Forecasting Analytics | Dr. Ahmed Aziz Ezzat

Final Project Report

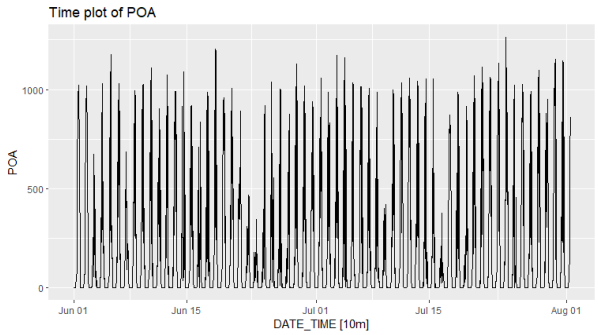
Team – Forecast Wizards

Name	**Contribution**
Simran Nair	Data Exploration
Dyuti Soudarapu	Data Visualization
Dhruvil Patel	Model Creation & Results
Pujita Vijayakumar	Model Creation & Results

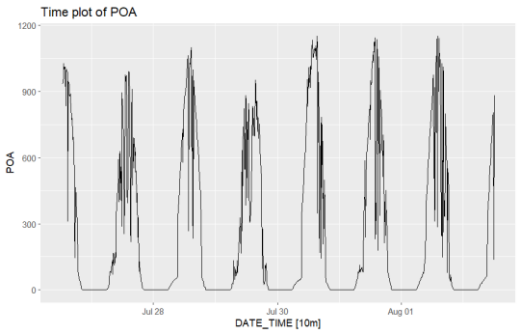
Contents

Data Exploration

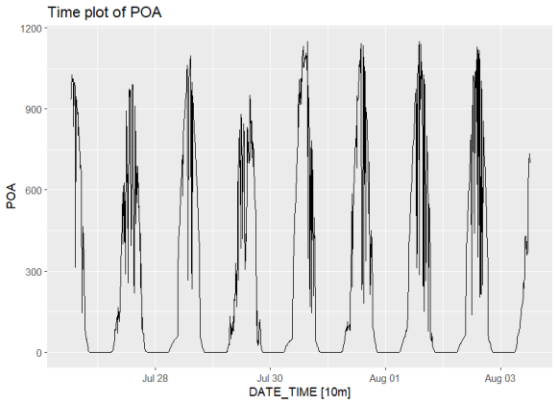
Time Series Plot



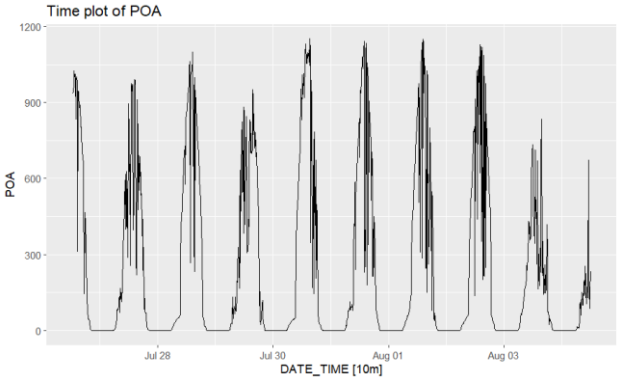
Dataset - 1



Dataset - 2



Dataset - 3

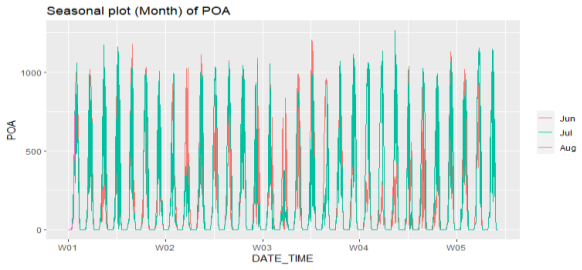


Dataset - 4

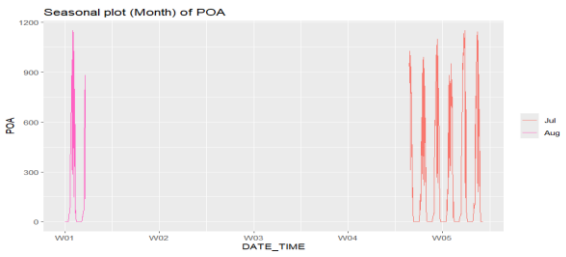
Insights:

- The plot reveals daily cycles with periods of high and low values, likely reflecting a 24-hour pattern. This is indicated by the regular intervals between the peaks which seem to align with daily timing.
- There is significant variability within each day, as shown by the fluctuating heights of the peaks and the depths of the troughs. This reflects different rates of activity or energy during different times of the day.
- The troughs, which could represent nighttime or periods of inactivity, are quite consistent, suggesting that the lows of the cycle are more stable than the highs.
- The peaks, which are during the daytime, show more variability. There are occasional spikes which exceed the general pattern of the peaks, indicating periods of exceptionally high readings or activity.
- The starting and ending values each day seem relatively stable, implying that the cycle resets to a similar baseline each day.

