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

Student's Name: Sreesha Pulipati **Mobile No: 8639196385**

Roll Number: B20119 **Branch:CSE**

1 a.

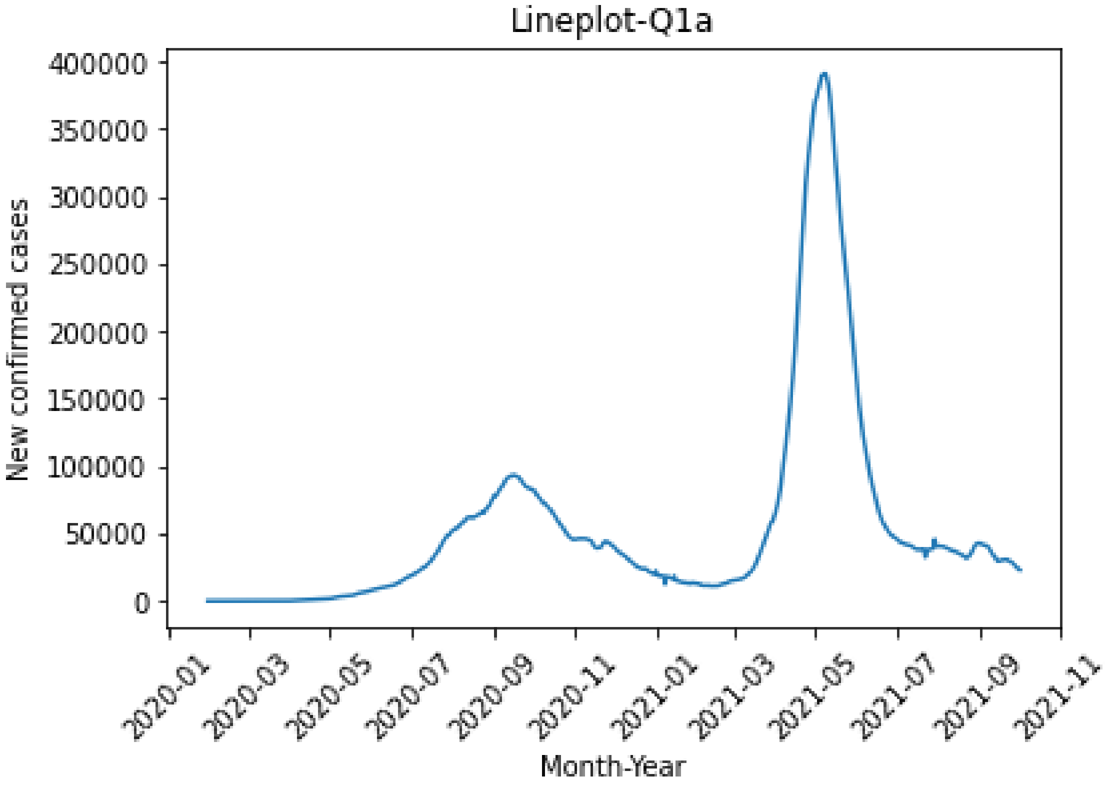


Figure 1 No. of COVID-19 cases vs. days

Inferences:

1. From the plot, we can infer that the cases increase day by day after some time the cases decrease and then again the cases increase.
2. The reason for such plot is because of the increase in cases due to the spread of the disease and then decreases due to some isolation in society and then again increases due to more spread of the disease.

3. The duration of the first wave is approximately 5 months and the second wave is approximately 4 months.

b. The value of the Pearson’s correlation coefficient is 0.999.

Inferences:

1. From the value of Pearson’s correlation coefficient, we can say that both the data sequences are strongly correlated to each other.
2. We generally expect observations (here number of COVID-19 cases) on days one after the other to be similar as they are highly correlated to each other so there is not much difference between the one day after the other.
3. If there are more cases then the spread will be high hence many people will be affected and this leads to increase in cases day by day , and when they are less cases , spread will be low and less cases appears. Hence there is high correlation between given sequence and one-day lagged sequence.

c.

