Universitat
Oberta
de Catalunya

Tipología y Ciclo de Vida del Dato

Práctica 2 – Web Scraping

Alumno: Rodrigo Baranda Castrillo

Índice de Contenidos

A tener en cuenta

La práctica se ha realizado en su totalidad de forma individual y el lenguaje utilizado ha sido Python. Al repositorio se subirán tanto el Notebook sobre el que se ha trabajado y ese mismo Notebook en formato *.py*. A lo largo de este documento se acompañarán las explicaciones con algunas gráficas/tablas que considero relevantes, sin embargo, el resto de la información está en los archivos mencionados.

Enlace de git: https://github.com/rbaranda10/CovidAnalysis

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

Para la realización de esta práctica se ha cambiado el dataset generado en la práctica 1. Inicialmente, se había realizado un Scraping de una serie de datos macroeconómicos extraídos de Expansión. Los datos que había descargado eran los siguientes:

	Año	País	Tasa Desempleo	Varianza Anual Desempleo	SMI €	Salario Medio
0	2005	España [+]	8,8%	-1,50	598,5 €	20.616€
1	2005	Reino Unido [+]	5,1%	0,40	1.134,7 €	44.361€
2	2005	Francia [+]	9,1%	0,20	1.217,9 €	30.521€
3	2005	Portugal [+]	8,8%	0,50	437,2 €	14.042€
4	2005	Estados Unidos [+]	4,9%	-0,50	655,4 €	30.252€
5	2005	Japón [+]	4,4%	-0,10	826,3 €	36.382€
6	2005	Australia [+]	5,1%	0,00	1.184,7 €	32.213€

Sin embargo, dado que sólo contaba con información de país, tasa de desempleo, varianza anual de desempleo, salario mínimo interprofesional y salario medio, he decidido escoger el un dataset de Kaggle.

El conjunto de datos contiene información acerca de la enfermedad *SARS-CoV-2* (en adelante, *COVID-19*). El enlace desde el que se ha descargado es el siguiente:

https://www.kaggle.com/shashwatwork/impact-of-covid19-pandemic-on-the-global-economy

Location	date	Total Cases	Total Deaths	Stringency	Population	GDP	HDI
Togo	2019-12-31	1.0	0.0	0.0	8278737.0	1429.813	0.503
Togo	2020-01-01	1.0	0.0	0.0	8278737.0	1429.813	0.503
Togo	2020-01-02	1.0	0.0	0.0	8278737.0	1429.813	0.503
Togo	2020-01-03	1.0	0.0	0.0	8278737.0	1429.813	0.503
Togo	2020-01-04	1.0	0.0	0.0	8278737.0	1429.813	0.503
...	...	...	...	...	...	...	...
Turks and Caicos Islands	2020-10-15	696.0	6.0	53.7	38718.0	NaN	NaN

A partir de esta información, se tratará de entender en qué partes del mundo el *COVID-19* ha tenido un mayor impacto, ver la evolución en número de casos y muertes a lo largo del tiempo, ver si los países que han tomado medidas más estrictas han obtenido mejores resultados que los que no, plantear diferencias llamativas entre países para ver si los datos pueden ser verdaderamente fiables, etc.

El dataset inicial contaba con 13 columnas, sin embargo, 4 de ellas no contenían información relevante, de hecho, tenían datos erróneos, como los que se muestran a continuación:

Unnamed: 9	Unnamed: 10	Unnamed: 11	Unnamed: 12	Unnamed: 13
#NUM!	#NUM!	#NUM!	17.477233	7.497754494
#NUM!	#NUM!	#NUM!	17.477233	7.497754494
#NUM!	#NUM!	#NUM!	17.477233	7.497754494
#NUM!	#NUM!	#NUM!	17.477233	7.497754494
#NUM!	#NUM!	#NUM!	17.477233	7.497754494

Estas columnas se han desechado por no ser relevantes en ningún posible análisis.

A continuación, se irán explicando los atributos de que consta el conjunto de datos y las operaciones/transformaciones que se han realizado sobre ellos con el objetivo de obtener unos datos más útiles a la hora de estudiarlos (*Nota: se irán añadiendo capturas de pantalla con trozos de código/resultados con el objetivo de que se entiendan más el conjunto de datos y las transformaciones*):

1. Location. Este campo es el país del que tenemos información. Disponemos de un total de 210 países en nuestro conjunto de datos.

```
In [14]: print("De un total de", total_countries, "países, sólo hay" , total_countries-countries_with_null_deaths, "para los que con
```

```
De un total de 210 países, sólo hay 22 para los que contamos con datos en todo el periodo.
```

   A partir de esta variable se han extraído la latitud y longitud de cada país (un punto central del país) para, a partir de ahí, realizar un mapa que nos permitiera representar cómo ha evolucionado la pandemia con el tiempo.

2. Date. Este campo es la fecha para la que se tienen datos de ese país. El primer día para el que tenemos datos es el 31/12/2019 y el último es el 19/10/2020. Éste será nuestro periodo de análisis.

```
In [108]: print(df['date'].min())
          print(df['date'].max())
```

```
2019-12-31
2020-10-19
```

   Cabe resaltar que no de todos los países contamos con datos diarios (i.e., España no proporcionaba datos los fines de semana). De hecho, del total de 210 países de que disponemos, sólo 22 tiene datos en todo el periodo temporal, los otros 188, tienen algún gap. Más adelante se mostrará que se ha realizado un análisis de la evolución de los muertos por la enfermedad a lo largo del tiempo, para lo cual se necesitaba contar con datos de todos los días. Lo que se ha hecho, por tanto, es generar registros para aquellos días de los que no se tuviera información, completando ese registro con los datos que hubiera del país en el día anterior (i.e., rellenar los datos del sábado y domingo con los que se tuvieran el viernes).

