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

Student's Name: Rustam Narayan **Mobile No: 8603861159**

Roll Number: B20128 **Branch:CSE**

1 a.

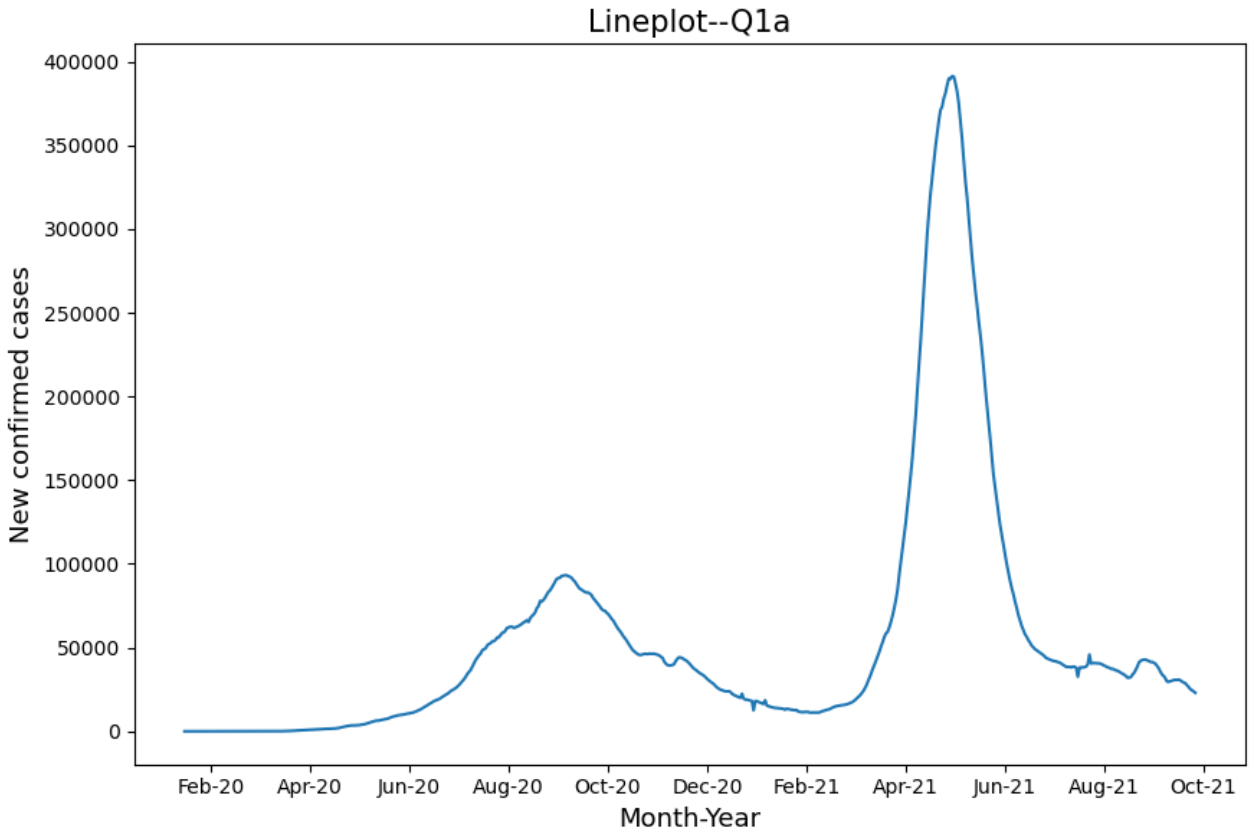

Figure 1 No. of COVID-19 cases vs. days

Inferences:

1. Yes, the day after another has similar trend in fresh number of covid cases.
2. Since from the graph we observe that there is no drastic change in number of cases in a day and around it instead the number of cases grow slowly and reduce slowly after first and second wave.
3. July 2020 to October 2020 was the duration of first wave.

b. The value of the Pearson's correlation coefficient is 0.991 (with one-day lagged time series).

Inferences:

1. From the value of Pearson's correlation coefficient, there is very strong the degree of correlation between the two-time sequences.
2. We generally expect observations (here number of COVID-19 cases) on days one after the other to be similar. It holds to a greater extent and the exception was the first and second wave.
3. It is because there is autocorrelation to some heuristic number of lag in time-series data.

c.

