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

Student's Name: Ankit **Mobile No: 9053437219**

Roll Number: B19236 **Branch: Mech**

1 a.

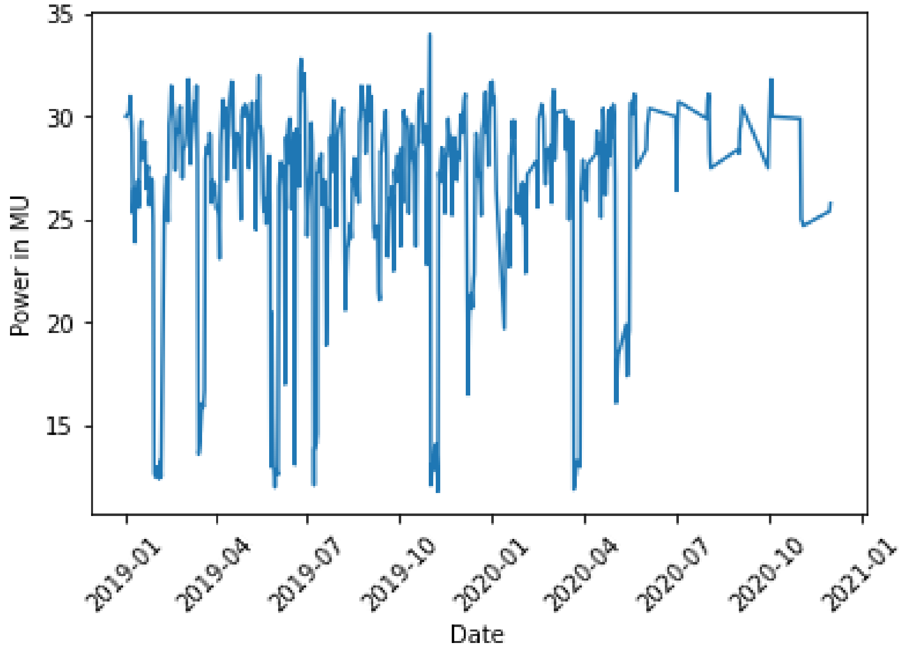

Figure 1 Power consumed (in MW) vs. days

Inferences:

1. **Infer from the plot whether the days one after the other have similar power consumption?**
   Yes, they do.
2. **State the reason behind inference 1.**
   We can see a cyclic dipping pattern in the line plot above.

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

Inferences:

1. **From the value of the Pearson's correlation coefficient, what do you infer about degree of correlation between the two-time sequences?**

   Original time sequence and one-day-lagged sequence have a high correlation of 0.76; they are closely related

2. **We generally expect observations (here power consumption) on days one after the other to be similar. To what extent does it hold true? Answer with respect to the value of Pearson's correlation coefficient.**

   As the Pearson correlation coefficient has a relatively high value, the assumption holds true.

3. **State the reason behind Inferences 1 and 2.**

   This is related to the seasonal behavior 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.

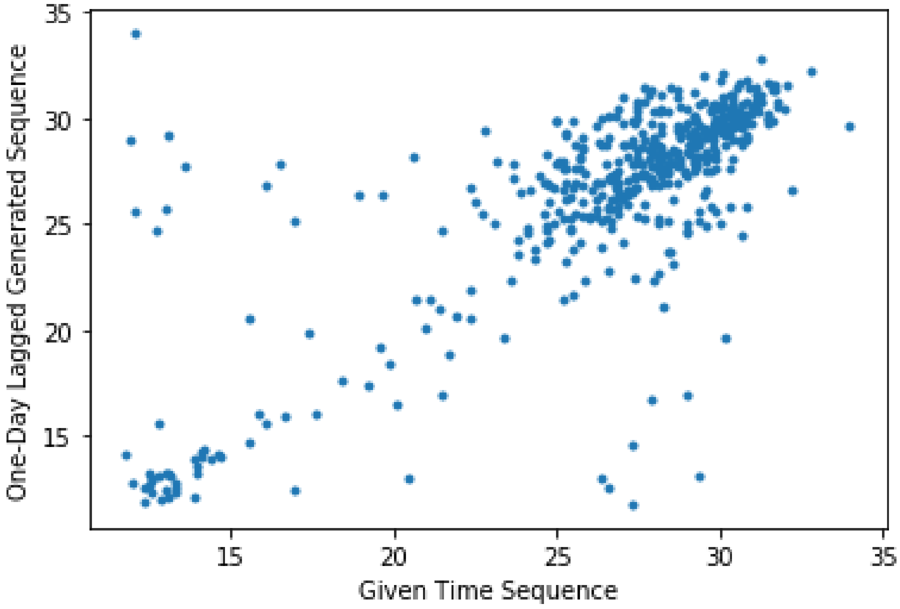

Figure 2 Scatter plot one day lagged sequence vs. given time sequence

Inferences:

1. **From the nature of spread of data points, what do you infer about the nature of correlation between the two sequences?**

   As we see the spread is very linear, which suggests a strong correlation between the variables.

