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 & Output:

```
[41] import pandas as pd
```

```
[42] df=pd.read_csv("height.csv")
```

```
[43] df.sample(5)
```

	Gender	Height
3066	Male	69.082703
7160	Female	64.920239
4897	Male	70.178307
139	Male	68.140590
8549	Female	61.660227

```
[44] df.Height.describe()
```

	Height
count	10000.000000
mean	66.367560
std	3.847528
min	54.263133
25%	63.505620
50%	66.318070
75%	69.174262
max	78.998742

dtype: float64

```
[45] # Z-score: Similar to std. deviation, it will give you a number, that tells how many std. deviation you are away from the mean
     df['zscore']=(df.Height-df.Height.mean())/df.Height.std()
     df.head(5)
```

	Gender	Height	zscore
0	Male	73.847017	1.943964
1	Male	68.781904	0.627505
2	Male	74.110105	2.012343
3	Male	71.730978	1.393991
4	Male	69.881796	0.913375

Next steps: Generate code with df | View recommended plots | New interactive sheet

```
[46] df[df['zscore']>3]
```

	Gender	Height	zscore

```
[46] df[df['zscore']>3]
```

	Gender	Height	zscore
994	Male	78.095867	3.048271
1317	Male	78.462053	3.143445
2014	Male	78.998742	3.282934
3285	Male	78.528210	3.160640
3757	Male	78.621374	3.184854

```
[47] df[df['zscore']<-3]
```

	Gender	Height	zscore
6624	Female	54.616858	-3.054091
9285	Female	54.263133	-3.146027

```
[48] #All Outliers

     df[(df.zscore<-3)|(df.zscore>3)]
```

	Gender	Height	zscore
994	Male	78.095867	3.048271
1317	Male	78.462053	3.143445

+ Code + Text

1317	Male	78.462053	3.143445
2014	Male	78.998742	3.282934
3285	Male	78.528210	3.160640
3757	Male	78.621374	3.184854
6624	Female	54.616858	-3.054091
9285	Female	54.263133	-3.146027

```
[49] df_no_outliers=df[(df.zscore>-3)&(df.zscore<3)]
     df_no_outliers.head()
```

	Gender	Height	zscore
0	Male	73.847017	1.943964
1	Male	68.781904	0.627505
2	Male	74.110105	2.012343
3	Male	71.730978	1.393991
4	Male	69.881796	0.913375

Next steps: Generate code with df_no_outliers | View recommended plots | New interactive sheet

```
[50] df.shape[0]-df_no_outliers.shape[0]
```

```
7
```

Conclusion:

In this practical, we applied the Z-Score method to detect outliers in a dataset of heights. By calculating how many standard deviations each data point deviated from the mean, we identified values with Z-Scores exceeding 3 as outliers. The method proved simple and effective for normally distributed data.

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?

- x is the data point (60 years),
- μ is the mean (35 years),
- σ is the standard deviation (8 years).

Calculation:

$z=(x - \mu)/\sigma = (60-35)/8 =$ **3.125**

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