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

Student's Name: Rajeev kumar **Mobile No: 9341062431**

Roll Number: b20124 **Branch:**CSE

1 a.

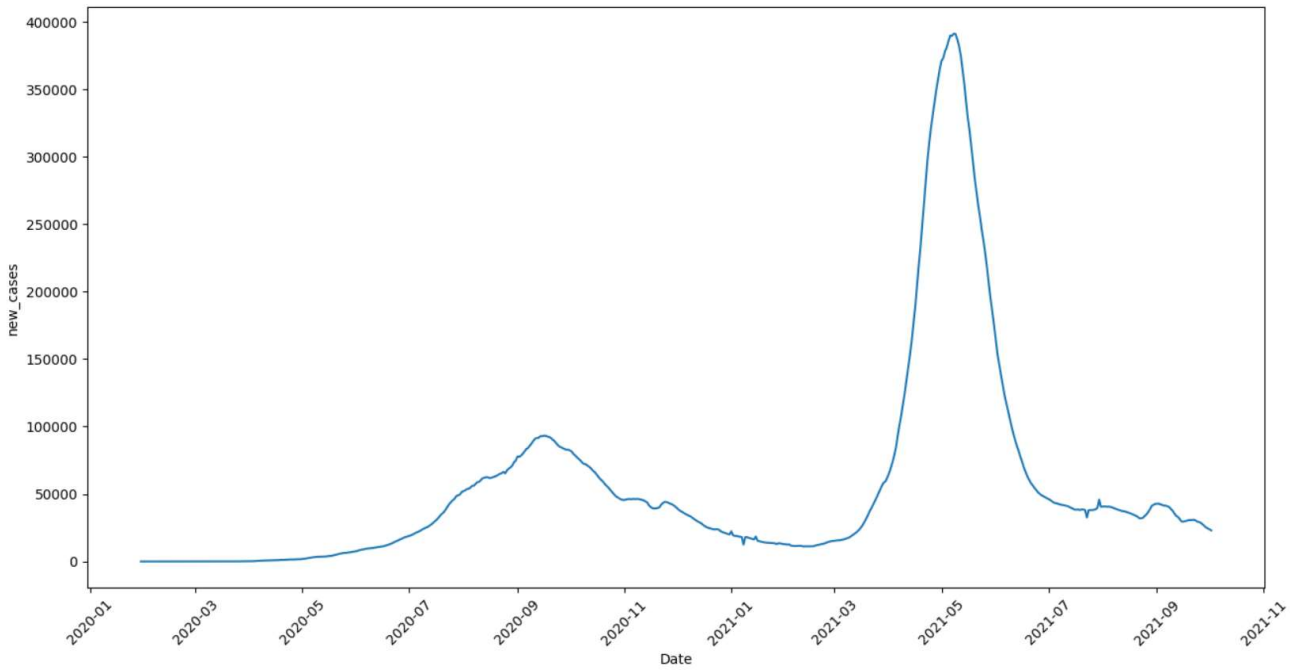

Figure 1 No. of COVID-19 cases vs. days

Inferences:

1. Yes, the days one after the other have similar power consumption
2. The reason for inference 1 is that it is a time series data.

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

Inferences:

1. Series is strongly correlated that's mean future value is highly dependent on past value.

2. We generally expect observations (here number of COVID-19 cases) on days one after the other to be similar. Since the correlation coefficient is very high so the next day no. of confirmed cases is dependent on the previous day data.

c.

