Outlier Removal using IQR

Power of IQR for Data Cleansing

In the realm of data analysis, outliers are like unruly outliers disrupting the harmony of your insights. However, fear not, for the Interquartile Range (IQR) comes to the rescue as a robust and effective tool for identifying and removing these data anomalies. In this article, we'll embark on a journey to understand the IQR method, its significance, and how it can be wielded to remove outliers with a practical Python example.

Understanding the Interquartile Range (IQR): The IQR is a statistical measure that quantifies the spread of data within the middle 50% of a dataset. It is the difference between the third quartile (75th percentile) and the first quartile (25th percentile). This measure is particularly robust against outliers since it focuses on the middle portion of the data distribution.

The Code Implementation:

Python3

```
import numpy as np

# Generate example data
data = np.array([23, 25, 22, 27, 21, 24, 26, 100, 23, 28, 22, 29])

# Calculate first and third quartiles
q1 = np.percentile(data, 25)
q3 = np.percentile(data, 75)

# Calculate the IQR
iqr = q3 - q1

# Set IQR multiplier
iqr_multiplier = 1.5

# Define lower and upper bounds
lower_bound = q1 - iqr_multiplier * iqr
upper_bound = q3 + iqr_multiplier * iqr

# Identify outliers
outliers = [x for x in data if x < lower_bound or x > upper_bound]

print("Original Data:", data)
print("Outliers detected using IQR:", outliers)
```

Explanation of the Code:

1. We import the NumPy library for numerical operations in Python.
2. We create an example dataset called `data` with a mix of normal and outlier values.
3. We calculate the first quartile (`q1`) and the third quartile (`q3`) using the `percentile` function from NumPy.
4. We calculate the IQR by subtracting `q1` from `q3` .
5. We set an IQR multiplier (`iqr_multiplier`) which determines how far from the quartiles the bounds will be placed. Commonly used values are 1.5 or 3.
6. We define the lower and upper bounds for outlier detection based on the IQR and multiplier.
7. We identify outliers by checking if a data point falls outside the defined bounds.
8. Finally, we print both the original dataset and the detected outliers.

Report An Issue

If you are facing any issue on this page. Please let us know.

Dash

All

Articles

Videos

Quiz