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

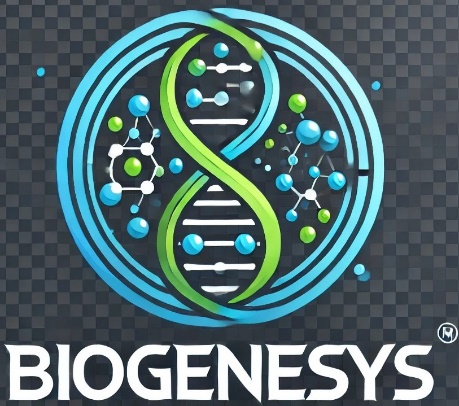
**Nombre del autor:** Fernando Gómez

**Email:** fernandomgomez8@gmai.com

**Cohorte:** DAPT-05

**Fecha de entrega:** 24/02/2025

**Institución:** Biogenesys

****

# Introducción

# El proyecto "Expansión Estratégica de laboratorios para investigación y vacunación de COVID-19" tiene como propósito identificar las ubicaciones más óptimas para la expansión de laboratorios farmacéuticos en Latinoamérica, específicamente en Colombia, Argentina, Chile, México, Perú y Brasil, para optimizar la respuesta a la pandemia y postpandemia mejorando el acceso a las vacunas.

# Este proyecto de análisis de datos incluye la realización de un análisis exploratorio de los datos, la aplicación de técnicas de limpieza y organización de datos, la mejora del acceso a los datos mediante operaciones ETL, y el desarrollo de dashboards interactivos para proporcionar una representación visual clara, concisa y fácilmente interpretable de la información clave, para apoyar la toma de decisiones informadas.

# Desarrollo del proyecto

### Metodología de carga, transformación y limpieza de datos con Python

Para la carga, transformación y limpieza de datos mediante lenguaje de programación Python (y sus librerías Numpy y Pandas) se leyó el archivo “data\_latinoamerica.csv” mediante pandas (pd.read\_csv) y luego se siguieron los siguientes pasos:

1. **Toma de contacto con los datos**: en un primer acercamiento a los datos disponibles, se comenzó por la lectura del archivo ”readme.txt”, que contiene un diccionario de datos donde se explica el contenido de cada columna del dataset.
2. **Verificación de la estructura del dataset**: se comprobó que el dataset tuviera la cantidad indicada de filas y columnas (12.216.057 y 50, respectivamente).
3. **Identificación de nulos**: se identificaron los totales de nulos de cada columna, y luego se estableció un criterio para identificar columnas con una cantidad de nulos mayor a 4 millones, resultando 26 columnas que verificaban dicho criterio.
4. **Filtrado de datos por países de interés:** se filtraron los datos para incluir únicamente los países de interés para el análisis: Argentina, Brasil, Chile, Colombia, México y Perú. Este filtrado se hizo por medio de la columna “location\_key”.
5. **Filtrado por fecha:** se estableció como parámetro que las fechas de la columna “date” sean mayores al 01-01-2021, dado que el volumen de datos más completo y relevante se registra a partir de dicha fecha. El dataset se redujo a 3,744 registros o filas y 50 columnas.
6. **Tratamiento de nulos remanentes:** se verificaron nuevamente los totales de nulos para cada columna. A continuación, se establecieron dos criterios para rellenar las columnas con datos faltantes:
   * Las 9 columnas con menos de 100 nulos se rellenaron con la media de la correspondiente columna.
   * Para las 3 columnas que tenían más de 100 nulos, se rellenaron los nulos con 0.

Luego de esto se verificó que efectivamente ya no hubiera nulos en el dataframe.

1. **Exportación a archivo .csv:** finalmente se exportó el dataframe depurado a un archivo.csv llamado “DatosFinalesFiltrados” para leerlo luego con Pandas y trabajar sobre él desde este punto en adelante.

