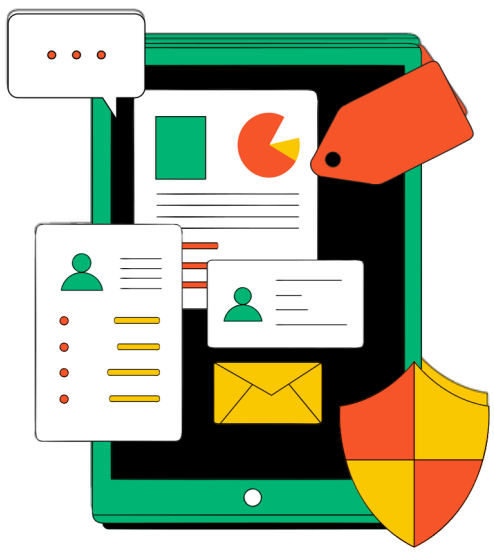
LECCIÓN 1

Lección 1: Visualización de datos

La visualización de datos es un rol importante en el análisis de los datos. A través de los gráficos se puede resumir grandes cantidades de información de forma comprensible para las personas, también es posible determinar qué tendencias tienen los datos y cómo se comportan con solo observar un gráfico. Utilizando Python y algunas de sus bibliotecas es posible crear muchos tipos de visualizaciones para representar la información de manera clara.

La visualización de datos es la representación gráfica de información y datos. Al utilizar elementos visuales como cuadros, gráficos y mapas, las herramientas de visualización de datos proporcionan una manera accesible de ver y comprender tendencias, valores atípicos y patrones en los datos. Nuestros ojos son atraídos por los colores y patrones. Podemos identificar rápidamente el rojo del azul o el cuadrado del círculo. Nuestra cultura es visual, lo que incluye todo tipo de cosas, desde arte y publicidad hasta televisión y películas. La visualización de datos es otra forma de arte visual que capta nuestro interés y mantiene nuestros ojos en el mensaje. Cuando vemos un gráfico, vemos rápidamente las tendencias y los valores atípicos. Si podemos ver algo, lo interiorizamos rápidamente. Es contar historias con un propósito. Si alguna vez haz visto una gigantesca hoja de cálculo de datos y no te fue posible ver una tendencia, sabes cuán eficaz puede ser una visualización.



La visualización es una herramienta cada vez más importante para darle sentido a las billones de filas de datos. A través de la visualización de datos se pueden contar historias mostrando la información de una forma clara, ordenada y fácil de entender en comparación con mostrar tablas de miles o millones de celdas que, para una persona, pueden carecer de sentido. Sin embargo, crear una buena visualización no es tan simple como adornar un gráfico para que se vea mejor o pegar la parte "informativa" de una infografía. La visualización eficaz de datos es un delicado equilibrio entre forma y función. La gráfica más simple podría ser demasiado aburrida para captar la atención del público o lograr que diga algo importante; la visualización más sorprendente podría fallar por completo a la hora de transmitir el mensaje correcto o podría decir mucho. Los datos y los elementos visuales deben trabajar juntos, y hay algo de arte en combinar un gran análisis con una gran narración. Cuando se crean visualizaciones, normalmente se piensa en gráficos de tendencias o de barras, si bien estos son útiles, no son la única forma de representar los datos, hay muchas maneras de crear vistas para enseñar la información.

Entre los tipos de visualizaciones generales tenemos:

Gráficos de líneas

Los gráficos de líneas permiten identificar tendencias en los datos en el transcurso del tiempo. Por ejemplo, el progreso de una tarea, las ventas a lo largo del tiempo, una variable física y muchas variables en general se pueden representar con gráficos de líneas. Por ejemplo, las ventas de viviendas por año pueden tener una visualización como en la figura 1:

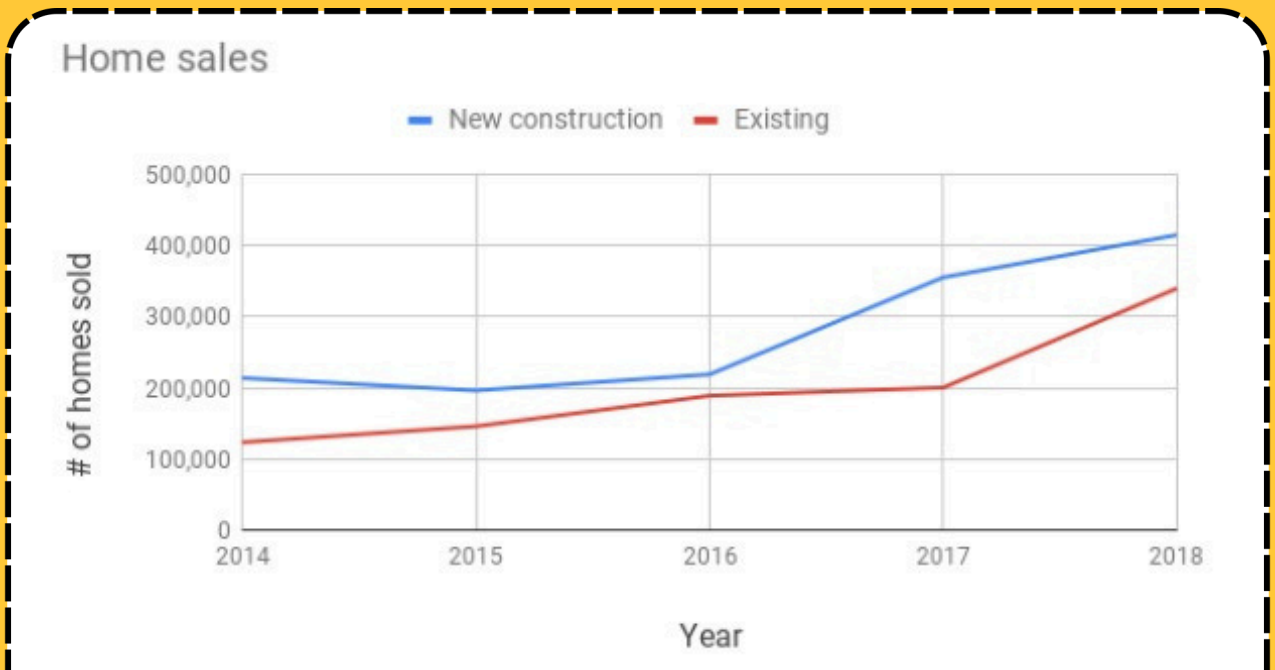


Figura 1 ejemplo de gráfico de líneas

Gráficos de áreas

Los gráficos de áreas permiten identificar tendencias en los datos en el transcurso del tiempo. Son parecidos a los gráficos de líneas, pero se diferencian en que el espacio situado debajo de las líneas está sombreado para que se aprecie mejor la magnitud de las tendencias. En la figura 2 se muestra un ejemplo de un gráfico de áreas.

