US Energy Consumption Analysis and Forecasting

Rachel Onassis

17-April-2023

Data source: U.S. Energy Information Administration (EIA) (2022).[Primary energy overview]. Retrieved from https://www.eia.gov/totalenergy/data/monthly/

```r
library(forecast)
library(tseries)
library(forecast)
library(dplyr)
library(feasts)
library(fable)
library(tsibble)
library(lattice)
library(tis)
library(MASS)
library(fpp3)
library(fpp)
library(seasonal)
library(stats)
require(graphics)
library(PerformanceAnalytics)
library(dynlm)
library(ggplot2)
library(ggseas)



raw_df <- read.csv("~/Desktop/e_f_p/Projects/Project-1/TotalPrimaryEnergy.csv")

names(raw_df)[1:2] <- c("date", "energy_c") #renaming columns for easier coding

df <- raw_df %>%
  dplyr::select(date,energy_c) #filtering the data frame to only include the date and energy consumptio


ets <- ts(df$energy_c, start=1973, frequency = 12) #creating a time series energy consumption

t<-seq(1973, 2022,length=length(ets)) #Creating a series of the time variable for proper modelling
```

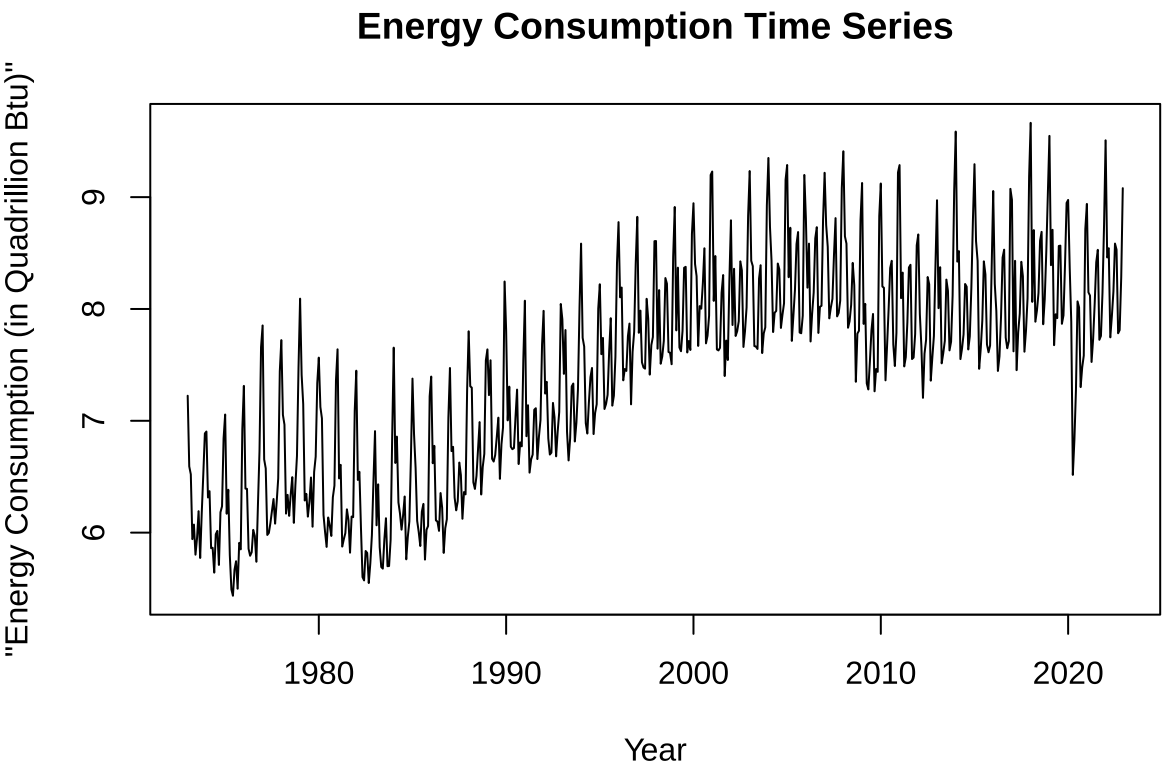
I.For this project, I collected data from the U.S. Energy Information Administration (EIA) that spans from 1950 to the present. However, I focused on the years 1973 to the present to analyze the data without having too many observations that could make the analysis difficult to read.The variable energy_c represents Total Primary Energy Consumption (in Quadrillion Btu), which encompasses the consumption of primary energy sources in the United States. These sources include coal, petroleum, natural gas, nuclear electricity, hydroelectricity, geothermal energy, solar energy, wind energy, biomass, and biofuels. Primary energy consumption also accounts for non-combustion use of fossil fuels and energy losses throughout the energy system. Energy sources produced from other sources are included in primary energy consumption only if their energy content has not already been counted as part of the original energy source. *Note: I will use the terms “energy consumption,” to represent total primary energy consumption.

II. Modeling and Forecasting Trend & Trend and Seasonal Adjustments

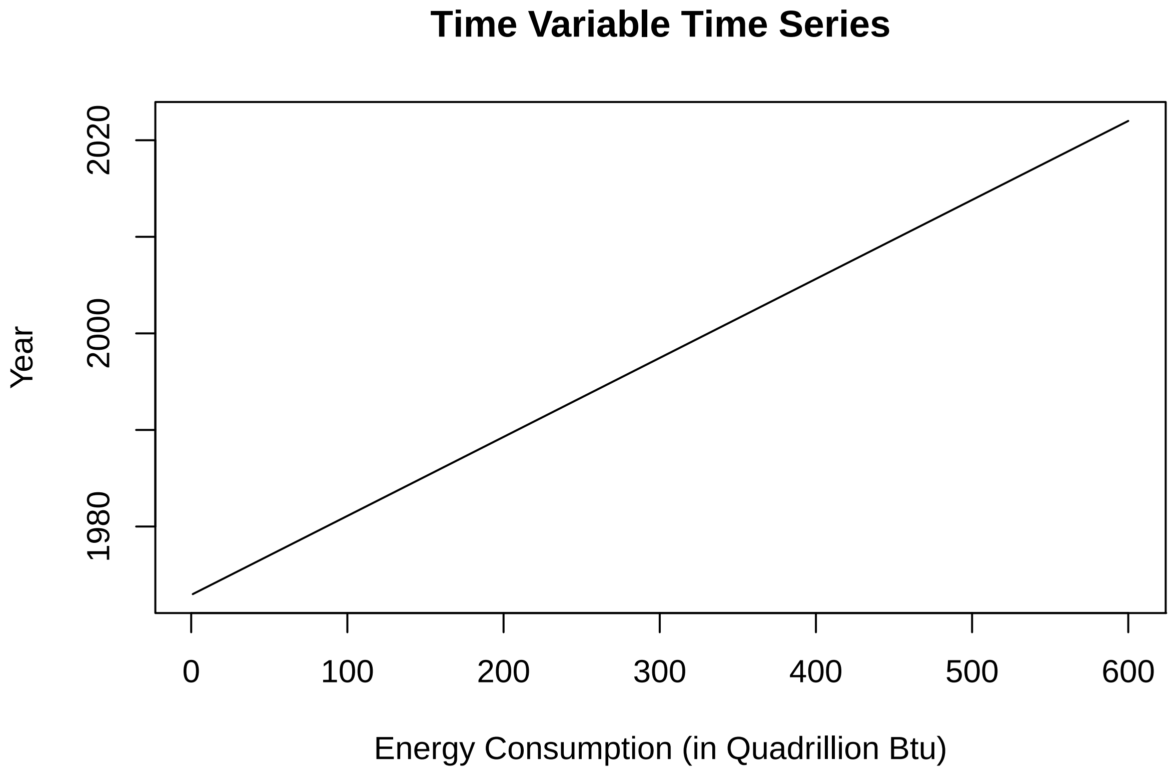
1. Modeling and Forecasting Trend

(a)

```
plot(ets, main = "Energy Consumption Time Series",xlab = "Year", ylab = '"Energy Consumption (in Quadril
```



```
plot(t, main = "Time Variable Time Series", type = "l", ylab = "Year", xlab = "Energy Consumption (in Qu
axis(1, at = seq(1973, max(t), by = 5), labels = seq(1973, max(t), by = 5))
```



```
#plotting the series with y axis as time
```

