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

Student's Name: Gayatri Shridhar Kapse **Mobile No: 8623916561**

Roll Number: B20199 **Branch:**EE

1 a.

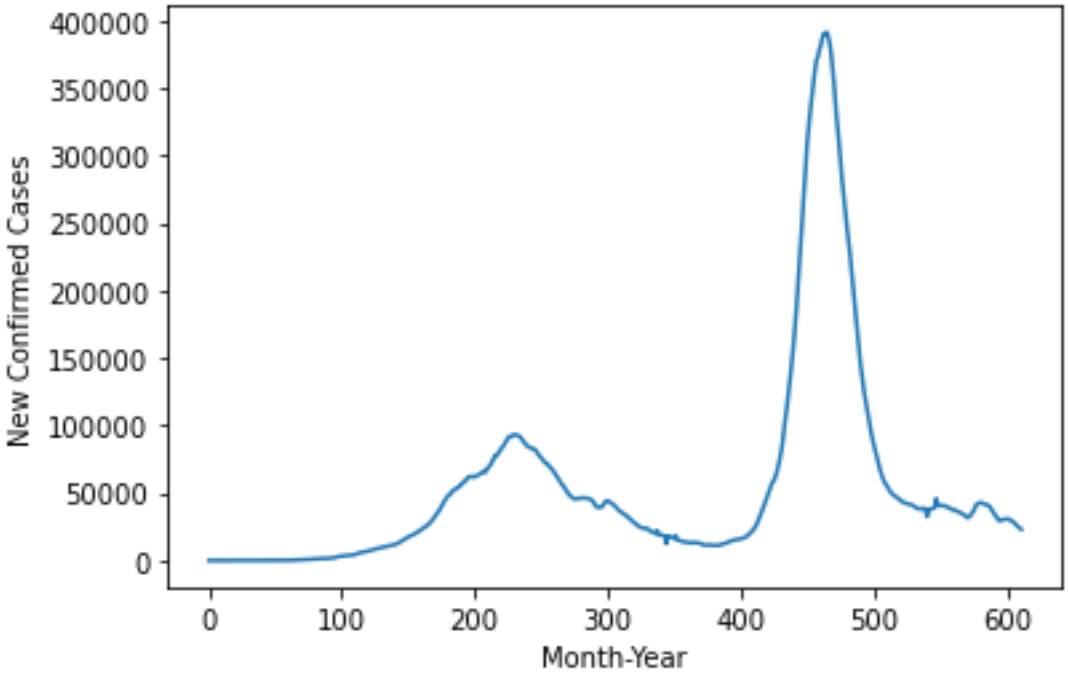

Figure 1 No. of COVID-19 cases vs. days

Inferences:

1. Infer from the plot whether the days one after the other have similar power consumption?
2. State the reason behind inference 1.
3. Infer from the plot the duration of first and second waves.

   Note: The plot above is for illustration purposes. Replace it with the plot obtained by you. Suitably rename x-axis and y-axis legends.

b. The value of the Pearson's correlation coefficient is: 0.99899677160

Inferences:

1. From the above value of pearson's correlation coefficient we infer that the one day-lag sequence is highly correlated with the given sequence, and also they are positively correlated.
2. In the above case its observed that the cases that day is similar to the day before as from the correlation coefficient we get that they are positively highly correlated, In this case its true that "Generally expected observations on days one after the other to be similar".
3. High value of correlation coefficient.

c.

