A

Mini Project Report

On

"Analysis & Estimation of Global Energy Prediction"

Submitted for partial fulfillment of requirement for the award of degree

of

Masters of Business Administration
(Artificial Intelligence and Data Science)

GRAPHIC ERA (DEEMED TO BE UNIVERSITY)
DEHRADUN (UTTARAKHAND)

Session: 2023-2025

Supervision by:
Supervisor Name: Mr. Chandraprakash Thukral
Designation: Asst. Professor

Submitted by:
Name: Kartikey Chaurasia
Roll No.: 1404320
Enrollment No.: GE-23144320

DEPARTMENT OF MANAGEMANT STUDIES
GRAPHIC ERA (DEEMED TO BE UNIVERSITY) DEHRADUN

DECLARATION

I hereby declare that the Internship/ Field Project entitled **"Analysis & Prediction of Global Energy Prediction**" submitted for the Degree of Master of Business Administration in Artificial Intelligence and Data Science, is my original work and the Mini Project has not formed the basis for the award of any degree, diploma, associateship, fellowship, or similar other titles. It has not been submitted to any other University or Institution for the award of any degree or diploma.

(Signature of Student)

Name of the Student – Kartikey Chaurasia

CERTIFICATE BY SUPERVISOR

I have the pleasure in certifying that Mr./Ms. Kartikey Chaurasia is a student of Graphic Era (Deemed to be University) of the Master's Degree in Business Administration (MBA) in AI&DS. his University Roll No is. 1404320

He has completed his Mini Project titled as "Analysis & Prediction of Global Energy Prediction" under my guidance.

I certify that this is his original effort & has not been copied from any other source. This project has also not been submitted in any other university for the purpose of award of any Degree.

This project fulfils the requirement of the curriculum prescribed by Graphic Era (Deemed to be University), Dehradun, for the said course.

I recommend this Mini Project for evaluation & consideration for the award of Degree to the student.

Signature: Name of the Guide:

Signature: Name of the Area Chair/ HOD:

ACKNOWLEDGEMENT

I express my sincere thanks to my project guide, Mr. Chandraprakash Thukral Designation Assistant Professor, Department of Management Studies for guiding me right form the inception till the successful completion of the project.

I also record my indebtedness to my supervisor, Prof. Chandraprakash Thukral, for his counsel and guidance during the preparation of this Mini Project.

I wish to record my sincere thanks to friends and classmates for their help and cooperation throughout our project.

(Signature of Student)

Name of the Student – Kartikey Chaurasia

Table of Contents

Table of Figures

Abstract

The global energy market is a pivotal component of the world economy, with energy prices influencing a wide range of economic activities and policies. This report delves into the complexities of predicting global energy prices, a task of significant importance for policymakers, businesses, and investors. By analysing historical trends, examining the factors that influence energy prices, and evaluating various prediction methodologies, this report aims to provide a comprehensive understanding of the dynamics at play.

Historical data reveals that energy prices are subject to considerable volatility due to factors such as geopolitical tensions, technological advancements, and regulatory changes. The supply and demand dynamics of energy resources, along with environmental considerations, further complicate price predictions. Traditional econometric models, alongside more recent machine learning approaches, offer varied strengths and weaknesses in forecasting accuracy.

This report includes detailed case studies of successful and unsuccessful energy price predictions, providing practical insights into the effectiveness of different methodologies. Additionally, it explores current trends in the energy market, such as the increasing prominence of renewable energy sources and evolving energy policies, to predict future price trajectories. The findings suggest that while no single prediction model can capture the full complexity of global energy prices, a hybrid approach that integrates multiple models may offer more robust predictions. The report concludes with recommendations for stakeholders to navigate the challenges and opportunities presented by the global energy market.

By offering a thorough analysis and estimation of global energy price prediction, this report aims to equip readers with the knowledge and tools necessary to make informed decisions in a volatile and critical sector.

Chapter 1 : Introduction

Introduction

Energy prices are a crucial component of the global economy, affecting everything from industrial production and transportation to household energy bills and national economic stability. Given the widespread impact of energy prices, accurate prediction and analysis are vital for policymakers, businesses, and consumers alike. Fluctuations in energy prices can lead to significant economic consequences, influencing inflation rates, trade balances, and the cost of goods and services. For instance, a surge in oil prices can increase transportation and production costs, leading to higher prices for a wide range of products.

The global energy market is characterized by its complexity and volatility, involving a mix of traditional fossil fuels (oil, natural gas, coal) and renewable energy sources (solar, wind, hydropower). Each of these energy sources has distinct supply and demand dynamics, regulatory environments, and technological advancements that affect their market prices. Fossil fuels have long dominated the energy market due to established infrastructures and relatively lower costs. However, the increasing emphasis on sustainability and environmental concerns is driving a significant shift towards renewable energy sources. This transition is adding layers of complexity to energy price predictions, as renewable energy adoption grows and interacts with existing energy infrastructures.

This project aims to conduct a comprehensive analysis and estimation of global energy price predictions. It seeks to examine historical energy price data to identify patterns and key events that have influenced price movements over the past decades. It also investigates the primary factors that influence energy prices, including supply and demand dynamics, geopolitical developments, technological innovations, and environmental regulations. Various methodologies used for predicting energy prices, such as econometric models, machine learning techniques, and hybrid approaches, will be assessed, highlighting their respective strengths and limitations. Additionally, the project presents case studies of both successful and unsuccessful energy price predictions to provide practical insights into the effectiveness of different prediction models. By analysing current trends in the energy market, with a focus on the increasing role of renewable energy and evolving energy policies, the project aims to project future price trends over the next decade. Finally, it offers strategic recommendations for stakeholders, including policymakers, businesses, and investors, on how to navigate the challenges and opportunities in the global energy market.

1.1. Technology Used

Machine Learning is an AI technique that teaches computers to learn from experience. Machine learning algorithms use computational methods to "learn" information directly from data without relying on a predetermined equation as a model.

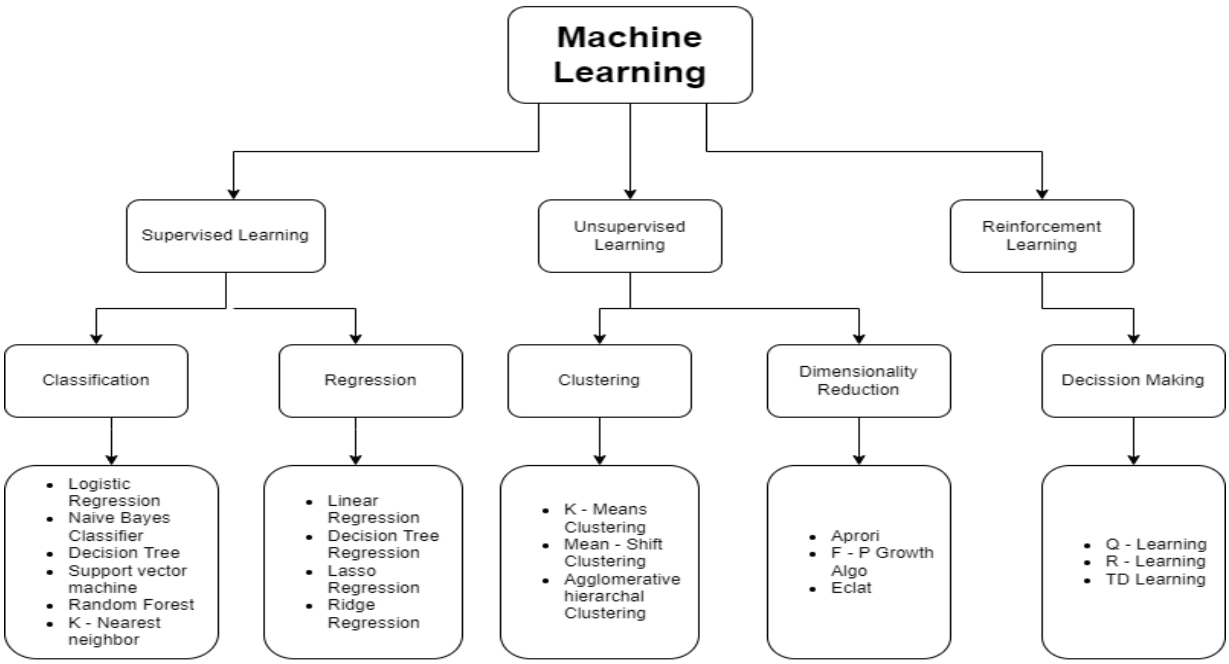


Figure 1 : Types of Machine Learning

1.1.1. Supervised Learning

In supervised learning, the training data consists of labelled examples, where each example includes both input features and corresponding target labels. The goal is to learn a mapping from input features to the desired output labels. The algorithm learns from the labelled data and then can make predictions or classifications on new, unseen data.

Examples of supervised learning algorithms include linear regression, decision trees, support vector machines (SVM), and neural networks.

