

## Práctica de Laboratorio: Análisis Exploratorio de Datos - Data Wrangling

Docente: [Ana María Cuadros Valdivia](#)

Para realizar el Análisis Exploratorio de datos, lo primero que deberíamos hacer es intentar responder a las siguientes preguntas (data wrangling):

### Paso 1: Analiza el comportamiento de tus datos.

- Un registro es una entidad, describa que representa un registro

	iso_code	continent	location	date	total_cases	new_cases	new_cases_smoothed	total_deaths	new_deaths	new_deaths_smoothed	...	male_smokers	handwashing_facilities	hospital_beds_per_million
0	AFG	Asia	Afghanistan	2020-01-05	0.0	0.0	Nan	0.0	0.0	Nan	...	Nan	37.75	
1	AFG	Asia	Afghanistan	2020-01-06	0.0	0.0	Nan	0.0	0.0	Nan	...	Nan	37.75	
2	AFG	Asia	Afghanistan	2020-01-07	0.0	0.0	Nan	0.0	0.0	Nan	...	Nan	37.75	

3 rows × 67 columns

Cada registro en el dataset representa la situación epidemiológica de COVID-19 en un país determinado (location) en una fecha específica (date). Por tanto, una fila contiene los datos diarios (o acumulados) como casos nuevos, muertes, índice de positividad, entre otros.

- ¿Cuántos registros hay?

```

  ✓  play  print(f"Número total de registros: {len(df)}")
  ↵  Número total de registros: 429435

```

429,435 registros El dataset contiene miles de filas. Esto sugiere una granularidad diaria y global, ya que se recopilan datos de muchos países a lo largo del tiempo.

- ¿Son demasiado pocos?

La cantidad de registros es considerable pero no excesiva para ser procesada

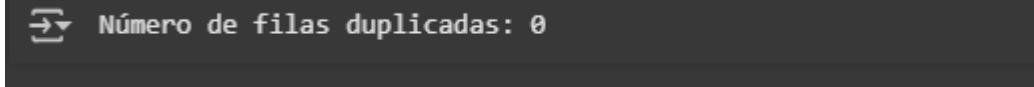
- ¿Son muchos y no tenemos Capacidad (CPU+RAM) suficiente para procesarlo?

Dado que el número total de registros es inferior a medio millón y que la estructura de los datos es tabular y limpia, este volumen es manejable en entornos de procesamiento estándar como Google Colab, Jupyter Notebooks, o servidores backend escritos en

Python. No es necesario recurrir a tecnologías de Big Data, lo que simplifica el desarrollo del backend del framework.

- ¿Hay datos duplicados?

No hay registros duplicados (`duplicated().sum() = 0`).

 Número de filas duplicadas: 0

Si se encuentran filas duplicadas, podrían generar sesgos en el análisis estadístico. Afortunadamente, la mayoría de los datasets de OWID están limpios y no suelen tener duplicados.

- ¿Qué datos son discretos y cuáles continuos?

```
0
iso_code          object
continent         object
location          object
date              object
total_cases       float64
...
population        int64
excess_mortality_cumulative_absolute  float64
excess_mortality_cumulative           float64
excess_mortality                 float64
excess_mortality_cumulative_per_million float64
67 rows × 1 columns
dtype: object
```

Las variables como new\_cases, new\_deaths, icu\_patients, etc., son **variables continuas**. En cambio, location y date son **discretas o categóricas**. Las primeras permiten cálculos numéricos, mientras que las segundas se usan para segmentación.

- Muchas veces sirve obtener el tipo de datos: texto, int, double, float
- ¿Cuáles son los tipos de datos de cada columna?

```
0
iso_code          object
continent         object
location          object
date              object
total_cases       float64
...
population        int64
excess_mortality_cumulative_absolute  float64
excess_mortality_cumulative           float64
excess_mortality                 float64
excess_mortality_cumulative_per_million float64
67 rows × 1 columns
dtype: object
```

El tipo de datos es fundamental para elegir los métodos estadísticos. La mayoría son float64, pero location y date son object, por lo que date podría convertirse a tipo fecha (datetime64) para operaciones temporales.

