Nobel Project

Eduardo Chamizo

2025-06-17

Resumen

Este proyecto indaga en las técnicas propias para la visualización y análisis exploratorio del conjunto de datos “Nobel Prize”, empleando diversas librerías para su consecución.

Objetivo del problema

Explicar los datos de manera tabular y visual, analizando patrones significativos y otros resultados destacables.

Librerías

- httr: conexión con APIs.
- jsonlite: maneja datos JSON.
- lubridate: manejo de fechas.
- dplyr: manipulación de datos.
- ggplot2: gráficos.
- tidyverse: conjunto de paquetes.
- maps: datos geográficos.
- countrycode: codificación de países.
- tidytext: análisis de texto.
- wordcloud: nubes de texto.

```
library(httr)
library(jsonlite)
library(lubridate)
library(dplyr)
library(ggplot2)
library(tidyverse)
library(maps)
library(countrycode)
library(tidytext)
library(wordcloud)
```

Carga de datos desde el API

```
url_laureados <- "http://api.nobelprize.org/v1/laureate.csv"

queryString_laureados <- list(gender = "All")

response_laureados <- VERB("GET", url_laureados, query = queryString_laureados, content_type("applicati

df_laureados <- read.csv(textConnection(content(response_laureados, "text")),
                         stringsAsFactors = FALSE)

df_laureados<-df_laureados[c(-1,-9,-10,-11,-15,-16)]

df_final<-df_laureados
```

Procesamiento de datos

Limpieza inicial

Cambiar strings vacíos por NAs

```
columnasstringvacions <- c(colnames(df_final))

df_final[columnasstringvacions] <- lapply(df_final[columnasstringvacions],
                                          function(x) ifelse(x == "", NA, x))

df_final <- df_final[df_final$gender=="male" | df_final$gender=="female", ]
```

Estandarización de nomenclatura de algunas columnas

Columnas con fechas

```
df_final$born <- ymd(df_final$born)

df_final$died <- ymd(df_final$died)

df_final$year <- ymd(paste0(df_final$year, "-01-01"))

df_final$age_at_death <- trunc(as.numeric(difftime(df_final$died, df_final$born,
                                                   units = "days")) / 365.25)
df_final$age_at_prize<-trunc(as.numeric(difftime(df_final$year,df_final$born,
                                                 units = "days"))/365.25)
```

Renombrar algunos países

```
df_final$bornCountry <- sub(".*\\(now ([^)]+)\\)", "\\1", df_final$bornCountry)

df_final <- df_final %>%
  mutate(
    bornCountry = case_when(
      bornCountry %in% c("England", "Scotland", "Wales", "Northern Ireland", "UK", "United Kingdom", "G
      bornCountry %in% c("USA", "U.S.", "U.S.A.", "United States", "United States of America") ~ "USA",
      bornCountry == "Faroe Islands (Denmark)" ~ "Denmark",
      bornCountry == "East Timor" ~ "Timor-Leste",
      bornCountry == "Guadeloupe France" ~ "France",
      bornCountry == "Trinidad and Tobago" ~ "Trinidad and Tobago",
      bornCountry == "the Netherlands" ~ "Netherlands",
      TRUE ~ bornCountry
    )
  )
```

Quitamos el "now" en el nombre de ex-países y sustituimos por el nombre final. Ello se realiza para facilitar la codificación posterior, necesaria para mostrar visualmente los países.

