**Carrera: Data Analytics**

**Módulo 4**

# **Biogenesys**

**Nombre del autor**: Gabriel Salomon

**Email:** gabrielsalomon023@hotmail.com

**Cohorte**: DA-FT08

**Fecha de entrega**: 17/11/2024



**Institución:**

Desde “Biogenesys”, realizamos un análisis exhaustivo del virus “COVID-19” en la base de datos “data\_latinoamerica.csv” durante los últimos 2 años de dicho dataset. Utilizando la herramienta de Business Intelligence “Power BI”, con el objetivo de obtener insights valiosos sobre el comportamiento del virus e identificar oportunidades de expansión en 6 países.

**Introducción**:

La empresa farmacéutica “Biogenesys” busca identificar ubicaciones estratégicas para expandir sus laboratorios en Latinoamérica. Este proyecto tiene como objetivo optimizar la respuesta a los efectos de la pandemia y post pandemia, mejorando el acceso a vacunas en la región. Para lograrlo, la estrategia está basada en un análisis integral de datos relacionados con la incidencia del COVID-19, las tasas de vacunación y la infraestructura sanitaria disponible en los países seleccionados: Argentina, Brasil, Chile, Colombia, México y Perú.

**Objetivos del proyecto**:

Para alcanzar las metas propuestas, el proyecto se guía por cuatro objetivos clave:

Análisis exploratorio de datos(EDA): Identificar tendencias y oportunidades mediante estadísticas descriptivas y visualizaciones, explorando las variables que impactan el COVID-19.

Limpieza y preparación de datos: Garantizar la calidad de los datos mediante técnicas de limpieza, minimizando inconsistencias y optimizando su confiabilidad para el análisis.

Optimización del proceso ETL: Implementar estrategias eficientes de extracción, transformación y carga para facilitar el manejo de los datos y acelerar la toma de decisiones.

Desarrollo de dashboards interactivos: Diseñar visualizaciones interactivas y dinámicas en Power BI para explorar los datos desde distintas perspectivas.

Los datos analizados aportarán insights esenciales sobre los países de interés y respaldarán la identificación de ubicaciones óptimas para la expansión de laboratorios y centros de vacunación. Este análisis, además, permitirá comprender las dinámicas demográficas y de salud pública en cada país, proporcionando un panorama completo para la toma de decisiones.

**Desarrollo del proyecto:**

El análisis comenzó con la exploración del dataset proporcionado por Ingeniería de Datos. Este contenía 12,216,057 registros y 50 columnas en formato CSV, lo que facilitó su manipulación respecto al dataset original de 21 GB. También se recibió un archivo complementario con información detallada sobre las variables seleccionadas, útil para la preparación y limpieza de los datos.

Carga y verificación inicial de datos:

Se utilizó Python para leer el archivo y comprobar la coherencia del dataset con las especificaciones iniciales.

A través de numpy y pandas, se verificaron dimensiones y se identificaron valores faltantes.

Se generaron filtros iniciales para centrar el análisis en los seis países seleccionados (Argentina, Brasil, Chile, Colombia, México y Perú).

Limpieza y filtrado de datos:

Se eliminó información irrelevante (datos asociados a subregiones) y se trabajó exclusivamente con registros relevantes para cada país.

Los pocos valores nulos se completaron con el valor 0 (cero).

Se estableció el 01/01/2021 como punto de inicio del análisis, eliminando datos irrelevantes previos.

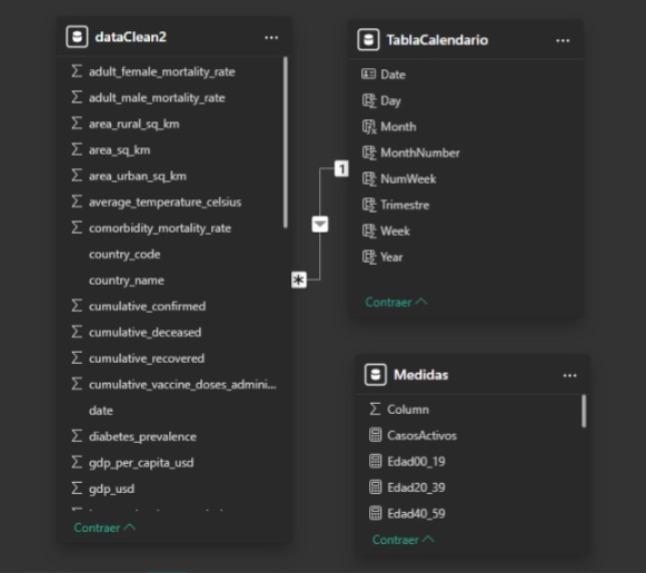
**Detalle de tablas:**

dataClean2:  
Es el nombre que le dimos a la tabla final antes de exportarla como archivo .csv desde Python, aquí están contenidos todos los datos que fueron brindados. Lo cual contenía una tabla con 50 columnas (country\_name, new\_confirmed, new\_decesed, date, etc.) y más de 12.000.000 de filas.

