

Capstone project

Module B: AI for Leaders - Milestone 3

March 9, 2025

Joshua j. mammi francis

Boston University

**Milestone Three**

Joshua J. Mammi Francis

Faculty of Computing & Data Sciences, Boston University

DX699 O2 AI for Leaders (Spring 2025)

Dr. Joshua Von Korff

March 9, 2025

**Milestone Three: Dataset**

**US Energy Generation 2001-2022**

*An Exploratory Data Analysis (EDA) of Renewable Energy Trends*

**Summary**

Exploratory data analysis (EDA) of power generation trends underscores the accelerating shift toward renewable energy in the United States, mirroring a global transition to more sustainable energy practices and reduced dependence on fossil fuels. By 2024, renewables contributed 24% of the country's total electricity generation, marking a notable rise from earlier years and highlighting the expanding role of technologies like solar, wind, and geothermal energy in the U.S. energy portfolio.

This transition is driven by environmental imperatives, advancements in clean energy technologies, shifting regulations, and changing market forces that have enhanced the economic appeal of renewables. Historically dominated by fossil fuels, which still accounted for 60% of electricity generation in 2023, the U.S. energy landscape is seeing a marked transformation, spurred by policies aimed at reducing greenhouse gas emissions and fostering investment in renewables. For instance, between 2014 and 2023, solar energy generation experienced an extraordinary increase of approximately 688%, fueled by government incentives and market demand.

EDA techniques—such as statistical methods, correlation analysis, and visualizations—play a crucial role in extracting insights from power generation data, enabling policymakers, investors, and stakeholders to recognize key trends and challenges. This analytical approach informs strategies for tackling energy efficiency issues, improving infrastructure, and ensuring a smooth transition to a sustainable energy future. However, this push for renewable energy is not without obstacles, including regulatory barriers, infrastructure demands, and the need for equitable workforce development in the clean energy sector, not to mention the changes in government priorities marked by the recent US elections.

To ensure a successful transition, collaboration between governments, industries, and communities is essential. Such efforts will influence the trajectory of renewable energy growth in the United States, shaping its contribution to climate change mitigation and its role in building a resilient, low-carbon economy.

The energy sector in the United States is undergoing a significant transformation, marked by increasing reliance on renewable energy sources and evolving seasonal patterns in power generation. Understanding these trends is critical for policymakers, industry stakeholders, and researchers who aim to address challenges related to energy sustainability, infrastructure, and environmental impact. However, existing approaches often lack a comprehensive view of the dynamics between different energy sources and their contributions over time.

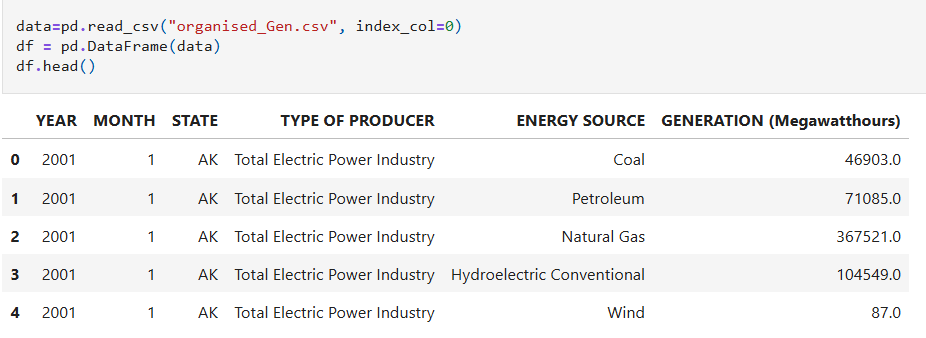
This report addresses the need for a detailed analysis of the U.S. electricity generation landscape by leveraging monthly power generation data, sourced from the U.S. Energy Information Administration (EIA), spanning the years 2001 to May 2022. The analysis focuses on uncovering trends in the growth of renewable energy resources, such as solar, wind, and geothermal, as well as examining seasonal variations in total electricity production.

By extracting meaningful insights from this data, this report aims to establish a foundation for the development of Machine Learning models. These models will enhance the ability to predict and track changes in the energy sector, with a particular emphasis on the types of generation resources by location. The ultimate goal is to provide actionable insights that can support informed decision-making and promote a smoother transition to a sustainable, low-carbon energy future.