- ¿Entre qué rangos están los datos de cada columna?, valores únicos, min, max

	iso_code	continent	location	date	total_cases	new_cases	new_cases_smoothed	total_deaths	new_deaths	new_deaths_smoothed	...	male_smokers	handwashing_facilities	hospital_beds_per_thousand	life_expectancy	human_development_index	population	excess_mortality_cumulative_absolute	excess_mortality_cumulative
count	429435	402910	429435	429435	4.118040e+05	4.101580e+05	4.088290e+05	4.118040e+05	410008.000000	409378.000000	...	243817.000000	161741.000000	290689.000000	380299.000000	319127.000000	4.264350e+05	1.341100e+04	
unique	255	6	255	1688	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
top	OVID_HIC	Africa	High-income countries	2021-10-22	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
freq	3028	95419	3028	261	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
mean	NaN	NaN	NaN	NaN	8.365232e+06	8.017300e+03	8.041020e+03	8.125957e+04	71.852138	72.000828	...	33.097758	50.046380	3.106895	73.702088	0.722178	1.520339e+08	5.004765e+04	
std	NaN	NaN	NaN	NaN	4.477582e+07	2.295649e+05	8.661511e+04	4.41901e+05	1388.322990	513.055665	...	13.803952	31.905238	2.549108	7.387914	0.149237	6.975498e+08	1.558691e+05	
min	NaN	NaN	NaN	NaN	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000	0.000000	...	7.000000	1.190000	0.100000	53.280000	0.300000	4.700000e-01	-3.772610e+04	
25%	NaN	NaN	NaN	NaN	6.280700e+03	0.000000e+00	0.000000e+00	0.000000e+00	0.000000	0.000000	...	22.800000	20.880000	1.300000	69.500000	0.600000	5.237780e+05	1.785000e+02	
50%	NaN	NaN	NaN	NaN	8.365300e+04	0.000000e+00	0.000000e+00	1.200000e+01	7.990000e+02	0.000000	...	33.100000	49.540000	2.500000	75.050000	0.740000	6.398393e+08	6.815200e+03	
75%	NaN	NaN	NaN	NaN	7.582270e+05	0.000000e+00	3.132000e+02	6.574000e+03	0.000000	3.140000	...	41.500000	82.500000	4.210000	79.460000	0.830000	3.26952e+07	3.912804e+04	
max	NaN	NaN	NaN	NaN	7.788688e+08	4.423823e+07	6.319491e+08	7.057132e+08	103719.000000	14817.000000	...	78.100000	100.000000	13.800000	88.750000	0.980000	7.975105e+09	1.349770e+04	

Se puede observar el valor mínimo, máximo y la cantidad de valores únicos por columna. Esto ayuda a detectar outliers y validar el rango esperado de cada variable, por ejemplo, si alguna tiene valores negativos donde no debería.

- ¿Todos los datos están en su formato adecuado?

```
df['date'] = pd.to_datetime(df['date'])
df.dtypes
```

	0
iso_code	object
continent	object
location	object
date	datetime64[ns]
total_cases	float64
...	...
population	int64
excess_mortality_cumulative_absolute	float64
excess_mortality_cumulative	float64
excess_mortality	float64
excess_mortality_cumulative_per_million	float64

67 rows × 1 columns  
dtype: object

Se debe convertir la columna date al formato adecuado de fecha para facilitar operaciones temporales como filtrado por año o mes.

- Los datos tienen diferentes unidades de medida?

Sí, las unidades varían: por ejemplo, new\_cases es número de personas, stringency\_index es un porcentaje del 0 al 100, y positive\_rate va de 0 a 1. Esto obliga a tener cuidado al comparar variables.

- Cuáles son los datos categóricos, ¿hay necesidad de convertirlos en numéricos?  
location es una variable categórica con múltiples países
- ¿Qué representa un registro?
  - Describe qué representa cada fila.
  - Si es una data etiquetada, como interpretas la información de las clases?

- ¿Hay niveles de granularidad de los datos? Por ejemplo, datos a nivel país, región, ciudad. Años, meses, días, horas, minutos, etc.

Cada fila representa una combinación país-fecha, lo que indica que hay granularidad diaria y por país. Esto es útil para hacer análisis temporales o comparativos regionales.

