Forecasting Energy Availability From Renewable Sources Using Weather Data

Aashik Sharif Basheer Ahamed - 011870531

Sheheryar Ahmad Pirzada - 011869749

Problem Statement

Forecasting renewable energy generation, specifically from solar panels, is crucial for optimizing energy management and supporting sustainable practices. Variations in weather conditions, such as temperature, solar radiation, and dew point, significantly impact energy production. This project aims to develop a predictive machine learning model that estimates solar energy generation using weather data, enabling better resource planning and decision-making.

Goals and Objectives

- **Analyze Weather Impacts:** Identify key meteorological factors like temperature, solar radiation, and dew point that influence solar energy generation and find correlation in between them.

- **Develop Predictive Models:** Build machine learning models (e.g., Random Forest, Gradient Boosting) to forecast solar energy output at least 24 hours in advance.

Seattle Weather Data

Source: Visual Crossing Weather

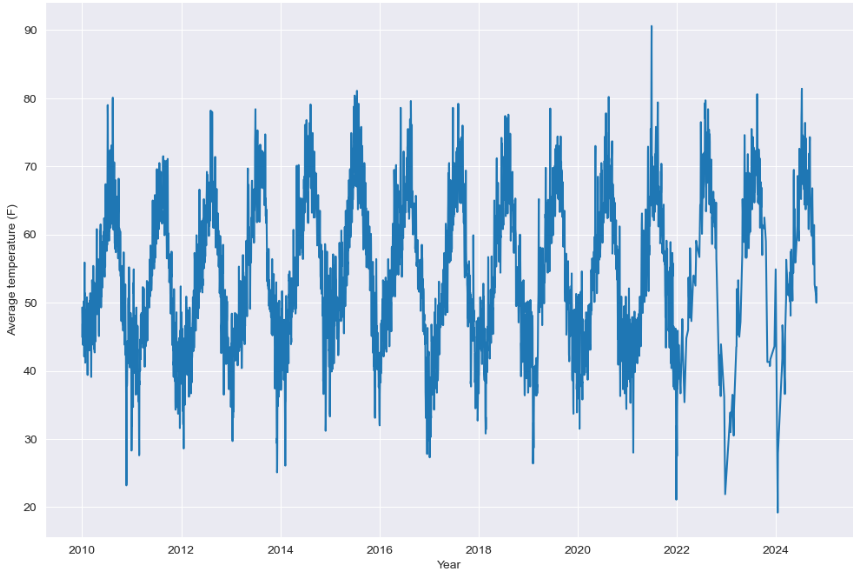
A Comprehensive Dataset for Seattle Weather data

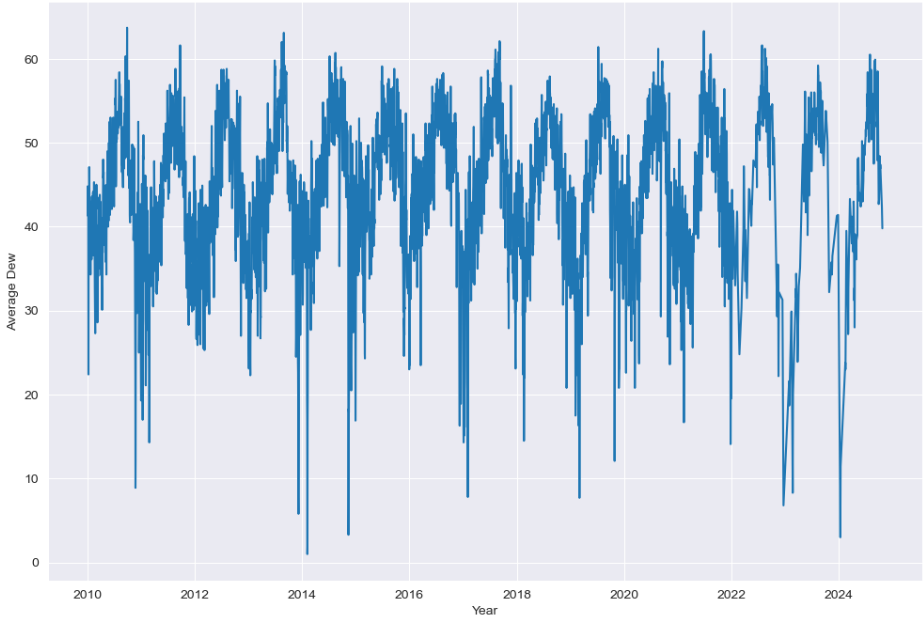
- Source: Visual Crossing
- No. of features: 32
- No. of samples: 5419
- Some key features include:
  - temp
  - dew
  - precip
  - snow
  - windspeed
  - visibility
  - cloudcover
  - solarradiation
  - solarenergy
- Date range: 2010 - 2024

