

Parcial II Procesamiento de datos a Gran Escala

- JUAN DAVID LOPEZ BECERRA
- NICOLÁS SAMUEL MARTIN VASQUEZ
- JUAN DIEGO GONZÁLEZ JIMENEZ



Un pequeño contexto



El dataset es de 2021 del
Hospital San Juan

Posee 2k de datos

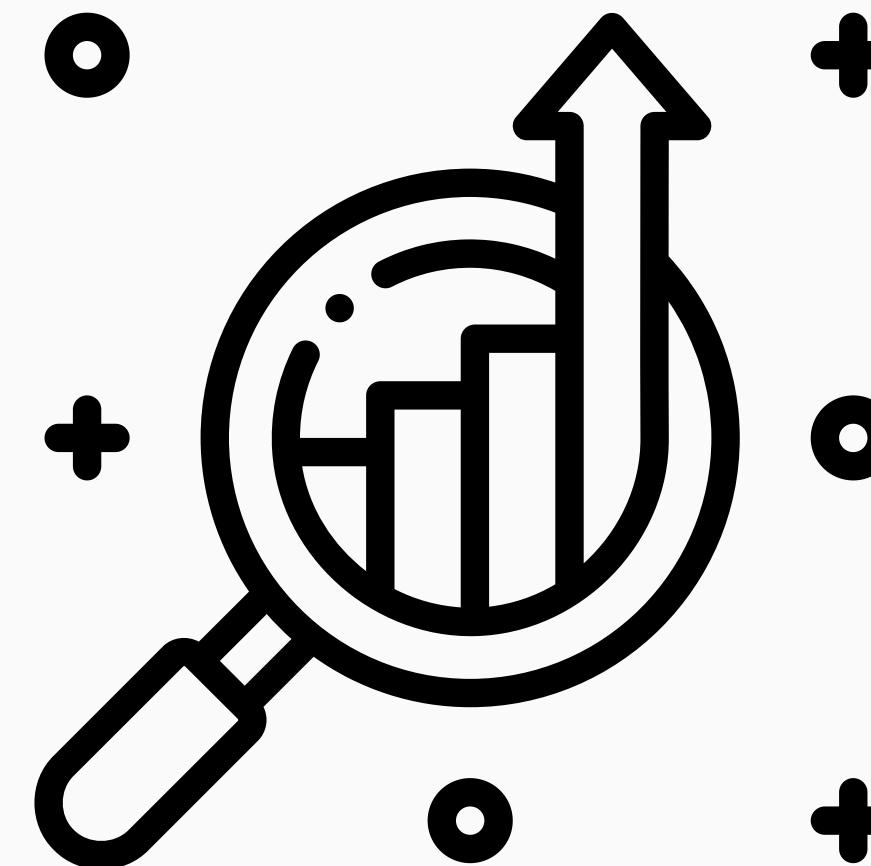


El dataset es de enero
2016 a febrero 2023

Posee 67k de datos



Exploración de los Datos



**Presentaremos tablas y
gráficos para entender
mejor los datos**



DataFrame Río Negro



Tipos de datos en los Atributos

#	Column	Non-Null Count	Dtype
0	DEPARTAMENTO	2074 non-null	object
1	MUNICIPIO	2074 non-null	object
2	AREA NACIMIENTO	2074 non-null	object
3	SEXO	2074 non-null	object
4	PESO (Gramos)	2074 non-null	int64
5	TALLA (Centímetros)	2074 non-null	int64
6	FECHA NACIMIENTO	2074 non-null	object
7	TIEMPO DE GESTACIÓN	2074 non-null	int64
8	NÚMERO CONSULTAS PRENATALES	2074 non-null	int64
9	TIPO PARTO	2074 non-null	object
10	MULTIPLICIDAD EMBARAZO	2074 non-null	object
11	EDAD MADRE	2074 non-null	object
12	EDAD PADRE	2074 non-null	object

dtypes: int64(4), object(9)

Indicamos que tipo de dato tenía establecido cada atributo

- **Object** indica que contiene texto o tipos mixtos (Números y texto)
- Nos indica que casi no hay nulos



Análisis Estadístico Atributos Cuantitativas

Estadísticas descriptivas:

	PESO (Gramos)	TALLA (Centímetros)	TIEMPO DE GESTACIÓN
count	2074.000000	2074.000000	2074.000000
mean	3121.619576	48.623915	38.724204
std	383.912788	1.954703	1.174222
min	570.000000	25.000000	25.000000
25%	2870.000000	47.000000	38.000000
50%	3105.000000	49.000000	39.000000
75%	3367.500000	50.000000	40.000000
max	4940.000000	55.000000	43.000000

NÚMERO CONSULTAS PRENATALES

count	2074.000000
mean	6.609450
std	2.256326
min	0.000000
25%	5.000000
50%	7.000000
75%	8.000000
max	19.000000

Usamos las cuatro variables numéricas que hay

- Diferencia considerable entre **peso mínimo y máximo**
- La *desviación estándar* esta **muy baja** en todos **menos en el peso**



Tabla de contingencia

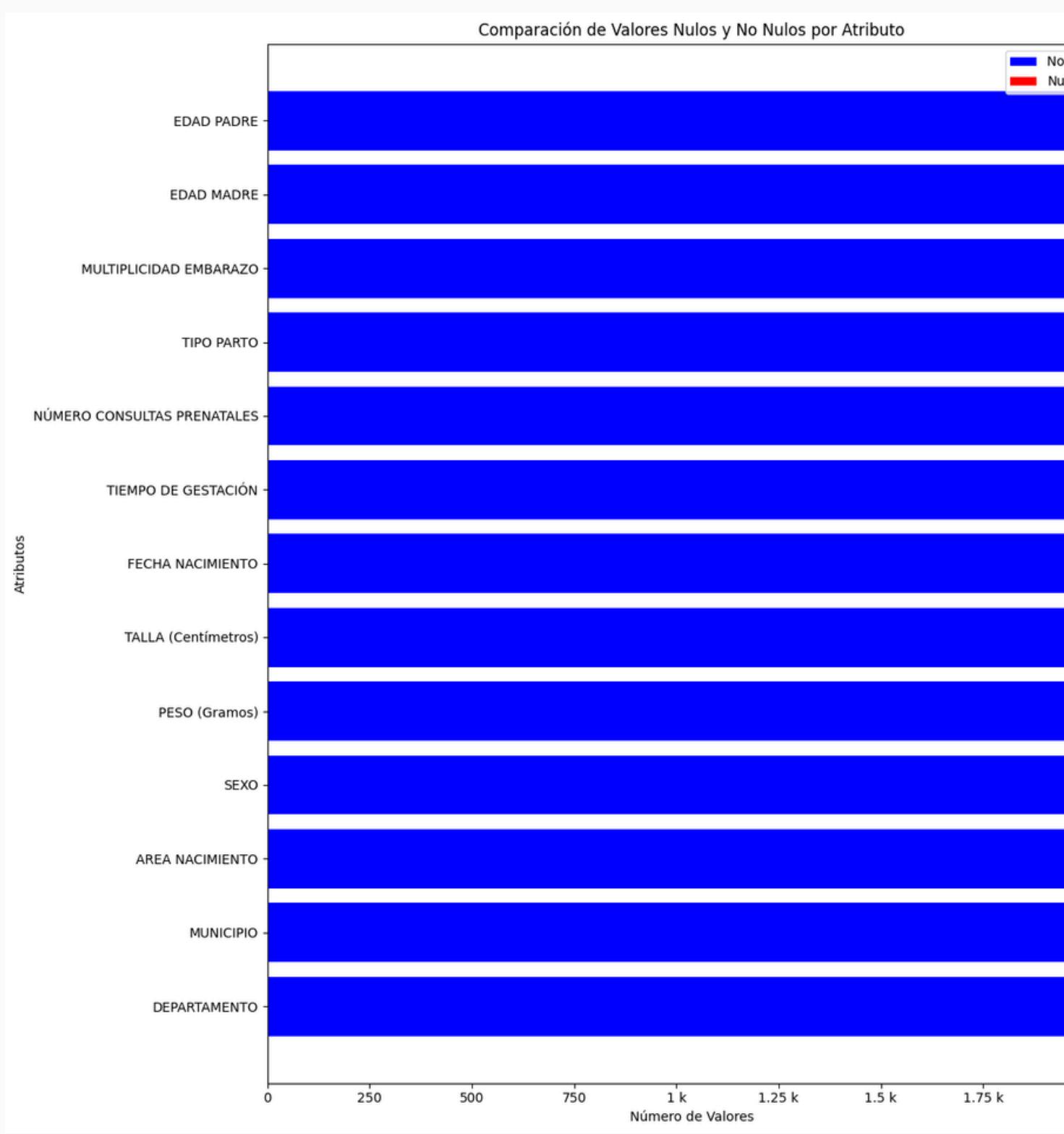
AREA NACIMIENTO	CABECERA MUNICIPAL	RURAL	DISPERSO
TIPO PARTO			
CESÁREA	760	3	
ESPONTÁNEO	1283	5	
INSTRUMENTADO	23	0	

Se realizada un análisis de contingencia entre Area de nacimiento y tipo de parto

- En la cabecera municipal, se registraron **más partos** espontáneos que cesáreas.
- En el área rural dispersa, los números son mucho más bajos.
- La mayoría de los partos son espontáneos en ambas áreas