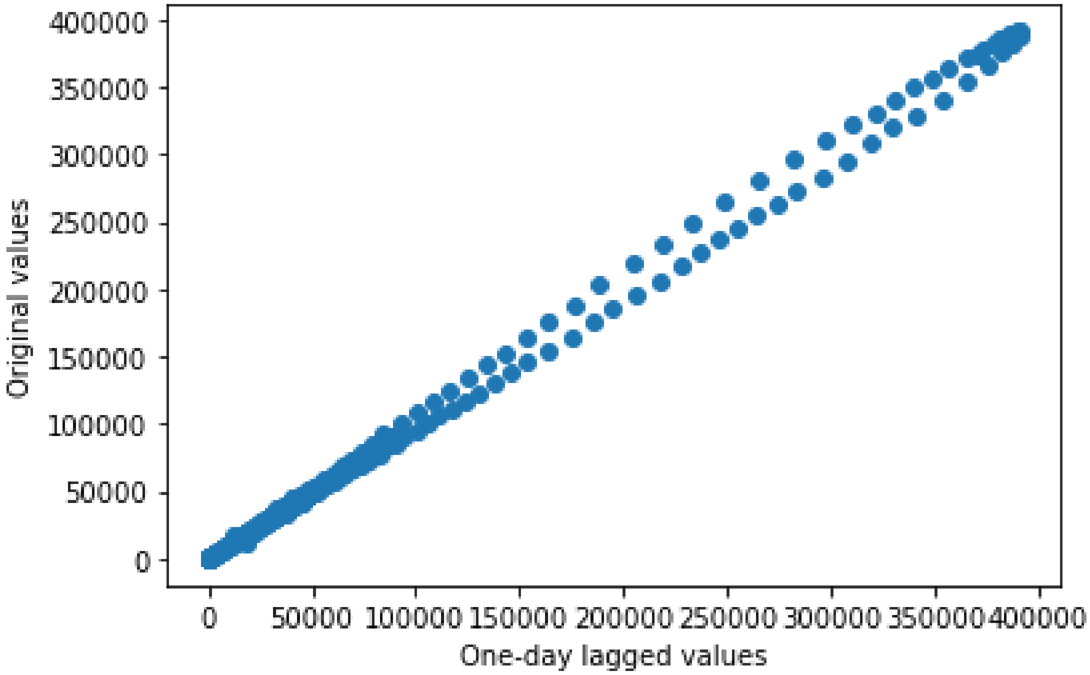
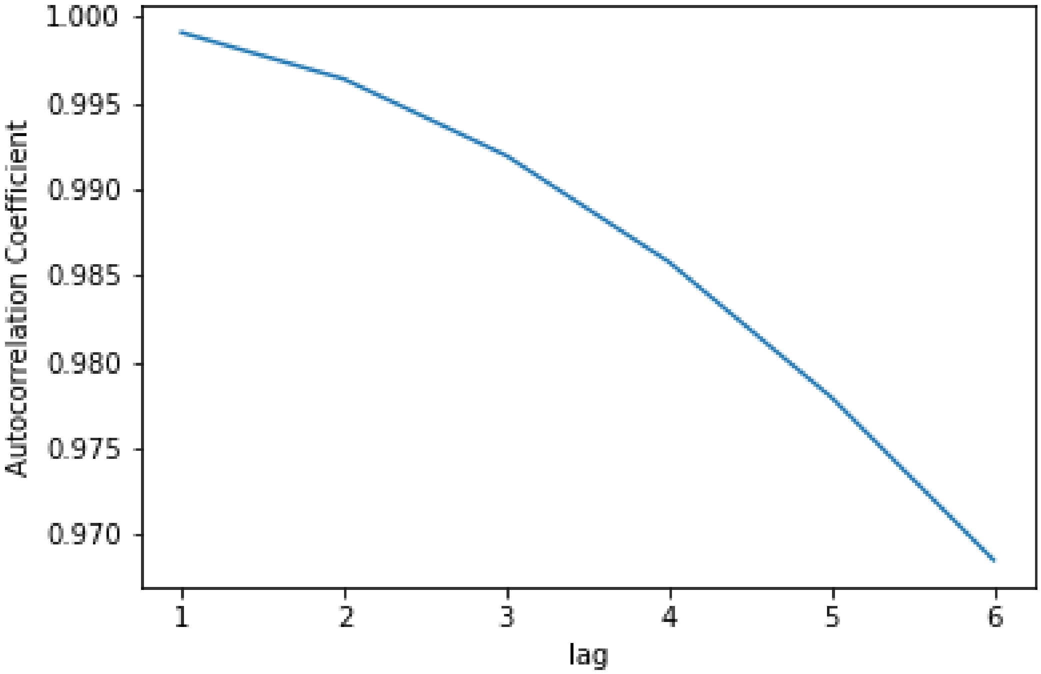


