Power Grid Load Forecasting using Machine Learning Approaches: Bachelor's Project

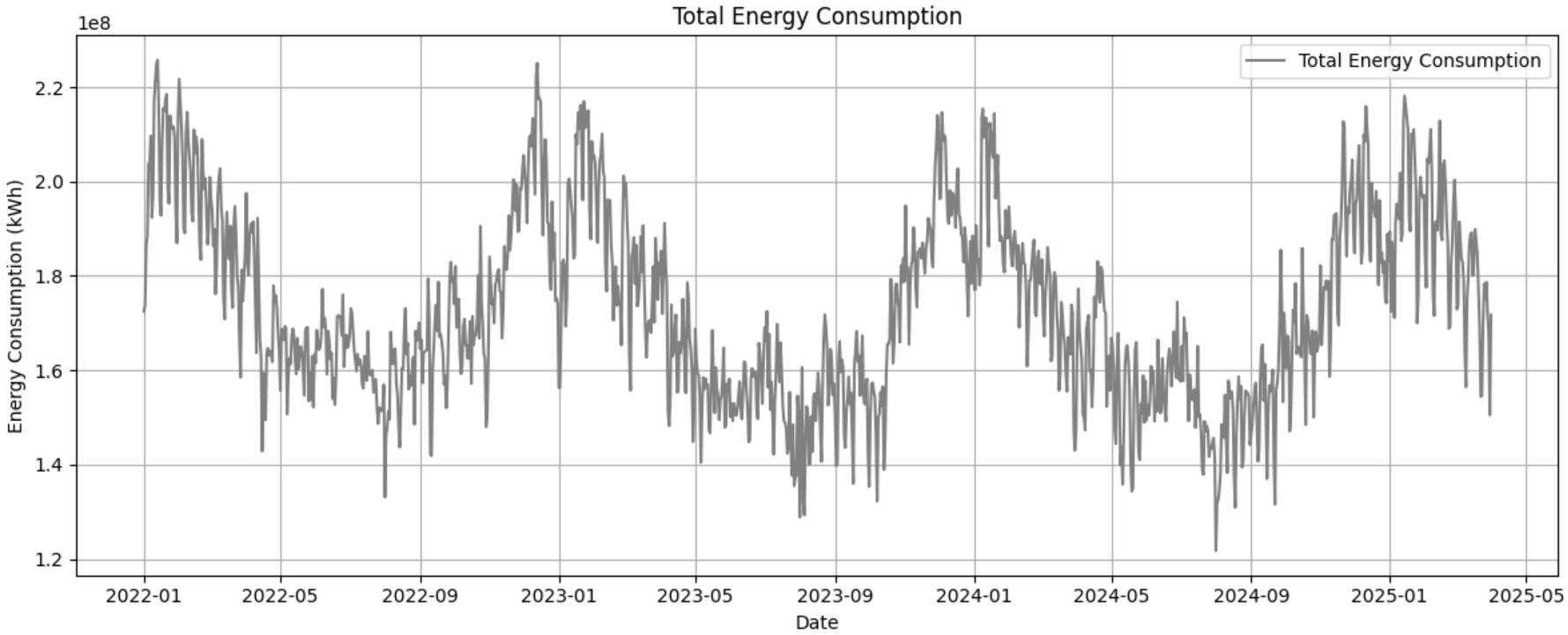
Supervised by Chanel Guillaume, Youssef Saïd, Clément Targe

Outline

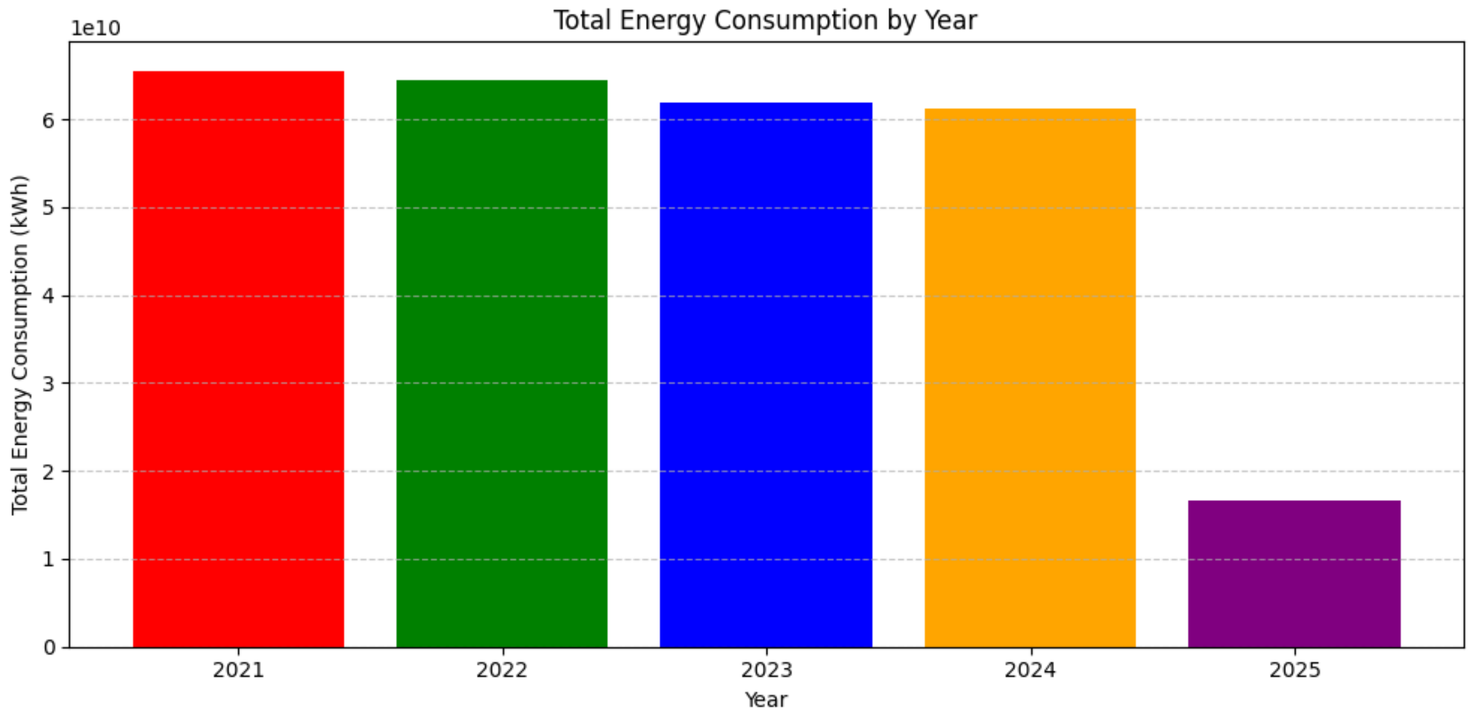
- I. Visualisation
- II. Data Cleaning
- III. Time Series Analysis
- IV. Models
- V. Evaluation
- VI. Challenges & Discussion

I. Visualisation

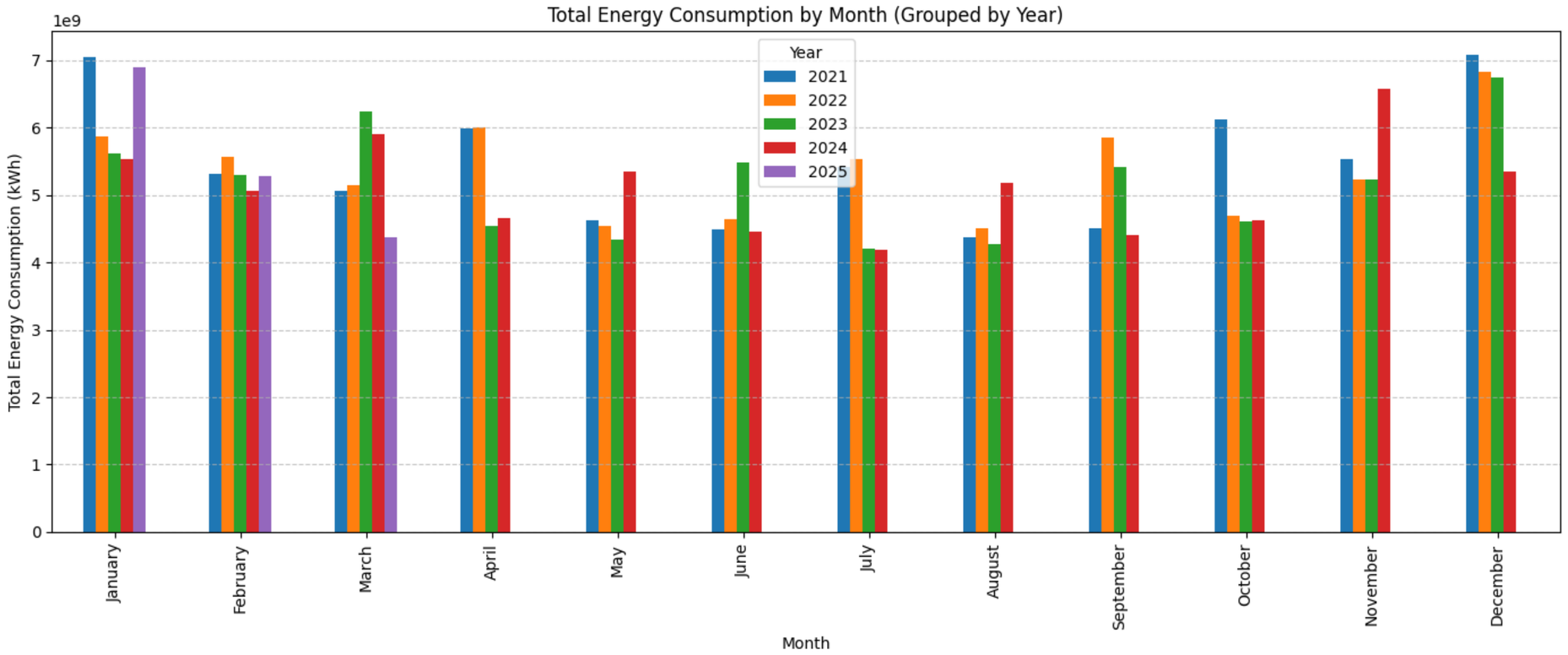
Total Energy Consumption Data: 2022 - 2025