Nulos vs No nulos en los atributos

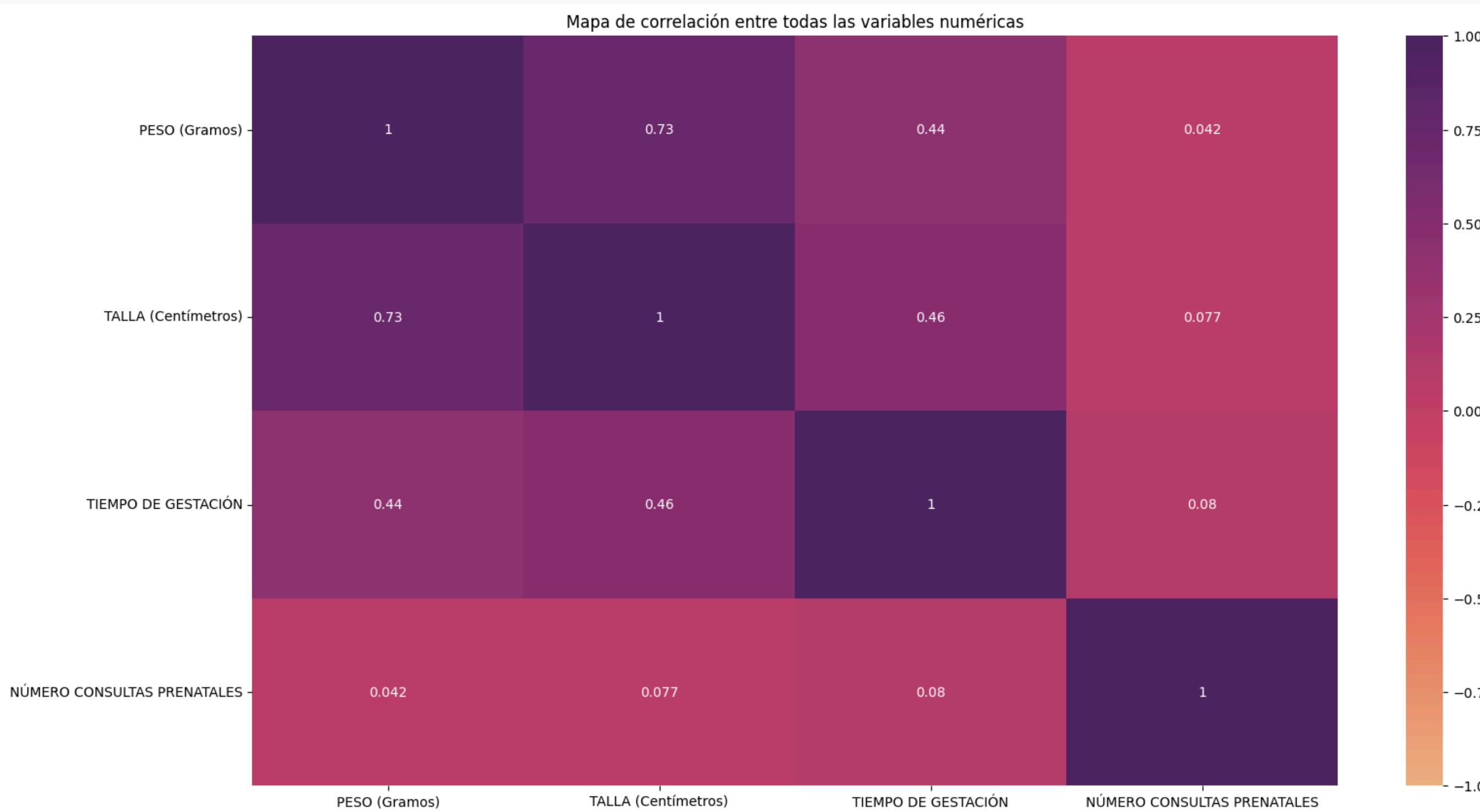


Se compararon los valores nulos y los nulos en los atributos para analizar la utilidad de algunos atributos.

- *Dataset **MUY** bueno.*
- *Pocos valores nulos, casi inexistentes*



Mapa de correlación entre todas las variables numéricas

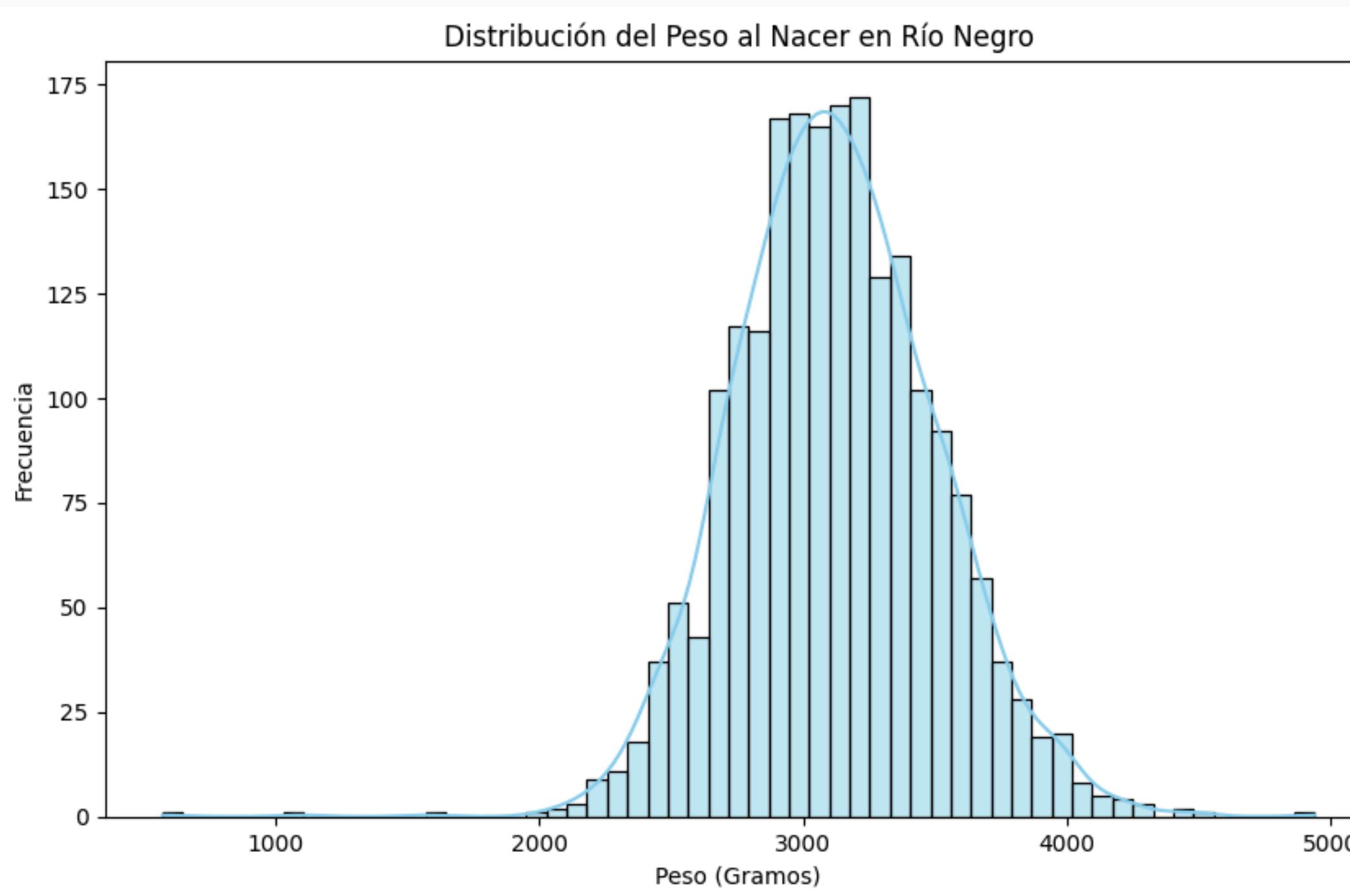


Se realiza una mapa de correlación entre las variables numéricas del dataset

- La correlación más intensa es la de Talla y Peso



Grafico Distribución del Peso al Nacer en Río Negro

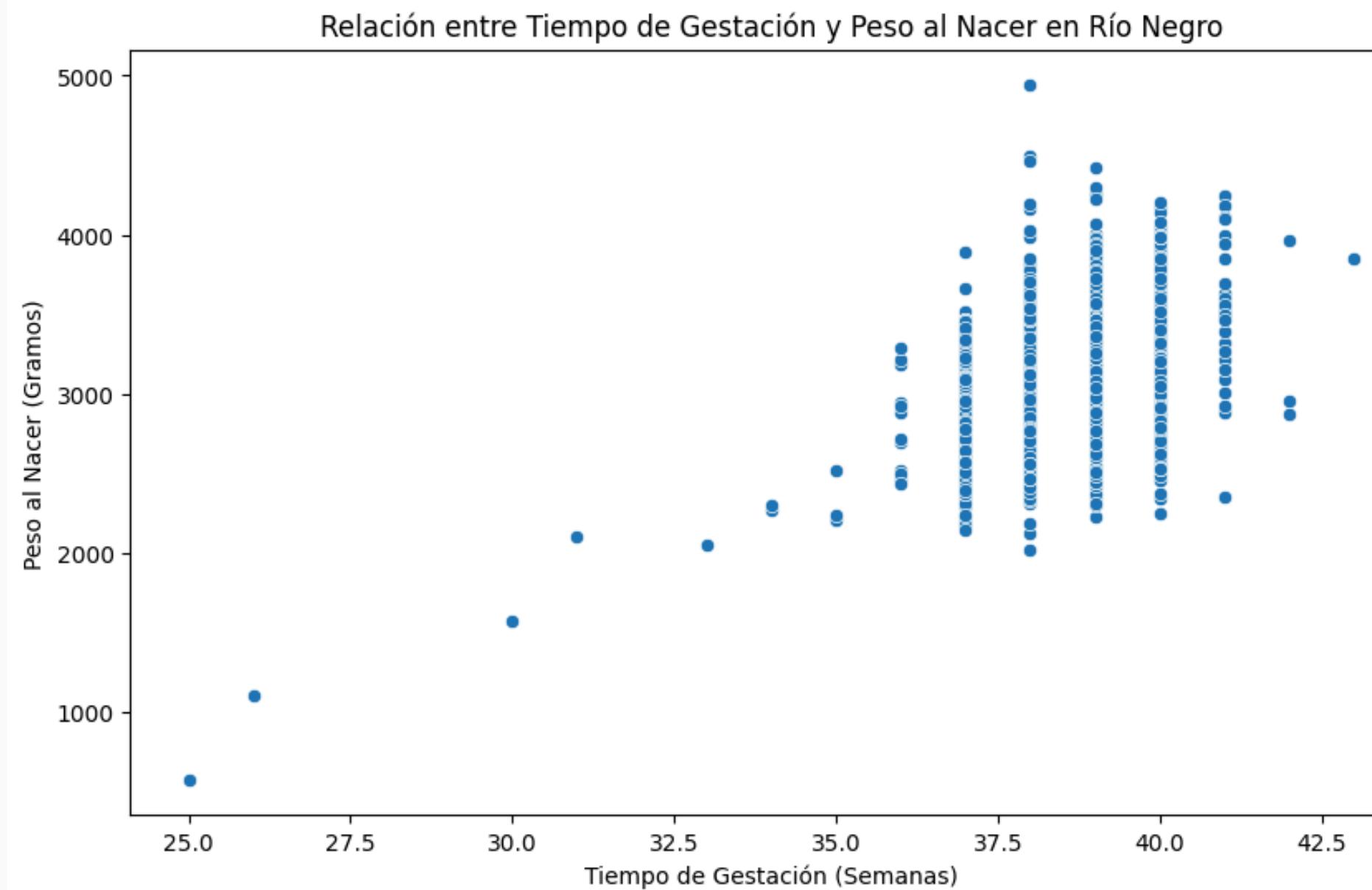


Se realiza una grafica de barras con curva para observar el comportamiento del peso de los niños al nacer:

- El **pico** esta como en los **3.100gr**
- Vemos valores interesantes como **menos de 100gr** y **casi 5.00gr**



