**Expansión Estratégica de Biogenesys**

**Nombre del autor**: Gonzalez Esther

**Email**: [esther.isa.gonza@gmail.com](mailto:esther.isa.gonza@gmail.com)

**Cohorte**: DAF-017

**Fecha de entrega**: 29-09-2025

# Introducción

# El presente informe resume el análisis exploratorio de datos (EDA) realizado sobre la base de casos de COVID-19 en distintos países de Latinoamérica. El estudio busca comprender no sólo la evolución de la pandemia a través de variables sanitarias como casos confirmados, fallecimientos y recuperados, sino también explorar cómo estos indicadores se relacionan con factores demográficos, económicos y contextuales que influyen en su comportamiento.

# Para este propósito se aplicaron técnicas de estadística descriptiva mediante el uso de librerías como Pandas y NumPy, que permitieron calcular medidas de tendencia central, dispersión y distribución, además de identificar valores atípicos y patrones relevantes. Asimismo, se incorporaron visualizaciones con Matplotlib y Seaborn, que facilitaron la interpretación de los datos y el hallazgo de correlaciones entre las variables más significativas.

# Este análisis constituye una primera etapa dentro de un proceso más amplio de construcción de tableros interactivos en Power BI, cuyo fin es ofrecer herramientas visuales que apoyen la toma de decisiones estratégicas en el ámbito sanitario. La información obtenida no solo permite evaluar el impacto de la pandemia en cada país, sino también establecer comparaciones entre ellos y proyectar escenarios que sirvan de base para la selección de lugares óptimos de inversión en infraestructura de salud, como la instalación de un laboratorio.

Objetivos :

* Realizar un análisis exploratorio de datos sobre la incidencia de COVID-19 y otros factores relevantes, identificando tendencias y oportunidades mediante estadísticas, mediciones y visualizaciones.
* Aplicar técnicas de limpieza de datos para asegurar la calidad de los datos, facilitando análisis y decisiones estratégicas confiables.
* Mejorar el acceso a los datos mediante operaciones eficientes de extracción, transformación y carga (ETL), aumentando la eficacia del análisis y la toma de decisiones.
* Desarrollar dashboards interactivos con visualizaciones eficientes, permitiendo explorar datos desde múltiples perspectivas para una toma de decisiones informada y estratégica.

# Desarrollo del proyecto

## 1. Metodología

El trabajo se realizó en dos fases:

* ETL (Extracción, Transformación y Carga): limpieza de duplicados, tratamiento de valores nulos, estandarización de formatos y preparación de nuevas variables derivadas.
* EDA (Análisis Exploratorio de Datos): análisis estadístico descriptivo y visualización de datos para identificar patrones, distribuciones y relaciones entre variables.
* Librerías utilizadas: Pandas, Numpy, Matplotlib y Seaborn.

## 2. Análisis ETL

En la etapa de limpieza se realizaron las siguientes acciones:

Las tareas realizadas fueron las siguientes:

1. Carga de librerías y dataset  
   Se utilizaron Pandas y NumPy como principales librerías para el manejo de datos.  
     
   El dataset fue cargado desde un archivo CSV con información de casos confirmados, fallecimientos, población y otros indicadores.
2. Exploración inicial  
     
   Se visualizaron las primeras filas con head(10) para comprobar la correcta lectura de la base.  
     
   Se utilizó info() y shape para conocer la cantidad de registros, columnas y tipos de datos.  
     
   El dataset inicial contenía 12.216.057 filas y 50 columnas.
3. Selección de países de interés  
     
   Se filtraron únicamente los países definidos para el análisis: Argentina, Brasil, Chile, Colombia, México y Perú.  
     
   Se verificó que el filtrado fuera correcto usando unique().
4. Filtrado temporal  
     
   Se convirtieron las fechas a formato datetime y se seleccionaron registros posteriores al 01/01/2021, para centrar el análisis en un período más reciente.
5. Revisión de valores nulos y no nulos  
     
   Se contaron los valores faltantes (isnull().sum()) y válidos (notnull().sum()) por columna.  
     
   Se detectaron columnas con alta proporción de valores nulos, especialmente en datos poblacionales y de vacunación.
6. Revisión de duplicados y claves  
     
   Se analizaron las columnas location\_key, country\_code y country\_name, eliminando aquellas que no aportan información adicional para el nivel de análisis.
7. Tratamiento de valores faltantes  
     
   Para variables críticas (new\_confirmed, new\_deceased, new\_recovered), se imputaron los valores faltantes con 0.  
     
   Adicionalmente, en la columna new\_recovered se completaron valores nulos con el promedio de recuperados por país, logrando así una imputación más representativa.
8. Exportación de datos limpios  
     
   El dataset resultante, con las transformaciones aplicadas, se guardó como data\_latinoamerica\_clean.csv, quedando con 3.744 filas y 48 columnas.

**3. Resultados**

* Reducción del dataset: de 12.2 millones de registros a 3.744, al seleccionar países y fechas relevantes.
* Número de columnas: de 50 a 48, tras eliminar redundancias (location\_key, country\_code).
* Valores nulos: se identificó alta proporción en algunas variables, las cuales fueron corregidas con imputación (0 o promedio por país).
* Dataset limpio y consistente: sin duplicados y listo para el análisis estadístico del Avance 2.

**4. Análisis EDA**

Todos los códigos utilizados se pueden ver en el archivo de python del trabajo

En este segundo avance se realizó la carga sobre el csv guardado en el avance 1 con los datos ya limpios para poder realizar el respectivo EDA.

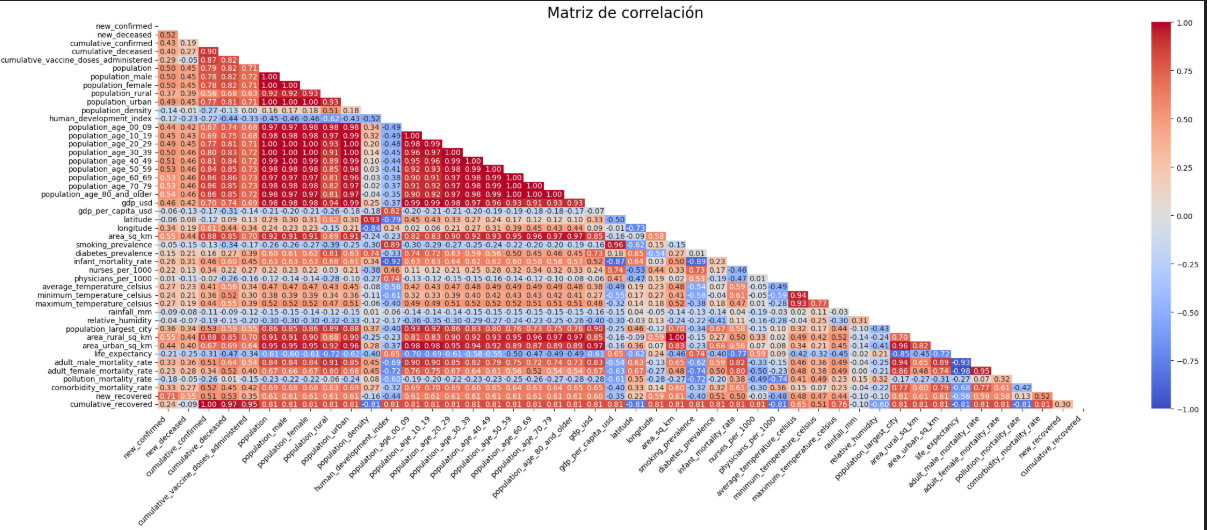
Se realizaron las siguientes tareas:

* Carga de datos: Se utilizó la librería Pandas para leer el archivo CSV en un DataFrame (df\_Latam).  
   df\_Latam = pd.read\_csv("data\_latinoamerica\_clean.csv")
* Revisión de duplicados: Se verificó si existían filas repetidas en la base.  
  df\_Latam.duplicated().sum()
* Exploración inicial: Se aplicó describe(include="all") para obtener un resumen estadístico general de todas las variables, incluyendo categóricas y numéricas.
* Revisión de nulos: Se calcularon los valores nulos y no nulos por columna.
* Revisión de tipos de datos: Se identificaron columnas categóricas (ej. país, fecha) y numéricas (ej. casos, fallecidos, recuperados).
* Tratamiento de valores negativos: Se detectaron valores negativos en variables de casos y muertes diarias, que fueron reemplazados por 0 para evitar distorsión en los análisis posteriores.
* Se detectaron valores negativos en casos y muertes diarias, corregidos a 0.
* Luego se procedió a realizar las gráficas correspondientes para continuar con el análisis