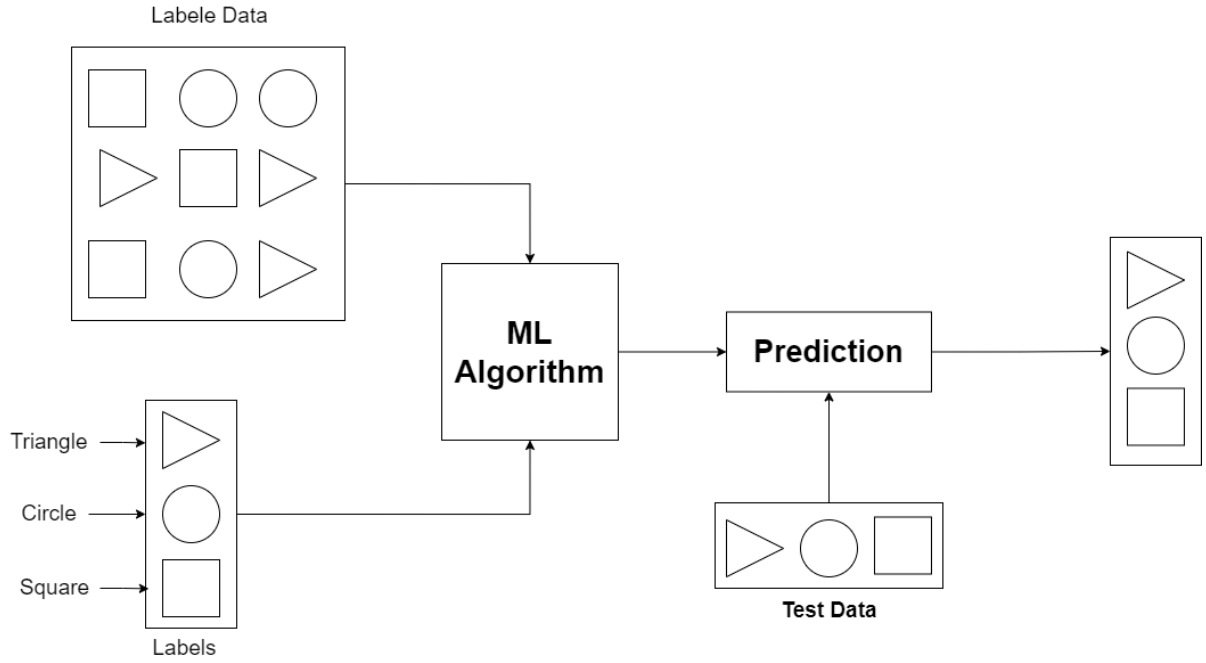


Figure 2 : Working of Supervised Learning

1.1.2. Unsupervised learning

- Unsupervised learning deals with unlabelled data, where the training data consists only of input features without any corresponding target labels. The objective of unsupervised learning is to discover underlying patterns, structures, or relationships within the data.
- Common tasks in unsupervised learning include clustering, where similar data points are grouped together, and dimensionality reduction, which aims to represent the data in a lower-dimensional space.
- Popular unsupervised learning algorithms include k-means clustering, hierarchical clustering, and principal component analysis (PCA).

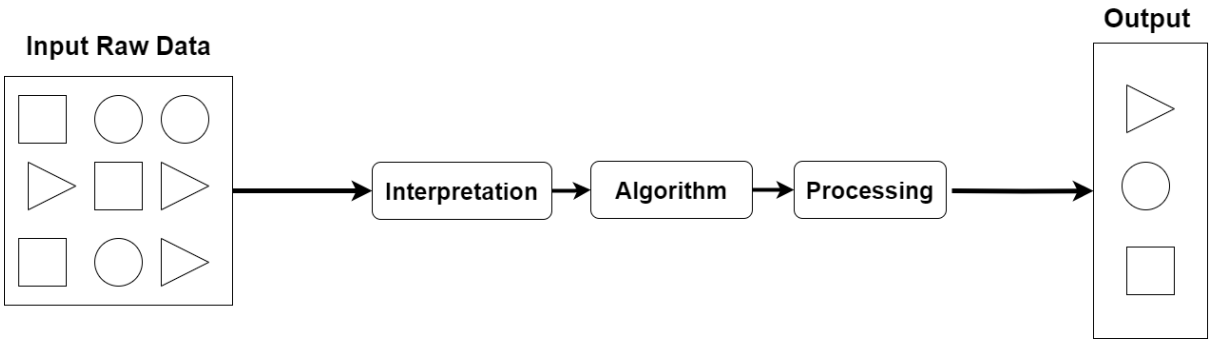


Figure 3 : Working of Unsupervised Learning

1.1.3. Reinforcement Learning

- Reinforcement Learning: Reinforcement learning involves an agent that learns to interact with an environment to maximize a reward signal. The agent learns through trial and error, exploring different actions and receiving feedback in the form of rewards or penalties.
- The goal is to discover an optimal policy or decision-making strategy that maximizes long-term cumulative rewards. Reinforcement learning has applications in robotics, game playing, autonomous vehicles, and many other domains. Notable algorithms in reinforcement learning include Q-learning, deep Q-networks (DQN), and policy gradient methods.

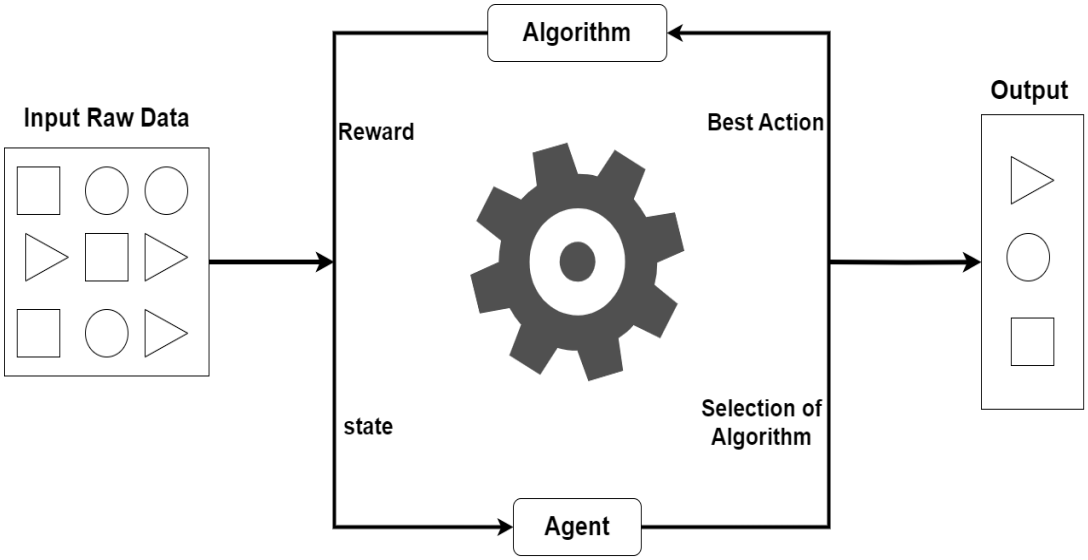


Figure 4 : Working of Reinforcement Learning

1.2. How Machine learning helps in Energy Sector?

Machine learning (ML) has emerged as a transformative technology in the energy sector, offering innovative solutions to complex problems and driving efficiency, sustainability, and cost savings. By leveraging vast amounts of data, machine learning algorithms can identify patterns, make predictions, and optimize operations in ways that were previously unattainable. Here are several key areas where machine learning is making a significant impact in the energy sector:

1. **Predictive Maintenance:** One of the most significant applications of machine learning in the energy sector is predictive maintenance. By analysing data from sensors embedded in equipment, machine learning models can predict when a piece of machinery is likely to fail. This allows energy companies to perform maintenance before a breakdown occurs, reducing downtime and maintenance costs. Predictive maintenance is particularly valuable in wind turbines, power plants, and oil rigs, where unexpected failures can lead to significant financial losses and safety hazards.

2. **Energy Price Forecasting:** Accurate energy price forecasting is crucial for energy producers, consumers, and traders. Machine learning algorithms can analyze historical price data, supply and demand trends, geopolitical factors, and weather patterns to predict future energy prices. These predictions help stakeholders make informed decisions about production, purchasing, and trading strategies. Machine learning models, such as neural networks and support vector machines, have proven to be effective in capturing the complexities of energy markets and improving forecasting accuracy.

3. **Demand Response Management:** Machine learning is playing a vital role in demand response management by predicting energy consumption patterns and enabling more efficient energy use. Smart grids equipped with machine learning algorithms can analyze real-time data from smart meters, weather forecasts, and historical consumption data to optimize energy distribution. This helps balance supply and demand, reduce energy wastage, and lower costs for both utilities and consumers. Demand response management is particularly important for integrating renewable energy sources, which can be intermittent and unpredictable.

4. **Renewable Energy Integration:** Integrating renewable energy sources into the power grid poses challenges due to their variable nature. Machine learning helps address these challenges by forecasting renewable energy generation based on weather data and historical performance. For instance, machine learning models can predict solar and wind energy output, enabling grid operators to plan for fluctuations and maintain grid stability. Additionally, ML algorithms can optimize the operation of energy storage systems, ensuring that excess energy generated during peak times is stored and used efficiently.

5. **Energy Efficiency Optimization:** Machine learning algorithms are being used to optimize energy consumption in buildings, industrial processes, and transportation systems. By analyzing data from sensors and control systems, machine learning can identify inefficiencies and recommend actions to reduce energy use. For example, in smart buildings, ML algorithms can adjust heating, ventilation, and air conditioning (HVAC) systems in real-time to maintain comfort while minimizing energy consumption. In industrial settings, machine learning can optimize production processes to reduce energy waste and improve overall efficiency.

6. **Anomaly Detection and Security:** Ensuring the security and reliability of energy infrastructure is critical. Machine learning is used to detect anomalies in energy systems that may indicate cyber-attacks, equipment malfunctions, or other issues. By continuously monitoring data from various sources, machine learning models can

identify unusual patterns and alert operators to potential threats. This proactive approach helps protect critical infrastructure and maintain the integrity of energy supply.

7. **Enhancing Customer Experience:** Machine learning is also enhancing the customer experience in the energy sector. Utilities can use ML algorithms to analyze customer data and provide personalized energy-saving recommendations. Additionally, machine learning can improve customer service by powering chatbots and virtual assistants that handle inquiries and provide support. This leads to higher customer satisfaction and engagement.

In summary, machine learning is revolutionizing the energy sector by improving predictive maintenance, energy price forecasting, demand response management, renewable energy integration, energy efficiency optimization, anomaly detection, and customer experience. As the technology continues to evolve, its applications in the energy sector are expected to expand, driving further innovation and sustainability.

Chapter 2 : Literature Review

Literature Review

The writing survey for the venture "Examination & Estimation of Worldwide Vitality Cost Forecast" looks at existing inquire about and thinks about related to vitality cost estimating, the variables affecting vitality costs, and the application of machine learning in the vitality segment. This survey gives a establishment for understanding the current state of information and distinguishes crevices that this venture points to address.

Energy Cost Forecasting

Energy cost determining is a well-researched range with various ponders investigating distinctive strategies. Conventional econometric models, such as time arrangement examination and relapse models, have been broadly utilized for this reason. For occasion, Pindyck (1999) utilized econometric models to analyze the instability of oil costs and concluded that geopolitical occasions altogether affect cost variances. Additionally, Serletis (1991) utilized vector autoregression (VAR) models to foresee characteristic gas costs, highlighting the significance of supply and request dynamics.

However, conventional models frequently battle to capture the complexity and non-linear nature of vitality markets. This has driven to the appropriation of more progressed procedures, counting machine learning and half breed models. A think about by Yu, Wang, and Lai (2008) illustrated that neural systems might outflank conventional models in determining rough oil costs, especially in capturing sudden showcase shifts. Zhang, Lai, and Wang (2008) assist amplified this work by coordination hereditary calculations with neural systems to improve forecast accuracy.