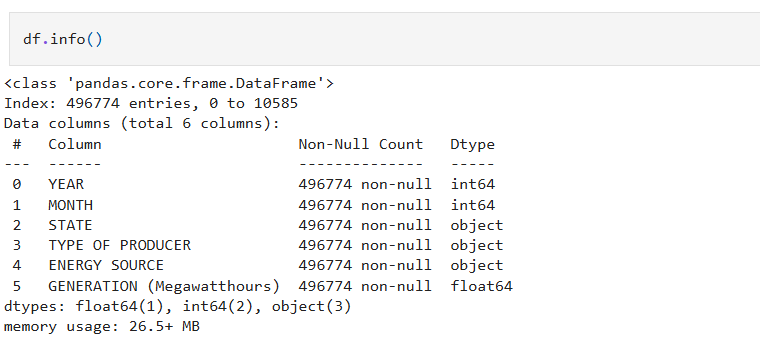
***Link to the dataset: https://www.kaggle.com/datasets/kevinmorgado/us-energy-generation-2001-2022/data***

`

Using Pandas dataframe the data stored in a local drive was imported and transformed into a pandas dataframe but the first column was excluded to prevent having double indices.

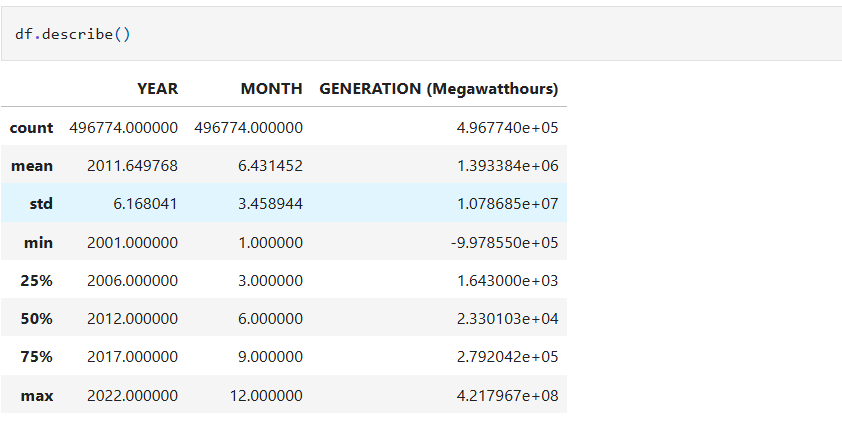


**Examining the dataset**



The exploratory data analysis (EDA) of the energy generation dataset, which contains no null or missing values, provides a robust foundation for extracting meaningful insights without the need for extensive data cleaning or preprocessing. The completeness of the dataset ensures that all records across the variables, including YEAR, MONTH, STATE, TYPE OF PRODUCER, ENERGY SOURCE, and GENERATION (Megawatthours), are intact, enabling a seamless analysis of trends and patterns. This allows the focus to shift toward identifying key insights, such as shifts in energy generation by source, seasonal variations, and regional contributions, without concerns about data gaps introducing potential bias or inaccuracies.

During the EDA, the absence of missing values also simplifies the use of statistical and visualization techniques. Key metrics such as mean, median, and standard deviation of energy generation across states and sources were calculated with confidence. Visualizations such as histogram and KDE plots, bar charts, and heatmaps offered a clear view of trends, particularly the increasing adoption of renewable energy sources over time.



The output of the df.describe() method provides key statistical insights into the dataset, which spans from 2001 to 2022, capturing nearly 500,000 rows of data. The YEAR column has a minimum value of 2001 and a maximum of 2022, indicating full coverage of the 21-year period. The mean year of approximately 2011.65 suggests a relatively even distribution of data across the timeline, with no major gaps or clustering. Similarly, the MONTH column ranges from 1 (January) to 12 (December), highlighting that the dataset comprehensively tracks monthly electricity generation without missing seasonal data.

The GENERATION (Megawatthours) column reveals important insights about the scale and variation of electricity production. The mean value of approximately 1.39 million megawatt-hours indicates substantial energy generation, though the standard deviation of about 10.79 million megawatt-hours underscores considerable variation, likely due to differences in state capacities, energy sources, and seasons. The minimum value of -997,855 megawatt-hours is unexpected and could indicate erroneous data or adjustments for net negative generation (e.g., exports or energy losses). The maximum value of over 421.79 million megawatt-hours suggests periods or regions of exceptionally high production, possibly reflecting large-scale fossil fuel or renewable projects.

