**Expansión Estratégica de Biogenesys con Python**

**Nombre del autor: Mariugenia Rincon**

**Email: rinconmariugenia2@gmail.com**

**Cohorte: DA-PT04**

**Fecha de entrega:** 23/12/2024

**Institución:** Empresa Biogenesys dedicada al sector salud.

Logo: Creado con inteligencia artificial.



Paleta de Colores:

Color principal: #0F2738

Color secundario: #00A7E2, #E9310F, #0860A0

Color terciario: #C5CAD2

Color de acento: #FEAC01, #7BD100

# Introducción

# El presente informe detalla los resultados obtenidos en el proyecto desarrollado para la empresa ficticia Biogenesys, líder en la industria de salud en América Latina. Este proyecto fue diseñado con el propósito de transformar datos relacionados con indicadores de salud sobre el COVID-19, factores de riesgo y estadísticas poblacionales en visualizaciones y análisis que respalden decisiones estratégicas en áreas clave.

# El principal objetivo del proyecto fue proporcionar herramientas analíticas y dashboards interactivos que permitan a Biogenesys evaluar la correlación entre distintos factores de riesgo, como prevalencia de enfermedades crónicas, mortalidad y nuevos casos, facilitando la identificación de patrones críticos y áreas de mejora.

# Entre los objetivos organizacionales alcanzados, destacan:

# Apoyo a la toma de decisiones informadas: Análisis detallado de indicadores de salud para priorizar estrategias y recursos en las áreas de mayor impacto.

# Visualización eficiente de datos: Creación de dashboards que consolidan métricas clave como tasas de mortalidad, prevalencia de enfermedades y nuevos casos, siguiendo principios de diseño centrados en la claridad y accesibilidad.

# Detección de patrones clave: Identificación de correlaciones entre factores de riesgo y su impacto en la salud poblacional, proporcionando a la organización una base sólida para la planificación y ejecución de programas de mejora.

# Desarrollo del proyecto

# 

# AVANCE 1:

Se realizó la carga del archivo con los datos proporcionados, se realizó una exploración de los datos y su contenido así como se verificaron las dimensiones del mismo.

Se realizó un filtrado para considerar sólo los países requeridos a saber: Colombia, Argentina, Chile, México, Perú y Brasil, además se filtraron las fechas para acotar solo a aquellas mayores a 1 de enero del 2021. Se constató que habían muchos datos nulos en algunos millones, por lo que se decidió hacer un nuevo filtro por Location\_key correspondiente a los paises seleccionados ya que era más representativo.

Se verificó de nuevo la cantidad de valores faltantes en el nuevo dataset acotado y se procedió a imputar con la media correspondiente a la mayoría de ellos, y los que tenían mayor cantidad de valores nulos se los reemplazó con cero.

El dataset final se redujo a 3744 filas y 50 columnas y se procedió a exportarlo como un archivo .csv de forma que en adelante se trabaje con ese y tener mejor eficiencia a la hora de cargarlo y manejarlo.

Se realizó una exploración estadística a todos los campos mediante .describe() para las variables numéricas y se acreó también una función para obtener la moda, varianza, y rango ya que éstos no son proporcionados por .describe()

# AVANCE 2:

Inicialmete, se colocó la columna Date como indice para poder hacer un análisis de los datos a lo largo del tiempo. Se creó la matriz de correlación para ver si existe relacion entre las variables numéricas.

En la gráfica anterior se presentan las variables con mayor correlación, para identificarlas más rápido, se definió un umbral cuyo valor absoluto sea mayor a 0.5 y se graficaron en un heatmap. Ya teniendo una idea de las variables que tienen mayor correlación, se podría hacer un análisis más detallado de estas variables, de esto se podría hacer en un análisis posterior y una toma de decisiones para la creación de un modelo.

Se crearon graficos de cajas e histogramas para las variables numéricas con el objetivo de verificar la distribución de los datos.

Para comprender el comportamiento de la temperatura con respecto a los nuevos confirmados y nuevos decesos se realizó un grafico de dispersión para cada uno agrupado por país, de él no se observa con claridad si existe relación o no, por lo que debería hacerse un análisis separado por país.

Adicionalmente se hicieron gráficas de comparación de la evolución de vacunas por pais con respecto al tiempo, y con respecto a grupos etarios, gráficas para comparar la tasa de mortalidad entre masculino y femenino, análisis de correlación entre los factores de riesgo e indicadores demográficos y gráficos para analizar la tasa de mortalidad con respecto a la diabetes y al hábito de fumar.

