Visualización para grandes volúmenes de datos

Mario Andrés Hernández Moreno

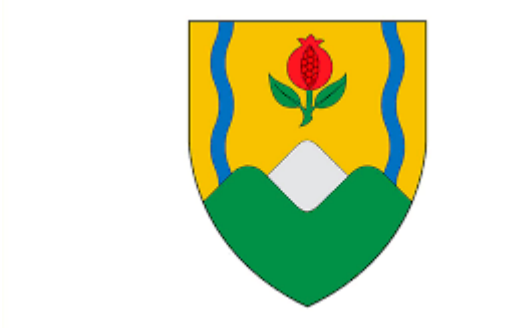
Actividad 3: Proyecto de Visualización de Datos

Mag. Mario Alejandro Bravo Ortiz

Universidad Autónoma de Manizales, Manizales
Especialización en Inteligencia Artificial
2024

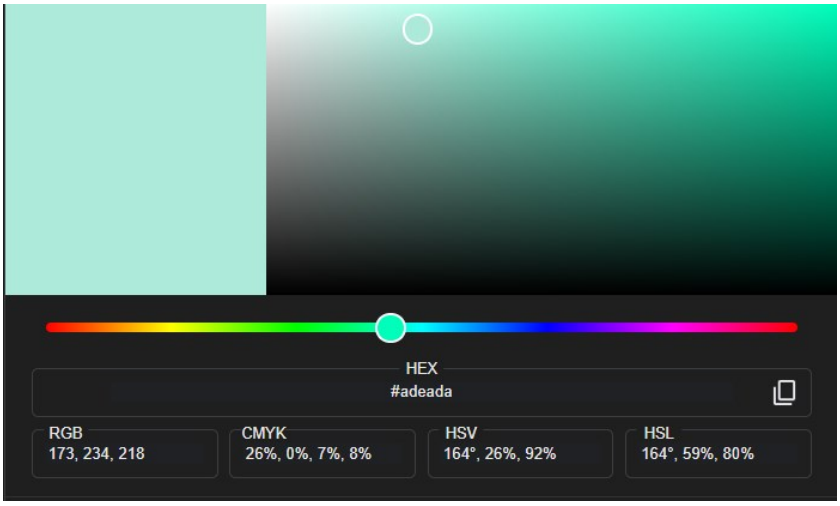
Contenido

1. Nacimientos por residencia departamento de Caldas



Para la visualización de grandes volúmenes de datos, se hizo uso de librerías como numpy, pandas, seaborn, matplotlib, metworkx, plotly, statsmodels, wordcloud y geopandas. Además, se utilizaron archivos json para obtener algunas coordenadas y características geográficas para poder graficar los mapas del departamento de Caldas con sus respectivos municipios.

También es importante mencionar que se usó un color estándar que fue el #adeada para tratar de unificar los gráficos estéticamente y llevar un orden en cuanto a la temática trabajada.

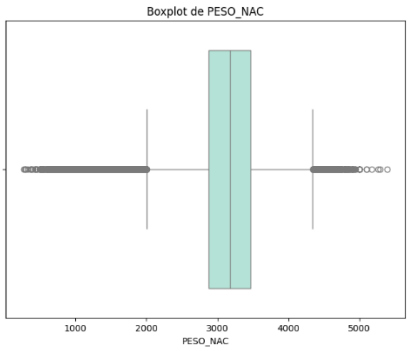


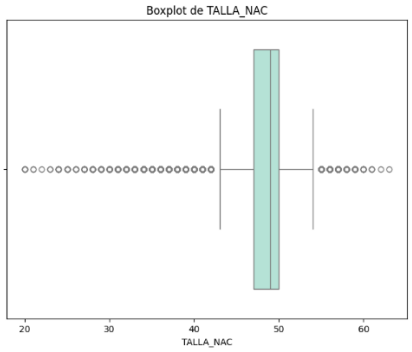
```
!pip install statsmodels
!pip install wordcloud
!pip install geopandas
```

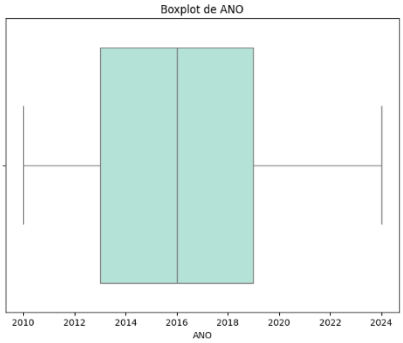
```
import os
import re #MANIPULACIONES DE CADENA
import math #CALCULOS
import numpy as np #MATRICES Y ARRGLOS DE DATOS
import pandas as pd #ESTRUCTURAS A ALTO DE DATOS Y MANIPULACION
import seaborn as sns #GRAFICAR VISUALIZACIONES
import matplotlib.patches as mpatches #VISUALIZACIONES
import matplotlib.pyplot as plt #VISUALIZACIONES
from matplotlib.colors import LinearSegmentedColormap
import networkx as nx
import plotly.express as px
import plotly.graph_objs as go
import plotly.offline as pyo # para exportar en html
import json
import plotly.graph_objects as go
from urllib.request import urlopen
with urlopen('https://gist.githubusercontent.com/john-guerra/43c7656821069d00dcbc/raw/be6a6e239cd5b5b803c6e7c2ec405b793a9064dd/Colombia.geo.json') as response:
    counties = json.load(response)
from wordcloud import WordCloud, STOPWORDS
from PIL import Image
import geopandas as gpd

from google.colab import files
```

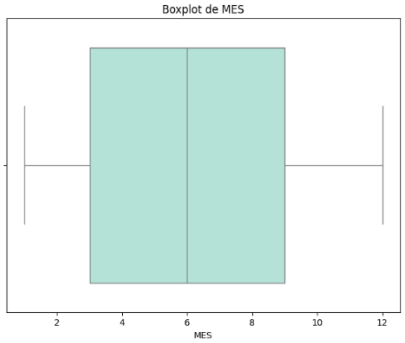
El DataFrame que se trabajó fue de 129938 filas y 16 columnas con la siguiente información: Departamento de Nacimiento, Municipio de Nacimiento, Área del Nacimiento, Sitio de la Parto, Sexo del nacido vivo, Peso del nacido vivo, al nacer, Talla del nacido vivo, al nacer, Año del nacimiento, Mes del nacimiento, Tiempo de gestación del nacido vivo, Número de consultas prenatales que tuvo la madre del nacido vivo, Tipo de parto de este nacimiento, Multiplicidad del embarazo, Edad de la madre a la fecha del parto y Número de hijos nacidos vivos que ha tenido la madre.

