Exploratory Data Analysis Report

# Team Lead: Melissa Best

# Recorder: Kevin Puduseril

# Spokesperson: Maxwell Tuttle

# Recap Background and Question

Our research question is:

What combination of state characteristics (political leaning, grid capacity, median income, existing energy mix, etc.) and renewable energy policy designs (e.g., Renewable Portfolio Standard targets, subsidy types) best predict a successful change in fossil fuel shares per state in the U.S.?

## Hypothesis:

States with liberal political climates, robust grid capacity, higher median incomes, and ambitious, well-designed renewable energy policies will demonstrate greater success in reducing fossil fuel energy shares.

## Prediction:

A machine learning model will show these features as significantly and positively correlated with fossil fuel reduction over the past decade.

# 

# Methods

## Datasets:

We used multiple datasets, as outlined in our proposal

* U.S. EIA Net Generation by State by Type of Producer by Energy Source (1990–2023):
  + [Energy Generation Data](https://www.eia.gov/electricity/data/state/)
  + Annual electricity generation by source (coal, natural gas, wind, solar, etc.).
  + Format: Excel (XLS) files.
* U.S. EIA State Energy Data System (SEDS):
  + [Energy Consumption Data](https://www.eia.gov/state/seds/seds-data-complete.php)
  + Includes energy consumption and production data by fuel type and state.
  + Format: Excel files (use\_all\_btu.xlsx and Codes\_and\_Descriptions.xlsx)
* DSIRE Database of State Incentives for Renewables & Efficiency:
  + State-level policy information (e.g., RPS presence, subsidy types)
  + Copied the information from [State Policy Map](https://ncsolarcen-prod.s3.amazonaws.com/wp-content/uploads/2025/04/RPS-CES-April2025.pdf) into a .csv file.
  + Copied the list of all policies by state from [State Energy Programs](https://programs.dsireusa.org/system/program).
* U.S. Census Bureau - American Community Survey (ACS):
  + State socioeconomic characteristics (median household income, education levels).
  + [S1901: Income in the Past 12 Months ... - Census Bureau Table](https://data.census.gov/table/ACSST1Y2023.S1901?q=median+household+income+by+state)
* MIT Election Data and Science Lab:
  + State-level political leaning and voting data (1976-2020)
  + [Presidential Elections](https://dataverse.harvard.edu/file.xhtml?fileId=10244938&version=8.0&toolType=PREVIEW)
  + [Senate Elections](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/PEJ5QU)

## Data Preprocessing:

* We merged datasets using state identifiers.
  + Median Income and Energy Generation
  + Median Income and Presidential Election Data
  + Median Income and Senate Election Data
  + U.S. EIA State Energy Data System and Presidential Election Data
  + U.S. EIA State Energy Data System, Presidential Election Data, DSIRE (Database of State Incentives for Renewables & Efficiency), and Median Income
* Grouped U.S. EIA State Energy Data System (SEDS) by Fossil Fuels and Renewable Energy:
  + Fossil Fuels have MSN Codes beginning with:
    - fossil\_prefixes = {"NG", "CL", "CO", "PA", "PC", "DF", "JF", "FF"}
  + Renewable Energy Sources have MSN Codes beginning with:
    - renewable\_prefixes = {"HY", "WD", "WS", "SO", "GE", "WY"}
* The outcome variable, % reduction in fossil fuel share, was calculated as:
  + delta\_fossil\_share = fossil\_share\_2010 - fossil\_share\_2022
* A binary “success” indicator was created, where 1 = state reduced fossil fuel share more than the national median.
* Policy variables (e.g., RPS strength, subsidy presence) were encoded as binary or numeric.

## EDA Methods:

We used the following approaches:

* Descriptive Statistics: Measures of central tendency and spread (mean, median, standard deviation).
* Categorical Analysis: Frequency and proportion tables for policy-related variables
* Visualization:
  + Fossil fuel reduction distribution (histogram).
  + Fossil vs. renewable generation over time (line chart).
  + Fossil fuel reduction by political leaning (Box plot).
  + Top states with highest renewable energy growth (Bar chart).
  + Energy Generation by Source over Time (line chart).
  + Energy Generation by Median Income Quartile (Bar chart).
  + Median Income and Total Energy Generation (Scatterplot).
  + Policy Creation by Year (Histogram).

# 

# Results

## Table 1

Summary of continuous variables (mean, median, standard deviation).

