Metadata for program_09.py

Author: Pin-Ching Li (li3106)

This metadata is written for the usage of program: program_09.py. program_09.py is a script for data quality checking. Datasets are checked by four quality checking tests: missing value, gross error, misplaced temperature data, and range problem of temperature data. These four scenarios would be fixed with different methods.

After the script run, there are two files and four plots created to record the result of data quality checking. Failedcheck.txt records how many data points have quality problem in the four scenarios mentioned above. Data_afterchecking.txt contains the clean datasets which gets rid of the problematic data points.

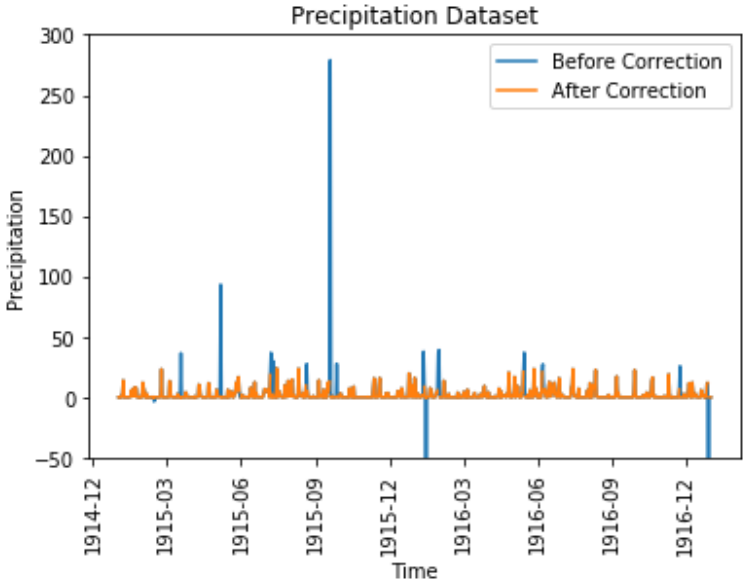
The table with number of corrections is shown below:

	Precip	Max Temp	Min Temp	Wind Speed
1. No Data	2.0	2.0	2.0	0.0
2. Gross Error	15.0	14.0	2.0	2.0
3. Swapped	0.0	4.0	4.0	0.0
4. Range Fail	0.0	5.0	5.0	0.0

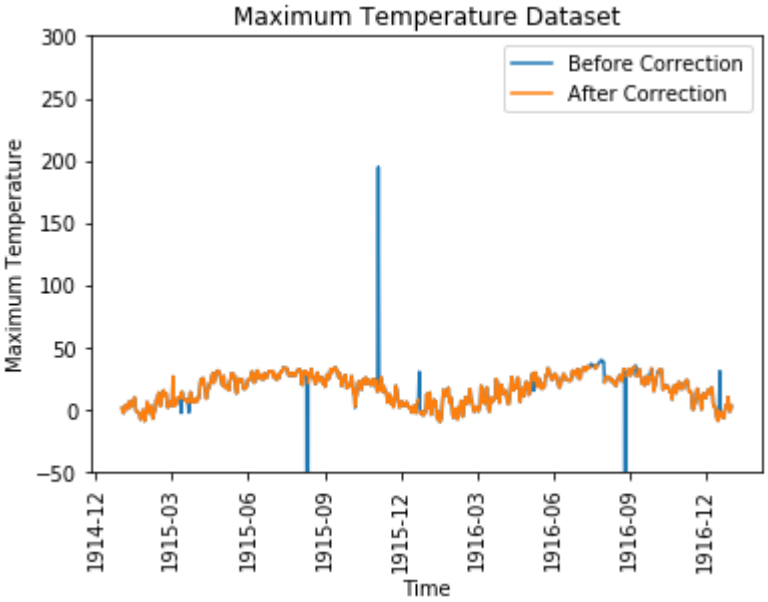
There are small amounts of missing data. Two missing values for each dataset except for Wind Speed dataset. The gross error seems to have a huge impact on Precipitation and Max Temperature datasets. It's hard to tell most of the gross error can be true or not for these two datasets. Some of them are very close to the criterion we set. For example, there are two obviously gross error in Precipitation dataset, where the Precip almost equals to 275 and 100. However, the rest 13 gross errors can possibly be taken as outliers in dataset (large but reasonable). The values of these errors are close to the accepted value. Therefore, the criterion of gross error can hugely influence one tail of the dataset distribution. At the other tail of dataset distribution, the minimum value seems to suffer less influence from the gross error. It is obvious that the Wind Speed dataset has the two gross errors. The other data points in this dataset sit within the criterion we gave.

By swapping misplaced value and applying range fail test, we can make dataset more consistent. The maximum value would always be larger than minimum value. Magnitudes of difference between max and min temperature are all selected within a reasonable range.

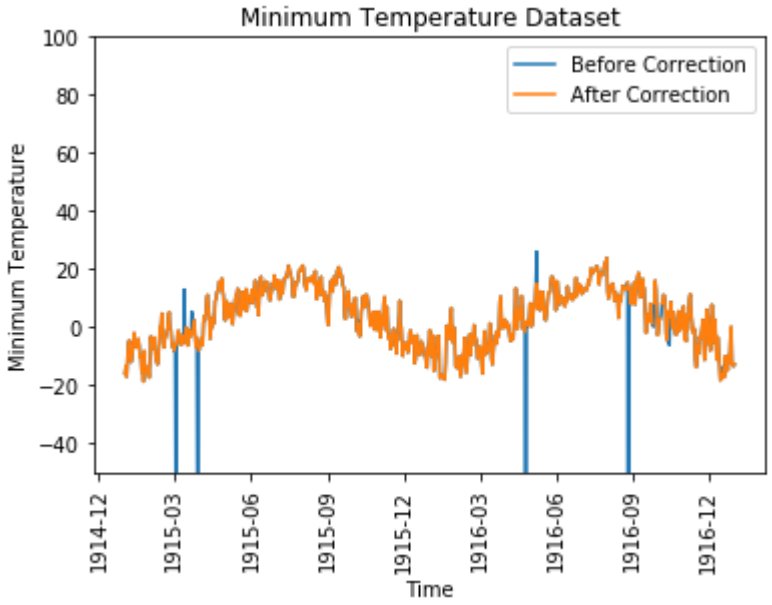
Plot of Precipitation dataset before and after correction is shown below. Two missing data points (-999) are removed and several gross errors are removed. It is obvious that the dataset after correction gets rid of the unreasonable spikes. However, some removed data points can still be argued if they are gross error or not. We need further information to delete them, such as the flags of datasets.



Plot of Maximum Temperature dataset before and after correction is shown below. The unreasonable spikes and missing data of dataset are removed. The dataset after correction seems to show a much smoother and periodic curve.



Plot of Minimum Temperature dataset before and after correction is shown below. Minimum Temperature dataset after correction is smoother than before. The missing value and some huge values are removed with respect to gross error and unreasonably huge difference between Maximum Temperature dataset and itself.



Wind Speed dataset contains less error than previous datasets. There are two gross errors. No missing data is within the wind speed dataset. This dataset is a relatively consistent in recording data from field.