**5. Gráficas**

### 5.1 Matriz de correlación

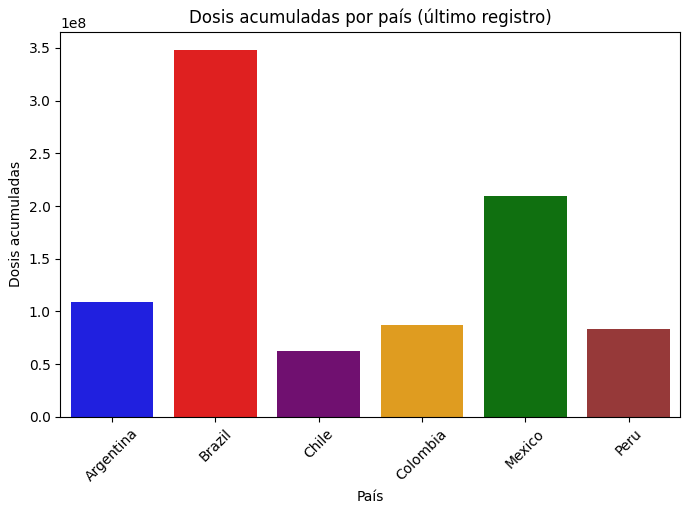
Se observa una alta correlación positiva entre los distintos grupos etarios de la población y entre el tamaño poblacional y el número de dosis administradas. También existe una relación importante entre vacunación y variables demográficas, confirmando que el tamaño poblacional influye directamente en la cantidad de dosis aplicadas.



1. Relación entre contagios y fallecimientos  
   Se observa una correlación positiva moderada (≈0.52) entre los casos confirmados diarios y los fallecimientos diarios.  
   Esto confirma la expectativa epidemiológica: a mayor número de contagios, tiende a aumentar la mortalidad.
2. Población y magnitud de la pandemia  
   La población total y sus desagregaciones etarias presentan correlaciones positivas con casos y muertes acumuladas (≈0.45 – 0.53).  
   Esto indica que los países más poblados reportaron más casos y muertes en términos absolutos.  
   Sin embargo, el análisis relativo (tasas por 100.000 habitantes) es imprescindible para evitar sesgos.
3. Vacunación y contagios  
   La vacunación acumulada tiene una correlación positiva baja a moderada (≈0.29) con los casos confirmados.  
   Esto sugiere que los países más afectados desplegaron campañas de vacunación más intensivas.  
   La relación no es directa: también dependen de factores como logística, disponibilidad de dosis y estrategias gubernamentales.
4. Densidad poblacional e índice de desarrollo humano (IDH)  
   Se detectan correlaciones negativas leves (≈ -0.11 a -0.13) con contagios y muertes.  
   Esto evidencia que ni la concentración urbana ni el nivel de desarrollo explican por sí solos la magnitud de la pandemia.  
   Algunos países con alta densidad lograron control gracias a políticas estrictas, mientras que otros con baja densidad tuvieron brotes relevantes.
5. PIB y capacidad de respuesta  
   El PIB total (gdp\_usd) muestra correlaciones positivas moderadas con casos y muertes (≈0.46).  
   En cambio, el PIB per cápita presenta correlaciones débiles o negativas.  
   Esto significa que los países con economías más grandes registraron más casos por mayor capacidad de testeo y registro, pero un mayor ingreso individual no garantiza menor mortalidad o contagios.
6. Mortalidad por género.  
   La mortalidad masculina y femenina están altamente correlacionadas entre sí (≈1).  
   Esto confirma que las diferencias de género no explican la mortalidad, sino que esta responde a condiciones sanitarias y socioeconómicas generales que afectan a toda la población.

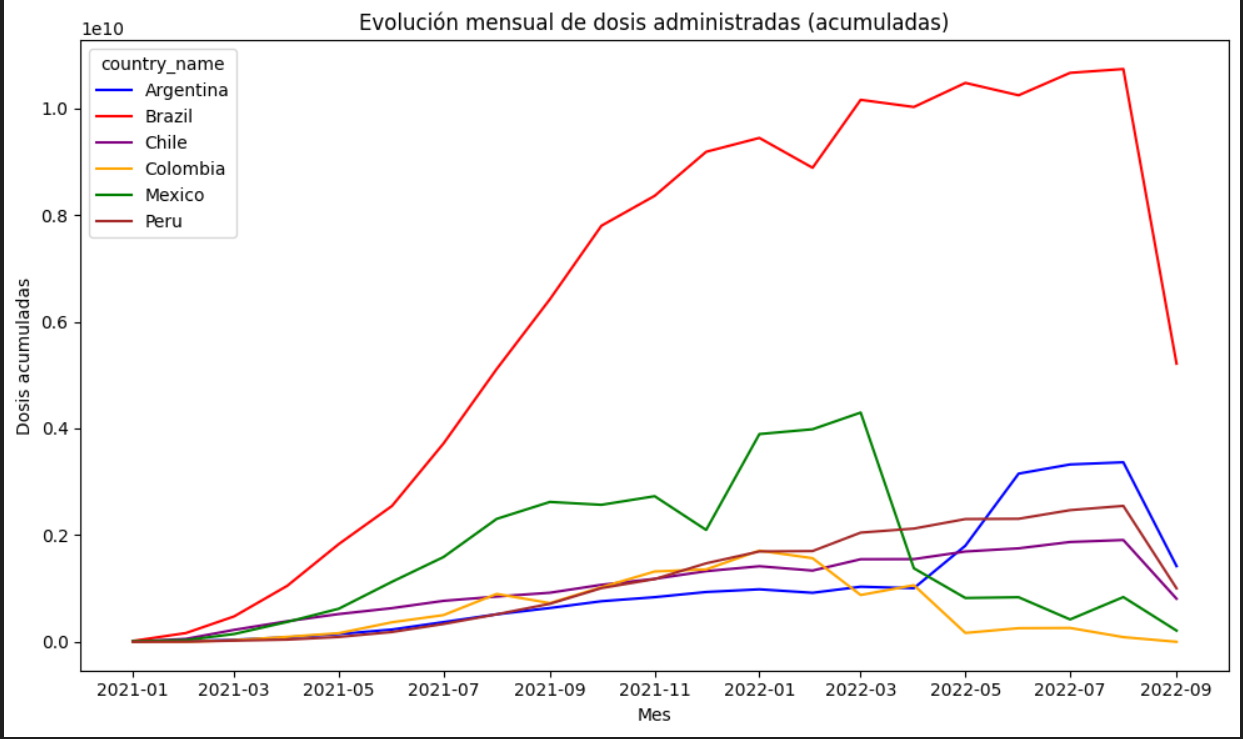
### 5.2 Dosis acumuladas por país

Brasil lidera ampliamente en dosis aplicadas, seguido por México y Argentina. Chile, Colombia y Perú muestran menores volúmenes.



### 5.3 Evolución mensual de dosis administradas

Brasil muestra el crecimiento más acelerado. México y Argentina mantienen una curva sostenida, mientras que Chile, Colombia y Perú presentan un aumento más gradual.



