Metadata

Lab09: Data Quality Check

Author: Alka Tiwari (tiwari13)

Github: https://github.com/Environmental-Informatics/09-data-quality-checking-roccabye

Script Name: program_09.py

Original data file: DataQualityChecking.txt

Program Description: This script uses daily climate data for a single site and checks the dataset for gross error, inconsistencies and range problems for the variables (Precipitation (mm), Maximum and Minimum Air Temperature(°C) and Wind Speed (m/s). It also creates plot for all the variables before and after data quality check.

Data Quality Checks performed

- Check 1: Remove No Data values.
  - Replace all values of -999 in this file with NaN values
    - Table 1 shows the number of data points that has been replaced with the index "1. No Data"; two data points of each variable precipitation, maximum and minimum air temperature has been replaced from negative NaN value. There has been no -999 value for Wind Speed.
- Check 2: Check for gross errors
  - Apply the following error thresholds: $0 \leq P \leq 25$mm; -25ºC≤ $T \leq 35$ºC, $0 \leq WS \leq 10$ m/s.
  - Replace values outside this range with NaN.
    - Table 1 shows the number of data points that has been replaced with the index "2. Gross Error"; 15 precipitation values, 14 maximum air temperature values, 2 minimum air temperature and 2 wind speed values have been replaced to NaN value, as they are not falling into the given threshold/range.
- Check 3: Swap Max Temp and Min Temp when Max Temp is less than Min Temp.
  - Check that all values of Max Temp are greater than for Min Temp for the current day's observations.
  - Where they are not, swap the values.
  - Table 1 shows the number of data points that has been replaced with the index "3. Swapped"; four values of maximum air temperature have been less than minimum air temperature, so swapped with each other
- Check 4: Check for daily temperature range exceedance.

- Identify days with temperature range (Max Temp minus Min Temp) greater than 25°C.
- When range is exceeded replace both Tmax and Tmin with NaN.
- Five values of maximum and minimum air temperature have failed the criteria, as shown in Table 1 with the index "4. Range"

	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	0	5	5	0

Table 1: Number of corrections for all three checks.

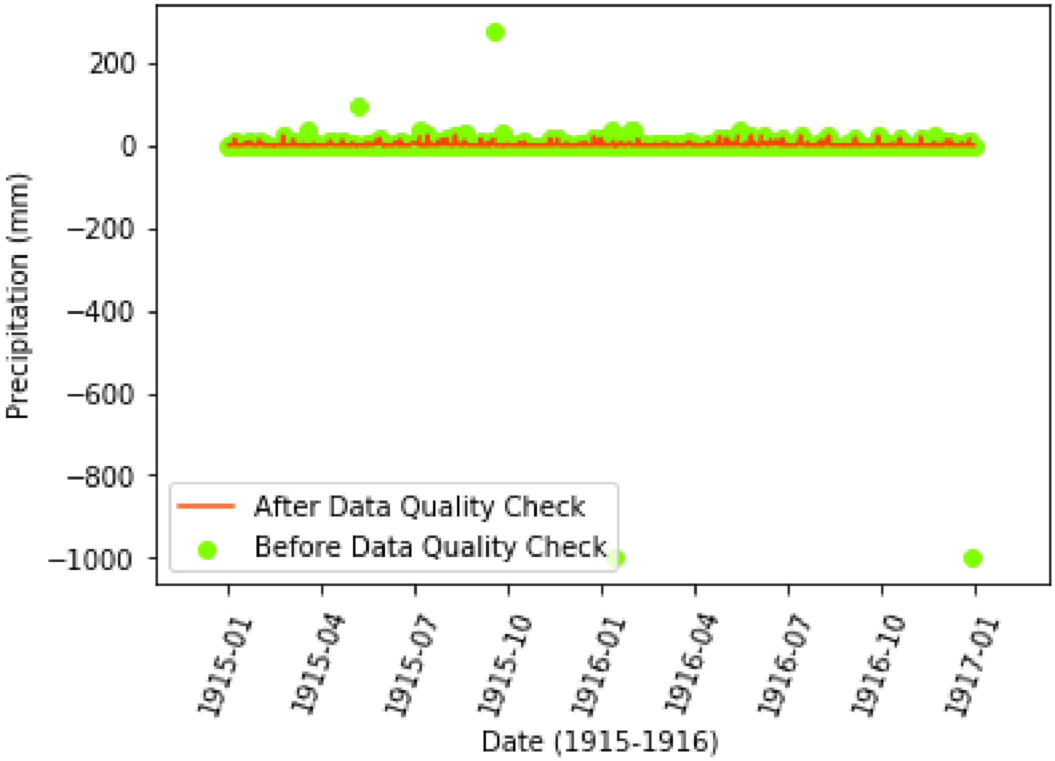

Figure 1: Precipitation data for the site before and after data quality check. For the time series two data points has no value and 15 data points have gross error i.e. values are greater than 25 mm, which has been replaced with no value after data quality check.

