Andrew Shaw
Midterm - 6011 Coding & Analysis
12 July 2025

The data collected and cleaned will be used to help determine what state-level factors explain the variation in solar photovoltaic (PV) capacity growth across the United States between 2014 and 2023. The time horizon selected provides a window into the more recent push towards solar capacity providing the most relevant data. Two sets of data come from the U.S. Energy Information Administration (EIA) the first of which is total state capacity in megawatt hours. The second are electricity prices across states over the time horizon. The next data comes from the Lawrence Berkeley National Laboratory providing the breakdown of renewable portfolio standards (RPS) as well as any specific solar carve-outs for states in the time horizon. The final data describes the makeup state legislatures have any bearing on increased RPS or solar capacity which comes from the National Conference of State Legislatures (NCSL).

The datasets are quite robust and will require the following libraries: readr, readxl, tidyr, dplyr, janitor, and ggplot2. The first data coming from EIA is a robust set containing individual generators, technology used, and nameplate capacity or total megawatt hours across all states. These present a particular challenge as they are broken down by year and thus will need to be compiled using a `for` loop after creating a variable as a list of the file paths to the raw .xlsx files. As the files are imported, the `for` loop creates a 'year' variable that takes last 4 characters from the file name. Collecting the appropriate data requires heavy use of the dplyr library. Filtering for the rows containing 'solar PV' technology, grouping by state, summarizing by summing the total megawatt hours by state, and finally mutating which adds a 'year' column and the previously saved year variable from the file name. The yearly summaries are then appended to a blank data frame initialized before the loop.

The next data from the Berkeley Laboratory provides an interesting challenge as the variables needed do not present as variables, rather as values in a catch-all column. This is simply imported from a single excel spreadsheet. Next a smaller data frame is created after using grepl() from base R to filter for only rows that contain the relevant variables such as 'solar carve-out' and 'Total RPS'. Finally, as this is a wide dataset, the rows need to pivot to appropriately join the first dataset. Using pivot_longer() from tidyr restructures the data so year and the policy present in a state by year format. Because state policy varies widely, it is difficult to divine meaning among states for variable such as solar carve-out. More valuable to the analysis is whether such a carve-out for solar exists. Using 'ifelse' statements and mutating to create new variable columns a binary variable denotes whether a solar carve-outs was active for the state in the given year in addition to the total RPS percentage, which is more comparable across states. This data is then cleaned up by summarizing the new variable columns and grouped by state ready to be joined with the other data.

The last two data sets are much simpler to compile. The state legislatures from NCSL come as a .csv. The data provides all the data in the proper format to match the other data. The only steps are to filter the appropriate columns ‘state’, ‘year’, ‘party_control’, and to ensure the variable names used to join the data match using clean_names() function from the janitor library. The average annual price of electricity coming from EIA, also a .csv, has two major cleaning tasks. The first being the state variable entries all have the prefix “All Sectors: “that can be removed by using dplyr to mutate the state column and str_remove() function to remove that string leaving only the state. The pivot_longer() function transforms the data into the same state by year format for annual electricity costs. With all the data cleaned they can now all be joined using dpylr and the left_join() function. Each dataset is joined using the common ‘state’ and ‘year’ variables.

The final variables in the analysis data frame are as follows: total megawatts, RPS target percentage, RPS binary, solar carve-out binary, party control, and electricity price in c/kwh. The total megawatts (mW) is the dependent variable ranging from 0.5 mW to 18,925 mW with the median at only 188.4 mW. The RPS target provides an indicator for how ambitious a state is in pursuing renewable energy with Vermont leading in the most recent data at 63%. The RPS active variable remains stable across the time horizon with thirty states actively pursuing renewable policies. The binary solar carve-out variable can show whether the presence of targeted policies help boost the growth of solar generation in a state.

Party control provides a look behind a common partisan narrative. In the time horizon, Republicans have held a trifecta in around 22-24 states and the data can help answer the question if that is relevant to solar energy policies. Texas, historically a Republican trifecta has seen immense solar growth in recent years. This could be an outlier or perhaps sentiment is shifting. Finally, electricity prices in c/kWh provide a good control for market conditions. Prices range from 7.13 to 39.72 and an average price of 11.79 c/kWh and using the breakdown by state can help determine if higher prices result in an increase of solar capacity or if the opposite is true.

The variables will be used in a linear regression model to examine how factors like policies, politics, and prices correlate with solar capacity growth across states over time. Using ggplot2 can help visualize and highlight some of the interesting data from the analysis. Line plots for some of the fastest growing states in solar like Texas, are a simple but striking example and the analysis might point to which factors have led to Texas’ explosive growth. A map feels particularly useful in this context to visualize the correlation between a high RPS target and total megawatt capacity of operable generators across states. Depending upon the results of analysis, there could be many visually striking options for visualizations. While there is some other data that would be excellent to feature here, such as new data center construction or proposals, data is hard to come by given the recency of such projects and the uncertain times for renewable policy. This analysis aims to help shed some light on which factors support solar adoption and understanding these correlations can design better policy.