### 

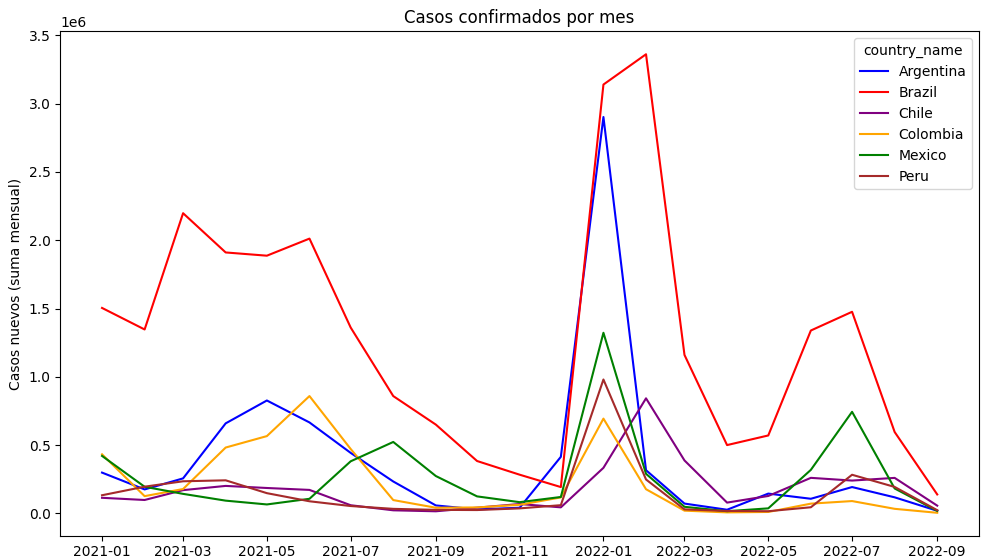
### 

### 

### 

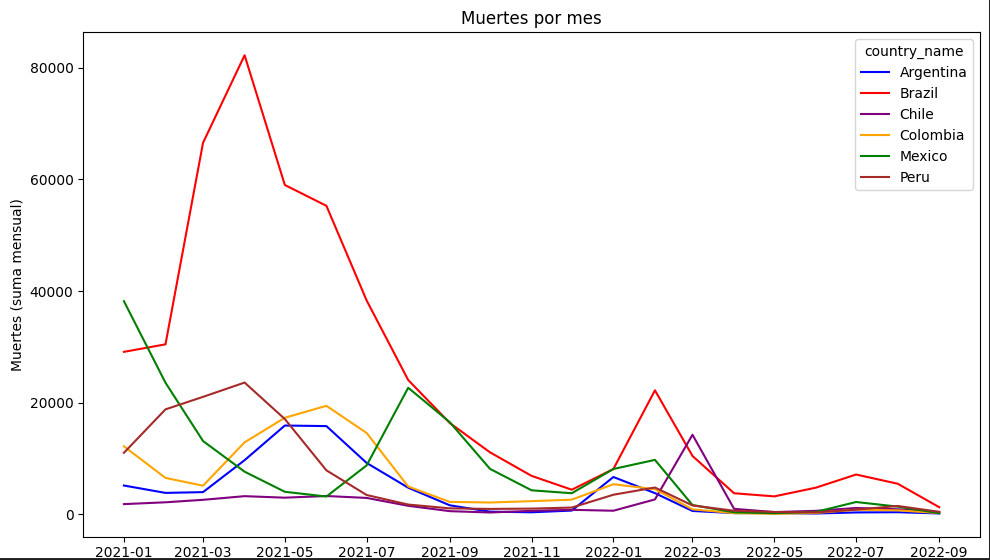
### 5.4 Casos confirmados por mes

Brasil concentra los picos más altos, seguido por Argentina en 2022 (ola Ómicron). El resto de países presenta valores más bajos, aunque con picos en las mismas fechas.



### 5.5 Muertes por mes

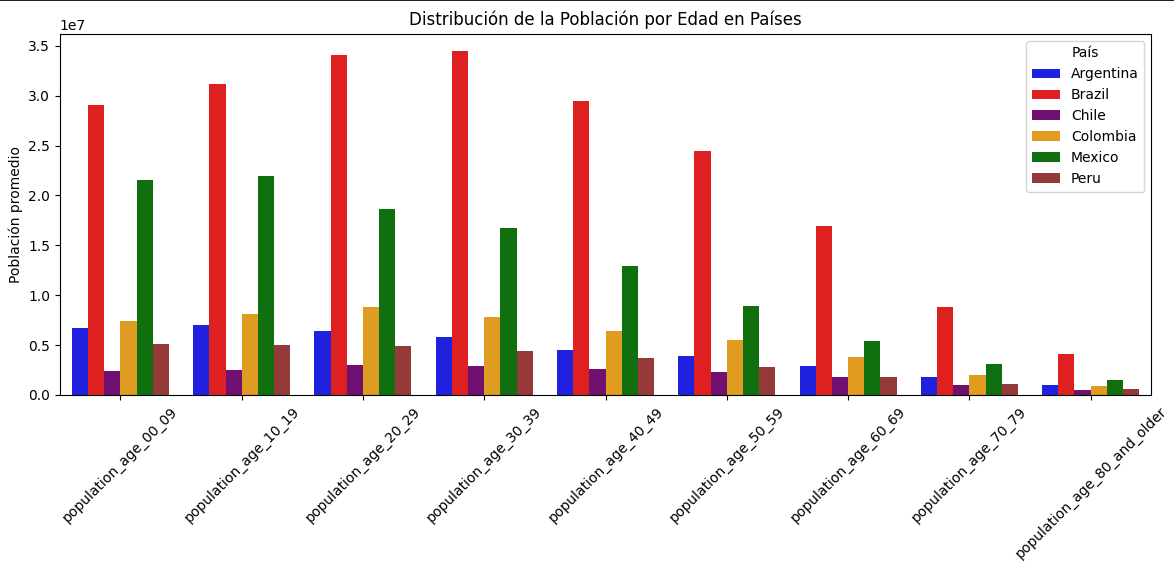
La mortalidad sigue un patrón similar a los casos. Brasil destaca con los niveles más altos, mientras que México, Argentina y Perú presentan cifras significativas. A medida que avanza la vacunación, se observa una reducción de las muertes.



### 

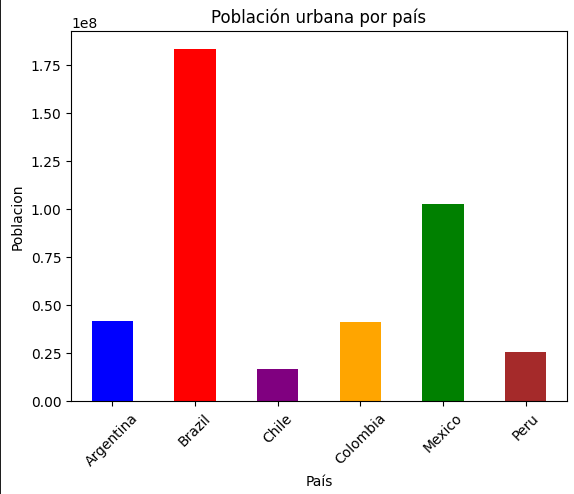
### 5.6 Distribución de la población por edad

Brasil y México concentran la mayor población en todos los grupos etarios. Esto explica su mayor necesidad de vacunas y también una mayor exposición a contagios.



### 5.7 Población urbana por país

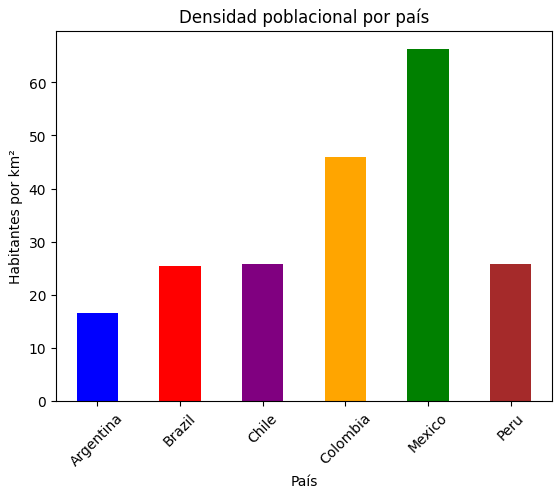
Brasil y México tienen las poblaciones urbanas más grandes, lo que favorece la propagación del virus por la concentración en ciudades.



### 

### 5.8 Densidad poblacional por país

México y Colombia presentan mayor densidad, lo cual eleva el riesgo de contagios. Argentina muestra la densidad más baja.



