Metadata file for assignment 09

1. Program information

Name: program_09.py

Created by: Gargeya Vunnava (GitHub: gargeyavunnava, Purdue username: vvunnava)

Short program description:

The program reads a space delimited text file ('DataQualityChecking_postchecking.txt') that contains date-wise meteorological data on precipitation (mm), min temperature (°C) , max temperature (°C) and wind speed (m/s). The program then processes the data to perform data quality checking and removes different types of errors as discussed in the next section. The program records the no of changes performed and stores the changes in a text file called 'Replaced_values_info.txt' and stores all the process data in a file called 'DataQualityChecking_postchecking'.

2. Description of the data quality checks performed

The program performs 4 types of data quality checks:

I. No data values: The program checks for the number -999 in the file and replaces it with a numpy NaN character. -999 is usually used for missing data are data that cannot be reported. Since it can influence a math operation performed on the data, they have to be converted into NaN charecters in prevent them messing with math operations performed. This check found 2 errors each in precipitation, max and min columns and fixed them by replacing -999 by NaN characters. (refer row 1 in table 1 and fig 1)

II. Gross errors: Each variable can have limits beyond which the data cannot exists. If the raw data has any values beyond these limits, they have to be removed and replaced with NaN characters. The program uses the following thresholds: $0 \leq$ precipitation ≤ 25; $-25 \leq$ min/max temperature ≤ 35, $0 \leq$ wind speed ≤ 10. The program found 15 gross errors in precipitation data, 14 errors in Max temp data and 2 each in Min temp and wind speed data. The program replaced all these gross errors with NaN characters. (refer row 2 in table 1 and fig 2)

III. Swapping: If the Max temperature value recorded for a date is lower than the min temperature value recorded for the same date (or vice versa), they need to be swapped to retain the meaning of min and max temperatures recorded. The program found 4 dates where swapping had to be done for min and max temperatures. (refer row 3 in table 1 and fig 3)

IV. Temperature range exceedance: If the difference between Max and Min temperature recorded for a date is greater than 25 °C, then both the Max and Min values are replaced with NaN characters. The program found 5 dates with this range exceedance problem and

fixed the data by assigning NaN characters to temperature data. (refer row 4 in table 1 and fig 4)

3. Table with the number of corrections made

	Precipitation	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	0.0	5.0	5.0	0.0

Table 1: Table with no of changes performed for each variable in the raw meteorological data file.

4. Plots for each variable before and after quality checking

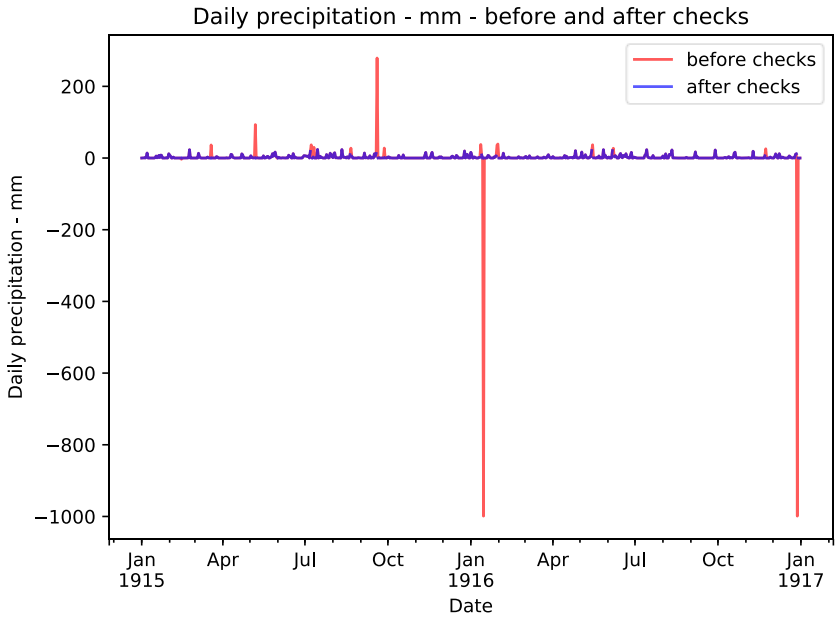

Fig 1: Daily precipitation values in mm before and after checks. 17 data points were modified: 2 No data errors and 15 gross errors.

The extremely deviated negative values shown by the red lines in fig 1 are the no data errors with -999 values. The large positive values shown by red line are due to gross errors which are not present in data shown by the blue line.

