**BIOGENESYS**

**Nombre del autor: Gina Correa**

**Email: ginagiulianacorrea@gmail.com**

**Cohorte: FT-DA01**

**Fecha de entrega:** 6/05/2024

**Institución:**

****

# Introducción

Biogenesys nos ha contratado para realizar un estudio que impulse su estrategia de expansión en Latinoamérica, enfocándonos específicamente en los mercados de Colombia, Argentina, Chile, México, Perú y Brasil. El objetivo principal de este proyecto es realizar un análisis exhaustivo de la incidencia de COVID-19 y otros factores relevantes utilizando técnicas de análisis exploratorio de datos (EDA), limpieza de datos, mejoras en el acceso a la información y el desarrollo de dashboards interactivos. Estas acciones tienen como fin proporcionar información detallada y confiable que respalde la toma de decisiones estratégicas e informadas dentro de Biogenesys.

Hasta el momento, hemos logrado identificar tendencias significativas en los datos relacionados con la pandemia y otros factores clave para la expansión en la región mencionada. Además, hemos mejorado la calidad de los datos mediante técnicas de limpieza y validación, y hemos implementado procesos eficientes de extracción, transformación y carga (ETL) para facilitar el acceso y la disponibilidad de los datos necesarios para el análisis.

El objetivo final de este proyecto es utilizar estas herramientas y técnicas para desarrollar dashboards interactivos que permitan a los directivos de Biogenesys explorar los datos desde múltiples perspectivas, identificar patrones y tendencias relevantes, y tomar decisiones estratégicas de manera más informada y oportuna. Esto les ayudará a gestionar de manera más efectiva su estrategia de expansión y estar preparados para responder rápidamente ante cualquier situación futura que pueda surgir en estos mercados latinoamericanos clave.

## Desarrollo del Proyecto

**Recopilación y Selección de Datos (Avance 1):**

* La metodología de recopilación y selección de datos se basó en la revisión del archivo "data\_latinoamérica.csv" para identificar las columnas y datos relevantes para el análisis de los países de interés para Biogenesys en Latinoamérica.
* Se seleccionaron variables relacionadas con la incidencia de COVID-19, factores socioeconómicos, demográficos y políticos.
* Luego, se procedió a la carga y transformación de los datos utilizando código Python en un notebook llamado "PIDA\_M4\_GinaGiuliana\_Correa.ipynb".
* Se importaron librerías, se leyó el archivo CSV, se verificó la cantidad de registros y columnas, se seleccionaron países de interés y se filtraron datos posteriores al 01/01/2021.
* Se compararon valores entre países para rellenar datos faltantes y se realizó una limpieza preliminar eliminando registros nulos y corrigiendo tipos de datos.
* Se examinaron las características del conjunto de datos para comprender la distribución de variables clave como la incidencia de COVID-19.
* Finalmente, los datos filtrados se guardaron en "DatosFinalesFiltrado.csv" y se aplicaron ciclos "for" para el cálculo de estadísticas descriptivas y otras métricas, permitiendo analizar la dispersión y consistencia de los datos.
* Contestamos las siguientes preguntas;
* ¿Qué implican estas métricas y cómo pueden ayudar en el análisis de datos?
* ¿Se muestran todas las estadísticas en todas las columnas durante el análisis?
* ¿Cuál es la razón de la respuesta anterior y cómo podría afectar la interpretación de los resultados obtenidos?

Estas métricas estadísticas proporcionan información importante sobre las diferentes columnas de tu conjunto de datos;

1. \*\*count:\*\* Indica el número de valores no nulos en cada columna. Esto es útil para identificar si hay valores faltantes en tus datos.

2. \*\*mean (promedio):\*\* Representa el valor promedio de los datos en cada columna. Puede ayudar a entender el valor típico o central de los datos.

3. \*\*std (desviación estándar):\*\* Mide la dispersión o variabilidad de los datos con respecto al promedio. Una desviación estándar alta indica que los datos están más dispersos alrededor del promedio.

