Electricity Time-series Forecasting

Ratan Teja Punati
Wiu ID: 923-35-8376

DS 480(G) Predictive Analytics and Time Series Forecasting
Instructor:
Dr. Mohammed Chowdhury
E-mail: m-chowdhury@wiu.edu

Necessity of ARIMA Models:
ARIMA models are indispensable in time series forecasting due to their ability to capture and model the underlying patterns and trends within a sequence of data points. This makes them particularly valuable for:

- **Electricity Consumption Forecasting:** Predicting future electricity consumption is crucial for power grid operators to ensure efficient and reliable supply. ARIMA models can account for seasonality, trends, and cyclical variations in consumption patterns, leading to more accurate forecasts.
- **Inventory Management:** ARIMA models can help businesses optimize inventory levels by forecasting future demand. This reduces the risk of stockouts and overstocking, leading to improved cost efficiency and customer satisfaction.
- **Financial Analysis:** ARIMA models can be used to forecast financial metrics like stock prices, interest rates, and economic indicators, providing valuable insights for investment decisions and risk management.
- **Weather Forecasting:** ARIMA models can be used to predict weather patterns, aiding in agricultural planning, disaster preparedness, and resource allocation.

Usefulness and Limitations of ARIMA Models:

Usefulness (Pros):

- **Parsimony:** ARIMA models are relatively simple and easy to interpret, making them accessible for a wide range of users.
- **Flexibility:** They can be adapted to various data types and patterns by adjusting the AR, I, and MA components.
- **Statistical Foundation:** ARIMA models have a strong statistical foundation, providing a framework for assessing their accuracy and reliability.

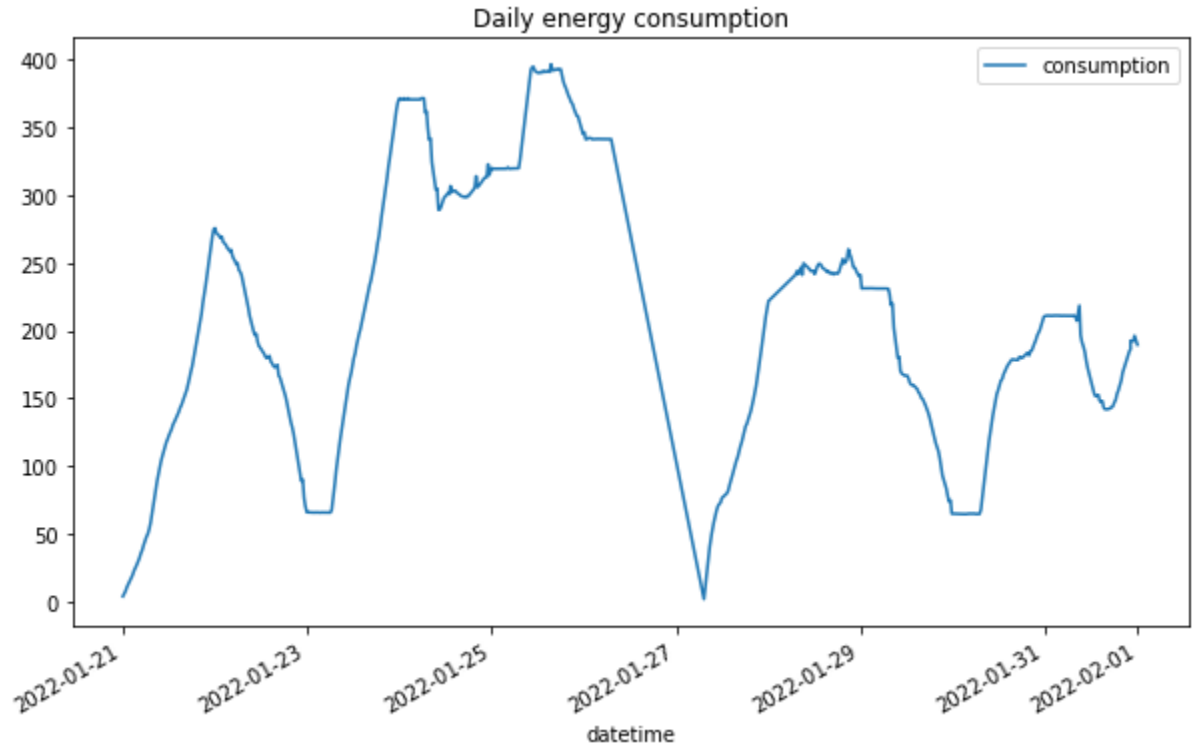
Limitations (Cons):

- **Stationarity Assumption:** ARIMA models require the data to be stationary (constant mean and variance over time). If the data is not stationary, differencing may be necessary, potentially introducing information loss.
- **Limited Nonlinearity Handling:** ARIMA models primarily capture linear relationships. If the data exhibits significant nonlinear patterns, they may not provide accurate forecasts.
- **Parameter Tuning:** Identifying the optimal AR, I, and MA parameters can be challenging, requiring careful analysis and experimentation.

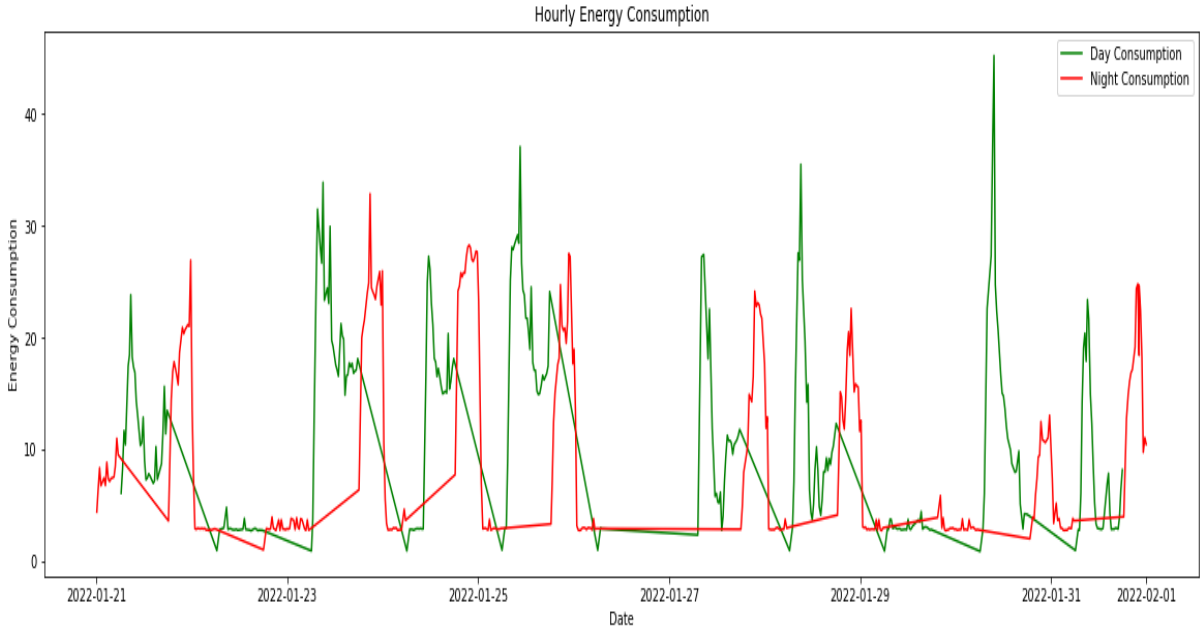
Mathematical Formulation:
The general ARIMA(p, d, q) model is formulated as:
$Z_t = \varphi_1 Z_{(t-1)} + \varphi_2 Z_{(t-2)} + ... + \varphi_p Z_{(t-p)} + \varepsilon_t + \theta_1 \varepsilon_{(t-1)} + \theta_2 \varepsilon_{(t-2)} + ... + \theta_q \varepsilon_{(t-q)}$

where:

