Energy Production Prediction

By: Dennis Mutuku, Wyclife Orimba, Elsie Ochieng and Allan Ngeiywa

Project Overview and Business Understanding

In this analysis, the goal is to develop a time series forecasting model to predict energy production using a dataset of historical energy production data to enable informed decision-making

BUSINESS PROBLEM

The phasing out of traditional energy sources might create a potential gap incapable of meeting the nation's energy demands.

Ensuring a smooth transition necessitates a prediction model that accurately anticipates energy production, guaranteeing a balance between the declining traditional sources and the burgeoning renewable ones.

Project Objectives

Integrate external variables such as weather data to enhance the model's predictive capabilities.

Develop a robust time series forecasting model that accurately predicts energy production patterns.

Create an adaptive model that can adjust to changing conditions and evolving production patterns.

Provide a user-friendly interface for stakeholders to access and interpret forecasted energy production data.

Contribute to cost savings, efficient resource utilization, and environmental sustainability through improved energy production forecasts.

Data understanding and Analysis

The data used was sourced from the Federal Reserve Economic Data (FRED) that contains frequently updated US macro and regional economic time series at annual, quarterly, monthly, weekly, and daily frequencies. The data contains the following features:

Date: Captures monthly timestamps from January 1939 to October 2023.

IPG2211A2N: Values showing energy produced for each month

Future Engineering

Feature engineering is an essential step in time series forecasting, as it involves transforming raw data into meaningful features that can be effectively utilized by the forecasting model.

We extracted two new features from the 'Date' column: 'Year' and 'Month'. These new features will provide valuable insights into the seasonal patterns of the time series data.

Exploratory Data Analysis and Visualization

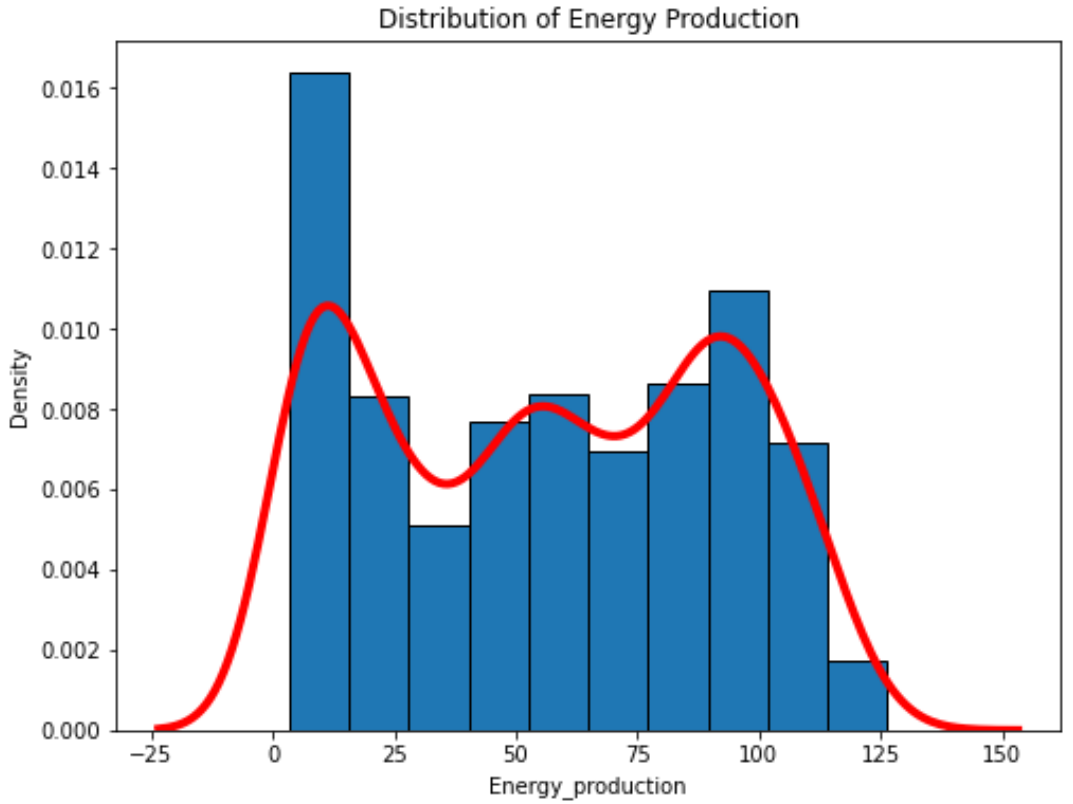
In this section, the focus is on obtaining crucial insights from the energy production dataset through Exploratory Data Analysis (EDA) and Visualization. As the United States shifts toward renewable energy sources, the primary objective is to forecast accurate energy patterns.

