ENERGYVISTA: A COMPREHENSIVE EXPLORATION OF THE ENERGY MARKET

MINOR PROJECT REPORT

By

AYUSH KUMAR (RA2211056010011)
MOHAMED IRSATH (RA2211056010055)
ANISH SARKAR (RA2211056010056)

Under the guidance of

DR. M.ANAND

In partial fulfilment for the Course

of

21CSS202T – FUNDAMENTALS OF DATA SCIENCE

in **Data Science and Business Systems**

FACULTY OF ENGINEERING AND TECHNOLOGY

SCHOOL OF COMPUTING

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

KATTANKULATHUR

NOVEMBER 2023

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

(Under Section 3 of UGC Act, 1956)

BONAFIDE CERTIFICATE

Certified that this minor project report for the course **21CSS202T FUNDAMENTALS OF DATA SCIENCE** entitled in "**EnergyVista: A Comprehensive Exploration of the Energy Market**" is the bonafide work of **Ayush Kumar (RA2211056010011) , Mohamed Irsath (RA2211056010055)** and **Anish Sarkar (RA2211056010056)** who carried out the work under my supervision.

SIGNATURE

Dr. M.Anand
Assistant Professor
Data Science and Business Systems
SRM Institute of Science and Technology
Kattankulathur

ABSTRACT

" EnergyVista: A Comprehensive Exploration of the Energy Market ” employs Python libraries to comprehensively analyze energy production and consumption trends. The project explores datasets, visualizes top energy-producing and consuming countries, and utilizes choropleth maps to highlight geographical distribution. Seasonal decomposition uncovers temporal patterns, while resource-specific analyses delve into coal, natural gas, nuclear, and petroleum production. A key component involves XGBoost for time series forecasting, predicting future energy consumption. Feature importance analysis reveals influential factors. The project concludes with actionable insights, presented in an interactive dashboard. The code is well-documented for clarity and adaptability.

"EnergyVista" is a vital tool, providing insights for policymakers and industry stakeholders, guiding decisions toward a sustainable and resilient energy future.

Keywords: Energy Production, Energy Consumption, Time Series Forecasting, Resource Analysis, Choropleth Maps, Sustainability, XGBoost, Global Energy Dynamics.

ACKNOWLEDGEMENT

We express our heartfelt thanks to our honorable **Vice Chancellor Dr. C. Muthamizhchelvan**, for being the beacon in all our endeavors.

We would like to express my warmth of gratitude to our **Registrar Dr. S. Ponnusamy,** for his encouragement.

We express our profound gratitude to our **Dean (College of Engineering and Technology) Dr. T. V. Gopal,** for bringing out novelty in all executions.

We would like to express my heartfelt thanks to Chairperson, School of Computing **Dr. Revathi Venkataraman,** for imparting confidence to complete my course project

We wish to express my sincere thanks to **Course Audit Professors Dr. Vadivu. G, Professor, Department of Data Science and Business Systems** for their constant encouragement and support.

We are highly thankful to my Course project Faculty **Dr.Anand , Assistant Professor , Department of Data Science and Business Systems** for his assistance, timely suggestion and guidance throughout the duration of this course project.

We extend my gratitude to our **HoD Dr. M Lakshmi, Professor, Department of Data Science and Business Systems** and my Departmental colleagues for their Support.

Finally, we thank our parents and friends near and dear ones who directly and indirectly contributed to the successful completion of our project. Above all, I thank the almighty for showering his blessings on me to complete my Course project.

TABLE OF CONTENTS

1. INTRODUCTION

In an era defined by the imperative for sustainable energy practices, understanding the complexities of global energy dynamics is crucial. "EnergyVista" stands as a robust Python-based tool, dedicated to a thorough exploration of extensive energy datasets spanning multiple years.

This project focuses on visualizing energy production and consumption patterns, unraveling temporal trends, conducting resource-specific analyses, and employing time series forecasting. By leveraging advanced analytical methods, EnergyVista seeks to illuminate the geographical distribution of energy production, temporal influences on global energy trends, and the role of diverse resources in shaping the energy market. The integration of XGBoost for forecasting adds a predictive dimension, offering glimpses into future consumption patterns.

Through actionable insights, EnergyVista aims to empower decision-makers, industry stakeholders, and researchers to make informed choices that promote sustainability, optimize resource utilization, and contribute to a resilient global energy landscape.

As we delve into the intricacies of EnergyVista, we embark on a journey to uncover hidden narratives within the data, offering a condensed yet comprehensive view of global energy dynamics and guiding us toward a sustainable and resilient future.

1.1 Motivation:

The driving force behind the "EnergyVista" project is a commitment to address key challenges in the global energy landscape. Motivated by the imperative for sustainability and efficient resource use, the project aims to analyze production and consumption trends. By identifying opportunities to reduce overproduction and minimize energy wastage, the project seeks to provide actionable insights for strategic decision-making. Ultimately, it aspires to contribute to a more sustainable and balanced global energy ecosystem.

1.2 Objectives:

The primary goal of the "EnergyVista" project is to conduct a thorough analysis of global energy production and consumption trends. The project aims to identify the existing disparities between production and consumption, with a specific focus on mitigating overproduction and minimizing energy wastage. Through advanced data analysis and time series forecasting, the project endeavors to provide insights that empower decision-makers to optimize energy resource allocation, promote sustainability, and formulate strategies to curtail unnecessary energy production. Ultimately, the goal is to contribute to a more efficient and sustainable global energy landscape.

1.3 Problem Statement:

The global energy landscape confronts challenges of imbalances between production and consumption, leading to overproduction and wastage. Predictive insights for strategic decision-making are lacking, while the sustainability imperative presses for more efficient resource utilization. The "EnergyVista" project aims to address these issues by analyzing global energy trends, mitigating overproduction, and offering forecasts to guide sustainable practices, ultimately contributing to a balanced and resilient energy future.

1.4 Challenges:

The "EnergyVista" project faces key challenges, including managing data variability, ensuring predictive accuracy in forecasting models, navigating resource-specific complexities, incorporating interdisciplinary perspectives, effectively communicating findings to stakeholders, and integrating advanced data science techniques. Overcoming these challenges is essential for the project's success in providing valuable insights into global energy dynamics and sustainability.

2. LITERATURE SURVEY

In this section, we discuss the previous research that has been conducted in an effort to Forecast Electricity Consumption. The information present in this section includes Data Cleaning, Data preprocessing, Data Visualisation, machine learning algorithms, used in our study.

- In the paper "A Survey on Electric Power Consumption Prediction Techniques" the publishers said that The prediction process of electric power consumption is divided into three categories based on time period of prediction. They are short-term prediction, medium-term prediction and long-term prediction. Short-term prediction: The time-period of short-term prediction takes for an hours to one-day ahead or a week. It aims at economic dispatch, optimal generator unit commitment, power distribution and load dispatching while addressing realtime control and security assessment. Mid-term prediction: The time-period of mid-term prediction is few weeks to a few months. The purpose of this type of prediction is to maintain system, purchasing energy, and price settlement so that demand and generation is balanced. Long-term prediction: The time-period of long-term prediction is a year to 10-20 years ahead. It aims at system expansion planning, i.e., generation, transmission and distribution. This prediction can also affect the purchase of new generating units.

