**Expansión Estratégica de Biogenesys con Python**

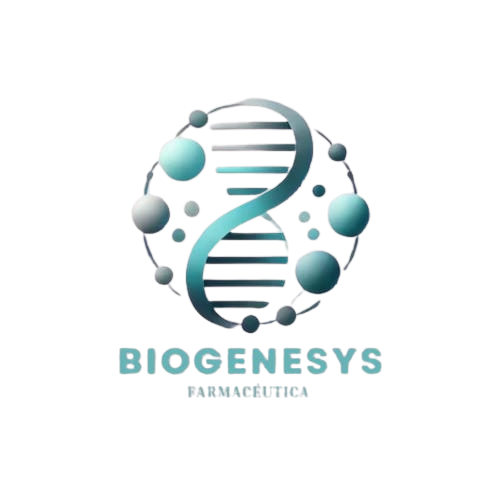
**Nombre del autor: Erica Solange Baumann**

**Email: ericabaumann@gmail.com**

**Cohorte: DA-PT03**

**Fecha de entrega: 04/11/2024**

**Institución: BIOGENESYS**



# 

# 

# Introducción

# Biogenesys es una empresa farmacéutica en proceso de expansión en América Latina, buscando responder de manera óptima a los desafíos de salud pública pospandemia. Su estrategia se enfoca en aumentar el acceso a servicios sanitarios mediante la identificación de ubicaciones estratégicas para establecer nuevos laboratorios y centros de vacunación.

# Este proyecto de expansión abarca seis países: Argentina, Brasil, Chile, Colombia, Perú y México. A través del análisis de datos en Python y el uso de Power BI, se busca responder de forma ágil y escalable a los efectos de la pandemia, optimizando el acceso a las vacunas mediante el despliegue de infraestructura en ubicaciones de alto impacto.

#### Desarrollo del Proyecto

#### Avance 1: Carga y Transformación de los Datos

En este primer avance, se centró en la preparación y limpieza del conjunto de datos proporcionado. El archivo original era considerable, con un peso de aproximadamente 21 GB, 22 millones de registros y 707 columnas. Dado su tamaño, el equipo de ingeniería de datos generó un subconjunto específico para Latinoamérica en formato CSV, con 12,216,057 filas y 50 columnas, lo que facilitó su manipulación y análisis para los objetivos del proyecto.

Métodos de carga y transformación de datos.

1. Carga del dataset: Utilizando Pandas, se cargó el archivo CSV optimizado, permitiendo acceder de manera más eficiente a los datos específicos de los países en los que BIOGENESYS desea expandir su infraestructura.

# Abrimos el dataset

data\_lat = pd.read\_csv('data\_latinoamerica.csv')

1. Filtrado de países: Los datos fueron filtrados para trabajar exclusivamente con los países de interés (Argentina, Brasil, Chile, Colombia, México y Perú), asegurando que el análisis se centre en los territorios más relevantes.

# Filtrado de países específicos

paises = ['Argentina', 'Brazil', 'Chile', 'Colombia', 'Mexico', 'Peru']

data\_paises = data\_lat[data\_lat['country\_name'].isin(paises)]

1. Gestión de valores nulos: Se verificaron y reemplazaron valores nulos en columnas clave, como new\_confirmed, mediante la media de cada país. Esto asegura que la información sea lo más completa posible sin introducir sesgos significativos.

# Ejemplo de imputación de valores nulos para la columna 'new\_confirmed' por país

for pais in paises:

print(pais)

# Cantidad de valores nulos antes de la imputación

d0 = data\_paises[data\_paises["country\_name"] == pais]["new\_confirmed"].isnull().sum()

# Cálculo de la media para reemplazar valores nulos

d = data\_paises[data\_paises["country\_name"] == pais]["new\_confirmed"].mean()

# Asignación de la media a los valores nulos

data\_paises.loc[data\_paises["country\_name"] == pais, "new\_confirmed"] = data\_paises[data\_paises["country\_name"] == pais]["new\_confirmed"].fillna(d)

# Verificación de valores nulos después de la imputación

d1 = data\_paises[data\_paises["country\_name"] == pais]["new\_confirmed"].isnull().sum()

print('cant\_na =', d0, 'prom =', d, 'cant\_na =', d1)

1. Transformación de datos: Se estandarizaron columnas de fechas y ajustaron los tipos de datos donde fue necesario para mejorar el rendimiento en fases futuras de análisis. Además, se realizaron configuraciones de Pandas para asegurar una revisión exhaustiva de las columnas y valores disponibles.

# Configuración para visualizar todas las columnas y filas en el entorno de trabajo

pd.set\_option('display.max\_columns', None)

pd.set\_option('display.max\_rows', None)

pd.set\_option('display.max\_colwidth', None)

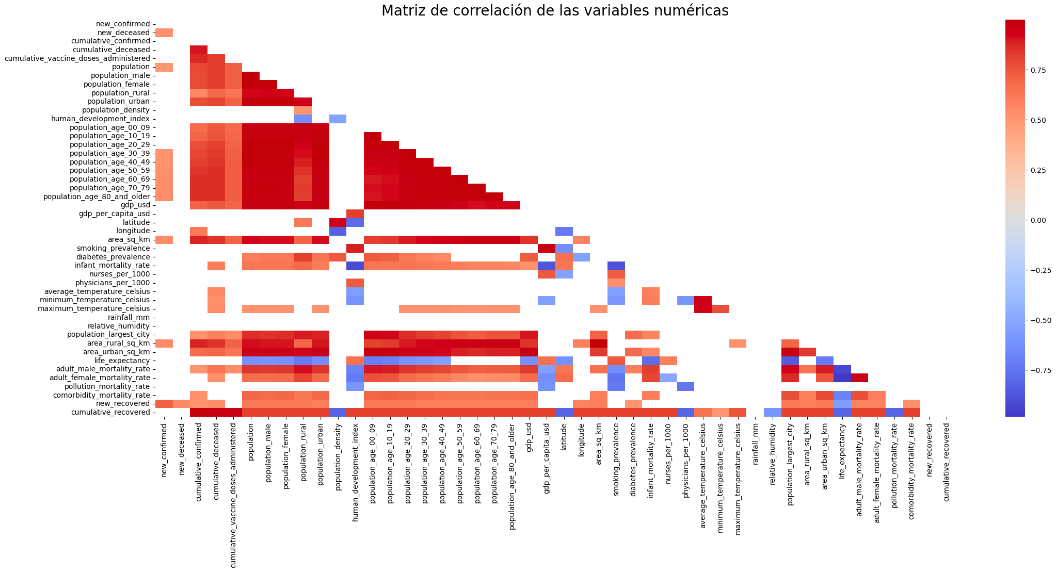
Resultados Iniciales: La limpieza de los datos permitió una reducción en la variabilidad de los registros, proporcionando un conjunto de datos robusto y listo para un análisis exploratorio y visualizaciones avanzadas, que guiarán las decisiones estratégicas de BIOGENESYS en su expansión en América Latina.

### Avance 2: Análisis Exploratorio de Datos (EDA) e Insights

### Con los datos ya limpios y preparados, procedimos a realizar un análisis exploratorio (EDA) detallado, con el objetivo de extraer insights relevantes que guíen la planificación estratégica para la expansión de BIOGENESYS. Mediante visualizaciones detalladas, buscamos identificar patrones, tendencias y posibles anomalías en la incidencia de COVID-19, las tasas de vacunación y la disponibilidad de infraestructura sanitaria. Estos análisis permiten a los directivos tomar decisiones fundamentadas sobre la ubicación de nuevos laboratorios y centros de vacunación, maximizando el impacto positivo en la salud pública.

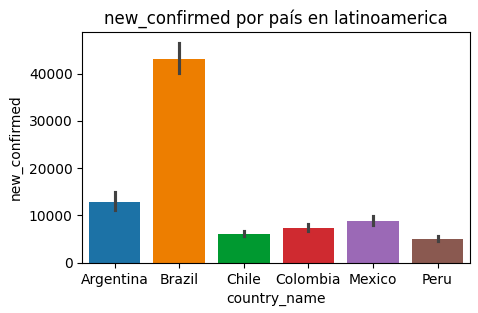
### Preparación de Datos Temporales: Iniciamos el EDA estableciendo la columna de fechas date como índice y asegurándonos de que los datos estuvieran en formato de fecha. Este paso es fundamental para el análisis temporal, permitiendo observar la evolución de los datos a lo largo del tiempo.

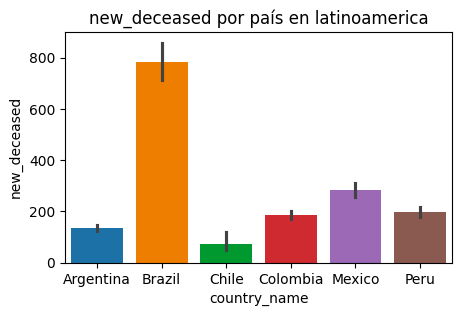
### Análisis de Correlación: Se generó una matriz de correlación visualizada con un heatmap, que reveló correlaciones significativas entre varias variables.