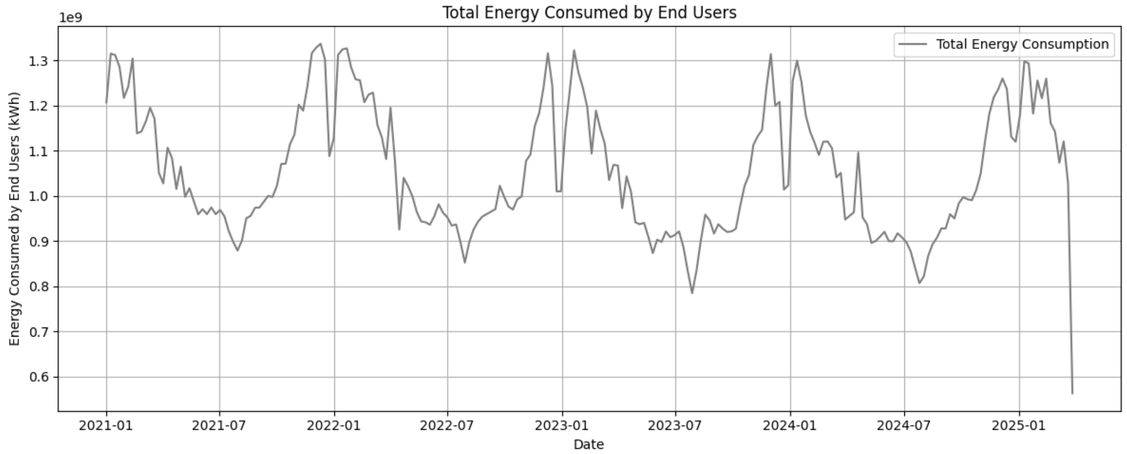
I. Visualisation

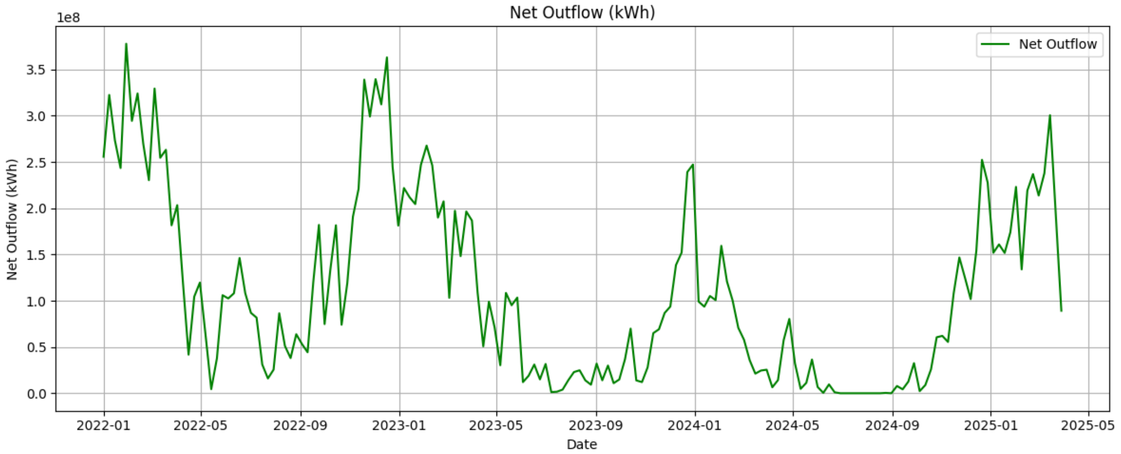


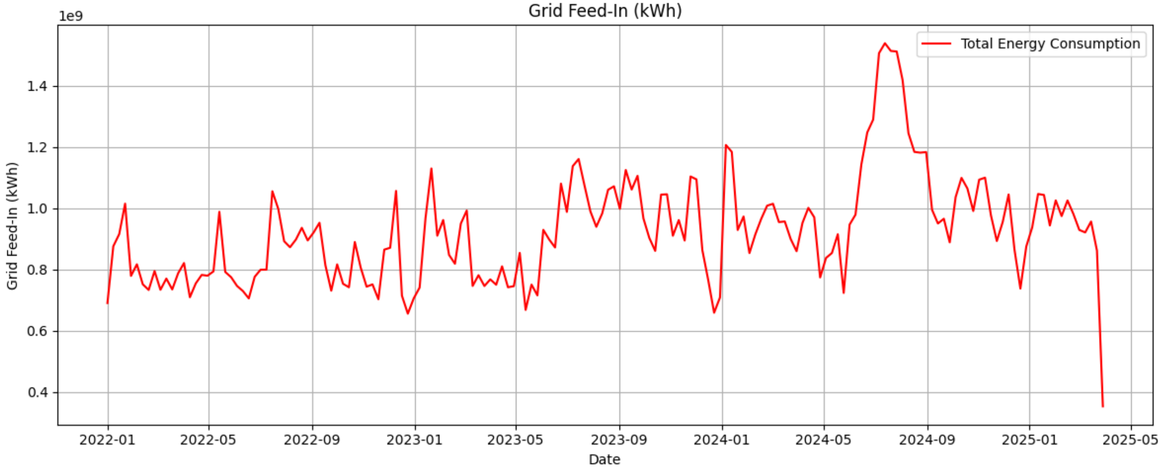
I. Visualisation

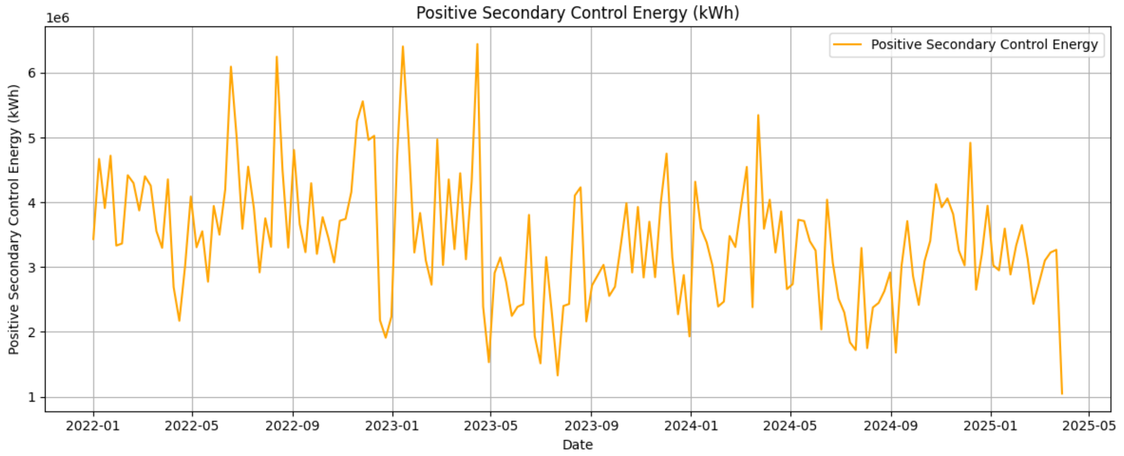


I. Visualisation of related variables



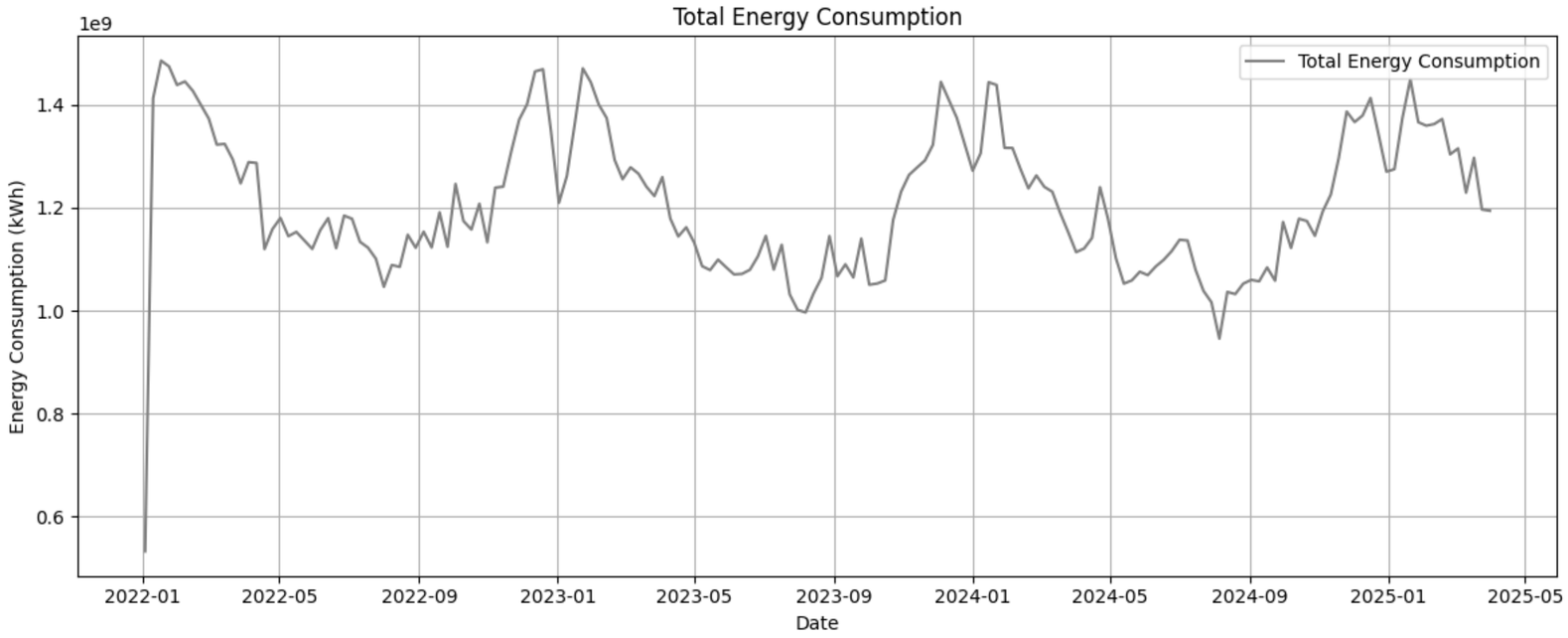






Variable to predict
Weekly sum

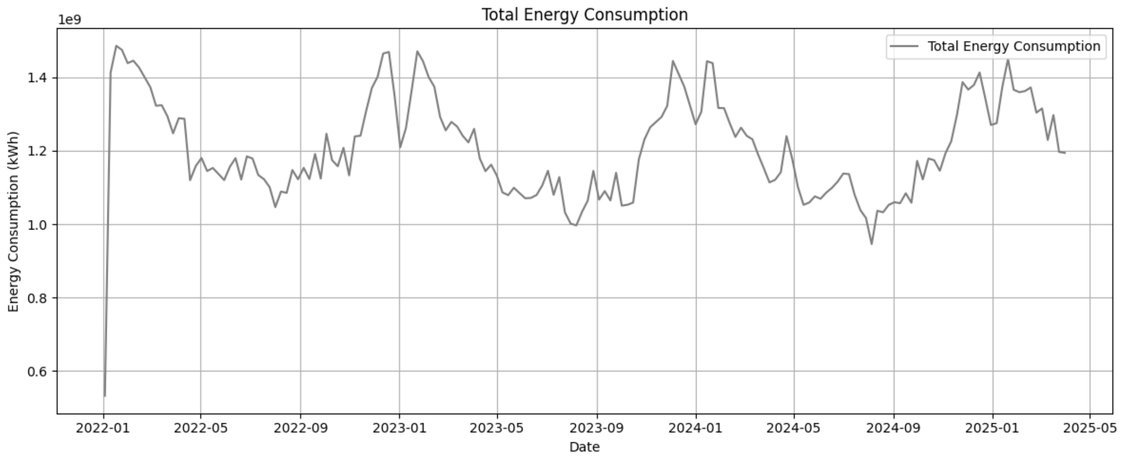
- The weekly predicted sum will be from Monday-Sunday, as is per EU energy regulations

