Pestaña 1

****

**Expansión Estratégica de Biogenesys con Python**

**Alumno: Hillebrand, Francisco Javier**

**Email:** [**franhillebrand@gmail.com**](mailto:franhillebrand@gmail.com)

**Cohorte: Data Analytics Part-Time 09**

**Fecha de Entrega: 10/10/2025**

**Institución: Biogenesys**

ÍNDICE

[**Introducción 3**](#_h829pkpouq4g)

[**EDA e Insights 3**](#_4mjywtd4idxc)

[Propiedades Estadísticas del Dataset 3](#_kw2r7xlgo3n1)

[Matriz de Correlaciones 4](#_1lz3uigs57c9)

[Visualización de algunos Insights 6](#_qljwssoddljh)

[Promedio de casos nuevos por país 6](#_3j3ft3nna7kj)

[Comparación de casos confirmados, densidad poblacional y dosis administradas por país 6](#_w4vcyyuhyx45)

[Correlación entre nuevos casos confirmados, aumento de las muertes y aumento de la población a lo largo del tiempo 7](#_95z9giemrj94)

[Comparación de la estructura sanitaria del país 8](#_ww0hzy8h8kdr)

[Comparación del porcentaje de mortalidad femenina y masculina 9](#_k06e0abaiudp)

[Identificación de Tendencias y Patrones 9](#_elvtgbf8ythh)

[Evolución de dosis administradas por mes de cada país 9](#_qkjvdq4ik0ax)

[Casos nuevos vs. temperatura promedio 10](#_fh1l7nhimscj)

[Infraestructura médica vs. mortalidad 11](#_t5h5ujffuv78)

[Relación entre infraestructura médica y dosis administradas 12](#_aip5zktjzqp8)

[Relación entre vacunación, aumento y recuperación de afectados 12](#_9d46ht9oldk6)

[Correlación entre fallecimientos de COVID-19, fallecimientos por comorbilidad, el porcentaje población preponderante a la diabetes y el porcentaje de fumadores en la población 14](#_un4a5xo34s2x)

[Análisis Temporal 14](#_6bpvcnzevk5j)

[Tasa de cambio de casos confirmados en general 14](#_v5zmgphjdng2)

[Promedio Móvil de Casos Confirmados por País 15](#_hd0y0hibuw99)

[Tasa de cambio de fallecimientos en general 17](#_ybtfme8ep8op)

[Promedio Móvil de Muertes por País 18](#_z6d70g6gb3ci)

[Análisis Espacial 21](#_sgin8w4h2e3u)

[Total de muertes por País 21](#_5d5mn8gkbp1o)

[Total de Casos Confirmados por País 21](#_quw5fw937nu1)

[Otras Correlaciones 22](#_cmzani1o817z)

[Relación entre casos confirmados, fallecimientos y las población urbana y rural 22](#_f3bfmsx360kr)

[Relación entre la la acumulación de fallecimientos y casos confirmados en las áreas urbanas y rurales 23](#_vq0t652tvhk3)

[**Análisis del Dashboard 23**](#_1b1i90cit8s4)

[**Desarrollo del Proyecto 24**](#_xuiin8o8tv59)

[AVANCE N°1 24](#_uo5tpf5nlosz)

[Diccionario y Limpieza de Datos 24](#_j39cn3om6l6c)

[Importar librerías 26](#_dao24fhtemau)

[Leer Archivo 27](#_3c5389qe0ma9)

[Compruebo la cantidad de registros y columnas 27](#_6ha80m1pdhej)

[Filtro el Dataframe para obtener solo los registros de los países donde Biogenesys busca expandirse 27](#_erz17q3y6pe7)

[Filtra los datos en fechas mayores a 2021-01-01 27](#_2pvw9ylj8ajq)

[Trabajo con valores nulos y faltantes 27](#_j8g7y72ewpzi)

[Identificar variables claves 28](#_j5swp3pl395t)

[Guardar los datos filtrados en un archivo CSV 29](#_tqxpzgba9iwg)

[Calculo estadísticas descriptivas 29](#_pxmrk6ctsd5j)

[Función que obtiene la mediana, varianza y rango de un conjunto de valores 30](#_o1dppvz9ig5e)

[AVANCE N°2 31](#_db9vqprfml44)

[Importar librerías 31](#_kr56plpo1zvp)

[Análisis Estadístico 31](#_97m43rk0z4to)

[Visualización de Datos con Matplotlib y Seaborn 32](#_iqkq7buavxdq)

[Promedio de Casos Confirmados por país 32](#_l2x8y4in2ura)

[Comparación de casos confirmados por país y su densidad poblacional 33](#_wtdxkhnc95f8)

[Correlación entre nuevos casos confirmados, aumento de las muertes y aumento de la población a lo largo del tiempo 35](#_g6uwlctuqsln)

[Comparación de la estructura sanitaria del país 36](#_xwid3kd23pyy)

[Comparación del porcentaje de mortalidad femenina y masculina 38](#_f6oergumm1a2)

[Identificación de Tendencias y Patrones 38](#_s6b5ok1oxnvo)

[Evolución de dosis administradas por mes de cada país 39](#_fay6hdnn8tsq)

[Casos nuevos vs. temperatura promedio 40](#_pst2rehzn8ib)

[Infraestructura médica vs. mortalidad 40](#_tn3k7uhwmvco)

[Relación entre infraestructura médica y dosis administradas 41](#_wqjfhlkhrzq0)

[Relación entre vacunación, aumento y recuperación de afectados 42](#_86g6z6xajzo4)

[AVANCE N°3 43](#_ralcilmub1gt)

[Importar Librerías 43](#_ilphavccsb44)

[EDA 43](#_r14uwkp9d6vc)

[Análisis Temporal 43](#_bxfjbpd761dm)

[Tasa de cambio de casos confirmados en general 44](#_r3bmxxxiitic)

[Promedio Móvil de Casos Confirmados por País y Promedio Móvil de Muertes por País 44](#_kxp3zvtqlr0u)

[Análisis Espacial 46](#_m6gdc0k9mlsd)

[Total de muertes por País 46](#_upkwi8f49t5k)

[Total de Casos Confirmados por País 46](#_rpo70ghvhpb6)

[Otras Correlaciones 47](#_fd7vidbzoqli)

[Relación entre casos confirmados, fallecimientos y las población urbana y rural 47](#_hx4rhxt7uce0)

[Relación entre la la acumulación de fallecimientos y casos confirmados en las áreas urbanas y rurales 47](#_pwywg2djkaqn)

[AVANCE N°4 47](#_wc4ly0kavni2)

[Conexión de Python con Power BI 47](#_u87wltcozrtq)

[Predicción de nuevos casos 49](#_p61resb0fhnq)

[Predicción de fallecimientos 52](#_h18hzqcmszm0)

[Creación de Dashboards en Power BI 53](#_50mdz3sp50z1)

[Visualizaciones Estáticas vs Interactivas 53](#_puij5yec8hl4)

[**Conclusiones y Recomendaciones 54**](#_lboqbhy7dpib)

[**Reflexión Personal 56**](#_gu4xoly81oq5)

# Introducción

El presente informe detalla el proceso y los resultados del análisis estratégico enfocado en la incidencia de COVID-19 en los países hacia los que pretende expandirse Byogenesis. El propósito central fue transformar datos brutos en recomendaciones accionables para informar decisiones críticas relacionadas con la salud pública y, específicamente en lo que respecta a la ubicación estratégica de futuros laboratorios y centros de vacunación.

Para alcanzar este objetivo, el proyecto se estructuró de tal manera que se pudieron cumplir los siguientes logros organizacionales clave:

* Calidad y Acceso a Datos (ETL y Limpieza): Se realizaron operaciones eficientes de Extracción, Transformación y Carga (ETL), aplicando técnicas de limpieza de datos para estandarizar formatos, manejar valores faltantes y asegurar que los datos fueran de alta calidad y completamente confiables para análisis posteriores.
* Análisis Exploratorio y Descubrimiento de Insights: Se completó un análisis exploratorio de datos (EDA) exhaustivo. Mediante el uso de estadísticas, mediciones (como la correlación) y visualizaciones, se lograron identificar tendencias claras en la incidencia de COVID-19, revelando oportunidades estratégicas para la intervención.
* Desarrollo de Herramientas para la Decisión: Se desarrollaron visualizaciones eficientes y dashboards interactivos (como se demuestra en el anexo de gráficos comparativos y mapas de calor) que permiten a los tomadores de decisiones explorar los datos desde múltiples perspectivas, facilitando una toma de decisiones informada y ágil sobre la demanda de vacunas, la logística de distribución y la infraestructura sanitaria requerida en los países analizados.

En resumen, este análisis ofrece insights valiosos para una expansión estratégica y sostenible de las operaciones de Byogenesis.

# EDA e Insights

## Propiedades Estadísticas del Dataset

En el apartado [Análisis Estadístico](#_97m43rk0z4to) se hizo un análisis de medidas de tendencia central y dispersión de las variables del dataset, y podemos destacar algunos insights:

1. Si tenemos en cuenta el promedio y la mediana de la acumulación de casos confirmados y la aparición de nuevos casos por día, Argentina y Brasil son aquellos que más casos acumulados y en promedio mayor cantidad de casos nuevos por día registran:
   1. Analizando el caso de Argentina, puede que esto tenga que ver con que fue el país al que menos vacunas se le administró en promedio, por ende el contagio fue mayor.
   2. En el caso de Brasil, puede que sea por el hecho de que es, por diferencia, el país con mayor población. A más población, más contagios.
2. Hay que observar el caso de México, ya que siendo el país con mayor densidad poblacional (lo que puede facilitar el contagio del virus), está entre los primeros 3 países que, en promedio, más casos confirmados y fallecimientos tuvieron. También por eso es el país que más vacunas administró, en promedio, detrás de Brasil.
3. Teniendo en cuenta la cantidad de muertes por día, y el acumulado de muertes a través del tiempo, Brasil es el país con los números más altos en estas variables. Teniendo esto en cuenta, hay que recordar que es el país que más contagios diarios registró.
4. Anteriormente mencionamos a Brasil y Argentina, ambos con muchos casos nuevos por día, y en el caso de Brasil, es el país que más fallecidos tuvo. Sin embargo, son dos de los primeros cuatro países con un mayor PBI per cápita, por lo que podrían costear más vacunas.
5. También es importante destacar que Brasil y Chile son los dos países que más enfermeros tienen cada 1.000 habitantes, un dato importante para la administración de vacunas.
6. Si bien Argentina fue uno de los países más afectados como se dijo en el primer insight, hay que tener en cuenta que es el país con la mayor cantidad promedio de médicos cada 1.000 habitantes, un dato importante para la administración de vacunas.

Si desea ver las medidas de cada variable por país, puede hacerlo en [Stats por país](https://docs.google.com/spreadsheets/u/0/d/1gZ1LtpInaxwzx_Va04LP_RJiQqScaXN5UI2LILenwfY/edit).

## Matriz de Correlaciones

A partir de la Matriz de Correlaciones generada en el apartado [Análisis Estadístico](#_97m43rk0z4to), puedo resaltar algunos insights:

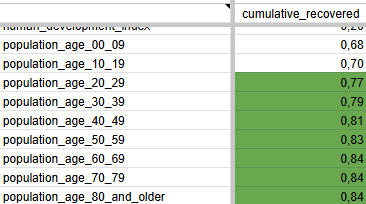
1. Existe una fuerte correlación positiva entre los casos positivos acumulados de COVID-19 y las muertes acumuladas por COVID-19, lo que indica que en las regiones o períodos con mayor número de contagios también se registran más muertes.

*Esto se percibe ya que las variables cumulative\_confirmed y cumulative\_deceased tienen una correlación positiva de 0,90.*

1. Los casos positivos acumulados y las muertes acumuladas aumentan en todas las edades a lo largo del tiempo, pero un poco más en personas de 60 años o más. Esto se percibe en la siguiente imagen.



También se puede destacar que el número de recuperados aumenta especialmente en la población de personas mayores a 40 años.



1. Existe una fuerte correlación positiva entre los casos positivos acumulados y los recuperados acumulados, lo cual refleja que en las regiones o períodos con mayor número de contagios también se registran más personas que lograron recuperarse.

*Esto se percibe ya que las variables cumulative\_confirmed y cumulative\_recovered tienen una correlación positiva* *de 0,95.*

1. Existe una fuerte correlación positiva entre las muertes acumuladas y los recuperados acumulados, lo cual refleja que en las regiones o períodos con mayor cantidad de muertes también se registran más personas que lograron recuperarse.

*Esto se percibe ya que las variables cumulative\_deceased y cumulative\_recovered tienen una correlación positiva* *de 0,90.*

1. A medida que aumenta la población de ambos géneros, aumenta el acumulado de casos positivos para ambos por igual.

*Esto se percibe ya que la correlación positiva entre cumulative\_confirmed y las variables population\_male y population\_female es de 0,78 para ambos.*

1. A medida que aumenta la población de ambos géneros, aumenta el acumulado de muertes para ambos por igual.

*Esto se percibe ya que la correlación positiva de cumulative\_deceased y las variables population\_male y population\_female es de 0,82 para ambos.*

1. A medida que aumenta la población de ambos géneros, aumenta el acumulado de recuperados para ambos por igual. Viendo los dos insights anteriores, vemos que el número de recuperados crece menos o igual con respecto a los casos confirmados y las muertes, siempre y cuando comparemos la correlación positiva de estas variables con la población de ambos géneros.

*Esto se percibe ya que la correlación positiva de cumulative\_recovered y las variables population\_male y population\_female es de 0,78 para ambos.*

1. Se puede destacar que la administración de vacunas aumenta al mismo tiempo que aumenta el acumulado de recuperados.

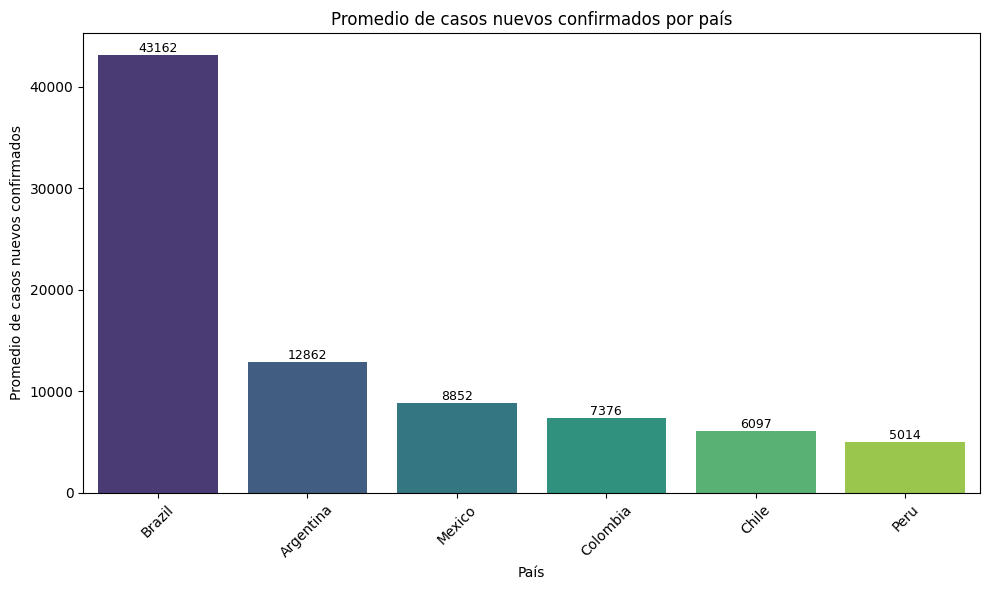
*Esto se percibe ya que la correlación positiva entre cumulative\_vaccine\_doses\_administered y cumulative\_recovered es de 0,88.*

Si desea ver las correlaciones entre todas las variables puede acceder al documento [correlaciones\_byogenesis](https://docs.google.com/spreadsheets/u/0/d/1Rqw4mkNgMMQiOCB9D1ubbil3v_dtuMqzVcFeYbivoDc/edit)

## Visualización de algunos Insights

### Promedio de casos nuevos por país

En este caso vemos lo anteriormente mencionado, Brasil y Argentina son los países que en promedio han tenido la mayor cantidad de nuevos casos por día.

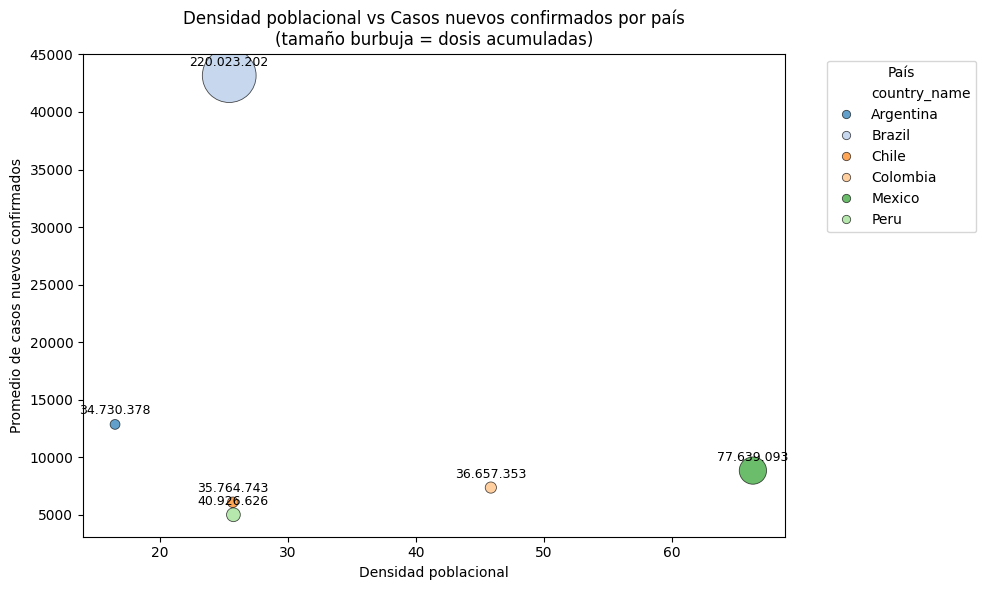


### Comparación de casos confirmados, densidad poblacional y dosis administradas por país

Vemos que los países que tienen mayor densidad poblacional no son aquellos que más casos confirmados por día tienen, la densidad poblacional no parece ser determinante para el contagio del virus.

