

Visualización para grandes volúmenes de datos

Mario Andrés Hernández Moreno

Actividad 4: Exploración de Software y Librerías para la Visualización de Grandes Volúmenes de Información

Mag. Mario Alejandro Bravo Ortiz

Universidad Autónoma de Manizales, Manizales
Especialización en Inteligencia Artificial

2024

Contenido

1. Nacimientos por residencia departamento de Caldas.....	3
2. Elementos interactivos:	5
2.1 Filtros:	5
2.2 Tarjeta:	5
2.3 Gráficas de caja:.....	5
2.4 Gráficas de dispersión:.....	6
2.5 Gráfica de violín:	7
2.6 Gráficas de áreas:	8
2.7 Gráfica de columnas agrupadas y de líneas:	8
2.8 Mapa de correlación entre variables:	9
2.9 Gráfica circular:	10
2.10 Gráficas de anillos:	11
2.11 Gráfica de líneas:	12
2.12 Gráfica de barras:	12
2.13 Infografía:	13
2.14 Nube de palabras:.....	14
2.15 Gráfica de barras apiladas:.....	14
2.16 Mapa de calor:.....	15
2.17 Mapa de burbujas:.....	16
3. Dashboard Final:	17
4. Visualización interactiva.....	18
5. Referencias.....	19

1. Nacimientos por residencia departamento de Caldas

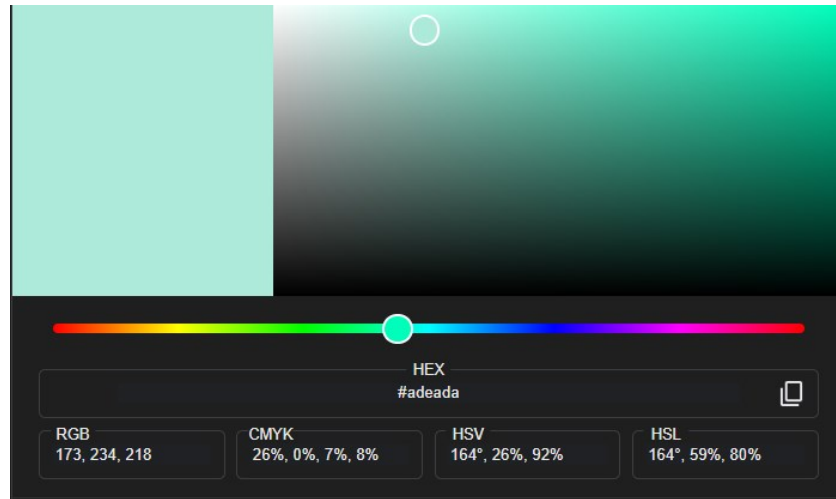


Para la visualización de grandes volúmenes de datos, se hizo uso de la herramienta de Power BI para la creación de un dashboard interactivo que facilite la comprensión de los datos y la toma de decisiones. Se utilizaron distintos tipos de gráficos que posee el visualizador y también se agregaron más objetos visuales a fin de personalizar el cuadro de mando.

En primer lugar se importaron los datos del DataFrame que habíamos preprocesado en Python en la anterior actividad con variables como: 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, Número de hijos nacidos vivos que ha tenido la madre, Municipio, Latitud y Longitud, cada variable debidamente etiquetada, con sus unidades de medida y previamente transformada.



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.



Para trabajar nuestro conjunto de datos fue necesario modificar algunas categorías de datos como el caso de Latitud, Longitud y Municipio para hacer uso del mapa de densidad o calor.



También se transformaron los datos de Mes para poder obtener sus respectivas etiquetas en una nueva columna llamada Mes_Como_Texo haciendo uso de una fórmula DAX.

```
1 Mes_Como_Texo =  
2 SWITCH(  
3     'Nacimientos_por_residencia_departamento_de_Caldas_LIMPIO'[MES],  
4     1, "Enero",  
5     2, "Febrero",  
6     3, "Marzo",  
7     4, "Abril",  
8     5, "Mayo",  
9     6, "Junio",  
10    7, "Julio",  
11    8, "Agosto",  
12    9, "Septiembre",  
13    10, "Octubre",  
14    11, "Noviembre",  
15    12, "Diciembre",  
16    BLANK()  
17 )
```

2. Elementos interactivos:

2.1 Filtros:

The image shows a set of interactive filters for a dashboard. It includes six filter components: 'SEXO' with buttons for 'Femenino', 'Indeterminado', and 'Masculino'; 'MES' with a dropdown menu showing 'Todas'; 'AÑO' with a range selector from 2010 to 2024; 'MUNICIPIO' with a dropdown menu showing 'Todas'; 'TIPO DE PARTO' with a dropdown menu showing 'Todas'; and 'AREANAC' with a dropdown menu showing 'Todas'.

