Balancing Economic Growth and Environmental Sustainability: A Data-Driven Analysis of Energy Consumption Patterns

A graduate project submitted in partial fulfillment of the requirements
For the degree of Master of Science in Business Analytics

In Collaboration with:
Denise Becerra
Josh Dennis
Krish Viswanadhan Nair
Kanak Sharma
Vishnu Sai Nandan Tummala

TABLE OF CONTENTS

Abstract

In the face of rising global energy demands and escalating climate change, nations are increasingly challenged to balance economic growth with environmental sustainability. This research investigates the complex relationships among development profiles, energy access, renewable energy adoption, and sustainable outcomes using rigorous, data-driven approaches. It will critically examine the assumptions that renewable energy either drives growth or imposes excessive economic costs. Through the classification of countries based on development and energy characteristics, and by analyzing key economic indicators alongside environmental metrics, this study aims to uncover the factors that enable successful energy is correlated with GDP growth, the role of trade and investment in economic performance, and the influence of energy consumption patterns on CO2 emissions. The findings offer actionable insights for policymakers, international organizations, and energy stakeholders seeking resilient and equitable strategies for sustainable development in a rapidly evolving global context.

Introduction

In a time where the pursuit of economic growth increasingly intersects with the urgency of environmental preservation, understanding the relationship between development, energy access, and sustainability outcomes is essential. As the world's energy demands rise and climate change accelerates, nations are faced with pressure to transition towards renewable energy while sustaining economic advancement and yet, the ways to achieve both still remain unclear and uneven. Policymakers often face conflicting narratives about how renewable energy adoption inevitably drives growth, or it imposes unacceptable economic burdens. This research challenges such assumptions by applying rigorous data-driven methods to decipher these complex relationships.

The significance of this study lies in its potential to offer actionable insights for the governments, international organizations, and energy stakeholders. By illuminating critical factors that enable or hinder sustainable energy transitions without sacrificing economic goals, this study supports the design of smarter, more reliant development strategies in the face of global environmental crises. Beginning with the research objectives and questions, this report progresses through the data collection and database management, methodology, results, and discussion, then culminates with a conclusion that offers key insights, this study will address the critical problem of how nations can strategically balance economic growth and environmental sustainability, amid inconsistent development profiles, unequal energy access, and varying environmental outcomes, by using data driven analysis to uncover the key drivers of successful energy transitions.

Research Objectives & Questions

Research Objectives

1. To classify countries based on development profiles, energy access, and labor force characteristics.
2. To analyze the relationship between economic growth, renewable energy adoption, and environmental outcomes.
3. To evaluate how energy consumption trends and natural resource factors influence CO_2 emissions per capita and find what influences the GDP of countries.

Research Questions

- How does the level of GDP growth vary with renewable energy use across high, moderate, and low impact classifications?
- What development profiles emerge across countries based on energy access and the labor force?
- Can countries be classified based on features related to energy consumption and GDP?
- How do renewable energy usage, electricity access (rural vs. urban), and forest area influence CO_2 emissions per capita?
- How do foreign direct investment, exports, imports, and domestic savings (as a % of GDP) influence a country's GDP (in current US dollars)?
- How have individual countries transitioned toward renewable energy, and what are the patterns in emissions?

Methodology

Database & Data

The database was constructed using data sourced from World Bank Group, specifically from the World Development Indicators (WDI) dataset. This dataset provides a comprehensive collection of internationally comparable statistics about global development, making it a robust foundation for cross-country and temporal analysis. For the purposes of the project, data was extracted for all countries with available records spanning the period from 2021 to 2023. The time frame was intentionally selected to capture recent trends and shifts in development patterns, particularly in the aftermath of the COVID-19 pandemic and during a period of ongoing global economic and environmental transition.

The indicators included in the dataset were selected through a process designed to ensure alignment with the research objectives of the project. This began with an extensive review of the World Development Indicators (WDI) database, which offers a vast array of global development metrics compiled by the World Bank. The focus was to identify variables that could provide meaningful insight into both economic performance and environmental sustainability across countries. Initial filtering prioritized indicators with consistent, up-to-date coverage from 2021 to 2023, as this time frame reflects a critical post-pandemic period characterized by both recovery efforts and renewed attention to sustainable development. From there, the list was narrowed based on relevance, data quality, and comparability, selecting variables that not only reflected current global priorities but also aligned with the analytical goals. Once the relevant series were identified, the data was exported in CSV format and subsequently imported into a SQLite 3 database for structured storage and efficient querying. To facilitate clarity and organization, the database was segmented into two primary tables: Economic Indicators and Environmental Indicators.

Furthermore, an additional table was created encompassing all countries and years represented in the data. This table serves as the foundation for an associative entity that enables flexible and relational queries across the economic and environmental dimensions of the data. By establishing this schema, the team was able to efficiently join tables and explore complex relationships across multiple indicators, time periods, and geographic regions.

Analytical Methods

Data Preprocessing and Validation

The data collection process began with each team member individually exploring the data source and suggesting key series to include. Once the data was assembled in the database, any variables with more than 70% null values were automatically excluded, as imputation in such cases would likely yield unreliable results. Similarly, countries that lacked meaningful data across all project years were removed, resulting in a final dataset of 106 countries. Missing data at the country level was addressed using a combination of linear interpolation, forward fill, and backward fill to preserve temporal continuity without biasing trends.

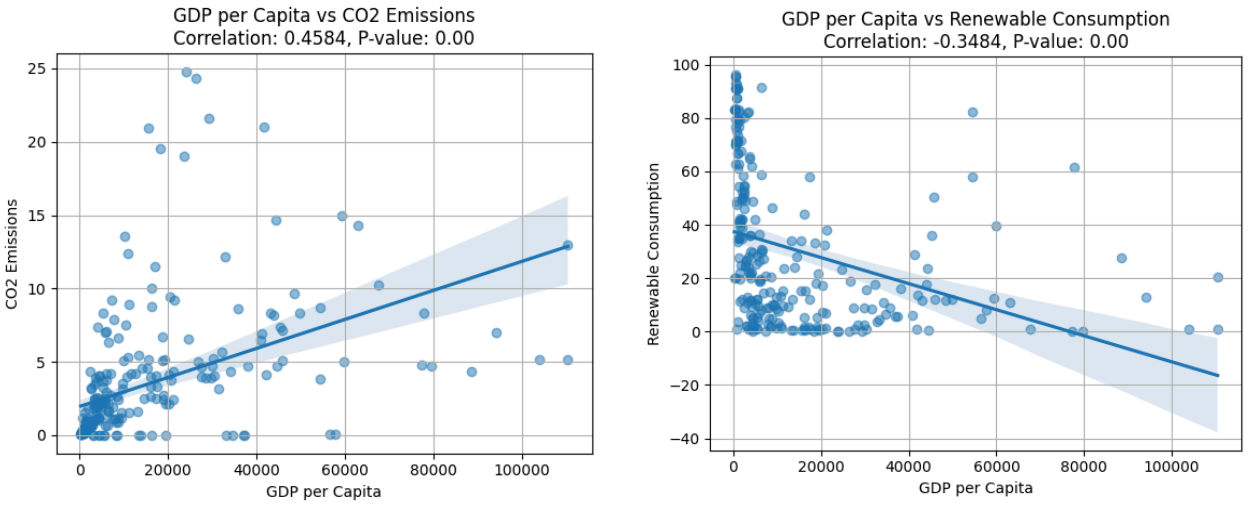
To prepare the data for analysis, multicollinearity was checked and a principal component analysis (PCA) was conducted for factor reduction. Depending on the requirements of the modeling approach, methods were either used imputation to address remaining null values or removed rows with missing data. The most frequently used imputation method was KNN imputation. For standardization, StandardScaler() from sklearn.preprocessing was applied across all numerical features to normalize scale. Lastly, an outlier analysis was conducted. This step often required human judgment to determine which countries had values that were unreasonably extreme and should therefore be excluded. These

variances were often attributed to low population inflating per capita values (U.S. Census Bureau, n.d.).

Methods & Tools for Data Analysis

Hypothesis Testing

Before proceeding with deeper analysis and development of predictive models, a hypothesis testing was conducted on the key variables identified during the data selection phase. This preliminary step was essential to verify the existence of statistically significant relationships between variables and to justify more advanced modeling approaches. Using standard statistical tests, pairwise correlations were evaluated to assess both the strength and direction of associations. The results confirmed that the selected variables exhibited strong statistical significance, with p-values at or near zero, indicating a very low probability that observed relationships occurred by chance. Moreover, correlation coefficients for these variables revealed moderate to strong positive or negative associations, providing a solid foundation for subsequent multivariate analysis and model construction.



Figures 1 & 2: LT- GDP per Capita vs CO_2 Emissions; RT- GDP vs Renewable Consumption

Classification Analysis

A multi-layered analytical framework was applied, progressing from preprocessing and feature engineering to unsupervised and supervised learning, emphasizing interpretability and robustness. For multicollinearity, the Variance Inflation Factor (VIF) was calculated for all predictors. Features with a VIF greater than 10 or infinite were dropped to reduce redundancy and enhance model generalization. Variables were retained based on their statistical independence and theoretical relevance to sustainability economics. When considering feature engineering, a new categorical variable named *Degradation_Label* was derived to capture the interaction between economic and environmental outcomes. Based on a logic that combines GDP growth (%) and CO_2 change (%), each country-year observation was assigned to one of four classes:

- High Degradation: High GDP growth and rising emissions
- Green Growth: High GDP growth and falling emissions
- Pollution Without Growth: Low/negative GDP growth and rising emissions
- Stable: Low GDP growth with flat or declining emissions

This target variable was later used for supervised classification tasks. When determining unsupervised clustering and dimensionality reduction and to group countries with similar sustainability profiles, K-Means Clustering (k=3) was conducted on three key standardized variables including renewable energy consumption, CO_2 emissions, and GDP growth. To test whether rapid economic growth correlates with environmental degradation, two ensemble classification models were trained on the Degradation_Label (High, Stable, Pollution w/o Growth, Green Growth)

- Random Forest Classifier: Leveraged for its interpretability and robustness. A GridSearchCV was performed with 5-fold Stratified K-Fold Cross-Validation to tune hyperparameters such as n_estimators, max_depth, and min_samples_leaf. Feature importance scores from this model helped identify the top predictors of environmental degradation.

- XGBoost Classifier: Employed as a high-performing gradient boosting model. It was tuned using GridSearchCV and stopped early to prevent overfitting. Feature importance scores from this model helped identify the top predictors of environmental degradation, and the confusion matrix helps evaluate the performance of a classification model by showing how well the predicted labels match the actual (true) labels.

Model performance was evaluated using accuracy, precision, recall, and F1-score. Accuracy was computed multiple times using accuracy_score() for different models on training and testing data for models like XGBoost and Random Forest. Precision, Recall, and F1-score were not explicitly computed with precision_score, recall_score, or f1_score functions but were likely included in classification_report(), which calculates all three. A confusion matrix was used to diagnose class-wise performance, showing high accuracy with almost perfect diagonal alignment. There was a minor misclassification between Class 1 and Class 2, involving just one sample. However, there was strong separation between all classes, indicating excellent model performance and good separation between the classes.

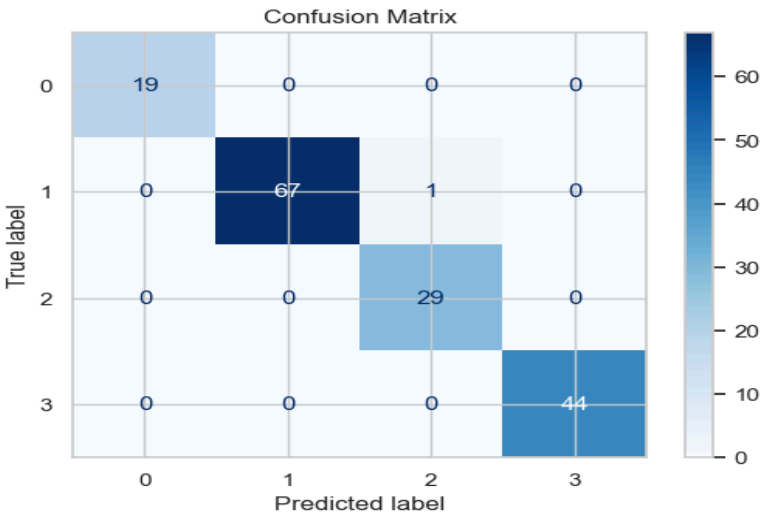


Figure 3: Confusion Matrix

Cross-validation accuracy was plotted across folds to confirm consistency. The code files include the use of cross_val_score() with cv=5, which confirms that the model was validated across multiple folds. This process ensures robustness and consistency in performance. It also helps detect overfitting by comparing training, testing, and fold scores.

