Global Renewable Energy Analysis and Prediction *with PCA, ARIMA, ARIMAX, and Random Forest*

December 8th, 2024
ISYE 6740 Final Project

Michael Daniels - *mdaniels33@gatech.edu*
Nicholas Jerdack - *njerdack3@gatech.edu*
Alexander Shropshire – *ashropshire6@gatech.edu*

Abstract

As the global demand for renewable energy accelerates, forecasting energy trends has become increasingly crucial for policymakers and investors. This project explores global renewable energy patterns and predicts future output and consumption by analyzing diverse countries with varying energy profiles. We identified key groupings among nations using unsupervised and supervised machine learning techniques, including Principal Component Analysis (PCA), clustering, and time-series forecasting. We developed scenario-based predictions for renewable energy growth.

Our findings reveal clear patterns, such as the clustering of oil-producing countries and renewable energy leaders and provide actionable insights into future energy transitions. By leveraging historical energy data, policy trends, and investment trajectories, our models offer flexible forecasts tailored to conservative and aggressive scenarios. These insights highlight opportunities for targeted investments and practical policy design, contributing to global sustainability and energy security.

Introduction & Problem Statement

The global shift toward renewable energy is a multifaceted challenge influenced by economic, political, and technological factors. Despite growing investments and policy efforts to reduce reliance on fossil fuels, forecasting renewable energy output remains complex due to nations' diverse energy profiles, historical trajectories, and strategic priorities. These disparities create significant uncertainty in projecting global energy transitions.

This project addresses these challenges by analyzing the current energy landscapes of key countries and forecasting their future renewable energy output under various scenarios. Our focus includes significant energy producers, polluters, and renewable energy leaders like the United States, China, India, and Scandinavian nations. Using unsupervised clustering techniques, we identify patterns in energy adoption, categorizing countries based on shared trends. Additionally, our supervised regression models using data in a time series sense provide detailed predictions, offering insights into the potential growth of renewable energy across different regions.

The insights generated through this analysis aim to guide policymakers, investors, and stakeholders in making strategic decisions that support global energy transitions. By accounting for uncertainties in investments and policies, our approach provides a robust framework for understanding and navigating the rapidly evolving renewable energy landscape.

Team 005: Nicholas Jerdack, Alexander Shropshire, Michael Daniels

Project Scope

The team focused on analyzing a diverse selection of countries to explore global renewable energy trends. These nations represent globally renowned clean energy leaders, major oil suppliers, and large-scale economies, offering a comprehensive global view of the varied energy strategies.

Countries Analyzed

The countries were selected and categorized as follows:

- **Major Oil Producers:**
  United States, Saudi Arabia, Qatar, Australia, Gabon, Angola, Libya, Russia, Iran, Iraq, Kuwait, Venezuela, Nigeria, Kazakhstan, Mexico, and Canada.
- **Renewable Energy Leaders:**
  Norway, Iceland, Denmark, Germany, Sweden, Finland, New Zealand, Costa Rica, Portugal, Spain, Austria, and Switzerland.
- **Large-Scale Economies and Emerging Energy Economies:**
  China, India, Brazil, and United Arab Emirates.

These countries were filtered from the dataset, and any rows with missing values in key features were removed to ensure a clean and robust analysis.

Literature Review

In Pham Xuan Hoa's study[3] on six developed Asian countries, the relationship between renewable energy consumption and variables like economic growth and fossil fuel use is explored using the Cobb-Douglas model, but without cross-validation. In contrast, Liu Xintian's work [4] analyzes the effects of El Nino on renewable energy in Japan, the U.S., and Australia using the same renewable energy dataset from 1965 to 2022 that we are using. Our study will expand on these approaches by analyzing a broader range of countries, incorporating additional variables like weather and greenhouse gas emissions, and applying time-series and machine learning models with cross-validation and parameter tuning.

According to Deloitte's studies of the industry [5], the renewable energy industry in 2023 witnessed strong growth in solar energy, while wind faced challenges due to supply chain, labor, and permitting issues. Utility-scale solar installations outpaced wind by a significant margin, reflecting a trend expected to continue into 2024 with strong federal investment. Despite these hurdles, the sector is positioned for further growth, driven by clean energy laws, tax incentives, and increasing corporate demand for renewable solutions.

Techniques Employed

The project utilized a combination of analytical techniques to identify patterns and predict future renewable energy trends:

Preprocessing Methodology

Data Overview and Preparation

The dataset for this analysis integrates renewable energy-related CSV files from Kaggle and time-series data from the World Bank API. It combines historical and categorical information with global metrics on energy use, economics, and demographics. This unified dataset provided the foundation for feature selection and time-series modeling.

Key variables included energy output, renewable energy investments, GDP, population, and energy consumption metrics. However, challenges such as missing data, multicollinearity, and non-country entities required a robust data preparation process to ensure relevance and usability.

Data Extraction

Kaggle Data
The Kaggle files were dynamically loaded into Pandas DataFrames. Filenames were sanitized to ensure consistent variable naming, removing special characters, numbers, and spaces. DataFrames were iteratively merged using outer joins on the standard keys Entity, Code, and Year.[1]

World Bank Data
Data was extracted using specific indicators and normalized to include essential fields such as Country, Year, and indicator values. Non-country entities, such as regional aggregates, were excluded to maintain a country-level focus.[2]

Data Cleaning & Preprocessing

- **Missing Values**: Forward filling was applied to address missing data within country-level time series, ensuring continuity while acknowledging the risk of over-smoothing abrupt changes. There is also a risk when imputing as smoothing techniques sometimes obscure abrupt changes or outliers.
- **Redundant Features**: To reduce redundancy, overlapping capacity measures (e.g., solar, wind, geothermal) and duplicate columns resulting from merging processes were removed.
- **Normalization**: Certain indicators, like energy consumption and GDP, were normalized by population to enhance cross-country comparability.
- **Entity Filtering**: Non-country entities were excluded to streamline the analysis to distinct nations.
- **Temporal Alignment**: Data was resampled annually, with interpolation applied to fill missing years where necessary, ensuring temporal consistency across countries. Also, recent data is underrepresented due to reporting delays, limiting insights into current trends.
- **Preliminary Feature Selection**: Variables relevant to renewable energy trends, economic indicators, and demographics were prioritized, while irrelevant or redundant fields were excluded to improve model performance.
- **Data Structuring**: The dataset was reorganized into a multi-index format with Year and Entity as indices, facilitating efficient time-series modeling and feature engineering.

The robust dataset combines trusted global sources and supports multivariate trend analysis. Future improvements could include integrating real-time or satellite-derived metrics to address temporal gaps, employing machine learning-based imputation for missing values, and incorporating composite indicators to capture complex relationships.

Unsupervised Learning

Principal Component Analysis (PCA)

PCA is a statistical technique used to reduce the dimensionality of datasets while retaining the most essential information. PCA simplifies complex datasets and highlights underlying patterns by transforming the original variables into smaller uncorrelated components. In this project, PCA was employed to analyze the energy profiles of different countries by compressing multiple indicators, such as energy output, renewable investments, and economic metrics, into a few principal components.

By reducing the data's complexity, PCA facilitated the identification of these patterns and informed subsequent clustering and predictive modeling steps. This approach helped uncover relationships between energy profiles that might not have been apparent in the raw data, providing a robust framework for grouping and analyzing nations based on their energy strategies.

Supervised Learning

ARIMA (AutoRegressive Integrated Moving Average)

An ARIMA model is a univariate time series model, that is a widely used statistical method for analyzing and forecasting time-series data, univariate meaning that only one time series is used. This model has three components, which are the AR, differencing (I), and MA components.

- AutoRegression (AR): Captures relationships between an observation and a certain number of lagged observations (past values).
- Integrated (I): Accounts for differencing the data to achieve stationarity, addressing trends or seasonality.
- Moving Average (MA): Models the dependency between observation and residual errors from a moving average model applied to previous observations (it regresses current error against previous errors).

The AR component regresses the current time value of a time series X against its past values. If the coefficients are θ_i, where $i = 1, \ldots, m$ with m being the length of the time series and θ_0 being the constant, the formulation for AR is:

$$x_t = \theta_0 + \theta_1 x_{t-1} + \theta_2 x_{t-2} + \ldots + \theta_m x_{t-m} + \varepsilon_t$$

where ε_t is often taken to be IID random normal with mean 0 and variance 1. This is basically a linear combination of the values in a time series, with the current value regressed against previous values. In an ARIMA model, the parameter p is used for the AR component. It determines how many past time series values the current time value is regressed against. The MA component does something similar, but with errors (difference between actual value and its prediction). The errors will be denoted by Z with coefficients α_i where $i = 1, \ldots, m$ with m being the same as in the AR portion and α_0 being the constant. The resultant formula that regresses the current error term against previous error terms is:

$$z_t = \alpha_0 + \alpha_1 z_{t-1} + \alpha_2 z_{t-2} + \ldots + \alpha_m z_{t-m} + \varepsilon_t$$

with ε_t taken as in the AR component above and α_0 being a constant. In an ARIMA model, the parameter q determines how many previous error terms the current error term is regressed against.

If both are used as models by themselves (not being part of an ARIMA model or as just part of an ARMA model, which combines these two aspects), they consider that the time series is stationary, which may not be the case for each country in this dataset. The differencing (I) component tries to account for this.

Instead of the time series being $x_1, x_2, \ldots, x_m$, differencing takes the difference between a value and its previous value, so there will be one less value in the time series with the result being $x_2 - x_1, x_3 - x_2, \ldots, x_t - x_{t-1}$. A stationary time series has constant mean and constant variance over time (no time varying mean and no time varying variance). Putting this with the sum of the AR component and MA component like in ARMA, the formula for ARIMA is:

$$x_t = c + \theta_1(x_2 - x_1) + \theta_2(x_3 - x_2) + \ldots + \theta_{m-1}(x_m - x_{m-1}) + \alpha_1(z_2 - z_1) + \alpha_2(x_3 - x_2) + \ldots + \alpha_{m-1}(x_m - x_{m-1}) + \varepsilon_t$$

This can be written compactly as:

$$x_t = c + \sum_{i=1}^{p-1} \theta_i (x_{t-i+1} - x_{t-i}) + \sum_{j=1}^{q-1} \alpha_j (z_{t-j+1} - z_{t-j}) + \varepsilon_t$$

The errors are unknown to begin with, so an ARIMA function first fits an AR model to the differenced data and gets the errors for the MA component from there.

In this project, ARIMA was the primary forecasting tool to predict renewable energy output for each country. It was particularly suited for modeling energy trends because it can manage non-stationary data by integrating different techniques (different amounts of differencing). The model was tuned for each country to optimize parameters, ensuring accurate predictions reflective of unique energy trajectories.

The benefits of ARIMA include its focus on temporal dependencies, which allowed it to effectively model long-term trends and fluctuations in renewable energy data. By adapting parameters to each country, ARIMA captured the nuances of diverse energy profiles. Despite its strengths, ARIMA also posed challenges, such as its reliance on linear relationships and assumptions of stationary data. These limitations were mitigated through careful data preprocessing, including differencing and normalization, to ensure meaningful and reliable predictions.

This combination of unsupervised and supervised techniques allowed the team to identify historical trends, understand underlying patterns, and generate data-driven forecasts for renewable energy adoption worldwide.

ARIMAX (ARIMA + X, where X = Exogenous Variables)

In theory, adding more unique features of information to a model, or exogenous variables, one should be able to improve the predictive ability of ARIMA. ARIMAX takes what was shown for ARIMA above and adds exogenous time series variables as more predictors. The formula becomes:

$$x_t = c + \sum_{i=1}^{p-1} \theta_i (x_{t-i+1} - x_{t-i}) + \sum_{j=1}^{q-1} \alpha_j (z_{t-j+1} - z_{t-j}) + \sum_{v=1}^{n} \sum_{k=1}^{m-1} \beta_{k,v} (b_{v,t-k+1} - b_{v,t-k}) - + \varepsilon_t$$

This adds a linear combination of each variable with differencing considered, where $\beta_{k,v}$ are coefficients for each of the differenced values of the exogenous time series $b_{v,t}$ with the rest the same as before.

