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

Student's Name: Sarthak Saptami Kumar Jha **Mobile No: 8825319259**

Roll Number: B20317 **Branch: CSE**

1 a.

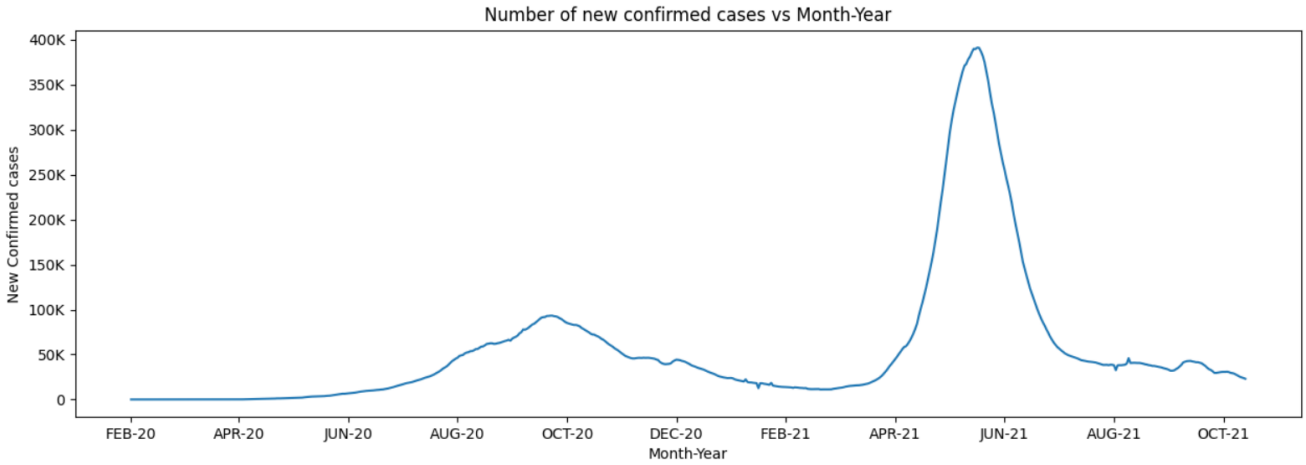


Figure 1 No. of COVID-19 cases vs. days

Inferences:

1. For most of the curve, the current day, and day before have very similar values, except for the point where there is a maximum, at such points there is high fluctuations in the values.
2. Since the data does not have any large fluctuations, we can say that the next and before data are quite similar in value.
3. The duration of the first wave is around 4 months from Aug-20 to Nov-20, while the second wave is around 3 months from Apr-21 to Jun-21.

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

Inferences:

1. We can infer that since there is very high correlation between time series data with one time lag.
2. This observation strongly holds that covid cases with one time lag, are highly correlated.

3. We can observe from the plot that, the data with one day difference does not have any high fluctuations, and therefore, we have very high correlation in data with one time lag.

c.

