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

Name: Pranjal Sharma

Roll Number: B20305

Mobile No: 7374065064

Branch:Mechanical Engineering

1 a.

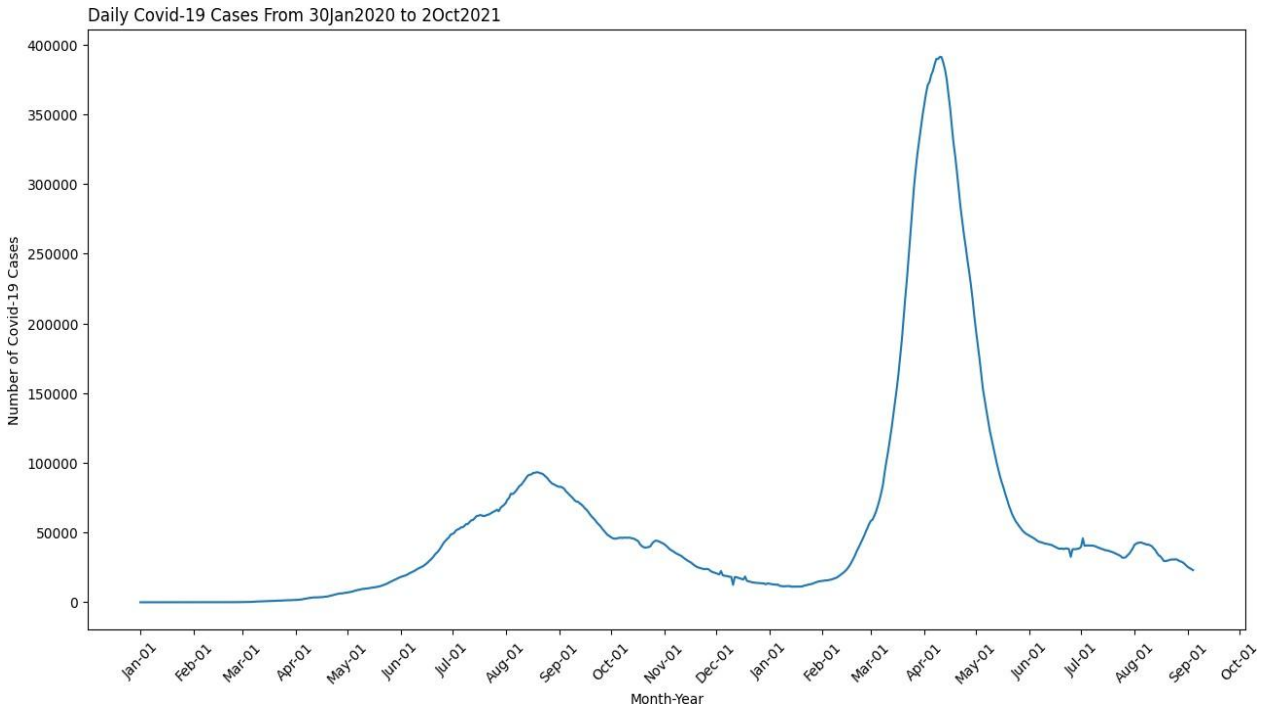

Figure 1 No. of COVID-19 cases vs. days

Inferences:

1. No, since during first and second wave number of covid cases are increasing rapidly and after the peak of wave cases are also decreasing rapidly.
2. Duration of first wave was around **8 months** and second wave was around **5 months**.

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

Inferences:

1. Two time series are very strongly correlated with each other. It means that future values are affected by past values.

2. Observations on days one after the other are very similar since, the value of correlation coefficient is very high **(0.999)** it means future observation will be high dependent on the past observations.

c.

