Detección de outliers

Cómo identificar y gestionar outliers de manera efectiva

Los outliers pueden ser desafiantes, pero entender su naturaleza y aprender a manejarlos adecuadamente es **esencial para obtener conclusiones precisas y tomar decisiones informadas**.
Estos datos atípicos pueden ser ventanas a oportunidades y amenazas ocultas en tus datos, por lo que abordarlos con cuidado es clave para un análisis efectivo.

Ahora que comprendes la relevancia de los outliers, es crucial aprender cómo identificarlos y gestionarlos de manera efectiva en tus análisis de datos. Algunas técnicas comunes incluyen:

- Consideración del **contexto** y conocimiento del dominio para determinar si un valor es genuinamente atípico o si representa una observación válida.
- Visualización de datos mediante **gráficos**, como diagramas de caja (boxplots) y diagramas de dispersión.

  ```
  import pandas as pd
  import matplotlib.pyplot as plt

  df = pd.DataFrame({"datos": [1.5, 2.3, 0.8, 1.2, 2.5, 3.3, 1.9, 2.2, -0.5, 3.6]})

  # Evalúo unos límites aceptables
  max_aceptable = 3.7
  min_aceptable = 0.5

  # Crear boxplot
  fig, ax = plt.subplots()
  ax.boxplot(df["datos"])
  ax.set_title("Gráfico Boxplot")
  ax.set_xlabel("Variable")
  ax.set_ylabel("Valor")

  # Agregar líneas de referencia para los valores aceptables
  ymin, ymax = ax.get_ylim()
  ax.hlines(y=max_aceptable, xmin=0.8, xmax=1.2, color='red', linestyle='dashed', lw=2)
  ax.hlines(y=min_aceptable, xmin=0.8, xmax=1.2, color='green', linestyle='dashed', lw=2)

  plt.show()
  ```

  En este diagrama de caja, los outliers son los círculos blancos

- Utilización de **métodos estadísticos**, como el rango intercuartílico y el criterio Z.

Dos métodos comunes para identificar outliers son el rango intercuartílico (IQR) y el criterio Z.

1. **Rango Intercuartílico (IQR):**
    - El IQR es la diferencia entre el tercer cuartil (Q3) y el primer cuartil (Q1) de un conjunto de datos. Matemáticamente, se expresa como (IQR = Q3 - Q1). En ese rango se encierran el 50% central de las observaciones de una variable.
    - Los outliers se identifican generalmente como valores que están por debajo de (Q1 - 1.5 * IQR) o por encima de (Q3 + 1.5 * IQR). Estos límites son a menudo llamados "bigotes" en un diagrama de caja (boxplot).
    - Si un valor cae fuera de estos límites, se considera un posible outlier.
2. **Criterio Z:**
    - El criterio Z se basa en la desviación estándar de un conjunto de datos. La fórmula del Z-score para un dato (X) en un conjunto de datos con media (μ) y desviación estándar \(σ) es

$$Z = (X - \mu)/\sigma$$

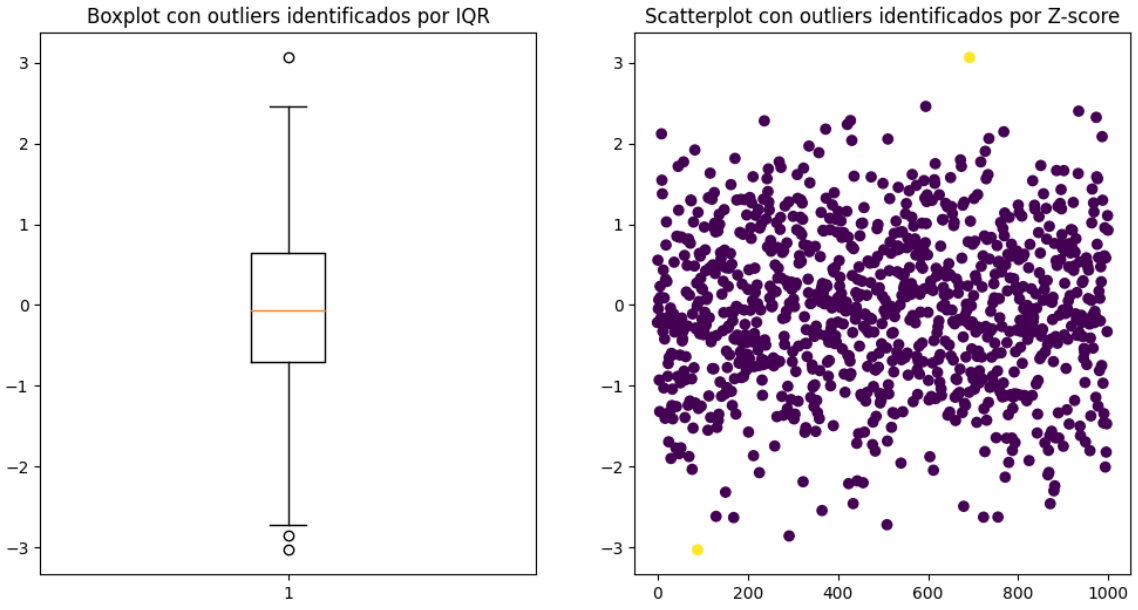
    - Los valores Z cercanos a cero indican que un dato está cerca de la media, mientras que valores Z grandes (positivos o negativos) indican que el dato está lejos de la media.
    - Usualmente, se considera que los valores con un Z-score superior a 3 o inferior a -3 son outliers.

Ambos métodos pueden ser implementados en Python con las bibliotecas NumPy y SciPy.

Ejemplo en Python utilizando NumPy y SciPy:

```
import numpy as np
from scipy import stats

# Generar datos de ejemplo
data = np.random.normal(0, 1, 1000)  # Datos normales con media 0 y desviación están
dar 1

# Método del rango intercuartílico (IQR)
Q1 = np.percentile(data, 25)
Q3 = np.percentile(data, 75)
IQR = Q3 - Q1
lower_limit = Q1 - 1.5 * IQR
upper_limit = Q3 + 1.5 * IQR
```

```
outliers_iqr = (data < lower_limit) | (data > upper_limit)

# Método del criterio Z
z_scores = np.abs(stats.zscore(data))
threshold = 3
outliers_z = z_scores > threshold

# Visualización de resultados
import matplotlib.pyplot as plt

plt.figure(figsize=(12, 6))

plt.subplot(1, 2, 1)
plt.boxplot(data)
plt.title('Boxplot con outliers identificados por IQR')

plt.subplot(1, 2, 2)
plt.scatter(range(len(data)), data, c=outliers_z, cmap='viridis')
plt.title('Scatterplot con outliers identificados por Z-score')

plt.show()
```



Este es solo un ejemplo básico. La elección de qué método usar depende del tipo de datos y de las características específicas de tu conjunto de datos. En algunos casos, puede ser útil utilizar ambos métodos para obtener una detección más robusta de outliers.