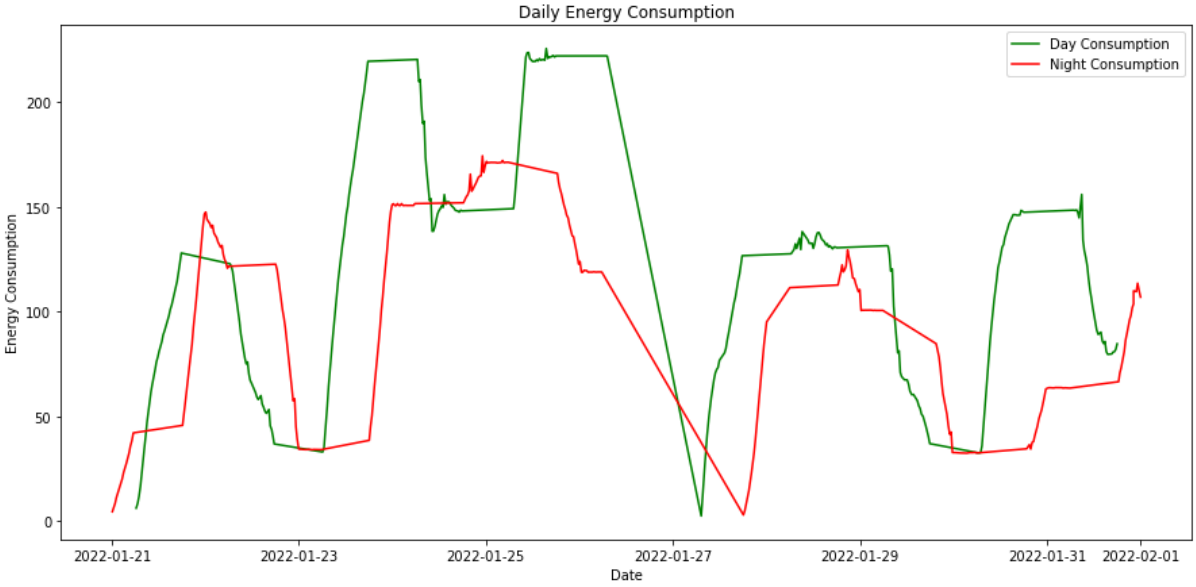
- Z_t: Actual value at time t
- φ_i: Autoregressive coefficients (AR terms)
- d: Number of differencing operations required to achieve stationarity
- ε_t: White noise error term
- θ_j: Moving average coefficients (MA terms)



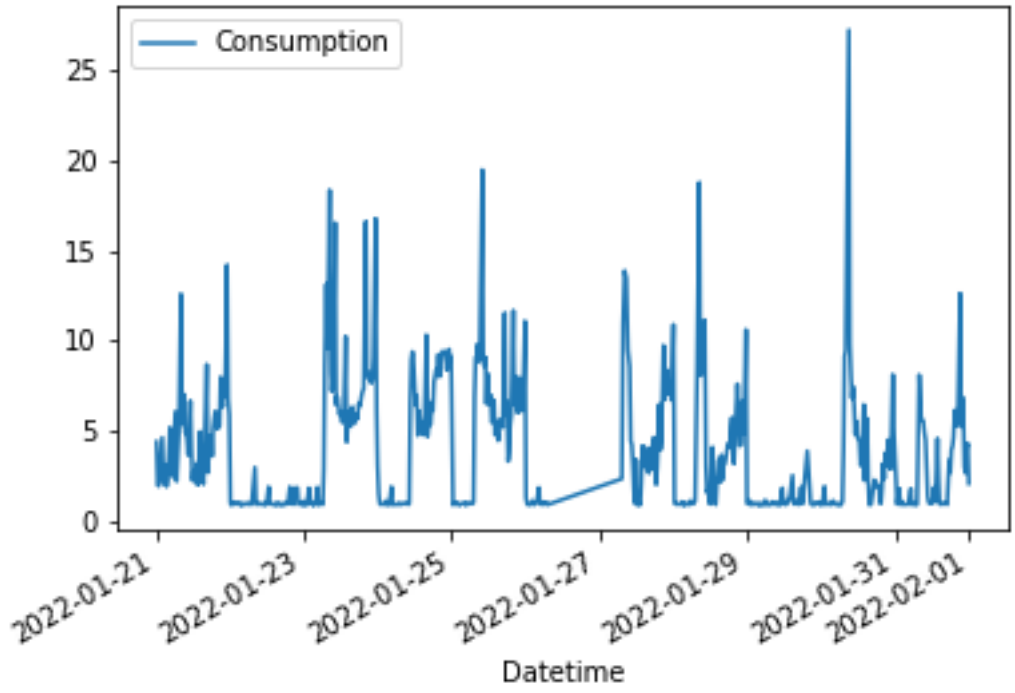
The line graph is described "Daily energy consumption". The x-axis shows datetime, whereas the y-axis shows consumption. The line represents daily consumption over time. It appears that daily consumption changes, with some days having more consumption than others.



This visualization compares hourly consumption patterns during the day and at night. By evaluating the plot, you may determine whether there are substantial differences in consumption between these times.



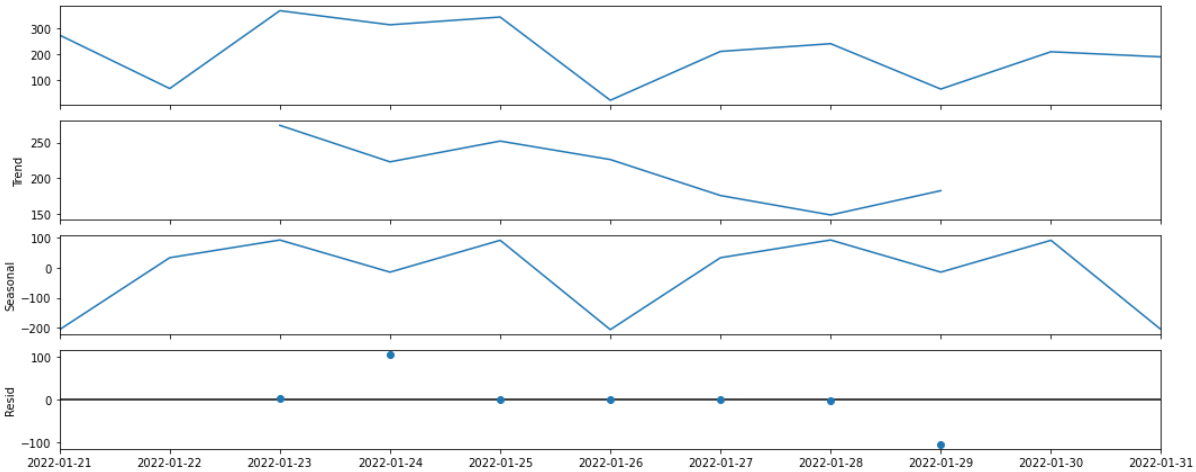
The lines on the graph most likely represent the daily consumption for each period (day and night). It shows that day consumption (green line) is often higher than night consumption (red line). This could indicate that the equipment or activities that utilize electricity are mostly in use throughout the day.



