**Expansión Estratégica de Biogenesys con Python**

Nombre del autor: Nahuel Undabarrena

Email: [nahuelunda@gmail.com](mailto:nahuelunda@gmail.com)

Cohorte: DA-FT12

Fecha de entrega: 07/04/2025

Institución: Biogenesis



**Introducción:**

La empresa farmacéutica Biogenesisbusca 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.

Biogenesis me ha contratado como Data Analyst para que realice un estudio que ayudará 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.

**Desarrollo del proyecto:**

*Importacion de los datos*

Lo primero que se realizó en el proyecto fue la importación del conjunto de datos utilizando Python desde Visual Studio Code (VSCode), lo cual permitió trabajar de manera flexible y eficiente con los datos. Previo a la carga del archivo, se importaron las principales librerías necesarias para el análisis: *Pandas* para la manipulación de estructuras de datos, *NumPy* para operaciones numéricas, y las bibliotecas de visualización *Seaborn* y *Matplotlib.pyplot* para la creación de gráficos y análisis exploratorio. Esta configuración inicial sentó las bases para todo el procesamiento y visualización de los datos realizados posteriormente.

*Limpieza de datos:*

Una vez importado el conjunto de datos en Python, el primer paso fue realizar un proceso exhaustivo de limpieza para asegurar la calidad del análisis posterior. La limpieza comenzó con una revisión general de la tabla, especialmente del listado de países. Esto se debió a que el análisis estaba centrado únicamente en un grupo específico de países, por lo que era necesario filtrar aquellos que no eran relevantes para los objetivos del estudio. Para ello, se generó una máscara basada en una lista previamente definida con los códigos de los países que se queria nconservar. Esta lista fue utilizada para conservar únicamente los registros correspondientes a dichos países.

Posteriormente, se procedió a trabajar con la columna *date*, que contenía la información temporal del dataset. Inicialmente, esta columna estaba en formato string, por lo que se aplicó un casteo para convertirla al tipo datetime. Una vez hecho esto, se generó una nueva máscara temporal para conservar únicamente los datos dentro de un rango de fechas específico, correspondiente al período relevante para el análisis del impacto del COVID-19 (2021 y 2022). Finalmente, se aplicaron ambas máscaras de forma simultánea, quedándonos con un subconjunto de datos más limpio, manejable y centrado en los objetivos del proyecto.

En una segunda etapa de limpieza, se intentó eliminar columnas y filas que tengan únicamente valores nulos, pero se verificó que no existían columnas o filas completamente vacías. A pesar de ello, se observó una gran cantidad de valores nulos distribuidos en múltiples columnas, lo cual requería una atención más específica. En particular, se identificaron muchas filas que correspondían a subregiones o subdivisiones administrativas dentro de los países, las cuales no contenían información relevante (en su mayoría estaban vacías) y, además, generaban ruido para el análisis agregado por país. Para resolver esto, se comparó la columna *location\_key* con la lista de códigos de países ya definidos (*countrieCodes*), y se filtraron aquellas filas en que el location\_key no coincidía con ninguno de estos códigos, eliminando así registros que no aportaban al análisis.

Luego de aplicar este filtro, el dataset quedó reducido a 3744 registros y 50 columnas, lo que representa una base de datos mucho más limpia y adecuada para el análisis. A partir de ahí, se detectó que la columna location\_key era ahora completamente redundante, ya que su contenido era idéntico al de la columna *country\_code*, por lo que se procedió a eliminarla para evitar duplicidad de información.

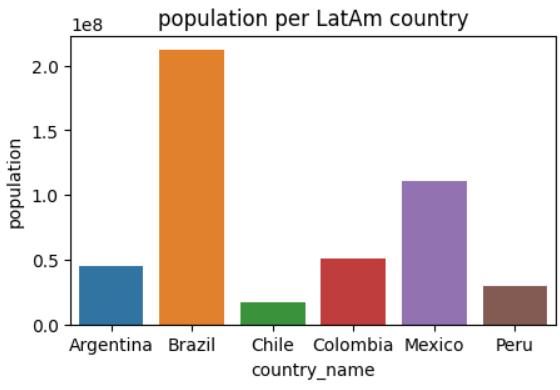
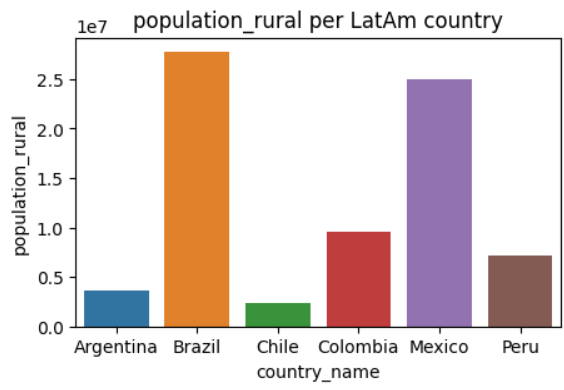
Para continuar con la imputación de valores faltantes, se calcularon las medianas de algunas columnas clave relacionadas con la evolución de la pandemia: 'new\_confirmed', 'new\_deceased', 'cumulative\_confirmed', 'cumulative\_deceased' y 'cumulative\_vaccine\_doses\_administered'. Estas medianas se utilizaron para reemplazar los valores nulos de dichas columnas, ya que la mediana es una medida robusta frente a valores extremos y proporciona una imputación razonable para datos faltantes en series temporales epidemiológicas.