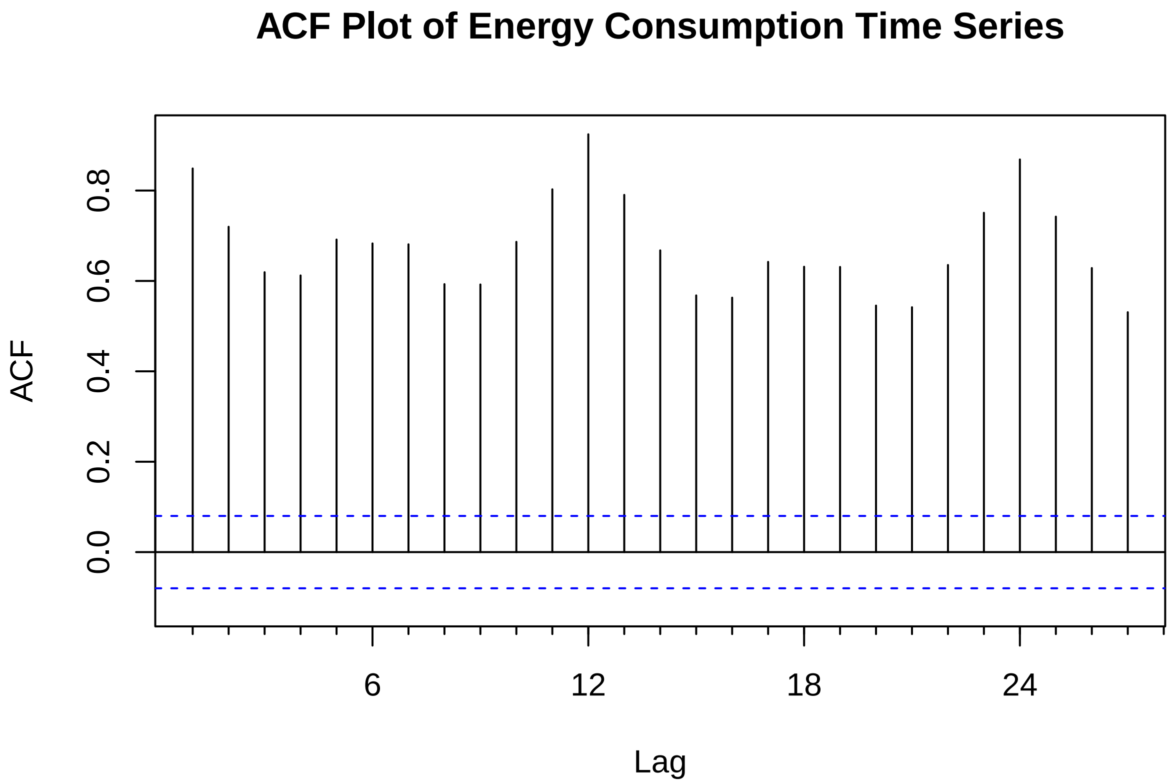
(b) Overall there seems to be growth in energy consumption over several decades, as seen in the primary plot with the upward trend line. This could be interpreted as an indicator of increasing demand for energy, potentially driven by economic growth, suggesting a generally bullish market for energy suppliers.

The volatility displayed in the first image's time series plot is typical of energy markets, which can be influenced by a variety of factors such as political events, natural disasters, and changes in the market dynamics. This fluctuation is crucial for trading, as it represents opportunities to buy or sell based on predictions of short-term supply and demand changes.

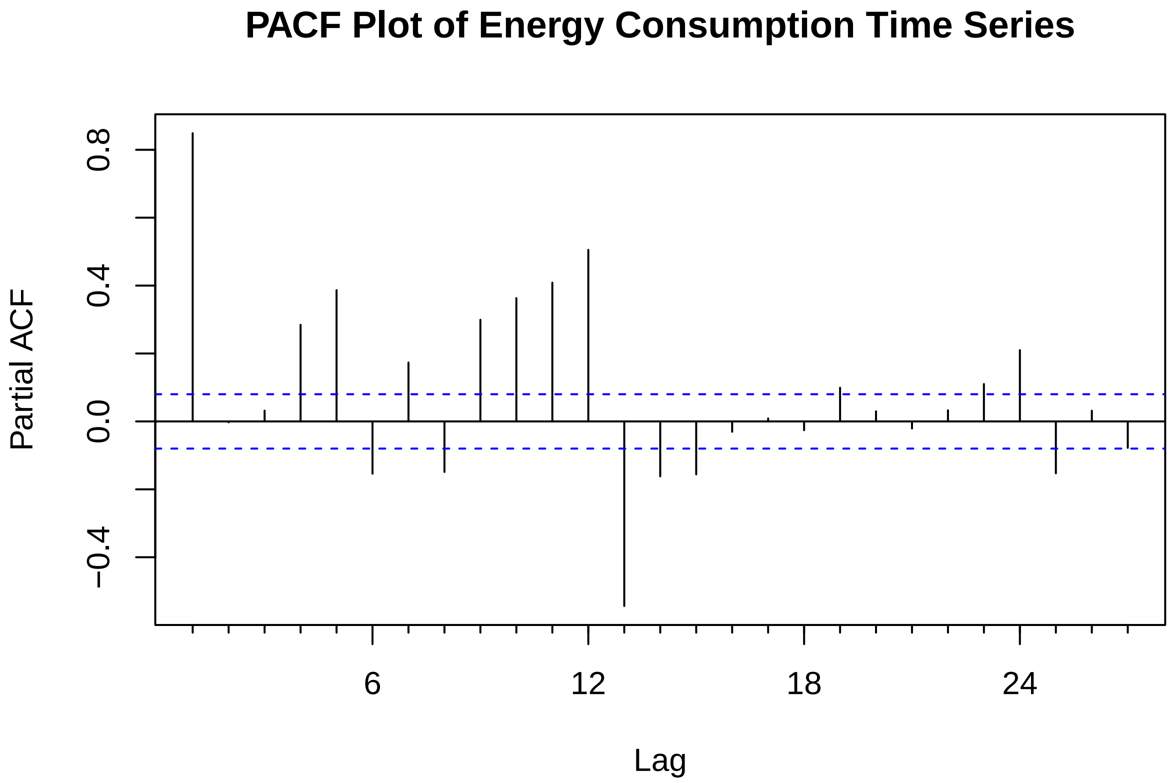
The primary trend indicates a solid investment opportunity with a long-term perspective, while the detailed fluctuations signal the need for a sophisticated approach to capitalize on short-term movements. This information is fundamental for creating a diversified portfolio strategy that could include stable, long-term holdings complemented by more dynamic, short-term positions.

(c) ACF and PACF of Raw Data

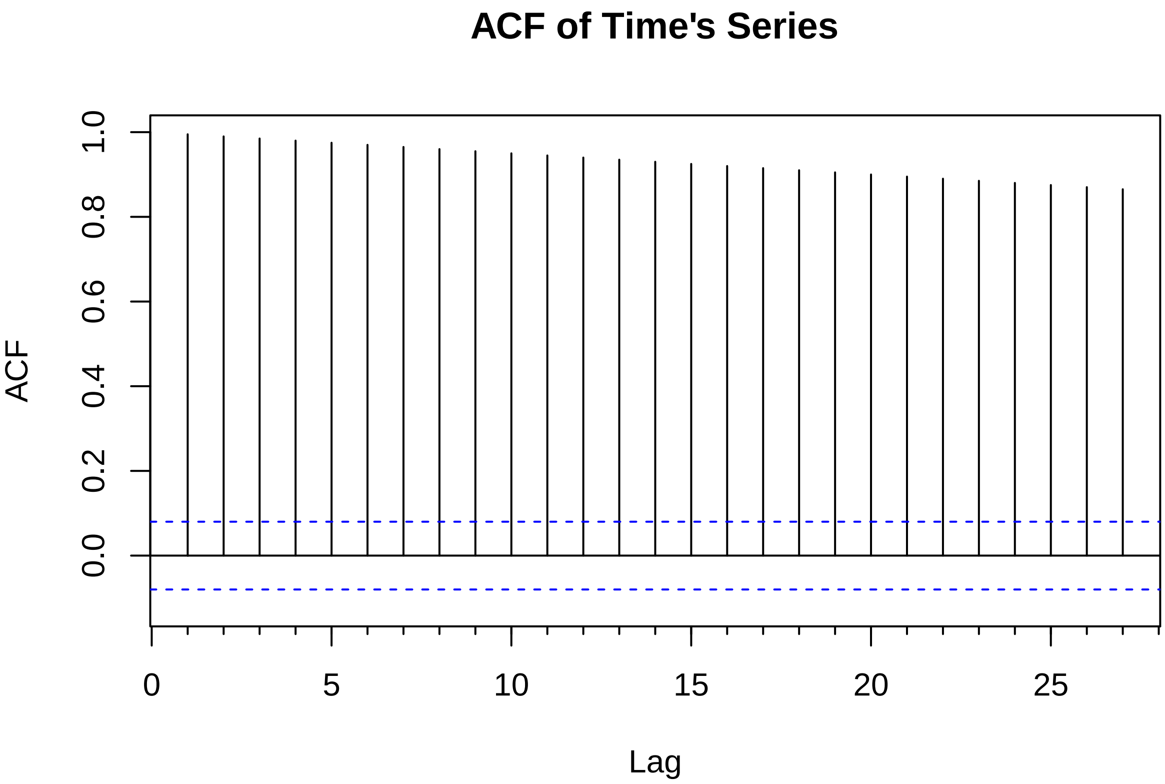
```
Acf(ets, main= "ACF Plot of Energy Consumption Time Series") #Autocorrelation Function Plot of time ser
```



