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

Student's Name: Mayank Bansal **Mobile No: +91963699345**

Roll Number: B20156 **Branch:CSE**

1 a.

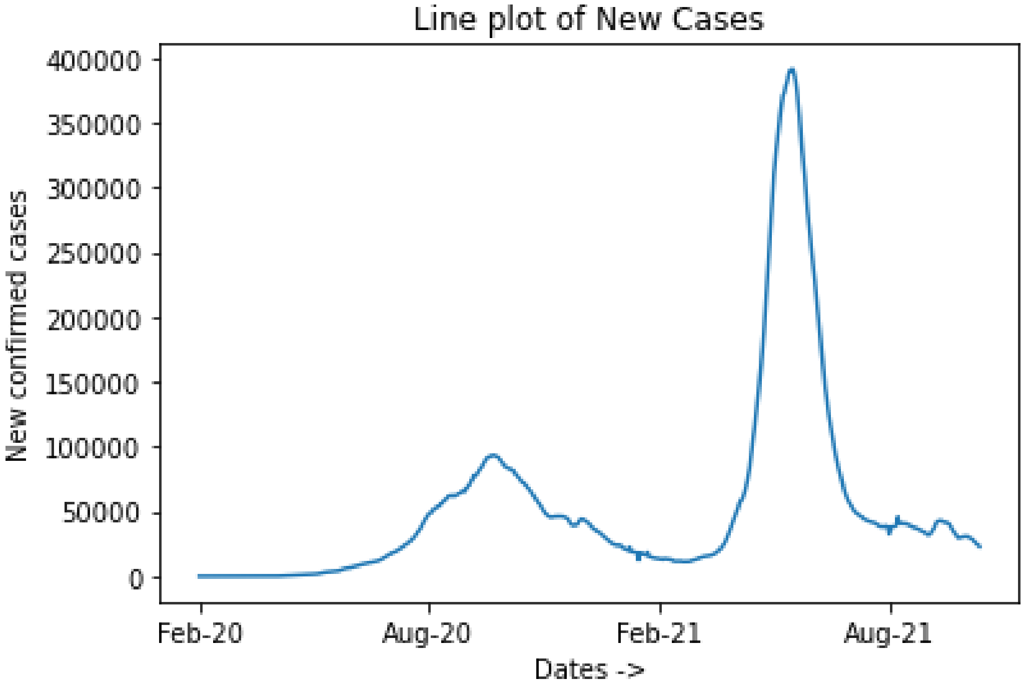

Figure 1 No. of COVID-19 cases vs. days

Inferences:

1. From the plot the number of cases one after the other have similar rise. Points at the start and end of first and second wave have a slight discrepancy.
2. Graph shows that number cases in coming days depend on the present number of cases.
3. First wave lies within Aug20-Oct20 and Second within May21-Aug21

b. The value of the Pearson's correlation coefficient is 0.998

Inferences:

1. Pearson’s correlation coefficient shows that value of correlation b/w both time series is very strong.
2. We generally expect results of days one after the other to be similar. It can be verified with correlation coefficient which is almost 1.
3. Because COVID spread was generally based on current number of patients in the open environment which were further affected by the past.

c.