Análisis exploratorio

Estadísticas básicas

```
summary(df_final[, c("born", "died", "year", "age_at_death")])
```

```
##       born                 died                 year
##  Min.   :1817-11-30   Min.   :1903-11-01   Min.   :1901-01-01
##  1st Qu.:1893-06-03   1st Qu.:1958-07-17   1st Qu.:1950-01-01
##  Median :1918-07-16   Median :1987-05-17   Median :1979-01-01
##  Mean   :1913-07-26   Mean   :1980-12-24   Mean   :1974-02-10
##  3rd Qu.:1939-02-24   3rd Qu.:2009-09-10   3rd Qu.:2003-01-01
##  Max.   :1997-07-12   Max.   :2025-05-28   Max.   :2024-01-01
##  NA's   :19           NA's   :298
##   age_at_death
##  Min.   : 39.00
##  1st Qu.: 74.00
##  Median : 82.00
##  Mean   : 80.69
##  3rd Qu.: 89.00
##  Max.   :103.00
##  NA's   :299
```

Métricas cuantitativas por género

```
df_final %>%
  group_by(gender) %>%
  summarise(
```

```
    mean_age_death = mean(age_at_death, na.rm = T),
    min_age_death = min(age_at_death, na.rm = T),
    max_age_death = max(age_at_death, na.rm = T),
    mean_age_prize = mean(age_at_prize,na.rm = T),
    min_age_prize = min(age_at_prize,na.rm=T),
    max_age_prize = max(age_at_prize,na.rm = T),
    n = n()
  )
```

```
## # A tibble: 2 x 8
##   gender mean_age_death min_age_death max_age_death mean_age_prize min_age_prize
##   <chr>           <dbl>         <dbl>         <dbl>          <dbl>         <dbl>
## 1 female           78.9            58           103           56.9            16
## 2 male             80.8            39           102           59.4            24
## # i 2 more variables: max_age_prize <dbl>, n <int>
```

Destaca la aplastante presencia varonil frente a la de mujeres. También, el hecho de que los hombres experimentan los dos eventos estudiados con posterioridad.

Top países con más apariciones

```
paises_top <- df_final %>%
  count(bornCountry, sort=T) %>%
  top_n(20)

paises_top
```

```
##     bornCountry   n
## 1           USA 297
## 2            UK 109
## 3       Germany  84
## 4        France  62
## 5        Sweden  30
## 6        Poland  29
## 7        Russia  29
## 8         Japan  28
## 9        Canada  21
## 10        Italy  20
## 11      Austria  19
## 12  Netherlands  19
## 13  Switzerland  19
## 14      Denmark  13
## 15       Norway  13
## 16        China  12
## 17      Hungary  11
## 18    Australia  10
## 19      Belgium   9
## 20        India   9
## 21 South Africa   9
```

Las principales potencias históricas son las que coronan el encabezado.