Since in our project we are going to use Linear Regression in past 10 years hourly data, This is Long-term prediction.

- Desley Vine, Laurie Buys, Peter Morris (2013) this paper describes about the electricity consumption and to identify the prospective electricity response in residential consumer's in usage of electricity models. The different response criteria of the residential customer's are recognized, overall conservation, peak demand reduction and its effectiveness highlighted.

- Rahimi, N., Park, S., Choi, W., Oh, B., Kim, S., Cho, Y. H., ... & Lee, D. (2023), In this article the objective is to provide a comprehensive review of ensemble solar power forecasting algorithms. The study aims to evaluate the performance of different ensemble methods in improving the accuracy of solar power forecasting. The article discusses various ensemble methods used in solar power forecasting, including simple averaging, weighted averaging, bagging, boosting, and stacking. The review also discusses the advantages and disadvantages of each method and provides insights into the factors that affect the accuracy of ensemble forecasting models. The article concludes by highlighting the importance of ensemble methods in improving the accuracy of solar power forecasting.

- Duke Ghosh and Joyashree Roy (2011) describes that it is a practice in India to engage consultants in firms. The review is about the usage and ways to improve the energy efficency. Further investigation suggest that the majority of these firms have either implemented the process to reduce the costs associated with energy cosumption or to ensure uninterrupted power supply.

- VijayaMohan Pillai.N (2001) describes about the forecasting of energy consumption. The number of continuous energy problems has made the need for accurate projection of electricity demand and the importance of the forecasting methods. The method proposes about class time series framework for forecasting.

3. REQUIREMENTS

3.1 Requirement Analysis

From the given scenario, we draw the following requirements:

1. Ensure access to reliable and up-to-date global energy datasets, with a focus on electricity data. Implement data cleaning procedures to maintain high data quality.
2. Utilize pandas and plotly for efficient data analysis and visually compelling comparisons of energy production and consumption across countries..
3. Implement XGBoost for accurate time series forecasting, fine-tuning the model for optimal performance.
4. Derive actionable insights into energy conservation opportunities through a combination of literature analysis and dataset exploration.
5. Use plotly to create clear and concise visualizations, aiding effective communication of key findings and recommendations to diverse stakeholders.
6. Develop a modular project structure to ensure flexibility and scalability, accommodating future enhancements and additional datasets.

3.2 Hardware Requirement

From the given scenario, we draw the following requirements:

1. Install Python, preferably the latest version.
2. Install necessary libraries and packages using a package manager like pip.
   pandas, plotly, xgboost, statsmodels, scikit-learn, matplotlib, seaborn, country_converter, calendar, datetime
3. Obtain the required energy datasets in a format compatible with pandas (e.g., CSV, Excel).
4. Use Jupyter Notebook or a Python Integrated Development Environment (IDE) for code execution.
5. Train the XGBoost model on historical energy data.

4. ARCHITECTURE AND DESIGN

1. Data Collection Module:

   Design: Implement scripts or functions to collect data from diverse sources, emphasizing electricity data. Ensure data integrity during the collection process.

2. Data Cleaning Module:

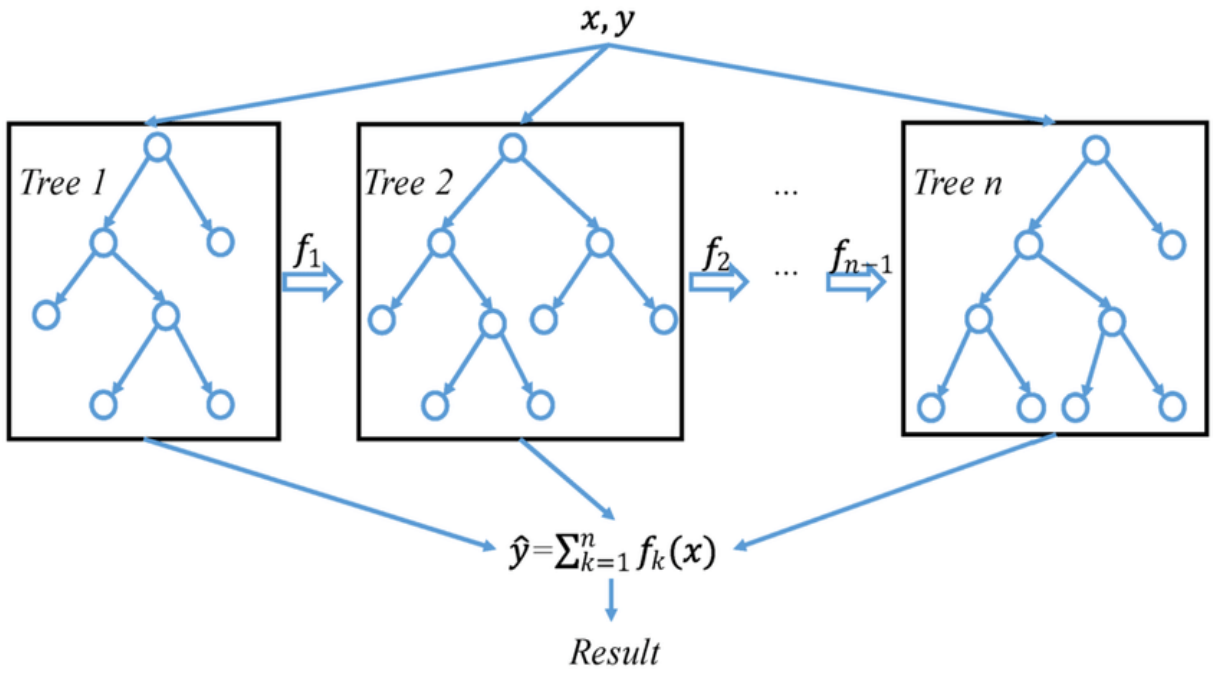
   Design: Develop procedures or functions to clean and preprocess the collected data. Address inconsistencies, handle missing values, and conduct quality checks.

3. EDA Module:

   Design: Utilize pandas and plotly for data exploration. Create functions for generating exploratory visualizations, facilitating in-depth analysis of energy datasets.

4. XGBoost Forecasting Module:

   Design: Implement XGBoost models for time series forecasting. Fine-tune hyperparameters and conduct model training on historical energy data.



5. Insights Derivation Module:

   Design: Combine literature analysis and dataset exploration results to derive actionable insights. Develop functions to extract key patterns and correlations.

6. Plotly Visualization Module:

   Design: Leverage plotly for creating interactive and clear visualizations. Design functions

to generate various plots, charts, and graphs for effective communication.

7. Reporting Module:

   Design: Compile visualizations and key findings into a comprehensive report or presentation. Design user-friendly interfaces for stakeholder interaction.