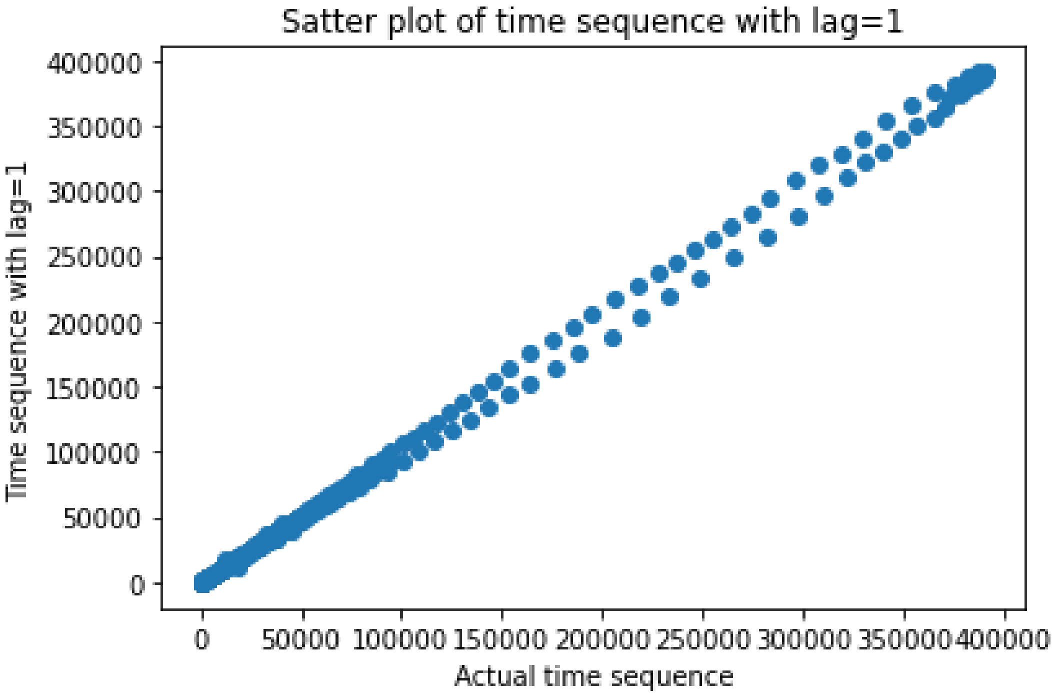

Figure 2 Scatter plot one day lagged sequence vs. given time sequence

Inferences:

1. From the nature of the spread of data points, what do you infer about the nature of correlation between the two sequences is very high(almost equal to 0.999).
2. Yes the scatter plot seem to obey the nature reflected by Pearson’s correlation coefficient calculated in 1.b
3. Yes because all the future cases were dependent on past patients in open environment.

Note: The scatter plot above is for illustration purposes. Replace it with the scatter plot obtained by you. Suitably rename x-axis and y-axis legends.

d.