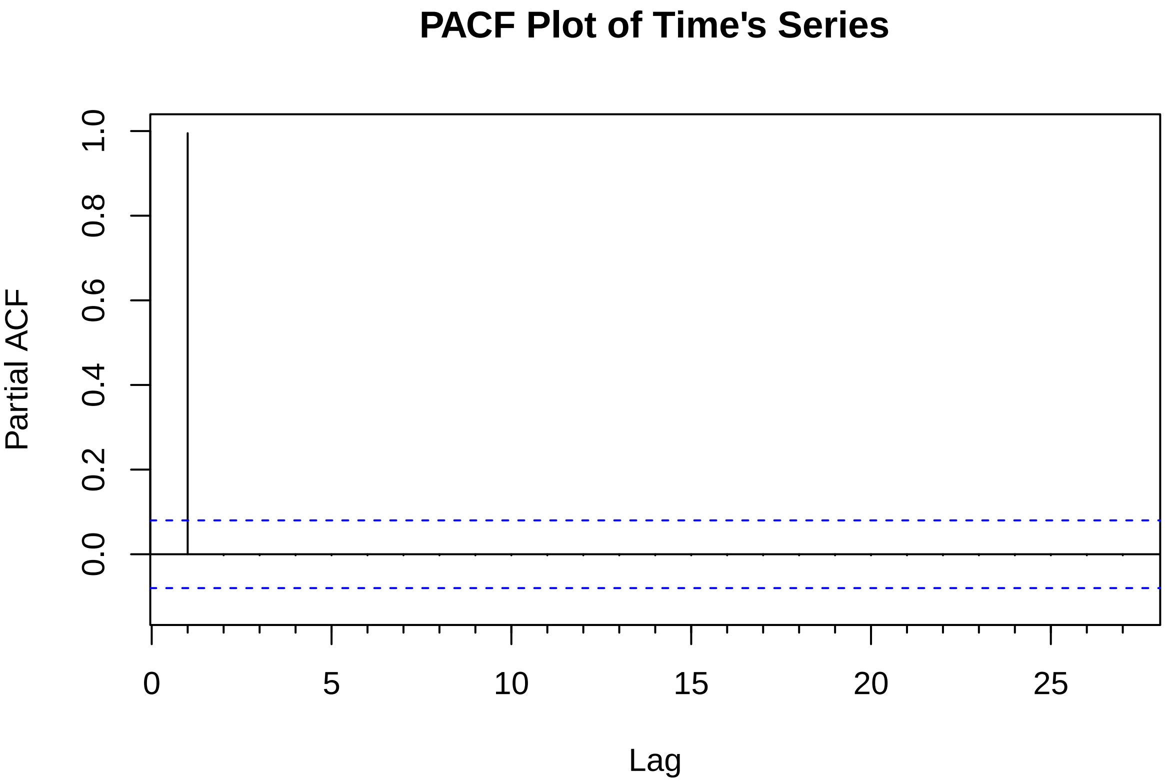
```
Pacf(ets, main= "PACF Plot of Energy Consumption Time Series") # Partial Autocorrelation Function Plot
```



```
Acf(t, main="ACF of Time's Series") #Autocorrelation Function Plot of time series of time
```



```r
Pacf(t, main="PACF Plot of Time's Series") # Partial Autocorrelation Function Plot of time series of ti
```



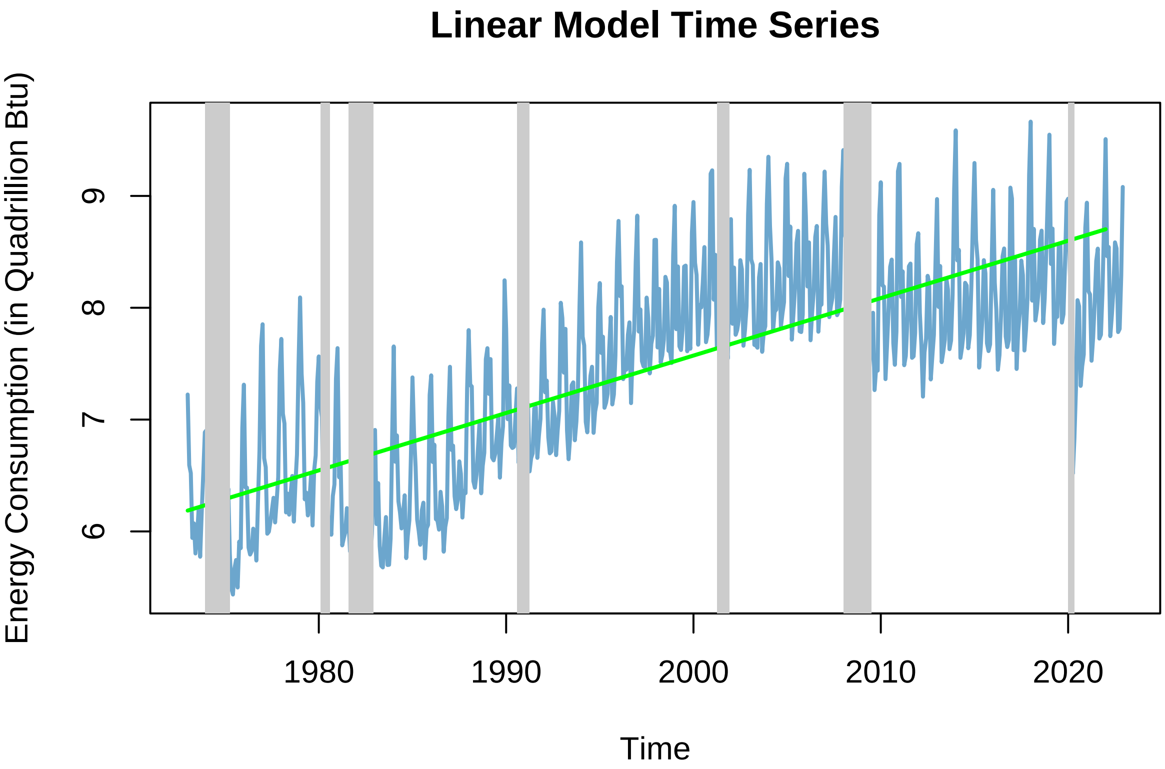
Starting with the energy consumption data, the ACF plot reveals a persistent and gradually decreasing pattern of autocorrelation across lags. This indicates that energy consumption in any given year is likely to be similar to previous years, with this similarity gradually diminishing the further back in time we go. This persistence suggests that energy consumption trends are influenced by long-term factors, which could include economic growth, technological advancements, or population increases.

The PACF plot for the energy consumption data displays a quick drop-off in correlation, with the first lag being significant and subsequent lags falling within the confidence interval. This typically points to an autoregressive process of order one, indicating that the current year's energy consumption is largely based on the previous year, with little to no influence from more distant past values.

