**Expansión estratégica de laboratorios para investigación y vacunación de COVID-19**

**Nombre del autor: Ernesto Castro**

**Email: ercastrob76@gmail.com**

**Cohorte: DA-FT08**

**Fecha de entrega:** 17-11-2024

**Institución:** 

# Introducción

# La empresa farmacéutica BIOGENESYS 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.

# Desarrollo del proyecto

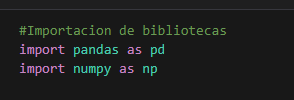
**Avance 1**

El primer paso realizado fue la selección de las columnas con las que se trabajaría en el análisis, en la siguiente tabla se muestran resaltadas en verde las columnas del data set original que se utilizaran para el análisis, las que no son usadas tienen el estado Eliminar y el criterio por el cual se eliminan.

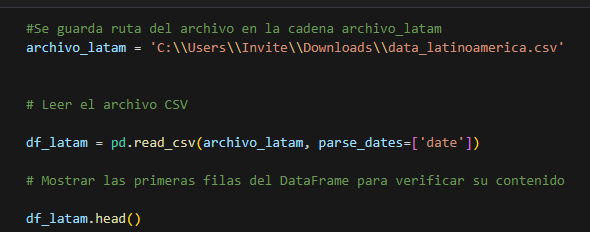
| **COLUMNA** | **Estado** | **Criterio de eliminación** |
| --- | --- | --- |
| location\_key' | Utilizar |  |
| date' | Utilizar |  |
| country\_code' | Eliminar | No relevante |
| country\_name' | Utilizar |  |
| new\_confirmed' | Utilizar |  |
| new\_deceased' | Utilizar |  |
| cumulative\_confirmed' | Utilizar |  |
| cumulative\_deceased' | Utilizar |  |
| cumulative\_vaccine\_doses\_administered' | Utilizar |  |
| population' | Utilizar |  |
| population\_male' | Eliminar | Baja correlacion-Incidencia |
| population\_female' | Eliminar | Baja correlacion-Incidencia |
| population\_rural' | Eliminar | Baja correlacion-Incidencia |
| population\_urban' | Eliminar | Baja correlacion-Incidencia |
| population\_density' | Utilizar |  |
| human\_development\_index' | Utilizar |  |
| population\_age\_00\_09' | Eliminar | Baja correlacion-Incidencia |
| population\_age\_10\_19' | Eliminar | Baja correlacion-Incidencia |
| population\_age\_20\_29' | Eliminar | Baja correlacion-Incidencia |
| population\_age\_30\_39' | Utilizar |  |
| population\_age\_40\_49' | Utilizar |  |
| population\_age\_50\_59' | Utilizar |  |
| population\_age\_60\_69' | Utilizar |  |
| population\_age\_70\_79' | Utilizar |  |
| population\_age\_80\_and\_older' | Utilizar |  |
| gdp\_usd' | Utilizar |  |
| gdp\_per\_capita\_usd' | Utilizar |  |
| latitude' | Eliminar | Baja correlacion-Incidencia |
| longitude' | Eliminar | Baja correlacion-Incidencia |
| area\_sq\_km' | Eliminar | No relevante |
| smoking\_prevalence' | Eliminar | demasiados valores nulos |
| diabetes\_prevalence' | Eliminar | demasiados valores nulos |
| infant\_mortality\_rate' | Utilizar |  |
| nurses\_per\_1000' | Utilizar |  |
| physicians\_per\_1000' | Utilizar |  |
| average\_temperature\_celsius' | Utilizar |  |
| minimum\_temperature\_celsius' | Eliminar | Baja correlacion-Incidencia |
| maximum\_temperature\_celsius' | Eliminar | Baja correlacion-Incidencia |
| rainfall\_mm' | Eliminar | Baja correlacion-Incidencia |
| relative\_humidity' | Eliminar | Baja correlacion-Incidencia |
| population\_largest\_city' | Eliminar | No relevante |
| area\_rural\_sq\_km' | Eliminar | No relevante |
| area\_urban\_sq\_km' | Eliminar | No relevante |
| life\_expectancy' | Eliminar | No relevante |
| adult\_male\_mortality\_rate' | Utilizar |  |
| adult\_female\_mortality\_rate' | Utilizar |  |
| pollution\_mortality\_rate' | Utilizar |  |
| comorbidity\_mortality\_rate' | Utilizar |  |
| new\_recovered' | Eliminar | demasiados valores nulos |
| cumulative\_recovered' | Eliminar | demasiados valores nulos |

Es conveniente aclarar que en un análisis con un data set previo se encontró que la correlación de algunas de las columnas era muy baja y por esto se descartan del análisis. Debido a que en este primer análisis se filtró principalmente por número de valores nulos también se reconsidero la inclusión de columnas inicialmente descartadas ya que eran relevantes de acuerdo a los criterios de estudio para la expansión.

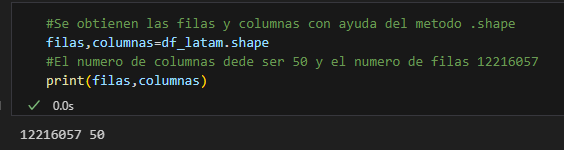
Para el tratamiento de los datos en python primero importamos las librerias necesarias

****

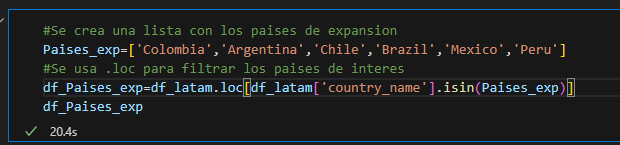
Se realiza la lectura del archivo .csv



Se comprueban las dimensiones del data set



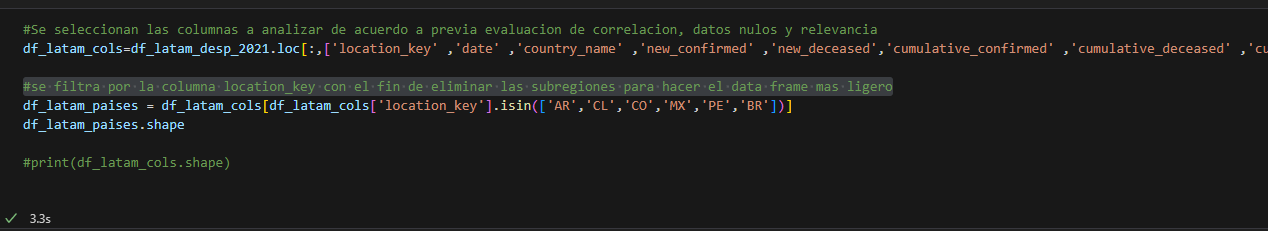
Se realiza un primer filtro para utilizar solo los países de expansión.