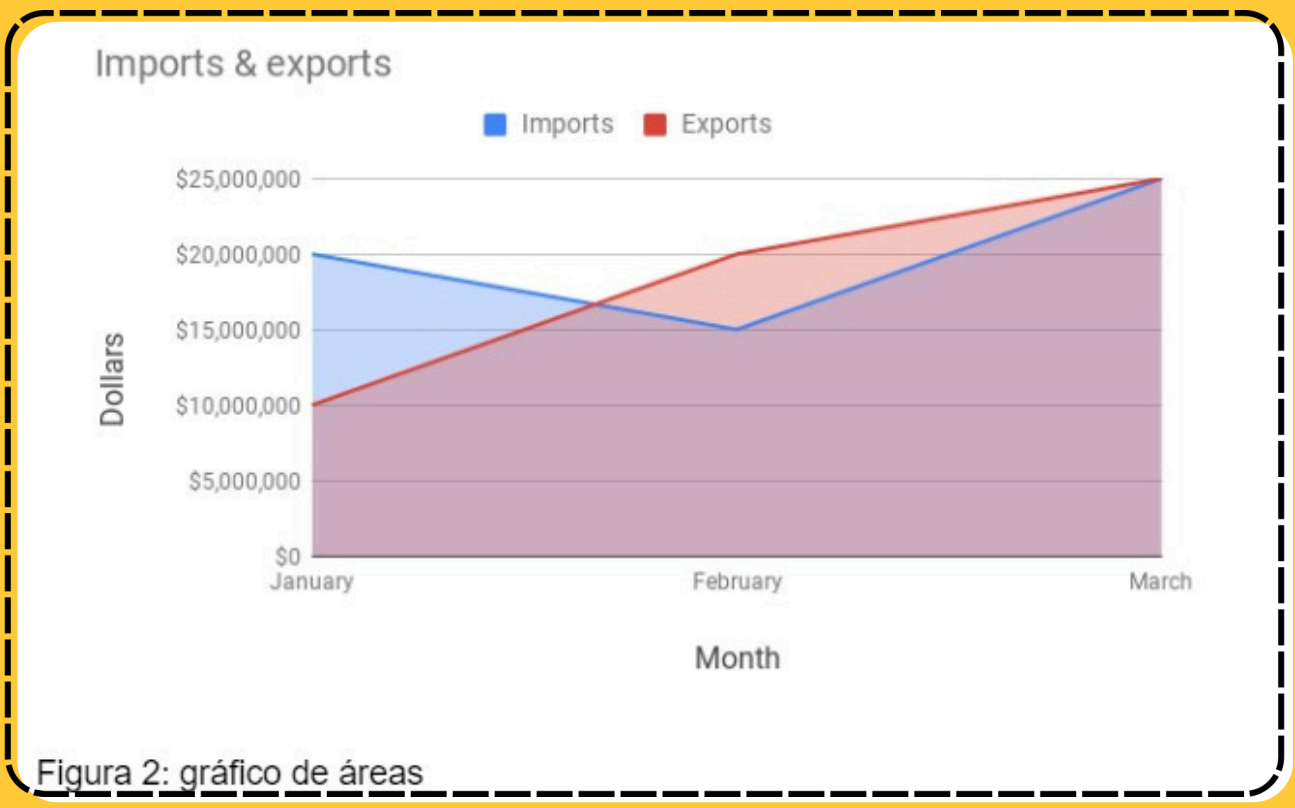


Figura 2: gráfico de áreas

Los gráficos de áreas tienen múltiples variantes, dependiendo de cómo se maneje el espacio. Es posible apilar las áreas para mostrar el progreso en el tiempo de una variable (grafico de línea) y la representación porcentual como un área, como el ejemplo de la figura 3.

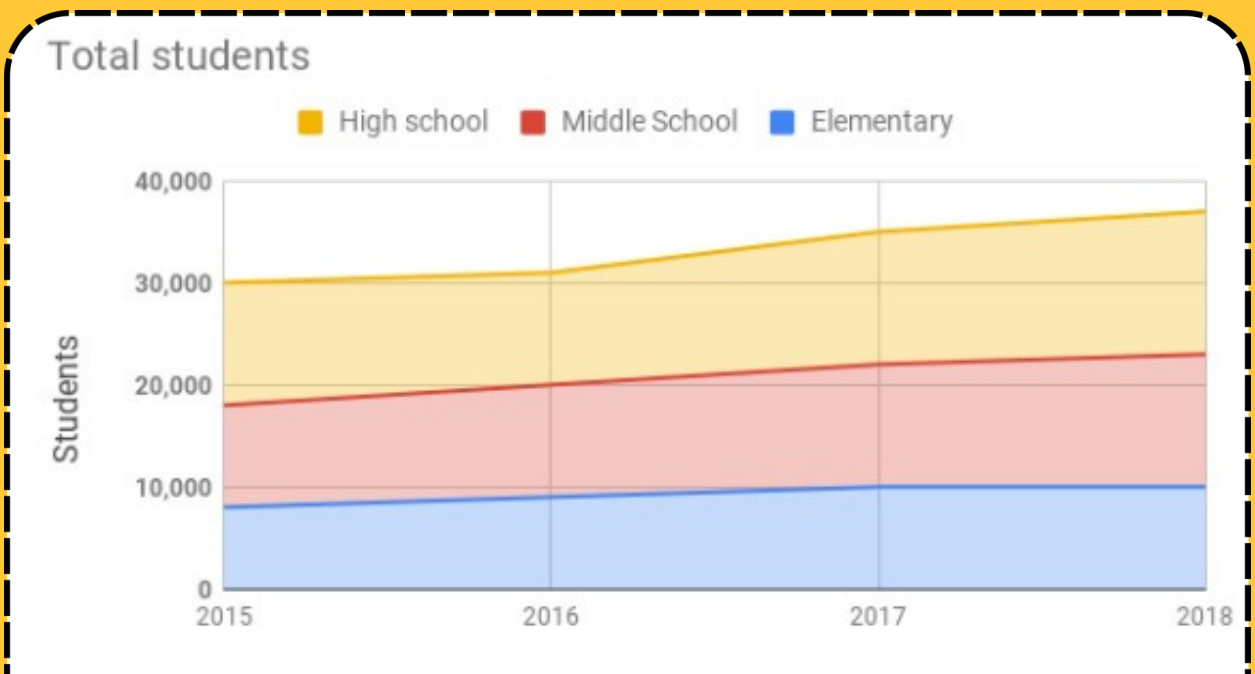


Figura 3: ejemplo de gráfico de áreas apiladas.