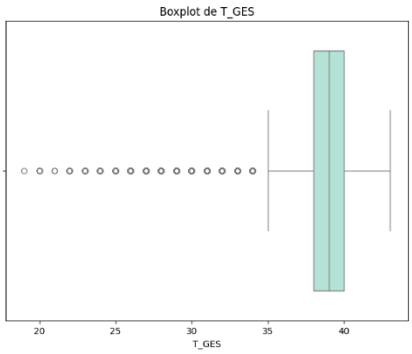
1.1 Gráficas de caja:

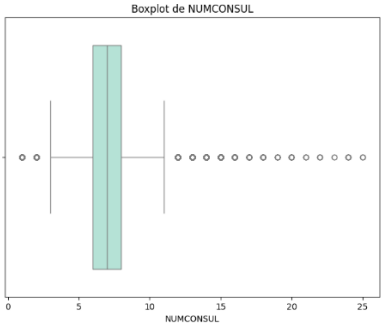


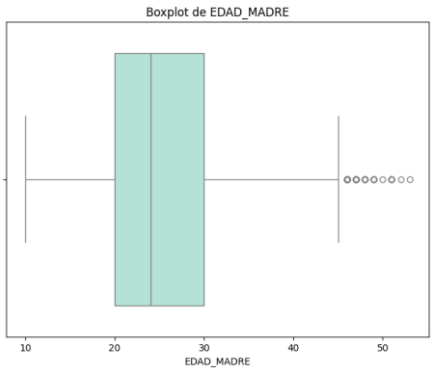






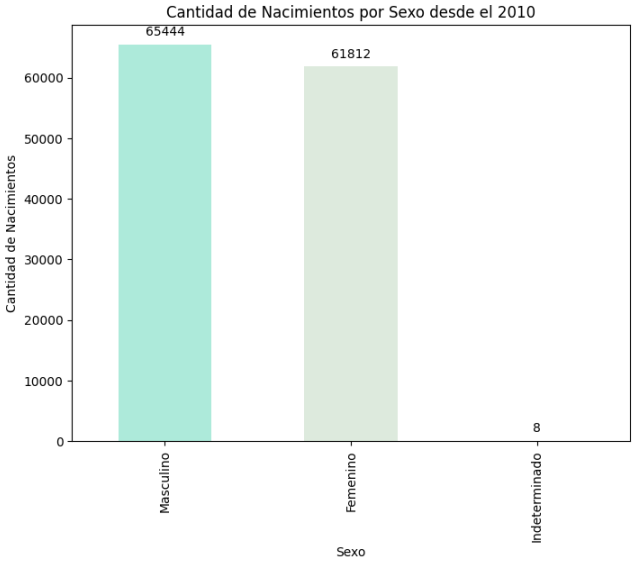






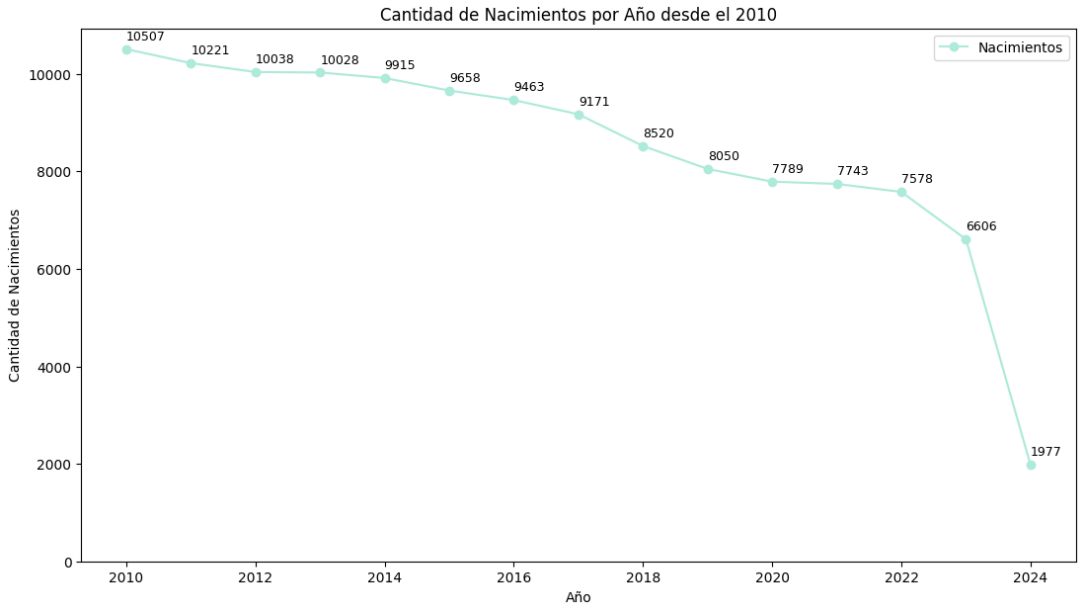
En este caso las gráficas de caja nos ayudan a visualizar si existen o no valores atípicos, fuera de los rangos esperados acorde al contexto del cual se está estudiando, también nos permiten identificar los rangos de los valores por ejemplo del peso de los recién nacidos para identificar bebés con peso extremadamente bajo o alto y la edad de la madre para identificar madres muy jóvenes o mayores.

1.2 Gráfica de barras:



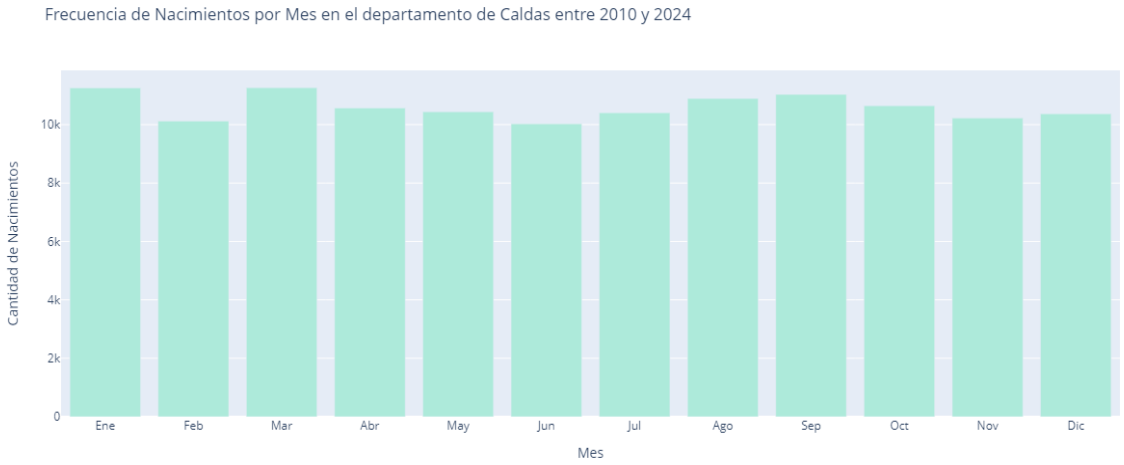
Para este caso usamos un gráfico de barras con sus respectivas etiquetas para comparar la cantidad de nacimiento por sexo, utilizando las variables categóricas con su respectivo agrupamiento, esto nos permite ver de una manera clara las cantidades facilitando la interpretación y logrando inferir que desde el 2010 han nacido más bebés de género masculino que bebés de género femenino.

