IC 272: DATA SCIENCE - III

LAB ASSIGNMENT – VI

Auto-regression

Student's Name: Gajraj Singh Chouhan **Mobile No: +91-9351159849**

Roll Number: B19130 **Branch: DSE**

1 a.

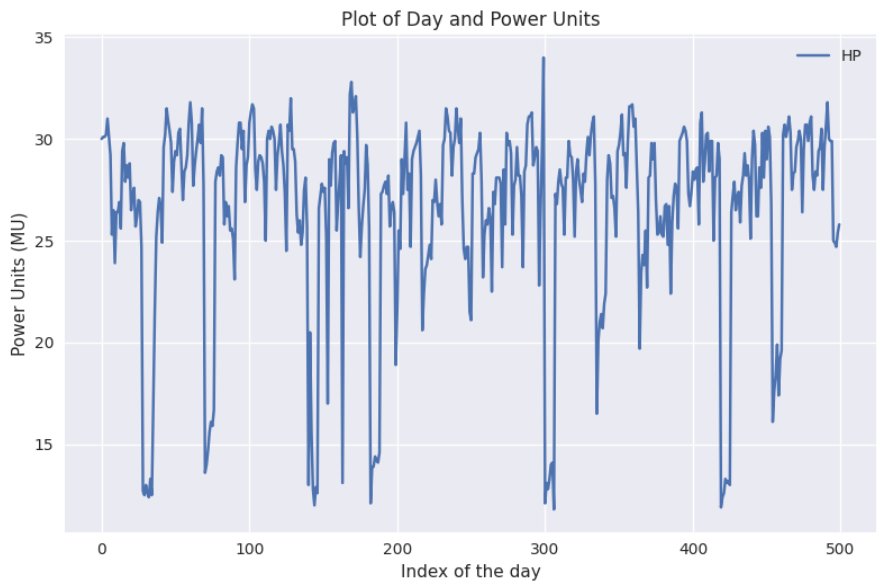

Figure 1 Line plot of index of day vs the power units at that day

Inferences:

1. The days one after the other have somewhat similar power consumption.
2. We can say they are similar, because the high Power Units (y axis) are generally followed by similarly high power units and the same goes for the lower power units.

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

Inferences:

1. According to Pearson's correlation coefficient , correlation between the two time sequences is strong.
2. Here the power consumption on days one after the other is very similar as we could infer from the Pearson's Correlation Coefficient.
3. For the correlation coefficient greater than 0.6, we can say the variables will be having strong linear correlation. Now as strong the coefficient is, more similar will be the variables (here the time sequences).

c.

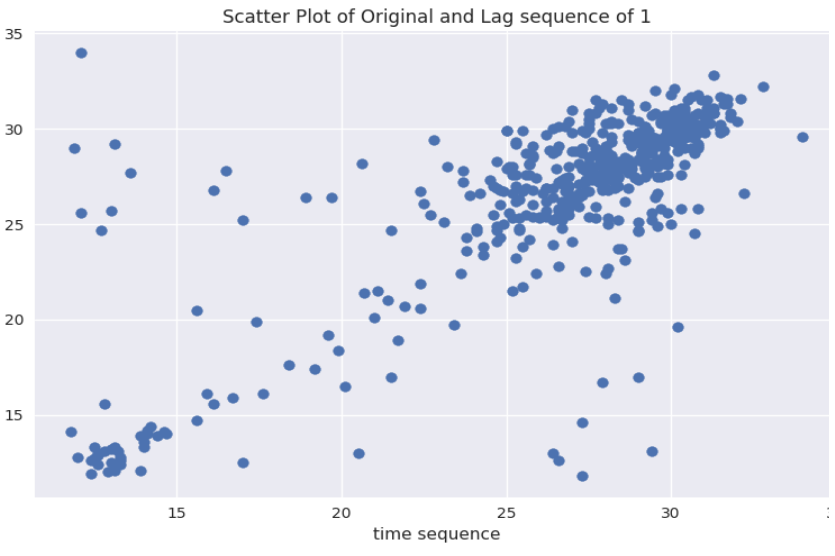

Figure 2 Scatter plot one day lagged sequence vs. given time sequence

Inferences:

1. From the spread of data, we can say the correlation is positive and it will be strong due to less spread for larger time stamps.
2. Yes, the scatter plot seems to obey the nature reflected by Pearson's correlation coefficient calculated in 1.b.
3. The strong correlation was also reflected by the high correlation coefficient of 0.767, so the plot follows the correlation coefficient.

d.