The quartile values further illustrate the disparity in generation levels. At the 25th percentile, production is as low as 1,643 megawatt-hours, while the 75th percentile sits at 279,204 megawatt-hours, indicating that most states or periods generate significantly less energy than the largest producers. The median (50th percentile) value of 23,301 megawatt-hours reinforces this skewness, with the majority of data points representing smaller-scale energy production.

In summary, the df.describe() output highlights the dataset's completeness over time and its substantial variation in energy production across regions and months. While the overall trends suggest a robust dataset, outliers (such as the negative generation value) warrant further investigation to ensure data accuracy and reliability for downstream analysis. These insights lay the groundwork for exploring seasonal patterns, regional differences, and the growing role of renewables in energy generation.

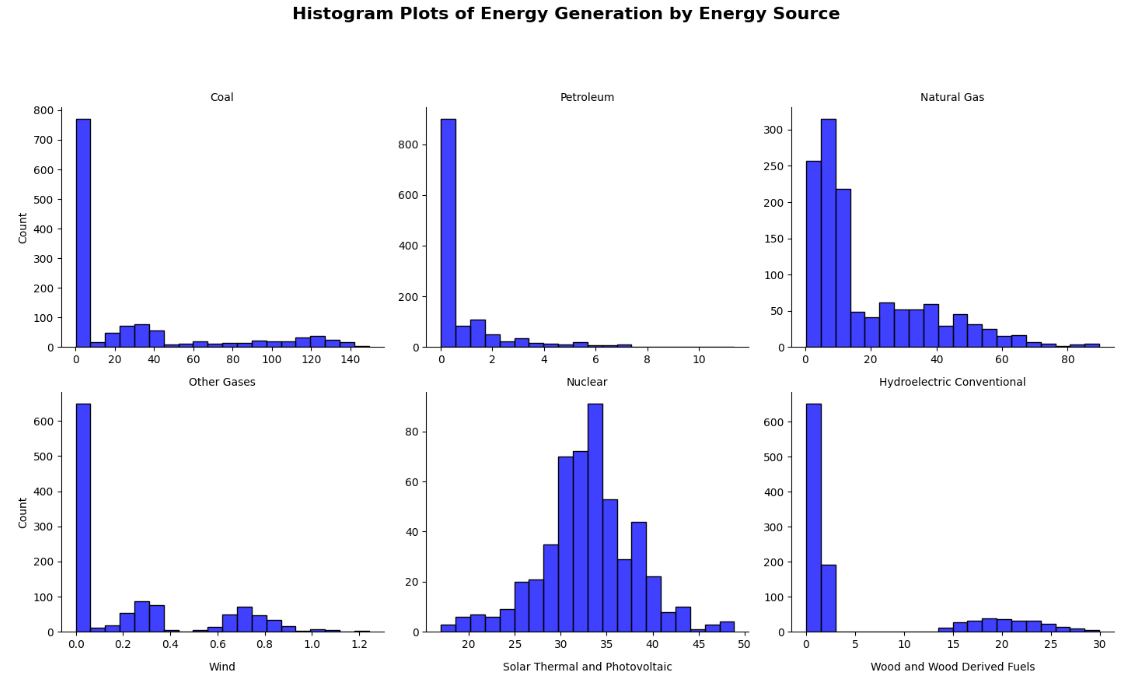
Refined the data frame, by applying a filter to isolate total energy generation, enabling the analysis of patterns in energy sources and types of producers. Entries labeled "Total Electric Power Industry" under TYPE OF PRODUCER and "Total" under ENERGY SOURCE were excluded to ensure accuracy and prevent complications during future groupby calculations.