### Conclusión sobre Transformaciones y Limpieza de Datos:

Luego de aplicar los procesos mencionados, el dataset se redujo a 3,744 registros o filas y 50 columnas, manteniendo su integridad estructural. Las transformaciones realizadas, como el filtrado y el rellenado de nulos, fueron claves para asegurar la calidad y consistencia de los datos, permitiendo un análisis confiable, claro y fundamentado. Estas medidas garantizaron que los datos analizados fueron representativos, relevantes y aptos para su análisis exploratorio y visualización interactiva en dashboards, facilitando una toma de decisiones informada para la expansión estratégica de los laboratorios de la firma “Biogenesys”.

# EDA e insights obtenidos mediante el Análisis Exploratorio de Datos y las Visualizaciones

#### Primer análisis y preparación de los Datos

En primer lugar, lo que se hizo fue obtener una descripción de las principales medidas estadísticas de cada una de las columnas del dataset (tanto numéricas como categóricas).

A continuación, se colocó la columna “date” como índice para facilitar análisis temporales, por lo que el dataset pasó a tener un índice y 49 columnas.

#### Matriz de correlación

Se creó una matriz de correlación de variables numéricas para obtener un panorama general de la correlación entre todas las columnas de nuestro dataframe. Para simplificar su visualización e interpretación, se decidió mostrar solo las correlaciones con un valor absoluto mayor o igual a 0,5. Además se ocultó el triángulo superior de la matriz, es decir, el que se encuentra sobre la diagonal principal.

#### Gráficos de columnas de variables numéricas

Se crearon gráficos de barras para representar los valores promedio de las principales variables numéricas para cada país.

De estas gráficas una de las informaciones más importantes que se obtuvo es que sólo Colombia y Brasil cuentan con datos de recuperados.

#### Gráficos para el análisis de la distribución de frecuencias de variables numéricas

Se graficaron histogramas (histplots), gráficos de cajas (boxplots) y gráficos de violín (violinplots) para tener un panorama general de la distribución de frecuencias de cada una de las variables numéricas de interés.

#### Correlación entre temperatura y casos de COVID-19

Se evidenció la relación entre la temperatura promedio de cada país y el incremento de casos y muertes por covid-19 mediante gráficos de dispersión (scatterplot). En estos gráficos puede observarse que la mayor cantidad de casos nuevos y muertes se dieron con temperaturas de entre 25 y 30 °C.

Chile y Argentina presentaron las temperaturas promedio más bajas, manteniéndose estables en términos de casos nuevos y muertes.

Esto probablemente podría observarse mejor con gráficos de líneas.

#### Análisis temporal de la vacunación y distribución etaria de la población

En un primer momento se hizo un gráfico de columnas para mostrar la distribución etaria del total de la población de Latinoamérica. Luego, se realizaron gráficos de líneas para cada grupo etario donde se muestran los promedios de las dosis de vacunas administradas a lo largo del tiempo en cada país.

De estos gráficos se deduce que, en todos los casos, en general las dosis administradas han ido en aumento, se observa también un hecho llamativo y es que México que era el segundo país que mayor cantidad de dosis mensuales contabilizaba (como cabría esperar por la población que posee), mostró un fuerte descenso hacia abril de 2022 donde llegó a ser el peor posicionado en este aspecto junto a Colombia. Por el contrario, Argentina que era uno de los países que acumulaba menos dosis mensuales, en abril de 2022 comienza un marcado aumento que dos meses después lo sitúa en segundo lugar, después de Brasil.

#### Análisis de la tasa de mortalidad por país

Mediante un gráfico de columnas se mostró la tasa de mortalidad por Covid-19 para cada país, siendo Perú el que mostró una tasa de mortalidad ampliamente mayor al resto de los países. El siguiente país con mayor tasa de mortalidad es Brasil, pero presentando valores más cercanos al resto de los países.

#### Análisis de la tasa de mortalidad masculina vs femenina

Mediante un gráfico de columnas apiladas se representó, para cada país, el contraste entre la tasa de mortalidad masculina y la femenina por covid-19, evidenciando que en todos los países la tasa de mortalidad masculina es cercana al doble de la femenina.