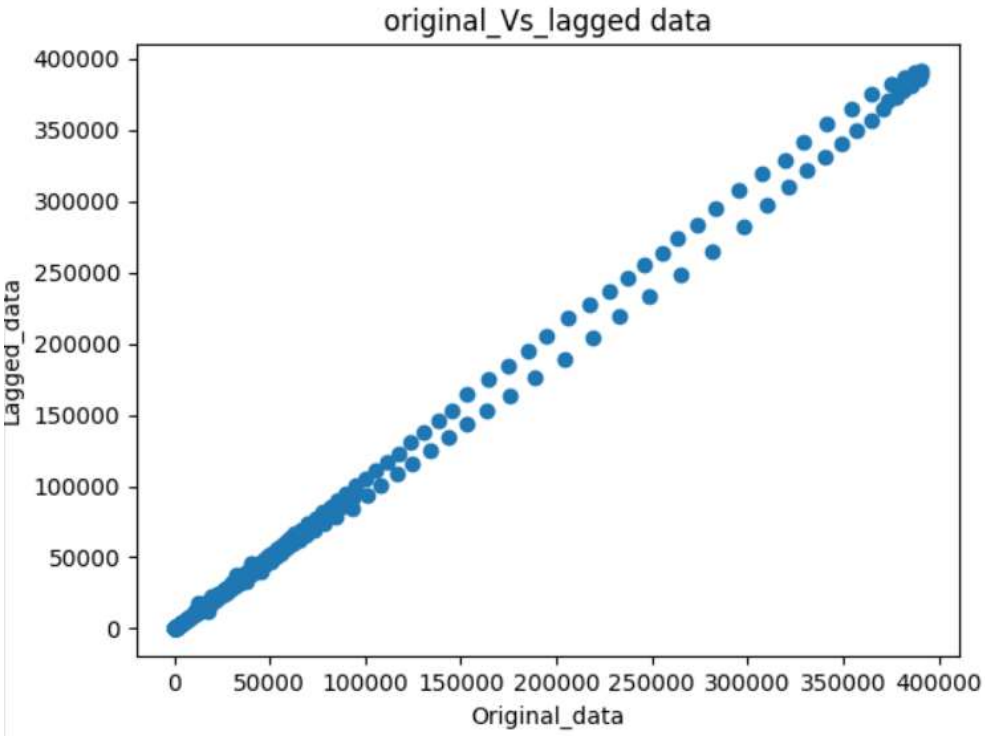

Figure 2 Scatter plot one day lagged sequence vs. given time sequence

Inferences:

1. Pearson's correlation is very high that's mean variable are highly dependent on each other.
2. the scatter plot seems to obey the nature reflected by Pearson's correlation coefficient.
3. The nature of environment and the mutant of corona virus depends on the previous day to larger extent and it does not change suddenly.

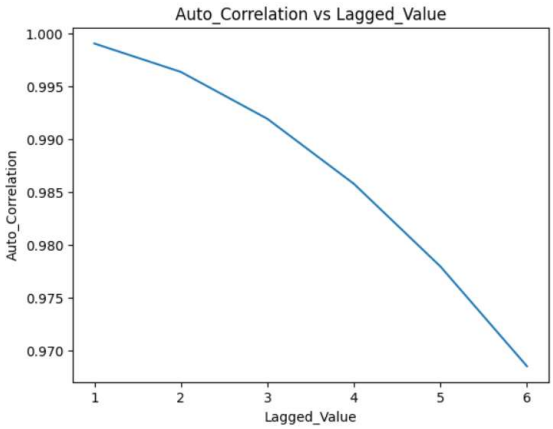

Figure 3 Correlation coefficient vs. lags in given sequence

1. **d.**

   inferences:

   1.The correlation coefficient value decreases with respect to increase in lags in time sequence**.**

   2. The reason behind the observed trend is that in that span the environment and the mutant of corona virus has got changed to a considerable extent.

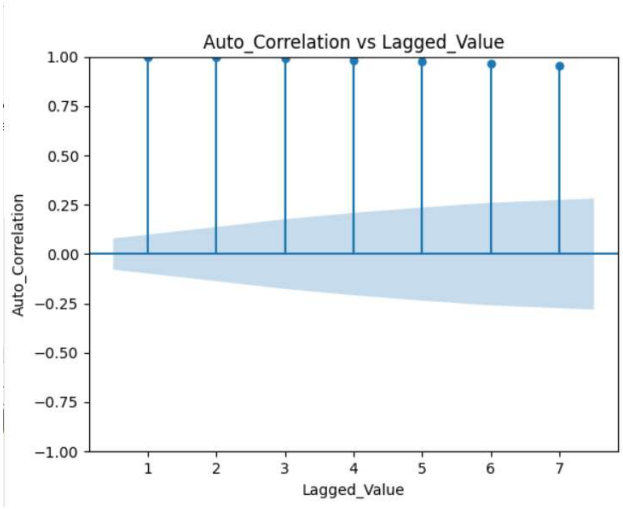

Figure 4 Correlation coefficient vs. lags in given sequence generated using 'plot_acf' function

Inferences:

1. Correlation coefficient decreases with respect to increase in lags in time sequence
2. The reason behind the observed trend is that in that span the environment and the mutant of corona virus has got changed to a considerable extent

2

a. The coefficients obtained from the AR model are; [59.955,1.03,0.26,0.028, -0.175, -0.152]

b. i.