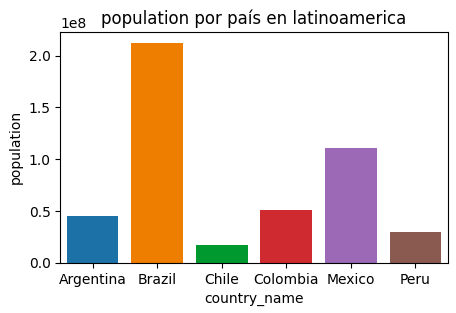


### Gráficos de Barras para Variables Demográficas: Se construyeron gráficos de barras para variables demográficas como new\_confirmed, new\_deceased, population, population\_male, population\_female, population\_rural, population\_urban, que dieron graficos similares a estos::

### 

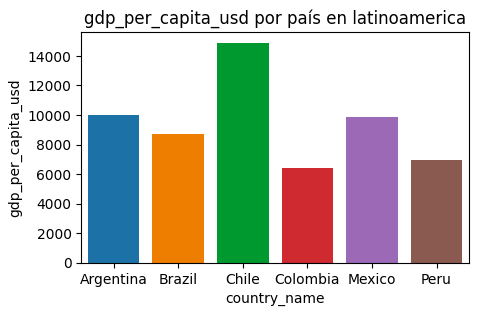


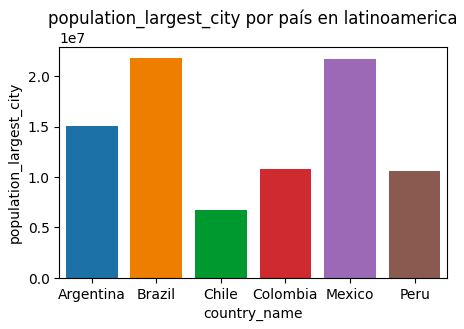


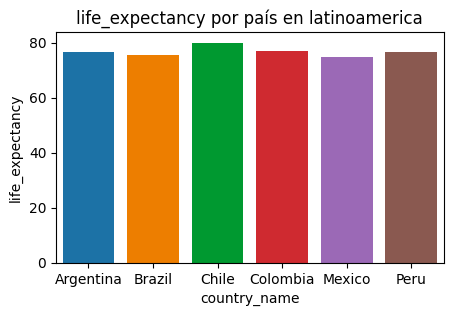


### Distribución Poblacional por Rango Etario: Se realizaron gráficos de barra para mostrar la distribución de la población en varios rangos etarios, con una distribución proporcional a la población total en cada país. Realice 9 gráficos que visualizan las edades de 0 a 9 años, de 10 a 19 años, de 20 a 29 años, de 30 a 39 años, de 40 a 49 años, de 50 a 59 años, de 60 a 69 años, de 70 a 79 años y de 80 años en adelante.

### Gráficos Económicos y de Salud: Se analizaron variables como gdp\_per\_capita\_usd, population\_largest\_city, y life\_expectancy.

****

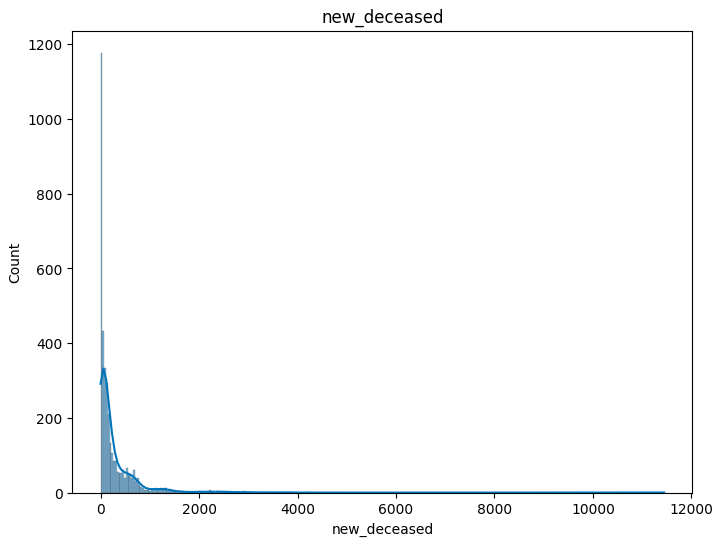
****

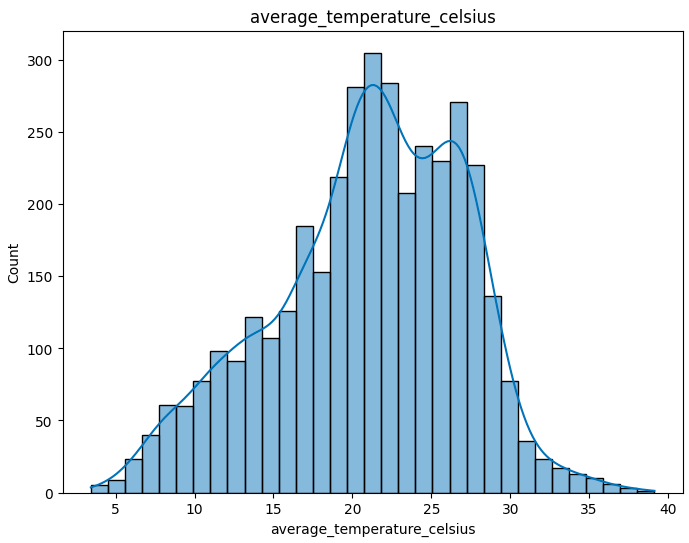
****

### Análisis de Distribuciones con Histplot: Utilizando gráficos histplot, observamos una distribución asimétrica en cumulative\_confirmed, lo cual es importante para identificar picos en casos acumulados.

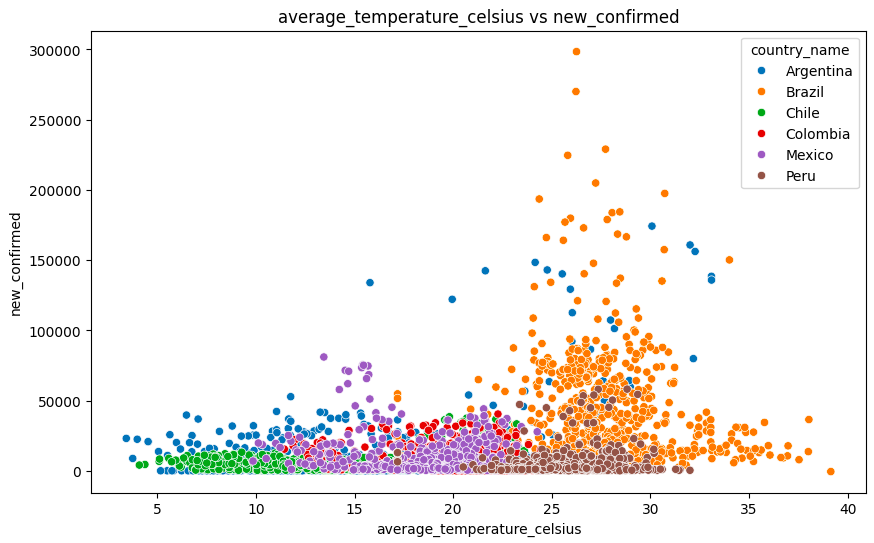
### 

### 

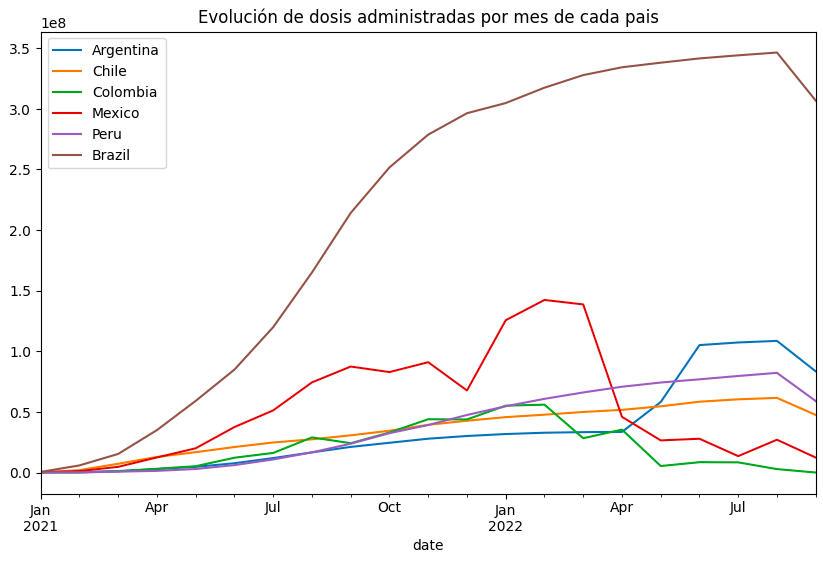


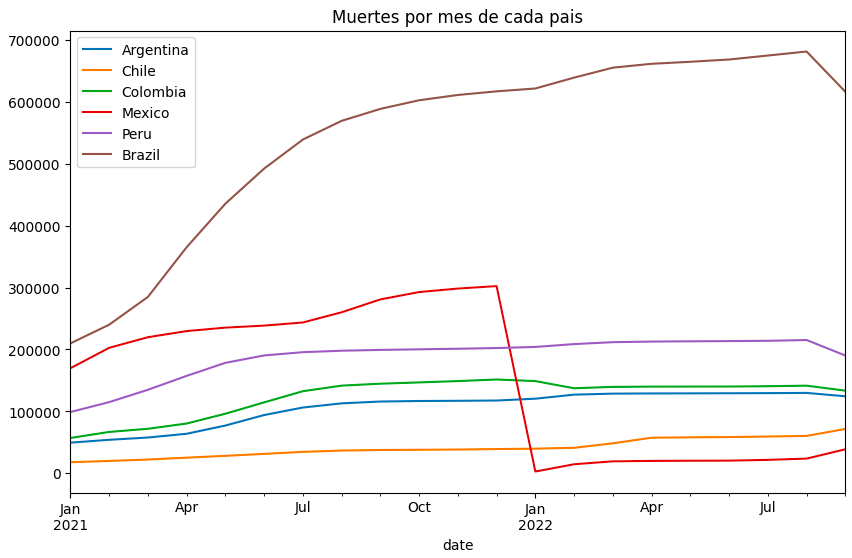


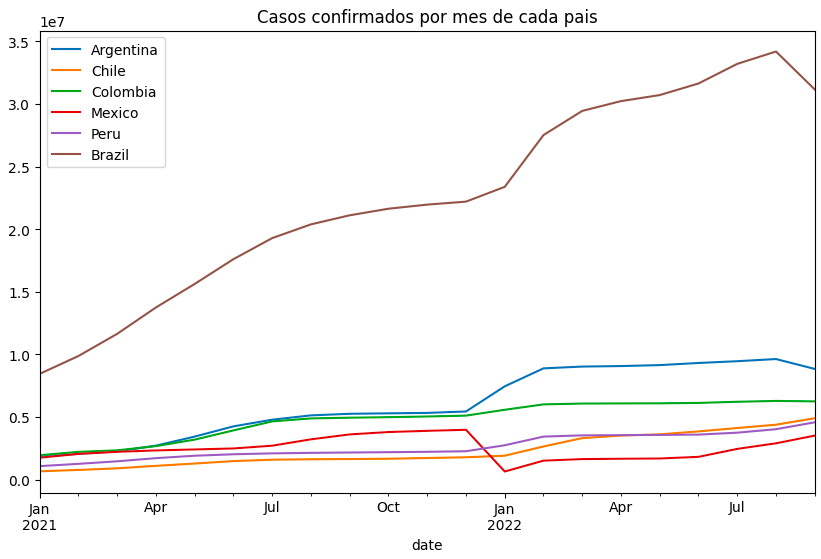
### Relación entre Temperatura y Nuevos Casos: Mediante gráficos de dispersión (scatterplot), analizamos la relación entre la temperatura y nuevos casos confirmados, encontrando una correlación débil.



