IC 272: DATA SCIENCE - III

LAB ASSIGNMENT – VI

Auto-regression

Student's Name: Amit Maindola **Mobile No: +91 7470985613**

Roll Number: B20079 **Branch: CSE**

1 a.

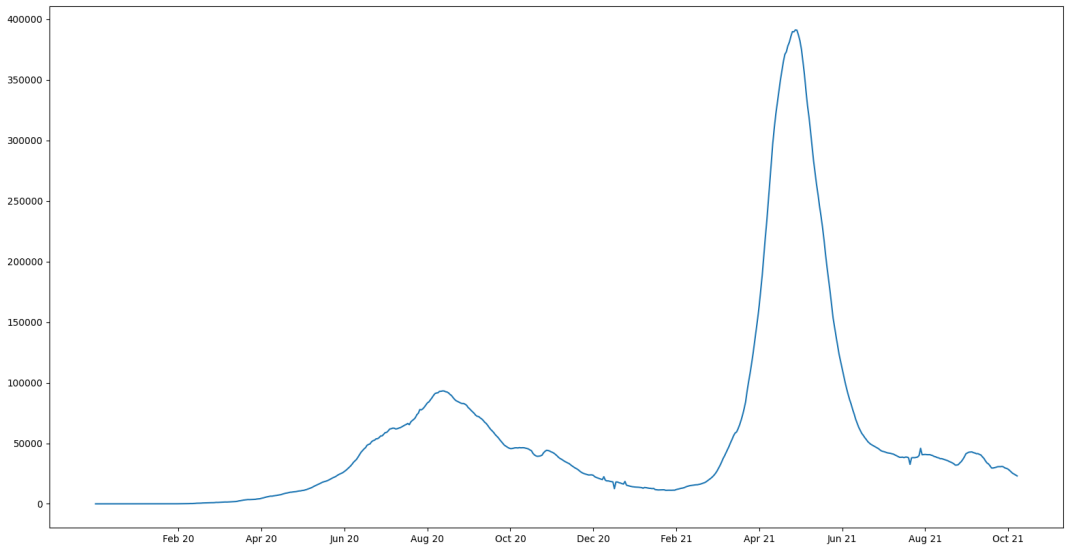

Figure 1 No. of COVID-19 cases vs. days

Inferences:

1. The successive days have similar no. of confirmed cases.
2. It can be observed that the curve we plotted is continuous therefore the consequent days have similar no. of confirmed cases.
3. The first wave was around August-2020 and the second wave was around May-2021.

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 sequence are highly correlated.
2. The no. confirmed cases on consequent days are almost similar as the correlation coefficient is almost equal to 1.
3. They are highly correlated and consequent days have similar covid cases because the coefficient of correlation is almost equal to 1.

c.

