US Energy Generated EDA

COMPUTER SCIENCE

SOFTWARE AND DATA MODELING

Yadika Dammagoni

Central Michigan University

TUESDAY, DECEMBER 6

Overview

The energy generated is measured in Megawatt hours. The dates of recording for this dataset ranges from the years 2001 to 2022 and is further broken down by month. The data is divided by each state including D.C. and by energy sources. Some example energy sources are Hydroelectric (dams), Wind, Coal, Natural Gas, and Nuclear. We also merged state population estimates into the dataset.

The main intention is to find the data of each state which use the different types of renewable energy sources and find which states produce the leading and smallest energy source. Correlation between the attributes etc. For this, we collected the datasets from the kaggle (Energy Production dataset, Average Monthly US State Temperature) and Hawai'i and Alaska datasets from the website. Apart from this, we also collected State population table data from Wikipedia to get the population of each state.

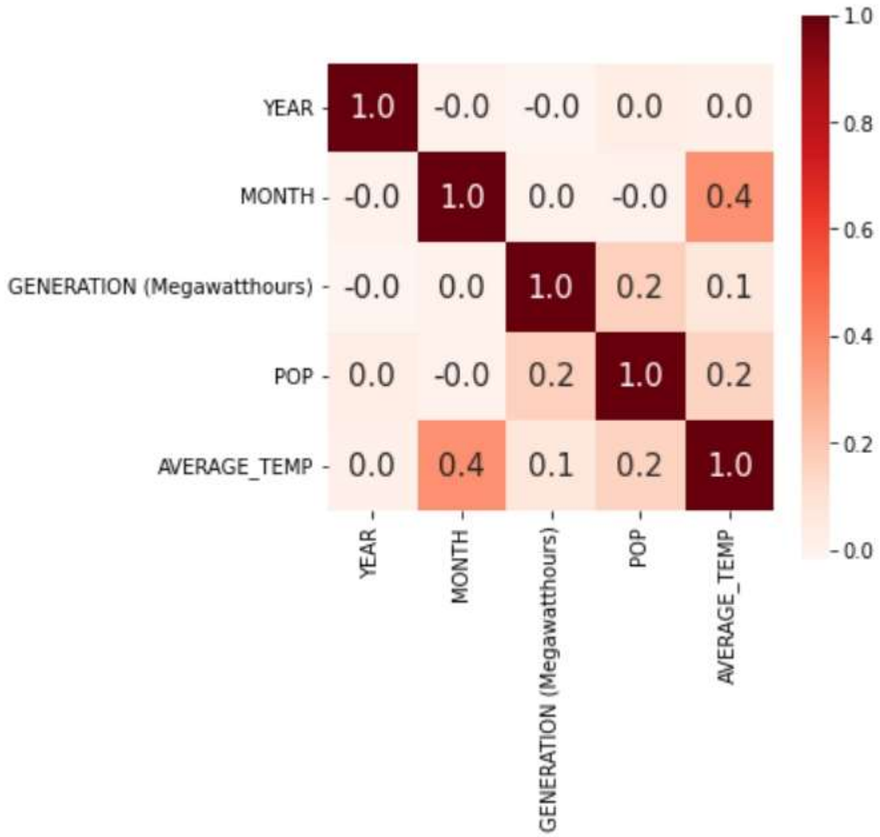
We have cleaned each dataset and removed unwanted columns, and null values in order to merge with other datasets. Initially, we needed to adjust the other datasets to be similar to the energy generation one to have columns for year and state. The scrapped table has a column for every state and rows for every year. So, in order to merge the data, it will need to be in a similar format. So, all the state columns were combined into a new data frame. Then we started combining the average temperatures and filling in the missing values. Similar to the state population table we have adjusted the Hawaii and Alaska datasets data we modified and made a final merged dataset.

Dataset

The columns of the dataset go in this order YEAR, MONTH, STATE, TYPE OF PRODUCER, ENERFY SOURCE, GENERATION, and (Megawatt hours). Out of these columns, there are two temporal (YEAR and MONTH) and one spatial (STATE) column. YEAR is the year when the record was recorded (2001 to 2022), MONTH is the month number when the record was recorded, STATE is the US state where the record was recorded, TYPE OF PRODUCER is what sets the power was produced, ENERGY SOURCE what was used to make the energy (Dam, Coal Plant, etc.), and GENERATION (Megawatt hours) is how much energy was generated. After merging the datasets, the final dataset contains columns YEAR, MONTH, STATE , TYPE OF PRODUCER(Total Electric power Industry, Electric Generators, Electric Utilities etc) these are the type of producers who procduces the enery, ENERGY SOURCE, GENERATION, and POP is population column which defines count of the people who uses the different types of energies. AVERAGE_TEMP is the column which records the average of the temperatures which generates from the different energy sources.