Figure 2 Scatter plot one day lagged sequence vs. given time sequence

Inferences:

1. From the nature of the spread of data points, we can infer that there is high correlation.
2. Yes the scatter plot seem to obey the nature reflected by Pearson's correlation coefficient calculated in 1.b.
3. From the plot we can observe that the plot is strictly increasing, as x values are increasing y values are increasing and this shows that there is high correlation and we can confirm this from coefficient value from 1b

d.



Inferences:

1. The correlation coefficient value decreases with respect to increase in lags in time sequence.
2. The correlation decreases with increases in lag series as the dependency on the past values decreases.

e.

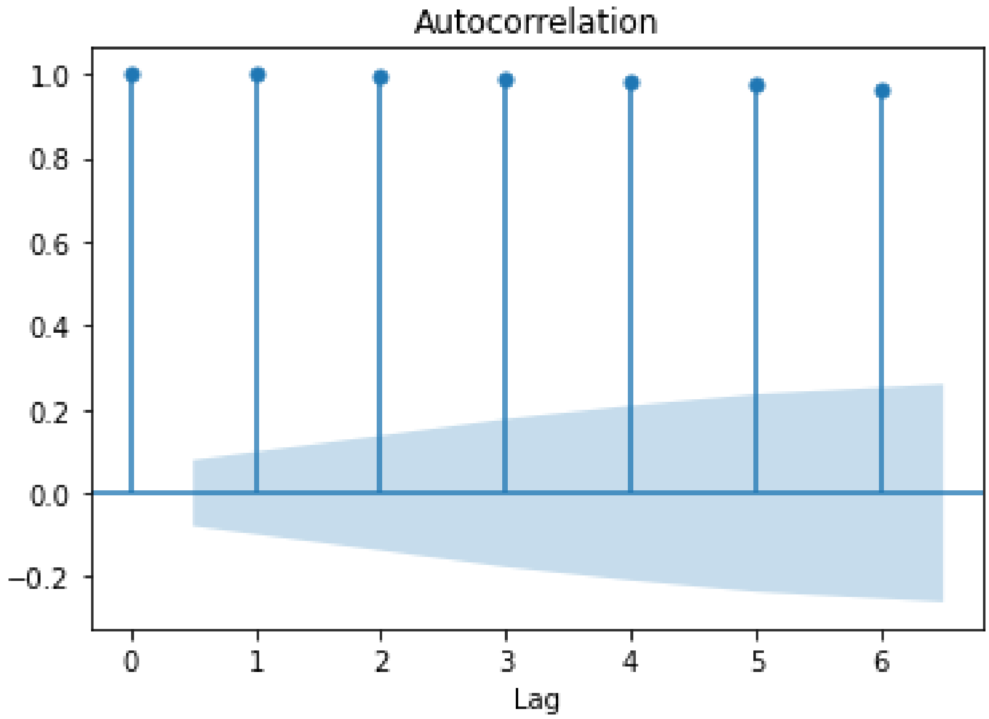


Figure 3 Correlation coefficient vs. lags in given sequence generated using 'plot_acf' function

Inferences:

1. The correlation coefficient value decreases with respect to lags in time sequence.
2. The correlation decreases with increases in lag series as the dependency on the past values decreases.