3. Total Cases. Esta variable contiene el acumulado de casos acumulados por un país a una fecha dada.
4. Total Deaths. Esta variable contiene el acumulado de muertos por *COVID-19* en un país a una fecha dada.
5. Stringency. Ésta variable es una métrica desarrollada por la Universidad de Oxford que tiene por objeto cuantificar cuánto de duras han sido las medidas tomadas para combatir la enfermedad. Por lo tanto, a mayor número de medidas tomadas: cierre de escuelas, transporte público, confinamiento domiciliario…, mayor será este índice. Éste índice va variando en función de las medidas que haya en ese momento. Referencia: https://www.civilsdaily.com/news/what-is-stringency-index/
6. Population. Ésta variable indica la población total del país. Se ha utilizado para, a partir de ella, calcular el número de casos y de muertes por cada 100*k* habitantes. Esto será importante a la hora de analizar los datos porque, de este modo, todos los países estarán en la misma escala.
7. GDP (*Gross Domestic Product*) – Producto Interior Bruto (PIB). Ésta variable indica el producto interior bruto de cada país.
8. HDI (Human Development Index). Éste es un índice que mide cuánto de desarrollado está un país. Es un índice que no busca representar sólo cuánto de desarrollado económicamente se encuentra un país, sino también en lo siguiente: "*a long and healthy life, being knowledgeable and have a decent standard of living*."
   En este punto se tratará de plantear un modelo de regresión univariante entre el PIB y el HDI para ver cuánto de correlados están y si se puede entender el uno a partir del otro.

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

Como ya se ha indicado en el punto anterior, los datos a analizar se encuentran en un único fichero CSV, por lo que no habrá que realizar tareas de integración de datos. Por otro lado, las variables seleccionadas serán las ya mencionadas, habiendo sido despreciadas cuatro variables que aparecían inicialmente en el dataset pero que contenían datos basura.

Pregunta 3. Limpieza de los datos.

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

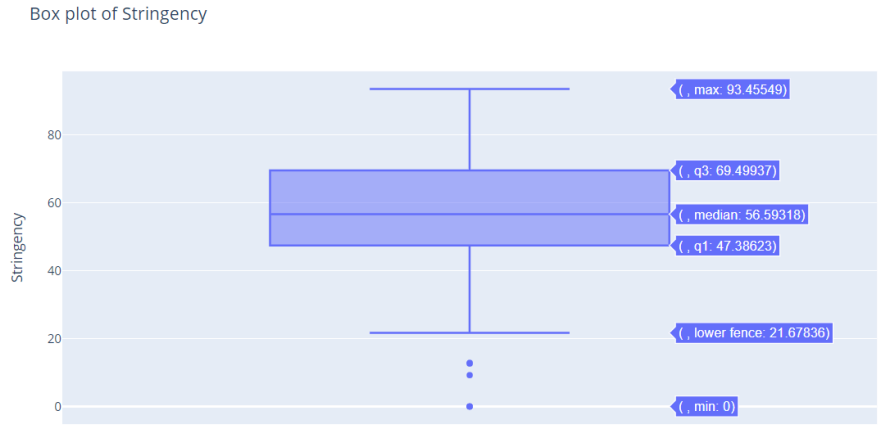
En este *dataset* se encuentran dos problemas relativos al problema de la existencia de nulos.

- Datos de casos y muertes en nulo porque se incluyen registros antes de que se hayan empezado a registrar datos de la enfermedad. En este caso, lo que se ha hecho es poner estos registros a 0 (detectar que son los primeros de ese país y ponerlos a 0).
- Datos de casos y muertes vacíos en mitad de los datos de un país. En estos casos lo que se ha hecho es completar el registro con el dato del anterior día del que hubiera dato disponible. También se ha valorado poner un valor medio entre el último día del que hubiera datos y el siguiente.

3.2. Identificación y tratamiento de valores extremos.

Respecto a valores extremos, se han representado una serie de *Boxplots*, dado que son una herramienta que permite detectar con facilidad si hay algún *outlier* que resulte especialmente llamativo.

Para el caso del índice *Stringency*, se han detectado 3 países que no tomaron ninguna medida. Además, estos países no han reportado HDI ni GDP. Se tendrán en cuenta a la hora de ver la evolución de muertes, pero, no se tendrán en cuenta al analizar la posible relación entre el HDI y el GDP.

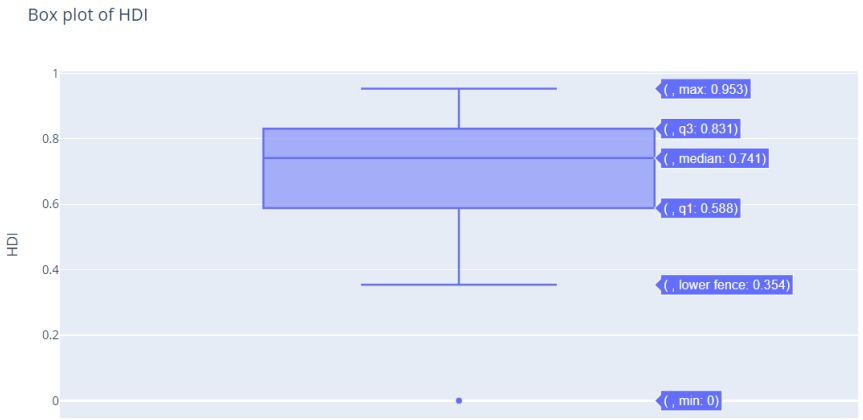



```
In [238]: search_outliers[search_outliers['Stringency']==0]
```

Out[238]:

	Location	Total Deaths	Total Cases	Stringency	GDP	HDI
95	Isle of Man	20.234742	304.859155	0.0	NaN	NaN
100	Jersey	25.827103	312.042056	0.0	NaN	NaN
113	Liechtenstein	0.883929	89.170404	0.0	NaN	0.916

En el caso del análisis de valores del HDI, se ha visto que Kosovo presenta un valor de 0. Sin embargo, dado que este valor puede ser correcto y no se ha encontrado evidencia que lo contradiga, se mantendrá.




```
search_outliers[search_outliers['HDI'] == 0]
```

	Location	Total Deaths	Total Cases	Stringency	GDP	HDI
104	Kosovo	216.231818	6117.909091	71.329	9795.834	0.0

El final del proceso de limpieza/preparación de datos es generar nuevas filas para aquellos registros para los que no tengamos fecha. Como se mencionaba anteriormente, estas nuevas filas se han generado con el dato que existiera en el registro previo o, en caso de que fuera el primer registro para ese país, se rellenarían a 0.

