Práctica 2: Limpieza y análisis de datos

Daniel Lugo Laguna Pablo Mora Galindo

Índice

PARTE 0. Descripción de la práctica a realizar

El objetivo de esta actividad será el tratamiento de un dataset, que puede ser el creado en la práctica 1 o bien cualquier dataset libre disponible en Kaggle (https://www.kaggle.com).

Los objetivos concretos de esta práctica son:

- Aprender a aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios o multidisciplinares.

- Saber identificar los datos relevantes y los tratamientos necesarios (integración, limpieza y validación) para llevar a cabo un proyecto analítico.
- Aprender a analizar los datos adecuadamente para abordar la información contenida en los datos.
- Identificar la mejor representación de los resultados para aportar conclusiones sobre el problema planteado en el proceso analítico.
- Actuar con los principios éticos y legales relacionados con la manipulación de datos en función del ámbito de aplicación.
- Desarrollar las habilidades de aprendizaje que les permitan continuar estudiando de un modo que tendrá que ser en gran medida autodirigido o autónomo.
- Desarrollar la capacidad de búsqueda, gestión y uso de información y recursos en el ámbito de la ciencia de datos.

PARTE I. Definición del dataset y objetivos del análisis

1. Descripción del dataset

¿Por qué es importante y a qué pregunta pretende responder?

El dataset elegido para esta práctica está titulado "COVID-19 World Vaccination Progress", recopilado diariamente en https://ourworldindata.org/ y referenciado en kaggle: https://www.kaggle.com/gpreda/covid-world-vaccination-progress

El conjunto de datos contiene el detalle de volúmenes de vacunaciones diarias de COVID-19 segregado por país, para gran parte de los países del mundo. Asímismo se muestran los diferentes tipos de vacunas suministradas en los mismos y el origen de los datos. Los campos presentes en el conjunto de datos son los siguientes:

- **country** - Nombre del país al que hacen referencia los datos.
- **iso_code** - Código ISO del país en cuestión.
- **date** - Fecha a la que corresponden los datos de vacunaciones.
- **total_vaccinations** - Número total de inmunizaciones para la fecha del registro y el pais en cuestión.
- **people_vaccinated** -Dependiendo de la pauta de inmunización del tipo de vacuna, una persona puedo recibir una o más dosis (típicamente dos). En cierto momento, el número de vacunaciones puede llegar a ser mayor que el número de personas vacunadas.
- **people_fully_vaccinated** - Número de personas con la pauta completa de dosis según la vacuna inoculada para la fecha del registro y el pais en cuestión.
- **daily_vaccinations_raw** - Número de dosis de la pauta de vacunación realizadas para la fecha del registro y el pais en cuestión. (Número sin corregir)
- **daily_vaccinations** - Número de dosis de la pauta de vacunación realizadas para la fecha del registro y el pais en cuestión.
- **total_vaccinations_per_hundred** número de vacunaciones respecto a la población total para la fecha del registro y el pais en cuestión.
- **people_vaccinated_per_hundred** Personas con al menos una dosis de la pauta de vacunación respecto a la población total para la fecha del registro y el pais en cuestión.

- **people_fully_vaccinated_per_hundred** Personas completamente inmunizadas respecto a la población total para la fecha del registro y el pais en cuestión.
- **daily_vaccinations_per_million** - Ratio en partes por millón (ppm) entre el número de vacunaciones y la población total para la fecha del registro y el pais en cuestión.
- **Vaccines** tipos de vacunas utilizadas en el país hasta la fecha.
- **Source name** - Fuente de la información de los datos (autoridad nacional, organización internacional, organización local)
- **source_website** - Sitio web de la fuente de datos.

Actualmente, el mundo sigue inmerso en la crisis originada por la pandemia de la COVID-19. Las consecuencias del virus están teniendo gran impacto tanto en la salud, como económicos y sociales a lo largo de gran parte de los paises del mundo. Sin embargo, la comunidad científica internacional lleva trabajando desde el estallido de la pandemia en distintas vacunas que permitan inmunizar a la población frente al virus. Actualmente existen ya vacunas lanzadas por distintos laboratorios: Astrazeneca, Pfizer, Moderna, Sputnik, Sinopharm. Actualmente, a fecha de Mayo de 2021, la mayoría de paises se encuentra en pleno proceso de vacunación, pero el avance en este proceso no avanza al mismo ritmo en todos los paises.

Este tema es **suficientemente relevante, por todas las consecuencias que ha traído el COVID-19, como para realizar un estudio en profundidad del proceso** y analizar los patrones que permitan comprender ciertos obstáculos en las vacunaciones por los distintos países.

Con el análisis de este conjunto de datos se pretende verificar y comparar estas diferencias en el proceso de vacunación entre territorios, así como realizar ciertas estimaciones y predicciones sobre la forma en que pueden evolucionar estas vacunaciones en el futuro próximo. Para ello se estudiará la elaboración de un modelo de regresión de series temporales, lo cual podría ser útil para proyectar el número de inmunizaciones con el ritmo de vacunación actual.

PARTE II: Análisis exploratorio y limpieza del dataset

2. Integración y selección de los datos de interés

En el dataset se estudia el progreso en la vacunación a través de distintos parámetros. Básicamente se trata de personas completamente inmunizadas, personas con al menos una dosis, ratio porcentual de personas vacunadas, dosis diarias inoculadas, etc. Asímismo se indica la fecha, el país y la fuente de los datos.

En primer lugar se procede a la carga de las librerías y el fichero de datos:

```
#tidyr y dplyr se utilizan a lo largo del documento
#para transformación de datos (filtrado, agrupación)
library(tidyr)
library(dplyr)
library(stringr)
#knitr: Para funciones específicas de formato de tablas en la exportación.
library(knitr)
library(ggplot2)
library(kableExtra)
library(lubridate)
library(data.table)
library(rworldmap)
library(countrycode)
library(corrplot)
```

```
library(nortest)
library(aTSA)
library(forecast)
```

A continuación se procede a la carga del dataset.

```
covid_file <-"country_vaccinations.csv"
covid_vac <-read.csv(covid_file,sep=",",encoding="UTF-8")
dim(covid_vac)
```

```
## [1] 20390    15
```

El fichero, tal como se indicó en el apartado anterior, contiene 15 variables con informes diarios de vacunación por pais. El número total de registros es de 20390. A continuación se incluye una pequeña muestra de los datos y el tipo de campo.

```
str(covid_vac)
```

```
## 'data.frame':    20390 obs. of  15 variables:
##  $ country                             : chr  "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan"
##  $ iso_code                            : chr  "AFG" "AFG" "AFG" "AFG" ...
##  $ date                                : chr  "2021-02-22" "2021-02-23" "2021-02-24" "2021-02-25" ...
##  $ total_vaccinations                  : num  0 NA NA NA NA NA 8200 NA NA NA ...
##  $ people_vaccinated                   : num  0 NA NA NA NA NA 8200 NA NA NA ...
##  $ people_fully_vaccinated             : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ daily_vaccinations_raw              : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ daily_vaccinations                  : num  NA 1367 1367 1367 1367 ...
##  $ total_vaccinations_per_hundred      : num  0 NA NA NA NA NA 0.02 NA NA NA ...
##  $ people_vaccinated_per_hundred       : num  0 NA NA NA NA NA 0.02 NA NA NA ...
##  $ people_fully_vaccinated_per_hundred: num  NA NA NA NA NA NA NA NA NA NA ...
##  $ daily_vaccinations_per_million      : num  NA 35 35 35 35 35 35 41 46 52 ...
##  $ vaccines                            : chr  "Oxford/AstraZeneca, Pfizer/BioNTech, Sinopharm/Beijing
##  $ source_name                         : chr  "World Health Organization" "World Health Organization"
##  $ source_website                      : chr  "https://covid19.who.int/" "https://covid19.who.int/" "
```

Se observa la existencia de 6 variables categóricas (fecha, país, tipo de vacunas y fuente de los datos) y 9 variables numéricas (todas las métricas de vacunación)

De cara al análisis que se pretende plantear, es necesario tener en cuenta diferentes dimensiones sobre los datos (espaciales, temporales y categorías de vacunas). Se dividirán por tanto los campos en los siguientes grupos de variables: localización, tiempo, métricas de vacunación, tipología de vacunas y fuentes.

- **Localización:** Se incluye tanto el nombre del pais *country* como el código ISO *iso_code* del mismo. Este segundo campo es redundante respecto a *country* pero será necesario para el enriquecimiento del dataset en este apartado.
- **Tiempo:** La única variable temporal es *Date*, la cual es fundamental para el análisis.
- **Métricas de vacunación:** Dado que el objetivo es verificar el ritmo de vacunación diario, la variable que mejor permite a priori mediar esta tendencia es *daily_vaccinations*. Dado que existe gran discordancia entre el volumen poblacional entre paises, de cara a las hipótesis a plantear así como el análisis de regresión, se utilizará 'daily_vaccinations_per_million' en lugar de 'daily_vaccinations' ya que el dato aparece en ppm, facilitando la comparación entre paises con un criterio normalizado. Se mantiene por tanto *daily_vaccinations_per_million*.

- **Tipología de vacunas:** Se incluye un listado de las vacunas que se están administrando para el pais y fecha. Los datos no contienen información de volumenes por cada vacuna individual, sino la suma de todas las vacunas que se aplican por pais. El campo *vaccines* puede ser relevante a nivel comparativo y se mantendrá.
- **Fuentes de datos:** Se asume que el trabajo realizado por *Our World in Data* es riguroso y, por tanto, más allá de garantizar la fidelidad del dato (procedente de fuentes autorizadas del país en cuestión), los campos de fuente serán eliminados para este análisis (*source_name*, *source_website*).

Se eliminarán los campos indicados durante el análisis de valores perdidos, de forma que se verifique en primer lugar si los mismos presentan bajo número de registros perdidos.

3. Limpieza de los datos y tratamiento de valores extremos

¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?

Antes de revisar la existencia de valores ya categorizados como NA, se comprobará la existencia de valores NA que puedan aparecer como cadena vacía (), interrogación o FALSE.

```
sum(colSums(covid_vac==""))
sum(colSums(covid_vac=="?"))
sum(colSums(covid_vac=="FALSE"))
```

```
## [1] NA
## [1] NA
## [1] NA
```

No existe ningún valor en el dataset igual a los mencionados, pero si existen ya valores categorizados como NA originalmente. Se comprueba el volumen total de estos valores.

```
paste("Total valores nulos: ", sum(is.na(covid_vac)))
paste("% valores nulos: ", ((sum(is.na(covid_vac))/prod(dim(covid_vac)))*100) %>%
        round(2))
```

```
## [1] "Total valores nulos:  71761"
## [1] "% valores nulos:  23.46"
```

Cerca de una cuarta parte de los valores del dataset son nulos. Se revisa a continuación el volumen de valores perdidos para todas las variables numéricas.

3.1 Variables numéricas

```
covid_vac %>%
    gather(key = "campo", value = "val") %>%
    mutate(is.missing = is.na(val)) %>%
    group_by(campo, is.missing) %>%
    summarise(num_NA = n()) %>%
    filter(is.missing==T) %>%
    select(-is.missing) %>%
    arrange(num_NA) %>% kable(caption="Resumen de volumen de NAs por campo")
```

Cuadro 1: Resumen de volumen de NAs por campo

campo	num_NA
daily_vaccinations	214
daily_vaccinations_per_million	214
total_vaccinations	8772
total_vaccinations_per_hundred	8772
people_vaccinated	9531
people_vaccinated_per_hundred	9531
daily_vaccinations_raw	10719
people_fully_vaccinated	12004
people_fully_vaccinated_per_hundred	12004

```
covid_na_det <- covid_vac[is.na(covid_vac$daily_vaccinations_per_million),
                          c("country","date","daily_vaccinations_per_million")]
```

Como puede comprobarse en la tabla anterior, los campos **daily_vaccinations** y **daily_vaccinations_per_million** tienen un volumen de valores perdidos significativamente inferior comparado con el resto de métricas (solo un 1 % frente al total de registros del dataset son valores perdidos de este campo). Por tanto, termina de confirmar el campo **daily_vaccinations_per_million** como métrica principal del análisis, tal como se indicó en el apartado 2.

