**Expansión Estratégica de Laboratorios Biogénesys para la investigación y vacunación de COVID-19**

**Nombre del autor: Ma. Florencia Yofre**

**Email:** mafloriarencia@gmail.com

**Cohorte:** DA-FT02

**Fecha de entrega:** lunes, 3 de junio de 2024

**Institución: Biogenesys** es una empresa farmacéutica innovadora dedicada a la investigación y desarrollo de medicamentos y vacunas. Con una sólida presencia global, se enfoca en mejorar la salud pública mediante soluciones médicas avanzadas. Comprometida con la excelencia científica y la responsabilidad social, Biogenesys trabaja en colaboración con gobiernos y organizaciones de salud para garantizar el acceso equitativo a tratamientos y vacunas, especialmente en regiones con alta necesidad. Su misión es llevar salud y esperanza a más personas, marcando una diferencia significativa en la vida de millones en todo el mundo.

# Introducción

**Propósito del Proyecto:**

El proyecto tiene como objetivo principal identificar las ubicaciones óptimas para la expansión de laboratorios farmacéuticos de Biogenesys en Latinoamérica.

Utilizando análisis de datos sobre la incidencia de COVID-19, tasas de vacunación y disponibilidad de infraestructuras sanitarias, se busca optimizar la respuesta ante la pandemia y postpandemia, mejorando el acceso a las vacunas en las regiones seleccionadas.

**Objetivos Organizacionales Alcanzados:**

1. **Identificación de Regiones Estratégicas:** Se han analizado seis países de Latinoamérica (Colombia, Argentina, Chile, México, Perú y Brasil) para determinar las ubicaciones más adecuadas para la expansión de laboratorios farmacéuticos, basándose en datos críticos de salud y economía.
2. **Optimización de la Respuesta a la Pandemia:** El análisis detallado de la incidencia de COVID-19 y las tasas de vacunación permitirá a Biogenesys desarrollar estrategias para mejorar la respuesta sanitaria en las regiones seleccionadas.
3. **Mejora del Acceso a Vacunas:** Con el análisis de la disponibilidad de infraestructuras sanitarias, se han identificado áreas con necesidades específicas de mejora, lo que permitirá a Biogenesys planificar inversiones en infraestructura sanitaria para aumentar la capacidad de vacunación.
4. **Toma de Decisiones Informadas:** La utilización de herramientas avanzadas de análisis de datos, como Python y Power BI, ha proporcionado una base sólida de información para la toma de decisiones estratégicas.
5. **Alineación con los Objetivos de Crecimiento:** Este proyecto respalda el objetivo de crecimiento de Biogenesys, permitiendo una expansión eficaz y eficiente en el mercado latinoamericano, lo que a su vez contribuirá a aumentar la presencia y el impacto de la empresa en la región.

En resumen, el proyecto permitirá a Biogenesys identificar oportunidades clave para la expansión y mejorar su capacidad para responder a futuras emergencias sanitarias, logrando así sus objetivos organizacionales de crecimiento y optimización en la entrega de servicios de salud.

# Desarrollo del proyecto

# Primer Avance:

**Proceso de Análisis de Datos:**

Recibí una base de datos muy extensa, con 12,216,057 filas y 50 columnas, en formato CSV. Esta contenía numerosos datos faltantes, nulos y atípicos.

Para asegurar la calidad y fiabilidad del análisis, realicé un riguroso proceso de limpieza y normalización de datos, seguido de un proceso de Extracción, Transformación y Carga (ETL).

1. **Limpieza de Datos:**
   * Identificación y corrección de datos faltantes y nulos, así como la gestión de valores atípicos, asegurando que los datos fueran precisos y completos.
   * Los datos fueron filtrados por países: Colombia, Argentina, Chile, México, Perú y Brasil, y por fechas mayores a 2021-01-01, reduciendo significativamente la cantidad de datos.
   * Los datos nulos fueron reemplazados por la media para mantener la consistencia sin sesgar los resultados.
   * Los valores negativos en columnas específicas fueron transformados a positivos mediante el valor absoluto.
2. **Normalización de Datos:**
   * Se normalizaron los datos para facilitar su análisis, lo cual es crucial para obtener resultados coherentes y significativos.

El objetivo de estos pasos fue mejorar significativamente la calidad de los datos, permitiendo realizar un análisis bien fundamentado que proporcione insights valiosos y, así, colaborar con la toma de decisiones estratégicas de expansión de la empresa en América Latina. Estoy convencido de que este robusto proceso de manejo de datos contribuirá a la formulación de estrategias más efectivas y alineadas con los objetivos de BIOGENESYS, facilitando así su crecimiento y consolidación en el mercado latinoamericano.

