

Vidyavardhini's College of Engineering and Technology

Department of Artificial Intelligence & Data Science

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 \boldsymbol{x}_i using the formula:

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

- 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.

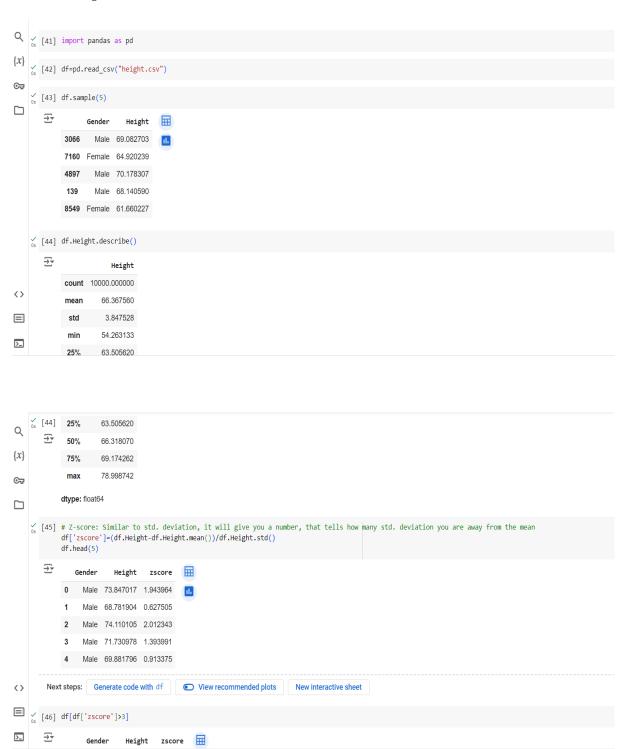
Limitations:

- 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.



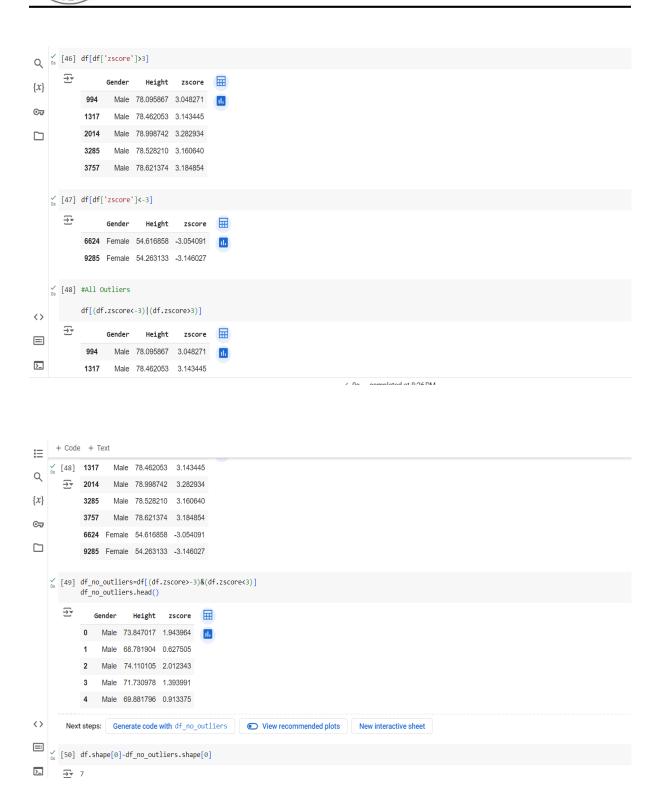
Vidyavardhini's College of Engineering and Technology Department of Artificial Intelligence & Data Science

Code & Output:





Vidyavardhini's College of Engineering and Technology Department of Artificial Intelligence & Data Science



Vidyavardhini's College of Engineering and Technology Department of Artificial Intelligence & Data Science

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.