Gráfico de barras

Un gráfico de barras es una forma de representar gráficamente datos numéricos mediante rectángulos verticales u horizontales, conocidos como barras. El tamaño de cada barra se ajusta proporcionalmente al valor que representa. Este tipo de gráficos proporcionan una comparación visual de cantidades o frecuencias, lo que facilita la interpretación de los datos. Estos gráficos sirven para comparar cantidades, analizar tendencias y presentar datos. (figura 4)

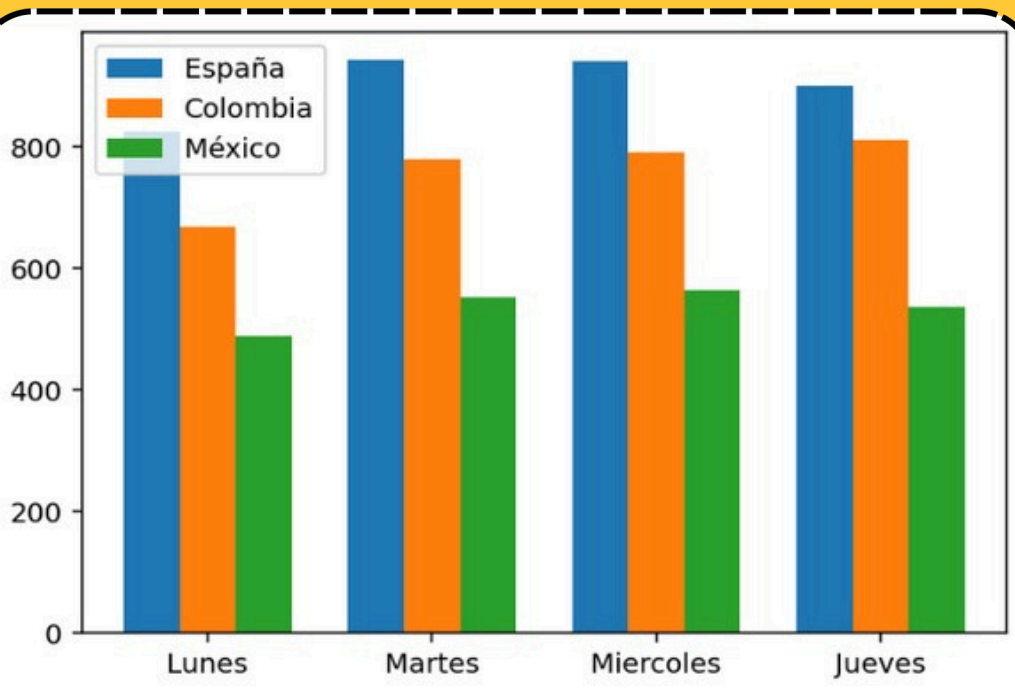


Figura 4: ejemplo de un gráfico de barras verticales.

De forma similar a los gráficos de área, las barras se pueden apilar para representar el porcentaje de los valores dentro de un todo, como es el caso de la figura 5:

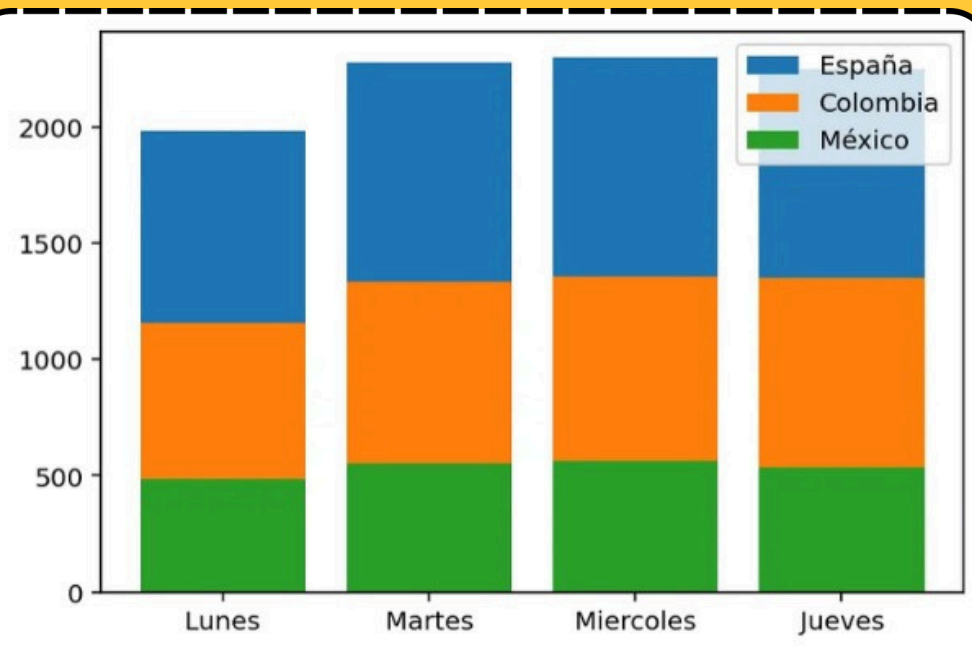


Figura 5: barras apiladas

Diagramas de cajas y bigotes

Un diagrama de cajas y bigotes es una manera conveniente de mostrar visualmente grupos de datos numéricos a través de sus cuartiles. Las líneas que se extienden paralelas a las cajas se conocen como «bigotes», y se usan para indicar variabilidad fuera de los cuartiles superior e inferior. Los valores atípicos se representan a veces como puntos individuales que están en línea con los bigotes. Los diagramas de cajas y bigotes se pueden dibujar vertical u horizontalmente. Normalmente utilizado en estadísticas descriptivas, los gráficos de cajas y bigotes son una excelente forma de examinar rápidamente uno o más conjuntos de datos gráficamente. Aunque parezcan primitivos en comparación con un Histograma o un Gráfico de Densidad, tienen la ventaja de ocupar menos espacio, lo cual es útil cuando se comparan distribuciones entre muchos grupos o conjuntos de datos. En un diagrama de cajas y bigotes se puede apreciar de forma explícita: Cuáles son los valores clave, tales como: el promedio, el percentil 25 medio, etc. Si hay valores atípicos y cuáles son sus valores. Si los datos son simétricos. Cuán estrechamente se agrupan los datos. Si los datos están sesgados y si es así, en qué dirección. En la figura 6 se muestra el contenido de un diagrama de cajas y bigotes.