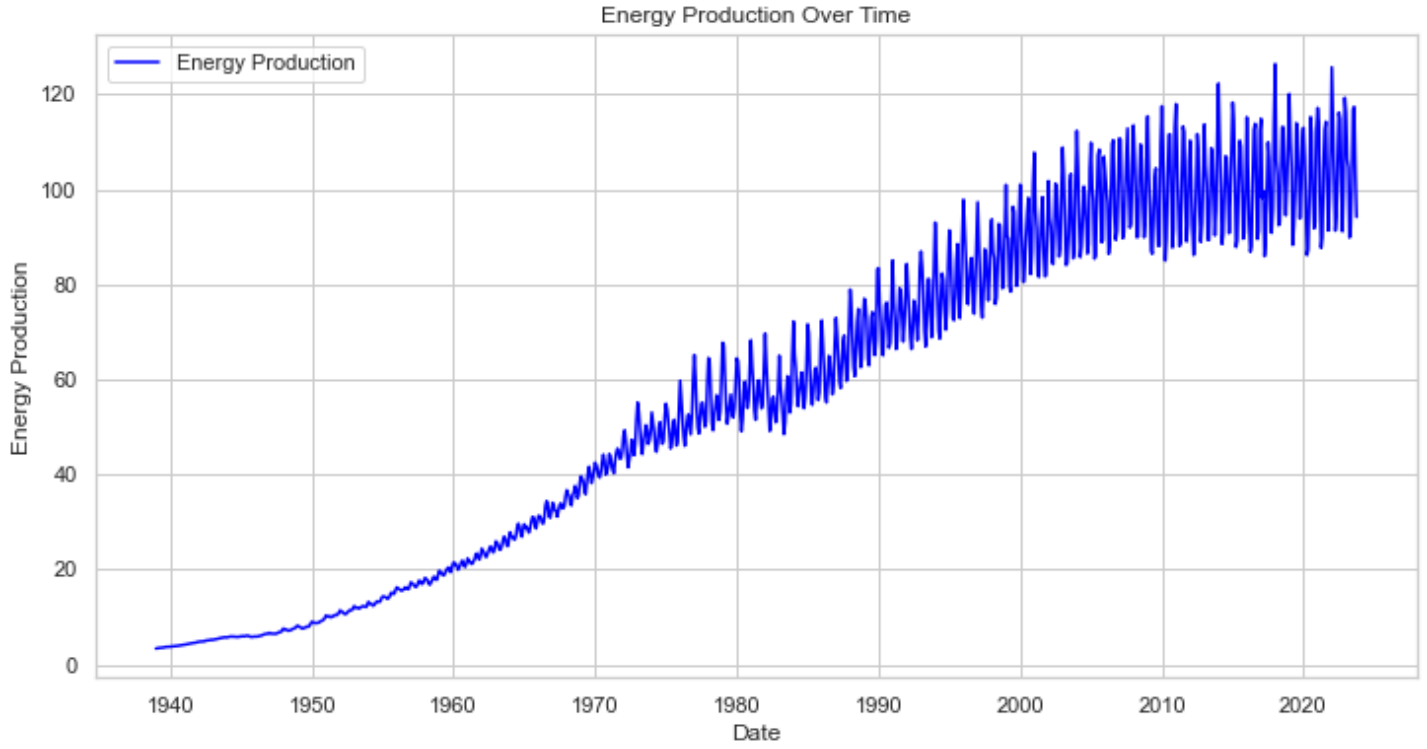
This analysis delves into understanding the dataset's temporal trends, statistical characteristics, and seasonal variations to uncover trends, patterns, and relationships.

By utilizing various visualization techniques and statistical methods, this exploration aims to inform the development of robust forecasting models and shed light on the dynamics of energy production during this transitional phase.

Univariate Analysis

A visual inspection of the data indicates that it does not have a normal distribution as shown below

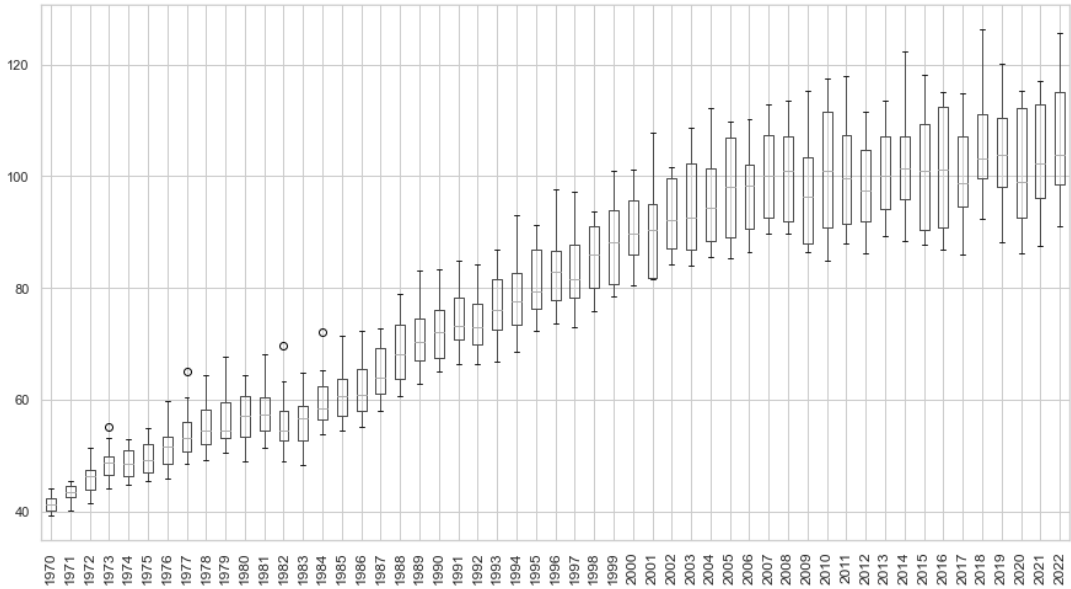




The plot shows that energy production has increased steadily over time.