Factors Impacting Vitality Prices

The writing recognizes a few key variables affecting vitality costs, counting supply and request flow, geopolitical occasions, mechanical progressions, and natural controls. Hamilton (2009) talked about the part of supply disturbances and geopolitical pressures in causing oil cost spikes. Kilian (2009) emphasized the affect of worldwide financial action on vitality request and costs, appearing that financial development in developing markets drives up vitality utilization and prices.

Technological progressions, especially in renewable vitality and vitality capacity, moreover play a significant part. Nemet (2006) highlighted how mechanical advancements in sun oriented photovoltaic (PV) cells have driven to noteworthy fetched decreases, affecting the competitiveness of sun based vitality. So also, IEA (2019) detailed that progressions in

battery capacity advances are pivotal for coordination renewable vitality into the lattice and stabilizing prices.

Environmental directions and approaches pointed at diminishing carbon emanations are progressively impacting vitality markets. Aldy and Stavins (2012) talked about the affect of carbon estimating and renewable vitality commands on vitality costs, noticing that such approaches can make showcase instabilities but too drive advancement and efficiency.

Machine Learning in the Vitality Sector

Machine learning has picked up impressive consideration in the vitality division for its capacity to handle huge datasets and complex designs. The writing on machine learning applications in vitality cost determining is broad. A think about by Ahmed et al. (2020) checked on different machine learning procedures utilized in vitality determining, counting bolster vector machines, choice trees, and profound learning models. The creators concluded that machine learning models for the most part outflank conventional strategies in terms of exactness and adaptability.

In the setting of prescient support, ML models have been effectively connected to foresee gear disappointments and optimize support plans. Carvalho et al. (2019) illustrated the adequacy of utilizing machine learning calculations to foresee issues in wind turbines, coming about in diminished downtime and support costs.

Furthermore, machine learning is essential in upgrading request reaction administration and joining renewable vitality sources. Hong et al. (2019) utilized machine learning models to anticipate power request and optimize the operation of keen networks, accomplishing critical vitality investment funds. Essentially, Bedi et al. (2018) connected ML methods to figure sun oriented and wind vitality generation, making a difference lattice administrators oversee changeability and keep up stability.

Gaps and Future Directions

Despite the headways in vitality cost estimating and the application of machine learning, a few holes stay. Numerous thinks about center on particular vitality sources or locales, constraining the generalizability of their discoveries. Furthermore, the integration of numerous vitality sources and the affect of developing advances, such as blockchain and the Web of Things (IoT), on vitality markets are regions that require assist exploration.

Future investigate ought to point to create more comprehensive models that consolidate a more extensive run of variables and use the most recent mechanical progressions. Cross breed

models that combine conventional econometric methods with machine learning hold guarantee for making strides expectation exactness and robustness.

In conclusion, the writing audit highlights the advance made in understanding and foreseeing vitality costs, the variables affecting these costs, and the part of machine learning in the vitality division. This extend builds on these bits of knowledge to give a more all encompassing investigation and estimation of worldwide vitality cost patterns, pointing to contribute to the continuous endeavors to improve forecast precision and advise decision-making in the vitality advertise.

Chapter 3: Data and Methodology

DATA AND METHODOLOGY

3.1. Data Collection

Initially, we collect a dataset for our Global Energy Prediction System from an online source Kaggle. Once the dataset is collected, it is divided into training data and testing data through a process known as data splitting. The training dataset is used for learning prediction model and testing data is used for evaluating the prediction model. For this project, 60% of data is used for training and 40% of data is used for testing. The dataset consists of 11 attributes.

```
energy_consumption_df
```

	Entity	Code	Year	Oil Consumption - EJ	Gas Consumption - EJ	Coal Consumption - EJ	Solar Consumption - EJ	Hydro Consumption - EJ	Nuclear Consumption - EJ	Wind Consumption - EJ	Geo Biomass Other - EJ
0	Algeria	DZA	1965	15.405264	7.430506	0.814101	0.0	1.111112	0.0	0.0	0.0
1	Algeria	DZA	1966	20.272721	7.719256	0.790841	0.0	0.986112	0.0	0.0	0.0
2	Algeria	DZA	1967	18.942046	7.488256	0.604760	0.0	1.138890	0.0	0.0	0.0
3	Algeria	DZA	1968	20.167318	7.873256	0.639651	0.0	1.563890	0.0	0.0	0.0
4	Algeria	DZA	1969	21.305947	10.351696	0.814101	0.0	1.002779	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...
5510	Zimbabwe	ZWE	2015	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5511	Zimbabwe	ZWE	2016	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5512	Zimbabwe	ZWE	2017	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5513	Zimbabwe	ZWE	2018	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5514	Zimbabwe	ZWE	2019	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

5515 rows × 11 columns

Figure 5 : Dataset

Data Fields

- Oil Consumption – EJ: Oil Consumption ExaJoule.
- Gas Consumption – EJ: Gas Consumption ExaJoule.
- Coal Consumption – EJ: Coal Consumption ExaJoule.
- Solar Consumption – EJ: Solar Consumption ExaJoule.
- Hydro Consumption – EJ: Hydro Consumption ExaJoule.
- Nuclear Consumption – EJ: Nuclear Consumption ExaJoule.
- Wind Consumption – EJ: Wind Consumption ExaJoule.
- Geo Biomass Other Consumption – EJ: Geo Biomass Other Consumption ExaJoule.

3.2. Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a crucial step in data analysis where various techniques and tools are employed to gain insights and understanding of the dataset. The primary goal of EDA is to summarize the main characteristics of the data and identify patterns, trends, and relationships between variables.

It involves visualizing the data using charts, graphs, and statistical measures, as well as performing data transformations and cleaning if necessary. EDA helps in detecting outliers, missing values, and anomalies, and assists in making informed decisions regarding data preprocessing, feature engineering, and modeling strategies. Overall, EDA plays a vital role in setting the foundation for further analysis and modeling in data-driven tasks.

1. Reading the csv file in jupyter notebook.

```
energy_consumption_df
```

	Entity	Code	Year	Oil Consumption - EJ	Gas Consumption - EJ	Coal Consumption - EJ	Solar Consumption - EJ	Hydro Consumption - EJ	Nuclear Consumption - EJ	Wind Consumption - EJ	Geo Biomass Other - EJ
0	Algeria	DZA	1965	15.405264	7.430506	0.814101	0.0	1.111112	0.0	0.0	0.0
1	Algeria	DZA	1966	20.272721	7.719256	0.790841	0.0	0.986112	0.0	0.0	0.0
2	Algeria	DZA	1967	18.942046	7.488256	0.604760	0.0	1.138890	0.0	0.0	0.0
3	Algeria	DZA	1968	20.167318	7.873256	0.639651	0.0	1.563890	0.0	0.0	0.0
4	Algeria	DZA	1969	21.305947	10.351696	0.814101	0.0	1.002779	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...
5510	Zimbabwe	ZWE	2015	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5511	Zimbabwe	ZWE	2016	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5512	Zimbabwe	ZWE	2017	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5513	Zimbabwe	ZWE	2018	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5514	Zimbabwe	ZWE	2019	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

5515 rows × 11 columns

Figure 6 : Reading the csv File

2. Then performing various EDA tasks on the dataframe like:-
    I. Data Cleaning: Remove or handle missing values, outliers, duplicates, and irrelevant data. This step ensures that the dataset is ready for analysis.
    II. Variable Identification: Determine the types of variables in the dataset, such as numerical, categorical, or ordinal variables. This step helps in understanding the nature of the data.

III. Univariate Analysis: Analyze individual variables one at a time. This includes calculating descriptive statistics, visualizing distributions using histograms, box plots, and summary tables.

IV. Bivariate Analysis: Explore relationships between pairs of variables. This involves comparing variables through scatter plots, correlation matrices, and cross-tabulations to identify any associations or patterns.

V. Multivariate Analysis: Examine interactions and relationships among multiple variables simultaneously. Techniques like heatmaps, cluster analysis, and dimensionality reduction methods can help in understanding complex relationships.

VI. Feature Engineering: Create new derived variables or transform existing variables to better represent the underlying patterns in the data. This step can involve techniques like scaling, normalization, or creating interaction terms.

VII. Data Visualization: Utilize various visual representations like bar charts, line plots, scatter plots, or heatmaps to visually explore the data and identify trends or patterns.

VIII. Summary Statistics: Calculate summary measures such as means, medians, standard deviations, and quantiles to gain a deeper understanding of the data distribution and characteristics.

IX. Draw Conclusions: Based on the findings from the exploratory analysis, draw meaningful insights, identify limitations, and determine the next steps for further analysis or modeling.

3. Checking for the missing values.

```
# Understanding the Basic Ground Information of My Data
def all_about_my_data(df):
    print("Here is some Basic Ground Info about your Data:\n")

    # Shape of the DataFrame
    print("Number of Instances:",df.shape[0])
    print("Number of Features:",df.shape[1])

    # Summary Stats
    print("\nSummary Stats:")
    print(df.describe())

    # Missing Value Inspection
    print("\nMissing Values:")
    print(df.isna().sum())

all_about_my_data(energy_consumption_df)
```

```
Here is some Basic Ground Info about your Data:

Number of Instances: 5515
Number of Features: 11

Summary Stats:
              Year  Oil Consumption - EJ  Gas Consumption - EJ  \
count  5515.000000           4139.000000           4145.000000
mean   1993.522393           1011.463696            548.258342
std      16.979473           4680.229995           2697.775400
min    1900.000000              0.571998              0.000000
25%    1981.000000             63.741789              8.403757
50%    1995.000000            142.531989             51.675541
75%    2007.500000            413.589372            223.762679
max    2019.000000          53619.924660          39292.467570

       Coal Consumption - EJ  Solar Consumption - EJ  Hydro Consumption - EJ  \
count            4145.000000             4145.000000             4145.000000
mean              737.574797                3.540386              162.425080
std              3686.446745               47.121578              792.287738
min                 0.000000                0.000000                0.000000
25%                 2.916669                0.000000                1.138297
50%                31.854595                0.000000               12.380565
75%               161.238449                0.006497               61.380605
max             44993.467100             1792.996379            10455.126740

       Nuclear Consumption - EJ  Wind Consumption - EJ  Geo Biomass Other - EJ
count               4145.000000            4145.000000             4145.000000
mean                 120.624554              11.684124               12.864392
std                  642.226313             119.280802               78.568993
min                    0.000000               0.000000                0.000000
25%                    0.000000               0.000000                0.000000
50%                    0.000000               0.000000                0.063889
75%                   21.301450               0.101389                3.102351
max                 7493.281090            3540.051401             1614.026957

Missing Values:
Entity                          0
Code                            0
Year                            0
Oil Consumption - EJ         1376
Gas Consumption - EJ         1370
Coal Consumption - EJ        1370
Solar Consumption - EJ       1370
Hydro Consumption - EJ       1370
Nuclear Consumption - EJ     1370
Wind Consumption - EJ        1370
Geo Biomass Other - EJ       1370
dtype: int64
```

