Program Name: program_09.py
Programmer Name: Tyler Field
Programmer Username: tfield

Program Overview:

The program loads in the file *DataQualityChecking.txt* and runs four checks for data quality. There are four variables in the dataset: Precipitation (mm), Maximum Temperature (°C), Minimum Temperature (°C), and Wind Speed (m/s). Summary statistics are output before and after each check to analyze the changes made, and this data is compiled into a final table summarizing the effect of each quality check. Additionally, plots for each variable, before and after each relevant check are outputted as PNG images for graphical confirmation of significant changes.

Data Quality Checking:

There are four data quality checks performed by this program:

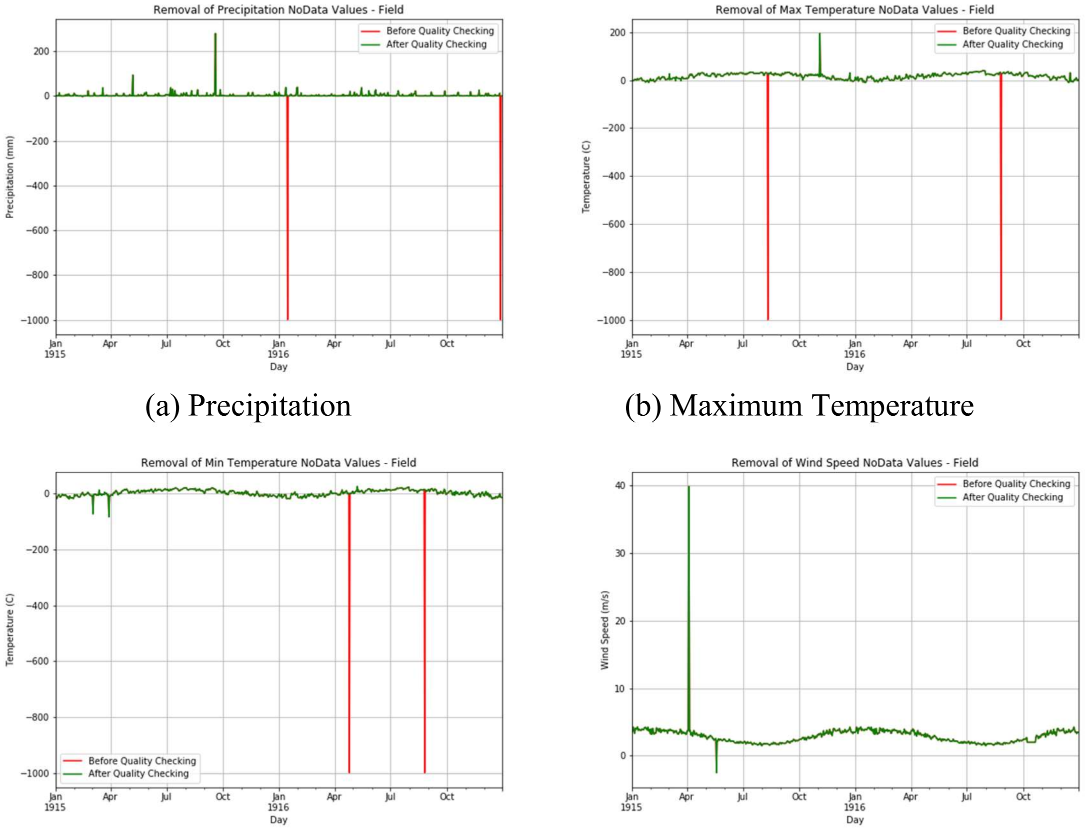
1. Removal of no data values (-999) from all four variables
2. Removal of gross error values from all four variables using the following ranges:
    a. $0 \leq P \leq 25$
    b. $-25 \leq T \leq 35$
    c. $0 \leq WS \leq 10$
3. Swapping the value for max and min temperature for a given day if the maximum temperature is less than the minimum temperature
4. Removal of days where the maximum temperature exceeds the minimum temperature by more than $25°C$

After each quality check, appropriate figures were generated and saved as PNG files. Each file is saved in the format **QC_[QC check number].[variable number]_[QC check type]_[variable name].png**. All removals were performed by converting the offending value to NaN. The change to the dataset after each quality check can be seen in Table 1.

