# Dealing With Outliers

## Outliers –

Outlier is the number or the number which stays out of the range calculated as below.

**[Q1 – 1.5\*IQR, Q3 + 1.5\*IQR],** yellow marking is the lowest number and the green marking is the largest number in the range specified. Any number that resides outside this range will be called an outlier.

Now let us understand what **IQR** is.

**IQR or Inter-Quartile Range** –

In a given list of numbers in our dataset, we first calculate the actual median value. It's called **Q2(50th percentile)**. Once we have **Q2**, we have two halves of the list. One half occurs before the actual median, one half occurs after. The median value of the first half is called **Q1(25th percentile)** and the median value of the second half is called **Q3(75th percentile)**, the difference between the **Q3** and **Q1** is our **interquartile range**(**IQR**).

**Example -**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Given List of Numbers | | | | | | | | |
| 4 | 4 | 10 | 11 | 15 | 7 | 14 | 12 | 6 |
| After Sorting | | | | | | | | |
| 4 | 4 | 6 | 7 | 10 | 11 | 12 | 14 | 15 |
|  |  |  | Actual Median Value | | |  |  |  |
| First half | | | | Q2 | Last Half | | | |
|  |  |  |  |  |  |  |  |
| 4 | 4 | 6 | 7 | 10 | 11 | 12 | 14 | 15 |
| Median Value of First half or Q1 | | | | Median Value of Last half or Q3 | | | |
| 5 | | | | 13 | | | |
| IQR (Q3 - Q1) | | | | | | | | |
|  |  |  |  | **8** |  |  |  |  |

Based on **[Q1 – 1.5\*IQR, Q3 + 1.5\*IQR]** range formula, our range is, -7 to 25. So any number if present will be treated as an outlier.

## Dealing Methods –

### Univariate method –

### One of the simplest methods for detecting outliers is the use of box plots. A box plot is a graphical display for describing the distribution of the data. Box plots use the median and the lower and upper quartiles.

Tukey’s method defines an outlier as those values of the data set that fall far from the central point, the median. The maximum distance to the center of the data that is going to be allowed is called the cleaning parameter. If the cleaning parameter is very large, the test becomes less sensitive to outliers. On the contrary, if it is too small, a lot of values will be detected as outliers.

**Multivariate Method –**

Outliers do not need to be extreme values. Therefore the univariate method does not always work well. The multivariate method tries to solve that by building a model using all the data available and then cleaning those instances with errors above a given value.

### Minkowski error –

Now, we are going to talk about a different method for dealing with outliers. Unlike the univariate and multivariate methods, it doesn’t detect and clean the outliers. Instead, it reduces the impact that outliers will have on the model.

The Minkowski error is a loss index that is more insensitive to outliers than the standard sum squared error. The sum squared error raises each instance error to the square, making a too big contribution of outliers to the total error. The Minkowski error solves that by raising each instance error to a number smaller than 2, for instance, 1.5. This reduces the contribution of outliers to the total error. For instance, if an outlier has an error of 10, the squared error for that instance will be 100, while the Minkowski error will be 31.62.