**Expansión Estratégica de Biogenesys con Python**

**Nombre del autor: Piras Mariana.**

**Email: marianapiras@hotmail.com**

**Cohorte: DAPT02**

**Fecha de entrega:** 14-09-2024

**Institución:** **BIOGENESYS,** es una empresa farmacéutica que busca identificar las ubicaciones óptimas para la expansión de laboratorios farmacéuticos, basándose en el análisis de datos de incidencia de COVID-19, tasas de vacunación y la disponibilidad de infraestructuras sanitarias. La meta es **optimizar la respuesta a los efectos de la pandemia y postpandemia con el fin de mejorar el acceso a las vacunas.**

# Introducción

El proyecto consiste en realizar un estudio para la empresa farmacéutica **BIOGENESYS** para ayudar en su estrategia de expansión en Latinoamérica específicamente en Colombia, Argentina, Chile, México, Perú y Brasil. Este proyecto es vital para que estén preparados y puedan tener una respuesta rápida ante cualquier situación futura que pueda surgir, la directiva ha propuesto que para poder hacer esta inversión necesitan ubicar regiones y recolectar datos que sean de importancia y que les ayuden a tomar decisiones. Por lo que como analista de datos debo utilizar las distintas herramientas vistas en la carrera para poder ayudar a los directivos a tomar decisiones en este proceso.

Los objetivos son concretos y medibles:

* 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.

Con los datos proporcionados ofreceré insights valiosos sobre posibles ubicaciones estratégicas para futuros laboratorios y centros de vacunación. En el análisis abordaré aspectos como la demanda de vacunas, la logística de distribución y la infraestructura sanitaria existente. Este análisis de tendencias y demografía es fundamental para comprender el entorno del mercado en los países de interés.

# Desarrollo del proyecto

El desarrollo del proyecto integrador se llevó a cabo en cuatro avances:

**AVANCE 1:**

**Carga y transformación de los datos**

Para el primer avance me centré en el data set proporcionado y en el archivo adicional "readme.txt" que contiene información detallada sobre las columnas seleccionadas. Teniendo presente la importancia de una adecuada preparación y limpieza de datos para el análisis subsiguiente.

Como primer paso leí el archivo "readme.txt" para poder familiarizarme con los datos del data set.

Posteriormente creé un entorno virtual de trabajo con el nombre .proyecto\_integrados\_m4\_venv y lo active con el comando .proyecto\_integrador\_m4\_ven\scripts\activate. Luego con el comando pip install, en la terminal, instale las librerías de pandas y numpy, también realice la carpeta de requierements.txt con el comando pip freeze>requierements.txt, donde estarán incluidas todas las dependencias que el proyecto necesita.

Una vez abierto mi espacio de trabajo en jupyter notebook con el nombre PIDA\_M4\_Mariana\_Piras.ipynb seleccione el kernel creado en primera instancia.

Importe pandas con el comando import pandas as pd, ya que esta librería la necesito para poder leer mi data set y numpy.

Leí el archivo data\_latinoamerica.csv en visual studio code con pd.read\_csv().

La siguiente pregunta nos pedía que corroboremos el número de filas y columnas del data set, para lo cual use el comando df.shape.

Para seleccionar los países donde se expandirán realicé .isin[] sobre la columna “country\_name”.

La siguiente consigna nos pedía filtrar los datos por fechas, pero previo a ello debí cambiar el tipo de datos de la columna “Fecha” ya que se encontraba como string por lo que lo pase a tipo datatime con .astype.

Luego verifique los datos faltantes por columnas y realice una limpieza donde los valores eran mayores de 4 millones de datos faltantes, con isnull().sum().

Aplico un segundo filtro a mis datos por la columna “location\_key”, con el objetivo de lograr conseguir in data set más manejable y con menos datos faltantes.

Para comparar a nivel país y rellenar los valores faltantes use un ciclo for para completar las los valores nulos de las columnas ( new\_confirmed, new\_deceased, cumulative\_confirmed, cumulative\_deceased, rainfall\_mm, elative\_himidity, average\_temperature\_celsius, mínimum\_temperature\_celsius, maximun\_temperature\_celsius) con las medias de cada país y para las comulmas de new\_recovered, cumulative\_recovered y cumulative\_vaccine\_doses\_administered no hace falta iterarlas con un ciclo ya que solo cambiaré los valores nulos por cero con la función fillna().

Una vez verificado que no tengo instancias vacías, guardo el data set filtrado por fechas en una carpeta “EXPORT”, para usarlo en los futuros análisis ya que es más pequeño y fácil de manejar. Con el comando .to\_csv(ruta\_guardado).

Luego reinicie el notebook para borrar un poco de memoria utilizada y volví a cargar pero esta vez leí el archivo filtrado.

Aplique un ciclo for para hacer el cálculo de las medidas descriptivas y métricas importantes con el comando . describe() para poder responder las preguntas proporcionadas.

Como último paso del avance numero 1 realice el cálculo de las medidas descriptivas que el método describe no proporciona, están fueron la mediana, varianza, rango y moda, mediante la creación de una función denominada describir\_datos(), donde utilizo un ciclo for para poder iterarlos, una vez calculados respondí las peguntas solicitadas.

**AVANCE 2:**

**Análisis Exploratorio – Visualización**

Ya con los datos limpios, en esta fase del proyecto, se vas a profundizar en el análisis exploratorio de datos para extraer insights valiosos que orienten en la planificación estratégica de la expansión de laboratorios farmacéuticos. A través de visualizaciones claras y detalladas, se buscará descubrir patrones, tendencias y anomalías en los datos de incidencia de COVID-19, tasas de vacunación y disponibilidad de infraestructura sanitaria. El objetivo es crear un cuadro completo que permita a los directivos tomar decisiones informadas sobre dónde ubicar los nuevos laboratorios y centros de vacunación para maximizar el impacto positivo en la salud pública.

Para comenzar con el avance 2 del proyecto, sobre el archivo utilizado en el avance anterior, importe las librerías necesarias para esta etapa que fueron, numpy, pandas, matplotlib, seaborn y os, también realice la importación de los warnings, para ignorarlos cuando los mismos me indiquen que métodos están por ser deprecados. Para proseguir con la importación de los datos finales filtrados para que a partir de ellos poder comenzar con el análisis estadístico y la realización de las visualizaciones de los datos.

Por medio de un ciclo for imprimí las estadísticas básicas de cada columna.

Antes de seguir con el análisis corroboré que tipo de dato era mi columna “date”, ya que necesitaba ponerla como índice de mi datafreme para poder realizar análisis que variaran con el tiempo y para ello necesitaba que los datos fueran del tipo date, efectivamente era de tipo object por lo que tuve que cambiarlos a date mediante el uso de to\_datetime de pandas y luego sí ponerla como índice de mi dataframe mediante set\_index().

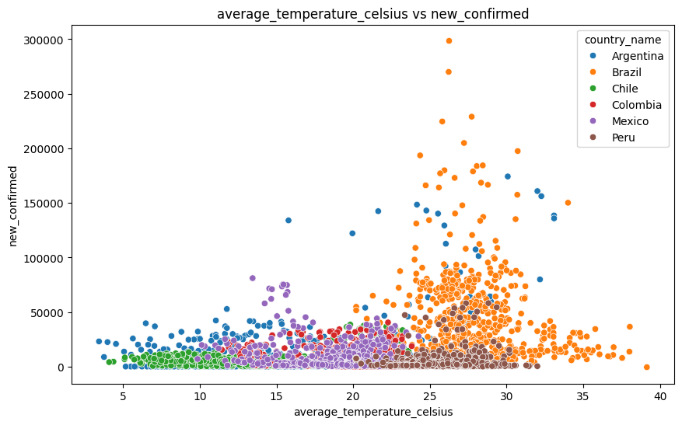
Paso seguido realizaré una matriz de correlación se definió un umbral cuyo valor absoluto sea mayor a 0.5 y se graficaron en un heatmap para ver si existe alguna relación entre las variables para tener en cuenta para un análisis posterior.

Ya teniendo una idea de las variables que tienen mayor correlación, se podría hacer un análisis más detallado de estas variables en un análisis posterior y una toma de decisiones basadas en los datos.

A continuación realice una rápida revisión gráfica de la gran mayoría de las columnas para ver cómo se comportan los datos, para esta primera aproximación utilicé barplots utilizando un ciclo for y la librería seaborn.

Repito el proceso del ciclo for pero en este caso realizó histogramas para ver la distribución de algunos datos: new\_confirmed", "new\_deceased", "cumulative\_confirmed", "cumulative\_vaccine\_doses\_administered", "average\_temperature\_celsius", "minimum\_temperature\_celsius", "maximum\_temperature\_celsius", "relative\_humidity".

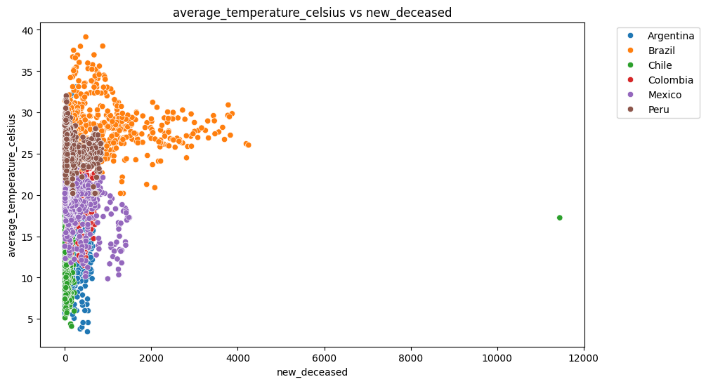
Realizó un análisis enfocado en algunas variables específicas, en este caso voy a analizar la variable average\_temperature y la variable new\_confirmed, con el fin de tratar de entender el comportamiento de los casos respecto a la temperatura.



Puede notarse que no hay correlación prácticamente entre los nuevos confirmados y el promedio de temperaturas, en parte estamos confirmando lo que se puede ver en el heatmap de principio al no tener coloración alguna.

Se puede observar que en algunos países la temperatura no influye en la cantidad de casos confirmados, sin embargo en otros países si se puede observar una relación entre la temperatura y la cantidad de casos confirmados.

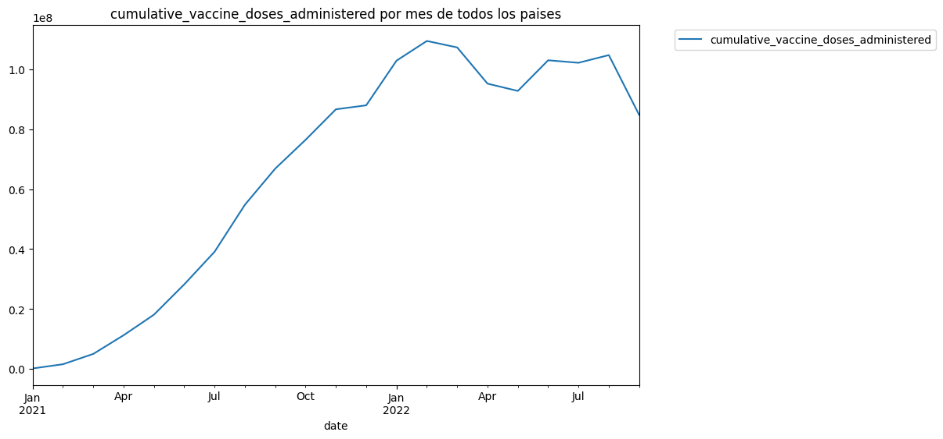
Como veo que la temperatura tiene alguna relación con los casos confirmados, voy a realizar un análisis de la temperatura promedio con la cantidad de muertes.



