Final Project for CIS 3252

Libin Varughese

Department of Computer Information Systems, California State Polytechnic University, Pomona

CIS 3252: Business Intelligence

Dr. Fadi Batarseh

November 27, 2022

Introduction

I am interested in how much people use, acquire, and produce energy. The world we live in now is constantly moving towards a more technological future in almost every country and that would require an ever-increasing need for more energy to support that growth. Whether that energy demand is met by solar, fossil fuels, wind, hydropower, nuclear or any other means is what really got me interested in this domain. I want to see what the future of the energy industry, what is going to power it, and how it will affect various countries. This is what motivated me to choose this domain for this project as it has been a passion of mine for quite some time. The energy domain is something I believe everyone should be invested in, even just a little bit, as it impacts everyone's lives in one way or another.

The energy industry is mainly focused on how to locate, produce, transport, and sell energy, in various forms like LNG (Liquid Natural Gas), and how different sectors and people consume the energy. They are involved with oil companies, energy companies, and governments to name a few, but they really have a foothold in every industry as energy is needed by every single entity in society, whether they are an individual or a business. They are also responsible for what the majority of the energy consumed is sourced from like fossil fuels or renewable sources like solar power. They can also initiate change in where the majority of energy comes from and in doing so, can also affect international issues like climate change and the global energy crisis.

Business Framing

My first analytical question is: "Do states with higher average retail rates have higher total sales than states with lower average retail rates?" I wish to investigate the relationship between the rates people pay for electricity and the total amount of sales a state collects from those it supplies. This question is important to answer because state governments and other interested entities like electricity suppliers can see how they are performing as sales can be seen as a key performance indicator or they can discover what is the optimal price point that will maximize sales while reducing their costs.

My second analytical question is: “Do states that produce higher shares of energy have higher consumption rates of energy than states with lower shares of energy production?” I wish to investigate the relationship between the amount of energy a state produces annually and the amount of energy it consumes annually. This particular question is important to answer because one can see which states produce more than they need and can possibly sell off the excess energy to neighboring states or to whoever desires such as foreign governments and companies. Governments can also issue subsidies to states where there is plenty of excess energy produced to encourage them to continue producing at a constant rate like how agriculture is subsidized in the United States.

Key Stakeholders

A key stakeholder in this would be the state(s) and federal government of the United States. They are interested in this as they would be able to possibly predict just how much they can make off sales of excess energy to other entities like neighboring states and foreign governments once they receive the results of how much energy the states actually produce and how much of it they consume. The impact they can have on this is also significant as they can pass legislation on how much energy a state can produce, how the energy is produced, what is a reasonable rate to charge for energy usage, and how the energy is to be distributed and transported from the suppliers to customers. Anytime something is traded between states or internationally, the federal and the state (to a certain degree) government is always involved and when dealing with a precious resource such as energy, the impact they have is quite significant.

Another key stakeholder would be the suppliers of the energy such as power plants and energy suppliers like Southern California Edison. They would be interested in this as they are the ones who produce and supply the power to the states, so to see how much they produce and sell annually, while also seeing how much their customers consume allows them to improve their business such as reducing the utility rate to encourage the growth of sales. They may also improve their production capabilities (building more plants, improving existing distribution methods) if it seems that their customers are consuming more than they can produce currently or reduce production if it seems they are producing much more than what is necessary. The impact they have is the most notable and direct as the data needed to answer these questions are originating from them, so any change they make within their business can result in minor to massive changes to the data regarding sales, production, and possibly consumption.

The last key stakeholder I have identified are the customers of the energy suppliers. They are interested in this as they are the ones who make up the energy sales and are also the end users (consumers). By analyzing the results, they may conclude that they are paying too much for their energy usage and may petition their government to enact policies to limit what they can be reasonably charged as a utility rate. Their impact is noticeable as they might reduce or increase their consumption after seeing how much their supplier provides, which will affect total sales and annual production of the energy companies.