2. **Does the scatter plot seem to obey the nature reflected by Pearson's correlation coefficient calculated in 1.b?**

   Yes, the spread is indicative of a high Pearson correlation.

3. **State reason behind inference 2.**

   A high positive Pearson correlation indicates that we can expect one variable to increase as the other does.

d.

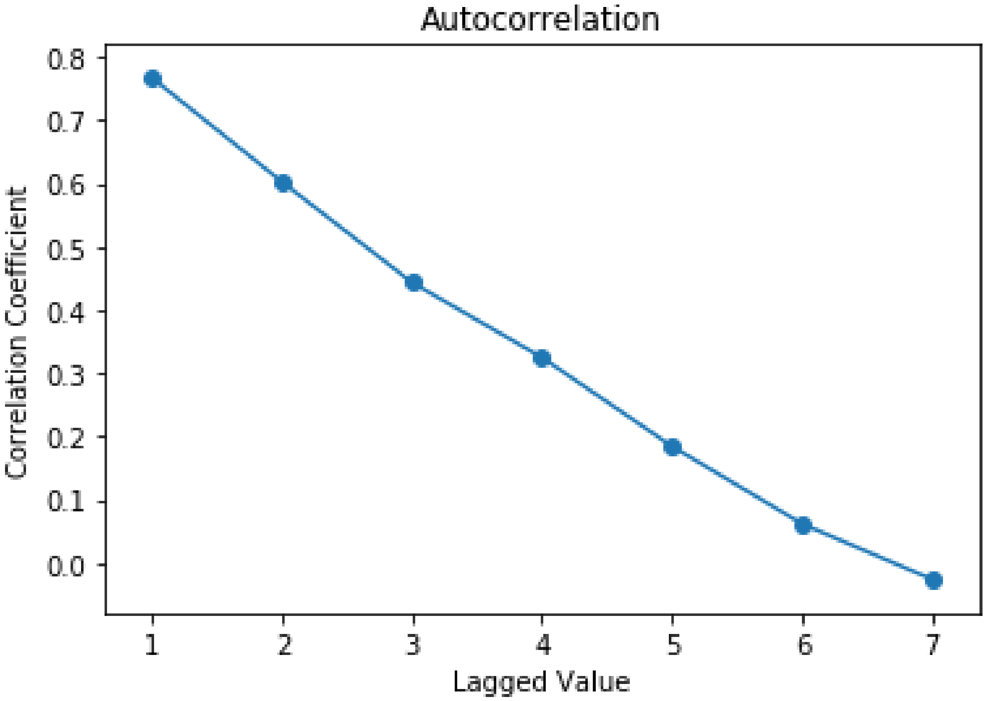

Figure 3 Correlation coefficient vs. lags in given sequence

Inferences:

1. **Infer the trend of correlation coefficient value with respect to increase in lags in time sequence.**
   As the number of lags increase, the coefficient values decrease.
2. **Explain the reason behind the observed trend.**
   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.

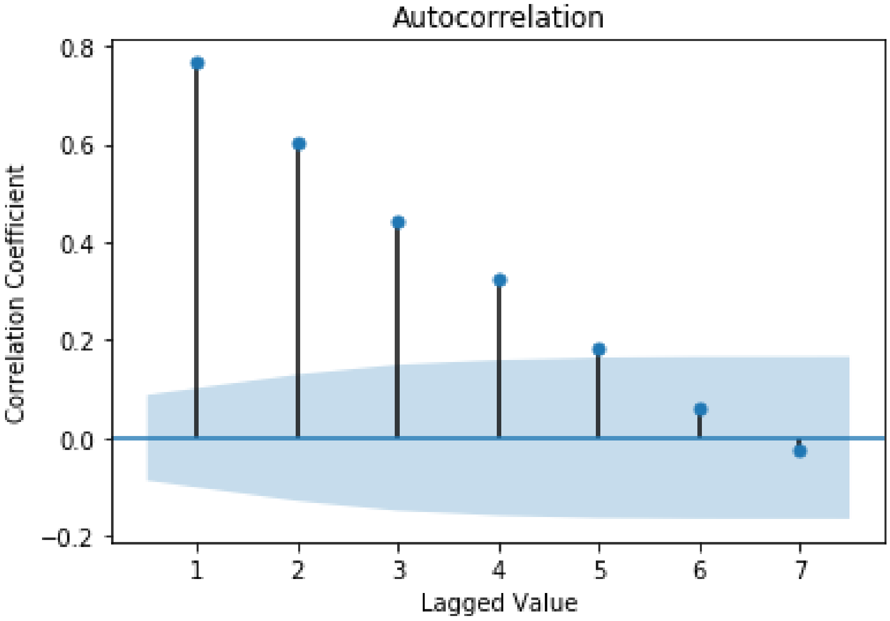

Figure 4 Correlation coefficient vs. lags in given sequence generated using 'plot_acf' function

