Spanish TDT movies report

Guillermo Diaz

2022-09-20

Índice

Introduction

GitHub

This line of code is quite magical. Change any `opts_chunk` here and it will change setting for 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 166 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.

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

df_no_desc %>%
  head(6) %>%
  knitr::kable() %>%
  kable_styling(bootstrap_options = "striped", full_width = T)
```

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.

```r
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	Ye
2022-09-13 00:19:00	Paramount Network	Vanilla Sky	Vanilla Sky	200
2022-09-13 00:42:00	Neox	Ruslan: la venganza del asesino	Driven to Kill	200
2022-09-13 01:10:00	laSexta	La mujer del pastor	The Pastor's Wife	201
2022-09-13 13:10:00	La 2	El sonido de un tambor	Cimarron: The Sound of a Drum	196
2022-09-13 16:05:00	TRECE	Comando secreto	The Secret War of Harry Frigg	196
2022-09-13 16:20:00	La 1	La cuchara de Elli	Tessa Hennig - Elli gibt den Löffel ab	201

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

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

Date	Channel	Spanish title	Original title	Year	Genre
2022-09-18 00:26:00	Neox	Tenemos que hablar	Tenemos que hablar	2016	Comedia
2022-09-18 00:35:00	Antena 3	Suplantación de identidad	The Cheating Pact	2013	Suspense
2022-09-18 00:53:00	Cuatro	Colonia V	The Colony	2013	Ciencia ficció
2022-09-18 01:15:00	La 1	Amor, ladrón, diamantes	Liebe, Diebe, Diamanten	2015	Drama
2022-09-18 01:30:00	TRECE	Sol naciente	Rising Sun	1993	Suspense
2022-09-18 01:45:00	Paramount Network	Shame	Shame	2011	Drama

I don't like to see the unit in the `Length` column. I also noticed to that film genres are in Spanish. Let's clean the length.

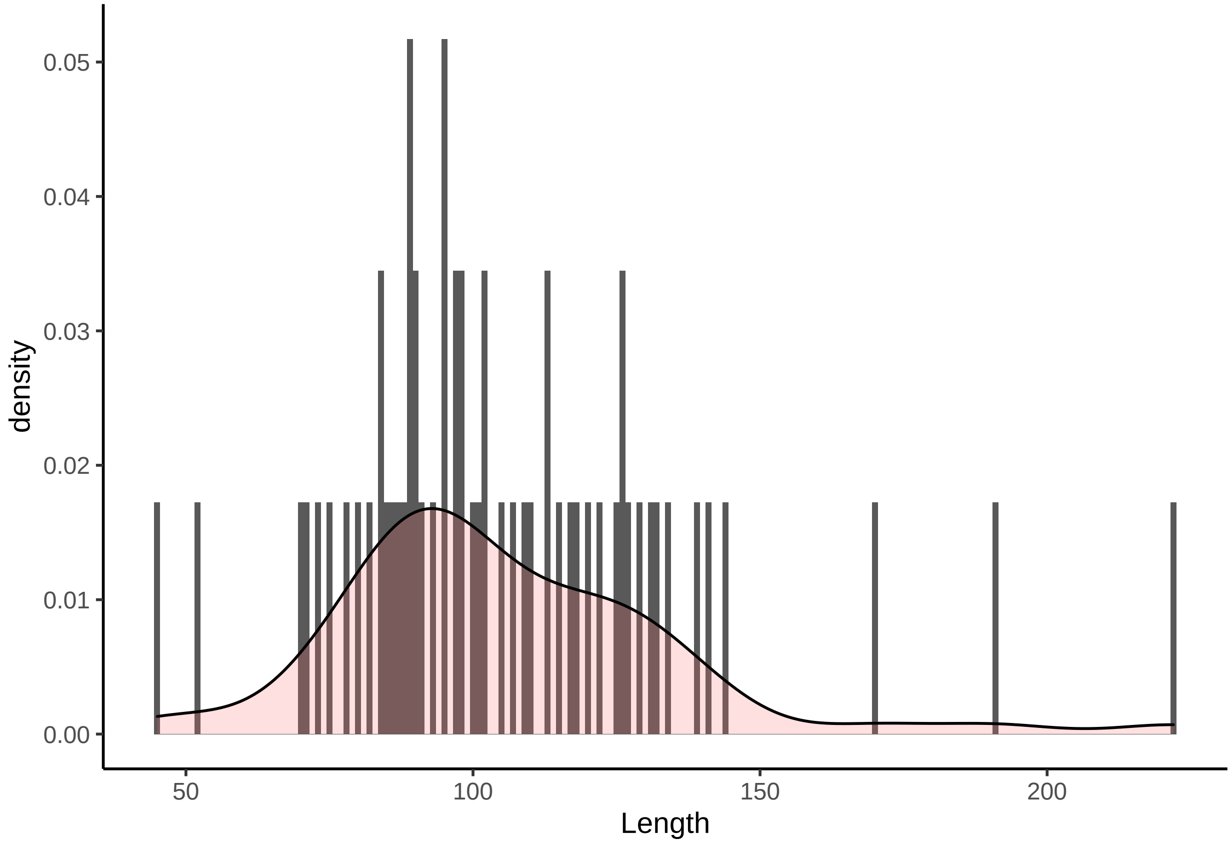
```r
library(readr)
df_clean$Length <- parse_number(df_clean$Length)
```