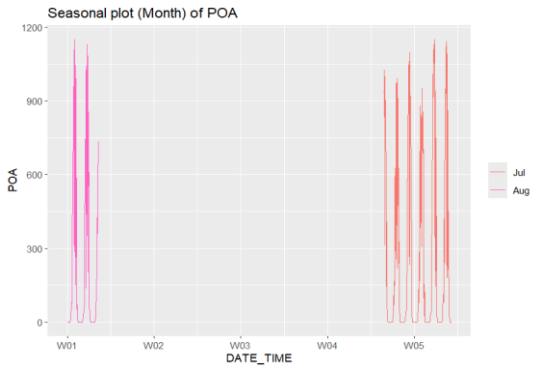
Monthly Seasonality



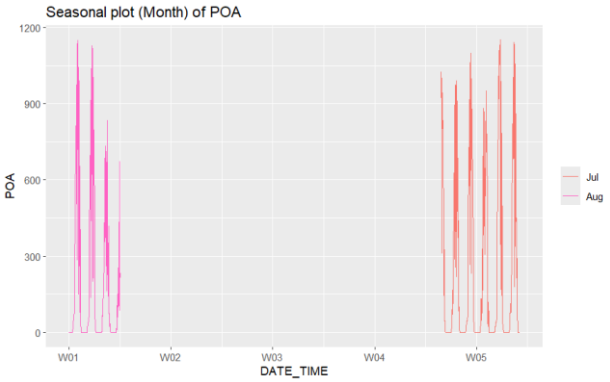
Dataset - 1



Dataset -2



Dataset – 3

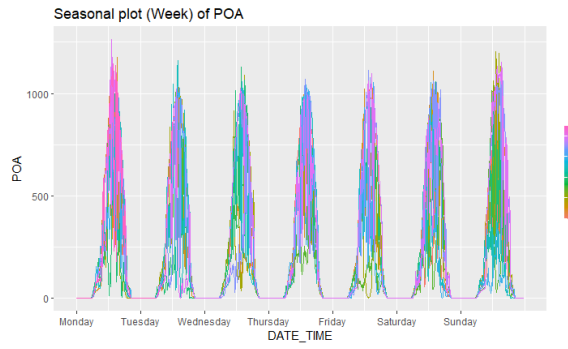


Dataset – 4

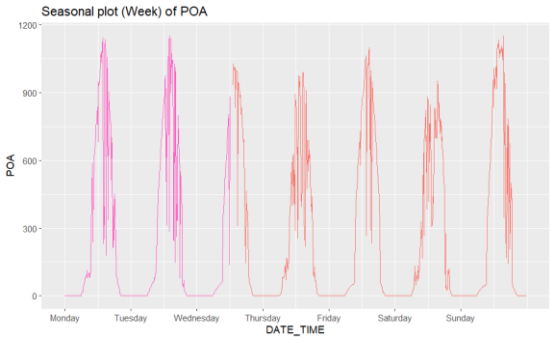
Insights:

- The plots show a comparison between months. Each month's data is distinguished by colour, allowing for a visual comparison of trends and patterns across June, July, and August.
- There is a strong similarity in the pattern from week to week across the different months, indicating that the factor being measured has a consistent weekly cycle.
- Although the weekly pattern is consistent across months, there are differences in the data's magnitude in some weeks. This could be due to weekly fluctuations in activity or output.

