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

Student's Name: Prashant Kumar **Mobile No: 8700350173**

Roll Number: B19101 **Branch: CSE**

1 a.

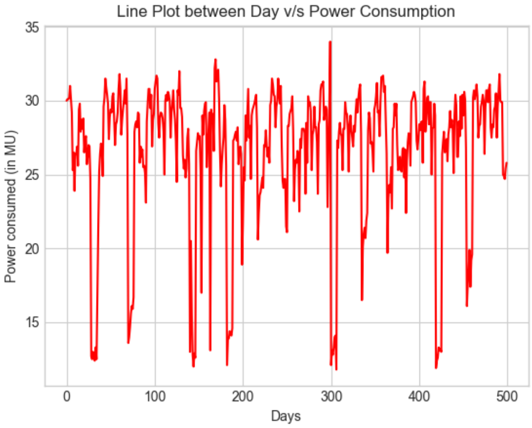

Figure 1 Power consumed (in MW) vs. days

Inferences:

1. The day one after the another has similar power consumption
2. Although graph has many spikes that indicate that value get suddenly increased or decreased. But most of the values lie between the range of 25-30.

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 infer about degree of correlation between the two-time sequences is high.
2. Here power consumption on days one after the other to be similar as we verified from the high value of Pearson's correlation. High value concludes that the one day lagged value is more correlated with the actual value.
3. This is because power consumption is almost similar for day one after the another.

c.

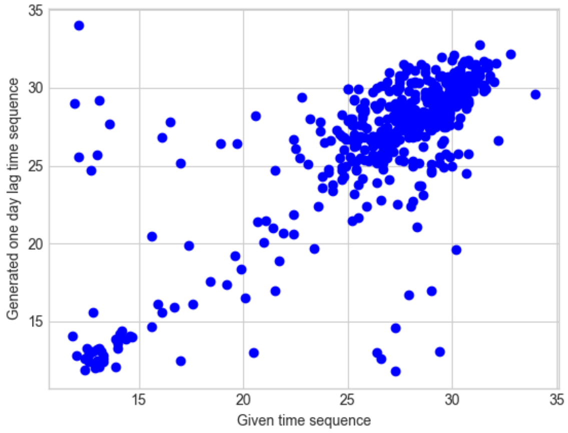

Figure 2 Scatter plot one day lagged sequence vs. given time sequence

Inferences:

1. From the nature of spread of data points, we can infer about the nature of correlation between the two sequences is that the sequence is linear and has high correlation between them.
2. Yes, the scatter plot seems to obey the nature reflected by Pearson's correlation coefficient calculated in 1.b.
3. As the correlation is also high, scatter plot is almost linear. This is because the power consumption is almost similar for the day one after the another.

d.

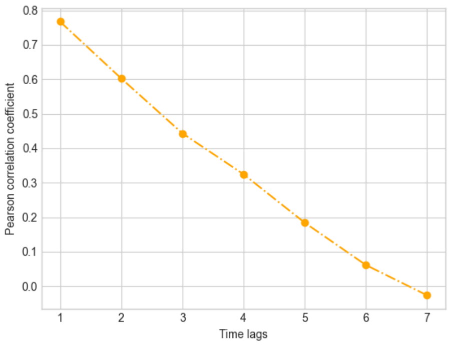

Figure 3 Correlation coefficient vs. lags in given sequence

Inferences:

1. Correlation coefficient value get decrease with respect to increase in lags in time sequence linearly.
2. Reason of the observed trend is as we increase the lags, the similarity between them get decrease. For starting the correlation coefficient is maximum shows that the sequence is almost similar. On increasing time lag trend of similarity with the original value get decrease.

e.

