**[Título del proyecto]**

**Nombre del autor: Gerardo Gabriel Morsucci Riquelme**

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

**Cohorte: DA-PT01**

**Fecha de entrega:**29/7/2024

**Institución: BIOGENESYS LABORATORIOS**

****

# Introducción

# El proyecto "Expansión Estratégica de Biogenesys con Python" tiene como propósito identificar las ubicaciones óptimas para la expansión de laboratorios farmacéuticos en Latinoamérica, específicamente en Colombia, Argentina, Chile, México, Perú y Brasil. Este análisis se basa en datos de incidencia de COVID-19, tasas de vacunación y disponibilidad de infraestructuras sanitarias. Los objetivos organizacionales alcanzados incluyen la realización de un análisis exploratorio de datos, la aplicación de técnicas de limpieza de datos, la mejora del acceso a los datos mediante operaciones ETL, y el desarrollo de dashboards interactivos. Estos esfuerzos están dirigidos a optimizar la respuesta a la pandemia y postpandemia, mejorando el acceso a las vacunas y facilitando una toma de decisiones informada y estratégica.

# 

# Desarrollo del proyecto

### Metodología de Recopilación y Selección de Datos

En el desarrollo del proyecto, se siguió una metodología rigurosa para la recopilación y selección de datos. A continuación, se detallan los pasos seguidos:

1. **Revisión Inicial**: Se comenzó corroborando el archivo readme.txt proporcionado por el equipo de ingeniería de datos, con el objetivo de entender el enfoque y la naturaleza de cada columna del dataset.
2. **Verificación de la Estructura del Dataset**: Se comprobó que el dataset tuviera la cantidad especificada de registros y columnas, que eran 12,216,057 y 50, respectivamente.
3. **Filtrado Inicial:** Se filtraron los datos para incluir únicamente los países especificados en el análisis: Argentina, Brasil, Chile, Colombia, México y Perú.
4. **Identificación de valores faltantes**: Se estableció un criterio para identificar columnas con una cantidad de valores faltantes mayor a 4 millones, resultando en 26 columnas con dicha cantidad de valores.
5. **Filtro por Clave de Ubicación**: Se utilizó la columna location\_key, que contiene las siglas ISO de los países, para filtrar los datos, ya que esta columna es fundamental para el dashboard.
6. **Filtrado por Fecha:** Se estableció como parámetro que las fechas en la columna date comenzarán a partir del 01-01-2021, dado que los datos relevantes empiezan en esa fecha.
7. **Tratamiento de Valores Faltantes:** Se verificaron nuevamente los valores faltantes. El dataset se redujo a 3,744 registros y 50 columnas. A continuación, se abordaron las columnas con datos faltantes:
   * 9 columnas con menos de 100 valores faltantes se rellenaron con la media de cada columna.
   * Para las otras 3 columnas con 584, 2,116 y 2,736 valores faltantes, se decidió rellenar los datos faltantes con 0.

### Conclusión sobre Transformaciones y Limpieza de Datos:

Tras aplicar los filtros y tratamientos mencionados, el dataset final quedó reducido a 3,744 registros y 50 columnas, manteniendo su integridad estructural. Las transformaciones realizadas, como el relleno de valores faltantes y el seteo de parámetros, fueron esenciales para asegurar la calidad y consistencia de los datos, permitiendo un análisis confiable y fundamentado. Estas medidas garantizaron que los datos utilizados fueran representativos y aptos para el análisis exploratorio y la visualización interactiva en los dashboards, facilitando una toma de decisiones informada para la expansión estratégica de Biogenesys.

# EDA e insights

### Insights Obtenidos en el Análisis Exploratorio de Datos y Visualizaciones

#### Preparación de los Datos

Como primera instancia, se hizo un parseo de la columna date para colocarla como índice. Luego, se filtraron las fechas para que abarcaran únicamente el año 2021. Esto se hizo con el fin de realizar un análisis de series de tiempo, para lo cual fue necesario crear las columnas week y month.

