Proyecto Integrador
BIOGENESYS Farmacéutica

M4 – DATA ANALYTICS 2.0

María Paz Camino – mpcamino@hotmail.com

DA_PT07 | 06-07-2025

Estructura y alcance del proyecto

En el presente trabajo se expone el desarrollo del proyecto integrador que involucra a la empresa farmacéutica **BIOGENESYS**, cuyo objetivo es identificar las ubicaciones óptimas para la expansión de laboratorios farmacéuticos en Latinoamérica, a partir del análisis de datos sobre la incidencia de COVID-19, tasas de vacunación e infraestructura sanitaria. La necesidad surge porque actualmente la empresa no dispone de indicadores claros que faciliten la toma de decisiones estratégicas basadas en datos, lo que resulta clave para responder de forma ágil ante los desafíos de la pandemia y la postpandemia. Como analista de datos, la tarea principal será limpiar, modelar y analizar los datos mediante herramientas como Power BI, generando visualizaciones interactivas que faciliten la obtención de insights para apoyar la expansión en Argentina, Brasil, Chile, Colombia, México y Perú.

El proyecto se desarrollará en cuatro etapas principales:

- Aplicar técnicas de limpieza de datos para asegurar la calidad de los mismos, facilitando el análisis y las decisiones estratégicas.
- 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.
- Mejorar el acceso a los datos mediante operaciones eficientes de extracción, transformación y carga (ETL).
- Desarrollar un Dashboard interactivo con visualizaciones eficientes, permitiendo explorar datos desde múltiples perspectivas para una toma de decisiones informada y estratégica.

En primer lugar, el informe presenta una introducción donde se resume brevemente el proyecto y los objetivos organizacionales alcanzados. Luego, se describe en forma detallada el proceso realizado en cada avance. Posteriormente, se establecen los insights obtenidos en el análisis exploratorio de datos y en las visualizaciones, así como también el análisis del Dashboard. Luego, se exponen los resultados principales junto con recomendaciones de futuras líneas de análisis que requieran atención adicional.

Finalmente, se incluye una reflexión personal sobre lo aprendido durante el proyecto y las habilidades adquiridas como futura Analista de Datos.

BIOGENESYS

FARMACÉUTICA

INFORME FINAL

1. Introducción

El presente informe tiene como objetivo analizar y sintetizar información clave sobre el impacto de la pandemia de COVID-19 en América Latina, con el fin de orientar decisiones estratégicas de expansión de la empresa BIOGENESYS Farmacéutica. A través de técnicas de limpieza, modelado y visualización de datos, se elaboró un conjunto de indicadores que permiten comprender la magnitud de los contagios, la capacidad de respuesta sanitaria y la distribución de variables demográficas y epidemiológicas en seis países: Argentina, Brasil, Chile, Colombia, México y Perú.

El análisis se desarrolló en varias etapas que incluyeron la depuración de un extenso volumen de datos, la exploración de patrones de correlación entre variables y la construcción de un Dashboard interactivo en Power BI que facilita la consulta y comparación de la información. Entre los principales hallazgos se destaca que Brasil y México concentran la mayor cantidad de casos confirmados y fallecimientos, acompañados por una capacidad logística de vacunación significativamente superior al resto de los países. Asimismo, México presenta la prevalencia más elevada de diabetes, lo que implica desafíos y oportunidades adicionales en términos de provisión de tratamientos y programas preventivos.

El conjunto de indicadores y visualizaciones generadas aporta una base objetiva que respalda la recomendación preliminar de priorizar estos dos mercados como destinos estratégicos de expansión. El informe concluye con un compendio de conclusiones, referencias complementarias y reflexiones sobre el proceso de análisis de datos efectuado.

2. Desarrollo del proyecto:

2.1. Primera Etapa: Técnicas de Limpieza

En esta primera fase del proyecto se llevaron a cabo las siguientes consignas:

1- Lectura del archivo *Readme.txt* y análisis de las columnas y datos que se van a utilizar para obtener un mayor conocimiento del dataset.

2- Creación del notebook llamado “PIDA_M4_Camino_Maria_Paz.ipynb”. Importación de las librerías que se necesitan para realizar el 1° avance del PI.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

3- Lectura archivo data_latinoamerica.csv con código Python en Visual Studio Code.

```
# Ruta al archivo CSV-Se lo va a llamar Latinoamérica. Acá se podrían haber filtrado las columnas y parseo de fechas.
df_Latinoamerica = pd.read_csv("data_latinoamerica.csv")

# Mostrar primeras filas en formato tabla
df_Latinoamerica.head(10)
```

4- Comprobación de cantidad de registros y columnas especificadas

```
# Cantidad de filas y columnas
filas, columnas = df_Latinoamerica.shape
print(f"El dataset tiene {filas} filas y {columnas} columnas.")

# Ver detalle de las columnas
print(df_Latinoamerica.columns)

# Ver el tipo de dato en las columnas
print(df_Latinoamerica.dtypes)
```

5- Selección países donde se expandirán: Colombia, Argentina, Chile, México, Perú y Brasil.

```
# Filtrado de Paises donde se quieren expandir
paises_LATAM = ['Colombia', 'Argentina', 'Chile', 'Mexico', 'Peru', 'Brazil']
filtro = df_Latinoamerica['country_name'].isin(paises_LATAM)

df_Latinoamerica_seleccion = df_Latinoamerica[filtro]

df_Latinoamerica_seleccion.head(5)

# Cantidad de filas y columnas del nuevo dataset
filas, columnas = df_Latinoamerica_seleccion.shape
print(f"El dataset filtrado tiene {filas} filas y {columnas} columnas.")
```

6- Filtro de datos: Fechas mayores a 2021-01-01.

```
# Conversion columna Date de string a datetime
df_Latinoamerica_seleccion["date"] =
pd.to_datetime(df_Latinoamerica_seleccion["date"])

# Verificacion de que los datos son fechas
print(df_Latinoamerica_seleccion["date"].dtypes)

# Filtrar todas las fechas mayores a 2021-01-01
filtro = df_Latinoamerica_seleccion["date"] > "2021-01-01"
df_Latinoamerica_mayor_2021 = df_Latinoamerica_seleccion[filtro]

# Ver resultado
print(f"El dataset filtrado tiene {df_Latinoamerica_mayor_2021.shape[0]} registros y
{df_Latinoamerica_mayor_2021.shape[1]} columnas.")

# Ver los cambios en las primeras 5 filas
df_Latinoamerica_mayor_2021.head(5)
```

7- Compara a nivel de país para llenar valores faltantes.

```
# Análisis de calidad de los datos - Columnas con mas de 4MM de registros
df_Latinoamerica_mayor_2021.isnull().sum()[df_Latinoamerica_mayor_2021.isnull().sum(
) > 4000000]

# Nuevo filtro por location Key
codigo_pais = ['AR','CL','CO','MX', 'BR', 'PE']
filtro = df_Latinoamerica_mayor_2021['location_key'].isin(codigo_pais)
df_Latinoamerica_mayor_2021_locationkey = df_Latinoamerica_mayor_2021[filtro]

df_Latinoamerica_mayor_2021_locationkey.head(10)

# Ver cantidad de registros y columnas
filas, columnas = df_Latinoamerica_mayor_2021_locationkey.shape
print(f"El dataset tiene {filas} filas y {columnas} columnas.")

# Ver nuevamente los nulos
df_Latinoamerica_mayor_2021_locationkey.isnull().sum()
```

8- Limpieza preliminar de los datos, eliminando registros nulos y corrigiendo los tipos de datos donde sea necesario, tratados con valores hacia adelante y hacia atrás.

```
# Completar los valores faltantes con el valor anterior (forward fill) por país
for col in df_Latinoamerica_mayor_2021_locationkey.columns:
    if df_Latinoamerica_mayor_2021_locationkey[col].isna().sum() > 0 and col != 
'country_name':
        df_Latinoamerica_mayor_2021_locationkey[col] = (
            df_Latinoamerica_mayor_2021_locationkey
            .groupby('country_name')[col]
            .transform(lambda x: x.fillna(method='ffill')))

# verifico que ninguna columna tenga nulos
df_Latinoamerica_mayor_2021_locationkey.isnull().sum()

# Rellenar cumulative_vaccione_doses_administrered con ffill + bfill por país
df_Latinoamerica_mayor_2021_locationkey["cumulative_vaccine_doses_administered"] = (
    df_Latinoamerica_mayor_2021_locationkey
    .groupby("country_name")["cumulative_vaccine_doses_administered"]
    .transform(lambda x: x.fillna(method='ffill').fillna(method='bfill')))

# Rellenar new_recovered con ffill + bfill por país
df_Latinoamerica_mayor_2021_locationkey["new_recovered"] = (
    df_Latinoamerica_mayor_2021_locationkey
    .groupby("country_name")["new_recovered"]
    .transform(lambda x: x.fillna(method='ffill').fillna(method='bfill')))

# Rellenar cumulative_recovered con ffill + bfill por país
df_Latinoamerica_mayor_2021_locationkey["cumulative_recovered"] = (
    df_Latinoamerica_mayor_2021_locationkey
    .groupby("country_name")["cumulative_recovered"]
    .transform(lambda x: x.fillna(method='ffill').fillna(method='bfill')))

# verifico que ninguna columna tenga nulos
df_Latinoamerica_mayor_2021_locationkey.isnull().sum()

# Se rellenan con 0 aquellas columnas que aun tienen nulos - Todos los datos son
nulos para un pais
df_Latinoamerica_mayor_2021_locationkey["new_recovered"] =
df_Latinoamerica_mayor_2021_locationkey["new_recovered"].fillna(0)
df_Latinoamerica_mayor_2021_locationkey["cumulative_recovered"] =
df_Latinoamerica_mayor_2021_locationkey["cumulative_recovered"].fillna(0)
```

9- Examen de las características básicas del dataset para comprender la distribución de las variables clave como incidencia de COVID-19 e identificación de aquellas que se consideran claves para el análisis.

```
df_Latinoamerica_mayor_2021_locationkey.head(5)
df_Latinoamerica_mayor_2021_locationkey.info()
# Análisis estadístico por columna
print(df_Latinoamerica_mayor_2021_locationkey.describe())
```

10- Guardado de datos filtrados en un archivo con el nombre DatosFinalesFiltrado.csv. Luego de otros cambios se guardó un nuevo archivo llamado DatosFinalesFiltrado1.csv.

```
df_Latinoamerica_mayor_2021_locationkey.to_csv('DatosFinalesFiltrado.csv')
print('Datos filtrados correctamente')

df_Latinoamerica_mayor_2021_locationkey.to_csv('DatosFinalesFiltrado1.csv')
print('Datos filtrados correctamente')
```

11- Aplicación bucles for y/o while para el cálculo de estadísticas descriptivas y otras métricas importantes que ofrece pandas por default.

```
# Análisis estadístico por columna
for i in df_Latinoamerica_mayor_2021_locationkey.columns:
    print(i)
    print(df_Latinoamerica_mayor_2021_locationkey[i].describe())
    print('-----------------------------------------------------')
```

¿Qué implican estas métricas y cómo pueden ayudar en el análisis de datos?

Nos ayudan a entender la distribución de los datos (mín., máx., media, desviación estándar, etc.), detectar outliers y evaluar la consistencia interna de cada variable.

¿Se muestran todas las estadísticas en todas las columnas durante el análisis?

No, el describe y el bucle para variables numéricas no aplica para todas. Por defecto, variables categóricas quedan excluidas.

¿Cuál es la razón de la respuesta anterior y cómo podría afectar la interpretación de los resultados obtenidos?

Porque conceptos como media, desviación estándar o rango no tienen sentido para datos categóricos (string). Si solo se miran los números, se pierde de vista cómo están distribuidas las categorías (Países, fechas) lo que puede sesgar las conclusiones.