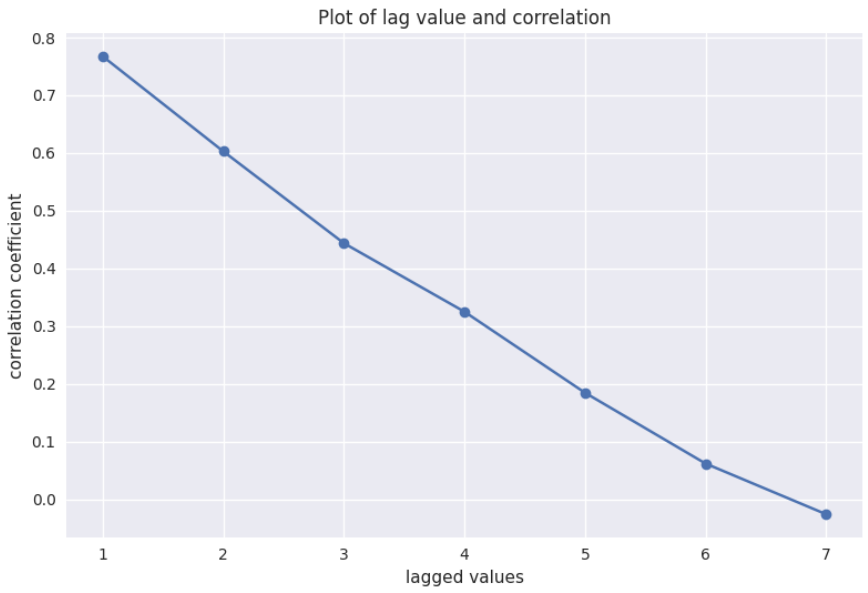

Figure 3 Plot of Correlation Coefficient

Inferences:

1. As the lag values increase, our correlation coefficient decreases from 0.767 to ~-0.1 when lag = 7.
2. As the lagged value increases, we can see the data becoming very much less similar as the data is kind of periodic in the sense that it switches from low to high value in approx 3-4 days. Hence at values>=3 the correlation is becoming weaker and weaker.

e.

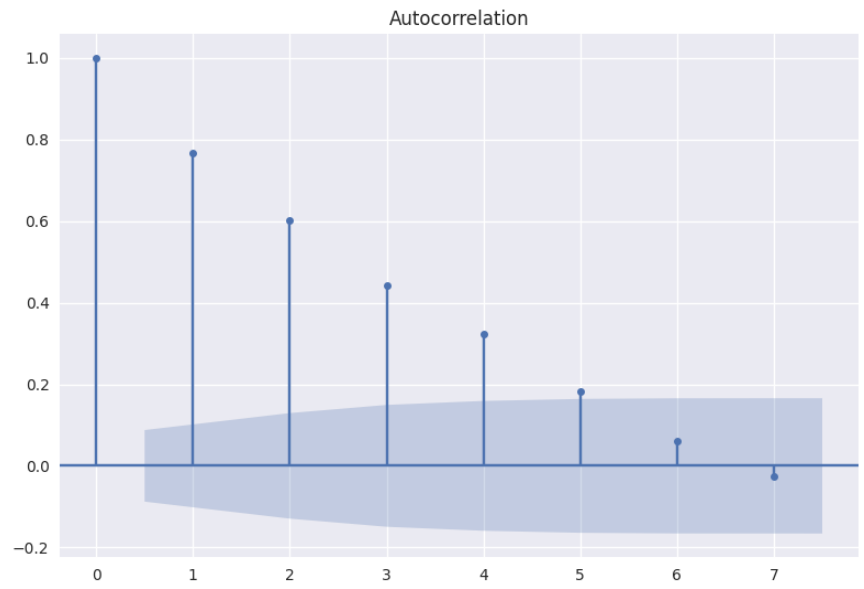

Figure 4 Correlation coefficient vs. lags in given sequence generated using 'plot_acf' function

