Data Science Project
Sub-Project – VI
Auto-regression

Student's Name: Anmol Bishnoi **Mobile No: 7042845211**

Roll Number: B19069 **Branch: CSE**

1 a.

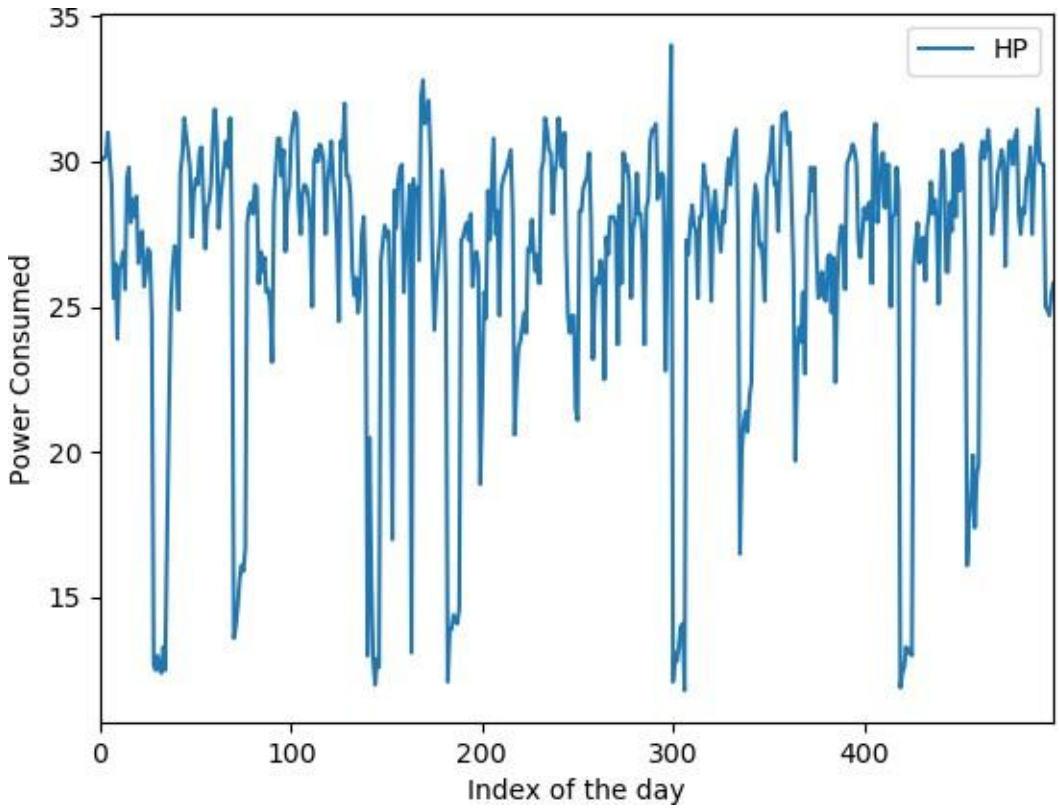

Figure 1 Power consumed (in MW) vs. days

Inferences:

1. Yes, the days one after the another, do have similar power consumptions except the few points where the power consumption drops massively.

2. It is clearly evident from the graph, that the power consumption more or less sees an increasing trend until it drops massively and then eventually rises to the same mean level and yet again continuing to show the increasing trend.

b. The value of the Pearson's correlation coefficient is **0.768**

Inferences:

1. Since the Pearson's coefficient lies between 0.5 and 1, we can say that the two time sequences have a high degree of correlation.
2. A high degree of correlation corresponding to the value of Pearson's coefficient that we got means that the power consumption on days one after the another follow a very similar pattern.

c.

