Práctica 2

0. Integrantes.

Bryan Steven Cortez Chichande

César Alexander Guzmán Vásquez

1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?

El tema principal del presente caso práctico es la pandemia de COVID-19 y su evolución a nivel mundial en temas de contagios, personas recuperadas, mortalidad y vacunación. Por tal razón, se dispuso la utilización de datos recopilados por Organizaciones autorizadas en el tema y cuyo período de tiempo sea el año 2021. En la búsqueda de los datos que describan cada indicador influyente en el tema antes mencionado, se recurrió a la utilización de dos conjuntos de datos, los cuales se pueden visualizar en su formato original en la carpeta "Datasets CSV", o importados en Excel para facilitar en su visualización, en la carpeta "Datasets Excel". A la vez, queremos comparar al Ecuador, frente a otros países para observar su desempeño en la pandemia, ya que es el país natal de los integrantes de este grupo, y nos resulta interesante compararlo con otros países.

A continuación, se muestra la descripción de los Datasets, obtenidos de la página Kaggle:

Covid-19 Global Dataset. Cifras actualizadas de casos diarios confirmados, fallecidos y activos de 218 países. Los atributos del conjunto de datos son los siguientes:

Tabla 1. Atributos del conjunto de datos COVID-19 Global Dataset

Nombre	Tipo
Fecha	Fecha
País	Cadena
Total de casos acumulados	Numérico
Casos diarios nuevos	Numérico
Casos activos	Numérico
Total de decesos acumulados	Numérico
Decesos diarios nuevos	Numérico

Número de registros: 145.221

COVID-19 World Vaccination Progress. Vacunación diaria y total contra COVID-19 en el mundo. Los atributos del conjunto de datos son los siguientes:

Tabla 2. Atributos del conjunto de datos COVID-19 World Vaccination Progress

Nombre	Tipo
País	Cadena
Código ISO	Cadena
Fecha	Cadena
Total de vacunaciones	Numérico
Personas vacunadas	Numérico
Personas completamente vacunadas	Numérico
Tasa de vacunaciones diarias	Numérico
Vacunaciones diarias	Numérico
Vacunaciones totales por ciento	Numérico
Personas vacunadas por ciento	Numérico
Personas completamente vacunadas por ciento	Numérico
Vacunaciones diarias por millón	Numérico
Vacunas	Cadena
Nombre de la fuente	Cadena
Nombre del sitio web	Cadena

Número de observaciones: 63.401

A continuación, se muestra el fragmento de código del programa en R, donde se importan los dataset:

```
28
29 # Importación de Datasets
30
31 A continuación procedemos a leer los archivos del directorio raíz del programa.
32
33 ```{r message= FALSE, warning=FALSE}
34 ds1.covid<-read.csv("./Covid-19-Global-Dataset.csv",header=T,sep=",")
35 ds2.covid<-read.csv("./COVID-19-World-Vaccination-Progress.csv",header=T,sep=",")
36 ```
```

Figura 1. Código utilizado para la lectura de los Datasets

La importancia de los datos recopilados en los dos conjuntos de datos es la información aportada en relación a la evolución de la pandemia actual tomando en cuenta por un lado las consecuencias (contagios, recuperaciones, mortalidad) y por el otro las herramientas reactivas contra la enfermedad y sus efectos (vacunación).

```
43
44 ```{r message= FALSE, warning=FALSE}
45 str(ds1.covid)
46 ```
```

```
'data.frame':   145221 obs. of  7 variables:
 $ date                  : chr  "2020-2-15" "2020-2-16" "2020-2-17" "2020-2-18" ...
 $ country               : chr  "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...
 $ cumulative_total_cases : num  0 0 0 0 0 0 0 0 0 1 ...
 $ daily_new_cases        : num  NA NA NA NA NA NA NA NA NA NA ...
 $ active_cases           : num  0 0 0 0 0 0 0 0 0 1 ...
 $ cumulative_total_deaths: num  0 0 0 0 0 0 0 0 0 0 ...
 $ daily_new_deaths       : num  NA NA NA NA NA NA NA NA NA NA ...
```

Figura 2. Descripción del primer conjunto de datos

```
49
50 ```{r message= FALSE, warning=FALSE}
51 str(ds2.covid)
52 ```
```

```
'data.frame':   63401 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 34 34 34 34 34 34 40 45 50 ...
 $ vaccines                          : chr  "Johnson&Johnson, Oxford/AstraZeneca, Pfizer/BioNTech,
Sinopharm/Beijing" "Johnson&Johnson, Oxford/AstraZeneca, Pfizer/BioNTech, Sinopharm/Beijing" "Johnson&Johnson,
Oxford/AstraZeneca, Pfizer/BioNTech, Sinopharm/Beijing" "Johnson&Johnson, Oxford/AstraZeneca, Pfizer/BioNTech,
Sinopharm/Beijing" ...
 $ source_name                       : chr  "World Health Organization" "World Health Organization" "World Health
Organization" "World Health Organization" ...
 $ source_website                    : chr  "https://reliefweb.int/sites/reliefweb.int/files/resources/weekly-
epidemiological-bulletin_w47.pdf" "https://reliefweb.int/sites/reliefweb.int/files/resources/weekly-
epidemiological-bulletin_w47.pdf" "https://reliefweb.int/sites/reliefweb.int/files/resources/weekly-
epidemiological-bulletin_w47.pdf" "https://reliefweb.int/sites/reliefweb.int/files/resources/weekly-
epidemiological-bulletin_w47.pdf" ...
```

Figura 3. Descripción del segundo conjunto de datos

La interrogante que se pretende responder es la siguiente:

- ¿Qué efectos positivos tiene el programa de vacunación mundial contra el COVID-19 tomando en cuenta las estadísticas de las consecuencias?
- ¿Cuál es el tipo de relación (Directamente/Inversamente proporcional) entre la evolución de las tasas de contagios, recuperaciones, muertes y los programas de vacunación?