Table 1a: Descriptive Statistics

| **Variable** | **Count** | **Mean** | **Std Dev** | **Median** | **Min** | **Max** | **Skew** | **Kurtosis** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **fossil\_share\_2010** | 52 | 67.74 | 24.99 | 69.51 | 0.00 | 99.03 | -0.86 | 0.51 |
| **fossil\_share\_2022** | 52 | 61.66 | 24.27 | 62.51 | 0.00 | 97.77 | -0.68 | 0.17 |
| **delta\_fossil\_share** | 52 | 6.08 | 8.47 | 5.73 | -18.45 | 23.65 | -0.17 | 0.59 |
| **median\_income** | 51 | 74,445 | 12,307 | 71,970 | 52,719 | 101,027 | 0.32 | -0.74 |

Context:

Table 1a: Descriptive Statistics provides an overview of the key continuous variables in the dataset, including fossil fuel shares in 2010 and 2022, the change in fossil fuel share (delta\_fossil\_share), and median household income. These statistics summarize the central tendency, spread, and distribution shape for each variable across all U.S. states.

Interpretation:

* Fossil Share 2010 (Mean = 67.74%): On average, about two-thirds of states’ electricity generation came from fossil fuels in 2010, though this share varies widely (0% to 99%), indicating significant differences in state energy mixes.
* Fossil Share 2022 (Mean = 61.66%): By 2022, fossil fuel dependency dropped by ~6 percentage points on average, reflecting a national shift toward renewable or cleaner sources.
* Fossil Share (Mean = +6.08%): The positive mean indicates an average reduction in fossil fuel share across states, but the range (-18.45% to +23.65%) shows some states increased fossil reliance while others drastically reduced it.
* Median Income (Mean = $74,445): States vary in median household income, which may influence energy policy effectiveness, as wealthier states often invest more in clean energy transitions.

Table 1b: By Success

| **Success Group** | **Fossil Share 2010 (Mean)** | **Fossil Share 2022 (Mean)** | **Δ Fossil Share (Mean)** | **Median Income (Mean)** |
| --- | --- | --- | --- | --- |
| **0** (below median reduction) | 67.40% | 67.88% | -0.48% | $74,946 |
| **1** (above median reduction) | 68.09% | 55.44% | +12.64% | $73,925 |

Context:

This table divides states into two groups based on their success in reducing fossil fuel share between 2010 and 2022. States with reductions greater than the median change are labeled as “1 (above median reduction)”, and those with smaller or negative changes are “0 (below median reduction)”.

Interpretation:

* Successful States (Group 1): These states reduced fossil fuel shares significantly (average reduction of 12.64 percentage points), showing strong policy shifts or investments in renewables. Their starting fossil fuel share in 2010 (68.09%) was slightly higher than less successful states but dropped to 55.44% by 2022.
* Less Successful States (Group 0): These states had a slight increase or negligible change in fossil fuel shares (-0.48% on average), remaining heavily dependent on fossil sources (67.88% in 2022).
* Income Differences: Median incomes between the two groups are relatively similar, suggesting that income alone may not fully explain fossil fuel reduction success. Other factors like political climate or policy design could be more influential.

## 

## Table 2

Frequency and percentage of states by Renewable Portfolio Standards (RPS) presence and political leaning.

## **Table 2a: Overall Distribution of Categorical Variables**

| **variable** | **level** | **percent** | **count** |
| --- | --- | --- | --- |
| Y\_OUTCOME | 0 | 50.980392 | 26 |
| Y\_OUTCOME | 1 | 49.019608 | 25 |
| rps\_present | 1.0 | 58.823529 | 30 |
| rps\_present | NaN | 27.450980 | 14 |
| rps\_present | 0.0 | 13.725490 | 7 |
| rps\_target\_quartile | NaN | 27.450980 | 14 |
| rps\_target\_quartile | Q1 | 25.490196 | 13 |
| rps\_target\_quartile | Q3 | 23.529412 | 12 |
| rps\_target\_quartile | Q4 | 11.764706 | 6 |
| rps\_target\_quartile | Q2 | 11.764706 | 6 |
| political\_bucket | Conservative-leaning | 52.941176 | 27 |
| political\_bucket | Liberal-leaning | 47.058824 | 24 |
| income\_quartile | Q4 | 25.490196 | 13 |
| income\_quartile | Q1 | 25.490196 | 13 |
| income\_quartile | Q3 | 23.529412 | 12 |
| income\_quartile | Q2 | 23.529412 | 12 |
| income\_quartile | NaN | 1.960784 | 1 |

Table 2a provides an overview of categorical variables that describe the policy, political, and socioeconomic landscape of U.S. states in the context of energy transitions. It reveals that states are almost evenly split between successful and unsuccessful outcomes (Y\_OUTCOME), indicating that progress in reducing fossil fuel dependency is not concentrated in a specific subset of states. Many states do not have Renewable Portfolio Standards (rps\_present), highlighting gaps in policy adoption that could influence clean energy performance.

