# Group 4 Progress Report

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

## Background

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 the following:

1. Will an evolution in the type of energy production impact a country's economic health?
2. Can energy production by source, overall consumption, and GDP be projected through the year 2050?

The key measurement used to rank countries in terms of development is the Human Development Index. According to the United Nations Development Program, "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] Countries are classified into four groups according to their HDI: "low human development" (Low), "medium human development" (Medium), "high human development" (High), and "very high human development" (Very High).[2] Generation and capacity will be used to quantify energy production, where generation represents the actual quantity of energy produced in terawatt-hours (TWh), while capacity is the maximum theoretical power output of all installed sources in gigawatts (GW). 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 value of all goods and services produced in a country for the given year, reported in 2017 international dollars.

## Data Preprocessing

The primary goal of the data preprocessing step was to transform each of the three identified data sources into a single table to be used in the modeling stage of the project. Prior to combining the data into a single table, each data source was investigated in detail to determine what data to keep from each source and what keys could be used between tables to join the data into a cohesive single source. The process was developed using separate R

[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.

scripts, and then combined into a single R script for creation and exportation of the final dataset to be used in subsequent modeling steps.

After looking across each dataset for a primary key, it was determined that each observation contained a three character country code (ISO Standard 3166[3]) that could be combined with the year as a unique identifier. The first dataset investigated for cleaning was the energy dataset. As the primary data source, the data points 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. As the project transitions into the modeling phase, decisions will be made on how to treat the incomplete data appropriately. The total observation count is 4,688 rows across 214 countries for years between 2000 and 2021. An example of the fully cleaned and joined dataset is presented in Figure 1.

| CountryCode | Year | Country.all | Region | CapClean | CapFossil | CapTotal | GenClean | GenFossil | GenTotal | Demand | Import | GDPPerCapita | HDI | HDICode |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AFG | 2000 | Afghanistan | SA | 0.19 | 0.03 | 0.22 | 0.31 | 0.16 | 0.47 | 0.57 | 0.1 | NA | 0.34 | Low |
| AFG | 2001 | Afghanistan | SA | 0.19 | 0.03 | 0.22 | 0.5 | 0.09 | 0.59 | 0.69 | 0.1 | NA | 0.34 | Low |
| AFG | 2002 | Afghanistan | SA | 0.19 | 0.03 | 0.22 | 0.56 | 0.13 | 0.69 | 0.79 | 0.1 | 1280.4631 | 0.36 | Low |
| AFG | 2003 | Afghanistan | SA | 0.19 | 0.04 | 0.23 | 0.63 | 0.31 | 0.94 | 1.04 | 0.1 | 1292.3335 | 0.38 | Low |

Figure 1: Sample of unified dataset

## Exploration and Preliminary Findings

### *Exploratory Visualizations*

Due to the scope of the problem statement and size of the dataset, a significant amount of time was spent performing exploratory visualizations. These visualizations assisted with narrowing down the overall and supporting statements. The following visuals and descriptions discuss significant takeaways from the initial visualizations stage.

Clean energy generation was investigated first to observe the real global trend in fossil vs clean energy output. Figure 2 appears to show increased momentum in clean energy production, where the slope of the trend line for *GenClean* (actual clean power in TWh) steepens in recent years. An additional observation is that generation of fossil fuels seems to be slowing down overall. The annual percent change in generation by type is shown in Figure 3 and shows 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.

The next section explores how clean energy production is dispersed among developing and developed nations.

