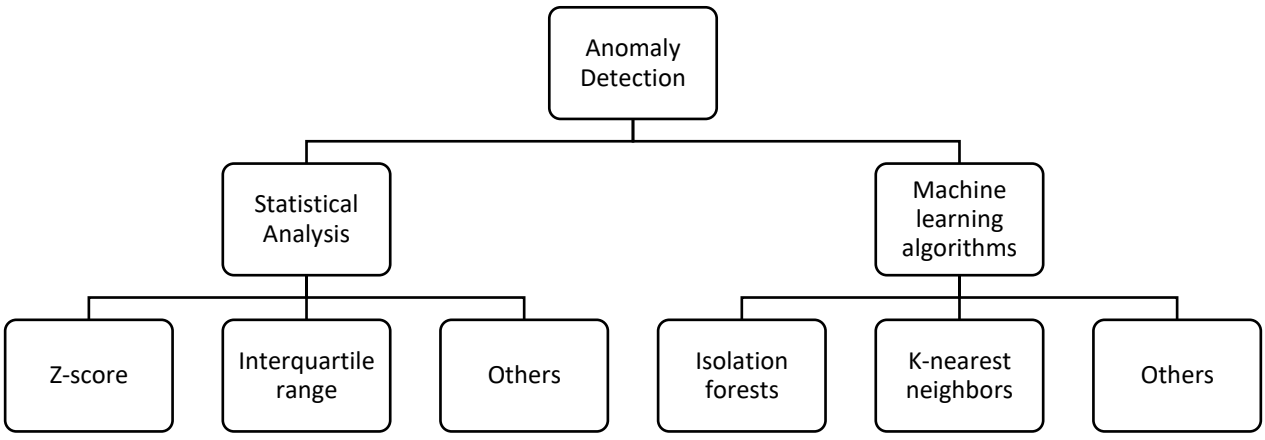
Anomaly Detection:

Anomaly is something different from what is normal, usual, or expected. In a dataset, anomalies are the observations that deviates from the usual or expected observations. The anomalies are is also known as the noises, outliers, or novelties in a data set.

In summer, the average temperature of Dhaka ranges from 26°C to 33°C. in somedays of a particular year, we may see the temperature near to the 20°C (i.e., 22°C) or near to the 40°C (i.e., 37°C) without any reason. In the temperature dataset of several years of Dhaka, the temperature of 22°C or 37°C should be identified as anomaly or outlier.

The anomaly detection is a technique to identify the observations that are deviated from the usual or expected observations. Often, anomalies are significantly different from the majority of the data. So, they can mislead to the conclusions in machine learning as well as data science. Moreover, a machine learning model may fit to the noise rather than the underlying pattern. As a result, we may see the poor performance of a machine learning algorithm. Again, anomaly detection is essential in financial transactions to identify potential fraud, in cybersecurity to detect malware, in healthcare to detect any potential emergency etc. Identifying and handling anomalies are crucial part prior to the applying a machine learning algorithm.

There is no rule of thumb to detect anomalies. We can use both statistical analysis and machine learning algorithms to detect anomalies. The following diagram showcases some techniques to identify anomalies.