#### Mapa de calor de indicadores demográficos y de salud

Se realizó un mapa de calor para analizar la correlación entre diferentes indicadores demográficos y de salud, como densidad de población, muertes por covid-19, prevalencia de diabetes, etc.

Se encontró que la prevalencia de diabetes tiene cierta correlación positiva con el acumulado de muertes por covid-19.

#### Análisis de la relación entre prevalencia de diabetes y tasa de mortalidad

Se realizó un gráfico de columnas superpuestas a fines de visualizar una comparativa entre la prevalencia de diabetes y la tasa de mortalidad por Covid-19. No se pudo extraer ninguna conclusión al respecto de este gráfico, ya que no se aprecia una relación claramente observable.

# Análisis de la relación entre prevalencia de tabaquismo y tasa de mortalidad

De la misma manera que en el punto anterior, se realizó otro gráfico de columnas superpuestas para comparar la prevalencia de tabaquismo y la tasa de mortalidad por Covid-19. En el gráfico se observó que no hay una relación clara observable entre ambas variables. Por ejemplo, Perú tiene la menor prevalencia de fumadores y la mayor tasa de mortalidad por Covid-19.

# Análisis temporal de nuevos casos y nuevas muertes para cada año en estudio

Para cada año contemplado en el estudio, es decir, 2021 y 2022 se realizaron gráficos de líneas que muestran la evolución semanal de nuevos casos confirmados y nuevas muertes en cada uno de los países. Esto mismo se repitió en otros gráficos de líneas, pero mostrando la evolución mensual. Finalmente se hizo un gráfico para cada variable, pero mostrando la evolución mensual contemplando los dos años del estudio conjuntamente.

De estos gráficos se deduce que:

* La mayor parte tanto de nuevos casos como de muertes se da en los primeros meses de cada año, para luego descender marcadamente y tener otro pico (aunque considerablemente menor) a mediados de cada año.
* Los picos de nuevos casos fueron bastante más altos en el año 2022. Por el contrario, los picos de muertes fueron más altos en 2021. Esto podría explicarse por la cuarentena más estricta aplicada en 2021 (menos contagios) y los efectos de la campaña de vacunación (menos muertes en 2022).
* En el gráfico semanal de muertes del año 2022 se observa un outlier en Chile que alcanza un pico muy marcado de 12000 muertes.

# Progreso de la vacunación y estrategias de vacunación por país

Se realizó un gráfico de barras para mostrar el avance de la vacunación, como el acumulado de dosis de vacunas administradas en cada país y un gráfico de columnas para comparar la estrategia de vacunación de cada país, entendida como la cantidad de dosis promedio administrada ´por habitante.

De estos gráficos se concluye que Brasil y luego México fueron los países en que más dosis de vacunas se han aplicado en total, sin embargo, son Chile y luego Perú los que muestran un mayor promedio de vacunas aplicadas por habitante.

# Análisis de la relación entre temperatura promedio y nuevos casos en Latinoamérica

Se grafico un gráfico de líneas de dos variables, temperatura promedio y nuevos casos de Covid-19 a lo largo del tiempo en Latinoamérica, para intentar encontrar una posible relación entre ambas.

En dicho gráfico puede verse que en general al disminuir la temperatura promedio aumentan los nuevos casos, excepto por el pico que se da aproximadamente en febrero de 2022. En ese punto se da un pico tanto de temperatura promedio como de casos.

# Urbanización y nuevos casos confirmados por país

Se realizó un gráfico de barras apiladas para mostrar la relación entre el porcentaje de población urbana y el porcentaje de población que representan los casos confirmados en cada país.

En este gráfico se observa un hecho llamativo: de los dos países más urbanizados que son México y Chile, uno de ellos (México) presenta uno de los porcentajes más bajos de casos confirmados acumulados mientras que el otro (Chile) presenta el porcentaje más alto de casos confirmados acumulados. Si tomamos en cuenta los demás países, en general se da que a mayor porcentaje de población urbanizada mayor porcentaje de casos confirmados.

