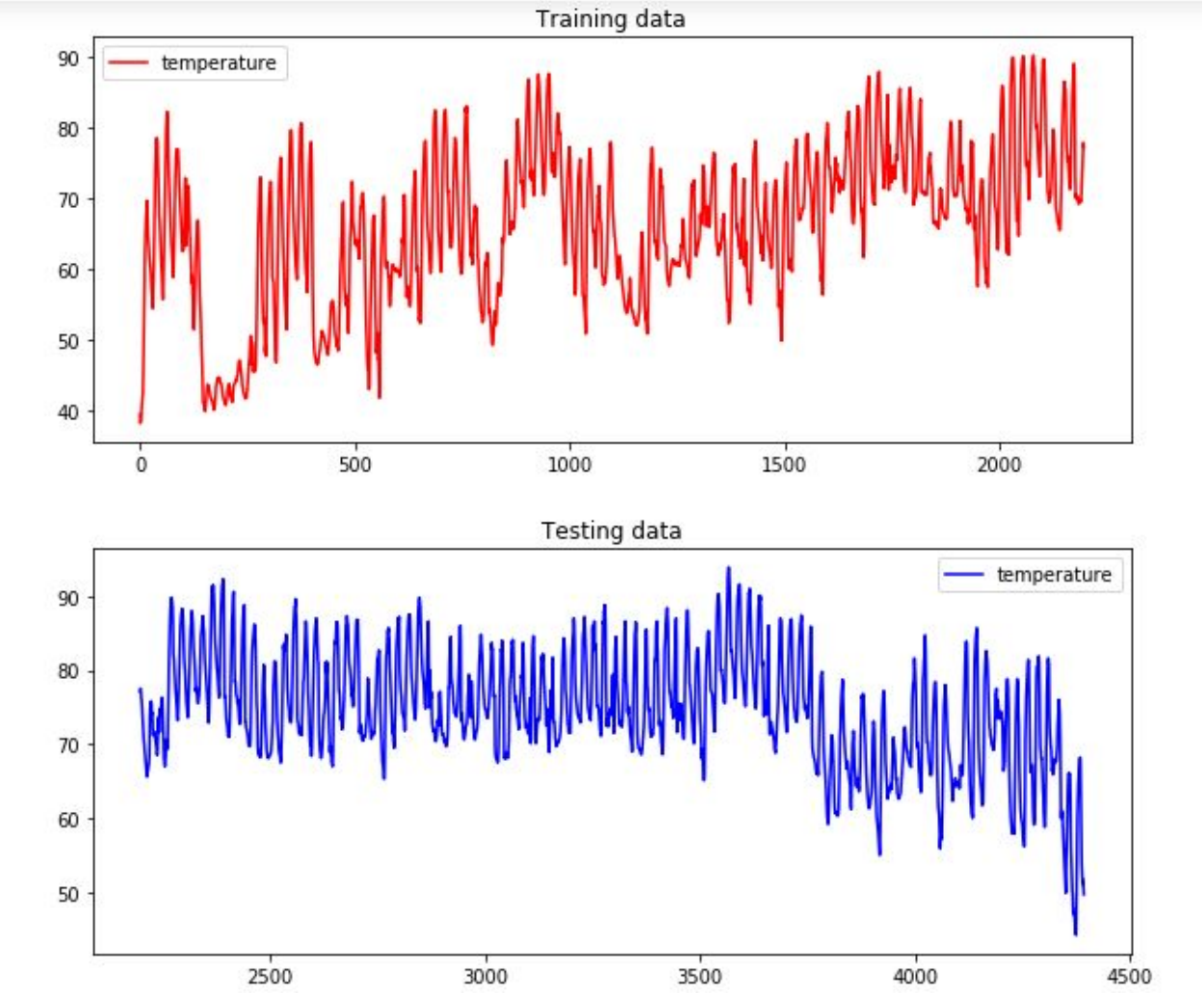
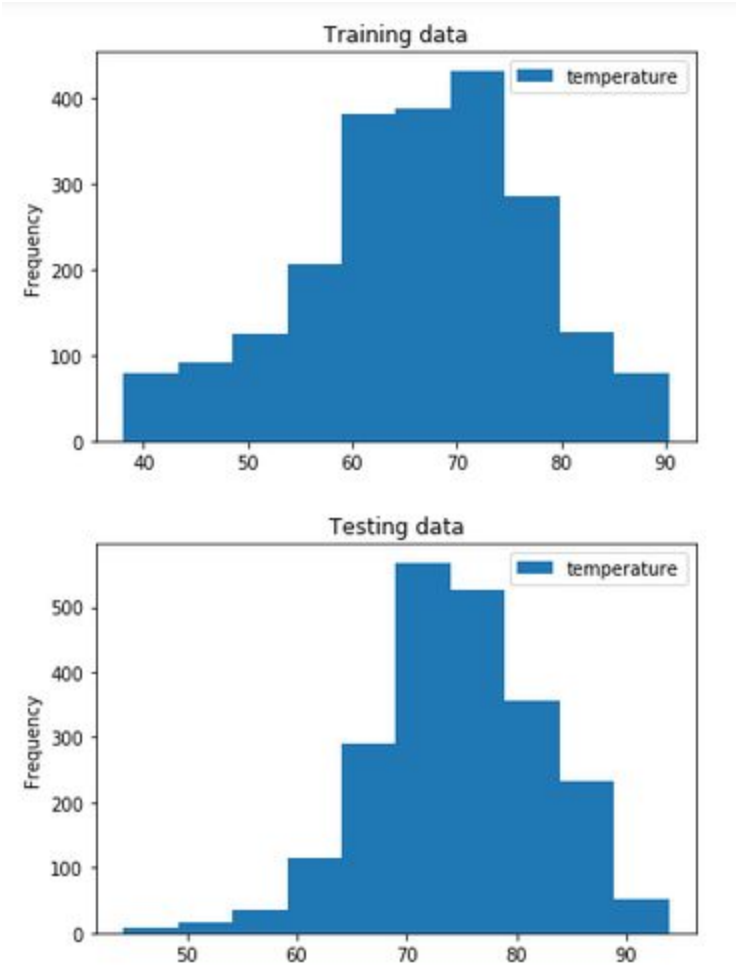
Results

