PROYECTO INTEGRADOR

“Expansión Estratégica de Biogenesys con Python”

Carrera: Data Analytics

Cohorte: DA-FT18

Estudiante: Melisa Rossi

Email: melirossi.mr@gmail.com

Fecha de entrega: 27 de Octubre de 2025

ÍNDICE

INSTITUCIÓN

La empresa BIOGENESYS, dedicada a la investigación y desarrollo farmacéutico, se encuentra en un proceso de expansión estratégica en Latinoamérica.

Su meta es determinar las ubicaciones óptimas para nuevos laboratorios farmacéuticos, basándose en el análisis de datos sobre incidencia de COVID-19, vacunación, factores demográficos, económicos y sanitarios.

INTRODUCCIÓN

El presente proyecto de Analítica de Datos se realizó con el propósito de evaluar el impacto de la pandemia de COVID-19 y sus factores correlacionados (demográficos, sanitarios y socioeconómicos) en seis países de Latinoamérica: Argentina, Brasil, Chile, Colombia, México y Perú. El objetivo organizacional principal es identificar las ubicaciones óptimas para la expansión de laboratorios farmacéuticos, minimizando riesgos operativos y maximizando el potencial de mercado.

El análisis se llevó a cabo mediante Python (Pandas, NumPy, Matplotlib y Seaborn) y herramientas de visualización en Power BI, abordando todas las etapas del ciclo analítico: limpieza, transformación, análisis exploratorio, modelado temporal y visualización interactiva.

OBJETIVO

- Comprender las relaciones entre las variables, identificar patrones y contrastar las diferencias estructurales entre los seis países latinoamericanos: Argentina, Brasil, Chile, Colombia, México y Perú.

DESARROLLO DEL PROYECTO

Se inicia el proyecto importando las librerías a utilizar y se procede a la carga del dataset brindado, el archivo csv “data_latinoamerica.csv”. Posteriormente se visualiza la estructura del dataset, para determinar los siguientes procesos.

El dataset original fue filtrado para obtener los seis países de interés y las fechas posteriores al 1 de enero de 2021, período de mayor relevancia para este análisis. Esto se llevó a cabo con funciones como “.isin()” para la variable con más de un valor a filtrar (caso de los países), y de manera más directa para la columna fecha, llamando a dicha columna y aplicando el período solicitado.

Realizado esto, se inicia con el proceso de normalización, donde se comienza con la visualización e identificación de valores nulos. Identificadas las columnas a tratar, se procede con la limpieza de estos datos.

En primer lugar se observaron dos columnas con una considerable cantidad de valores nulos, “new_recovered” y “cumulative_recovered”. Se decide eliminar estas columnas, ya que los nulos superaban, prácticamente en dos tercios, la cantidad de datos aprovechables. Para esto, se aplicó un filtro a todo el dataframe, para eliminar aquellas columnas (mediante la función “.drop()”) con valores nulos mayor al 50%.

Otras columnas, como las referentes a casos y muertes presentaban unos poco valores nulos, por lo que se decide rellenarlos con el valor “0”, ya que se supone que estos valores podrían estar al comienzo de la toma de registros, y reemplazarlos por otros valores no resultaría necesariamente significativo.

Para la columna de dosis de vacunas acumuladas se aplicó el relleno mediante Forward Fill, que utiliza el último valor conocido para rellenar ese dato faltante. Como aún se evidenciaban 103 valores nulos, se procedió a rellenar estos con “0”.

También se aplicó el relleno mediante Forward Fill para las variables climáticas que presentaban algunos datos inexistentes. Sin embargo, como en la columna “rainfall_mm“ aún aparecía un valor nulo, se aplicó un Backward Fill, utilizando el primer valor no nulo que se encuentre.