Data Preprocessing

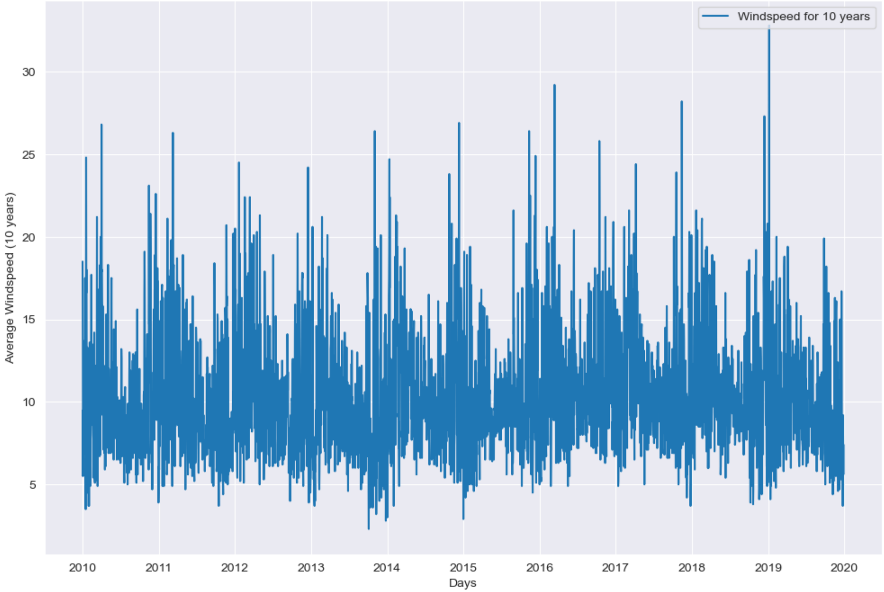
- Replacing missing / na values with appropriate values.
- Applied various interpolation techniques between temperature data.
- Processing date column into day, month and year columns.
- Scaling and Normalization
- Feature Engineering & Visualization