Task1

- The training and testing datasets are both stationary.
- This is because the observations in a stationary time series are not dependent on time and do not have trend or seasonal effects as we see in the graphs below



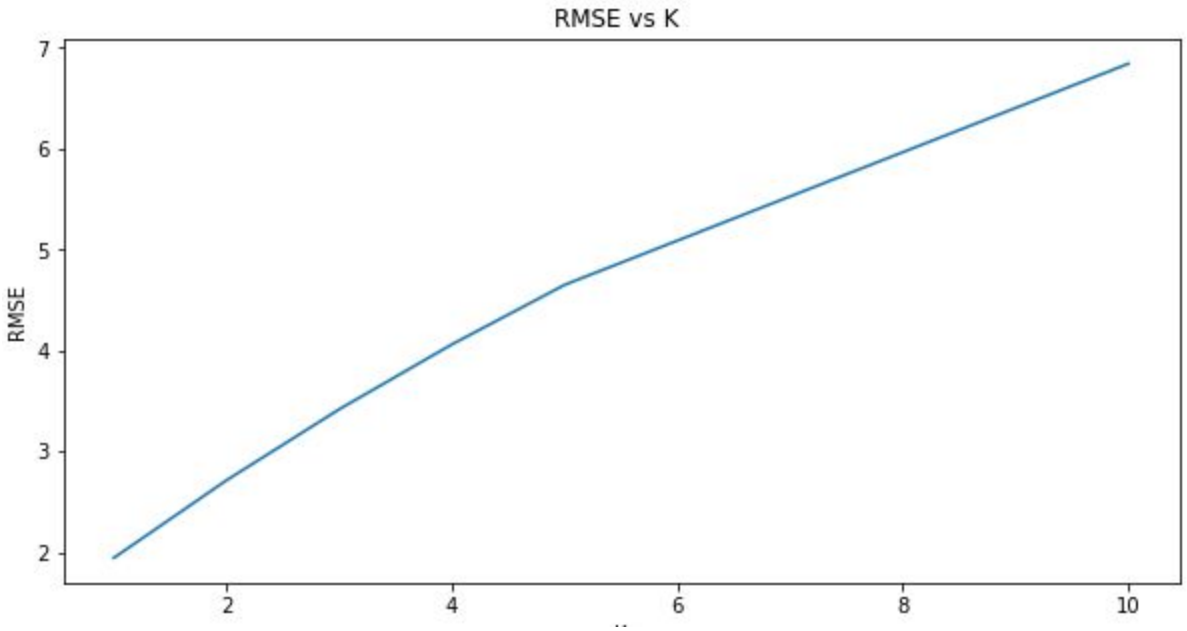
- The histogram looks Gaussian with a bit of a long tail which is skewed. This implies that we have a stationary dataset otherwise we'd see something much less 'bell-shaped' due to trending and/or seasonality (e.g., we'd see more data plotted to the left or right).