Inferences:

1. **Infer the trend of correlation coefficient value with respect to lags in time sequence.**
   As the number of lags increase, the coefficient values decrease.
2. **Explain the reason behind the observed trend.**
   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. **From the value of RMSE value comment how accurate is persistent model for the given time series.**
   It is reasonably accurate, since it's giving an error of approx. 12% relative to the average data point in the series
2. **State the reason behind Inference 1.**
   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.**

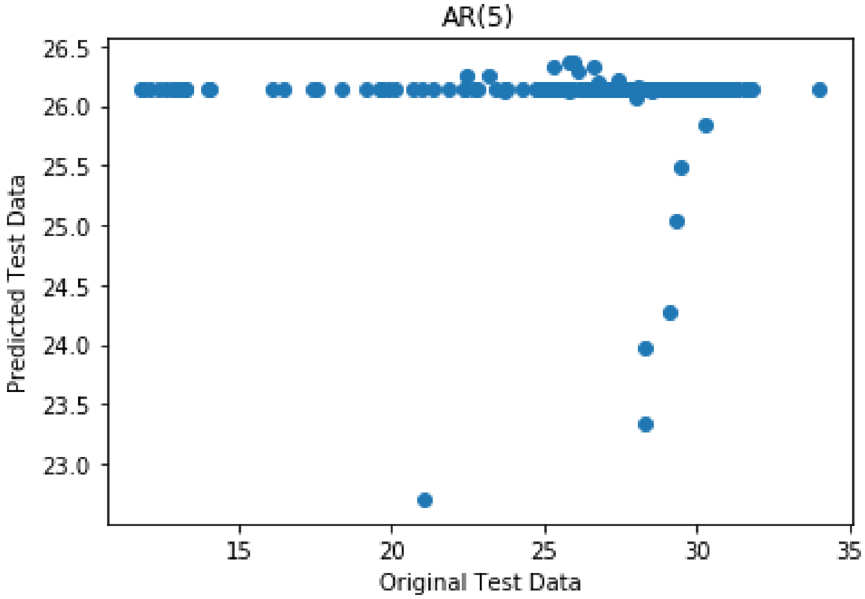

Figure 5 Predicted test data time sequence vs. original test data sequence

The RMSE between predicted power consumed for test data and original values for test data is **4.5378**

Inferences:

1. **From the value of RMSE value comment how accurate is the model for the given time series.**
   It is less accurate.
2. **State the reason behind Inference 1.**
   For this model, RMSE is quite significant with respect to the data.
3. **From the plot of predicted test data time sequence vs. original test data sequence comment how reliable is the model for future predictions with suitable reasons.**
   It would be preferred to use the persistence model in favor 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. **Based on RMSE value, compare the accuracy between the current model and model used in question 2.**
   As we see, RMSE for persistence model is lesser, hence it has better accuracy than 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	**4.532**
15	**4.56**
25	**4.515**

Inferences:

1. **Infer the trend of RMSE with respect to increase in lags in time sequence.**
   RMSE is decreasing.
2. **State the reason behind Inference 1.**
   RMSE decreases because we continue to take weighted input from data further back in time series, which improves our predictive model.

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, comment did use heuristics for calculating optimal number of lags improve the prediction accuracy of the model?**
   Using heuristic value certainly helps get a good estimate of the optimal number of lags for minimal RMSE.
2. **State the reason behind Inference 1.**
   As we know, optimal lag value is determined by: $C > 2/\sqrt{T}$, where C is correlation and T is length of data set. All lags with correlations following this rule are considered for the model. By help of this formula, we can conclude that heuristic value is helpful in determining optimal lag. I have no clue how this formula is derived though.

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. **Compare the prediction accuracies obtained without and with heuristic for calculating optimal lag with respect to RMSE values.**
   Prediction accuracy is higher in non-heuristic lag value.
2. **State the reason behind Inference 1.**
   $Accuracy\ \alpha\ 1/RMSE$, as we see RMSE is lower for non-heuristic lag value of p=25.