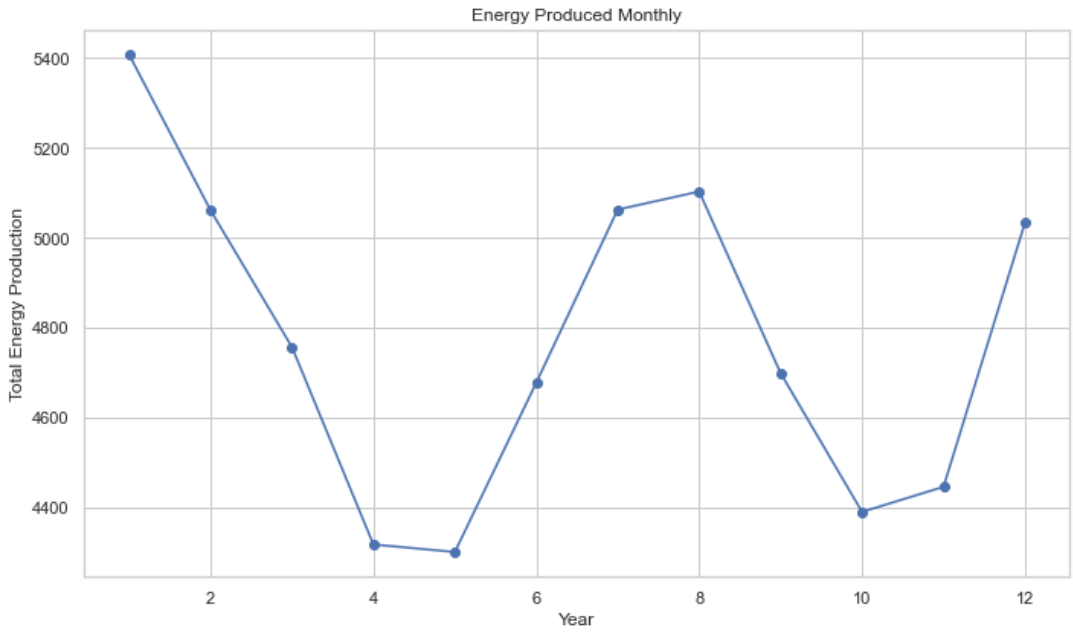
Box plot showing energy production over the years

- From the boxplots we see that the median energy production has increased steadily over time, from around 40 units in 1970. This suggests that the typical power plant is producing more energy today than it was in the past.



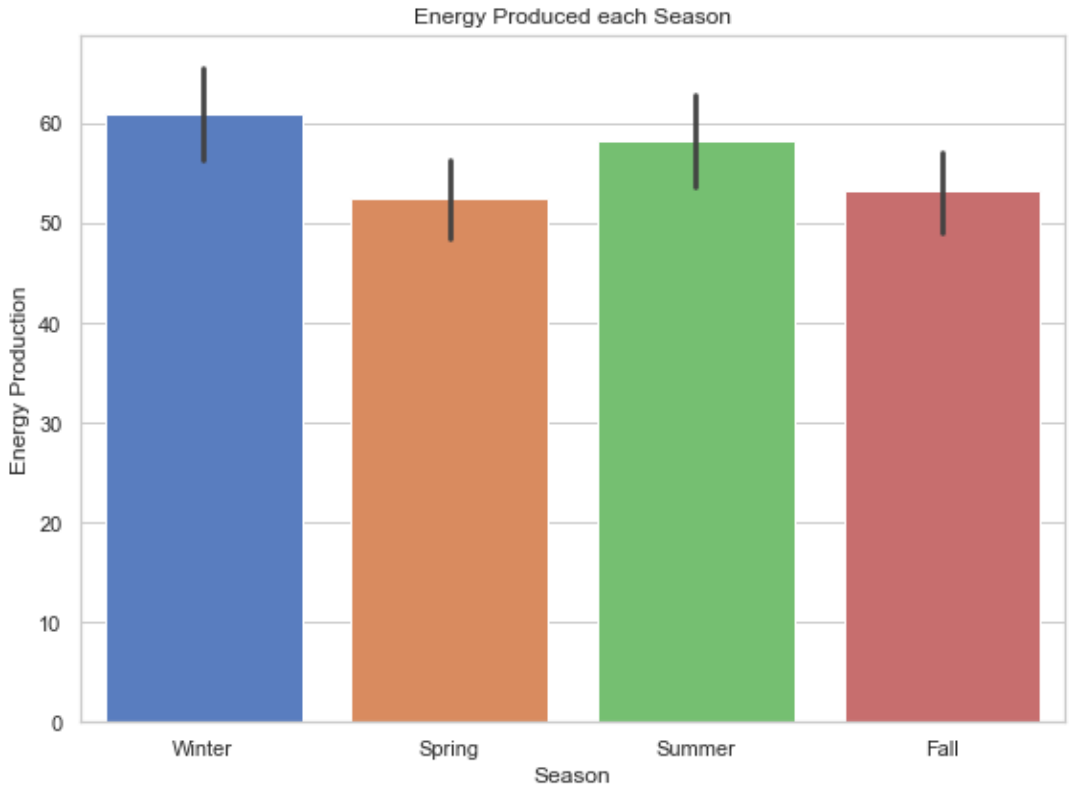
Bivarite Analysis

- In this section, we look at the relationship between monthly energy production and the ever-changing weather seasons
- From the line graph below we noted that energy demand was lowest in April, May, October and November



