Outliers and Anomaly Detection

# Outlier Detection

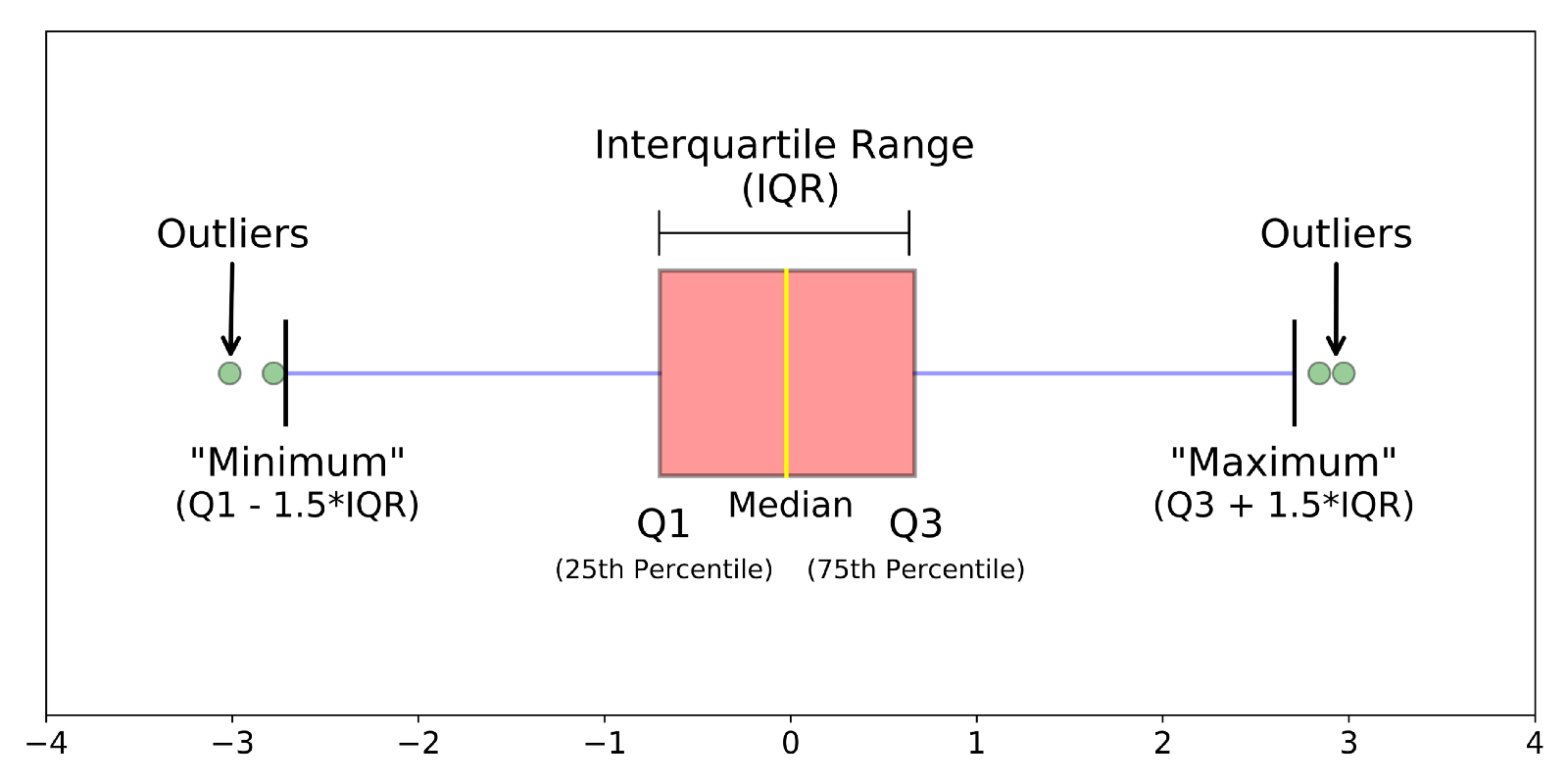
## IQR Method

This is the simplest, nonparametric outlier detection method in a one-dimensional feature space. Here outliers are calculated by means of the *IQR* (Interquartile Range).

