**EXPANSIÓN ESTRATÉGICA DE BIOGENESYS**

**Nombre del autor: Edwin Alberto Echeverri Córdoba**

**Email: eechevec@gmail.com**

**Cohorte: DA-FT14**

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

**Institución:** Empresa Farmacéutica BIOGENESYS



# 

# Introducción

La empresa farmacéutica BIOGENESYS ha puesto en marcha una iniciativa para identificar las ubicaciones más estratégicas para la expansión de sus laboratorios en América Latina. Esta decisión se fundamenta en un análisis riguroso de datos relacionados con la incidencia de COVID-19, las tasas de vacunación y la disponibilidad de infraestructura sanitaria en la región. El objetivo principal es fortalecer la capacidad de respuesta ante los efectos de la pandemia y la postpandemia, promoviendo así un mejor acceso a vacunas y servicios de salud en las zonas más necesitadas.

Como Data Analyst asignado al proyecto, mi responsabilidad es desarrollar un estudio que respalde esta estrategia de expansión, centrándose en seis países clave: Argentina, Brasil, Chile, Colombia, México y Perú. La dirección ha propuesto esta inversión con el propósito de identificar regiones prioritarias y generar información clave que respalde la toma de decisiones basada en datos.

**Objetivos del análisis**

Los objetivos definidos para este estudio son claros y medibles:

* Realizar un análisis exploratorio de datos relacionados con la incidencia de COVID-19 y otros factores relevantes, identificando patrones, tendencias y oportunidades mediante estadísticas descriptivas, métricas clave y visualizaciones.
* Aplicar técnicas de limpieza y depuración de datos para garantizar la calidad y fiabilidad de la información, elemento fundamental para generar conclusiones precisas y útiles para la estrategia empresarial.
* Optimizar los procesos de extracción, transformación y carga (ETL) para mejorar el acceso, manejo y actualización de los datos, aumentando la eficiencia del análisis y la capacidad de respuesta ante nuevos escenarios.
* Diseñar dashboards interactivos y visualmente eficaces, que permitan explorar los datos desde diversas perspectivas y faciliten una toma de decisiones ágil, estratégica y orientada a resultados.

**Impacto esperado**

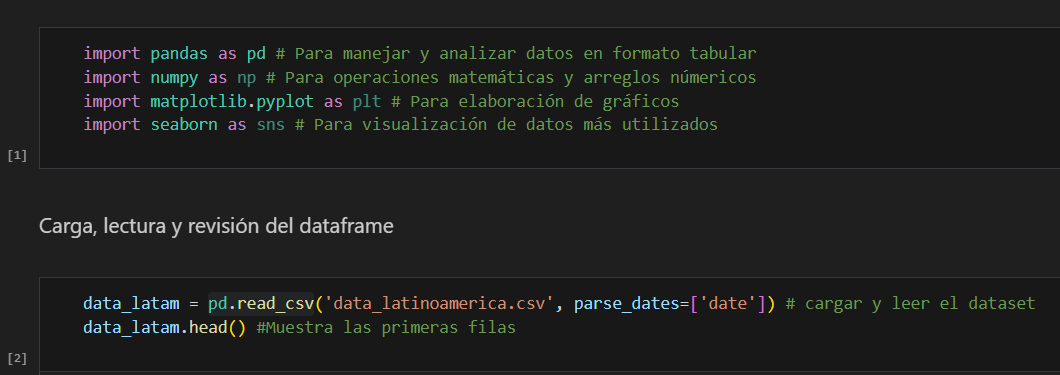
Este análisis proporcionará insights valiosos sobre posibles ubicaciones estratégicas para la instalación de nuevos laboratorios y centros de vacunación. A través del estudio de variables epidemiológicas, demográficas y de infraestructura, se generará una comprensión profunda del entorno regional, lo cual será determinante para el éxito de la expansión de BIOGENESYS en los mercados latinoamericanos.