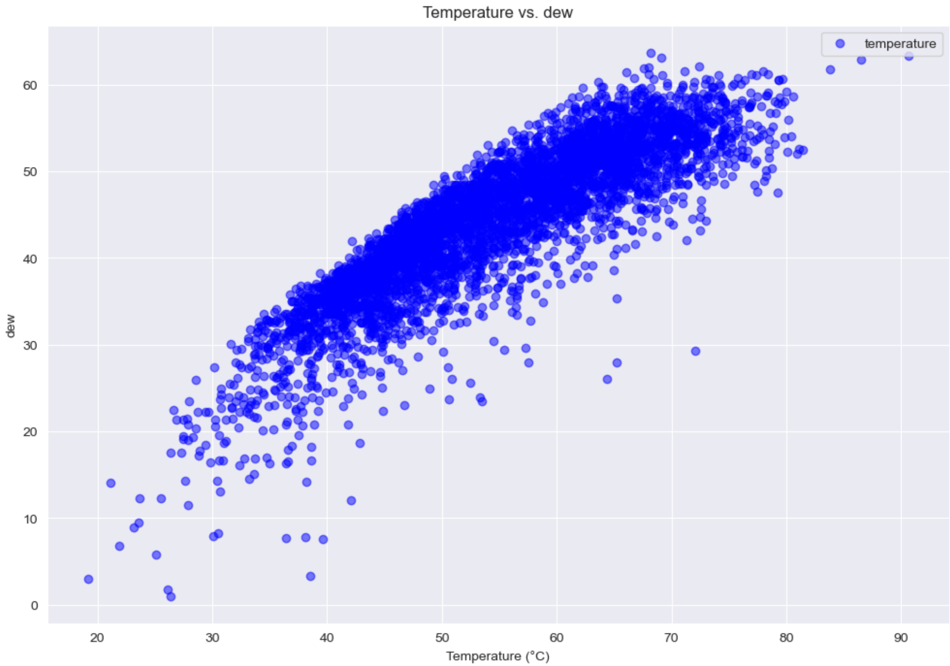
Data Visualisation



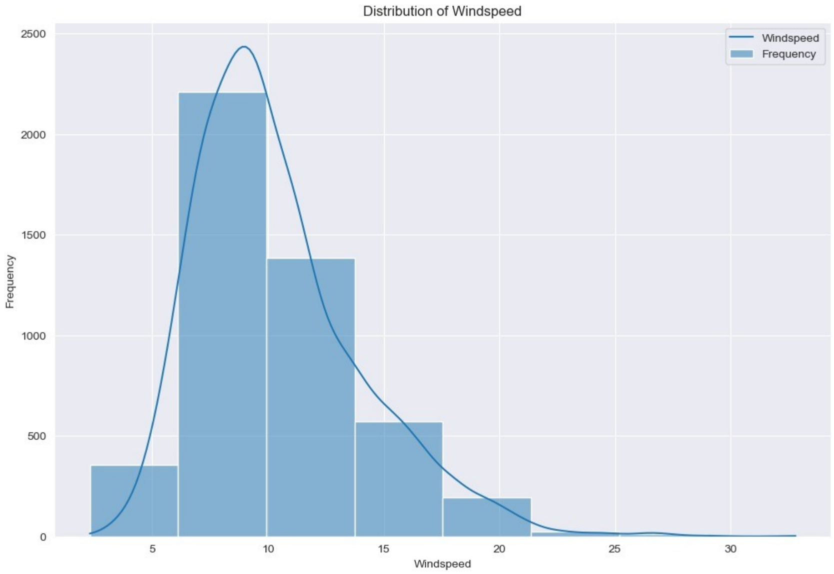


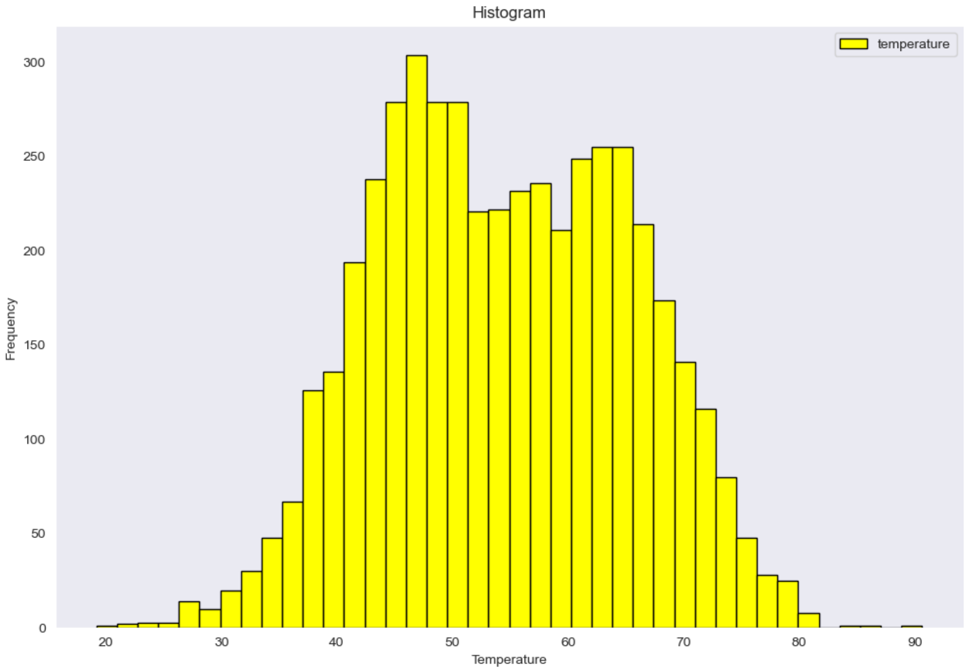
Data Visualisation



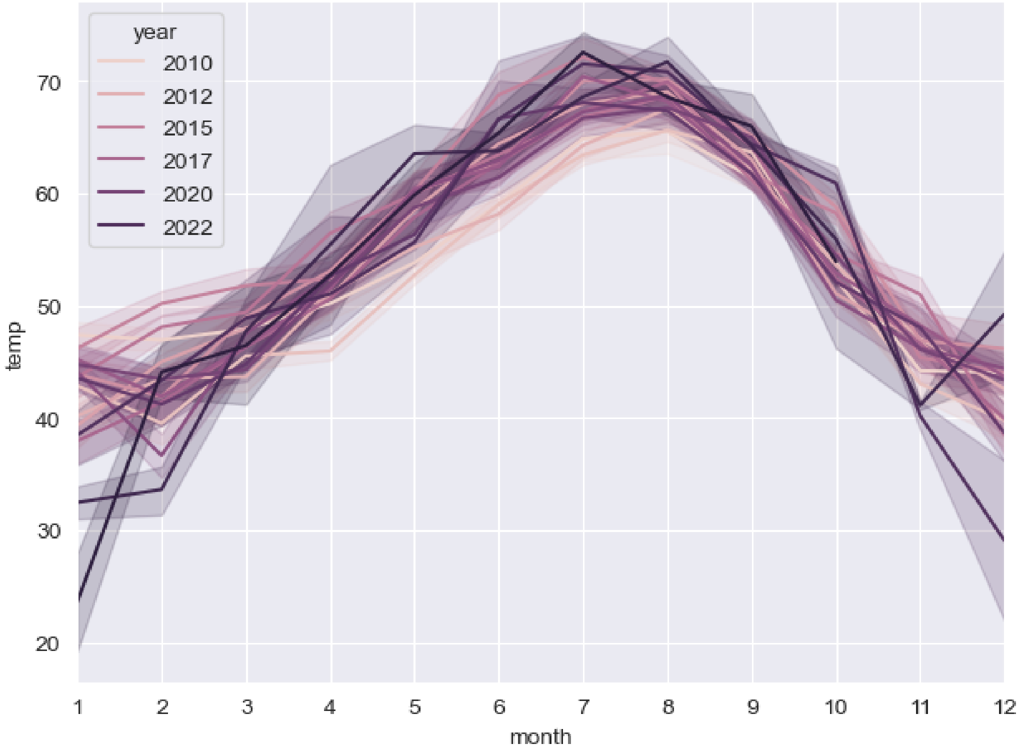


