ENERGY CONSUMPTION ANALYSIS

Group 7: Prateek Singh
MS in Data Science

Sumanth Reddy Thandra
MS in Data Science

Research Questions?

1. How has energy consumption evolved from 2002 to 2018, and what are the key seasonal or cyclical patterns?

2. How well do SARIMA models perform in short-term vs. long-term energy consumption forecasting, and what are their limitations and how they compare in forecasting energy consumption, and which model provides the best balance between accuracy and complexity?

Description of the Dataset:

The PJME_hourly dataset is a univariate time series with hourly electricity consumption (in MW) from the PJM East region. PJM Interconnection operates the power grid across parts of 13 U.S. states

- **Total Records: 145,366 hourly** observations.
- **Time Span:** From **Jan 1, 2002 - Jun 1, 2018**

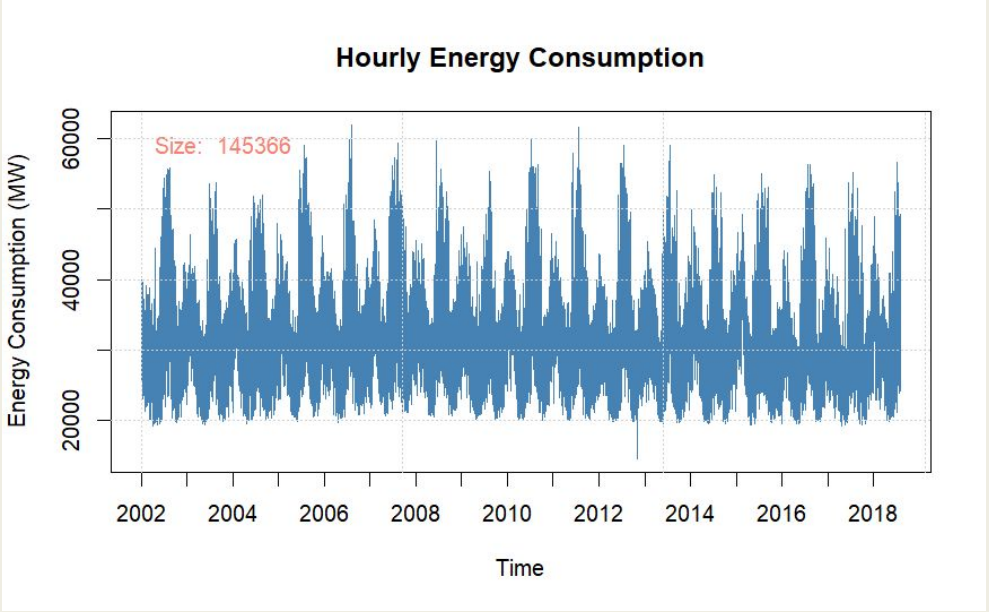
Data Preprocessing:

Load PJME_hourly and then Datetime column is converted to POSIXct format for proper date-time handling.

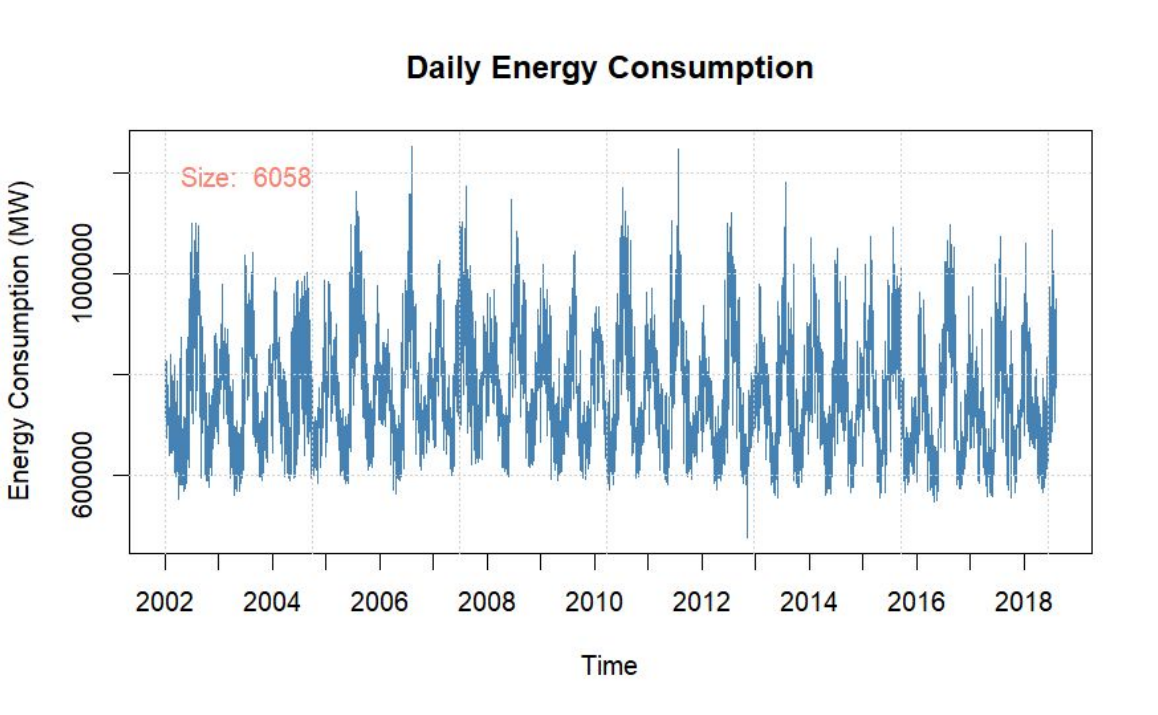
Hourly Energy Consumption:

- The hourly data is **too extensive and noisy** to observe any long term trends.
- The data appears to follow constant mean and constant variance, suggesting a stationary process **(weak stationarity).**
- Hence we check for stationarity
- So, converting to daily data also helps **reduce dimensionality** and improves model performance in forecasting.
- **Aggregation** will preserves essential structure and remaking the data easier to interpret and model.
- **Lag choices (24, 52)** in tests suggest daily/weekly patterns

Test	p-value	Conclusion
Augmented Dickey-Fuller	0.01	Stationary
KPSS	0.01	Non-stationary
Phillips-Perron	0.01	Stationary

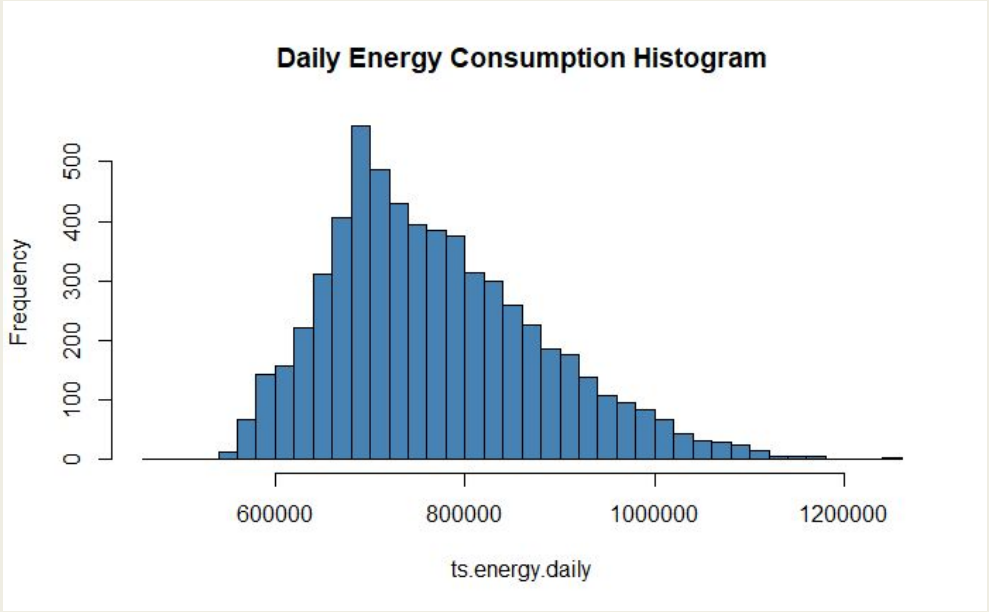


Daily Energy Consumption:

