Impact of Developed Countries' Clean Energy Production on Developing Nations

Liam O'Donnell, Will Coughlin, Joshua (Josh) Geiger, Stephanie Poole, Jonathan Ho

https://github.com/MGT-6203-Fall-2023-Edx/Team-4/

Introduction

The clean energy transition has been a politically and emotionally tumultuous issue in countries around the world. This project seeks to understand to what degree developed nations influence less-developed nations, specifically whether a change in primary energy sources from highly developed countries affects less developed countries. Other topics of interest are: (1) will an evolution in the type of energy production impact a country's economic health; and (2) can energy capacity by source guide clean energy investment decisions?

Countries in the same region may exhibit similar trends in energy production by type over time. Government and private energy firms may be interested in understanding the relationship between energy type and GDP to drive investment. If a highly developed country's energy sources can impact a less developed country's energy sources, partnerships and investments can be strategically optimized. By analyzing these trends, leaders of developing countries could focus effort on which alternative sources of energy to pursue and predict potential demand and GDP increases. Finally, nonprofits within the energy sector could use the data to determine opportunities in energy source type by country to drive investment and development strategy.

Data Background

The primary metric for gauging countries on the basis of development is the Human Development Index. As defined by the United Nations, "Human Development Index (HDI) is a summary measure of average achievement in key dimensions of human development: a long and healthy life, being knowledgeable, and having a decent standard of living. The HDI is the geometric mean of normalized indices for each of the three dimensions."[1] Using the calculated HDI, each country is also sorted into a discrete classification to describe its development level: "Low," "Medium," "High," and "Very High."[2] One may consider Low and Medium HDI Countries to be "Developing", while High and Very High HDI countries are "Developed."

Energy production will be quantified with generation in terawatt-hours (TWh) and capacity in gigawatts (GW). While generation represents the actual quantity of energy produced, capacity is the maximum theoretical power output of all installed sources. Capacity will represent output that has been committed to since it is unaffected by external factors such as demand and outages, unlike generation. Finally, Gross Domestic Product (GDP) will be used as an economic metric, separate from HDI, for each of the selected countries. This represents the

[1] United Nations, "Human Development Index," Human Development Reports, n.d., https://hdr.undp.org/data-center/human-development-index#/indicies/HDI.

[2] United Nations, "Data Reader's Guide," Human Development Reports, n.d., https://hdr.undp.org/reports-and-publications/2020-human-development-report/data-readers-guide.

value of all goods and services produced in a country for the given year, reported in 2017 international dollars. Furthermore, this value is presented on a per-capita basis to adjust for population size (GDP per capita).

Chiefly, the preprocessing stage was required to transform the multiple data sources into a single table to be used for further exploration and modeling. During this process, each separate source was examined to decide which data to keep and which keys could be used to merge the individual tables into a single set. This procedure was developed in purpose-built R scripts and then combined into a single R script that performs all steps needed in creation and exporting of the final datasets. It was determined after some preliminary modeling that more variables may be of interest than initially anticipated. A “wide” dataset export was then added to provide the data modeler with more potential variables of interest.

Upon examining each dataset for a primary key, it was observed that each source used a three character country code (ISO 3166[3]) that could be combined with the year as a unique identifier. Of these datasets, the first dataset considered was the energy dataset. As the primary data source, the key variables extracted from this dataset were capacity and generation by clean energy and fossil fuel energy. The country's energy demand and energy imports were also retained. GDP per capita and HDI values were extracted from the second and third datasets, respectively. Additionally, a country name lookup table was created that contains each unique country name indexed by the country code from the HDI dataset. Each of the four cleaned tables were joined using a full join to ensure no rows were removed until the modeling phase. It is clear from the unified table that not every country has all data available for every year. The total observation count is 4,688 rows across 214 countries for years between 2000 and 2021.

Data Exploration Insights

Exploratory data visualizations were produced to narrow the focus of the primary and supporting business objectives. This section presents significant takeaways from this stage.

Clean energy generation was investigated first to observe the real global trend in fossil versus clean energy output. Figure 1 suggests an increased momentum in clean energy production, where the slope of the trend line for *GenClean* (actual clean power in TWh) increases in recent years. An additional observation is generation of fossil fuels appears to be slowing down overall. The annual percent change in generation by type is shown in Figure 2 and illustrates that clean energy generation consistently increases with less volatility than fossil fuel generation.