8. Testing and Validation Module:

   Design: Develop testing scripts to validate the reliability and accuracy of the forecasting model. Conduct thorough validation against historical data.

   Internet Connectivity:

5. IMPLEMENTATION

Data Preprocessing:

In [1]:
```
from pandas.plotting import parallel_coordinates
import plotly.express as px
from functools import reduce
from plotly.subplots import make_subplots
import plotly.graph_objects as go
import plotly.figure_factory as ff
import plotly.io as pio
from plotly.offline import plot
```

In [2]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import country_converter as coco
import calendar
import datetime
```

In [3]:
```
data = pd.read_excel('World Energy Overview.xlsx')
data['Date'] = pd.to_datetime(data['Date'])

production = pd.read_csv('Production_Data/Production_Total.csv')
coal_production = pd.read_csv('Production_Data/Production_Coal.csv')
NaturalGas_production = pd.read_csv('Production_Data/Production_NaturalGas.csv')
Neuclear_production = pd.read_csv('Production_Data/Production_Nuclear+renewables.csv')
Petrolium_production = pd.read_csv('Production_Data/Production_Pertroleum.csv')

consumption = pd.read_csv('Consumption_Data/Consumption_Total.csv')
coal_consumption = pd.read_csv('Consumption_Data/Consumption_Coal.csv')
NaturalGas_consumption = pd.read_csv('Consumption_Data/Consumption_NaturalGas.csv')
Neuclear_consumption = pd.read_csv('Consumption_Data/Consumption_Neuclear+renewables.csv')
Petrolium_consumption = pd.read_csv('Consumption_Data/Consumption_Petroleum.csv')
```

In [4]:
```
data.head()
```

Production Analysis:

In [14]:
```
from statsmodels.tsa.seasonal import seasonal_decompose
results = seasonal_decompose(total_production['Total_Production'])

fig_1 = px.line(y=results.observed, x=results.observed.index, title='Global Energy Production Rate',width=850, height=450)
fig_1.update_xaxes(rangeslider_visible=True)
fig_1.update_xaxes(title_text='Date')  # Set x-axis label
fig_1.update_yaxes(title_text='Total Production')  # Set y-axis label
fig_1.update_xaxes(rangeslider_visible=True)
fig_1.show()

# Figure 2
fig_2 = px.line(y=results.trend, x=results.trend.index, title='Production Trend', width=850, height=400)
fig_2.update_xaxes(title_text='Date')  # Set x-axis label
fig_2.update_yaxes(title_text='Trend')  # Set y-axis label
fig_2.show()
```

Consumption Analysis:

In [21]:
```
total_consumption0 = data.filter(['Total Primary Energy Consumption', 'Date'], axis=1)
total_consumption = total_consumption0.rename(columns={'Total Primary Energy Consumption':'Total_Consumption'})
total_consumption = total_consumption.set_index('Date')
```

```
results = seasonal_decompose(total_consumption['Total_Consumption'])

fig_1 = px.line(y=results.observed, x=results.observed.index, title='Global Energy Consumption Rate',width=850, height=450)
fig_1.update_xaxes(rangeslider_visible=True)
fig_1.update_xaxes(title_text='Date')  # Set x-axis label
fig_1.update_yaxes(title_text='Total Consumption')  # Set y-axis label
fig_1.show()
fig_2 = px.line(y=results.trend, x=results.trend.index, title='Consumption Trend',width=850, height=400)
fig_2.update_xaxes(title_text='Date')  # Set x-axis label
fig_2.update_yaxes(title_text='Total Consumption')  # Set y-axis label
fig_2.show()
```

Consumption vs Production:

In [23]:
```
data20 = data.filter(['Total Primary Energy Production','Total Primary Energy Consumption', 'Date'], axis=1)
data2 = data20.rename(columns={'Total Primary Energy Consumption':'Total_Consumption',
                               'Total Primary Energy Production':'Total_Production'})
data2 = data2.set_index('Date')
data2.head()

fig = px.line(data2, x=data2.index, y=data2.columns,
              title='Energy Production vs Consumption',width=850, height=500)
fig.update_xaxes(rangeslider_visible=True)
fig.show()
```

Energy Forecasting using Gradient Boost:

In [24]:
```
df = pd.read_csv('PJME_hourly.csv')
df = df.set_index('Datetime')
df.index = pd.to_datetime(df.index)
```

In [25]:
```
df.plot(style='.',
        figsize=(15, 5),
        title='PJM Energy (in MW) over time')
plt.show()
```

Untitled

In [26]:
```
train = df.loc[df.index < '01-01-2015']
test = df.loc[df.index >= '01-01-2015']
```

In [27]:
```
fig, ax = plt.subplots(figsize=(15, 5))
train.plot(ax=ax, label='Training Set', title='Train/Test Split')
test.plot(ax=ax, label='Test Set')
ax.axvline('01-01-2015', color='black', ls='--')
ax.legend(['Training Set', 'Test Set'])
plt.show()
```

In [28]:
```
def create_features(df):
    df = df.copy()
    df['hour'] = df.index.hour
    df['dayofweek'] = df.index.dayofweek
    df['quarter'] = df.index.quarter
    df['month'] = df.index.month
    df['year'] = df.index.year
    df['dayofyear'] = df.index.dayofyear
    df['dayofmonth'] = df.index.day
    df['weekofyear'] = df.index.isocalendar().week
    return df

df = create_features(df)
```

In [29]:
```
train = create_features(train)
test = create_features(test)

features = ['dayofyear', 'hour', 'dayofweek', 'quarter', 'month', 'year']
target = 'PJME_MW'

