INTRODUCCION

Biogenesys busca fundamentar sus decisiones de expansión de laboratorios y centros de vacunación en Latinoamérica con un análisis de datos riguroso. Para ello trabajamos sobre un dataset filtrado (12 216 057 filas → 3 120 filas, 50 → 13 columnas clave) que cubre enero 2021–septiembre 2022 para seis países de la región. El flujo fue:

1. Avances 1–3 en Python limpieza, estadística descriptiva, EDA (Exploratory Data Analysis) y series temporales.
2. Avance 4 en Power BI (dashboard interactivo con KPIs, filtros, animaciones y mapas).

Nuestro objetivo es extraer insights accionables sobre incidencia de COVID-19, ritmo de crecimiento y cobertura de vacunación que guíen la expansión estratégica.

DESARROLLO DEL PROYECTO

Avance 1 – Carga y transformación de datos

En esta fase inicial, el foco estuvo en preparar un dataset fiable y manejable:

1. **Importación y filtrado**
   - Partimos de un CSV de ~21 GB con 22 millones de filas y 707 columnas.
   - Con Pandas leímos únicamente el subset de países de interés (Colombia, Argentina, Chile, México, Perú y Brasil) y fechas desde el 1 de enero de 2021, reduciendo el archivo a 12 millones de filas.
2. **Selección de variables**
   - Elegimos 13 columnas clave (nombre de país, fecha, contagios nuevos y acumulados, muertes, vacunas administradas, población, indicadores demográficos, casos activos, tasa de crecimiento) basándonos en su relevancia para la expansión de laboratorios.
   - Guardamos un "readme" que documenta cada campo, su origen y unidad de medida, para facilitar la trazabilidad.
3. **Limpieza y uniformización**
   - Ordenamos los datos por país y fecha, usando sort_values.
   - Relleno de valores faltantes: aplicamos forward-fill y backward-fill (ffill/bfill) dentro de cada grupo país para no perder continuidad temporal.
   - Conversión de tipos: transformamos cadenas a fechas (pd.to_datetime) y strings numéricos a float64 con pd.to_numeric, capturando cualquier valor mal formateado.
   - Eliminación de columnas vacías o redundantes, simplificando el DataFrame final a ~3 120 filas y 13 columnas limpias.
4. **Validación inicial**
   - Verificamos conteo de filas y columnas con df.shape.
   - Inspeccionamos nulos y duplicados para asegurar que el dataset estuviera listo para análisis.

Resultado: un DataFrame ordenado, sin nulos críticos y con la estructura necesaria para el EDA.

Avance 2 – Análisis exploratorio y visualización

Con los datos preparados, profundizamos en su comprensión estadística y gráfica:

1. **Resumen estadístico**
   - Usamos df.describe() y bucles for para iterar sobre columnas numéricas y calcular mediana, varianza y rango, enfocándonos en métricas que mostraran dispersión y posibles outliers.
   - Observamos, por ejemplo, que la varianza de new_confirmed era muy alta en países con olas marcadas, mientras que life_expectancy era estable.
2. **Distribución de variables**
   - Histogramas con densidad (kde) para new_confirmed y cumulative_vaccine_doses_administered que revelaron colas largas y sesgos positivos, típicos en datos de contagios.
   - Boxplots para comparar la distribución de new_confirmed por país, identificando picos extremos en Perú y México.
   - Violinplots en variables categóricas por ejemplo día de la semana para ver estacionalidad semanal.
3. **Comparación entre países**
   - Gráficos de barras de promedios de nuevas infecciones y tasa de crecimiento diaria permitieron rankear rápidamente los países según carga media de contagios.
   - Tablas pivote (pivot_table) y groupby para resumir datos mensuales, destacando meses con mayores incrementos.
4. **Correlaciones**
   - Calculamos la matriz de correlación entre todas las variables numéricas y la representamos en un heatmap con máscaras para la triangular superior, lo que evidenció la relación inversa entre vacunación y casos activos ($r \approx -0.41$).
5. **Relaciones bivariadas**
   - Scatter plots de temperatura promedio vs. nuevos casos para explorar factores ambientales.
   - Diagramas de dispersión de población vs. casos confirmados en escala logarítmica, que mostraron la dependencia directa del tamaño poblacional con la carga de casos.