Random Forest as both Variable Selection Tool & Standalone Model

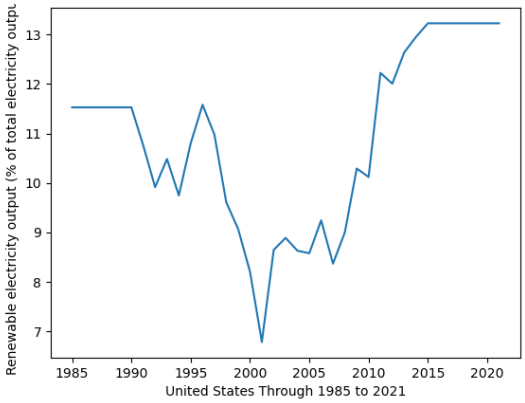
For each country in this dataset, random forest was used in the regression sense, instead of the classification sense. It was used as a model to compare with ARIMA and ARIMAX, but also as a variable selection method for ARIMA and ARIMAX. The basis of a random forest is the idea of a decision tree. A decision tree takes the predictor (exogenous variables) data and does a binary split of the data on one variable. This results in two child nodes that each have an independent set of the data. Combining those nodes results in the original data. This is then repeated at each node with the data at that node. This process keeps repeating until the data has split into a satisfactory number of leaf nodes. A prediction can then be made using this tree for a new datapoint by following the corresponding nodes that have the fulfilling inequalities that match the datapoint. The resultant prediction is the average training response value at the corresponding leaf node. The purpose of random forest is to fit a number of these trees. At the end, the prediction is the average of all the predictions of the trees.

Additional Techniques:

Team 005: Nicholas Jerdack, Alexander Shropshire, Michael Daniels

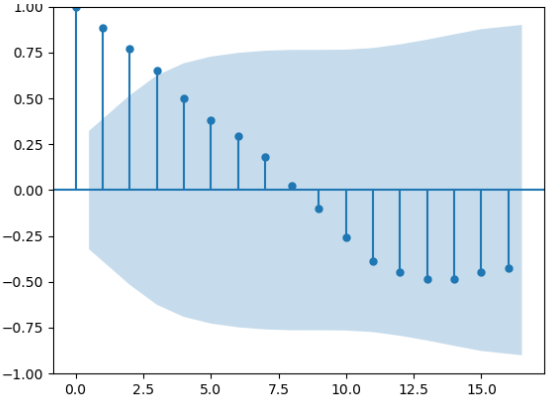
Stationarity Testing

Stationarity for each country's response data for the columns Renewable electricity output (% of total electricity output) and Renewable energy consumption (% of total final energy consumption) were checked through Time Series plots along with ACF and PACF plots as well as using the Augmented Dickey Fuller hypothesis test. The plots and Augmented Dickey Fuller test were also done for the first differenced data, to see if first order differencing made the data stationary. This was also done in the case of the response time series having the natural log of it taken. The log was tried to see if this would account for any varying variance. Plots for the United States are shown below for the original data:

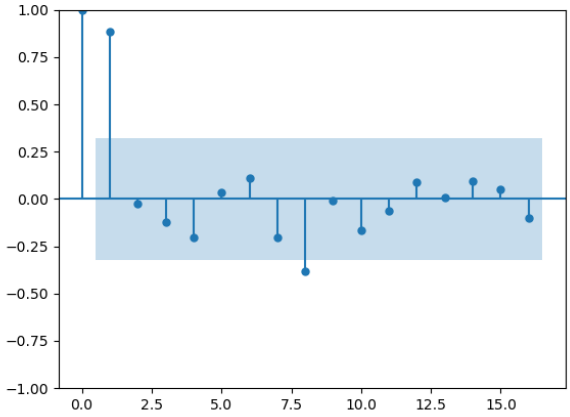


Historical Renewable Output Data for United States 1895-2021

ACF



PACF



For the Augmented Dickey Fuller test, the null hypothesis is that the time series contains a unit root, meaning that at least the mean is varying through time, and the time series is not likely stationary. Rejecting the null hypothesis means that there is no unit root in the time series, meaning that at least the mean does not vary with time indicating that the time series is likely stationary. The table for the p-values of the Adjusted Dickey Fuller test are shown in the Appendix for both response variables, each variable differenced, each variable having the natural log taken, and each variable having the natural log taken and differenced for each of the incorporated countries.

Parameter Tuning

Each of these three models was parameter tuned. For ARIMA and ARIMAX, the three parameters p, d, and q were tuned. The parameter p and q could take on integer values between 0 and 10. This defined how many previous values and previous errors each value in the time series should be regressed against. For the parameter d, defining the level of differencing in the data, it could take on integer values between 0 and 2. A value of two meant that each value in the time series x would have the value two time points before it subtracted from it.

Random Forest also had parameter tuning for a select number of its hyperparameters. There are many hyperparameters in random forest, but five were determined best to tune with most of the rest kept at their default values. The five that were tuned were n_estimators (which defined the number of trees that were fit in the random forest taking values 50, 100, 200, 300, 400, and 500), max_depth (which defined the maximum depth of each tree and could take values of None, 10, and 20), min_samples_split (which defined the minimum number of samples in a node where a split could be made and took values 2, 5, 10, and 15), min_samples_leaf (which defined the minimum number of samples that could be at each leaf node and could take values 1, 2, 5, 10, and 15), and max_features (which defined how many features could be picked from at each node to do another split and could take values of square root, logarithm base 2, and None). One parameter that was not left at its default was bootstrap because this is a time series dataset, so we do not want any bootstrapping done.

Cross-Validation

To conduct cross validation for a time series, it would have to be done by a slightly different method than cross validation for a non-time series dataset. For a non-time series dataset, a certain percentage of the data can be chosen to be in the training and testing datasets with these picked randomly to be in each. This can't be done with a time series dataset because the data is naturally ordered. This means the train/test data splitting must be done to keep the data order intact. Also, to keep the independence of the train versus test datasets, a gap must be made between them. An example for this dataset would be a country with data from 1965 to 2022. One instance of the train test split could be 1965 to 2010 in the training dataset and 2012 to 2022 in the testing dataset, leaving a gap of 1.

In the random forest case, the training dataset is further split in the same way in different split amounts. Each split would be tested on each combination of the parameters and the average mean squared error for each parameter combination would be obtained (seven splits of this kind were used). The parameter combination with the smallest average mean squared error would be the chosen parameter values.

In the case of ARIMA and ARIMAX, the best p, d, and q are found on the current split of the dataset by finding the combination that gives the smallest BIC value. Then cross validation is used on different splits of the training dataset to find the average mean squared error for just that parameter combination. The testing set used was the last 4 observations of the data, while the raining set was the rest of the data. Since this is being used in different countries, each country has a varying amount of data in the training set because different countries may have different amounts of data (not all data for each country starts at 1965).

Though we dive into our modeling methodologies and summarize results later in the paper, the result set from the techniques described above have corresponding outputs in the Appendix to view as an accompaniment to our summary insights.

Exploratory Data Analysis

Team 005: Nicholas Jerdack, Alexander Shropshire, Michael Daniels

Overview

The exploratory data analysis (EDA) phase provides an integrated view of global renewable energy investment, electricity output, and global socioeconomic patterns from 2000 to 2020. This phase aimed to identify trends, interactions, and anomalies that highlight both the progress and challenges in renewable energy transitions. Insights were derived from analyzing the dataset's geographic, temporal, and economic dimensions.

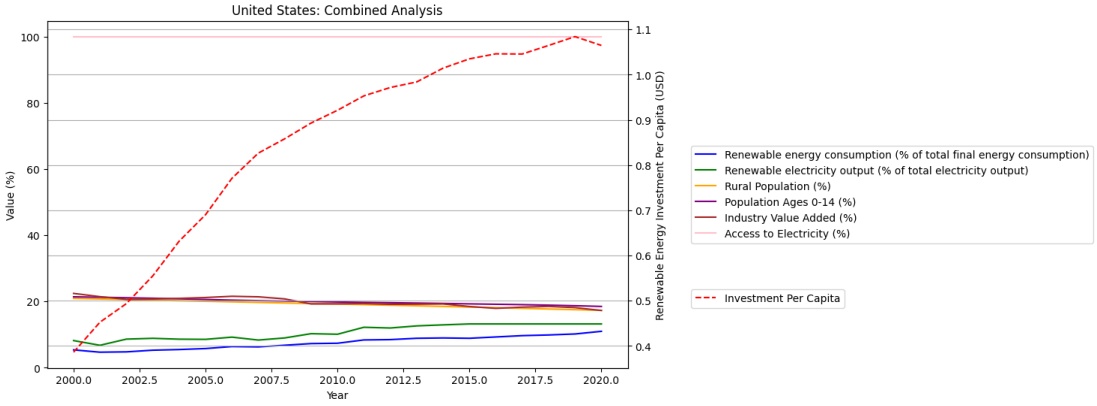
Geographic Trends

Pathways to Energy Transformation

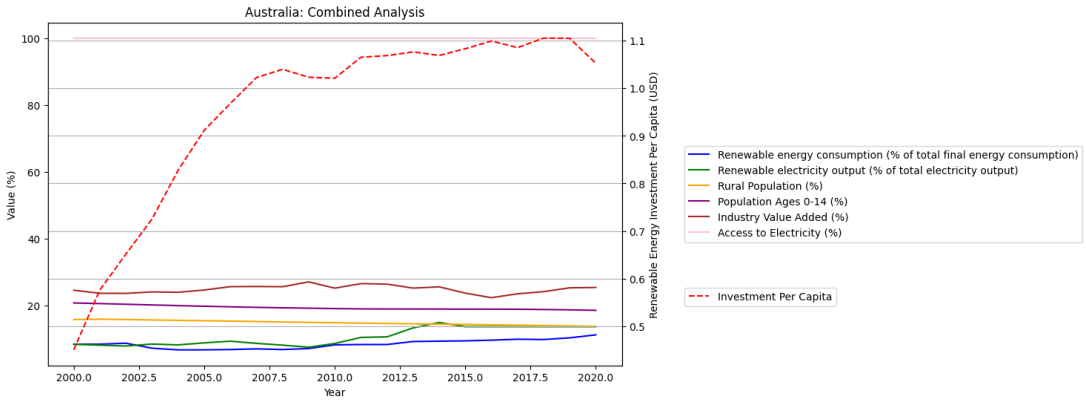
Countries across the globe exhibit distinct patterns in their adoption and implementation of renewable energy technologies. Key findings include:

United States and Australia:

Renewable energy investments rose steadily, plateauing in the mid-2010s. This leveling-off reflects a shift from infrastructure expansion to operational optimization, with future emphasis on advanced storage systems and grid improvements.



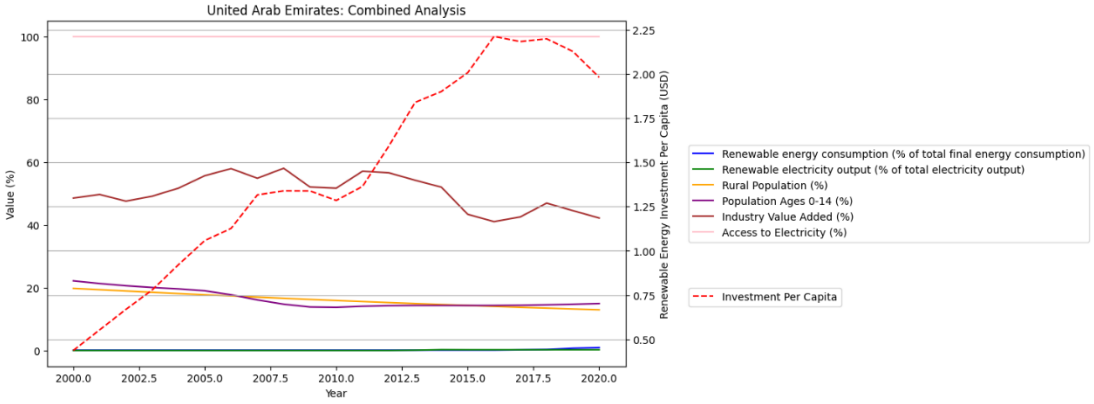
Multi-Variable Time Series: United States