2. Integración y selección de los datos de interés a analizar.

En esta etapa se realizaron las siguientes actividades:

a) Identificación de la escala de tiempo de los dos conjuntos de datos.

Figura 4. Obtención de los valores únicos del campo Año

Debido a que la etapa de vacunación en la mayoría de países se intensifico en el año 2021, se decidió trabajar con los datos del año antes mencionado.

b) Obtención de los valores año y mes derivados del campo ***date***, para realizar agregaciones y filtrado de datos a partir de los mismos.

```{r message= FALSE, warning=FALSE}
ds1.covid$date=as.Date(ds1.covid$date)
ds1.covid$year=as.integer(format(ds1.covid$date, format="%Y"))
ds1.covid$month=as.integer(format(ds1.covid$date, format="%m"))
```

Figura 5. Operaciones sobre el primer conjunto de datos

```{r message= FALSE, warning=FALSE}
ds2.covid$date=as.Date(ds2.covid$date)
ds2.covid$year=as.integer(format(ds2.covid$date, format="%Y"))
ds2.covid$month=as.integer(format(ds2.covid$date, format="%m"))
```

Figura 6. Operaciones sobre el segundo conjunto de datos

c) Realización de operaciones de agrupamiento tomando en cuenta las columnas country, year y month. Además, debido a que los datos eran valores acumulados en relación al incremento de las fechas, se realizó un filtrado un filtrado de las fechas máximas por cada país, año y mes. Dicha operación ayudo a no obtener datos agrupados erróneos, ya que si se efectuaba una operación de suma por país, año y mes; los valores que se obtendrían no explicarían la verdadera realidad. Finalmente, se llevó a cabo un proceso de selección de los atributos o columnas relevantes para el estudio.

```{r message= FALSE, warning=FALSE}
ds1.covid.modf = data.frame(ds1.covid %>%
                  group_by(country,strftime(date, "%Y-%m")) %>%
                  filter(date == max(date)))
```

Figura 7. Filtrado de datos del primer conjunto de datos

```r
```{r message= FALSE, warning=FALSE}
ds1.covid.modf=data.frame(ds1.covid.modf[8:9],ds1.covid.modf[2:7])
```
```

Figura 8. Selección de datos del primer conjunto de datos

```r
```{r message= FALSE, warning=FALSE}
ds2.covid.modf = data.frame(ds2.covid %>%
                    group_by(country,strftime(date, "%Y-%m")) %>%
                    filter(date == max(date)))
```
```

Figura 9. Filtrado de datos del segundo conjunto de datos

```r
```{r message= FALSE, warning=FALSE}
ds2.covid<-ds2.covid %>% select(c(1,3:12,16,17))
```
```

Figura 10. Primera selección de columnas del segundo conjunto de datos

```r
```{r message= FALSE, warning=FALSE}
ds2.covid.modf=data.frame(ds2.covid.modf[12:13],ds2.covid.modf[1],ds2.covid.modf[3:11])
```
```

Figura 11. Segunda selección de columnas del segundo conjunto de datos

d) Fusión de los conjuntos de datos a partir de los campos similares, los mismos que se muestran a continuación:

- country
- year
- month

```r
```{r message= FALSE, warning=FALSE}
ds.covid.cst <- merge(ds1.covid.modf, ds2.covid.modf,by=c("country","year","month"))
```
```

Figura 12. Fusión de los dos conjuntos de datos

3. Limpieza de los datos.

La fase de limpieza de los datos ayuda a corregir, normalizar, escalar y eliminar valores que no siguen un formato completo o válido.

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

El conjunto de datos tiene 1071 ceros distribuidos entre las columnas numéricas relevantes.

```r
```{r message= FALSE, warning=FALSE}
cant.ceros <- data.frame(colSums(ds.covid.cst[4:17] == 0, na.rm = T))
names(cant.ceros) <- c("cantidad")
```

Vemos la cantidad total de ceros.

```{r message= FALSE, warning=FALSE}
sum(cant.ceros$cantidad)
```
```

```
[1] 1071
```

Figura 13. Obtención de la cantidad de ceros del conjunto de datos

Además, el conjunto de datos tiene una gran cantidad de valores vacíos (6336), debido al contexto del mismo. Hay varias columnas cuyos valores no han sido recuperados debido a su variabilidad y el tiempo de recopilación.

```r
```{r message= FALSE, warning=FALSE}
cant.na <- data.frame(colSums(is.na(ds.covid.cst[4:17])))
names(cant.na) <- c("cantidad")
```

Vemos la cantidad total de valores NA.

```{r message= FALSE, warning=FALSE}
sum(cant.na$cantidad)
```
```

```
[1] 6336
```

Figura 14. Obtención de la cantidad de valores vacíos del conjunto de datos

Debido a que los valores vacíos en este conjunto de datos en particular no pueden pasar por un proceso de imputación o completamiento utilizando alguna técnica o algoritmo especializado, se procedió a remplazar dichos valores por cero.

```r
```{r message= FALSE, warning=FALSE}
ds.covid.cst[is.na(ds.covid.cst)] <- 0
```
```

Figura 15. Remplazo de valores vacíos por ceros

3.2 Identificación y tratamiento de valores extremos.

En el contexto del conjunto de datos, aparecen varios ceros debido al tiempo de recopilación de los mismo. Por ejemplo, en los primeros meses de la pandemia no había un proceso de

vacunación; por lo tanto, en este tipo de casos van a aparecer ceros. Otro caso similar se presenta con los países que no tienen estadísticas claras al inicio de la pandemia, ya sea en la cifra de contagios, muertes, recuperaciones, vacunados, entre otros. En base a lo antes expuesto, se considera que no se debe llevar a cabo un proceso para el tratamiento de los valores extremos.

