Reporte final

Título: Análisis estadístico de nacidos vivos en Hospital Manuel Uribe Ángel:

Estudiante: Andrea Natalia Gutiérrez Calderón

Fecha:08/08/2025

Profesor: Andrés Fabian Leal Archila

1. INTRODUCCION

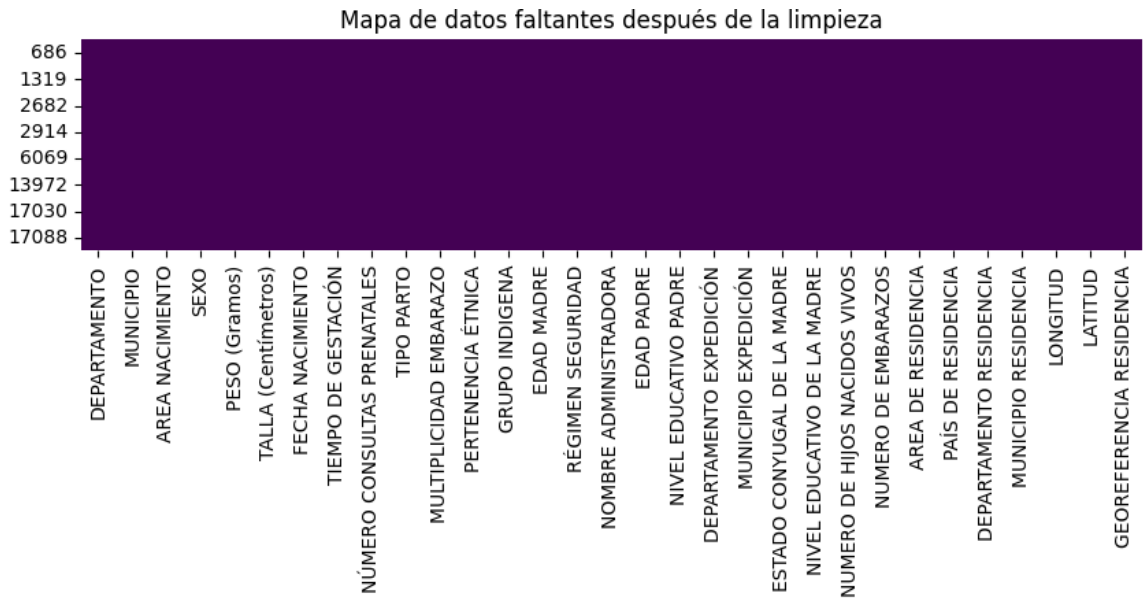
Este análisis exploratorio de datos se enfoca en un subconjunto del dataset de nacimientos. Se prestará especial atención a seis variables clave: SEXO, PESO (Gramos), TIPO PARTO, TALLA (Centímetros), NUMERO DE HIJOS NACIDOS VIVOS y RÉGIMEN SEGURIDAD. A través de visualizaciones y estadísticas descriptivas, buscaremos comprender las características de los nacimientos en función de estas variables, explorando la distribución por género, el análisis del peso y la talla al nacer, los diferentes tipos de parto y su frecuencia, la paridad de las madres y la distribución de los regímenes de seguridad social.

2. CARGA Y EXPLORACIÓN INICIAL DE DATOS

Se cargó el archivo CSV en un DataFrame de pandas y se realizó una exploración preliminar para comprender la estructura del conjunto de datos. Esta exploración incluyó la revisión de las primeras filas para observar el formato y contenido de las variables, así como la obtención de información general para identificar tipos de datos y verificar la existencia de valores no nulos. El dataset original cuenta con 17.299 registros y 31 columnas. También se generó un resumen estadístico de las variables numéricas, que permitió conocer medidas de tendencia central, dispersión y rango; por ejemplo, el peso promedio al nacer se estimó en aproximadamente 3.057 gramos.