En la gráfica anterior se ve un comportamiento similar al de los casos confirmados, en algunos países la temperatura no influye en la cantidad de muertes, sin embargo en otros países si se puede observar una relación entre la temperatura y la cantidad de muertes.

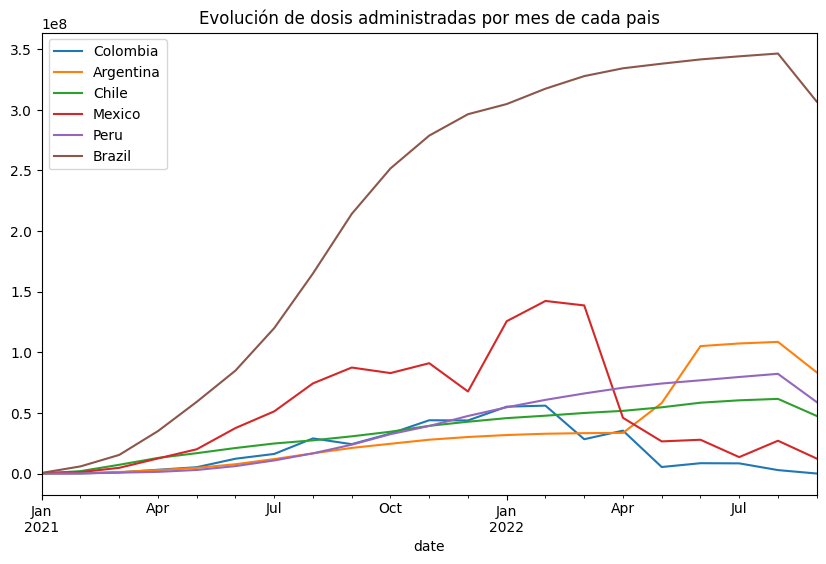
Como siguiente paso veo cómo se comportan las variables en un estudio longitudinal, es decir a lo largo del tiempo.

Por lo tanto, se hace una primera aproximación y observaremos las dosis administradas por mes en todos los países.



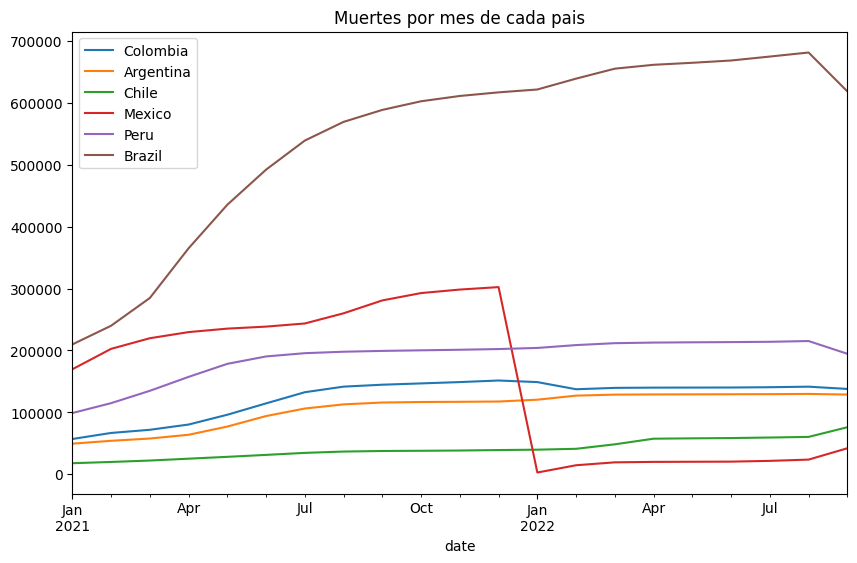
De dicha gráfica se puede deducir que hubo un aumento de la cantidad administrada de dosis en los países de interés (Argentina, Chile, Colombia, México, Perú y Brasil), sin embargo éste es un análisis general.

Al ver que el gráfico anterior no proporciona tanta información realice un análisis más detallado de las dosis administradas por mes en cada país y obtuve lo siguiente.



Se puede observar que en algunos países la cantidad de dosis administradas ha ido aumentando a lo largo del tiempo, sin embargo en otros países la cantidad de dosis administradas ha ido disminuyendo. Esto puede ser debido a muchos factores, como la cantidad de vacunas disponibles, la cantidad de personas que se han vacunado, entre otros.

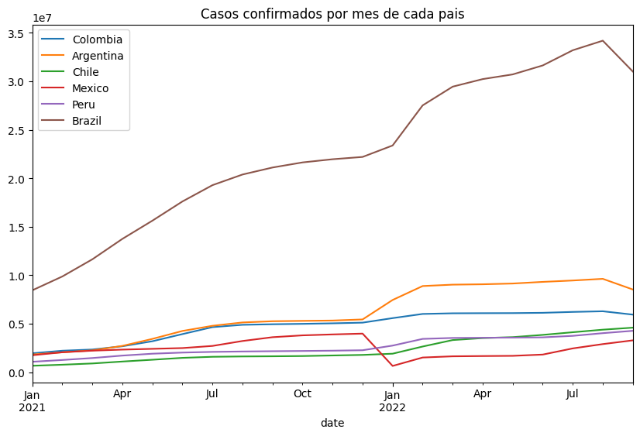
* Análisis de las muertes acumuladas por mes en cada país.



El país que más muertes ha tenido es Brasil, seguido de México, Colombia, Argentina, Perú y Chile. Sin embargo, esto no quiere decir que estos países sean los que peor han manejado la pandemia, ya que hay que tener en cuenta la cantidad de habitantes de cada país además de otros factores y cantidad de datos que tenemos.

La disminución de casos en México puede ser una pérdida de datos o una disminución real de casos (coincide esta disminución con el aumento que hubo en ese periodo de la administración de vacunas), esto se puede analizar en un estudio posterior.

* Análisis de casos confirmados por mes en cada país.



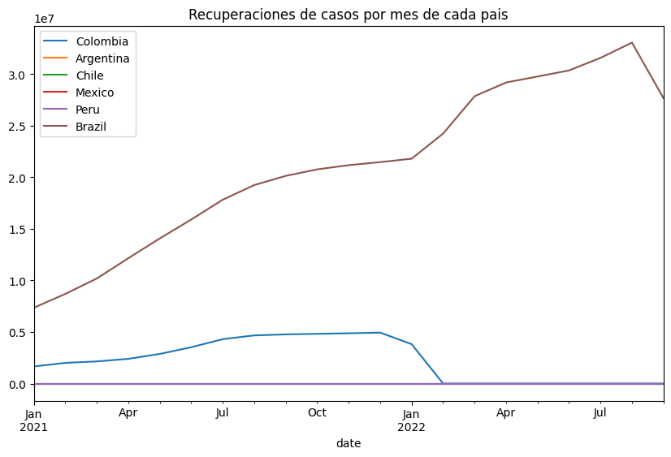
El país que más casos confirmados ha tenido es Brasil, seguido de México, Argentina, Colombia, Chile y Perú. Esto no quiere decir que manejaran peor la pandemia, ya que, como mencioné anteriormente, hay que tener en cuenta la cantidad de habitantes de cada país además de otros factores y cantidad de datos que tenemos.

Brasil muestra un aumento significativo en los casos confirmados a lo largo del tiempo. Esto sugiere que Brasil tuvo un brote más severo o una mayor cantidad de casos confirmados en comparación con los otros países. Los otros países (Argentina, Chile, Colombia, México y Perú) muestran fluctuaciones en los casos confirmados, pero sus números permanecen relativamente más bajos que los de Brasil durante todo el período.

Cada país tiene diferentes picos y valles en sus líneas, lo que puede indicar brotes específicos o medidas de control que afectaron la propagación del virus en diferentes momentos.

Las diferencias en las tendencias pueden reflejar la diversidad políticas de salud pública y las respuestas gubernamentales a la pandemia en cada país.

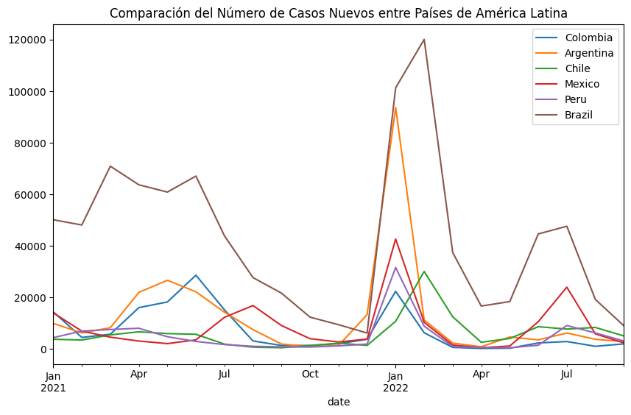
* Recuperación de casos por mes en cada país.



El país con mayor recuperación ha sido Brasil, pero es relativo ya que depende de la cantidad de habitantes y otros parámetros, sin embargo en esta gráfica se puede apreciar que solo tres países tienen valor para este parámetro, México, Chile y Argentina no poseen datos.

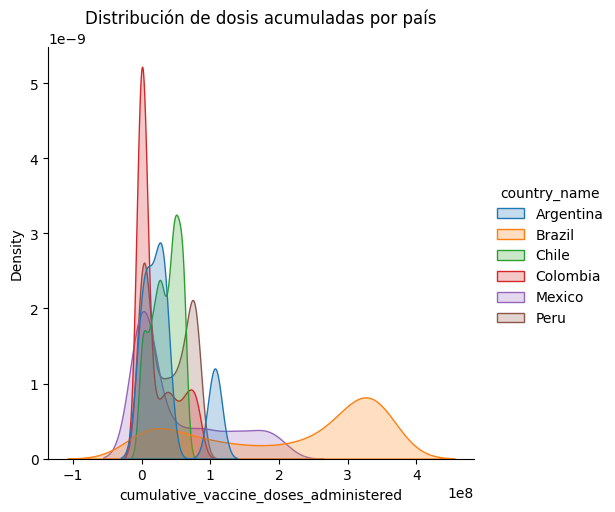
Para obtener mayor claridad, se aplicó la función value\_counts(), revelando 2736 observaciones con un valor de 0.0, dentro de un total de 1009 valores únicos en la columna. Esta distribución de los datos implica que no es posible derivar conclusiones sólidas ni confiables.

* Comparación de casos nuevos entre los países seleccionados.



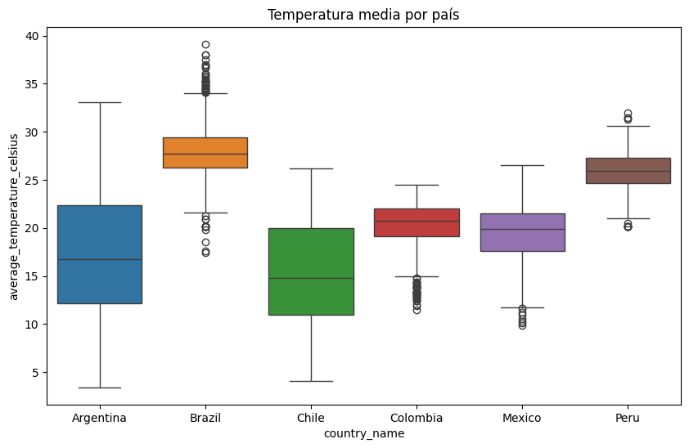
De los países que mejor manejó la pandemia fue Chile, seguido de Perú y Colombia, se puede observar que la cantidad de muertes y casos confirmados es menor en estos países, además de no ser tan desproporcionada la cantidad de vacunas administradas. Argentina, México y Brasil han tenido mayores fluctuaciones.