Grafico de dispersión entre Peso al Nacer y Tiempo de Gestación

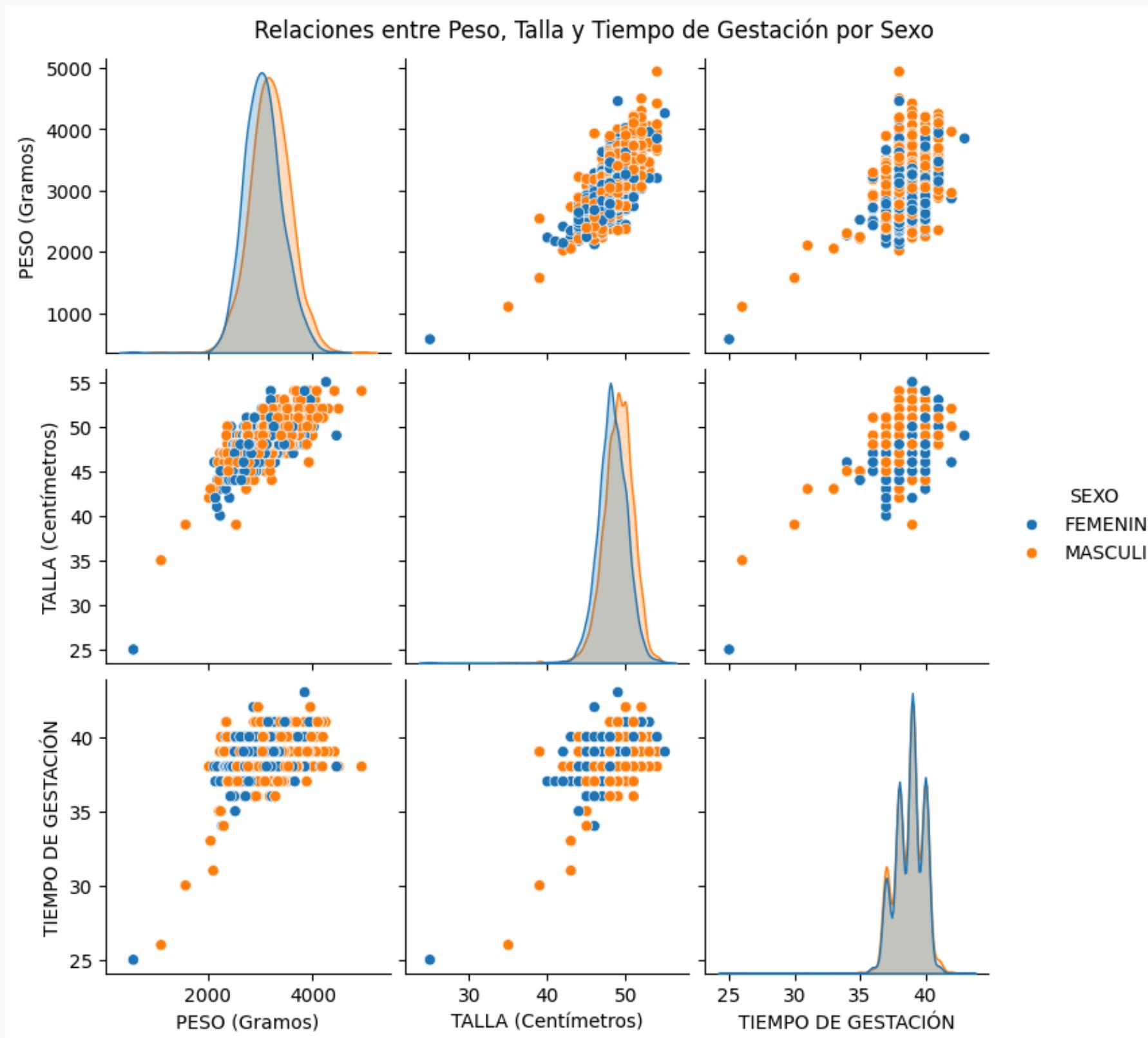


Quisimos comprar la dispersión de los datos entre Peso y tiempo de gestión

- Se observa que a menor tiempo de gestación los bebés pesan menos
- La mayoría de los bebés tienen un tiempo superior a 36 semanas y mas de 2.500gr



Pair plot con Diferenciación por Sexo

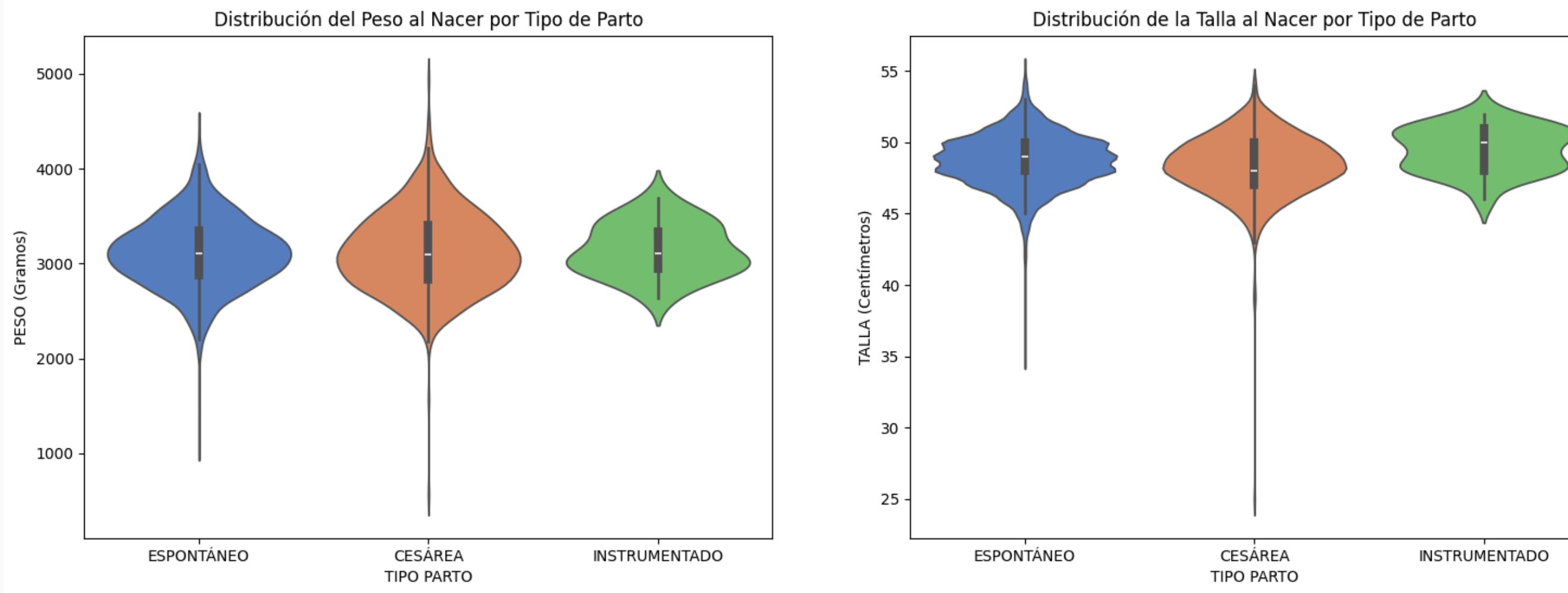


Se hizo un pair plot para comparar el sexo en los atributos de peso, talla y tiempo de gestación

- Vemos que la comparación por sexo es bastante similar
- En las graficas de dispersión se observan datos bastante atípicos
- En las gráficas de curvas casi no se presentan diferencias



Gráfico de Violín para Comparar Distribuciones



Hicimos un grafico de violín entre el tipo de parte y las variables de talla y peso

- Lo mas ancho es donde se concentran la mayoría de datos
- La líneas son la variación de los datos



DataFrame Bucaramanga



Tipos de datos en los Atributos

#	Column	Non-Null Count	Dtype
0	DEPARTAMENTO	65577	non-null object
1	MUNICIPIO	65577	non-null object
2	AREA NACIMIENTO	0	non-null float64
3	SITIO NACIMIENTO	65575	non-null object
4	CODIGO INSTITUCION	65549	non-null float64
5	SEXO	0	non-null float64
6	PESO (Gramos)	65570	non-null float64
7	TALLA (Centímetros)	65570	non-null float64
8	HORA NACIMIENTO	65574	non-null object
9	PARTO ATENDIDO POR	65575	non-null object
10	TIEMPO DE GESTACIÓN (Semanas)	65567	non-null float64
11	NUMERO CONSULTAS PRENATALES	65567	non-null float64
12	TIPO PARTO	65575	non-null object
13	MULTIPLICIDAD EMBARAZO	65575	non-null object
14	APGAR1	65504	non-null float64
15	APGAR2	65504	non-null float64
16	GRUPO SANGUINEO	65430	non-null object
17	FACTOR RH	0	non-null float64
18	PERTENENCIA ÉTNICA	65575	non-null object
19	GRUPO INDIGENA	20	non-null object
20	EDAD MADRE	65349	non-null float64
21	ESTADO CONYUGAL MADRE	65349	non-null object
22	NIVEL EDUCATIVO MADRE	65191	non-null object

23	ULTIMO AÑO APROBADO MADRE	63970	non-null	float64
24	PAIS RESIDENCIA	65191	non-null	object
25	departamento_residencia	65191	non-null	object
26	MUNICIPIO RESIDENCIA	65191	non-null	object
27	AREA RESIDENCIA	0	non-null	float64
28	LOCALIDAD	22271	non-null	object
29	BARRIO	63183	non-null	object
30	CENTRO POBLADO	0	non-null	float64
31	RURAL DISPERSO	1736	non-null	object
32	NÚMERO HIJOS NACIDOS VIVOS	64964	non-null	float64
33	FECHA ANTERIOR HIJO NACIDO VIVO	34296	non-null	object
34	NÚMERO EMBARAZOS	64964	non-null	float64
35	RÉGIMEN SEGURIDAD	64964	non-null	object
36	NOMBRE ADMINISTRADORA	60651	non-null	object
37	EDAD PADRE	64780	non-null	float64
38	NIVEL EDUCATIVO PADRE	64955	non-null	object
39	ULTIMO AÑO APROBADO PADRE	60729	non-null	float64
40	COMUNA	64964	non-null	object
41	NOMCOMUNA	47908	non-null	object
42	NUM NOMBCOMUNA	47951	non-null	object
43	GRUPO EDAD MADRE	64964	non-null	object
44	CURSO DE VIDA MADRE	64964	non-null	object
45	GRUPO EDAD PADRE	64780	non-null	object
46	CURSO DE VIDA PADRE	64780	non-null	object
47	AÑO	64964	non-null	float64
48	BARRIO_VER	64964	non-null	object
49	ciudad_geo	46594	non-null	object
50	ORDEN	57904	non-null	float64
51	DIA SEMANA	64964	non-null	object
52	MES	64964	non-null	object