2

a. The coefficients obtained from the AR model are 599.55, 1.037, 0.262, 0.028 ,-0.175 ,-0.152

b. i.

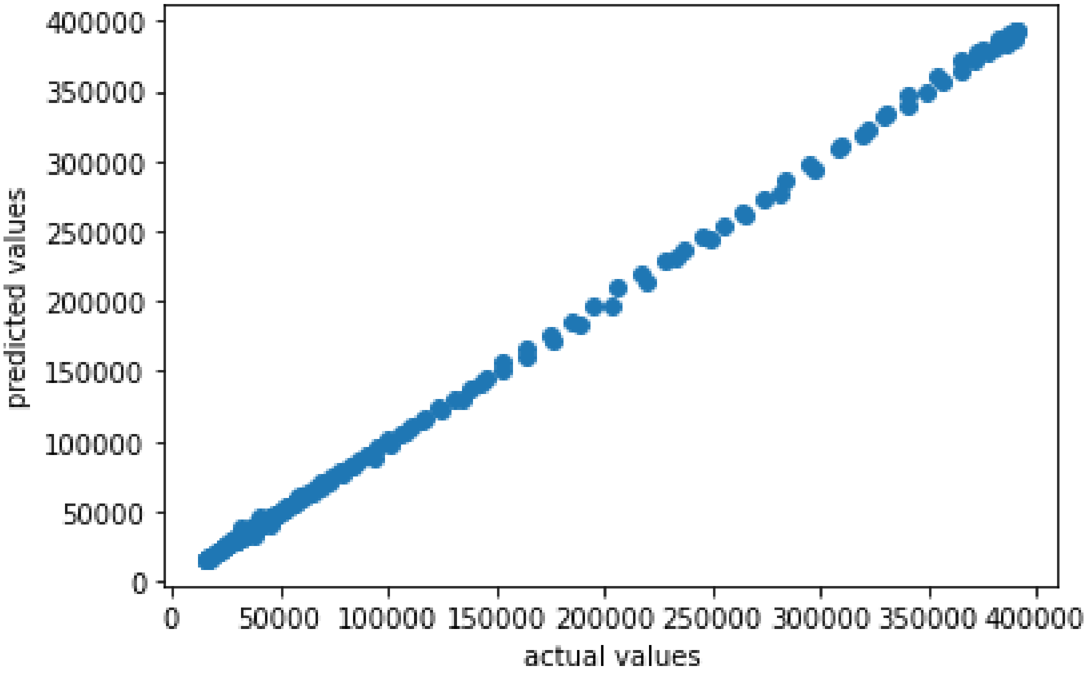


Figure 4 Scatter plot actual vs. predicted values

Inferences:

1. From the nature of the spread of data points, the nature of the correlation between the two sequences is strong.
2. Yes, the scatter plot seems to obey the nature reflected by Pearson's correlation coefficient calculated in 1.b.
3. Predicted values are calculated based on the past values. Graph shows that predicted values and actual values are highly correlated hence present values depend on the past values and we can confirm this from the correlation obtained in 1b.

ii.