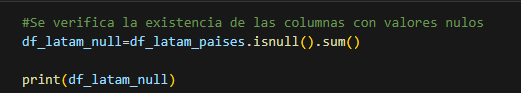
Se realiza un segundo filtro en el cual solo se tienen en cuenta datos posteriores al 01-01-2021



Se crea un nuevo data frame que solo incluya las columnas seleccionadas para el análisis y luego se filtra por la columna location\_key para eliminar las subregiones del data frame y que este sea de un tamaño más manejable.

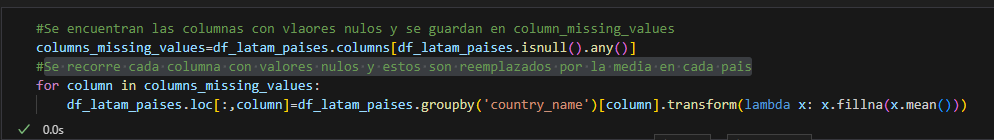


Luego se procedió a verificar los valores nulos por columnas y el resultado se muestra en la tabla después de la captura de pantalla del código.

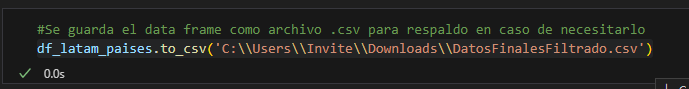


| **Columnas con datos nulos** | **valores nulos** | **Acción** |
| --- | --- | --- |
| new\_confirmed | 21 | reemplazar con media |
| new\_deceased | 21 | reemplazar con media |
| cumulative\_confirmed | 21 | reemplazar con media |
| cumulative\_deceased | 21 | reemplazar con media |
| cumulative\_vaccine\_doses\_administered | 584 | reemplazar con media |
| average\_temperature\_celsius | 42 | reemplazar con media |

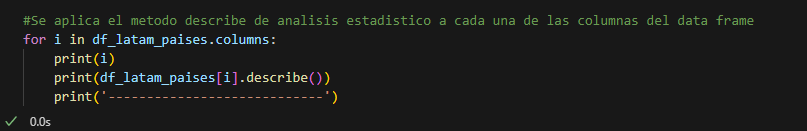
Se procede a reemplazar los valores nulos por la media por país.



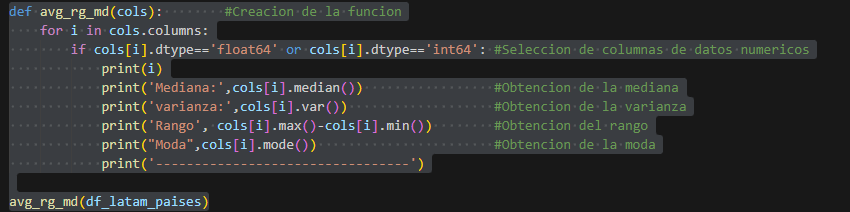
Se guarda el data frame como archivo csv para respaldo en caso de necesitarlo.



Se realiza un cálculo de métricas descriptivas

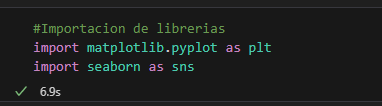


Se crea una función para tener la mediana, la varianza y el rango.