Political leaning is fairly balanced between liberal- and conservative- leaning states, suggesting that energy transition success cannot be attributed solely to political ideology. Additionally, the even distribution across income quartiles shows that both wealthier and less affluent states are represented, implying that economic capacity is not the only driver of success. Together, these findings suggest that a complex mix of factors—beyond just income or political orientation—affects energy policy outcomes and fossil fuel reduction progress.

## 

## **Table 2b: Categorical Variables by Success**

| **variable** | **level** | **outcome** | **count** | **percent** | **p\_value** |
| --- | --- | --- | --- | --- | --- |
| rps\_present | NaN | 0 | 8 | 30.769231 | 0.357965 |
| rps\_present | 0.0 | 0 | 5 | 19.230769 | 0.357965 |
| rps\_present | 1.0 | 0 | 13 | 50.000000 | 0.357965 |
| rps\_present | NaN | 1 | 6 | 24.000000 | 0.357965 |
| rps\_present | 0.0 | 1 | 2 | 8.000000 | 0.357965 |
| rps\_present | 1.0 | 1 | 17 | 68.000000 | 0.357965 |
| rps\_target\_quartile | Q1 | 0 | 8 | 30.769231 | 0.644408 |
| rps\_target\_quartile | Q2 | 0 | 2 | 7.692308 | 0.644408 |
| rps\_target\_quartile | Q3 | 0 | 5 | 19.230769 | 0.644408 |
| rps\_target\_quartile | Q4 | 0 | 3 | 11.538462 | 0.644408 |
| rps\_target\_quartile | NaN | 0 | 8 | 30.769231 | 0.644408 |
| rps\_target\_quartile | Q1 | 1 | 5 | 20.000000 | 0.644408 |
| rps\_target\_quartile | Q2 | 1 | 4 | 16.000000 | 0.644408 |
| rps\_target\_quartile | Q3 | 1 | 7 | 28.000000 | 0.644408 |
| rps\_target\_quartile | Q4 | 1 | 3 | 12.000000 | 0.644408 |
| rps\_target\_quartile | NaN | 1 | 6 | 24.000000 | 0.644408 |
| political\_bucket | Conservative-leaning | 0 | 13 | 50.000000 | 0.881908 |
| political\_bucket | Liberal-leaning | 0 | 13 | 50.000000 | 0.881908 |
| political\_bucket | Conservative-leaning | 1 | 14 | 56.000000 | 0.881908 |
| political\_bucket | Liberal-leaning | 1 | 11 | 44.000000 | 0.881908 |
| income\_quartile | Q1 | 0 | 8 | 30.769231 | 0.697142 |
| income\_quartile | Q2 | 0 | 5 | 19.230769 | 0.697142 |
| income\_quartile | Q3 | 0 | 5 | 19.230769 | 0.697142 |
| income\_quartile | Q4 | 0 | 7 | 26.923077 | 0.697142 |
| income\_quartile | NaN | 0 | 1 | 3.846154 | 0.697142 |
| income\_quartile | Q1 | 1 | 5 | 20.000000 | 0.697142 |
| income\_quartile | Q2 | 1 | 7 | 28.000000 | 0.697142 |
| income\_quartile | Q3 | 1 | 7 | 28.000000 | 0.697142 |
| income\_quartile | Q4 | 1 | 6 | 24.000000 | 0.697142 |
| income\_quartile | NaN | 1 | 0 | 0.000000 | 0.697142 |

Table 2b examines the relationship between categorical variables—such as RPS presence, political leaning, and income quartiles—and energy transition success (Y\_OUTCOME). The data shows that states with and without RPS policies are similarly distributed between successful and unsuccessful outcomes, suggesting that simply having an RPS does not guarantee success.

