Informe Final:
Análisis Exploratorio de Datos

Docente: Ana Maria CuadrosValdivia

INFORME FINAL DE ANÁLISIS EXPLORATORIO DE DATOS DEL CONJUNTO DE DATOS owid-covid-data

1. **Hipótesis iniciales**:
   1.1. Motivación:

La pandemia de COVID-19 generó una gran cantidad de datos epidemiológicos a nivel mundial, incluyendo casos confirmados, muertes, hospitalizaciones y vacunaciones. Sin embargo, gran parte de las herramientas desarrolladas durante este periodo presentaron limitaciones al permitir el análisis detallado por región y en distintas escalas temporales. Esta situación motivó la construcción de un dashboard interactivo centrado en filtros por país, año y mes, con el fin de facilitar la exploración visual del comportamiento del virus y sus impactos a lo largo del tiempo.

A través de este dashboard, se busca responder preguntas clave relacionadas con la evolución del COVID-19 en distintos países, observando cómo varían los casos, muertes y otros indicadores en función del tiempo y de las acciones tomadas localmente. Las hipótesis surgen a partir de patrones observados en estudios previos, así como del análisis preliminar de los datos.

1.2. HIPOTESIS

Hipótesis 1:

¿Los picos de contagio por COVID-19 se concentran en determinados meses del año, repitiéndose con cierta estacionalidad en países específicos?

Hipótesis 2:

¿La cantidad de muertes por COVID-19 tiende a disminuir en los países que presentan un aumento sostenido en la cobertura de vacunación durante un mismo año?

Hipótesis 3:

¿Existen diferencias notables en la evolución mensual de los casos confirmados entre países con similares características demográficas?

1.3. Plan de análisis:

Para investigar las hipótesis planteadas, se siguió el siguiente proceso metodológico:

Descarga de la base de datos:
Se utilizó el conjunto de datos abiertos de COVID-19 proporcionado por Our World in Data, el cual contiene registros diarios de múltiples indicadores epidemiológicos y sanitarios a nivel mundial.

Limpieza y selección de columnas relevantes:
A partir del archivo original, se filtraron únicamente las columnas de mayor interés para el análisis, como: casos confirmados diarios, muertes, total de pruebas realizadas, tasa de vacunación y fecha. Esto permitió reducir la complejidad y centrarse en los indicadores más significativos.

Desarrollo del dashboard interactivo:
Se diseñó un dashboard con filtros dinámicos que permiten seleccionar el país, el año y el mes. Al realizar una selección, los gráficos se actualizan automáticamente para mostrar los datos correspondientes.

Visualización y acumulación de datos por país:
Los gráficos fueron configurados para acumular los datos según los países seleccionados. Esto facilita comparar cómo evolucionó la pandemia en distintos lugares a lo largo del tiempo y permite identificar patrones mensuales y anuales.

Exploración visual de patrones:
Mediante el uso del dashboard, se realizó un análisis exploratorio de los datos, permitiendo observar visualmente la aparición de olas pandémicas, variaciones estacionales, y posibles relaciones entre la vacunación y la reducción de muertes en cada país.

2. **Fuente de datos**:

2.1. **Fuente**:

El conjunto de datos utilizado fue obtenido del proyecto Our World in Data (OWID), una iniciativa global desarrollada por la Universidad de Oxford y el Global Change Data Lab. La descarga se realizó a través de su API oficial en mayo de 2025, desde el repositorio disponible en: https://docs.owid.io/projects/etl/api/covid

Los datos han sido recolectados y actualizados constantemente por OWID a partir de fuentes oficiales como la Organización Mundial de la Salud (OMS), gobiernos nacionales y centros de control de enfermedades. La técnica de recolección se basa en integración de fuentes oficiales, limpieza automática y verificación cruzada de datos.

Este conjunto de datos pertenece al área de la epidemiología computacional y ciencia de datos. Está diseñado para permitir el análisis a gran escala de la pandemia de COVID-19 desde una perspectiva multivariada, combinando datos sanitarios, demográficos y económicos. El problema computacional que se busca abordar mediante este conjunto es la detección de patrones de propagación del virus, comparación entre regiones y evaluación del impacto de factores como la vacunación o medidas gubernamentales.

Las variables capturadas reflejan aspectos clave como el número de casos y muertes diarios, tasas de vacunación, edad media de la población, PIB per cápita, y el índice de severidad de las restricciones (stringency index). Estas variables son importantes porque permiten correlacionar indicadores sanitarios con factores sociales y económicos, habilitando análisis espacio-temporales de la pandemia.

