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

Student's Name: Kuldeep Jain Dugar **Branch:**

Roll Number: B20112 **CSE**

Mobile No: 8986388665

1 a.

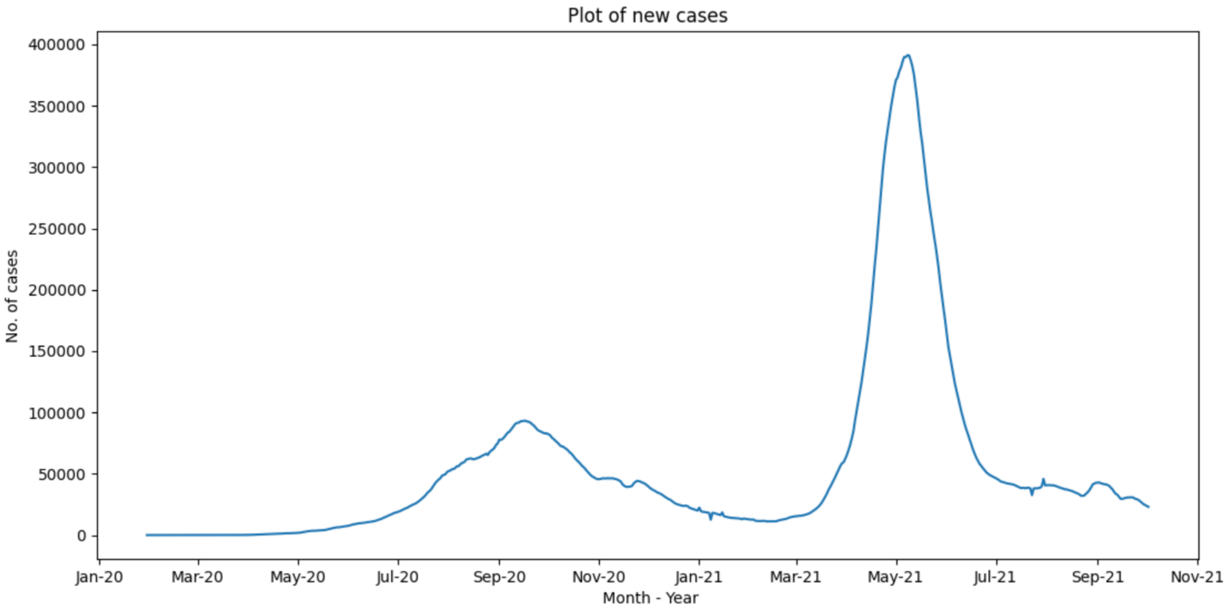


Figure 1 No. of COVID-19 cases vs. days

Inferences:

1. In the plot most values are similar to the previous one except the ones at the time of a wave
2. The trend in number of new cases depends on the trend in the past dates (whether increasing, decreasing or constant)
3. First wave – August 2020 to October 2020.
4. Second Wave – April 2021 to June 2021

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

Inferences:

1. From the value of Pearson's correlation coefficient, we infer that it is very strong.
2. We generally expect observations (here number of COVID-19 cases) on days one after the other to be similar. It holds to a great extent as we can see the value of pearson coefficient to be almost 1.
3. Reason- our assumption that the observations at previous time steps are significant to predict the value at the next time step is true in this case

c.