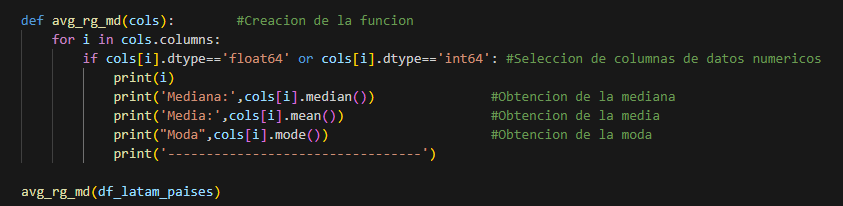


**Avance 2**

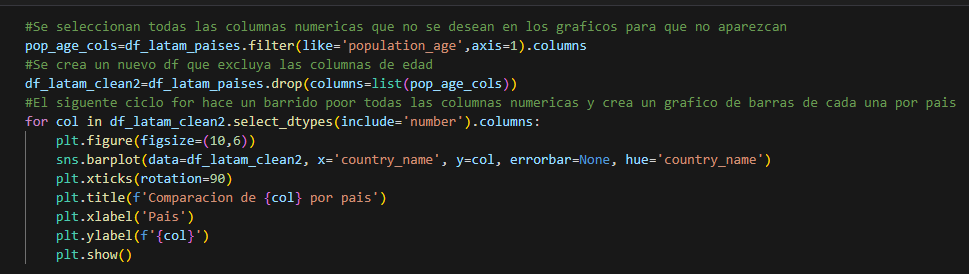
Importación de librerías

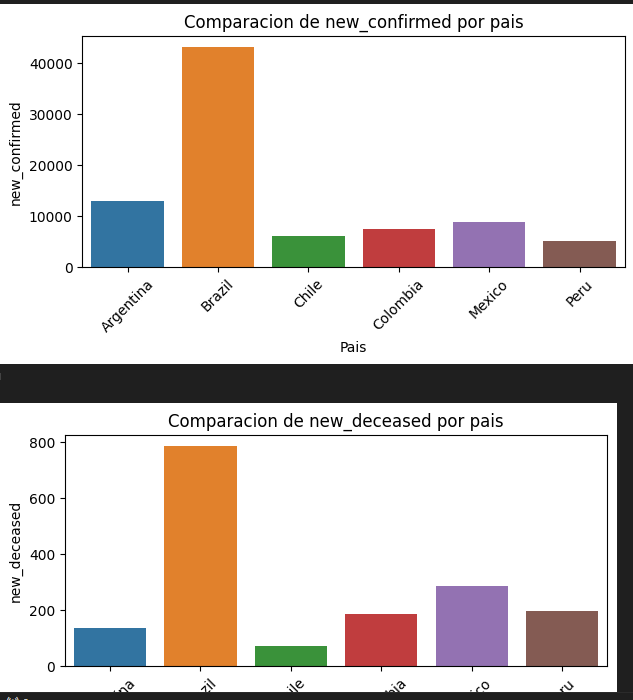
****

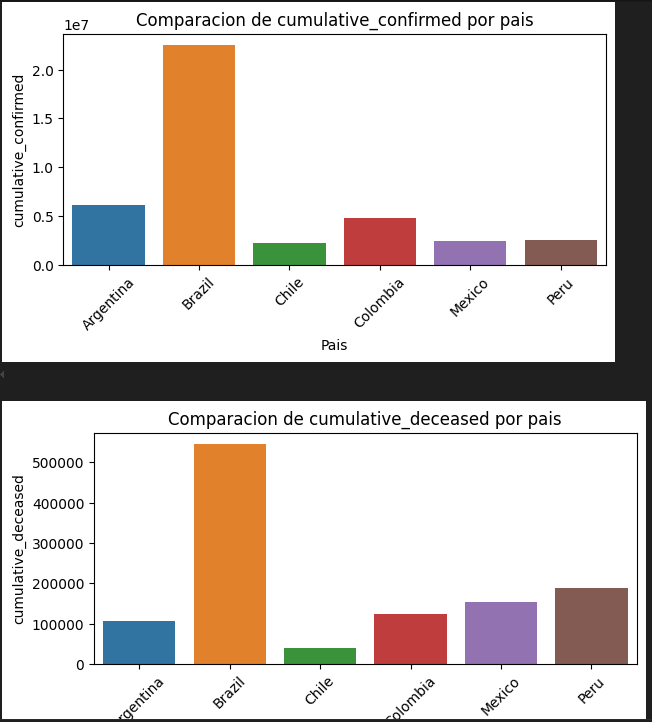
Medidas de tendencia central

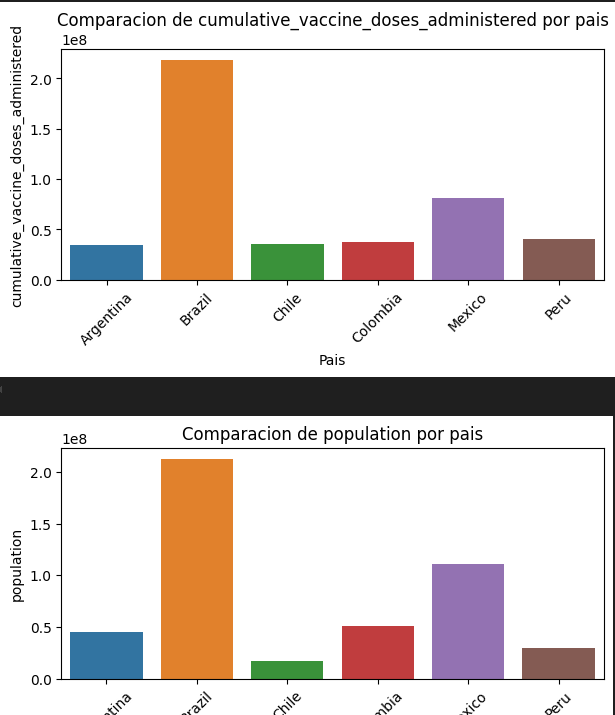


Visualización de datos -gráficos de barra países contra variables.