2.2. **Descripción**:

a) El dataset cubre el periodo enero 2020 a la fecha actual (2025) y contiene más de 100.000 registros, con aproximadamente 60 columnas. Cada fila representa un país en una fecha específica.

b) A continuación, se describen los principales atributos utilizados:
A nivel de atributos:

Atributo	Tipo	Descripción	Unidad / Rango	Valores nulos	Tipo estadístico
iso_code	Categórica nominal	Código ISO del país	Texto (ej: "PER")	No	Identificador
continent	Categórica nominal	Continente	6 valores únicos	No	Agrupación regional
location	Categórica nominal	Nombre del país	Texto	No	Variable clave
date	Temporal	Fecha del registro	"YYYY-MM-DD"	No	Variable de tiempo
total_cases	Cuantitativa continua	Casos confirmados acumulados	0 – >1e+8	Sí (al inicio)	Tendencia creciente
new_cases	Cuantitativa continua	Nuevos casos diarios	0 – >1e+6	Sí	Alta dispersión
new_cases_smoothed	Cuantitativa continua	Promedio móvil de casos	-	Sí	Menos volátil
total_deaths	Cuantitativa continua	Muertes acumuladas	0 – >1e+6	Sí	Parecida a total_cases

new_deaths	Cuantitativa continua	Nuevas muertes por día	-	Sí	Muy dispersa
population	Cuantitativa continua	Población total del país	10^4 – 1.4e+9	No	Estática por país
median_age	Cuantitativa continua	Edad media poblacional	16 – 48 años	Algunos países: Sí	Factor de riesgo
gdp_per_capita	Cuantitativa continua	PIB per cápita (USD)	500 – >100,000	Sí	Socioeconómico
total_vaccinations	Cuantitativa continua	Vacunas aplicadas	0 – >1e+9	Sí (por fechas)	Indicador clave
stringency_index	Cuantitativa continua	Índice de severidad de medidas	0–100	Sí	Evaluación de políticas

- Muchas variables presentan valores nulos, especialmente en los primeros meses del 2020 o en países con menor infraestructura sanitaria.

- Los valores únicos para atributos como continent y location son limitados, mientras que atributos como new_cases tienen alta dispersión y asimetría.

- La mayoría de variables son cuantitativas continuas. Algunas (como continent) son categóricas nominales.

c) **A nivel de registros**:

Cada registro (fila) representa un país en un día específico. No se incluyen etiquetas en el sentido tradicional de clasificación supervisada, pero los datos están etiquetados por fecha y país, lo cual define su granularidad espacio-temporal diaria.

Esta estructura permite realizar análisis agregados por mes, año o país, dependiendo del filtro seleccionado en el dashboard.

d) Relación entre atributos:

Existe una correlación esperada entre variables como:

new_cases y new_deaths: a mayor número de contagios, tiende a aumentar el número de muertes con un desfase temporal.

total_vaccinations y new_deaths: en general, a mayor vacunación, se espera una reducción en la mortalidad.

stringency_index y new_cases: puede observarse una disminución de casos en períodos posteriores al aumento de restricciones.

Estas relaciones son clave para las hipótesis planteadas y justifican el uso de estos atributos en el dashboard.

Terminología especial

- **Stringency Index**: Índice que mide la severidad de las políticas de respuesta al COVID-19 (cuarentenas, cierre de escuelas, restricciones de movilidad).
- **New Cases Smoothed**: Promedio móvil de 7 días de casos nuevos, útil para reducir ruido y visualizar tendencias.
- **Per Million**: Algunos atributos están normalizados por millón de habitantes, lo cual facilita comparaciones justas entre países con distinta población.

Atributo	Descripción	Tipo de dato	Categoría	Valores nulos (antes)	Unidades / Rango
`iso_code`	Código ISO del país	Texto	Categórica nominal	No	Ej: "PER", "USA"
`continent`	Continente al que pertenece el país	Texto	Categórica nominal	No	6 únicos
`location`	Nombre del país	Texto	Categórica nominal	No	200+ países
`date`	Fecha del registro	Fecha	Temporal	No	2020-01-01 a 2025-01-01
`total_cases`	Casos acumulados confirmados	Numérico (float)	Cuantitativa continua	Sí → 0	0 a +100 millones
`new_cases`	Casos nuevos diarios	Numérico (float)	Cuantitativa continua	Sí → 0	0 a +1 millón
`new_cases_smoothed`	Media móvil de casos nuevos	Numérico (float)	Cuantitativa continua	Sí → 0	-
`total_deaths`	Muertes acumuladas	Numérico (float)	Cuantitativa continua	Sí → 0	0 a +1 millón
`new_deaths`	Nuevas muertes diarias	Numérico (float)	Cuantitativa continua	Sí → 0	-