TablaCalendario:  
Creamos esta tabla para realizar dentro de ella la creación de columnas que consideramos relevantes para nuestro análisis, y poder trabajar con mayor comodidad a lo largo del tiempo. Las columnas creadas fueron Date, Year, Month, MontNumber, Trimestre, Day, Wee, NumWeek, Day. Esta tabla fue conectada con “dataClean2” mediante el campo “Date”, el cual establecimos como Tabla de fechas para utilizar como eje en todo el dashboard.

Medidas:  
Dentro de esta tabla creamos 20 medidas que consideramos relevantes e interesantes para nuestro análisis y que nos brinde mayor flexibilidad a la hora de crear gráficos interactivos. Las medidas DAX creadas fueron: CasosActivos, Edad00\_19, Edad20\_39, Edad40\_59, Edad60\_79, EfectividadVacunacion, Mayores80, NuevosCasos, NuevosFallecidos, NuevosRecuperados, PromedioFallecidos, PromedioNuevosCasos, PromedioRecuperados, SexoFemenino, SexoMasculino, TasaFatalidad, TotalCasos, TotalDosis, TotalFallecidos, TotalRecuperados.

De esta manera nos quedó el modelo:



Análisis descriptivo inicial:

Se generaron estadísticas resumidas con .info() para evaluar la distribución de las variables clave.

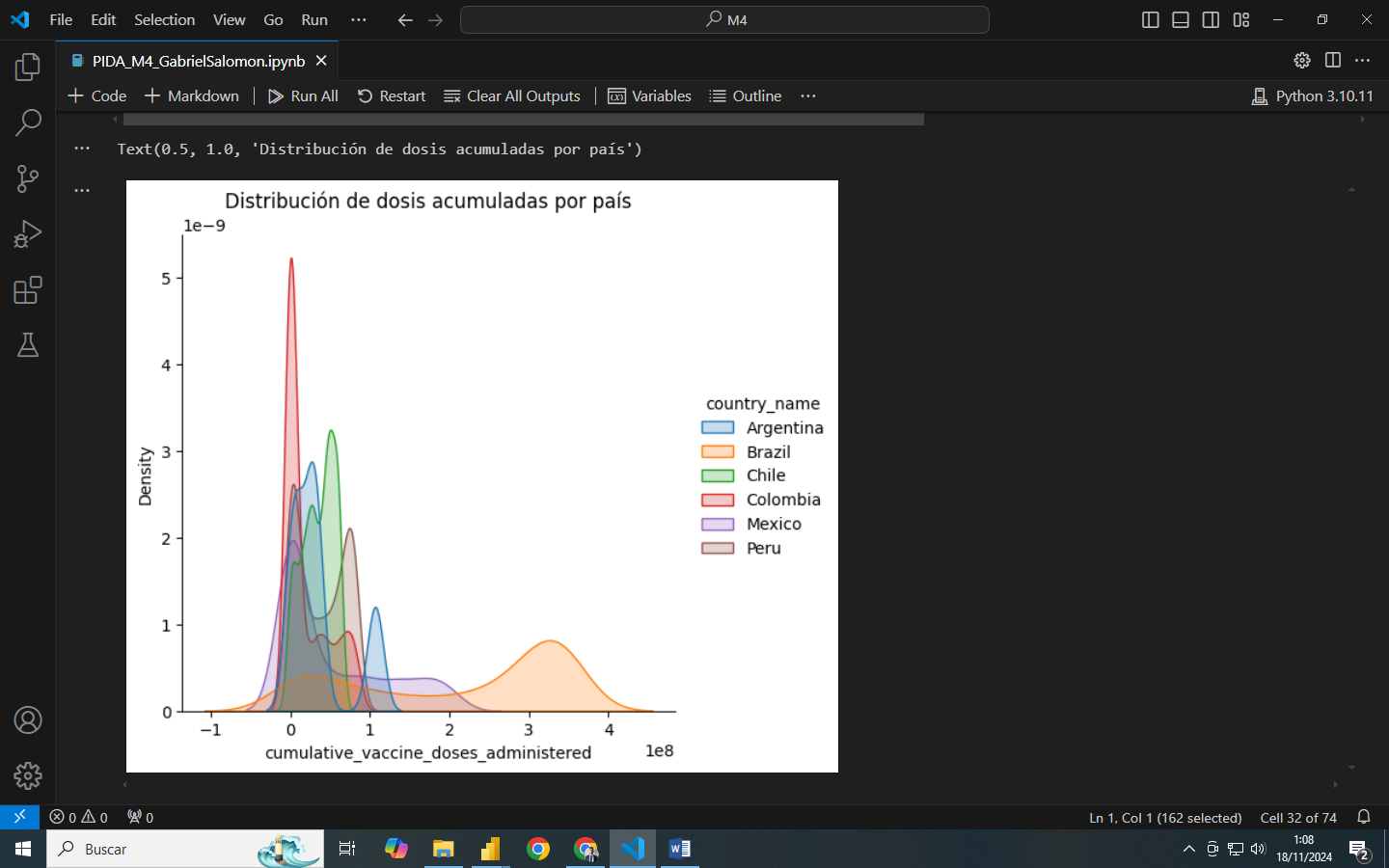
Los resultados se visualizaron utilizando gráficos básicos de dispersión, histogramas y diagramas de caja para identificar patrones iniciales en los datos.