# Distribución etaria de la población de cada país

Se realizó un gráfico de torta para cada país a fines de mostrar el porcentaje de su población que corresponde a cada grupo etario (tomando intervalos de 10 años).

En estos gráficos se observa que en todos los países aproximadamente la mitad de la población tiene menos de 40 años.

# Análisis temporal de la tasa de mortalidad del Covid-19

Se realizó un gráfico de líneas que muestra el avance temporal de la tasa de mortalidad del Covid-19 (acumulado de muertes/población total) para cada uno de los países.

En dicho gráfico se revelan varios hechos interesantes/llamativos: se observa que en México en enero de 2021 la tasa de mortalidad bajo abruptamente hasta casi 0, para luego volver a aumentar lentamente. Se observa también que en Argentina, Brasil y Colombia la tasa de mortalidad tuvo un comportamiento muy similar. Se observa además que la curva de mortalidad en Chile venía teniendo un comportamiento similar a Argentina, Brasil y Colombia hasta febrero de 2021 cuando se da un punto de inflexión y pasa a tener la mortalidad más alta después de Perú.

# Análisis temporal de la tasa de letalidad del Covid-19 por país

Se realizó un gráfico de líneas que muestra el avance temporal de la tasa de letalidad del Covid-19 (acumulado de muertes/acumulado de casos confirmados) para cada uno de los países.

En dicho gráfico se observa que Argentina y Chile presentan las menores tasas de letalidad. Además, se observa que México y Perú presentan tasas de letalidad significativamente más altas que el resto de los países y que las curvas de ambos países presentan un punto de inflexión en enero de 2021, donde bajan abruptamente. También se puede hacer una observación más global de las curvas de todos los países y concluir que la tasa de letalidad del Covid-19 ha ido disminuyendo con el paso del tiempo.

# Análisis temporal global de la tasa de letalidad del Covid-19 en Latinoamérica

Se realizó un gráfico de líneas para mostrar la variación temporal de la tasa de letalidad promedio de toda Latinoamérica.

En el gráfico se ve que la letalidad global de Latinoamérica fue decreciendo de manera constante, produciéndose un descenso abrupto en enero de 2022. Esto marca un claro éxito de la campaña de vacunación a nivel global de Latinoamérica.

# Relación entre la cobertura de vacunación y la reducción de casos en Latinoamérica

Se hizo un gráfico de líneas que muestra la variación mes a mes de dos variables: los nuevos casos confirmados y el porcentaje de población vacunada.

En dicho gráfico se observa que al principio de la serie temporal analizada la vacunación y los nuevos casos crecían a la par, luego de mayo de 2021 mientras crecía la vacunación disminuían los nuevos casos.

En enero de 2022 la vacunación continuaba creciendo, pero de igual manera se dio el mayor pico de casos nuevos, que luego comenzaron a descender de la misma manera abrupta que aumentaron y en abril de 2022 los casos nuevos ya se estabilizaron nuevamente en los mismos valores previos al pico de enero de ese mismo año.

En julio de 2022 se da otro pico de casos (aunque mucho más leve que el de enero) más allá de que la vacunación seguía en aumento.

# Relación entre la cobertura de vacunación y la reducción de la tasa de letalidad en Latinoamérica

Se realizó un gráfico de líneas donde se muestra de forma conjunta la evolución mes a mes de ambas variables: porcentaje de población vacunada y tasa de letalidad (porcentaje de casos letales).

De esta manera se mostró de manera aún más evidente la eficiencia de la campaña de vacunación.

# Evolución de casos activos vs. recuperaciones en Latinoamérica

Se realizó un gráfico de líneas que muestra de forma conjunta la evolución mes a mes de casos activos (casos confirmados-muertes-recuperados) y de recuperados.