- ¿Están todas las filas completas o tenemos campos con valores nulos?
  - En caso que haya demasiados nulos: ¿Queda el resto de información útil?. Se debe agregar o combinar sus datos.

df.isnull().sum()	
iso_code	0
continent	26525
location	0
date	0
total_cases	17631
...	...
population	0
excess_mortality_cumulative_absolute	416024
excess_mortality_cumulative	416024
excess_mortality	416024
excess_mortality_cumulative_per_million	416024
67 rows × 1 columns	
dtype: int64	

Varias columnas como icu\_patients, reproduction\_rate y total\_tests pueden tener valores nulos. Si una columna tiene demasiados nulos, podría descartarse o imputarse si tiene importancia estratégica

¿Siguen alguna distribución?  
Usa describe() y analiza los valores.

df.describe()																
	date	total_cases	new_cases	new_cases_smoothed	total_deaths	new_deaths	new_deaths_smoothed	total_cases_per_million	new_cases_per_million	new_cases_smoothed_per_million	...	male_smokers	handwashing_facilities	hospital_beds_per_thousand	...	...
count	429435	4.118946e+05	4.101596e+05	4.089290e+05	4.118048e+05	410608.000000	409378.000000	411934.000000	410159.000000	409529.000000	...	243817.000000	161741.000000	290629.000000	1	1
mean	2023-01-01 01:06:25.463691000	7.365292e+06	8.017360e+03	8.041026e+03	8.125957e+04	71.652139	72.060028	112096.199420	122.357073	122.713852	...	33.097758	50.649390	3.106895	...	...
min	2023-01-01 00:00:00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000	0.000000	0.000000	0.000000	0.000000	...	7.700000	1.190000	0.100000	...	...
25%	2023-01-01 00:00:00	6.280750e+03	0.000000e+00	0.000000e+00	4.300000e+01	0.000000	0.000000	1916.100000	0.000000	0.000000	...	22.600000	20.860000	1.300000	...	...
50%	2023-01-20 00:00:00	6.365300e+04	0.000000e+00	1.200000e+01	7.990000e+02	0.000000	0.000000	29145.480000	0.000000	0.000000	...	27.900000	33.100000	49.540000	2.500000	...
75%	2023-01-08 00:00:00	7.582720e+05	0.000000e+00	3.132900e+02	9.574000e+03	0.000000	3.140000	156770.190000	0.000000	56.250000	...	41.500000	82.500000	4.210000	...	...
max	2024-01-14 00:00:00	7.750668e+06	4.423623e+07	6.319461e+06	7.057132e+06	103719.000000	14817.000000	703598.600000	241758.230000	34536.890000	...	78.100000	100.000000	13.800000	...	...
std	NaN	4.477582e+07	2.296648e+05	8.661611e+04	4.41197e+05	1368.322990	515.636565	16240.412405	1508.778585	559.701663	...	13.853952	31.905236	2.549168	...	...
8 rows × 63 columns																

Se utilizó el método .describe() para obtener una visión numérica general de las variables numéricas del conjunto de datos. Este comando proporciona estadísticas clave como la media, desviación estándar, mínimos, máximos y percentiles de cada variable. A través de esta información, es posible observar si existe un sesgo en los datos. Por ejemplo, si la media y la mediana de new\_cases o new\_deaths difieren considerablemente, podría indicar la presencia de **outliers positivos** (como olas epidémicas en algunos países).

Complementando este análisis, se visualizaron histogramas para algunas variables clave: new\_cases, new\_deaths, positive\_rate y stringency\_index. En los histogramas se observa que las distribuciones **no son normales**; tienden a estar **asimétricamente sesgadas hacia la izquierda**, con una concentración significativa de registros en valores cercanos a cero, y una cola larga hacia la derecha.

Este comportamiento es esperable en datos epidemiológicos, donde muchos países tienen días con muy pocos casos o muertes, mientras que otros (como India, EE.UU. o Brasil) tienen días con cifras muy altas. Esta asimetría indica que al aplicar modelos estadísticos, puede ser necesario usar transformaciones como logaritmos o normalización, especialmente si se busca aplicar métodos que asumen distribuciones normales.