Data Visualisation

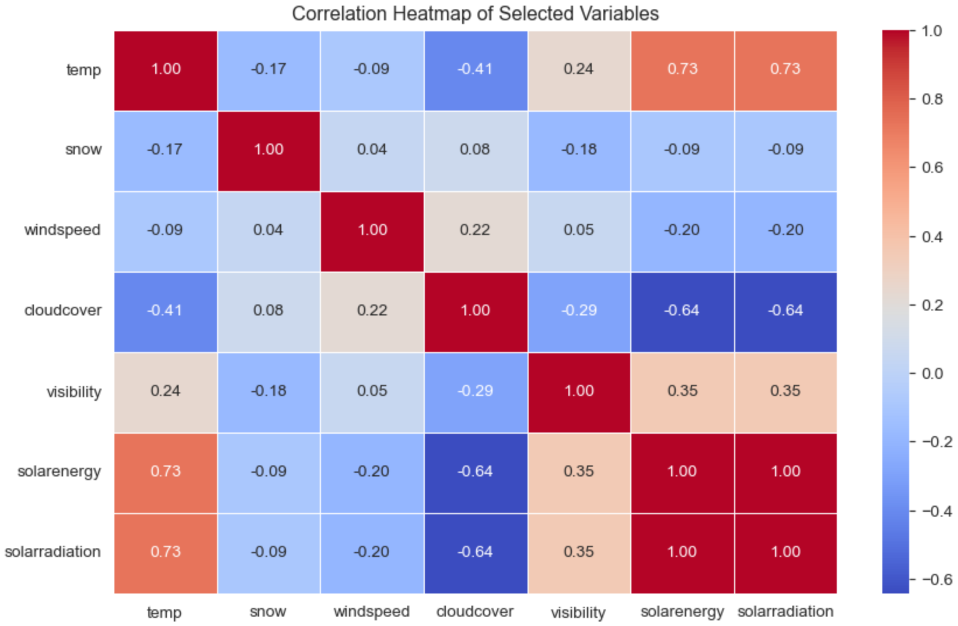


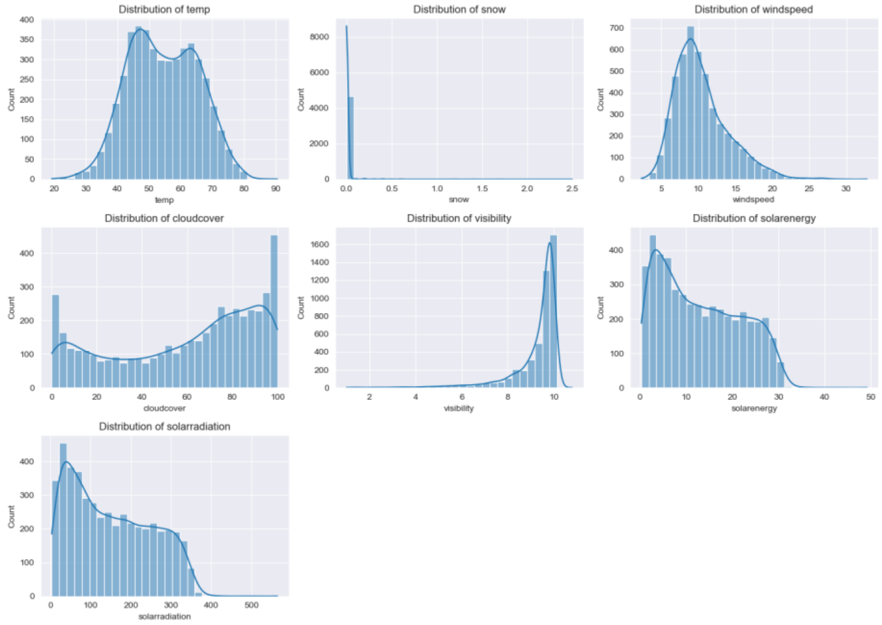


Data Visualisation

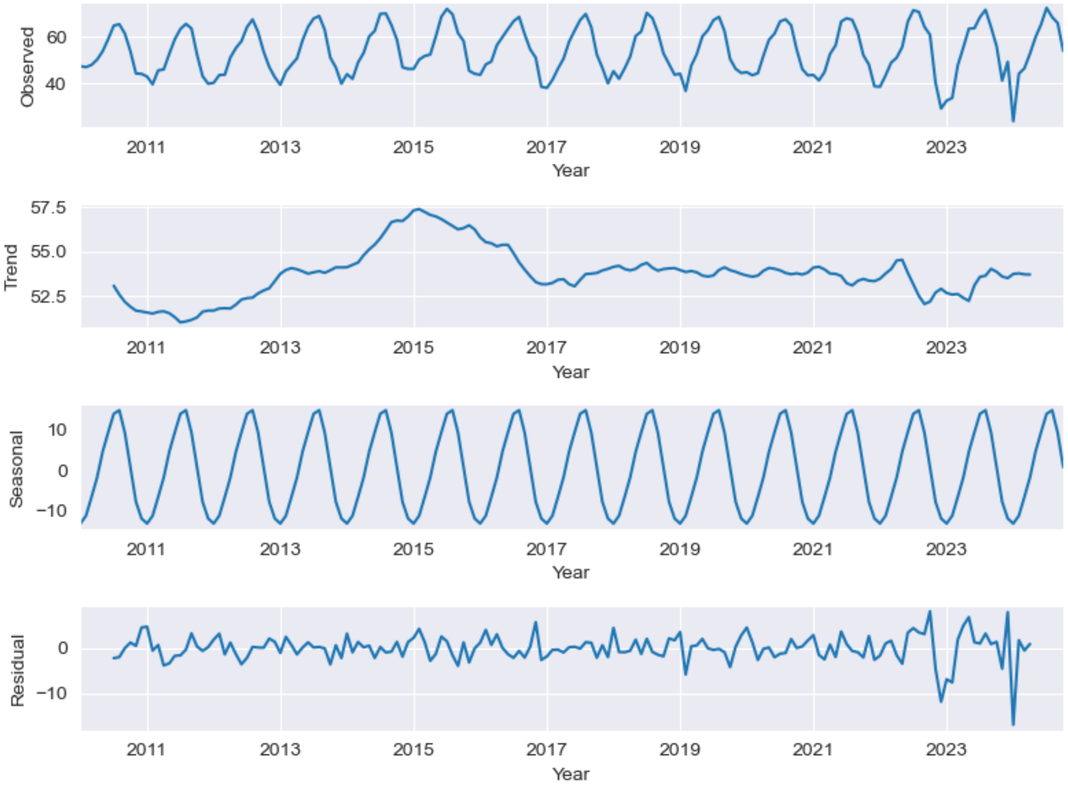


Exploratory Data Analysis (EDA)