3. IDENTIFICACIÓN Y MANEJO DE VALORES FALTANTES

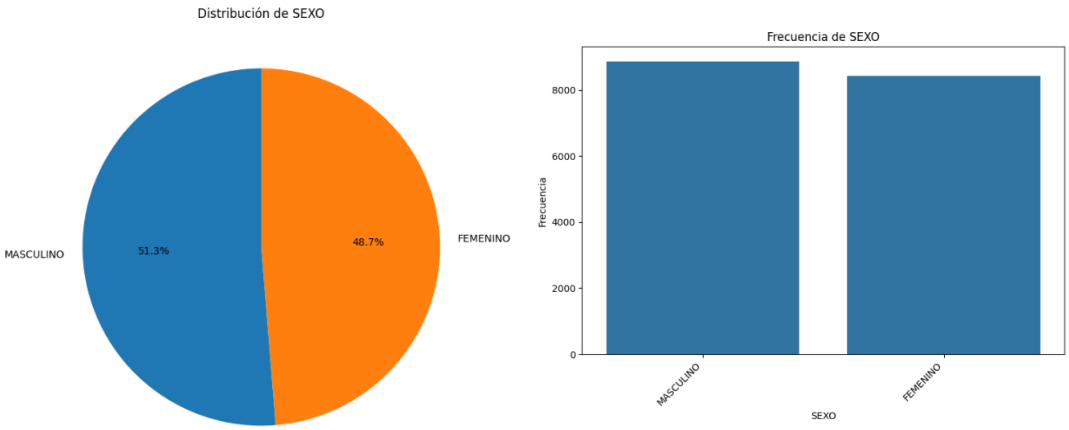
Se detectó la presencia de valores faltantes en varias columnas, siendo especialmente significativa la ausencia de datos en “GRUPO INDÍGENA” (17.290 valores nulos). También se encontraron vacíos en “NOMBRE ADMINISTRADORA” (1.655), “EDAD PADRE” (58), “NIVEL EDUCATIVO PADRE” (310), “LONGITUD” (4), “LATITUD” (4) y “GEOREFERENCIA RESIDENCIA” (5). Para un análisis inicial, se optó por eliminar las filas con datos faltantes en estas columnas, lo que redujo temporalmente la muestra a 8 registros. En etapas posteriores se emplearon estrategias de limpieza distintas con el fin de conservar una mayor cantidad de información.

