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

Student's Name: Nikhil **Mobile No: 8949463760**

Roll Number: B20219 **Branch:ComputerScience&Engineering**

1

a.

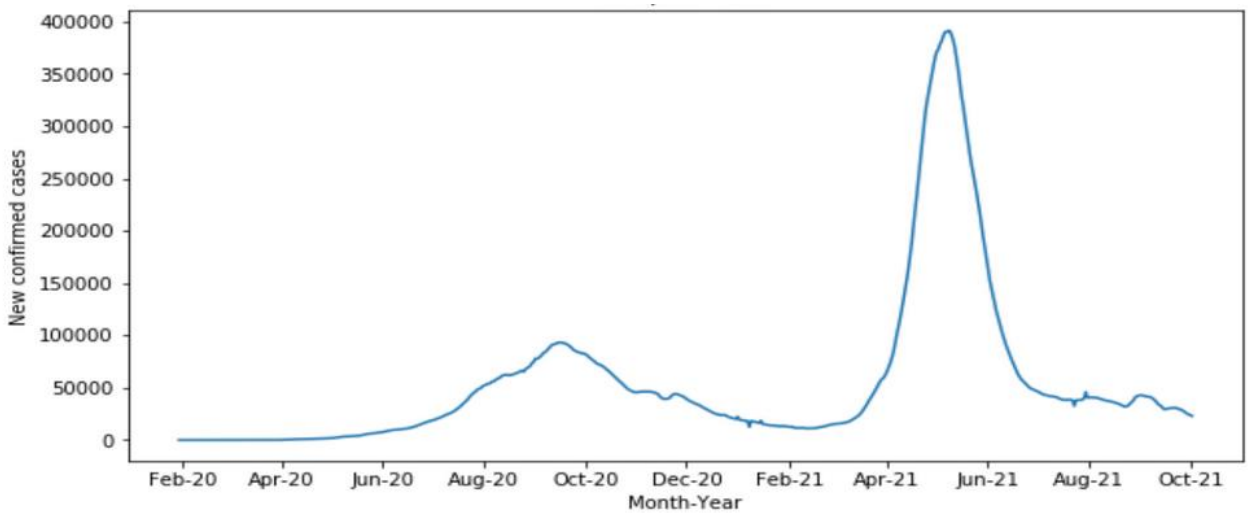

Figure 1 No. of COVID-19 cases vs. days

Inferences:

1. Days one after the other do have similar number of covid cases.
2. This is because the plot is a definite curve which is not stationary.
3. The first wave hit around June-2020 and ended around feb-2021 while the second wave hit at around March-2021 and end around Aug=2021.

b.

The value of the Pearson's correlation coefficient is 0.999

Inferences:

1. From the degree of correlation coefficient, we can see that it is very close to one, so the number of cases on day one after the other are highly similar.
2. We generally expect observations (here number of COVID-19 cases) on days one after the other to be similar. Here is the same case. Observations on day one after the other are highly similar and it can be seen through a high correlation coefficient of 0.999.
3. State the reason behind Inferences 1 and 2.

c.

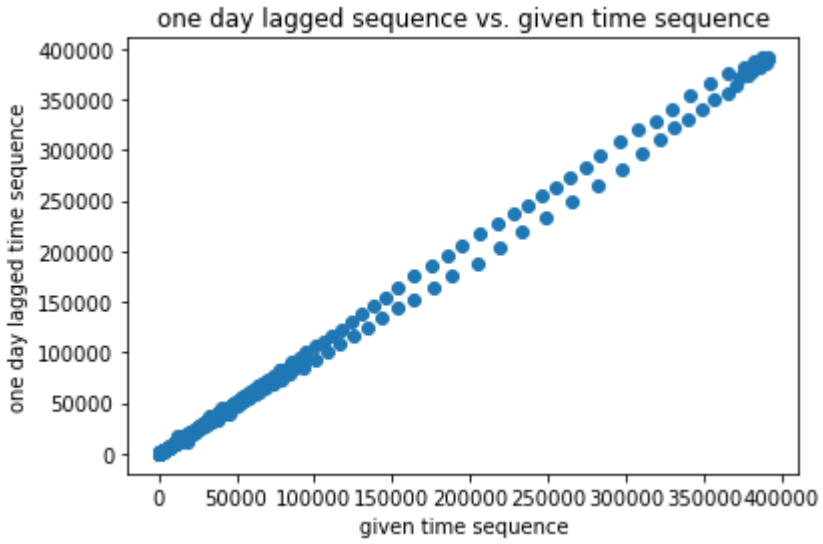

Figure 2 Scatter plot one day lagged sequence vs. given time sequence

Inferences:

1. From the nature of the spread of data points, we observe that the 2 attributes have very high positive correlation coefficient.
2. The scatter plot seems to directly obey the nature reflected by the Pearson's correlation coefficient.
3. High and positive correlation coefficient means if one attribute increases, it is highly likely that the other attribute also increases, and that is what we see in the scatter plot: when the given time

sequence has higher value, so does the one day lagged time sequence has. That's why the scatter plot obeys the nature reflected by the Pearson's correlation coefficient.

d.

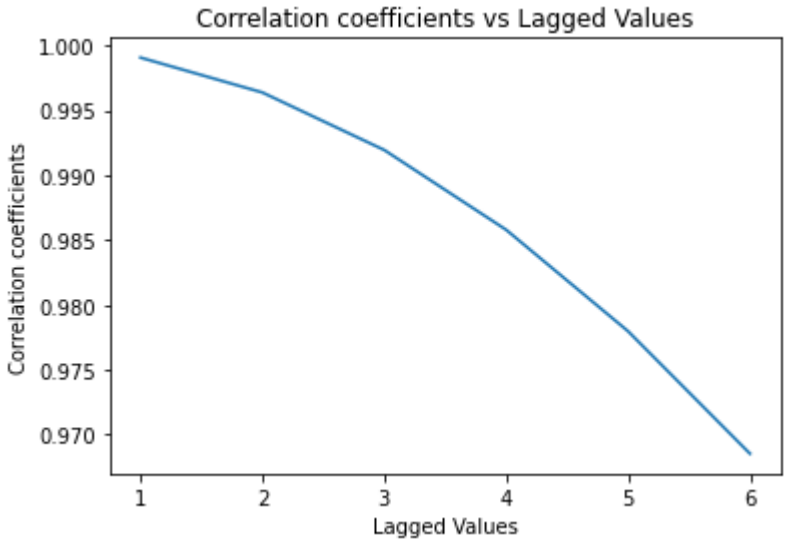

Figure 3 Correlation coefficient vs. lags in given sequence

Inferences:

1. The correlation coefficient values show a decreasing trend with respect to increase in the lag.
2. This is because one-day lagged sequence has some <= 1 correlation with the two-day lagged sequence so, in turn, the original time sequence has a correlation with two-day lagged time sequence smaller than or equal to correlation with one day lagged time sequence and so on. That's why we see a decreasing trend.

e.

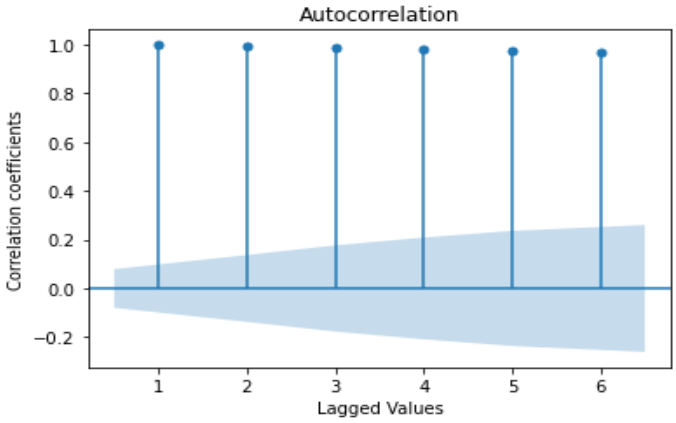

Figure 4 Correlation coefficient vs. lags in given sequence generated using 'plot_acf' function

Inferences:

1. The correlation coefficient values show a decreasing trend with respect to increase in the lag.
2. This is because one-day lagged sequence has some <= 1 correlation with the two-day lagged sequence so, in turn, the original time sequence has a correlation with two-day lagged time sequence smaller than or equal to correlation with one day lagged time sequence and so on. That's why we see a decreasing trend.