Figure 7 : Missing values in the Dataset

4. Now, we first create a copy of the original DataFrame og_df to ensure the original data remains unchanged. Next, we identify and remove rows where the 'country' column has the value 'World' to exclude aggregate or global data, focusing only on country-specific data. This is achieved by filtering the DataFrame to find the relevant index positions and then dropping these rows directly from the copy using the drop method. Finally, we display the first few rows of the modified DataFrame to verify the changes.

```
world_df = og_df.copy()
index_names = world_df[world_df['country'] == 'World'].index
world_df.drop(index_names, inplace = True)
world_df.head()
```

	country	year	oil_consumption	gas_consumption	coal_consumption	solar_consumption	hydro_consumption	nuclear_consumption	wind_consumption
0	Algeria	1965	15.405264	7.430506	0.814101	0.0	1.111112	0.0	0.0
1	Algeria	1966	20.272721	7.719256	0.790841	0.0	0.986112	0.0	0.0
2	Algeria	1967	18.942046	7.488256	0.604760	0.0	1.138890	0.0	0.0
3	Algeria	1968	20.167318	7.873256	0.639651	0.0	1.563890	0.0	0.0
4	Algeria	1969	21.305947	10.351696	0.814101	0.0	1.002779	0.0	0.0

Figure 8 : Removed Aggregate Global Data to Focus on Country-Specific Insights

5. Now we calculate the annual percentage change for each energy resource from 1965 to 2019 and visualizes these changes in bar plots for the period from 1999 to 2019, facilitating the analysis of trends and fluctuations over time.

```
# Calculating Percentage Change for Each Resource Over Years from 1965 to 2019
i = 0
for source in sources:
    values = world_df_bup[source]
    # Calculate and Replace Percentage Change
    world_df_bup[f'Percentage Change in {source} (%)'] = values.pct_change()*100
    # Changing the Figure Size to Make the Axis Readable
    plt.figure(i,figsize=(15,3))
    b = sns.barplot(x='year',y=f'Percentage Change in {source} (%)',data=world_df_bup)
    plt.title(f'Percentage Change in {source} over the World from 1999 to 2019')
    plt.ylim(-10, 100)
    plt.show()
    i+=1
```