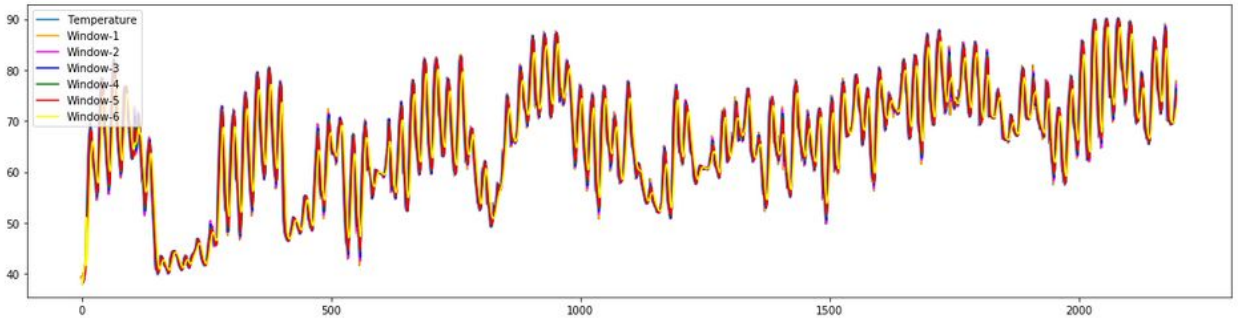
Task2

1. The rolling function in pandas is used to calculate the window and calculate the average. It has been modified to take in the previous k values for calculating the statistic.
2. RMSE values have been calculated for various values of k, and we see that there is an increase in RMSE values as k increases.
3. This is because, the data under consideration is temperature data and the results seem logical as there is less variation in temperature when the nearest neighbours are considered. Taking larger window size would lead to more variance being considered for forecasting and would hence lead to more error. In other words, it makes more sense to consider the past one or 2 hours of weather data which would not vary much over a period of 2 hours, than the past 12 hours.
4. We get k=1 to have the lowest RMSE value.
   (k=1) RMSE: 1.945
   (k=2) RMSE: 2.715
   (k=3) RMSE: 3.418

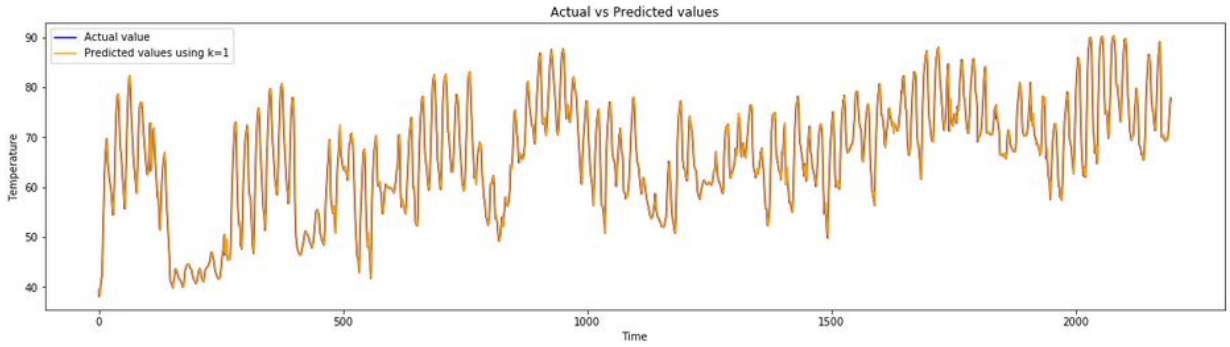
(k=4) RMSE: 4.062
(k=5) RMSE: 4.652
(k=10) RMSE: 6.842