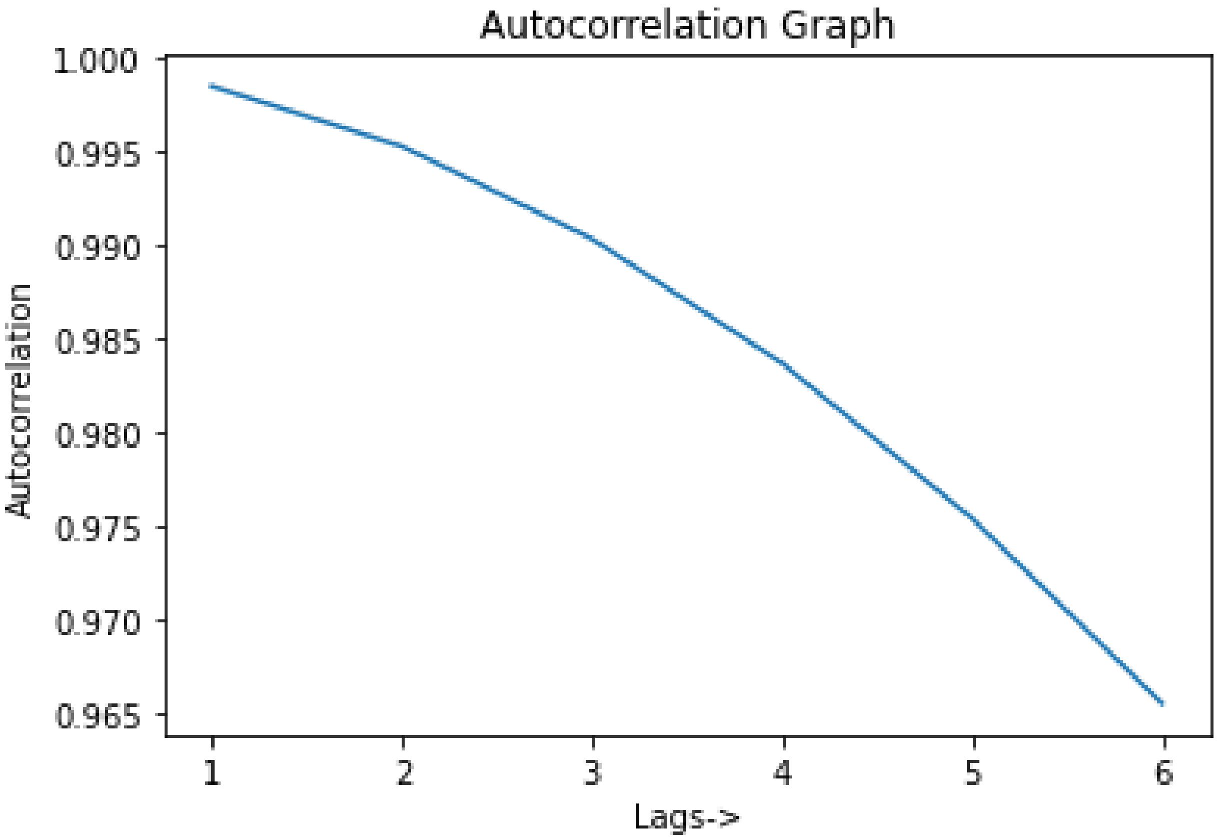

Figure 3 Correlation coefficient vs. lags in given sequence

Inferences:

1. The correlation coefficient value decreases gradually with respect to increase in lags in time sequence.

2. As we keep increasing the number of lag values, the dependency of output on lag value is low ,therefore it decreases.

e.

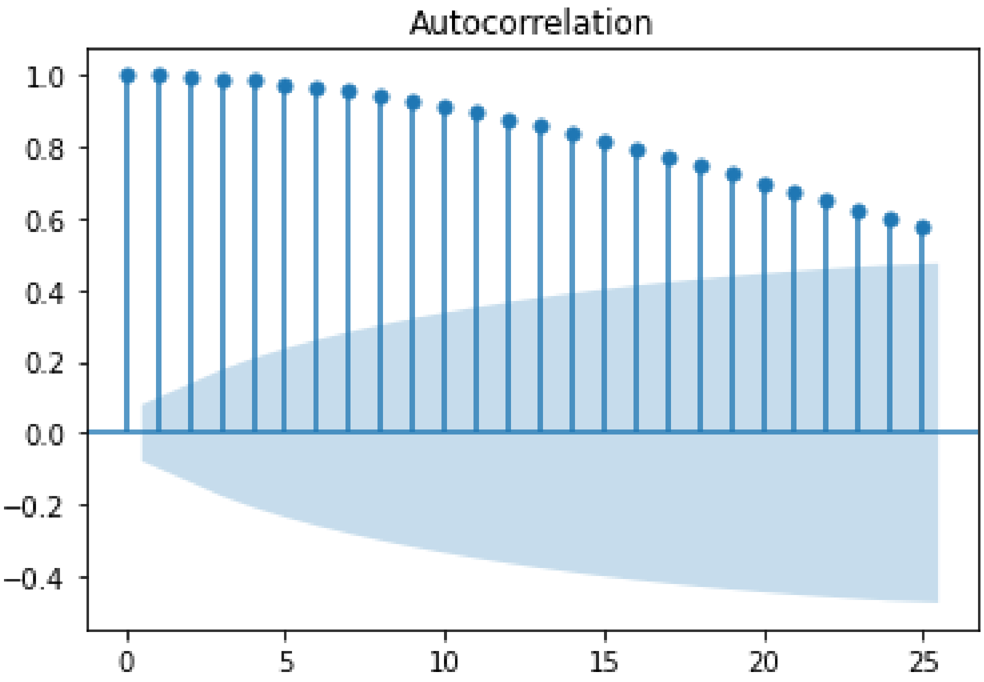

Figure 4 Correlation coefficient vs. lags in given sequence generated using 'plot_acf' function

Inferences:

1. The correlation coefficient value decreases gradually with respect to increase in lags in time sequence.
2. As we keep increasing the number of lag values, the dependency of output on lag value is low ,therefore it decreases.

2

a. The coefficients obtained from the AR model are

[59.954, 1.037, 0.262, 0.027, -0.175, -0.152]

b. i.

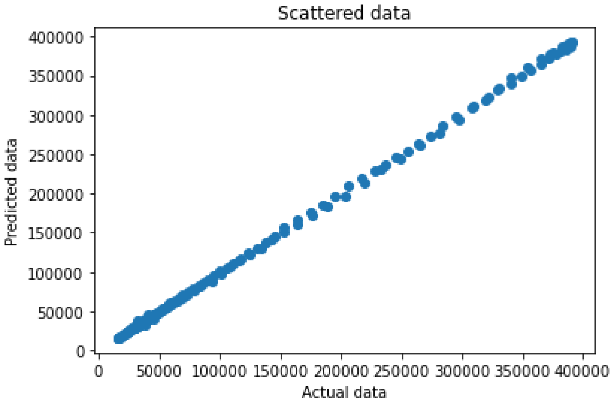

Figure 5 Scatter plot actual vs. predicted values

Inferences:

1. From the nature of the spread of data points, the nature of the correlation between the two sequences is very strong.
2. Yes the scatter plot seem to obey the nature reflected by Pearson's correlation coefficient calculated in 1.b .
3. As the lag is increased, more variables are added to our regression model and it inherently improves the fit.

ii.

