Data Checking

Creator: Tianle Xu

1. Description of the program:

This code is used to check the quality of a certain data, which contains daily climate data from a single site. Columns in order is date, precipitation (mm), minimum air temperature (Celsius), maximum air temperature (Celsius) and wind speed (m/s).

All values of -999 will be replaced by NumPy NaN values and the number of replaced values will be count.

Then, the the error thresholds: $0 \leq P \leq 25$, $-25 \leq T \leq 35$, $0 \leq WS \leq 10$ will be applied to the data and the data outside this range will be replaced by NaN, and the number of replaced values will be count as well.

In addition, check whether all values in the ‘max temp’ are greater than those in the ‘min temp’ column. If not, the values will be swapped, and the number of values replaced of each data type will be record.

Finally, identify days with temperature range greater than 25 Celsius. These values bigger than 25 Celsius will be replaced by NaN and record the number of replaced values.

2. Description the performance and influence of the data quality checking

From the table as follow, precipitation, minimum temperature and maximum temperature all have two none values while wind speed data doesn’t have none value data.

15 values of precipitation data, 14 of minimum temperature, 2 of maximum temperature and 2 of wind speed outside of the range are replaced, respectively.

Four values in minimum temperature and maximum temperature are swapped.

There are 5 days with temperature range greater than 25 Celsius.

3.

