Visualización de múltiples distribuciones

Cómo mostrar varios conjuntos de datos al mismo tiempo para compararlos y entenderlos mejor.

Gráficos y Variables

Visualizar múltiples distribuciones, como por ejemplo la variación de temperatura durante los meses, requiere técnicas especiales mas allá de histogramas o gráficos de densidad. Necesitamos métodos que nos puedan mostrar múltiples distribuciones al mismo tiempo, como *boxplots*, *violin plots*, y *ridgeline plots*.

Variable de respuesta

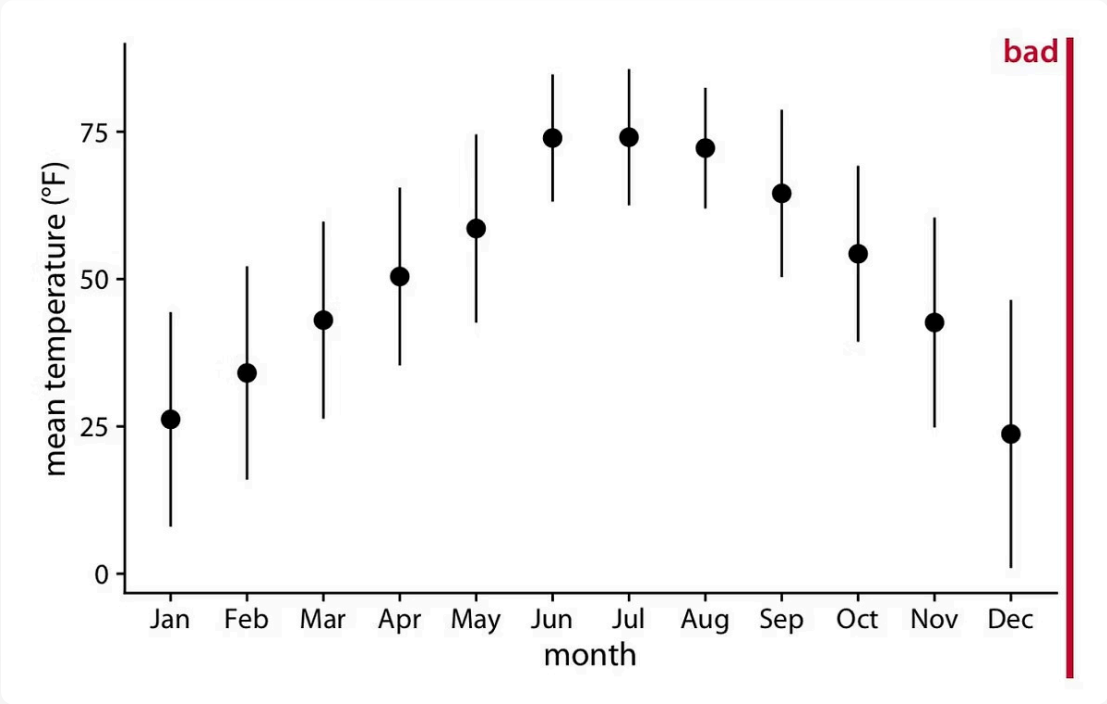
La variable cuya distribución queremos mostrar (ej. temperatura).