Diversidad geográfica

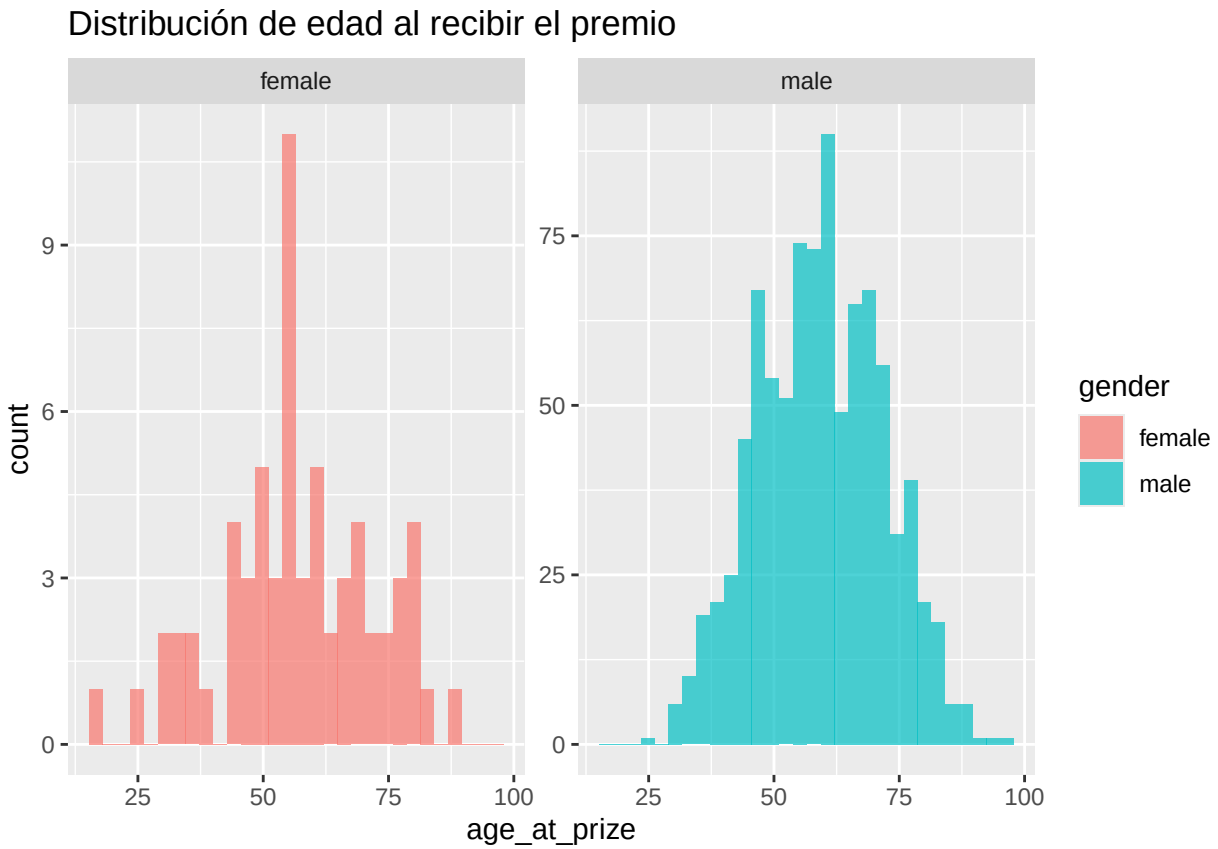
```
df_final %>%
  group_by(category) %>%
  summarise(
    n_paises = n_distinct(bornCountry),
    pais_dominante = names(sort(table(bornCountry), decreasing=T))[1])
```

```
## # A tibble: 6 x 3
##   category   n_paises pais_dominante
##   <chr>         <int> <chr>
## 1 chemistry        37 USA
## 2 economics        21 USA
## 3 literature       46 France
## 4 medicine         36 USA
## 5 peace            47 USA
## 6 physics          30 USA
```