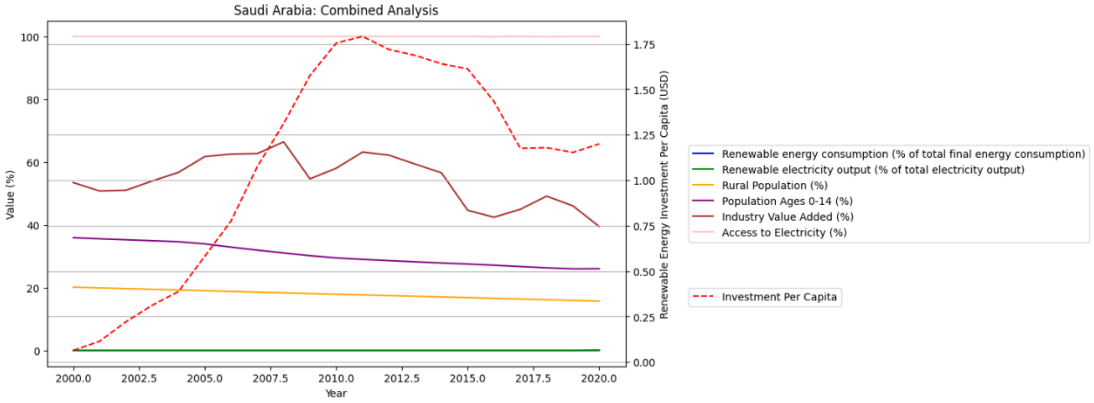
Multi-Variable Time Series: Australia

Oil-Dependent Economies:

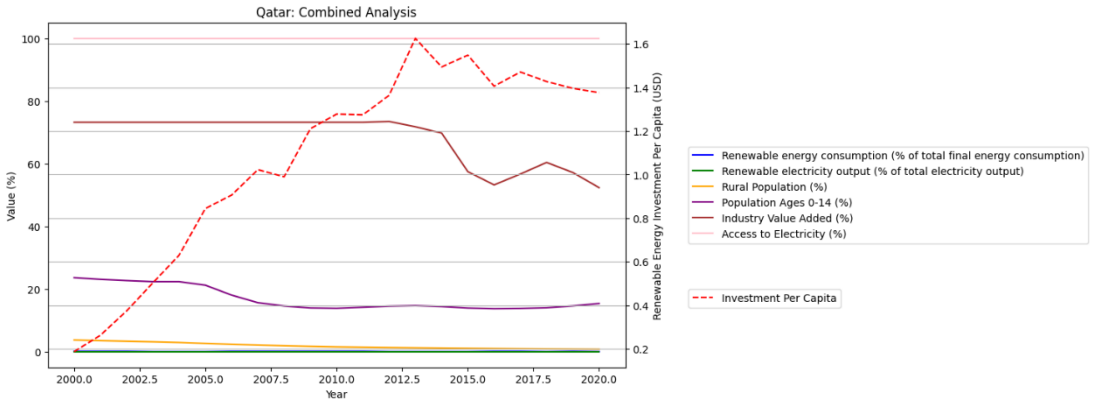
Nations like Saudi Arabia, Qatar, and the United Arab Emirates experienced peaks in renewable energy investments between 2010 and 2015, driven by economic diversification goals. However, subsequent declines reveal systemic challenges, such as entrenched fossil fuel infrastructure and market volatility.



Multi-Variable Time Series: United Arab Emirates



Multi-Variable Time Series: Saudi Arabia

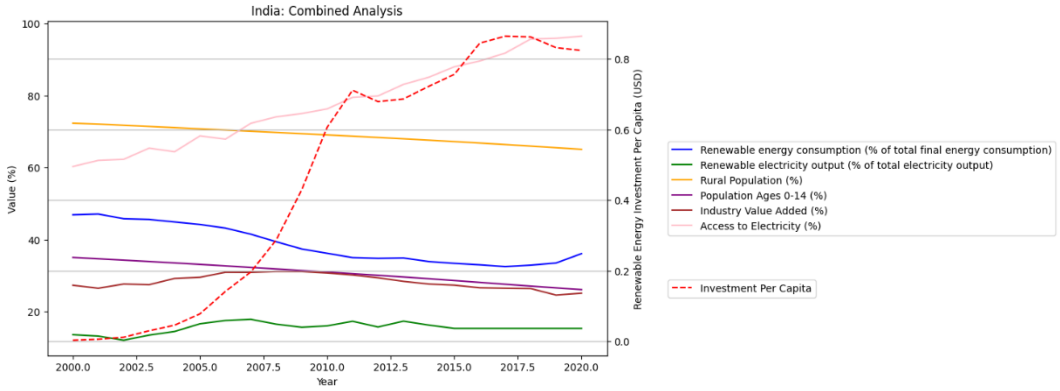


Multi-Variable Time Series: Qatar

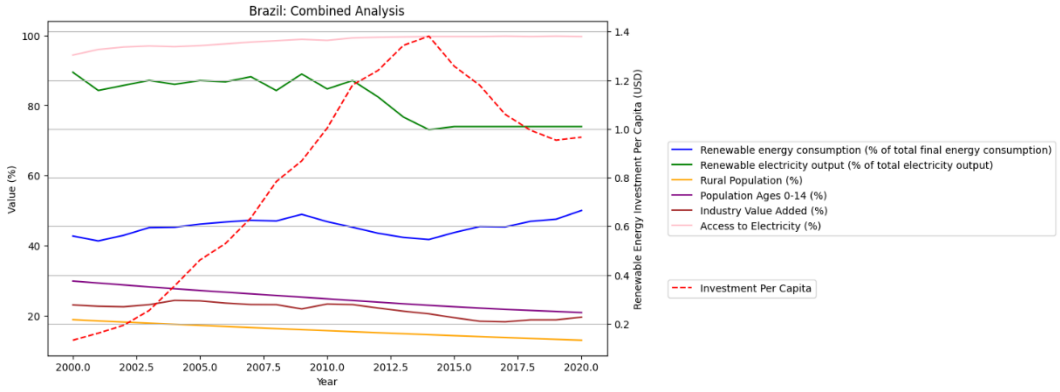
BRICS Economies:

Countries like China, Brazil, and India showed varying progress:

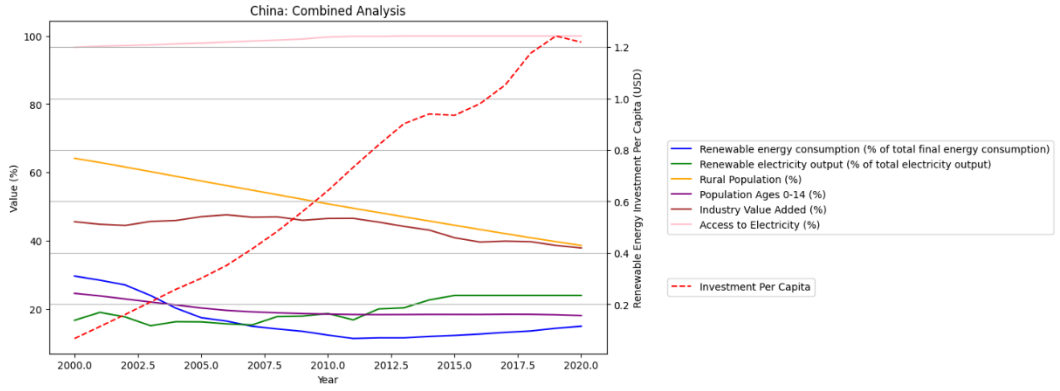
- China: Substantial investments and rising electricity output aligned with industrial policies to lead global clean energy innovation.
- Brazil: A reliance on hydropower ensured consistent renewable energy output, though diversification into other sources remains limited.
- India: Gains in rural electrification and renewable energy integration were slower, reflecting infrastructure and policy challenges.



Multi-Variable Time Series: India



Multi-Variable Time Series: Brazil

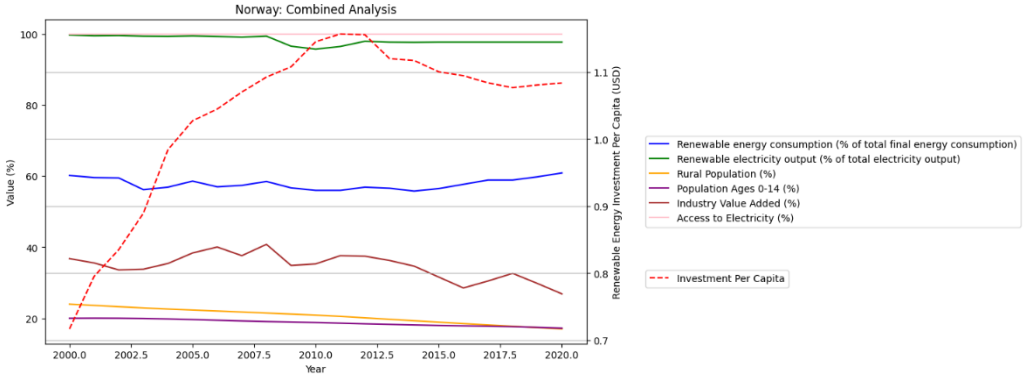


Multi-Variable Time Series: China

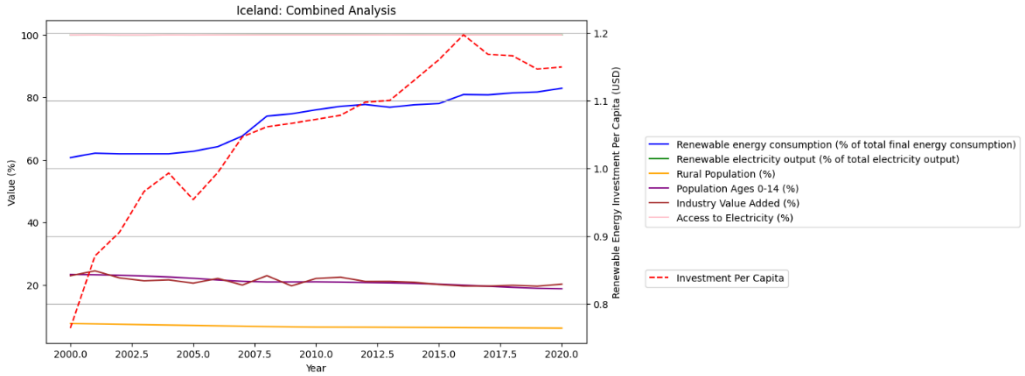
Renewable Energy Leaders:

Iceland and Norway serve as models of successful transitions, with high reliance on geothermal and

hydropower resources. These examples highlight the advantages of abundant natural resources but underscore the challenges of replicating such models in less resource-endowed regions.



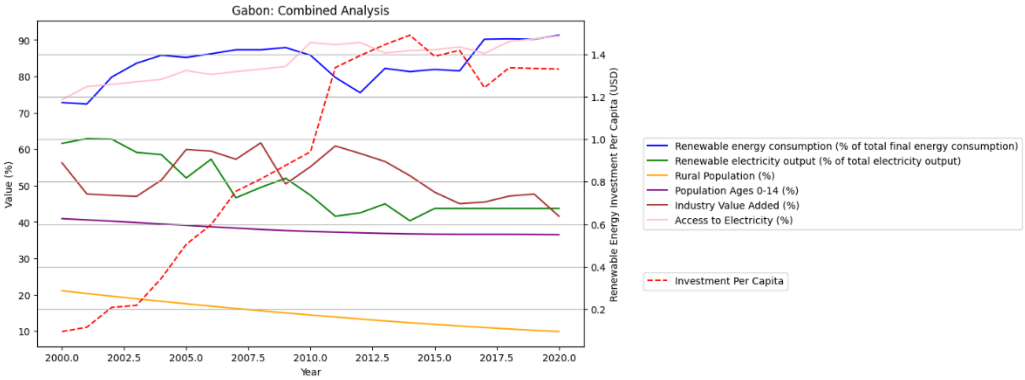
Multi-Variable Time Series: Norway



Multi-Variable Time Series: Iceland

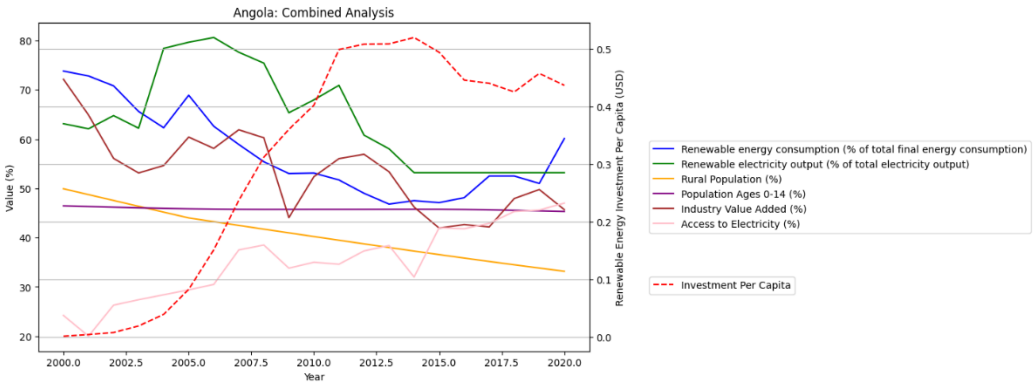
Sub-Saharan Africa:

Progress was inconsistent, with countries like Angola and Gabon facing unique challenges. In Angola, post-2015 increases in renewable energy consumption were driven by international aid rather than domestic policies. In Gabon, fluctuating outputs pointed to inefficiencies in infrastructure and grid management.