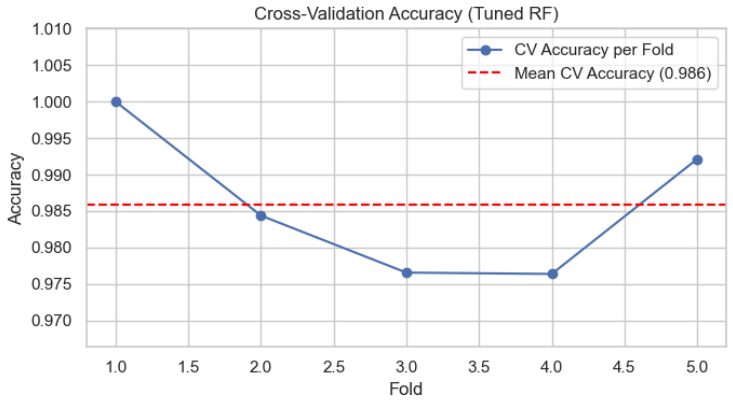


Figure 4: Cross Validation for Random Forest

The cross-validation results for the tuned Random Forest model demonstrate consistently high accuracy across all five folds, with a mean CV accuracy of 0.986. Although slight variability is observed across folds, ranging from just under 0.98 to slightly above 1.00, the model maintains strong performance throughout. The narrow fluctuation suggests robust generalization and stability across different subsets of the training data. Importantly, the absence of significant performance drops in any fold indicates that the model is not overly sensitive to data splits, further supporting its reliability for deployment on unseen data.

Learning Curves were used to assess training/validation convergence and overfitting: Code File includes use of learning_curve() and train sizes vary from 10% to 100% of the data. Training and validation accuracy are plotted. Interpretation provided:

- High training accuracy (1.0), indicating excellent fit on training data.
- Validation accuracy rises and stabilizes, indicating low overfitting.
- Narrow gap between curves suggests strong generalization.

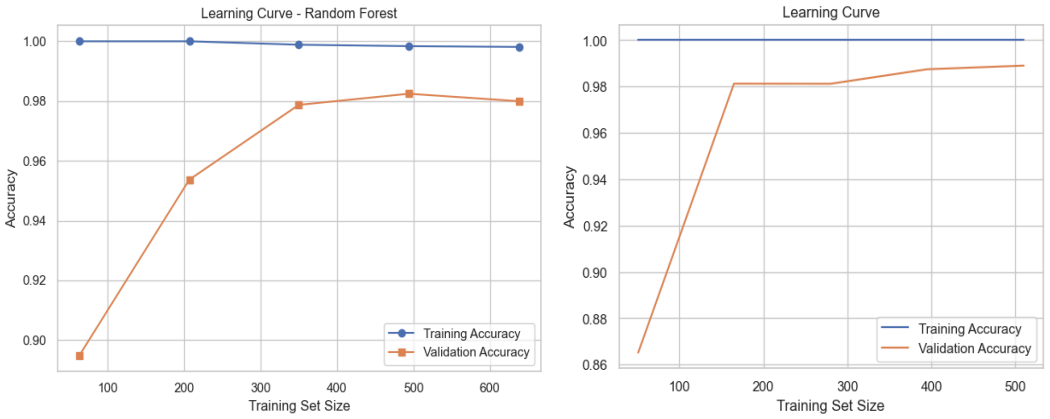
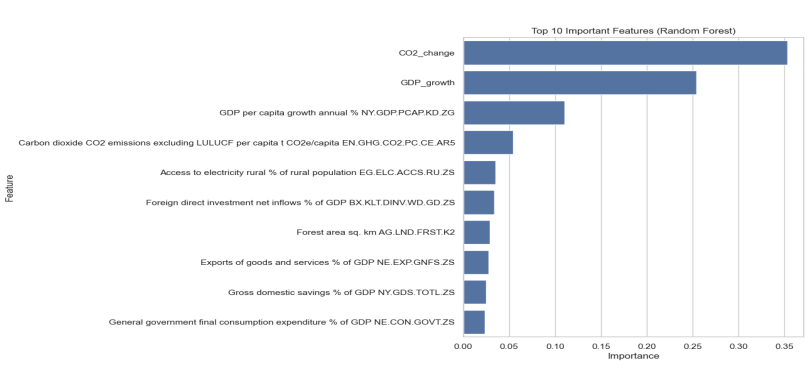


Figure 5 & 6: Learning Curve for Random Forest and XG Boost

The learning curves for both Random Forest and XGBoost models reveal high overall performance with limited overfitting. For the Random Forest model, training accuracy remained consistently close to 1.0, while validation accuracy improved steadily and stabilized around 0.98. The small and narrow gap between training and validation curves suggests good generalization and minimal overfitting. In the case of XGBoost, although the training accuracy is perfect across all sample sizes, indicating potential overfitting at smaller training sizes, the validation accuracy quickly rises and stabilizes above 0.98 as more data is introduced. This convergence between curves signifies that XGBoost also generalizes well when trained on sufficient data. Overall, both models demonstrate strong predictive performance and robustness across varying training sizes. For model interpretability and feature importance, both Random Forest and XGBoost models consistently highlighted CO_2 emissions, GDP growth as dominant features influencing the degradation label.

Figure 7: Feature Importance



The Random Forest model highlighted several key features influencing environmental degradation predictions. CO_2 change emerged as the most influential predictor, followed closely by GDP growth and GDP per capita growth. These results reinforce the strong link between economic expansion and environmental outcomes. Additional top features included carbon dioxide emissions per capita, access to electricity in rural areas, and foreign direct investment, indicating that both environmental pressure and development metrics play significant roles. Notably, forest area also appeared among the top contributors, suggesting the relevance of natural resource coverage in the model's predictions. Overall, the feature importance distribution confirms that a combination of economic indicators and emission-related factors predominantly drives the model's classification of degradation levels. These consistently appeared as the top degradation label predictors, confirming that economic activity and environmental indicators play key roles.

Clustering: Hierarchical & K-Means Analysis

To classify countries based on various indicators of world development, energy access, labor force structure, and environmental impact, both hierarchical clustering and k-means clustering were employed in this study. These methods use the exploration of underlying patterns on unseen data through unsupervised machine learning techniques and make it possible to derive meaningful insights on country grouping. A correlation heatmap was visualized to identify dimensionality of manually selected variables.

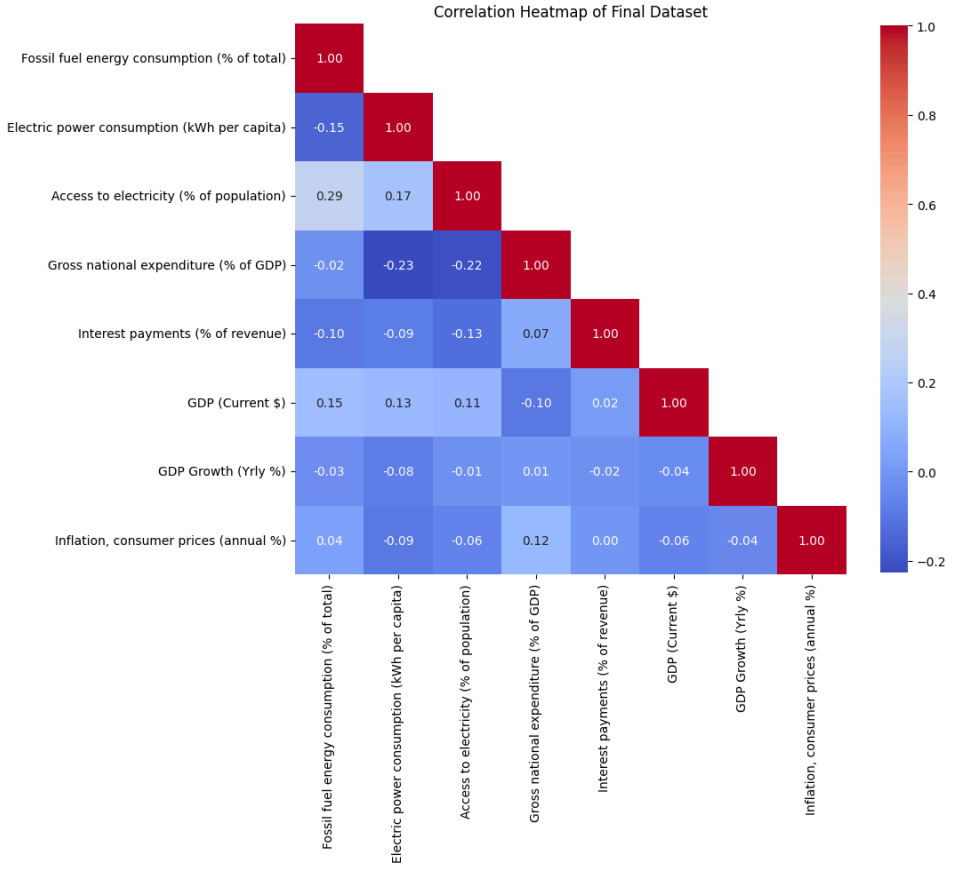


Figure 8: Correlation between features in Hierarchical Clustering Analysis

Hierarchical clustering utilizes a tree-like structure, also known as a dendrogram, that nests relationships among countries based on their feature similarities. This unsupervised machine learning technique allows the identification of natural divisions within the data without requiring a predetermined number of clusters. In this study, a hierarchical analysis was implemented to answer how countries be classified based on features related to energy consumption and GDP, and begins by treating each country as its own cluster and eventually merging the closest pairs based on a chosen distance metric, Euclidean Distance alongside Ward's linkage method. This dendrogram helps visualize the optimal number of clusters by examining the linkage distances where meaningful splits occur. In the dendrogram below five distinct clusters were formed among countries based on factors like fossil fuel consumption, electric power consumption, access to electricity, GDP, etc.

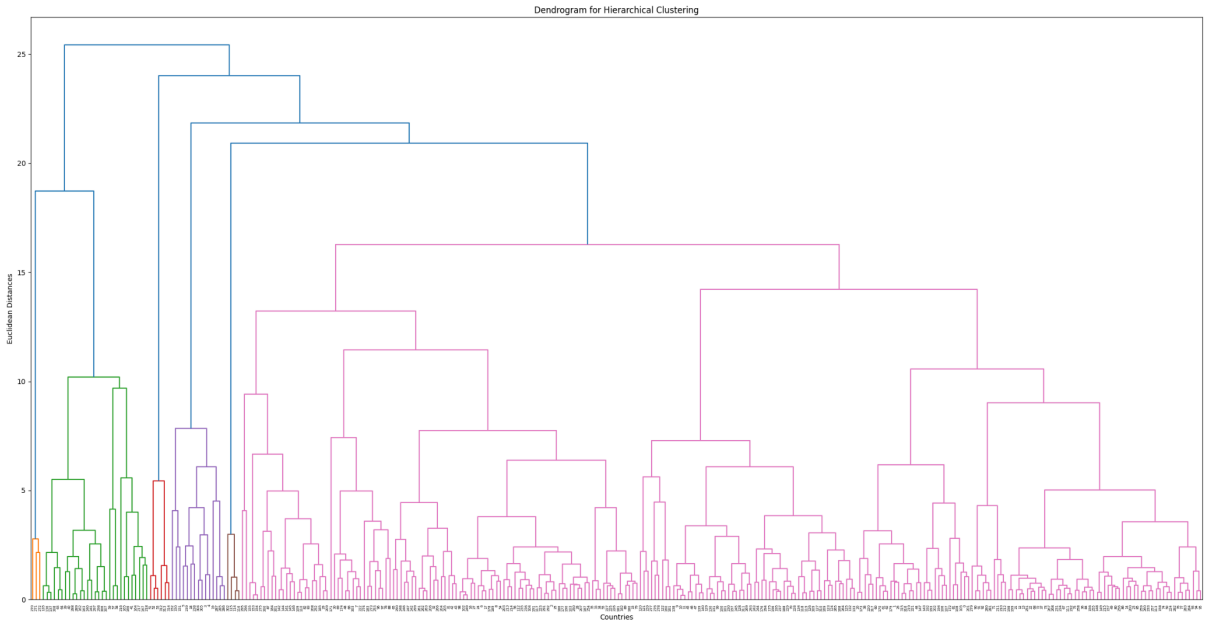


Figure 9: Dendrogram for Optimal Clusters

To conceptualize the dendrogram into a more digestible manner, a bar chart visualization of the count of countries in each cluster was visualized. Here it shows that 7 clusters were formed with cluster one and cluster three having the most countries grouped together based on their similarities on energy consumption, access to electricity and so on.