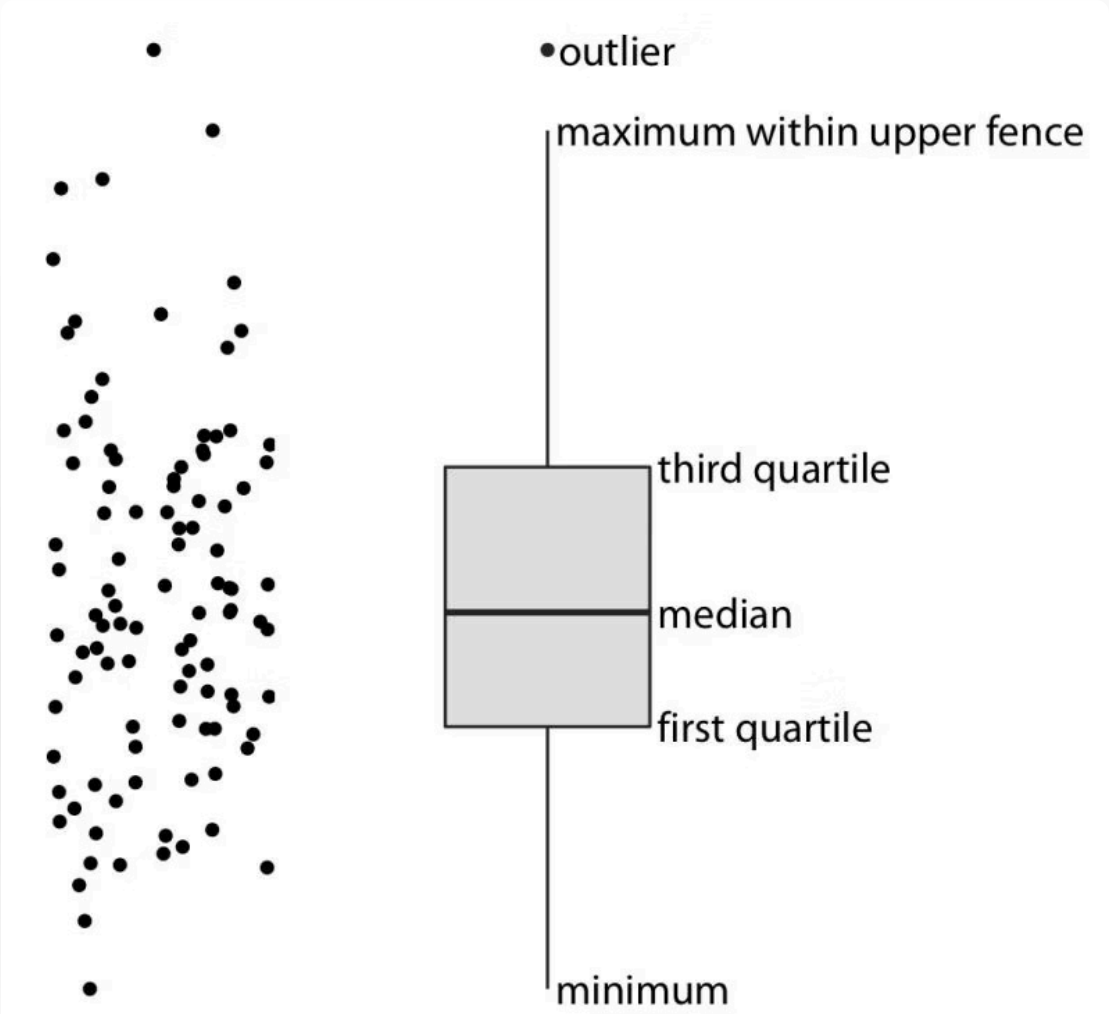
Variable de agrupación

Define cómo se agruparán los datos de la variable de respuesta (ej. meses del año).

Puntos y barras de error

Este gráfico muestra la temperatura en Lincoln, NE, usando puntos para las medias y barras de error para la variación. Este método es problemático y pierde información. No queda claro qué representan las barras: si indican el desvío estándar, el margen de error, las temperaturas máximas y mínimas, u otra medida.





