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

1 a.

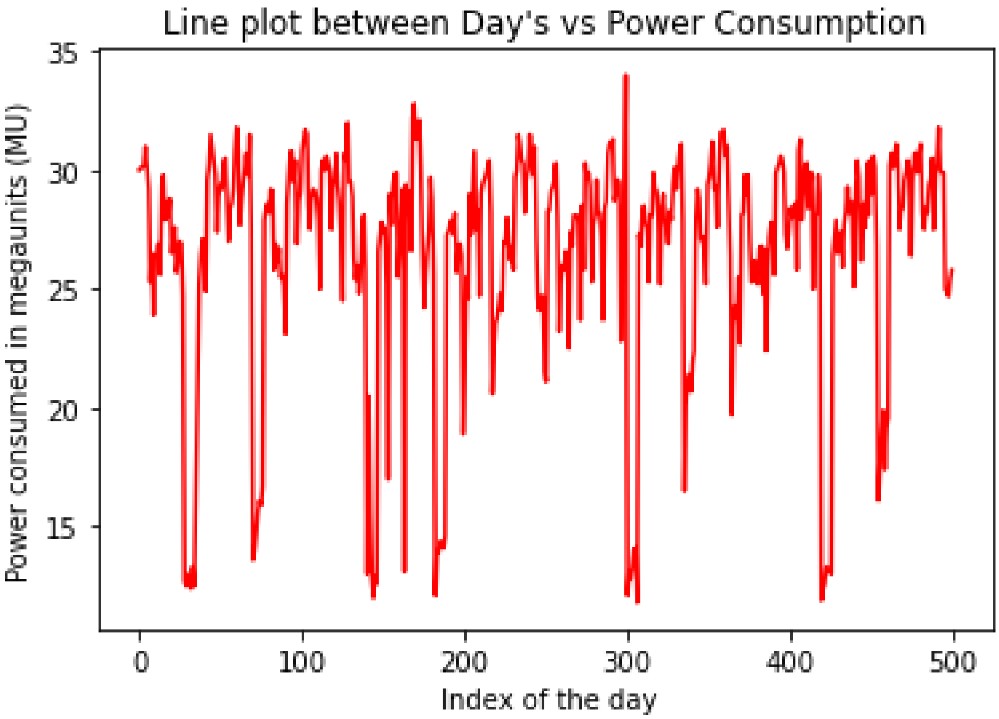

Figure 1 Power consumed (in MW) vs. days

Inferences:

1. Power consumption for most of the days is between 25 MU to 30MU. On some days the Consumption is unusually low.
2. The reason behind the power consumption to remain in a fixed range is that the daily household pretty much uses the same amount of energy each day. The irregularity occurs due to various reasons. But one possibility could be the data is not measured correctly. It may be the noise as well.

b. The value of the Pearson's correlation coefficient is 0.76709.

Inferences:

1. The value of Pearson's correlation is high between the time series and a time series with lag=1.
2. Since the Pearson's correlation is has a pretty decent positive value, we can infer that both the series are related closely that is the Power consumption on a particular day is consistent with the power consumption on the day before that. Since, the Person's correlation is high and positive it also signifies the same that both are closely and more linearly related to each other.
3. Since we expect that the power consumption on a particular day for a household id pretty much the same for most of the days, we can easily explain why the Pearson's correlation is high and positive value.

c.