12- Creación de una función que permite obtener la mediana, varianza y el rango.

```
def obtener_estadisticas(columna):
    # Devuelve la mediana, varianza y rango de una serie de datos numéricos
    mediana = columna.median()
    varianza = columna.var()
    rango = columna.max() - columna.min()
    return mediana, varianza, rango

# luego se usaría asi
columna_vacunas_administradas =
df_Latinoamerica_mayor_2021_locationkey['cumulative_vaccine_doses_administered']
mediana, varianza, rango = obtener_estadisticas(columna_vacunas_administradas)

print(f'Mediana: {mediana}')
print(f'Varianza: {varianza}')
print(f'Rango: {rango}')
```

¿Qué representa la mediana?

La mediana es el valor central de la distribución.

¿Cómo varía la dispersión de los datos en el conjunto de datos analizado, en términos de la varianza y el rango?

La varianza y el rango indican qué tan dispersos están los datos alrededor de la mediana. Una varianza alta y un rango amplio implican gran variabilidad.

¿Qué nos puede indicar esto sobre la consistencia o la variabilidad de los datos en relación con la mediana?

Si la varianza y el rango son elevados, la variable presenta una distribución amplia y más dispersa. Si son bajos, los datos son más consistentes alrededor de la mediana.

2.2. Conclusión de la Etapa 1: Limpieza y Transformación de Datos

En esta primera etapa, se partió de un dataset inicial que contenía 12.216.057 registros y 50 columnas, lo que requería una cuidadosa selección y preparación para garantizar la calidad y relevancia de la información.

A través de diversos filtros, primero se acotó el alcance geográfico para considerar únicamente los países de interés en Latinoamérica (Colombia, Argentina, Chile, México, Perú y Brasil), lo que redujo el dataset a 11.970.289 registros y 50 columnas. Posteriormente, se limitó el análisis temporal a registros con fecha posterior al 1 de enero de 2021, alcanzando un total de 7.537.296 registros y 50 columnas.

Se aplicó además un último filtrado para garantizar que cada registro estuviera asociado a una clave de ubicación (Location_key), lo que resultó en un dataset final de 3.744 registros y 50 columnas. De esta manera, se logró trabajar sobre una base clara, acotada y preparada para un análisis más preciso.

A nivel de calidad de datos, se implementó una estrategia de imputación para garantizar la consistencia interna del dataset:

- Se reemplazaron los datos nulos utilizando el dato anterior y/o posterior correspondiente a cada país.
- Aquellos registros que no disponían de información fueron asignados a 0 para mantener la integridad de la base.

Por último, la versión final del dataset filtrado y transformado se guardó bajo el nombre DatosFinalesFiltrado.csv.

3. EDA e Insights

En este apartado se presentan los principales hallazgos obtenidos a partir del análisis exploratorio de datos (EDA). Mediante el uso de técnicas estadísticas y visualizaciones, se identificaron patrones y relaciones relevantes que servirán de base para orientar las decisiones estratégicas sobre la expansión de laboratorios farmacéuticos y centros de vacunación.

3.1. Barplot: Variables clave por país

Se realizaron 12 gráficos de barras comparativos entre países que permiten visualizar variables clave para el análisis estratégico. A continuación, se detallan las principales observaciones con los valores aproximados:

- **Población total**: Brasil presenta la mayor cantidad de habitantes, superando los 210 millones, seguido de México con aproximadamente 110 millones, y Colombia y Perú con 50 y 29 millones respectivamente. Argentina alcanza unos 45 millones, mientras que Chile se sitúa en torno a los 20 millones.

- **Población por género**: La distribución es equilibrada en todos los países, con proporciones similares entre hombres y mujeres.

- **Índice de Desarrollo Humano (IDH):** Los valores son parejos y elevados, en un rango de 0,75 a 0,85, siendo Chile y Argentina los de mayor índice y México y Perú ligeramente por debajo.

- **Población por grupo etario:** En todos los países se observa un predominio de adultos (30-69 años), con un volumen que varía entre 20 y 120 millones según el país. Los niños y adolescentes representan un grupo importante, mientras que los adultos mayores constituyen el segmento más reducido.

- **Casos confirmados acumulados de COVID-19**: Brasil lidera con más de 27 millones de casos, seguido por Argentina con alrededor de 8 millones y Colombia con aproximadamente 5 millones. México y Perú se sitúan en valores intermedios, cercanos a 4 millones cada uno.

- **Muertes acumuladas**: Brasil alcanza el registro más alto con 490 mil fallecimientos, seguido por México con cerca de 178 mil, Perú con 120 mil, Colombia con 115 mil y Argentina con valores menores, alrededor de 80 mil.

- **Prevalencia de tabaquismo**: Chile es el país con mayor prevalencia, con más de 35%, seguido por Argentina con cerca de 20%.

- **Prevalencia de diabetes**: México presenta la tasa más elevada, cercana al 13%, seguido de Brasil con un valor próximo al 11%.

- **Dosis acumuladas de vacunas administradas**: Brasil supera los 348 millones de dosis, México alcanza alrededor de 210 millones, mientras que Argentina y Colombia aplicaron un promedio de 90 millones cada uno.

- **Temperatura promedio anual**: Brasil registra la temperatura más alta, cercana a los 27 °C, seguido de Colombia y México con unos 21 °C, mientras que Argentina y Chile presentan valores inferiores, alrededor de 13 °C.

- **Población rural vs urbana**: En todos los países predomina la población urbana. Brasil cuenta con una población urbana superior a 170 millones, México con más de 90 millones y Argentina con cerca de 40 millones de habitantes urbanos.

- **GDP per cápita (USD):** Chile se destaca como el país con mayor producto bruto per cápita, superando los 15.mil USD, seguido de Argentina y México con valores de 10 mil, Brasil 9 mil USD, y Perú con alrededor de 7 mil USD.

Estas visualizaciones, presentadas en la Figura 1, permiten identificar de manera integrada principalmente patrones demográficos y socioeconómicos. La información obtenida constituye una base objetiva y detallada que contribuirá a la decisión al momento de seleccionar las ubicaciones estratégicas para la instalación de nuevos laboratorios y centros de vacunación.

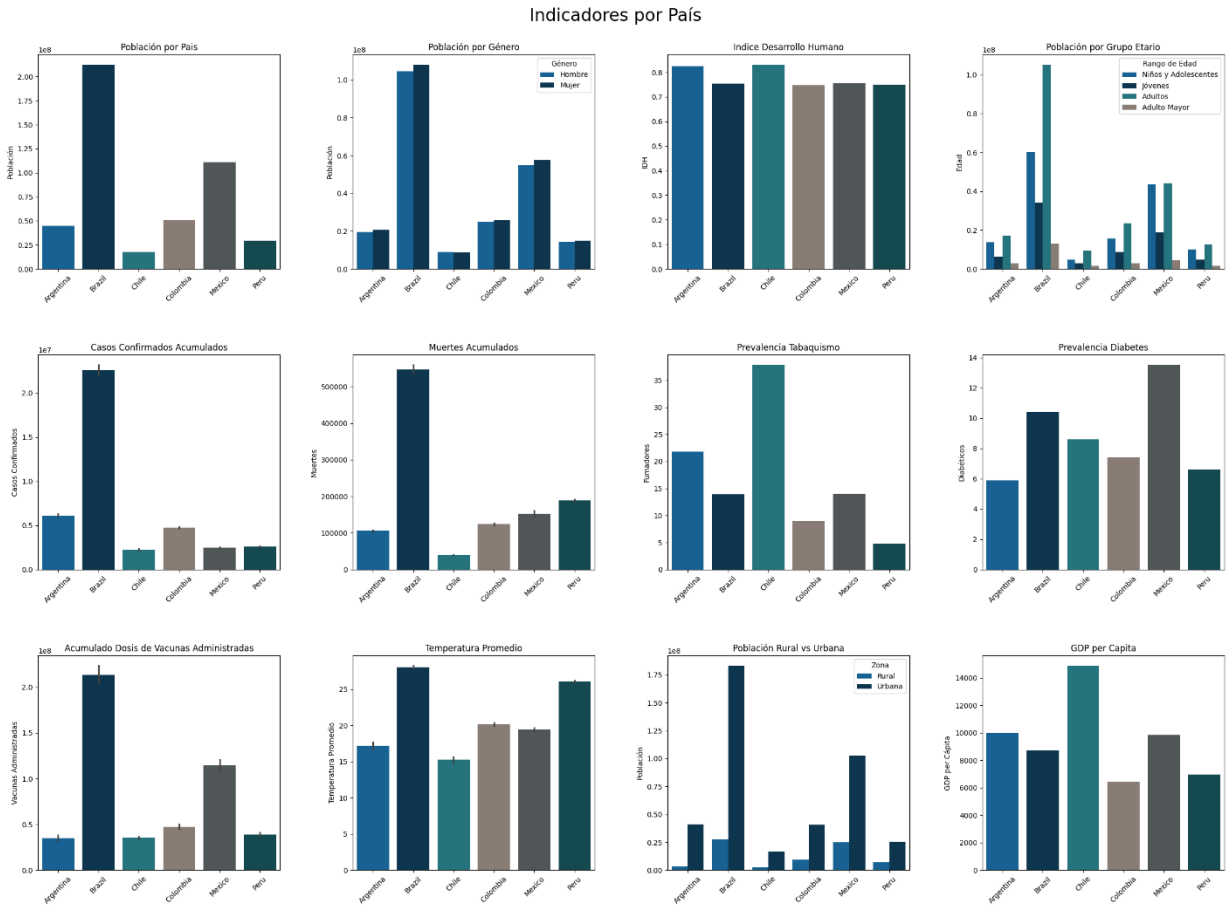


Figura 1: Barplots de variables clave por país

3.2. Heatmap: Matriz de correlación (Valores > 0.5)

Se calculó la matriz de correlación de Pearson (Figura 2) entre algunas de las variables numéricas consideradas clave, aplicando una submáscara para mostrar solo el triángulo inferior dado que la correlación es simétrica. De este modo se facilita la lectura evitando duplicaciones. Asimismo, se filtraron los valores con un coeficiente absoluto mayor a 0,5, lo que permite centrar el análisis en las relaciones más relevantes y evitar el ruido de correlaciones débiles. A continuación, se detallan las principales observaciones

- **Confirmados acumulados – Muertes acumuladas** (0.90): Muy alta correlación positiva. Los países con más casos confirmados tienen más muertes.

- **Población – Confirmados acumulados** (0.79) y Muertes acumuladas (0.82): Fuerte correlación positiva. A mayor población, más casos y fallecimientos, indicando un efecto de escala poblacional.

- **Diabetes – Población** (0.60): Relación positiva moderada, posiblemente vinculada con el tamaño poblacional y prevalencia proporcional.

- **Temperatura promedio – Muertes acumuladas** (0.56): Correlación positiva moderada. Puede reflejar un patrón climático relacionado con la distribución geográfica de los países más afectados.

- **GDP per cápita – IDH** (0.82): Alta correlación positiva, consistente con el hecho de que el PBI per cápita es un componente del Índice de Desarrollo Humano.

- **Tabaquismo – GDP per cápita (0.96) y Tabaquismo – IDH (0.89):** Relaciones muy fuertes y positivas. Sugiere que los países con mayor desarrollo e ingreso presentan prevalencia de tabaquismo más alta.

- **IDH – Temperatura promedio** (-0.56): Correlación negativa moderada. Los países con IDH más alto tienden a tener temperaturas medias más bajas.

- **Tabaquismo – Temperatura promedio** (-0.54): Correlación negativa moderada. Puede reflejar factores culturales y geográficos.