[3] ISO. “ISO 3166 — Country Codes,” n.d. https://www.iso.org/iso-3166-country-codes.html.

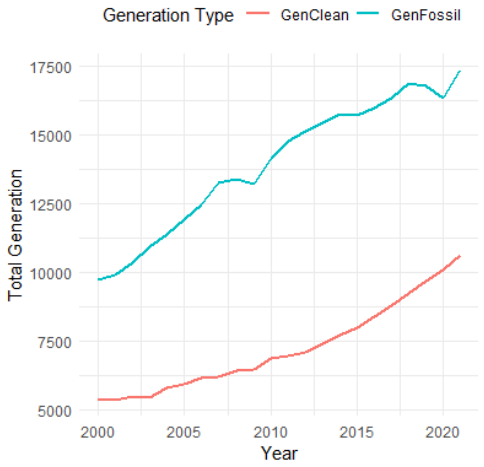

Figure 1: Clean and fossil fuel generation by year

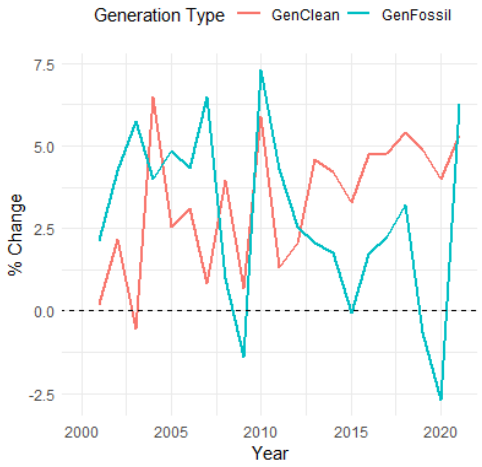

Figure 2: Percent change by generation source per year

From the cleaned dataset, a column of data referred to in this report as *CapCleanProportion* was produced. This is the ratio of clean energy capacity to total capacity for a country in a given year. Capacity was the focus rather than generation, for a representation of commitment and capability. The plot of *CapCleanProportion* by year and *HDICode* in Figure 3 exhibits a clear upward trend in clean energy capacity proportion for High and Very High HDI countries. Overall, the trends for Low and Medium HDI countries appear to be flat. In the last 10 years of observations, the trend for Medium HDI countries appears to trend upward, but Low HDI countries appear to exhibit no significant change.

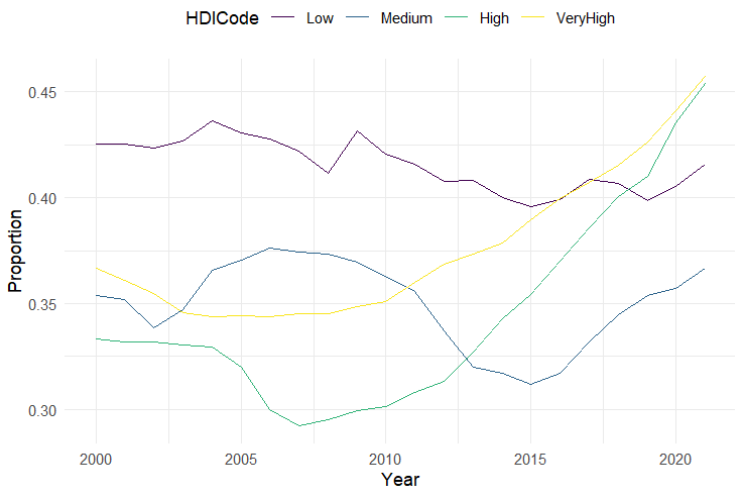

Figure 3: *CapCleanProportion* by year (2000 – 2021) and HDI Group

The Effect of Developed Countries' Habits on Developing Countries

Multiple Linear Regression

An initial attempt was made to predict *CapCleanProportion* for developing countries using values from developed countries and several other possible explanatory variables from the dataset. Backward Stepwise regression, LASSO and Elastic Net were applied to build multiple linear regression models for the years 2006, 2011, and 2018. Leave-One-Out Cross-Validation was conducted to measure performance. While the Adjusted R^2 of the selected model for each year was acceptable, the predictions obtained from them were not. Instead, time series techniques were applied to model these values to capture each year within the dataset. The next section will give an overview of these results.

