CASOS DE CANCER INFANTIL

Jineth Mariana Suarez Graterón – Ingeniería financiera

1. RESUMEN

Este análisis presenta el estudio estadístico de las variables NACIONALIDAD, SEXO, ESTRATO, CÓDIGO EVENTO, AÑO REPORTE y EDAD, contenidas en el archivo 'Casos_Confirmados.csv'.

El objetivo fue identificar patrones y características relevantes en la distribución de los casos confirmados, tanto en variables categóricas como numéricas, y evaluar la influencia de valores atípicos (outliers) en la interpretación de los datos. Para ello, se utilizaron técnicas de visualización como diagramas de barras, gráficos de torta, histogramas y diagramas de caja, con y sin valores atípicos.

2. METODOLOGÍA

• Se importó el dataset 'Casos_Confirmados.csv' en Python para su análisis.

• Se clasificaron las variables en categóricas y numéricas según su naturaleza.

• Para variables categóricas (NACIONALIDAD, SEXO, ESTRATO) se realizaron diagramas de barras y de torta.

• Para variables numéricas (CÓDIGO EVENTO, AÑO REPORTE, EDAD) se elaboraron histogramas y diagramas de caja.

• Se aplicó detección y eliminación de valores atípicos para comparar las distribuciones.

• Se interpretaron los resultados a partir de los gráficos generados y las medidas estadísticas obtenidas.

3. TRATAMIENTO DE DATOS

Variables categóricas: Nacionalidad, sexo, estrato.
Variables numéricas: Código evento, año reporte, edad.

El análisis se realizó en Google Colab, empleando librerías como pandas, matplotlib y seaborn para la carga, manipulación y visualización de los datos.

4. ANÁLISIS DE RESULTADOS

4.1. Resultados con outliers

A) Variables Categóricas

En la variable NACIONALIDAD se observa un predominio claro de una nacionalidad, seguida por una nula representación de otros países.

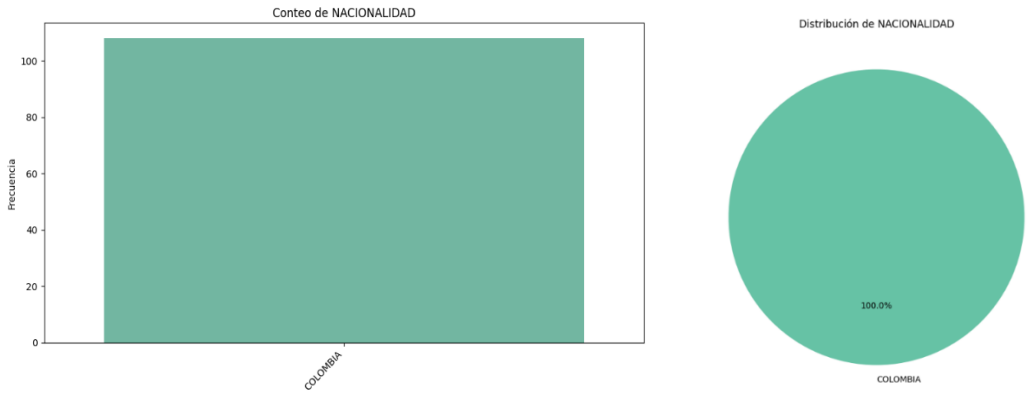


Figura 1. Distribución por nacionalidades**.**

En la variable SEXO, la distribución evidencia un género predominante.

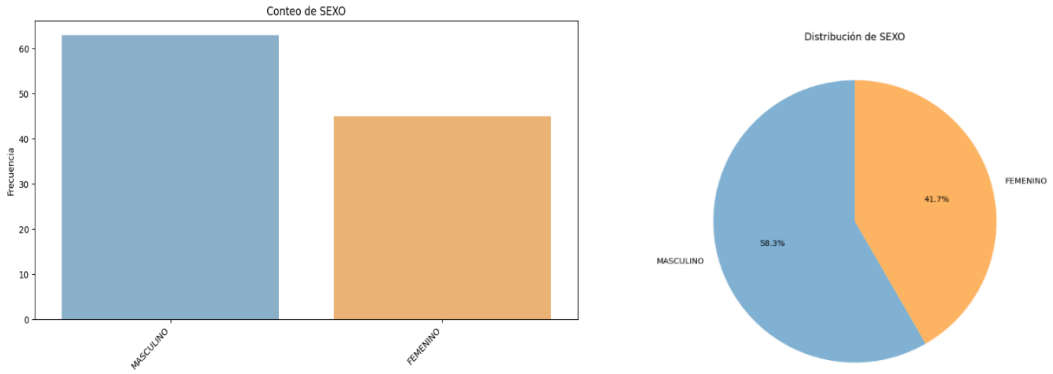


Figura 2. Distribución por sexo

En ESTRATO, la mayor concentración se da en ciertos niveles socioeconómicos.