5. Below we have the actual values vs predicted values for various window sizes

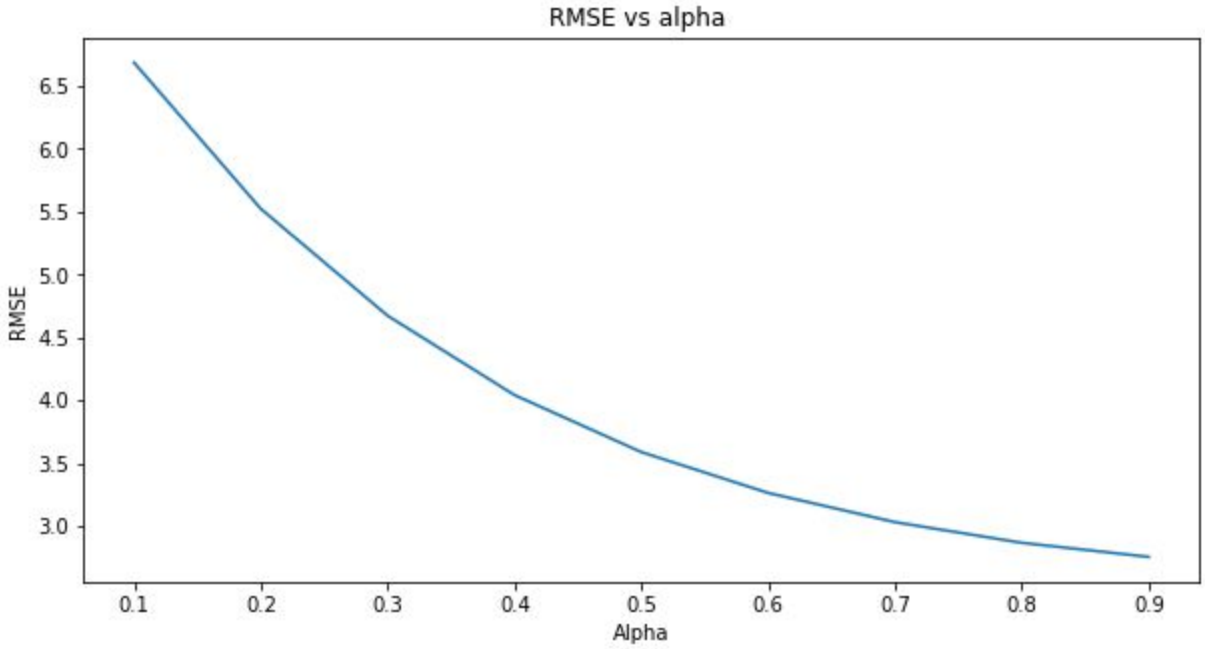


6. The plot of actual vs predicted values also indicates the same.

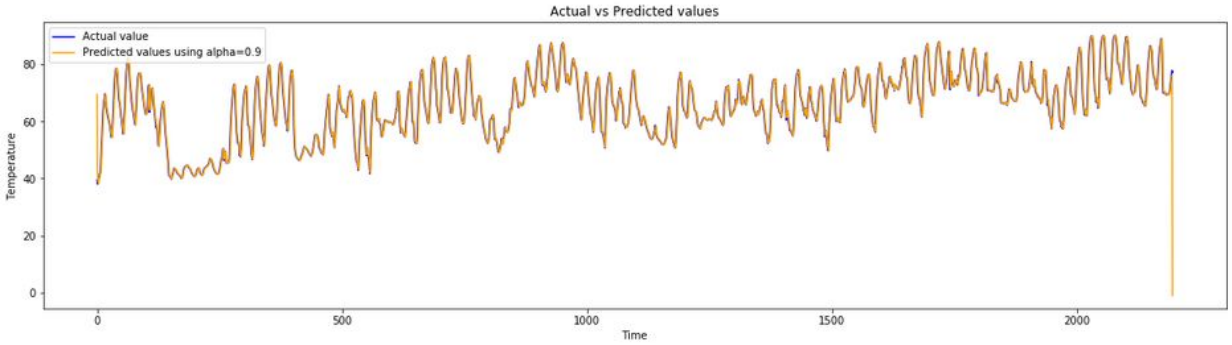


Task3

1. An exponential smoothing model is applied.
2. The RMSE for alpha=0.1 is 6.686
3. Plotting RMSE values for different values of alpha, we get alpha=0.9 to have the lowest RMSE of 2.757