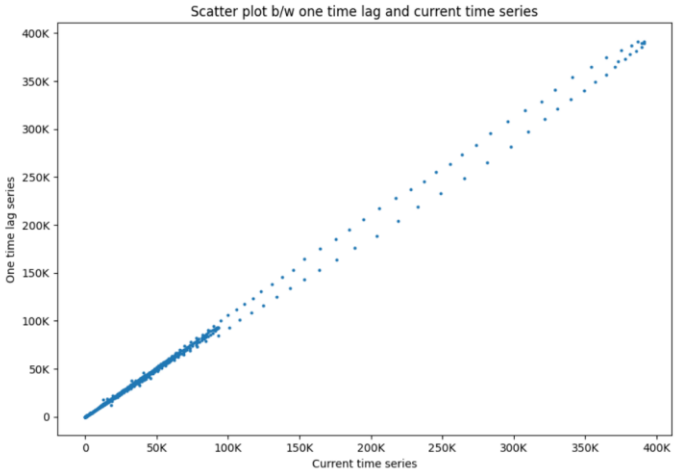
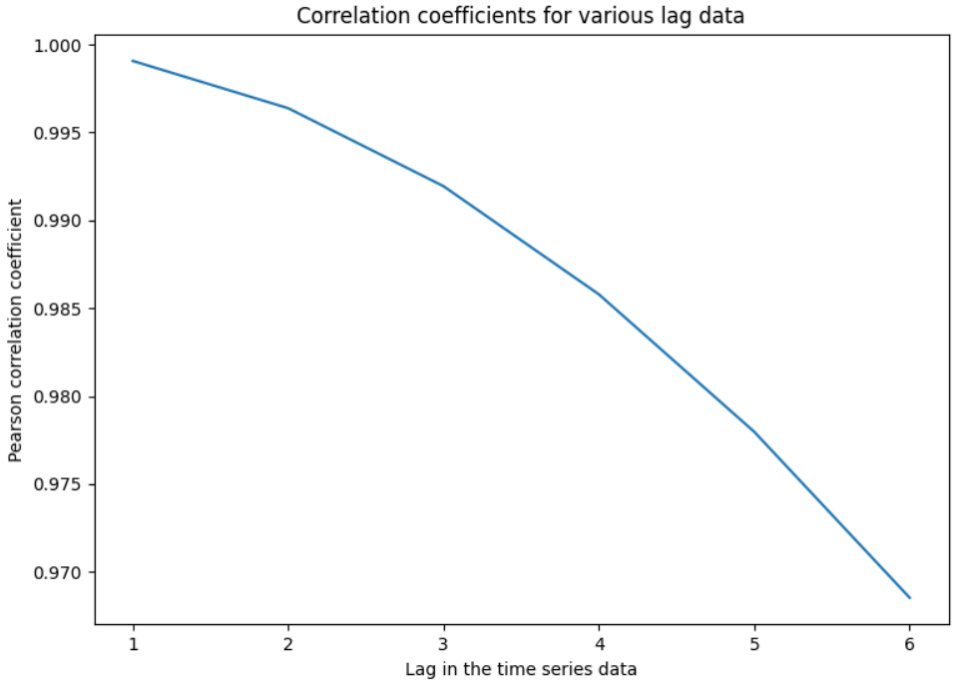


Figure 2 Scatter plot one day lagged sequence vs. given time sequence

Inferences:

1. We can clearly see that, the scatter plot has data points which are aligned equally from the axes, inferring strong positive correlation.
2. Yes, the scatter plot clearly follows the high correlation coefficient calculated in 1.b
3. Since we calculated that there is very high correlation between time series data and one time lag series data, the scatter plot also depicts the same.

d.



Inferences:

1. We can see that with increased time lag, there is a rapid decline in the Pearson correlation coefficient.
2. With increased time lag, there is lesser correlation between the current data and lagged data and the lagged data starts getting much more unrelated compared to the current data.

e.