Time Series Decomposition for Temperature



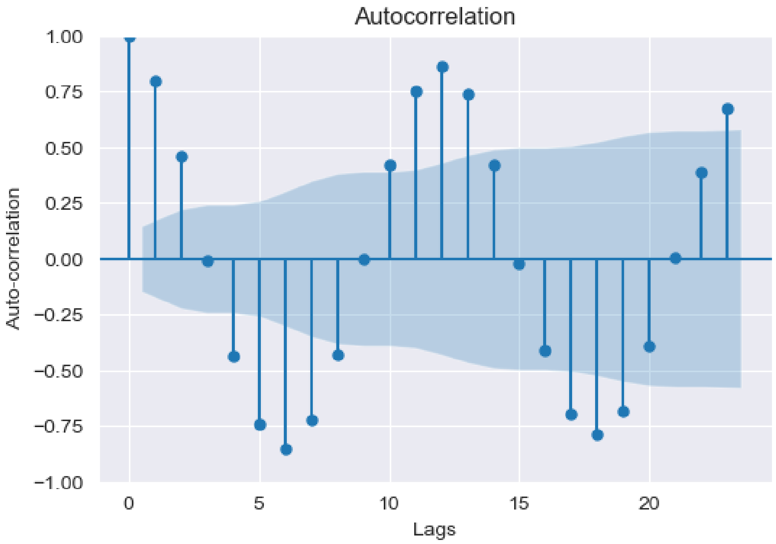
CHECKING STATIONARY OF THE DATA

2]:

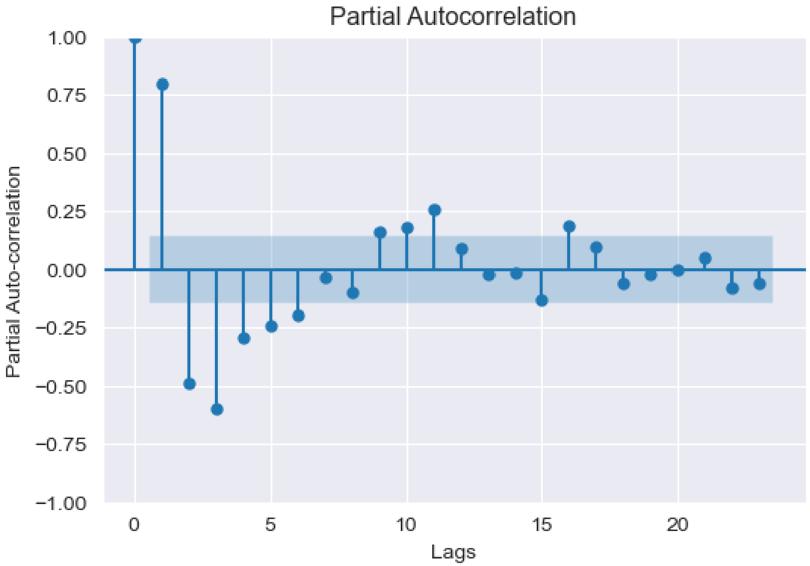
```
# Perform ADF test
adf_result = adfuller(data['temp'])

print('ADF Statistic: %.2f' % adf_result[0])
print('ADF p-value: %.4f' % adf_result[1])
```

```
ADF Statistic: -2.15
ADF p-value: 0.2262
```

Autocorrelation and partial Autocorrelation

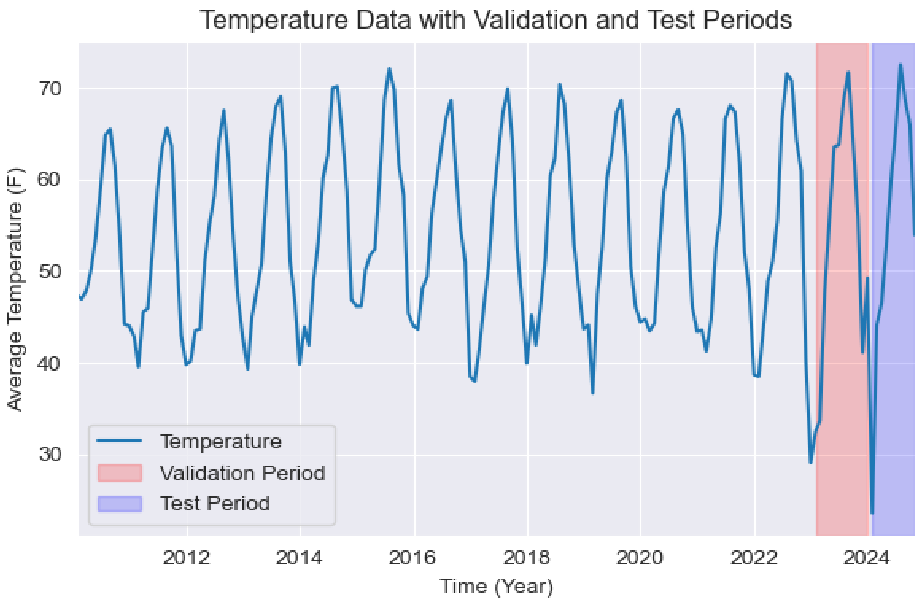


From the graph, it can be said that at lag = 1 the autocorrelation is significant, and the shaded area represents the upper and lower bounds for critical value where the null hypothesis cannot be rejected



