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

Student's Name: Tarun Singla **Mobile No: 8872526396**

Roll Number: b19198 **Branch: EE**

1 a.

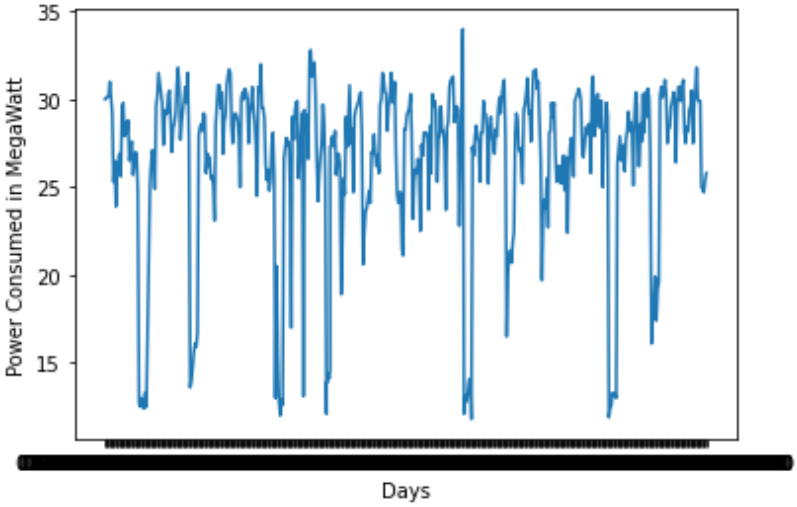

Figure 1 Power consumed (in MW) vs. days

Inferences:

1. The days one after the other do have similar power consumption.
2. Though there are continuous spikes in the graph from but most of the values are in the range from 25 – 30.

b. The value of the Pearson's correlation coefficient is **0.7650**

Inferences:

1. From the value of the Pearson's correlation coefficient the are highly correlated and it shows positive

correlation.

2. They are similar we can deduce x(t) from x(t-1) as it has high correlation coefficient.
3. Because there is not much change in day to day basis of population. So there is not much change in power consumption.

c.

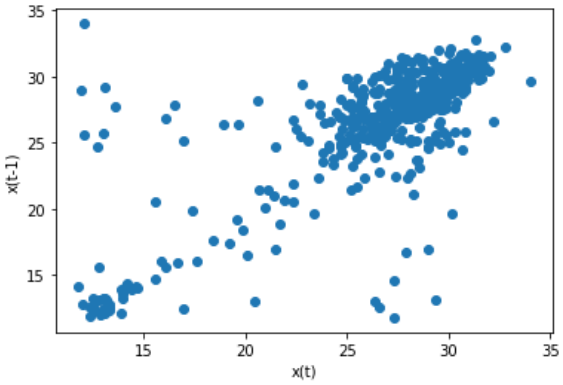

Figure 2 Scatter plot one day lagged sequence vs. given time sequence

Inferences:

1. Most of the values are on y = x so they are highly correlated. We can infer that from high correlation coefficient.
2. Yes, the scatter plot obeys the nature reflected by Pearson correlation coefficient. As we can see from the part b.
3. We can see that with high correlation coefficient and from graph as most of the values are along y = x and plot in 1a infers the same

.

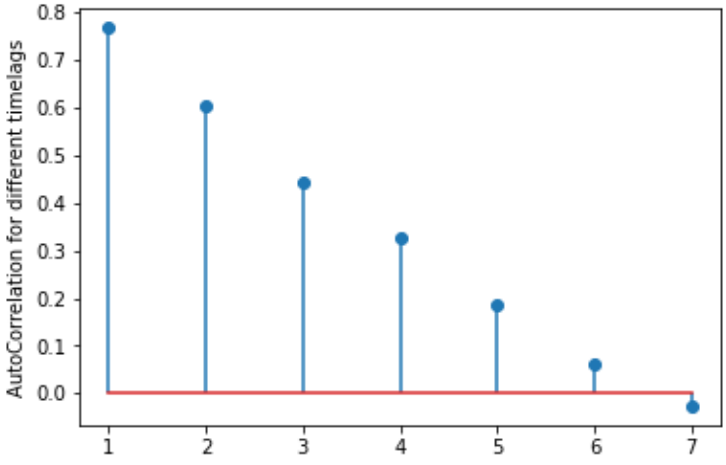

Figure 3 Correlation coefficient vs. lags in given sequence

d.