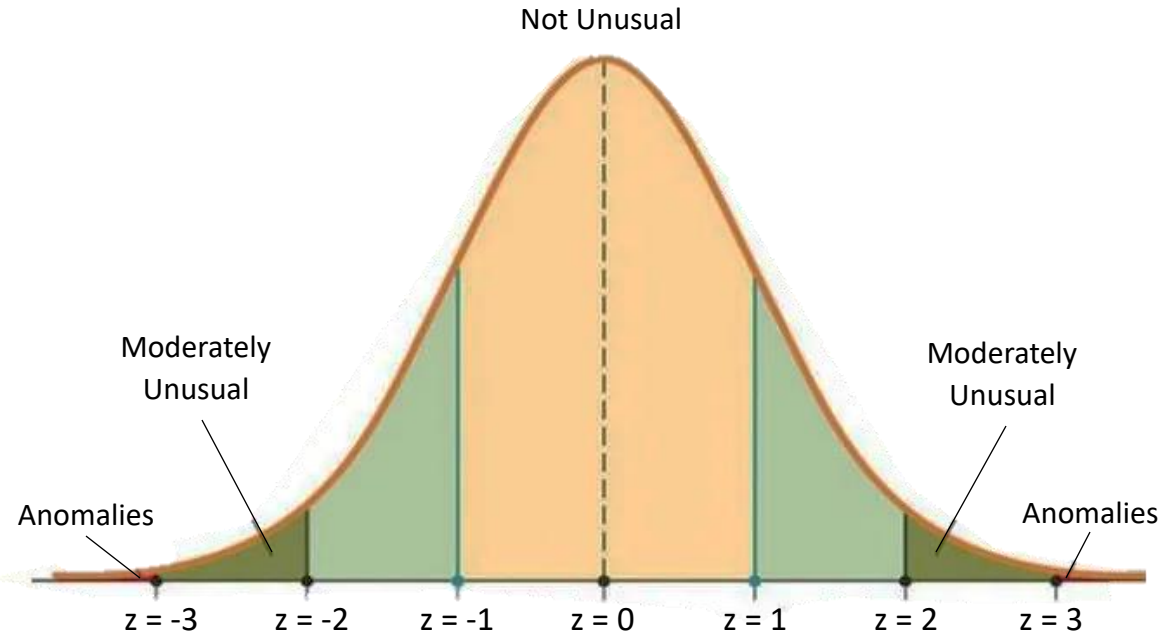


Techniques of Anomaly Detection

Statistical analysis for anomaly detection:

Z-score: The method z-score requires to use the standard deviation. It determines the number of standard deviations a specific observation is away from the mean in a particular feature. Generally, we define an observation as an anomaly whose z-score is greater than a threshold in absolute term. The threshold value may be 2 or 3.

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

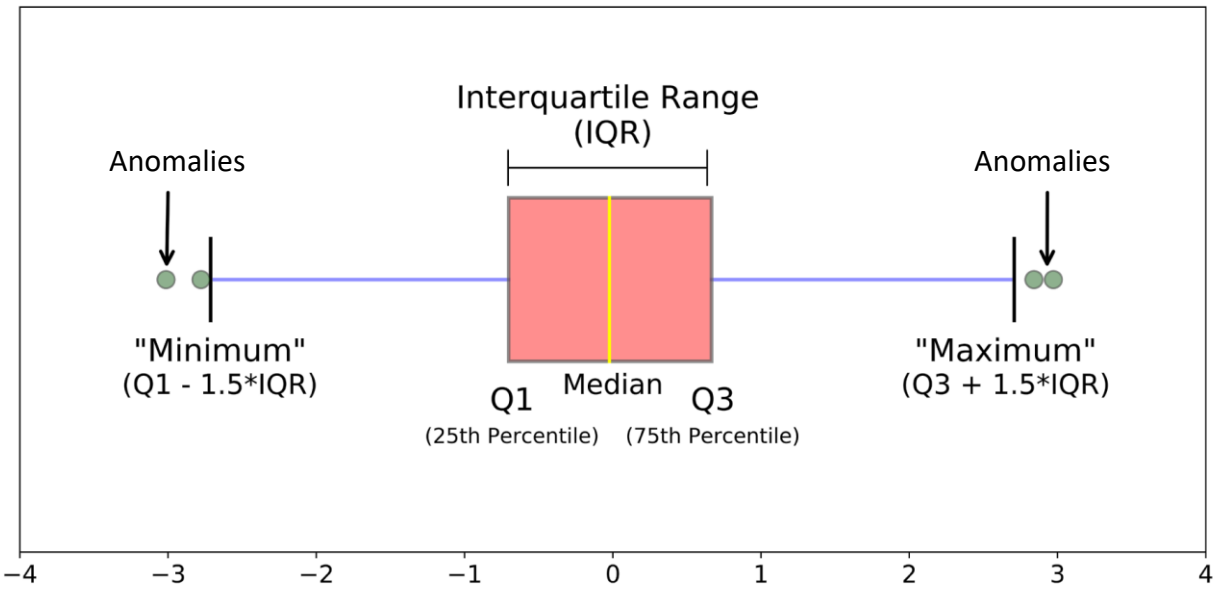


In Python, we can use scipy.stats.zscore to calculate Z-score.

Interquartile range (IQR): It is the range between the first quartile (Q_1) and the third quartile (Q_3). Any observation beyond some multiplier of IQR is considered as the outlier. A popular multiplier is 1.5.

$$IQR = Q_3 - Q_1$$

The observations below the $Q_1 - 1.5 \times IQR$ and above the $Q_3 + 1.5 \times IQR$ represent outliers. This method can be shown graphically using the boxplot.