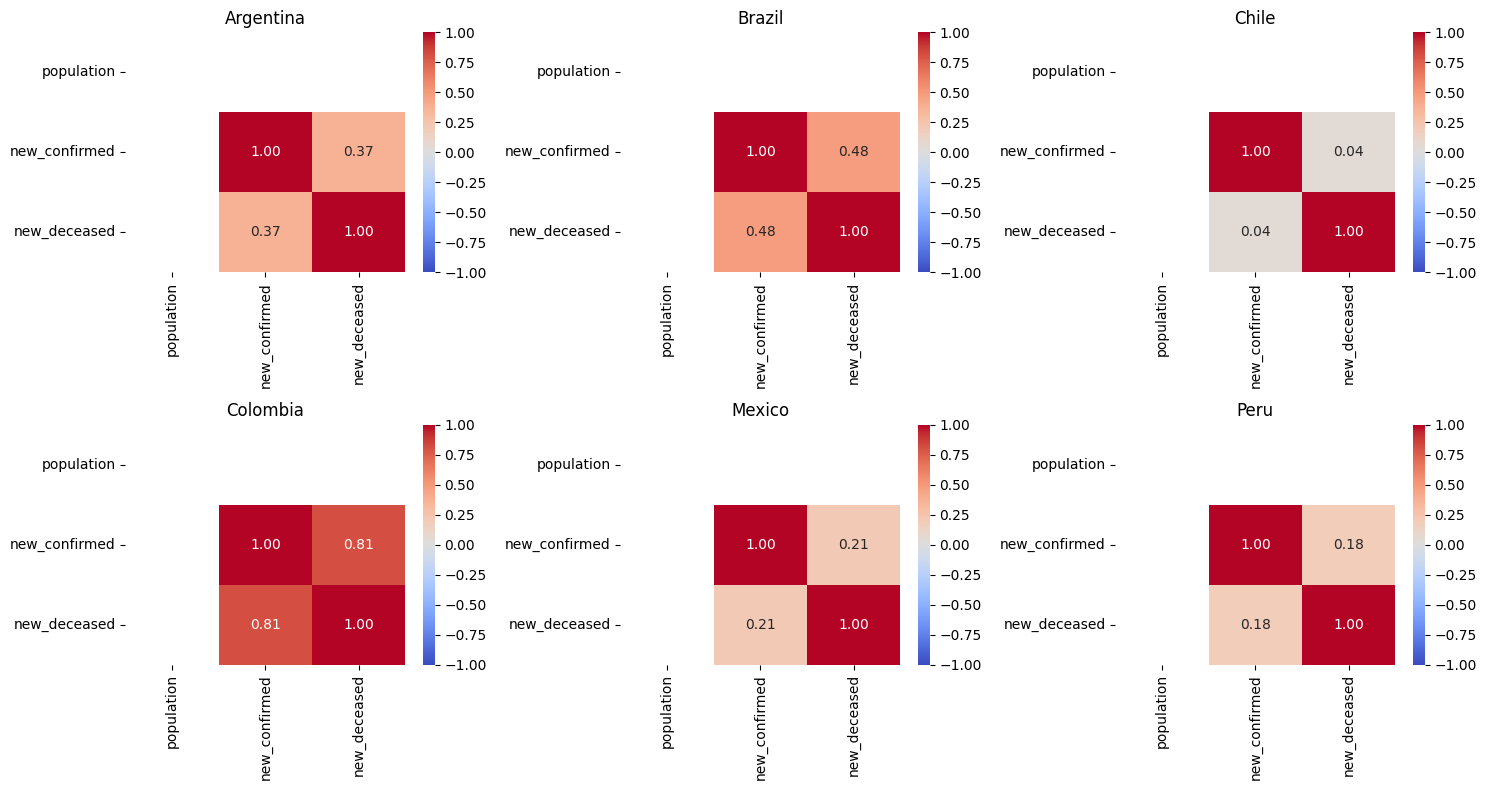
Si tenemos en cuenta también la administración de vacunas, vemos en el siguiente gráfico, que México fue el segundo país que más vacunas administró por lo que no aumentaron los casos.

Sigue siendo interesante el caso de Brasil, donde se administran más vacunas, y aún así han tenido el promedio más alto de casos confirmados por día, por lo que en este país puede haber una oportunidad para ofrecer vacunas. Otro punto a destacar, es que se le debe dar prioridad a la Argentina, es el país al que menos vacunas se le administraron, y el número promedio de casos confirmados por día es de los más altos.



### Correlación entre nuevos casos confirmados, aumento de las muertes y aumento de la población a lo largo del tiempo

En este caso hice un heatmap para evaluar la correlación entre las variables new\_confirmed, new\_deceased y population por país



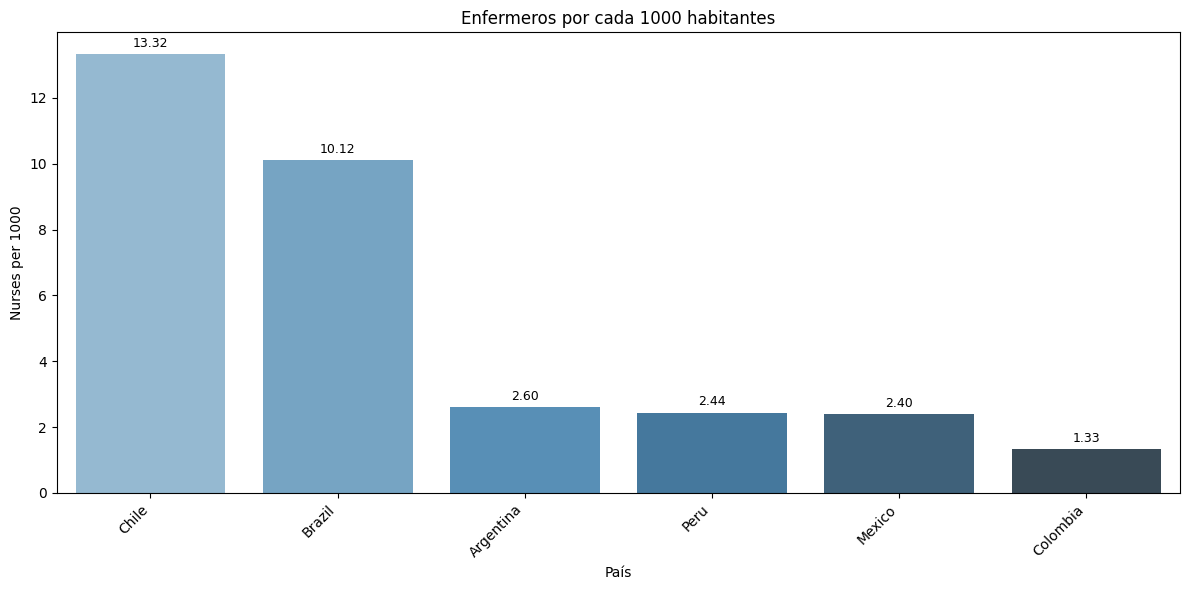
Lo más destacable aquí es que en Colombia hay una fuerte correlación positiva entre los casos confirmados y el número de muertes por día, por ende, estas dos variables aumentan a la vez, una correlación que no es tan fuerte en los demás países.

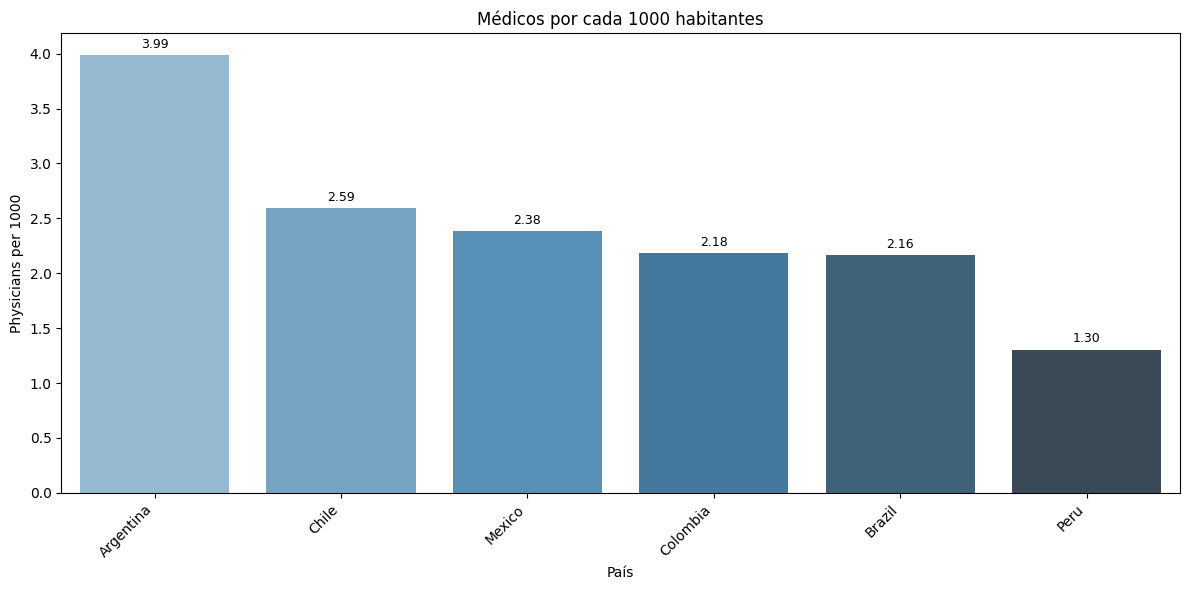


### Comparación de la estructura sanitaria del país

Hay que destacar la situación de Chile, es el primer país con mayor cantidad promedio de enfermeros por 1000 habitantes, y el segundo en cantidad promedio de médicos por 1000 habitantes, quizás por ello tuvo menor cantidad de casos confirmados por día. Además es interesante tener esto en cuenta para la administración de vacunas.

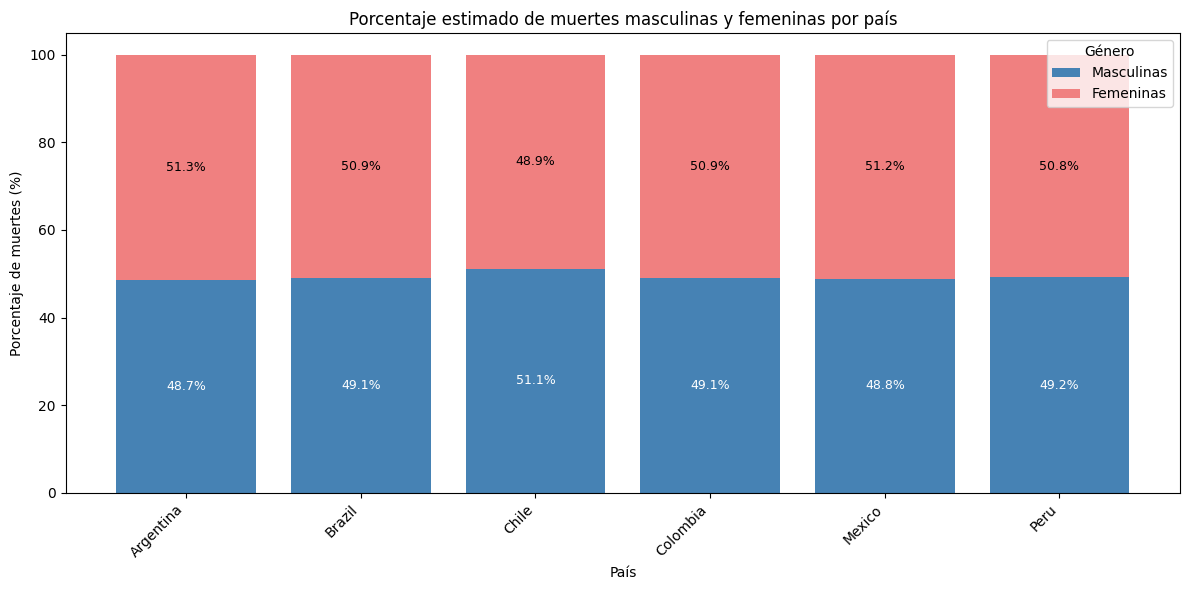
Brasil fue el país que más dosis administró y puede que tenga que ver con qué tiene uno de los mejores números promedio de médicos por 1000 habitantes. Y además, es el segundo país en cuanto a la cantidad de enfermeros por 1000 habitantes.





### Comparación del porcentaje de mortalidad femenina y masculina

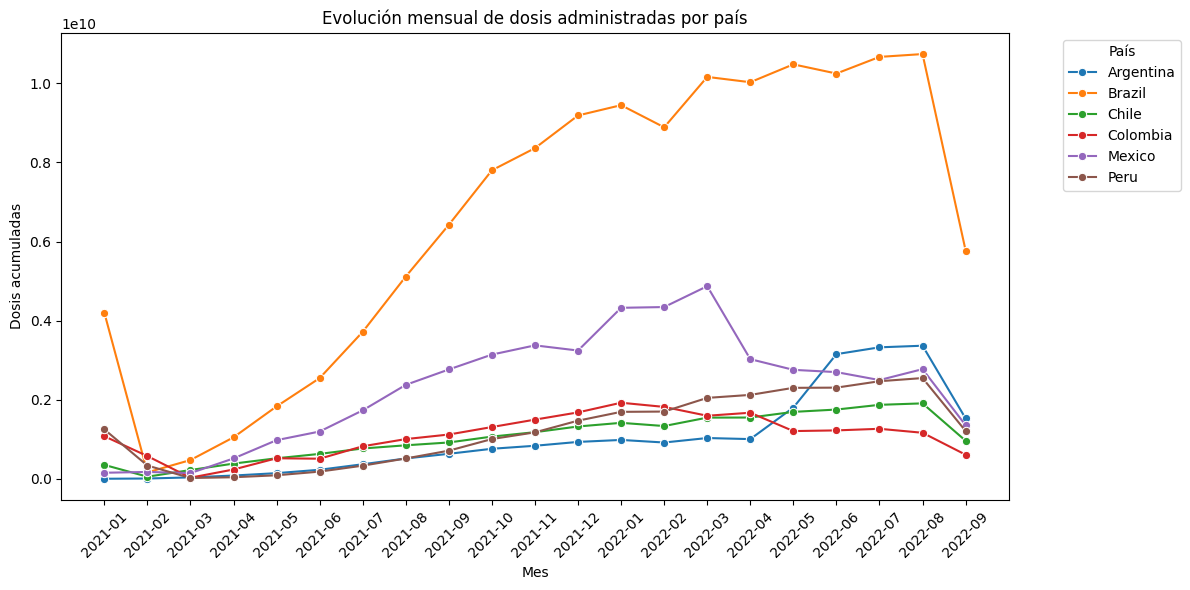
Por lo que vemos, la enfermedad afectó casi por igual a ambos géneros, entonces las muertes por género no creo que sea una variable clave para tomar la decisión hacia donde expandirse.



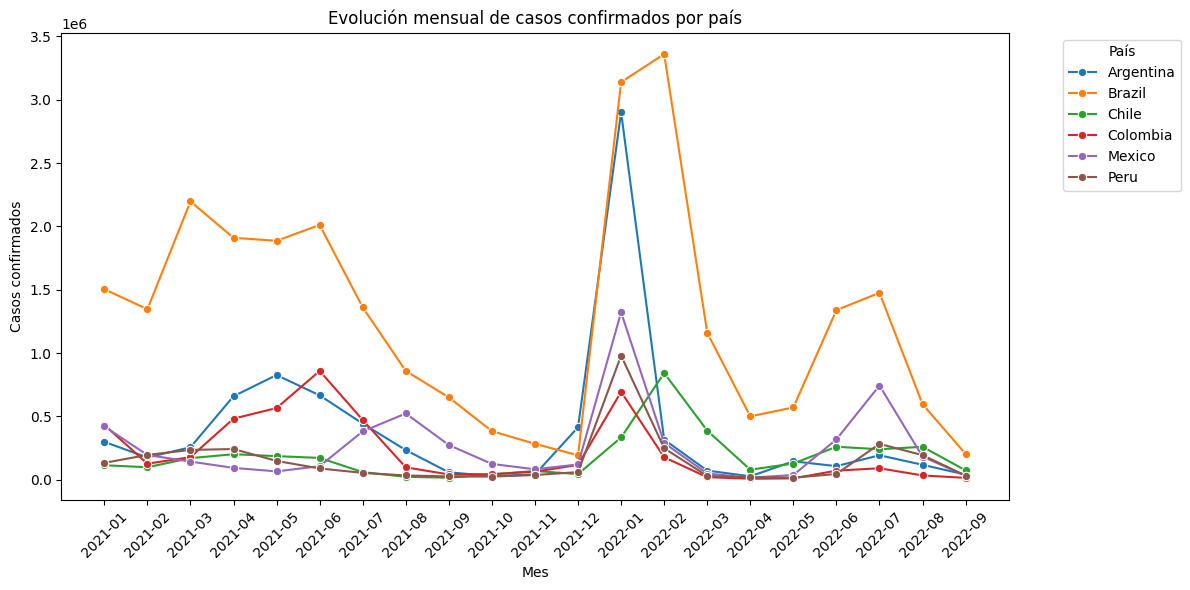
## Identificación de Tendencias y Patrones

### Evolución de dosis administradas por mes de cada país

En este gráfico destaco el aumento de dosis administradas en diciembre de 2021 en Argentina, mismo mes donde vemos que hubo una suba en la cantidad de casos (segundo gráfico).

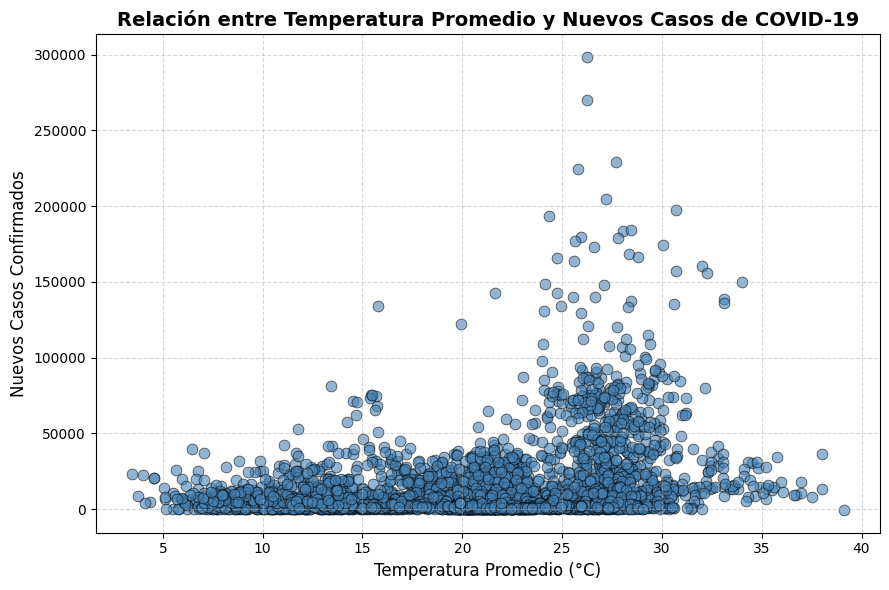


En abril de 2021, todos los países comenzaron a administrar más vacunas, mes en el que se fue estabilizando la cantidad de contagios, al menos hasta diciembre de ese año. Ya en diciembre, subieron los casos y la administración de vacunas también.



