Part Two

The table below contains the statistics requested, displayed as floating point numbers.

	DataType	Averages	Standard Deviation	Minimum Value	Maximum Value
0	mm	1.553125	3.328302	0.000000	23.200000
1	Temperature_outside	11.138877	5.355042	-1.810000	26.380000
2	Temperature_range (low)_outside	7.865634	4.878930	-4.100000	18.700000
3	Temperature_range (high)_outside	15.522535	7.034981	1.500000	38.500000
4	Humidity	48.519774	5.188886	37.000000	59.000000
5	Temperature_indoor	21.827885	2.058307	18.040000	29.210000
6	Temperature_range (low)_indoor	20.555932	2.405125	14.900000	28.200000
7	Temperature_range (high)_indoor	23.532768	1.702157	19.700000	31.100000
8	Baro	1009.996338	9.868772	979.600000	1035.600000

Manually Change Values Within the CSV and Observe Changes To Statistics

The objective is to introduce outliers into the data. The first outlier will be a value which introduces a material change (becomes the new Max Value) which can easily be picked up using traditional statistical tests, for example the Grubb Test.

The second two outliers are more nuanced, and sneak past the Grubb Test undetected. What I hope to show is that statistical methods for identifying outliers are important, however they alone are insufficient to pick up all spurious values that may be encountered.

From “indoor-temperature-1617.csv”, I will change:

i. DateTime “2016-10-09” **Humidity** Value from 54 to 90 (Material Outlier)
ii. DateTime “2017-01-01” **Humidity** value from 45 to 57 (Nuanced Outlier)
iii. DateTime “2017-09-01” **Humidity** value from 55 to 45 (Nuanced Outlier)

From “outside-temperature-1617.csv”, I will change:

i. DateTime “2016-10-09” **Temperature** Value from 10.66 to 40 degrees (Material Outlier)
ii. DateTime “2017-01-01” **Temperature** Value from 4.9 to 17 degrees (Nuanced Outlier)
iii. DateTime “2017-09-01” **Temperature** Value from 14.78 to 5 degrees (Nuanced Outlier)

New Data Table

	DataType	Averages	Standard Deviation	Minimum Value	Maximum Value
0	mm	1.553125	3.328302	0.000000	23.200000
1	Temperature_outside	11.236511	5.585707	-1.810000	40.000000
2	Temperature_range (low)_outside	7.865634	4.878930	-4.100000	18.700000
3	Temperature_range (high)_outside	15.522535	7.034981	1.500000	38.500000
4	Humidity	48.632768	5.647397	37.000000	90.000000
5	Temperature_indoor	21.827885	2.058307	18.040000	29.210000
6	Temperature_range (low)_indoor	20.555932	2.405125	14.900000	28.200000
7	Temperature_range (high)_indoor	23.532768	1.702157	19.700000	31.100000
8	Baro	1009.996338	9.868772	979.600000	1035.600000

Humidity Changes

Inspecting Data Only

i. Having introduced the value 90, which is significantly higher than the previous Maximum Value of 59. It is conceivable that one is able to spot the Outlier using the Maximum Value when comparing it against the Standard Deviation and Average. This is not overly empirical, however it can act as a good first line of defence for spotting extreme case data.
ii. The Average has barely moved (48.52 to 48.63), therefore it is not informative when viewed in isolation.
iii. The Standard Deviation has changed marginally from 5.2 to 5.6. Again, not informative when viewed in isolation.

