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

Student's Name: Abhay Gupta

Mobile No: 9511334630

Roll Number: B20075

Branch:CSE

1 a.

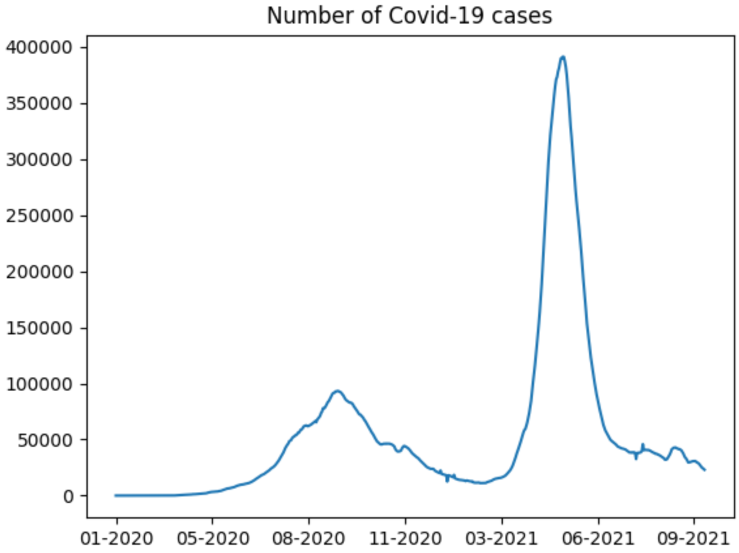

Figure 1 No. of COVID-19 cases vs. days

Inferences:

1. The days after the another have similar number of cases
2. Because the plot if continuous, also, as the number of cases on consecutive days cannot change suddenly. So they are similar
3. The first wave is from May 2020 to Nov 2020 and the second wave is from March 2021 to June 2021.

b. The value of the Pearson's correlation coefficient is **0.99906**

Inferences:

1. As the correlation value is close to 1. It can be inferred that the two time sequences are positively correlated.
2. It can be said that observations one after the other are similar as correlation value is very close to 1.
3. As the correlation is very close to one, which implies the high dependency of time sequences on each other.

c.

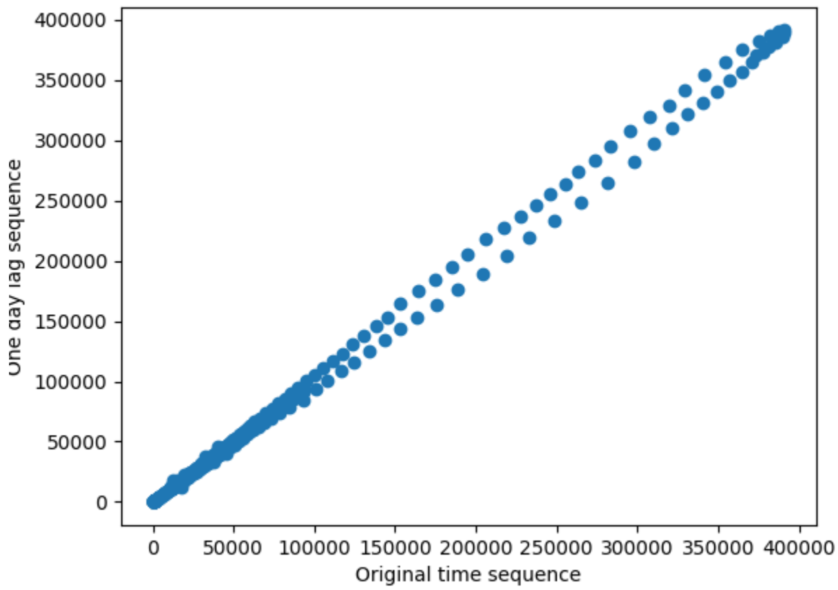

Figure 2 Scatter plot one day lagged sequence vs. given time sequence

Inferences:

1. As, one sequence increases with the increase of the other, we can say that they are positively correlated.

2. The scatter plot seem to obey the nature reflected by Pearson’s correlation coefficient calculated in 1.b as both signifies that correlation value is positive and close to 1.
3. From 1b it can be seen that correlation value is 0.999, from graph also it can be seen that correlation is close to 1 and it’s positive.

d.