### Análisis Temporal: Realizamos análisis a lo largo del tiempo de dosis administradas mensualmente y la evolución de casos confirmados, muertes y recuperaciones.



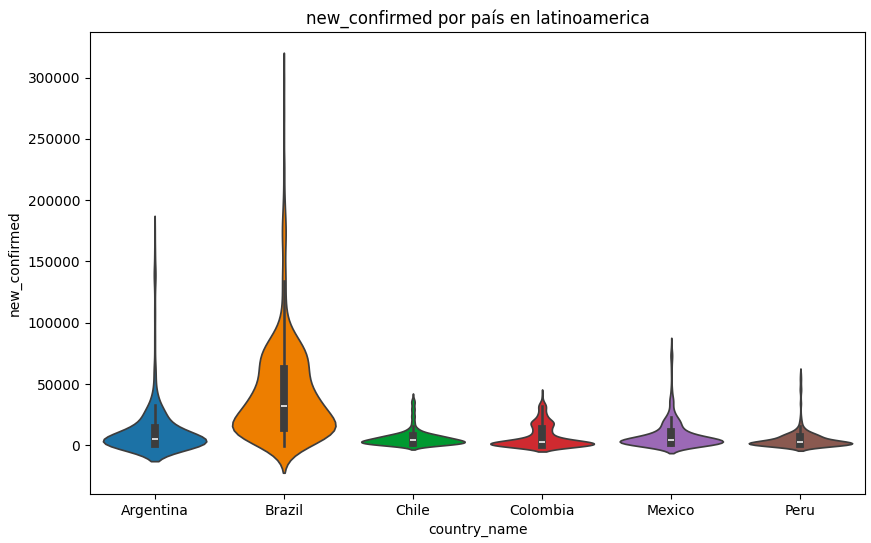


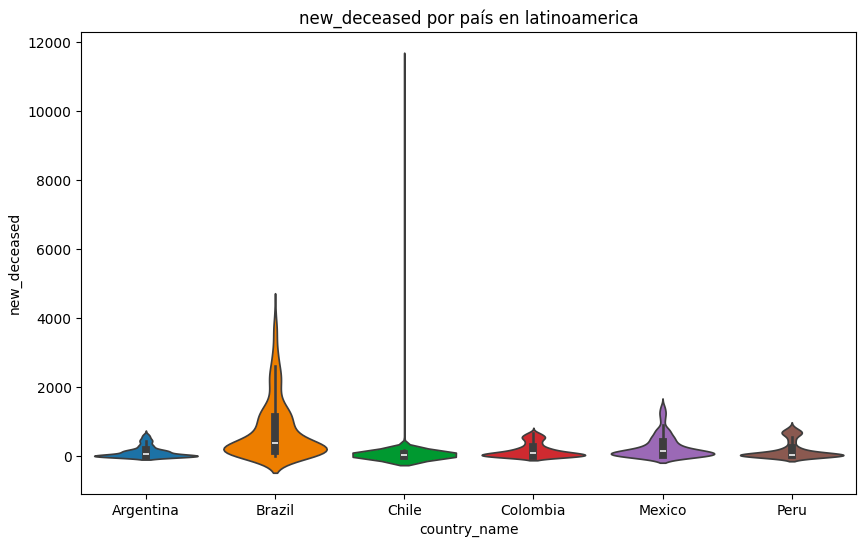


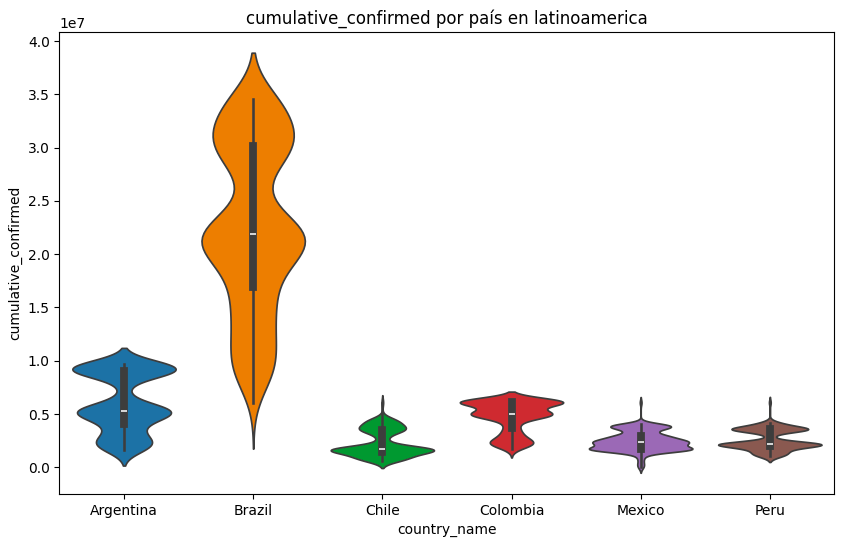
### Boxplot de Temperaturas Promedio: Se construyó un boxplot de temperaturas promedio, permitiendo observar variaciones significativas entre países y épocas del año.

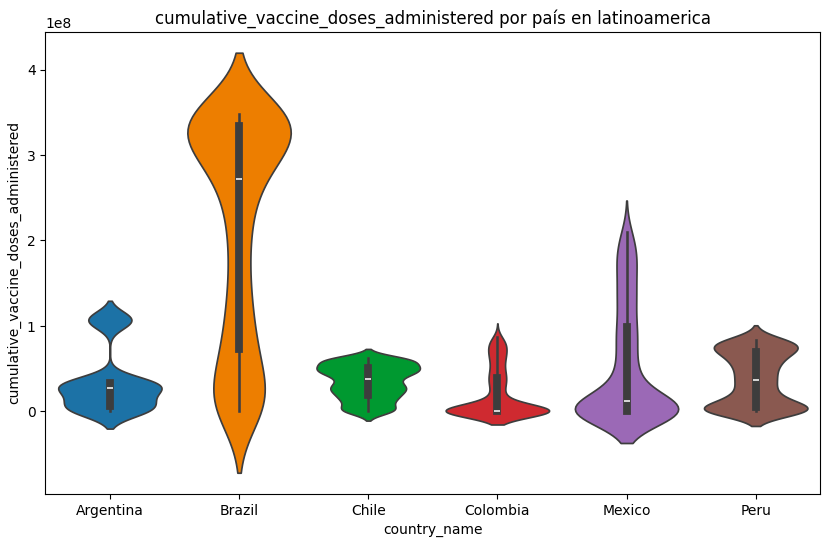
### 

### Análisis de Variabilidad con Violinplot: Los violinplots revelaron una alta variabilidad en los casos en Brasil y Argentina, indicando días con picos en los casos.