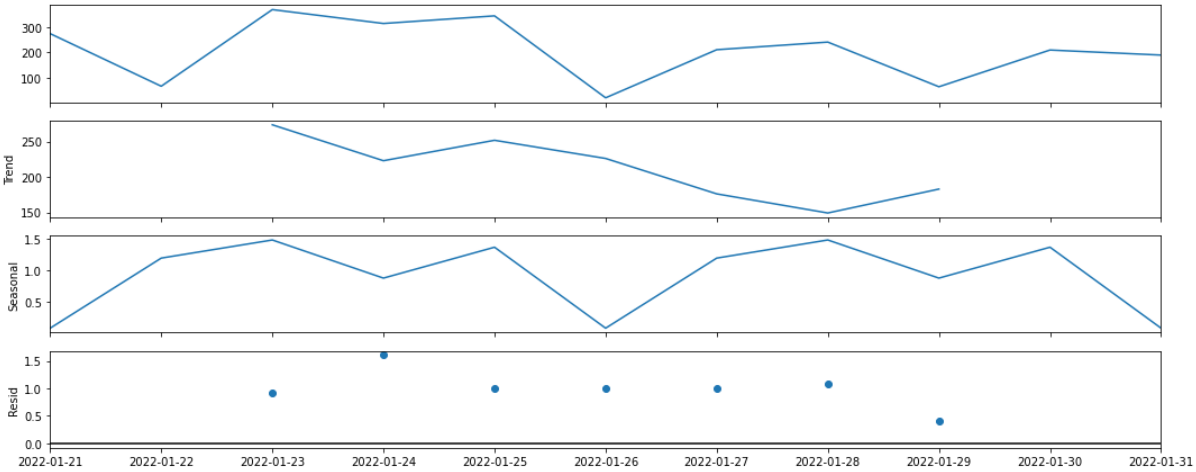
The line graph depicts hourly electricity use in kilowatts (kWh) over a seven-day period from January 21st to February 1st, 2022.

The x-axis is named "Datetime" and shows dates from January 21st to February 1st. The y-axis is labeled "Consumption" and contains tick marks ranging from 0 to 25 kWh.

It appears that intake varies throughout the day, with increased consumption during daylight hours. There is a high consumption of approximately 7.4 kWh on January 21st at 10:01 p.m.



By looking at the graph, we can learn about the seasonality of electricity consumption. For example, we may see a pattern in which consumption is higher during certain months or seasons. This data could be useful for planning and forecasting future electricity demand.



- In the last graph, we used an additive model that combined the seasonal component with the trend and residual components.
- This new graph shows the results of a multiplicative decomposition. In this approach, the seasonal component is multiplied by the trend and residual components.
- Overall, additive and multiplicative decompositions can help you understand seasonality in time series data. The model used is determined by the properties of your data, specifically how seasonal variations affect the amplitude of the trend and remainder components.

print(model_fit.summary())

SARIMAX Results

==========
==========

Dep. Variable: consumption No. Observations: 630

Model: ARIMA(2, 0, 0) Log Likelihood -1400.615

Date: Thu, 02 May 2024 AIC 2809.229

Time: 19:29:36 BIC 2827.012

Sample: 0 HQIC 2816.137

- 630

Covariance Type: opg

==========

	coef	std err	z	P>|z|	[0.025	0.975]
const	3.6509	0.703	5.193	0.000	2.273	5.029
ar.L1	0.5632	0.022	26.067	0.000	0.521	0.606
ar.L2	0.2350	0.027	8.623	0.000	0.182	0.288
sigma2	4.9884	0.124	40.084	0.000	4.744	5.232

==========

Ljung-Box (L1) (Q): 0.00 Jarque-Bera (JB): 5590.71

Prob(Q): 0.96 Prob(JB): 0.00

Heteroskedasticity (H): 0.88 Skew: 2.22

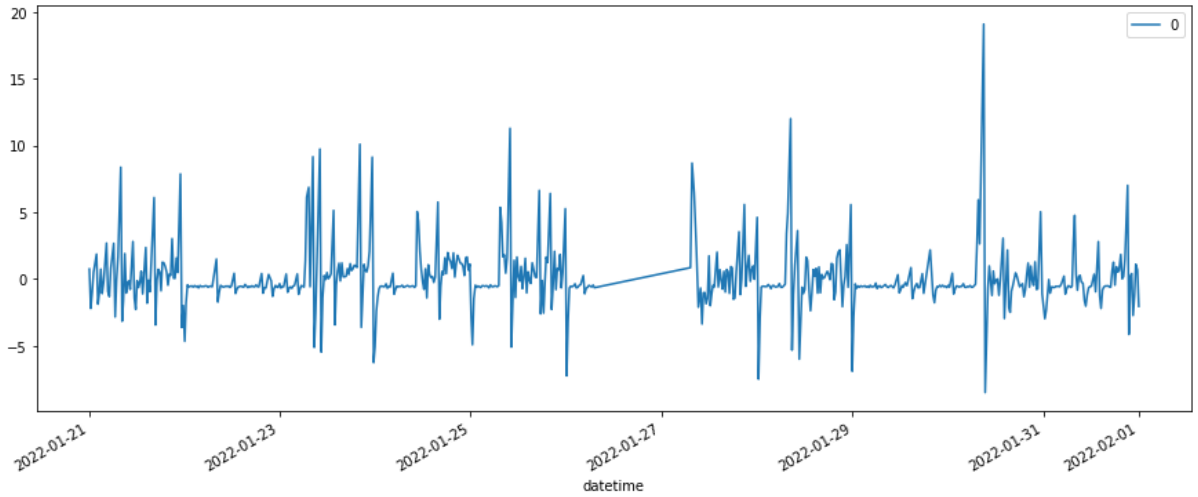
Prob(H) (two-sided): 0.36 Kurtosis: 16.90