The first and the third quartile (*Q1, Q3*) are calculated. An outlier is then a data point xi that lies outside the interquartile range. That is:

Equation

Using the interquartile multiplier value *k*=1.5, the range limits are the typical upper and lower whiskers of a box plot.



## Z-Score

Z-score is a parametric outlier detection method in a one or low dimensional feature space.

This technique assumes a Gaussian distribution of the data. The outliers are the data points that are in the tails of the distribution and therefore far from the mean. How far depends on a set threshold zthrfor the normalized data points zi calculated with the formula:

Equation

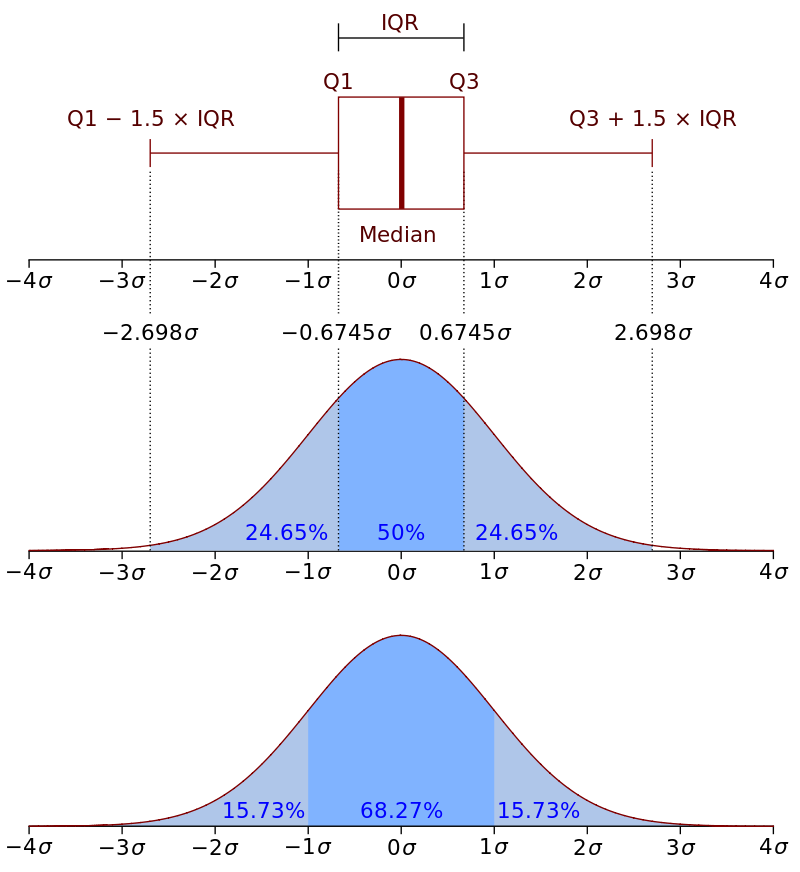
where xi is a data point, μ is the mean of all xi and is the standard deviation of all xi.

An outlier is then a normalized data point which has an absolute value greater than zthr. That is:

Equation

Commonly used zthr values are 2.5, 3.0 and 3.5.

Relation between IQR and Z Score methods



## Modified Z-score method

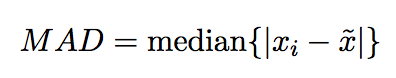
The limitation of the percentile-based methods: As the number of observations increases, so does the number of observations considered outliers; After all, using a percentile based method will always flat-out reject a certain percentage of our observations.

Knowing that our data is roughly normally distributed, we could use the **Z-score method**, by which we would consider points to be outliers based on how much they deviate from the mean value; However, the mean is not a robust statistic; It is heavily influenced by outliers, meaning that the outliers we are trying to detect would affect the method itself.