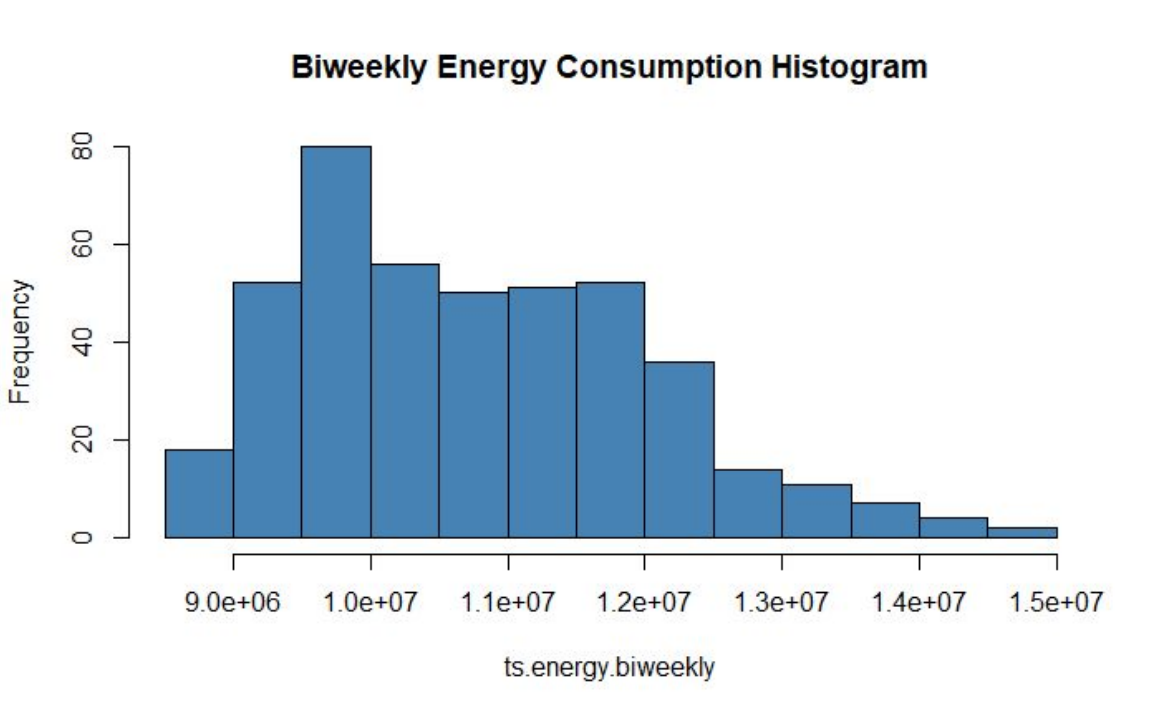


- After converting to daily format, the dataset size becomes more manageable with **6058 observations**
- Daily consumption values show a **roughly normal distribution**
- **Still extensive**, but easier to visualize compared to hourly data
- **Clear seasonal patterns** emerge — regular peaks and dips suggest recurring behavior
- **Signs of a weekly/biweekly trend**, indicating periodicity
- **Lag 11–18 (days)** from tests suggests the tests are accounting for autocorrelation up to **~ 1.5 - 2.5 weeks**.

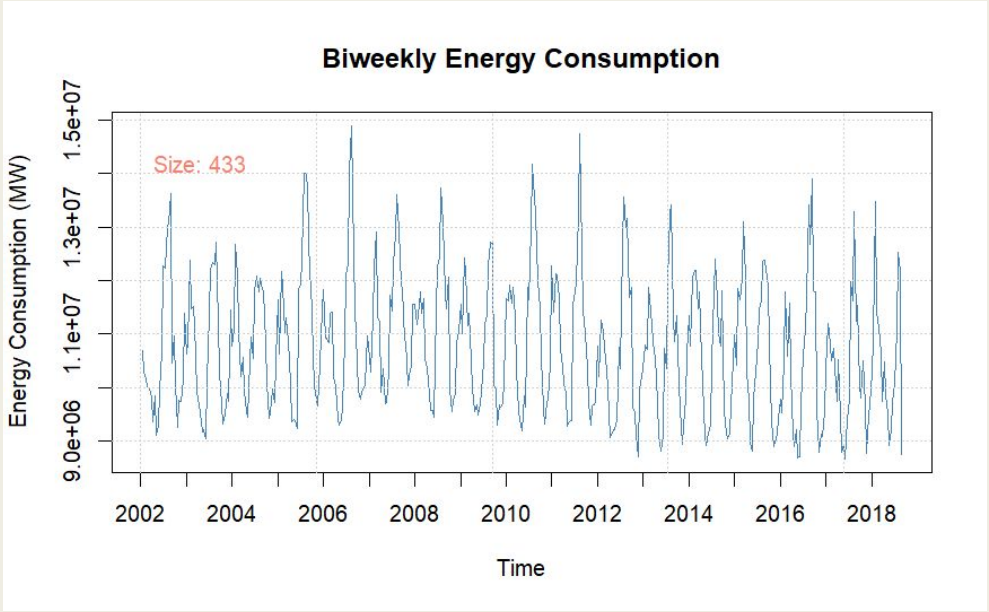
Test	p-value	Conclusion
Augmented Dickey-Fuller	0.01	Stationary
KPSS	0.01	Non-stationary
Phillips-Perron	0.01	Stationary



Bi-Weekly Energy Consumption:



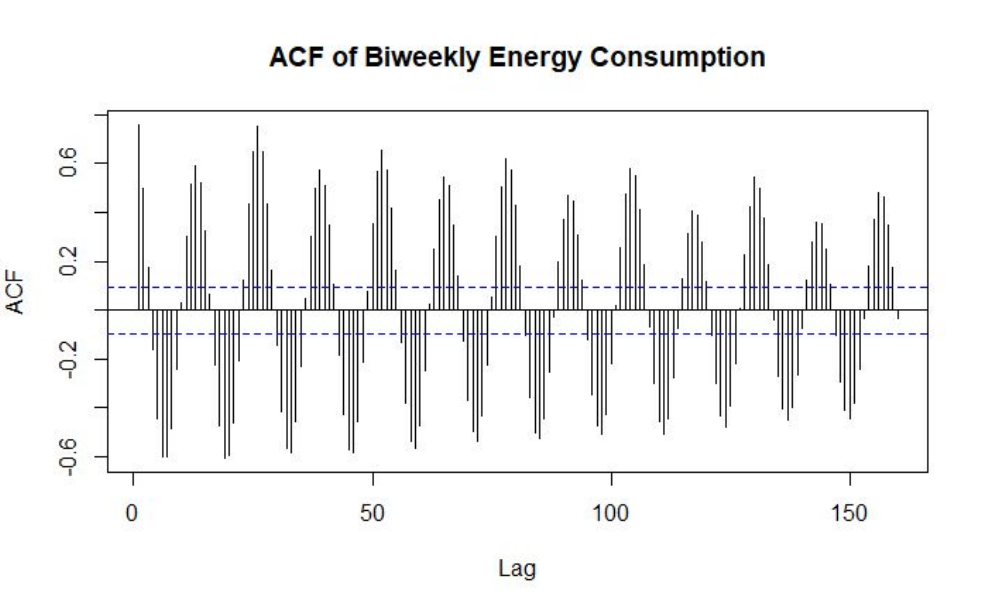
- The dataset now has **433 observations,** which is manageable for plotting and analysis
- **Clear seasonal trends** are evident in the data, suggesting recurring patterns over time
- The data **does not show** a general **upward or downward trend or any polynomial trend** in the long term
- Instead, the data more appropriately follows **a seasonal trend**, rather than a deterministic trend (no consistent long-term increase or decrease)

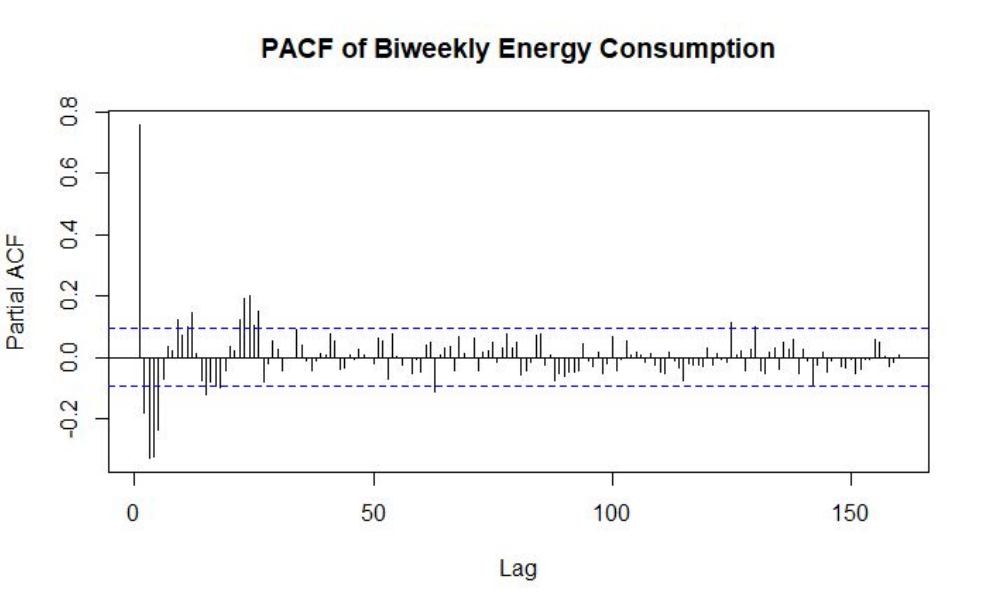


Checking for stationarity for Stochastic Trend

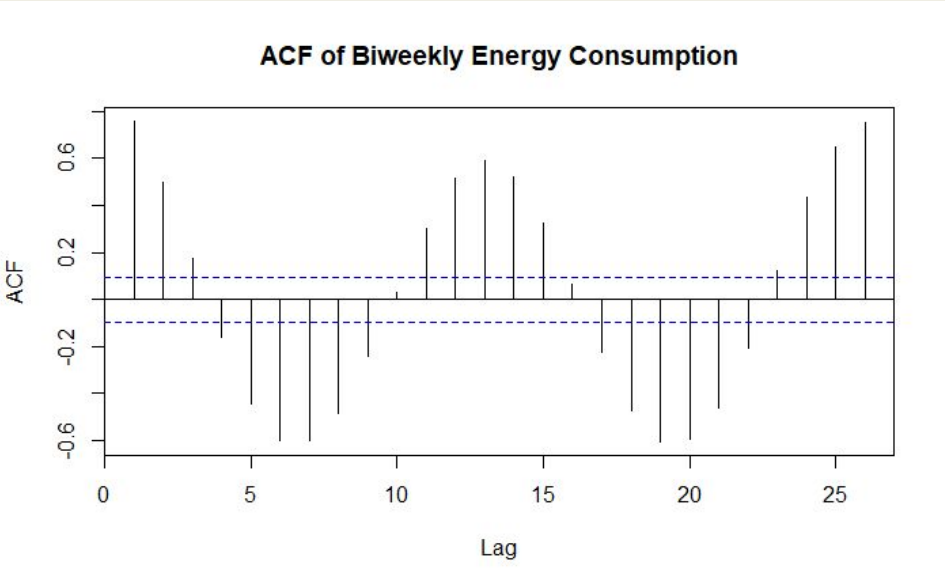
- ADF & PP Tests (p = 0.01) → Strong evidence against unit root (data is stationary).
- KPSS Test (p = 0.1) → Fails to reject level stationarity (supports stationarity)
- The decomposed ACF and PACF, show strong seasonality.