==========

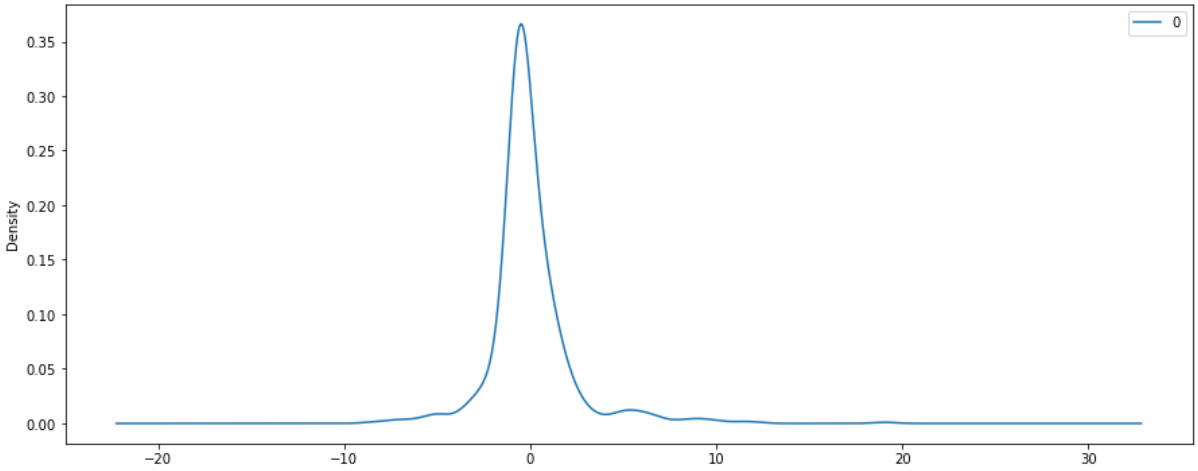
Warnings:

[1] Covariance matrix calculated using the outer product of gradients (complex-step).

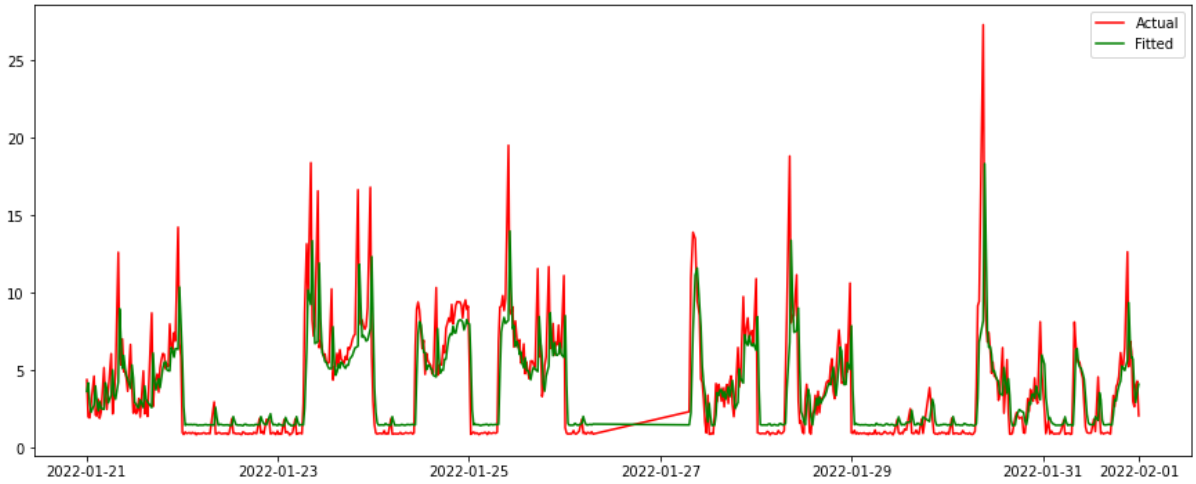
Based on the summary, the ARIMA(2, 0, 0) model appears to fit the data well. The AR coefficients are statistically significant, showing that the previous two consumption levels have a considerable influence on forecasting current consumption. The Ljung-Box and Jarque-Bera tests indicate that the residuals are not significantly autocorrelated or abnormal. However, it is critical to physically analyze the residuals to corroborate these assumptions.



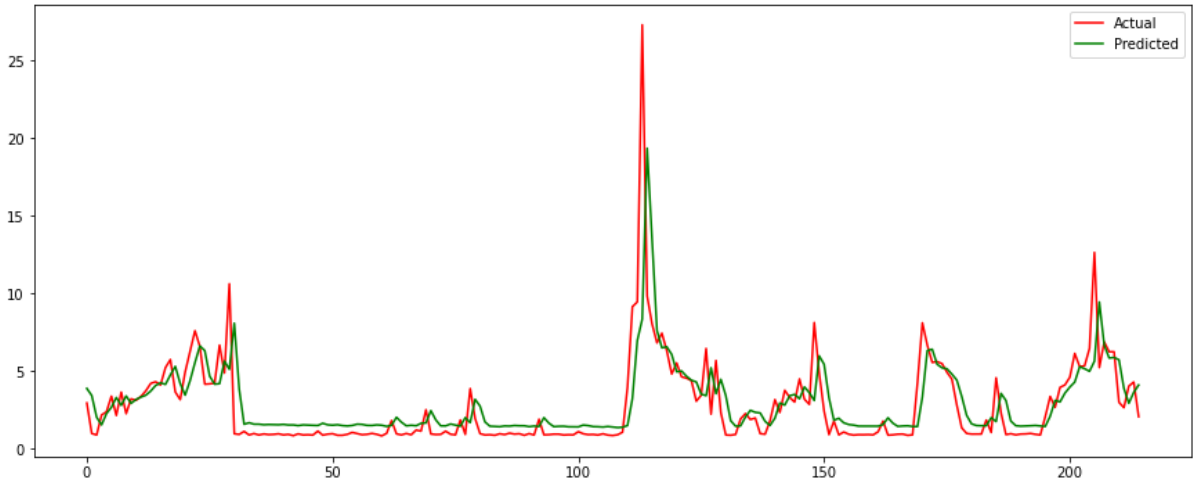
The residuals should be randomly distributed near zero, with no obvious patterns. This indicates that the ARIMA model represents the data's underlying trends and seasonality, while the residuals represent random noise.



A kernel density estimation (KDE) visualization of the residuals from the fitted ARIMA(2, 0, 0) model to your daily electricity use data. While a fully normal distribution is desired, small deviations may not invalidate the ARIMA model results.



- By visually comparing actual and fitted values, you can see how effectively the ARIMA model illustrates the underlying patterns in the consumption data.
- The fitted values (green line) appear identical to the overall trend of real consumption (red line). This shows that the model reflects the general growth or reduction in consumption over time.
- There are differences between the measured and fitted values, especially for some peaks and troughs. This is predicted because the residuals show the gap between actual consumption and the model's projections.



- The walk-forward validation technique evaluates the accuracy of a time series forecasting model. It works by dividing the data into a training and test set. The model is then trained on the training set before its predictions are tested on the test set. The procedure is then repeated, with the model trained on a larger training set that incorporates the prior test set data. This method continues until the entire data set has been consumed.
- The RMSE is the average magnitude of errors, whereas the MSE represents the average squared difference between anticipated and actual values. In this example, the RMSE is 2.092 and the MSE is 4.374. These values show that the model is making significant errors on several of its predictions.