# Desarrollo del proyecto

**Introducción al Análisis de Datos**

En la primera etapa del proyecto, se trabajó con un dataset proporcionado la empresa, el cual contiene 12.216.057 filas y 50 columnas en formato CSV. Esta versión procesada facilita la manipulación del dataset original, que tiene un tamaño de 21 GB. Además, se incluyó un archivo adicional con descripciones detalladas de las columnas seleccionadas para una preparación y limpieza de datos adecuada.

Una vez familiarizados con el contenido del dataset, se creó un notebook en Python utilizando Visual Studio Code. Se importaron las librerías necesarias y se cargó el archivo CSV, verificando la coincidencia de registros y columnas mediante la función np.shape de Numpy (ver archivo “PIDA\_M4\_Edwin\_Echeverri.ipynb”).



**Filtros Iniciales y Limpieza de Datos**

El primer paso fue filtrar la información correspondiente a los países seleccionados (Argentina, Brasil, Chile, Colombia, México y Perú), aplicando una máscara que permitiera focalizar el análisis. Posteriormente, se evaluó la calidad de los datos identificando columnas con más de 4 millones de valores faltantes. La columna location\_key presentaba duplicados de ubicaciones debido a la presencia de regiones y subregiones, por lo cual se refinó el filtro para incluir solo las instancias correspondientes a los países seleccionados.

Se aplicó un segundo filtro temporal para utilizar datos posteriores al 1 de enero de 2021. En cuanto a los valores nulos, se adoptaron diferentes estrategias: para columnas con pocos faltantes se implementó el uso del último valor valido como forma de disminuir el impacto de este reemplazo de datos en el dataframe en general, mientras que en aquellas con altos porcentajes de valores nulos se optó por rellenarlos con ceros, a fin de conservar datos útiles de otros países que si contenían dicha información faltante. Esta decisión se tomó teniendo en cuenta que esos ceros podrían ser reemplazados en el futuro con información actualizada.