Para este dashboard utilizamos algunos filtros para poder analizar desde distintos puntos de vista la información que teníamos, se implementaron diferentes tipos de segmentación para las variables de Sexo, Mes, Año, Municipio, Tipo de Parto y Área de nacimiento, al ajustar estos filtros, los gráficos se actualizan dinámicamente para mostrar información específica sobre el rango seleccionado, lo que facilita el análisis detallado de los datos.

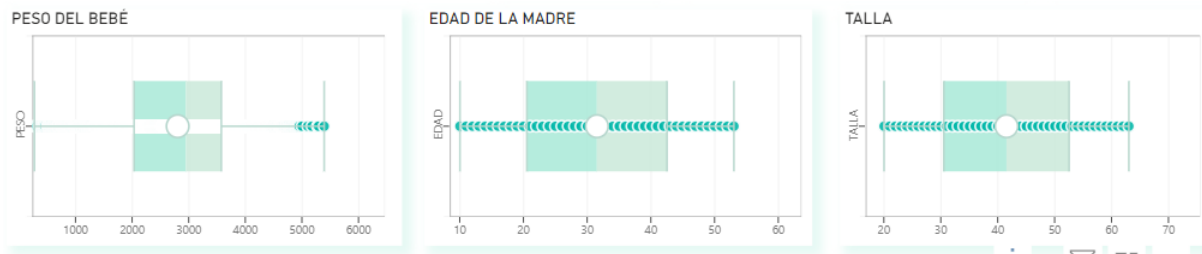
2.2 Tarjeta:



Para este tablero empleamos una tarjeta dinámica que nos muestra el número total de nacimientos a medida que se interactúa con los filtros o demás gráficos.

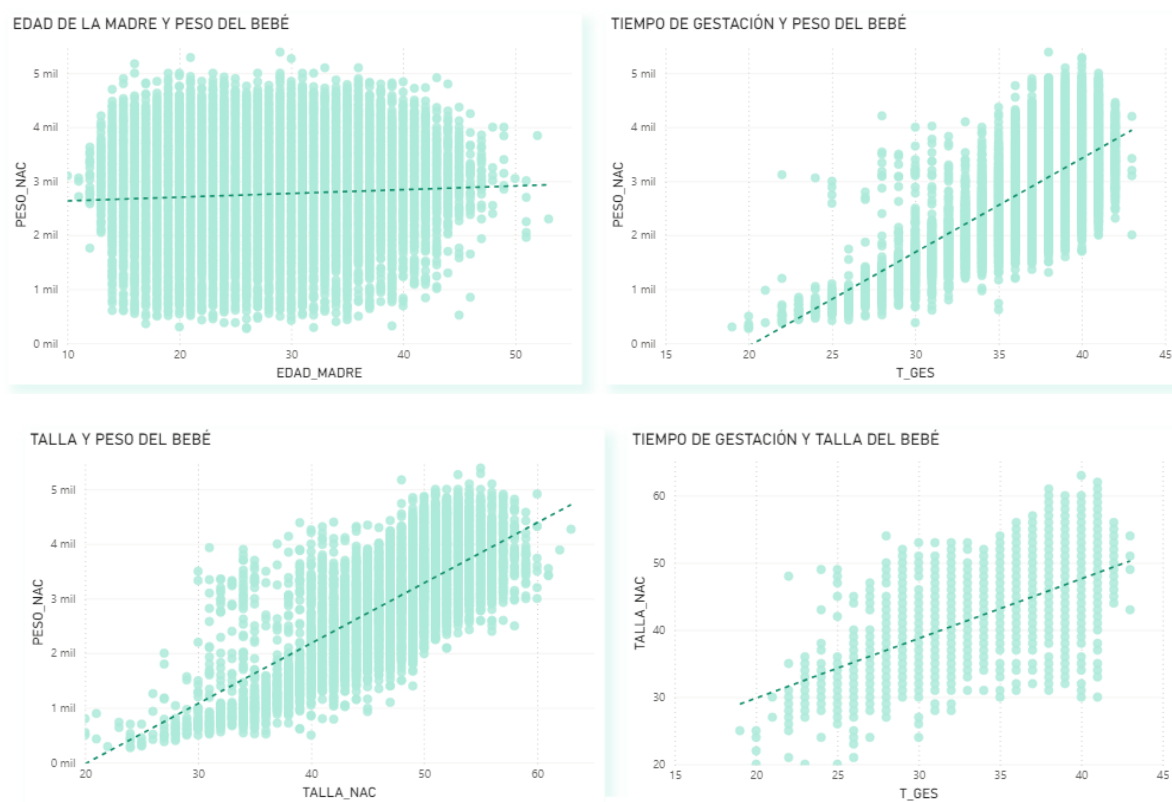
2.3 Gráficas de caja:





Estos gráficos 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.