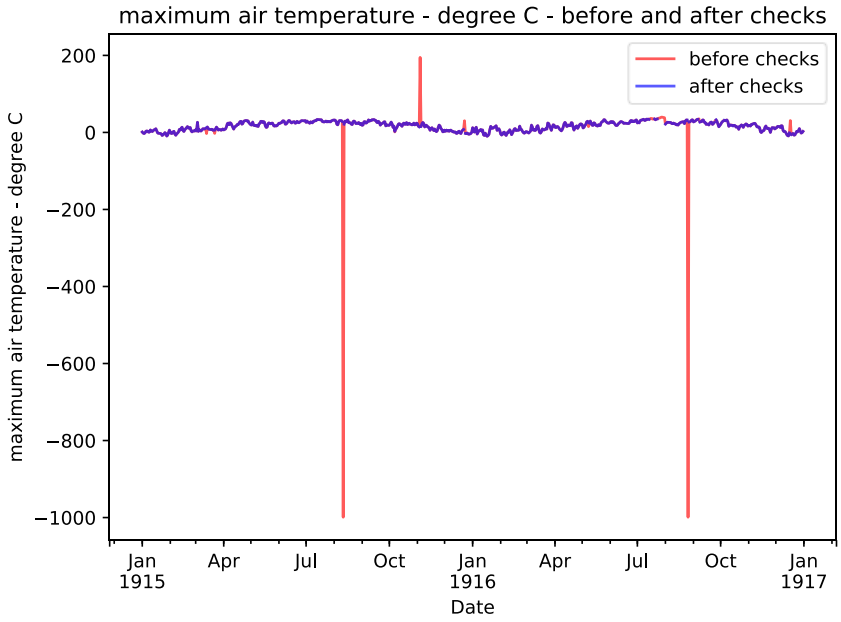

Fig 2: Daily maximum air temperature values in °C before and after checks. 14 data points were modified: 2 No data errors, 4 swapped errors and 8 range fail errors.

The extremely deviated negative values shown by the red lines in fig 2 are the no data errors with -999 values. Other positive and negative deviations shown by red line are due to gross errors which are not present in data shown by the blue line representing the corrected data.

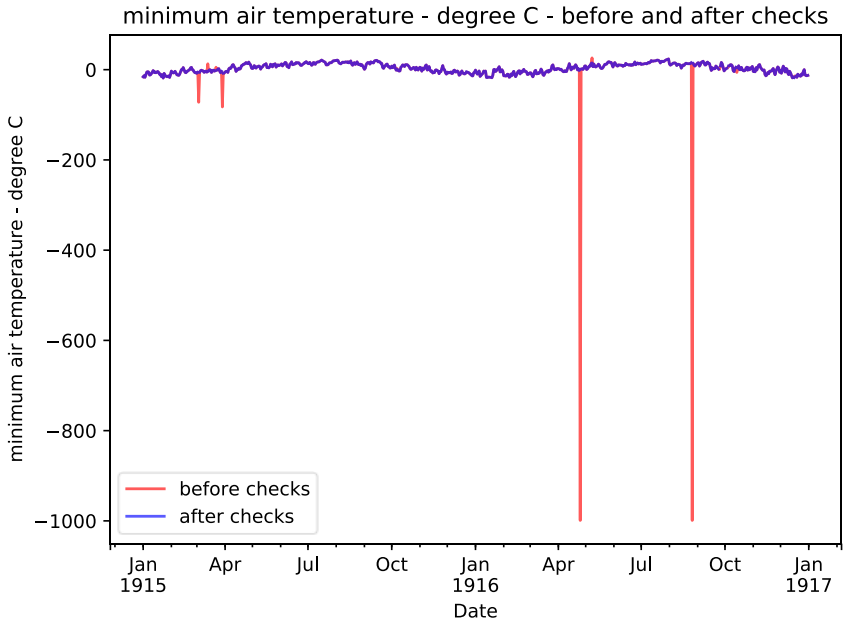

Fig 3: Daily minimum air temperature values in °C before and after checks. 14 data points were modified: 2 No data errors, 4 swapped errors and 8 range fail errors.

The extremely deviated negative values shown by the red lines in fig 3 are the no data errors with -999 values. The positive and negative deviations shown by red line are due to gross errors which are not present in data shown by the blue line. When compared with fig 2, the red line in fig 3 has significant negative gross errors which makes sense as it represents daily min temperatures while fig2 represents daily max temperatures.

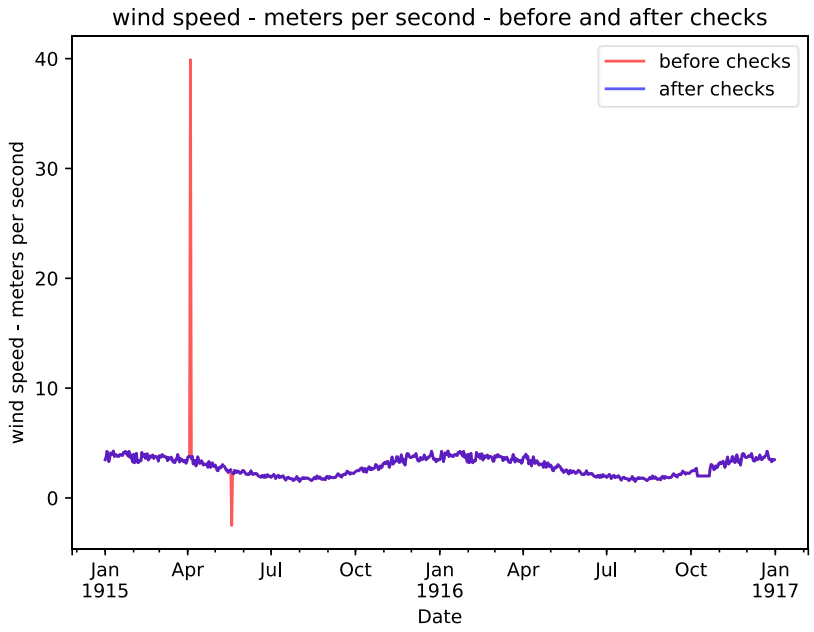

Fig 4: Daily wind speed in m/s before and after checks. 2 data points related to gross errors were modified.

The daily wind speed data had the least modifications performed. There were just 2 gross errors as shown by the big red spikes in figure 4.