Figure 9 : Annual Percentage Change for Each Energy Resource

6. Visualization of the data.

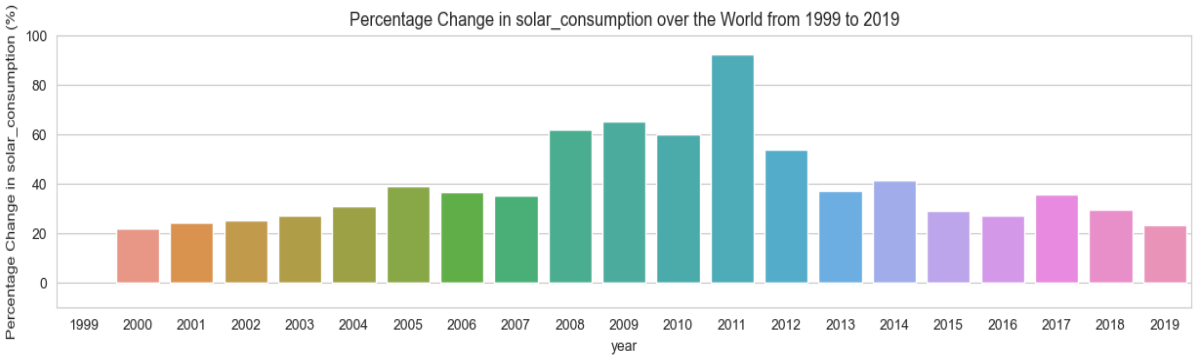


Figure 10 : %Change in Solar_Consumption

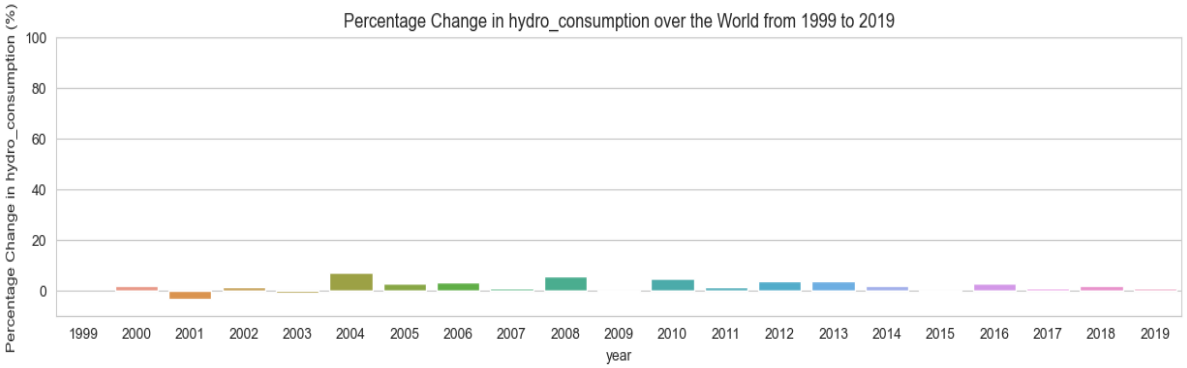


Figure 11 : %Change in Hydro_Consumption

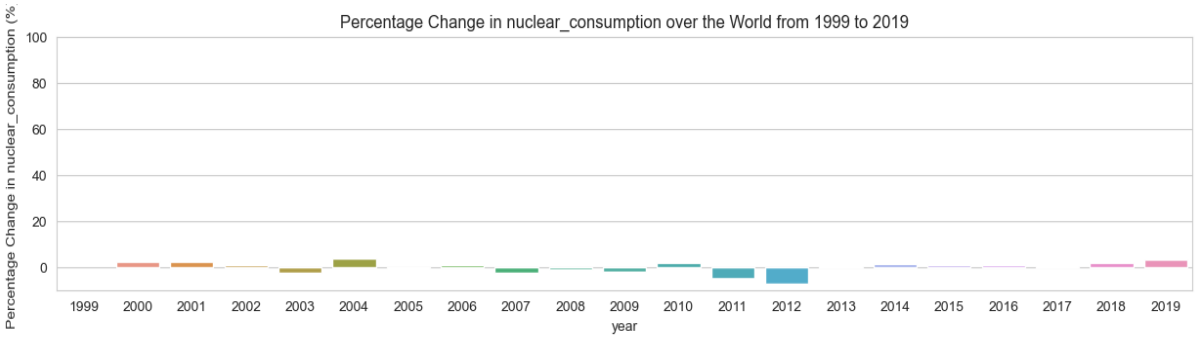


Figure 12 : %Change in Nuclear_Consumption

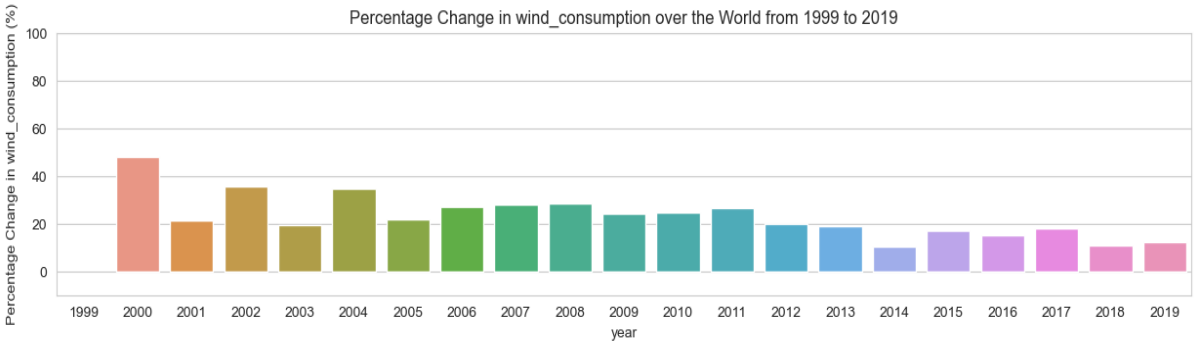


Figure 13 : %Change in Wind_Consumption

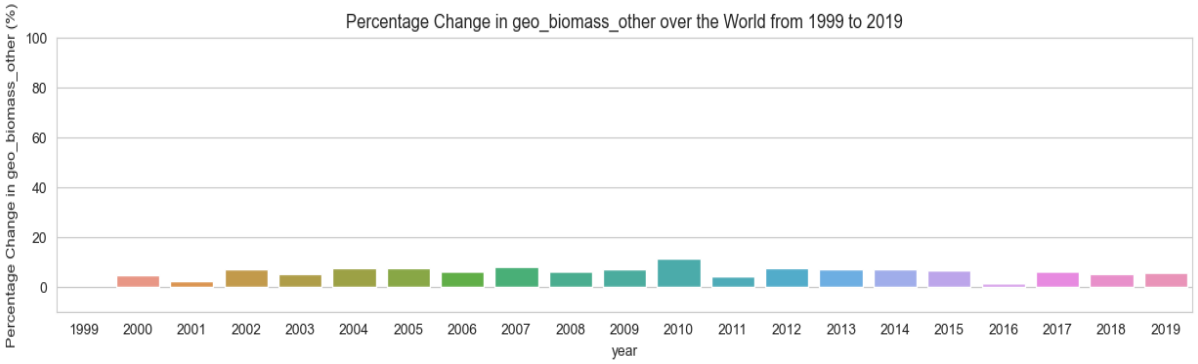


Figure 14 : %Change in Geo_Biomass_Other_Consumption

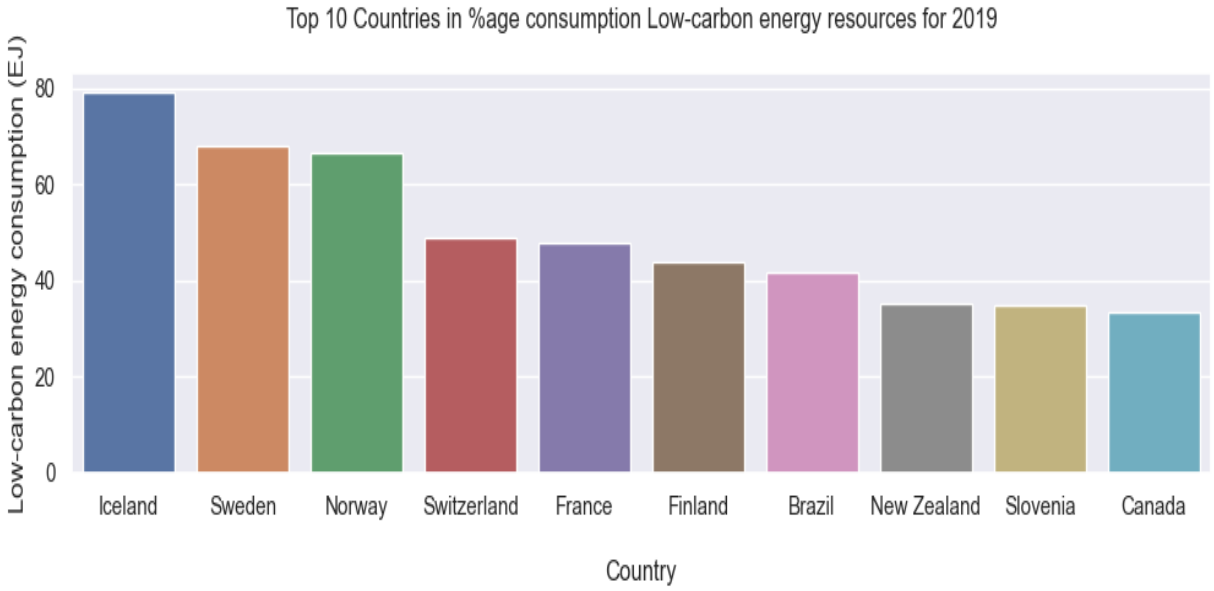


Figure 15 : Top Countries with Maximum Renewable Energy and Nuclear Energy Consumption

Figure 16 : Energy Trends of Small Countries having Big Impact

Figure 17 : Top Energy Consumers

3.3. Algorithms Used

3.3.1. Polynomial Regression

- Polynomial regression is a type of regression analysis where the relationship between the independent variable x x and the dependent variable y y is modeled as an nth degree polynomial.
- This method is useful for capturing non-linear relationships that cannot be effectively represented by a simple linear model.
- It allows for more flexibility by fitting a polynomial curve to the data points, which can represent complex patterns and trends.

Polynomial Function:

- The polynomial function is a mathematical expression that includes terms of the independent variable raised to different powers.
- Unlike linear regression, which fits a straight line to the data, polynomial regression fits a curve that can bend and capture more intricate relationships.
- The degree of the polynomial (the highest power of x x in the equation) determines the flexibility and complexity of the model.

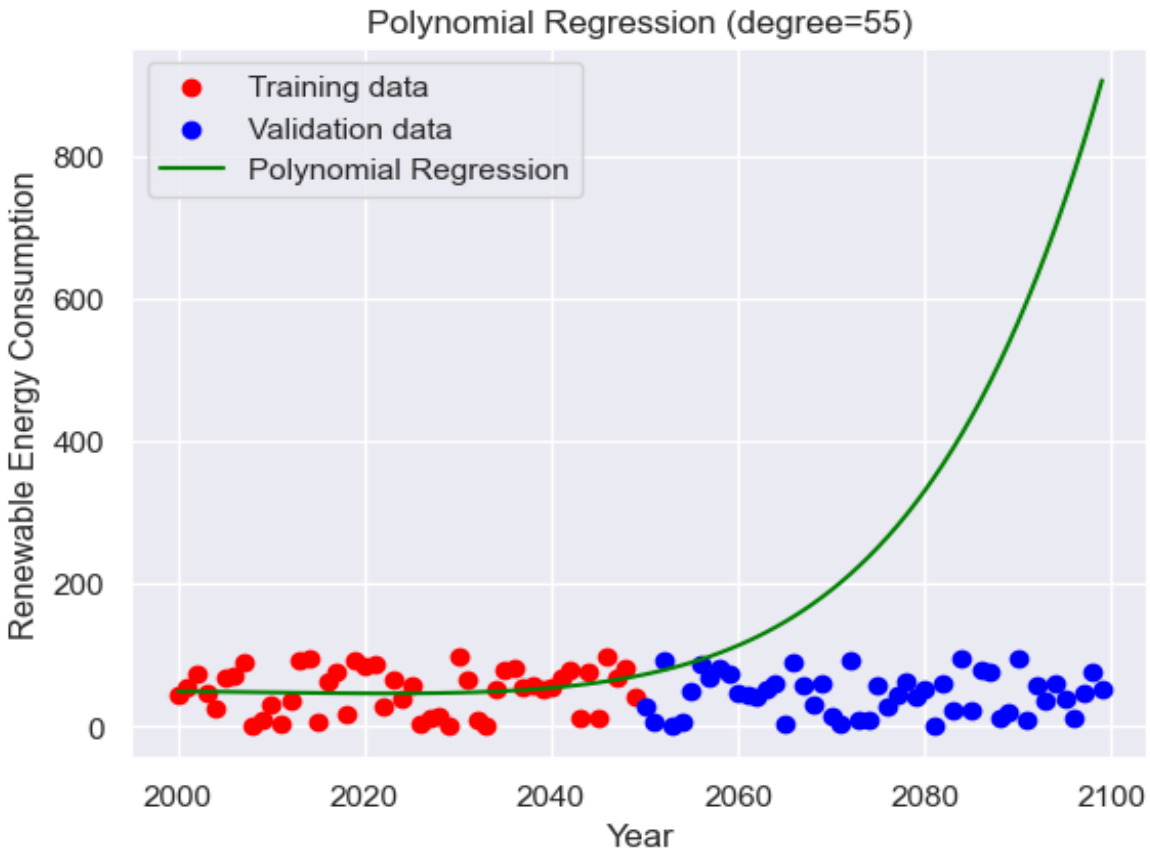


Figure 18 : Polynomial Function

Formula

y = b0 + b1*x + b2*x^2 + …..+bn*x^n

Here,

y: Dependent Variable (Target Variable)

x: Independent Variable (Predictor Variable)

b0: Represents the bias, or intercept

b1, b2, \ldots, bn: Coefficients for the polynomial terms

n: Degree of the polynomial

3.3.2. ARIMA Time Series Model

- The ARIMA (AutoRegressive Integrated Moving Average) model is a popular and powerful tool for time series forecasting.
- It combines three components: Autoregressive (AR), Integrated (I), and Moving Average (MA), to model and predict future points in a time series.
