**Student’s Name: Vision Aggarwal**

**Roll Number: B20171**

**Mobile No: 637629099**

**Branch:Dse**

# a.

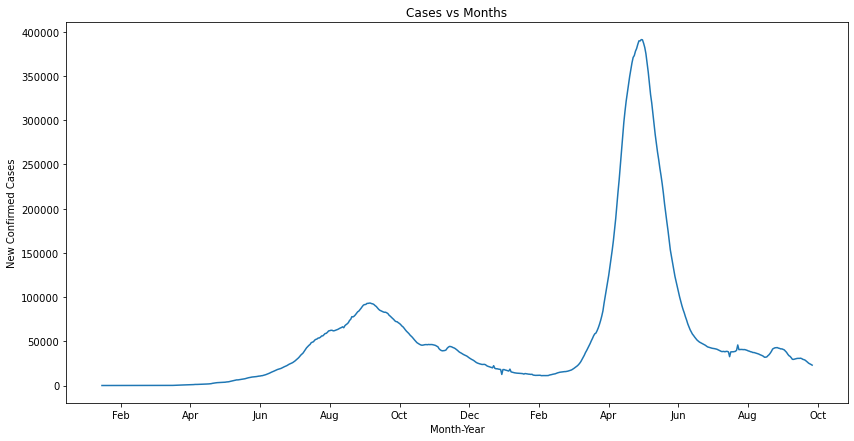


Figure 1 No. of COVID-19 cases vs. days

# Inferences:

1. Yes, completely.
2. Because the correlation between the original cases and lag-1 cases is almost 100% are have nearly the same values.
3. The duration of first wave seems to be 3 months (Aug to Oct) and that of third wave also seems to be 3 months (Apr to June).

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

# Inferences:

1. The degree of correlation is almost one between original and lag-1 series.
2. We generally expect observations (here number of COVID-19 cases) on days one after the other to be similar. In this case, it happens almost 100% correctly based on the correlation coefficient.
3. Reason is that the correlation coefficient is almost 1 which means that the observation is always nearly dependent on the previous day itself.

**c.**

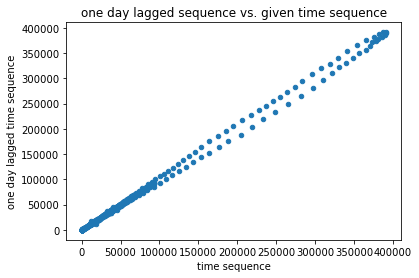


Figure 2 Scatter plot one day lagged sequence vs. given time sequence

# Inferences:

1. We can clearly observe that the graph resembles the y=x line to a great extent showing that both have very high correlation.
2. The graph clearly supports the high Pearson correlation obtained before.
3. Because the graph shows that both attributes have very high high correlation as y is increasing as x increases.

**d.**

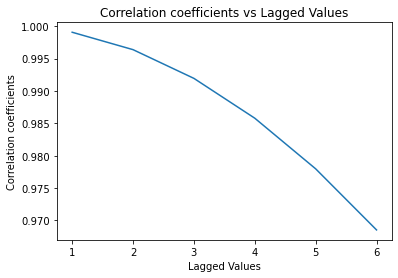


Figure 3 Correlation coefficient vs. lags in given sequence

**Inferences:**

1. As the lags increases, the correlation between lag values and present values decreases.
2. The reason is that the present value is most dependent on the nearest past and dependency decreases as difference of time increases.

**e.**

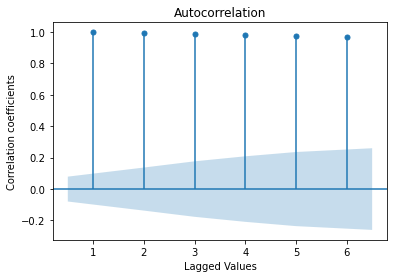


Figure 4 Correlation coefficient vs. lags in given sequence generated using 'plot\_acf' function

# Inferences:

1. As the lags increases, the correlation between lag values and present values decreases.
2. The reason is that the present value is most dependent on the nearest past and dependency decreses as the difference of time increases.

**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.**

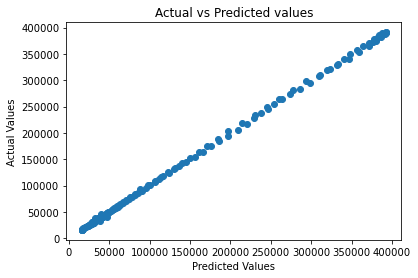


Figure 5 Scatter plot actual vs. predicted values

# Inferences:

1. The two sequences have very strong positive correlation.
2. The scatter plot seem to obey the nature reflected by Pearson’s correlation coefficient completely.