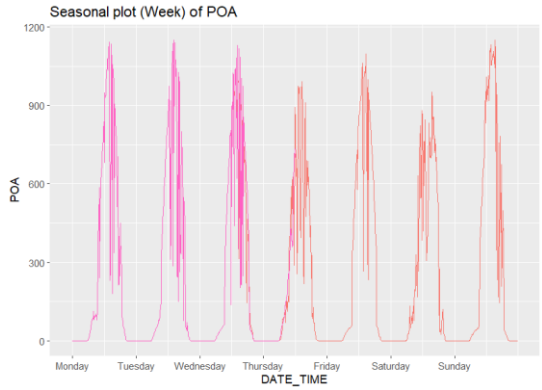
Weekly Seasonality



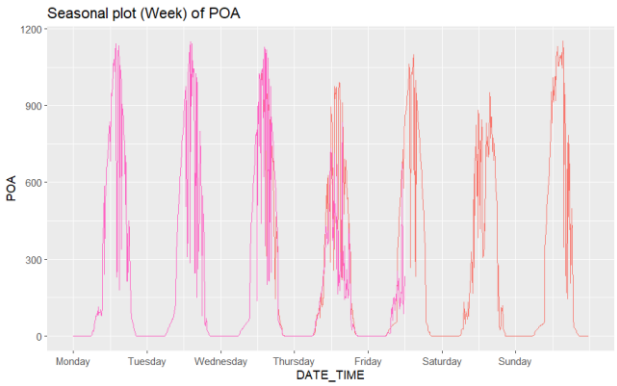
Dataset – 1



Dataset – 2



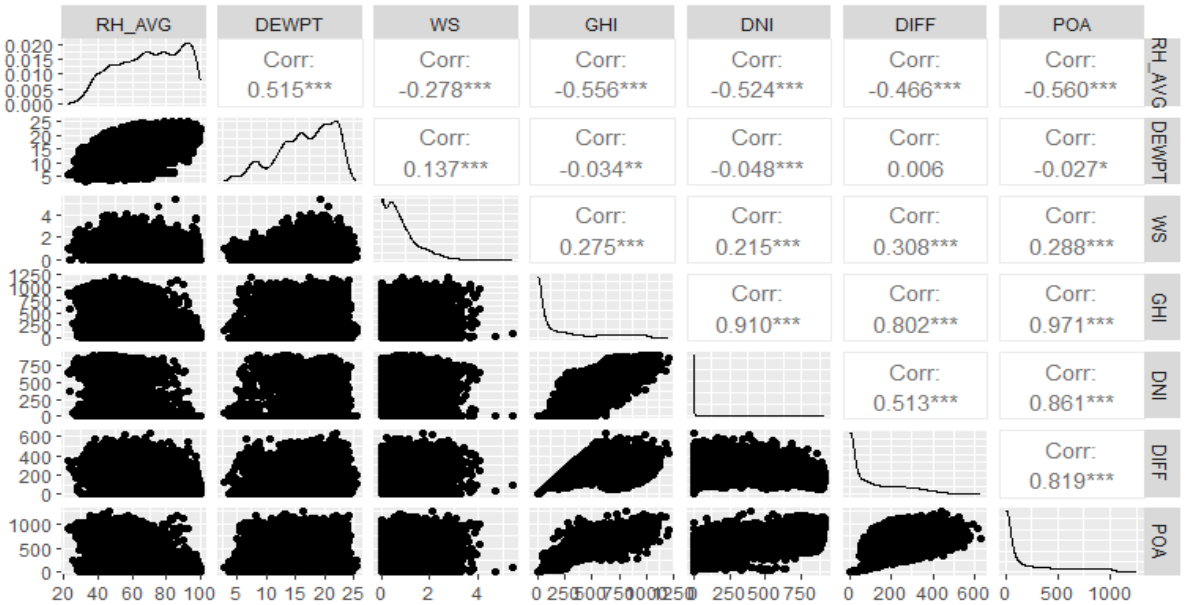
Dataset – 3



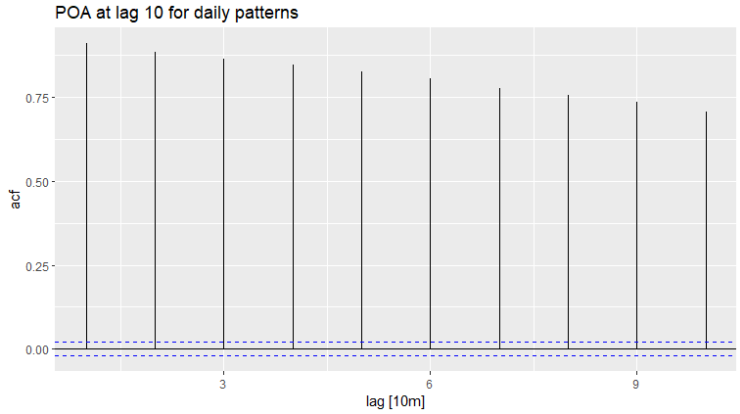
Dataset – 4

Insights:

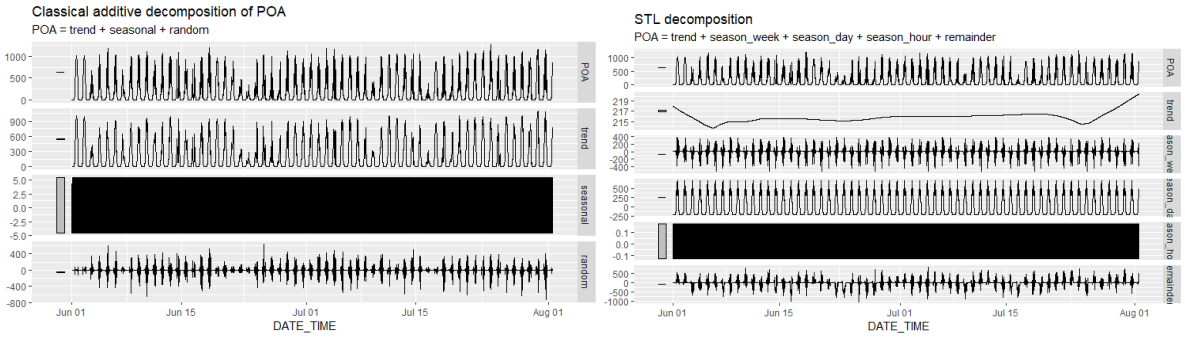
- There is a clear daily pattern, with peaks and troughs indicating higher and lower POA values. This pattern repeats every week, suggesting a strong daily seasonality in the data.
- Different weeks are color-coded, showing a high degree of similarity across the weeks, which indicates consistency in weekly patterns.
- The plots often show a distinct pattern for the weekend days (Saturday and Sunday), which may have either higher or lower POA values compared to weekdays. This implies different operational schedules or activity levels during weekends.
- Certain weekdays, like Tuesdays and Wednesdays, sometimes show different levels of activity. It suggests that the middle of the week could have factors influencing the POA differently than the start or end of the week.