3.3 Exportación del conjunto de datos resultante

El conjunto de datos resultante de las actividades anteriores se procede a almacenar en un archivo de fácil lectura (CSV) para futuros análisis.

```
176
177 ```{r message= FALSE, warning=FALSE}
178 write.csv(x=ds.covid.cst, file="dataset_covid.csv")
179 ```
```

Figura 16. Guardado del conjunto de datos

4. Análisis de los datos.

4.1. Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).

Para temas prácticos, vamos a enfocarnos en analizar cuatro países con los cuales haremos las comparativas. Para ellos escogimos a Ecuador, porque es el país de los dos miembros que integran este equipo; España, que es el único país desarrollado de habla hispana y el cual sirve de referencia para las comparativas respecto a los países de Latinoamérica; Alemania, por ser el país más rico de la Unión Europea y uno de los que aparentemente sobrellevaron al inicio la pandemia y por último, Estados Unidos, que es el país más rico del mundo, pero que tuvo bastantes dificultades sobrellevando la pandemia, al menos al inicio.
Por tanto, vamos a proceder analizando que tal lo ha hecho nuestro país frente a las grandes potencias mundiales.

```
185 ```{r message= FALSE, warning=FALSE}
186 ds.covid.info_ec <- ds.covid.cst[ds.covid.cst$country == "Ecuador",]
187 ds.covid.info_es <- ds.covid.cst[ds.covid.cst$country == "Spain",]
188 ds.covid.info_de <- ds.covid.cst[ds.covid.cst$country == "Germany",]
189 ds.covid.info_us <- ds.covid.cst[ds.covid.cst$country == "United States",]
190 ```
```

Figura 17. Separación de los datos en los países seleccionados

4.2. Comprobación de la normalidad y homogeneidad de la varianza.

Vamos a proceder comprobando los valores que toman las variables cuantitativas provienen de una población distribuida normalmente, para ello utilizaremos la prueba de normalidad de Anderson-Darling.

Para ello, vamos a comprobar cada variable si obtiene un p-valor superior al nivel de significación prefijado α = 0, 05. Si esto se cumple, entonces se considera que variable en cuestión sigue una distribución normal.

```r
```{r message= FALSE, warning=FALSE}
library(nortest)


alpha = 0.05
col.names = colnames(ds.covid.cst)
for (i in 1:ncol(ds.covid.cst)) {
  if (i == 1) cat("Variables que no siguen una distribución normal:\n")
  if (is.integer(ds.covid.cst[,i]) | is.numeric(ds.covid.cst[,i])) {
    p_val = ad.test(ds.covid.cst[,i])$p.value
    if (p_val < alpha) {
      cat(col.names[i])

      if (i < ncol(ds.covid.cst) - 1) cat(", ")
      if (i %% 3 == 0) cat("\n")
    }
  }
}
```
```

```
Variables que no siguen una distribución normal:
year, month,
cumulative_total_cases, daily_new_cases, active_cases,
cumulative_total_deaths, daily_new_deaths, total_vaccinations,
people_vaccinated, people_fully_vaccinated, daily_vaccinations_raw,
daily_vaccinations, total_vaccinations_per_hundred, people_vaccinated_per_hundred,
people_fully_vaccinated_per_hundred, daily_vaccinations_per_million
```

Figura 18. Comprobación de la normalidad de las variables

Continuamos con la homogeneidad de varianzas mediante la aplicación de un test de Fligner-Killeen. Para ello, estudiaremos esta homogeneidad en cuanto a los grupos conformados por los casos acumulados y las muertes acumuladas. En el siguiente test, la hipótesis nula consiste en que ambas varianzas son iguales.

```r
```{r message= FALSE, warning=FALSE}
fligner.test(cumulative_total_cases ~ cumulative_total_deaths, data = ds.covid.cst)
```
```

```
        Fligner-Killeen test of homogeneity of variances

data:  cumulative_total_cases by cumulative_total_deaths
Fligner-Killeen:med chi-squared = 1746.2, df = 1524, p-value = 5.77e-05
```

Figura 19. Comprobación de la homogeneidad de la varianza

Como podemos ver, el valor obtenido de p-value es inferior al aceptado de 0.05, desestimamos que la hipótesis de que las varianzas de ambas muestras sean homogéneas.

4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.

Primero, comenzaremos con un análisis visual de los datos del último año. Viendo de la distribución de casos diarios, casos acumulados, decesos y vacunación en el tiempo en Ecuador.

Como se puede visualizar en la primera gráfica, vemos que a partir de mayo de 2021, cuando se posicionó el nuevo gobierno, comenzaron a descender los casos nuevos casos, gracias a la implementación del nuevo plan de vacunación. Lo mismo se puede ver en las otras gráficas que se marca una diferencia a partir de mayo en las tendencias, empezando desde abajo hasta llegar al punto de la nueva ola formada por las variantes Delta y Omicron.

