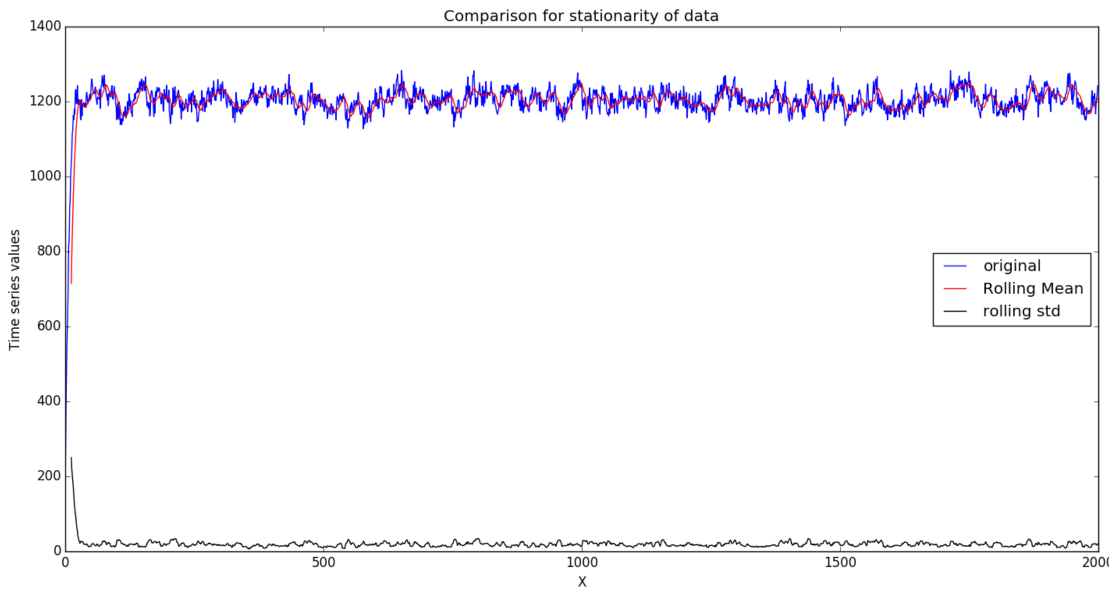
IOT Project 3 - Forecasting

Data Observation: Before applying the different models, stationarity of the dataset should be checked. Hence the given data is plotted and observed for trend, seasonality and variability

If it has a constant mean and variance, the data is assumed stationary

Here is a plot of the data :



As can be seen that data doesn't show any trends and has constant mean and variance, we can assume it to be stationary and carry out the analysis with different models on it. We will ignore the starting 16 values since they are not consistent with rest of the data. Out of the remaining 1984 values, last 500 will be used as test data and remaining as training data.

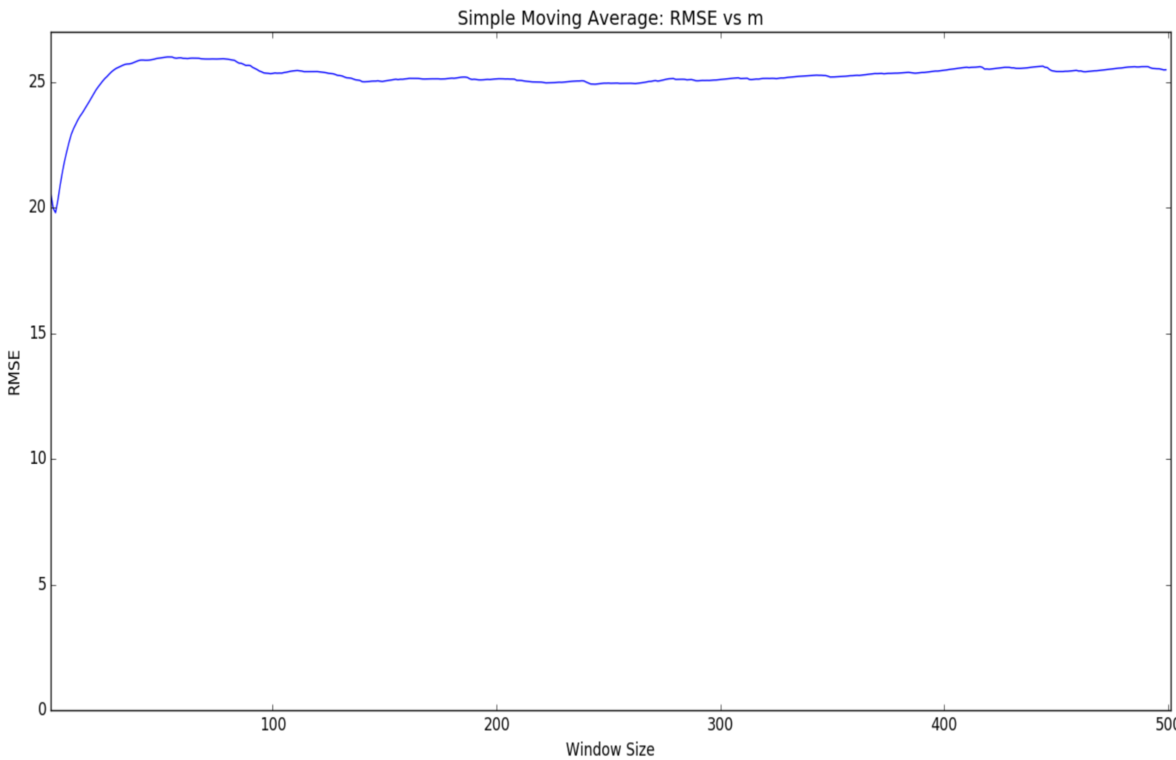
Task 1:

1. **Simple moving average model to the training dataset for a given m**
2. **Calculate the error, i.e., the difference between the predicted and original value in the training data set, and compute the root mean squared error (RMSE)**
3. **Repeat above two steps by varying m and calculate the RMSE**
4. **Plot RMSE vs m. Select m based on the lowest RMSE value. For the best value of m plot the predicted values against the original values**

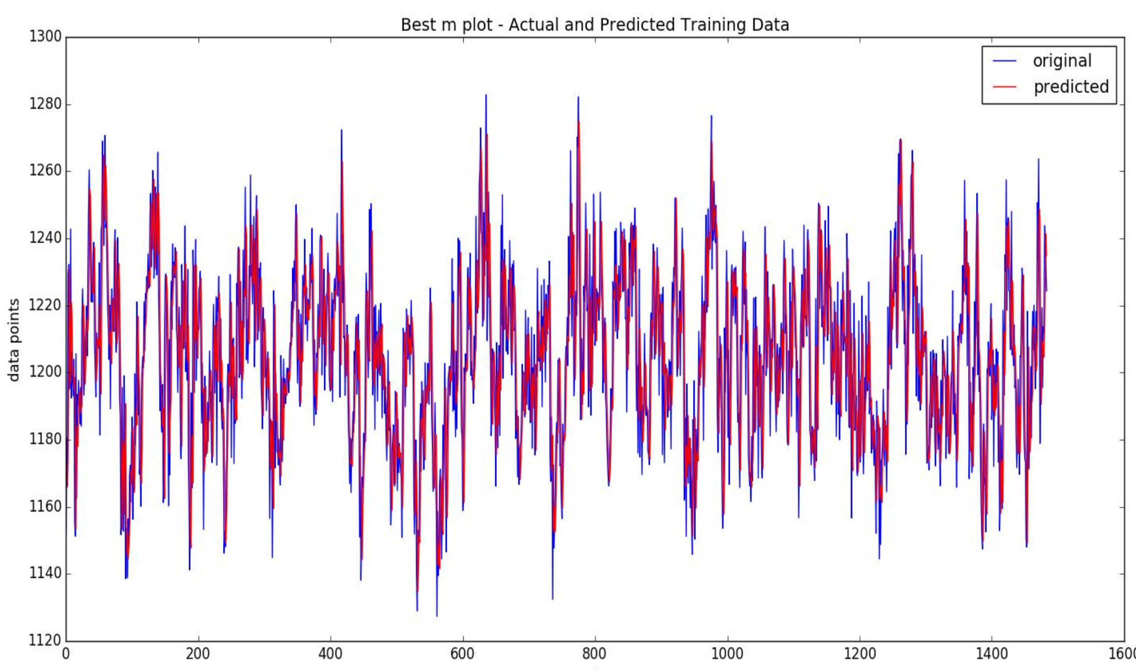
   Answers:

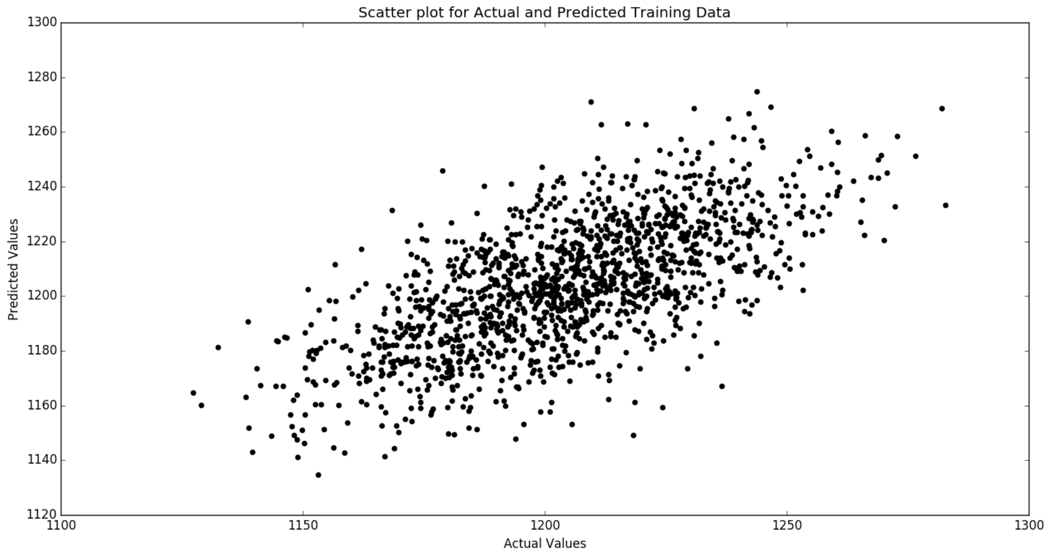
   1. The submitted code has an implementation for simple moving average for a given m value. An example run is done for m = 4 and its RMSE is calculated.
   2. The error calculation and RMSE calculation is done in the code. E.g. for an example run for m =4 on the training set, the calculated RMSE = `20.2781`

3. m is varied from 1 to 500 and simple moving average is run on the range. The RMSE is plotted for each m. Minimum RMSE is 19.8099439293 at m=3.
4. Plot for RMSE vs m. Minimum RMSE is found at m =3



For best value of m i.e. 3, following is the original against vs predicted values plot on training data



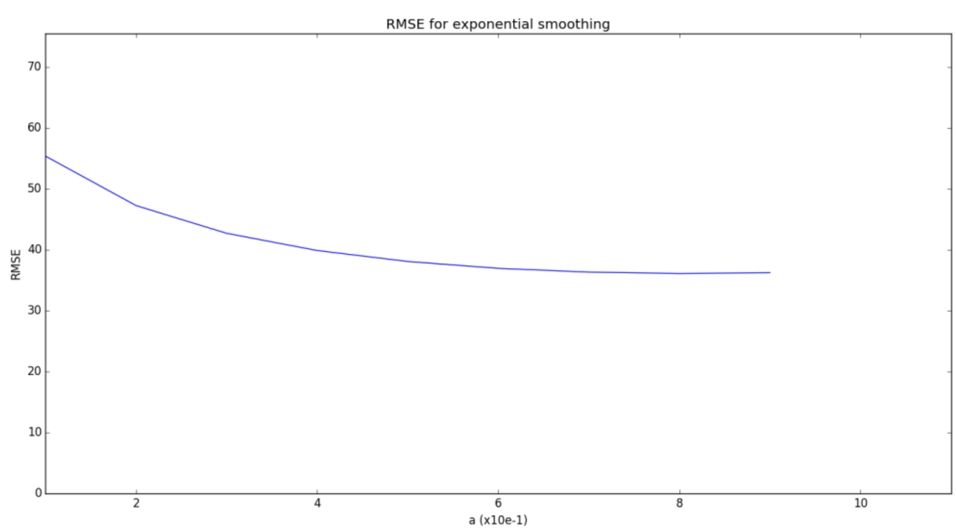


Conclusions: We can see that with varying value of m, the RMSE is increasing and hence it can be concluded that, the time series values are best predicted using the values in their recent past (last three values). Also, with m =3, the RMSE obtained is minimum. This result is returned by the code and also can be seen from the graph plotted above. Therefore, average of past 3 values best predicts the current value. So, in order to forecast on test data, we will use m =3