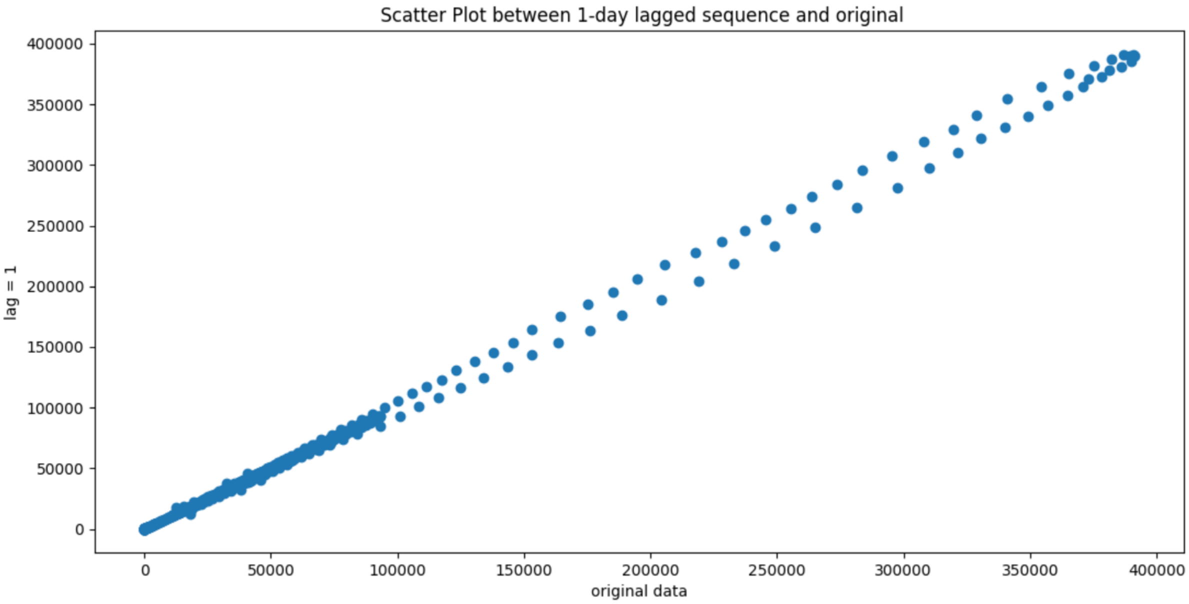


Figure 2 Scatter plot one day lagged sequence vs. given time sequence

Inferences:

1. The nature is that both are strongly related due to high coefficient.
2. Does the scatter plot seem to obey the nature reflected by Pearson's correlation coefficient calculated in 1.b? - YES
3. Accept a few almost the graph follows y = x.

d.

