PRAC1 - Visualización de datos

Autor: Manuel Fernández

Diciembre 2022

Contents

Cargamos las librerías que emplearemos durante el ejercicio.

Cargamos el conjunto de datos COVID-19 y vemos la dimesión del mismo

```
df_covid <- read.csv("./data/covid.csv",sep=",", na.strings="",stringsAsFactors=FALSE)
dim(df_covid)
```

```
## [1] 241034     67
```

Miramos estádisticos básicos para las variables del conjunto de datos

```
summary(df_covid)
```

```
##    iso_code          continent           location            date            total_cases
##  Length:241034      Length:241034      Length:241034      Length:241034      Min.   :        1
##  Class :character   Class :character   Class :character   Class :character   1st Qu.:     4957
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :    55460
##                                                                             Mean   :  4772239
##                                                                             3rd Qu.:   568860
##                                                                             Max.   :646926650
##                                                                             NA's   :13949
##    new_cases       new_cases_smoothed  total_deaths       new_deaths    new_deaths_smoothed
##  Min.   :      0   Min.   :      0    Min.   :      1   Min.   :    0   Min.   :    0.00
##  1st Qu.:      0   1st Qu.:      5    1st Qu.:    117   1st Qu.:    0   1st Qu.:    0.00
##  Median :     47   Median :     88    Median :   1257   Median :    1   Median :    1.29
##  Mean   :  12172   Mean   :  12202    Mean   :  76325   Mean   :  133   Mean   :  133.63
##  3rd Qu.:    911   3rd Qu.:   1082    3rd Qu.:  10343   3rd Qu.:   13   3rd Qu.:   14.86
##  Max.   :4083952   Max.   :3439392    Max.   :6646094   Max.   :17733   Max.   :14859.29
##  NA's   :14234     NA's   :15433      NA's   :33326     NA's   :33410   NA's   :34594
##  total_cases_per_million new_cases_per_million new_cases_smoothed_per_million
##  Min.   :     0          Min.   :     0.00     Min.   :    0.00
##  1st Qu.:  1175          1st Qu.:     0.00     1st Qu.:    1.17
##  Median : 11434          Median :     6.37     Median :   17.94
##  Mean   : 65303          Mean   :   186.08     Mean   :  186.28
##  3rd Qu.: 78112          3rd Qu.:    92.08     3rd Qu.:  131.08
##  Max.   :690004          Max.   :228872.02     Max.   :36421.83
##  NA's   :14984           NA's   :15269         NA's   :16463
##  total_deaths_per_million new_deaths_per_million new_deaths_smoothed_per_million reproduction_rate
##  Min.   :   0.00          Min.   :  0.00         Min.   :  0.00                  Min.   :-0.06
##  1st Qu.:  31.75          1st Qu.:  0.00         1st Qu.:  0.00                  1st Qu.: 0.73
##  Median : 226.60          Median :  0.01         Median :  0.18                  Median : 0.96
##  Mean   : 713.63          Mean   :  1.34         Mean   :  1.34                  Mean   : 0.92
```

```
##  3rd Qu.:1058.77          3rd Qu.:  0.89         3rd Qu.:  1.25                  3rd Qu.: 1.14
##  Max.   :6389.68          Max.   :554.20         Max.   :148.64                  Max.   : 5.69
##  NA's   :34348            NA's   :34432          NA's   :35611                   NA's   :62457
##   icu_patients     icu_patients_per_million hosp_patients    hosp_patients_per_million
##  Min.   :    0.0   Min.   :  0.00           Min.   :     0   Min.   :   0.00
##  1st Qu.:   27.0   1st Qu.:  3.39           1st Qu.:   193   1st Qu.:  35.73
##  Median :  121.0   Median :  8.99           Median :   763   Median :  91.47
##  Mean   :  747.4   Mean   : 18.38           Mean   :  3964   Mean   : 148.22
##  3rd Qu.:  508.0   3rd Qu.: 23.77           3rd Qu.:  2973   3rd Qu.: 195.59
##  Max.   :28891.0   Max.   :180.68           Max.   :154497   Max.   :1526.85
##  NA's   :208350    NA's   :208350           NA's   :205360   NA's   :205360
##  weekly_icu_admissions weekly_icu_admissions_per_million weekly_hosp_admissions
##  Min.   :   0.0        Min.   :  0.00                   Min.   :     0.0
##  1st Qu.:  37.0        1st Qu.:  3.15                   1st Qu.:   287.8
##  Median : 173.0        Median :  7.25                   Median :  1013.0
##  Mean   : 406.6        Mean   : 12.58                   Mean   :  4712.4
##  3rd Qu.: 545.0        3rd Qu.: 16.51                   3rd Qu.:  4463.8
##  Max.   :4838.0        Max.   :222.90                   Max.   :153977.0
##  NA's   :233080        NA's   :233080                   NA's   :221790
##  weekly_hosp_admissions_per_million  total_tests          new_tests        total_tests_per_thousand
##  Min.   :  0.00                     Min.   :0.000e+00   Min.   :       1   Min.   :    0.00
##  1st Qu.: 30.86                     1st Qu.:3.647e+05   1st Qu.:    2244   1st Qu.:   43.59
##  Median : 71.35                     Median :2.067e+06   Median :    8783   Median :  234.14
##  Mean   : 96.78                     Mean   :2.110e+07   Mean   :   67285   Mean   :  924.25
##  3rd Qu.:131.41                     3rd Qu.:1.025e+07   3rd Qu.:   37229   3rd Qu.:  894.37
##  Max.   :701.93                     Max.   :9.214e+09   Max.   :35855632   Max.   :32925.83
##  NA's   :221790                     NA's   :161647      NA's   :165631     NA's   :161647
##  new_tests_per_thousand new_tests_smoothed new_tests_smoothed_per_thousand positive_rate
##  Min.   :  0.00         Min.   :       0   Min.   :  0.00                  Min.   :0.00
##  1st Qu.:  0.29         1st Qu.:    1486   1st Qu.:  0.20                  1st Qu.:0.02
##  Median :  0.97         Median :    6570   Median :  0.85                  Median :0.06
##  Mean   :  3.27         Mean   :  142178   Mean   :  2.83                  Mean   :0.10
##  3rd Qu.:  2.91         3rd Qu.:   32205   3rd Qu.:  2.58                  3rd Qu.:0.14
##  Max.   :531.06         Max.   :14769984   Max.   :147.60                  Max.   :1.00
##  NA's   :165631         NA's   :137069     NA's   :137069                  NA's   :145107
##  tests_per_case      tests_units        total_vaccinations  people_vaccinated
##  Min.   :      1.0   Length:241034      Min.   :0.000e+00   Min.   :0.000e+00
##  1st Qu.:      7.1   Class :character   1st Qu.:1.296e+06   1st Qu.:7.372e+05
##  Median :     17.5   Mode  :character   Median :9.444e+06   Median :4.756e+06
##  Mean   :   2403.6                      Mean   :3.036e+08   Mean   :1.379e+08
##  3rd Qu.:     54.6                      3rd Qu.:6.246e+07   3rd Qu.:2.938e+07
##  Max.   :1023631.9                      Max.   :1.303e+10   Max.   :5.468e+09
##  NA's   :146686                         NA's   :172497      NA's   :175505
##  people_fully_vaccinated total_boosters      new_vaccinations   new_vaccinations_smoothed
##  Min.   :1.000e+00       Min.   :1.000e+00   Min.   :       0   Min.   :       0
##  1st Qu.:6.131e+05       1st Qu.:1.473e+05   1st Qu.:    3959   1st Qu.:     547
##  Median :4.162e+06       Median :2.902e+06   Median :   30440   Median :    5852
##  Mean   :1.208e+08       Mean   :6.659e+07   Mean   :  892332   Mean   :  370976
##  3rd Qu.:2.719e+07       3rd Qu.:1.860e+07   3rd Qu.:  230946   3rd Qu.:   44057
##  Max.   :5.032e+09       Max.   :2.638e+09   Max.   :49677470   Max.   :43690352
##  NA's   :178158          NA's   :203202      NA's   :184218     NA's   :97868
##  total_vaccinations_per_hundred people_vaccinated_per_hundred people_fully_vaccinated_per_hundred
##  Min.   :  0.0                  Min.   :  0.00                Min.   :  0.00
##  1st Qu.: 28.9                  1st Qu.: 19.66                1st Qu.: 13.12
```

```