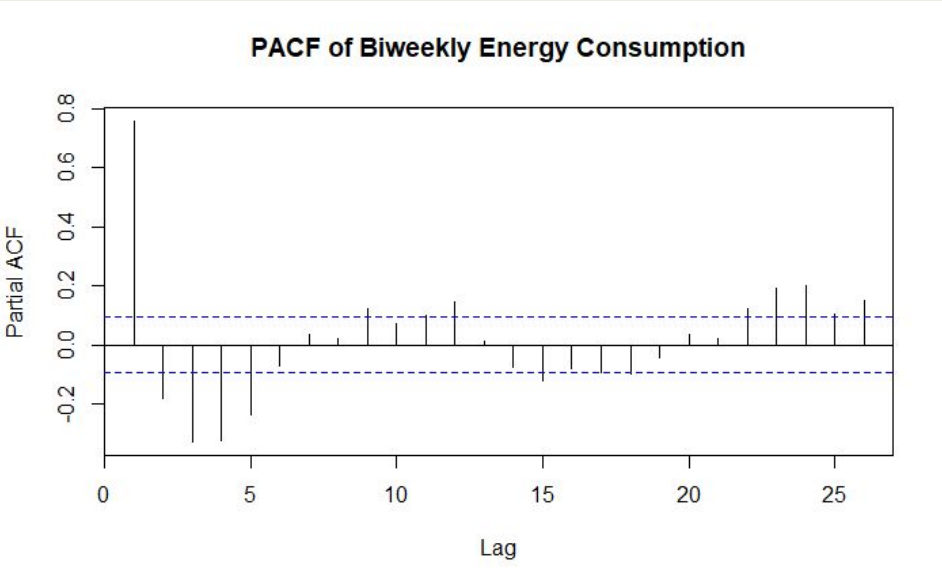
Test	p-value	Conclusion
ADF	0.01	Stationary
KPSS	0.1	Stationary
PP	0.01	Stationary





ACF, PACF, EACF Analysis:





AR/MA

	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	x	x	x	x	x	x	x	x	x	o	x	x	x	x
1	x	x	x	x	x	x	x	x	x	o	x	x	x	x
2	x	x	o	o	o	o	o	o	o	o	o	o	o	o
3	x	x	o	o	o	o	o	o	o	o	x	o	o	o
4	x	x	x	o	o	o	o	o	o	o	o	o	o	o
5	x	x	x	x	o	o	o	o	o	o	x	o	o	o
6	x	x	o	x	x	x	o	o	o	o	x	o	o	o
7	x	x	x	x	o	x	o	o	o	o	x	o	o	o

- For Non-Seasonal part
  - ACF (decay with no sharp cutoff) → **MA(1), MA(2)**
  - PACF (sharp cutoff at Lag 1)→ **AR(1)**
  - EACF → (o) dominates after 2, 2
  - AR(2): Lags beyond 2 become insignificant
    - → **AR(2) may be adequate.**
  - MA: Lags beyond 2 become insignificant
    - → **MA(2) likely sufficient**.
- For Seasonal part
  - ACF (decays slowly with not sharp cutoff till lag 14)
    - → Differencing might be needed
  - PACF (Cuts off after lag 1)
    - → AR(1) may be adequate

Candidate Modelling: (Based on ACF, PACF, EACF)

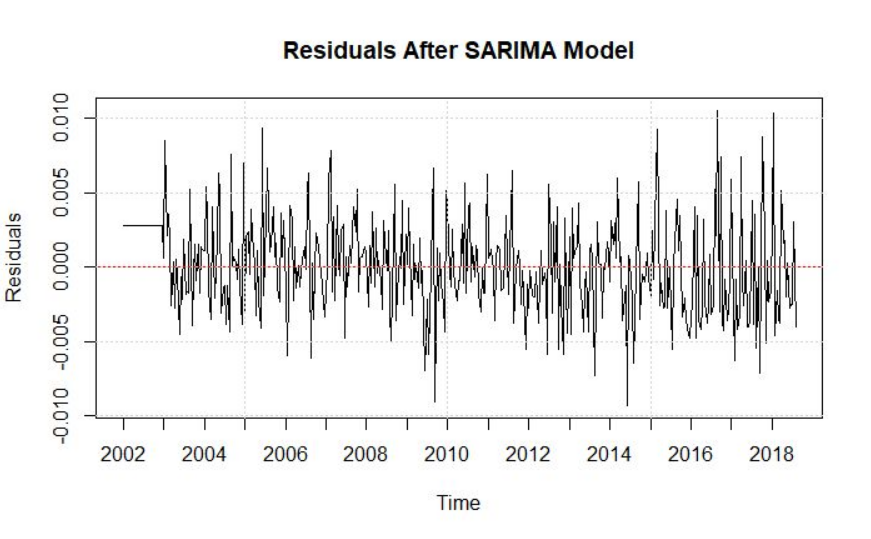
	Model <chr>	AIC <dbl>	BIC <dbl>
3	ARMA(2,0,2)	12831.61	12856.02
7	Auto.ARIMA	12831.61	12856.02
6	ARMA(2,1,2)	12852.81	12873.14
2	ARMA(2,0,1)	12890.38	12910.72
1	ARMA(2,0,0)	12968.22	12984.49
4	ARMA(2,1,0)	12993.40	13005.60
5	ARMA(2,1,1)	12995.34	13011.60

Best models: ARMA(2,0,2), ARMA(2,1,2), ARMA(2,0,1)

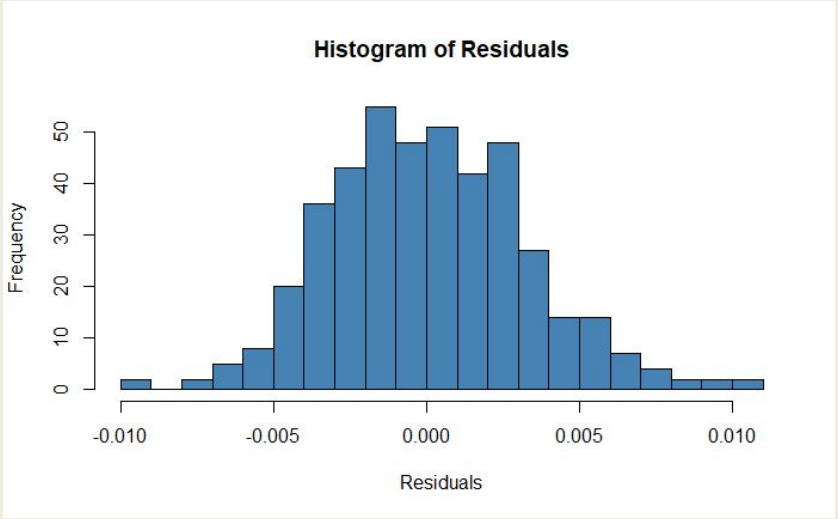
	Model <chr>	AIC <dbl>	BIC <dbl>
2	SARIMA(2,0,2)(1,1,1)	11982.01	12010.05
3	SARIMA(2,0,2)(1,1,2)	11982.73	12014.79
4	SARIMA(2,0,1)(0,1,1)	11983.02	12003.05
5	SARIMA(2,0,1)(1,1,1)	11983.71	12007.75
6	SARIMA(2,0,1)(1,1,2)	11984.29	12012.33
1	SARIMA(2,0,2)(0,1,1)	11984.88	12008.92
7	Auto.ARIMA	12022.14	12042.17

Best models: SARIMA(2,0,1)[0,1,1], SARIMA(2,0,2)[1,1,1], SARIMA(2,0,1)(1,1,1)

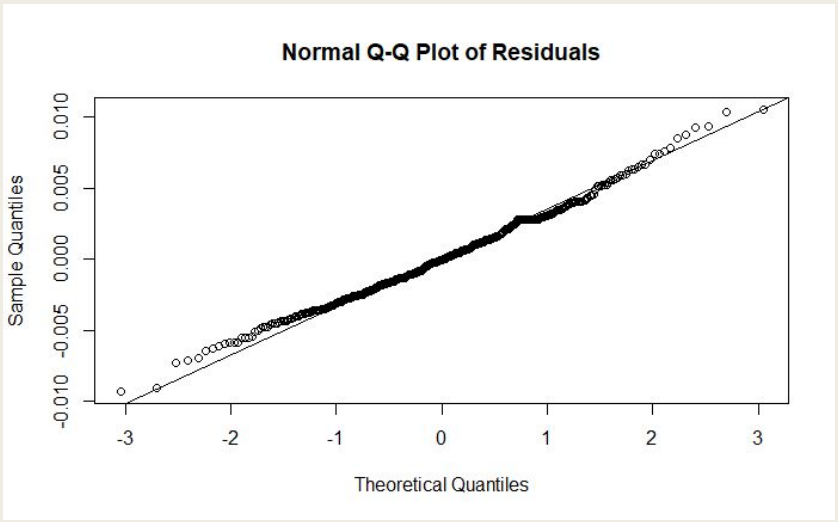
Model Diagnostic on Final Model: **SARIMA(2,0,1)(0,1,1)[26]**