#### Análisis de Series de Tiempo

Una vez realizada la preparación, se observaron varios patrones en los gráficos de línea propuestos. En el gráfico semanal, se identificó que Brasil fue el país más afectado en términos de casos de COVID-19. No debemos olvidar considerar su densidad poblacional al interpretar estos datos. También se observó que Argentina tuvo un pico más grande que los demás países en las últimas semanas del año.

#### Tasa de Crecimiento de Casos

En un gráfico de tasa de crecimiento (%) de casos confirmados de COVID-19, se observó un pico significativo entre enero y febrero de 2022. Para el inicio de marzo, la tasa de crecimiento había vuelto a índices normales.

#### Vacunación

Se analizó la cantidad de vacunas colocadas a través de un gráfico de barras, donde Brasil aparecía como el líder claro. Sin embargo, un análisis más detallado que incluyó la densidad poblacional reveló que Brasil era el país que menos dosis por habitante había colocado, con alrededor de 1.65 dosis por habitante. Chile fue el líder en este aspecto, con alrededor de 3.6 dosis por habitante.

#### Distribución Poblacional

A través de un gráfico de torta poblacional, se identificó a México como el país con la natalidad más elevada, con casi un 20% de los habitantes teniendo entre 0 y 9 años.

#### Correlación entre Temperatura y Casos de COVID-19

Se observó la relación entre la temperatura de cada país y el incremento de casos mediante un scatterplot. Inicialmente, se identificaron outliers en Chile, los cuales fueron eliminados del dataset. Brasil mostró la mayor correlación entre aumento de temperatura y aumento de casos nuevos, aunque esta correlación no era lo suficientemente fuerte como para ser decisiva. Chile y Argentina presentaron temperaturas más bajas, manteniéndose estables en términos de casos nuevos.

#### Cobertura de Vacunas y Reducción de Casos

Un análisis a través de un gráfico de líneas con dos ejes Y mostró una alta correlación entre la cobertura de vacunas y la reducción de casos nuevos. Al llegar al pico de vacunación, los casos disminuyeron notablemente.

#### Enfermedades Crónicas

Se realizó un análisis de enfermedades crónicas producidas por el tabaco y la diabetes. Los gráficos de barras mostraron que la diabetes tiene un impacto significativamente mayor en México y Brasil. En cuanto al tabaquismo, Chile y Argentina son los más afectados.

#### Mortalidad y Letalidad

* Mortalidad por COVID-19: México mostró la mayor tasa de mortalidad, seguido por Brasil y Chile. Esta disparidad sugiere que puede ser complicado establecer una correlación clara entre las condiciones preexistentes y la mortalidad por COVID-19.
* Tasa de Letalidad: Todos los países mostraron un comportamiento similar en términos de tasa de letalidad, con una tendencia general a la disminución de la fatalidad al analizar los datos a lo largo del tiempo.

Estos insights fueron fundamentales para identificar las ubicaciones óptimas para la expansión de laboratorios farmacéuticos de Biogenesys, priorizando los países con mayores necesidades y desafíos en términos de incidencia de COVID-19, tasas de vacunación y recursos sanitarios.

# 

# Análisis del dashboard

# En el dashboard creado en Power BI, la navegación es intuitiva y está diseñada para facilitar el acceso a la información clave. Utiliza las flechas de navegación para moverte entre las diferentes páginas del dashboard. También se ha incorporado una flecha encerrada en un círculo que permite volver a la página de portada en cualquier momento.

# El dashboard está diseñado para conectar con el usuario aplicando técnicas de storytelling, como el patrón Z, y utiliza una paleta de colores adecuada para cada situación. Cada país analizado tiene un color asignado que se mantiene constante a lo largo del análisis, lo que facilita la identificación y comparación de datos entre países. En particular, la variable que representa las muertes se destaca con un color gris, respetando su seriedad.

# La información en el dashboard siempre se alinea con las conclusiones planteadas en este proyecto. Está completamente enfocado y dedicado a explicar las razones detrás de los resultados obtenidos. La última pestaña del dashboard se centra específicamente en las conclusiones y muestra los insights más importantes del trabajo.

# Para darle más profundidad y riqueza a los datos, se realizaron análisis complementarios de inteligencia de tiempo y exhaustivos estudios poblacionales. Estos análisis adicionales proporcionan una visión más completa y detallada de los datos, apoyando de manera sólida las conclusiones y recomendaciones del proyecto.