Finalizado el proceso de transformación de datos, se ha pasado de tener un *dataframe* con 50.418 filas a tener uno de 61.740.

iso_code	Location	date	Total Cases	Total Deaths	Stringency	Population	GDP	HDI	Continent_Country	Country	Continent	Lat-Long	Deaths 100k Hab	Cases 100k Hab
TGO	Togo	2019-12-31	1.0	0.0	0.0	8278737.0	1429.813	0.503	(TG, AF)	TG	AF	(8.7800265, 1.0199765)	0.0	0.012079
TGO	Togo	2020-01-01	1.0	0.0	0.0	8278737.0	1429.813	0.503	(TG, AF)	TG	AF	(8.7800265, 1.0199765)	0.0	0.012079
TGO	Togo	2020-01-02	1.0	0.0	0.0	8278737.0	1429.813	0.503	(TG, AF)	TG	AF	(8.7800265, 1.0199765)	0.0	0.012079

Pregunta 4. Análisis de datos.

A lo largo de este apartado se analizarán los siguientes puntos:

- Evolución de Muertes por 100k habitantes con el tiempo a nivel mundial. En este punto se presentará un mapamundi dinámico sobre el que se podrá ver cómo han ido evolucionando los muertos por 100k habitantes con el tiempo. El objetivo de este punto es ver por qué zonas se distribuye más el virus y de una forma más mortal y, también, plantear si esto puede tener sentido o no.
- Evolución de Muertes por continente. En este punto se mostrarán una serie de líneas de tendencia que permitirán ver de una forma clara cómo ha evolucionado la pandemia con el tiempo en cada continente. Al igual que en el punto anterior, será interesante ver si esto puede tener sentido con los datos de población y sanitarios de cada continente.
- Evolución de Muertos por cada 100k habitantes por país. De este modo se podrá hacer "*zoom*" o poner el foco en países concretos cuyos datos hayan resultado llamativos.
- Análisis de Muertos por 100k habitantes respecto a la severidad de las medidas tomadas en el país y el índice de desarrollo del país. En este punto se pintará una gráfica tridimensional con el objetivo de detectar cuál es la principal causa que limita la mortalidad del virus.
- Regresión Lineal entre el GDP y el HDI para ver si existe una fuerte relación de dependencia lineal entre ambas variables o si, por el contrario, sí que tienen cierta independencia (lo cual indicaría que la variable HDI, que busca medir la calidad de vida independientemente del GDP, tiene sentido).
- Análisis de Varianza (Homocedasticidad) entre los distintos países e intersemanales de un mismo país para ver si tendría sentido plantear algún modelo que se ajuste a series temporales (verificar si son o no estacionarios).

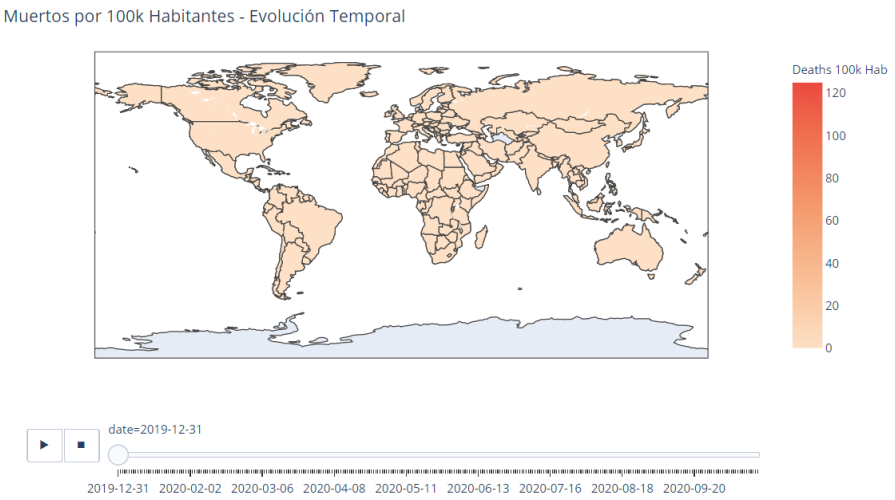
4.1. Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).

En este punto se incluirán aquellos análisis que hayan estado marcados por un alto componente de visualización y que, por lo tanto, no haya tests/contrastes, etc., que respalden la información presentada. Ese tipo de análisis se presentarán en los puntos 4.2 y 4.3.

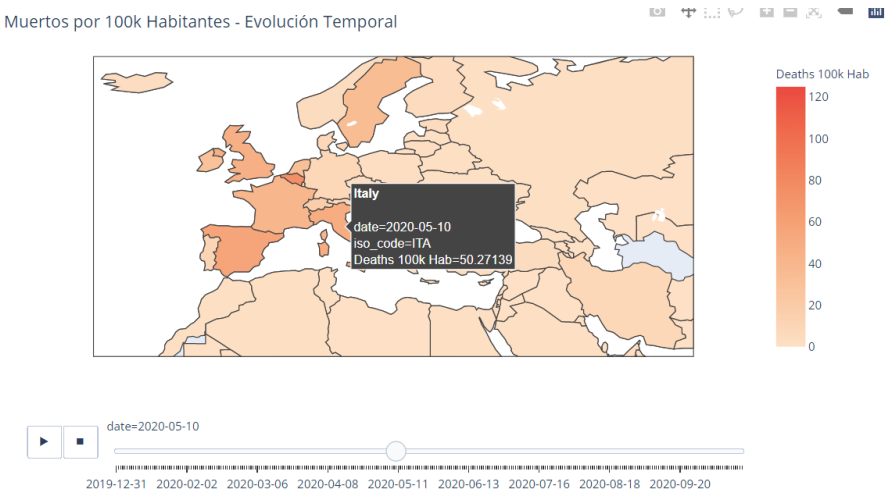
Evolución de Muertes por 100k habitantes con el tiempo a nivel mundial.