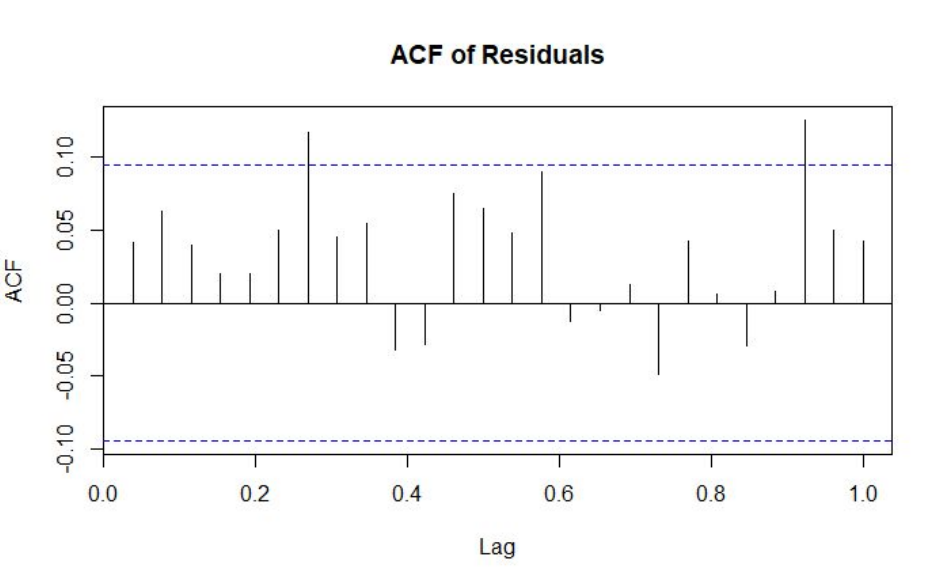


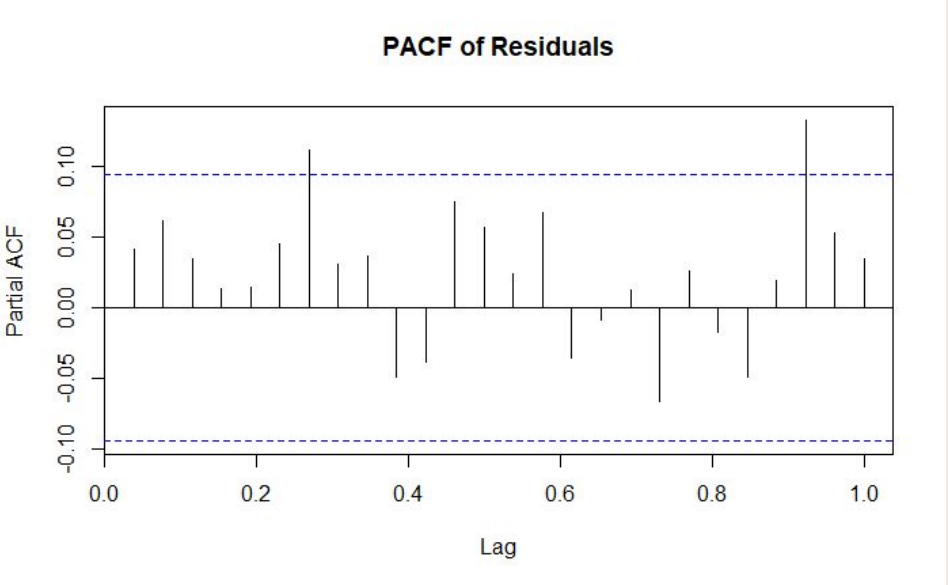
- The Energy consumption was in ten's of millions
- Performed **Double Log for transformation**

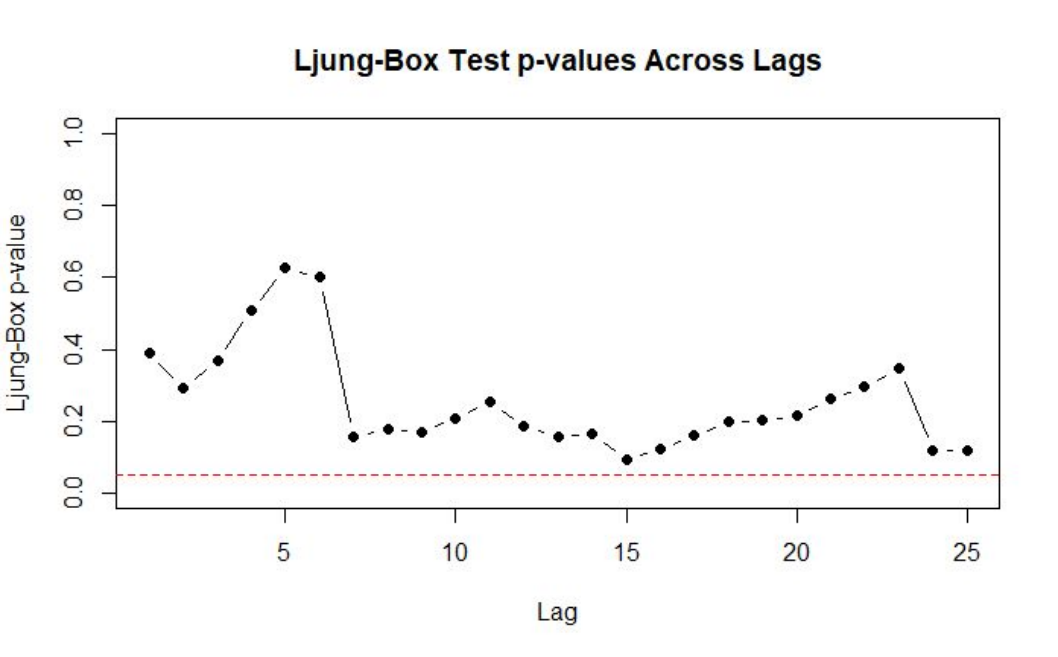


- **Residual Mean:** The residuals have a mean close to zero, indicating no significant bias in the model.
- **Homoscedasticity:** Residuals seem to have constant variance, but there are signs of **increased fluctuations after 2012-13.**
- **Shappiro-Wilk test:** Confirms close to normality (p-value = 0.04306)