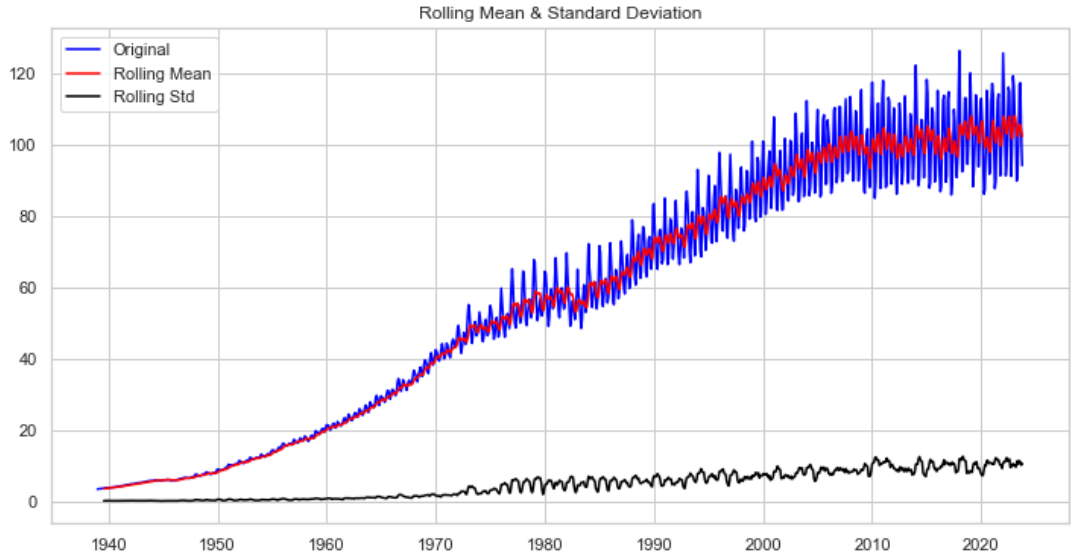
Energy Produced each Season

- The graph below illustrates energy production by season showing Fall and Spring having lowest energy produced further highlighting Fall and Spring as periods with the lowest energy production.
- This reaffirms that seasons featuring milder weather conditions exhibit lower consumption, consequently resulting in reduced production



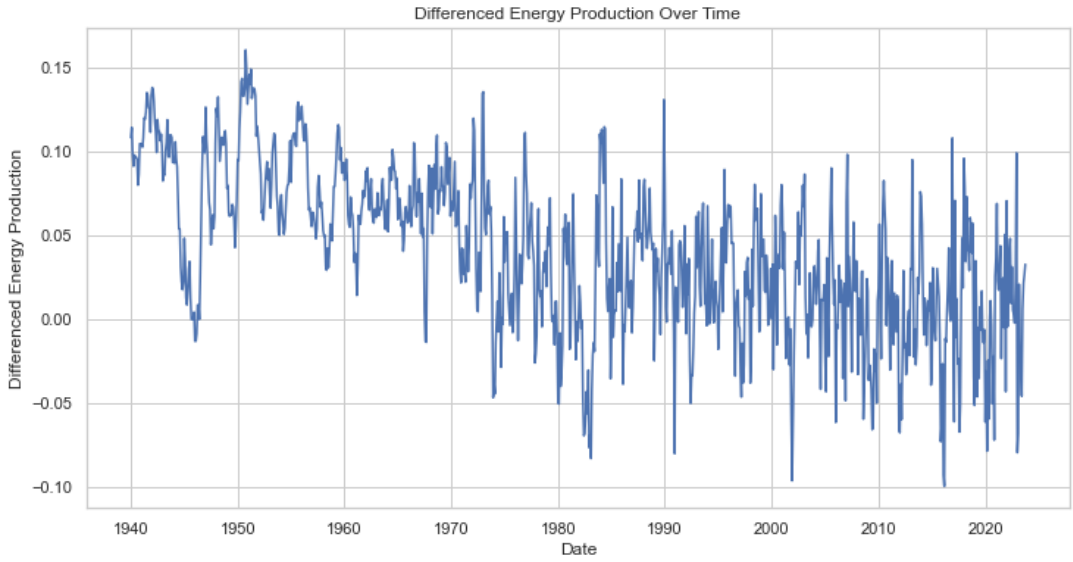
Time Series Modeling

- We first used rolling mean and AD Fuller test to check for stationarity of the data. The results showed that the data was not stationary as shown by the below plot



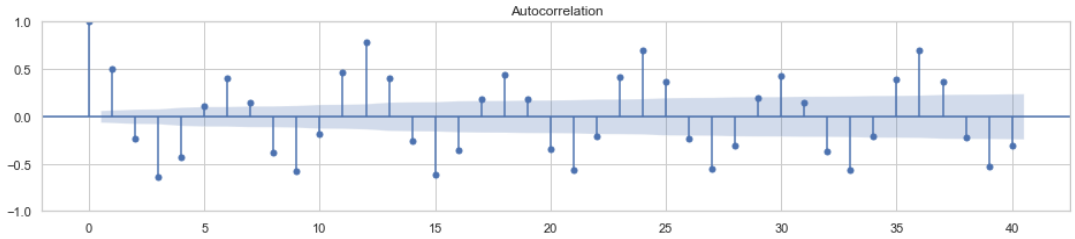
Differencing and decomposition

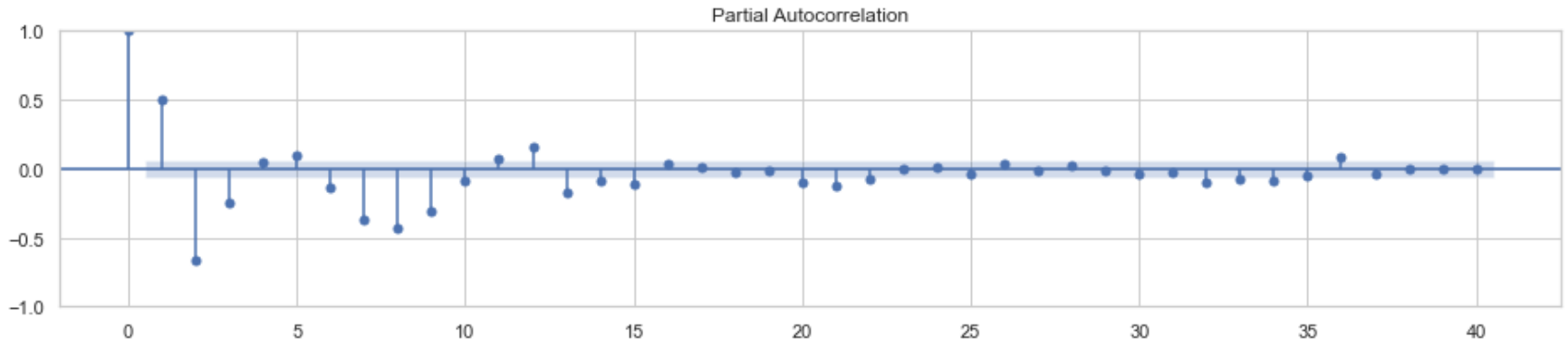
- Since our data was not stationary, we used differencing to check for seasonality and log transformation to make the data stationary. The resulting plot is as shown below:
- The AD fuller test after doing this confirmed that the data is stationary.