NOTA: Para ver con mejor detalle las gráficas, puede hacerlo desde el archivo html generado por R.

```
```{r message= FALSE, warning=FALSE}
plot(ds.covid.info_ec$daily_new_cases)

plot(ds.covid.info_ec$cumulative_total_cases)

plot(ds.covid.info_ec$cumulative_total_deaths)

plot(ds.covid.info_ec$people_fully_vaccinated)

```
```

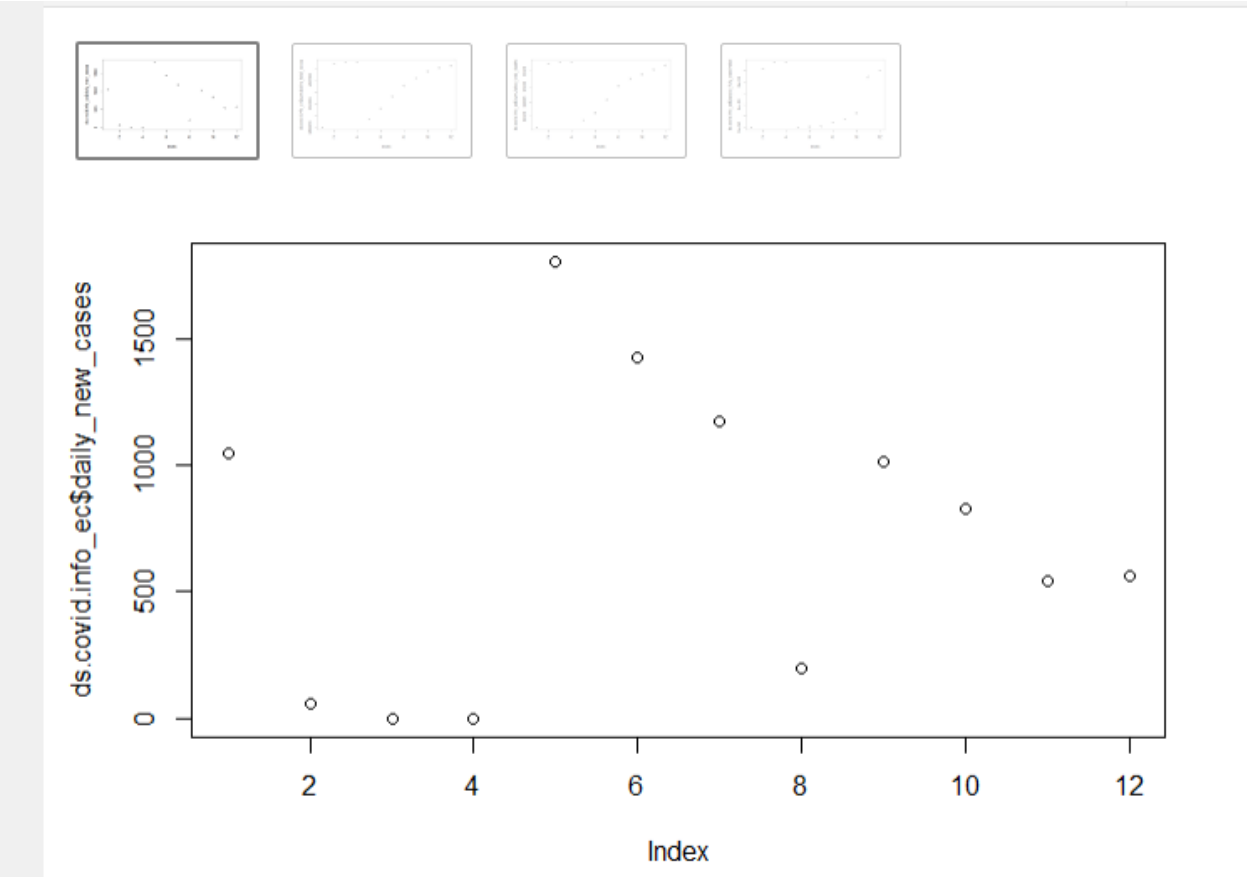

Figura 20. Gráficos del año 2021 de Ecuador

En el caso de España, podemos ver una tendencia totalmente diferente y errática de los nuevos casos diarios, ya que no hay una tendencia clara, se puede ver un repunte en el verano, pero luego un descenso, sin embargo, en octubre por alguna razón vemos dispararse los contagios a nivel récord en todo el año. En el caso de los contagios se ve algo similar a Ecuador, con la marca de cada ola de la pandemia, sin embargo, en el caso de las muertes se ve una más abultada la ola que la del Ecuador. Y por otro lado en la vacunación se ve como a partir de julio empieza una vacunación exponencial.

```
283 ```{r message= FALSE, warning=FALSE}
284 plot(ds.covid.info_es$daily_new_cases)
285
286 plot(ds.covid.info_es$cumulative_total_cases)
287
288 plot(ds.covid.info_es$cumulative_total_deaths)
289
290 plot(ds.covid.info_es$people_fully_vaccinated)
291
292 ```
```