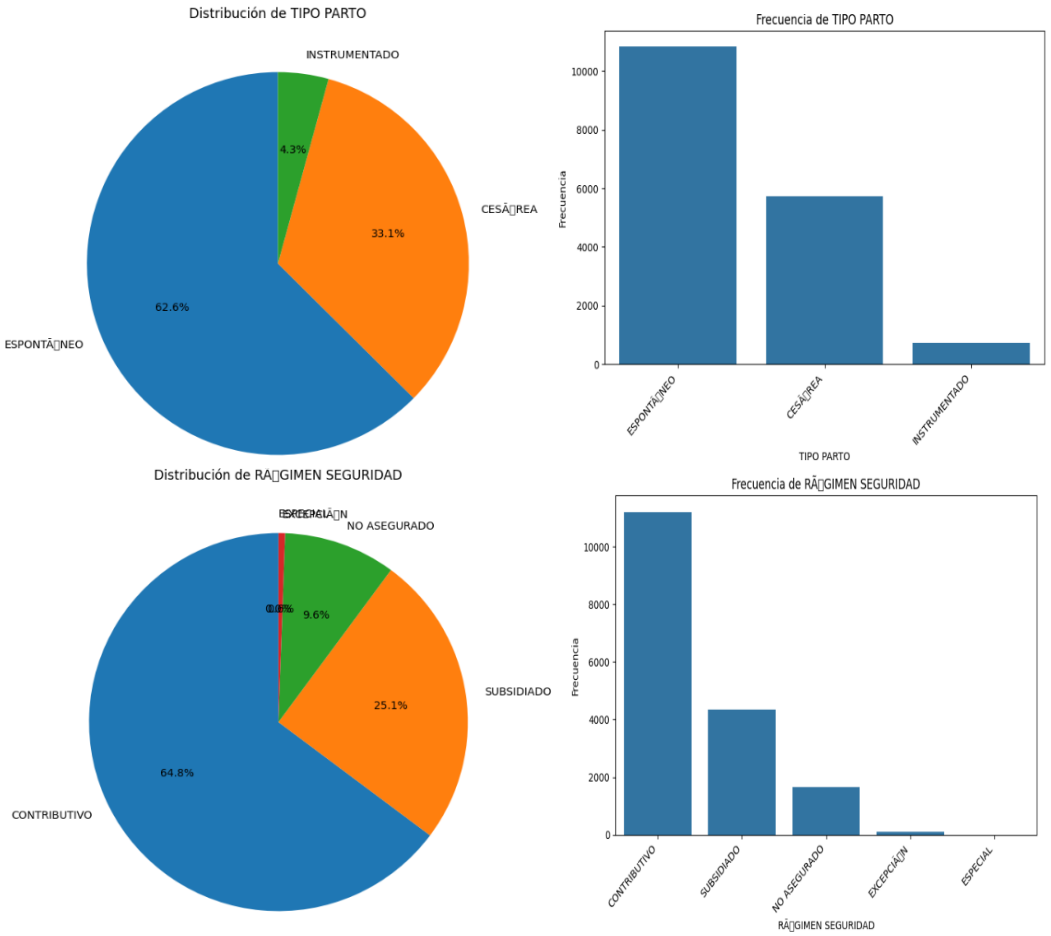


4. VISUALIZACIÓN DE DATOS (VARIABLES SELECCIONADAS)

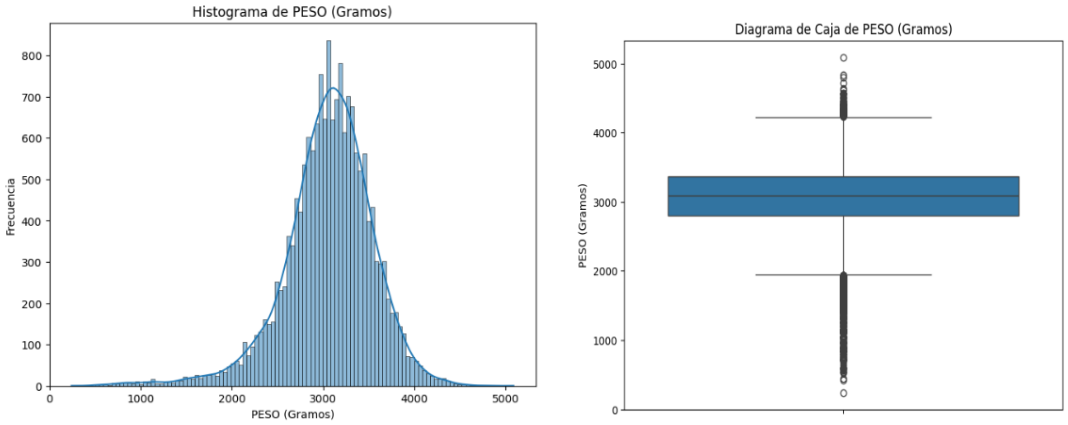
Se analizaron seis variables específicas: SEXO, PESO (gramos), TIPO DE PARTO, TALLA (centímetros), NÚMERO DE HIJOS NACIDOS VIVOS y RÉGIMEN DE SEGURIDAD.