4. \*\*min (mínimo):\*\* Es el valor más pequeño en cada columna. Puede proporcionar información sobre el rango mínimo de los datos.

5. \*\*25%, 50% (mediana), 75% (percentil 75):\*\* Estos percentiles dividen los datos ordenados en cuatro partes iguales. La mediana (50%) indica el valor que separa la mitad superior de la mitad inferior de los datos. El percentil 25 (Q1) es el valor por debajo del cual cae el 25% de los datos, y el percentil 75 (Q3) es el valor por debajo del cual cae el 75% de los datos.

6. \*\*max (máximo):\*\* Es el valor más grande en cada columna. Proporciona información sobre el rango máximo de los datos.

No, no se muestran todas las estadísticas en todas las columnas durante el análisis. En general, las estadísticas se muestran para las columnas numéricas, ya que tienen sentido para cálculos como promedio, desviación estándar, mínimo, máximo, etc. Las columnas que contienen datos no numéricos (como cadenas o fechas) no se incluyen en este tipo de análisis.

La razón por la que no se muestran todas las estadísticas en todas las columnas es que algunas métricas, como el promedio o la desviación estándar, solo tienen sentido para datos numéricos. Incluir estas estadísticas para columnas no numéricas no tendría significado y podría generar errores o resultados poco útiles.

Esta selección de estadísticas puede afectar la interpretación de los resultados porque proporciona una visión general de las propiedades numéricas de tus datos, como la dispersión, la centralidad y los valores extremos. Sin embargo, no proporciona información detallada sobre la distribución de los datos (por ejemplo, si siguen una distribución normal), la presencia de valores atípicos, la correlación entre variables, entre otros aspectos importantes que pueden requerir análisis más detallados y específicos. Es importante combinar estas estadísticas con visualizaciones y técnicas de análisis más avanzadas para obtener una comprensión completa de tus datos.

* ¿Qué representa la mediana?
* ¿Cómo varía la dispersión de los datos en el conjunto de datos analizado, en términos de la varianza y el rango?
* ¿Qué nos puede indicar esto sobre la consistencia o la variabilidad de los datos en relación con la mediana?

La mediana es un valor que divide un conjunto de datos ordenados en dos partes iguales. Esto significa que el 50% de los datos están por encima de la mediana y el otro 50% están por debajo de la mediana. La mediana es una medida de tendencia central que es menos sensible a los valores extremos o atípicos que el promedio (media aritmética). Si los datos están ordenados de menor a mayor, la mediana será el valor que ocupa la posición central.

La variabilidad de los datos en un conjunto de datos se puede entender a través de medidas como la varianza y el rango.

1. \*\*Varianza:\*\* La varianza es una medida de dispersión que indica cuánto se alejan los valores individuales de la media. Una varianza alta indica una mayor dispersión de los datos alrededor de la media, mientras que una varianza baja indica una dispersión más cercana a la media. Si la varianza es cero, significa que todos los valores son iguales y no hay dispersión.

2. \*\*Rango:\*\* El rango es la diferencia entre el valor máximo y el valor mínimo en un conjunto de datos. Es una medida de la amplitud total de los datos. Un rango amplio indica una mayor variabilidad en los valores observados.

La relación entre la mediana y la dispersión de los datos puede indicar la consistencia o la variabilidad de los datos de la siguiente manera:

- Si la mediana está cerca del centro del rango intercuartílico (IQR) y la varianza es baja, los datos tienden a estar agrupados cerca de la mediana, lo que sugiere una consistencia y poca variabilidad en los datos.

- Si la mediana está en el centro del rango intercuartílico pero la varianza es alta, los datos pueden estar distribuidos más uniformemente alrededor de la mediana, lo que indica cierta variabilidad pero una consistencia relativa en la posición central de los datos.

