**Student’s Name:** Vishal Sharma

**Roll Number:** B20239

**Mobile No:** 9540140310

**Branch:** Data Science and Engineering

# a.Chart, line chart, histogram Description automatically generated

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

# Inferences:

1. The no of cases day after have a similar rise, except those days where the first and second wave started.
2. It is clear from the graph that the number of cases depend on the number of cases a day before.
3. The first wave started around 20thof August and lasted till October. While the second wave was at its peak around May.

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

# Inferences:

1. Both the time sequences, original and delayed data have a very strong correlation.
2. We generally expect that the number of covid case are like those on the day before. Thus, they have a very strong correlation.
3. It is due to the assumed fact that the no of rise in covid cases is similar to the day before.

**c.**

Chart, scatter chart

Description automatically generated

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

# Inferences:

1. From the above scatter plot, both the data sets are strongly correlated but not perfect 1.
2. Yes, the scatter plot seems to obey the value Pearson’s correlation found in 1.b
3. It is because, there was a sudden rise in the number of covid cases in the first and second wave.

**d.**

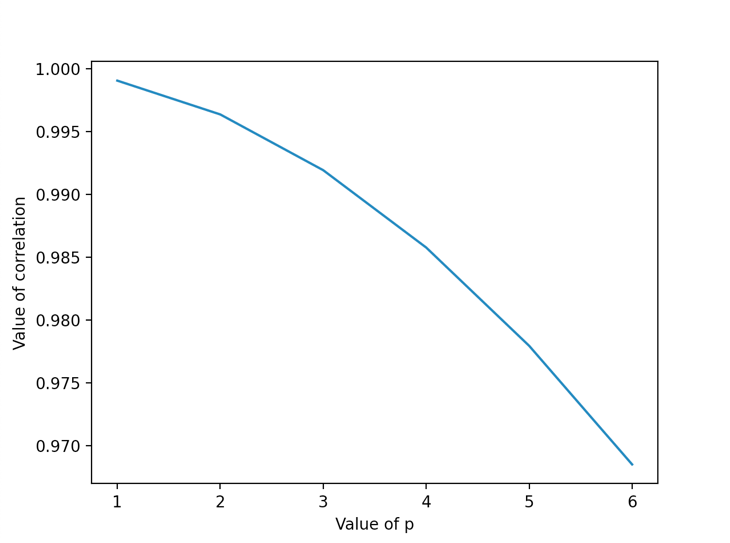
****

Figure 3 Correlation coefficient vs. lags in given sequence

**Inferences:**

1. The value of correlation decreases as the value of ‘lag’ decreases.
2. As the number of lags keeps increasing, the data does not match and it ‘hangs out’, and does not overlap.

**e.**

**Chart, histogram

Description automatically generated**

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

# Inferences:

1. The value of correlation decreases as the value of ‘lag’ decreases.
2. As the number of lags keeps increasing, the data does not match and it ‘hangs out’, and does not overlap.
3. The coefficients obtained from the AR model are 59.954, 1.036, 0.261, 0.027, -0.175, -0.152

**b. i.**

Chart, line chart, scatter chart

Description automatically generated

Figure 5 Scatter plot actual vs. predicted values

# Inferences:

1. From the above scatter plot, they have a very strong correlation between them.
2. Yes, the above scatter plot satisfies the obtained Pearson’s coefficient value.
3. As the lag increases, more and more variables are increased and thus the accuracy keeps increasing.

**ii.**

# Chart, line chart Description automatically generated

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

# Inferences:

1. From the above plot, though the data is accurate, and errors are less, there is still a scope of improvement and the errors can be reduced further.

**iii.**

The RMSE(\%) and MAPE between predicted power consumed for test data and original values for test data are 1.824 and 1.574 respectively.

**Inferences:**

1. From the errors, it is clear that though the errors are less, it is still high for a model to be reliable.
2. Because for a larger value of ‘lag’, the errors will decrease.

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

|  |  |  |
| --- | --- | --- |
| **Lag value** | **RMSE (%)** | **MAPE** |
| 1 | 5.372 | 3.446 |
| 5 | 1.824 | 1.574 |
| 10 | 1.685 | 1.519 |
| 15 | 1.611 | 1.496 |
| 25 | 1.703 | 1.535 |

Chart, bar chart

Description automatically generated

Figure 7 RMSE(%) vs. time lag

**Inferences:**

1. The RMSE error decreases when the lag is increased from 1 to 15, but a slight increase when the lag is increased from 15 – 25. Also, the error reduces quickly from 1 to 5.
2. This is because, the error reduces as the value of ‘lag’ is increased. But the increase in error is because at ‘lag’ = 25, the model becomes quite complex and can’t fit the data more accurately.

Chart, bar chart

Description automatically generated

Figure 8 MAPE vs. time lag

**Inferences:**

1. The RMSE error decreases when the lag is increased from 1 to 15, but a slight increase when the lag is increased from 15 – 25. Also, the error reduces quickly from 1 to 5.
2. This is because, the error reduces as the value of ‘lag’ is increased. But the increase in error is because at ‘lag’ = 25, the model becomes quite complex and can’t fit the data more accurately.

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 respectively.

**Inferences**:

1. Based on the value of error, the value of heuristic for calculation the optimal number of ‘lags’ did not increase the accuracy of the model. It can be seen that the value of RMSE for ‘lag’ = 10 is less than the error at optimal lag.
2. Because as the value of lag keeps increasing, the error becomes random, and it becomes hard to predict the error value.
3. The prediction accuracies obtained without heuristic is less as compared to error values found with heuristic.