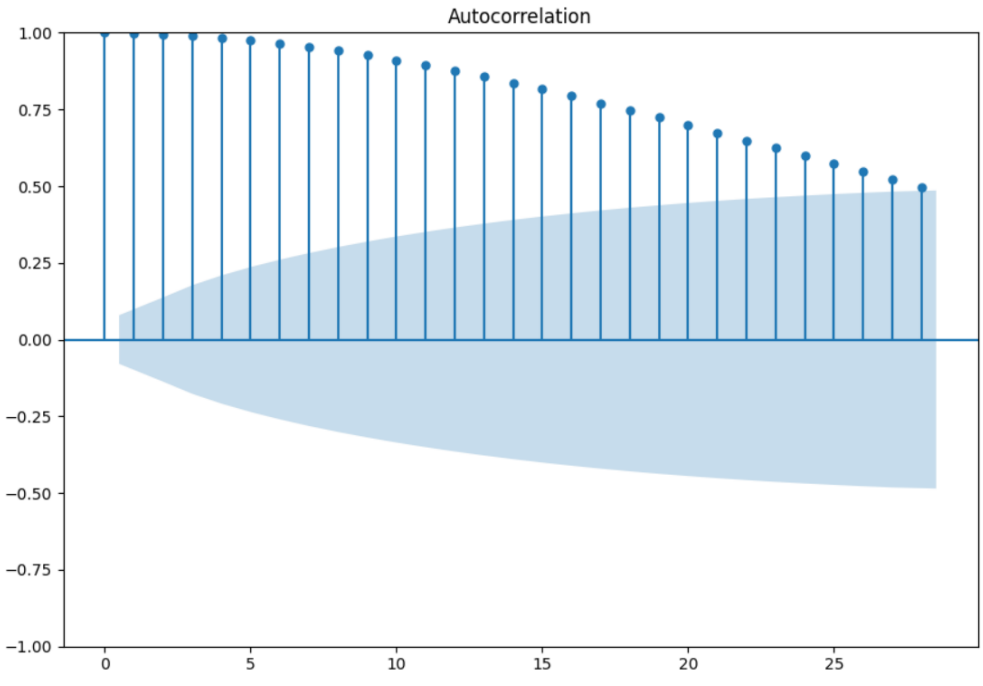


Figure 4 Correlation coefficient vs. lags in given sequence generated using 'plot_acf' function

Inferences:

1. The correlation coefficient falls rapidly with increase in lag days.
2. As we increase the number of lag in the time series, the correlation falls as more fluctuations and variations in the data cause the correlation to decrease.

2

a. The coefficients obtained from the AR model are 59.955, 1.037, 0.262, 0.028, -0.175, -0.152.

b. i.

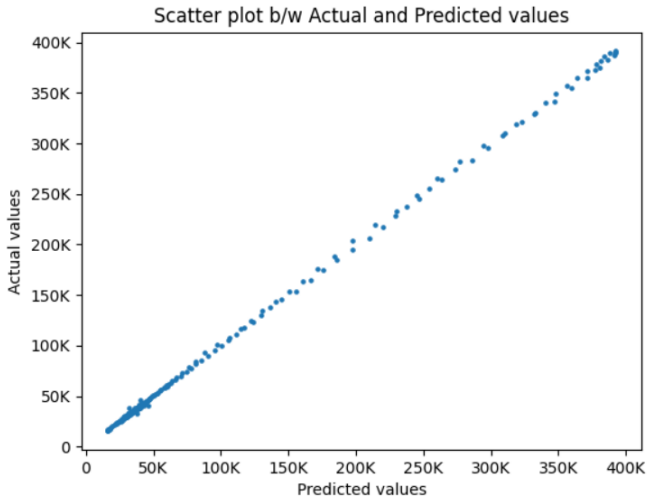


Figure 5 Scatter plot actual vs. predicted values

Inferences:

1. We can say that there is strong positive correlation between the predicted and actual values.
2. Yes the scatter plot depicts a very strong correlation, which is close to the Pearson correlation coefficient calculated in 1b

3. Since we are making 1-step ahead predictions and using 5 lag values, we are therefore getting strong correlation statistics.

ii.