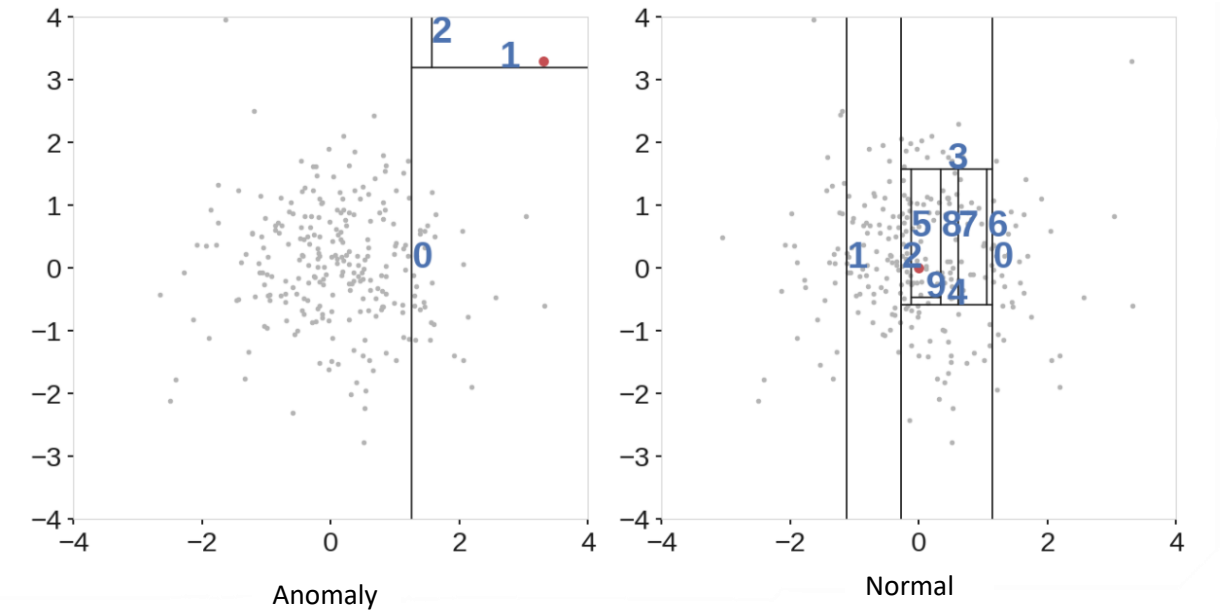

Interquartile Range
(IQR)
Anomalies
Anomalies
"Minimum"
(Q1 - 1.5*IQR)
Q1
(25th Percentile)
Median
Q3
(75th Percentile)
"Maximum"
(Q3 + 1.5*IQR)
−4
−3
−2
−1
0
1
2
3
4

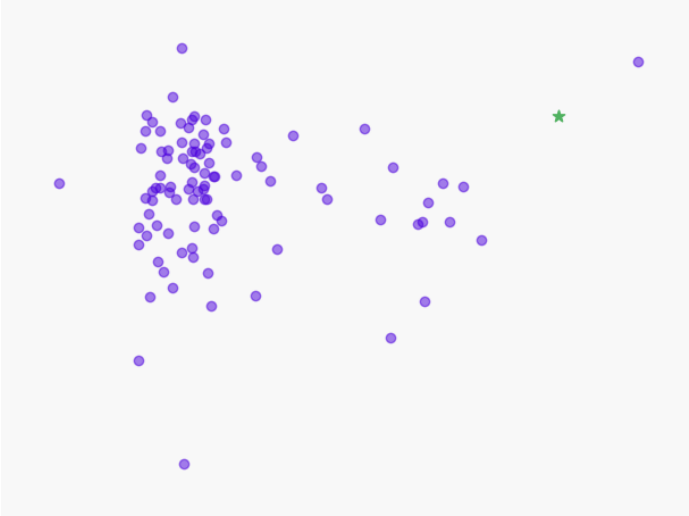
Machine learning algorithm for anomaly detection:

Isolation forests: The goal of isolation forests is to isolate anomalies. It is an unsupervised machine learning algorithm. This algorithm is based on the decision tree. Each tree in isolation forests is called isolation tree (iTree). The algorithm consists in:

i. Select a feature randomly.
ii. The dataset is partitioned randomly to obtain two subsets of the data.
iii. Until an observation is isolated, the previous two steps are repeated.
iv. Repeat previous steps recursively.

As the ensemble model like Random Forest, a forest is made up of many isolation trees (iTrees) whose results are combined to obtain a better result. The anomalies tend to be the leaves closest to the root of the tree. An anomaly scores 1 is assigned to normal observations and -1 to anomalies.





To build isolation forests in Python, we can use sklearn.ensemble.IsolationForest.