Inferences:

1. As the lag values increase, our correlation coefficient decreases from 0.767 to -0.1.
2. As the lagged value increases, we can see the data becoming very much less similar as the data is kind of periodic in the sense that it switches from low to high value in approx 3-4 days. Hence at values>=3 the correlation is becoming weaker and weaker.

2. The RMSE between predicted power consumed for test data and original values for test data is 3.1917.

Inferences:

1. The persistent model is very much accurate for the given time series.
2. As it's RMSE is lower than the autoReg model we will use, we can say it's very much accurate.

3 a.

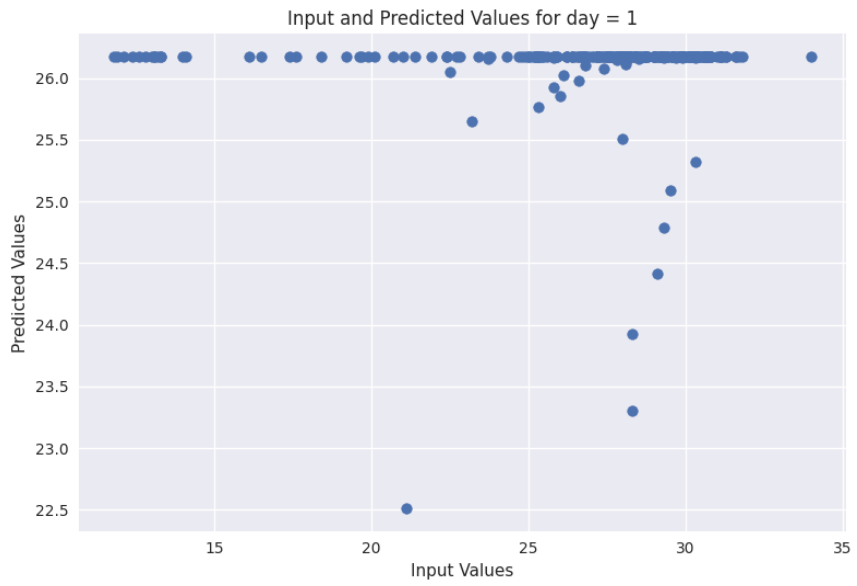

Figure 5 Predicted test data time sequence vs. original test data sequence

The RMSE between predicted power consumed for test data and original values for test data is 4.536.

Inferences:

1. According to this model the accuracy is quite weak as seen in the graph.
2. The predicted value remain constant mostly with high spread for the input values, which is not true hence the low accuracy.
3. For future predictions of the given time series the model will pretty much have low accuracy, hence it will less reliable.
4. It is less accurate than the persistence model.

b.

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

Lag value	RMSE
1	4.536
5	4.537
10	4.526
15	4.555
25	4.514

Inferences:

1. RMSE first decreased then increased till 15 then again decreased till 25 with respect to increase in lags in time sequence.
2. Due to less accuracy of the model, the RMSE is increasing ‘cause the model gives constant (not always but almost like ~26 as seen) high values which differ from the actual high and low values of the graph thus increasing the RMSE.

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. Yes, using heuristics for calculating optimal number of lags improves the prediction accuracy of the model.
2. As using the heuristic value = 5, the RMSE decreased hence the accuracy increased.

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 4.514

Inferences:

1. The heuristic value has a lower accuracy than the non heuristic one.
2. As the RMSE value is lower for lag = 25, we can say the non heuristic one is more accurate.