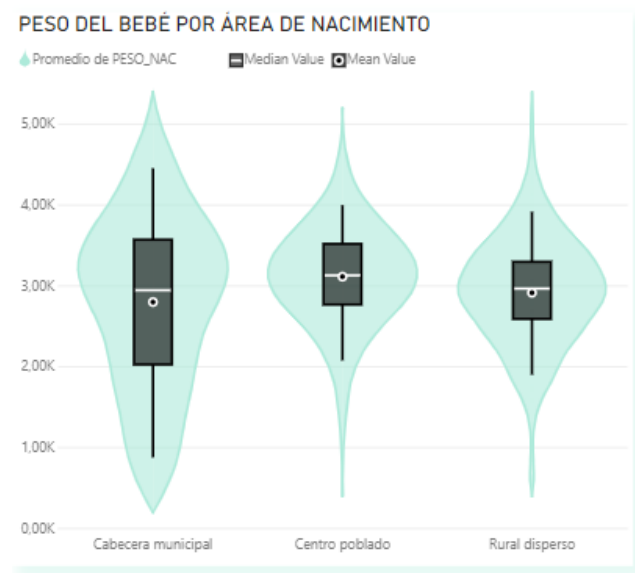
2.4 Gráficas de dispersión:



Para este caso utilizamos gráficos de dispersión para entender las relaciones entre algunas variables como la edad de la madre y el peso del nacido, donde parece existir una leve relación directa pero que no es muy clara, sin embargo, esto nos ayuda a analizar la variabilidad de los

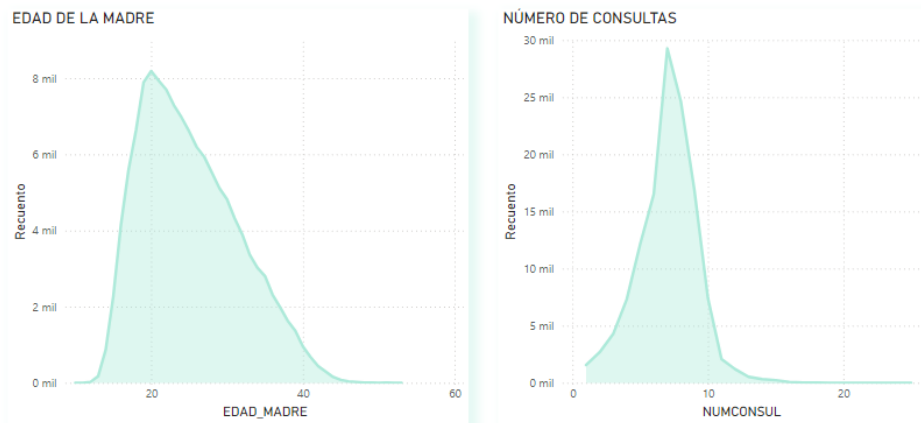
datos. Por otro lado, las relaciones entre semanas de gestación con el peso del bebé y la talla, indican una relación positiva lo que indica que, a mayor tiempo de gestación, mayor es el peso del bebé y también la talla. Esto también nos indica que los bebés prematuros suelen tener pesos o tallas considerablemente más bajos, pero a partir de cierto tiempo de gestación, el aumento del peso y de la estatura podría desacelerarse. 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.

2.5 Gráfica de violín:



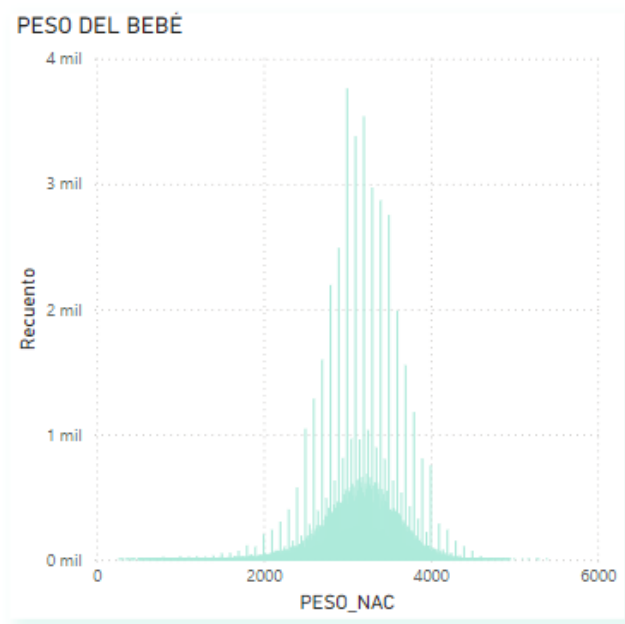
Para esta gráfica fue necesario instalar el objeto visual llamado Violín plot (1.3.0) con el que se logró relacionar el peso del bebé con el área de nacimiento y comparar cómo varían estas variables dependiendo del lugar donde nacen los bebés, lo que nos ofrece una representación detallada de la distribución 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.

2.6 Gráficas de áreas:



En este caso comparamos la distribución o rango de edad de la madre y del número de consultas de las gestantes, lo que 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, mientras que para el número de consultas el punto de mayor concentración apunta hacia 7 números de consultas en promedio antes del parto.