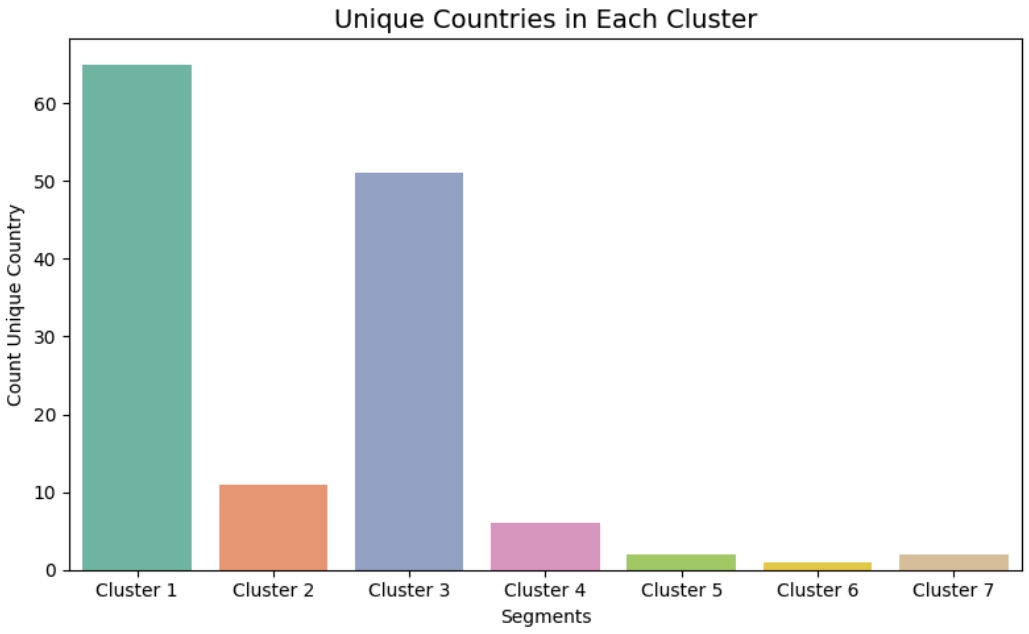


Figure 10: Barchart Unique Count of Countries by Cluster

The second predictive model, K-means clustering, was implemented to answer the research question regarding what development profiles emerge across countries based on energy access and the labor force. K-means cluster analysis is an unsupervised algorithm in unsupervised machine learning. This algorithm tries to group instances, or observations, into segments in a way that each observation is compared to another common observation known as a centroid. This then minimizes the sum of squares within the cluster and occurs in iterative steps for a predefined number of clusters.

To identify the ideal number of clusters (k), an elbow plot was utilized to assess the points at which adding an additional cluster yields diminishing improvements in model performance. This method helps strike an effective balance between model complexity and interpretability, minimizing the risk of underfitting or overfitting the data. Based on the elbow plot, two distinct clusters were identified as the most appropriate segmentation for this data.

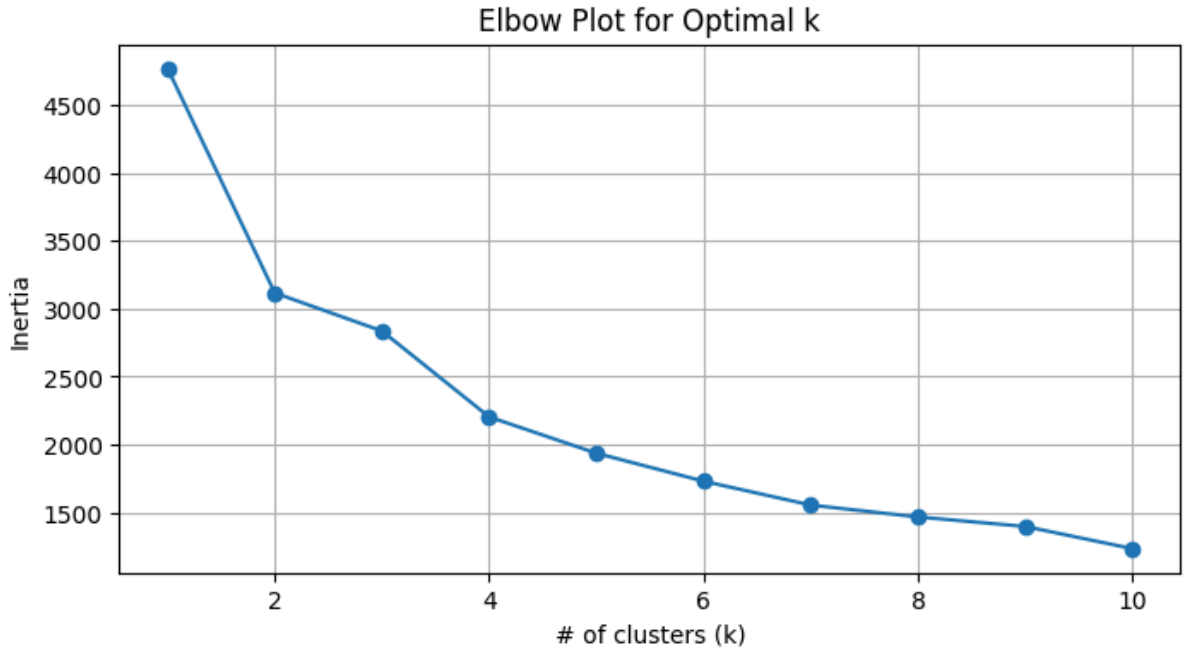


Figure 11: Elbow Plot for Optimal K

Multiple Linear Regression

The multiple linear regression model explored how renewable energy consumption, electricity access, and forest area impacts CO_2 emissions per capita. The independent variables include renewable energy consumption as a percentage of total energy use, access to electricity in both rural and urban areas, and total forest area in square kilometers; the dependent variable is CO_2 emissions per capita. The model was trained on a cleaned dataset extracted from the environmental indicators table. To ensure consistency in scale across features, standard scaling was applied. An 80/20 train-test split was used to evaluate model performance and generalizability.

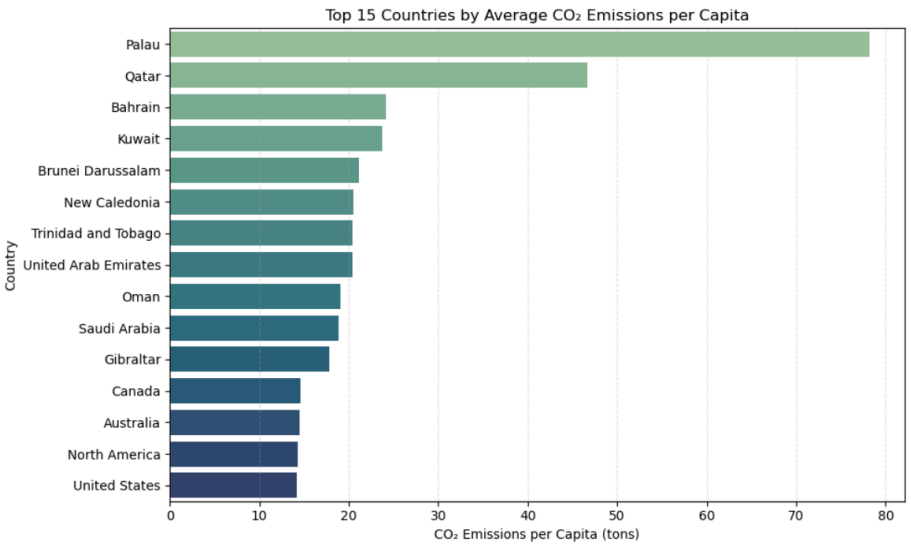


Figure 12: Shows the top 15 countries with average CO2 Emissions per capita

Additionally another analysis was performed that examines how foreign direct investment (FDI), exports, imports, and gross domestic savings, as a percentage of GDP, influence a country's overall GDP in current U.S. dollars. The data was sourced from the economic indicators table within a structured SQLite database. A total of complete and clean records were selected using SQL queries that filtered out any rows with missing values in key columns. These columns include FRDI, exports, imports, gross domestic savings, and GDP.

All the data was converted to float format to prepare for numerical modeling. A multiple linear regression (MLR) model was then developed, with GDP (in current USD) as the target variable and FDI, exports, imports, and savings as the predictor variables.

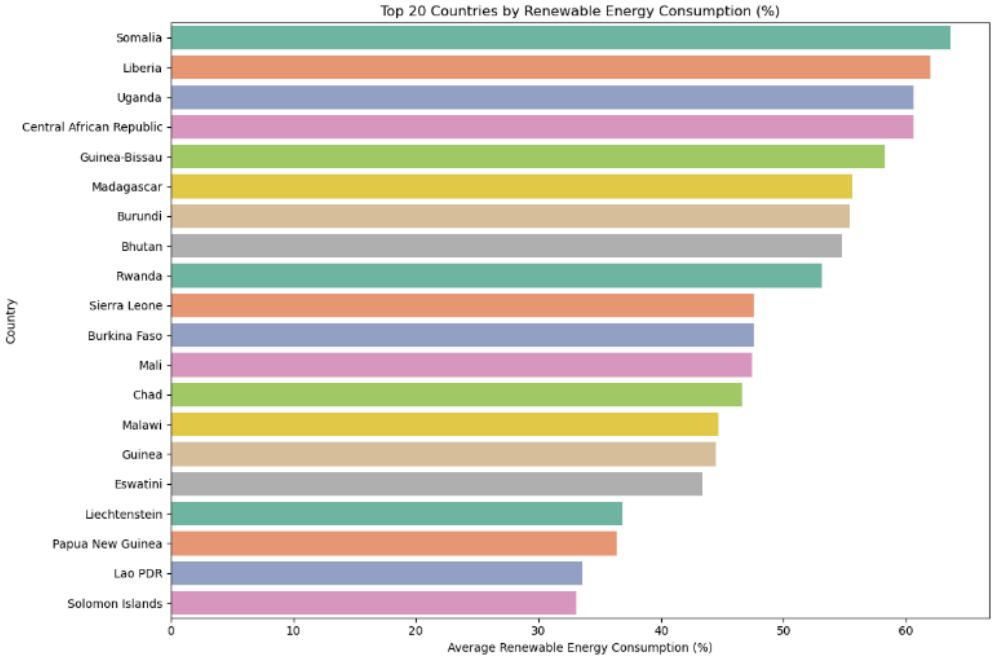


Figure 13: Top 20 countries by renewable energy consumption(%)

Time Series Analysis

The time series analysis focused on understanding the trends and interrelationships between renewable energy consumption and CO_2 emissions from 2021 to 2023 across selected countries. The countries analyzed included China, India, Brazil, Japan, Russia, South Africa, Germany, the United States, and the United Kingdom. The mean renewable energy consumption and CO_2 emissions for each country over the time period were calculated and normalized to enable direct comparison. Deriving insights from this time series is important to the study because it will reveal how renewable energy consumption and CO_2 emissions evolve over time, helping identify long-term trends, policy impacts, and shifts in energy

usage. This understanding is essential for assessing the effectiveness of sustainability initiatives and guiding future environmental and economic strategies.

Results

This study analyzed how economic growth and renewable energy adoption relate to environmental outcomes across countries from 2021 to 2023. The results are presented through cluster analysis, classification modeling, and feature interpretation. For classification, countries were grouped into High, Moderate, and Low Impact clusters using K-Means on standardized indicators such as GDP growth, CO_2 emissions, and renewable energy use. The resulting clusters were labeled:

- High Impact: High economic growth at the cost of environmental sustainability.
- Moderate Impact: Balanced adoption of renewables with stable growth.
- Low Impact: lagging on renewable adoption and emission mitigation.

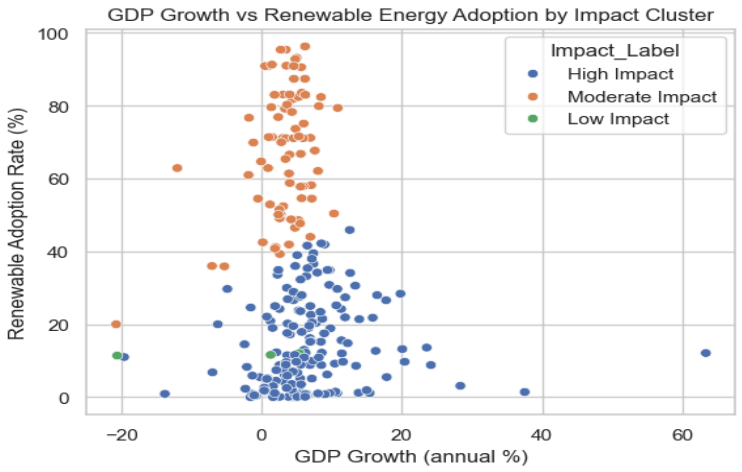


Figure 14: GDP Growth vs Renewable Energy Adoption by Impact Cluster

Degradation Classification: Using GDP and CO_2 change, a rule-based degradation label was created- Green Growth, High Degradation, Pollution Without Growth, and Stable. This label became the target variable for supervised learning models.

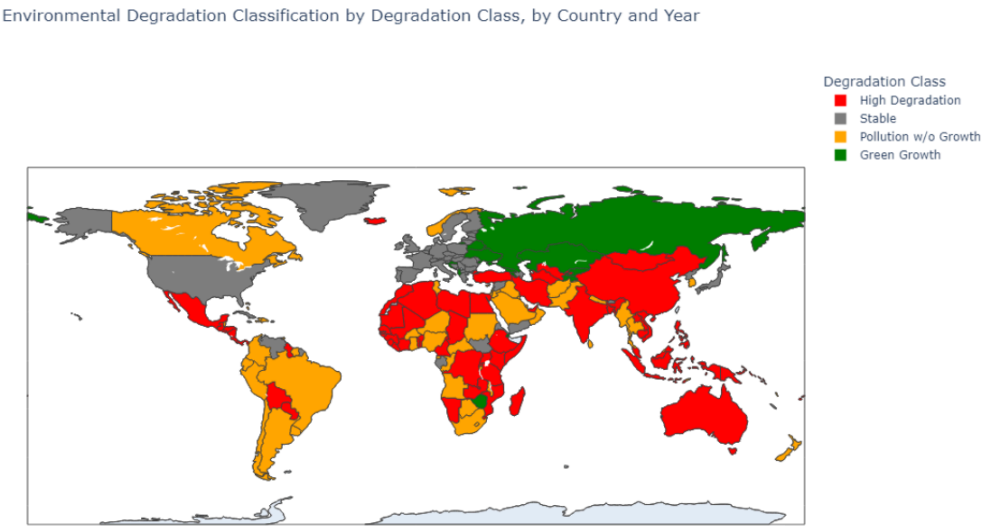


Figure 15: Environmental Degradation Classification by Degradation Class, Country, and Year

Supervised Model Results: Two ensemble models—Random Forest and XGBoost—were trained to predict the degradation label. For accuracy, both models achieved over 98% and were validated using 5-fold cross-validation.