### Casos nuevos vs. temperatura promedio

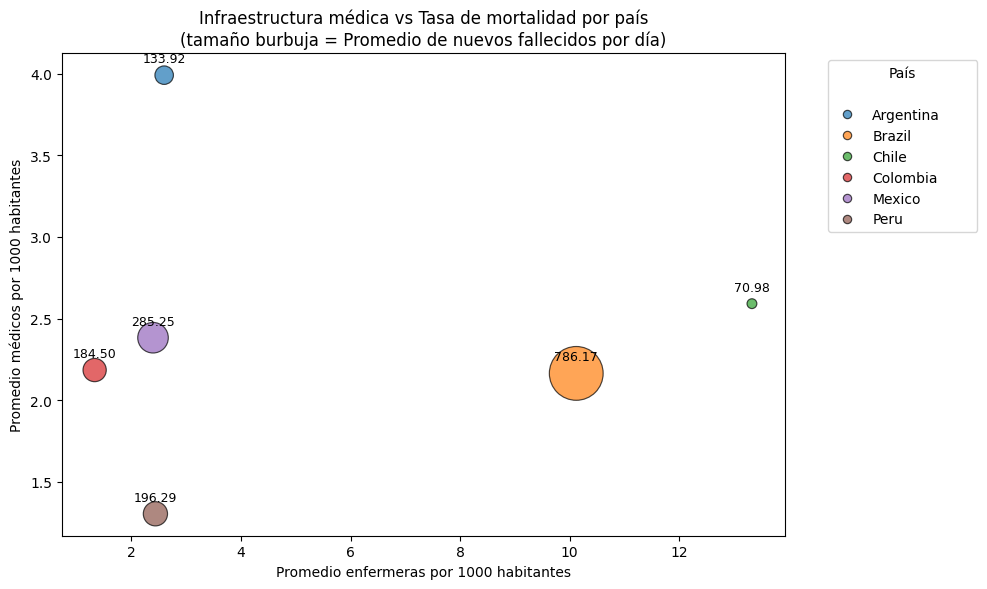
Rápidamente podemos observar que en las temperaturas más cálidas, el número de nuevos casos confirmados aumenta bastante, especialmente entre los 20° y los 35°. Puede que haya una relación entre el aumento de la temperatura, y el aumento de los casos.



### Infraestructura médica vs. mortalidad

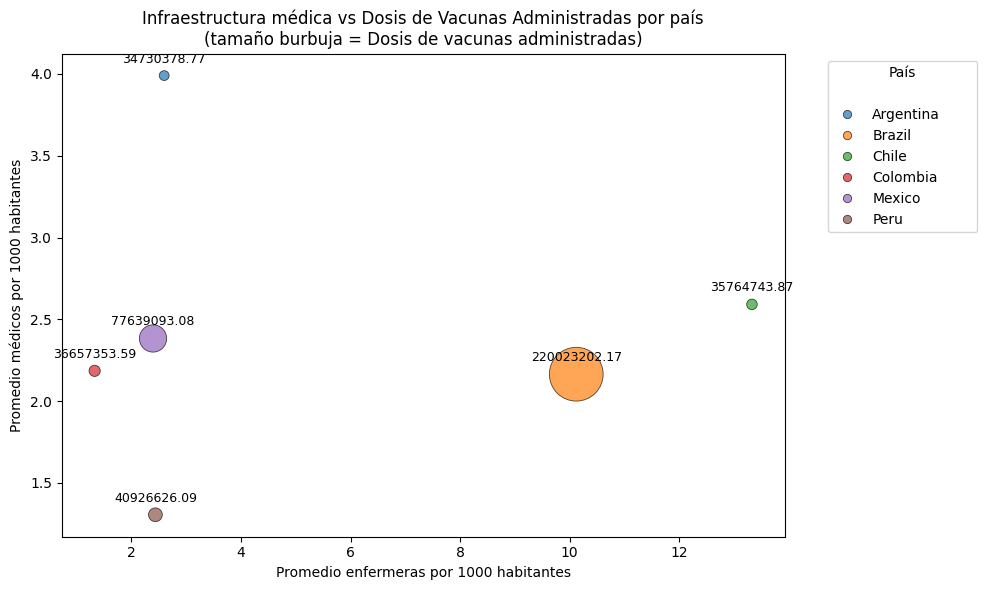
Este gráfico vuelve a destacar la infraestructura médica chilena, que es la más preparada y por ello considero que es el país con menor cantidad promedio de muertes por día.

Viendo este gráfico puede que Brasil no puede con tantos casos nuevos por día como vimos antes, y quizás por ello es el país con el mayor promedio de muertes al día.



### Relación entre infraestructura médica y dosis administradas

Con este gráfico, podemos observar que no hay una clara correlación entre la cantidad de médicos y enfermeros, y la cantidad de dosis administradas. Por ejemplo, en este caso puede que una mejor infraestructura médica, puede lograr una mayor administración de vacunas (es el caso de Brasil). Pero no es siempre así, ya que médicos y enfermeros no solo se dedican a administrar vacunas, sino también a atender a las personas afectadas. Por eso vemos que Argentina pese a tener el mejor promedio de médicos cada 1000 habitantes, es el país que menos vacunas administró.

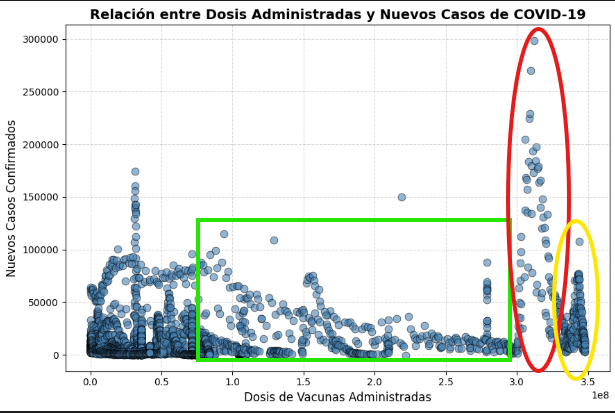


También hay que tener en cuenta que el dataset no tiene información sobre la cantidad de vacunas que compró cada país, claramente un país que compra más vacunas, puede administrar aún más. La falta de este dato (que considero importante), demuestra por qué no hay un claro patrón entre la cantidad de médicos y enfermeros, y las vacunas administradas.

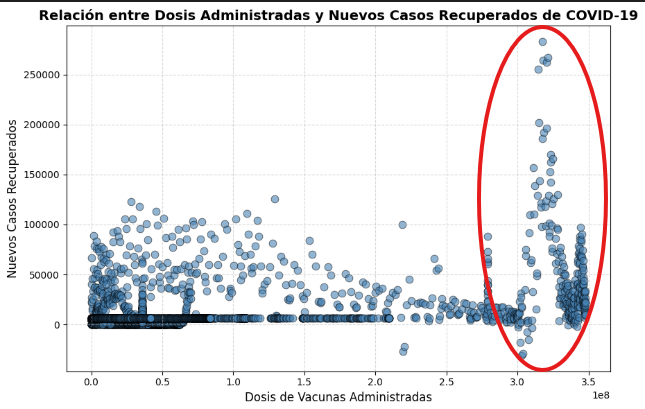
### Relación entre vacunación, aumento y recuperación de afectados

Como era de esperarse, a medida que suben los casos, suben las dosis administradas. Igualmente, vemos en este gráfico que hay una disminución en los nuevos casos a medida que sube la administración de vacuna (rectángulo verde).

También vemos que luego de un pico de nuevos casos confirmados (círculo rojo), aumentó aún más la administración de vacunas (círculo amarillo).

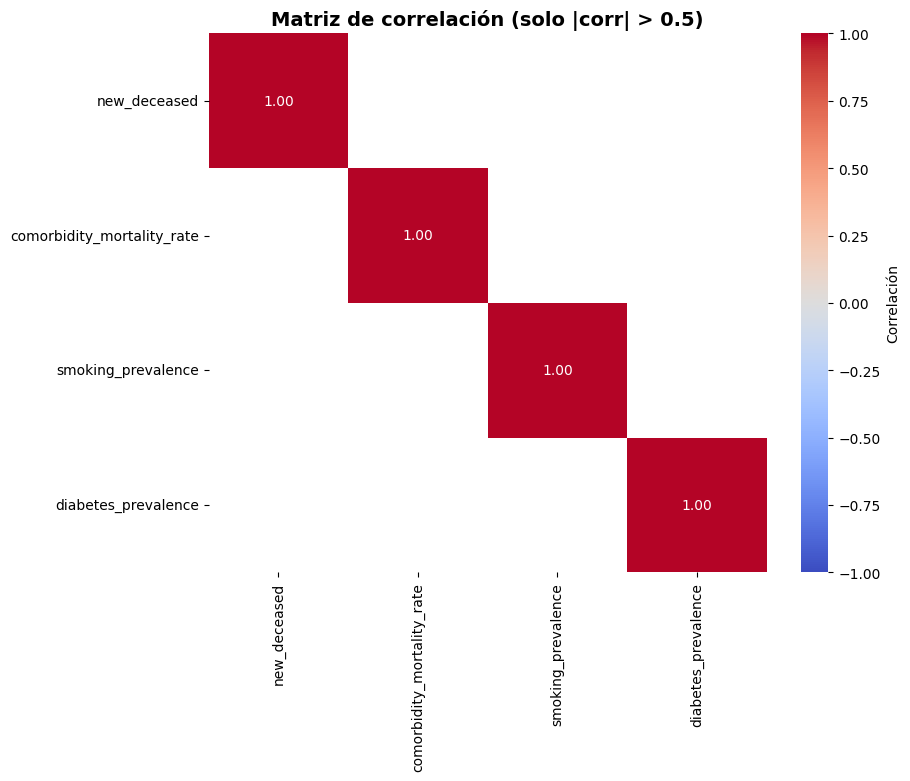


En este segundo gráfico vemos como aumentan la cantidad de recuperados a medida que aumentan las dosis de vacuna administradas.



### Correlación entre fallecimientos de COVID-19, fallecimientos por comorbilidad, el porcentaje población preponderante a la diabetes y el porcentaje de fumadores en la población

En este caso vemos que no hay una fuerte correlación (ni positiva ni negativa), entre los fallecidos de COVID-19, los fallecidos por comorbilidad, los fumadores de la población y la población que tiene preponderancia a la diabetes.



## Análisis Temporal

### Tasa de cambio de casos confirmados en general

Haciendo un análisis de la serie temporal de la variable new\_confirmed encontré clara tendencia a la baja en la incidencia de casos a lo largo del tiempo. La serie inició con una fase de alta volatilidad, marcada por un pico de, aproximadamente, 300.000 casos nuevos, que representó el máximo histórico. Posteriormente, la serie experimentó un declive sostenido y una reducción en la volatilidad (según el gráfico de diferencia simple, cuyas líneas son de color rojo), lo que indica que la magnitud de los cambios diarios se redujo drásticamente. Hacia la etapa final del año 2022, la incidencia de 'new\_confirmed' alcanzó un estado de estabilidad, sugiriendo que la situación se controló y se mantuvo en niveles menores.

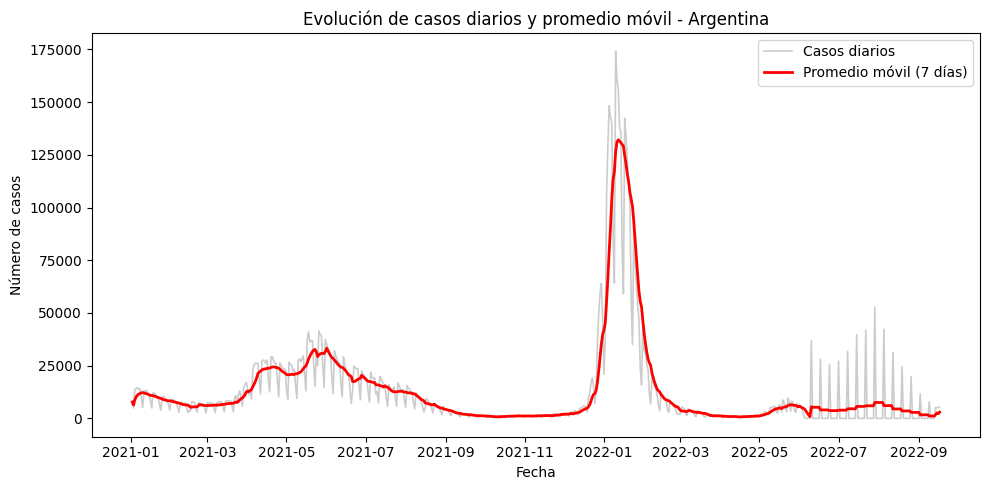


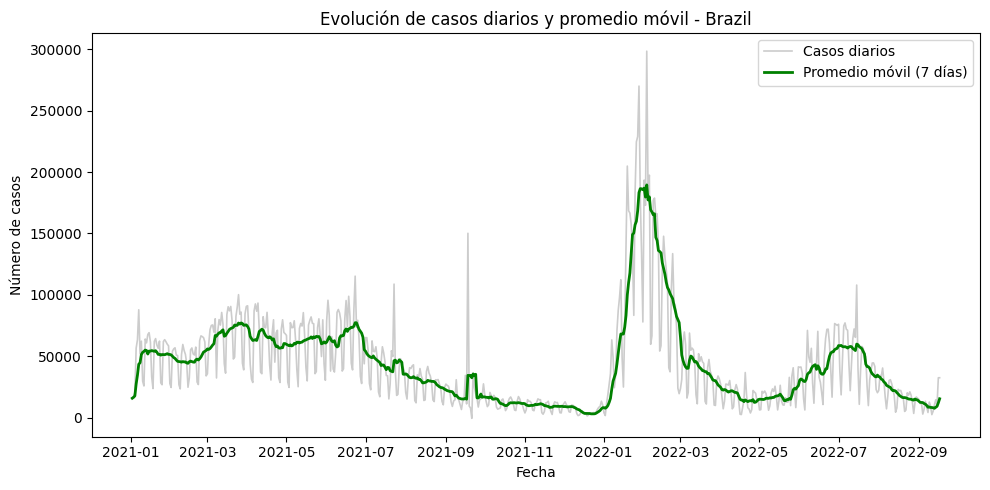
### Promedio Móvil de Casos Confirmados por País

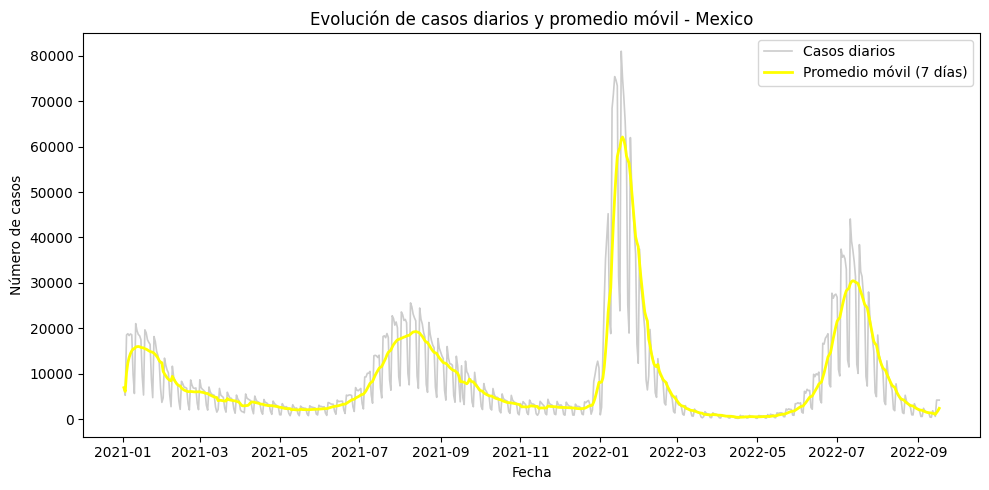
En este caso haremos un análisis más detallado, y observaremos cómo ha avanzado la enfermedad a través del tiempo en cada país.

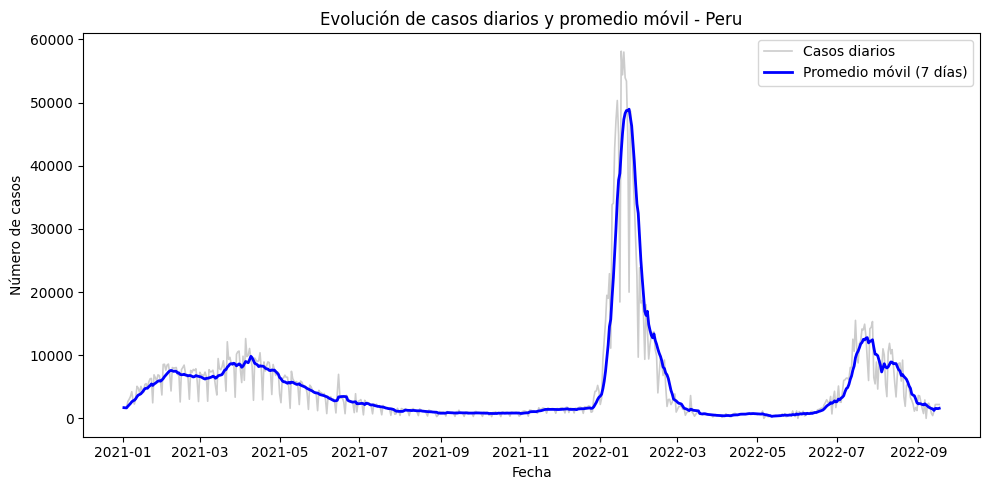
Considero que hay un comportamiento parecido en todos los países, en enero suele haber un aumento en el número de casos, estos meses suelen ser los más cálidos en todos los países menos México, esto demuestra la correlación anteriormente expuesto, donde vemos que suelen haber más casos en temperaturas más cálidas. También veo un aumento, no tan pronunciado, a partir de junio (mes cálido en México), el cual suele aumentar hasta julio aproximadamente.

