**INFORME PROYECTO INTEGRADOR 4**

**Nombre del autor: Ezequiel Gonzalez**

**Email: ezeqiel111@gmail.com**

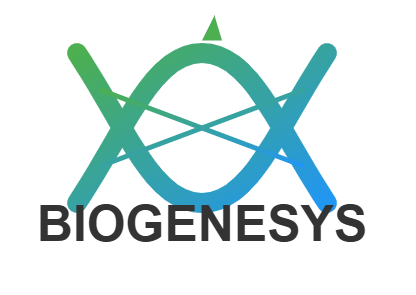
**Cohorte: DAFT-14**

**Fecha de entrega:** 06/09/2025

**Institución:**

**BIOGENESYS** es una destacada empresa farmacéutica comprometida con la innovación en el sector de la salud. Su misión principal se centra en el desarrollo, producción y distribución de soluciones farmacéuticas avanzadas, incluyendo vacunas, con el objetivo de mejorar la salud pública y la calidad de vida de las personas.

En el contexto actual, marcado por los desafíos sanitarios globales y la necesidad de una respuesta rápida ante crisis como la pandemia de COVID-19, BIOGENESYS busca expandir estratégicamente su alcance. La empresa invierte en la investigación y el análisis de datos para identificar las ubicaciones más óptimas para el establecimiento de nuevos laboratorios y centros de vacunación. Esta expansión no solo busca optimizar la producción y la logística de distribución, sino también fortalecer la capacidad de respuesta ante futuras emergencias sanitarias y garantizar un mejor acceso a las vacunas y tratamientos esenciales en regiones clave.

El presente análisis de datos, enfocado en Latinoamérica, se enmarca en la visión de BIOGENESYS de liderar una respuesta proactiva y basada en evidencia para los desafíos de salud más apremiantes de n**uestro tiempo.**

# 

# 

# 

# 

# 

# 

# 

# Introducción

# El propósito central de este proyecto es identificar las ubicaciones más óptimas para la expansión de laboratorios farmacéuticos y centros de vacunación de BIOGENESYS en Latinoamérica. Esto se logrará mediante un análisis de datos detallado sobre la incidencia de COVID-19, las tasas de vacunación y la disponibilidad de infraestructura sanitaria. La meta es optimizar la respuesta ante futuras emergencias sanitarias y mejorar el acceso a vacunas y tratamientos esenciales, guiando las decisiones de inversión de la empresa con base en evidencia sólida.

# 

# Desarrollo del proyecto

# BIOGENESYS es una destacada empresa farmacéutica comprometida con la innovación en el sector de la salud. Su misión principal se centra en el desarrollo, producción y distribución de soluciones farmacéuticas avanzadas, incluyendo vacunas, con el objetivo de mejorar la salud pública y la calidad de vida de las personas.

# En el contexto actual, marcado por los desafíos sanitarios globales y la necesidad de una respuesta rápida ante crisis como la pandemia de COVID-19, BIOGENESYS busca expandir estratégicamente su alcance. La empresa invierte en la investigación y el análisis de datos para identificar las ubicaciones más óptimas para el establecimiento de nuevos laboratorios y centros de vacunación. Esta expansión no solo busca optimizar la producción y la logística de distribución, sino también fortalecer la capacidad de respuesta ante futuras emergencias sanitarias y garantizar un mejor acceso a las vacunas y tratamientos esenciales en regiones clave.

# El presente análisis de datos, enfocado en Latinoamérica, se enmarca en la visión de BIOGENESYS de liderar una respuesta proactiva y basada en evidencia para los desafíos de salud más apremiantes de nuestro tiempo.