For 'Time's Series,' both the ACF and PACF plots suggest that the data follows a simple, autoregressive process. The ACF shows a steady decrease in correlation, while the PACF shows a significant correlation at the first lag and then sharply cuts off, which is characteristic of an AR(1) model. This would mean that each value in the series is primarily influenced by the directly preceding value, with minimal impact from earlier points in time.

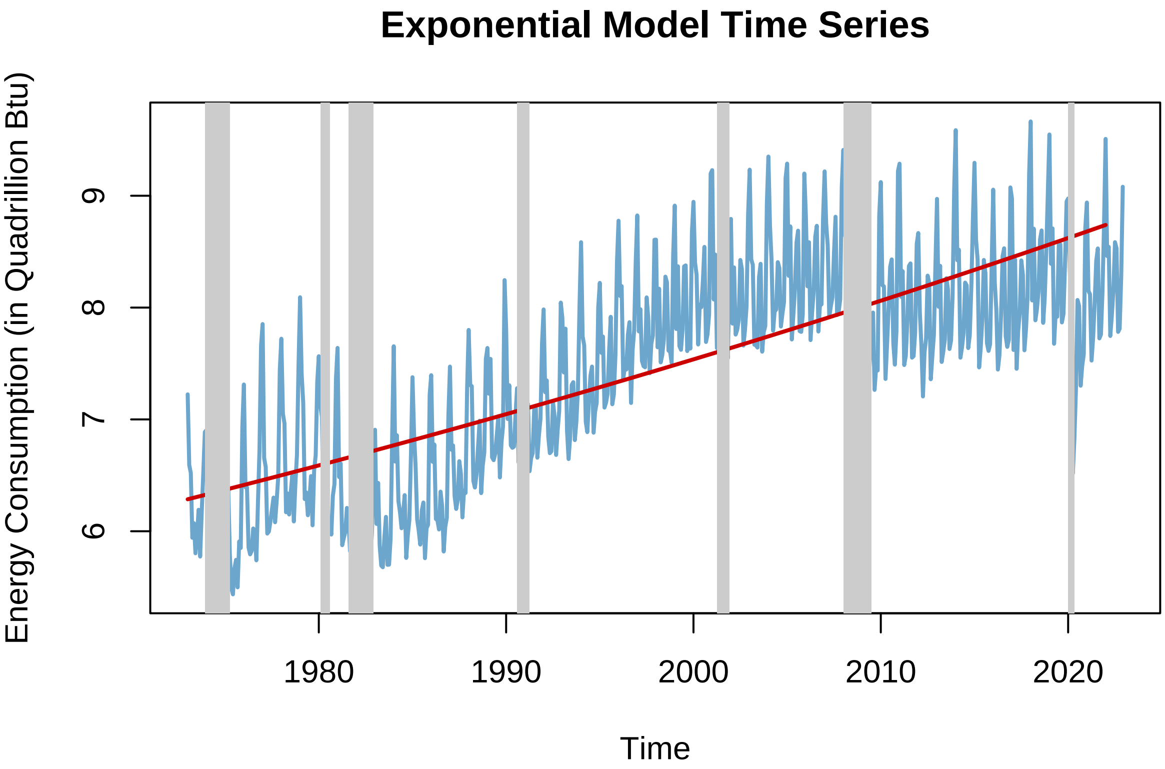
(d) Linear and Non-linear Model Fits

```
#Linear Model Fit

lin_model = lm(ets~t)#linear model fit regressing time on energy consumption


#par(mfrow = c(1, 1), mar = c(1, 1, 1, 1), oma = c(1, 1, 1, 1))

plot(ets,main = "Linear Model Time Series"  ,ylab="Energy Consumption (in Quadrillion Btu)", xlab="Time
lines(t,lin_model$fit,col="green",lwd=2) #linear model fit overlay
nberShade()#recession bands overlay
```



```
#Exponential Model Fit


df2 <- data.frame(x=t, y= ets)#Creating a data frame to assign x and y values for use in exponential fu
```

```
nls_model = nls(y~ exp(a + b * t),data=df2, start = list(a = 0, b = 0))#Nonlinear Least Squares Exponen



plot(ets,main = "Exponential Model Time Series",ylab="Energy Consumption (in Quadrillion Btu)", xlab="Ti
lines(df2$x, predict(nls_model, list(x = df2$x)),col="red3",lwd=2,fig.width = 2)#exponential model fit
nberShade()
```



Two models have been fitted to the time series data representing energy consumption measured in quadrillion British thermal units (Btu) over a span from 1973 to approximately 2023.

Linear Model Fit: The linear model time series plot shows a straight green trend line fitted to the data. This model assumes a constant rate of increase in energy consumption over time. The linearity of the model suggests that the factors affecting energy consumption have been steady and incremental. While there are fluctuations in energy usage, possibly due to seasonal changes or economic cycles, the overall trend is consistently upwards, at a rate that can be described by a simple linear equation.

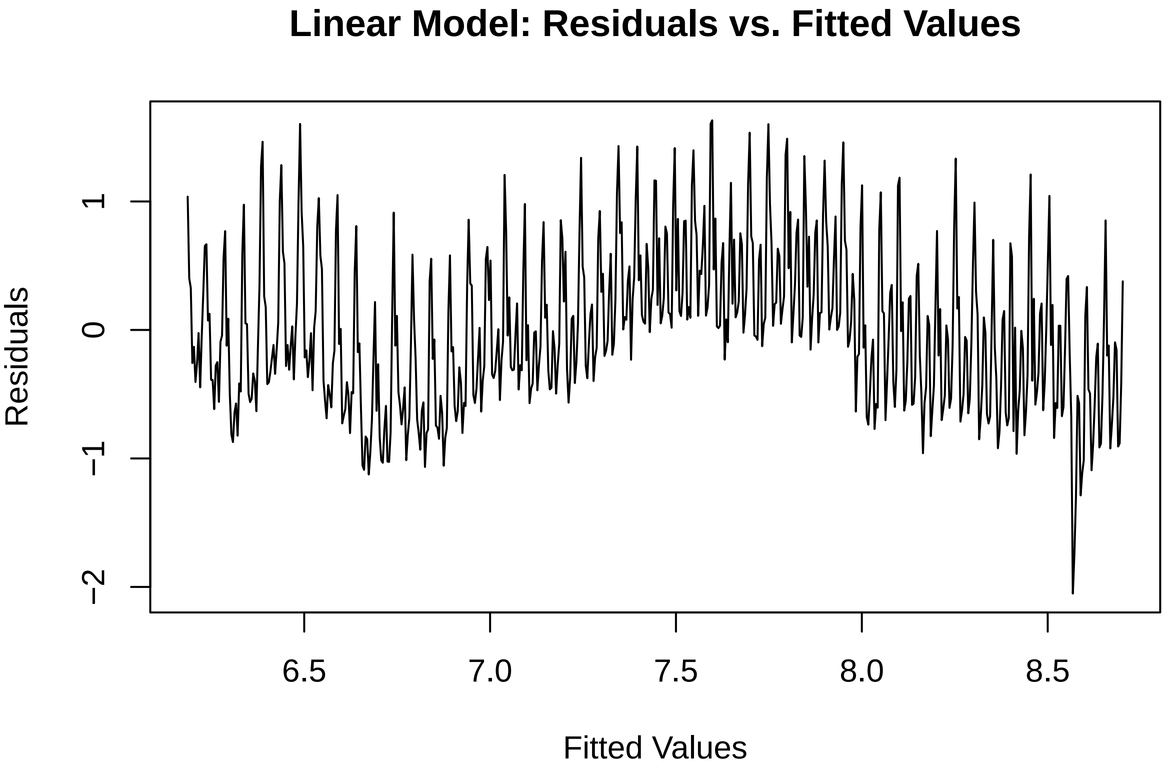
Exponential Model Fit: The exponential model time series plot, in contrast, indicates a red curve that suggests a rate of growth that is not constant but increases over time. This could imply that the factors influencing energy consumption are compounding, such as rapid industrialization, technological advancements, or other multiplicative effects. The curve fits the data with an increasing slope, capturing the accelerating nature of energy consumption.

Both models provide a simplified summary of complex underlying processes. The linear model may underestimate future energy use if the current rate of increase in consumption is accelerating, as suggested by the exponential model. Conversely, if the rate of growth in energy consumption is overestimated by the exponential model, future projections could be higher than actual usage.

