# **Visualización de datos para análisis de datos**

La visualización de datos es una herramienta fundamental en el análisis de datos. A menudo, los datos pueden ser difíciles de entender simplemente mirando una tabla de números, pero cuando se representan gráficamente, los patrones y las tendencias pueden ser más evidentes y fáciles de interpretar. Además, las visualizaciones son una forma efectiva de comunicar resultados e ideas a otros.

A la hora de crear gráficos y visualizaciones, es importante tener en cuenta algunas mejores prácticas, cómo seleccionar la visualización adecuada para el tipo de datos que se está representando, asegurarse de que la visualización sea fácil de leer y evitar exagerar o distorsionar los datos. También es importante etiquetar claramente los ejes y proporcionar títulos informativos para las visualizaciones.

Entre los tipos de gráficos más comunes se encuentran los gráficos de pastel, los diagramas de caja, los gráficos de dispersión, los gráficos de barras, los gráficos de áreas y los histogramas.

Los gráficos de pastel son útiles para mostrar la proporción de diferentes categorías en un conjunto de datos, mientras que los diagramas de caja son útiles para visualizar la distribución y los valores atípicos en un conjunto de datos numéricos. Los gráficos de dispersión son excelentes para mostrar la relación entre dos variables, mientras que los gráficos de barras y de áreas son útiles para comparar valores entre diferentes categorías. Los histogramas son útiles para mostrar la distribución de datos numéricos.

En Python, hay varias bibliotecas de visualización de datos disponibles, como Seaborn, Matplotlib y Plotly. Seaborn es una biblioteca de visualización de datos basada en Matplotlib, que proporciona una sintaxis más fácil de usar y gráficos más atractivos. Matplotlib es una biblioteca de visualización de datos más antigua y más establecida, y proporciona una amplia gama de gráficos personalizables. Plotly es una biblioteca de visualización de datos interactiva que permite la creación de visualizaciones interactivas en línea.

## Tipos de gráficos

### Gráficos de pastel

Son útiles para mostrar la proporción de diferentes categorías en un conjunto de datos. Se utilizan mejor cuando el número de categorías es pequeño y cuando las proporciones son significativamente diferentes entre sí. Por ejemplo, se puede utilizar un gráfico de pastel para mostrar la proporción de ingresos de una empresa que provienen de diferentes fuentes de ingresos, como ventas en línea, ventas en tiendas físicas, servicios y otros.

### Diagramas de caja

También conocidos como boxplots, son útiles para mostrar la distribución de un conjunto de datos y para identificar posibles valores atípicos o outliers. Se utilizan mejor cuando se tienen datos numéricos y cuando se quiere comparar varias distribuciones de datos. Por ejemplo, se puede utilizar un diagrama de caja para comparar la distribución de las calificaciones de diferentes grupos de estudiantes en un examen.

### Gráficos de dispersión

Son útiles para mostrar la relación entre dos variables numéricas. Se utilizan mejor cuando se quiere identificar patrones o tendencias en los datos y cuando se quiere determinar si existe una relación entre dos variables. Por ejemplo, se puede utilizar un gráfico de dispersión para mostrar la relación entre la cantidad de horas de estudio y las calificaciones de los estudiantes en un examen.

### Gráficos de barras

Son útiles para mostrar la frecuencia o la cantidad de diferentes categorías en un conjunto de datos. Se utilizan mejor cuando se quiere comparar la frecuencia o la cantidad de diferentes categorías y cuando se tienen datos categóricos. Por ejemplo, se puede utilizar un gráfico de barras para mostrar la cantidad de ventas de diferentes productos en una tienda durante un período de tiempo determinado.

### Gráficos de áreas

Son útiles para mostrar la distribución acumulada de un conjunto de datos y para comparar la cantidad o la frecuencia de diferentes categorías en el tiempo. Se utilizan mejor cuando se tiene un conjunto de datos que se distribuye a lo largo del tiempo o cuando se quiere mostrar la evolución de la frecuencia de diferentes categorías. Por ejemplo, se puede utilizar un gráfico de áreas para mostrar la cantidad de ventas de diferentes productos en una tienda durante un año.

