Name:	Pranita Kumbhar
Roll No:	70
Class/Sem:	TE/V
Experiment No.:	6
Title:	Implementation of outlier detection technique.
Date of Performance:	28/8/25
Date of Submission:	04/9/25
Marks:	
Sign of Faculty:	

Aim: The aim of this experiment is to detect outliers in a dataset using the Z-Score Method, a statistical technique that identifies points that deviate significantly from the mean of the dataset.

Objective: To assess the effectiveness of the Z-Score method in recognizing anomalous data points in normally distributed datasets.

Theory:

The **Z-Score method** is a statistical technique used to identify outliers in a dataset by measuring how far each data point deviates from the mean in terms of standard deviations. It is particularly effective when the data follows a normal distribution. The Z-Score for a data point x_i is calculated as:

$$Z_i = \frac{x_i - \mu}{\sigma}$$

Where μ is the mean and σ is the standard deviation of the dataset. The Z-Score tells us how many standard deviations a data point is away from the mean. Typically, if the absolute value of the Z-Score exceeds a threshold (commonly 3), the data point is considered an outlier.

This method works well for detecting outliers when data is symmetrically distributed but may not be ideal for skewed or heavy-tailed distributions. It is simple, interpretable, and widely used in fields like finance, healthcare, and quality control, where identifying unusual observations is crucial for making informed decisions.

3. **Algorithm:**

The **Z-Score method** is based on the standard score, which indicates how many standard deviations a data point is from the mean of the data. Data points with Z-Scores beyond a certain threshold (usually 3 or -3) are flagged as outliers.

Steps:

1. **Input**: A dataset X with n observations.
2. **Compute the Mean** μ and **Standard Deviation** σ of the dataset.
3. **Calculate the Z-Score** for each data point x_i using the formula:

$$Z_i = \frac{x_i - \mu}{\sigma}$$

4. **Flag outliers**: Any data point for which $|Z_i| > threshold$ (typically 3) is considered an outlier.
5. **Output**: A list of outliers.

Advantages:

- Simple to implement and interpret.
- Works well when the data is normally distributed.

Limitations:

- The Z-Score method assumes a normal distribution. For non-Gaussian distributions, this method may not be appropriate.
- Sensitive to small datasets; a few extreme values can significantly skew the mean and standard deviation.

Code:

```
import numpy as np
import pandas as pd

data = {
   'Values': [10, 12, 14, 15, 15, 17, 20, 50, 60, 14, 15, 16, 100]
}

df = pd.DataFrame(data)

df['Z-Score'] = (df['Values'] - df['Values'].mean()) / df['Values'].std()

threshold = 2.5


df['Outlier'] = np.where(df['Z-Score'].abs() > threshold, True, False)

print(df)

outliers = df[df['Outlier'] == True]
print("\nOutliers detected:")
print(outliers)
```

Output:

```
   Values  Z-Score  Outlier
0      10 -0.659088    False
1      12 -0.583929    False
2      14 -0.508770    False
3      15 -0.471190    False
4      15 -0.471190    False
5      17 -0.396031    False
6      20 -0.283292    False
7      50  0.844095    False
8      60  1.219891    False
9      14 -0.508770    False
10     15 -0.471190    False
11     16 -0.433611    False
12    100  2.723074     True

Outliers detected:
   Values  Z-Score  Outlier
12    100  2.723074     True
```

Conclusion:

Given a dataset of customer ages with a mean of 35 years and a standard deviation of 8 years, a customer is 60 years old. Using the Z-Score method, determine if this customer's age is an outlier with a threshold of 3. What is the Z-Score for this data point, and is it considered an outlier?

Z-Score formula:

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

Where:

- X=60
- μ=35
- σ=8

calculation :

$$Z=60-35/8$$
$$=25/8$$
$$=3.125$$

Since the Z-Score threshold for determining an outlier is 3, the Z-Score of 3.125 exceeds this threshold. Therefore, the customer's age of 60 is considered an outlier.