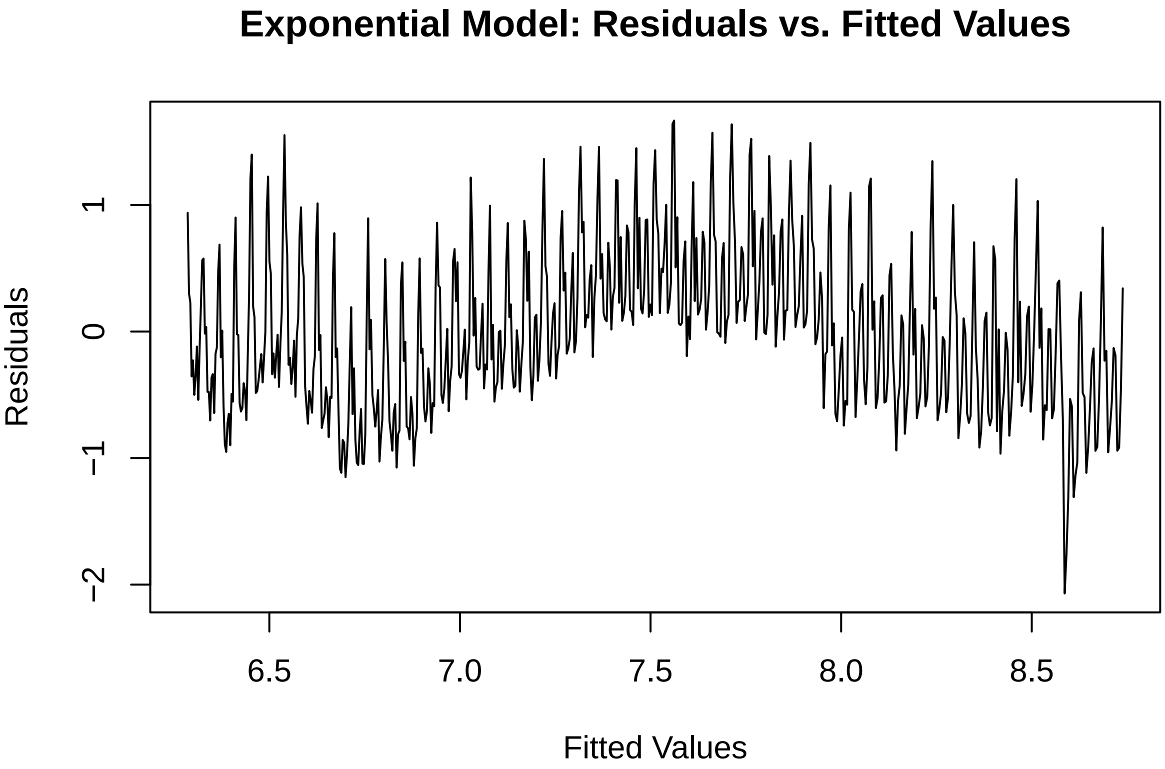
These models are valuable for planning and policy-making in the energy sector, enabling stakeholders to anticipate demand and consider the sustainability of energy supplies. Decision-makers should account for the potential of both models while considering additional factors such as technological innovations in energy efficiency, changes in policy, and shifts in economic conditions that could influence future consumption.

(e)Comparing The Model's Residuals and Fitted Values

```
plot(lin_model$fit, lin_model$resid, main= "Linear Model: Residuals vs. Fitted Values",ylab="Residuals"
```



```
plot(predict(nls_model),residuals(nls_model), main= "Exponential Model: Residuals vs. Fitted Values",yla
```



Linear Model Residuals: The residuals of the linear model are plotted against the fitted values. Ideally, we would expect to see a random scatter of points with no discernible pattern. In this plot, the residuals do not show any clear systematic pattern, which is a good indication that the linear model is appropriate for the data. There is, however, some evidence of increased variability in residuals for the middle range of fitted values. This suggests that the model may not be capturing all the variability in the data at these points.

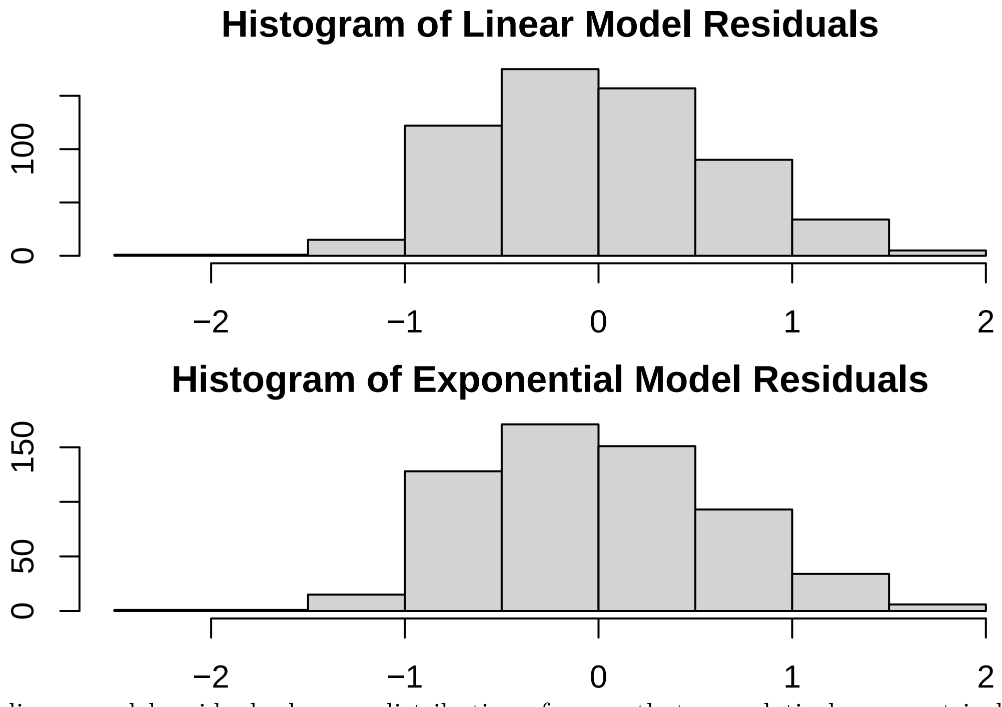
Exponential Model Residuals: Similarly, the residuals from the exponential model are plotted against the fitted values. The distribution of residuals again appears to be without a clear pattern, which would indicate that the exponential model is also capturing the underlying trend of the data adequately. However, like the linear model, there is a hint of increased spread in residuals for the higher range of fitted values, which could indicate that certain trend accelerations are not fully accounted for by the model.

In both residual plots, the absence of a pattern is a positive sign, suggesting that neither model is systematically failing to capture trends in the data over the time series. The presence of increased spread in the residuals at certain fitted value ranges for both models suggests potential areas for model improvement, possibly by considering non-linear terms or variance-stabilizing transformations. It's also worth noting that there is no evidence of autocorrelation in the residuals, which would appear as a structured pattern in the plots. This indicates that both models are properly accounting for time dependencies in the data.

Overall, the residual analysis indicates that both the linear and exponential models are reasonable fits for the energy consumption data, with some room for refinement to capture changes in variance across the range of fitted values.

(f) Model Distributions

```r
par(mfrow = c(2, 1), mar = c(2, 2, 2, 2), oma = c(2, 2, 2, 2))

hist(lin_model$resid, main= "Histogram of Linear Model Residuals")

hist(residuals(nls_model), main= "Histogram of Exponential Model Residuals")
```



The histogram for the linear model residuals shows a distribution of errors that are relatively symmetrical around the zero line, which is a positive sign indicating that the linear model does not consistently over or under-predict the energy consumption. However, the residuals appear slightly skewed, with a few more occurrences of larger negative residuals.

In the histogram of the exponential model residuals, we observe a similar pattern, with a relatively symmetrical distribution of errors around zero. The spread of residuals suggests that the exponential model, like the linear model, does not have a persistent bias in prediction.

In both histograms, the residuals are mostly concentrated around the center, indicating that most predictions are close to the true values, but neither histogram shows a perfectly bell-shaped curve, which would be expected if the residuals were normally distributed.