Del gráfico se puede hacer una lectura general y decir que al aumentar el número de casos activos también fue aumentando el de recuperados, lo que evidencia que de los casos activos cada vez más pacientes terminaron por recuperarse.

# Análisis del dashboard

# El dashboard creado en Power BI, tiene una forma de navegación muy intuitiva, con fácil acceso a la información más relevante obtenida durante el análisis exploratorio. El mismo cuenta con una portada con botones para acceder a cualquiera de las páginas que conforman el informe, cuenta también con flechas de navegación que permiten al usuario moverse entre sus diferentes páginas y tiene además una flecha encerrada en un contenedor circular que permite regresar a la portada desde cualquier página del informe.

# El dashboard está diseñado teniendo en cuenta técnicas básicas de storytelling, como el patrón Z, y utiliza una paleta de colores sobrios y adecuada a cada gráfico para hacerlo más fácilmente comprensible. Cada país que forma parte del análisis se representa con un color que se mantiene constante en todos los gráficos que muestran variables por país, lo que facilita su identificación.

# La información expuesta en el dashboard concentra los gráficos considerados más relevantes de todo el análisis exploratorio y los divide en 7 temáticas (una por página) para facilitar su comprensión y asimilación. Estas temáticas son: panorama general, tendencias temporales, vacunación, letalidad, análisis por temperatura, población y prevalencia de enfermedades.

# Todas las páginas del informe cuentan con filtros (ubicados en el lateral izquierdo) que permiten filtrar los datos por año, mes y país.

# Conclusiones y Recomendaciones

**Conclusión General**

El EDA (análisis exploratorio de datos) y las visualizaciones interactivas en el dashboard evidenciaron una clara necesidad de ampliar la cantidad de laboratorios farmacéuticos en ciertos países de Latinoamérica. Todo el análisis reveló diferencias importantes en el impacto del COVID-19 y las coberturas de vacunación entre los países en estudio. Estos hallazgos son claves para formular estrategias de expansión de laboratorios farmacéuticos con el objetivo de optimizar la respuesta a los efectos de la pandemia y postpandemia mejorando el acceso a las vacunas.

#### Conclusiones particulares (estratégicas)

* La mayor tasa de letalidad del Covid-19 en Latinoamérica se presenta en México (aprox. 5%), siendo en este país notablemente más alta que en el resto de los países.
* La mayor parte tanto de nuevos casos como de muertes se da en los primeros meses de cada año, para luego descender marcadamente y tener otro pico (aunque considerablemente menor) a mediados de cada año.
* La tasa de mortalidad femenina es considerablemente menor a la masculina.
* Los picos de nuevos casos fueron bastante más altos en el año 2022. Por el contrario, los picos de muertes fueron más altos en 2021. Esto podría explicarse por la cuarentena más estricta aplicada en 2021 (menos contagios) y los efectos de la campaña de vacunación (menos muertes en 2022).
* De estos gráficos se concluye que Brasil y luego México fueron los países en que más dosis de vacunas se han aplicado en total (lo cual es esperable porque son los países con mayor población), sin embargo, son Chile y luego Perú los que muestran un mayor promedio de vacunas aplicadas por habitante.
* La mayor urbanización y la densidad poblacional favorecen el contagio y por ende la aparición de mayor cantidad de nuevos casos de Covid-19.
* Las mayores tasas de letalidad del Covid-19 en Latinoamérica son la de México y Perú.
* Existe una clara relación inversamente proporcional entre la cobertura de vacunación y la tasa de letalidad por Covid-19.
* Colombia, seguido por Perú y México son los países con menor cantidad de enfermeros per cápita.
* México, perú y Colombia son los tres países de los seis en estudio que presentan los menores ingresos per cápita e índice de desarrollo humano.

#### Respuesta a la consigna principal del Proyecto

Si bien sería recomendable expandir laboratorios farmacéuticos para favorecer la vacunación contra el Covid-19 en toda Latinoamérica, se considera más urgente hacerlo en México, siguiendo por Colombia, Brasil en el orden de prioridad.

