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

Student's Name: Sumit Kumar **Mobile No: 7549233722**

Roll Number: B19118 **Branch:CSE**

1 a.

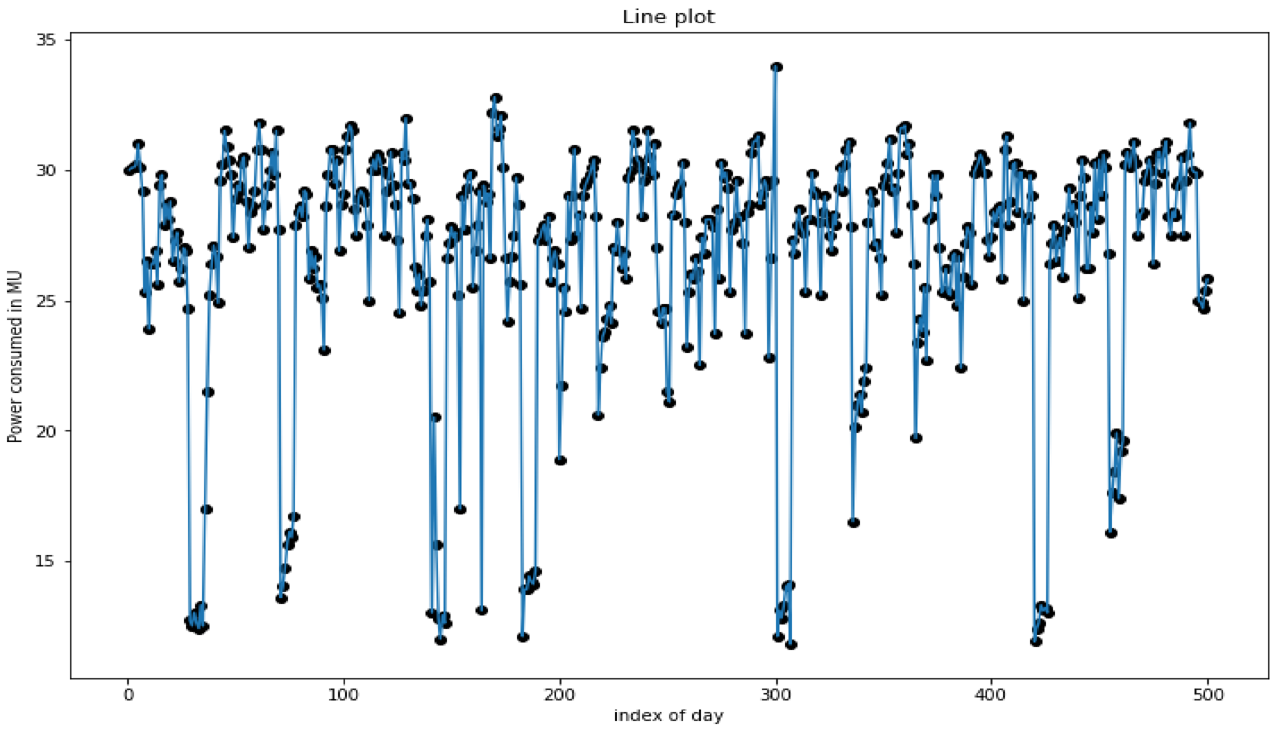

Figure 1 Power consumed (in MW) vs. days

Inferences:

1. The Power consumption for most of the days are nearly similar, in range 25-30 MU, but for some days the Power Consumption is significantly low.
2. The above plot shows that there is a sudden fall in the power consumption maybe due to the power cut or other technical issues.

b. The value of the Pearson’s correlation coefficient is **0.7675**

Inferences:

1. The two time series are highly correlated(positively) from the value of the correlation coefficient.
2. From the value of the correlation coefficient, the two time series is highly correlated and we can say that the power consumption on the days one after the another to be similar and it holds for most of the days. The value of the lag data increases with the increase in the given data.
3. The two time series are very highly dependent because the absolute value of the Pearson’s correlation coefficient is high and it shows the positive correlation and highly dependent.

c.