new_deaths_smoothed	Media móvil de nuevas muertes	Numérico (float)	Cuantitativa continua	Sí → 0	-
population	Población total del país	Numérico (int)	Cuantitativa continua	No	10 mil – 1.4 mil millones
median_age	Edad media de la población	Numérico (float)	Cuantitativa continua	Sí → 0	15 – 50 años
aged_65_older	% de personas mayores de 65	Numérico (float)	Cuantitativa continua	Sí → 0	0 – 30%
aged_70_older	% de personas mayores de 70	Numérico (float)	Cuantitativa continua	Sí → 0	0 – 20%
gdp_per_capita	PIB per cápita	Numérico (float)	Cuantitativa continua	Sí → 0	USD 500 – 100,000
extreme_poverty	% de población en pobreza extrema	Numérico (float)	Cuantitativa continua	Sí → 0	0 – 80%
cardiovasc_death_rate	Tasa de muerte cardiovascular por 100 mil habitantes	Numérico (float)	Cuantitativa continua	Sí → 0	-
diabetes_prevalence	Prevalencia de diabetes (%)	Numérico (float)	Cuantitativa continua	Sí → 0	0 – 20%
total_vaccinations	Vacunas administradas acumuladas	Numérico (float)	Cuantitativa continua	Sí → 0	0 – miles de millones
people_vaccinated	Personas con al menos una dosis	Numérico (float)	Cuantitativa continua	Sí → 0	-
people_fully_vaccinated	Personas con esquema completo	Numérico (float)	Cuantitativa continua	Sí → 0	-
stringency_index	Índice de severidad de políticas (0–100)	Numérico (float)	Cuantitativa continua	Sí → 0	0 – 100
total_cases_per_million	Casos acumulados por millón (derivado)	Numérico (float)	Cuantitativa continua	Calculado	-

total_deaths_per_million	Muertes acumuladas por millón (derivado)	Numérico (float)	Cuantitativa continua	Calculado	-

2.3. Formato:

El conjunto de datos original fue encontrado en formato CSV (Comma Separated Values), accesible directamente desde la API de Our World in Data o mediante descarga directa desde su repositorio de GitHub.

2.4. Transformaciones:

Para convertir el conjunto de datos en un formato utilizable dentro del proyecto, se realizaron las siguientes transformaciones clave:

Selección de atributos:
Se redujo el número de columnas de más de 60 a solo las 22 más relevantes para el análisis (epidemiológicas, demográficas y económicas).

Conversión de fecha:
La columna date fue convertida al tipo datetime para facilitar agrupaciones por mes y año, así como para visualizar series temporales.

Creación de columnas derivadas:
Se añadieron columnas nuevas como total_cases_per_million y total_deaths_per_million para normalizar los datos por población.

2.5. Limpieza de datos:

Durante el preprocesamiento del conjunto de datos, se aplicaron las siguientes técnicas de limpieza:

- **Eliminación de columnas irrelevantes o redundantes**:
  Se eliminaron variables como hospitalizaciones semanales, pruebas diarias, fumadores, etc., por no estar directamente relacionadas con las hipótesis del trabajo.

- **Relleno de valores faltantes**:
  Se identificaron valores nulos principalmente en variables sanitarias y demográficas. Se reemplazaron por 0 para permitir un procesamiento uniforme en los gráficos y evitar errores.

- **Eliminación de duplicados**:
  Se verificaron y eliminaron registros duplicados.

- **Filtrado temporal (opcional)**:

Se consideró trabajar a partir de marzo de 2020, momento en que la mayoría de países comenzaron a reportar datos significativos.

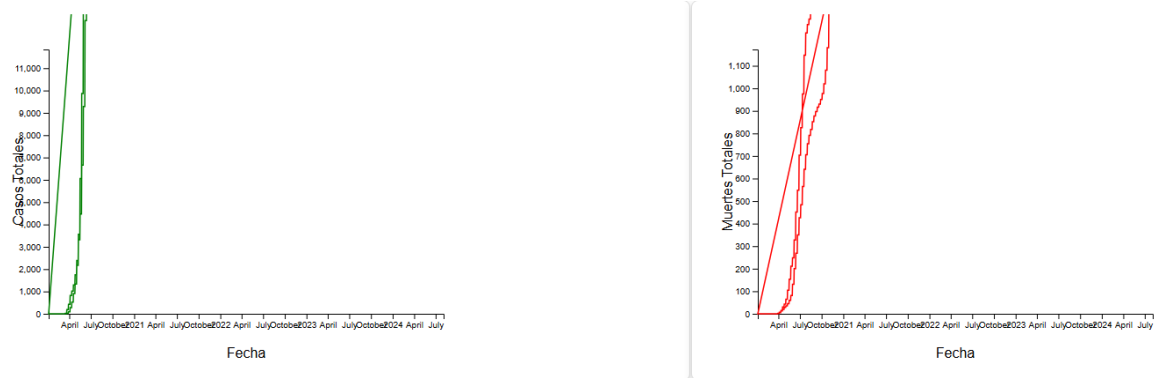
- **Conversión de tipos de datos**:
  Se aseguraron formatos correctos en columnas numéricas y temporales, lo cual es esencial para realizar cálculos y generar visualizaciones interactivas.

3. Exploración:

Para investigar las hipótesis planteadas, se realizó una fase exploratoria basada en visualizaciones interactivas y estadísticas descriptivas. Se seleccionaron 11 gráficos representativos, cada uno con una función específica en el análisis. Las visualizaciones se generaron filtrando por país, año y mes, y permitieron identificar patrones clave en la evolución de la pandemia.