Preparación del dataset final:

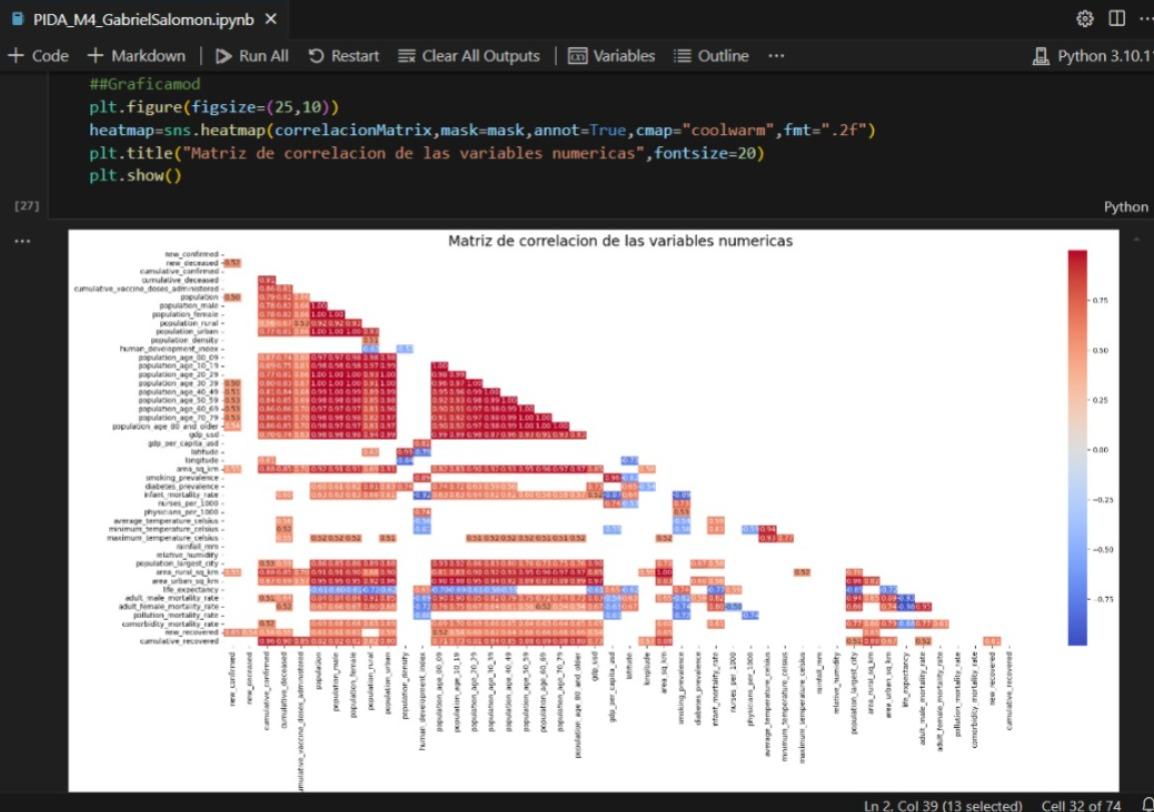
Los datos procesados se guardaron en un archivo limpio para evitar repetir el proceso de limpieza en análisis posteriores.