- ARIMA is especially useful for understanding and forecasting time series data that exhibits patterns over time, such as trends and seasonality.

Working of ARIMA Time Series Model

The ARIMA (Autoregressive Integrated Moving Average) model is a powerful and widely used tool for time series forecasting. It combines three main components—Autoregressive (AR), Integrated (I), and Moving Average (MA)—to model and predict future points in a time series. Here's a detailed breakdown of how ARIMA works:

Components of ARIMA

Autoregressive (AR) Component:

- The AR component uses the dependency between an observation and a number of lagged observations (previous time steps).
- It captures the relationship between the current value and the values at prior time steps.
- The order of the AR component is denoted by p p, which indicates the number of lagged observations included in the model.

Integrated (I) Component:

- The I component represents the differencing of raw observations to make the time series stationary, i.e., removing trends or seasonality.
- Differencing is a technique to remove trends by subtracting the previous observation from the current observation.
- The order of differencing is denoted by d d, which indicates how many times the data have been differenced to achieve stationarity.

Moving Average (MA) Component:

- The MA component uses the dependency between an observation and a residual error from a moving average model applied to lagged observations.
- It captures the relationship between the current value and the residual errors from prior time steps.
- The order of the MA component is denoted by q q, which indicates the number of lagged forecast errors included in the model.

Steps to Build an ARIMA Model

Identification:

- Determine if the time series is stationary. If not, apply differencing to make it stationary.
- Use plots like Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) to identify appropriate values for p p (AR order) and q q (MA order).
- Select the order of differencing d d based on the number of times the data need to be differenced to achieve stationarity.

Estimation:

- Estimate the parameters p p, d d, and q q using statistical techniques such as Maximum Likelihood Estimation (MLE).

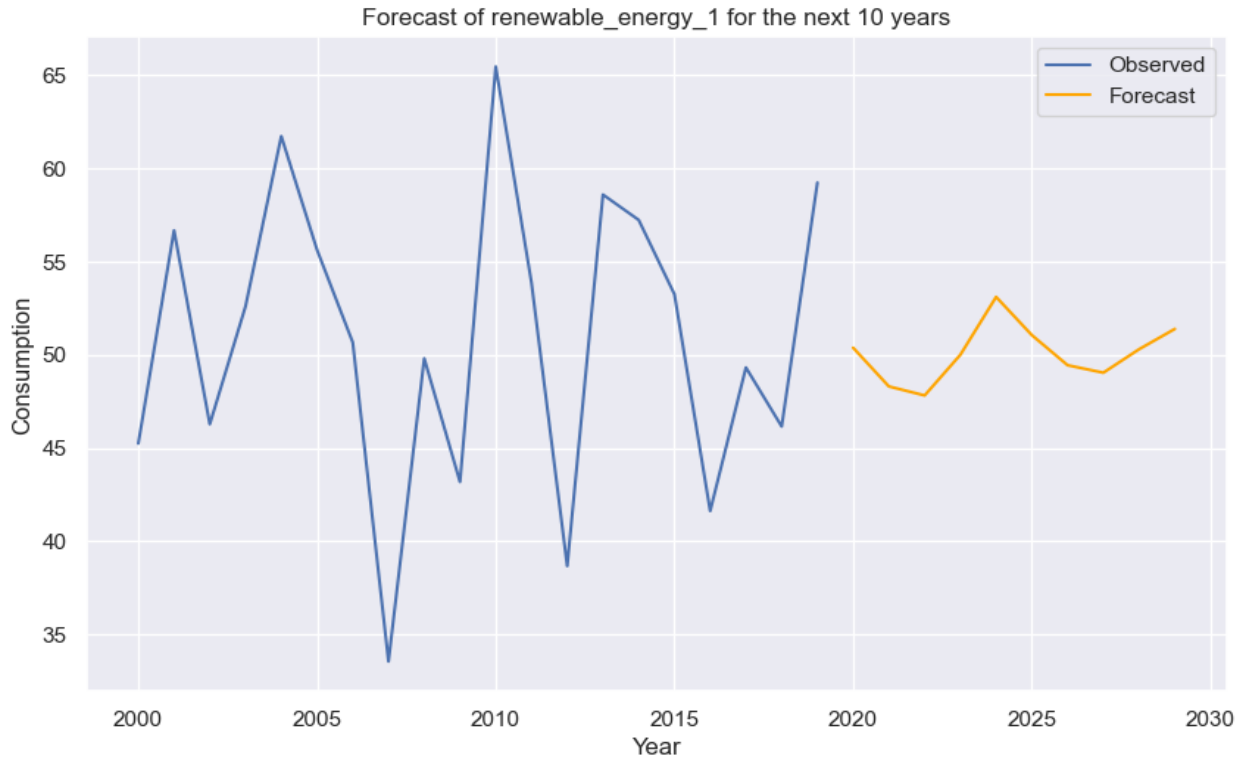
Diagnostic Checking:

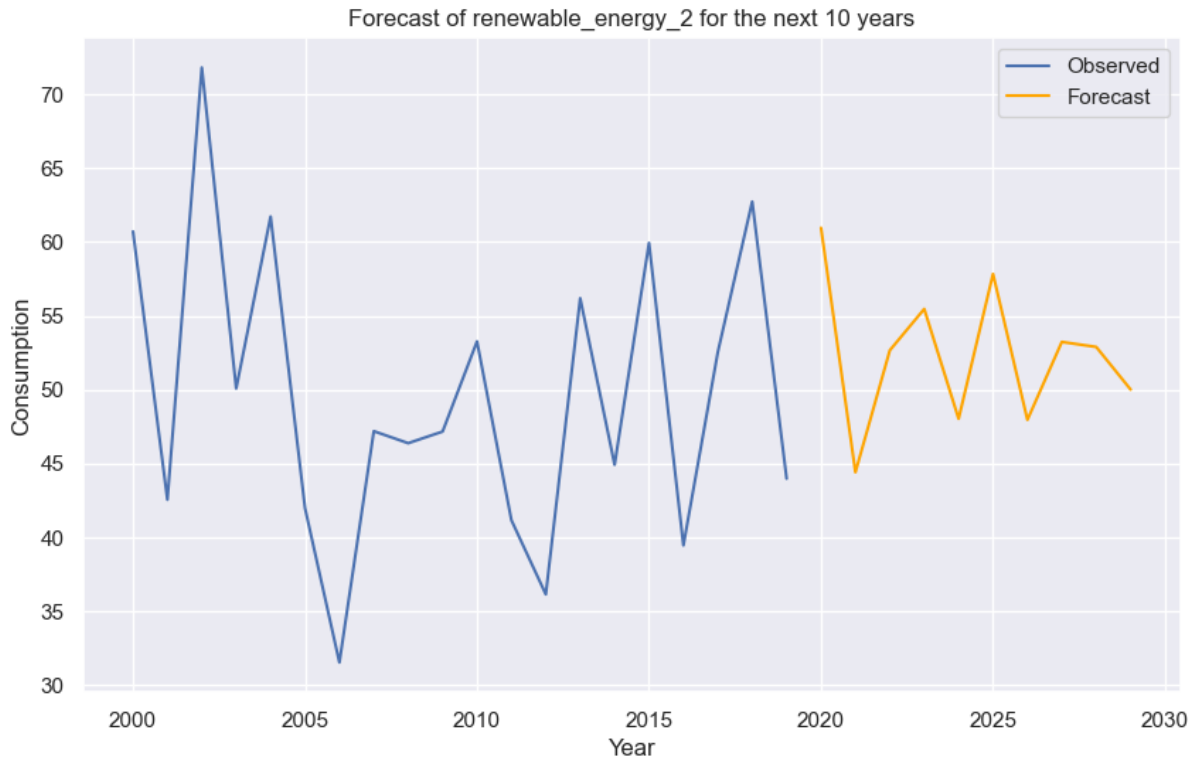
- Analyse the residuals of the fitted model to ensure they resemble white noise (i.e., they are uncorrelated and have a constant mean and variance).
- Perform statistical tests like the Ljung-Box test to validate the adequacy of the model.

Forecasting:

- Use the fitted ARIMA model to forecast future values in the time series.
- Generate forecast intervals to assess the uncertainty in the predictions.

Example: Suppose we have a dataset which can be plotted as follows –





There has been a promising growth of low carbon energies consumption. Assuming that there is a one to one mapping between consumption and production we are extrapolating to say that if consumption of a particular resource goes down then production also goes down, and vice versa. The net estimated fossil fuels consumption can distributed among the low carbon resources to meet the Paris climate goal.

Chapter 4 : Development of the model

Development of the model

Exploratory Data Analysis, Polynomial Regression & ARIMA Time Series Model is used to build a Predictive Model.

4.1. How to apply algorithm in the Project?

Import the necessary libraries: In Python, start by importing the required libraries such ass scikit-learn (sklearn) for machine learning algorithms and NumPy and pandas for data manipulation –

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
from sklearn.preprocessing import PolynomialFeatures
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
from statsmodels.tsa.stattools import adfuller
from statsmodels.tsa.stattools import acf, pacf
from statsmodels.tsa.seasonal import seasonal_decompose
from statsmodels.tsa.arima_model import ARIMA
from matplotlib.pylab import rcParams
from statsmodels.tsa.stattools import adfuller
```

- Load and pre-process the data Load your dataset into a panda DataFrame. Perform necessary pre-processing steps such as handling missing values, encoding categorical variables, or scaling numeric features.
- Split the data into training and testing sets: Split your data into training and testing sets to evaluate the performance of your model.
- Create and train the model: Create an instance of the ARIMA model and train it using the training data.
- Evaluate the model: Evaluate the model's performance on the testing set by calculating metrics such as mean squared error (MSE) and R-squared.

- Use the trained model for predictions: After training the ARIMA model, use it to make predictions on new, unseen data.
- Using Polynomial Regression for Additional Analysis: Apply polynomial regression to capture non-linear relationships in the data.
- Diagnostic Checking and Model Validation: Perform diagnostic checking to ensure model assumptions are met and validate the model using statistical tests.

4.2. Library Used

4.2.1. Pandas

- Pandas is a Python open-sourced library used for working with data sets. It has functions for analyzing, cleaning, exploring, and manipulating data.
- The name "Pandas" has a reference to both "Panel Data", and "Python Data Analysis" and was created by Wes McKinney in 2008.
- Pandas allows us to analyze big data and make conclusions based on statistical theories.
- Pandas can clean messy data sets, and make them readable and relevant. Relevant data is very important in data science.
- Pandas helps in finding correlation between two or more columns, average value, max value, min value, and many more things.
- Pandas are also able to delete rows that are not relevant, or contains wrong values, like empty or NULL values. This is called cleaning the data.

4.2.2. Numpy

- NumPy is a widely used Python library that is primarily used for working with arrays. It was created by Travis Oliphant in 2005 and is an open-source project. The name "NumPy" is short for "Numerical Python."
- One of the main advantages of NumPy is its efficient handling of arrays. In Python, lists can be used as arrays, but they can be slow for certain operations. NumPy provides the ndarray (n-dimensional array) object, which is designed to be faster and more efficient than traditional Python lists. The ndarray object comes with a wide range of supporting functions, making it easy to perform various operations on arrays.

- In addition to basic array manipulation, NumPy also offers functions for linear algebra, Fourier transforms, and matrix operations. This makes it a valuable tool for scientific computing and data analysis tasks.
- NumPy's array object is commonly used in data science due to its speed and efficient memory utilization. It allows for faster computation and better resource management, which is crucial when working with large datasets or performing complex mathematical operations.
- Overall, NumPy provides a powerful array object and a comprehensive set of functions that are essential for numerical computing in Python. Its versatility and performance make it a popular choice among data scientists and researchers.

4.2.3. Matplotlib

- Matplotlib is a low-level graph plotting library in python that serves as a visualization utility.
- It was created by John D. Hunter and is open source and we can use it freely.
- Matplotlib is mostly written in python, a few segments are written in C, Objective-C and JavaScript for Platform compatibility.

4.2.4. Sklearn

Scikit-learn, also known as sklearn, is a powerful and widely used library for machine learning in Python. It offers a wide range of efficient tools and algorithms for tasks such as classification, regression, clustering, and dimensionality reduction. Scikit-learn provides a consistent and user-friendly interface for implementing these machine learning models.

Some important features of scikit-learn:

- Comprehensive set of algorithms: Scikit-learn includes a vast collection of machine learning algorithms. It covers popular algorithms such as support vector machines (SVM), random forests, gradient boosting, k-means clustering, and many more. This allows users to choose the most appropriate algorithm for their specific task.
- Ease of use: Scikit-learn is designed to be user-friendly and accessible to everyone. It provides a simple and intuitive API that allows users to quickly

implement machine learning models without extensive programming knowledge. The library also offers detailed documentation and examples to help users get started easily.

- Extensibility and interoperability: Scikit-learn is built on top of other popular Python libraries such as NumPy, SciPy, and matplotlib. This makes it easy to integrate with other data processing and visualization tools in the Python ecosystem. It also allows users to leverage the functionalities of these libraries along with scikit-learn for more advanced data analysis tasks.
- Consistency and reusability: Scikit-learn follows a consistent interface for all its algorithms, making it easy to switch between different models without significant changes to the code. This consistency enables reusability of code and promotes good software engineering practices in machine learning projects.
- Performance and scalability: Scikit-learn is designed to be efficient and scalable. It implements optimized algorithms and data structures, making it suitable for handling large datasets and complex computations. Additionally, scikit-learn provides tools for model selection, feature extraction, and preprocessing, which contribute to the overall performance of machine learning workflows.

4.2.5. Seaborn

- Seaborn is a Python visualization library that is specifically designed for creating statistical plots. It builds upon the functionality of the matplotlib library and integrates closely with pandas, a popular data manipulation library in Python.
- The main goal of Seaborn is to enhance the visual appeal of statistical graphs and make them more engaging. It achieves this by providing a wide range of attractive default styles and color palettes that can be easily applied to plots. These aesthetic enhancements make it easier to communicate and interpret data visually.
- One of the key features of Seaborn is its dataset-oriented approach to plotting. It offers a high-level API that allows users to create various types of statistical visualizations, such as scatter plots, bar plots, box plots, and heatmaps, with just a few lines of code. This flexibility enables users to explore different visual representations of the same variables, facilitating a deeper understanding of the underlying data.

- Seaborn also offers advanced features for statistical visualization, such as the ability to incorporate additional dimensions through color or size encoding, visualize patterns in categorical data, and create complex multi-plot grids for exploring relationships between multiple variables.
- Overall, Seaborn is a powerful visualization library that simplifies the creation of statistical plots in Python. Its integration with pandas and matplotlib, along with its emphasis on aesthetics and dataset-oriented plotting, makes it a popular choice for data analysts and scientists looking to effectively communicate insights from their data.

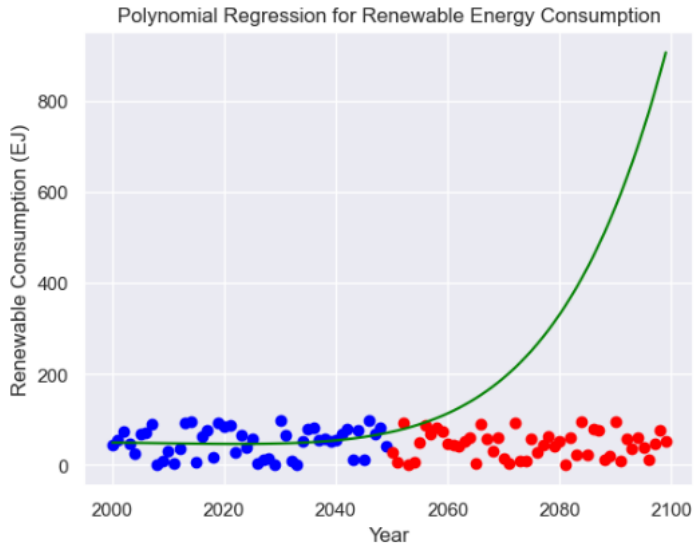
4.2.6. Statsmodels

- Statsmodels is a Python library for estimating and testing statistical models.
- It supports various models including linear regression, generalized linear models, and time series analysis models like ARIMA.
- Statsmodels integrates seamlessly with pandas and NumPy data structures.
- It provides tools for seasonal decomposition, autocorrelation, and partial autocorrelation functions.
- Statsmodels offers detailed model summaries, statistical metrics, parameter estimates, and diagnostic tools.
- It integrates with matplotlib and seaborn for visualizing results and diagnostics, supporting effective communication of insights.

Chapter 5 : Result

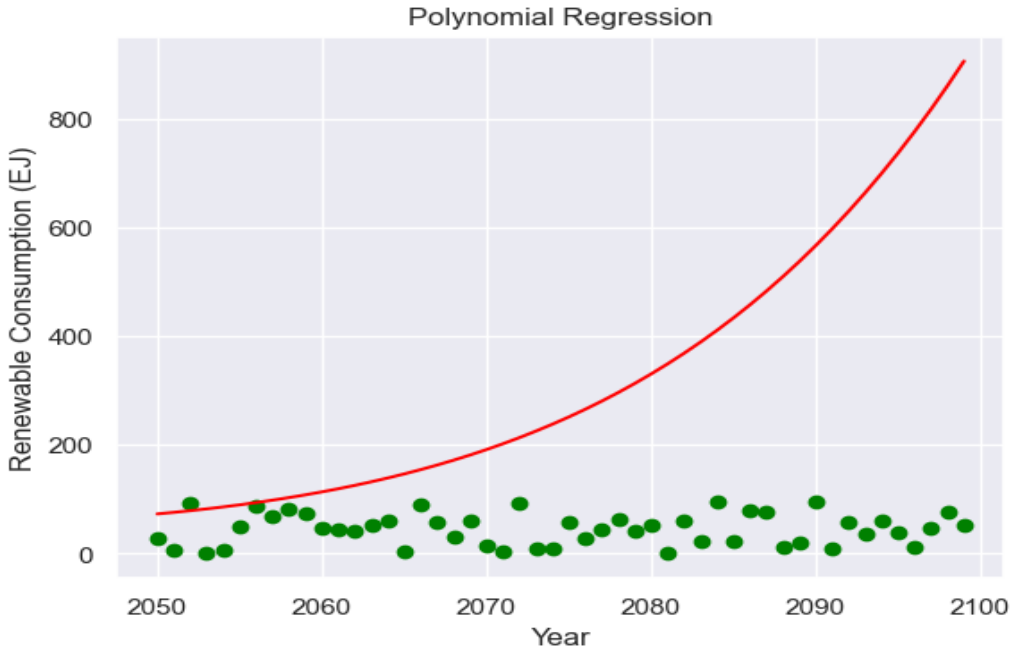
5.1. Result Of Polynomial Regression Model