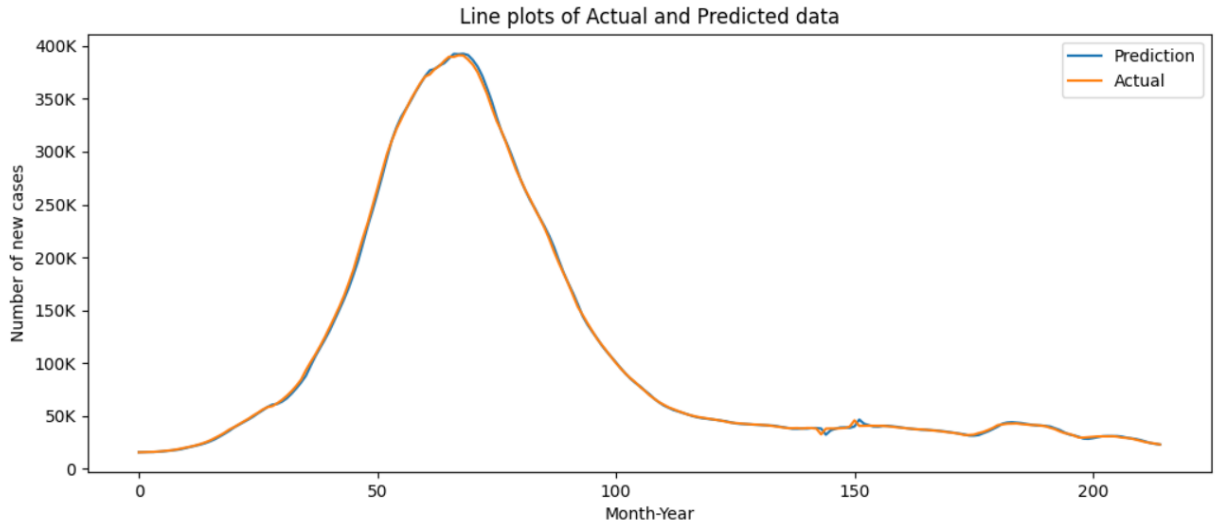


Figure 6 Predicted test data time sequence vs. original test data sequence

Inferences:

1. We can say that the model is very reliable for future predictions as the line plot clearly shows that the model has predicted the data with high accuracy.

iii.

The RMSE(\%) and MAPE between predicted number of cases for test data and original values for test data are 1.825 and 1.575.

Inferences:

1. From the RMSE and MAPE values, we can infer that since they are both less than 2%, the model is reliable for predictions.
2. Since we are using 1-step ahead predictions along with using 5 lag values, we are getting much reliable predictions from the model.

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

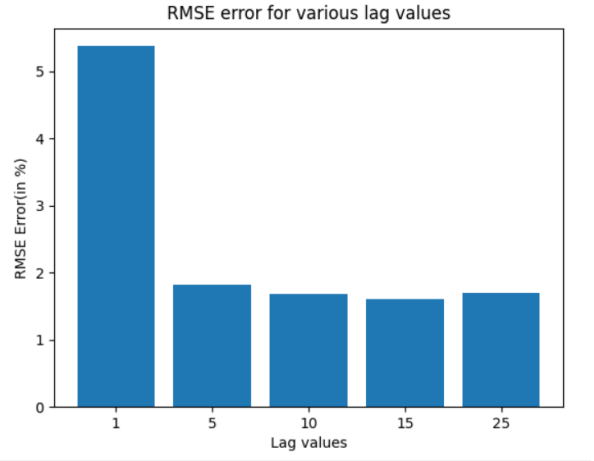


Figure 7 RMSE(%) vs. time lag

Inferences:

1. There is a general trend of decline in RMSE with increase in time lag.
2. As we start increasing lag, we are taking into account much more data for further prediction and thus we have lower RMSE.

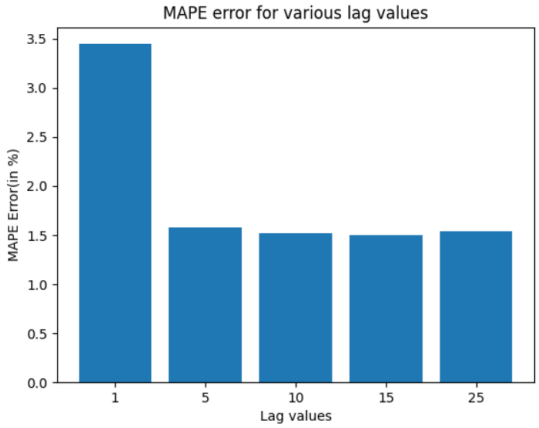


Figure 8 MAPE vs. time lag

Inferences:

1. There is a general trend of decline in MAPE with increase in time lag.
2. As we start increasing lag, we are taking into account much more data for further prediction and thus we have lower MAPE.

4

The heuristic value for the optimal number of lags is 0.136 and the 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.

Inferences:

1. We encounter poor RMSE and MAPE values after using lag values, calculated from autocorrelation.
2. Using the heuristic value, we include till only those lag values, which might contribute significantly to the prediction of data using autoregression, but in this case, it could lead also lead to overfitting.
3. If we do not use heuristic value, and calculate the errors for say 100 lag values, we get RMSE as 1.949 and MAPE as 2.673, while if we calculate heuristic value, we infer that till 77 lag values is sufficient, and thus we get RMSE as 1.759 and 2.026.