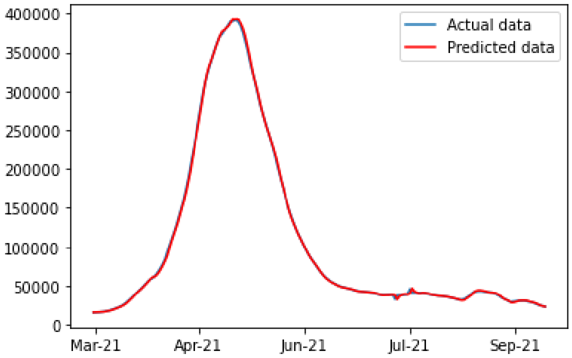

Figure 6 Predicted test data time sequence vs. original test data sequence

Inferences:

1. From the plot of predicted test data time sequence vs. original test data sequence, the model is very much reliable.

iii.

The RMSE(%)=1.825 and MAPE= 1.575

Inferences:

1. From the value of RMSE(\%) and MAPE the model is quite accurate but still it can be improved further.
2. We can take into consideration more lag values which can decrease RMSE values.

3

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

Lag value	RMSE (%)	MAPE
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

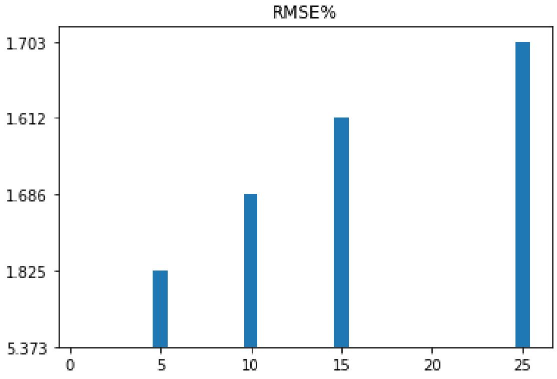

Figure 7 RMSE(%) vs. time lag

Inferences:

1. The RMSE(%) decreases quickly from 1 to 25 but after that there is less decrement with respect to increase in lags in time sequence but also there is slight increase when lag values change from 15 to 25.
2. It is because a complex model is needed to fit our data more accurately so when the lag is increased from 1 to 5 the accuracy improves significantly but accuracy becomes gradual after that.

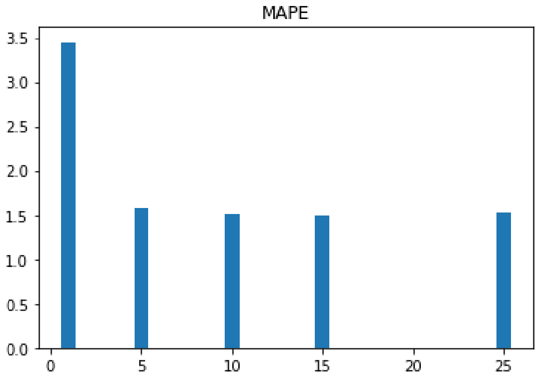

Figure 8 MAPE vs. time lag

Inferences:

1. The MAPE decreases quickly from 1 to 25 but then decreases gradually with respect to increase in lags in time sequence but there is a slight increase in MAPE value from lags=15 to 25.
2. It is because a complex model is needed to fit our data more accurately so when the lag is increased from 1 to 5 the accuracy improves significantly but then the increase in accuracy in gradual.

4

The heuristic value for the optimal number of lags is 77

The RMSE(%) =1.759and MAPE =2.026.

Inferences:

1. Based upon the RMSE(%) and MAPE value, the heuristics for calculating the optimal number of lags didn’t improve the prediction accuracy of the model significantly as we can see the MAPE for lag=5,10,15,25 was less than that for optimal lag.
2. Because as we keep increasing the lag, after certain time the pattern of RMSE vs lag will become random.
3. The prediction accuracies obtained without heuristic is less as compared to with heuristic for calculating optimal lag with respect to MAPE values , the prediction accuracies are almost same with respect to RMSE values in both cases.