2.7 Gráfica de columnas agrupadas y de líneas:



Esta gráfica 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)

2.8 Mapa de correlación entre variables:



Para elaborar esta gráfica fue necesario utilizar el objeto visual de Python mediante un script para seleccionar las variables a relacionar y graficar con el mismo código que se implementó en la anterior actividad.

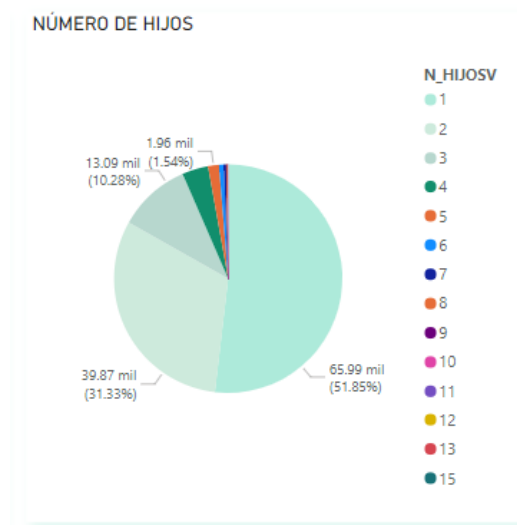
```
# Calculamos la matriz de correlación
correlation_matrix = df.corr()

colors = ["#FFFFFF", "#ADEADA"] # De blanco a verde agua
cmap = LinearSegmentedColormap.from_list("custom_cmap", colors)

# Creamos el mapa de calor
plt.figure(figsize=(12, 10))
sns.heatmap(correlation_matrix, annot=True, cmap=cmap, fmt=".2f", cbar_kws={'shrink': 0.8}, annot_kws={"size": 8})
plt.title('Mapa de Correlación entre variables')
plt.show()
```

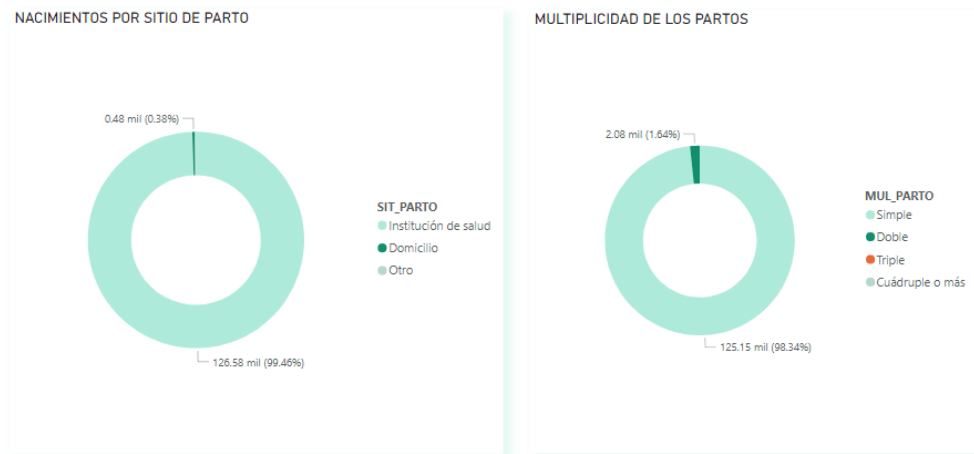
El mapa de correlación nos ayuda a identificar múltiples relaciones entre las variables que maneja nuestro modelo como lo podemos evidenciar con correlaciones positivas entre en la talla del recién nacido y su peso, el tiempo de gestación y el peso al nacer, al igual que con la talla, y otras en menor 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.

2.9 Gráfica circular:



En este caso podemos visualizar la cantidad de hijos que han tenido las gestantes y entender mejor su proporción, donde el 51.85% sólo ha tenido un solo hijo incluido el presente, el 31.33% ha tenido 2 hijos, el 10.28% ha tenido 3 hijos y así respectivamente hasta llegar al caso de una madre con 15 hijos en Manzanares.

