IC 272: DATA SCIENCE - III
LAB ASSIGNMENT – VI
Auto-regression

Student's Name: Jai Prakash Yadav **Mobile No: 9306871378**

Roll Number: B19247 **Branch: ME**

1 a.

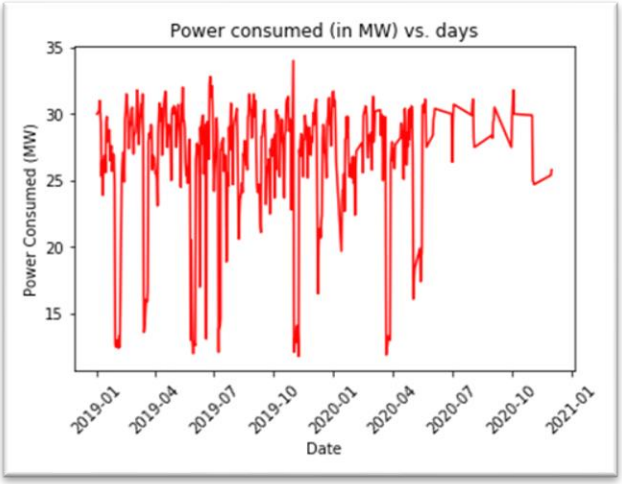

Figure 1 Power consumed (in MW) vs. days

Inferences:

1. Except a few days the days one after the other have similar power consumption.
2. Generally, people consume the same amount of power every day. The days with low power consumption maybe due to power cuts or other technical reasons.

b. The value of the Pearson's correlation coefficient is 0.768

Inferences:

1. From the value of the Pearson's correlation coefficient, we can say that the two time series are highly correlated (positively).
2. We generally expect observations (here power consumption) on days one after the other to be similar. On the basis of correlation coefficient, it holds for most of the days.
3. The two time series are highly dependent, therefore has a strong correlation coefficient.

c.

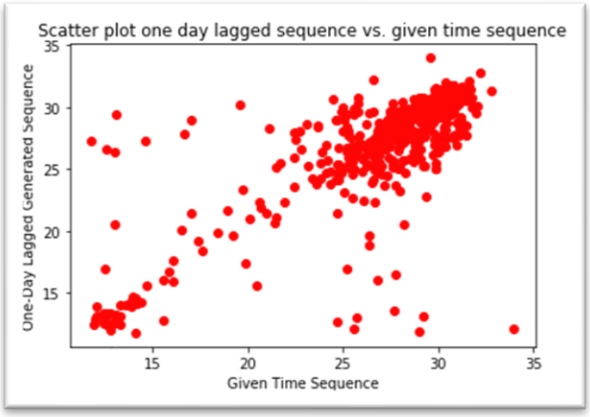

Figure 2 Scatter plot one day lagged sequence vs. given time sequence

Inferences:

1. From the nature of spread of data points, the correlation between the two sequences is very high.
2. Yes, the scatter plot seems to obey the nature reflected by Pearson's correlation coefficient calculated in 1.b.
3. The value of power consumed are almost similar for most of the days. Therefore, the correlation between the two are strong and positive. Also, they have a strong linear relationship.

d.

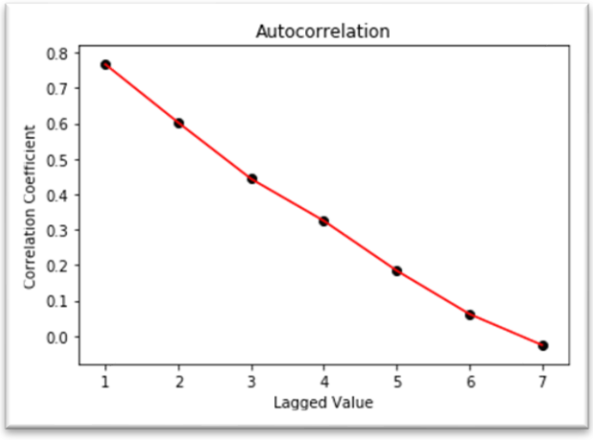

Figure 3 Correlation coefficient vs. lags in given sequence

Inferences:

1. The correlation coefficient value decreases with increase in lags in time sequence for the initial lags.
2. As the no of lags increases the lagged time series dependency with original time series data decreases.

e.