- Usa medidas estadísticas:
  - Medidas de tendencia central: media aritmética, geométrica, armónica, mediana, moda, desviación estándar.
  - Correlación y covarianza: permite entender la relación entre dos variables aleatorias.

```

→ --- new_cases ---
Media: 8017.36
Mediana: 0.00
Moda: 0.0
Desviación estándar: 229664.87

--- new_deaths ---
Media: 71.85
Mediana: 0.00
Moda: 0.0
Desviación estándar: 1368.32

--- positive_rate ---
Media: 0.10
Mediana: 0.06
Moda: 0.01
Desviación estándar: 0.12

--- stringency_index ---
Media: 42.88
Mediana: 42.85
Moda: 11.11
Desviación estándar: 24.87
  
```

Se calcularon las principales medidas de tendencia central y dispersión para algunas de las variables más relevantes del conjunto de datos.

- **La media** representa el valor promedio.
- **La mediana** muestra el valor central de la distribución, menos sensible a los valores extremos.
- **La moda** indica el valor más frecuente, aunque en variables continuas puede no existir.
- **La desviación estándar** cuantifica la variación de los datos respecto a la media.

En la mayoría de los casos (como new\_cases y positive\_rate), se observa que la media es superior a la mediana, lo que confirma una distribución **sesgada positivamente**, es decir, con una cola larga hacia la derecha.

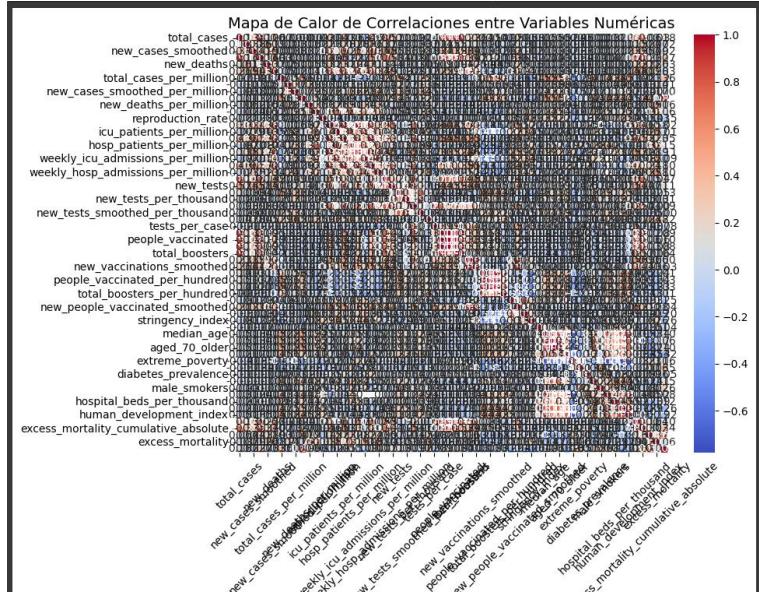
- ¿Hay correlación entre features (características)?

	new_cases	new_deaths	stringency_index	positive_rate
total_cases	0.127604	0.099206	-0.081409	0.041080
new_cases	1.000000	0.505723	0.007153	0.074801
new_cases_smoothed	0.376576	0.187672	0.017095	0.181523
total_deaths	0.156200	0.162214	-0.021345	0.035230
new_deaths	0.505723	1.000000	0.059048	0.048759
...	...	...	...	...
population	0.148152	0.215584	0.114992	-0.074870
excess_mortality_cumulative_absolute	0.322467	0.371079	-0.046687	0.101114
excess_mortality_cumulative	0.035304	0.180617	0.069153	0.243402
excess_mortality	0.074270	0.263156	0.260314	0.383756
excess_mortality_cumulative_per_million	0.016120	0.025834	-0.260050	0.224192

Se calculó la **matriz de correlación** para identificar qué variables están estadísticamente relacionadas.

- Por ejemplo, existe una **correlación positiva alta entre new\_cases y new\_deaths**, lo cual es lógico ya que a mayor número de contagios, es probable que aumenten las muertes.
- También puede encontrarse una relación entre positive\_rate y total\_tests: si el número de pruebas baja, la tasa de positividad sube.

Estas correlaciones son importantes para entender la dinámica de la pandemia y para evitar usar variables altamente correlacionadas en modelos predictivos por riesgo de multicolinealidad.