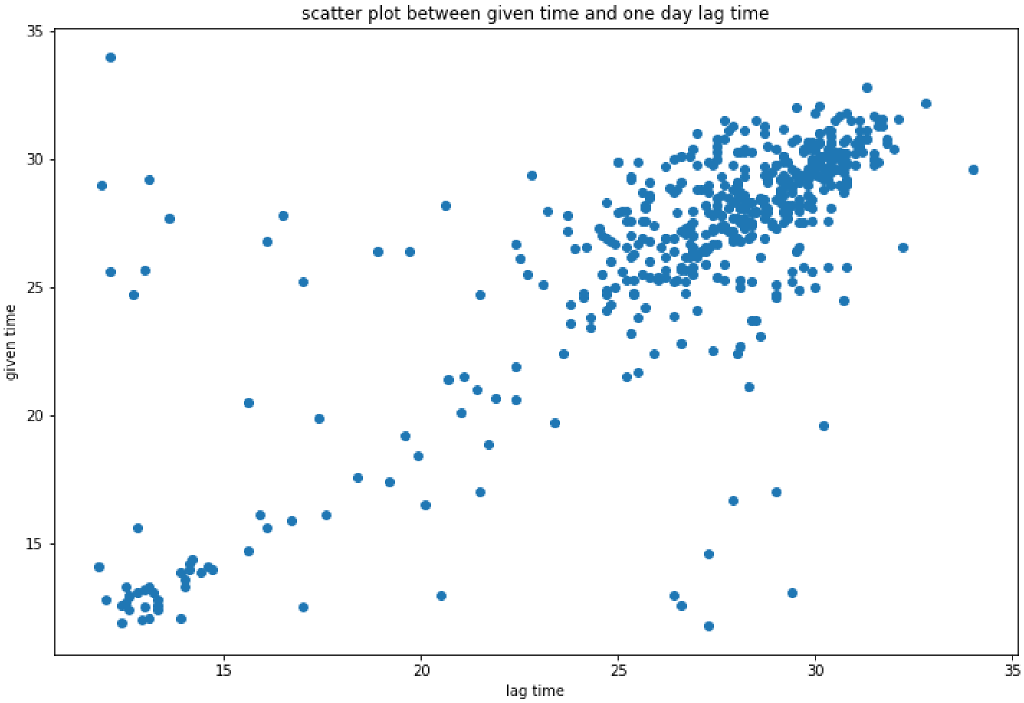

Figure 2 Scatter plot one day lagged sequence vs. given time sequence

Inferences:

1. From the above scatter plot, we can say that the correlation coefficient of the data is very high and positively correlated. As the value of lag time data increases, the given time data also increases. There is a stronger relationship between the data.
2. Yes, the scatter plot seem to obey the nature reflected by Pearson's correlation coefficient calculated in 1.b
3. The value of the by Pearson's correlation coefficient shows that the data is highly correlated. Pearson's correlation coefficient is found to be high and positive which shows that as the lag time data increases, the value of the given time data also increases which can be easily seen in the above scatter plot.

d.