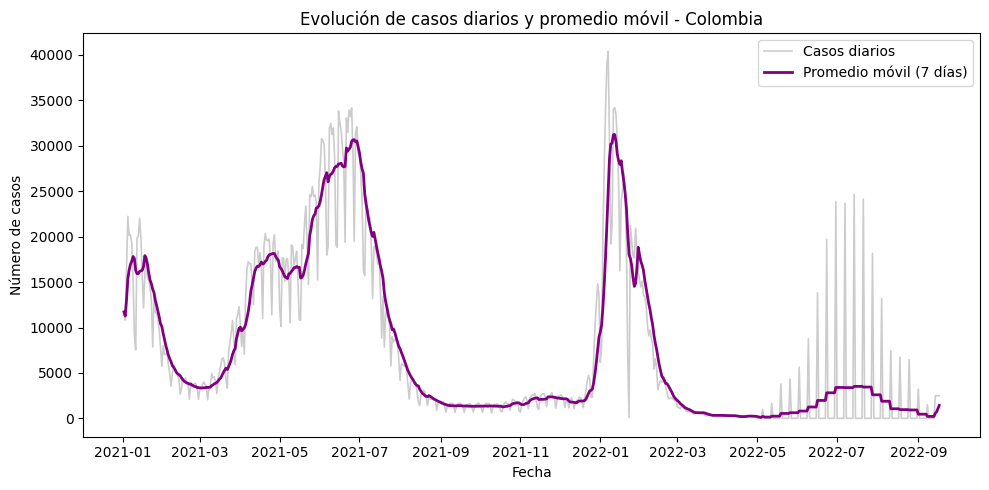
Además hay que hacer hincapié en lo siguiente: A partir de agosto y hasta diciembre, suelen disminuir los casos.

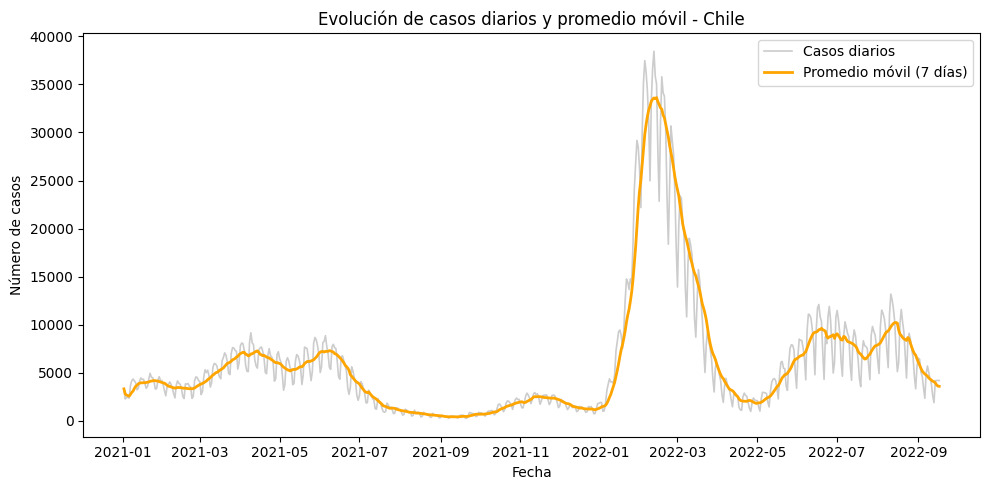






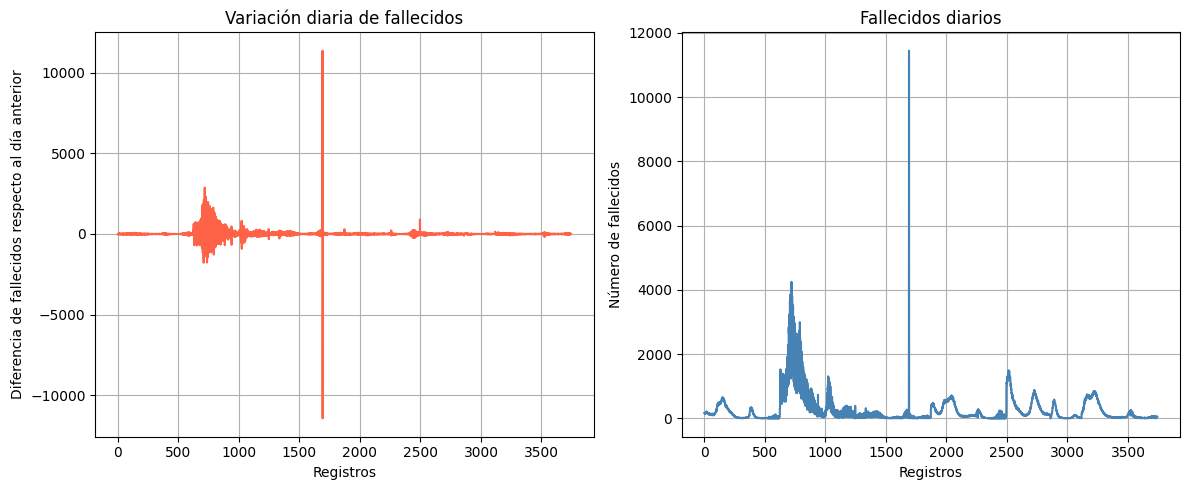






### Tasa de cambio de fallecimientos en general

Hice un análisis parecido al análisis de los casos confirmados, pero esta vez con la cantidad de fallecidos diarios y su variación. Aquí pude observar una dinámica similar de brote seguida de una estabilización. El gráfico de "Fallecidos diarios" (derecha) muestra una serie con valores principalmente bajos, pero con varios picos (el mayor superando los 11.000 fallecidos), que representan los momentos de mayor impacto de la pandemia. Por su parte, el gráfico de "Variación diaria de fallecidos" (izquierda) exhibe grandes fluctuaciones, con variaciones positivas y negativas que alcanzan ±10.000, lo cual es atípico e indica una alta volatilidad. A pesar de estos picos y la volatilidad, la serie muestra una clara estabilización a lo largo del tiempo, donde la incidencia y la variación diarias se reducen significativamente, manteniéndose en niveles entre 1000 y 2000 durante la mayor parte del período final.

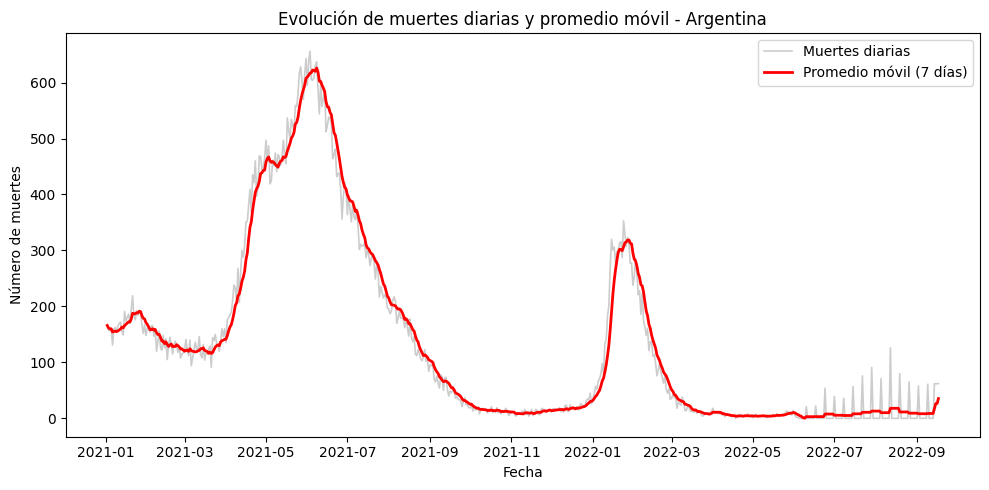


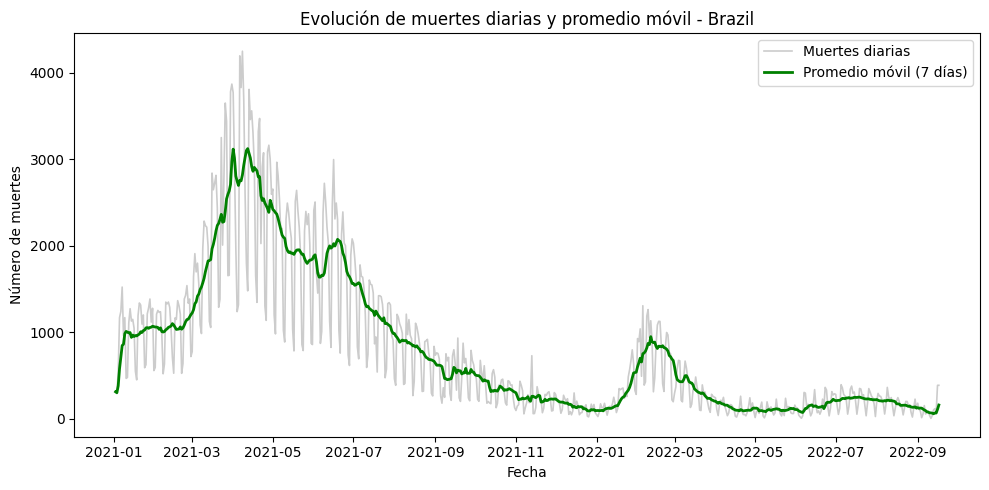
### Promedio Móvil de Muertes por País

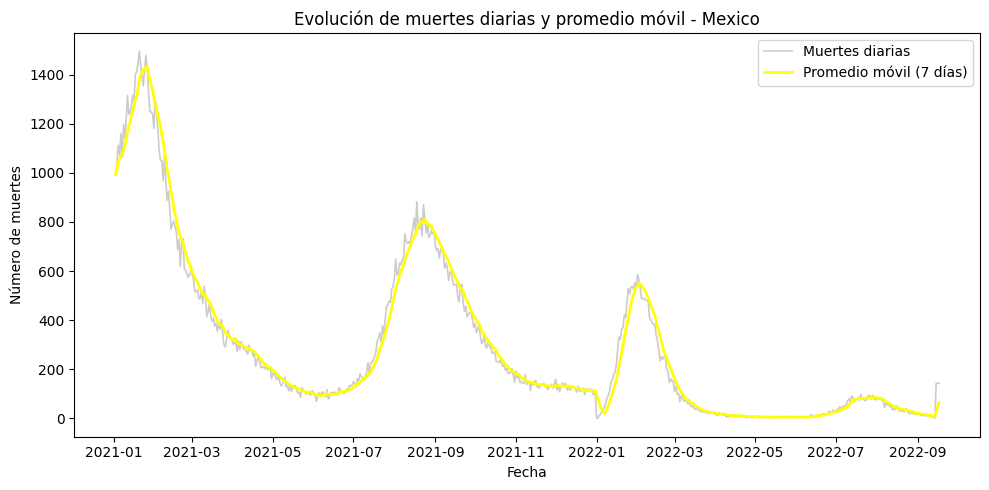
En general, podemos decir que las muertes disminuyeron en 2022, sin contar Chile que tuvo un fuerte pico en abril de ese año. El resto de los países tuvo el pico de fallecimientos en 2021, en los primeros siete meses del año.

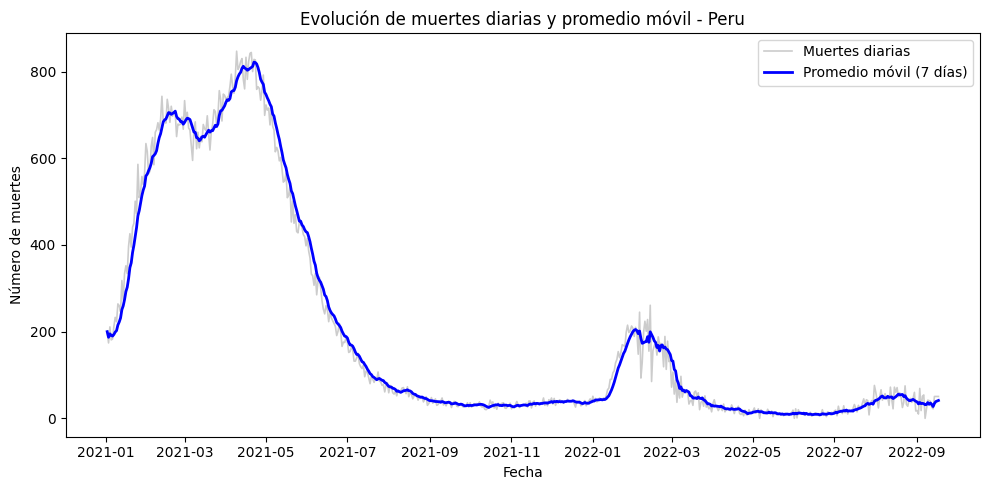
En México feo un patrón de aumentos de fallecidos en enero y julio de ambos años. En febrero y septiembre disminuyen las muertes.

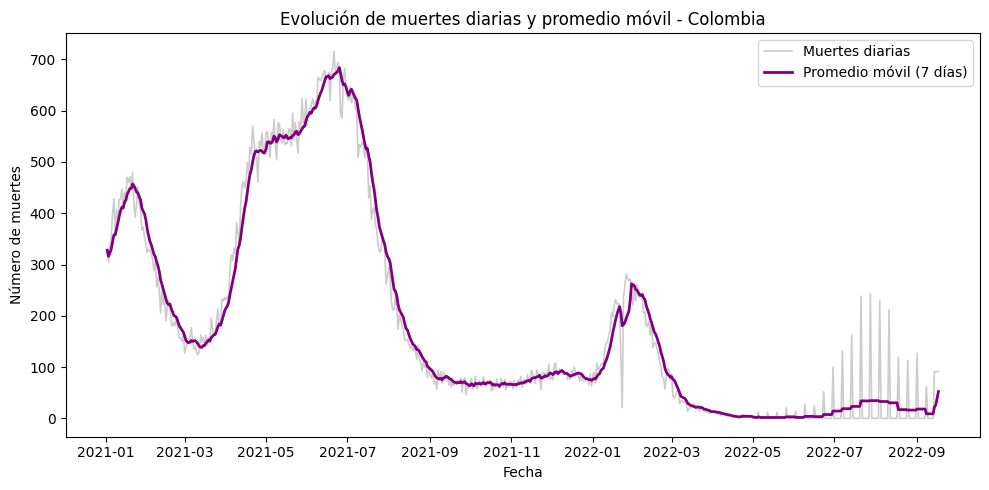
En Perú, Colombia y Brasil, en enero de ambos años, hubo un aumento de fallecimientos.

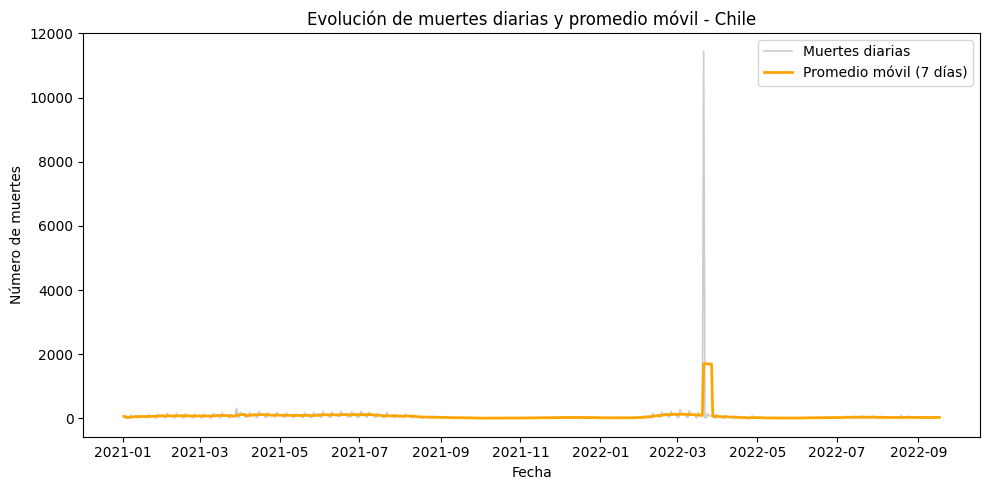










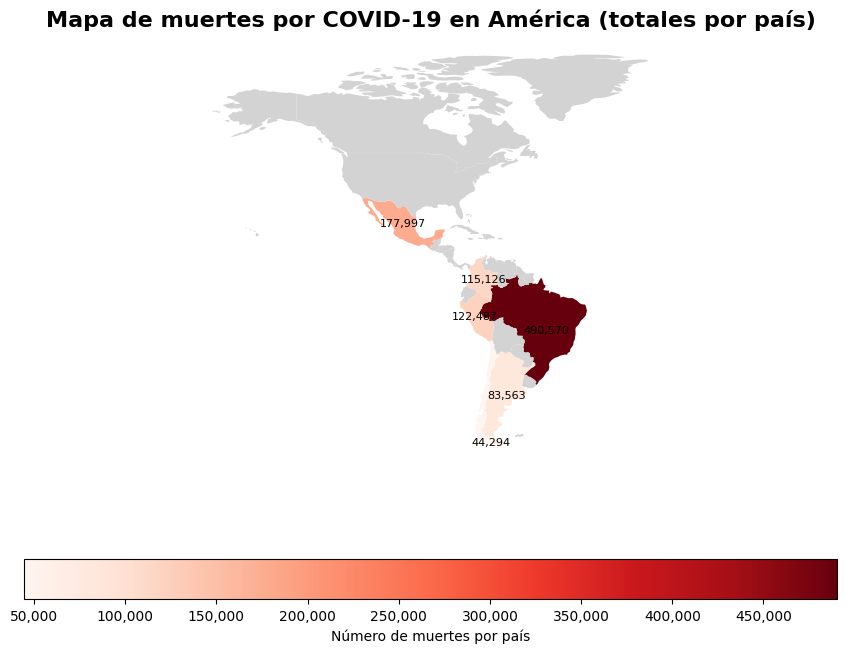


## Análisis Espacial

### Total de muertes por País

En este caso hice un análisis por País, y en el mapa vemos claramente que Brasil fue el país al que más le afectó la enfermedad, entre este país y el segundo país con más muertes (México) hay más de 300 mil casos de diferencia.