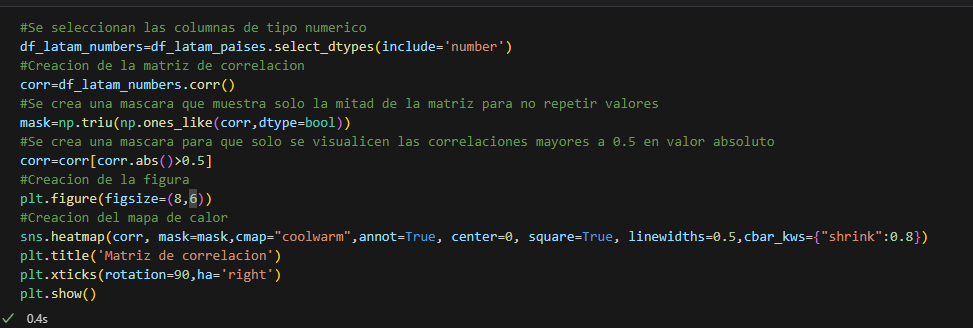


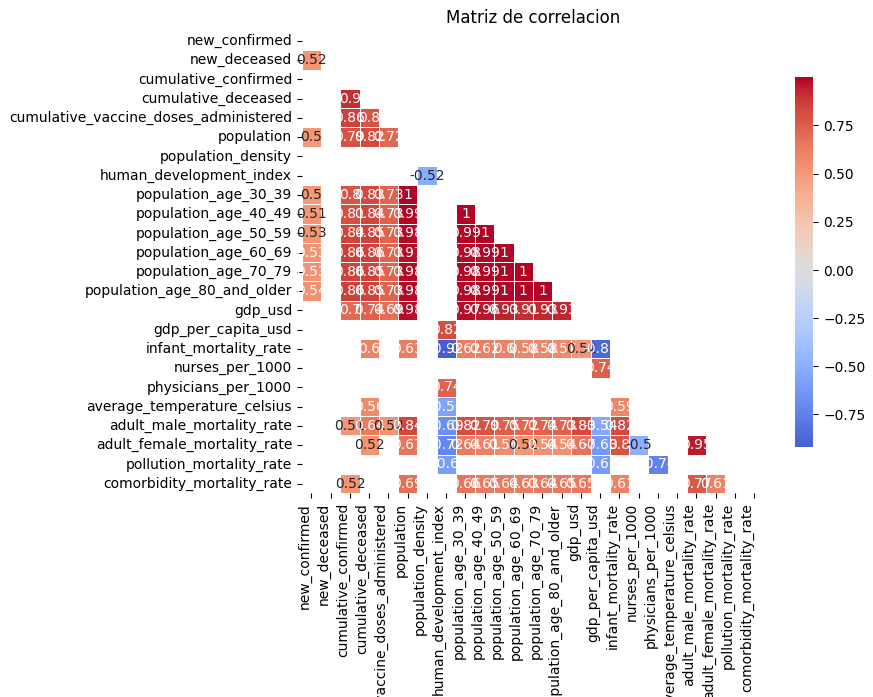




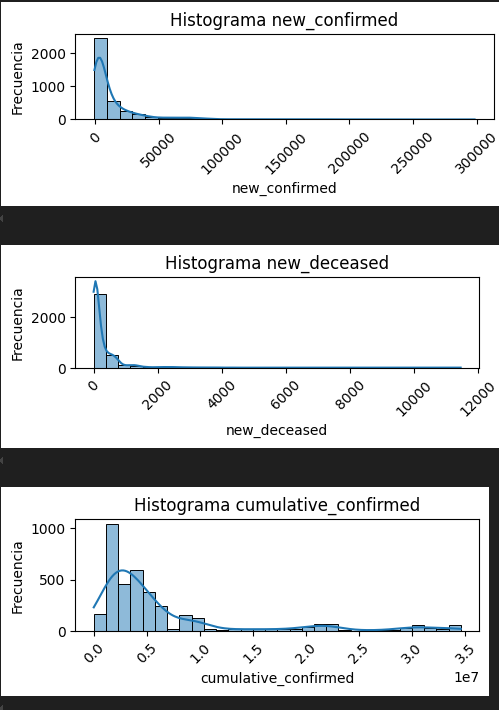


# Matriz de correlación en mapa de calor

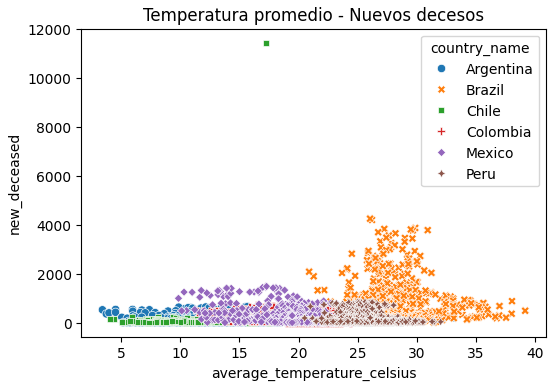




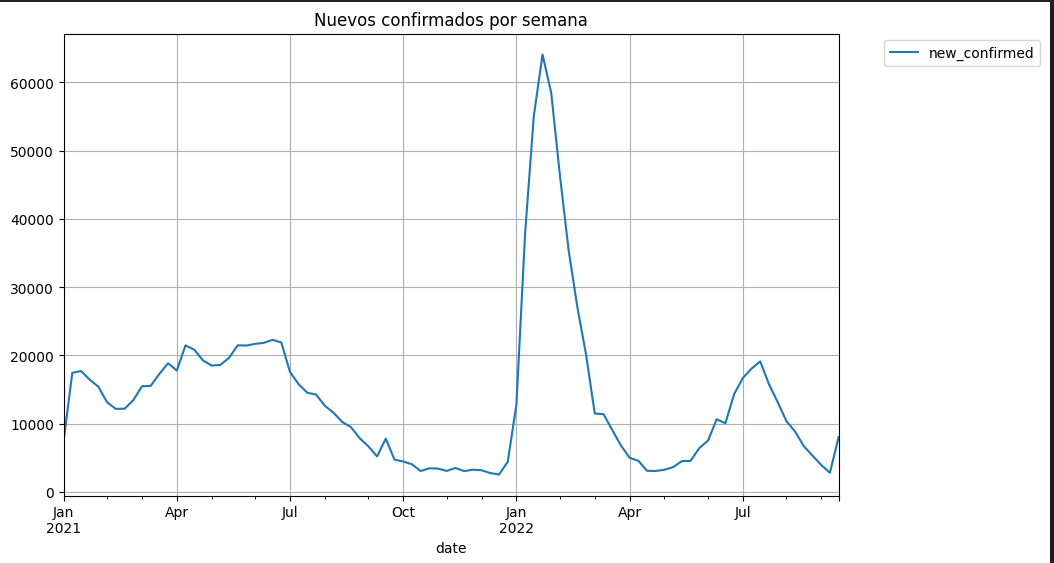
# Histogramas de columnas con variación.