X_train = train[features]
y_train = train[target]
```

Untitled

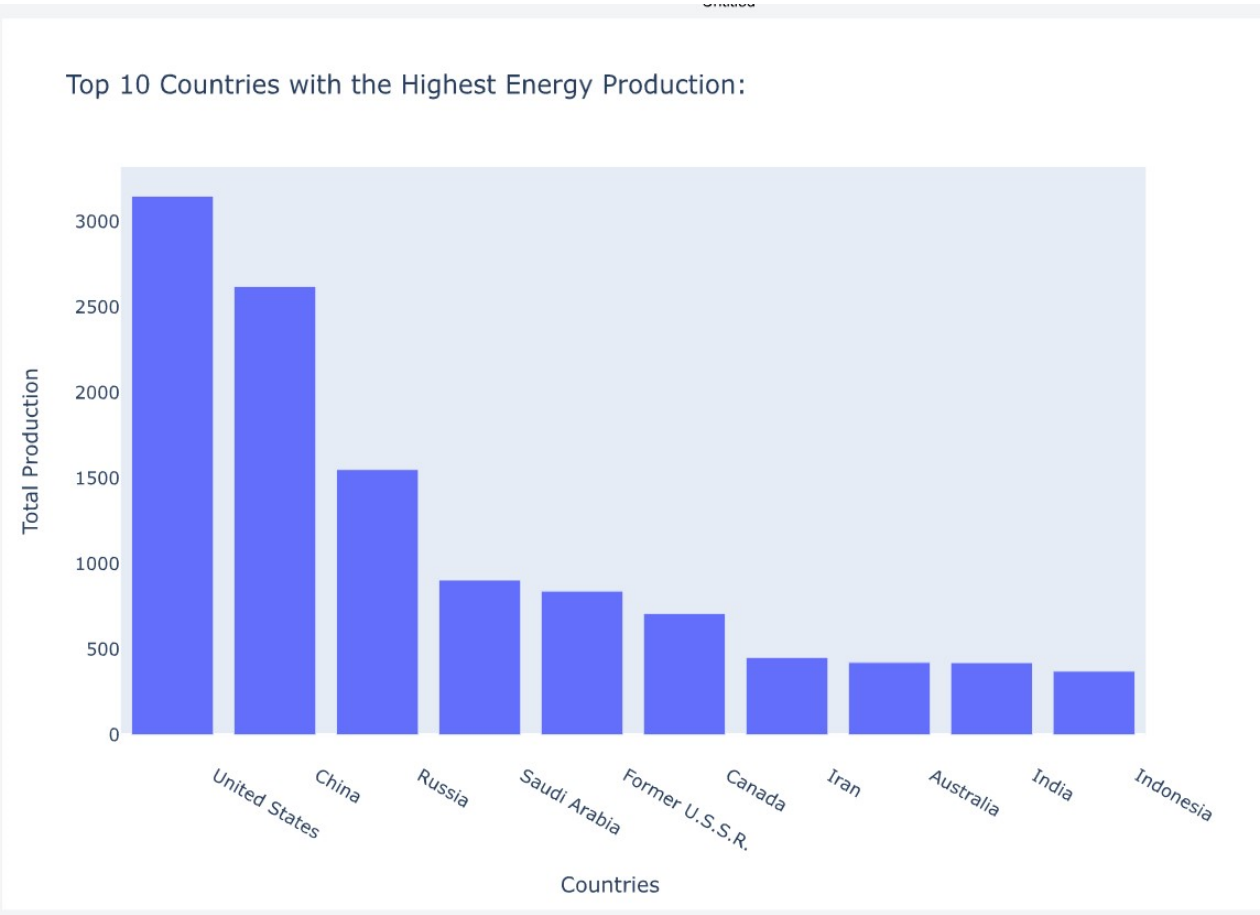
```
X_test = test[features]
y_test = test[target]
```

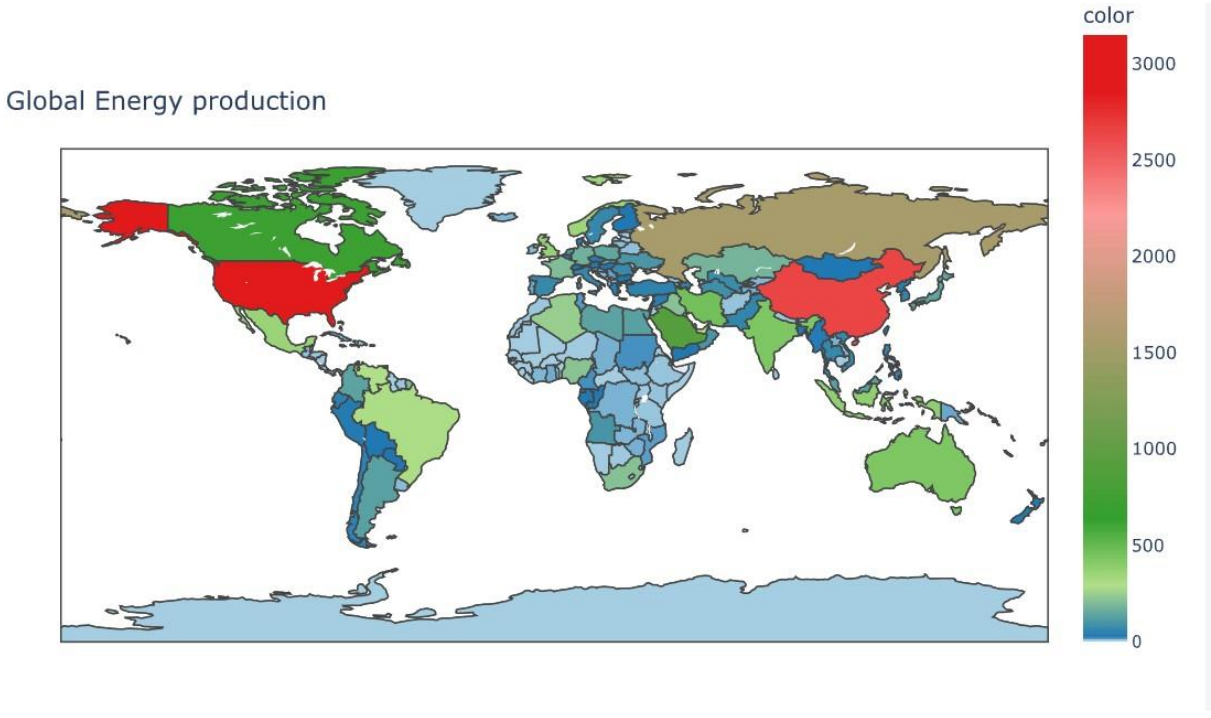
In [30]:
```
import xgboost as xgb
from sklearn.metrics import mean_squared_error