### Histogramas

Son útiles para mostrar la distribución de un conjunto de datos numéricos y para identificar posibles valores atípicos o outliers. Se utiliza mejor cuando se tiene un gran conjunto de datos y cuando se quiere identificar patrones en la distribución de los datos. Por ejemplo, se puede utilizar un histograma para mostrar la distribución de las calificaciones de un examen.

## Herramientas de visualización

La visualización de datos es una parte clave del análisis de datos, ya que ayuda a entender y comunicar los patrones y tendencias que se encuentran en los datos. Existen varias herramientas de visualización de datos en Python, pero tres de las más populares son seaborn, matplotlib y plotly.

### Seaborn

Es una biblioteca de visualización de datos basada en matplotlib que se utiliza para crear gráficos estadísticos atractivos y eficientes en Python. Esta biblioteca es ideal para crear gráficos complejos de múltiples variables y permite personalizar fácilmente los gráficos utilizando paletas de colores, estilos de trama y elementos de estilo.

### Matplotlib

Es una biblioteca de visualización de datos muy potente y versátil que se utiliza para crear gráficos de alta calidad en Python. Esta biblioteca permite crear una amplia variedad de gráficos, desde simples gráficos de líneas hasta gráficos más complejos como mapas y diagramas de barras apiladas. Matplotlib también es altamente personalizable y cuenta con una gran comunidad de usuarios que comparten diferentes soluciones para crear gráficos personalizados.

### Plotly

Es una biblioteca de visualización de datos interactiva en línea que permite crear gráficos interactivos, diagramas y dashboards. Es una biblioteca de código abierto y se puede usar tanto en Python como en R. Plotly es una herramienta muy poderosa para crear visualizaciones interactivas, y es especialmente útil para compartir gráficos y datos en línea en tiempo real.

## Mejores prácticas al graficar datos

Algunas de las mejores prácticas a considerar al crear gráficos y visualizaciones para análisis de datos son:

### Mantener la simplicidad

Demasiada información en un solo gráfico puede dificultar la interpretación. Es mejor dividir los datos en varias visualizaciones o usar técnicas como la agrupación o la agregación de datos para reducir la cantidad de información. Asegurarse de que los gráficos sean fáciles de entender y no estén sobrecargados de información. Utilizar sólo los elementos necesarios para contar la historia.

### Usar la paleta de colores adecuada

Los colores deben usarse con cuidado y con un propósito específico. Los colores brillantes pueden distraer de la información importante, mientras que los colores más sutiles pueden ser difíciles de distinguir. Es importante elegir colores que resalten la información que se desea comunicar. También es importante tener en cuenta que algunos colores pueden tener connotaciones culturales y que su elección puede afectar la percepción de los datos.

### Utilizar títulos y etiquetas descriptivas

Las etiquetas deben ser claras y descriptivas, y los ejes deben estar escalados de manera adecuada para representar los datos de manera precisa. Las leyendas deben ser concisas y descriptivas para facilitar la comprensión de la información.

### Elegir el tipo de gráfico adecuado

Es importante elegir el tipo de gráfico que mejor represente los datos y que permita una fácil interpretación.Seleccionar el tipo de gráfico que mejor se adapte a los datos y a la historia que se quiere contar. Algunos tipos de gráficos funcionan mejor para ciertos tipos de datos, como los histogramas para la distribución de frecuencia. Por ejemplo, un gráfico de barras es ideal para comparar valores numéricos, mientras que un gráfico de líneas es mejor para mostrar tendencias a lo largo del tiempo.

### Evitar la distorsión y la falta de contexto

Los gráficos deben ser precisos y no deben distorsionar la información para hacer que los datos parezcan más favorables. Es importante evitar técnicas como cambiar la escala de los ejes o manipular los valores de los datos para hacer que los resultados parezcan más impresionantes. Además, es importante proporcionar contexto para ayudar a los lectores a entender la importancia de los datos.

### Probar diferentes opciones

No hay una sola forma correcta de crear un gráfico. Es importante experimentar con diferentes opciones y ajustar el gráfico según sea necesario para presentar la información de manera clara y concisa.