En cuanto a las columnas relacionadas al clima y ambiente (como temperatura, humedad, lluvia, etc.), se optó por una estrategia distinta: los valores nulos fueron rellenados con los valores del día anterior utilizando el método de propagación hacia adelante (ffill), lo que permite mantener la coherencia temporal de estas variables sin introducir sesgos arbitrarios. En la columna new\_recovered, que representa los nuevos recuperados por día, se asumió que los valores nulos correspondían a cero recuperaciones reportadas ese día, y por ende, fueron reemplazados directamente por 0.

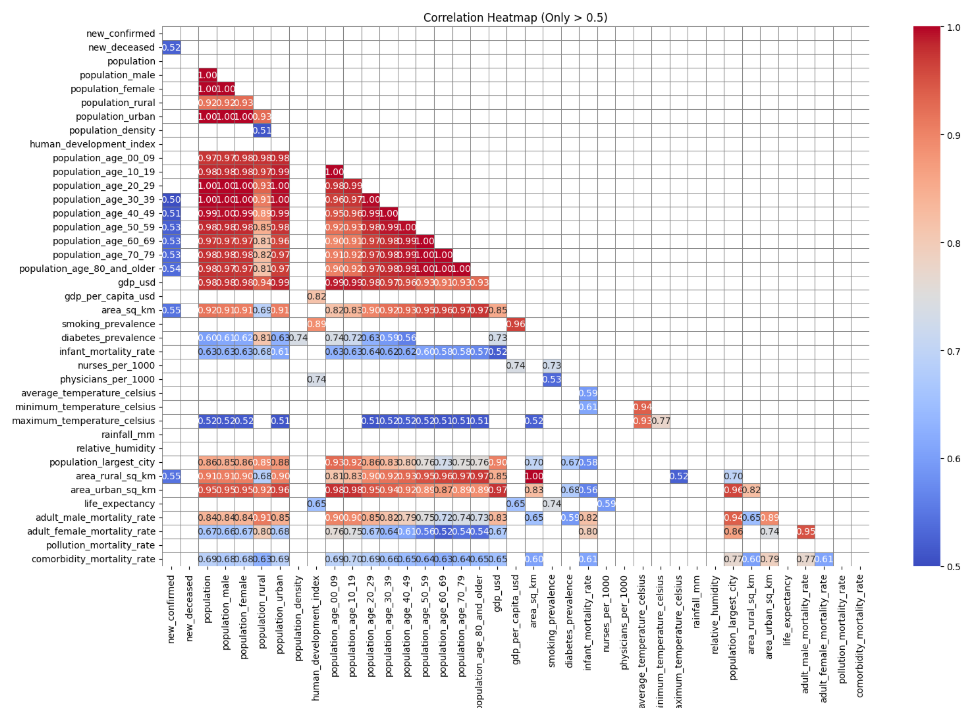
Finalmente, como parte de una optimización técnica del dataset, se realizó un cambio en los tipos de dato numéricos, pasando de float64 a float32. Esta conversión permitió reducir significativamente el consumo de memoria sin comprometer la precisión requerida para el análisis, especialmente útil cuando se trabaja con grandes volúmenes de datos y múltiples operaciones analíticas.

*Creando gráficos en Python:*

Antes de comenzar con la creación de visualizaciones, fue necesario realizar un análisis estadístico preliminar sobre las variables numéricas del conjunto de datos. Para ello, se implementaron funciones personalizadas en Python que permitieran calcular estadísticas descriptivas como la mediana, la varianza y el rango de las columnas seleccionadas.

