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

Student's Name: Deepak Kumar **Mobile No: 9817959477**

Roll Number: B20191 **Branch: Electrical Engineering**

1 a.

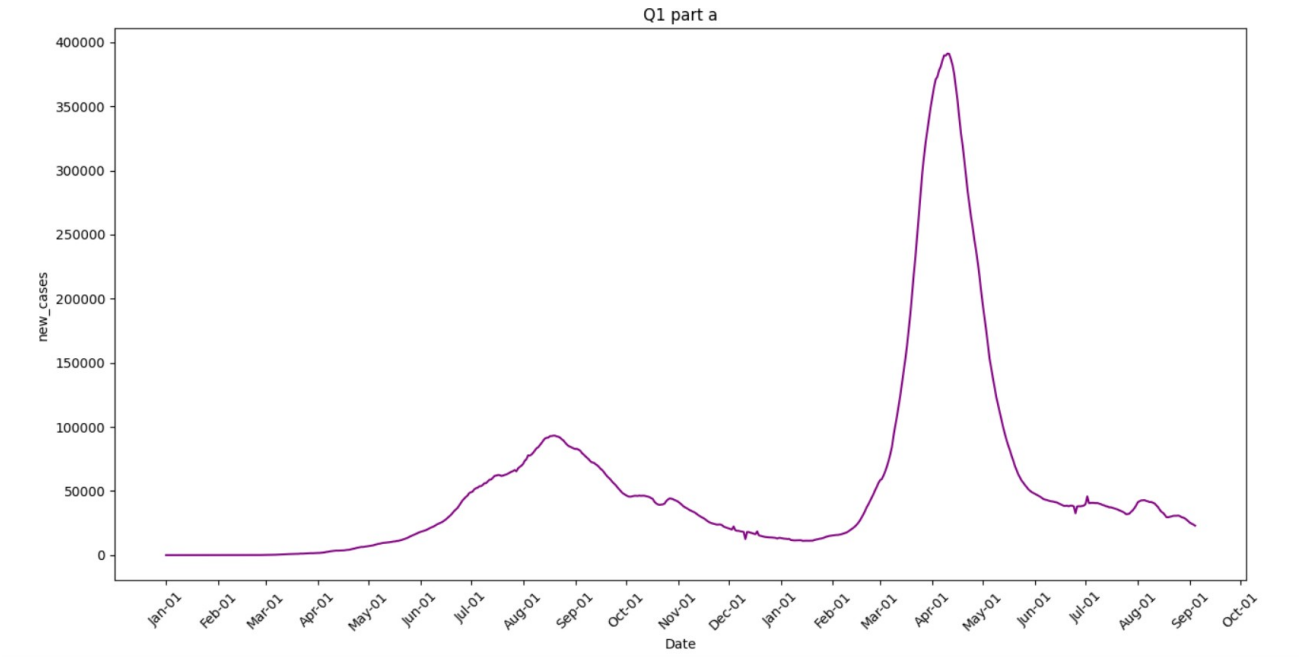

Figure 1 No. of COVID-19 cases vs. days

Inferences:

1. No, as we can see that the number of cases in second wave are increasing more rapidly and after its peak, cases are decreasing more rapidly than that of first wave.
2. The duration of first wave is around 8 months which is less than duration of second wave (6 months) and peak value of second wave is greater than that of first wave.

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

Inferences:

1. Two time series are very strongly correlated with each other. It means that future values are affectedby past values.
2. As the value of Pearson correlation coefficient is very high, it means that observations on days one after the other are very similar and future observations will be highly dependent on previous observations.

c.