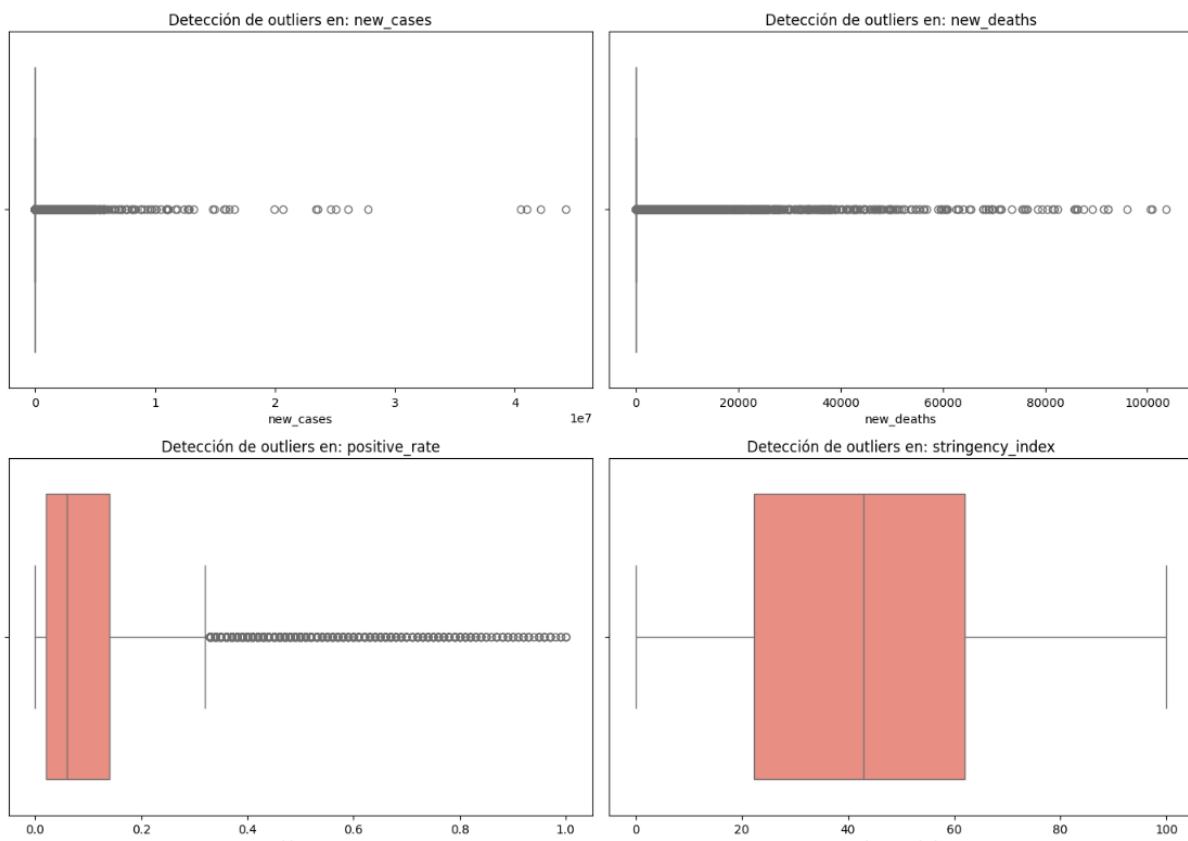


- Las new\_cases y new\_deaths tienen una celda roja oscura con valor 0.85, esto significa que están **fuertemente relacionadas**: cuando aumentan los casos, también aumentan las muertes.
- Si alguna celda está en azul fuerte con valor cercano a -0.5 o menos, indica que una variable tiende a disminuir cuando la otra aumenta.

Este gráfico es clave para seleccionar las variables más independientes para un modelo, evitar redundancias, y entender cómo se relacionan los factores en la evolución del COVID-19.

## Paso 2. Análisis de outliers

- ¿Cuáles son los Outliers? (unos pocos datos aislados que difieren drásticamente del resto y “contaminan” ó desvían las distribuciones)



Se realizó una inspección visual de los datos utilizando **boxplots** para detectar **outliers**, es decir, valores que se alejan drásticamente del resto de la distribución. Los gráficos muestran claramente que variables como new\_cases y new\_deaths contienen valores extremos **mucho mayores** que el resto de los datos, especialmente en países altamente poblados como India o Estados Unidos durante los picos de la pandemia.