Boxplots

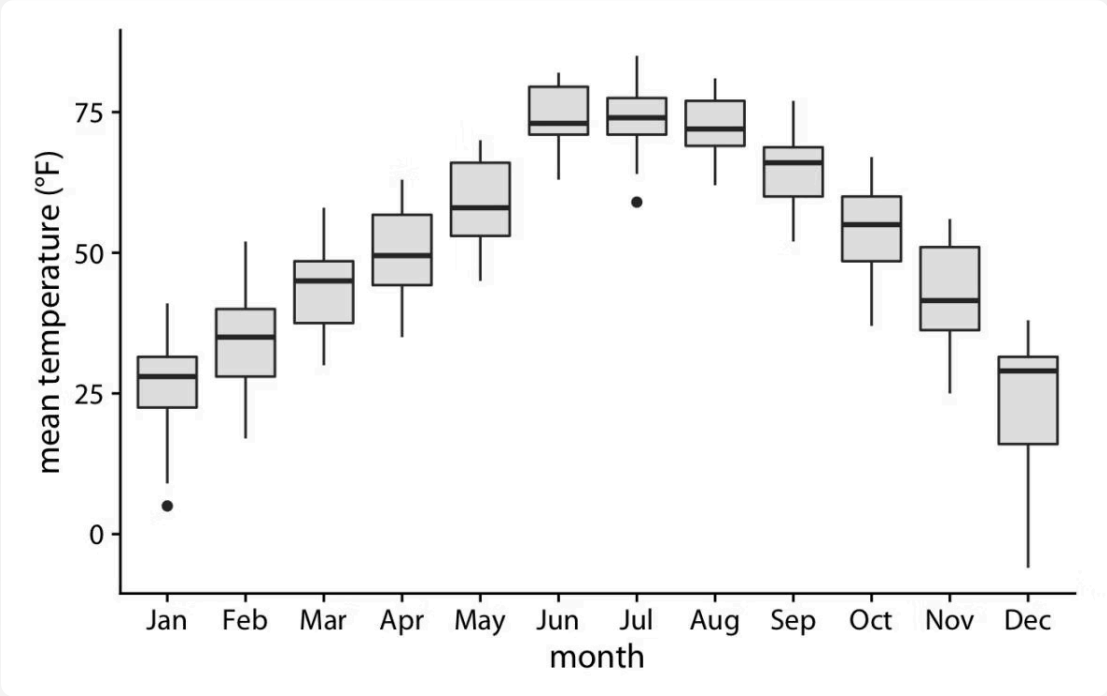
Proveen de una manera clara y estandarizada una forma de visualizar los datos divididos en cuartiles.

También muestran outliers y son fáciles de dibujar a mano.

La caja cubre el 50% de los datos.

