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

Student's Name: Sanditi Goutham Reddy **Mobile No: 6300507972**

Roll Number: B20130 **Branch: CSE**

1 a.

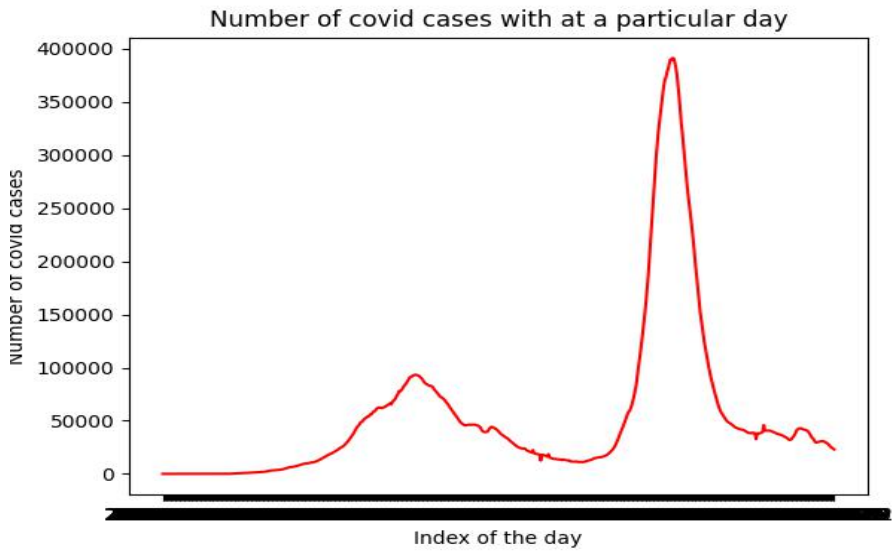


Figure 1 No. of COVID-19 cases vs. days

Inferences:

1. From the plot we can infer that the cases increases day by day after some time it the cases decreases and then again the cases increases.
2. The reason for the increasing in cases due to the spread of the disease and then decreases due to some isolation in society and then again increases due to more spread of the disease.
3. the plot the duration of first wave is approximately 5 months and second wave is approximatrly 4 months.

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

Inferences:

1. By the value of pearson correlation coefficient we can say that both the data sequences are highly correlated to each other.

2. We generally expect observations (here number of COVID-19 cases) on days one after the other to be similar as they are highly correlated to each other so there is no much difference between the one day after the other.

c.

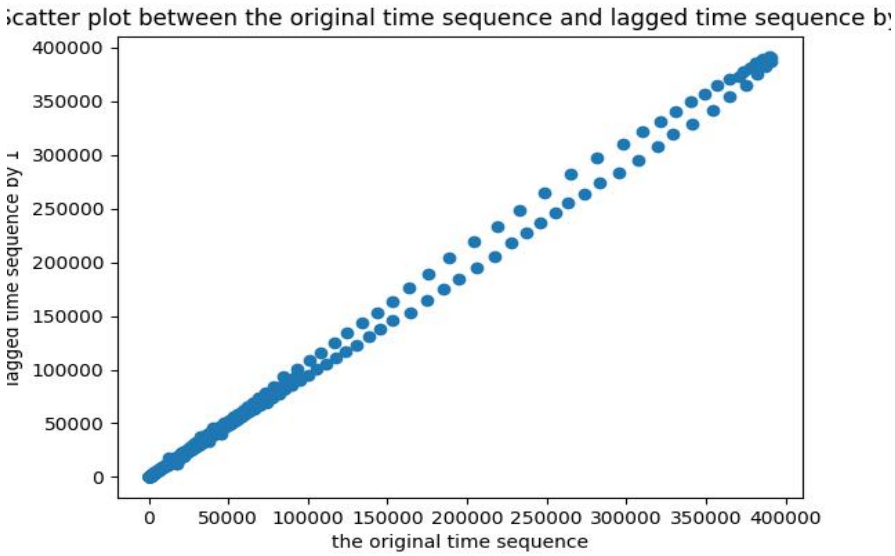


Figure 2 Scatter plot one day lagged sequence vs. given time sequence

Inferences:

1. Based on nature of the spread of data points, we can say that it the nature of correlation between the two sequences are highly correlated.
2. the scatter plot seems to obey the nature reflected by Pearson's correlation coefficient calculated in 1b.
3. In this they are highly correlated as the one sequence leads to increase the sequence of the other data and they are approximately at 45 degrees so they are very much related to each other.

d.

