**Informe para la empresa Biogenesys Laboratory**

**Nombre del autor: María Verónica Sofía Villagra Vélez**

**Email: verovillagrav@gmail.com**

**Cohorte: DA\_PT02**

**Fecha de entrega:** cuándo se completó y entregó el informe

**Institución:**

La empresa farmacéutica BIOGENESYS tiene como objetivo identificar las ubicaciones más adecuadas para expandir sus laboratorios, utilizando un análisis detallado de la incidencia de COVID-19, las tasas de vacunación y la disponibilidad de infraestructuras sanitarias. Su propósito es optimizar la respuesta frente a los efectos de la pandemia y la postpandemia, mejorando así el acceso a las vacunas y fortaleciendo la capacidad de respuesta ante futuras emergencias sanitarias.

Logotipo

Descripción generada automáticamente

# Introducción

El proyecto se centró en brindar asistencia al Laboratorio Biogenesys mediante un estudio y análisis de la situación del COVID-19, con el objetivo de apoyar la toma de decisiones para la estrategia de expansión en Latinoamérica, específicamente en Colombia, Argentina, Chile, México, Perú y Brasil. Comprender el contexto de cada país y su respuesta ante la pandemia es crucial para diseñar planes que mejoren el acceso a vacunas en el futuro.

# Desarrollo del proyecto

# El proyecto comenzó con la carga de un archivo CSV denominado ‘data\_latinoamerica’ en Visual Studio Code para su visualización y análisis con Python. Para esto, instalamos e importamos las librerías necesarias como Numpy, Pandas, Matplotlib y Seaborn, que fueron fundamentales para realizar cálculos, transformaciones y visualizaciones. Además, creamos un archivo requirements.txt para documentar las librerías y sus versiones.

# El dataset original que nos proporcionó el Laboratorio Biogenesys tenía más de 22 millones de registros y 707 columnas, pero recibimos una versión acotada a los países de Latinoamérica, con aproximadamente 12 millones de registros y 50 columnas. Al cargar el archivo, generamos un DataFrame en Jupyter Notebook que contenía 12,216,057 filas y 50 columnas.

# Como nuestro análisis se centra en seis países específicos —Argentina, Brasil, Chile, Colombia, México y Perú—, filtramos el dataset para trabajar únicamente con los datos de estos países, reduciendo el DataFrame a 11,970,289 registros. Este filtro se aplicó porque dichos países son el foco de nuestra estrategia de expansión.

# En una primera etapa, realizamos comprobaciones iniciales para verificar la cantidad de filas y columnas, los nombres de las mismas, y la existencia de valores faltantes o nulos. Descubrimos columnas con más de 2.5 millones de datos faltantes, lo que nos llevó a desarrollar una estrategia para manejar estos valores y mejorar la calidad de los datos. Eliminamos las filas con location\_key que no coincidían con los países de interés (AR, BR, CL, CO, PE, MX) y nos deshicimos de registros zonales o repetidos.

# Además, filtramos los datos para incluir solo aquellos registrados a partir del 1 de enero de 2021, lo que redujo el dataset a 3,744 filas. En este punto, decidimos abordar los valores nulos. Las columnas con más del 50% de datos faltantes, como ‘new\_recovered’ y ‘cumulative\_recovered’, fueron reemplazadas por cero. Para el resto de las columnas, optamos por reemplazar los faltantes con la media.

# Una vez completada la limpieza, realizamos análisis estadísticos preliminares. Utilizamos funciones como .head() y .describe() para obtener medidas estadísticas clave como el conteo, la media, la desviación estándar, entre otras. También creamos una función personalizada para calcular la mediana, varianza, rango y moda en las columnas numéricas. Estas primeras exploraciones revelaron alta dispersión de los datos en varias columnas, por lo que recomendamos un análisis cuidadoso.