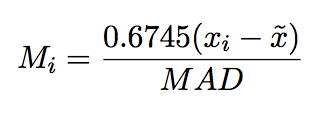
What if we take the same method but, instead of using the mean and standard deviation we use the median and the deviation from the median? The median is a robust statistic, meaning it will not be greatly affected by outliers. This is called the **Robust Z-score method**, and instead of using standard deviation, it uses the **MAD (Median Absolute Deviation).**

We will need:

* x̃ , which is just the median of the sample
* MAD, which is calculated by taking the absolute difference between each point and the median, and then calculating the median of those differences.



We can calculate the Modified Z-score like this:



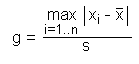
*0.6745 is the 0.75th quartile of the standard normal distribution, to which the MAD converges to*.

Now we can calculate the score for each point of our sample! As a rule of thumb, we’ll use the score of 3.5 as our cut-off value; This means that every point with a score above 3.5 will be considered an outlier.

## Grubbs’ Test

Grubbs' outlier test checks normally distributed data for outliers. This implies that one has to check whether the data show a normal distribution before applying the Grubbs test. The Grubbs test always checks the value which shows the largest absolute deviation from the mean. If an outlier has been identified and removed, the test must not be repeated without adapting the critical value.

The application of the test is quite simple and straightforward: one searches the maximum of the absolute differences between the values xi and the mean http://www.statistics4u.com/fundstat_eng/img/sc_16_xq.png. The result is divided by the standard deviation of the sample. If the resulting test statistic g is greater than the critical value, the corresponding value can be regarded to be an outlier. An extract of the critical values is shown in the following table:



|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| n | gcrit α=0.05 | gcrit α=0.01 |  | n | gcrit α=0.05 | gcrit α=0.01 |  | n | gcrit α=0.05 | gcrit α=0.01 |
| 3 | 1.1543 | 1.1547 |  | 15 | 2.5483 | 2.8061 |  | 80 | 3.3061 | 3.6729 |
| 4 | 1.4812 | 1.4962 |  | 16 | 2.5857 | 2.8521 |  | 90 | 3.3477 | 3.7163 |
| 5 | 1.7150 | 1.7637 |  | 17 | 2.6200 | 2.8940 |  | 100 | 3.3841 | 3.7540 |
| 6 | 1.8871 | 1.9728 |  | 18 | 2.6516 | 2.9325 |  | 120 | 3.4451 | 3.8167 |
| 7 | 2.0200 | 2.1391 |  | 19 | 2.6809 | 2.9680 |  | 140 | 3.4951 | 3.8673 |
| 8 | 2.1266 | 2.2744 |  | 20 | 2.7082 | 3.0008 |  | 160 | 3.5373 | 3.9097 |
| 9 | 2.2150 | 2.3868 |  | 25 | 2.8217 | 3.1353 |  | 180 | 3.5736 | 3.9460 |
| 10 | 2.2900 | 2.4821 |  | 30 | 2.9085 | 3.2361 |  | 200 | 3.6055 | 3.9777 |
| 11 | 2.3547 | 2.5641 |  | 40 | 3.0361 | 3.3807 |  | 300 | 3.7236 | 4.0935 |
| 12 | 2.4116 | 2.6357 |  | 50 | 3.1282 | 3.4825 |  | 400 | 3.8032 | 4.1707 |
| 13 | 2.4620 | 2.6990 |  | 60 | 3.1997 | 3.5599 |  | 500 | 3.8631 | 4.2283 |
| 14 | 2.5073 | 2.7554 |  | 70 | 3.2576 | 3.6217 |  | 600 | 3.9109 | 4.2740 |

There is a one-sided alternative which allows to test either the minimum xmin or the maximum xmax of the entire data set. The test statistics calculates according to the following formulas:

http://www.statistics4u.com/fundstat_eng/img/grubbs_statistic_onesided.png

A value can be regarded an outlier if the statistic g is greater than the critical value. Please note that in the case of the one-sided test the critical values are different. An extract is given below:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| n | gcrit α=0.05 | gcrit α=0.01 |  | n | gcrit α=0.05 | gcrit α=0.01 |  | n | gcrit α=0.05 | gcrit α=0.01 |
| 3 | 1.1531 | 1.1546 |  | 15 | 2.4090 | 2.7049 |  | 80 | 3.1319 | 3.5208 |
| 4 | 1.4625 | 1.4925 |  | 16 | 2.4433 | 2.7470 |  | 90 | 3.1733 | 3.5632 |
| 5 | 1.6714 | 1.7489 |  | 17 | 2.4748 | 2.7854 |  | 100 | 3.2095 | 3.6002 |
| 6 | 1.8221 | 1.9442 |  | 18 | 2.5040 | 2.8208 |  | 120 | 3.2706 | 3.6619 |
| 7 | 1.9381 | 2.0973 |  | 19 | 2.5312 | 2.8535 |  | 140 | 3.3208 | 3.7121 |
| 8 | 2.0317 | 2.2208 |  | 20 | 2.5566 | 2.8838 |  | 160 | 3.3633 | 3.7542 |
| 9 | 2.1096 | 2.3231 |  | 25 | 2.6629 | 3.0086 |  | 180 | 3.4001 | 3.7904 |
| 10 | 2.1761 | 2.4097 |  | 30 | 2.7451 | 3.1029 |  | 200 | 3.4324 | 3.8220 |
| 11 | 2.2339 | 2.4843 |  | 40 | 2.8675 | 3.2395 |  | 300 | 3.5525 | 3.9385 |
| 12 | 2.2850 | 2.5494 |  | 50 | 2.9570 | 3.3366 |  | 400 | 3.6339 | 4.0166 |
| 13 | 2.3305 | 2.6070 |  | 60 | 3.0269 | 3.4111 |  | 500 | 3.6952 | 4.0749 |
| 14 | 2.3717 | 2.6585 |  | 70 | 3.0839 | 3.4710 |  | 600 | 3.7442 | 4.1214 |

## Extreme Studentized Deviate

The generalized extreme Studentized deviate (ESD) test is used to detect one or more outliers in a univariate data set that follows an approximately normal distribution. The primary limitation of the **Grubbs test and the Tietjen-Moore test is that the suspected number of outliers, k, must be specified exactly.** If k is not specified correctly, this can distort the conclusions of these tests. On the other hand, the **generalized ESD test only requires that an upper bound for the suspected number of outliers be specified**.

Given the upper bound, r, the generalized ESD test essentially performs r separate tests: a test for one outlier, a test for two outliers, and so on up to r outliers.

The generalized ESD test is defined for the hypothesis:

|  |  |
| --- | --- |
| H0: | There are no outliers in the data set |
| Ha: | There are up to *r* outliers in the data set |
| Test Statistic: | Compute  ***R1=maxi|xi−x¯|/s***  with x¯ and *s* denoting the sample mean and sample standard deviation, respectively. Remove the observation that maximizes |xi−x¯| and then recompute the above statistic with *n* - 1 observations.Repeat this process until *r* observations have been removed. This results in the *r* test statistics *R1*, *R2*, ..., *Rr*. |

Corresponding to the *r* test statistics, compute the following *r* critical values

λi= tn−i−1,p(n−i)/ √ (n−i−1+t2n−i−1,p)(n−i+1)

where *i* = 1, 2, ..., *r*, tν,p is the 100*p* percentage point from the *t* distribution with ν degrees of freedom and

p=1−α/2(n−i+1).

The number of outliers is determined by finding the **largest *i* such that *Ri* > λi.**

Note that although the generalized ESD is essentially Grubbs test applied sequentially, there are a few important distinctions:

* + The generalized ESD test makes appropriate adjustments for the critical values based on the number of outliers being tested for that the sequential application of Grubbs test does not.
  + If there is significant masking, applying Grubbs test sequentially may stop too soon. The example below identifies 3 outliers at the 5% level when using the generalized ESD test. However, trying to use Grubbs test sequentially would stop at the first iteration and declare no outliers.
  + Grubbs test allows one-sided tests (i.e., you can specify a minimum test or the maximum test) in addition to two-sided tests (both the minimum and the maximum value are tested). The generalized ESD test is restricted to two-sided tests.