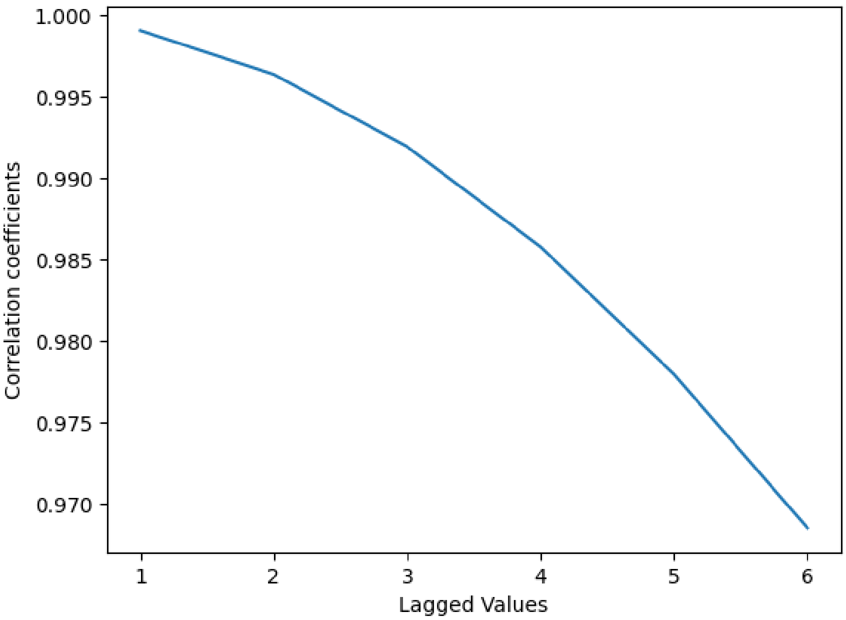

Figure 3 Correlation coefficient vs. lags in given sequence

Inferences:

1. As the lags increase the correlation coefficients decrease.
2. The reason for the above trend is, the value at any time t depends more on the previous value which is near to it as compared to the one which is far from it.

e.

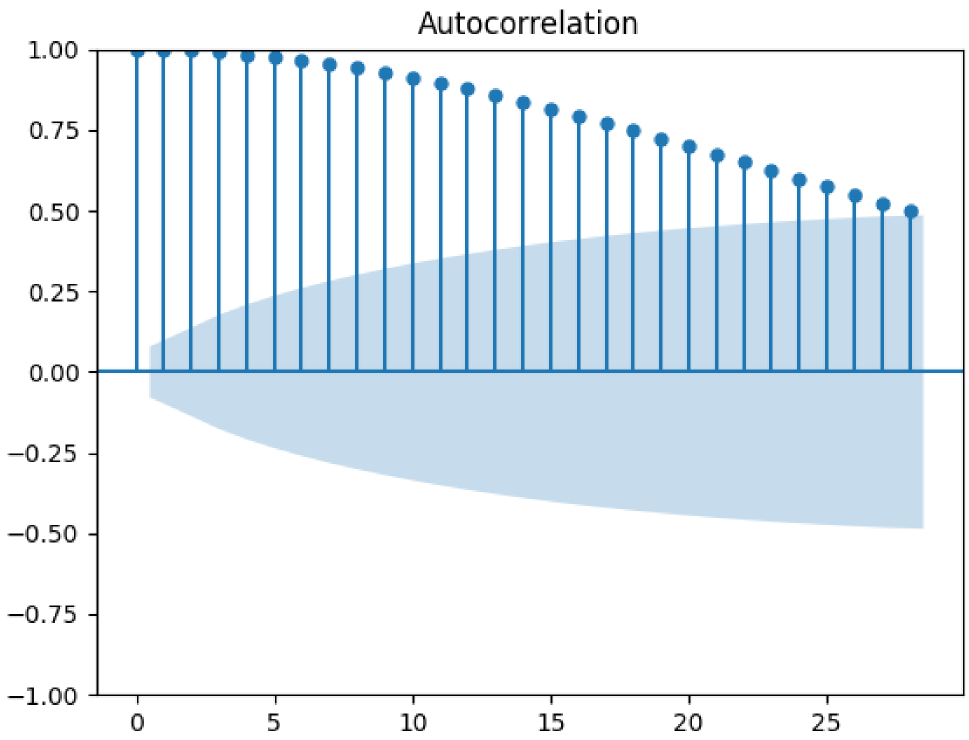

Figure 4 Correlation coefficient vs. lags in given sequence generated using 'plot_acf' function

Inferences:

1. As the lags increase the correlation coefficients decrease.
2. The reason for the above trend is, the value at any time t depends more on the previous value which is near to it as compared to the one which is far from it.

2

a. The coefficients obtained from the AR model are :- 59.954, 1.036, 0.261, 0.027, 0.175, -0.152

b. i.