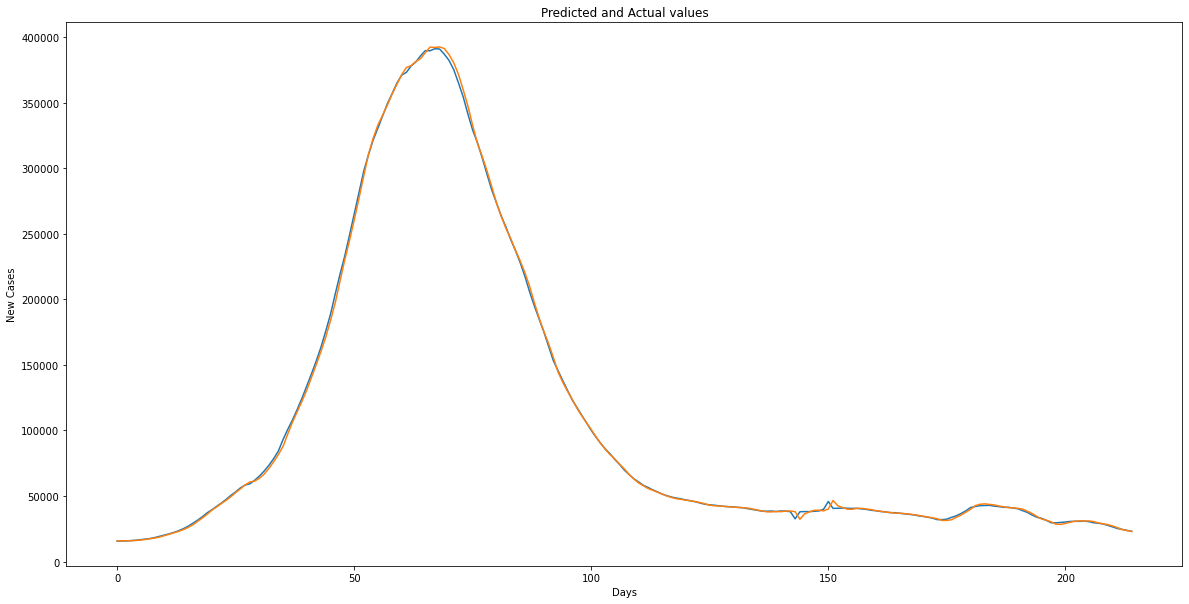
 **ii.**

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 if it only depends upon past observations then this model will be highly useful but if future depends upon other conditions like vaccination also then this model can give wrong results..

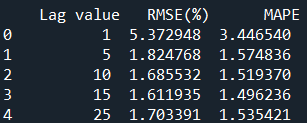
**iii.**

The RMSE(%) and MAPE between predicted power consumed for test data and original values for test data are 1.8247% and 1.57% respectively.

**Inferences:**

1. From the value of RMSE(\%) and MAPE value, We can see that the model is highly accurate.
2. Because both the errors are less than 2%.

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

****

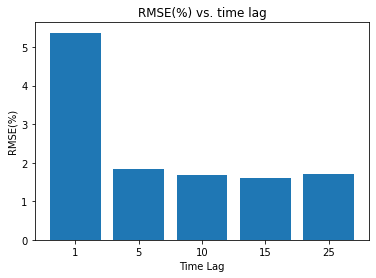


Figure 7 RMSE(%) vs. time lag

**Inferences:**

1. As the lags increases, the value of rmse error decreases.
2. Because the no. of attributes used for regression model are increasing, which leads to higher accuracy.

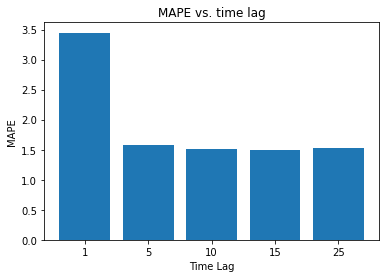


Figure 8 MAPE vs. time lag

**Inferences:**

1. As the lags increases, the value of MAPE error decreases.
2. Because the no. of attributes used for regression model are increasing, which leads to higher accuracy.

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 2.02 respectively.

**Inferences**:

1. Based upon the RMSE(%) and MAPE value, We can say that heuristics for calculating the optimal number of lags doesn’t improve the prediction accuracy of the model.
2. Because when we are using 77 time lags for autoregression model, accuracy of our model slightly decreases because such a big time lag values have very low correlation to present time and so many attributes may lead to unnecessary overfitting.
3. Accuracy for model using heuristics have reduced slightly.