**Expansión Estratégica de BIOGENESYS en Latinoamérica: Análisis de Datos para Optimizar la Ubicación de Laboratorios Farmacéuticos**

**Nombre del autor: Isaac Manuel Pérez Cuadrado**

**Email: imperez@correo.unicordoba.edu.co**

**Cohorte: DA-ft 14**

**Fecha de entrega:** 09-06-2025

**Institución:**



# Introducción

La empresa farmacéutica **Biogenesys**, especializada en soluciones sanitarias innovadoras, ha decidido emprender una expansión estratégica hacia América Latina, con el objetivo de fortalecer su red de laboratorios y centros de vacunación. Ante los desafíos provocados por la pandemia de COVID-19 y sus efectos post-pandemia, la compañía busca establecerse en ubicaciones clave que le permitan actuar con rapidez, eficacia y precisión ante futuras crisis sanitarias.

Para apoyar esta decisión, se ha diseñado un proyecto de análisis de datos orientado a **identificar regiones óptimas en seis países latinoamericanos**: Argentina, Brasil, Chile, Colombia, México y Perú. El estudio se basa en el procesamiento y análisis de más de 12 millones de registros sobre la incidencia de COVID-19, cobertura de vacunación y variables sanitarias y demográficas.

Utilizando herramientas como **Python (Pandas, NumPy, Matplotlib, Seaborn)** y **Power BI**, el proyecto se divide en cuatro fases: carga y limpieza de datos, análisis exploratorio y visualización, análisis temporal y de correlaciones, e integración de dashboards interactivos. Cada fase tiene como objetivo extraer **insights concretos** que ayuden a la toma de decisiones estratégicas para la inversión en infraestructura sanitaria.

Con este proyecto, Biogenesys espera contar con un panorama claro de las **zonas de mayor necesidad y oportunidad**, optimizando así el impacto positivo en la salud pública de la región.

# Desarrollo del proyecto

### ****Metodología de Recopilación y Selección de Datos****

Para el presente análisis, se partió de un dataset original de aproximadamente **21 GB**, compuesto por más de **22 millones de registros y 707 columnas**. Este archivo contenía información sanitaria y demográfica de múltiples países del mundo. Dada su magnitud, el equipo de ingeniería de datos de Biogenesys realizó una selección y transformación inicial, generando un archivo más manejable, data\_latinoamerica.csv, con **12.216.057 registros y 50 columnas**, centrado exclusivamente en países de América Latina.

Adicionalmente, se entregó un documento complementario, readme.txt, con el detalle y la descripción de cada una de las columnas incluidas. Este archivo resultó clave para comprender la estructura del dataset y planificar el análisis.

Los países seleccionados para el estudio de expansión de Biogenesys fueron:

* **Argentina**
* **Brasil**
* **Chile**
* **Colombia**
* **México**
* **Perú**

Se filtraron los registros correspondientes a estos países, considerando solo las fechas **posteriores al 1 de enero de 2021**, periodo más representativo en términos de vacunación y control de la pandemia.

### ****Transformación y Limpieza de Datos****

Una vez cargado el archivo en Python, se procedió con un exhaustivo proceso de transformación y limpieza de datos. Las principales etapas fueron:

1. **Verificación de estructura**, asegurando que el dataset contenía la cantidad esperada de filas y columnas.
2. **Filtrado por países y fechas** de interés para el análisis.
3. **Tratamiento de valores nulos**, utilizando técnicas como:
   * Imputación por valores promedio o anteriores (forward fill).
   * Eliminación de registros con alta proporción de datos faltantes si no eran esenciales.
4. **Conversión de tipos de datos** (fechas, numéricos y categóricos).
5. **Cálculo de nuevas variables**, como tasas, proporciones, acumulados mensuales, entre otros.
6. **Estadísticas exploratorias** con Pandas y funciones personalizadas.

No obstante, durante este proceso se identificaron **errores importantes cometidos por el equipo encargado de la integración de la base de datos**. Particularmente, se detectaron **inconsistencias graves en las variables acumulativas**, como cumulative\_recovered, lo que **afectó la calidad y precisión de varios cálculos críticos**.

Además, se encontró que las columnas new\_recovered y cumulative\_recovered contenían **una cantidad significativa de datos nulos**, lo que **las volvió prácticamente inutilizables** para el análisis directo. Ante esta situación, se decidió aplicar una estimación de new\_recovered utilizando la fórmula:

new\_recovered = new\_confirmed - new\_deceased

Si bien esta estimación **no fue del todo precisa**, permitió al menos obtener una **aproximación útil** para avanzar en los análisis, particularmente en los cálculos de recuperación y evolución de la pandemia en los países evaluados.