Political leaning (liberal vs. conservative) and income quartiles (Q1-Q4) also show nearly equal representation across both groups, with high p-values indicating no statistically significant difference. These findings suggest that categorical factors alone do not strongly influence success and that other continuous measures, like fossil fuel share reduction, may better explain variations in performance.

## 

## **Table 2c: Continuous Variables by Success**

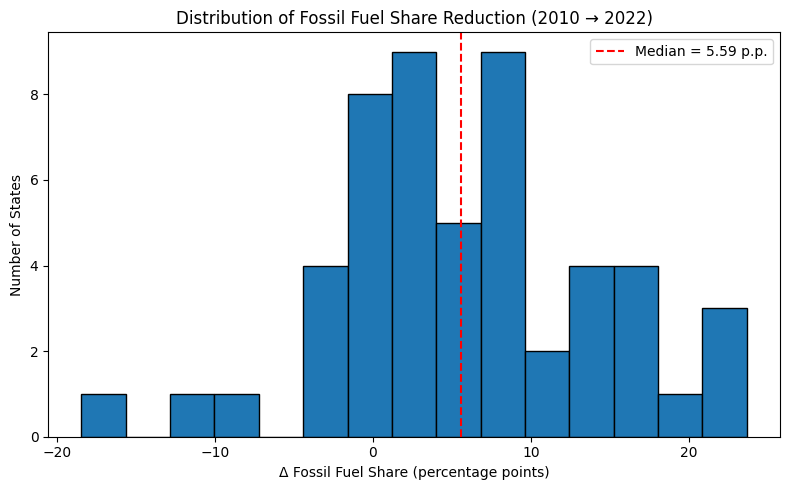
|  | **Y\_OUTCOME** | |
| --- | --- | --- |
| **Metric** | **0** | **1** |
| fossil\_share\_2010\_count | 26.00 | 25.00 |
| fossil\_share\_2010\_mean | 67.40 | 67.95 |
| fossil\_share\_2010\_std | 28.06 | 22.48 |
| fossil\_share\_2010\_median | 71.49 | 65.72 |
| fossil\_share\_2010\_min | 0.00 | 16.43 |
| fossil\_share\_2010\_max | 99.03 | 97.74 |
| fossil\_share\_2022\_count | 26.00 | 25.00 |
| fossil\_share\_2022\_mean | 67.88 | 55.08 |
| fossil\_share\_2022\_std | 26.05 | 21.38 |
| fossil\_share\_2022\_median | 72.72 | 56.38 |
| fossil\_share\_2022\_min | 0.00 | 9.43 |
| fossil\_share\_2022\_max | 97.77 | 90.97 |
| delta\_fossil\_share\_count | 26.00 | 25.00 |
| delta\_fossil\_share\_mean | -0.48 | 12.87 |
| delta\_fossil\_share\_std | 5.40 | 5.23 |
| delta\_fossil\_share\_median | 0.37 | 11.76 |
| delta\_fossil\_share\_min | -18.45 | 5.81 |
| delta\_fossil\_share\_max | 5.65 | 23.65 |
| median\_income\_count | 25.00 | 25.00 |
| median\_income\_mean | 73902.80 | 73924.96 |
| median\_income\_std | 13261.50 | 10472.60 |
| median\_income\_median | 71798.00 | 71970.00 |
| median\_income\_min | 52719.00 | 59673.00 |
| median\_income\_max | 96346.00 | 94991.00 |
| political\_score\_count | 25.00 | 25.00 |
| political\_score\_mean | -0.01 | -0.03 |
| political\_score\_std | 0.20 | 0.19 |
| political\_score\_median | -0.00 | -0.02 |
| political\_score\_min | -0.30 | -0.40 |
| political\_score\_max | 0.33 | 0.37 |
| rps\_target\_pct\_count | 13.00 | 17.00 |
| rps\_target\_pct\_mean | 33.96 | 34.69 |
| rps\_target\_pct\_std | 21.65 | 27.32 |
| rps\_target\_pct\_median | 40.00 | 25.20 |
| rps\_target\_pct\_min | 8.50 | 2.00 |
| rps\_target\_pct\_max | 70.00 | 100.00 |

Table 2c highlights key differences in continuous variables between successful and unsuccessful states. The most notable factor is the reduction in fossil fuel share—successful states achieved an average decrease of about 12.9 percentage points, while unsuccessful states showed virtually no improvement.