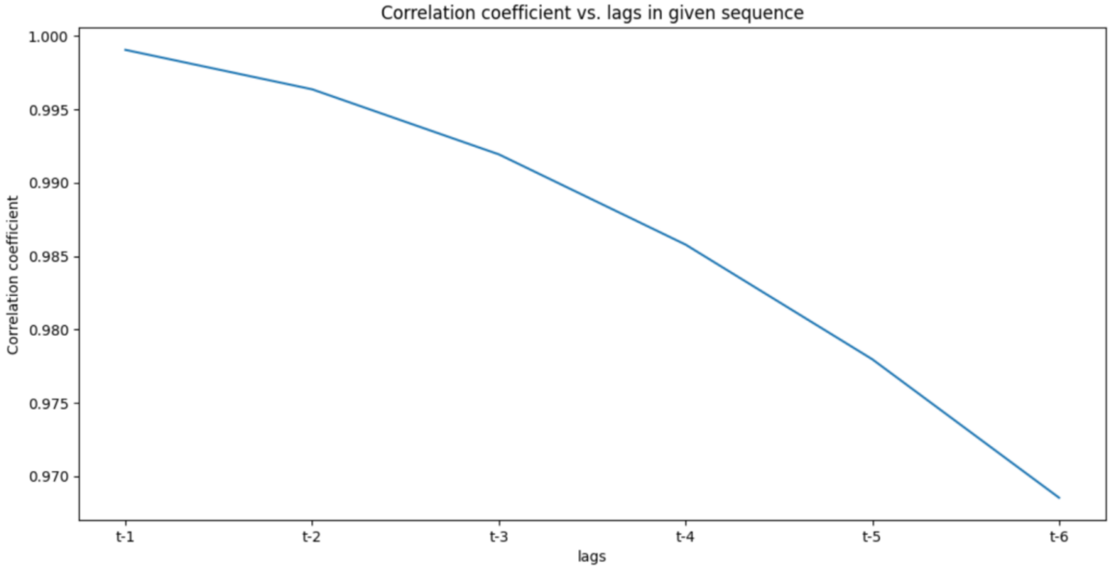


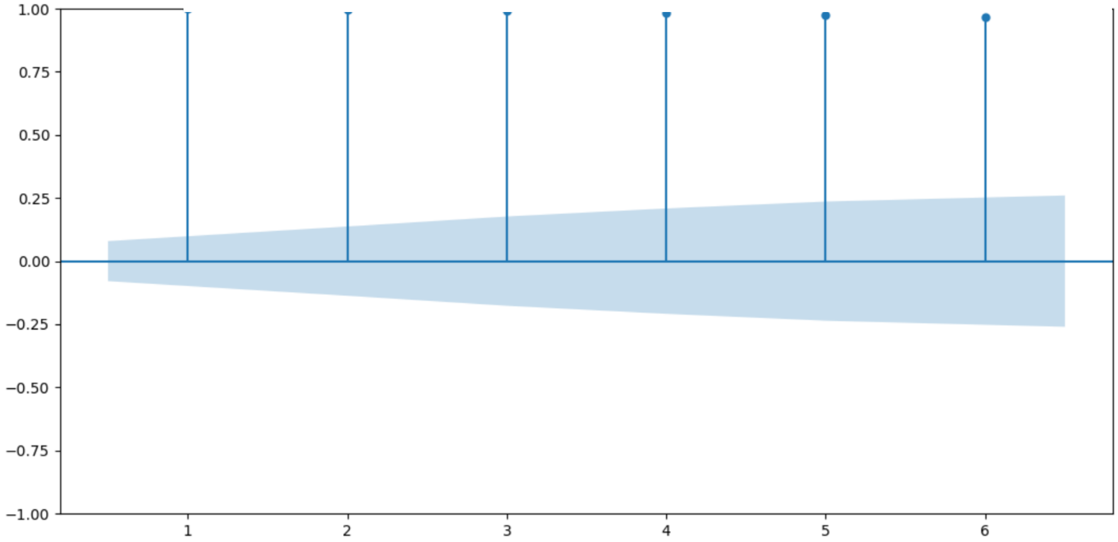
Figure 3 Correlation coefficient vs. lags in given sequence

Inference

1. Trend of correlation coefficient value with respect to increase in lags in time sequence.- Decreasing
2. Because the series hang out at the ends and do not overlap.

e.

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



Inferences:

1. trend of correlation coefficient value with respect to lags in time sequence.- decreasing
2. Because the series hang out at the ends and do not overlap

2

a. The coefficients obtained from the AR model are;
[5.995, 1.036, 2.617, 2.75, -1.753, -1.52]

b. **i.**

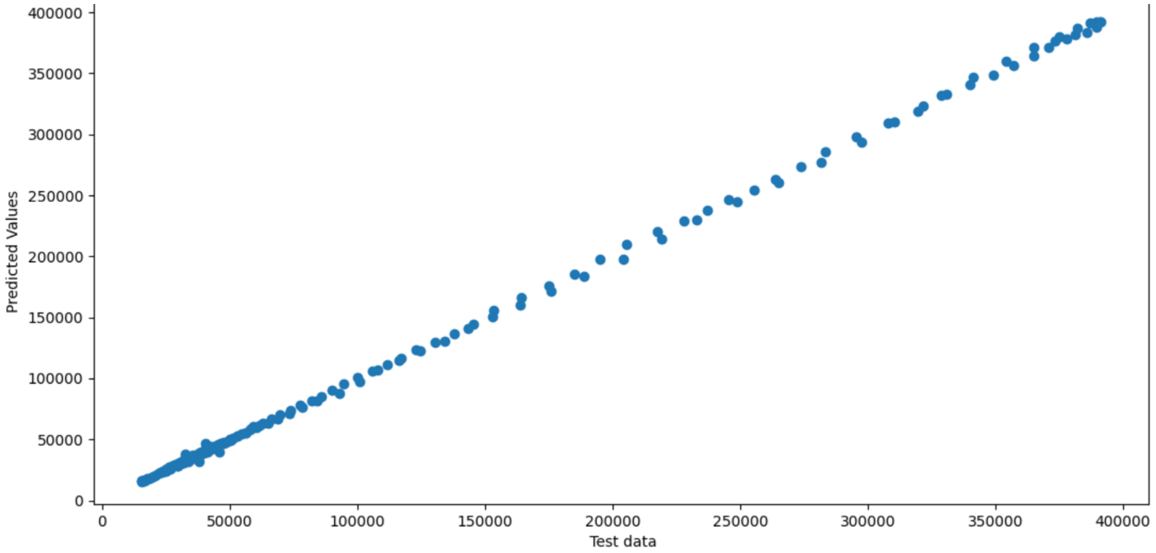


Figure 5 Scatter plot actual vs. predicted values

Inferences:

1. From the nature of the spread of data points, the nature of the correlation between the two sequences is very strong
2. Does the scatter plot seem to obey the nature reflected by Pearson's correlation coefficient calculated in 1.b?- YES
3. Since the lag has increased to 5 it is more approproiate and better fit.

ii.