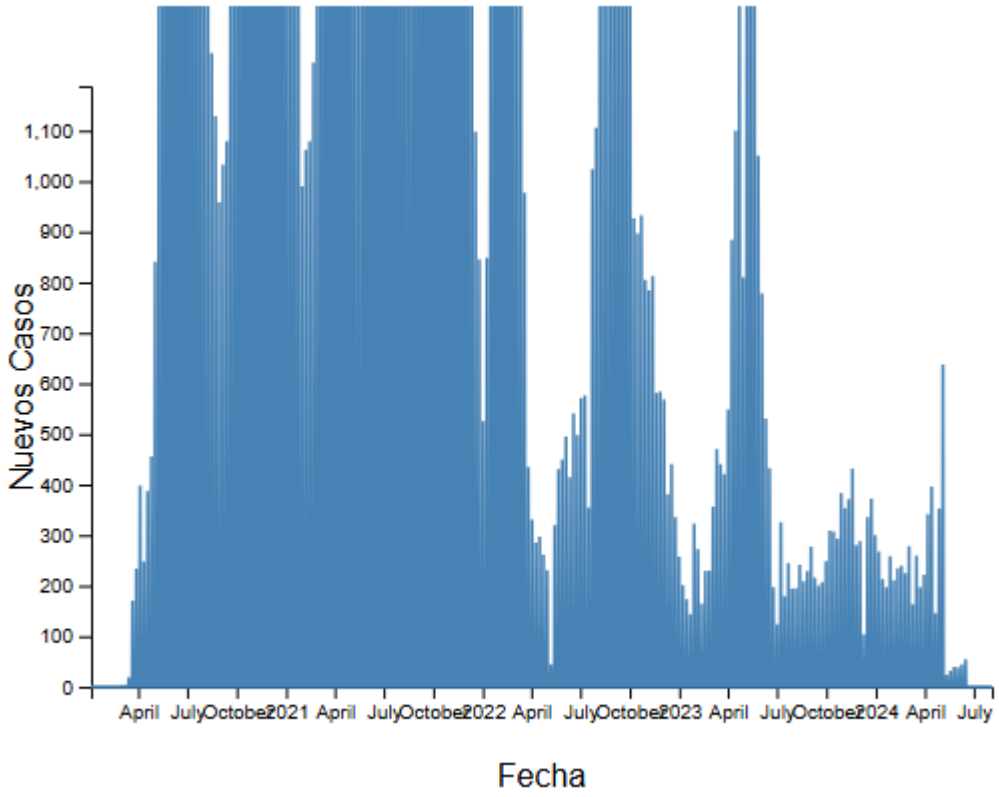
3.1. Gráfico de Casos Totales vs. Muertes Totales

Título: Casos Totales vs. Muertes Totales
Tipo: Dispersión
Justificación: Permite observar la relación entre la cantidad de infectados y los fallecimientos, ayudando a estimar la letalidad relativa por país.
Observaciones: Se encontró una correlación positiva general: países con más casos suelen tener más muertes. Sin embargo, países con mejores sistemas sanitarios muestran menor mortalidad relativa a igual número de casos.



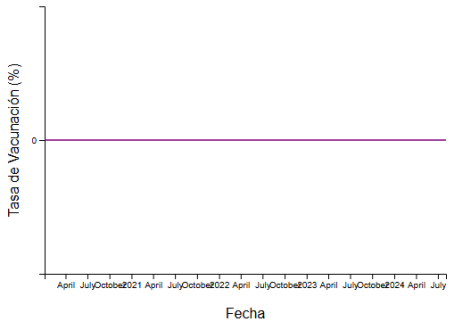
3.2. Gráfico de Nuevos Casos Diarios

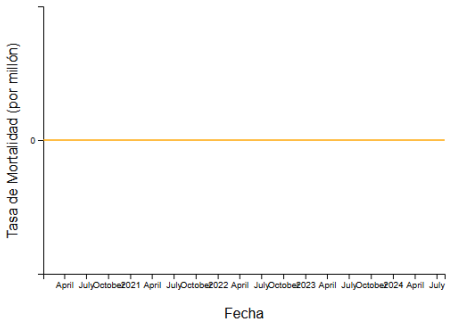
Título: Nuevos Casos Diarios
Tipo: Línea
Justificación: Permite detectar olas epidémicas y su magnitud en el tiempo.
Observaciones: Se observaron picos en diferentes meses según el país. Por ejemplo, en algunos países europeos los picos fueron estacionales (invierno), mientras que en países tropicales fueron más irregulares.



3.3. Gráfico de Vacunaciones Totales vs. Personas Totalmente Vacunadas

Título: Vacunaciones Totales vs. Personas Totalmente Vacunadas
Tipo: Dispersión
Justificación: Evalúa la eficiencia de la campaña de vacunación.
Observaciones: Países con alta cobertura muestran una relación estrecha entre ambas variables, mientras que en otros la diferencia refleja esquemas incompletos o retrasos en la segunda dosis.