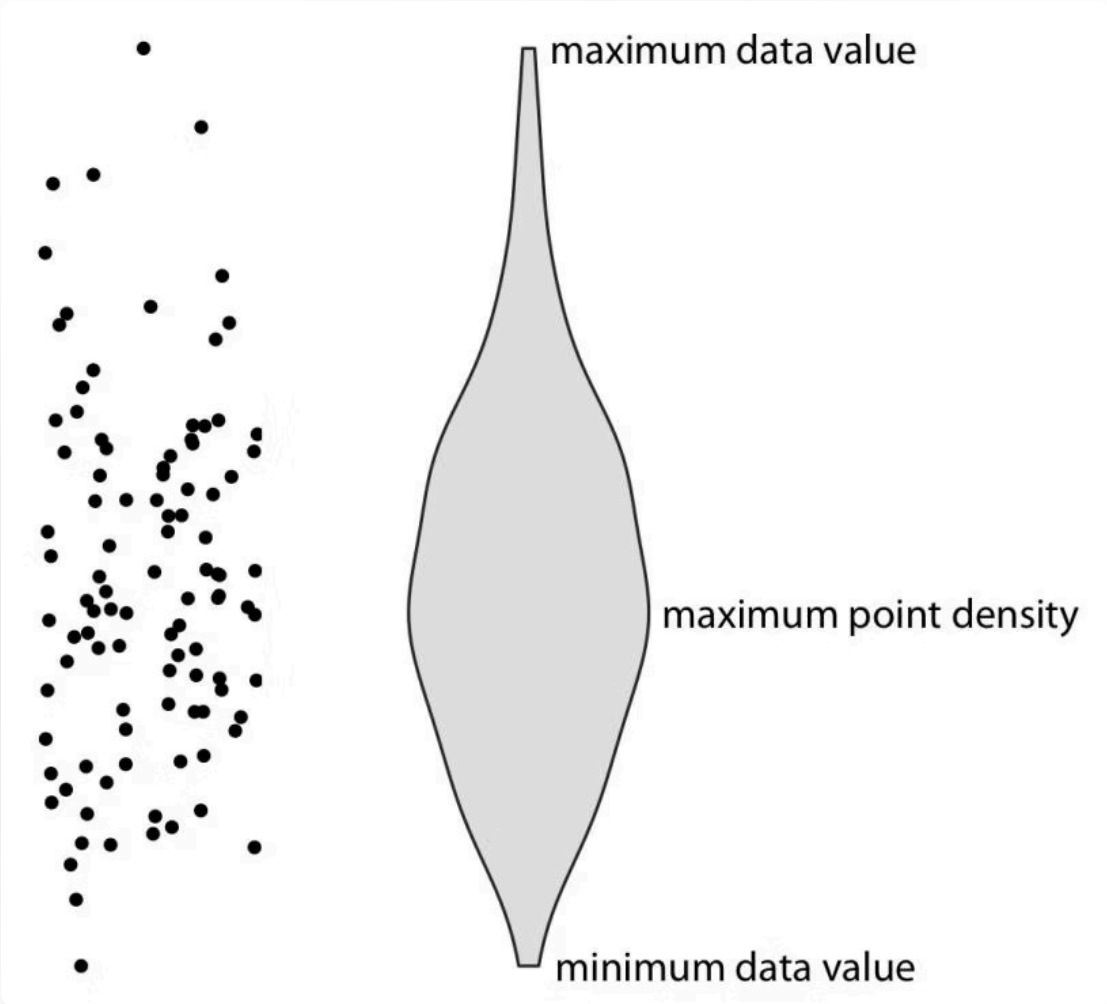
Múltiples distribuciones con *boxplots*

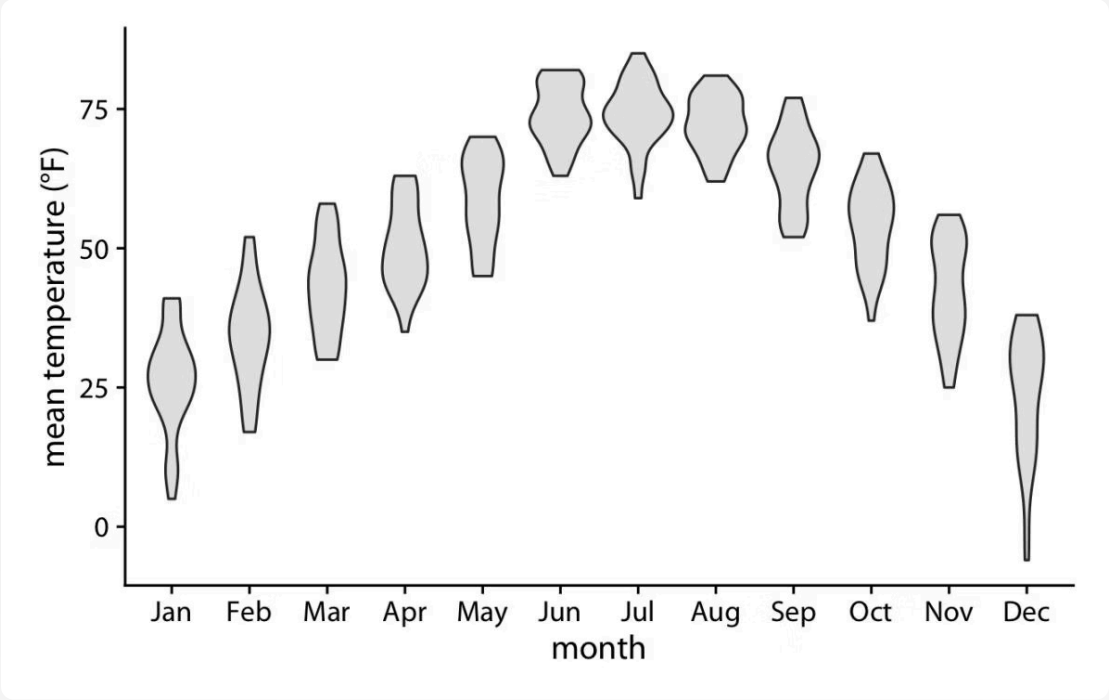
En este gráfico podemos ver la variación de temperaturas a lo largo de los meses. También el sesgo que hay en diciembre y el equilibrio de datos en julio.