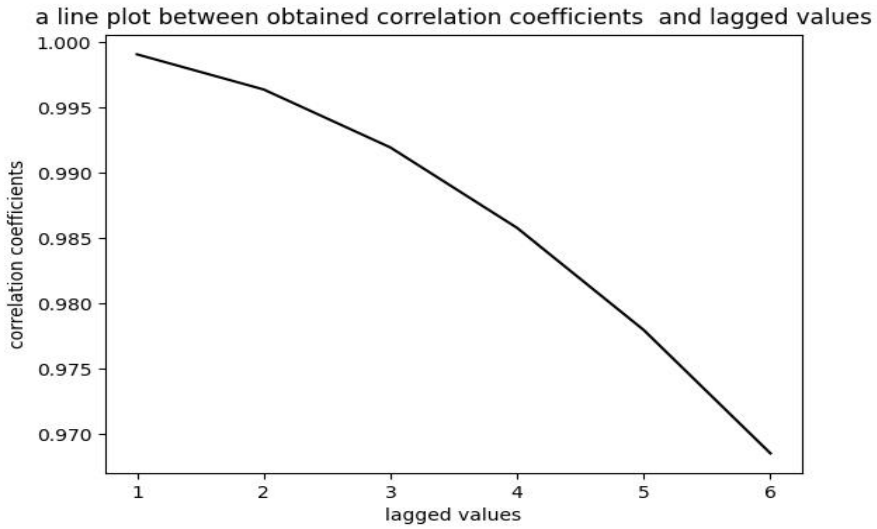


Figure 3 Correlation coefficient vs. lags in given sequence

Inferences:

1. As the lag values increases the correlation between the original and the respective lag series decreases.
2. The correlation decreases with increases in lag series as the dependency on the past values decreases.

e.

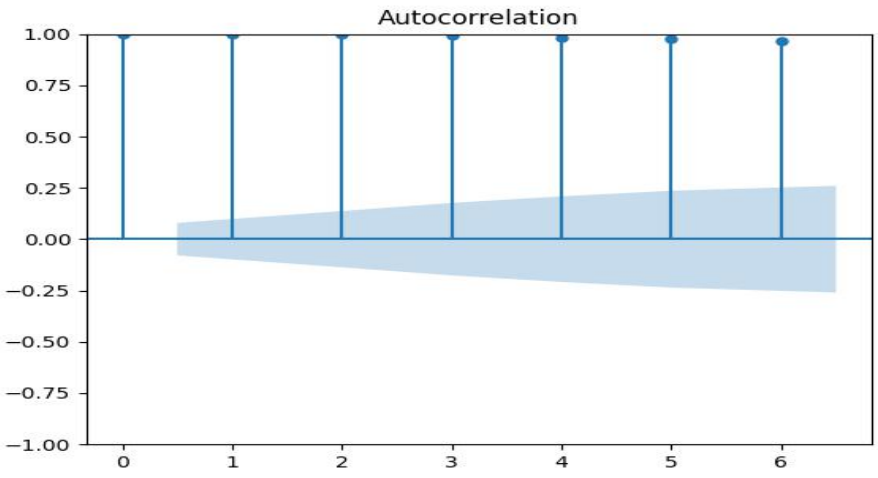


Figure 4 Correlation coefficient vs. lags in given sequence generated using 'plot_acf' function

Inferences:

1. As the lag values increases the correlation between the original and the respective lag series decreases.
2. The correlation decreases with increases in lag series as the dependency on the past values decreases

2

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

b. i.

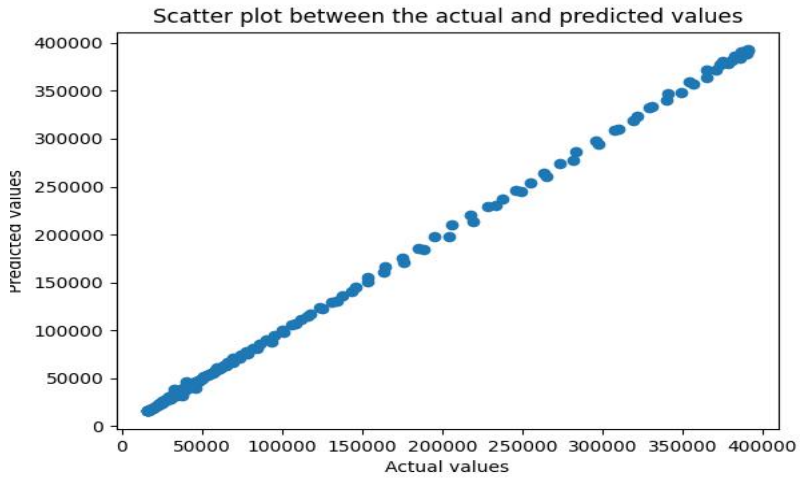
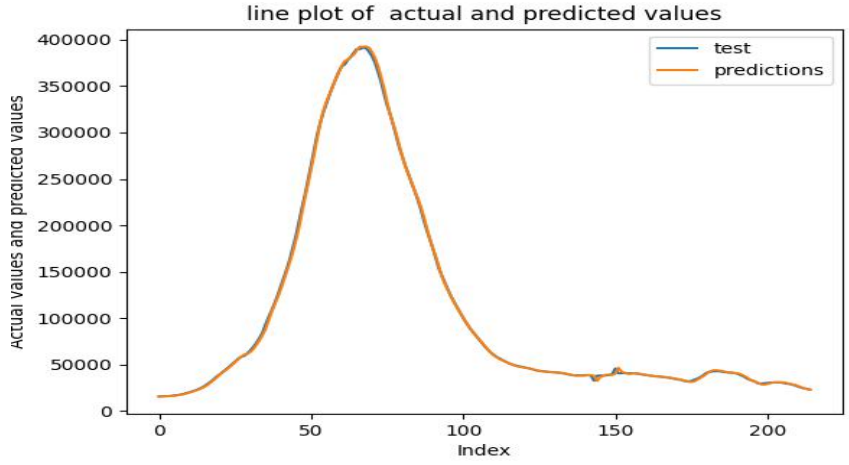