Classification Report:

	precision	recall	f1-score	support
Green Growth	1.00	1.00	1.00	19
High Degradation	1.00	0.99	0.99	68
Pollution w/o Growth	0.97	1.00	0.98	29
Stable	1.00	1.00	1.00	44
accuracy			0.99	160
macro avg	0.99	1.00	0.99	160
weighted avg	0.99	0.99	0.99	160

Figure 16: Classification Report for XGBoost

=== Classification Report ===

	precision	recall	f1-score	support
Green Growth	1.00	0.95	0.97	19
High Degradation	1.00	0.99	0.99	68
Pollution w/o Growth	1.00	1.00	1.00	29
Stable	0.96	1.00	0.98	44
accuracy			0.99	160
macro avg	0.99	0.98	0.99	160
weighted avg	0.99	0.99	0.99	160

Figure 17: Classification Report for Random Forest

Confusion Matrix and Learning curves (as shown in the report above) were also used, and they showed high training and validation accuracy with minimal overfitting. For Feature Importance, top predictors across models included- CO_2 change, GDP growth, forest area, and access to rural electricity. These features confirm the dual influence of economic and environmental factors.

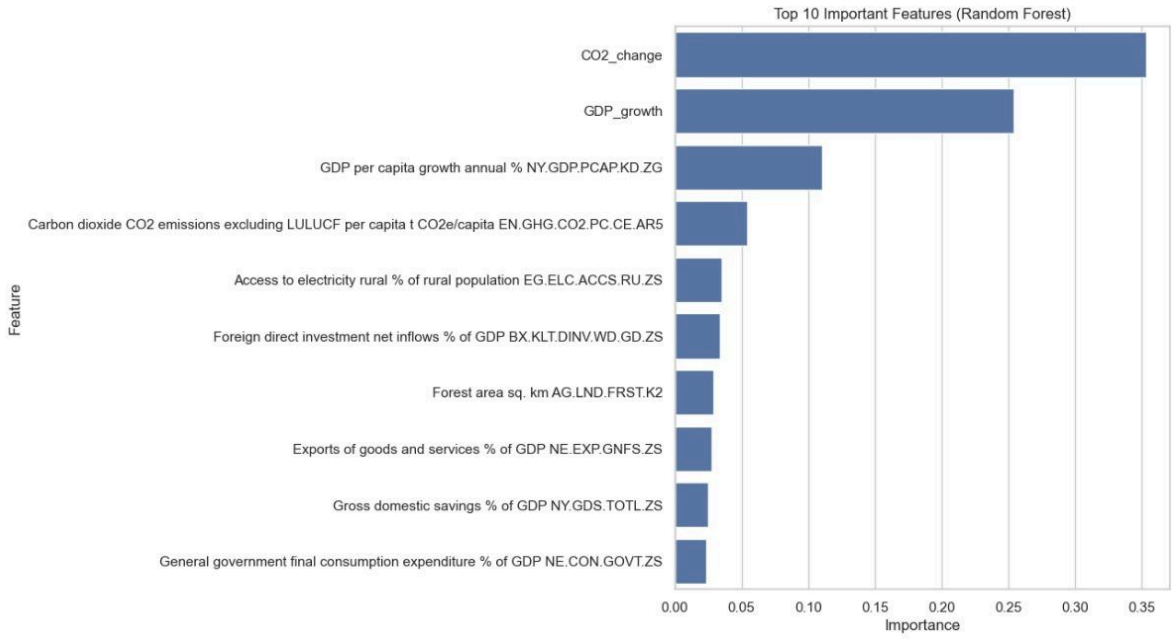


Figure 18: Feature Importance.

These results confirm that economic development indicators and environmental pressure metrics are strong predictors of a country's sustainability trajectory.

The multiple linear regression model revealed varying impacts of the predictor variables on CO_2 emissions per capita. Our findings show that renewable energy usage and rural electricity access reduce CO_2 emissions, while urban access and forest area slightly increase it, overall indicating variations in CO_2 emissions per capita, and a need for additional features to improve accuracy as we can see in the graph that Outliers like China and the US

strongly affect model accuracy with R^2 of 0.3696. These results suggest that while some variables have directional influence, other unaccounted factors likely play a much larger role in determining CO_2 emissions.

When a secondary analysis was completed, the model estimated that foreign direct investment and exports have negative impacts on GDP, while imports and savings show positive relationships, with exports and savings having the strongest effects. However, the scatter plot revealed substantial variance, indicating that predictions often deviate from the ideal trend. With an R^2 score of 0.3822, the model suggests that the four economic indicators explained the variation in GDP, highlighting that additional variables are needed for more accurate predictions.

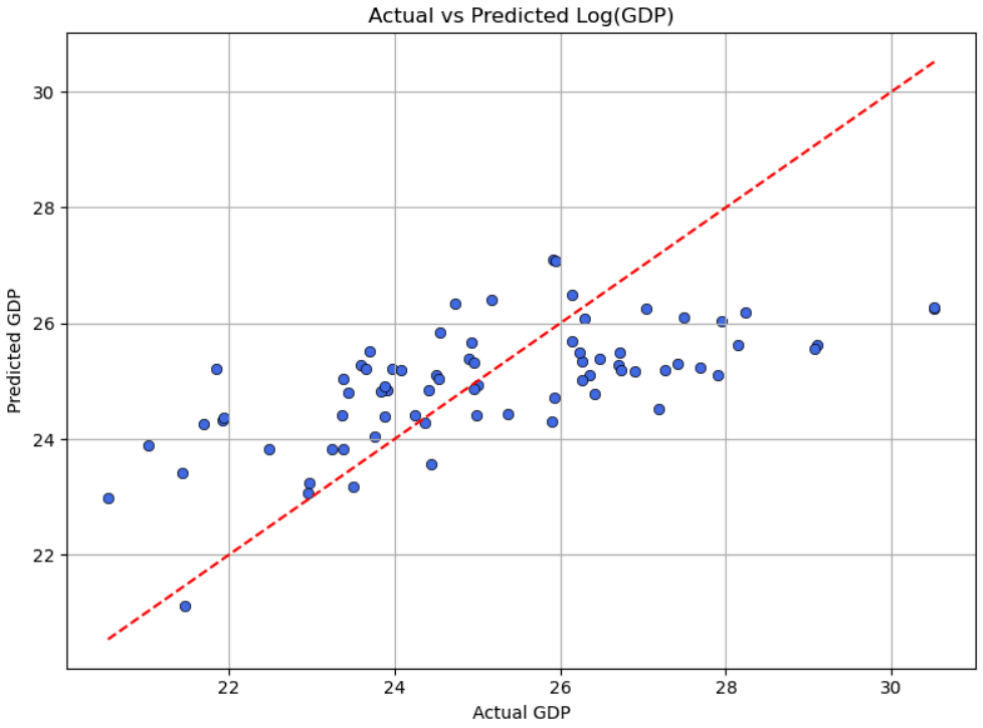


Figure 19: Predicted vs Actual GDP

The low-medium R^2 score suggests that many other factors (e.g., population, innovation, infrastructure, education, political stability) may play more significant roles.

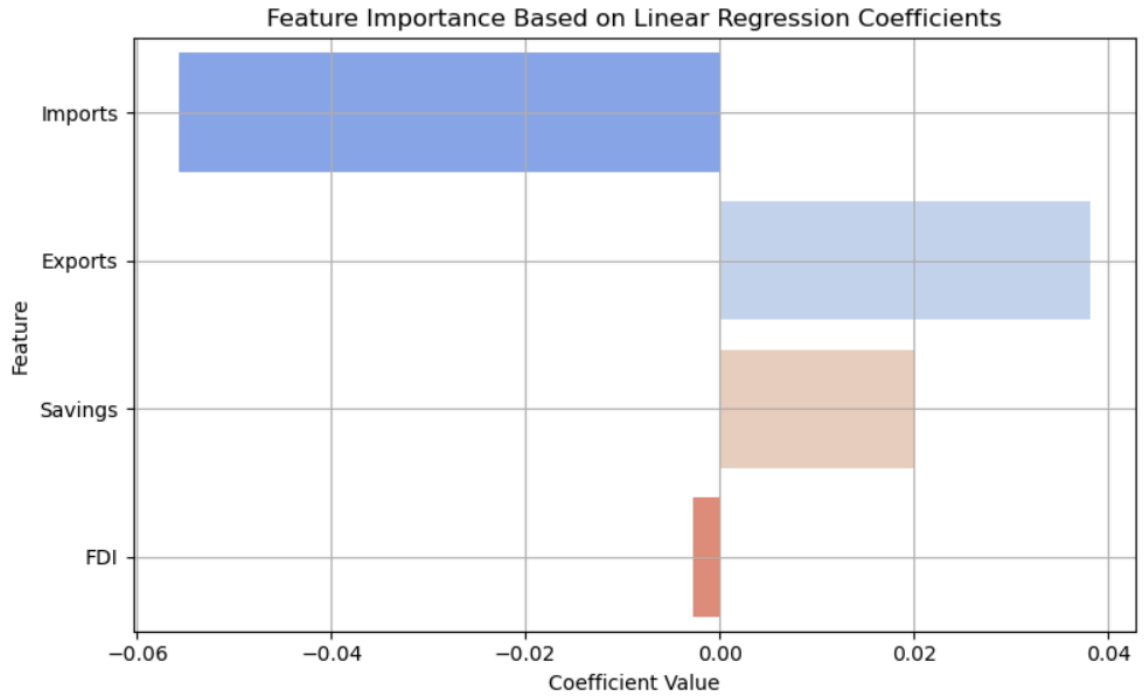


Figure 20: Shows feature importance based on Linear Regression coefficients chosen

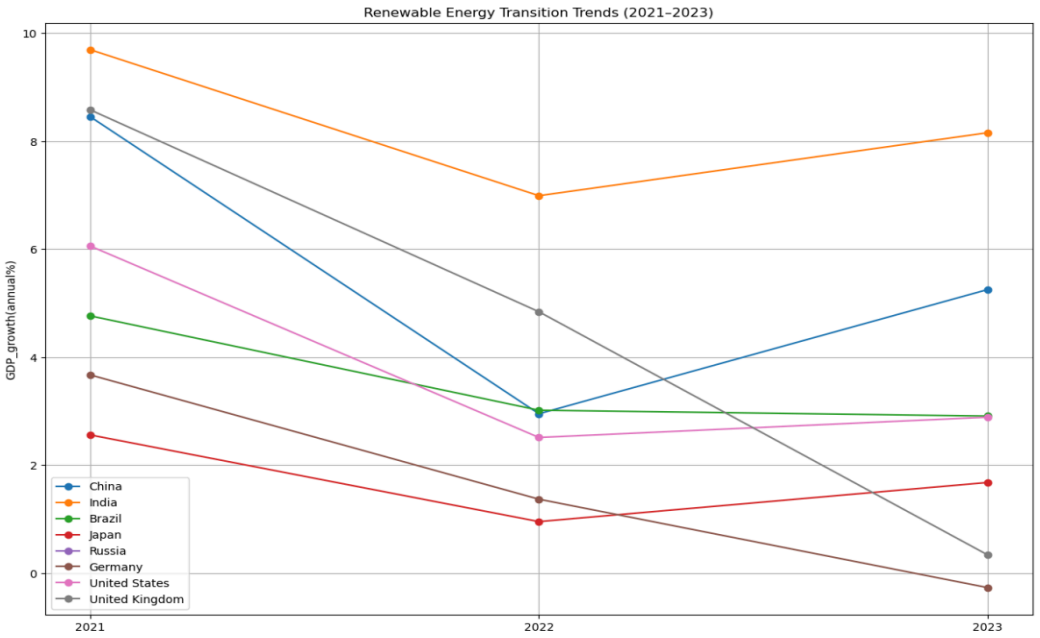
From the figure above we can see that imports have the strongest (negative) impact on log-GDP in the model. Exports and Savings are positively associated with GDP. FDI seems less significant in this dataset or model context.The MLR model serves as a basic exploratory step but should be expanded with more variables and potentially nonlinear methods for better performance.

In time series analysis, a side-by-side bar chart was generated to visualize these comparisons. The chart indicates that countries such as Brazil and Germany demonstrate relatively higher normalized renewable energy usage and lower emissions, whereas China and the United States showed high values in both categories, indicating continued reliance on fossil fuels despite renewable energy efforts. To analyze trends over time, a linear regression model was applied to the CO_2 emissions data for selected countries. In the case of the United

States, the regression slope suggested a slightly downward trend in CO_2 emissions from 2021-2023, potentially reflecting the impact of environmental policies or changes in energy sourcing. Additionally, a pivot-based time series transformation was applied to renewable energy consumption data. This enabled the application of clustering techniques like KMeans to group countries based on similar patterns in their renewable energy adoption. The clustering showed that countries could be broadly grouped into low, medium, and high renewable adoption trajectories.