## **Dbscan (Density Based Spatial Clustering of Applications with Noise)**

Dbscan is a density based clustering algorithm, it is focused on finding neighbors by density (MinPts) on an ‘n-dimensional sphere’ with radius ɛ. A cluster can be defined as the maximal set of ‘density connected points’ in the feature space.

Dbscan then defines different classes of points:

* **C*ore point****:* **A** is a core point if its neighborhood contains at least the same number or more points than the parameter MinPts.
* ***Border point***: **C** is a border point that lies in a cluster and its neighborhood does not contain more points than MinPts, but it is still ‘*density reachable’*by other points in the cluster.
* ***Outlier***: **N** is an outlier point that lies in no cluster and it is not ‘*density reachable’* nor ‘*density connected’* to any other point. Thus this point will have “his own cluster”.

A cluster satisfies two properties:

* All points within the cluster are mutually density-connected.
* If a point is density-reachable from any point of the cluster, it is part of the cluster as well.

Dbscan estimates the number of clusters by itself, there is no need to specify the number of desired clusters, it is an unsupervised machine learning model.

Outliers (noise) will be assigned to the -1 cluster. After tagging those instances, they can be removed or analyzed.

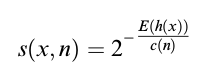
## Isolation Forest

Last but not least, isolation forests are an effective method for detecting outliers or novelties in data. It is a relatively novel method based on binary decision trees.

Isolation forest’s basic principle is that outliers are few and far from the rest of the observations. To build a tree (training), the algorithm randomly picks a feature from the feature space and a random split value ranging between the maximums and minimums. This is made for all the observations in the training set. To build the forest a tree ensemble is made averaging all the trees in the forest.

Then for prediction, it compares an observation against that splitting value in a “node”, that node will have two node children on which another random comparisons will be made. The number of “*splittings*” made by the algorithm for an instance is named: “*path length*”. As expected, outliers will have shorter path lengths than the rest of the observations.

An outlier score can computed for each observation:



**outlier score**

Where *h(x)*is the path length of the sample *x*, and *c(n)*is the ‘unsuccessful length search’ of a binary tree (the maximum path length of a binary tree from root to external node) *n*is the number of external nodes. After giving each observation a score ranging from 0 to 1; 1 meaning more outlyingness and 0 meaning more normality. A threshold can be specified (ie. 0.55 or 0.60)

**Z-Score pros:**

* It is a very effective method if you can describe the values in the feature space with a gaussian distribution. (Parametric)
* The implementation is very easy using pandas and scipy.stats libraries.

**Z-Score cons:**

* It is only convenient to use in a low dimensional feature space, in a small to medium sized dataset.
* Is not recommended when distributions can not be assumed to be parametric.

**Dbscan pros:**

* It is a super effective method when the distribution of values in the feature space can not be assumed.
* Works well if the feature space for searching outliers is multidimensional (ie. 3 or more dimensions)
* Sci-kit learn’s implementation is easy to use and the documentation is superb.
* Visualizing the results is easy and the method itself is very intuitive.

**Dbscan cons:**

* The values in the feature space need to be scaled accordingly.
* Selecting the optimal parameters eps, MinPts and metric can be difficult since it is very sensitive to any of the three params.
* It is an unsupervised model and needs to be re-calibrated each time a new batch of data is analyzed.
* It can predict once calibrated but is strongly not recommended.

**Isolation Forest pros:**

* There is no need of scaling the values in the feature space.
* It is an effective method when value distributions can not be assumed.
* It has few parameters, this makes this method fairly robust and easy to optimize.

**Isolation Forest cons:**

* Visualizing results is complicated.
* If not correctly optimized, training time can be very long and computationally expensive.

# Anomaly Detection – Time Series

## Moving Average Method

In this method with the help of moving average of past data, present day value is estimated. Moving average can be exponential Moving average or Simple Moving average. Exponential moving average gives more weight to recent data unlike equal weightage given by simple moving average.