Chequeamos los nulos mediante “.isnull().sum()” y visualizamos que nuestro dataframe ya no contiene valores faltantes. Ahora bien, se continúa con la transformación de los datos. En primer lugar, se convierte la columna “date” de tipo objeto a tipo fecha, y luego, las columnas categóricas se las transformó de 'object' a tipo 'category' (categóricas). Algunas variables o columnas numéricas serán reemplazadas de datos de tipo 'float64' a 'int64', principalmente aquellas que refieren a personas o población.

Para finalizar, volvemos a visualizar cómo quedó la estructura de nuestro dataframe, para corroborar que todos los pasos se hayan ejecutado exitosamente.

En fin, el proceso de limpieza y transformación de datos no es solo requisito técnico, sino un paso fundamental para garantizar la validez y agilidad de los datos, como también para tener un buen enfoque de las etapas analíticas posteriores.

La aplicación de filtros estratégicos aseguró que el análisis se concentrara en el periodo de mayor relevancia, mientras que la eliminación selectiva de variables con alto volumen de nulos (new_recovered y cumulative_recovered) previene la introducción de sesgos o ruido estadístico.

La imputación diferenciada de nulos y aplicación de técnicas secuenciales (Forward Fill y Backward Fill) para ciertas variables, garantiza la usabilidad y confiabilidad de los datos. Finalmente, la conversión de tipos de datos optimiza el dataframe y prepara las variables para el modelado y visualización.

En resumen, se obtuvo un dataframe limpio, sin valores faltantes y estructuralmente optimizado, listo para proporcionar insights precisos sobre la evolución de la pandemia y sus factores correlacionados en Latinoamérica.

EDA E INSIGHTS

Inicialmente se realizó un análisis estadístico descriptivo, mediante bucle for, para obtener la media, desviación estándar, valor máximo y mínimo de distintos grupos de variables. Si bien esto puede realizar mediante la función ".describe()", de esta manera se pudieron conseguir las medidas especificadas por cada país, y por grupo de variables relacionadas.

El análisis descriptivo de los datos permitió caracterizar las condiciones demográficas, sanitarias, económicas y epidemiológicas de los seis países latinoamericanos seleccionados: Argentina, Brasil, Chile, Colombia, México y Perú.

Esta fase buscó establecer una base sólida para comprender las diferencias estructurales entre las naciones, evaluar su capacidad de respuesta sanitaria y detectar patrones que orienten las decisiones estratégicas de BIOGENESYS respecto a su expansión regional.

Se encontró que Brasil domina las cifras absolutas de contagios, muertes y vacunación, lo que se corresponde con su gran tamaño poblacional. Argentina muestra altos niveles de variabilidad, pero con una mortalidad controlada. México combina una alta incidencia con una mortalidad elevada. Chile destaca por su eficiencia en vacunación y estabilidad sanitaria, mientras, Perú y Colombia presentan patrones intermedios, con una respuesta sostenida y menor volatilidad en los indicadores.

Para profundizar el estudio de los indicadores, se realizó un análisis exploratorio acompañado de diversas visualizaciones. Se implementaron gráficos de barras, histogramas, mapas de calor, diagramas de dispersión, gráficos de líneas y boxplots, entre otros, con el fin de interpretar la dinámica temporal y los factores subyacentes que incidieron en la evolución de la pandemia y la vacunación en la región.

De este análisis encontramos que:

Gráficos de barra: comparativa entre países

Los gráficos de barra permitieron realizar comparaciones directas entre los países en distintos indicadores clave:

- **Nuevos casos y muertes:**
  Brasil concentra la mayor cantidad de nuevos contagios (298408) y fallecimientos (4200), seguido por México y Argentina.
  En contrapartida, Chile y Perú exhiben cifras considerablemente menores, lo que evidencia diferencias tanto en magnitud poblacional como en la capacidad de respuesta sanitaria.

- **Vacunación**:
  Brasil y México lideran en dosis de vacunas administradas, seguidos por Colombia.
  Los países con menores volúmenes absolutos de vacunación fueron Chile, Perú y Argentina, aunque proporcionalmente su cobertura fue alta en relación con la población total.