### 5.9 PIB total

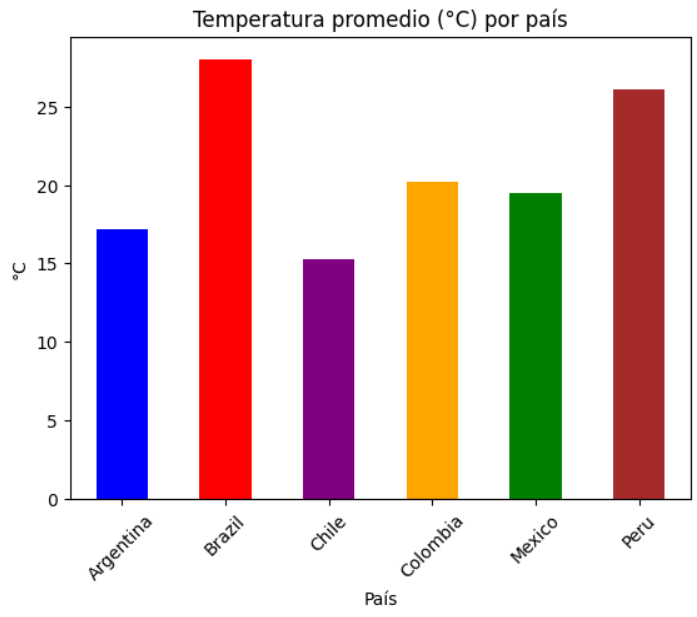
### PIB total: Brasil y México destacan como las economías más grandes.

### 

### 

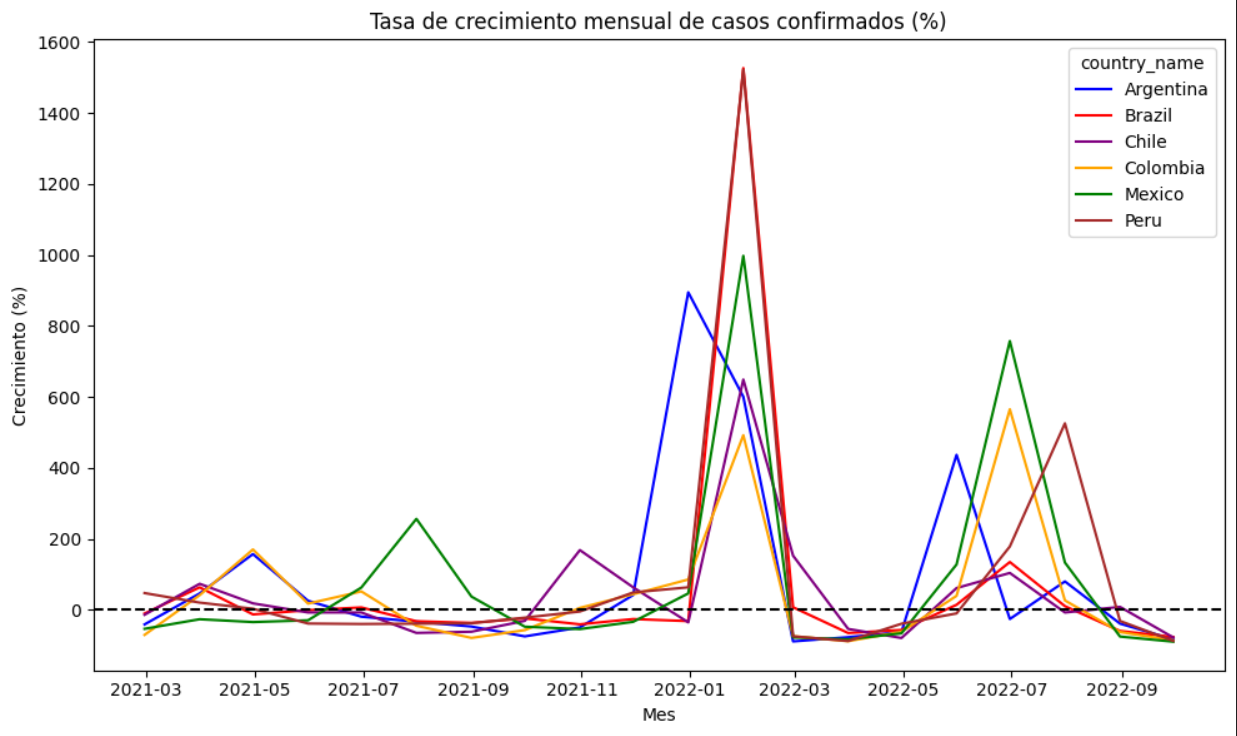
### 5.10 Temperatura promedio por país

Brasil y Perú registran temperaturas más altas, mientras que Argentina y Chile son más fríos, lo que puede influir en la estacionalidad de los brotes.



### 5.11 Tasa de crecimiento mensual de casos confirmados

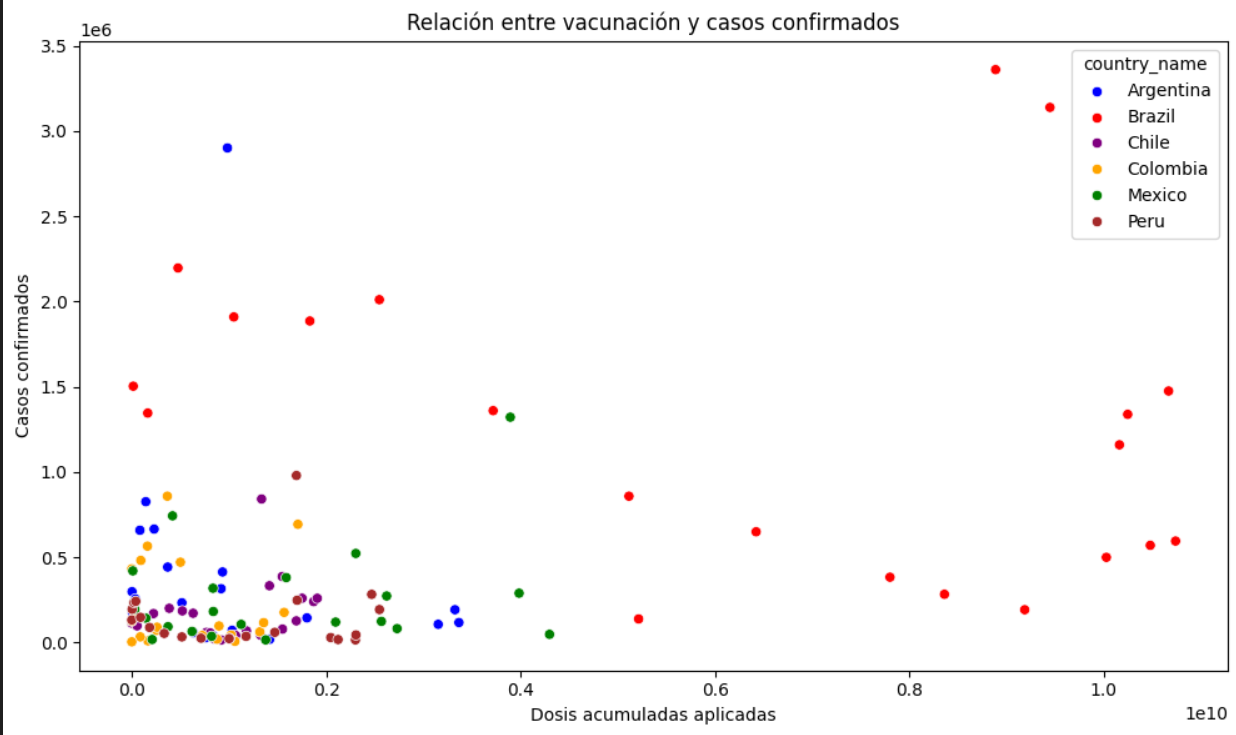
El mayor crecimiento relativo ocurrió a inicios de 2022 en Perú, Brasil y Argentina, coincidiendo con la ola Ómicron.