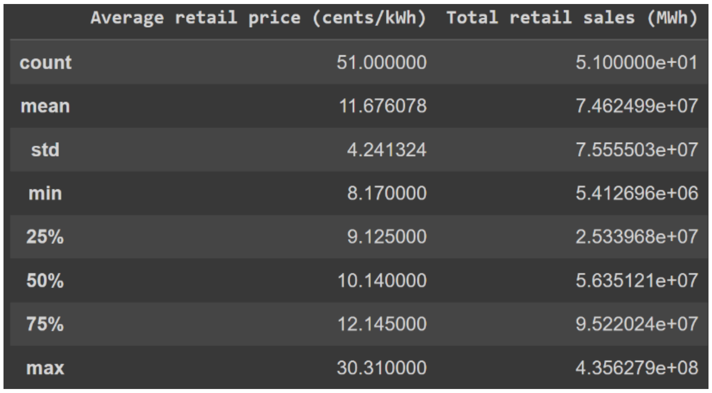
Understanding Data

The first dataset, which is an Excel sheet I created from the 2021 state electricity profiles data listed in spreadsheet form on the website and titled “Dataset_for_Final_Project”, I am using comes from the U.S. Energy Information Administration’s website. The data has five variables and their values as five columns and has a total of 51 instances. The first column is the ‘Name’ variable and has a variable type of object, and it is the state’s name. The second column is the ‘Average retail price (cents/kWh)’ variable, which has a variable type of float and is the average retail rate customers pay for electricity in their state in cents per kilowatt hour. The third column is the ‘Net summer capacity (MW)’ variable, which has the variable type of integer, and it is the maximum output that the electricity producing equipment can produce during the summer peak time period and is measured in megawatts. The fourth column is ‘Net generation (MWh)’ variable, which has the variable type of integer, and it is the total amount of electricity produced by the state for the year. The fifth and final column is the ‘Total retail sales (MWh)’ variable, which also has the variable type of integer, and it is the total sales the state made over the year in megawatt hours.

The second dataset, which is an Excel CSV file and titled “SelectedStateRankingsData,” I am using also comes from the U.S. Energy Information Administration and is the State Total Energy Rankings of 2020. The dataset contains six variables and their values as six columns and has a total of 51 instances, however one of the instances is for Washington D.C. The first column is the ‘State’ variable, which has a variable type of object, and it is the state’s name written as an abbreviation. The second column is the ‘Production, U.S. Share’ variable, which has a variable type of float, and it is the amount of energy the state produces in comparison to all the other states for the year. The third column is the ‘Production, Rank’ variable, which has a variable type of integer, and is just where the state ranks when compared to the rest of the states of the U.S. in terms of overall energy production amount. The fourth column is the ‘Consumption per Capita, Million Btu’ variable, which has a variable type of integer, and it is the amount of energy one person in the state consumed for the year measured in BTUs (British Thermal Unit). The fifth column is the ‘Consumption per Capita, Rank’ variable, which has a variable type of integer, and is how the state ranks to the other states in terms of consumption. The sixth column is the ‘Expenditures per Capita, Dollars’ variable, which has the variable type of integer, and is how much one person from the state spent on energy for the year. The seventh and final column is the ‘Expenditures per Capita, Rank’ variable, which has a variable type of integer, and is how the state ranks to the other states in terms of expenditures for energy.

Findings

Question #1: Do states with higher average retail rates have higher total sales than states with lower average retail rates?



	Average retail price (cents/kWh)	Total retail sales (MWh)
count	51.000000	5.100000e+01
mean	11.676078	7.462499e+07
std	4.241324	7.555503e+07
min	8.170000	5.412696e+06
25%	9.125000	2.533968e+07
50%	10.140000	5.635121e+07
75%	12.145000	9.522024e+07
max	30.310000	4.356279e+08

The above summary statistics are from the first dataset and show us that the mean retail price/rate that Americans pay for electricity is 11.68 cents and this price point results in a mean total retail sale of 74,624,990 megawatt hours. It also shows us that the minimum (lowest) retail price is 8.17 cents, and that the maximum (highest) retail price is 30.31 cents. The lowest price results in a total retail sale of 5,412,696 megawatt hours and the highest price results in a total retail sale of 435,627,900 megawatt hours. The statistics also provide us the standard deviation of both columns and the 25%, 50%, and 75% quartiles of both quartiles. The count refers to the number of instances in the dataset, which is the same for both columns.