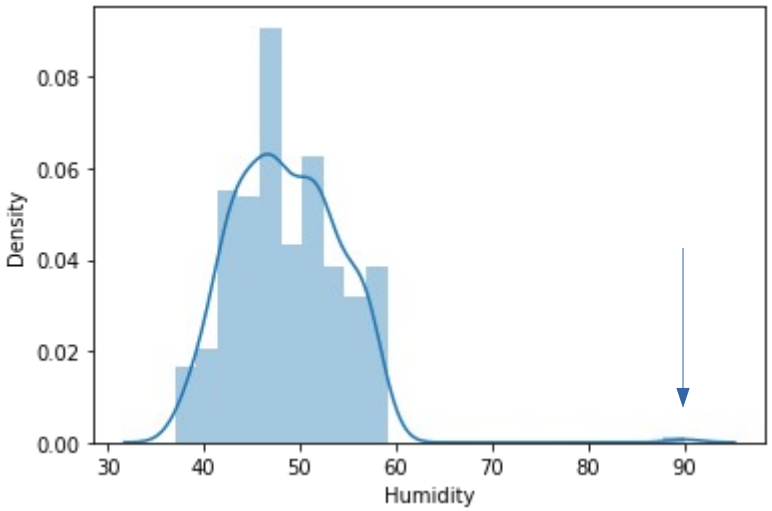
Grub Test

$$G = \frac{\max_{i=1,\ldots,N} \left|Y_i - \bar{Y}\right|}{s}$$

DateTime "2016-10-09" HumidityValue from 54 to 90

G (usually denoted with a Z), for our spurious value of 90 is 7.3221. Therefore, P(X<90) = 1, and P(X≥90) = 0. This is significant, as it is telling us that there is a 0% probability that 90 appears using the average and standard deviation. Therefore, we are able to detect our spurious outlier using the statistics produced by the data.

The concept is illustrated below using the distribution curve. The arrow is pointing to X = 90. As we can see, the probability density is approximately zero.



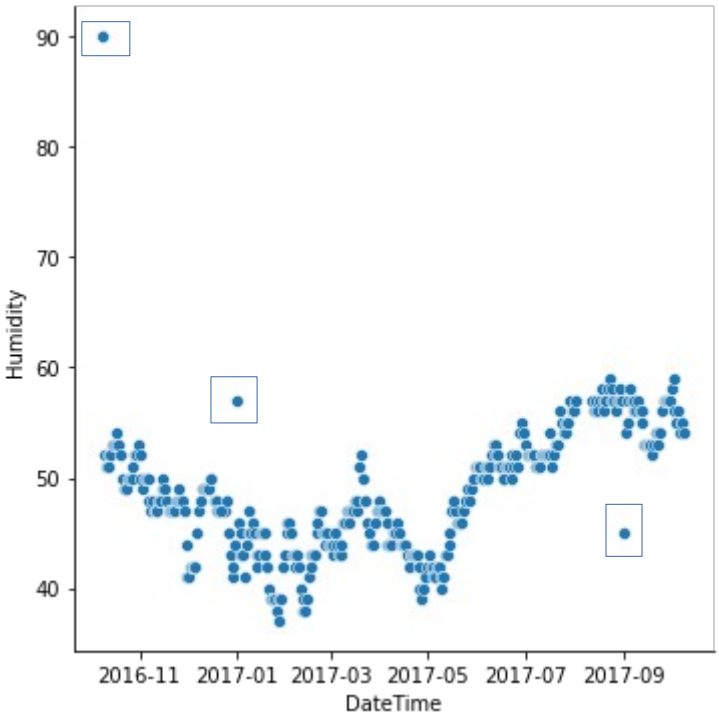
DateTime "2017-01-01" Humidity value from 45 to 57.

G for our spurious value of 57 is 1.484. Therefore, P(X<57) = 93.11%, and P(X≥90) = 6.89%. I will use a critical value of 5%, and as such I cannot reject the null hypothesis that X is not an outlier.

DateTime "2017-09-01" Humidity value from 55 to 45

G for our spurious value of 45 is -0.64632. Therefore, P(X<45) = 26%, and P(X≥45) = 74%. This is quite comfortably within the distribution for values we would expect.

Why have I assumed that 57 and 45 are outliers at their respective dates? Below is a scatter plot, in which the three outliers discussed are marked in boxes. Visualising is an extra tool in addition to using statistics, to quickly spot outliers hidden in the data, especially when the function is nonlinear.



Temperature Value (Outside)