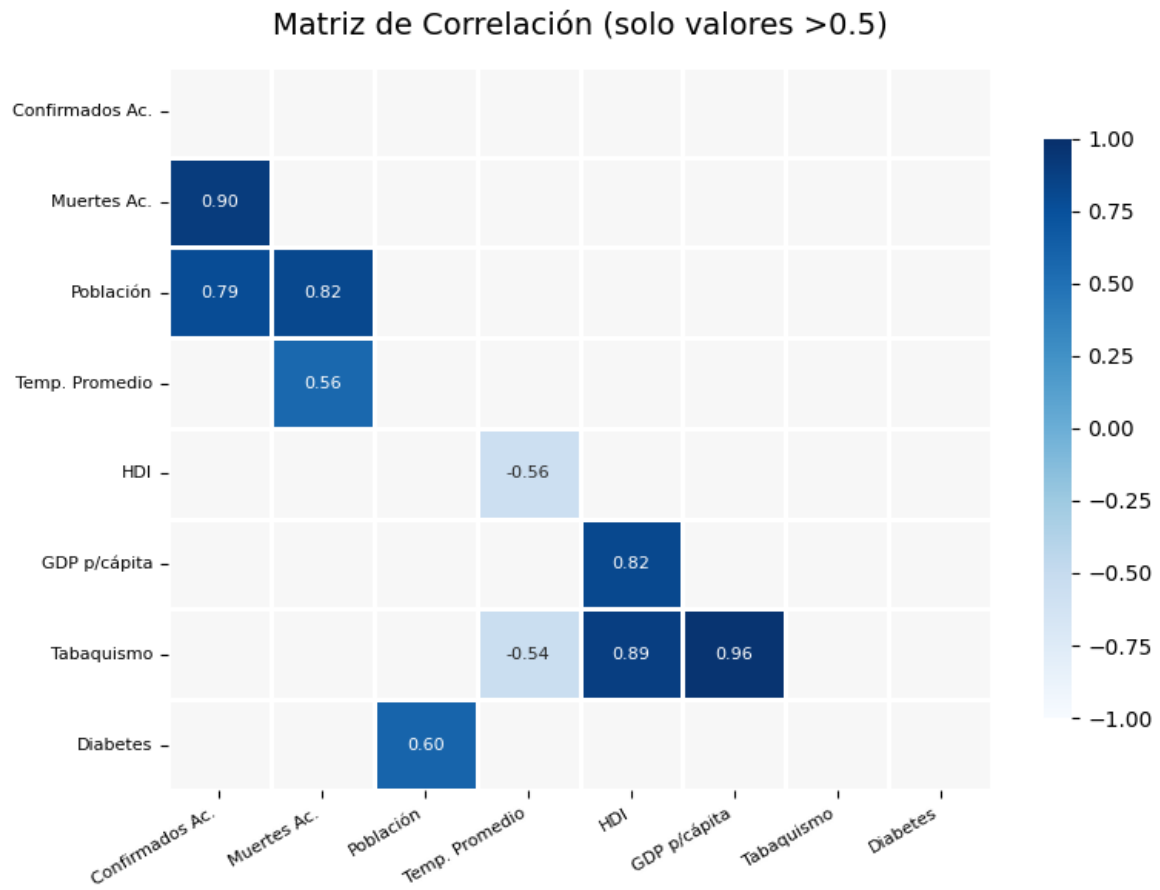


Figura 2: Heatmap: Matriz de correlación de variables críticas

3.3. Histogramas: Variables clave que tienen cambios en los valores

Los histogramas muestran la distribución de las variables clave analizadas:

- **Población, casos confirmados, muertes y vacunas administradas** presentan distribuciones sesgadas a la derecha, indicando que la mayoría de los países tienen valores bajos o medios, con pocos casos de valores muy altos (Ej., Brasil).

- **IDH y GDP** per cápita muestran concentraciones en rangos específicos, con menor dispersión (IDH entre ~0,75 y 0,83).

- **Tabaquismo y diabetes** presentan distribuciones multimodales, reflejando diferencias marcadas entre países.

- **Temperatura promedio** se distribuye de manera más uniforme, con tendencia a valores intermedios (~20 °C).

Estos patrones (Figura 3) permiten identificar la heterogeneidad entre países y ayudan a contextualizar los análisis comparativos.

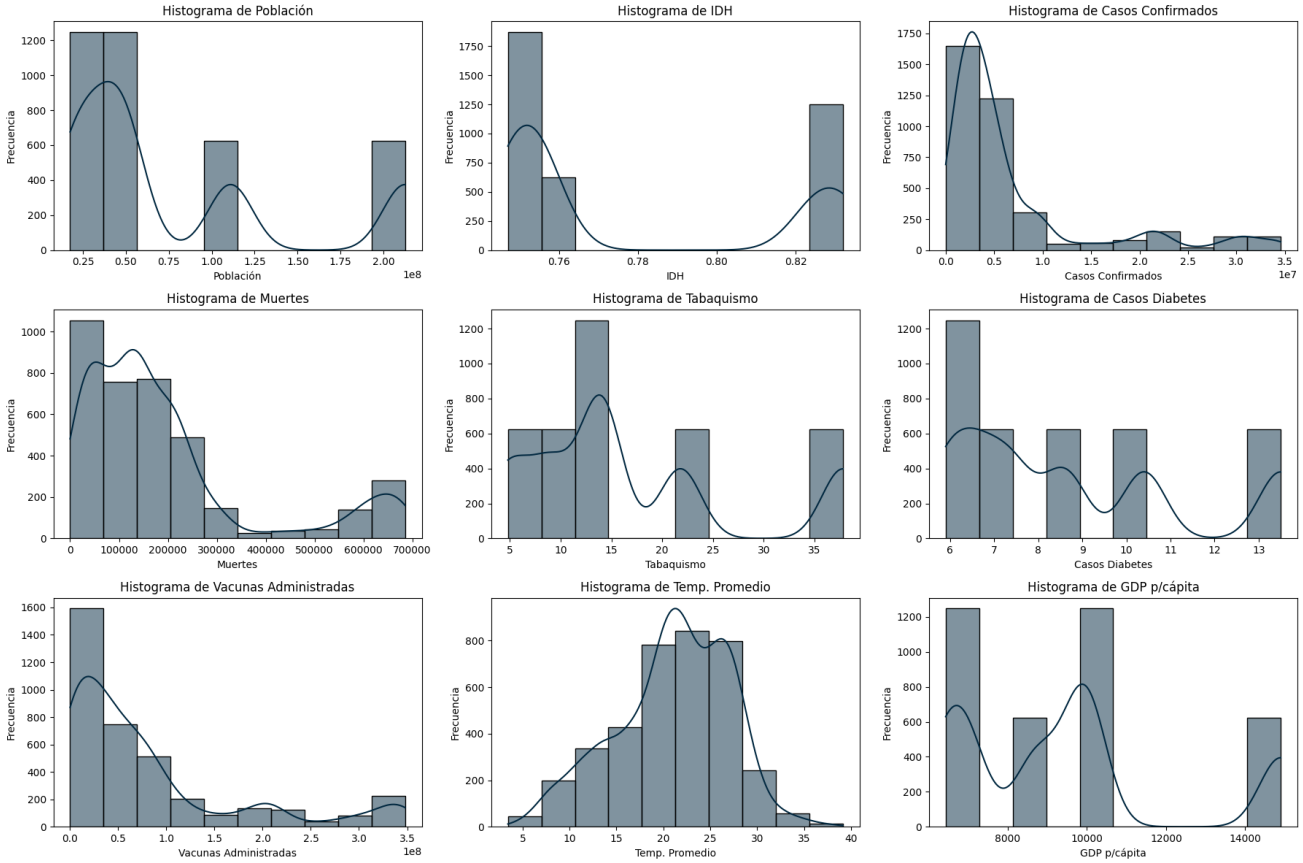


Figura 3: Histograma de variables clave

3.4. Scatterplot: Temperaturas medias vs casos confirmado / muertes

En estos gráficos de dispersión (Figura 4) se analiza la relación entre la temperatura promedio y el número de casos confirmados (izquierda) y de muertes (derecha). No se observa una tendencia lineal clara entre temperatura y casos o muertes.

La mayoría de los registros se concentran en temperaturas medias entre 20 °C y 30 °C, con un rango amplio de valores de casos y muertes. Esto sugiere que la temperatura promedio, en este nivel agregado, no es un predictor directo del volumen de contagios o decesos. Para países con temperaturas menores a 20 °C, la cantidad de casos y muertes tiende a mantenerse en rangos más bajos.

Outliers relevantes: en Casos Confirmados, se identifican varios puntos con más de 250.000 casos, asociados principalmente a Brasil y México, que tienen poblaciones más grandes. En Muertes, hay un registro por encima de 10.000 muertes que resalta como un outlier, probablemente vinculado a Brasil, dado su peso poblacional y epidemiológico.

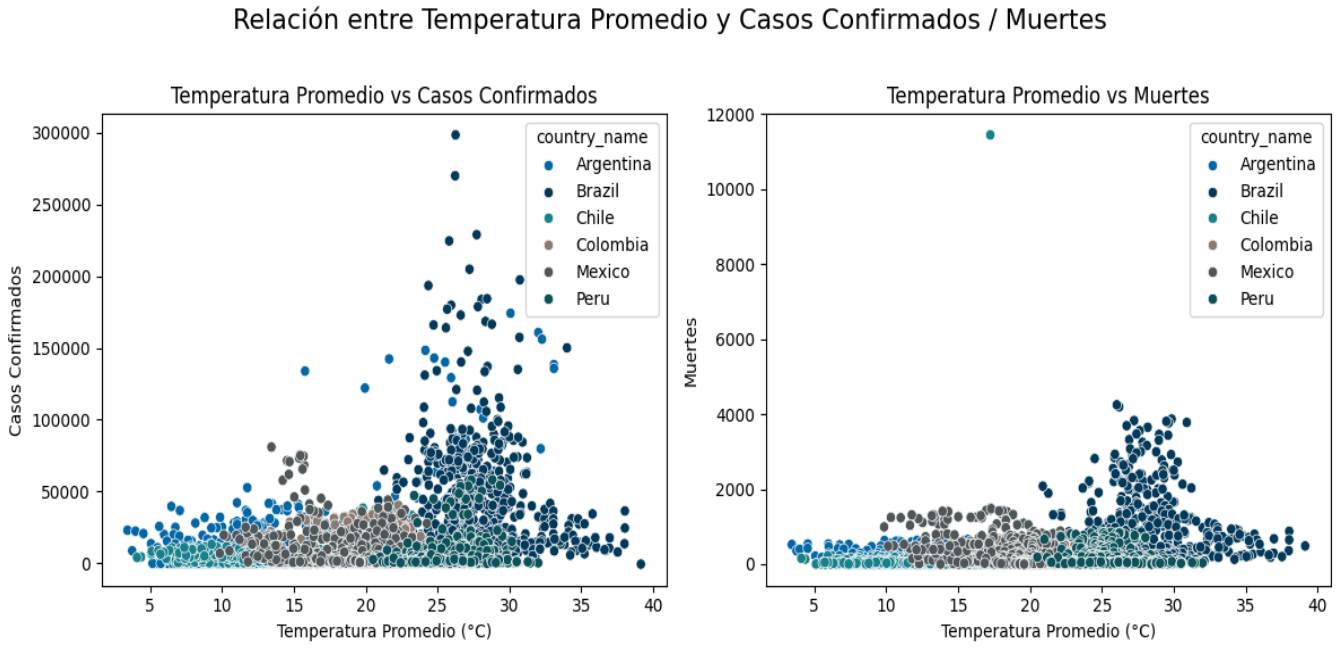


Figura 4: Scatterplot Temperaturas promedio vs Casos Confirmados / Muertes

3.5. Barplot: Dosis Totales de Vacunas Administradas por País (Valor Máximo)

El gráfico 5 muestra que Brasil lidera ampliamente en cantidad de dosis administradas (alrededor de 350 millones), seguido por México con aproximadamente 210 millones. Argentina ocupa el tercer lugar, mientras que Chile registra el menor número de dosis aplicadas entre los países comparados. Estas diferencias reflejan probablemente no solo, el tamaño poblacional sino también, la capacidad logística de vacunación.