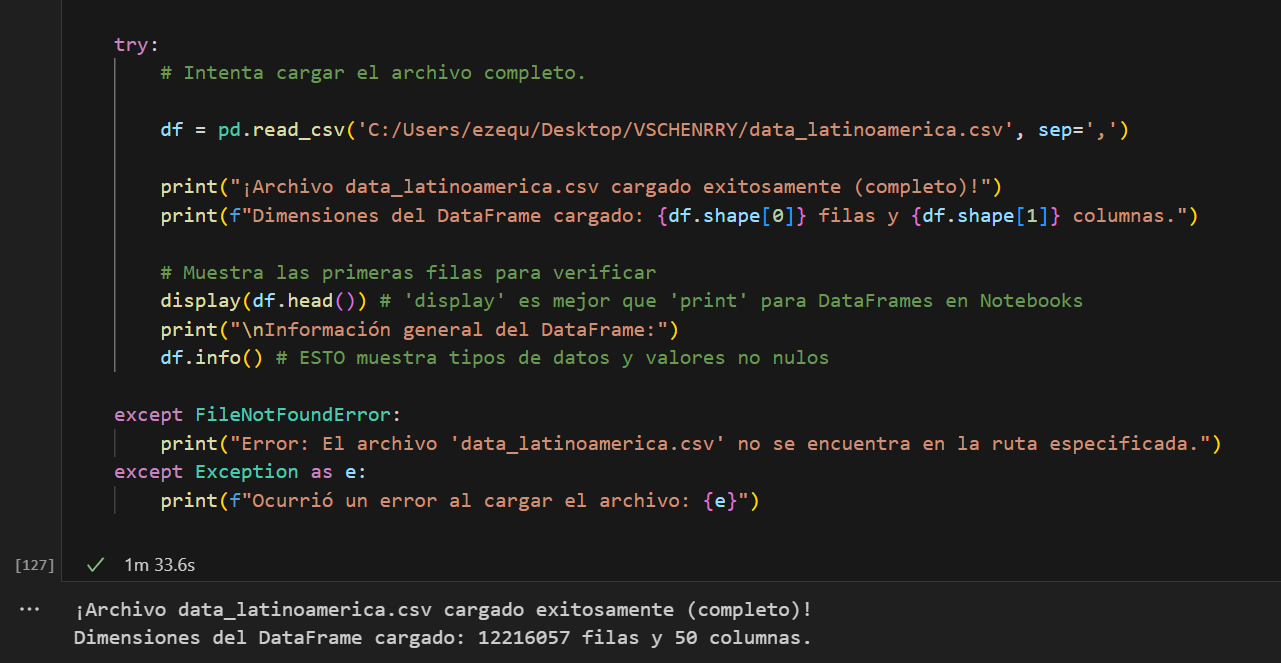
### Metodología y Preparación de Datos

# Para este análisis estratégico, se utilizó un conjunto de datos integral que originalmente comprendía aproximadamente 21 GB con 22 millones de registros y 707 columnas. Conscientes de la magnitud y la necesidad de un enfoque específico, el equipo de ingeniería de datos preprocesó un subconjunto relevante, enfocado exclusivamente en países de Latinoamérica. Este nuevo archivo, denominado data\_latinoamerica.csv, contiene un volumen manejable de 12,216,057 filas y 50 columnas, complementado por un archivo readme.txt que detallaba la información de cada columna seleccionada.

# El proceso de preparación y limpieza de datos, fundamental para asegurar la confiabilidad del análisis, se llevó a cabo meticulosamente siguiendo las siguientes fases:

**AVANCE 1**

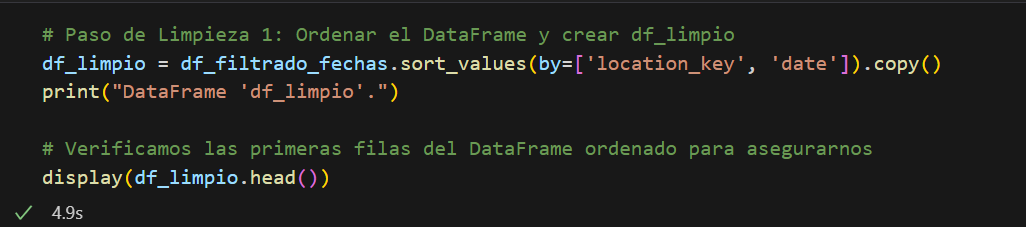
# Extracción y Carga Inicial: El archivo data\_latinoamerica.csv fue cargado en un entorno de Jupyter Notebook en Visual Studio Code utilizando la librería Pandas. A pesar de su tamaño reducido, se requirió una gestión eficiente de la memoria para su carga completa.



# 

# Filtrado Estratégico: Se aplicaron filtros precisos para concentrar el análisis en los países de interés para la expansión: Colombia, Argentina, Chile, México, Perú y Brasil. Adicionalmente, se filtraron los datos para incluir solo registros posteriores al 1 de enero de 2021, lo que permitió un enfoque en el período más relevante de la pandemia y post-pandemia. Se optó por mantener la granularidad de las sub-regiones (location\_key) inicialmente para un análisis más detallado, pero posteriormente se consolidó el dataset a nivel de país principal (~3700 filas) para la mayoría de las visualizaciones comparativas.