Se toma una muestra de los 214 valores perdidos del dataset.

```
  covid_na_det %>%
  group_by(country) %>%
  summarise(no = n()) %>%
  arrange(desc(no)) %>% slice(1:5)
```

```
## # A tibble: 5 x 2
##   country        no
##   <chr>       <int>
## 1 Guinea          3
## 2 Brunei          2
## 3 Afghanistan     1
## 4 Albania         1
## 5 Algeria         1
```

El volumen de registros perdidos para cada país (grupos sobre los que se aplicarán los análisis estadísticos posteriormente) es muy bajo, solo hay dos países que tengan más de un registro perdido.

Estos NAs pueden deberse distintas razones: a un dato no publicado por el país para el día concreto, a una fecha muy inicial donde aun no se había comenzado a vacunar a la población o a datos de poca calidad por la fuente original. En todo caso, el análisis posterior estará centrado en países europeos y ninguno va a tener más de un valor perdido.

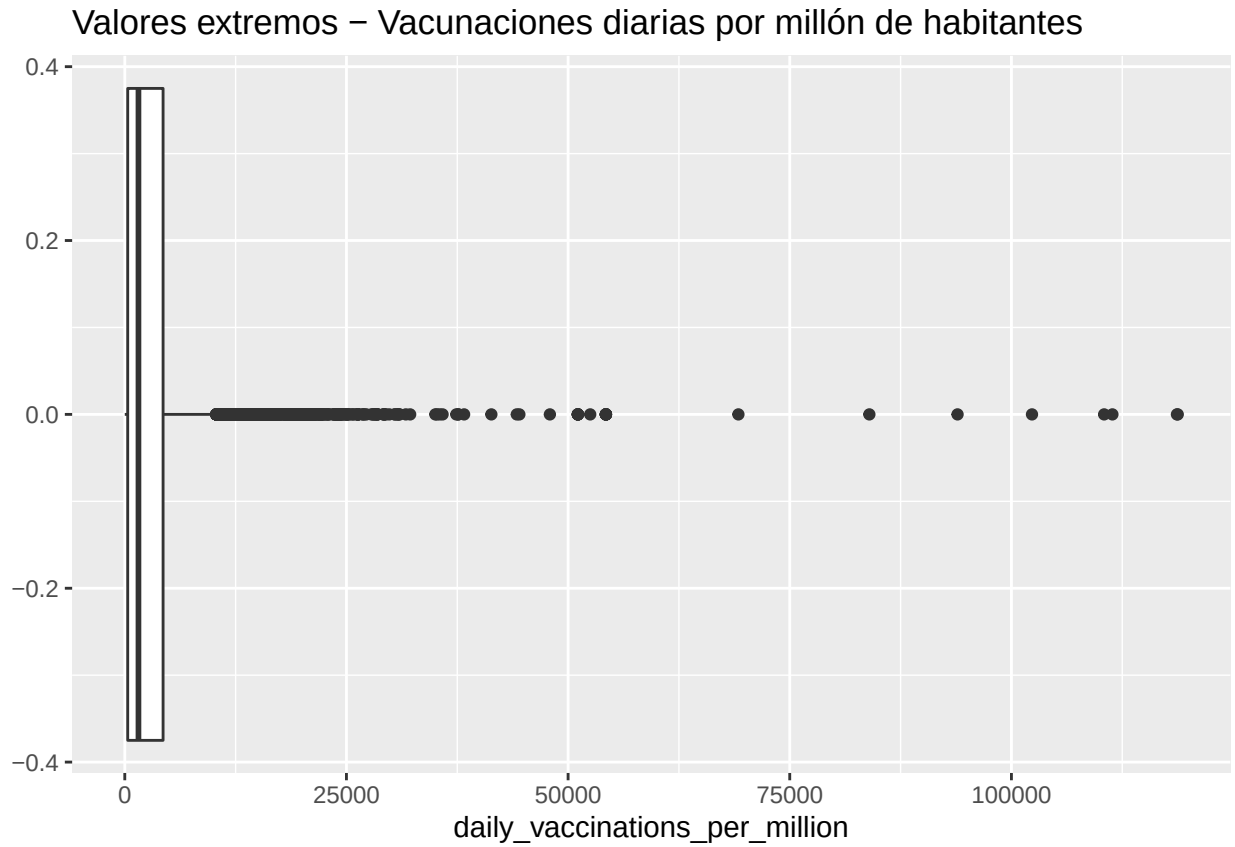
Dicho lo anterior, se imputará a 0 para todos los NAs de **daily_vaccinations_per_million** y **daily_vaccinations** (tomando como hipótesis que se deben a ninguna vacunación realizada ese día en ese país).

```
covid_vac_clean <- covid_vac
covid_vac_clean$daily_vaccinations_per_million[is.na(covid_vac_clean$daily_vaccinations_per_million)]<-
```

```
  0
covid_vac_clean$daily_vaccinations[is.na(covid_vac_clean$daily_vaccinations)] <-
  0
```

En cuanto a los valores extremos de estas dos variables mencionadas, se estudiarán a partir de un diagrama de caja para la variable *daily_vaccinations_per_million* (ya que *daily_vaccinations* está contenida en esta última).

```
p <- ggplot(covid_vac_clean, aes(daily_vaccinations_per_million))
p + geom_boxplot() +
  ggtitle("Valores extremos - Vacunaciones diarias por millón de habitantes")
```



Se observan valores extremos en la gráfica anterior. Se revisarán estos valores en más detalle.

```
OutVals = boxplot(covid_vac_clean$daily_vaccinations_per_million, plot=FALSE)$out
out_ind <- which(covid_vac_clean$daily_vaccinations_per_million %in% c(OutVals))
covid_out <- covid_vac_clean[out_ind,c("country","date","daily_vaccinations",
                                       "daily_vaccinations_per_million") ]
head(arrange(covid_out,desc(daily_vaccinations_per_million)), n = 10)
```

```
##             country       date daily_vaccinations
## 1            Bhutan 2021-03-28              91636
## 2            Bhutan 2021-03-29              91568
## 3            Bhutan 2021-03-27              85949
## 4            Bhutan 2021-03-30              85229
```

```
## 5            Bhutan 2021-03-31              78953
## 6            Bhutan 2021-04-01              72473
## 7            Bhutan 2021-04-02              64799
## 8            Bhutan 2021-04-03              53400
## 9  Falkland Islands 2021-02-08                189
## 10 Falkland Islands 2021-02-09                189
##    daily_vaccinations_per_million
## 1                          118759
## 2                          118671
## 3                          111389
## 4                          110456
## 5                          102322
## 6                           93924
## 7                           83979
## 8                           69206
## 9                           54264
## 10                          54264
```

```
covid_vac_clean %>%
  group_by(country) %>%
  count() %>%
  filter(country == 'Bhutan' || country == 'Spain')
```

```
## # A tibble: 2 x 2
## # Groups:   country [2]
##   country     n
##   <chr>   <int>
## 1 Bhutan     61
## 2 Spain     142
```

Para los valores más extremos, se presentan dos casuísticas. Por un lado Bután, donde la vacunación se realizó masivamente en un período más corto de tiempo y con un volumen de población relativamente bajo (en torno a 765.000 personas). Por otro lado para las islas malvinas, el volumen de población no llega a 3000 personas, por lo que las vacunaciones por millón de habitantes se disparan, a pesar de que las diarias no llegan a 200 en los casos mostrados.

A partir de estos ejemplos extremos y, dado que los datos se han obtenido de fuentes oficiales gubernamentales en la gran mayoría de casos, se decide no realizar imputación ni eliminación de valores extremos.

3.2 Variables categóricas

Se comprueba que para las variables categóricas (country, date, vaccines, source_name y source_website) no existen valores perdidos, por tanto no es necesario realizar ninguna eliminación y/o imputación de datos.

```
sapply(covid_vac_clean, function(x) sum(is.na(x)))
```

```
##                        country                       iso_code
##                              0                              0
##                           date              total_vaccinations
##                              0                           8772
##               people_vaccinated         people_fully_vaccinated
##                           9531                          12004
```

```
##              daily_vaccinations_raw                         daily_vaccinations
##                                10719                                          0
##      total_vaccinations_per_hundred               people_vaccinated_per_hundred
##                                8772                                       9531
## people_fully_vaccinated_per_hundred             daily_vaccinations_per_million
##                               12004                                          0
##                            vaccines                                source_name
##                                   0                                          0
##                      source_website
##                                   0
```

Antes de continuar el análisis se mantendrán únicamente las variables relevantes para este estudio, en base a los razonamientos realizados anteriormente.

```
covid_vac_redux <- covid_vac_clean[ , names(covid_vac_clean) %in% c("country",
                                          "iso_code","date",
                                          "daily_vaccinations_per_million",
                                          "vaccines")]
```

PARTE III: Preprocesado y análisis de los datos

4 Preprocesado de los datos

4.1 Creación de variables adicionales

Antes de proceder a seleccionar los grupos de trabajo es necesario preprocesar el dataset para obtener dos nuevas variables que serán útiles para el análisis.

Comenzando con la variable *vaccines*, que contiene una lista plana de vacunas.

```
head(covid_vac_redux[c(1,3,5)],5)
```

```
##       country       date                                                vaccines
## 1 Afghanistan 2021-02-22 Oxford/AstraZeneca, Pfizer/BioNTech, Sinopharm/Beijing
## 2 Afghanistan 2021-02-23 Oxford/AstraZeneca, Pfizer/BioNTech, Sinopharm/Beijing
## 3 Afghanistan 2021-02-24 Oxford/AstraZeneca, Pfizer/BioNTech, Sinopharm/Beijing
## 4 Afghanistan 2021-02-25 Oxford/AstraZeneca, Pfizer/BioNTech, Sinopharm/Beijing
## 5 Afghanistan 2021-02-26 Oxford/AstraZeneca, Pfizer/BioNTech, Sinopharm/Beijing
```

Para poder analizar el uso o no de cada vacuna por separado, se crearán variables dummy que indiquen si la vacuna se usa (1) o no (0) en las vacunaciones diarias del pais. Se separan en primer lugar los valores.

```
covid_vac_redux$vaccines <- str_replace_all(covid_vac_redux$vaccines, " ", "")
vaccine <- vector()
for (i in unique(covid_vac_redux$vaccines)){
  for (j in strsplit(i, ",")){
    vaccine <- c(vaccine, j)
  }
}
```

La lista de vacunas utilizadas se muestra por tanto a continuación.

```r
vaccine_used <- unique(vaccine)
# Vacunas utilizadas:
vaccine_used
```

```
##  [1] "Oxford/AstraZeneca" "Pfizer/BioNTech"    "Sinopharm/Beijing"
##  [4] "Sinovac"            "SputnikV"           "Johnson&Johnson"
##  [7] "Moderna"            "Covaxin"            "Sinopharm/Wuhan"
## [10] "Abdala"             "Soberana02"         "QazVac"
## [13] "Sinopharm/HayatVax" "CanSino"            "EpiVacCorona"
## [16] "RBD-Dimer"
```

