Author – Eric Kong
Date – 3/30/2020
File Information – Automated Data Quality Checking with Python

Program Name – program_09
Program Description – The script is a modified template provided by the instructor. It performs an automated data quality checking by using functions. Raw and processed data is plotted to show the results.

Data Quality Checks & Effects – The script performs an automated data quality checking by using functions. First, there is an error check for “no data” entries. Then values outside thresholds for the 4 different data types within the set is replaced with NaN. Some entries for minimum and max temp are in the wrong column, so those are swapped to the correct position during the third check. Finally, the difference in high and low temp for the day is calculated to determine if it is outside the acceptable difference. The results of these data quality checks can be seen below in Table 1. The number in each column represents the number of errors that occurred for the specific data during the check (I.e 15 data points fell outside of the acceptable range during the gross error check of the precipitation data).

Each error check affects the contents of the file in different ways. The ‘no data’ check ensures that the “no data” value that is equal to -999 is replaced with NaN so that it does not mess with the results when plotted. This is the overall goal for each data check. The user wants results without errors and outliers. You do not want the person viewing the results of your work to see an outlier and start assuming why it is there. All of the figures, besides Figure 4, shows two blue stars at -999, due to the no data values.

The gross error changes data outside of an established threshold to NaN so values that are consider too high or low appear. The final results, Figures 1- 4, including the gross error checks can be seen at the bottom of this document. There are obvious values that are not “no data” but fall outside the acceptable range.

Recorded data has the possibility of containing the right data in the wrong location. The swap check ensures that min and max temperature are not flipped. According to Table 1, this happened four times, but this is hard to see on the plots.

Temperature changes throughout the day, especially in the Midwest, but it should not rise or fall 25 Celsius. The range fail check prevents this error from making its way into the results. This is also hard to see on the the Figure 1 and 2 because minimum and maximum of the day are not on one plot.