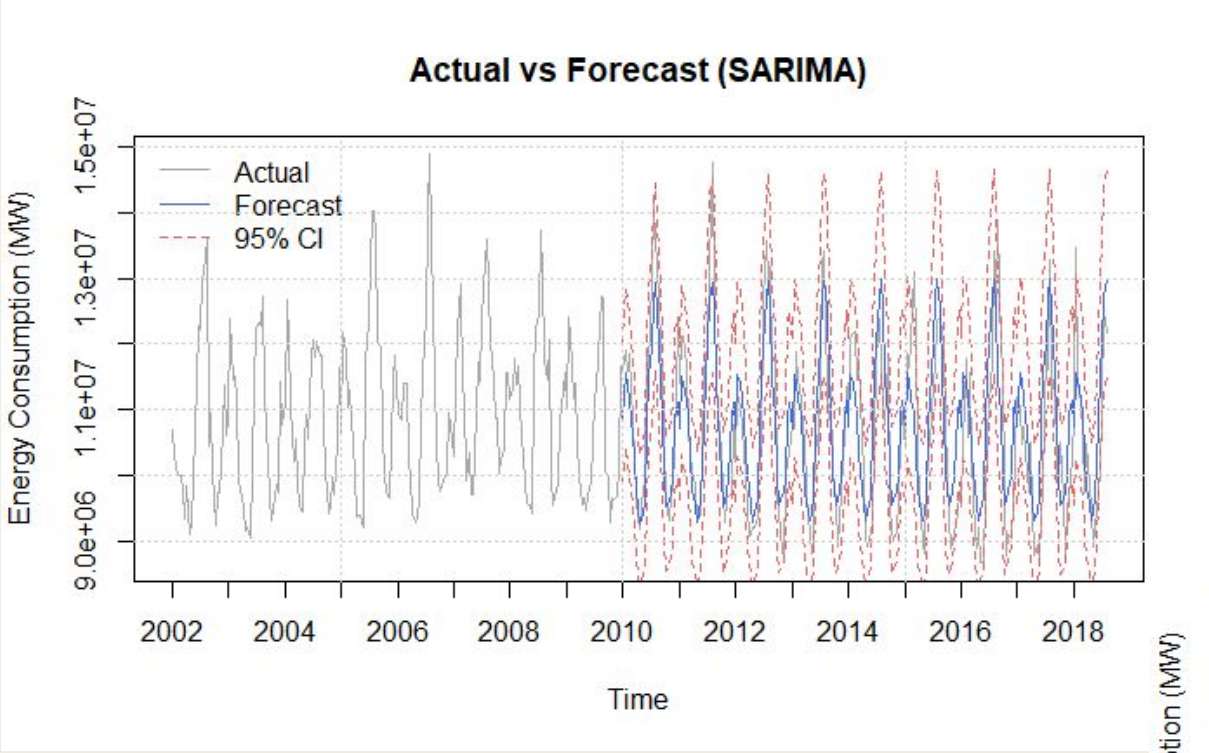






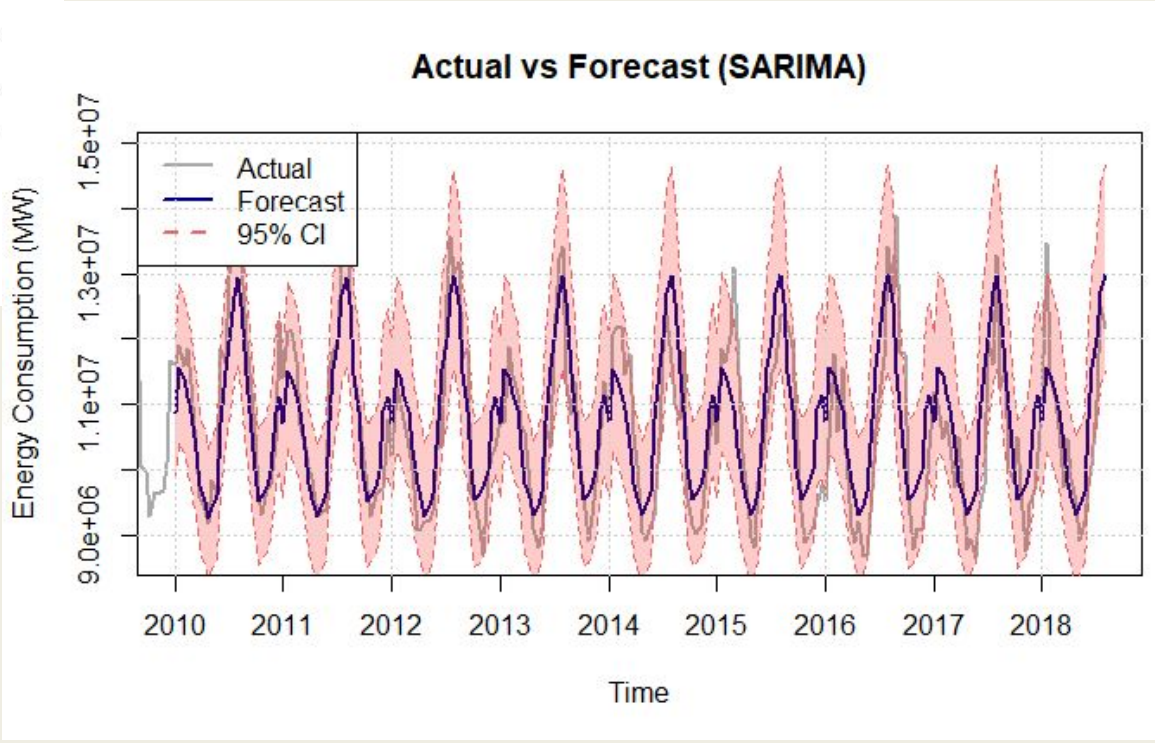
- **ACF and PACF plots** show no significant autocorrelation.
- **Box-Ljung test** confirms the independence of residuals (p-value = 0.2432).
- **Ljung-Box** p values across suggest independence as well

Prototype Modelling: **SARIMA(2,0,1)(0,1,1)[26]**



Metric <chr>	Value <dbl>
MAE	585420.028
RMSE	739120.764
MAPE (%)	5.492
R-squared	0.681

- Training on **~50%** data.
- **2002-2010 approx**
- **208 observations**
- Testing on rest



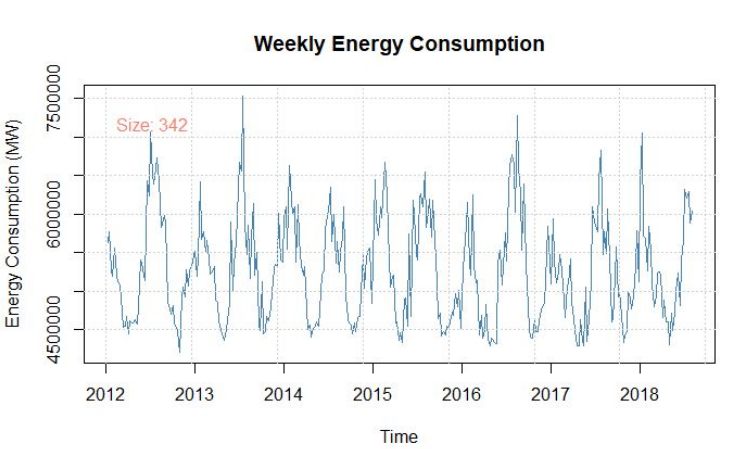
Conclusion:

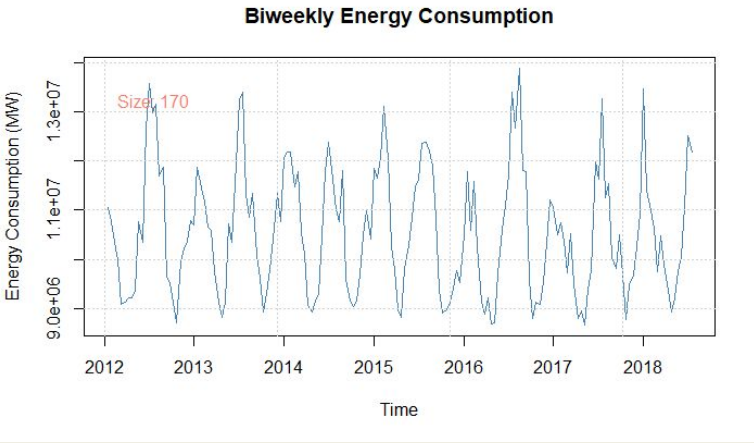
- The model **SARIMA(2, 0, 1)(0, 1, 1)[26]** performed pretty well.
- A **double log transformation** was necessary due to the high scale of energy data.
- The residuals for the model **show increasing variance after 2012–13** in the residual analysis.
- Training on 50% of the data and testing on the rest gave an **R-squared of 68.1%**.
- **Overfitting didn't help** with accuracy.

- We are testing data from around **2011 onwards**.
- There might be some additional factors impacting the data from **2012 onwards**.

- It **should be a good idea to split the series** and perform further analysis to get **short-term trend** analysis for the data after 2012.

Post 2011 Weekly / Bi-Weekly Data (Stochastic Trend):





Test	p-value	Conclusion
ADF	0.01	Stationary
KPSS	0.1	Stationary
PP	0.01	Stationary

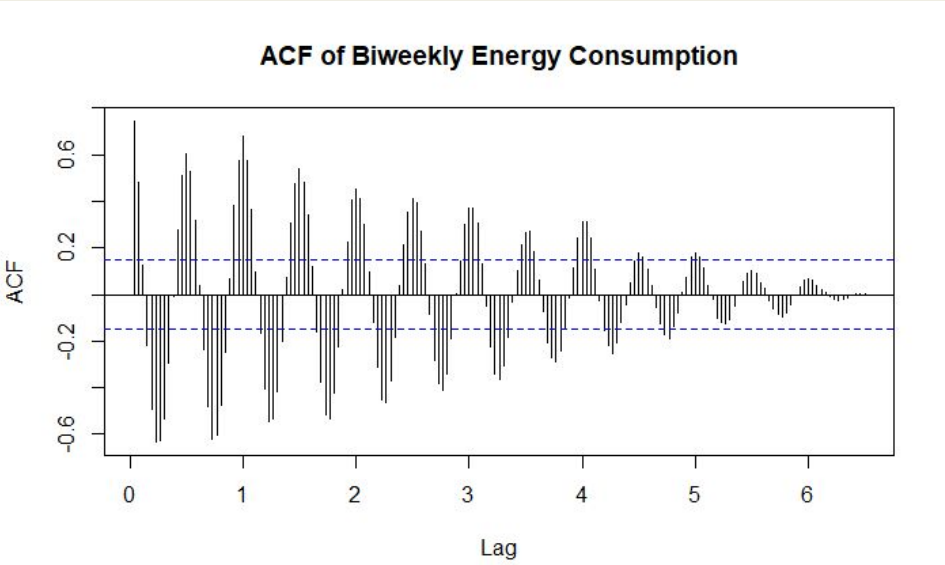
Both series are **stationary** with same **p-value level significance level**

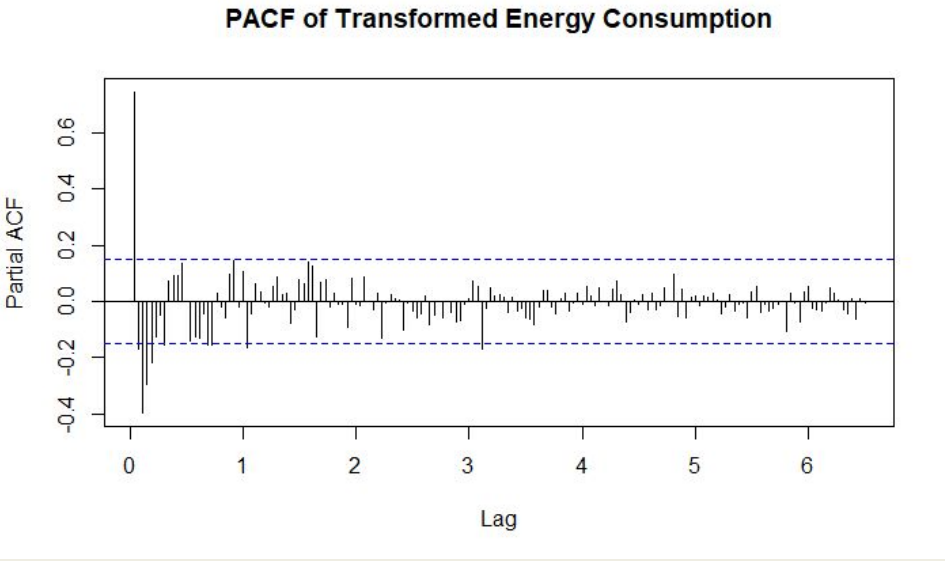
ADF and KPSS tests favor the **biweekly series** for stronger **stationarity**