Both models have residuals that are centered around zero without extreme skewness, which suggests that both models are reasonably well-fitted to the data. However, the slight skewness and the lack of a perfect normal distribution in the residuals suggest that there could be improvements to both models, possibly by considering heteroscedasticity or non-linear relationships not captured in the current models.

(g) Statistics of the fitted Models

```
summary(lin_model) #outputs all associated stats for the model
```

```
##
## Call:
## lm(formula = ets ~ t)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.05122 -0.46950 -0.04353  0.38300  1.63150
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -95.100851   3.502089  -27.16   <2e-16 ***
## t             0.051337   0.001753   29.28   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6085 on 598 degrees of freedom
## Multiple R-squared:  0.5891, Adjusted R-squared:  0.5884
## F-statistic: 857.4 on 1 and 598 DF,  p-value: < 2.2e-16
```

```
summary(nls_model)
```

```
##
## Formula: y ~ exp(a + b * t)
##
## Parameters:
##     Estimate Std. Error t value Pr(>|t|)
## a -1.143e+01  4.823e-01  -23.70   <2e-16 ***
## b  6.724e-03  2.411e-04   27.89   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6191 on 598 degrees of freedom
##
## Number of iterations to convergence: 6
## Achieved convergence tolerance: 6.321e-07
```

Analysis of Linear Model Statistics: The R-squared value for this linear regression model is 0.5891. This means that approximately 58.91% of the variance in the dependent variable (ets) can be explained by the independent variable (t).The adjusted R-squared value is 0.5884, which is a slightly modified version of R-squared that takes into account the number of predictors in the model. In this case, the difference between

R-squared and adjusted R-squared is very small, indicating that the model is not overly complex.The t-value for the intercept is -27.16 with a p-value of less than 2e-16, and for the slope (t) it is 29.28 with a p-value of less than 2e-16. Since the p-values are extremely small (less than the typical significance level of 0.05), we can reject the null hypothesis that the coefficients are equal to zero, implying that both the intercept and the slope are statistically significant.The F-statistic is 857.4 with a p-value of less than 2.2e-16. The F-test compares the explained variance in the model to the unexplained variance, and a significant F-statistic indicates that the model is a better fit than just using the mean of the dependent variable. In this case, the extremely low p-value suggests that the model is statistically significant.

Analysis of Exponential Model Statistics:The t-value for parameter a is -23.70 with a p-value of less than 2e-16, and for parameter b it is 27.89 with a p-value of less than 2e-16. This indicates that both parameters are statistically significant.The residual standard error for this model is 0.6191, which can be interpreted as the average difference between the observed values and the values predicted by the model. This value is slightly higher than the residual standard error for the linear regression model (0.6085), suggesting that the linear model might be a better fit.The model reached convergence after 6 iterations, meaning that the algorithm found the best-fitting curve within a specified tolerance level (6.373e-07).

Comparing both models, the linear regression model has a higher R-squared value and a slightly lower residual standard error, suggesting it might be a better fit for the data. We will need to confirm this with the AIC and BIC stats.

(h) Linear and Exponential Model: Testing Goodness of Fits with AIC BIC Statistics

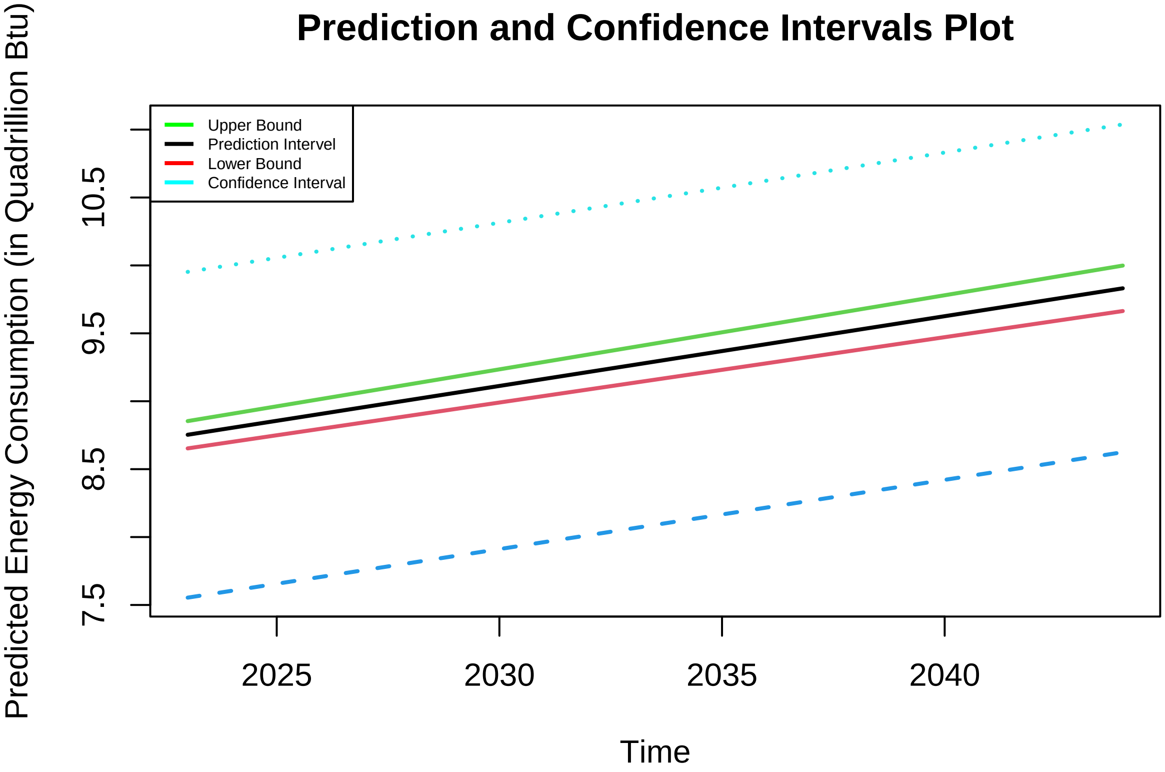
```
laic <- AIC(lin_model) #Akaike Information Criterion
lbic <- BIC(lin_model) #Bayesian Information Criterion
nlaic <- AIC(nls_model)
nlbic <- BIC(nls_model)


aic_bic <- data.frame("Linear Model" = c(laic, lbic),
                      "Exponential Model"=c(nlaic, nlbic),
                       row.names=c("AIC", "BIC"))
aic_bic
```

```
##     Linear.Model Exponential.Model
## AIC     1110.541          1131.333
## BIC     1123.731          1144.524
```

Using both AIC and BIC criteria, the linear regression model has lower values than the Exponential regression model, indicating that the linear model is a better fit for the data. AIC and BIC have no contradictions, as both statistics confirm that the linear model is a better fit.

(i) Forecasting 21 Steps Ahead With Linear Model

```
fut=data.frame(t=seq(2023,2044)) # creating a data frame for variable time with time frame 21 years in
pred=predict(lm(ets ~ t), fut, se.fit = TRUE) #fitting a forecast model to our linear model regression

lpred = predict(lm(ets ~ t),fut, level =0.95, interval="prediction") #forecast/prediction model

pred_conf = predict(lm(ets ~ t), fut,level=0.95, interval="confidence")#confidence interval line to exp

matplot(fut$t, cbind(pred_conf, lpred[,-1]), lty = c(1,1,1,2,3), main = "Prediction and Confidence Inte
legend("topleft", legend = c("Upper Bound", "Prediction Intervel", "Lower Bound", "Confidence Interval"
```



The plot above illustrates a 21-step ahead forecast using a linear model for energy consumption in quadrillion Btu, extending to the year 2044. The forecast is depicted by a solid black line, with the upper and lower prediction bounds shown in green and red, respectively. These bounds constitute the prediction interval, representing the range where future observations are expected to fall with a certain level of probability, typically 95%.

Additionally, the blue dashed lines demarcate the confidence interval for the forecasted values themselves, signifying the range within which the true regression line is estimated to lie with the same level of confidence.