# Manejo de Valores Faltantes (NaN): Se identificaron y abordaron los valores nulos (NaN) en columnas críticas como new\_confirmed, new\_deceased, cumulative\_vaccine\_doses\_administered, new\_recovered, cumulative\_recovered, y las métricas de temperatura/clima. Para las series temporales, se priorizó el relleno por país (location\_key) utilizando métodos como ffill() (relleno hacia adelante) y bfill() (relleno hacia atrás) para mantener la coherencia temporal. En los casos donde persisieron nulos en métricas de "nuevos eventos" (new\_confirmed, new\_deceased, new\_recovered), estos fueron rellenados con cero, bajo el supuesto de "no reportado". Para los escasos nulos restantes, especialmente en cumulative\_recovered y rainfall\_mm, que no pudieron ser inferidos temporalmente, se optó por un relleno final con cero, asegurando la completitud del dataset sin distorsionar la información sensible.



# Corrección y Estandarización de Datos: Se identificaron y corrigieron valores atípicos o ilógicos, como los valores negativos en new\_confirmed y new\_recovered, los cuales fueron ajustados a cero para reflejar la realidad del conteo. La columna de date fue convertida al formato datetime y establecida como índice principal del DataFrame para facilitar el análisis de series temporales.

# Generación de Métricas Clave: Se calcularon nuevas métricas esenciales como casos activos y tasa de crecimiento diaria de casos, proporcionando una visión más profunda de la dinámica de la pandemia.

# Conclusión de las Transformaciones y Limpieza de Datos:

# Este riguroso proceso de preparación de datos fue fundamental para transformar un conjunto de datos crudo y extenso en un recurso limpio, estructurado y confiable. El resultado es el archivo DatosFinalesFiltrado.csv, un dataset listo para ser utilizado en plataformas de visualización interactiva como Power BI. La calidad y la consistencia de los datos garantizadas a través de estas transformaciones son la base sobre la cual se construirán los análisis estratégicos y las recomendaciones para la expansión de BIOGENESYS, asegurando que las decisiones de inversión estén fundamentadas en evidencia sólida y precisa.

# 

# EDA e insights

# 1- Tasas de vacunación: evolución e impacto

# 

# Se observa claramente una aceleración constante en la administración de dosis acumuladas a lo largo del tiempo en todos los países, reflejando el progreso de las campañas de vacunación.

# 

# Brasil se destaca como el país con el mayor volumen absoluto de dosis administradas, lo cual es coherente con su tamaño poblacional y la alta incidencia de casos. Esto sugiere una gran capacidad logística de distribución y administración de vacunas ya establecida. de la severidad y la fatalidad de la enfermedad.

2- Relacin vacunacion - mortalidad

Los gráficos de línea revelan una tendencia general clara: a medida que las dosis de vacunas administradas acumuladas aumentaron, la suma mensual de nuevas muertes tendió a disminuir. Esta relación es un fuerte indicador del impacto positivo de la vacunación en la reducción de la severidad y la fatalidad de la enfermedad.

3-Análisis de la Variabilidad de Temperatura y Potencial de Contagios:

Los gráficos de distribución de la temperatura media indican que Chile y Argentina presentan un rango de temperatura media más amplio, lo que se traduce en una mayor variabilidad estacional a lo largo del año.

Desde una perspectiva de salud pública general, las estaciones de otoño e invierno suelen estar asociadas con un aumento en la incidencia de enfermedades respiratorias debido a factores como el mayor tiempo en interiores, la humedad y el frío. Esto podría hacer que estos países sean más propensos a repuntes de contagios de COVID-19 durante esos meses.

Sin embargo, al observar los gráficos de evolución mensual de nuevos casos confirmados por país, no se puede establecer una correlación estacional clara y consistente ligada directamente a los cambios de temperatura en el período analizado. Los picos de contagio de COVID-19 en este dataset parecen estar más influenciados por otros factores, como la aparición de nuevas variantes, la adopción de medidas sanitarias o el avance de la vacunación

# 

# 

# Análisis del dashboard

### Análisis del Dashboard de Power BI

# Puntos Fuertes y Mensajes Clave:

# Mensaje Central Directo: El título "EXPANSION: BRASIL Y CHILE" inmediatamente comunica el objetivo principal del análisis, lo cual es muy efectivo.

# KPIs Resumen Clave: Las tarjetas de "ESPERANZA DE VIDA" y "MORTALIDAD INFANTIL" proporcionan un contexto de salud pública de alto nivel, que es relevante para la misión de BIOGENESYS.

# Enfoque en Países Estratégicos: Los botones de filtro de "PAIS" (Brasil y Chile) son muy útiles para la interactividad y para permitir a los directivos profundizar en los datos de estos dos países prioritarios.

# Visualización de Personal de Salud: El gráfico de barras de "PERSONAL DE SALUD CADA 1000 HABITANTES" es muy bueno. Muestra la densidad de médicos y enfermeras, una métrica vital para la infraestructura. Aquí se ve claramente que Chile y Argentina tienen una alta densidad de profesionales, lo que apoya la elección de Chile como un centro de operaciones (especialmente por el talento humano).