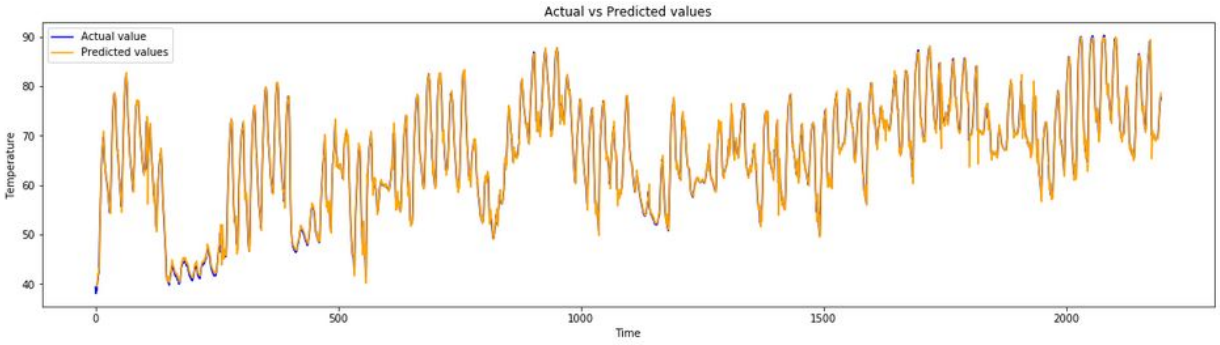
4. We plot the actual and predicted values taking alpha=0.9 and we see that the model performs well as the actual and the predicted values almost overlap.



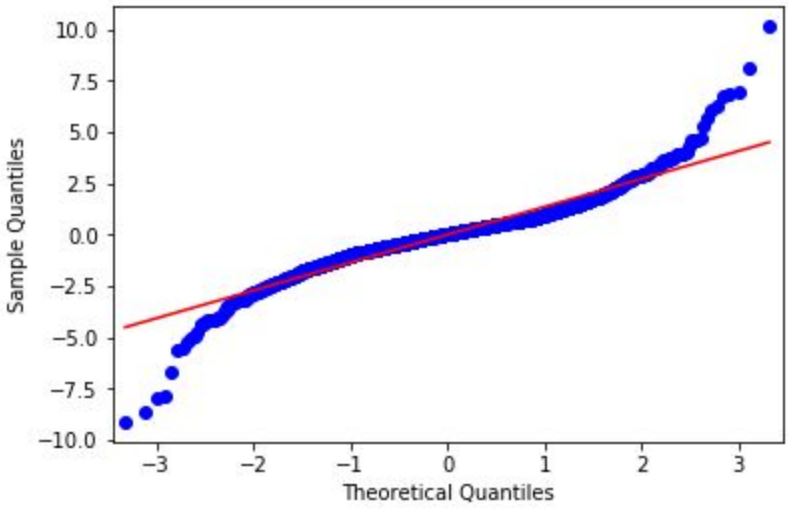
5. Values of α close to one have less of a smoothing effect and give greater weight to recent changes in the data which is suitable to temperature data and we see that alpha=0.9 gives us good results.