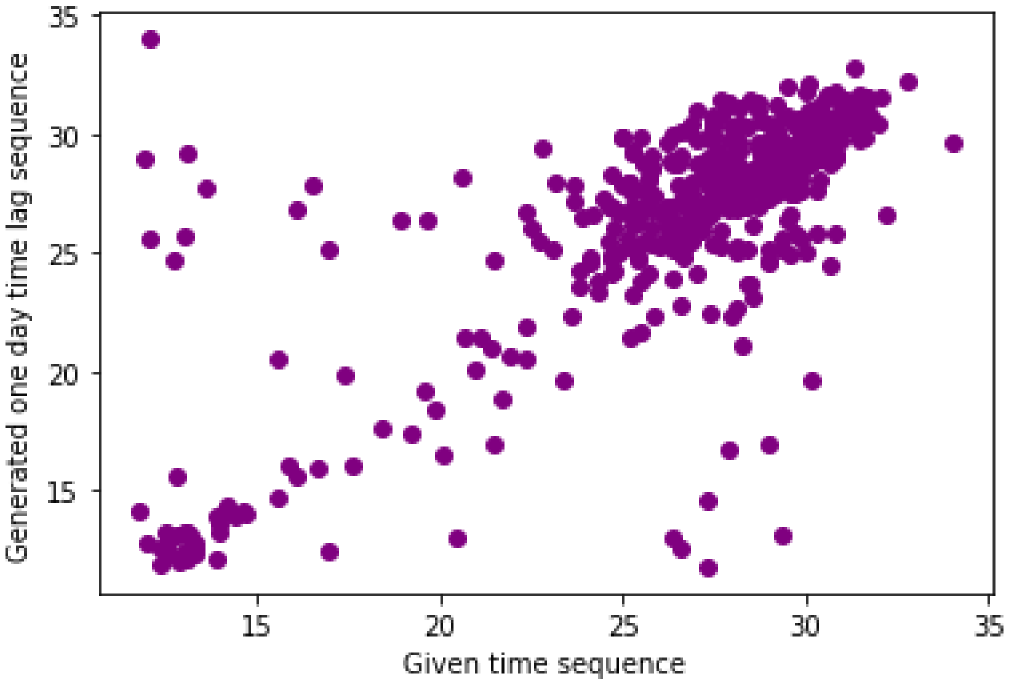

Figure 2 Scatter plot one day lagged sequence vs. given time sequence

Inferences:

1. Since the regression line fits the data pretty closely, we can infer that the Pearson's correlation is high and since the slope is positive, we can also infer that the Correlation will also be a positive value.
2. Yes, the scatter plot seem to obey the nature reflected by Pearson's correlation coefficient calculated in 1(b).
3. As, from the scatter plot (observing the distribution) we can see that the data is close to a linear distribution i.e. if one increases then the other also increases and vie-a-versa. This implies that there is a positive correlation between them and since the line fits the data pretty well, it also suggests that the correlation is high. So, the observation matches the calculation done In previous part.

d.

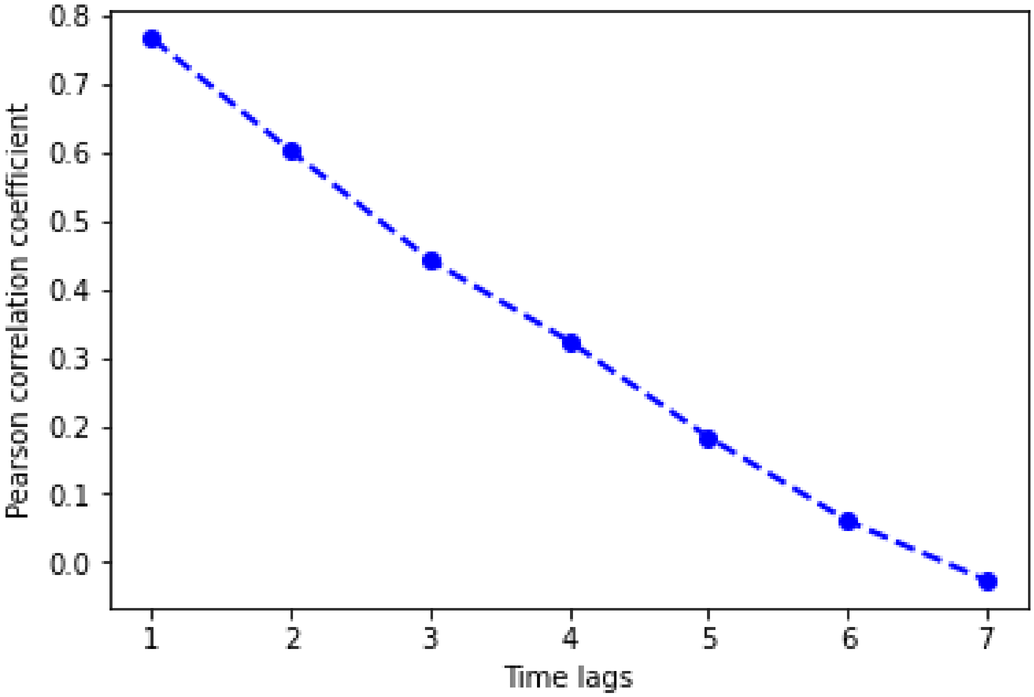

Figure 3 Correlation coefficient vs. lags in given sequence

Inferences:

1. As expected, the correlation coefficient decreases with the increase in time lag.
2. This is due to the fact that. As we go down the time series, the data keeps on becoming less and less related i.e. in a long run due to various external factors we consume different amount of Energies in our houses. A simple example is consumption in summers is far greater than the consumption

winters. Another one could be increase in the Electricity prices which leads to less power consumption in common households.

e.