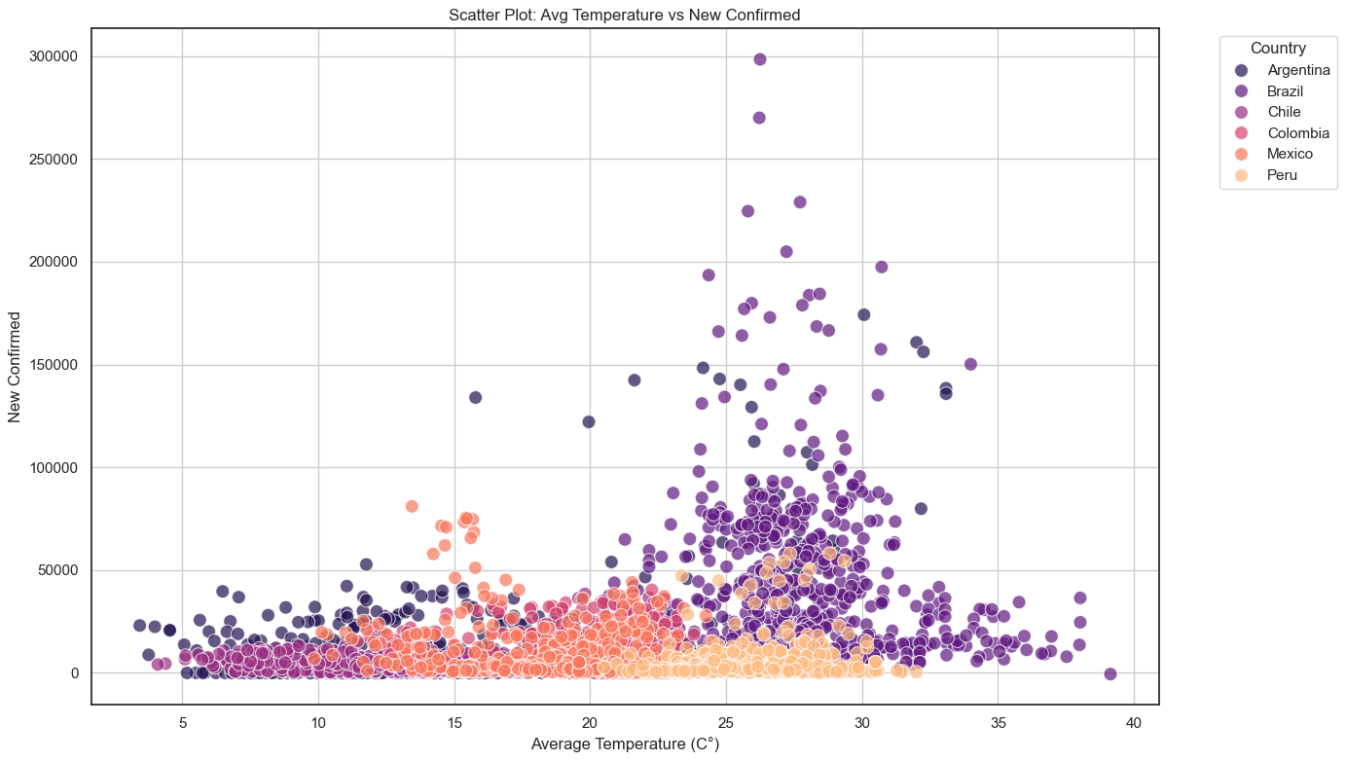
A partir de ahí, se dio inicio a la etapa de visualización de los datos, con el objetivo de interpretar de forma más intuitiva las relaciones y tendencias contenidas en el dataset. Como primer paso, se elaboraron una serie de gráficos de barras por país, iterando sobre todas las columnas numéricas mediante un ciclo for. Esto permitió generar automáticamente múltiples gráficos, uno para cada variable, facilitando así la identificación de diferencias sustanciales entre países en aspectos como casos confirmados, muertes, vacunación, recuperación, etc. Gracias a esta visualización, se pudo detectar rápidamente, por ejemplo, que Brasil destacaba con números significativamente más altos en casos confirmados, muertes y recuperados, lo cual se explica por su alta densidad de población. También se observó un dato curioso: México, pese a tener menos población total que Brasil, presentaba un nivel casi equivalente de población rural, lo que dio lugar a reflexiones sobre la distribución geográfica de los habitantes y su impacto en la gestión de la pandemia.******

Posteriormente, se generó un heatmap (mapa de calor) para observar la correlación entre todas las variables numéricas del conjunto de datos. Este gráfico permitió identificar qué variables se movían en conjunto y cuáles no tenían relación entre sí. **Si bien visualmente puede parecer denso, este heatmap fue útil para destacar relaciones como la baja correlación entre casos confirmados y fallecimientos, lo cual valida algunas intuiciones previas y ayuda a dirigir el enfoque del análisis a las variables con mayor influencia mutua.**