```
# Visualising the Polynomial Regression Training Results
# Train
plt.scatter(X_train, y_train, color = 'blue')
# Test
plt.scatter(X_val, y_val, color = 'red')
plt.plot(YEARS, lin2.predict(poly.fit_transform(YEARS)), color = 'green')
plt.title('Polynomial Regression for Renewable Energy Consumption')
plt.xlabel('Year')
plt.ylabel('Renewable Consumption (EJ)')

plt.show()
```




```
# Fitting Polynomial Regression to the Dataset
degree = 55
poly = PolynomialFeatures(degree=degree)
X_poly_train = poly.fit_transform(X_train)

lin2 = LinearRegression()
lin2.fit(X_poly_train, y_train)

# Transforming the Validation Set
X_poly_val = poly.transform(X_val)

# Predicting on the Validation Set
y_pred = lin2.predict(X_poly_val)
```

```
# Visualising the Polynomial Regression Validation Results
plt.scatter(X_val, y_val, color = 'green')

plt.plot(X_val, y_pred, color = 'red')
plt.title('Polynomial Regression')
plt.xlabel('Year')
plt.ylabel('Renewable Consumption (EJ)')

plt.show()
```




```
# Evaluating RMSE Score
rmse = np.sqrt(mean_squared_error(y_val,y_pred))
# Evaluating r2 Score
r2 = r2_score(y_val,y_pred)
print(rmse)
print(r2)
```

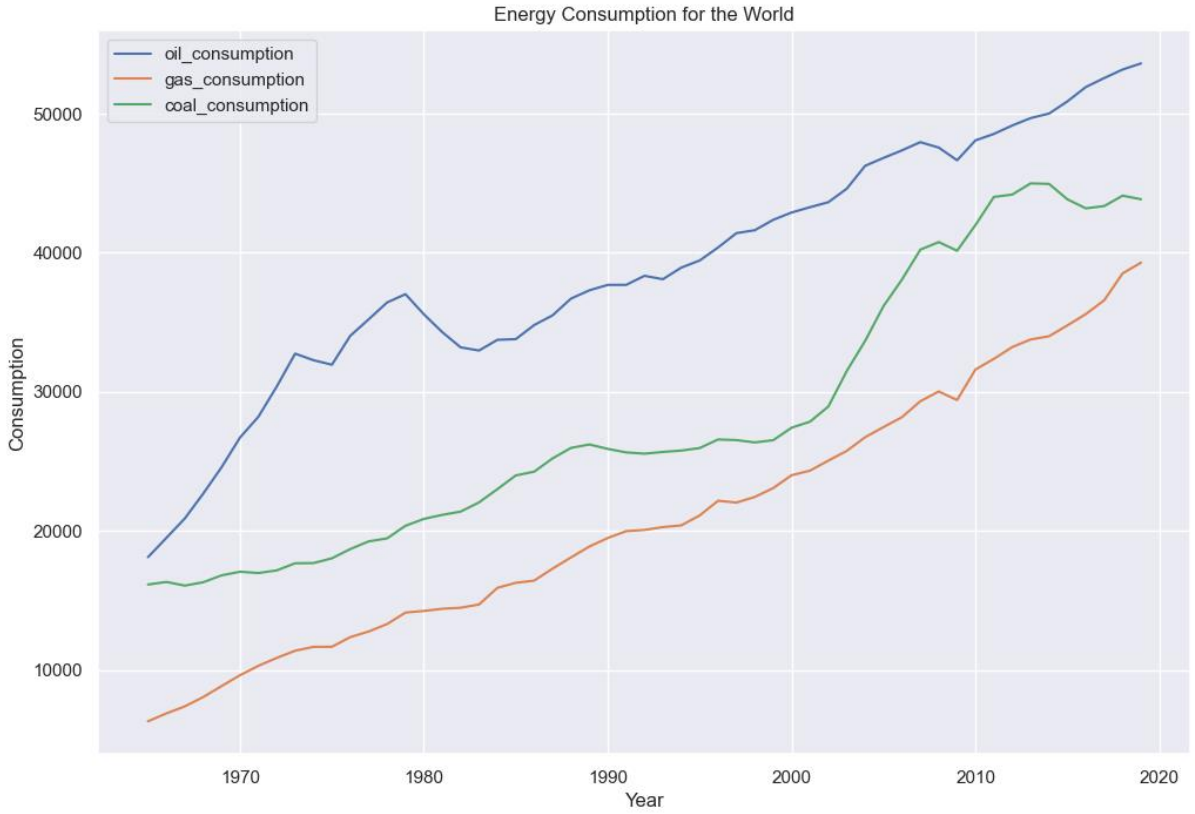
```
368.8223135653636
-163.82787638658638
```

Our R2 score is close to 1, but RMSE score shows a lot of deviation in the predicted values in Exajoules. We see an overfitting here. It also depends on the fact that we only have data for the past 55 years.

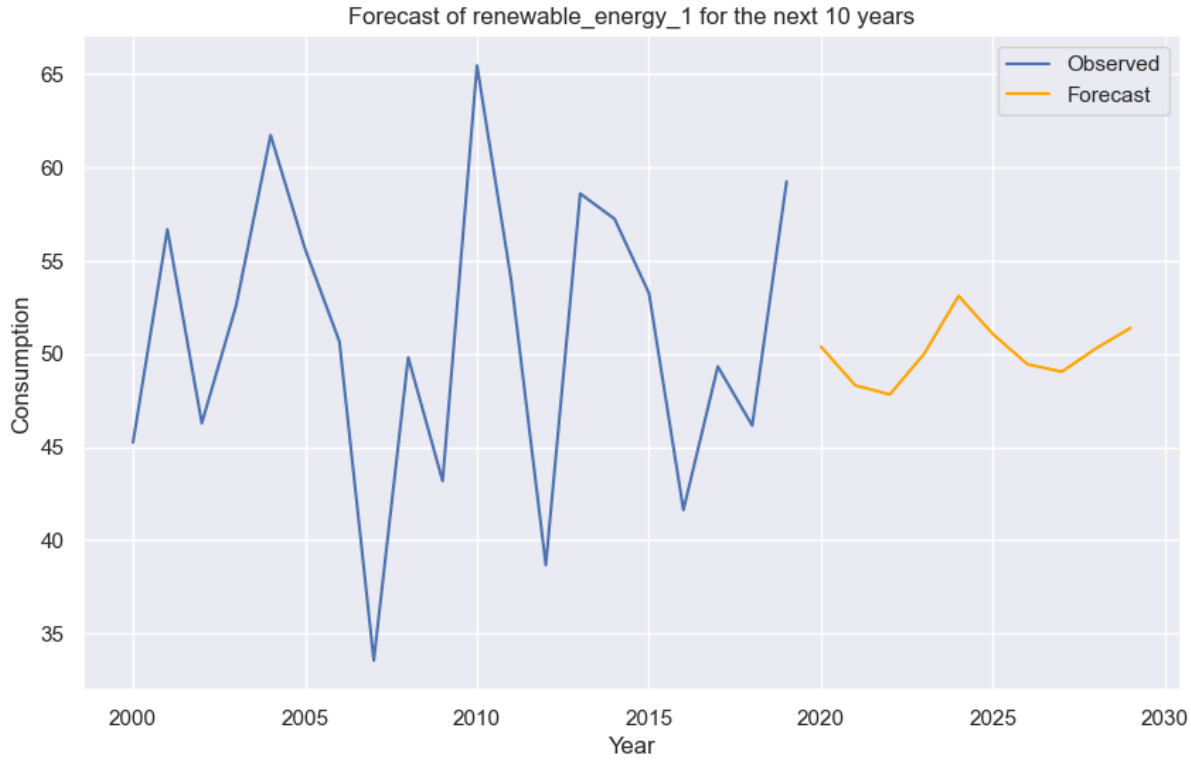
5.2. Result of ARIMA Time Series Model

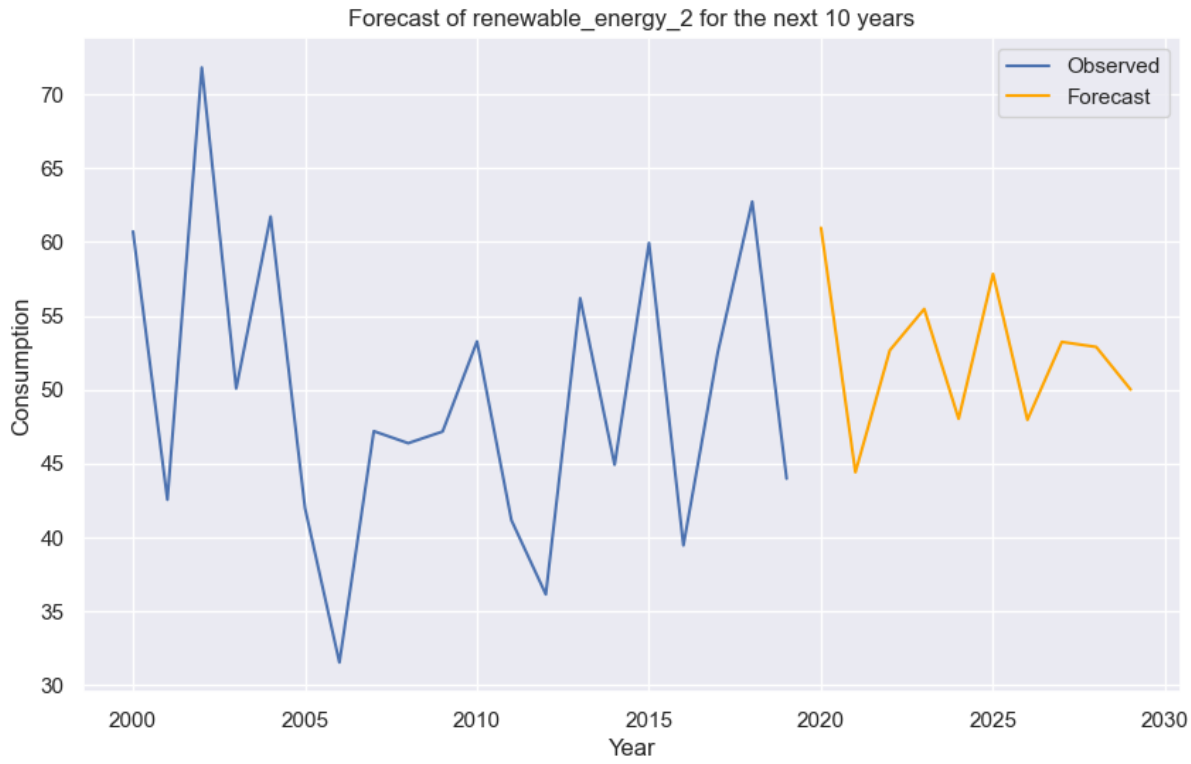
```
import matplotlib.pyplot as plt

# List of Energy Sources
energy_sources = ['oil_consumption', 'gas_consumption', 'coal_consumption']

# Print the Columns of the DataFrame to Check for Discrepancies
print("Columns in DataFrame:", world_consumption_df.columns)

# Initialize the Plot
plt.figure(figsize=(12, 8))

# Loop through Each Energy Source
for energy in energy_sources:
    if energy in world_consumption_df.columns:
        # Plot the Average Consumption of the Energy Source Per Year
        plt.plot(world_consumption_df.groupby('year')[energy].mean(), label=energy)
    else:
        print(f"Warning: Column '{energy}' not found in the DataFrame")

# Add Title and Labels
plt.title('Energy Consumption for the World')
plt.xlabel('Year')
plt.ylabel('Consumption')

# Add a Legend
plt.legend()

# Show the Plot
plt.show()
```

```
Columns in DataFrame: Index(['country', 'oil_consumption', 'gas_consumption', 'coal_consumption',
       'solar_consumption', 'hydro_consumption', 'nuclear_consumption',
       'wind_consumption', 'geo_biomass_other'],
      dtype='object')
```



```python
# Sample Data Creation (Remove this Part if your DataFrame is Already Defined)
data = {
    'year': [2000 + i // 5 for i in range(100)],
    'renewable_energy_1': np.random.rand(100) * 100,  # Random Data for Illustration
    'renewable_energy_2': np.random.rand(100) * 100
}
world_consumption_df = pd.DataFrame(data)

# Define the Types of Renewable Energy
renewable_energy = ['renewable_energy_1', 'renewable_energy_2']

# Number of Years to Predict
number_of_years_to_predict = 10

# Forecast the Next 10 Years
energy_results_ARIMA = get_forecast(world_consumption_df, renewable_energy, number_of_years_to_predict, plot=True)

# Estimate Total Consumption for Each Energy Type
tot_cons_renewable_energy_wise = get_energy_consumption_estimation(number_of_years_to_predict, energy_results_ARIMA)

print(tot_cons_renewable_energy_wise)
```





5.4. Findings

There has been a promising growth of low carbon energies consumption. Assuming that there is a one to one mapping between consumption and production we are extrapolating to say that if consumption of a particular resource goes down then production also goes down, and vice versa. The net estimated fossil fuels consumption can distributed among the low carbon resources to meet the Paris climate goal.

Conclusion

In conclusion, the project on global energy estimation has provided valuable insights into the prediction and analysis of energy prices worldwide. By leveraging advanced statistical and machine learning techniques, we have explored historical data, identified trends, and built models capable of forecasting future energy prices. The integration of various methodologies such as machine learning algorithms, time series analysis, and statistical modeling, particularly using tools like ARIMA and polynomial regression, has enhanced our understanding of energy market dynamics.

Through the application of these models, we have demonstrated their effectiveness in capturing and predicting complex patterns in energy pricing. Moreover, tools like Seaborn for visualization and Statsmodels for statistical analysis have played pivotal roles in interpreting and communicating our findings. This project underscores the importance of data-driven approaches in decision-making within the energy sector, offering stakeholders valuable insights for strategic planning and risk management.

Looking ahead, further refinement and validation of these models could enhance their predictive accuracy, supporting more informed decision-making in the face of evolving energy markets. As global energy demands continue to shift, the methodologies and insights developed in this project will remain crucial for navigating the complexities of the energy landscape in the years to come.

Chapter 6 : References

References

1. International Energy Agency (IEA). World Energy Outlook. Retrieved from [IEA Official Website].

2. U.S. Energy Information Administration (EIA). International Energy Outlook. Retrieved from [EIA Official Website].

3. BP. BP Statistical Review of World Energy. Retrieved from [BP Official Website].

4. Global Energy Assessment (GEA). Retrieved from [GEA Official Website].

5. Renewable Energy Policy Network for the 21st Century (REN21). Renewables Global Status Report. Retrieved from [REN21 Official Website].

6. "Energy Policy" journal. Available at [Publisher's Website].

7. "Energy Economics" journal. Available at [Publisher's Website].

8. "Renewable and Sustainable Energy Reviews" journal. Available at [Publisher's Website].

9. Schobert, H.H. Energy and Society: An Introduction. Publisher, Year.

10. Smil, V. Energy and Civilization: A History. Publisher, Year.

Analysis & Estimation of Global Energy Prediction

by Kartikey Chaurasia

Submission date: 31-May-2024 07:33PM (UTC+0530)
Submission ID: 2392529598
File name: ysis_Estimation_of_Global_Energy_Prediction_Report_Kartikey.pdf (1.82M)
Word count: 5462
Character count: 32702

Analysis & Estimation of Global Energy Consumption

ORIGINALITY REPORT

28%	23%	10%	20%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	fastercapital.com Internet Source	5%
2	www.cemca.org Internet Source	1%
3	open-innovation-projects.org Internet Source	1%
4	www.coursehero.com Internet Source	1%
5	www.imensosoftware.com Internet Source	1%
6	Submitted to Istanbul Aydin University Student Paper	1%
7	Submitted to University of Hertfordshire Student Paper	1%
8	Submitted to National Institute of Business Management Sri Lanka Student Paper	1%
9	Submitted to Bradley University Student Paper	1%