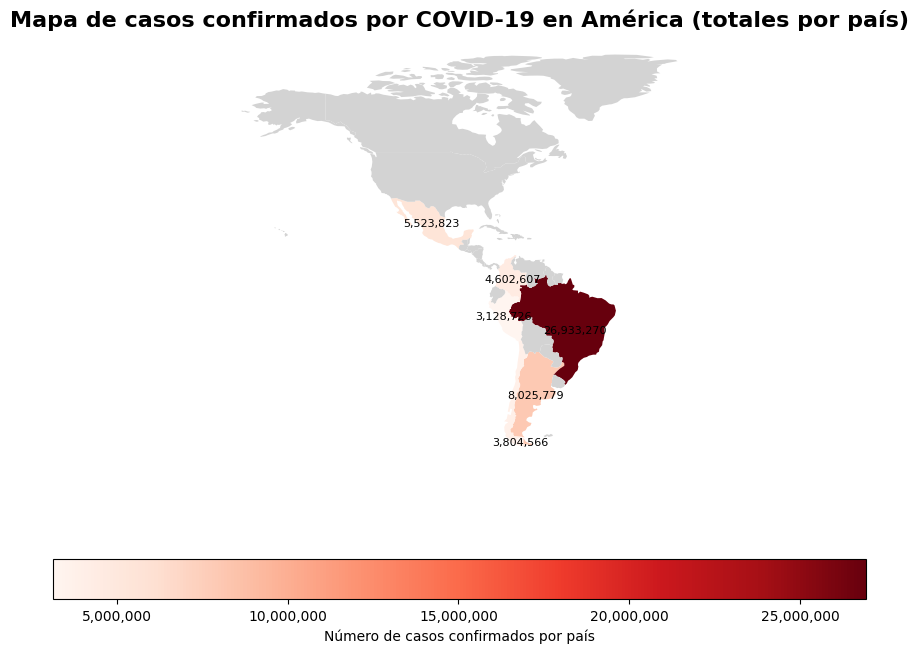
El país menos afectado fue Chile con 44.294 muertes, el país con la mejor infraestructura médica de todos.



### Total de Casos Confirmados por País

Nuevamente Brasil es el país más afectado con más de 26 millones de casos confirmados, hay una diferencia de más de 18 millones de casos en comparación al segundo país con más casos confirmados (Argentina).

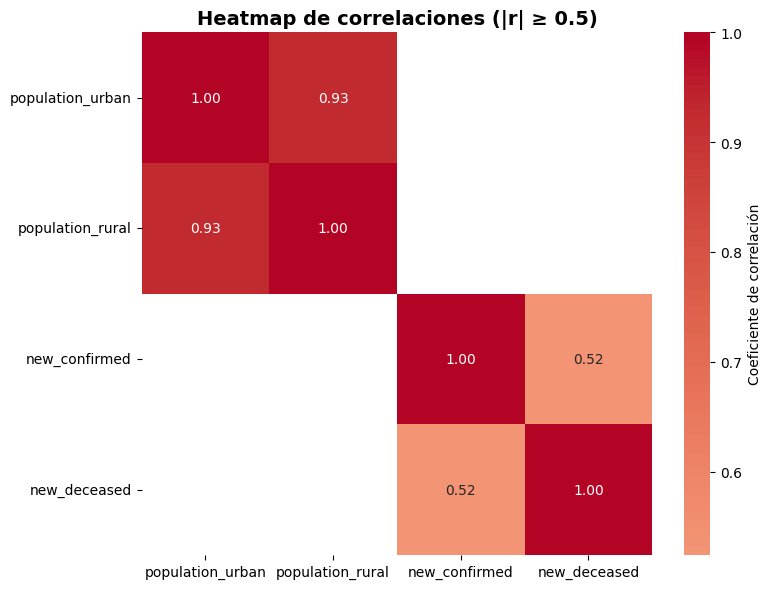
Observé un dato interesante en Argentina, es el segundo país con menos muertes, pero el segundo que más casos confirmados tuvo.



## Otras Correlaciones

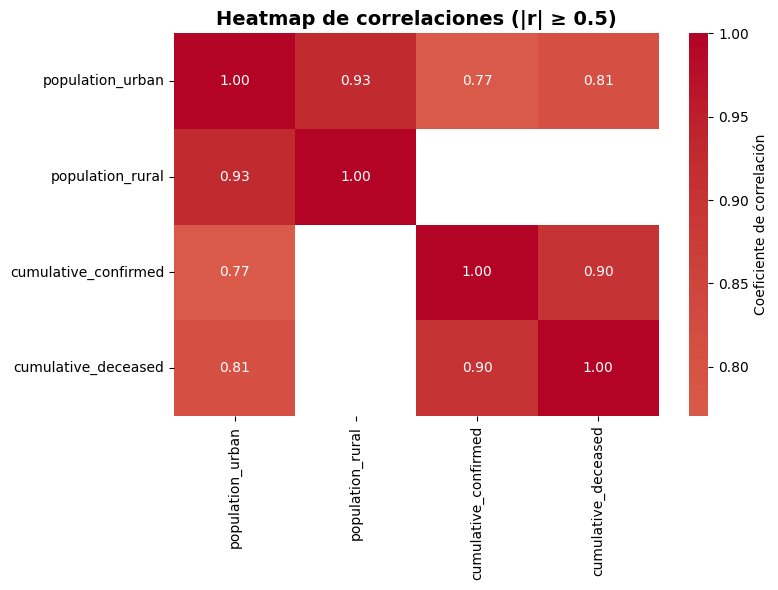
### Relación entre casos confirmados, fallecimientos y las población urbana y rural

Quise investigar si hay una correlación entre las poblaciones urbanas y rurales, y los casos confirmados y los fallecimientos, pero luego de graficar el heatmap pude observar que no, o que al menos no hay una fuerte correlación positiva ni negativa.



### Relación entre la la acumulación de fallecimientos y casos confirmados en las áreas urbanas y rurales

Se puede observar que en las poblaciones urbanas más grandes, hay una mayor cantidad de casos confirmados, y también una mayor cantidad de fallecidos, en comparación a las poblaciones rurales.



# Análisis del Dashboard

El dashboard del proyecto lo encontrará en [PI.pbix](https://drive.google.com/open?id=1ne1Lfio4HAXwYxSesOTjCmJZN4f_Ldxn)

El dashboard se divide en 3 tableros:

* Análisis General: En este tablero verá (leyendo de izquierda a derecha en cada fila, comenzando desde la esquina superior izquierda) un análisis de algunas variables determinantes para entender el impacto del COVID-19 en cada país. Algunas de estas variables son: el número de casos confirmados, el número de fallecidos, la cantidad de dosis administradas, la infraestructura sanitaria, etc.
* Análisis Poblacional: En este tablero verá el impacto de la enfermedad pero teniendo en cuenta el número total de habitantes en cada país y el PBI per cápita de cada uno.
* Análisis Predictivo: En este tablero hice una estimación de posibles casos confirmados y posibles fallecimientos en el futuro, teniendo en cuenta los datos del dataset. Esta información será útil para decidir hacia qué país debe expandirse Byogenesis.

En los 3 tableros puede seguir el patrón Z para observar el contenido.

# Desarrollo del Proyecto

## AVANCE N°1

### Diccionario y Limpieza de Datos

En este Proyecto, el Diccionario de Datos nos fue proporcionado en la consigna, pero si tenemos en cuenta que este informe se va a presentar al cliente, me parece importante que el diccionario esté disponible para él.

Sin embargo, hice algunos cambios en el tipo de datos de las columnas que se verán reflejados en el diccionario:

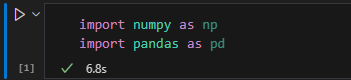
* La columna *date* es de tipo datetime.
* Cambio los tipos de datos de las columnas *country\_code, location\_key* y *country\_name* a string.
* Las columnas que almacenan los datos acerca de la población serán enteros, ya que no son necesarios los decimales. Menos la columna *population\_density* que se mantendrá como float.
* Las columnas *new\_confirmed, new\_deceased, cumulative\_confirmed, cumulative\_deceased, cumulative\_vaccine\_doses\_administered, new\_recovered y cumulative\_recovered*, también serán de tipo entero.

El código utilizado para hacer estos cambios lo puede ver en [PIDA\_M4\_Francisco\_Hillebrand.ipynb](https://drive.google.com/open?id=1WHrB51UynhIkW96AmxLUrjQMSHfFU9A8).

| **Name** | **Type** | **Description** | **Example** |
| --- | --- | --- | --- |
| date | datetime | ISO 8601 date (YYYY-MM-DD) of the datapoint. Only has dates greater than 2021-01-01 | 3/30/2020 |
| population | integer | Total count of humans | 51606633 |
| population\_male | integer | Total count of males | 25846211 |
| population\_female | integer | Total count of females | 25760422 |
| rural\_population | integer | Population in a rural area | 9568386 |
| urban\_population | integer | Population in an urban area | 42038247 |
| population\_density | float | Population per squared kilometer of land area | 529.3585 |
| human\_development\_index | float | Composite index of life expectancy, education, and per capita income indicators | 0.903 |
| population\_age\_00\_To\_79 | integer | Estimated population between the ages of ${lower} and ${upper}, both inclusive | 42038247 |
| population\_age\_80\_and\_older | integer | Estimated population over the age of 80 | 477081 |
| gdp | float [USD] | Gross domestic product; monetary value of all finished goods and services | 24450604878 |
| gdp\_per\_capita | float [USD] | Gross domestic product divided by total population | 1148 |
| new\_confirmed | integer | Count of new cases confirmed after positive test on this date | 34 |
| new\_deceased | integer | Count of new deaths from a positive COVID-19 case on this date | 2 |
| cumulative\_confirmed | integer | Cumulative sum of cases confirmed after positive test to date | 6447 |
| cumulative\_deceased | integer | Cumulative sum of deaths from a positive COVID-19 case to date | 133 |
| new\_recovered | integer | Count of new recoveries from a positive COVID-19 case on this date | 13 |
| cumulative\_recovered | integer | Cumulative sum of recoveries from a positive COVID-19 case to date | 133 |
| latitude | float | Floating point representing the geographic coordinate | 30.9756 |
| longitude | float | Floating point representing the geographic coordinate | 112.2707 |
| area\_sq\_km | float [squared kilometers] | Area encompassing this region | 3729 |
| area\_rural\_sq\_km | float [squared kilometers] | Area encompassing rural land in this region | 3729 |
| area\_urban\_sq\_km | float [squared kilometers] | Area encompassing urban land this region | 3729 |
| life\_expectancy | float [years] | Average years that an individual is expected to live | 75.722 |
| smoking\_prevalence | float [%] | Percentage of smokers in population | 16.9 |
| diabetes\_prevalence | float [%] | Percentage of persons with diabetes in population | 13.3 |
| infant\_mortality\_rate | float | Infant mortality rate (per 1,000 live births) | 9.8 |
| adult\_male\_mortality\_rate | float | Mortality rate, adult, male (per 1,000 male adults) | 143.719 |
| adult\_female\_mortality\_rate | float | Mortality rate, adult, female (per 1,000 male adults) | 98.803 |
| pollution\_mortality\_rate | float | Mortality rate attributed to household and ambient air pollution, age-standardized (per 100,000 population) | 13.3 |
| comorbidity\_mortality\_rate | float [%] | Mortality from cardiovascular disease, cancer, diabetes or cardiorespiratory disease between exact ages 30 and 70 | 16.6 |
| nurses\_per\_1000 | float | Nurses and midwives (per 1,000 people) | 5.8974 |
| physicians\_per\_1000 | float | Physicians (per 1,000 people) | 1.609 |
| location\_key | string | Unique string identifying the region | US\_CA\_06001 |
| country\_code | string | ISO 3166-1 alphanumeric 2-letter code of the country | US |
| country\_name | string | American English name of the country, subject to change. | Brazil |
| cumulative\_vaccine\_doses\_administered | integer | Cumulative sum of vaccine doses administered to persons | 923449 |
| average\_temperature | float [celsius] | Recorded hourly average temperature | 11.22 |
| minimum\_temperature | float [celsius] | Recorded hourly minimum temperature | 1.74 |
| maximum\_temperature | float [celsius] | Recorded hourly maximum temperature | 19.42 |
| rainfall | float [millimeters] | Rainfall during the entire day | 51 |
| relative\_humidity | float [%] | The amount of water vapor present in air expressed as a percentage of the amount needed for saturation at the same temperature | 43.09 |
| population\_largest\_city | integer | Population in the largest city of the country | 91256325 |

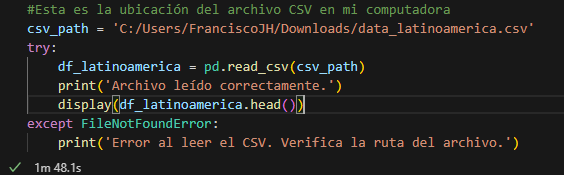
### Importar librerías

Importo las librerías numpy y pandas con el siguiente código



El código también se puede corroborar en el archivo[PIDA\_M4\_Francisco\_Hillebrand.ipynb](https://drive.google.com/open?id=1WHrB51UynhIkW96AmxLUrjQMSHfFU9A8), que encontrará en la carpeta de Drive.

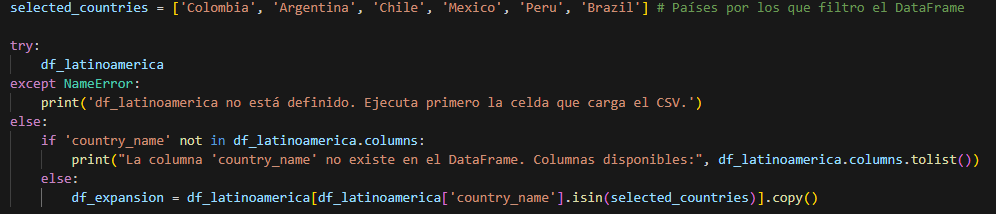
### Leer Archivo



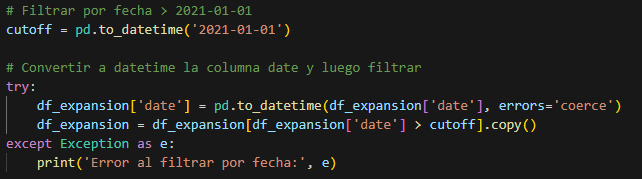
### Compruebo la cantidad de registros y columnas



### Filtro el Dataframe para obtener solo los registros de los países donde Biogenesys busca expandirse



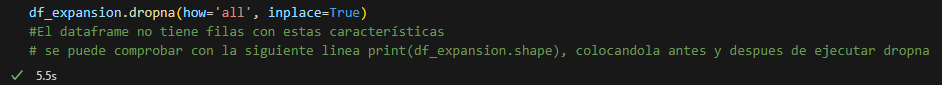
### Filtra los datos en fechas mayores a 2021-01-01



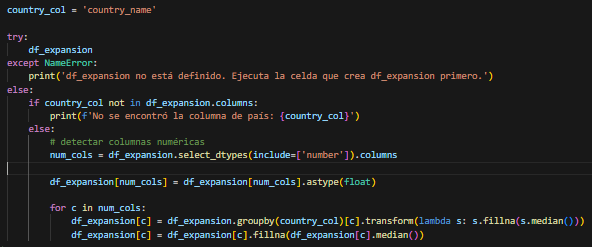
### Trabajo con valores nulos y faltantes

Para estos valores tome las siguientes decisiones:

* Se eliminarán los registros que tengan valores nulos en todas las columnas.



* Para los valores faltantes, decidí agrupar los registros por país, y reemplazarlos por la mediana de cada columna del dataframe. Si la mediana por país no se encuentra, entonces lo reemplazo por la mediana global de la columna.

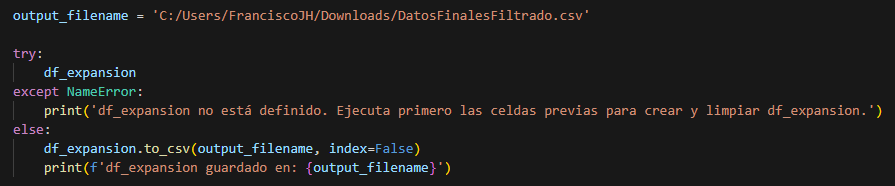


### Identificar variables claves

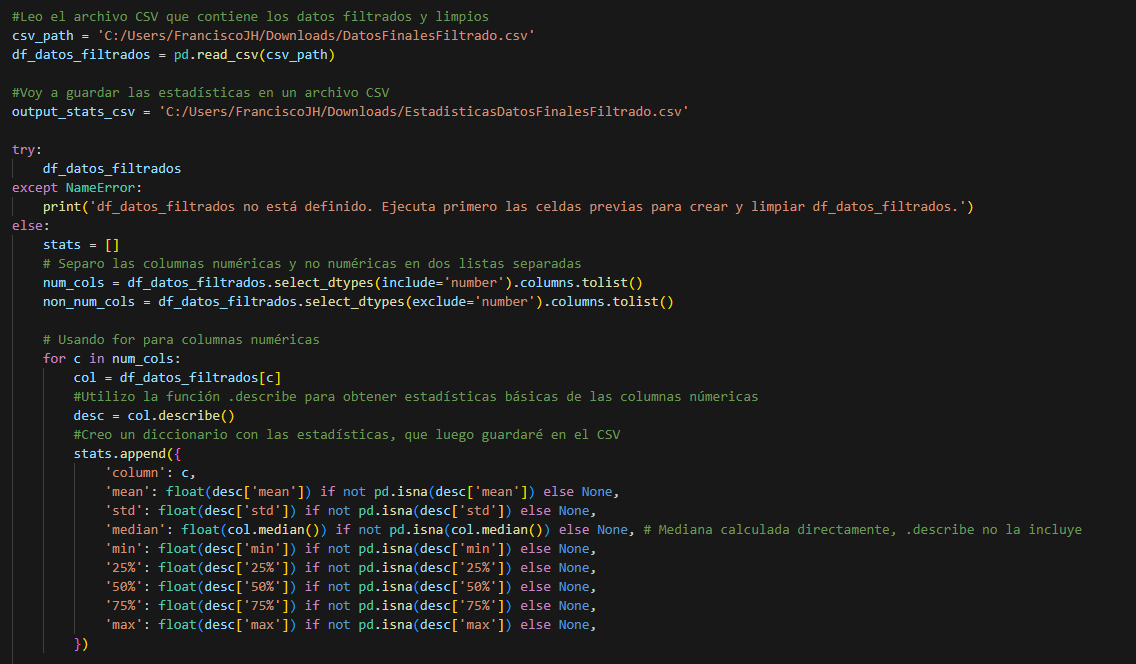
Considere que las variables claves para cumplir los objetivos propuestos para mi análisis son:

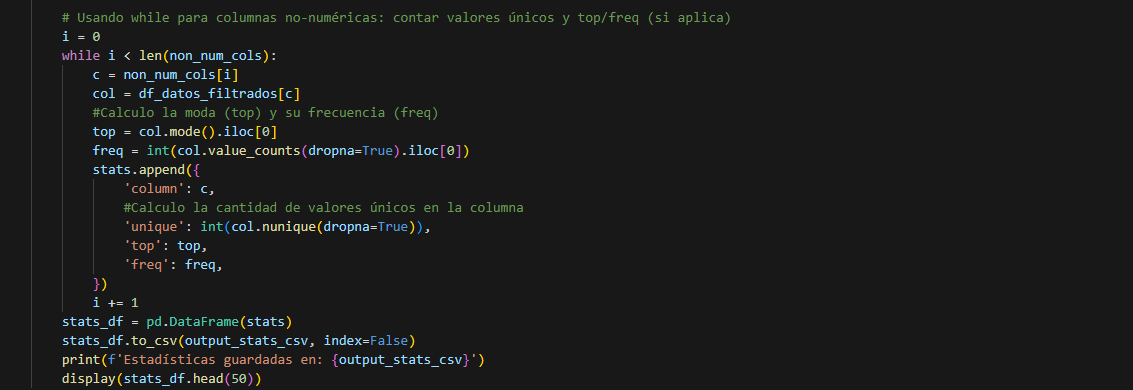
* **date:** Necesaria para hacer un análisis temporal del avance del COVID-19 y de las vacunación contra él.
* **population, population\_male, population\_female, urban\_population, population\_density y rural\_population:** Importante para entender el verdadero impacto del COVID-19 y las vacunas en la población y sus diferentes clasificaciones.
* **population\_age:** Estas columnas son muy importantes para descubrir el impacto de la enfermedad en las distintas edades.
* **new\_confirmed:** Necesaria para conocer el avance de la enfermedad.
* **new\_deceased:** Necesario para conocer el impacto negativo de la enfermedad.
* **new\_recovered y cumulative\_recovered:** Importante para conocer la respuesta de la población a la enfermedad.
* **nurses y physicians\_per\_1000:** Necesario para conocer si el sistema médico se encuentra o no preparado.
* **country\_name:** Para analizar los países correctos.
* **cumulative\_vaccine\_doses\_administered:** Es necesario saber cuantas vacunas se suministraron hasta una fecha específica.
* **cumulative\_confirmed y cumulative\_deceased:** Para conocer la tasa de aumento de contagios y de muertes por COVID-19.