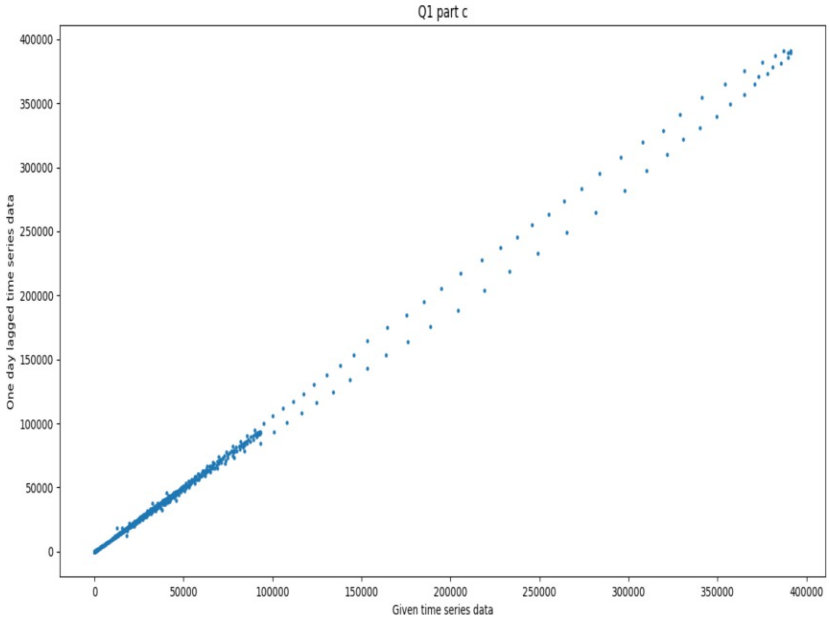

Figure 2 Scatter plot one day lagged sequence vs. given time sequence

Inferences:

1. Correlation between to variables is **very strongly** and **positive**, means two variables are highly dependent on each other.
2. By this scatter plot we can observe that the correlation coefficient should be almost 1, and we found the same as expected in Q1b.

d.

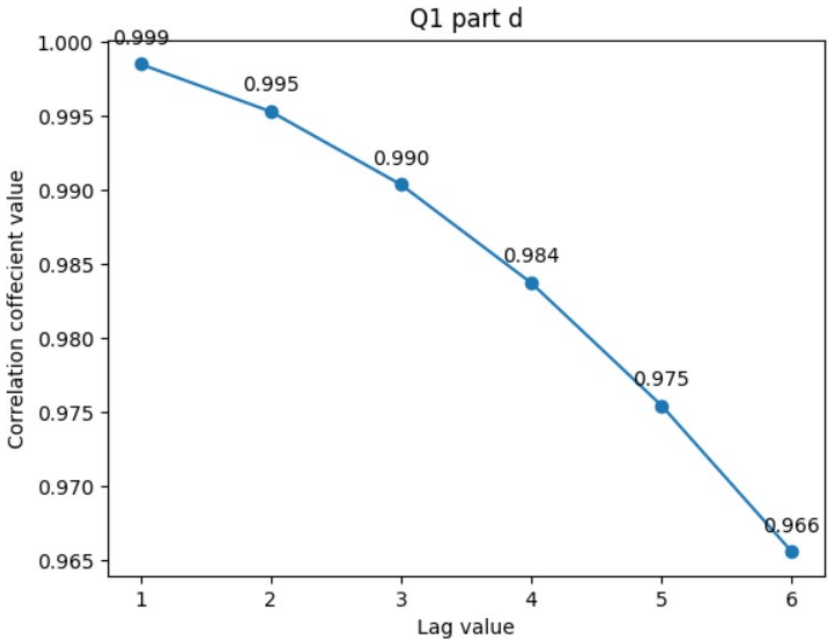

Figure 3 Correlation coefficient vs. lags in given sequence

Inferences:

1. Correlation coefficient is **decreases** as lag value **increases**.
2. When data have a trend, the autocorrelations for small lags tend to be large and positive because observations nearby in time are also nearby in size. So, the ACF of trended time series tend to have positive values that slowly decrease as the lags increase.

e.

