**PROYECTO FINAL DE ANÁLISIS DE DATOS**

**NIVEL EXPLORADOR**

**ANÁLISIS DE DATOS SOBRE LOS NACIMIENTOS EN EL HOSPITAL GENERAL DE MEDELLÍN EN LOS AÑOS 2022 Y 2023**

**SUSANA GALINDO ROLDÁN**

**DAYANA TORRES BARRERA**

**LUIS FELIPE VANEGAS CASAS**

**2025**

*Estudiantes Bootcamp análisis de datos nivel explorador, institución universitaria tecnológico de Antioquia,*

[*felipevanegasc@gmail.com*](mailto:felipevanegasc@gmail.com)

[*dayatorresb2@gmail.com*](mailto:dayatorresb2@gmail.com)

[*susanagalindoia@gmail.com*](mailto:susanagalindoia@gmail.com)

**RESUMEN**

Este proyecto busca analizar los datos de nacimientos registrados en el Hospital General de Medellín durante los años 2022 y 2023. A través de herramientas como Python, Google Colab y bibliotecas como Pandas y Matplotlib, se hizo un análisis exploratorio para entender mejor las características de las madres y los bebés al momento del parto. Se examinaron variables como la edad de la madre, tipo de parto, nivel educativo, régimen de seguridad social, peso al nacer y tiempo de gestación. El análisis permitió identificar patrones importantes que podrían ser útiles para futuras decisiones en salud pública.

**ABSTRACT**

This project aims to analyze birth data from the General Hospital of Medellín during the years 2022 and 2023. Using tools like Python, Google Colab, and libraries such as Pandas and Matplotlib, an exploratory analysis was carried out to better understand the characteristics of mothers and newborns at the time of delivery. Variables such as mother's age, delivery type, education level, health system affiliation, birth weight, and gestation period were examined. The analysis revealed significant patterns that could support future decisions in public health.

**INTRODUCCIÓN**

El nacimiento de un niño es un evento que está influenciado por muchos factores como la edad de la madre, el acceso a servicios de salud, el nivel educativo, entre otros. En este proyecto se analiza una base de datos pública extraída del portal de datos abiertos del gobierno de Colombia, enfocada en nacimientos registrados en el Hospital General de Medellín. Este análisis pretende descubrir patrones que permitan comprender mejor las condiciones en las que nacen los bebés en este centro hospitalario y, al mismo tiempo, poner en práctica los conocimientos adquiridos durante el curso.

**OBJETIVO GENERAL:**

Consolidar las habilidades adquiridas en el bootcamp análisis de datos nivel explorador.

**OBJETIVOS ESPECÍFICOS:**

Realizar un proyecto donde se apliquen los conocimientos sobre análisis exploratorio de datos, limpieza de datos, análisis estadístico y gráficos.

Evidenciar el uso de herramientas para extracción, transformación, visualización y limpieza de datos como Google Colab, Python y sus librerías.

Realizar el análisis de los resultados y hallazgos obtenidos mediante el desarrollo del proyecto.

**DESARROLLO DEL PROYECTO**

**Fuente de datos**

La base de datos fue tomada de datos abiertos del Gobierno de Colombia y contiene más de 6.200 registros correspondientes a nacimientos ocurridos en el Hospital General de Medellín en los años 2022 y 2023. Se trabajó con 32 columnas que representan diversas variables, tanto de la madre como del recién nacido.