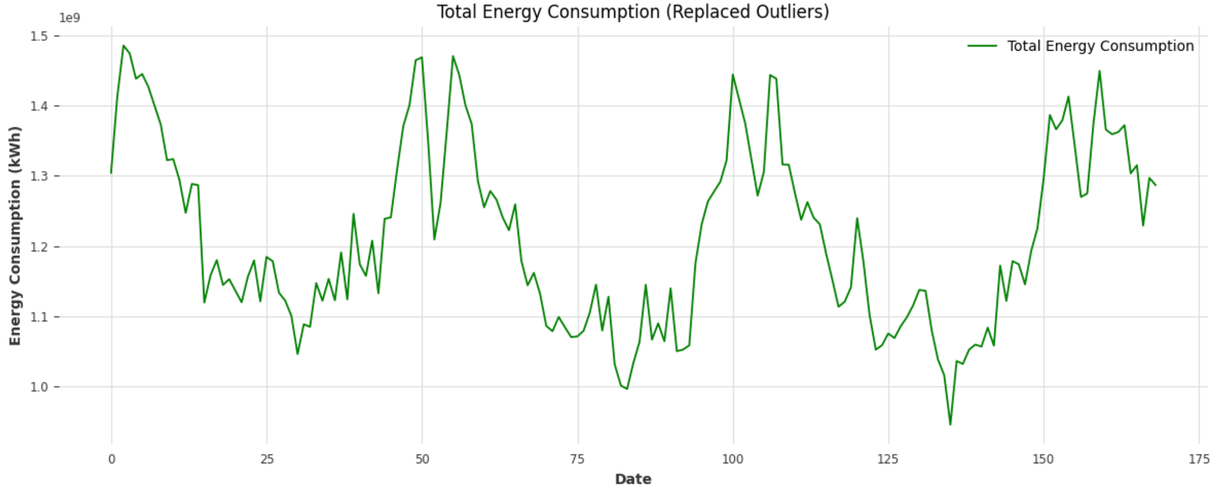


Variable to predict

underreported values (self-caused)

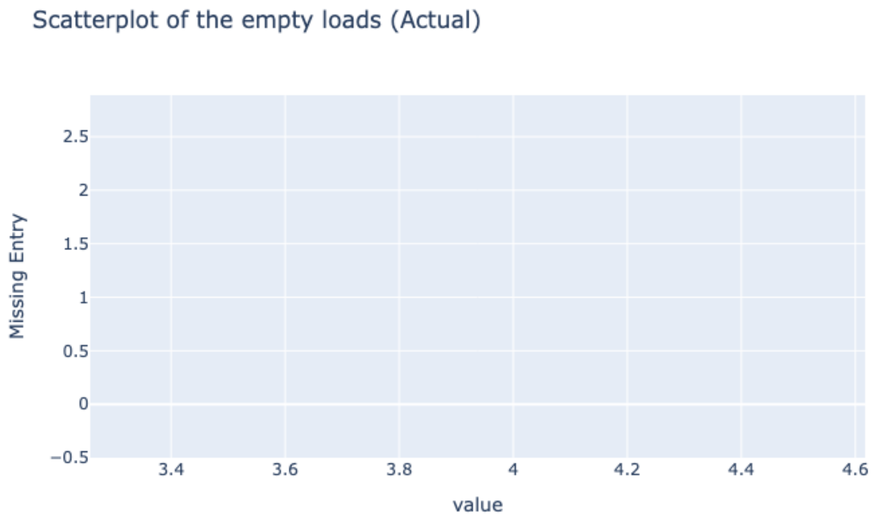
- After summing weekly from Monday, I removed the underreported values (first and last not fully summed values) and replaced them with the rolling mean in that period (w=14)





II. Data Cleaning

II. Data Cleaning



Check min, max, std for anomalies

```
print(df.describe())
```

✓ 0.2s

```
       Total Energy Consumed by End Users (kWh)  \
count                              1.138520e+05
mean                               1.557924e+06
std                                2.965795e+05
min                                8.924723e+05
25%                                1.326569e+06
50%                                1.545676e+06
75%                                1.751344e+06
max                                2.399117e+06
```

Is each index unique?

```
df.index.is_unique
```

[2536] ✓ 0.0s

```
True
```

I cleaned the data by converting all values to numeric and checking for missing or invalid values. None were found.
I verified that each timestamp was unique.
Summary statistics looked normal, with no extreme outliers.

III. Time Series Analysis

III. Time Series Analysis

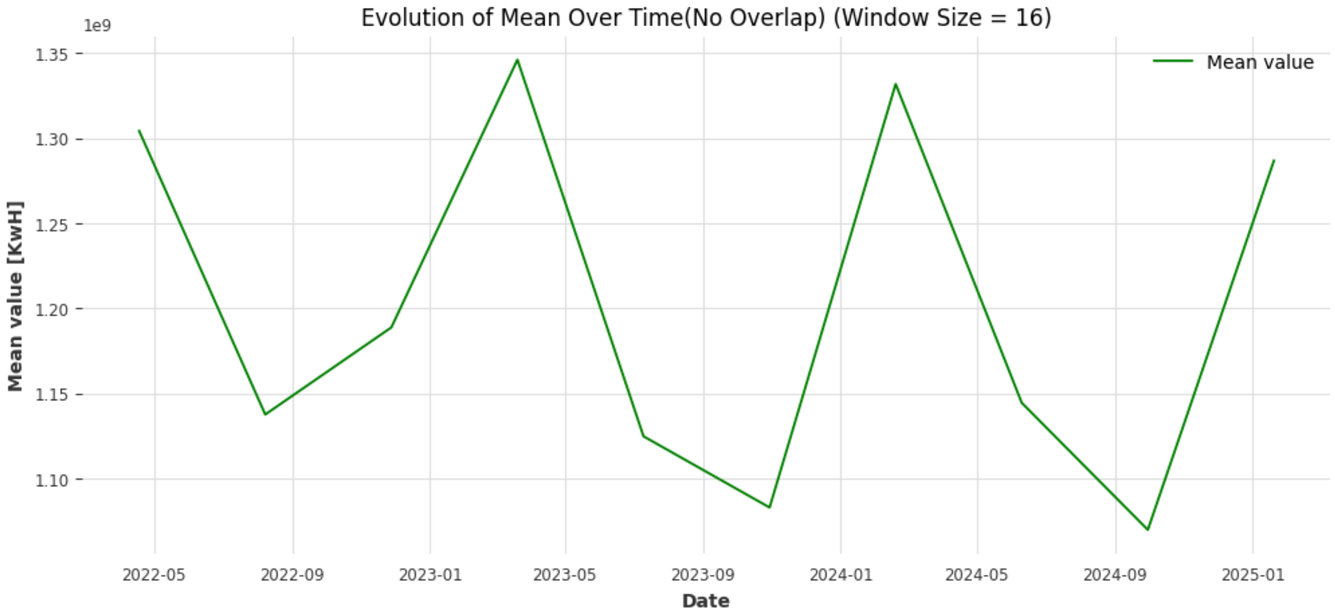
- I. ACF: which lags are most important, does my data have a moving average?
- II. ADF Test: is my data stationnary?
- III. Periodogram: Smoothing, and is my data white noise?
- IV. After modeling, residual analysis, are they (white noise/normal)?

Tests and Forecast

- Tested for stationarity
- Tested for moving average order
- Tested for White Noise and Spectral analysis
- Testing for normality
- Modeling AR, ARMA, ARIMA, SARIMA, SARIMAX

Evolution of Mean over Time

Moving Average



Autocovariance function

- Measures the linear dependence of a time series with a lagged version of itself
- Quantifies how much the values of a time series at different points in time are related to each other, specifically in terms of their covariance

The covariance function for equally spaced data $y_1, \ldots, y_n$ is defined as:

$$c_h = \frac{1}{n-h-1} \sum_{i=1}^{n-h} (y_i - \bar{y})(y_{i+h} - \bar{y}), \quad h = 0, 1, \ldots, n-2,$$

where $\bar{y}$ is the sample mean. The correlogram (ACF) is a graph of $\hat{\rho}_h = \frac{c_h}{c_0}$ against lag h [4].