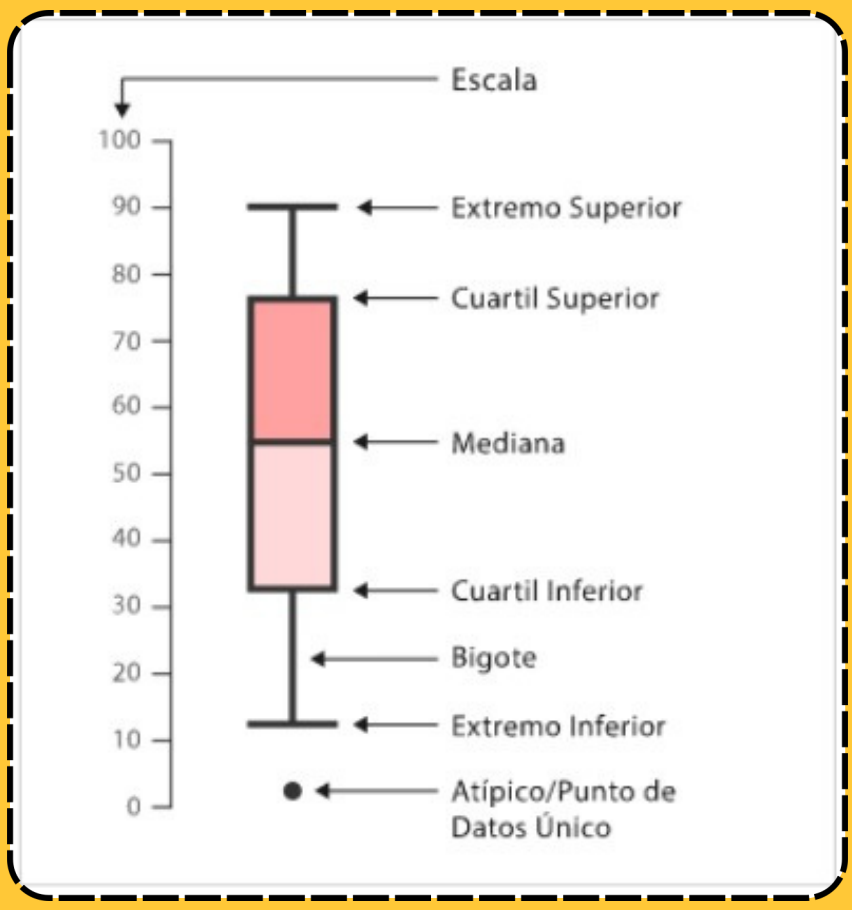


Figura 6: Diagrama de cajas y bigotes

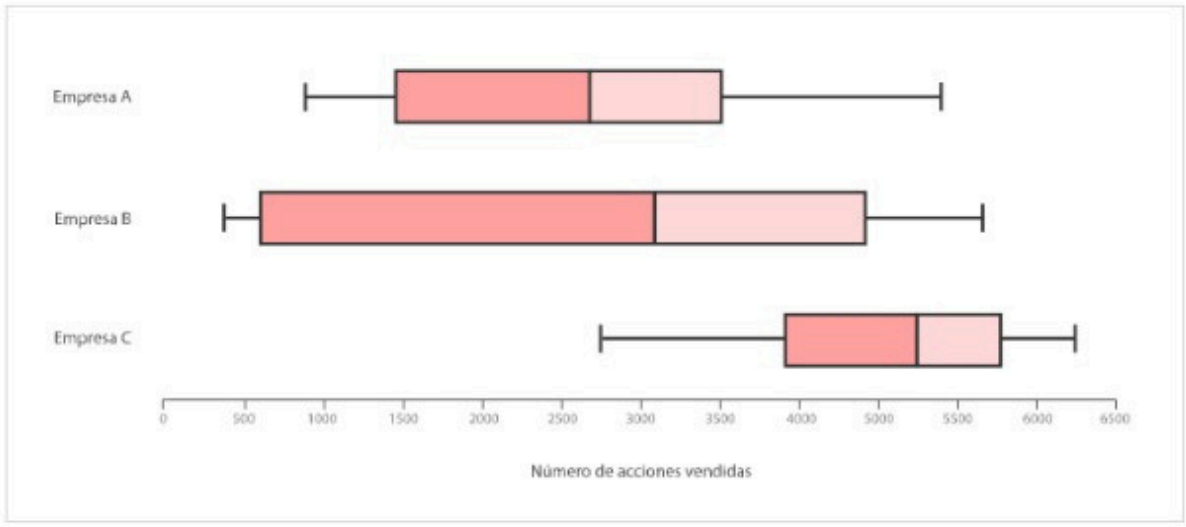


Figura 7: Ejemplo de uso del diagrama de cajas y bigotes.

Diagramas de violín

Un diagrama de violín se utiliza para visualizar la distribución de los datos y su densidad de probabilidad. Este gráfico es una combinación de un diagrama de cajas y bigotes y un diagrama de densidad girado y colocado a cada lado, para mostrar la forma de distribución de los datos. La barra negra gruesa en el centro representa el intervalo intercuartil, la barra negra fina que se extiende desde ella, representa el 95 % de los intervalos de confianza, y el punto blanco es la mediana. Los diagramas de cajas y bigotes están limitados a su visualización de los datos, ya que su simplicidad visual tiende a ocultar detalles significativos sobre cómo se distribuyen los valores en los datos. Por ejemplo, con los diagramas de cajas y bigotes no puedes ver si la distribución es bimodal o multimodal. Si bien los diagramas de violín incluyen más información, pueden estar mucho más abarrotados que los diagramas de cajas y bigotes. En la figura 8 se muestra la descripción y en la figura 9 un ejemplo de uso.