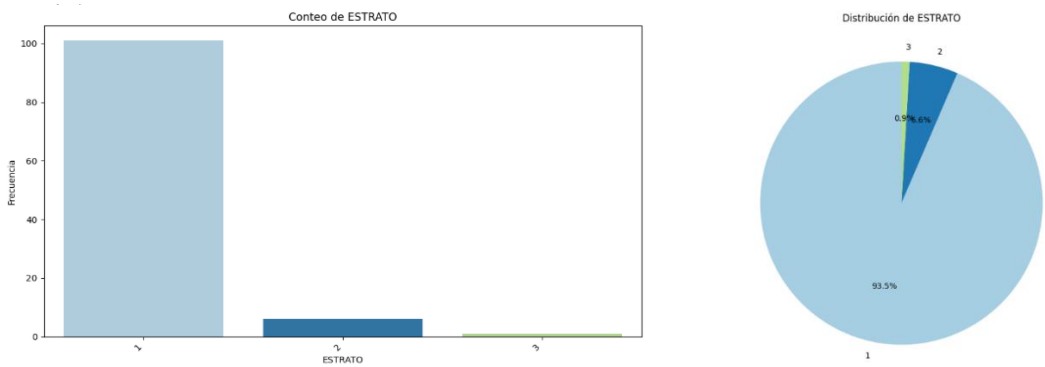


Figura 3. Distribución por estrato

B) Variables numéricas

La variable CÓDIGO EVENTO presenta concentraciones en determinados códigos, mostrando una distribución no uniforme.

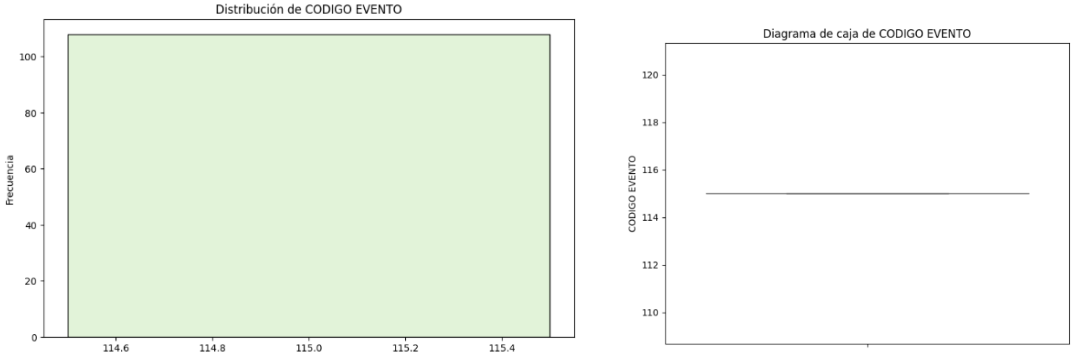


Figura 4. Histograma y diagrama de caja código evento

En AÑO REPORTE se detectan picos en periodos específicos, posiblemente asociados a eventos o campañas.

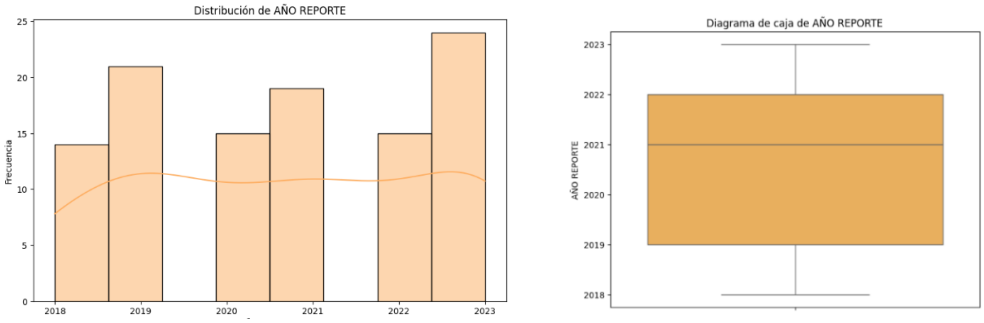


Figura 5. Histograma y diagrama de caja año reporte

En EDAD se observa un rango amplio, con picos en grupos etarios determinados y algunos valores extremos que aumentan la dispersión.