Multi-Variable Time Series: Gabon

Team 005: Nicholas Jerdack, Alexander Shropshire, Michael Daniels



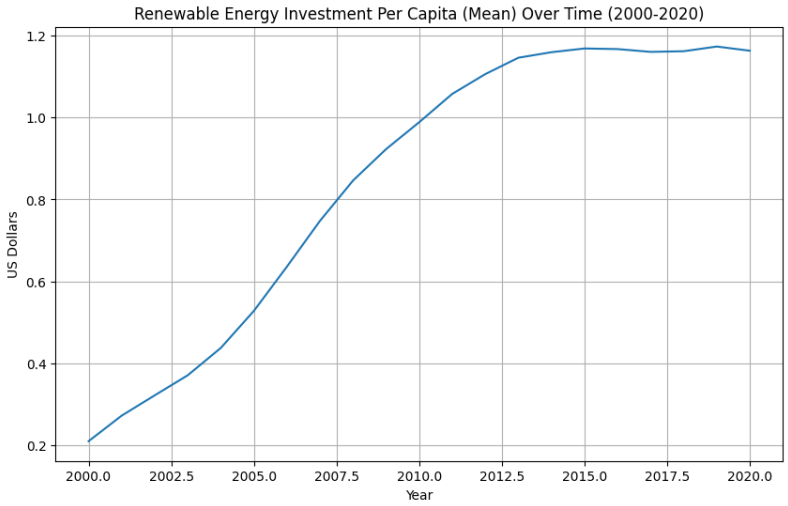
Multi-Variable Time Series: Angola

Temporal Patterns

Shifting Priorities in Energy Investment

Globally, renewable energy investments rose steadily from 2000 to 2015, driven by favorable policies and international climate agreements. However, the post-2015 stagnation suggests a shift from capacity-building to operational maintenance, influenced by both economic pressures such as the 2008 financial crisis and the COVID-19 pandemic as well as a need for reinvigorated financing mechanisms and policy incentives to sustain long-term progress.

Energy use per capita stabilized globally after 2010, reflecting a shift toward energy efficiency and changing consumption patterns. Meanwhile, GDP per capita grew steadily until the mid-2010s, after which economic volatility introduced challenges for maintaining renewable energy investments.

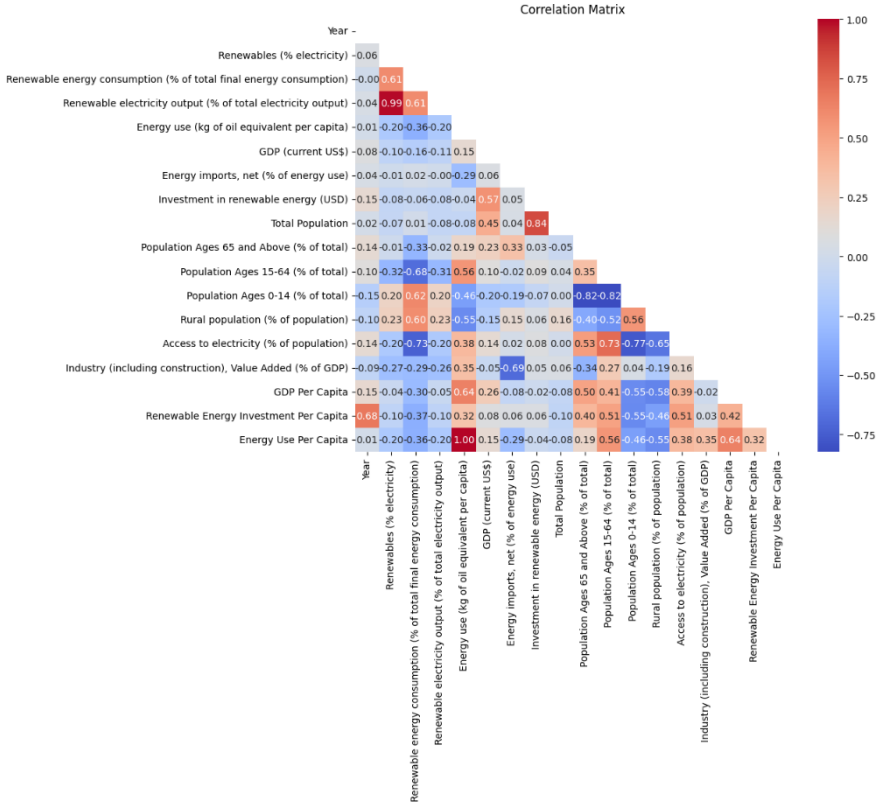


Average Renewable Energy Investment Per Capita is steadily increasing from 2000 through 2012, before plateauing for a ~8+ years

Key Interactions and Relationships

Correlation Analysis

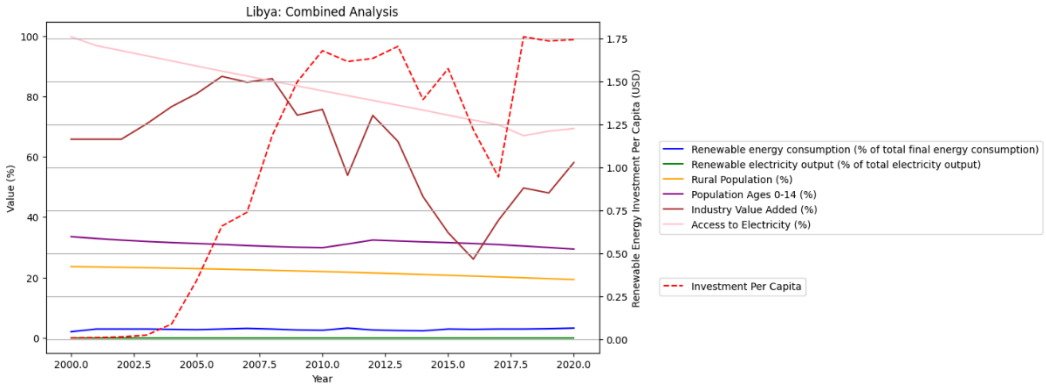
- **Renewable Electricity Output and Renewable Energy Consumption**:
  A near-perfect correlation highlights the direct relationship between production and consumption, as expected. Both features, our chosen target variables for the modeling process, are key measures in understanding how a nation is doing in sustainable energy.
- **GDP Per Capita and Renewable Investment Per Capita**:
  A moderate positive correlation indicates wealthier nations are better positioned to invest in renewable energy. However, exceptions like Costa Rica demonstrate how robust policies can drive renewable adoption even in less wealthy nations.
- **Renewable Investment and Fossil Fuel Consumption**:
  A negative correlation underscores the transition from fossil fuels to renewables, though progress varies. For instance, China leads in renewable investment but remains the largest consumer of coal, exemplifying the dual challenges of transitioning from fossil fuels (Complete correlation matrix below).



All-Country Multi-variable Pearson's Correlation Coefficient Matrix

Outliers and Systemic Barriers

Anomalies reveal barriers to renewable energy transitions. In Libya, there is political instability causing volatile investment patterns and disrupting long-term planning. In Angola, there is increased renewable energy consumption with minimal investment suggesting a reliance on external funding. In Gabon, there is erratic electricity output which highlights technical and financial constraints in integrating renewables into national grids. In India, there are persistent gaps between electricity output and access which underscores the complexity of addressing energy poverty while scaling renewable infrastructure. See combined analyses charts for India, Angola, and Gabon above and Libya below:



Multi-Variable Time Series: Libya

EDA Conclusions

The findings reveal that a complex interplay of economic, political, and infrastructural factors influences renewable energy transitions:

- Wealthier nations exhibit more substantial progress but require consistent policy support and technological innovation for long-term success.
- Emerging economies face unique challenges in balancing economic growth with sustainable energy transitions, such as volatile infrastructure and policies.
- Outliers emphasize the importance of governance, international cooperation, and tailored strategies to address country-specific barriers.

These insights provide a foundation for understanding global energy dynamics and inform targeted recommendations for accelerating renewable energy adoption.

PCA Analysis

Clustering Insights

The Principal Component Analysis (PCA) and clustering techniques provided critical insights into renewable energy adoption and investment trends over time. To track these changes, the PCA analysis focused on four key years: **2000**, **2005**, **2010**, and **2015**. These snapshots highlight how countries shifted their energy strategies and investment priorities across two decades. Additionally, the analysis explored mean renewable energy use by continent and examined specific countries with unique energy profiles to uncover nuanced trends.

Analysis of PCA Clusters

The PCA transformed complex energy datasets into two principal components, capturing most of the variability in the data (75-80% across years). By visualizing countries in the PCA space, clusters emerged based on renewable energy investment levels, energy output, and dependence on fossil fuels. These clusters represent distinct groups of countries.

Key Observations from the PCA Axes

Renewable Energy Development (x-axis):

- Countries further to the right on the x-axis exhibit higher renewable energy adoption and infrastructure development levels.
- Clean energy leaders like Norway, Iceland, and Denmark dominate this space, reflecting their advanced renewable energy systems.

Oil Production (y-axis):

- Countries higher on the y-axis are characterized by significant oil production.
- The Gulf States, including Saudi Arabia, Qatar, and Kuwait, are clustered at the top left, representing their dominance in oil production but limited renewable energy adoption.

Notable Country Positions

Norway:
Norway stands out as a unique case in the top right corner. It is both a leading oil producer and a global leader in renewable energy. This reflects Norway's balanced approach to leveraging its natural oil reserves while aggressively investing in hydropower and other renewable energy technologies.

Gulf States:
Countries like Saudi Arabia, Qatar, Kuwait, and the UAE are grouped in the **top left corner**. Their high oil production levels dominate their energy profiles, and their limited investment in renewable energy leaves them lagging in renewable development. Despite some diversification efforts, these nations have yet to make considerable progress in renewable energy transitions.

Clean Energy Leaders:
Nations like Norway, Iceland, Sweden, and Denmark consistently emerged as leaders in renewable energy. Their dominance in the PCA space reflects their heavy reliance on geothermal, hydropower, and wind energy. Notable trends include Norway and Iceland maintaining their strong position due to their robust infrastructure and abundant natural resources. At the same time, other European nations like Sweden and Denmark showed incremental shifts toward greater energy efficiency.

BRICS Economies:
Countries such as China, India, Brazil, and South Africa displayed dynamic changes in their PCA positioning. For example, China's rapid industrialization led to significant investments in renewable energy, positioning it closer to renewable leaders by 2015. Conversely, India and Brazil showed more gradual shifts due to infrastructure and policy limitations.

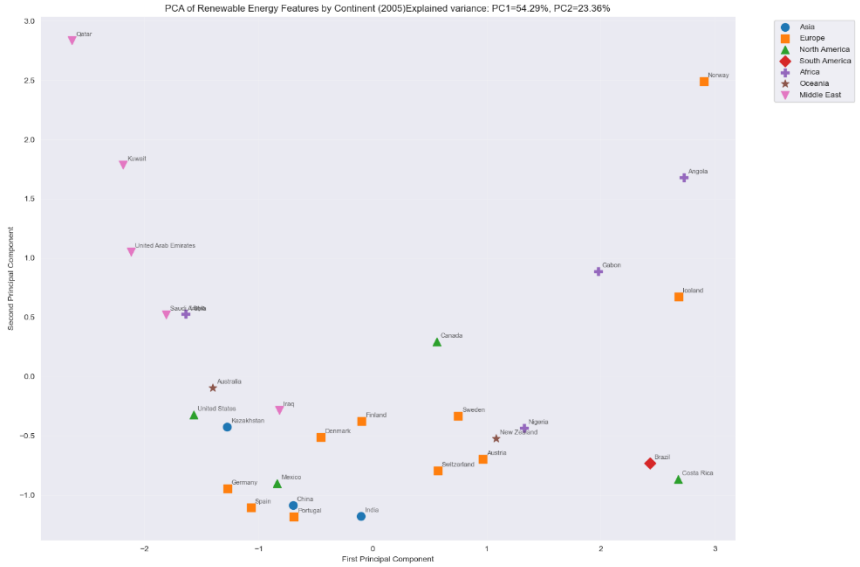
Changes Across Key Years

2000:
Countries were broadly separated by their dependency on fossil fuels versus renewables. Oil-exporting nations dominated the lower end of renewable energy contributions, while Europe and Scandinavia stood out as clean energy pioneers.