Visualización de datos

Pirámide de edades por género

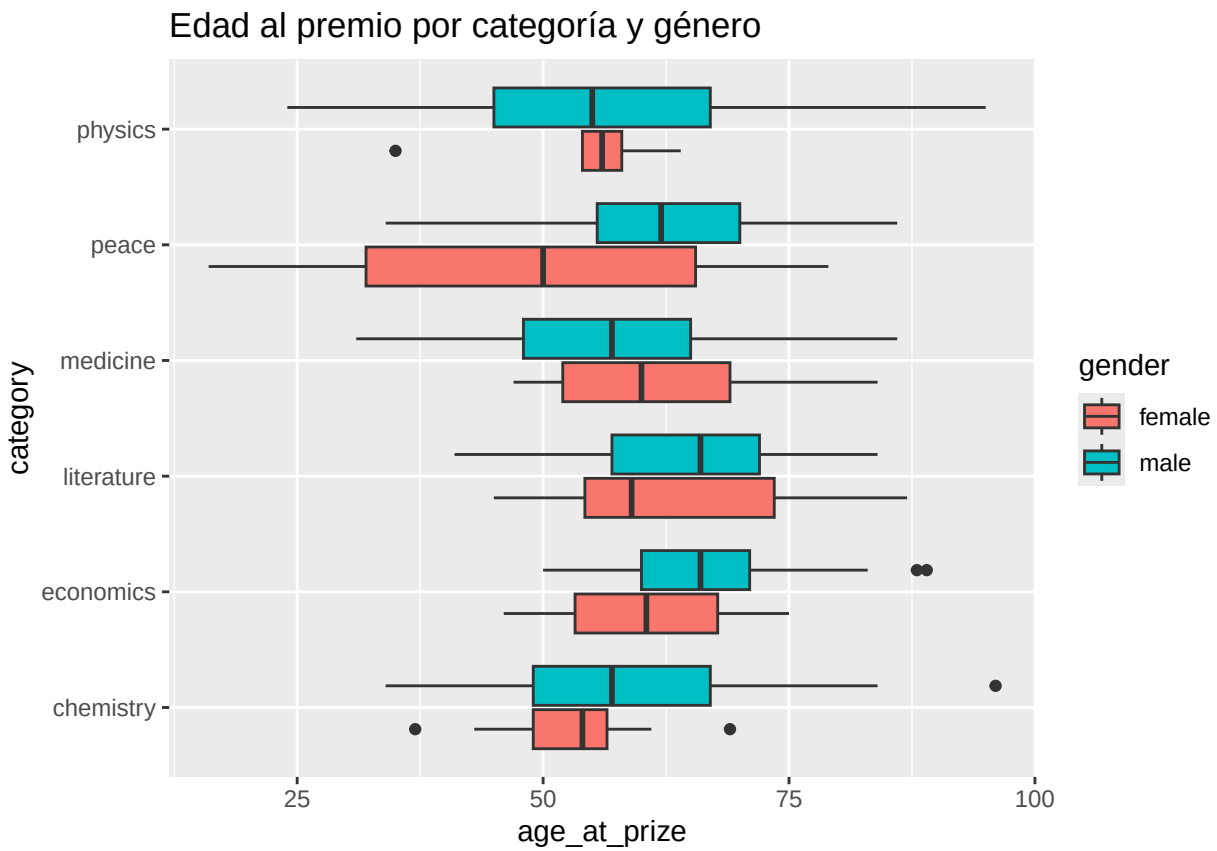
```
ggplot(df_final, aes(age_at_prize, fill=gender)) +
  geom_histogram(alpha=0.7, position="identity") +
  facet_wrap(~gender, scales="free_y") +
  labs(title="Distribución de edad al recibir el premio")
```



Se confirma con este gráfico lo que anteriormente se concluía de manera tabular.