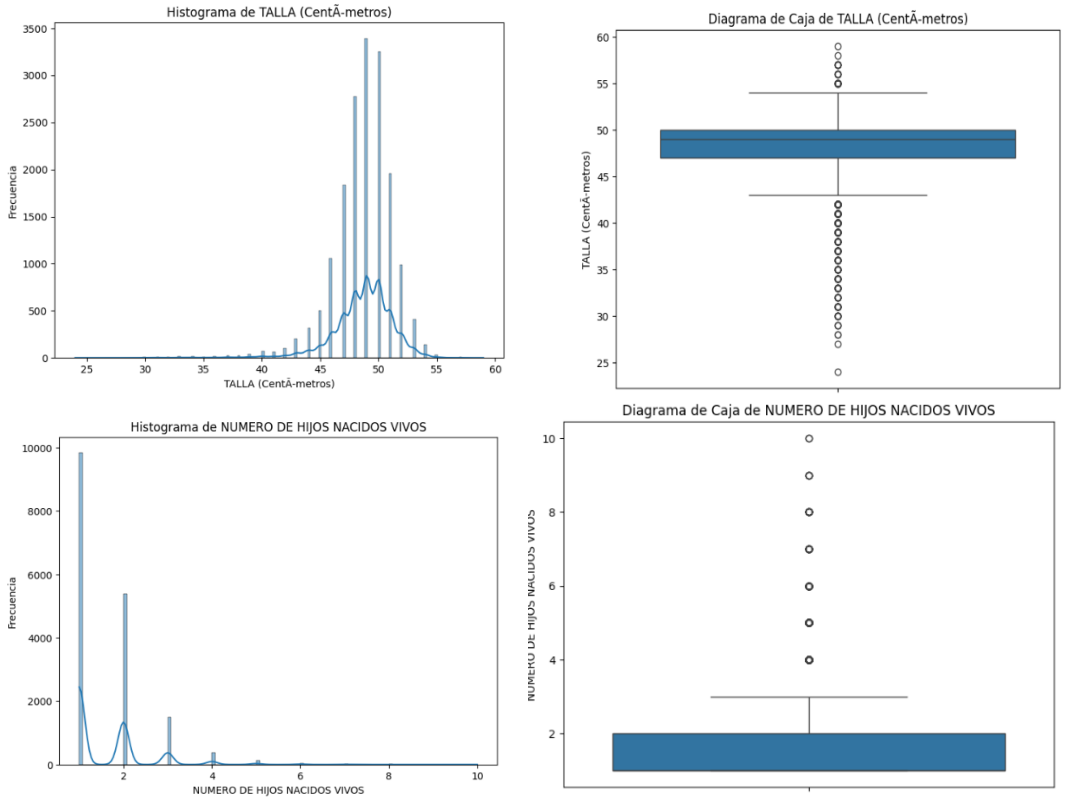
- En las variables categóricas, la distribución fue la siguiente: SEXO (51,3 % masculino y 48,7 % femenino), TIPO DE PARTO (62,6 % espontáneo, 33,1 % cesárea y 4,3 % instrumentado) y RÉGIMEN DE SEGURIDAD (64,8 % contributivo, 25,1 % subsidiado y 9,6 % no asegurado).





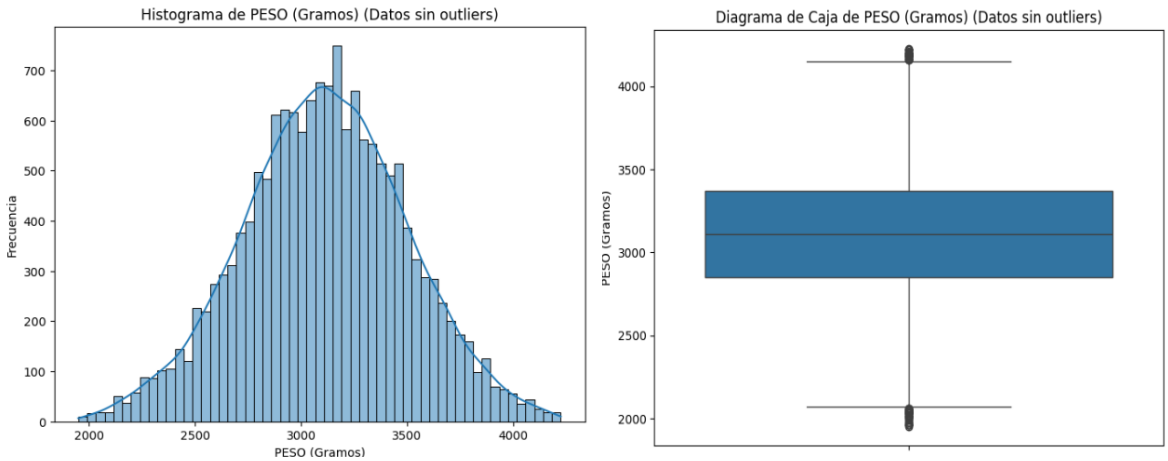
- En las variables numéricas, los histogramas y diagramas de caja evidenciaron la dispersión de los datos y la presencia de posibles valores atípicos.

