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

Student's Name: Ankit Pal Singh **Mobile No: 9149024234**

Roll Number: B20181 **Branch:EE**

1 a.

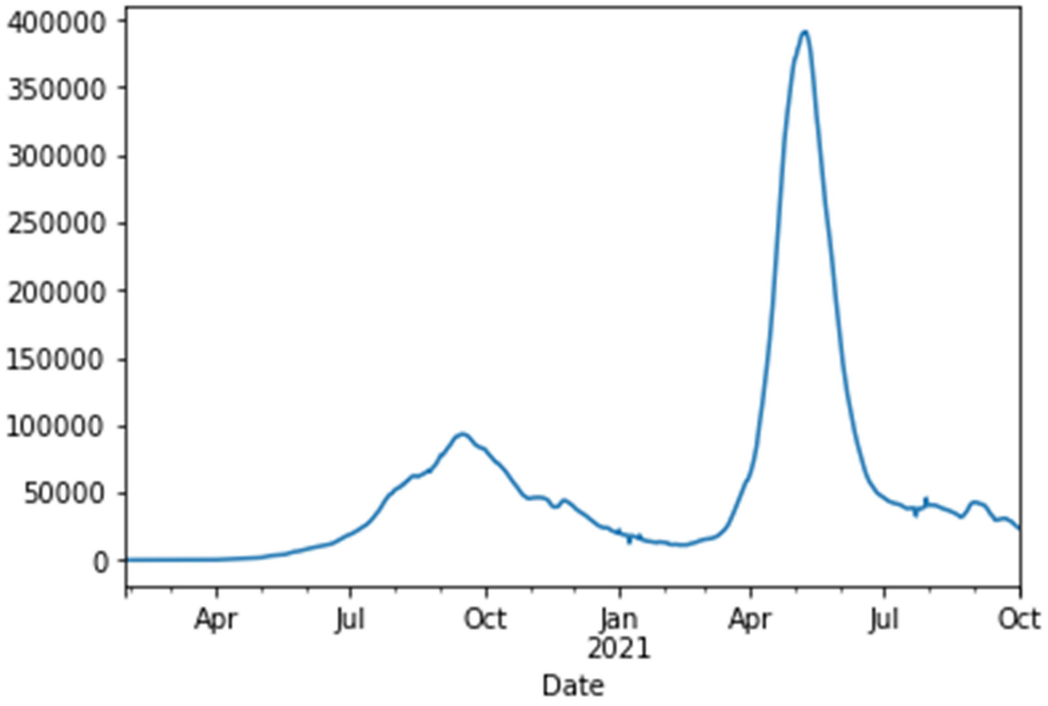


Figure 1 No. of COVID-19 cases vs. days

Inferences:

1. the days one after the other have a similar number of COVID-19 cases .
2. The reason for the above inference is as at the two peaks the value covid cases followed each other very efficiently
3. The duration of the first wave is in between July to October and the duration for the second wave is March to july.

b. The value of the Pearson's correlation coefficient is 0.99 .

Inferences:

1. The correlation is highly positive and approaches to 1 , which means when the value of one time series increases then the value of other time series increases and vice versa.
2. As the two peaks (first represent the first covid wave and other represent 2nd covid wave) , for the first wave the number of covid cases are similar to one after the other and same holds for the 2nd wave (so basically the relation is observe at both peaks).

c.