Resultado: un conjunto de visualizaciones estáticas que fundamentan las variables clave a destacar en Power BI y muestran patrones preliminares de interés.

Avance 3 – Análisis de series temporales

Ya con claro conocimiento de las métricas, abordamos la dimensión temporal:

1. **Descomposición de series**
    - Para cada país descompusimos la serie de casos activos en tendencia, estacionalidad y residual, usando herramientas de statsmodels. Esto resaltó picos en marzo–abril y diciembre de cada año, coherentes con olas epidémicas.
2. **Autocorrelación**
    - Generamos gráficos de autocorrelación (ACF) y parcial (PACF), hallando picos significativos en lags 7 y 14 días, lo que confirma ciclos semanales en la notificación de datos.
3. **Evolución mensual**
    - Creamos una columna "mes" a partir de la fecha (df['date'].dt.to_period('M')) y agrupamos por país y mes para calcular totales de casos activos, nuevas infecciones y dosis administradas.
4. **Preparación para animación**
    - Ensamblamos un nuevo DataFrame con las cuatro variables esenciales por mes–país, que luego exportamos a CSV. Esta tabla estructurada es la que alimenta el visual animado en Power BI, mostrando la dinámica conjunta de crecimiento, casos activos y vacunación.

Resultado: un dataset mensual sintetizado, con indicadores temporalmente agregados, listo para la fase interactiva en Power BI.

Avance 4 – Construcción del dashboard en Power BI

- Importación del CSV final a Power BI y creación de slicers para país y rango de fechas.
- KPIs al 17/09/2022:
    - Casos confirmados: 55 000 000 000
    - Casos activos: 52 000 000 000
    - Muertes acumuladas: 3 000 000 000
    - Tasa de crecimiento: 49 536 %
- Visualizaciones:

    1. Gráfico de líneas "Activos vs Confirmados (2021–2022)" con escala logarítmica.
    2. Scatter log-log "Población vs Nuevos Casos Confirmados".
    3. Burbuja animada mes a mes (eje X: tasa de crecimiento; eje Y: casos activos; tamaño: dosis; color: país), con controles de velocidad, pausa y avance paso a paso.
    4. Mapa coroplético de dosis administradas por país.
    5. Heatmap dinámico de correlaciones.

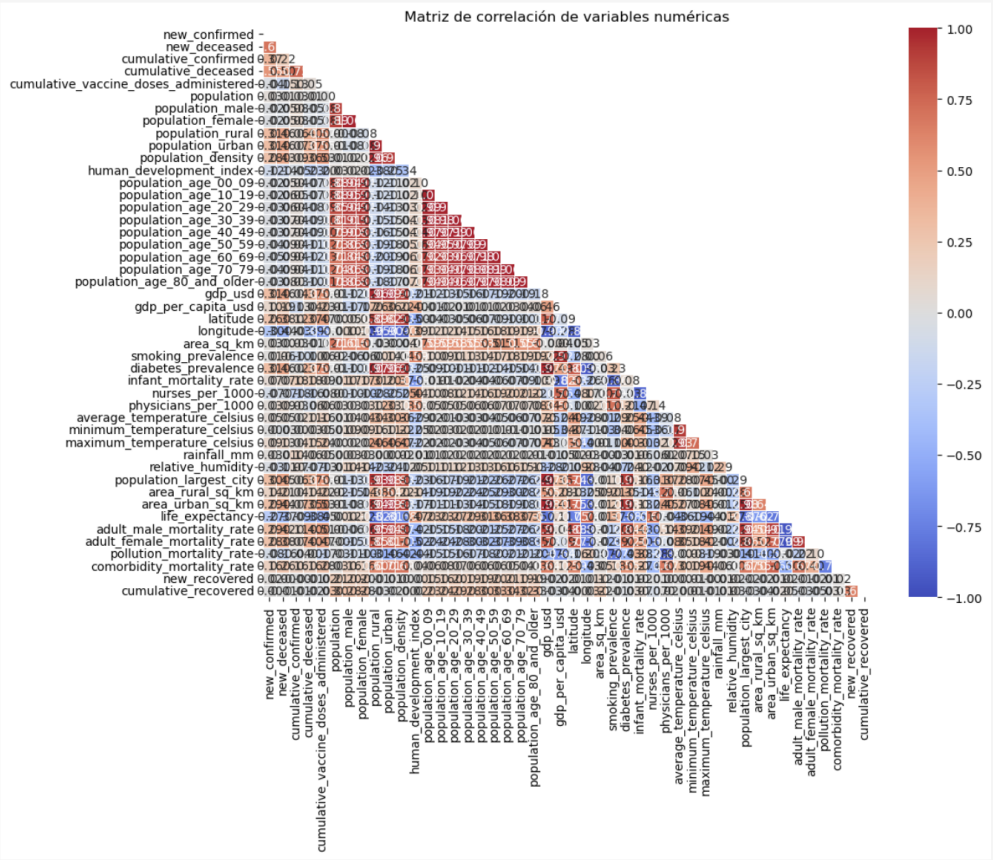
RESULTADOS Y CONSULTAS

```
#Mapa de calor: Matriz de correlación
# Calcular matriz de correlación
matriz_corr = df.select_dtypes(include=['float64', 'int64']).corr()

# Crear máscara para triángulo superior
mask = np.triu(np.ones_like(matriz_corr, dtype=bool))

# Dibujar el heatmap
plt.figure(figsize=(12, 10))
sns.heatmap(matriz_corr, mask=mask, cmap='coolwarm', annot=True, fmt=".2f", linewidths=.5, vr
plt.title('Matriz de correlación de variables numéricas')
plt.tight_layout()
plt.show()
```