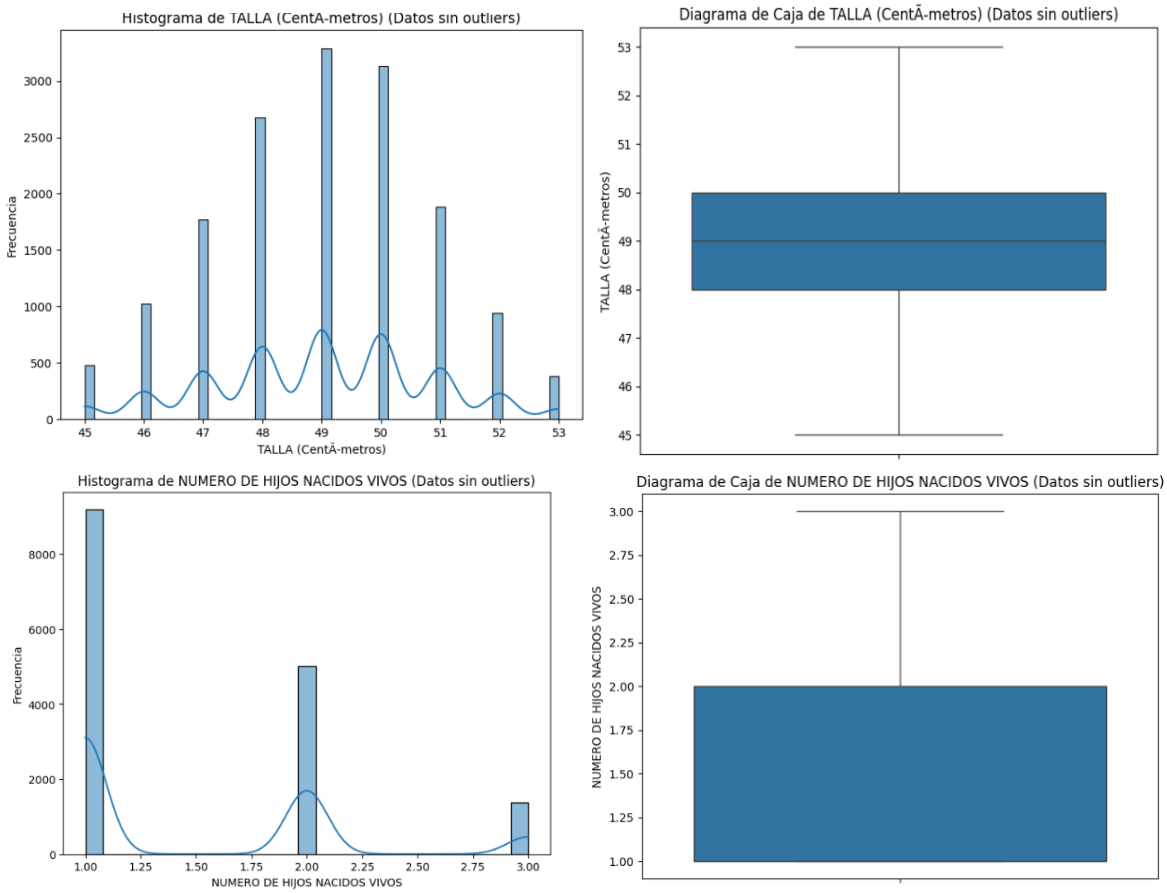




5. DETECCIÓN Y ELIMINACIÓN DE VALORES ATÍPICOS

Se identificaron outliers en las variables “PESO (gramos)”, “TALLA (centímetros)” y “NÚMERO DE HIJOS NACIDOS VIVOS” mediante el método del rango intercuartílico. La exclusión de estos valores redujo el número de registros de 17.299 a 15.578, eliminándose un total de 1.721 filas. Tras esta depuración, los rangos de las variables se ajustaron; por ejemplo, el peso al nacer pasó a oscilar entre 1.950 y 4.225 gramos.





6. ANÁLISIS DE TEST DE NORMALIDAD

Para la variable PESO (gramos), las pruebas de Shapiro-Wilk, Kolmogorov-Smirnov (Lilliefors), Anderson-Darling y Jarque-Bera arrojaron valores p muy bajos (en su mayoría cercanos a 0,0000), lo que lleva a rechazar la hipótesis nula de normalidad. Aunque el estadístico de Shapiro-Wilk (0,9990) es cercano a 1, el tamaño de la muestra hace que incluso pequeñas desviaciones sean estadísticamente significativas. El estadístico de Anderson-Darling (1,0371) supera algunos valores críticos según el nivel de significancia considerado, confirmando que la distribución no es perfectamente normal.

En el caso de la variable TALLA (centímetros), los resultados fueron aún más contundentes: todas las pruebas produjeron valores p de 0,0000, y el estadístico de Anderson-Darling (195,9395) excedió ampliamente los valores críticos. Esto indica una desviación clara respecto a la normalidad. El estadístico de Shapiro-Wilk (0,9684) es menor que el de peso, lo que sugiere una distribución más alejada de la normal.