Estas observaciones **no deben eliminarse sin análisis contextual**, ya que en este caso los outliers:

- **No son errores de carga**, sino **valores reales**, confirmados por reportes oficiales (por ejemplo, más de 100 mil casos diarios en India en mayo de 2021).
- Reflejan **eventos críticos en la pandemia**, como olas epidémicas o crisis sanitarias.

```
new_cases: 41774 outliers detected
new_deaths: 27708 outliers detected
```

- ¿Podemos eliminarlos? ¿Es importante conservarlos?
- son errores de carga o son reales?

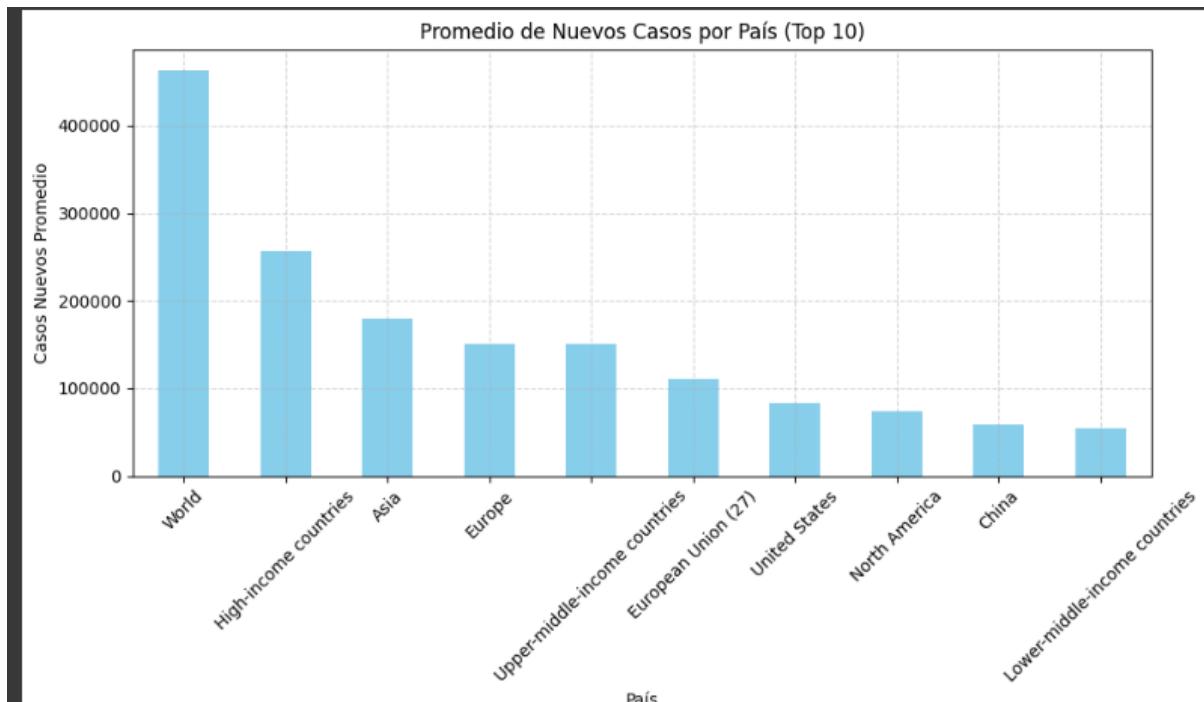
- **Sí hay outliers**, especialmente en casos y muertes.
- **No deben eliminarse automáticamente**: en contextos como este (epidemiológicos), los valores extremos **aportan información valiosa**.
- Es más adecuado **tratar los outliers de forma diferenciada**: por ejemplo, segmentarlos, escalarlos, o analizarlos por separado en modelos predictivos.