Test	Weekly Series	Biweekly Series	Better Stationarity
ADF Test	DF = -7.894	DF = -8.571	Biweekly (more -ve score)
KPSS Test	KPSS = 0.04996	KPSS = 0.04428	Biweekly (lower KPSS score)
Phillips-Perron Test	Z(α) = -85.679	Z(α) = -58.024	Weekly (more -ve score)

Although, post 2011 data **did not show stronger stationarity** than our original data full data

ACF, PACF, EACF Analysis:





- ❖ Similar components were concluded from long term trend.
- ❖ AR component decays a lot faster post 2011 so AR(1) should be sufficient

- For Non-Seasonal part
  - ACF (decay with no sharp cutoff) → **MA(1), MA(2)**
  - PACF (sharp cutoff at Lag 1)→ **AR(1)**
  - EACF → (o) dominates after 2, 2
  - AR(2): Lags beyond 2 become insignificant
    - → **AR(2) may be sufficient.**
  - MA: Lags beyond 2 become insignificant
    - → **MA(2) likely sufficient**.
- For Seasonal part
  - ACF (decays slowly with not sharp cutoff till lag 14)
    - → Differencing might be needed
  - PACF (Cuts off after lag 1)
    - → AR(1) may be adequate

```
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x x o x x x x x x o x  x  x  x
1 x x o x x x x x x o x  x  x  x
2 x x o o o o o x o o o  o  o  o
3 x o o o o o o o o o o  o  o  o
4 x x o o o o o o o o o  o  o  o
5 x x o o o o x o o o o  o  o  o
6 x o x o o o x o o o o  o  o  o
7 x x x o x x x o o o o  o  o  o
```

Candidate Modelling: (Based on ACF, PACF, EACF)

	Model <chr>	AIC <dbl>	BIC <dbl>
8	ARMA(2,0,2)	5064.824	5083.638
9	Auto.ARIMA	5064.824	5083.638
7	ARMA(2,0,1)	5085.821	5101.500
5	ARMA(1,0,2)	5114.198	5129.877
3	ARMA(1,0,0)	5127.421	5136.828
6	ARMA(2,0,0)	5124.838	5137.381
4	ARMA(1,0,1)	5127.187	5139.731
2	ARMA(0,0,2)	5135.915	5148.458
1	ARMA(0,0,1)	5179.251	5188.659

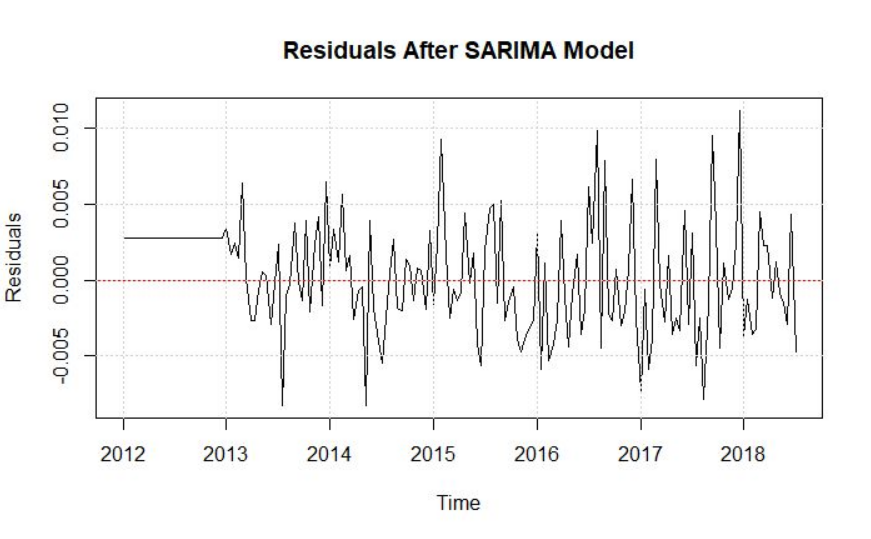
Best models: ARMA(2,0,2), Auto.ARMA(2,0,2), ARMA(2,0,1)

	Model <chr>	AIC <dbl>	BIC <dbl>
7	Auto.ARIMA	4294.834	4306.713
5	SARIMA(2,0,1)(1,1,2)	4290.744	4308.562
6	SARIMA(0,1,2)(0,0,1)	4288.874	4309.662
4	SARIMA(2,0,1)(1,1,1)	4296.010	4310.859
2	SARIMA(2,0,2)(0,1,1)	4291.909	4312.698
3	SARIMA(2,0,2)(1,1,1)	4290.867	4314.625
1	SARIMA(2,0,2)(1,0,2)	4296.876	4314.695

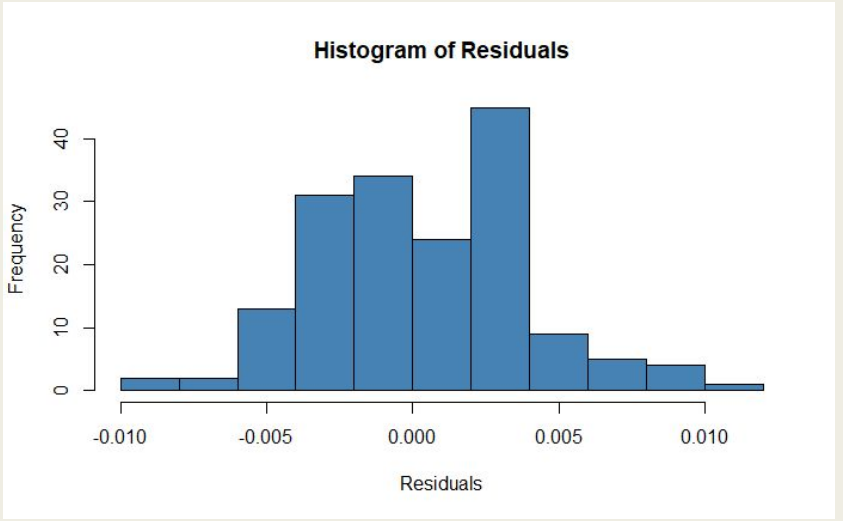
1. SARIMA(0,1,2)(0,0,1) → Lowest BIC
2. Auto.ARIMA(1,0,1)(0,1,1)
3. SARIMA(2,0,1)(1,1,2)

- Finalized model**: SARIMA(1,0,1)(0,1,1)**
  - Series already stationary.
  - No need for differencing

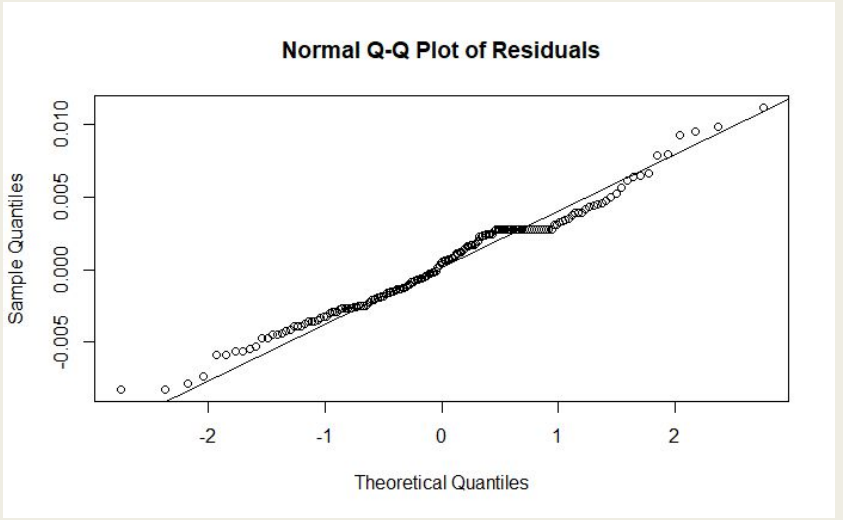
Model Diagnostic on Final Model: **SARIMA(1,0,1)(0,1,1)[26]**