En síntesis, tanto el peso como la talla, aun después de eliminar outliers, no siguen una distribución normal de forma estricta, resultado coherente con la evidencia gráfica obtenida mediante los gráficos Q-Q.

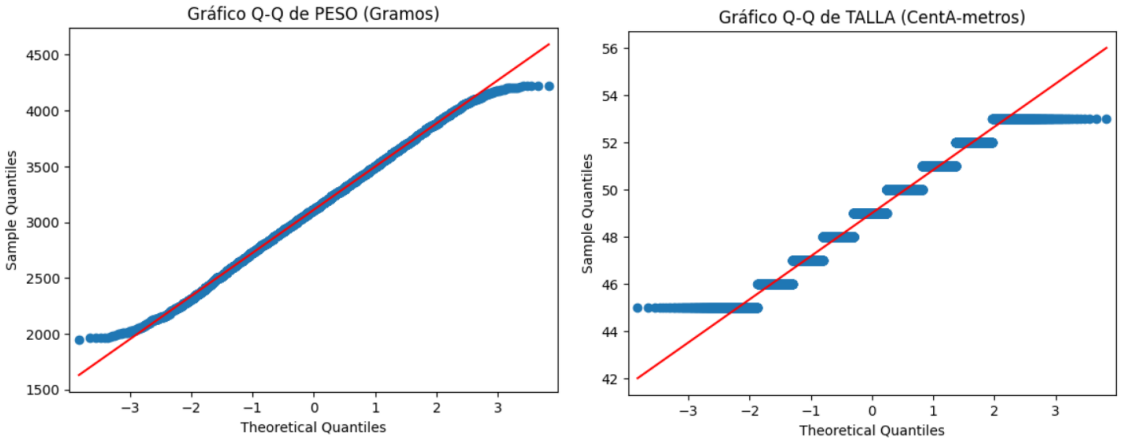
```
Resultados de los Tests de Normalidad:

--- Variable: PESO (Gramos) ---
Shapiro-Wilk: Estadístico=0.9990, p-valor=0.0000
Kolmogorov-Smirnov (Lilliefors): Estadístico=0.0084, p-valor=0.0080
Anderson-Darling: Estadístico=1.0371
  Nivel de significancia 15.0%: Valor crítico=0.5760
  Nivel de significancia 10.0%: Valor crítico=0.6560
  Nivel de significancia 5.0%: Valor crítico=0.7870
  Nivel de significancia 2.5%: Valor crítico=0.9180
  Nivel de significancia 1.0%: Valor crítico=1.0920
Jarque-Bera: Estadístico=12.6448, p-valor=0.0018

--- Variable: TALLA (CentÃmetros) ---
Shapiro-Wilk: Estadístico=0.9684, p-valor=0.0000
Kolmogorov-Smirnov (Lilliefors): Estadístico=0.1172, p-valor=0.0010
Anderson-Darling: Estadístico=195.9395
  Nivel de significancia 15.0%: Valor crítico=0.5760
  Nivel de significancia 10.0%: Valor crítico=0.6560
  Nivel de significancia 5.0%: Valor crítico=0.7870
  Nivel de significancia 2.5%: Valor crítico=0.9180
  Nivel de significancia 1.0%: Valor crítico=1.0920
Jarque-Bera: Estadístico=136.8839, p-valor=0.0000
```