Team 005: Nicholas Jerdack, Alexander Shropshire, Michael Daniels

Year 2000: 2-Component Principal Component Analysis of renewable energy variables by continent

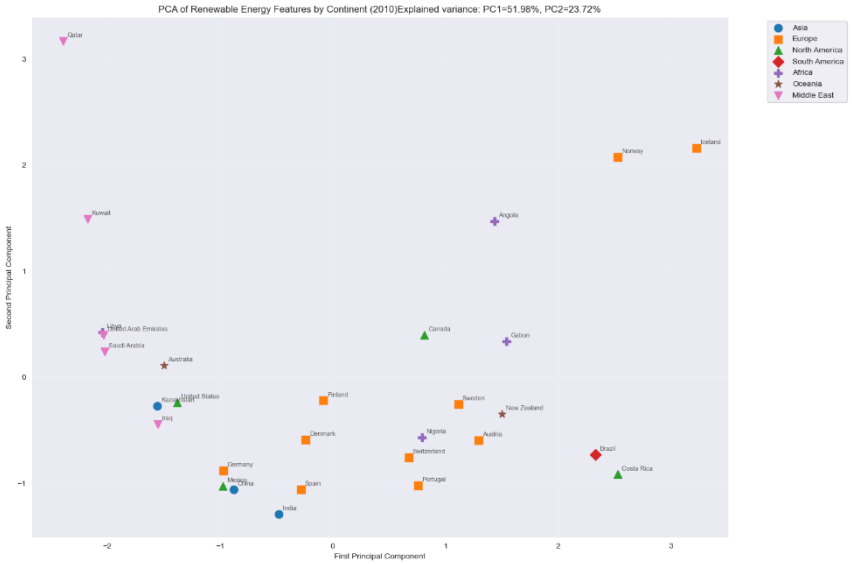
2005:
Gradual diversification in the Middle East was observed, though still limited. European nations such as Germany and Portugal began showing notable increases in renewable investments.



Year 2005: 2-Component Principal Component Analysis of renewable energy variables by continent

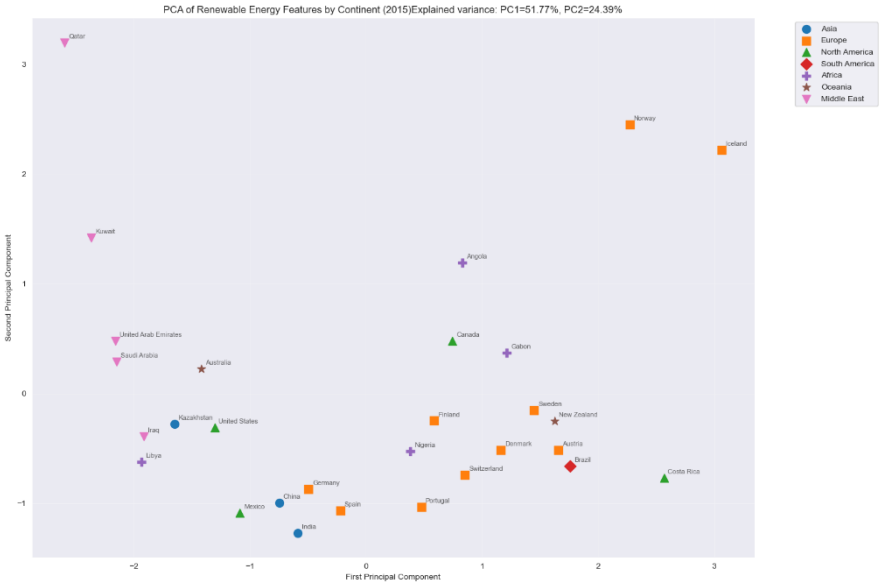
2010:
Emerging economies like China and Brazil showed upward movement in renewable adoption. Meanwhile, the Middle East remained concentrated in the oil-dependent cluster, with little visible progress.

Team 005: Nicholas Jerdack, Alexander Shropshire, Michael Daniels



Year 2010: 2-Component Principal Component Analysis of renewable energy variables by continent

2015:
A clear distinction emerged between nations advancing rapidly in renewable energy (e.g., China and Brazil) and those stagnating (e.g., Angola, Libya, and other oil exporters). European countries continued solidifying their leadership, with nations like Sweden and Austria achieving significant efficiency gains.



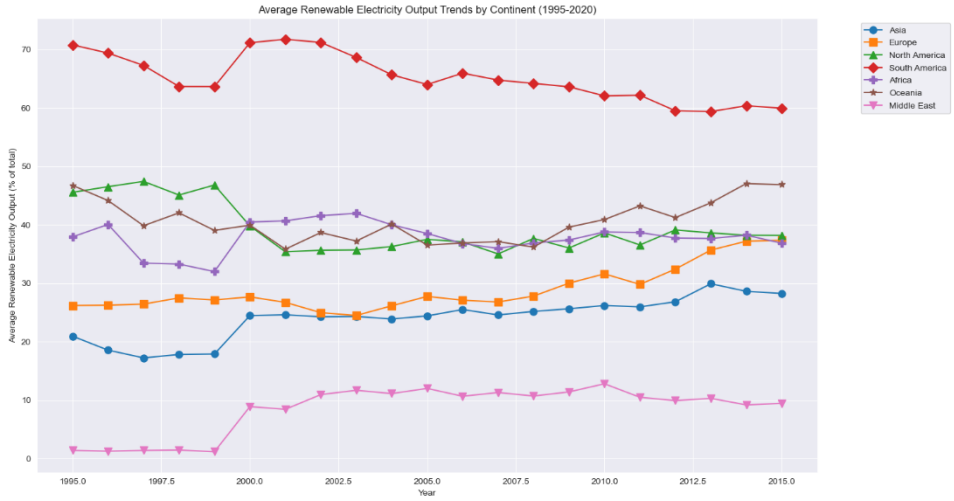
Year 2015: 2-Component Principal Component Analysis of renewable energy variables by continent

Mean Renewable Energy Use by Continent

To complement the PCA, the mean renewable energy use by continent (1995–2020) was analyzed:

- **Europe:** Consistently led renewable electricity output, reflecting strong policy support and investment.
- **North America:** Showed gradual improvements, driven primarily by the United States and Canada.

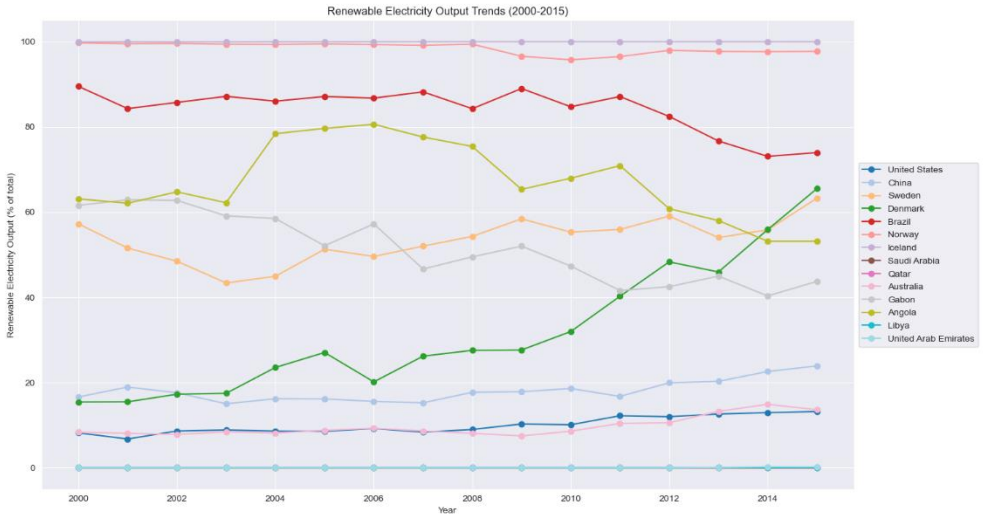
- **Asia:** Displayed variability, with China's rapid progress contrasting with slower adoption in other nations.
- **Africa and the Middle East:** Continued to lag, with minimal renewable contributions. Exceptions include Angola, which showed sharp increases post-2010 due to international aid.



From 1995: Average Renewable Electricity Output by continent

Country-Specific Insights

- **Norway:** A unique case that balances significant oil production with leadership in renewable energy. This duality highlights Norway's ability to capitalize on its oil wealth while maintaining its commitment to clean energy development.
- **Gulf States:** Positioned as strong oil producers but weak in renewable energy adoption, reinforcing their dependence on fossil fuels.
- **China:** Demonstrated noteworthy progress, moving rightward on the x-axis over time, driven by aggressive solar and wind investments.
- **Brazil:** Maintained reliance on hydropower but struggled to diversify into other renewable sources, keeping it closer to the middle of the PCA space.



From 2000: Average Renewable Electricity Output by key country

Team 005: Nicholas Jerdack, Alexander Shropshire, Michael Daniels

Summary of Findings

The PCA and clustering analyses revealed the following key insights:

- **Renewable Energy Development (x-axis)** and **Oil Production (y-axis)** provided a meaningful framework for understanding global energy strategies.
- **Norway** stands out in the top right quadrant for its rare balance of oil production and renewable energy leadership.
- **Gulf States** remained concentrated in the **top left**, underscoring their entrenched reliance on fossil fuels.
- **Emerging Economies** showed dynamic shifts toward renewable energy, reflecting diverse progress and challenges.

These findings underscore the global disparities in renewable energy adoption and provide a roadmap for targeted interventions to accelerate energy transitions.

Modeling Methodologies

The objective of modeling after analyzing historical trends was to predict the most accurate future values for Renewable Electricity Output (% of Total Electricity Output) and Renewable Energy Consumption (% of Total Final Energy Consumption) using ARIMA, Random Forest, and ARIMAX models. The following methodology outlines our modeling processes, technical decisions, and insights derived from the results.

Data Preparation

The preprocessing steps were detailed earlier. Here, we focus on the rationale for using natural logarithmic transformations alongside regular scaling. Log transformations stabilize variance, mitigate heteroscedasticity, and reduce the impact of outliers, particularly for countries with significant discrepancies in renewable energy metrics. Log scaling enhances the suitability of time series models like ARIMA and ARIMAX by transforming the data into a more normalized distribution. Both regular and log-transformed datasets were used in all subsequent modeling steps to compare the impact of these transformations on predictive accuracy.

Stationarity Testing

Stationarity is a fundamental assumption for ARIMA-based models. To ensure the suitability of the time series data, we applied the following methods to assess and enforce stationarity:

1. **Dickey-Fuller Test**: Initial tests indicated widespread non-stationarity in both target variables across countries. For example, under regular scaling, the Dickey-Fuller p-values for Renewable Electricity Output in China and Brazil were 0.73 and 0.93, respectively. Such high p-values required something like differencing to address the presence of unit roots.
2. **First Differencing**: After differencing, stationarity was confirmed with p-values below 0.05 for both countries, reflecting a successful transformation.
3. **ACF/PACF Visual Analysis**: Pre-differencing plots revealed slow decay in autocorrelations, while post-differencing plots showed sharp truncations at low lags, confirming stationary patterns.

Going into more detail below, we used a Dickey Fuller p-value threshold of 0.05 meaning that if the value was above that, the countries did not have rejected null hypotheses that the data is nonstationary.

The variable Renewable electricity output (% of total electricity output) had a median (more than half of the countries) reflecting varying mean and/or variance with time, which is not ideal for many time series

models [see breakdown of exact values in the Appendix]. For the first differenced Renewable electricity output (% of total electricity output) data, the median p-value was 0.0000001. This p-value is much less than 0.05, indicating that either more than half of the countries had significant p-values or using first order differencing more than half of the time made the data likely stationary. This may give an indication that the ARIMA and ARIMAX models (that parameter tune the differencing to make the data stationary) will have differencing of the variable data for at least one of the countries, meaning the value for d will be more than 0. ACF and PACF plots (autocorrelation function and partial autocorrelation function plots) were also made, which usually for all countries agreed with the ADF test. For the ACF and PACF for the differenced data, the plots also agreed with the median ADF results. From all this, it seems like first order differencing of the Renewable electricity output (% of total electricity output) could be used to make the data stationary.