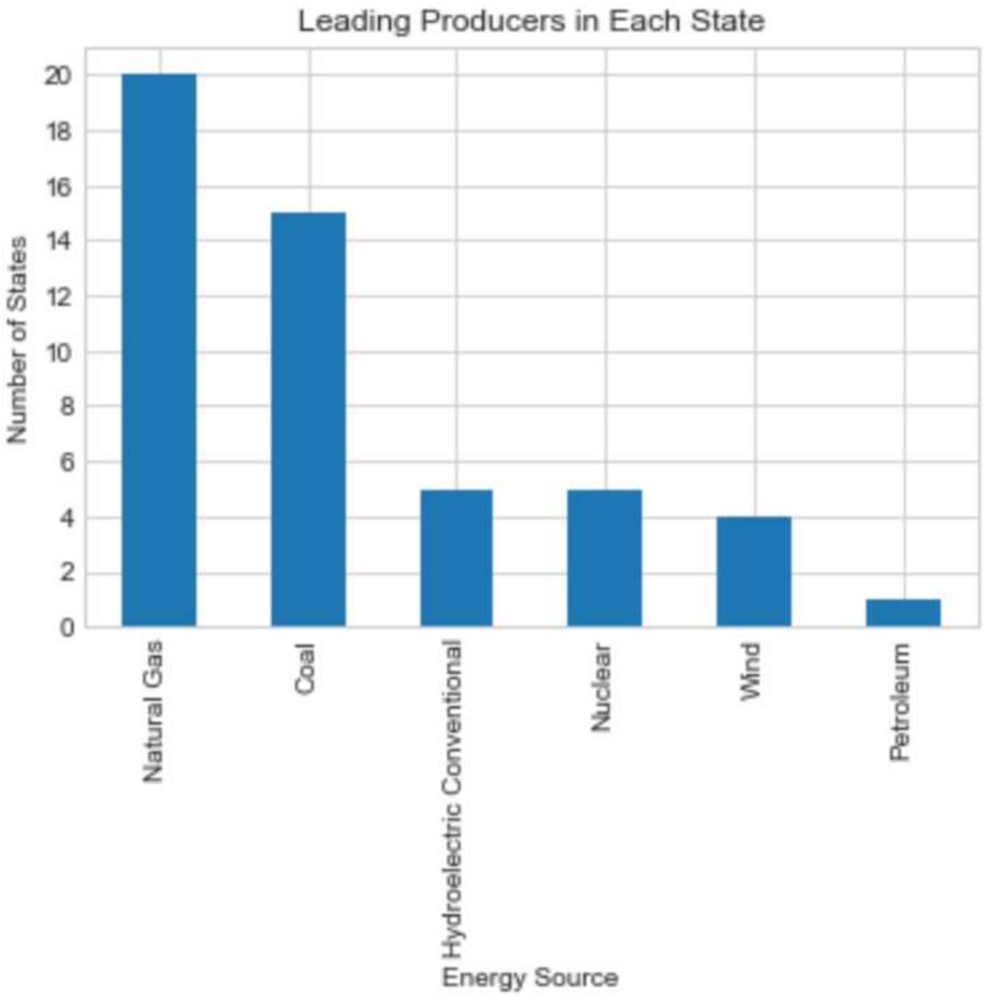
EDA

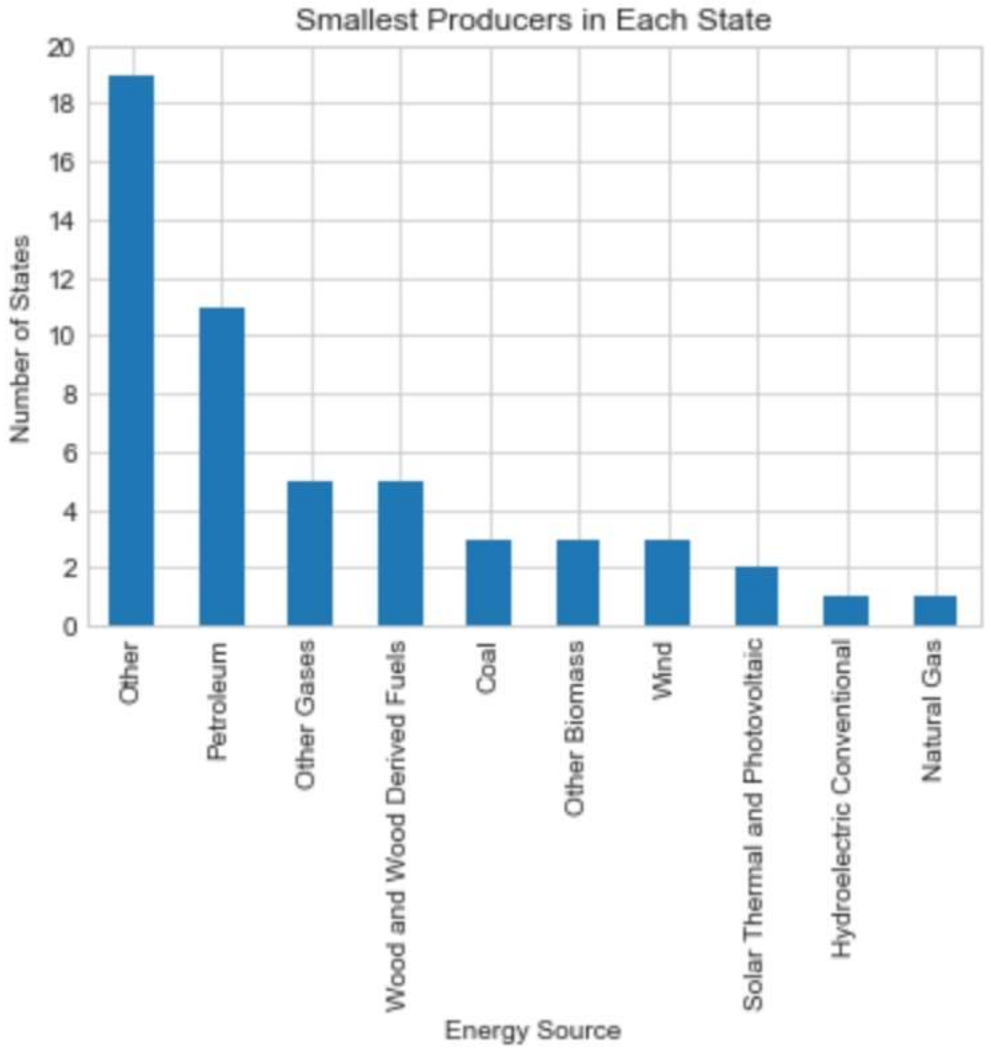
Looking through the dataset, a few points need to check regarding the correlation. From the below plot, we can identify the correlation between the temperature and the month, and also population correlates with the energy generation. Calculated the energy metrics of average temperature like mean, which we got the results (`53.32712104847148`), these results help us to know briefly about the dataset.



Looking through the dataset, it was found that from 2001 to 2015 Coal was the source that generated the most electricity per year and then from 2016 until 2021, Natural Gas produced the most per year. For the least produced per year, Pumped Storage was by far the leader since it was always losing energy. The reason for this, is because Pumped Storage stores energy and is losing a new amount per year. So, not including Pumped Storage gives a better insight on what sources produce the least. From 2001 until 2013, Solar Thermal and Photovoltaic energy was the least and from 2014 to 2017 and from 2019 until 2021, Other Gases produced the least with Other producing the least in 2018.

Which energy source was the leading producer in each state was also investigated. Specifically for this data, we focused on the year 2021. Here are graphs that show the results.