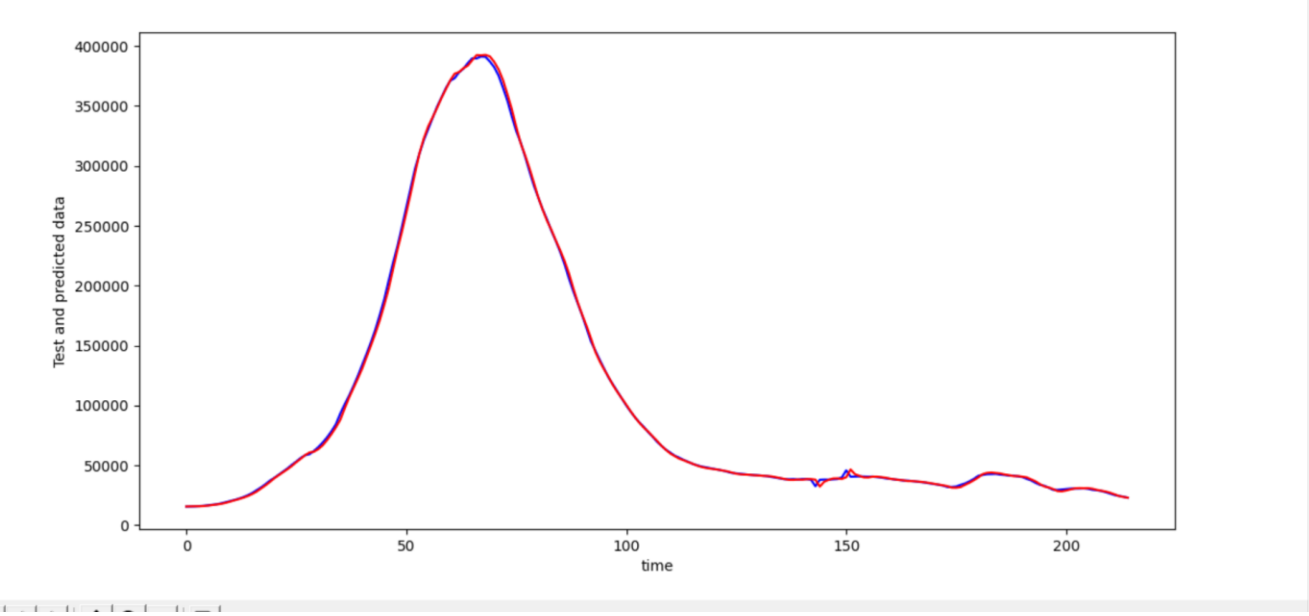


Figure 6 Predicted test data time sequence- red vs. original test data sequence-blue

Inferences:

1. From the plot of predicted test data time sequence vs. original test data sequence we can comment the model is reliable for future predictions with suitable reasons.

iii.

The RMSE(\%) and MAPE between predicted power consumed for test data and original values for test data are . 1.824 , 1.574

Inferences:

1. From the value of RMSE(%) and MAPE value we can say the model for the given time series is. Very accurate
2. The trend of new cases depends on the past cases

3 [5.372948489180813, 1.8247684769389676, 1.685531934880951, 1.6119348114120344, 1.7033914119397975]

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

Lag value	RMSE (%)	MAPE
1	5.372	**3.44**
5	1.824	**1.57**
10	1.685	**1.52**
15	1.611	**1.49**
25	1.703	**1.53**