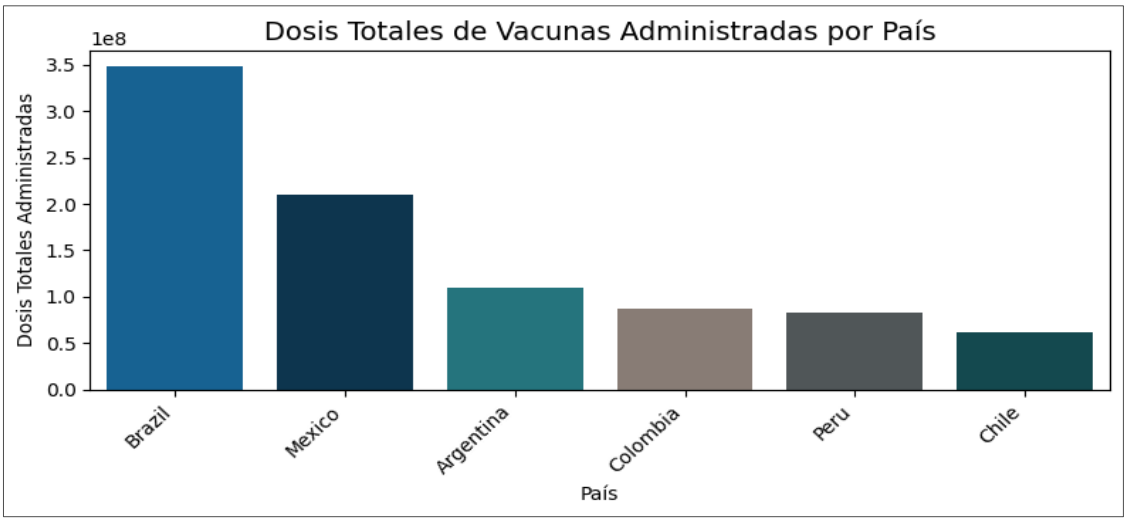
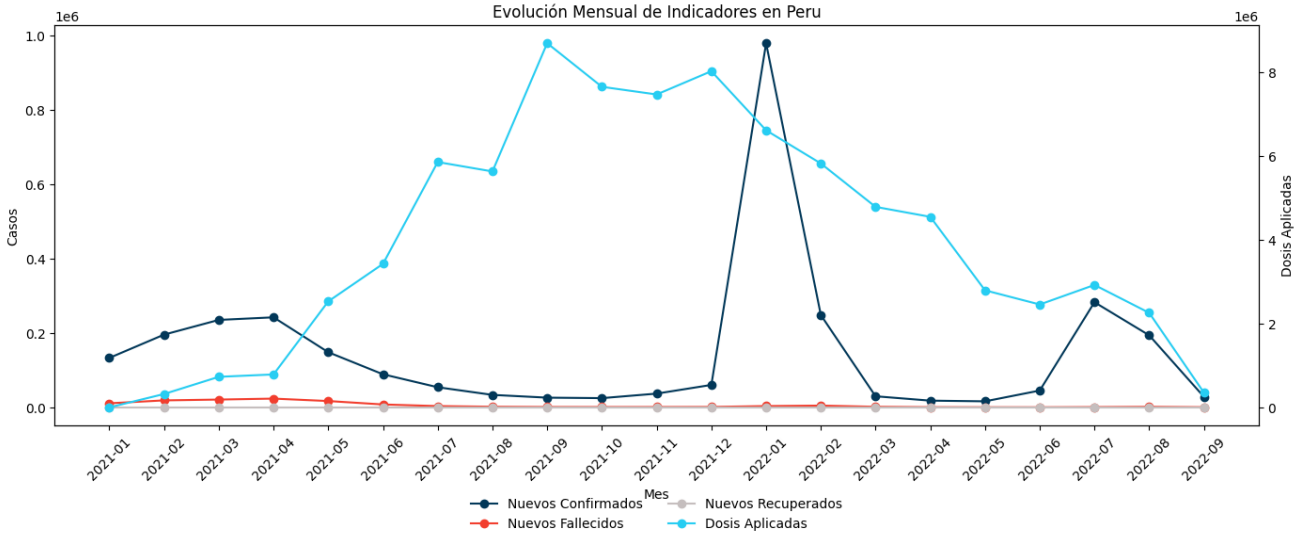
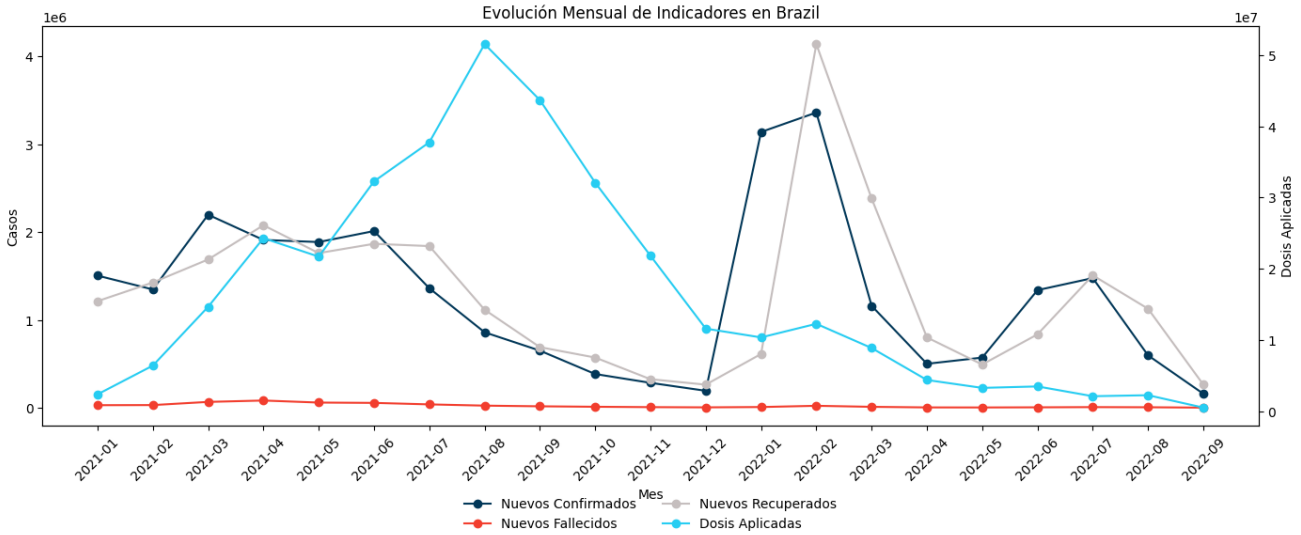


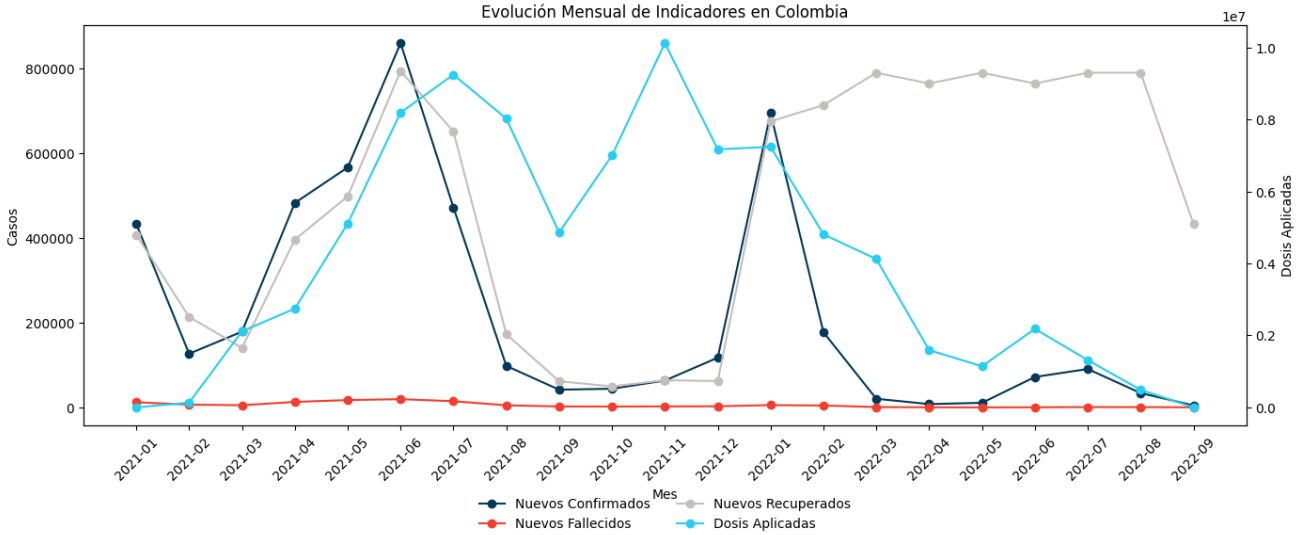
Figura 5: Scatterplot Temperaturas promedio vs Casos Confirmados / Muertes

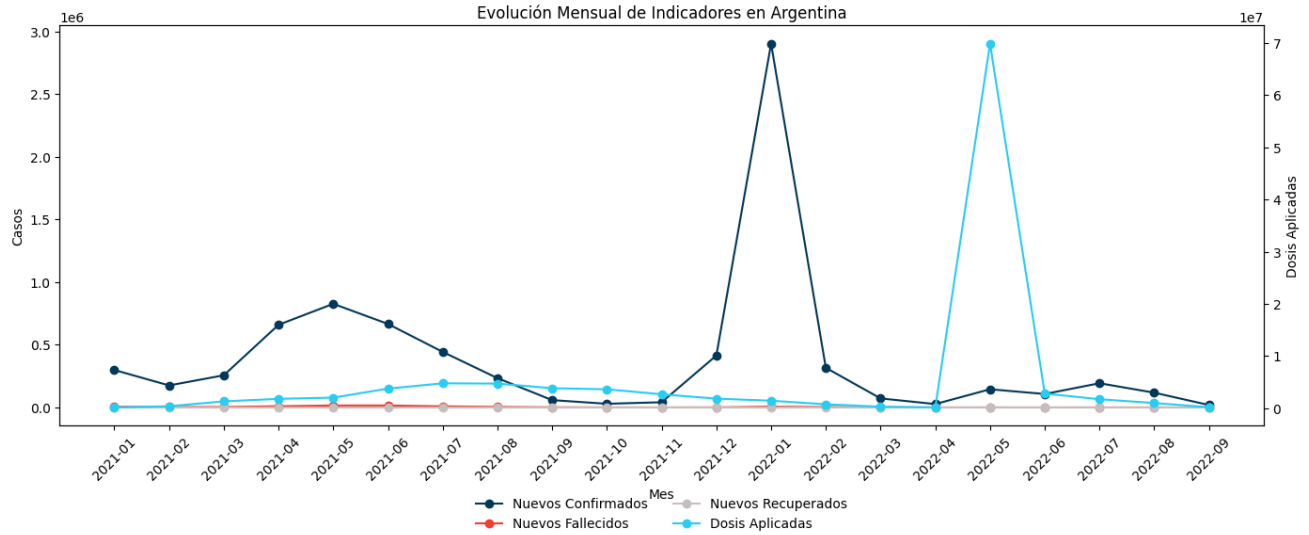
3.6. Lineplot: Dosis administradas, casos confirmados, muertes y recuperados por país por mes