Median household income and political score are relatively similar across both groups, indicating that socioeconomic status and political leaning alone do not explain success. Additionally, while successful states tend to have higher Renewable Portfolio Standards (RPS) target percentages, the limited coverage of this data suggests that policy enforcement and energy sector shifts are more critical drivers of performance.

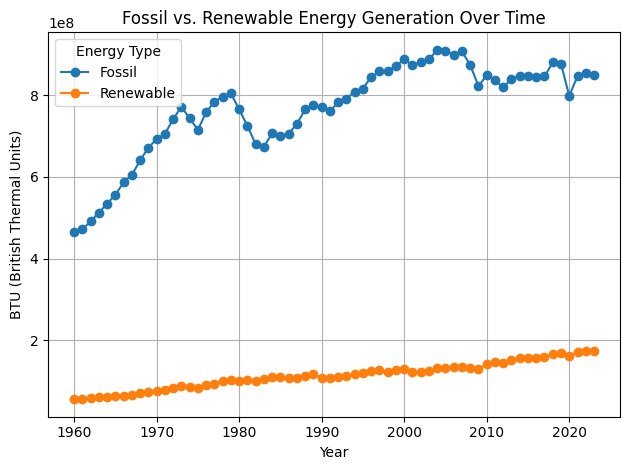
## Figures

### Fossil fuel reduction distribution (histogram).



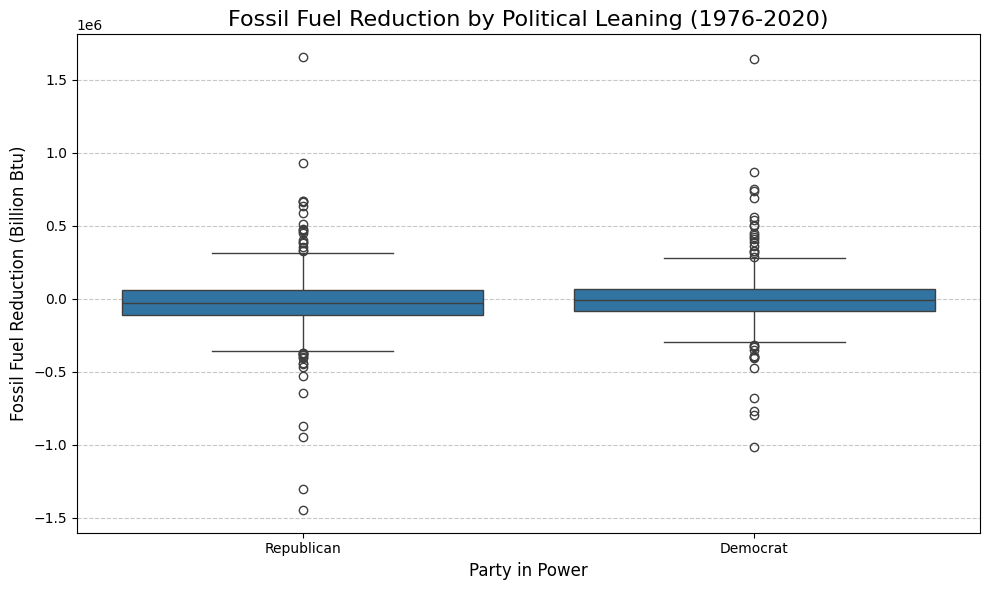
The histogram illustrates how U.S. states have changed their reliance on fossil fuels between 2010 and 2022. Overall, most states achieved a reduction, with a median decrease of approximately 5.6 percentage points, indicating a general shift toward cleaner energy sources. However, the distribution reveals significant variation: while some states reduced their fossil fuel share by more than 20 percentage points, others saw little change or even increases, highlighting uneven progress across the country. This variation likely reflects differences in state energy policies, infrastructure, and resource availability, emphasizing that while the national trend is moving away from fossil fuels, progress is not uniform across all states.

