Name:	Sunit Sunil Khaire
Roll No:	19
Class/Sem:	TE/V
Experiment No.:	6
Title:	Implementation of outlier detection technique.
Date of Performance:	
Date of Submission:	
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.

Output:

```
import numpy as np

# Sample dataset
data = np.array([25, 35, 45, 60, 75, 40, 30, 65, 50, 20])

# Step 2: Compute the mean and standard deviation
mean = np.mean(data)
std_dev = np.std(data)

# Step 3: Calculate the Z-Score for each data point
z_scores = [(x - mean) / std_dev for x in data]

# Step 4: Flag outliers based on the threshold (here, 3)
threshold = 3
outliers = [data[i] for i in range(len(data)) if abs(z_scores[i]) > threshold]

# Step 5: Output the list of outliers
print(f"Mean: {mean}")
print(f"Standard Deviation: {std_dev}")
print(f"Z-Scores: {z_scores}")
print(f"Outliers: {outliers}")
```

```
Mean: 44.5
Standard Deviation: 17.09532099727876
Z-Scores: [-1.1406629921195408, -0.5557076115454173, 0.029247769028706173, 0.9066808398898913, 1.7841139107510766, -0.26322992125835554, -0.8481853018
32479, 1.1991585301769532, 0.3217254593157679, -1.4331406824066024]
Outliers: []
```

Conclusion:

In this experiment, the Z-Score method effectively detected outliers by identifying data points that deviated significantly from the mean in a normally distributed dataset. This statistical approach proved to be a reliable technique for recognizing anomalous values, allowing for the isolation of potential errors or unusual patterns in the data. The results confirm that the Z-Score method is a valuable tool for outlier detection in datasets that follow a normal distribution

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?

Given:

- Mean age of dataset = 35 years
- Standard deviation = 8 years
- Data point = 60 years
- Z-Score threshold = 3

Z-Score formula:

$\mathbf{z = (X - \mu) / \sigma}$

Where:

- z is the Z-Score,
- X is the data point (60 years),
- μ is the mean (35 years),
- σ is the standard deviation (8 years).

Calculation:

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

Since the Z-Score is 3.125, which is greater than the threshold of 3, the data point (60 years) is considered an outlier.