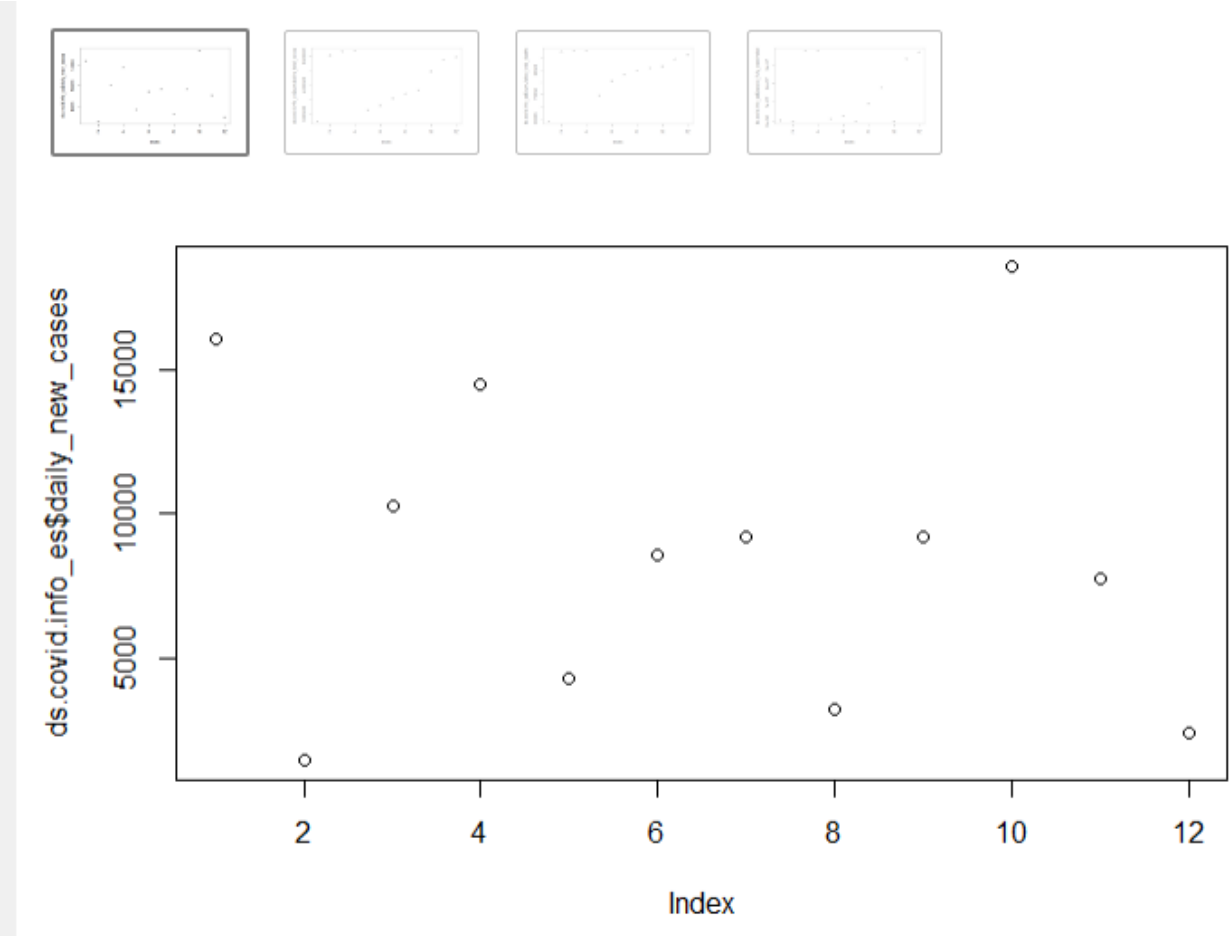

Figura 21. Gráficos del año 2021 de España

En el caso de Alemania vemos algo totalmente diferente, su pico más alto de nuevos casos diarios fue en mayo, justo a puertas del verano, sin embargo y curiosamente, vemos que en junio

comienza un descenso drástico, hasta que en noviembre vemos escalar un poco los contagios, presuntamente por la nueva variante omicron. En el gráfico de los casos y muertes acumuladas, se ven similares a España, con las marcadas olas que se dieron este año, y la vacunación igualmente desde junio se ve como empieza a escalar.

```
297
298 ```{r message= FALSE, warning=FALSE}
299 plot(ds.covid.info_de$daily_new_cases)
300
301 plot(ds.covid.info_de$cumulative_total_cases)
302
303 plot(ds.covid.info_de$cumulative_total_deaths)
304
305 plot(ds.covid.info_de$people_fully_vaccinated)
306
```