Dado que el condicionante principal del *dataset* escogido es que es una serie temporal, los análisis realizados tendrán que ir siempre ligados a la variable tiempo. En primer lugar, se ha preparado un mapa sobre el que se puede avanzar y retroceder en el periodo temporal y ver el estado de muertes por cada 100k habitantes en todo el mundo.

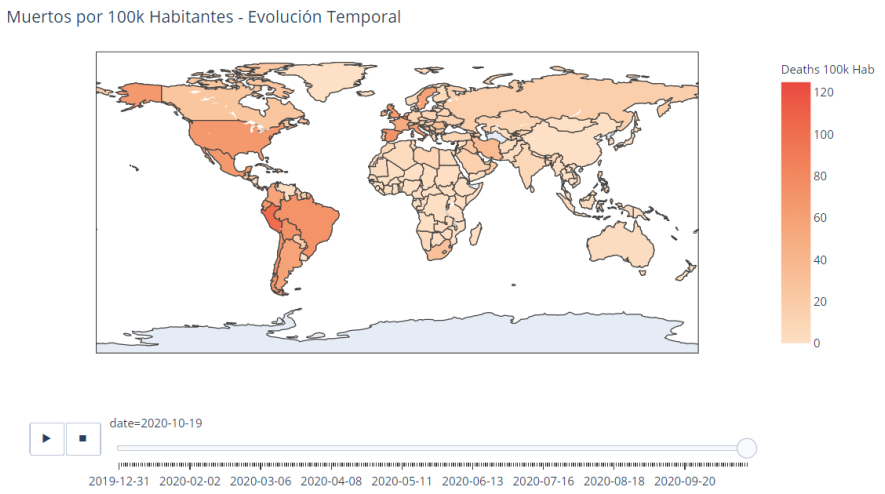
Punto inicial:



Sobre el mapa se puede ir realizando zoom para ver, por ejemplo, Europa, cómo ha ido evolucionando:



La situación final sería la siguiente:



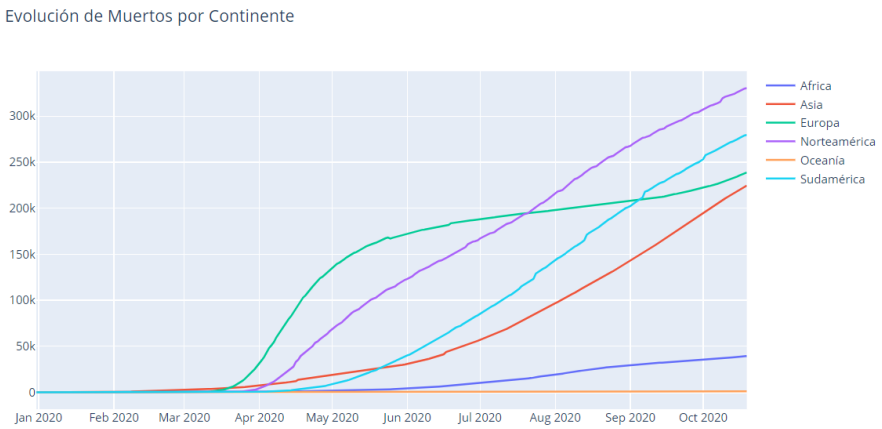
Viendo el resultado final de este mapa, por ejemplo, se puede ver que, según indican los datos, las zonas más afectadas por el COVID-19 serían Estados Unidos, Sudamérica y el sur de Europa. Resulta llamativo ver cómo, en zonas en las que las capacidades sanitarias son sensiblemente peores que las de Europa, como pueden ser África; o en el foco inicial del virus, China (0.329 muertes por 100k habitantes VS 72.23

muertes por 100k habitantes de España), los datos de muertes son muy inferiores a los de los países más golpeados.

En este punto cabe cuestionarse la capacidad de registro de datos o voluntad de compartir libremente la información por cada uno de los países porque resulta llamativo ver los mencionados números de China o de África en general.

Evolución de Muertes por continente

Realizando el análisis de muertos por continente, se puede obtener una información parecida a la ya contada, pero seguramente más entendible visualmente:

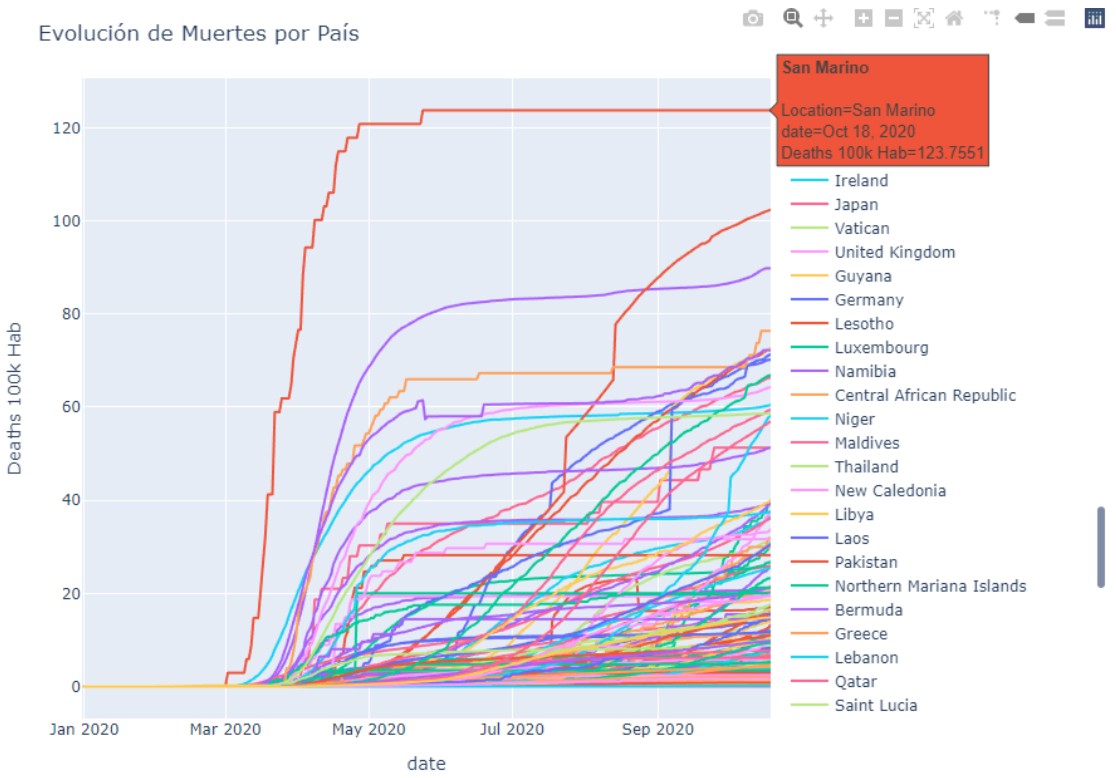


Aquí se puede ver que, por ejemplo, en África el número de muertos reportado era aproximadamente 6 veces menor al que tenía Europa y 8 veces menor que el de Norteamérica. Por otro lado, resalta el trabajo realizado por Oceanía a la hora de contener la pandemia.

Una de las dificultades a la hora de analizar estos datos y presentar conclusiones es que, aparte del mencionado problema de calidad o veracidad del dato, es que hay intangibles como por ejemplo el punto geográfico en que está situado Oceanía que son difíciles de ponderar con los datos disponibles.

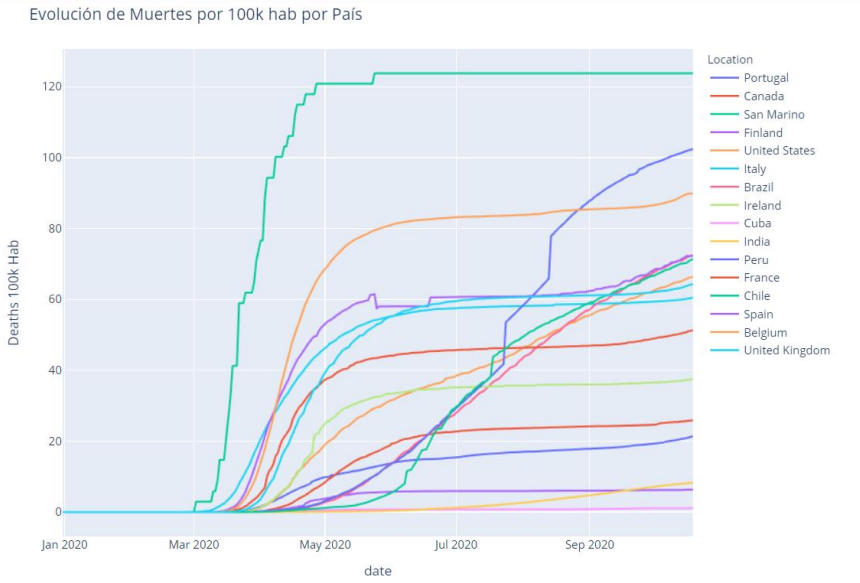
Evolución de Muertes por 100k habitantes con el tiempo a nivel mundial.

La información mostrada en forma de mapa y, posteriormente agrupada por continente, también se ha extraído por países.

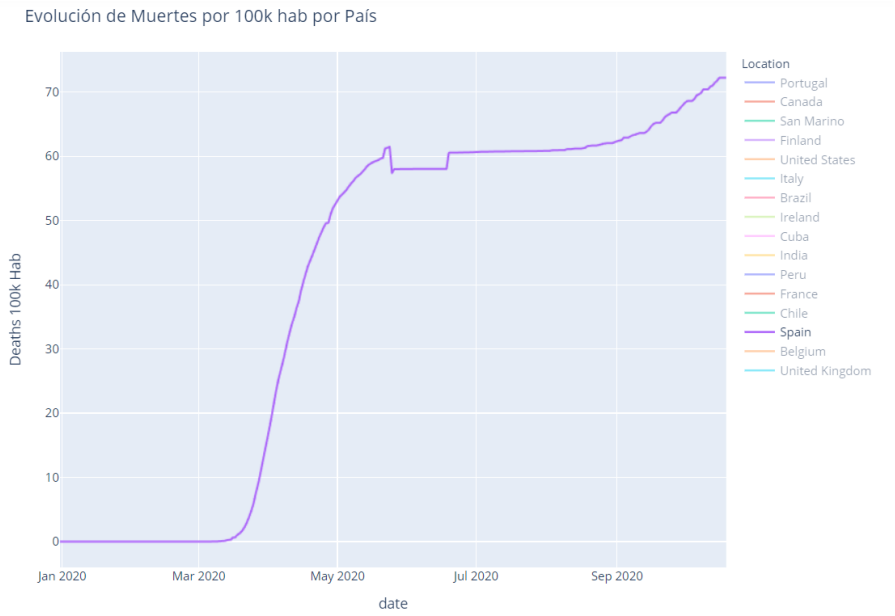


De la gráfica mostrada, resultan llamativos dos países: el elevadísimo número de muertos por 100k habitantes de San Marino que, hasta el 18/10/2020 era el primero y, por otro lado, los datos de Burundi o Solomon Islands con prácticamente 0 muertes.

Dado que mostrar todos los países genera una gráfica con mucho ruido y poco entendible, se muestra la gráfica con los datos de unos pocos:

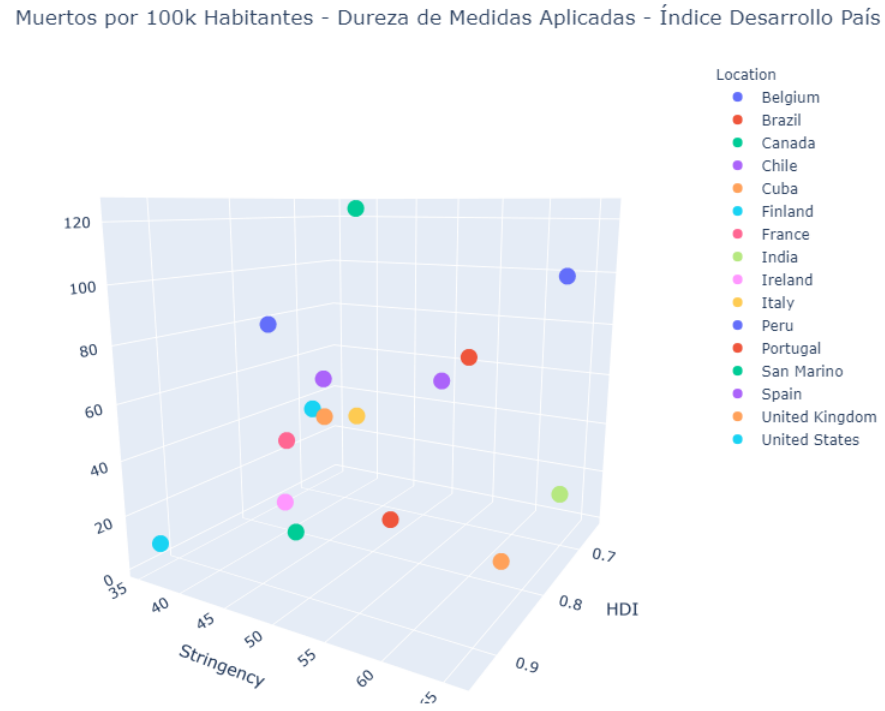


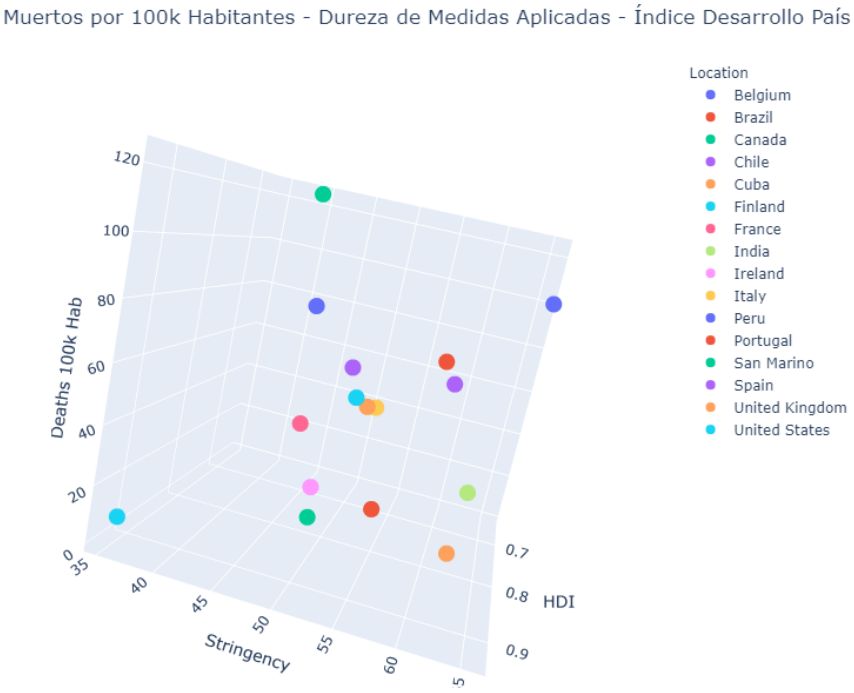
En esta gráfica, por ejemplo, se puede ver que España es uno de los países punteros en la estadística de muertos por cada 100k habitantes. Resulta llamativo, por ejemplo, que se puede ver en la gráfica de España que incluso hay un descenso de las muertes. Esto resalta la cuestionable calidad del dato de que disponemos: dependemos de lo que registre y quiera reportar cada país.



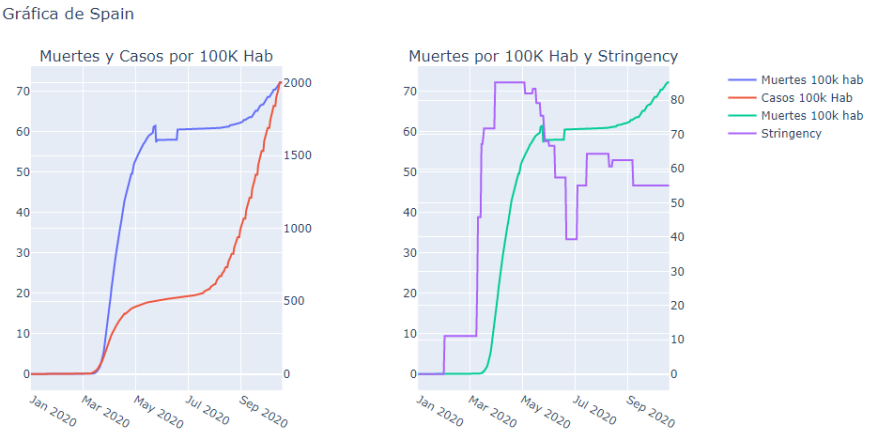
Análisis de Muertos por 100k habitantes respecto a la severidad de las medidas tomadas en el país y el índice de desarrollo del país

En las dos siguientes gráficas mostradas, se puede ver que no parece haber una relación directa entre dureza de las medidas aplicadas (*Stringency*) y números de muertos por 100k habitantes. De hecho, la correlación entre estas dos variables (Stringency-Muertes) es de 0.08, la cual está demasiado próxima a 0 para considerarla significativa. En este sentido, se podría pensar que quizá los países más golpeados fueran aquellos que tienen un HDI más bajo por, seguramente, tener más dificultades para alcanzar ciertas necesidades sanitarias. Sin embargo, la correlación entre HDI-Muertes es de 0.20. Este valor, si bien está lejos de ser especialmente elevado, puede proporcionar dos lecturas: (i) existe un mayor volumen de contagios y muertes en países desarrollados porque es más difícil limitarles sus libertades personales o (ii) los países con menor HDI no reportan los datos consistentemente, de forma o no premeditada, de tal modo que su situación parece mejor de la que es.





El último punto que se ha representado ha sido el relativo a la relación entre Muertos-Restricciones. Por ejemplo, en esta gráfica de España se ve que, una vez se alcanzó un máximo local de muertes, se decidió reducir las restricciones; en el mes siguiente, la pendiente de la curva de muertes ya volvía a repuntar.