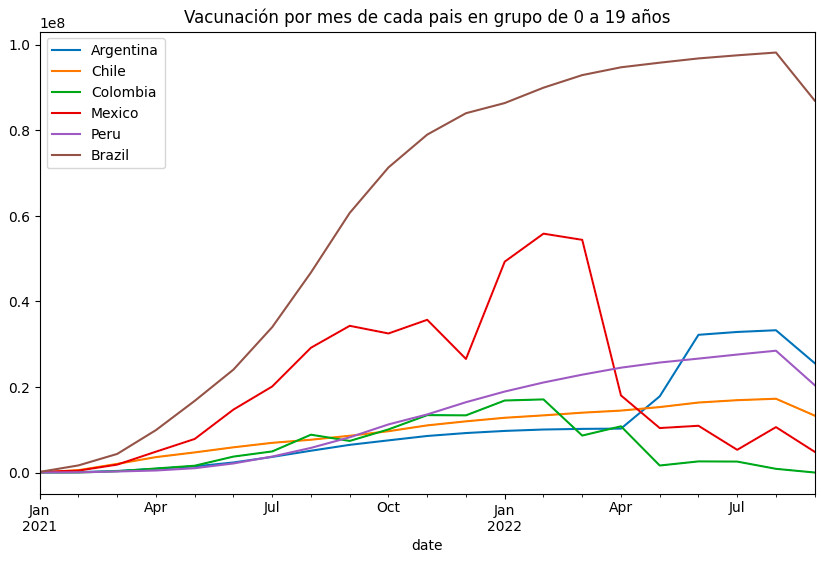


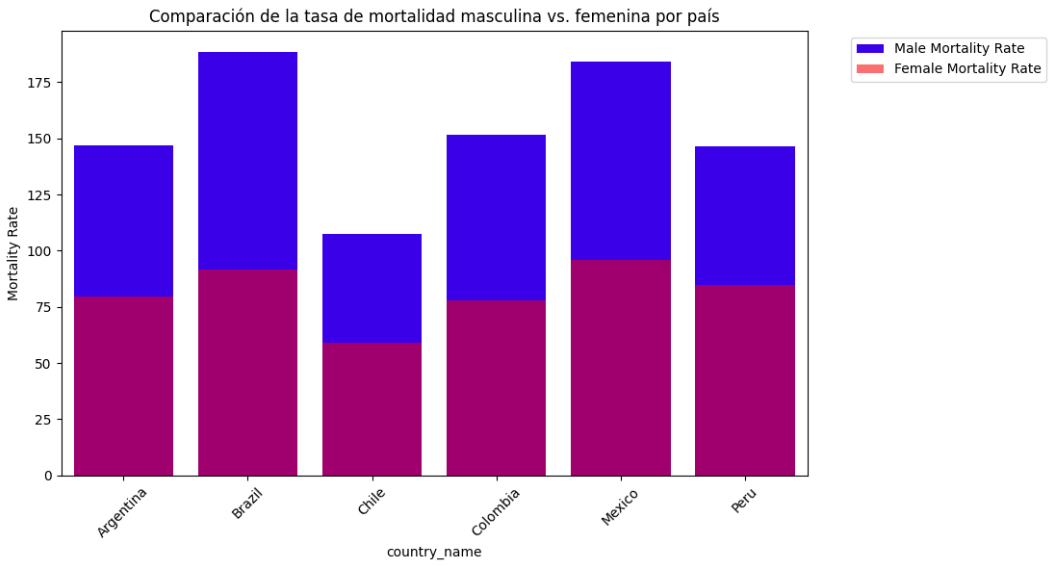




### Distribución General de la Población y Mortalidad: Se analizaron rangos etarios en la población, vacunación por mes por rangos etarios, y la comparación de la tasa de mortalidad masculina y femenina, revelando una mayor mortalidad en hombres.

### 

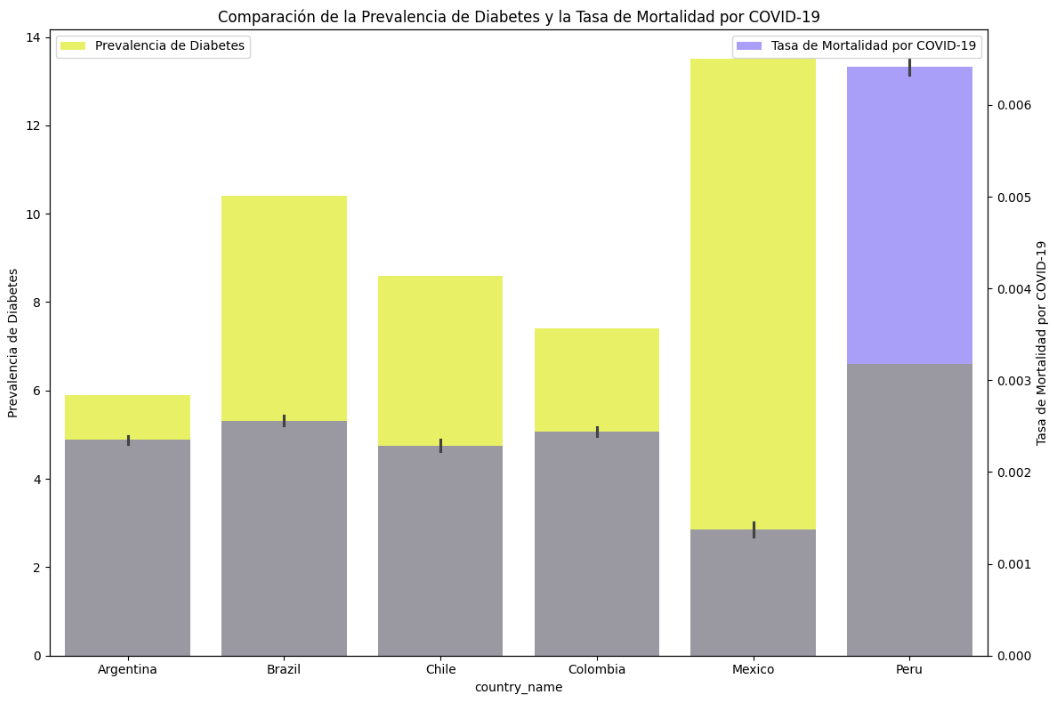




### Factores de Riesgo y Preexistencias: Se generó una matriz de correlación para indicadores de salud, mostrando que el tabaquismo y la diabetes son factores comunes. Luego generamos barplots para analizar la relación entre el fumar y la tasa de muertes por COVID, y de la relación entre la presencia de diabetes y la tasa de muertes por COVID.

### 

### 



### Resultados de lo visualizado en el Avance 2: Con base en todos los análisis, concluimos que Brasil, México y Colombia requieren más recursos, dadas sus altas cifras de casos confirmados y muertes, además de una tasa de vacunación baja en relación con la población. Esto proporciona una guía para la expansión de BIOGENESYS en áreas donde se maximizará el impacto en salud pública.

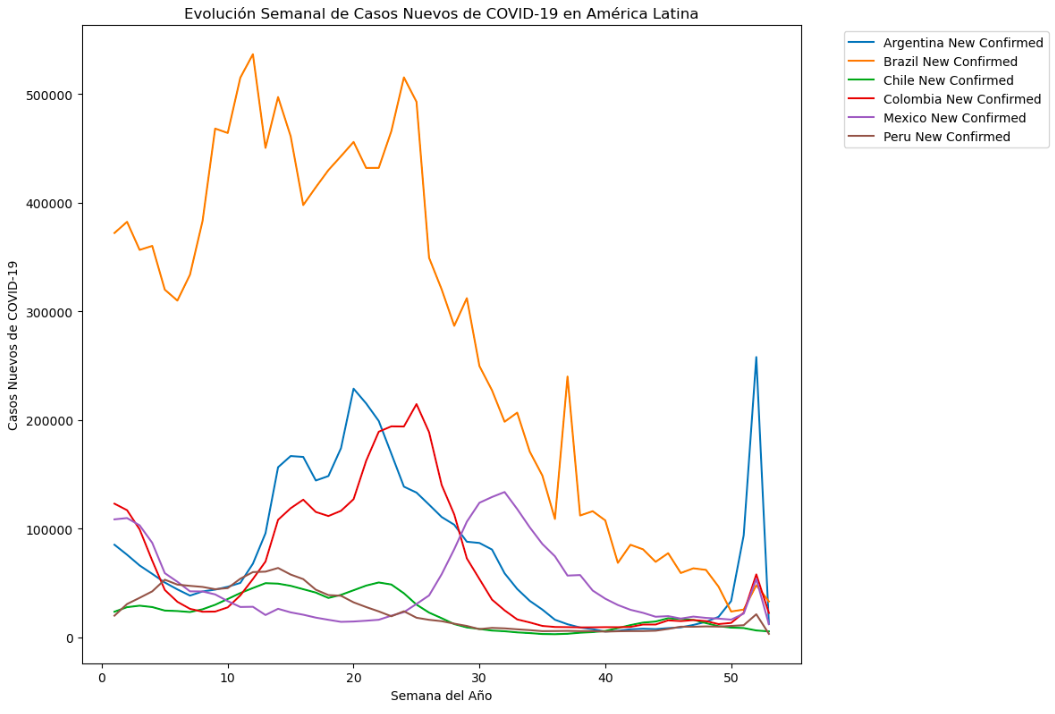
### Avance 3: Visualización Avanzada y Análisis Temporal

### Avance 3: Análisis Exploratorio de Datos (EDA) con Numpy y Pandas

En este avance, realizamos un análisis más profundo de los datos sobre la incidencia de COVID-19. El objetivo fue pulir y preparar los datos para visualizaciones avanzadas que nos ayuden a identificar las ubicaciones estratégicas para la expansión de los laboratorios de BIOGENESYS.

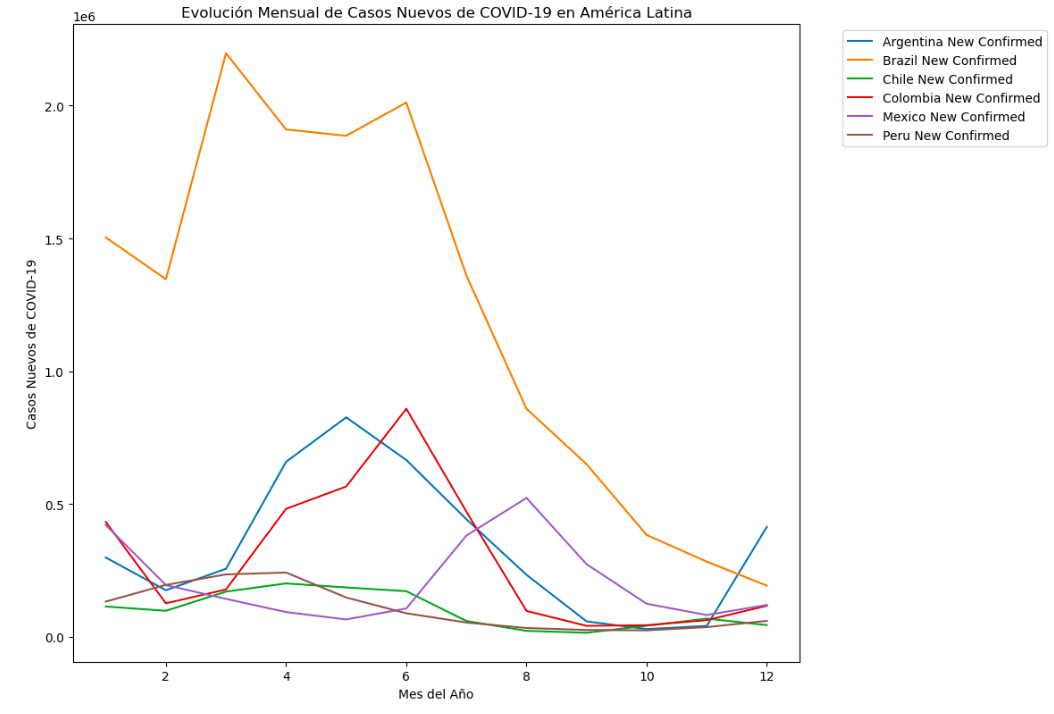
**Creación de Columnas Temporales**: Utilizando la función .isocalendar(), generamos nuevas columnas como week para facilitar el análisis temporal semanal.  
# Creación de la columna 'week' a partir de la fecha

data['week'] = data['date'].dt.isocalendar().week

****

**Análisis Mensual de Casos Nuevos**: También realizamos el análisis de casos nuevos por mes, permitiendo identificar variaciones en tendencias mensuales.  
  