Inferences:

1. Correlation Coefficient decrease with increase in lagged values.
2. This is because as we increase the lagged values. More Lagged values will make data unrelated.

e.

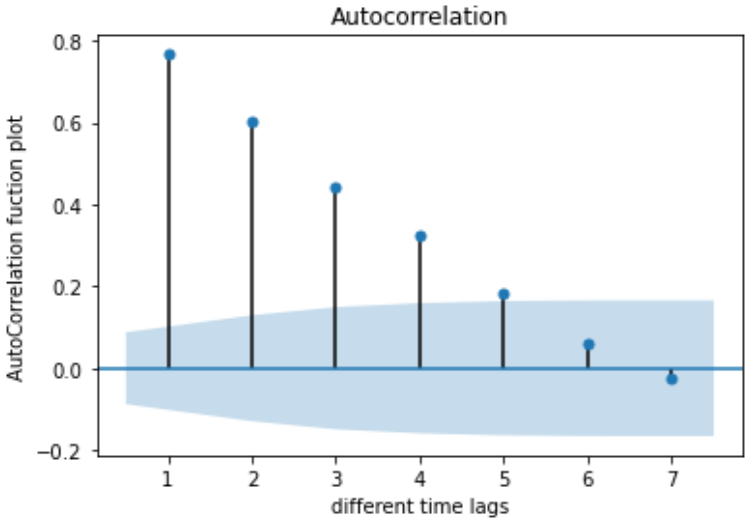

Figure 4 Correlation coefficient vs. lags in given sequence generated using 'plot_acf' function

Inferences:

3. Correlation Coefficient decrease with increase in lagged values.
4. This is because as we increase the lagged values. More Lagged values will make data unrelated.

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

Inferences:

1. RMSE error is not quite high so persistent model is not bad it can be used for prediction.
2. RMSE error is low. It is predicting correctly, and modal is good.

3 a.

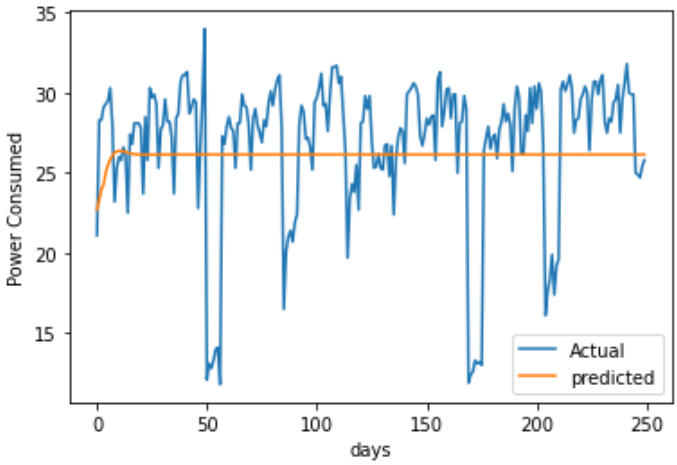

Figure 5 Predicted test data time sequence vs. original test data sequence

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

Inferences:

1. This model is not good.
2. Most of the values are very far away from the predicted values.
3. This model is good for future prediction As there is very much difference in predicted and Actual values.
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	4.539
5	4.538
10	4.526
15	4.556
25	4.514

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 his 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 **4.538**

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 and RMSE is **4.514**

The optimal number of lags using heuristics for calculating optimal lag is 5 and value is **4.537**.

Inferences:

1. The prediction accuracy obtained without heuristic is better.
2. Heuristic approach works on the value of autocorrelation and number of observations. It's quite possible that current value is highly dependent on some lag value but due to a smaller number of observations it gets neglected.

Guidelines for Report (Delete this while you submit the report):

- **The plot/graph/figure/table should be centre justified with sequence number and caption.**
- **Inferences should be written as a numbered list.**
- **Use specific and technical terms to write inferences.**
- **Values observed/calculated should be rounded off to three decimal places.**
- **The quantities which have units should be written with units.**