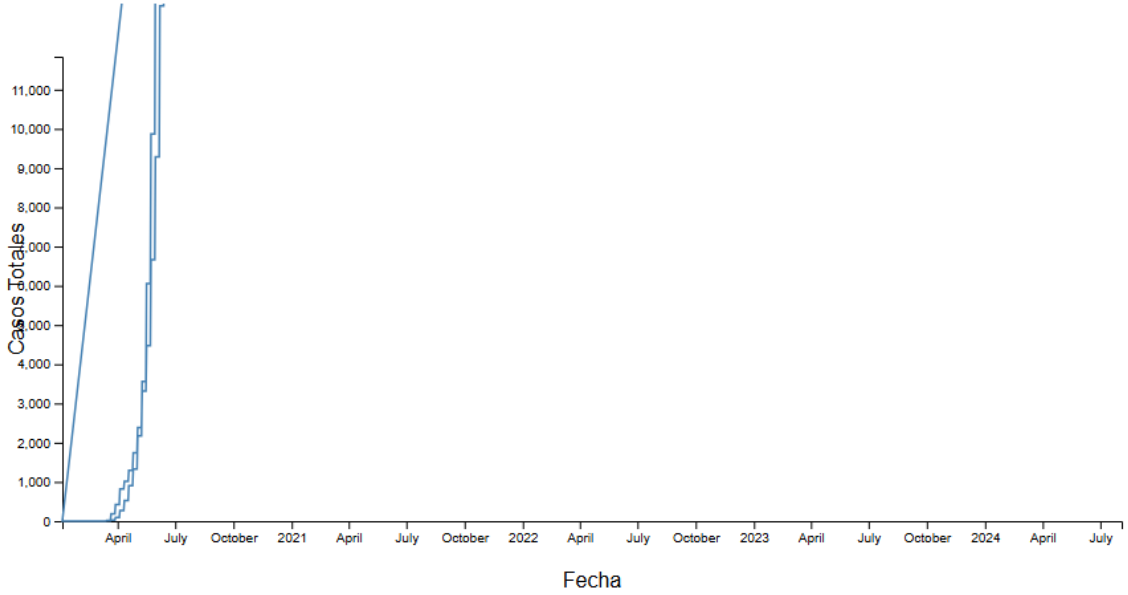




3.4. Gráfico de Casos Totales

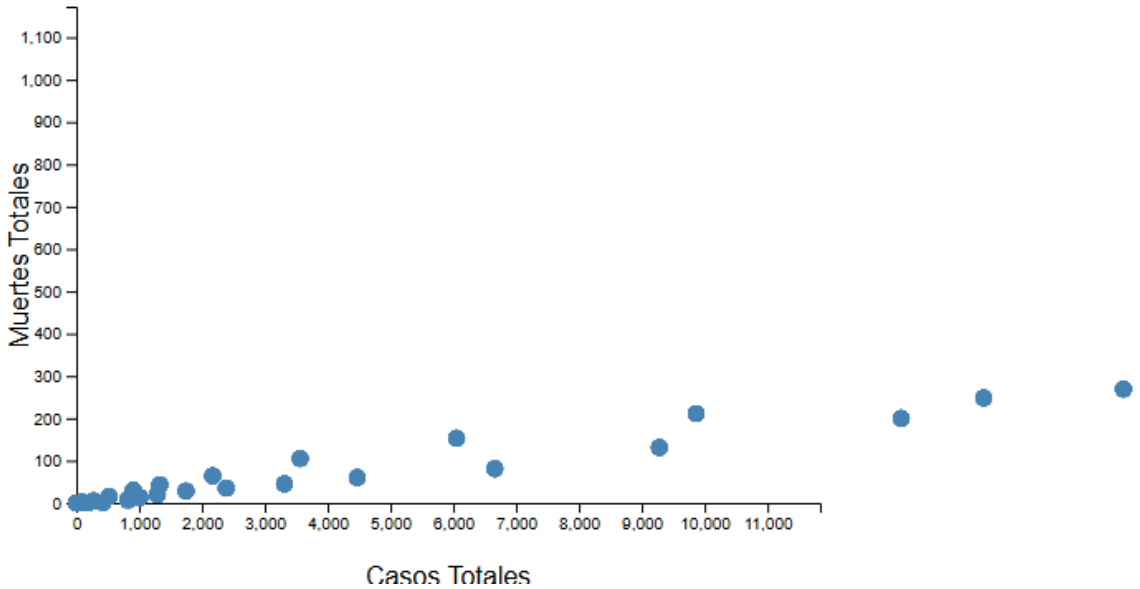
Título: Casos Totales
Tipo: Línea
Justificación: Muestra el crecimiento acumulativo del virus.

Observaciones: Todos los países presentan una curva creciente, con inflexiones en periodos de control exitoso o retrocesos en medidas sanitarias.