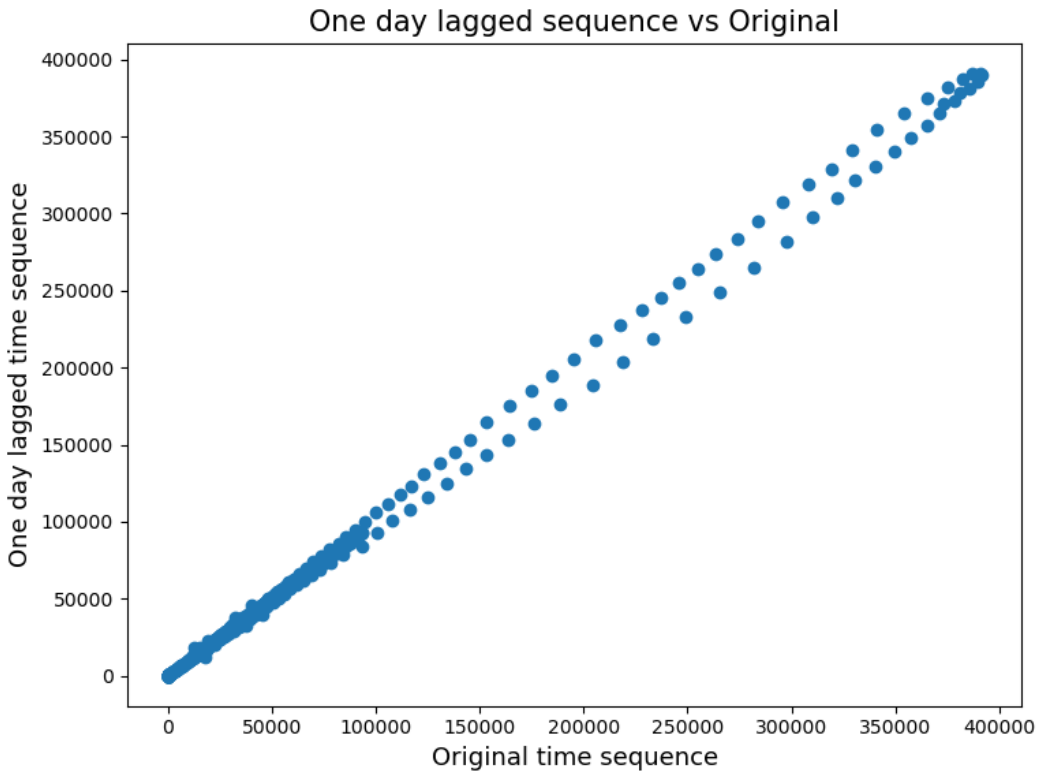

Figure 2 Scatter plot one day lagged sequence vs. given time sequence

Inferences:

1. From the nature of the spread of data points, there is strong positive correlation between the two sequences.
2. Yes, the scatter plot seems to obey the nature reflected by Pearson's correlation coefficient calculated in 1.b.
3. It is because there is autocorrelation to some heuristic number of lag in time-series data.

d.

