**EXPANSIÓN ESTRATÉGICA DE BIOGENESYS 🚀**

**Nombre del autor: Bianca De Robertis**

**Email: biancaderobertiss19977@gmail.com**

**Cohorte: DA-PT02**

**Fecha de entrega: 16/09/2024**

**Institución:** Empresa Farmaceútica BIOGENESYS

# Introducción

# La empresa farmacéutica BIOGENESYS busca identificar las ubicaciones óptimas para la expansión de sus laboratorios, basándose en un análisis de datos de incidencia de COVID-19, tasas de vacunación y la disponibilidad de infraestructuras sanitarias. Su meta es optimizar la respuesta a los efectos de la pandemia y postpandemia con el fin de mejorar el acceso a vacunas.

Como Data Analyst de este proyecto, debo realizar un estudio que ayudará a la estrategia de expansión en Latinoamérica, específicamente en los países de Argentina, Brasil, Chile, Colombia, México y Perú. La dirección ha propuesto hacer esta inversión para ubicar regiones y recolectar datos que sean de importancia para la toma de decisiones informadas.

Los objetivos son concretos y medibles:

* Realizar un análisis exploratorio de datos sobre la incidencia de COVID-19 y otros factores relevantes, identificando tendencias y oportunidades mediante estadísticas, mediciones y visualizaciones.
* Aplicar técnicas de limpieza de datos para asegurar la calidad de los datos, facilitando análisis y decisiones estratégicas confiables.
* Mejorar el acceso a los datos mediante operaciones eficientes de extracción, transformación y carga (ETL), aumentando la eficacia del análisis y la toma de decisiones.
* Desarrollar dashboards interactivos con visualizaciones eficientes, permitiendo explorar datos desde múltiples perspectivas para una toma de decisiones informada y estratégica.

Se utilizarán los datos proporcionados para ofrecer insights valiosos sobre posibles ubicaciones estratégicas para futuros laboratorios y centros de vacunación. Este análisis de tendencias y demografía será fundamental para comprender el entorno del mercado en los países de interés.

# Desarrollo del proyecto

# En el primer avance, nos centramos en el dataset proporcionado por Ingeniería de Datos de 12,216,057 filas y 50 columnas en formato de CSV, facilitando la manipulación y el análisis del dataset original. Además, contamos con un archivo adicional que contiene información detallada sobre las columnas seleccionadas para una adecuada preparación y limpieza de datos.

# Una vez que adquirimos mayor conocimiento del dataset,creamos un notebook en Python e importamos las librerías vistas y necesarias para encarar este análisis. Leemos el archivo csv con código Python en Visual Studio Code y comprobamos que el dataset cargado coincide con la cantidad de registros y columnas especificadas con np.shape (numpy).

Posteriormente, filtro la data para los países seleccionados y comparamos el nivel de país para llenar los valores faltantes.

A partir de este primer filtrado y con los valores resultantes, vemos que la columna location\_key repite muchas locaciones y es la razón de tantos datos faltantes, por lo que hacemos un nuevo filtro por esta columna pero solo con instancias que correspondan a los países seleccionados y no a sus regiones o subregiones: AR, CL, CO, MX, PE, BR.

Posteriormente, filtramos los datos en fechas mayores a 2021-01-01 y evaluamos la cantidad de vacíos que tenemos para eliminarlos o rellenarlos. Las columnas que tienen alta cantidad de nulos las relleno con 0 ya que no me daría un promedio confiable. Se decide no eliminar toda la columna ya que hay países que sí tienen datos y lo mismo ocurre con las filas, perdiendo información. Hay que tener en cuenta que los 0 en el futuro pueden cambiar y pueden ser rellenados con información recolectada o la columna se puede ir completando en la medida que se necesite.

Finalmente, guardo los datos filtrados en un nuevo archivo para poder reutilizarlo y evitar repetir el proceso de limpieza. Verifico si existe un directorio llamado "export" en el directorio actual. Si no existe, lo creo.

Reiniciamos el notebook para liberar memoria y cargar el archivo filtrado para continuar con el análisis. Cargo el archivo CSV en un DataFrame de pandas y le solicito a Python que me brinde información detallada sobre el DataFrame data\_latinoamerica\_paises. Luego imprimo las estadísticas principales columna por columna usando un ciclo for y while.

En el segundo avance, teniendo los datos limpios se profundiza en el análisis exploratorio de datos para extraer insights valiosos. Con visualizaciones claras y detalladas, se busca descubrir patrones, tendencias y anomalías en los datos de incidencia de COVID-19.

Como primer acercamiento, se realiza un ciclo for para recorrer las columnas y obtener la media,mediana,desviación estándar,mínimo y máximo de cada columna. Luego se crea una matriz de correlación para ver si existe alguna relación entre las variables para tener en cuenta un análisis posterior. Se crea la matriz de correlación de las variables numéricas, se crea máscara para el triángulo superior y se muestra valores superiores a 0.5 es decir donde hay una correlación fuerte. Sin embargo tambien se puede mostrar valores negativos.