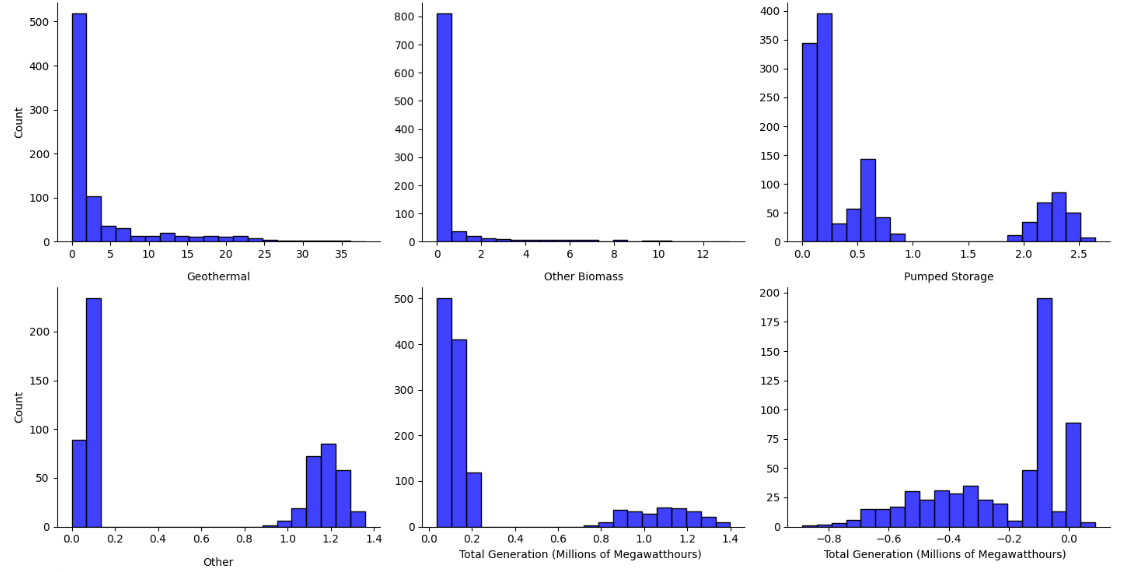
****

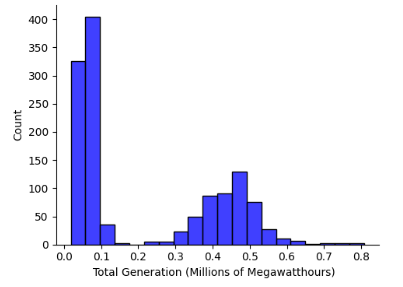
**Univariate analysis**

**Histogram plots**

The histogram plots provide valuable insights into the distribution and variability of energy generation across different categories, such as energy sources, states, or time periods. By visualizing the frequency of energy generation levels, the plots reveal patterns that are otherwise difficult to discern from raw data. For instance, they may highlight that most energy generation values cluster around lower production levels, while a smaller number of high-output instances represent outliers, reflecting major power plants or large-scale projects. Additionally, histograms capture seasonal fluctuations or disparities in energy output between renewable and non-renewable sources, showing conventional or non-renewable distributions to be more shewed and only the renewable distribution shows some symmetry in its distribution centered around 33 million MegaWatts. This visual tool not only aids in identifying the central tendency and spread of the data but also emphasizes the diversity within the energy sector, which can inform targeted strategies for infrastructure development and resource allocation.