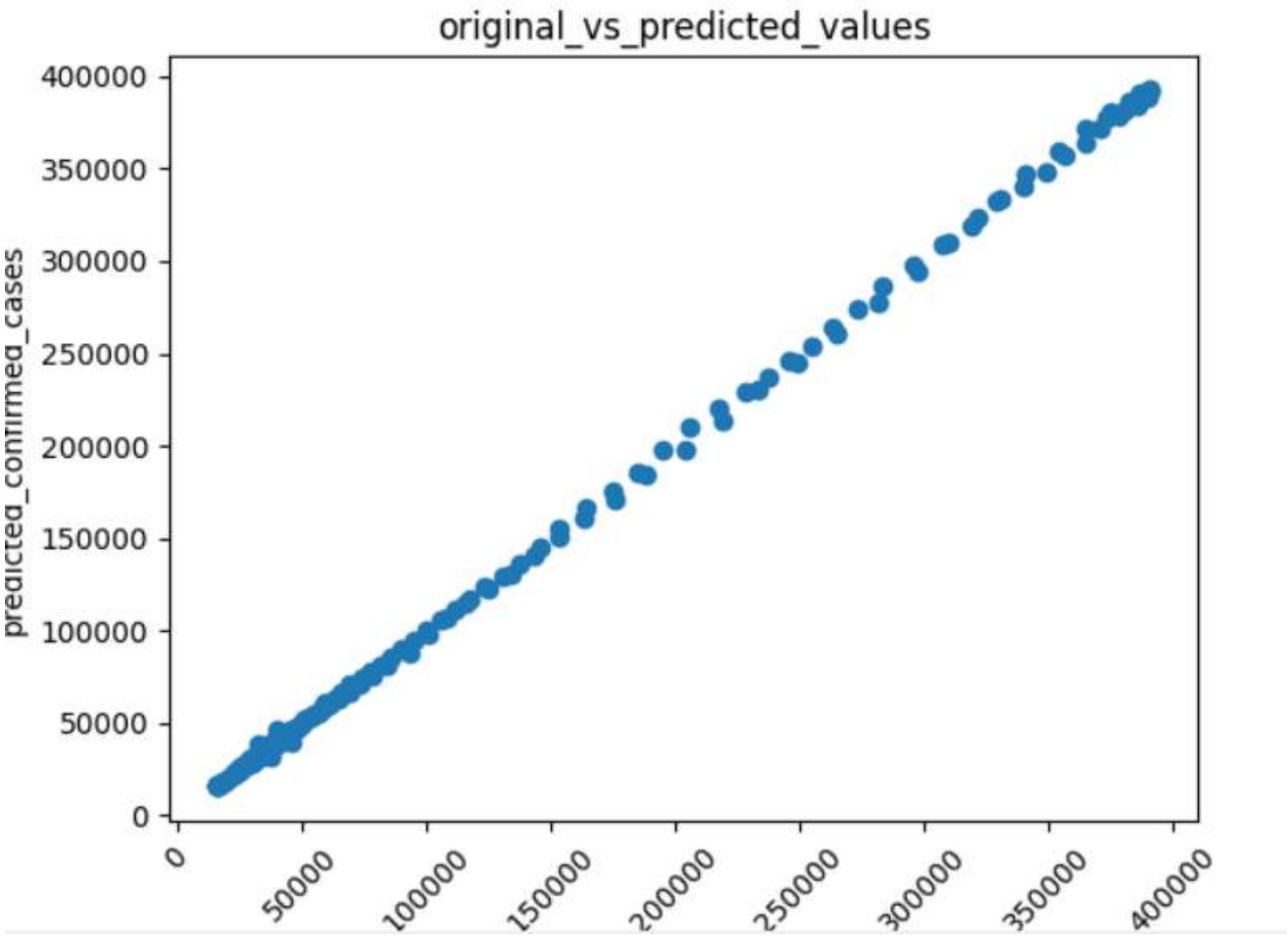

Figure 5 Scatter plot actual vs. predicted values

Inferences:

1. From the nature of the spread of data points, the nature of correlation between the two sequences is positive and its magnitude also seems high.
2. Yes, scatter plot seems to obey the nature reflected by Pearson's correlation coefficient calculated.
3. The nature of environment and the mutant of corona virus depends on the previous day to larger extent and it does not change suddenly. Also the span we used as lag is 5 days and in that span these changes are not expected to be much.

ii.

