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.

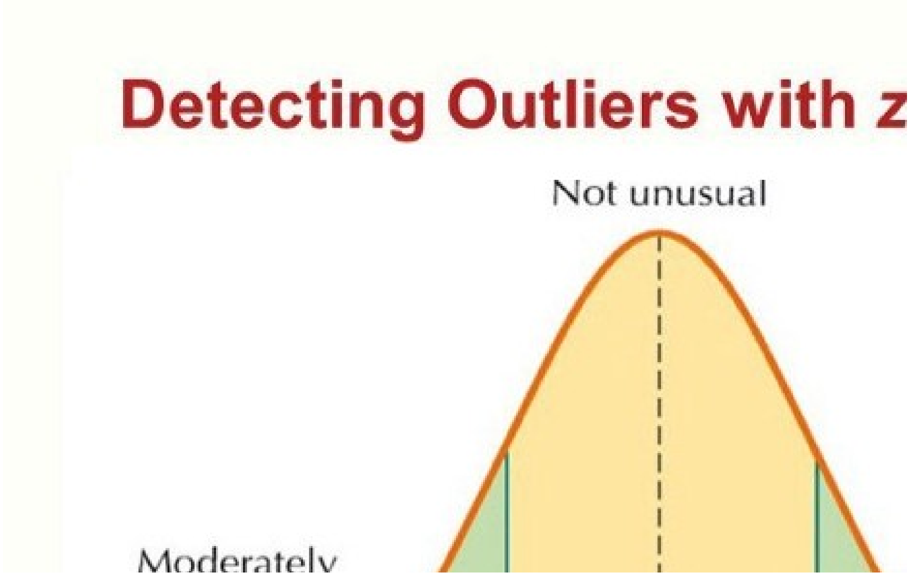
Below are some of the techniques of detecting outliers

- 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.*

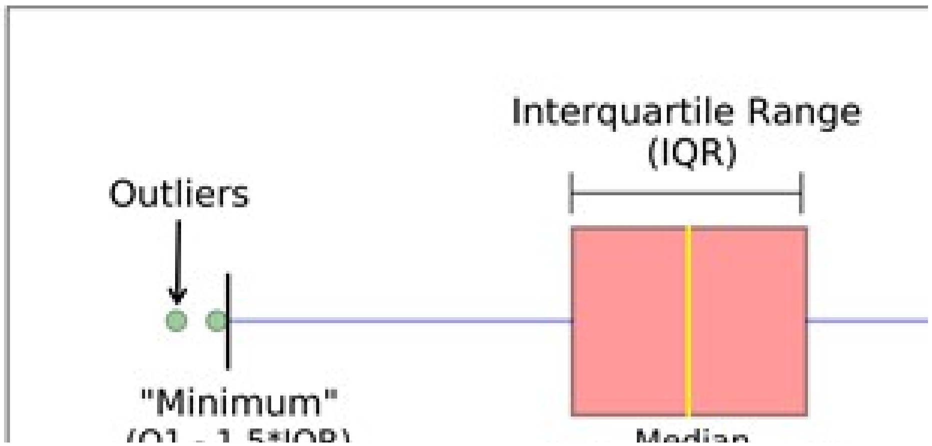


Steps:

- loop through all the data points and compute the Z-score using the formula (Xi-mean)/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