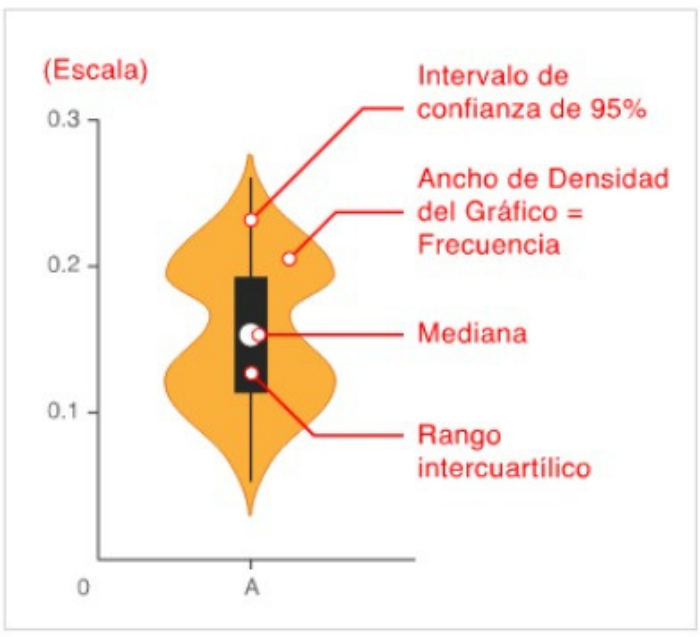


Figura 8: Estructura de un diagrama de violín

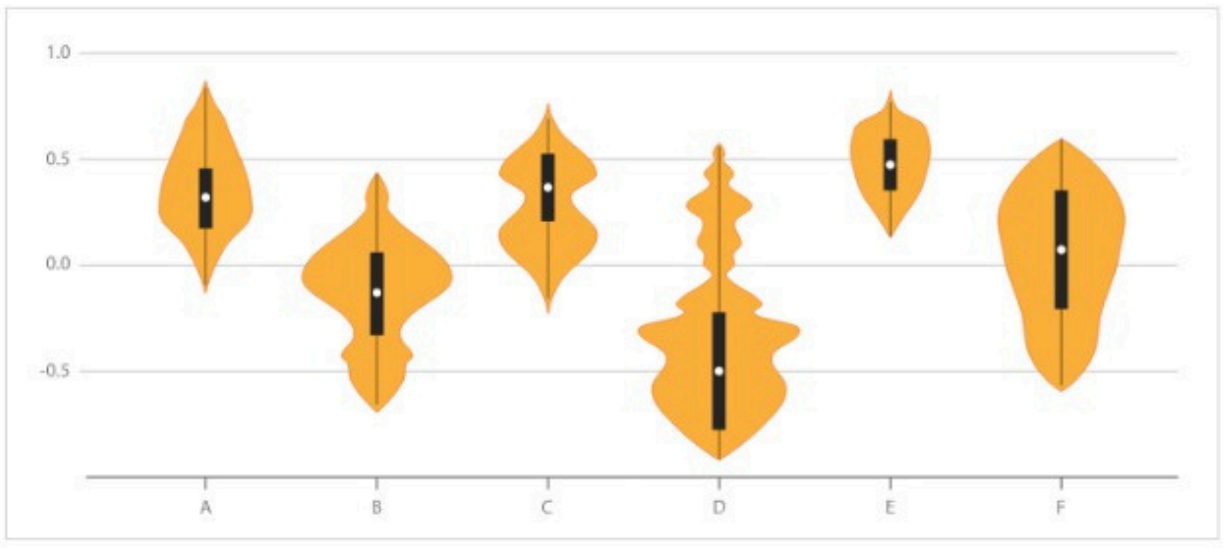


Figura 9: Ejemplo de un diagrama de violín.

Gráfico de dispersión

Los gráficos de dispersión muestran solamente los puntos de datos de una variable. Esto es útil para mostrar relaciones. Para la correlación, los gráficos de dispersión ayudan a mostrar la fuerza de la relación lineal entre dos variables. Para la regresión, los gráficos de dispersión suelen incorporar una línea de ajuste. En la figura 10 se muestra un gráfico de dispersión, en el que se incorporan varias variables empleando colores. De esta forma es posible saber si hay correlaciones y tendencias en los datos.