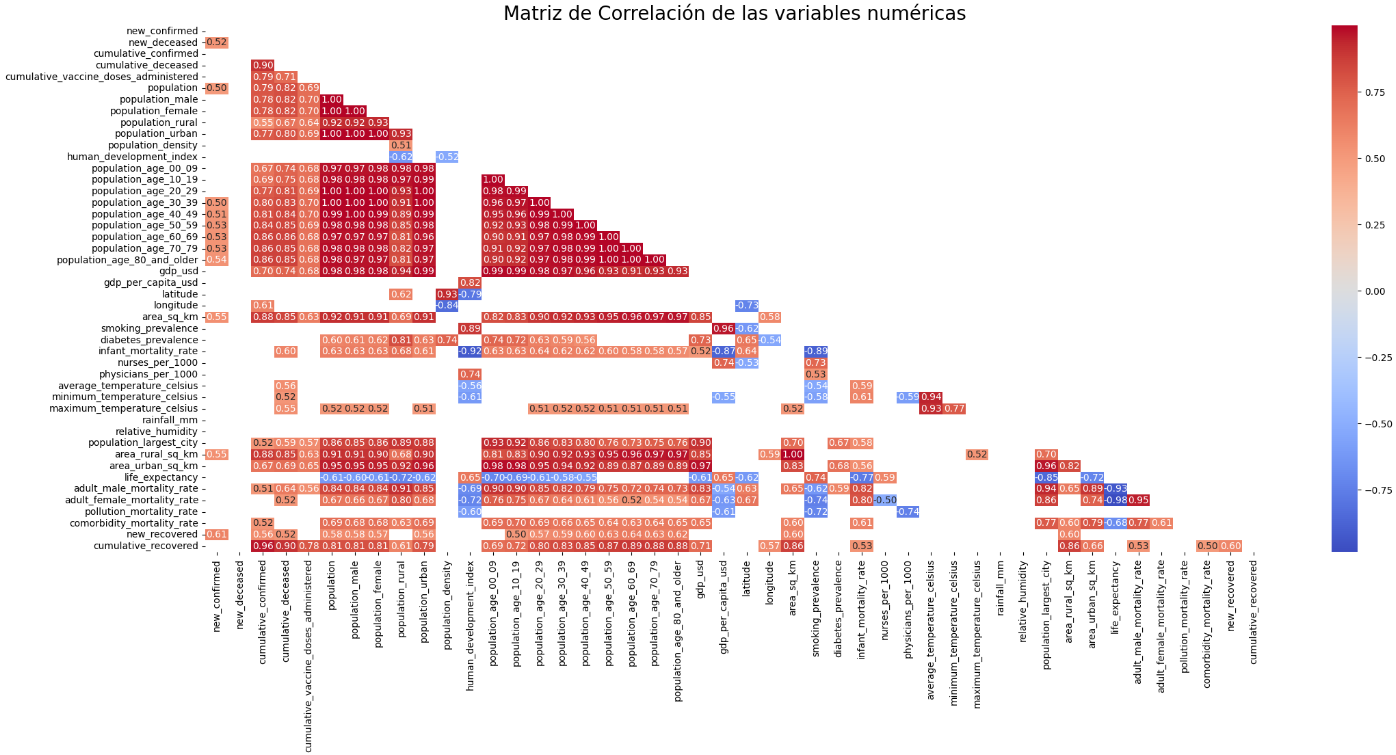


Finalmente, se exportó un nuevo archivo CSV con los datos filtrados para reutilizarlos sin repetir el proceso de limpieza. Se utilizaron funciones como .describe() para obtener estadísticas básicas, aunque se identificó la necesidad de visualizar estos datos mediante gráficos para facilitar su interpretación.

**Análisis Exploratorio de Datos (EDA)**

En la segunda fase, se profundizó en el análisis exploratorio de datos. Se desarrollaron visualizaciones a través de las librerías Seaborn y Matplotlib para detectar patrones, tendencias y anomalías.

Se construyó una matriz de correlación para identificar relaciones significativas entre variables, enfocándose en correlaciones absolutas superiores a 0.5. Posteriormente, se utilizaron barplots para examinar visualmente variables seleccionadas. Entre los hallazgos iniciales se observó que Brasil presenta cifras más elevadas por su extensión geográfica, Chile tiene una mayor prevalencia de tabaquismo y México registra altos niveles de diabetes.

