Imagen que contiene Forma

Descripción generada automáticamente

**PROYECTO AGORA**

**ENTENDIMIENTO DE LOS DATOS –**

**ESTADÍSTICAS VITALES**

**DOCUMENTO INICIAL**

**2024**

**Control de cambios**

|  |  |  |  |
| --- | --- | --- | --- |
| **VERSIÓN** | **FECHA** | **MODIFICACIÓN** | **RESPONSABLE** |
| 1.3 | 24 de abril de 2024 | Modificación del documento | Luis Carlos Chicaíza |
| 1.5 | 07 de mayo de 2025 | Ajustes del documento con base a los comentarios recibidos | Luis Carlos Chicaíza |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Contenido

[1. Introducción 5](#_Toc165993329)

[2 Metodología 6](#_Toc165993330)

[2.1 Exploración de los datos 6](#_Toc165993331)

[3. Fuente principal de los datos 7](#_Toc165993332)

[3.1 Origen y procedencia de los datos 7](#_Toc165993333)

[3.2 Descripción general de ESTADÍSTICAS VITALES 8](#_Toc165993334)

[3.2.1 Ajustes requeridos en nombres de las fuentes 9](#_Toc165993335)

[3.2.2 Ajustes requeridos en registros de resumen en la exportación 9](#_Toc165993336)

[3.3 Estadísticas descriptivas de las variables 10](#_Toc165993337)

[3.3.1 Esquema para la exploración y análisis descriptivo de los datos 11](#_Toc165993338)

[3.3.2. Listado detallado de la fuente 12](#_Toc165993339)

[3.3.3 Análisis preliminares 22](#_Toc165993340)

[3.4 Análisis de correlación 24](#_Toc165993341)

[3.4.1 ¿Qué es el Coeficiente de Correlación? 24](#_Toc165993342)

[3.5 Visualización de datos 26](#_Toc165993343)

[3.5.1 Variable: EdadMadreMomentoDiagnostico 26](#_Toc165993344)

[3.5.2 Variable: SemanasGestacion 27](#_Toc165993345)

[3.5.3 Variable: Peso 29](#_Toc165993346)

Ilustraciones

[Imagen 1. Carpetas por años de la fuente ESTADÍSTICAS VITALES 7](#_Toc165993325)

[Imagen 2. Valores no numéricos 9](#_Toc165993326)

[Imagen 3. Datos inconsistentes variable Peso 9](#_Toc165993327)

[Imagen 4. Matriz de correlación de variables numéricas 24](#_Toc165993328)

## Introducción

El presente documento ofrece una visión detallada de los principales hallazgos derivados del proceso de exploración inicial de los datos del sistema llamado ESTADISTICAS VITALES. En esta fase inicial de análisis, se identificó la fuente de datos y se examinaron las variables contenidas en ella y sus características fundamentales.

La fuente de datos denominada ESTADÍSTICAS VITALES, comprende información recopilada durante el periodo comprendido entre 2014 a 2023. Este periodo permite analizar una variedad de contextos y condiciones que resultan ser importantes para la comprensión y el análisis de las tendencias relacionadas.

Durante la fase inicial de exploración, se realizó un análisis exhaustivo de las variables contenidas en el conjunto de datos del sistema ESTADÍSTICAS VITALES. Se aplicaron técnicas estadísticas descriptivas básicas para examinar la distribución, la variabilidad y las tendencias presentes en los registros disponibles. Este análisis preliminar sirvió como punto de partida para comprender mejor la naturaleza de los datos y así poder identificar posibles áreas de interés y enfoque para análisis posteriores.

En estas secciones, se presenta la metodología de exploración de los datos, como los principales resultados obtenidos en esta fase inicial, incluyendo las características de las variables analizadas, las tendencias y patrones identificados en los datos del sistema ESTADÍSTICAS VITALES.