1.3 Gráfica de líneas:



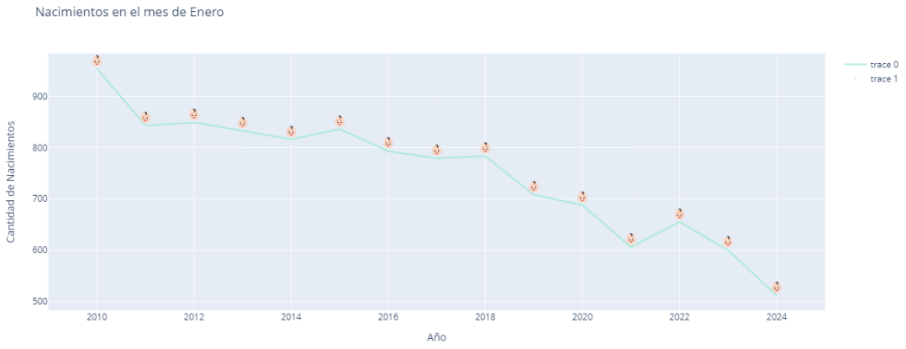
Para este caso utilizamos un gráfico de líneas donde podemos ver la evolución de la cantidad de nacimientos por años desde el 2010, con una disminución considerable, esto nos facilita identificar tendencias temporales, la progresión, las fluctuaciones en la tasa de natalidad a lo largo de los años, detectar cambios abruptos, comparar periodos e incluso llegar a realizar proyecciones futuras.

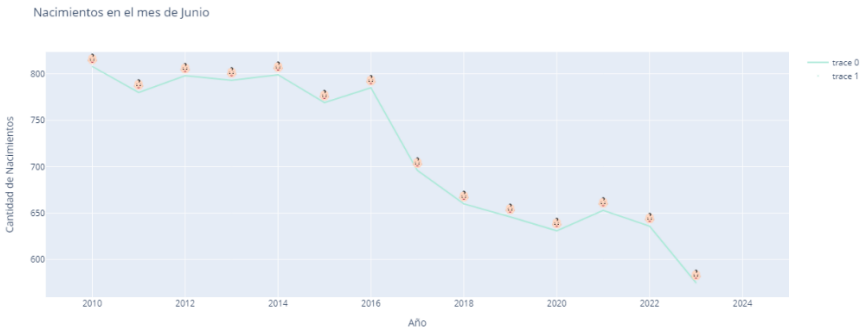
1.4 Gráfica de barras:



Para esta gráfica podemos comparar de manera visual y clara la diferencia de nacimientos por mes, por ejemplo podemos analizar que en marzo fue el mes en que más nacieron bebés desde el 2010 y Junio el mes en que menos bebés nacieron, logrando interactuar con la gráfica y facilitando la interpretación de los datos.

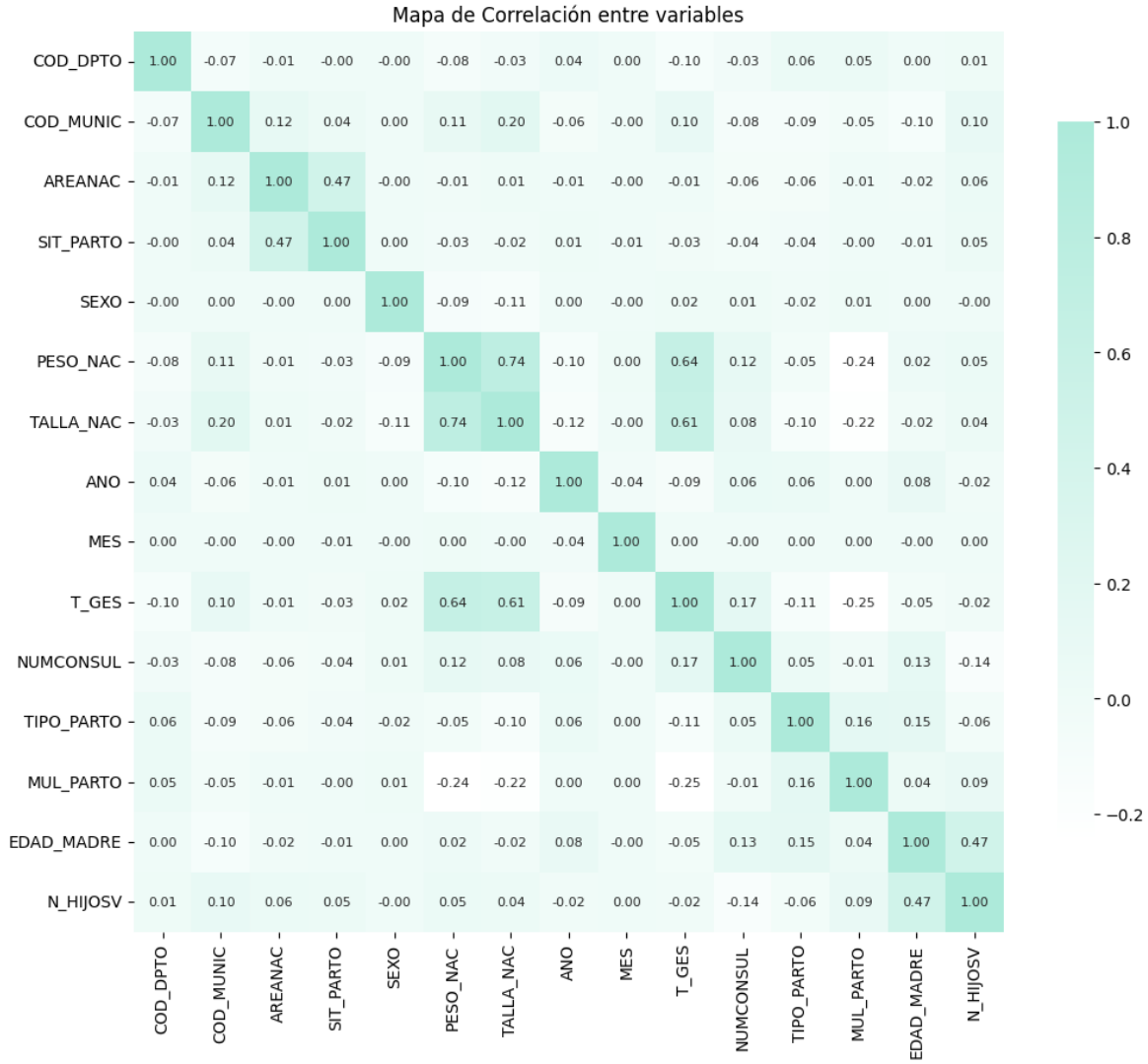
1.5 Gráficas de líneas:





En este caso comparamos la evolución de los nacimientos por cada mes a lo largo de los años, esto nos permite identificar tendencias por segmentación de cada mes del año, el cómo ha evolucionado la natalidad, facilita la comparación entre meses, la identificación de anomalías y puede ayudar a respaldar análisis de impactos externos y facilitar proyecciones. También se podría hacer el ejercicio de relación de cada año con sus respectivos meses.

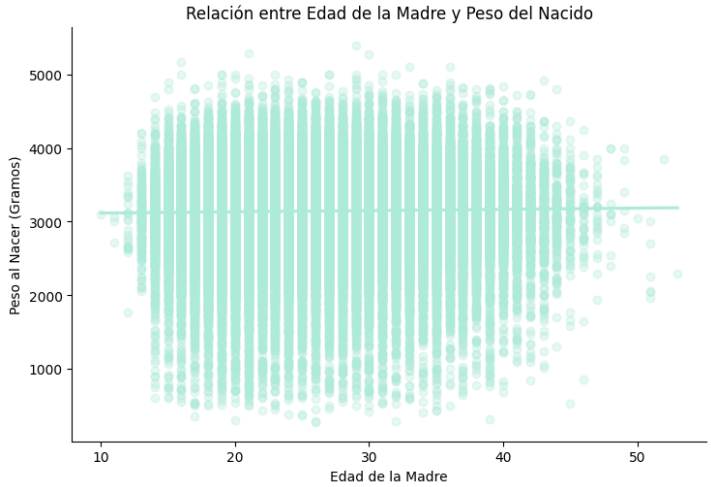
1.6 Matriz correlación:



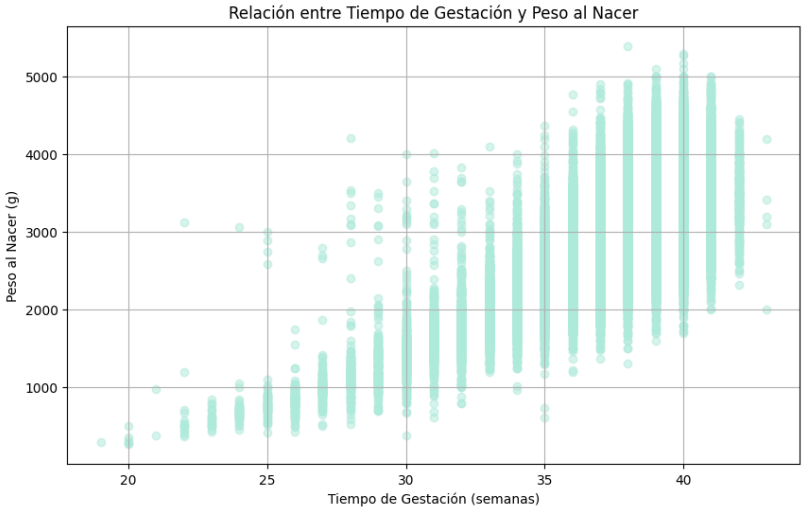
El mapa de correlación nos ayuda a identificar múltiples relaciones entre las variables que maneja nuestro df como lo podemos evidenciar con correlaciones positivas entre en la talla del

recién nacido y su peso, el tiempo de gestación y el peso al nacer, al igual que con la talla, y otras en menor medida pero relevantes como la relación entre la edad de la madre y el número de hijos que ha tenido, aportando a la identificación de patrones sociodemográficos, comparando categorías y aclarando el panorama para seguir analizando los datos.

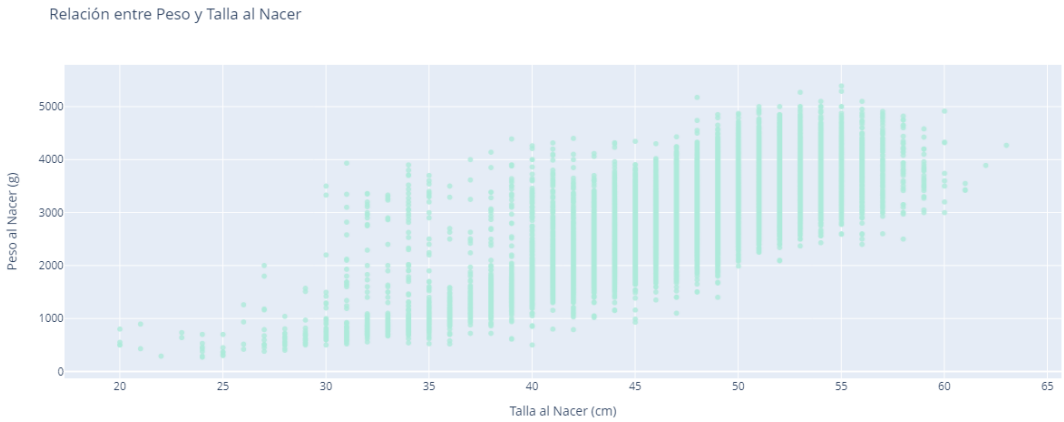
1.7 Gráficos de dispersión:



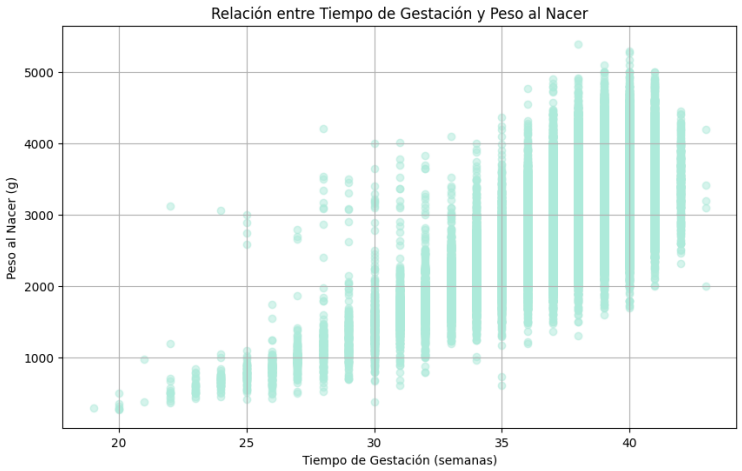
En esta gráfica podemos explorar la relación entre la edad de la madre y el peso del nacido, a mi parecer hay una leve relación directa pero que no es muy clara, sin embargo, esto nos ayuda a analizar la variabilidad de los datos.