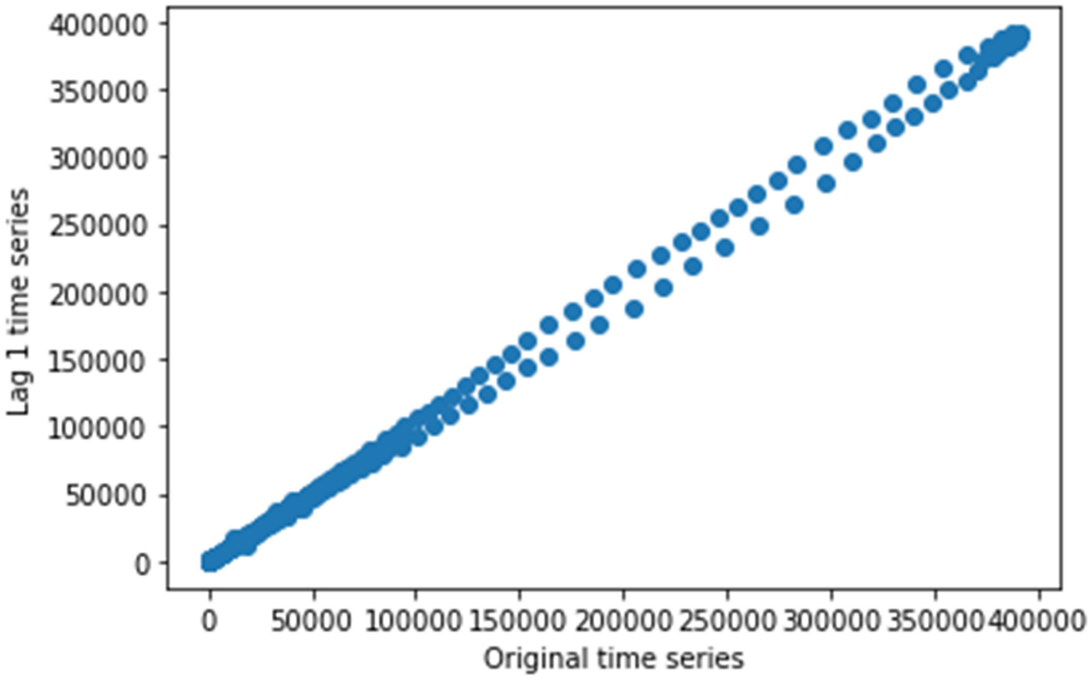
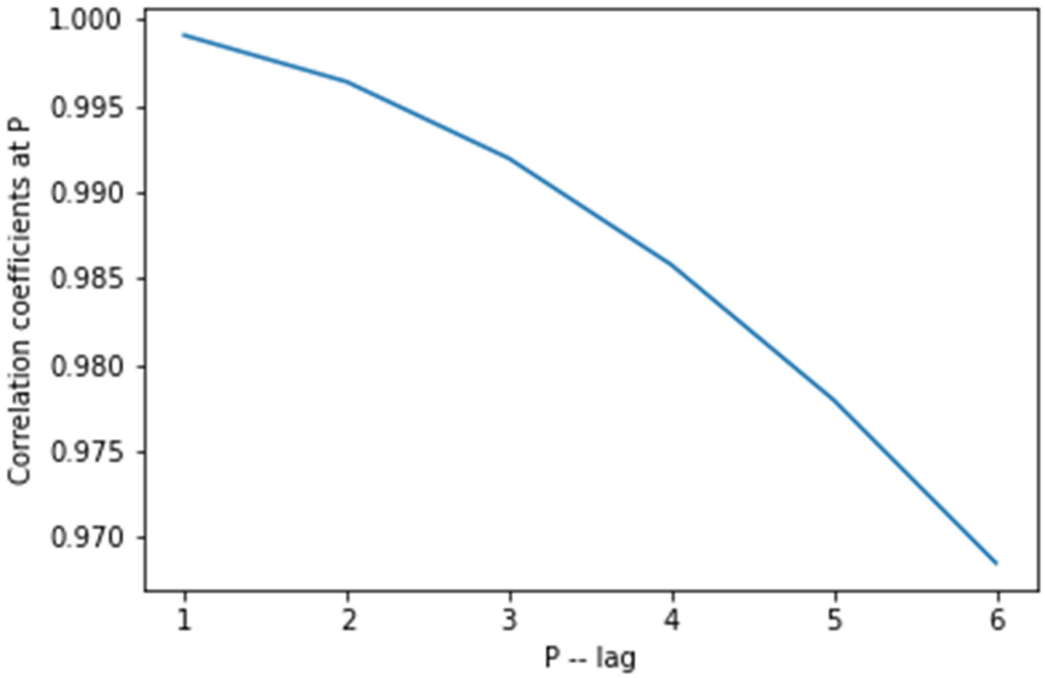


Figure 2 Scatter plot one day lagged sequence vs. given time sequence

Inferences:

1. Since the spread is linear so it means high correlation value exist between the two variables original time series and the lag 1 time series.
2. Scatter plot seems to obey the nature reflected by correlation very well.
3. The reason for the above inference is because the scatter plot is linear and our correleation value is highly positive.



d.

Figure 3 Correlation coefficient vs. lags in given sequence

Inferences:

1. As the value of P (the lag) increases the value of correlation coefficient decreases.
2. The reason behind the above inference is as the data is more dependent on previous value (upto very little extent) let's say till p =1 .

e.