Table 1. Results of Data Checking

	Precip	Min Temp	Max 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

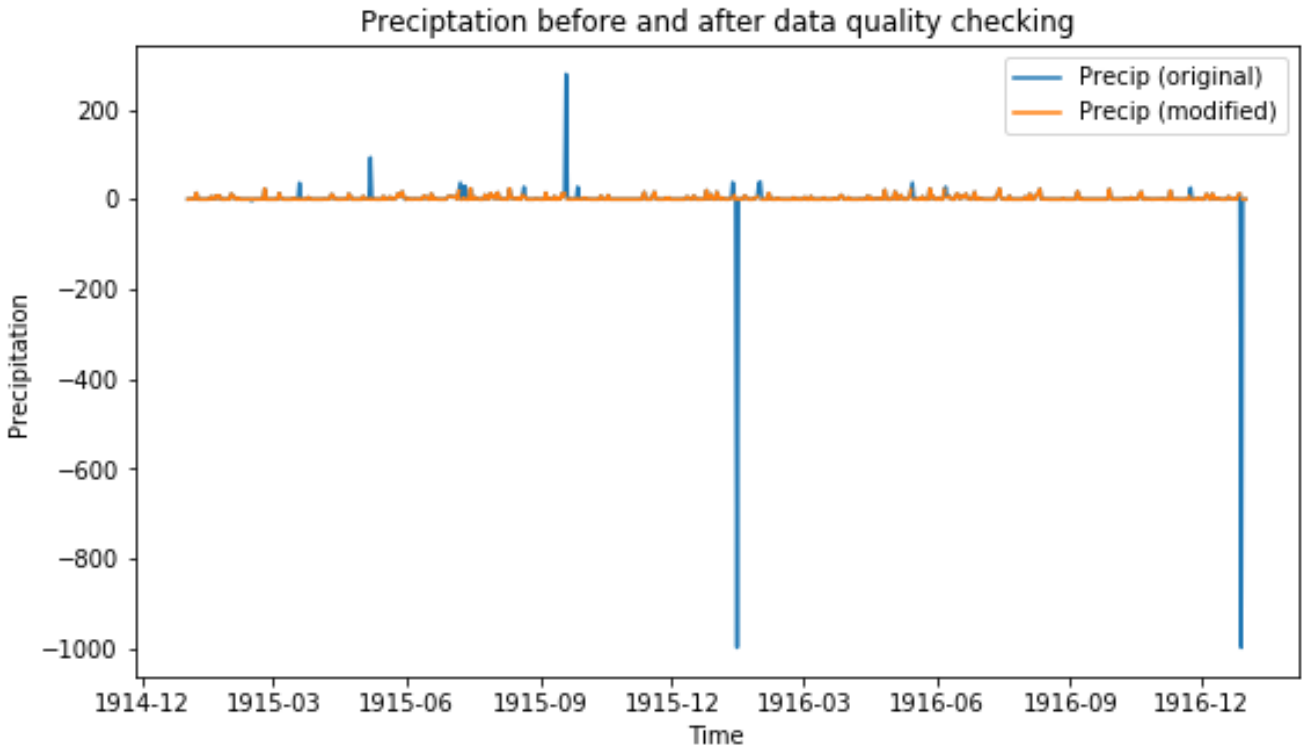

Fig. 1. Comparison of Original and Modified Precipitation Data

From the plot, it is obvious that in the original data, almost all data are the same as that after checking while just several data are out of reasonable range of precipitation value and even two are below zero, which is impossible. After data quality checking, all unreasonable values are replaced by Nan value. The modified data would be good material for people to do relative study.