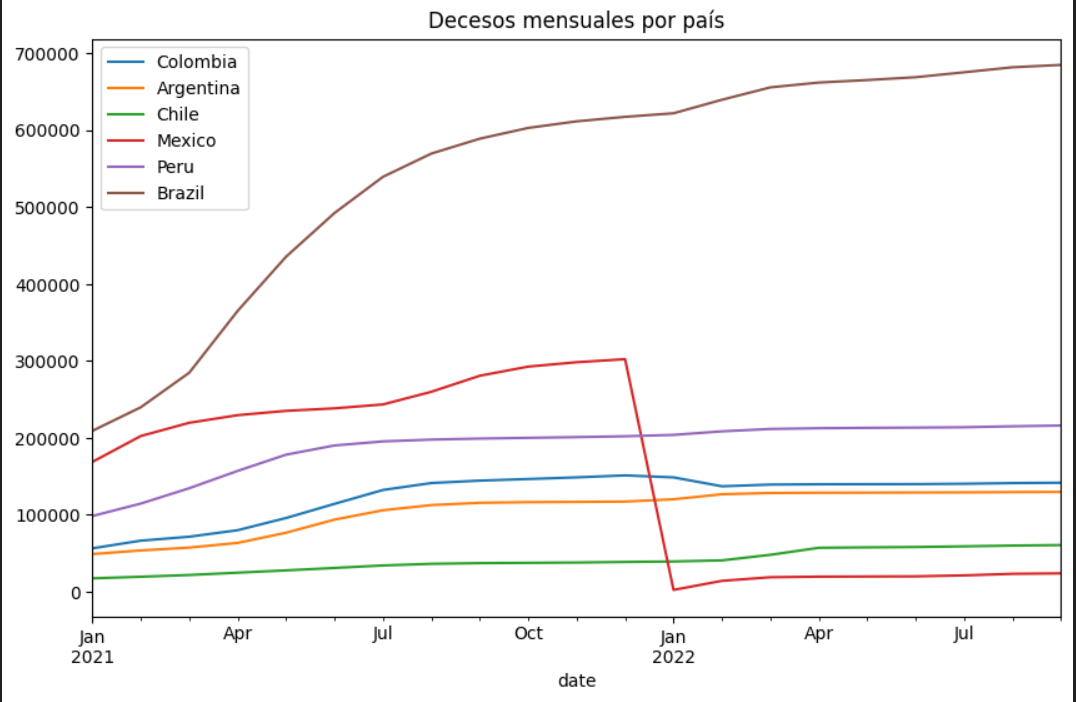


Se generaron histogramas y scatter plots para analizar la distribución de variables como la temperatura promedio y su relación con nuevos casos confirmados y fallecimientos. Se identificó que en algunos países la temperatura no influye directamente, mientras que en otros podría tener un impacto.

**Análisis Temporal y Demográfico**

Se realizó un análisis temporal con gráficos de líneas para observar la evolución mensual de las dosis administradas y decesos. En algunos países se detecta una tendencia creciente en la vacunación, mientras que en otros dicha tendencia disminuye, posiblemente por factores como disponibilidad de vacunas y campañas nacionales.

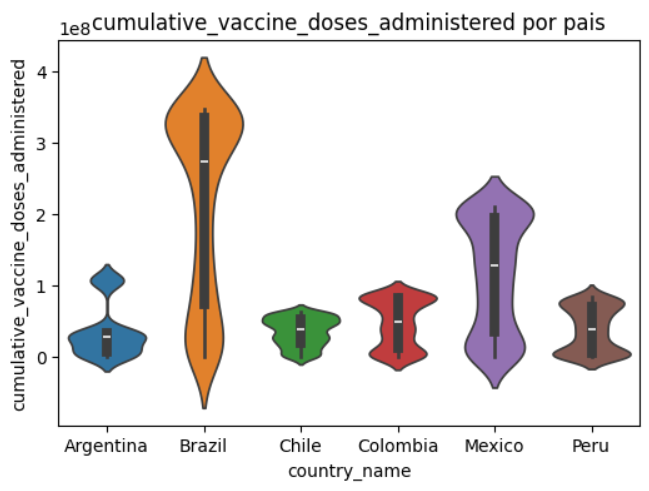
También se observó que Brasil presenta el mayor número de decesos y casos confirmados. En contraste, Chile muestra indicadores más favorables tanto en casos como en mortalidad, lo que podría indicar una gestión efectiva de la pandemia.



Además, se evaluó la distribución de la población por edad y se analizaron los grupos más vacunados, así como la mortalidad por género. Se exploraron enfermedades preexistentes como tabaquismo y diabetes, identificándolas como factores asociados a una mayor mortalidad en los países analizados.

**Análisis Avanzado y Series Temporales**

En el penúltimo avance, se aplicaron técnicas avanzadas de análisis de series temporales con Pandas y Numpy. Se estudió el comportamiento semanal y mensual de los casos nuevos, observándose picos asociados a eventos sociales. Aunque la tendencia general muestra una disminución, persisten alzas puntuales.



Respecto al progreso de vacunación, Brasil lidera en número total de dosis administradas, aunque esto se explica en parte por su mayor población. Se analizan también indicadores como porcentaje de población urbana y dosis aplicadas por cada 100 habitantes, donde Chile muestra un rendimiento superior.