##  Median :103.0                  Median : 56.29                Median : 47.91
##  Mean   :106.3                  Mean   : 48.51                Mean   : 43.44
##  3rd Qu.:171.0                  3rd Qu.: 75.17                3rd Qu.: 69.88
##  Max.   :379.7                  Max.   :128.76                Max.   :126.76
##  NA's   :172497                 NA's   :175505                NA's   :178158
##  total_boosters_per_hundred new_vaccinations_smoothed_per_million new_people_vaccinated_smoothed
##  Min.   :  0.00             Min.   :     0                        Min.   :       0
##  1st Qu.:  1.60             1st Qu.:   279                        1st Qu.:     115
##  Median : 22.05             Median :  1133                        Median :    1574
##  Mean   : 28.14             Mean   :  2396                        Mean   :  137919
##  3rd Qu.: 50.87             3rd Qu.:  3261                        3rd Qu.:   14909
##  Max.   :139.71             Max.   :117113                        Max.   :21072182
##  NA's   :203202             NA's   :97868                         NA's   :97974
##  new_people_vaccinated_smoothed_per_hundred stringency_index population_density    median_age
##  Min.   : 0.00                              Min.   :  0.00   Min.   :    0.137   Min.   :15.1
##  1st Qu.: 0.00                              1st Qu.: 27.29   1st Qu.:   37.312   1st Qu.:22.3
##  Median : 0.03                              Median : 45.37   Median :   87.324   Median :30.6
##  Mean   : 0.10                              Mean   : 45.94   Mean   :  456.094   Mean   :30.6
##  3rd Qu.: 0.11                              3rd Qu.: 64.39   3rd Qu.:  214.243   3rd Qu.:39.1
##  Max.   :11.71                              Max.   :100.00   Max.   :20546.766   Max.   :48.2
##  NA's   :97974                              NA's   :62189    NA's   :31058       NA's   :46802
##  aged_65_older   aged_70_older   gdp_per_capita     extreme_poverty  cardiovasc_death_rate
##  Min.   : 1.14   Min.   : 0.53   Min.   :   661.2   Min.   : 0.10    Min.   : 79.37
##  1st Qu.: 3.53   1st Qu.: 2.06   1st Qu.:  4449.9   1st Qu.: 0.60    1st Qu.:170.05
##  Median : 6.70   Median : 4.03   Median : 12951.8   Median : 2.20    Median :243.96
##  Mean   : 8.79   Mean   : 5.55   Mean   : 19526.4   Mean   :13.65    Mean   :261.55
##  3rd Qu.:14.18   3rd Qu.: 8.68   3rd Qu.: 27936.9   3rd Qu.:21.40    3rd Qu.:329.94
##  Max.   :27.05   Max.   :18.49   Max.   :116935.6   Max.   :77.60    Max.   :724.42
##  NA's   :48851   NA's   :47818   NA's   :47277      NA's   :114808   NA's   :47349
##  diabetes_prevalence female_smokers   male_smokers   handwashing_facilities
##  Min.   : 0.99       Min.   : 0.10   Min.   : 7.70   Min.   :  1.19
##  1st Qu.: 5.31       1st Qu.: 1.90   1st Qu.:21.60   1st Qu.: 20.86
##  Median : 7.20       Median : 6.30   Median :33.10   Median : 49.84
##  Mean   : 8.39       Mean   :10.68   Mean   :32.81   Mean   : 50.90
##  3rd Qu.:10.59       3rd Qu.:19.30   3rd Qu.:41.30   3rd Qu.: 83.24
##  Max.   :30.53       Max.   :44.00   Max.   :78.10   Max.   :100.00
##  NA's   :37284       NA's   :94244   NA's   :96253   NA's   :145610
##  hospital_beds_per_thousand life_expectancy human_development_index   population
##  Min.   : 0.10              Min.   :53.28   Min.   :0.39            Min.   :4.700e+01
##  1st Qu.: 1.30              1st Qu.:69.50   1st Qu.:0.60            1st Qu.:8.368e+05
##  Median : 2.50              Median :75.05   Median :0.74            Median :6.948e+06
##  Mean   : 3.09              Mean   :73.61   Mean   :0.72            Mean   :1.408e+08
##  3rd Qu.: 4.20              3rd Qu.:79.07   3rd Qu.:0.84            3rd Qu.:3.370e+07
##  Max.   :13.80              Max.   :86.75   Max.   :0.96            Max.   :7.975e+09
##  NA's   :68500              NA's   :19936   NA's   :51908           NA's   :1035
##  excess_mortality_cumulative_absolute excess_mortality_cumulative excess_mortality
##  Min.   : -37726.1                    Min.   :-28.45              Min.   :-95.92
##  1st Qu.:     53.2                    1st Qu.:  0.74              1st Qu.:  0.03
##  Median :   6301.8                    Median :  7.64              Median :  7.42
##  Mean   :  49771.5                    Mean   : 10.09              Mean   : 14.22
##  3rd Qu.:  36120.8                    3rd Qu.: 15.77              3rd Qu.: 19.39
##  Max.   :1240683.3                    Max.   : 76.55              Max.   :376.71
##  NA's   :232947                       NA's   :232947              NA's   :232918
##  excess_mortality_cumulative_per_million
```

```
##  Min.   :-1984.28
##  1st Qu.:   37.34
##  Median :  852.11
##  Mean   : 1411.09
##  3rd Qu.: 2261.22
##  Max.   :10032.84
##  NA's   :232947
```