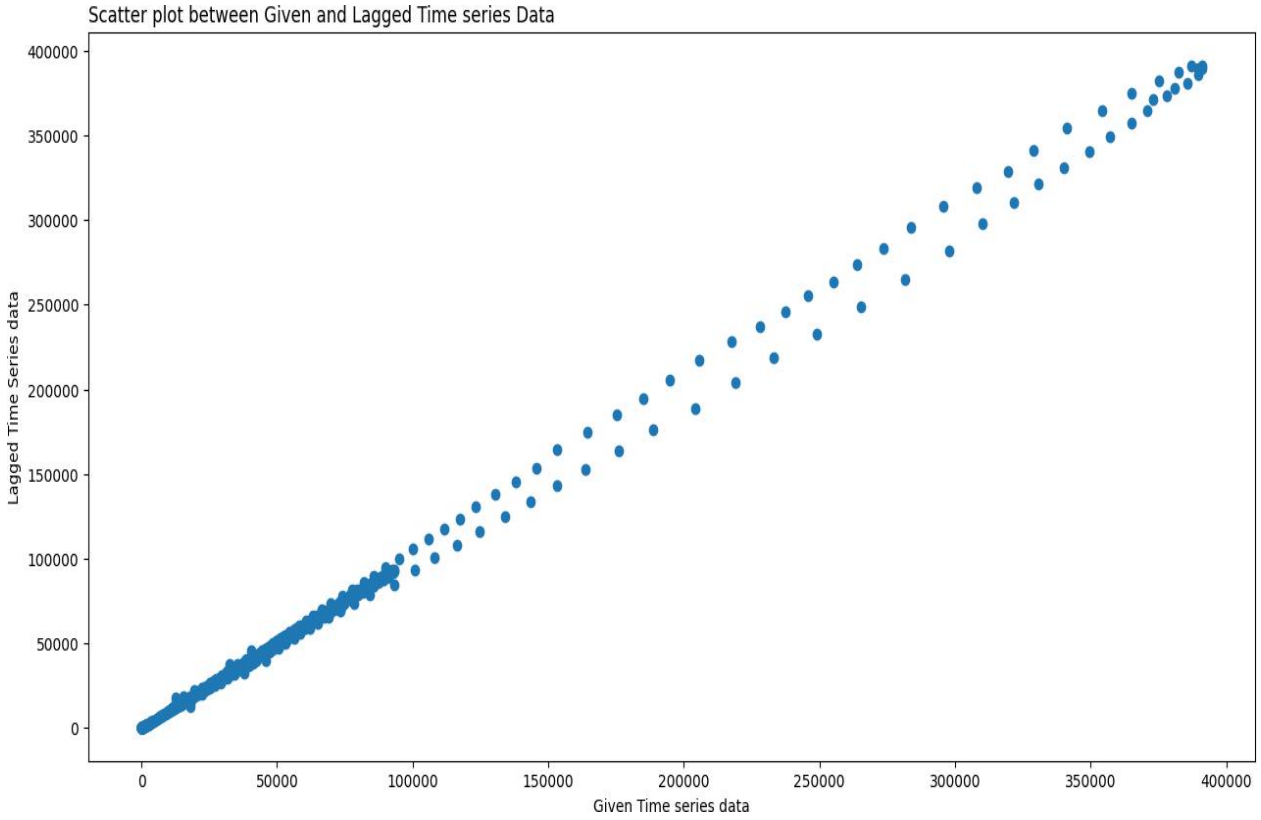

Figure 2 Scatter plot one day lagged sequence vs. given time sequence

Inferences:

1. Correlation between to variables is **very strongly** and **positive**, means two variables are highly dependent on each other.
2. Yes, completely.
3. Since, datapoints in the graph are almost in a straight with slop 1 and originating from 0, which means if one variable is increasing then other variable will also increase by almost same value. Which shows that variables are highly correlated with each other.

d.