Durante la limpieza y transformación de datos, una variable que también atrajo mucho mi atención fue cumulative\_vaccine\_doses\_administered, que registra el total de dosis de vacunas aplicadas hasta cada fecha. Para obtener una mejor perspectiva diaria, se extrajo esta información en una tabla separada y se creó la variable vaccine\_doses\_daily, que representa las dosis diarias aplicadas. Esta nueva variable se calculó restando el valor acumulado de cada día con el del día anterior.

Sin embargo, al igual que con otras variables acumulativas, se detectaron inconsistencias y valores erráticos en los datos diarios resultantes. Se aplicaron correcciones y filtros para mitigar estos errores, logrando así un conjunto de datos más confiable que permitió realizar análisis y visualizaciones aceptables para la toma de decisiones.

### ****Conclusión sobre la Limpieza y Transformación****

El proceso de limpieza y transformación de los datos fue clave para depurar la información y **minimizar el impacto de los errores heredados en la base de datos original**. La identificación temprana de inconsistencias, especialmente en las variables acumulativas, y la decisión de aplicar estimaciones alternativas, permitieron mantener la integridad general del análisis.

A pesar de las limitaciones y supuestos necesarios en algunas columnas, el dataset final (DatosFinalesFiltrado.csv) logró consolidar **información confiable, estandarizada y contextualizada**, lista para su uso en análisis exploratorios, visualización avanzada y reportes ejecutivos en Power BI.

Este trabajo de depuración permitió asegurar que las decisiones estratégicas de expansión propuestas por Biogenesys se basaran en datos con un nivel aceptable de precisión, maximizando así el impacto positivo en los sistemas de salud de los países seleccionados.

# EDA e insights

El análisis exploratorio de datos (EDA) realizado sobre el dataset limpio y transformado permitió identificar patrones críticos, relaciones entre variables, y diferencias sustanciales entre los seis países latinoamericanos analizados: **Argentina, Brasil, Chile, Colombia, México y Perú**. Este proceso fue fundamental para comprender el contexto de la pandemia y postpandemia, evaluar necesidades sanitarias, y guiar decisiones estratégicas para la expansión de laboratorios y centros de vacunación de Biogenesys.

### ****1. Tendencias Generales del COVID-19****

* **Picos de casos confirmados** se observaron en todos los países entre mediados de 2021 y principios de 2022, evidenciando la ola más agresiva vinculada a las variantes Delta y Ómicron.
* **Brasil y México** registraron el mayor número de casos acumulados, seguidos por Argentina. Estos países también mostraron mayor dispersión de datos (alta varianza), lo que indica una evolución de la pandemia menos uniforme en sus territorios.
* **Chile** presentó un control temprano de los casos, posiblemente vinculado a una política de vacunación más agresiva en etapas iniciales.

### ****2. Vacunación: cobertura y velocidad****

* **Chile lideró** la cobertura de vacunación en términos proporcionales a su población, seguido por **Argentina**.
* En contraste, **Colombia y Perú** mostraron una progresión más lenta, especialmente en los primeros seis meses de campaña.
* La **correlación positiva** entre cobertura de vacunación y disminución de casos activos comenzó a ser visible a partir del segundo semestre de 2021.

### ****3. Infraestructura y variables demográficas****

* Al analizar la relación entre **urbanización** y propagación del virus, se observó que las zonas más densamente pobladas (como Ciudad de México, São Paulo y Lima) concentraron mayores niveles de contagio.
* Se identificó una **fuerte correlación entre mortalidad y comorbilidades** como la diabetes y enfermedades respiratorias, siendo Perú y México los países más impactados por estas condiciones.
* La **distribución etaria** reveló que países con poblaciones más envejecidas (como Chile y Argentina) tuvieron tasas de hospitalización más altas, mientras que las muertes en edades medias fueron más frecuentes en Brasil y México.

### ****4. Insights derivados de visualizaciones****

* **Mapas de calor de correlación** revelaron una asociación significativa entre el descenso de temperatura media y el aumento de casos en regiones como el sur de Chile y Argentina.
* Los **diagramas de dispersión** entre temperatura y casos confirmados reforzaron la estacionalidad del virus, sugiriendo que los meses fríos coincidieron con repuntes significativos de contagio.
* El **boxplot de temperatura media por país** mostró que Brasil mantiene temperaturas más estables durante todo el año, lo que podría explicar una curva de contagios menos marcada estacionalmente.
* Los **gráficos de líneas por país** evidenciaron que las **muertes por mes** tuvieron comportamientos más variables que los casos confirmados, lo que puede estar relacionado con deficiencias hospitalarias temporales o saturación del sistema de salud.