Se realizan diagramas de barras donde cada barra representa un país y **La altura de cada barra representa el número de nuevos casos confirmados:** Es una cantidad discreta asociada a cada categoría, no una frecuencia dentro de un rango.

Se decide realizar histogramas para ver la distribución de algunos datos y varias visualizaciones de datos con Matplotlib y Seaborn para la identificación de tendencias y patrones.

Comienzo generando un gráfico de dispersión de la temperatura promedio vs los nuevos confirmados para ver la relación entre estas variables y el comportamiento de los casos respecto a la temperatura. Aquí se puede observar que en algunos países la temperatura no influye directamente y en otros parece ser que si. De esta forma, se realiza un análisis de la temperatura promedio con la cantidad de muertes, viendo un comportamiento similar al de los casos confirmados.

Ahora, hago un estudio longitudinal de cómo se comportan las variables a lo largo del tiempo. Como primera aproximación se grafica con líneas las dosis totales administradas por mes en todos los países, observando en general que la cantidad en los 6 países ha ido en aumento a lo largo del tiempo.

Para un análisis detallado, vemos las dosis administradas por mes en cada país. Es un promedio del acumulado aunque también se podría hacer el máximo por mes para contar con el valor más alto alcanzado. Con esta tendencia, se ve que en algunos países la cantidad de dosis ha ido aumentando a lo largo del tiempo mientras que en otras va disminuyendo. Esto puede darse por varios factores, como la cantidad de vacunas disponibles, la cantidad de personas que se han vacunado en las campañas nacionales, entre otras.

Hago lo mismo para el acumulado de muertes, donde noto que el país con más decesos es Brasil mientras que en México hubo una baja muy brusca que puede tratarse de una pérdida de datos o una disminución real de datos con estrategias gubernamentales. Para estos análisis hay que tener en cuenta la cantidad de habitantes en cada país como otros factores, incluido la cantidad de datos que tenemos.

Teniendo en mente lo anterior, notamos que el país con más casos confirmados es Brasil también, junto al que más recuperados tuvo. Esta última variable se encuentra incompleta en muchos países, por lo que es difícil hacer un análisis de esta métrica o de cualquiera que la incluya, como los casos activos.

Como conclusión inicial y a través de las visualizaciones, podemos afirmar que Chile es el país que mejor manejó la pandemia ya que la cantidad de muertes y los casos confirmados son menores, además de que no es desproporcionada la cantidad de vacunas administradas. Por otro lado, Brasil tiene una cantidad de muertes y casos confirmados mucho mayor al resto de los país y la cantidad de dosis de vacunas administradas es muy desproporcionada según el gráfico de densidad y se puede interpretar como una vacunación retrasada (los demás países se encuentran en el orden).

Cada gráfico realizado en Python se personaliza con etiquetas, títulos y leyendas para que sea más fácil el análisis tanto para el que los crea como para el que puede consultarlos en el futuro. De esta forma, las visualizaciones terminan siendo más informativas y atractivas.

Por otro lado, se realizó un boxplot con la temperatura media por país y aunque Chile tiene rangos diferentes de temperatura que podrían haberlo afectado, se observa que manejo mejor la pandemia que los demás países. Esto implica que los nuevos centros deben crearse en países donde directamente hay más casos confirmados y muertes, que es donde se necesita ayuda.

Se generan violínplots para ver cómo se comportan las variables generando las alertas necesarias y se realizan gráficos de barra para ver la distribución de la población por distintos grupos de edad y cuales son los grupos etarios más vacunados. Por otro lado, revisamos la mortalidad por género para identificar la población más vulnerable en cada país de forma que puedan vacunarse primero.

Luego verificamos la vacunación por mes de cada país en grupos de 0 a 19 años, 20 a 39, 40 a 59, 60 a 79 y de 80 en adelante. Revisamos la mortalidad por genero para identificar la población más vulnerable en cada país, con esto ya podriamos tener una idea de cuál sería la población a vacunar primero. De acuerdo a la gráfica de mortalidad, se puede observar que en todos los paises la cantidad de muertes es mayor en los hombres que en las mujeres, por lo que se podría decir que los hombres son más vulnerables a la enfermedad o están menos protegidos.

Finalmente, analizo las enfermedades preexistentes y comorbilidades. Veo si existe una correlación entre las variables y si hay alguna enfermedad que sea más común en los países seleccionados. De la matriz obtenida podemos observar que fumar y tener diabetes son las enfermedades preexistentes que más afectan a la población en estos países, causando el mayor número de muertes. Filtrando por países, las dos enfermedades muestran un comportamiento similar pero se puede observar que la diabetes ha causado más muertes que fumar.

En el anteúltimo avance, realizamos un análisis más profundo de los datos relacionados con la incidencia de COVID-19. Se realiza un estudio exploratorio utilizando técnicas avanzadas de Pandas y Numpy con series temporales para comprender la evolución de elementos específicos del conjunto de datos.