**Definición de columnas**

* año: Hace referencia al año al que corresponde la información generada
* periodo de reporte: Hace referencia al número del trimestre del año en el que su pública la información
* sexo: Sexo del nacido vivo
* peso gramos: Peso del nacido vivo, al nacer
* talla centimetros: Talla del nacido vivo, al nacer
* fecha\_nacimiento: Fecha del Nacimiento
* parto\_atendido\_por: Por quien fue atendido el parto
* tiempo de gestación: Tiempo de gestación del nacido vivo
* número consultas prenatales: Número de consultas prenatales que tuvo la madre del nacido vivo
* tipo\_parto: Tipo de parto en este nacimiento
* multiplicidad embarazo: Multiplicidad del embarazo
* apgar 1: Apgar del nacido vivo (al minuto). El Apgar es una prueba rápida que se realiza a los recién nacidos para evaluar su estado físico justo después del nacimiento. Se evalúan cinco características: Aspecto (color de la piel), Pulso (frecuencia cardíaca), Grimace (respuesta refleja), Actividad (tono muscular), y Respiración (esfuerzo respiratorio).
* apgar: Apgar del nacido vivo (a los 5 minutos)
* grupo sanguíneo: Hemoclasificación del nacido vivo
* factor\_rh: Factor RH
* pertenencia étnica; De acuerdo con la cultura, pueblo o rasgos físicos, el nacido vivo es reconocido por sus padres como:
* edad madre: Edad de la madre a la fecha del parto
* estado conyugal madre: Estado conyugal de la madre
* nivel educativo madre: Nivel educativo de la madre
* último \_no\_ aprobado madre: Último año de estudio que aprobó la madre
* país residencia: País de residencia de la madre
* departamento residencia: Departamento de residencia de la madre
* municipio residencia: Municipio de residencia habitual de la madre
* área residencial: Área de residencia habitual de la madre
* localidad: Barrio de residencia habitual de la madre
* número hijos nacidos vivos: Número de hijos nacidos vivos que ha tenido la madre incluido el presente
* número embarazos: Número de embarazos incluido el presente
* régimen seguridad: Entidad administradora en salud a que pertenece la madre
* nombre administradora: Entidad administradora en salud a que pertenece la madre
* edad padre: Edad del padre a la fecha de nacimiento del hijo
* nivel educativo padre: Último año de estudio que aprobó el padre
* profesión certificador: Profesión de quien certifica el nacimiento

**Proceso de limpieza y análisis**

1. Carga y exploración inicial: Se cargó la base de datos usando Pandas. Se revisó la estructura general, identificando valores nulos, columnas poco útiles y datos inconsistentes.
2. Limpieza de datos: Se eliminaron columnas irrelevantes, se normalizaron formatos y se trabajó con valores faltantes para garantizar un análisis más claro.
3. Análisis gráfico:

* Histograma de la edad de las madres.
* Gráficos de barras para tipo de parto, nivel educativo y régimen de salud.
* Gráficos de dispersión y correlación para tiempo de gestación y peso del bebé.
* Mapa de calor para visualizar correlaciones entre variables numéricas.

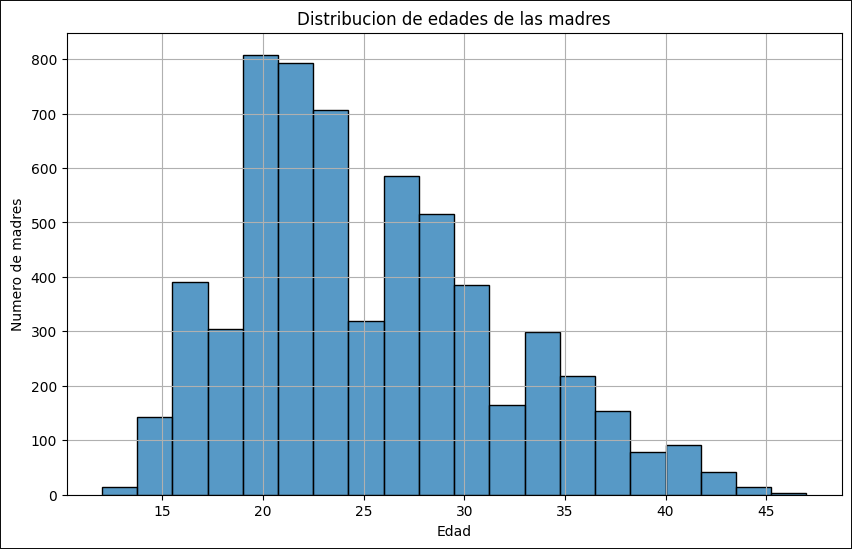
**Hallazgos:**

Durante el análisis de los datos de nacimientos en el Hospital General de Medellín para los años 2022 y 2023, se identificaron varios patrones relevantes que reflejan tanto las condiciones sociodemográficas de las madres como factores relacionados con el estado de salud de los recién nacidos. A continuación, se presentan los hallazgos más destacados:

* Edad de las madres: La mayoría de los nacimientos se concentra en mujeres jóvenes, especialmente entre los 20 y 30 años, con un pico notable entre los 20 y 22 años. Este dato sugiere que la maternidad se presenta comúnmente en la etapa temprana de la adultez. También se identificaron casos de maternidad adolescente y de madres mayores de 35 años, aunque en menor proporción.
* Tipo de parto: El parto natural o vaginal es el más frecuente, lo que puede estar relacionado con buenas prácticas en salud materna. Sin embargo, se observó que el número de cesáreas también es considerable, sobre todo en mujeres mayores de 30 años, lo cual podría deberse a recomendaciones médicas según el perfil de riesgo o a decisiones clínicas relacionadas con la edad y condiciones de salud.
* Nivel educativo de la madre: Predomina el nivel educativo medio (bachillerato completo), seguido por la educación técnica o tecnológica. Se encontró una menor proporción de madres con educación universitaria, lo que podría asociarse con una postergación de la maternidad en mujeres con mayor escolaridad. Esto también refleja un contexto donde muchas mujeres deciden ser madres antes de ingresar o culminar estudios superiores.
* Régimen de afiliación en salud: La mayoría de las madres están afiliadas al régimen subsidiado, lo que evidencia un alto uso del sistema de salud público. Esta situación puede estar relacionada con condiciones socioeconómicas que limitan el acceso al régimen contributivo, indicando una población mayoritariamente de estratos bajos o medios.
* Peso al nacer: Se identificó que la gran mayoría de los bebés nacen con un peso dentro del rango normal (entre 2.500 y 3.999 gramos). Sin embargo, también se observaron casos de bajo peso al nacer (menos de 2.500 gramos), los cuales podrían estar relacionados con nacimientos prematuros, bajo nivel de control prenatal o condiciones maternas de salud.
* Relación entre tiempo de gestación y peso del bebé: Se encontró una correlación positiva entre estas dos variables; es decir, a mayor tiempo de gestación, mayor es el peso del bebé al nacer. Este hallazgo es consistente con el desarrollo natural del embarazo, donde los bebés ganan más peso en las últimas semanas de gestación.
* Edad de la madre y tipo de parto: Se evidenció que las madres más jóvenes tienden a tener partos naturales, mientras que en las mayores de 30 años aumentan los casos de cesáreas. Esto puede estar relacionado con mayores precauciones médicas y con condiciones físicas que cambian con la edad.
* Distribución de nacimientos por mes: Al observar la distribución mensual, se notó una relativa estabilidad en el número de nacimientos a lo largo del año, aunque con ligeras variaciones. Esto podría indicar que no hay una estacionalidad marcada, aunque sería interesante contrastarlo con más años para confirmar una tendencia.
* Datos atípicos y vacíos: Durante la limpieza de los datos se encontraron algunas inconsistencias, como registros con valores nulos o categorías repetidas. Esto resalta la importancia de contar con procesos adecuados de recolección y registro de información en las instituciones de salud.
* Relación entre variables numéricas: El análisis de correlaciones mostró que, además del peso y la gestación, otras variables como la edad de la madre tienen una relación débil con las demás, lo cual sugiere que cada factor aporta de forma distinta al perfil general de los nacimientos.

**Analisis y graficos**

Histograma de distribución de edad de las madres

****

* **Análisis:** La mayoría de los nacimientos se concentra en mujeres entre los **20 y 30 años**, con un pico especialmente entre **20 y 22 años**. Hay una curva a la derecha, lo que indica menos nacimientos en edades mayores.