### Guardar los datos filtrados en un archivo CSV



### Calculo estadísticas descriptivas





Una vez calculadas las estadísticas, respondo las preguntas:

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

* mean (media): valor promedio; mide tendencia central, sensible a outliers.
* median (mediana): otra medida de tendencia central, menos sensible a outliers en comparación con la media.
* std (desviación estándar): dispersión alrededor de la media; permite evaluar variabilidad.
* min/max: valores extremos; detectan posibles outliers o errores de entrada.
* percentiles (25%, 50% (mediana), 75%): resumen de la distribución; la mediana es útil cuando la distribución es asimétrica.
* unique (para columnas no numéricas): cantidad de valores distintos; útil para decidir si tratar como variable categórica o textual.
* top/freq (modo y frecuencia): moda y su frecuencia; útil para imputar categóricas o entender concentración.

Estas medidas, permiten:

* Decidir transformaciones, imputaciones (mediana vs media), y detectar columnas con poca información.
* Identifican outliers y errores de captura.
* Permiten comparar dispersión entre países y priorizar limpieza o análisis.

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

No; las funciones aplicadas distinguen tipos: para columnas numéricas se calculan media, distribución estándar, media, mínimo, máximo y los percentiles; para no-numéricas se calculan moda y su frecuencia, y la cantidad de valores únicos. Por lo tanto, no todas las métricas se aplican a todas las columnas.

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

Estadísticas como la media requiere datos numéricos. Aplicarlas a columnas no numéricas no tiene sentido y produce errores o resultados engañosos. Para variables categóricas se usan conteos y la moda.

Por eso es importante decidir métricas por tipo de variables. Es necesario inspeccionar outliers antes de sacar conclusiones.

### Función que obtiene la mediana, varianza y rango de un conjunto de valores



1. ¿Qué representa la mediana?

La mediana es el valor que se encuentra en el centro del conjunto de datos una vez ordenados.

Divide la muestra en dos partes iguales: 50% de los datos son menores o iguales a la mediana y 50% son mayores o iguales; además, es más robusta frente a valores extremos que la media.

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

La varianza mide cuánto se alejan los datos, en promedio, respecto a la media.

* Una varianza alta significa que los datos están muy dispersos.
* Una varianza baja indica que los datos están concentrados alrededor de la media.

El rango mide simplemente la diferencia entre el valor máximo y el mínimo.

* Es una medida de dispersión simple, pero puede verse afectada por valores atípicos.

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

* Si el rango y la varianza son pequeños, los datos están consistentes, concentrados cerca de la mediana, lo que indica poca variabilidad.
* Si el rango y la varianza son grandes, los datos son variables, lo que significa que, aunque la mediana marque un valor central, los valores individuales se alejan bastante de ese centro.

## AVANCE N°2

### Importar librerías



### Análisis Estadístico

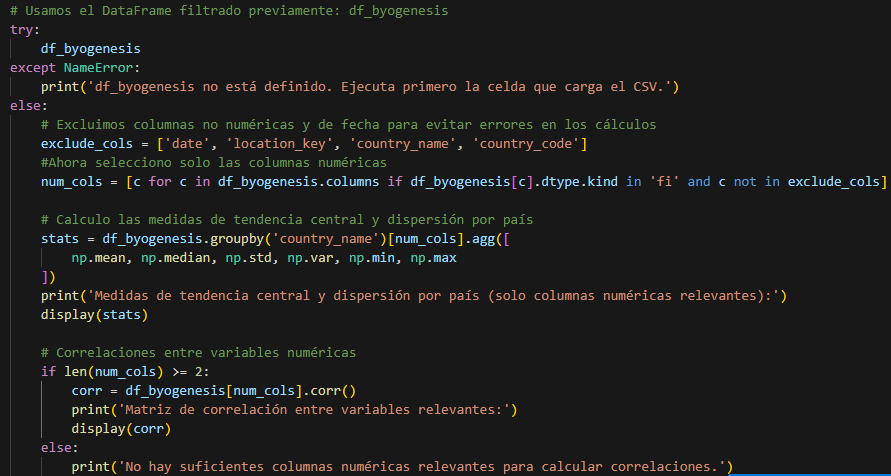
En este caso hice los siguientes pasos:

* Agrupe los registros por país, y para cada variable calculé las siguientes medidas:
  + Media
  + Mediana
  + Desviación estándar
  + Varianza
  + Mínimo
  + Máximo

Para observar los valores obtenidos, puedo hacerlo en [Stats por país](https://docs.google.com/spreadsheets/u/0/d/1gZ1LtpInaxwzx_Va04LP_RJiQqScaXN5UI2LILenwfY/edit)

* Cree una matriz de correlaciones entre todas las variables del dataset, la puede ver en [correlaciones\_byogenesis](https://docs.google.com/spreadsheets/u/0/d/1Rqw4mkNgMMQiOCB9D1ubbil3v_dtuMqzVcFeYbivoDc/edit).

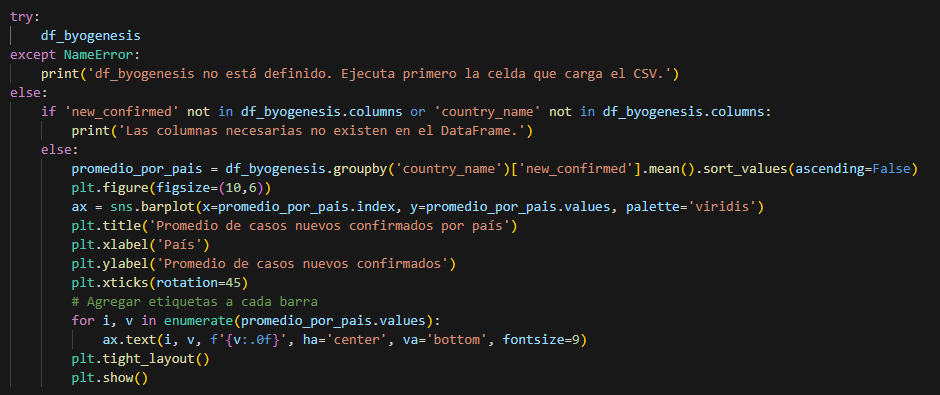
Los insights encontrados a partir de este análisis lo podrá ver en el apartado [EDA e Insights](#_4mjywtd4idxc)



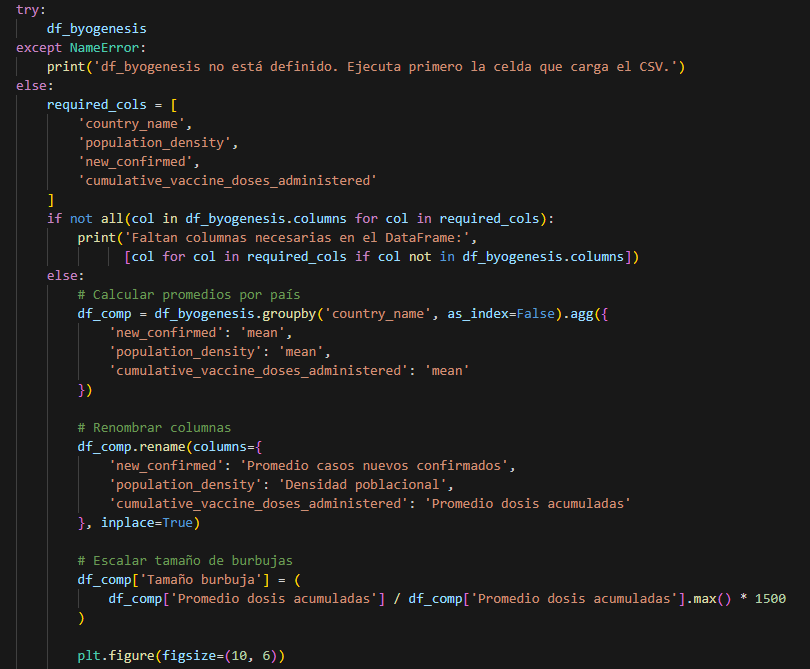
### Visualización de Datos con Matplotlib y Seaborn

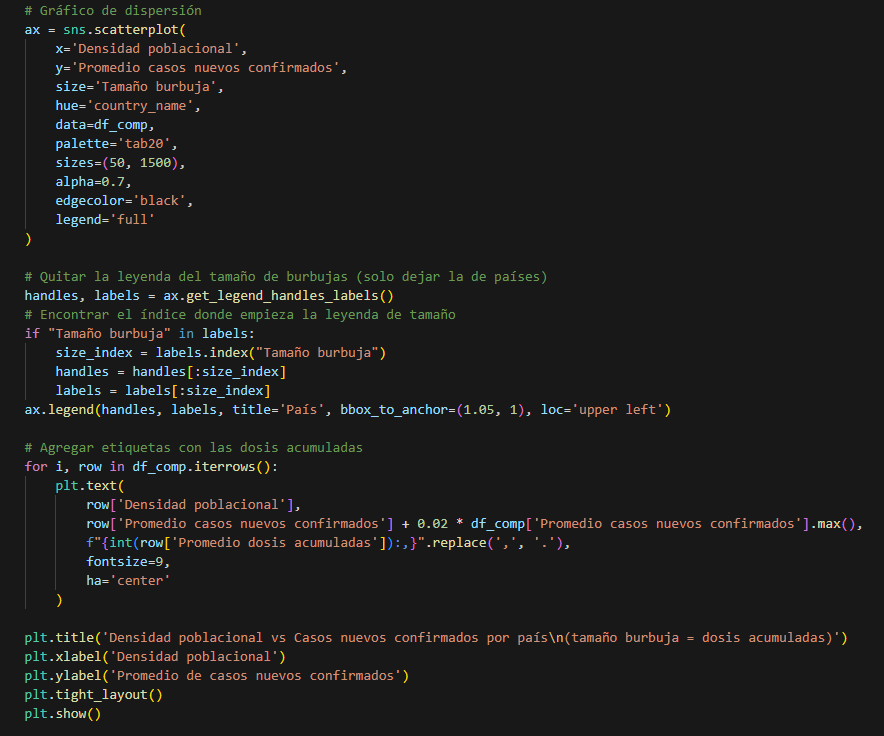
En este apartado se verá el código utilizado para realizar los gráficos del apartado [Visualización de los Insights](#_qljwssoddljh)

#### [Promedio de Casos Confirmados por país](#_l2x8y4in2ura)

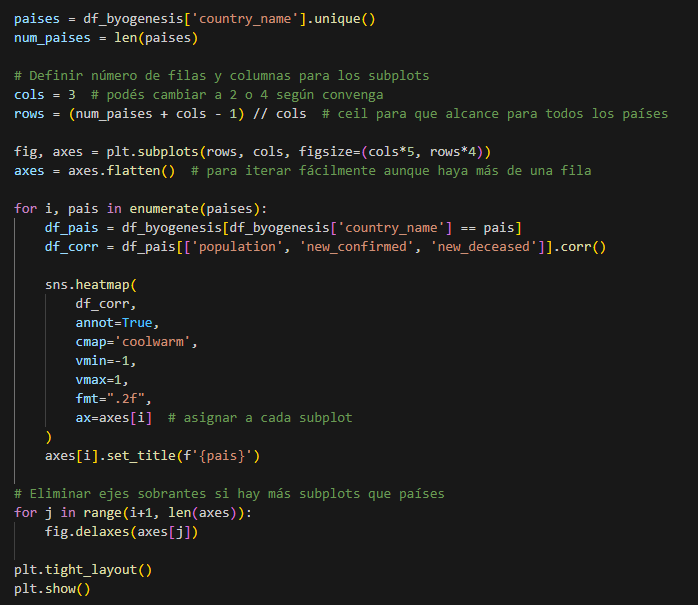


#### [Comparación de casos confirmados por país y su densidad poblacional](#_w4vcyyuhyx45)

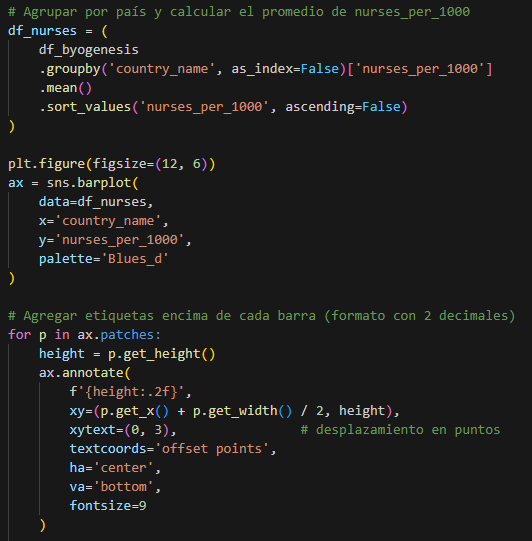


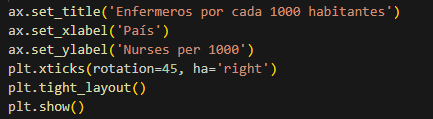


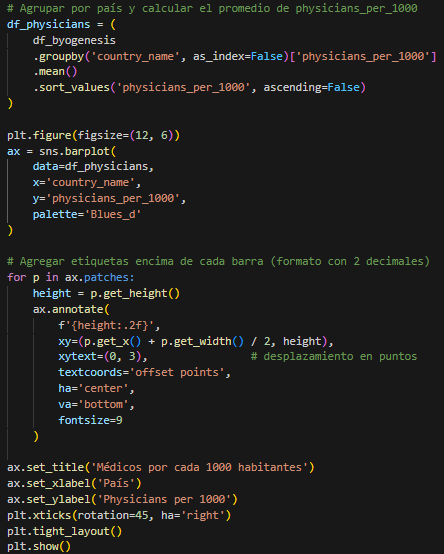
#### [Correlación entre nuevos casos confirmados, aumento de las muertes y aumento de la población a lo largo del tiempo](#_95z9giemrj94)



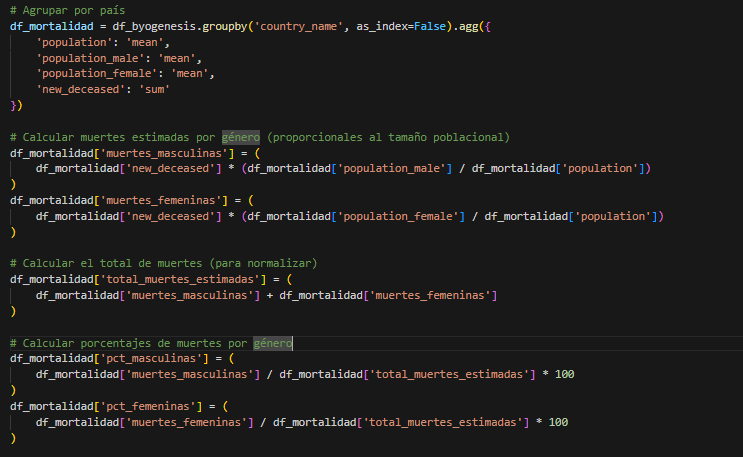
#### [Comparación de la estructura sanitaria del país](#_ww0hzy8h8kdr)