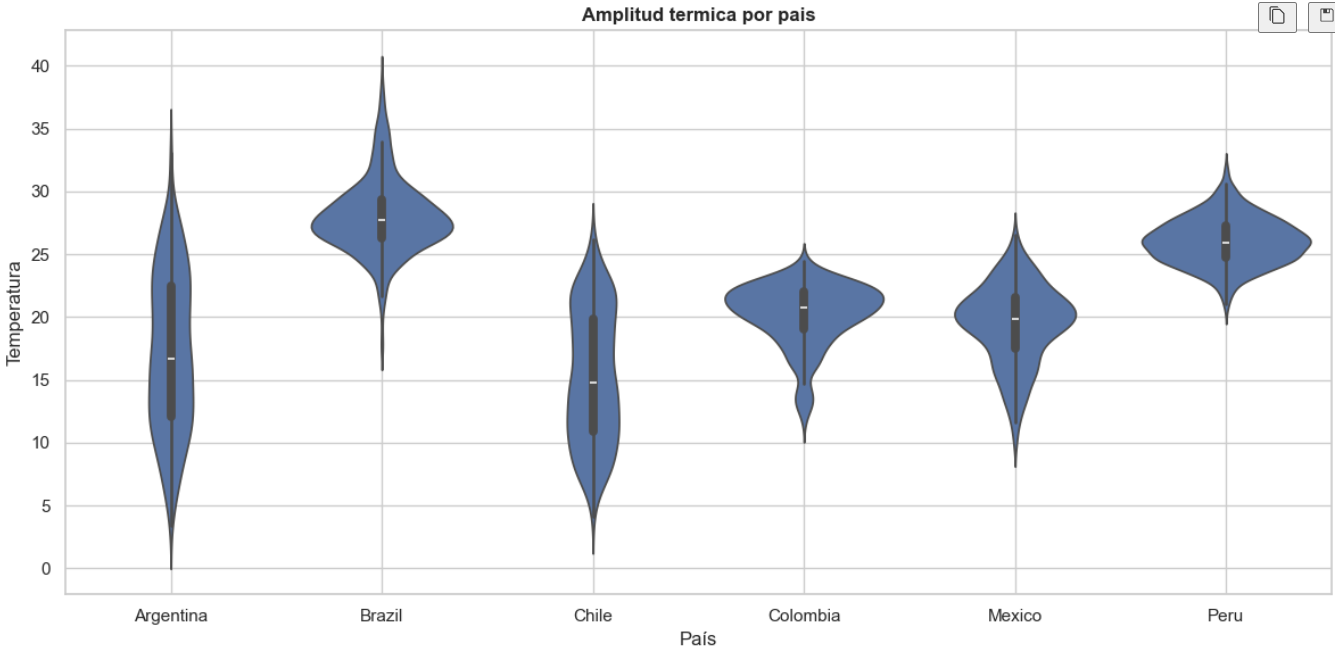
****

Durante la exploración, también se identificaron las columnas cuyo contenido variaba a lo largo del tiempo, es decir, aquellas con valores dinámicos o cambiantes. Estas columnas fueron almacenadas en una lista llamada *changingValuesColumns*, con la intención de graficar histogramas que permitieran ver la distribución de sus valores. Tras graficados los mismos se pudo tener una mejor visualización de la distribución de los datos.

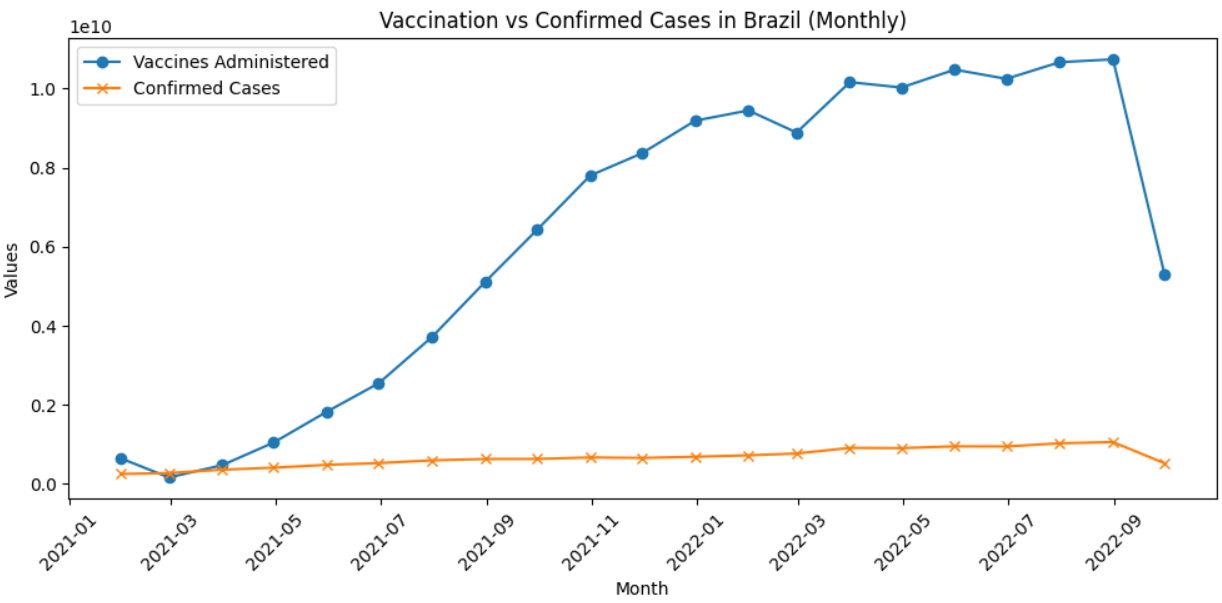
Más adelante, se elaboraron una serie de gráficos para investigar la relación entre variables climáticas y los impactos de la pandemia. Se pusieron en relación las muertes confirmadas y los casos confirmados de COVID-19 con la temperatura media de cada país. Esto permitió explorar visualmente si existía alguna relación entre el clima y la propagación o letalidad del virus.



En línea con estos análisis, se hizo un gráfico de violines para mostrar la amplitud térmica de cada país y añadir contexto. Asimismo se diseñaron también gráficos de líneas para comparar la evolución de los casos confirmados frente a los fallecimientos por país. Estas visualizaciones permitieron ver claramente la tendencia temporal y la distancia entre ambas curvas, lo que a su vez reveló la eficacia (o falta de ella) de las medidas sanitarias adoptadas en ciertos momentos.