# AVANCE 3:

Este avance se enfocó en realizar un análisis exploratorio de datos (EDA) con Numpy y Pandas.

Como una primera exploraión se decidió analizar el comportamiento de algunas variables de interés (new\_confirmed, new\_deceased) de manera semanal y anual, para esto se agruparon los datos por semana y por año.

Se realizó una gráfica de líneas agrupadas por pais, y se observó que el comportamiento de nuevos confirmado y de muertes es similar semanal, mensual y anualmente, donde al inicio de cada gráfica se observa mayoría de macos y van disminuyendo a lo largo del tiempo en todos los paises, excepto México que tuve un repunte de nuevo casis al final del año.Este comportamiento podría deberse a la implementación de vacunas y a los avances de la medicina en cuanto iban encontrando nuevos tratamientos para sanar a las personas con COVID-19.

Luego se realizaron gráficas de dispersión para evaluar si existía relación entre los nuevos casos confirmados y la temperatura de cada pais, del cual se determinó que no existe relación significativa.

Se realizó un gráfico de columnas para verificar cómo están distribuidos la cantidad de personas en grupos etarios, resultando que los grupos con mayor población son de las edades entre 10 años y 29años, y el menor el grupo de personas con 80 años o más, a medida que aumenta la edad se reduce la cantidad de personas en dicho grupo.

Se realizó un gráfico de columnas paraevaluar las estrategias de vacucación de cada país mediante la tasa de vacunación, donde resultó que Chile es el país que más vacunas administró, hay que destacar que el gráfico muestra tasas de más del 100% y esto puede deberse a que el cálculo considera el número total de dosis administradas, pero no toma en cuenta que muchas personas reciben más de una dosis (por ejemplo, esquemas de vacunación que requieren dos dosis o incluso refuerzos).

También se ccreó un gráfico para ver la evolución de los nuevos casos confirmados y la temperatura a lo largo del tiempo, en el cual no se observó ninguna coincidencia entre el aumento o descenso de la temperatura y los casos confirmados.

Se realizó un análisis del porcentaje de poblacion en zonas urbanas y casos confimados en esas zonas por país, aunque anteriormente se observó que Chile tuvo la tasa más alta de vacunación, en este ultimo grafico se observa que tuvo el mayor porcentaje de casos confirmados, ésto puede deberse a que:

- Los países con mayor porcentaje de población urbana suelen tener una mayor transmisión del virus, y Chile tiene una alta proporción de su población viviendo en áreas urbanas.

- El porcentaje de vacunación alto puede no haber coincidido con el inicio del brote en Chile. Si el virus ya se había propagado ampliamente antes de alcanzar un alto nivel de vacunación, la tasa de casos confirmados habría sido más alta.

- Chile implementó una estrategia de pruebas amplia y eficiente, lo que podría haber resultado en la identificación de una mayor proporción de casos confirmados en comparación con otros países, donde los casos pueden haber sido subestimados.

Se crearon gráficos para analizar la prevalencia de condiciones creexistentes en países con altas y aajas tasas de mortalidad, en ellos se observa que la diabetes afecta mucho mas a México y Brasil, por lo que se podría decir que estos paises son los que más han sufrido por la diabetes. Respecto a fumar se puede observar que afecta más a Chile y Argentina.

Para evaluar la evolución de la vacunación en América Latina y la reducción de casos de COVID-19 se realizó un gráfico de líneas donde se representó con una línea los casos nuevos a lo largo del tiempo y otra línea para el porcentaje de población vacunada, se observó que los casos fueron disminuyendo a medida que se vacunaban más personas, de igual forma en este gráfico el porcentaje asciende a más de 100% por lo que se explicó anteriormente.

# 

# EDA e insights

# En el contexto del proyecto desarrollado para Biogenesys, se llevaron a cabo diversas técnicas de EDA que proporcionaron insights valiosos sobre la evolución del COVID-19 en países de América Latina. A continuación, se detallan los hallazgos más destacados:

# 

# Correlación entre Variables Clave:

# Se identificaron correlaciones significativas entre variables como la prevalencia de diabetes, la tasa de tabaquismo y las tasas de mortalidad por COVID-19. Estos hallazgos sugieren que las comorbilidades juegan un papel importante en la gravedad de la enfermedad.