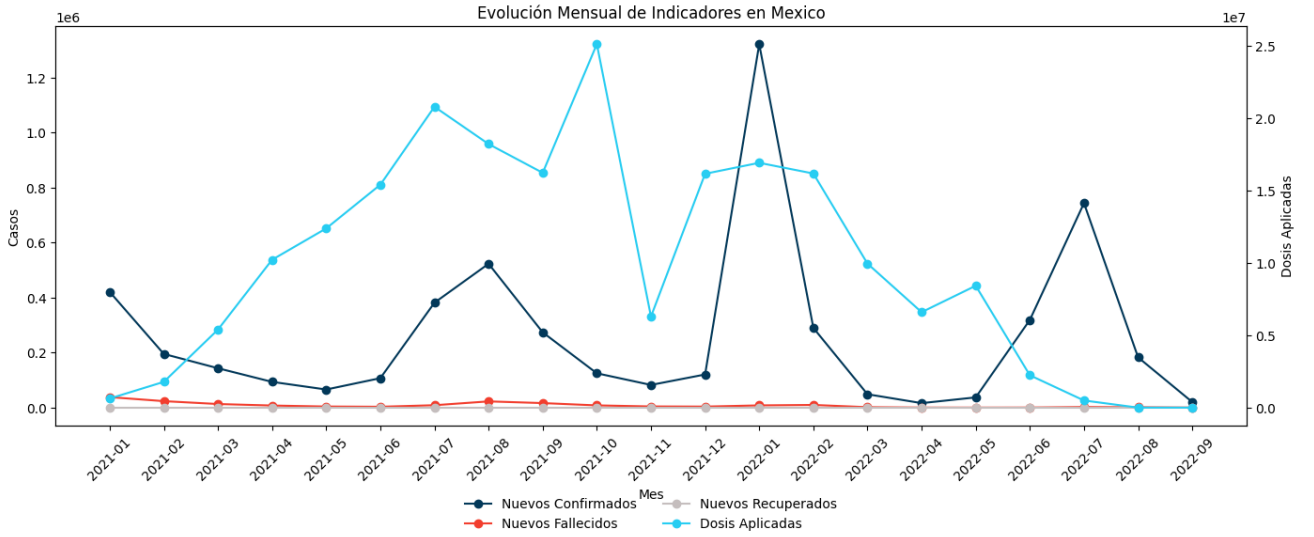
Se realizaron gráficos (Figura 6) de líneas por país (Perú, Brasil, Colombia, Argentina, México y Chile) que permiten comparar la evolución mensual de los principales indicadores de la pandemia. En cada caso, se visualizan en simultáneo los nuevos casos confirmados, los nuevos recuperados, los nuevos fallecidos y la cantidad de dosis de vacunas aplicadas, lo que facilita identificar tendencias, picos de contagios y períodos de vacunación más intensos.

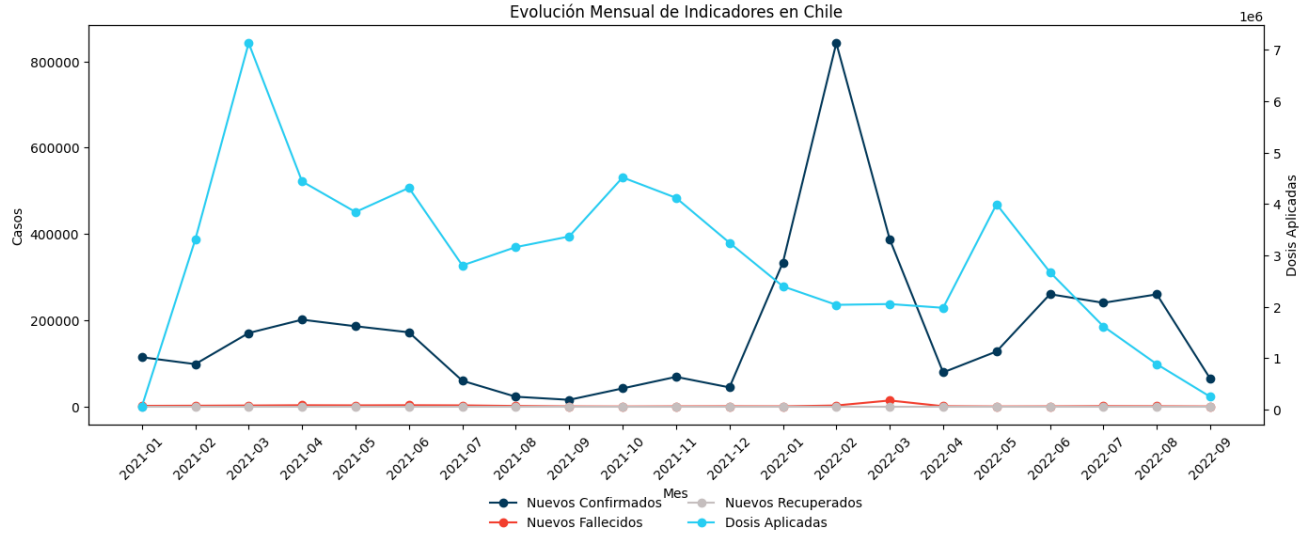

Evolución Mensual de Indicadores en Peru
1e6
1e6
Casos
Dosis Aplicadas
Mes
Nuevos Confirmados
Nuevos Recuperados
Nuevos Fallecidos
Dosis Aplicadas


Evolución Mensual de Indicadores en Brazil
1e6
1e7
Casos
Dosis Aplicadas
Mes
Nuevos Confirmados
Nuevos Recuperados
Nuevos Fallecidos
Dosis Aplicadas


Evolución Mensual de Indicadores en Colombia
1e7
Casos
Dosis Aplicadas
Mes
Nuevos Confirmados
Nuevos Recuperados
Nuevos Fallecidos
Dosis Aplicadas


Evolución Mensual de Indicadores en Argentina
1e6
1e7
Casos
Dosis Aplicadas
Mes
Nuevos Confirmados
Nuevos Recuperados
Nuevos Fallecidos
Dosis Aplicadas


Evolución Mensual de Indicadores en Mexico
1e6
1e7
Casos
Dosis Aplicadas
Mes
Nuevos Confirmados
Nuevos Recuperados
Nuevos Fallecidos
Dosis Aplicadas


Evolución Mensual de Indicadores en Chile
1e6
Casos
Dosis Aplicadas
Mes
Nuevos Confirmados
Nuevos Recuperados
Nuevos Fallecidos
Dosis Aplicadas

A continuación se detalla lo que puede observarse en cada gráfico en particular.

- **Perú**: Los nuevos confirmados muestran picos en enero y julio de 2022. Los recuperados acompañan la tendencia de contagios en menor volumen. En tanto, los fallecidos se mantienen estables en niveles bajos. Las dosis aplicadas crecen sostenidamente hasta fines de 2021, evidenciándose luego un descenso progresivo.

- **Brasil**: Los casos confirmados presentan máximos en marzo de 2021 y enero de 2022, con caídas posteriores. Los recuperados siguen una tendencia similar a los contagios. Los fallecidos se mantienen estables, con leves repuntes en los picos de casos. Las dosis aplicadas crecen sostenidamente durante 2021, con un aumento pronunciado a inicios de 2022.

- **Colombia**: Los confirmados tienen picos en junio de 2021 y enero de 2022 y los recuperados siguen una evolución similar. Los fallecidos presentan una curva estable, con leves repuntes en picos de contagio. Las dosis aplicadas crecen rápidamente hasta mediados de 2021 y luego se estabilizan en valores altos.

- **Argentina**: Los casos confirmados muestran picos marcados en enero de 2022 y julio de 2022. Los recuperados acompañan la evolución de los contagios. Los fallecidos permanecen relativamente bajos en comparación con los casos. Las dosis aplicadas se concentran en dos grandes periodos de aplicación en 2021 y mediados de 2022.

- **México**: Los nuevos confirmados presentan varios picos, destacando uno importante en enero de 2022. Los recuperados siguen un patrón similar al de los contagios. Los fallecidos se mantienen en niveles bajos y estables. Las dosis aplicadas tienen un aumento constante durante 2021 y oscilaciones en 2022.

- **Chile**: Los confirmados alcanzan un máximo en enero de 2022. Los recuperados muestran la misma dinámica que los contagios. Los fallecidos mantienen niveles bajos y estables. Las dosis aplicadas presentan un crecimiento rápido en la primera mitad de 2021 y disminuyen paulatinamente después.

En términos generales, Brasil y México fueron los países que presentaron los mayores volúmenes de nuevos casos confirmados a lo largo del período analizado, mientras que Chile y Perú registraron los niveles más bajos de contagios. Esta misma tendencia se observa en la cantidad de nuevos recuperados, donde Brasil concentró la mayor cantidad de casos y Perú y Chile se ubicaron con cifras considerablemente menores. Respecto a los nuevos fallecidos, Brasil nuevamente se destacó por mantener los valores más altos de mortalidad mensual, en contraste con Chile y Argentina, que mostraron las líneas más bajas. Finalmente, al analizar las dosis de vacunas aplicadas, Brasil lideró con la mayor cantidad administrada en casi todos los meses, seguido de México, mientras que Chile y Perú fueron los países con menos dosis totales aplicadas en el período observado.

3.7. Boxplot: Temperaturas promedio por país

El boxplot (Figura 7) muestra que Brasil y Perú concentran las temperaturas medias más elevadas, con medianas cercanas a 27 °C, mientras que Argentina y Chile presentan valores más bajos, alrededor de 16–17 °C, y mayor dispersión en los datos. Colombia y México tienen rangos intermedios, con medianas próximas a 20 °C. Se observa la presencia de varios outliers, especialmente en Brasil, que reflejan registros extremos de temperatura media en algunas regiones.

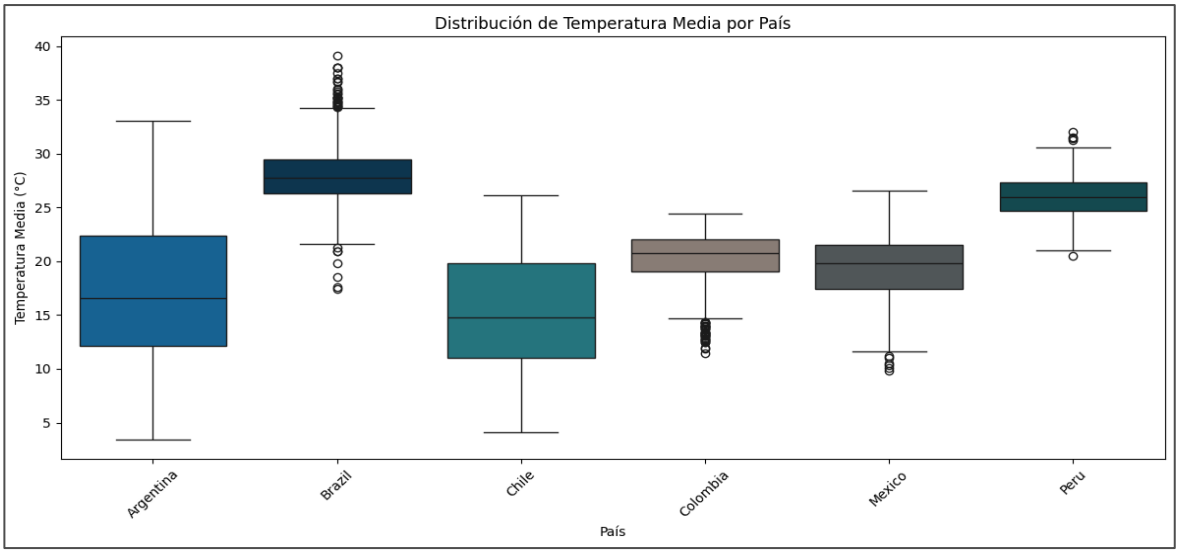


Figura 7: Boxplot Temperaturas promedio por país

Al relacionar la distribución de la temperatura media con los gráficos de muertes y casos confirmados analizados previamente, no se observa una asociación clara ni directa entre temperaturas más altas y un mayor o menor número de fallecimientos. Por ejemplo, Brasil, con temperaturas medias elevadas, presenta la mayor cantidad de muertes, pero Perú, que también muestra temperaturas altas, registra cifras de mortalidad más bajas en comparación. Esto coincide con lo evidenciado en la matriz de correlación, donde la relación entre temperatura promedio y muertes acumuladas era solo moderada (0,56), sugiriendo que la temperatura podría ser uno de muchos factores que interactúan, pero no un determinante principal del impacto de la pandemia en los países analizados.