Interpretation

Source

use one, understand it really well

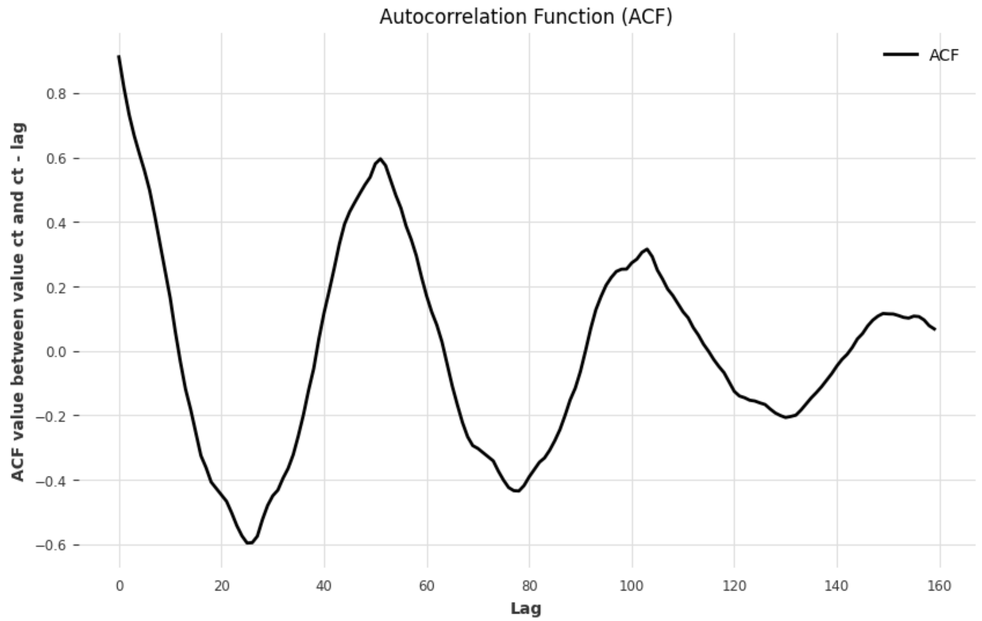
According to the Time Series Analyis book, To summarise: for causal and invertible ARMA models the ACF and PACF have the following properties:

	AR(p)	**MA(q)**	**ARMA(p,q)**
ACF	Tails off	Cuts off after lag q	Tails off
PACF	Cuts off after lag p	Tails off	Tails off

This gives an approach to identifying AR and MA models based on the ACF and PACF, and suggests how to choose p or q. [4]

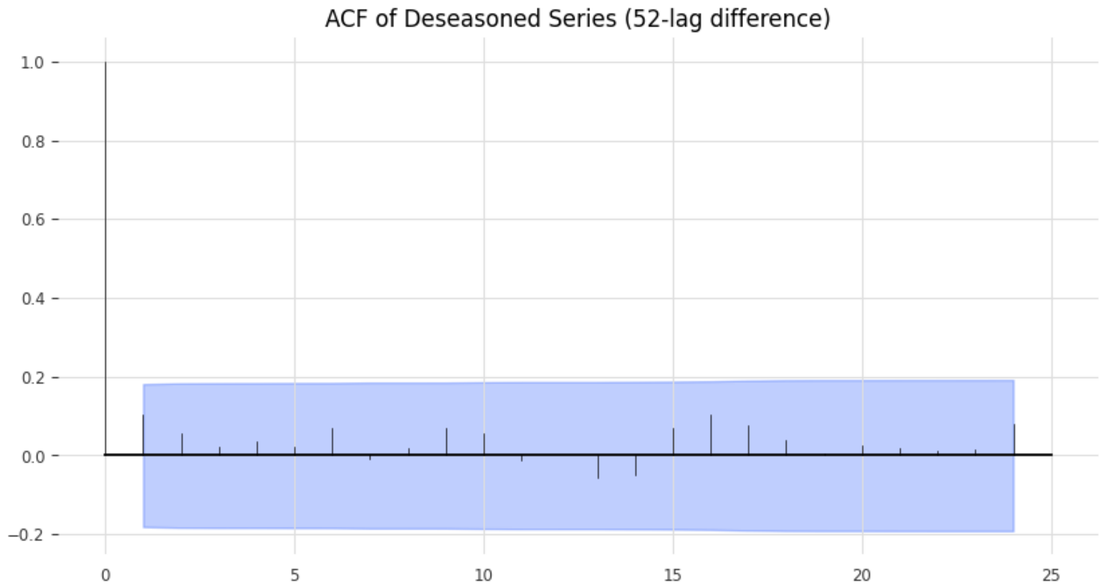
Autocovariance function of original data

Plotted lags for 150-200



ACF tails off, no sharp drop, is cyclical (=> potential seasonality)
High at lag 1, then rapidly decays, likely suggesting an AR(1) model peaking every time at a 52-week cycle

Testing for MA order
ACF of deseasoned data



The ACF values of the deseasoned ts are low, this does not suggest a ARMA model

q = 0

Testing for stationarity

A stationary time series is a time series whose statistical properties like mean, variance, and autocorrelation are all constant over time

KPSS Test: The KPSS test is used to test the null hypothesis that a time series is stationary. It does this by estimating the test statistic:

$$C(l) = \frac{1}{\sigma^2(l)} \sum_{t=1}^{n} S_t^2, \quad \text{where } S_t = \sum_{j=1}^{t} e_j$$

Here, $e_1, \ldots, e_n$ are the residuals from regressing Y_t on a constant or a linear trend (depending on whether testing for level or trend stationarity), and $\sigma^2(l)$ is a long-run variance estimate using a truncation lag l. [4] The test is interpreted as follows:

If kpss result is low ==> stationary

Testing for stationarity

A stationary time series is a time series whose statistical properties like mean, variance, and autocorrelation are all constant over time

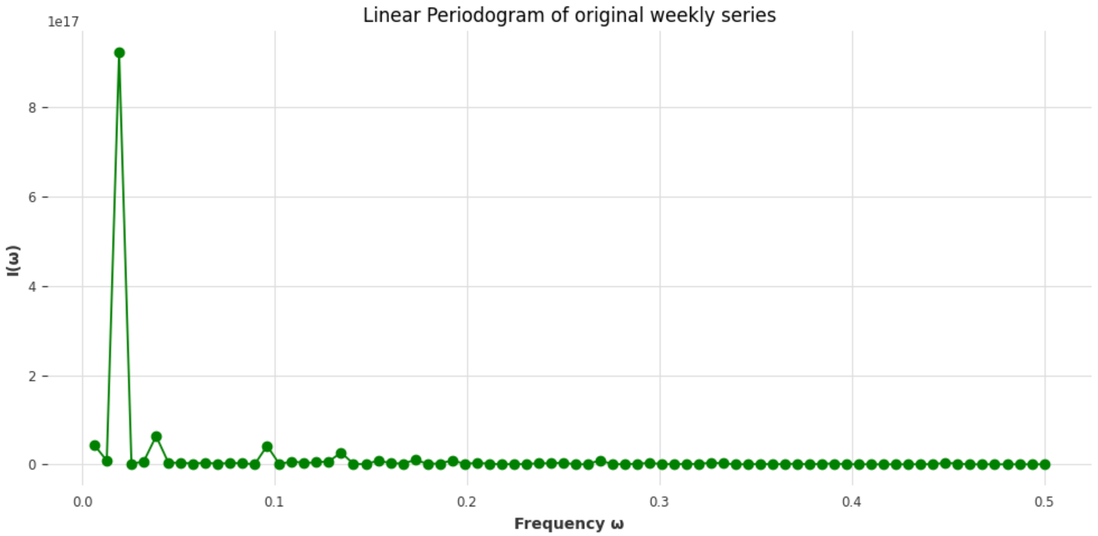
If kpss result is low ==> stationary
interpretation of p-value and kpss statistic, use kpss function instead

No need for ARIMA then (instead of ARMA), d=0

Linear Periodogram

Results

- The spectral decomposition shows that the highest frequency is of 52 weeks, this clearly demonstrates a yearly seasonality.
- The other frequencies are much weaker




```
significant_frequencies = [(w, I) for w, I in zip(freqs, I_vals) if I > threshold]
for freq, power in significant_frequencies:
    print(f"Frequency: {freq:.5f}, Power: {power:.2e}, Period ≈ {1/freq:.2f} weeks")
```

✓ 0.0s

```
Frequency: 0.00641, Power: 4.29e+16, Period ≈ 156.00 weeks
Frequency: 0.01923, Power: 9.24e+17, Period ≈ 52.00 weeks
Frequency: 0.03846, Power: 6.33e+16, Period ≈ 26.00 weeks
Frequency: 0.09615, Power: 4.15e+16, Period ≈ 10.40 weeks
Frequency: 0.13462, Power: 2.65e+16, Period ≈ 7.43 weeks
```

Linear Periodogram
Definition

(a) If $y_1, \ldots, y_n$ is an equally-spaced time series, its periodogram ordinate for ω is defined as

$$I(\omega) = |d(\omega_j)|^2$$

this means that:

$$I(\omega) = \frac{1}{n}\left[\left(\sum_{t=1}^{n} y_t \cos(2\pi\omega t)\right)^2 + \left(\sum_{t=1}^{n} y_t \sin(2\pi\omega t)\right)^2\right], \quad 0 < \omega \leq \frac{1}{2}$$

Spectral decomposition

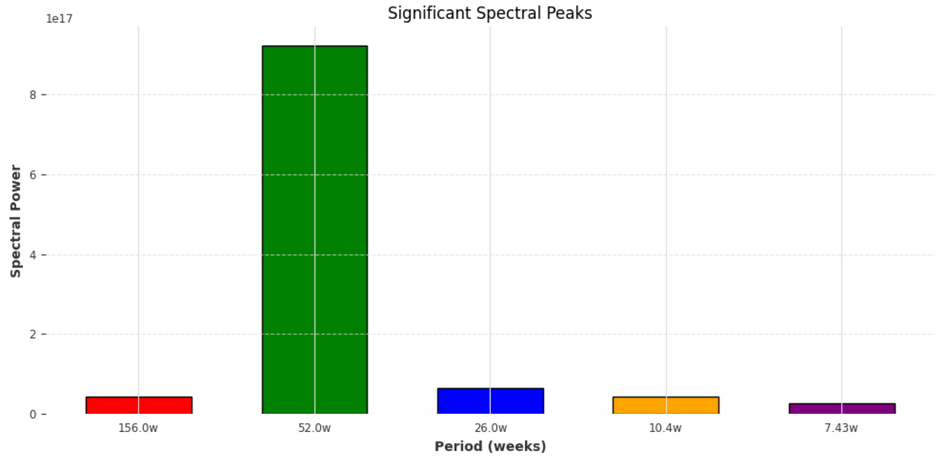
Results

- The spectral decomposition shows that the highest frequency is of 52 weeks, this clearly demonstrates a yearly seasonality.
- The other frequencies are much weaker

```
significant_frequencies = [(w, I) for w, I in zip(freqs, I_vals) if I > threshold]
for freq, power in significant_frequencies:
    print(f"Frequency: {freq:.5f}, Power: {power:.2e}, Period ≈ {1/freq:.2f} weeks")
```