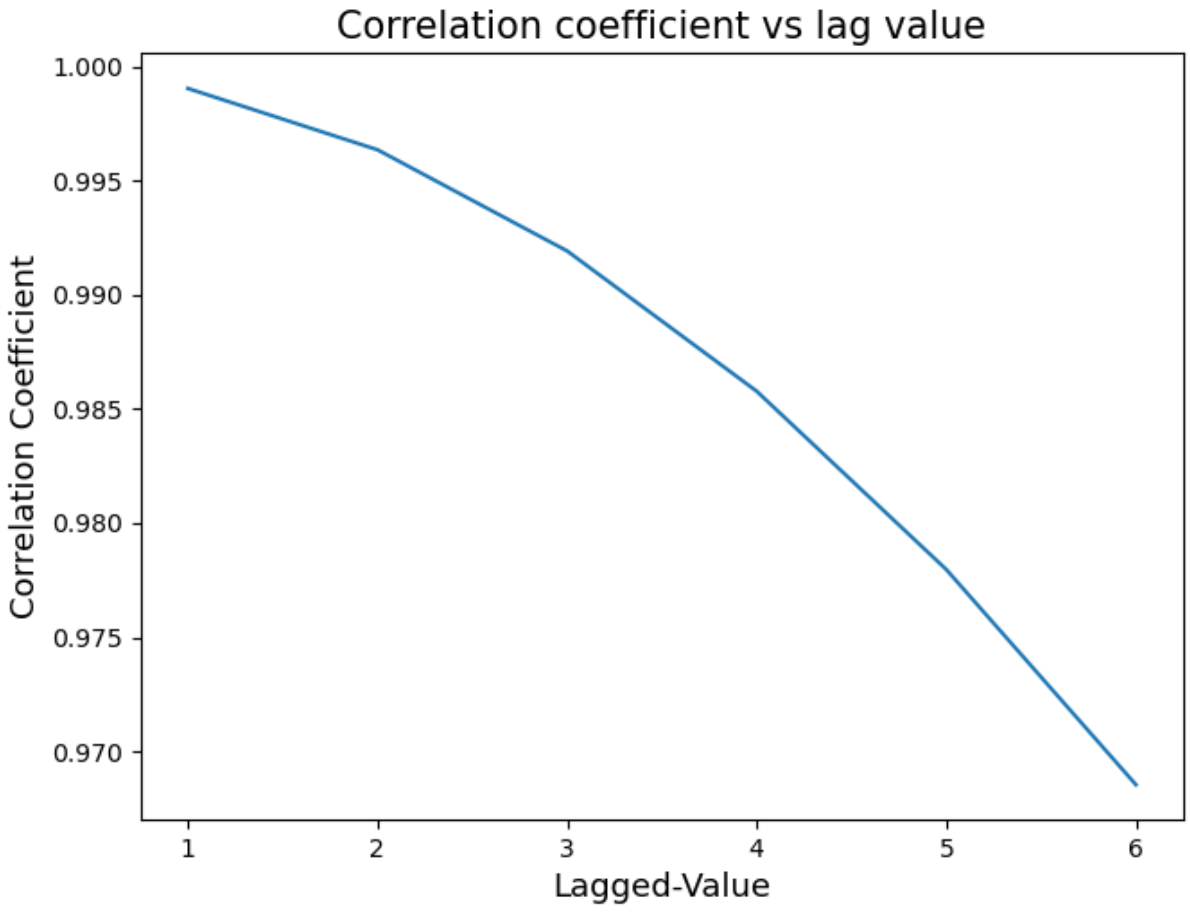

Figure 3 Correlation coefficient vs. lags in given sequence

Inferences:

1. Correlation coefficient value decreases with respect to increase in lags in time sequence.
2. Since covid data is a time series data and there is some trend in time series data and due to which we can predict the next value based on the previous entries and this is possible due to autocorrelation in time series data which decreases slowly with the value of lag .

e.