# 

# Conclusiones y Recomendaciones

**Conclusión General**

El análisis exploratorio de datos y las visualizaciones interactivas en los dashboards mostraron una clara necesidad de expandir la infraestructura de laboratorios farmacéuticos en regiones específicas de Latinoamérica. La información recolectada y analizada reveló diferencias significativas en la incidencia de COVID-19, las tasas de vacunación y la capacidad de respuesta sanitaria entre los países estudiados. Estos hallazgos son esenciales para formular estrategias de expansión que maximicen el impacto positivo en la salud pública.

#### Conclusiones Estratégicas

1. **Políticas de Vacunación y Eficiencia:**
   * Chile vacunó 3.6 veces a cada habitante, demostrando una alta eficiencia en la campaña de vacunación.
   * Perú y Argentina también mostraron buenos niveles de vacunación, con 2.7 y 2.4 veces respectivamente.
   * Brasil, México y Colombia presentan menores tasas de vacunación por habitante, con 1.65, 1.8 y 1.7 respectivamente, lo que indica áreas prioritarias para la mejora.
2. **Inversión en Salud y Recursos Humanos:**
   * Argentina destaca por tener la mayor cantidad de médicos, seguido de Chile.
   * Chile lidera en la cantidad de enfermeras, seguido por Brasil.
   * Perú y Colombia están en los últimos lugares en términos de personal médico y de enfermería, lo que sugiere una necesidad crítica de inversiones en recursos humanos.
3. **Incidencia de COVID-19:**
   * Brasil y Argentina presentan las mayores cifras de nuevos casos de COVID-19, con más de 40,000 y alrededor de 12,000 casos respectivamente.
   * México, Colombia, Chile y Perú muestran cifras significativamente menores, pero aún relevantes.
4. **Tasas de Mortalidad:**
   * México y Colombia tienen altas cifras de fallecimientos nuevos, con alrededor de 9,000 y 7,500 respectivamente.
   * Chile y Perú también presentan cifras preocupantes, con 5,000 y 4,500 fallecimientos nuevos.
5. **Problemas Sociales y Económicos:**
   * México lidera en la prevalencia de diabetes, seguido por Brasil y Chile.
   * Chile tiene la mayor prevalencia de tabaquismo, seguido por Argentina.
   * Estos problemas de salud pública subrayan la necesidad de un enfoque integral que no solo se centre en la pandemia de COVID-19 sino también en otras enfermedades crónicas.
   * En el carácter económico Brasil y México son los que más invirtieron en la pandemia, pero debido a su densidad esto no es suficiente.
6. **Densidad y Habitantes:**
   * La densidad de población y el número total de habitantes son factores críticos a considerar para la expansión. Países con alta densidad y grandes poblaciones como Brasil y México pueden beneficiarse más de la presencia de nuevos laboratorios.

#### Respuesta Principal del Proyecto

**Ubicaciones Óptimas en orden de prioridad para la Expansión de Laboratorios Farmacéuticos:**

* **Brasil**: Dada la alta incidencia de casos y la necesidad de mejorar las tasas de vacunación.
* **México**: Alta mortalidad y prevalencia de enfermedades crónicas hacen que la expansión sea urgente.
* **Colombia**: Con altas cifras de mortalidad y bajas tasas de vacunación, es crucial mejorar la infraestructura sanitaria.
* **Perú**: A pesar de una buena tasa de vacunación, la falta de personal médico sugiere la necesidad de más laboratorios.
* **Argentina y Chile:** Aunque tienen mejores tasas de personal sanitario y vacunación, siguen siendo estratégicos debido a su alta población y prevalencia de enfermedades crónicas.

Estas conclusiones permiten a Biogenesys tomar decisiones informadas para la expansión estratégica de sus laboratorios, optimizando así la respuesta a la pandemia y mejorando el acceso a las vacunas y la atención sanitaria en Latinoamérica.