# Creación de la columna 'month' para análisis mensual

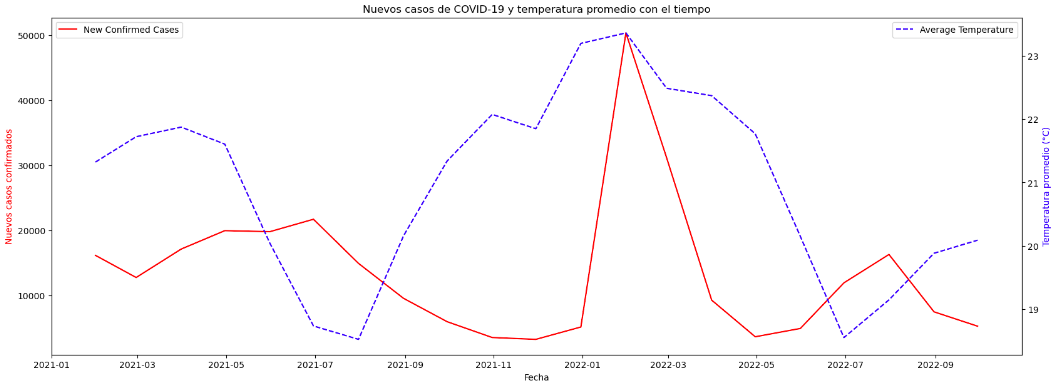
data['month'] = data['date'].dt.month

****

**Relación entre Temperatura Promedio y Casos Nuevos**: Generamos gráficos para examinar cómo las variaciones en la temperatura promedio se relacionan con los casos nuevos confirmados. Observamos que una disminución en la temperatura coincide con aumentos en los casos, especialmente en marzo de 2022, posiblemente debido a un aumento en la vacunación en Brasil.

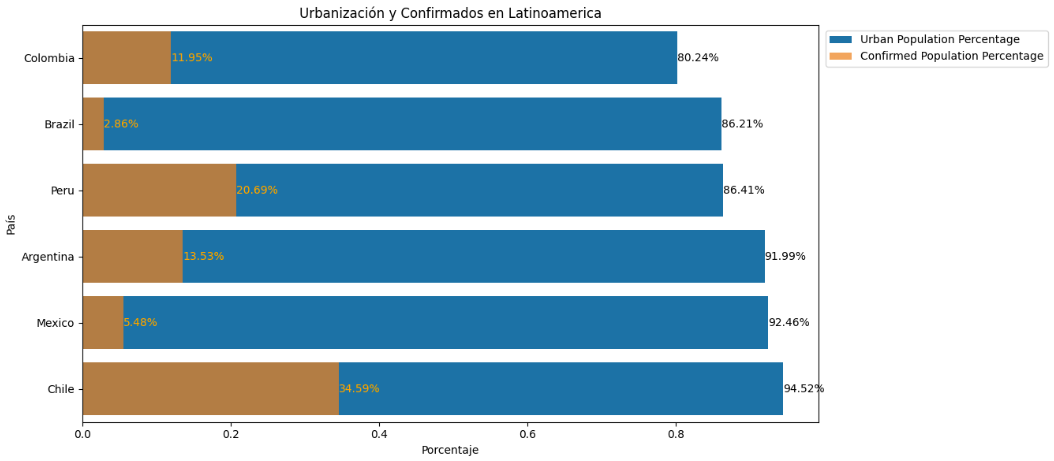
# Gráfico de temperatura promedio y casos nuevos confirmados

sns.scatterplot(x='average\_temperature', y='new\_confirmed', data=data)

****

**Análisis de Urbanización y Propagación de Casos**: Investigamos cómo la tasa de urbanización se relaciona con los casos confirmados, observando que Chile presenta un alto impacto en la propagación, a pesar de ser uno de los países con mejor control de la pandemia.  
  
# Gráfico de tasa de urbanización y casos confirmados

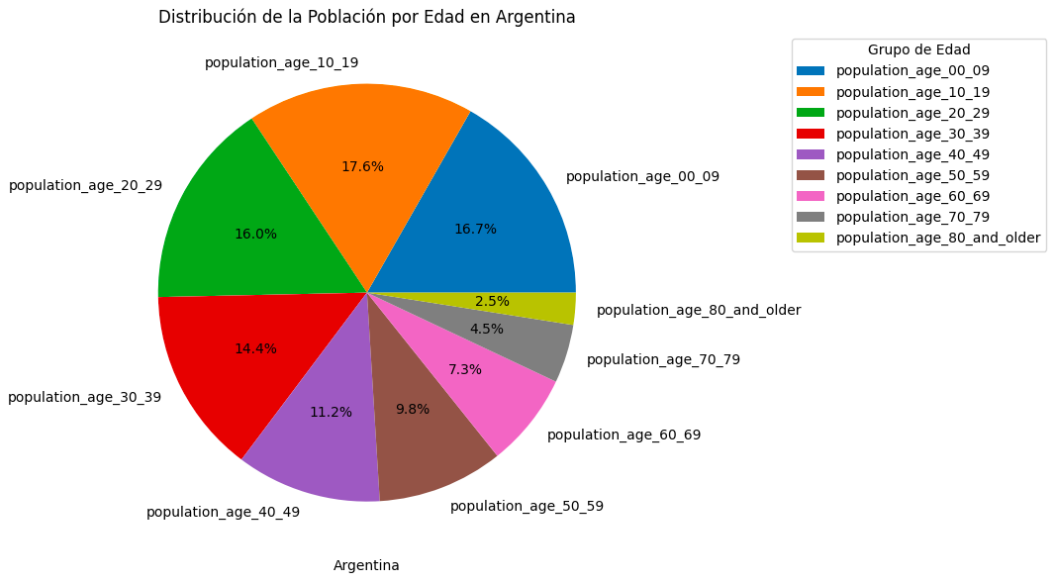
sns.scatterplot(x='urban\_population', y='new\_confirmed', data=data)

****

**Distribución de la Población por País**: Se crearon gráficos para visualizar la distribución de la población en cada país. Aquí mostramos el ejemplo de Argentina.

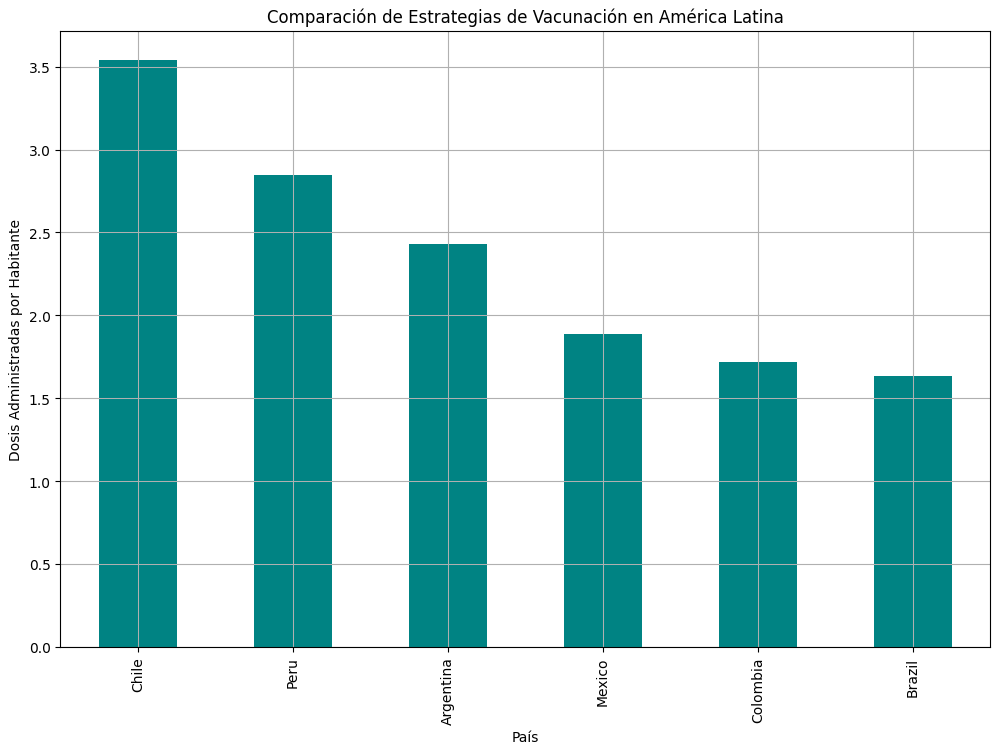
# Gráfico de barras de distribución de población en Argentina

data[data['country\_name'] == 'Argentina']['population'].plot(kind='bar')

****

**Dosis de Vacuna por Habitante**: Un gráfico relevante fue el de dosis de vacuna por habitante, donde Chile destaca en primer lugar, mientras que Brasil tiene la menor cantidad de dosis.  
  