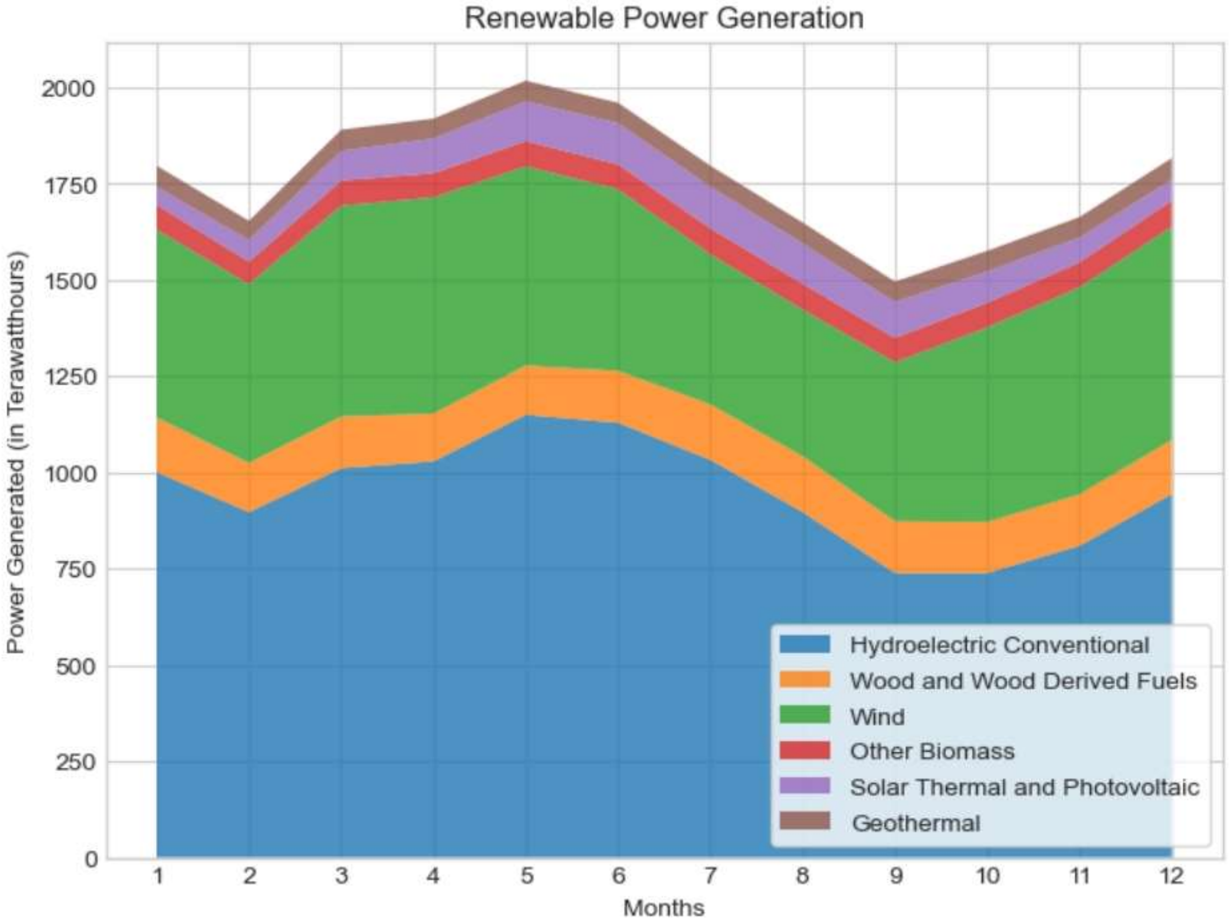


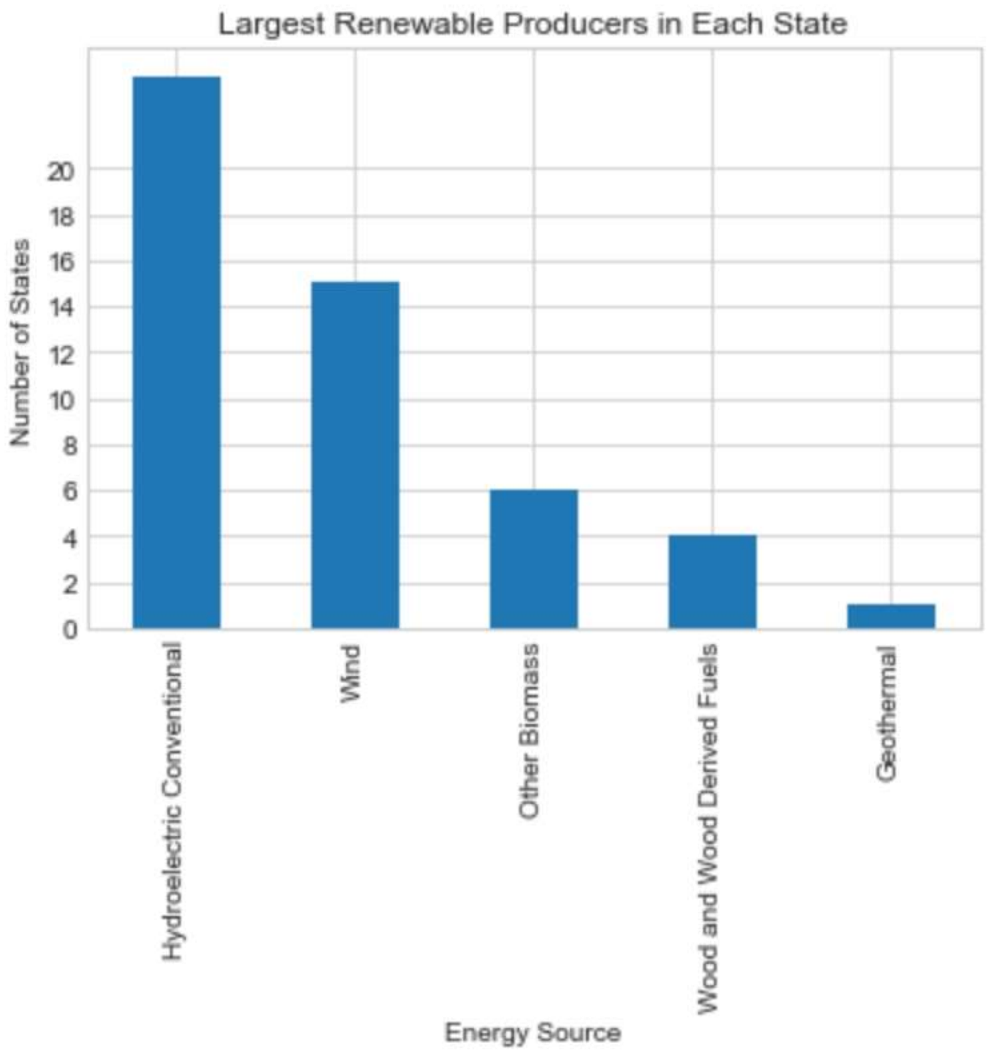


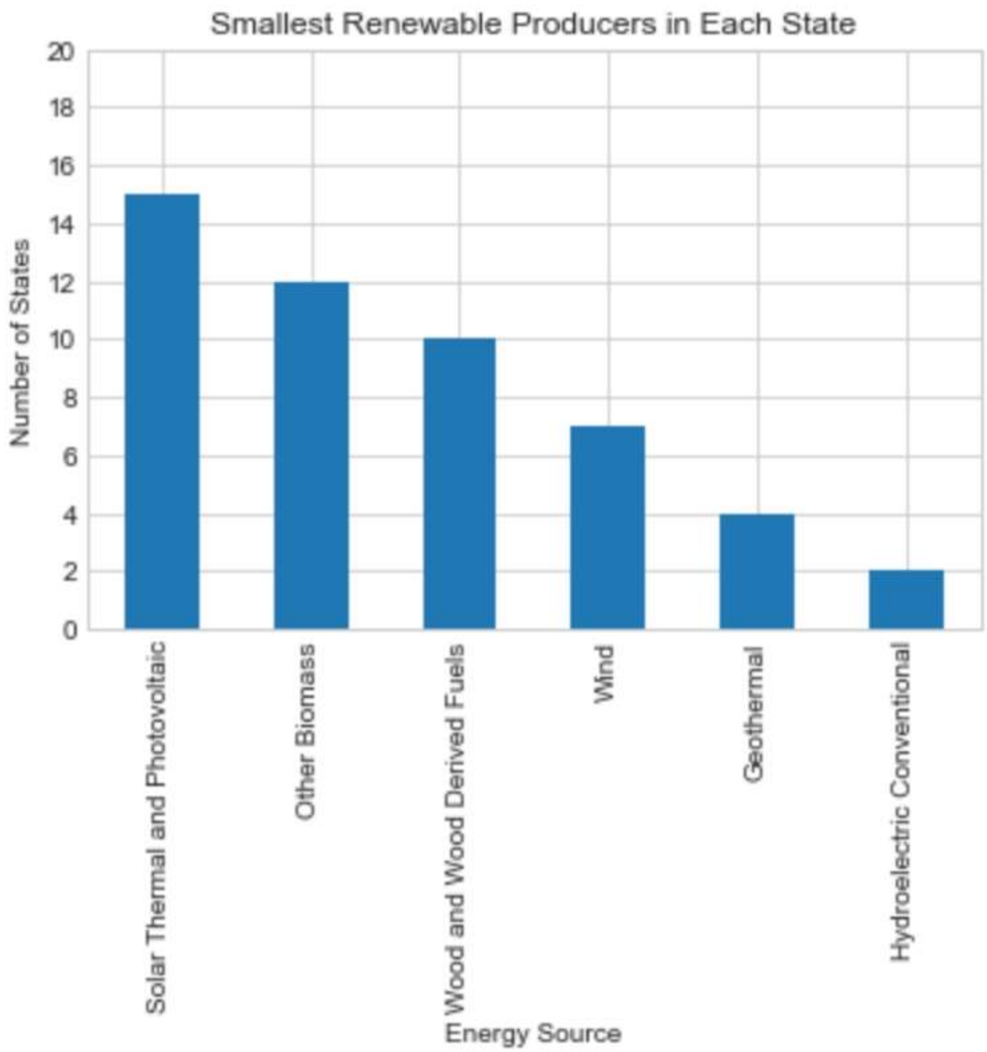
Interestingly, states tend to vary more with energy sources that produce the least energy and tend to use similar sources that produce the most energy.