Vemos que el conjunto de datos dispone de muchos nulos. Guardamos el total de filas en una variable, para poder calcular el porcentaje de valores faltantes por cada columna.

```
df_size <- count(df_covid)$n
df_size
```

```
## [1] 241034
```

Vemos cuantos valores perdidos tenemos en el conjunto de datos

```
colSums(is.na(df_covid))
```

```
##                                  iso_code                                 continent
##                                         0                                     13557
##                                  location                                      date
##                                         0                                         0
##                               total_cases                                 new_cases
##                                     13949                                     14234
##                         new_cases_smoothed                              total_deaths
##                                     15433                                     33326
##                                new_deaths                        new_deaths_smoothed
##                                     33410                                     34594
##                     total_cases_per_million                     new_cases_per_million
##                                     14984                                     15269
##             new_cases_smoothed_per_million                    total_deaths_per_million
##                                     16463                                     34348
##                     new_deaths_per_million               new_deaths_smoothed_per_million
##                                     34432                                     35611
##                           reproduction_rate                               icu_patients
##                                     62457                                    208350
##                    icu_patients_per_million                              hosp_patients
##                                    208350                                    205360
##                   hosp_patients_per_million                    weekly_icu_admissions
##                                    205360                                    233080
##          weekly_icu_admissions_per_million                    weekly_hosp_admissions
##                                    233080                                    221790
##         weekly_hosp_admissions_per_million                               total_tests
##                                    221790                                    161647
##                                 new_tests                    total_tests_per_thousand
##                                    165631                                    161647
##                      new_tests_per_thousand                         new_tests_smoothed
##                                    165631                                    137069
##             new_tests_smoothed_per_thousand                              positive_rate
##                                    137069                                    145107
##                               tests_per_case                               tests_units
##                                    146686                                    134246
##                           total_vaccinations                          people_vaccinated
##                                    172497                                    175505
```

```
##                     people_fully_vaccinated                                 total_boosters
##                                      178158                                         203202
##                                new_vaccinations                       new_vaccinations_smoothed
##                                      184218                                          97868
##             total_vaccinations_per_hundred                    people_vaccinated_per_hundred
##                                      172497                                         175505
##        people_fully_vaccinated_per_hundred                       total_boosters_per_hundred
##                                      178158                                         203202
##      new_vaccinations_smoothed_per_million                    new_people_vaccinated_smoothed
##                                       97868                                          97974
## new_people_vaccinated_smoothed_per_hundred                           stringency_index
##                                       97974                                          62189
##                           population_density                                     median_age
##                                       31058                                          46802
##                                aged_65_older                                  aged_70_older
##                                       48851                                          47818
##                                gdp_per_capita                               extreme_poverty
##                                       47277                                         114808
##                         cardiovasc_death_rate                           diabetes_prevalence
##                                       47349                                          37284
##                                female_smokers                                 male_smokers
##                                       94244                                          96253
##                         handwashing_facilities                      hospital_beds_per_thousand
##                                      145610                                          68500
##                               life_expectancy                         human_development_index
##                                       19936                                          51908
##                                    population             excess_mortality_cumulative_absolute
##                                         1035                                         232947
##                    excess_mortality_cumulative                                excess_mortality
##                                      232947                                         232918
##       excess_mortality_cumulative_per_million
##                                      232947
```