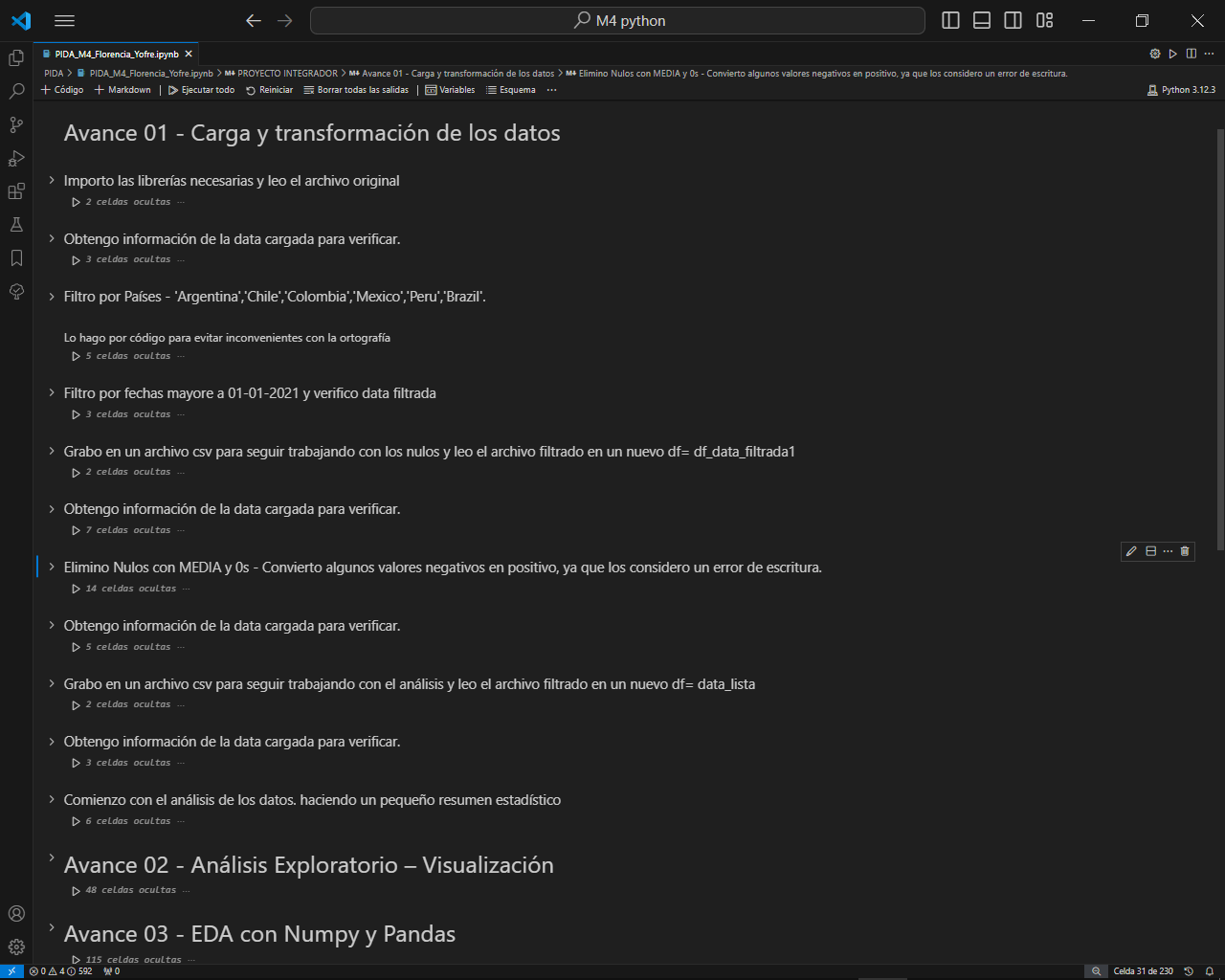
En este primer avance, se plantearon y respondieron varias preguntas clave:

* ¿Qué implican estas métricas y cómo pueden ayudar en el análisis de datos?
  + Al aplicar bucles for y/o while en Python junto con Pandas, pude calcular métricas estadísticas descriptivas como el conteo, la media, la desviación estándar y los percentiles para cada columna de nuestros datos. Estas métricas son esenciales para entender rápidamente las características principales de nuestro conjunto de datos y pueden ayudar significativamente en nuestro análisis.
* ¿Se muestran todas las estadísticas en todas las columnas durante el análisis?
  + No todas las estadísticas se mostraban para todas las columnas, ya que no todas las métricas eran relevantes para cada tipo de datos o columna específica.
* ¿Cuál es la razón de la respuesta anterior y cómo podría afectar la interpretación de los resultados obtenidos?
  + La selección selectiva de estadísticas podría influir en la interpretación de los resultados:
    - Para columnas categóricas, algunas métricas numéricas no eran aplicables, resultando en la exclusión de esas estadísticas.
    - La ausencia de ciertas métricas podría indicar posibles problemas en los datos, como valores nulos o inconsistencias, lo que afectaría la interpretación de los resultados.

Desarrollé una función personalizada que nos permite obtener la mediana, la varianza y el rango de nuestros datos.

**Análisis de Mediana, Varianza y Rango:**

* **¿Qué representa la mediana?**
  + La mediana es el valor central de un conjunto de datos ordenado. Representa el punto en el que la mitad de los datos es menor y la otra mitad es mayor. Es una medida de tendencia central que es menos sensible a valores atípicos que la media.
* **¿Cómo varía la dispersión de los datos en el conjunto de datos analizado, en términos de la varianza y el rango?**
  + La varianza mide la dispersión de los datos en torno a la media, indicando qué tan dispersos están los valores. Un valor alto de varianza indica que los datos están muy dispersos. El rango es la diferencia entre el valor máximo y el mínimo, proporcionando una medida de la amplitud de los datos. Un rango amplio indica una gran dispersión.
* **¿Qué nos puede indicar esto sobre la consistencia o la variabilidad de los datos en relación con la mediana?**
  + Si la varianza y el rango son altos, significa que hay una alta variabilidad en los datos, lo que puede indicar inconsistencias. Si son bajos, los datos son más consistentes y están más agrupados alrededor de la mediana. Esto puede ser crucial para determinar la fiabilidad de los datos y su interpretación.