# Distribución de Nuevos Casos y Decesos a lo Largo del Tiempo:

# El análisis temporal mostró picos en los nuevos casos confirmados y decesos durante ciertos períodos, seguidos de disminuciones que coinciden con la implementación de campañas de vacunación masiva. Este patrón resalta la efectividad de las vacunas en la reducción de la propagación y letalidad del virus.

# Impacto de la Urbanización en la Propagación del Virus:

# Se observó que países con mayor porcentaje de población urbana presentaron tasas más altas de casos confirmados. Este hallazgo indica que las áreas urbanas, debido a su densidad poblacional, pueden ser más susceptibles a la rápida propagación del virus.

# Estrategias de Vacunación y su Efectividad:

# Chile destacó por su alta tasa de vacunación, superando el 100% debido a la administración de dosis adicionales o refuerzos. Este enfoque proactivo se correlacionó con una disminución más rápida en los casos nuevos, evidenciando la importancia de una estrategia de vacunación robusta.

# Distribución Etaria de la Población y Vulnerabilidad:

# Los grupos etarios entre 10 y 29 años representan la mayor proporción de la población en los países analizados. Sin embargo, las tasas de mortalidad fueron más elevadas en grupos de mayor edad, lo que subraya la necesidad de enfoques específicos para proteger a las poblaciones más vulnerables.

# Relación entre Temperatura y Propagación del Virus:

# No se encontró una relación significativa entre las variaciones de temperatura y el número de casos confirmados o decesos. Este resultado sugiere que factores climáticos pueden tener un impacto limitado en la propagación del COVID-19, en comparación con otros factores como las medidas de salud pública y el comportamiento social.

# Estos insights proporcionan una comprensión más profunda de cómo diversos factores han influido en la dinámica de la pandemia en América Latina. Las visualizaciones desarrolladas permiten a Biogenesys identificar áreas críticas y orientar estrategias para mitigar el impacto del COVID-19 en la región.

Según el análisis de datos presentado en el proyecto, los países más adecuados para **la expansión de los laboratorios Biogenesys** serían aquellos que presentan una combinación de alta necesidad de servicios de diagnóstico y tratamiento, junto con condiciones favorables para la implementación de nuevas operaciones. Basándonos en los hallazgos, estos factores podrían guiar la elección:

Brasil

Justificación:

Alta densidad poblacional y grandes áreas urbanas con tasas elevadas de casos confirmados.

La amplia población genera una gran demanda de servicios de salud, especialmente en regiones urbanas donde la propagación del virus ha sido más severa.

Es uno de los países más grandes de la región, lo que abre oportunidades para una expansión a gran escala.

México

Justificación:

Gran población urbana con alta incidencia de comorbilidades (como diabetes y tabaquismo), lo que representa una mayor demanda de diagnóstico y tratamiento.

México es un mercado clave en América Latina por su tamaño y acceso al Tratado entre México, Estados Unidos y Canadá (T-MEC), facilitando el comercio regional.

Argentina

Justificación:

Demuestra picos de casos significativos, especialmente en áreas urbanas, lo que sugiere que aún hay oportunidades para mejorar el acceso a servicios de salud.

Un sistema de salud con base sólida, que podría facilitar la adopción de tecnologías y servicios innovadores de Biogenesys.

Colombia

Justificación:

Es otro país con alta densidad poblacional en áreas urbanas y una necesidad creciente de atención médica avanzada.

Su posición estratégica en América Latina facilita la expansión hacia otros países de la región.

Factores a considerar antes de la expansión:

Demanda de servicios de diagnóstico: Priorizar países con alta prevalencia de enfermedades crónicas y niveles de mortalidad por COVID-19.

Infraestructura de salud existente: Países con sistemas de salud pública y privada desarrollados facilitarán la adopción de nuevos servicios.

Entorno regulatorio: Elegir países con marcos regulatorios favorables para la entrada de nuevas empresas de biotecnología.

Factores económicos y logísticos: Considerar el costo de operación, la disponibilidad de recursos humanos calificados y la facilidad de distribución de insumos médicos.