Figure 1.1

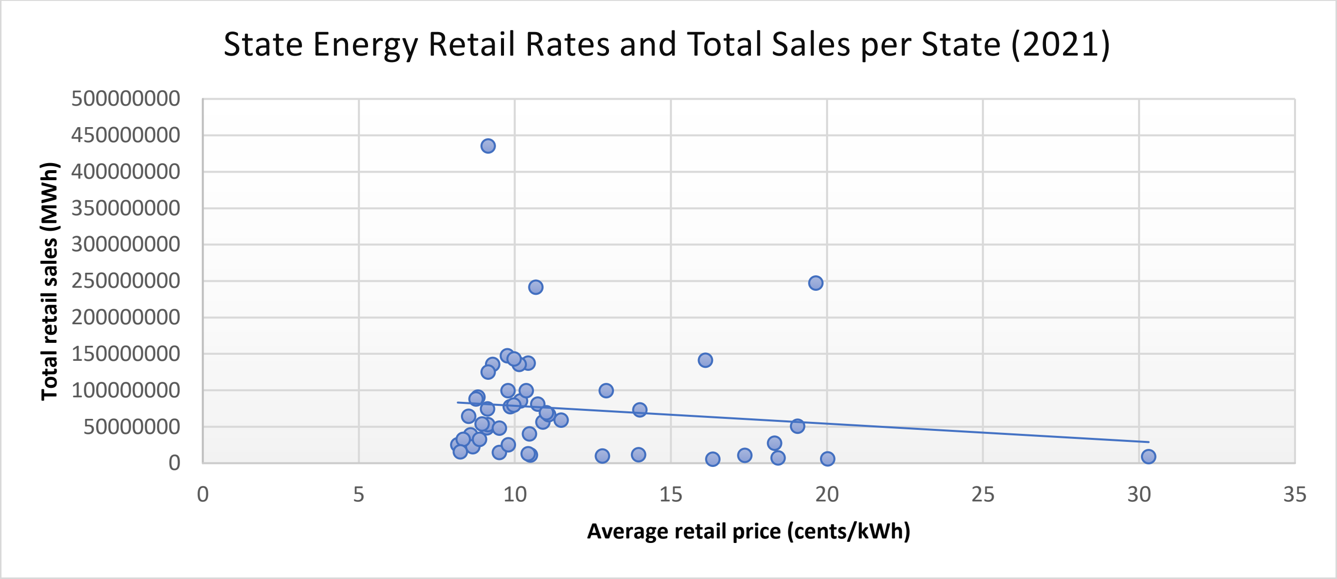
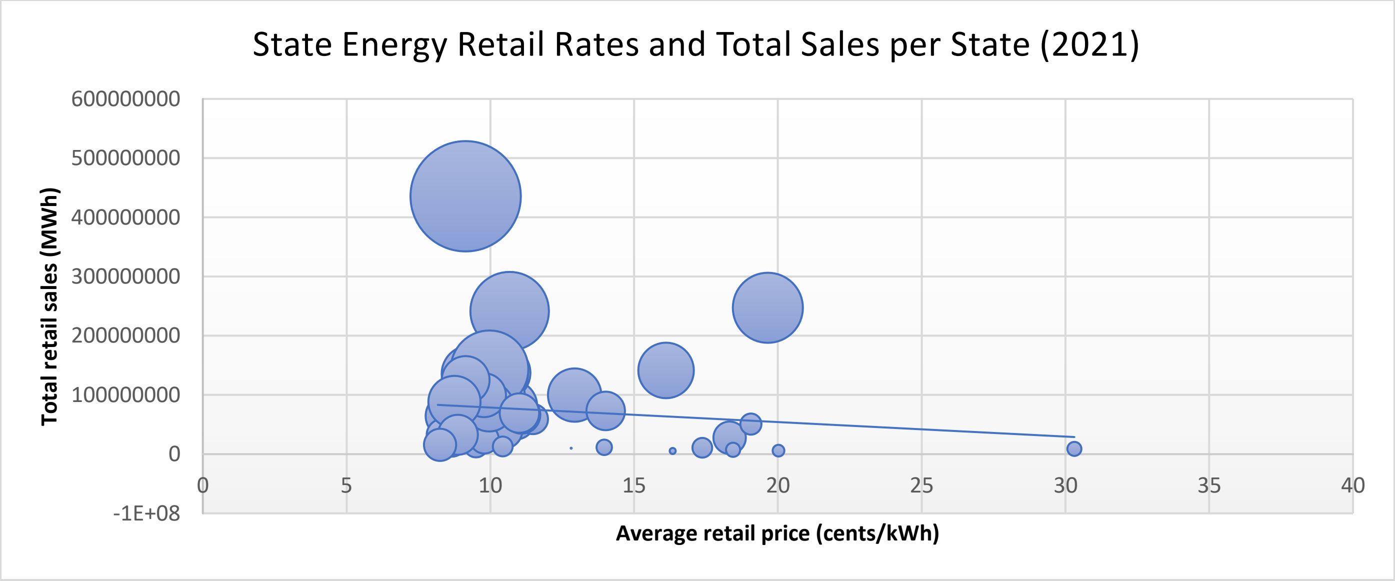
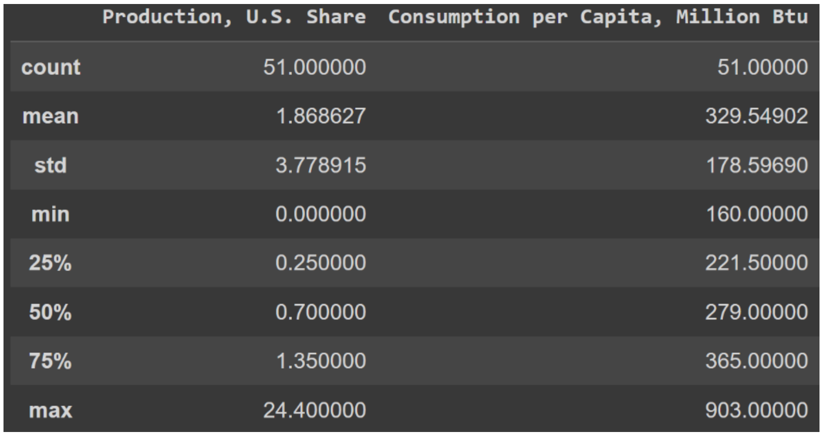