Exploración visual inicial:

Se desarrollaron gráficos preliminares con Matplotlib y Seaborn para observar relaciones entre variables como temperatura promedio, incidencia de COVID-19, y cantidad de dosis administradas.



Se detectaron correlaciones fuertes (mayores a 0.5) mediante una matriz de correlación visualizada en un heatmap, lo que ayudó a seleccionar variables relevantes para análisis posteriores.



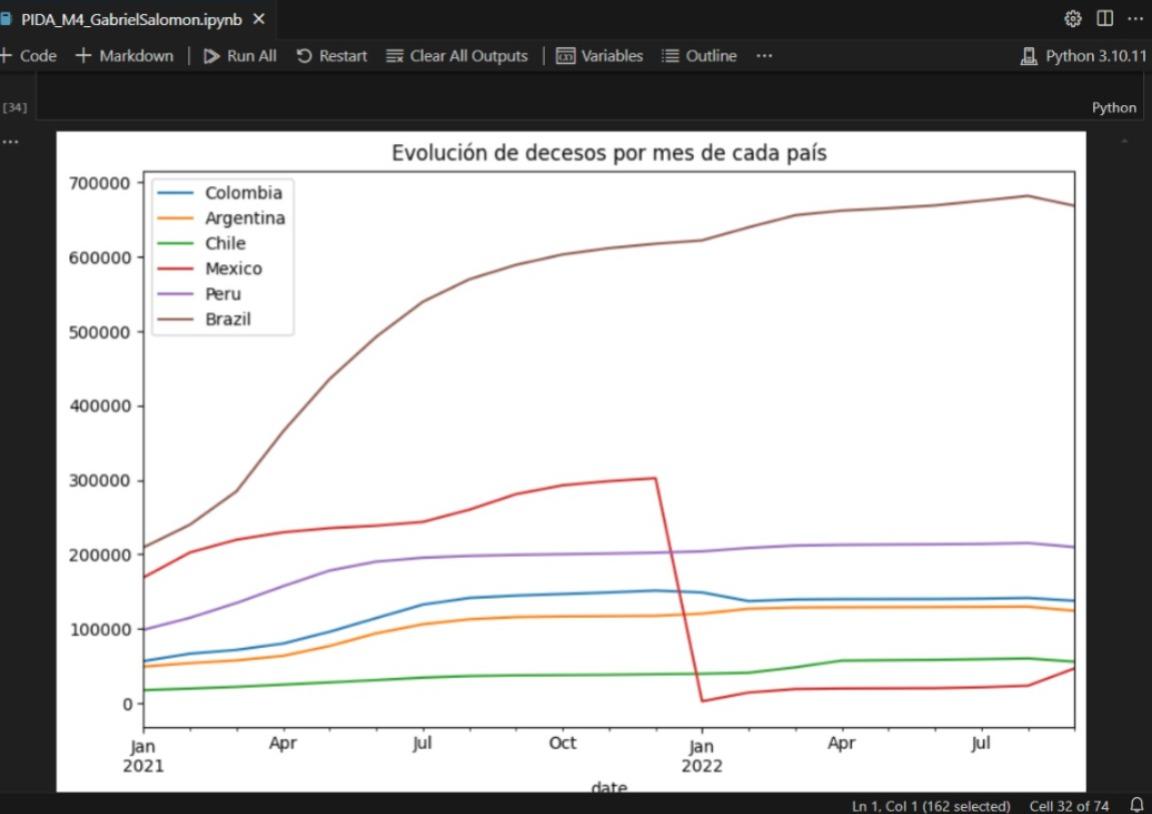
Análisis exploratorio avanzado:

Con los datos limpios, se profundizó en el análisis exploratorio utilizando técnicas avanzadas para extraer insights significativos.

Tendencias temporales:

Se estudió la evolución mensual de los casos confirmados, decesos y dosis administradas. Esto permitió identificar picos relacionados con factores estacionales y eventos específicos.