Time Series Modeling

To reduce the overall problem size, the dataset was broken up by continent. Exploratory multiple linear regression showed that when *Continent* was treated as a categorical variable to predict *CapCleanProportion*, some of its coefficients were significantly different from zero. Thus, a separation of subsets based on this factor may account for intrinsic properties and lead to more robust, specialized models.

Furthermore, each continent was separated into its constituent High and Low Development countries. *CapCleanProportion* was calculated as a total over the whole category within that continent, resulting in two time series per continent: *CapCleanProportion* for Low Development countries and *CapCleanProportion* for High Development Countries. Using correlation analysis, it was found that grouping Low and Medium *HDICodes* into a single Low Development group and High and VeryHigh *HDICodes* into a single High Development group was broadly suitable. The remaining steps in this section were carried out for each continent.

Each pair of time series was split with the most recent 5 years held out for testing. Then, with the training set, cross-correlation of Low Development *CapCleanProportion* against lags of High Development *CapCleanProportion* was inspected. The maximal lag length to consider was determined with a correlation coefficient cutoff of 0.3, as a heuristic. Figure 4 shows the cross-correlation plots for Africa. Observe that the correlation coefficient for the Low Development series and the High Development series with lags from 0 to 5 indicate moderate correlation with values greater than 0.3. Further lags show far weaker correlation, however it appears to grow stronger again with lag lengths of 10 and 11. To account for the series lengths and to favor simpler models, lag lengths with high magnitude relative to the total number of observations were ignored.

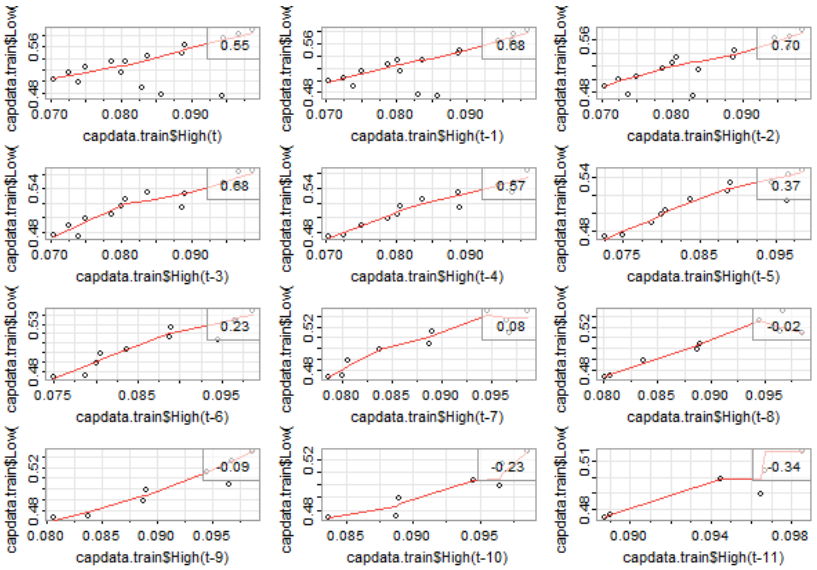

Figure 4: Low Development vs High Development *CapCleanProportion* values in Africa

Next, a baseline ARIMA model was built for Low Development *CapCleanProportion.* This baseline model was built with the appropriate number of years trimmed from the start of the training series to account for the maximum High Development series lag length. This trimming further reduces the size of an already small training set. The 'auto.arima' function in the forecasts package fits multiple models on the given series and selects the best model based on the desired criteria. Corrected AIC (AICc) was used to correct for small samples. For this selected model, residual diagnostics and a hypothesis test for autocorrelation within the series were checked to ensure model validity.[4]

After fitting a valid univariate model, the values for High Development *CapCleanProportion* were introduced as an exogenous series. Supplying the series to the 'auto.arima' and 'Arima' functions in the 'forecasts' package will create Regression with ARIMA Errors models.[5] For each possible lag length from 0 to the maximum selected in the previous step, the optimal model was built based on AICc. Then, the model representing the optimal lag length was selected, again using AICc. Other types of models for capturing the relationship were considered, but the available tools in the R, as well as the interpretability of the Regression with ARIMA Errors model, as described by the creator of the forecasts package,[6] supported the selection of this technique.