****

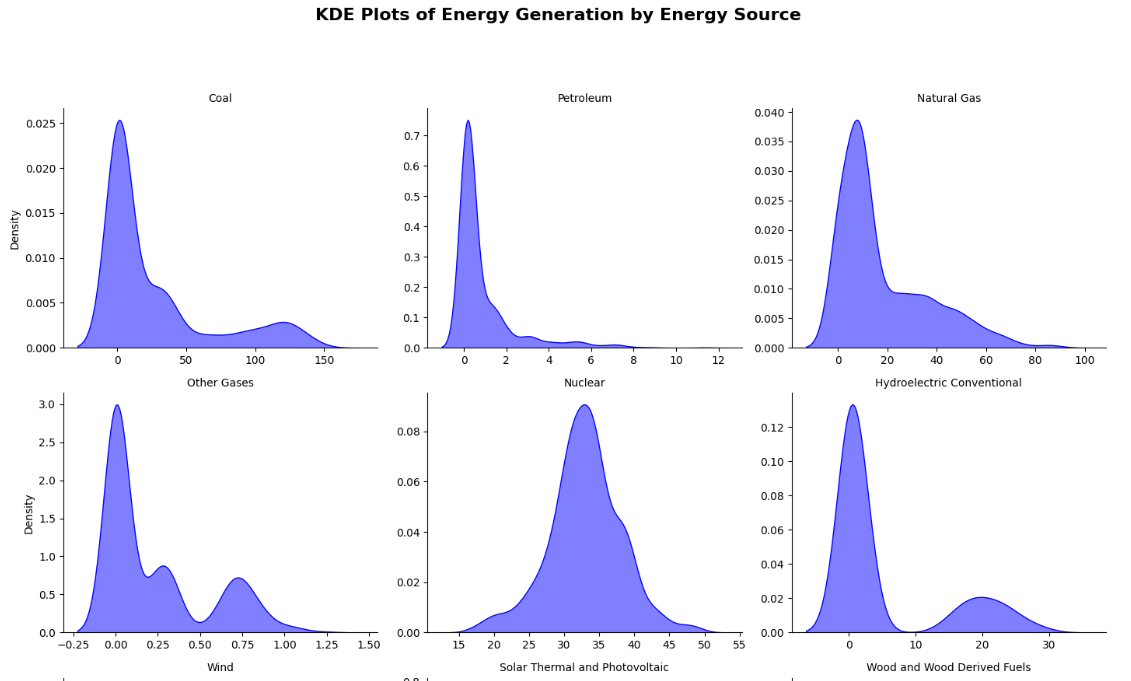
****

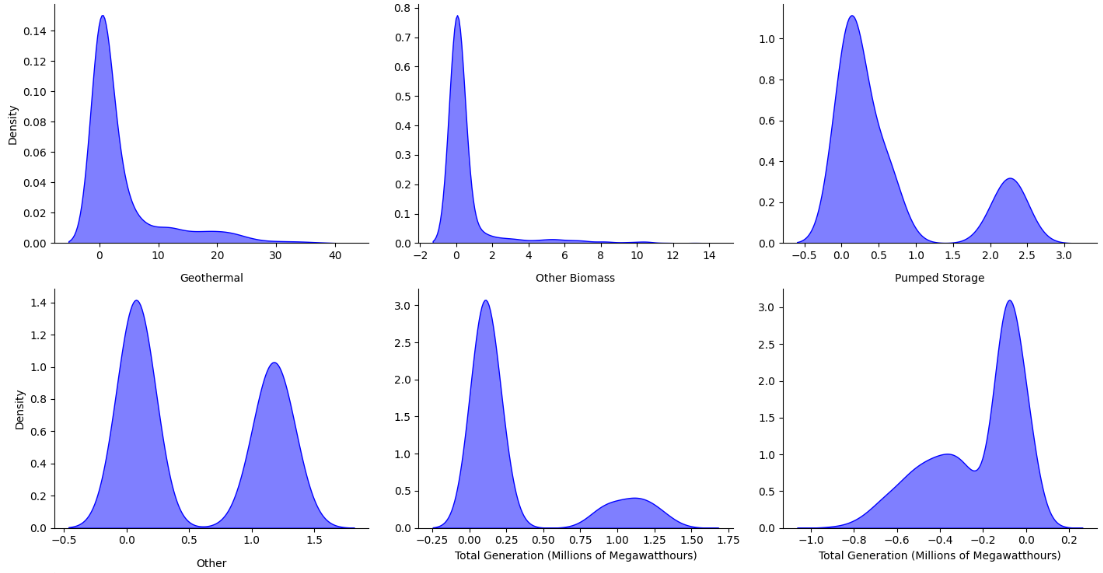
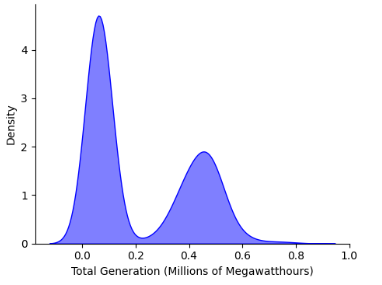
****

**KDE Plots**

The KDE plots of the data further

The Kernel Density Estimate (KDE) plots provide a smooth representation of the distribution of energy generation values within the dataset, offering insights into the concentration and spread of data for various categories such as energy sources, states, or time periods. Unlike histograms, KDE plots show continuous curves, further demonstrating the skewness of the conventional forms of energy production and the symmetry of the Solar, wing and geothermal sources of energy production concentrated around 30 to 35 million MegaWatts of energy.