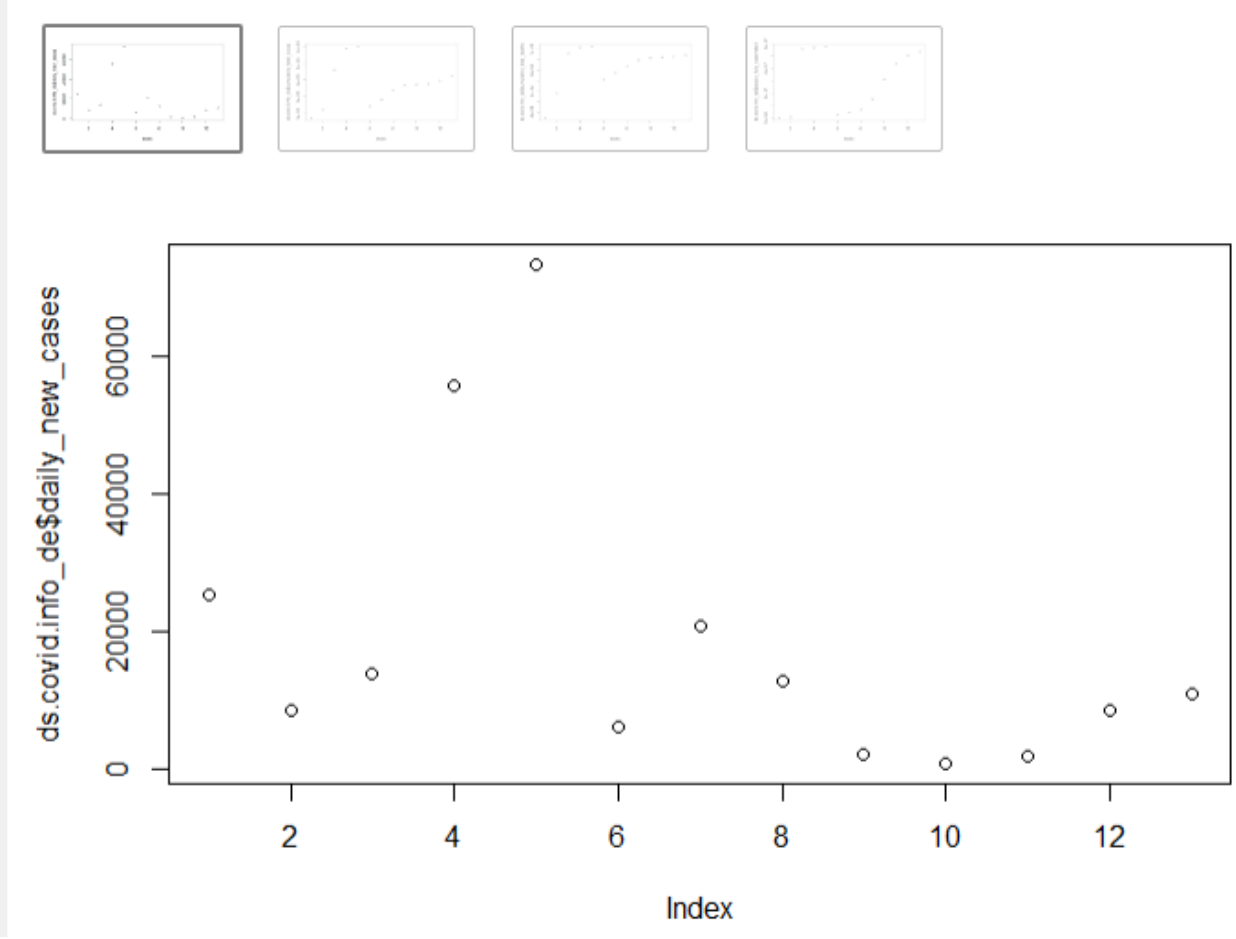

Figura 22. Gráficos del año 2021 de Alemania

Por último, en el caso de Estados Unidos, en los nuevos casos diarios, se ve un comportamiento errático como el de España, pero con la diferencia que su pico más alto está en enero, probablemente porque se asumían las consecuencias del descuido de Trump en el manejo de la pandemia, y las medidas tomadas por Biden todavía no tenían efecto. En cuanto al resto de gráficas, el comportamiento es similar al de Alemania, se marcan las olas, y la vacunación empieza a crecer bastante a partir de junio.

```r
```{r message= FALSE, warning=FALSE}
plot(ds.covid.info_us$daily_new_cases)

plot(ds.covid.info_us$cumulative_total_cases)

plot(ds.covid.info_us$cumulative_total_deaths)

plot(ds.covid.info_us$people_fully_vaccinated)
```

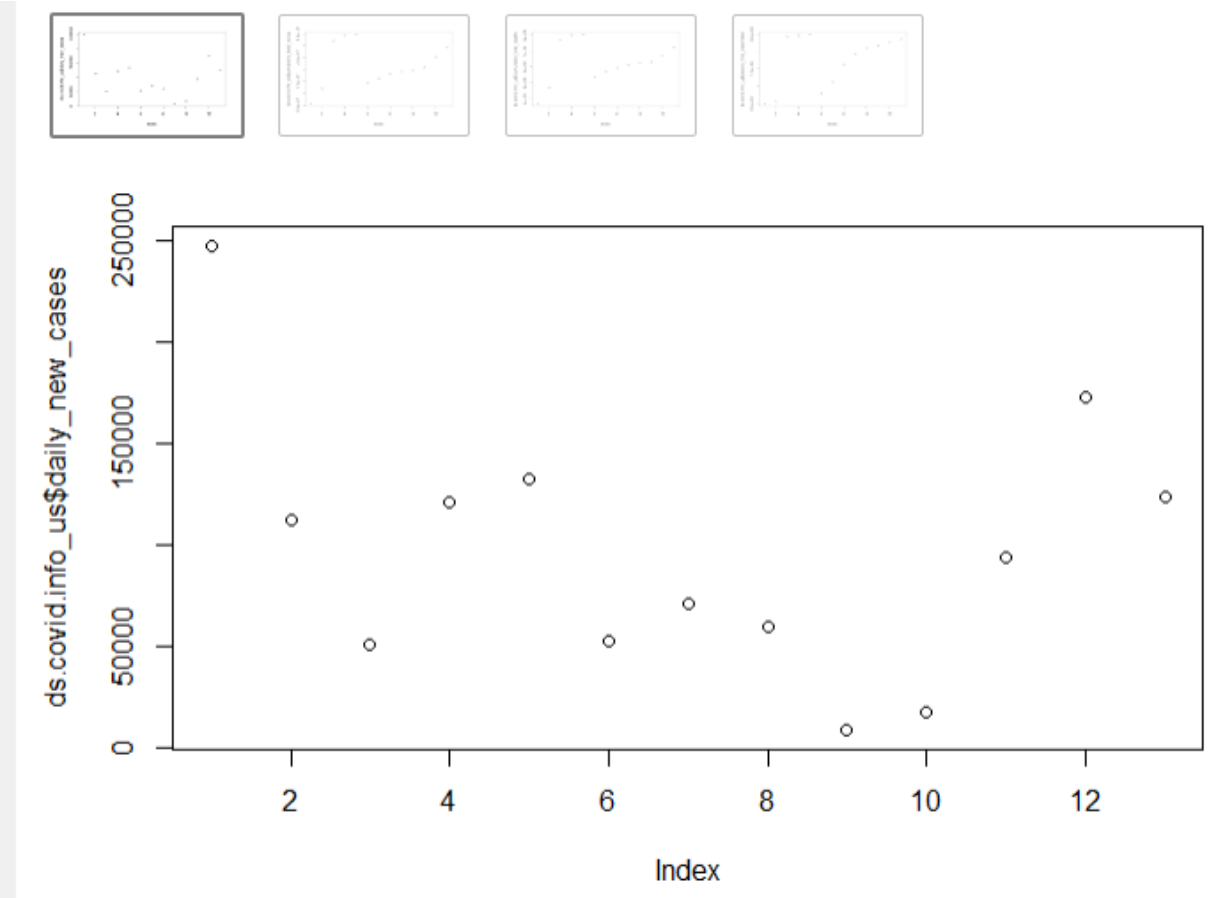

Figura 23. Gráficos del año 2021 de Estados Unidos

- Correlación de variables

Procedemos a realizar un análisis de correlación entre todas las variables para determinar cuáles de ellas ejercen una mayor influencia sobre el total de muertes acumuladas. Para lo cual usaremos el método de Spearman, puesto que hemos visto que tenemos datos que no siguen una distribución normal, por lo que no nos sirve el de Pearson. Adicionalmente cada coeficiente de correlación también muestra el p-valor asociado, para dar más información.
```

```r
corr_matrix <- matrix(nc = 2, nr = 0)
colnames(corr_matrix) <- c("estimate", "p-value")

for (i in 1:(ncol(ds.covid.cst) - 1)) {
  if (is.integer(ds.covid.cst[,i]) | is.numeric(ds.covid.cst[,i])) {
    spearman_test = cor.test(ds.covid.cst[,i],
                             ds.covid.cst[,7],
                             method = "spearman")
    corr_coef = spearman_test$estimate
    p_val = spearman_test$p.value

    pair = matrix(ncol = 2, nrow = 1)
    pair[1][1] = corr_coef
    pair[2][1] = p_val
    corr_matrix <- rbind(corr_matrix, pair)
    rownames(corr_matrix)[nrow(corr_matrix)] <- colnames(ds.covid.cst)[i]
  }
}

print(corr_matrix)
```

```
                                  estimate       p-value
year                           -0.03329326  1.408399e-01
month                           0.07477283  9.290738e-04
cumulative_total_cases          0.95792371  0.000000e+00
daily_new_cases                 0.78042855  0.000000e+00
active_cases                    0.79028569  0.000000e+00
cumulative_total_deaths         1.00000000  0.000000e+00
daily_new_deaths                0.79687080  0.000000e+00
total_vaccinations              0.55973068 7.359973e-162
people_vaccinated               0.52898636 1.275459e-141
people_fully_vaccinated         0.53093772 7.629395e-143
daily_vaccinations_raw          0.53223948 1.153569e-143
daily_vaccinations              0.76062117  0.000000e+00
total_vaccinations_per_hundred  0.32997725  5.903873e-51
people_vaccinated_per_hundred   0.31117394  3.192217e-45
people_fully_vaccinated_per_hundred 0.33617337  6.183155e-53
daily_vaccinations_per_million  0.16762742  8.295854e-14
```

