

Data Cleaning

Identifying Outliers: Z-score method

Before learning about the z-score method for removing the outliers, we are going to answer some questions related to the outliers. These questions are:

What are outliers?

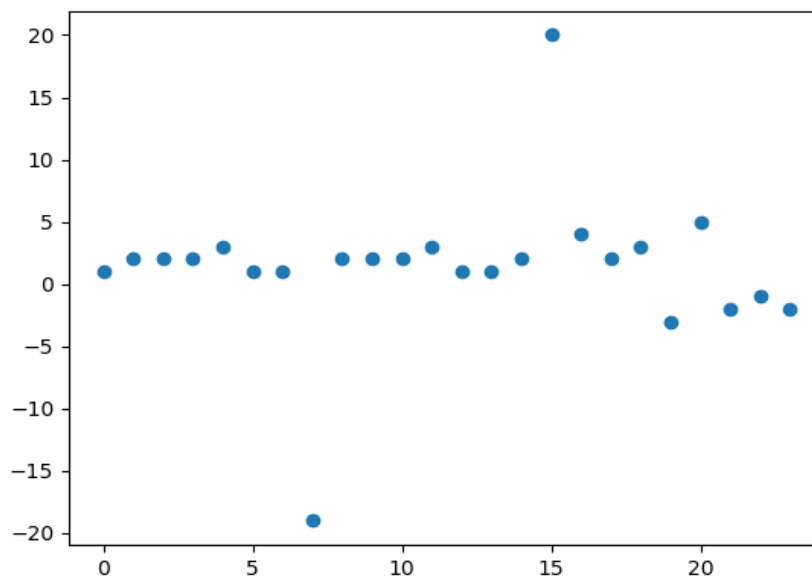
Where do they come from?

Why outliers are bad?

How to deal with outliers?

What are outliers?

Formally, an outlier is a data point that lies at a larger distance from other data points.



In the above diagram, you can see that the data points that lie at -19 and 20 are far away from all other points. These can be depicted as outliers. But we

need to use certain methods to find out the outliers in an accurate manner. We are going to learn the z-score method to identify the outliers.

Where do they come from?

The most common reasons for getting outliers in a dataset can be:

Human Error.

Natural deviation in the population.

Instrument error.

Change in system behaviour

Why outliers are bad?

If outliers are present in the dataset, then they produce variability in the data.

Let's say we take 10 numbers: [1, 2, 1, 1, 3, 2, 4, 2, 15, 3]

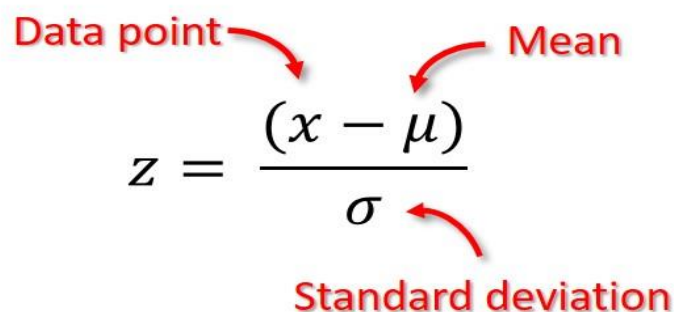
- If we take the mean of this then it comes out to be 3.4
- And if we remove the outlier 15 from this data then the mean is 2.1
- This is a drastic change in the mean.
- Many other metrics like standard deviation and variation also change in the presence of outliers and show deviated results.

How to deal with outliers?

The most common univariate method to treat outliers is using the z-score.

Using a z-score the data is first normalized and then a threshold is defined which can help in eliminating the outlier data point from the rest of the data.

The formula for the z-score is:



The diagram shows the z-score formula with red arrows pointing to its components: 'Data point' points to x , 'Mean' points to μ , and 'Standard deviation' points to σ .

$$z = \frac{(x - \mu)}{\sigma}$$

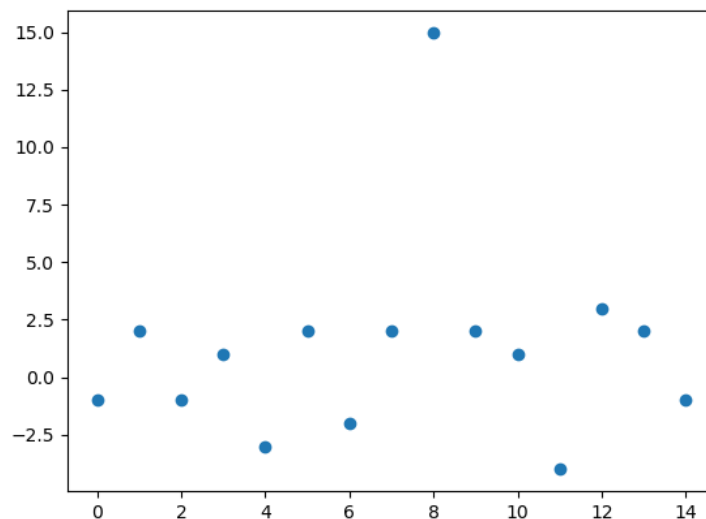
Iterative Algorithm

1. Convert the data to z-score.
2. A data point is an outlier if it exceeds some standard deviation threshold (often -3 and 3).
3. Remove the outliers and repeat until no more outliers are found.

Let's take an example and find out if any outlier is present in the dataset or not. The list "data" contains the data point for the temperature in a city for the last 15 days. We need to find out if some unusual data or outlier is present in this dataset.

```
data = [-1, 2, -1, 1, -3, 2, -2, 2, 15, 2, 1, -4, 3, 2, -1]
```

If we plot this dataset then we get the below graph.



We can see that one data point is far away from the others. Let's use the z-score method to find out if this point is an outlier or not. We are going to use the upper threshold as 3 and the lower threshold as -3.

```
upper_threshold = 3  
lower_threshold = -3
```

Then find the mean and standard deviation to calculate the z-score.

```
mean = np.mean(data)
std = np.std(data)
```

Now, for each data point, we are going to calculate the z-score and print the outlier if present.

```
for i in data:
    z = (i-mean)/std
    if z > upper_threshold or z < lower_threshold:
        print(i)
```

The output is 15. It means that 15 is the data point which is an outlier in the dataset. That's how we can find the outliers in a dataset.