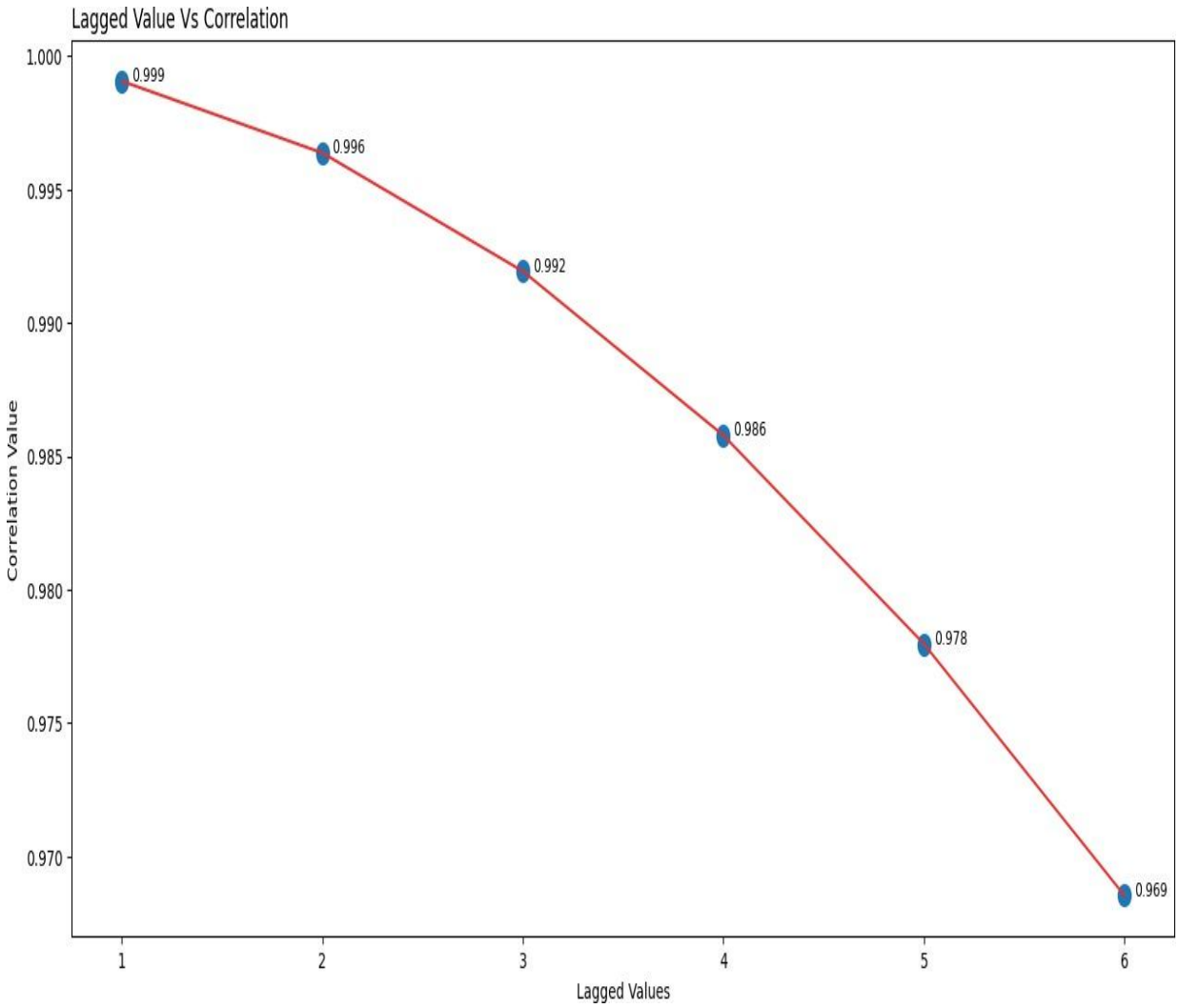

Figure 3 Correlation coefficient vs. lags in given sequence

Inferences:

1. Correlation coefficient is **decreases** as lag value **increases**.
2. When data have a trend, the autocorrelations for small lags tend to be large and positive because observations nearby in time are also nearby in size. So, the ACF of trended time series tend to have positive values that slowly decrease as the lags increase.

e.