### ****5. Anomalías y oportunidades****

* Se identificaron **anomalías en la carga de datos acumulativos**, especialmente en variables como cumulative\_recovered, que interrumpían la correcta interpretación de tendencias. Este hallazgo reforzó la necesidad de usar métricas derivadas o estimadas para análisis más confiables.
* A pesar de las inconsistencias, la estimación de new\_recovered permitió detectar **mejoras progresivas en la recuperación** de pacientes, con un crecimiento estable hacia finales de 2022.

### ****Conclusión del EDA****

El EDA permitió extraer insights críticos para Biogenesys, ayudando a entender la evolución del COVID-19 en Latinoamérica desde una perspectiva sanitaria, climática y demográfica. La combinación de análisis estadístico y visualizaciones detalladas reveló qué países presentan mayor urgencia de inversión, cuáles tienen mejor estructura sanitaria, y dónde las intervenciones pueden generar un impacto más efectivo.

Estos hallazgos constituyen una base sólida para el diseño de dashboards interactivos en Power BI, permitiendo a los directivos tomar decisiones estratégicas respaldadas por datos.

# Análisis del dashboard

El **dashboard interactivo desarrollado en Power BI** consta de **cuatro páginas o secciones principales**, pensadas para ofrecer una visión integral del impacto del COVID-19 en América Latina durante los años 2021 y 2022. El diseño está orientado a facilitar la navegación, interpretación y extracción de conclusiones para la toma de decisiones estratégicas, especialmente en lo relativo a salud pública y planificación de recursos.

### ****Página 1 – Portada****

Esta hoja cumple la función de introducción visual al informe. Contiene:

* **El logo institucional o corporativo**, según corresponda.
* El título del trabajo:  
  **"INFORME DE SITUACIÓN DE LA SARS-CoV-2 (COVID-19) EN AMÉRICA LATINA 2021–2022"**.
* **Botones interactivos** que permiten navegar directamente a cada una de las páginas del informe, facilitando una experiencia de usuario ordenada y amigable.

Esta página no contiene visualizaciones de datos, sino que actúa como menú de acceso a las distintas secciones temáticas del dashboard.

### ****Página 2 – Informe General de Contagios****

Esta sección analiza el comportamiento general de la pandemia en seis países latinoamericanos:

* **Indicadores clave (KPIs)**:
  + Casos confirmados totales: **52.03 millones**
  + Muertes totales: **1.04 millones**
  + Tasa de mortalidad por cada 100.000 habitantes: **222.03**
* **Mapa geográfico con burbujas proporcionales**:  
  Visualiza la magnitud de los contagios por país, permitiendo identificar de manera inmediata los focos de mayor afectación.
* **Gráfico de torta sin centro**:  
  Representa la proporción de casos confirmados por país, en relación con el total regional.
* **Gráfico de líneas (evolución mensual)**:  
  Muestra cómo evolucionaron los casos y muertes mes a mes durante 2021 y 2022, resaltando los picos epidémicos más relevantes.
* **Tabla comparativa de indicadores por año**:  
  Ofrece una vista detallada por país y año, útil para análisis cuantitativo y comparativo.

**Conclusión de esta página**:  
Se evidencia que países como Brasil y México concentraron la mayor carga de casos y muertes. Las curvas mensuales reflejan los efectos de las olas epidémicas y muestran cómo, a lo largo del tiempo, las tasas de mortalidad comenzaron a disminuir progresivamente, en parte gracias a la implementación de campañas de vacunación.

### ****Página 3 – Perfil Demográfico y Factores de Riesgo****

Esta sección integra variables demográficas y sociales para comprender mejor los factores estructurales que condicionaron el impacto del COVID-19:

* **Distribución poblacional por edad (rangos de 10 años hasta >80)**:  
  Permite detectar qué países tienen mayor proporción de población vulnerable (adultos mayores).
* **Gráfico de torta sin centro – Zona geográfica (rural vs urbana)**:  
  Indica cómo se distribuye la población, dato clave para entender desigualdades en el acceso a servicios de salud.
* **Gráfico de factores de riesgo vs tasa de mortalidad**:  
  Muestra la relación entre **tabaquismo, diabetes, tasa de comorbilidad** y las muertes por COVID, observándose correlaciones significativas.
* **Gráfico de infraestructura sanitaria (médicos/enfermeras por cada 1.000 hab. vs mortalidad infantil)**:  
  Sirve como aproximación a la capacidad de respuesta sanitaria de cada país.
* **Indicadores agregados**:
  + Total población analizada: **466.11 millones**
  + Tasa de contagio por cada 100.000 personas: **11.16k**

**Conclusión de esta página**:  
Los países con mayor proporción de población envejecida, más comorbilidades y menor cobertura sanitaria presentan mayores tasas de mortalidad. También se destacan las desigualdades estructurales entre zonas rurales y urbanas.