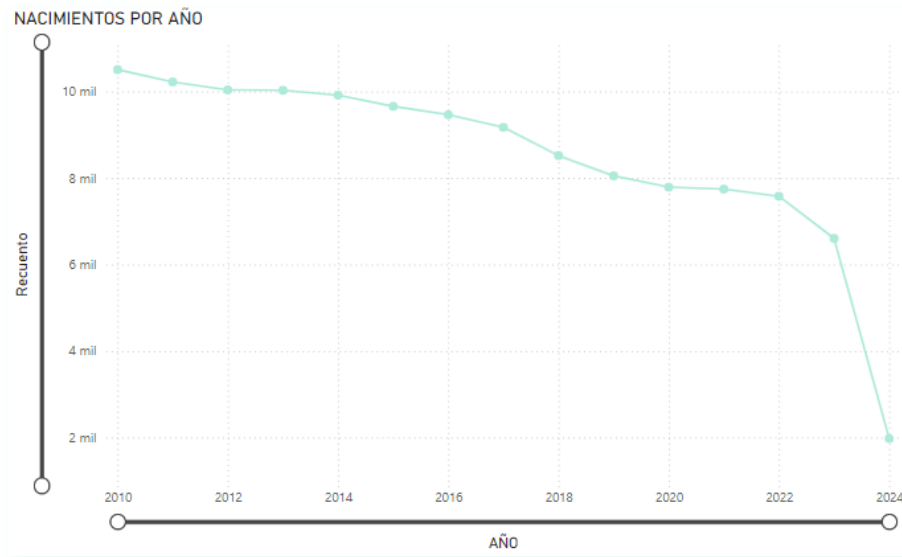
2.10 Gráficas de anillos:



En los gráficos de anillos tenemos la distribución del sitio del parto, este gráfico es ideal porque nos muestra cómo se distribuyen 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 99.46% el resto se distribuye entre el domicilio y otros lugares no especificados en el modelo.

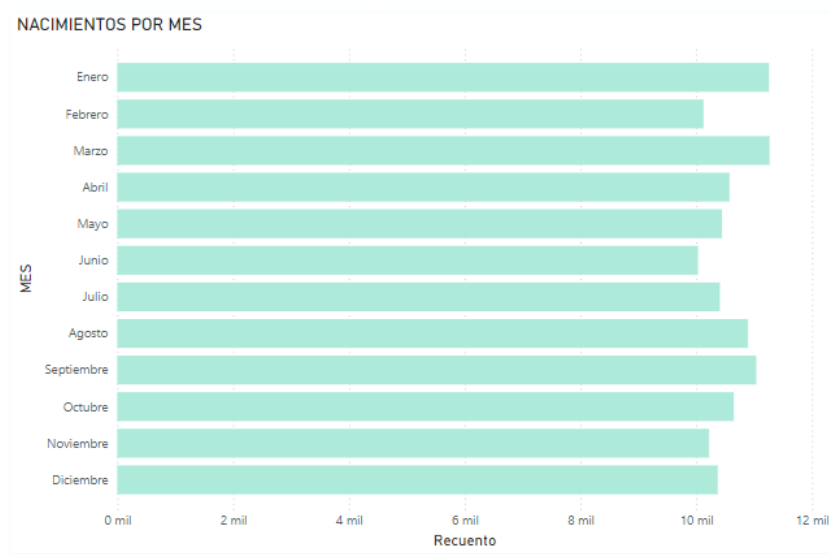
Por otro lado, de manera interactiva podemos analizar la distribución de la multiplicidad del embarazo, es una forma clara y directa de mostrar cómo se distribuyen los diferentes tipos de embarazo dentro de un total, siendo el 98.3% (125151) de embarazos de tipo simple, el 1.64% (2083) de tipo doble, 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.

2.11 Gráfica de líneas:



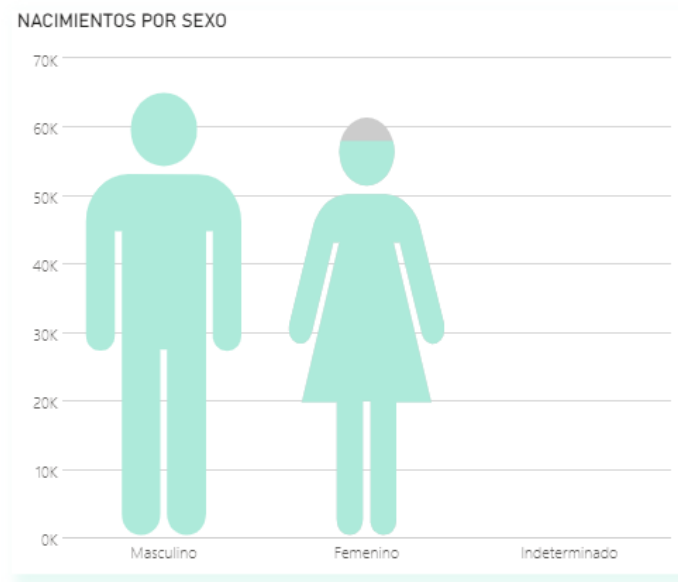
Para este caso utilizamos un gráfico de líneas donde podemos ver la evolución de la cantidad de nacimientos por años desde el 2010, con una disminución considerable, además se le agregó al gráfico un control deslizante, esto nos facilita identificar tendencias temporales, la progresión, las fluctuaciones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

2.12 Gráfica de barras:



Para esta gráfica podemos comparar de manera visual, dinámica 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.