- Si la mediana está más cerca del extremo inferior o superior del rango intercuartílico y/o la varianza es alta, los datos pueden estar más dispersos alrededor de la mediana, lo que sugiere una mayor variabilidad y menos consistencia en la posición central de los datos.

En resumen, la mediana nos da una idea de dónde se concentran los valores centrales de un conjunto de datos, mientras que la varianza y el rango nos informan sobre la dispersión y amplitud de los datos. La relación entre estas medidas puede proporcionar información valiosa sobre la consistencia o variabilidad de los datos en relación con la posición central representada por la mediana.

En conclusión, la metodología empleada garantizó datos de calidad y relevancia para respaldar la estrategia de expansión de Biogenesys, facilitando un análisis más preciso. La preparación y limpieza de datos fueron etapas críticas para resultados íntegros y toma de decisiones informadas.

**Análisis Exploratorio de Datos (Avance 2):**

* Tras la limpieza de datos, se realizó un análisis exploratorio más profundo para extraer insights valiosos que guíen la planificación estratégica de la expansión.
* Se utilizaron visualizaciones claras y detalladas para descubrir patrones, tendencias y anomalías en la incidencia de COVID-19, tasas de vacunación e infraestructura sanitaria.
* Se exploraron las propiedades estadísticas del conjunto de datos con Pandas y Numpy.
* Para la visualización se emplearon herramientas como Matplotlib y Seaborn.
* Se representó la distribución de la incidencia y las tasas de vacunación mediante histogramas y gráficos de densidad.
* Se compararon regiones o países con gráficos de barras.
* Durante este análisis se identificaron tendencias a largo plazo, patrones estacionales o geográficos, y la influencia de factores como la temperatura en la propagación de la enfermedad.
* También se examinó cómo varían las tasas de vacunación en diferentes regiones o momentos del año.

Estas visualizaciones y análisis estadísticos permitieron crear un panorama detallado de la situación, proporcionando información sólida para la toma de decisiones estratégicas en la expansión de laboratorios farmacéuticos en Latinoamérica.

**Análisis Exploratorio Detallado (Avance 3):**

* Se realizó un análisis más profundo de los datos relacionados con la incidencia de COVID-19.
* El objetivo principal fue preparar los datos para una visualización avanzada que permita identificar ubicaciones estratégicas para la expansión de laboratorios.
* Se utilizaron técnicas avanzadas de Pandas y Numpy, centrándose en el análisis de series temporales para comprender la evolución de elementos específicos del conjunto de datos.
* Se identificaron tendencias, estacionalidad y patrones temporales relevantes.
* Se realizaron análisis de autocorrelación y descomposición de series temporales.
* Se generaron gráficas como: evolución de casos activos vs. recuperados, tasa de crecimiento (%), relación entre cobertura de vacunación y reducción de casos, nuevos casos y temperatura promedio, entre otros.
* Se investigaron posibles correlaciones entre variables para comprender mejor las interacciones entre los datos.
* Se aplicaron funciones personalizadas a cada columna para profundizar en el análisis y obtener insights valiosos para recomendaciones basadas en datos.

**Integración en Power BI (Avance 4):**

* Se sintetizó el análisis en dashboards interactivos y reportes para facilitar la toma de decisiones estratégicas de expansión de laboratorios y centros de vacunación.
* Se importó el conjunto de datos preparado y analizado en Python a Power BI ("DatosFinalesFiltrado2.csv").
* Este paso trasladó el análisis técnico a una plataforma que permite la visualización interactiva y la comprensión intuitiva de los datos por parte de los directivos.
* Los dashboards de Power BI permiten a los usuarios explorar los datos, filtrar información y realizar análisis en tiempo real.
* Se realizaron visualizaciones claras y precisas que facilitan la identificación de patrones, tendencias y áreas prioritarias para la toma de decisiones estratégicas.
* La integración de los resultados del análisis en Power BI ha sido un paso crucial para el éxito del proyecto. Los dashboards y reportes generados proporcionan a la empresa farmacéutica herramientas.