2

a. The coefficients obtained from the AR model are 59.955, 1.037, 0.2672, 0.028, -0.175, -0.152.

b.

i.

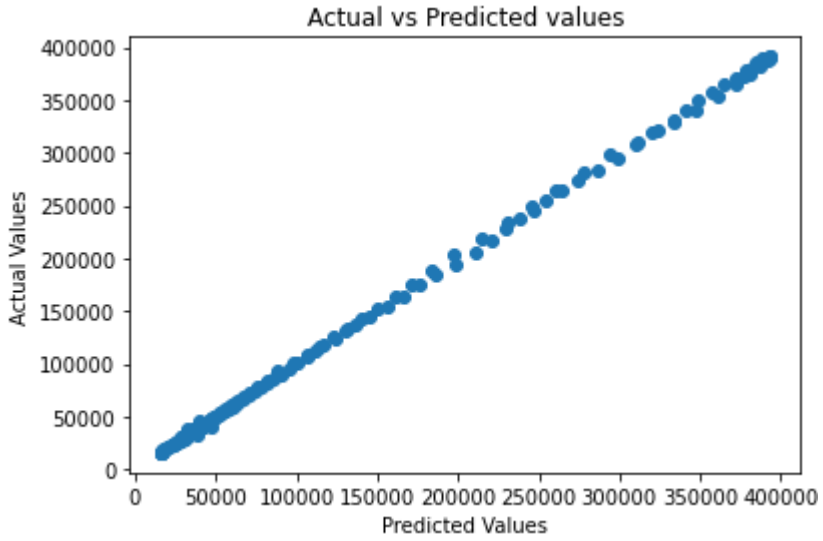

Figure 5 Scatter plot actual vs. predicted values

Inferences:

1. From the nature of the spread of data points, we see that the attributes have a very, very high correlation coefficient.
2. The scatter plot seems to strictly obey the correlation coefficient calculated before.
3. As the lag is increased, more variables are added to our regression model and it inherently improves the fit

ii.

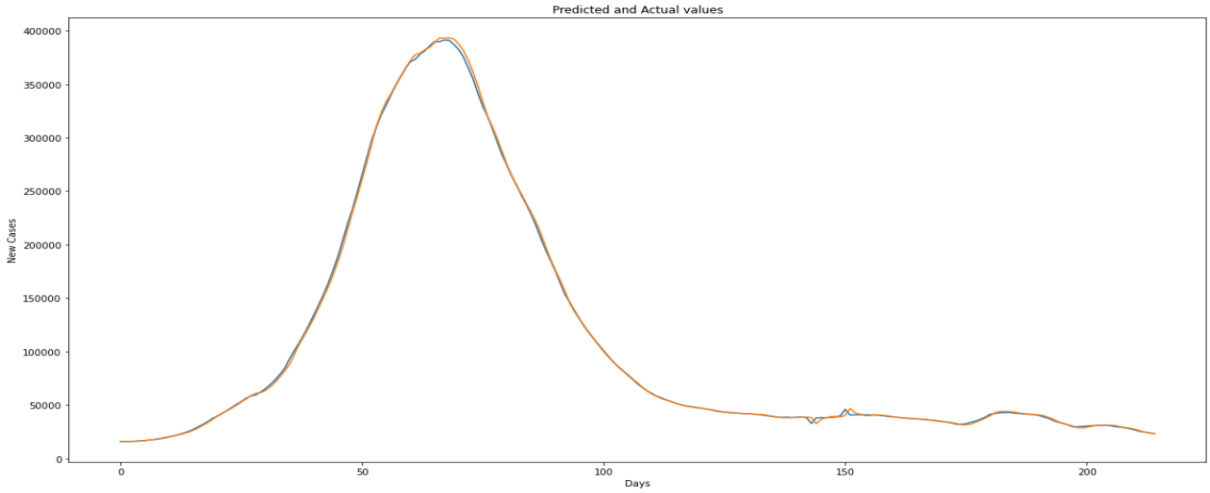

Figure 6 Predicted test data time sequence vs. original test data sequence

Inferences:

1. From the plot of predicted test data time sequence vs. original test data sequence our model is not so suitable and reliable for future predictions because even if it seems to give higher accuracy there is further room for improvement.

iii.