Se calcula la tasa mortalidad para los países seleccionados, destacando que Perú y México presentan peores indicadores, mientras que Argentina y Chile tienen valores más bajos. Se subraya la necesidad de completar los datos de recuperados para mejorar la precisión del análisis de casos activos.

**Integración en Power BI**

En la fase final, se integraron los hallazgos en un dashboard interactivo en Power BI, conectando el archivo CSV procesado en Python. Este dashboard permite una visualización clara de los resultados, facilitando la identificación de regiones prioritarias para la expansión de laboratorios.

Se compararon además las ventajas de las visualizaciones interactivas frente a gráficas estáticas, resaltando el valor de los dashboards digitales en entornos de toma de decisiones estratégicas.

Este informe resume los avances logrados en el análisis de datos para apoyar la estrategia de expansión de BIOGENESYS en Latinoamérica, proporcionando una base sólida para decisiones informadas en salud pública y logística sanitaria.

# Insigths

**Identificación de Países Prioritarios**

* Brasil presenta la mayor carga de casos confirmados y muertes, lo que lo convierte en un país prioritario para la instalación de nuevos laboratorios o centros de vacunación. Sin embargo, esta necesidad debe equilibrarse con su gran extensión territorial y desafíos logísticos.
* Chile se posiciona como un referente en gestión sanitaria: menores tasas de letalidad y una administración eficiente de vacunas, lo que lo convierte en un entorno favorable para modelos piloto o expansión basada en eficiencia.
* México y Perú muestran indicadores preocupantes en letalidad y tasas de mortalidad, lo que señala la necesidad urgente de reforzar infraestructura sanitaria.

**Variables Críticas en la Pandemia**

* Se identificó una alta correlación entre diabetes, tabaquismo y mortalidad por COVID-19, lo que permite enfocar campañas sanitarias preventivas en estos grupos vulnerables.
* La urbanización parece estar asociada a una mejor contención de la pandemia en países como Chile, lo que sugiere priorizar áreas urbanas densas en futuras expansiones.
* La relación entre temperatura y casos confirmados varía por país, lo que indica que los factores climáticos deben analizarse localmente antes de diseñar estrategias.

**Tendencias Temporales y Comportamiento Estacional**

* Se observaron picos estacionales en los casos confirmados (especialmente tras festividades), lo que sugiere momentos clave para intervención sanitaria o campañas de vacunación.
* Las dosis administradas muestran una curva creciente en la mayoría de los países, pero con variaciones importantes. Este dato es clave para sincronizar el suministro de vacunas con la demanda proyectada.

**Brechas en la Calidad de Datos**

* La falta de datos de personas recuperadas limita el análisis de casos activos, afectando la precisión de predicciones y recursos requeridos.
* Algunas columnas del dataset presentaban altos niveles de valores nulos, por lo que se debieron tomar decisiones de imputación que deben revisarse y actualizarse con nuevas fuentes de datos.

**Segmentación Demográfica y Género**

* Se identificaron grupos etarios y géneros más afectados, permitiendo desarrollar planes de acción segmentados por perfil poblacional.
* Este tipo de análisis también ayuda a definir políticas de vacunación focalizadas, maximizando su impacto y eficiencia.

# Análisis del dashboard

A partir de los insights obtenidos durante el análisis en Python, se construyó un dashboard en Power BI que integra las visualizaciones más relevantes, asumiendo el rol de Data Analytics. Adicionalmente, se generaron nuevas visualizaciones aprovechando la versatilidad y facilidad que ofrece Power BI para crear gráficos y cálculos dinámicos. Cabe destacar que ciertas relaciones resultaron más evidentes al utilizar esta herramienta, posiblemente debido a la familiaridad con su entorno de trabajo.

El reporte interactivo está compuesto por una portada y dos hojas principales: “Datos Generales” y “Datos Pandemia”.

