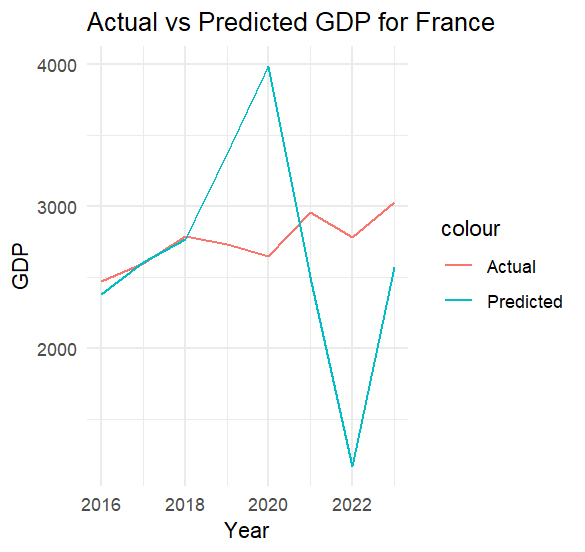
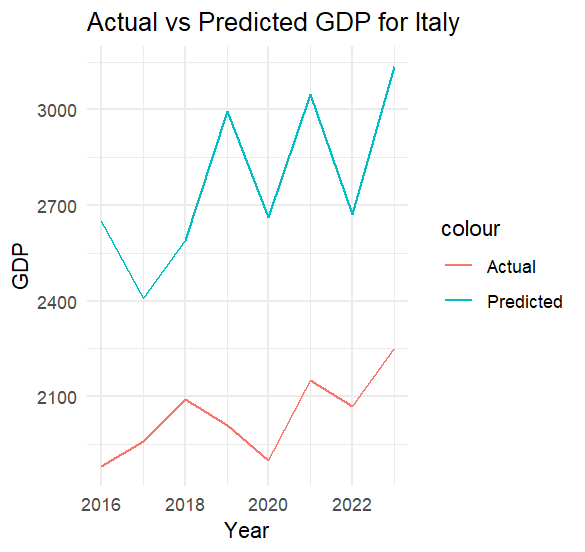
Model Implementation:

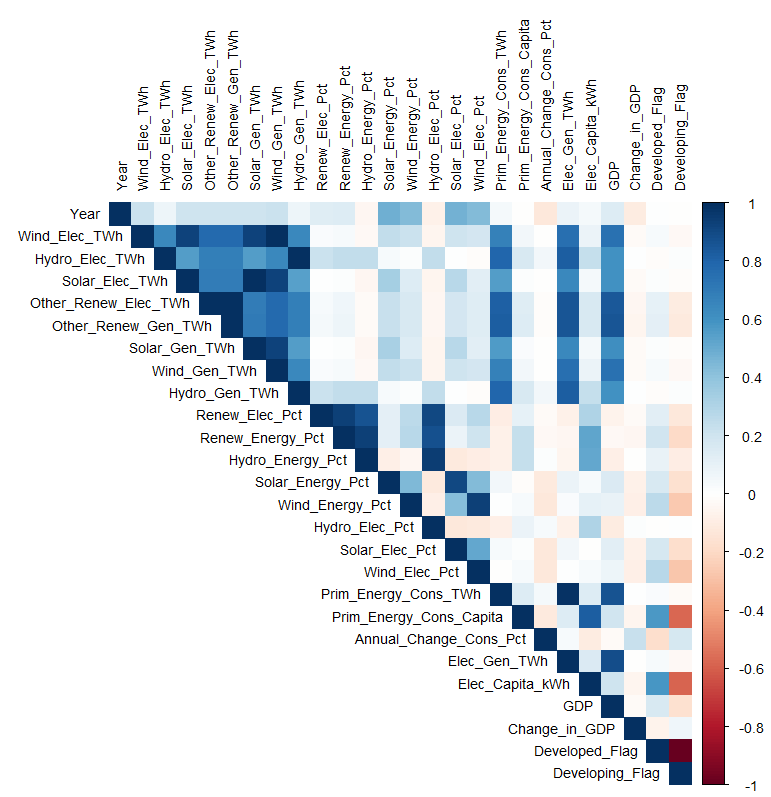
After carefully looking through the data and cleaning it, the first thing that came to our mind was using linear regression. Given that the GDP was linear and so were all the other variables. The plotted graphs all showed the data was going in a linear fashion so we started with that. When we ran the linear regression,the data seemed inflated with the R squared being 0.99 which does not make sense. We decided to dive deeper into it and tried to predict the GDP using linear regression and compared it to the actual.