Vamos a ver el porcentaje de datos que no tienen nulos.

```
covid_not_nil <- sapply(df_covid, function(x) sum(!is.na(x))/df_size * 100)
covid_not_nil
```

```
##                                    iso_code                                    continent
##                                   100.000000                                   94.375482
##                                    location                                         date
##                                   100.000000                                  100.000000
##                                 total_cases                                   new_cases
##                                    94.212850                                   94.094609
##                           new_cases_smoothed                                total_deaths
##                                    93.597169                                   86.173735
##                                  new_deaths                          new_deaths_smoothed
##                                    86.138885                                   85.647668
##                       total_cases_per_million                         new_cases_per_million
##                                    93.783450                                   93.665209
##                 new_cases_smoothed_per_million                      total_deaths_per_million
##                                    93.169843                                   85.749728
##                       new_deaths_per_million                 new_deaths_smoothed_per_million
##                                    85.714878                                   85.225736
##                             reproduction_rate                                  icu_patients
```

```
##                                 74.087888                                 13.559913
##                 icu_patients_per_million                              hosp_patients
##                                 13.559913                                 14.800402
##               hosp_patients_per_million                       weekly_icu_admissions
##                                 14.800402                                  3.299949
##        weekly_icu_admissions_per_million                      weekly_hosp_admissions
##                                  3.299949                                  7.983936
##        weekly_hosp_admissions_per_million                               total_tests
##                                  7.983936                                 32.936017
##                                 new_tests                     total_tests_per_thousand
##                                 31.283138                                 32.936017
##                     new_tests_per_thousand                          new_tests_smoothed
##                                 31.283138                                 43.132919
##           new_tests_smoothed_per_thousand                             positive_rate
##                                 43.132919                                 39.798120
##                            tests_per_case                               tests_units
##                                 39.143025                                 44.304123
##                       total_vaccinations                          people_vaccinated
##                                 28.434578                                 27.186621
##                   people_fully_vaccinated                            total_boosters
##                                 26.085946                                 15.695711
##                           new_vaccinations                     new_vaccinations_smoothed
##                                 23.571778                                 59.396600
##            total_vaccinations_per_hundred               people_vaccinated_per_hundred
##                                 28.434578                                 27.186621
##       people_fully_vaccinated_per_hundred                  total_boosters_per_hundred
##                                 26.085946                                 15.695711
##     new_vaccinations_smoothed_per_million               new_people_vaccinated_smoothed
##                                 59.396600                                 59.352622
## new_people_vaccinated_smoothed_per_hundred                         stringency_index
##                                 59.352622                                 74.199076
##                           population_density                                median_age
##                                 87.114681                                 80.582822
##                              aged_65_older                             aged_70_older
##                                 79.732735                                 80.161305
##                             gdp_per_capita                           extreme_poverty
##                                 80.385755                                 52.368546
##                       cardiovasc_death_rate                       diabetes_prevalence
##                                 80.355883                                 84.531643
##                             female_smokers                              male_smokers
##                                 60.900122                                 60.066630
##                       handwashing_facilities                    hospital_beds_per_thousand
##                                 39.589436                                 71.580773
##                            life_expectancy                       human_development_index
##                                 91.728968                                 78.464449
##                                population           excess_mortality_cumulative_absolute
##                                 99.570600                                  3.355128
##                 excess_mortality_cumulative                           excess_mortality
##                                  3.355128                                  3.367160
##    excess_mortality_cumulative_per_million
##                                  3.355128
```

Creamos un subconjunto con las columnas que son susceptibles de ser empleadas en el estudio.

```
df <- df_covid[,c("iso_code"
                  ,"continent"
                  ,"location"
                  ,"date"
                  ,"total_cases"
                  ,"new_cases"
                  ,"total_deaths"
                  ,"new_deaths"
                  ,"people_vaccinated"
                  ,"people_fully_vaccinated"
                  ,"new_vaccinations"
                  ,"population_density"
                  ,"median_age"
                  ,"aged_65_older"
                  ,"aged_70_older"
                  ,"gdp_per_capita"
                  ,"extreme_poverty"
                  ,"cardiovasc_death_rate"
                  ,"diabetes_prevalence"
                  ,"male_smokers"
                  ,"female_smokers"
                  ,"life_expectancy"
                  ,"population")
               ]
```

Cargamos el conjunto de datos de los vuelos

```
df_vuelos <- read.csv("./data/airports.csv",sep=",", na.strings="",stringsAsFactors=FALSE)
summary(df_vuelos)
```

```
##    Airport.ID        Name               City             Country              IATA
##  Min.   :    1   Length:7698        Length:7698        Length:7698        Length:7698
##  1st Qu.: 1993   Class :character   Class :character   Class :character   Class :character
##  Median : 4068   Mode  :character   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 5171
##  3rd Qu.: 7729
##  Max.   :14110
##      ICAO              Latitude         Longitude           Altitude       Timezone
##  Length:7698        Min.   :-90.000   Min.   :-179.877   Min.   :-1266   Length:7698
##  Class :character   1st Qu.:  6.908   1st Qu.: -78.975   1st Qu.:   63   Class :character
##  Mode  :character   Median : 34.086   Median :   6.376   Median :  352   Mode  :character
##                     Mean   : 25.808   Mean   :  -1.391   Mean   : 1016
##                     3rd Qu.: 47.240   3rd Qu.:  56.001   3rd Qu.: 1203
##                     Max.   : 89.500   Max.   : 179.951   Max.   :14472
##      DST            Tz.database.time.zone     Type              Source
##  Length:7698        Length:7698           Length:7698        Length:7698
##  Class :character   Class :character      Class :character   Class :character
##  Mode  :character   Mode  :character      Mode  :character   Mode  :character
##
##
##
```

Vemos que el conjunto de vuelos no tiene nulos.

```
df_num_airports <- as.data.frame(table(df_vuelos$Country))
head(df_num_airports)
```

```
##             Var1 Freq
## 1    Afghanistan   22
## 2        Albania    5
## 3        Algeria   44
## 4 American Samoa    2
## 5         Angola   25
## 6       Anguilla    1
```

```
names(df_num_airports)[1]<-"country"
names(df_num_airports)[2]<-"num_airports"
colnames(df_num_airports)
```

```
## [1] "country"      "num_airports"
```

Extraemos las localizaciones que tenemos del conjunto de datos.

```
uniques_locations <- unique(df$location)
```

Vamos a comparar los paises del cojunto de datos "aeropuerto" con las localizaciones covid. Comparamos ambos valores empleando la distancia *levenshtein*.