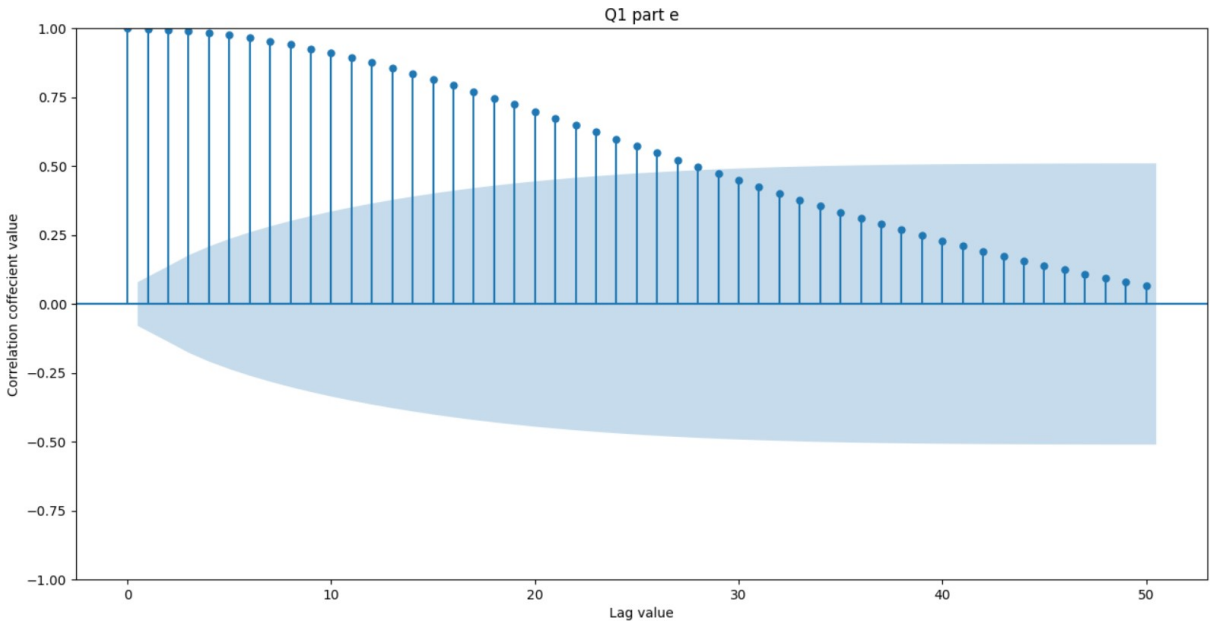

Figure 4 Correlation coefficient vs. lags in given sequence generated using 'plot_acf' function

Inferences:

1. Value of correlation coefficient **decreases** as lag value **increases**.
2. Autocorrelations for small lags tend to be large because observations nearby in time are also nearbyin size. So, the ACF of trended time series tend to have positive values that slowly decrease as the lags increase.

2

a. The coefficients obtained from the AR model are [5.995, 1.037, 0.262, 0.028, -0.175, -0.152].

b. i.