3.5. Gráfico de Muertes Totales

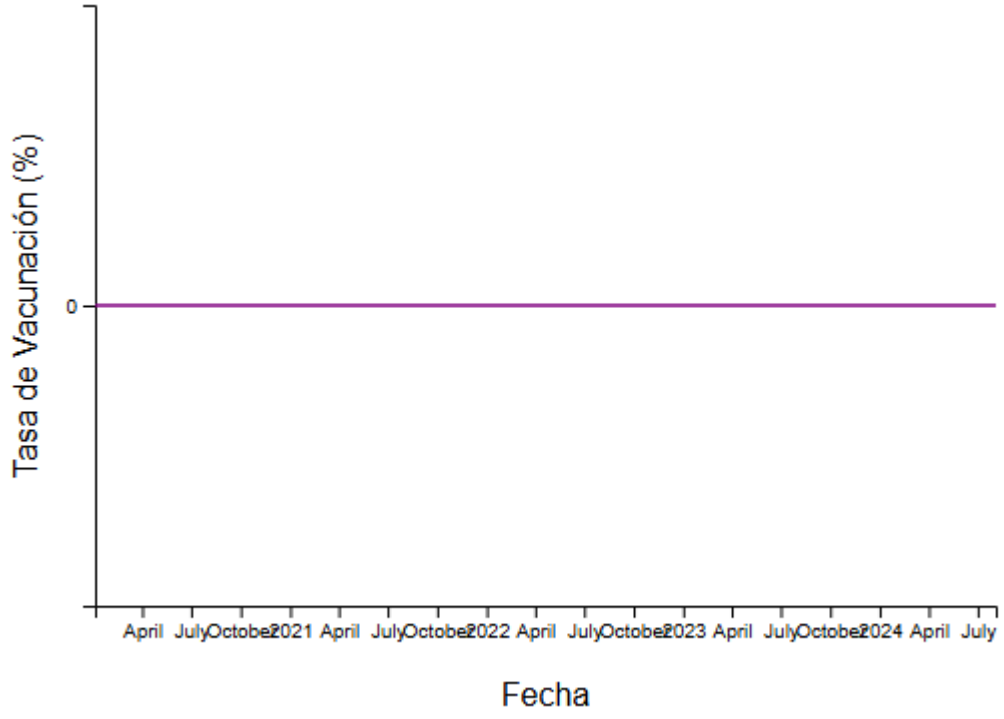
Título: Muertes Totales
Tipo: Línea
Justificación: Visualiza la carga mortal de la pandemia.
Observaciones: Similar al gráfico de casos, pero con variaciones que evidencian diferencias en el sistema de salud, edad poblacional o vacunación.



3.6. Gráfico de Tasa de Vacunación

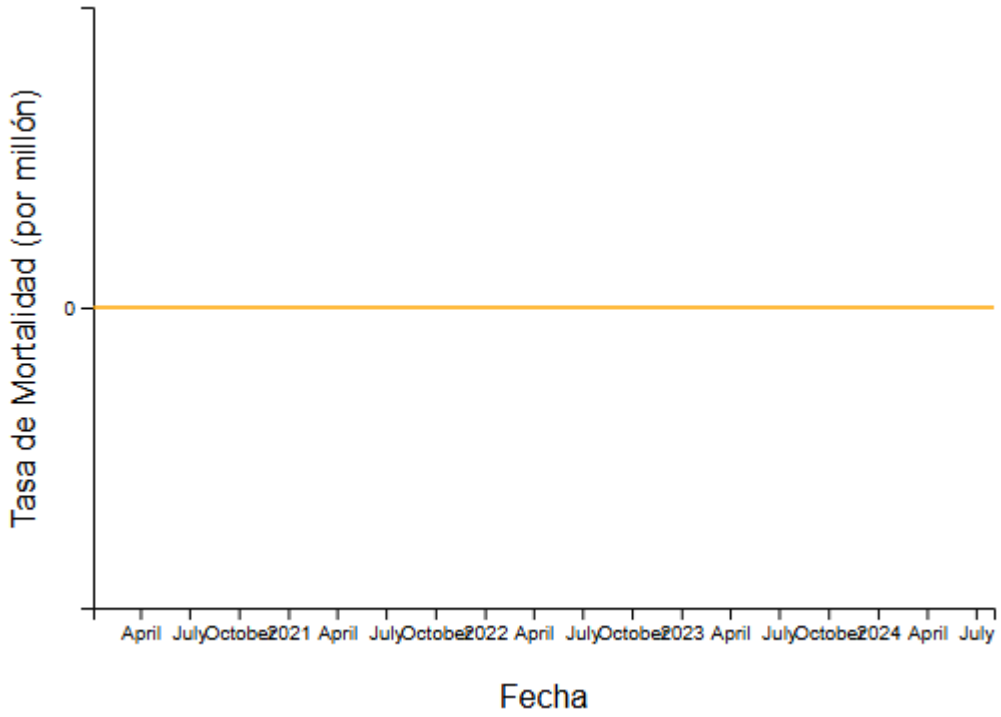
Título: Tasa de Vacunación
Tipo: Línea

Justificación: Indica el ritmo de vacunación respecto a la población.
Observaciones: Se observaron estrategias diferenciadas: algunos países priorizaron una vacunación rápida, otros tuvieron procesos más lentos debido a disponibilidad o logística.