### 

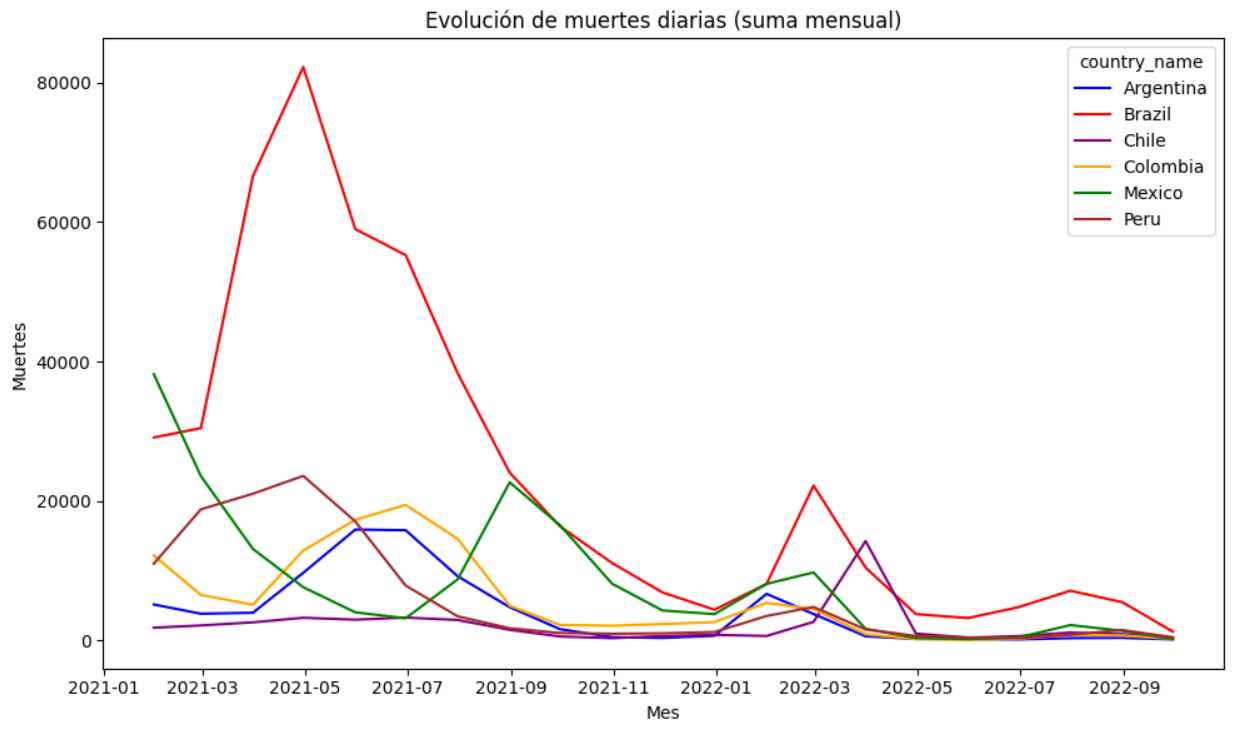
### 5.12 Relación entre vacunación y casos confirmados

Aunque los países con más vacunas (Brasil y México) también presentaron más casos, la reducción de la mortalidad sugiere que la vacunación tuvo un efecto más fuerte en evitar muertes graves que en frenar contagios.



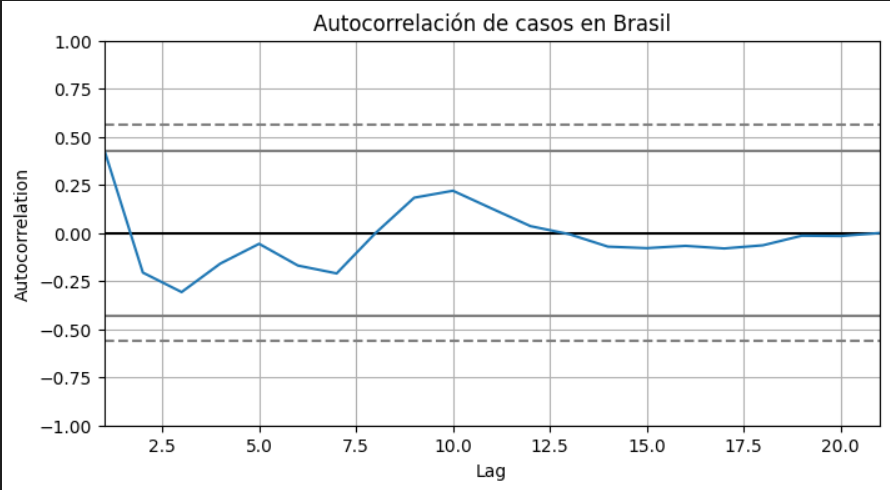
### 5.13 Evolución de muertes diarias (suma mensual)

Las curvas confirman que el número de muertes descendió tras el avance de la vacunación, especialmente en Brasil y Argentina.

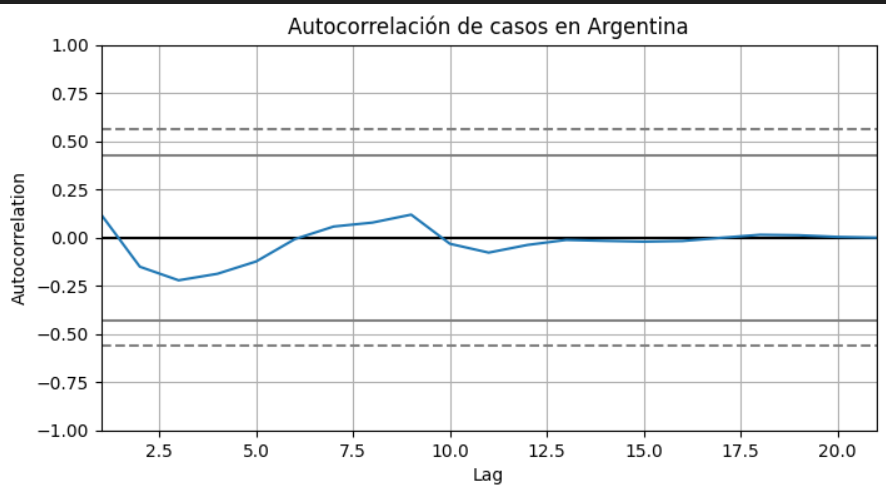


### 5.14 Autocorrelación de casos

* Brasil: los contagios no siguen un patrón predecible a lo largo del tiempo.

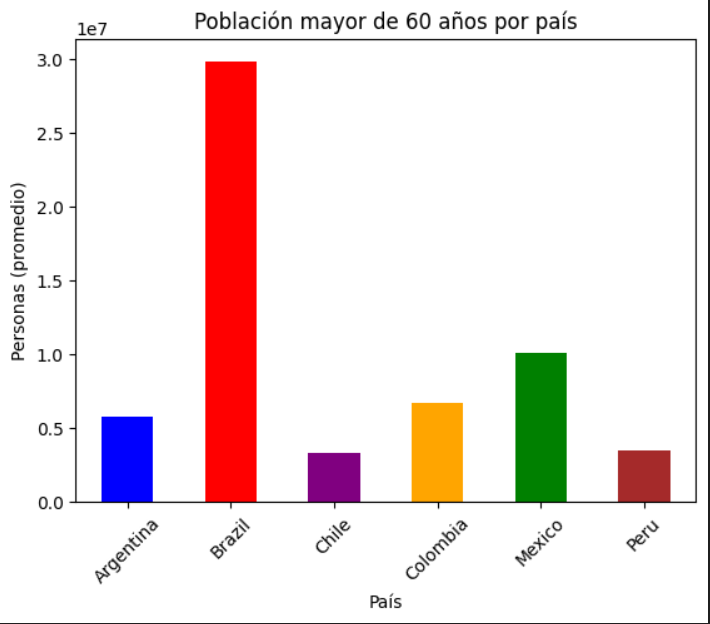


* Argentina: la autocorrelación es débil, reflejando variabilidad en los brotes.



### 5.15 Población mayor de 60 años ( Esto se realizó con la creación de funciones)

Brasil y México concentran la mayor población adulta mayor, explicando en parte sus altas cifras de mortalidad.