# Gráfico de dosis por habitante

sns.barplot(x='country\_name', y='doses\_per\_capita', data=data)

****

**Incidencia de Fumar y Diabetes**: Se analizaron los factores de riesgo, observando que la diabetes afecta principalmente a México y Brasil, mientras que fumar es más prevalente en Chile y Argentina. No obstante, la mortalidad por COVID-19 no mostró una correlación directa con estos factores.  
  
# Gráficos de diabetes y tabaquismo

sns.barplot(x='country\_name', y='diabetes\_prevalence', data=data)

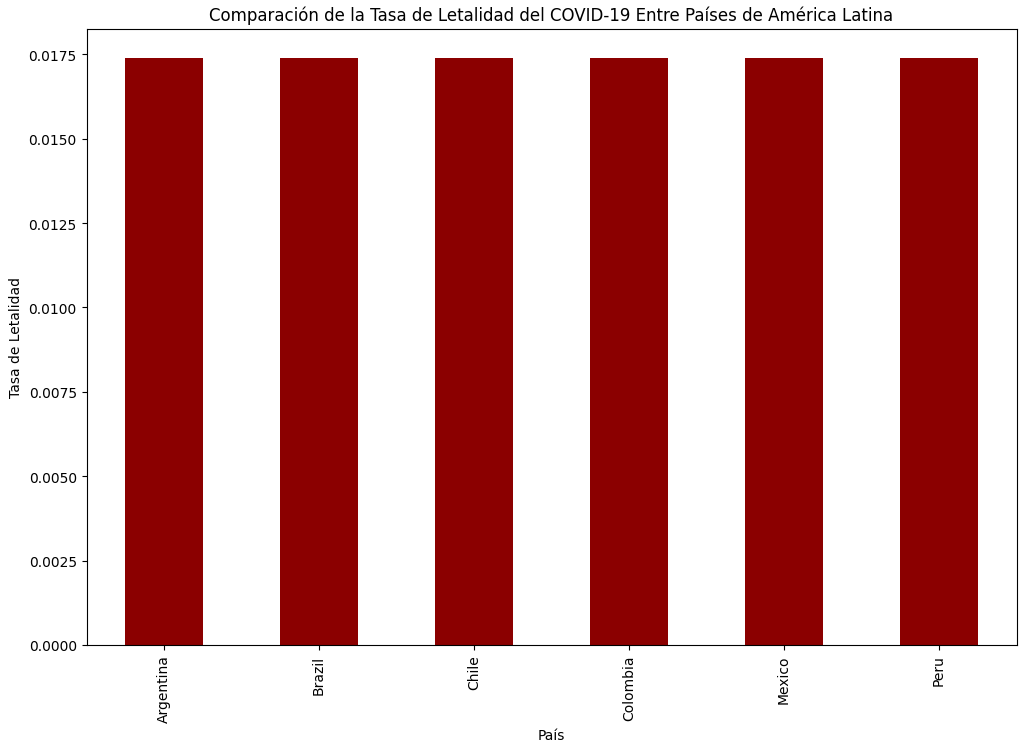
sns.barplot(x='country\_name', y='smoking\_prevalence', data=data)

**Tasa de Mortalidad por COVID**: Calculamos la tasa de mortalidad (muertes/población) en cada país, observando patrones similares en las tasas de mortalidad por COVID.

# Cálculo y gráfico de tasa de mortalidad por COVID

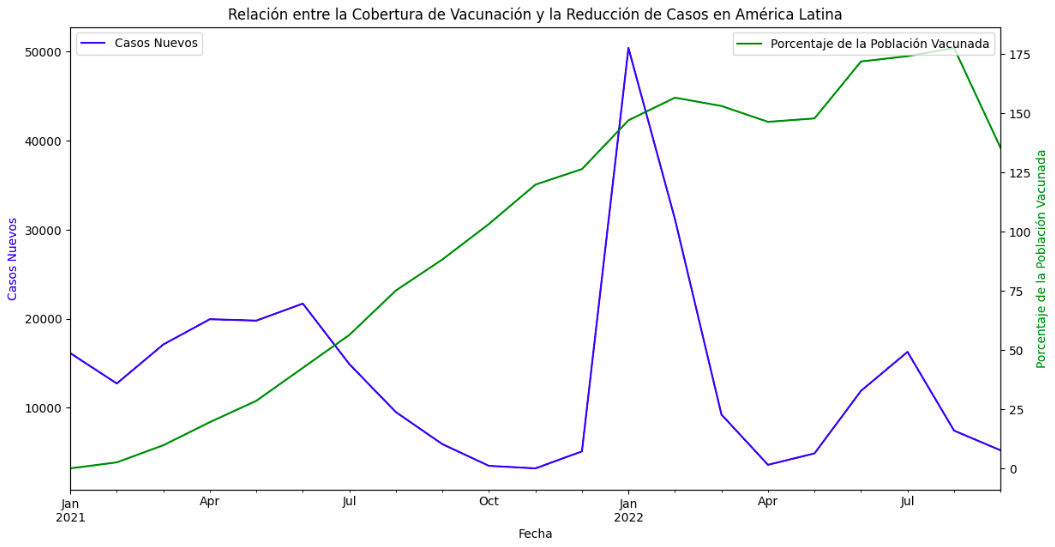
data['mortality\_rate'] = data['cumulative\_deceased'] / data['population']

sns.barplot(x='country\_name', y='mortality\_rate', data=data)

****

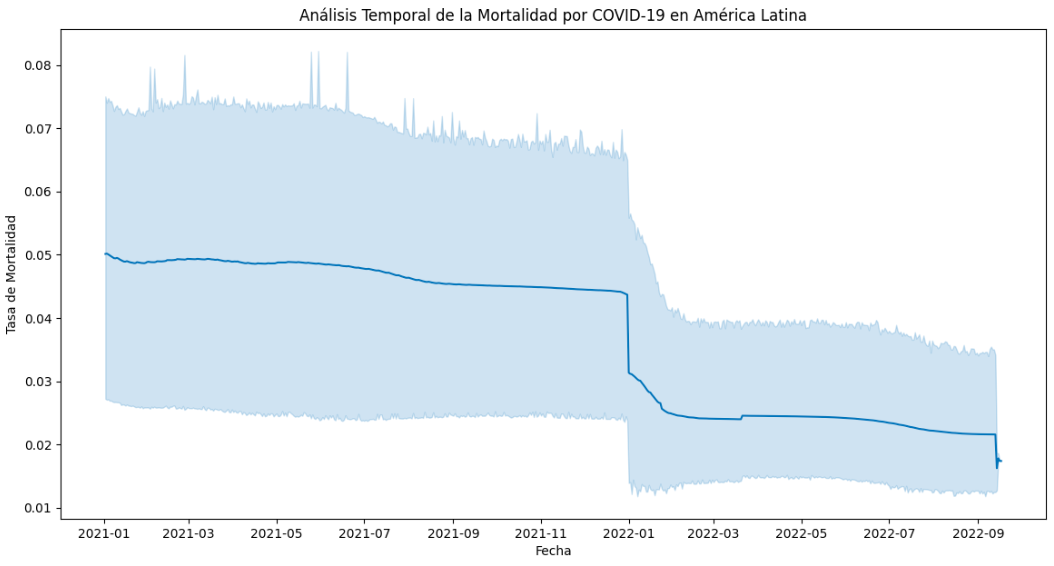
**Relación entre Vacunación y Reducción de Casos**: Otro gráfico importante mostró cómo el aumento en la vacunación se relaciona con una reducción en los casos nuevos.  
  
# Gráfico de vacunación acumulada y reducción de casos

sns.lineplot(x='date', y='cumulative\_vaccine\_doses', hue='country\_name', data=data)

****

**Evolución de la Tasa de Mortalidad en el Tiempo**: Observamos que la tasa de mortalidad disminuyó gradualmente con el aumento de la vacunación en el tiempo.  
# Gráfico de tasa de mortalidad en el tiempo

sns.lineplot(x='date', y='mortality\_rate', hue='country\_name', data=data)

****

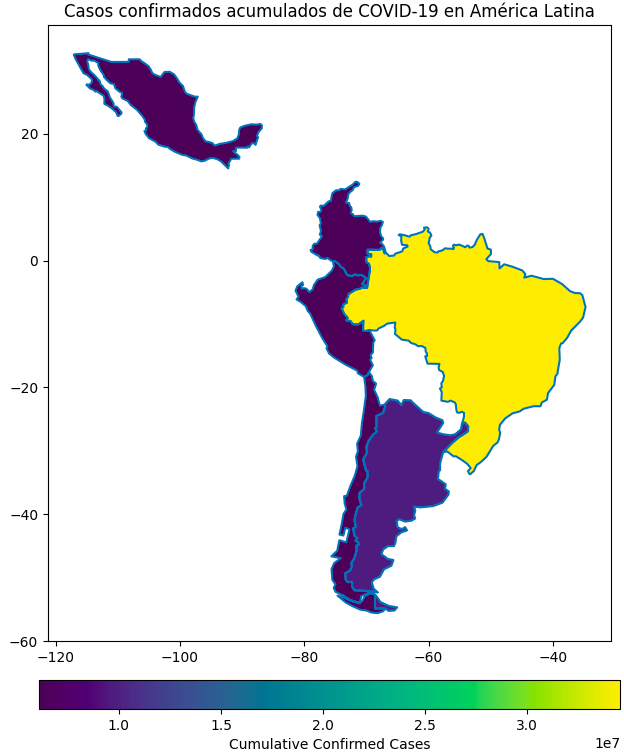
**Casos Activos y Recuperados Acumulados**: Se graficaron las cantidades acumuladas de casos activos y recuperados, mostrando el crecimiento de ambas métricas en el tiempo.  
# Gráficos de casos activos y recuperados

sns.lineplot(x='date', y='cumulative\_active\_cases', hue='country\_name', data=data)

sns.lineplot(x='date', y='cumulative\_recovered', hue='country\_name', data=data)

**Mapa de Calor de Variables Clave**: Finalmente, se utilizó un mapa de calor para visualizar casos confirmados, muertes, vacunas administradas y recuperaciones, lo cual facilita la identificación de patrones a nivel general.  
  