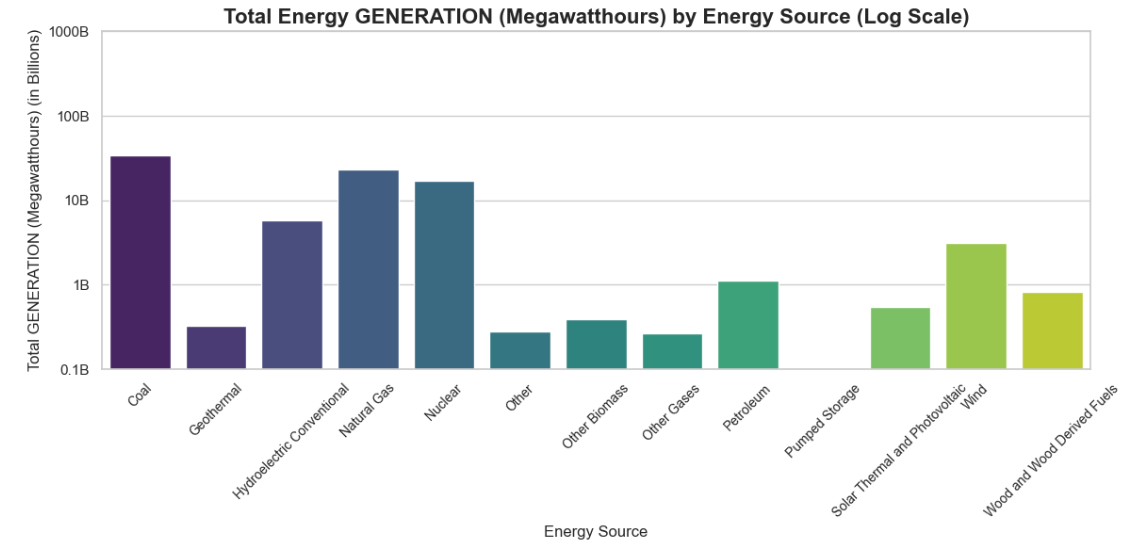


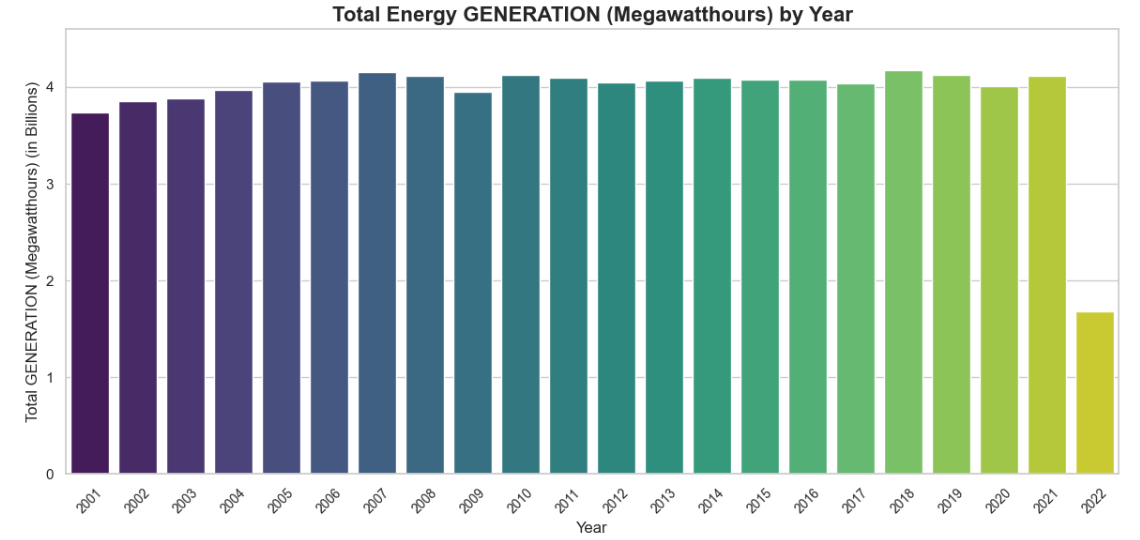
**Bi-variate analyses**

**Barcharts**

A bivariate analysis of the dataset using a bar chart, with energy sources on the horizontal axis and energy generation amounts on the vertical axis, provides a clear comparison of how much electricity each source contributes. This visualization highlights the differences in energy output between renewable and non-renewable sources, making it easy to see which types of energy dominate the generation landscape. For example, fossil fuels such as coal and natural gas show significantly higher bars, reflecting their historical dominance in total power generation. Meanwhile, renewable sources like wind, solar, and hydroelectric power reveal their growing contributions through increasing bar heights.