Boxplot de edades por categoría y género

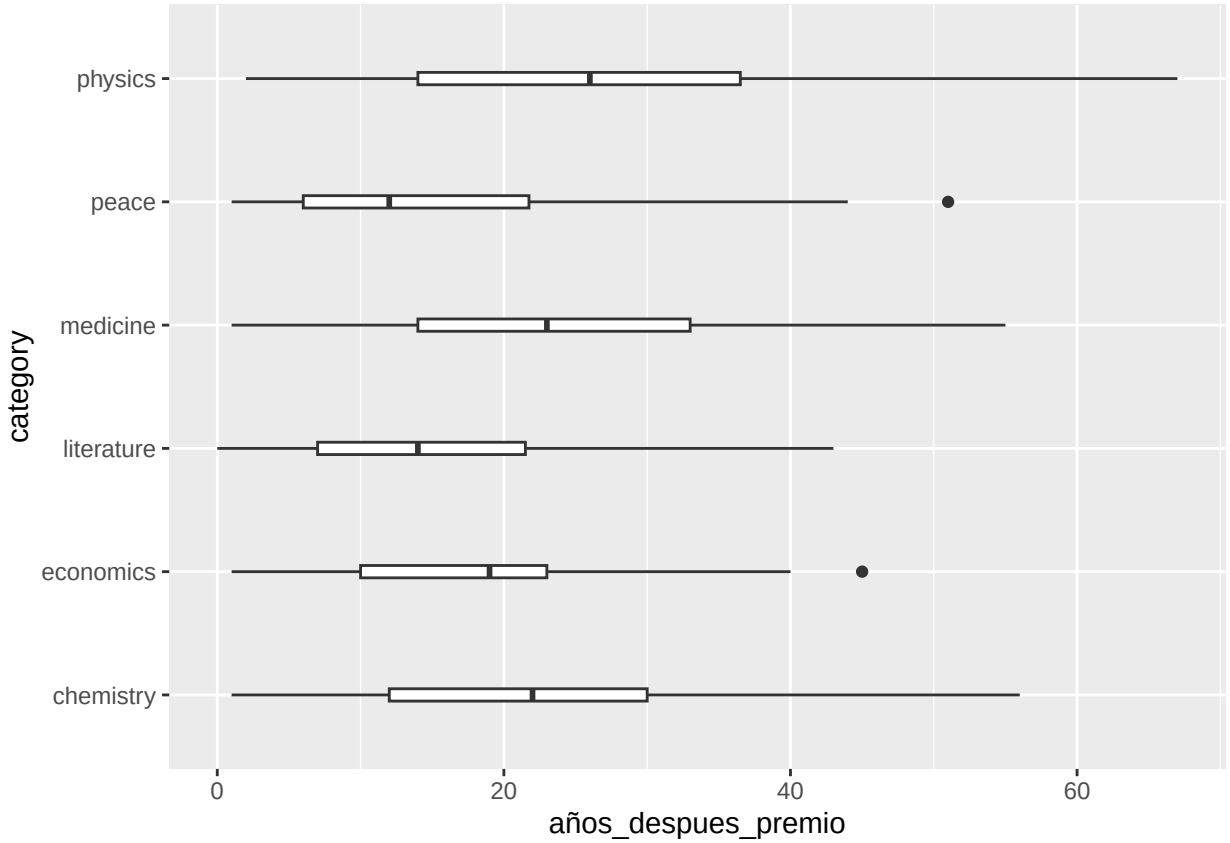
```
ggplot(df_final, aes(category, age_at_prize, fill=gender)) +
  geom_boxplot() +
  coord_flip() +
  labs(title="Edad al premio por categoría y género")
```



La edad de premiación típica oscila entre los 25 y 70 años. Especialmente en física, hay más jóvenes premiados que en otras categorías.