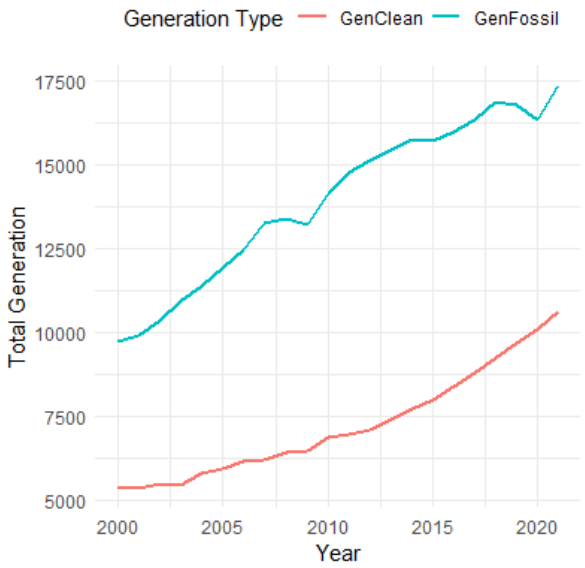


Figure 2: Clean and fossil fuel generation by year

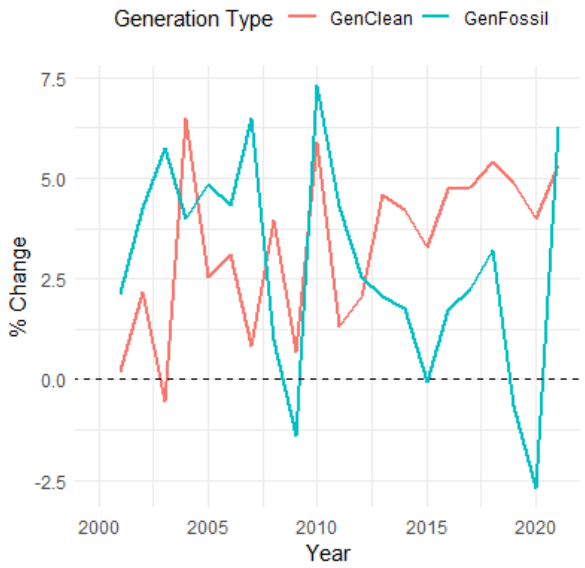


Figure 3: Percent change by generation source per year

*Differences Between HDI Groups*

From the cleaned dataset, a column of data referred to from here on out as *CapCleanProportion* was produced. This is the ratio of clean energy capacity to total capacity for a country in a given year. In this process, capacity was looked at, rather than generation, for a representation of commitment and capability.

After removing missing data, it was found that the annual means for *CapCleanProportion* were 0.401, 0.456, 0.265, and 0.391 for the Low, Medium, High, and Very High HDI groups, respectively. Interestingly, the annual mean was higher in the Low and Medium groups than for High and Very High. Then the focus was narrowed to the last year on record (2021) to determine if similar patterns could be detected. The means of *CapCleanProportion* in 2021 were 0.398, 0.445, 0.318, and 0.477 for the Low, Medium, High, and Very High groups, respectively. It appears that in 2021 the behavior was not drastically different, but overall, countries in the High and Very High groups had higher shares of clean capacity than the corresponding historical average. Figure 4 shows boxplots of *CapCleanProportion* in the year 2021 for each of the four groups (also referred to here and throughout as *HDICode*). These plots exhibit long whiskers and wide interquartile ranges, which indicate significant numbers of observations near the extremes of each group. These within-group distributions were examined and will be discussed later in this section.