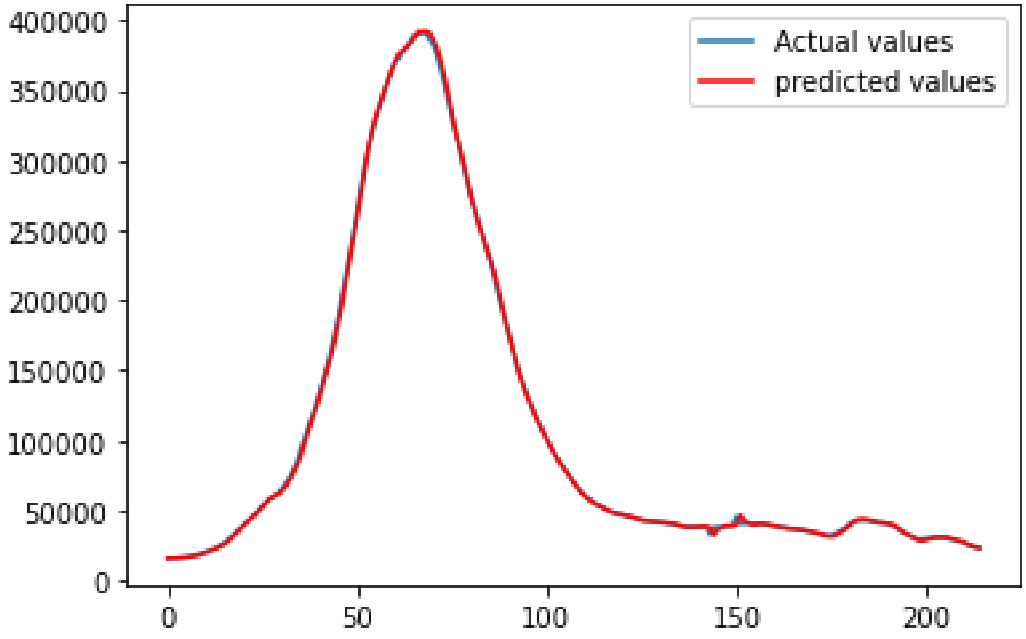


Figure 5 Predicted test data time sequence vs. original test data sequence

Inferences:

1. From the plot of predicted test data time sequence vs. original test data sequence , we can observe that both sequences are highly correlated hence the model is highly reliable for future predictions.

iii.

The RMSE(\%) and MAPE between predicted power consumed for test data and original values for test data are 1.825 and 1.575(in percentage).

Inferences:

1. From the value of RMSE(\%) and MAPE value the model for the given time series is very accurate as the error is less.
2. As the error is low we can use the data for future prediction as there is very less deviation between the predicted and original values.

3

Table 1 RMSE (%) and MAPE between predicted and original data values wrt lags in time sequence

Lag value	RMSE (%)	MAPE in percentage
1	5.373	3.447
5	1.825	1.575
10	1.686	1.519
15	1.612	1.496
25	1.703	1.535

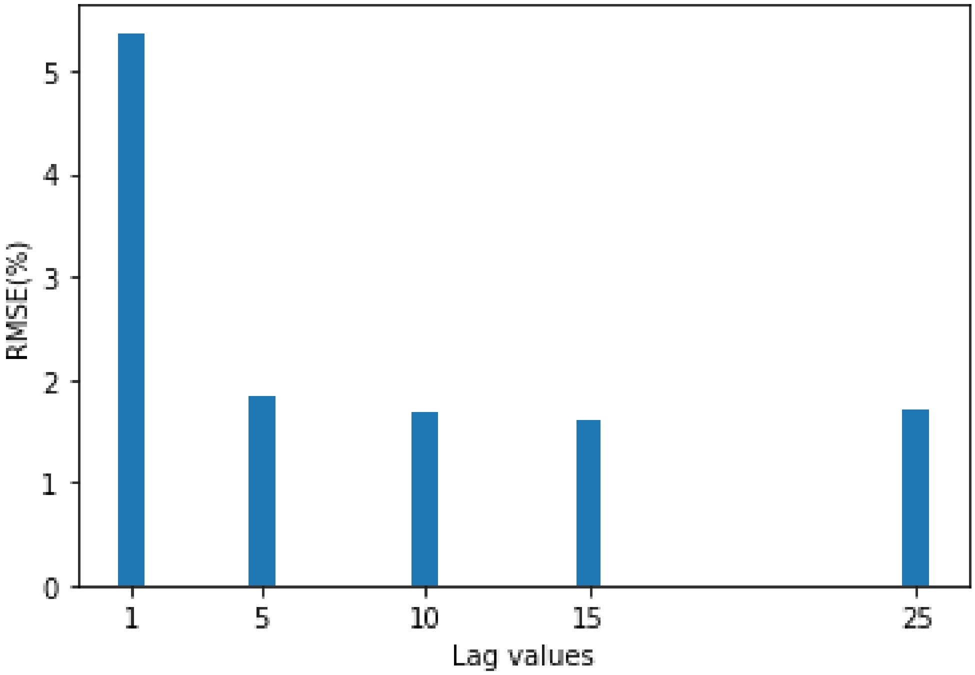


Figure 6 RMSE(%) vs. time lag

Inferences:

1. From the above plot the RMSE(%) decreases with respect to increase in lags in time sequence. But RMSE(%) increases with high lag values.
2. As the time lag increases the dependency of present value increases on past values so that the RMSE decreases but as lag values are much high there will be decrease in dependency and this leads to increase in RMSE Values.

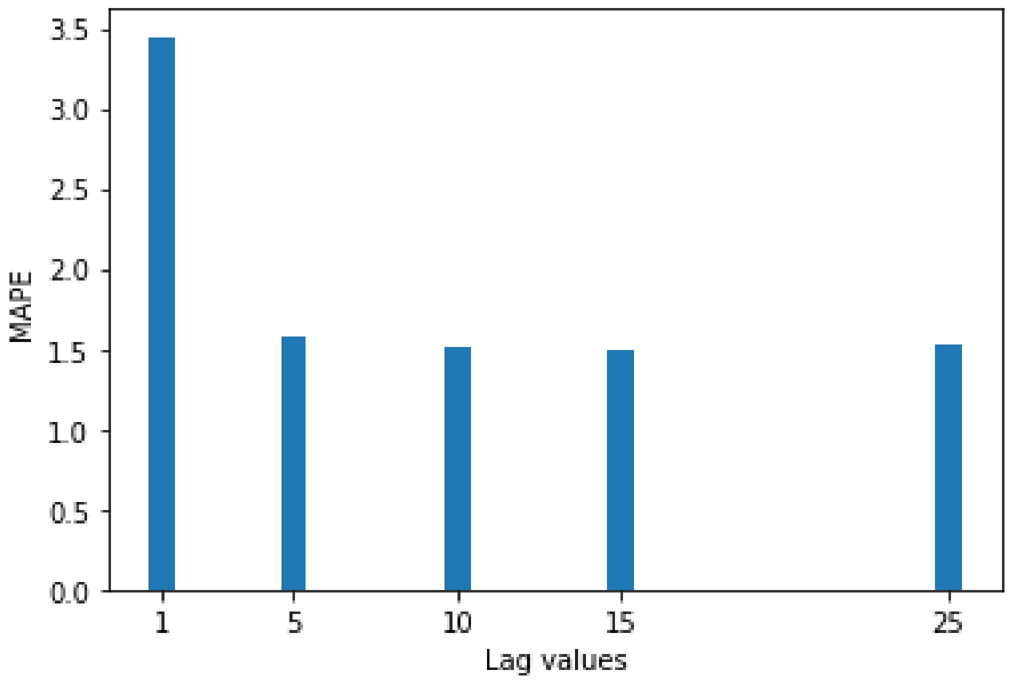


Figure 7 MAPE vs. time lag

Inferences:

1. From the above plot the MAPE decreases with respect to increase in lags in time sequence. But MAPE increases with high lag values.
2. As the time lag increases the dependency of present value increases on past values so that the MAPE decreases but as lag values are much high there will be decrease in dependency and this leads to increase in MAPE Values.

4

The heuristic value for the optimal number of lags is 77

The RMSE(%) and MAPE value between test data time sequence and original test data sequence are 1.759 and 2.026(in percentage).

Inferences:

1. Yes, based upon the RMSE(%) and MAPE value, heuristics for calculating the optimal number of lags improved the prediction accuracy of the model.
2. The RMSE with optimal lag is around 5.3-1.6 range where our calculated RMSE is 1.759 which is within the range and the MAPE value with optimal lag is around 3.446-1.496 where as heuristic MAPE is 2.206 which is also in the range of MAPE.
3. The prediction accuracies obtained with the heuristic for calculating optimal lag with respect to RMSE(%) and MAPE values is less compared to that obtained without heuristic value.