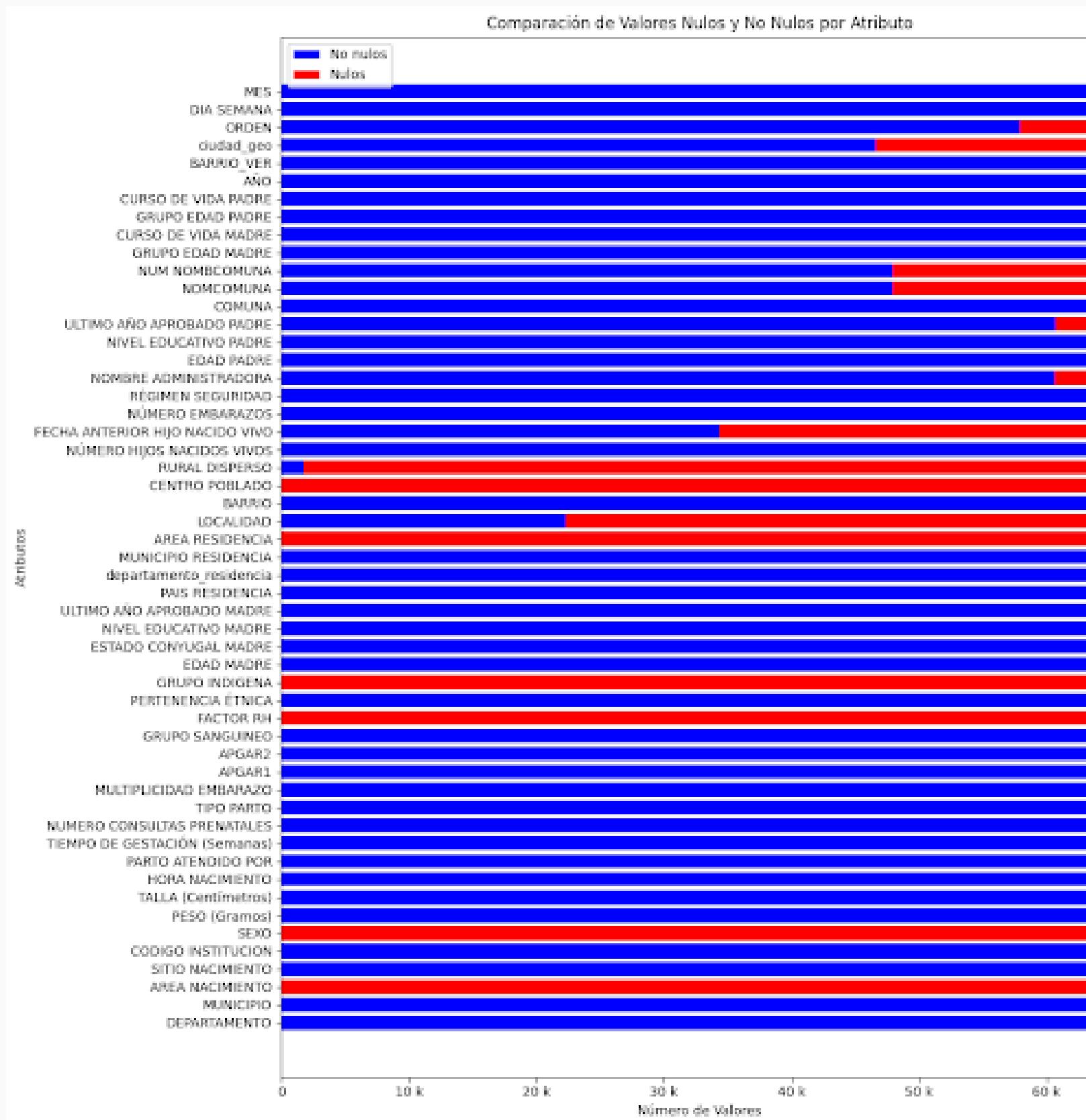
dtypes: float64(20), object(33)

Indicamos que tipo de dato tenia establecido cada atributo

- Vemos que hay muchos atributos con 0 no nulos
- Vemos que es un dataset mas complejo que el anterior
- Solo hay tipo float y object



Nulos vs No nulos en los atributos

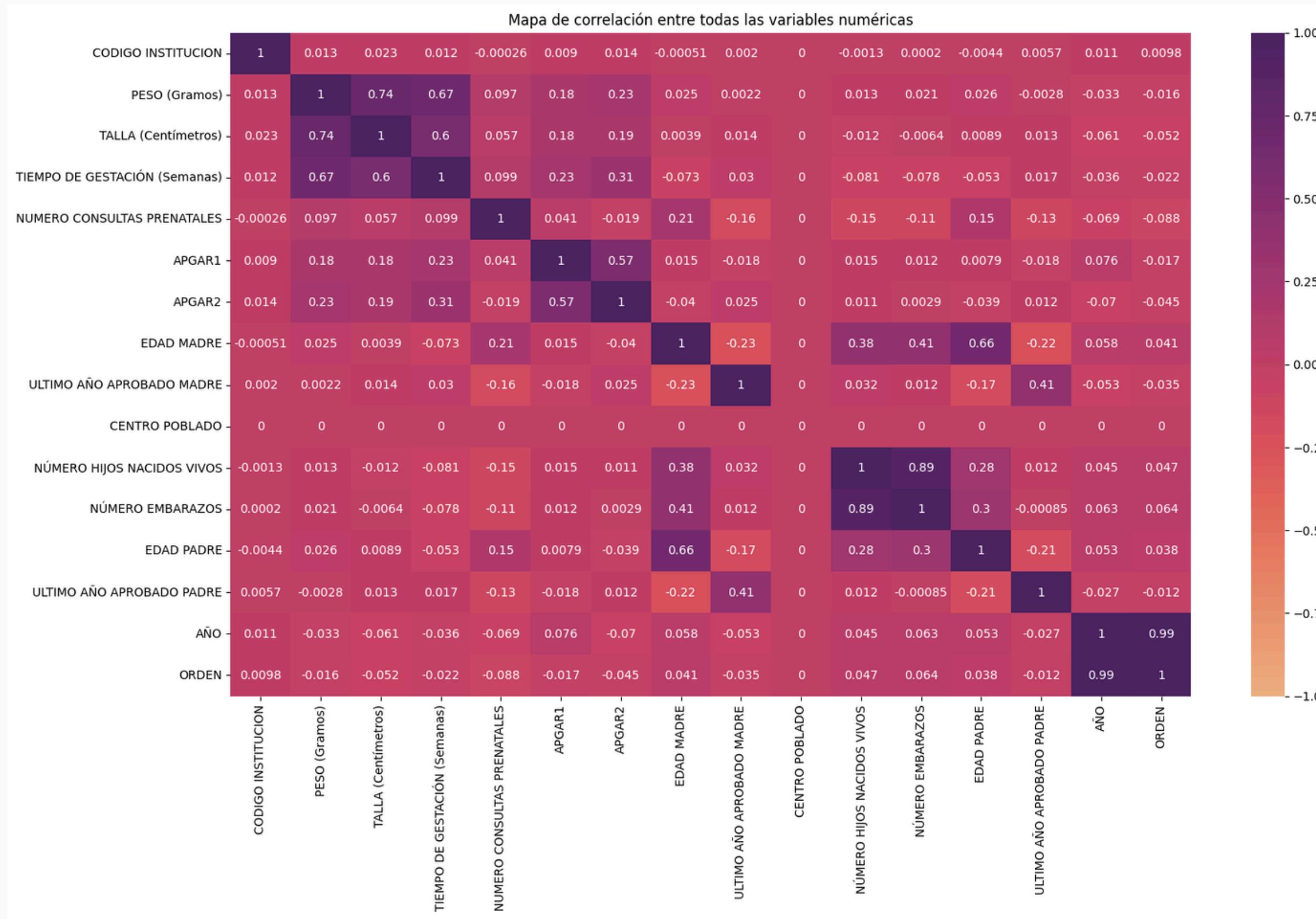


Se compararon los valores nulos y los nulos en los atributos para analizar la utilidad de algunos atributos.

- Vemos que hay atributos prácticamente inútiles
- Hay varios que tiene una cantidad considerable de nulos
- Hay atributos sin nulos