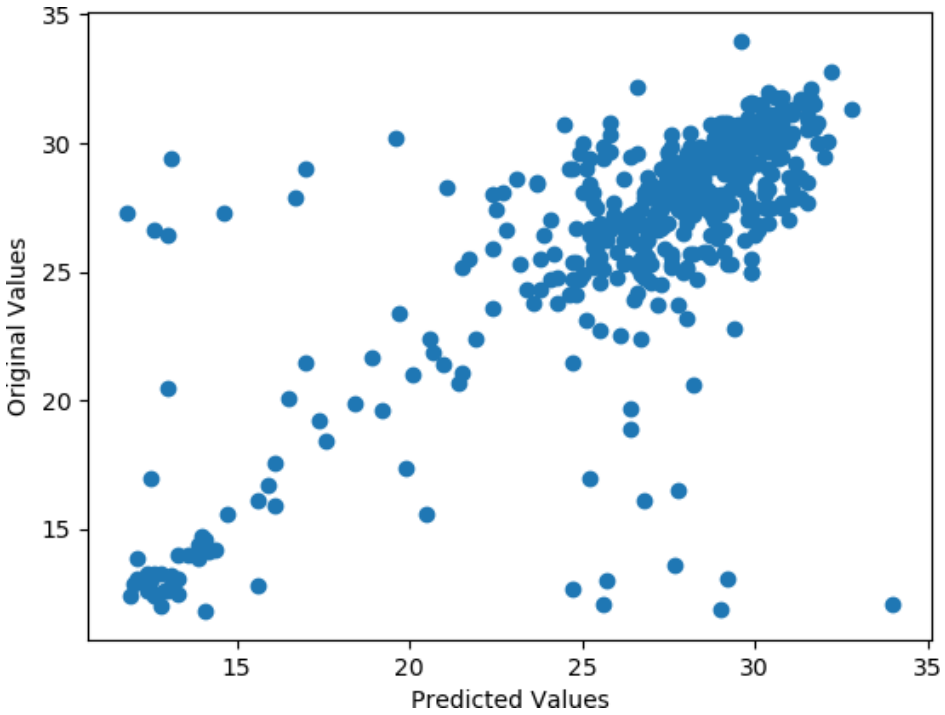

Figure 2 Scatter plot one day lagged sequence vs. given time sequence

Inferences:

1. Since most points lie on and around the straight x=y, we can safely say that the two sequences have a very high degree of correlation.
2. The scatter plot also obeys the nature reflected by Pearson's correlation coefficient calculated in 1.b
3. We can say so because the greater the Pearson's coefficient, the more will be the degree of correlation, and the points of the scatter plot will resemble a straight line to a greater extent.

d.

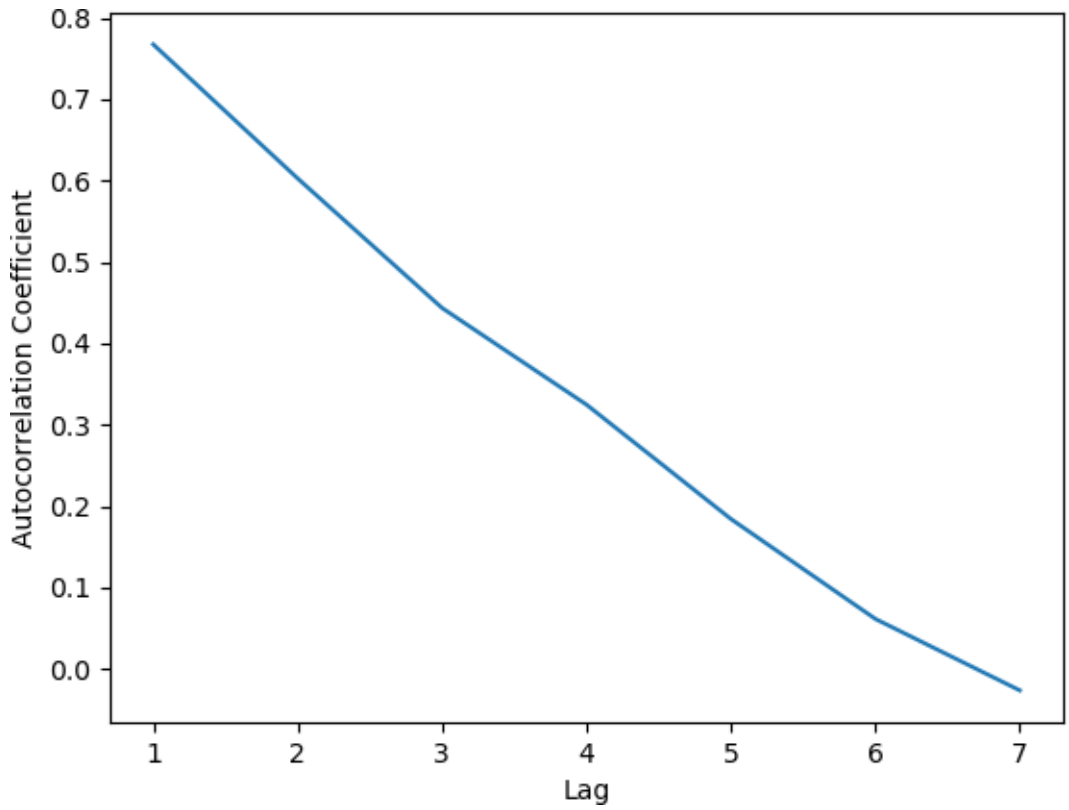

Figure 3 Correlation coefficient vs. lags in given sequence

Inferences:

1. The correlation coefficient decreases as the amount of lag is increased.
2. This happens because the values at any given time might be similar to the values at the time just after it, but as the amount of time between the observations is increased, the values need not be similar and instead become more and more dissimilar.

e.

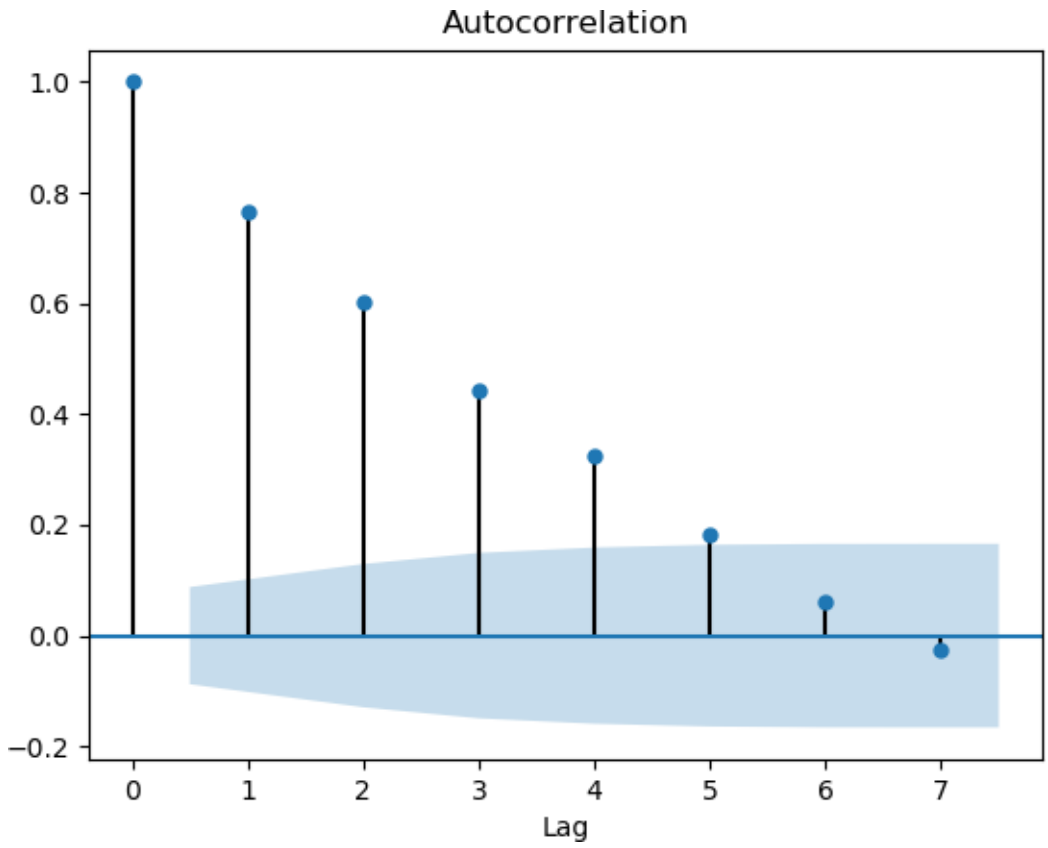

Figure 4 Correlation coefficient vs. lags in given sequence generated using 'plot_acf' function

Inferences:

1. The correlation coefficient decreases as the amount of lag is increased.
2. This happens because the values at any given time might be similar to the values at the time just after it, but as the amount of time between the observations is increased, the values need not be similar and instead become more and more dissimilar.

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

Inferences:

1. Hence, the persistence model turns out to be pretty accurate in predicting the future values in this case.
2. We say so because RMSE tells us the average amount of difference between the true value and the predicted value and an RMSE of 3.198 is very good.

3 a.