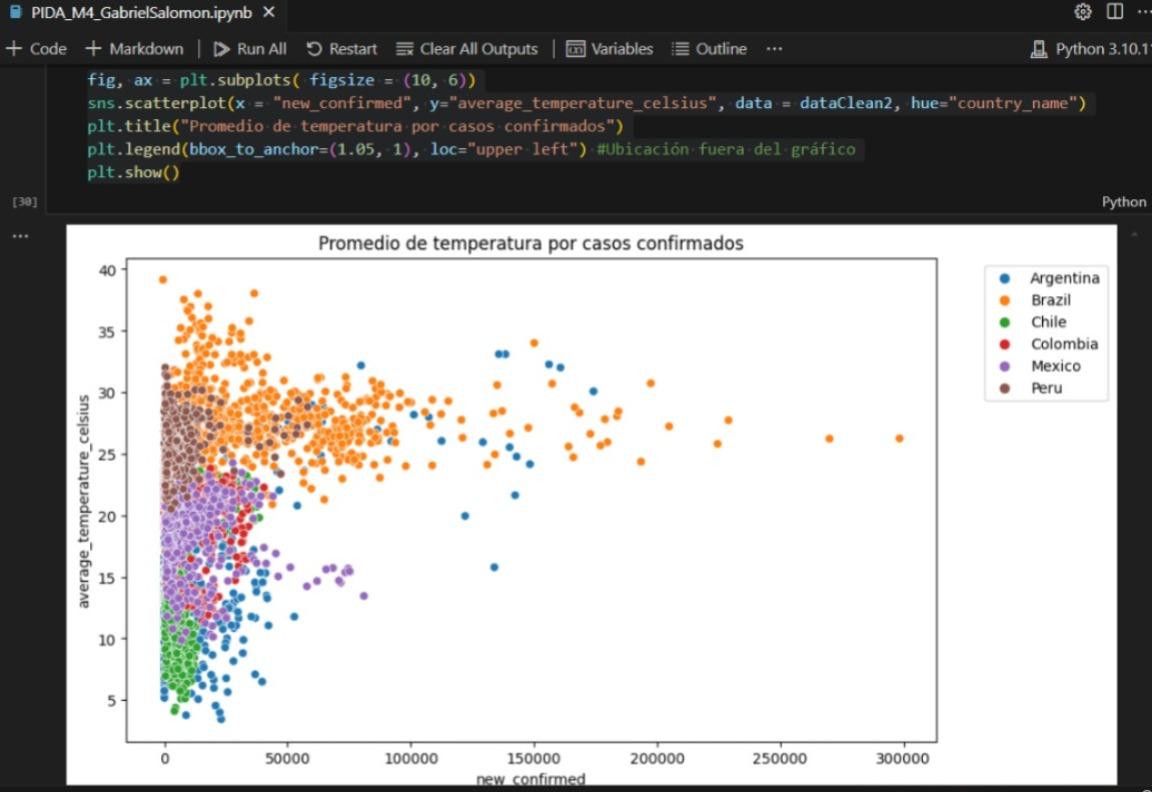
Brasil mostró los números más altos de incidencia y mortalidad, mientras que Chile destacó por su manejo efectivo de la pandemia, con una alta proporción de vacunación y bajas tasas de mortalidad.



Relación entre temperatura y COVID-19:

Se analizó la relación entre la temperatura promedio y los casos confirmados. Aunque algunos países mostraron correlaciones débiles, en otros se detectaron patrones más definidos.