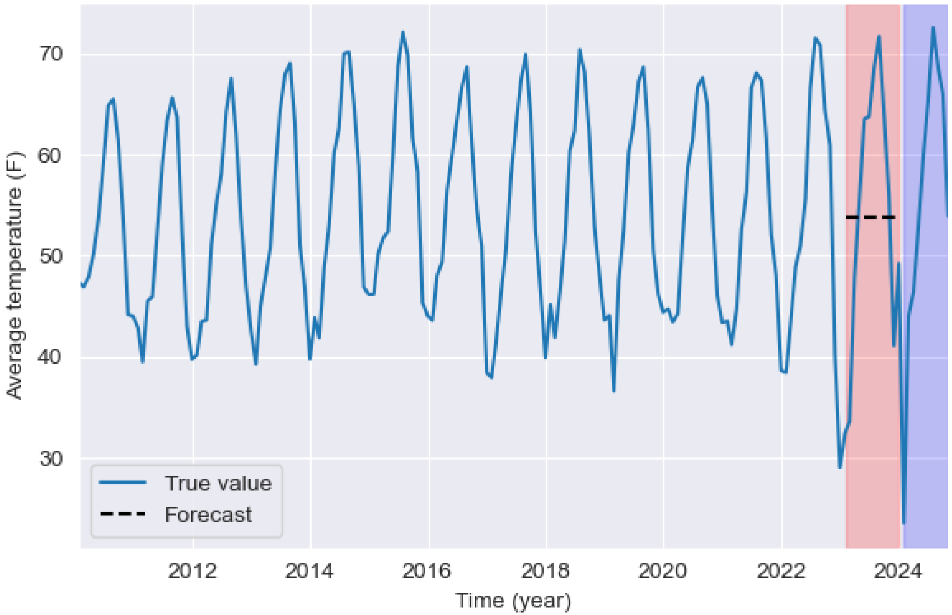
The shaded area represent the upper and lower bounds for critical values where null hypothesis cannot be rejected and it can be said that null hypothesis can be rejected only for lag = 1.

Data Split

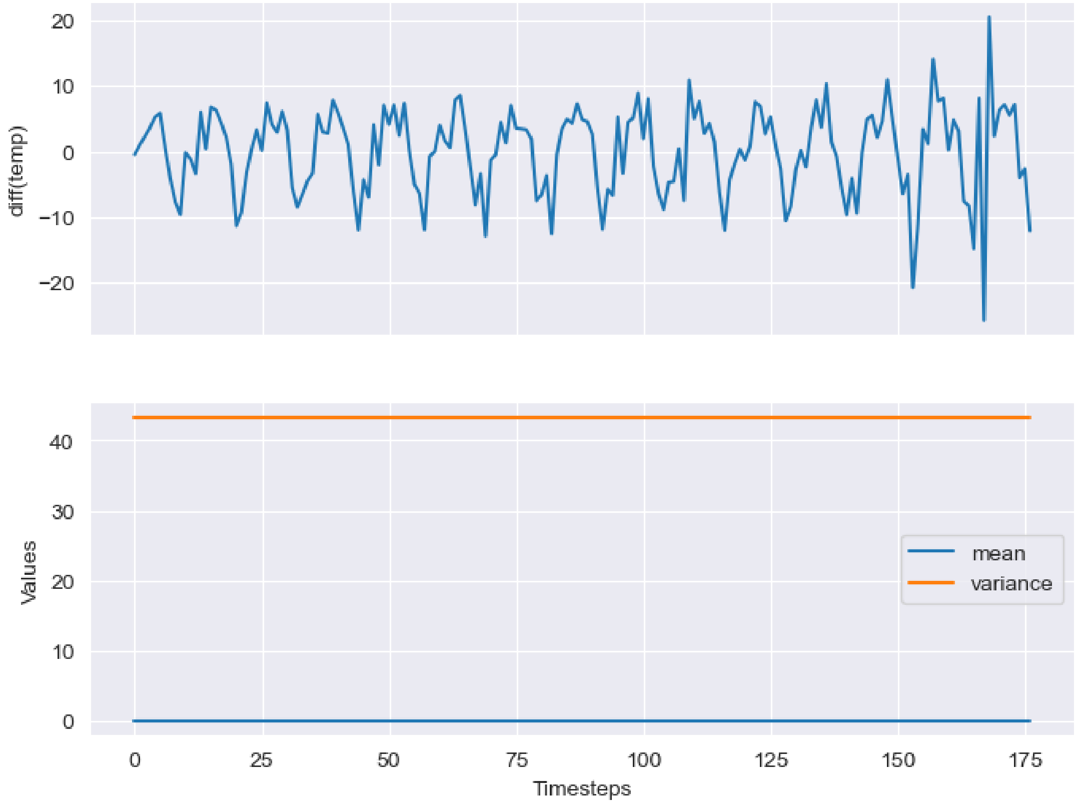


The temperature data with validation and test periods is plotted in the graph. The validation period lies in the range of year 2023 where the temperature has reached the maximum of 70 degree centigrade and test data is from 2020 to rest.