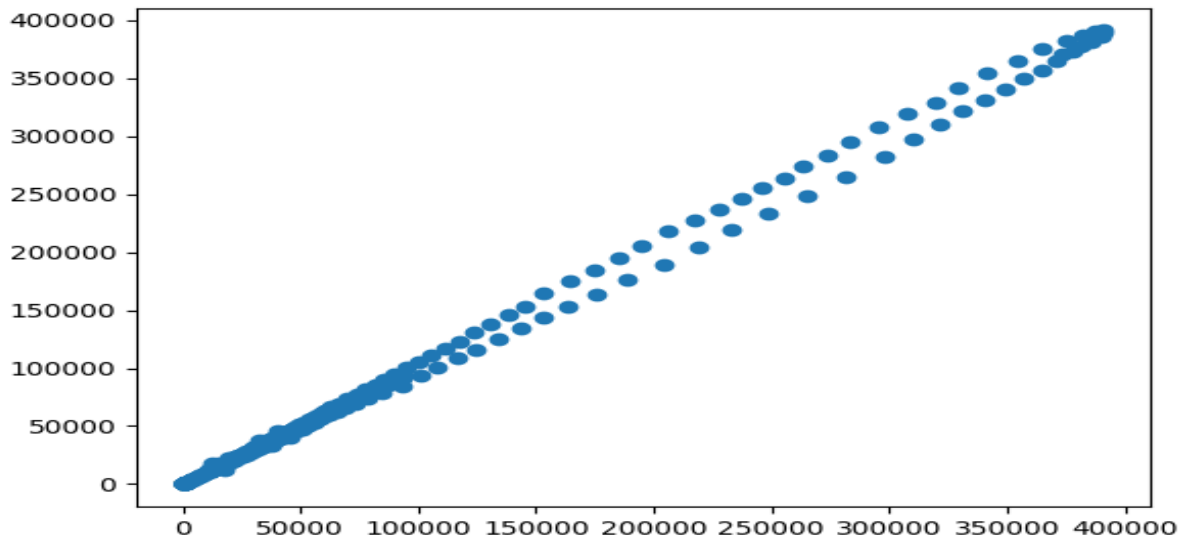

Figure 2 Scatter plot one day lagged sequence vs. given time sequence

Inferences:

1. From the nature of the spread of data points, the correlation is nearly 1.
2. Yes, the scatter plot seems to obey the nature reflected by Pearson's correlation coefficient.
3. As it can be observed that the curve is around the diagonal of the plot its correlation must be near 1.

d.

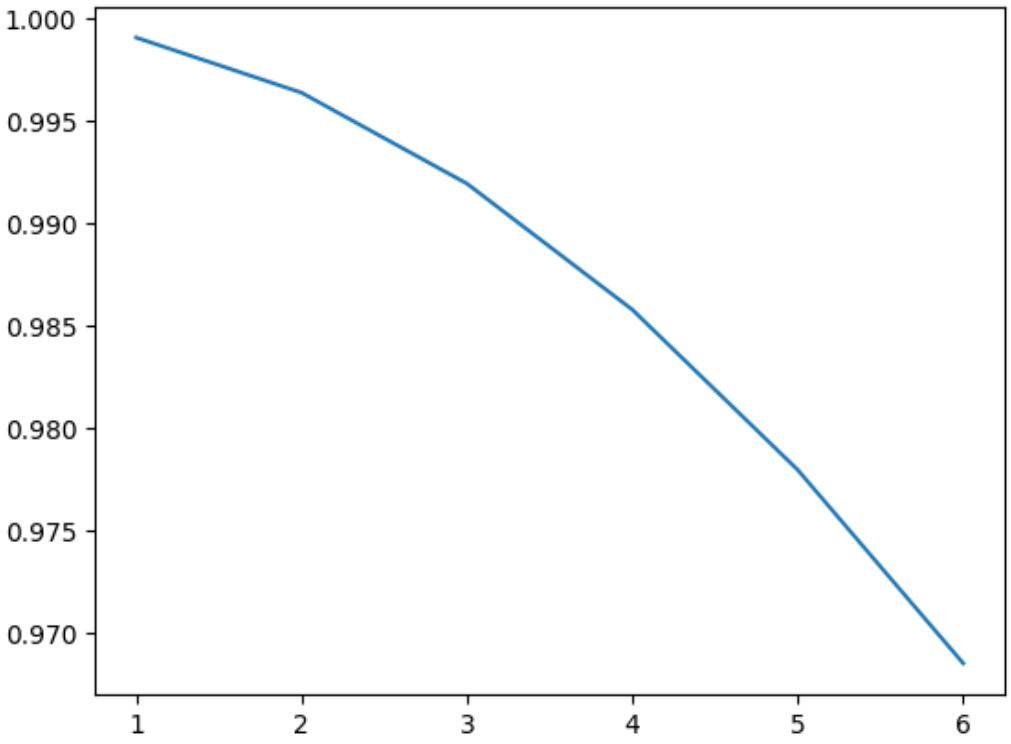

Figure 3 Correlation coefficient vs. lags in given sequence

Inferences:

1. The value of correlation coefficient decreases with increase in the no. of lags.
2. As the covid cases on consequent days are nearly same, the correlation coefficient between two consequent days will be more than that of correlation between two days with more gap in between.

e.