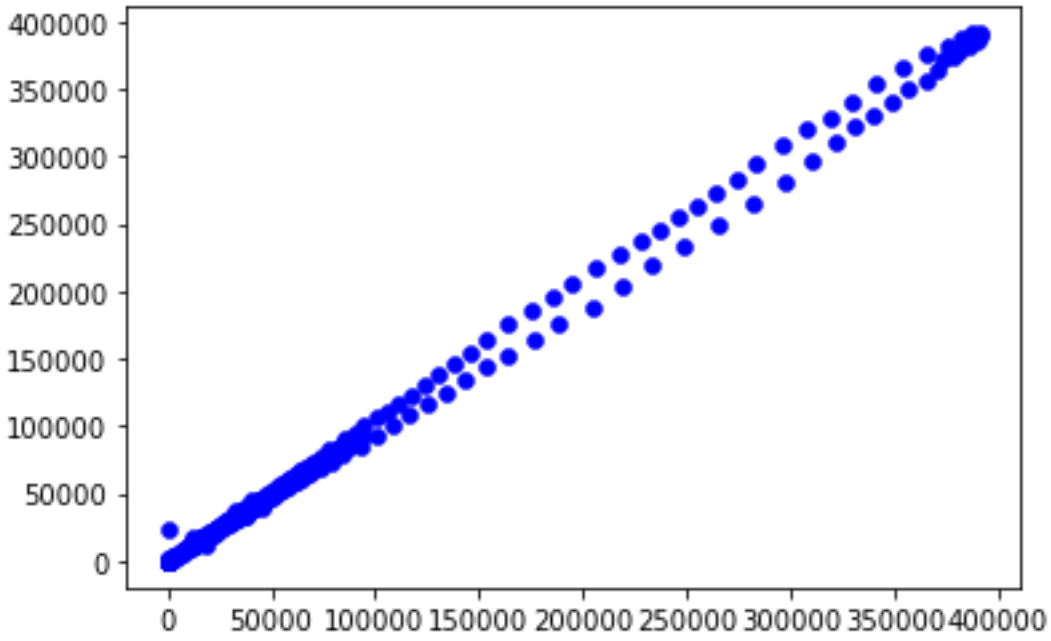

Figure 2 Scatter plot one day lagged sequence vs. given time sequence

Inferences:

1. From the above scatter plot we observe that as the value of given time sequence is increasing the one day lag sequence is also increasing this tells us that the two sequences are positively correlated and also as the scatter plot looks like a straight line this tells that they are highly correlated.
2. Yes the scatter plot makes it totally clear why we got high correlation coefficient in the in the previous question.

3. From both the scatter plot and pearson's correlation coefficient its clear that the lagged sequence is highly correlated to the given sequence.

e.