Figure 5 Scatter plot actual vs. predicted values

Inferences:

1. From the nature of the spread of data points, the nature of the correlation between the two sequences is highly correlated to each other.
2. the scatter plot seems to obey the nature reflected by Pearson's correlation coefficient calculated in 1.b.
3. In this they are highly correlated as the one sequence leads to increase the sequence of the other data and they are approximately at 45 degrees so they are very much related to each other.

ii.



Predicted vs actual observations

Inferences:

1. The predicted and the original data are coinciding to each other which indicates they are highly correlated to each other so that they can be sued for future prediction.

iii. The RMSE(\%) and MAPE between predicted power consumed for test data and original values for test data are **1.825 and 1.575** .

Inferences:

1. As the error is very low for both the process i.e., RMSE and MAPE so we can say that the predicted values are accurate
2. As the error is low we can use the data for future prediction as there is very less deviation between the predicted and original 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**

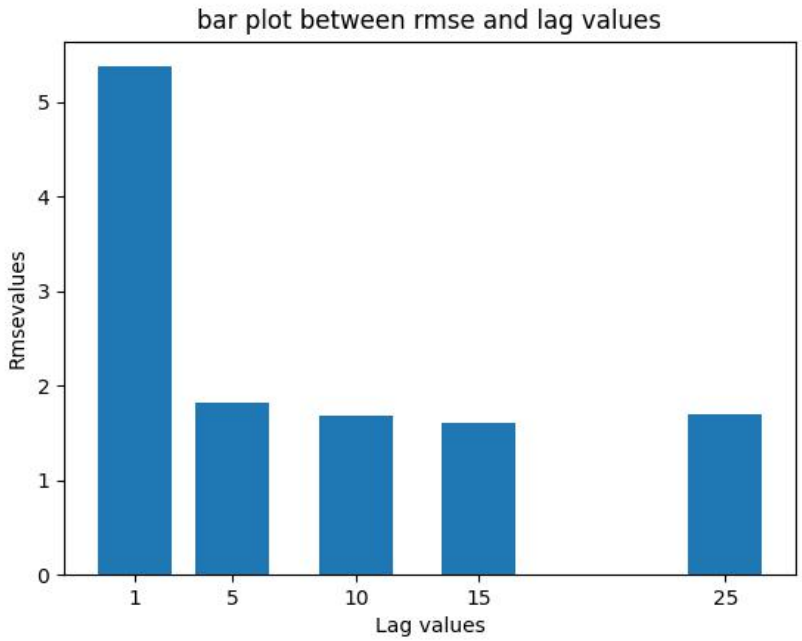


Figure 7 RMSE(%) vs. time lag

Inferences:

1. As the time lag increases the RMSE values decreases but on highly increasing the lag value the RMSE values again increases.
2. As the time lag increases the dependency of present value increases on past values so that the RMSE decreases but too much dependency also leads to increase in RMSE Values.

.

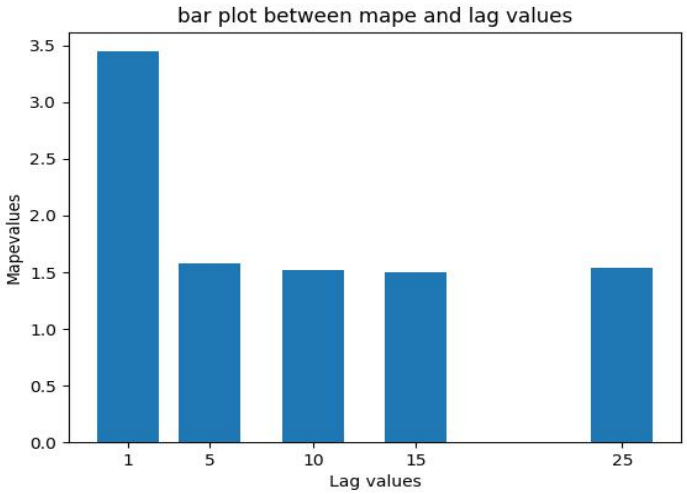


Figure 8 MAPE vs. time lag

Inferences:

1. As the time lag increases the MAPE values decreases but on highly increasing the lag value the MAPE values again increases.
2. As the time lag increases the dependency of present value increases on past values so that the MAPE decreases but too much dependency also leads to increase in MAPE Values

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. Based upon the RMSE(%) and MAPE value, heuristics for calculating the optimal number of lags improve the prediction accuracy.
2. The RMSE with optimal lag is around 5.3-1.6 range where our calculated RMSE is 1759 which is within the range and the MAPE value with optimal lag is around 3.446-1.496 where as heuristic MAPE is 2.206 which is also in the range of MAPE.