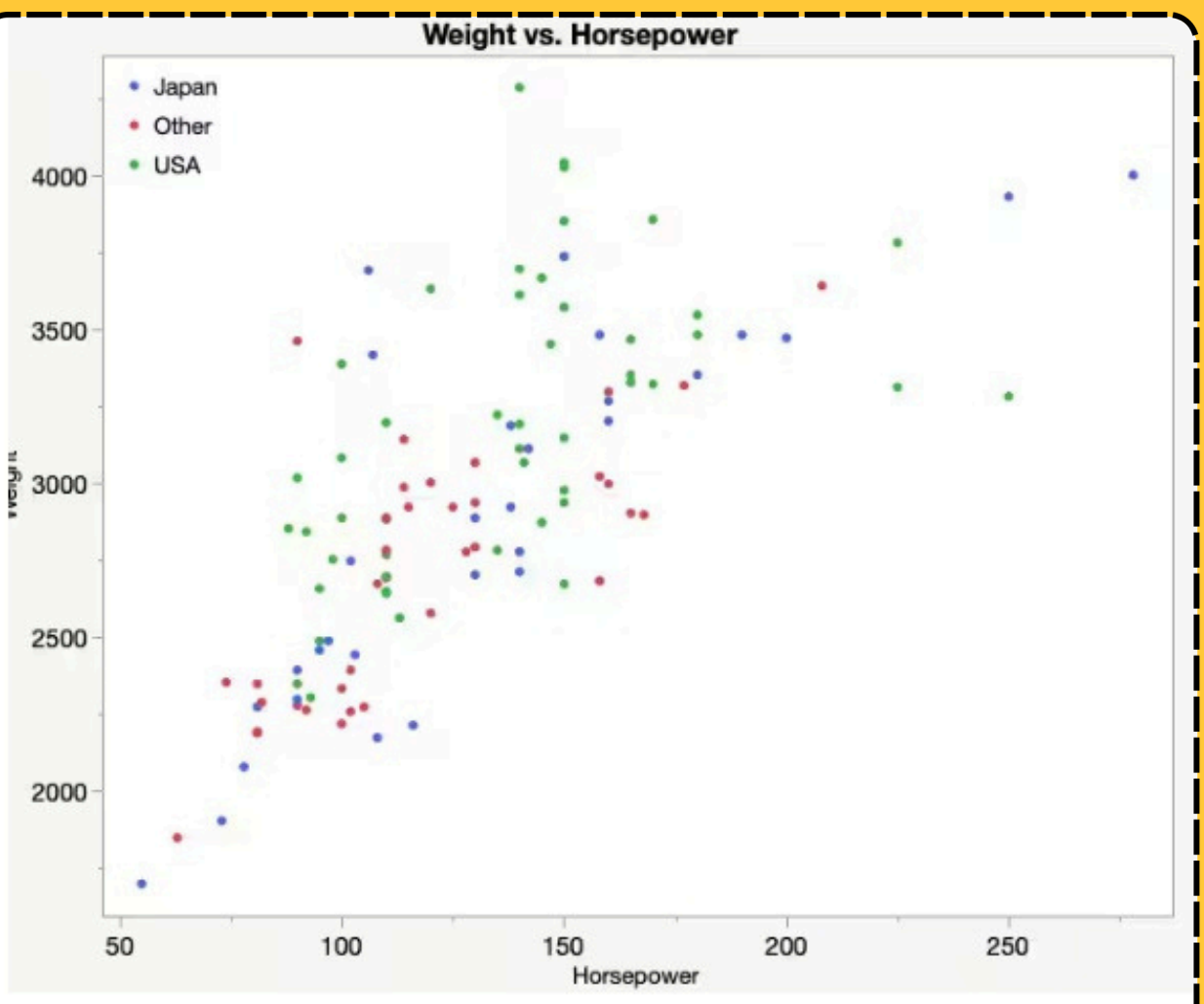


Figura 10: ejemplo de gráfico de dispersión.

Gráfico de burbujas:

Un gráfico de burbujas es un gráfico de varias variables que es un cruce entre un Gráfico de Dispersión y un Pirámide de Población.

Al igual que un diagrama de dispersión, los gráficos de burbujas utilizan un sistema de coordenadas cartesianas para trazar puntos a lo largo de una cuadrícula donde los ejes X e Y son variables separadas. Sin embargo, a diferencia de un gráfico de dispersión, a cada punto se le asigna una etiqueta o una categoría. Cada punto trazado representa entonces una tercera variable por el área de su círculo. Los colores también se pueden utilizar para distinguir entre categorías o para representar una variable de datos adicional. El tiempo se puede mostrar ya sea por tenerlo como variable en uno de los ejes o por animar las variables de datos que cambian con el tiempo. Los gráficos de burbujas se utilizan normalmente para comparar y mostrar relaciones entre círculos etiquetados/categorizados, mediante el uso de posicionamiento y dimensiones. El cuadro general de los gráficos de burbujas puede utilizarse para analizar patrones/correlaciones. Demasiadas burbujas pueden hacer que el gráfico sea difícil de leer, por lo que los gráficos de burbujas tienen una capacidad limitada de tamaño de datos. Esto puede ser remediado por la interactividad: hacer clic o voltear sobre burbujas para mostrar información oculta, tener la opción de reorganizar o filtrar categorías agrupadas.
Al igual que en los gráficos de área proporcional, los tamaños de los círculos deben dibujarse basándose en el área del círculo, no en su radio o diámetro. No solo el tamaño de los circulos cambiara exponencialmen-te, sino que esto dará lugar a interpretaciones erróneas por el sistema visual humano.

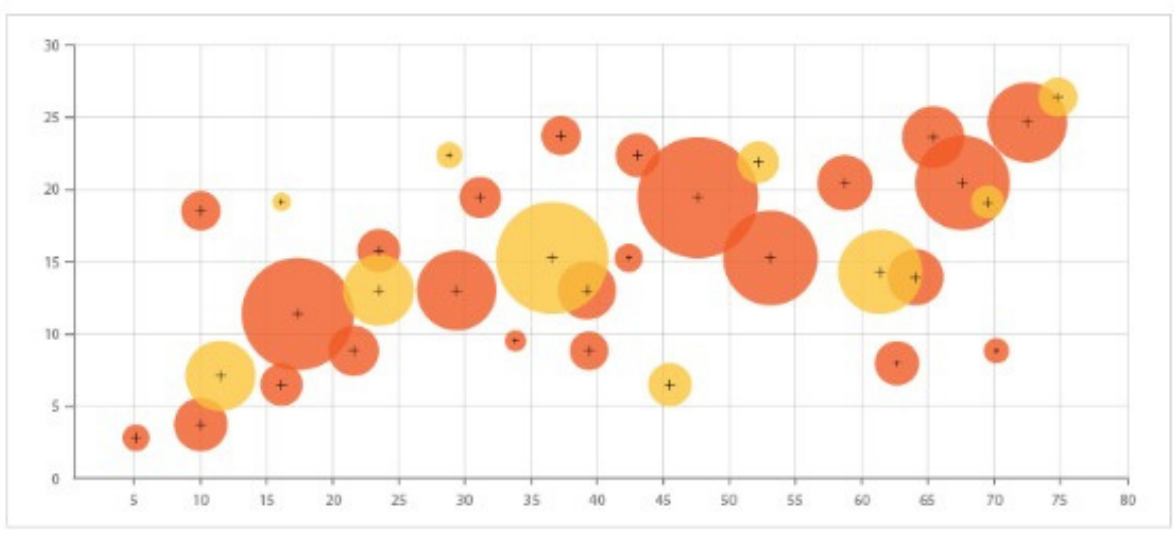


Figura 11: Gráfico de burbujas.

Gráfico de bala:

Usado normalmente para mostrar datos de rendimiento, las funciones del gráfico de bala, son como las del Gráfico de Barras, pero van acompañados de elementos visuales adicionales para aunar más contexto. Originalmente, los gráficos de bala fueron desarrollados por Stephen Few como una alternativa a los indicadores y medidores de tablero, porque a menudo no mostraban suficiente información, eran menos eficientes en el espacio y estaban llenos de «basura del gráfico». El valor de datos principal se codifica por la longitud en la barra principal en el centro del gráfico, que se conoce como la «medida de característica». El marcador de línea que se ejecuta perpendicularmente a la orientación del gráfico se conoce como «medida comparativa» y se utiliza como marcador de destino para comparar con el valor de la medida de característica. Así que si la barra principal ha pasado la posición de la medida comparativa, sabe que ha alcanzado su objetivo. Las barras de color segmentadas detrás de la medida de característica se utilizan para mostrar puntuaciones de rangos cualitativos. Cada tono de color (los tres tonos de gris en el ejemplo anterior) se utilizan para asignar una clasificación de rango de rendimiento. Así por ejemplo, pobre, promedio y bueno. Cuando se utilizan los gráficos de bala, es ideal mantener el número máximo de rangos a cinco.