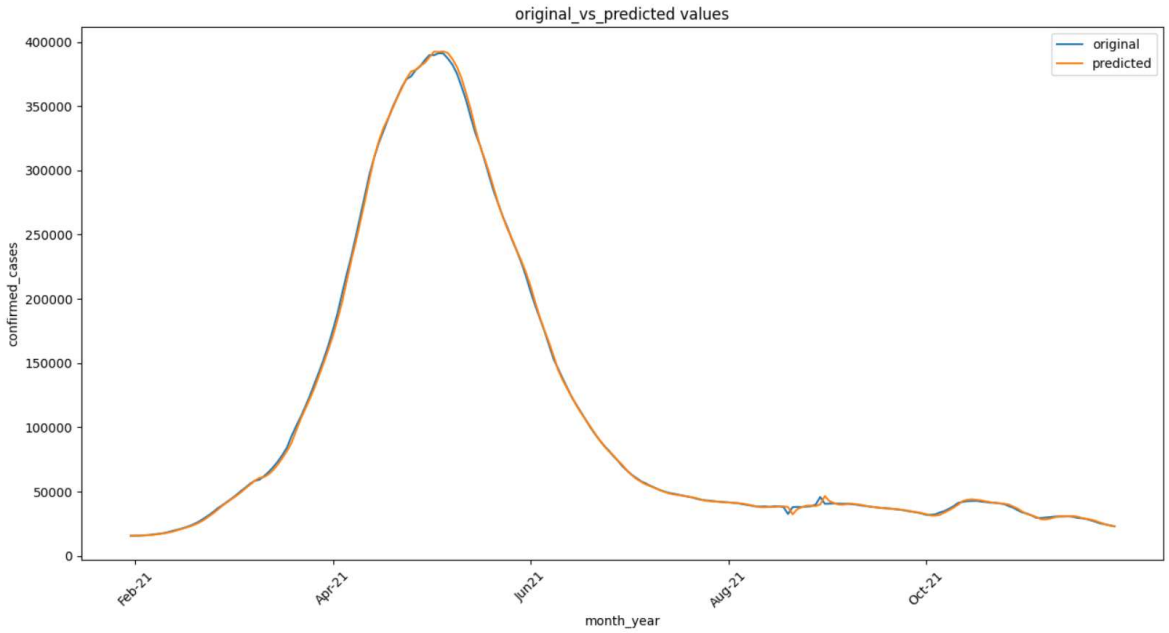

Figure 6 Predicted test data time sequence vs. original test data sequence

Inferences:

1.From the plot of predicted test data time sequence vs. original test data sequence we can say that the model is reliable for future predictions because it has a very less rmse value which can be compensated.

iii.

The RMSE(\%) and MAPE between predicted power consumed for test data and original values for test data are 1.8245,1.574 respectively.

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

Lag value	RMSE (%)	MAPE
1	**5.27%**	**3.44**
5	**1.82%**	**1.57**
10	**1.68%**	**1.51**
15	**1.61%**	**1.49**
25	**1.70%**	**1.53**