Renewable energy sources were also looked at. After looking through all the different types of energy sources, these were the ones that were identified as being renewable: Hydroelectric Conventional, Wood and Wood Derived Fuels, Wind, Other Biomass, Solar Thermal and Photovoltaic, and Geothermal. From this, Hydroelectric leads the list for highest production for all renewable sources from 2001 to 2018. After 2018, Wind leads the list. Solar produces the least overall from 2001 to 2013 and then Geotherm does the least from 2014 to 2021. The next four figures show the energy production of each renewable source in more detail.

ENERGY SOURCE	MEAN GENERATION (Megawatthours)
Hydroelectric Conventional	309510.57
Wind	245882.78
Geothermal	193372.77
Solar Thermal and Photovoltaic	57305.17
Wood and Wood Derived Fuels	57229.02
Other Biomass	17782.58

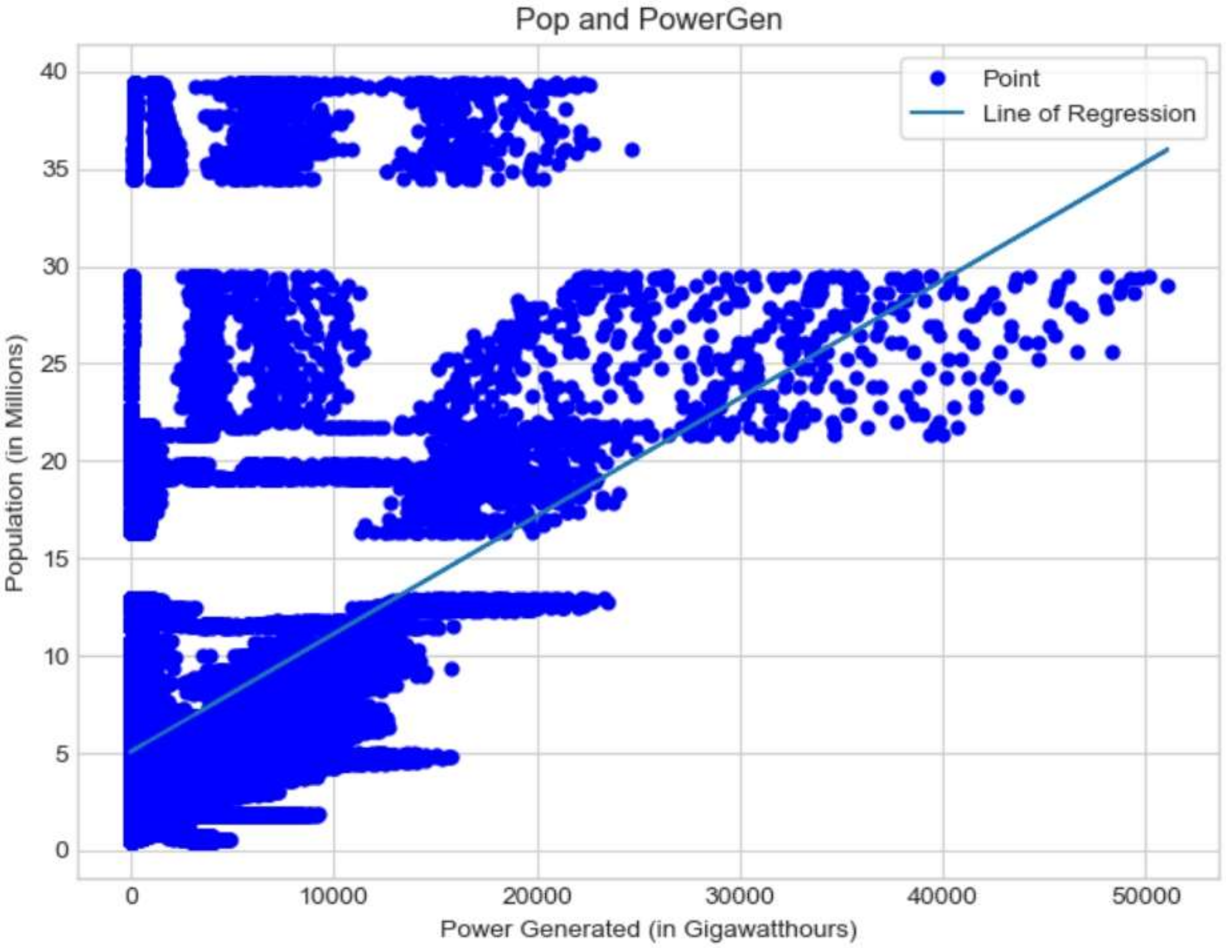



Largest Renewable Producers in Each State
Number of States
0
2
4
6
8
10
12
14
16
18
20
Hydroelectric Conventional
Wind
Other Biomass
Wood and Wood Derived Fuels
Geothermal
Energy Source



From these, one can see that Hydroelectric produces the most, followed by wind. Also, in the graphs, one can see that Hydroelectric and wind fluctuate between the months, with May to June being the best for Hydroelectric and that March to April is the best for Wind. Each state's top and bottom renewable electricity source was also found.

Finally, the State's population was investigated to see what correlates with it. From this, it was found that it correlates positively somewhat with energy generation, positively a tiny with average temperature, and almost no correlation with year and month. From these, a graph was made that plots population vs power generation. There is also included a regression line.



One interesting feature of this graph is how population divides the graph into 3 parts: 0 to around 13 million, 16 to 30 million, and then 34 to 40 million. Another interesting feature is how the points above 30 million people generate a similar amount of power compared to points below 13 million.

Conclusion