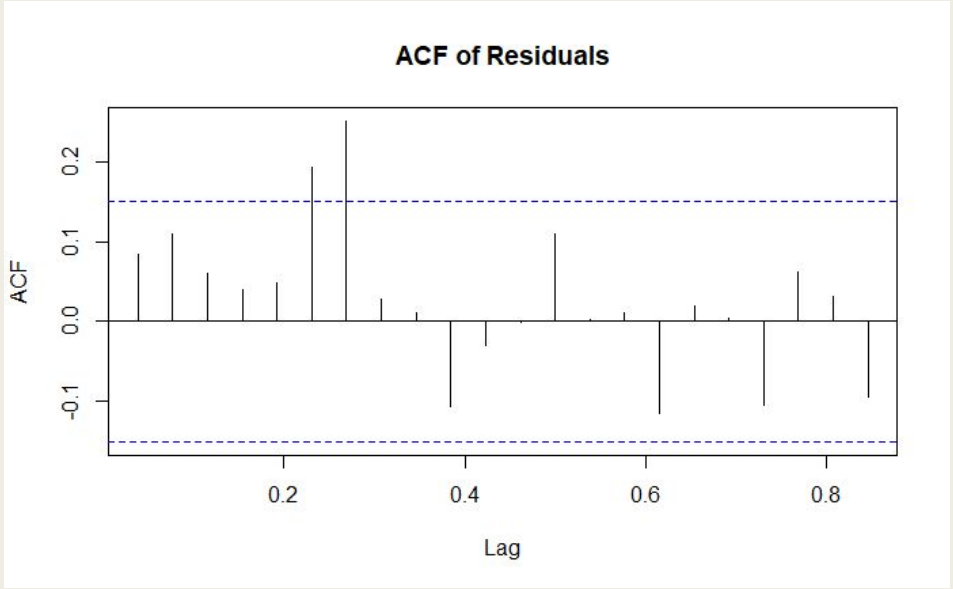


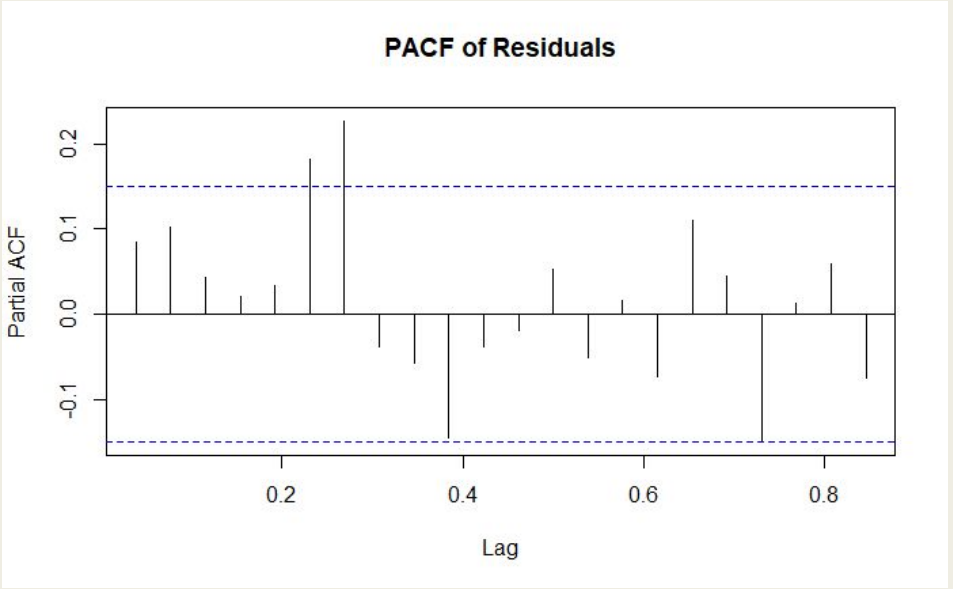
- The Energy consumption was in ten's of millions
- Performed **Double Log for transformation**

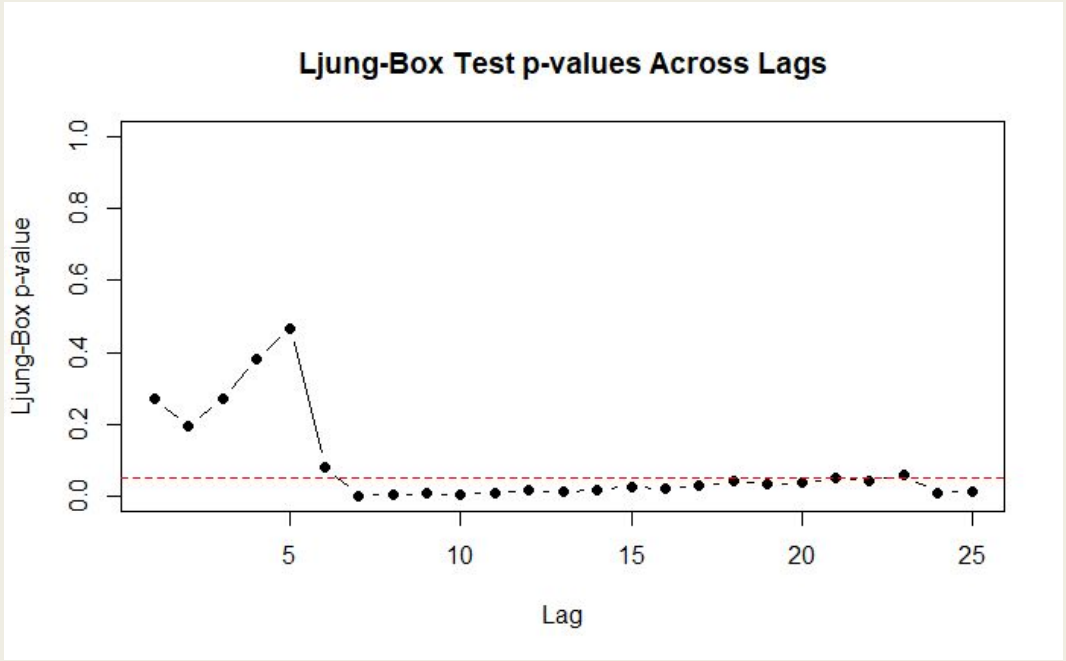


- **Residual Mean:** The residuals have a mean close to zero, indicating no significant bias in the model.
- **Homoscedasticity:** Residuals seem to have constant variance as well
- **Shappiro-Wilk test:** Confirms close to normality (p-value = 0.03193)



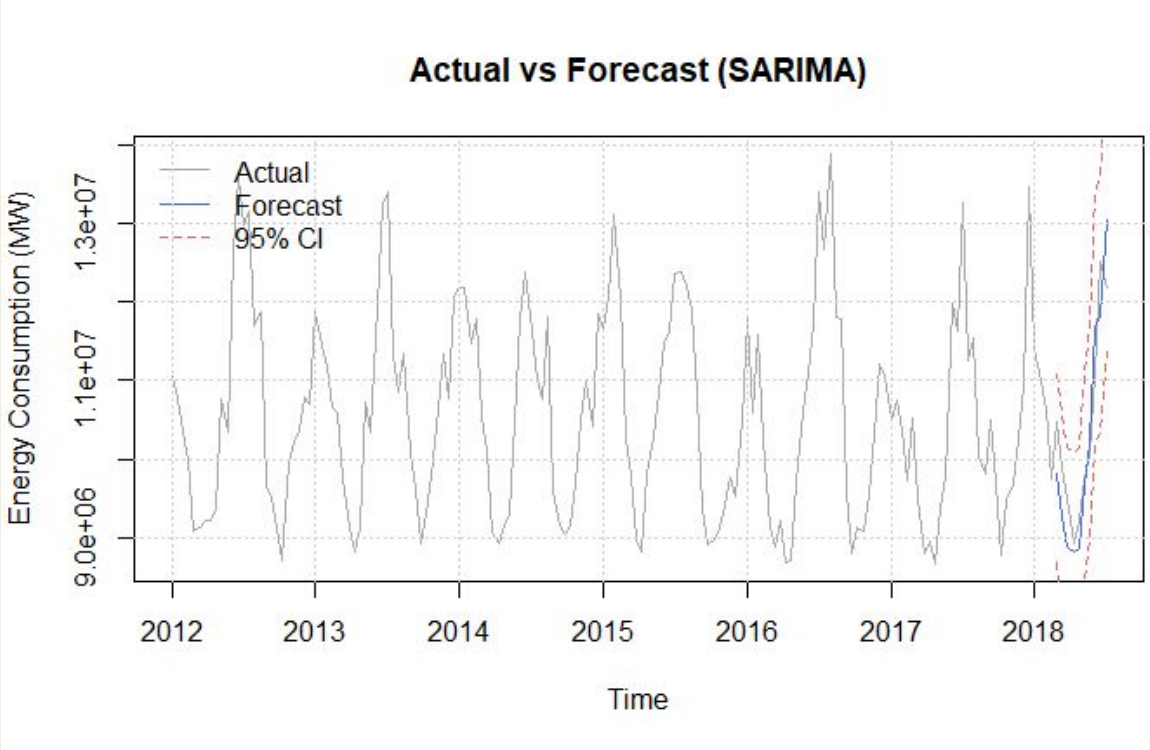






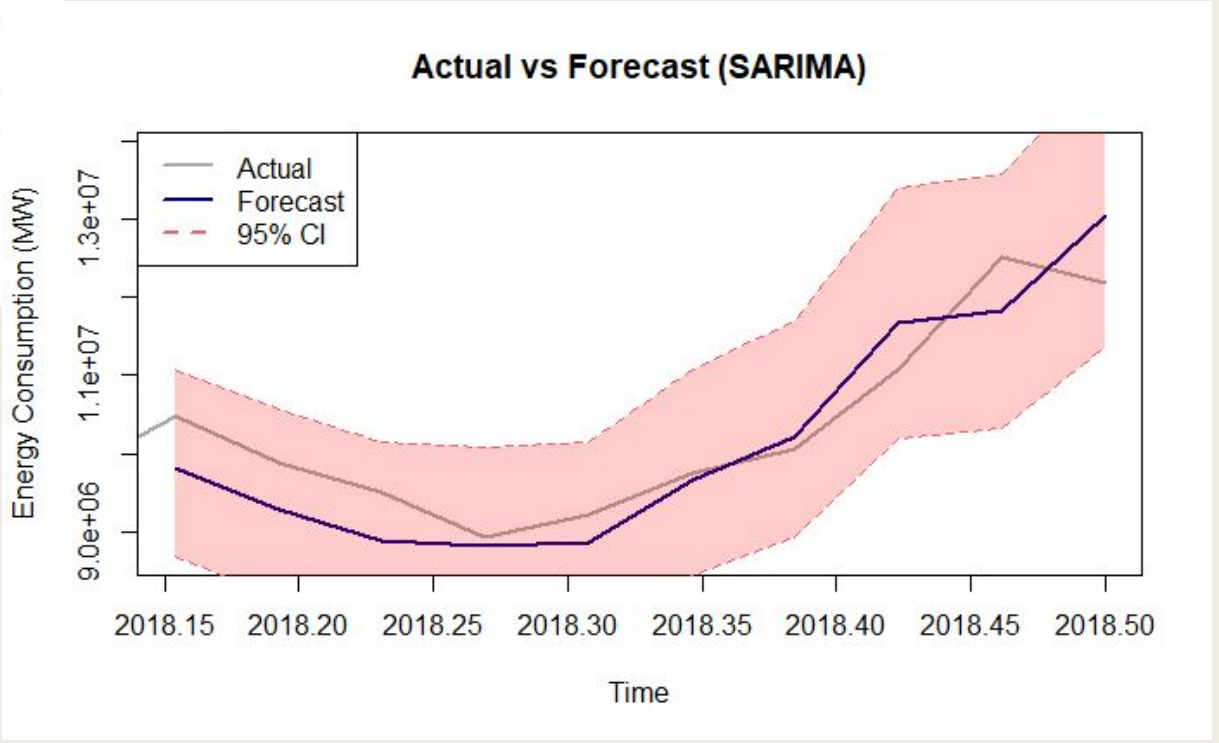
- **ACF and PACF plots** show no significant autocorrelation.
- **Box-Ljung test** confirms the independence of residuals **(p-value = 0.06157).**
- **Ljung-Box** p values across suggest independence as well.
- Although there are some minor correlations around lag 7, suggesting trend at ~**3.5 months**