Antes cargamos las librerías necesarias:

```
if(!require("stringdist")){
  install.packages('stringdist')
}
```

```
## Loading required package: stringdist
```

```
##
## Attaching package: 'stringdist'
```

```
## The following object is masked from 'package:magrittr':
##
##     extract
```

```
library(stringdist)
```

```
df_num_airports$equivalente<-""
df_num_airports$levenshtein<-99

for (i in 1:nrow(df_num_airports)) {
  for (j in 1:length(uniques_locations)){
    country<-as.character(df_num_airports[i, "country"])
    location<-as.character(uniques_locations[j])
    levenshtein <- stringdist(country,location,method = "lv")
     if (levenshtein<=df_num_airports[i, "levenshtein"]){
      df_num_airports[i,"equivalente"]<-location
      df_num_airports[i,"levenshtein"]<-levenshtein
    }
  }
}
```

Vemos aquellos que tiene una distancia *levenshtein* menor o igual a dos.

```
head(df_num_airports[df_num_airports$levenshtein<=2,])
```

```
##                country num_airports          equivalente levenshtein
## 1          Afghanistan           22          Afghanistan           0
## 2              Albania            5              Albania           0
## 3              Algeria           44              Algeria           0
## 5               Angola           25               Angola           0
## 6             Anguilla            1             Anguilla           0
## 8 Antigua and Barbuda            2 Antigua and Barbuda           0
```

Quitamos aquellos nombres que tienen una distancia *levenshtein* alta

```
df_num_airports$equivalente[df_num_airports$levenshtein>1] <- NA
```

Vemos los que se quedaron fuera

```
countries_out <- uniques_locations[!(uniques_locations %in% df_num_airports$equivalent)]
countries_out
```

```
##  [1] "Africa"                          "Andorra"
##  [3] "Asia"                            "Bonaire Sint Eustatius and Saba"
##  [5] "Congo"                           "Curacao"
##  [7] "Czechia"                         "Democratic Republic of Congo"
##  [9] "England"                         "Eswatini"
## [11] "Europe"                          "European Union"
## [13] "High income"                     "International"
## [15] "Kosovo"                          "Liechtenstein"
## [17] "Low income"                      "Lower middle income"
## [19] "Micronesia (country)"            "Monaco"
## [21] "North America"                   "North Macedonia"
## [23] "Northern Cyprus"                 "Northern Ireland"
## [25] "Oceania"                         "Pitcairn"
## [27] "San Marino"                      "Scotland"
## [29] "Sint Maarten (Dutch part)"       "South America"
## [31] "Timor"                           "Tokelau"
## [33] "United States Virgin Islands"    "Upper middle income"
## [35] "Vatican"                         "Wales"
## [37] "World"
```

Extramos los datos que vamos a usar del conjunto de vuelos, para juntarlo con el dataset del COVID-19.

```
df_num_airports_clean <- df_num_airports[!is.na(df_num_airports$equivalente),c("num_airports","equivale
colnames(df_num_airports_clean)<-c("airports","location")
head(df_num_airports_clean)
```

```
##   airports           location
## 1       22        Afghanistan
## 2        5            Albania
## 3       44            Algeria
## 5       25             Angola
## 6        1           Anguilla
## 8        2 Antigua and Barbuda
```

Por curiosidad, veamos a ver los paises que tiene más vuelos

```
head(df_num_airports_clean[order(-df_num_airports_clean$airports),])
```

```
##     airports      location
```

```
## 224     1512 United States
## 38       430        Canada
## 12       334     Australia
## 28       264        Brazil
## 177      264        Russia
## 80       249       Germany
```

Y a continuación el pais que tiene más nuevos casos

```
df[order(-df$new_cases),c("location")][1]
```

```
## [1] "World"
```