Mapa de correlación entre todas las variables numéricas

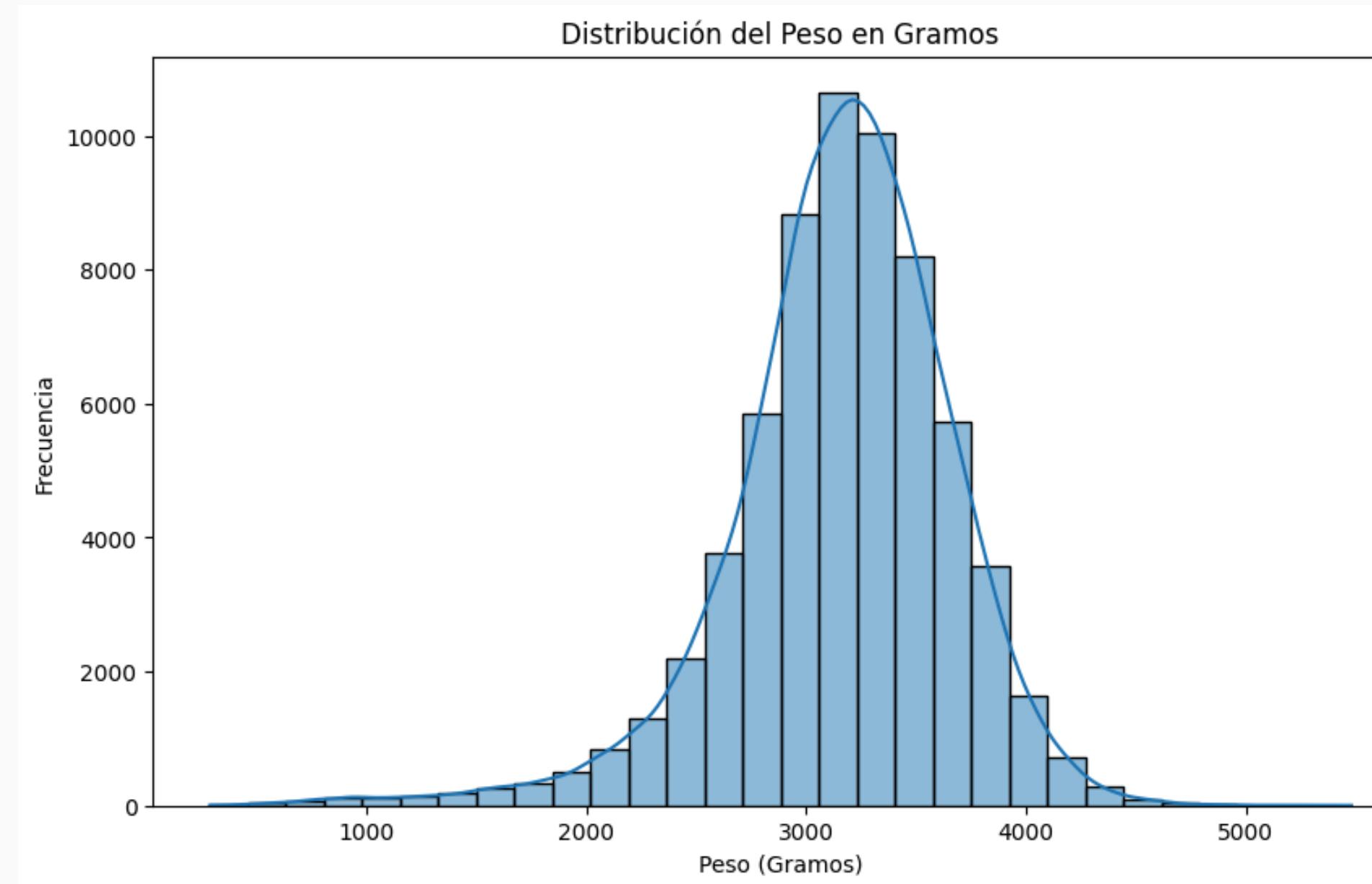


Se realiza una mapa de correlación entre las variables numéricas del dataset

- La mayoría de la correlaciones muy neutras
- Hay pocas correlaciones fuertes



Grafico Distribución del Peso al Nacer en Bucaramanga

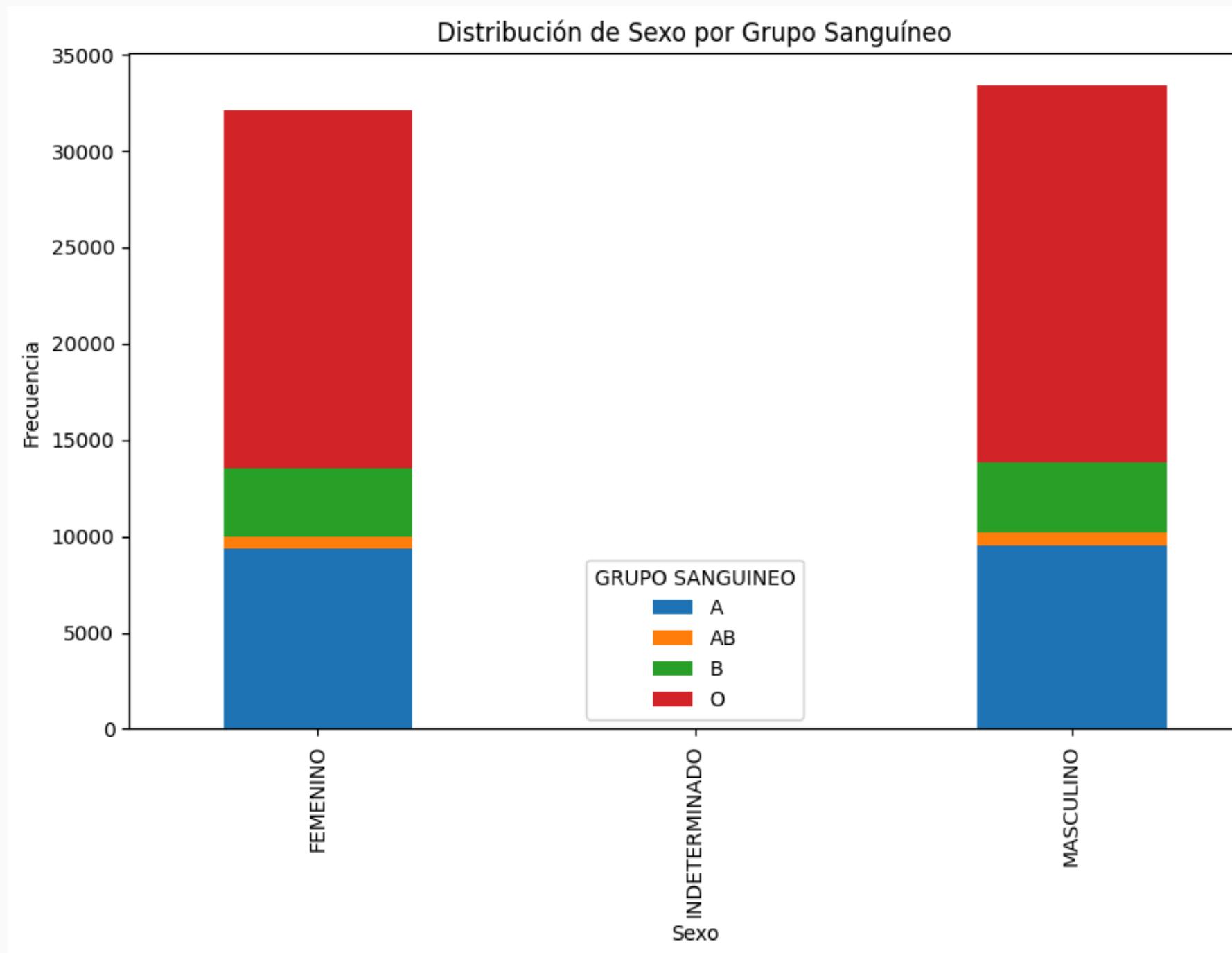


Se realiza una grafica de barras con curva para observar el comportamiento del peso de los niños al nacer:

- El **pico** esta como en los **3.100gr, pico similar al de Río Negro**
- Vemos valores constantes desde menos de 1.000gr hasta 5.000gr



Gráfica de barras apilada de Sexo por Grupo Sanguíneo

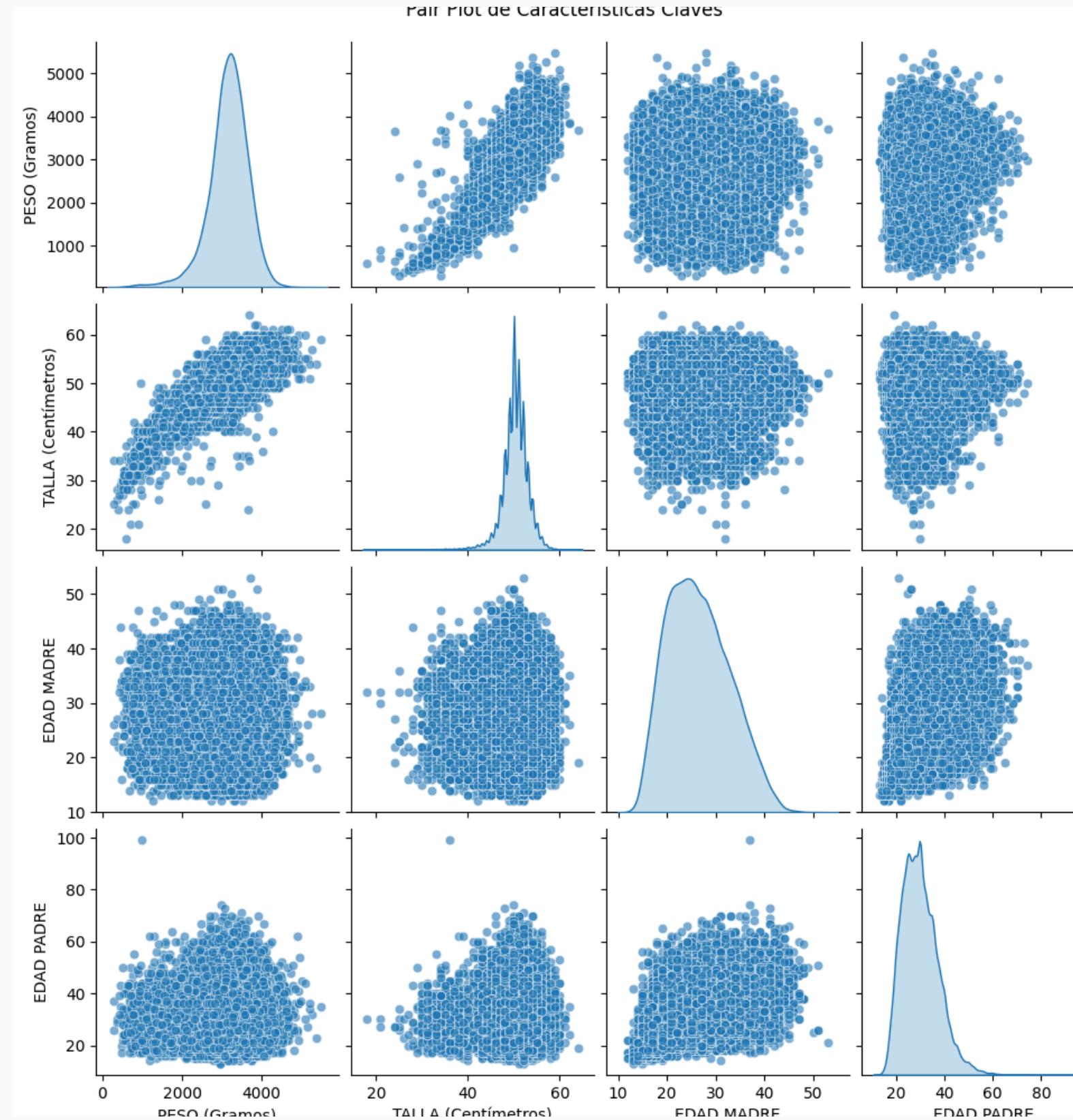


Hicimos un gráfico de barras apiladas para comparar el seco según grupo sanguíneo

- El grupo sanguíneo más común es el O tanto en masculino y femenino
- Los AB son muy pocos



Pair plot de algunos atributos relacionados

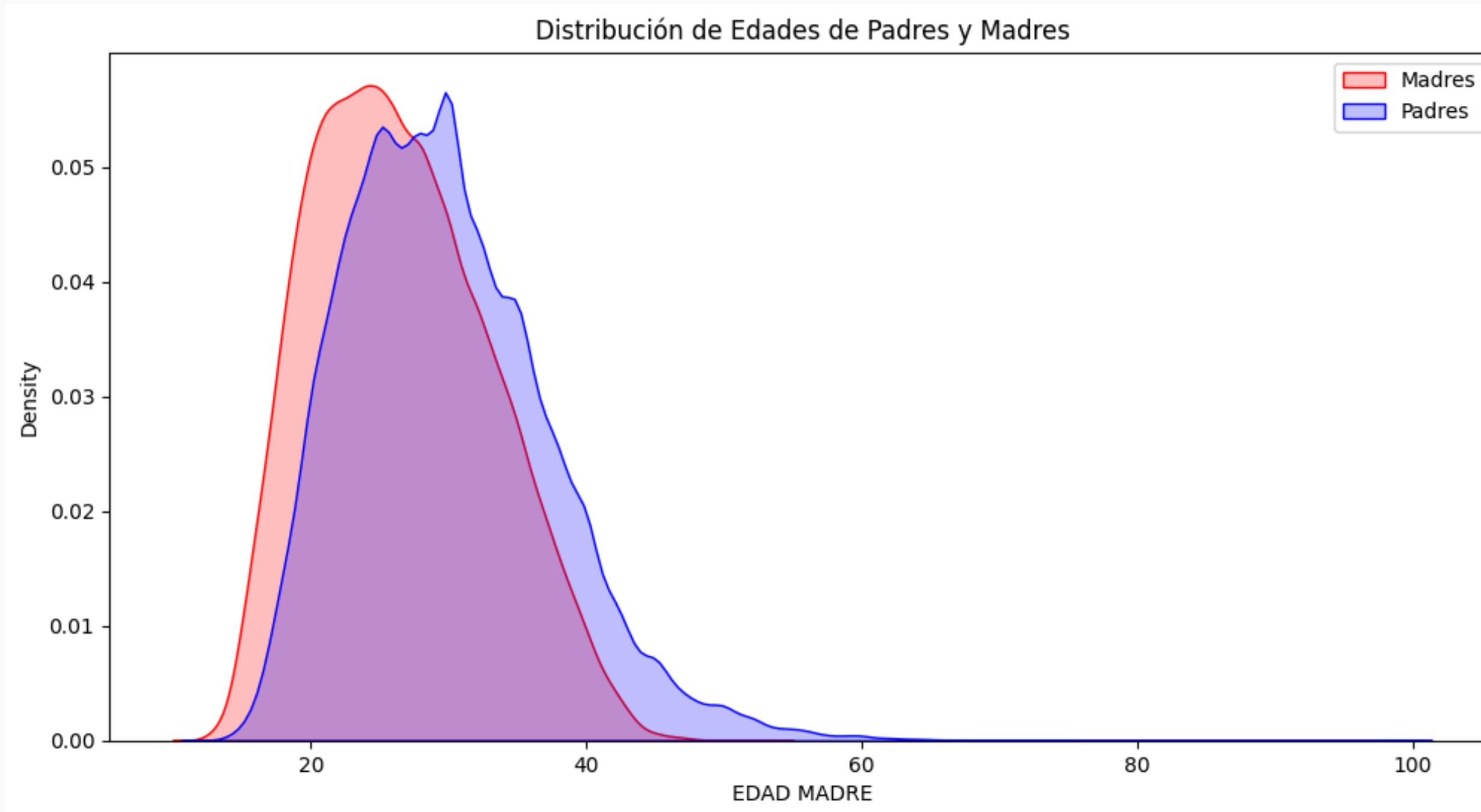


Hicimos un pair plot para comparar las distribuciones de los atributos principales

- Existe una **correlación positiva** entre el **peso** y la **talla**
- La distribución de las **edades de los padres** están **sesgadas hacia edades más jóvenes**
- La mayoría de los datos para el peso se concentran alrededor de 3000 a 4000 gr
- Para la talla alrededor de 40 a 50 cm