## Metodología

Con el propósito de realizar un análisis exhaustivo de los datos provenientes de la fuente ESTADISTICAS VITALES proporcionada, se busca aplicar diversas técnicas de análisis exploratorio de los datos. Algunas técnicas son las siguientes:

### 2.1 Exploración de los datos

**Análisis de estadísticas descriptivas**

Calcular la media, mediana, desviación estándar y los percentiles, para así comprender la distribución y la variabilidad de las variables numéricas del conjunto de datos.

**Análisis de correlación**

Explorar la relación entre pares de variables numéricas mediante el cálculo de coeficientes de correlación, como el coeficiente de correlación de Pearson. Con el objetivo de identificar posibles asociaciones lineales o no lineales entre las variables.

**Análisis de frecuencias**

Examinar la distribución de valores únicos en variables categóricas para comprender la proporción de cada categoría y detectar posibles desequilibrios o sesgos en los datos.

**Visualización de datos**

Generar gráficos y visualizaciones, como histogramas, diagramas de dispersión y gráficos de barras, para representar gráficamente la distribución de los datos y las relaciones entre variables.

# Fuente principal de los datos

## 3.1 Origen y procedencia de los datos

A continuación, se lista los datos asociados a la base de datos recibida para el proyecto AGORA:

|  |  |
| --- | --- |
| Nombre de la fuente en el proyecto: | ESTADISTICAS VITALES |
| Número de archivos o tablas1: | 10 carpetas con la información por año |
| Nombre del archivo original de la fuente entregado2: | Vitales.zip |
| Ruta del archivo original y archivos de trabajo3: | //caoba-access.virtual. .dfs.core.windows.net/vitales/vitales\_parquet\_{año} |

Tabla 1. Descripción inicial de la fuente ESTADISTICAS VITALES

Para el procesamiento de los datos, se definió unos contendores en la infraestructura con la información por año. Esta información está consolidada desde el año 2014 hasta el año 2023.

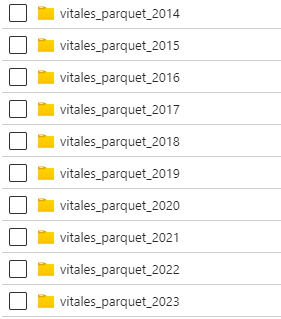


Imagen 1. Carpetas por años de la fuente ESTADÍSTICAS VITALES

El procesamiento se realizará sobre estas carpetas para poder analizar de una fecha específica o consolidar todas las fechas para un mayor análisis.

## 3.2 Descripción general de ESTADÍSTICAS VITALES

A continuación, se detalla los registros por año de los registros en la fuente:

|  |  |  |  |
| --- | --- | --- | --- |
| Base | #  de Registros | # de Columnas | ID Únicos |
| VITALES\_2014 | 658806 | 12 | 653223 |
| VITALES\_2015 | 653218 | 12 | 647581 |
| VITALES\_2016 | 639072 | 12 | 633596 |
| VITALES\_2017 | 644325 | 12 | 638906 |
| VITALES\_2018 | 618110 | 12 | 612827 |
| VITALES\_2019 | 579880 | 12 | 574982 |
| VITALES\_2020 | 556059 | 12 | 551313 |
| VITALES\_2021 | 548042 | 12 | 543367 |
| VITALES\_2022 | 519897 | 12 | 515132 |
| VITALES\_2023 | 40731 | 12 | 40369 |

Tabla 2. Descripción inicial de la fuente ESTADÍSTICAS VITALES

El total de registros es de 5’458.140 registros en la fuente.

### 3.2.1 Ajustes requeridos en nombres de las fuentes

La información recibida abarca un periodo que comprende desde 2014 hasta 2023. No fue necesario aplicar ajuste en los nombres de las variables, no se evidencio un numero o nombre distinto de las columnas a través del tiempo. También la información recibida no evidencia registros iguales o duplicados.