* Distribución de dosis acumuladas por país.



De los países que peor manejaron la pandemia fue Brasil, seguido de México, Colombia, Argentina, Perú y Chile, se puede observar que la cantidad de muertes y casos confirmados es mayor en éste país, además de ser desproporcionada la cantidad de vacunas administradas. Brasil a pesar de aplicar muchas dosis de vacunas es el país con mayor cantidad de casos y muertes confirmadas, sin embargo como mencioné anteriormente hay que tener en cuenta el número de habitantes de cada país.

Del gráfico del promedio de temperatura versus el número de nuevos confirmados por país:

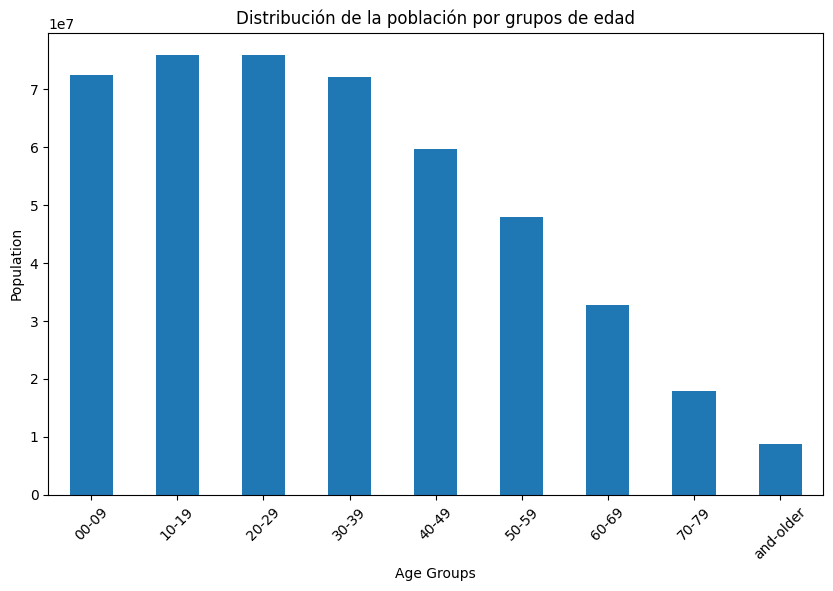


Se deduce que Chile y Argentina tienen rangos diferentes de temperatura y eso podría haber afectado la cantidad de casos confirmados y muertes, se puede observar que chile ha manejado mejor la pandemia que los demás países.

Esto poco a poco he llegado a concluir que los países donde se podría crear los nuevos centros serían donde hay más casos confirmados y muertes, ya que es donde se necesita más ayuda.

Pese a nuestro anterior análisis, no puedo tomar una decisión definitiva, ya que no he hecho un análisis detallado de los datos, por lo que recomiendo hacer un análisis más detallado y un estudio de los datos para poder tomar una decisión más acertada.

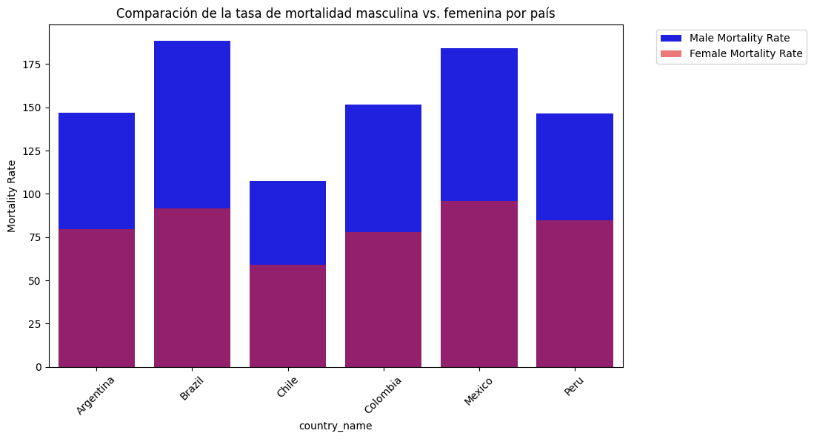
Seguiré viendo cómo se comportan las variables de a través de violinplot y analizando la distribución de la población por grupos de edad en los países seleccionados, para tratar de entender cómo se comporta la población en cada país y en un futuro poder tomar decisiones más acertadas.



Esta gráfica es interesante porque nos permite ver la población por grupos de edad en cada país, lo que nos puede ayudar a entender cómo se comporta la población y también nos genera la idea de que podríamos graficar la cantidad de vacunas por cada grupo etario.

Luego realicé las gráficas de vacunación por mes en cada grupo etario, las gráficas fueron muy similar a la evolución de las dosis administradas por mes en cada país, donde Brasil claramente tiene un aumento significativo.

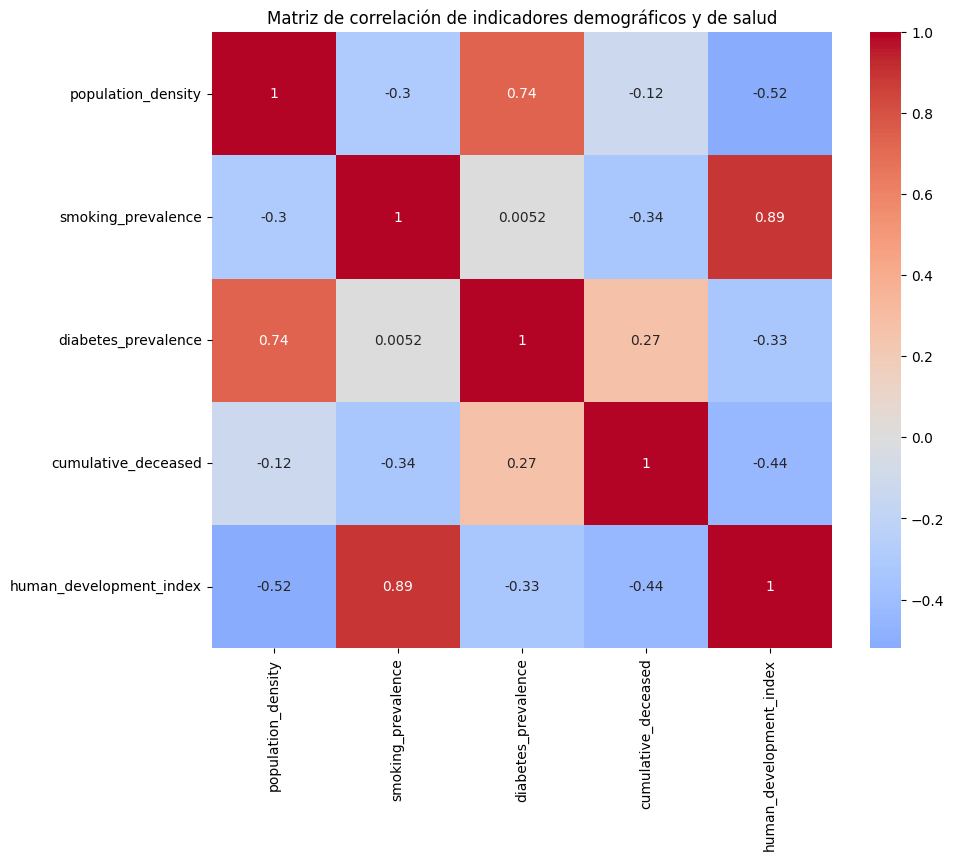
Reviso la mortalidad por genero para identificar la población más vulnerable en cada país, con esto ya podríamos tener una idea de cuál serían la población a vacunar primero.



De acuerdo a la gráfica de mortalidad, se puede observar que en todos los países la cantidad de muertes es mayor en los hombres que en las mujeres, por lo que se podría decir que los hombres son más vulnerables a la enfermedad o están menos protegidos.

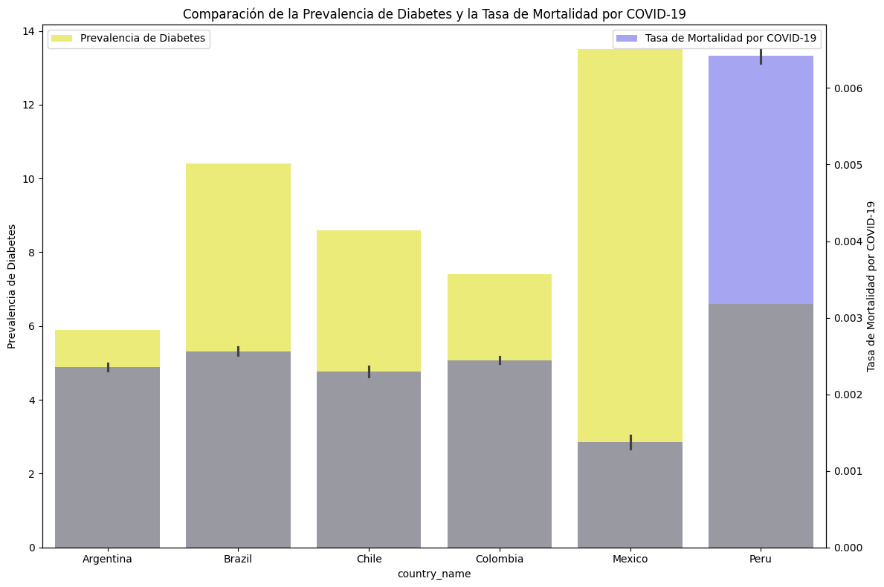
Si observo las enfermedades preexistentes o comorbilidades podría ver si hay alguna enfermedad que sea más común en los países seleccionados y que haya causado más muertes.

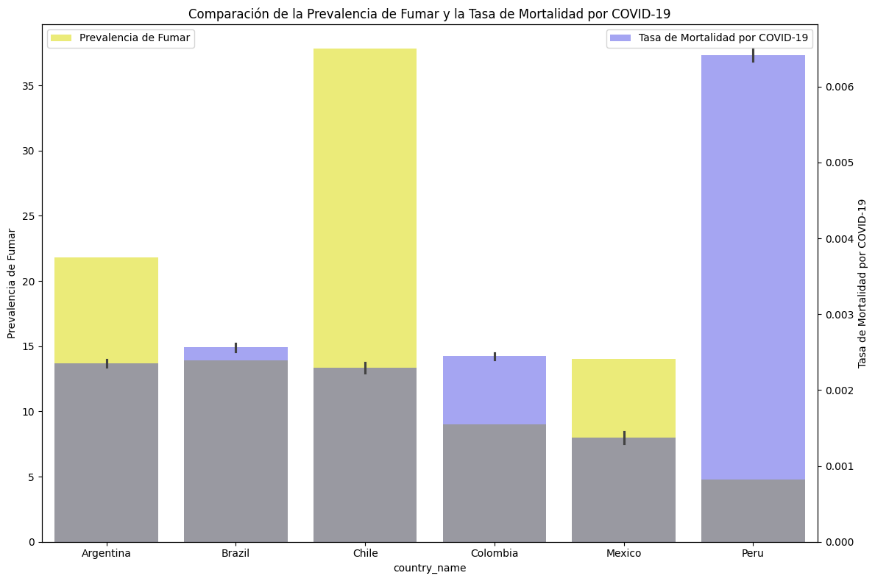
Una manera más detallada de ver las variables y la correlación entre ellas es a través de una matriz.



