TEST TASK

# WEATHER DATA ANALYSIS

Script attached

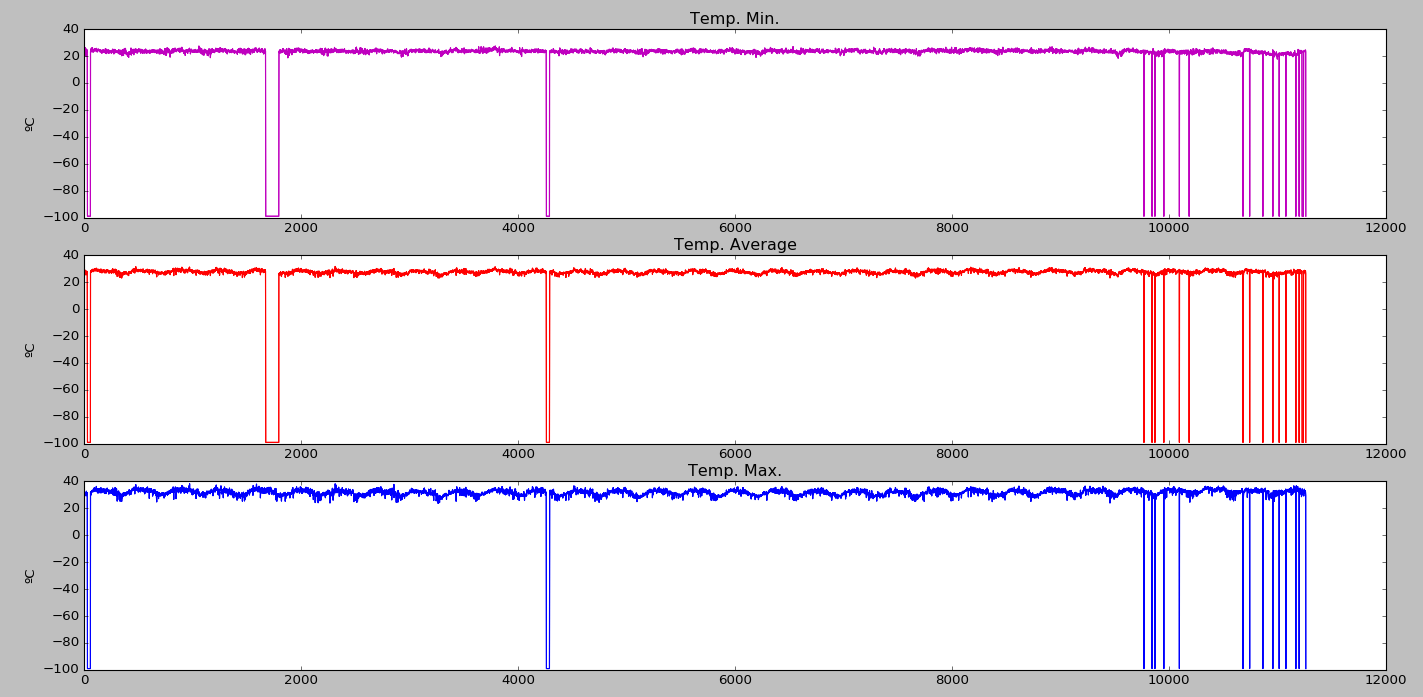
## 1st Question

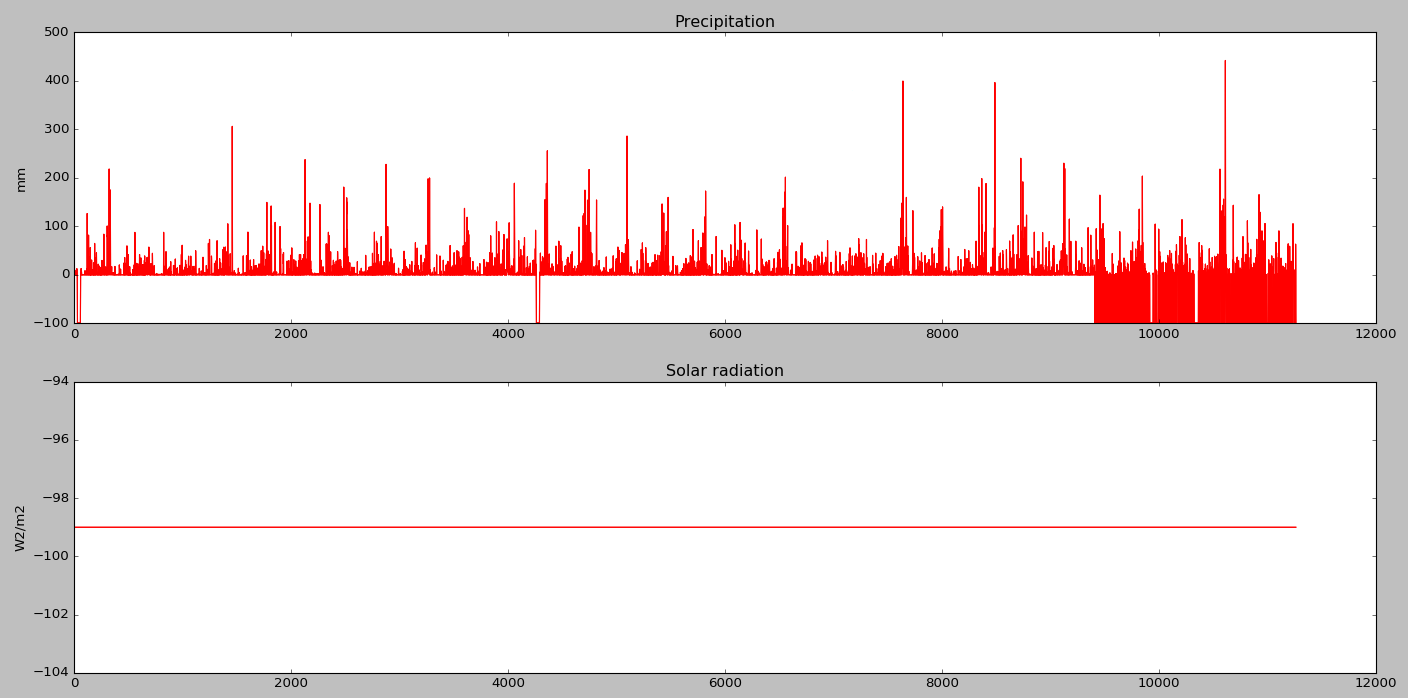
I converted the WTD to CSV

DATE column does not have date-time format so I could not figure out the time steps and the time range.