# EDA e insights

# [Deberás detallar los insights obtenidos en el Análisis Exploratorio de Datos y visualizaciones]

**Población y Contagios: Una Relación Ineludible con Nuancias:**

* Brasil, con su vasta población de 212,6 millones de habitantes, registra la mayor cantidad de casos confirmados durante el período 2021-2022, con un total de 34,8 millones de contagios.
* En contraste, Chile, con una población de 19,7 millones de habitantes, presenta una cantidad relativamente menor de contagios (10,2 millones) durante el mismo período.
* Esta disparidad destaca la importancia de considerar la densidad poblacional como un factor determinante en la propagación del virus.

**Vacunación: Un Escudo Efectivo con Diferencias Significativas:**

* Chile se destaca por su eficiente campaña de vacunación, logrando una tasa de vacunación completa del 87,2% de su población durante el período 2021-2022.
* En comparación, Perú presenta una tasa de vacunación completa mas baja,, lo que podría explicar parcialmente su mayor tasa de letalidad.

**Dosis Aplicadas: Una Barrera a la Transmisión con Patrones Evidentes:**

* Se observa una tendencia clara: a mayor cantidad de dosis de vacunas aplicadas, menor número de contagios.
* Esto evidencia la eficacia de la vacunación como estrategia para reducir la transmisión del virus y proteger a la población.

**Temperatura: Un Factor No Determinante con Nuevos Análisis:**

* El análisis no encuentra una correlación significativa entre la temperatura y el número de contagios o muertes por COVID-19 durante el período 2021-2022.

**Muertes: Un Pico Inicial y una Tendencia Decreciente con Detalles Temporales:**

* Se observa un pico inicial en el número de muertes por COVID-19 en todos los países durante el primer semestre de 2021.
* A partir del segundo semestre de 2021 y durante el año 2022, se observa una tendencia descendente en el número de muertes, coincidiendo con la implementación de las campañas de vacunación y la aplicación de medidas de control.
* Esta tendencia positiva destaca la importancia de las medidas de salud pública para mitigar el impacto de la pandemia.

**Tasa de Letalidad: Perú en Alerta con Posibles Causas:**

* Perú presenta la tasa de letalidad más alta entre los países analizados durante el período 2021-2022, con un 3,1% de letalidad.
* Esta alta tasa podría estar asociada a una menor cobertura de vacunación, como se mencionó anteriormente, y a una gestión deficiente de la pandemia.

**Datos de Recuperados: Un Vacío Informativo con Implicaciones:**

* Solo se cuenta con información de recuperados para Brasil y Colombia durante el período 2021-2022.
* La falta de datos de recuperados en los demás países limita la comprensión completa de la situación epidemiológica y la evaluación del impacto real de la pandemia.

# 

# Análisis del dashboard

Recorriendo el Dashboard:

Página de Inicio:

* La página de inicio ofrece una descripción general de la compañía farmacéutica y presenta una barra de navegación para acceder fácilmente a las diferentes secciones del dashboard.

Página de Países:

* Esta página se centra en datos específicos de cada país, ofreciendo un análisis integral de los casos de COVID-19, la cobertura de vacunación y diversos indicadores de salud.
* Puntos clave a destacar:
  + Comparación de nuevos contagios y muertes entre países.
  + Análisis de las tasas de administración de vacunas.
  + Examen de las tasas de mortalidad por género y grupos de edad.
  + Exploración de las tendencias de recuperación.
  + Visualización de los patrones de aplicación de vacunas.

Página de Fechas:

* La página de fechas permite a los usuarios examinar las tendencias a lo largo del tiempo, utilizando un filtro por año para explorar períodos específicos.
* Puntos clave a destacar:
  + Análisis de las tasas de recuperación anuales.
  + Examen de las tendencias de mortalidad anuales.
  + Exploración de los patrones anuales de administración de vacunas.
  + Identificación de tendencias en nuevos casos por año.