# Reflexión personal

### Reflexión sobre lo Aprendido y Habilidades Adquiridas

#### Aprendizajes

# Durante el desarrollo de este proyecto, he profundizado en diversas áreas críticas del análisis de datos, particularmente en el contexto de la salud pública y la gestión de pandemias. Algunos de los aprendizajes más significativos incluyen:

# Gestión de Grandes Volúmenes de Datos: Trabajar con datasets extensos requirió un enfoque meticuloso para garantizar la integridad y precisión de los datos.

# Análisis Exploratorio de Datos (EDA): El EDA fue crucial para identificar patrones, tendencias y outliers, proporcionando una base sólida para el análisis posterior.

# Limpieza y Transformación de Datos: Aprendí la importancia de técnicas de limpieza y transformación de datos para asegurar que los análisis sean fiables y representativos.

# Visualización de Datos: El desarrollo de dashboards interactivos y gráficos de barras facilitó la interpretación de datos complejos, permitiendo una comunicación efectiva de los hallazgos.

# Toma de Decisiones Informada: Integrar datos de diferentes fuentes y contextos me permitió hacer recomendaciones estratégicas basadas en evidencia sólida.

#### Reflexión sobre el Proceso y Mejoras

# Si tuviera que volver a empezar este proyecto, hay varios aspectos que consideraría hacer de manera diferente para mejorar la eficiencia y la profundidad del análisis:

# Planeación Inicial más Detallada: Dedicar más tiempo a la planificación inicial del proyecto, estableciendo claramente los objetivos y los métodos desde el principio.

# Automatización de Procesos: Implementar más técnicas de automatización en la limpieza y transformación de datos para reducir errores manuales y aumentar la eficiencia.

# Mayor Integración de Datos Externos: Incluir datos adicionales de fuentes externas para enriquecer el análisis, como datos económicos y demográficos más detallados.

# Análisis Predictivo: Integrar modelos predictivos desde el inicio para anticipar tendencias futuras y no solo analizar datos históricos.

# En resumen, este proyecto ha sido una experiencia enriquecedora que ha mejorado mis habilidades y conocimientos como Analista de Datos. Si bien el enfoque actual ha sido efectivo, hay oportunidades para refinar y optimizar el proceso para futuros proyectos, asegurando resultados aún más robustos y útiles.

# EXTRA CREDIT

# Los extra credits cumplidos fueron 3:

* **Explorar el uso de funciones de orden superior para una manipulación eficiente de los datos.**

A través de funciones de orden superior, se logró un EDA más exhaustivo y completo. La función groupby() fue muy utilizada durante todo el proyecto para agrupar los datos por diferentes variables clave, lo que permitió un análisis más detallado y preciso.

* **Crear mapas utilizando la información de latitud y longitud disponible en tus datos.**

**Herramientas Utilizadas:**

Para la visualización geográfica de los datos, se emplearon dos herramientas principales: Folium y GeoPandas. Ambas herramientas permitieron generar mapas que ayudaron a visualizar la distribución y magnitud de los casos de COVID-19, muertes y vacunas administradas en los países analizados.

**Mapas con Folium:**

Con Folium, se creó un mapa reajustable del mundo donde se marcaron puntos en los países a analizar: Argentina, Brasil, Chile, Colombia, México y Perú. Cada punto se representó con diferentes tonos de colores según el heatmap propuesto.

**Desafíos y Resultados:**

Aunque el mapa resultó interactivo y flexible, el uso de diferentes colores no fue óptimo para el entendimiento general. La coloración no proporcionó una clara distinción entre los diferentes niveles de incidencia de COVID-19, muertes o vacunaciones.

La interpretación de los datos se vio comprometida debido a la complejidad en la paleta de colores utilizada.

**Mapas con GeoPandas:**

Con GeoPandas, se lograron representar de manera más efectiva las variables clave: casos confirmados, muertes y vacunas administradas. Se generaron mapas estáticos que mostraban estos datos con una escala de colores clara y precisa.