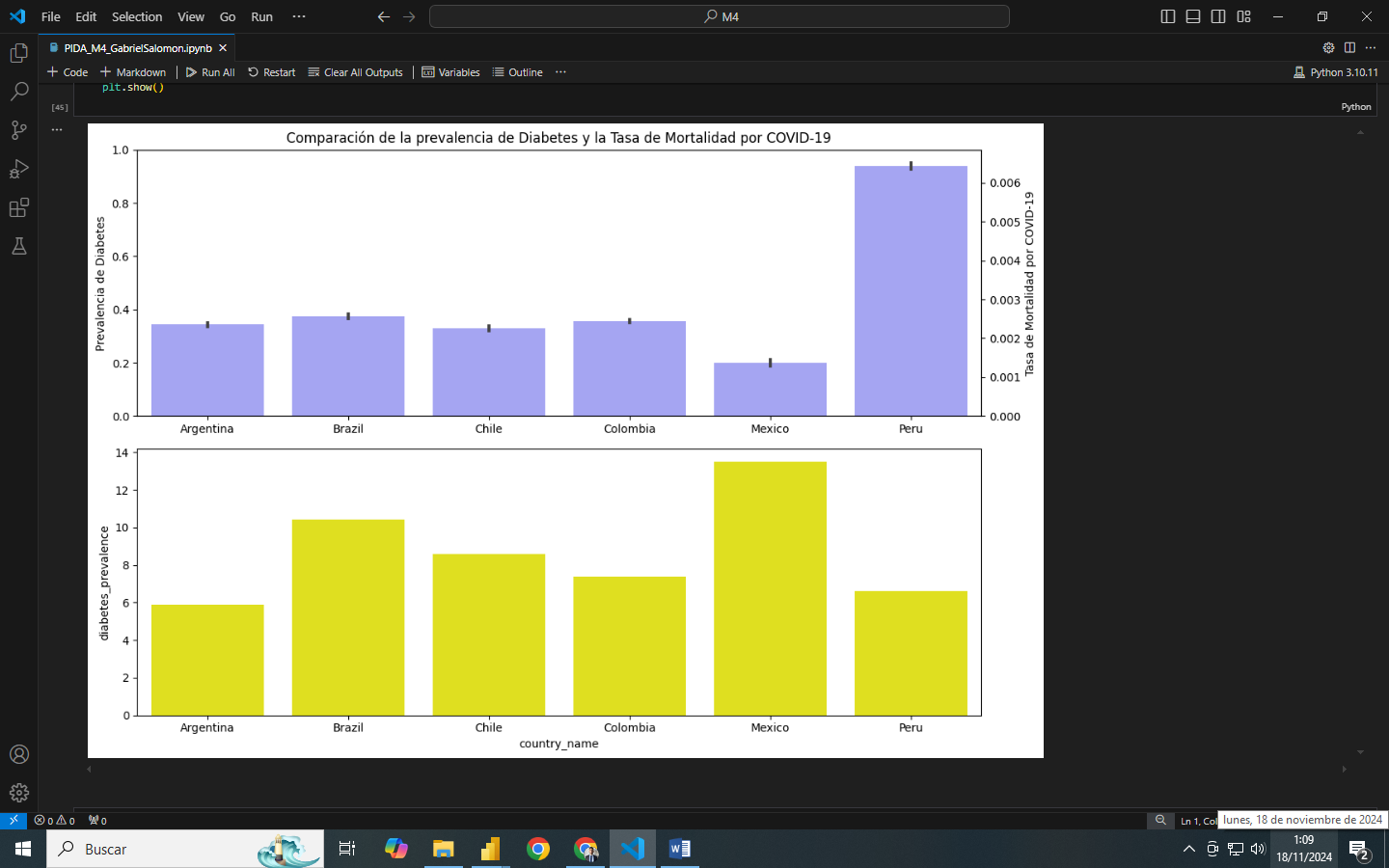
Estudios similares se realizaron con la temperatura y las muertes acumuladas, encontrando comportamientos paralelos.



Mortalidad por género y grupos etarios:

Se utilizaron gráficos de barras y violinplots para estudiar la mortalidad por género y la distribución de la población vacunada en distintos grupos de edad.

Se identificó que los grupos más afectados varían entre países, pero las comorbilidades como la diabetes y el tabaquismo tuvieron un impacto significativo en todos los países analizados.



Índices de desempeño sanitario:

Se calculó la tasa de letalidad (decesos/confirmados) y la tasa de mortalidad (decesos/población). México y Perú presentaron las tasas de letalidad más altas, mientras que Argentina y Chile tuvieron un mejor desempeño relativo.

Visualización y análisis en Power BI

En la última etapa del proyecto, se consolidaron los hallazgos en un dashboard interactivo desarrollado en Power BI.

Este incluyó:

Población y vacunación:

Visualizaciones dinámicas que permiten observar la población urbana y rural, los grupos etarios vacunados y la evolución de las dosis administradas en cada país.

Se detectaron inconsistencias en los datos poblacionales, lo que sugiere posibles vacíos en la calidad del dataset que deben abordarse en futuros análisis.

Performance del COVID-19:

Gráficos interactivos para comparar casos confirmados, decesos y recuperados a lo largo del tiempo.

Mapas geoespaciales con indicadores de vacunación y tasas de fatalidad para identificar áreas críticas en cada país.

Índice de expansión:

Se diseñó un índice combinando la tasa de letalidad y las vacunas administradas. Este índice destacó las áreas donde la expansión de laboratorios tendría mayor impacto, priorizando Brasil, México y Colombia.

**Análisis general del tablero**:

Creamos un tablero completo en Power BI el cual nos permite analizar el rendimiento de Biogenesys, que además de un informe general contiene un enfoque especial en la evolución de los casos y en incidencias del COVID-19. Haciendo foco en los insights mas valiosos, a lo amplio del tablero podremos apreciar indicadores clave como pueden ser la variación del virus a lo largo de los meses evaluados, un mapa para ver la distribución por país, si afectó mas a la zona rural o a la ciudad, etc. Con segmentaciones precisas para su análisis por país, o por año.

Iniciamos con una portada sencilla la cual nos brinda una presentación de la empresa “Biogenesys” con una breve descripción, además contiene su logo y 3 botones interactivos para acceder al informe principal, al informe de variaciones o de incidencia.