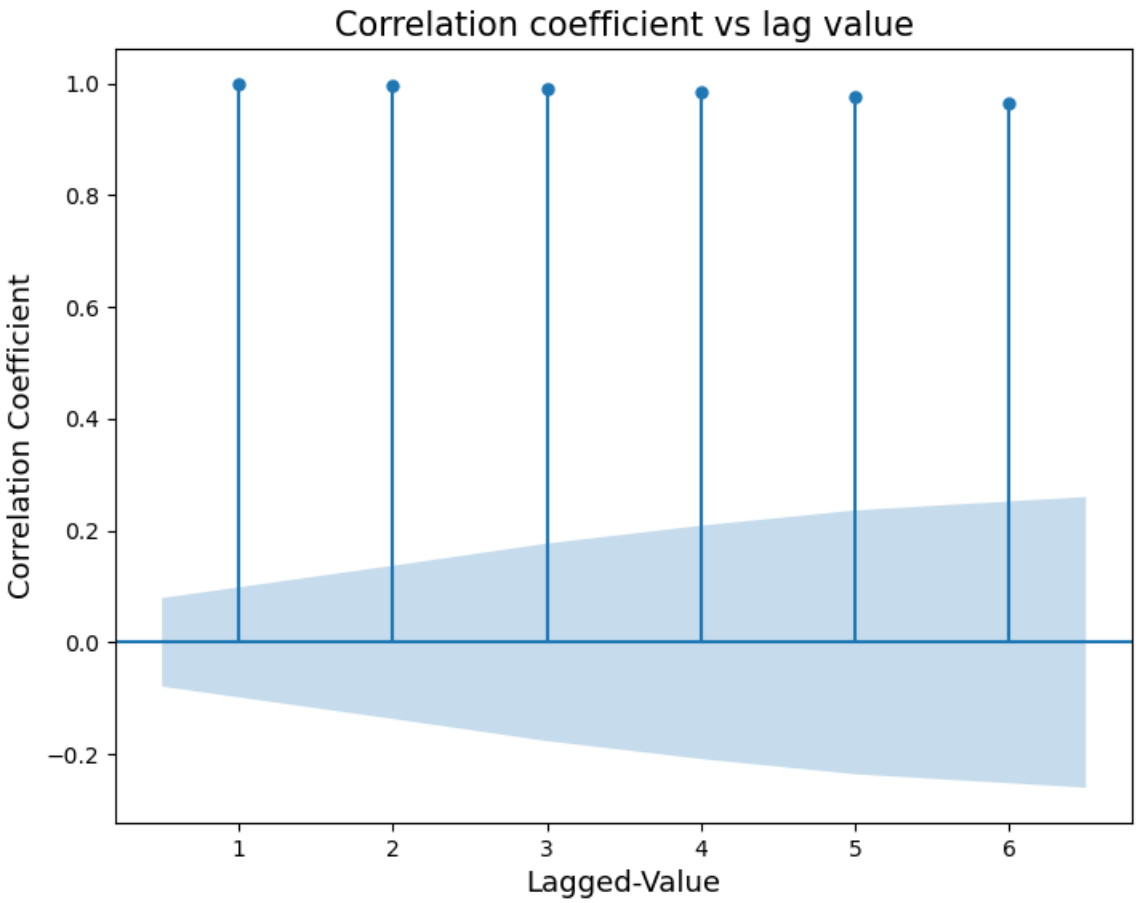

Figure 4 Correlation coefficient vs. lags in given sequence generated using 'plot_acf' function

Inferences:

1. Correlation coefficient value decreases very slowly with respect to lags in time sequence.

2. Since covid data is a time series data and there is some trend in time series data and due to which we can predict the next value based on the previous entries and this is possible due to autocorrelation in time series data which decreases slowly with the value of lag.

2

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

b. i.