Gráfica de curvas para Edades de Padres y Madres

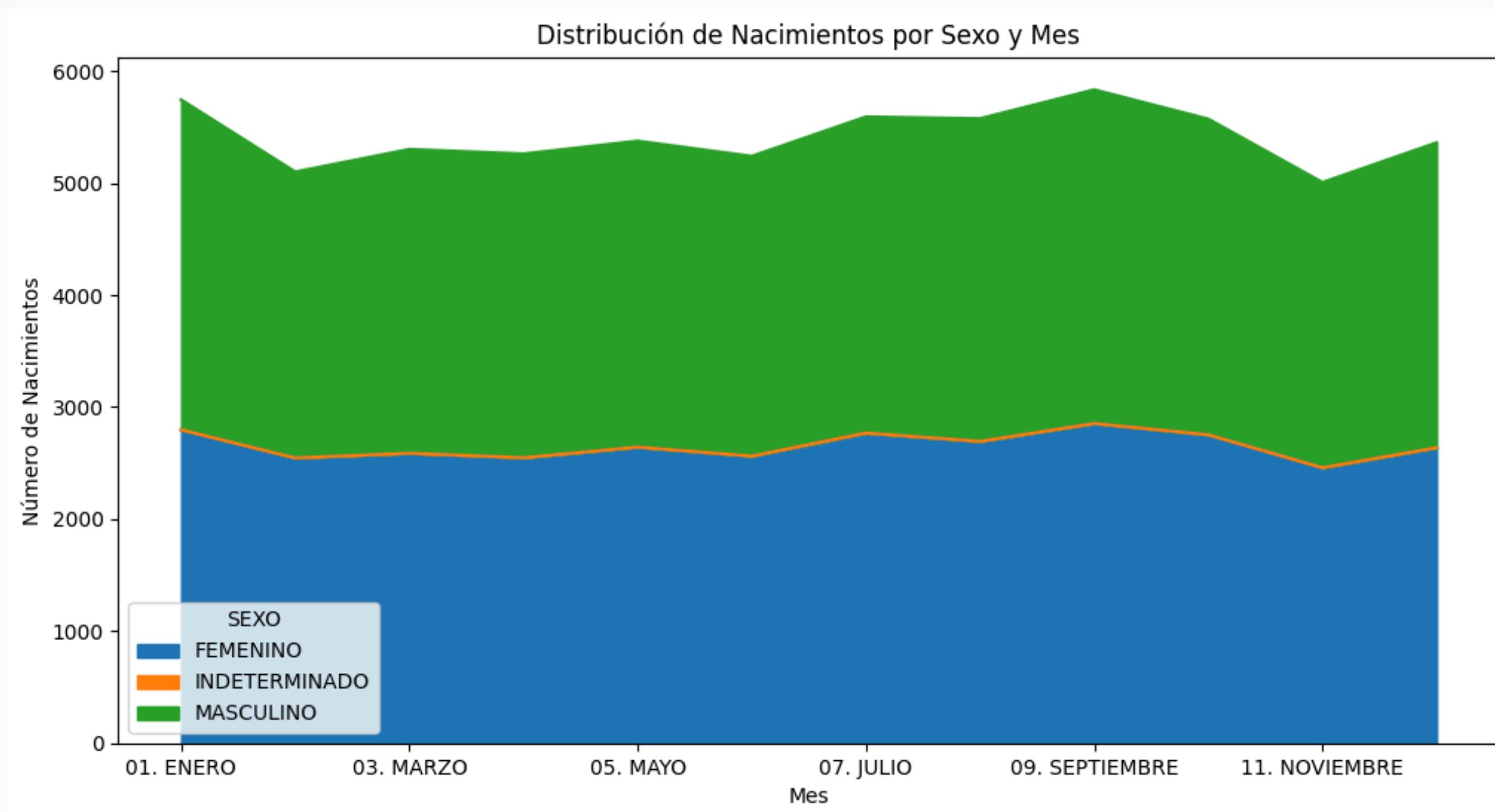


Hicimos un gráficos de curvas comprando las edades de los padres

- La tendencia de las madres es tener edades mas jóvenes
- Los papás tienen edades mas viejas
- Los papás tienen rango mas grande de edades



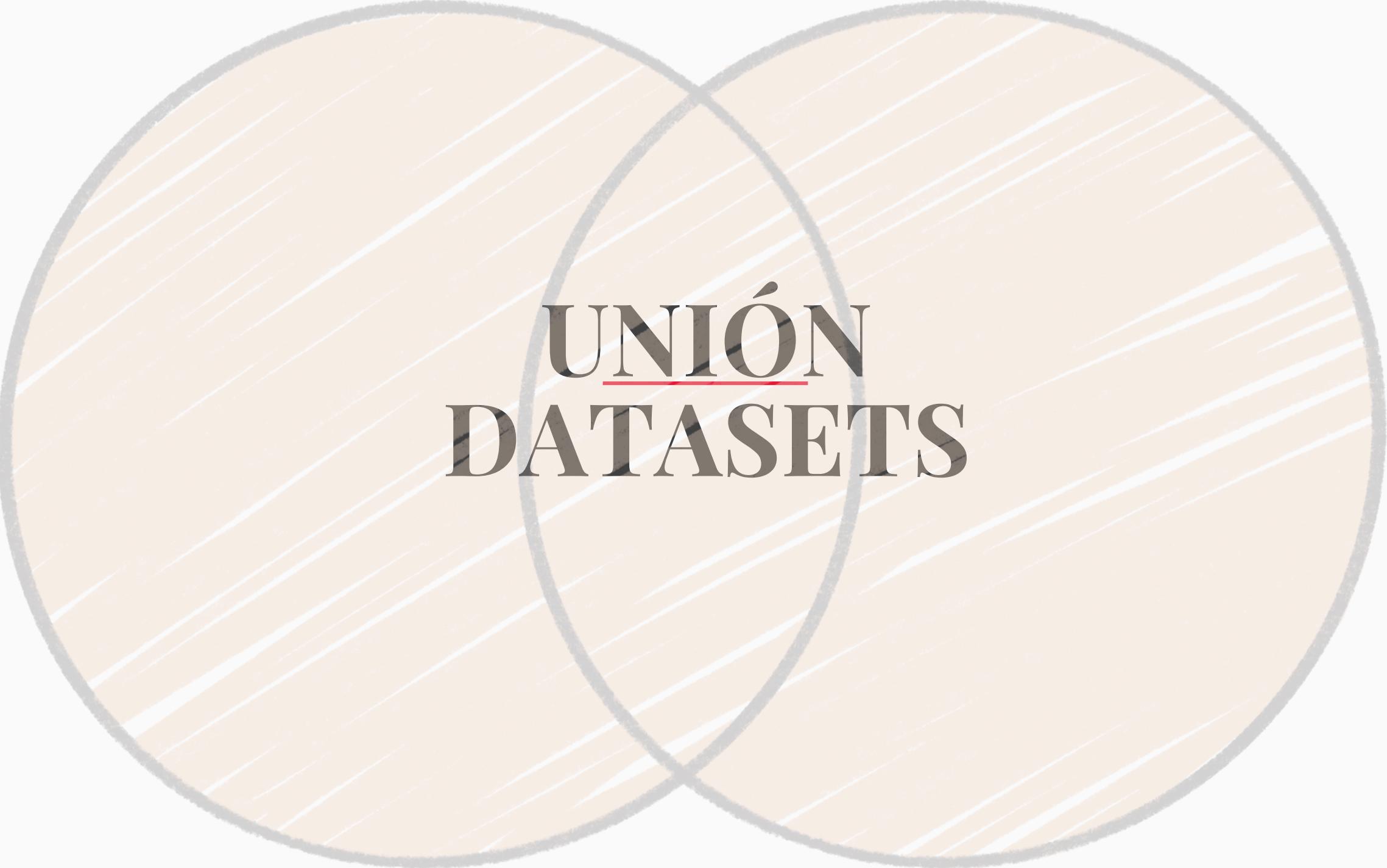
Gráfica de Nacimientos por Sexo y Mes



Hicimos un gráfico apilado para comprar el numero de nacimientos a lo largo del año por el sexo

- Se ve que nacen mas niños masculinos
- El pico de meses de masculinos es en septiembre
- El pico de los nacimientos femeninos es en julio-septiembre





UNIÓN DATASETS



Concatenación de los dataframes

Se tuvo que analizar que columnas coincidían entre los datasets y renombrar algunas..

```
# Renombrar las columnas en df para que coincidan con los nombres en dff
df_renamed = df.rename(columns={
    'TIEMPO DE GESTACIÓN (Semanas)': 'TIEMPO DE GESTACIÓN',
    'NUMERO CONSULTAS PRENATALES': 'NÚMERO CONSULTAS PRENATALES'
})

# Definir las columnas comunes a ambos DataFrames
columns_common = [
    'DEPARTAMENTO', 'MUNICIPIO', 'AREA NACIMIENTO', 'SEXO', 'PESO (Gramos)',
    'TALLA (Centímetros)', 'TIEMPO DE GESTACIÓN', 'NÚMERO CONSULTAS PRENATALES',
    'TIPO PARTO', 'MULTIPLICIDAD EMBARAZO', 'EDAD MADRE', 'EDAD PADRE'
]

# Seleccionar solo las columnas comunes en ambos DataFrames
df_final1 = df_renamed[columns_common]
dff_final1 = dff[columns_common]

# Concatenar los DataFrames
newdf = pd.concat([df_final1, dff_final1], ignore_index=True)

# Verificar los resultados
print("Dimensiones del DataFrame final:", newdf.shape)
print(newdf.head())
```



Transformación y limpieza de datos

ANTES

```
Valores nulos por columna:  
DEPARTAMENTO          0  
MUNICIPIO              0  
AREA NACIMIENTO        0  
SEXO                   2  
PESO (Gramos)           7  
TALLA (Centímetros)     7  
TIEMPO DE GESTACIÓN    10  
NÚMERO CONSULTAS PRENATALES 10  
TIPO PARTO              2  
MULTIPLICIDAD EMBARAZO   2  
EDAD MADRE              228  
EDAD PADRE              797  
dtype: int64
```



DESPUÉS

```
Valores nulos por columna después de la imputación:  
DEPARTAMENTO          0  
MUNICIPIO              0  
AREA NACIMIENTO        0  
SEXO                   0  
PESO (Gramos)           0  
TALLA (Centímetros)     0  
TIEMPO DE GESTACIÓN    0  
NÚMERO CONSULTAS PRENATALES 0  
TIPO PARTO              2  
MULTIPLICIDAD EMBARAZO   2  
EDAD MADRE              0  
EDAD PADRE              0  
dtype: int64
```

Tipos de datos de las columnas:

```
DEPARTAMENTO          object  
MUNICIPIO              object  
AREA NACIMIENTO        object  
SEXO                  object  
PESO (Gramos)           float64  
TALLA (Centímetros)     float64  
TIEMPO DE GESTACIÓN    float64  
NÚMERO CONSULTAS PRENATALES  float64  
TIPO PARTO              object  
MULTIPLICIDAD EMBARAZO   object  
EDAD MADRE              object  
EDAD PADRE              object  
dtype: object
```