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

Análisis de Homocedasticidad.

Supongamos que quisiéramos utilizar algún modelo de series temporales: Auto Regresivo (AR), Media Móvil (MA), Auto Regresivo Media Móvil (ARMA)…para tratar de predecir la evolución de la pandemia con el tiempo. El primer estos modelos de series temporales suelen partir de la base de que los datos con los que se trabaje deben ser estacionarios: las variables estadísticas más relevantes permanecen invariantes con el tiempo (media, desviación estándar, covarianza). Por ejemplo, se analizará la varianza entre distintos países, para ver si éstos tienen homocedasticidad o no.

Para ello se utilizará el Test de Levene, tanto para los Casos por país como para las Muertes:

$$Levene's Test = \begin{cases} H_0 : Homocedasticidad \\ H_1 : Heterocedasticidad \end{cases}$$

```
# Método que devuelve el resultado del test de Levene para el análisis de la varianza entre los diferentes años de nuestro

spain_cases = df[df['Location'] == 'Spain']['Total Cases']
france_cases = df[df['Location'] == 'France']['Total Cases']
portugal_cases = df[df['Location'] == 'Portugal']['Total Cases']
italy_cases = df[df['Location'] == 'Italy']['Total Cases']
finland_cases = df[df['Location'] == 'Finland']['Total Cases']

scipy.stats.levene(spain_cases, france_cases, portugal_cases, italy_cases, finland_cases)
```

```
LeveneResult(statistic=125.83472705073596, pvalue=2.1027160009018174e-92)
```

```
# Método que devuelve el resultado del test de Levene para el análisis de la varianza entre los diferentes años de nuestro

spain_cases = df[df['Location'] == 'Spain']['Total Deaths']
france_cases = df[df['Location'] == 'France']['Total Deaths']
portugal_cases = df[df['Location'] == 'Portugal']['Total Deaths']
italy_cases = df[df['Location'] == 'Italy']['Total Deaths']
finland_cases = df[df['Location'] == 'Finland']['Total Deaths']

scipy.stats.levene(spain_cases, france_cases, portugal_cases, italy_cases, finland_cases)
```

```
LeveneResult(statistic=117.82684399984105, pvalue=3.3185781818236316e-87)
```

En ambos casos se obtiene un p-valor de 0 prácticamente (hay magnitudes del orden de -92), lo cual rechaza la hipótesis nula de que hay homogeneidad de varianza en los conjuntos.

Sin embargo, es cierto que, dado que cada país tomará unas medidas muy diferentes ante el COVID-19, no es creíble pretender realizar un modelo que funcione a nivel global. Para ver si podría tener sentido realizarlo a nivel nacional, se ha estudiado la homocedasticidad para los datos de España de entre el 01/05/2020 y el 01/06/2020. Nuevamente obtenemos un p-valor inferior a 0.05, lo cual indica que no hay homocedasticidad ni entre los datos de un mismo mes.

```
spain = df.copy()
```

```
spain['date'] = pd.to_datetime(spain["date"])
spain = spain[spain['Location'] == 'Spain']
```

```
mayo_spain = spain[(spain['date'] >= '2020-05-01')  & (spain['date'] <= '2020-06-01')]

primera_semana = mayo_spain['Total Deaths'][0:8]
segunda_semana = mayo_spain['Total Deaths'][8:16]
tercera_semana = mayo_spain['Total Deaths'][16:24]
cuarta_semana = mayo_spain['Total Deaths'][24:32]
```

```
scipy.stats.levene(primera_semana, segunda_semana, tercera_semana, cuarta_semana)
```

```
LeveneResult(statistic=3.9997416719979055, pvalue=0.01727932236350042)
```

Además del análisis de homocedasticidad, se ha aplicado el test de Dickey Fuller en los tres casos, para ver si los datos eran estacionarios (al no haber homocedasticidad, ya sabíamos que no, pero se ha verificado).

$$DickeyFuller's Test = \begin{cases} H_0 : NoEstacionaria \\ H_1 : Estacionaria \end{cases}$$

El p-valor obtenido en cada caso ha sido de:

- Total de casos por países: p-value = 0.856707
- Total de muertes por países: p-value = 0.427399
- Total de muertos en el mes de mayo en España: p-value = 0.192414

De este punto, lo que se concluye es la enorme dificultad a la hora de tratar de implementar un modelo que se ajuste de una forma óptima a los datos y permita realizar predicciones de Casos/Muertes en el futuro. Esta dificultad es debida a la propia naturaleza de la enfermedad y las medidas reactivas que obliga a imponer frente a ella y, por otro lado, la ya mencionada cuestionable veracidad de los datos.

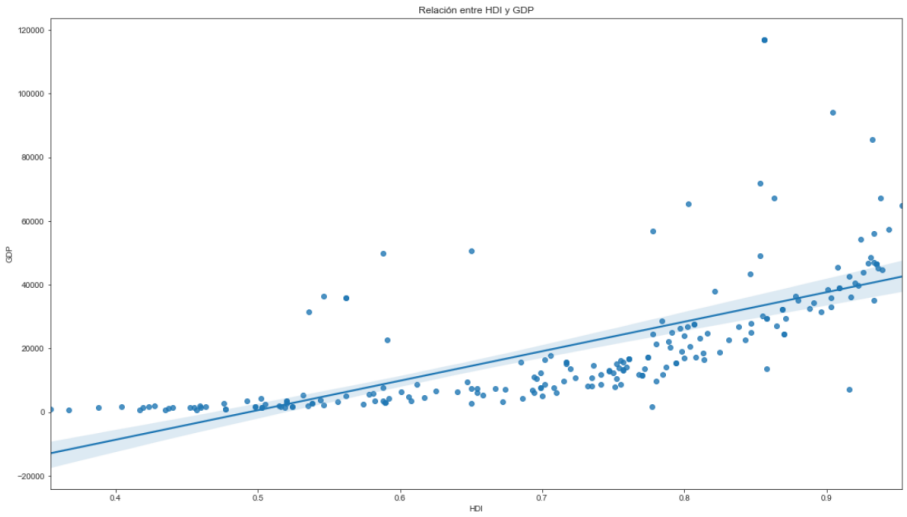
4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.