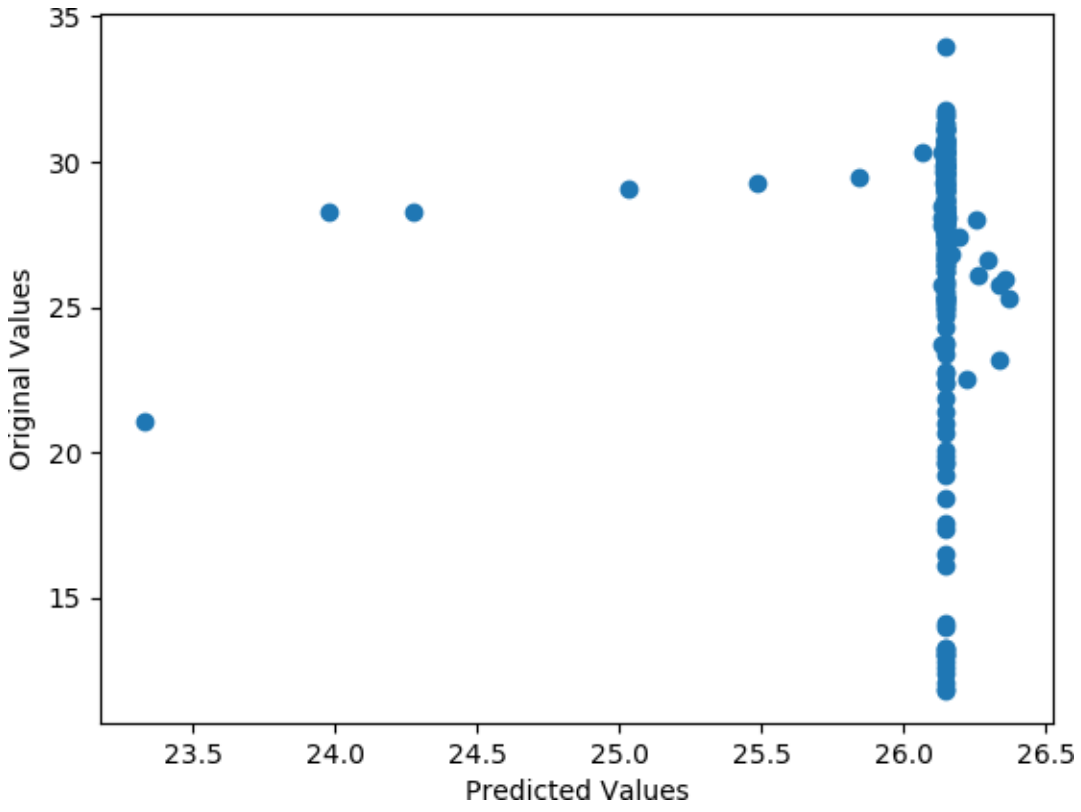

Figure 5 Predicted test data time sequence vs. original test data sequence

The RMSE between predicted power consumed for test data and original values for test data is **4.528**

Inferences:

1. Hence, this model is quite accurate in predicting the future values in this case.
2. We say so because RMSE tells us the average amount of difference between the true value and the predicted value and an RMSE of 4.528 is very good.
3. The model is not at all reliable for making future predictions, since the model is making all future predictions based on the values at previous timestamps that the model itself predicted and this value has already converged. Hence, the model will give the same value for every timestamp in the future.
4. On the basis of RMSE values, the persistence model turns out to be more accurate when compared to the AutoReg model.

b.

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

Lag value	RMSE
1	4.532
5	4.528
10	4.523
15	4.553
25	4.511

Inferences:

1. The RMSE first decreases and then starts to increase again.
2. Adding values with bigger lags that have no correlation with the value that we need to predict to the AutoReg model makes it more and more inaccurate.

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 **4.528**

Inferences:

1. Using heuristics did save us a lot of time while building the auto regression model in deciding how many time lags to consider which we would have tried to find out by manually comparing the RMSE at different numbers and then finding the least one.
2. This is because using heuristics, we were able to directly determine what all values of lag were relevant with a big enough correlation to the points that we needed to predict the values of.

d.

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. The prediction accuracy at a lag value of 10 was actually slightly greater than the model built with a lag value of 5
2. The difference between the RMSE values at the two different lags was less than 0.005 and this can be simply accounted to the randomness in the data.