Pasamos al informe general, el cual contiene 6 tarjetas con indicadores clave de rendimiento como lo son los casos totales, el total de fallecidos, el total de recuperados, los nuevos casos, los nuevos fallecidos y la efectividad por vacunación.

Además nos vamos a encontrar con un gráfico de columnas 100% apiladas haciendo referencia a las zonas afectadas por país, un gráfico de mapa visualizando los países que estamos evaluando, un gráfico de líneas que compara el total de contagios, recuperados y fallecidos mes a mes, y un gráfico de columnas apiladas que muestra la totalidad de los casos comparándolos entre un año y otro.

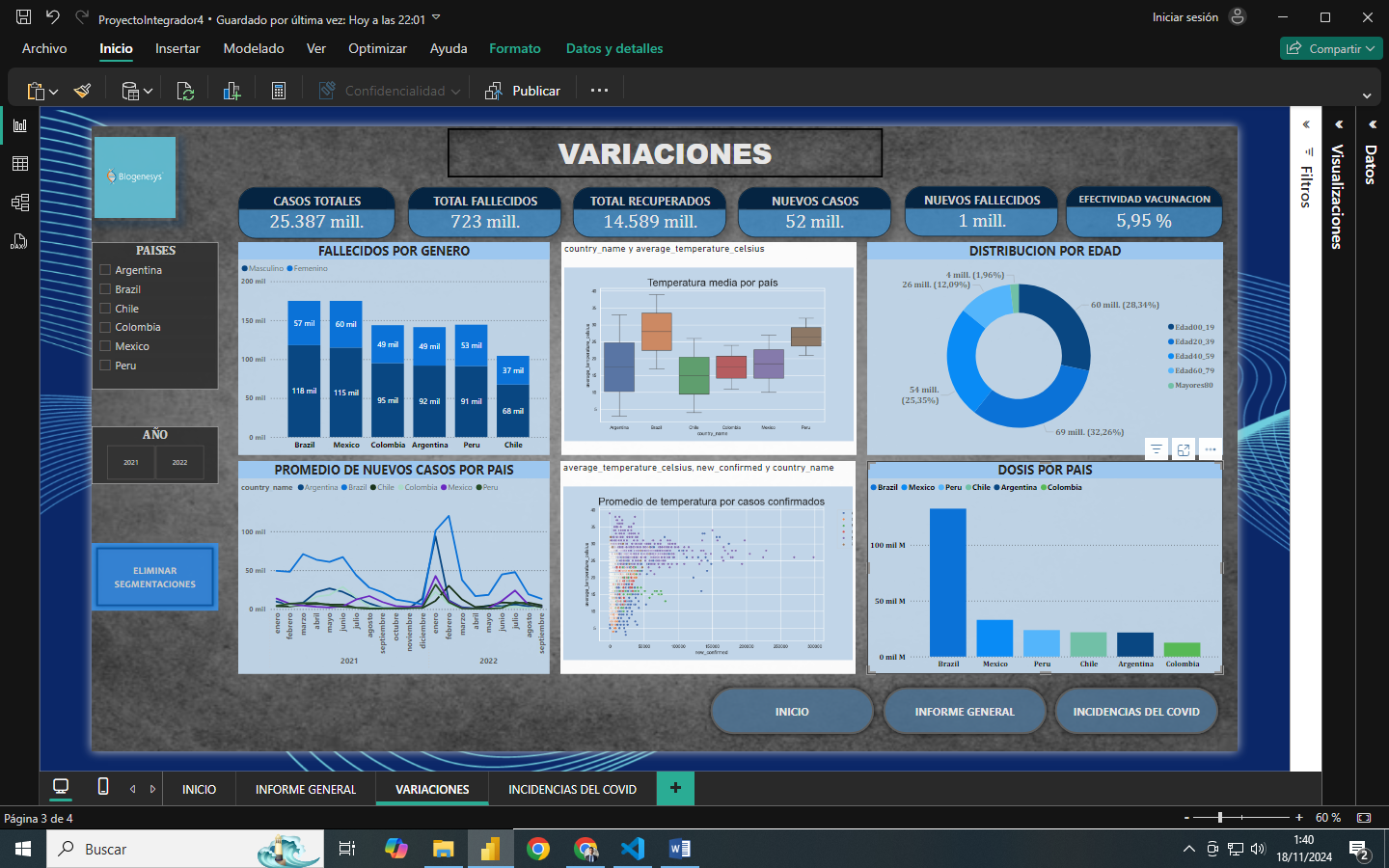
En el costado izquierdo se ubican los segmentadores para poder aplicar el filtro necesario para observar insights detallados a necesidad, donde se podrá elegir por año, o por el país. Debajo de todos ellos, se podrá observar un botón para eliminar las segmentaciones aplicadas.