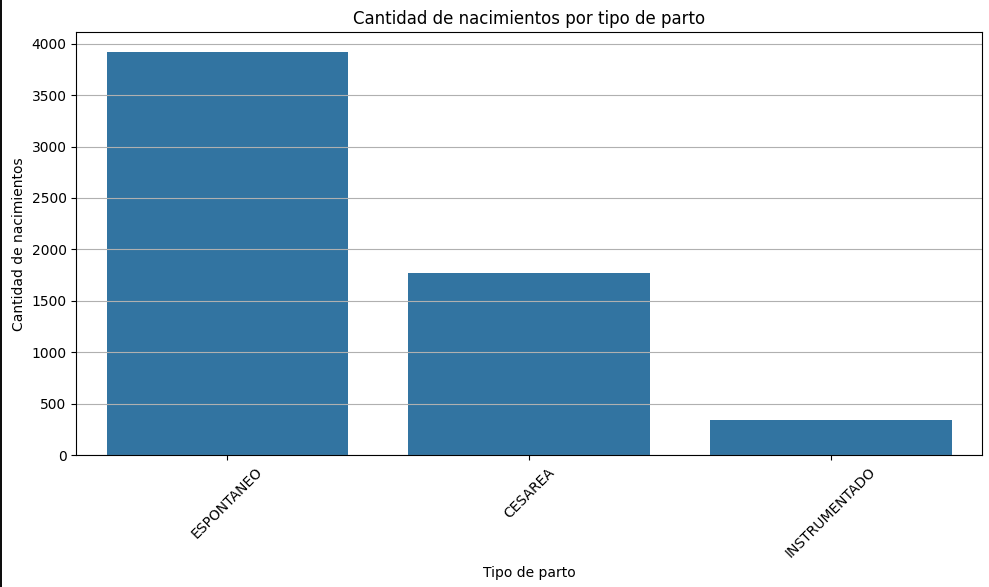
**🎓** Nivel educativo de la Madre

Gráfico

El contenido generado por IA puede ser incorrecto.

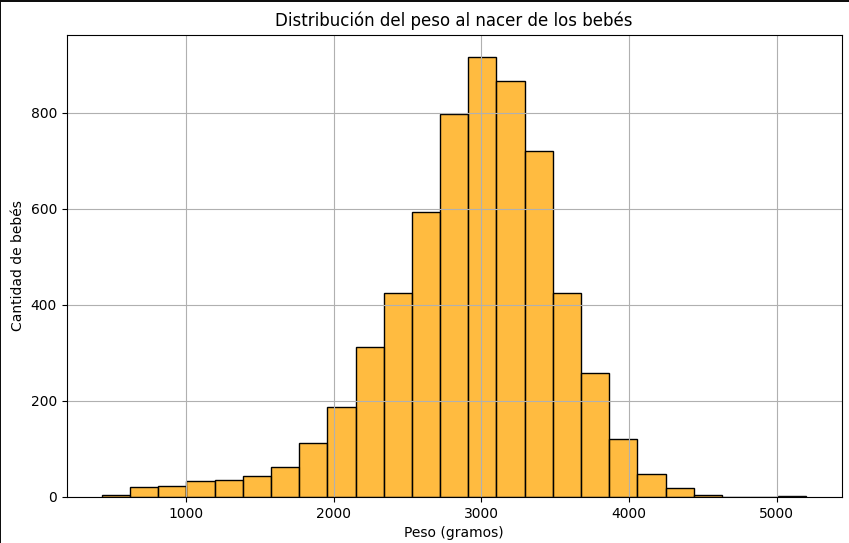
**Análisis:** Se observa que el mayor número de madres tiene **educación media** (bachillerato), seguido por **educación técnica o tecnológica**. La proporción con educación universitaria es menor, la decisión de maternidad en mujeres es tardía con más años de estudio.

**🏥** Cantidad de nacimientos por tipo parto



**Análisis:** Predomina el **parto natural o vaginal**, aunque hay una proporción considerable de **cesáreas**. Esto puede sugerir un acceso regular a servicios de salud obstétrica.

Distribución de peso al nacer de los bebés



**Análisis:** La mayoría de bebés nacen con un peso dentro del rango normal (2.500 – 3.999 g). Hay un pequeño grupo de bebés con **bajo peso al nacer** (<2.500 g), lo que puede estar relacionado con embarazos prematuros o factores de salud materna.

Edad de las madres según el tipo de parto

Gráfico, Gráfico de cajas y bigotes

El contenido generado por IA puede ser incorrecto.

**Análisis:** Las **cesáreas** son más comunes en madres mayores de 26 años, mientras que el **parto vaginal** predomina en mujeres jóvenes. Esto concuerda con prácticas médicas donde la edad es un factor para decidir el tipo de parto.

Nacimientos por régimen de seguridad social

Gráfico, Gráfico en cascada

El contenido generado por IA puede ser incorrecto.

**Análisis:** La mayoría de los nacimientos se dan en madres con régimen **subsidiado**, seguido del **contributivo**. Esto indica una población con alto uso del sistema de salud pública, posiblemente de estratos bajos o medios