# 

# Análisis del dashboard

# El dashboard creado con Power BI consta de 5 páginas, las cuales se decriben a continuación:

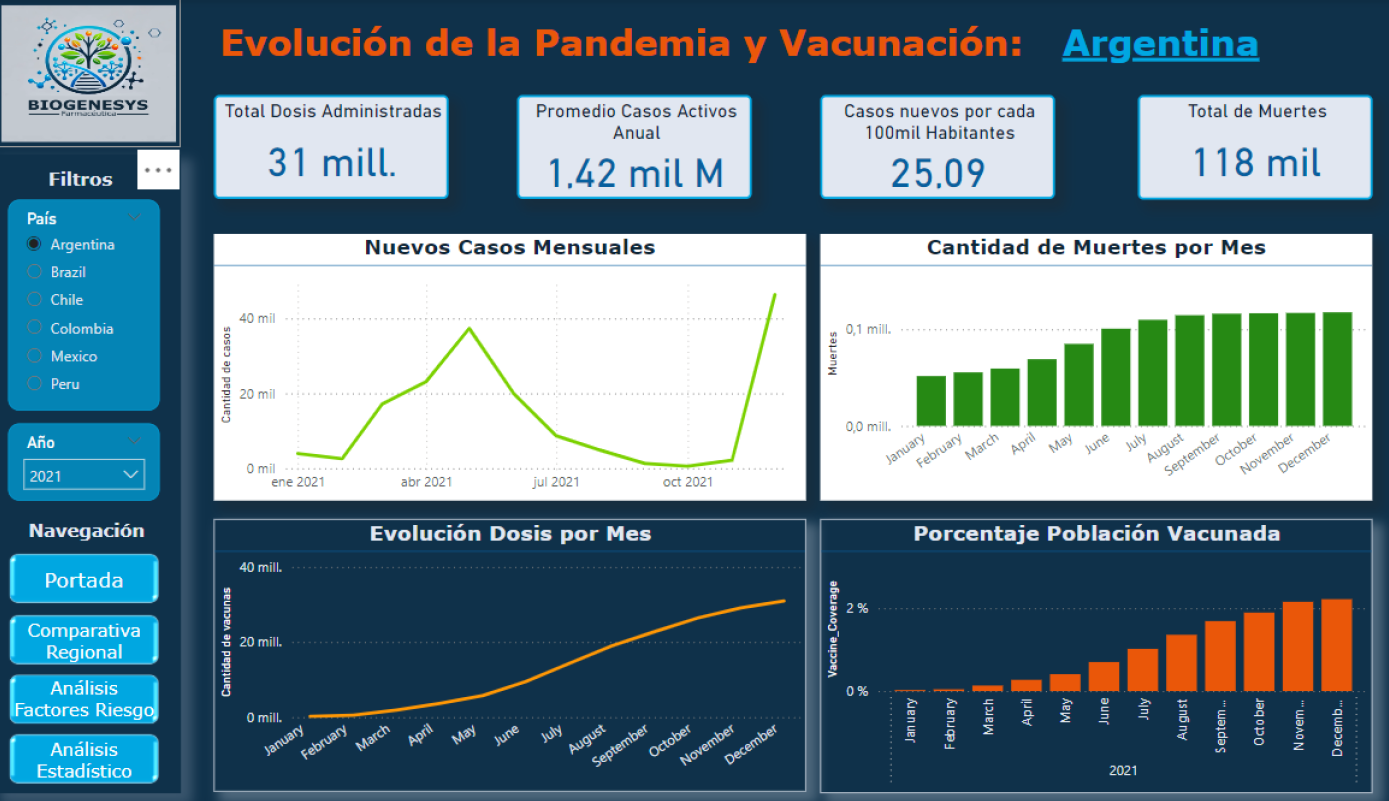
Página 1 - Portada

Contiene como imagen principal el logo creado con inteligencia artificial para la empresa ficticia, además se muestra una breve descripción de los objetivos del análisis de los datos, y finalmente contiene cuatro botones de navegación que redirigen al usuario a cuatro dashboards creados para obtener insights de interés para la empresa. Ver imagen a continuación.