# Finalmente, guardamos el nuevo DataFrame en un archivo CSV denominado DatosFinalesFiltrado. Optamos por corregir los datos faltantes en lugar de eliminarlos, lo que nos permitió mejorar la calidad del dataset y facilitar su análisis en futuras fases del proyecto. El objetivo general de este proceso fue garantizar la integridad y confiabilidad de los datos para tomar decisiones estratégicas informadas.

# 

# EDA e insights

# Análisis Exploratorio de Datos (EDA) e Insights:

# Con los datos ya limpios, realizamos un análisis exploratorio de datos (EDA) para obtener insights clave que orienten la planificación estratégica de expansión de la empresa. Nuestro objetivo es identificar patrones, tendencias y anomalías en los datos relacionados con la incidencia de COVID-19, las tasas de vacunación y la disponibilidad de infraestructura sanitaria. Esto permitirá a los directivos tomar decisiones informadas sobre la ubicación de nuevos laboratorios y centros de vacunación, optimizando el impacto en la salud pública.

# Comenzamos ajustando la columna de fechas (date) como índice, lo que facilitó el análisis temporal. Posteriormente, generamos una matriz de correlación visualizada con un heatmap, revelando correlaciones fuertes entre variables como:

# Cumulative\_confirmed y Cumulative\_deceased

# Cumulative\_recovered y Cumulative\_confirmed

# GDP\_per\_capita\_usd y Human\_development\_index

# Area\_urban\_square\_kilometer y Population\_largest\_city

Imagen que contiene Escala de tiempo

Descripción generada automáticamente

Construimos gráficos de barras para analizar columnas tales como: new\_confirmed’, ‘new\_deceased’, ‘population’, ‘population\_male’, ‘population\_female’, ‘population\_rural’, 'population\_urban’, que dieron cuadros parecidos a éste (Con Brasil sobresaliendo debido a su mayor población):

Gráfico, Gráfico de barras

Descripción generada automáticamente Gráfico, Gráfico de barras

Descripción generada automáticamenteGráfico, Gráfico de barras

Descripción generada automáticamente

# También generamos gráficos para diferentes rangos etarios y variables como el GDP per cápita y la población en áreas rurales y urbanas. Estos gráficos revelaron que Brasil y México tienen las mayores poblaciones, lo que sugiere un mercado prometedor. México y Colombia destacan por su alta densidad poblacional, mientras que Chile, con su elevado PIB per cápita, y Argentina, con una alta tasa de enfermeros y médicos respectivamente por cada 1000 habitantes, también deben ser considerados en la estrategia de expansión;

Gráfico, Gráfico de barras

Descripción generada automáticamente Gráfico, Gráfico de barras

Descripción generada automáticamente

Gráfico, Gráfico de barras

Descripción generada automáticamente

Gráfico, Gráfico de barras

Descripción generada automáticamente Gráfico, Gráfico de barras

Descripción generada automáticamente Gráfico, Gráfico de barras

Descripción generada automáticamente Gráfico, Gráfico de barras

Descripción generada automáticamente

# Un análisis con gráficos histplot mostró distribuciones asimétricas en columnas como cumulative\_confirmed, con una gran dispersión y varios outliers. También utilizamos scatterplots para estudiar la relación entre la temperatura y los nuevos casos confirmados, sin encontrar una correlación fuerte.

Gráfico, Histograma

Descripción generada automáticamente

# Realizamos análisis temporales para visualizar la evolución de las dosis de vacunas administradas por mes en cada país, así como el número de casos confirmados, muertes y recuperaciones. Estos gráficos sugieren que Chile gestionó eficazmente la vacunación y prevención, destacándose en comparación con los otros países.

# Además, creamos boxplots para analizar la variabilidad de la temperatura promedio y violinplots que indicaron una alta dispersión de casos en países como Brasil y Argentina, sugiriendo la presencia de picos o brotes importantes en ciertos períodos.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

# También examinamos la distribución etaria de la población mediante gráficos de barras, y analizamos la vacunación por mes según rangos etarios. Un barplot comparativo mostró que la tasa de mortalidad es más alta en hombres que en mujeres.