2.13 Infografía:



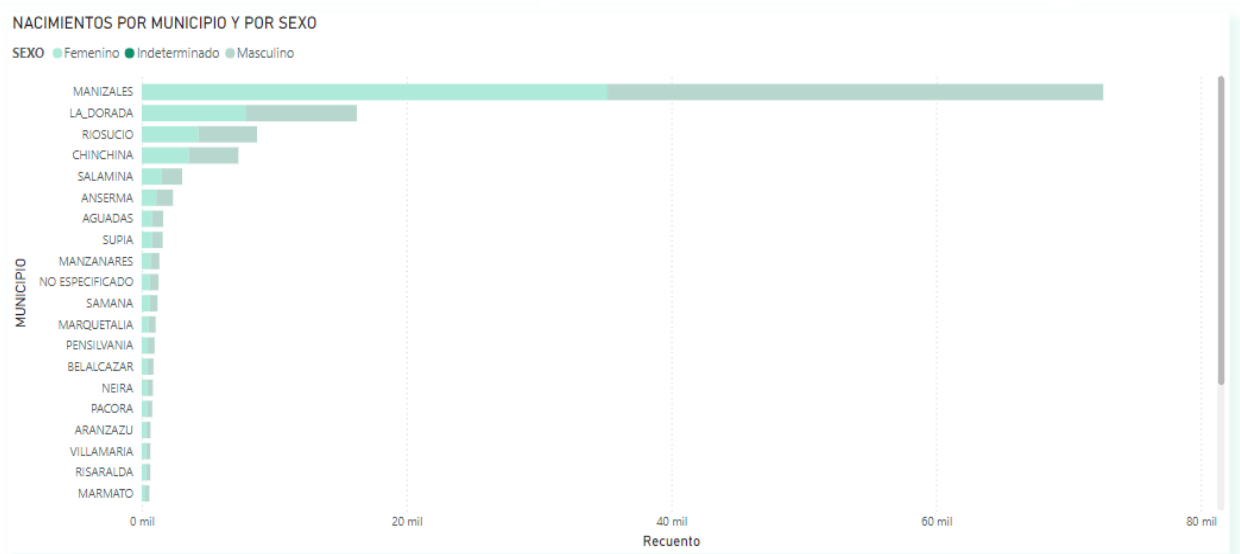
Este gráfico hace uso del objeto visual Infographic Designer 1.9.7 y nos permite personalizar la visualización con íconos relacionados con los nacimientos por sexo, facilitando la comprensión de los datos y dando a conocer que han nacido más de 65.000 bebés de género masculino y un poco más de 61.000 bebés de género femenino.

2.14 Nube de palabras:



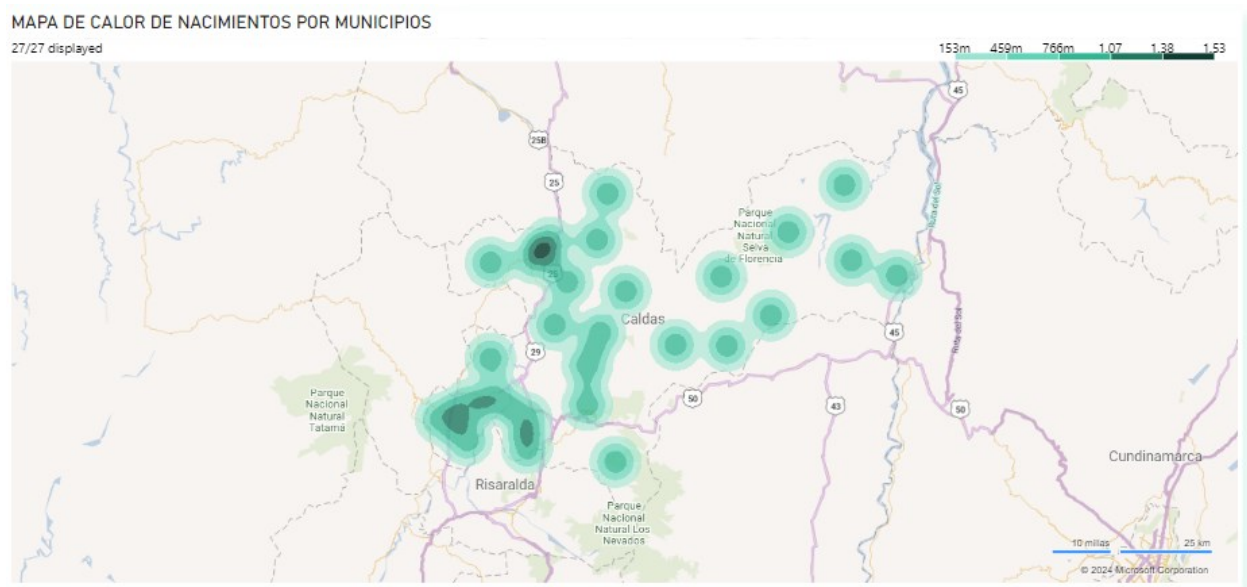
En este caso se utilizó el objeto visual Word Cloud by Powerviz donde aplicamos una máscara con la silueta del departamento de Caldas y graficamos los nombres de los municipios con mayor frecuencia de nacimientos, esto facilita la interpretación en cuanto a frecuencia con la detección de patrones visuales.