To analyze how countries are transitioning toward renewable energy, a line plot was created to display GDP growth for selected countries from 2021 to 2023. Countries included in this comparison were China, India, Brazil, Japan, Russia, Germany, the United States, and the United Kingdom. The graph reveals that GDP growth rates varied across countries, but many showed steady or rising trends during this time, reflecting potential economic resilience alongside energy transitions. Notably, countries like India and China maintained positive GDP growth, which could indicate the integration of renewable energy without sacrificing economic performance. On the other hand, some developed nations exhibited more modest or volatile changes.

Figure 21: Renewable Energy Transition Trends over 3 years



To uncover patterns among countries, K-Means clustering was applied to normalized renewable energy consumption data from 2021 to 2023. Countries were grouped into three clusters based on their energy transition patterns:

- *Fast Adopters*: Countries with high and growing renewable energy shares.

- *Moderate Adopters*: Countries with stable or modest growth in renewable energy.

- *Slow Adopters*: Countries with consistently low renewable energy usage.

A scatter plot was generated comparing renewable energy values from 2021 vs. 2023, color-coded by cluster. The plot clearly shows separation between the three clusters, offering insight into global disparities in renewable energy adoption. For instance, Fast Adopters may include European countries and Brazil, whereas large emitters like China or India may appear in the Moderate or Slow categories depending on transition pace.

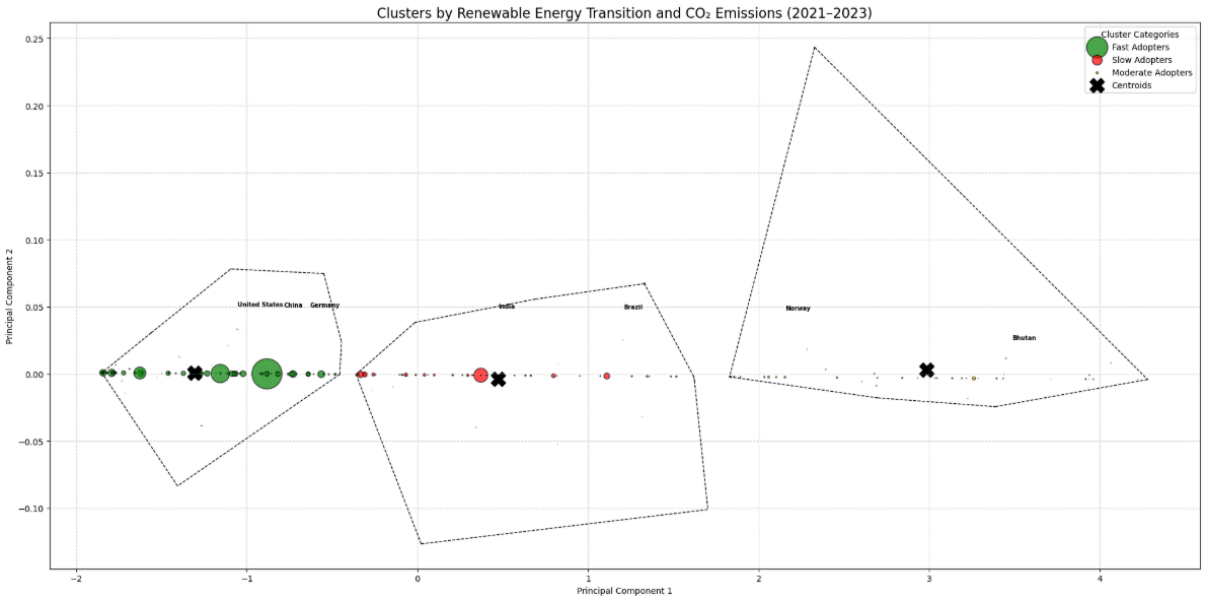


Figure 22: Clustering on Renewable Energy Transition and CO_2 Emissions over time

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
Albania	Burkina Faso	Albania	Argentina	Iceland	Sri Lanka	China
Australia	Congo, Dem. Rep.	Armenia	Barbados	Tonga		United States
Austria	Cote d'Ivoire	Bangladesh	Lebanon			
Barbados	Kenya	Bolivia	Russian Federation			
Belgium	Madagascar	Bosnia and Herzegovina	Tajikistan			

Figure 23: Table of Top 5 Countries in each Hierarchical Cluster

Seven clusters were formed in the hierarchical clustering model, with clusters one and cluster three having the most countries with the likes of Austria and Belgium in the first and Armenia and Bolivia in the latter. China and the United States of America were grouped together due to their similarities in economic growth, electric power consumption, and fossil fuel consumption.

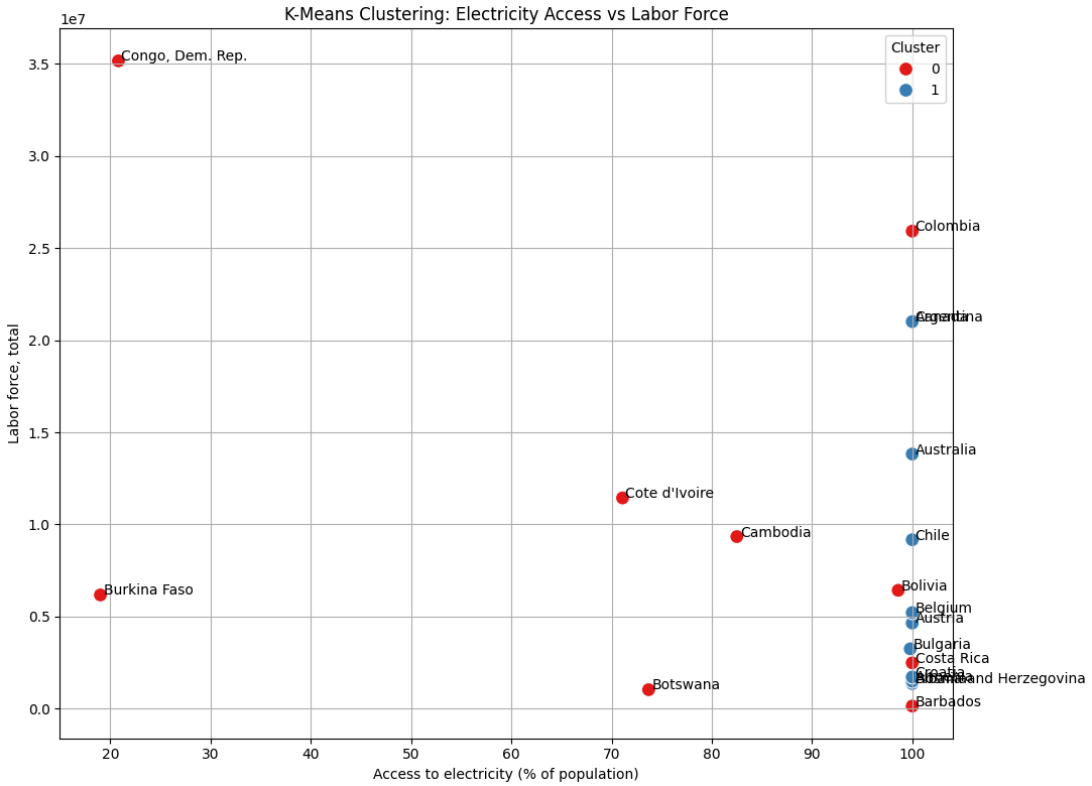


Figure 24: K-Means Clusters as a Scatter Plot

In the K-means Cluster Analysis, Austria and Belgium were also grouped together as in the hierarchical analysis dispute different feature selection playing roles in the analysis. This analysis differs from the previous, in that only two clusters were formed. This is due to the limitations of the algorithm, if more clusters were added, it would hinder the model and reduce how well the clusters are formed, challenging the integrity of the predictive model on unseen data. When examining labor force as a percent of population and access to electricity, the Republic of Congo was a finding of key interest. Although grouped on the first cluster, the Democratic Republic of Congo has a much higher labor force with lower access to electricity, opposite to other countries within the same segment, most likely due to its lack of infrastructure and the sociopolitical issues the country has. qualifying it as an outlier. Other outliers such as Bangladesh, Indonesia, Brazil, and Nigeria were also excluded to minimize the skewness of the dataset.

Discussion

Overview of Models Used

Multiple Linear Regression (MLR) for GDP and CO_2 emissions. Classification models (e.g., logistic regression or decision trees) to predict sustainability outcomes.Clustering (e.g., K-Means) to group countries by energy and economic profiles. Clustering results depended heavily on the chosen scaling and normalization method. Examining global energy consumption over the past 3 years using time series showed gradual shifts toward renewable sources, but the pace varied by region.

The objective was to assess if countries switching to renewable resources impacted their economies. Countries are also classified and grouped for targeted insights. The models confirmed in providing us with directional insights such as renewable energy generally reduces CO2 emissions. The results were consistent with global findings, renewable energy is

shown to reduce emissions. Outliers like China and the United States had high carbon emissions and high GDP. If focused on renewable energy resources, the CO2 emissions can be reduced and a sustainable energy consumption can be achieved. Our limitations included finding consistent data in the same period globally. To confirm that all countries that opted for renewable energy resources were not negatively affected we require more data of the global trends.

Conclusion

The analysis revealed a clear tension between rapid economic growth and sustainable energy use. For the first Objective, we examined the relationship between economic growth and renewable energy adoption. Encouragingly, we found that sustainable growth is achievable. Countries like Germany and Brazil are leading examples. They show that it is possible to grow economically while actively integrating renewable energy into national policy. In the second Objective, our analysis highlighted China and the United States as major outliers. What sets them apart are consistently high levels of GDP, electric power consumption, fossil fuel dependency, and correspondingly high CO_2 emissions. These factors make them unique in terms of their global environmental impact. Objective three revealed that while traditional economic indicators like exports and domestic savings did influence GDP, we also discovered that urban electricity access plays a critical role in shaping CO_2 emissions. This finding emphasizes how infrastructure and access contribute to environmental outcomes, especially when fossil fuels are involved.

Finally, across all three objectives, our results emphasize the urgent need for balanced development strategies. It's not enough to focus on economic output, we need policies that align economic ambition with environmental responsibility.

References

Awan, A. A. (2024, February 9). *SQL NOT IN OPERATOR: A comprehensive guide for beginners*. DataCamp. https://www.datacamp.com/tutorial/sql-not-in-operator

Bureau, U. C. (2024, October 15). *International database*. Census.gov. https://www.census.gov/programs-surveys/international-programs/about/idb.html

Flowchart. SmartDraw. (n.d.). https://www.smartdraw.com/flowchart/

Get started with shiny. Shiny. (n.d.). https://shiny.posit.co/getstarted.html

Learn. scikit. (n.d.). https://scikit-learn.org/stable/

O'Sullivan, C. (2024, April 1). *Feature selection with hierarchical clustering for interpretable models*. Medium. https://medium.com/data-science/feature-selection-with-hierarchical-clustering-for-interpretable-models-a091802f24e0

Ryzhkov, E. (2020, July 23). *5 stages of data preprocessing for K-means clustering*. Medium. https://medium.com/@evgen.ryzhkov/5-stages-of-data-preprocessing-for-k-means-clustering-b755426f9932

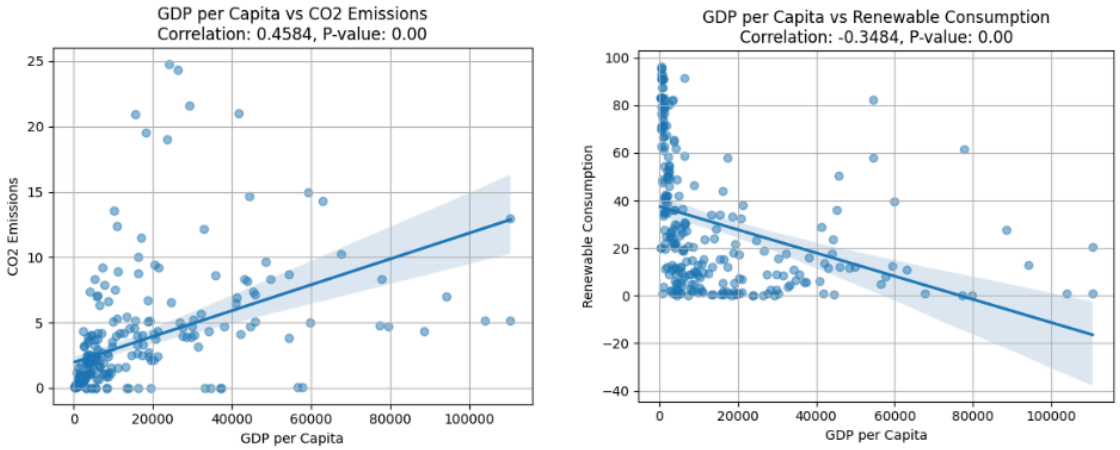
SQL Group by: Intermediate SQL - mode. Mode Resources. (2016, May 23). https://mode.com/sql-tutorial/sql-group-by