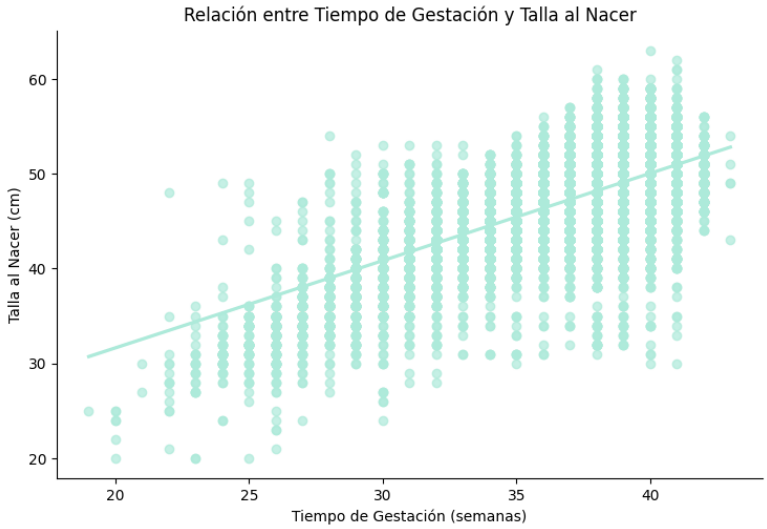
Para esta gráfica si encontramos una correlación positiva entre las semanas de gestación que tiene la madre y el peso del bebé al nacer, lo que indica que, a mayor tiempo de gestación, mayor es el peso del bebé. Esto también nos indica que los bebés prematuros suelen tener pesos considerablemente más bajos, pero a partir de cierto tiempo de gestación, el aumento del peso podría desacelerarse.



La gráfica de dispersión nos ayuda a identificar la relación entre el peso y la talla al nacer, ya que ambos son indicadores clave de la salud neonatal, por lo que a mayor tamaño (talla), generalmente, mayor es el peso. También podemos observar outliers, como recién nacidos con un peso bajo para su talla (posible indicador de crecimiento intrauterino restringido) o, al contrario, bebés con un peso inusualmente alto para su estatura.

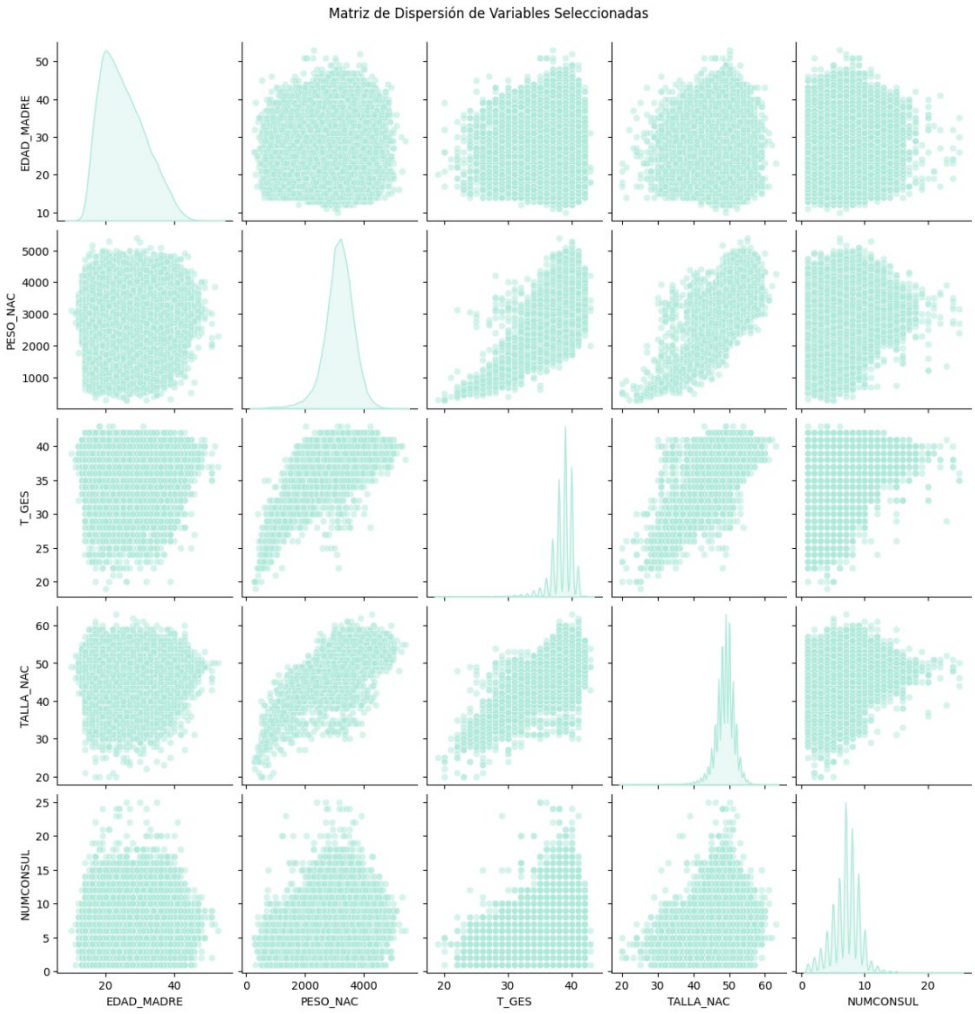


Para esta gráfica si encontramos una correlación positiva entre las semanas de gestación que tiene la madre y el peso del bebé al nacer, lo que indica que, a mayor tiempo de gestación, mayor es el peso del bebé. Esto también nos indica que los bebés prematuros suelen tener pesos considerablemente más bajos, pero a partir de cierto tiempo de gestación, el aumento del peso podría desacelerarse.