Página de Condiciones:

* Esta página se adentra en las condiciones de salud de las poblaciones de los distintos países, proporcionando información valiosa sobre los factores de riesgo y las posibles intervenciones.
* Puntos clave a destacar:
  + Visualización de la prevalencia del tabaquismo por país.
  + Análisis de la prevalencia de la diabetes en todos los países.
  + Examen de las tasas de mortalidad específicas por género.
  + Exploración de la mortalidad por contaminación del aire.
  + Identificación de países con alta mortalidad relacionada con la comorbilidad.
  + Mapa interactivo que muestra los casos confirmados, las muertes y los recuperados por país al colocar el cursor sobre ellos.

Página de Promedios:

* La página de promedios resume las métricas clave de todos los países, proporcionando una visión general concisa de las tendencias generales.
* Puntos clave a destacar:
  + Visualización de las tasas de reducción de casos.
  + Análisis del promedio de población recuperada.
  + Examen de las tasas promedio de mortalidad por país.
  + Exploración de la disponibilidad promedio de enfermeras y médicos.
  + Identificación de países con altas tasas de crecimiento de casos.

Las funciones interactivas del dashboard permiten a los usuarios explorar los datos en profundidad, identificando patrones y tendencias que pueden no ser evidentes en visualizaciones estáticas.

La capacidad de filtrar datos por país y fecha proporciona una vista granular de la situación, permitiendo a los usuarios centrarse en áreas de interés específicas.

La presentación clara y concisa del dashboard facilita la comunicación de datos complejos a una amplia gama de interesados.

# Conclusiones y Recomendaciones

Considerando los resultados obtenidos, se identifican las siguientes conclusiones estratégicas para la expansión de Biogenesys en Latinoamérica:

**Países Prioritarios:**

* **Chile:** Su eficiente manejo de la pandemia, alta tasa de vacunación y enfoque en la salud pública lo convierten en un mercado atractivo para la expansión de Biogenesys.
* **Brasil:** A pesar de la alta densidad poblacional y el número de contagios, la gran cantidad de habitantes y la creciente demanda de soluciones de salud podrían ser un factor favorable para la expansión

**Factores a Considerar:**

* **Políticas de vacunación:** Es importante evaluar las políticas de vacunación de cada país para identificar oportunidades de colaboración y desarrollo de nuevos productos y servicios.
* **Inversión en salud:** Se debe considerar la inversión en salud de cada país para determinar el potencial de mercado y la demanda de soluciones de diagnóstico y tratamiento.
* **Problemas sociales y económicos:** Los problemas sociales y económicos, como la pobreza y la desigualdad, pueden afectar el acceso a la atención médica y la adopción de nuevas tecnologías.

**Recomendación Principal:**

En base a los insights obtenidos, se recomienda que Biogenesys priorice su expansión en **Chile**, considerando su política sanitaria. Sin embargo, se debe realizar un análisis más profundo de la situación en **Brasil**, tomando en cuenta su gran población y creciente demanda de soluciones de salud.

**Consideraciones Finales:**

Es importante destacar que esta recomendación se basa en los datos y análisis disponibles al momento. Se recomienda realizar un seguimiento continuo y considerar otros factores relevantes para la toma de decisiones estratégicas de expansión.

# Reflexión personal

# Este proyecto implicó un gran aprendizaje con momentos bastantes complejos, se me dificulto bastante,pero sin lugar a dudas este módulo ha sido el que más me ha gustado, es impresionante la gratificación que uno siente al aprender tantos conceptos tan rápido, y sentir que uno es capaz de resolver. Este módulo me encantó, adopté mucho más la metodología autodidacta y estas últimas semanas han sido de estudio y comprensión de este lenguaje el cual espero utilizar por mucho tiempo en mis empleos. También es bueno mencionar que el Profe de este módulo estuvo atento y predispuesto a solucionar todas mis inquietudes. Las últimas semanas han sido fascinantes.