Many countries showed what was shown in the medians as above for Renewable electricity output (% of total electricity output). Countries like Australia, Gabon, Angola, United Arab Emirates, and Costa Rica already had statistically significant p-values for the undifferenced column data. In these cases, differencing was not needed to make the data likely stationary. There ADF tests for the first order differenced data were then not needed. For the rest of the countries, the p-values for the undifferenced data and differenced data followed the same trend as what the medians showed. This showed that the Renewable electricity output (% of total electricity output) variable was stationary for some countries and non-stationary for others. For the plots, countries like Australia, Angola, Gabon, United Arab Emirates, and Costa Rica showed different results. This gave the indication that the undifferenced data for most countries could likely be nonstationary. In cases, the ADF test and the plots could lead to differing conclusions. In those cases, intuition of the data would need to be used. It would be a safer assumption to assume nonstationary data and let the ARIMA and ARIMAX models later on deal with it.

	Regular Scale Models - Medians Across Key Countries			
Target Variable	Dickey Fuller	1st Differenced Dickey Fuller	ACF/PACF Visual Judgement	Same Post-Differencing
Renewable electricity output (% of total electricity output)	0.21	0.0000001	Likely Non-Stationary	Likely Stationary
Renewable energy consumption (% of total final energy consumption)	0.64	0.09	Likely Non-Stationary	Likely Stationary

Median Stationarity Test Results Across Key Countries: Regular Scale

	Natural Log Scale Models - Medians Across Key Countries			
Target Variable	Dickey Fuller	1st Differenced Dickey Fuller	ACF/PACF Visual Judgement	Same Post-Differencing
Renewable electricity output (% of total electricity output)	0.34	0.00001	Likely Non-Stationary	Likely Stationary
Renewable energy consumption (% of total final energy consumption)	0.63	0.03	Likely Non-Stationary	Likely Stationary

Median Stationarity Test Results Across Key Countries: Natural Log Scale

For the variable Renewable energy consumption (% of total final energy consumption), the original data median for the selected countries gave a p-value of 0.64 for the ADF test, which is more than half of the time above the threshold of 0.05 meaning that the data was likely nonstationary like the previous variable [see breakdown of exact values in the Appendix]. On the other hand, for the first order differencing of the Renewable energy consumption (% of total final energy consumption) data, the ADF test gave median p-value of 0.09, which more than half of the time is also above the threshold of 0.05 meaning the first order differencing of Renewable energy consumption (% of total final energy consumption) has a median that is still likely nonstationary. For this variable, first order differencing did not make the data likely stationary more than half of the time meaning that more than half of the countries had higher p-values or higher orders of differencing may be needed, which may be able to be found through parameter tuning of ARIMA and ARIMAX, or a different way of making the data stationary may be needed. For the ACF and PACF plots, the original column of data usually showed that the data is likely nonstationary, which agrees

with the ADF test. On the other hand, the first order differenced ACF and PACF plots showed that usually, the data was likely stationary, giving an opposite conclusion to the ADF test. The ADF p-value was closer to 0.05 for the median, which could indicate that the data may be closer to stationarity for many countries than the p-value indicates.

For this variable and the countries Brazil, Libya, and Costa Rica, the undifferenced data showed to already be likely stationary from the ADF test p-values. In terms of the differenced data, the countries China, Qatar, Gabon, Angola, India, and the United Arab Emirates did not have p-values for the differenced data that indicated likely stationary data. In terms of the plots of the undifferenced data, Brazil, Saudi Arabia, Libya, United Arab Emirates, and Costa Rica gave opposite results. For the differenced data, the plots for China, Qatar, Gabon, Angola, India, and the United Arab Emirates gave opposite results. Again, the ADF test and the plots could indicate different conclusions, so it would be best for the ARIMA and ARIMAX models to best determine what is needed.

Because of the phenomenon described, natural log transformations, used to stabilize the variance, reduced the need for differencing in some cases. For example, Gabon's Renewable Output time series exhibited stationarity without differencing under log scaling. While the overall trends remained consistent between regular and log-transformed data, the latter often provided smoother series, aiding downstream modeling.

For different countries, it would be optimal to make each variable stationary by different methods. In some cases, logging the data may lead to better results, but in others, keeping the data as it is may be better. A custom approach for each was proven to be best rather than a single broad-stroke approach.

Model Development

ARIMA Modeling

ARIMA was chosen as the baseline model because it can capture univariate temporal dependencies. A grid search was used to optimize (P, D, Q) parameters over ranges of P = [0, 10], D = [0, 2], and Q = [0, 10]. Parameter configurations varied across countries, reflecting heterogeneity in temporal structures. For example, Renewable Electricity Output in Brazil required (P, D, Q) = (1, 0, 2), capturing its moderate autocorrelation patterns. At the same time, India's Renewable Energy Consumption was best modeled with (2, 0, 1), indicating a higher degree of persistence in lag dependencies. ARIMA consistently delivered impressive performance across both regular and log-transformed data. For Renewable Electricity Output under regular scaling, ARIMA achieved a median test MSE of 0.82 across all countries, with particularly low values for Norway (0.30) and the United States (0.09).

Random Forest was employed as a predictive model and a feature-selection tool for ARIMAX. Its ability to handle high-dimensional data and identify essential predictors made it valuable for identifying exogenous variables. The most frequently selected predictors included macroeconomic indicators (e.g., GDP, investment) and demographic factors (e.g., rural and urban population). For instance, "Energy Imports" and "Industry" emerged as critical drivers of Renewable Electricity Output in the United States, while "GDP" and "Rural Population" were highly significant for India. While Random Forest performed adequately as a standalone model, it often fell short compared to ARIMA. For Renewable Electricity Output, Random Forest recorded a median test MSE of 0.09 across all countries under regular scaling, marginally outperforming ARIMA in a few cases. However, it consistently underperformed for Renewable Energy Consumption, with a median test MSE of 2.51 under the same scale.

ARIMAX extended ARIMA by incorporating exogenous variables identified through Random Forest. Parameter tuning for ARIMAX followed the same process as ARIMA, with (P, D, Q) optimized for each country. While the inclusion of external factors aimed to enhance predictive accuracy, ARIMAX often underperformed compared to ARIMA. For instance, in the United States, ARIMAX recorded a test MSE of 14.91 for Renewable Electricity Output under regular scaling, significantly higher than ARIMA's 0.09.

Similar trends were observed under log scaling, where ARIMAX exhibited test MSE values exceeding 57 for certain countries. These results suggest that while exogenous variables provide valuable context, their integration may have introduced noise or overfitting in some cases, particularly in data-sparse regions.

As shown in both tables below, different amounts of variables were picked for each country. For the ARIMA and ARIMAX models, the p, d, and q parameters were the same in all cases with a handful of exceptions. This gave the indication that having the random forest picked variables did not help in modeling the variables too much because adding the exogenous variables would be expected to cause the ARIMAX model p, d, and/or q to decrease. This might indicate that the predictors need to be transformed or made stationary like the response variable or that these variables are not needed.

	Renewable electricity output (% of total electricity output)		
	Regular Scale Models		
Country Name	ARIMA (P, D, Q)	ARIMAX (P, D, Q)	Chosen Variables
United States	(1, 0, 1)	(1, 0, 1)	Energy imports
China	(1, 0, 1)	(1, 0, 1)	Industry
Brazil	(1, 0, 2)	(1, 0, 2)	Older Popul., Rural Popul.
Norway	(1, 0, 1)	(1, 0, 1)	GDP, Investment, Total Popul., Rural Popul.
Saudi Arabia	(1, 0, 0)	(1, 0, 0)	Total Popul., Rural Popul.
Australia	(1, 0, 1)	(1, 0, 1)	Investment, Older Popul.
Gabon	(1, 0, 1)	(1, 0, 1)	Energy imports, Investment, Middle-Age Popul.
Angola	(1, 0, 1)	(1, 0, 1)	Energy use, Rural Popul.
India	(1, 0, 1)	(1, 0, 1)	GDP, Older Popul., Rural Popul.
United Arab Emirates	(1, 0, 1)	(1, 0, 1)	Investment, Rural Popul.
Costa Rica	(1, 0, 1)	(1, 0, 1)	Energy use, Energy imports, GDP, Investment, Older Popul. , Rural

Renewable Output Parameter Tuning Results Across Key Countries: Regular Scale

	Renewable energy consumption (% of total final energy consumption)		
	Regular Scale Models		
Country Name	ARIMA (P, D, Q)	ARIMAX (P, D, Q)	Chosen Variables
United States	(1, 0, 1)	(1, 0, 1)	Investment, Older Popul., Rural Popul.
China	(3, 0, 0)	(3, 0, 0)	GDP, Older Popul., Younger Popul., Rural Popul.
Brazil	(1, 0, 1)	(1, 0, 1)	Energy Imports, Younger Popul., Rural Popul., Industry
Norway	(1, 0, 2)	(1, 0, 2)	Rural Popul.
Iceland	(1, 0, 1)	(1, 0, 1)	Energy Imports, Younger Popul., Rural Popul.
Saudi Arabia	(0, 0, 0)	(0, 0, 0)	Energy use, GDP, Energy imports, Investment, Older Popul., Industry
Qatar	(1, 0, 0)	(1, 0, 0)	Energy use, Younger Popul., Rural Popul.
Australia	(1, 0, 1)	(1, 0, 1)	GDP
Gabon	(1, 0, 1)	(1, 0, 1)	Older Popul., Younger Popul., Rural Popul.
Angola	(1, 0, 1)	(1, 0, 1)	Rural Popul.
Libya	(1, 0, 0)	(1, 0, 0)	Energy use
India	(2, 0, 1)	(2, 0, 1)	GDP, Energy imports, Older Popul., Younger Popul.
United Arab Emirates	(1, 0, 1)	(1, 0, 1)	Investment, Older Popul., Younger Popul.
Costa Rica	(1, 0, 1)	(1, 0, 1)	GDP, Energy imports, Older Popul., Younger Popul., Rural Popul.

Renewable Consumption Parameter Tuning Results Across Key Countries: Regular Scale

From the variables that random forest found as most important to use, the tables below show how many countries used each variable across the 2 different target variable modeling efforts.

Renewable Output Variable **Renewable Consumption Variable**

Variable Name	# countries that used variable
Rural Population	7
Investment	5
Older Population	4
Energy Imports	3
GDP	3
Total Population	2
Energy use	2
Industry	1
Middle-aged population	1

Variable Name	# of countries that used variable
Rural population	9
Younger Population	8
Older population	7
GDP	5
Energy imports	5
Investment	3
Energy Use	3
Industry	2

In the cases of the natural log transformed models, the variables investment, energy use, energy imports, and rural population had at least 5 countries with these variables in the models. In the unlogged case, it had been variables that related to population seen above. The only one for the logged case used related to population was the % rural population. Otherwise for the logged model, it was more monetary variables. Like before, the ARIMA and ARIMAX models had the same parameter values, showing that having the predictors may not have changed the model or helped. This will have to be explored more in terms of the MSE values.

Evaluation Metrics and Results

Mean Squared Error (MSE) was used as the primary evaluation metric, reflecting its sensitivity to significant errors (a critical consideration in energy forecasting). Train and Test MSE values were calculated for all models across countries and scales, revealing insight about the countries, each approach, and how both interact.

Model Comparisons

In terms of ARIMA, random forest, and ARIMAX, the model that did the best in terms of training MSE for the unlogged **Renewable electricity output (% of total electricity output)** response was ARIMA in most cases, since the median had the smallest value. Random forest was not too far behind, but ARIMAX was a good quantity behind. In terms of testing MSE, random forest did the best followed by ARIMA. ARIMAX was even further behind. This all confirms what was found above in that ARIMA and ARIMAX had the same hyperparameter values in all country cases, proving more that the exogenous variables did not help. It was expected that random forest would do the worst in most cases, since it is not a model made specifically for time series, but it did as well as ARIMA in training MSE but did better in terms of testing MSE (which is more important).

In terms of the training MSE, the countries United States, Norway, Australia, India, and Costa Rica did the best with ARIMA. Some of the other countries like Gabon and Angola had extremely high MSE values for ARIMA and ARIMAX. The highest training error was for the ARIMA model for Angola and ARIMA still had the smallest median MSE showing that it did better in most cases. In terms of the testing MSE, ARIMA did better for Norway and Costa Rica. Random forest did better for most of the rest. As the median table showed, this is unexpected since it is not a model made for time series data.