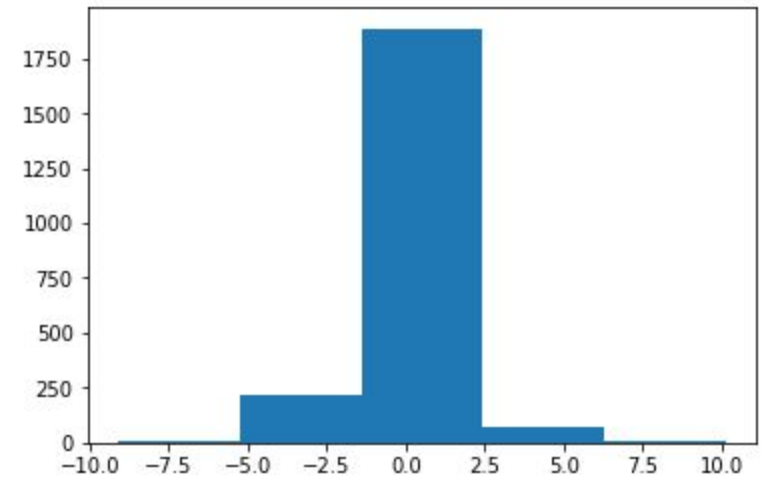
Task4

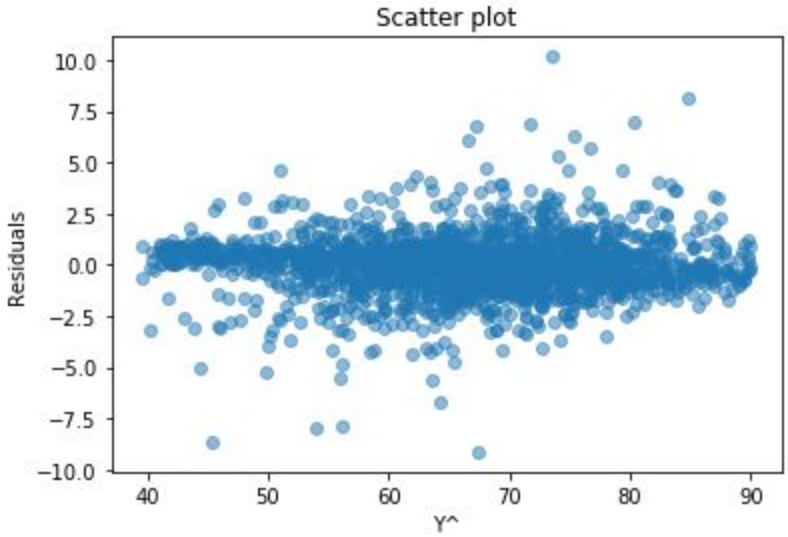
1. We plot the PACF to determine the order p of the AR value.
2. We calculate the PACF coefficients until we get the first zero coefficient. We then fix the order p of the AR(p) model to the lag prior to getting the zero coefficient. As in the case of autocorrelation, partial autocorrelation value is zero if it falls within the confidence interval bounds. In this case, we get an order of 3.
3. The parameters obtained are -
   Coefficients:
   const 1.764205
   L1.temperature 1.720000
   L2.temperature -0.798661
   L3.temperature 0.052104
4. We get an RMSE of 1.359 indicating that the model has been trained well.
5. We plot the actual and the predicted values and we see that the model performs very well



6. Residual analysis -
   a. The qqplot indicates deviates from 45 degree line and hence it is not a normal distribution. The histogram also indicates the same.







7. When a pattern is observed in a residual plot, a linear regression model is probably not appropriate for the data, suggesting a better fit for a non-linear model. We don't see any correlation in the residuals as we don't see a pattern.
8. Chi-square test. Based on reading and observation, we see that the chi-square test is not a very powerful test to check normality and is not dependable. The results vary based on bin size, bin width, value counts per bin and also depend on the method used to calculate expected frequencies.
9. The chi-distribution table gives a value of 0.71 for degree of freedom =2 and alpha=0.95. The null hypothesis is that the residuals follow the normal distribution.
10. The null hypothesis is accepted if $\chi^2 < \chi_{0.95,l-c}$.
    $\chi^2 obtained = 6043.7$ *and* $\chi_{0.95,l-c} = 0.71$
    This implies that the residuals don't follow a normal distribution.

Task5

1. We compare all the models on testing data to obtain the best model. Each model has been tested on testing data at the end of each task.
2. AR(p) - RMSE : 18.425
   Exponential smoothing model - RMSE : 2.427
   Simple moving average model - RMSE : 1.921
3. Based on the performance on the testing data, we see that the simple moving average has the least RMSE and hence is the best model for our data. We see that the simplest algorithm gave us the best results as our data is meteorological data which does not vary a lot. Hence just taking the average of the previous data points will be most suitable for our data.