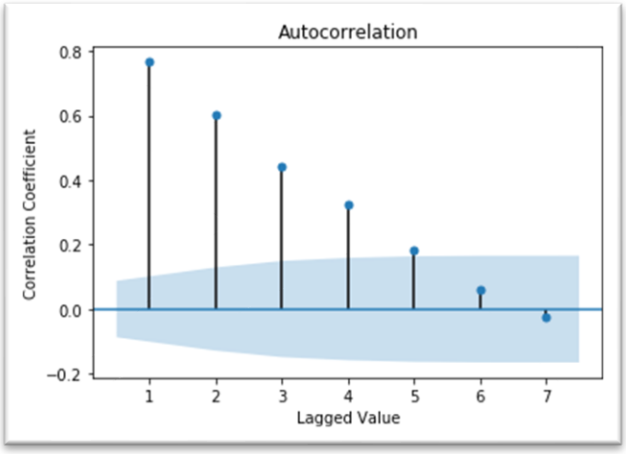

Figure 4 Correlation coefficient vs. lags in given sequence generated using 'plot_acf' function

Inferences:

1. The correlation coefficient value decreases with increase in lags in time sequence for the initial lags.
2. As the no of lags increases the lagged time series dependency with original time series data decreases.

2 The RMSE between predicted power consumed for test data and original values for test data is 3.198.

Inferences:

1. From the value of RMSE value accuracy of persistent model for the given time series is high.
2. The values of adjacent values in the time series are similar. The correlation between the two are also high. So, the persistence model can give an accurate result here.

3 a.

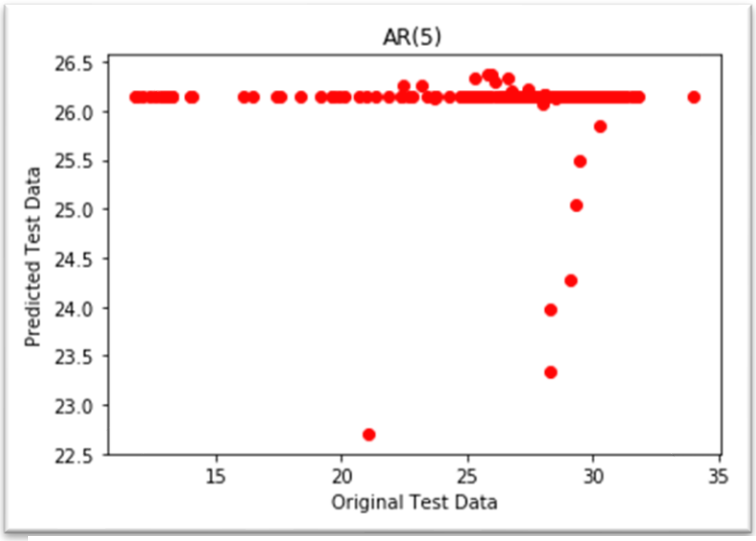

Figure 5 Predicted test data time sequence vs. original test data sequence

The RMSE between predicted power consumed for test data and original values for test data is 4.538.

Inferences:

1. From the RMSE value, accuracy of the model for the given time series is moderate.
2. The Autoregression gives a linear relationship between the input and output data which may not be always true.
3. The autoregression model predicts most values in the range (26,26.5), whereas the test data has values between (10,35).
4. On the basis of RMSE value, the accuracy of persistence model > the accuracy of AR (5) model.

b.

Table 1 RMSE between predicted and original data values wrt lags in time sequence

Lag value	RMSE
1	4.537
5	4.537
10	4.526
15	4.556
25	4.514

Inferences:

1. We cannot observe any specific relationship of RMSE and Lag but for the initial lags (around 60 in this case) as they had almost similar RMSE value but after that RSME started to increase rapidly.
2. RMSE depends upon Auto Regression till the p lags.

c. The heuristic value for optimal number of lags is 5.

The RMSE value between test data time sequence and original test data sequence is 4.537.

Inferences:

1. B1. Based upon the RMSE value, heuristics for calculating optimal number of lags didn't improve the prediction accuracy of the model much.
2. Decisions made using heuristic approach may not necessarily be optimal.

d.

The optimal number of lags without using heuristics for calculating optimal lag is 25.

The optimal number of lags using heuristics for calculating optimal lag is 5.

Inferences:

1. With respect to RMSE values, the prediction accuracy without heuristic for calculating optimal lag is slightly greater than prediction accuracy with heuristic for calculating optimal lag.
2. Decisions made using heuristic approach may not necessarily be optimal.