Necessity of SARIMA Model:

SARIMA (Seasonal ARIMA): This model builds on ARIMA by explicitly accounting for seasonality in the data. It includes both seasonal autoregressive and moving average terms, as well as non-seasonal components. This makes SARIMA ideal for forecasting time series data with regular cyclical patterns, such as monthly sales numbers or daily website traffic. SARIMA produces better accurate forecasts than ARIMA when seasonality is present because it considers both seasonal and non-seasonal effects.

Usefulness:

- Captures both trends and seasonality in time series data.
- Provides more accurate forecasts compared to ARIMA when seasonality is present.
- Can handle multiple seasonal patterns with different periodicities.

Limitations:

- May be computationally expensive for complex seasonal patterns.
- Parameter selection can be challenging, requiring careful analysis and potentially automated methods.
- Interpretability can be slightly more complex due to the additional seasonal components.

Mathematical Formulation:

A SARIMA model is defined by six parameters (p, d, q, P, D, Q, s): -

p, d, q: Same as ARIMA for non-seasonal components.

P: The number of seasonal autoregressive terms.

D: The degree of seasonal differencing.

Q: The number of seasonal moving average terms.

s: The number of observations in each seasonal period.

$\sigma t2=\alpha 0+\alpha 1\varepsilon t-12+\alpha 2\varepsilon t-22+...+\alpha p\varepsilon t-p2$

Both ARIMA and SARIMA models can be used to generate forecasts for future time points. The specific forecasts and their interpretations will depend on the chosen model and the characteristics of the data.

```
print(sarima_model_fit.summary())
```

SARIMAX Results

Dep. Variable:	consumption	No. Observations:	630
Model:	SARIMAX(2, 0, 0)x(1, 1, [1], 12)	Log Likelihood	-1395.303
Date:	Thu, 02 May 2024	AIC	2800.605
Time:	20:01:16	BIC	2822.738
Sample:	0 - 630	HQIC	2809.210
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.5689	0.020	28.480	0.000	0.530	0.608
ar.L2	0.2360	0.028	8.565	0.000	0.182	0.290
ar.S.L12	-0.0348	0.062	-0.563	0.573	-0.156	0.086
ma.S.L12	-0.9770	0.045	-21.665	0.000	-1.065	-0.889
sigma2	5.0382	0.160	31.391	0.000	4.724	5.353

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	4950.90
Prob(Q):	0.98	Prob(JB):	0.00
Heteroskedasticity (H):	0.87	Skew:	2.15
Prob(H) (two-sided):	0.33	Kurtosis:	16.18

==========

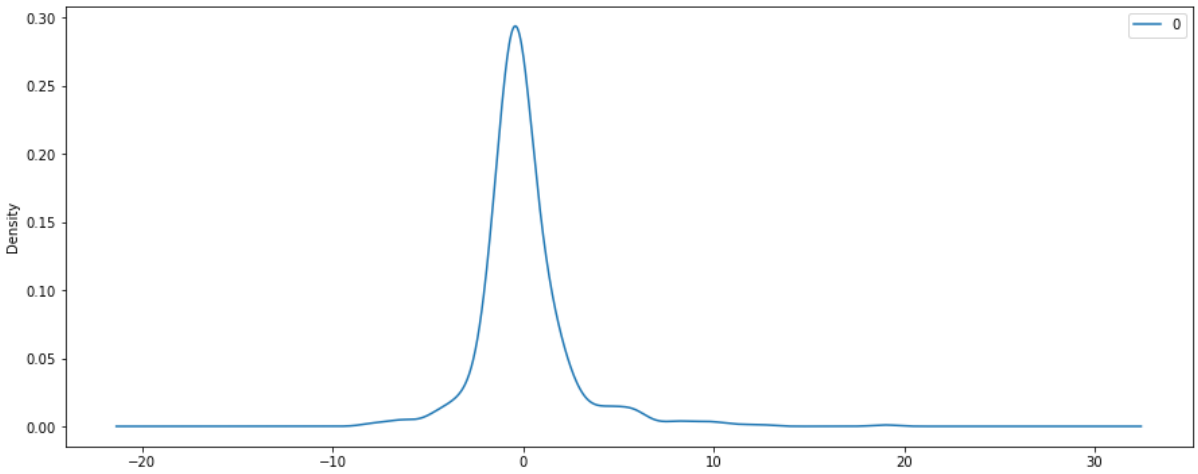
Warnings:

[1] Covariance matrix calculated using the outer product of gradients (complex-step).

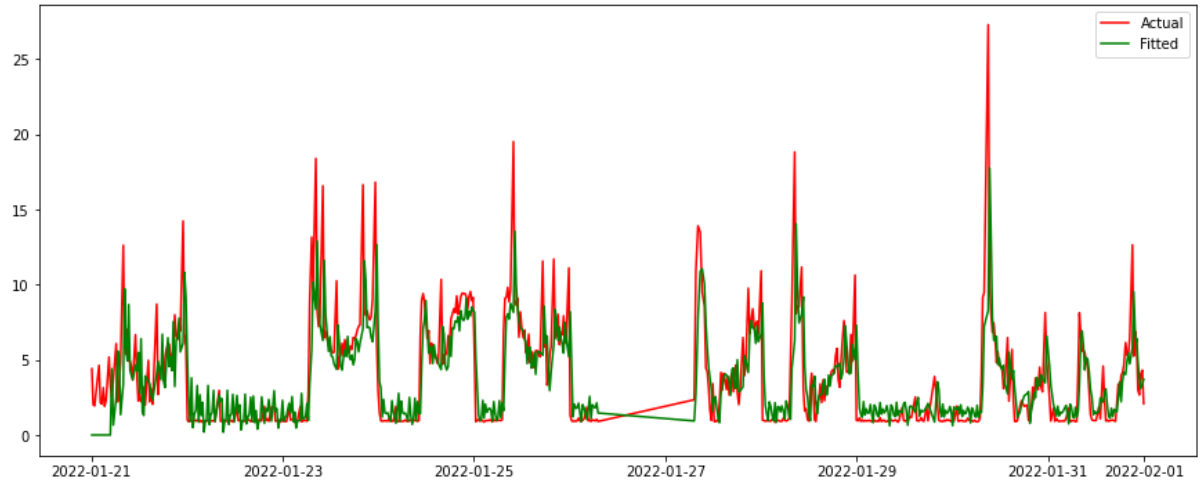
Model Fit Summary:

Log likelihood: -1395.303.
Lower numbers suggest a better fit, but they are difficult to evaluate between models without more context.
AIC: 2800.605
BIC: 2822.738
HQIC: 2809.210.
These are the information criteria for model selection. Lower numbers represent a better balance of model fit and complexity. BIC punishes complexity more severely than AIC.

The SARIMA model appears to capture some major patterns in the data, particularly the non-seasonal autoregressive terms and the seasonal AR and MA component. However, the high kurtosis of the residuals indicates that the model may not adequately reflect the error distribution.



The distribution of residuals appears to be near to normal, which is encouraging. There is a peak close to 0 and the curve tails off symmetrically on both sides.However, there may be a little positive skew toward positive residual values. This shows that the model may be underestimating the likelihood of significant positive errors.