En la hoja “Datos Generales”, se visualiza información demográfica clave de los países seleccionados, incluyendo:

* Población total, masculina y femenina
* Promedio de ingresos per cápita
* Distribución por grupos etarios
* Porcentaje de población rural y urbana
* Personal de salud disponible
* Matriz de correlación entre morbilidades

Esta información es fundamental para que la empresa comprenda el contexto socioeconómico y sanitario de los países evaluados, lo cual es clave para definir estrategias de expansión y desarrollo de negocio.

La hoja “Datos Pandemia” muestra un análisis detallado sobre la gestión y evolución de la pandemia en cada país. Se incluyen:

* Series temporales sobre casos confirmados, decesos y recuperados
* Acumulados por variable
* Mapa temático que refleja la distribución geográfica de los casos, decesos y recuperados como acumulados
* Gráfico de barras con el nivel de vacunación segmentado por grupo etario

Este dashboard, sustentado en el análisis previo realizado con Python, permite consolidar la información y facilita una interpretación visual clara y precisa. Todo lo anterior contribuye a elaborar conclusiones valiosas que serán presentadas a la dirección de la empresa, proporcionando una base sólida para la toma de decisiones estratégicas basadas en datos.

# Conclusiones y Recomendaciones

# Tras la finalización del estudio exploratorio basado en el dataset proporcionado, se concluye que los países más adecuados para la expansión estratégica de la empresa BIOGENESYS en Latinoamérica son: México, Colombia, Perú y Brasil.

# Estos países presentan en conjunto:

# Las tasas de letalidad más elevadas por COVID-19

# Altos niveles de casos confirmados y decesos acumulados

# Un bajo porcentaje de vacunas administradas en relación con su población

# Estos indicadores sugieren que las políticas de vacunación en dichos países han sido insuficientes o no han contado con los recursos necesarios para enfrentar eficazmente la pandemia. Por tanto, la presencia de BIOGENESYS podría contribuir significativamente a optimizar la respuesta sanitaria y mejorar el acceso a la vacunación, con potencial impacto positivo en los indicadores de salud pública a mediano y largo plazo.

Otras conclusiones adicionales serían las siguientes:

* Considerar a Chile como modelo de buenas prácticas para replicar en otros países de la región.
* Integrar variables de salud preexistente (diabetes y tabaquismo) en la planificación sanitaria.
* Fortalecer el monitoreo de datos temporales para anticipar brotes y optimizar la distribución de recursos.
* Enriquecer el dataset con información adicional de recuperados y datos regionales más granulares para mejorar la precisión del análisis.

# Reflexión personal

A lo largo del desarrollo de este proyecto integrador, he atravesado un proceso de aprendizaje profundo y desafiante. Trabajar con un dataset tan grande y complejo me permitió poner en práctica una gran parte de los conocimientos adquiridos a lo largo de la carrera, especialmente en cuanto a limpieza, transformación y análisis de datos en Python.

Personalmente, todavía me resulta retador realizar un análisis completo en Python, sobre todo cuando se trata de construir visualizaciones claras e intuitivas. Aunque reconozco el enorme potencial de esta herramienta para el tratamiento de datos y la generación de métricas, siento que Power BI me resulta mucho más accesible para la parte visual, ya que me permite interactuar de forma más rápida y precisa con los datos. Esta preferencia también se debe a que ya había trabajado previamente con Power BI, por lo que su lógica me resulta más familiar.

Más allá de los desafíos técnicos, valoro profundamente la experiencia que este proyecto significó. Requirió muchas horas de trabajo, organización y análisis consciente, pero al mismo tiempo me permitió afianzar herramientas, descubrir nuevas metodologías y, sobre todo, comprobar cuánto he crecido en este recorrido académico. Me voy con la certeza de que este es solo el comienzo, y con muchas ganas de seguir aprendiendo y perfeccionándome en el fascinante mundo del análisis de datos.