Tipos de datos de las columnas después de la corrección:

```
DEPARTAMENTO          object  
MUNICIPIO              object  
AREA NACIMIENTO        object  
SEXO                  object  
PESO (Gramos)           float64  
TALLA (Centímetros)     float64  
TIEMPO DE GESTACIÓN    float64  
NÚMERO CONSULTAS PRENATALES  float64  
TIPO PARTO              object  
MULTIPLICIDAD EMBARAZO   object  
EDAD MADRE              float64  
EDAD PADRE              float64  
dtype: object
```

Respuesta a Preguntas



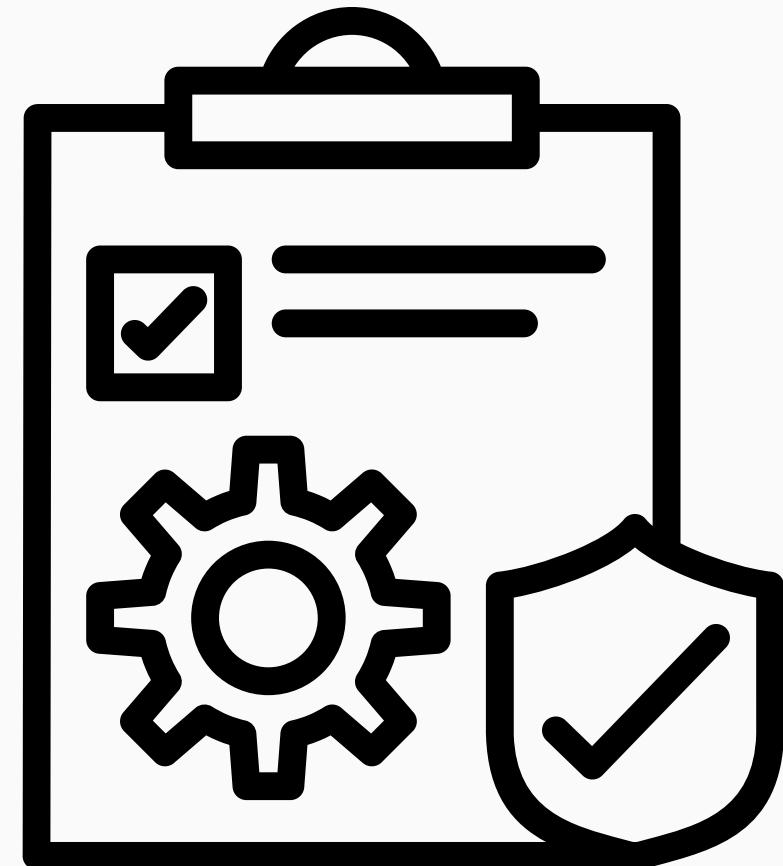
- **Pregunta 1:** ¿Cómo cambia la salud de los bebés entre Bucaramanga y Rionegro?
Importancia: Es esencial para identificar diferencias en la atención médica y condiciones socioeconómicas que afectan a los recién nacidos. Esta comparación ayudará a desarrollar estrategias específicas para mejorar la atención prenatal y neonatal en cada región, y a implementar intervenciones localizadas que optimicen los resultados de salud de los recién nacidos y mejoren los procesos de atención materna e infantil
- **Pregunta 2:** ¿Qué factores están asociados con los nacimientos prematuros (período de gestación menor a 37 semanas) en Bucaramanga y Río Negro?
Importancia: Reducir la morbilidad y mortalidad neonatal. Conocer estos factores permitirá diseñar intervenciones específicas y asignar recursos de manera más efectiva para apoyar a las mujeres embarazadas de alto riesgo.



Reporte de Calidad de Datos

Técnicas propuestas para tratar valores faltantes:

- Eliminación de filas o columnas
- Imputación de valores
- Utilización de modelos de imputación
- Codificación de valores faltantes
- Corrección de valores de entrada
- Formateo de fecha
- Detección de valores atípicos



Transformación y limpieza de datos

Por el momento tratamos de disminuir la dramática cantidad de valores nulos.

Para esto tomamos algunas consideraciones:

- Filtrado de valores atípicos: identificar cualquier valor que parezca estar significativamente fuera de lo común.
- Imputación de valores faltantes: Uso de mediana y valor más frecuente.
- Eliminación de columnas: Remoción de columnas con > 90% de datos faltantes

Transformación y limpieza de datos

- Revisar estadísticas descriptivas para identificar valores atípicos
- Calcular el percentil 95 de la columna
- Imputar valores de peso atípicos con la mediana

Estadísticas descriptivas:			
	PESO (Gramos)	TALLA (Centímetros)	TIEMPO DE GESTACIÓN
count	2074.000000	2074.000000	2074.000000
mean	3121.619576	48.623915	38.724204
std	383.912788	1.954703	1.174222
min	570.000000	25.000000	25.000000
25%	2870.000000	47.000000	38.000000
50%	3105.000000	49.000000	39.000000
75%	3367.500000	50.000000	40.000000
max	4940.000000	55.000000	43.000000

NÚMERO CONSULTAS PRENATALES			
count		2074.000000	
mean		6.609450	
std		2.256326	
min		0.000000	
25%		5.000000	
50%		7.000000	
75%		8.000000	
max		19.000000	

Descripción del problema de analitica a resolver

El objetivo principal es identificar diferencias significativas en la salud de los recién nacidos entre Bucaramanga y Río Negro, así como los factores asociados con los nacimientos prematuros en ambas ciudades. Esto permitirá desarrollar estrategias específicas para mejorar la atención prenatal y neonatal, así como asignar recursos de manera efectiva para apoyar a las mujeres embarazadas de alto riesgo

Técnicas de Análisis Propuestas del problema de analitica a resolver

Comparación de Técnicas:

1. Análisis de Clasificación:

- Para la clasificación, se pueden comparar algoritmos como Support Vector Machines (SVM) y Random Forest.
- SVM es una técnica poderosa para separar clases en conjuntos de datos complejos, mientras que Random Forest es robusto y puede manejar características irrelevantes o redundantes.

2. Análisis de Regresión:

- En el análisis de regresión, se pueden comparar modelos como Random Forest, Gradient Boosting, Linear Regression y Neural Networks.
- Gradient Boosting y Neural Networks han mostrado buen desempeño en términos de precisión y exactitud, mientras que Linear Regression ofrece tiempos de entrenamiento más rápidos.

Machine Learning: Regression

“En este estudio, aplicamos varias técnicas de Machine Learning para predecir el peso de los bebés al nacer, utilizando variables clave como el tiempo de gestación, el número de consultas prenatales, y la edad de los padres.”



Modelos de Regresión y Configuración

Para nuestro análisis, evaluamos los siguientes modelos:

- Random Forest
- Gradient Boosting
- Linear Regression
- Neural Networks

Cada modelo fue seleccionado por su capacidad de manejar diferentes aspectos y complejidades de nuestros datos.



Evaluación de Rendimiento de los Modelos

Modelo	MAE	MSE	RMSE	R ²	MAPE	Explained Variance	Training Time (s)
Random Forest	322.65	166800.88	408.41	0.346	10.63%	0.346	14.44
Gradient Boosting	287.6	132768.21	364.37	0.479	9.50%	0.479	3
Linear Regression	291.03	136248.58	369.12	0.466	9.65%	0.466	0.07
Neural Network	288.84	135066.33	367.51	0.47	9.48%	0.472	206.61



Análisis Detallado del Rendimiento

El análisis de rendimiento revela:

- **Gradient Boosting** y **Neural Network** ofrecen la mejor precisión y exactitud, con los valores más bajos en MAE, MSE y RMSE.
- **Gradient Boosting** demuestra ser el más eficiente en explicar la variabilidad de los datos con el mayor R^2 y Explained Variance.
- **Linear Regression** destaca por su rapidez en el entrenamiento, siendo el más rápido entre los modelos evaluados.