Correlation and Autocorrelation

- The PACF plot shows that the energy production data is significantly correlated with the previous lag, but not with any of the subsequent lags. This suggests that the energy production data is an AR(1) process, meaning that the current value of the data is linearly dependent on the previous value of the data.
- The ACF plot shows that the energy production data is significantly correlated with the first few lags, but the correlation decays as the lag increases.





PACF plot

Model 1: PMDARIMA

- PMDARIMA is python library used for automating the process of choosing the best ARIMA model for uni-variate time series data.
- It provides an auto-arima function that automatically selects the optimal ARIMA model for a given time series.

SARIMAX Results

Dep. Variable:	y	No. Observations:	804
Model:	SARIMAX(0, 1, 2)x(2, 1, [1], 12)	Log Likelihood	2088.479
Date:	Tue, 12 Dec 2023	AIC	-4164.957
Time:	01:12:32	BIC	-4136.917
Sample:	0 - 804	HQIC	-4154.180
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	-0.4655	0.031	-14.831	0.000	-0.527	-0.404
ma.L2	-0.5148	0.031	-16.616	0.000	-0.576	-0.454
ar.S.L12	-0.0744	0.075	-0.996	0.319	-0.221	0.072
ar.S.L24	-0.1040	0.050	-2.060	0.039	-0.203	-0.005
ma.S.L12	-0.5255	0.071	-7.431	0.000	-0.664	-0.387
sigma2	0.0003	1.21e-05	24.435	0.000	0.000	0.000

Ljung-Box (L1) (Q):	1.97	Jarque-Bera (JB):	242.89
Prob(Q):	0.16	Prob(JB):	0.00
Heteroskedasticity (H):	4.76	Skew:	-0.07
Prob(H) (two-sided):	0.00	Kurtosis:	5.71

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

SARIMAX Results

Dep. Variable:	resid	No. Observations:	804
Model:	SARIMAX(0, 1, 2)x(2, 1, [1], 12)	Log Likelihood	2088.479
Date:	Tue, 12 Dec 2023	AIC	-4164.957
Time:	01:12:40	BIC	-4136.917
Sample:	07-01-1939 - 06-01-2006	HQIC	-4154.180
Covariance Type:	opg		

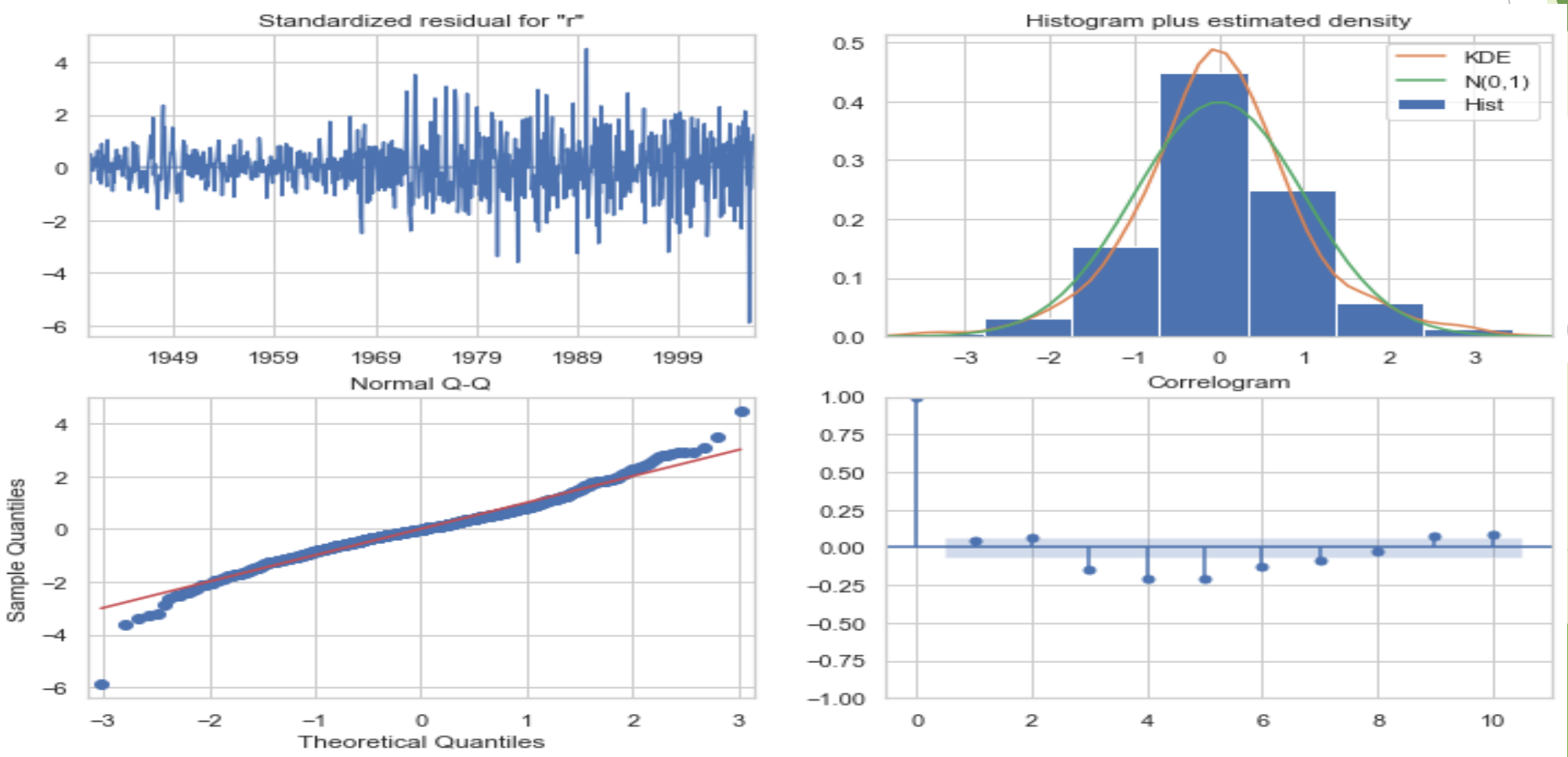
	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	-0.4655	0.031	-14.831	0.000	-0.527	-0.404
ma.L2	-0.5148	0.031	-16.616	0.000	-0.576	-0.454
ar.S.L12	-0.0744	0.075	-0.996	0.319	-0.221	0.072
ar.S.L24	-0.1040	0.050	-2.060	0.039	-0.203	-0.005
ma.S.L12	-0.5255	0.071	-7.431	0.000	-0.664	-0.387
sigma2	0.0003	1.21e-05	24.435	0.000	0.000	0.000