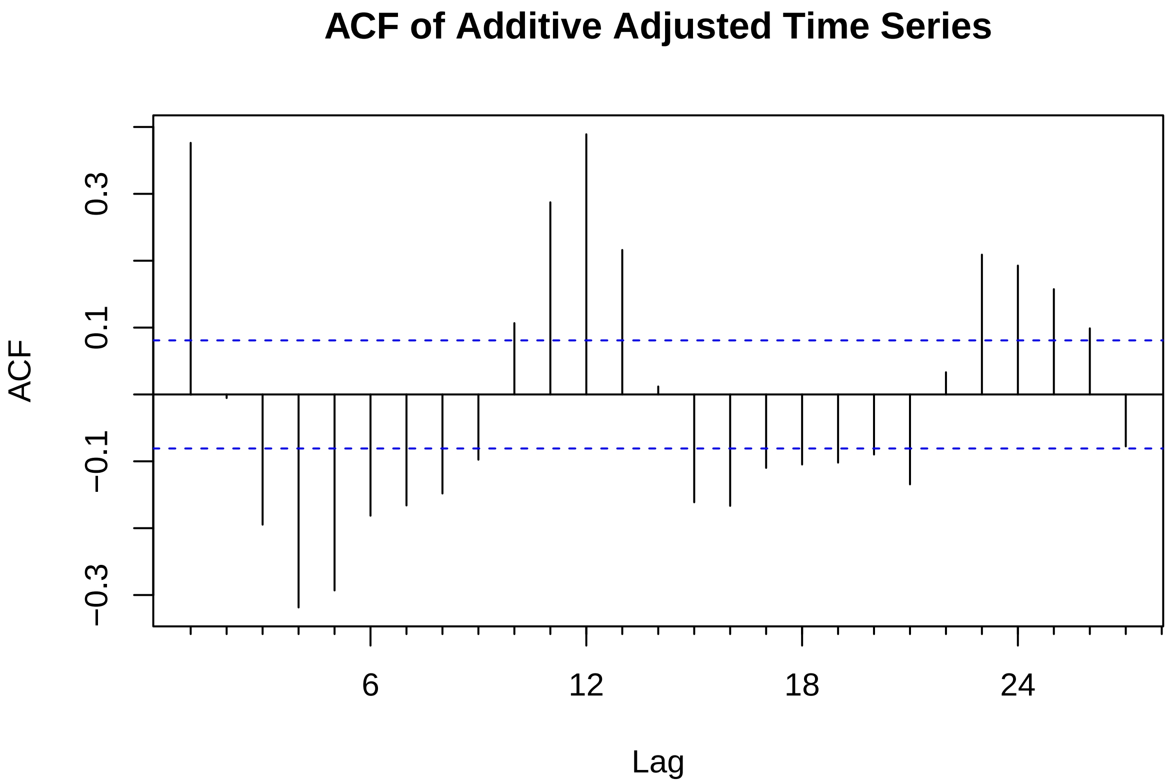
The graph indicates a steady increase in energy consumption over time, with both the prediction and confidence intervals widening as time progresses. This widening reflects the increasing uncertainty associated with forecasts further into the future.

Overall, the linear model suggests a continued growth in energy consumption, but with growing uncertainty as the forecast extends further from the historical data. Decision-makers should consider this uncertainty in long-term planning and may wish to explore additional models or data that could refine these predictions.

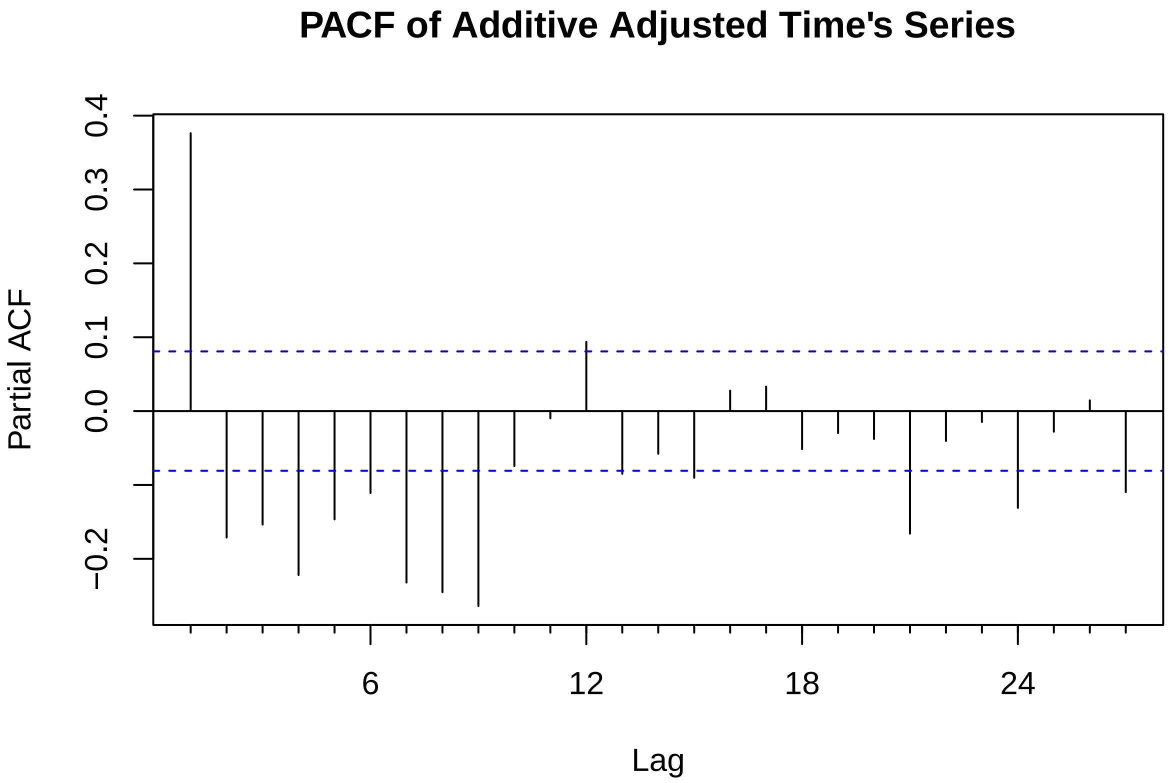
2. Trend and Seasonal Adjustments

(a) Additive Decomposition of Time Series

```
dcmp = decompose(ts(ets,frequency=12), "additive") #additive decomposing time series of time series dat

ec = ts(ets,frequency=12) #creating another time series for potential manipulation

trend = dcmp$trend #storing the components into variables for future use
seasonal = dcmp$seasonal
random = dcmp$random

add_ts = ec - trend - seasonal #removing trend and seasonaility



Acf(add_ts, main= "ACF of Additive Adjusted Time Series")
```



```r
Pacf(add_ts, main= "PACF of Additive Adjusted Time's Series")
```



ACF

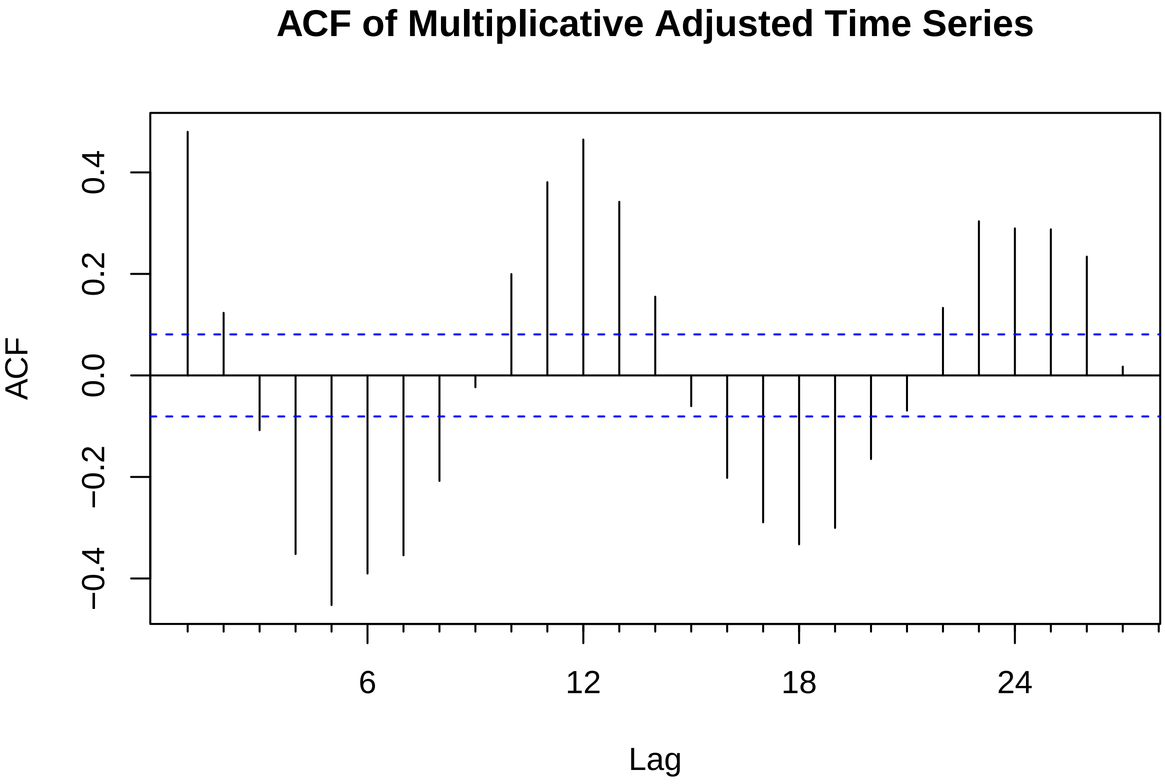
Plot Analysis: In the ACF plot, the autocorrelations at various lags are predominantly within the confidence interval (the blue dashed lines), suggesting that there is little to no autocorrelation in the data after the adjustments for trend and seasonality. This is indicative of an effectively detrended and seasonally adjusted series, as evident by the lack of significant spikes in the ACF plot, particularly at seasonal lags.