Pasamos al tablero de “Variaciones” que hace foco en distintas variabilidades que consideramos como datos importantes y de gran relevancia, donde vamos a encontrar las mismas 6 tarjetas indicadoras principales. Una característica importante de éste apartado es que no solo tenemos gráficos creados en Power BI, sinó que también importamos gráficos desde Python.

Tendremos un gráfico de columnas apiladas comparando la cantidad de fallecidos por género, a su lado insertamos un de Python que nos muestra la temperatura media por país, también a su lado encontramos un gráfico de anillos mostrando la distribución de edad de la población.

En la parte inferior izquierda del tablero tenemos un gráfico de líneas comparando por mes, el promedio de nuevos casos por país, además hay otro gráfico importado desde Python pero éste es de dispersión para poder observar el promedio de temperatura por casos confirmados. Por último nos queda un gráfico más de columnas apiladas comparando la cantidad de dosis por país. De igual manera todos los datos de esta tabla se pueden segmentar, conteniendo los mismos segmentadores que fueron establecidos por defecto.



Por último pasamos al tablero de “Incidencias del COVID” que hace foco en puntos críticos con insights valiosos para cualquier análisis que deseen implementar, allí también vamos a encontrar las mismas 6 tarjetas indicadoras principales. Además en éste apartado también tenemos gráficos creados en Power BI y Python como complemento.

Primeramente en la parte superior izquierda vamos a observar un gráfico de embudo que va mostrando los nuevos fallecidos por mes. En el medio de la pantalla aún en la parte superior, vamos a ver un gráfico de barras apiladas indicando la taza de fatalidad, y a su derecha sigue un gráfico importado desde Python que nos va a mostrar la distribución de nuevos casos confirmados.

En la parte inferior izquierda del tablero tenemos un gráfico de áreas mostrando la tasa de fatalidad por casos, vemos hacia el medio y tenemos un gráfico de líneas mostrando los nuevos recuperados por mes, y por último en la parte derecha nos queda un gráfico más importado desde Python que nos revela la distribución de dosis por país.

Llegando así a completar el “Patrón Z” recorriendo el gráfico con datos valiosos y claros, de fácil interpretación.



Como norma general se hace la mención de que cada sección del tablero contiene sus botones de navegabilidad por el mismo, los correspondientes segmentadores para sus gráficos y debajo un botón para deshacer las segmentaciones aplicadas. De esta forma nuestro público podrá disfrutar de una correcta interacción mostrando sus valores clave de manera simple.

**Conclusiones y recomendaciones:**

Los hallazgos del proyecto permiten identificar las ubicaciones óptimas para la expansión de Biogenesys.

Países prioritarios:

Brasil, México y Colombia presentan las mayores tasas de letalidad y los menores niveles de vacunación relativa, lo que sugiere un alto impacto potencial de los nuevos laboratorios.

Chile como modelo:

Aunque Chile no es prioritario para la expansión, sus datos indican un manejo eficiente de la pandemia que puede servir como referencia para estrategias en otros países.

Validación de datos:

Es necesario verificar los cálculos de Perú, donde se detectaron posibles inconsistencias en los datos de tasa de letalidad.

Recomendaciones futuras:

Ampliar el análisis incorporando datos sobre infraestructura sanitaria, distribución logística de vacunas y factores socioeconómicos, complementando la evaluación actual para robustecer la estrategia de expansión.

**Reflexión final:**

Este proyecto destacó la importancia del manejo integral de datos y herramientas como Python y Power BI para apoyar decisiones estratégicas basadas en evidencia. Aunque aún encuentro más intuitivo trabajar con Power BI para visualizaciones, seguiré profundizando mis habilidades en Python para integrar ambas herramientas en futuros análisis.

Ha sido un viaje apasionante adentrarme en el mundo de Power BI.

**Reflexión personal**:

Ha sido un viaje apasionante adentrarme en el mundo de Power BI. Este proyecto me ha permitido crecer tanto a nivel técnico como en mi capacidad para analizar datos y generar insights valiosos. Estoy especialmente orgulloso de haber podido crear reportes personalizados y de haber descubierto la complementación que puedo implementar junto con Python. Aunque sé que aún tengo mucho por aprender, estoy emocionado por seguir explorando las posibilidades de esta herramienta.

Herramientas utilizadas:

* Visual Studio Code
* Python ( matplotlib, seaborn, numpy, pandas )
* Power BI
* DAX