Subsequently, performance of the baseline ARIMA model and the best exogenous predictor model were compared using the test set. Figures 5 and 6 visualize the predicted values and prediction intervals given by each model compared to the actual values in the test set. In the case of Africa, the models appear to be visually identical. In support, accuracy

[4] The Pennsylvania State University, "3.2 Diagnostics," STAT 510 Applied Time Series Analysis, https://online.stat.psu.edu/stat510/lesson/3/3.2.
[5] Rob J Hyndman and George Athanasopoulos, *Forecasting: Principles and Practice*, chap. 9, sec. 2, https://otexts.com/fpp2/.
[6] Rob J Hyndman, "The ARIMAX Model Muddle," Hyndsight, https://robjhyndman.com/hyndsight/arimax/.

metrics of each showed very slight differences with the exogenous predictor model having a slightly lower root mean squared error (RMSE). Significant increases observed in the error metrics between training and testing indicate overfitting and the capture of random patterns due to the very small size of the time series.

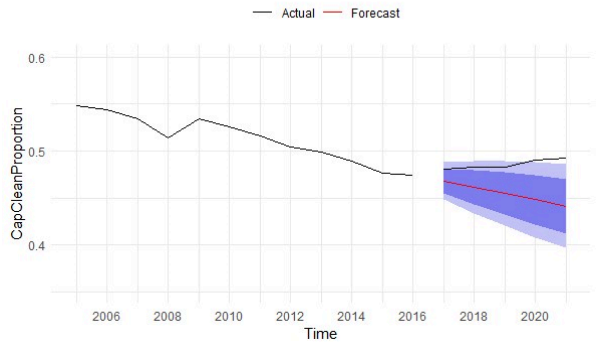

Figure 5: Forecasts from baseline Africa model (ARIMA(0,1,0) with drift)

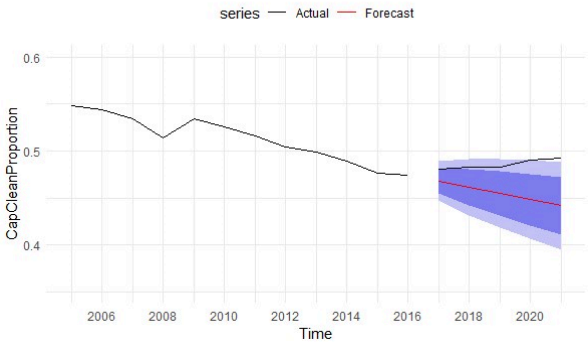

Figure 6: Forecasts from Africa model with exogenous predictor (Regression with ARIMA(0,1,0) errors)

A case where the exogenous model performed significantly worse than the simple model is that of Asia. A comparison of Figures 7 and 8 shows severe deviation from the true values when introducing exogenous predictors.

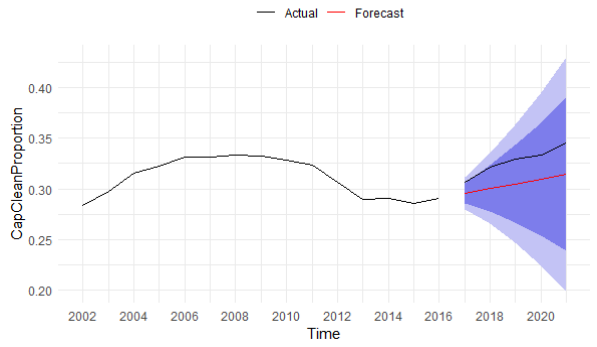

Figure 7: Forecasts from baseline Asia model (ARIMA(0,2,0))

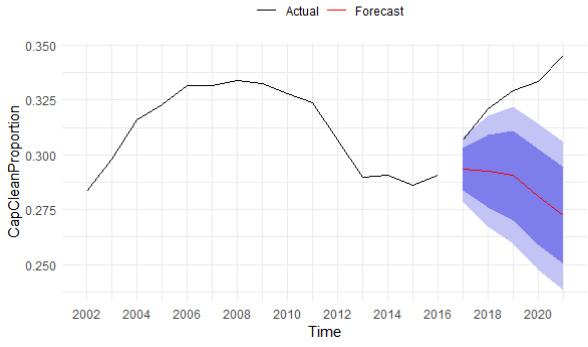

Figure 8: Forecasts from Asia model with exogenous predictor (Regression with ARIMA(2,0,0) errors)

Finally, the best available model for each continent was retrained on the full dataset. These completed models provide a launchpad for forecasting and any other future refinement. Most notably, only Africa and North America showed any potential for use of *CapCleanProportion* from High Development countries to predict that of Low Development countries.