3.8. Heatmap: Mapa de calor de Métricas por país

El mapa de calor (Figura 8) muestra que Brasil es el país con valores máximos en población, casos confirmados y fallecidos (todas en 1), mientras que Chile se destaca por el máximo IDH y GDP per cápita (ambos en 1). Argentina exhibe un valor alto en IDH (0,92), aunque en el resto de métricas se mantiene en rangos intermedios o bajos. México alcanza el valor máximo en prevalencia de diabetes (1) y un nivel medio en población. Perú resalta por una temperatura media elevada (0,85), pero en la mayoría de las métricas presenta valores reducidos. Colombia no muestra predominio claro en ninguna variable, manteniéndose en rangos medios o bajos.

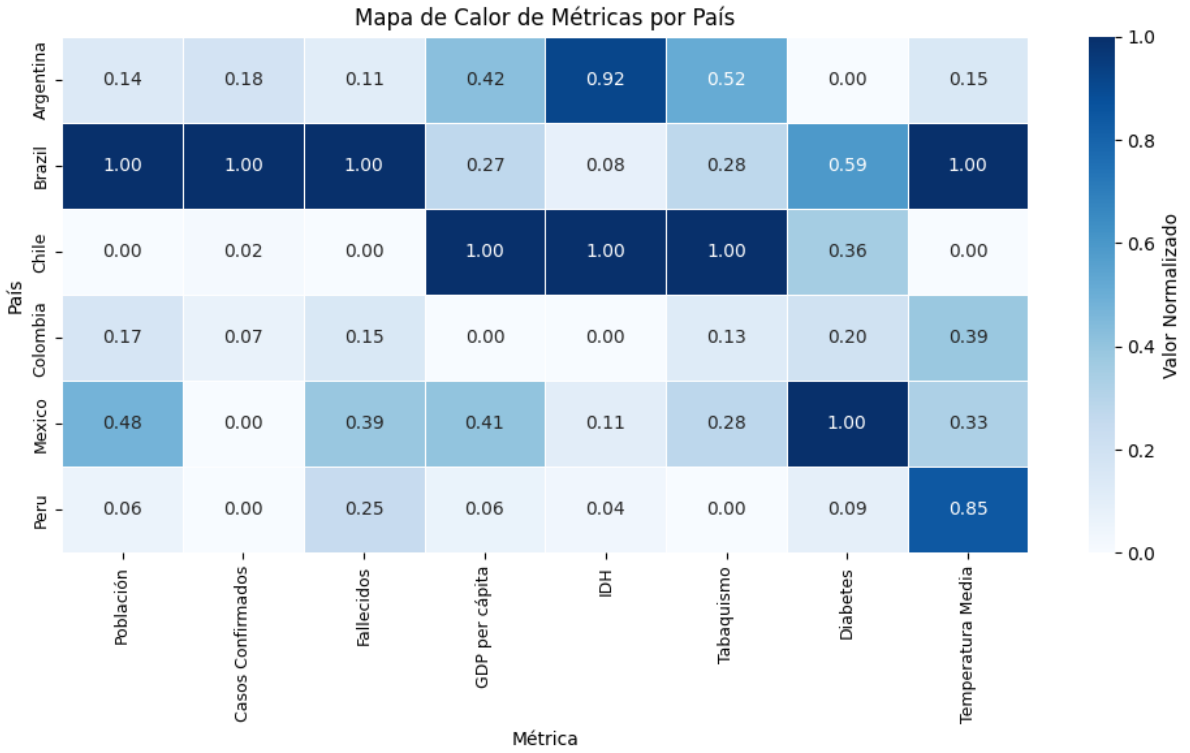


Figura 8: Heatmap:Métricas por país

3.9. Barplot (apiladas): Tasa de mortalidad adulta femenina y masculina por país

El gráfico (Figura 9) muestra que México y Brasil presentan las tasas de mortalidad adulta más altas tanto en hombres como en mujeres, mientras que Chile registra los valores más bajos. En todos los países, la mortalidad masculina supera a la femenina de manera consistente.

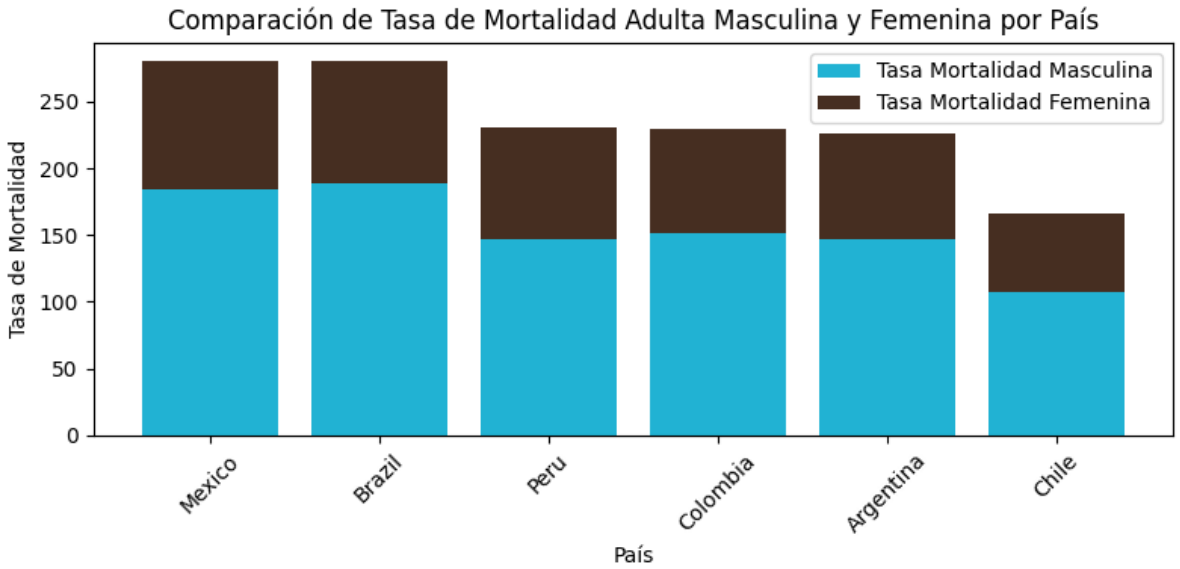


Figura 9: Barplot apiladas: Tasa de mortalidad adulta por género por país

3.10. Barplot (agrupadas): Prevalencia de diabetes y tasa de mortalidad adulta por país

El gráfico (Figura 10) muestra que México lidera en prevalencia de diabetes y tiene, junto con Brasil, las tasas promedio de mortalidad adulta más altas. En contraste, Chile presenta la menor tasa de mortalidad, pese a mantener una prevalencia intermedia de diabetes.

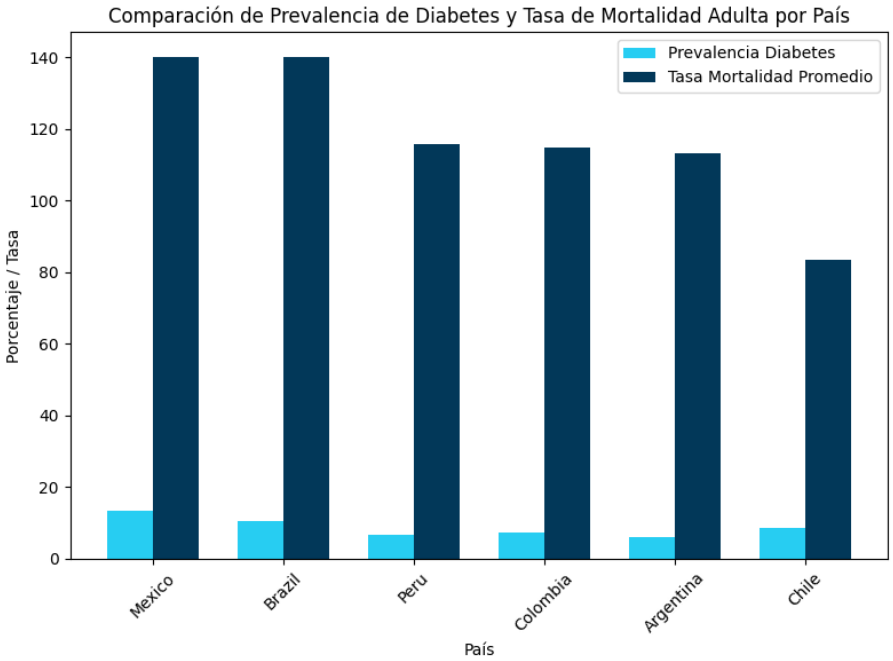


Figura 10: Barplot: Diabetes y Tasa de mortalidad adulta por país

3.11. Lineplot: Evolución trimestral de Nuevos Casos Confirmados y Recuperados (Brasil, México y Argentina).

Brasil muestra consistentemente la mayor cantidad de casos confirmados, con un pico marcado en el primer trimestre de 2022. México y Argentina tienen volúmenes mucho menores, también con aumentos en el mismo período. Las líneas de recuperados en Argentina y México son casi nulas, probablemente por falta de registro. En general, se observa un patrón de incremento fuerte a inicios de 2022 y descenso posterior en los tres países.

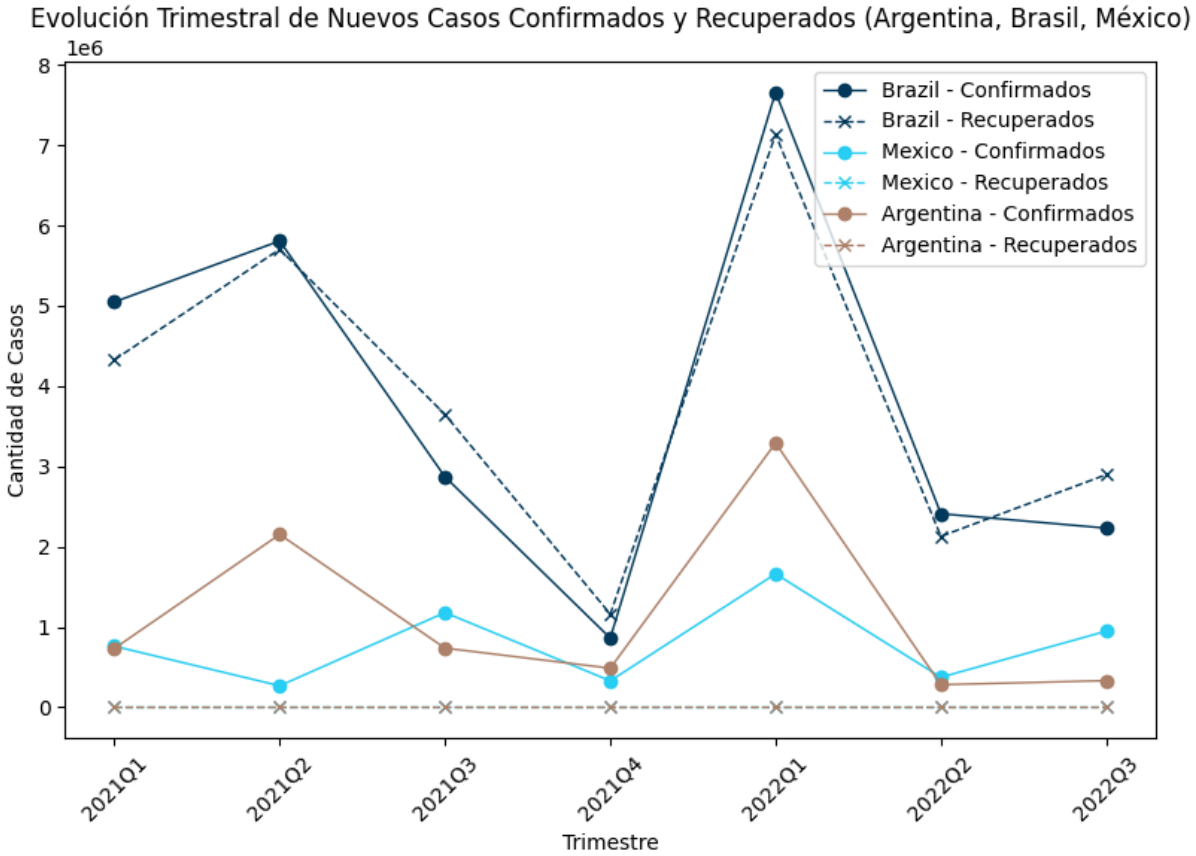


Figura 11: Lineplot: Evolución Trimestral Confirmados/Recuperados

3.12. Relación entre cobertura de vacunación y nuevos casos confirmados.

La cobertura de vacunación creció de forma sostenida mientras los nuevos casos mostraron picos marcados, sobre todo a inicios de 2022, y luego disminuyeron. No se observa una relación directa inmediata entre el aumento de cobertura y la reducción de casos en todos los períodos.