Longevidad post-premio

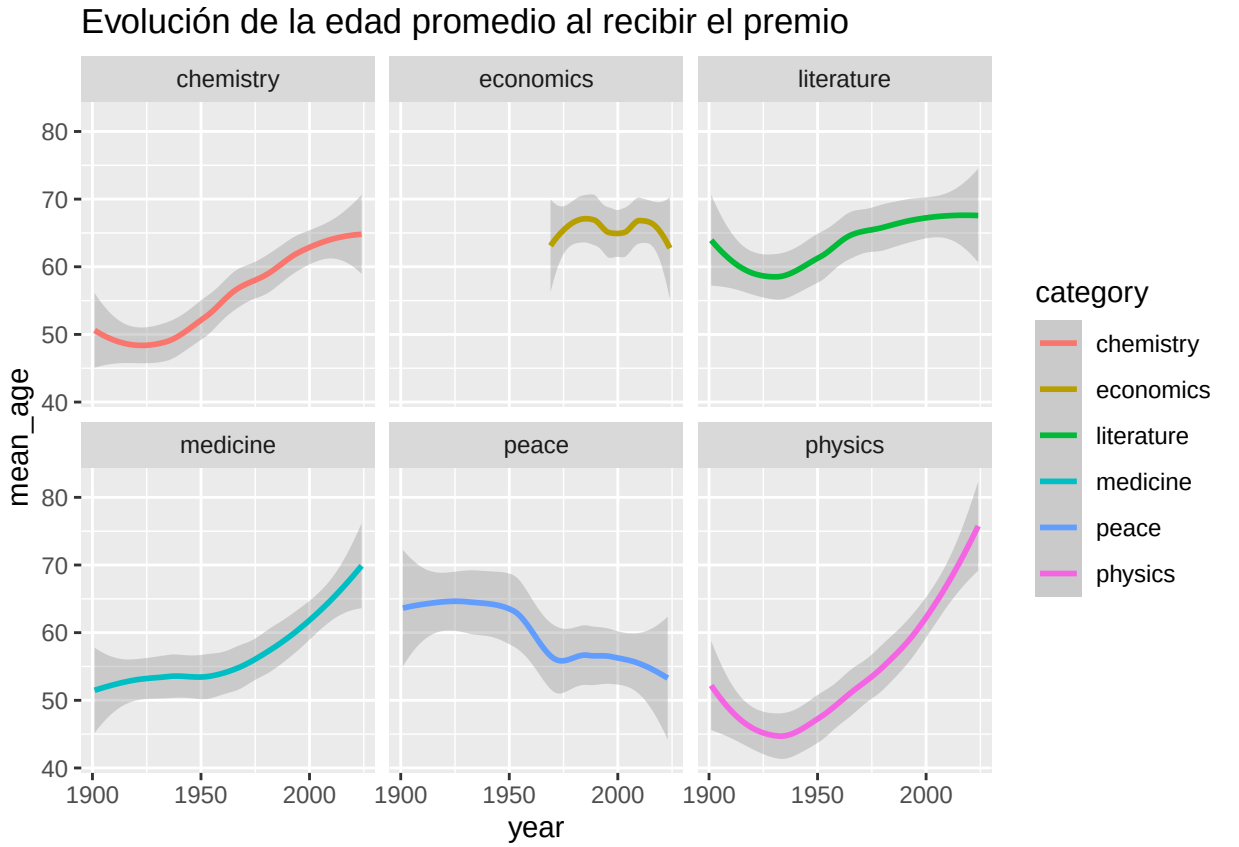
```
df_final %>%
  filter(!is.na(age_at_death)) %>%
  mutate(años_despues_premio = age_at_death - age_at_prize) %>%
  ggplot(aes(category, años_despues_premio)) +
  geom_boxplot(width=0.1) +
  coord_flip()
```



Los físicos/as, al soler recibir el premio más prematuramente como se vio antes, tienden a durar más. Esta afirmación nos hace pensar que el gráfico es consistente con el anterior.

Gráfico de líneas para la variación de la edad promedio al recibir el premio

```
df_final %>%
  group_by(year, category) %>%
  summarise(mean_age = mean(age_at_prize, na.rm=T)) %>%
  ggplot(aes(year, mean_age, color=category)) +
  geom_smooth(method="loess") +
  facet_wrap(~category) +
  labs(title="Evolución de la edad promedio al recibir el premio")
```



Mientras la tendencia general ha sido que los premiados cada vez sean más longevos, algo que podría achacarse a la mayor esperanza de vida; destacan los premiados por la paz, posiblemente fundamentado en el mayor componente activista que caracteriza a los jóvenes.