Se crea una nueva tabla con variables dummy sobre el uso de la vacuna por paises.

```r
# Inspecionamos su uso:
vaccine_data_val <- data.frame(matrix(ncol = length(vaccine_used), nrow = 0))
for (i in covid_vac_redux$vaccines){
  vaccine_data_val<- rbind(vaccine_data_val, Vectorize(grepl, USE.NAMES = TRUE)
                           (vaccine_used, str_replace_all(i," ","")))
}
vaccine_data_val[vaccine_data_val == TRUE] = 1
vaccine_data_val[vaccine_data_val == FALSE] = 0
colnames(vaccine_data_val) <- paste0(unique(vaccine))
```

Se integran estas variables dummy con el dataset principal del análisis.

```r
covid_vac_redux_enc <- bind_cols(covid_vac_redux, vaccine_data_val)
covid_vac_redux_encoded <- subset(covid_vac_redux_enc, select=-c(vaccines))
```

Por otro lado es interesante también disponer del continente al que pertenece el país en cuestión. Para ello se hace uso de una función específica que lo calcula a partir del código ISO del mismo.

```r
# Representamos el total de vacunaciones en cada país de cada continente,
# añadiendo una nueva columna "continent":
covid_vac_redux_encoded <- covid_vac_redux_encoded %>%
  mutate("continent" = countrycode(sourcevar = covid_vac_redux_encoded[, "country"],
                                   origin = "country.name",destination = "continent",
                                   warn=FALSE))
```

Los países que no se han encontrado coincidencias exactas están contenidos en otros (por ejemplo, los datos de Reino Unido incluyen los de Gales, Escocia, Irlanda del norte, Islas feroe,etc) o no se encuentra el código ISO. Se eliminan estos países.

```r
remove_countries <- c('Northern Ireland','England','Wales','Scotland',
                      'Falkland Islands','Faeroe Islands',
                      'Isle of Man','Cayman Islands','Saint Helena',
                      'Saint Lucia','Saint Vincent and the Grenadines',
                      'Saint Kitts and Nevis','Timor','Kosovo')
covid_vac_redux_encoded <- covid_vac_redux_encoded %>%
  filter(!country %in% remove_countries)
```

Por último, se exporta el dataset limpio y enriquecido sobre el que se realizará los análisis posteriores de este documento.

```
write.csv(covid_vac_redux_encoded,"country_vaccinations_clean.csv",
          row.names = FALSE)
```

4.2 Análisis de los datos

4.2.1 Selección de grupos de estudio Con los datos ya procesados, es necesario obtener los subgrupos sobre los que se realizará el análisis. La primera división a realizar será a nivel de país. La comparación de este estudio se va a centrar principalmente en la comparación de España con otros países de Europa. Por ello, se obtendrán grupos con datos de vacunaciones diarias para España, Italia y Francia. También se generará un subconjunto para las vacunaciones de Estados Unidos, a modo de comparativa.

```
covid_vac_redux_encoded_spain <- covid_vac_redux_encoded[covid_vac_redux_encoded$
                                                         country == 'Spain',]
covid_vac_redux_encoded_Italy <- covid_vac_redux_encoded[covid_vac_redux_encoded$
                                                         country == 'Italy',]
covid_vac_redux_encoded_France <- covid_vac_redux_encoded[covid_vac_redux_encoded$
                                                         country == 'France',]
covid_vac_redux_encoded_USA <- covid_vac_redux_encoded[covid_vac_redux_encoded$
                                                    country == 'United States',]
```

Se realizará tambien una división de grupos por continentes para Europa y América.

```
covid_vac_redux_encoded_Europe <- covid_vac_redux_encoded[covid_vac_redux_encoded$continent == 'Europe'
```

```
covid_vac_redux_encoded_America <- covid_vac_redux_encoded[covid_vac_redux_encoded$continent == 'America
```

4.2.2 Comprobación de normalidad y homogeneidad de la varianza

Test de normalidad de los datos

Primero se realizará una comprobación de normalidad de los datos. Se realizará el test de normalidad sobre el conjunto de datos sin separación entre grupos, en concreto para la variable *daily_vaccinations_per_million*.

Para ello se aplicará el test de lilliefors, dado que el número de muestras es mayor a 50.

```
lillie.test(covid_vac_redux_encoded$daily_vaccinations_per_million)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  covid_vac_redux_encoded$daily_vaccinations_per_million
## D = 0.25408, p-value < 2.2e-16
```

El p-valor obtenido está muy por debajo del nivel de significancia (establecido por defecto en 0.05), por lo tanto se rechaza la hipótesis nula de normalidad en los datos respecto a la variable *daily_vaccinations_per_million*.

Test de homogeneidad de la varianza

Para poder comprobar si se asume homocedasticidad o heterocedasticidad, es necesario aplicar un contraste sobre la varianza.

Se aplica un test bilateral sobre la varianza con la siguiente formulación:

$H_0 : \sigma^2 = {\sigma^2}_0$

$H_1 : \sigma^2 \neq {\sigma^2}_0$

El nivel de significancia se ha establecidado en $\alpha = 0.05$

Se realizará una comparación entre los datos de vacunaciones diarias para Italia y para España, de cara a verificar el compartamiento de la varianza entre estos dos grupos. Se aplica la funcion var.test para realizar este test sobre las variables en análisis.

```
var.test(covid_vac_redux_encoded_spain$daily_vaccinations_per_million,
         covid_vac_redux_encoded_Italy$daily_vaccinations_per_million,
alternative = 'two.sided',conf.level = 0.95)
```

```
##
##  F test to compare two variances
##
## data:  covid_vac_redux_encoded_spain$daily_vaccinations_per_million and covid_vac_redux_encoded_Italy
## F = 0.99398, num df = 141, denom df = 149, p-value = 0.9723
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.7173178 1.3793982
## sample estimates:
## ratio of variances
##          0.9939773
```

Por tanto se rechaza la hipótesis nula y se puede considerar que la varianza de *daily_vaccinations_per_million* para España e Italia no es diferente y se asume homocedasticidad.

4.2.3 Uso de vacuna por países. Análisis de correlaciones

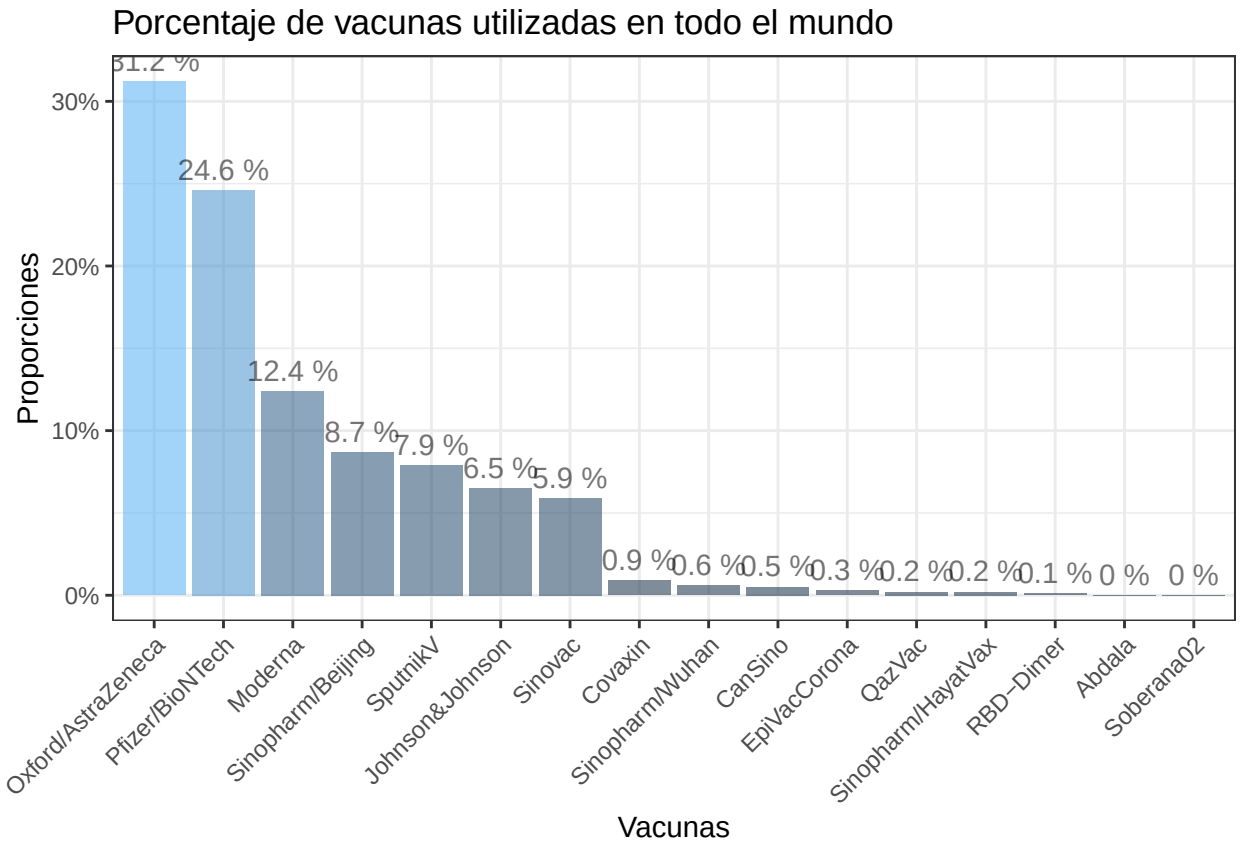
Distribución geográfica de uso de vacunas

El primer análisis planteado en este estudio describe la distribución de las distintas marcas de vacunas que se inoculan en los distintos paises del mundo. Asímismo, se describen qué tipos de vacunas se suelen administrar de forma combinada a nivel de país.

El primer paso es revisar que marca se ha utilizado de forma más frecuente, sumando todos los días y en todos los países considerados, durante la campaña de vacunaciones.

```
vaccine_data_val %>%
  summarise_all(sum) %>%
  gather(key="Vaccine_name", value="Vaccine_count") %>%
  mutate(Vaccine_count1=round(Vaccine_count/sum(Vaccine_count),3)) %>%
  ggplot(mapping=aes(x=reorder(Vaccine_name,-Vaccine_count1), y=Vaccine_count1,
                     fill =Vaccine_count1, alpha=0.7)) +
  geom_col() +
```

```
  labs(x = "Vacunas", y = "Proporciones",
       title  = "Porcentaje de vacunas utilizadas en todo el mundo") +
  geom_text(aes(label = paste(Vaccine_count1*100,"%")), vjust=-0.5) +
  scale_y_continuous(labels = scales::percent) +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        legend.position = "None")
```