Understanding excel macros codes. Tutorialspoint. (n.d.). https://www.tutorialspoint.com/excel_macros/excel_macros_understanding_codes.htm

W3schools.com. W3Schools Online Web Tutorials. (n.d.-a). https://www.w3schools.com/python/

What is a flowchart? process flow diagrams & maps | ASQ. (n.d.). https://asq.org/quality-resources/flowchart

Appendix



Figures 1 & 2: LT- GDP per Capita vs CO_2 Emissions; RT- GDP vs Renewable Consumption

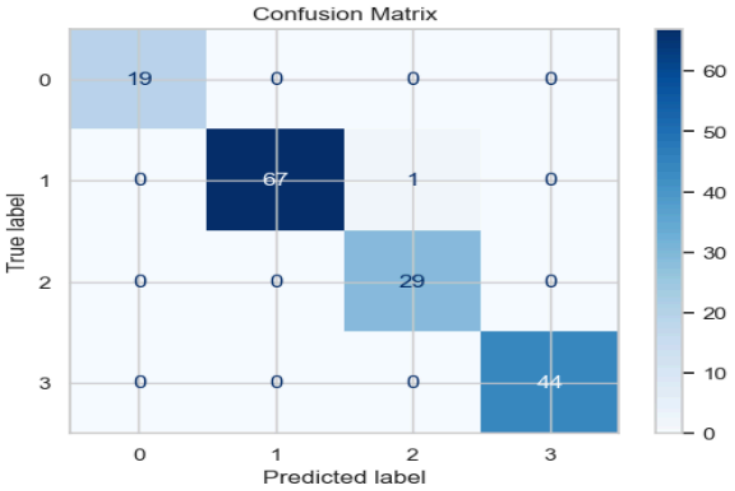


Figure 3: Confusion Matrix

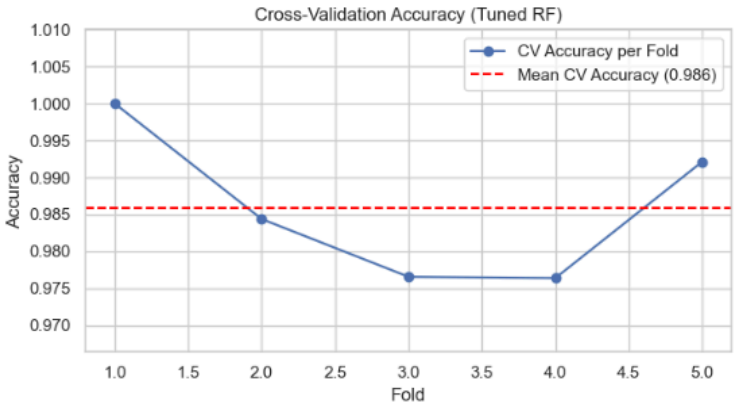


Figure 4: Cross Validation for Random Forest

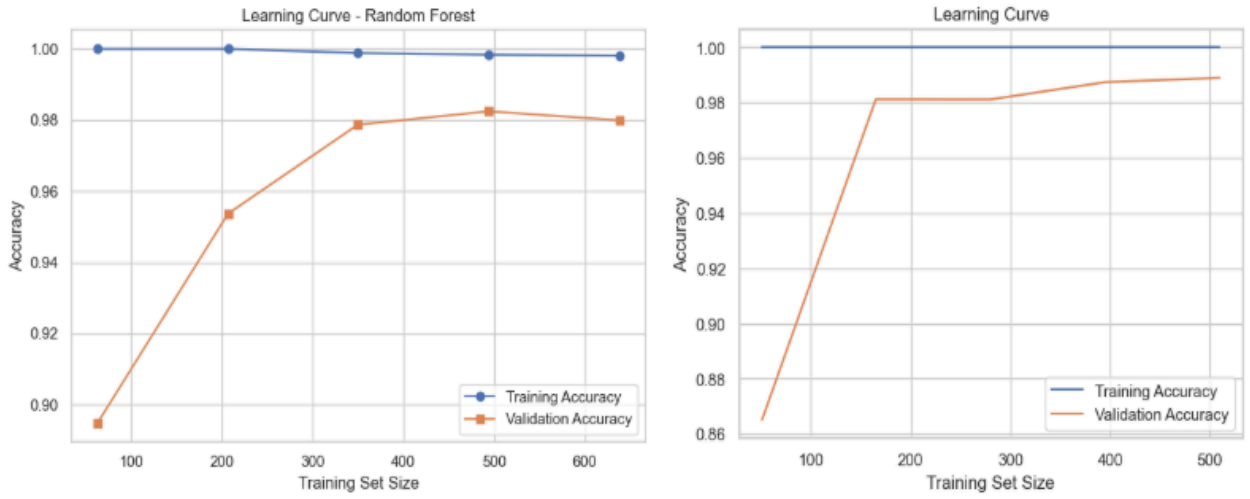
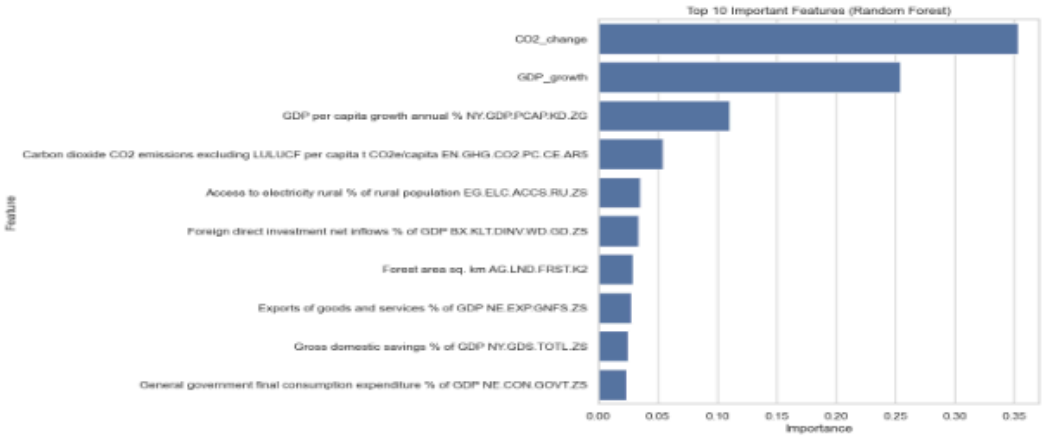