Mapa mundial

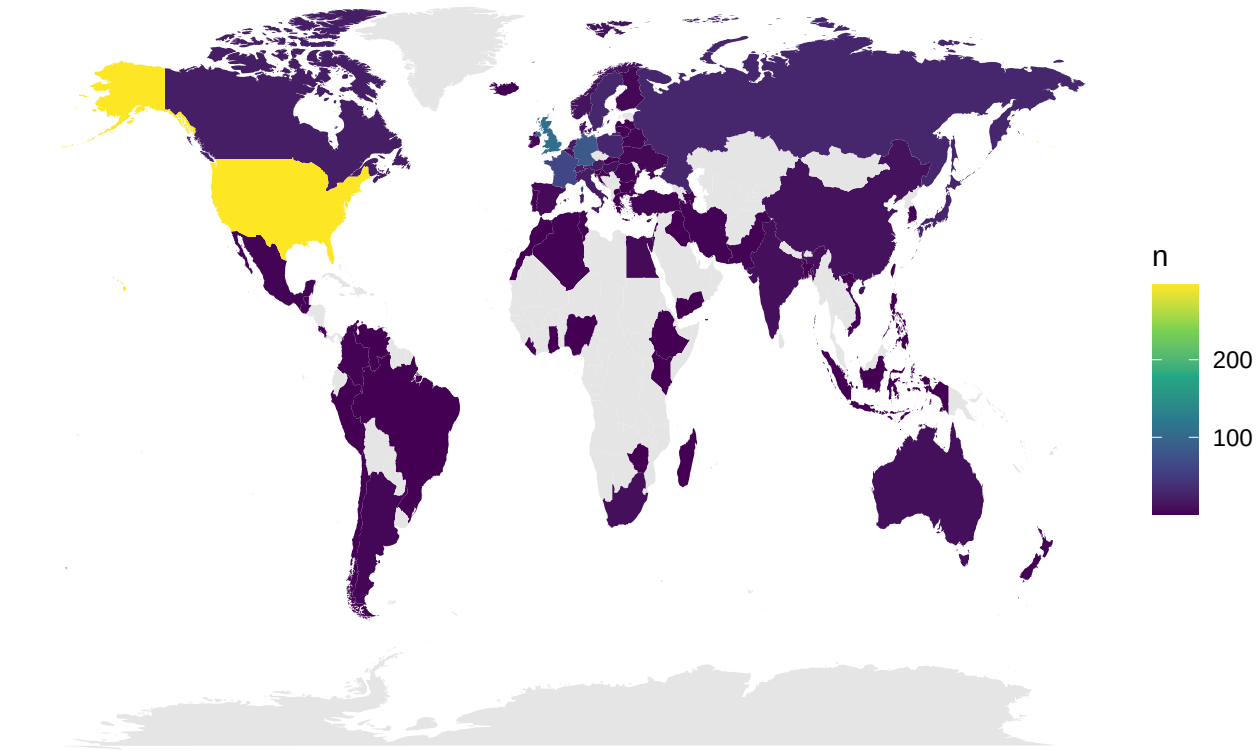
```
world_map <- map_data("world") %>%
  mutate(region = case_when(
    region == "USA" ~ "United States",
    region == "UK" ~ "United Kingdom",
    TRUE ~ region))

nobel_by_country <- df_final %>%
  count(bornCountry) %>%
  mutate(
    region = case_when(
      TRUE ~ countrycode(bornCountry, "country.name", "country.name", warn = FALSE))) %>%
  filter(!is.na(region))

world_nobel <- world_map %>%
  left_join(nobel_by_country, by = "region")

ggplot(world_nobel, aes(long, lat, group = group, fill = n)) +
  geom_polygon() +
  scale_fill_viridis_c(na.value = "grey90") +
```

```
  theme_void() +
  labs(title = "Distribución mundial de premios Nobel por país de nacimiento")
```



Visualización gráfica del número de premios por país.

Análisis de texto

Nube de texto según la categoría del premio (en nuestro caso “peace”)

```
motivaciones_words <- df_final %>%
  unnest_tokens(word, motivation) %>%
  anti_join(stop_words) %>%
  count(category, word, sort=T)

motivaciones_words %>%
  filter(category == "peace") %>%
  with(wordcloud(word, n, max.words=50))
```

Las palabras más frecuentes son “peace”, “international” y “efforts”. Esto concuerda con la coyuntura en que muchos de estos premios se otorgaron: tras dos guerras mundiales, la creación de la ONU, descolonización... Resulta interesante mencionar también la aparición de la palabra “nuclear” en un contexto ajeno a la física o química.

Conclusiones

- Se puede establecer una tendencia temporal en algunas de las características estudiadas, como la edad al recibir el premio.
- El mapa facilita mucho el reconocimiento de potencias no sólo históricas y económicas, sino también intelectuales.