Denise Dodd

Solar Company Using Regression Models to Diversify Territories and Channels

INTRODUCTION

Clean energy has been on the rise for the past several years as governments, companies, and individuals work to combat the growing threat of climate change. Clean energy provides an eco-sustainable alternative to energy provided by methods which emit CO2 emissions into the atmosphere. My company specifically focuses on the solar energy as we provide solar panels and generators to clients across the United States. *In this project, I will research what additional opportunities would be available to our company outside of the U.S. solar sector. Are there other countries that it would be advantageous for us to provide solar to? Is there another branch of clean energy that is on the rise and could follow some of the same framework that the company already has in place?* As I proceed through this project, I will be looking for connections between solar energy and other forms of clean energy. I will also be searching for countries that have already proven to be viable consumers of solar energy. To prevent the risk of plateauing in the U.S. solar market, it will be good for my company to consider diversifying into other territories or clean energy channels.

THE DATA

The dataset for this project is titled Renewable Energy Worldwide: 1965~2022 and can be found via Kaggle at https://www.kaggle.com/datasets/belayethossainds/renewable-energy-world-wide-19652022?select=02+modern-renewable-energy-consumption.csv. This data set details generation of solar energy, wind energy, hydro energy, and other clean energy for several different countries.

In milestone 2, additional data was added to the data frame detailing the number of yearly sun-hours in each country. This dataset, titled Sunshine Duration by City, was obtained from Kaggle as well can be found here: https://www.kaggle.com/datasets/prasertk/sunshine-duration-by-city I joined the two datasets on the country column and checked for nulls to ensure accuracy of the join.

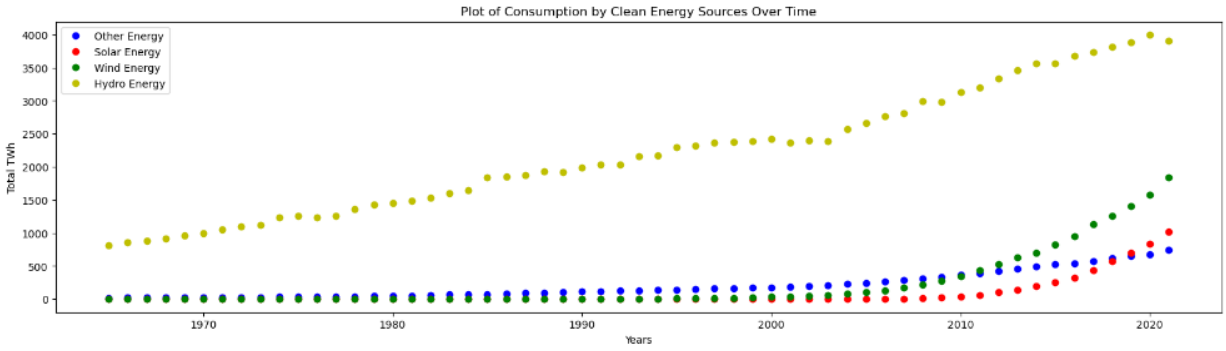
This additional data point will add context to which countries have the greatest supply of solar which can be harvested for consumption.

SUMMARY

This project was divided into three milestones. The first two milestones involved exploring the data through graphs (Milestone 1) and data cleaning (Milestone 2). In the final milestone (Milestone 3), I perform a variety of regression analyses.

MILESTONE 1

In the first Milestone, I explored the data with several visualizations to determine the scope of the data and if there are any preexisting relationships between the variables. The most enlightening visualization was a scatterplot that showed generation of the four clean energy sources (hydro, wind, solar and other) over time. This scatterplot showed me that there has been a steep increase in solar energy, as well as all other forms of clean energy, during the last 10 years in the dataset. For the first 40+ years represented in the dataset many forms of clean energy were close to nonexistent in many of the countries in this dataset. I will consider this when I clean the data in Milestone 2. Also, this graph shows that solar and wind energy seem to have a linear relationship and that both are increasing at a steady rate. This tells me that my company could branch out into wind energy, and it will not cannibalize our current solar energy markets.



MILESTONE 2

The second milestone focused on cleaning and preparing the data. The first several steps of this milestone pertained to incorporating the sun hour data into the data frame. I first uploaded the sun hour dataset referenced above and aggregated it by mean yearly sun hours grouping by country. I assigned this aggregation to a new data frame which I joined with my original data frame from Milestone 1 on the Entity/Country columns. After the join, I checked for nulls to ensure that every country in my original data frame and a corresponding sun hours entry from my added dataset. I then dropped the "Country" and "Code" cols as they are repetitive with the "Entity" column. The final part of cleaning the join data is to rename the columns. Each dataset had a "Years" column which represented a different set of data. I renamed these to accurately call out what yearly data each column is displaying.

Now that the datasets have been joined and I have incorporated sun hours into my data frame, my next step was to drop several rows. As noted above, in Milestone 1 I learned that the data trends differently in the last 10 years of the dataset than in the beginning of the dataset where clean energy had not yet emerged as a viable form of energy. For many countries in the first years of this dataset, they generated 0 terawatt hours of solar or wind energy. As clean energy becomes more popular and countries are building structures to support clean energy, data from a time prior to clean energy becomes irrelevant. Therefore, I dropped several rows of data and only kept data that from on or after 2011.