Table 1: Number of Changed Values after each Quality Check

Quality Check	Precipitation	Maximum Temperature	Minimum Temperature	Wind Speed
1. No Data	2	2	2	0
2. Gross Error	15	14	2	2
3. Swapped	0	4	4	0
4. Range Fail	0	5	5	0

As can be seen in the No Data row of Table 1, two values were missing from each of the first three variables. This led to the noticeable difference before and after the quality check in Figure 1a, Figure 1b, and Figure 1c, but no change in Figure 1d, since Wind Speed had no missing values.



(a) Precipitation

(b) Maximum Temperature

(c) Minimum Temperature

(d) Wind Speed

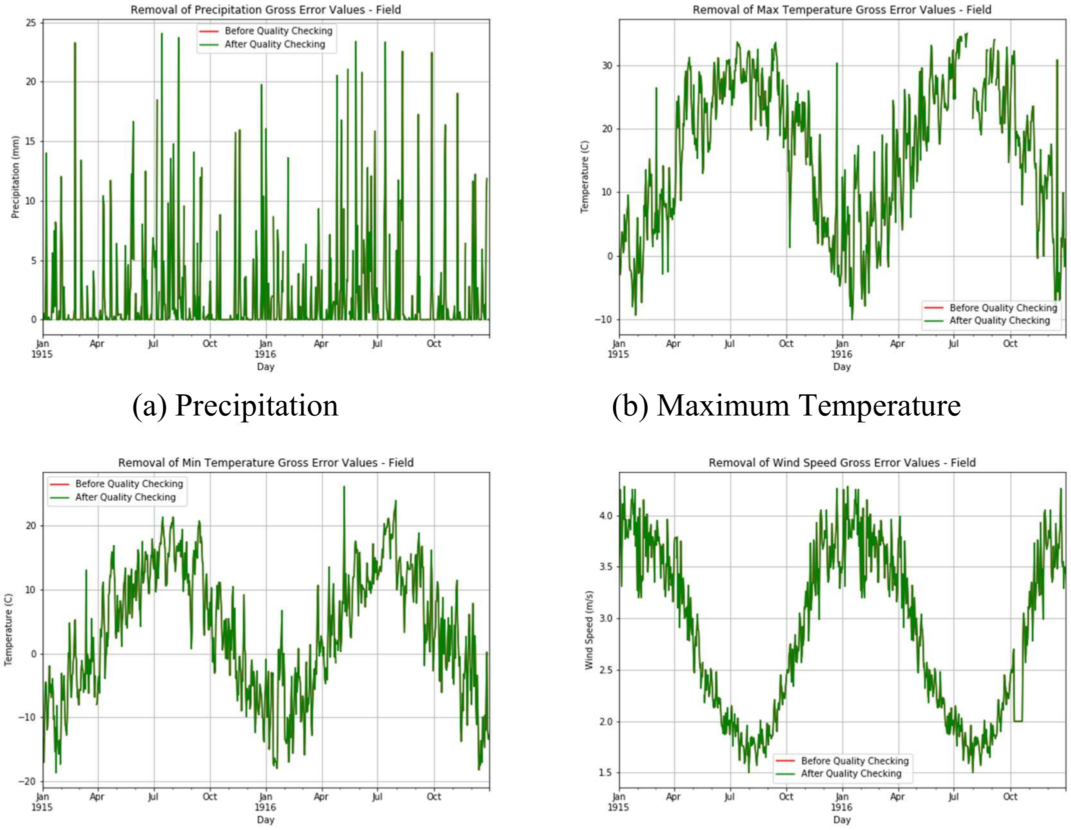
Figure 1: Pre- & Post- Data Quality Check #1

All four variables had at least two gross errors removed during quality check #2, as seen in the Gross Error row of Table 1. Although it is not easily possible to observe the red, pre-quality check lines in (c) Minimum Temperature (d) Wind Speed

Figure 2, due to the overlap of the green, post-quality check lines, the post-quality check lines can be seen to be within the prescribed boundaries defined at the beginning of this section. Additionally, using the max and min lines from the *describe()* method before and after the quality check is performed, its success can be verified. This is shown in Table 2.