Figura 24. Correlación de las variables

Como podemos ver las variables más correlacionadas con el total de muertes acumuladas en función de su proximidad con los valores -1 y +1 son el total de casos acumulados, los casos activos, los nuevos casos diarios, las muertes diarias y las vacunaciones diarias.

- Prueba de Contraste

La segunda prueba estadística que vamos a realizar, será una prueba de contraste de hipótesis, la cual plantea que "La proporción de muertes por Covid-19 es menor en los países desarrollados?"

Para ello, usaremos la comparativa de Ecuador, con España, Alemania y Estados Unidos, que tienen un PIB per cápita entre 6, 10 y 12 veces mayor respectivamente, por lo que en teoría deberían tener un mejor sistema de salud.

```
```{r message= FALSE, warning=FALSE}

t.test(ds.covid.info_ec$cumulative_total_deaths, ds.covid.info_es$cumulative_total_deaths, alternative = "less")

t.test(ds.covid.info_ec$cumulative_total_deaths, ds.covid.info_de$cumulative_total_deaths, alternative = "less")

t.test(ds.covid.info_ec$cumulative_total_deaths, ds.covid.info_us$cumulative_total_deaths, alternative = "less")
```
```

```
        Welch Two Sample t-test

data:  ds.covid.info_ec$cumulative_total_deaths and ds.covid.info_es$cumulative_total_deaths
t = -18.602, df = 13.889, p-value = 1.632e-11
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -45009.42
sample estimates:
mean of x mean of y
 30167.58  79887.25


        Welch Two Sample t-test

data:  ds.covid.info_ec$cumulative_total_deaths and ds.covid.info_de$cumulative_total_deaths
t = -9.7556, df = 12.691, p-value = 1.467e-07
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -43732.88
sample estimates:
mean of x mean of y
 30167.58  83622.15


        Welch Two Sample t-test

data:  ds.covid.info_ec$cumulative_total_deaths and ds.covid.info_us$cumulative_total_deaths
t = -16.837, df = 12.016, p-value = 5.058e-10
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -536844.7
sample estimates:
mean of x mean of y
 30167.58 630561.85
```

Figura 25. Resultados de las pruebas de contraste

Si tomamos un α=0.05, entonces vemos que nuestrop-value en todos los casos es menor que el valor de significación fijado. Por tanto vemos que se cumple lo propuesto en la hipótesis.