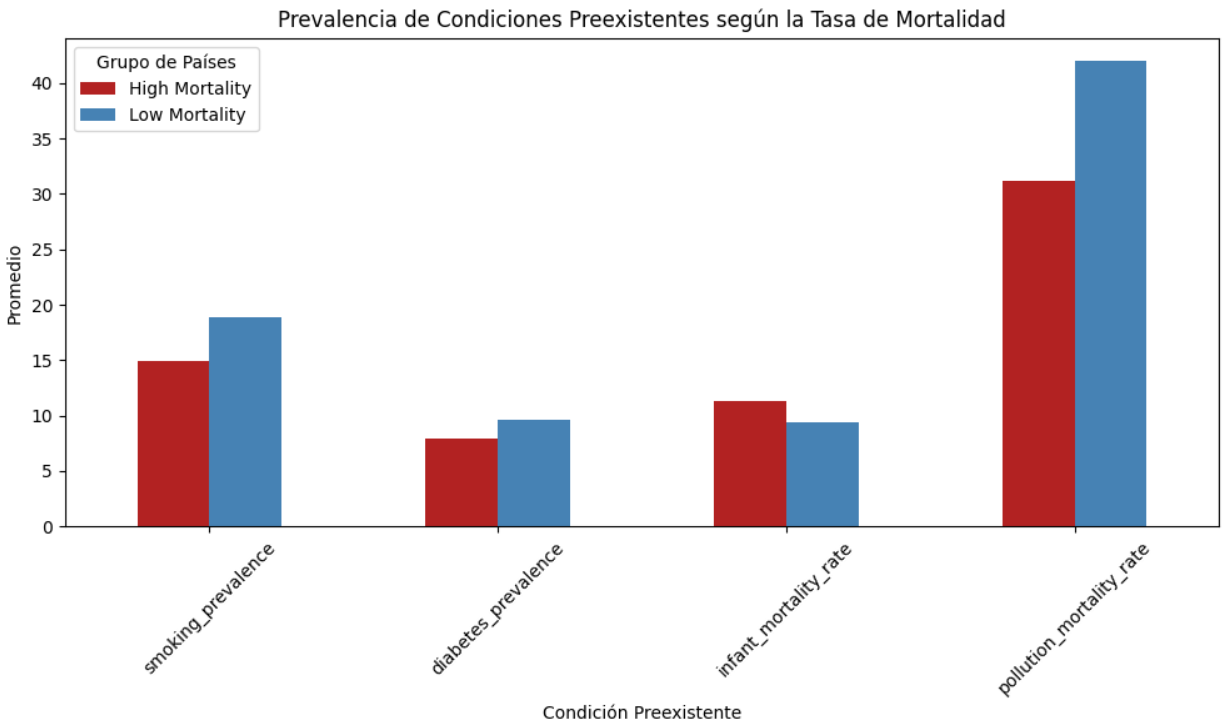
Complementando lo anterior, se elaboró un gráfico de líneas que mostraba, por país, la evolución conjunta de los casos confirmados y las vacunas administradas, con el objetivo de detectar si existía una sincronización o retraso entre los picos de casos y el aumento en la vacunación.



A su vez, se volvieron a utilizar gráficos de dispersión, pero esta vez incorporando la dimensión temporal para mostrar la progresión en la relación entre vacunación y casos confirmados a lo largo del tiempo. Esta visualización fue particularmente valiosa para observar cómo cambiaba la relación entre ambas variables conforme avanzaba la pandemia y se intensificaban las campañas de vacunación.

Más tarde, se abordó el análisis demográfico mediante un gráfico de barras que representaba la distribución de la población por edad en cada país. Esta visualización permitió observar de forma clara qué países tienen una población más envejecida, lo cual puede ser un factor de riesgo en situaciones sanitarias críticas como una pandemia global.

Una parte interesante del análisis fue la visualización de condiciones preexistentes (como diabetes, por ejemplo) y su relación con la tasa de mortalidad. Para esto, se categorizaron los países en dos grupos: “high mortality” y “low mortality”, basándose en la cantidad de muertes asociadas a cada condición. Posteriormente, se calcularon promedios de prevalencia de condiciones preexistentes dentro de cada grupo, lo cual permitió observar qué condiciones de salud eran más comunes en los países con mayor tasa de letalidad. Aunque el gráfico puede parecer complejo, resultó útil para identificar posibles factores de riesgo estructurales.



Además, se estudió el impacto de la urbanización en la propagación del virus. Para ello, se construyó una visualización que ponía en relación el porcentaje de población urbana con la cantidad de casos confirmados, explorando la hipótesis de que la densidad poblacional urbana puede haber sido un factor de aceleración en el contagio.

Finalmente, se realizaron análisis temporales de mortalidad por COVID-19 desglosados por país, con visualizaciones que permitieron estudiar la evolución de las muertes a lo largo del tiempo, facilitando comparaciones directas entre países y revelando patrones estacionales o estructurales en la gestión de la pandemia.

*Creacion de Dashboard en PowerBI:*

Antes de comenzar con la creación de los dashboards en Power BI, fue necesario realizar una última etapa de preparación en los datos: el ajuste de los tipos de dato de varias columnas usando Power Query. Este paso fue fundamental para asegurar que Power BI pudiera interpretar correctamente la información y generar visualizaciones precisas. Se revisaron y modificaron columnas que contenían fechas, valores numéricos y categóricos, garantizando que cada una estuviera en el formato adecuado para su análisis posterior.

Una vez finalizado el proceso de limpieza, análisis y visualización preliminar, se trasladó el trabajo a Power BI con el objetivo de construir un dashboard interactivo y comprensible que permitiera explorar los datos relevantes del impacto del COVID-19 desde diferentes perspectivas. Para organizar la información de manera clara y eficiente, se decidió estructurar el reporte con una pantalla principal (Home) y tres tableros adicionales, cada uno con un enfoque temático particular.

*HOME:*

La primera página del reporte cumple la función de portada y menú de navegación. Desde allí, los usuarios pueden acceder a los tres paneles informativos principales. Se incorporó un diseño limpio con títulos representativos y botones que facilitan la interacción y dirigen al usuario de forma intuitiva a las diferentes secciones del dashboard.