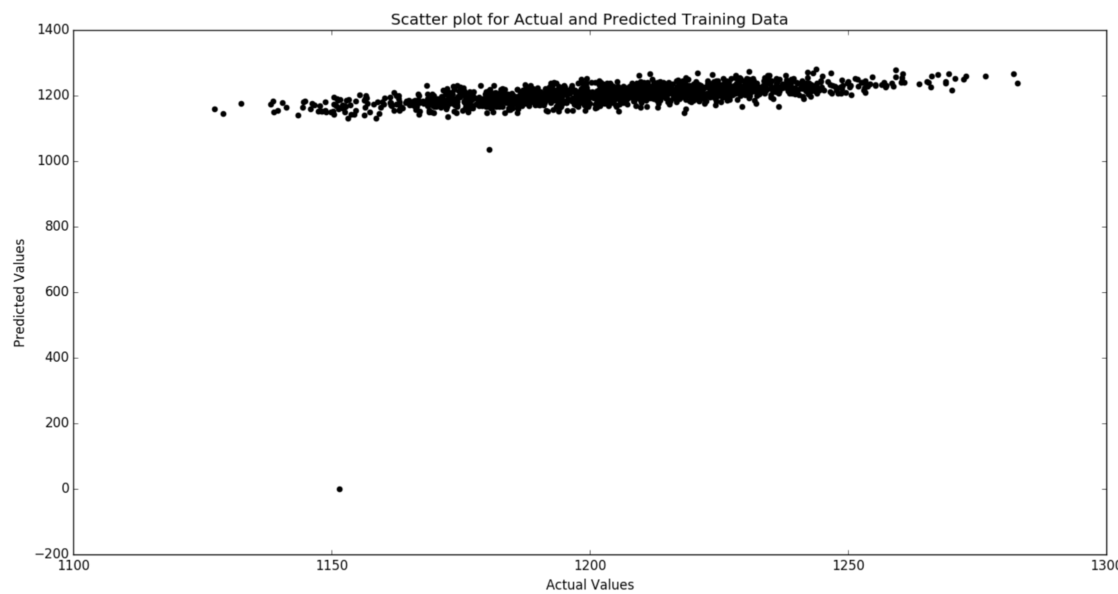
Task 2

1. **Apply exponential smoothing model on the training data**
2. **Repeat the steps to find the best value of α**
3. **Select α with minimum RMSE. Plot original data against training data for best α**
   <u>Answers</u>

1. In the code implementation, α is varied from 0.1 to 1 and the predicted values are found on the training data. Also, RMSE is calculated for each of α values and plotted



2. Minimum RMSE is found at α = 0.9 with value 36.124 (`minexpRmse is: 36.1243422529`)
3. Selecting minimum value of α = 0.9. We plot the training data and predicted values



Conclusions: As can be seen from the plot and the values returned from the code implementation, the minimum RMSE occurs at α = 0.9 and hence it will be used to forecast values on test data. Also the RMSE value is quite big (double of the RMSE of first model). Hence it may not be the best candidate for prediction model.

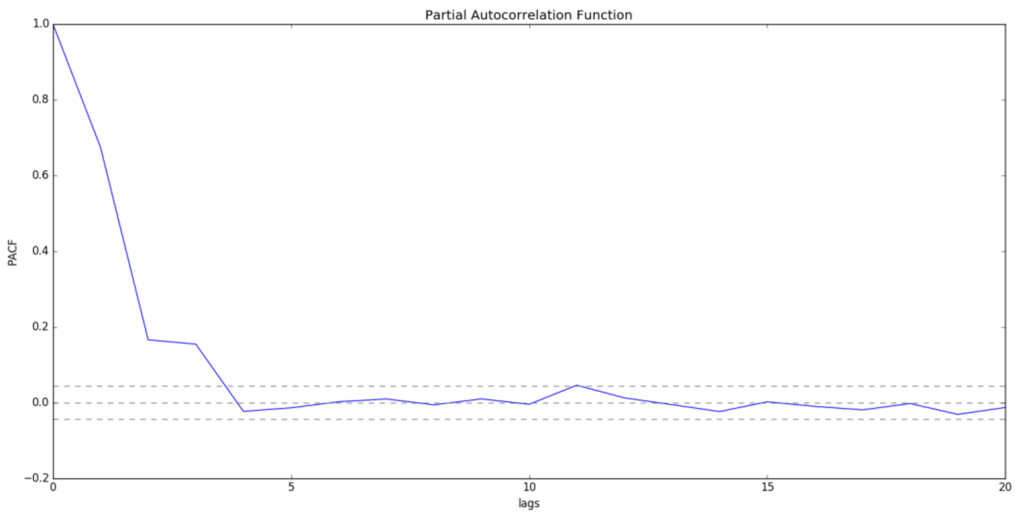
Task 3 AR (p) model

1. **Apply the autoregressive algorithm AR (p) to the training data set for a given value p.**