With the data cleaned, I created a correlation matrix of the data to determine if there is a relationship between any of the variables and the strength of those relationships. I was pleased to see that there is a strong positive relationship between wind and solar energy which confirms what I saw in the scatter plot in Milestone 1. There are a lot of strong, positive relationships in this dataset which puts the thought of multicollinearity in the back of my mind. I will have to consider that when selecting a regression model.

	Year	Geo Biomass Other - TWh	Solar Generation - TWh	Wind Generation - TWh	Hydro Generation - TWh	Avg_Yearly_Sun(Hrs)
Year	1.000000e+00	0.089656	0.183939	0.122239	0.021614	-9.284958e-15
Geo Biomass Other - TWh	8.965592e-02	1.000000	0.790165	0.869896	0.694347	-9.989743e-02
Solar Generation - TWh	1.839389e-01	0.790165	1.000000	0.899925	0.643077	-2.103263e-02
Wind Generation - TWh	1.222394e-01	0.869896	0.899925	1.000000	0.742122	-1.040472e-02
Hydro Generation - TWh	2.161420e-02	0.694347	0.643077	0.742122	1.000000	-7.171316e-02
Avg_Yearly_Sun(Hrs)	-9.284958e-15	-0.099897	-0.021033	-0.010405	-0.071713	1.000000e+00

My last step of Milestone 2 was to make dummy variables so the data is prepared for regressions in Milestone 3. I made dummy variables of the categorial “Country” column as I am hopeful that my model will be able to predict countries that are apt to consume solar energy and therefore would be good expansion territories for my company. I dropped the first dummy column to prevent multicollinearity, but there are still many dummy variables. This combined with what I learned while reviewing the confusion matrix tells me that I will have to be very conscious of multicollinearity when selecting my regression model.

MILESTONE 3

In Milestone 3, I was able to utilize this data to complete several regression models. To prepare for these regressions, I split the data into a target variable, which was Solar Generation, and all the remaining variables as feature variables. I then divided the data into an 80/20 training/test split. Using this data, I created an instance of a Linear Regression model and fitted it using the training data. I then ran a predictive analysis and assessed the efficacy of this analysis based on a series of metrics (R-Squared, Root Mean Squared Error, and Mean Absolute Error.) The metrics prove this to be an effective model, but knowing that I had concerns of multicollinearity, I also ran a ridge regression using the same using a similar process as the linear regression. The ridge regression had a slightly higher R-squared value and a lower RMSE value. Therefore, I chose to proceed with the ridge regression and calculate the 5 countries that are predicted to have the highest generation of solar energy. These would be viable territories for my country to research when considering new countries to branch out to.

In addition to diversifying territories, I was also interested in diversifying channels of clean energy. Solar and wind energy are highly correlated and my graph in Milestone 1 shows that they will not cannibalize one another. Therefore, I ran similar linear and ridge regressions on wind energy. Again, the ridge regression proved to be slightly more accurate, so I used this model to determine the top 5 predicted wind consuming countries.

	Country	Predicted Solar TWh
0	Japan	170.230103
1	Spain	165.012284
2	United Kingdom	164.738244
3	Italy	164.678260
4	Germany	164.162328

	Country	Predicted Wind TWh
0	United States	345.942446
1	Japan	295.863387
2	Italy	284.035542
3	India	281.293638
4	South Korea	281.281791

CONCLUSION

In conclusion, this project accomplished the preliminary work of determining if there are any viable countries or additional channels of clean energy that my company could branch out to. Solar and wind energy are both rising in popularity, are highly correlated, and do not cannibalize one another. *Wind energy appears to be a good complement to solar energy.* The next steps in considering this would be to determine if the solar and wind energy frameworks are compatible. Can our solar energy batteries and generators also be used to store wind energy? Will power companies agree to comparable terms for wind versus solar energy? Wind energy cannot be harvested via solar panels, so what will the upfront cost be in building windmills? This project has determined that it is worth considering branching out into wind energy, but there is still much additional research that must be completed before recommendations can be made.

Similarly, this project has completed the preliminary work of determining countries that could be interested in solar and wind energy as they are predicted to have high generation. *Both Japan and Italy are in the top 5 predicted wind and solar consuming countries. I would recommend that my company researches these territories further.* Who are our competitors in these countries? Is there already a framework set up for

solar/wind generation and harvesting that my company could tap into? What are the governmental regulations and incentives for clean energy in these countries? Again, this project has accomplished the task of narrowing down and identifying potential viable territories, but additional research will need to be completed before final decisions are made.

At this point, more research will need to be completed regarding the infrastructure, federal regulations, and practicalities of bringing our company to other countries and branches of clean energy. I would also suggest my company produces its own research and data as the information used in this report is unvalidated data obtained from Kaggle. *Overall, this project completed its objective of identifying potentially viable channels of clean energy and countries that are predicted to have high generation of specific clean energy channels.*