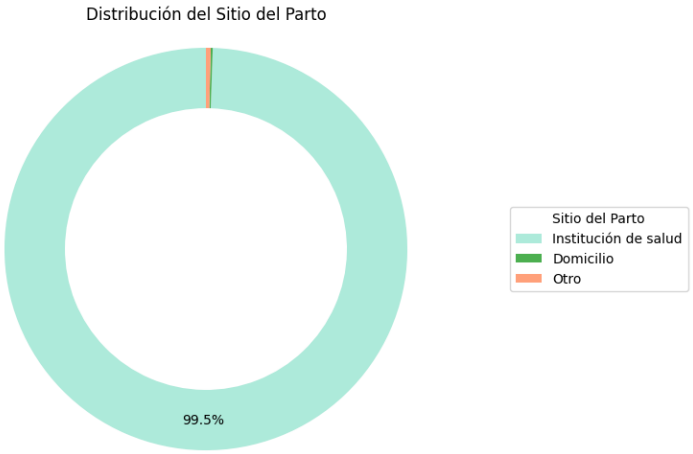
Para esta gráfica es ideal analizar la relación entre el tiempo de gestación en semanas y la talla del bebé al nacer ya que, a mayor tiempo en el útero, el bebé tiende a crecer más, indicando una correlación positiva, aunque también podemos encontrar algunos outliers donde los bebés tienen una talla mucho menor o mayor de lo esperado para su tiempo de gestación. Otro aspecto importante a destacar es que el gráfico de dispersión ayuda a identificar riesgos neonatales. Por ejemplo, si los bebés con un tiempo de gestación menor de 37 semanas (prematuros) tienen una talla significativamente menor, esto puede alertar sobre la necesidad de atención neonatal especializada en algunos municipios de Caldas.

1.8 Gráfica pairplot o Matriz de Dispersión:

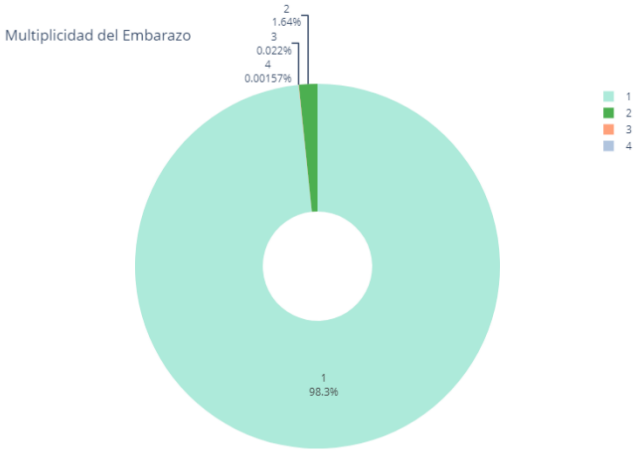


En este caso podemos visualizar simultáneamente las relaciones en algunas variables, como las que ya habíamos evidenciado con el peso y el tiempo de gestación, la talla y el peso, pero también podemos visualizar un ligera relación no tan directa entre el número de consultas y el tiempo de gestación, así mismo podemos ver que hay otras variables que no tienen una relación clara como el número de consultas prenatales y la edad de la madre, pues la matriz de dispersión facilita la detección visual de patrones complejos y de valores atípicos, facilitando la toma de decisiones basada en datos.

1.9 Gráficas de torta:



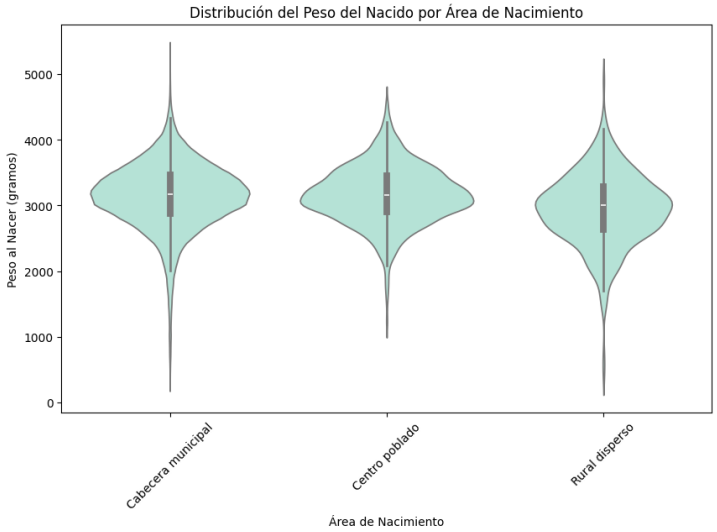
En el gráfico de torta tenemos la distribución del sitio del parto, este gráfico es ideal porque nos muestra cómo se distribuyen las los sitios de parto, entendiendo la frecuencia de cada tipo de lugar en que nacieron los bebés, comparando fácilmente las distintas categorías, que para este caso la gran mayoría de partos se dieron en una institución de salud con un 99.5% el resto se distribuye entre el domicilio y otros lugares no especificados en el df.



En este caso tenemos la gráfica de torta de manera interactiva para analizar la distribución de la multiplicidad del embarazo, es una forma clara y directa de mostrar cómo se distribuyen los diferentes tipos de embarazo dentro de un total, siendo el 98.3% (125151) de embarazos de tipo

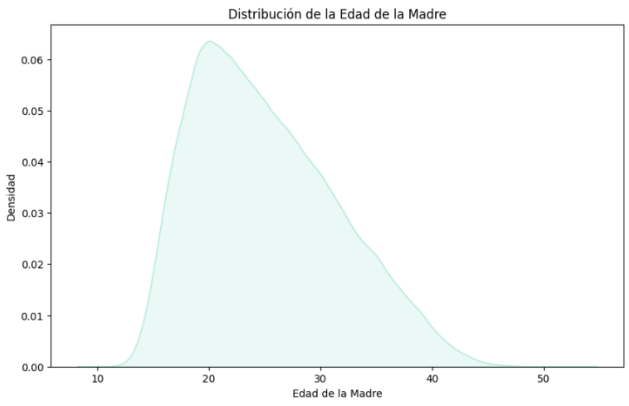
simple, el 1.64% (2083) de tipo doble, el 0.022% (28) de tipo triple y el 0.00157% (2) de tipo cuádruple o más lo que resalta la prevalencia de partos simples frente a múltiples.