Suggested:  
Upon thoroughly examining and cleaning the data, our initial approach was to utilize linear regression, as both the GDP and the other variables exhibited linear trends. The plotted graphs indicated a linear relationship, leading us to proceed with this method. However, the results from the linear regression analysis appeared inflated, with an R-squared value of 0.99, which seemed implausibly high. Consequently, we decided to investigate further by comparing the predicted GDP values from the linear regression model to the actual GDP figures.



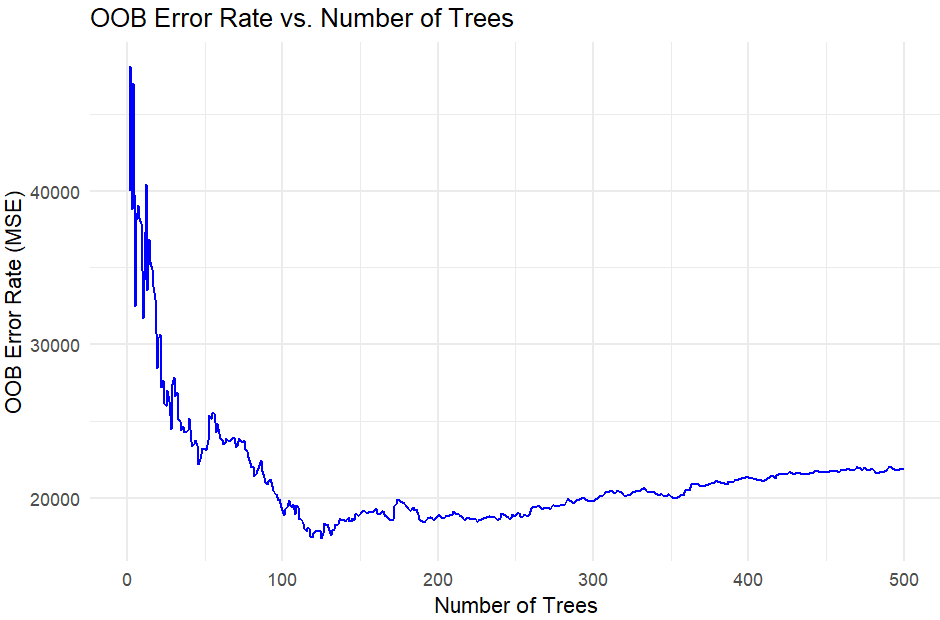
We can see that the predictions are all over the place and they are not accurate at all. The next step we decided to take was to check the correlation between variables and themselves. From the correlation matrix we could see that the variables are highly correlated with each other and therefore there is multicollinearity. We found some interesting discoveries like that some variables are negatively correlated with GDP like bioenergy reliance (in some regions). We also made a VIF test just to make sure and it also agreed with our hypothesis. Wr concluded that linear regression is not suitable for this data and decided to go with random forest algorithm