En esta matriz puedo observar que el fumar y tener diabetes son las enfermedades preexistentes que más afectan a la población en los países seleccionados, por lo que se podría decir que estas enfermedades son las que más han causado muertes en dichos países, sin embargo esto no quiere decir que sean las únicas enfermedades que han causado muertes sino que son las más comunes.

De acuerdo a esto veré cómo se comporta la mortalidad en los países seleccionados por enfermedades preexistentes o comorbilidades.





Las dos enfermedades muestran un comportamiento similar en los países seleccionados, sin embargo se puede observar que la diabetes es la enfermedad que más ha causado muertes en dichos países, seguido del fumar.

Como conclusión de este avance podemos decir que los países que más necesitan ayuda son Brasil, México y Colombia, ya que son los que más casos confirmados y muertes han tenido, además de ser los países que menos vacunas han administrado.

Las enfermedades preexistentes que más han afectado a la población en los países seleccionados son el fumar y la diabetes, por lo que se podría decir que estas enfermedades son las que más han causado muertes, sin embargo no quiere decir que sean las únicas enfermedades que hayan causado muertes sólo son con las que contamos datos.

**AVANCE 3:**

**EDA con Numpy y Pandas**

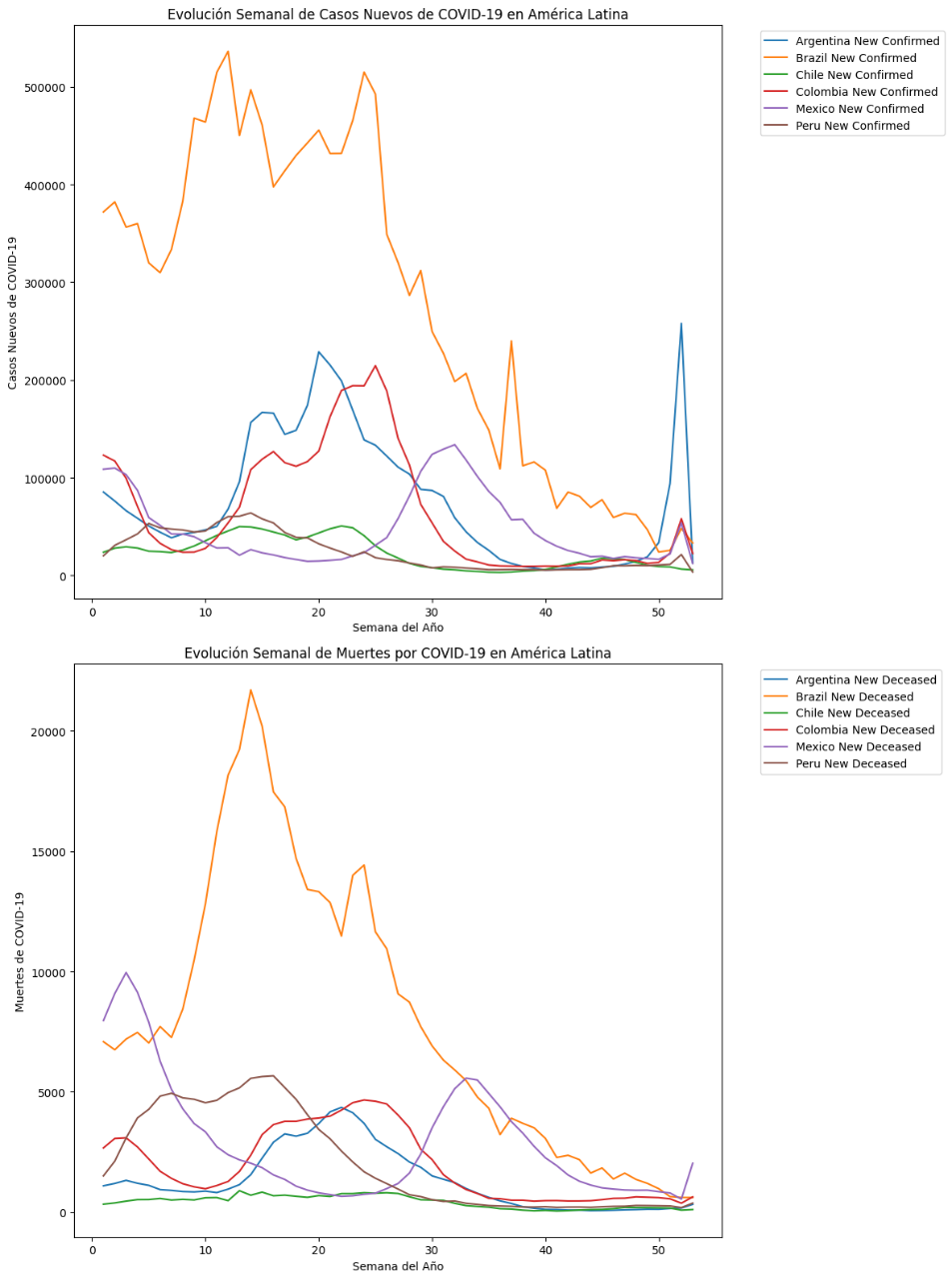
En este avance del proyecto, realizaré un análisis más profundo y sofisticado de los datos relacionados con la incidencia de COVID-19. El objetivo en este avance será pulir y preparar los datos para una visualización avanzada que permita identificar con precisión las ubicaciones más estratégicas para la expansión de los laboratorios farmacéuticos. Este paso me acercará a ofrecer recomendaciones basadas en datos sólidos que podrán guiar eficazmente las decisiones de inversión de la empresa.

Para comenzar con la realización del avance número 3, importe las librerías necesarias que fueron Pandas, Numpy, Matplotlib y Seaborn, posteriormente importe los datos para trabajar sobre ellos usando los datos finales filtrados.

Pedí la información del dataset para corroborar de que tipo es la columna date ya que en este avance voy a realizar un análisis de series temporales y es necesario que el tipo de dato sea de tipo “date”, viendo que el tipo era object procedí con el cambio del tipo mediante la función to\_datetime() y paso seguido setie la columna “date” como índice, para realizar el análisis de serie temporal mencionado, mediante la función .set\_index().

Creo una columna adicional en mi data con el número de semana usando la función .isocalendar().week. Y realizo un nuevo dataframe pero con los datos del año 2021 solamente ya que es el año más fuerte de pandemia y con mayores datos.

Armo un dataframe en base al recortado en el paso anterior. Lo agrupo por country\_name y week y sobre new confirmed y new deceased los sumos y reseteo en índice.

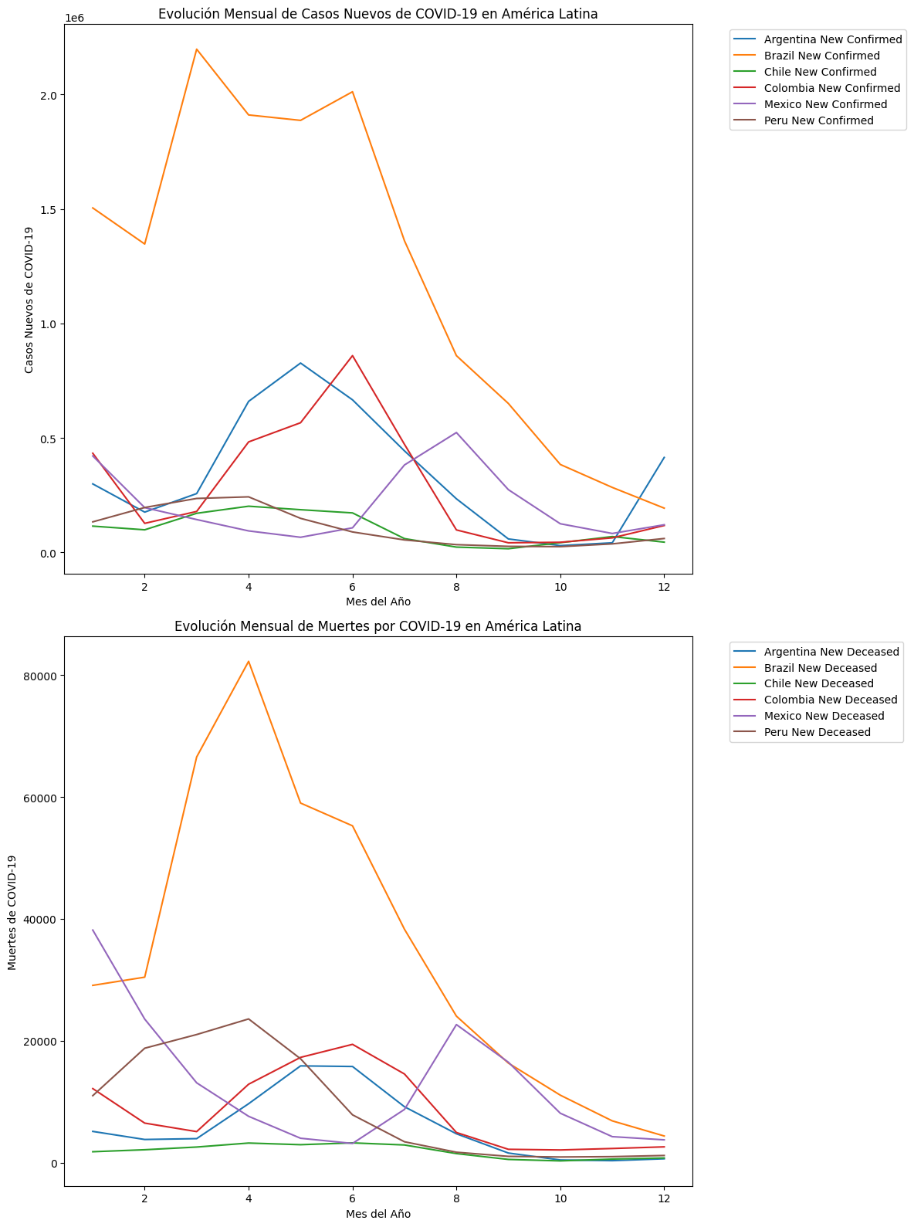
Analicé cómo evolucionan los casos de COVID-19 en América Latina a lo largo del tiempo por semana.

De los dos gráficos realizados se puede apreciar que en ambos Brasil tiene una curva muy dispar con los demás países.

En el caso del gráfico de nuevos casos confirmados se puede apreciar que Brasil tuvo un pico muy elevado hacia el comienzo pero luego los casos fueron disminuyendo, en el caso de Argentina se puede ver un aumento de los casos en fin de año, si bien Colombia y México tienen algunos picos entre las semanas 10 y 40 luego descienden. En el caso de Chile se nota un número relativamente bajo, comparado con los otros países, y bastante estable de nuevos casos.

Para el caso del gráfico de muertes tiene una curva similar al gráfico de casos confirmados en Brasil, pero en el caso de Argentina las curvas difieren, ya que el pico de contagios que se produjo hacia fin de año fue elevado pero no hubo tantas muertes, en general en todos los países puede notarse una baja en las muertes llegando a fin de año, esto puede deberse a las politicas de vacunación de cada país, y a las mejoras en el manejo de la pandemia.

