Exploratory Analysis

Nathan Bush

2022-11-08

### REPORT

In this analysis I compare the political leanings of states in the US with their energy usage statistics and emissions. I am interested in understanding whether state political views effect the type of energy states they are using. This is important because it can effect and inform policy, show whether or not states are taking advantage of their natural resources, and help us understand a few of the dynamics behind energy systems and how politics plays a role in these systems. If I find that conservative leaning states tend to use less renewable energy than their state’s natural resources provide, or that more liberal states tend to rely excessively on renewable energy when their state is relatively sparse in these resources, it can be infered that more effective policy needs to be implemented to incentivize states to act more efficiently regarding their energy generation.

The data I use in this analysis comes from three sources. My first data set is energy usage statistics on a wide variety of metrics in each state. This is taken from the US Energy Information Administration. Key variables of interest here are the percentages of renewable energy, solar energy, petroleum, natural gas, and coal and as a percentage of the total energy use in the state. This data set also includes metrics of each state such as the average temperature, annual precipitation, and CO2 emissions. My second and third datasets are county voting data on the 2020 presidential election taken from the MIT Election and Data Science lab and the Harvard Dataverse. The key variables here are State, County, Candidate, # of votes in each county, and total votes. My third dataset is table that I webscraped from Wikipedia showing who each state’s governor is and which political party they are associated with. My hypothesis is that a state who voted Republican in the 2020 election and has a republican governor is more conservative than a state that has a democratic governor but voted republican in the 2020 election. I will be able to test whether or not there is any significant difference in energy usage based on these observed differences by adding the governor variable to my analysis.

Cleaning and merging this data was a timely process. I cleaned the energy data by getting rid of unnecessary rows and columns generated in excel and deleting variables which were unrelated to my analysis. In the energy usage data set, a few of the data points had \* for values less than 0.5 which the original creators of the data set deemed as statistically equal to 0. I changed these values to 0. I also adjusted the names of a few variables to be easier to work with. I deleted total US observations and the Washington DC variable. I transposed key rows and columns to get the data in tidy format. Other adjustments are notated in my code. For the presidential data, I filtered out all observations for votes other than those for Joe Biden and Donald Trump. I also summarized all vote types (early, absentee, in-person, etc…), and created a new variable that summarized total votes across all counties within states, and created a dummy variable for party. The governor’s data was fairly simple to web scrape and import into r. I deleted a couple of unnecessary variables, named key variables appropriately, and merged the datasets.

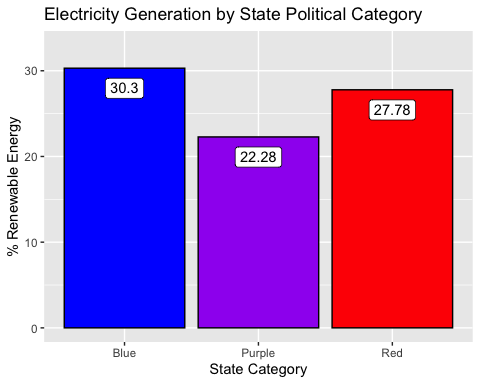
I merged the data using state as the unique ID and then got to work creating 3 state political categories. Blue, Red, and Purple, I used dummy variables for governor (1 if republican, 0 if not), for president (1 for voting republican, 0 if not). I then created a variable for state political category where (Blue = 0 for governor dummy & 0 for president dummy, Red = 1 for governor dummy and 1 for president dummy, Purple = 1 in either dummy, 0 in the other). I then created another variable for mean renewable energy percentage based on state political category. The units of analysis in my final merged data set vary between variables and are specific to each respective visualization.

I did employ a few transformations throughout the course the project. These are outlined in each visualization, but are primarily based upon adjusting state energy measurement statistics to account for differing state populations. I did not remove any outliers in the raw data, but do so below in a few visualizations.