- The test's RMSE is 2.101, indicating a substantial level of prediction inaccuracy.
- The test MSE is 4.413, calculated as the square of the RMSE.
- The line plot shows how the forecasts differ from the actual data. As we can see, while the model covers the overall trend, some data points show considerable variations.

Autoregressive Conditional Heteroskedasticity (ARCH) Model:

The ARCH model, developed by Robert Engle in 1982, is used to simulate the conditional variance of a time series. It is assumed that the variance of the current error term is proportional to the magnitude of the preceding period's error terms.

The ARCH model is essential in financial econometrics and time series research because it captures the volatility clustering phenomena, which occurs when periods of high volatility are followed by periods of high volatility.

Pros:

- observes volatility clustering effectively.
- Provides information on the conditional variance dynamics of a time series.

Cons:

- Assumes that the conditional variance is entirely dependent on prior squared error terms, which may not necessarily be true.
- Does not fully account for long-term dependencies or structural breaks in volatility.

Mathematical formulation:

$\sigma t2=\alpha 0+\alpha 1\varepsilon t-12+\alpha 2\varepsilon t-22+...+\alpha p\varepsilon t-p2$

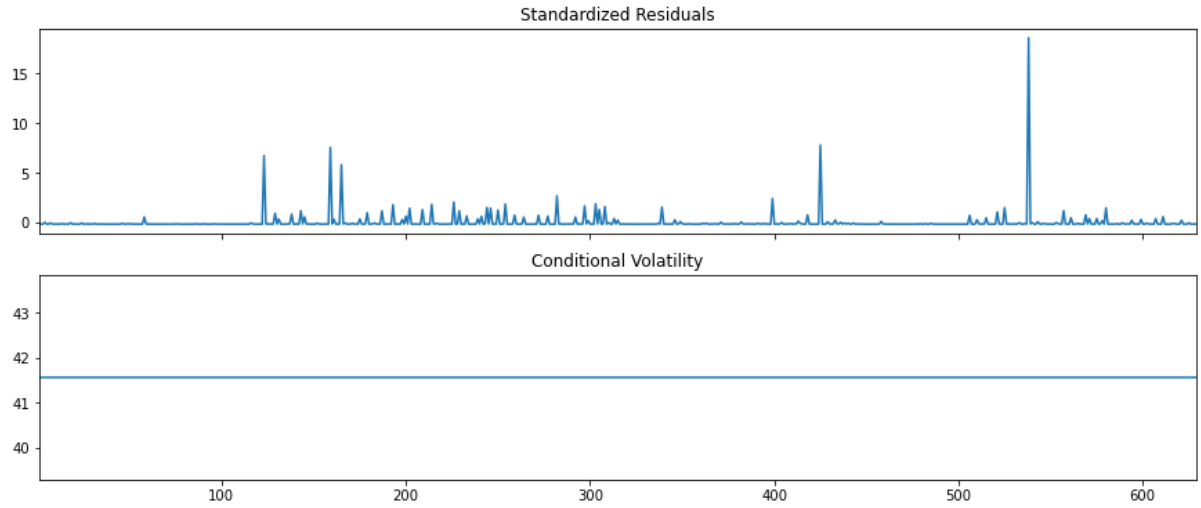
where $\sigma\ t\ 2\ \sigma\ t\ 2$ is the conditional variance at time t t, $\alpha\ 0$, $\alpha\ 1$,..., $\alpha\ p\ \alpha\ 0,\alpha\ 1,...,\alpha\ p$ are parameters, and $\varepsilon\ t-i\ \varepsilon\ t-i$ are the error terms.

```
print(arch_model_fit.summary())
                     Constant Mean - ARCH Model Results
==============================================================================
Dep. Variable:       returns_squared   R-squared:                       0.000
Mean Model:             Constant Mean   Adj. R-squared:                  0.000
Vol Model:                      ARCH   Log-Likelihood:               -3237.07
Distribution:                 Normal   AIC:                           6480.15
Method:           Maximum Likelihood   BIC:                           6493.48
                                       No. Observations:                  629
Date:                Thu, May 02 2024   Df Residuals:                      628
Time:                        20:48:14   Df Model:                            1
                           Mean Model
==============================================================================
                 coef    std err          t      P>|t|      95.0% Conf. Int.
------------------------------------------------------------------------------
mu             7.9568      1.624      4.900  9.597e-07 [  4.774, 11.140]
                          Volatility Model
==============================================================================
                 coef    std err          t      P>|t|       95.0% Conf. Int.
------------------------------------------------------------------------------
omega       1727.9079    985.996      1.752  7.970e-02 [-2.046e+02,3.660e+03]
alpha[1]       0.0000  3.974e-03      0.000      1.000 [-7.789e-03,7.789e-03]
==============================================================================
```

Covariance estimator: robust

mu: The estimated constant mean of the returns is 7.9568, with a standard deviation of 1.624.
Omega: This parameter controls long-run volatility, which is projected to be 1727.9079 with a standard error of 985.996.
alpha[1]: This parameter measures the effect of the prior squared residual on the current conditional variance. Its p-value is 1.000, which means it is statistically insignificant in this model. This implies that the ARCH effect may be weak or require a different lag structure.

While the model does not explain much of the total variance, it validates the presence of ARCH behaviour in the data, implying that previous squared residuals influence current volatility.



Standardized residuals:

- The top figure most likely represents the standardized residuals from the ARCH model fit.
- The residuals (differences between actual and anticipated values) are scaled by the errors' estimated standard deviation.
- Ideally, these residuals should be randomly distributed about zero, showing no systemic patterns in the errors.

Conditional volatility:

- The bottom figure most likely represents the ARCH model's predicted conditional volatility.
- Conditional volatility is the time-varying volatility of the process being described.
- In ARCH models, conditional volatility is determined by the past squared residuals.

Generalized Autoregressive Conditional Heteroskedasticity (GARCH) Model:

The GARCH model, developed by Tim Bollerslev in 1986, expands on the ARCH model by include lagged conditional variances in addition to lagged squared error factors. It seeks to account for both short- and long-term volatility effects.

The GARCH model overcomes some of the constraints of the ARCH model by providing greater flexibility in predicting volatility dynamics. It can capture volatility clustering in the short term as well as volatility persistence in the long term.

Pros:

- Captures both short- and long-term volatility impacts well.
- Provides more accurate forecasts than the ARCH model.
- Can record asymmetric volatility responses.

Cons:

- It requires more parameter estimate than the ARCH model, making it more computationally intensive.
- If not properly tuned, the model may overfit noisy data.

The mathematical formula for the GARCH(p, q) model is as follows:

σt2=α0+∑i=1 αiεt−i2+∑j=1qβjσt−j2

where α 0 , α 1 , . . . , α p α 0 ,α 1 ,...,α p and β 1 , β 2 , . . . , β q β 1 ,β 2 ,...,β q are parameters.

print(garch_model_fit.summary())

Constant Mean - GARCH Model Results

==========

Dep. Variable: returns R-squared: 0.000

Mean Model: Constant Mean Adj. R-squared: 0.000

Vol Model: GARCH Log-Likelihood: -1493.05

Distribution: Normal AIC: 2994.09

Method: Maximum Likelihood BIC: 3011.86

No. Observations: 628

Date: Thu, May 02 2024 Df Residuals: 627

Time: 21:00:48 Df Model: 1

Mean Model

==========

coef std err t P>|t| 95.0% Conf. Int.