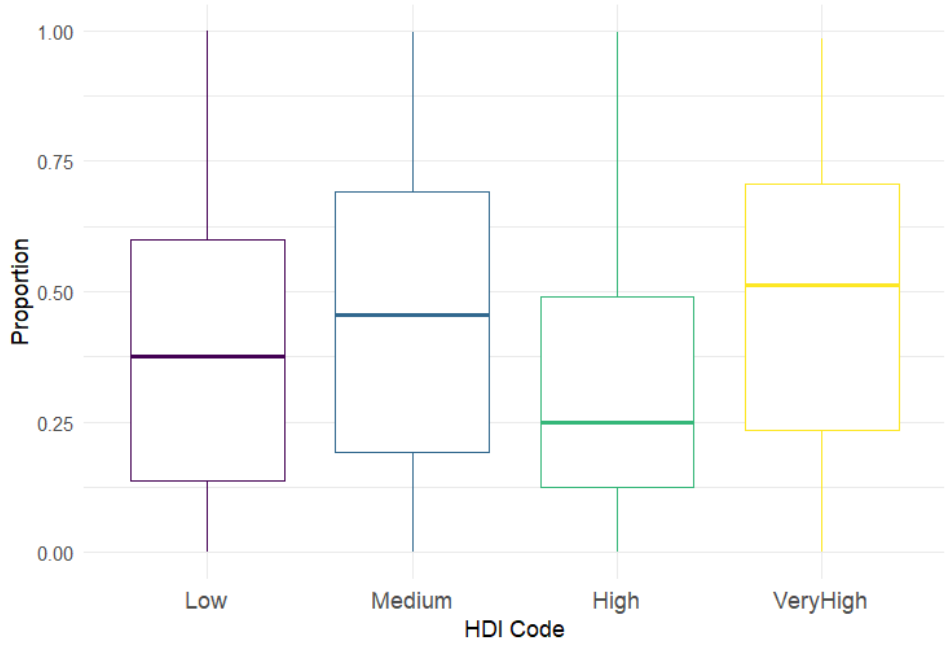


Figure 4: Proportion of clean energy capacity by HDI group

The question of difference in *CapCleanProportion* group means was considered next. An ANOVA test was performed with 2021 *CapCleanProportion* as the dependent variable, and *HDICode* as the independent variable. The results in Figure 5 show that *CapCleanProportion* means between HDI Groups are significantly different at significance level $\alpha = .05$. A subsequent Tukey's Multiple Comparison Test yielded that the means of *CapCleanProportion* were different for the Very High and High groups.

```
             Df Sum Sq Mean Sq F value Pr(>F)
HDICode       3   0.74 0.24680   2.774 0.0428 *
Residuals   180  16.01 0.08896
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 5: ANOVA results for *CapCleanProportion* explained by *HDICode*