### Fossil vs. renewable generation over time (line chart).



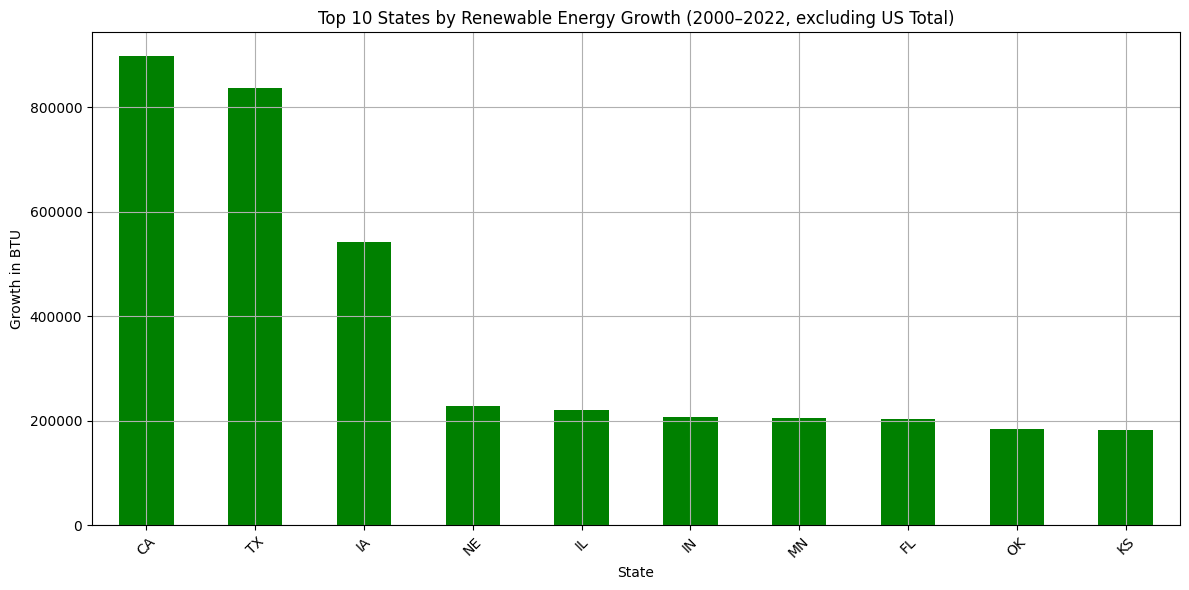
The chart illustrates the historical trend of U.S. energy generation from fossil fuels and renewable sources between 1960 and 2022. Fossil fuels have long dominated energy production, peaking around 2007, but have since plateaued or slightly declined. In contrast, renewable energy started from a low baseline but has shown steady and consistent growth, particularly since the 2000s. This divergence after 2010 signals a gradual energy transition, driven by technological advancements, policy incentives, and growing environmental awareness. While fossil fuels still lead in absolute output, the rise of renewables highlights a shifting energy landscape focused on sustainability.

### Box plot: Fossil fuel reduction by political leaning.



The box plot illustrates the distribution of fossil fuel reduction across U.S. states from 1976 to 2020, categorized by political leaning. Overall, Democratic-leaning states show a slightly higher median reduction in fossil fuel consumption compared to Republican-leaning states, suggesting marginally stronger efforts toward reduction. However, the wide spread and overlapping ranges in both groups indicate high variability, with several outliers reflecting years of significant increases or decreases in fossil fuel use. While the data hints at a modest trend of greater reductions in Democratic states, political affiliation alone does not fully account for state-level energy changes, and additional factors like economic structure, geography, and federal policies likely play key roles.

### Bar chart of top states with highest renewable energy growth.



The bar chart highlights the top 10 U.S. states with the highest growth in renewable energy consumption from 2000 to 2022, excluding the national total. California leads the nation, reflecting its aggressive clean energy policies and investments in solar, wind, and geothermal power. Texas follows closely, driven largely by its massive expansion in wind energy. Iowa ranks third, benefitting from its substantial wind energy infrastructure. Other states like Nebraska, Illinois, and Indiana show moderate but meaningful increases, while Minnesota, Florida, Oklahoma, and Kansas round out the top 10. The chart underscores how both strong policy support and natural resource availability have shaped renewable energy growth across the country.