Table 2: Pre- & Post- Quality Check #2 Max and Min Values

	Precipitation	**Maximum Temperature**	**Minimum Temperature**	**Wind Speed**
Pre-Min	-3.475	-10.080	-82.600	-2.500
Post-Min	0	-10.080	-18.630	1.500
Pre-Max	279.000	194.800	26.100	39.900
Post-Max	24.050	34.960	26.100	4.280



(a) Precipitation

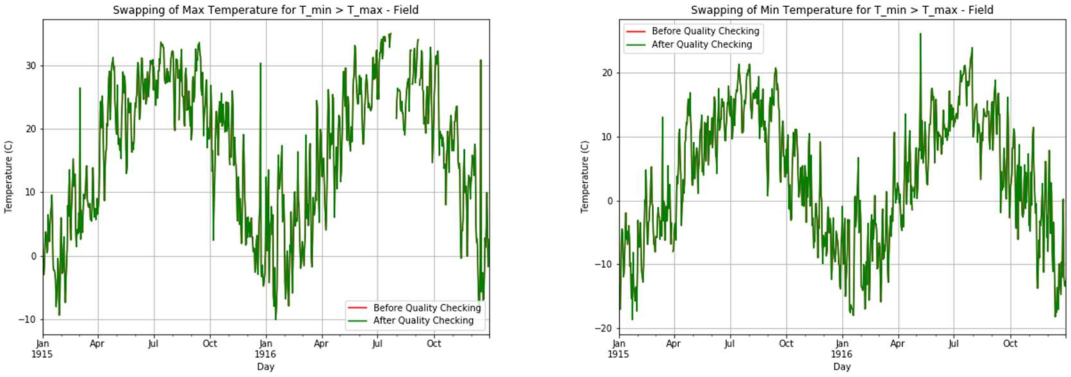
(b) Maximum Temperature

(c) Minimum Temperature

(d) Wind Speed

Figure 2: Pre- & Post- Data Quality Check #2

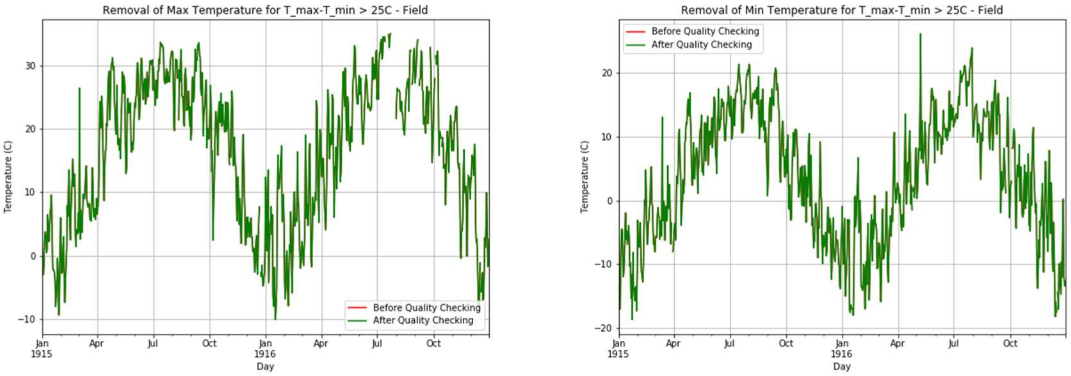
As data quality checks 3 and 4 only pertain to the maximum and minimum temperature variables, no figures are generated for precipitation and wind speed, since there will be no change beyond the post-quality check #2 lines shown in (c) Minimum Temperature
(d) Wind Speed
Figure 2a and (c) Minimum Temperature (d) Wind Speed
Figure 2d. Additionally, data quality checks 3 and 4 are performed in unison on both temperatures, resulting in the equal numbers for the two temperatures in the Swapped and Range Fail rows of Table 1, and zeros for the Precipitation and Wind Speed variables in those rows, since the quality checks do not affect them.



(a) Maximum Temperature (b) Minimum Temperature

Figure 3: Pre- & Post- Data Quality Check #3

Figure 3 shows the swapping of maximum and minimum temperatures when the maximum is less than the minimum, but no discernable change can be seen, likely due to the high frequency of data and only four values being swapped. This is also what occurs in (a)
Maximum Temperature (b) Minimum Temperature
Figure 4, although five values are removed due to the range being greater than $25°C$.



(a) Maximum Temperature (b) Minimum Temperature

Figure 4: Pre- & Post- Data Quality Check #4

By summing the values in each column of Table 1, the total number of changes made to each variable can be calculated, as shown in Table 3. This can provide feedback on which subsystems may need improvements or may have failed.

Table 3: Total Quality Check Changes Made to Each Variable

	Precipitation	**Maximum Temperature**	**Minimum Temperature**	**Wind Speed**
Total Number of Quality Improvements	17	25	13	2