Parameter Evaluation



MAE = 10.95 (degrees Fahrenheit)
MAPE = 23.27 %
RMSE = 12.69 (degrees Fahrenheit)
MSE = 160.95 (degrees Fahrenheit squared)
R2 = -0.00

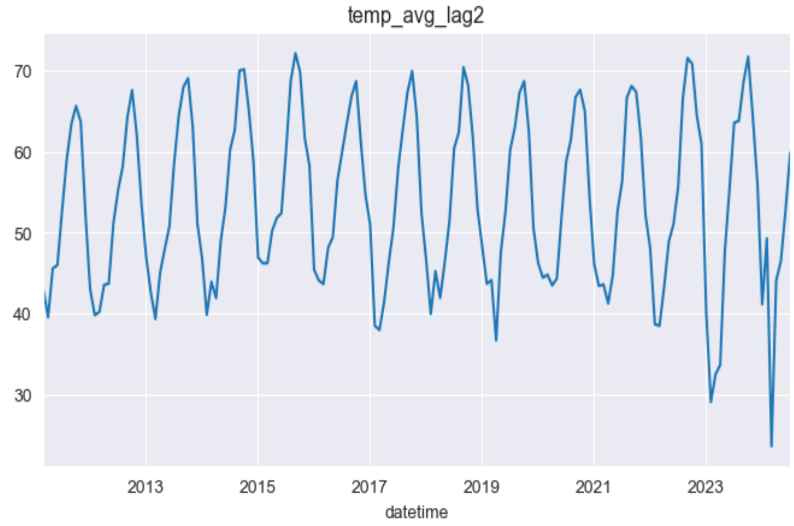


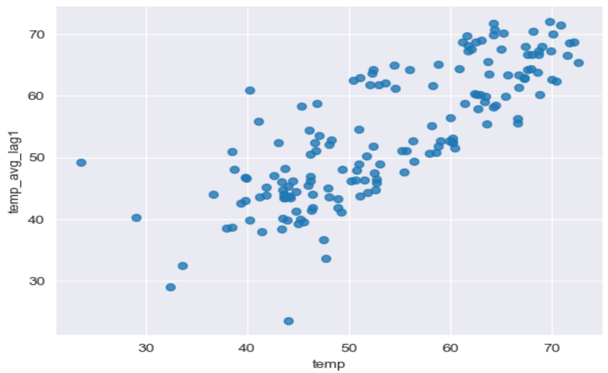
ADF Statistic: -5.93
ADF p-value: 0.000

Model building

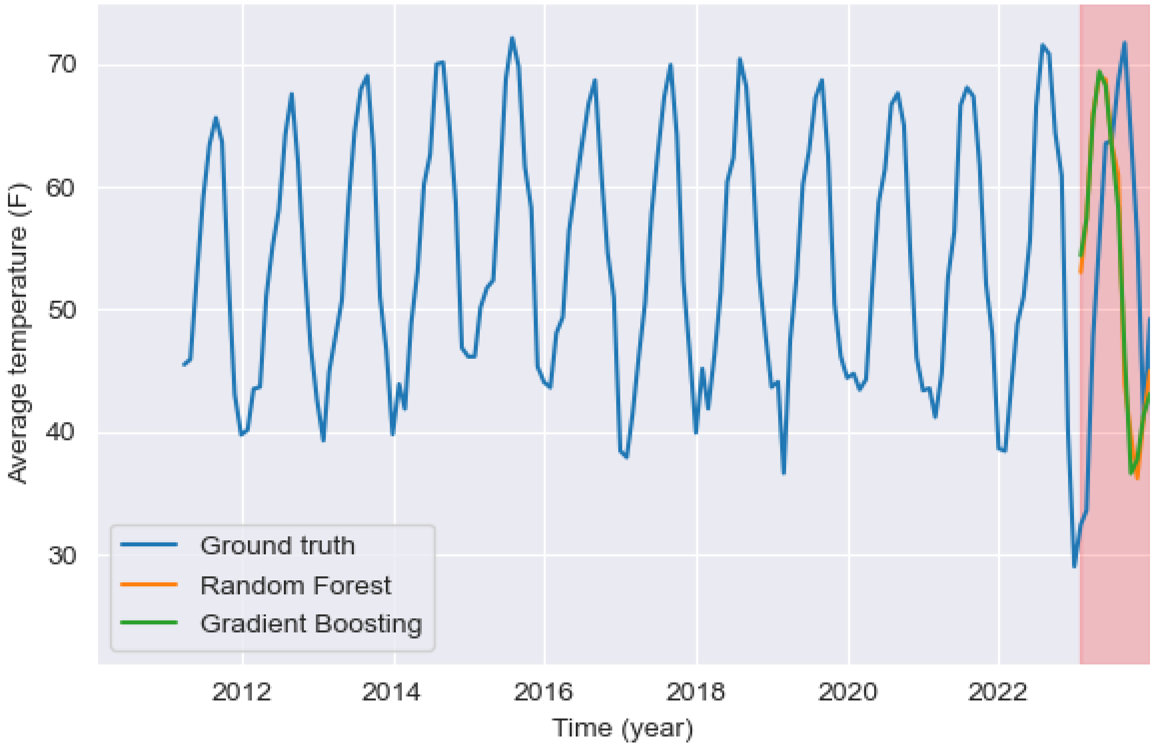
```
SARIMAX Results
==============================================================================
Dep. Variable:          temp              No. Observations:        178
Model:            ARIMA(2, 0, 1)          Log Likelihood        -510.925
Date:           Sun, 01 Dec 2024          AIC                   1031.851
Time:                  04:10:31           BIC                   1047.759
Sample:              01-31-2010           HQIC                  1038.302
                   - 10-31-2024
Covariance Type:            opg
==============================================================================
            coef          std err          z          P>|z|       [0.025      0.975]
------------------------------------------------------------------------------
const      53.7964         0.345        155.952      0.000       53.120      54.472
ar.L1       1.6534         0.035         46.855      0.000        1.584       1.723
ar.L2      -0.9083         0.031        -28.956      0.000       -0.970      -0.847
ma.L1      -0.7662         0.062        -12.440      0.000       -0.887      -0.646
sigma2     17.9104         1.238         14.462      0.000       15.483      20.338
===================================================================================
Ljung-Box (L1) (Q):              18.30   Jarque-Bera (JB):          339.67
Prob(Q):                          0.00   Prob(JB):                    0.00
Heteroskedasticity (H):           3.11   Skew:                       -0.67
Prob(H) (two-sided):              0.00   Kurtosis:                    9.63
===================================================================================
```