### 3.2.2 Ajustes requeridos en registros de resumen en la exportación

Al analizar los valores no numéricos de los datos originales de ESTADÍSTICAS VITALES se identificó que en las variables 'SemanasGestacion' y ´FechaCertificadoAAAAMM´ hay 2 datos erróneos, mientras que en la variable 'Peso' hay 64.630, porque estas variables deberían ser catalogadas como variables numéricas.

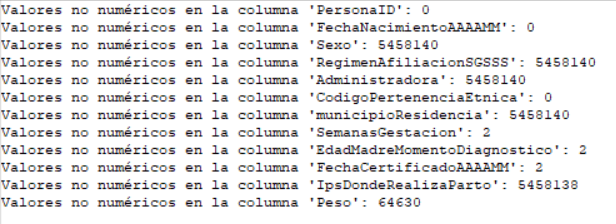


Imagen 2. Valores no numéricos

Al comprobar estos registros se valido que la información tenia estos datos atípicos, que iban a afectar directamente en el cálculo de las estadísticas básicas

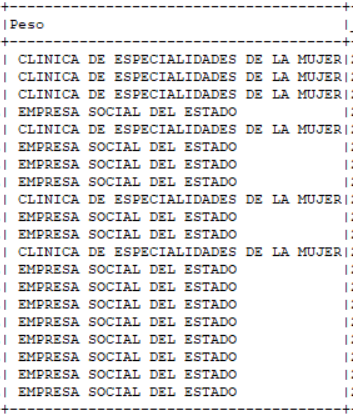


Imagen 3. Datos inconsistentes variable Peso

En total se eliminaron 64.632 registros de la muestra original, que se componen de 64.630 registros de Peso y 2 registros que compartían el mismo error en las columnas ‘SemanasGestacion' y 'EdadMadreMomentoDiagnostico', por lo cual la fuente final quedo con un total de 5.393.508 registros.

## 3.3 Estadísticas descriptivas de las variables

En esta fase se realizará el análisis exploratorio de los datos de la fuente ESTADÍSTICAS VITALES. Esta fase busca obtener una compresión más profunda de la naturaleza y las características básicas de las variables que componen este conjunto de datos.

A través de algunas técnicas de estadística y visualizaciones, se explorará la distribución, la variabilidad y las posibles relaciones entre las diferentes variables. También se buscará identificar algunos patrones, tendencias y anomalías que puedan proporcionar información importante para futuros análisis y la toma de decisiones.

Este análisis permitirá generar una contextualización adecuada de los datos, y sentar las bases para la limpieza de los datos adecuada para el entorno de la investigación y realizar análisis mucho más robustos en etapas posteriores a este proyecto

### 3.3.1 Esquema para la exploración y análisis descriptivo de los datos

A continuación, se define el siguiente esquema para realizar la exploración y el análisis descriptivo de los datos. Para cada variable de la fuente de datos ESTADÍSTICAS VITALES se debe diligenciar la siguiente tabla.

|  |  |
| --- | --- |
| **FUENTE** | **Nombre de la fuente / Nombre del archivo o tabla** |
| **Campo 1** | Nombre del campo tal como aparece en el archivo fuente |
| Descripción | Tipo de dato (fecha, texto categórico, texto no categórico, numérico nominal o numérico ordinal.) |
| Periodo | Periodo de la base de datos |
| Extensión | Formato del archivo |
| Volumen (#filas) | Cantidad de registros |
| Campos (#columnas) | Numero de atributos |
| **Campos ESTADÍSTICAS VITALES** | Descripción de todos los atributos de la base de datos |
| **Campo n** | nombre del campo en la fuente de datos |
| Alias | Nombre de la variable |
| Descripción | Descripción o propósito del atributo. |
| Tipo | Tipo de dato (fecha, texto categórico, texto no categórico, numérico nominal o numérico ordinal.) |
| Dominio | Rango de los datos o valor esperado. |
| Estadísticas básicas | Se calculan las estadísticas básicas o la frecuencia de las variables. |
| Observaciones | Información relevante de la variable a tener en cuenta |