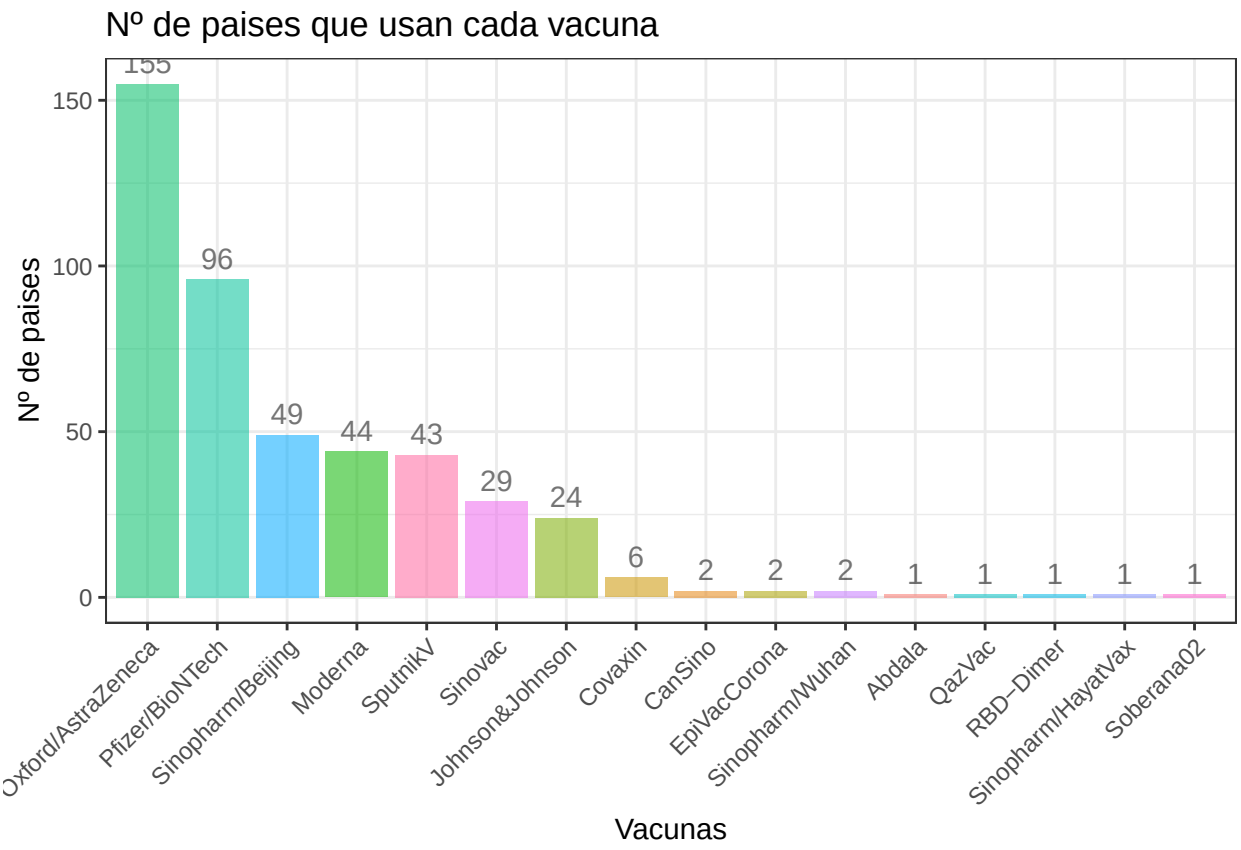


En un 31.2 % de los días, para todos los países considerados, Astrazeneca ha estado presente en la campaña de vacunaciones, seguido por Pfizer con un 24.6 % y Moderna con un 12.4 %. Llama la atención como ciertas vacunas tienen un porcentaje casi nulo de frecuencia de uso. Por entender este hecho, se mostrará el número de países que utilizan cada tipo de vacuna.

```
vaccine_in_countries<- covid_vac_redux_encoded[,c(1,5:20)] %>%
  group_by(country)%>%
  summarise_all(sum)
```

```
data <- data.frame("No_of_countries"= apply(vaccine_in_countries[-1],2,
                                            function(c)sum(c!=0)))
cbind("Vaccine"=row.names(data),data) %>%
  ggplot(mapping=aes(x=reorder(Vaccine, -No_of_countries), y=No_of_countries,
                     fill = Vaccine, alpha=0.5))+
  geom_col() +
  labs(x = "Vacunas", y = "Nº de paises",
       title  = "Nº de paises que usan cada vacuna")+
  geom_text(aes(label = No_of_countries), vjust=-0.5)+
```

```
  theme_bw()+
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        legend.position = "None")
```




```
for (i in (grep("Soberana02", covid_vac_redux$vaccine, ignore.case=TRUE))){
  country_Soberana02 <- covid_vac_redux$country[i]
}
```

Se observa como 157 países del total de 199 han utilizado la vacuna “Oxford/AstraZeneca”. También destaca la existencia de otras tantas vacunas que tan solo la han utilizado un país. como por ejemplo “Soberana02”, utilizada únicamente en Cuba.

```
country_Soberana02
```

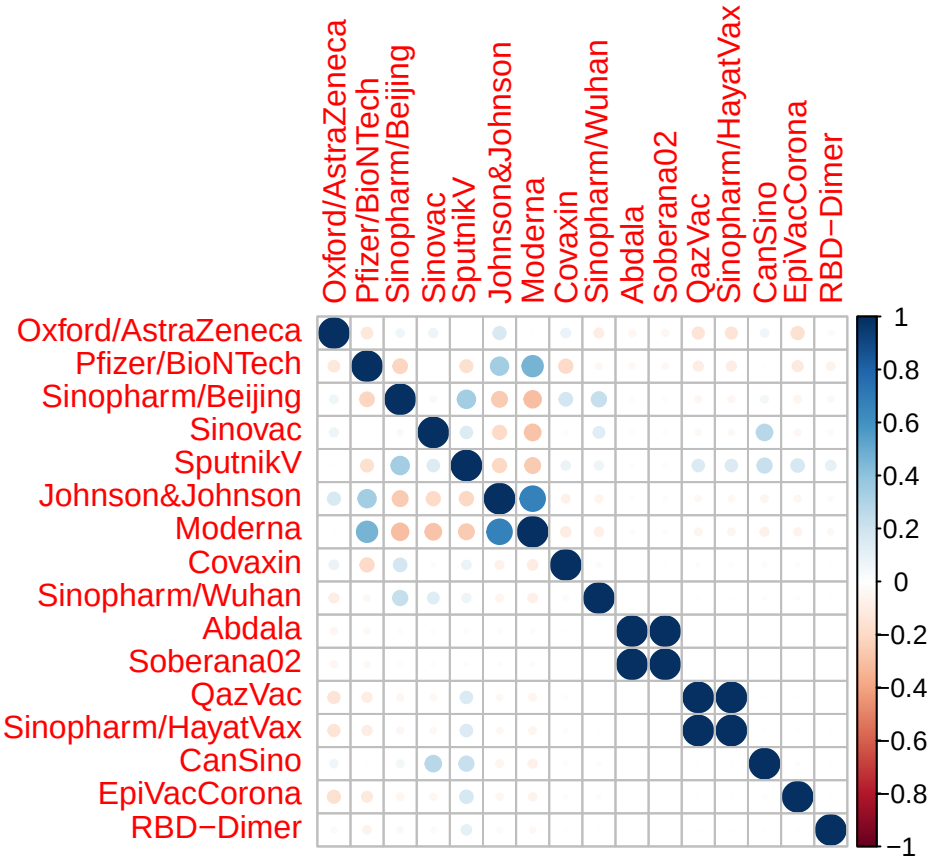
```
## [1] "Cuba"
```

Análisis de correlaciones

Una vez revisada la distribución geográfica de uso de vacunas de forma invidivual, se procederá a analizar la correlación que existe en el uso de vacunas. Es decir, **cuáles de ellas se suelen administrar de forma conjunta durante la campaña de vacunación de un país.**

Primero se analizará a nivel mundial.

```
cor_vac <- corrplot(cor(covid_vac_redux_encoded[,c(5:20)],method = 'pearson'))
```

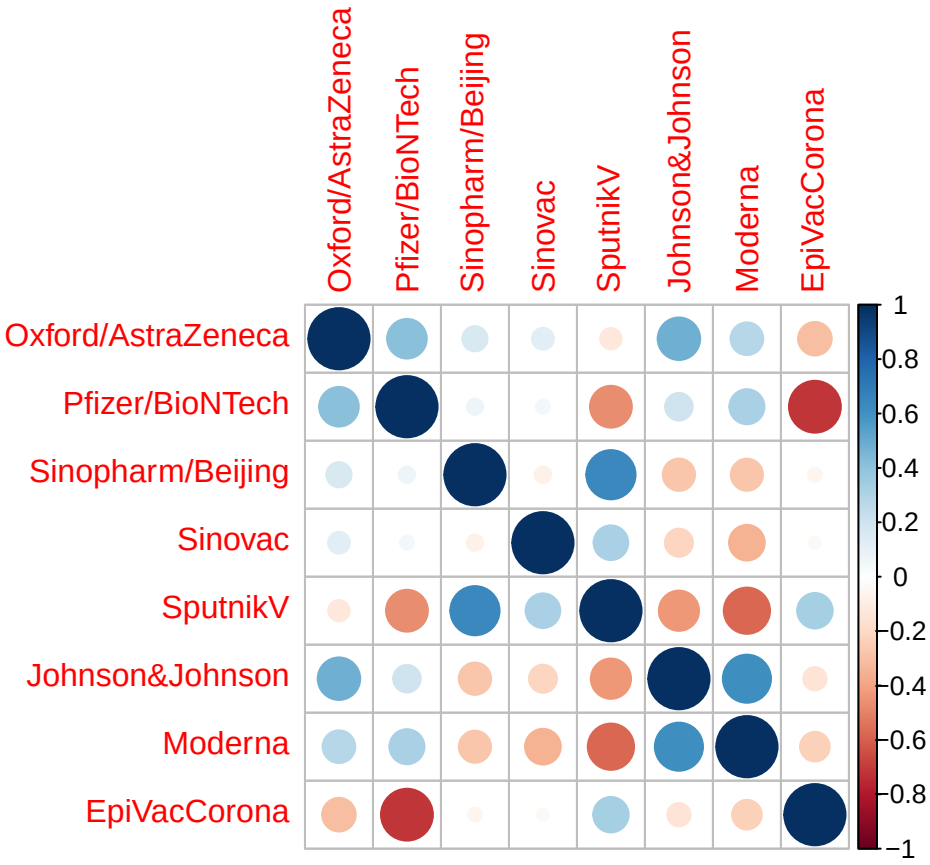


De los resultados anteriores, se puede extraer que las únicas vacunas que muestran correlación máxima se dan en casos minoritarios. Destaca la antes mencionada Soberana02, la cual se administra en Cuba conjuntamente con Abdala. Para el resto de marcas, destaca la combinación de Moderna y Johnson&Johnson, que tienen una correlación positiva significativa, por lo que es habitual verlas juntas. En la gran mayoría de países europeos aparece dicha combinación.

Dado que muchas vacunas tienen un uso más local o regional, se generará el gráfico anterior separado para América y para Europa. Se mostraran solo las vacunas que tengan uso en al menos un país del continente.

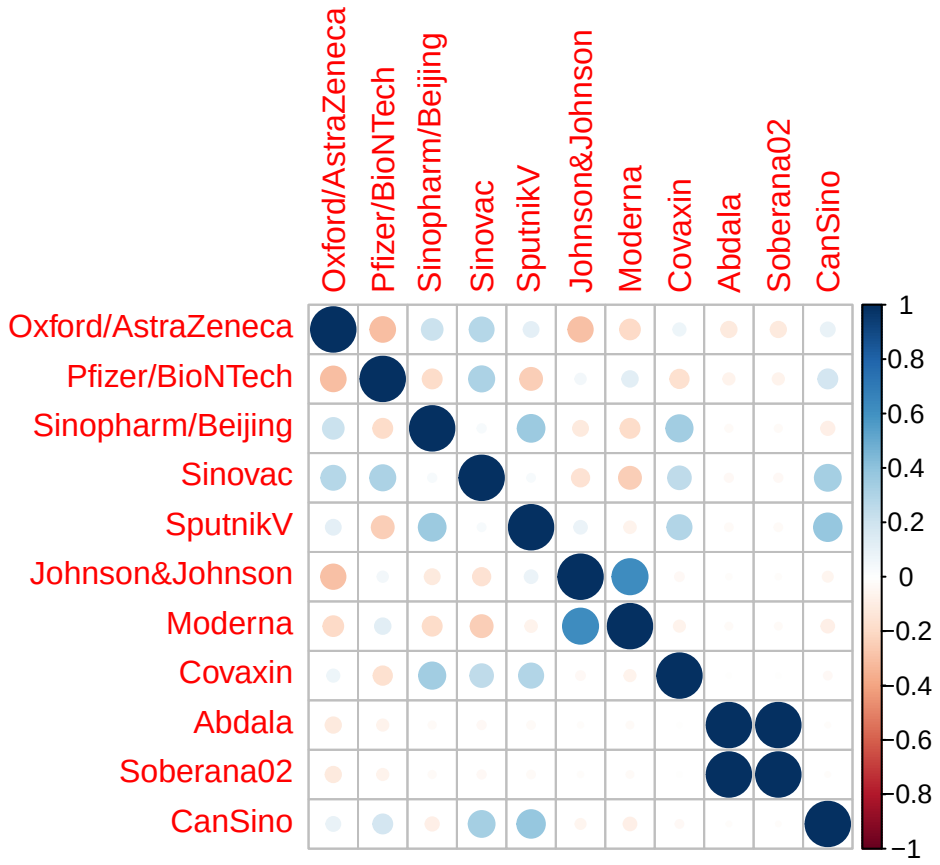
Por lo que, para **Europa**:

```
cor_vac <- corrplot(cor(covid_vac_redux_encoded_Europe[,c(5:11,19)],
                        method = 'pearson'))
```



Y para **América**:

```
cor_vac <- corrplot(cor(covid_vac_redux_encoded_America[,c(5:12,14,15,18)],
                        method = 'pearson'))
```



Comparando los gráficos anteriores, las tendencias generales observadas en el gráfico global se mantienen, aunque existen también matices significativos.