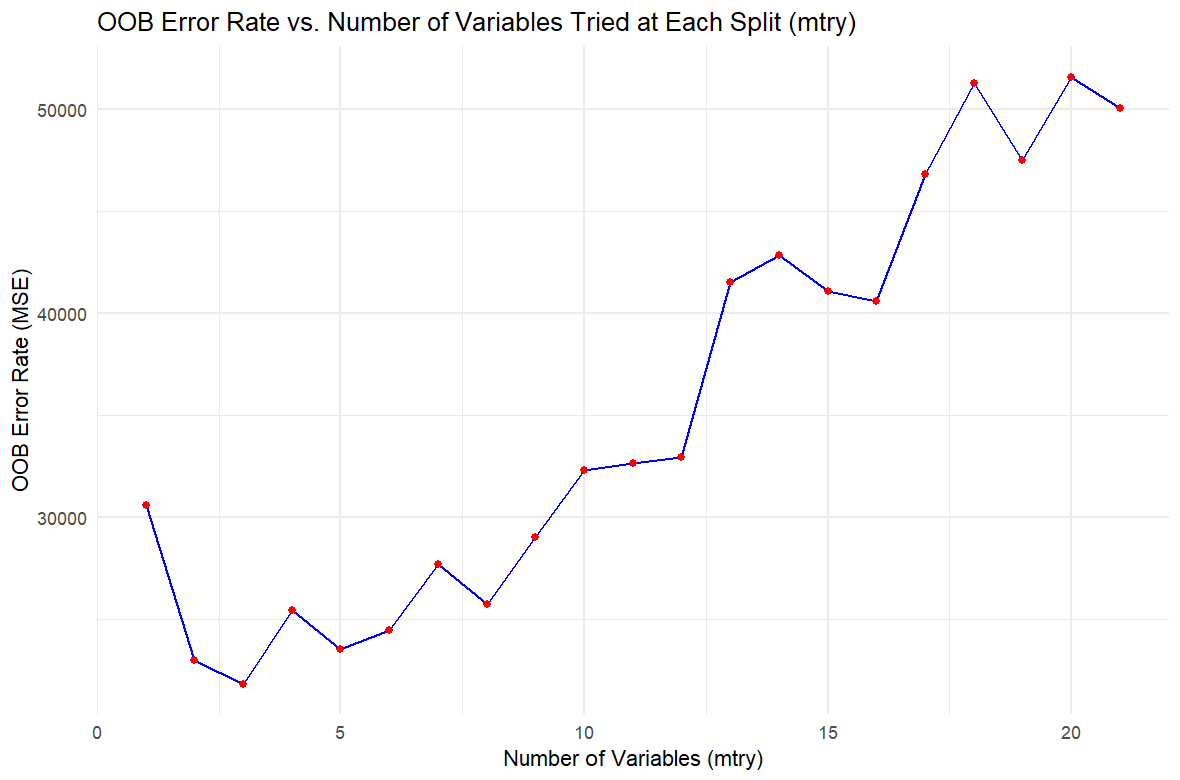
Suggested: We observed that the predictions were highly inconsistent and lacked accuracy. To address this, our next step was to examine the correlation between the variables. The correlation matrix revealed a high degree of multicollinearity, with variables being strongly correlated with each other. Interestingly, we discovered that certain variables, such as bioenergy reliance in specific regions, were negatively correlated with GDP. To confirm our findings, we conducted a Variance Inflation Factor (VIF) test, which supported our hypothesis. Consequently, we concluded that linear regression was unsuitable for this dataset and opted to use the Random Forest algorithm instead.



For this project we decided to implement our model on 5 regions and picked an economic powerhouse in each region to represent it. For Asia we chose India. Germany for Europe. Brazil for South America. Saudi Arabia for the Middle East and Canada for North America. For each country we split the data into an 80/20 split and we used random forest to create 500 trees. We observed the OOB error rate for each tree and picked the optimal number of trees to use. After that we tried finding the best number of features to try combinations with also known as the Mtry. The Mtry is the best number of variables to use after each split. We also plotted these results to make it easier to understand and interpret.

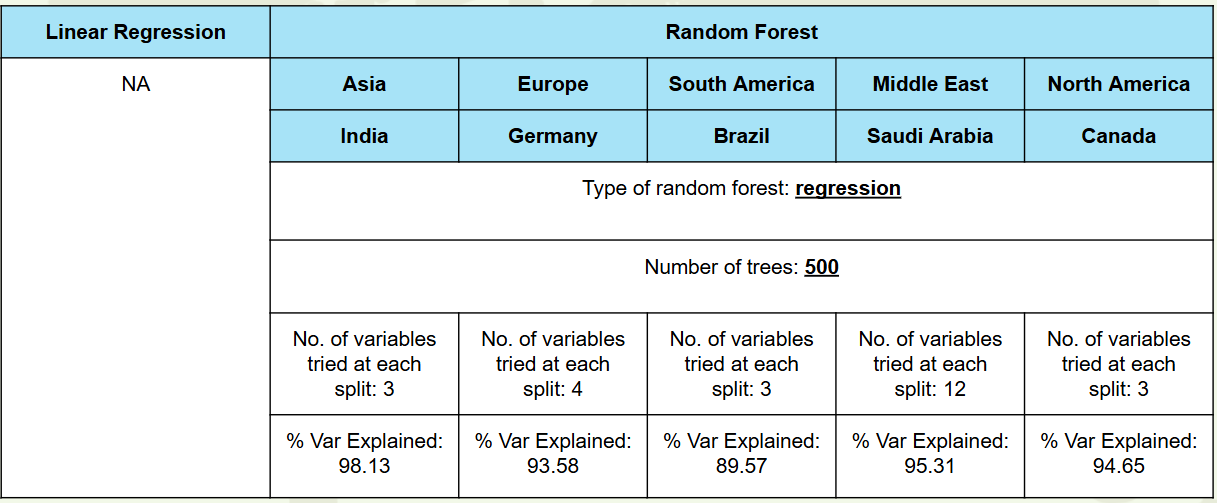
Suggested: For this project, we implemented our model across five regions, selecting a leading economy to represent each one. We chose India for Asia, Germany for Europe, Brazil for South America, Saudi Arabia for the Middle East, and Canada for North America. For each country, we divided the data into an 80/20 split and used the Random Forest algorithm to create 500 trees. We monitored the Out-Of-Bag (OOB) error rate for each tree and determined the optimal number of trees to employ. Following this, we identified the best number of features to test combinations with, known as the Mtry—the ideal number of variables to use after each split. To facilitate understanding and interpretation, we plotted these results.