Table 1: Number of Corrections Made During Error Checks

	Precip	Max Temp	Min Temp	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

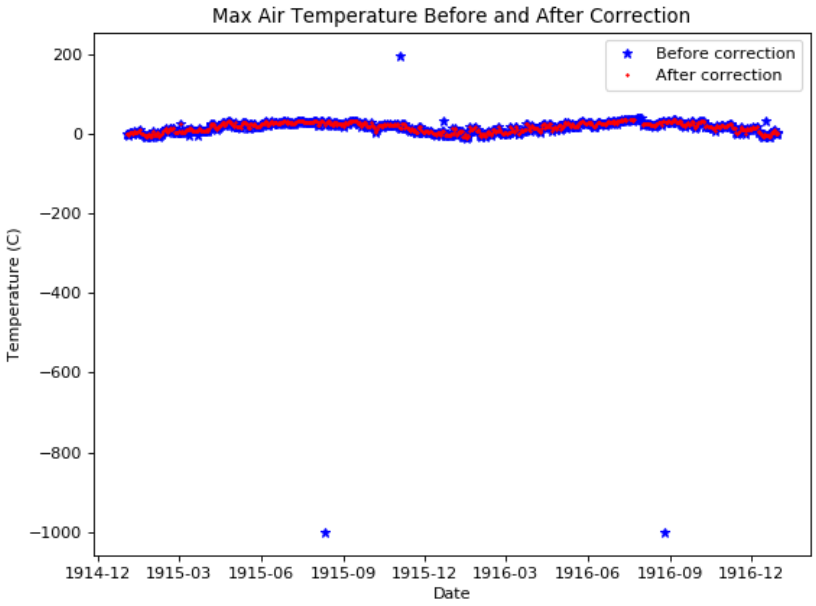

Figure 1: Plot of max air temperature data before and after error check

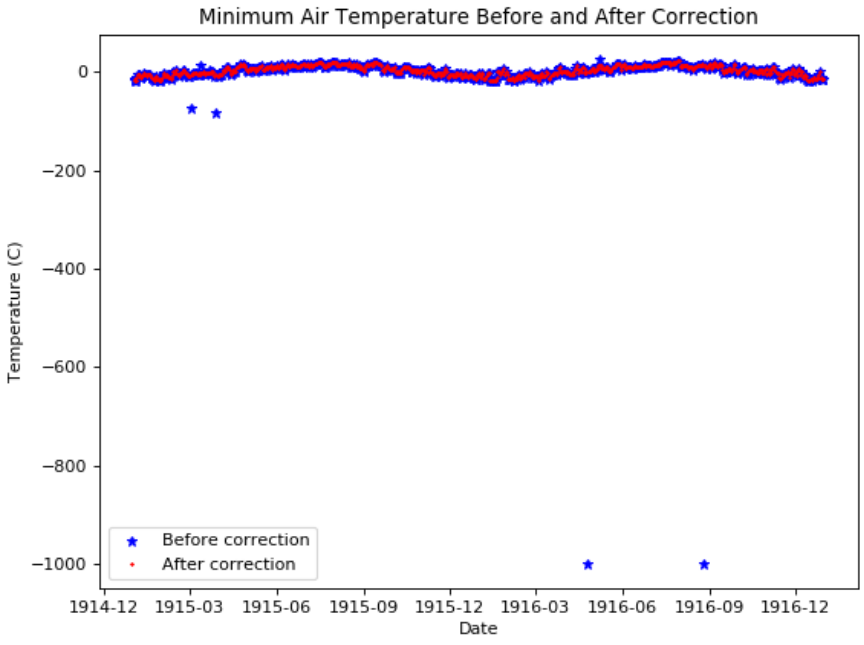

Figure 2: Plot of minimum air temperature data before and after error check

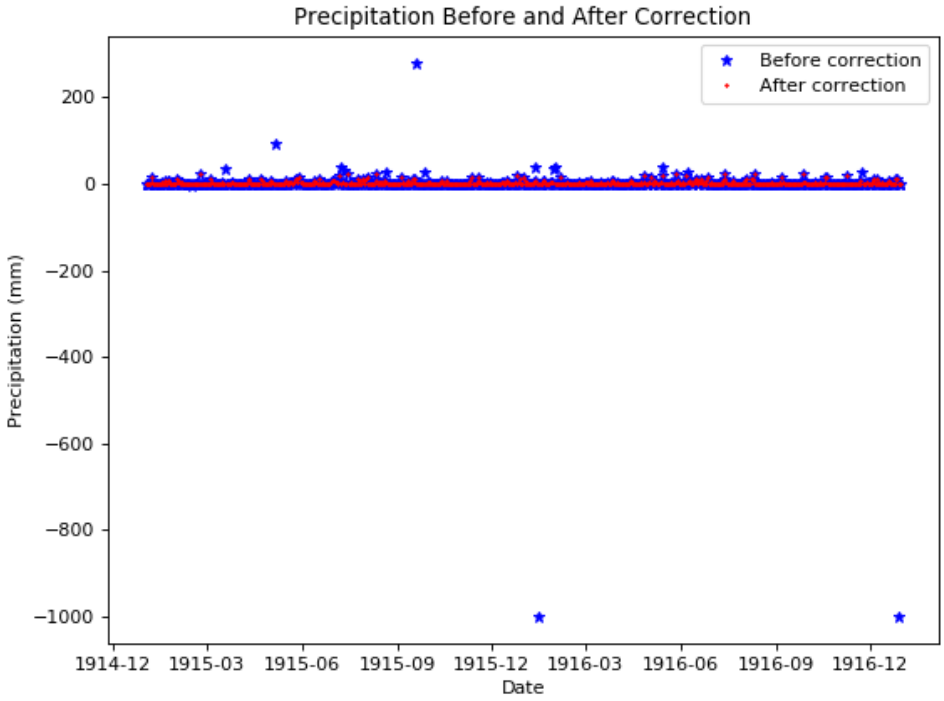

Figure 3: Plot of precipitation data before and after error check

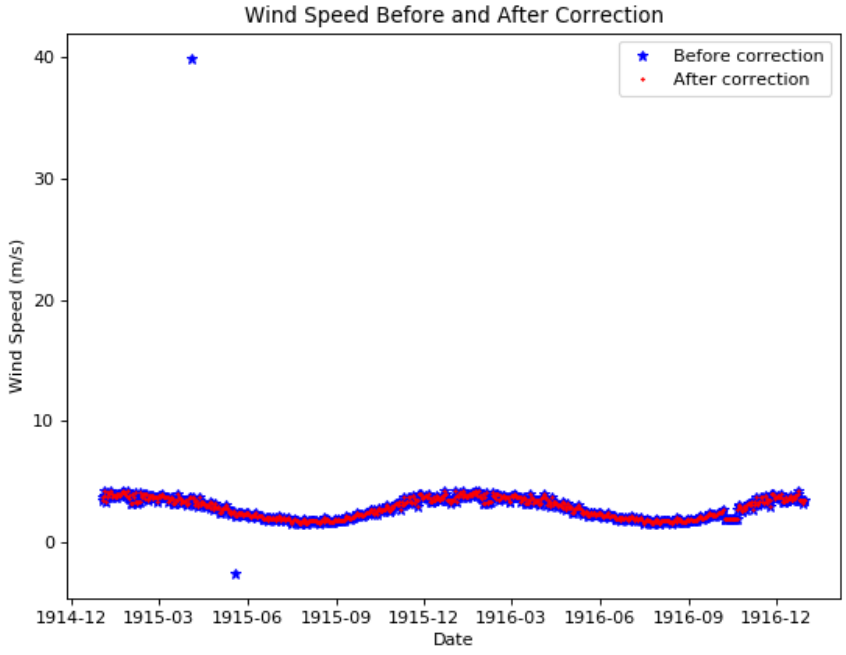

Figure 4: Plot of wind speed data before and after error check