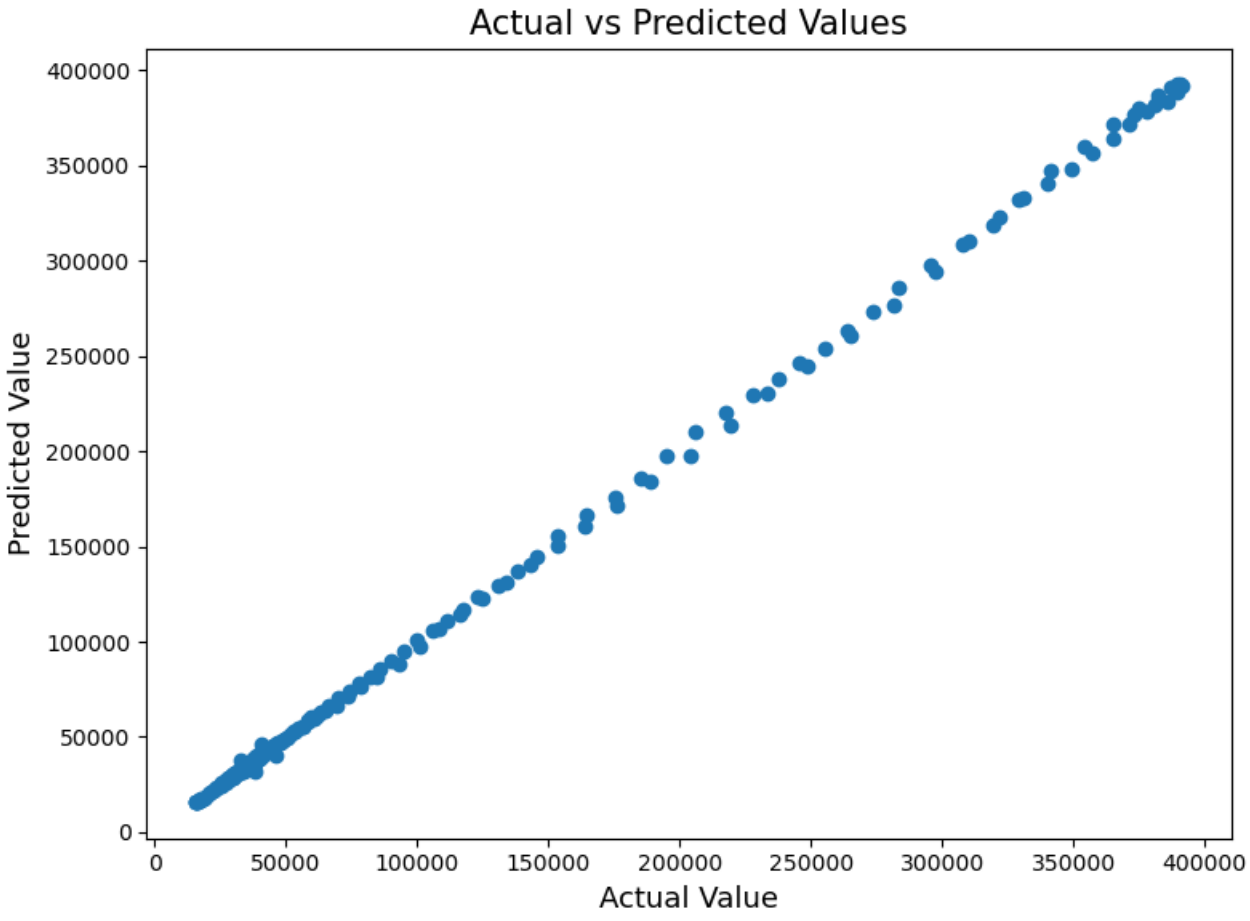

Figure 5 Scatter plot actual vs. predicted values

Inferences:

1. From the nature of the spread of data points, there is strong positive correlation between the two sequences.
2. Yes, the scatter plot seems to obey the nature reflected by Pearson's correlation coefficient calculated in 1.b.
3. It is because there is autocorrelation to some heuristic number of lag in time-series data.

ii.

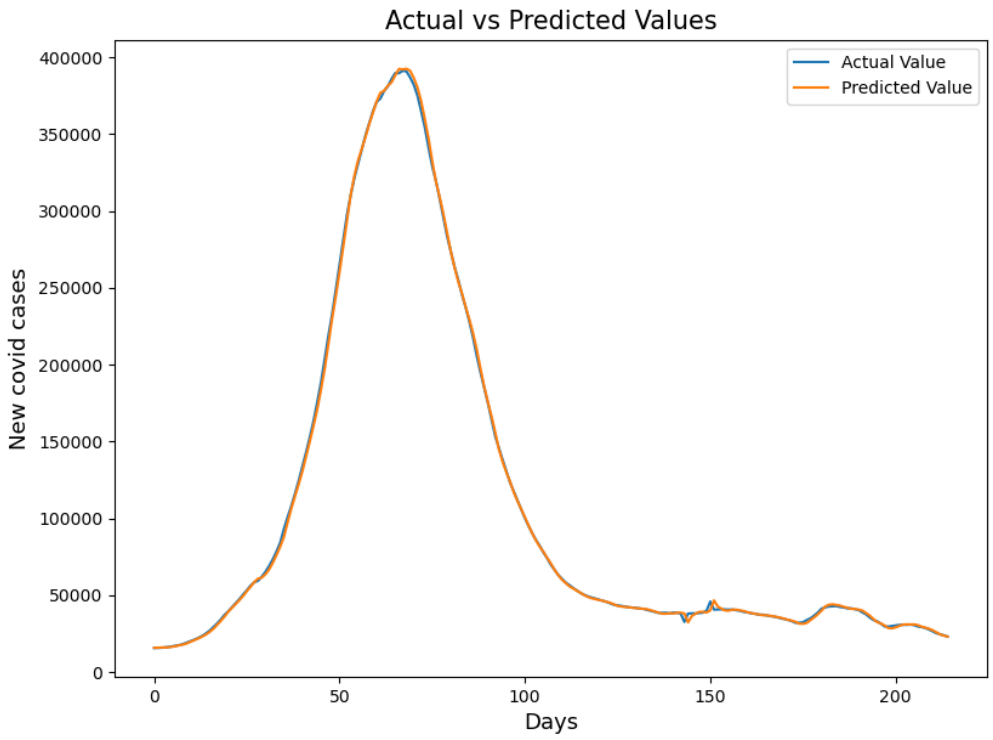

Figure 6 Predicted test data time sequence vs. original test data sequence

Inferences:

1. From the plot of predicted test data time sequence vs. original test data sequence we observe that the predicted values almost exactly match with the original values. Based on this plot we can say that this method is highly reliable.

iii.