Otras buenas prácticas incluyen el uso adecuado de los títulos y etiquetas de ejes, la elección adecuada del tipo de gráfico según la naturaleza de los datos y el propósito de la visualización, y la atención a los detalles estéticos como los colores, la legibilidad y la simplicidad. También al elegir los colores para un gráfico, es importante seleccionar colores que sean fácilmente distinguibles entre sí y que no sean demasiado brillantes o estridentes. Además, es importante evitar el uso de gráficos tridimensionales excesivamente complicados que pueden dificultar la comprensión de los datos.

Por ejemplo, supongamos que queremos graficar la cantidad de ventas de una tienda por mes en el último año. Usando la librería Matplotlib, podemos crear un gráfico de líneas para representar esta información. Primero, importamos Matplotlib y creamos dos listas, una para los meses y otra para las ventas correspondientes:

| import matplotlib.pyplot as plt  meses = ["Enero", "Febrero", "Marzo", "Abril", "Mayo", "Junio",  "Julio", "Agosto", "Septiembre", "Octubre", "Noviembre", "Diciembre"] ventas = [10000, 12000, 15000, 18000, 20000, 22000,  25000, 28000, 30000, 32000, 35000, 38000] |
| --- |

Luego, creamos el gráfico de líneas y agregamos etiquetas y leyendas adecuadas:

| plt.figure(figsize=(15,5)) plt.plot(meses, ventas) plt.title("Ventas por mes en el último año") plt.xlabel("Meses") plt.ylabel("Ventas") plt.show() |
| --- |
|  |

Este código producirá un gráfico de líneas que muestra la cantidad de ventas por mes en el último año. La elección de un gráfico de líneas es apropiada para mostrar tendencias a lo largo del tiempo y las etiquetas y leyendas claras hacen que la información sea fácil de interpretar.

| plt.pie(ventas, labels=meses, autopct='%1.1f%%') plt.title('Ventas Mensuales') plt.show() |
| --- |
|  |

Este código generará un gráfico de pastel que muestra la proporción de ventas de cada mes en relación con el total de ventas. Las etiquetas en el gráfico indican el mes correspondiente y el porcentaje de ventas que representa en relación con el total.

Otro ejemplo sería, suponer que tenemos un conjunto de datos que representa el número de ventas mensuales de dos tiendas diferentes. Queremos comparar las ventas de cada tienda y ver cuál de las dos tiene más éxito.

Primero, creamos dos listas de datos, una para cada tienda:

| ventas\_tienda1 = [12000, 15000, 18000, 20000, 21000, 23000, 25000, 27000, 28000, 30000, 32000, 34000] ventas\_tienda2 = [10000, 13000, 14000, 17000, 19000, 21000, 24000, 25000, 27000, 28000, 29000, 31000] |
| --- |

Luego, podemos usar la biblioteca Matplotlib de Python para crear un gráfico de barras que muestre las ventas mensuales de cada tienda.

| import matplotlib.pyplot as plt  meses = range(1, 13)  plt.bar(meses, ventas\_tienda1, color='green', label='Tienda 1') plt.bar(meses, ventas\_tienda2, color='blue', label='Tienda 2')  plt.title('Ventas Mensuales de Tienda 1 y Tienda 2') plt.xlabel('Mes') plt.ylabel('Ventas') plt.legend()  plt.show() |
| --- |
|  |

En este ejemplo, creamos una lista de números que representan los meses del año (1 a 12) y luego usamos la función bar de Matplotlib para crear barras que representan las ventas mensuales de cada tienda. También agregamos un título, etiquetas de eje y una leyenda para que la gráfica sea más fácil de entender.

La gráfica resultante nos permitirá visualizar claramente las ventas mensuales de cada tienda y compararlas de forma sencilla.

En conclusión, la visualización de datos es una parte fundamental en el análisis de datos, ya que nos permite entender mejor los patrones y tendencias presentes en los datos. Con una buena selección de gráficos y herramientas adecuadas, podemos presentar nuestros hallazgos de manera clara y efectiva, lo que facilita la toma de decisiones basadas en datos.