot() +
  geom_sf(aes(fill = Movies)) +
  theme_void() +
  theme(legend.position = "top") +
  labs(fill = "Number of movies:") +
  guides(fill = guide_legend(nrow = 2, byrow = TRUE))
```



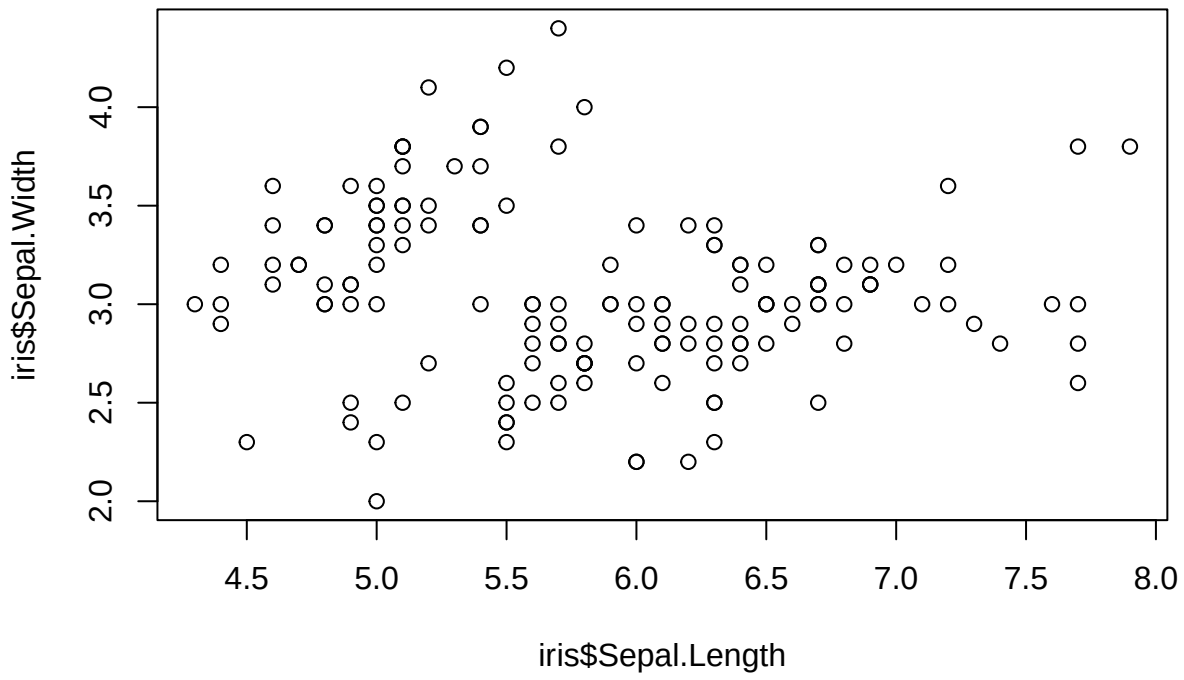
Con `results = 'hold'` lo que hacemos es sacar todos los resultados del tirón, sin partir el cuadro de código cada vez que hay un resultado que mostrar[1].

```
data(iris)
a <- 1+1
print(a)
head(iris)
```

```
## [1] 2
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

Note that the **echo = FALSE** parameter was added to the code chunk to prevent printing of the R code that generated the plot.

[1] Me marco un pie de página absolutamente escandaloso.




```
plot(iris$Sepal.Length, iris$Sepal.Width)
```

```
data("HairEyeColor")
chisq.test(HairEyeColor[, , 2])
```

```
##
##  Pearson's Chi-squared test
##
## data:  HairEyeColor[, , 2]
## X-squared = 106.66, df = 9, p-value < 2.2e-16
```

```
library(vcd)
```

¡Cuidado con el nombre de las variables!

```
library(ggplot2)
data(iris)

ggplot(iris, aes(Sepal.Width, Sepal.Legth)) +
  geom_point()
```

Real Stuff with TDT data

```
df <- read.csv("https://raw.githubusercontent.com/GuilleDiaz7/Automatic-Web-Scraping-of-Spanish-TDT-Film
               fileEncoding = "UTF-8")
library(tidyr)
df_clean <- df %>%
  drop_na()
library(dplyr)
head(
  df_clean %>% select(
  1:8
  )
)
```

```
##             date_time           channel                   sp_title
## 1 2022-09-18 00:26:00              Neox        Tenemos que hablar
## 2 2022-09-18 00:35:00          Antena 3 Suplantación de identidad
## 3 2022-09-18 00:53:00            Cuatro                 Colonia V
## 4 2022-09-18 01:15:00              La 1   Amor, ladrón, diamantes
## 5 2022-09-18 01:30:00             TRECE              Sol naciente
## 6 2022-09-18 01:45:00 Paramount Network                     Shame
##            original_title year               genre        country  length
## 1      Tenemos que hablar 2016             Comedia          España  91 min
## 2       The Cheating Pact 2013 Suspense / Thriller Estados Unidos  85 min
## 3              The Colony 2013     Ciencia ficción          Canadá  95 min
## 4 Liebe, Diebe, Diamanten 2015               Drama       Alemania  90 min
## 5              Rising Sun 1993 Suspense / Thriller Estados Unidos 129 min
## 6                   Shame 2011               Drama    Reino Unido 101 min
```

Para crear *listas*:

- Un elemento
- Otro elemento
  - Otro elemento más

1. Elemento 1
2. Elemento 2
3. Elemento 3

Create tables the handy way

```
library(palmerpenguins)
penguins %>%
  head() %>%
  knitr::kable()
```

Para escribir una cita en bloque

> Cara antigua. Es **code block** en la opción *Format* del Editor Visual.