Avanzando a *violin plots*

Como ya no es necesario dibujar a mano, hubo un avance hasta estos gráficos que se pueden usar en escenarios similares y proveen mejor detalle sobre la densidad de los datos y distribuciones bimodales.

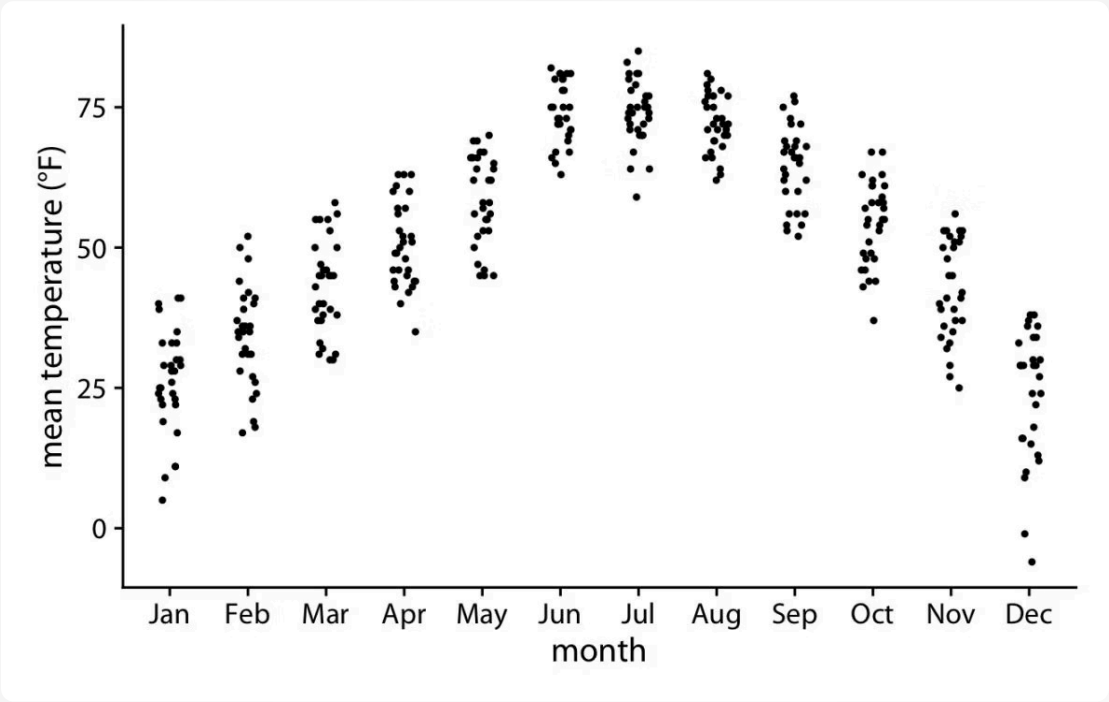




Múltiples distribuciones con *violin plots*

Siguiendo con las temperaturas en Licoln, NE, podemos ver que noviembre tiene dos modas en aprox. 35 y 50 grados Farenheit.

> Antes de usar *violin plots* para visualizar distribuciones, es importante verificar que cada grupo tenga suficientes datos como para justificar la representación de las densidades de puntos mediante líneas suavizadas.”



Strip Charts y *Jittering*

Como los *violin plots* se derivan de estimaciones de densidad, presentan limitaciones similares. En particular, pueden dar la impresión de que existen datos donde en realidad no los hay, o de que el conjunto de datos es muy denso cuando en verdad es bastante disperso.