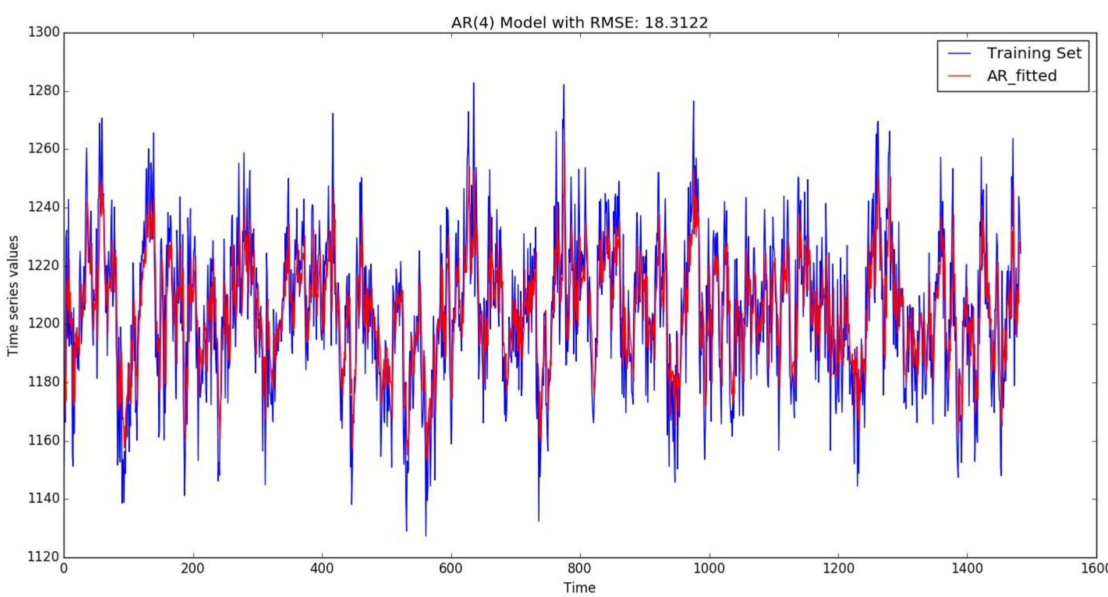
2. **Select p by plotting the partial autocorrelation function**
3. **Estimate the parameters of the AR (p) model. Provide RMSE value and a plot the predicted values against the original values**

<u>Answers</u>

1. Autoregressive model is applied using statsmodel library in python. ARIMA function is used to apply AR with parameters (p,d,q)=(p,0,0) and PACF function is used to plot PACF
2. PACF function is used for a lag of 20 with method as yule walker. Using this the p value is obtained as 4



3. The parameters for the AR(4) are found to be
   [1.20412795e+03 5.42060881e-01 7.98105665e-02 1.70751438e-01 -2.63419621e-02]
   Using ARIMA, the calculated value of RMSE for AR(4) model is 18.3122.
   The plot of original values against the predicted values:



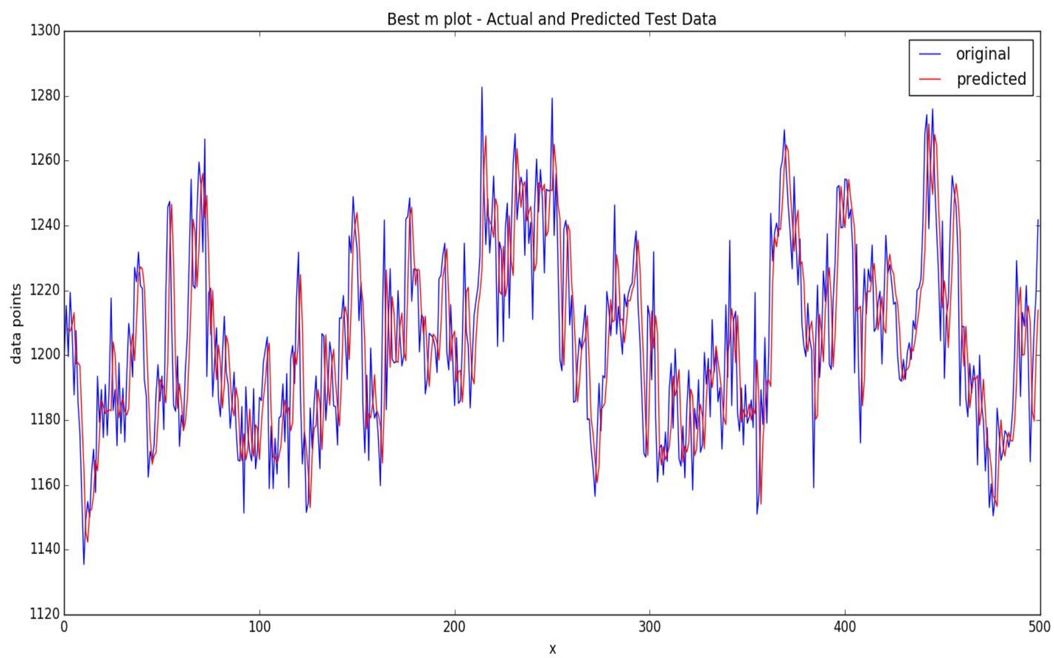
Conclusion: This model gives an RMSE of 18.312 which is by far the best value. This is a good candidate for being the best forecasting model.

Task 4

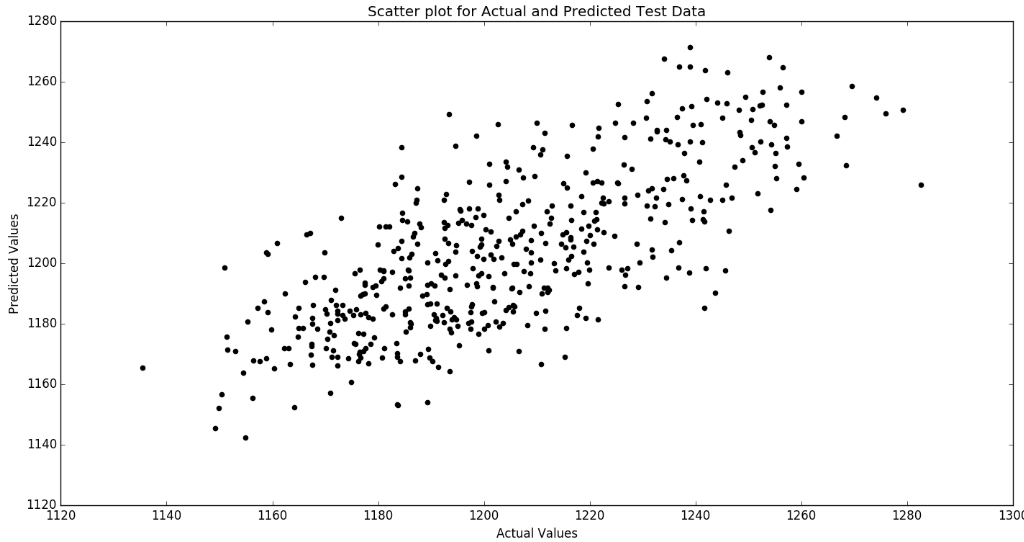
Run all three models on the test data, and chose the best one

After running the three models with their best parameter values, we get the following plots