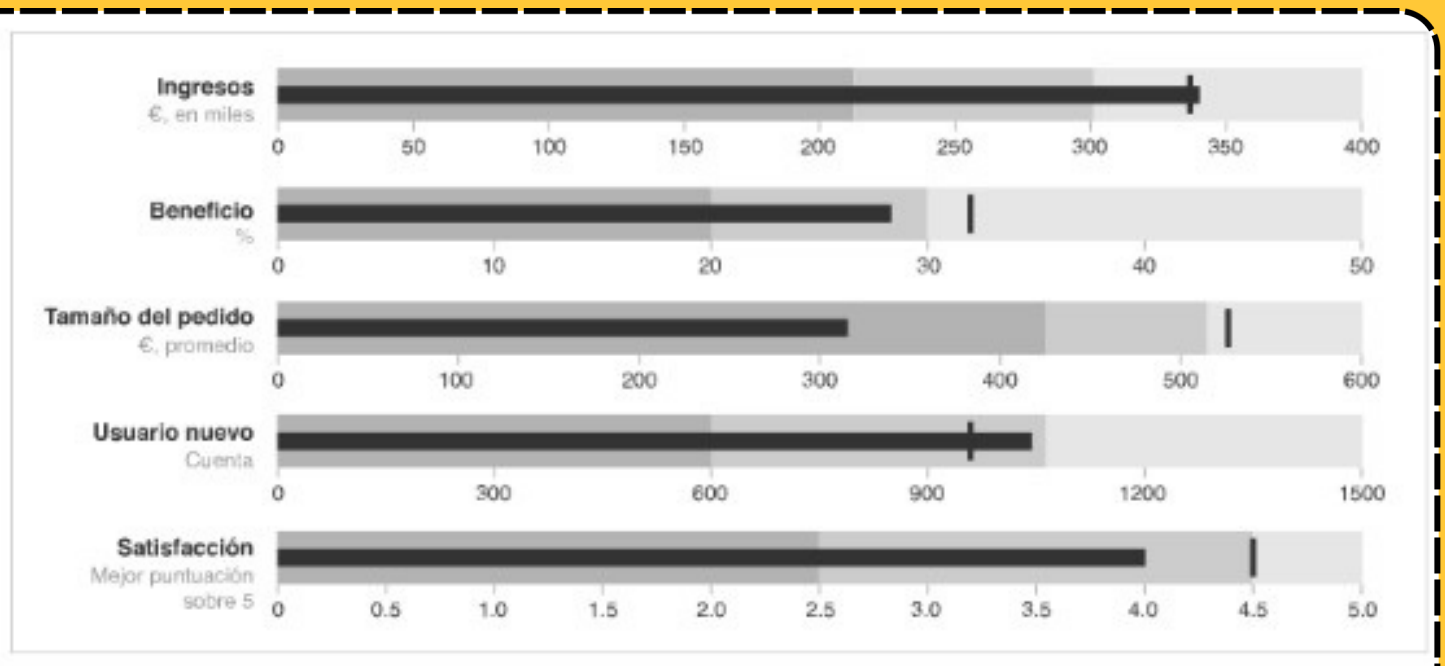


Figura 12: gráfico de bala.

Histogramas:

Un histograma visualiza la distribución de los datos a lo largo de un intervalo continuo o un período de tiempo determinado. Cada barra en un histograma representa la frecuencia tabulada en cada intervalo/bin. El área total del histograma es igual al número de datos. Los histogramas ayudan a dar una estimación de dónde se concentran los valores, cuáles son los extremos y si hay vacíos o valores inusuales. También son útiles para dar una visión aproximada de la distribución de probabilidad.