# Mapa de calor para casos confirmados, muertes, vacunas y recuperaciones

sns.heatmap(data[['cumulative\_confirmed', 'cumulative\_deceased', 'cumulative\_vaccine\_doses', 'cumulative\_recovered']].corr(), annot=True, cmap='coolwarm')



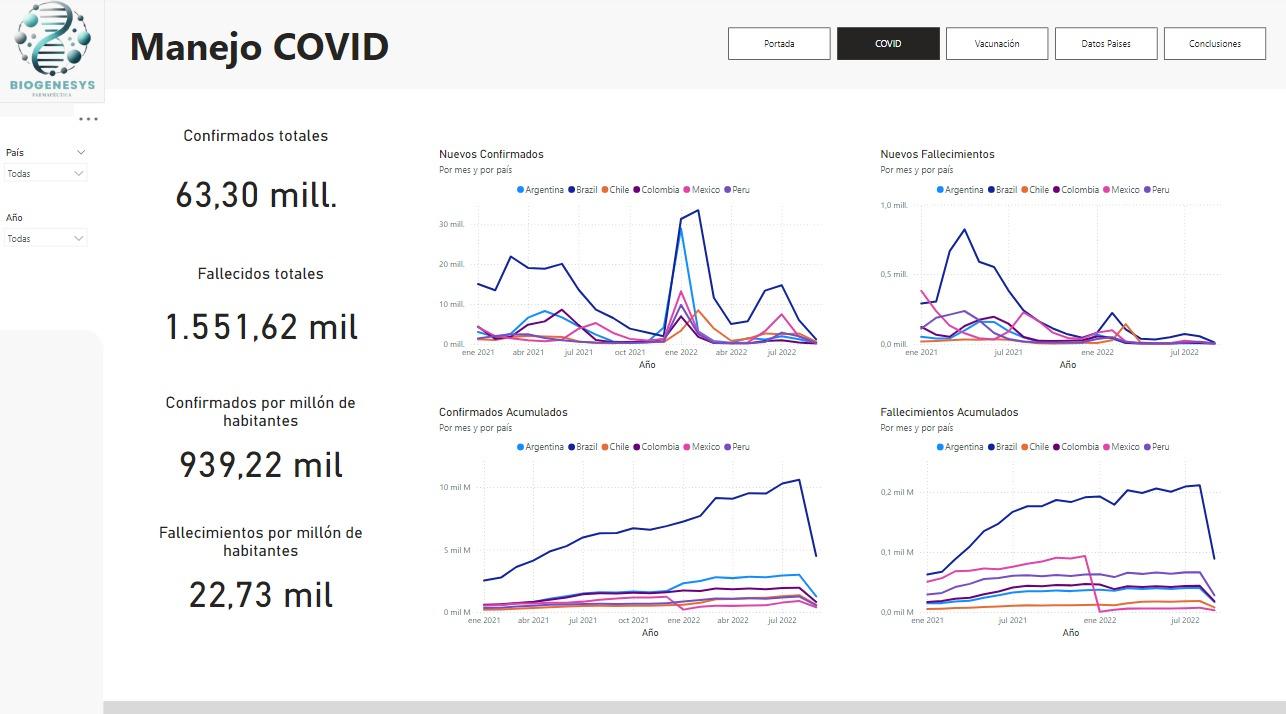
### Avance 4: Integración y Análisis en Power BI

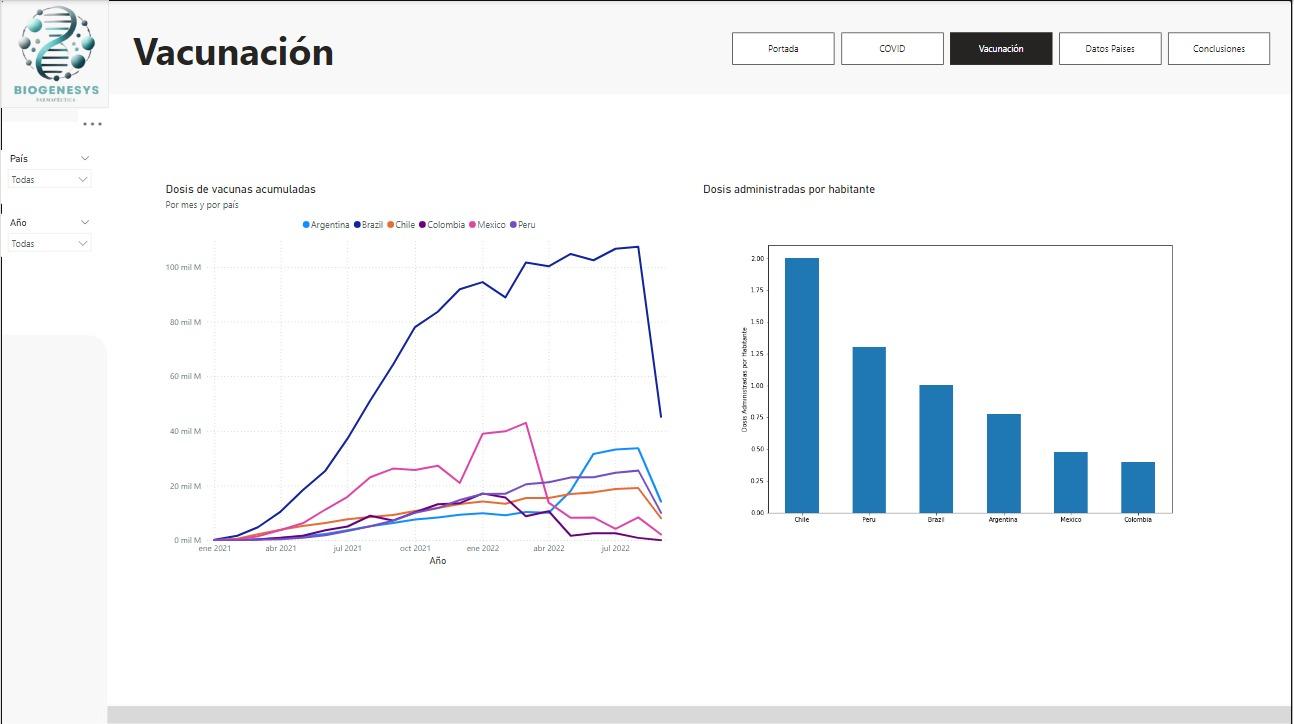
1. Preparación y Carga de Datos El dataset limpio se exportó a un archivo CSV para su uso en Power BI, facilitando la creación de paneles interactivos. Se incluyeron gráficos clave como la evolución de casos y el análisis de infraestructura sanitaria.

2. Paneles de Power BI

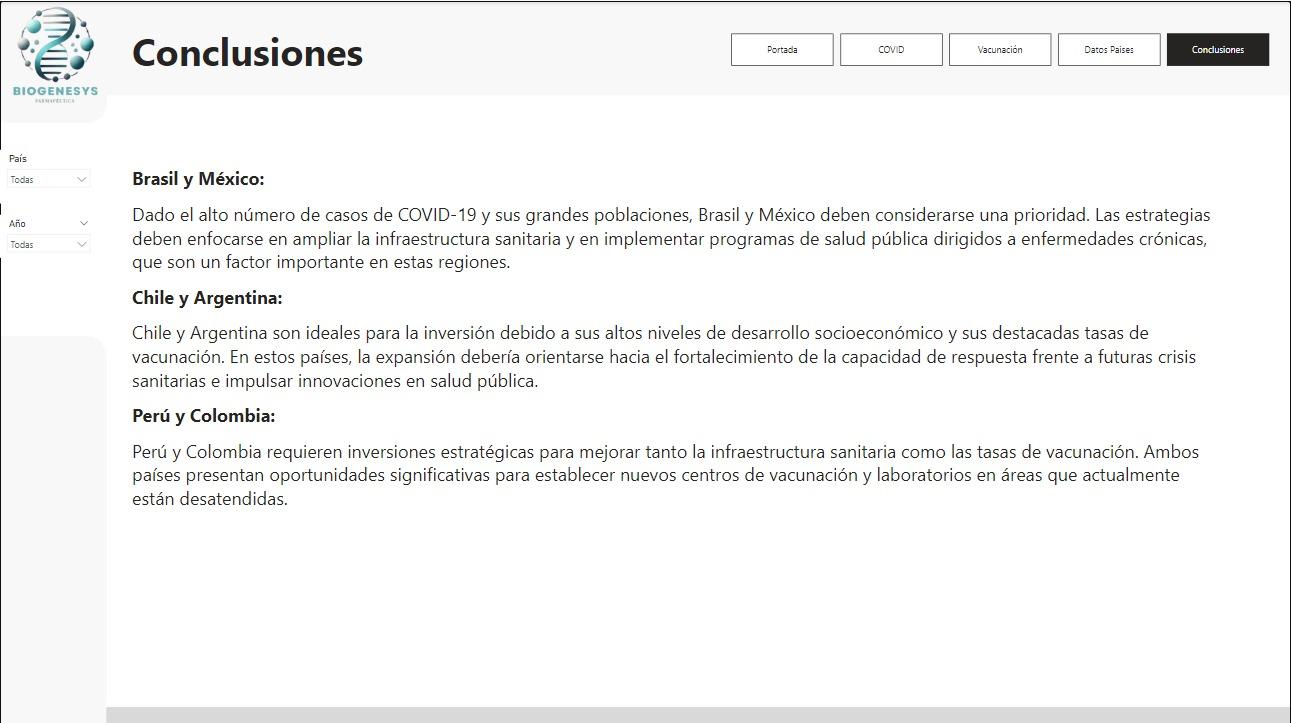
* Panel de Portada: Se indica nombre del Proyecto, y las pestañas COVID, VACUNACION, DATOS PAISES, Y CONCLUSIONES.
* Panel COVID: Muestra métricas de casos confirmados, fallecidos y tasas de vacunación con segmentadores para años y países.
* Panel VACUNACION: En esta pestaña se muestran datos relacionados a la vacunación de cada país: cantidad de dosis totales aplicadas y dosis por persona por páis.
* Panel DATOS PAIS: Datos referidos a la situación de cada país: población, densidad de población, población rural y urbana, composición etaria, Índice de Desarrollo Humano, etc.
* Panel CONCLUCIONES: El análisis confirmó que las políticas de vacunación y el acceso a infraestructura médica desempeñan un papel crucial en la gestión de la pandemia.