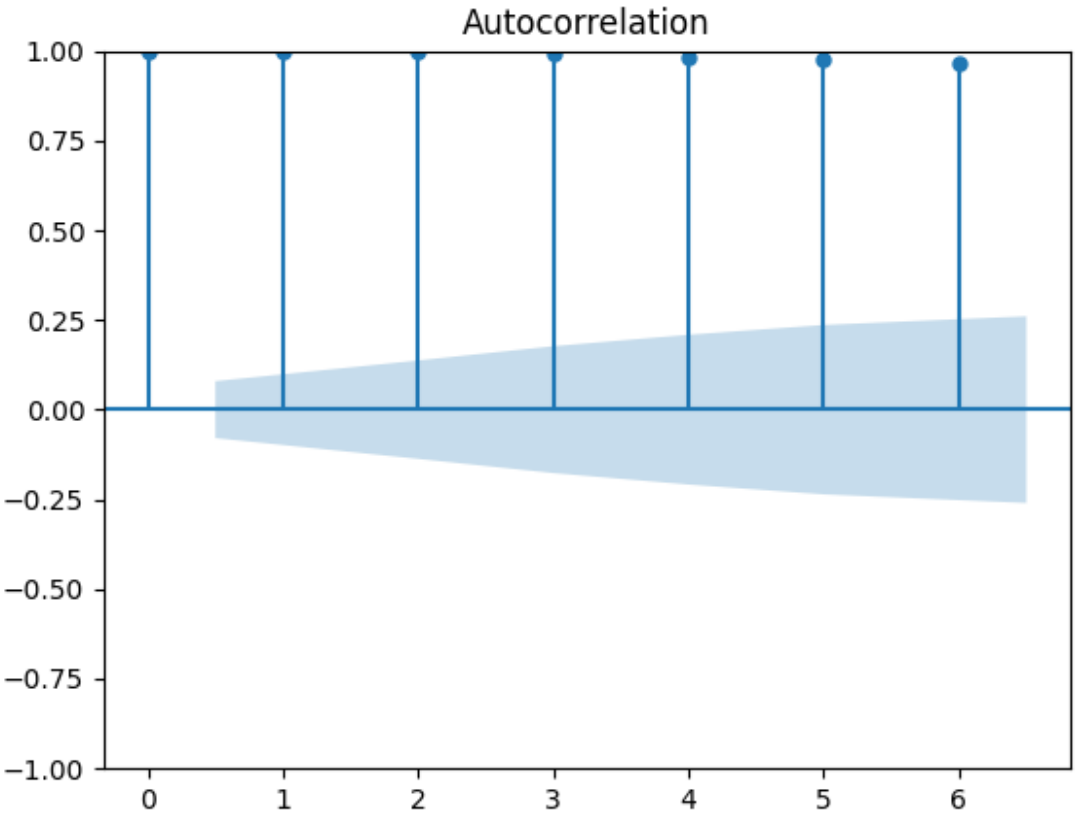

Figure 4 Correlation coefficient vs. lags in given sequence generated using 'plot_acf' function

Inferences:

1. The correlation coefficient is almost equal to 1 but is decreasing continuously.
2. As the covid cases on consequent days are nearly same, the correlation coefficient between two consequent days will be more than that of correlation between two days with more gap in between.

2

a. The coefficients obtained from the AR model are: 59.955, 1.036, 0.262, 0.027, -0.175, -0.152 respectively.

b.

i.