# build the regression model
reg = xgb.XGBRegressor(base_score=0.5, booster='gbtree',
                       n_estimators=1000,
                       early_stopping_rounds=50,
                       objective='reg:linear',
                       max_depth=3,
                       learning_rate=0.01)
reg.fit(X_train, y_train,
        eval_set=[(X_train, y_train), (X_test, y_test)],
        verbose=100)
```

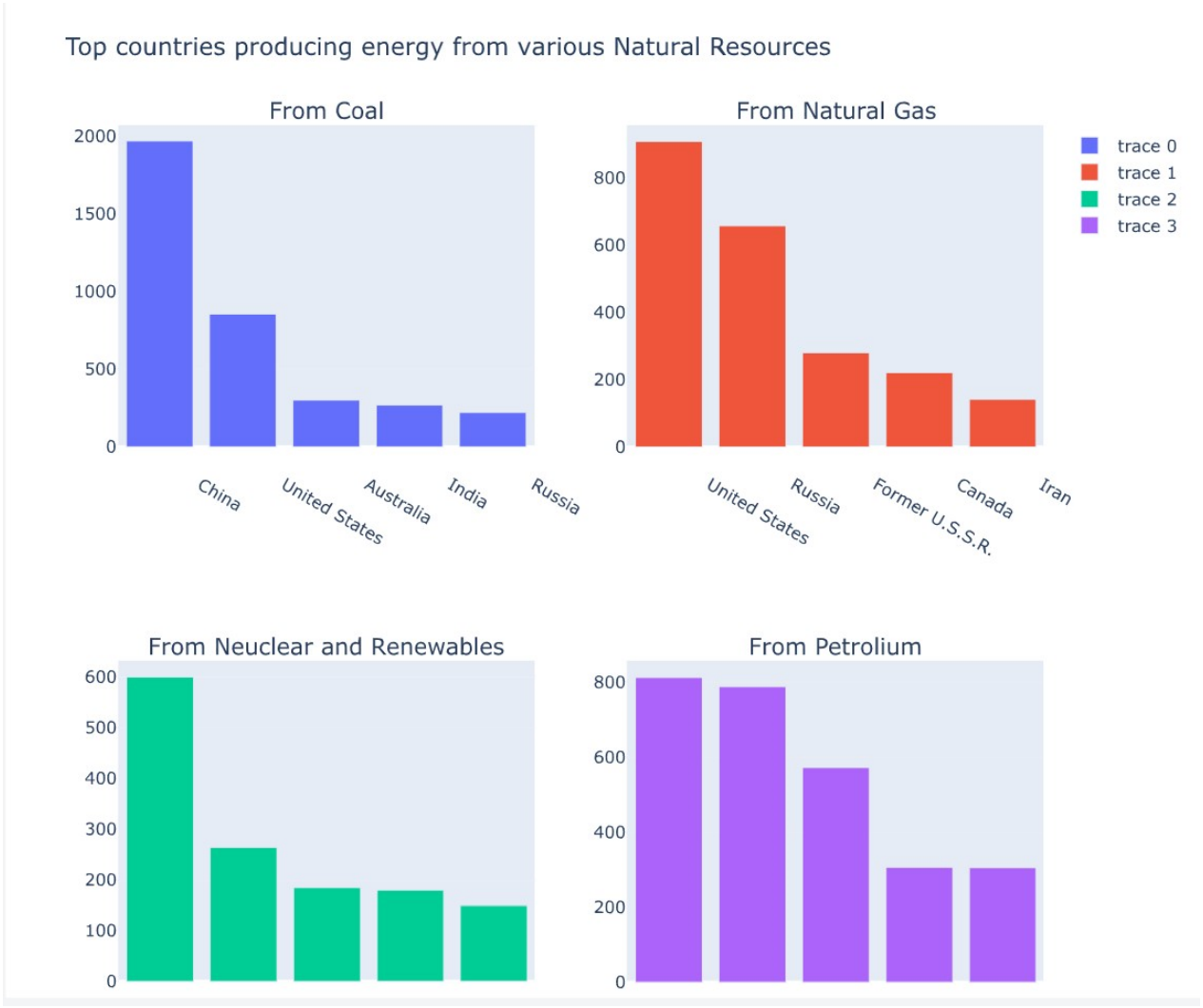
In [32]:
```
test['Forecast'] = reg.predict(X_test)
df = df.merge(test[['Forecast']], how='left', left_index=True, right_index=True)
ax = df[['PJME_MW']].plot(figsize=(15, 5))
df['Forecast'].plot(ax=ax, style='.')
plt.legend(['Truth Data', 'Forecast'])
ax.set_title('Past Consumption and Forecast')
plt.show()
```

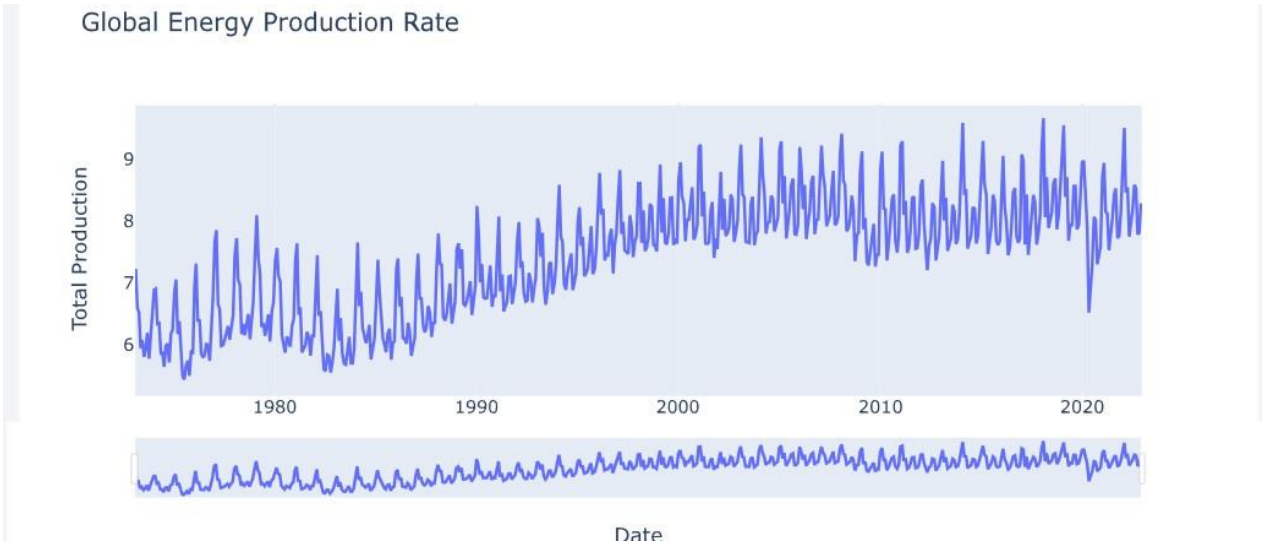
6. RESULTS AND DISCUSSION

Production Anaysis:

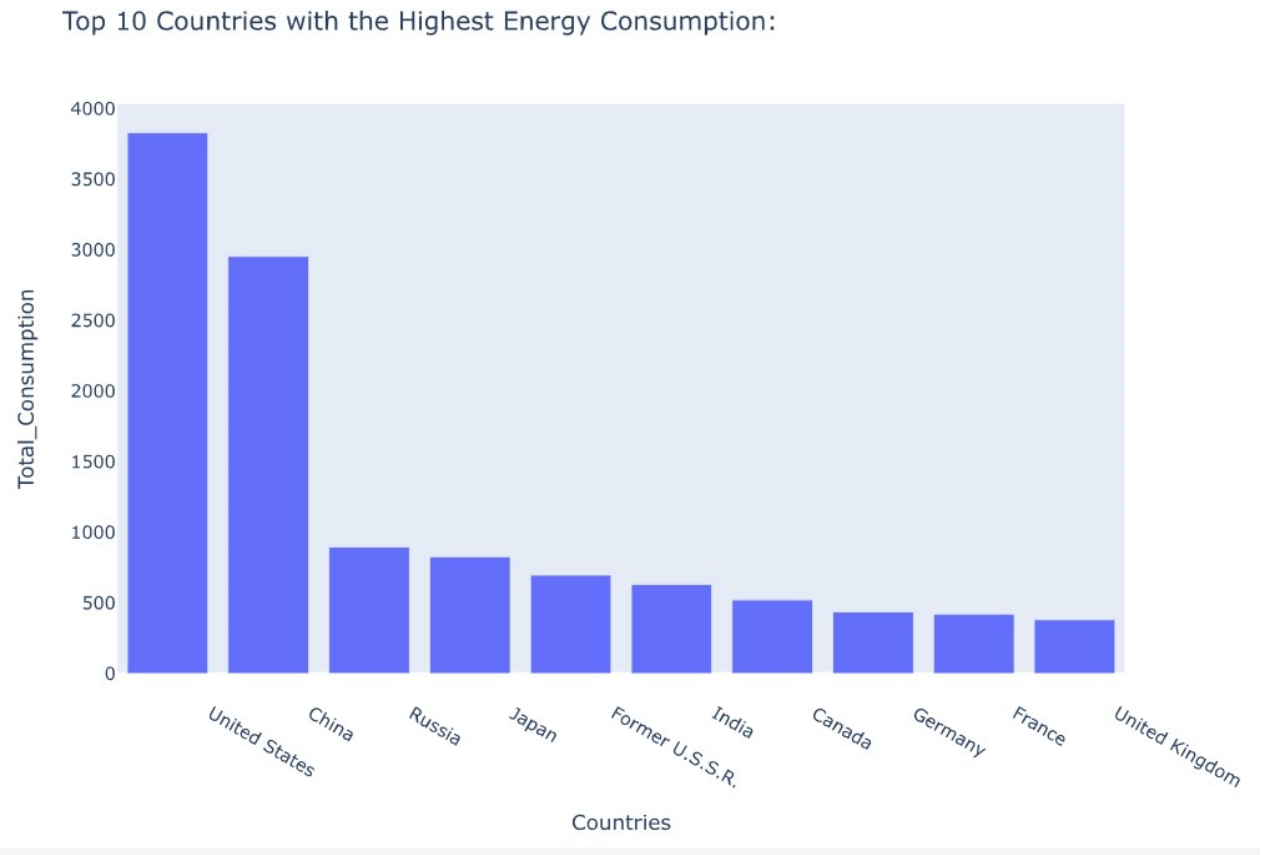


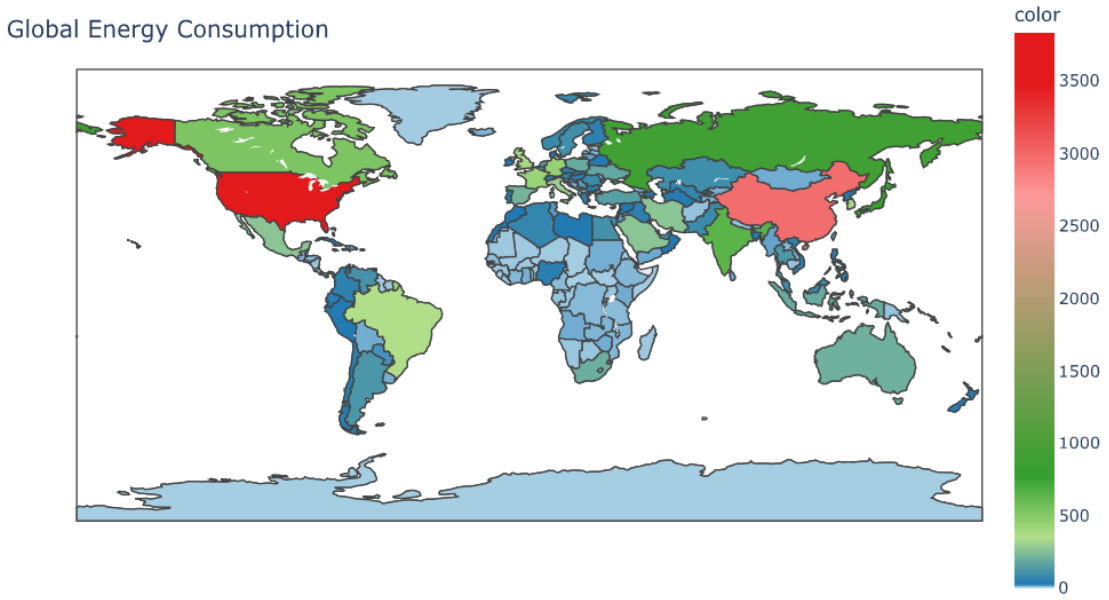


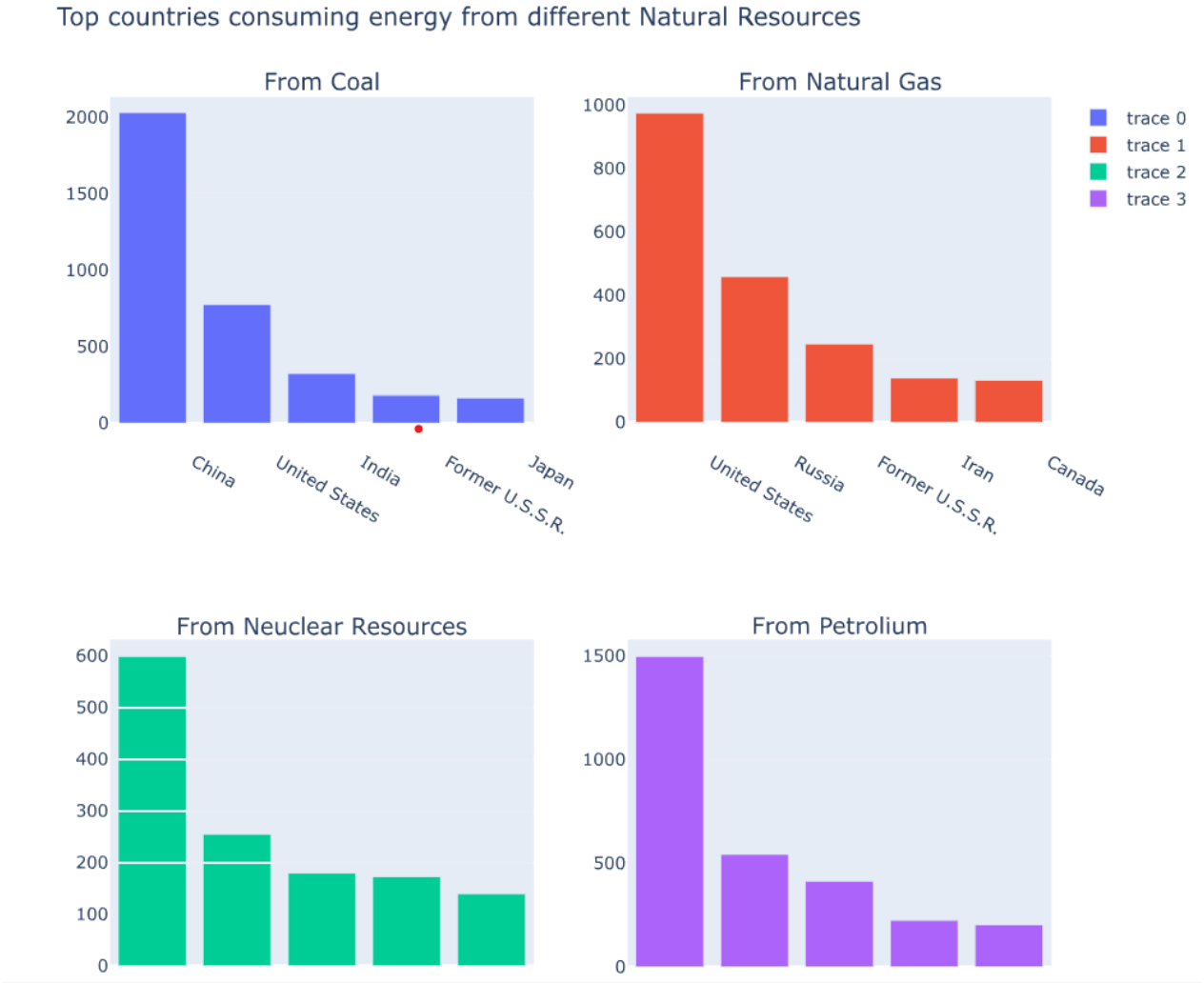

Top countries producing energy from various Natural Resources
From Coal
2000
1500
1000
500
0
China
United States
Australia
India
Russia
From Natural Gas
800
600
400
200
0
United States
Russia
Former U.S.S.R.
Canada
Iran
trace 0
trace 1
trace 2
trace 3
From Neuclear and Renewables
600
500
400
300
200
100
0
From Petrolium