# EXTRA CREDIT

Para enriquecer el análisis y proporcionar una perspectiva más completa de la situación, se incorporan mapas geoespaciales utilizando la información de latitud y longitud disponible en los datos. Se emplea la biblioteca Folium para crear mapas interactivos que visualizan la distribución geográfica de casos confirmados, muertes, vacunados y recuperados por COVID-19 en los países de interés.

**Metodología:**

1. **Filtrado de Datos:** Se seleccionan las columnas relevantes para el análisis geoespacial: "latitude", "longitude", "cumulative\_deceased", "cumulative\_recovered" y "country\_name". Se eliminan filas con valores nulos para garantizar la calidad de los datos.
2. **Creación del Mapa Base:** Se crea un mapa utilizando Folium, centrándolo en una ubicación inicial (latitud 0, longitud 0) y estableciendo un zoom inicial de 2. Este zoom puede ajustarse según las necesidades de visualización.
3. **Generación de Capas HeatMap:** Se crean dos capas HeatMap independientes para visualizar la distribución espacial de las muertes y los casos recuperados por COVID-19.
   * **HeatMap Muertes:** Representa la densidad de muertes por COVID-19 en cada región. Se configura con:
     + Opacidad mínima: 0.5
     + Valor máximo: valor máximo en la columna "cumulative\_deceased"
     + Radio: 15
     + Desenfoque: 10
     + Zoom máximo: 1
     + Nombre: "Muertes"
   * **HeatMap Recuperados:** Representa la densidad de casos recuperados por COVID-19 en cada región. Se configura de manera similar a la capa de muertes, utilizando el valor máximo de la columna "cumulative\_recovered" y el nombre "Recuperados".
4. **Adición de Capas al Mapa:** Se añaden ambas capas HeatMap al mapa principal utilizando el método "add\_child". Esto permite visualizar la distribución espacial de las muertes y los casos recuperados de forma superpuesta.
5. **Información al Pasar el Cursor:** Se implementa la funcionalidad "Tooltip" para mostrar información detallada al pasar el cursor sobre cada punto del mapa. La información incluye el nombre del país, el número de muertes y el número de casos recuperados.
6. **Control de Capas:** Se agrega un control de capas (Layers Control) al mapa, permitiendo a los usuarios activar o desactivar la visualización de las capas HeatMap de forma individual, facilitando la comparación y el análisis.
7. **Visualización del Mapa:** Se muestra el mapa interactivo generado, permitiendo a los usuarios explorar la distribución geográfica de las muertes y los casos recuperados por COVID-19 en los países de interés.

**Beneficios de la Incorporación de Mapas Geoespaciales:**

* **Visualización Clara:** Los mapas HeatMap proporcionan una representación visual clara y comprensible de la distribución espacial de las muertes y los casos recuperados por COVID-19.
* **Identificación de Patrones:** Permite identificar patrones y tendencias en la distribución geográfica de la enfermedad, facilitando la comprensión de la dinámica de la pandemia en diferentes regiones.
* **Toma de Decisiones Informadas:** Brinda información valiosa para la toma de decisiones estratégicas relacionadas con la expansión de laboratorios, centros de vacunación y otras iniciativas de Biogenesys en Latinoamérica.
* **Análisis Comparativo:** Permite comparar la distribución geográfica de muertes y casos recuperados entre diferentes países, identificando áreas con mayor o menor impacto de la pandemia.

**Conclusiones:**

La incorporación de mapas geoespaciales al análisis enriquece significativamente la comprensión de la situación epidemiológica en Latinoamérica. La visualización de la distribución geográfica de muertes y casos recuperados por COVID-19 proporciona información valiosa para la toma de decisiones estratégicas por parte de Biogenesys, permitiendo optimizar la ubicación de laboratorios, centros de vacunación y otras iniciativas para mitigar el impacto de la pandemia y mejorar la salud pública en la región.