# Gráficos de dispersión.

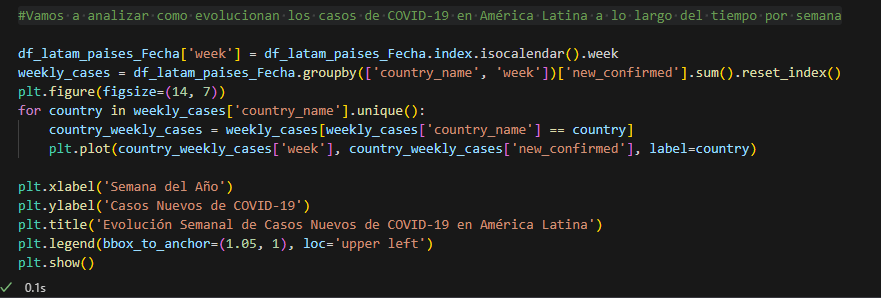


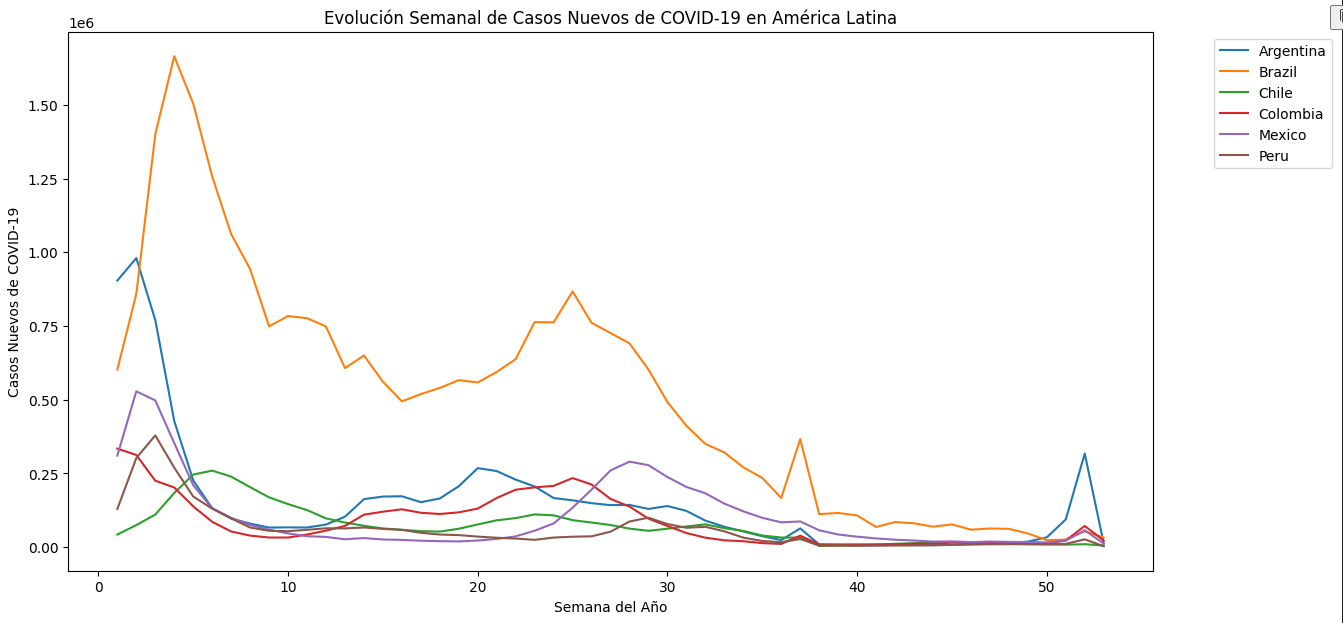
# Gráfico de nuevos confirmados por semana



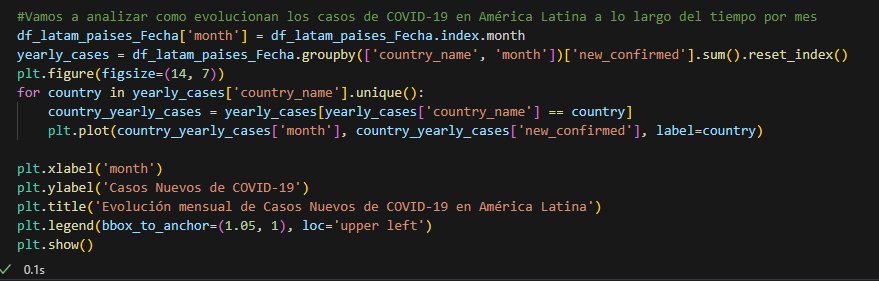
**Avance 3**

Evolución semanal de nuevos casos de COVID-19





Evolución mensual de nuevos casos de COVID-19



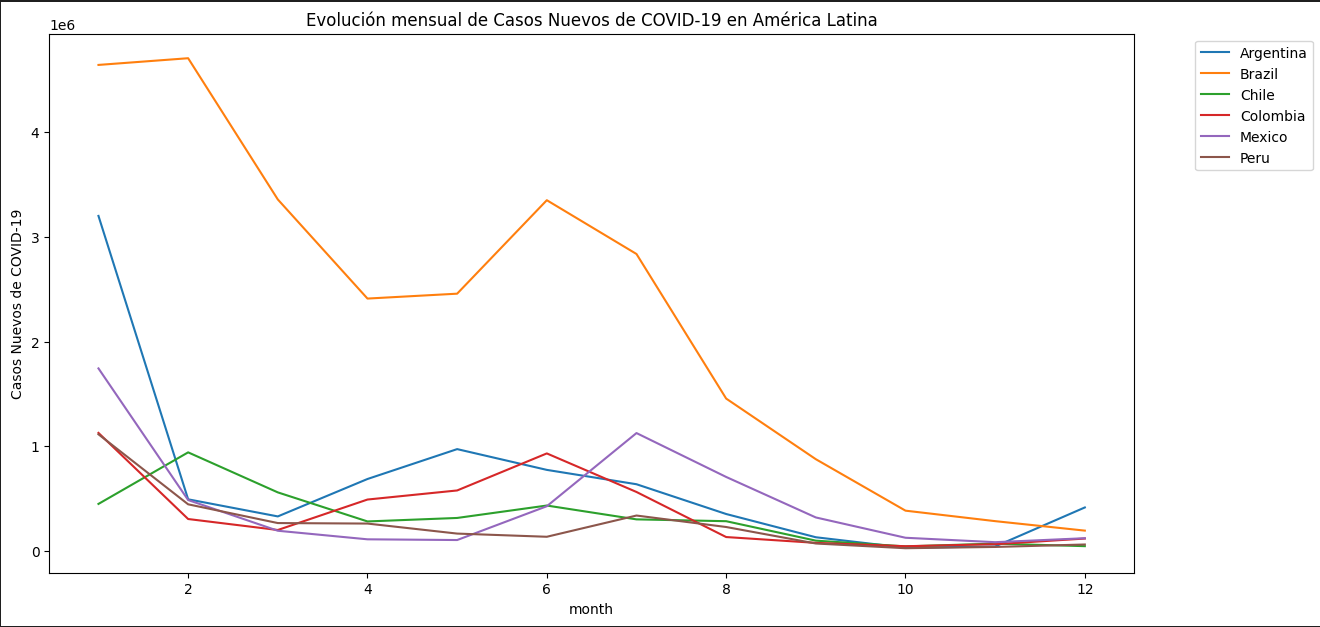


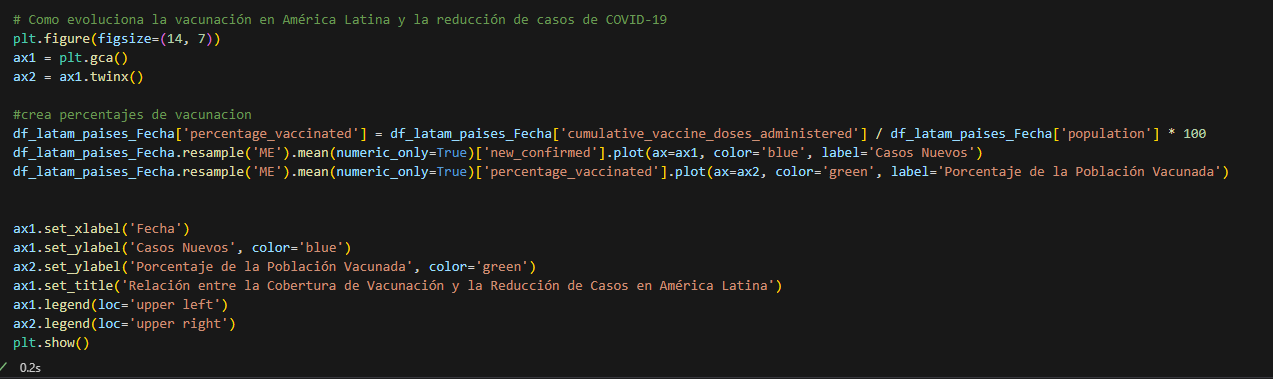
Gráfico de nuevos casos y temperatura promedio en el tiempo