### Paso 3: Visualización

- Las variables que podemos representar son:
  - Variables categóricas: Gráfico de barras y circular

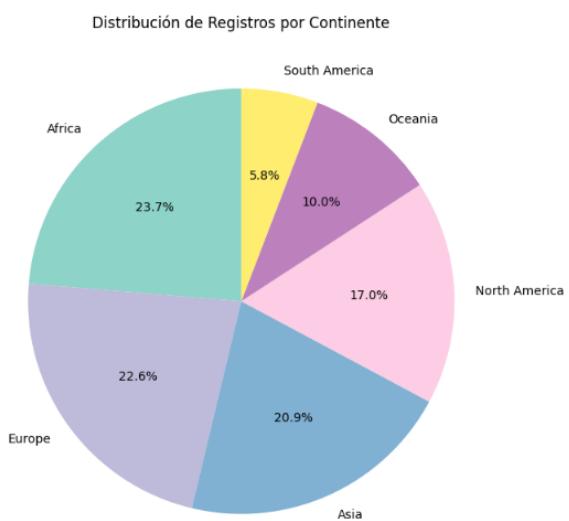
- Variables numéricas: Una variable: histogramas, dos variables: boxplot

Gráfico de barras: comparar cantidades de una variable.



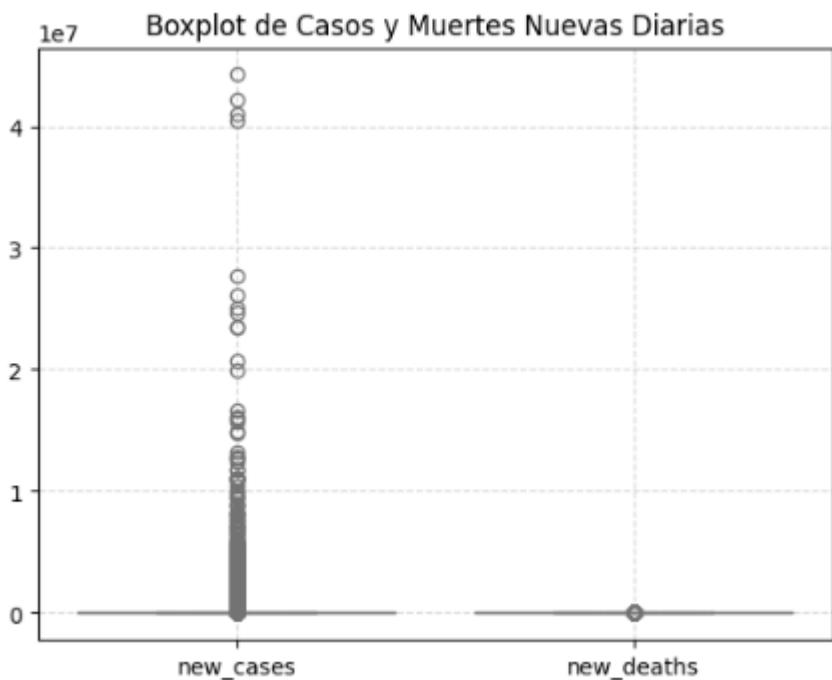
Este gráfico de barras compara el **promedio de nuevos casos diarios** entre los 10 países más afectados. Visualmente, se observa qué naciones han experimentado los mayores niveles de contagio en promedio. Los países con mayor población tienden a destacar, como India o Estados Unidos.

Gráfico circular: para representar porcentajes y proporciones.



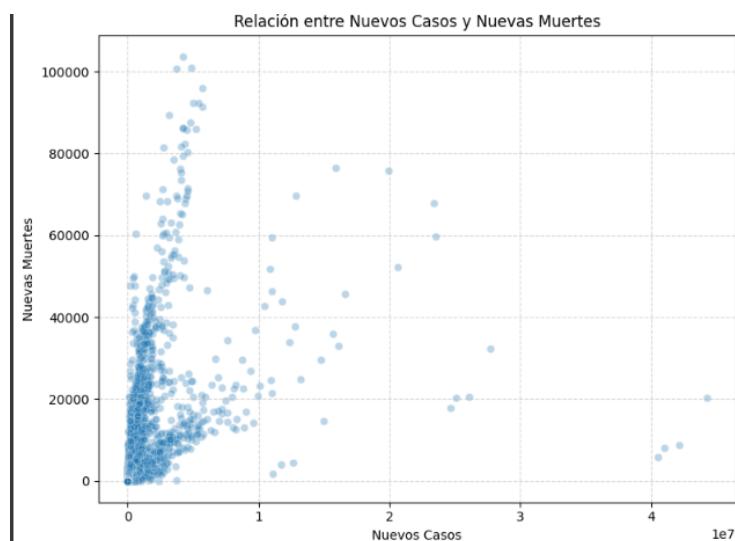
El gráfico circular representa la **proporción de registros por continente**, mostrando visualmente cómo se distribuyen los datos. Europa, Asia y América suelen concentrar la mayor parte de los registros, lo cual refleja una mayor cobertura de datos en estas regiones.