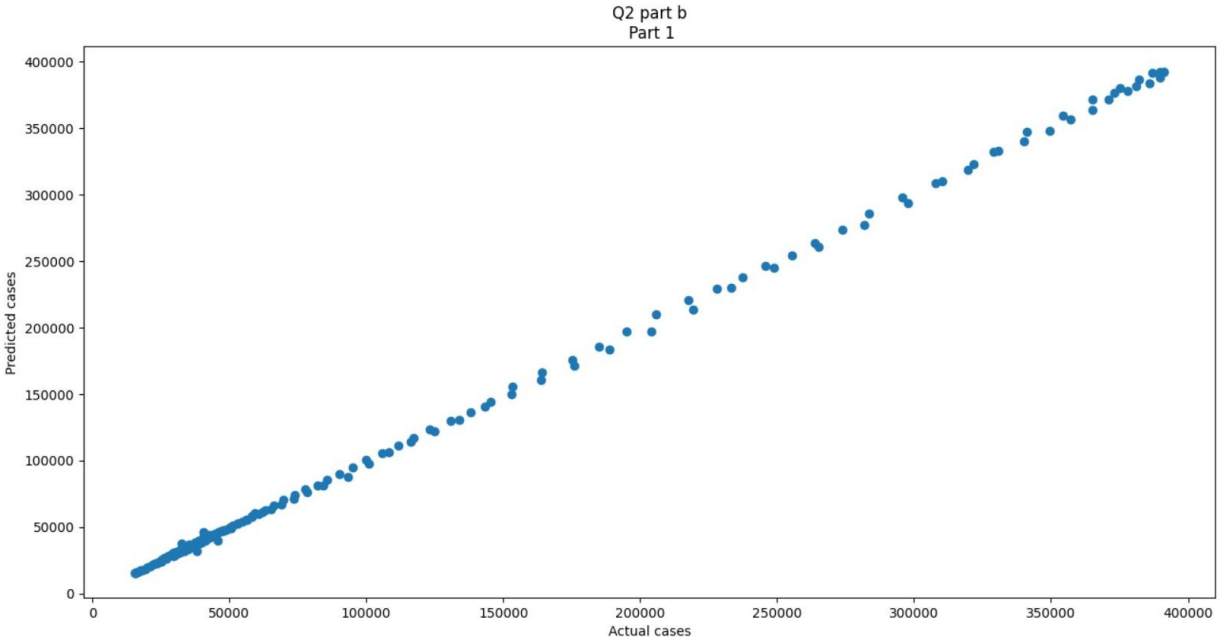

Figure 5 Scatter plot actual vs. predicted values

Inferences:

1. From the nature of the spread of data points, the two sequences have very strong positive correlation.
2. Yes, the scatter plot seems to obey the nature reflected by Pearson's correlation coefficient calculated in 1.b.
3. The data points are dense around the Y=X line.

ii.

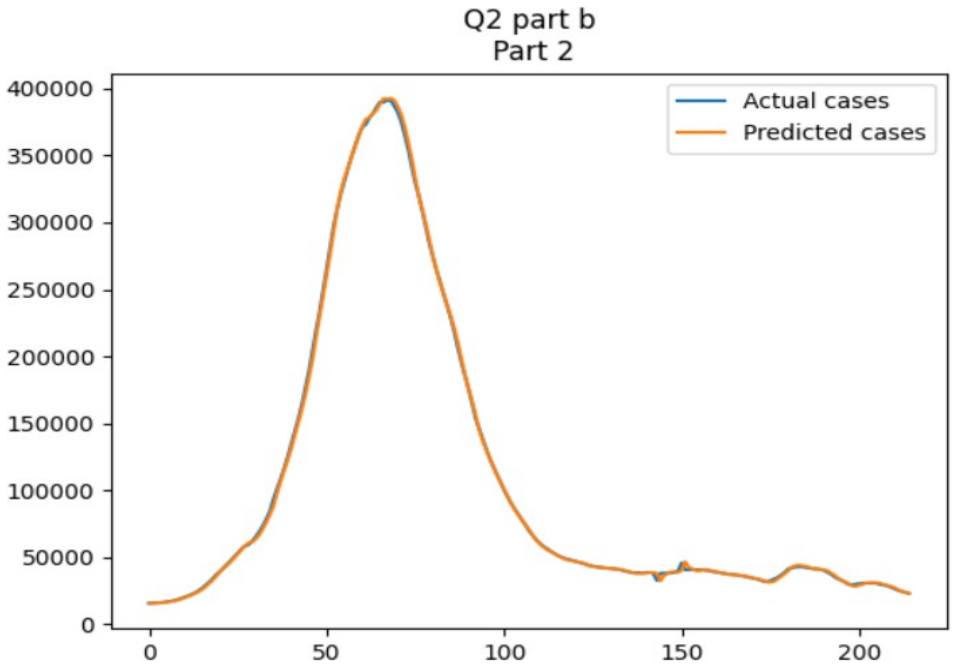

Figure 6 Predicted test data time sequence vs. original test data sequence

Inferences:

1. From the graph we can see predicted values are quite accurate so we can say that our model is reliable.
2. About future prediction if it only depends upon past observations then this model will be highly useful but if future depends upon other conditions also then this model can give better results.

iii.

RMSE is 1.825% and MAPE is 0.016.

Inferences:

1. As the RMSE and MAPE values are too small, the model for the given time series is quite accurate.
2. RMSE and MAPE values are nothing but error in our prediction. As the error is too small, our model is reliable for this series.