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

Inferences:

1. From the value of RMSE(\%) and MAPE value we can comment the autoregression method is highly accurate in predicting number of covid cases.
2. It is because there is autocorrelation to some heuristic number of lag in given time-series data and autoregression method of prediction makes the same assumption in prediction.

3

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

Lag value	RMSE (%)	MAPE
1	5.373	3.446
5	1.825	1.574
10	1.686	1.519
15	1.612	1.496
25	1.703	1.535

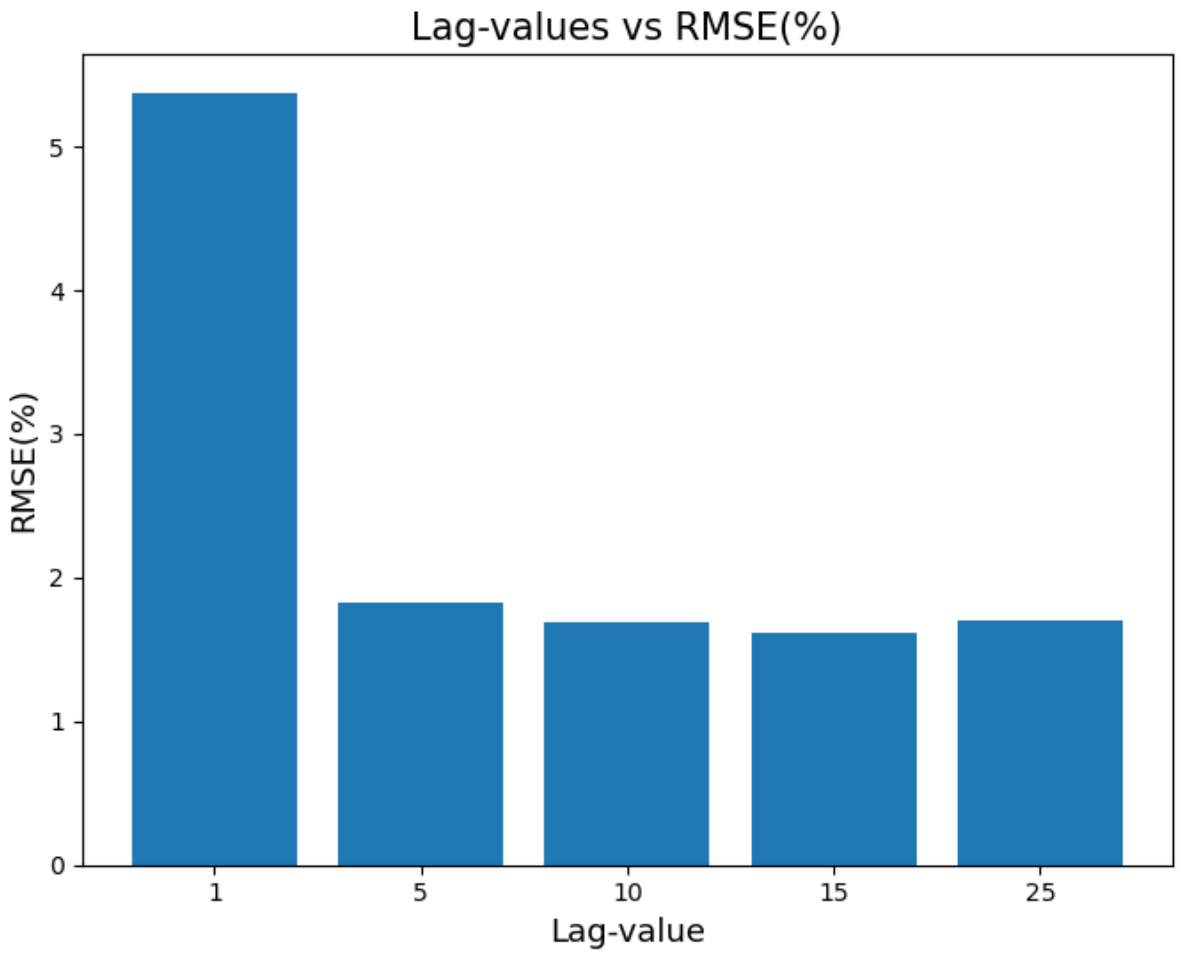

Figure 7 RMSE(%) vs. time lag

Inferences:

1. RMSE(%) decreases slowly with respect to increase in lags in time sequence.

2. The autoregression method becomes more accurate if we include a greater number of inputs in model building and so as the lag-value increases the prediction error decreases.