- **Distribución poblacional:**
  La jerarquía demográfica se mantiene constante: Brasil supera los 213 millones de habitantes, seguido de México (111 millones), Colombia (51 millones) y Argentina (45 millones).
  Perú y Chile completan el grupo con 29 y 18 millones respectivamente.
  Estas diferencias poblacionales explican gran parte de la variación observada en los casos, muertes y vacunación totales.

- **Estructura urbana y rural:**
  Brasil y México concentran tanto la mayor población rural como urbana, reflejando una dualidad demográfica que influye en la propagación del virus.
  Chile y Argentina destacan por una mayor urbanización y, en consecuencia, una exposición más controlada.

- **Densidad poblacional:**
  México presenta la densidad más alta, seguido por Colombia, mientras que Argentina es el país con menor densidad, lo que puede explicar una menor propagación comunitaria.

- **Desarrollo económico:**
  El PIB per cápita sitúa a Chile como el país más próspero, seguido por Argentina y México, mientras que Colombia y Perú muestran los valores más bajos.
  Esto sugiere que la capacidad de inversión en salud pública y la resiliencia económica fueron factores determinantes en la respuesta a la crisis sanitaria.

- **Condiciones sanitarias y hábitos de riesgo:**
  En prevalencia de tabaquismo, Chile (37,8%) y Argentina (21,8%) presentan los valores más altos.
  En prevalencia de diabetes, México lidera (13,5%), seguido por Brasil (10,4%), lo que los posiciona como países de mayor riesgo ante infecciones respiratorias graves.

- **Mortalidad infantil y expectativa de vida:**
  Chile muestra la menor tasa de mortalidad infantil y la mayor expectativa de vida, mientras que Brasil y Colombia exhiben tasas más elevadas.
  Esto sugiere que los países con mejor acceso sanitario presentan una mayor longevidad.

- **Condiciones ambientales:**
  Brasil y Perú son los países más cálidos (28 °C y 26 °C respectivamente), mientras que Argentina y Chile presentan temperaturas más bajas y mayor variabilidad térmica.
  En cuanto a precipitaciones, Perú y Colombia registran los valores más altos, y la humedad relativa es mayor en zonas tropicales.

Histogramas y distribuciones

Los histogramas de nuevos casos confirmados y dosis de vacunación acumuladas evidencian patrones comunes de comportamiento:

- En los nuevos casos, la mayoría de los registros se concentran en valores bajos, con colas derechas largas que reflejan picos epidémicos esporádicos y severos.

- En las vacunas acumuladas, se observa un patrón escalonado de crecimiento: un aumento abrupto al inicio de las campañas y una estabilización progresiva posterior. Los picos y mesetas corresponden a diferentes fases de la vacunación o ritmos desiguales entre países.

Mapas de calor (Heatmaps)

El análisis de correlación entre variables reveló vínculos esperables y otros de interés particular:

- Las variables demográficas y de conteo acumulado presentan las correlaciones más fuertes (≥ 0,8), como son el caso de la población total con los casos y muertes acumuladas.

- Existe una correlación positiva fuerte (0,77) entre la población urbana y los casos confirmados, más alta que la rural (0,55), lo cual es coherente con la mayor exposición en entornos densamente poblados.

- Si observamos las variables del clima, notamos que los nuevos confirmados y nuevas muertes tienen una correlación débil en relación a la temperatura. La temperatura promedio no parece ser un predictor fuerte de la cantidad de nuevos casos en la región.

- Se observó una correlación negativa entre las mejores condiciones de salud presentan una longevidad superior.eranza de vida y las tasas de mortalidad infantil o por comorbilidades, lo que confirma que los países con mejores condiciones de salud presentan una longevidad superior.