3

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

Lag value	RMSE (%)	MAPE
1	5.373	0.035
5	1.825	0.016
10	1.685	0.015
15	1.612	0.015
25	1.703	0.016

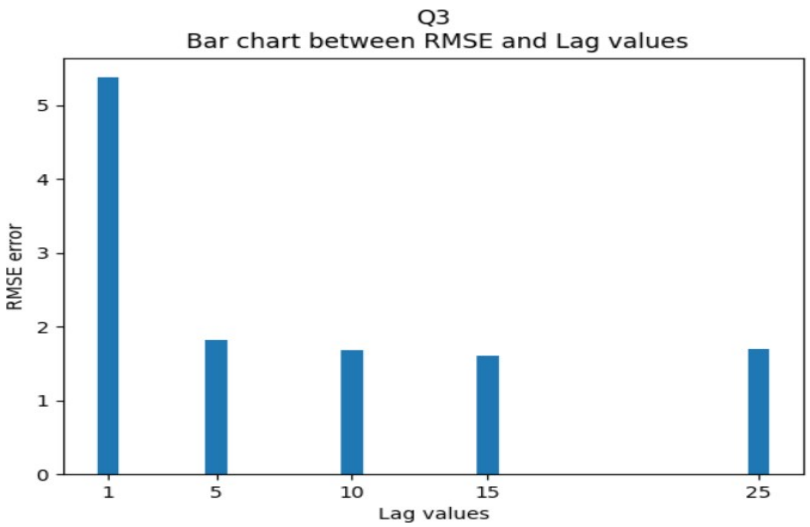

Figure 7 RMSE(%) vs. time lag

Inferences:

1. RMSE(%) value is decreasing with the increase in lags in time sequence.

2. Increasing lag means future predictions will depend upon the previous value up to that lag. Recent past value helps more in predicting future values but as we go very deep in past then some value didn't represent future observations correctly as a result these values contributes to error.

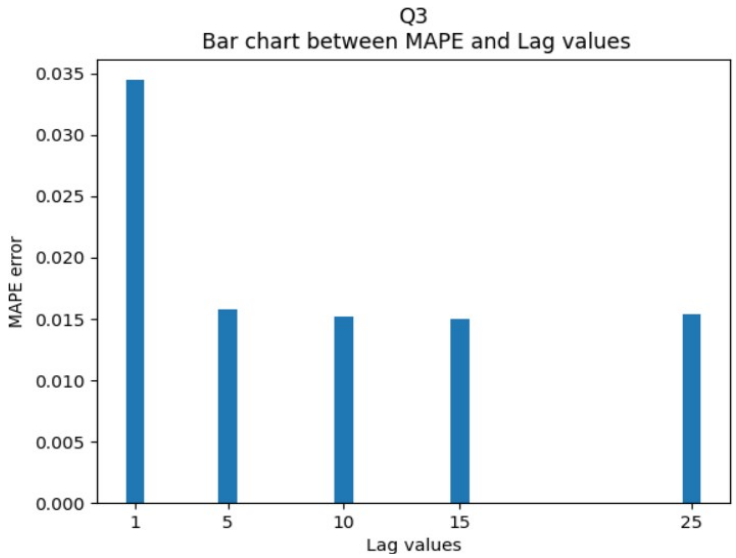

Figure 8 MAPE vs. time lag

Inferences:

1. MAPE value is decreasing as we increase lags in time sequence.

2. Increasing lag means future predictions will depend upon the previous value up to that lag. Recent past value helps more in predicting future values but as we go very deep in past then some value didn't represent future observations correctly as a result these values contributes to error.

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 **0.020** respectively.

Inferences:

1. Based upon the RMSE (%) and MAPE value, it seems heuristics for calculating the optimal number of lags didn't improve the prediction accuracy of the model.
2. Optimal lag is 77 which means future observations depends upon 77 previous. Recent past value helps more in predicting future values but as we go very deep in past then some value didn't represent future observations correctly as a result these values contributes to error.