2.15 Gráfica de barras apiladas:



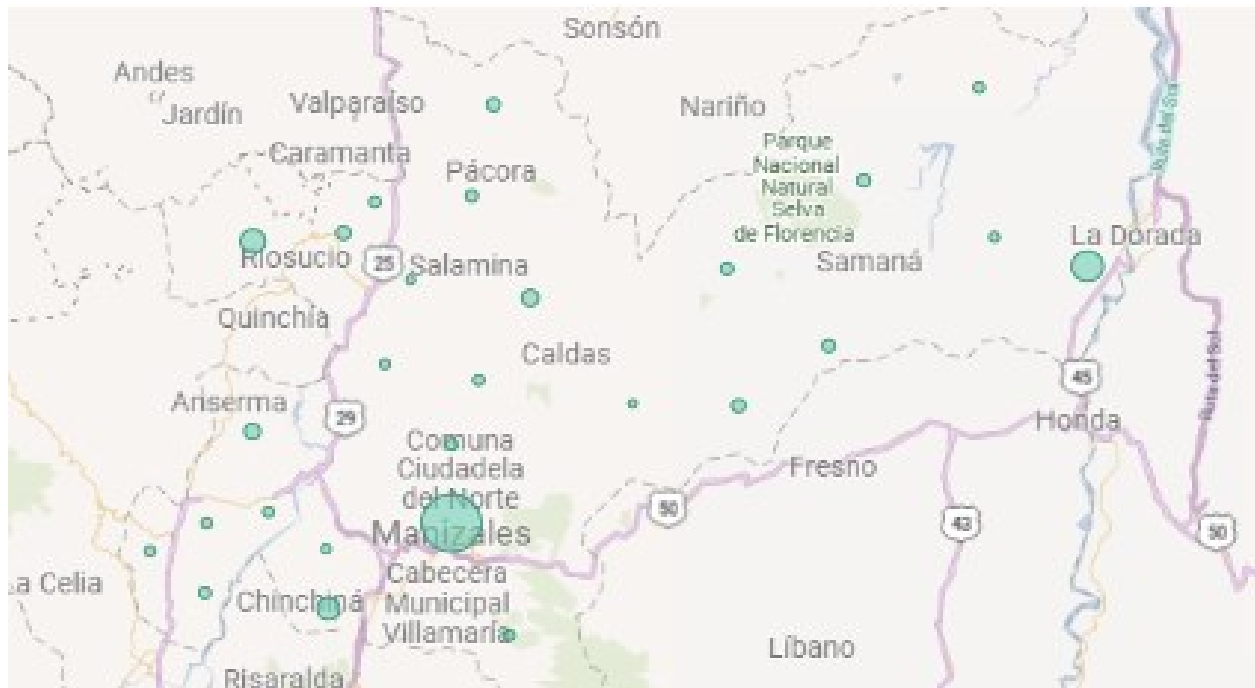
Para la gráfica de barras apiladas relacionamos la cantidad de nacimientos por municipios con el sexo de los recién nacidos, lo que facilita comparar la diferencia de nacimientos en cada lugar, siendo Manizales el municipio donde más nacen bebés desde el 2010 quizá respaldado por la concentración de 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.