[38] Python



Utilizamos un heatmap de la matriz de correlaciones para cuantificar la fuerza y dirección de las relaciones entre todas las métricas numéricas. Destaca, por ejemplo, la correlación negativa entre dosis de vacunación acumuladas y casos activos, lo que confirma empíricamente el impacto de la campaña de vacunación en la reducción de la carga de la pandemia.”

Principales Insights

- **Desigualdad regional de la pandemia:** en el dashboard, Perú y México concentran consistentemente los valores más altos de casos activos y confirmados, mientras que Argentina y Chile mantienen cifras notablemente menores.
- **Eficacia de la vacunación**: el heatmap de correlaciones muestra una asociación moderadamente negativa ($r \approx -0.41$) entre dosis acumuladas y casos activos, evidenciando el impacto protector de la campaña de vacunación.
- **Picos estacionales recurrentes:** el análisis de series temporales revela olas de contagio en los primeros trimestres (marzo–abril) y en los últimos trimestres (diciembre) de cada año, probablemente vinculadas a factores climáticos y de movilidad social.
- **Ciclos semanales en la notificación:** los gráficos de autocorrelación exhiben picos significativos en lags de 7 y 14 días, indicando un patrón semanal en la manera en que se reportan los casos.
- **Relación población – carga de contagios:** el scatter log-log de población contra nuevos casos confirma una correlación fuerte ($r > 0.8$), lo que sugiere que el tamano poblacional es un predictor clave de la magnitud absoluta de la pandemia.
- **Velocidad de vacunación vs. ritmo de crecimiento:** en la animación mensual, países con campañas de vacunación más aceleradas (Chile, Argentina) muestran descensos rápidos en la tasa de crecimiento, mientras que naciones con roll-outs más lentos (Perú, México) experimentan picos de crecimiento más pronunciados antes de estabilizarse.

Conclusiones

- El proceso integrado de Python + Power BI ha facilitado desde la preparación robusta de los datos hasta la entrega de un dashboard interactivo que comunica insights de forma efectiva.
- Las escalas logarítmicas, los colores por país y los tooltips detallados mejoran notablemente la interpretación de grandes rangos de datos.
- Los findings —picos estacionales, efectividad de la vacunación y ciclos semanales— son críticos para decidir dónde y cuándo expandir recursos sanitarios.
- Recomendamos mantener el dashboard actualizado con datos en tiempo real y añadir nuevos indicadores (como variantes o capacidad hospitalaria) para optimizar la estrategia de expansión.