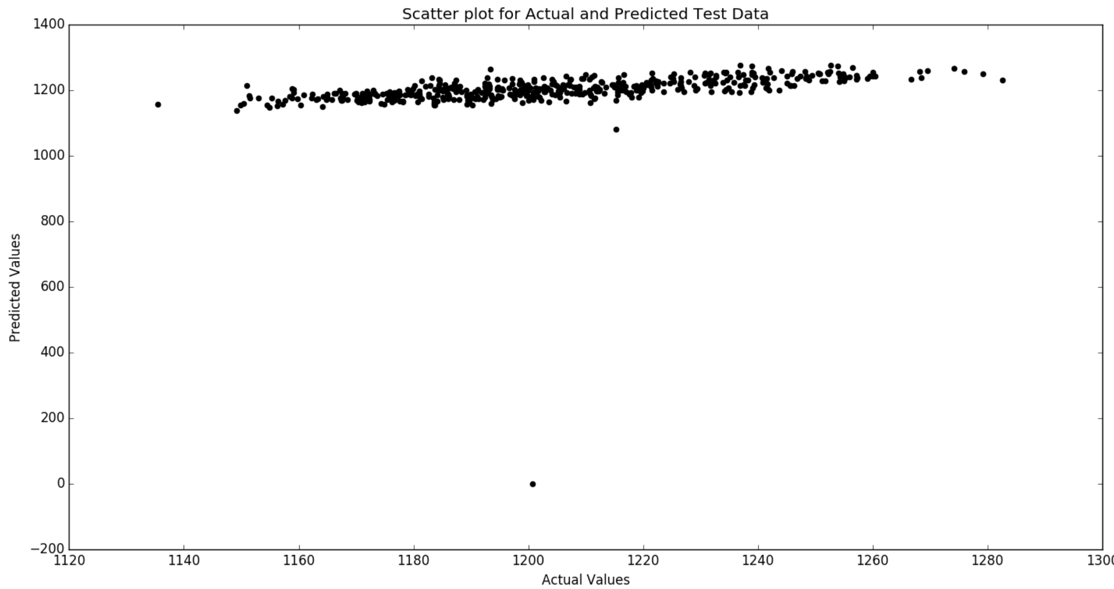
1. Simple moving average with m=3



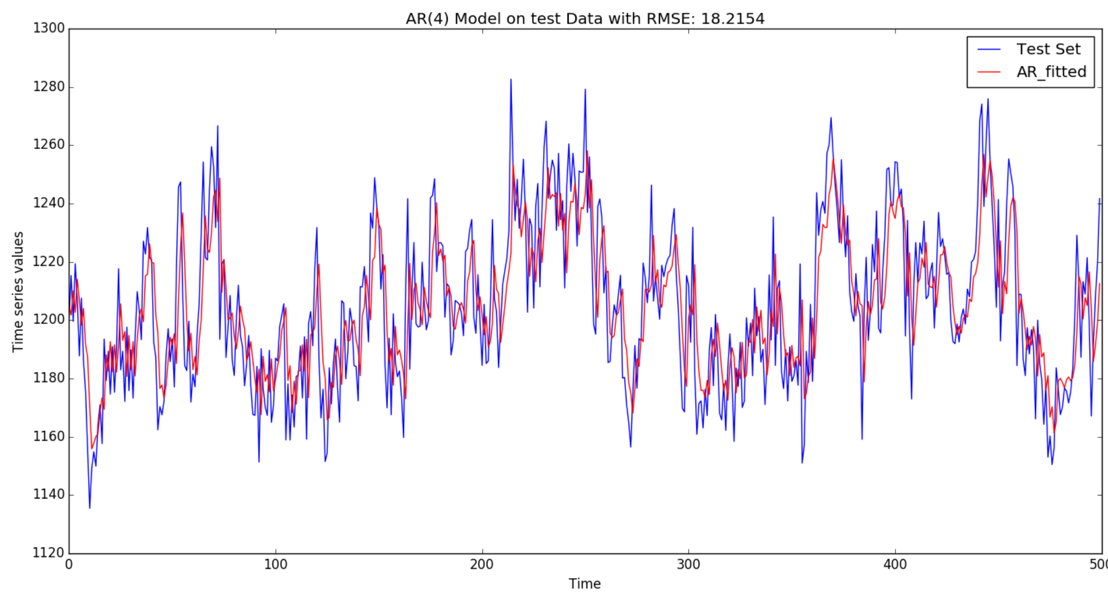
The RMSE obtained from the code for m=3 on test data is 19.5527. Simple moving Average RMSE (Test data) for m = 3 : 19.5527. Looking at the plot, it looks to be a good model choice. Let’s analyze others. Also, a scatter plot comparing the actual and predicted values on test data is given below for simple moving average model for m =3.



2. Exponential smoothing
   Exponential smoothing is applied on test data with α = 0.9 and following value of RMSE and plot are obtained. `Exponential smoothing RMSE for test Data = 57.4917.`
   As can be seen this model is giving a huge RMSE. Also, some values are predicted quite badly (e.g. value near 1200) as can be seen from the scatter plot comparing actual and predicted values. So, this model is not the best forecasting model



3. AR(4) model



With AR(4) model, the RMSE value obtained on predicting test data is 18.2154 which is superior when compared to simple moving average and exponential smoothing model.

Conclusion: We can clearly see that the best prediction results on test data with minimum RMSE is given by model no.3 i.e. AR(4) model and hence it is the best forecasting model for the given data set.