Clean Energy Generation/Capacity as a Predictor for GDP

To address the supporting business question on whether the use of clean energy affects a country's GDP, the group analyzed *CapClean* and *GenClean*. For exploratory data analysis (EDA), the group expanded upon the graph shown in Figure 1 to include *GDPPerCapita*. The results are shown in Figure 9.

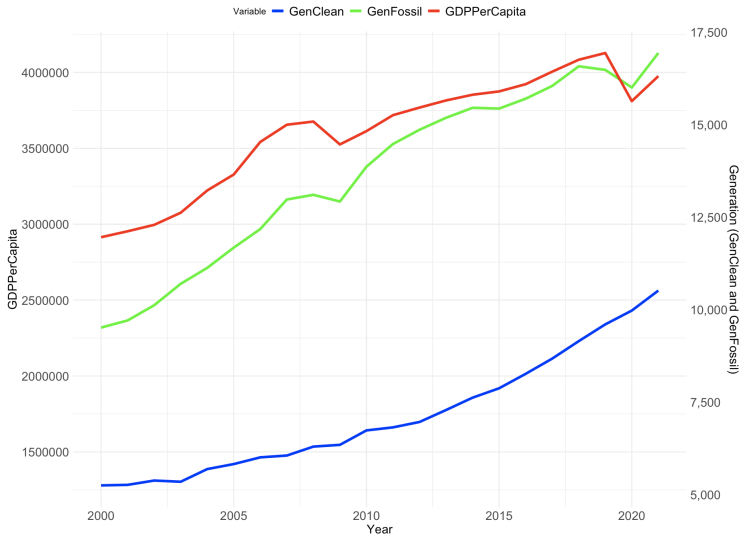

Figure 9: Aggregate Production Data by Year

Based on Figure 9, it may be reasonable to conclude that overall generation of fossil fuels strongly correlates with a country's GDP more than clean energy generation. However, it is important to consider that the figure includes data from all countries. To investigate the statistical relationship between GDP and clean energy generation, a linear regression model was constructed on GDP per capita and clean energy generation. Although clean energy shows statistical relevance at alpha equal 0.05, the near-zero coefficient indicates no significant relationship. Upon reassessment, it is evident that a holistic linear regression model could not be applied due to the lack of independence and identical distribution (i.i.d.) among data points, amplified by variations in different countries' clean energy consumption.

To solve the issues of the data points not being i.i.d. and the model having vastly different values of *GenClean* and *CapClean*, the group focused on one year (2012). A linear regression was performed with *GDPPerCapita* as the target variable using *CapCleanProportion (CapClean_prop* in figures in this section*)*, region code, and total energy import. The scatter plot matrix is shown in Figure 10. Based on the irregular shapes shown in the scatter plot matrix, the data set was scaled and redistributed using a Box-Cox transformation. A similar linear model was run, with this version using the scaled and normally distributed data. *CapCleanProportion* remained statistically irrelevant at alpha equal to 0.05. The Q-Q plot of this model shown in Figure 11 has an irregular shape and confirms that alternative methods need to be explored.

