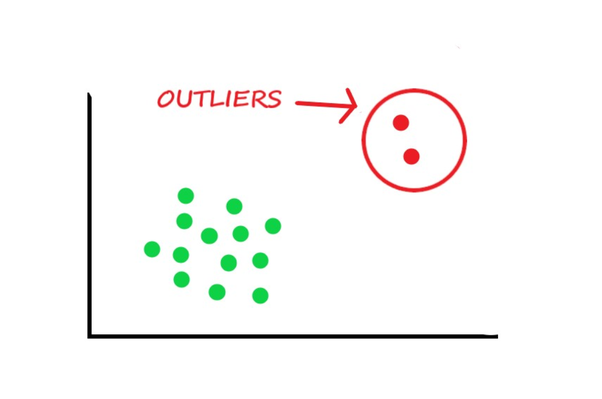
1. What are Outliers?

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. Anomalies, or outliers, can be a serious issue when training machine learning algorithms or applying statistical techniques. They are often the result of errors in measurements or exceptional system conditions and therefore do not describe the common functioning of the underlying systemOutlier detection, also known as anomaly detection, is a common task for many data science teams. It is the process of identifying data points that have extreme values compared to the rest of the distribution.



Applications:

data quality monitoring, identifying price arbitrage in finance, detecting cybersecurity attacks, healthcare fraud detection, banknote counterfeit detection and more

**Types of Outliers**

### **Global Outliers**

Also known as “Point Anomalies,” are kind of outliers which deviate significantly from the rest of the data. Measurement will be called a global outlier if it diverges from the distribution of data regardless of the features because that measurement is far off the global distribution. It is the simplest type of outlier and is found in the majority of cases.

A global outlier remains distinct from other data points by representing its outliers. It can be better explained by considering a real-life example dataset of credit card fraud detection, which contains transactional data of a bank’s customer who holds a credit card. If we consider the daily transaction amount by the customer as one of the attributes, then a transaction with a very high amount as compared to the normal range of the individual’s expenditure will be considered as a Point or Global outlier.

### **2. Contextual Outliers**

If a data instance is anomalous in a specific context, then it would be called a contextual outlier or a conditional outlier. Therefore, a contextual outlier will represent a small group of outliers in itself (having some similar features) as compared to a significantly larger group of observations. The value might, however, be seen as normal in a different context.

The idea of a context is induced by the structure of the dataset and should be specified as a part of problem formulation. The alternative of applying a contextual outlier technique is decided by the meaningfulness of the contextual outliers in the target domain where it has to be applied.

### **3. Collective Outliers**

When a subset of observations in a dataset, as a collection, deviates significantly from the entire dataset, it is called a collective outlier. It is not necessary that each instance within the collective outlier is also an outlier. When seeking outlier detection, it is very important to keep the context in mind because sometimes, a point or collective outlier can also be a contextual outlier given the context of the study.

Analogy

A plane landing on a highway is a **global** outlier because it’s a truly rare event that a plane would have to land there. If the highway was congested with traffic that would be a **contextual** outlier if it was happening at 3 a.m. when traffic doesn’t usually start until later in the morning when people are heading to work. And if every car on the freeway was moving to the left lane at the same time that would be a **collective** outlier because although it’s definitely not rare that people move to the left lane, it is unusual that all cars would relocate at the same exact time.

A banking customer who normally deposits no more than $1000 a month in checks at a local ATM suddenly makes two cash deposits of $5000 each in the span of two weeks is a **global anomaly**because this event has never before occurred in this customer’s history. The time series data of their weekly deposits would show an abrupt recent spike. Such a drastic change would raise alarms as these large deposits could imply illicit commerce or money laundering.

A sudden surge in order volume at an eCommerce company, as seen in that company’s hourly total orders for example, could be a **contextual outlier** if this high volume occurs outside of a known promotional discount or high volume period like Black Friday. Could this stampede be due to a pricing glitch which is allowing customers to pay pennies on the dollar for a product?