- Las tasas de mortalidad por contaminación y por comorbilidades también se asocian positivamente (aprox. 0,42), reflejando un efecto conjunto de vulnerabilidad ambiental y sanitaria.

Diagramas de dispersión

Los scatterplots permitieron explorar relaciones entre variables continuas:

- **Casos confirmados vs. Temperatura media**:
  No se detectó una relación clara. Más bien, la temperatura define el rango en el que se mueve la pandemia en cada país. Los casos tienden a concentrarse en temperaturas moderadas (20–30 °C), sin evidencia de que el calor reduzca o aumente los contagios.
  Los picos más altos corresponden a Brasil y Argentina, con brotes de hasta 300000 casos diarios.

- **Muertes vs. Temperatura media**:
  Siguiendo un patrón similar, no se observó una correlación fuerte entre temperatura y mortalidad.
  Los valores más altos se concentraron en Brasil (hasta 4200 muertes diarias) en un rango de temperatura de 25°C a 33°C. México no presenta picos significativos, pero sí se evidencia una concentración en un rango de entre 10°C y 20°C.

- **Diabetes vs. Mortalidad promedio:**
  Se identificó una tendencia positiva clara: los países con mayor prevalencia de diabetes (México y Brasil) también presentan las tasas de mortalidad más elevadas.
  Chile destaca como un caso atípico positivo, con buena gestión sanitaria y baja mortalidad a pesar de una prevalencia moderada.

Gráficos de líneas: análisis temporal

Los gráficos de evolución temporal permitieron observar la dinámica del COVID-19 y sus efectos:

- **Casos confirmados por mes**:
  Todos los países comparten un patrón estacional con dos grandes picos epidémicos: uno en mediados de 2021 y otro, más pronunciado, a inicios de 2022.
  Brasil y Argentina fueron los más afectados, con más de 3 millones de casos mensuales en los picos más altos.

- **Muertes por mes**:
  El comportamiento fue similar, aunque las muertes muestran una tendencia decreciente sostenida desde mediados de 2021.
  El descenso refleja la efectividad del proceso de vacunación y la mejora en la capacidad de respuesta sanitaria.
  En 2022, pese a un incremento en los contagios, las muertes no aumentaron proporcionalmente, lo que evidencia el efecto protector de la inmunización.

- **Dosis de vacunación acumuladas**:
  Brasil encabeza con la curva más pronunciada, seguido por México.
  En un segundo grupo, Argentina, Chile, Perú y Colombia presentan curvas similares, con una vacunación más gradual y estable.

Boxplots y distribuciones climáticas

Los boxplots de temperatura media por país muestran diferencias notables:

- Brasil y Perú: temperaturas altas y estables, con poca variación interna.
- México y Colombia: temperaturas moderadas y consistentes.
- Argentina y Chile: mayor variabilidad térmica, coherente con su geografía extensa y diversidad climática.

Los outliers detectados podrían representar registros extremos o errores de medición, por lo que conviene analizarlos antes de modelizar tendencias climáticas.

Distribución por Grupos Etarios y Mortalidad por Género

El análisis de la población por grupos etarios confirma el predominio demográfico de Brasil en todos los segmentos, seguido por México.

México y Argentina poseen una base poblacional joven (0–19 años) particularmente amplia, lo que podría representar tanto una oportunidad de crecimiento económico como un desafío sanitario en contextos de alta movilidad social.

Por su parte, Brasil concentra su mayor volumen en el rango de 20–39 años, correspondiente a la fuerza laboral principal, lo que influye directamente en su dinamismo económico y en la propagación de enfermedades contagiosas.

En contraste, Chile presenta una estructura más equilibrada en la vejez, con mayor proporción de adultos mayores, coherente con su alta expectativa de vida.

En cuanto a la mortalidad por género, se observa que la tasa masculina duplica aproximadamente a la femenina, patrón consistente con lo observado en la mayoría de las regiones del mundo durante la pandemia.