2.16 Mapa de calor:



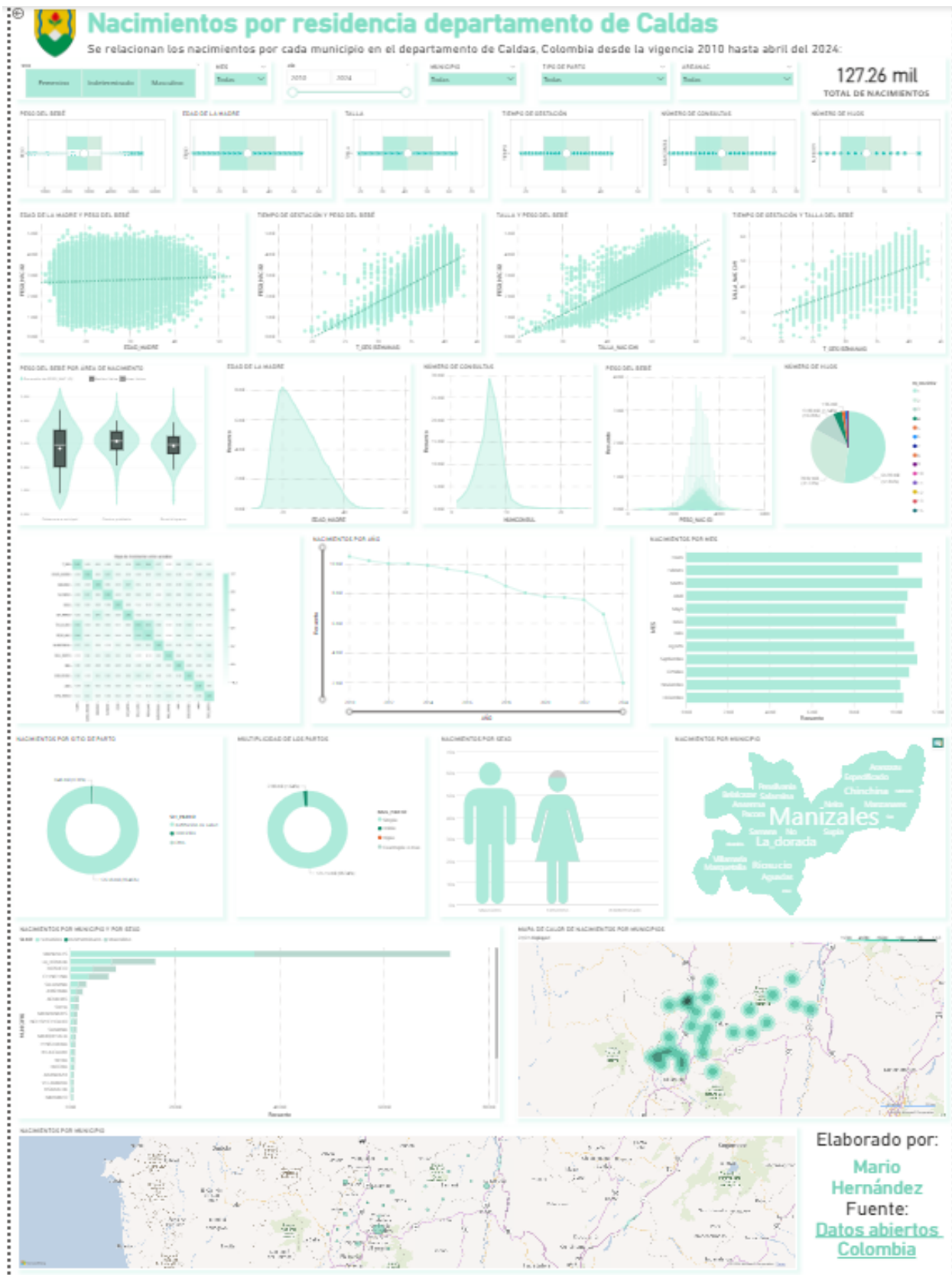
Este mapa de calor funciona gracias al recurso de Heat Map de Microsoft Corporation y los valores de la Latitud y la Longitud, funciones geográficas que nos facilitan comprender la ubicación del departamento de Caldas y de cada uno de los 27 municipios respecto a Colombia y al mundo, en el mapa podemos observar la concentración de nacimientos por cada municipio y diferenciar gráficamente su densidad.

2.17 Mapa de burbujas:



Finalmente realizamos el mapa de burbujas para representar de otra manera la cantidad de nacimientos por nacimientos en cada municipio de Caldas.

3. Dashboard Final:



4. Visualización interactiva

Los elementos interactivos del dashboard tienen un papel importante para facilitar la comprensión, el análisis y el storytelling de los datos, pues permite al usuario ver valores precisos y detalles adicionales al mover el cursor sobre los puntos de datos, también facilita el filtrado de la información de acuerdo a sus intereses donde el usuario puede resaltar datos relacionados en otros, permitiendo explorar correlaciones entre diferentes variables o responder a preguntas específicas.

Otra característica del dashboard es que a medida que los usuarios interactúan con las diferentes selecciones, los gráficos se actualizan en tiempo real, proporcionando retroalimentación visual inmediata sobre cómo diferentes factores impactan los datos, que se complementan con las leyendas dinámicas que aclaran el panorama, lo que facilita la identificación de tendencias, por ejemplo, si se selecciona un municipio en el mapa, el usuario puede ver de inmediato cómo se comportan variables como el peso y la talla de los bebés o la edad de las madres en ese lugar específico pero si el usuario lo desea también puede analizar cada gráfico por separado y simplificar la cantidad de información a su conveniencia.

El uso de estos gráficos interactivos mejora enormemente la comprensión de los datos al permitir una navegación clara, fácil de entender y la exploración de información a profundidad. Los usuarios pueden ajustar los filtros para personalizar su análisis y utilizar los gráficos interconectados para identificar patrones, correlaciones y distribuciones de manera rápida y efectiva. Además, el uso de mapas y visualizaciones intuitivas permite interpretar la información de forma más clara, maximizando la utilidad del dashboard para la toma de decisiones o la investigación.

5. Referencias

Datos Abiertos Gobierno de Colombia. (2024). Nacimientos por residencia departamento de Caldas [Data set]. Recuperado de: https://www.datos.gov.co/Salud-y-Proteccion-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>