APPLICATIONS OF AI – ASSIGNMENT 1 / PART_2

The statistics computed for each component of the weather data from each csv file are as follows,

STATISTICS

Parsed File Name: **indoor-temperature-1617.csv**

	Humidity	Temperature	Temperature_range (low)	Temperature_range (high)
Mean	48.52	21.83	20.56	23.53
Max	59	29.21	28.2	31.1
Min	37	18.04	14.9	19.7
Std_Dev	5.18	2.06	2.4	1.7

Parsed File Name: **outside-temperature-1617.csv**

	Temperature	Temperature_range (low)	Temperature_range (high)
Mean	11.14	7.87	15.52
Max	26.38	18.7	38.5
Min	-1.81	-4.1	1.5
Std_Dev	5.35	4.87	7.02

Parsed File Name: **barometer-1617.csv**

	Baro
Mean	1010.0
Max	1035.6
Min	979.6
Std_Dev	9.86

Parsed File Name: **rainfall-1617.csv**

	mm
Mean	1.55
Max	23.2
Min	0.0
Std_Dev	3.32

CSV FILE MODIFICATION

I modified the file "**indoor-temperature-1617.csv**" by introducing an outlier on its "Temperature" field. I replaced one value which was originally "23.93" with "2100.93" that we normally should not read as indoor temperature value from a sensor so basically the new value became an outlier.

After this modification I run my program again and calculated the summary statistics on the modified CSV file. The below are the comparison between the old and the new results for the "temperature" field.

Parsed File Name: indoor-temperature-1617. csv

	Humidity	Temperature (OLD)	Temperature (NEW)	Temperature_range (low)	Temperature_range (high)
Mean	48.52	21.83	27.7	20.56	23.53
Max	59	29.21	2100.93	28.2	31.1
Min	37	18.04	18.04	14.9	19.7
Std_Dev	5.18	2.06	110.37	2.4	1.7

As can be seen above, when we look at the "Temperature (NEW)" field we can spot that Max value is 2100.93 which is an unexpected value measured for a temperature, and the Standard Deviation "110.37" is around 4 times of the Mean "27.7" which is not normal as well. So, we can easily conclude that there is at least one outlier here. By looking at the Max value we can conclude that at least there is one outlier which has the value of "2100.93". For this particular example, we can say that the statistics are enough to identify the incorrect data.