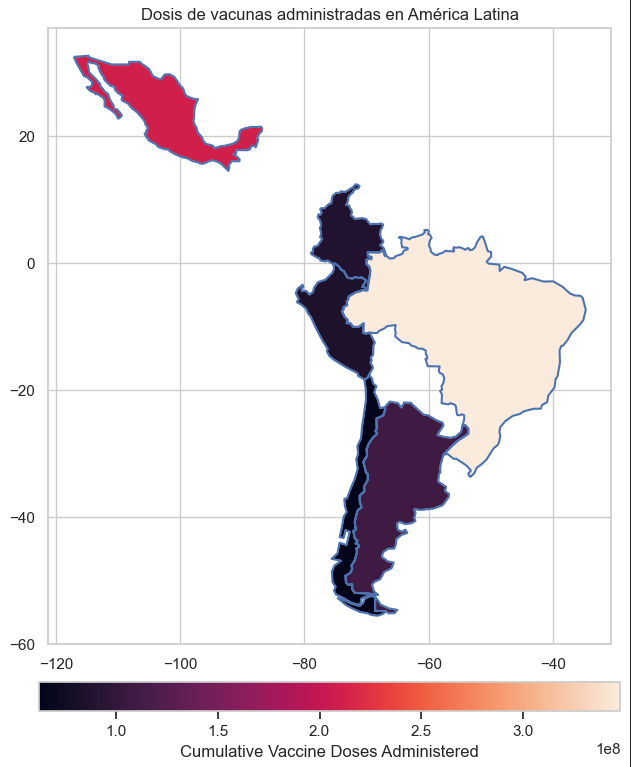
**Beneficios de GeoPandas**

Claridad y Precisión: Los mapas estáticos generados con GeoPandas proporcionaron una representación clara y precisa de los datos.

Foco en los Países Analizados: Al mostrar únicamente los países a analizar, se eliminó el ruido visual y se focalizó la atención en las áreas de interés.

**Ampliación del Análisis:** Estos mapas fueron una excelente forma de ampliar el análisis realizado durante el proyecto, proporcionando una visión geográfica detallada que complementó los hallazgos obtenidos a través de otras visualizaciones.

Ejemplo a continuación:



* **Modelado de Machine Learning**

**Las variables empleadas en el modelo fueron:**

* **new\_confirmed**
* **new\_deceased**
* **new\_recovered**

#### 

#### 

#### Enfoque del Modelo

Se planteó una regresión logística con el objetivo de predecir cuántos de los futuros contagios terminarían en muerte en ausencia de vacunación. Para ello, se transformaron los valores de la variable new\_deceased en binarios:

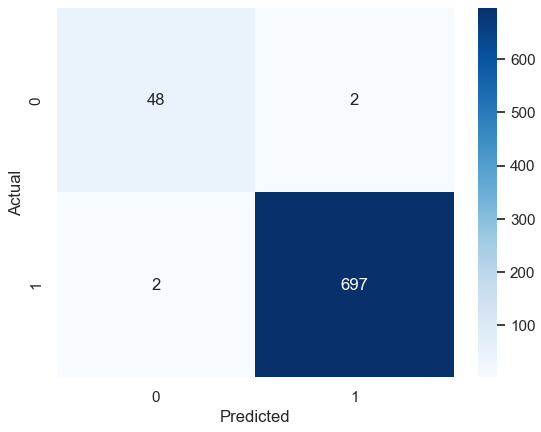
* **1: Si alguien murió.**
* **0: Si nadie murió.**

#### Métrica de Evaluación

La **métrica** de evaluación seleccionada fue una **matriz de confusión**. Se utilizó una muestra total de 750 registros. Los resultados obtenidos fueron los siguientes:

* **48 True Positives:** Casos donde el modelo predijo correctamente que alguien moriría (valor binario de 1).
* **697 True Negatives**: Casos donde el modelo predijo correctamente que nadie moriría (valor binario de 0).

Estos resultados reflejan la capacidad del modelo para distinguir entre casos que resultarán en muerte y aquellos que no, proporcionando una herramienta valiosa para predecir los impactos de futuros contagios en ausencia de vacunación.



* **Uso de funciones de inteligencia de tiempo en PowerBI**

Se utilizó la función de inteligencia de tiempo DATEADD para crear un gráfico de líneas que muestra la tasa de crecimiento porcentual de los nuevos casos confirmados de COVID-19.