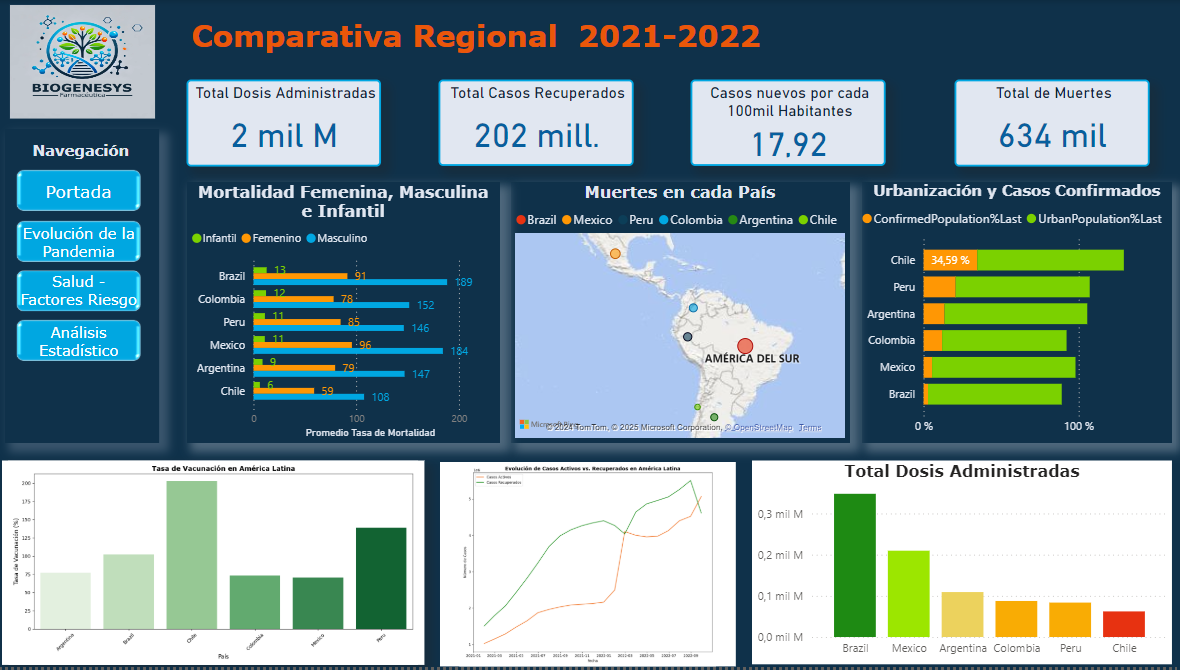
Página 2 - Evolución de la Pandemia

Este dashboard está diseñado para explorar y evaluar variables de interés país por país y por año, de forma que el usuario pueda usar los filtros destinado para ello y analizar campos críticos para la evaluación de la Pandemia país por país, este dashboard permite la seleccion de un país y un año a la vez. En él puede observarse la evolución mensual de los casos nuevos de personas con COVID-19, la cantidad de muertes, la cantidad de dosis administradas por mes y el porcentaje de personas vacunadas. Finalmente, contiene botones de navegación, para que el usuario se desplace con facilidad a las otras páginas. Ver imagen a continuación.