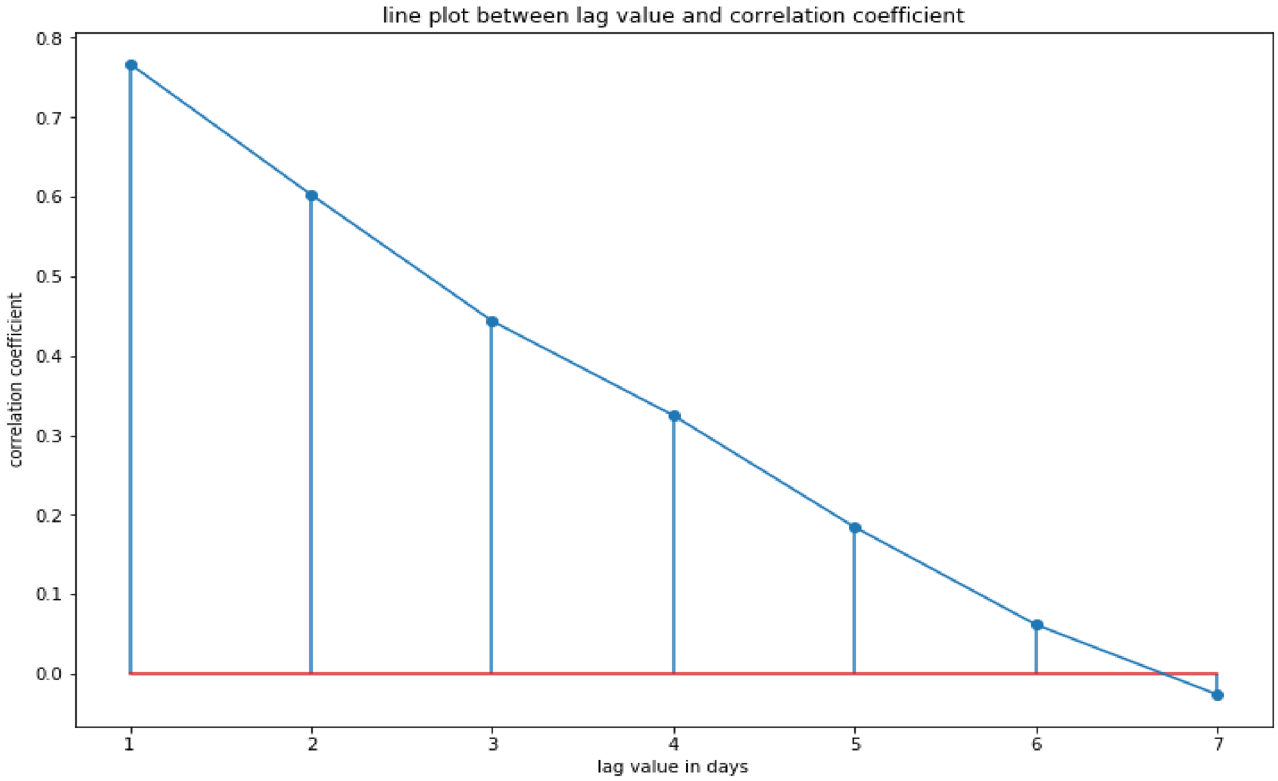

Figure 3 Correlation coefficient vs. lags in given sequence

Inferences:

1. As the number of the lag value increases, the correlation coefficient decreases.
2. We can observe that as the number of the lag value is increasing, the dependency between the lagged time series data and the original data is decreasing.

e.

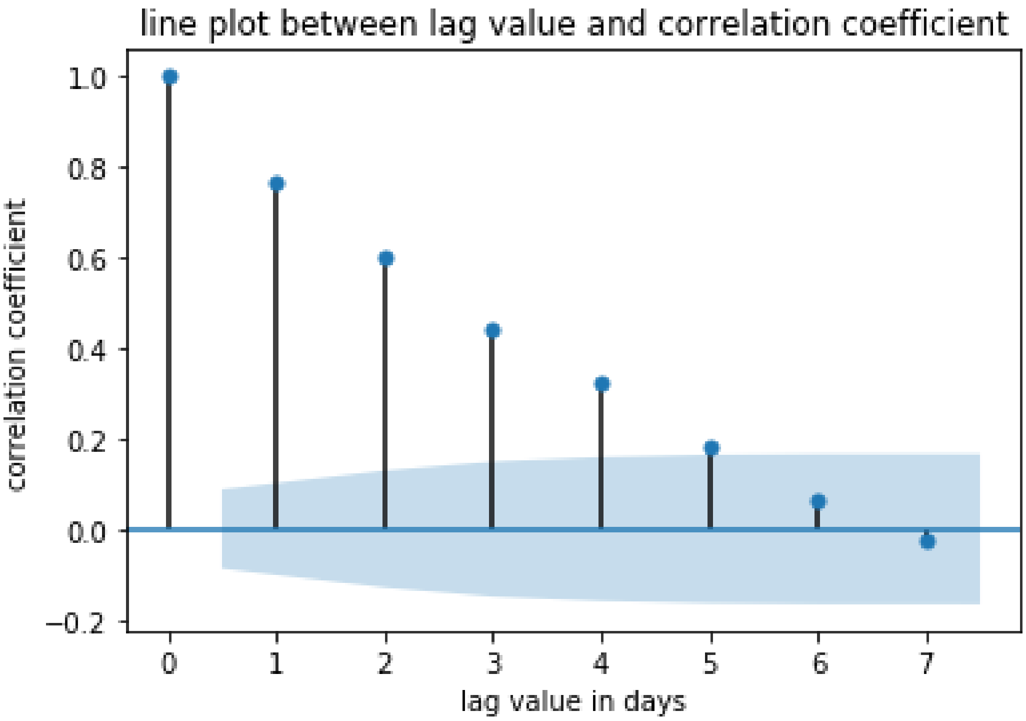

Figure 4 Correlation coefficient vs. lags in given sequence generated using 'plot_acf' function

Inferences:

1. As the number of the lag value increases, the correlation coefficient decreases.
2. We can observe that as the number of the lag value is increasing, the dependency between the lagged time series data and the original data is decreasing.

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 we can say that the accuracy of the persistent model for the given time series data is high.
2. The lag value for the persistent model is 1 and from the above plots we can say that the data of the adjacent values of the time series is almost similar and the correlation coefficient is also high therefore the persistent model gives the accurate result here.

3 a.