On this project we have performed all the concepts which we have learnt in the class like data cleaning, data wrangling, pandas and conducting the EDA part. After merging the all the datasets into single dataset, we performed operations on it. Calculated the mean average temperatures and correlation between each attribute, and also identified the results and plots which energy source is used highest and lowest in each state in the US. Calculated the mean generation of the each energy source like wind, geothermal etc. Plotted the graph between population and power generation with the regression line which divides into the 3 parts and gives the results.

Appendix A: How We Created the Dataset

Appendix A is provided as a Jupyter notebook included with the discussion board post.

Appendix B: What We Did and Learned

What I Did:

- Assisted in gathering requirements and consolidating data into a single dataset.
- Discussed and removed unnecessary columns to ensure data accuracy.
- Web scraped data from Wikipedia, including state population and average monthly temperature datasets, and integrated this data into the main Energy Dataset.
- Conducted exploratory data analysis (EDA) focusing on renewable energy sources, energy metrics, and correlations with population and temperature.
- Calculated key energy metrics, such as mean and average temperatures, and identified the highest and lowest energy generation values.

What I Learned:

- Reinforced concepts of data cleaning, wrangling, and exploratory data analysis (EDA) using pandas and Python.
- Gained practical experience in integrating and analyzing diverse datasets.
- Deepened understanding of how energy data intersects with population and temperature variables.
- Improved skills in web scraping, data integration, and calculating energy metrics

References

Morgado, R. (2022). *US Energy Generation 2001-2022* (Version 2) [Data set]. Kaggle. Retrieved October 8, 2022, from https://www.kaggle.com/datasets/kevinmorgado/us-energy-generation-2001-2022

NOAA National Centers for Environmental information. (2022) *Climate at a Glance: Statewide Mapping*. Retrieved on November 28, 2022, from https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/statewide/mapping

Wong, J. (2022). *Average Monthly Temperature by US State* (Version 1) [Data set]. Kaggle. Retrieved September 16, 2022, from https://www.kaggle.com/datasets/justinrwong/average-monthly-temperature-by-us-state