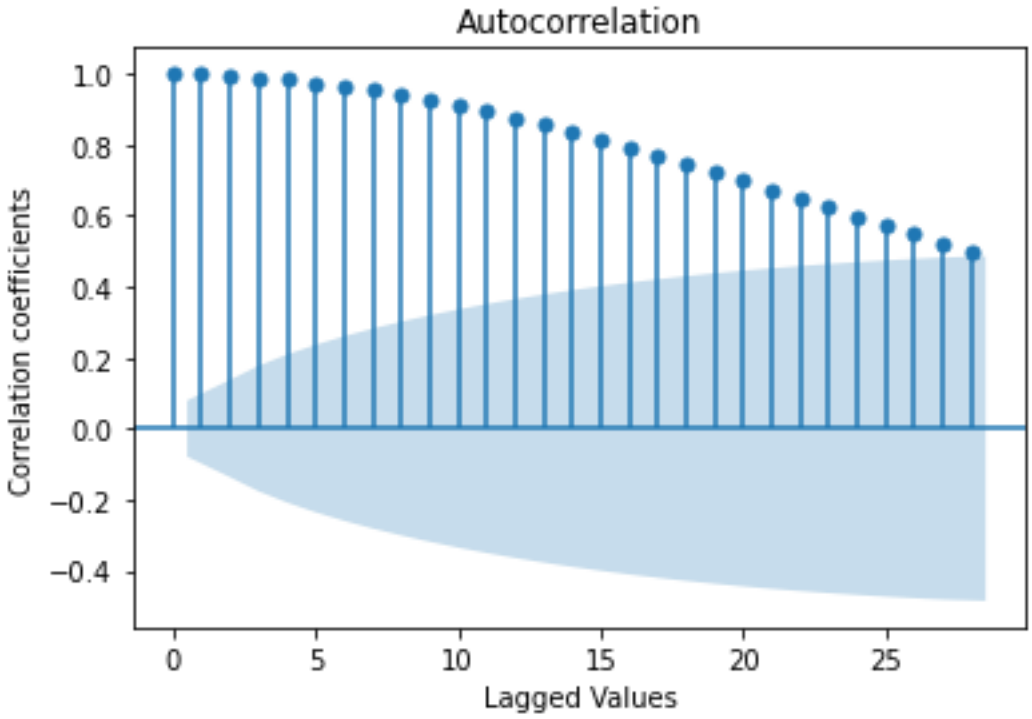

Figure 3 Correlation coefficient vs. lags in given sequence generated using 'plot_acf' function

Inferences:

1. From the above graph we observe that as the value of lag increases we the correlation coefficients decreases, this tells us that as the lag value increases the relation between the lagged sequence and the given sequence is decreasing.
2. The above trend is seen because the dependency of data observed today on the days before will decrease as we go in past , as its hard to exactly predict a data today just on what happened last month or two months ago.

d.

```
1. The correlation between lag 2 sequence and original is:0.9962419074852337

2. The correlation between lag 3 sequence and original is:0.9917489864794244

3. The correlation between lag 4 sequence and original is:0.9855554656451243

4. The correlation between lag 5 sequence and original is:0.9777141408998662

5. The correlation between lag 6 sequence and original is:0.968278061970016
```

Inferences:

1. We observe that correlation coefficient for lag=2 is higher than the correlation coefficient at lag=6.
2. The above trend is seen because the dependency of data observed today on the days before will decrease as we go in past, as its hard to exactly predict a data today just on what happened last month or two months ago.

2

a. The coefficients obtained from the AR model are;

```
[ 5.99548333e+01,1.03675933e+00,2.61712336e-01,2.75612628e-02,
 -1.75391955e-01 -1.52461366e-01]
```

b. i.

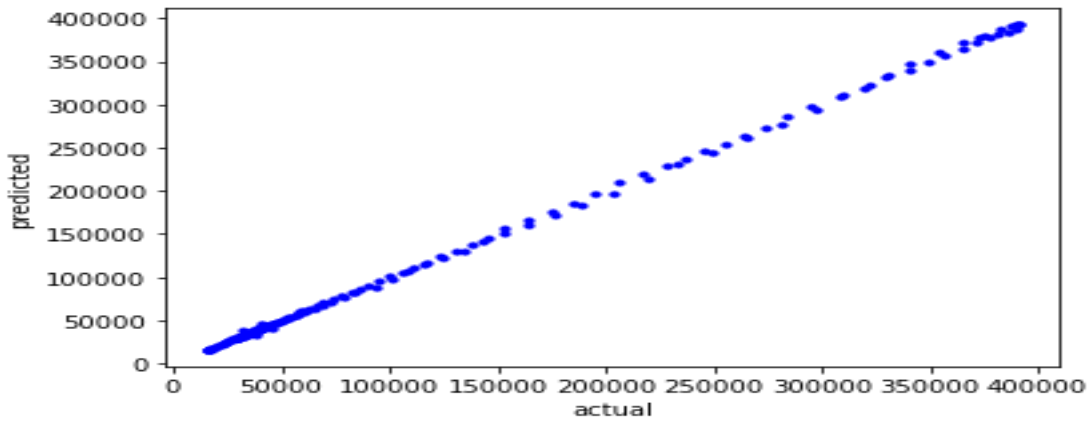

Figure 4 Scatter plot actual vs. predicted values

Inferences:

1. From the nature of the spread of data points, we observe that our prediction values are close to actual values.
2. Yes, the above plot seems to obey the correlation coefficient that we calculated in the question 1.
3. From the plot we are actually comparing the original value and the value that we got after creating the AR model so ,The weights that we obtained is used in the AR model and in this the weights obtained seem to decrease as the Lag value is higher , so this is exactly what we inferred in first question

ii.