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

Date	Channel	Spanish title	Original title	Year	Genre
2022-09-18 00:26:00	Neox	Tenemos que hablar	Tenemos que hablar	2016	Comedia
2022-09-18 00:35:00	Antena 3	Suplantación de identidad	The Cheating Pact	2013	Suspense
2022-09-18 00:53:00	Cuatro	Colonia V	The Colony	2013	Ciencia ficció
2022-09-18 01:15:00	La 1	Amor, ladrón, diamantes	Liebe, Diebe, Diamanten	2015	Drama
2022-09-18 01:30:00	TRECE	Sol naciente	Rising Sun	1993	Suspense
2022-09-18 01:45:00	Paramount Network	Shame	Shame	2011	Drama

How long are the films show in TV?

```r
library(ggplot2)
df_clean %>%
  drop_na() %>%
  ggplot(aes(x = Length)) +
    geom_histogram(aes(y=..density..), binwidth = 1) +
    geom_density(alpha = 0.2, fill = "#FF6666") +
    theme_classic()
```



¿How many unique values there are in each variable?

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

```
##       date_time         channel        sp_title original_title           year
##             151              11             157            157             59
##           genre         country          length    description
##              17               9              56            157
```

So, if there are 166 instances, why do we only have 157 movies? I guess some of them were broadcasted more than once. These are the ones:

```r
df_clean %>%
  group_by(`Spanish title`) %>%
  summarise(Emisiones = n()) %>%
  filter(Emisiones > 1) %>%
  arrange(desc(Emisiones)) %>%
  knitr::kable()
```

Spanish title	Emisiones
Querido fotogramas	3
¡Viven!	2
Breakdown	2
Cómo entrenar a tu dragón 2	2
Diario de Greg 3: Días de perros	2
Grace Kelly: Los millones perdidos	2
Indiana Jones y la última cruzada	2
Se llamaba Grace Kelly	2

Mapping the data

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

```
unique(df_clean$Country)
```

```
## [1] NA               "España"         "Estados Unidos" "Canadá"
## [5] "Alemania"       "Reino Unido"    "Italia"         "Corea del Sur"
## [9] "Francia"
```

```
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 of America",
      Country == "Italia" ~ "Italy",
      Country == "Reino Unido" ~ "United Kingdom",
      Country == "Corea del Sur" ~ "South Korea",
      Country == "Francia" ~ "France",
      is.na(Country) == TRUE ~ "Sin datos"
    )
  )

df_map
```