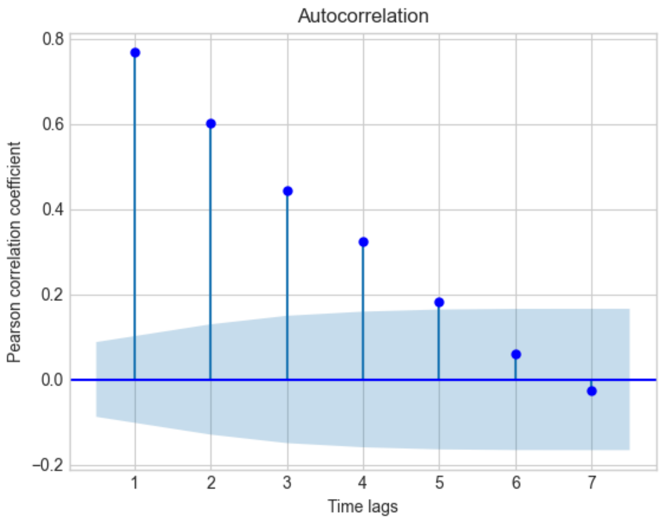

Figure 4 Correlation coefficient vs. lags in given sequence generated using 'plot_acf' function

Inferences:

1. Correlation coefficient value get decrease with respect to increase in lags in time sequence linearly.
2. Reason of the observed trend is as we increase the lags, the similarity between them get decrease. For starting the correlation coefficient is maximum shows that the sequence is almost similar. On increasing time lag trend of similarity with the original value get decrease.

2 The RMSE between predicted power consumed for test data and original values for test data is **3.192**.

Inferences:

1. From the value of RMSE value, we can say that the persistent model for the given time series has decent accuracy.

2. Reason is similar that the power consumption is almost similar on day one after the another. Since correlation is also high which means high correlated. And also the persistence model also used the (t-1)th as a predicted output of (t)th data. So, we can say that the accuracy is decent

3 a.

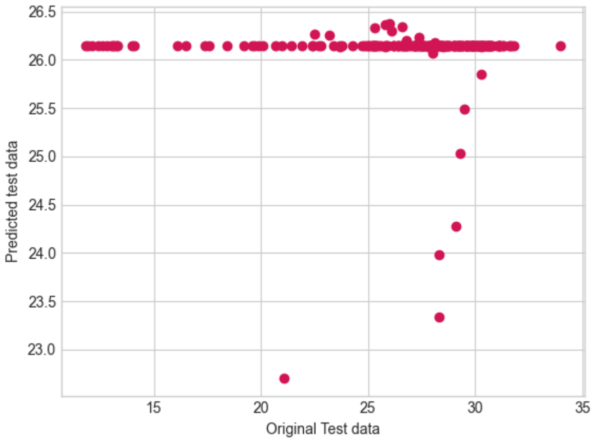

Figure 5 Predicted test data time sequence vs. original test data sequence

The RMSE between predicted power consumed for test data and original values for test data is **4.537**.

Inferences:

1. From the value of RMSE we can say that the model for the given time series is not quite good.
2. Reason behind this is that the model is predicting value between 26-26.5 for all inputs. Persistence model predict quite better than this model.
3. From the plot of predicted test data time sequence vs. original test data sequence we can say that the model is less reliable for future predictions. Since, persistence model is quite better.
4. On the basis of RMSE value, we can say that the persistence model has good accuracy as compared to the current 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. There is no regular trend as the RMSE get increase on 15 ad decrease on 25. But usually RMSE is decreasing.
2. Reason behind this, is that on increasing the time lag we are considering all lagged data value that are more correlated to predict the value.

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. Although the RMSE that we calculated through heuristic didn't decreased the value as compared to previous one. But it preserved the relation with the previous time lags and the model is balanced.
2. In the c part the lag is 25 which will consider more data as an input to predict the output. Hence, give good prediction accuracy as compared to lag with value equal to 5.

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. The prediction accuracies obtained without heuristic for calculating optimal lag with respect to RMSE values is high as compared to the with heuristic.
2. Reason is that the lag value equal to 25 which will consider more data as an input to predict the output. Hence, give good prediction accuracy as compared to lag with value equal to 5.