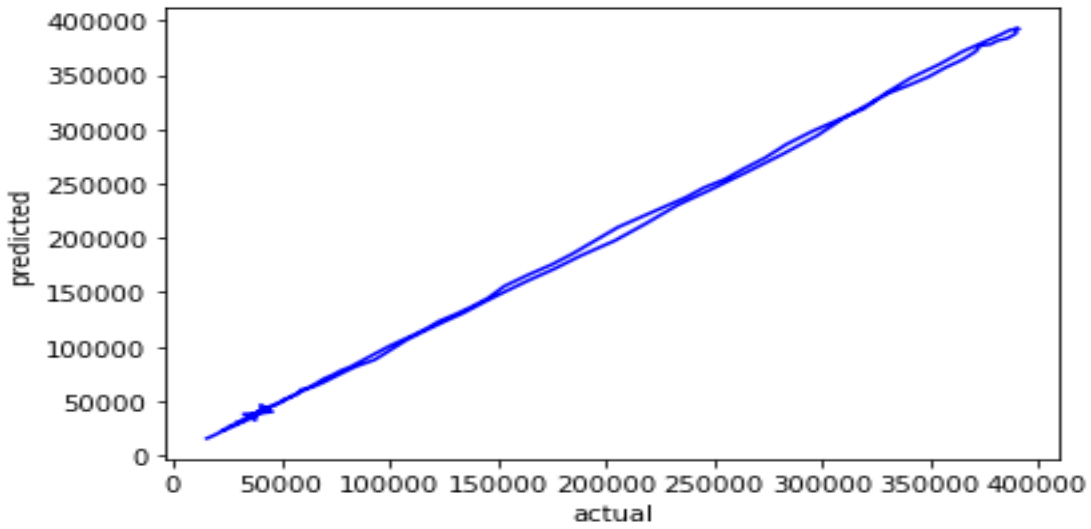

Figure 5 Predicted test data time sequence vs. original test data sequence

Inferences:

1. From the plot of predicted test data time sequence vs. original test data sequence we observe that this model is totally reliable for future predictions because from the plot we infer that the predicted values are almost close to the actual values.

iii.

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

```
RMSE(%): 1.824768476938991
MAPE: 1.5748363824058171
```

Inferences:

1. Low values of RMSE indicate that the data is acceptably accurate.
2. As RMSE and MAPE values are indication of error that's why lower the value higher will be the accuracy.

3

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

Lag value	RMSE (%)	MAPE
1	5.372948	3.446540
5	1.824768	1.574836
10	1.685532	1.519370
15	1.611935	1.496236
25	1.703391	1.535421