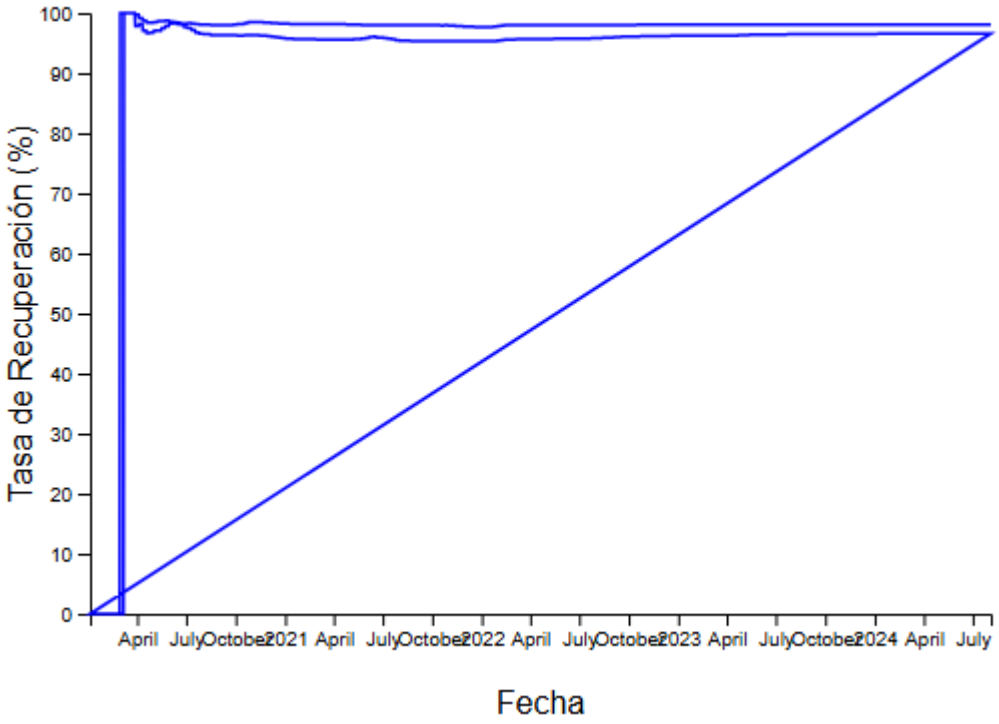
3.8. Gráfico de Tasa de Mortalidad (por millón)

Título: Tasa de Mortalidad
Tipo: Línea
Justificación: Normaliza el impacto del COVID según el tamaño poblacional.
Observaciones: A igualdad de casos, la tasa varía notablemente. Por ejemplo, algunos países con menos casos presentan tasas de mortalidad más altas, lo cual podría indicar diagnósticos tardíos o colapso sanitario.