Ljung-Box (L1) (Q):	1.97	Jarque-Bera (JB):	242.89
Prob(Q):	0.16	Prob(JB):	0.00
Heteroskedasticity (H):	4.76	Skew:	-0.07
Prob(H) (two-sided):	0.00	Kurtosis:	5.71

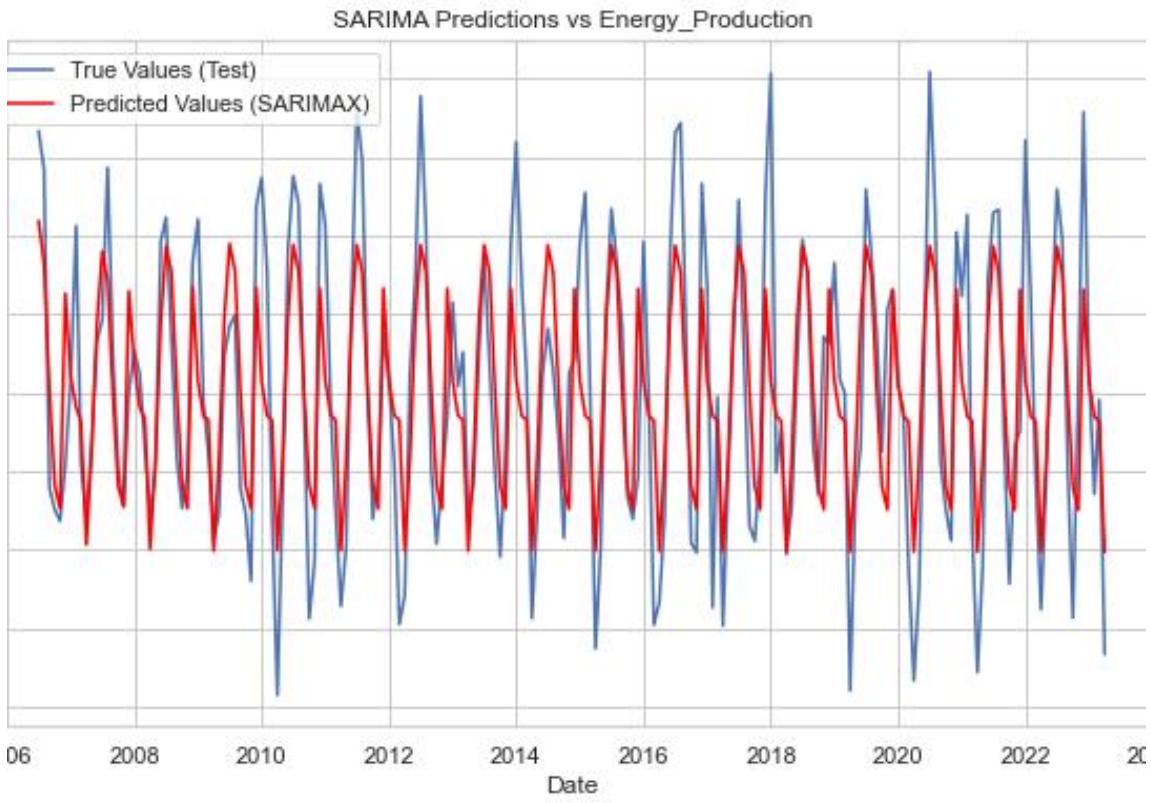
Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