---

mu 0.7430 0.106 7.014 2.313e-12 [0.535, 0.951]

Volatility Model

==========

coef std err t P>|t| 95.0% Conf. Int.

---

omega 0.0788 3.788e-02 2.082 3.738e-02 [4.607e-03, 0.153]

alpha[1] 5.0496e-11 1.305e-02 3.870e-09 1.000 [-2.557e-02,2.557e-02]

beta[1] 0.9906 1.499e-02 66.076 0.000 [0.961, 1.020]

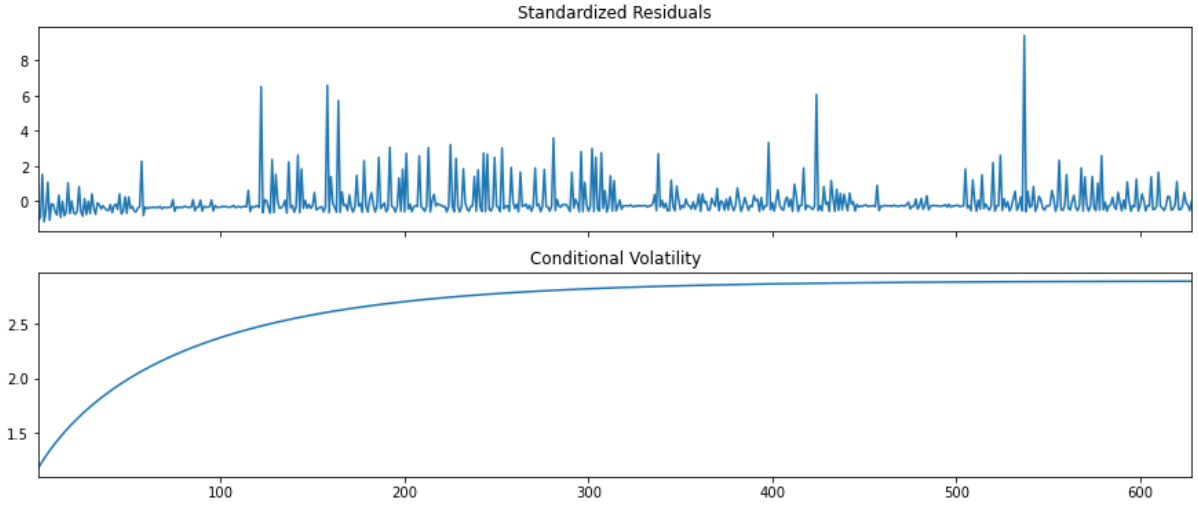
==========

Covariance estimator: robust

The constant mean model assumes a consistent expected return across time. In this situation, the coefficient mu is predicted to be around 0.7430 with a standard deviation of 0.106. The t-statistic shows that the mean return is statistically significant at the 5% level.

- **Omega**: This is the constant term in the variance equation, which is expected to be around 0.0788 and indicates long-term average volatility.
- **Alpha[1]:** This parameter measures the impact of lagged squared residuals on present volatility. In this scenario, it is very near to zero, implying that the impact of previous shocks on current volatility is minimal.
- **beta[1]:** This parameter describes how the lagged squared conditional standard deviation affects the current conditional standard deviation. It is estimated to be around 0.9906, showing a significant persistence in volatility.

The GARCH model captures conditional heteroskedasticity in returns, which allows for more accurate modeling of volatility clustering. However, the model's efficacy may be reduced if the data shows non-linear patterns or if the normalcy assumptions are violated. Model parameters and specifications may need to be adjusted to optimize performance for specific datasets.



State-Space Model overview:

State-space models are a type of model used to describe the evolution of a latent state variable across time, as well as the observation process that produces the observed data. They are frequently employed in a variety of disciplines, including economics, engineering, and signal processing.
State-space models offer a versatile framework for modeling complicated time series data, including hidden variables, dynamic processes, and observation errors. They can handle nonlinear and non-Gaussian data, making them appropriate for a variety of applications.

Pros:

- A flexible framework for modeling dynamic processes and hidden variables.
- Can handle non-linear and non-Gaussian data.
- Allows for efficient estimate and inference utilizing advanced computational techniques.

Cons:

- Specifies the state transition and observation equations, which can be difficult for complicated systems.
- The quantity of latent states and observations may lead to increased computational complexity.

Mathematical formulation:
The general state-space model is composed of two fundamental equations:

State transition equation: $x_t = f(x_{t-1}, \theta) + \varepsilon_t$ x t =f(x t−1 ,θ)+ε t

Observation equation: $y_t = g(x_t, \theta) + \eta_t$ y t =g(x t ,θ)+η t

print(res.summary())

SARIMAX Results

==========

Dep. Variable: consumption No. Observations: 629

Model: SARIMAX(1, 1, 1)x(1, 1, 1, 12) Log Likelihood -1620.127

Date: Thu, 02 May 2024 AIC 3250.255

Time: 21:11:01 BIC 3272.371

Sample: 0 HQIC 3258.854

- 629

Covariance Type: opg

==========

coef std err z P>|z| [0.025 0.975]