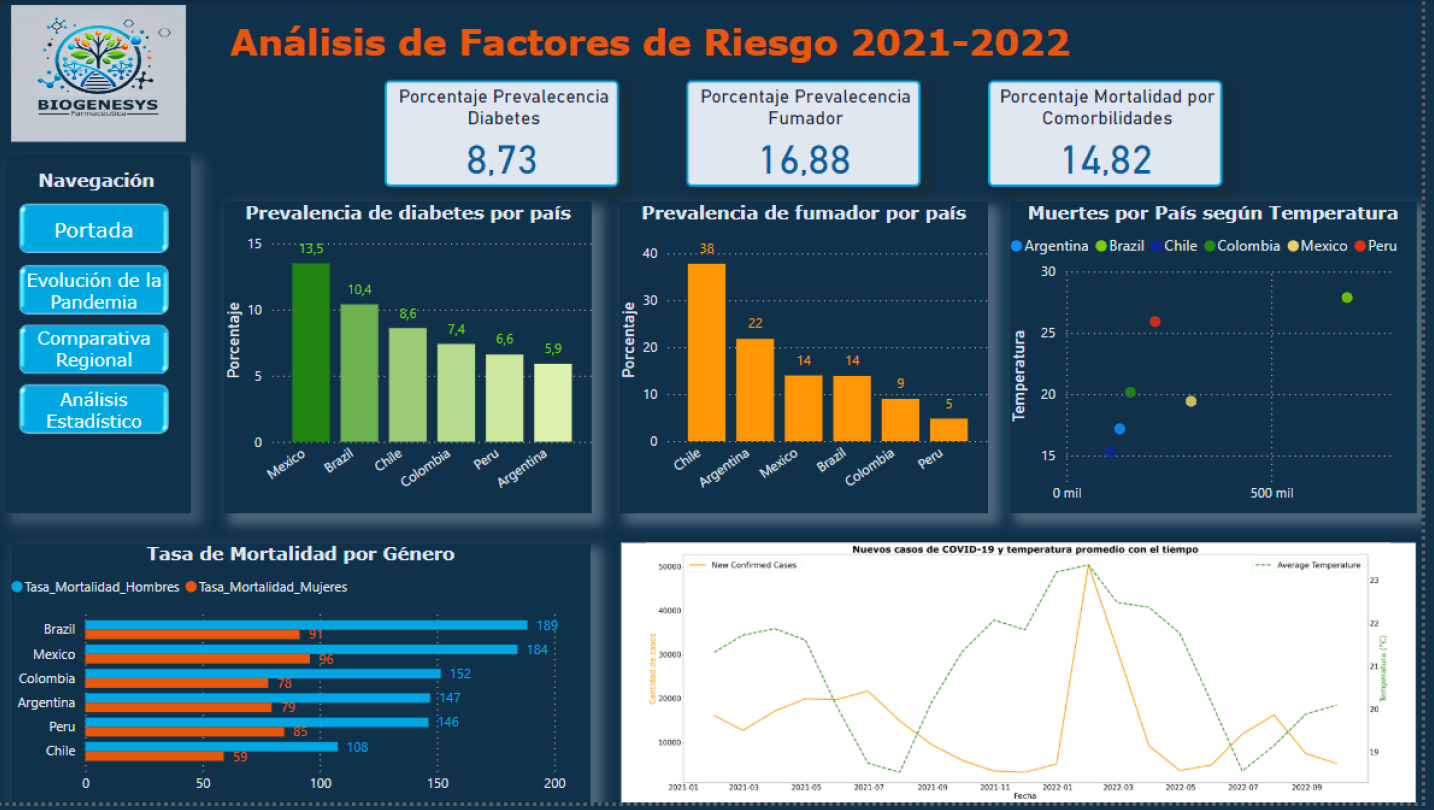
Página 3 - Comprativa Regional

Este dashboard está orientado a evaluar el comportamiento de variables de interés, englobándolas en los años 2021-2022 (unicos años proporcionados en los datos) y para todos los paises en conjunto, por lo cual es una visión a latinoamérica como un todo. Debido a ésto, no contiene filtros o segmentadores, sin embargo, es posible elegir algún país desde los gráficos interactivos y destacarlo en todos los gráficos para su evaluación en particular, sobretodo para ver los detalles en los KPI de la parte superior. Contiene en total cuatro KPIs y cuatro gráficos interactivos de Power BI, y dos gráficos no interactivos creados mediante integración con Python, dichos gráficos se muestran no solo crean para observar la potencia de dicha integración, sino que al ser un dashboard que no necesita ser en realidad interactivo, resalta la facilidad de crear los gráficos mediante scripts de Python que a veces resulta más sencillo de crear que con power BI. Finalmente, contiene botones de navegación, para que el usuario se desplace con facilidad a las otras páginas. Ver imagen a continuación.