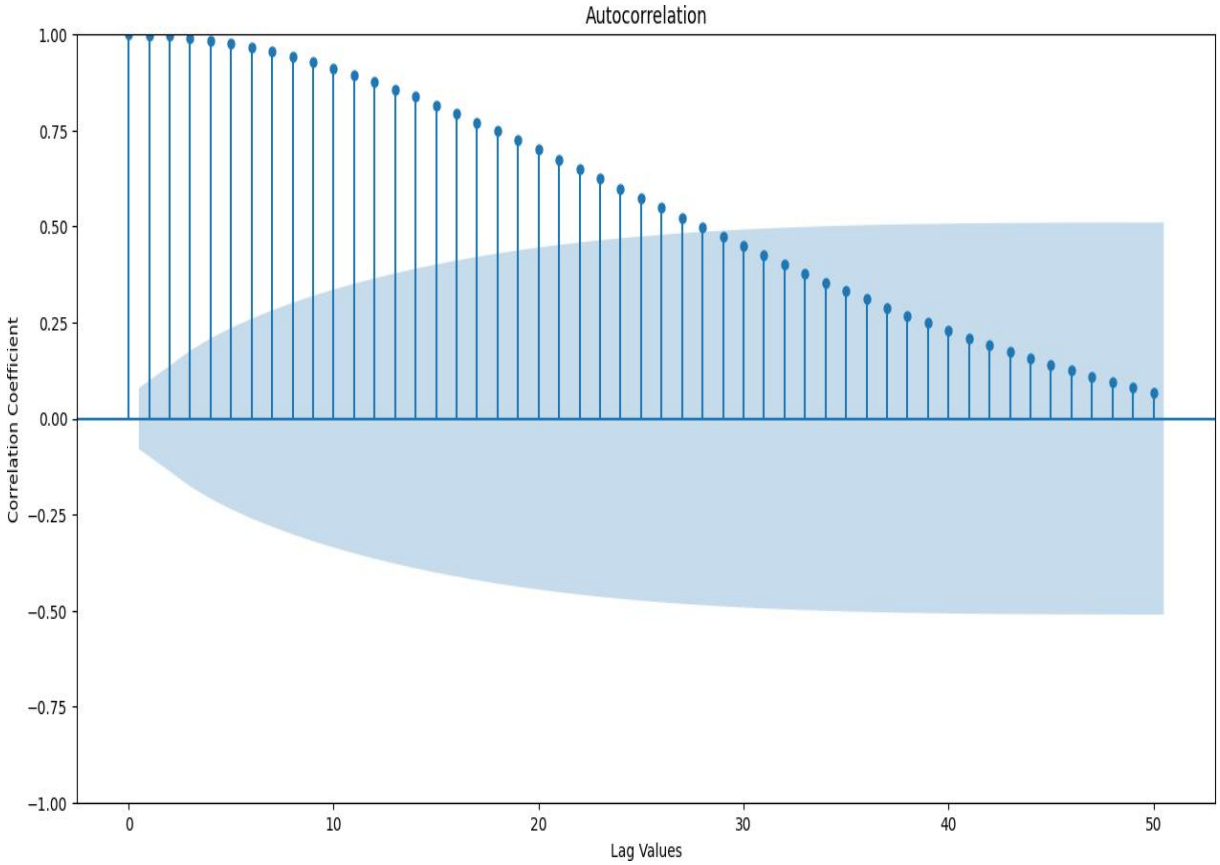

Figure 4 Correlation coefficient vs. lags in given sequence generated using 'plot_acf' function

Inferences:

1. Value of correlation coefficient **decreases** as lag value **increases**.
2. Same reason as explained above.

2

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

b. i.