1.10 Gráfica de violín:



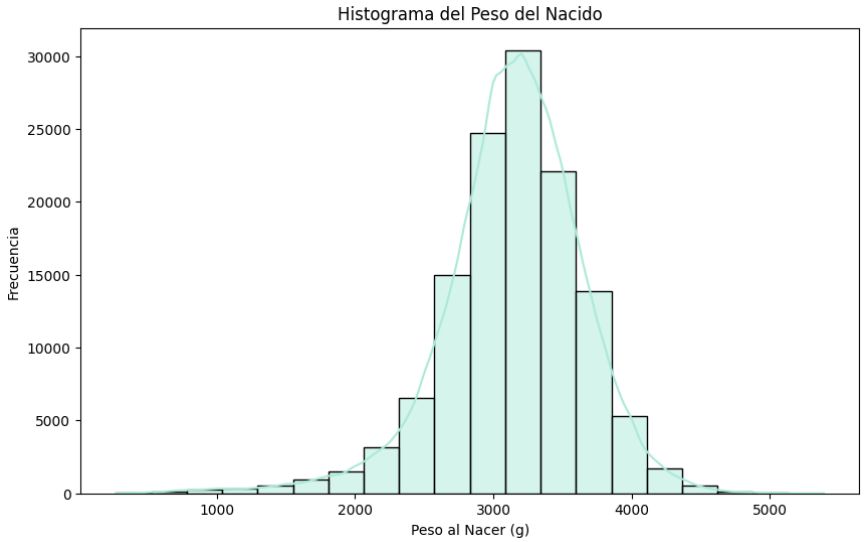
Para este gráfico de violín podemos ver la distribución del peso del nacido por el área de nacimiento, lo que nos ofrece una representación detallada de la distribución completa de los datos, si son simétricos o existen sesgos (hacia pesos más bajos o más altos) o si hay multimodalidad (varias concentraciones de valores), destacando tanto la densidad, variabilidad, concentración y forma de la distribución en cada categoría. Además, facilita la comparación entre áreas, permite identificar posibles outliers.

1.11 Gráfico de dispersión:



El uso de la gráfica de densidad nos permite ver hacia que rango de edad se concentra la mayoría de las madres gestantes, la asimetría (si la mayoría de las madres son jóvenes o de mayor edad) y la curtosis (concentración de los datos en torno a la mediana), existe un pico pronunciado en torno a los 20-30 años, lo que sugiere que la mayor parte de las madres son jóvenes.

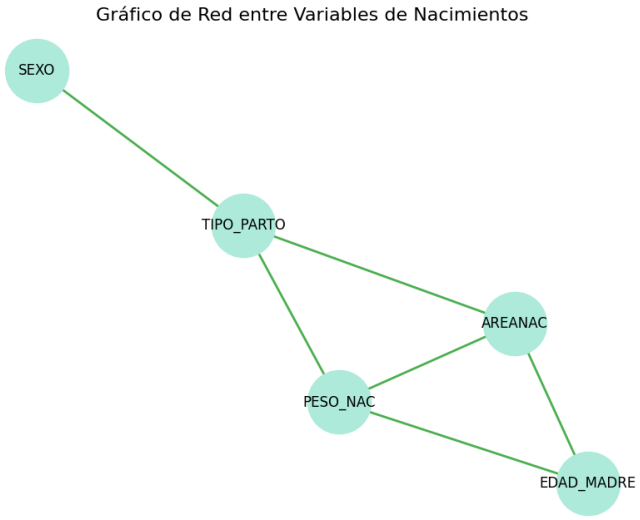
1.12 Histograma:



Este histograma nos permite ver la cantidad de nacidos vivos dentro de la distribución del peso lo que facilita entender que tan dispersos están los datos, indican una distribución normal, con una alta cantidad de bebés con pesos entre 2500 y 3500 g, también indica que hay bebés con algunos pesos por debajo de lo esperado que seguramente son prematuros o con un peso muy alto que se lo puede analizar así:

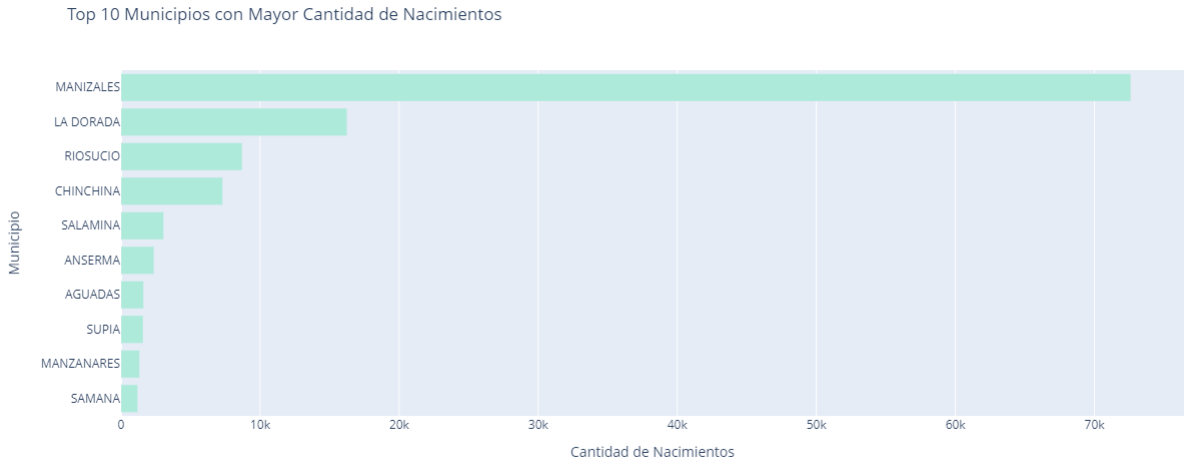
- Bajo peso al nacer (menor de 2500 gramos)
- Peso normal (2500-4000 gramos)
- Macrosomía (más de 4000 gramos)

1.13 Grafo:



Esta es otra gráfica que nos permite analizar las relaciones y patrones en el contexto de nacimientos, pues se muestra cómo diferentes características de los datos están interconectadas, por ejemplo, podemos observar que el peso al nacer tiene relación directa con ciertas variables como la edad de la madre, el área de nacimiento y el tipo de parto por lo que la edad de la madre puede influir en el peso del bebé al nacer. Las madres de mayor edad tienden a tener embarazos más complejos, lo que puede afectar el peso. Por otro lado, el área de nacimiento (urbano o rural) podría estar relacionada con el peso del bebé, debido a las diferencias en acceso a atención médica y nutrición entre las áreas. Otro aspecto importante es que, dependiendo del área de nacimiento, ciertos tipos de parto pueden ser más comunes. Por ejemplo, en áreas rurales puede haber más partos naturales o en casa debido a la limitada infraestructura de salud.

