@author: Fred Coerver

This module takes a pandas frame as input, together with several hyperparameters.

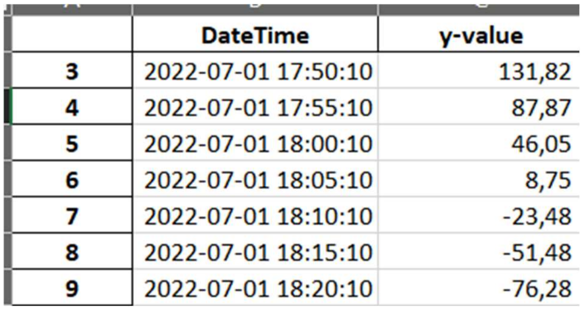
The pandas data frame must have a straight forward x and y column including an index. Not more Not less.

The output are two pandas dataframes. One table with statistical information of the input and whether a value x is identified as a spike or dip in the dataset, based on the input parameters. And the second is a list of spike. More details about the output below. The algorithm makes a window of datapoints from x - windowsizeleft to x + windowsizeright

Linear regression is performed on this x-range with the give y-values. The regression value : y = slope*x + intercept

This window is iterated over the pandas frame and the regression values are put into the output table, along with the other statistical values.

Input dataframe structure :



	DateTime	y-value
3	2022-07-01 17:50:10	131,82
4	2022-07-01 17:55:10	87,87
5	2022-07-01 18:00:10	46,05
6	2022-07-01 18:05:10	8,75
7	2022-07-01 18:10:10	-23,48
8	2022-07-01 18:15:10	-51,48
9	2022-07-01 18:20:10	-76,28

1) pandas index as int
2) x or DateTime value format %Y-%m-%d %H:%M:%S {working on an update that also accepts floats as x}. Column name must be "DateTime" if Timescale is true and must be next/right to the index
3) y {float or int}. Columnn name is free.

e.g. Regression is internally in the model executed on :

y=[115730244, 117300778, 116863076, 108493789, 97012607, 95430706]

x=[1657540800, 1657544400, 1657548000, 1657555200, 1657558800, 1657562400]

The data frame can have more columns, but the script only analyse 1 columns at a time.

So in case you want to analyse more columns, you need to call the script for each column and slicing

the data frame in a way that the input frame complies to 1), 2) and 3)

Hyperparameters

ignore_startsamples = 5 (in case you want to omit starting rows from your calculation of the dataset. This parameter does not delete the rows! = default is 5. In case you increase the windowsizeleft, you might need to increase this value as well)

ignore_endsamples = 3 (in case you want to omit starting rows from your calculation of the dataset. This parameter does not delete the rows! In case you increase the windowsizeright, you might need to increase this value as well)

P1inc = 10 ---> In case the standard deviation of a regression window is P1inc-times higher as the previous std, then this x value is marked as a spike

accuracy = 4 ---> If a y-value exceeds the regressed value +- accuracy * std then this x-value is marked as spike

window_sizeleft = 3 ---> the regresssionwindow is #windowsizeleft x-values from the analysed x-point and #windowsizeright x-values from the analyzed x-point. So suppose the algorythm calculates whether x[j] is a spike, then the window x[j-windowsizeleft] until x[j+windowsizeright] is considered as the window of x-points. From these x- an y-point the regression is calculated, and the regression value is compared with the actuel y-value

window_sizeright = 3 ---> amount of datapoint right from the x-value, which is subject for analyzing whether it is a spike/dip or not

ignoreendsamples = 3 ---> the Endresult output dataframe is capped with [ignorestartsamples:-ignoreendsamples] before returning to main

Timescale ---> If x-values are different from time then any string is valid, in case the x-values need to be in timeformat then this parameter is True.

OUTPUT:

1) Pandas dataframe with format columns:

```
   'DateTime', 'y-value', 'KPI_name', 'Regr_value', 'Value', 'Unixtime',
     'Std', 'spike', 'spikevalue', 'Slope', 'Intercept', 'Diff',
     'Regr_value_plus_std', 'Regr_value_min_std'
```

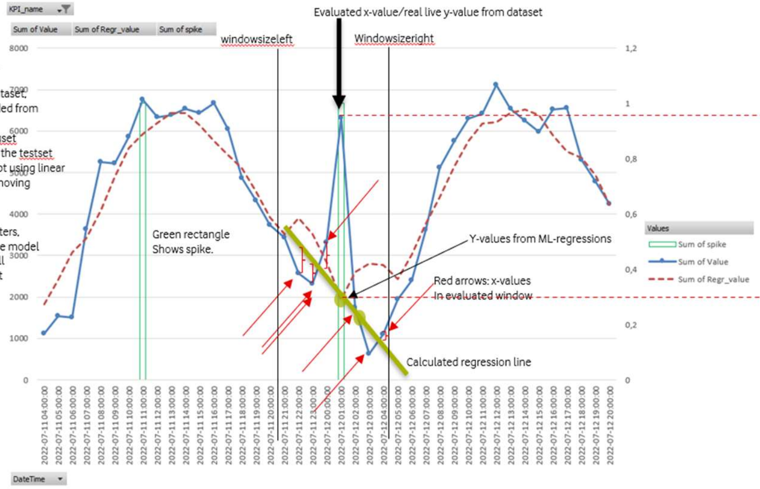
2) Pandas dataframe as 1) but then sorted and 'spike' ==1 or ==-1

"""

Fast Anomaly Detection

● Each datapoint is evaluated.
The datapoint itself is
excluded from regression dataset,
Just like the testset is excluded from
The trainingset in ML
So the window is the trainingset
And the evaluated x-value is the testset
In essence it is an ML concept using linear
Regression adaptive to the moving
Datasetvalues.

Depending on hyperparameters,
Accuracy and windowsize the model
Indicates whether a point will
be classified as a spike or not



Evaluated x-value/real live y-value from dataset

windowsizeleft

Windowsizeright

Green rectangle
Shows spike.

Y-values from ML-regressions

Red arrows: x-values
In evaluated window

Calculated regression line

Difference between
actual value and model
Value.

`show differences
Between regression
line and real datapoints
, which is the basis
for calculating the
Standard deviation