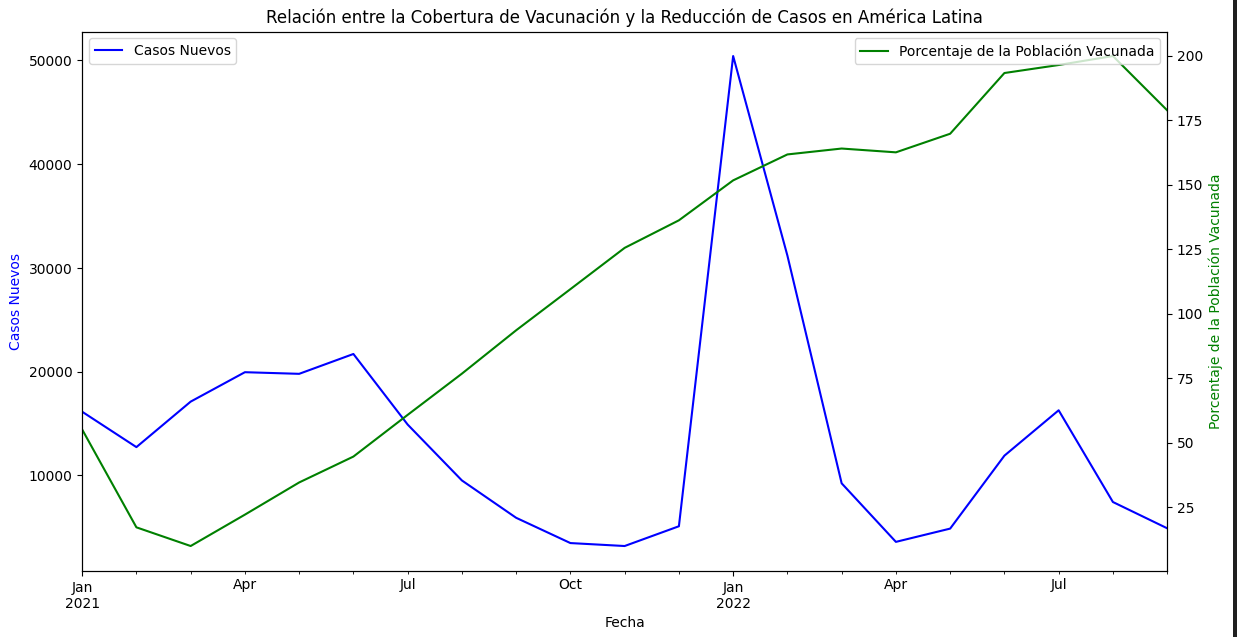
# 

# Comparación de estrategias de vacunación por países

# Comparación de la tasa de mortalidad por países.

Relación cobertura de vacunación y nuevos casos





# 

# EDA e insights

# **Avance 2**

De los gráficos de barras observados se puede ver que los países con más muertes son Brasil, México y Colombia a pesar de que Argentina presenta más nuevos casos que Colombia, la cantidad de muertes es inferior a la de Colombia, Mexico y Peru.

De las vacunas acumuladas y la población se puede decir que en Chile y Perú las dosis acumuladas superan la población lo cual significa que algunas personas ya recibieron una segunda dosis, y hasta una tercera en el caso de Chile, mientras en países como Colombia, Argentina y México aún no se cubre toda la población con la primera dosis. En Brasil la cantidad de vacunas aplicadas es aproximadamente igual a la población lo que indica que la primera dosis está probablemente cubierta.

Otro caso que llama la atención es ver como la tasa de mortalidad masculina supera a la femenina.

En relación al número de médicos por cada 1000 habitantes Argentina tiene la delantera casi doblando a los demas paises seguido de Chile, los demas paises presentan un número más bien parejo en este dato, en cuanto a las enfermeras por cada 1000 habitantes Chile tiene la ventaja seguido de Brasil esta puede ser en parte la razón de porqué Chile y Argentina presentan el menor número de muertes.

En el gráfico de nuevos confirmados por semana se puede ver un pico cercano a febrero de 2022 lo cual es interesante pero no hay una explicación clara hasta ahora del porqué de este caso.

**Avance 3**

Los gráficos de evolución de nuevos casos tanto semanal como mensual a pesar de no tener la misma forma tal cual si muestran una gran similitud, es decir se notan tendencias de aumento y decremento en los mismos periodos. Lo que sí se puede concluir de las gráficas es la disminución de casos nuevos en el tiempo al final del periodo de la muestra.

En el gráfico de evolución de nuevos casos y temperatura promedio se puede observar la coincidencia de sus máximos en la gráfica, lo que lleva a pensar que si hay una relación estrecha entre estas dos variables, además estos ocurren justamente cerca de febrero de 2022 lo cual puede ser la explicación del pico mencionado en el avance 2.