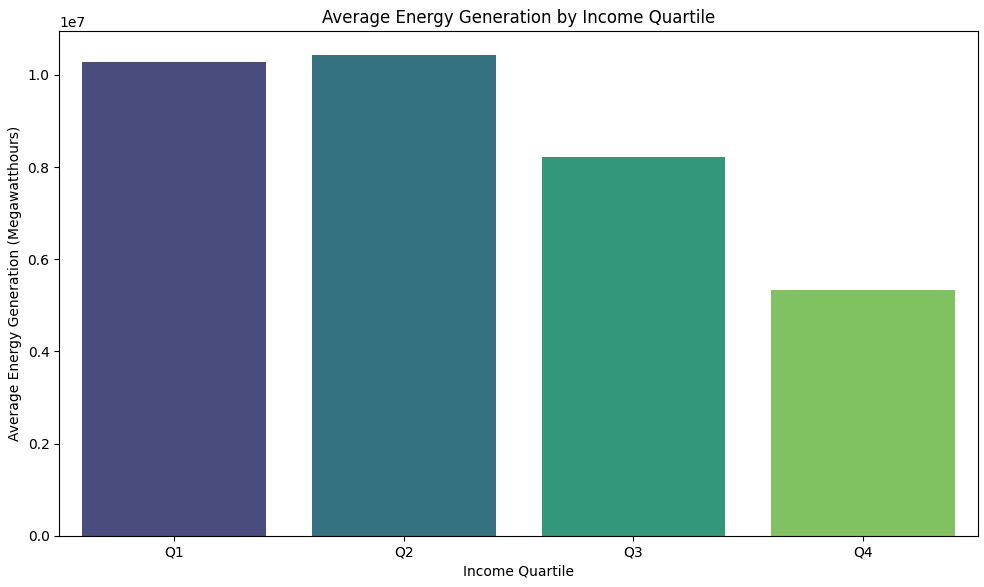
### Energy Generation by Source over Time

## 

The above chart shows the change in energy generation by different sources over the course of 33 years. Some key takeaways are:

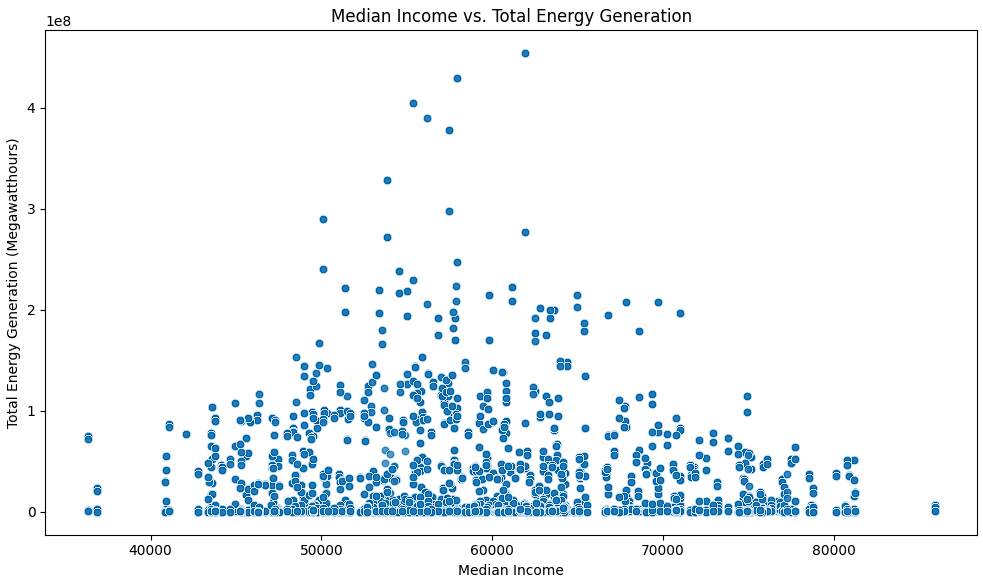
* Coal powered energy generation has drastically decreased, starting roughly around the year 2005
* Natural Gas powered energy generation has continuously increased throughout the observation of the data
* Solar Thermal and Photovoltaic powered energy generation began increasing slowly, starting roughly in 2012
* Wind powered energy generation began steadily increasing starting roughly around the year 2005
* Nuclear powered and hydroelectric energy generation have remained relatively constant throughout the observation of this data