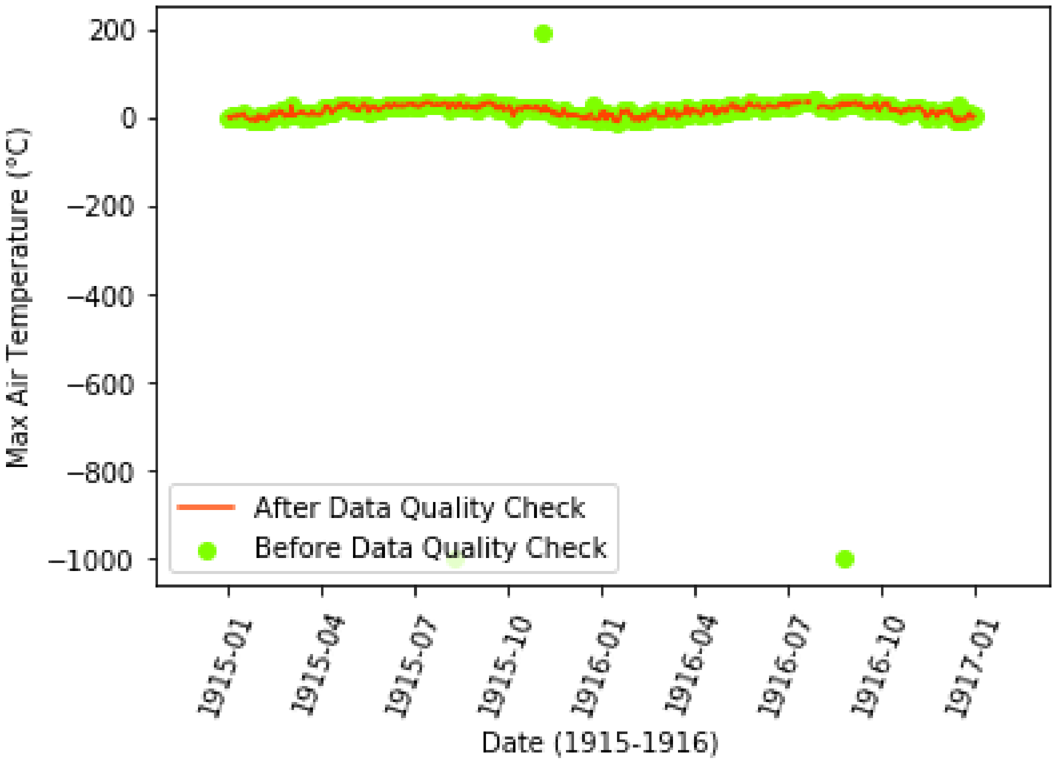

Figure 2: Maximum Air Temperature data for the site before and after data quality check. For the time series two data points has no value and 14 data points have gross error i.e. values are greater than 35°C and less than -25°C, which has been replaced with no value after data quality check. Also, four data points have values greater than consecutive maximum air temperature, hence swapped in the dataset after data quality check, and five data points fail the range criteria (temperature difference more than 25°C which has been replaced with no values after data quality check.