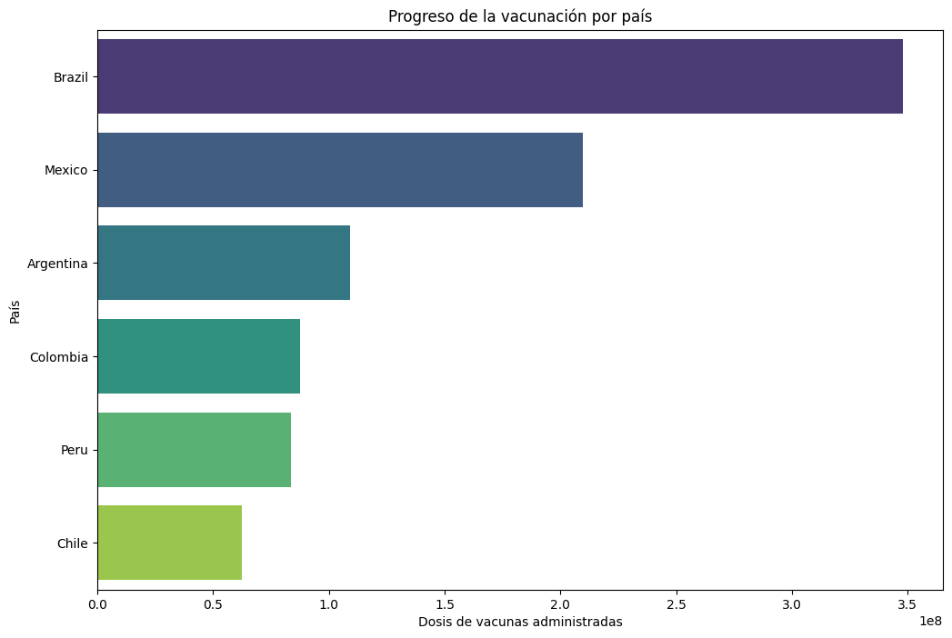
Seguido a éste análisis hice lo mismo pero esta vez por mes.



El análisis que se puede sacar de estos dos últimos gráficos es el mismo que en los anteriores, solo que el nivel de detalles de los últimos es menor al estar agrupados por mes.

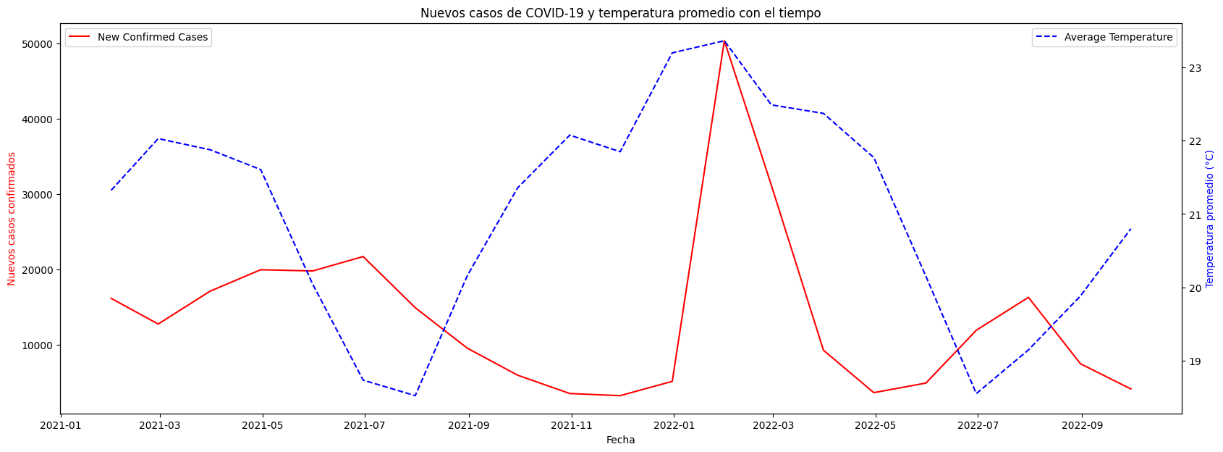
Si bien la cantidad de casos confirmados y de muertes disminuyen a lo largo del tiempo, no indica el término de la pandemia, debido a que ambos valores siguen siendo altos y en algunos casos con picos elevados hacía finales del año.

* Analizando el progreso de la vacunación por país:



Se puede apreciar que Brasil es el país que para la fecha ha administrado más dosis de vacunas seguido por México. Argentina, Colombia, Perú y Chile. Dicha información de contradice con los hallazgos anteriores, pero debemos tener en cuenta la cantidad de población y la cantidad de vacunas disponibles en cada país.

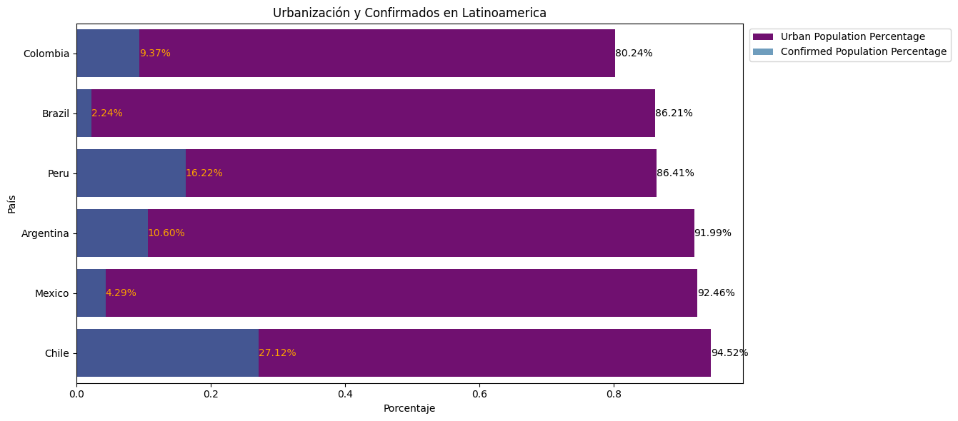
* Relación entre los nuevos casos y la temperatura promedio para los países seleccionados:



Podemos observar que la temperatura promedio ha ido aumentando y los casos confirmados han ido disminuyendo sin embargo hubo un pico antes del mes de marzo pero disminuyó rápidamente, esto puede ser por la cantidad de vacunas administradas o por la cantidad de personas que se han vacunado.

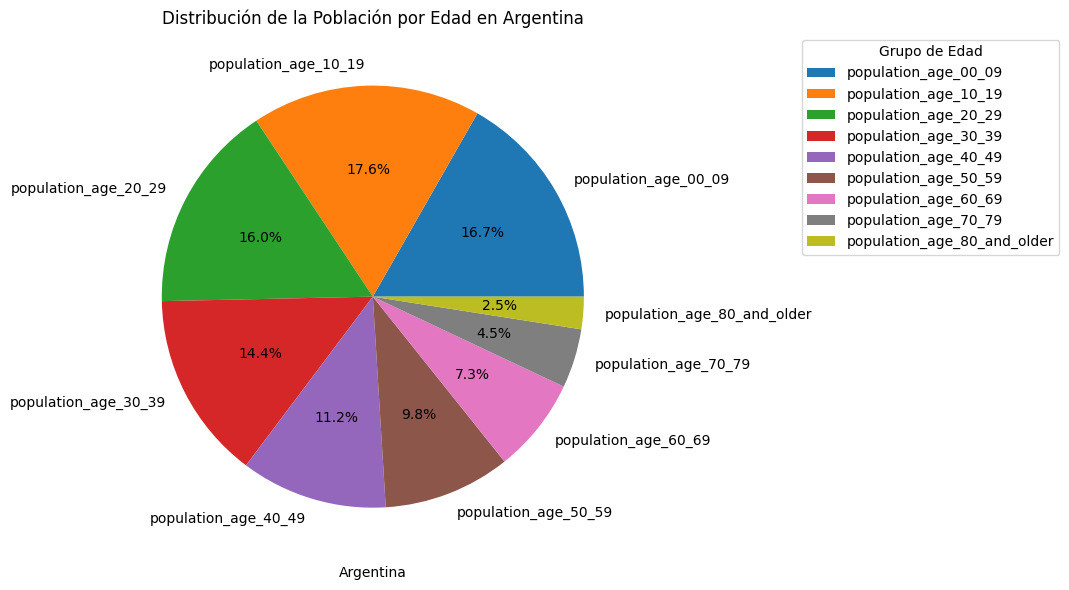
Si bien no son gráficas muy similares se puede notar cierta correlación sobre todo en el pico marzo del 2022.

Como cada país tiene un impacto debido a la urbanización y la densidad de la población, analice la densidad de la población en cada país. Para ello primero cree dos nuevas columnas a mi dataset: la de “urban\_population\_percentage” y “confitmed\_populationpercentage”, para la primera dividí la columna de population\_urban por la population y para la segunda dividí la cummulative\_confirmed por la population, para poder graficar el impacto de la urbanización en la propagación del covid-19.



El país con mayor impacto en la propagación fue chile, aunque es extraño ya que chile fue uno de los países que mejor manejó la pandemia, sin embargo esto puede ser debido a la densidad de población en chile.

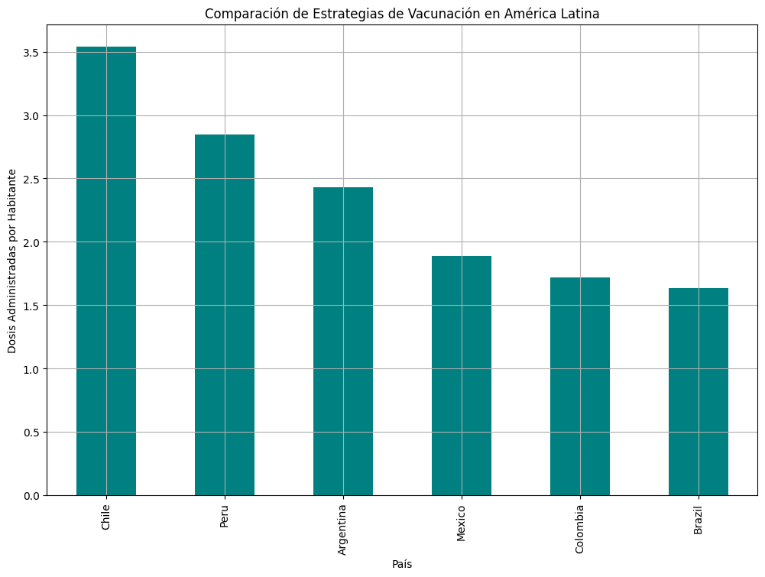
* Análisis de la densidad de la población por edad.



Pongo a modo de ejemplo uno de ellos, pero los gráficos muestran claramente la desigualdad en cantidades de población por grupos de edad en los países seleccionados aunque no proporcione muchas cosas interesantes. Es bueno tenerlo en cuenta para futuros análisis.

Sin embargo puedo empezar a ver cómo fueron de acertadas las estrategias de vacunación para mitigar la mortalidad y los casos confirmados.