El último análisis realizado ha sido ver si existe una relación de dependencia entre la variable GDP (Gross Domestic Product) y la variable HDI (Human Development Index). A priori, sabemos por lo que se ha mencionado anteriormente de lo que es la variable HDI que si bien no hay una relación de causalidad entre ambas, sí que deberían están correladas.

	Total Cases	Total Deaths	Stringency	Population	GDP	HDI	Deaths 100k Hab	Cases 100k Hab
Total Cases	1.000000	0.911620	0.092107	0.294433	0.051201	0.077945	0.290247	0.277336
Total Deaths	0.911620	1.000000	0.102633	0.241778	0.081342	0.124177	0.431604	0.279414
Stringency	0.092107	0.102633	1.000000	0.066595	-0.043232	-0.076592	0.090933	0.132230
Population	0.294433	0.241778	0.066595	1.000000	-0.062222	-0.010631	-0.001806	-0.029030
GDP	0.051201	0.081342	-0.043232	-0.062222	1.000000	0.646490	0.202349	0.285919
HDI	0.077945	0.124177	-0.076592	-0.010631	0.646490	1.000000	0.231342	0.209031
Deaths 100k Hab	0.290247	0.431604	0.090933	-0.001806	0.202349	0.231342	1.000000	0.601016
Cases 100k Hab	0.277336	0.279414	0.132230	-0.029030	0.285919	0.209031	0.601016	1.000000

Viendo la tabla de correlaciones, ya podemos ver que la correlación más fuerte que existe entre las variables de nuestro dataset (después de la de Casos-Muertes), es la de GDP-HDI.

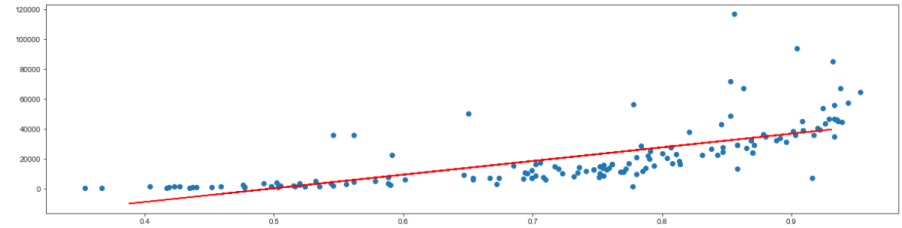
Para realizar esta regresión ya se ha eliminado el valor que veíamos previamente de Kosovo, con un HDI de 0.



Tras realizar una regresión lineal entre las variables HDI y GDP, se ha obtenido una recta de la forma:

$$y = -45164 + 91308X$$

Esta ecuación significaría que, para un país sin ningún tipo de desarrollo humano, el producto interior bruto por persona sería negativo, lo cual es imposible conceptualmente. El error absoluto cometido es de 11.506 puntos, que, si tenemos en cuenta que la mayoría de valores oscilan entre 20.000 y 30.000, es una desviación bastante grande.



```
Mean Absolute Error:  11506.875027526485
Mean Squared Error:  407514406.4582078
```

En esta recta se puede ver que la mayoría de errores cometidos es porque la recta está desplazada hacia arriba, fundamentalmente debido a los países con un elevado GDP en proporción al HDI (es decir, países ricos económicamente, pero no como humanos, según este índice). Una propuesta para mejorar este modelo sería analizar la distribución del GDP y eliminar aquellos valores que estuvieran más lejos de 3 desviaciones estándar, por ejemplo. De este modo, los errores respecto a estos casos aumentarían, pero en términos generales el rendimiento mejoraría.

Pregunta 5. Representación de los resultados a partir de tablas y gráficas.

A lo largo del ejercicio se han ido añadiendo distintas imágenes relativas a los análisis realizados, conque considero redundante volver a incluir la misma información.

Pregunta 6. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?

A lo largo de esta práctica se ha buscado realizar un análisis de la evolución del COVID-19 a nivel mundial. Se han llegado a conclusiones que pueden resultar llamativas, tales como que el volumen de muertos en Norte América a fecha 19/10/2020 fuera 8 veces superior a la de África en la misma fecha, o que no existe una fuerte correlación entre Restricciones y Número de muertos ni Desarrollo del País y Número de muertos, cuando los hechos sí que parecen respaldar, al menos, lo primero (la limitación de libertades de movimiento de las personas ha reducido la pendiente de la curva).

Por otro lado, se ha analizado la tipología de datos de que disponíamos, para ver cuánto de fáciles/manipulables eran, en caso de que se quisiera aplicar un modelo predictivo sobre ellos. El resultado ya lo he mencionado anteriormente: ni los datos entre países ni los de un mismo país intersemanales son estacionarios. Esto significa que las principales medidas estadísticas del dataset van variando mucho en poco tiempo, lo cual prácticamente imposibilita la aplicación de modelizaciones o machine learning para resolver este problema en la actualidad.

Por lo tanto, una de las conclusiones aplicada al campo de la Ciencia del Dato es la dificultad de tener información veraz y accesible. Este suele ser un punto a tener en cuenta en cualquier empresa, incrementando la dificultad en las empresas de mayor tamaño y que no hayan sido tradicionalmente *data-driven*; imaginemos esto mismo a nivel mundial, cuando, además, los países no tenían obligatoriedad alguna de facilitar la información. El resultado de esto ya lo hemos visto, países en que desciende el número de muertos de un día para otro (España), etc., es decir → Poca calidad de dato.

Considero que, si bien es cierto que los resultados a nivel científico/estadístico no son potentes, se ha tratado de realizar un trayecto por todos los puntos relevantes del conjunto de datos entendiendo sus posibles problemas, resolviéndolos cuando ha sido posible, y presentándolos de una forma sencilla e interactiva (las gráficas del notebook/html) para un tercero.

Pregunta 7. Código: Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos. Si lo preferís, también podéis trabajar en Python.

https://github.com/rbaranda10/CovidAnalysis