Gráfico, Gráfico de barras

Descripción generada automáticamente

# Generamos una matriz de correlación de indicadores demográficos y de salud, que reveló que el fumar y la diabetes son factores preexistentes clave en la mortalidad por COVID-19. Esto fue respaldado por barplots que analizaron la relación entre el hábito de fumar, la diabetes y la tasa de muertes por COVID, mostrando que ambos factores pueden agravar la enfermedad.

# En conclusión, el análisis sugiere que Brasil, México y Colombia son los países que requieren mayor asistencia, ya que han registrado el mayor número de casos confirmados y muertes, además de tener un menor suministro de vacunas en relación con su población.

# Análisis del dashboard

# Portada: Mi dashboard consta de una portada con una breve descripción del trabajo y con siete botones que nos permitirán navegar a lo largo del Dashboard.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Mapa

Descripción generada automáticamente Interfaz de usuario gráfica, Sitio web

Descripción generada automáticamente

Interfaz de usuario gráfica, Aplicación, Gráfico de barras

Descripción generada automáticamente Interfaz de usuario gráfica, Gráfico, Aplicación

Descripción generada automáticamente Escala de tiempo

Descripción generada automáticamente

# Conclusiones y Recomendaciones

# [Realiza una síntesis de los resultados obtenidos tanto con las visualizaciones como con el dashboard (una conclusión general teniendo en cuenta todo el proceso). Además deberás enumerar las conclusiones estratégicas considerando, por ejemplo, las políticas de vacunación por país, inversión en salud, habitantes, densidad, problemas sociales y económicos, etc. NO TE OLVIDES DE INCLUIR LA RESPUESTA PRINCIPAL DEL PROYECTO: identificar las ubicaciones óptimas para la expansión de laboratorios farmacéuticos]

# Reflexión personal

# [Breve reflexión sobre lo aprendido durante el proyecto y las habilidades adquiridas como Analista de Datos. Además, reflexiona sobre la siguiente pregunta: Si tuvieras que volver a empezar este proyecto ¿Lo harías de la misma manera? ¿Cambiarías algo?]

Feliz por lo aprendido y por el recorrido hecho. Muy orgullosa de mi trabajo final que me costo mucho tiempo, frustraciones, lagrimas, estrés, desvelos. Siempre digo que cuando mas conocemos sobre algo, más nos damos cuenta de lo mucho que desconocemos. Nos damos cuenta que sabemos muy poquito para los programas que hemos vistos.

Y en cuanto a como lo haría si no lo tendría que hacer de vuelta, lo hice en varias partes y lo volví hacer diversas ocasiones. Desde haber hecho aproximadamente unas siete veces en donde trabaje en Visual Studio Code ( que no me fue desconocido porque lo habia trabajado con otros programas), por algún extraño motivo se me ponía error en la parte final del avance tres, por lo que lo hice varias veces y en esos procesos de rehacerlos los fui borrando y perdí información valiosa.

Y en cuanto a la presentación en power bi, me gusto mucho el enganche, pero lo sufrí en cuestión que se me desprogramaba cuando subía mi archivo csv se me desprograma, lo programe luego en power query manualmente, pero me seguían dando en algunos gráficos incorrectos por lo que fue muy frustrante. Y luego vi que se podía cargar el archivo cvs de manera “personalizada” o algo así no recuerdo el término. Cargue nuevamente mi archivo y luego me puse a dividir las columnas a partir de la coma, y luego reemplace el punto por la coma y le agregue formato a mi archivo para trabajarlo. Es esta acción definitivamente ahorraría mucha frustración.

Y luego fue más sencillo la parte de armar el dashboard, si me ponía nerviosa la cuestión de tomar mis propias decisiones y saber si es lo correcto o no.

# 