1.14 Gráfica de barras:

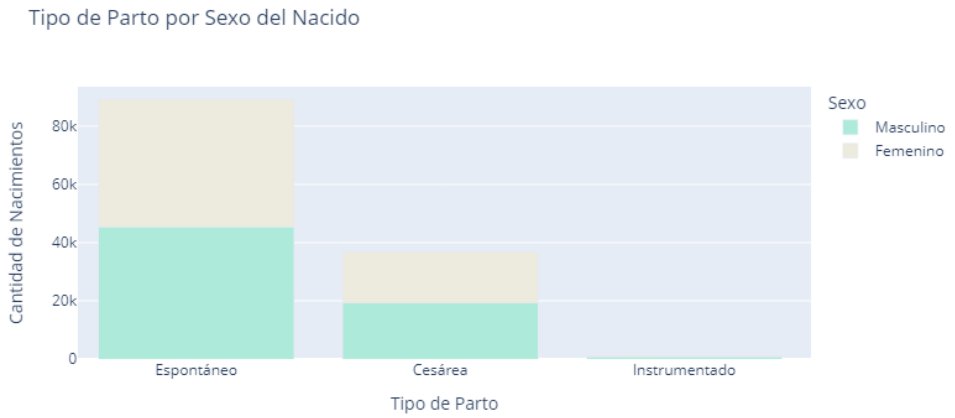


Para la gráfica de barras interactiva codificamos los nombres de los municipios del departamento de Caldas y graficamos en torno a los 10 con mayor cantidad de nacimientos, lo que facilita comparar la diferencia de nacimientos en cada lugar, siendo Manizales el municipio donde más nacen bebés desde el 2010 quizá respaldado por la concentración de personas en la ciudad y la cantidad de instituciones de salud, en segundo lugar tenemos a la Dorada y en tercer lugar a Riosucio, este gráfico facilita la comparación directa, clara y jerárquica y podría servir para analizar patrones geográficos y desigualdades sociodemográficas.

1.15 Mapas:

Este mapa de calor y con funciones geográficas nos facilita comprender la ubicación del departamento de Caldas respecto a Colombia y al mundo, nos indica una representación espacial con el total de los nacimientos que se han dado desde el 2010, esto podría usarse en el ejercicio de conocer la cantidad de nacimientos de los demás departamentos de Colombia, lo que indicaría áreas críticas, con acciones necesarias en cuanto sobrepoblación, o determinar patrones de comportamiento regionales en distintas áreas geográficas como la densidad de población, el acceso a servicios de salud, o características socioeconómicas específicas de la región.

1.16 Gráfica de barras apiladas:

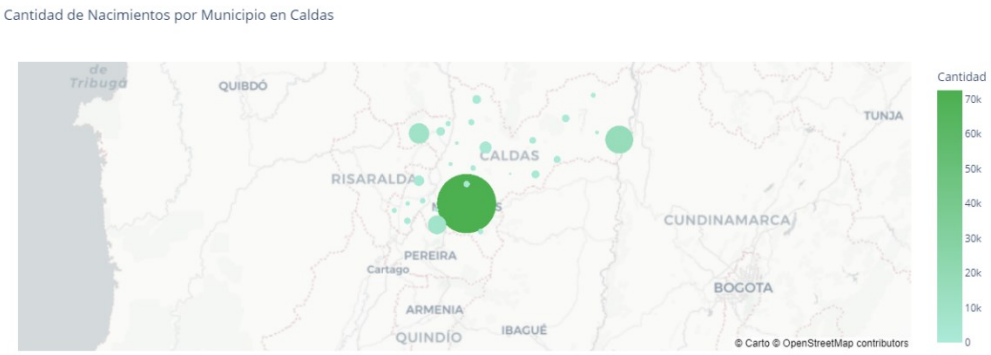


Para este caso usamos una tabla cruzada para grafica el tipo de parto por sexo del nacido, logrando entender un poco la relación entre el sexo y el tipo de parto, donde podemos observar que la mayoría de parto han sido espontáneos haciendo referencia a la manera natural de nacimiento, seguidos de cesárea y por último con un parto instrumentado, en cada tipo de parto, la distribución de sexo está equitativa podríamos decir, aunque con pequeñas diferencias donde el sexo masculino predomina.

1.17 Nube de palabras:

En esta nube de palabras aplicamos una máscara con la silueta del departamento de Caldas y grafiqué los nombres de los municipios con mayor frecuencia de nacimientos, esto facilita la interpretación en cuanto a frecuencia con la detección de patrones visuales

1.18 Mapa:



Finalmente, en este mapa de calor, graficamos la cantidad de nacimientos por los 27 municipios usando la latitud y longitud de la base de datos del DANE y con ello usamos las burbujas para representar la frecuencia de nacimientos en Caldas, todo de manera interactiva usando el tamaño y color de cada punto como indicador de concentración.

2. Visualización personalizada

La visualización personalizada nos ayuda significativamente a comprender mejor los datos sobre los nacimientos en el departamento de Caldas ya que adapta la presentación de los datos a las respuestas que buscamos encontrar, por ejemplo, permite resaltar aspectos únicos de la región, como la distribución geográfica o las tendencias demográficas locales de los municipios, también facilita la comprensión más profunda y rápida de la información, especialmente para aquellos que no son expertos en análisis de datos.

Por otro lado, la visualización personalizada puede enfocarse en indicadores clave o en la búsqueda de insights específicos, como la búsqueda de la tasa de natalidad por municipio o tendencias en el tipo de parto, lo que puede ayudar a organismos de control a tomar decisiones más acertadas basadas en datos.

También es importante mencionar que las visualizaciones personalizadas pueden adaptarse a la audiencia específica, ya sea personal médico, funcionarios públicos o la comunidad en general. Esto asegura que la información se presente de la manera más clara y relevante para cada grupo.

Por último, esta práctica nos permite comparar múltiples variables a nuestro interés y simultáneamente tal como lo evidenciamos en el respectivo ejercicio y además, pueden proporcionar tanto visiones generales como detalles específicos, permitiendo a los usuarios explorar los datos desde diferentes perspectivas.

3. Referencias

Datos Abiertos Gobierno de Colombia. (2024). Nacimientos por residencia departamento de Caldas [Data set]. Recuperado de: https://www.datos.gov.co/Salud-y-Protecci-n-Social/Nacimientos-por-residencia-departamento-de-Caldas/p3zy-i3aq/about_data

DANE. (2005). Tabla de municipios. Recuperado de: https://www.dane.gov.co/files/censo2005/provincias/subregiones.pdf