A publicly traded company’s stock is never a static thing, even when prices are relatively stable and there isn’t an overall trend, and there are minute fluctuations over time. If the stock price remained at exactly the same price (to the penny) for an extended period of time, then that would be a **collective outlier**. In fact, this very thing occurred to not one, but several tech companies on July 3 of this year on the Nasdaq exchange when the stock prices for several companies – including tech giants Apple and Microsoft – were listed as $123.45.

2. Why do they occur?

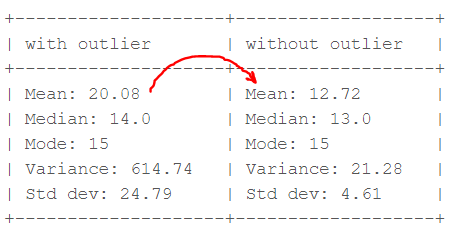
* An outlier may occur due to the variability in the data, or due to experimental error/human error.
* They may indicate an experimental error or heavy skewness in the data(heavy-tailed distribution).

3. What do they affect?

In statistics, we have three measures of central tendency namely Mean, Median, and Mode. They 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 ½ 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:

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.

computation with and without outlier (Image by author)

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)

**4.1 Detecting outliers using Boxplot:**

A box plot is a standard way to visualize the quartiles for numerical values in data. Quartiles divide numerical data into four groups:

1. The first quartile is the middle number between the minimum and the median, so 25 percent of the data falls below this point.
2. The second quartile is the median, which means that 50 percent of the data falls below this point.
3. The third quartile is the middle number between the maximum and the median, so 75 percent of the data falls below this point.
4. The fourth quartile is the highest 25 percent of the data

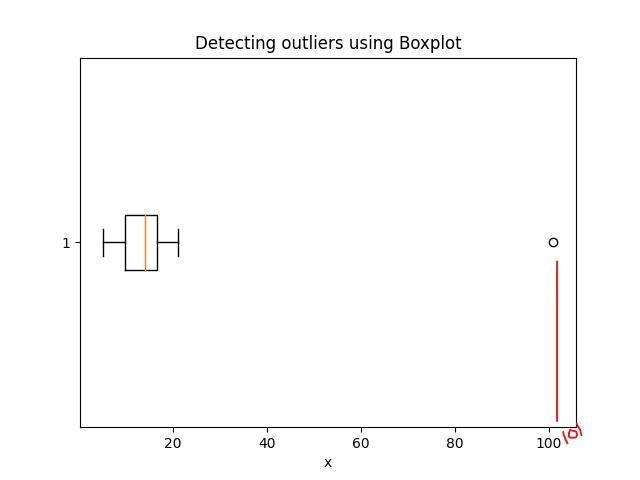
Python code for boxplot is:

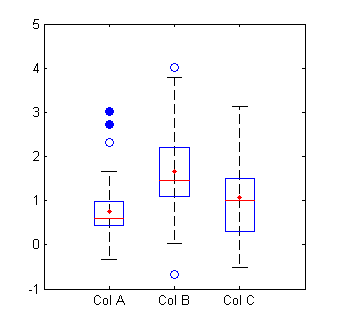
import matplotlib.pyplot as plt

plt.boxplot(sample, vert=False)

plt.title("Detecting outliers using Boxplot")

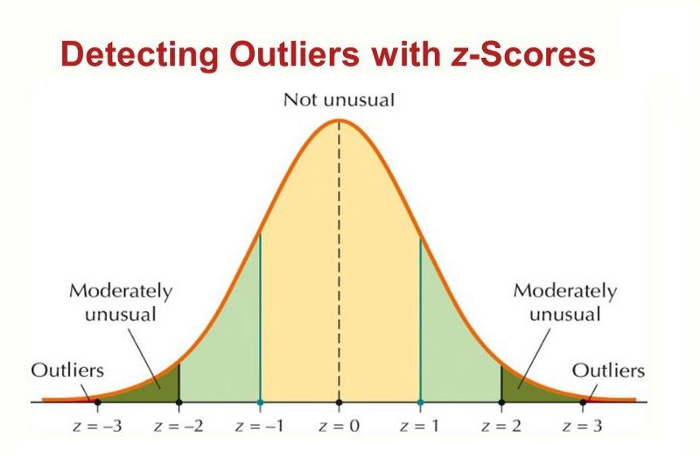
plt.xlabel('Sample')

