Code: ABE 65100N | **Name:** Environmental Informatics | **CRN:** 15495
Name of Student: Jibin Joseph

Assignment 09
Data Quality Checking

Name of the Program: program_09.py
Name of Program Creator: Jibin Joseph (joseph57)
Description of the Program:
The program imports the text file as a dataframe and the date contained in the file is used as index. The program performs four checks on the data and a quality check is completed. Also, the count of checks are also recorded and stored as a csv file.

Input: DataQualityChecking.txt (time series data containing four variables from 1915-01-01 to 1916-12-31)
Output: four plots, quality checked data and failure checks in the form of csv file

Main Four Checks include:

1. Removes the No Data (-999) values and replaces it with NaN
2. Removes Gross Error in Precipitation, Maximum Temperature, Minimum Temperature and Wind Speed and replaces with with NaN
3. Swapping of Maximum and Minimum Temperature
4. Daily Temperature exceedance of more than 25 ^{0}C. Such data are replaced by NaN

These checks have improved the quality of the data. This is evident from statistics of raw and final data as shown below.

Raw data.....

	Precip	Max Temp	Min Temp	Wind Speed
count	731.000000	731.000000	731.000000	731.000000
mean	0.288098	14.167227	0.548413	2.904172
std	53.773216	54.738379	53.477046	1.597814
min	-999.000000	-999.000000	-999.000000	-2.500000
25%	0.000000	6.735000	-4.080000	2.045000
50%	0.000000	18.560000	3.610000	2.910000
75%	2.237500	26.195000	11.875000	3.600000
max	279.000000	194.800000	26.100000	39.900000

Fig 1.a

All processing finished.....

	Precip	Max Temp	Min Temp	Wind Speed
count	714.000000	710.000000	722.000000	729.000000
mean	2.070588	16.296930	3.515904	2.860837
std	4.291815	11.267967	9.852835	0.798721
min	0.000000	-10.080000	-18.630000	1.500000
25%	0.000000	6.670000	-3.922500	2.050000
50%	0.000000	18.185000	3.680000	2.910000
75%	1.950000	25.822500	11.915000	3.600000
max	24.050000	34.960000	23.900000	4.280000

Fig 1.b

Fig 1. Descriptive Statistics of Raw Data and Final data

The mean of precipitation and minimum temperature has considerably changed after data quality checking. The quality checking removed 15 precipitation data and 14 maximum temperature points in Gross Error category. Also, there were 4 errors and 5 errors in swapped and range fail category. This has lead variation in descriptive statistics of final data (Fig 1.b) compared to raw data (Fig 1.a).

Table 1. Summary of Data Fail 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

After the checks, the program plot all the four variables with before and after correction values.

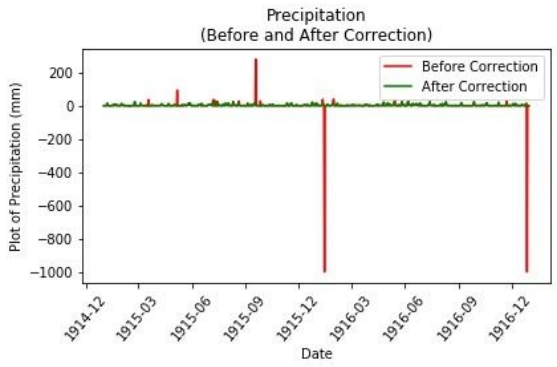

Fig 2.a Comparison of Precipitation

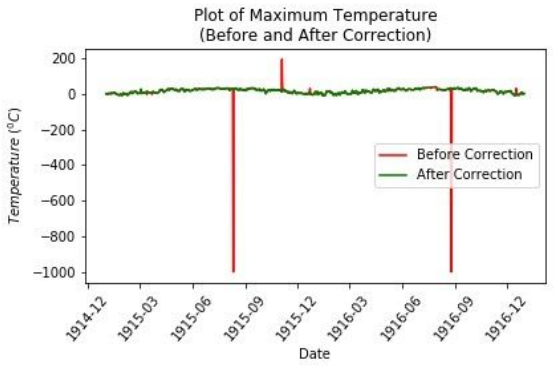

Fig 2.b Comparison of Max Temperature

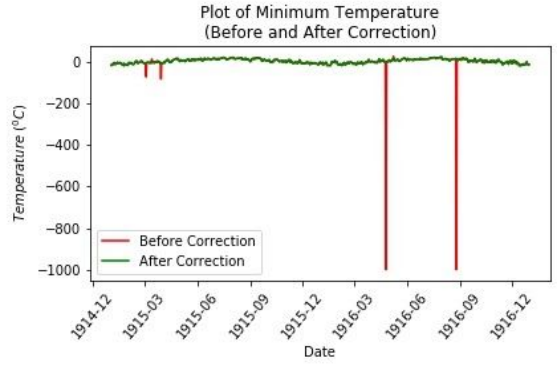

Fig 2.c Comparison of Min Temperature

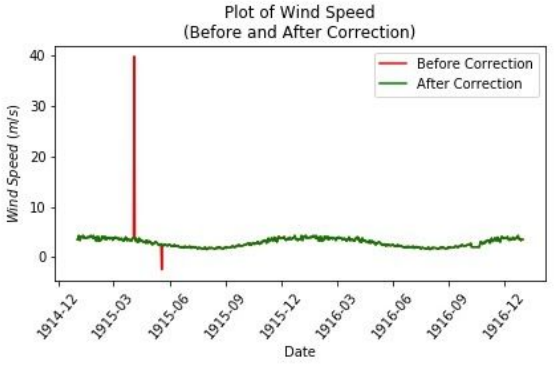

Fig 2.b Comparison of Wind Speed

Fig 2 Comparison of four variables (precipitation, maximum temperature, minimum temperature and wind speed) before and after data quality checking.