*Tabla 3: Descripción estadística de las variables*​

### 3.3.2. Listado detallado de la fuente

Teniendo en cuenta el esquema mencionado anteriormente, se genera un cuadro detallado describiendo la fuente de datos y la descripción de cada una de las variables identificadas.

|  |  |
| --- | --- |
| **FUENTE** | **INFORMACIÓN ESTADÍSTICAS VITALES** |
| Descripción | Contiene los datos del sistema ESTADÍSTICAS VITALES |
| Período | 2014 - 2023 |
| Extensión | .parquet |
| Volumen (#filas) | 5.458.140 |
| Campos (#columnas) | 12 |
| Fecha recepción | 7 de diciembre de 2023 |
| **Campo 1** | PersonaID |
| Alias | ID de la persona |
| Descripción | Corresponde al identificador del paciente anonimizado. |
| Tipo | Categórico Nominal |
| Dominio | [ENTERO POSITIVO]  Existen 5’411.296 valores únicos. |
| Estadísticas Básicas |  |
| Observaciones | * Existe una PersonaID = 78’325.368, con 302 registros |
| **Campo 2** | FechaNacimientoAAAAMM |
| Alias | Fecha de nacimiento del paciente. |
| Descripción | Las fechas de nacimiento están registradas con el formato AAAAMM que refleja año y mes de nacimiento |
| Tipo | Categórico / Fecha |
| Dominio | [AAAAMM] |
| Estadísticas Básicas |  |
| **Campo 3** | Sexo |
| Alias | Sexo del usuario o paciente |
| Descripción | Sexo del usuario o paciente. |
| Tipo | Categórico Nominal |
| Dominio | [FEMENINO, MASCULINO, INDETERMINADO]  Existen 3 categorías únicas. |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los registros tienen valores "MASCULINO" con 2766228 y "FEMENINO" con 2626848. * Solo hay 432 registros con el valor "INDETERMINADO". |
| **Campo 4** | RegimenAfiliacionSGSSS |
| Alias | Régimen de afiliación al sistema de salud. |
| Descripción | Muestra los tipos de régimen de afiliación de los individuos al sistema de salud |
| Tipo | Categórico Nominal |
| Dominio | [S – SUBSIDIADO, C – CONTRIBUTIVO, E – EXCEPCION, NR - NO REPORTADO, P – ESPECIAL]  Existen 5 categorías únicas. |
| Estadísticas Básicas |  |
| Observaciones | * La mayoría de los registros están asociados con el régimen "S - SUBSIDIADO", con un recuento de 2.979.605. * Le sigue el régimen "C - CONTRIBUTIVO" con 2.195.023 registros. * Hay una cantidad considerable de registros con "E - EXCEPCION" con 120221 registros. * Los valores "NR - NO REPORTADO" y "P - ESPECIAL" tienen recuentos más bajos, con 96.894 y 1.765 registros respectivamente. |
| **Campo 5** | Administradora |
| Alias | Entidad de salud del individuo. |
| Descripción | Representa la entidad de salud que administra el régimen de afiliación del individuo, se conforma con el código y nombre de la administradora. |
| Tipo | Categórico Nominal |
| Dominio | [ALFANUMERICO de 6 posiciones y Nombre] |
| Estadísticas Básicas |  |
| Observaciones | • Hay varias administradoras con un número significativo de registros como "SALUD TOTAL"(309.285), "COOSALUD"(28.5301), "EPS SURA"(282189), "NUEVA EPS"(253.746). |
| **Campo 6** | CodigoPertenenciaEtnica |
| Alias | Código asociado a la pertenencia étnica del individuo. |
| Descripción | Indica la pertenencia étnica de los individuos |
| Tipo | Categórico Nominal |
| Dominio | [1:6] |
| Estadísticas Básicas |  |
| Observaciones | * La categoría "6" es la más frecuente, con 4.946.179 registros. * Las categorías "5" y "1" también tienen un número considerable de registros con 247.591 y 194.558 respectivamente. |