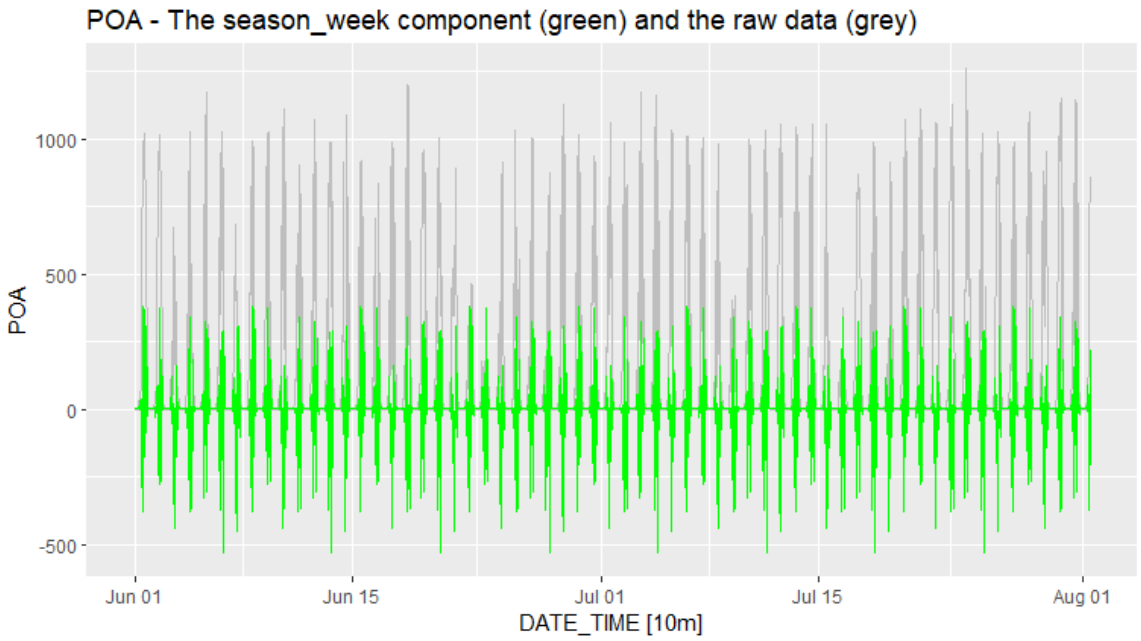


The graph was constructed with the aim of identifying correlations among the variables and the target variable. This is done for the first dataset only, as other datasets show nearly the same correlation values.



Above graph is just to visualize ACF before creating models





The data decomposition is only done on the first dataset just to have insights of the presence of both daily and weekly seasonality within the data set. This insight enables us to make informed decisions regarding the appropriate models to capture these underlying trends effectively. Here, we decided to use three models which are the Multiple Linear Model, ARIMA-X Model and Dynamic Regression Model. Then we will decide to use one of the models which has higher accuracy.

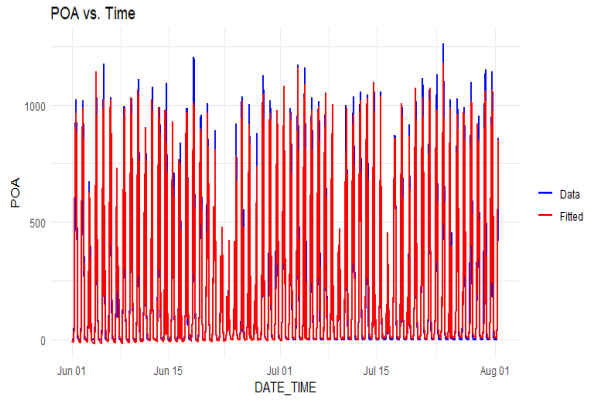
Model Creation and Selection

Three models - Multiple Linear Model, ARIMA-X Model, Dynamic Regression Model and Random Forest Models were utilized. Employing the segmented averages method, the data was grouped based on the same time-of-day, regardless of the date, into single segments. Subsequently, the average value for each segment was computed to generate future values for exogenous variables. These average values were then used to forecast for the upcoming 12-hour period (test data). We have done ADF and Ljung Box Test to do evaluation of the models.

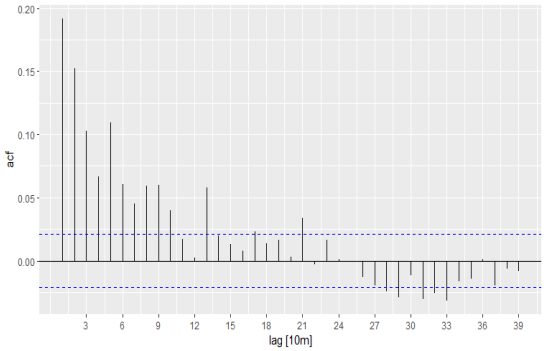
Multiple Linear Regression Model (Approach 1)

Y = a1X1 + a2X2 + ... + a7X7

Y = Output, X= Inputs, a= Regression Coefficient



Model Fitting



Residual ACF Plot

Stationary Test

```
        Augmented Dickey-Fuller Test

data:  LM_RESI$.resid
Dickey-Fuller = -8.2167, Lag order = 10, p-value = 0.01
alternative hypothesis: stationary
```

Insights: With a p-value of 0.01, which is less than the typical significance level of 0.05, we reject the null hypothesis and conclude that the time series (or residuals) is stationary.