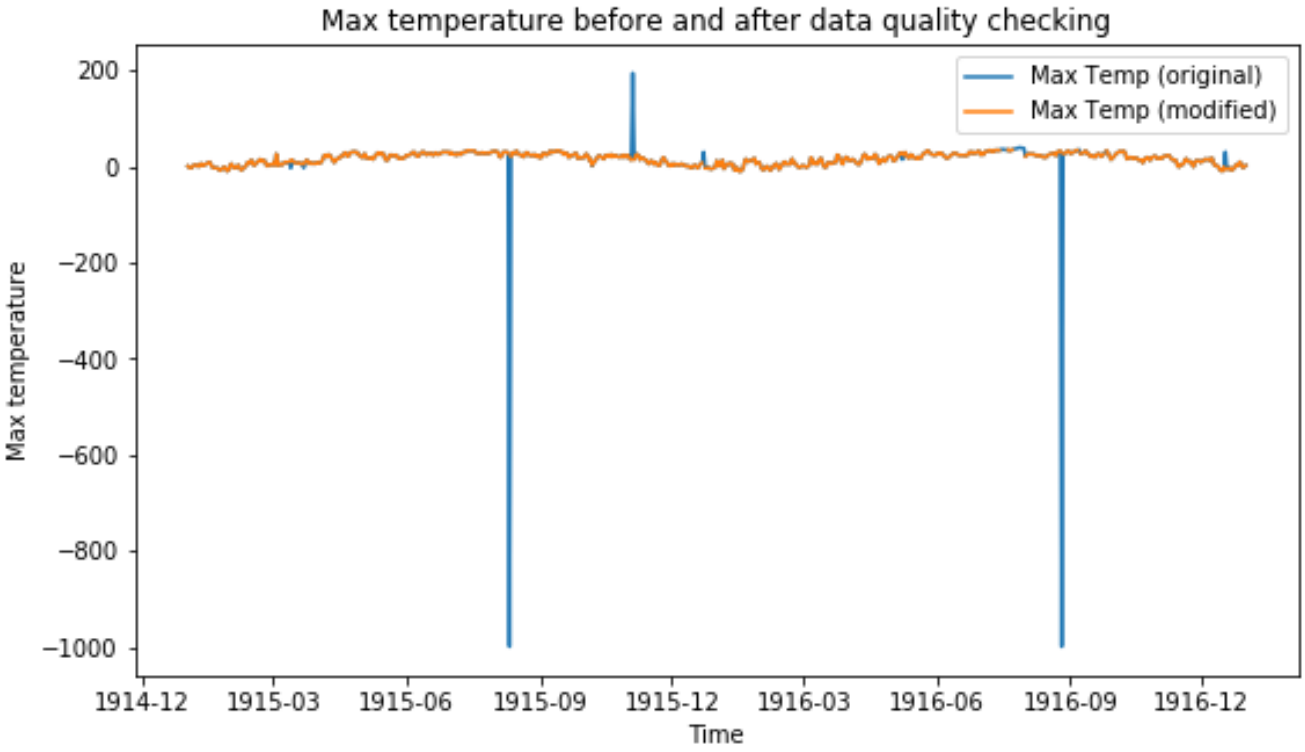

Fig. 2. Comparison of Original and Modified Max Temperature Data

The plot above illustrates that the most original data is the same as the modified data. Several of them are outside of the reasonable temperature range (-25~35 Celsius) in this certain place. Combined with the following plot, which shows minimum temperature data, some minimum values are bigger than maximum values as shown in above figure, so these values are swapped. After checking and replacement, the final data are good enough for future use.

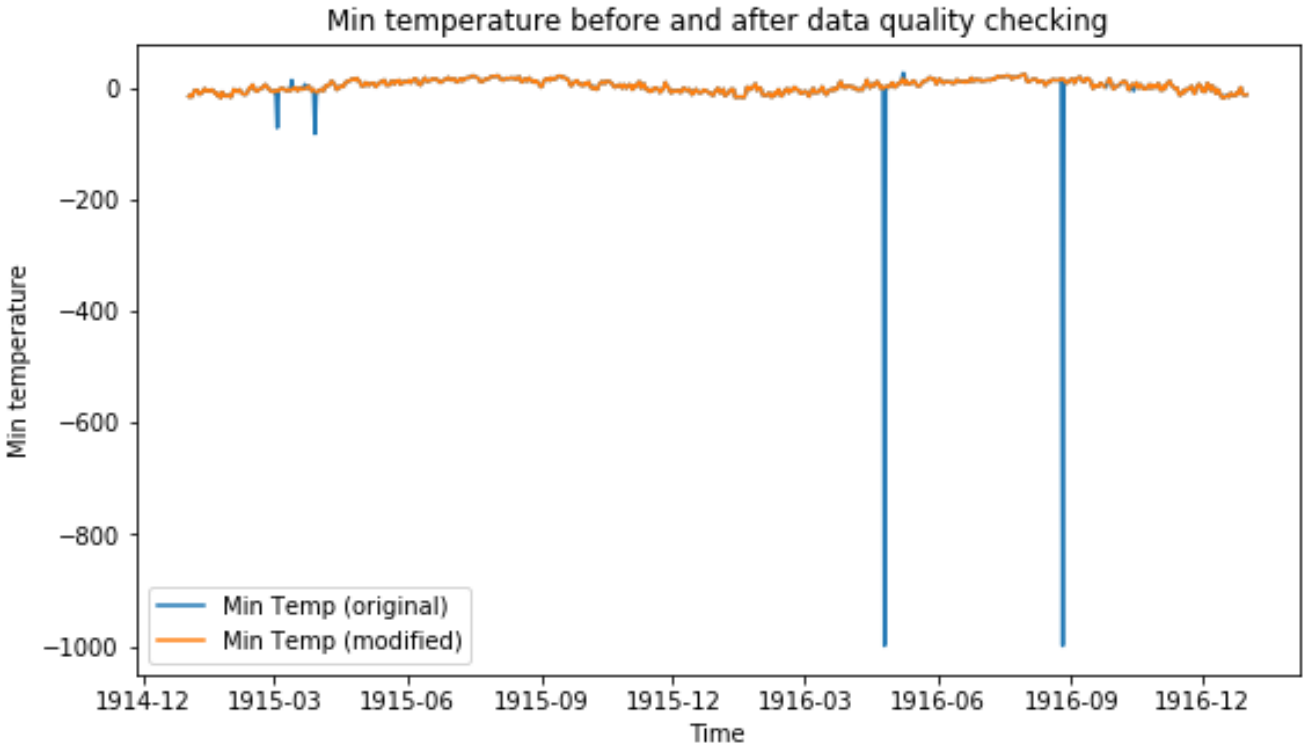

Fig. 3. Comparison of Original and Modified Min Temperature Data

This plot shows minimum temperature data. As mentioned in the caption of maximum temperature's plot, the data outside of range are replaced and several unreasonable values are swapped with those in maximum temperature data. The smooth plot of minimum data shows the great quality of the data.