| **Campo 7** | municipioResidencia |
| Alias | Municipio de residencia del individuo |
| Descripción | Representa el municipio de residencia de los individuos. Los municipios están representados por un código y un nombre, lo que permite un análisis de la distribución geográfica de los nacimientos. |
| Tipo | Categórico Nominal |
| Dominio | [código y nombre municipio] |
| Estadísticas Básicas |  |
| Observaciones | * Los municipios más comunes son "Bogotá"(745.085), "Cali"(213786), "Medellín"(206.926), "Barranquilla"(192.188) y "Cartagena"(138.116), como se evidencia por los recuentos más altos. |
| **Campo 8** | SemanasGestacion |
| Alias | Semanas de gestación de la madre. |
| Descripción | Indica la duración del embarazo en semanas |
| Tipo | Numérico |
| Dominio | [15: 99] |
| Estadísticas Básicas |  |
| Observaciones | •La media de semanas de gestación es aproximadamente 38.66 semanas.  •La desviación estándar es de aproximadamente 5.11 semanas, lo que indica la variabilidad en las semanas de gestación.  •El valor mínimo es de 15 semanas.  •El valor máximo es 99 semanas, lo cual es inusual y podría indicar un error en los datos o quizás situaciones muy excepcionales. |
| **Campo 9** | EdadMadreMomentoDiagnostico |
| Alias | Edad de la madre en el momento del diagnóstico |
| Descripción | Indica la edad de la madre al momento del diagnóstico |
| Tipo | Numérico |
| Dominio | [10: 54] |
| Estadísticas Básicas |  |
| Observaciones | •La media de la edad de la madre en el momento del diagnóstico es de aproximadamente 25.82 años.  • La desviación estándar es de aproximadamente 8.33 años.  • La edad mínima es 10 años, lo cual podría ser inusual.  • La edad máxima es 54 años. |
| **Campo 10** | FechaCertificadoAAAAMM |
| Alias | Fecha del certificado. |
| Descripción | Indica el año y el mes en que se emitió el certificado de nacimiento. |
| Tipo | Categórico/ Fecha |
| Dominio | [AAAAMM] |
| Estadísticas Básicas |  |
| Observaciones | En general, las fechas que mas registros tienen estan porporcioalmente distribuidas, la fecha con mayor numero de casos se asocia a “201409” con 59.434 registros. |
| **Campo 11** | IpsDondeRealizaParto |
| Alias | IPS en donde se realizo el parto. |
| Descripción | Representa el establecimiento de salud donde se llevó a cabo el parto. |
| Tipo | Categórico Nominal |
| Dominio | [Código y nombre de IPS] |
| Estadísticas Básicas |  |
| Observaciones | • La IPS "050010464801 - CLINICA EL ROSARIO S.A." tiene el mayor número de registros con 97.269 registros.  • Una proporción significativa de registros están clasificados como "NO DEFINIDO",con 74.224 registros. |
| **Campo 12** | Peso |
| Alias | Peso del recién nacido |
| Descripción | Indica el peso del bebé al nacer |
| Tipo | Numérico |
| Dominio | [1000: 9999] |
| Estadísticas Básicas |  |
| Observaciones | • La media del peso del recién nacido es de aproximadamente 3143.120 gramos.  • La desviación estándar es de aproximadamente 747.08 gramos, lo que indica variabilidad en el peso de los recien nacidos.  • El valor mínimo de peso es de 1000 gramos.  • El valor máximo de peso es 9999 gramos, lo cual es inusual y probablemente indica la presencia de datos erróneos o faltantes. |