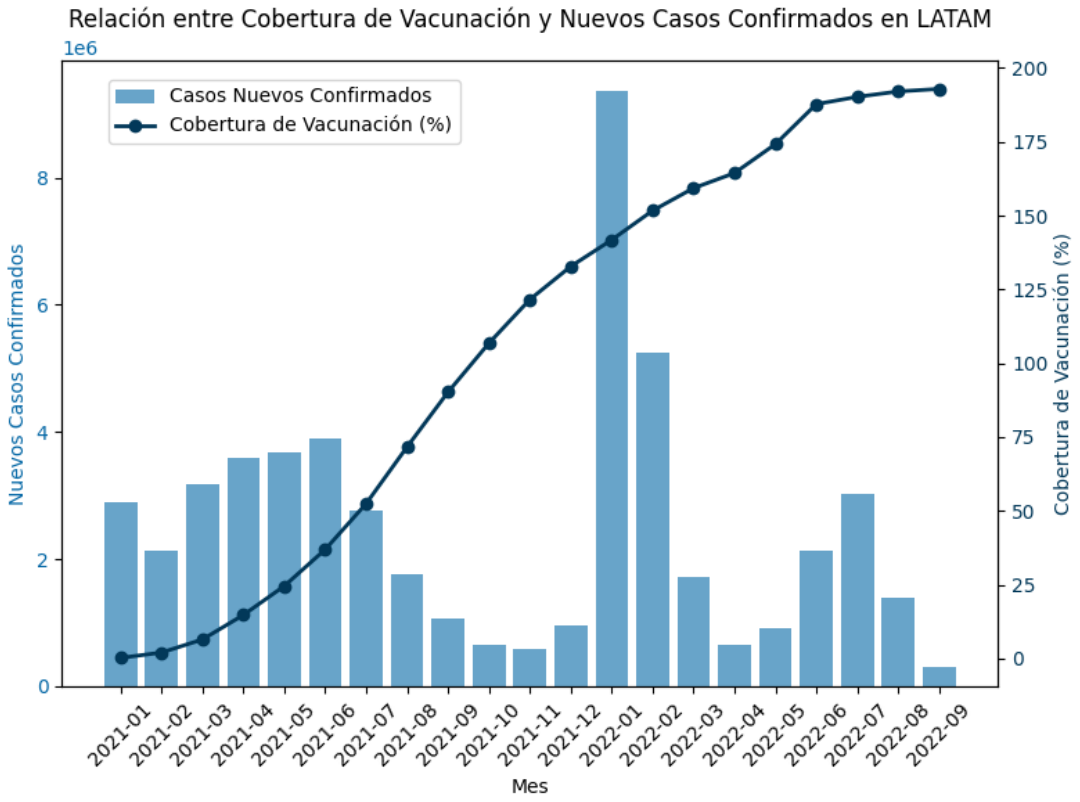
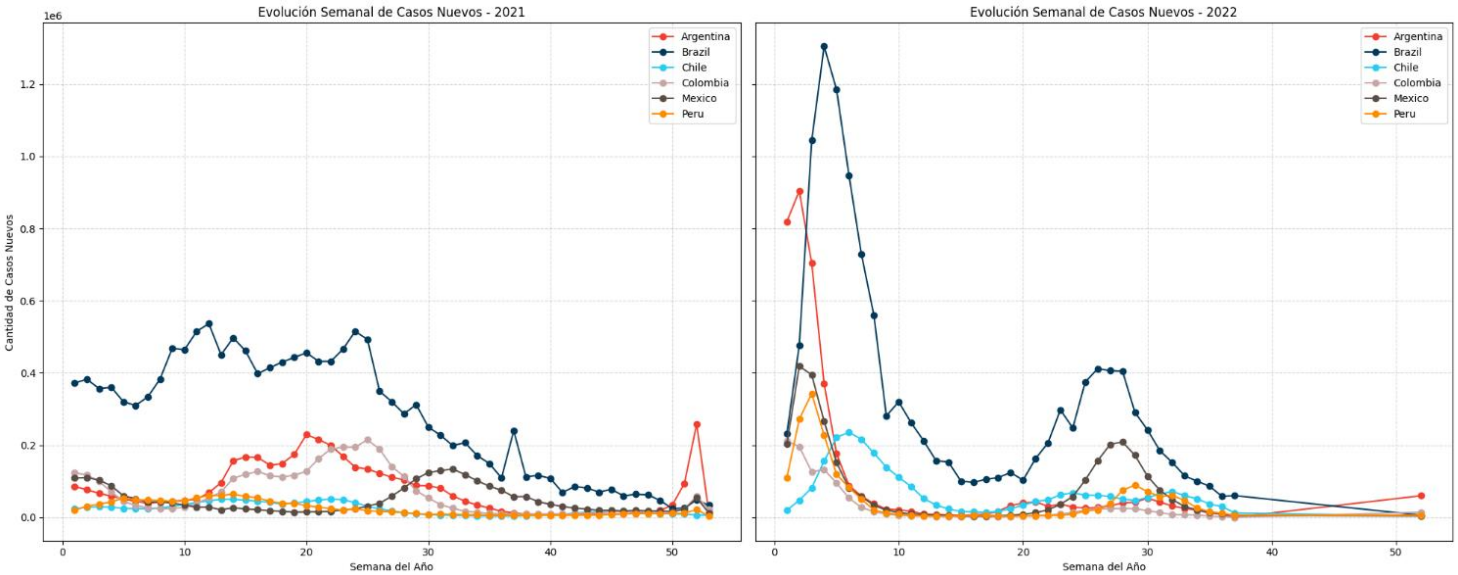


Figura 12: Cobertura de Vacunación / Nuevos Casos Confirmados

3.13. Evolución semanal de casos nuevos por país en 2021 – 2022

En 2021, Brasil mantuvo niveles elevados de casos durante casi todo el año con varios picos intermedios, mientras que los demás países mostraron fluctuaciones más moderadas. Argentina y Colombia presentaron aumentos alrededor de la mitad del año, mientras México tuvo un repunte hacia el tercer trimestre. En 2022, todos los países registraron un pico muy alto de casos en las primeras semanas, especialmente Brasil y Argentina, coincidiendo con la expansión de Ómicron. Luego, se observa un descenso sostenido y marcado en la mayoría de los países, con niveles mucho más bajos en el resto del año.

Figura 13: Evolución semanal de casos 2021/2022 por país



3.14. Evolución mensual de dosis nuevas aplicadas y muertes por país

En el primer gráfico (Figura 14) se observa que Brasil lideró de manera sostenida la aplicación de nuevas dosis durante la mayor parte del período, con un pico importante entre junio y septiembre de 2021. México también mostró un volumen alto, aunque más distribuido en el tiempo. Argentina tuvo un comportamiento particular con un pico muy marcado en mayo de 2022, que contrasta con su ritmo más estable en los meses anteriores. El resto de los países mantuvieron niveles más bajos y relativamente constantes de aplicación de dosis a lo largo del tiempo.

En el segundo esquema (Figura 15) se observa que Brasil tuvo los picos más altos de mortalidad, especialmente entre marzo y junio de 2021, cuando superó ampliamente a los demás países. México y Perú también registraron niveles elevados en la primera mitad de 2021. A partir de mediados de ese año, todas las curvas muestran un descenso progresivo y sostenido de las muertes mensuales, que coincide en el tiempo con el aumento de la cobertura de vacunación visto en el gráfico anterior. Desde 2022 en adelante, los valores se mantuvieron mucho más bajos en todos los países.