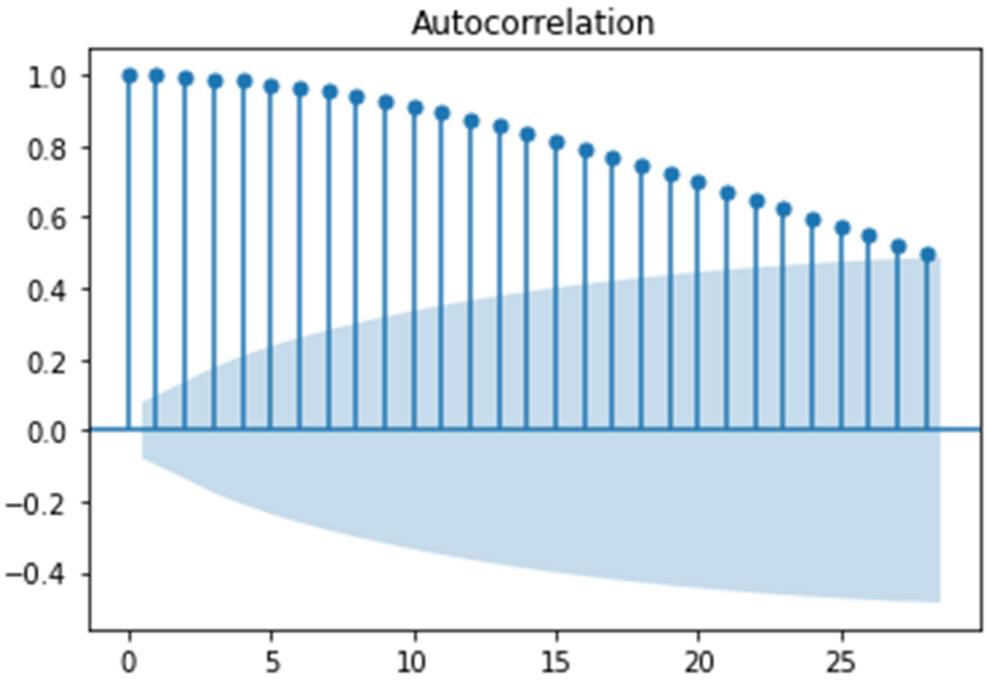


Figure 4 Correlation coefficient vs. lags in given sequence generated using 'plot_acf' function

Inferences:

1. As the value of P (the lag) increases the value of correlation coefficient decreases.
2. The reason behind the above inference is as the data is more dependent on previous value and those dependency is upto those past values which are very near .

2

a. The coefficients obtained from the AR model are [5.99548333e+01 1.03675933e+00 2.61712336e-01 2.75612628e-02 and -1.75391955e-01 -1.52461366e-01];

b. i.