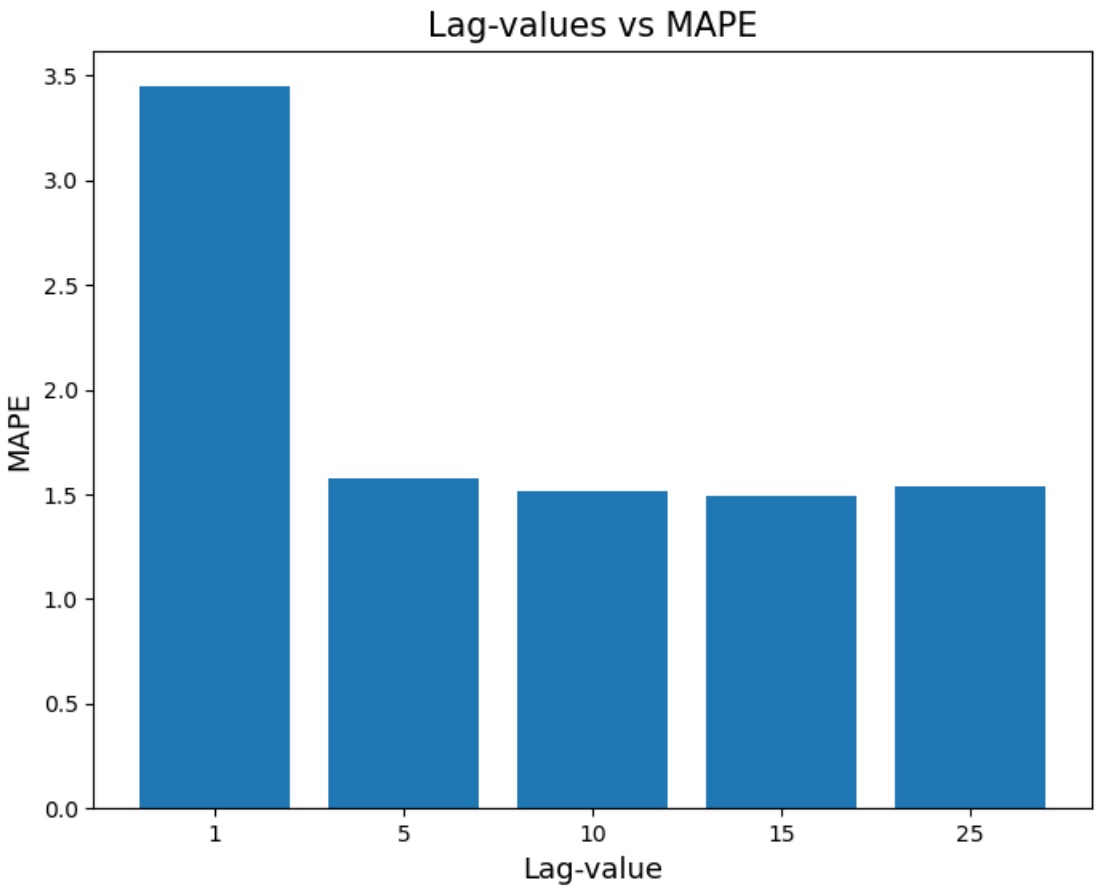

Figure 8 MAPE vs. time lag

Inferences:

1. MAPE value decreases with respect to increase in lags in time sequence.
2. The autoregression method becomes more accurate if we include a greater number of inputs in model building and so as the lag-value increases the prediction error decreases.

4

1. The heuristic value for the optimal number of lags is 77.

2. 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 did not improve the prediction accuracy of the model.
2. It is because the value of $2/\sqrt{T}$ in this case is 0.1003 which is very low and if we consider the lag-series with low correlation into account for prediction of next step then the prediction error will increase.
3. On comparing the prediction accuracies obtained without and with the heuristic, we observe that the optimal lag-value is 15.