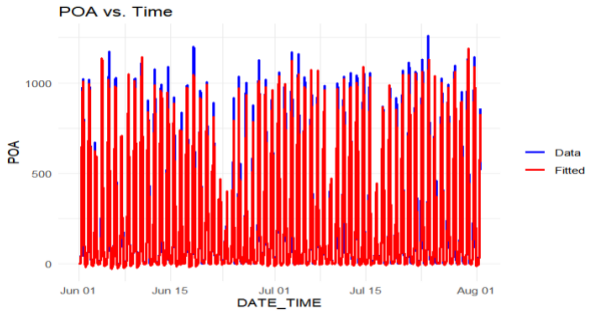
Ljung_Box Test

```
> ljung_box(LM_RESI$.resid)
     lb_stat     lb_pvalue
3.751557e+01 9.068628e-10
```

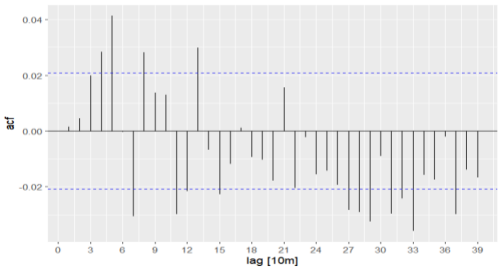
Insights:

Since the p-value is high (greater than 0.05), it indicates that there is no significant auto-correlation present in the residuals of the model. This suggests that the model adequately captures the autocorrelation structure in the data, and the residuals appear to be random and uncorrelated. Therefore, the model may provide a good fit to the data, and the assumptions of the model regarding residual auto-correlation may be satisfied.

ARIMA-X (Approach 2)



Model Fitting



ACF Residual Plot

Stationary Test

```
        Augmented Dickey-Fuller Test

data:  ARIMAX_RESI$.resid
Dickey-Fuller = -20.908, Lag order = 20, p-value = 0.01
alternative hypothesis: stationary
```

Insights With a p-value of 0.01, which is less than the typical significance level of 0.05, we reject the null hypothesis and conclude that the time series (or residuals) is stationary.

Ljung_Box Test

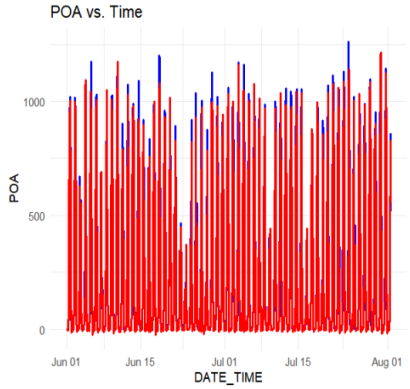
```
   lb_stat lb_pvalue
0.0226904 0.8802650
```

Insights This suggests that there is significant auto-correlation present in the residuals of your model. A p-value close to zero indicates strong evidence against the null hypothesis of no autocorrelation, suggesting that the auto correlation is statistically significant.

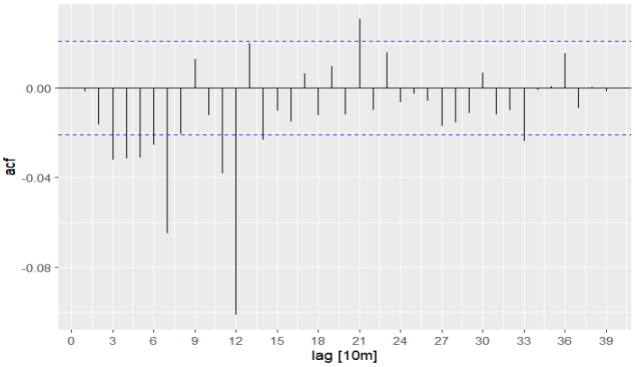
Parameters

ARIMA(1,0,2)(0,0,1)[6] is used
Non-seasonal part (ARIMA(1,0,2))
Autoregressive order (p): 1
Differencing (d): 0
Moving average order (q): 2
Seasonal part (ARIMA(0,0,1)[6])
Seasonal autoregressive order (P): 0
Seasonal differencing (D): 0
Seasonal moving average order (Q): 1
Seasonal period (s): 6