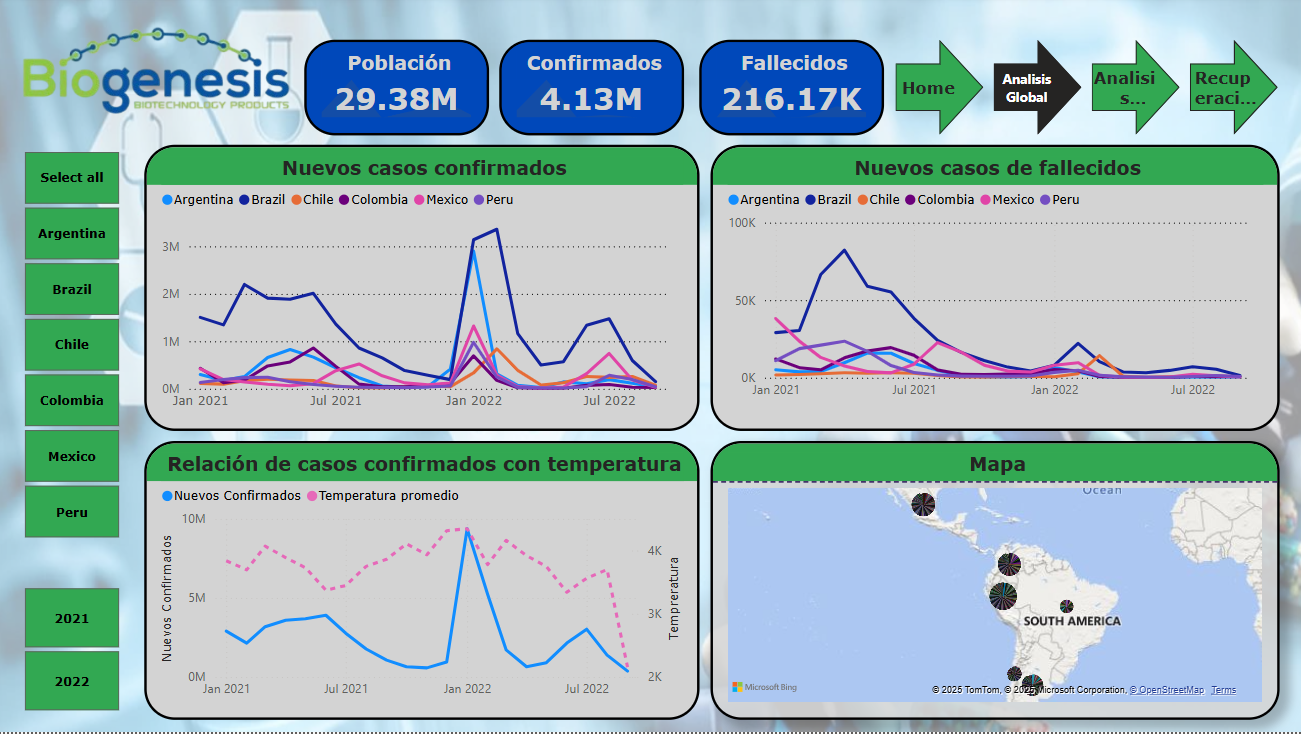


*ANÁLISIS GLOBAL:*

Este primer tablero ofrece una visión general de la evolución de la pandemia a nivel mundial. Entre las visualizaciones principales se incluyen:

* Un gráfico de líneas que muestra la evolución temporal de los casos confirmados.
* Otro gráfico de líneas para observar el comportamiento de los casos fallecidos a lo largo del tiempo.
* Un gráfico que explora la relación entre la temperatura promedio y los casos confirmados, permitiendo analizar si existen patrones entre el clima y la propagación del virus.
* Un mapa temático que facilita la visualización geográfica de la cantidad total de muertes por país.

Este panel también incorpora segmentadores por país y por año, lo que permite adaptar el análisis a distintas regiones y periodos temporales.