Boxplot: representa los datos numéricos a través de sus cuartiles pudiendo representar los outliers.



El boxplot compara la **dispersión** de los casos nuevos frente a las muertes nuevas. Se observa claramente la **mayor dispersión y presencia de outliers** en los new\_cases, con muchos valores extremos (días con brotes severos). Las muertes tienen una distribución más concentrada.

Scatterplot: muestra el grado de relación entre dos variables.



El scatterplot permite visualizar la **relación entre casos y muertes diarias**. Aunque la mayoría de los puntos están concentrados cerca del origen (valores bajos), se nota una **tendencia lineal creciente**: a mayor número de casos, también se incrementan las muertes, lo que sugiere una correlación positiva.

#### Paso 4. Encuentra un problema potencial en tus datos.

- Si es un problema de tipo supervisado:
  - ¿Cuál es la columna de “salida”? ¿binaria, multiclasificación?
  - ¿Está balanceado el conjunto salida?

Sí, se puede plantear como un **problema supervisado** si el objetivo es predecir alguna variable de salida, como:

- new\_cases (número de casos nuevos)
- new\_deaths (muertes nuevas)
- positive\_rate (tasa de positividad)

En ese caso, el conjunto de entrenamiento incluiría columnas como stringency\_index, total\_tests, reproduction\_rate, etc., como **features o entradas**, y una de las variables anteriores como **salida (label)**.

- ¿Cuáles parecen ser features importantes? ¿Cuáles podemos descartar?

#### Importantes:

- reproduction\_rate: refleja directamente la velocidad de propagación del virus.
- stringency\_index: muestra la severidad de las políticas públicas, podría afectar los casos.
- positive\_rate: correlaciona con número de pruebas y nivel de contagio.

#### Poco útiles o con datos faltantes masivos:

- icu\_patients y hosp\_patients: tienen muchos valores nulos y están presentes solo en ciertos países.

- La distribución, tendencia de las variables varía en el tiempo?

El análisis temporal muestra que las variables como new\_cases, stringency\_index y positive\_rate presentan **cambios evidentes en el tiempo**, por ejemplo:

- Picos de casos en 2020, 2021 (olas de contagio).
  - Cambios en la severidad de las medidas (stringency\_index decrece tras las vacunaciones).
  - Variación estacional en el positive\_rate.
- 
- ¿Hay algún problema notable con la calidad de los datos?

Sí, los principales son:

- **Valores nulos** en columnas críticas (icu\_patients, hosp\_patients, total\_tests).
- **Outliers reales** que deben analizarse por contexto.
- **Diferente granularidad** por país (algunos tienen más frecuencia de reporte que otros).
- **Formatos inconsistentes** en algunas columnas si no se filtran bien.

- 
- ¿Existe alguna relación sorprendente entre las variables?

Sí. Por ejemplo:

- El índice de medidas restrictivas (`stringency_index`) **no siempre** se correlaciona con menos contagios. Esto puede sorprender, pero refleja factores como **cumplimiento ciudadano, capacidad de testeo y vacunación**.
- Algunos países muestran alta tasa de positividad a pesar de bajos `new_cases`, lo que indica **subregistro o baja cantidad de pruebas**.



## Conclusión

### ¿Qué podemos aprender de este análisis?

Este análisis permitió identificar que el conjunto de datos:

- Tiene gran potencial para problemas supervisados y series temporales.
- Contiene outliers reales, datos faltantes y patrones estacionales.
- Permite extraer relaciones relevantes, pero también impone desafíos como el desbalance de clases y la heterogeneidad geográfica.

Antes de construir un modelo predictivo, es clave **limpiar, balancear, y ajustar por tiempo y región** para evitar sesgos y sobreajustes.