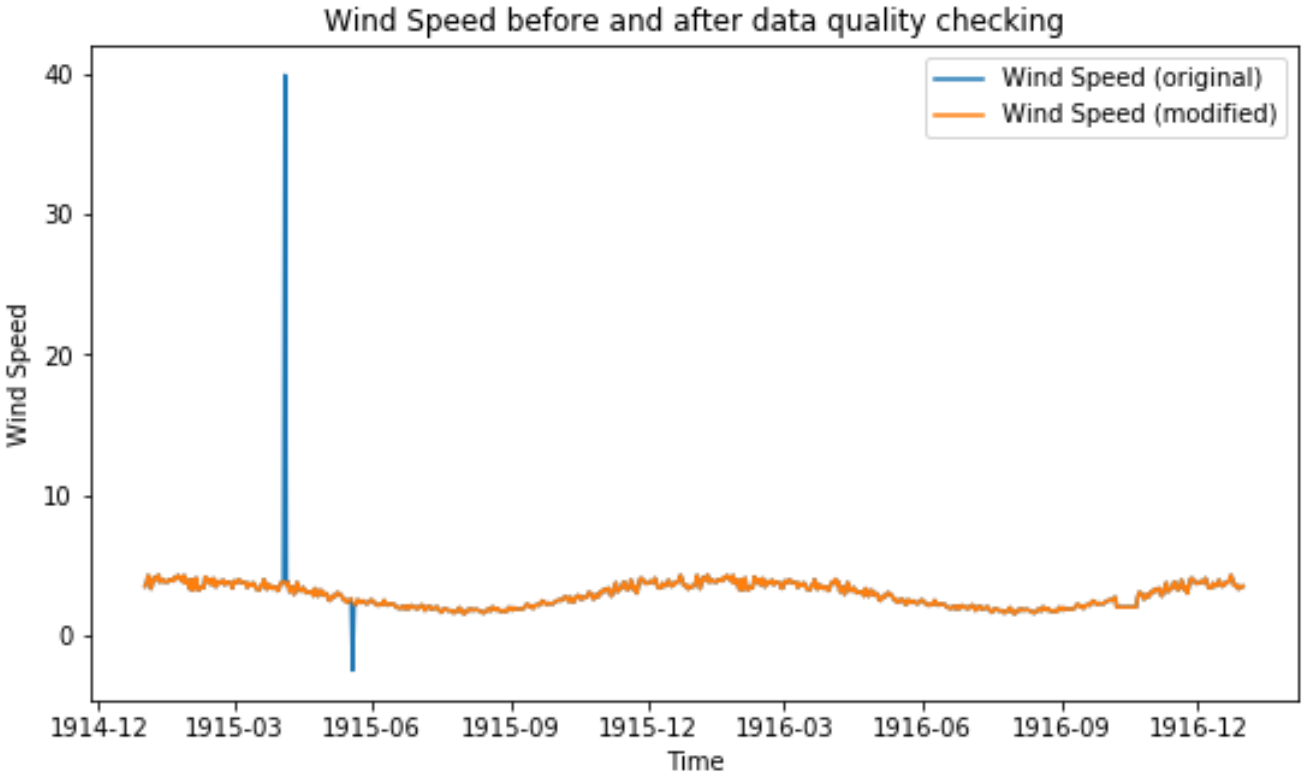

Fig. 4. Comparison of Original and Modified Wind Speed Data

By the plot above, the main problem is from the extreme values as most of data are fit well with the modified data. After data quality checking work, all the values unreasonable are replaced by none values. Thus, the final data could be used to future analysis.