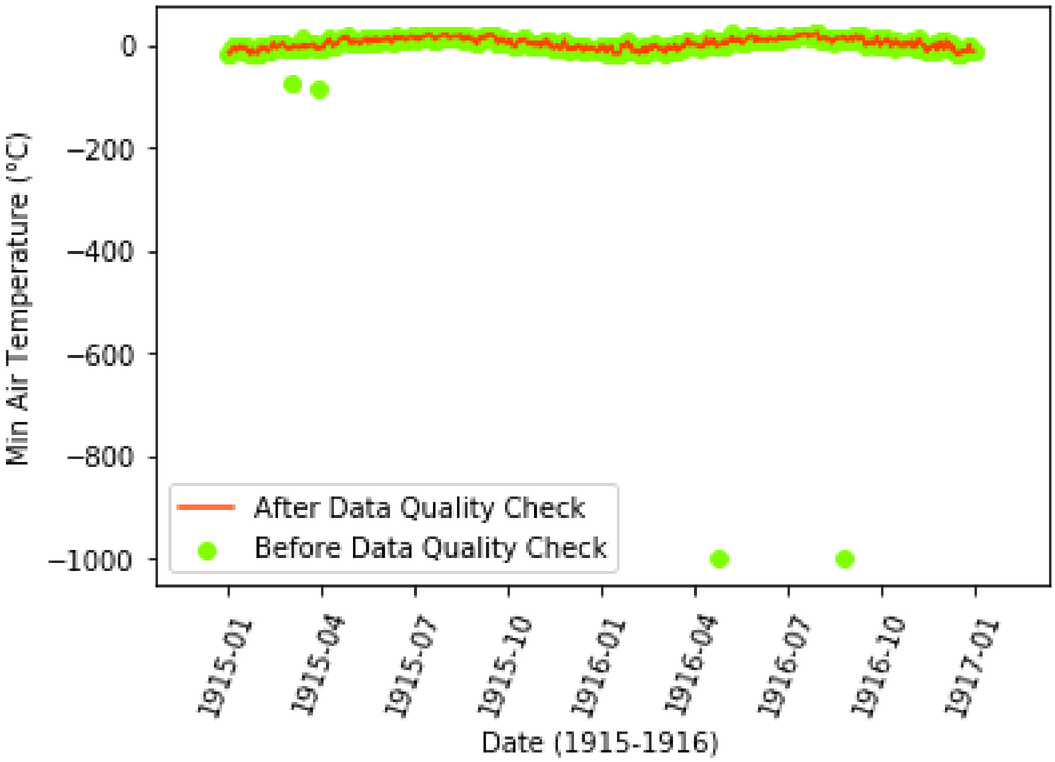

Figure 3: Minimum Air Temperature data for the site before and after data quality check. For the time series two data points has no value and 14 data points have gross error i.e. values are greater than 35ºC and less than -25ºC which has been replaced with no values in after data quality check dataset. Also, four data points have values greater than consecutive maximum air temperature, hence swapped in the dataset after data quality check, and five data points fail the range criteria (temperature difference more than 25ºC) which has been replaced with no values after data quality check.

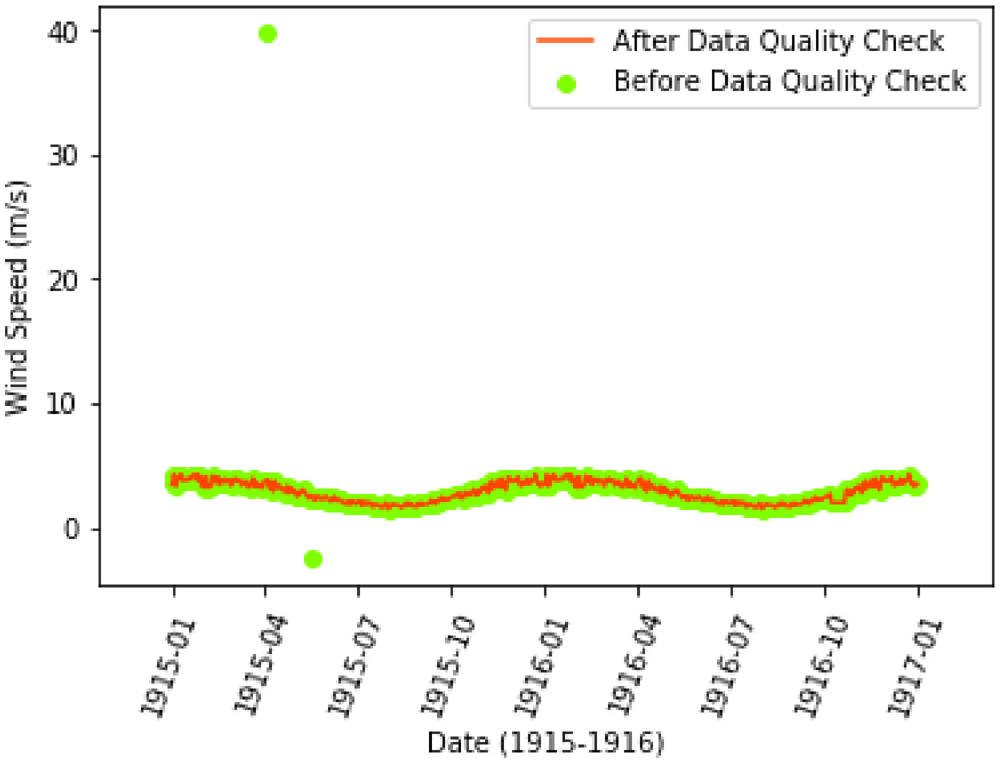

Figure 4: Wind Speed for the site before and after data quality check. For the time series two data points have gross error i.e. values are greater than less than 0 m/s, which has been replaced with no value.