

CLEAN DATA

REMOVE OUTLIERS

1. What are Outliers?

We all have heard of the idiom ‘odd one out which means something unusual in comparison to the others in a group.

Similarly, an Outlier is an observation in a given dataset that lies far from the rest of the observations.

That means an outlier is vastly larger or smaller than the remaining values in the set.

2. Why do they Occur?

An outlier may occur due to the variability in the data, or due to experimental error/human error.

They may indicate an experimental error or heavy skewness in the data (heavy-tailed distribution).

3. What do they affect?

In statistics, we have three measures of central tendency namely Mean, Median, and Mode as help us describe the data.

Mean is the accurate measure to describe the data when we do not have any outliers present.

Median is used if there is an outlier in the dataset.

Mode is used if there is an outlier AND about $\frac{1}{2}$ or more of the data is the same.

‘Mean’ is the only measure of central tendency that is affected by the outliers which in turn impacts Standard deviation.


Example 1:

Consider a small dataset,

sample= [15, 101, 18, 7, 13, 16, 11, 21, 5, 15, 10, 9]

By looking at it, one can quickly say '101' is an outlier that is much larger than the other values.

+-----+-----+	
with outlier	without outlier
+-----+-----+	
Mean: 20.08	Mean: 12.72
Median: 14.0	Median: 13.0
Mode: 15	Mode: 15
Variance: 614.74	Variance: 21.28
Std dev: 24.79	Std dev: 4.61
+-----+-----+	



From the above calculations, we can clearly say the Mean is more affected than the Median.

4. Detecting Outliers

If our dataset is small, we can detect the outlier by just looking at the dataset.

But what if we have a huge dataset, how do we identify the outliers then?
We need to use visualization and mathematical techniques.

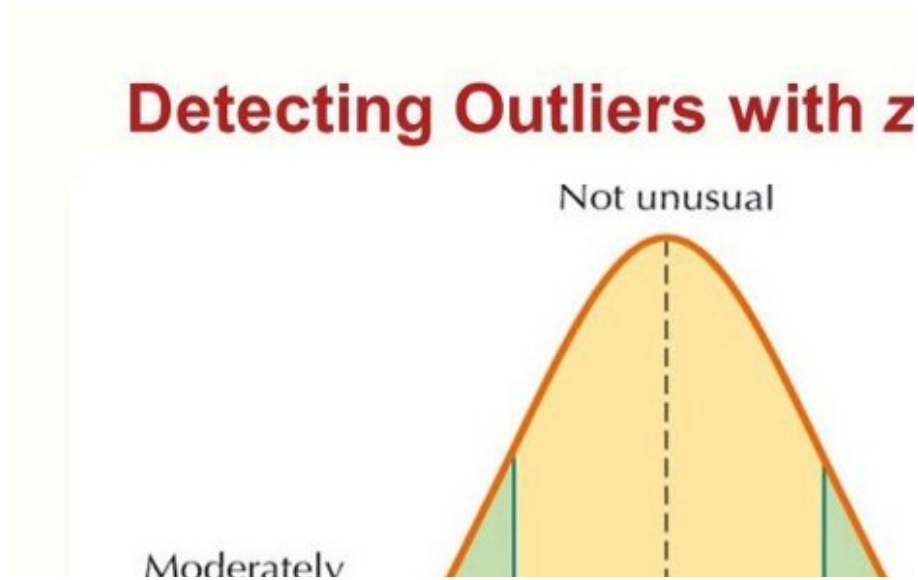
Below are some of the techniques of detecting outliers

- Boxplots
- Z-score
- Inter Quantile Range(IQR)

4. Detecting Outliers using Box Plot

4. Detecting Outliers using Z Score

Note : Any data point whose Z-score falls out of 3rd standard deviation is an outlier.

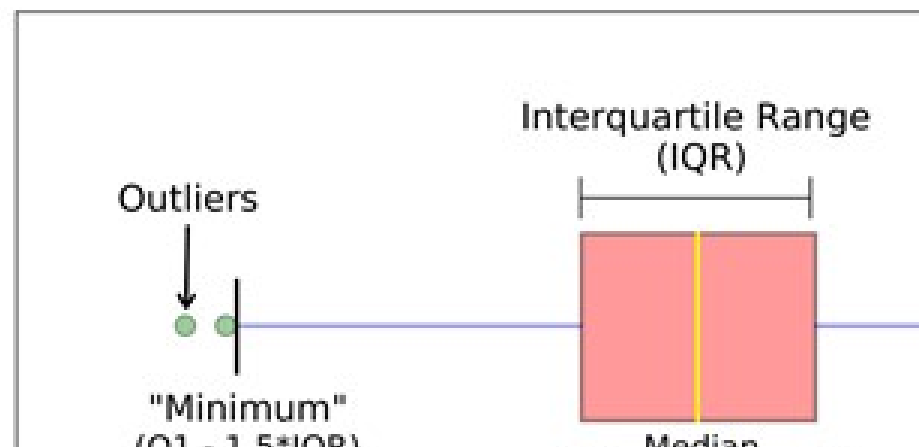


Steps:

- loop through all the data points and compute the Z-score using the formula $(X_i - \text{mean}) / \text{std.}$
- Define a threshold value of 3 and mark the datapoints whose absolute value of Z-score is greater than the threshold as outliers.

4. Detecting Outliers using the Inter Quartile Range (IQR)

Note : Data Points that lie 1.5 times of IQR above Q3 and below Q1 are outliers.



Steps:

- sort the dataset in ascending order
- calculate the 1st and 3rd quartiles(Q1, Q3)
- compute $IQR = Q3 - Q1$
- compute lower bound = $(Q1 - 1.5 * IQR)$, upper bound = $(Q3 + 1.5 * IQR)$
- loop through the values of the dataset and check for those who fall below the lower bound and above the upper bound and mark them as outliers

5. Handling Outliers

5.1 Trimming/Remove the outliers

In this technique, we remove the outliers from the dataset. Although it is not a good practice to follow. Python code to delete the outlier and copy the rest of the elements to another array.

5.2 Quantile based flooring and capping

In this technique, the outlier is capped at a certain value above the 90th percentile value or floored at a factor below the 10th percentile value.

5.3 Mean/Median imputation

As the mean value is highly influenced by the outliers, it is advised to replace the outliers with the median value