✓ 0.0s

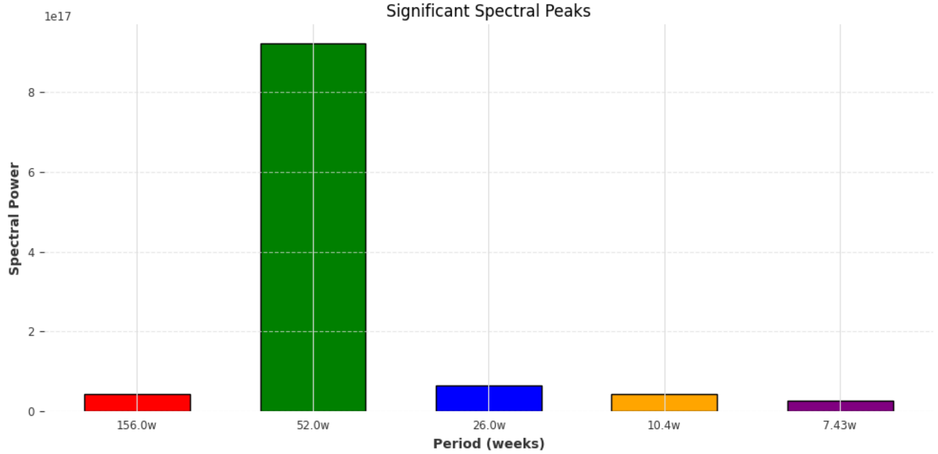
```
Frequency: 0.00641, Power: 4.29e+16, Period ≈ 156.00 weeks
Frequency: 0.01923, Power: 9.24e+17, Period ≈ 52.00 weeks
Frequency: 0.03846, Power: 6.33e+16, Period ≈ 26.00 weeks
Frequency: 0.09615, Power: 4.15e+16, Period ≈ 10.40 weeks
Frequency: 0.13462, Power: 2.65e+16, Period ≈ 7.43 weeks
```

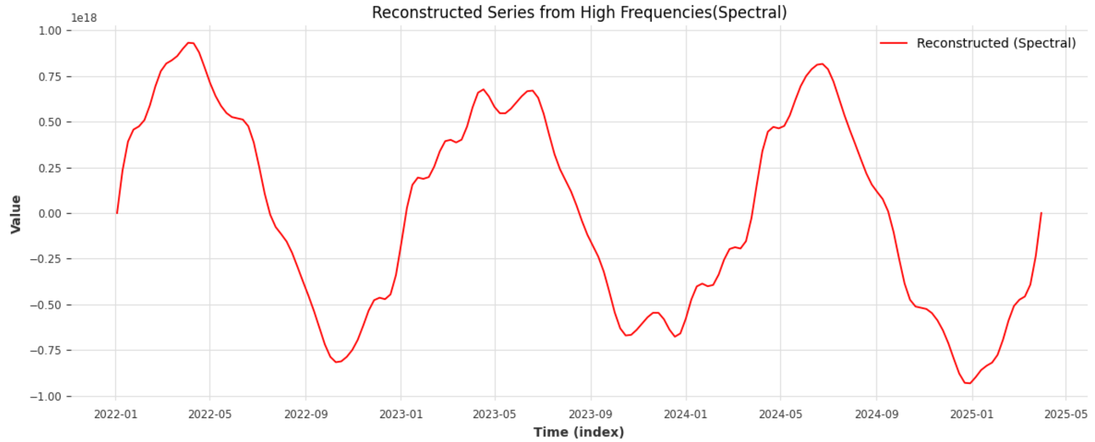


Reconstructed series from most significant frequencies

Result

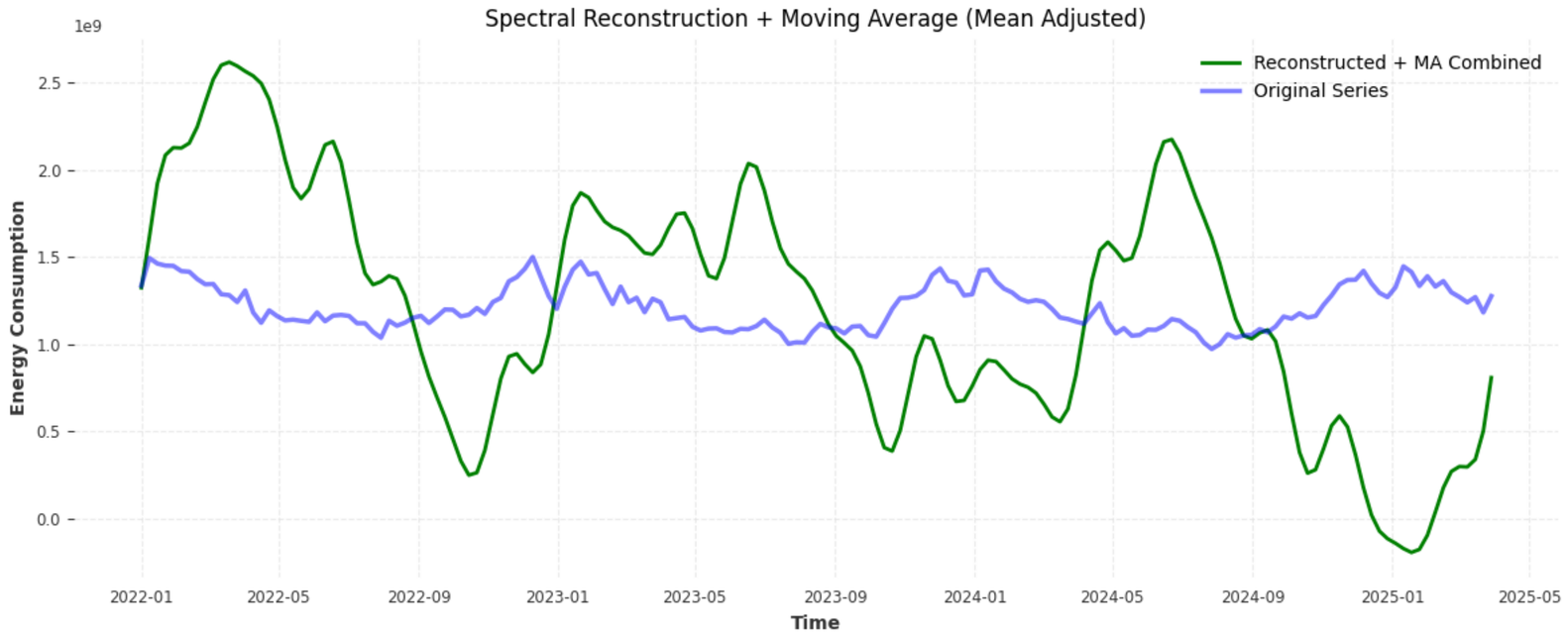
- The spectral decomposition shows that the highest frequency is of 52 weeks, this clearly demonstrates a yearly seasonality.
- Explain why differenced vs original frequencies similar, rigerous argument





Reconstructed series from most significant frequencies

Result



Testing for white noise

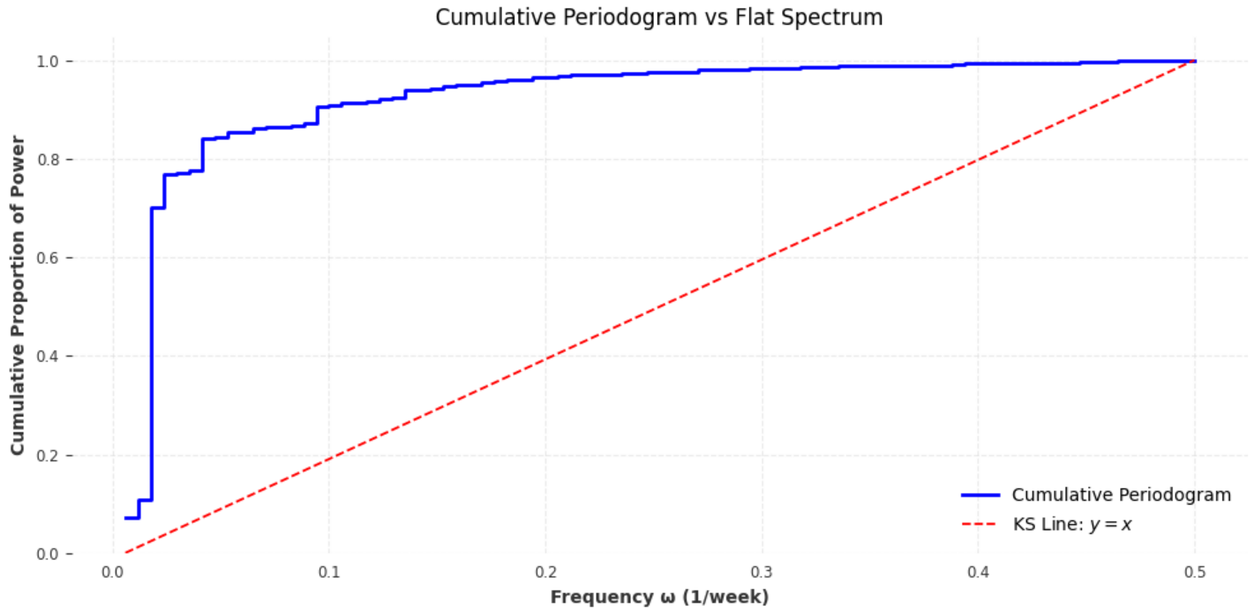
Cumulative Periodogram compared to Gaussian White noise spectrum

(c) The cumulative periodogram

$$C_r = \frac{\sum_{j=1}^{r} I(\omega_j)}{\sum_{l=1}^{m} I(\omega_l)}, \quad r = 1, \ldots, m$$

is a plot of $C_1, \ldots, C_m$ against the frequencies ω_j for $j = 1, \ldots, m$. [4]

According to Davidson, Gaussian and non-Gaussian white noise has a flat spectrum [4]



will be redone after removing periodicity

the frequency is very high near 1/52 weeks, suggesting 52 seasonality again, and also proving that our data is again not white noise