The figures above show an example of our process for India. From the figure we can see that the best number of trees to use for India is around 120 trees. These are the optimal numbers. After that we used the OOB Error rate vs Mtry graph to see the best number of variables to use after each split. In India's case it is 4 as 4 as the lowest OOB error rate. We repeated this process for the other 4 countries.

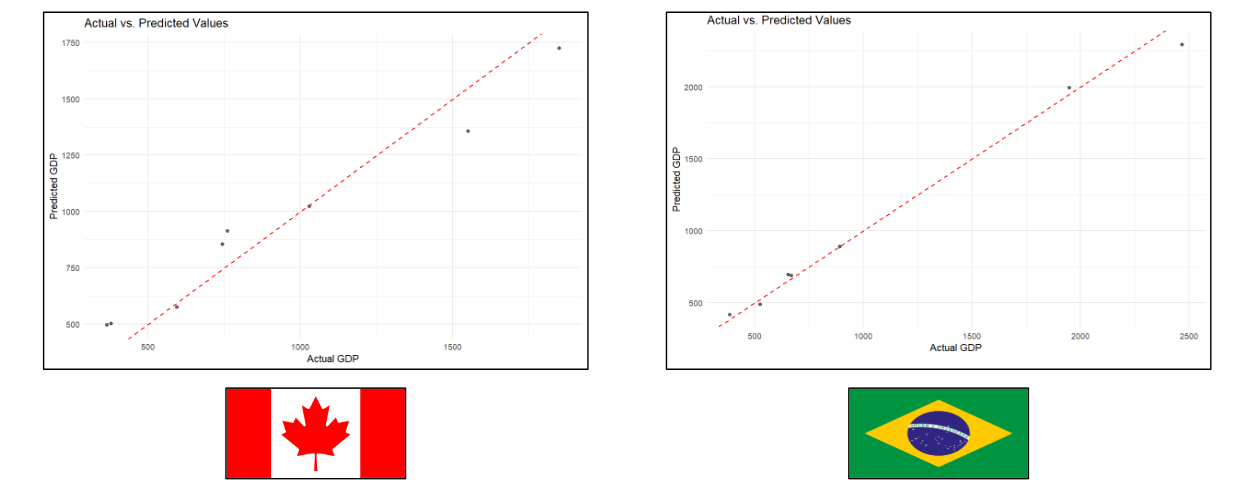
Suggested: The figures above illustrate our process for India. The optimal number of trees for India was determined to be approximately 120. We then utilized the OOB Error rate vs. Mtry graph to identify the best number of variables to use after each split, which for India, was found to be 4, achieving the lowest OOB error rate. This methodology was subsequently applied to the other four countries.