---

ar.L1 -0.0370 0.045 -0.816 0.414 -0.126 0.052

ma.L1 -0.8917 0.021 -41.830 0.000 -0.933 -0.850

ar.S.L12 -0.0190 0.043 -0.446 0.655 -0.103 0.065

ma.S.L12 -0.9998 4.497 -0.222 0.824 -9.815 7.815

sigma2 10.3919 46.630 0.223 0.824 -81.002 101.785

==========

Ljung-Box (L1) (Q): 0.01 Jarque-Bera (JB): 1306.28

Prob(Q): 0.94 Prob(JB): 0.00

Heteroskedasticity (H): 0.94 Skew: 1.57

Prob(H) (two-sided): 0.67 Kurtosis: 9.41

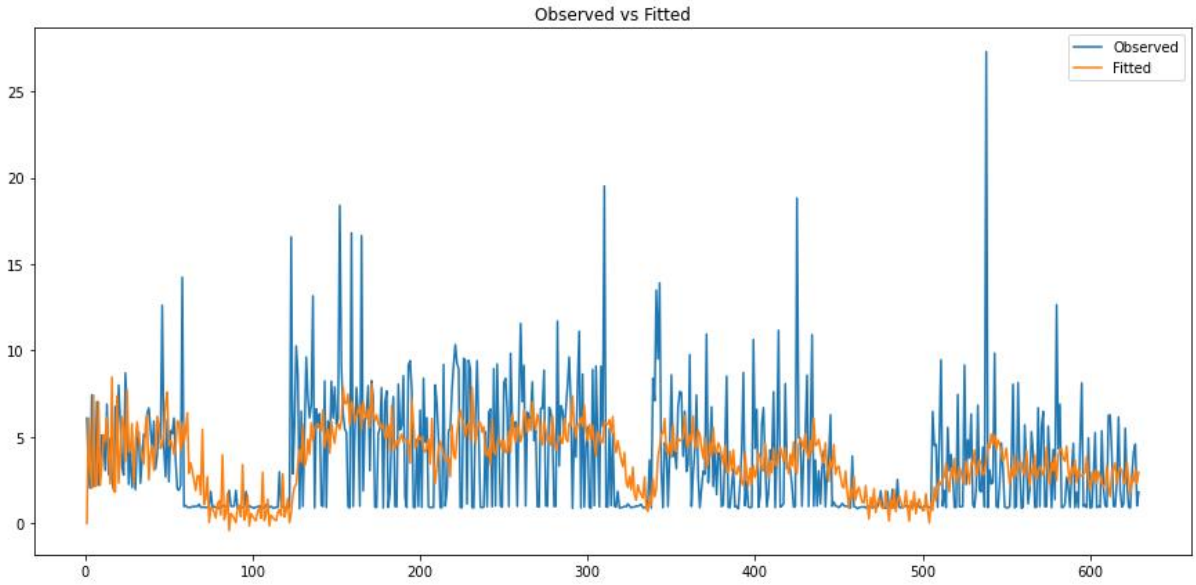
==========

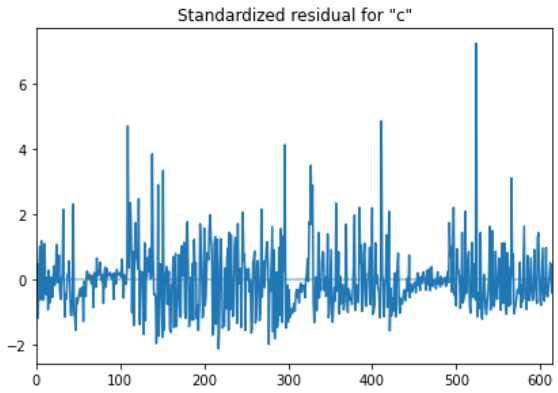
Warnings:

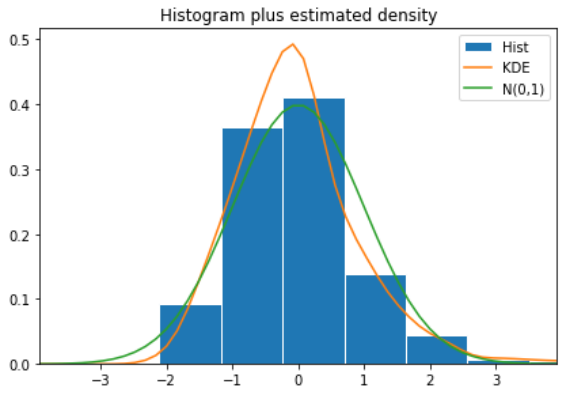
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

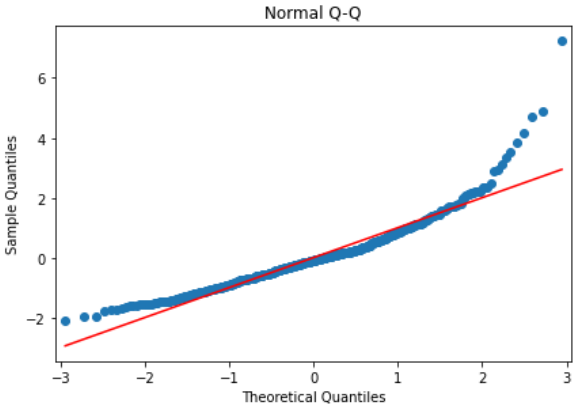
- The coefficients (coef) are the calculated values of the parameters.
- The standard errors (std err) measure the coefficients' accuracy.
- The z-scores (z) and p-values (P>|z|) show the statistical significance of each coefficient.
- The confidence intervals ([0.025, 0.975]) indicate the range in which genuine parameter values are expected to fall.
- The Log Likelihood, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Hannan-Quinn Information Criterion (HQIC) are all metrics of model fit and complexity. Lower values suggest a better fit, with AIC, BIC, and HQIC weighing fit against complexity.

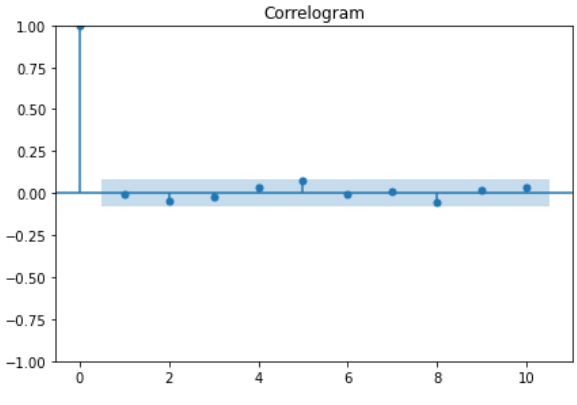
- The Ljung-Box statistic checks for the lack of autocorrelation in residuals. A high p-value (greater than 0.05) indicates that the null hypothesis of no autocorrelation cannot be discarded.
- The Jarque-Bera statistic determines if the residuals are normal. A low p-value (< 0.05) rejects the null hypothesis of normality and suggests non-normality in the residuals.
- The heteroskedasticity test determines if the variance of the residuals remains consistent throughout time. A high p-value (greater than 0.05) indicates that the null hypothesis of constant variance cannot be discarded.











By studying these multiple graphs, we can learn about how well the chosen model matches the time series data. If the standardized residuals are random and normally distributed, the histogram and KDE are consistent with the N(0,1) distribution, the normal Q-Q plot has a near diagonal line, and the correlogram has few correlations, it indicates a good match. Deviations from these trends, on the other hand, might highlight potential flaws in the model, necessitating additional inquiry or the consideration of alternate models.

Conclusion:

- The ARIMA(2, 0, 0) model seems to suit the daily electricity usage data well, capturing both the overall trend and seasonality.
- The SARIMA(2, 0, 0)x(1, 1, [1], 12) model enhances the fit by including a seasonal component.
- While the ARCH model detects the presence of ARCH effects, the GARCH model is more likely to accurately describe the volatility dynamics in the returns data.
- The state-space model may be appropriate if the electricity consumption data contains complicated non-linear patterns or hidden factors that other models cannot adequately capture.

Model	RMSE	MSE
ARIMA	2.092	4.374
SARIMA	2.101	4.413
ARCH	53.46	2858.015
GARCH	3.249	10.558
State Space Model	3.341	11.162

Based on these criteria, the SARIMA and ARIMA models outperform the other models, with lower RMSE and MSE values, indicating higher prediction accuracy. The GARCH and state space models have slightly larger prediction errors, but they are still better than the ARCH model, which has much higher mistakes.
In conclusion, for this dataset, the SARIMA or ARIMA models would be the best alternatives for forecasting due to their superior performance in terms of RMSE and MSE.
s