Juntamos ambos conjuntos de datos

```
df <- merge(x = df, y = df_num_airports_clean, by = "location")
head(df)
```

```
##      location iso_code continent       date total_cases new_cases total_deaths new_deaths
## 1 Afghanistan      AFG      Asia 2022-02-26      173146        62         7585          6
## 2 Afghanistan      AFG      Asia 2021-02-23       55646        29         2435          2
## 3 Afghanistan      AFG      Asia 2021-02-24       55664        18         2436          1
## 4 Afghanistan      AFG      Asia 2021-10-26      156071        31         7262          2
## 5 Afghanistan      AFG      Asia 2021-02-26       55696        16         2442          4
## 6 Afghanistan      AFG      Asia 2021-10-27      156124        53         7266          4
##   people_vaccinated people_fully_vaccinated new_vaccinations population_density median_age
## 1                NA                     NA               NA              54.422       18.6
## 2                NA                     NA               NA              54.422       18.6
## 3                NA                     NA               NA              54.422       18.6
## 4                NA                     NA               NA              54.422       18.6
## 5                NA                     NA               NA              54.422       18.6
## 6                NA                     NA               NA              54.422       18.6
##   aged_65_older aged_70_older gdp_per_capita extreme_poverty cardiovasc_death_rate
## 1         2.581         1.337       1803.987              NA               597.029
## 2         2.581         1.337       1803.987              NA               597.029
## 3         2.581         1.337       1803.987              NA               597.029
## 4         2.581         1.337       1803.987              NA               597.029
## 5         2.581         1.337       1803.987              NA               597.029
## 6         2.581         1.337       1803.987              NA               597.029
##   diabetes_prevalence male_smokers female_smokers life_expectancy population airports
## 1                9.59           NA             NA           64.83   41128772       22
## 2                9.59           NA             NA           64.83   41128772       22
## 3                9.59           NA             NA           64.83   41128772       22
## 4                9.59           NA             NA           64.83   41128772       22
## 5                9.59           NA             NA           64.83   41128772       22
## 6                9.59           NA             NA           64.83   41128772       22
```

Ahora vemos que dentro del dataset original disponemos del número de aeropuertos:

```
head(df[,c("airports","new_cases","date","location")])
```

```
##   airports new_cases       date    location
## 1       22        62 2022-02-26 Afghanistan
## 2       22        29 2021-02-23 Afghanistan
## 3       22        18 2021-02-24 Afghanistan
## 4       22        31 2021-10-26 Afghanistan
## 5       22        16 2021-02-26 Afghanistan
## 6       22        53 2021-10-27 Afghanistan
```