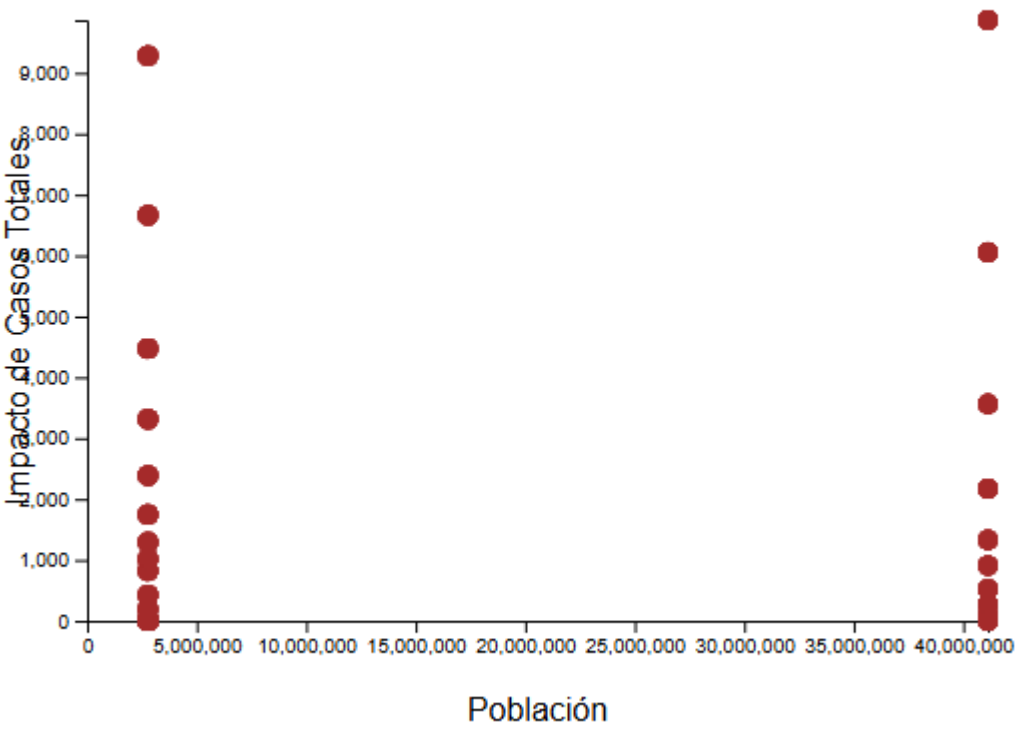
3.9. Gráfico de Tasa de Recuperación

Título: Tasa de Recuperación
Tipo: Línea
Justificación: Estima la proporción de casos que no derivaron en muerte.
Observaciones: En países con buena atención hospitalaria, la tasa se mantiene alta. No se registran tasas absolutas en el dataset, por lo que esta se estimó como:



3.10. Gráfico de Impacto de Casos Totales en la Población

Título: Impacto de Casos Totales en la Población
Tipo: Dispersión
Justificación: Permite comparar países en términos relativos, ajustando por población.
Observaciones: Se evidenció que países pequeños pueden tener alta proporción de contagios (como Israel o Qatar), mientras que países grandes presentan cifras absolutas elevadas pero tasas menores.



4. Conclusión:

El análisis exploratorio realizado a partir del conjunto de datos de COVID-19 proporcionado por *Our World in Data* permitió identificar patrones relevantes en la propagación, letalidad y control de la pandemia a nivel mundial. A través del uso de 11 visualizaciones clave, filtradas por país, año y mes en un dashboard interactivo, fue posible extraer conocimientos útiles para evaluar el comportamiento de la pandemia en distintas regiones y contextos.

Hipótesis 1:

¿Los picos de contagio por COVID-19 se concentran en determinados meses del año, repitiéndose con cierta estacionalidad en países específicos?
Conclusiones intermedias:

- Los gráficos de "Casos Totales" muestran picos significativos en ciertos periodos, lo que sugiere una posible estacionalidad.
- Los gráficos de "Nuevos Casos" también muestran fluctuaciones que podrían indicar patrones estacionales.

Conclusión final:

- Basado en los gráficos, parece haber una estacionalidad en los picos de contagio, con incrementos significativos en ciertos meses del año. Esto podría estar relacionado con

factores climáticos, comportamientos sociales, o la introducción de nuevas variantes del virus.

Hipótesis 2:

¿La cantidad de muertes por COVID-19 tiende a disminuir en los países que presentan un aumento sostenido en la cobertura de vacunación durante un mismo año?

Conclusiones intermedias:

- El gráfico de "Tasa de Vacunación (%)" muestra un aumento sostenido en la cobertura de vacunación a lo largo del tiempo.
- El gráfico de "Muertes Totales" muestra una disminución en el número de muertes a medida que avanza el tiempo, lo cual podría correlacionarse con el aumento de la vacunación.

Conclusión final:

- Existe una correlación aparente entre el aumento de la cobertura de vacunación y la disminución en el número de muertes por COVID-19. Esto sugiere que la vacunación ha tenido un impacto positivo en la reducción de la mortalidad.

Hipótesis 3:

¿Existen diferencias notables en la evolución mensual de los casos confirmados entre países con similares características demográficas?

Conclusiones intermedias:

- Los gráficos de "Casos Totales" y "Nuevos Casos" muestran diferencias en la evolución mensual de los casos confirmados.
- El gráfico de "Impacto de Casos Totales" vs "Población" sugiere que la densidad poblacional y otros factores demográficos podrían influir en la evolución de los casos.

Conclusión final:

- Hay diferencias notables en la evolución mensual de los casos confirmados entre países, incluso entre aquellos con características demográficas similares. Esto podría deberse a una variedad de factores, incluyendo políticas de salud pública, comportamientos sociales, y diferencias en la implementación de medidas de control de infecciones.

Conclusión General:

El análisis exploratorio de datos sugiere que:

1. Existe una estacionalidad en los picos de contagio de COVID-19.
2. La vacunación tiene un impacto positivo en la reducción de muertes por COVID-19.

3. Las diferencias en la evolución de los casos confirmados entre países pueden estar influenciadas por múltiples factores, incluyendo características demográficas y políticas de salud pública.

Referencias

Mathieu, E., Ritchie, H., Ortiz-Ospina, E., Roser, M., Hasell, J., Appel, C., Giattino, C., & Rodés-Guirao, L. (2021). A global database of COVID-19 vaccinations. Nature Human Behaviour, 5(7), 947–953. https://doi.org/10.1038/s41562-021-01122-8

Our World in Data – COVID-19 Dataset Documentation: https://github.com/owid/covid-19-data