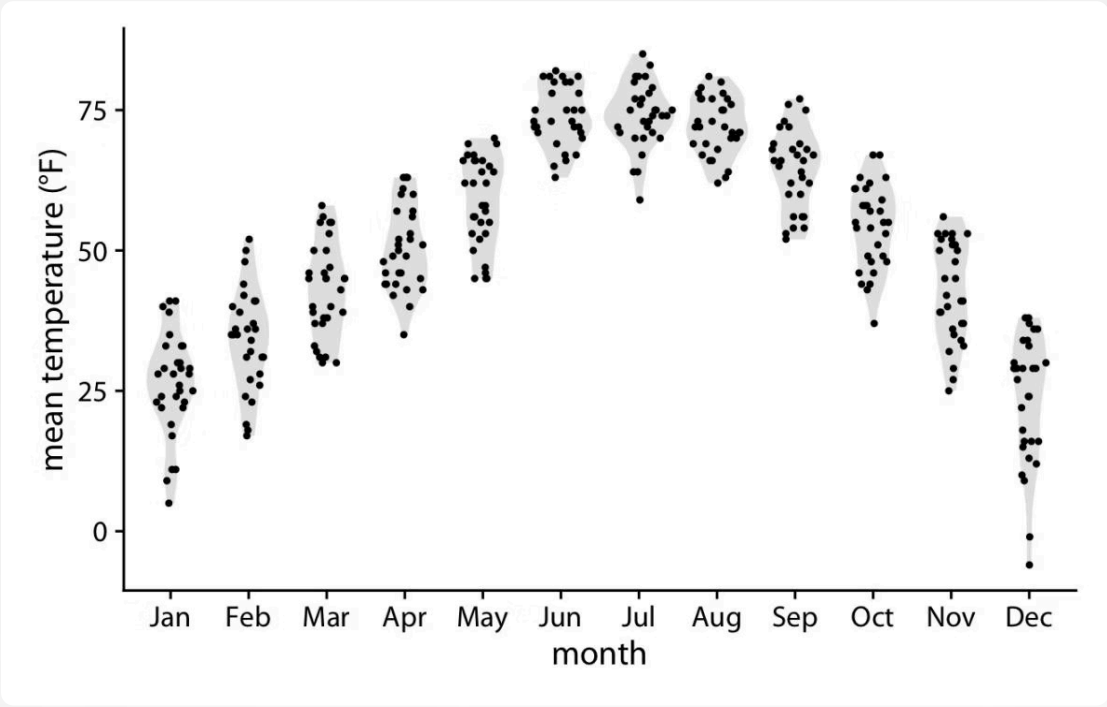
Para evitar estos problemas, podemos simplemente graficar todos los puntos individuales directamente como *dots*. A este tipo de gráfico se lo llama *strip chart*.

Los *strip charts* funcionan bien en principio, siempre que cuidemos de no superponer demasiados puntos. Una solución sencilla al *overplotting* es dispersar los puntos a lo largo del eje x agregando un poco de ruido aleatorio en esa dimensión. Esta técnica se conoce como *jittering*.

Sina plots: un enfoque híbrido

Los gráficos *Sina Plots* combinan lo mejor de los gráficos de violín y los puntos dispersos (*jittered points*). Muestran los datos individuales al mismo tiempo que visualizan la distribución, esparciendo los puntos de manera proporcional a la densidad.

La siguiente figura superpone un gráfico *strip chart* sobre un *violin plot* para ilustrar esta relación.



Ridgeline plots: distribuciones horizontales

Los gráficos de crestas visualizan distribuciones a lo largo del eje horizontal, escalonadas verticalmente. Se asemejan a las líneas de crestas montañosas y son excelentes para mostrar tendencias a lo largo del tiempo, como la duración de las películas o los patrones de votación política.

La figura 9-9 muestra las temperaturas de Lincoln en un gráfico de crestas, lo que hace que los datos bimodales (por ejemplo, los grupos de temperaturas de noviembre) resulten más intuitivos que en los *violin plots.*

En la Figura 9-12 se muestran cómo los patrones de votación en la Cámara de Representantes de EE.UU. se han vuelto cada vez más polarizados. Los puntajes *DW-NOMINATE* se utilizan para comparar los patrones de votación de los representantes entre partidos y a lo largo del tiempo.