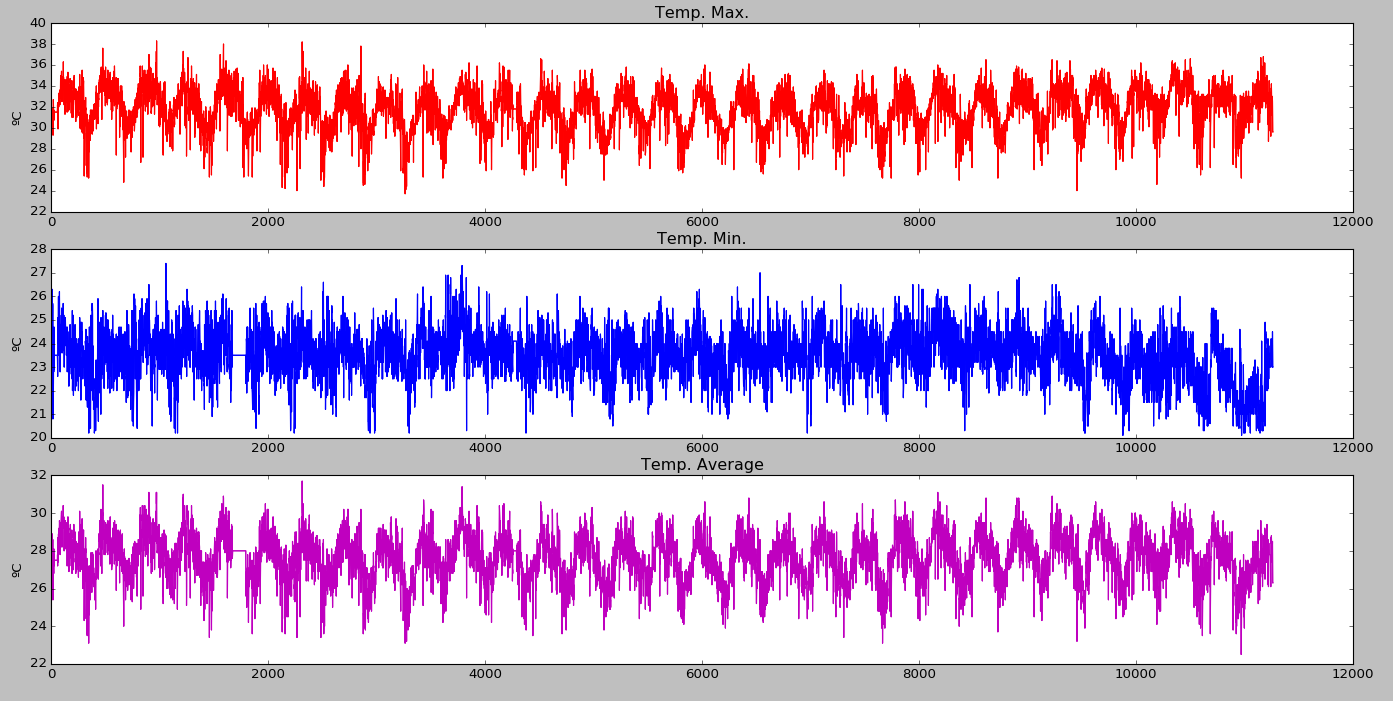
## 2nd Question

There are anomalies. The pattern is cyclical (31 cycles).