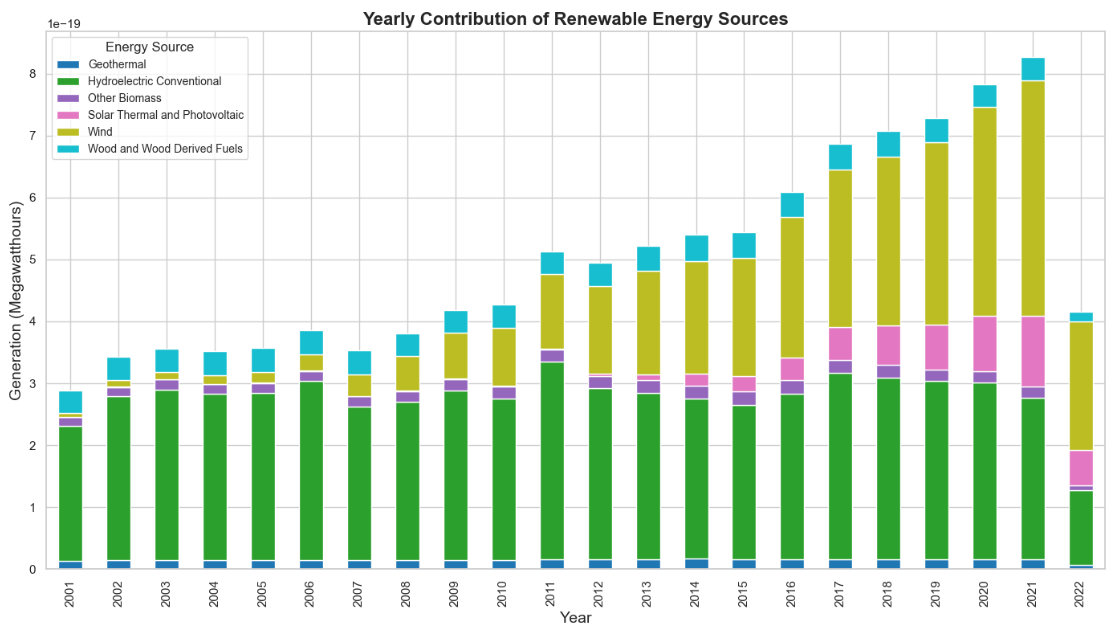


Another bivariate analysis of the dataset using a bar chart, with the year on the horizontal axis and the generation amount on the vertical axis, provides a clear visualization of energy production trends over time. This approach highlights how total electricity generation has evolved year by year, allowing for easy identification of growth patterns, declines, or plateaus in energy production. The chart below reveals an almost steady increase in total production with slight dips in years such as 2009, 2017 and 2020, and a major dip in the year 2021 probably due to incomplete data collection.



**Barchart showing the components of the yearly renewable energy bars**

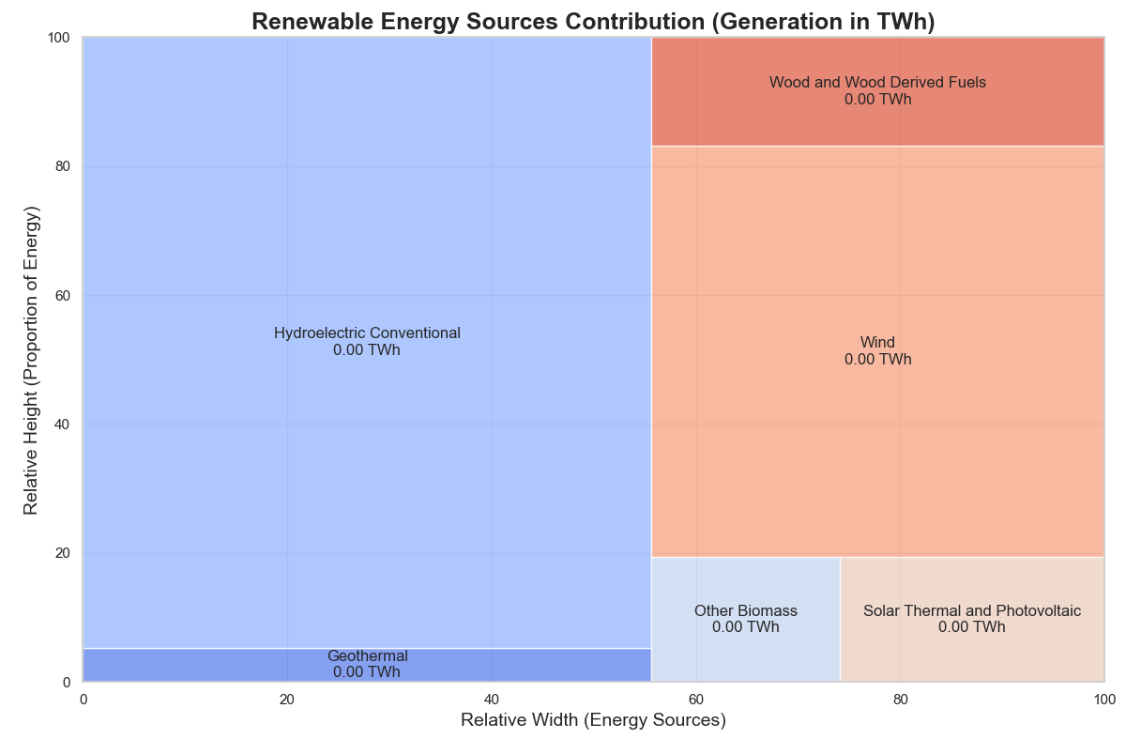
In the chart above, it is challenging to discern the specific contribution of individual renewable energy sources to the yearly total energy production. This lack of clarity makes it difficult to understand how each source, such as wind, solar, or hydroelectric power, has grown over time and contributed to the overall trend. To address this limitation, a chart that breaks down and displays the individual energy sources for each year is essential. Such a visualization would provide a more detailed perspective, highlighting the distinct impact of each renewable source and enabling a comprehensive analysis of their contributions to the nation's energy transition.