800
600
400
200
0

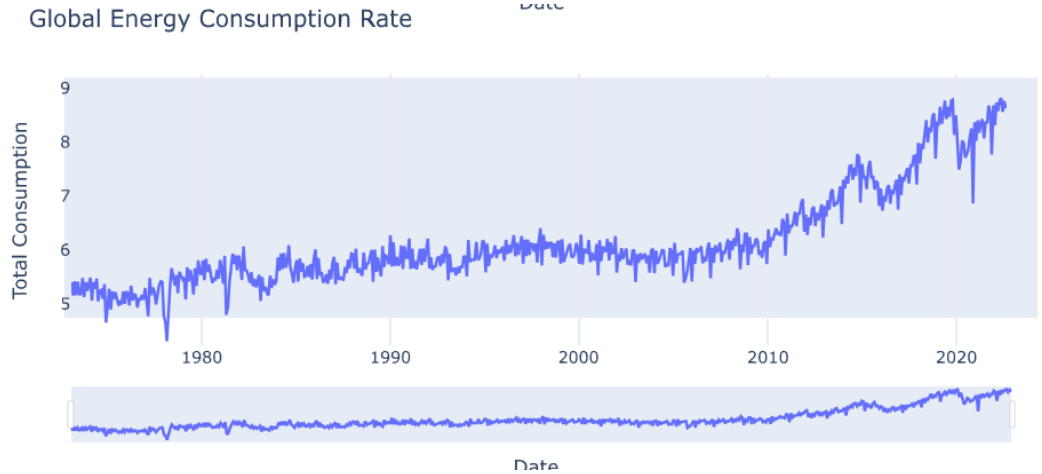

Global Energy Production Rate
Total Production
9
8
7
6
1980
1990
2000
2010
2020
Date

Consumption Analysis:

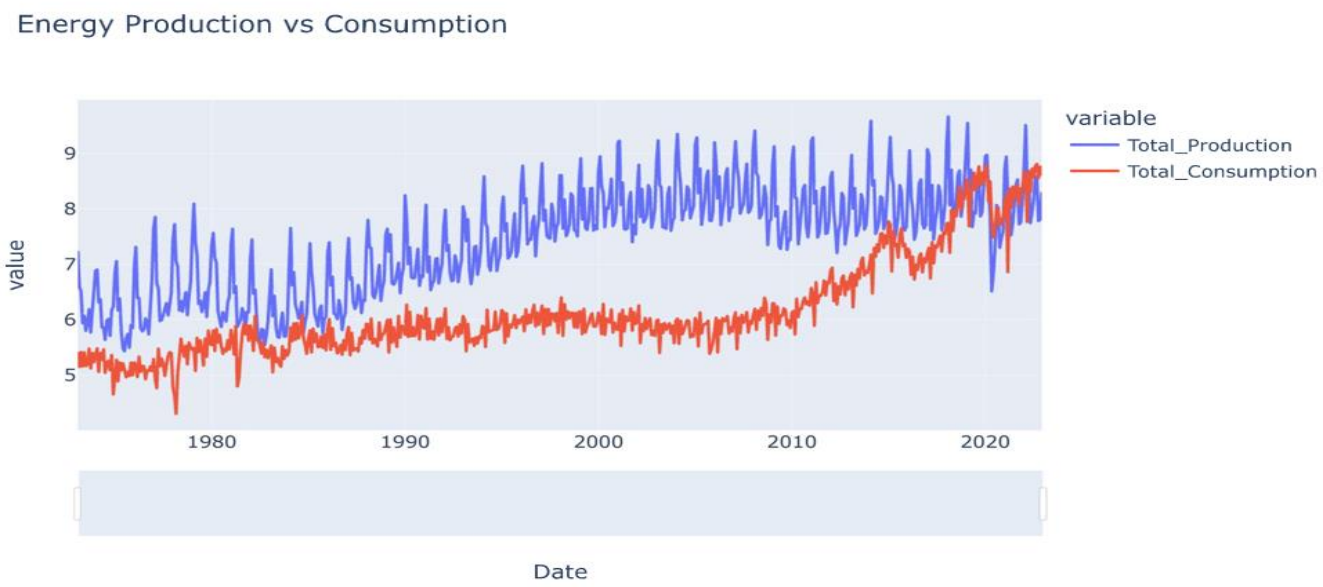




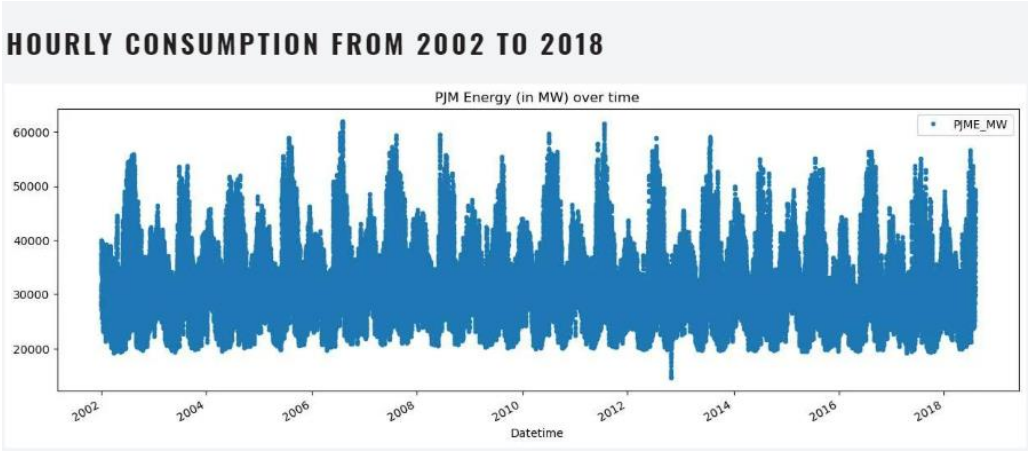

Top countries consuming energy from different Natural Resources
From Coal
2000
1500
1000
500
0
China
United States
India
Former U.S.S.R.
Japan
From Natural Gas
1000
800
600
400
200
0
United States
Russia
Former U.S.S.R.
Iran
Canada
trace 0
trace 1
trace 2
trace 3
From Neuclear Resources
600
500
400
300
200
100
0
From Petrolium
1500
1000
500
0


Global Energy Consumption Rate
Total Consumption
9
8
7
6
5
1980
1990
2000
2010
2020
Date

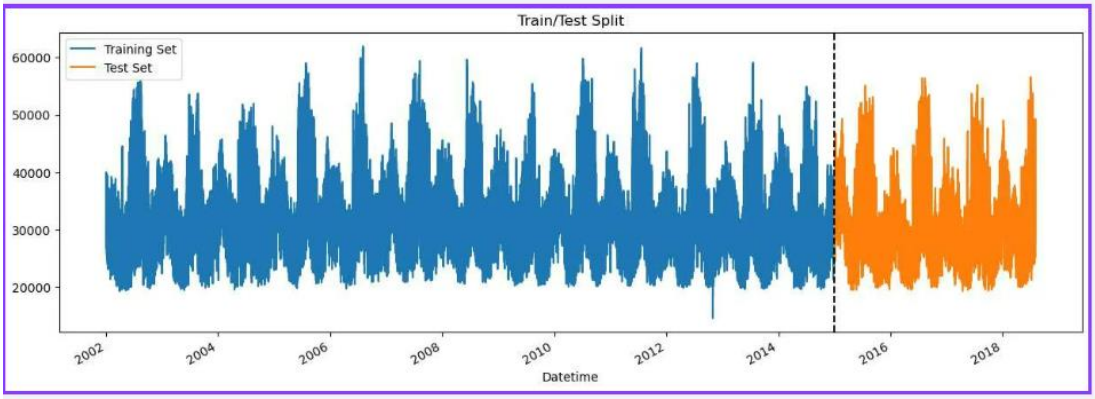
Production vs Consumption:



Energy Forecasting using Gradient Boost:



AFTER SEPERATING TEST CASE AND TRAIN CASE



FORECASTED DATA

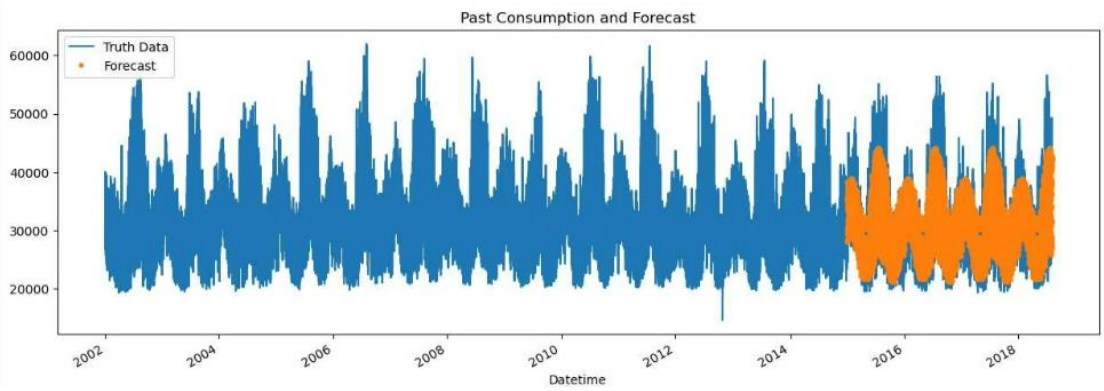
In [34]:
```
df
```

Out[34]:

Datetime	PJME_MW	hour	dayofweek	quarter	month	year	dayofyear	dayofmonth	weekofyear	Forecast
2002-01-01 01:00:00	30393.0	1	1	1	1	2002	1	1	1	NaN
2002-01-01 02:00:00	29265.0	2	1	1	1	2002	1	1	1	NaN
2002-01-01 03:00:00	28357.0	3	1	1	1	2002	1	1	1	NaN
2002-01-01 04:00:00	27899.0	4	1	1	1	2002	1	1	1	NaN
2002-01-01 05:00:00	28057.0	5	1	1	1	2002	1	1	1	NaN