Figure 1.2



In Figure 1.1, each point in the scatter plot has a x value of the average retail price for the state in cents per kilowatt hour and a y value of the states total retail sales for the year in megawatt hours. Figure 1.2, which is a bubble chart, has the same x and y values, but the size of each bubble is based on the net generation of each state in megawatt hours. The bigger the bubble, the more electricity it generated. Both visualizations convey to us that most states have an average retail price of around 8 to 11 cents per kilowatt hour and that many of them have a total retail sale amount of 30,000,000 to 120,000,000 megawatt hours. The trendline and the way the points are clustered on Figures 1.1 and 1.2 suggest that the higher the state's average retail price, the lower the state's total retail sales will be. However, there are a few outliers that are possibly messing up the data in both Figures 1.1 and 1.2 and Figure 1.2 seems to indicate that the more electricity a state generates, the more retail sales it makes based on the bubble's size and position on the plot.

Question #2: Do states that produce higher shares of energy have higher consumption rates of energy than states with lower shares of energy production?



	Production, U.S. Share	Consumption per Capita, Million Btu
count	51.000000	51.00000
mean	1.868627	329.54902
std	3.778915	178.59690
min	0.000000	160.00000
25%	0.250000	221.50000
50%	0.700000	279.00000
75%	1.350000	365.00000
max	24.400000	903.00000

The above summary statistics are from the second dataset and show us that the mean production amount of energy each state makes is about 1.87% of the total amount of energy produced in the U.S. and that one person in each state consumes on average 329.54902 million BTUs of energy a year. It also shows us that the minimum (lowest) percentage of energy production a state had was 0%, and that the maximum (highest) percentage was 24.4%. The lowest percentage results in a single person consuming 160 million BTUs of energy and the highest percentage results in a single person consuming 903 million BTUs of energy. The statistics also provide us the standard deviation of both columns and the 25%, 50%, and 75% quartiles of both quartiles. The count refers to the number of instances in the dataset, which is the same for both columns.