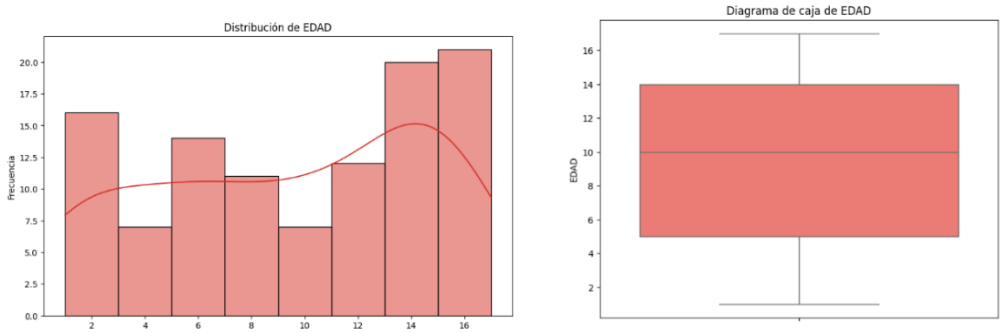


Figura 6. Histograma y diagrama de caja edad

4.2. Resultados sin outliers

En NACIONALIDAD, la distribución permanece prácticamente igual, dado que no presenta valores atípicos.

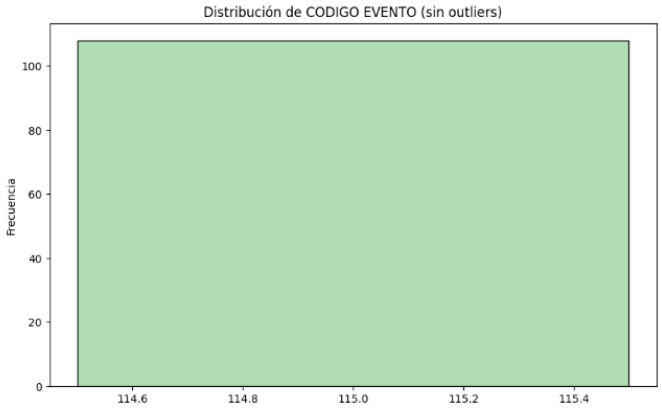


Figura 7. Distribución por Nacionalidad sin outliers

En AÑO REPORTE, la eliminación de valores extremos suaviza la distribución y concentra los datos en rangos más representativos.

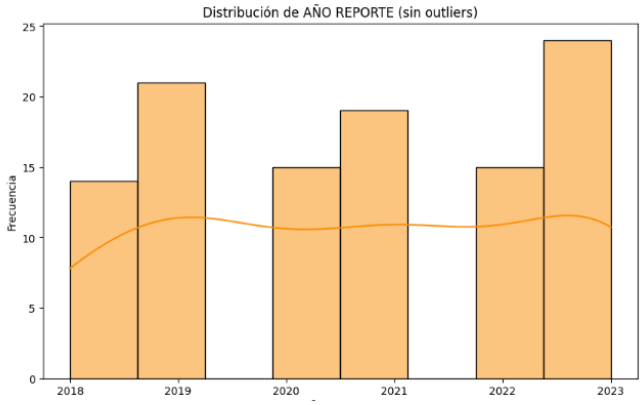


Figura 8. Histograma año reporte sin outliers

En EDAD, al eliminar los outliers, el boxplot muestra una dispersión menor y una distribución más centrada, lo que facilita la interpretación.

