

Análisis Exploratorio de Datos

Fuente: who.csv o who2.csv (datos de la OMS sobre tuberculosis)

1. Carga y Estructura de los Datos

El análisis comienza con la carga de un archivo CSV que contiene información agregada por país, año, sexo, edad y tipo de tuberculosis.

- Variables iniciales:
 - country, iso2, iso3, year
 - Columnas adicionales codificadas como new_sp_m014, new_ep_f25, etc., que combinan:
 - Tipo
 - sexo
 - grupo de edad

2. Transformación a Formato Largo

Se realiza una transformación con `.melt()` para llevar los datos a formato largo (long format), necesario para análisis más flexibles, obteniendo una fila por combinación única de país, año, tipo de TB, sexo, edad y número de casos.

- Ejemplo de variable descompuesta:
 - new_sp_m014 → type = sp, sex = m, age = 014, cases = n

3. Limpieza y Conversión de Datos

Se eliminan los valores faltantes (NaN) en la columna cases, se convierten los valores de cases a tipo entero (int), posteriormente se extraen las subcolumnas prefix, type, sex, age a partir de la columna "variable", posteriormente se eliminan columnas innecesarias, dejando un dataframe limpio con las siguientes variables clave:

- country, iso2, iso3, year, type, sex, age, cases

4. Filtrado por Tipo: TB Pulmonar Positiva (sp)

Se crea un subconjunto `df_sp` que contiene solo los casos del tipo “pulmonar positiva” (sp), por ser el más representativo, transmisible y comparable.

5. Tendencia Global de Casos

Se agrupan los datos por año y se suman los casos para obtener la evolución temporal total, y así se visualiza con un gráfico de línea (lineplot), mostrando:

- Disminución progresiva de los casos globales.
- Picos posibles alrededor de los años 1995–2000, luego tendencia a la baja.

6. Casos por Sexo

Se realiza un análisis por sexo agrupando por año y sexo, generando una visualización con líneas separadas por sexo.

7. Top 10 Países con Mayor Carga Acumulada

Agrupé los datos por país y se suman los casos totales, se seleccionan los 10 países con mayor número acumulado.

8. Evolución por Edad

Se hizo la agrupación por year y age, los grupos etarios con mayor número de casos:

- Adultos jóvenes y de mediana edad (15–54 años).
- Muestra la evolución por cohortes a lo largo del tiempo.

9. Evolución Temporal en India

Generé un subconjunto para el país “India”, y grafiqué los casos totales por año, el resultado, la India ha mostrado una disminución sostenida en el tiempo, aunque mantiene cifras elevadas.

10. Distribución por Tipo de TB y Sexo

Realicé una agrupación por type y sex, el resultado:

- El tipo pulmonar positiva (sp) es el más frecuente en ambos sexos.
- Mayor número de casos nuevamente en hombres.

11. Promedio Anual de Casos por País

Se hizo una agrupación por país, cálculo del promedio anual, fue similar al acumulado, pero permite identificar países con carga sostenida, no sólo por población: India lidera también esta métrica.

12. Heatmap Casos por País y Año

Se construye una tabla dinámica (pivot) con países en filas, años en columnas y valores de casos, se filtran solo los países con más de 8 años de datos disponibles, el heatmap evidencia:

- Persistencia de casos en países específicos.
- Mejoras notables en otros.
- Potenciales outliers en ciertos años.

Principales Patrones y Hallazgos

- Tendencia global decreciente en número de casos de TB.
- Desigualdad por sexo: hombres con mayor carga en todos los periodos.
- Concentración geográfica: pocos países representan la mayor parte de los casos.
- Carga en edad productiva: riesgo sanitario y económico.
- Algunos países muestran progresos claros, pero otros no presentan mejoras consistentes.