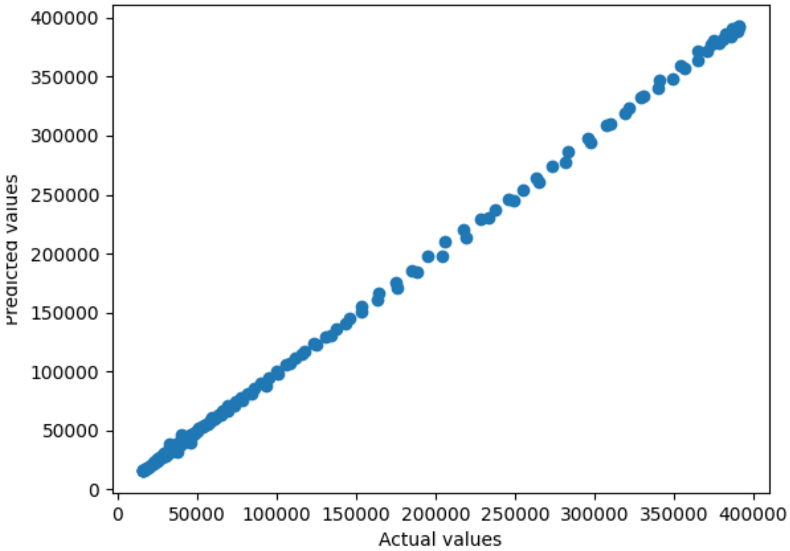

Figure 5 Scatter plot actual vs. predicted values

Inferences:

1. Both the sequences are positively and strongly correlated.
2. From the scatter plot it can be predicted that accuracy of the predicted data is quite well.

ii.

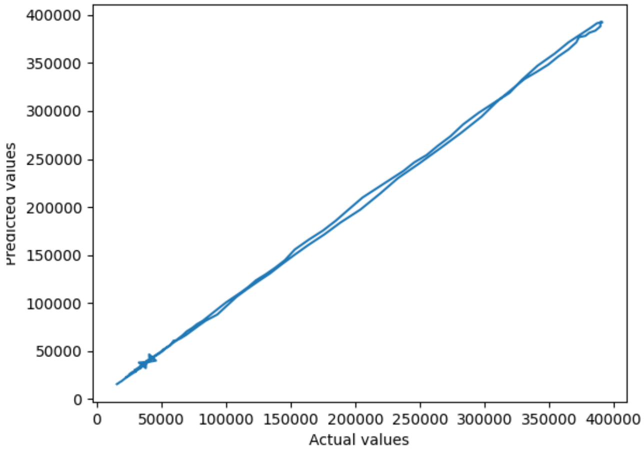

Figure 6 Predicted test data time sequence vs. original test data sequence

Inferences:

1. As the line plots are almost coinciding each other and they are close to the line y = x. So, the predicted data is reliable.

iii.

The RMSE(\%) and MAPE between predicted values of test data and original values for test data are :-

RMSE(%) = 1.824

MAPE = 1.574

Inferences:

1. From the value of RMSE(\%) and MAPE in can be inferred that the data is quite accurate.
2. As the error lies between 1% - 2% which is quite less, so we can say that the predicted data is reliable.

3

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

Lag value	RMSE (%)	MAPE
1	**5.372**	**3.446**
5	**1.824**	**1.574**
10	**1.685**	**1.519**
15	**1.611**	**1.496**
25	**1.703**	**1.535**

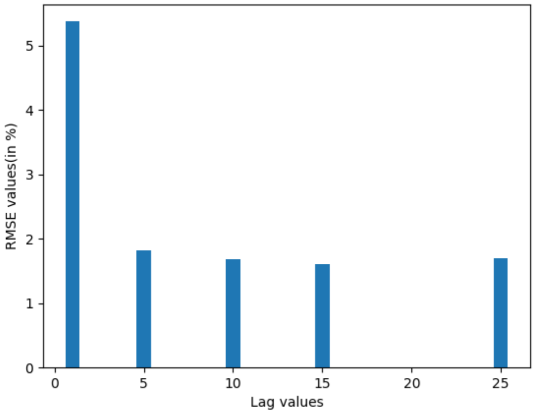

Figure 7 RMSE(%) vs. time lag

Inferences:

1. Firstly, RMSE value decreases till 15, then increase at p=25.
2. The reason for the above inference is that data gets overfit.

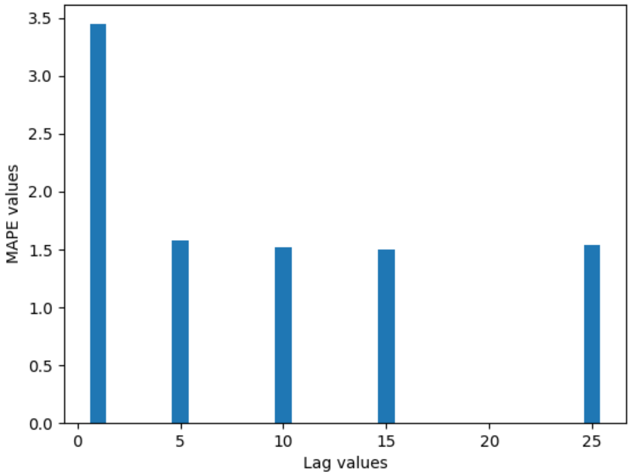

Figure 8 MAPE vs. time lag

Inferences:

1. Firstly, RMSE value decreases till 15, then increase at p=25.
2. The reason for the above inference is that data gets overfit .

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**.

Inferences:

1. As the RMSE value decreases, so heuristic value for the optimal number of lags increases the accuracy.