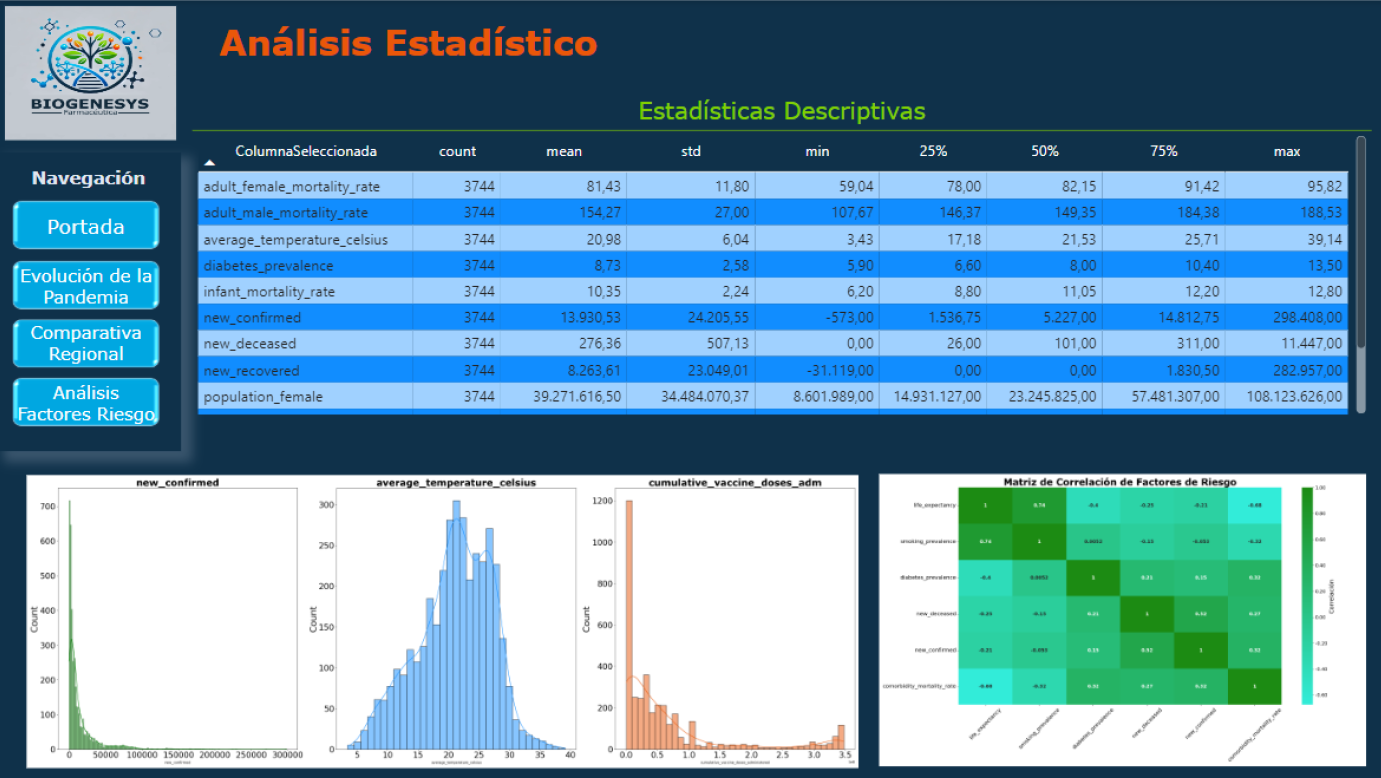
Página 4 - Análisis de Factores de Riesgo

Acá también se muestra una visión de latinoamérica mediante gráficos comparativos de los países, por lo cual tampoco se requirió de unar segmentadores por país ni por años. El objetivo de este dashboard es el análisis de factores de riesgo como enfermedades preexistentes y comorbilidades, la temperatura de los países y el género de las personas afectadas por COVID-19. Contiene tres KPIs y cuatro gráficos interactivos, y un gráfico no interactivos creado mediante un script de Python, todo esto para evaluar cómo inciden estos factores en el desarrollo de la pandemia. Finalmente, contiene botones de navegación, para que el usuario se desplace con facilidad a las otras páginas. Ver imagen a continuación.



Página 5 - Análisis Estadístico

Está orientado a observar medidas estadísticas, las cuales se crearon mediante scripts de Python por su facilidad y porque Power BI no dispone de las gráficas como histogramas, boxplots, matriz de correlación entre otros. También se muestra una tabla con las medidas de tendencia central y de variación para el análisis de las variables cuantitativas, el ser muchas medidas se decidió exportar el dataframe generado en Python mediante la función describe() y acortar el tiempo y esfuerzo en la creación de las mismas. La matriz de correlación sugiere una relación entre enfermedades crónicas como la diabetes y las tasas de mortalidad, lo que confirma la importancia de las comorbilidades en la gravedad de la enfermedad y variables como la edad y el género también parecen estar asociadas con diferentes resultados clínicos. Finalmente, contiene botones de navegación, para que el usuario se desplace con facilidad a las otras páginas. Ver imagen a continuación.



# Conclusiones y Recomendaciones

# El proyecto desarrollado para Biogenesys permitió obtener un análisis profundo sobre la evolución del COVID-19 en países seleccionados de América Latina (Colombia, Argentina, Chile, México, Perú y Brasil). A través de técnicas de análisis exploratorio de datos (EDA) y visualizaciones interactivas en dashboards, se destacaron patrones clave que informan la toma de decisiones estratégicas para la expansión de laboratorios farmacéuticos.

# Resultados Principales

# Correlaciones Críticas: Se identificaron correlaciones significativas entre la prevalencia de enfermedades crónicas (como diabetes y tabaquismo) y las tasas de mortalidad. Esto evidencia que las comorbilidades son determinantes clave en el impacto del COVID-19.