* Comparación de estrategias de vacunación



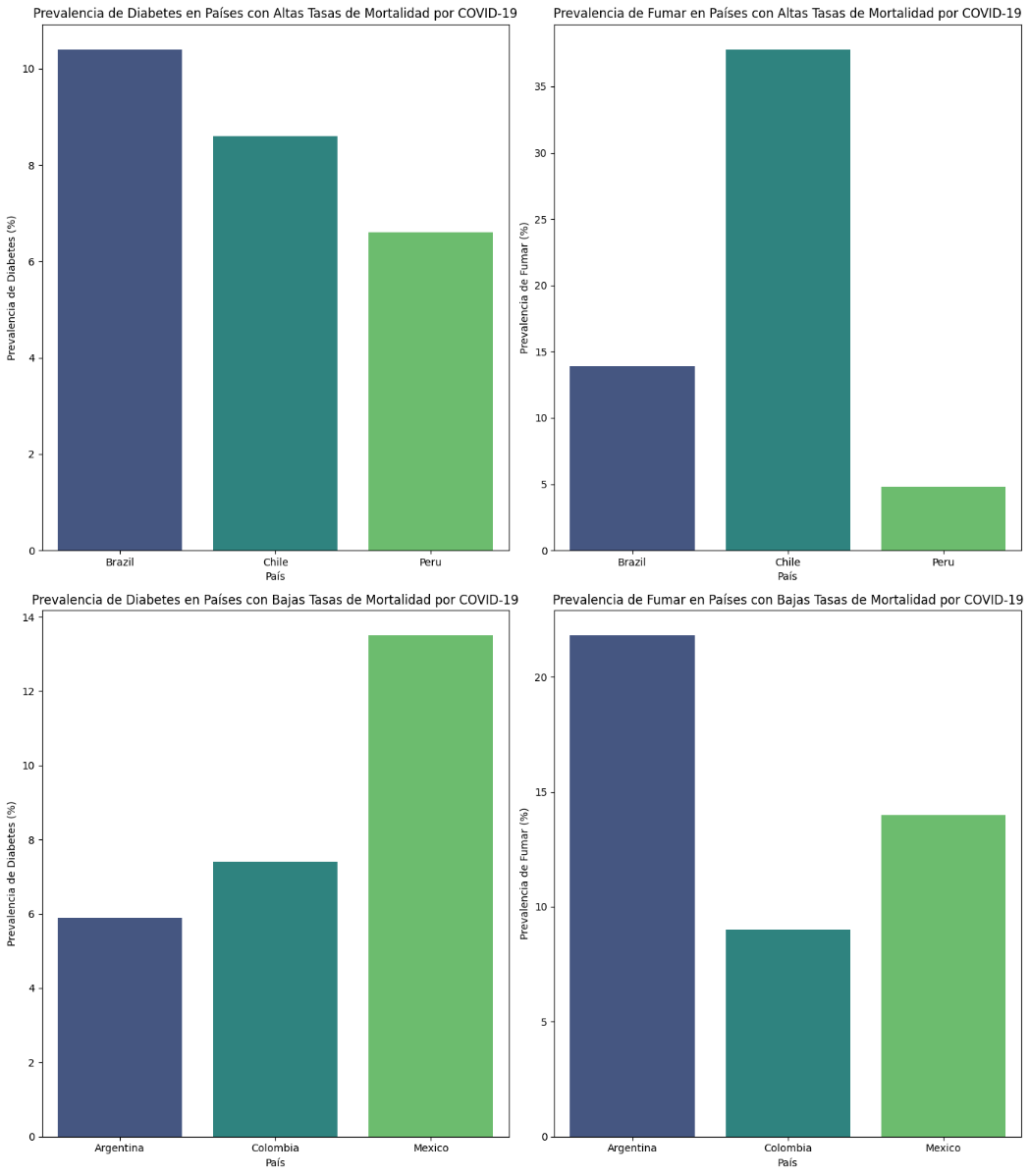
# 

Se puede apreciar que Chile sobresale en la cantidad de dosis administradas por habitante, lo que podría reflejar una estrategia de vacunación más agresiva o eficiente en comparación con los demás países representados.

Nuevamente se puede observar el buen resultado de las políticas de chile y una clara falla en Colombia y Brasil.

Con esto se puede ir definiendo los países que necesitan más ayuda y donde se deberían crear los nuevos centros.

* Prevalencia de Condiciones Preexistentes en Países con Altas y Bajas Tasas de Mortalidad.



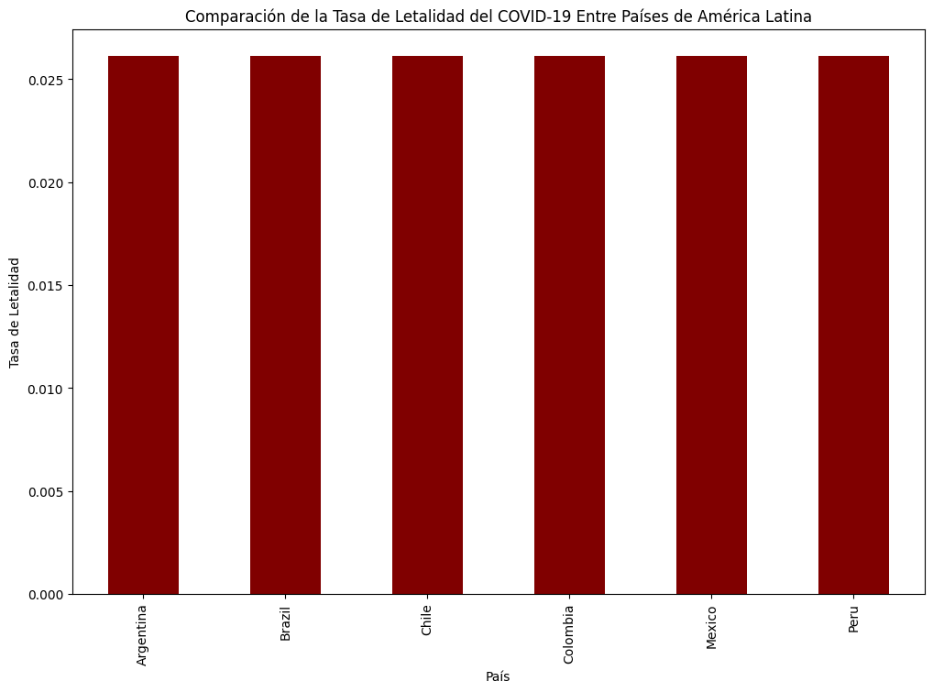
Ahora identifico que la diabetes afecta mucho más a México y Brasil, por lo que se podría decir que estos países son los que más han sufrido por la diabetes. Respecto a fumar se puede observar que afecta más a Chile y Argentina, por lo que se podría decir que estos países son los que más han sufrido por el fumar.

Sin embargo mientras Brasil aparece entre los países con mayor tasa de mortalidad, México aparece entre los menores.

Lo mismo pasas entre Chile (mayor) y Argentina (menor).

Por lo que difícilmente pueda entenderse una correlación entre Situaciones Preexistentes y Mortalidad por Covid-19.

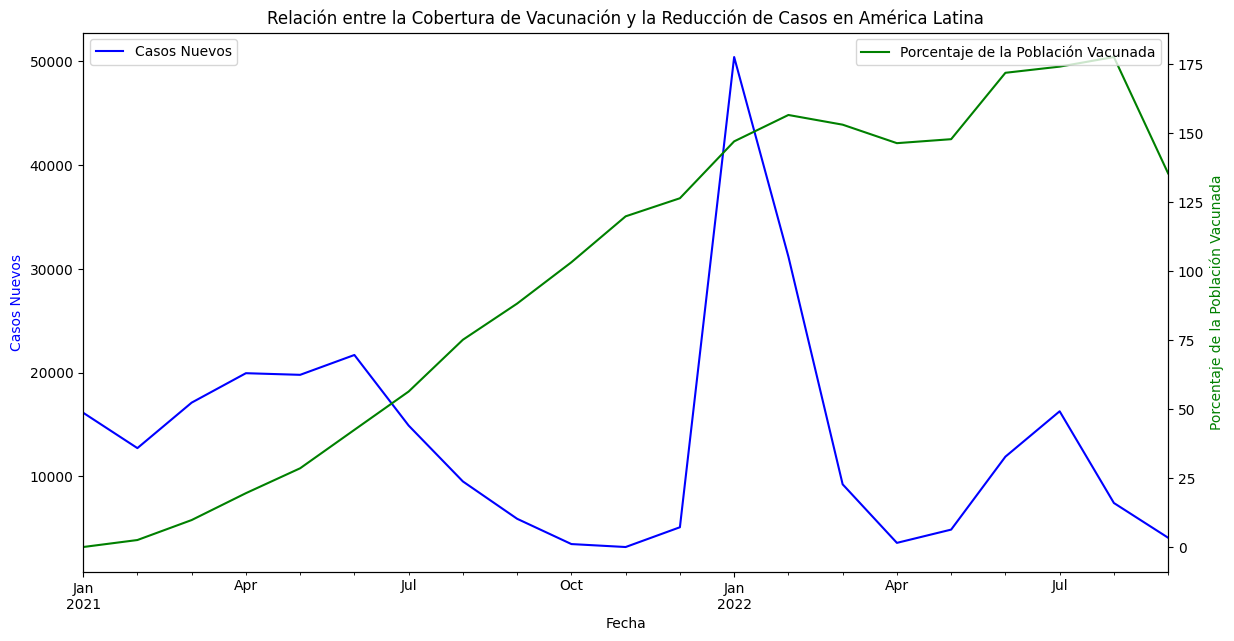
Viendo esto, me enfocaré en ver la letalidad por país.



Todos los países han tenido un comportamiento similar en cuanto a la letalidad, sin embargo apoyándose con las otras gráficas se puede observar que ha disminuido.

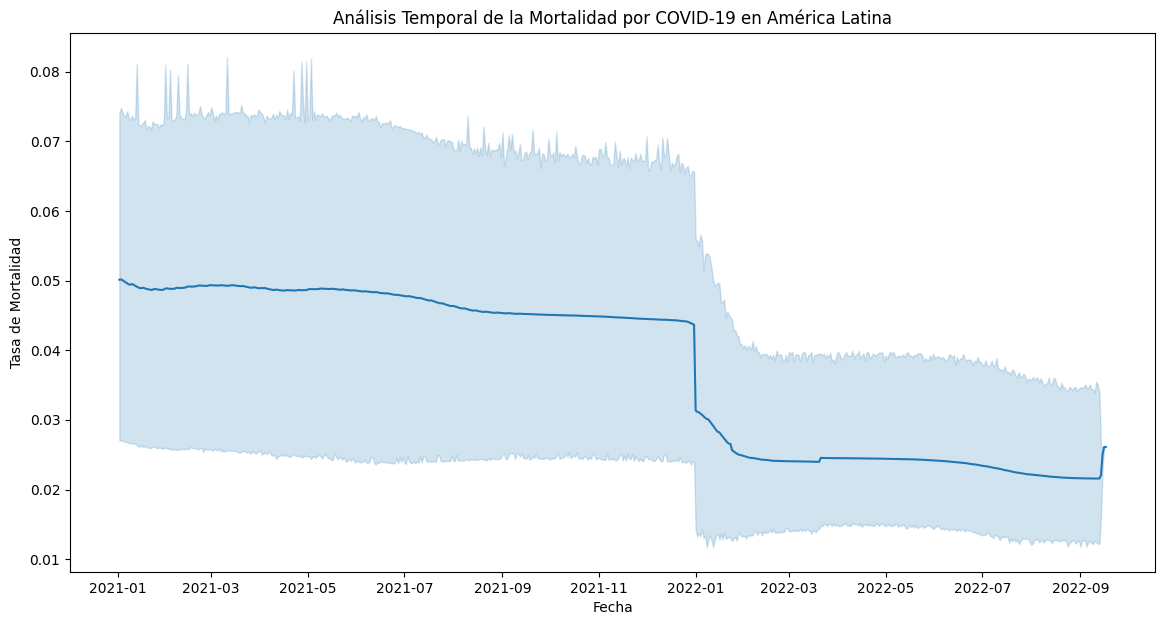
Los casos nuevos y las muertes han variado debido a la vacunación y a las políticas tomadas por los países, además de que se ha disminuido la cantidad de personas que se han contagiado y fallecido.