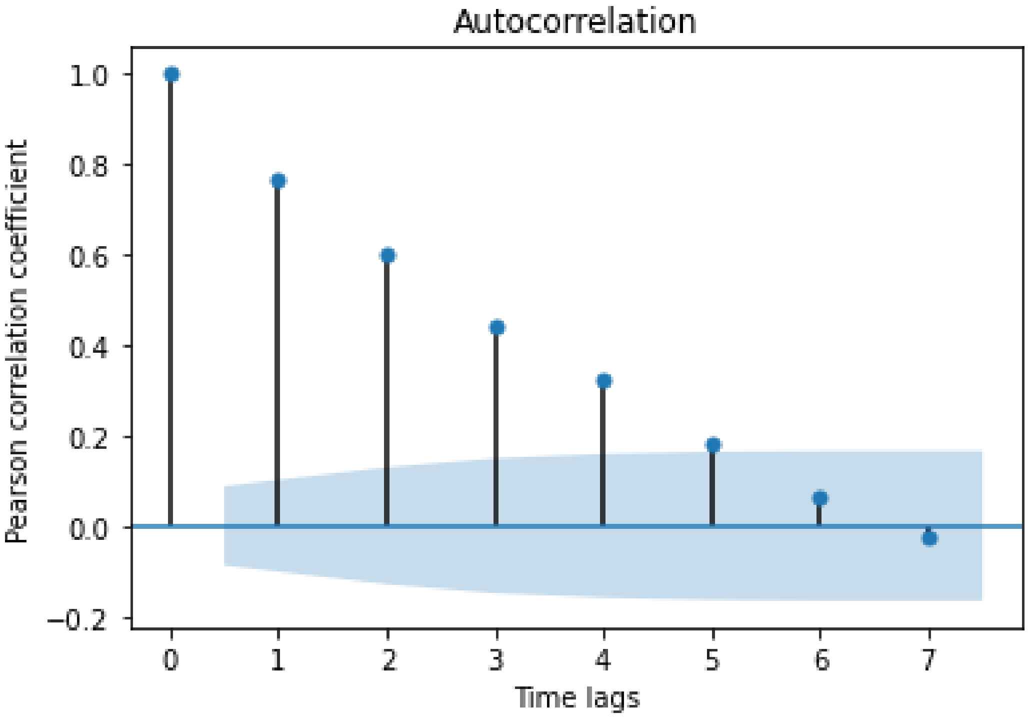

Figure 4 Correlation coefficient vs. lags in given sequence generated using 'plot_acf' function

Inferences:

1. As described in the previous part trend is exactly the same i.e. decrease in correlation with increase in lag width. In fact, the graph obtained from this library also matches the graph we coded.
2. The reason will also be the same as described in previous part i.e. there are some external factors that affect the time series data over a long period and they are not much significant in short period hence, there is greater variability in the time series as the time lag increases

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

Inferences:

1. From the RMSE value of the persistence model we can infer that this model is good. Since the RSE value between the test data and the predicted values is low like 15% if we compare it with the average of the power consumption, we can infer that model is good.
2. In the persistence model we are only considering that the value at a particular time series index id equal to the value at the previous timestamp. Since, we know that for a short period of time the power consumption remains approximately the same, the persistence model performs well.

$$y(t+1) = y(t) \quad , \qquad Persistence\ Model$$

3 a.