# Volumen de la Pandemia: Los gráficos de "CASOS CONFIRMADOS" (parece un gráfico de proporción o treemap por país) y "DECESOS TOTALES" son cruciales.

# En "CASOS CONFIRMADOS", es evidente que Brasil es el país con el mayor volumen absoluto de casos, lo que refuerza la idea de una alta demanda y una oportunidad de mercado masiva.

# En "DECESOS TOTALES", Brasil también muestra el mayor número de fallecimientos, lo que subraya la escala del impacto de la pandemia en ese país.

# Avance de la Vacunación: El gráfico de área de "VACUNAS ADMINISTRADAS" muestra la evolución temporal del esfuerzo de vacunación. Esto es clave para ver la capacidad logística y el alcance de las campañas en cada país.

# Sugerencias para Mejorar la Claridad y el Impacto:

# Gráfico "CASOS CONFIRMADOS": El título "BRASIL Y MÉXICO" puede ser un poco confuso si el gráfico muestra a todos los países. Considera un título más general como "Casos Confirmados por País" o "Distribución de Casos Confirmados en Latinoamérica". Si el gráfico es un 100% apilado, la "proporción" de Brasil dominará, pero los valores absolutos pueden ser difíciles de leer en el treemap. Asegúrate de que los números (27 mil M, 8 mil M, etc.) sean claros si se refieren a millones de casos o billones.

# Consistencia en la Etiquetado: Asegúrate de que todas las etiquetas de los ejes y títulos sean consistentes en unidades y claridad. Por ejemplo, en "VACUNAS ADMINISTRADAS", el eje Y debería ser 0 mil M, 5 mil M, 10 mil M si son "millones", o simplemente 0, 5M, 10M con un rótulo "Dosis Acumuladas (Millones)" en el eje.

# Integración de la Narrativa: Considera añadir un pequeño cuadro de texto en cada sección principal o al lado de los gráficos clave que resuma el *insight* principal que el directivo debería llevarse de esa visualización específica. Por ejemplo, junto al gráfico de casos confirmados: "Brasil representa el mayor volumen de demanda de servicios de salud."

# En resumen, tu dashboard es un buen punto de partida. Transmite efectivamente el mensaje principal y ya presenta visualizaciones clave. Con los ajustes de claridad y etiquetado, será una herramienta muy poderosa.

# KPIs Resumen de Alto Nivel: Las tarjetas grandes en la parte superior ("CONFIRMADOS: 52 mill.", "DECESOS: 1 mill.", "RECUPERADOS: 37 mill.") son muy efectivas. Proporcionan un resumen instantáneo y de gran impacto de la magnitud de la pandemia en los países analizados. Esto es excelente para la toma de decisiones rápidas.

# Filtros de Año: Los filtros de "Año" (2021, 2022) son muy útiles para la interactividad. Permiten a los directivos segmentar rápidamente los datos por período, lo que facilita el análisis de tendencias anuales.

# Proporción de Confirmados vs. Decesos: El gráfico de anillo (donut chart) "PROPORCION CONFIRMADOS VS DECESOS" es una visualización inteligente para mostrar la tasa de letalidad global de forma inmediata (1.95%). Es un dato muy revelador.

# Evolución de Casos Confirmados: El gráfico de líneas "CASOS CONFIRMADOS" muestra la evolución de los casos por país a lo largo del tiempo. Las líneas son claras y permiten ver los picos y valles para cada nación.

# Vacunas Administradas (por Mes): El gráfico de barras "VACUNAS ADMINISTRADAS" presenta el volumen de dosis administradas mensualmente. Es una buena forma de visualizar el ritmo de la campaña de vacunación a lo largo del tiempo.

# Personal Médico: El gráfico de barras "PERSONAL MEDICO" es muy relevante para la infraestructura, comparando la densidad de enfermeras y doctores por país. Aquí se visualiza bien qué países tienen más o menos personal de salud.

# Sugerencias para Mejorar la Claridad y el Impacto:

# Consistencia de Unidades en KPIs: Asegúrate de que las unidades en los KPIs (millones) sean claras. Los 52 mill., 1 mill., 37 mill. son intuitivos, pero si los datos suben mucho, considera si se necesitan abreviaturas como B para billones si se superan los miles de millones.