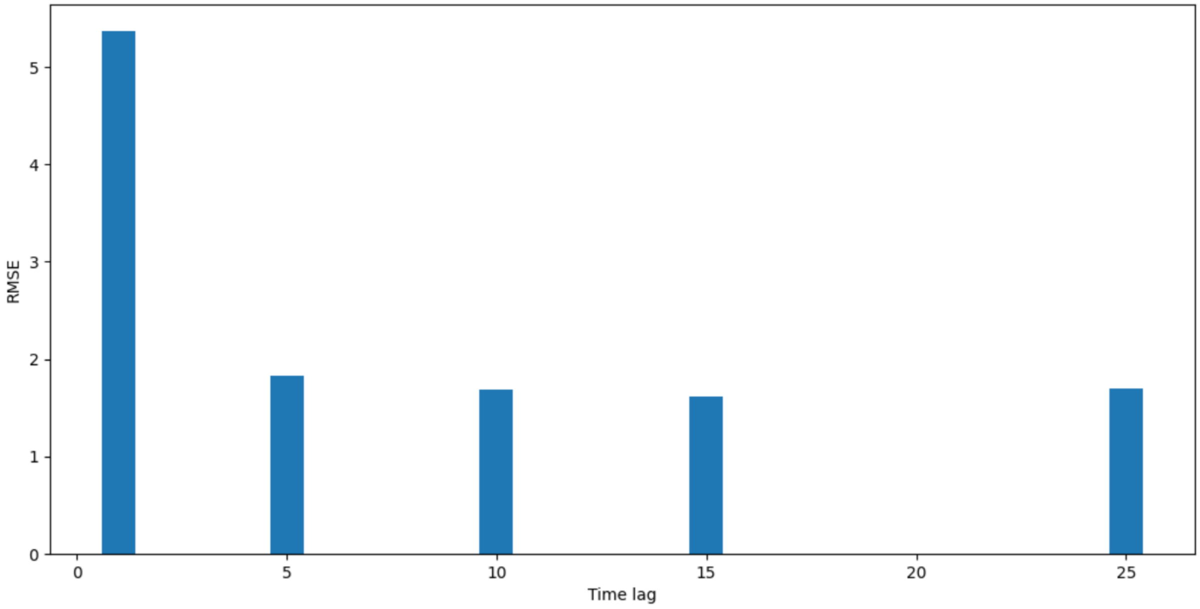


Figure 7 RMSE(%) vs. time lag

Inferences:

1. The RMSE(%) rapidly from 1 to 5 but then decreases gradually with respect to increase in lags in time sequence.
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.

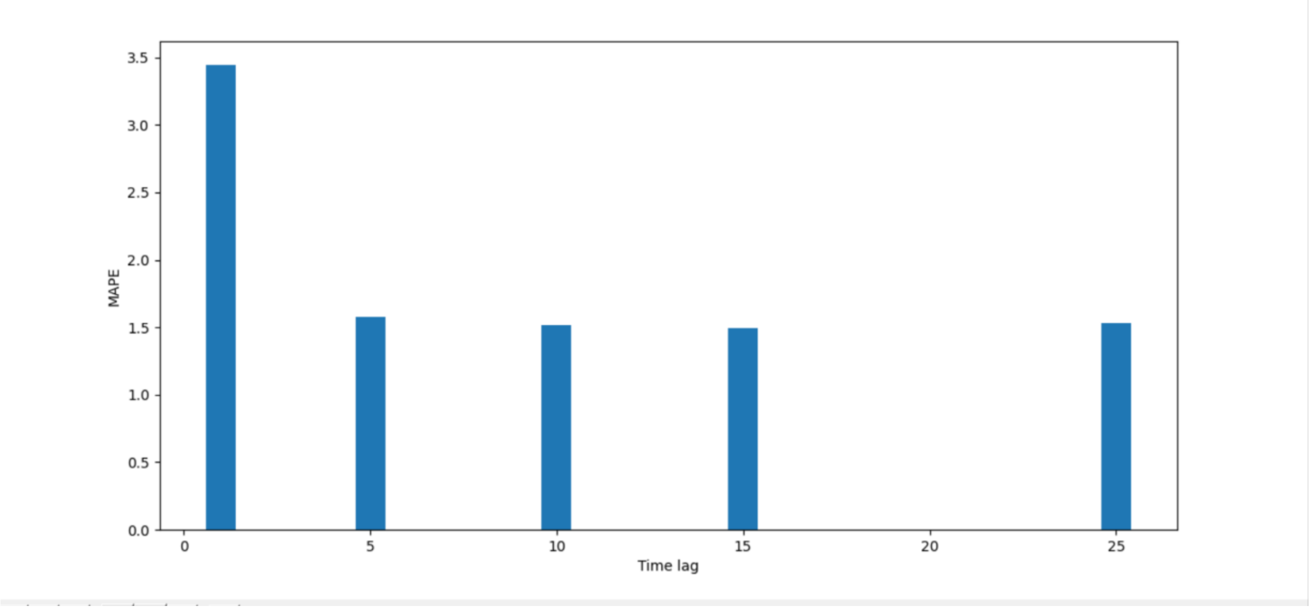


Figure 8 MAPE vs. time lag

Inferences:

1. The MAPE value rapidly from 1 to 5 but then decreases gradually with respect to increase in lags in time sequence

   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 78

The RMSE(%) and MAPE value between test data time sequence and original test data sequence are

1.768,2.0752

Inferences:

1. Based upon the RMSE(%) and MAPE value, did heuristics for calculating the optimal number of lags improve the prediction accuracy of the model?- no.
2. Because as we keep increasing the lag, after certain time the pattern of RMSE vs lag will become random and we can also see that as the observations are made for every day AR(78) doesn't make sense than that of a lag of around one day
3. It was 1.70 for n = 25 and on increasing it to 78, the rmsevalue has become 1.768

- .