Esta decisión se debe a que este país presenta una combinación de los siguientes factores: la mayor tasa de letalidad del Covid-19 de la región, una densidad de población y urbanización muy importantes, y una tasa de vacunación promedio menor a dos (2) dosis por habitante.

Los siguientes países en orden de prioridad para la expansión de laboratorios farmacéuticos son Colombia y Brasil, respectivamente. Esta afirmación se basa en las siguientes conclusiones arrojadas por los datos analizados: estos países tienen el menor promedio de dosis de vacunas aplicadas por habitante, además Colombia tiene el menor número de enfermeros per cápita de la región y los índices más bajos de ingreso per cápita y de desarrollo humano, así como una alta densidad poblacional. Tanto Colombia como Brasil presentan altas tasas de mortalidad y letalidad por Covid-19. Por su parte Brasil cuenta con grandes urbes que favorecerían el contagio de la enfermedad, además de ser considerable la prevalencia de otra enfermedad como la diabetes.

# Reflexión personal

Durante el desarrollo de este proyecto, he adquirido un conocimiento más profundo en áreas clave relacionadas con el análisis de datos, especialmente en el ámbito de la salud pública y la gestión de pandemias, lo cual me resultó de sumo interés. Entre los aprendizajes más destacados como analista de datos puedo listar los siguientes:

* Gestión de grandes volúmenes de datos: manejar conjuntos de datos extensos requirió un enfoque riguroso para preservar la calidad y exactitud de la información.
* Análisis exploratorio de datos (EDA): este paso fue esencial para detectar patrones, identificar tendencias y reconocer valores atípicos, estableciendo una base sólida para análisis más avanzados.
* Procesamiento y transformación de datos: comprobé la importancia de realizar procesos de limpieza y transformación de datos para garantizar resultados confiables y representativos.
* Visualización de datos: diseñé un dashboard interactivo y gráficos que facilitaron la interpretación de datos complejos, ayudando a comunicar los resultados de manera clara y concisa.
* Toma de decisiones basada en datos: la integración de información diversa me permitió formular recomendaciones estratégicas respaldadas por evidencia concreta.

Si tuviera la oportunidad de comenzar este proyecto nuevamente, haría una planificación Inicial más exhaustiva y a conciencia, incluso me empaparía más del lenguaje y de la forma de lectura y comprensión de los indicadores básicos del ámbito de la salud pública.

Por otro lado, también consideraría sumar datos adicionales, como indicadores socioeconómicos y demográficos más detallados, para enriquecer el análisis y darle más valor a los insights hallados.

# EXTRA CREDIT

Durante el desarrollo del proyecto he cumplido con los siguientes puntos del “extra credit”:

* Explorar el uso de funciones de orden superior para una manipulación eficiente de los datos: mediante el uso de funciones de orden superior, fue posible realizar un análisis exploratorio de datos (EDA) más profundo y completo. La función groupby() desempeñó un papel fundamental a lo largo del proyecto, facilitando la agrupación de los datos según diversas variables clave, lo que permitió llevar a cabo un análisis más preciso y detallado.

Crear mapas utilizando la información de latitud y longitud disponible en tus datos: se creó utilizando “Folium”, un mapa ajustable donde se marcaron puntos para cada país incluido en el análisis: Argentina, Brasil, Chile, Colombia, México y Perú. Cada punto se representó con diferentes tonalidades de color de acuerdo a los mapas de calor indicados (confirmados, muertos, recuperados, vacunas administradas). Si bien se cumplió con este punto para el trabajo, el mapa no resulta demasiado útil para diferenciar y comprender la incidencia de las diferentes variables que se quisieron representar sobre él.

* Se utilizaron funciones de inteligencia de tiempo para realizar gráficos de tendencia temporal en PowerBI, incluso se creó una tabla de fechas a partir de la fecha más antigua y la más reciente incluida en el análisis, y se activó su relación con la tabla de hechos en el modelo.