En primer lugar en Europa se están usando actualmente 8 marcas diferentes de vacunas, mientras que en América este número asciende a 11. Sin embargo, si se elimina Abdala y Soberana02, solo usadas en Cuba, el número total es casi idéntico.

Otro punto llamativo es la combinación Pfizer y Astrazeneca. Mientras que en Europa existe cierto grado de correlación positiva entre ambas (es decir, se suelen administrar juntas), en America ocurre lo contrario, ya que la correlación es ligeramente negativa, por lo que se autoexcluyen más que se combinan. Por otro, la sinergia entre Moderna y Johnson&Johnson se mantiene en ambos continentes, siendo significativa tanto en América como en Europa.

4.2.4 Contraste de hipótesis. Diferencia de medias entre grupos A continuación se compararán los valores medios diarios de vacunaciones por millón entre Europa e Italia, de esta manera se pretende comprobar si en los últimos meses el ritmo de vacunación ha sido mayor en Italia que en España. Se ha escogido Italia como país de comparación a España por ser similares en la evolución del COVID-19, volumen de contagios, tipos de vacunas administradas, etc.

Se plantea por tanto la siguiente pregunta de investigación: *¿Las vacunaciones diarias por millón (daily_vaccinations_per_million) son mayores en Italia que en España?*

El contraste a realizar en este caso es de **dos muestras sobre la diferencia de medias**, ya que se pretende comparar parámetros de dos poblaciones diferentes: las vacunaciones por millón en Italia y en España. Las muestras son **independientes** entre sí, es decir, las vacunaciones en un país y en otro no dependen la una de la otra.

Por el teorema del límite central, dado que ambas muestras son de tamaño grande (150 registros en la muestra de Italia y 142 en la muestra de España y por tanto n>30) y que se desea realizar un contraste sobre la

media, **se puede considerar normalidad en los datos**. Por tanto, asumiendo la hipótesis de normalidad en la distribución, el test es paramétrico ya que se establecen afirmaciones sobre parámetros (media) de la mencionada distribución.

El contraste se puede formular de la siguiente manera:

$H_0 : \mu_{\text{vac\_Italia}} = \mu_{\text{vac\_España}}$

$H_1 : \mu_{\text{vac\_Italia}} > \mu_{\text{vac\_España}}$

```
t.test(covid_vac_redux_encoded_Italy$daily_vaccinations_per_million,
       covid_vac_redux_encoded_spain$daily_vaccinations_per_million,
       var.equal=FALSE,
       alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  covid_vac_redux_encoded_Italy$daily_vaccinations_per_million and covid_vac_redux_encoded_spai
## t = -1.084, df = 289.22, p-value = 0.8604
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -811.9144       Inf
## sample estimates:
## mean of x mean of y
##  3359.573  3681.472
```

A la vista de los resultados, el p-valor es de 0.8604, por encima del nivel de significancia establecido en 0.05. En consecuencia la hipótesis nula se acepta y por tanto se puede afirmar con un 95 % de confianza que la media de vacunaciones por millón diaria en Italia y en España son iguales.

Esto tiene sentido dada la similitud entre los países y cómo se ha desarrollado la campaña de vacunación. Además, influye el hecho de que el reparto de las mismas esté gestionado por la Unión Europea, indicando un reparto relativamente equitativo con una gestión nacional similar del suministro.

4.2.5 Análisis de series temporales de vacunación A continuación,vamos a intentar predecir la progresión mundial del proceso de vacunación a lo largo del tiempo, es decir, predecir sus valores futuros. Y para ello, el primer requisito es que nuestras series temporales sean estacionarias, lo cual se suele conseguir aplicando logaritmos (para corregir heterocedasticidad), diferencia regular (eliminar tendencia), diferencia estacional (eliminar componente estacional), etc.

Como vamos a tratar con datos de series temporales, probablemente, a parte de agrupar las vacunaciones diarias (en ppm) por fecha, será útil añadir nuevas características de fecha al conjunto de datos modificado:

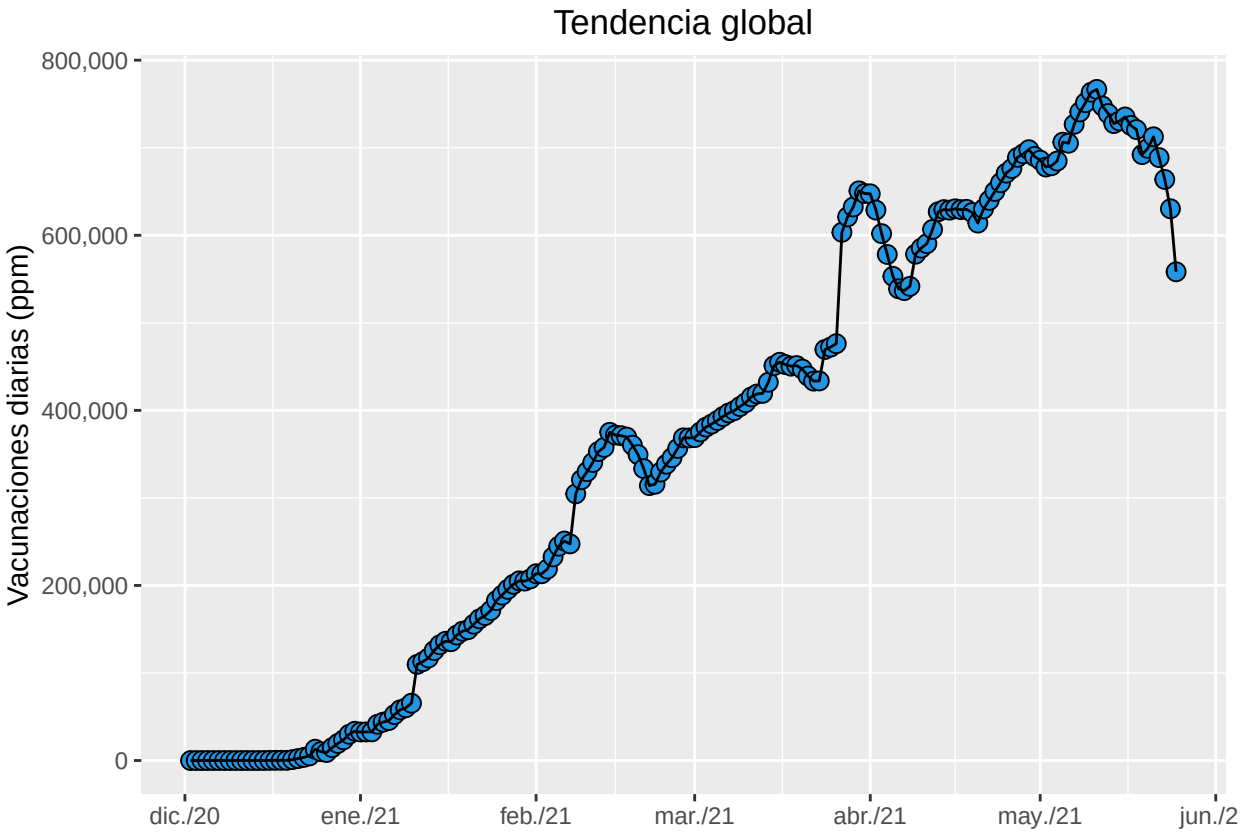
```
covid_vac_redux_TS <- covid_vac_redux[,3:4] %>%
  group_by(date) %>%
  summarise (daily_vaccinations_per_million=sum(daily_vaccinations_per_million))

# Convertimos nuestra variable date de tipo character a tipo date
covid_vac_redux_TS$date <- as.Date(covid_vac_redux_TS$date)

covid_vac_redux_TS['vac_test'] <- covid_vac_redux_TS['daily_vaccinations_per_million']

covid_vac_redux_TS['year'] <- as.numeric(format(covid_vac_redux_TS$date,'%Y'))
covid_vac_redux_TS['month'] <- as.numeric(format(covid_vac_redux_TS$date,'%m'))
```

```
covid_vac_redux_TS['day'] <- as.numeric(format(covid_vac_redux_TS$date,'%d'))
tail(covid_vac_redux_TS)
```

```
## # A tibble: 6 x 6
##   date       daily_vaccinations_per_million vac_test  year month   day
##   <date>                              <dbl>    <dbl> <dbl> <dbl> <dbl>
## 1 2021-05-20                         699279   699279  2021     5    20
## 2 2021-05-21                         712509   712509  2021     5    21
## 3 2021-05-22                         688740   688740  2021     5    22
## 4 2021-05-23                         663778   663778  2021     5    23
## 5 2021-05-24                         630270   630270  2021     5    24
## 6 2021-05-25                         558295   558295  2021     5    25
```

Ahora, mostramos la tendencia global de vacunación:

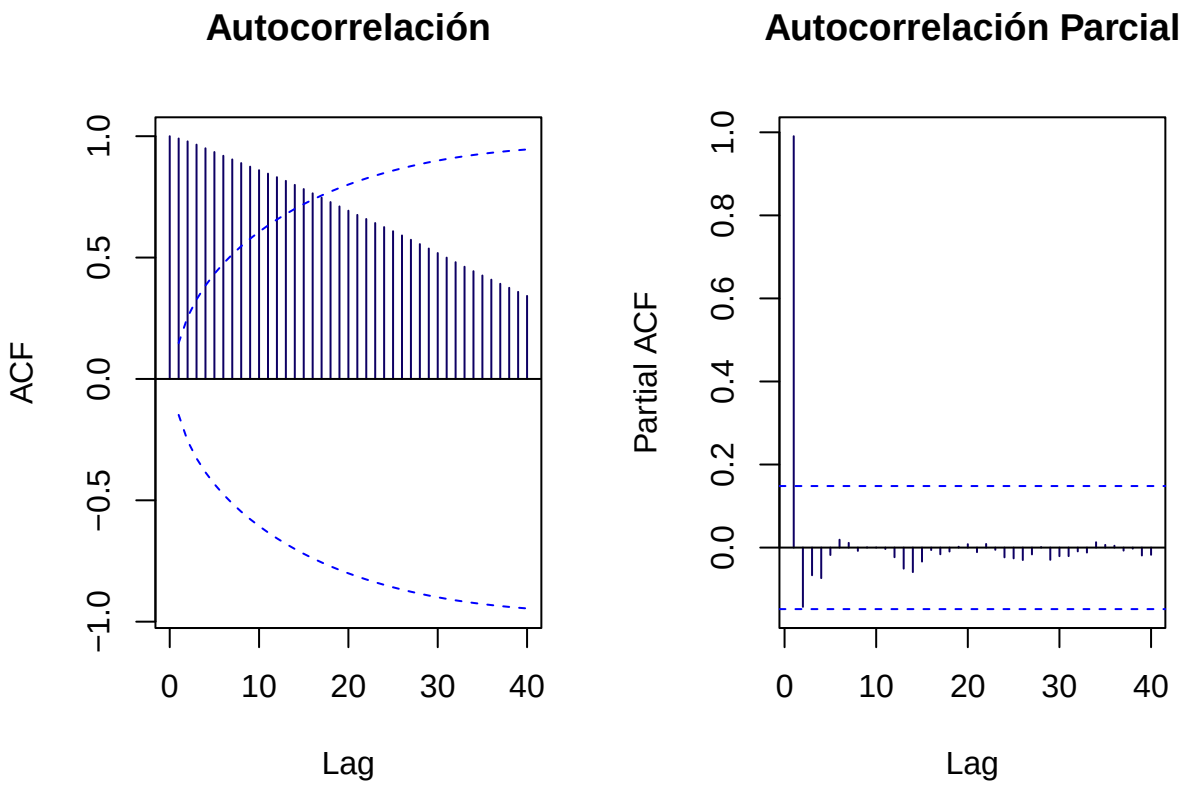
```
ggplot(covid_vac_redux_TS) +
  geom_point(aes(x=date, y=daily_vaccinations_per_million), size=3, pch=21,
             bg = 20, lwd = 1) +
  geom_line(aes(x=date, y=daily_vaccinations_per_million)) +
  scale_y_continuous(labels = scales::comma)+
  scale_x_date(date_breaks = "1 month", date_labels = "%b/%y")+
  labs(x=element_blank(),y ="Vacunaciones diarias (ppm)",
       title  = "Tendencia global", col=element_blank())+
  theme(plot.title = element_text(hjust = 0.5))
```



Y procedemos a realizar un análisis estacionario, donde las ACF proporcionan información sobre cómo

una observación influye en las siguientes y las PACF proporcionan la relación directa existente entre observaciones separadas por k retardos:

```
par(mfrow=c(1,2))

acf(covid_vac_redux_TS$daily_vaccinations_per_million, ci.type = "ma",
    lag.max=40, ci.col = "blue", main="Autocorrelación", col = "#0E0064")

pacf(covid_vac_redux_TS$daily_vaccinations_per_million, lag.max = 40,
     ci.col = "blue", main="Autocorrelación Parcial", col = "#0E0064")
```



Al analizar el ACF, podemos ver que los valores disminuyen lentamente hasta llegar a 0, siendo esta la primera señal de una serie no estacionaria en la media. Siguiendo los datos del gráfico de tendencia global, las series reflejan una tendencia creciente. Por lo que procedemos a comprobarlo con la prueba ADF (Dickey-Fuller test).

Dicha prueba se utiliza, como acabamos de comentar, para comprobar si el atributo *daily_vaccinations_per_million* representa una estacionaria de una serie temporal.

$H_0 = \phi = 0$

$H_1 \neq \phi = 0$

$$\phi = 0$$

significa que nuestra serie temporal es un proceso aleatorio, mientras que si

$$\phi \neq 0$$

$$(-1 < 1 + \phi < 1)$$

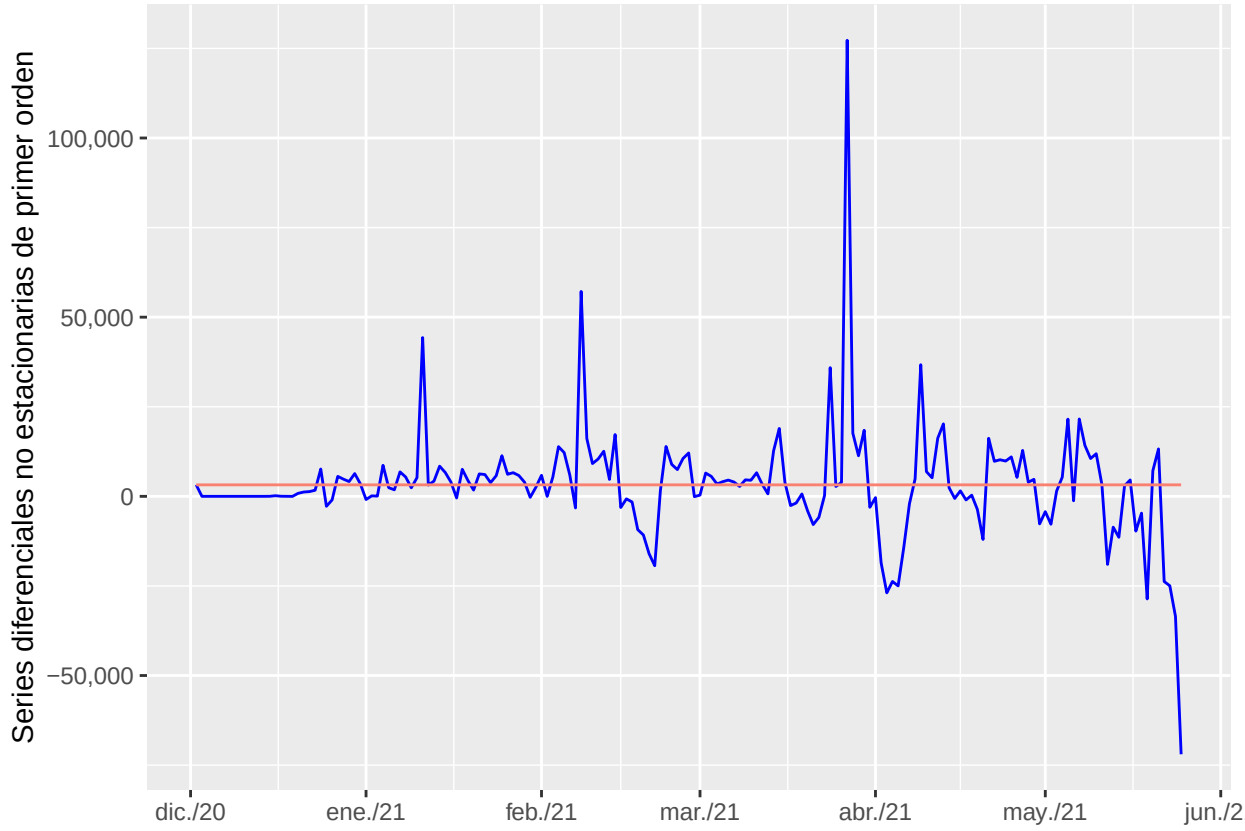
obtenemos un proceso estacionario. De este modo, necesitamos que el valor p sea inferior a 0.05 para proceder con ACF y PACF.

```
adf.test(covid_vac_redux_TS$daily_vaccinations_per_million)
```

```
## Augmented Dickey-Fuller Test
## alternative: stationary
##
## Type 1: no drift no trend
##      lag     ADF p.value
## [1,]   0  1.4297   0.960
## [2,]   1  0.5649   0.805
## [3,]   2  0.2580   0.717
## [4,]   3 -0.2776   0.564
## [5,]   4  0.0312   0.652
## Type 2: with drift no trend
##      lag   ADF p.value
## [1,]   0 -1.28   0.597
## [2,]   1 -1.31   0.588
## [3,]   2 -1.37   0.567
## [4,]   3 -1.53   0.512
## [5,]   4 -1.43   0.546
## Type 3: with drift and trend
##      lag    ADF p.value
## [1,]   0  0.564   0.990
## [2,]   1 -0.633   0.975
## [3,]   2 -1.048   0.928
## [4,]   3 -1.891   0.619
## [5,]   4 -1.381   0.834
## ----
## Note: in fact, p.value = 0.01 means p.value <= 0.01
```

A la vista del resultado, en ningún caso el p-valor es inferior a 0.05, por lo que procedemos a realizar diferentes transformaciones de nuestra serie temporal para convertirla en estacionaria, lo que nos permitirá construir un modelo ARIMA.

```
diff_daily = diff(covid_vac_redux_TS$daily_vaccinations_per_million)
diff_daily <- c(NaN,diff_daily)
covid_vac_redux_TS$diff1 <- diff_daily
covid_vac_redux_TS$diff1[is.na(covid_vac_redux_TS$diff1)] <- mean(covid_vac_redux_TS$diff1,
                                                                  na.rm = TRUE)

ggplot(covid_vac_redux_TS) +
  geom_line(aes(x=date, y=diff1), col="blue") +
  geom_line(aes(x=date, y=mean(diff1)), col="salmon")+
  scale_y_continuous(labels = scales::comma)+
  scale_x_date(date_breaks = "1 month", date_labels = "%b/%y")+
  labs(x=element_blank(),y ="Series diferenciales no estacionarias de primer orden",
       col=element_blank())
```



Para comprobar la estacionariedad de la serie, realizamos de nuevo la prueba de Dickey-Fuller:

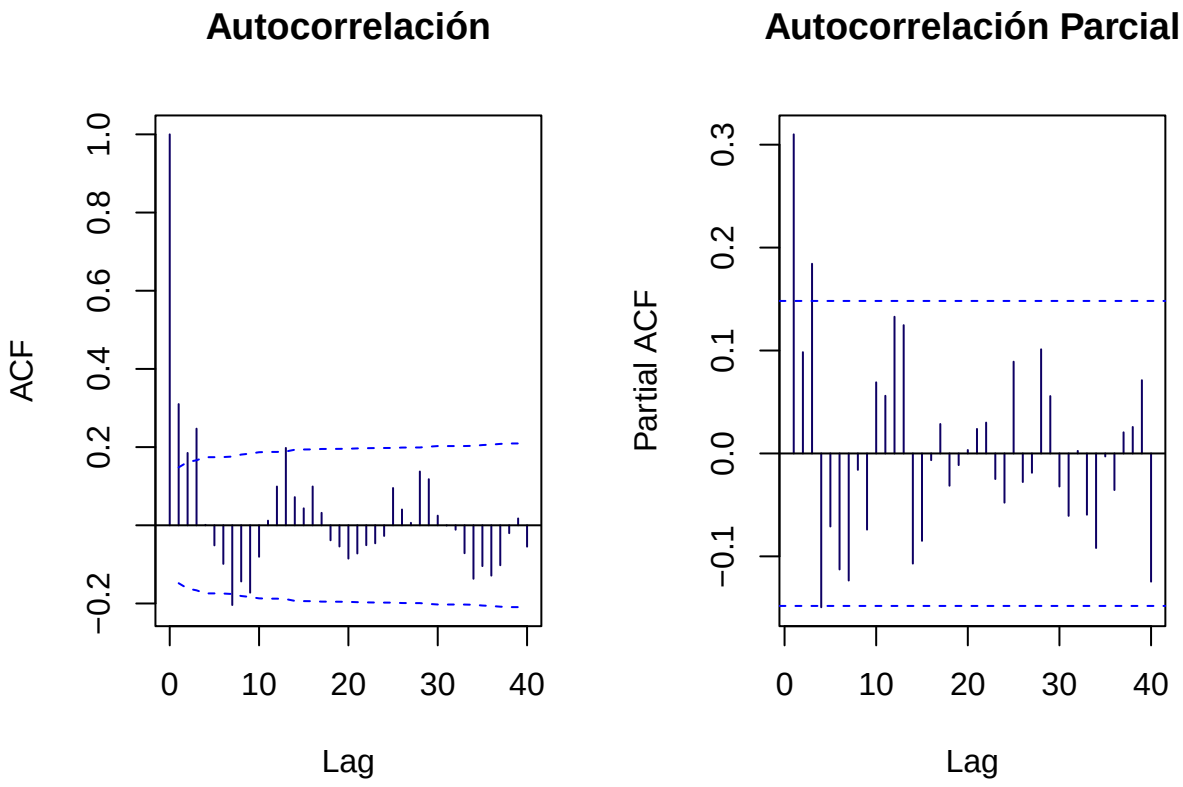
```
adf.test(covid_vac_redux_TS$diff1)
```

```
## Augmented Dickey-Fuller Test
## alternative: stationary
##
## Type 1: no drift no trend
##      lag   ADF p.value
## [1,]   0 -8.13    0.01
## [2,]   1 -5.53    0.01
## [3,]   2 -3.54    0.01
## [4,]   3 -3.94    0.01
## [5,]   4 -4.03    0.01
## Type 2: with drift no trend
##      lag   ADF p.value
## [1,]   0 -8.33    0.01
## [2,]   1 -5.68    0.01
## [3,]   2 -3.58    0.01
## [4,]   3 -4.06    0.01
## [5,]   4 -4.21    0.01
## Type 3: with drift and trend
##      lag   ADF p.value
## [1,]   0 -8.39  0.0100
## [2,]   1 -5.75  0.0100
## [3,]   2 -3.65  0.0304
```

```
## [4,]   3 -4.12  0.0100
## [5,]   4 -4.27  0.0100
## ----
## Note: in fact, p.value = 0.01 means p.value <= 0.01
```

Ahora, el p-valor obtenido en la prueba ADF es muy inferior a 0.05. Sin embargo, podemos sospechar que la tendencia y las fluctuaciones de la varianza aumentan, por lo que procedemos a la diferenciación no estacional de segundo orden.