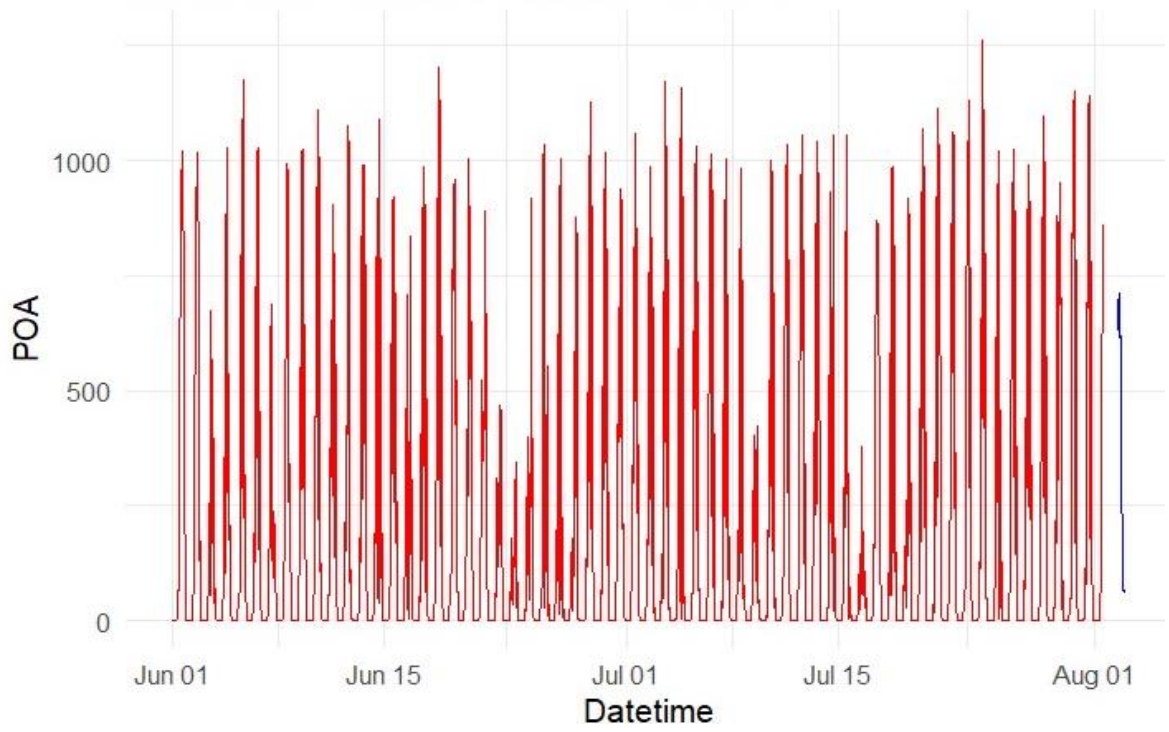
Dynamic Regression (Approach 3)



Model Fitting



ACF Residual Plot



Final Plot of Actual Data & Forecasted Values for Data_S1 using Dynamic Regression

Stationary Test

```
        Augmented Dickey-Fuller Test

data:  cleaned_residuals
Dickey-Fuller = -25.589, Lag order = 20, p-value = 0.01
alternative hypothesis: stationary
```

Insights: With a p-value of 0.01, which is less than the typical significance level of 0.05, we reject the null hypothesis and conclude that the time series (or residuals) is stationary.

Ljung_Box Test

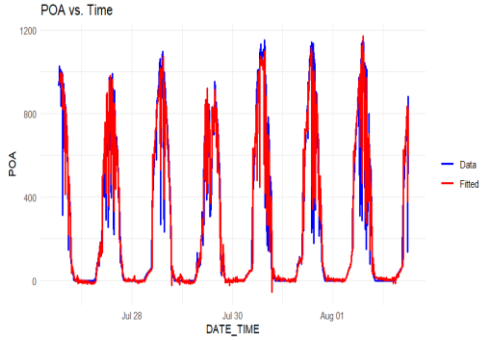
```
> ljung_box(DR_RESI$.resid)
   lb_stat  lb_pvalue
0.01982727 0.88802058
```

Insights: Since the p-value is high (greater than 0.05), it indicates that there is no significant auto-correlation present in the residuals of the model. This suggests that the model adequately captures the auto-correlation structure in the data, and the residuals appear to be random and uncorrelated. Therefore, the model may provide a good fit to the data, and the assumptions of the model regarding residual auto-correlation may be satisfied.

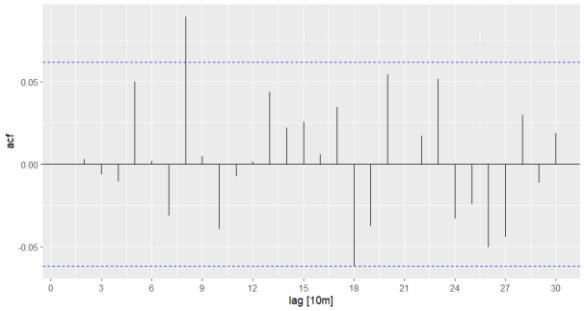
Parameters

ARIMA(3,0,0) model is used
Autoregressive terms (p = 3)
Differencing to make it stationary (d = 0)
Moving average terms (q = 0)

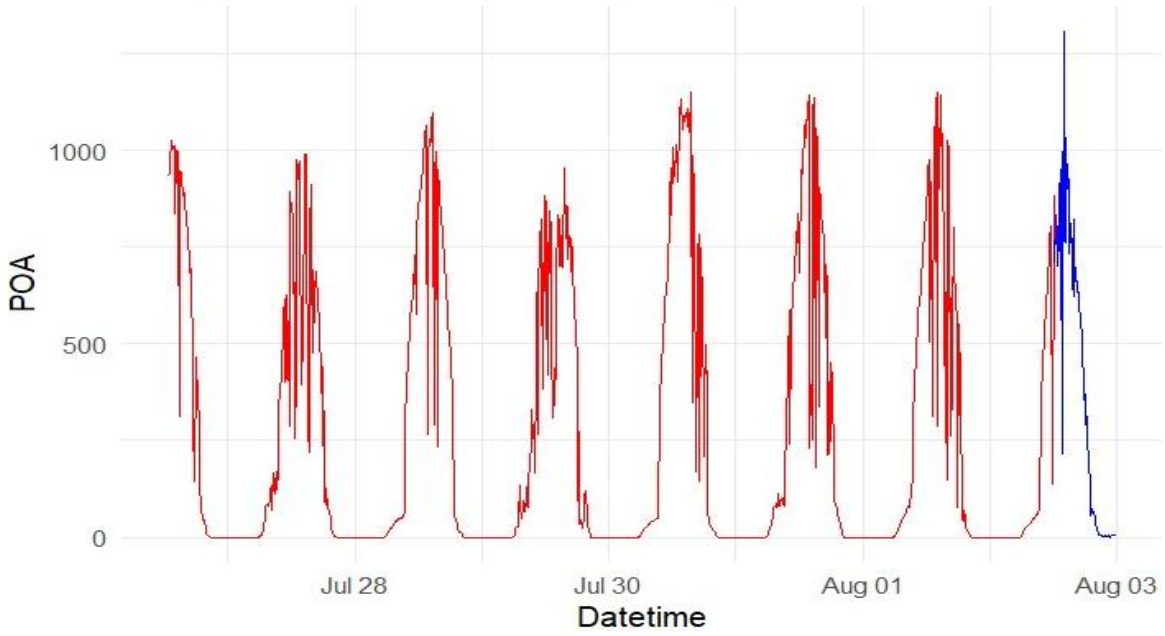
Continued Dynamic Regression Model for the 2nd Submission