Interpretation

Result

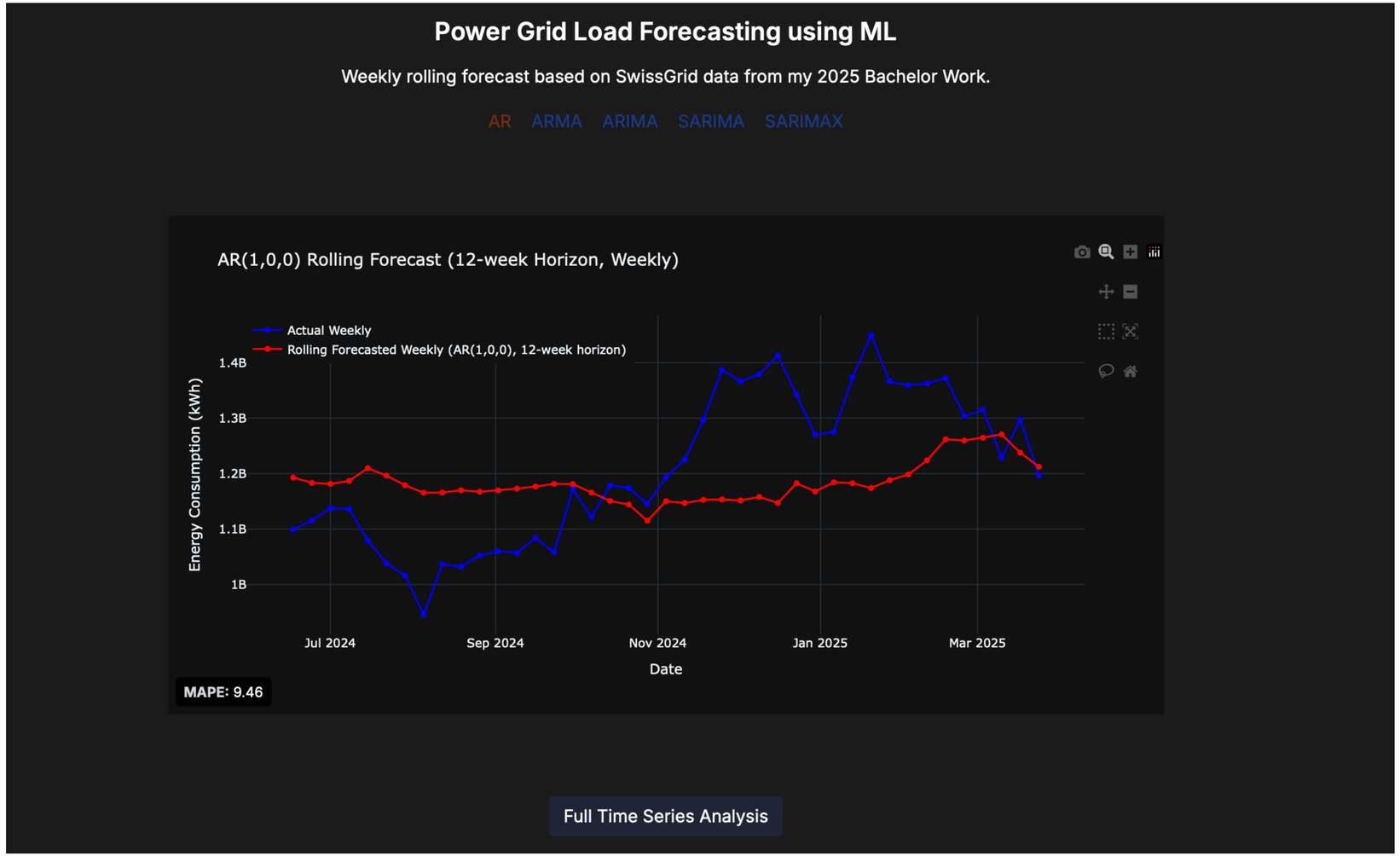
SARIMA (1,0,0 [52])

- From the ljung-box plot and Cumulative periodogram, we concluded our TS is not white noise
- So from the stationarity test, we concluded that d is likely 0, from deseasoned ACF, we concluded that q = 0, and from PACF we concluded that p is likely 1 (plot cumu. period for des ACF)
- From Spectral Analysis, we concluded that the seasonality is clearly 52 weeks, these important results will help us create our forecasting models.

IV. Models

AR Models on website

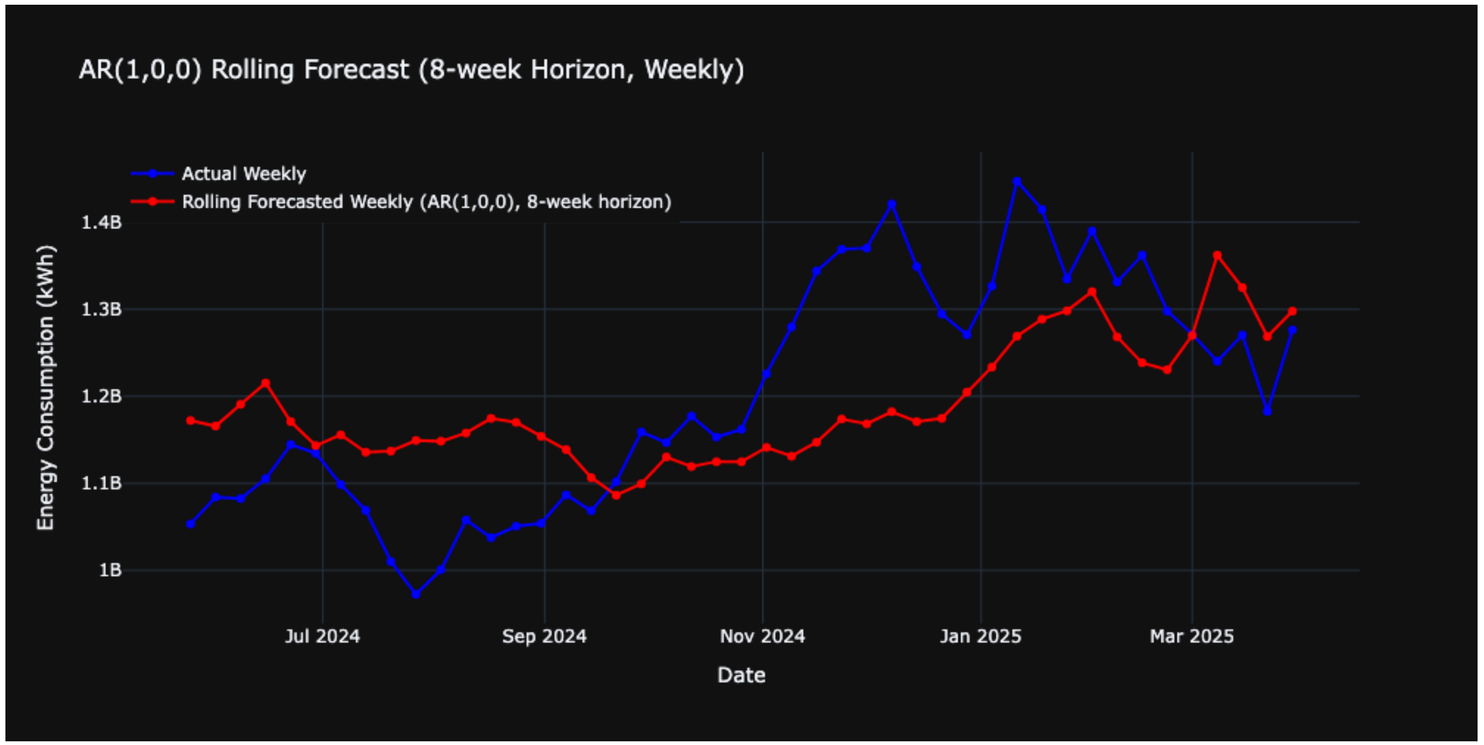
if remove seasonality if correlationnot energy prices



AR

Weekly sum

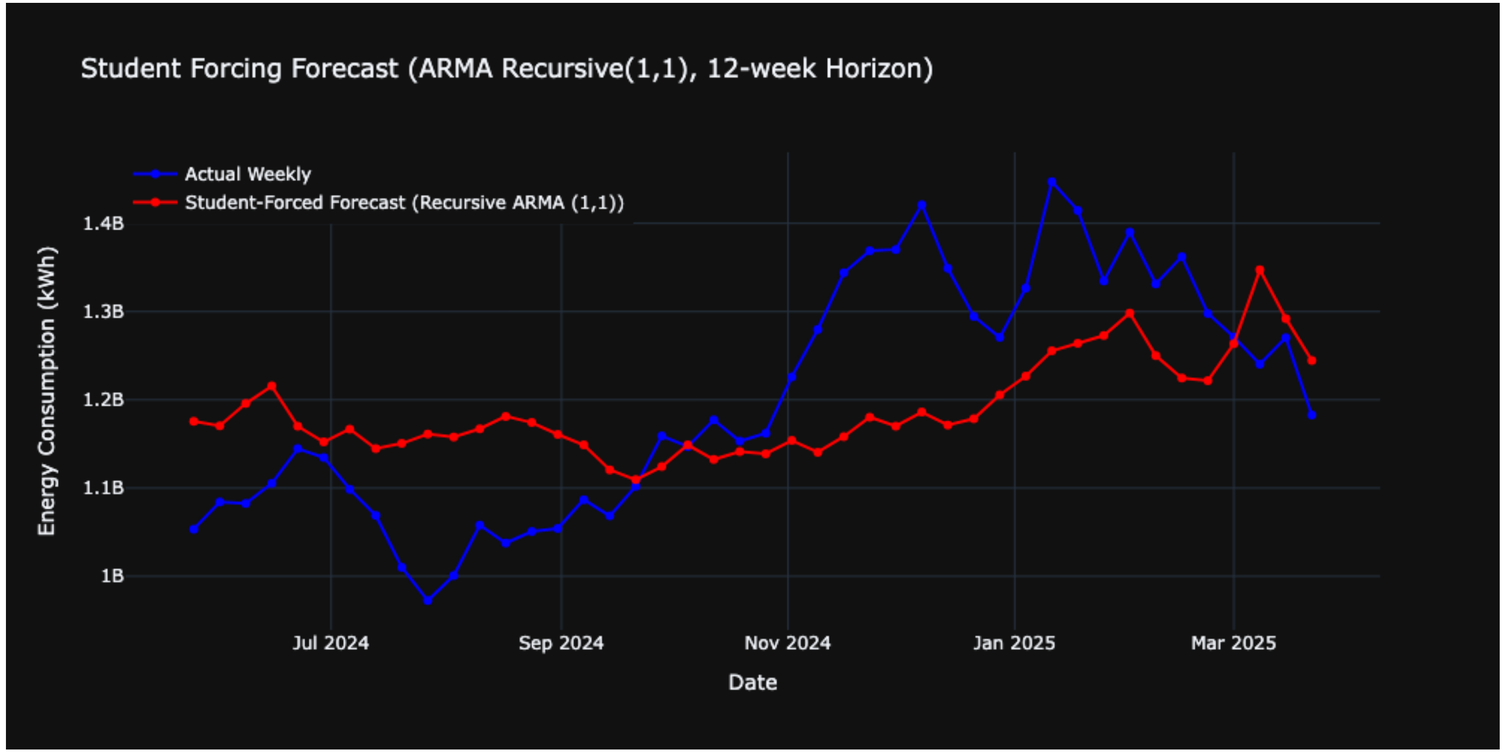
$$X_t = \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t$$