Comenzamos con el análisis del comportamiento semanal y anual de los casos nuevos en América Latina. Se observa que en las primeras semanas tenemos un pico y este alza puede ser por las fiestas en donde la gente probablemente se reunió y bajo los cuidados. Finalmente la tendencia baja, aproximadamente en la semana 45. Lo mismo sucede con los meses, particularmente con el mes de enero. A pesar que los casos confirmados disminuyen en el tiempo, no quiere decir que la pandemia haya terminado ya que la cantidad de nuevos confirmados presenta picos.

Con respecto al progreso de vacunación por país, Brasil es el que tiene mayor dosis administradas pero debe tenerse en cuenta la mayor población, como mencionamos anteriormente. Por otro lado, la temperatura promedio vs los casos nuevos confirmados nos muestra que la temperatura ha ido aumentando y los casos disminuyendo pero hubo un pico antes del mes de marzo 2022 que disminuye rápidamente y puede deberse a la cantidad de vacunas administradas o las personas que se han vacunado.

En tanto, con la visualización del porcentaje de población urbana por país vemos que el que tiene mayor impacto fue Chile y probablemente la pandemia se contuvo en la ciudad y no en la zona rural. Otra muestra del buen resultado de las políticas de Chile, es a través del gráfico de dosis administradas cada 100 personas en cada país (cada 100 da 350 dosis, probablemente porque son vacunas con más de una dosis por persona).

Con la Tasa de Letalidad calculada con decesos/confirmados, podemos decir que los países con peor performance fueron Perú y México, mientras que Argentina y Chile tienen los valores más bajos. Con la Tasa de Mortalidad en Latam vemos que los casos confirmados y los decesos han disminuido mientras que los recuperados han aumentado en el tiempo, debido a la vacunación y las políticas tomadas por los países.

Hay que tener en cuenta que para analizar los casos activos (confirmados-decesos-recuperados) se debe tener los datos completos de los recuperados en todos los países y eso es información que puede buscarse en el futuro para mejorar el análisis.

En el último avance, se integran y se presentan los hallazgos analíticos en Power BI con el objetivo de sintetizar el análisis realizado con un reporte interactivo que facilite la toma de decisiones estratégicas para la expansión de laboratorios. De esta forma, se conecta Power BI con el .csv creado y exportado de Python y se diseña un dashboard que muestra de manera efectiva los resultados facilitando la identificación de áreas prioritarias para la expansión. Además, se comparan los beneficios de utilizar visualizaciones interactivas en dashboards digitales vs las gráficas estáticas más útiles para reportes impresos.

Este análisis detallado permite la identificación de patrones y tendencias adicionales a las ya mencionadas en los avances anteriores, así como también la evaluación de la efectividad de distintas estrategias de vacunación y el impacto de factores como la urbanización y las condiciones preexistentes en la propagación de COVID-19.

# 

# Análisis del dashboard

Teniendo en cuenta los insights encontrados en Python, se integran las visualizaciones más importantes tomando el rol de Data Analytics y otras nuevas que son posibles gracias a la facilidad que existe para generar gráficos y medidas en Power BI.

El reporte cuenta con una PORTADA y 5 hojas (DATA, DOSIS DE VACUNAS, INDICES, %DEDATOS y CONCLUSIÓN). En la hoja de DATA encontramos información general como la población por grupo etario, la población por país, la muerte acumulada por año y la evolución de la vacunación contra el COVID-19.

En la hoja de DOSIS DE VACUNAS me encuentro con la Evolución de Dosis administradas por mes realizado con un gráfico de PYTHON donde indica esta evolución por mes y año. También, 2 tarjetas donde indica la cantidad de vacunas administradas y en otra la población total. En la tercera hoja, tengo índices que me ayudan a seguir profundizando en el analisis para llegar a insights de relevancia con lo solicitado por la empresa. Se encuentran los gráficos de Índice de accesibilidad, índice de demanda, índice de expansión y mortalidad por polución.

En la cuarta hoja, se encuentran los porcentajes de población vacunada y población recuperada.

Por último en la ultima hoja se encuentra la conclusión que me brindó este realizar este dashboard.

# Conclusiones y Recomendaciones

# Al finalizar este primer estudio de los datos con el dataset otorgado, se llega a la conclusión de que los países en donde debe expandirse la empresa BIOGENESYS en su estrategia dentro de Latinoamérica son México, Colombia y Brasil. En líneas generales y recapitulando, se puede decir que son los países que tienen las mayores tasas de fatalidad del COVID-19, más casos confirmados y decesos y el menor porcentaje de vacunas administradas según su población. Esto implica que las políticas de vacunación no han sido las mejores y/o no han tenido la cantidad de vacunas necesarias para afrontar la pandemia.

Estas ubicaciones seleccionadas son donde la empresa farmacéutica lograría optimizar la respuesta a la pandemia y mejorar el acceso a las vacunas, pudiendo representar en el futuro un cambio en la performance y en los índices que poseen estos países.

# Reflexión personal

# En cuanto a PYTHON esta muy bueno para realizar análisis rapidamente, pero me parece mucho más rápido e interactivo hacer los gráficos en PowerBI.

Me encuentro muy contenta y entusiasmada por todo lo que vimos en este último módulo. Se que me queda mucho por practicar, pero creo que de esto se trata para ser un gran Analista de Datos.