...	...	...	...	...	...	...	...	...	...	...
2018-08-02 20:00:00	44057.0	20	3	3	8	2018	214	2	31	42575.265625
2018-08-02 21:00:00	43256.0	21	3	3	8	2018	214	2	31	42522.703125
2018-08-02 22:00:00	41552.0	22	3	3	8	2018	214	2	31	40804.101562
2018-08-02 23:00:00	38500.0	23	3	3	8	2018	214	2	31	37933.539062
2018-08-03 00:00:00	35486.0	0	4	3	8	2018	215	3	31	31588.259766

145372 rows × 10 columns

Forecasted along with actuals:



7. CONCLUSION

In summary, this comprehensive energy analysis project unveils pivotal insights into global energy dynamics using Python and diverse data visualization techniques. Through meticulous data exploration and visualization, it highlights the top energy-producing and consuming countries, delves into various resource-specific analyses, and forecasts future energy consumption using advanced models like XGBoost.

The interactive visualizations, from choropleth maps to line plots, offer a comprehensive understanding of global energy trends. Notably, the project's ability to decipher production and consumption patterns for coal, natural gas, nuclear, and petroleum resources contributes significantly to its depth.

Additionally, the seasonal decomposition analysis provides a temporal perspective, while the feature importance analysis in forecasting models underscores crucial factors influencing energy consumption trends.