# Eje Y de Gráficos de Línea: Para los "CASOS CONFIRMADOS" de línea, el eje Y (0 mil., 1 mil., 2 mil., 3 mil.) parece referirse a millones. Si es así, asegúrate de que la etiqueta del eje Y sea explícita (ej. "Casos Mensuales (Millones)"). La misma sugerencia aplica para el gráfico de "VACUNAS ADMINISTRADAS" (ej. "Dosis Administradas (Millones)").

# Gráfico de Proporción: Para "PROPORCION CONFIRMADOS VS DECESOS", considera si la leyenda es inmediatamente clara para "Total\_Confirmados" y "Total\_decesos". Quizás un texto que diga "Casos Confirmados" y "Decesos" o "Mortalidad" sea más directo.

# Interacción de Gráficos: Asegúrate de que los filtros de Año (2021, 2022) y la selección de países en la leyenda de "CASOS CONFIRMADOS" interactúen adecuadamente con todos los demás gráficos de la página (KPIs, Vacunas Administradas, Personal Médico).

# En general, esta página es muy efectiva para dar una visión de alto nivel de la pandemia y las capacidades de respuesta. Las visualizaciones son apropiadas para los datos y comunican bien los mensajes clave.

# 

# Conclusiones y Recomendaciones

# A lo largo de este proyecto, hemos guiado a la empresa farmacéutica BIOGENESYS a través de un riguroso análisis de datos para identificar oportunidades estratégicas de expansión de laboratorios y centros de vacunación en Latinoamérica.

# Comenzamos con la preparación y limpieza de un dataset extenso (data\_latinoamerica.csv), transformando millones de registros brutos en un conjunto de datos limpio y estructurado (DatosFinalesFiltrado.csv). Este proceso incluyó el manejo de valores nulos, la corrección de inconsistencias y la creación de métricas clave como los casos activos y la tasa de crecimiento diaria.

# Posteriormente, realizamos un profundo Análisis Exploratorio de Datos (EDA) y Visualización utilizando Python (Pandas, NumPy, Matplotlib, Seaborn y Statsmodels). Gracias a diversos gráficos (histogramas, gráficos de línea, mapas de calor, boxplots, violinplots y diagramas de dispersión), pudimos:

# Comprender la evolución de la pandemia, identificando picos y valles, y la notable disociación entre casos y muertes en 2022 debido a la vacunación y la evolución viral (ej., variante Ómicron).

# Analizar las tasas de vacunación, su progresión y su impacto en la reducción de la mortalidad.

# Examinar la demografía y la infraestructura sanitaria de los países, revelando estructuras poblacionales (ej., poblaciones jóvenes en México/Perú, envejecidas en Chile) y densidades de personal médico.

# Investigar correlaciones entre indicadores demográficos y de salud, resaltando interconexiones críticas.

# Finalmente, integramos todos estos hallazgos en Power BI, diseñando un dashboard interactivo que traduce el análisis técnico en *insights* accionables para los directivos. Este dashboard permite visualizar de forma intuitiva los volúmenes de casos y vacunación, la capacidad de infraestructura y la demografía, sustentando las recomendaciones de expansión.

# En síntesis, este proyecto ha demostrado cómo un enfoque de análisis de datos robusto y visualizaciones claras puede guiar eficazmente decisiones de inversión estratégicas, permitiendo a BIOGENESYS optimizar su respuesta a desafíos de salud pública y maximizar su impacto positivo en la región.

# 

# Reflexión personal

# Este proyecto ha sido una valiosa experiencia práctica, consolidando habilidades fundamentales en el análisis de datos. Entre los aprendizajes clave, destaco:

# Manejo de Datos a Gran Escala: La experiencia con un dataset de millones de registros reforzó la importancia de la eficiencia en el procesamiento de datos con Pandas y el manejo de recursos computacionales (ej., optimización de carga de CSV).

# Limpieza de Datos Profunda: Se profundizó en técnicas de identificación y tratamiento de valores nulos y la corrección de inconsistencias lógicas

# Análisis Exploratorio y Visualización: La creación de una amplia gama de gráficos permitió no solo visualizar datos, sino también extraer *insights* complejos y patrones ocultos, mejorando la capacidad de contar historias con datos.

# Análisis de Series Temporales: La aplicación de la descomposición y autocorrelación proporcionó una comprensión más profunda de las tendencias y estacionalidades en datos dinámicos.

# Comunicación de Resultados: La integración en Power BI y el diseño de dashboards interactivos subrayaron la importancia de traducir el análisis técnico en herramientas accionables para la toma de decisiones gerenciales.

# Si tuviera que volver a empezar este proyecto, en general, mantendría la misma metodología paso a paso, ya que demostró ser efectiva para abordar la complejidad del dataset y las consignas. Sin embargo, realizaría algunos cambios:

# 

# 