Proyecto INT M4

Nombre del autor: Juan Sebastian Barreto Benavides

Email: Jsbarreto.sparking@gmail.com

Cohorte: DA-FT13

Fecha de entrega: 12/05/2025

Institución:Farmotudo

Introducción

los objetivos del proyecto y alcanzados son:

- aplicar tecnicas de carga,filtrado,limpieza y transformación de datos utilizando python y la librerias Pandas y Numpy
- aplicar herramientas avanzadas de estadistica y visualizacion de datos
- integrar python con power bi para importar y analizar el cojunto de datos y preparado para el dashboard interactivo de los resultados

Desarrollo del proyecto

avanze 1

1. Filtrado por país

- Se seleccionaron los países: **Colombia, Argentina, Chile, México, Perú y Brasil**.
- Se aplicó un filtro usando `.isin()` para quedarnos solo con registros de estos países.

2. Exploración inicial del dataset

- Se revisaron:
  - Tamaño del dataset antes y después del filtrado.
  - Fechas mínimas y máximas.
  - Columnas disponibles.
  - Tipo de datos de la columna `date`.

3. Filtrado por fecha

- Se excluyeron registros anteriores al **1 de enero de 2021** para mantener datos recientes y relevantes.

4. Eliminación de nulos extremos

- Se eliminaron:
  - Filas y columnas completamente vacías (`dropna`).
  - Columnas con más de **4 millones** de valores nulos.

5. Filtrado por códigos de país

- Se mantuvieron únicamente las filas donde el código de país (`location_key`) fuera uno de los siguientes: `AR`, `CO`, `CL`, `MX`, `PE`, `BR`.

6. Limpieza de la columna de fechas y nulos

- Se convirtió la columna `date` a formato datetime.
- Se rellenaron valores nulos con:
  - **Promedio** (para columnas numéricas).
  - **Forward fill (ffill)** y **backward fill (bfill)** para mantener consistencia.

7. Análisis exploratorio general

- Se utilizaron `describe()` y `info()` para conocer:
  - Distribución de datos.
  - Cantidad de valores nulos.
  - Medidas estadísticas básicas.

8. Selección de columnas clave para análisis

- Se eligieron columnas numéricas relevantes como:
  - `new_confirmed`, `new_deceased`, `population`, `gdp_usd`, `life_expectancy`
- **Justificación**:
  - Permiten análisis proporcionales, económicos y demográficos del impacto del COVID.

9. Guardado del dataset limpio

- Se exportó el dataset limpio como `DatosFinalesFiltrado.csv`.

10. Estadísticas descriptivas

- Se calcularon automáticamente **media, mediana, mínimo, máximo y desviación estándar** para cada columna numérica usando bucles.

11. Función personalizada de estadística

- Se definió una función para calcular:
  - **Media**, **mediana**, **varianza** y **rango** de cualquier columna numérica.

avance 2

1. Estadísticas generales por país

- Se calcularon:
  - Promedio, mediana, desviación estándar, mínimo y máximo
  - Para columnas clave relacionadas con: casos, muertes y vacunación.

2. Matriz de correlación

- Se construyó una matriz de correlación entre variables como:

    - Casos diarios/acumulados, muertes, densidad poblacional, esperanza de vida, diabetes, tabaquismo, etc.
- **Objetivo**: descubrir relaciones entre salud, demografía y economía.

3. Histograma y densidad de nuevos casos confirmados

- Se generaron histogramas por país con densidad para observar:
    - Distribución de nuevos contagios diarios.
    - Posibles outliers (valores atípicos) y concentración de datos.

avance 3

1. Verificación de valores nulos

```
print(data.isnull().sum())
print(data.isnull().values.any())
```

Analizaste si el DataFrame contenía valores nulos. Esto es importante para garantizar la calidad de los datos antes de realizar análisis más profundos.

2. Resumen por país (últimos valores acumulados)

```
tabla_resultado = data_latinoamerica_final.groupby("country_name").agg({
    "cumulative_confirmed": "last",
    "cumulative_recovered": "last"
}).reset_index()
```

Agrupaste los datos por país y seleccionaste el último dato disponible de casos **confirmados** y **recuperados** acumulados. Esto permitió tener un resumen actualizado por país.

3. Cálculo de Casos Activos Estimados y comparación visual

```
data_latinoamerica_final["casos_activos_estimados"] = (
    data_latinoamerica_final["cumulative_confirmed"]
    - data_latinoamerica_final["cumulative_recovered"]
```

```
    - data_latinoamerica_final["cumulative_deceased"]
)
```

Calculaste los **casos activos estimados** como la diferencia entre confirmados, recuperados y fallecidos. Luego comparaste estos datos con los recuperados en un gráfico de línea para **Chile**.

📉 Gráfico:

- Línea naranja = Casos activos estimados
- Línea verde = Recuperados
- Permite ver visualmente la evolución de la pandemia.