                                            Detecting outlier using Boxplot (Image by author)



**4.2 Detecting outliers using the Z-scores**

**Criteria:**any data point whose Z-score falls out of 3rd standard deviation is an outlier.

Detecting Outliers with Z-scores

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.

import numpy as np

outliers = []

def detect\_outliers\_zscore(data):

thres = 3

mean = np.mean(data)

std = np.std(data)

# print(mean, std)

for i in data:

z\_score = (i-mean)/std

if (np.abs(z\_score) > thres):

outliers.append(i)

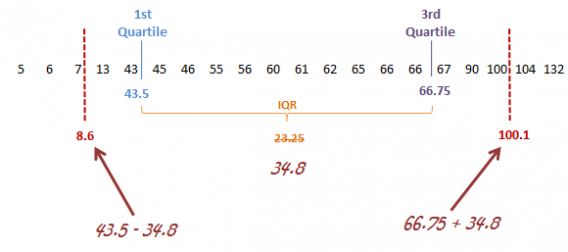
return outliers# Driver code

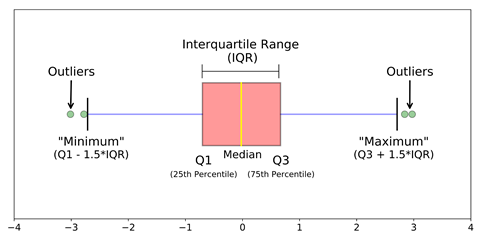
sample\_outliers = detect\_outliers\_zscore(sample)

print("Outliers from Z-scores method: ", sample\_outliers)

The above code outputs: **Outliers from Z-scores method: [101]**

**4.3 Detecting outliers using the Inter Quantile Range(IQR)**



                                                                IQR to detect outliers

**Criteria:** 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

Python Code:

outliers = []

def detect\_outliers\_iqr(data):

data = sorted(data)

q1 = np.percentile(data, 25)

q3 = np.percentile(data, 75)

# print(q1, q3)

IQR = q3-q1

lwr\_bound = q1-(1.5\*IQR)

upr\_bound = q3+(1.5\*IQR)

# print(lwr\_bound, upr\_bound)

for i in data:

if (i<lwr\_bound or i>upr\_bound):

outliers.append(i)

return outliers# Driver code

sample\_outliers = detect\_outliers\_iqr(sample)

print("Outliers from IQR method: ", sample\_outliers)

The above code outputs: **Outliers from IQR method**

5. Handling Outliers

Below are some of the methods of treating the outliers

* Trimming/removing the outlier
* Quantile based flooring and capping
* Mean/Median imputation

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.

# Trimming

for i in sample\_outliers:

a = np.delete(sample, np.where(sample==i))

print(a)

# print(len(sample), len(a))

The outlier ‘101’ is deleted and the rest of the data points are copied to another array ‘a’.

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.

Python code:

# Computing 10th, 90th percentiles and replacing the outliers

tenth\_percentile = np.percentile(sample, 10)

ninetieth\_percentile = np.percentile(sample, 90)

# print(tenth\_percentile, ninetieth\_percentile)b = np.where(sample<tenth\_percentile, tenth\_percentile, sample)

b = np.where(b>ninetieth\_percentile, ninetieth\_percentile, b)

# print("Sample:", sample)

print("New array:",b)

The above code outputs: **New array:**[15, 20.7, 18, 7.2, 13, 16, 11, 20.7, 7.2, 15, 10, 9]

The data points that are lesser than the 10th percentile are replaced with the 10th percentile value and the data points that are greater than the 90th percentile are replaced with 90th percentile value.

Mean/Median imputation

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

Python Code:

median = np.median(sample)# Replace with median

for i in sample\_outliers:

c = np.where(sample==i, 14, sample)

print("Sample: ", sample)

print("New array: ",c)

# print(x.dtype)

Visualizing the data after treating the outlier

plt.boxplot(c, vert=False)

plt.title("Boxplot of the sample after treating the outliers")

plt.xlabel("Sample")