7. VISUALIZACIÓN DE NORMALIDAD

Los gráficos Q-Q mostraron que, si bien la mayoría de los datos se alinean con la tendencia de una distribución normal teórica, existen desviaciones en los extremos. Esto confirma que las distribuciones de peso y talla no son perfectamente normales, reforzando los resultados de las pruebas estadísticas.



8. CONCLUSIONES

- Las variables categóricas como SEXO, TIPO PARTO y RÉGIMEN SEGURIDAD muestran distribuciones claras, con una ligera predominancia de nacimientos masculinos, una alta frecuencia de partos espontáneos y la mayoría de los nacimientos cubiertos por los regímenes de seguridad social contributivo y subsidiado. Las variables numéricas, como PESO y TALLA, presentan

distribuciones que se asemejan a una campana, aunque las pruebas de normalidad indican desviaciones significativas, probablemente influenciadas por el gran tamaño de la muestra y la naturaleza específica de los datos biométricos. El número de hijos nacidos vivos sugiere que es común que las madres tengan uno o dos hijos.

- El dataset presenta un número considerable de valores faltantes en algunas columnas, especialmente en 'GRUPO INDIGENA' y 'NOMBRE ADMINISTRADORA'. Si bien se realizó una limpieza inicial eliminando filas con nulos para ciertas visualizaciones, un análisis más profundo podría requerir estrategias de imputación o considerar el impacto de estos valores faltantes en análisis específicos.
- Se identificaron y eliminaron outliers en las variables numéricas clave (PESO, TALLA, NUMERO DE HIJOS NACIDOS VIVOS) utilizando el método IQR. Esta limpieza redujo el tamaño del dataset pero permitió obtener una visión más clara de la distribución central de estas variables. Es importante considerar el contexto de estos outliers; por ejemplo, pesos o tallas extremadamente bajos podrían indicar nacimientos prematuros o con complicaciones.
- A pesar de la eliminación de outliers, las pruebas estadísticas y los gráficos Q-Q indican que las variables numéricas continuas (PESO y TALLA) no siguen una distribución perfectamente normal. Esto es una observación importante para la selección de métodos estadísticos en análisis posteriores que asuman normalidad.

En general, el análisis exploratorio nos ha permitido obtener una comprensión sólida de las características principales del dataset de nacimientos, identificar áreas con datos faltantes o outliers y entender las relaciones iniciales entre algunas variables. Estos hallazgos son fundamentales para guiar análisis más profundos o la construcción de modelos predictivos.