### Energy Generation by Median Income Quartile



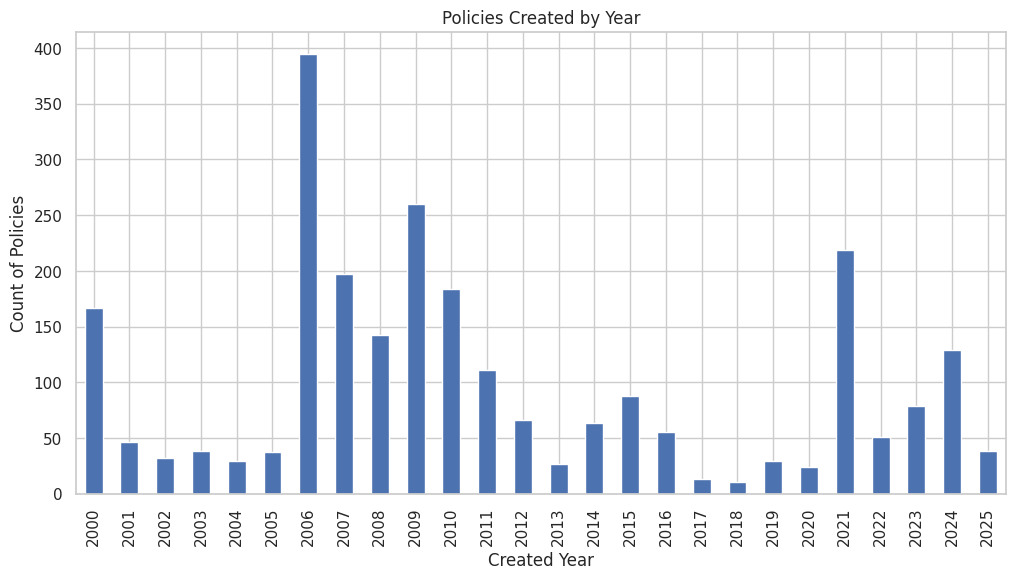
The above chart shows the breakdown of energy generation by median income quartiles, using state level data. One key takeaway is that it appears that the lower quartiles for median income generate more energy than those in the higher median income quartiles. A large explanation for this could be that the leading energy generation source, coal, is likely to exist in lower income areas.

### Median Income and Total Energy Generation (Scatterplot)



The above chart shows a scatterplot of the total energy generation versus the median income for each state for each year. This shows us that the data for these two datasets appears to be relatively normal, following the ideal bell curve.

### Policy Creation by Year (Histogram)



The above Histogram shows the number of clean energy policies created per year according to DSIRE data. It shows large numbers of policies created in 2006, 2009 and 2021.

# 

# Discussion and Next Steps

## Key Takeaways:

Our analysis focused on how energy production trends differ across states and political affiliations, particularly with respect to the transition from fossil fuels to renewable sources. Key takeaways from our study show a consistent national decline in fossil fuel-based energy production over the last two decades, accompanied by varying degrees of renewable energy adoption across states. The energy generation by source figure illustrates this long-term shift, where fossil fuel generation steadily decreased while renewable sources—-especially wind and solar—gained momentum. Notably, our analysis of political affiliation reveals that states leaning Democratic have experienced a sharper reduction in fossil fuel reliance compared to Republican-leaning states. The box plot of fossil fuel reduction by political leaning and Table 2 further underscore disparities in renewable energy growth.

## Next Steps:

Our next steps for this project, the data cleaning and pre-processing plan, will include many things. The first of the bunch will be creating a correlation matrix and potentially some collinearity visualizations on our dataset. We currently have only pieced together a few of the datasets, so further work to combine them all will need to be completed before continuing with this step. We also need to identify and deal with missing data. From our initial findings, there does not appear to be a large amount of missing data, but, when we take a closer look, we will decide whether to remove missing data elements or impute. The next thing we will do will be to perform advanced feature engineering (e.g., interaction terms).

# Appendix - Data Dictionary:

| **Data Source** | **Variables** |
| --- | --- |
| 1976-2020-president | Year |
| Description - presidential election data | State\_long |
|  | State |
|  | office |
|  | party\_detailed |
|  | candidatevotes |
|  | totalvotes |
|  | party\_simplified |
| 1976-2020-senate | Year |
| Description - Senate election data | State\_long |
|  | State |
|  | office |
|  | party\_detailed |
|  | candidatevotes |
|  | totalvotes |
|  | party\_simplified |
| state\_energy\_programs | Name |
| Description - State energy programs and incentives for states including the year it was implemented | State |
|  | Category |
|  | Policy\_Incentive\_Type |
|  | Created |
|  | Last\_Updated |
| median\_income | Year |
| Description - median income by state for 1984-2019 | State |
|  | Income |
| energy\_generation | YEAR |
| Description - energy generation by source for each state 1990-2023 | STATE |
|  | TYPE OF PRODUCER |
|  | ENERGY SOURCE |
|  | GENERATION (Megawatthours) |
| energy\_consumption | Year |
| Description - energy consumption by msn code and state 2000-2023 | State |
|  | FossilFuel |
|  | Renewable |