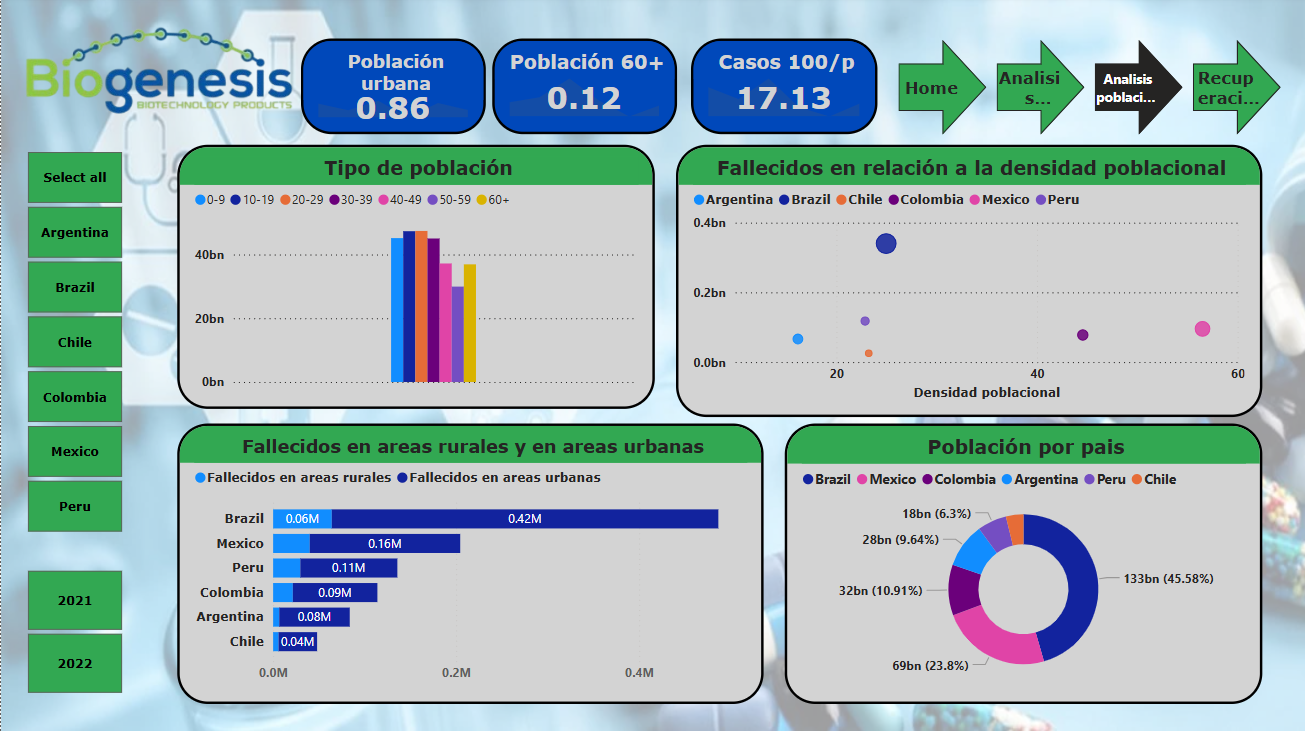


*ANÁLISIS POBLACIONAL:*

El segundo dashboard se enfoca en variables demográficas y urbanísticas para entender cómo distintos perfiles poblacionales se vieron afectados por la pandemia. En él se incluyen:

* Una tarjeta con la cantidad de población urbana.
* Visualizaciones que representan la distribución etaria de la población (por grupos de edad), lo que permite identificar posibles poblaciones de riesgo.
* Un gráfico que muestra la cantidad de casos confirmados cada 100 personas, facilitando comparaciones estandarizadas entre países.
* Una comparación entre la cantidad de fallecidos en zonas urbanas versus zonas rurales, brindando una perspectiva sobre cómo el entorno puede influir en los resultados sanitarios.
* Un scatter plot que muestra la relación entre la densidad poblacional y los fallecidos por Covid.
* Un gráfico que presenta la población total de cada país, como contexto general.

También en este tablero se incorporaron segmentadores interactivos por país y año.

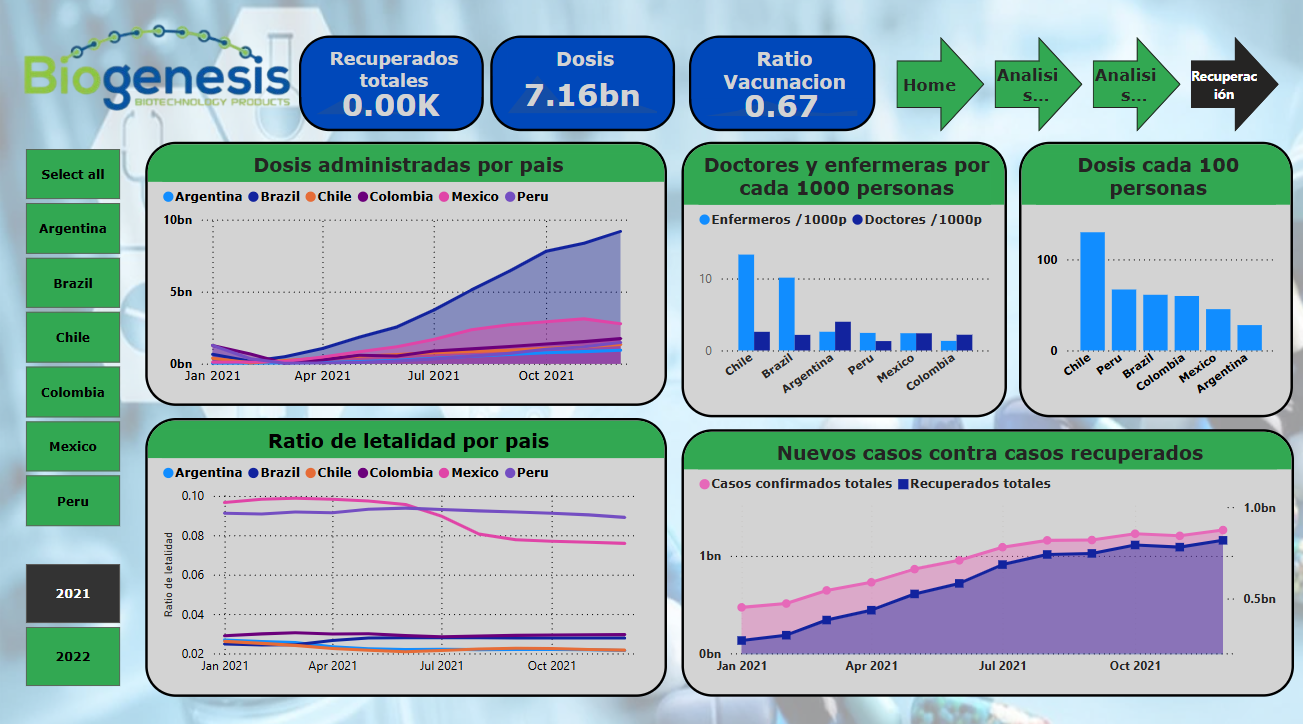


*RECUPERACIÓN:*

El tercer y último tablero está centrado en el análisis de recuperación, vacunación y capacidad sanitaria. Aquí se presentan indicadores clave relacionados con los esfuerzos por contener la pandemia y atender a los pacientes. Entre las visualizaciones se incluyen:

* Gráficos de barras que muestran la cantidad de dosis administradas cada 100 personas, país por país, y una tarjeta mostrando las dosis totales entregadas.
* Una visualización que refleja la cantidad de doctores y enfermeros cada 1000 habitantes, como medida de la capacidad del sistema de salud.
* Cálculo del ratio de vacunación alcanzado por cada país.
* Evolución mensual del ratio de letalidad del COVID-19, es decir, el porcentaje de casos confirmados que resultaron en fallecimientos.
* Gráficos de líneas que representan tanto la cantidad de dosis administradas mes a mes como la comparación entre nuevos casos y recuperaciones, permitiendo observar el avance en la superación de la enfermedad.

Como en los paneles anteriores, se incorporaron segmentadores por país y por año para filtrar la información y personalizar el análisis.



**Análisis y conclusión:**

Durante el análisis se identificó una relación clara entre la temperatura y la cantidad de casos confirmados de COVID-19. En muchos países, los picos de contagio coincidieron con aumentos significativos de la temperatura promedio. Esta correlación puede responder a múltiples factores, como cambios en el comportamiento social durante climas más cálidos, una mayor movilidad o relajación de medidas sanitarias en temporadas de calor, o incluso posibles implicancias biológicas y ambientales en la propagación del virus. Si bien la relación no implica causalidad directa, sí establece un patrón muy relevante y significativo que no se puede pasar por alto.

Otro hallazgo relevante fue la diferencia entre muertes registradas en áreas urbanas frente a áreas rurales. A pesar de que la concentración poblacional en zonas urbanas podría justificar esta diferencia en términos absolutos, se observó que, incluso en países con gran proporción de población rural, como Colombia, las muertes en zonas urbanas seguían siendo desproporcionadamente más altas. La densidad poblacional no explica por sí sola este fenómeno, ya que muchos de los países analizados presentan una población distribuida de forma relativamente equitativa. Esto sugiere que otros factores asociados a la vida urbana, como la velocidad de propagación, el acceso desigual a servicios de salud o la mayor exposición, pudieron tener un rol determinante en el impacto de la pandemia.

En cuanto al personal sanitario, Perú y Mexico aparecen como unos de los países con menor cantidad de enfermeros y médicos por cada mil habitantes. Este déficit podría estar directamente relacionado con la capacidad de respuesta de sus sistemas sanitarios. Sin embargo, se observó una diferencia significativa en el manejo de la situación entre ambos países. Mexico logró una respuesta temprana con una administración eficiente de vacunas, lo que le permitió controlar la situación y reducir la letalidad en menos tiempo. Perú, en cambio, inició su campaña de vacunación con un retraso de aproximadamente tres meses respecto al promedio del grupo de países analizados. Esta demora tuvo consecuencias: la tasa de letalidad en Perú fue más alta y su descenso fue más lento. A pesar de haber sido uno de los países que más dosis administró en términos totales, la demora inicial parece haber condicionado negativamente los resultados. A esto se suma su estructura demográfica, con alta proporción de personas menores de 20 años y mayores de 60, los dos grupos más vulnerables a los efectos de la pandemia.

Chile representa el caso opuesto: con una población relativamente envejecida —especialmente en el grupo de mayores de 60— logró disminuir rápidamente la letalidad del virus. Este éxito se vincula a su alta densidad de personal médico, que le permitió implementar estrategias de contención y atención más efectivas. Brasil, por su parte, aunque en números totales aparezca siempre en los primeros puestos debido a su tamaño poblacional (tiene aproximadamente el 45% de toda la población considerada en este análisis), mostró una capacidad de acción rápida y eficaz. Su respuesta, en términos relativos, fue de las más destacadas, lo cual se refleja en la evolución de sus indicadores.

Otro dato que juega a favor de Brasil es que pese a haber sido el país que más casos confirmados tuvo, tuvo menos casos cada 100 habitantes que Chile y Argentina. Siendo el de estos últimos 34.73 y 28.62 respectivamente y 20.26 el de Brasil.

Las campañas de vacunación fueron muy efectivas y pese a que en el 2022 hubo una mayor cantidad de casos confirmados, el ratio de letalidad en todos los países fue mucho mas bajo.

Finalmente, si se tuviera que recomendar una expansión del mercado de Biogénesis, los datos sugieren que el lugar con mayor necesidad estratégica sería la zona urbana de Perú. Este país no solo enfrenta limitaciones en personal sanitario, sino también una población especialmente vulnerable por su composición etaria. La alta mortalidad registrada y la respuesta tardía ante la crisis revelan un sistema de salud que podría beneficiarse significativamente con apoyo adicional, convirtiéndolo en un punto crítico para fortalecer la capacidad de respuesta ante futuras situaciones similares.