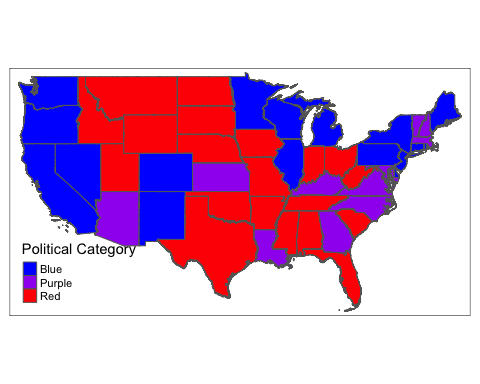
My primary style of analysis in this project is descriptive. I am primarily interested in the differences in renewable energy usage across state by political category, which is a descriptive measure. Initially, I had hoped to be able to use this data, to describe this question first, then potentially use regression analysis as a step towards causal inference in order to explain some potential factors that cause states to use more renewable energy than others. However, this leap to causal inference is a big one, and requires data I do not have. My data on renewable energy is limited to the percent of renewable energy a state generates to power their electricity needs as a percent of their total electricity related energy consumption. Energy data in particular is tricky to deal with because states are consistently buying and selling energy to other states based on their individual energy demands. This is unobserved in my data. Ideally, for a complete causal analysis, I would have the amount of energy bought and sold by each state, how each type of energy bought and sold was produced(renewable, natural gas, petroleum…), the amount of each type of energy that each state is using that it produced domestically, the amount of total energy used, produced, and sold by each state, and an index of some sort for the capacity of each state to produce renewable energy. Obviously, I do not have this data and am limited to what I do have. Since there are so many aspects of the ideal energy data for causal analysis that are absent from my data set, I will focus simply on descriptive relationships during my econometric analysis in this report.

In addition, my key variable of interest, Renewable Energy, is defined as the amount of renewable energy generated as a percentage of total electricity consumption. Therefore, it can be defined as both energy consumption and generation, which gets a bit convoluted in the analysis at times. Something to be aware of.

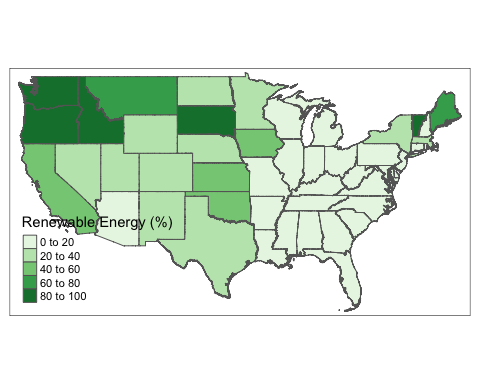
### Visualization 1

 Based on the above graph, we can see the percentage of electricity among groups of states that is generated by renewable energy. We observe that purple states (states with governors whose political party does not align with the way the state voted in the 2020 election) actually produce less of their electricity using renewable energy than either blue states (states with a democratic governor who voted for Joe Biden in 2020) or red states (states with a republican governor who voted for Donald Trump in 2020). This is an interesting result. One hypothesis for this could be that purple states are less efficient at passing legislation due to their conflicting political idealogies than either blue or red states. If renewable energy is objectively in the best interest of agents, passing renewable energy legislation would only be held up by one political attempting to spite the other for leverage purposes.

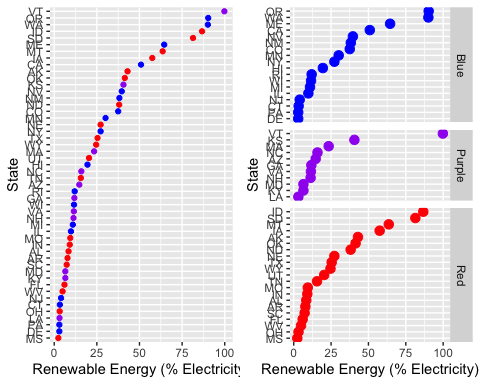
### Visualization 2

 This map is showing us the state category of each state in the US by color. We can use this as a reference for our results going forward. I removed Alaska and Hawaii from this map for visual reasons. Alaska is red, Hawaii is blue.

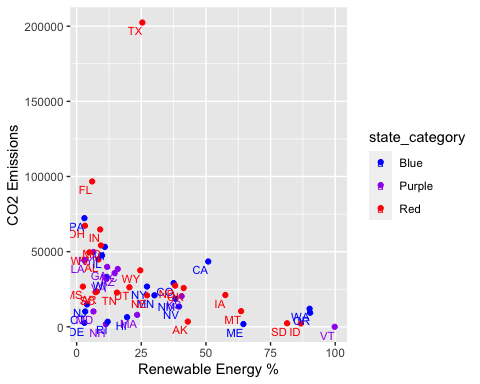
### Visualization 3

 This map shows the renewable energy percentage by state. Dark states have higher renewable energy percentages. We can see that states west of the Mississippi and concentrated in the northwest and northeast corners of the country seem to have higher percentages of renewable energy production. Again, Alaska and Hawaii have been removed for visual reasons here. Alaska has 43% of its electricity generated by renewable energy, and Hawaii has 19%.