Figure 5 & 6: Learning Curve for Random Forest and XG Boost

Figure 7: Feature Importance



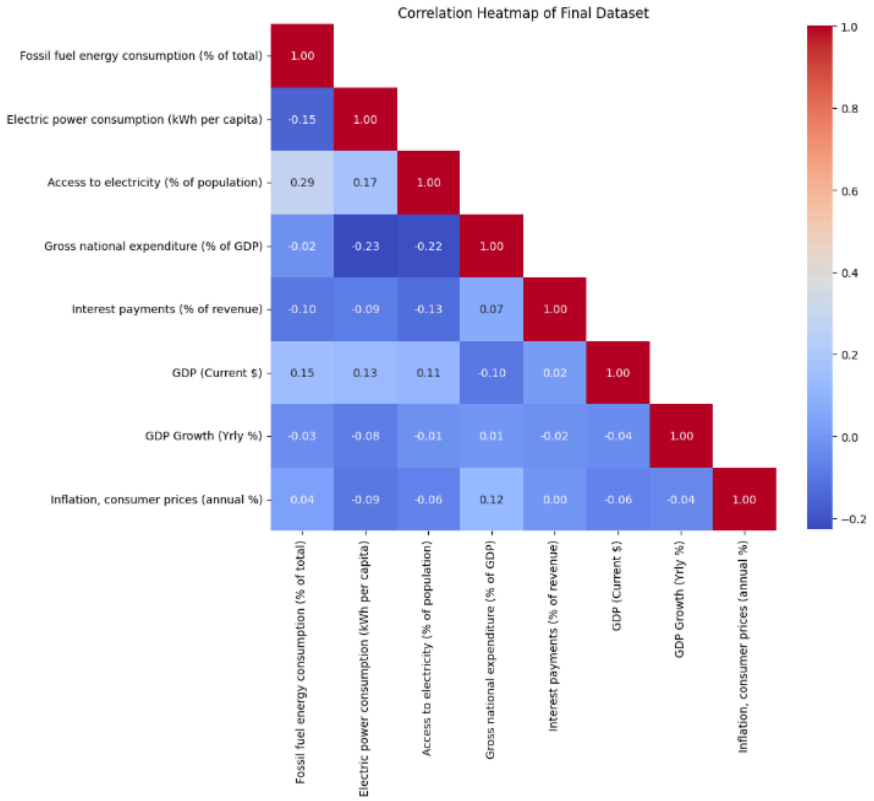


Figure 8: Correlation between features in Hierarchical Clustering Analysis

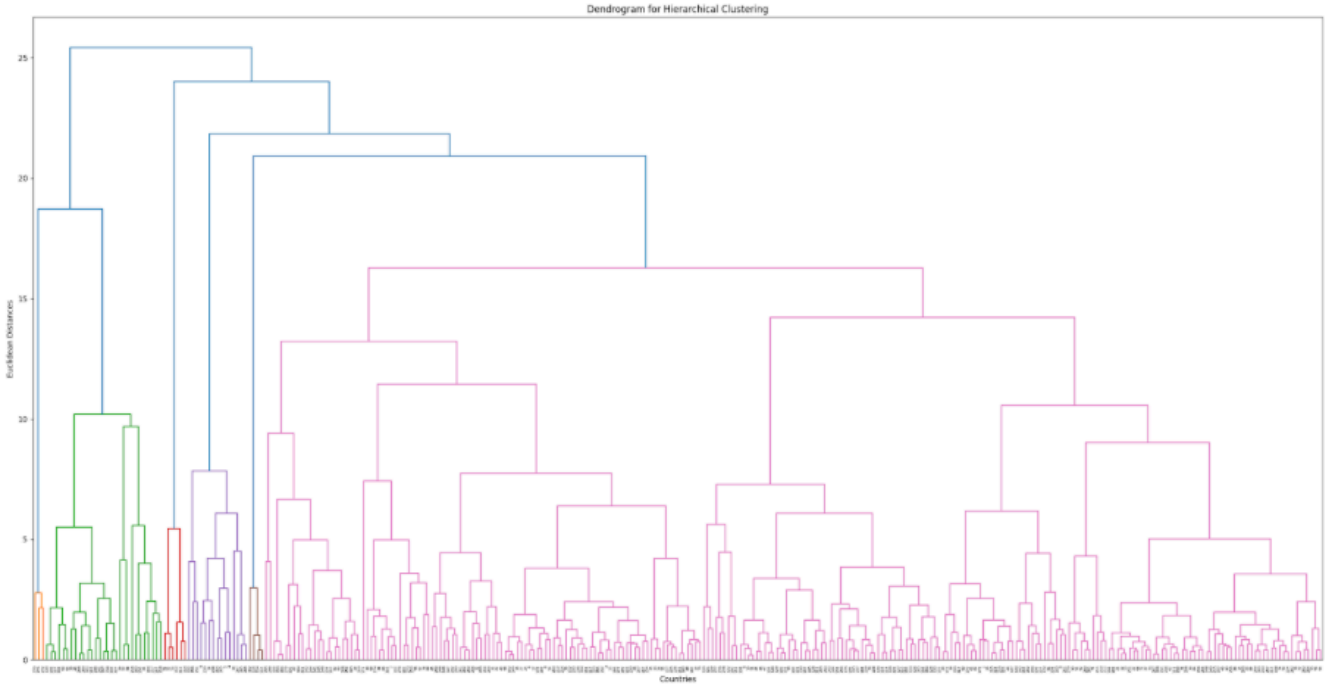


Figure 9: Dendrogram for Optimal Clusters

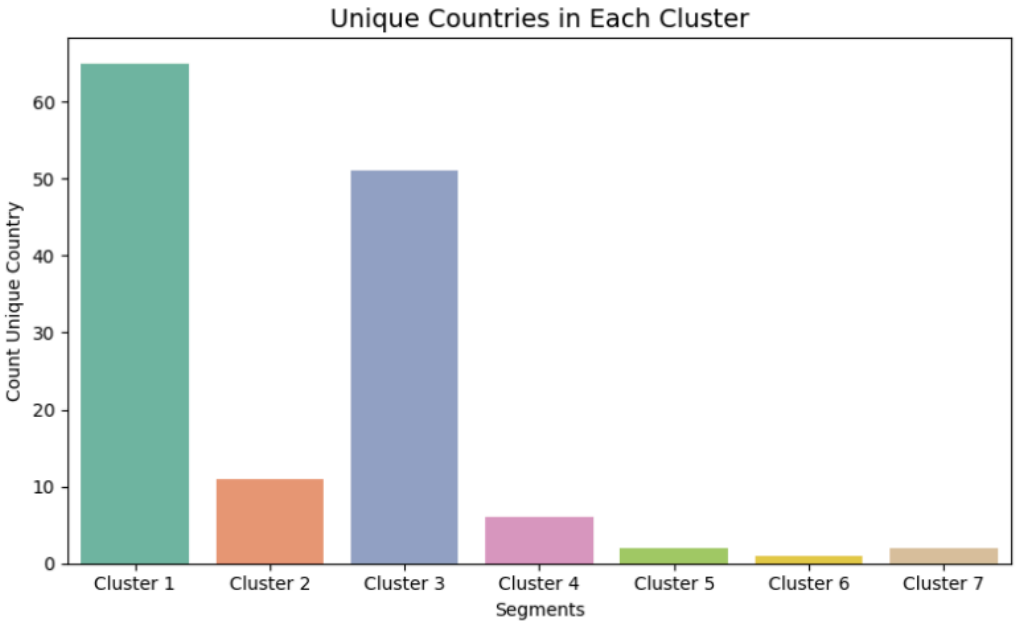


Figure 10: Barchart Unique Count of Countries by Cluster

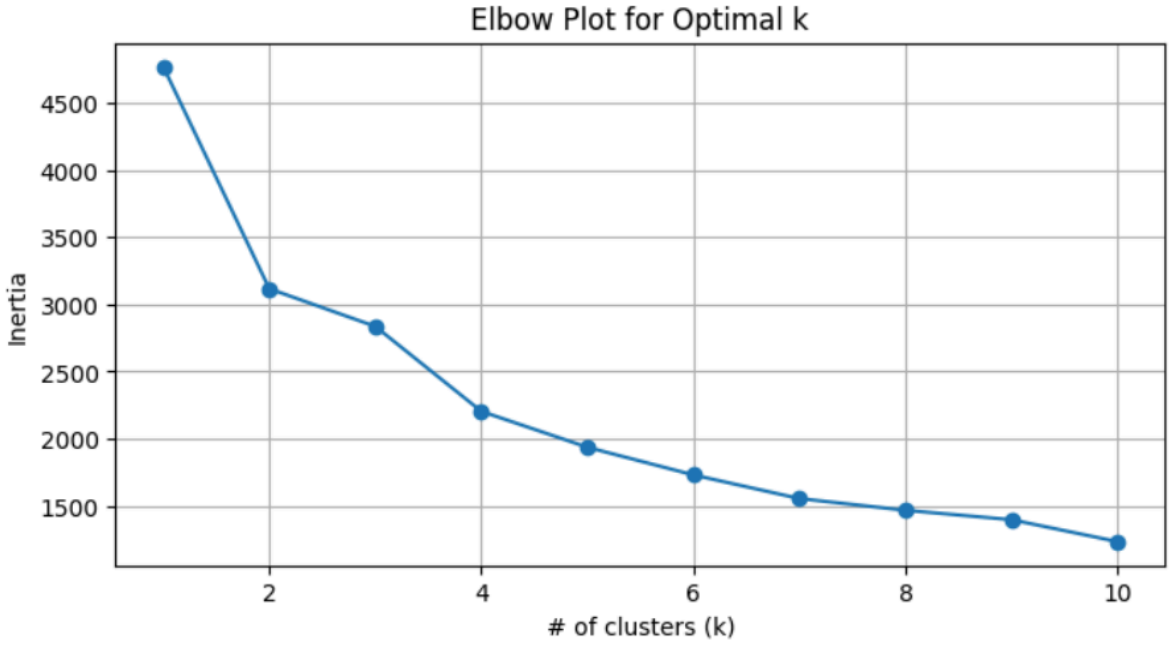


Figure 11: Elbow Plot for Optimal K

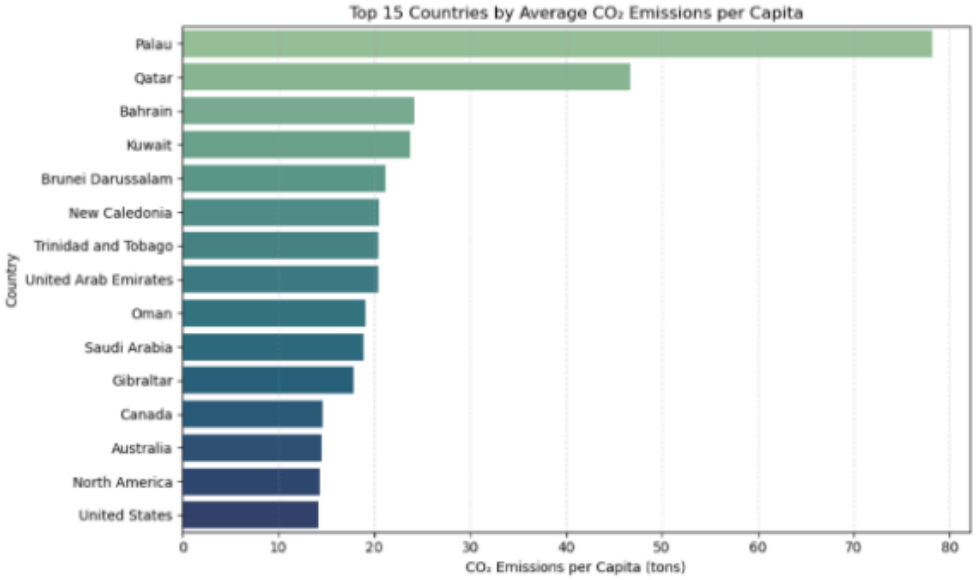


Figure 12: Shows the top 15 countries with average CO2 Emissions per capita

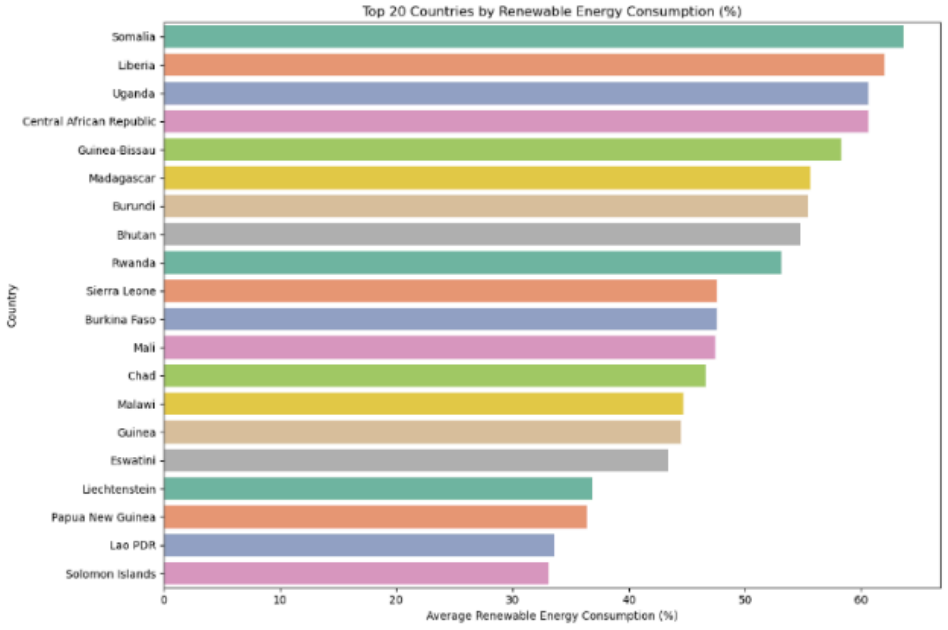


Figure 13: Top 20 countries by renewable energy consumption(%)

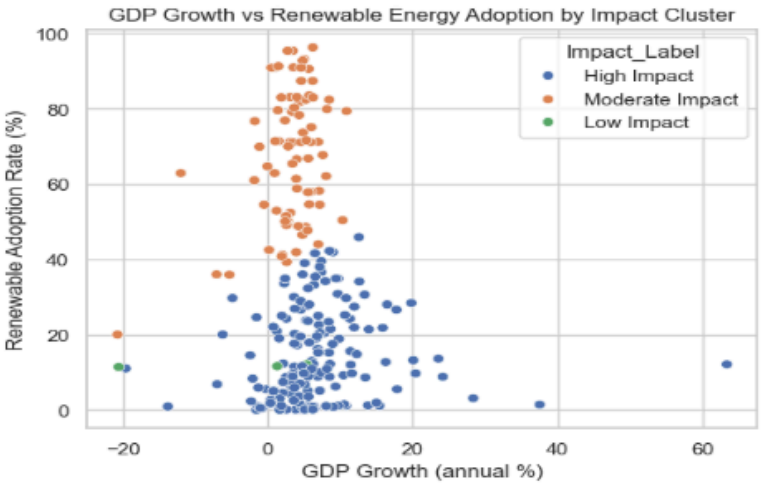


Figure 14: GDP Growth vs Renewable Energy Adoption by Impact Cluster

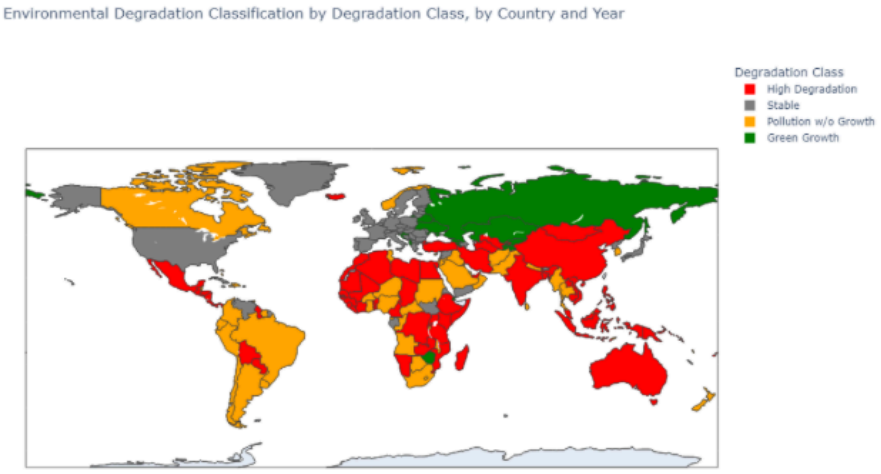


Figure 15: Environmental Degradation Classification by Degradation Class, Country, and Year

Classification Report:

	precision	recall	f1-score	support
Green Growth	1.00	1.00	1.00	19
High Degradation	1.00	0.99	0.99	68
Pollution w/o Growth	0.97	1.00	0.98	29
Stable	1.00	1.00	1.00	44
accuracy			0.99	160
macro avg	0.99	1.00	0.99	160
weighted avg	0.99	0.99	0.99	160

Figure 16: Classification Report for XGBoost

=== Classification Report ===

	precision	recall	f1-score	support
Green Growth	1.00	0.95	0.97	19
High Degradation	1.00	0.99	0.99	68
Pollution w/o Growth	1.00	1.00	1.00	29
Stable	0.96	1.00	0.98	44
accuracy			0.99	160
macro avg	0.99	0.98	0.99	160
weighted avg	0.99	0.99	0.99	160

Figure 17: Classification Report for Random Forest

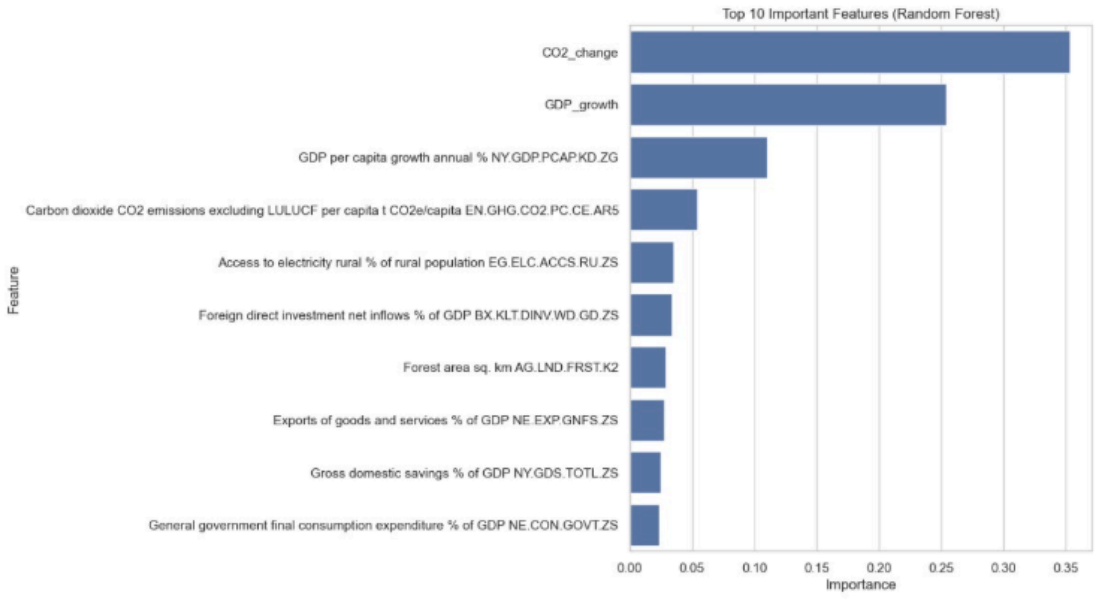


Figure 18: Feature Importance.