SARIMAX Model Summary

Diagnostic plots showing the model results



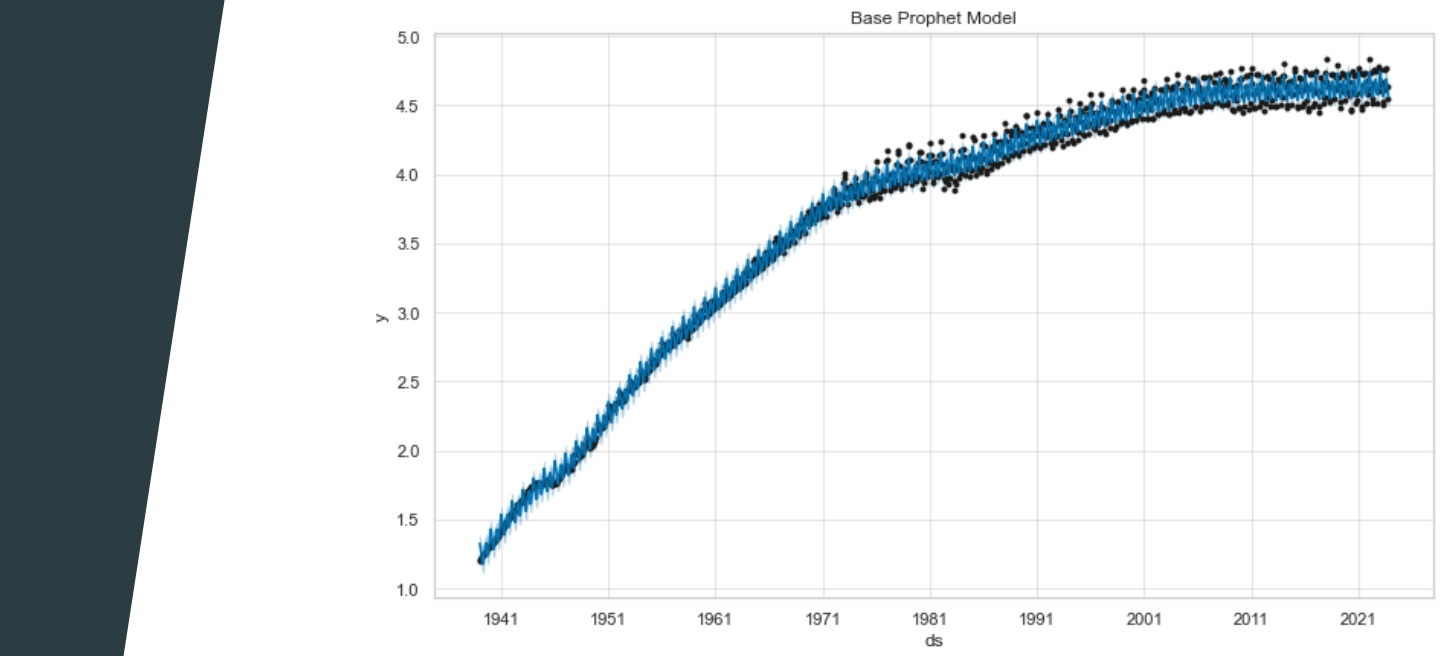
Model Evaluation



- From the plot of predictions and actual values we see that the predicted values track the actual values closely, with only a few minor deviations. The model also captures the overall trend of the data well.

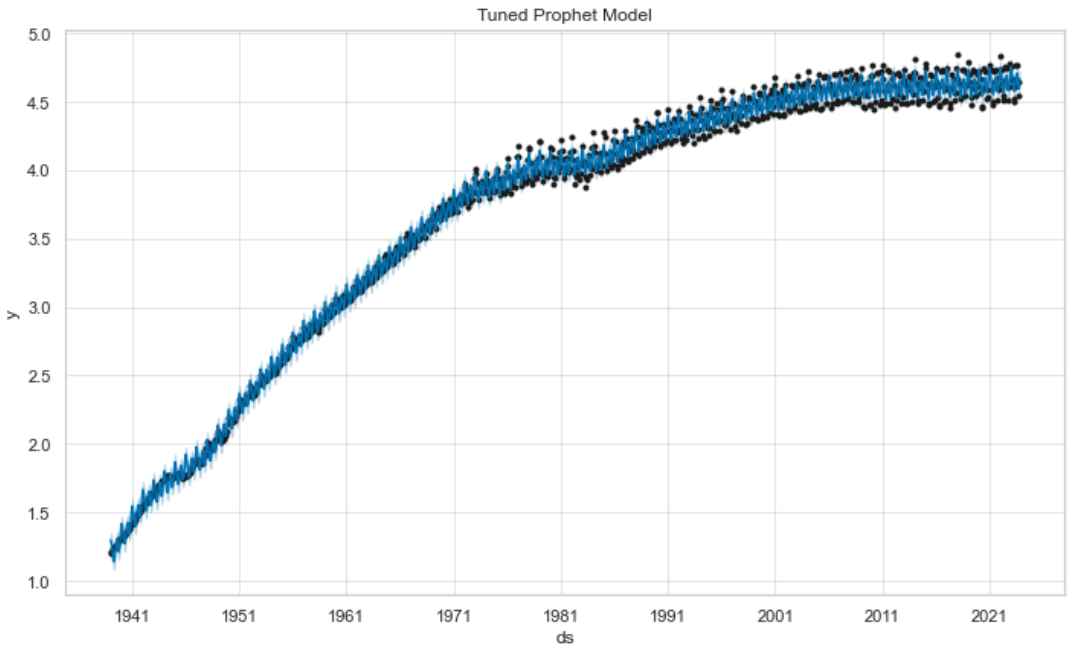
Model 2: Facebook Prophet

- Facebook Prophet is a forecasting tool designed for time series data that exhibits patterns such as trends and seasonality.