Model Fitting



ACF Residual Plot



Final Plot of Actual Data & Forecasted Values for Data_S2

Stationary Test

```
        Augmented Dickey-Fuller Test

data:  DR_RESI
Dickey-Fuller = -8.5481, Lag order = 8, p-value = 0.01
alternative hypothesis: stationary
```

Insights: With a p-value of 0.01, which is less than the typical significance level of 0.05, we reject the null hypothesis and conclude that the time series (or residuals) is stationary.

Ljung_Box Test

```
      lb_stat   lb_pvalue
0.000288727 0.986443023
```

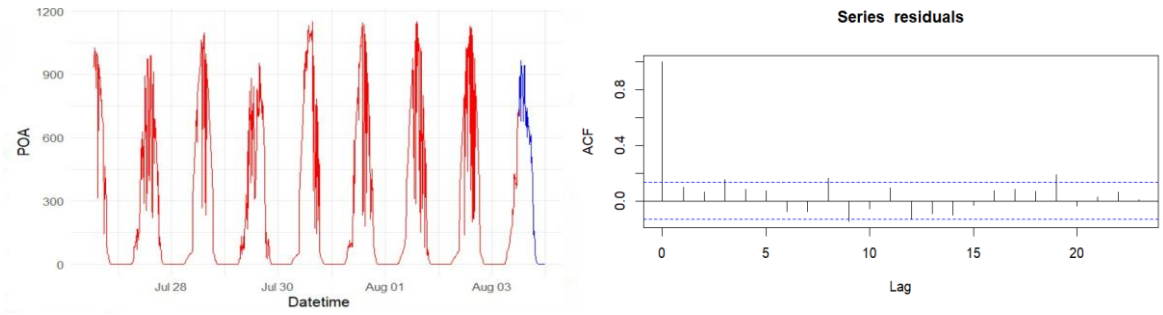
Insights: After observing a satisfactory fit and examining the ACF plot, we opted to go with the model.

Parameters

ARIMA(4,0,0)(2,0,0)[6] is used in the model
Non-seasonal part: ARIMA(4,0,0)
AR order (p) = 4 (there are 4 autoregressive terms)
Differencing (d) = 0 (no differencing)
MA order (q) = 0 (there are no moving average terms)
Seasonal part: ARIMA(2,0,0)[6]
Seasonal AR order (P) = 2 (there are 2 seasonal autoregressive terms)
Seasonal differencing (D) = 0 (no seasonal differencing)
Seasonal MA order (Q) = 0 (there are no seasonal moving average terms)
Seasonal period (s) = 6 (there are 6 observations per season)

Random Forest for the 3rd submission (Approach 4)

Random forest is a flexible and non-linear model that can capture complex relationships between predictors and the target variable. It works well with high-dimensional data and is robust to outliers and noise. Therefore, we decided to go with Random Forest.



Final Plot of Actual Data & Forecasted Values for Data_S3 using Random Forest

Stationary Test (Augmented Dickey-Fuller Test)

```
Dickey-Fuller = -5.1859, Lag order = 7, p-value = 0.01
alternative hypothesis: stationary
```

Insights: With a p-value of 0.01, which is less than the typical significance level of 0.05, we reject the null hypothesis and conclude that the time series (or residuals) is stationary.