This method can be executed in two ways **:**

### Moving Average on Daily Basis

This is very basic and easy to execute method among all the methods we’re going to discuss here.In this method the moving average of previous days is considered to be the expected value on present day. The next step includes checking whether estimated value is within predefined confidence band. Confidence Band is a range and we arrived at that defining it as a multiple of standard deviation from previous day’s Moving Average

### Moving Average on Weekly Basis

This method is slight modification in the method explained above. In this method, data from previous weekdays is taken into account. i.e to estimate value on present Monday, data from previous Mondays is useful. The confidence band is multiple of standard deviation of previous weekdays data.

## ARIMA

This is a bit more sophisticated method compared to above method. The ARIMA (Autoregressive integrated moving average) method is better than above explained methods because it uses the combination of auto-regression and moving average to estimate the value on present day.

## Prophet

Prophet” was Published by Facebook which uses additive regression model. This model helps in detecting anomalies .Prophet automatically detects changes in trends by selecting change points from the data and also do some modification in seasonal component(Year,Month) by some techniques like Fourier Transform.

Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data. Prophet is robust to missing data and shifts in the trend, and typically handles outliers well.

To train the model we will define basic hyperparameters some interval\_width and changepoint\_range. They can be used to adjust the width of the boundary.

Then we set as outliers everything higher than the top and lower the bottom of the model boundary. In addition, the set the importance of outlier as based on how far the dot from the boundary.

## SH-ESD

**Algorithm**

This two step process allows SH-ESD to detect both global anomalies that extend beyond the expected seasonal minimum and maximum and local anomalies that would otherwise be masked by the seasonality.

This modified algorithm is implemented on “Remainder” Component of time series.

Residual = Time series — Median — Seasonality

**“Why Remainder?”**

The “Remainder” term in the equation above is the “Unexplained Part” of the time series. The residual has a unimodal distribution that is amenable to the application of anomaly detection techniques such as ESD([Extreme Studentized Deviate](https://www.itl.nist.gov/div898/handbook/eda/section3/eda35h3.htm)).

**Residual Calculation** :

As explained above the residual calculation is important for unexplained behavior purpose. In calculation of Residual , Median and Seasonality was removed from original value.

Here seasonality is calculated using[STL](https://en.wikipedia.org/wiki/Decomposition_of_time_series) variant method(uses Loess method to find the seasonality). Median is used instead of trend to avoid spurious anomalies.

**“Why Seasonality is removed?”**

Consider a movie booking application where sales of movie tickets goes high on every Saturday, then high value on Saturday in particular week should not come as a surprise to a marketer. Hence, seasonality is removed to avoid fake anomalies due to seasonal behavior.

**“What to do with the residual component ?”**

We performed the Modified ESD test on Residual data in order to find the outliers. This Modified ESD test is a two step process :

. The first step in this test is calculation of modified z-score. It’s a measure of how many mean deviations below or above the median a raw score. Higher the z-score values means higher the mean deviation from median.

Modified Z-Score = (X-Median)/MAD

X = Original data

MAD = Mean Absolute Deviation

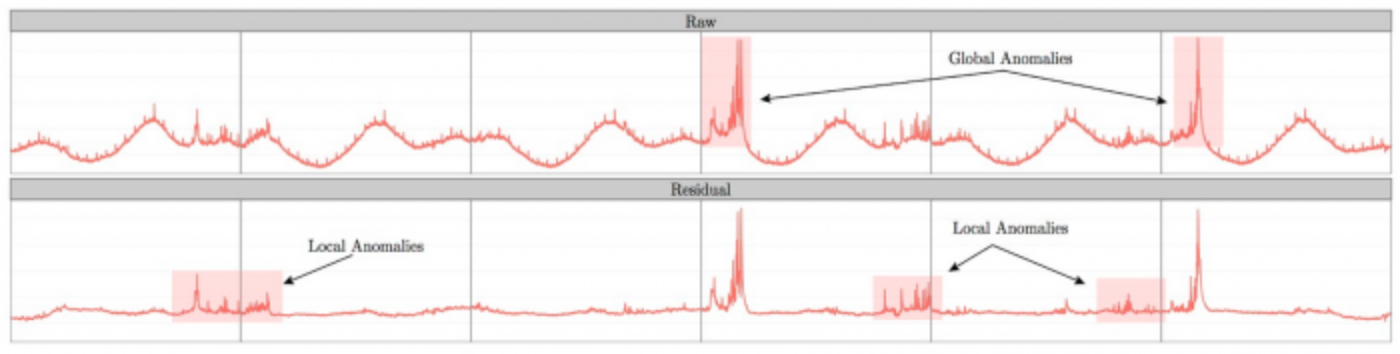
Note : Median and MAD(Mean Absolute Deviation) terms are used instead of “Mean” and “Standard Deviation” unlike general z-score. The reason being later has more effect of outliers on them. This explains the “modified” part in the above term.