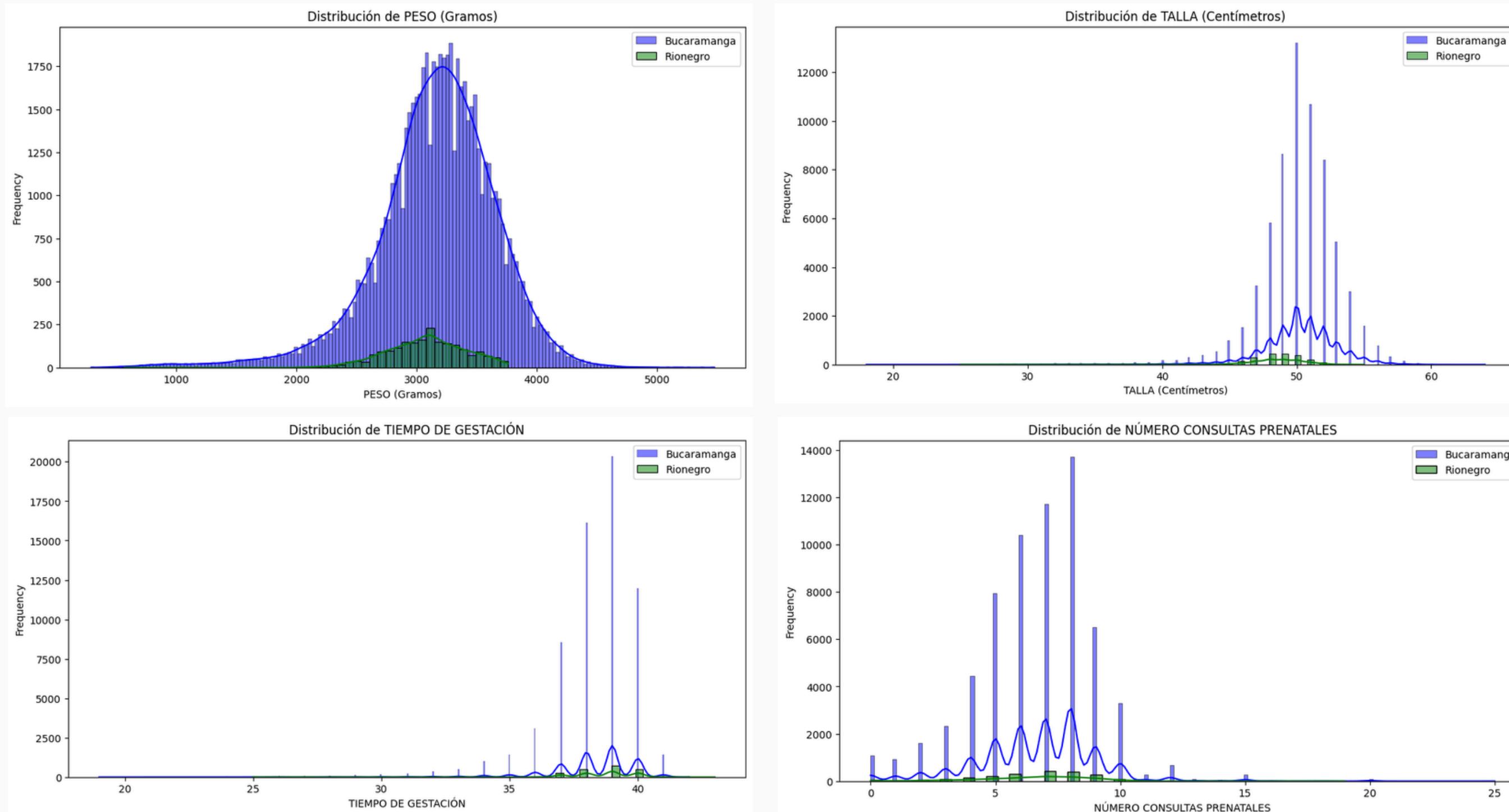


Respuesta a Preguntas

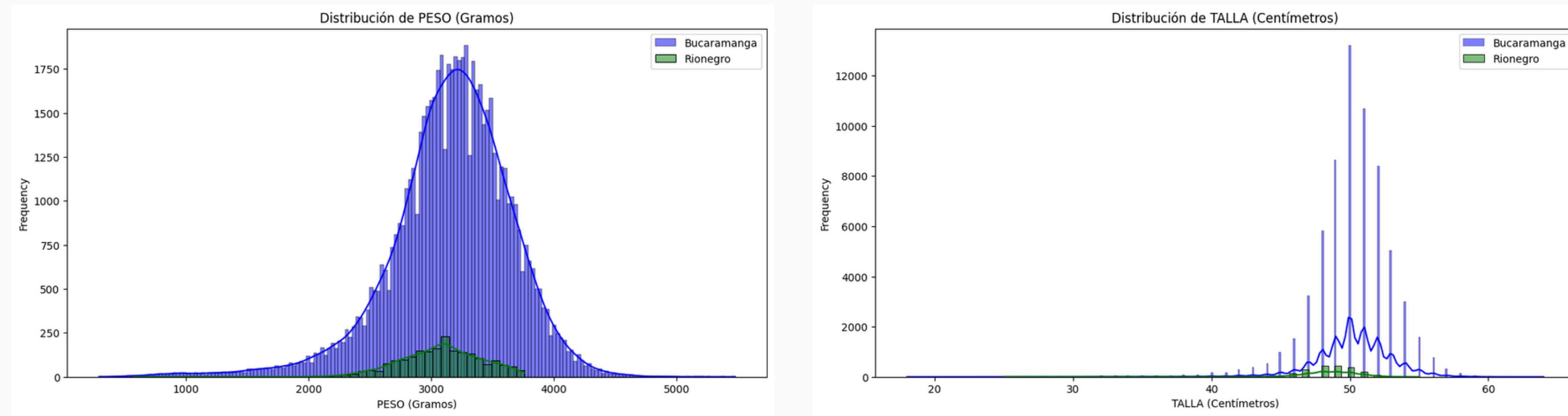
- ¿Cómo cambia la salud de los bebés entre Bucaramanga y Rionegro?
- ¿Qué factores están asociados con los nacimientos prematuros en Bucaramanga y Rionegro?



¿Cómo cambia la salud de los bebés entre Bucaramanga y Rionegro?



Análisis de Peso y Talla al Nacer

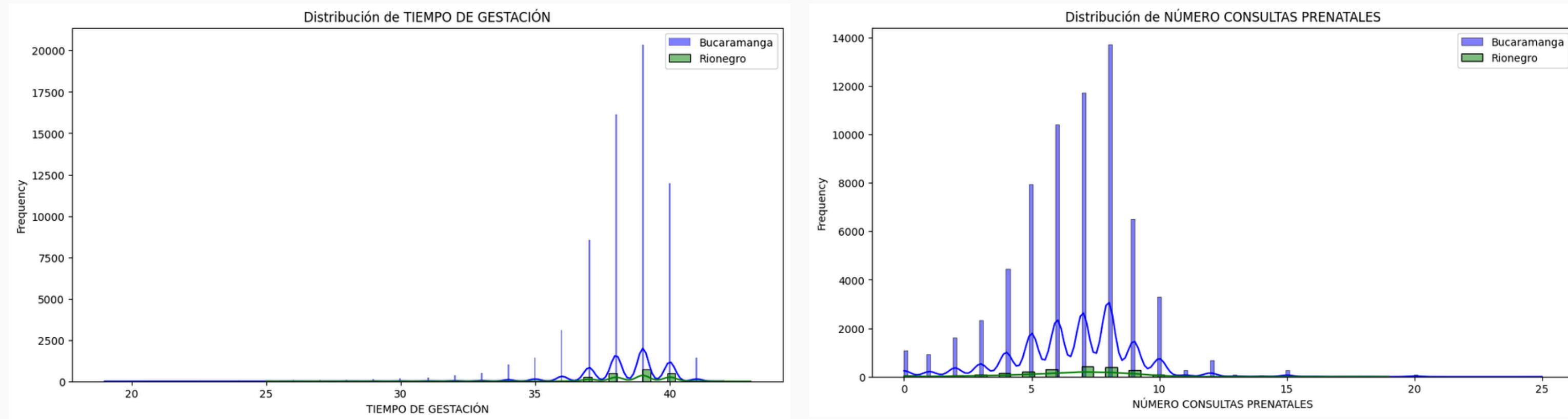


- **"Peso al Nacer:** En Bucaramanga, el peso promedio al nacer es de 3157.54 gramos, comparado con 3080.40 gramos en Rionegro."
- **"Talla al Nacer:** Los bebés en Bucaramanga tienen una talla promedio de 50.16 centímetros, superior a los 48.62 centímetros en Rionegro."

"Estas diferencias sugieren variaciones en el desarrollo fetal, influenciadas posiblemente por la nutrición y otros factores socioeconómicos."



Tiempo de Gestación y Acceso a Atención Prenatal



- "El tiempo de gestación es similar en ambos municipios, con una mediana cercana a las 39 semanas."
- "El número medio de consultas prenatales es aproximadamente 6.6 en ambos municipios, indicando un acceso uniforme a la atención prenatal."



¿Cómo cambia la salud de los bebés entre Bucaramanga y Rionegro?

En general, aunque existen algunas variaciones entre **Bucaramanga y Rionegro** en términos de peso y talla al nacer, el cuidado prenatal medido por las consultas parece ser uniformemente distribuido, sugiriendo que las diferencias observadas podrían ser influenciadas por factores socioeconómicos o ambientales más que por la calidad de la atención médica prenatal.



¿Qué factores están asociados con los nacimientos prematuros en Bucaramanga y Rionegro?



Identificamos varios factores clave asociados con los nacimientos prematuros en las dos ciudades estudiadas. A través de nuestro análisis, descubrimos que:

- **Peso y Talla al Nacer:** Existe una fuerte correlación positiva (0.78) entre estas medidas al nacer y el tiempo de gestación. Menores pesos y tallas están comúnmente asociados con nacimientos prematuros.
- **Número de Consultas Prenatales:** Observamos una correlación positiva débil (0.19) entre el número de consultas prenatales y un mayor tiempo de gestación.



Conclusiones Generales del Estudio Comparativo

Este estudio ha revelado diferencias significativas en los patrones de nacimiento entre Bucaramanga y Río Negro, destacando variaciones sociodemográficas y en la atención médica que reflejan la necesidad de enfoques personalizados en políticas de salud pública.



Observaciones Clave del Proyecto

- **Variabilidad en los Datos:** Hemos encontrado variaciones notables en aspectos clave como el peso al nacer y el número de consultas prenatales entre las ciudades estudiadas.
- **Calidad de Datos:** La limpieza y preparación de datos fueron cruciales para asegurar la precisión de nuestros análisis.
- **Uso de Tecnologías de Machine Learning:** La aplicación de modelos de regresión y análisis estadístico ha mejorado nuestra comprensión de los factores influenciadores en los patrones de nacimiento.



Recomendaciones para Mejorar la Atención Prenatal

- **Fortalecimiento de la Atención Prenatal:** Mejorar la calidad y frecuencia de la atención prenatal para asegurar que todas las mujeres embarazadas reciban el cuidado necesario.
- **Educación y Concienciación:** Implementar programas educativos para aumentar la conciencia sobre la importancia de la atención prenatal.
- **Investigación Continua:** Continuar con la investigación para explorar las causas de las variaciones observadas y evaluar la efectividad de nuevas intervenciones.
- **Colaboración Intersectorial:** Promover la colaboración entre diferentes sectores para abordar las necesidades de salud identificadas de manera más efectiva.
- **Uso de Datos en Tiempo Real:** Integrar tecnologías que permitan el uso de datos en tiempo real para mejorar la respuesta a las necesidades de salud materna e infantil.



Conclusión Final del Proyecto

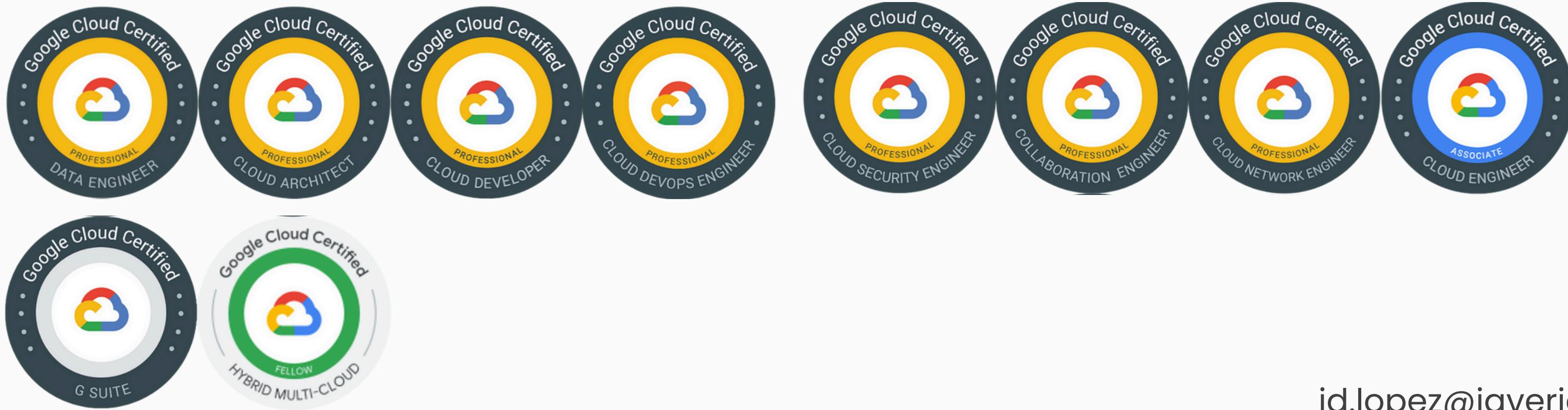
El análisis de datos a gran escala ha proporcionado insights valiosos que pueden informar y mejorar las políticas de salud pública. Este estudio subraya la importancia de adaptar las intervenciones de salud pública a las condiciones locales específicas para mejorar efectivamente la salud de madres y bebés en Bucaramanga, Río Negro y otras regiones similares.



Invitación a Certificarse con GCP.

Los exámenes de Google Cloud tienen un costo de alrededor de **200 Dólares**.

Si desean obtener el entrenamiento necesario para certificarse con Google Cloud y hacer el examen de forma **100% gratuita**. Me pueden contactar directamente, hay cupos limitados.



jd.lopez@javeriana.edu.co

Gracias

Referencias

- Nacidos Hospital San Juan de Dios Rionegro, Ant. Año 2021 By Container: Datos.gov.co Year: 2022 URL: https://www.datos.gov.co/dataset/Nacidos-Hospital-San-Juan-de-Dios-Rionegro-Ant-A-o/ru2a-86h9/data_preview
- 43. Nacidos Vivos en Municipio de Bucaramanga enero 2016 a febrero 2023 By Secretaría de Container: Datos.gov.co Year: 2019 URL: https://www.datos.gov.co/Salud-y-Protecci-n-Social/43-Nacidos-Vivos-en-Municipio-de-Bucaramanga-enero/x5xp-9w4b/about_data