* Evolución de la vacunación en América Latina y la reducción de casos de COVID-19



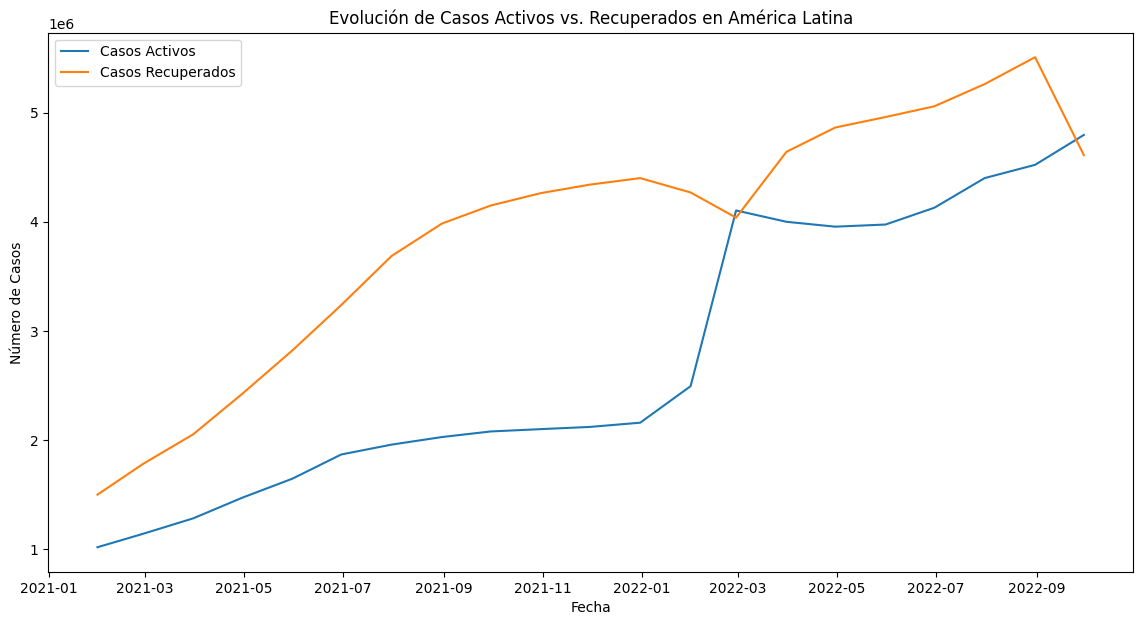
Los datos parecerían indicar que cuando se incrementó la población vacunada la cantidad de casos que venía creciendo comenzó a estabilizarse en valores de inicio de pandemia. Sin embargo en enero del 2022 hubo un pico muy marcado de nuevos casos.

* Análisis de cómo evoluciono la mortalidad:



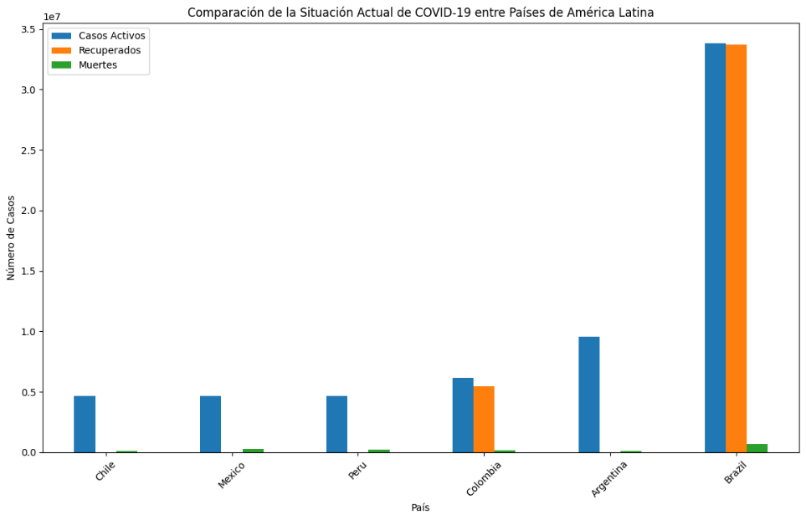
En esta gráfica se puede apreciar que la mortalidad por Covid 19 fue cayendo de manera pronunciada en la medida en que avanzó la campaña de vacunación en América Latina.

* Cómo evolucionan los casos activos y recuperados en América Latina.



La cantidad de casos activos fue creciendo mientras crecía a la par la cantidad de casos recuperados. Puede notarse que el pico de los casos activos coincide con una baja de los casos recuperados.

* En cuanto a la comparación de la situación actual:



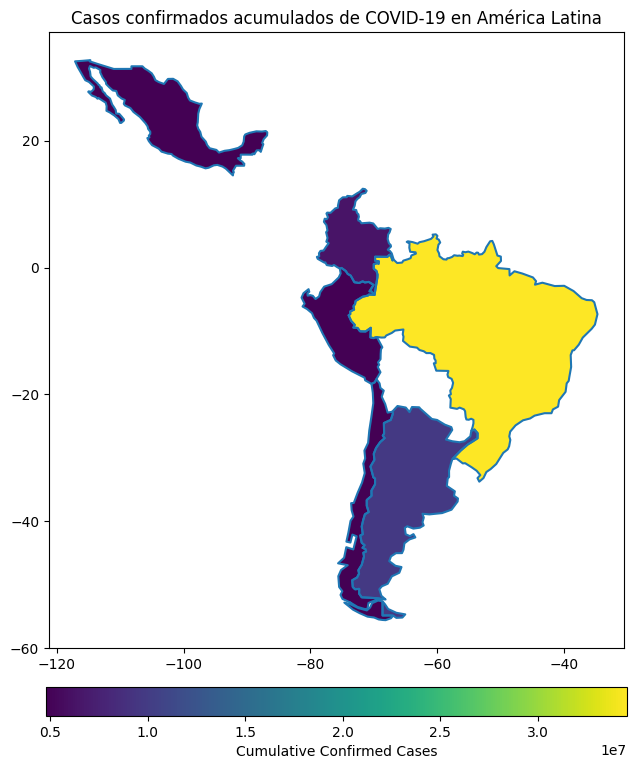
Por algún motivo el dataset no está persistiendo los datos acumulados de recuperados. Lo cual no nos permite ver la cantidad de recuperados actuales de cada país.

Para tener un mejor contexto graficaré la cantidad de vacunas administradas, los casos confirmados, muertes y recuperados.

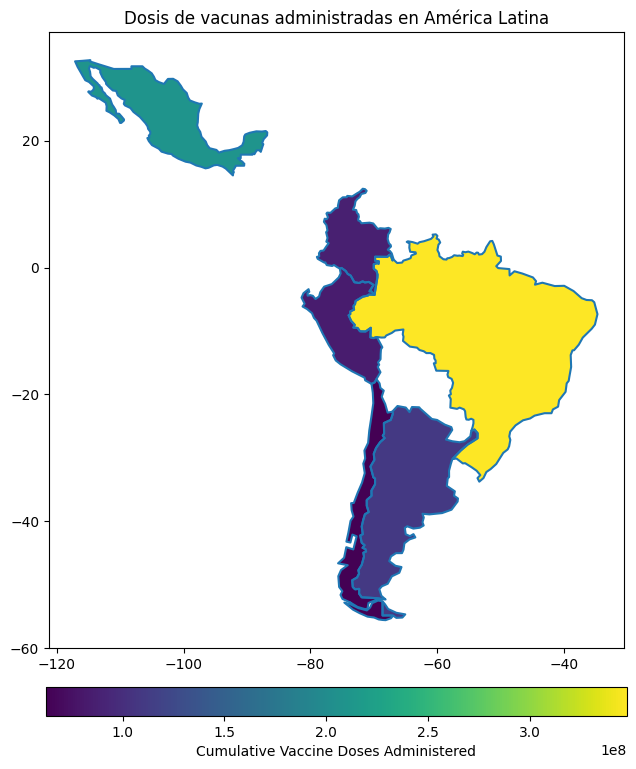
Para ello debí instalar la librería folium que se utiliza para crear mapas interactivos y visualizaciones geoespaciales.

Pero como el mapa mostrado no es muy detallado, realizó un análisis más detallado de las dosis administradas, casos confirmados, muertes y recuperados en cada país pero con gráficos más básicos manteniendo los mapas.

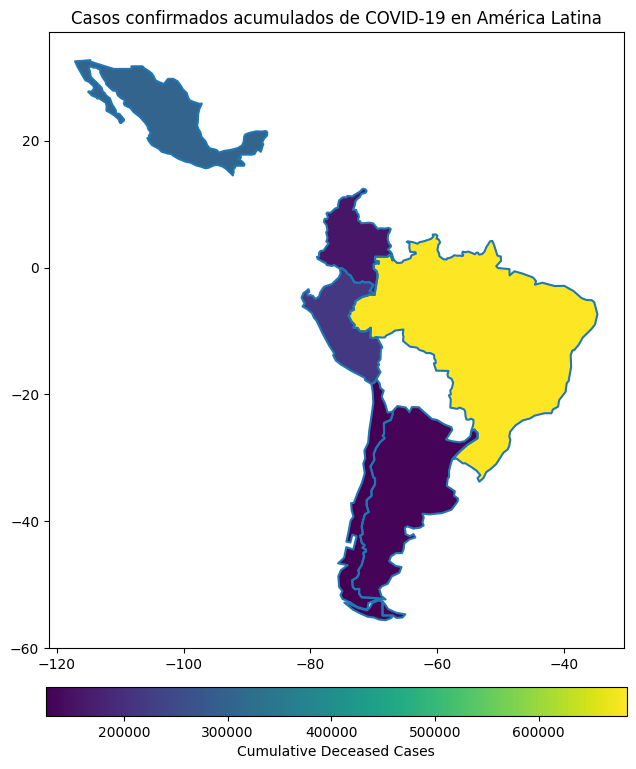
Para ello instalo geopandas y descargo el dataframe de geo desde la página web para conseguir un mapa más detallado y poder realizar mejores análisis. Los gráficos obtenidos son los siguientes:



En este gráfico puede notarse que Brasil tiene la mayor cantidad de casos confirmados acumulados en comparación con los restos de países estudiados, pero como mencioné anteriormente hay que tener en cuenta la densidad de la población para poder analizar más profundamente.



En éste mapa puede distinguirse que Brasil es el país que más dosis de vacunas administro seguido de México y Argentina.



Como en los mapas anteriores Brasil lleva la mayor cantidad acumulada de muertes y es seguido por México.

**AVANCE 4:**

**Aplicaciones Prácticas - Integración en Power BI**

Esta fase final del proyecto es clave para tener éxito, por lo tanto, me enfoqué en integrar, presentar los hallazgos analíticos y hacer reporting utilizando Power BI para la visualización. Con objetivo es sintetizar el análisis realizado en las fases anteriores en dashboards interactivos y reportes que faciliten la toma de decisiones estratégicas para la expansión de laboratorios y centros de vacunación. Estos insights ayudarán a la empresa farmacéutica a priorizar áreas de inversión basadas en la incidencia de COVID-19.