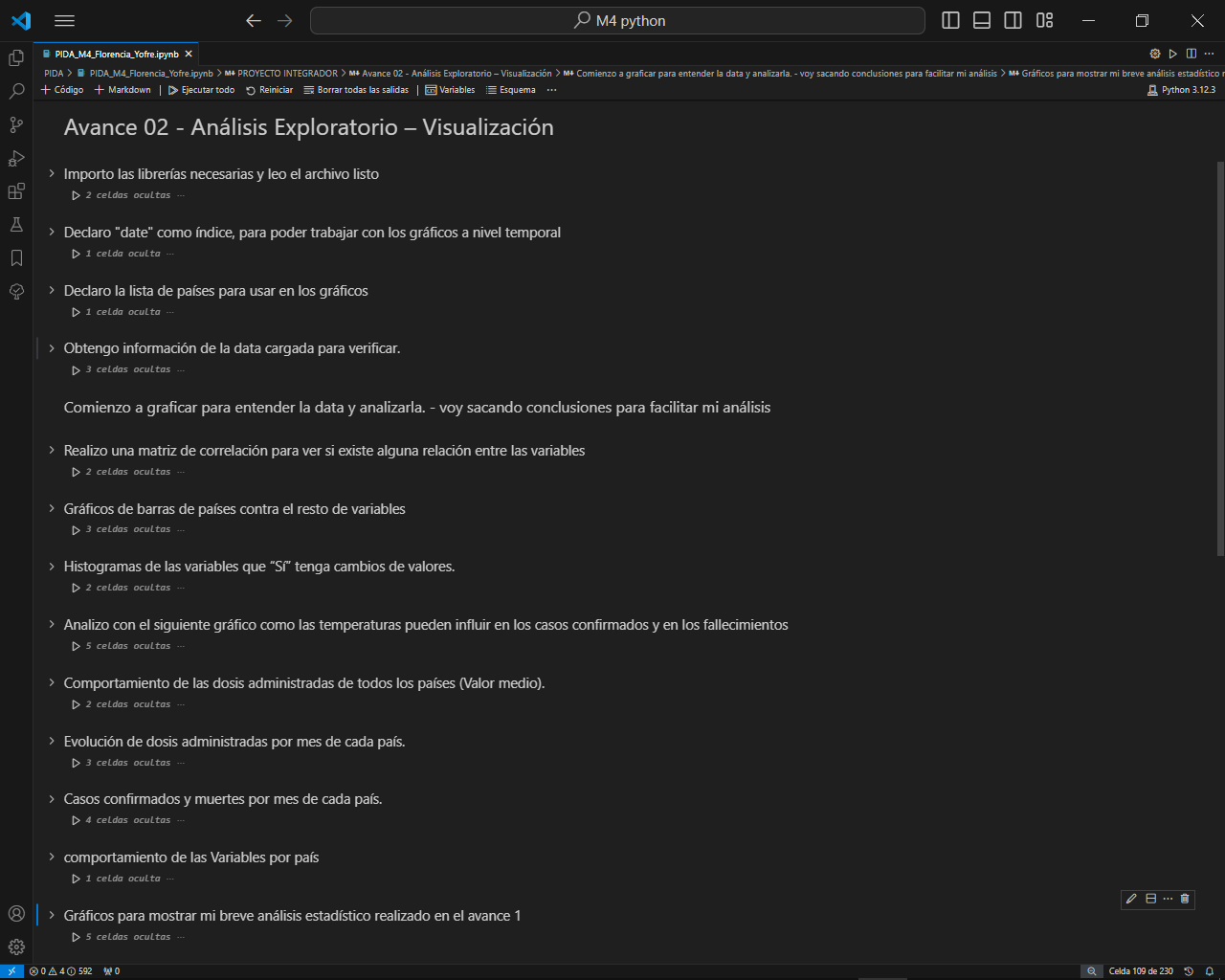
# Segundo Avance:

Con la data limpia, se realizaron visualizaciones con gráficos para identificar tendencias y patrones más claramente. Trabajé con cuatro librerías principales:

* Pandas y Numpy: Para explorar las propiedades estadísticas de los datos.
* Matplotlib y Seaborn: Para la visualización de datos.

Algunos de los gráficos que se realizaron fueron:

* Histogramas y gráficos de densidad: Para entender la distribución de la incidencia de COVID-19 y las tasas de vacunación.
* Gráficos de barras: Para comparar diferentes regiones.
* Mapas de calor: Para identificar correlaciones entre diferentes variables.
* Gráficos de dispersión: Para explorar posibles relaciones entre las variables.



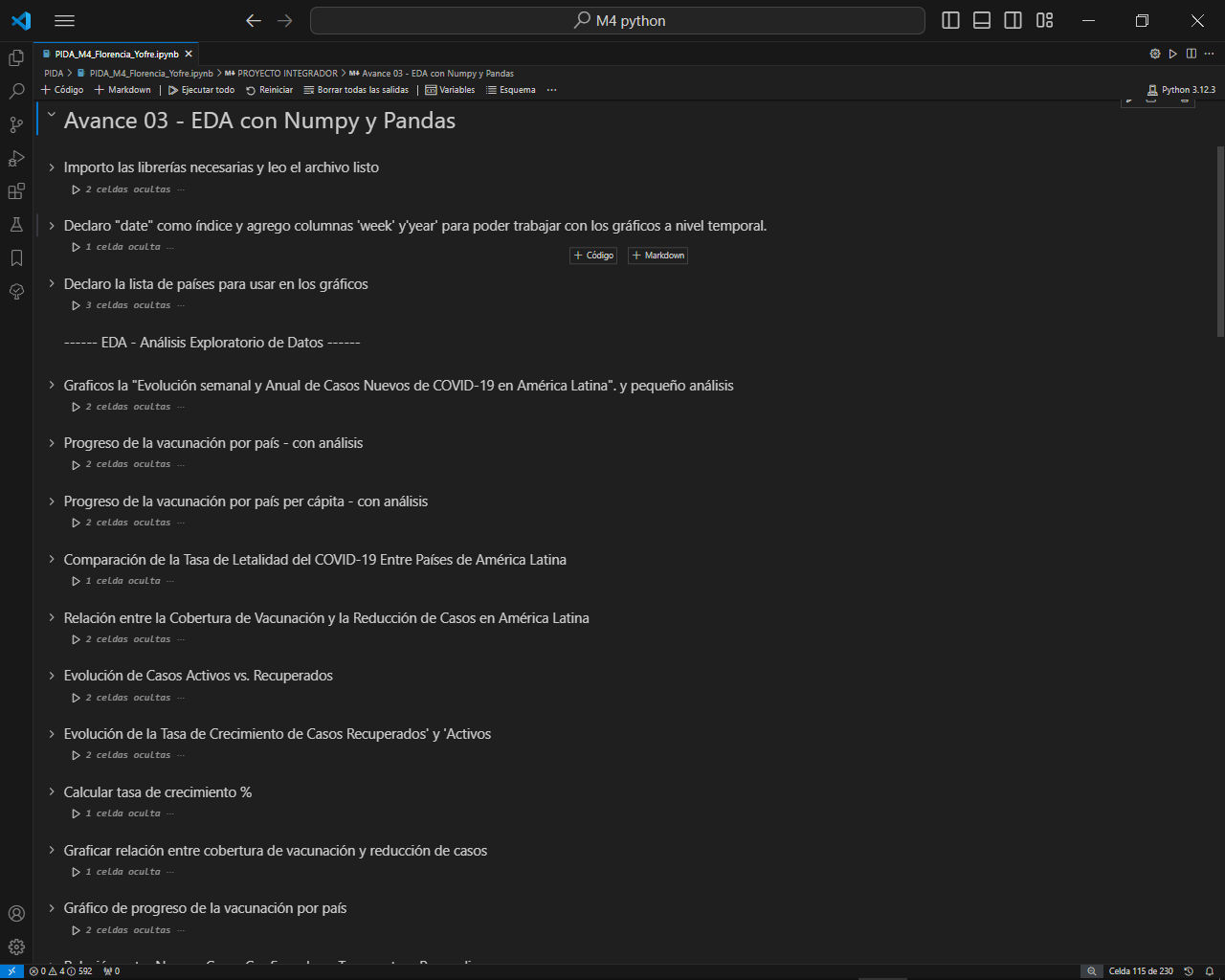
# Tercer Avance:

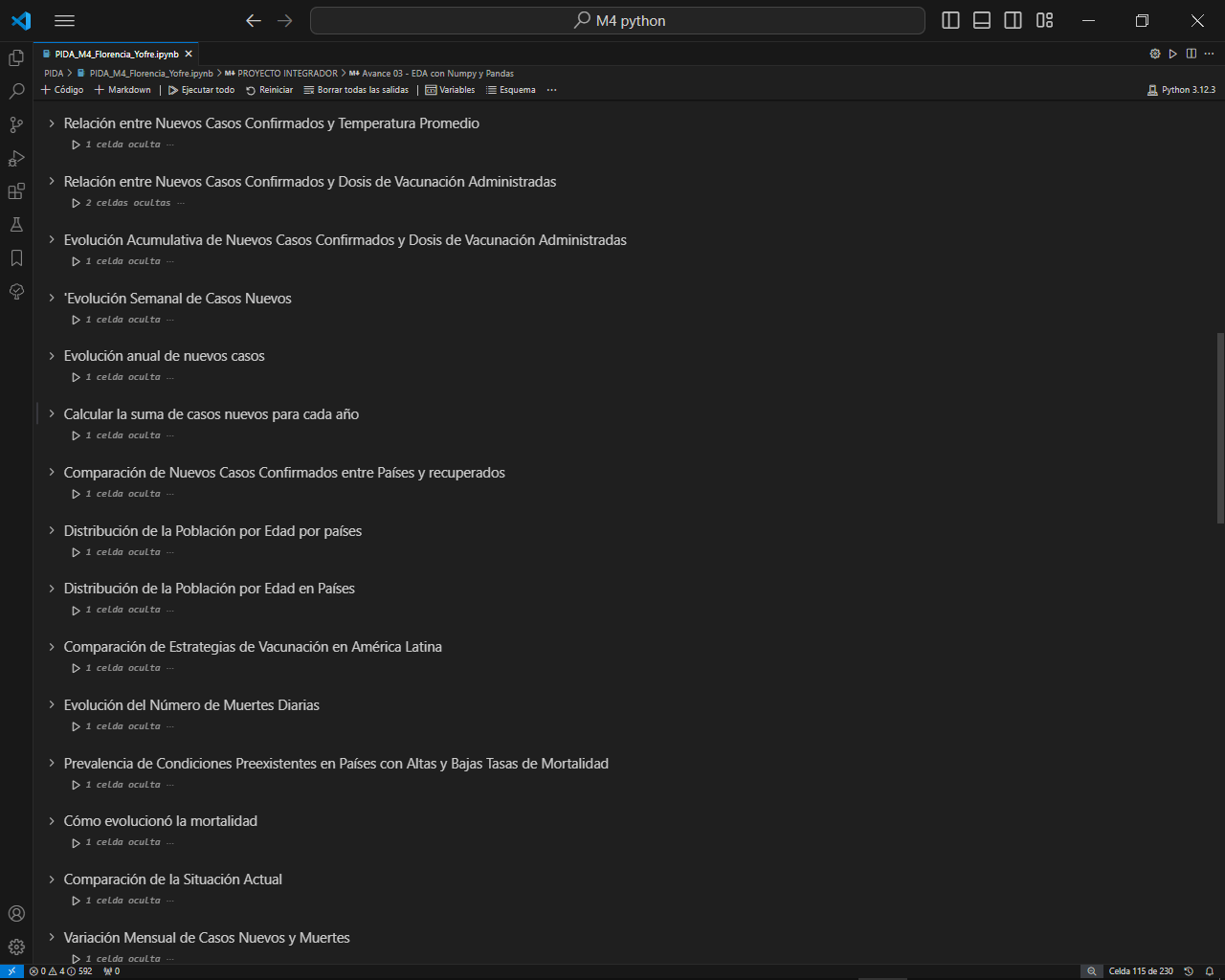
Con la información más clara y organizada, se pudo avanzar hasta el tercer avance y realizar el Análisis Exploratorio de Datos (EDA). En este tercer avance, realicé un análisis más profundo y sofisticado de los datos relacionados con la incidencia de COVID-19, centrándome en el análisis de series temporales para comprender la evolución de elementos específicos del conjunto de datos. Algunas de las gráficas generadas fueron:

Se generaron las siguientes gráficas, entre otras:

* Evolución de Casos Activos vs. Recuperados: Para entender la dinámica de la pandemia y el progreso de las recuperaciones.
* Tasa de Crecimiento (%): Para analizar el ritmo al que se propagó el COVID-19 en diferentes países.
* Relación entre la Cobertura de Vacunación y la Reducción de Casos: Para evaluar la efectividad de las campañas de vacunación.
* Progreso de la Vacunación por País: Para comparar el avance de la vacunación entre los países seleccionados.
* Nuevos Casos y Temperatura Promedio: Para investigar si existen correlaciones entre el clima y la incidencia de nuevos casos.
* Impacto de la Urbanización en la Propagación del COVID-19: Para analizar cómo la densidad poblacional influye en la propagación del virus.
* Evolución Semanal de Casos Nuevos: Para identificar tendencias y picos en la aparición de nuevos casos semanalmente.
* Evolución Anual de Casos Nuevos: Para observar la evolución de la pandemia a lo largo del tiempo.
* Distribución de la Población por Edad en Países: Para comprender cómo las características demográficas pueden afectar la vulnerabilidad ante el COVID-19.
* Comparación de Estrategias de Vacunación en América Latina: Para analizar y comparar las diferentes políticas de vacunación implementadas.
* Evolución del Número de Muertes Diarias: Para evaluar la mortalidad diaria a lo largo del tiempo.
* Prevalencia de Condiciones Preexistentes en Países con Altas y Bajas Tasas de Mortalidad: Para identificar posibles factores de riesgo.
* Variación Mensual de Casos y Muertes: Para observar fluctuaciones mensuales en los datos.

Además, investigué posibles correlaciones entre las distintas características para obtener una visión más completa y detallada de los datos.





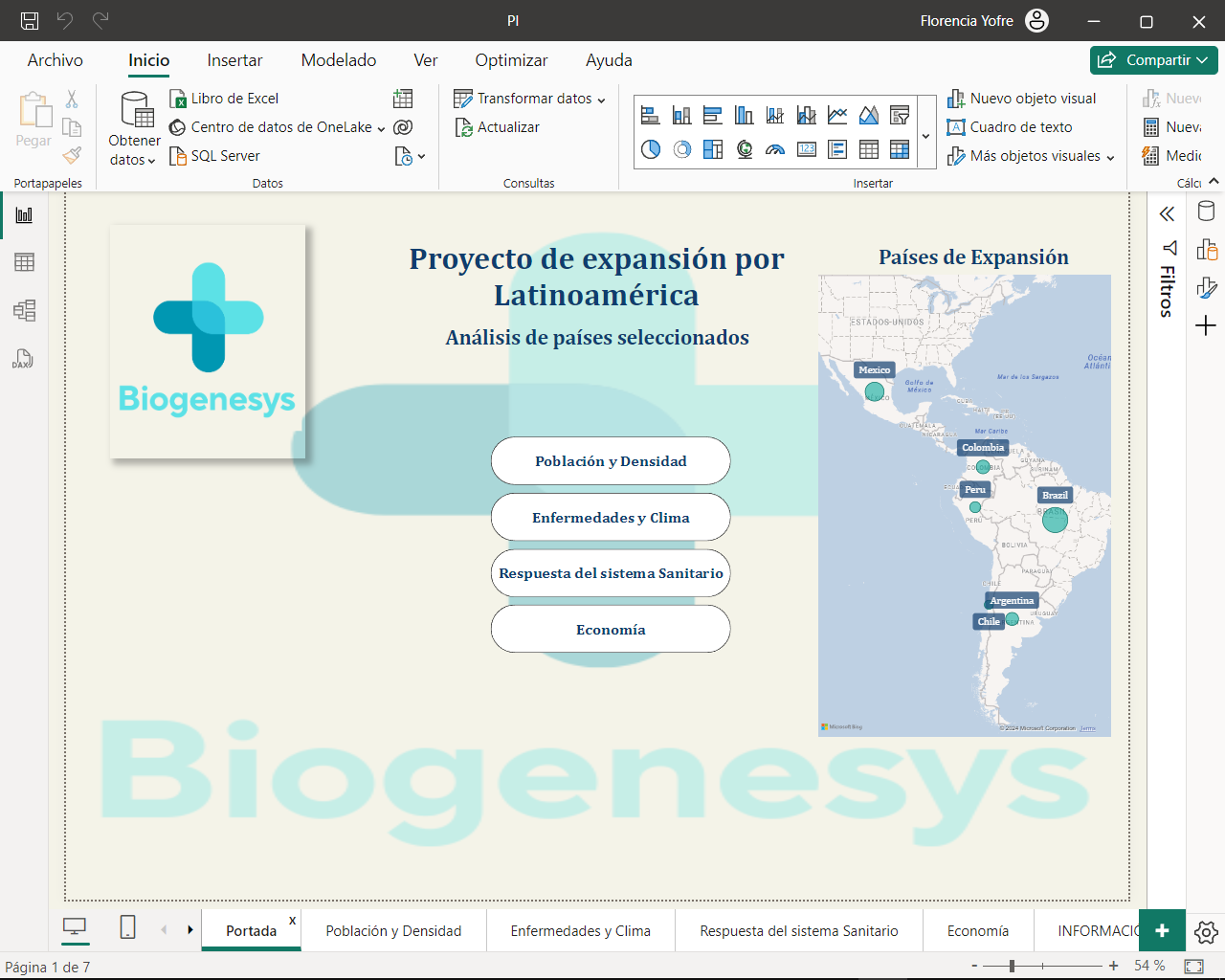
# Cuarto Avance:

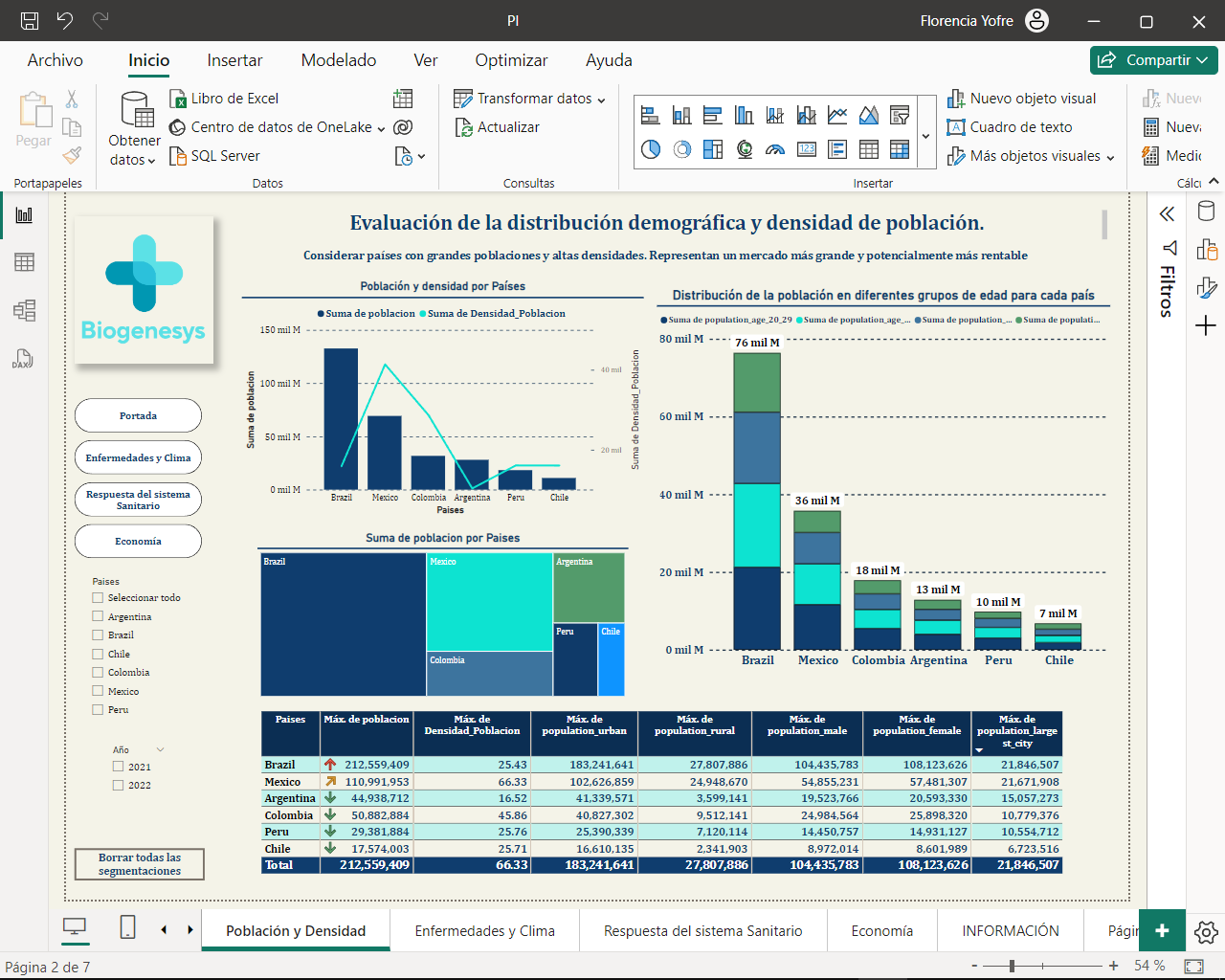
Con el análisis de datos ya finalizado, integré toda la información y los resultados obtenidos en Power BI. Esta integración permitió:

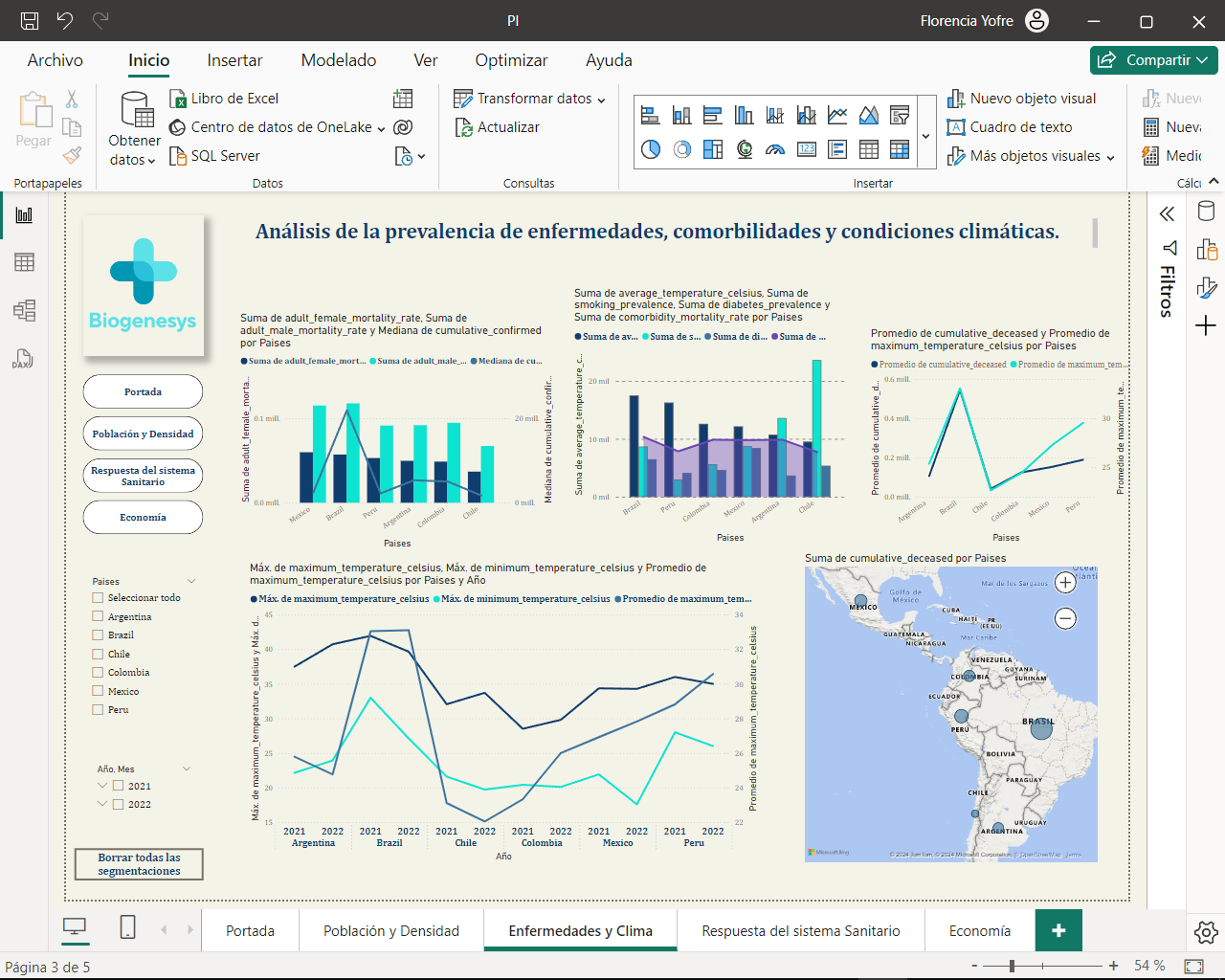
1. Crear Paneles Interactivos: Que facilitan la visualización dinámica de los datos y permiten a los usuarios explorar los resultados de manera más intuitiva.
2. Análisis Comparativo en Tiempo Real: Que permite comparar rápidamente diferentes escenarios y métricas clave.