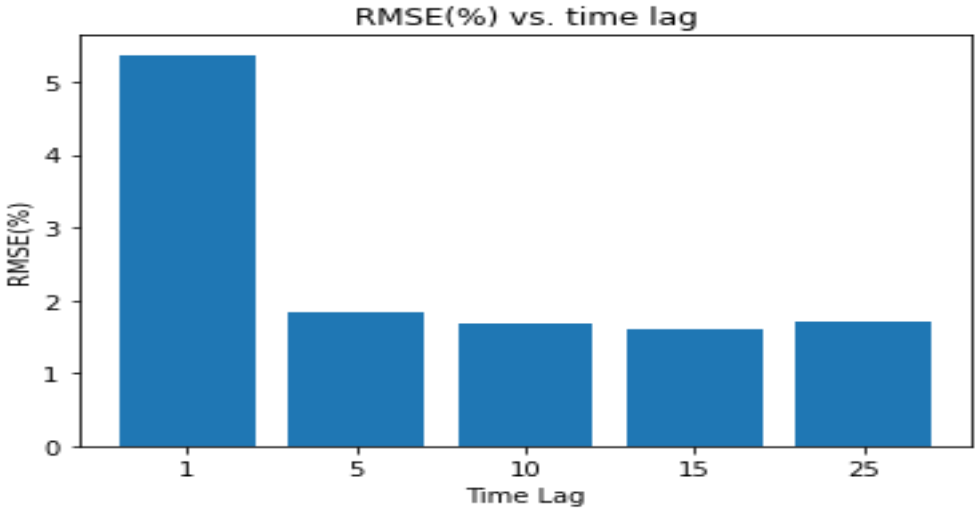

Figure 6 RMSE(%) vs. time lag

Inferences:

1. Mostly it is observed that as the lag increases the RMSE(%) decreases.
2. Y(lag=1) = a + bx

   .

   .

   Y(lag=25) = Wo +W1* X1 + W2 *X2 + W3*X3 +.....+W24 X24.

   From above it clearly indicates that the RMSE value is increasing as the lag increases.

   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

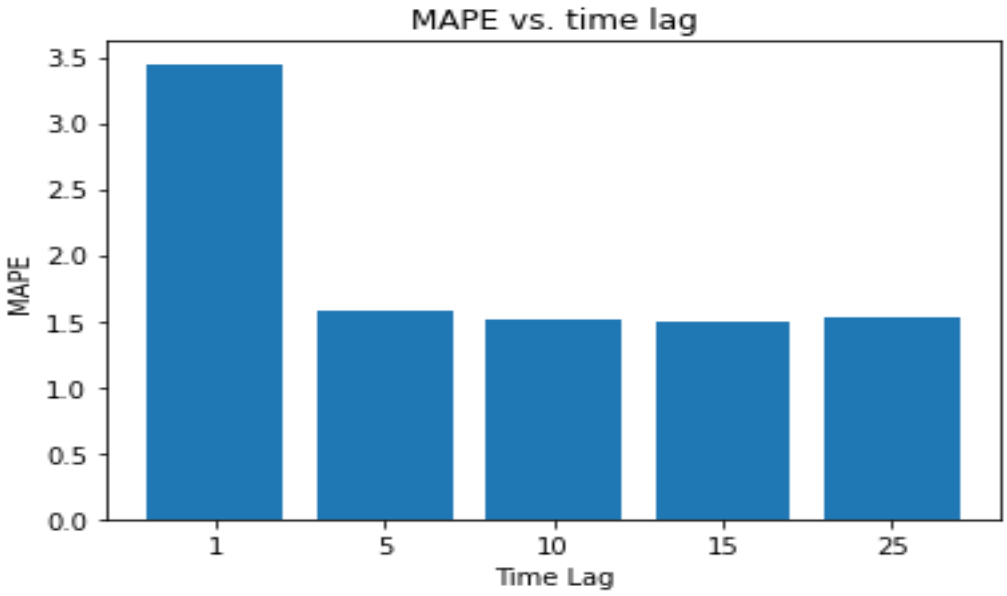

Figure 7 MAPE vs. time lag

Inferences:

1. The MAPE decreases quickly from 1 to 5 but then decreases gradually with respect to increase in lags in RMSE 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

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 are

```
RMSE(%): 1.7593780528857152
MAPE: 2.026443905282761
```

Inferences:

1. Based upon the RMSE(%) and MAPE value, the heuristics for calculating the optimal number of lags didn't improve the prediction accuracy of the model significantly as we can see the RMSE(%) for lag=10 was less than that for optimal lag.

2. Because as we keep increasing the lag, after certain RMSE the pattern of RMSE vs lag will become random and we can also see that as the observations are made for every day AR(77) doesn't make sense than that of a lag of around one day.

3. The prediction accuracies obtained without heuristic is less as compared to with heuristic for calculating optimal lag with respect to RMSE(%) and MAPE values. 8