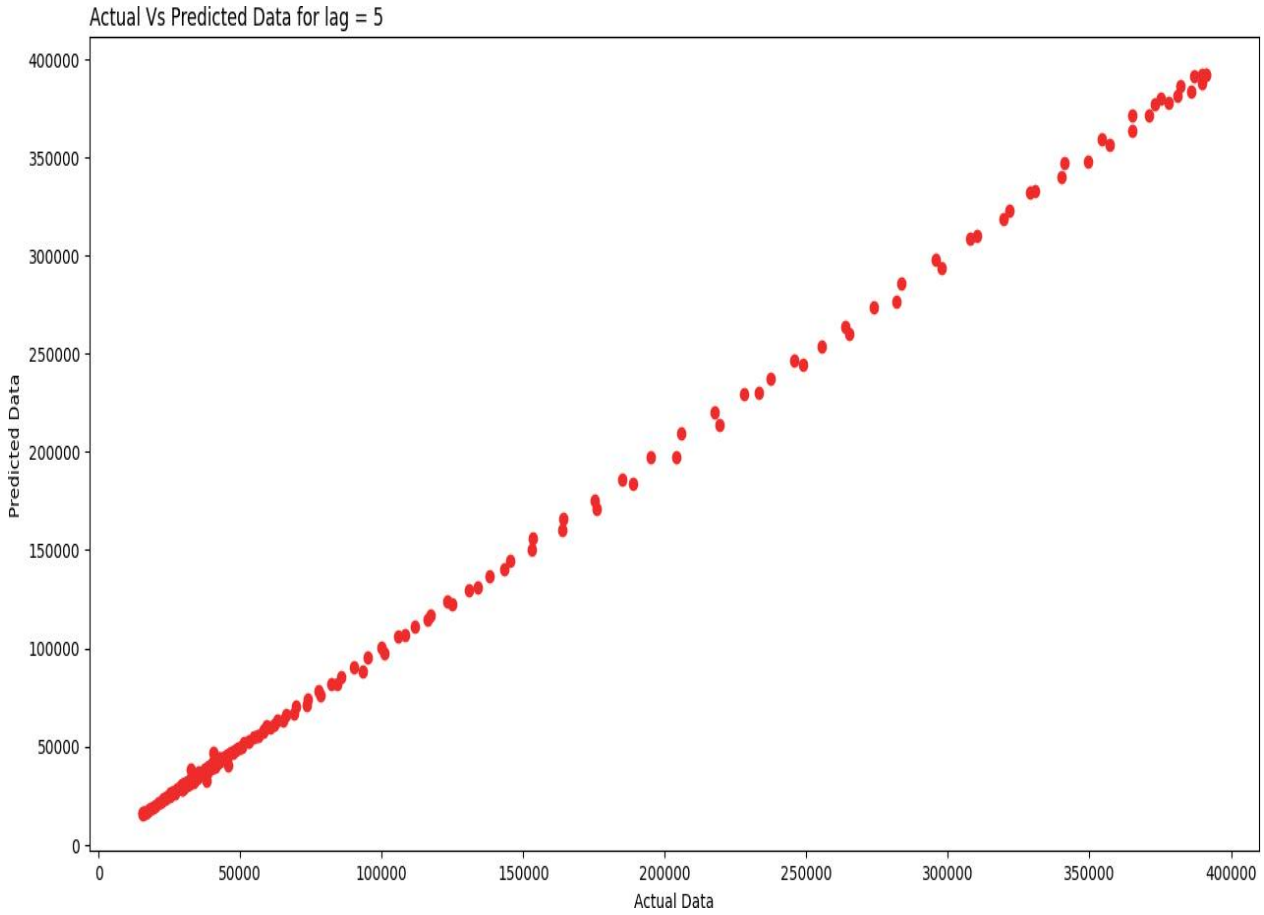

Figure 5 Scatter plot actual vs. predicted values

Inferences:

1. Two variables have very strong positive correlation.
2. Yes, completely.
3. From the graph we can see if one variable is increasing then other is also increasing, which means that they are highly correlated.

ii.