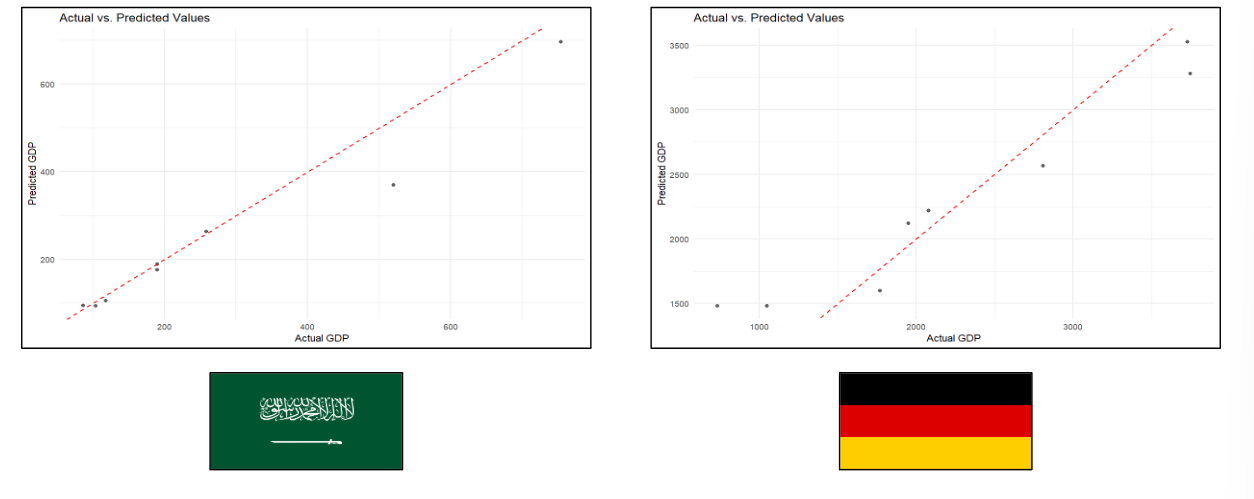


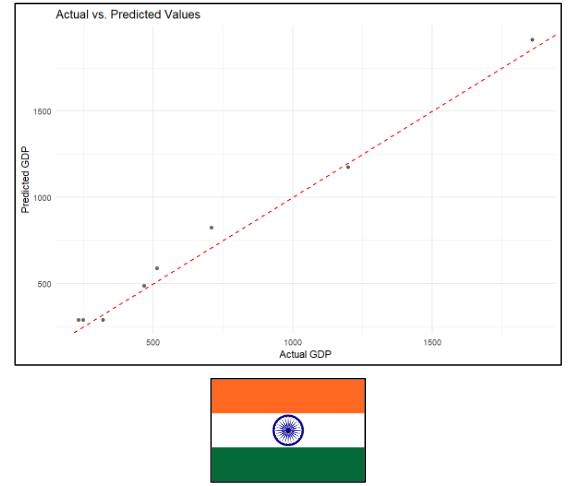
After applying the process to all 5 countries, we gathered the data and put it in the table above. We can see that we used the regression type for the random forest algorithm. We also tested the best number of trees for all 500 countries and found out that all of them except India are more accurate the more trees are made so the optimal number of trees for all the countries not named India were 500. We then took this information and tried to see the optimal number of variables to try at each split (Mtry). With the number of variables tried at each split being relatively low (3 or 4) for most trees, the model has a balanced approach to learning patterns in the data without overfitting. As for Saudi Arabia, we can see it has the largest number of Mtry which means the algorithm looks at 12 variables after each split leading to lower bias but higher variance. The percentage of variance explained (VAR explained) is a metric used to describe how well a model accounts for the variability in the target variable. The higher the number the better the model is. We can see that the VAR explained is really good in all models. Given all the data we have we decided to plot the predicted vs actual GDP for all the countries.

Suggested: After applying the process to all five countries, we compiled the data into the table above. We employed the regression type for the Random Forest algorithm. Upon testing the optimal number of trees, we discovered that, except for India, the accuracy of predictions increased with more trees. Consequently, the optimal number of trees for all countries, except India, was 500. Using this information, we then determined the optimal number of variables to try at each split (Mtry). Generally, the number of variables tried at each split was relatively low (3 or 4), allowing the model to balance learning patterns without overfitting. In the case of Saudi Arabia, the highest Mtry value was observed, with the algorithm considering 12 variables after each split. This led to lower bias but higher variance.

The percentage of variance explained (VAR explained) is a metric used to describe how well a model accounts for the variability in the target variable; higher values indicate better model performance. Our results showed high VAR explained values across all models. Based on this data, we decided to plot the predicted versus actual GDP for all the countries.







We can see that the random forest predictions are mostly accurate, especially in the case of Brazil. We can see the model is a lot more accurate than the linear model and that is why we decided to take the data from random forest and then use it on ARIMA for forecasting.