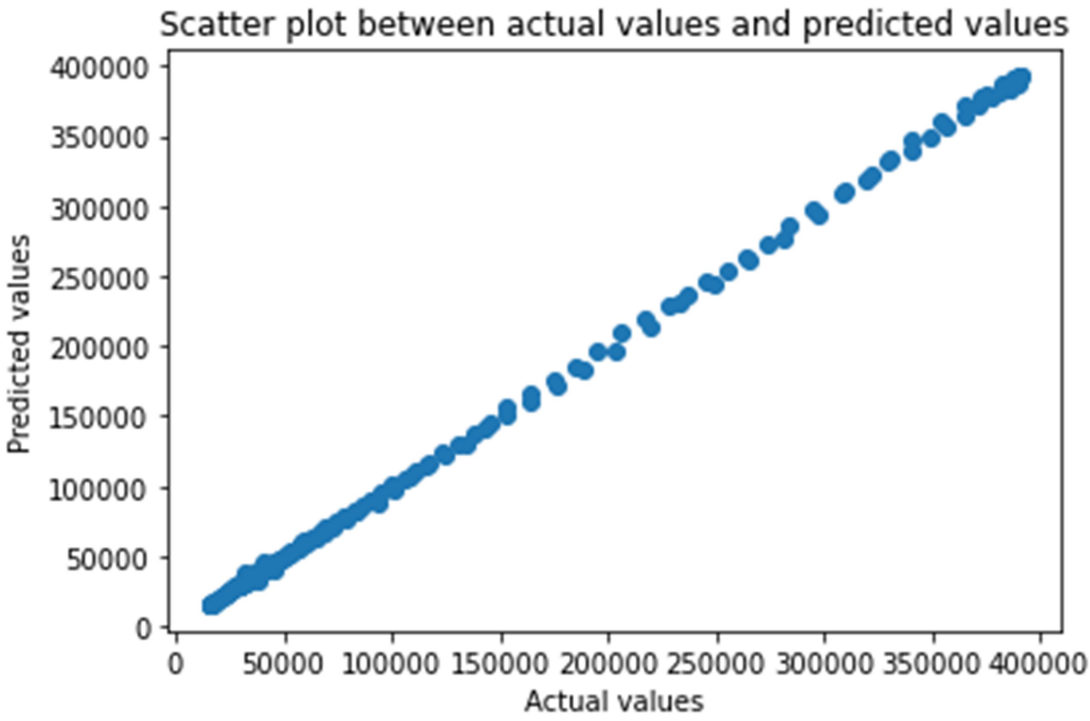
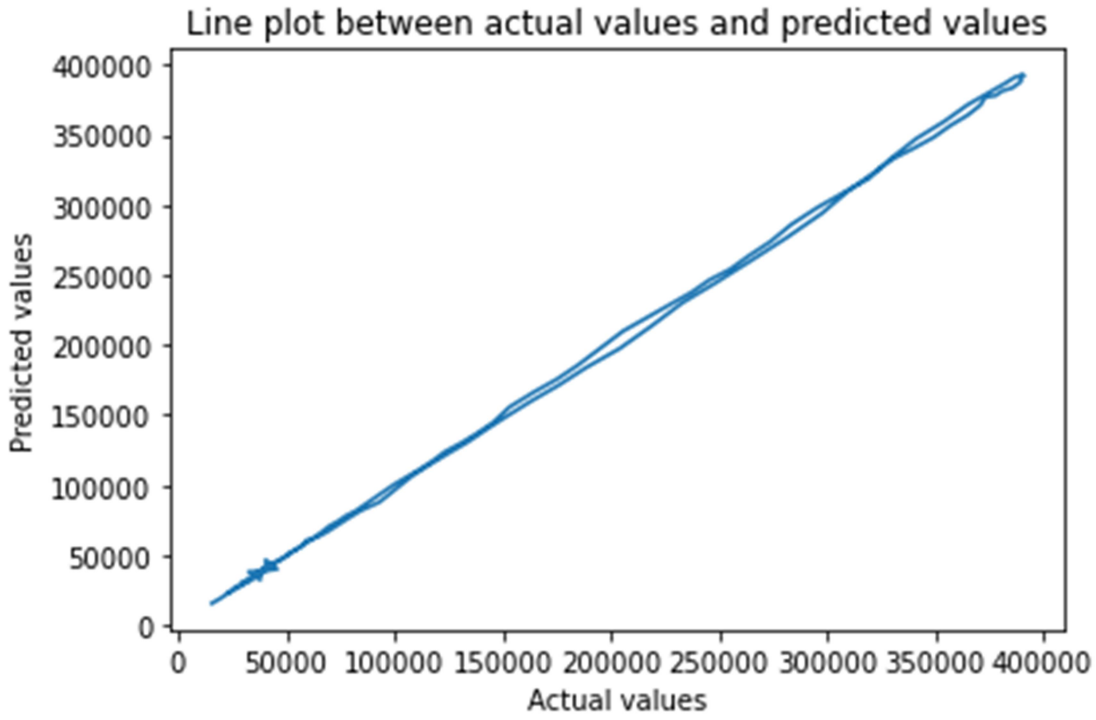


Figure 5 Scatter plot actual vs. predicted values

Inferences:

1. The value of correlation coefficient is highly positive as the scatter plot is linear.
2. Scatter plot seems to obey the nature reflected by correlation very well
3. The reason for the above inference is because the scatter plot is linear and our correlation value is highly positive



ii.

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

Inferences:

1. The model is good for predicting the future values because from the figure 6 the predicted data and original data highly matches.

iii.

The RMSE(\%) and MAPE between predicted power consumed for test data and original values for test data are 1.824 , 1.57 respectively.

Inferences:

1. The model is not that much reliable , but the model can predict upto some extent very well.
2. The reason for inference 1 is a good model has rmse lie in between 0.2 to 0.5 but our model have rmse somewhere around 1.8.

3

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

Lag value	RMSE (%)	MAPE
1	5.37	3.44
5	1.82	1.57
10	1.68	1.51
15	1.61	1.49
25	1.70	1.53