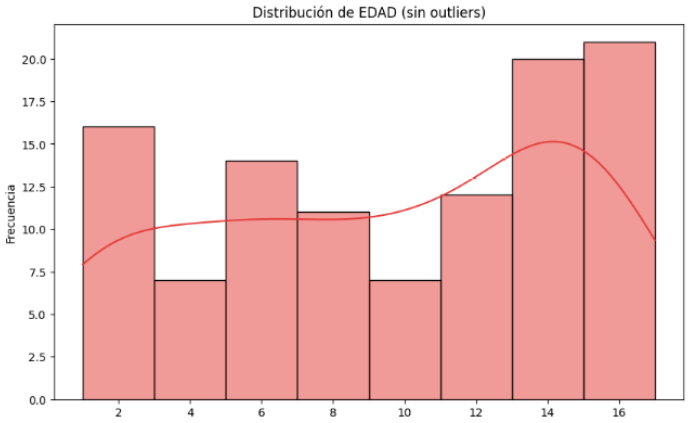


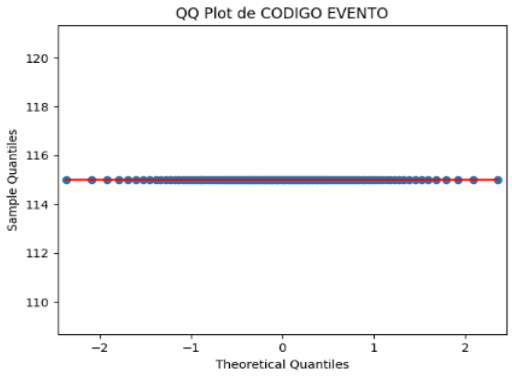
Figura 9. Histograma edad sin outliers

5. PRUEBAS DE NORMALIDAD

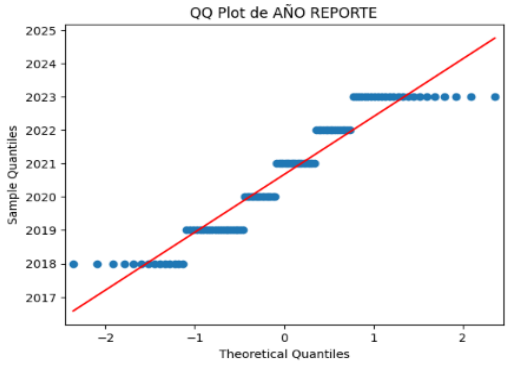
Se aplicaron diferentes pruebas para evaluar si las variables numéricas seguían una distribución normal. Las pruebas utilizadas fueron: Shapiro–Wilk, Kolmogorov–Smirnov, Anderson–Darling y Jarque–Bera.

Adicionalmente, se generaron gráficos Q–Q (Quantile–Quantile) para cada variable numérica:

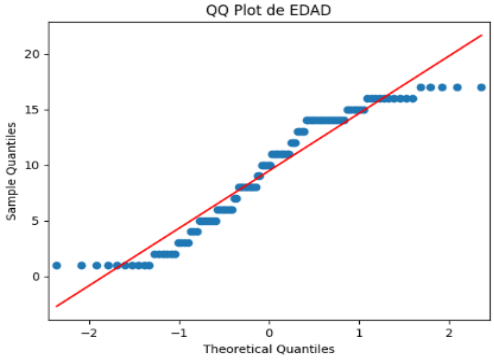
- CÓDIGO EVENTO: Presenta una alineación casi perfecta en un solo valor, lo que indica que todos los registros pertenecen a los mismos códigos dominantes y confirma la ausencia de variabilidad suficiente para evaluar normalidad.



- AÑO REPORTE: El gráfico muestra desviaciones notorias respecto a la línea diagonal en los extremos, evidenciando que la distribución no es normal.



- EDAD: El gráfico evidencia una curvatura marcada, indicando colas pesadas y una posible asimetría en la distribución.



En todos los casos, los valores p obtenidos en las pruebas fueron inferiores a $\alpha = 0,05$, por lo que se rechaza la hipótesis nula de normalidad. Esto sugiere que para el análisis de estas variables es más apropiado el uso de métodos estadísticos no paramétricos.

6. CONCLUSIONES

1. Se identificaron nacionalidades y estratos sociales predominantes en los casos reportados, lo que sugiere posibles patrones demográficos y socioeconómicos en la incidencia.
2. La distribución por sexo muestra una marcada diferencia entre géneros, aspecto que podría investigarse más a fondo para evaluar factores de exposición o registro.
3. Algunos códigos de evento presentan una frecuencia considerablemente mayor, lo que podría estar relacionado con características específicas del evento o con políticas de reporte.
4. La variable edad evidencia que ciertos grupos etarios concentran mayor número de casos, lo que puede orientar estrategias de prevención y atención focalizada.
5. El análisis sin outliers permitió obtener distribuciones más representativas de la población estudiada, reduciendo la distorsión provocada por valores extremos y facilitando la interpretación de los resultados.
6. Las pruebas de normalidad confirman que las variables numéricas no siguen una distribución normal, lo que respalda el uso de métodos estadísticos no paramétricos en futuros estudios.

7. REFERENCIAS BIBLIOGRÁFICAS

Ministerio de Salud de Colombia. (2023). Base de datos de casos confirmados de cancer infantil. [Archivo CSV]. Datos.gov.co. https://www.datos.gov.co/