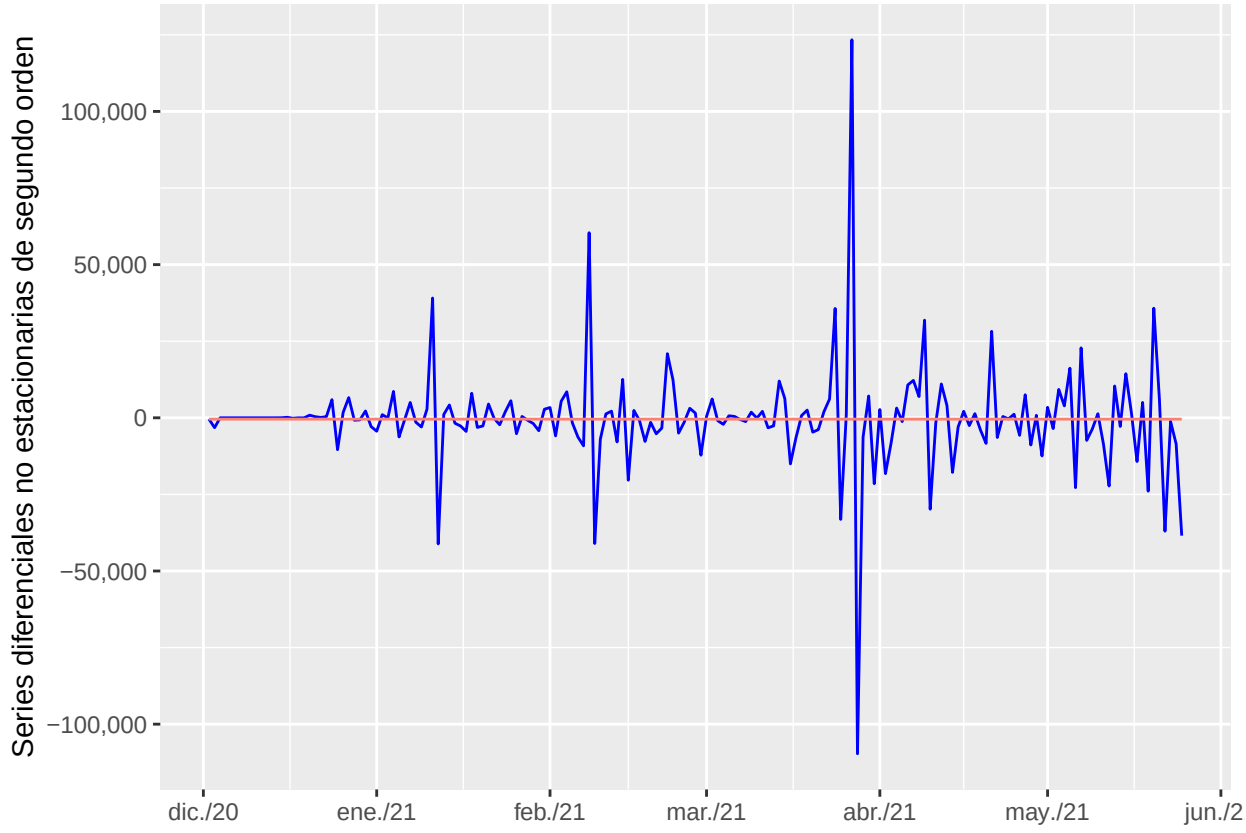
```
par(mfrow=c(1,2))

acf(covid_vac_redux_TS$diff1, ci.type = "ma", lag.max=40, ci.col = "blue",
    main="Autocorrelación", col = "#0E0064")

pacf(covid_vac_redux_TS$diff1, lag.max = 40, ci.col = "blue",
     main="Autocorrelación Parcial", col = "#0E0064")
```



Después de la diferenciación no estacionaria de primer orden, el ACF parece seguir disminuyendo lentamente hasta llegar a 0, aunque la diferenciación nos ayudó a solucionarlo en cierta medida. Intentamos hacer la diferenciación de segundo orden para ver si obtenemos mejores resultados:

```
diff_daily2 = diff(covid_vac_redux_TS$diff1)
diff_daily2 <- c(NaN,diff_daily2)
covid_vac_redux_TS$diff2 <- diff_daily2
covid_vac_redux_TS$diff2[is.na(covid_vac_redux_TS$diff2)] <- mean(covid_vac_redux_TS$diff2,
                                                                  na.rm = TRUE)
```

```
ggplot(covid_vac_redux_TS) +
  geom_line(aes(x=date, y=diff2), col="blue") +
  geom_line(aes(x=date, y=mean(diff2)), col="salmon")+
  scale_y_continuous(labels = scales::comma)+
  scale_x_date(date_breaks = "1 month", date_labels = "%b/%y")+
  labs(x=element_blank(),
       y ="Series diferenciales no estacionarias de segundo orden", col=element_blank())
```



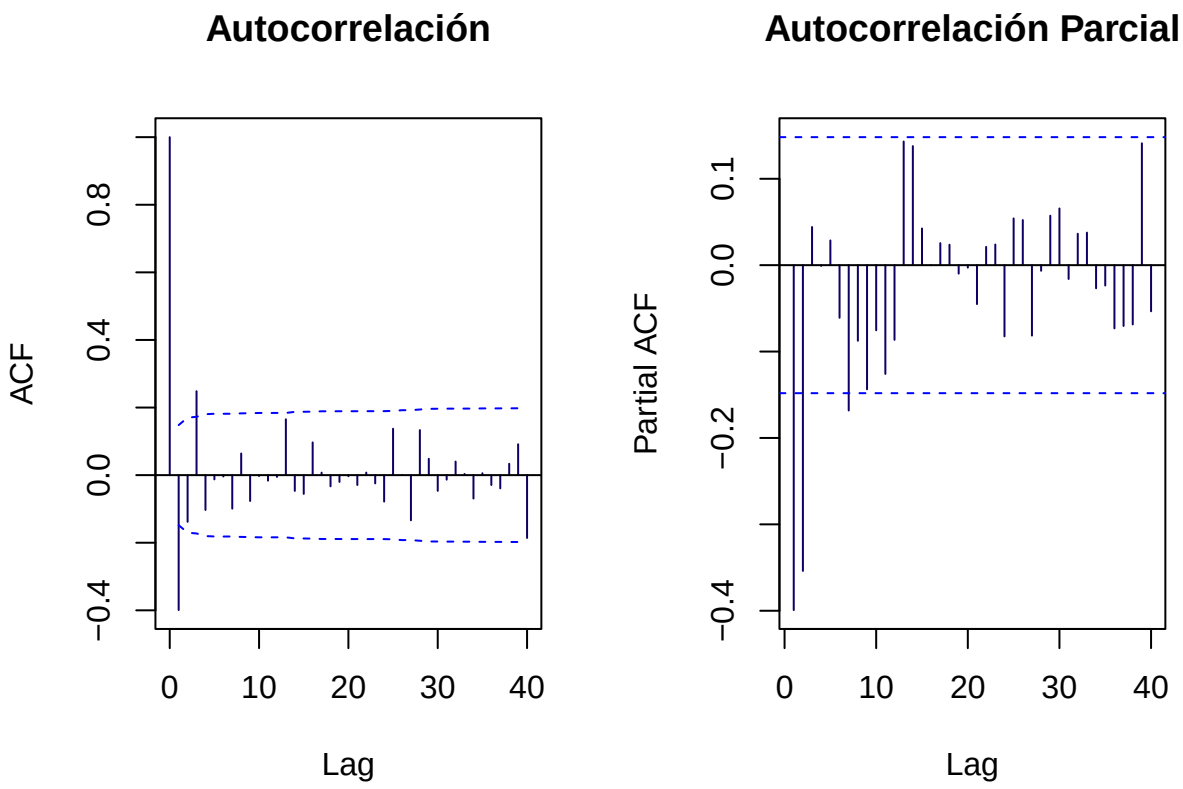

```
adf.test(covid_vac_redux_TS$diff2)
```

```
## Augmented Dickey-Fuller Test
## alternative: stationary
##
## Type 1: no drift no trend
##      lag    ADF p.value
## [1,]   0 -19.99    0.01
## [2,]   1 -15.95    0.01
## [3,]   2  -9.62    0.01
## [4,]   3  -7.64    0.01
## [5,]   4  -6.22    0.01
## Type 2: with drift no trend
##      lag    ADF p.value
## [1,]   0 -19.95    0.01
## [2,]   1 -15.93    0.01
```

```
## [3,]   2  -9.61    0.01
## [4,]   3  -7.63    0.01
## [5,]   4  -6.21    0.01
## Type 3: with drift and trend
##      lag    ADF p.value
## [1,]   0 -19.94    0.01
## [2,]   1 -15.95    0.01
## [3,]   2  -9.65    0.01
## [4,]   3  -7.68    0.01
## [5,]   4  -6.26    0.01
## ----
## Note: in fact, p.value = 0.01 means p.value <= 0.01
```

Este es sin duda el resultado esperado. No se detecta ninguna tendencia evidente en el gráfico, la media, representada como línea roja, es constante. Aplicando el test ADF sobre diff2 obtenemos como respuesta un p-valor muy pequeño (menor que 0.05), que nos indica que la serie es estacionaria.

```
par(mfrow=c(1,2))

acf(covid_vac_redux_TS$diff2, ci.type = "ma", lag.max=40, ci.col = "blue",
    main="Autocorrelación", col = "#0E0064")

pacf(covid_vac_redux_TS$diff2, lag.max = 40, ci.col = "blue",
     main="Autocorrelación Parcial", col = "#0E0064")
```



