TIC

LECCIÓN 1

Lección 1: Análisis exploratorio de los datos

Análisis exploratorio de los datos

El análisis exploratorio de los datos consiste en intentar descubrir qué explican los datos. Consiste en tratar de localizar patrones y relaciones entre los datos. Para llevar a cabo el análisis exploratorio se requiere indagar ¿Qué tipo de datos se tienen? ¿Son numéricos, son textos, son categorías, contienen información de fechas, horas o registros de tiempo? ¿Qué columnas hay y qué significa cada una? ¿Qué significan las filas? ¿Están completos los datos para todas las filas y columnas? O ¿hay falta de información? Todas estas preguntas permiten determinar el contenido de los datos de forma general y orientar las decisiones para el análisis de los datos. Con las bases establecidas, el siguiente paso es la visualización. Se pueden emplear gráficos y diagramas para representar visualmente la información. Un histograma puede revelar la distribución de datos numéricos, mientras que un diagrama de dispersión puede mostrar relaciones entre variables. Estas representaciones gráficas son claves para identificar patrones de manera intuitiva.

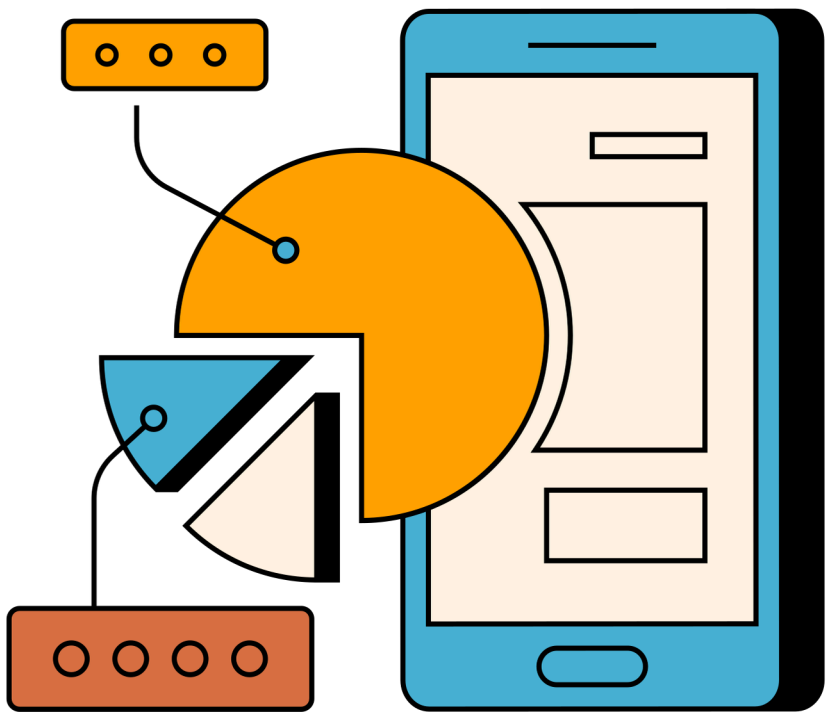
Visualización de los datos para explorar

La visualización de datos desempeña un papel fundamental en la comprensión de conjuntos de datos que contienen variables numéricas enteras. Cuando trabajamos con datos de este tipo, gráficos como histogramas y diagramas de caja permiten observar la distribución de los valores, identificar patrones y detectar posibles valores atípicos. Estas representaciones visuales proporcionan una visión rápida y clara de la variabilidad y tendencias en los datos numéricos. En el caso de variables categóricas, la visualización se centra en la frecuencia de cada categoría. Gráficos de barras y diagramas circulares son herramientas efectivas para ilustrar la proporción de cada categoría en el conjunto de datos. Este enfoque permite identificar la prevalencia de ciertos grupos y entender la distribución de las categorías.

Cuando nos enfrentamos a datos que contienen texto, la exploración se vuelve más desafiante pero igualmente crucial. Técnicas como nubes de palabras y análisis de sentimientos pueden revelar patrones y tendencias en el lenguaje utilizado. Estas visualizaciones permiten comprender la frecuencia de palabras clave, identificar temas recurrentes y evaluar la tonalidad general del texto. La exploración de datos textuales es esencial en ámbitos como el análisis de opiniones en redes sociales, la revisión de comentarios de clientes y la comprensión de grandes conjuntos de datos de texto para la investigación cualitativa.

Exploración con técnicas estadísticas

Además de las visualizaciones, existen diversas técnicas que enriquecen el análisis exploratorio de datos. El análisis estadístico descriptivo es esencial para obtener medidas resumidas, como la media, la mediana y la desviación estándar, que proporcionan una comprensión cuantitativa más profunda de la distribución de los datos. La matriz de correlación permite identificar relaciones entre variables numéricas, mientras que las pruebas de hipótesis pueden validar o refutar suposiciones sobre el conjunto de datos. La segmentación, mediante la cual se divide el conjunto de datos en subgrupos según ciertas características, facilita la identificación de patrones específicos



Enfoque en las variables

Según las características (o variables, o columnas) se pueden determinar cuatro tipos de análisis exploratorios:

- **Univariante no gráfico:** Consiste en tomar una sola columna (o una sola variable) y describirla con estadística, por ejemplo, determinar el máximo, el mínimo, la media, la mediana, la desviación estándar y los cuartiles con el fin de entender su rango y sus valores. **Univariante gráfico:** Este análisis
- complementa a la exploración univariante no gráfica, ya que como se mencionó anteriormente, los gráficos permiten entender de forma fácil y rápida el comportamiento numérico. En esta etapa se grafica cada una de las variables como son diagramas, histogramas, gráficos de barras, conteos, diagramas de cajas, entre otros. **No gráfico multivariante:** en
- este caso se realiza la relación entre dos o mas variables con el fin de ver como se relacionan. Se pueden tabular los datos, graficar en diagrama cartesiano, gráficos de tortas, de barras, análisis de correlación con el fin de establecer cómo interactúan las variables. **Gráfico multivariante:** Consiste en
- tomar conjuntos de variables agrupadas y crear gráficos para validar o refutar hipótesis. En este paso se hace visual el análisis multivariante.

Algunos gráficos que permiten complementar los análisis son:

- Diagramas de
- dispersión Gráficos de
- varias variables
- Gráficos de burbujas
  Mapas de calor