The chart above clearly illustrates that wind energy capacity has experienced steady growth over the years, alongside solar, thermal, and photovoltaic energy sources, which have also shown consistent increases. In contrast, hydroelectricity capacity has remained stagnant throughout the observed period, as has geothermal energy, which has exhibited little to no significant change. This disparity highlights the varying rates of development across renewable energy sources, with some advancing rapidly while others maintain a more static contribution to the overall energy mix.

**Square Area Graph**

The square area graph below effectively highlights the significance of each renewable energy source's contribution to total energy generation. Hydroelectricity stands out as the dominant source, occupying the largest area in the graph, signifying its substantial role in renewable energy production. Following hydroelectricity, wind energy emerges as the second-largest contributor, reflecting its steady growth in recent years. Next is wood and wood-derived fuels, which still hold a notable position in the renewable energy mix. Solar energy follows, showcasing its increasing importance, while other biomass contributes a slightly smaller but meaningful portion. Lastly, geothermal energy occupies the smallest area, indicating its relatively limited contribution compared to the other renewable sources. This visualization provides a clear representation of the varying levels of reliance on different renewable energy sources.



**Conclusion**

The exploratory data analysis (EDA) of the energy generation dataset has provided significant insights into the trends, patterns, and variations shaping the U.S. energy sector. By leveraging comprehensive and clean data, the analysis uncovered key observations, such as the increasing contribution of renewable energy sources like wind and solar, the stagnation of traditional renewables such as hydroelectric and geothermal power, and the steady decline in reliance on fossil fuels. Seasonal and regional trends, as well as the growing influence of modern technologies and policies, further enrich the understanding of the sector's evolution.

Through visualizations such as line plots, bar charts, KDE plots, and treemaps, the EDA highlighted the diversity in energy production levels across states, energy sources, and years. This analysis also illuminated the dynamic relationships between variables, revealing the broader trajectory toward sustainability and the transition to a low-carbon energy economy. Importantly, the insights gained from this EDA establish a solid foundation for future predictive modeling efforts, enabling stakeholders to forecast trends, allocate resources efficiently, and inform policies to support a resilient and sustainable energy infrastructure. The findings underscore the critical role of renewable energy in meeting climate goals while addressing emerging challenges in the energy landscape.

**References**

Mogardo, K. (2023). *US Energy Generation 2001-2022*. Kaggle. https://www.kaggle.com/datasets/kevinmorgado/us-energy-generation-2001-2022/data

Wikipedia contributors. (2023, July 28*). Energy policy of the United States. In Wikipedia. Retrieved* *from* <https://en.wikipedia.org/wiki/Energy_policy_of_the_United_States>.

BloombergNEF & Business Council for Sustainable Energy. (2025, February 20). *New study shows American sustainable energy technologies are ready to meet increasing energy demand. Retrieved from* https://about.bnef.com/blog/new-study-shows-american-sustainable-energy-technologies-are-ready-to-meet-increasing-energy-demand/

Knaflic, C. N. (2015). Storytelling with data (C. N. Knaflic, Ed.). John Wiley & Sons