Es importante destacar la importancia de esta etapa para trasladar el análisis técnico a una plataforma que facilite la comprensión intuitiva de los datos por parte de los directivos. Identificando áreas prioritarias para la expansión.

El primer paso realizado fue en Python donde cree un script para poder realizar la conexión posterior con Power BI, para ello importe las librerías que se necesitaban, leí el archivo de datos finales filtrados.

Paso seguido me dirigí a Power Bi y cargue el archivo desde un script de Python, una vez lograda la carga, realice un boceto a mano del dashboard que queria realizar, teniendo en cuenta el objetivo del mismo que era poder responder a la consigna de negocio solicitada, identificar las ubicaciones óptimas para la expansión de la farmacéutica Biogenesys, basándome en la incidencia del COVID-19, las tasas de vacunación y la infraestructura sanitaria.

Una vez diseñado el dashboard que quería presentar, continúe con el paso de importar los gráficos desde el reporte de Python, para lo cual dentro del editor de código utilizado realice los script que me permitirían pasar los gráficos creados a Power Bi, una vez listos los cargue en Power BI, para ello antes analicé que gráfico llevaba desde python a la herramienta de visualización, tuve en cuenta que los mismos iban a ser estáticos, por los que los gráficos que quería que fueran dinámicos y permitieran filtrado ya sea por año o por país los realicé directamente en la herramienta. Para los mismos antes de empezar a crearlos tuve que realizar unas medidas rápidas que luego me permitieran lograr las visualizaciones que quería representar.

Una vez plasmados los gráficos, procedí a darle el formato a mi lienzo y visualizaciones.

Por ultimo pero no menos importante saque las conclusiones pertinentes al análisis realizados en todo el transcurso del desarrollo del proyecto integrador.

# EDA e insights

Del análisis exploratorio de los datos y de las visualizaciones obtuve una seria de insights que servirán para las conclusiones finales del reporte.

Se puede notar a simple vista en las visualizaciones que Brasil ha sido el país que más se despegó con respecto a los otros cinco países que formaban parte del análisis, teniendo las curvas más altas y con mayores resultados ya sea en cantidad de contagios, nuevos confirmados y recuperados. Con lo cual podría aproximarse a una posible conclusión de que es al país que más ayuda necesita, pero sin embargo una toma de decisión basada solo en estos parámetros seria incorrecta ya que también hay que tener en cuenta y comparar la cantidad de habitantes que dispone, la infraestructura sanitaria y el nivel socio económico de la población.

# 

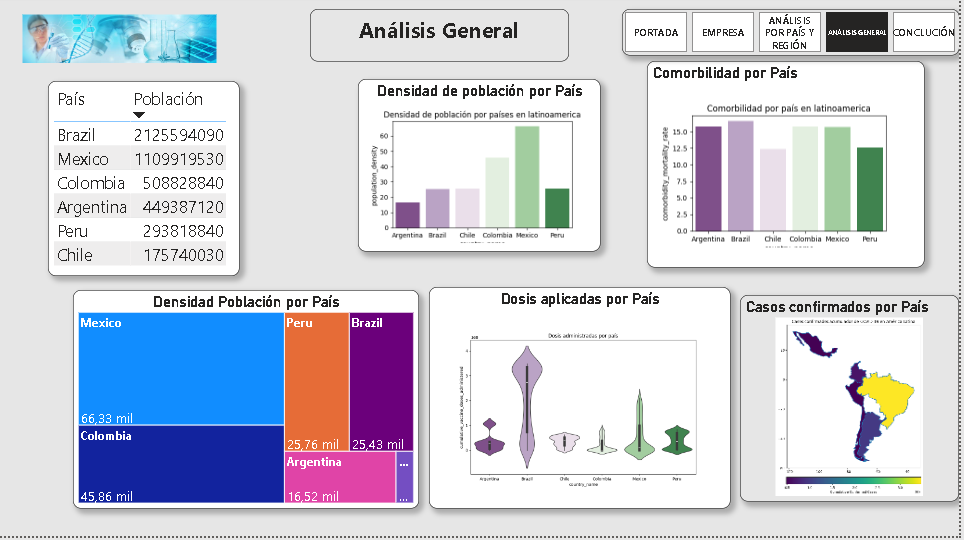
# Análisis del dashboard

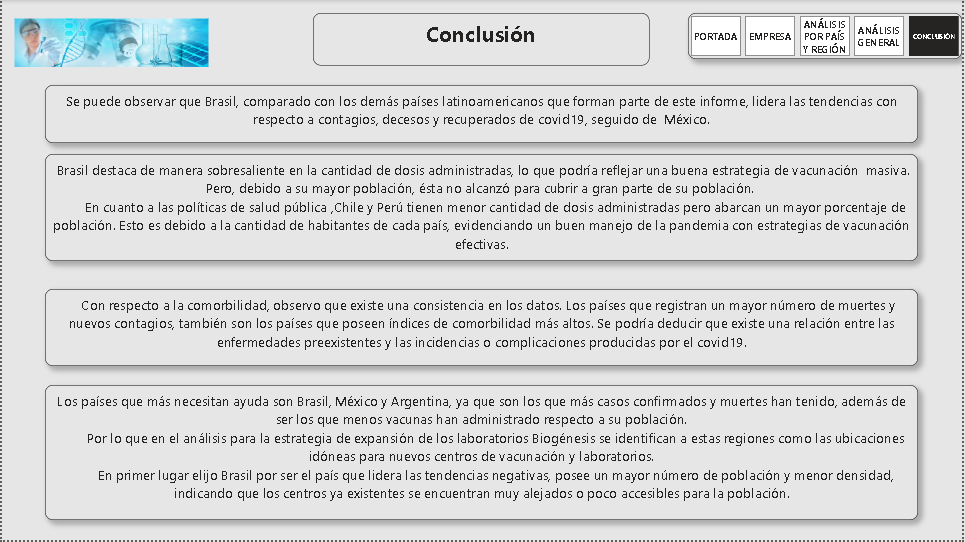
El dashboar consta de cuatro páginas de reporte: la primera es una página de portada donde se encuentran los botones interactivos que permiten navegar por las demás páginas del reporte según el usuario desee ver y analizar.

En la página número dos se encuentra un abreve descripción de la empresa para la cual se llevó a cabo el análisis exploratorio de los datos.



En la número tres se encuentran las visualizaciones por región y estacionalidad en la misma se podrá encortar dos filtros que permiten seleccionar el país y el año de interés para poder sacar las conclusiones pertinentes.

En la cuarta página se encuentran las visualizaciones estáticas del reporte donde se hace un enfoque más general a nivel de los seis países que conformaban el análisis de los datos.

Y por último pero no menos importante se encuentra la página de conclusiones basadas en el análisis de las visualizaciones presentadas.

# Conclusiones y Recomendaciones

1. Luego de analizar los diversos datos obtenidos, puedo observar que Brasil, comparado con los demás países latinoamericanos que forman parte de este informe, lidera las tendencias con respecto a contagios, decesos y recuperados de covid19, seguido de México, que también posee un número elevado de nuevos contagios y decesos. Si bien ambos países lideran en cantidad de muertes y nuevos contagios, se puede observar una gran diferencia entre ambos. Brasil posee un vasto número de población, pero con una densidad menor a los demás países latinoamericano. En contrapartida, México posee la mitad de población que Brasil, pero su densidad supera ampliamente a los países analizados.
2. Según las gráficas se puede analizar que Brasil destaca de manera sobresaliente en la cantidad de dosis administradas, lo que podría reflejar una buena estrategia de vacunación masiva. Pero, debido a su mayor población, ésta buena estrategia de vacunación no alcanzó para cubrir a gran parte de su población.
3. En cuanto a las políticas de salud pública, Chile y Perú tienen menor cantidad de dosis administradas pero abarcan un mayor porcentaje de población. Esto es debido a la cantidad de habitantes de cada país, evidenciando un buen manejo de la pandemia con estrategias de vacunación efectivas.
4. Respecto a la comparación del número de casos nuevos a través de los meses, se puede observar que durante diciembre del 2021 y enero del 2022, se produjeron picos de contagios que marcaron un aumento significativo de nuevos casos en todos los países. Brasil y Argentina parecen tener un mayor número de casos en comparación con los demás países, con números más altos a lo largo del tiempo. Sin embargo, las muertes no siguen la misma tendencia.
5. Durante diciembre del 2021, los decesos no han marcado un aumento significativo. Esto podría entenderse por un correcto manejo del sistema de salud, por el buen desempeño de políticas de vacunación y por prácticas de autocuidado por parte de la población.
6. Continuando con las conclusiones, de acuerdo a la gráfica de mortalidad femenina y masculina, se puede observar que en todos los países la cantidad de muertes es mayor en los hombres que en las mujeres, por lo que se podría decir que los hombres son más vulnerables a la enfermedad, están menos protegidos o sus cuidados han sido más precarios comparados con los femeninos.
7. Con respecto a la comorbilidad, observo que existe una consistencia en los datos. Los países que registran un mayor número de muertes y nuevos contagios, también son los países que poseen índices de comorbilidad más altos. Se podría deducir que existe una relación entre las enfermedades cardiopulmonares e inmunosupresoras, desarrollando un alto factor de riesgo ante las complicaciones producidas por el covid-19.
8. Considero que las personas que poseen comorbilidades forman un grupo de riesgo, considerados prioridad para la vacunación.
9. Se puede decir que los países que más necesitan ayuda son Brasil, México y Argentina, ya que son los países que más casos confirmados y muertes han tenido, además de ser los países que menos vacunas han administrado respecto a su población.
10. En el análisis para la estrategia de expansión de los laboratorios Biogénesis se identifican las regiones de Brasil, México y Argentina como las ubicaciones idóneas para nuevos centros de vacunación y laboratorios.
11. En primer lugar se elijo Brasil por ser el país que lidera las tendencias negativas, posee un mayor número de población y menor densidad, indicando que los centros ya existentes se encuentran muy alejados o poco accesibles para la población.
12. Tanto México como Argentina requieren estrategias de vacunación más agresivas ya que también poseen un alto índice de casos confirmados de covid-19 y una mala estrategia de vacunación con menor cantidad de dosis administradas. Argentina posee la menor cantidad de dosis administradas, la peor estrategia de vacunación y se encuentra en segundo lugar en casos confirmados después de Brasil.
13. La población prioritaria para los nuevos centros de vacunación de Biogenesys, serían los hombres y las personas con comorbilidades, ya que poseen mayores factores de riesgo. Y luego se abarcara toda la población.

# Reflexión personal

La realización de este proyecto integrador fue muy enriquecedor, ya que no solamente pude poner en práctica los aprendido en el módulo 4 de la carrera de data analystisc sobre Python sino que me permitido ver como se relacionan todos los temas visto a lo largo de la carrera y lo importante que son cada uno de ellos en el proceso de la toma de decisiones basadas en datos.

Este proyecto al igual que el anterior fue muy satisfactorio con respecto a formar lazos de compañerismos más fuertes con mis compañeros, en particular con dos de ellos que nos apoyamos mucho, nos reunimos a la distancia y pudimos llevar adelante la realización del mismo, analizar y sacar conclusiones entre los tres.

Si bien en lo personal tuve un par de inconvenientes respecto a que la computadora con que cuento para estudiar, si bien funciona cuando trabajo con Power BI le cuesta más trabajo y se ralentizo bastante.

# 