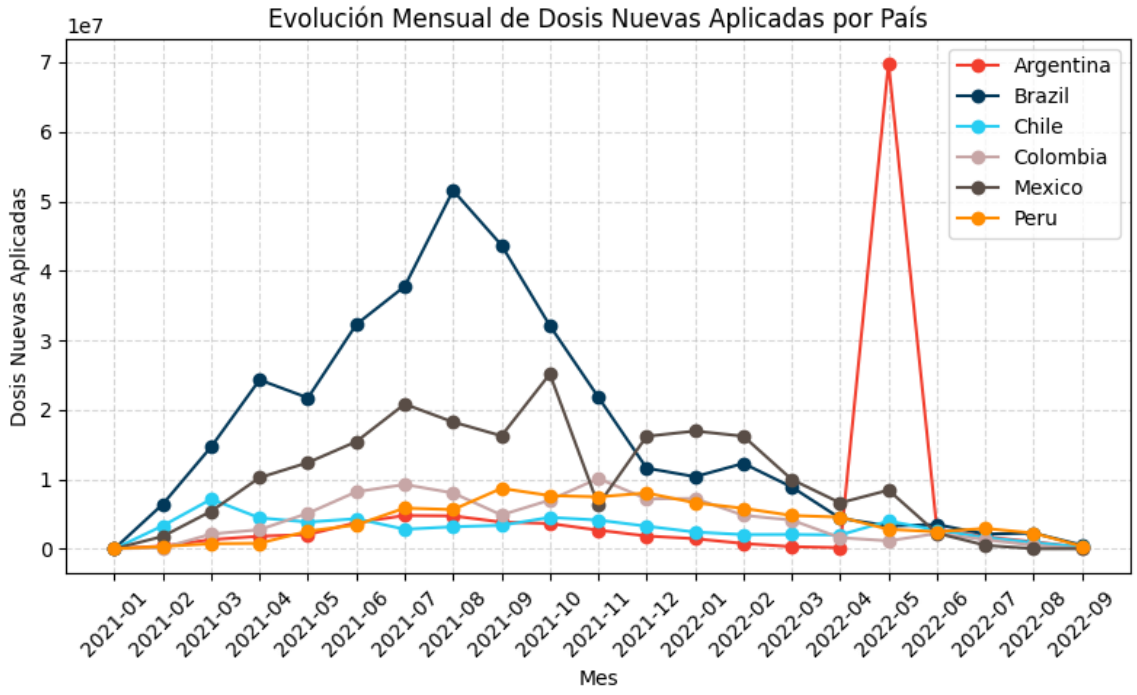


Figura 14: Evolución mensual de Dosis Nuevas aplicadas por país

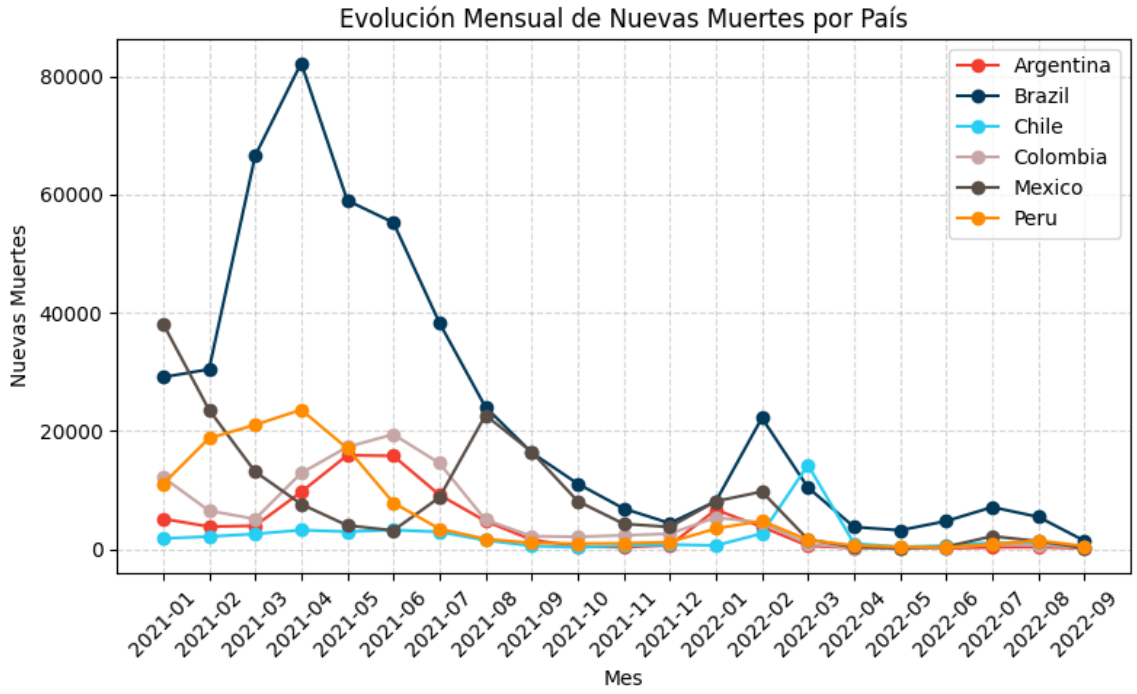
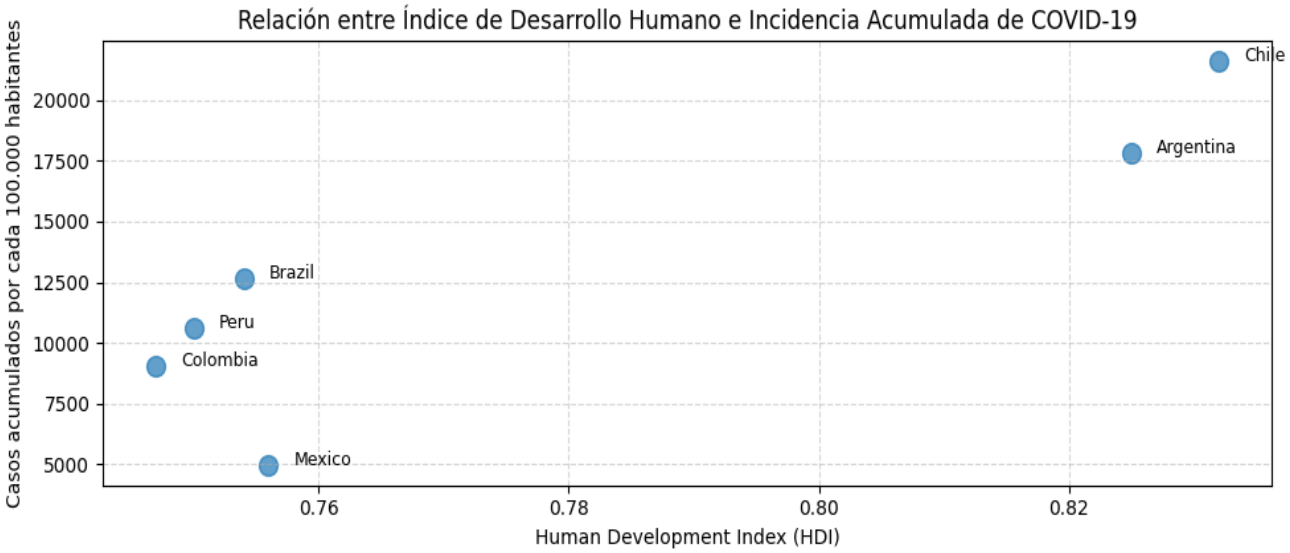


Figura 15: Evolución mensual de Muertes por país

3.15. HDI vs Incidencia Ajustada por Población (cada 100.000 hab.)

Este gráfico (Gráfico 16) muestra que los países con mayor Índice de Desarrollo Humano, como Chile y Argentina, registraron una incidencia acumulada más alta de casos por cada 100.000 habitantes. En cambio, países con HDI más bajo, como México, Colombia y Perú, presentaron valores menores. Esto sugiere que un mayor desarrollo humano puede estar asociado a una mayor capacidad de detección y registro de casos, más que necesariamente a mayor propagación.



4. Análisis de Dashboard

Luego del análisis realizado, se desarrolló un Dashboard interactivo en Power BI para la empresa BIOGENESYS Farmacéutica, con el propósito de integrar y visualizar indicadores demográficos, epidemiológicos y sanitarios relevantes. Este tablero permite explorar de manera dinámica la información procesada, identificar patrones clave y fundamentar decisiones estratégicas sobre la expansión de laboratorios en Latinoamérica ante el contexto de la pandemia y la postpandemia de COVID-19.

4.1. Navegación del Dashboard

El informe cuenta con una portada de presentación, seguida de cuatro secciones principales que pueden explorarse mediante los botones superiores:

Perfil Demográfico: muestra indicadores clave de población, distribución etaria, prevalencia de enfermedades crónicas y nivel socioeconómico. Incluye segmentadores por país y año.

Correlaciones y Dispersión: presenta relaciones entre variables mediante mapas de calor, correlaciones y diagramas de dispersión. Este apartado muestra datos consolidados de todos los países, sin segmentación individual.

Evolución Epidemiológica: expone la evolución mensual de casos confirmados, dosis aplicadas y fallecimientos, con segmentadores que permiten filtrar por país y año. Además de tarjetas con datos de casos confirmados, fallecidos y dosis aplicadas.

Conclusiones: sintetiza recomendaciones preliminares de expansión hacia Brasil y México, con argumentos basados en los datos analizados.

En esta última sección se incorporó un botón interactivo que, al activarse, despliega un panel con Recursos Complementarios. Allí se incluyen enlaces directos a publicaciones y documentos de referencia que fundamentan el análisis y aportan contexto adicional.

5. Conclusiones y recomendaciones

A partir de los datos analizados, recomendaría priorizar Brasil y México como opciones principales para la expansión de la farmacéutica. Estas recomendaciones se sustentan en varios factores:

- **Tamaño poblacional y volumen de casos:** Ambos países concentran las mayores poblaciones y registraron el mayor número de casos confirmados y muertes, lo que evidencia una alta demanda potencial de insumos médicos, vacunas y tratamientos.
- **Capacidad logística demostrada:** Lideraron en dosis totales de vacunas administradas, mostrando infraestructura sanitaria y operativa capaz de absorber y distribuir grandes volúmenes de productos farmacéuticos.
- **Prevalencia de enfermedades crónicas**: México, en particular, presenta la mayor prevalencia de diabetes, lo que incrementa las oportunidades para programas de prevención, diagnóstico y tratamiento de enfermedades no transmisibles.
- **Indicadores económicos:** Si bien Chile cuenta con el mayor GDP per cápita, el tamaño de mercado de Brasil y México es sustancialmente mayor, ofreciendo más escala comercial.

En todos los casos la pandemia tuvo un impacto profundo, generando una mayor conciencia social y gubernamental sobre la importancia de invertir en salud pública y privada.

6. Reflexión personal

La propuesta del proyecto integrador de este módulo me resultó sumamente interesante y enriquecedora. Sin embargo, considero que sería necesario extender el tiempo de cursada del Módulo 4, ya que las lecturas contienen un volumen importante de contenido que requiere un periodo adecuado para su verdadera internalización y apropiación. El proceso de enseñanza-aprendizaje no sólo demanda tiempo, sino también práctica constante, y en el afán de cumplir con los avances semanales puede perderse la oportunidad de profundizar y consolidar los conocimientos.

En el desarrollo de trabajo, pude practicar y afianzar conceptos fundamentales del análisis de datos, desde la carga de las bases hasta el diseño del Dashboard en Power BI.

En mi profesión, suelo trabajar con análisis de datos utilizando herramientas más básicas como Excel y sus gráficos, por lo que esa experiencia previa me ayudó a interpretar los resultados con mayor claridad. También me sirvió de apoyo para decidir qué variables resultaban clave y el modo de organizar la información de manera coherente dentro del proyecto.

Si tuviera que empezarlo nuevamente, creo que seguiría el mismo camino (necesitaría mucho más tiempo). Cada etapa presentó desafíos, pero el proceso que utilicé me permitió resolverlos adecuadamente y avanzar sin contratiempos.

7. EXTRA CREDIT

Avance 1: Funciones de orden superior

Son funciones que reciben otras funciones como argumento o devuelven otras funciones. Aplica para herramientas como:

map(): Recorre cada elemento de una colección (lista, serie, etc.) y aplica una función a cada uno, devolviendo una nueva colección con los resultados.

```
# Columna de casos confirmados y que se convierte en “alto” o “bajo”
nivel_contagio = map(lambda x: "alto" if x > 100000 else "bajo",
df_Latinoamerica_mayor_2021_locationkey["cumulative_confirmed"])
df_Latinoamerica_mayor_2021_locationkey["nivel_contagio"] = list(nivel_contagio)

# Visualizamos resultado
print(df_Latinoamerica_mayor_2021_locationkey[["country_name",
"cumulative_confirmed", "nivel_contagio"]].head(10))
```

filter(): Recorre la colección y guarda únicamente los elementos que cumplan con la condición definida en la función.

```
# nombres de país donde los casos acumulados son mayores a 1.000.000
paises_altos_niveles = list(filter(lambda pais:
df_Latinoamerica_mayor_2021_locationkey[df_Latinoamerica_mayor_2021_locationkey["cou
ntry_name"] == pais]["cumulative_confirmed"].sum() > 1000000,
                           df_Latinoamerica_mayor_2021_locationkey["country_name"].u
nique()))

# Ver qué países cumplieron la condición
print(paises_altos_niveles)
```

reduce() (de functools): Recorre toda la colección para reducirla a un único valor, aplicando una operación definida (sumas, multiplicaciones, concatenaciones, etc.).

```
# suma total de casos confirmador para todos los registros
from functools import reduce
total_contagios = reduce(lambda a, b: a + b,
df_Latinoamerica_mayor_2021_locationkey["new_confirmed"])

print(total_contagios)
```

apply(): Es una versión adaptada para Series o DataFrames que te permite aplicar una función a cada elemento (de una columna) o cada fila para crear nuevas columnas o realizar transformaciones específicas.

```
# Crear columna categórica para casos nuevos
df_Latinoamerica_mayor_2021_locationkey["nivel_casos_nuevos"] =
df_Latinoamerica_mayor_2021_locationkey["new_confirmed"].apply(
    lambda x: "bajo" if x < 5000 else ("moderado" if x < 10000 else "alto")
)

# Verificación
print(df_Latinoamerica_mayor_2021_locationkey[["new_confirmed",
"nivel_casos_nuevos"]].head(10))
```