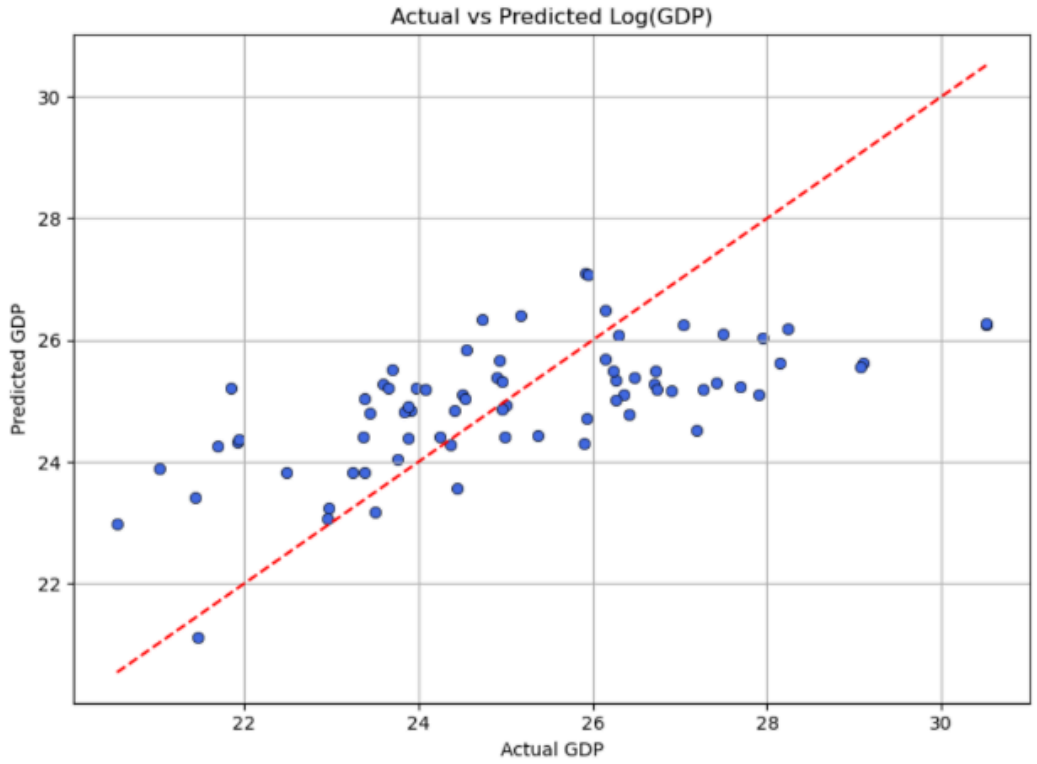


Figure 19: Predicted vs Actual GDP

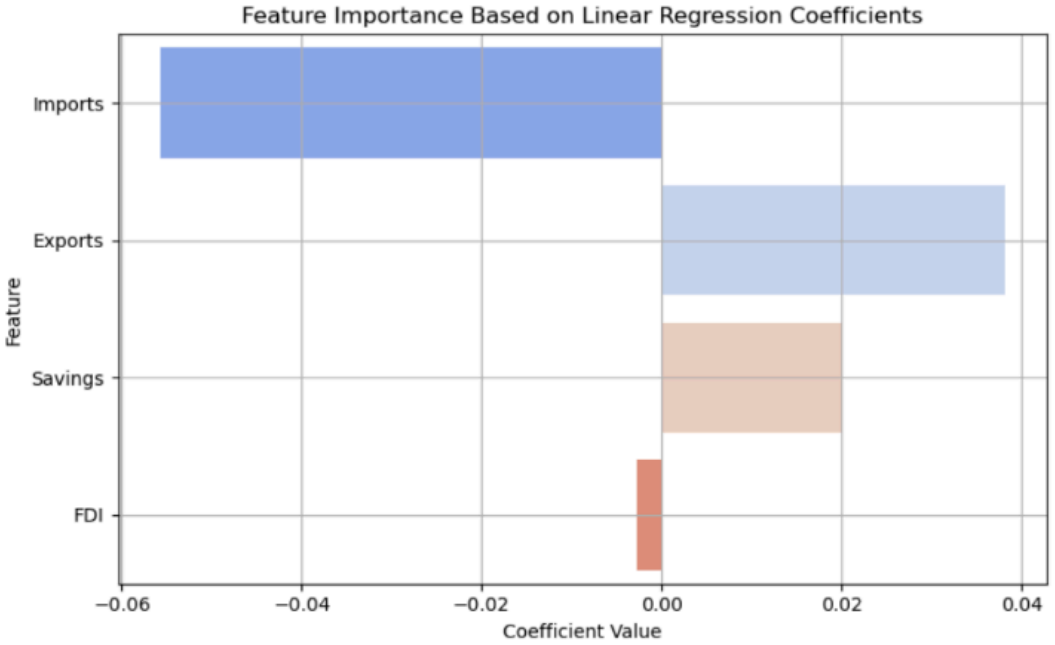


Figure 20: Shows feature importance based on Linear Regression coefficients chosen

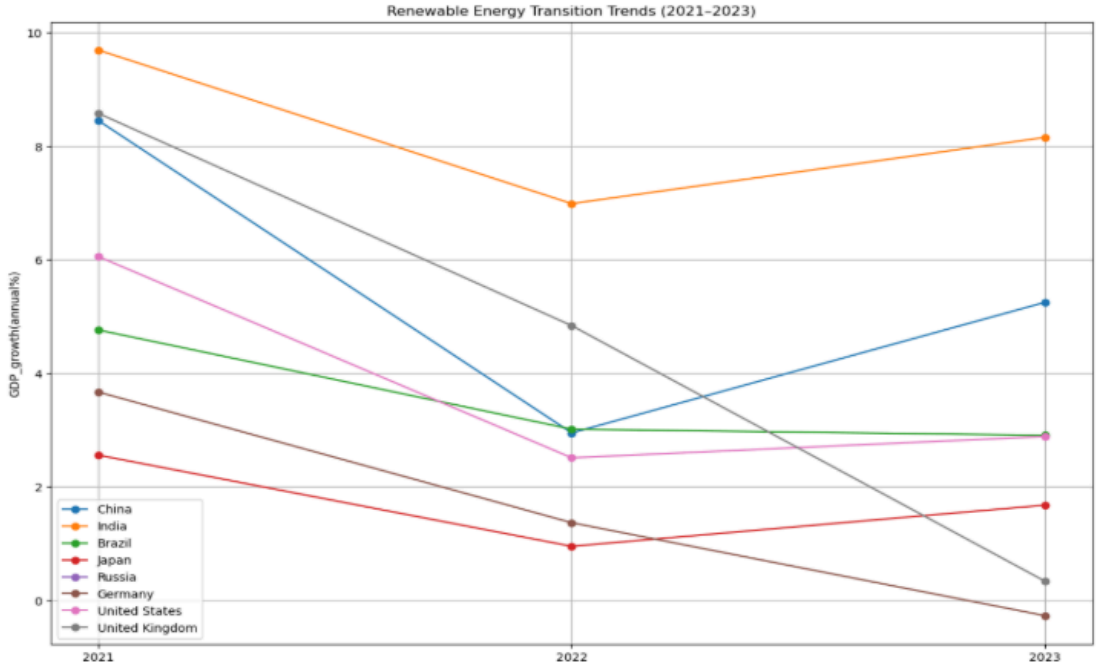


Figure 21: Renewable Energy Transition Trends over 3 years

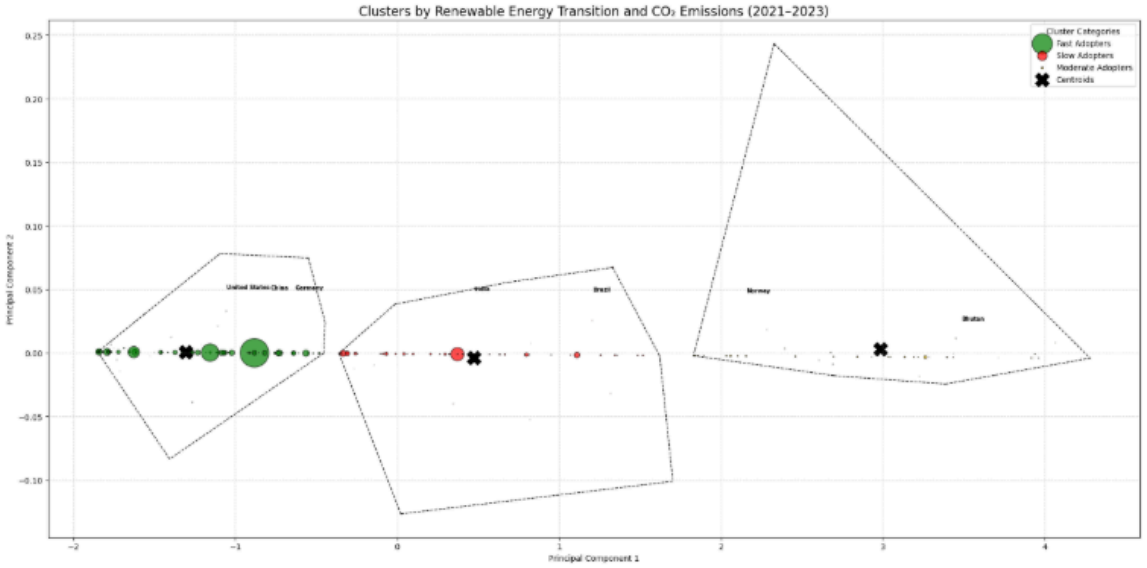


Figure 22: Clustering on Renewable Energy Transition and CO_2 Emissions over time

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
Albania	Burkina Faso	Albania	Argentina	Iceland	Sri Lanka	China
Australia	Congo, Dem. Rep.	Armenia	Barbados	Tonga		United States
Austria	Cote d'Ivoire	Bangladesh	Lebanon			
Barbados	Kenya	Bolivia	Russian Federation			
Belgium	Madagascar	Bosnia and Herzegovina	Tajikistan			

Figure 23: Table of Top 5 Countries in each Hierarchical Cluster

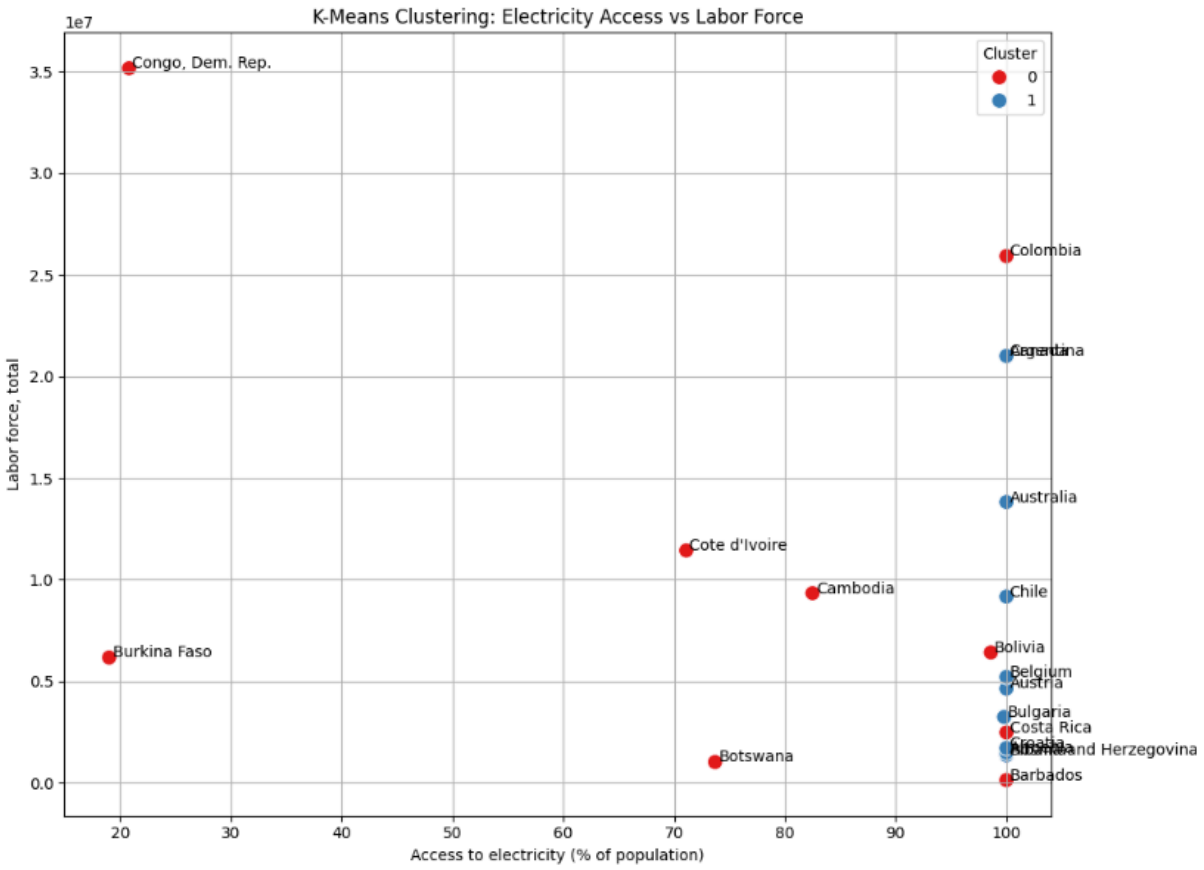


Figure 24: K-Means Clusters as a Scatter Plot