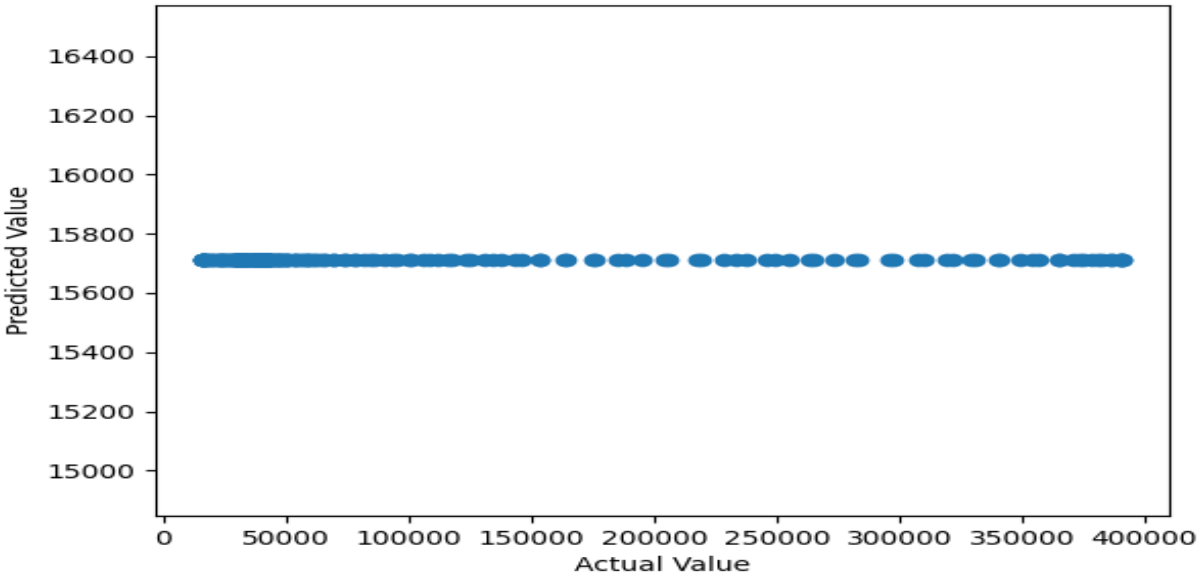

Figure 5 Scatter plot actual vs. predicted values

Inferences:

1. From the nature of the spread of data points, the nature of the correlation between both the sequences is high i.e. nearly 1.
2. Yes, the scatter plot seem to obey the nature reflected by Pearson's correlation coefficient.
3. Our model predicts correctly, as the correlation is almost equal to 1, therefore the model predicts good.

ii.