En síntesis, el conjunto, las visualizaciones confirman que:

- Las tendencias epidemiológicas y temporales fueron similares entre los países, aunque con magnitudes distintas.

- Los factores estructurales (densidad, urbanización, comorbilidades) incidieron más que el clima en la propagación del virus.

- La respuesta sanitaria —vacunación masiva y políticas de control— explica la reducción sostenida de la mortalidad en 2022.

- Los países con mejor infraestructura sanitaria y mayor desarrollo económico (Chile, Argentina, Brasil) mostraron los procesos de estabilización más tempranos.

En definitiva, el análisis exploratorio visual permitió contextualizar las diferencias entre los países, identificar patrones comunes y validar la coherencia interna de los datos, generando insumos fundamentales para las decisiones estratégicas de expansión de BIOGENESYS en América Latina.

ANÁLISIS DEL DASHBOARD

El dashboard desarrollado en Power BI permite explorar de manera dinámica los resultados del análisis realizado sobre los indicadores sanitarios, demográficos y económicos de los seis países latinoamericanos seleccionados: Argentina, Brasil, Chile, Colombia, México y Perú.

Desde la portada principal, el usuario puede abrir directamente el informe presionando sobre el logo. Además, puede seleccionar qué tipo de visualización desea consultar mediante botones interactivos, facilitando una navegación intuitiva y ordenada. Cada sección del dashboard aborda una dimensión específica del análisis —como la evolución de casos y muertes, el avance de la vacunación, o los factores socioeconómicos asociados— y puede recorrerse libremente según el interés del usuario.

Además, en todas las páginas se incluyó un ícono del logo en la esquina superior izquierda, que permite regresar fácilmente al panel principal, optimizando la experiencia de uso.

Varios de los gráficos fueron generados mediante scripts de Python integrados en Power BI, lo que permitió crear visualizaciones más personalizadas y analíticamente robustas. Entre ellas se destacan los gráficos de líneas para el seguimiento temporal de contagios y vacunación, los diagramas de dispersión que relacionan variables de salud y condiciones ambientales, y los heatmaps de correlación entre indicadores demográficos y epidemiológicos.

El dashboard ofrece así una visión integral y jerarquizada del comportamiento de la pandemia y su contexto regional, combinando datos sanitarios, climáticos y poblacionales. De esta forma, permite identificar rápidamente patrones, comparar países, detectar tendencias y respaldar la toma de decisiones estratégicas en torno a la expansión de BIOGENESYS en Latinoamérica.

CONCLUSIONES

El análisis integral de los datos epidemiológicos, demográficos y económicos permitió comprender con profundidad la evolución del COVID-19 y su impacto en América Latina, así como las condiciones estructurales de cada país que pueden influir en futuras estrategias sanitarias y de inversión.

Los resultados evidencian que Brasil es el país con mayor magnitud en todos los indicadores absolutos —población, casos, muertes y vacunación—, lo cual es coherente con su tamaño demográfico. México y Argentina le siguen, presentando comportamientos epidemiológicos similares, aunque con diferencias notables en mortalidad y ritmo de vacunación.

Chile se destaca por su mayor expectativa de vida, mejor control sanitario y elevado PIB per cápita, lo que lo posiciona como un referente regional en eficiencia de gestión de salud pública.

Perú y Colombia, por su parte, muestran contextos más vulnerables desde lo económico, aunque con respuestas sanitarias estables y adaptativas.

El estudio temporal revela una sincronía epidemiológica regional, con dos grandes picos de contagios (mediados de 2021 y principios de 2022), seguidos por una disminución sostenida en la mortalidad.

Este patrón refleja el efecto positivo de la vacunación masiva y la mejora en la capacidad de respuesta de los sistemas de salud.

Asimismo, las correlaciones entre variables confirman que la densidad urbana, las comorbilidades (diabetes, tabaquismo) y el desarrollo económico fueron factores determinantes en la propagación y el control del virus, mientras que las variables climáticas mostraron una influencia marginal.