The residual plot in Figure 6 demonstrates light tails, which likely indicates that the distribution is not normal. While this may not affect the conclusion of the ANOVA test, the Shapiro-Wilk test was conducted on each group to check for normality. The *p*-value of each HDI group test was less than .05, suggesting that none of the *CapCleanProportion* distributions were normal. This is further illustrated by the density curves in Figure 7 when examining the distributions by group.

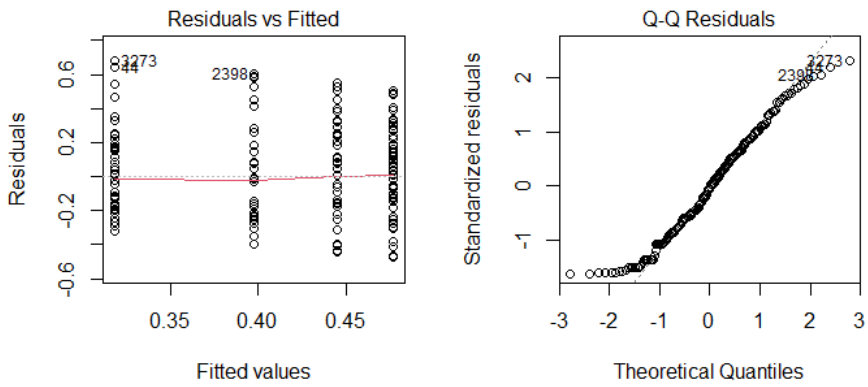


Figure 6: Diagnostic residual plots for ANOVA results in Figure 5

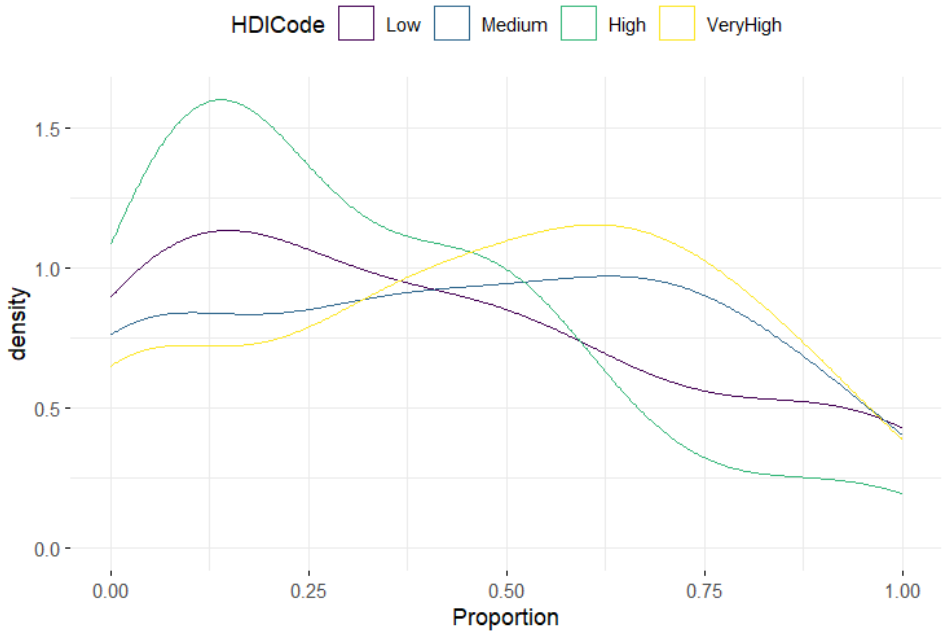


Figure 7: Distribution of *CapCleanProportion* by HDI Group

Due to the likelihood of non-normality, analogous nonparametric tests were conducted. The Kruskal-Wallis test results in Figure 8 showed again that *CapCleanProportion* group means are significantly different. A Pairwise Wilcoxon Rank Sum Test also concluded that the means of *CapCleanProportion* were different for the Very High and High groups.