For the **Renewable energy consumption (% of total final energy consumption)** ARIMA did the best in terms of training MSE, but ARIMAX did the best in terms of testing MSE for the median of the MSE

values for all the countries. This does not agree with what was found before in terms of the training MSE since the parameters were always the same. One would have expected that adding exogenous variables would decrease the ARIMA p, d, and/or q since exogenous variables would help to account for some variability in the model, but that was not the case from looking at the parameters. Looking at the testing MSE, the exogenous variables seemed to help.

To investigate it in finer detail, the Appendix has the country-level details in table format.

	Regular Scale Models - Medians Across Key Countries					
	Train MSE			*Test MSE*		
Target Variable	**ARIMA**	**Random Forest**	**ARIMAX**	**ARIMA**	**Random Forest**	**ARIMAX**
Renewable electricity output (% of total electricity output)	2.74	2.96	5.88	0.82	0.09	23.54
Renewable energy consumption (% of total final energy consumption)	1.34	2.60	8.36	2.08	2.51	1.64

Median Train and Test MSE Results across all countries for all models tested: Regular Scale

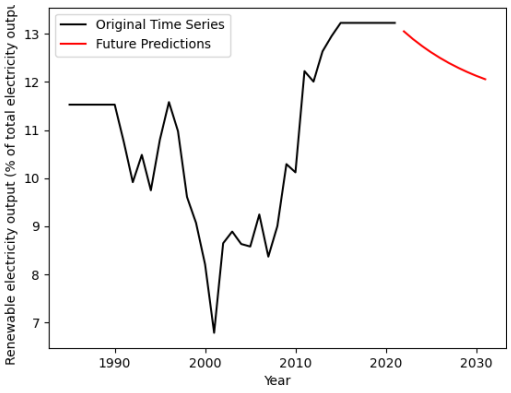
	Natural Log Scale Models - Medians Across Key Countries					
	Train MSE			*Test MSE*		
Target Variable	**ARIMA**	**Random Forest**	**ARIMAX**	**ARIMA**	**Random Forest**	**ARIMAX**
Renewable electricity output (% of total electricity output)	0.01	0.01	0.02	1.41	0.22	57.64
Renewable energy consumption (% of total final energy consumption)	0.00	0.01	0.04	4.86	3.98	15.52

Median Train and Test MSE Results across all countries for all models tested: Natural Log Scale

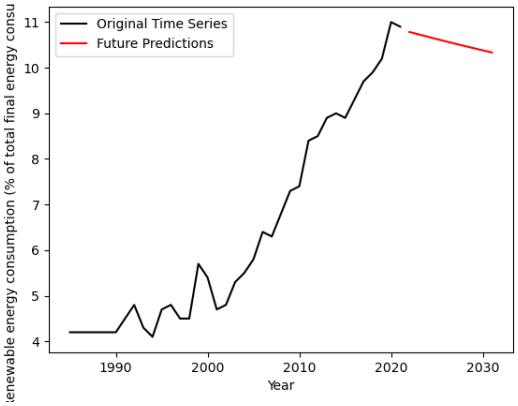
Since the parameter output differencing was never used, the fit models are really ARMA and ARMAX models that did not use differencing. These were described in the theoretical sections above. For all these variables, transformation, country combinations, it could be seen that a variety of models did better for many different cases. This proves the point that the better models would widely vary in these three idea combinations, showing that each country should be modeled with its own model. The best model found for each country in terms of testing MSE could be used to make future predictions for each variable.

Key Insights and Implications

Renewable Output



Renewable Consumption



10-Year Future Prediction Plots of the United States: Both metrics set to decrease from local highs

1. **ARIMA's Strength as a Baseline Model:**
   ARIMA demonstrated robust performance across variables and scales, consistently outperforming ARIMAX and Random Forest. Its test MSE values were lower and more stable across countries, reinforcing ARIMA's reliability as a forecasting tool for univariate time series with minimal need for external information.
2. **Challenges with ARIMAX:**
   Despite the theoretical advantages of incorporating exogenous variables, ARIMAX often underperformed relative to ARIMA. This was particularly evident under log scaling, where ARIMAX recorded a median test MSE of 57.64 for Renewable Electricity Output. These findings highlight the challenges of effectively integrating external factors, which may introduce noise or require more sophisticated feature selection and regularization techniques.
3. **Effectiveness of Exogenous Variables:**
   Random Forest identified meaningful predictors, such as "Energy Imports," "GDP," and "Rural Population," that shaped renewable energy trends. These insights have significant policy implications, particularly for countries like the United States, where energy imports and industrial activity were primary drivers of Renewable Electricity Output. However, ARIMAX's inability to translate these predictors into improved forecasting accuracy underscores the need for further refinement in how exogenous variables are incorporated.
4. **The Role of Log Scaling:**
   Log transformations stabilized variance and reduced skewness, improving model performance in several cases. For example, Norway's test MSE for Renewable Energy Consumption decreased from 7.71 under regular scaling to 7.05 under log scaling. However, the benefits of log scaling were inconsistent, particularly for ARIMAX, suggesting that its utility is model- and context-dependent.
5. **Country-Specific Variability:**
   Model performance varied significantly across countries, reflecting the impact of data quality and regional energy dynamics. Angola and Gabon posed challenges due to data sparsity or irregularities. Their persistently high MSE values underscore the importance of localized modeling strategies and the need for high-quality, granular data. Conversely, countries with well-structured energy markets, such as Norway and the United States, demonstrated low MSE values, indicating strong model applicability. These discrepancies highlight the need for tailored approaches in regions with poor data quality and volatile energy policies and economies.
6. **Policy and Practical Implications:**
   The results suggest that while ARIMA is effective for identifying historical trends and generating short-term forecasts, its utility in policymaking may be limited without including external factors. Random Forest's identification of key drivers, such as investment and energy imports, highlights potential areas for intervention to improve renewable energy adoption. That said, we must be conscious that modeling factors are much different than the action-oriented shifting of international energy behaviors, economies, and infrastructure.
7. **Future Improvements:** There are several improvements that should be explored to achieve more depth in our existing process. First, we could consider extending the list of potential parameters exposed to our grid search code which, even if at a higher computational expense, will receive more to choose from in its search for the optimal set. Second, we could seek to make exogenous variables stationary or apply data stationary methods differently to test the magnitude of the potentially improved results. We might also seek to assess the tradeoff of using an evaluation metric beyond MSE and consider what other error metrics might mean for our model comparison.

Conclusion

In conclusion, this project highlights the predictability of renewable energy output and consumption, emphasizing the noteworthy influence of economic and political forces on progress. Denmark's example vividly illustrates the relationship between investment effectiveness and renewable energy adoption. Over the past 20 years, Denmark has transitioned from relying on less than half of its energy from renewable

sources to nearly exclusively using renewables, demonstrating how sustained, strategic investments can yield transformative outcomes.

Our findings reaffirm that publicly available data can provide meaningful insights into global energy trends. However, while the data enables us to identify patterns and forecast trajectories, the challenge remains in determining each nation's most impactful, action-oriented strategies at any moment. As nations grow wealthier and more capable of pioneering renewable energy, the need for localized, data-driven approaches becomes increasingly essential. This study is a foundation for future work that combines analytical tools with targeted policy and industrial strategies to accelerate the global transition toward sustainable energy systems.

References

[1] Genoni, Maria Eugenia, et al. The World Bank. (n.d.). *World Bank Open Data - World Development Indicators.* Retrieved from http://api.worldbank.org/v2. Accessed October 2024.

[2] Hossain, B. (2022). *Renewable energy worldwide 1965–2022 [Data set].* Kaggle. Retrieved from https://www.kaggle.com/datasets/belayethossainds/renewable-energy-world-wide-19652022. Accessed October 2024.

[3]Hoa, Pham Xuan, et al. "Determinants of the renewable energy consumption: The case of Asian countries." Heliyon, vol. 9, no. 12, Dec. 2023, https://doi.org/10.1016/j.heliyon.2023.e22696.

[4]Liu, Xintian. "Impacts of El Nino on renewable energy industry and fishery and aquaculture industry around the Pacific." Theoretical and Natural Science, vol. 37, no. 1, July 2024, https://doi.org/10.54254/2753-8818/37/20240218.

[5]Motyka, M., Thomson, J., Hardin, K., & Amon, C. (2023, December 4). *2024 renewable energy industry outlook: Renewables set for a variable-speed takeoff as historic investment, competitiveness, and demand propel their development, while also exacerbating grid, supply chain, and workforce challenges*. Deloitte Research Center for Energy & Industrials. https://www2.deloitte.com/us/en/insights/industry/renewable-energy/renewable-energy-industry-outlook.html

Python Packages Used:

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... & Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825), 357-362. https://doi.org/10.1038/s41586-020-2649-2

McKinney, W. (2010). Data structures for statistical computing in Python. In *Proceedings of the 9th Python in Science Conference* (pp. 51-56). https://doi.org/10.25080/Majora-92bf1922-00a

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3), 90-95. https://doi.org/10.1109/MCSE.2007.55

Python Software Foundation. (2023). *pathlib — Object-oriented filesystem paths*. Python Documentation. https://docs.python.org/3/library/pathlib.html

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with Python. In *Proceedings of the 9th Python in Science Conference* (pp. 57-61). https://doi.org/10.25080/Majora-92bf1922-011

Smith, T., & Davidson-Pilon, C. (2019). *pmdarima: ARIMA and Seasonal-ARIMA models in Python*. https://www.alkaline-ml.com/pmdarima/

Nikolay Manchev. (2020). tscv: Time series cross-validator. GitHub Repository. https://github.com/NiklasRosenstein/tscv

Python Software Foundation. (2023). *ast — Abstract syntax trees*. Python Documentation. https://docs.python.org/3/library/ast.html

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830. https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

Appendix

MSE Tables:

Summaries

	Regular Scale Models - Medians Across Key Countries					
	Train MSE			*Test MSE*		
Target Variable	**ARIMA**	**Random Forest**	**ARIMAX**	**ARIMA**	**Random Forest**	**ARIMAX**
Renewable electricity output (% of total electricity output)	2.74	2.96	5.88	0.82	0.09	23.54
Renewable energy consumption (% of total final energy consumption)	1.34	2.60	8.36	2.08	2.51	1.64

	Natural Log Scale Models - Medians Across Key Countries					
	Train MSE			*Test MSE*		
Target Variable	**ARIMA**	**Random Forest**	**ARIMAX**	**ARIMA**	**Random Forest**	**ARIMAX**
Renewable electricity output (% of total electricity output)	0.01	0.01	0.02	1.41	0.22	57.64
Renewable energy consumption (% of total final energy consumption)	0.00	0.01	0.04	4.86	3.98	15.52

Regular Scale

	Renewable electricity output (% of total electricity output)					
	Regular Scale Models					
	Train MSE			*Test MSE*		
Country Name	**ARIMA**	**Random Forest**	**ARIMAX**	**ARIMA**	**Random Forest**	**ARIMAX**
United States	1.05	1.68	2.48	0.39	0.01	14.91
China	2.83	2.70	9.28	1.36	0.19	49.44
Brazil	17.14	15.03	37.77	17.16	0.92	105.24
Norway	0.83	1.10	1.02	0.30	0.35	30.65
Saudi Arabia	0.00	0.00	0.00	0.00	0.00	0.00
Australia	1.13	3.22	2.07	0.82	0.16	13.13
Gabon	292.05	22.17	2628.92	1.99	0.02	159.70
Angola	40051.95	53.11	2976.31	24.44	2.63	247.50
India	2.74	3.75	88.31	0.33	0.00	16.44
United Arab Emirates	0.00	0.01	0.00	0.00	0.00	0.02
Costa Rica	11.20	21.07	161.53	18.38	18.79	929.49

Natural Log Scale

	Renewable electricity output (% of total electricity output)					
	Natural Log Models					
	Train MSE			*Test MSE*		
Country Name	**ARIMA**	**Random Forest**	**ARIMAX**	**ARIMA**	**Random Forest**	**ARIMAX**
United States	0.01	0.02	0.03	0.70	0.05	46.57
China	0.01	0.01	0.03	1.99	0.20	80.18
Brazil	0.00	0.00	0.00	4.09	0.59	71.75
Norway	0.00	0.00	0.00	0.22	0.33	32.96
Iceland	0.00	0.00	0.00	0.00	0.00	0.00
Australia	0.01	0.03	0.02	1.25	0.27	31.61
Gabon	0.02	0.01	0.90	1.57	0.02	68.71
Angola	3.87	0.01	0.62	16.29	0.24	140.06
India	0.02	0.01	0.29	0.32	0.00	24.76
Costa Rica	0.00	0.00	0.02	19.15	19.81	1568.14