### Conclusiones Generales y Recomendaciones Estratégicas

### Conclusiones y Recomendaciones

Para sintetizar los resultados del análisis, nos enfocamos en los factores clave que influencian la expansión de laboratorios en Brasil, Argentina, Chile, Perú, Colombia y México. Estos factores, que incluyen incidencia de COVID-19, tasas de vacunación, infraestructura sanitaria, datos demográficos, indicadores socioeconómicos y factores ambientales, proporcionan un panorama integral que permite identificar áreas prioritarias y formular recomendaciones estratégicas.

#### Síntesis de Resultados

1. **Incidencia de COVID-19**: Brasil y México presentan las cifras más elevadas en casos confirmados y fallecimientos, lo que sugiere una alta carga de enfermedad y un impacto considerable en sus sistemas de salud. Estos países presentan una necesidad urgente de intervención con recursos adicionales para mitigar la propagación y mejorar la respuesta sanitaria. En cambio, Chile y Argentina, con tasas de vacunación relativamente altas y una gestión de la pandemia más efectiva, evidencian un control más firme sobre el COVID-19, lo cual puede reducir su necesidad de intervención directa.
2. **Tasas de Vacunación**: Chile se destaca por una alta tasa de vacunación, reflejando un sistema de salud eficiente y con alta capacidad de respuesta ante la pandemia. La situación de vacunación en Perú y Colombia, en cambio, presenta oportunidades de mejora. Invertir en programas de vacunación en estos países podría potenciar el control sobre futuras olas de contagio y aumentar la resistencia comunitaria a la transmisión del virus.
3. **Disponibilidad de Infraestructura Sanitaria**: A pesar de contar con una infraestructura más robusta, Brasil y México pueden estar enfrentando sobrecargas en sus hospitales y personal debido al alto volumen de casos de COVID-19. La densidad de hospitales y médicos en estos países indica una capacidad instalada que, con apoyo, puede optimizarse. Por otro lado, Perú y Colombia presentan deficiencias en personal médico y disponibilidad de camas, lo cual señala la necesidad de mejorar la infraestructura para reducir la vulnerabilidad de sus sistemas de salud, especialmente en áreas rurales.
4. **Datos Demográficos**: La gran población en Brasil y México, tanto en zonas urbanas como rurales, implica una necesidad diversa de servicios de salud en entornos urbanos y remotos. Esto justifica una estrategia que incluya laboratorios móviles o centros de salud itinerantes en áreas rurales, además de la ampliación de infraestructura en ciudades. En contraste, Chile y Argentina tienen poblaciones más concentradas en áreas urbanas, lo cual facilita la distribución de recursos y sugiere que una mayor inversión en centros urbanos podría ser suficiente para atender sus necesidades de salud.
5. **Indicadores Socioeconómicos**: Los altos índices de desarrollo humano y PIB per cápita en Chile y Argentina reflejan una capacidad de inversión en salud que puede facilitar la implementación de infraestructura avanzada. Esto les permite ser autosuficientes en la expansión de sus sistemas de salud. Perú y Colombia, en cambio, enfrentan desafíos socioeconómicos que limitan la capacidad de sus gobiernos para invertir en salud pública de manera sostenible. En estos países, es fundamental que cualquier inversión en infraestructura esté alineada con estrategias de sostenibilidad y colaboración público-privada para maximizar el impacto y la eficiencia de los recursos.
6. **Factores Ambientales y de Salud Pública**: Brasil y México muestran altas tasas de enfermedades crónicas, como diabetes y afecciones relacionadas con la contaminación. Estos factores complican la gestión de la salud pública, ya que aumentan la vulnerabilidad de la población frente a enfermedades infecciosas y otras comorbilidades. En Chile, donde la tasa de comorbilidad es menor, la intervención sanitaria puede ser más sencilla y menos costosa en términos de mitigación de riesgos.

#### Recomendaciones Estratégicas

**Inversión en Infraestructura y Personal Médico**:

* Para Brasil y México, se recomienda un enfoque de apoyo a la infraestructura hospitalaria existente y una expansión estratégica en zonas de alta incidencia de COVID-19. La contratación de personal médico adicional y la construcción de unidades de atención especializadas en áreas urbanas y rurales pueden ayudar a reducir la presión sobre el sistema de salud.
* En Perú y Colombia, es crucial invertir en infraestructura básica y personal médico, enfocándose en áreas rurales y semiurbanas que carecen de servicios adecuados. Estas inversiones reducirían las brechas en la capacidad de respuesta sanitaria y mejorarían la cobertura de atención.

**Refuerzo de Programas de Vacunación**:

* Aumentar los programas de vacunación en Perú y Colombia fortalecería su respuesta frente al COVID-19 y otras posibles epidemias. Esto implicaría no solo la disponibilidad de vacunas, sino también la implementación de campañas educativas para promover la aceptación de la vacuna.
* En Brasil y México, una estrategia para aumentar las tasas de vacunación en áreas rurales y remotas permitiría reducir la carga de casos y mejorar la protección comunitaria, especialmente en zonas con alta prevalencia de comorbilidades.

**Centros de Atención Adaptados a Población Urbana y Rural**:

* Dado que Brasil y México tienen una gran población dispersa en áreas urbanas y rurales, se recomienda el desarrollo de centros móviles de salud o clínicas itinerantes que puedan adaptarse a las necesidades de cada comunidad. Esto permitiría una cobertura más amplia sin necesidad de infraestructura permanente en cada localidad.
* Para Chile y Argentina, la concentración de la población en áreas urbanas permite optimizar recursos invirtiendo en centros de atención de alta capacidad en ciudades, lo que facilitaría el acceso a la mayoría de la población.

**Estrategias de Salud Pública para Enfrentar Comorbilidades**:

* Brasil y México, con altas tasas de diabetes y afecciones por contaminación, deberían desarrollar programas específicos de prevención y manejo de enfermedades crónicas, integrados con las estrategias de control de COVID-19. La creación de unidades de atención preventiva en salud pública y campañas de concienciación podrían ayudar a reducir la carga de estas enfermedades a largo plazo.
* Chile, con menores tasas de comorbilidad, puede mantener su enfoque en la promoción de la salud y prevención, facilitando un sistema de salud ágil y preparado para enfrentar futuras amenazas sanitarias.

**Colaboración Público-Privada para Sostenibilidad**:

Dada la variabilidad en la capacidad de inversión de cada país, se sugiere una estrategia de colaboración público-privada en Perú y Colombia para hacer sostenible la expansión de la infraestructura sanitaria. Esto podría incluir subsidios, incentivos fiscales para empresas de salud y alianzas con organizaciones internacionales que apoyen la creación de centros de salud a largo plazo.

#### Reflexión Personal

A lo largo de este proyecto, he adquirido una serie de habilidades en el análisis y transformación de bases de datos utilizando diversas librerías de Python. Aprendí a crear gráficos y a usarlos de manera efectiva en archivos de Jupyter Notebook, trabajando todo en Visual Studio Code. Comprendí que, para manejar grandes volúmenes de datos, Python es una herramienta poderosa por su capacidad para gestionar datasets extensos.

También aprendí a integrar la información obtenida en Python, incluyendo gráficos, en dashboards de Power BI. Este proceso me enseñó la importancia de la visualización de datos en la toma de decisiones empresariales, especialmente en el contexto de la posible expansión de un laboratorio en América Latina.

Uno de los aspectos más valiosos de esta experiencia fue enfrentarme a los desafíos que surgieron en el camino. La aparición de datos con exceso de decimales y valores repetidos en distintas filas me obligó a hacer una revisión minuciosa, lo que me permitió aprender más sobre la limpieza y validación de datos. Estas dificultades me mostraron que, a través de la práctica, se adquiere un conocimiento más profundo sobre el recorrido que hace un analista de datos.

El proyecto fue una oportunidad positiva para aprender y realmente disfruté cada avance. Cada pequeño logro, desde la creación de un gráfico hasta la resolución de problemas complejos, fue una fuente de satisfacción. Me gustó mucho ver cómo cada componente encajaba en el panorama general del análisis de datos.

Si tuviera que comenzar este proyecto de nuevo, dedicaría más tiempo a la limpieza y análisis detallado de los datos. Abordaría los problemas de calidad de datos de manera más exhaustiva, asegurándome de que los registros sean precisos y consistentes. Además, aprovecharía la oportunidad de profundizar aún más en los análisis, para poder ofrecer recomendaciones mejor fundamentadas.

En definitiva, este proyecto no solo me permitió adquirir nuevas habilidades técnicas, sino que también me enseñó la importancia de la colaboración con compañeros e instructores. La consulta y el intercambio de ideas con mis pares y profesores amplió nuestro conocimiento y resolvió dudas que surgieron en el camino. Ha sido una experiencia enriquecedora que disfruté y que me dejó lecciones valiosas para el futuro.