Both Seasonal Tren



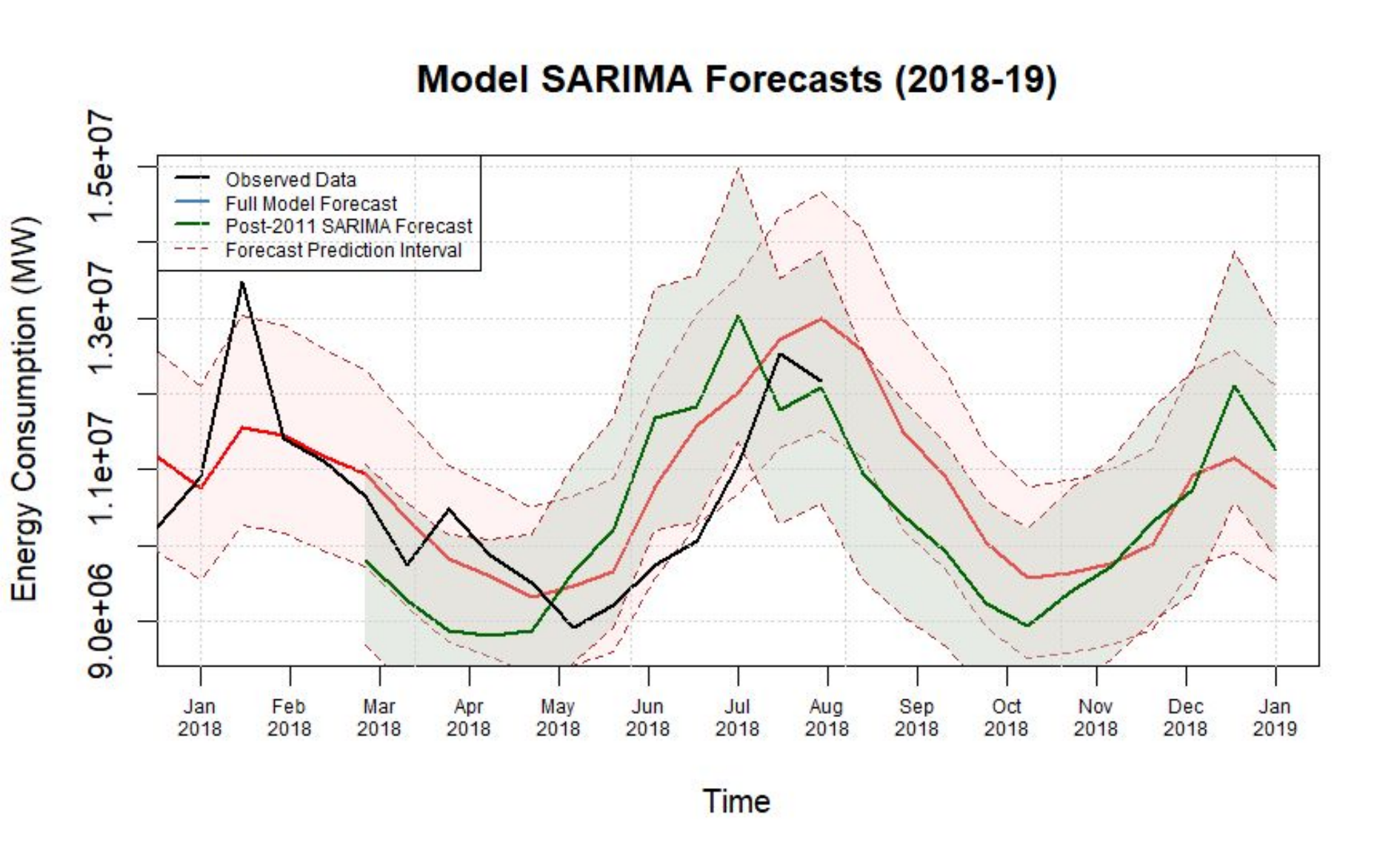
Metric <chr>	Value <dbl>
MAE	474926.190
RMSE	544114.244
MAPE (%)	4.451
R-squared	0.778

- Training on **~94%** data.
- **2012-2017 approx**
- **160 observations**
- **2018** had missing data
- **Tested** on that

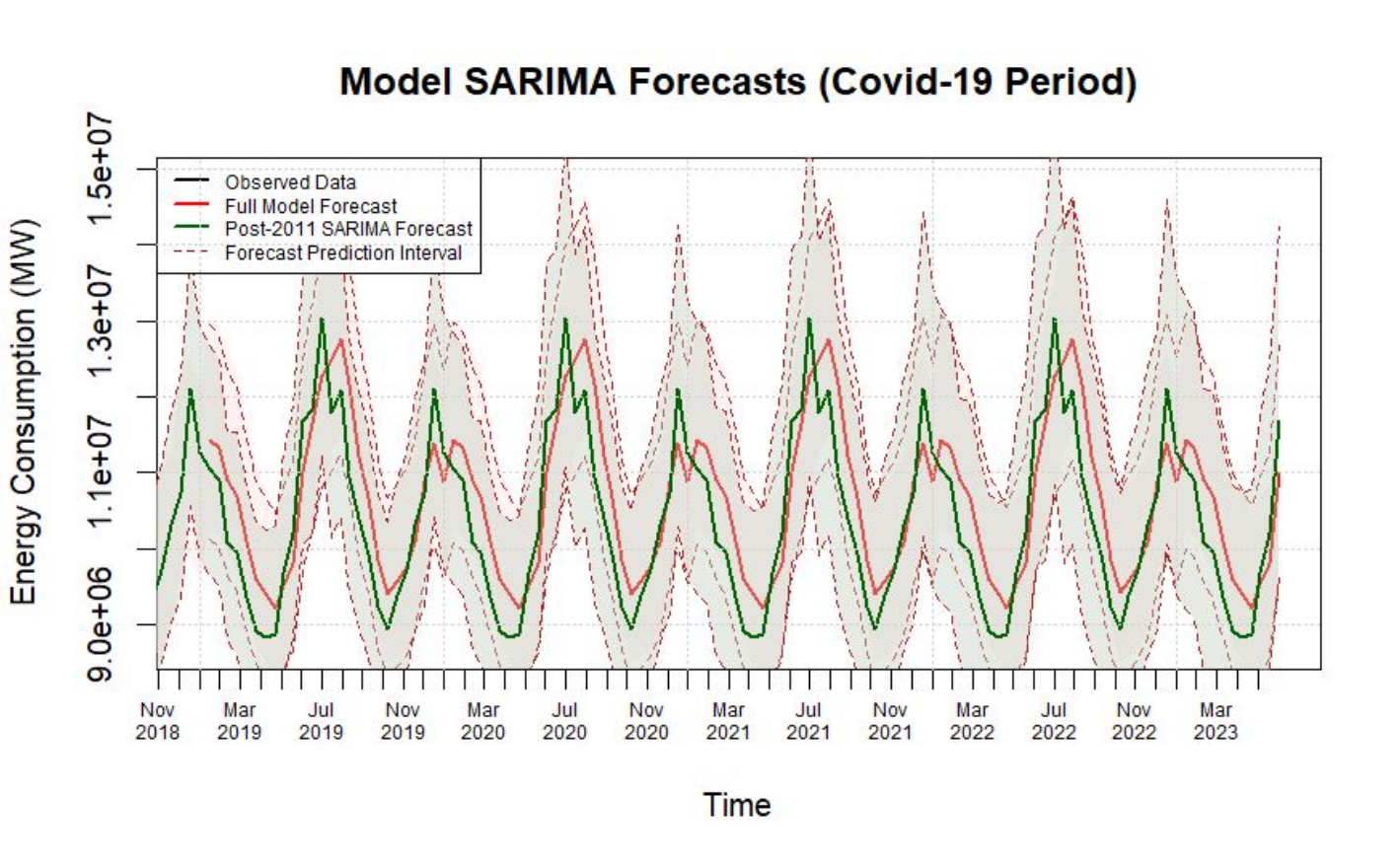


Visualizing both Seasonal Trends

Long-term trend still dominant



Forecasted Trend for Covid 19 - Era



- Post-2018 period includes the impact of COVID-19.
- Forecasting during the COVID era used weighted model averaging: **75% long-term, 25% short-term**.

Thank You