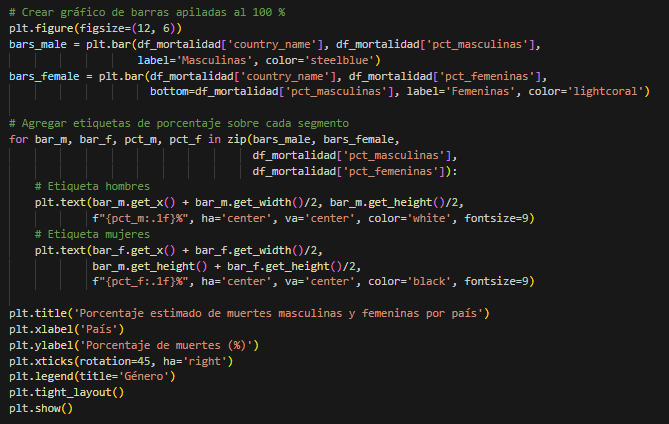






#### [Comparación del porcentaje de mortalidad femenina y masculina](#_k06e0abaiudp)

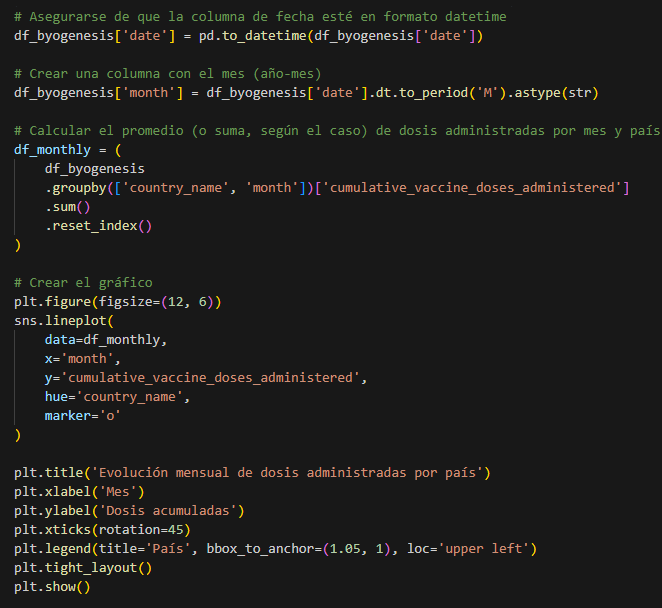


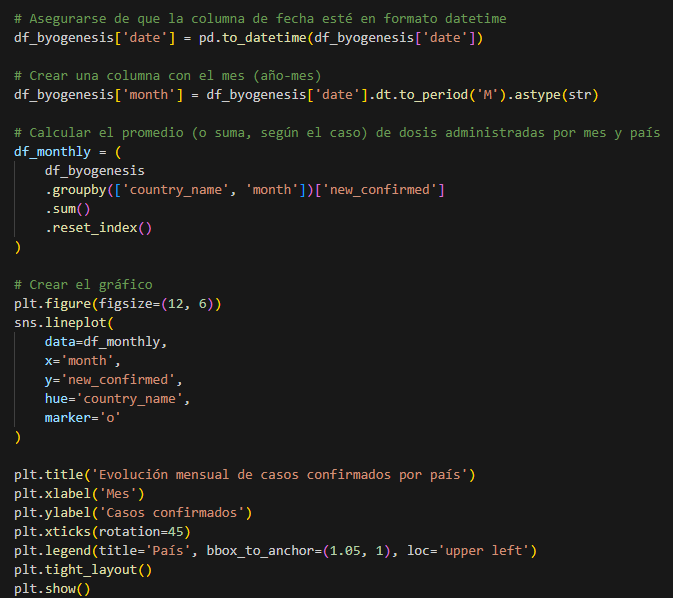


### Identificación de Tendencias y Patrones

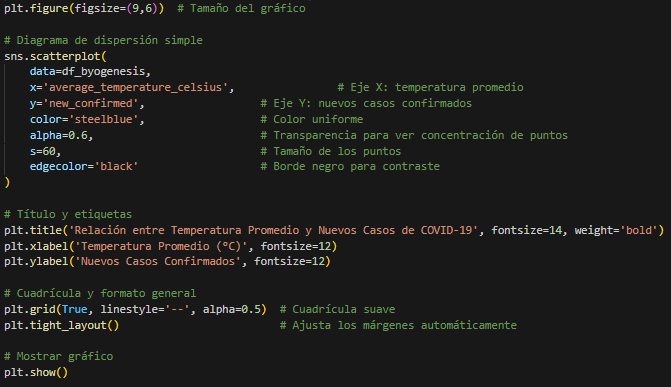
En este apartado se verá el código utilizado para realizar los gráficos del apartado [Identificación de Tendencias y Patrones](#_elvtgbf8ythh)

#### [Evolución de dosis administradas por mes de cada país](#_qkjvdq4ik0ax)

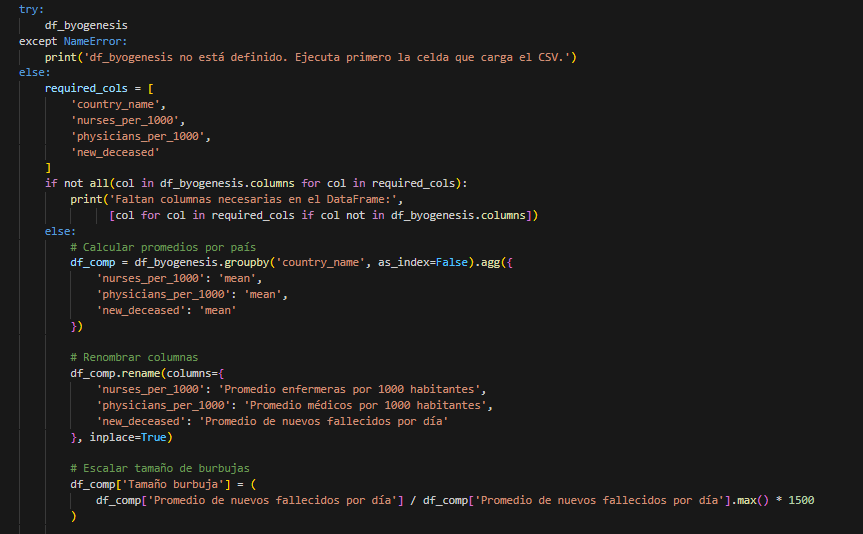


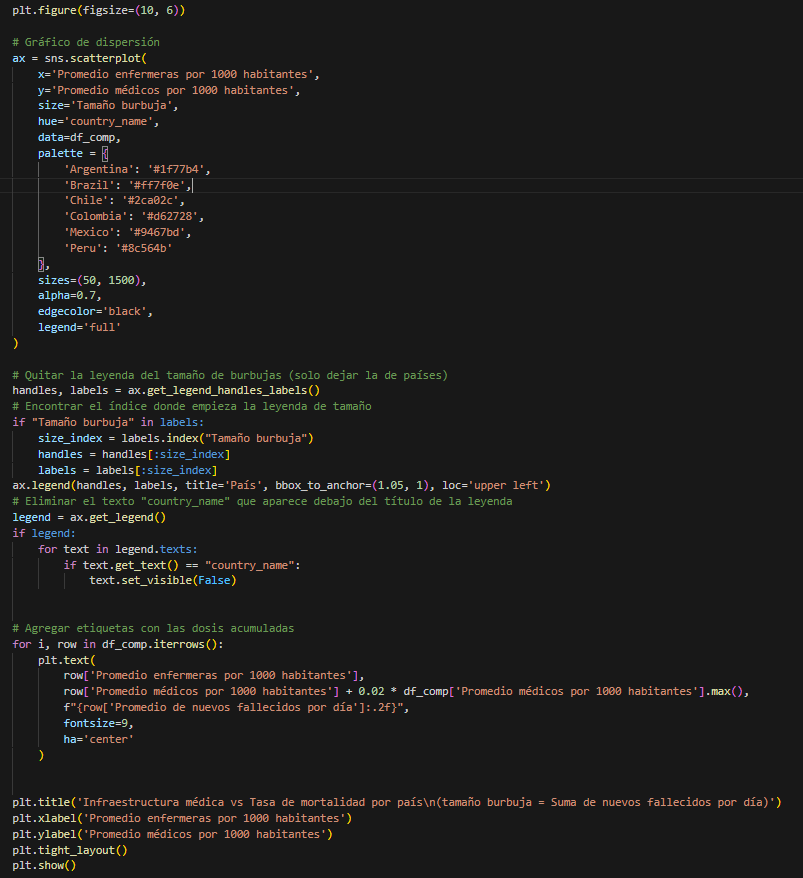


#### [Casos nuevos vs. temperatura promedio](#_fh1l7nhimscj)

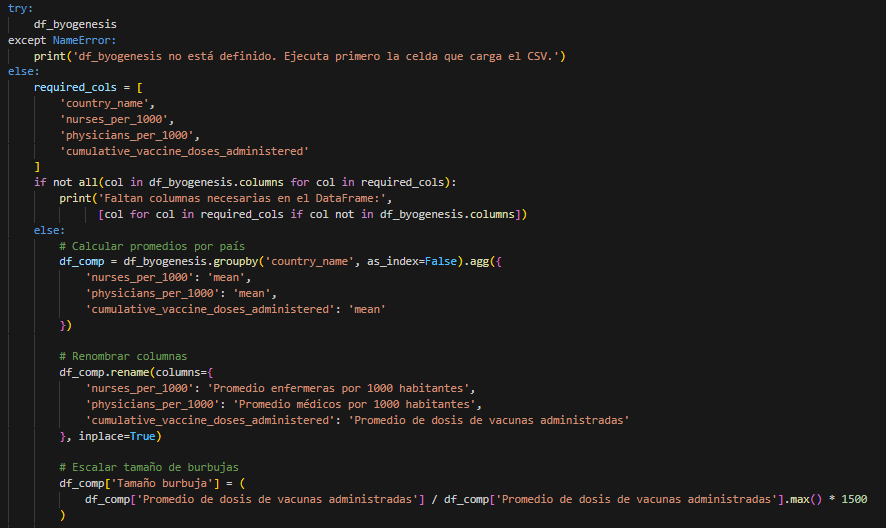


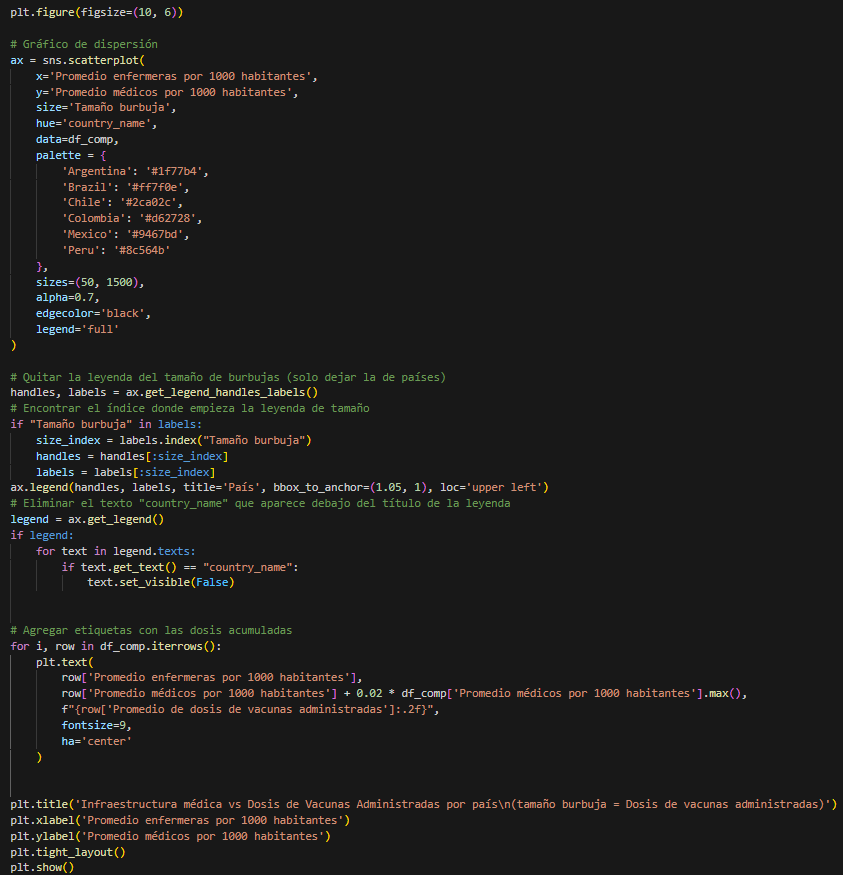
#### [Infraestructura médica vs. mortalidad](#_t5h5ujffuv78)



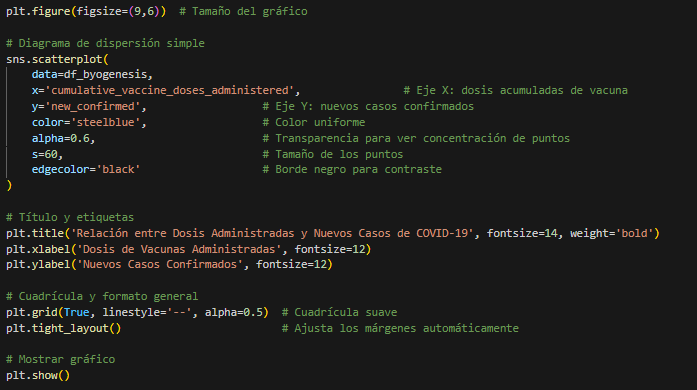


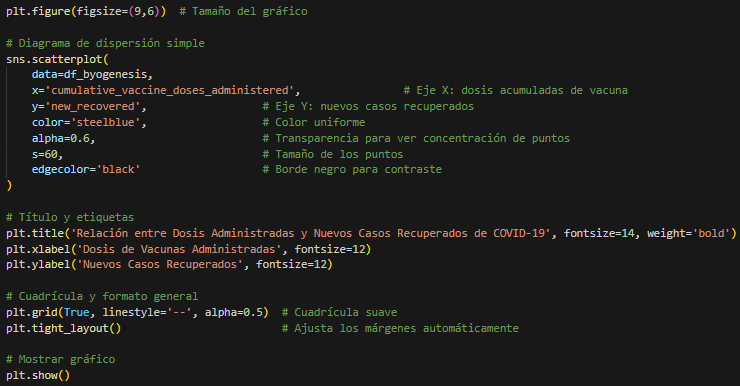
#### [Relación entre infraestructura médica y dosis administradas](#_aip5zktjzqp8)





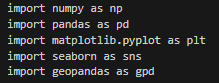
#### [Relación entre vacunación, aumento y recuperación de afectados](#_9d46ht9oldk6)





## AVANCE N°3

### Importar Librerías

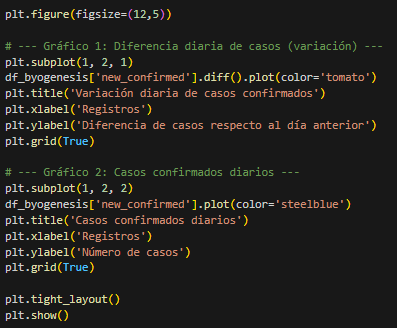


### EDA

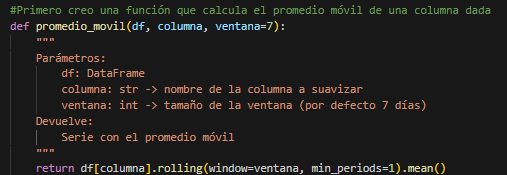
#### Análisis Temporal

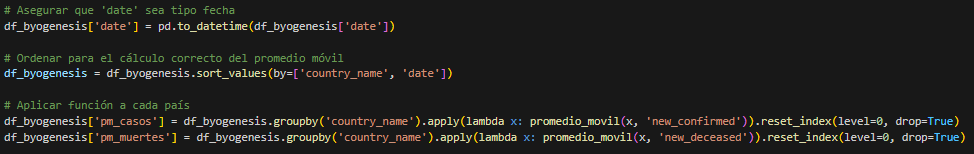
En esta sección dejo el código utilizado para obtener los gráficos que se muestran en la sección [Análisis Temporal](#_6bpvcnzevk5j).

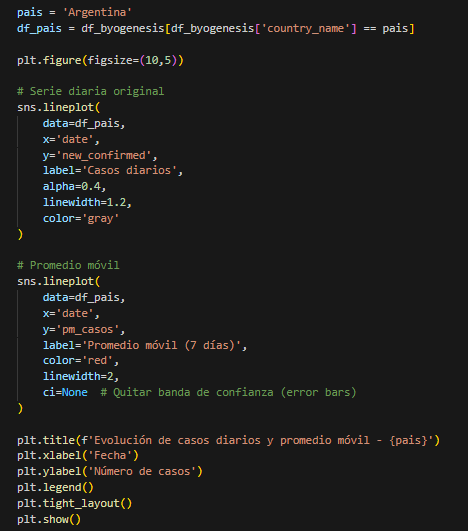
##### [Tasa de cambio de casos confirmados en general](#_v5zmgphjdng2)

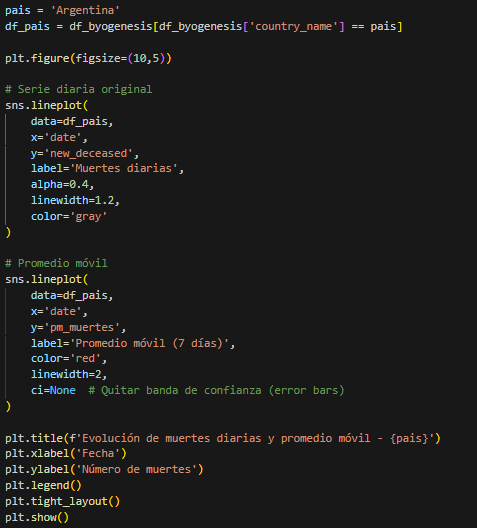


##### [Promedio Móvil de Casos Confirmados por País](#_hd0y0hibuw99) y [Promedio Móvil de Muertes por País](#_z6d70g6gb3ci)



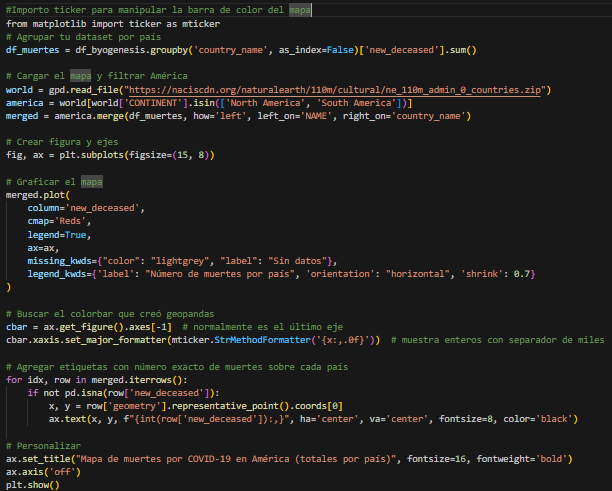




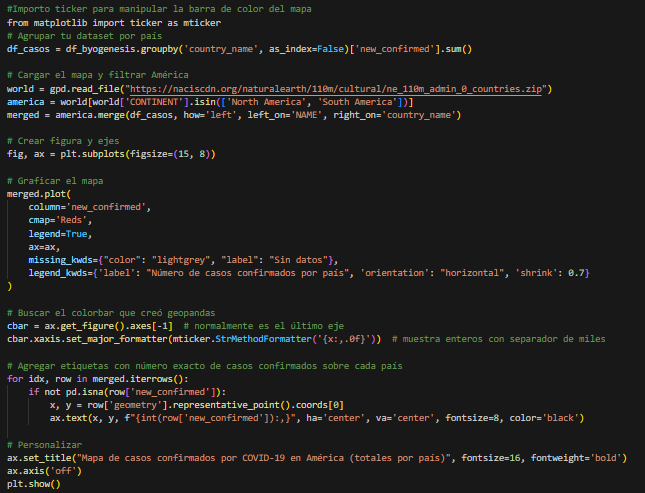


#### Análisis Espacial

##### [Total de muertes por País](#_5d5mn8gkbp1o)

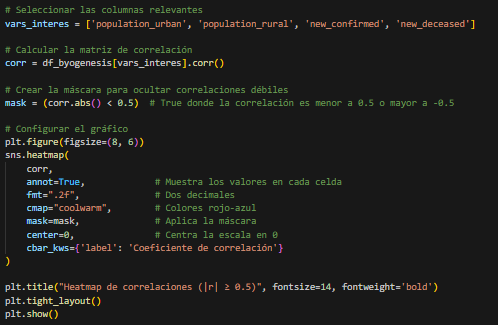


##### [Total de Casos Confirmados por País](#_quw5fw937nu1)

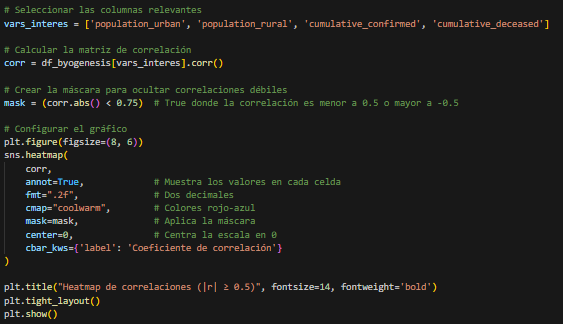


#### Otras Correlaciones

##### [Relación entre casos confirmados, fallecimientos y las población urbana y rural](#_f3bfmsx360kr)



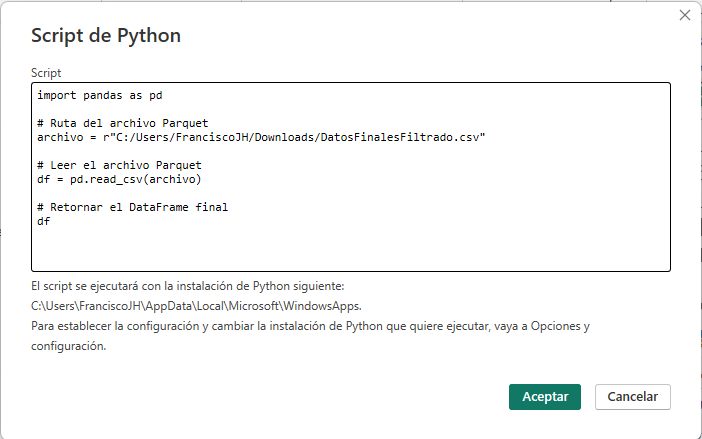
##### [Relación entre la la acumulación de fallecimientos y casos confirmados en las áreas urbanas y rurales](#_vq0t652tvhk3)