En conjunto, los hallazgos ofrecen una visión sólida del contexto post-pandemia y permiten identificar a Chile, Argentina y Brasil como los países más propicios para el fortalecimiento o expansión de la infraestructura farmacéutica, por su equilibrio entre población, capacidad sanitaria y potencial económico.

RECOMENDACIONES ESTRATÉGICAS

- Priorización geográfica:
  Dirigir la expansión inicial de BIOGENESYS hacia Brasil, Chile y Argentina, donde el volumen de mercado y la estabilidad sanitaria garantizan una base sólida para operaciones de laboratorio y distribución.

- Segmentación sanitaria y logística:
  Aprovechar la concentración urbana para optimizar la cadena de suministro y distribución de productos farmacéuticos, focalizando en regiones metropolitanas con alta densidad poblacional.

- Fortalecimiento de infraestructura en países emergentes:
  En Perú y Colombia, impulsar acuerdos institucionales orientados al refuerzo de la infraestructura sanitaria, contribuyendo al desarrollo local y ampliando la presencia regional de la empresa.

- Análisis de comorbilidades:
  Diseñar productos y programas enfocados en diabetes, enfermedades respiratorias y cardiovasculares, que presentan prevalencias significativas en la región.

- Continuidad del monitoreo epidemiológico:
  Mantener un sistema de vigilancia de datos en tiempo real que integre indicadores de salud, demografía y clima, permitiendo anticipar brotes y ajustar estrategias de mercado.

- Integración de indicadores sociales y ambientales:
  Considerar factores como contaminación ambiental, hábitos de consumo y acceso a servicios médicos para futuras fases del análisis predictivo.

En síntesis, los resultados obtenidos permiten delinear una estrategia de expansión sólida, basada en evidencia y sustentada en indicadores sanitarios, demográficos y económicos.

El análisis respalda que Brasil, Chile y Argentina representan las mejores oportunidades inmediatas para el crecimiento de BIOGENESYS, mientras que Perú y Colombia ofrecen un potencial significativo a mediano plazo, especialmente mediante alianzas estratégicas que fortalezcan la infraestructura de salud.

El estudio confirma que la toma de decisiones informadas —sustentadas en datos limpios, comparables y actualizados— puede marcar la diferencia entre una expansión exitosa y una operación de riesgo.

BIOGENESYS, al adoptar una visión analítica e integrada de la región, no solo podrá optimizar su rendimiento comercial, sino también contribuir al fortalecimiento de los sistemas sanitarios latinoamericanos, reforzando su compromiso con la salud pública y la innovación farmacéutica.

REFLEXIÓN PERSONAL

Este proyecto representó una experiencia enriquecedora tanto en lo técnico como en lo analítico. Aplicar herramientas de Python, Pandas, Numpy y visualización con Matplotlib y Seaborn permitió transformar un conjunto de datos masivo en información estructurada, visual y comprensible.

Más allá de los resultados estadísticos, el desafío estuvo en interpretar los datos desde una perspectiva estratégica, conectando los hallazgos con la realidad sanitaria, económica y social de Latinoamérica.

A lo largo del proceso, aprendí la importancia de la limpieza y normalización de datos, el uso de visualizaciones para descubrir patrones, y la relevancia del contexto al momento de realizar interpretaciones. Cada gráfico o métrica dejó de ser un resultado aislado para convertirse en un argumento que respalda una decisión.

Asimismo, este proyecto consolidó la comprensión de que el rol del analista de datos no se limita al procesamiento técnico, sino que es un puente entre los datos y la acción, ayudando a que las empresas tomen decisiones más informadas, eficientes y humanas.

Finalizo el trabajo con una visión más integral de cómo la ciencia de datos puede generar impacto real, especialmente cuando se orienta al bienestar social y al fortalecimiento de los sistemas de salud.