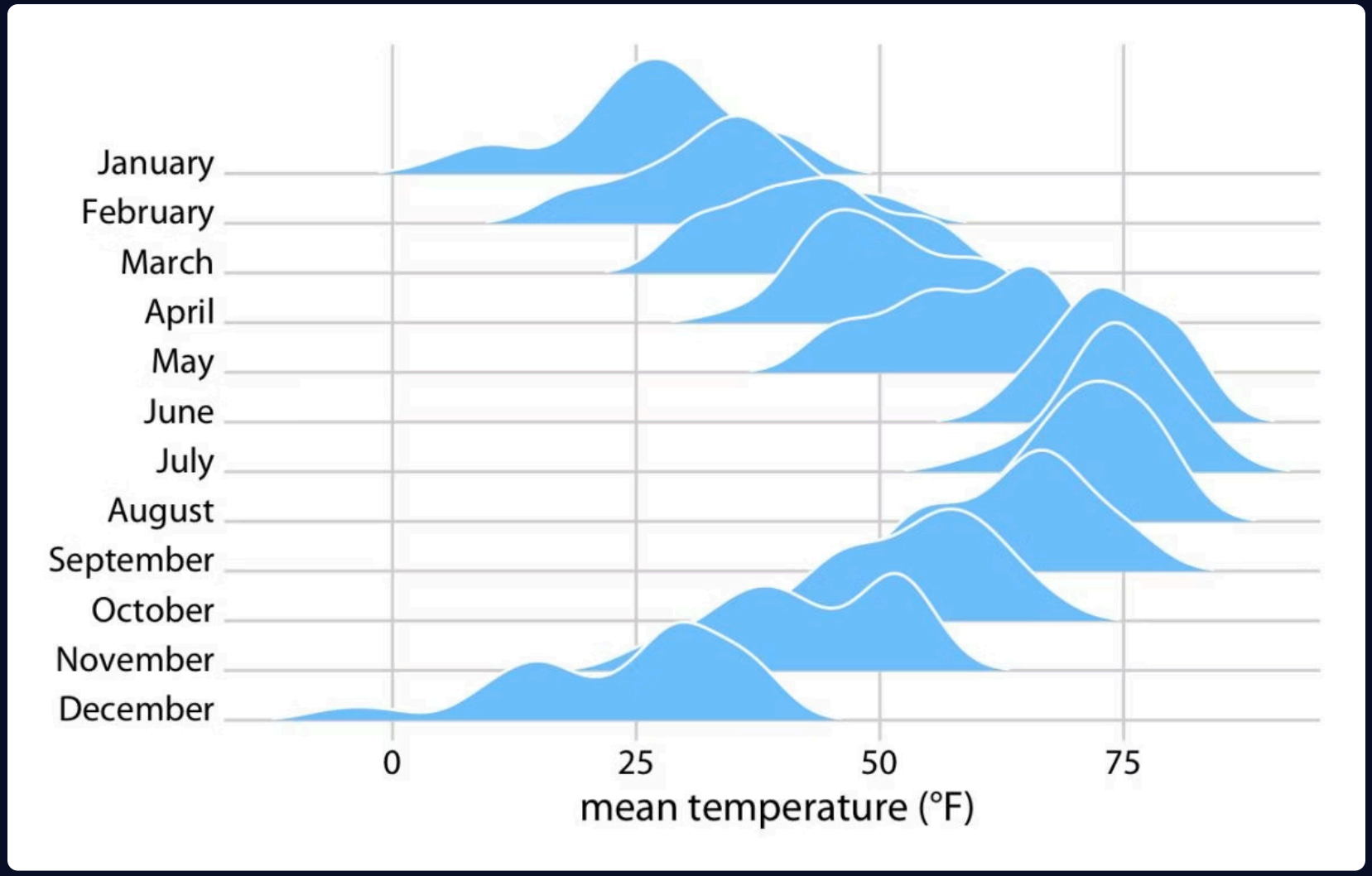


Figura 9-9.