Figure 2.1

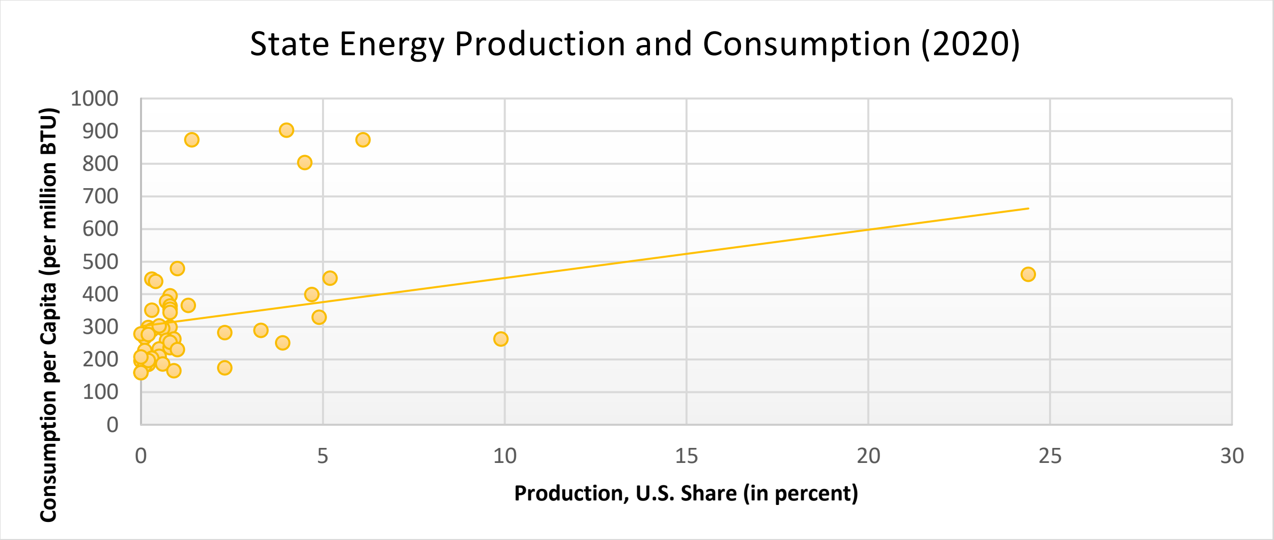
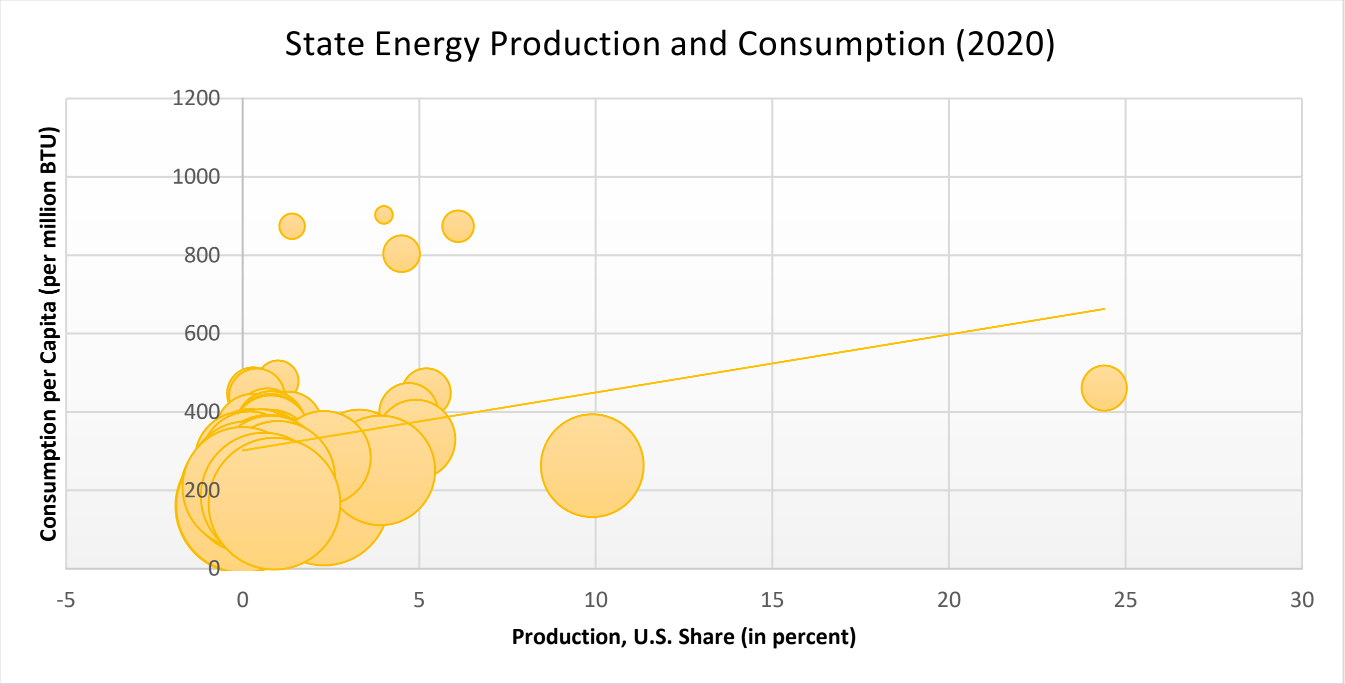


