Name: Anooshka Bajaj

1 a.

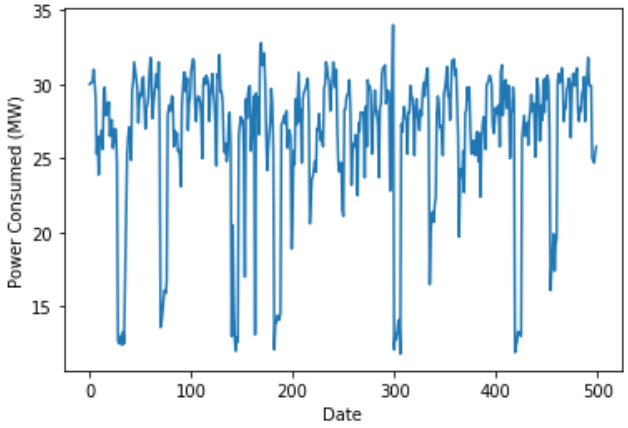

Figure 1 Power consumed (in MW) vs. days

Inferences:

1. Yes, the days one after the other have similar power consumption barring few days (it can be noise).
2. It is because the daily household pretty much consumes the same amount of power each day.

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

Inferences:

1. The degree of correlation between the two time sequences is high.
2. Since the Pearson's correlation is a close to 1 positive value, we can infer that both the series are related closely i.e. the Power consumption on a particular day is consistent with the power consumption on the day before that.
3. Value of Pearson's correlation is close to 1 which means data is strongly correlated. Its because power consumption on a particular day is almost same for most of the days.

c.

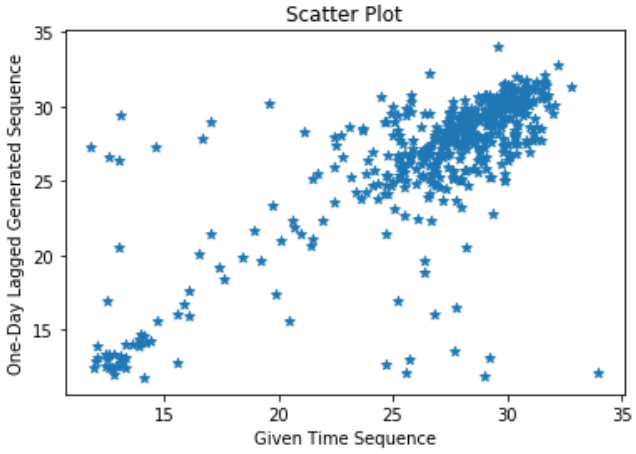

Figure 2 Scatter plot one day lagged sequence vs. given time sequence

Inferences:

1. From the nature of spread of data points, it can be inferred that data is positively correlated and the value of correlation coefficient is quite high.
2. Yes, the scatter plot seems to obey the nature reflected by Pearson's correlation coefficient calculated in 1.b.
3. There is a positive correlation and the correlation is also quite high (as it is close to linear distribution).

d.

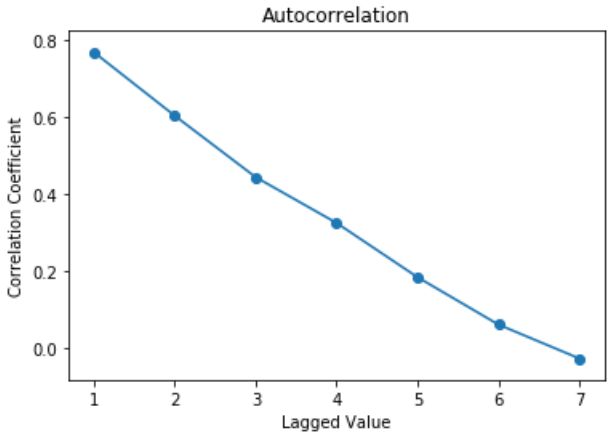

Figure 3 Correlation coefficient vs. lags in given sequence

1. As the value of lag increases, correlation coefficient decreases.
2. This means that the power consumption is more related to just its previous values rather than values more days back.

e.

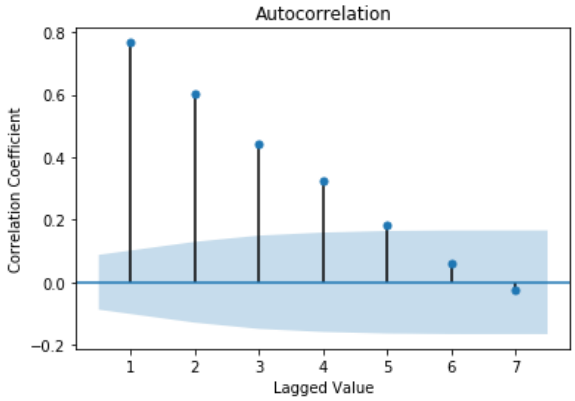

Figure 4 Correlation coefficient vs. lags in given sequence generated using 'plot_acf' function

Inferences:

1. As the value of lag increases, correlation coefficient decreases.
2. Power consumption is more related to just its previous values rather than values more days back.

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

Inferences:

1. The value of RMSE is low meaning that the model is quite good for this particular data.
2. In the persistence model we are only considering that the value at a particular time series index is equal to the value at the previous timestamp. For a short period of time the power consumption remains approximately the same thus, the persistence model performs well.

3 a.

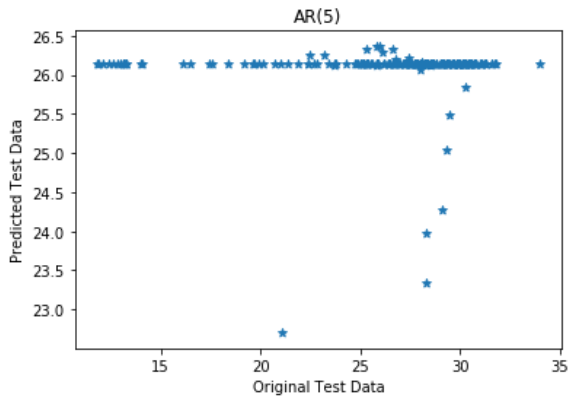

Figure 5 Predicted test data time sequence vs. original test data sequence

The RMSE between predicted power consumed for test data and original values for test data is: 4.5377

Inferences:

1. From the value of RMSE value we can see that the model is not very accurate. It has decent accuracy.
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. As its RMSE value is greater than that of question 2, this means that this model is less accurate than persistence model.

b.

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

Lag value	RMSE
1	**4.539**
5	**4.537**
10	**4.531**
15	**4.559**
25	**4.514**

Inferences:

1. RMSE values are pretty much same with a small amount of difference. Sometimes, it increase and sometimes it decrease. Thus, no particular trend of RMSE with respect to increase in lags in time sequence.
2. For different values of Lag, correlation is different and hence RMSE error has no specific trend.

c. The heuristic value for optimal number of lags is 5 lags.

The RMSE value between test data time sequence and original test data sequence is 4.5377

Inferences:

1. The RMSE value is pretty much same.
2. Without using heuristics the lags with minimum RMSE is 25 (optimal lag). In this case we use 5 as optimal lag after calculating the heuristic value. From the RMSE value of both the lags

we can see that there is not much. But to reduce computation Lags = 5 should be used. Hence, model performance will increase.

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. Without heuristics: RMSE = 4.514
   With heuristics: RMSE = 4.537
2. The accuracy in terms of RMSE is less when not using heuristics. But 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.