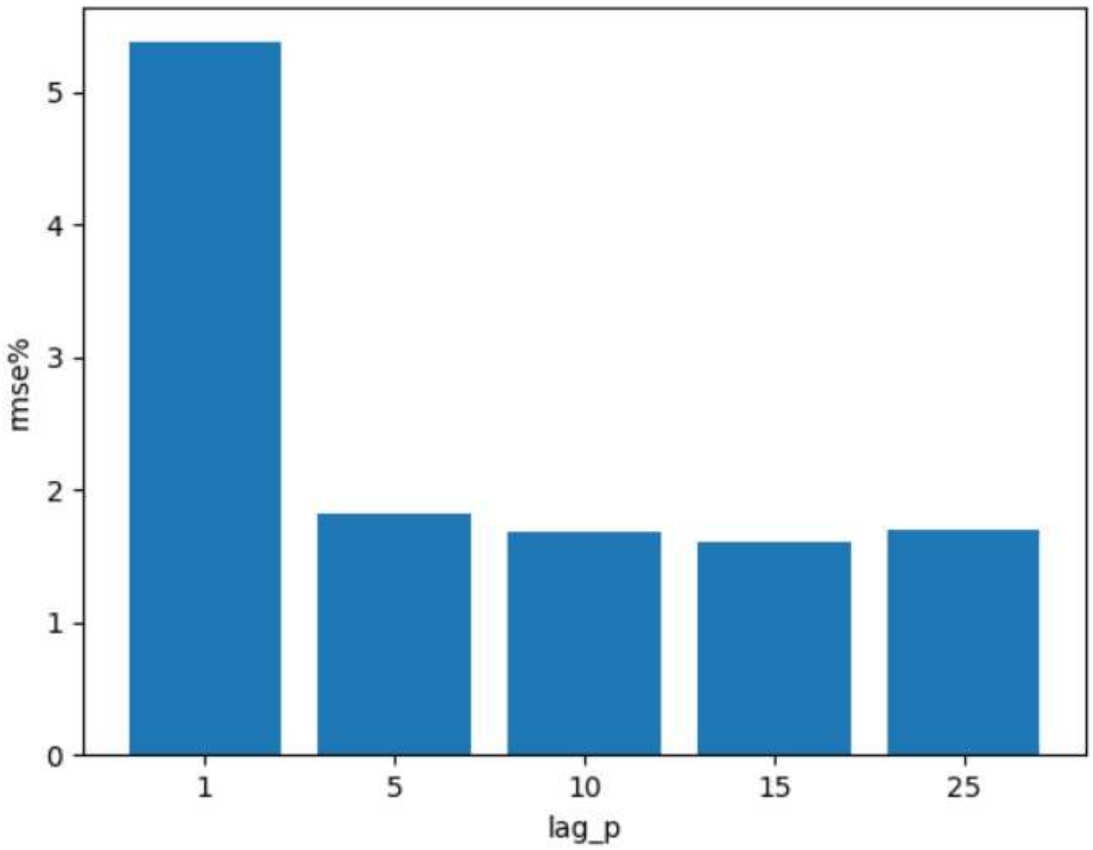

Figure 7 RMSE(%) vs. time lag

Inferences:

1. Lag decreases as rmse value increases.
2. ncreasing the lag means we are considering more environmental changes and others into consideration which increases its prediction accuracy.

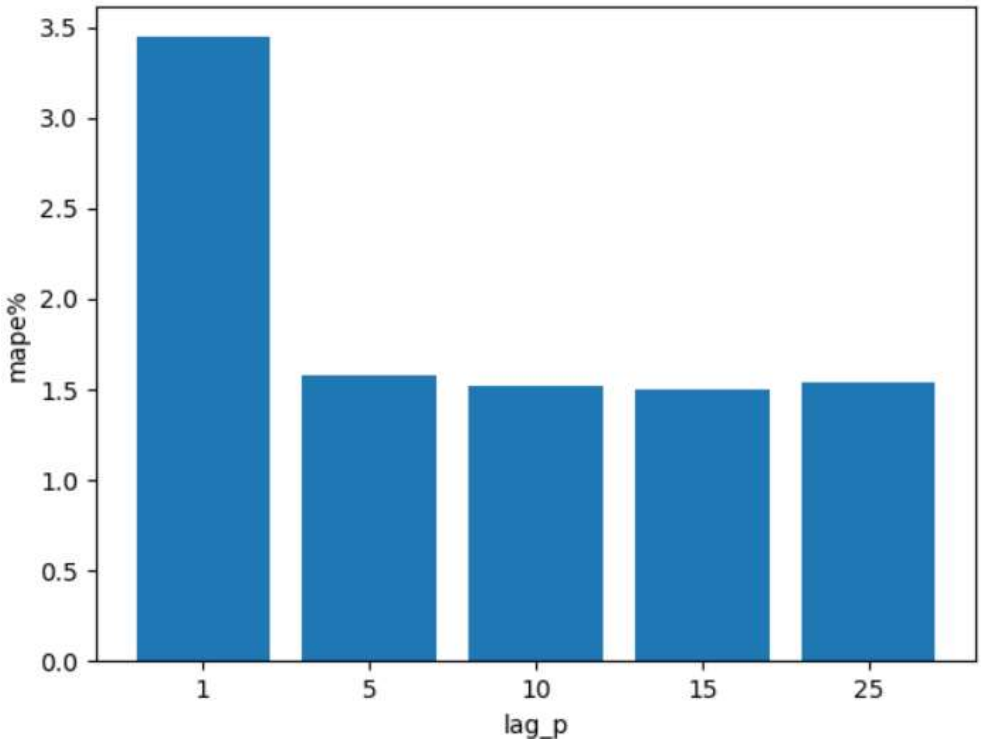

Figure 8 MAPE vs. time lag

Inferences:

1. The trend of RMSE(%) with respect to increase in lags in time sequence is that it decreases with increase in lag s

4

The heuristic value for the optimal number of lags is 76.

The RMSE(%) and MAPE value between test data time sequence and original test data sequence are1.7517,2.000 respectively.

Inferences:

1. Based upon the RMSE(%) and MAPE value, heuristics for calculating the optimal number of lags improves the prediction accuracy of the model to a certain extent.

2. This helps a lot in data which are too huge and on that starting with lag 1 and then increasing it one by one to choose the optimal value of p(lag) is not a good idea.
3. The RMSE and MAPE values with heuristics is lesser than lag 1 but more than the others