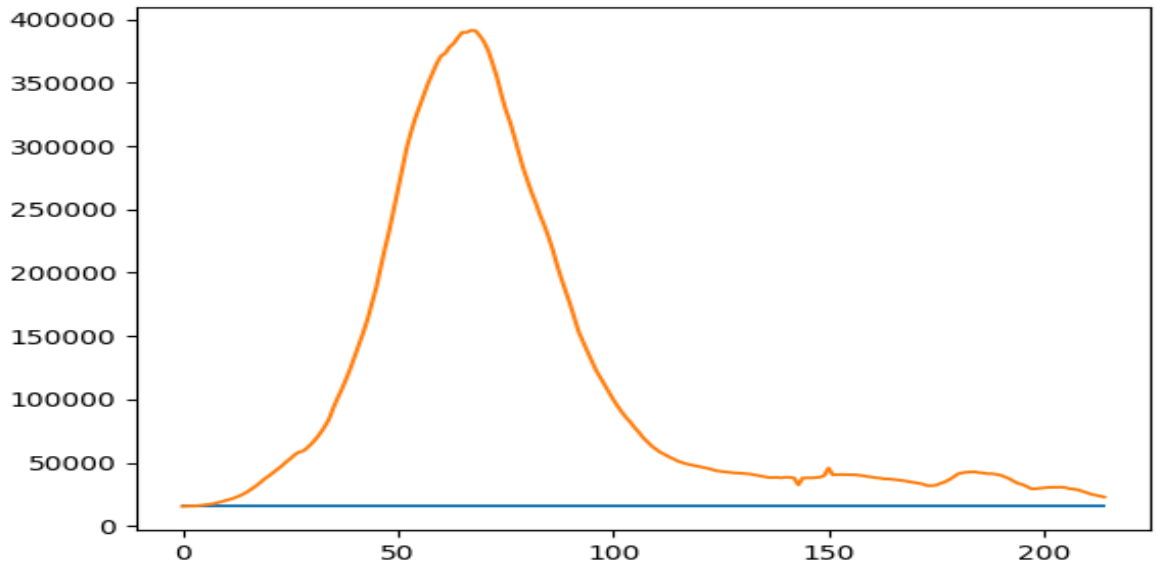

Figure 6 Predicted test data time sequence vs. original test data sequence

Inferences:

1. Our model is reliable for future predictions as it gives almost same values which are in test data.

iii.

The RMSE is 1.825 and MAPE is 1.574, between predicted confirmed cases for test data and original values for test data.

Inferences:

1. From the value of RMSE and MAPE value we can say that our model good and reliable.
2. Both the values i.e. RMSE and MAPE are below 2 which means our model predicts correctly.

3

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

Lag value	RMSE (%)	MAPE
1	5.372	3.446
5	1.824	1.574
10	1.686	1.519
15	1.612	1.496
25	1.703	1.535

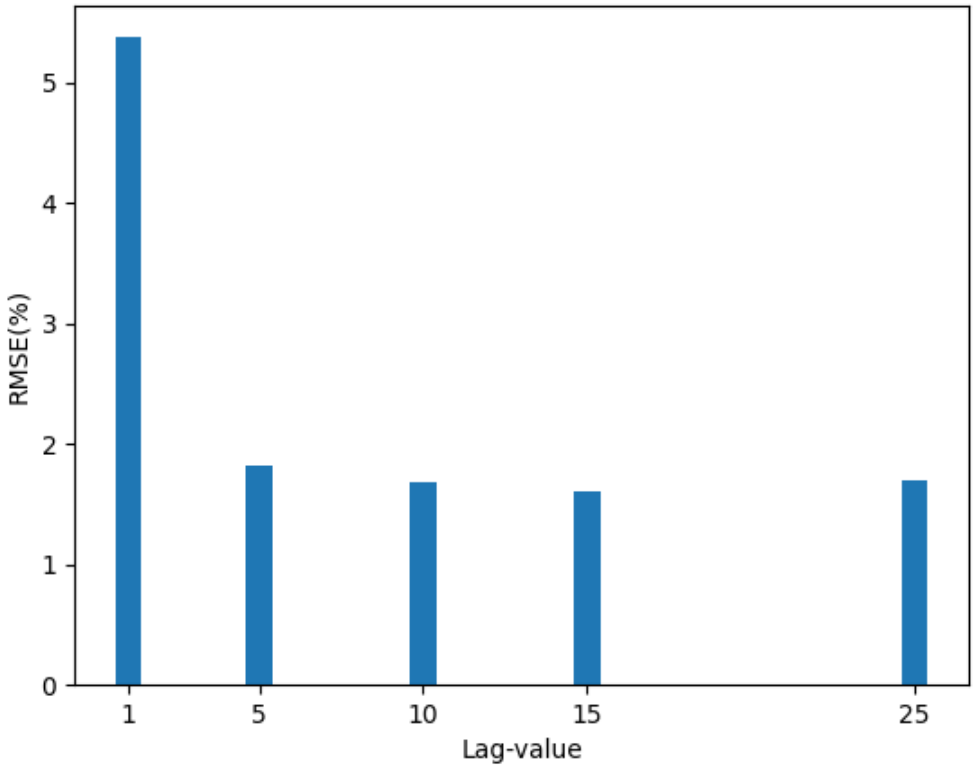

Figure 7 RMSE vs. time lag

Inferences:

1. RMSE decreases quickly from 1 to 5 but after that it decreases gradually with increase in lag value.

2. The reason behind inference 1 is that the complex model needed to fit our data more accurately therefore when the lag is increased from 1 to 5 the accuracy improves significantly but after that the increase is gradual

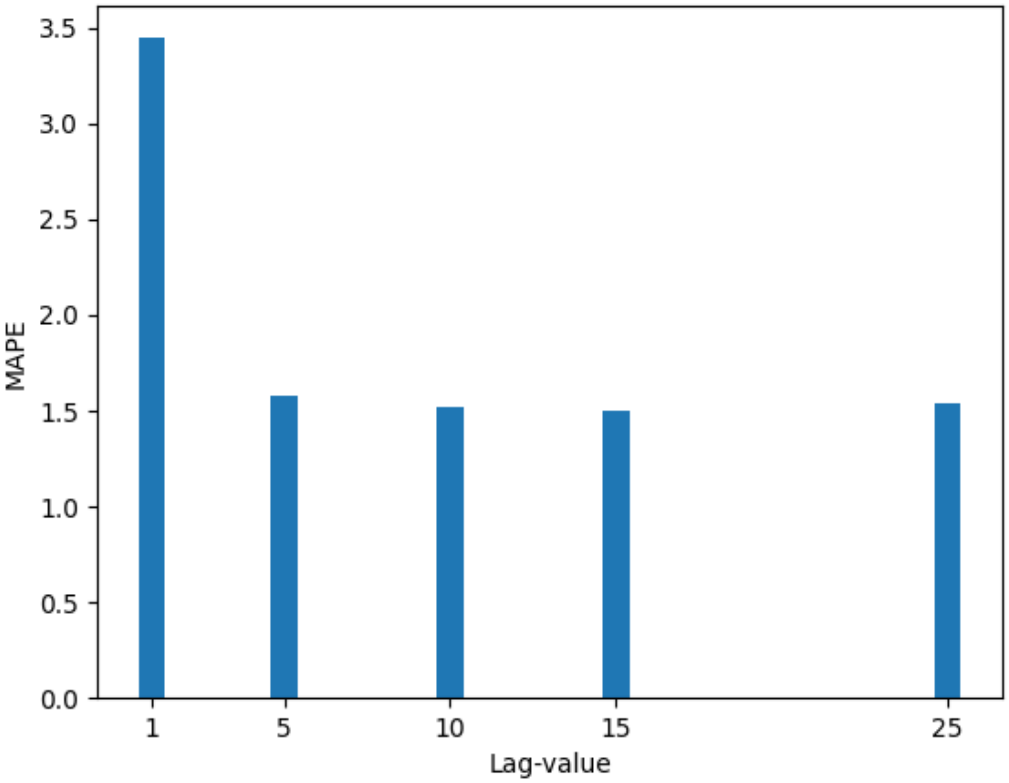

Figure 8 MAPE vs. time lag

Inferences:

1. MAPE first decreases with the increase in lags then it increases with increase in lags.
2. The reason behind inference 1 is that a complex model is needed to fit our data more accurately therefore when the lag is increased from 1 to 5 the accuracy improves significantly but after that the increase is gradual.

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 respectively are 1.76 and 2.026

Inferences:

1. Based upon the RMSE and MAPE value, the optimal lag value didn't improve the prediction accuracy of the model, as it can be seen that the RMSE for lag value 10 is less than that of the optimal lag value.
2. Because as we keep increasing the lag, after certain time the pattern RMSE vs Lag will become random and we can also see that as the observation are made for every day AR(77) does not make sense than that of a lag of around one day.
3. The prediction accuracies obtained without heuristic is less as compared to with heuristic for calculating optimal lag with respect to RMSE and MAPE values.