```
        Kruskal-Wallis rank sum test

data:  CapCleanProportion by HDICode
Kruskal-Wallis chi-squared = 8.5472, df = 3, p-value = 0.03596
```

Figure 8: Kruskal-Wallis Test results for *CapCleanProportion* explained by HDI Code

Next, trends of clean energy capacity by HDI group over time will be evaluated. The plot of *CapCleanProportion* by year and HDI code in Figure 9 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.

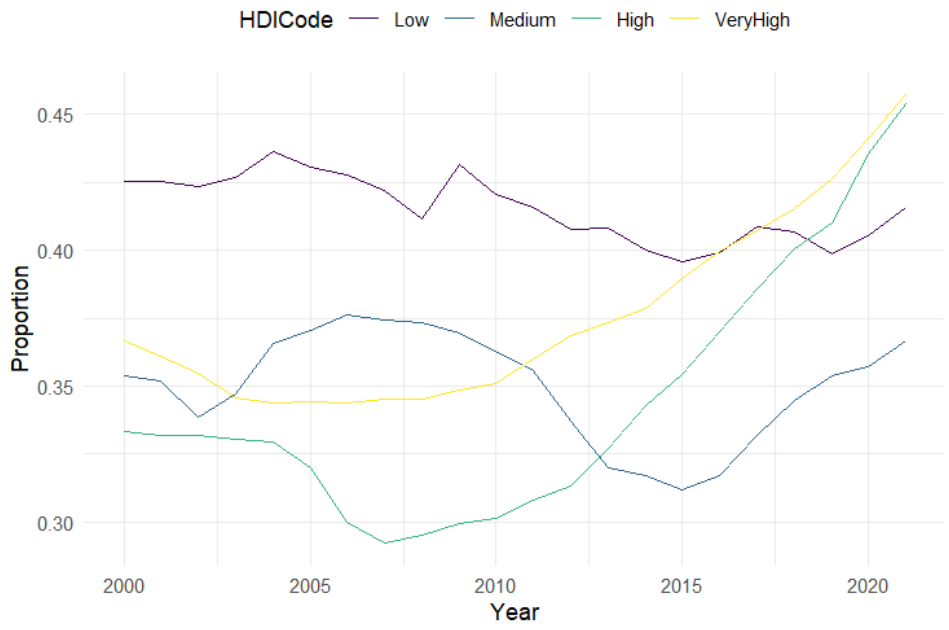


Figure 9: *CapCleanProportion* by year (2000 – 2021) and HDI Group

In the last 10 years of observations, the trend for Medium HDI countries appears to be more upward, but Low HDI countries appear to have exhibited no significant change.

## Next Steps

After analysis of clean energy capacity by HDI group, a decision on how to choose groups to represent the developed and less-developed nations on a binary basis will need to be made. One significant discovery that was made is that there is no evident trend in clean energy share for low-HDI nations. Therefore, less-developed countries may have to be represented by the medium HDI group. The team will also explore if trends become more significant when low and medium HDI are aggregated as a single representation of less-developed countries. For developed countries, trends may be analyzed with the high and very high HDI observations combined under a single “high-development” group.

These steps will enable modeling that will address the primary business question. As stated in the Background section, the goal will be to determine the relationship between developed and less-developed countries’ clean energy production. Research has begun into using linear regression or autoregression with exogenous variables and lag.[4] It may also be

[4] Christoph Hanck et al., *Introduction to Econometrics with R*, 2023, chap. 14, sec. 5, https://www.econometrics-with-r.org/.

found useful to perform multiple regression analysis to explore which other factors can predict clean energy capacity for less-developed nations. Given the datasets available, membership in economic or political unions, a country's continent, or other geographic features are available for consideration.

Looking forward to the supporting research questions, the team wishes to discover the effect of changing clean energy production share on a country's GDP. Regression analysis usage is anticipated to describe what, if any, relationship exists. Finally, the viability of forecasting clean energy production and GDP into the future will be explored using techniques like ARIMA or VAR.[5] The goal for these secondary analyses is that they provide valuable insights into the effect of clean energy transition on economic growth for less-developed countries.

To date, the project has progressed in accordance with the Gantt chart submitted in the project proposal and updated as shown in Figure 10. In the following weeks, the team will implement, adapt, and refine the models listed above and have a deeper understanding of the primary and supporting business questions.

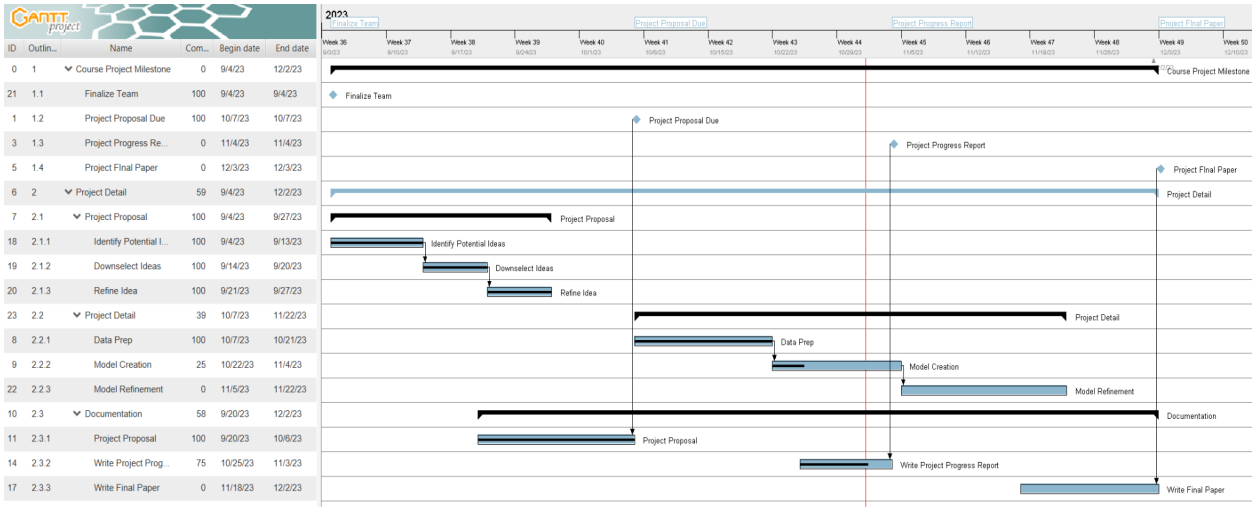


Figure 10: Team 4 Project Gantt Chart

[5] Hanck et al., *Introduction to Econometrics with R*, chap. 16, sec 1.