**Informe de Análisis de Datos de Nacimiento**

**Carlos Andrés Vageon Galvis**

**Introducción**

## Este informe presenta un análisis exploratorio de un conjunto de datos sobre Nacimientos Hospital General de Medellín 2021, con el objetivo de comprender las características de los nacimientos registrados, identificar posibles problemas en los datos y visualizar las distribuciones de las variables, tanto categóricas como numéricas

**Resumen del Análisis**

Las etapas del análisis incluyeron:

* Carga e inspección inicial del dataset.
* Identificación y eliminación de filas con valores faltantes.
* Análisis visual de variables categóricas y numéricas.
* Detección y eliminación de valores atípicos (outliers) en variables numéricas.
* Evaluación de la normalidad de las variables mediante pruebas estadísticas y gráficos Q-Q.

# 1. Inspección Inicial

El dataset contenía 10.038 registros y 32 columnas. Se identificaron valores faltantes en columnas como:

- PESO (Gramos)

- TALLA (Centímetros)

- TIEMPO DE GESTACIÓN

- APGAR1, APGAR2

- ÚLTIMO AÑO APROBADO MADRE

- LOCALIDAD

- EDAD PADRE

El resumen estadístico permitió obtener una descripción básica de las variables numéricas.

# 2. Valores Faltantes y Limpieza

Las columnas con mayor cantidad de valores faltantes fueron:

- EDAD PADRE

- LOCALIDAD

- ÚLTIMO AÑO APROBADO MADRE

Se utilizó el método df.dropna() para eliminar filas incompletas, reduciendo el dataset a 8.969 registros limpios.

# 3. Análisis de Variables Categóricas

Se utilizaron gráficos de barras y circulares para visualizar la distribución de las siguientes variables:

- SEXO: distribución casi equitativa.

- PARTO ATENDIDO POR: mayoritariamente por médicos.

- TIPO DE PARTO: principalmente espontáneo.

- MULTIPLICIDAD DE EMBARAZO: predominio de embarazos simples.

- GRUPO SANGUÍNEO, FACTOR RH, PERTENENCIA ÉTNICA: mayoría "Ninguno de los anteriores".

También se analizaron:

ESTADO CONYUGAL MADRE, NIVEL EDUCATIVO MADRE, ÁREA DE RESIDENCIA, RÉGIMEN DE SEGURIDAD, NIVEL EDUCATIVO PADRE, PROFESIÓN DEL CERTIFICADOR.

# 4. Análisis de Variables Numéricas y Detección de Outliers

Histogramas y diagramas de caja evidenciaron la presencia de valores atípicos en las siguientes variables:

- PESO (Gramos)

- TALLA (Centímetros)

- TIEMPO DE GESTACIÓN

- NÚMERO DE CONSULTAS PRENATALES

- APGAR1, APGAR2

- EDAD MADRE, EDAD PADRE

- ÚLTIMO AÑO APROBADO MADRE

- NÚMERO DE HIJOS NACIDOS VIVOS

- NÚMERO DE EMBARAZOS

# 5. Eliminación de Outliers y Visualización Posterior

Se aplicó el método del rango intercuartílico (IQR) para eliminar outliers, lo que redujo el dataset a 4.844 registros.

Posteriormente, los histogramas y box plots mostraron distribuciones más centradas y limpias.

# 6. Pruebas de Normalidad y Gráficos Q-Q

Se aplicaron las siguientes pruebas estadísticas para evaluar la normalidad:

- Shapiro-Wilk

- Kolmogórov-Smirnov

- Anderson-Darling

- Jarque-Bera

Los gráficos Q-Q y las pruebas anteriores indicaron que la mayoría de las variables numéricas no siguen una distribución normal.

## ****Conclusiones****

El análisis exploratorio permitió:

* Realizar una limpieza profunda del dataset.
* Comprender la estructura y distribución de las variables.
* Detectar y eliminar valores atípicos.
* Identificar la **no normalidad** en la mayoría de las variables numéricas.

Este diagnóstico inicial es clave para futuras etapas de análisis estadístico y modelado. Como pasos siguientes se recomienda:

* Explorar relaciones entre variables.
* Aplicar técnicas de transformación para mejorar la normalidad.
* Realizar análisis inferenciales o predictivos según los objetivos del estudio.

**Bibliografía**

**https://www.datos.gov.co/Salud-y-Protecci-n-Social/Nacimientos-Hospital-General-de-Medell-n-2021/w4jn-n4vq/about\_data**