	Renewable energy consumption (% of total final energy consumption)					
	Natural Log Models					
	Train MSE			*Test MSE*		
Country Name	**ARIMA**	**Random Forest**	**ARIMAX**	**ARIMA**	**Random Forest**	**ARIMAX**
United States	0.01	0.02	0.05	2.61	1.68	0.31
China	0.00	0.02	0.00	0.13	3.98	15.52
Brazil	0.00	0.00	0.02	4.93	7.24	51.67
Norway	0.00	0.00	0.00	7.01	7.05	4.05
Iceland	0.00	0.00	0.05	6.33	4.61	56.17
Australia	0.01	0.01	0.05	4.86	3.41	1.09
Gabon	0.00	0.00	0.08	87.67	85.67	26.48
Angola	0.01	0.01	0.04	33.96	41.34	299.51
Libya	0.01	0.01	5.67	0.13	0.07	0.83
India	0.00	0.00	0.00	2.89	3.37	1.84
Costa Rica	0.00	0.01	0.03	1.94	1.37	18.05

Stationarity Tables:

Summaries

	Regular Scale Models - Medians Across Key Countries			
Target Variable	**Dickey Fuller**	**1st Differenced Dickey Fuller**	**ACF/PACF Visual Judgement**	**Same Post-Differencing**
Renewable electricity output (% of total electricity output)	0.21	0.0000001	Likely Non-Stationary	Likely Stationary
Renewable energy consumption (% of total final energy consumption)	0.64	0.09	Likely Non-Stationary	Likely Stationary

	Natural Log Scale Models - Medians Across Key Countries			
Target Variable	**Dickey Fuller**	**1st Differenced Dickey Fuller**	**ACF/PACF Visual Judgement**	**Same Post-Differencing**
Renewable electricity output (% of total electricity output)	0.34	0.00001	Likely Non-Stationary	Likely Stationary
Renewable energy consumption (% of total final energy consumption)	0.63	0.03	Likely Non-Stationary	Likely Stationary

Team 005: Nicholas Jerdack, Alexander Shropshire, Michael Daniels

Regular Scale

	Renewable electricity output (% of total electricity output)			
	Regular Scale Models			
Country Name	Dickey Fuller	1st Differenced Dickey Fuller	ACF/PACF Visual Judgement	Same Post-Differencing
United States	0.21	0.00	Likely Non-Stationary	Likely Stationary
China	0.81	0.00	Likely Non-Stationary	Likely Stationary
Brazil	0.92	0.00	Likely Non-Stationary	Likely Stationary
Norway	0.55	0.00	Likely Non-Stationary	Likely Stationary
Saudi Arabia	0.77	0.00	Likely Non-Stationary	Likely Stationary
Australia	0.02	Not needed	Likely Non-Stationary	Likely Stationary
Gabon	0.00	Not needed	Likely Non-Stationary	Likely Stationary
Angola	0.00	Not needed	Likely Non-Stationary	Likely Stationary
India	0.31	0.00	Likely Non-Stationary	Likely Stationary
United Arab Emirates	0.00	Not needed	Likely Non-Stationary	Likely Stationary
Costa Rica	0.03	Not needed	Likely Non-Stationary	Likely Stationary

	Renewable energy consumption (% of total final energy consumption)			
	Regular Scale Models			
Country Name	Dickey Fuller	1st Differenced Dickey Fuller	ACF/PACF Visual Judgement	Same Post-Differencing
United States	1.00	0.00	Likely Non-Stationary	Likely Stationary
China	0.65	0.09	Likely Non-Stationary	Likely Stationary
Brazil	0.04	Not needed	Likely Non-Stationary	Likely Stationary
Norway	0.47	0.00	Likely Non-Stationary	Likely Stationary
Iceland	0.98	0.03	Likely Non-Stationary	Likely Stationary
Saudi Arabia	0.94	0.00	Likely Stationary	Likely Stationary
Qatar	0.11	0.36	Likely Non-Stationary	Likely Stationary
Australia	0.99	0.00	Likely Non-Stationary	Likely Stationary
Gabon	0.63	0.22	Likely Non-Stationary	Likely Stationary
Angola	0.63	0.31	Likely Non-Stationary	Likely Stationary
Libya	0.03	Not needed	Likely Non-Stationary	Likely Stationary
India	0.67	0.23	Likely Non-Stationary	Likely Stationary
United Arab Emirates	0.95	0.74	Likely Stationary	Likely Stationary
Costa Rica	0.03	Not needed	Likely Non-Stationary	Likely Stationary

Natural Log Scale

	Renewable electricity output (% of total electricity output)			
	Natural Log Models			
Country Name	Dickey Fuller	1st Differenced Dickey Fuller	ACF/PACF Visual Judgement	Same Post-Differencing
United States	0.30	0.00	Likely Non-Stationary	Likely Stationary
China	0.73	0.00	Likely Non-Stationary	Likely Stationary
Brazil	0.93	0.00	Likely Non-Stationary	Likely Stationary
Norway	0.55	0.00	Likely Non-Stationary	Likely Stationary
Iceland	0.35	0.20	Likely Non-Stationary	Likely Stationary
Australia	0.87	0.00	Likely Non-Stationary	Likely Stationary
Gabon	0.01	Not needed	Likely Non-Stationary	Likely Stationary
Angola	0.00	Not needed	Likely Non-Stationary	Likely Stationary
India	0.33	0.00	Likely Non-Stationary	Likely Stationary
Costa Rica	0.03	Not needed	Likely Non-Stationary	Likely Stationary

	Renewable energy consumption (% of total final energy consumption)			
	Natural Log Models			
Country Name	Dickey Fuller	Differencing Dickey Fuller	ACF/PACF Visual Judgement	Same Post-Differencing
United States	1.00	0.00	Likely Non-Stationary	Likely Stationary
China	0.59	0.21	Likely Non-Stationary	Likely Stationary
Brazil	0.04	Not needed	Likely Non-Stationary	Likely Stationary
Norway	0.45	0.00	Likely Non-Stationary	Likely Stationary
Iceland	0.97	0.03	Likely Non-Stationary	Likely Stationary
Australia	0.97	0.00	Likely Non-Stationary	Likely Stationary
Gabon	0.63	0.24	Likely Non-Stationary	Likely Stationary
Angola	0.79	0.27	Likely Non-Stationary	Likely Stationary
Libya	0.06	0.19	Likely Non-Stationary	Likely Stationary
India	0.67	0.00	Likely Non-Stationary	Likely Stationary
Costa Rica	0.02	Not needed	Likely Non-Stationary	Likely Stationary

Parameter Tuning Tables:

Regular Scale

	Renewable electricity output (% of total electricity output)		
	Regular Scale Models		
Country Name	**ARIMA (P, D, Q)**	**ARIMAX (P, D, Q)**	**Chosen Variables**
United States	(1, 0, 1)	(1, 0, 1)	Energy imports
China	(1, 0, 1)	(1, 0, 1)	Industry
Brazil	(1, 0, 2)	(1, 0, 2)	Older Popul., Rural Popul.
Norway	(1, 0, 1)	(1, 0, 1)	GDP, Investment, Total Popul., Rural Popul.
Saudi Arabia	(1, 0, 0)	(1, 0, 0)	Total Popul., Rural Popul.
Australia	(1, 0, 1)	(1, 0, 1)	Investment, Older Popul.
Gabon	(1, 0, 1)	(1, 0, 1)	Energy imports, Investment, Middle-Age Popul.
Angola	(1, 0, 1)	(1, 0, 1)	Energy use, Rural Popul.
India	(1, 0, 1)	(1, 0, 1)	GDP, Older Popul., Rural Popul.
United Arab Emirates	(1, 0, 1)	(1, 0, 1)	Investment, Rural Popul.
Costa Rica	(1, 0, 1)	(1, 0, 1)	Energy use, Energy imports, GDP, Investment, Older Popul. , Rural

	Renewable energy consumption (% of total final energy consumption)		
	Regular Scale Models		
Country Name	**ARIMA (P, D, Q)**	**ARIMAX (P, D, Q)**	**Chosen Variables**
United States	(1, 0, 1)	(1, 0, 1)	Investment, Older Popul., Rural Popul.
China	(3, 0, 0)	(3, 0, 0)	GDP, Older Popul., Younger Popul., Rural Popul.
Brazil	(1, 0, 1)	(1, 0, 1)	Energy Imports, Younger Popul., Rural Popul., Industry
Norway	(1, 0, 2)	(1, 0, 2)	Rural Popul.
Iceland	(1, 0, 1)	(1, 0, 1)	Energy Imports, Younger Popul., Rural Popul.
Saudi Arabia	(0, 0, 0)	(0, 0, 0)	Energy use, GDP, Energy imports, Investment, Older Popul., Industry
Qatar	(1, 0, 0)	(1, 0, 0)	Energy use, Younger Popul., Rural Popul.
Australia	(1, 0, 1)	(1, 0, 1)	GDP
Gabon	(1, 0, 1)	(1, 0, 1)	Older Popul., Younger Popul., Rural Popul.
Angola	(1, 0, 1)	(1, 0, 1)	Rural Popul.
Libya	(1, 0, 0)	(1, 0, 0)	Energy use
India	(2, 0, 1)	(2, 0, 1)	GDP, Energy imports, Older Popul., Younger Popul.
United Arab Emirates	(1, 0, 1)	(1, 0, 1)	Investment, Older Popul., Younger Popul.
Costa Rica	(1, 0, 1)	(1, 0, 1)	GDP, Energy imports, Older Popul., Younger Popul., Rural Popul.

Natural Log Scale

	Renewable electricity output (% of total electricity output)		
	Natural Log Models		
Country Name	**ARIMA (P, D, Q)**	**ARIMAX (P, D, Q)**	**Chosen Variables**
United States	(1, 0, 1)	(1, 0, 1)	Energy imports
China	(1, 0, 1)	(1, 0, 1)	Rural Popul., Industry
Brazil	(1, 0, 1)	(1, 0, 1)	Energy Use, Older Popul., Rural Popul.
Norway	(1, 0, 1)	(1, 0, 1)	GDP, Investment, Younger Popul.
Iceland	(5, 0, 4)	(5, 0, 4)	Younger Popul.
Australia	(1, 0, 1)	(1, 0, 1)	Energy imports, Investment, Older Popul.
Gabon	(1, 0, 1)	(1, 0, 1)	Energy imports, Investment, Older Popul., Rural Popul.
Angola	(1, 0, 1)	(1, 0, 1)	Energy use, Rural Popul.
India	(1, 0, 1)	(1, 0, 1)	GDP, Older Popul., Rural Popul.
Costa Rica	(1, 0, 1)	(1, 0, 1)	Energy use, GDP, Energy imports, Investment, Older Popul.

	Renewable energy consumption (% of total final energy consumption)		
	Natural Log Models		
Country Name	**ARIMA (P, D, Q)**	**ARIMAX (P, D, Q)**	**Chosen Variables**
United States	(1, 0, 2)	(1, 0, 2)	GDP, Investment, Younger Popul., Rural popul.
China	(2, 0, 1)	(2, 0, 1)	Energy use, GDP, Energy imports, Older popul.
Brazil	(1, 0, 1)	(1, 0, 1)	Energy imports, Investment, Older popul., Industry
Norway	(1, 0, 2)	(1, 0, 2)	GDP, Investment, Rural Popul.
Iceland	(1, 0, 1)	(1, 0, 1)	Energy use, Energy imports, Investment, Younger Popul., Rural Popul.
Australia	(1, 0, 1)	(1, 0, 1)	Industry
Gabon	(1, 0, 2)	(1, 0, 2)	Energy use, Energy imports, Investment
Angola	(1, 0, 1)	(1, 0, 1)	Energy use, Investment, Rural Popul.
Libya	(1, 0, 0)	(1, 0, 0)	Energy use, Younger Popul.
India	(2, 0, 1)	(2, 0, 1)	Energy use, GDP, Energy imports, Older popul.
Costa Rica	(1, 0, 1)	(1, 0, 1)	Energy imports, Younger Popul., Rural Popul.