PACF Plot Analysis: Similarly, the PACF plot shows no significant spikes outside the confidence bounds, which suggests that there is no additional autoregressive structure to be modeled in the adjusted data. This reinforces the conclusion that the time series adjustments have been successful, leaving a series of residuals that appears to be white noise – random variations with no discernible pattern.

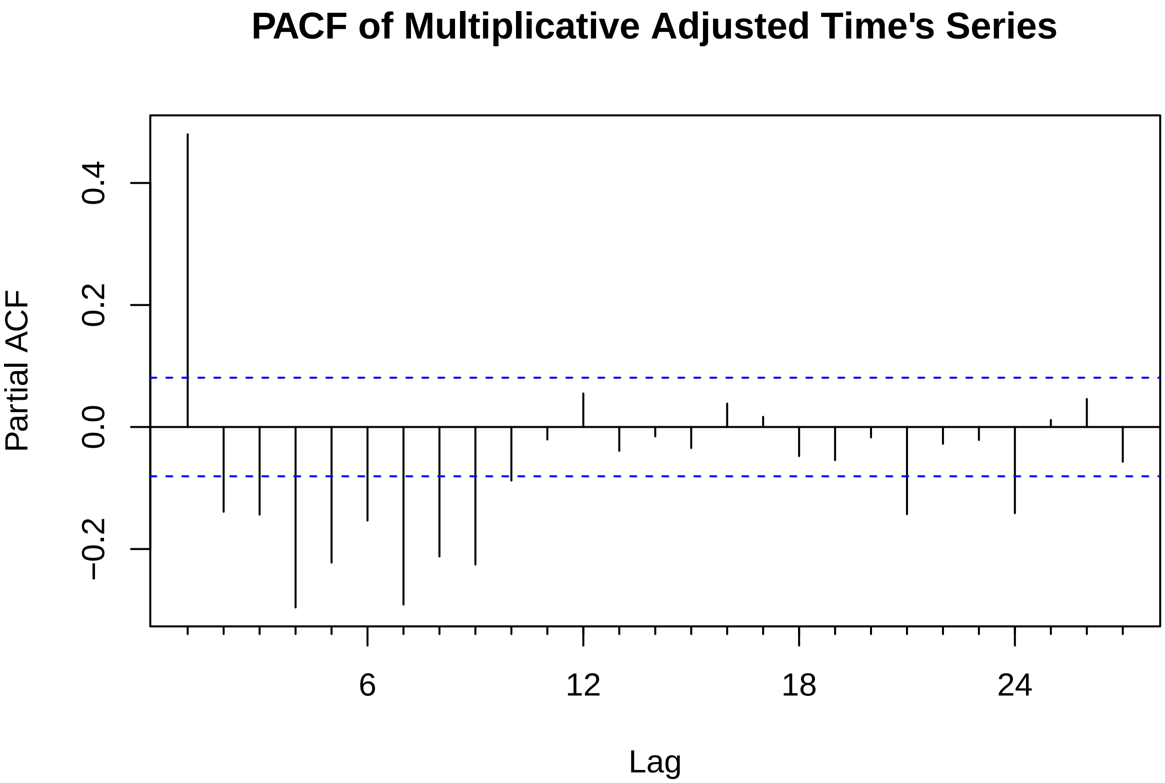
Overall, the analysis of the ACF and PACF plots for the additive adjusted time series implies that the linear model, after accounting for trend and seasonality, captures the systematic components of the time series well. What remains is seemingly random noise, which is an ideal outcome for the residuals of a well-fitted time series model.

(b)Multiplicative Decomposition of Time Series

```
Mdcmp = decompose(ts(ets,frequency=12), "multiplicative") #Multiplicative decomposing time series of ti

Mtrend = Mdcmp$trend
Mseasonal = Mdcmp$seasonal
Mrandom = Mdcmp$random


Mult_ts = (ec/Mtrend)/Mseasonal

Acf(Mult_ts, main= "ACF of Multiplicative Adjusted Time Series")
```



```
Pacf(Mult_ts, main= "PACF of Multiplicative Adjusted Time's Series")
```



The ACF and PACF plots for the multiplicative adjusted time series suggest that after seasonal and trend adjustments have been made, the data exhibits minimal autocorrelation.

In the ACF plot, most autocorrelations at various lags fall within the confidence bounds (blue dashed lines), indicating no significant autocorrelation. However, there are a few lags where the ACF is slightly above the upper confidence bound, but these do not exhibit a systematic pattern that would suggest a seasonal or trend component has been overlooked.

The PACF plot shows that all partial autocorrelations are within the confidence bounds, suggesting that there is no additional autoregressive behavior in the residuals of the time series after the multiplicative adjustments have been made.

Overall, both the ACF and PACF plots indicate that the multiplicative model has accounted for the trend and seasonal variations in the data well, leaving behind residuals that resemble white noise, meaning they do not exhibit any clear patterns or structures. This is typically the desired outcome when the goal is to have a time series where the only remaining variations are random and cannot be further modeled.

III. In conclusion, the extensive analysis of the energy consumption time series through various statistical models and diagnostics has indicated that the linear model, despite its simplicity, provides a statistically better fit compared to its counterparts. However, the forecasts generated by this model, characterized by wide confidence intervals, suggest considerable uncertainty in the predictions, especially as the forecast horizon extends. This level of uncertainty underscores the potential for enhancing the model's predictive power.

The presence of large confidence intervals could indeed be a reflection of unaccounted cyclical patterns inherent in the data. These cycles could stem from economic factors, policy changes, technological advancements, or other periodic phenomena that influence energy consumption over time. Incorporating these cyclical effects into the model could lead to a more nuanced understanding of the underlying processes and yield more accurate forecasts.

Moreover, an exploration into the stochastic properties of the time series has revealed that the residuals, after adjustments for trend and seasonality, exhibit properties close to white noise in both additive and multiplicative models. This is reassuring, as it implies that the major systematic components of the time series have been adequately captured. Nonetheless, the residual analysis hints at potential improvements in the modeling of the random components. A dynamic modeling approach, such as ARIMA (Autoregressive Integrated Moving Average) or SARIMA (Seasonal ARIMA), could offer a means to better encapsulate these dynamics and refine the forecast.

Additional investigation into the nature of the residuals through heteroscedasticity tests and exploring volatility clustering with models like GARCH (Generalized Autoregressive Conditional Heteroskedasticity) may further improve the accuracy and reliability of the predictive model.

In summary, while the linear model serves as a decent baseline, the breadth of the confidence intervals in the forecasts calls for a more sophisticated approach that accounts for cyclicality and dynamic error structures. Enhancing the model with these elements could significantly tighten the prediction intervals, reduce forecast uncertainty, and provide a more robust tool for policymakers and stakeholders in the energy sector.