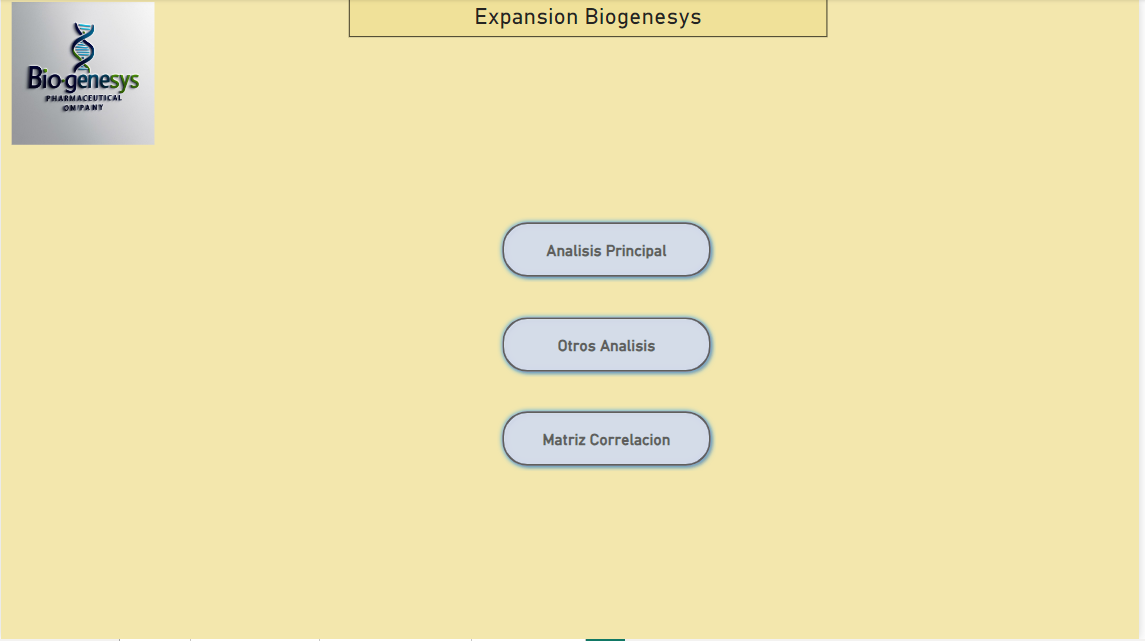
El análisis de estrategias de vacunación confirma la apreciación realizada del avance 2 observando que Chile ha superado la aplicación de la tercera dosis seguida de Perú que ha cubierto la segunda y va en camino a cubrir hasta la tercera, luego esta Argentina también en la segunda dosis seguida de México, Colombia y Brasil. La diferencia en el orden de los países entre los análisis del segundo y tercer avance se debe muy probablemente al reemplazo de datos nulos por la media en el acumulado de vacunas ya que este es un valor que se acumula y debe ser siempre creciente en cambio en este caso se usa el máximo lo cual debe ser el valor acertado o al menos un valor más cercano al correcto. Esta situación evidencia el cuidado que se debe tener al tratar con valores nulos.

Al comparar la tasa de mortalidad se aprecia que Perú tiene la tasa más alta seguido de México, Colombia, Brasil, Chile y Argentina, la explicación más razonable para esto puede ser que la relación de médicos y enfermeras por cada 1000 habitantes es mayor en Argentina y Chile lo cual ayudó a controlar la situación. Esta inferencia se refuerza al ver que la mortalidad más alta coincide con la tasa más baja de médicos por 1000 habitantes y es en Perú.

En el último gráfico de este avance de relación entre tasa de vacunación contra nuevos casos se puede observar como los nuevos casos se reducen a lo largo del tiempo aunque existe un pico del cual se habló antes y que puede estar relacionado con un incremento de la temperatura en ese momento.

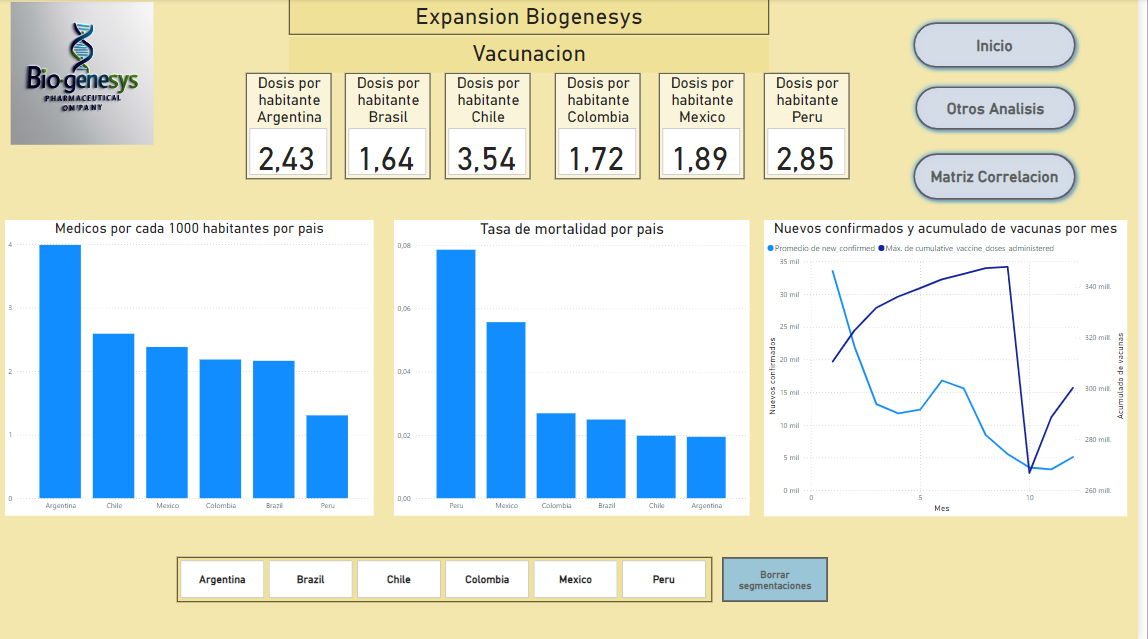
# Análisis del dashboard

La primera diapositiva del dashboard corresponde a un menú de inicio para navegar por las diferentes páginas del mismo con ayuda de los botones.