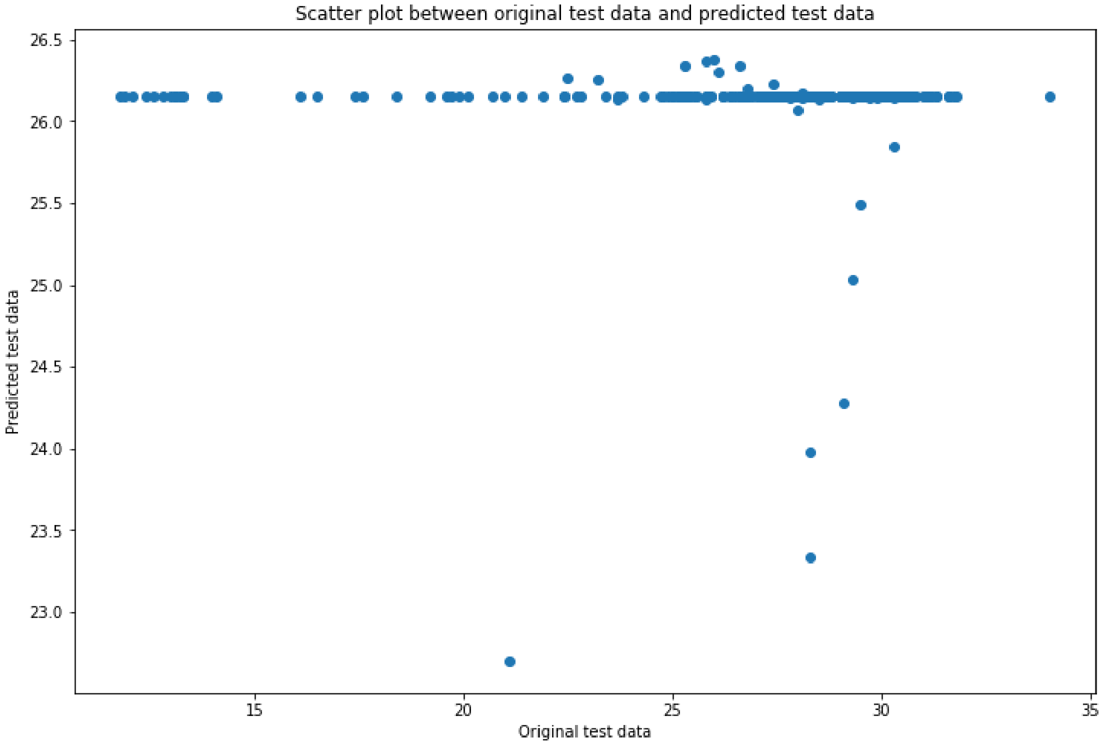

Figure 5 Predicted test data time sequence vs. original test data sequence

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

Inferences:

1. From the value of RMSE value, we can say that the accuracy of the predicted power consumed is moderate.
2. The above RMSE value shows how good the data is fitting in the regression line. The Autoregression for lag value = 5 is giving the linear relationship between the input and the output data which may not be always true.
3. From the above Autoregression model, the predicted data is mostly lying around 26.2 whereas the input data is in the range 10-35. And also the correlation analysis shows that there is no or very low correlation between the original test data and the predicted test data which shows that on increasing the value of the original data, the predicted data is nearly constant. The predicted data is nearly same for any value. So, we can say that this model is not reliable for the future predictions.
4. The RMSE value (3.198) of the persistence model is lower as compared to the RMSE value(4.537) of the Autoregression model(5). So, the accuracy of the persistence model is higher.

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. The RMSE value is nearly same for lag value 1 and 5 but for lag value 10 it decreases. But for lag value 15, it increases and the decreases to its lowest for lag value 25. So, we can say that as the lag value increases, the RMSE value first decreases then increases and then again decreases.
2. The RMSE value depends on the Autoregression till the p lags but here the predicted data is mostly lying around 26, therefore there is no any drastic change in the value of 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. No, the prediction accuracy of the model is not improved by the heuristics approach for calculating the optimal number of lags.
2. It may not be necessary that the heuristics approach is the best way to compute the optimal solution. The heuristics approach is used to get the value upto which the correlation with the original time series is satisfied. It may not be necessary that it computes the value which gives the least RMSE value.

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 RMSE value obtained without heuristics is 4.5141 whereas the RMSE value obtained using the heuristics approach is 4.537. So, the accuracy computed without the heuristics is high.
2. It may not be necessary that the heuristics approach is the best way to compute the optimal solution. The heuristics approach is used to get the value upto which the correlation with the original time series is satisfied. It may not be necessary that it computes the value which gives the least RMSE value.