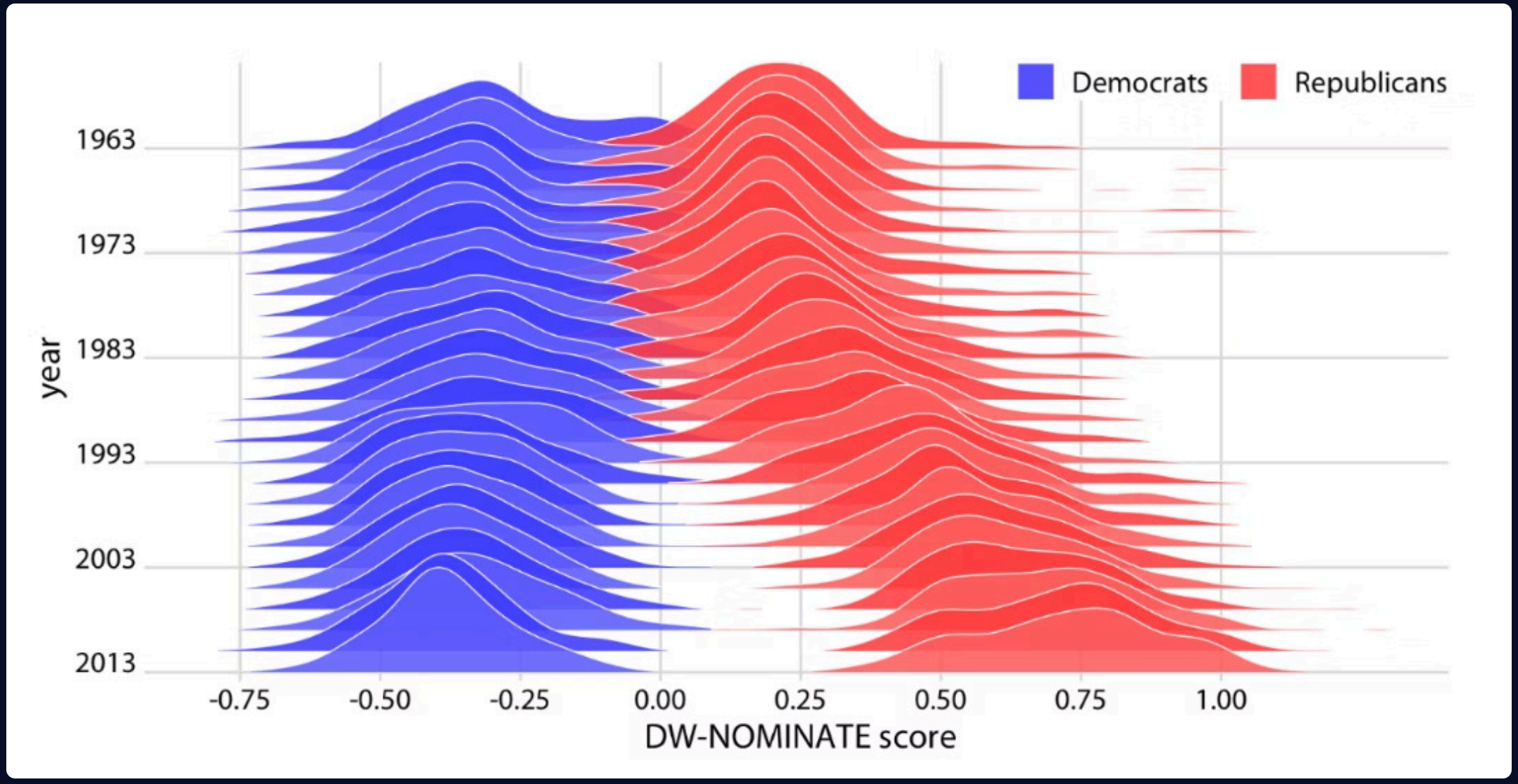


Figura *9-12.*