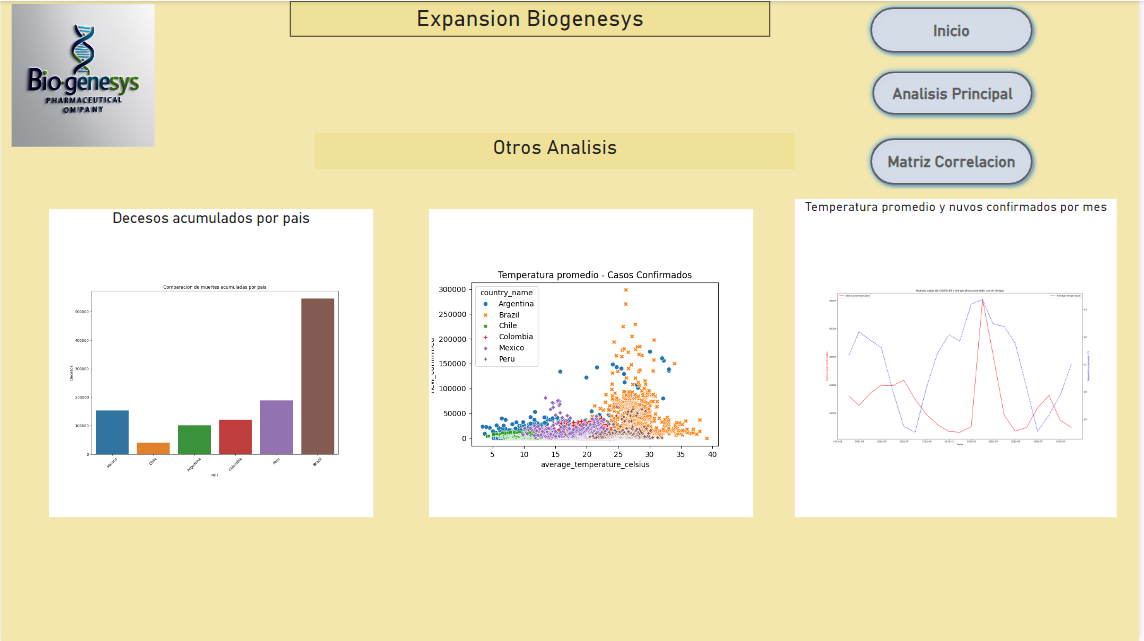


En la segunda página se realiza el análisis principal considerando el objetivo del proyecto, en esta se presentan tarjetas con la información de vacunación y tres gráficos con el fin de remarcar el efecto de la tasa de médicos por cada 1000 habitantes y se compara con el impacto que genera la tasa de mortalidad por país se cuenta con un tercer gráfico de líneas que muestra el efecto de las vacunas con una disminución de los nuevos casos como una evolución mensual.

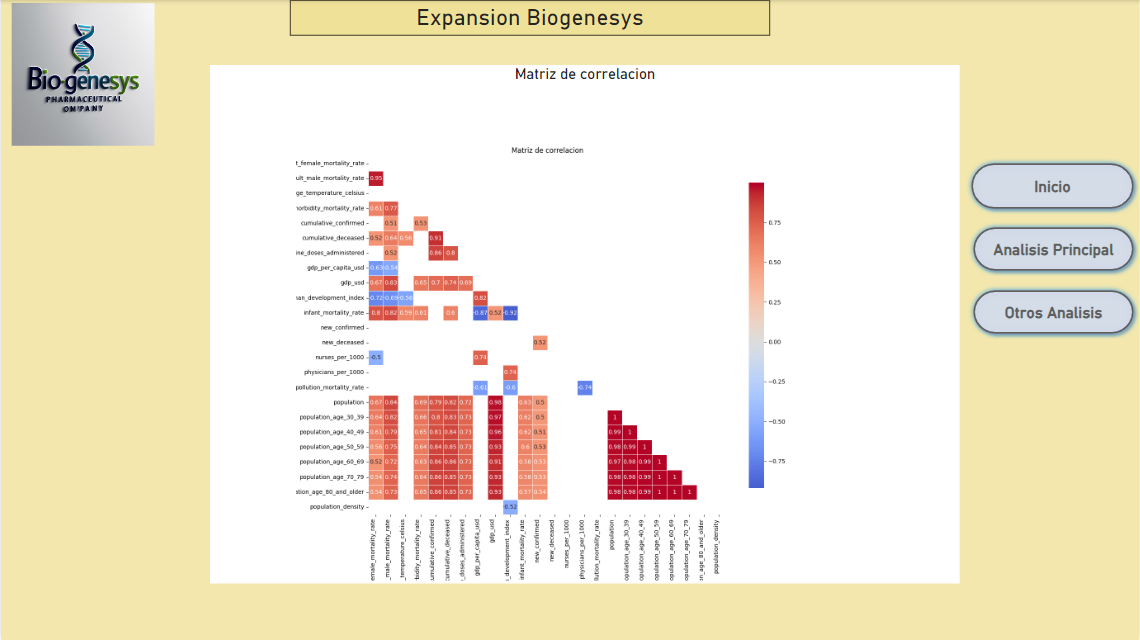
Se cuenta con segmentación por país y botones de navegación.



En la tercera página se incluyen gráficos generados con Python con el fin de analizar la tasa de muertes por país y la relación entre temperatura promedio y nuevos casos mediante dos tipos de gráfico.



Por último se encuentra una página con un solo gráfico debido al tamaño, es un mapa de calor (heatmap), que muestra una matriz de correlación de varias variables numéricas del dataset. Todas las páginas cuentan con botones de navegación.



# Conclusiones y Recomendaciones

En general las observaciones y análisis realizados llevan a inferir los siguiente:

Los países con mejor manejo de la situación tanto por estrategia de vacunación como por menor cantidad de decesos son Chile y Argentina, no es de extrañar que estos mismos dos países presentan los mayores indices de medicos y enfermeras por cada mil habitantes con esta claridad sobra decir que el impacto de nuevos laboratorios en dichos países muy probablemente no tengan el mismo impacto que en los otros cuatro.

De acuerdo a lo anterior los países candidatos a expansión serían entonces Brasil, Colombia, México y Perú; sin embargo en el caso de Perú la campaña de vacunación ha sido muy buena por lo que probablemente la necesidad de apoyo no es tan alta. Dicho esto nos quedan como principales candidatos Brasil, Colombia y México.

Dependiendo de la cantidad de laboratorios que la empresa esté dispuesta a abrir entendiendo que no pueden ser seis o de lo contrario no se haría este análisis se organizan en orden prioritario los tres países restantes así:

primero Brasil, segundo México y tercero Colombia y el argumento principal es la cantidad de decesos en cada uno de estos países ya que el principal objetivo debe ser que los laboratorios ayuden a evitar la pérdida de más vidas.

# Reflexión personal

# La realización de este proyecto ha sido bastante enriquecedora como futuro analista de datos ya que enfrenta al estudiante a una serie de situaciones que lo incitan a desarrollar y reforzar habilidades clave en el desempeño de la profesión.

# 