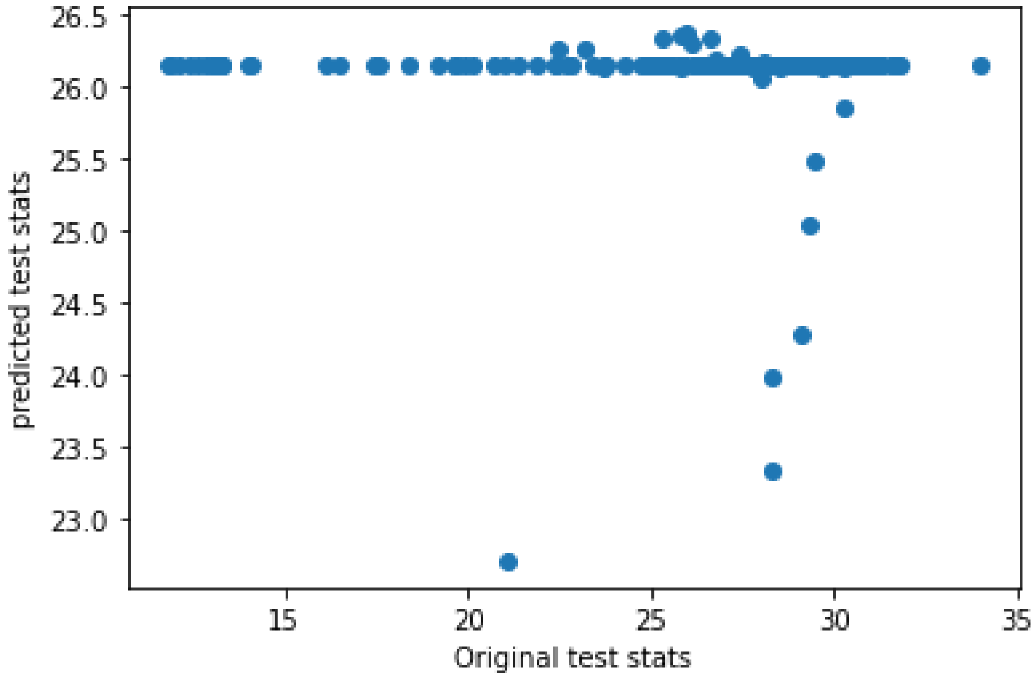

Figure 5 Predicted test data time sequence vs. original test data sequence

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

Inferences:

1. From the RMSE value which when represented as the percentage of average power consumption comes out to be 20%, we can infer that model is not so bad but is not so good either. The point to

note here is RMSE alone cannot be used to tell how good the data is i.e. we need to also look at the distribution of actual v/s predicted data and from the above graph the distribution is not so good.

2. From the plot we can infer that the model is not so good as the output is pretty much the same for any input value. Thus, it will perform badly in predicting future values.
3. By looking at the RMSE values we can say that persistence model is better than the Auto regression model with 5 Time stamps lag. This can also be inferred from the fact that the data becomes less and less relatable as we increase the time lag.

b.

Table 1 RMSE between predicted and original data values w.r.t lags in time sequence

Lag value	RMSE
1	**4.53666**
5	**4.53700**
10	**4.52628**
15	**4.55582**
25	**4.51413**

Inferences:

1. The RMSE values pretty much remains the same.
2. The reason is that our data is pretty much the same as we discussed in previous questions as well. So, no matter how many time-lags we consider the RMSE value doesn't change significantly.

c. The heuristic value for optimal number of lags is 0.126.

The RMSE value between test data time sequence and original test data sequence is 4.537

Inferences:

1. The model performance in terms of accuracy value of the RMSE value doesn't change much. But in terms of fast computation it may increase significantly.
2. Considering the optimal lags without calculation heuristic value to be the lags with minimum RMSE we obtain 10 as the optimal lags and 5 as optimal lags after calculating the heuristic value. From the RMSE value of both the lags we can see that there is not much difference though the value RMSE

value for Lags = 10 is slightly less but still it is not significant. So, we can use the Lags = 5 as it will reduce the computation very much. Hence, model performance will increase.

d.

Considering the optimal Lag Value as the lags with minimum RMSE without calculating the heuristic value.

The optimal number of lags without using heuristics for calculating optimal lag is: 10.

The optimal number of lags using heuristics for calculating optimal lag is: 5.

Inferences:

1. As already discussed, the accuracy in terms of RMSE doesn’t change much i.e. they differ only by a small value and that too is not significant. This implies in terms of accuracy or the RMSE value both the models perform almost the same. But with the Lag value as 5 the number of computations will be reduced to a great extent hence that model performs well in terms of time complexity.