Ljung_Box Test

```
data:  residuals
X-squared = 49.85, p-value = 0.0002327
```

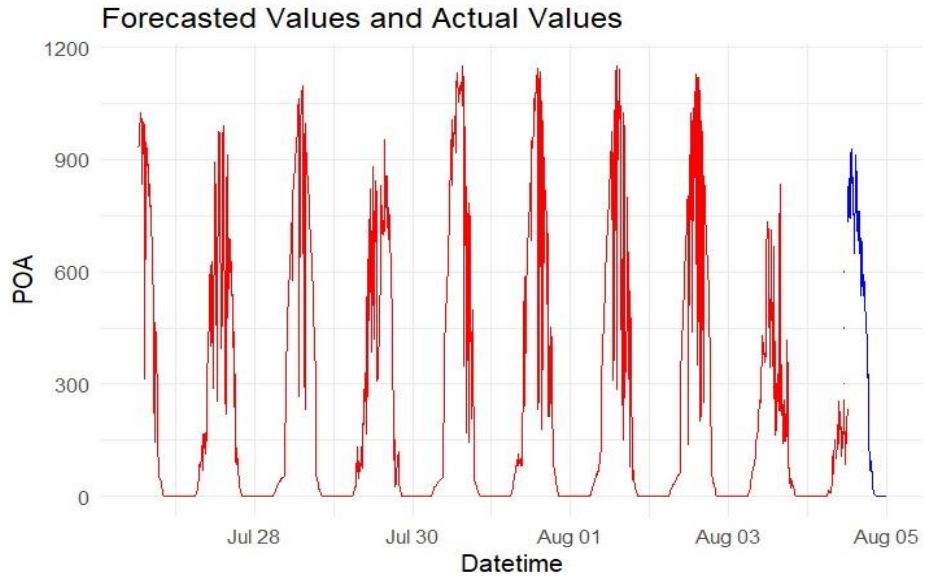
Parameters

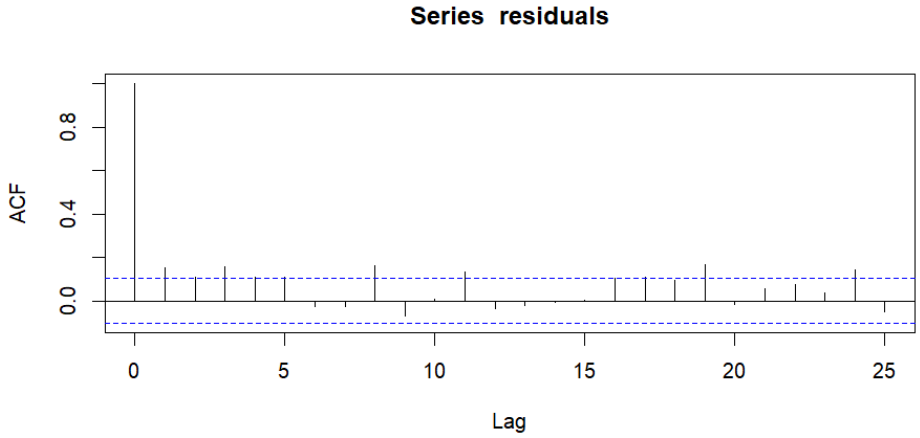
- **n_tree**: Denotes the quantity of trees incorporated into the forest. A value of 100 has been selected for n_tree
- **m_try**:This parameter controls the number of features randomly sampled at each split when growing each tree in the forest m_try = 3 is taken. It is usually taken square root of the variables

Using Baseline Models in Forecasting Comparison (Data_S4)

Model Used	MAE	RMSE	Comments
Simple Average	262.35	299.59	The Simple Average method shows better performance than both the Seasonal Naive and Drift methods in terms of RMSE, but still lags behind the Naive method.
Naive	63.06	168.14	The Naive method outperforms the other methods in terms of both RMSE and MAE, suggesting that the most recent past observation is a strong predictor for the next value.
Seasonal Naive	430.67	544.64	The Seasonal Naive method has the highest RMSE and MAE, which might indicate that the seasonal pattern assumed does not align well with the actual seasonal variations.
Drift method	367.24	464.95	The Drift method also performs poorly compared to the Naive method, suggesting that a linear trend does not accurately capture the future values beyond the immediate next steps.

Result and Conclusion





Final Plot of Actual Data & Forecasted Values for Data_S4 using Random Forest

Stationary Test (Augmented Dickey-Fuller Test)

```
data:  residuals
X-squared = 73.317, p-value = 5.186e-08
```

Parameters

- **n_tree**: Denotes the quantity of trees incorporated into the forest. A value of 60 has been selected for n_tree
- **m_try**:This parameter controls the number of features randomly sampled at each split when growing each tree in the forest m_try = 3 is taken. It is usually taken square root of the variables

Accuracy cannot be calculated for the TEST due to unavailability of the actual values. Data is divided into 70% train and 30% test. The model was trained for the train dataset and forecasted for the test.

	TSLM	ARIMA-X	DR	Random Forest
MAE 1	42.7	47.6	40.8	-
RMSE 1	83.5	86.4	80.4	-
MAE 2	60.8	73.1	57.6	-
RMSE 2	105	116	105	-
MAE 3	50.3	60.7	51.6	48.86
RMSE 3	96.6	97.5	101	101
MAE 4	35.5	33.5	35.5	30.65
RMSE 4	68.3	60.7	68.3	69.28

Conclusion

- Based on the analysis conducted on the dataset, it can be concluded that Random Forest demonstrates suitability for the provided data due to its capability to capture non-linear relationships among the variables. The model's robustness in handling complex interactions between predictors makes it a preferable choice in this context.
- ARIMA-X and Dynamic Regression models, adept at capturing seasonality patterns, may not be as effective in scenarios where non-linear relationships exist among the variables.
- One significant challenge encountered in the analysis is the incorporation of cloudy conditions to enhance the predictive capability of the models. Cloudy conditions can significantly impact solar energy generation, but integrating this factor into the forecasting models remains a complex task.
- Future research and model development efforts may be directed towards devising innovative approaches to incorporate cloudy conditions into predictive modelling frameworks, thereby enhancing the accuracy and reliability of future energy generation forecasts.