ARMA

Weekly sum

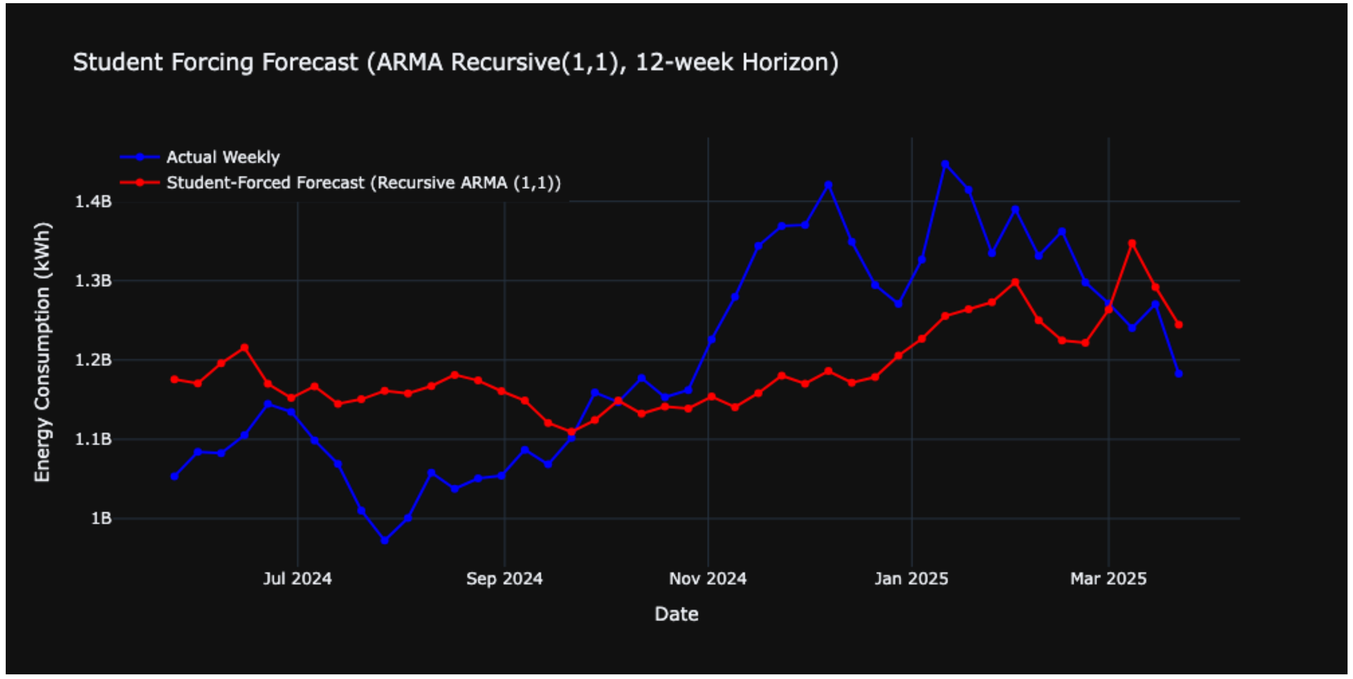
$$X_t = \varepsilon_t + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}.$$



ARIMA

Weekly sum

$$\left(1 - \sum_{i=1}^{p'} \alpha_i L^i\right) X_t = \left(1 + \sum_{i=1}^{q} \theta_i L^i\right) \varepsilon_t$$



SARIMA

Definition

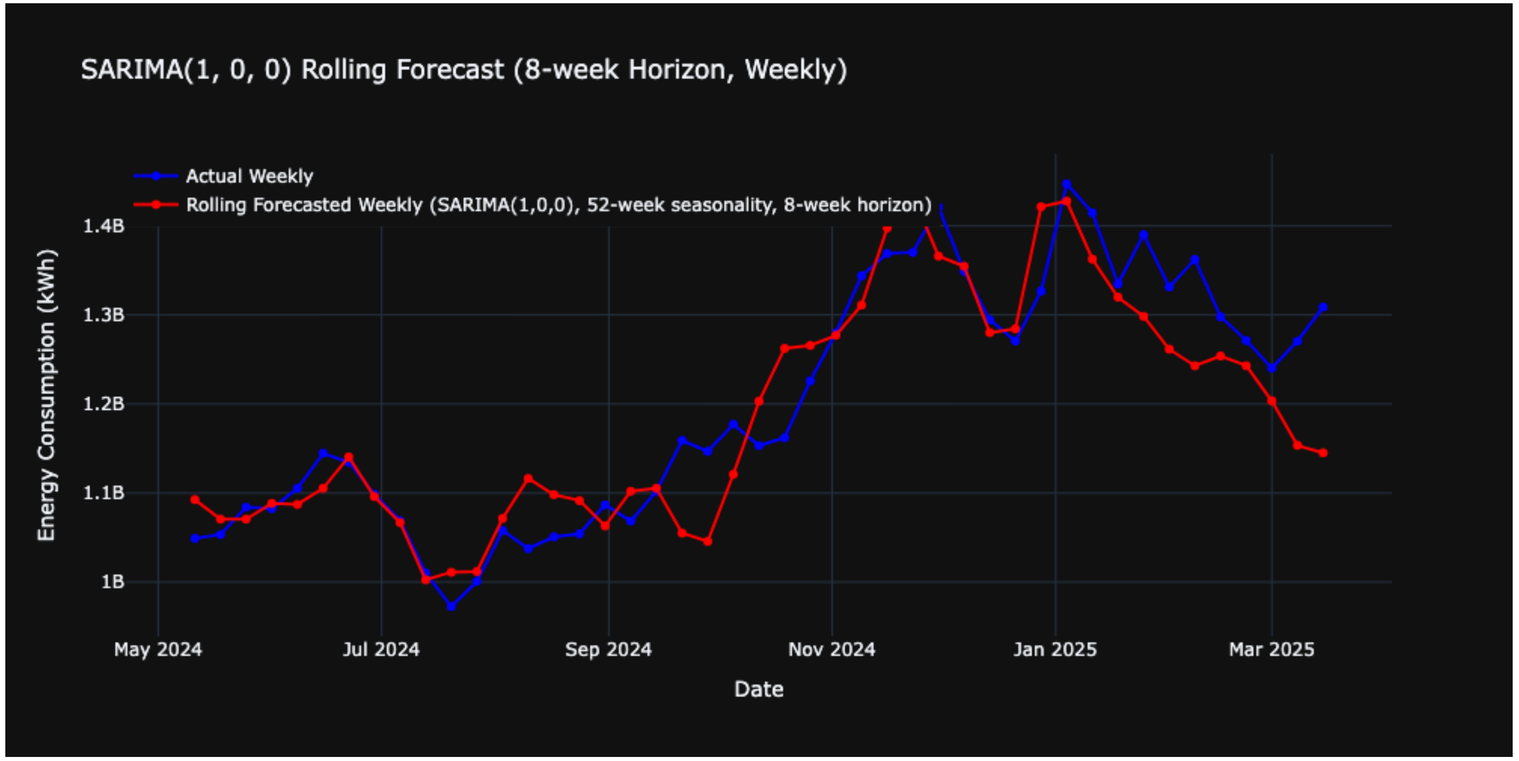
The multiplicative seasonal autoregressive moving average model $\mathrm{SARIMA}(p,d,q)\times(P,D,Q)_s$ is

$$\Phi_P(B^s)\phi(B)(I-B)^d(I-B^s)^D Y_t = \alpha + \Theta_Q(B^s)\theta(B)\varepsilon_t,$$

where $\{\varepsilon_t\}$ is Gaussian white noise. The ordinary autoregressive and moving average components are represented by the operators $\phi(B)$ and $\theta(B)$, respectively; the seasonal autoregressive and moving average components by $\Phi_P(B^s)$ and $\Theta_Q(B^s)$, of orders P and Q; and the ordinary and seasonal difference components by $(I-B)^d$ and $(I-B^s)^D$ of orders d and D.