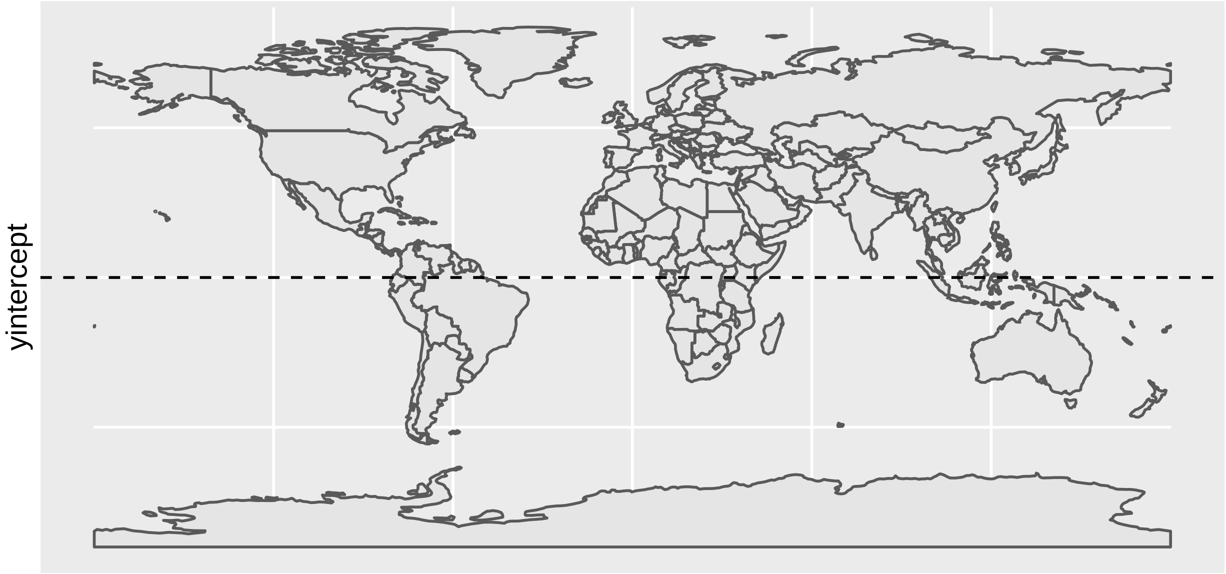
```
## # A tibble: 9 x 2
##   Country                  Movies
##   <chr>                     <int>
## 1 Germany                       3
## 2 Canada                        1
## 3 South Korea                   1
## 4 Spain                         8
## 5 United States of America     39
## 6 France                        1
## 7 Italy                         1
## 8 United Kingdom                4
## 9 Sin datos                   108
```

Then, we plot a map

```
library(sf)
library(rnaturalearth)

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

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



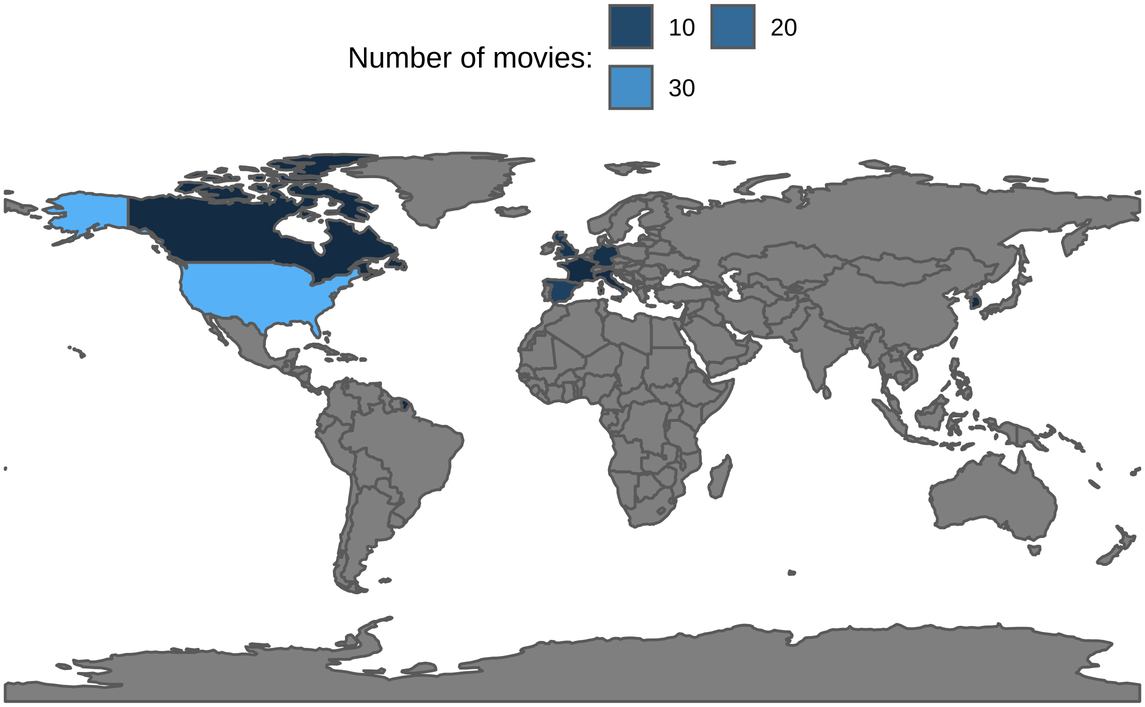
The latter was an empty world map. We are goint to fill it with our data. We create a suitable data frame.

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

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

Let's see how it looks filled.

```
library(ggplot2)
library(sf)
df_world %>%
  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))
```


Number of movies:
10
20
30