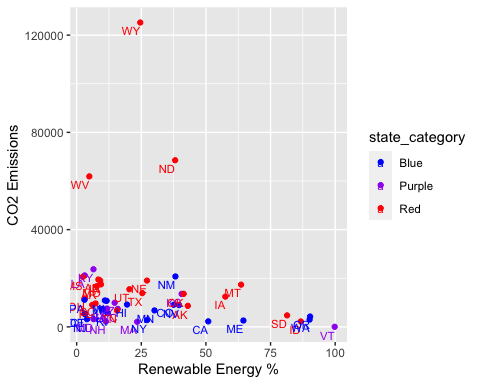
### Visualization 4

 In the above graph, I show the percentage of renewable energy by state in two dot plots. The plot on the left shows Renewable Energy percentages across all 50 states. We can see that the amount of renewable energy each states uses spans a the entire range between 0% and 100%, although it seems appears that roughly 2/3 of states use less than 25% renewable energy. On the right had side, I show the same graph, but separated by state political category. We do appear not see any noteworthy trends based upon political category. Although Vermont, a purple state, appears to be a bit of an outlier in the purple category, and is ran on 100% renewable energy. Very cool.

### Visualization 5

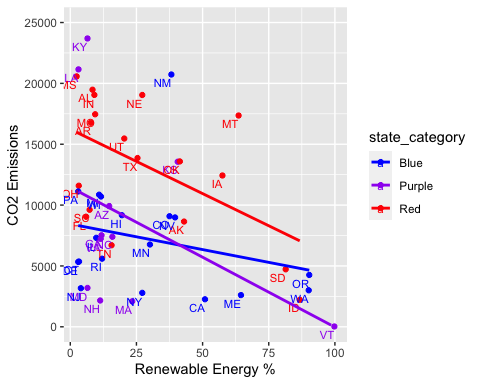
 In this graph I am looking to see if there is a relationship between Renewable Energy used and CO2 emissions. Intuition leads me to believe that there should be a negative linear relationship between the amount of renewable energy used and the amount of CO2 emitted. Based on this graph, we see that Texas is by far the biggest Co2 emitter in the US, despite using 25% renewable energy. Next, I will divide CO2 emissions by population size to see CO2 emissions per capita.

### Visualization 6



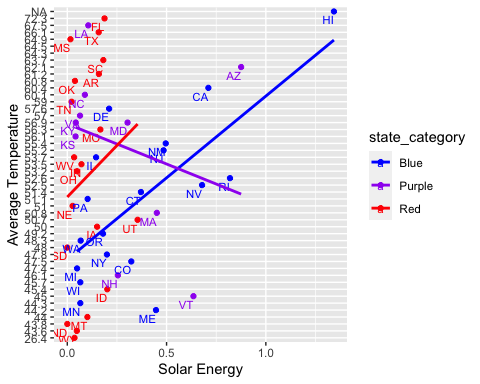
Now I get a graph showing CO2 emissions per capita and Renewable energy %. We see that Wyoming, West Virginia, and South Dakota are clear outliers and are significantly above the national average of CO2 emissions per capita. It is worth noting that all three of these states are red. Texas has entered the middle of the pack now that we adjust for state population. Let’s remove these three outliers and add local regressions by state political category to observe any potential relationship.

### Visualization 7

 Finally, we see that there is a negative linear relationship between CO2 emissions per capita and the percentage of renewable energy each state uses. This relationship is similar within each state political category. One observation worth noting is that of the 20 states with CO2 emissions per capita of less than 7500 metric tons, just 3 of these states are “red”.

I included the three graphs above to show my data visualization process and why it is useful to visualize your data with graphs. In visualization 6, we clearly saw that Texas was an outlier, calling for a transformation of our CO2 emissions variable to account for state population. Then, in visualization 7, we clearly see that there are three outlier states which will severely impact our analysis unless removed. Finally, we arrive at visualization 8 which gives us a more complete and holistic picture of the true relationship between renewable energy % and CO2 emissions.

### Visualization 8



In this graph I show the relationship between solar energy and average temperature by state. Intuition would lead us to believe that states with higher average temperatures employ more solar energy because they likely have more sunlight. I add local regressions to my graph to test whether or not this is the case. I find evidence that blue and red states are indeed employing more solar energy in states with hotter temperatures, while purple states are showing the opposite trend. Perhaps blue and red states are acting rationally in their own best interest, while purple states are unable to do so because of an entangled political environment. This MAY lend credence to our initial hypothesis that states with conflicting political interests are less efficient than states with more uniform political environments.