Overall, this project's robust analyses and visualizations contribute meaningfully to the discourse on global energy sustainability and resilience.

8. REFERENCES

1. Electricity load forecasting: a systematic review
   Isaac Kofi Nti, Moses Teimeh, Owusu Nyarko-Boateng & Adebayo Felix Adekoya
   Journal of Electrical Systems and Information Technology volume 7, Article number: 13 (2020)
   Published: 09 September 2020
2. Data Analytics in the Electricity Sector – A Quantitative and Qualitative Literature Review
   June 2020Energy and AI 1(3):100009
   DOI:10.1016/j.egyai.2020.100009
   LicenseCC BY 4.0
3. "Load Profiles and Their Use in Electricity Settlement"- Elexon guide 7th November, 2013.
   S.K. Sinha, S.K. Soonee, S.S. Barpanda, K.K. Ram Eastern Region Power Demand Scenario- A forecast GMPS-2000.
4. Analysis of energy consumption, emission and saving opportunities in an educational institute in northeast India. volume 4, pages375–388 (2020)
   S. Acharya, A. Shil, C. Debbarma, J. Reang, R. Chakraborty & A. Ghosh
   Published: 14 July 2020
5. Dataset - https://www.kaggle.com/datasets/akhiljethwa/world-energy-statistics
6. Dataset - https://www.kaggle.com/datasets/robikscube/hourly-energy-consumption?select=PJM_Load_hourly.csv