. The second step starts with removing the values with the maximum deviation from median. We again calculated the new values for Z-Score. After applying the confidence on this new modified z-score values will give us the actual outliers. The values in original data corresponding to the index of values in outliers in residual data are anomalies.

# Packages

## Twitter – AnomalyDetection

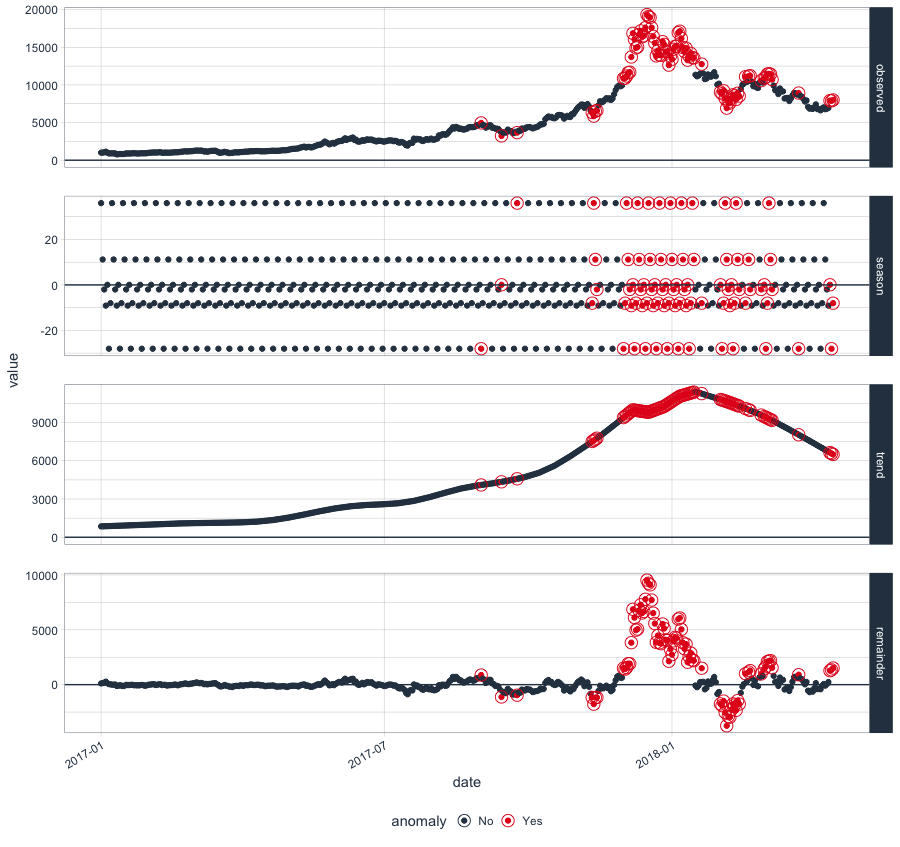
The primary algorithm, Seasonal Hybrid ESD (S-H-ESD), builds upon the Generalized ESD test [3] for detecting anomalies. S-H-ESD can be used to detect both global and local anomalies. This is achieved by employing time series decomposition and using [robust statistical metrics](http://www.wiley.com/WileyCDA/WileyTitle/productCd-0470129905.html), viz., median together with ESD. In addition, for long time series such as 6 months of minutely data, the algorithm employs piecewise approximation. This is rooted to the fact that trend extraction in the presence of anomalies is non-trivial for anomaly detection .