SARIMA

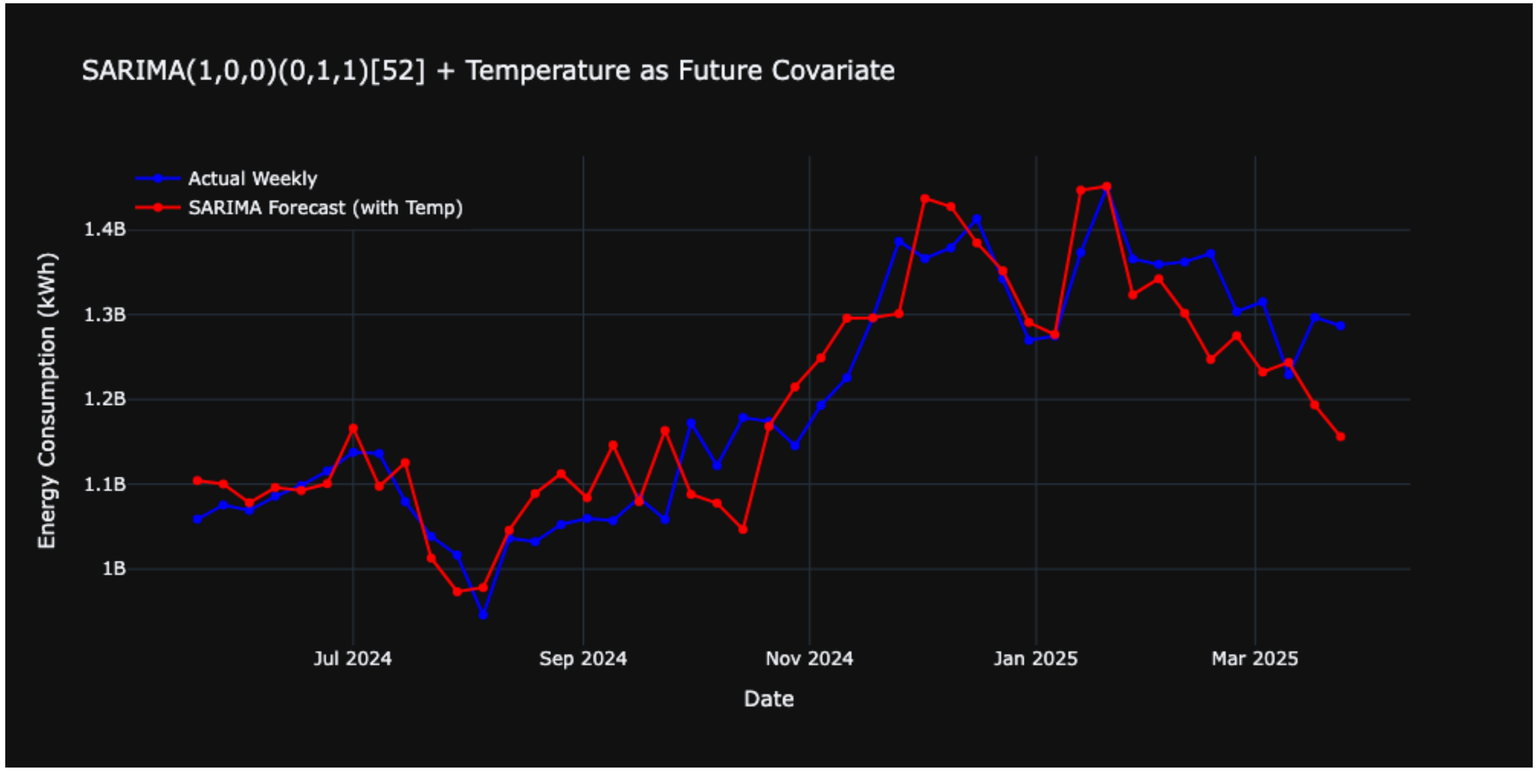
Weekly sum



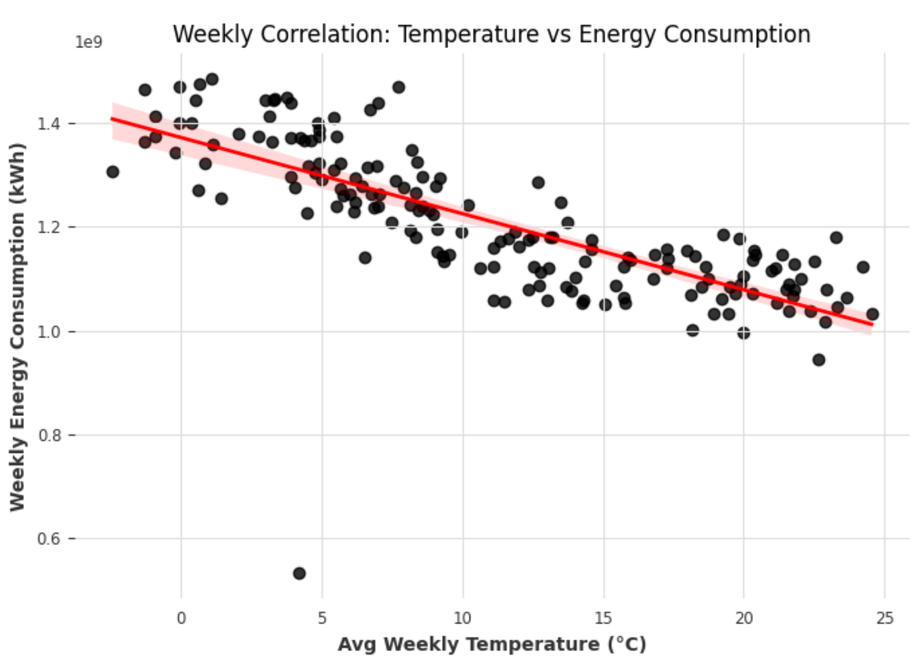
SARIMAX

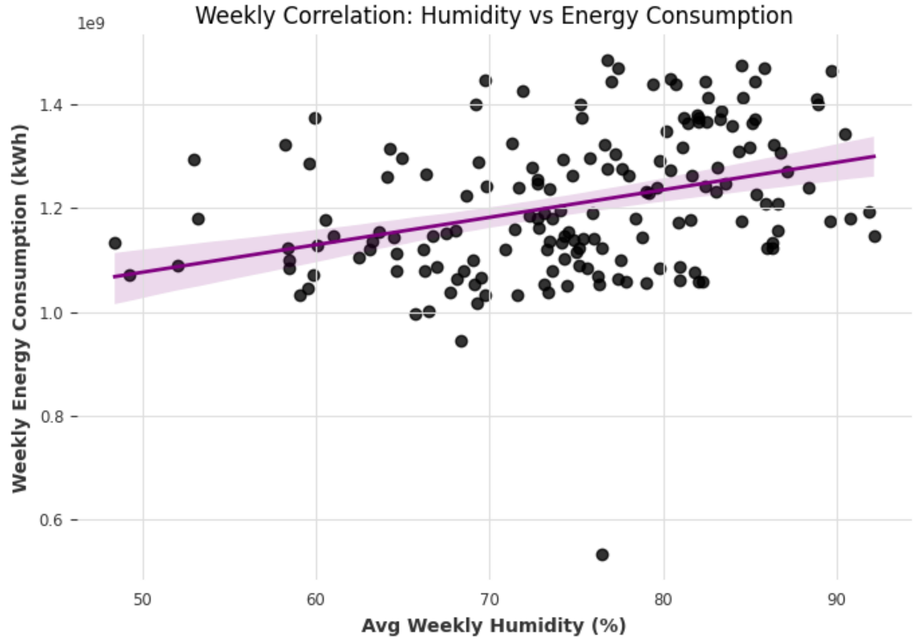
Weekly sum

$$\Theta(L)^{p}\theta(L^{s})^{P}\Delta^{d}\Delta_{s}^{D}y_{t} = \Phi(L)^{q}\phi(L^{s})^{Q}\Delta^{d}\Delta_{s}^{D}\epsilon_{t} + \sum_{i=1}^{n}\beta_{i}x_{t}^{i}$$



Correlations

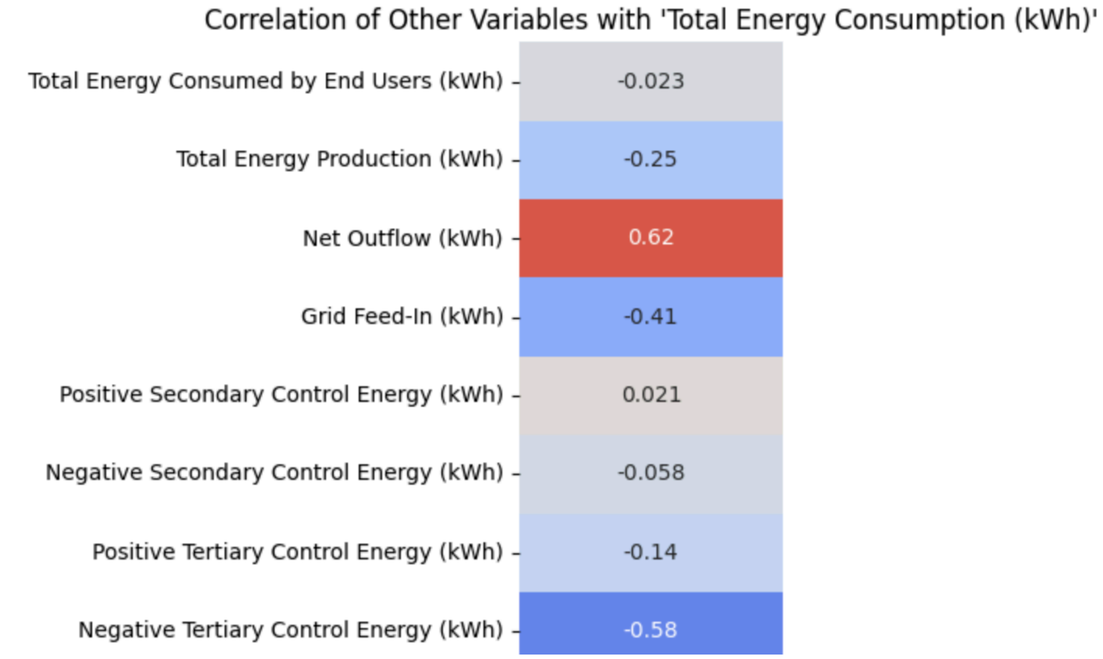


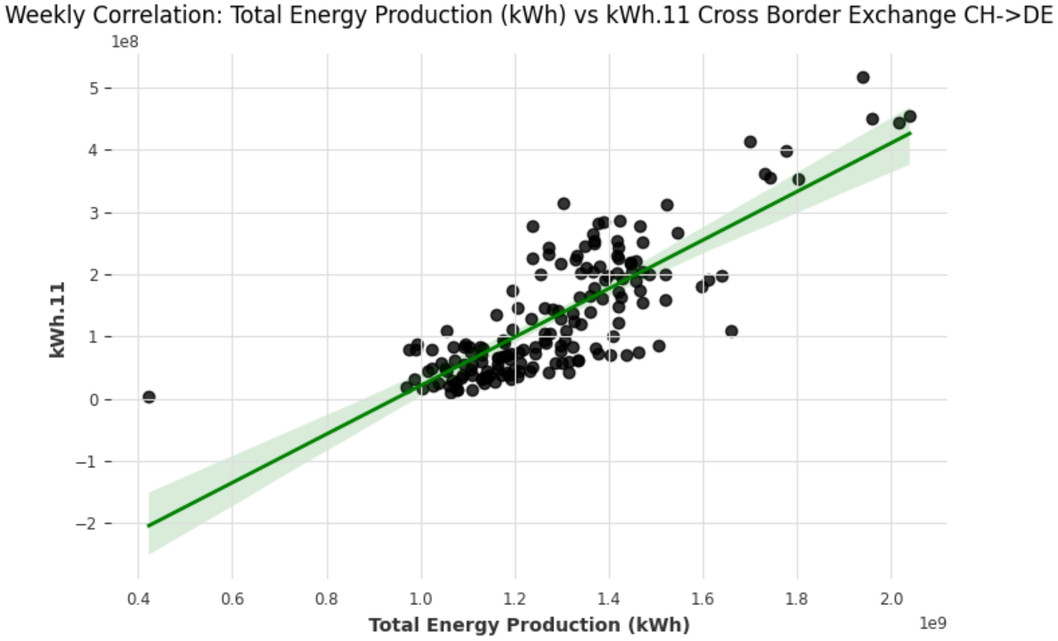


✅ Pearson correlation:

	energy
temperature	-0.750454
humidity	0.350662

Correlations with other variables in the same data file






```
Pearson correlation: 0.8053
```

V. Evaluation

MAPE
Evaluation Metric

7.1.1 Definition

The mean absolute percentage error (MAPE) is a metric used to evaluate the accuracy of a forecasting model. It calculates the average of the absolute percentage errors between the predicted values and the actual values. Lower MAPE values indicate more accurate forecasts.

The formula is:

$$\text{MAPE} = \frac{100\%}{n} \sum_{t=1}^{n} \left| \frac{y_t - \hat{y}_t}{y_t} \right|$$

where y_t is the true value and $\hat{y}_t$ is the predicted value at time t. [15]

MAPE
Evaluation Metric

Model Comparison – MAPE Scores

Model	MAPE (%)
AR(1)	8.01
ARMA(1,1)	8.12
ARIMA(1,1,1)	8.16
SARIMA(1,0,0 [0,1,1,52])	3.53
SARIMAX(1,0,0 [0,1,1,52]) with Temp	3.86

We concluded that the SARIMA (1,0,0,[52]) is best

IV. Challenges and discussion

Thanks for coming !