Tabla 4: Descripción datos fuente

### 3.3.3 Análisis preliminares

**Variables numéricas**

* La media de las semanas de gestación es de 38.66 semanas.
* El valor mínimo las semanas de gestación es de 15 semanas y el máximo es de 99 semanas, lo que podría indicar algunos valores atípicos o errores en los datos.
* La edad media de la madre en el momento del diagnóstico es de 25.83 años.
* La edad mínima de la madre es de 10 años y la máxima es de 54 años. Estos valores extremos pueden indicar la presencia de madres adolescentes o de mayor edad.
* El peso medio al nacer es de 3143.12 gramos.
* Los valores del peso mínimo y máximo son de 1000 y 9999 gramos, respectivamente., lo que puede incluir datos extremos como bebés con bajo peso al nacer con 1000 gramos o bebés con un peso superior al promedio. Estos datos pueden indicar valores atípicos o errores en los datos.

**Variables categóricas**

* En la variable “Sexo” La mayoría de los registros tienen valores "MASCULINO" con 2766228 y "FEMENINO" con 2626848 y solo hay 432 registros con el valor "INDETERMINADO".
* En la variable “RegimenAfiliacionSGSSS”la mayoría de los registros están asociados con el régimen "S - SUBSIDIADO", con un recuento de 2.979.605 (57% de los datos), le sigue el régimen "C - CONTRIBUTIVO" con 2195023 registros. (42% de los datos), también hay una cantidad considerable de registros con "E - EXCEPCION" con 120221 registros y los valores "NR - NO REPORTADO" y "P - ESPECIAL" tienen recuentos más bajos, con 96894 y 1765 registros respectivamente.
* Hay varias administradoras con un número significativo de registros como "SALUD TOTAL"(309285), "COOSALUD"(285301), "EPS SURA"(282189), "NUEVA EPS"(253746).
* En la variable "CodigoPertenenciaEtnica" la categoría "6" es la más frecuente, con 4946179 registros, las categorías "5" y "1" también tienen un número considerable de registros con 247.591 y 194.558 respectivamente y las categorías "3", "9","4" y "2" tienen recuentos mucho más bajos.
* Los municipios más comunes son "Bogotá"(745.085), "Cali"(213.786), "Medellín"(206926), "Barranquilla"(192.188) y "Cartagena"(138.116), como se evidencia por los recuentos más altos.
* Hay varias IPS con un número significativo de registros. La IPS "050010464801 - CLINICA EL ROSARIO S.A." tiene el mayor número de registros con 97269 registros.
* Una proporción significativa de registros están clasificados como "NO DEFINIDO”, con 74.224 registros.
* Hay varias IPS con un número considerable de registros, lo que indica que los partos están distribuidos entre diferentes instituciones de salud.

**Comentarios adicionales**

* No se incluye en el análisis la columna “\_\_null\_dask\_index\_\_” puesto que solo es un id único para cada registro.
* Para generar adecuadamente las estadísticas básicas de las variables numéricas identificadas se eliminaron 64.632 registros de la muestra original.
* Para una mayor profundidad de los datos se puede analizar los resultados de la exploración de los datos realizada en databricks.

## 3.4 Análisis de correlación

Se realizó una exploración de la relación entre pares de variables numéricas mediante el cálculo de matriz de correlación, como el coeficiente de correlación de Pearson. Con el objetivo de identificar posibles asociaciones lineales o no lineales entre las variables.

### 3.4.1 ¿Qué es el Coeficiente de Correlación?

El coeficiente de correlación mide la relación lineal entre dos variables numéricas. El coeficiente de correlación de Pearson, que es comúnmente utilizado y probablemente lo que se calcula en la matriz, varía entre -1 y +1. Aquí está cómo interpretar estos valores:

* +1: Una correlación perfectamente positiva. Esto significa que a medida que una variable aumenta, la otra también aumenta en una relación lineal perfecta.
* -1: Una correlación perfectamente negativa. Significa que a medida que una variable aumenta, la otra disminuye en una relación lineal perfecta.
* 0: No hay correlación lineal. Las dos variables no se influyen entre sí de manera lineal.