The figure below shows large global anomalies present in the raw data and the local (intra-day) anomalies that S-H-ESD exposes in the residual component via our statistically robust decomposition technique.[](https://g.twimg.com/blog/blog/image/figure_raw_residual_global_local.png)

Besides time series, the package can also be used to detect anomalies in a vector of numerical values. We have found this very useful as many times the corresponding timestamps are not available. The package provides rich visualization support. The user can specify the direction of anomalies, the window of interest (such as last day, last hour) and enable or disable piecewise approximation. Additionally, the x- and y-axis are annotated in a way to assist with visual data analysis.

## Anomalize

The first step is the time series decomposition using time\_decompose(). The measured value or the numerical value on which detection needs to be performed for a particular group is decomposed into four columns that are **observed**, **season**, **trend**, and **remainder**. The default method used for decomposition is **stl**, which is a seasonal decomposition utilizing a Loess smoother.



### How It Works

anomalize has three main functions:

* [time\_decompose()](https://business-science.github.io/anomalize/reference/time_decompose.html): Separates the time series into seasonal, trend, and remainder components
* [anomalize()](https://business-science.github.io/anomalize/reference/anomalize.html): Applies anomaly detection methods to the remainder component.
* [time\_recompose()](https://business-science.github.io/anomalize/reference/time_recompose.html): Calculates limits that separate the “normal” data from the anomalies!

Suppose we want to determine which daily download “counts” are anomalous. It’s as easy as using the three main functions ([time\_decompose()](https://business-science.github.io/anomalize/reference/time_decompose.html), [anomalize()](https://business-science.github.io/anomalize/reference/anomalize.html), and [time\_recompose()](https://business-science.github.io/anomalize/reference/time_recompose.html)) along with a visualization function, [plot\_anomalies()](https://business-science.github.io/anomalize/reference/plot_anomalies.html).

If you’re familiar with Twitter’s AnomalyDetection package, you can implement that method by combining [time\_decompose(method = "twitter")](https://business-science.github.io/anomalize/reference/time_decompose.html) with [anomalize(method = "gesd")](https://business-science.github.io/anomalize/reference/anomalize.html). Additionally, we’ll adjust the trend = "2 months" to adjust the median spans, which is how Twitter’s decomposition method works.

### STL - Seasonal and Trend decomposition using Loess and IQR Method

Loess regression is the most common method used to smoothen a volatile time series, it fits multiple regression in local neighborhood, you can also say that the data is divided and regression is applied to each part, which is useful in time series because we know the bound of time which is the X variable in this case. This method works well in the case where the trend dominates the seasonality of the time series.

Here trend is long-term growth that happens over many observations and seasonality is the cyclic pattern occurring on a daily cycle for a minute or an hour or weekly.

### Piece Wise Median (Twitter Anomaly Detection method) and SH – ESD Method

There is a second technique which you can use for seasonal decomposition in time series based on median that is the Twitter method which is also used AnomalyDetection package. It is identical to STL for removing the seasonal component. The difference is in removing the trend is that it uses piece-wise median of the data(one or several median split at specified intervals) rather than fitting a smoother.

This method works well where seasonality dominates the trend in time series.

After the time series analysis is complete and the remainder has the desired characteristics to perform anomaly detection which again creates three new columns.

remainder\_l1 : The lower limit of the remainder.

remainder\_l2 : The upper limit of the remainder.

anaomaly : Column is telling whether the observation is an anomaly or not.

Anomalies are high leverage points that distort the distribution. The anomalize implements two methods that are resistant to high leverage points:

IQR: Inner Quartile Range

GESD: Generalized Extreme Studentized Deviate Test

### IQR

It is a similar method used in tsoutliers() function of the forecast package. In IQR a distribution is taken and 25% and 75% inner quartile range to establish the distribution of the remainder. Limits are set by default to a factor of 3 times above, and below the inner quartile range, any remainder beyond the limit is considered as an anomaly.

### GESD

In GESD anomalies are progressively evaluated removing the worst offenders and recalculating the test statistics and critical values, or more simply you can say that a range is recalculated after identifying the anomalies in an iterative way.