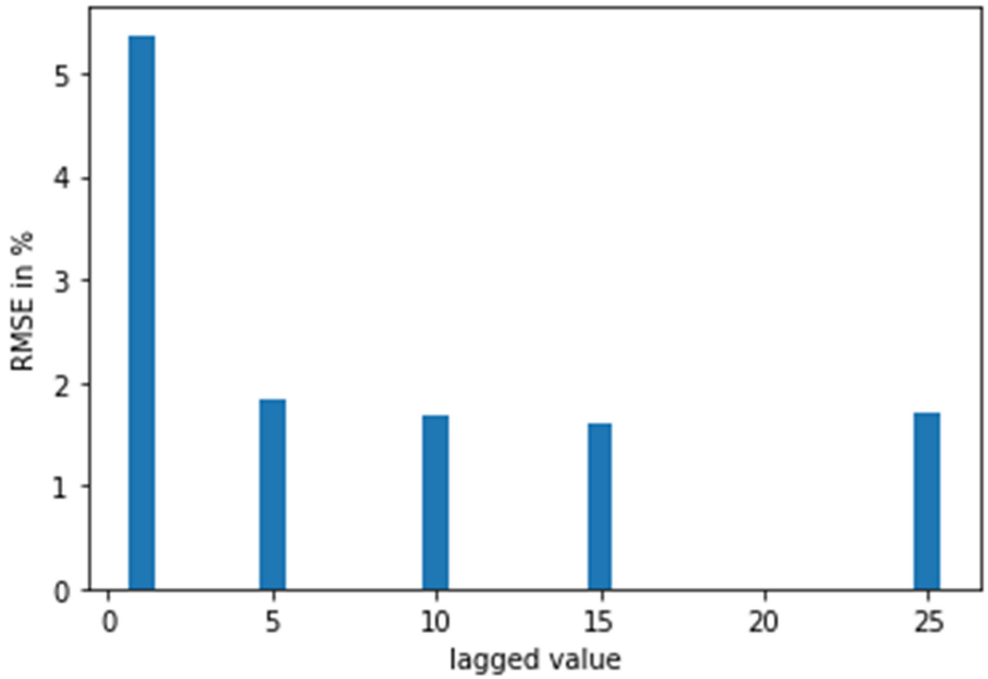


Figure 7 RMSE(%) vs. time lag

Inferences:

1. As the value of lags increases the value of RMSE decreases which means the with some high value of P the model with predictability can be made.
2. The reason for the above inference Is that the data can predict better if we make use high value of p that is the values are more dependent on that value of P.

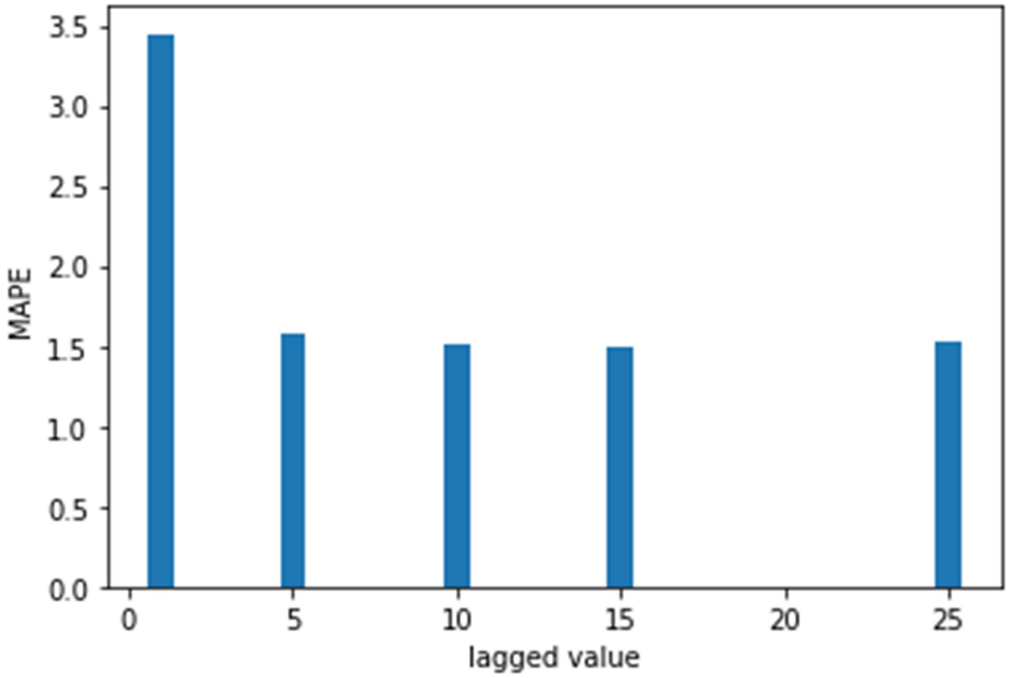


Figure 8 MAPE vs. time lag

Inferences:

1. As the value of lags increases the value of RMSE decreases which means the with some high value of P the model with predictability can be made.
2. The value of Mape increases but upto to a value , after that it becomes constant.
3. The reason for the above inference Is that the data can predict better if we make use high value of p that is the values are more dependent on that value of P.

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 are1.75 and 2.02

Respectively.

Inferences:

1. Yes , the model with heuristic value of P can predict better .
2. The reason for the above inference is the value of RMSE and MAPE with that p value decreases