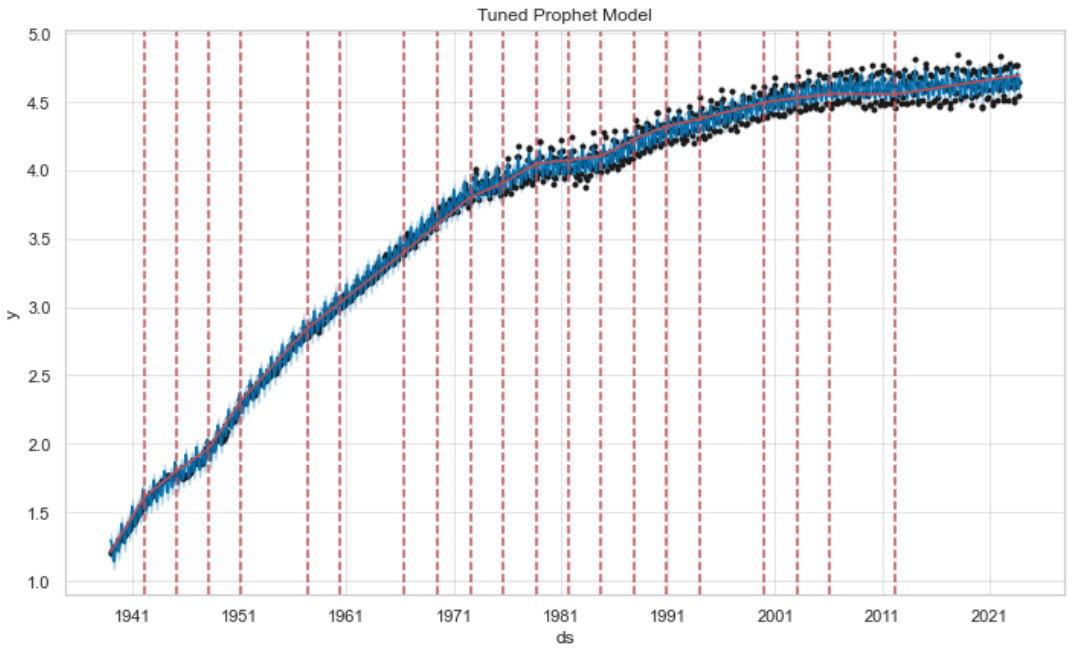
Model 3:Tuning Facebook Prophet Model

- When tuning a Facebook Prophet model, the focus is on adjusting hyperparameters and optimizing the model for better forecasting performance. By default, Prophet automatically detects changepoints in the time series.



Adding Changepoints

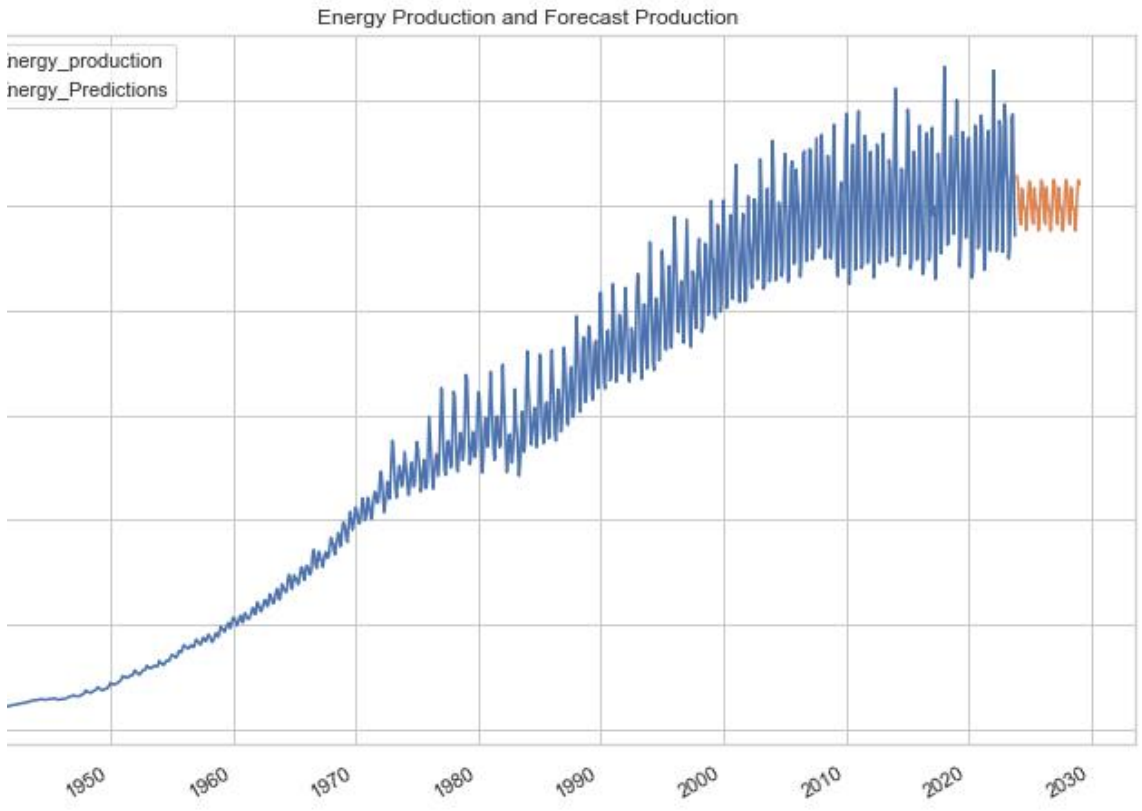
- Changepoints determines how flexible the model is to the overall changepoints fit across the entire dataset.
- The red dotted lines indicate that the Prophet model has identified significant changes in the trend or seasonality of the data at those points in time. These changes could be due to a variety of factors, such as technological advancements, changes in government policy, or economic conditions



Models Observations

- Our evaluation of models centered on the Mean Squared Error (MSE), a metric gauging the average squared difference between actual and estimated values. A lower MSE signifies superior predictive accuracy.
- the PMDARIMA model emerges as the best model having the lowest MSE. The best-performing model having the lowest MSE. This outcome implies that the PMDARIMA model offers superior predictive capability compared to the base Prophet and tuned Prophet models.

Forecasting Future Energy Production



- From the this plot, it appears that the model managed to capture some of the trend of energy production and short term fluctuations such as the seasonal peaks and valleys.
- The model however appears to underestimate the uncertainty in the forecasts which may be as a result of insufficient data, or lack of other factors that influence energy production such as economic fluctuations.

Conclusion

- The study aimed at predicting energy production trends, enabling a smooth transition towards renewable sources while emphasizing accurate forecasts in the face of shifting demand and supply dynamics.
- Despite promising performance by the selected PMDARIMA model, the study also revealed limitations. The models demonstrated an underestimation of uncertainty in predicted values, suggesting potential gaps in accounting for diverse factors influencing energy production.

Recommendations

Incorporating other factors: When devising energy production strategies, it's crucial for the United States to factor in seasonal variations in weather

Balancing Renewable Energy Reliability: The transition towards renewable energy sources demands a comprehensive approach

Optimizing Infrastructure Management: Strategic planning for energy grid maintenance or significant alterations should be scheduled during periods of low energy demand, notably in April, May, October, and November