Inspecting Data Only

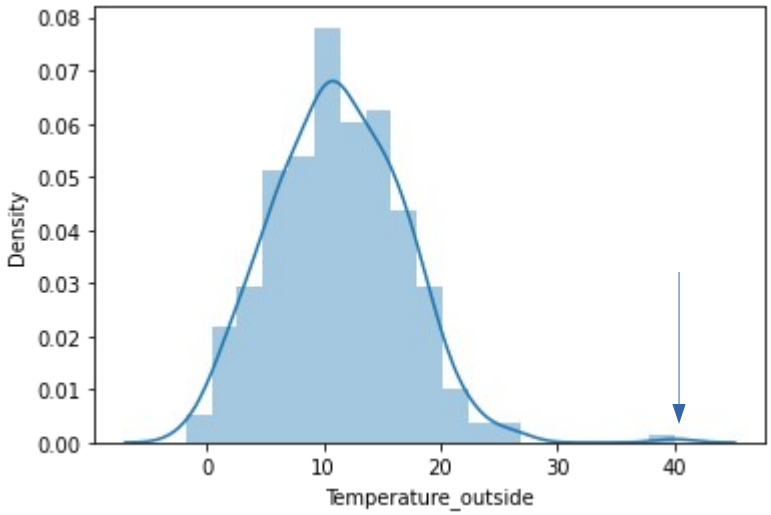
i. Having introduced the value 40, which is significantly higher than the previous maximum value of 26.38. The new maximum value is probably identifiable as an outlier from a good understanding of the average and standard deviation.
ii. The Average has barely moved, from 11.14 to 11.24, this is not informative.
iii. The Standard Deviation has changed marginally from 5.35 to 5.59, again not informative.

Grub Test

DateTime "2016-10-09" Temperature Value from 10.66 to 40

G for our spurious value of 40 is 5.155. Therefore, $P(X<40) = 1$, and $P(X\geq40) = 0$. This is significant, as it is telling us that there is a 0% probability that 40 appears using the average and standard deviation. Therefore, we are able to detect our spurious outlier using the statistics produced by the data.

The concept is illustrated below using the distribution curve. The arrow is pointing to X = 40. As we can see, the probability density is approximately zero.



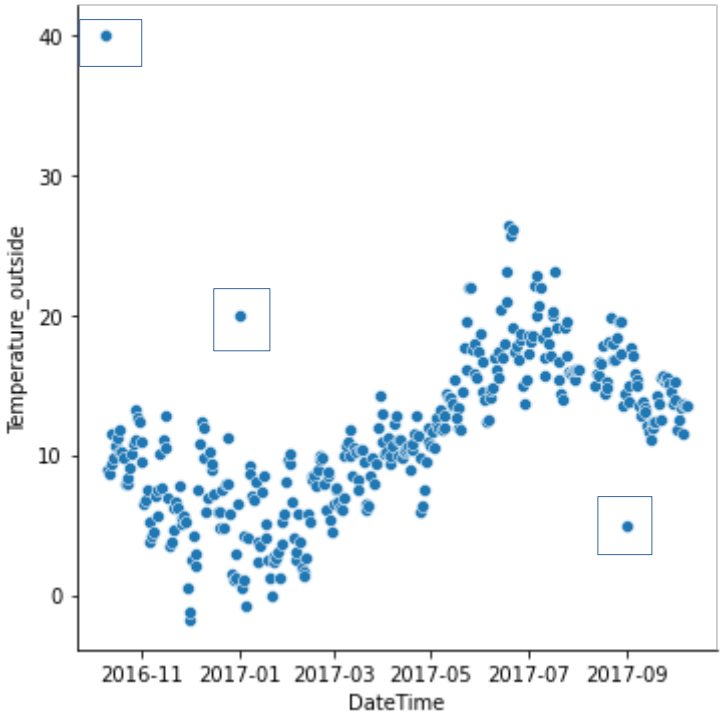
DateTime "2017-01-01" Temperature Value from 4.9 to 17 degrees

G for our spurious value of 17 is 1.034. Therefore, P(X<17) = 85 %, and P(X≥17) = 15%. I will use a critical value of 5%, and as such I cannot reject the null hypothesis that X is not an outlier.

DateTime "2017-09-01" Temperature Value from 14.78 to 5 degrees

G for our spurious value of 5 is -1.12. Therefore, the P(X<5) = 13 %, and P(X≥5) = 87%. This is quite comfortably within the distribution of values we would expect.

Again, let's observe a scatter plot, to show that all three values are in fact outliers.



Conclusion

1) Outliers which exist at the extreme ends of the distribution, i.e. that are clearly identifiable from the Maximum and Minimum Values in the table, can be contextualised using averages and standard deviation. This is not rigorous, but is a useful first defence.
2) If a function is nonlinear, then distribution data fails to pick up outliers entirely. When faced with this, we need to inspect the data visually.