### ****Página 4 – Vacunación****

La última sección del dashboard se centra en el análisis de los datos de vacunación y su impacto:

* **Total de vacunas aplicadas**: **900.13 millones**
* **Gráfico de torta por país**:  
  Muestra la distribución proporcional de vacunas aplicadas, siendo Brasil el principal actor por su población.
* **Gráfico de correlación (vacunación vs muertes)**:  
  Permite identificar cómo la mayor cobertura de vacunación se relaciona con una menor tasa de mortalidad.
* **Evolución diaria de casos confirmados por país**:  
  Gráfico detallado que permite identificar patrones nacionales específicos.
* **Promedio acumulado de vacunación total**:  
  Visualización que muestra el progreso combinado del proceso de inmunización en los países analizados.

**Conclusión de esta página**:  
La vacunación tuvo un efecto visible en la reducción de muertes, aunque el ritmo y cobertura variaron entre países. Esto pone en evidencia la necesidad de reforzar políticas públicas de inmunización y logística sanitaria en zonas rezagadas.

### 🧭 Conclusión general

Este dashboard ofrece una experiencia **narrativa, interactiva y visualmente accesible** para explorar la situación del COVID-19 en América Latina. Los insights extraídos permiten comprender no solo la magnitud de la pandemia, sino también los factores estructurales que la agravan. Además, proporciona bases sólidas para tomar decisiones estratégicas, como priorizar la expansión de laboratorios o reforzar el sistema sanitario en zonas de mayor vulnerabilidad.

# Conclusiones y Recomendaciones

A lo largo del desarrollo de este proyecto se abordó de manera integral la situación del COVID-19 en seis países de América Latina durante los años 2021 y 2022, mediante análisis exploratorio de datos, transformaciones y limpieza rigurosa de la base, y finalmente la visualización de resultados en un dashboard interactivo en Power BI.

Los resultados obtenidos permiten no solo entender la magnitud del impacto de la pandemia, sino también establecer relaciones entre variables sociales, demográficas, sanitarias y económicas, lo cual facilita una toma de decisiones informada en relación a políticas públicas y estrategias privadas, como la **expansión de laboratorios farmacéuticos**.

La limpieza y transformación de los datos fue clave, especialmente debido a errores en variables acumulativas y la existencia de valores nulos en columnas esenciales como new\_recovered y comulative\_recovered. A pesar de esto, se logró estimar nuevos recuperados de forma aceptable y mantener un análisis robusto.

# Reflexión personal

## Reflexión personal sobre el proyecto

Este proyecto representó una oportunidad invaluable para profundizar en el manejo de grandes volúmenes de datos reales y complejos, así como para aplicar diversas técnicas de análisis, limpieza y visualización que son fundamentales en el trabajo cotidiano de un Analista de Datos.

Aprendí a gestionar datasets extensos con millones de registros, a identificar problemas comunes como valores nulos o inconsistencias en variables acumulativas, y a desarrollar soluciones creativas para estimar datos faltantes. Además, la integración de visualizaciones interactivas mediante Power BI me permitió comprender la importancia de presentar información de manera clara y dinámica para facilitar la toma de decisiones estratégicas.

Desde el punto de vista técnico, reforcé habilidades en Python (pandas, numpy), manipulación y transformación de datos, así como en la creación de dashboards efectivos. También desarrollé una mirada crítica para interpretar estadísticas y detectar patrones o anomalías que pueden influir en decisiones empresariales de alto impacto.

### Reflexión sobre el proceso

Si tuviera que volver a comenzar este proyecto, consideraría inicialmente dedicar más tiempo a la revisión y validación de la calidad de los datos proporcionados, particularmente sobre las variables acumulativas, para evitar errores que luego afecten análisis posteriores. Implementar controles de calidad más rigurosos desde el inicio optimizaría el proceso y la confianza en los resultados.

También exploraría la incorporación temprana de herramientas automatizadas de limpieza y ETL para facilitar la gestión de datos masivos, y quizás destinaría más esfuerzo en diseñar visualizaciones que integren más dimensiones (como factores socioeconómicos o demográficos) desde etapas iniciales.

En resumen, aunque el camino recorrido fue sólido y nos permitió alcanzar los objetivos, siempre hay espacio para mejorar en planificación y preparación de datos, que son la base para un análisis confiable y eficiente.