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

Student's Name: Rashmi

Mobile No: 7015331137

Roll Number: B19218

Branch: Engineering Physics

1 a.

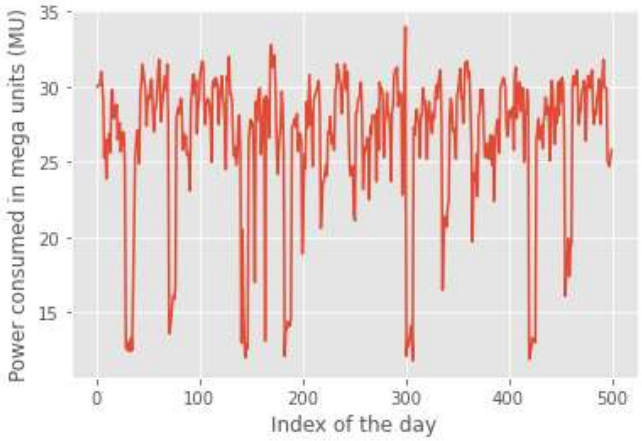

Figure 1 Power consumed (in MW) vs. days

Inferences:

1. The days one after the other have similar power consumption.
2. We can see a cyclic dipping pattern in the line plot above.

b. The value of the Pearson's correlation coefficient is **0.767**.

Inferences:

1. Original time sequence and one-day-lagged sequence have a high correlation of 0.767, so they are closely related with positive correlation.
2. They are similar we can deduce x(t) from x(t-1) as it has high correlation coefficient.
3. This is related to the seasonal behaviour of electricity usage as well as assumption that electricity usage won't suddenly increase between a short period of time. This fact is dictated by the high Pearson correlation.

c.

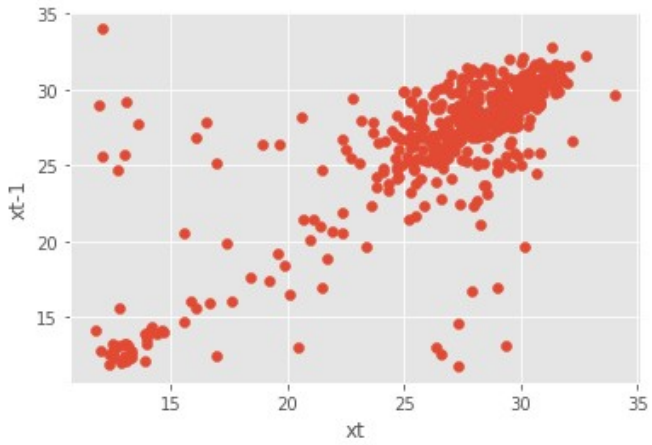

Figure 2 Scatter plot one day lagged sequence vs. given time sequence

Inferences:

1. Most of the values are on y = x so they are highly correlated. We can infer that from high correlation coefficient.
2. The scatter plot seems to obey the nature reflected by Pearson's correlation coefficient calculated in 1.b
3. A high positive Pearson correlation indicates that one variable increase as the other increases.

d.

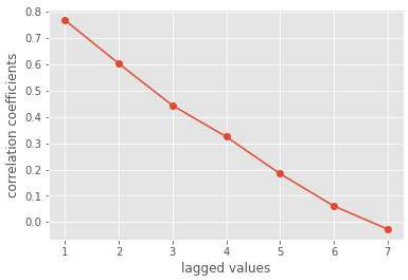

Figure 3 Correlation coefficient vs. lags in given sequence

Inferences:

1. Correlation Coefficient decreases with increase in lagged values.
2. If we start to increase the time-lag then, the dependence on previous data becomes more unreliable as it gives a larger window of opportunity for the data to fluctuate.

e.

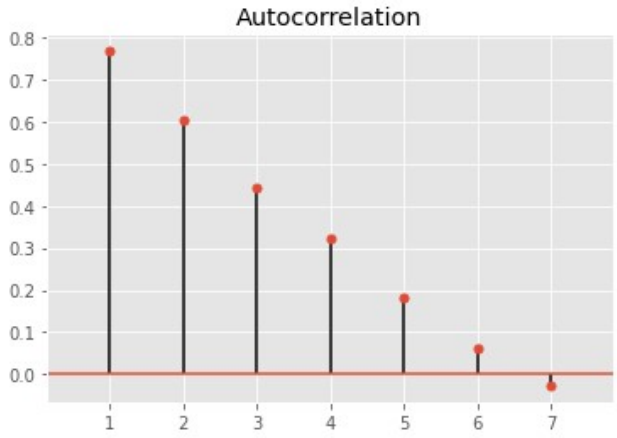

Figure 4 Correlation coefficient vs. lags in given sequence generated using 'plot_acf' function

Inferences:

1. Correlation Coefficient decreases with increase in lagged values.
2. If we start to increase the time-lag then, the dependence on previous data becomes more unreliable as it gives a larger window of opportunity for the data to fluctuate.

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

Inferences:

1. It is reasonably accurate, since it's giving an error of approx. 12% relative to the average data point in the series.

2. Persistent model can be thought of as the simplest autoregression model (with time lag 1). As we see in the graphs, plotted above, for time-lag 1, the correlation is high, which leads to relatively okay predictions of the persistent model.

3 a.

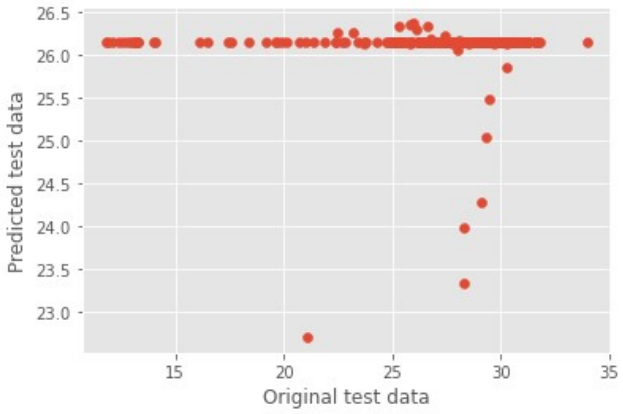

Figure 5 Predicted test data time sequence vs. original test data sequence

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

Inferences:

1. It is less accurate.
2. Most of the values are very far away from the predicted values.
3. It would be preferred to use the persistence model in favour of this one; because as we see in the graph, this model abandons a linear relationship between the variables right around the central tendencies of the data.
4. On the basis of RMSE value, persistent model in Q2 is better as RMSE values is lower as compared to this model.

b.

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

Lag value	RMSE
1	4.5367
5	4.537

10	4.5263
15	4.5558
25	4.5141

Inferences:

1. For small lag values, the RMSE decreases on increasing the lag value. After certain value, RMSE suddenly increases. For further value, nature of RMSE is unpredictable.
2. As we increase the lag value then, our model tries to cover more information from data set and there is a high variance in output. If we take further large value then, overfitting of curve happens. In this case, our model loses its generalizability and there is a high error.

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. Based upon the RMSE value, heuristics for calculating optimal number of lags didn't improve the prediction accuracy of the model much.
2. Heuristic value only helps in deciding the best lag value. The prediction accuracy can only be improved if we add some non-linear prediction ability to our model.

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 accuracy obtained without heuristic is better.
2. $Accuracy\ \alpha$ 1/ $RMSE$, as we see RMSE is lower for non-heuristic lag value of p=25 so accuracy is high.