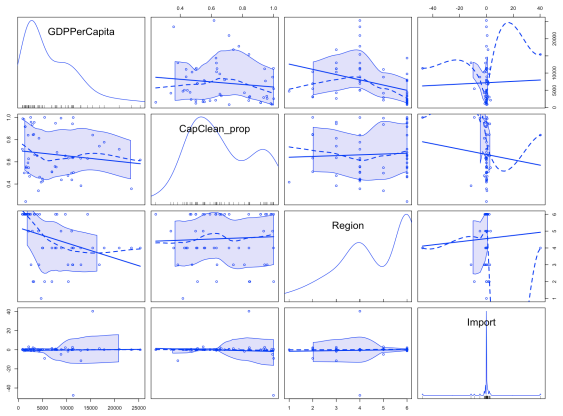

Figure 10: Regression variable scatterplot matrix

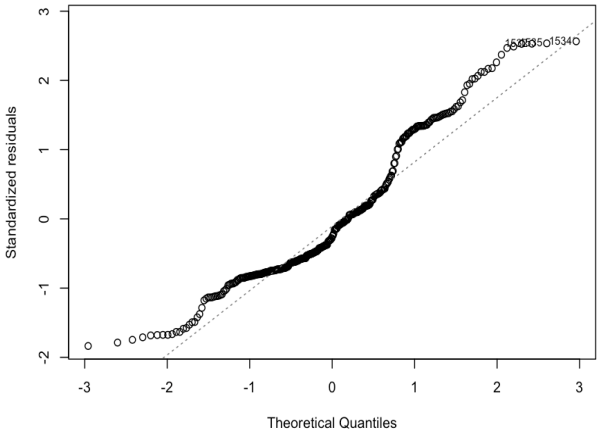

Figure 11: Regression QQ plot

Given the suboptimal results of multiple linear regression, the group transitioned to using a random forest model on the same set of data. While the random forest's predicted values were within 1 standard deviation on the validation dataset (2013), the RMSE increased significantly when compared to the regression model's performance on the same validation data. This was an indicator that the random forest model was overfit.

The modeling processes in this section were repeated using *GenClean* and *GenClean_prop* (calculated in a similar manner to *CapCleanProportion*). Both models exhibit characteristics similar to *CapCleanProportion*. As a final check to determine correlation, countries were grouped into one line and the median of *CapCleanProportion* was regressed against *GDPPerCapita* via random forest. The random forest exhibited a RMSE outside of 2 standard deviations of the mean value. Given the random forest and linear regression results on several dataset variations, the group determined that clean energy generation was not an adequate predictor for a country's GDP per capita. Further analysis is needed to determine if GDP detail can be expanded to support a statistically significant predictive model.

Guiding Clean Energy Investment Decisions

One of the secondary research questions investigated was guiding clean energy investment decisions. The initial approach taken to support the proposal was to forecast clean energy growth. The ARIMA models were used as the first attempt to provide insight on this question. As a reminder, the ARIMA approach discussed earlier aggregated clean energy data by continent. When using these models, the forecasts provided trivial guidance exhibiting a linear trend from the last datapoint with growing prediction intervals as the forecast length grew. An alternate approach was then taken to attempt to answer this question.

The second approach was to examine individual countries. The hypothesis was that looking solely at clean energy proportions by continent could bias the forecast based on the

largest clean energy countries in a given continent. Examining individual countries would allow a discrete approach to determining if there are valuable insights to guiding clean energy investments. To test this, each country's clean energy capacity proportion was plotted as a heatmap for all years and sorted by maximum clean energy proportion observed. Countries with absolute changes below 1 were excluded. The resulting heatmap (Figure 12) revealed distinct visual changes for certain countries, prompting further analysis using Holt-Winters exponential smoothing without seasonality. The findings aligned with the ARIMA approach, indicating a monotonic increase in clean capacity proportion and emphasizing trivial extrapolation of data with widening prediction intervals.

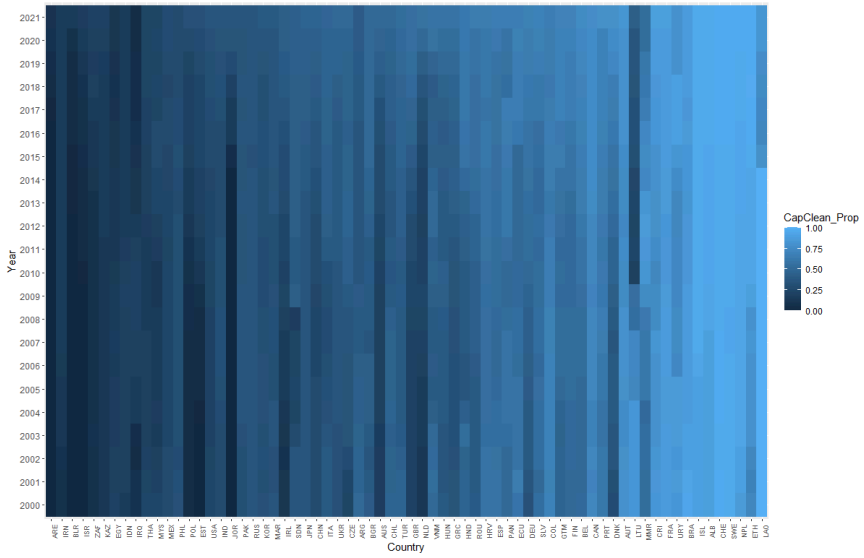

Figure 12: Heatmap of *CapCleanProportion* over time by Country

Still seeking to guide the investment potential of clean energy, the team pivoted to an approach adapting some financial modeling techniques. The clean energy capacity was treated as a country's "stock" price. For each year, the resulting change in clean energy capacity was computed as the "stock return." A final "compound return" was calculated for each country. After filtering the data in the same manner as above, the "compound return" data was plotted as a bar chart, as shown in Figure 13, grouped by HDI code and sorted based on the "compound return" for each country. The bars of the chart are filled with a logarithmic gradient color scale of the maximum clean energy capacity from the yearly dataset.