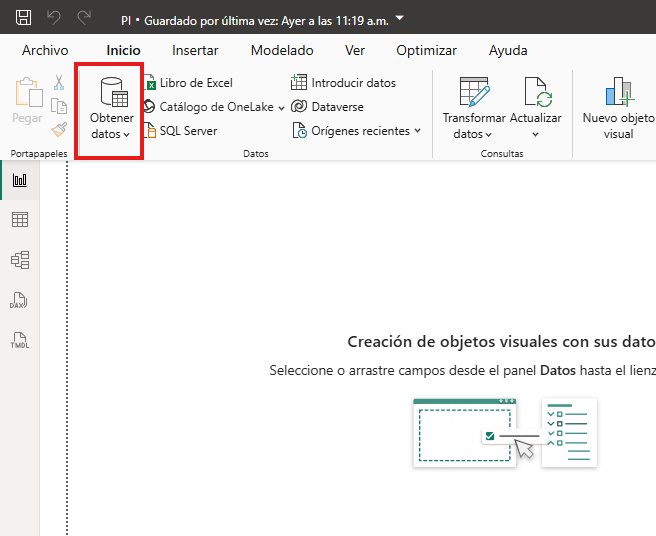
## AVANCE N°4

### Conexión de Python con Power BI

Para importar el CSV a Power BI utilice el siguiente script.



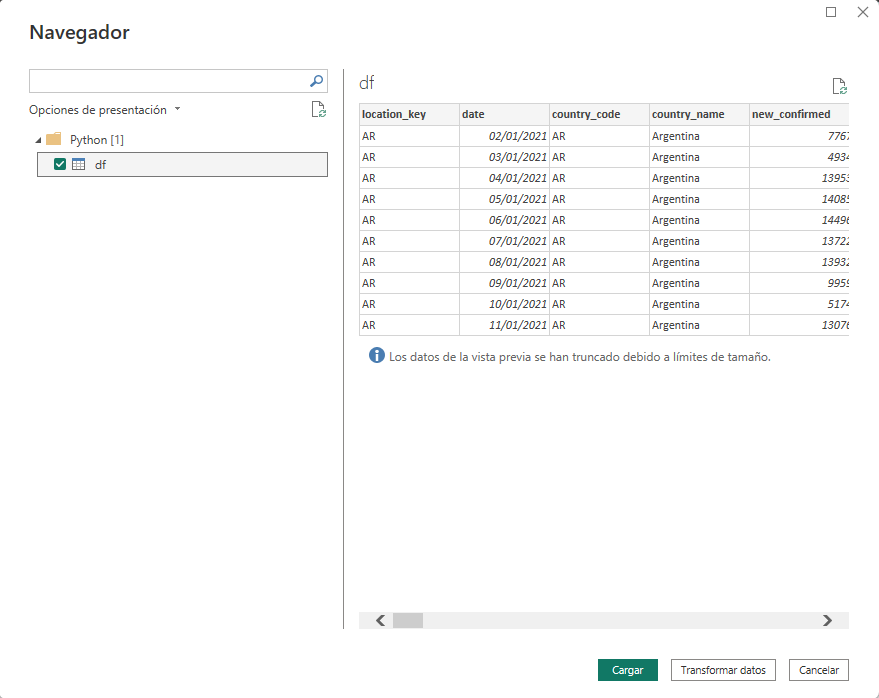
Esto lo hice desde la opción “Obtener datos”, en la pestaña “Inicio”.



Desde allí elegí la opción “Script de Python”, donde cargué el script anterior.

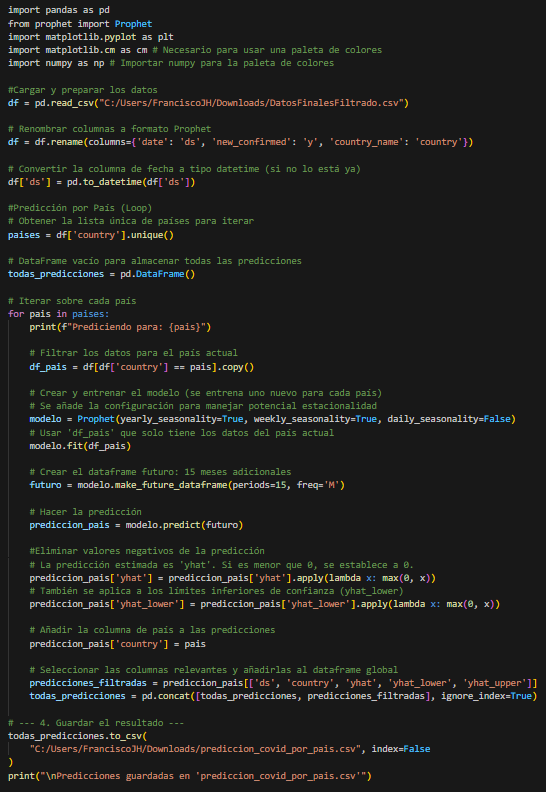


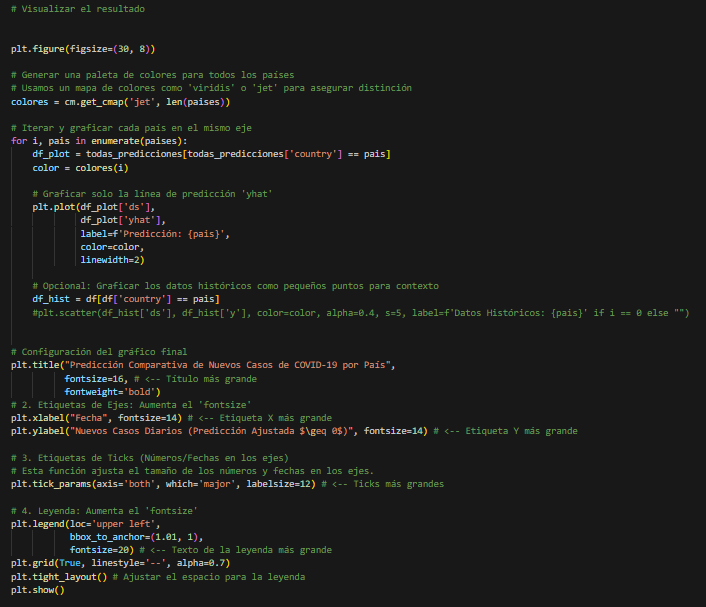
Una vez establecida la conexión, seleccione mi Dataframe, y cargue los datos.



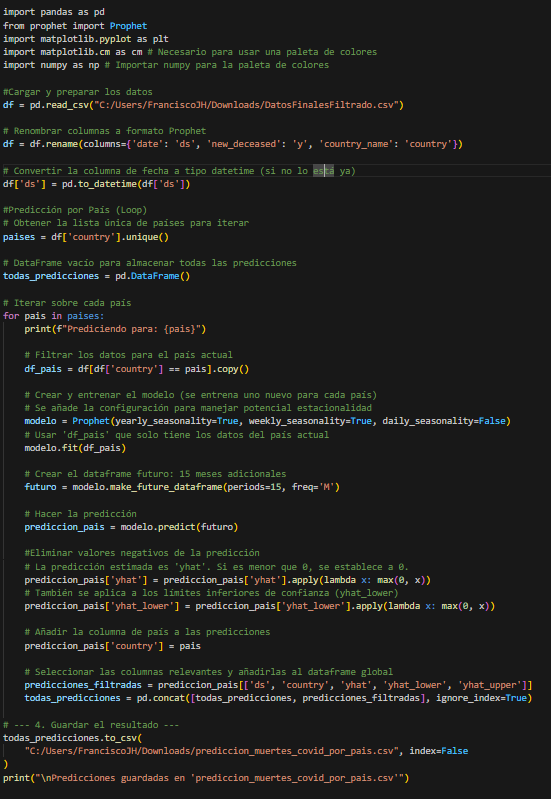
### Predicción de nuevos casos

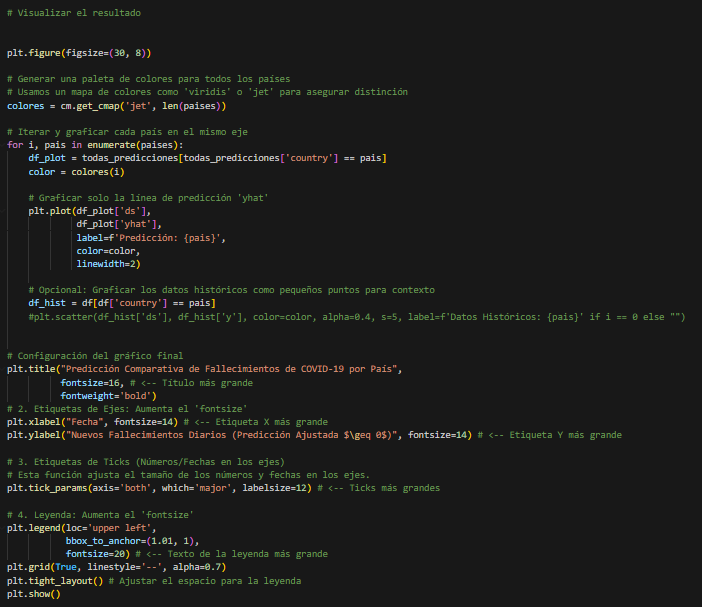
Este fue el código utilizado para generar una predicción de nuevos posibles casos confirmados de COVID-19. La predicción se ve en el tablero de Power BI.





### Predicción de fallecimientos





### Creación de Dashboards en Power BI

El dashboard lo encontrará en [PI.pbix](https://drive.google.com/open?id=1ne1Lfio4HAXwYxSesOTjCmJZN4f_Ldxn).

### Visualizaciones Estáticas vs Interactivas

| Visualización | Ventajas |
| --- | --- |
| Estática | * Ideales para reportes impresos o presentaciones fijas. * Consistencia y control total del formato. * No requieren conexión a Internet ni plataformas adicionales. * Adecuadas para auditorías o registros históricos. |
| Interactiva | * Los usuarios pueden filtrar, seleccionar rangos de fechas o hacer drill-down en los datos. * La interacción ayuda a descubrir patrones, anomalías y relaciones que no se ven en gráficos fijos. * Los dashboards pueden conectarse a fuentes de datos en tiempo real. * Cada perfil (gerente, analista, técnico) puede visualizar los datos relevantes para su área. |

# Conclusiones y Recomendaciones

Luego de realizar el análisis, podemos observar los siguientes resultados:

* Brasil es el país con más casos y fallecimientos, por ende, también fue el país que más vacunas administró.
* México es el segundo país que más fallecidos registró, y por ende el segundo país que más vacunas otorgó. Hay que tener en cuenta que es el país con mayor densidad poblacional, algo muy importante teniendo en cuenta que el COVID-19 se contagia fácilmente.
* Argentina es el segundo país con más casos confirmados, pero el segundo con menos fallecidos, quizás se deba a que fue el tercer país que más vacunas administró. Además, es el país con menor densidad poblacional.
* Colombia es el segundo país con mayor densidad poblacional.
* Chile es el país con el promedio de médicos y enfermeros cada 1000 habitantes seguido por Brasil. En el caso de Chile, es el país con mejor expectativa de vida, y en el caso brasilero, su expectativa de vida es similar al resto de países (menos Chile).
* Brasil y México son los países con el PBI más alto.
* Los contagios y fallecidos suben en temperaturas de entre 20°C y 30°C.
* Los casos y fallecimientos suben en países con mayor población.
* Brasil y México son los países con mayor población de riesgo (60 años o más).
* El orden de los países en cuanto a mayor PBI per cápita son:
  + Chile
  + Argentina
  + México
  + Brasil
  + Perú
  + Colombia
* A medida que se administran más vacunas, ha subido la cantidad de recuperados.
* Se estima que Brasil tendrá la mayor cantidad de fallecidos para enero de 2023, y el segundo país con mayor cantidad de nuevos casos para la misma fecha.
* Se estima que para enero de 2023, Argentina tendrá la mayor cantidad de casos confirmados, y será el segundo país con más fallecimientos.
* Se estima que Colombia será el tercer país con más casos confirmados y el segundo en fallecimientos para enero de 2023.
* Se estima que México será el país con menos casos confirmados y menos fallecimientos a enero de 2023.
* Al igual que México, Chile será el país con menos fallecimientos a enero de 2023. Y se estima que para esa fecha será el segundo país con menos contagios.

Teniendo en cuenta estos resultados, como analista de datos, recomiendo a Byogenesis lo siguiente:

* Expandirse principalmente a Brasil: Recomiendo que la mayor cantidad de laboratorios se instalen en este país. Esto debido a que: fue el país más afectado en el pasado y además se estima que sea uno de los más afectados a futuro, por ello necesitaran muchas dosis de vacunas, las cuales Byogenesis puede otorgar. Además, Brasil es el país con el mayor PBI, por lo que es más probable que pueda costear las vacunas que nosotros ofrecemos. Otro dato importante de este país es la buena cantidad de médicos y enfermeros cada 1000 habitantes que posee, los cuales son necesarios para la administración de vacunas. También hay que tener en cuenta que es un país cálido, lo cual, como vimos, aumenta la probabilidad de contagios. Por último, es importante destacar que el futuro de Brasil es riesgoso.
* Uno de los dos países donde más laboratorios pondría detrás de Brasil, sería México: Recomiendo esto porque es el segundo país con el mejor PBI, y el tercero con el mejor PBI per cápita, por lo que es más probable que su gobierno costee un programa de vacunación. Es importante destacar que es el país de mayor densidad poblacional y además el segundo país con mayor población de riesgo (personas de 60 años o más).
* El segundo país donde más laboratorios pondría detrás de Brasil y junto a México sería Argentina: Fue el segundo país con más casos, pero fue el segundo con menos fallecidos, esto gracias a que fue el tercer país que más vacunas administró. Es decir, que Argentina conoce los beneficios de la vacuna, por lo que es probable que las solicite. Además, las estimaciones futuras de Argentina no son alentadoras.
* El cuarto país donde más invertiría en la instalación de laboratorios sería Colombia: Un país cálido cuya proyección a futuro no es buena (en cuanto a fallecimientos y nuevos casos), por lo que es probable que necesite vacunas. Además es el segundo país con mayor densidad poblacional, por lo que es probable que aumenten los contagios. También es importante destacar que es el país con el peor número de médicos y enfermeros por cada 1000 habitantes, por lo que no posee demasiado personal sanitario para administrar vacunas.
* El quinto país donde invertiría sería Perú: Su PBI es el más bajo de los 6 países, y el PBI per cápita es el segundo más bajo, por lo que considero que al gobierno peruano le costaría invertir en un programa de vacunación. Además, su proyección a futuro no es ideal, pero tampoco terrible, ya que es el cuarto país en la estimación de posibles fallecidos y casos confirmados para enero de 2023. También hay que destacar que es uno de los países con menor presencia de médicos y enfermeros por cada 1000 habitantes. Por último, es importante destacar que es el segundo país con menor población de riesgo.
* El país en el que menos invertiría sería Chile: Este es un país que tiene una buena infraestructura sanitaria, pero se estima que a futuro sea de los menos afectados. Además fue el país menos afectado, por ende el que menos vacunas administró.

En resumen, invertiría en los países siguiendo este orden:

1. Brasil
2. México y Argentina
3. Colombia
4. Perú
5. Chile

La estrategia general recomendada para Biogenesys consiste en priorizar la instalación de laboratorios en países con alta necesidad sanitaria (gran cantidad de casos y fallecimientos en el pasado y a futuro) y capacidad económica suficiente para sostener programas de vacunación, comenzando por Brasil y México. En paralelo, mantener presencia limitada en países con menor proyección de casos, como Chile.

# Reflexión Personal

Este proyecto validó que la calidad de los datos es la base de toda decisión estratégica, lo que requirió una dedicación significativa en la fase de limpieza y ETL. Como Analista de Datos, consolidé habilidades clave. Adquirí conocimientos en el modelo Prophet para la predicción de series de tiempo, gestionando activamente las limitaciones de los datos al extrapolar. Demostré la capacidad de realizar análisis multidimensionales al combinar y correlacionar variables diversas como la demografía, la incidencia de la enfermedad y el clima, lo cual fue esencial para formular la recomendación final. Finalmente, desarrollé la destreza de crear visualizaciones estratégicas en Python (heatmaps, gráficos comparativos), necesarias para la toma de decisiones. Además conecté este lenguaje a Power BI, asegurando que las conclusiones se puedan visualizar de manera interactiva, y así apoyar a la toma de decisiones.

Además, si tuviese que empezar nuevamente este proyecto, me concentraría aún más en el conocimiento de las variables, para hacer el proceso de limpieza de datos más eficiente, algo que me costó al realizar este proyecto.