The RMSE(\%) and MAPE between predicted power consumed for test data and original values for test data are 1.885 and 1.587 respectively.

Inferences:

1. From the value of RMSE(\%) and MAPE value we see that our model is not so accurate.
2. Because there is further room for improvement by increasing the value of p.

3

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

Table 1 RMSE (%) and MAPE between predicted and original data values w.r.t lags in time sequence

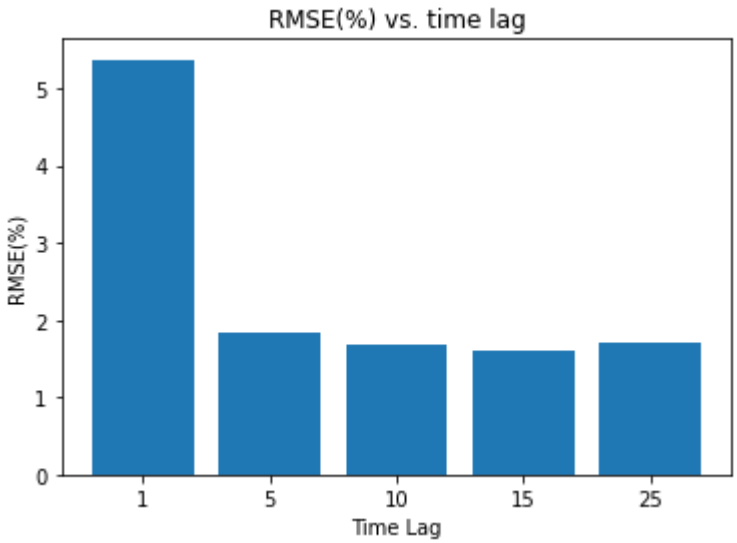

Figure 7 RMSE(%) vs. time lag

Inferences:

1. The RMSE (%) decreases rapidly from time lag 1 to 5 but after that the decrease is very slight or no decrease.
2. This is because a more complex model is needed to fit the data more accurately so when lag is increased from 1 to 5 the accuracy is improved but after that the accuracy almost remains constant.

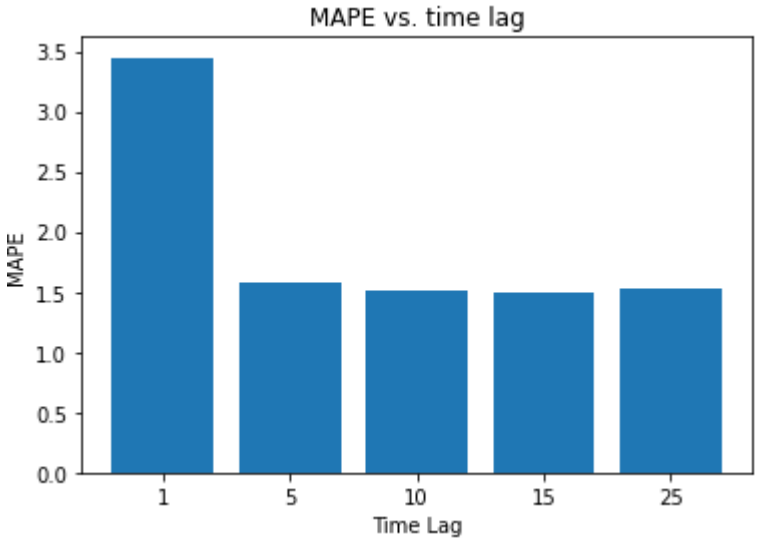

Figure 8 MAPE vs. time lag

Inferences:

1. The MAPE value decreases rapidly from time lag 1 to 5 but after that the decrease is very slight or no decrease.
2. This is because a more complex model is needed to fit the data more accurately so when lag is increased from 1 to 5 the accuracy is improved but after that the accuracy almost remains constant.

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 1.759 and 2.026.

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 time 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.

Guidelines for Report (Delete this while you submit the report):

- **The plot/graph/figure/table should be centre justified with sequence number and caption.**
- **Inferences should be written as a numbered list.**
- **Use specific and technical terms to write inferences.**
- **Values observed/calculated should be rounded off to three decimal places.**
- **The quantities which have units should be written with units.**