Cómo resultado se obtiene la matriz de correlación mostrada en la siguiente imagen:

Gráfico, Gráfico de rectángulos

Descripción generada automáticamente

Imagen 4. Matriz de correlación de variables numéricas

* Diagonal: La diagonal principal (de arriba a la izquierda a abajo a la derecha) siempre muestra 1.00, porque es la correlación de cada variable consigo misma.
* Peso y SemanasGestación: Un valor de 0.70 indica una correlación positiva entre estas dos variables. Esto sugiere que, en general, a mayores semanas de gestación, mayor es el peso del recién nacido.
* Otros Valores cercanos a cero: Estos valores indican que no hay una correlación lineal detectable entre las variables comparadas. Por ejemplo, la edad de la madre en el momento del diagnóstico no tiene correlación lineal con las semanas de gestación o el peso del recién nacido, según los datos.

## 3.5 Visualización de datos

Se generaron gráficos y visualizaciones, como histogramas, y diagramas de bigotes y gráficos de barras, para representar gráficamente la distribución de los datos. Debido al tamaño al tamaño de la fuente existen varias técnicas para hacer estas visualizaciones en Datos BigData.

### 3.5.1 Variable: EdadMadreMomentoDiagnostico

|  |  |  |
| --- | --- | --- |
| Escala lineal | * Rango de Edades: Las edades se extienden desde 0 hasta más de 100, lo cual es poco común y sugiere posibles errores de entrada de datos o valores atípicos, ya que las edades superiores a 100 son extremadamente raras. * Frecuencias: La mayoría de las edades se concentran en el rango inferior del espectro, con la frecuencia más alta en las primeras barras del histograma, que parecen incluir las edades desde el nacimiento hasta alrededor de los 50 años. * Forma de la Distribución: La distribución de las edades parece sesgada hacia la derecha, indicando que hay más registros de poblaciones más jóvenes y disminuye a medida que aumenta la edad. * Escala de Frecuencia: La escala vertical está en notación científica (por ejemplo, 1e6), lo que indica el número de registros para cada rango de edad. Por ejemplo, una barra que llegue a 2.5e6 indica 25 millones de registros para ese rango de edad. * Valores Anómalos: Las barras que se extienden más allá de la edad típica de vida humana sugieren la presencia de valores anómalos o errores en los datos. | * La mediana está cerca de la parte inferior de la caja, lo que podría indicar que la distribución de las edades es asimétrica con una cola un poco más larga hacia los valores más altos. * Hay varios valores atípicos que están muy por encima del tercer cuartil y los bigotes. Esto podría deberse a errores de entrada de datos, a una población con una cantidad de personas de edad avanzada, o a otros factores específicos del conjunto de datos. |
| Escala logarítmica | * Concentración de Datos: La mayoría de los datos se concentran alrededor del rango de 1.4 en la escala logarítmica. Esto indica que hay una edad específica (o un rango de edades) que es más común en el conjunto de datos, después de aplicar la transformación logarítmica. * En general, se concentran la mayoría de los datos en un rango específico. | * Mediana: La línea verde en el centro de la caja representa la mediana de la distribución del log\_edad. Está situada alrededor de 1.4 en la escala logarítmica, indicando que el valor de la mediana de la edad transformada cae alrededor de este punto. * Cuartiles: El borde inferior de la caja representa el primer cuartil (Q1), y el borde superior representa el tercer cuartil (Q3). El espacio entre estos dos representa el rango intercuartílico (IQR), que contiene la mitad central de los datos. La caja parece ser simétrica alrededor de la mediana, lo que sugiere cierta simetría en la distribución del log\_edad. * Valores Atípicos (Outliers): Los fuera del primer cuartil y tercer cuartil indican valores atípicos en los extremos de la distribución. |