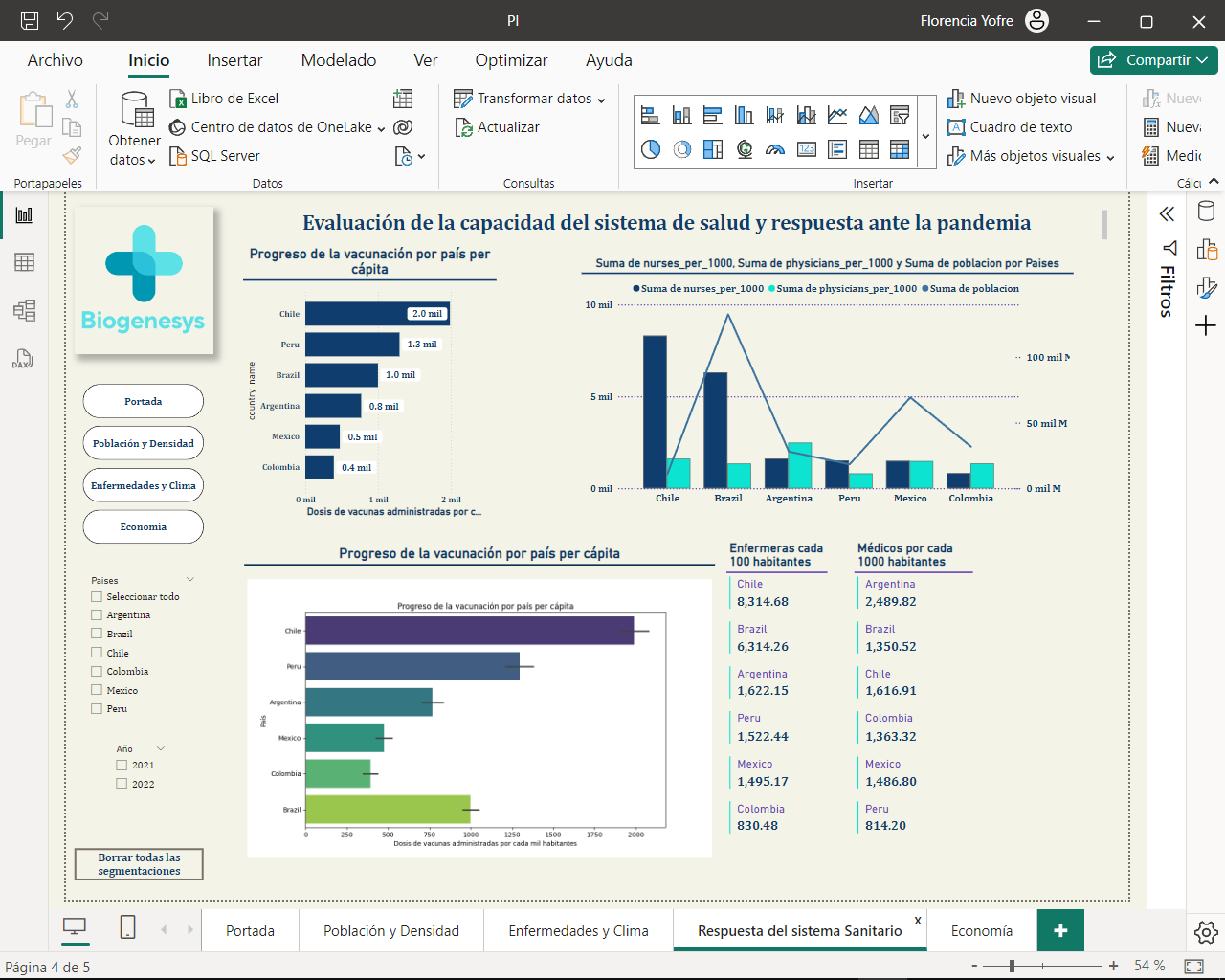
El dashboard realizado en Power BI incluye las siguientes secciones:

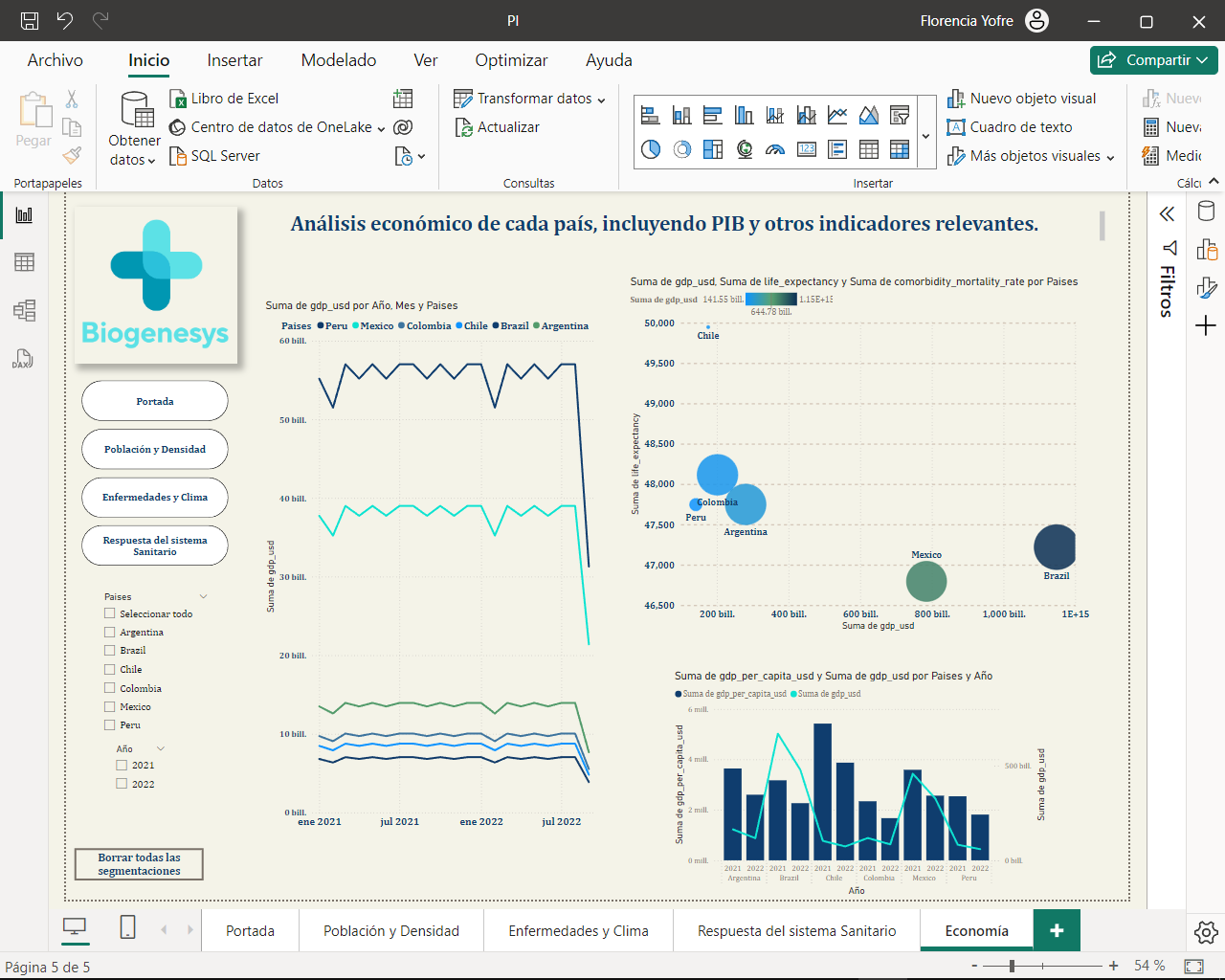
1. Portada: Introducción al proyecto de expansión por Latinoamérica, destacando el análisis de los países seleccionados.
2. Población y Densidad: Análisis demográfico y de densidad poblacional en los países seleccionados.
3. Enfermedades y Clima: Relación entre enfermedades prevalentes, condiciones climáticas y la incidencia del COVID-19.
4. Respuesta del Sistema Sanitario: Evaluación de la infraestructura sanitaria y la capacidad de respuesta ante la pandemia.
5. Economía: Análisis económico y su impacto en la capacidad de respuesta y recuperación de los países.











**Conclusiones y Recomendaciones**

Síntesis de los Resultados:

El análisis detallado y las visualizaciones en el dashboard han permitido obtener una visión clara y precisa de la situación sanitaria en los países analizados. A continuación, se resumen las conclusiones principales:

1. **Identificación de Ubicaciones Óptimas**: Brasil, México y Argentina se identificaron como los países con mayor necesidad y potencial para la expansión de laboratorios debido a su alta incidencia de COVID-19 y menor cobertura de vacunación.
2. **Infraestructura Sanitaria**: Es esencial invertir en mejorar la infraestructura sanitaria en regiones específicas de Colombia y Perú, donde se identificaron deficiencias significativas.
3. **Estrategias de Vacunación**: Es necesario intensificar las campañas de vacunación en áreas urbanas de Chile y México, donde la propagación del virus es más rápida.

**Respuesta Principal del Proyecto**:

Las ubicaciones óptimas para la expansión de laboratorios farmacéuticos de Biogenesys en Latinoamérica son:

1. **Brasil**: Debido a su alta población y la necesidad de mejorar la cobertura de vacunación.
2. **México**: Por su gran tamaño demográfico y la infraestructura sanitaria existente que requiere mejoras.
3. **Argentina**: Por su combinación de alta incidencia de COVID-19 y la oportunidad de mejorar la infraestructura sanitaria y la cobertura de vacunación.

**Reflexión Personal**

Durante este proyecto, he adquirido habilidades esenciales en el análisis de datos, limpieza de datos, uso de herramientas avanzadas como Python y Power BI, y en la interpretación de insights significativos para la toma de decisiones estratégicas. Si tuviera que volver a empezar este proyecto, me enfocaría aún más en la automatización de procesos de limpieza de datos y en la implementación de funciones de orden superior para optimizar el análisis. Esto podría acelerar el proceso y mejorar la eficiencia del análisis de grandes volúmenes de datos.

**EXTRA CREDIT**

Aunque no implementé todas las mejoras en esta etapa, investigué sobre funciones de orden superior que pueden ser muy útiles en futuros proyectos. Algunas de estas funciones incluyen:

1. Funciones de Map: Para aplicar operaciones a todos los elementos de un iterable.
2. Funciones de Filter: Para seleccionar elementos que cumplan ciertas condiciones.
3. Funciones de Reduce: Para realizar agregaciones en una lista.
4. Funciones Lambda: Para definiciones rápidas y simples de funciones.
5. Funciones de Iteración (Itertools): Para manipulaciones eficientes de iterables.

Estas mejoras potenciales contribuirían a hacer el análisis de datos aún más robusto y eficiente.

Este informe detalla de manera completa y precisa el proceso de análisis y las conclusiones del proyecto de expansión estratégica de Biogenesys en Latinoamérica, proporcionando un marco sólido para la toma de decisiones estratégicas informadas.