4. Análisis de correlación entre variables

```
corr = data_latinoamerica_final[corr].corr()
sns.heatmap(corr, annot=True, cmap="coolwarm", fmt=".2f")
```

Seleccionaste variables relevantes (como nuevos casos, densidad poblacional, expectativa de vida, prevalencia de diabetes, etc.) y calculaste una **matriz de correlación**.

El **heatmap** te permitió observar qué variables están más relacionadas entre sí, lo cual es clave para entender **factores de riesgo o protección**.

5. Clasificación de nuevos casos según promedio

```
media_casos = data_latinoamerica_final["new_confirmed"].mean()
```

Calculaste la **media de nuevos casos** en el conjunto de datos.

Luego definiste una función para clasificar si los nuevos casos eran **“Altos”** o **“Bajos”** según si estaban por encima o por debajo de la media, y creaste una nueva columna `clasificacion_nuevos_casos`.

```
data_latinoamerica_final["clasificacion_nuevos_casos"] = ...
```

Esto es útil para segmentar datos y analizar comportamientos diferentes entre los países o momentos con alta o baja incidencia.

6. Filtrado de datos clasificados como "Altos"

```
alta_clasificacion = data_latinoamerica_final[data_latinoamerica_final["clasificacion_nuevos_casos"] == "Alto"]
```

Filtraste y mostraste solo los registros clasificados como de **alto número de nuevos casos**, para analizarlos por separado.

7. Verificación adicional

Revisaste nuevamente:

- Valores nulos con `isnull().sum()`
- La estructura de los datos con `print(data_latinoamerica_final)`
- La frecuencia de cada país con `value_counts()`

EDA e insights

- se obtuvo casos acumulados por paises para saber que pais tiene mas casos
- dosis de vacunas de cada pais
- los recuperados de cada pais
- poblacion que tiene cada pais
- evolucion de casos activos vs recuperados en colombia
- relacion entre vacunacion y vasos confirmados
- promedio de nuevos casos confirmados por pais

Análisis del dashboard

- dashboard tiene una portada donde lo lleva graficas donde tienen diferentes graficos
- tiene boton de devolver a la portada
- tiene filtrado de paises como AR,BR,CL,CO,MX,PE
- al filtrar todo se cambia

Conclusiones y Recomendaciones

1. **Casos Confirmados Acumulados**

   Brasil presenta la mayor cantidad de casos confirmados acumulados, lo que destaca su relevancia en el monitoreo de la pandemia en la región. Argentina también muestra un número significativo, seguido por Perú y Colombia, lo que permite identificar los países con mayor impacto sanitario.

2. **Dosis de Vacunas Administradas**

   La distribución de las dosis de vacunas administradas se encuentra con mas porcentaje de brasil y de segundo lugar es de mexico

3. **Nuevos Recuperados**

   Los países presentan cifras destacables pero los que mas recuperados tiene es el pais de brasil con 5,565,456 y en segundo es mexico con 2,475,518

4. **Población Total**

   Los datos permiten visualizar la población total de cada país de primer lugar esta brasil y segundo mexico

5. **Promedio de casos confirmados**

   el promedio de casos confirmados mas alto es chile y argentina en promedio los demas paises son mas bajos

6. **Demanda de Vacunas**

   Los países analizados (Argentina, Brasil, Chile, Colombia, México y Perú) presentan altos niveles de casos confirmados acumulados, especialmente Brasil y Argentina. Esto indica una **alta demanda potencial de vacunas**, principalmente en regiones con mayor densidad poblacional y afectación. La distribución uniforme de dosis administradas entre los países también sugiere que todos mantienen una necesidad constante de suministros para continuar sus campañas de inmunización.

7. **Relacion alta con la vacunacion y casos confirmados**

   brasil tiene una relacion alta con la vacunacion y los casos confirmados en comparacion del resto del pais

8. **Nuevos fallecidos por pais**

   el pais con mas fallecidos tiene es bracil con una cifra demasiada alta a comparacion con los demas es bracil y el segundo es mexico

CONCLUSIONES:

Brasil representa una oportunidad estratégica clave para la expansión dela empresa farmacéutica, debido a su elevada población y a la alta cantidad de casos confirmados acumulados de COVID-19. Estos factores lo convierten en el país con mayor demanda potencial de productos médicos y vacunas en la región. Además, el elevado número de dosis de vacunas administradas y recuperados evidencia un sistema de salud activo y una población dispuesta a acceder a tratamientos.

Reflexión personal

A lo largo de este proyecto he aprendido muchísimo, viendo videos, investigando y luchando para que cada paso del análisis saliera bien. Ha sido un proceso acelerado en el que he tenido que adaptarme rápido, resolver errores y aplicar nuevos conceptos constantemente. Aunque a veces siento que el ritmo no me permite practicar tanto como me gustaría, reconozco que este esfuerzo me ha permitido avanzar mucho más de lo que imaginaba. Me encantaría poder dedicar más tiempo a practicar, reforzar lo aprendido y así seguir mejorando mis habilidades de forma más sólida.