Suggested: The predictions generated by the Random Forest model demonstrated high accuracy, particularly for Brazil. This marked improvement over the linear model validated our decision to leverage the Random Forest data for further forecasting using the ARIMA model. By combining these two approaches, we aimed to enhance the reliability of our predictions.

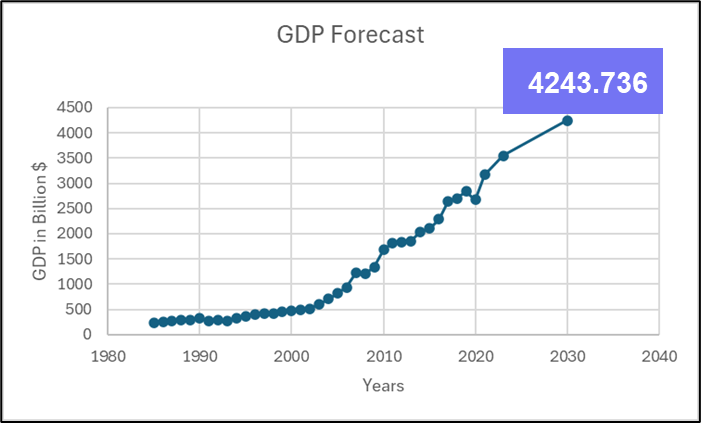
—----------------------------------------------------------------—-------------------------------------------------------

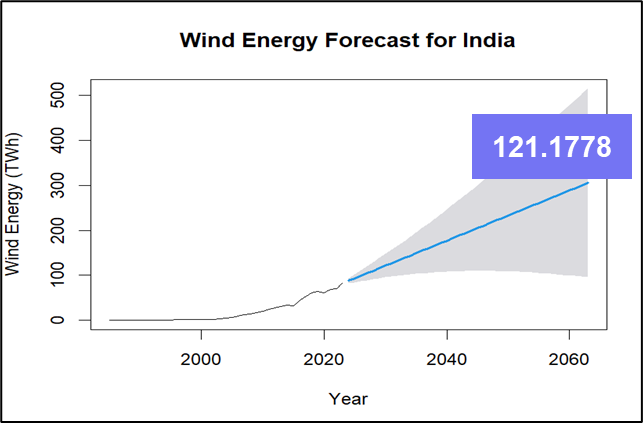
—----------------------------------------------------------------—-------------------------------------------------------

—----------------------------------------------------------------—-------------------------------------------------------

**Forecasting:**

For forecasting GDP, we focused on India for the year 2030, given that many countries have climate goals set for this year (United Nations, 2015). First, we analyzed the decision tree produced by the Random Forest model to identify the input parameters influencing the prediction. In the case of India, wind energy in terawatt-hours (TWh) was the key predictor. Subsequently, we used the ARIMA model to forecast the value of wind energy in 2030, plotting a line graph with a 95% confidence interval. Since the forecasted wind energy value was outside the range of our training data, we calculated the offset and input it into the Random Forest model to estimate the additional GDP that would result if this wind energy amount was realized. These values were then saved in separate CSV files, and their graph was plotted to illustrate the forecasted GDP.

****

****

**Conclusion:**

This project set out to create predictive models for GDP, leveraging renewable energy data from various countries and years. The outcomes hold significant importance for both policymakers and businesses. Renewable energy, although not the only factor, plays a notable role in influencing GDP. Specifically, wind energy stood out as the most critical factor impacting economic growth, underscoring the potential benefits of investments in wind energy.

To achieve our goal, extensive data wrangling was necessary to build a reliable dataset. This step was crucial to ensure the precision and dependability of our predictive models. We chose the Random Forest model for training and testing the data, which proved effective in forecasting GDP trends. Our model particularly focused on the economies of regions with the highest renewable electricity consumption, as noted in the 2022 report, providing valuable insights.

The challenges posed by multicollinearity and variability in GDP trends across different countries were managed through meticulous data handling and model selection. The results are instrumental for policymakers designing effective carbon tax and renewable energy incentives and for businesses evaluating economic risks and opportunities linked to renewable energy. By understanding the impact of renewable energy production on GDP, stakeholders can make informed decisions that drive sustainable economic growth.

In summary, our project highlights the significant role of renewable energy in shaping economic outcomes and offers essential insights for future policy and business strategies. These findings support the creation of targeted incentives and investments that can promote both economic and environmental benefits.