### 3.5.2 Variable: SemanasGestacion

|  |  |  |
| --- | --- | --- |
| Escala lineal | * Rango de SemanasGestación: Las semanas de gestación se extienden desde 15 hasta 99, lo cual es poco común y sugiere posibles errores de entrada de datos o valores atípicos. * Frecuencias: La mayoría de las semanas de gestación se concentran en el rango inferior del espectro, con las frecuencias más altas en 38, en general se distribuyen los datos entre 30 y 40 semanas. * Forma de la Distribución: La distribución de las semanas de gestación indican que hay más registros cercanos a la media. * Valores Anómalos: Las barras que se extienden más allá de las semanas de gestación promedio, sugieren la presencia de valores anómalos o errores en los datos. | * Hay varios valores atípicos que están muy por encima del tercer y primer cuartil. Esto podría deberse a errores de entrada de datos o de digitación. |
| Escala logarítmica | * Concentración de Datos: La mayoría de los datos se concentran alrededor del rango de 1.6 en la escala logarítmica. Esto indica que hay una cantidad específica de semanas de gestación que es más común en el conjunto de datos. * En general, se concentran la mayoría de los datos en un rango específico. | * Mediana: La línea verde en el centro de la caja representa la mediana de la distribución del log\_SemanasGestacion. Está situada alrededor de 1.6 en la escala logarítmica, indicando que el valor de la mediana de las semanas de gestación transformada cae después de este punto. * Cuartiles: El borde inferior de la caja representa el primer cuartil (Q1), y el borde superior representa el tercer cuartil (Q3). El espacio entre estos dos representa el rango intercuartílico (IQR), que contiene la mitad central de los datos. La caja parece ser simétrica alrededor de la mediana, lo que sugiere cierta simetría en la distribución del log\_ SemanasGestacion. * Valores Atípicos (Outliers): Los datos fuera del primer cuartil y tercer cuartil indican valores atípicos en los extremos de la distribución. |

### 3.5.3 Variable: Peso

|  |  |  |
| --- | --- | --- |
| Escala lineal | * Rango de Peso: Los pesos se extienden desde 150 hasta alrededor de 9999 gramos, lo cual es poco común y sugiere posibles errores de entrada de datos o valores atípicos. * Frecuencias: La mayoría de los pesos se concentran en el rango inferior del espectro, con las frecuencias más altas en el centro del del histograma, que parecen incluir los rangos entre 2000 y 4000 gramos. * Forma de la Distribución: La distribución de los pesos indican que hay más registros cercanos a la media. * Valores Anómalos: Las barras que se extienden más allá del Peso promedio sugieren la presencia de valores anómalos o errores en los datos. | * Hay varios valores atípicos que por fuera del tercer cuartil y primer cuartil. Esto podría deberse a errores de entrada de datos, a unos recién nacidos con valores atípicos en relación con el peso promedio. |
| Escala logarítmica | * Concentración de Datos: La mayoría de los datos se concentran alrededor del rango de 3.5 en la escala logarítmica. Esto indica que hay un peso específico (o un rango de pesos) que es más común en el conjunto de datos, después de aplicar la transformación logarítmica. * En general, se concentran la mayoría de los datos en un rango específico. | * Mediana: La línea verde en el centro de la caja representa la mediana de la distribución del log\_Peso Está situada alrededor de 3.5 en la escala logarítmica, indicando que el valor de la mediana de la edad transformada cae alrededor de este punto. * Cuartiles: El borde inferior de la caja representa el primer cuartil (Q1), y el borde superior representa el tercer cuartil (Q3). El espacio entre estos dos representa el rango intercuartílico (IQR), que contiene la mitad central de los datos. La caja parece ser simétrica alrededor de la mediana, lo que sugiere cierta simetría en la distribución del log\_Peso. * Valores Atípicos (Outliers): Los datos fuera del primer cuartil y tercer cuartil indican valores atípicos en los extremos de la distribución. |