Model building



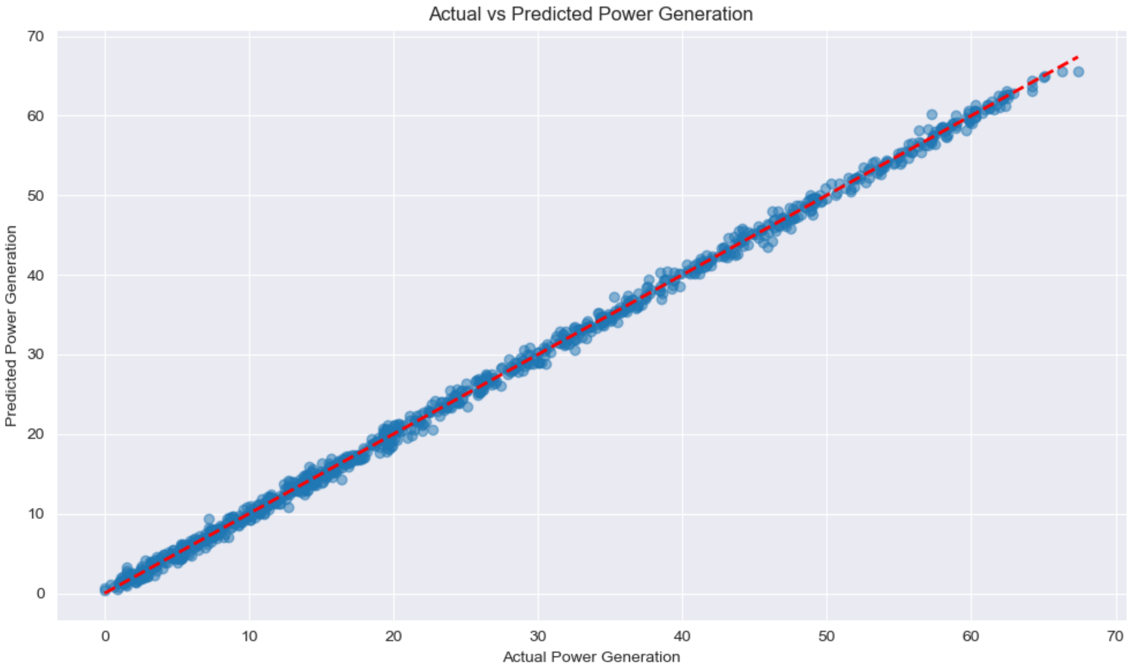


datetime	temp	temp_avg_lag1	temp_avg_lag2	temp_avg_lag3	temp_avg_lag4	temp_avg_lag5	temp_avg_lag6	temp_avg_lag7
2011-03-31	45.529032	39.521429	42.874194	44.019355	44.196667	53.796774	61.446667	65.506452
2011-04-30	45.956667	45.529032	39.521429	42.874194	44.019355	44.196667	53.796774	61.446667
2011-05-31	52.722581	45.956667	45.529032	39.521429	42.874194	44.019355	44.196667	53.796774
2011-06-30	59.023333	52.722581	45.956667	45.529032	39.521429	42.874194	44.019355	44.196667
2011-07-31	63.419355	59.023333	52.722581	45.956667	45.529032	39.521429	42.874194	44.019355
...	...	...	...	...	...	...	...	...
2024-03-31	46.437500	44.066667	23.566667	49.250000	41.125000	55.925000	64.193750	71.722222
2024-04-30	52.727273	46.437500	44.066667	23.566667	49.250000	41.125000	55.925000	64.193750
2024-05-31	59.853846	52.727273	46.437500	44.066667	23.566667	49.250000	41.125000	55.925000
2024-06-30	65.381818	59.853846	52.727273	46.437500	44.066667	23.566667	49.250000	41.125000



Random Forest
MAE = 3.85 (degrees Celsius)
MAPE = 9.83 %
RMSE = 5.45 (degrees Celsius)
MSE = 29.66 (degrees Celsius squared)
R2 = 0.83

Gradient Boosting
MAE = 4.71 (degrees Celsius)
MAPE = 11.85 %
RMSE = 6.41 (degrees Celsius)
MSE = 41.13 (degrees Celsius squared)
R2 = 0.76



Improved Mean Squared Error (with new features): 0.4702
Improved R-squared Score (with new features): 0.9985
Cross-validated R^2 (with new features): 0.9973 ± 0.0029

Future Work

- Incorporate Additional Features

  - Integrate other relevant environmental variables like: Precipitation Probability, Altitude, Seasonality Adjustments
- Improve the Model and apply the same concept for wind turbine power with windspeed

  - Hybrid Models: Combine Random Forest with models like Gradient Boosting or Neural Networks for better performance. Time-Series Models: Experiment with ARIMA and LSTM
- Expand the Dataset

  - Include data from multiple geographic locations to test the model's robustness across different climates and conditions.

Thank you!