# Impacto de la Vacunación: Se observó una tendencia generalizada de disminución en los casos confirmados y las tasas de mortalidad a medida que aumentaban las tasas de vacunación. Chile lideró en el porcentaje de vacunación, pero también presentó el mayor porcentaje de casos confirmados, probablemente por factores como alta densidad urbana y estrategias efectivas de detección.

# Distribución de Nuevos Casos: En todos los países, los casos confirmados y las muertes tuvieron picos al inicio del período analizado, disminuyendo hacia el final del 2021, excepto en México, que presentó un repunte de casos hacia finales del año.

# Factores Urbanos y Poblacionales: Los países con una alta proporción de población urbana (Chile y Argentina) tuvieron mayores tasas de transmisión. Además, los grupos etarios más afectados variaron, siendo más prevalentes los casos en adultos jóvenes (20-40 años).

# Condiciones Socioeconómicas: Factores como la densidad poblacional, la prevalencia de tabaquismo y las desigualdades en acceso a la salud explicaron las diferencias en los impactos del COVID-19 entre los países analizados.

# Relación entre Clima y Casos: No se observó una correlación significativa entre la temperatura y el número de casos confirmados, lo que sugiere que otros factores tienen un mayor impacto en la transmisión del virus.

El análisis demuestra que las ubicaciones óptimas para la expansión de los laboratorios farmacéuticos Biogenesys deben considerar una combinación de factores sanitarios, sociales y económicos. La vacunación y el manejo de comorbilidades son aspectos cruciales para mitigar la carga del COVID-19. Además, el acceso a infraestructura y población urbana densa ofrecen oportunidades estratégicas para maximizar el impacto de la expansión.

Según el análisis integral, Chile, México y Brasil emergen como los países más adecuados para la expansión inicial de Biogenesys.

Chile: Consolidar presencia debido a su infraestructura avanzada y políticas efectivas.

México: Potencial de mercado alto con necesidades urgentes de manejo de comorbilidades.

Brasil: Expansión estratégica en regiones de alta población y menor acceso a la salud.

Estas decisiones deben estar respaldadas por inversiones en educación sanitaria, alianzas con gobiernos locales y programas de investigación y desarrollo para enfermedades prevalentes.

En general, una combinación de Chile, Brasil y México podría ser ideal para un plan de expansión inicial de los laboratorios de Biogenesys, mientras que Argentina y Colombia pueden considerarse como fases posteriores o expansiones complementarias. Estos países permitirían a Biogenesys generar un impacto significativo en la región mientras optimizan sus recursos.

# Reflexión personal

# Durante el desarrollo del proyecto para Biogenesys, he adquirido valiosas habilidades técnicas y analíticas que fortalecen mi perfil como Analista de Datos. Uno de los aprendizajes más importantes fue la capacidad de transformar grandes volúmenes de datos sin procesar en información accionable mediante técnicas de limpieza, imputación y visualización. A través del manejo de herramientas como Python, Pandas, Matplotlib y Power BI, desarrollé un enfoque integral que permitió explorar relaciones complejas entre factores de salud y demográficos.

# Además, aprendí a identificar correlaciones significativas, interpretar patrones en gráficos y generar insights clave para la toma de decisiones estratégicas. Este proyecto también fortaleció mis habilidades en el diseño de dashboards efectivos, aplicando principios de diseño centrados en el usuario para garantizar que la información crítica sea accesible y comprensible para los stakeholders.

# En cuanto a habilidades blandas, el proyecto me enseñó a priorizar tareas, manejar desafíos como la presencia de datos incompletos y desarrollar una mentalidad orientada a la solución. También reforcé mi capacidad para comunicar hallazgos complejos de manera clara y persuasiva, lo cual es esencial en roles analíticos.

Si tuviera la oportunidad de comenzar este proyecto nuevamente, aprovecharía la experiencia adquirida para optimizar varios aspectos del proceso. Por ejemplo, dedicaría más tiempo a la fase inicial de exploración y documentación de los datos, ya que entender profundamente la estructura del dataset y sus limitaciones desde el principio podría haber acelerado el flujo de trabajo posterior. También implementaría técnicas más avanzadas de imputación, como modelos predictivos o algoritmos de machine learning, para manejar los datos faltantes en lugar de enfoques simples como el promedio o el reemplazo por ceros.

# 