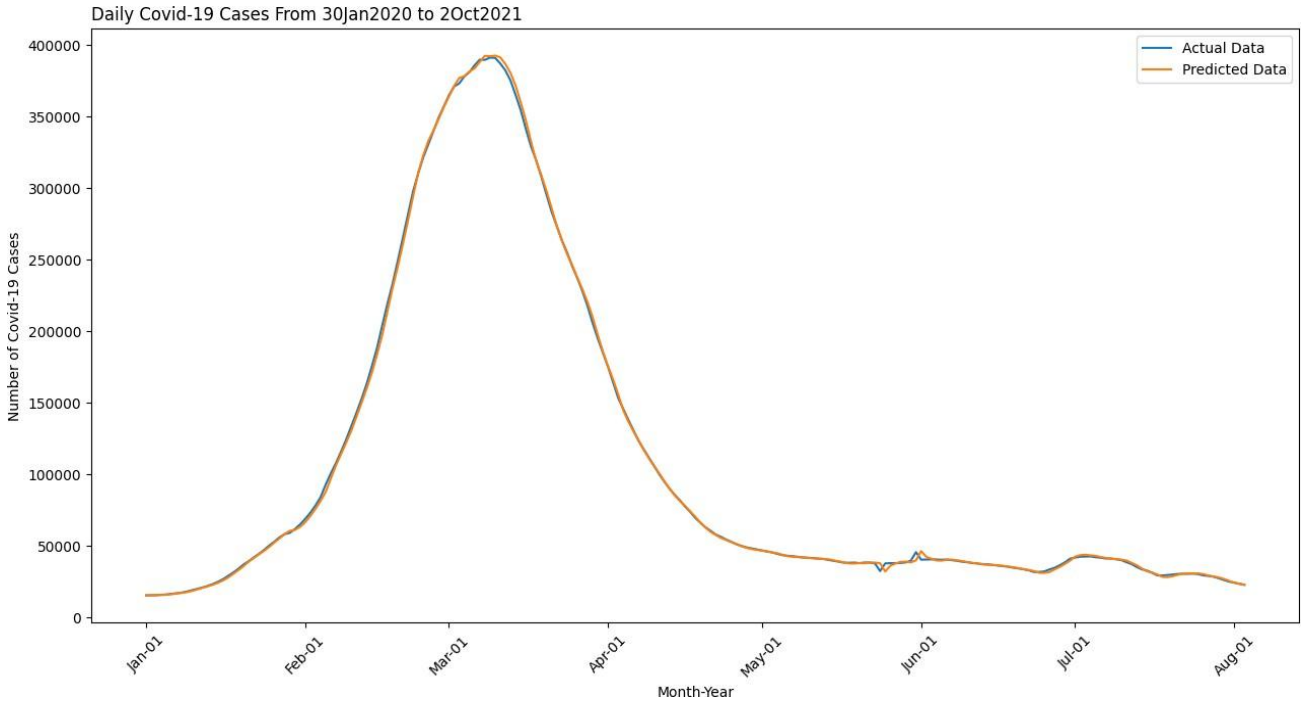

Figure 6 Predicted test data time sequence vs. original test data sequence

Inferences:

1. Model is very highly reliable as from the graph we can see predicted values are quite accurate. About future prediction of Covid-cases if it only depends upon past observations then this model will be highly useful but if the future depends upon other conditions (like presence of vaccines or lockdown) also then this model may give wrong results.

iii.

The RMSE(\%) and MAPE between predicted Covid-cases for test data and original values for test data are **1.825%** and **0.016** respectively.

Inferences:

1. Model is highly accurate.
2. Low value of RMSE and MAPE represents that the difference between actual and predicted data is very small. Which means data predicted by model is almost same as actual data.

3

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

Lag value	RMSE (%)	MAPE
1	5.373%	0.034
5	1.825%	0.016
10	1.686%	0.015
15	1.612%	0.015
25	1.703%	0.015

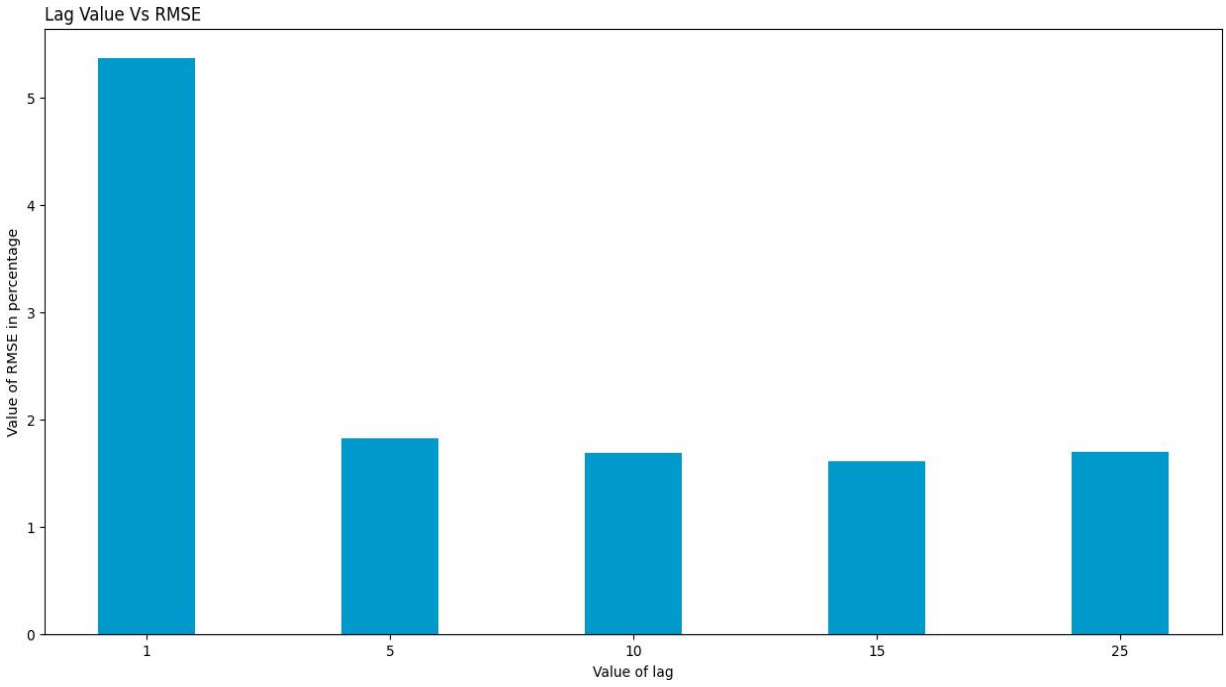

Figure 7 RMSE(%) vs. time lag

Inferences:

1. With the **increase** in lag value RMSE value **decreases** but **after certain value of lag** RMSE value starts **Increasing.**
2. By increasing lag values, we are incorporating more past values to predict future. Recent past values contribute more in accurately predicting future so, by increasing lag value RMSE decreases. But after certain lag values contributions in predicting future decreases to a certain level that increases the RMSE, that's why we see increase in RMSE after certain value of lag.

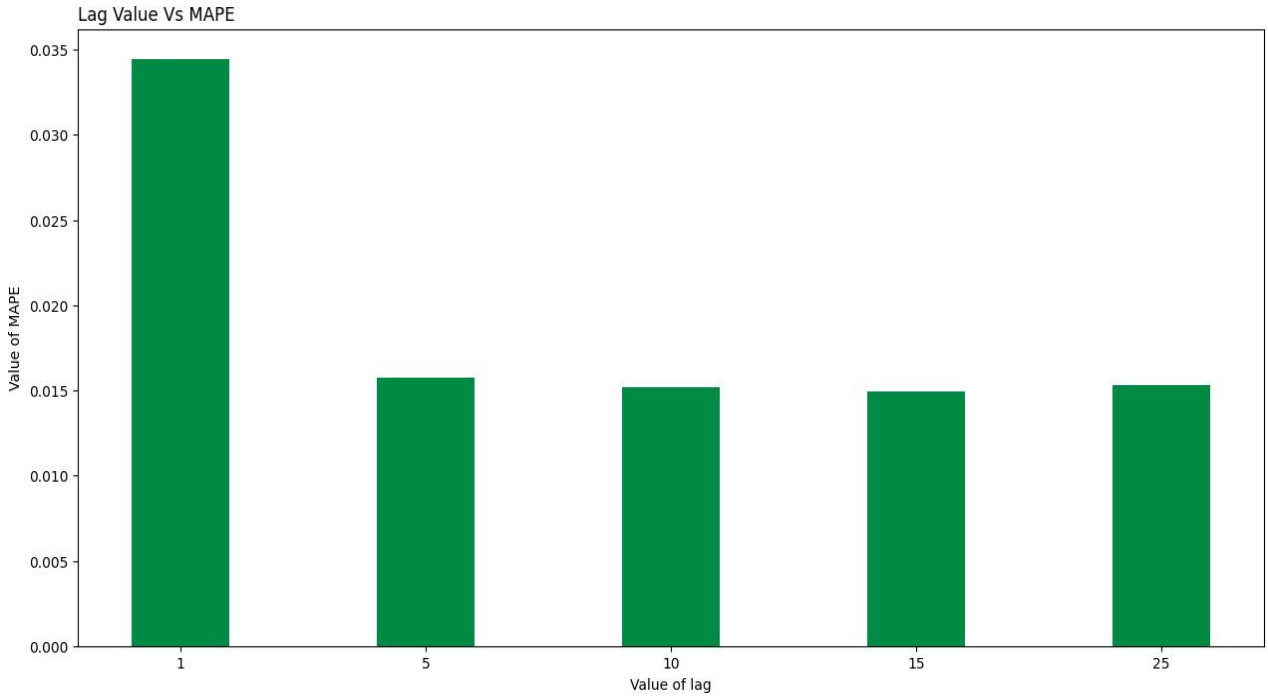

Figure 8 MAPE vs. time lag

Inferences:

1. With the **increase** in lag value, MAPE value is **decreases**.
2. Same as explained above.

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%** and **0.021** respectively.

Inferences:

1. Yes.
2. Optimal lag is 78 which means future observations depends upon 78 previous. Recent past value helps more in predicting future values; hence it increases the prediction accuracy.
3. Prediction accuracies is **high** in which values obtained with the heuristic for calculating optimal lag then values obtained without.