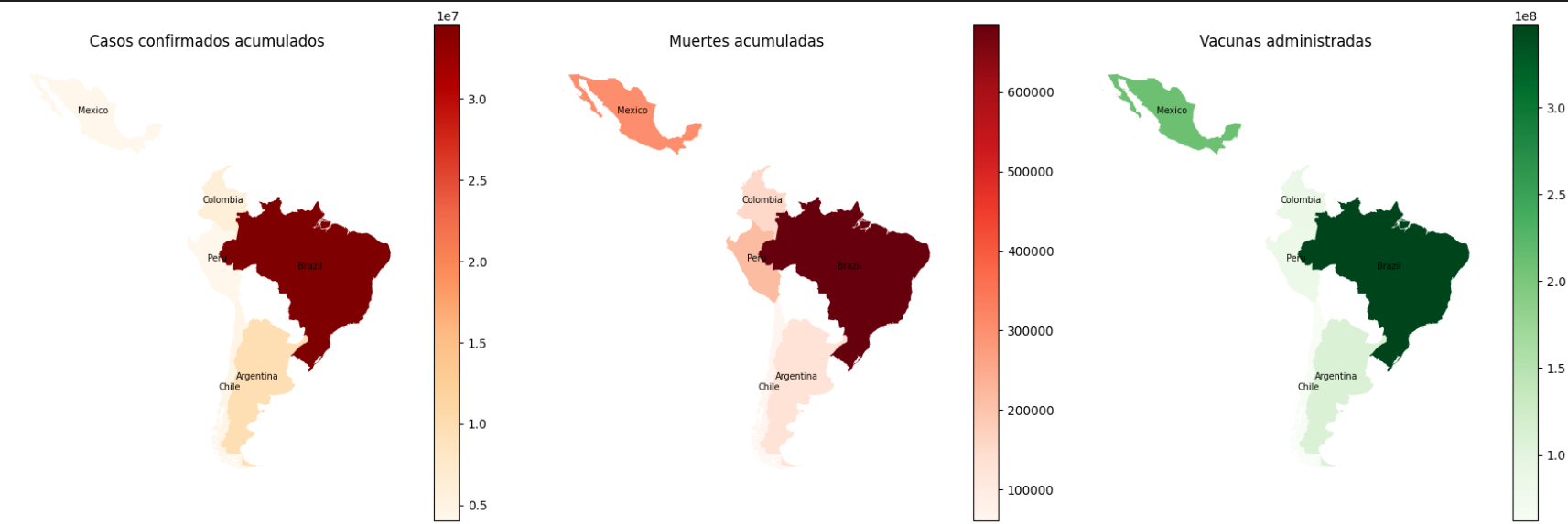
### 5.16 PBI per cápita ( Esto se realizó con la creación de funciones)

Se observa que Chile lidera con el valor más alto, seguido de Argentina y México. En cambio, Colombia y Perú presentan los niveles más bajos, lo que indica una menor capacidad económica individual en comparación con los demás países.

### 

### 5.17 Mapas comparativos

* Casos confirmados acumulados: Brasil lidera ampliamente.
* Muertes acumuladas: Brasil y México destacan.
* Vacunas administradas: nuevamente Brasil y México muestran los mayores volúmenes.



**6. Dashboard en Power BI**

## Metodología de preparación y modelado en Power BI

1. **Conexión del dataset limpio**
   * Importar la base de datos ya depurada en Python/Excel.
   * Verificar la correcta carga en Power BI.
2. **Limpieza final en Power Query**  
   * Colocar los títulos como encabezados.
   * Corroborar que la columna date esté en tipo Fecha.
   * Verificar que las variables numéricas estén en tipo Número entero o Decimal según corresponda.
   * Eliminar valores nulos o reemplazarlos según criterio.
   * Quitar duplicados para asegurar consistencia.

**3. Luego procedemos a desarrollar el tablero:**

El tablero desarrollado en Power BI tiene como objetivo realizar un análisis comparativo de indicadores de salud, vacunación, clima y economía/demografía en países de Latinoamérica, con el fin de identificar patrones y determinar el país más conveniente para la instalación de un laboratorio.

Se integraron fuentes de datos con métricas epidemiológicas, demográficas y socioeconómicas, y se diseñaron medidas DAX que permiten calcular indicadores clave de forma dinámica e interactiva.

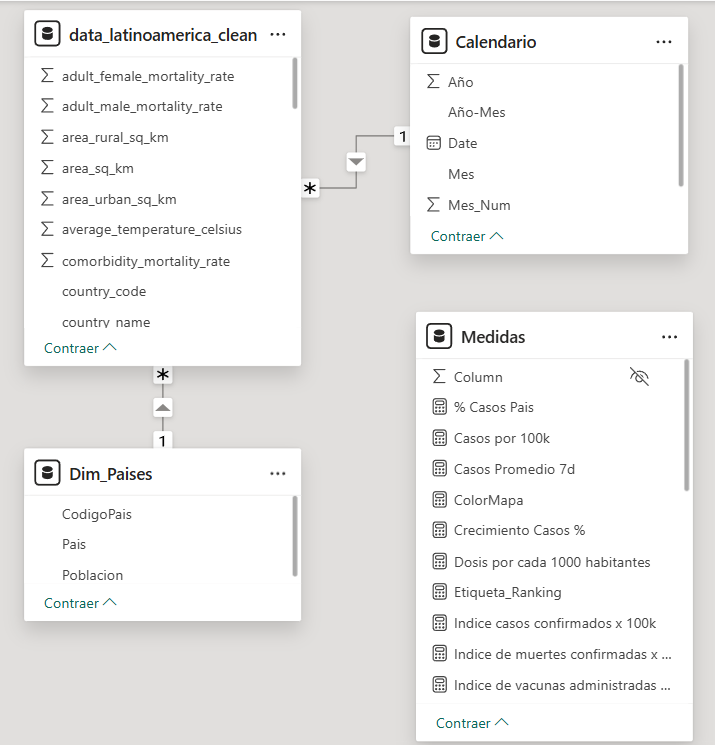
## 4. Estructura del Modelo de Datos

El tablero se apoya en un modelo relacional que integra la tabla de hechos con las dimensiones de país y calendario, garantizando consistencia en los cálculos y la flexibilidad en el análisis temporal y geográfico.

Las relaciones se estructuraron de la siguiente manera:

* Tabla de hechos: data\_latinoamerica\_clean, que contiene información epidemiológica, climática y socioeconómica.
* Dimensión de países: Dim\_Paises, con información descriptiva de cada país (código, nombre, población).
* Dimensión calendario: Calendario, que permite desglosar y analizar los datos por año, mes o fecha específica.
* Carpeta de Medidas: se diseñó para centralizar los indicadores calculados mediante fórmulas DAX, organizados en subcarpetas temáticas (Salud, Vacunas, Clima, Economía y KPIs).

A continuación, se presenta el diagrama del modelo relacional utilizado:



### Se armaron las siguientes medidas con cálculos en DAX para realizar el análisis en Power BI

### 4.1 Clima

Temperatura Promedio: indicador de referencia para analizar la relación entre clima y propagación de enfermedades.

### 4.2 Economía y Demografía

PIB per cápita  
Población promedio de país  
Población total  
Superficie país

4.3 Salud

% Casos País  
Casos por 100k  
Casos Promedio 7d  
Crecimiento de Casos %  
Índice de casos confirmados x 100k  
Índice de muertes confirmadas x 100k  
Muertes por 100k  
Muertes Promedio 7d  
Tasa Mortalidad Diaria (%)  
Total casos confirmados  
Total muertes

### 4.4 Vacunas

Dosis por cada 1000 habitantes  
Índice de vacunas administradas x 100 habitantes  
Vacunas % Población  
Vacunas Administradas

### 4.5 KPIs e Indicadores

Índice Integrador: métrica que combina variables de salud, vacunación y economía.  
Índice Integrador Máximo: referencia para calcular rankings.  
País Más Conveniente: devuelve el país con el mejor puntaje.

## 

## 5. Visualizaciones

El tablero incluye distintos tipos de gráficos y tarjetas que muestran:

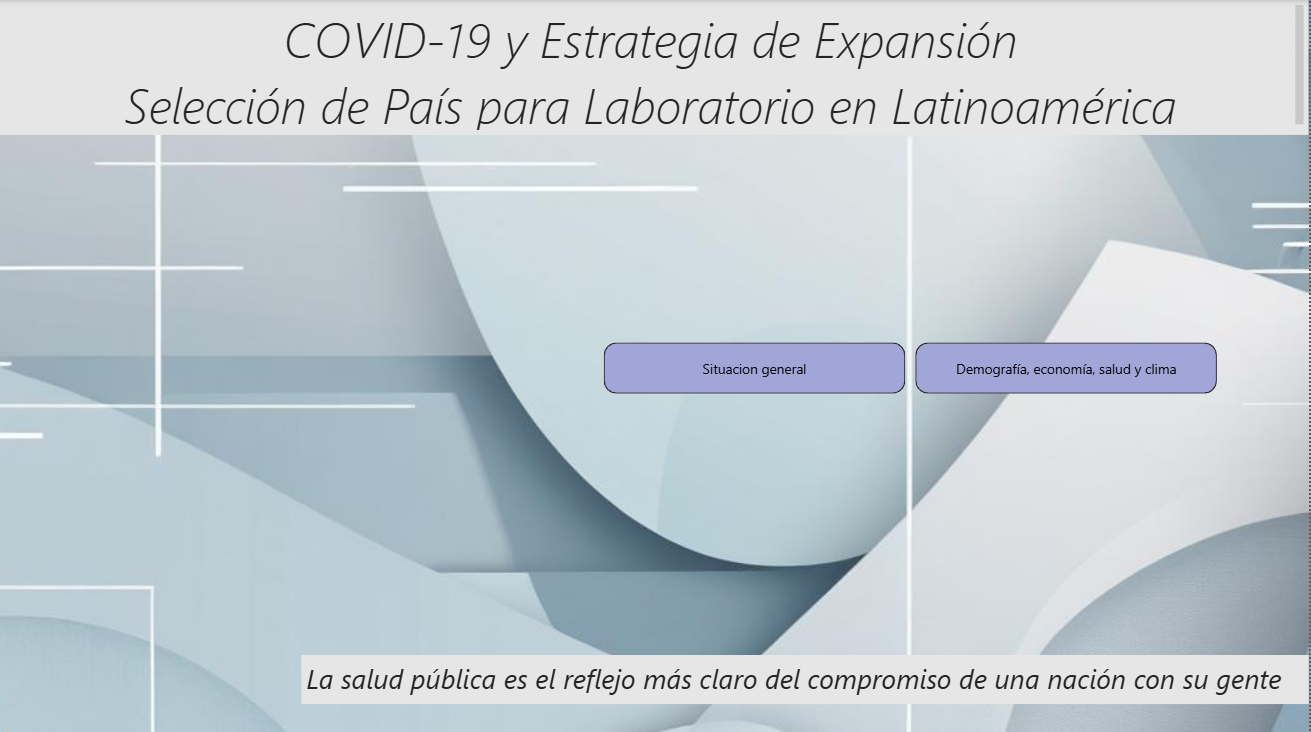
* KPIs principales: casos confirmados, muertes, vacunación.
* Mapas coropléticos: para comparar por país según color o tamaño.
* Rankings y tablas: para visualizar el país mejor posicionado según el índice integrador.
* Gráficos de tendencia: para analizar la evolución de casos, muertes y vacunas a lo largo del tiempo.

En las proximas imagenes se explican con imágenes del tablero realizado

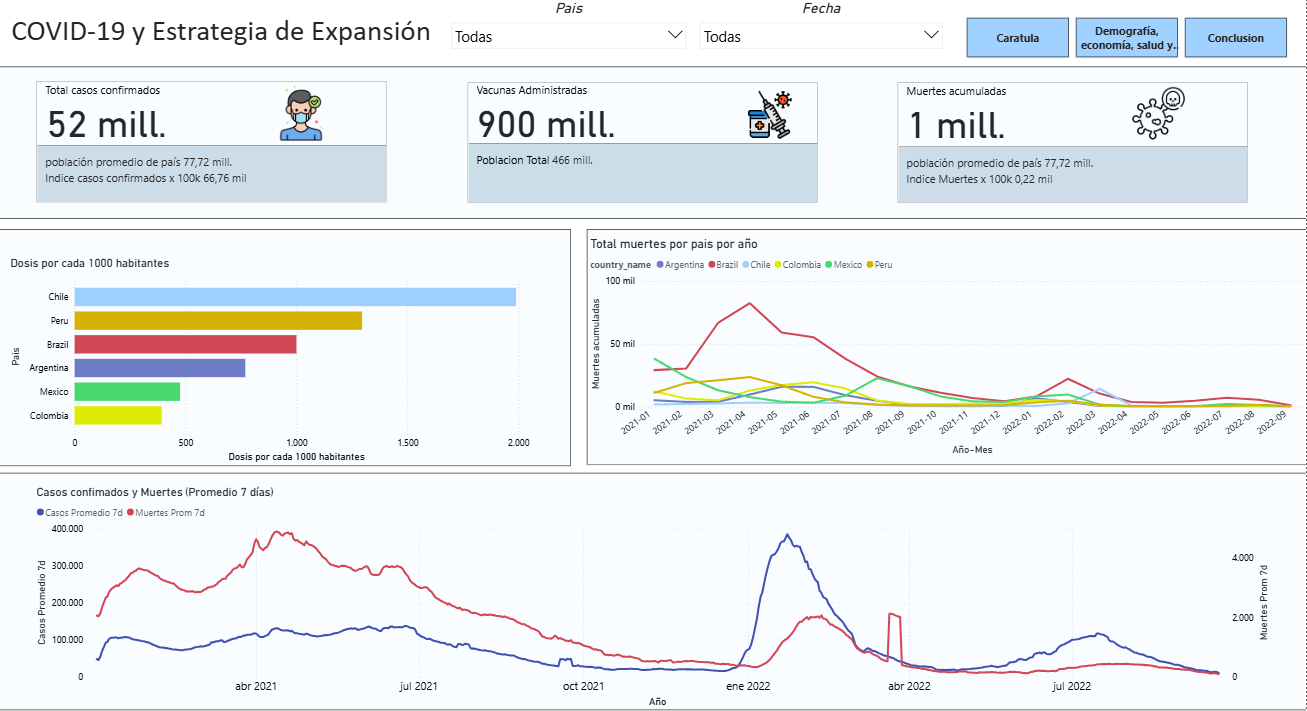
## Carátula

* Se presenta el título del proyecto: “COVID-19 y Estrategia de Expansión: Selección de País para Laboratorio en Latinoamérica”.
* Se incluyen botones de navegación que permiten moverse entre las secciones:  
  Situación general (visión epidemiológica y sanitaria).  
  Demografía, economía, salud y clima (visión integral y análisis comparativo).
* Frase de contexto: “La salud pública es el reflejo más claro del compromiso de una nación con su gente”.

Es el punto de partida, desde aquí navegamos a los dashboards principales.



## Situación general

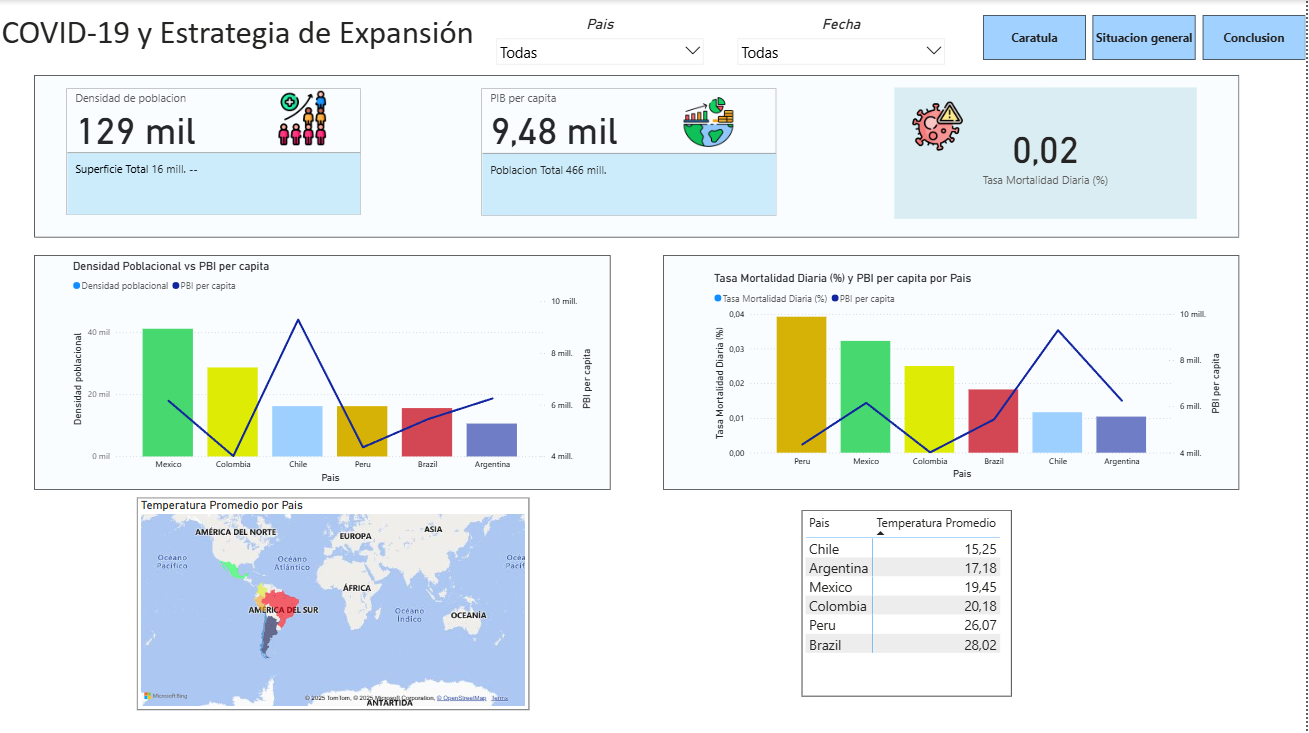


Esta página nos muestra la fotografía sanitaria de la región:

1. KPIs principales (arriba)  
   * Total de casos confirmados: 52 millones.
   * Vacunas administradas: 245 mil millones de dosis.
   * Total de muertes: 1 millón.  
      Cada KPI está acompañado de indicadores derivados como casos por cada 100k habitantes o porcentaje de vacunación sobre la población.
2. Gráfico de barras – Dosis por cada 1000 habitantes  
   * Chile lidera en vacunación proporcional, seguido de Perú y Brasil.
   * Colombia y México aparecen rezagados.
3. Gráfico de líneas – Muertes totales por país  
   * Se visualiza la evolución temporal.
   * Brasil y México presentan picos altos al inicio, mientras que los demás países mantienen curvas más bajas.
4. Gráfico combinado – Casos confirmados y muertes (promedio 7 días)  
   * Línea azul: casos promedio 7d.
   * Línea roja: muertes promedio 7d.
   * Permite observar la correlación entre olas de contagio y la mortalidad asociada.

Interpretación: esta sección ayuda a entender la magnitud de la pandemia, la capacidad de respuesta sanitaria y la eficacia de las campañas de vacunación en cada país.

3) Demografía, economía, salud y clima



Aquí se integran variables que van más allá de lo sanitario:

1. KPIs superiores  
   * Densidad poblacional: 129 mil.
   * PIB per cápita: 9,48 mil.
   * País más conveniente: Chile (según el Índice Integrador).
2. Gráfico – Densidad poblacional vs PIB per cápita  
   * Relaciona riqueza y concentración poblacional.
   * México y Colombia muestran alta densidad; Chile y Argentina, menor densidad pero PIB más sólido.
3. Mapa y tabla – Temperatura promedio por país  
   * Se observa la influencia del clima, con Brasil y Perú más cálidos y Chile/Argentina más fríos.
4. Gráfico – Tasa de mortalidad diaria vs PIB per cápita  
   * Relación inversa: países con menor PIB tienden a mostrar mayor mortalidad diaria.
5. Ranking – País más conveniente para expansión.
   * Se visualiza el Índice Integrador, resultado de combinar PIB, mortalidad y clima.
   * Chile aparece como el país más favorable, seguido de México y Argentina.

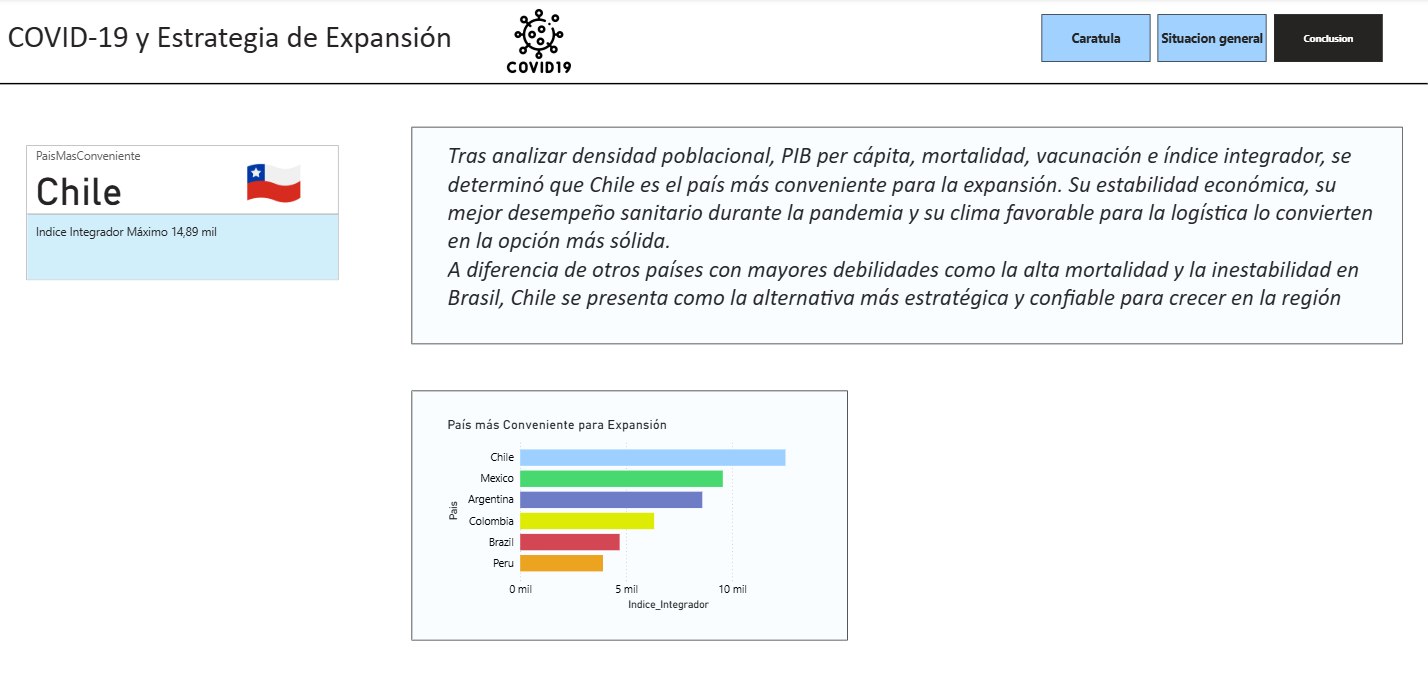
Interpretación: esta sección muestra que no basta con mirar la pandemia, sino que la decisión debe integrar economía, población y condiciones climáticas. Chile surge como el país óptimo porque combina:

* Buen PIB per cápita.
* Mortalidad relativamente baja.
* Temperatura moderada y estable.

Navegación y uso

* Los slicers (arriba) permiten filtrar por país o por fecha para personalizar el análisis.
* Los botones (Carátula, Situación general, Demografía…) funcionan como menú interactivo.
* Cada visualización está vinculada a las medidas DAX creadas, lo que asegura consistencia en los cálculos.

## 6. Conclusiones y recomendaciones



## Argumento económico: PIB per cápita vs. PIB total

* El PIB total refleja el tamaño de la economía de un país, pero no necesariamente indica el nivel de bienestar de su población ni la equidad en el acceso a servicios de salud.
* El PIB per cápita, en cambio, mide el ingreso promedio por habitante y es un mejor indicador de poder adquisitivo, estabilidad social y calidad de vida.
* Chile lidera el PIB per cápita en la región, lo que implica:  
  + Mayor capacidad de inversión en salud y ciencia.
  + Un entorno más sólido para el desarrollo de proyectos de investigación.
  + Estabilidad macroeconómica en comparación con países de mayor PIB total pero con desigualdades profundas (ejemplo: Brasil).

Índice Integrador: permitió sintetizar variables clave en un único indicador comparable.

Elegir Chile significa priorizar un ecosistema más equitativo y eficiente para investigación, en lugar de solo un mercado más grande pero con mayores desigualdades.

### Condiciones climáticas y costos de conservación

* Chile posee un clima templado a frío en varias regiones, lo que representa una ventaja natural para la conservación y transporte de vacunas, que requieren refrigeración constante.
* En países más cálidos como Brasil o Perú, los costos de cadena de frío se elevan significativamente, impactando directamente en la logística y gasto operativo del laboratorio.
* La localización en Chile permite reducir costos energéticos asociados al almacenamiento y mantener estándares de calidad de forma más eficiente.

Esto no solo disminuye costos, sino que además mejora la seguridad de conservación de insumos biomédicos.

### Estabilidad y entorno sanitario

* Chile se destaca por su sistema de salud más organizado en comparación con otros países de la región.
* Ha mostrado altos niveles de cobertura de vacunación y respuesta sanitaria eficiente durante la pandemia.
* Su estabilidad política y regulatoria lo convierte en un entorno atractivo para la inversión a largo plazo.

# 