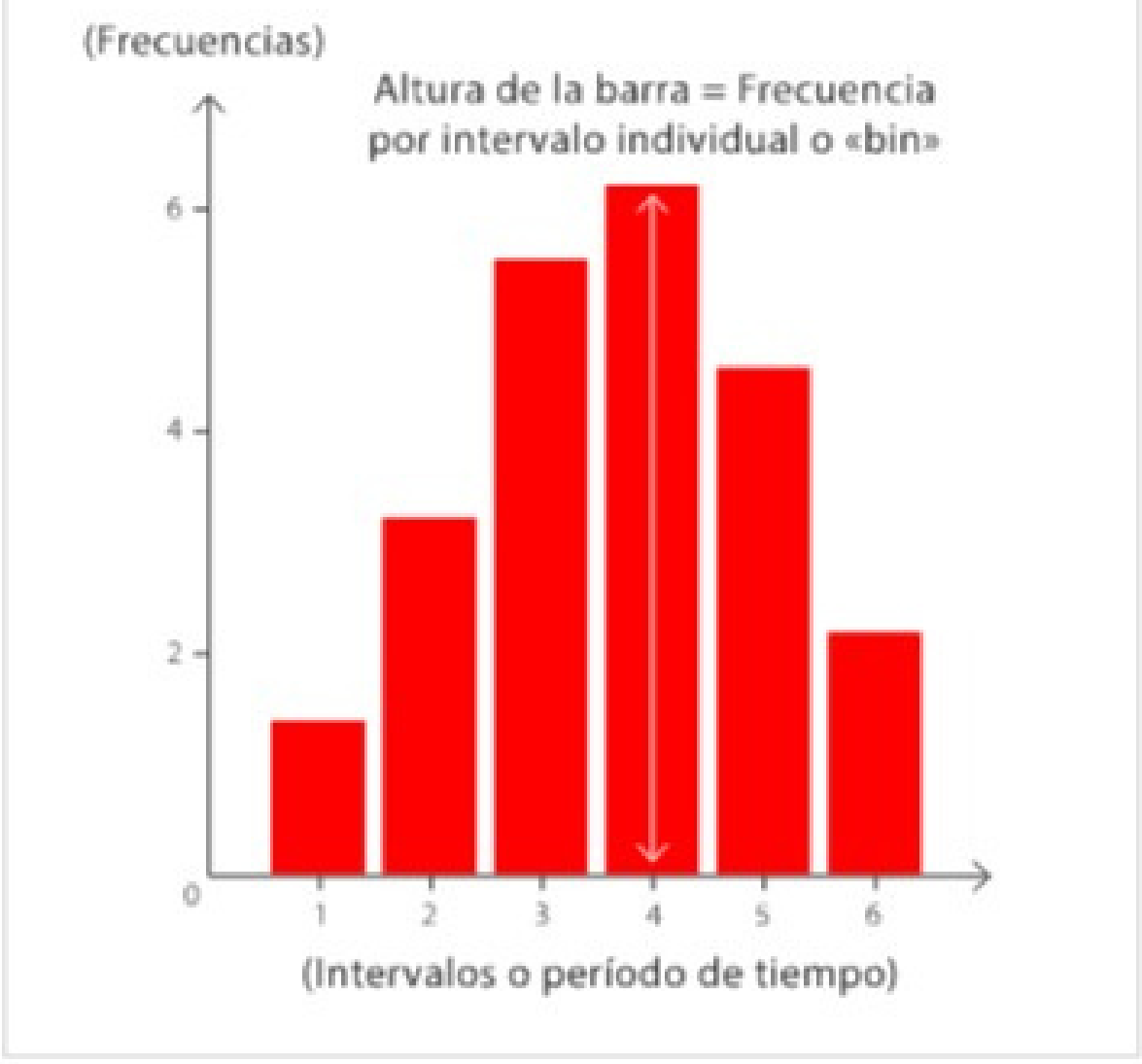


Figura 13: Ejemplo de un histograma

Gráfico de densidad

Un gráfico de densidad visualiza la distribución de datos en un intervalo o período de tiempo continuo. Este gráfico es una variación de un Histograma que usa el suavizado de cerner para trazar valores, permitiendo distribuciones más suaves al suavizar el ruido. Los picos de un gráfico de densidad ayudan a mostrar dónde los valores se concentran en el intervalo. Una ventaja de los gráficos de densidad sobre los histogramas es que son mejores para determinar la forma de distribución porque no se ven afectados por el número de contenedores utilizados (cada barra utilizada en un histograma típico). Un histograma que consta de solo 4 compartimientos no producirá una forma de distribución lo suficientemente distinguible como lo haría un histograma de 20 compartimientos. Sin embargo, con los gráficos de densidad esto no es un problema.

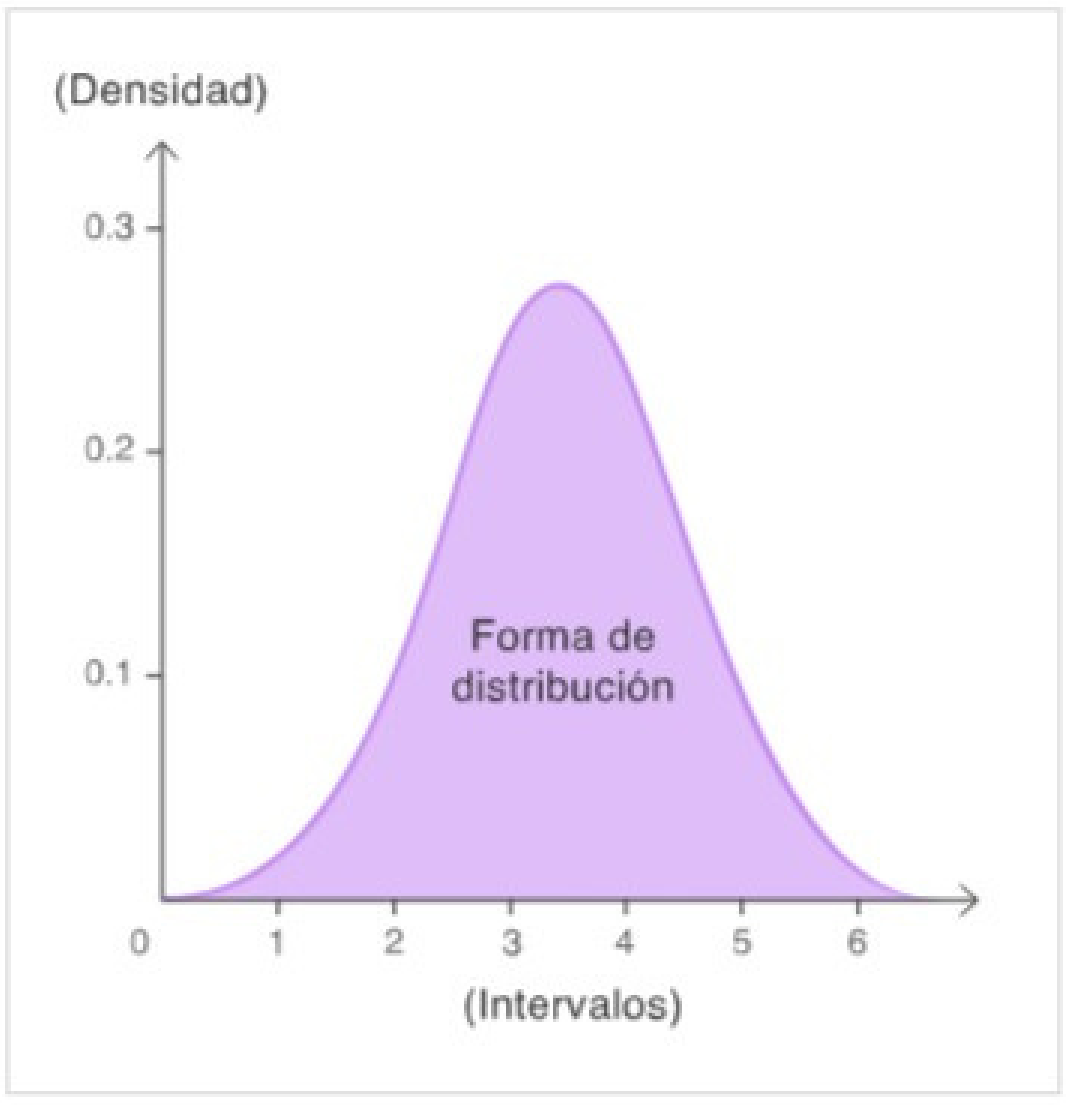


Figura 14: gráfico de densidad.