Una vez esto, ajustamos el modelo:

```
auto.arima(covid_vac_redux_TS$daily_vaccinations_per_million, seasonal=FALSE,
           trace=TRUE)
```

```
##
##  Fitting models using approximations to speed things up...
##
##  ARIMA(2,1,2) with drift         : 3824.473
##  ARIMA(0,1,0) with drift         : 3840.912
##  ARIMA(1,1,0) with drift         : 3823.516
##  ARIMA(0,1,1) with drift         : 3825.83
##  ARIMA(0,1,0)                    : 3845.975
##  ARIMA(2,1,0) with drift         : 3824.152
##  ARIMA(1,1,1) with drift         : 3821.277
##  ARIMA(2,1,1) with drift         : 3824.257
##  ARIMA(1,1,2) with drift         : 3822.913
##  ARIMA(0,1,2) with drift         : 3827.471
##  ARIMA(1,1,1)                    : 3820.305
##  ARIMA(0,1,1)                    : 3827.863
##  ARIMA(1,1,0)                    : 3824.177
##  ARIMA(2,1,1)                    : 3823.239
##  ARIMA(1,1,2)                    : 3821.839
##  ARIMA(0,1,2)                    : 3829.138
##  ARIMA(2,1,0)                    : 3823.839
##  ARIMA(2,1,2)                    : 3823.462
##
##  Now re-fitting the best model(s) without approximations...
##
##  ARIMA(1,1,1)                    : 3838.928
##
##  Best model: ARIMA(1,1,1)
```

```
## Series: covid_vac_redux_TS$daily_vaccinations_per_million
## ARIMA(1,1,1)
##
## Coefficients:
##          ar1      ma1
##       0.7451  -0.4174
## s.e.  0.1003   0.1221
##
## sigma^2 estimated as 217937355:  log likelihood=-1916.39
## AIC=3838.79   AICc=3838.93   BIC=3848.26
```

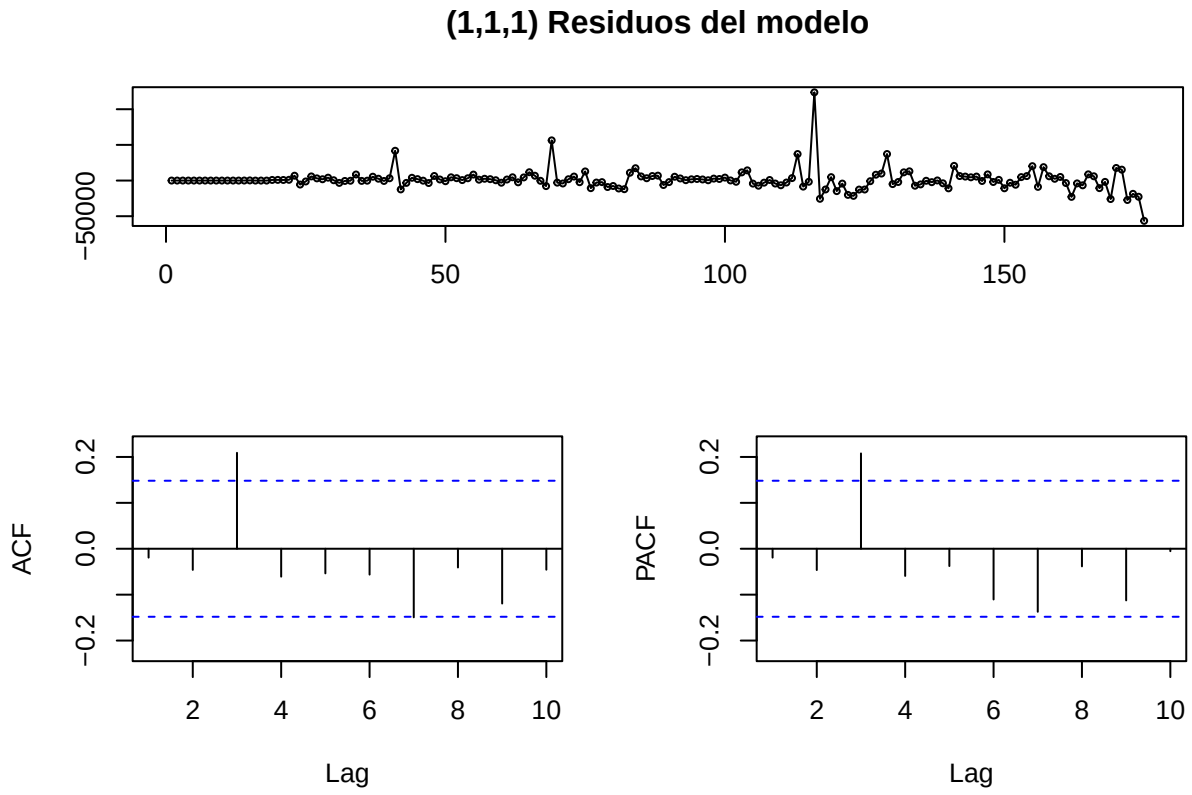
La función auto.arima realiza la función ARIMA de forma automática, va iterando y buscando el modelo que más se ajusta a los datos, dándonos una respuesta del mejor modelo seleccionado. El resultado después de sucesivas iteraciones se queda con el modelo que tuvo el menor valor de AIC, es decir, el modelo (1,1,1).

Usando la notación ARIMA, el modelo ajustado se puede escribir como:

$$\hat{Y}_{d_t} = 0{,}7451 Y_{t-1} - 0{,}4174 e_{t-1} + E$$

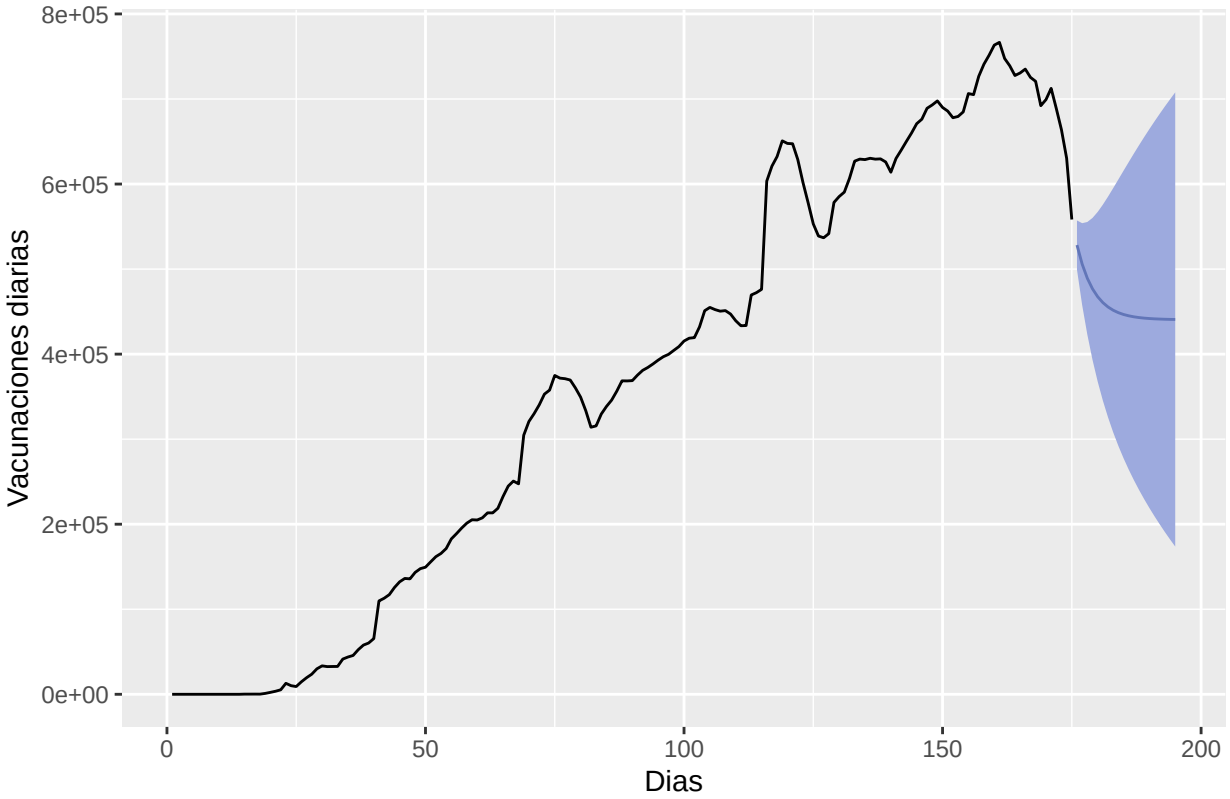
, donde E es un error y la serie original se diferencia con la orden 1.

```
modeloarima<-auto.arima(covid_vac_redux_TS$daily_vaccinations_per_million,
                        seasonal=FALSE)
tsdisplay(residuals(modeloarima), lag.max=10, main='(1,1,1) Residuos del modelo')
```



Se observa en los residuos que los valores perdidos están dentro del límite de significancia (gráfico ACF). Procedemos a realizar una estimación de este modelo, haciendo un forecast en el que vamos a predecir las vacunaciones diarias para el próximo mes con un nivel de significancia de un 95 %.

```
prediccion <- forecast(modeloarima,h=20,level=95)
forecast::autoplot(prediccion,xlab="Dias",ylab="Vacunaciones diarias",main="",
                   fcol="#596dd5")
```



Hemos definido en 20 (h=20), ya que el hecho de añadir más pasos (días) incrementa aún más la incertidumbre. Según las predicciones realizadas por el modelo ARIMA (1,1,1), se produce una disminución del número de vacunaciones. No esperamos que este valor caiga de una manera tan pronunciada como indica dicha predicción, pero es cierto que incluso los valores reales muestran una reducción en los últimos días. Esto podría deberse al proceso de fabricación de vacunas o a otros factores externos, como por ejemplo la estabilización del número de vacunas diarias. No obstante, llegamos a la conclusión de que la información actual no es suficiente para explicar estos cambios en el proceso de vacunación diario, sin la adición de externalidades que permitan mejorar el comportamiento del modelo.

PARTE IV: Conclusiones

A continuación se plantea un resumen de las conclusiones extraídas a lo largo del documento:

- En un **31.2 % de los días**, para todos los países considerados en el estudio, **Astrazeneca ha estado presente en la campaña de vacunaciones**, seguido por Pfizer con un 24.6 % y Moderna con un 12.4 %.
- **En Europa se están usando actualmente 8 marcas diferentes de vacunas**, mientras que **en América este número asciende a 11**. Aunque dos de ellas se administran solo en Cuba, por lo que la configuración es similar en ambos continentes.
- **La combinación Pfizer y Astrazeneca suele ser frecuente en las campañas de vacunación europeas** (0.42). En America ocurre lo contrario, ya que la correlación es ligeramente negativa (-0.31), por lo que se autoexcluyen más que se combinan.
- **La sinergia entre Moderna y Johnson&Johnson se mantiene en ambos continentes**, siendo significativa tanto en América como en Europa (correlación de en torno a 0.62 en ambos continentes)

- Se puede afirmar con un 95 % de confianza que **la media de vacunaciones por millón diaria en Italia y en España es igual**. Esto tiene sentido dada la similitud entre los países y cómo se ha desarrollado la campaña de vacunación. Además, influye el hecho de que el reparto de las mismas esté gestionado por la Unión Europea, indicando un reparto relativamente equitativo con una gestión nacional similar del suministro.
- Con respecto al análisis de series temporales, **con los métodos utilizados se nos permite obtener una idea mundial de la posible evolución de la curva de vacunaciones diarias y tomar medidas preventivas frente a ello**. O incluso a nivel nacional se podría comparar los resultados con otros países y deliberar que línea de vacunación es más efectiva.A pesar de esto, contando únicamente con la información de la serie histórica, no es suficiente para alcanzar una predicción suficientemente precisa sobre el compartamiento futuro.

Tabla de contribuciones

Contribuciones	Firma
Investigación previa	Daniel Lugo Laguna, Pablo Mora Galindo
Redacción de las respuestas	Daniel Lugo Laguna, Pablo Mora Galindo
Desarrollo código	Daniel Lugo Laguna, Pablo Mora Galindo