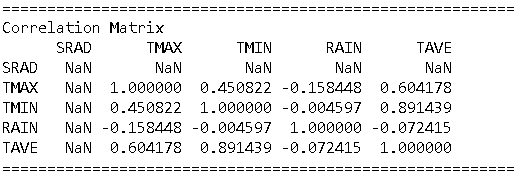
Raw Data Set



Temperature – after preparation

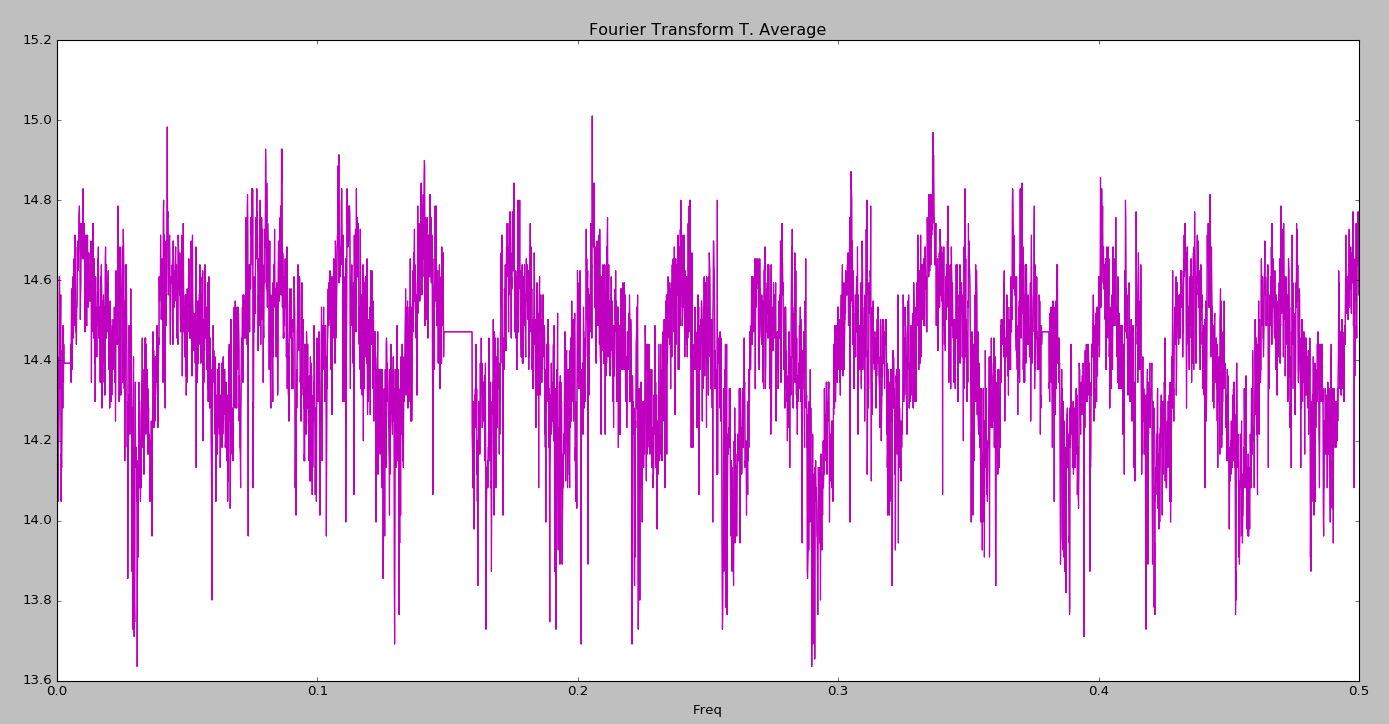
## 3rd Question

The correlation matrix has been produced after “cleaning” the data set.



## 4rd Question

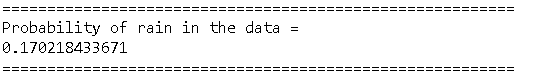
I could not figure out the time steps.



## 5th Question

I could not figure out the time steps.

The probability of precipitation (>10 mm) is:



## 6th Question

I could not figure out the time steps. I could not see the date & time format in source file (WTD file)

# DEEP LEARNING

## 1st Question

1. Selecting the best predictors using Receiver Operating Chart (ROC) (values close to 1)
2. Selecting tuning parameters Visually (Outliers)
3. Selecting tuning parameter programmatically. Algorithms: Genetic algorithm, Simplex Method or an hybrid method.
4. Re-sampling training data
5. Selecting the simplest model compatible with a good performance (good predictability)

## 2nd Question

Softmax Layer is the Exponential Function Normalized between (0,1).It is a linear function that classifies the data. It is normally used in the latest layer

ReLU Layer is a non-linear function that allows superposition. However at high values of X, the prediction is always either 0 or 1, limiting the predictability.

## 3rd Question

Advantages

* Easier to “clean” and evaluate
* Easier to take counter measures to avoid over fitting
* Faster

Disadvantages

* Potentially poorer predictability

## 4rd Question

One can expect a reasonable range of temperature, rainfall etc. Values outside this range are considered abnormal values. Please see the script attached. One can fill the gaps with interpolating functions (for simplicity I filled the gaps with an average value).

One can use the concept of return period (T). For instance, one can exclude values with T≥10 years. One can calculate this value if the data allows or can use external studies in the area where the data has been obtained.

# MISCELLANEOUS

## 1st and 2nd Question

I take n = number of sample points, and p = number of observations.

There is a Mean Square Error (MSE) per each observation (MSEi i=1…p) .This is irrespective whether n > p or n < p as :

One can select the model with the lowest MSE

The model might very well predict the training data and consequently produce a low MSEm, but there is a danger that the real MSE, when one compare with the real outcome, is not the minimum.

There techniques to select the minimum MSEreal:

* Validation set approach = Randomly splitting the data into training data and validating data
* LOOCV method
* K-fold cross validation
* etc

## 3rd and 4th Question

Precision Number (P)

Single precision floating number = 36 bits

Double precision floating number = 64 bits

Unit conversion

1 byte = 8 bits

1 MB = 106 bytes

1 GB = 109 bytes

Quantity of numbers = N ( for a matrix i,j, N=i\*j)

3. question= 2.7 GB

4 question = N = 2\*109 numbers = ixj