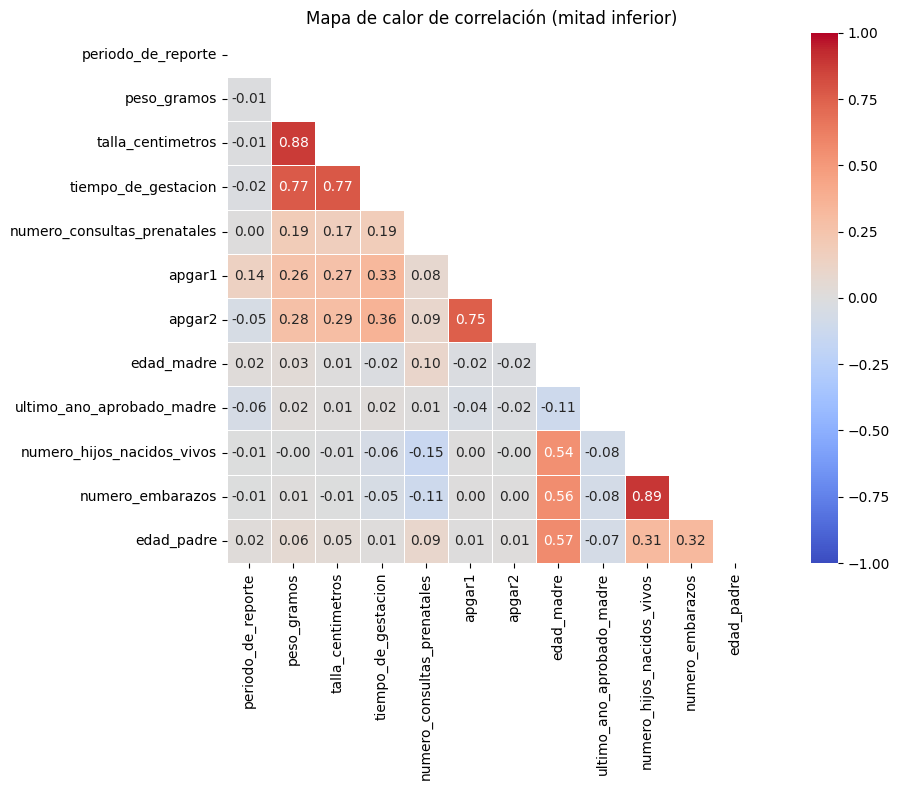
Relación entre el tiempo de gestación y peso del bebé

Gráfico

El contenido generado por IA puede ser incorrecto.

**Análisis:** Hay una **correlación positiva**: a mayor tiempo de gestación, mayor peso del bebé. Esto es esperado, ya que el crecimiento fetal se intensifica en las últimas semanas.(Desarrollo del bebé)

Mapa de calor de correlación entre variables numéricas



**Análisis:** El **peso del bebé** y el **tiempo de gestación** son las variables con **mayor correlación positiva**. La **edad de la madre** tiene una **correlación baja** con las demás variables numéricas,es decir,que una mejor atención prenatal se relaciona con mejores resultados neonatales.

**CONCLUSIONES**

Este proyecto permitió poner en práctica los conocimientos adquiridos durante el Bootcamp de análisis de datos a través del estudio de una base de datos real sobre nacimientos en el Hospital General de Medellín. Al trabajar con los datos, se pudo evidenciar la importancia de realizar procesos de limpieza, exploración y visualización para obtener información valiosa.

Una de las principales conclusiones es que la mayoría de los nacimientos se da en mujeres jóvenes entre los 20 y 30 años, en su mayoría con nivel educativo medio y afiliadas al régimen subsidiado. Estos datos reflejan un contexto social donde muchas madres aún tienen acceso limitado a educación superior y servicios privados de salud. Además, se observó que las cesáreas son más frecuentes en mujeres mayores de 30 años, lo que puede estar relacionado con factores médicos y decisiones clínicas.

El análisis también permitió comprobar que existe una relación positiva entre el tiempo de gestación y el peso del bebé, lo cual es coherente con el desarrollo normal del embarazo. A través de gráficos y estadísticas se logró interpretar estos patrones de forma visual y comprensible, mostrando cómo el análisis exploratorio puede ayudar a entender mejor situaciones reales.

En general, este proyecto no solo cumplió con el objetivo de aplicar herramientas como Python, Pandas y Matplotlib, sino que también ayudó a desarrollar habilidades para interpretar datos, comunicar hallazgos de manera clara y comprender cómo el análisis puede apoyar la toma de decisiones en temas de salud pública a futuro.

Para los datos donde estaban relacionados la ubicación geográfica de la gestantes no se relacionaron los datos ya que en la columna localidad se encontraron datos de barrio ,municipio ,centro,vereda lo que no es congruente para sacar datos y analizarlos de una manera clara y concluyente.