Aunque si comparamos la proporción de las muertes acumuladas con respecto a la población de cada país, vemos que únicamente Alemania, es el único que realmente tiene una proporción baja de muertes, ya que los otros países sobrepasan el 2%, por lo que podemos decir que no basta la riqueza del país, sino también la gestión y la cultura de cada país para evitar mayor proporción de decesos.

```r
```{r message= FALSE, warning=FALSE}

sum(ds.covid.info_ec$cumulative_total_deaths)

sum(ds.covid.info_es$cumulative_total_deaths)

sum(ds.covid.info_de$cumulative_total_deaths)

sum(ds.covid.info_us$cumulative_total_deaths)

p_ec <- (sum(ds.covid.info_ec$cumulative_total_deaths) / 17640000) * 100

p_es <- (sum(ds.covid.info_es$cumulative_total_deaths) / 47350000) * 100

p_de <- (sum(ds.covid.info_de$cumulative_total_deaths) / 83240000) * 100

p_us <- (sum(ds.covid.info_us$cumulative_total_deaths) / 329500000) * 100

str(p_ec)

str(p_es)

str(p_de)

str(p_us)
```
```
[1] 362011
[1] 958647
[1] 1087088
[1] 8197304
 num 2.05
 num 2.02
 num 1.31
 num 2.49
```

Figura 26. Proporcionalidad de muertes según la población de cada país

- Regresión Lineal

Con el objetivo de realizar futuras predicciones, se plantea realizar un modelo de regresión lineal utilizando regresores tanto cuantitativos como cualitativos con el que poder realizar predicciones sobre las muertes acumuladas.

Por tanto, vamos a proceder creando varios modelos de regresión utilizando las variables que estén más correlacionadas con las muertes acumuladas, como ya lo vimos en apartados anteriores. Una vez planteados los modelos, podremos escoger el mejor utilizando como criterio de mayor coeficiente de determinación (R2).
```

```r
```{r message= FALSE, warning=FALSE}

cumulative_total_deaths = ds.covid.cst$cumulative_total_deaths
cumulative_total_cases = ds.covid.cst$cumulative_total_cases
active_cases = ds.covid.cst$active_cases
daily_new_cases = ds.covid.cst$daily_new_cases
daily_new_deaths = ds.covid.cst$daily_new_deaths
daily_vaccinations = ds.covid.cst$daily_vaccinations
country = ds.covid.cst$country


model1 <- lm(cumulative_total_deaths ~ cumulative_total_cases + active_cases + daily_new_cases + daily_new_deaths + ds.covid.cst$daily_vaccinations  + country, data = ds.covid.cst)

model2 <- lm(cumulative_total_deaths ~ daily_new_cases + daily_new_deaths + daily_vaccinations  + country, data = ds.covid.cst)

model3 <- lm(cumulative_total_deaths ~ cumulative_total_cases + active_cases + country, data = ds.covid.cst)

model4 <- lm(cumulative_total_deaths ~ active_cases + daily_new_cases + daily_vaccinations  + country, data = ds.covid.cst)

model5 <- lm(cumulative_total_deaths ~ cumulative_total_cases + daily_new_cases + daily_vaccinations  + country, data = ds.covid.cst)

```
```

Figura 27. Definición de los modelos

Una vez planteados los modelos con las diferentes combinaciones de variables, continuamos con la comparativa de los coeficientes R2.

```r
```{r message= FALSE, warning=FALSE}

coeficients_table <- matrix(c(1, summary(model1)$r.squared,
                              2, summary(model2)$r.squared,
                              3, summary(model3)$r.squared,
                              4, summary(model4)$r.squared,
                              5, summary(model5)$r.squared),
                            ncol = 2, byrow = TRUE)


colnames(coeficients_table) <- c("Modelo", "R^2")
coeficients_table
```
```

```
     Modelo       R^2
[1,]      1 0.9893315
[2,]      2 0.9387971
[3,]      3 0.9891282
[4,]      4 0.9416355
[5,]      5 0.9887859
```

Figura 28. Comparativa de los coeficientes R2

Como vemos, al final la mejor combinación, es el primer modelo, que usa todas las variables, sin embargo, el quinto modelo, donde prescindimos de los casos activos, está muy cerca del resultado del modelo 1, por lo que podemos concluir que esa variable realmente no es tan importante a pesar que guarda una alta correlación con el número de muertes acumuladas.

5. Conclusiones

En esta práctica, se han cumplido los objetivos del curso, al haber puesto en práctica todo lo aprendido en la materia, empezando desde la preparación de los datos, eliminando valores nulos, filtrando, reordenando las columnas y juntándolos en un nuevo dataset que puede ser usado por cualquiera persona interesada en ampliar su conocimiento.

Como hemos podido observar, se han realizado varios tipos de pruebas estadísticas sobre el dataset construido a partir de dos datasets más pequeños, que nos ha permitido tener una visión más amplia del problema que nos ha permitido en lo posible cumplir con el objetivo propuesto inicialmente.

Mediante cálculos, tablas y gráficos hemos podido ver las principales variables que afectan en la incidencia de muertes causadas por Covid-19, su contraste y los modelos que mejor se adaptarían para predecir futuros comportamientos de la pandemia.

Para concluir, hemos visto que el Ecuador dentro todo y a pesar de ser un país subdesarrollado, frente a otros países que los hemos comparado, como las naciones hermanas de España, Alemania y Estados Unidos, lo ha hecho relativamente bien en el año 2021 en lo que fue la tercera y cuarta ola de la pandemia. Confirmando el hecho de que otros factores como la gestión de los gobiernos y factores culturales tienen también una relativa importancia en el problema.

6. Recursos

Guitart Hormigo, I. (2019). Introducción a la limpieza y análisis de los datos. In L. Subirats Maté, D. O. Pérez Trenard, & M. Calvo González. Barcelona: UOC.

Guitart Hormigo, I. (2019). Introducción al ciclo de vida de los datos. In L. Subirats Maté, D. O. Pérez Trenard, & M. Calvo González. Barcelona: UOC.

7. Enlaces

Código. A continuación, se indica el link al repositorio en GitHub:

https://github.com/cesalexguz/PRA2_Grupal_Cortez_Guzman

Dataset. A continuación, se indica el link al repositorio en Zenodo:

https://zenodo.org/record/5811072#.Yc5scWjMKUl

Video. A continuación, se indica el link al video explicativo:

https://drive.google.com/file/d/1byciERfvTahT7mTOdDGQ1Fdcfaqn2Qnb/view?usp=sharing