Para el análisis emplearemos este conjunto de datos:

```
summary(df)
```

```
##    location           iso_code          continent             date            total_cases
##  Length:205718      Length:205718      Length:205718      Length:205718      Min.   :       1
##  Class :character   Class :character   Class :character   Class :character   1st Qu.:    4975
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :   53656
##                                                                             Mean   : 1273212
##                                                                             3rd Qu.:  475369
##                                                                             Max.   :99230740
##                                                                             NA's   :6918
##    new_cases         total_deaths       new_deaths      people_vaccinated   people_fully_vaccinated
##  Min.   :      0   Min.   :      1   Min.   :   0.00   Min.   :0.000e+00   Min.   :1.000e+00
##  1st Qu.:      0   1st Qu.:    120   1st Qu.:   0.00   1st Qu.:5.701e+05   1st Qu.:4.886e+05
##  Median :     42   Median :   1149   Median :   0.00   Median :4.048e+06   Median :3.471e+06
##  Mean   :   3222   Mean   :  20680   Mean   :  36.05   Mean   :2.658e+07   Mean   :2.324e+07
##  3rd Qu.:    711   3rd Qu.:   8412   3rd Qu.:  10.00   3rd Qu.:1.632e+07   3rd Qu.:1.405e+07
##  Max.   :1355244   Max.   :1083362   Max.   :4529.00   Max.   :1.305e+09   Max.   :1.273e+09
##  NA's   :7189      NA's   :23805     NA's   :24045     NA's   :153916      NA's   :156371
##  new_vaccinations   population_density    median_age    aged_65_older    aged_70_older
##  Min.   :       0   Min.   :    0.137   Min.   :15.10   Min.   : 1.144   Min.   : 0.526
##  1st Qu.:    3043   1st Qu.:   36.253   1st Qu.:22.60   1st Qu.: 3.526   1st Qu.: 2.142
##  Median :   20591   Median :   85.129   Median :30.60   Median : 6.704   Median : 4.032
##  Mean   :  240738   Mean   :  375.513   Mean   :30.65   Mean   : 8.763   Mean   : 5.544
##  3rd Qu.:  104680   3rd Qu.:  212.865   3rd Qu.:38.70   3rd Qu.:14.178   3rd Qu.: 8.643
##  Max.   :24741000   Max.   :20546.766   Max.   :48.20   Max.   :27.049   Max.   :18.493
##  NA's   :161885     NA's   :10906       NA's   :21058   NA's   :23107    NA's   :22074
##  gdp_per_capita     extreme_poverty cardiovasc_death_rate diabetes_prevalence  male_smokers
##  Min.   :   661.2   Min.   : 0.10   Min.   : 79.37        Min.   : 0.990      Min.   : 7.70
##  1st Qu.:  4227.6   1st Qu.: 0.50   1st Qu.:167.29        1st Qu.: 5.290      1st Qu.:21.60
##  Median : 12951.8   Median : 2.00   Median :242.65        Median : 7.200      Median :31.40
##  Mean   : 19686.5   Mean   :12.98   Mean   :260.07        Mean   : 8.428      Mean   :32.38
##  3rd Qu.: 27936.9   3rd Qu.:18.90   3rd Qu.:329.63        3rd Qu.:10.680      3rd Qu.:41.10
##  Max.   :116935.6   Max.   :77.60   Max.   :724.42        Max.   :30.530      Max.   :76.10
##  NA's   :22275      NA's   :86235   NA's   :21617         NA's   :15586       NA's   :66999
##  female_smokers  life_expectancy   population           airports
##  Min.   : 0.1    Min.   :53.28   Min.   :1.952e+03   Min.   :   1.0
##  1st Qu.: 1.9    1st Qu.:69.02   1st Qu.:1.121e+06   1st Qu.:   3.0
##  Median : 6.2    Median :74.99   Median :7.489e+06   Median :  11.0
##  Mean   :10.6    Mean   :73.46   Mean   :3.934e+07   Mean   :  37.6
##  3rd Qu.:19.1    3rd Qu.:78.92   3rd Qu.:2.816e+07   3rd Qu.:  29.0
##  Max.   :44.0    Max.   :84.86   Max.   :1.426e+09   Max.   :1512.0
##  NA's   :64990   NA's   :1305
```

```
head(df,10)
```

```
##       location iso_code continent       date total_cases new_cases total_deaths new_deaths
## 1  Afghanistan      AFG      Asia 2022-02-26      173146        62         7585          6
## 2  Afghanistan      AFG      Asia 2021-02-23       55646        29         2435          2
## 3  Afghanistan      AFG      Asia 2021-02-24       55664        18         2436          1
## 4  Afghanistan      AFG      Asia 2021-10-26      156071        31         7262          2
## 5  Afghanistan      AFG      Asia 2021-02-26       55696        16         2442          4
## 6  Afghanistan      AFG      Asia 2021-10-27      156124        53         7266          4
## 7  Afghanistan      AFG      Asia 2022-03-06      174582       251         7623          1
```

```
## 8  Afghanistan      AFG      Asia 2021-02-25       55680        16         2438          2
## 9  Afghanistan      AFG      Asia 2021-09-14      154180        86         7171          2
## 10 Afghanistan      AFG      Asia 2022-03-04      174214         0         7619          0
##    people_vaccinated people_fully_vaccinated new_vaccinations population_density median_age
## 1                 NA                     NA               NA              54.422       18.6
## 2                 NA                     NA               NA              54.422       18.6
## 3                 NA                     NA               NA              54.422       18.6
## 4                 NA                     NA               NA              54.422       18.6
## 5                 NA                     NA               NA              54.422       18.6
## 6                 NA                     NA               NA              54.422       18.6
## 7            4952744                4281934               NA              54.422       18.6
## 8                 NA                     NA               NA              54.422       18.6
## 9                 NA                     NA               NA              54.422       18.6
## 10                NA                     NA               NA              54.422       18.6
##    aged_65_older aged_70_older gdp_per_capita extreme_poverty cardiovasc_death_rate
## 1          2.581         1.337       1803.987              NA                597.029
## 2          2.581         1.337       1803.987              NA                597.029
## 3          2.581         1.337       1803.987              NA                597.029
## 4          2.581         1.337       1803.987              NA                597.029
## 5          2.581         1.337       1803.987              NA                597.029
## 6          2.581         1.337       1803.987              NA                597.029
## 7          2.581         1.337       1803.987              NA                597.029
## 8          2.581         1.337       1803.987              NA                597.029
## 9          2.581         1.337       1803.987              NA                597.029
## 10         2.581         1.337       1803.987              NA                597.029
##    diabetes_prevalence male_smokers female_smokers life_expectancy population airports
## 1                 9.59           NA             NA           64.83   41128772       22
## 2                 9.59           NA             NA           64.83   41128772       22
## 3                 9.59           NA             NA           64.83   41128772       22
## 4                 9.59           NA             NA           64.83   41128772       22
## 5                 9.59           NA             NA           64.83   41128772       22
## 6                 9.59           NA             NA           64.83   41128772       22
## 7                 9.59           NA             NA           64.83   41128772       22
## 8                 9.59           NA             NA           64.83   41128772       22
## 9                 9.59           NA             NA           64.83   41128772       22
## 10                9.59           NA             NA           64.83   41128772       22
```

```
write.csv(df,"./data/clean_dataset.csv")
```