## Econometric Analysis

## reg\_1 reg\_2 reg\_3  
## Dependent Var.: Renewables Renewables new\_solar  
##   
## (Intercept) 108.2\*\*\* (22.50) 126.3\*\*\* (24.66) 0.0235 (0.1902)  
## avg\_temp -1.138\* (0.4762) -1.358\* (0.5174) 0.0085\* (0.0041)  
## GDP 0.0019 (0.0060) -0.0055 (0.0110) -1.8e-6 (5.06e-5)  
## precip -0.5504\* (0.2371) -0.6718\* (0.2661) -0.0040\* (0.0020)  
## co2\_pop -0.0005\* (0.0002)   
## Solar 0.0143 (0.0186)   
## trump\_vote 13.42 (14.44) -0.2704\* (0.1122)  
## rep\_gov 1.919 (11.39) 0.1157 (0.0894)  
## trump\_vote x rep\_gov -4.701 (18.03) -0.0810 (0.1414)  
## \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
## S.E. type IID IID IID  
## Observations 49 47 47  
## R2 0.30943 0.40378 0.43903  
## Adj. R2 0.26339 0.27826 0.35489  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

### Motivation

In my econometric analysis, I will be using regression analysis as a descriptive tool. I am interested in examining the relationships between a couple of independent variables within my data and my key variable of interest, the amount of renewable energy that each state generates as a percentage of its total electricity consumption. Primarily, I am interested in the differences between the amount of renewable energy by state categories. I can use regression analysis, similarly to the visualizations above, to see if there is any discernible difference between renewbable energy by state political category. I can also analyze other variable in order to see if any of them have predictive power in determining the amount of renewable energy a state generates for its electricity. It is highly likely that many of them will, due to reverse causality issues.

### Methods & Results

#### Regression 1

Based on Regresssion 1 and the the table shown above it, we see that both average temperature and precipitation have statistically significant negative coefficients at the 10% level. A 1 unit increase in the average temperature corresponds with a -1.138 change in the percent of renewable energy a state uses. A 1 unit increase in the amount of precipitation a state receives corresponds to a -0.5504 percent change in the amount of renewable energy a state uses. This is likely due to a lack of sunshine, and the may mean that solar energy is a relatively more popular form of renewable energy generation, an interesting result. The effect of GDP is statistically 0. This is a simple regression, and does not take into account political factors yet.

#### Regression 2

In this equation, I fully interact binary variables that take 1 if a state voted for Trump in the 2020 election and 1 if a state has a Republican Governor in order to discern differences in renewable energy by state political category. I also add continuous variables for CO2 emissions per capita. In this model, we see that a state that both voted for Trump in 2020 and has a republican governor is associated with 4.701% less renewable energy in comparison to states with just a trump vote, just a republican governor, or neither. However, this result is not statistically significant. It is interesting to realize that this is the only coefficient with a negative value, suggesting that blue and purple states both are predicted to have higher levels of renewable energy than red states, although again we should take this with a grain of salt due to the lack of significance. We also observe that a 1 unit increase in CO2 emissions per capita is associated with a .0004 percent decline in renewable energy, significant at the 10% level.

#### Regression 3

For my final regression, I use the amount of solar energy per capita as my dependent variable in an attempt to see if states are acting in their best interest regarding energy generation regardless of their political category. This furthers my analysis from visualization 8. We observe that in fact, states with a trump vote for president and a democratic governor (one type of purple state) are associated with 0.2704 less solar energy per capita, significant at the 10% level. All other state categories have either positive coefficients or coefficients that are statistically zero. This is a very interesting finding, and suggests that in fact, these types of states may not be acting in their best interests regarding solar energy. Obviously, due to the limitations in my data, and other exogenous factors, interpret this result lightly, but it is certainly worth noting. A 1 unit increase in the amount of precipitation a state receives is associated with a -0.0040 unit decrease in the amount of solar energy per capita, a result that is to be expected. A 1 unit increase in average temperature is associated with a 0.0085 unit increase in the amount of solar energy per capita that a state uses, significant at the 10% level.

### Conclusion

In my analysis, I examined the differences between the amount of renewable energy generated as a percentage of total electricity consumption between state political categories in a descriptive manner to see if there were any potential discrepencies. I do find that on average, blue states have the highest percentage of renewable energy, while purple states have the least. Potential evidence exists that purple states may not be acting in accordance with their objective best interests in terms of energy production, one hypothesis being that they have a more diverse political climate which makes it more difficult to pass legislation of any sort. I also show that there are three distinct outliers regarding CO2 emissions per capita in the US, all red states.