Figure 2.2 (on next page)



In Figure 2.1, each point in the scatter plot has a x value of the amount of energy production a state has to the total U.S. amount expressed as a percent and a y value of consumption per capita (amount that one person consumes) in millions of BTUs. Figure 2.2, which is a bubble chart, has the same x and y values, but the size of each bubble is based on the state's ranking in terms of energy consumption per capita. In this case, the smaller the bubble, the more energy the average person of that state consumes. Both visualizations convey to us that most states have an average production percentage of around 0.20 to 1.00 and that the average American person consumes about 200-300 million BTUs of energy. The way the points are clustered on Figures 2.1 and 2.2 and their respective trendlines suggest that states that produce more energy do consume more energy than states that produce less. The outliers may seem to skew the results, but the reason for that is due to many states having similar production and consumption amounts. However, if one looks away from the cluster towards the rest of the points, there seems to be an upward trend supporting the finding that states that produce more energy also consume more as well.

Conclusion

For the first question, my findings suggest that states that have higher average retail prices/rates have lower total retail sales than states with average retail rates. It disproved my initial thoughts on the topic as the findings showed an opposite result from what I stated in the question . My suggestion based on my analysis and the results would be that if a state wanted to increase the amount of retail sales for electricity within itself, it would have to charge lower prices/rates for it. People usually speak with their wallet, and it is clear based on my analysis and results that people will spend less overall when something becomes more and more expensive, which in this case, is electricity and its price in cents per kilowatt hour.

For the second question, my findings suggests that states that have a higher share of energy production also have a higher rate of energy consumption compared to states that have lower shares of energy production. It proved correct my initial thoughts on the subject as my findings presented results that agree with what I stated in the question. My suggestion based on my analysis and results would be that if a state does increase its energy production capabilities and overall amount produced, it should anticipate a higher rate of consumption as well from the average person living within the state as that is what my results also suggests. It seems that the increase in supply also led to an increase in the demand. This might be that people are conserving less energy since they know that their state is producing more energy and therefore, they are allowed to use more of it without much consequence to the overall supply.

Reflection

My experience with this project was an overall positive one and I learned two key things from completing it, which helped enhance my experience with it. The first one is that analytics projects take time because they are an iterative process, which means that halfway through the project, one may realize that the wrong questions are being asked and will have to start from the beginning once again. By going through the process several times, one starts to learn more about the subject they are analyzing and makes connections and decisions based on information they may have not discovered during the first iteration. The iterative process makes it so that the analyst really understands what they are doing, why they are doing it, what are the proper questions to ask for the current situation and come out of it with the most accurate findings and best tailored solution. I do not believe I would have really learned this had I not experienced it for myself during this project.

The second key thing I learned during the completion of this project is that sometimes analyst have to do the best they can with the data they have acquired. One will not always find the data or datasets that will perfectly align with they are trying to solve or discover. They may only have the data that best aligns with it and from there, they must extract the relevant information from it to complete their task. The data they need may not be in one location, but in several, and the analyst will need to either combine the data together to receive a more complete picture from the extracted information or use the different datasets separately to answer different, but related questions. I personally experienced this as I had to use two different datasets to answer each question individually, but both the questions and datasets were about the same domain. Experiencing this firsthand showed me that research and data retrieval during the entirety of the project may be needed to solve any possible obstacles that may show up.