Figure 13, with "cumulative return," serves as a valuable tool for guiding investment strategy. For volume-based investments, identifying countries in green with a high "cumulative return" on the bar chart is recommended. Alternatively, for impact-driven investments, choosing countries with lower HDI codes and relatively low "cumulative return" outside the red gradient offers viable options. The graphical representation allows for diverse interpretations based on desired investment strategies.

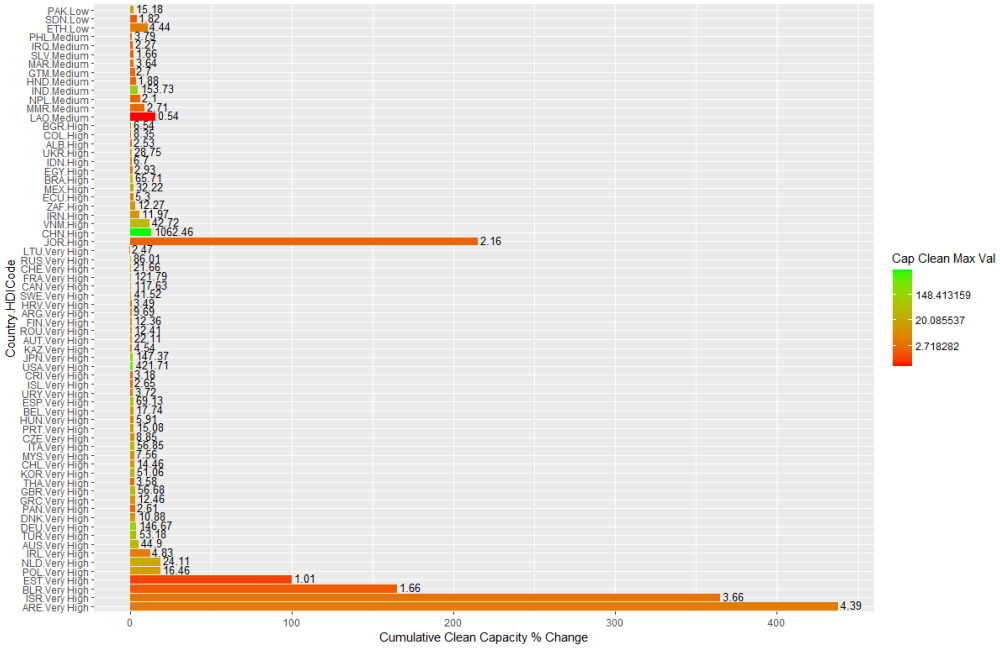

Figure 13: Cumulative Clean Capacity Change (%) by Country

Conclusion

Principally, the objective was to assess the impact of developed nations on less-developed nations, particularly how changes in primary energy sources from highly developed countries affect less developed countries. The primary research question was first analyzed using multiple linear regression techniques (e.g. Backward Stepwise regression, LASSO, Elastic Net) but resulted in poor predictive quality. Subsequently, time series modeling comparing a baseline ARIMA model and the best exogenous predictor model was applied with test and training sets. Accuracy metrics revealed minor differences in training for some but significant error increases between training and testing for most models. This indicates overfitting and rendered the models poorly suited for forecasting. Ultimately the team did not identify a reliable model to predict that a change in primary energy sources from highly developed countries affects less developed countries.

This conclusion extended to supporting questions, including the influence of clean energy on a country's GDP and using forecasts for clean energy investment decisions. The use of the existing models to produce forecasts resulted in no viable models that can adequately capture true patterns with a high degree of confidence. To enhance forecast robustness and reduce overfitting, a larger time series dataset is necessary such as including monthly or weekly points. The team shifted to treating clean energy capacity as a country's "stock" price, yielding a final "compound return" for each country and offering a different perspective on country growth and investment potential. This analysis addressed a secondary research question by creating a foundation for future work that may suit different investor types or business studies.