Predicting Solar Panel Effectiveness
Christopher (Will) Schenk
Data Science Project

Introduction:
Last year, I found an extensive amount of data published by the United States Energy Information Agency on every component of the electricity grid in the United States. This project pinpointed solar panels and attempted to apply different techniques learned in class. I discovered three data sets whose intersection provided an opportunity to map monthly efficiencies of all 3,894 large solar power plants in the United States and then applied Logistic Regression and K-Means Clustering to analyze the data I formulated.

Approach:
After searching through many publications, I identified three valuable sets and refined the power plants to just solar, roughly 4,000 out of 25,000 total power plants of all types. The EIA-860 Generators publication provided latitudinal and longitudinal data on each generator (multiple within each power plant). The EIA-860 Plants publication provided plant-level attributes. The EIA-923 publication provided the monthly power output of each plant for the year 2020.

Most valuable, the EIA-860 publication provided a nameplate capacity feature, the electric full-load intended output in megawatts per hour which I used to approximate the maximum of each solar power plant. By taking the monthly electricity generation of each plant, dividing it by the number of days in that particular month, and then dividing it by 24, I obtained a standard hourly output for each power plant per month. By taking these values, which number 48,000, one for each month and each solar plant, and dividing by the nameplate capacity, I obtained an efficiency value and normalized the number between 0 and 1.

To improve this efficiency metric, I worked with solar insolation data, the amount of sunlight that hits the ground at any given time, and the number of sunlight hours in a day for each plant, but after many hours of work, I found my projections to provide inaccurate data, and ultimately decided that continuing this approach would not add value to by project, thus stopping and moving with the original efficiency metric.

Now, I had a data set containing the efficiency for each solar power plant for each month of the year, which I concatenated with locational coordinates, American states, Yearly generation, and the sector name (Utility, local, ect.).

By plotting the longitudinal and latitudinal coordinates of each solar plant on a scatter plot, the map of the United States becomes visible. To prevent a skew in data, I deleted solar plants in Hawaii.

I attempted to use linear regression and logistic regression with various hyperparameters and arrangements but had difficulty developing a useful model. After much time, I found the best to be a Logistic Regression model where I used the efficiency values to predict the American state in which a solar power plant was locate

Following this, I applied K-Means Clustering to group the power plants based both on efficiency and location, and having found success, I continued to zoom into regions populated by large clusters of solar plants.

By focusing the K-Means model on specific regions within the United States, even more valuable trends emerged, which I will explain in the results section.

To improve my model, I worked on predicting the location of solar plants based on monthly efficiencies using a metric less obscure than the American state's feature. Potentially, by forming clusters and assigning each a labeled integer, I could predict the general location based on efficiency. I found this model only provides accurate predictions with overfitting. To find another metric for location, I discovered a tool known as geohashing. This allows the coordinate system on earth to be translated into a bounding box metric. The size of each bounding box is given by 'precision'. The hashes associated with each bounding box can be easily translated into general coordinate positions and plotted on a graph.

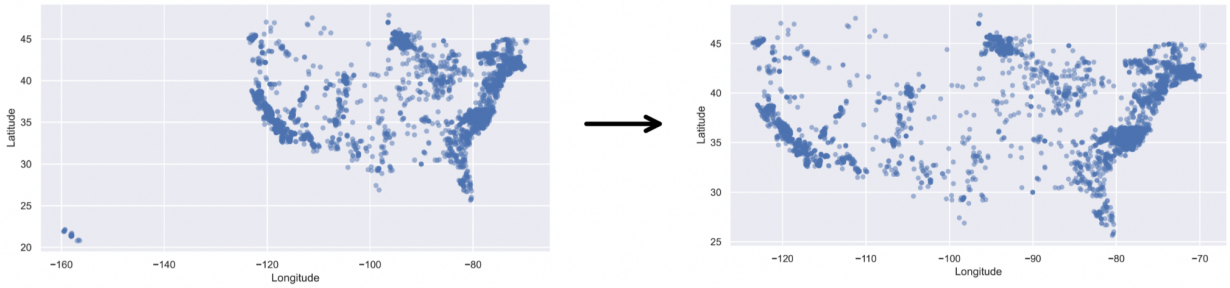
After many hours of researching and testing different geohash libraries and their implementations, I found a library that could efficiently bound my coordinate positions into hashes. I attempted to use logistic regression to predict these bounding boxes based on efficiency but ran into countless syntax and structural errors. This was since geohashes were objects whose components were negatively reactive to slight changes. For this reason, I was unable to apply this logistic regression model.

Instead, I worked with the code to generate a visually compelling heatmap of solar efficiencies.

Results:

Data Collection and Cleaning:

I found very accurate latitudinal and longitudinal coordinates which I refined to the main-land United States.

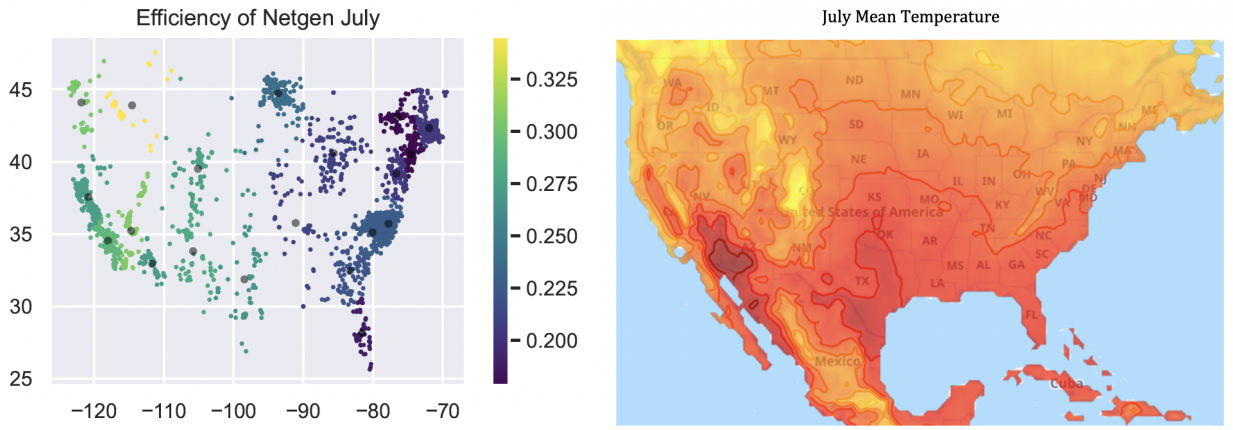


After formulating and normalizing the efficiencies, their statistics were not yet informative since different regions of the United States have very different efficiencies for every month.

	Netgen January	Netgen February	Netgen March	Netgen April	Netgen May	Netgen June	Netgen July	Netgen August	Netgen September	Netgen October	Netgen November
count	3893.000000	3893.000000	3893.000000	3893.000000	3893.000000	3893.000000	3893.000000	3893.000000	3893.000000	3893.000000	3893.000000
mean	0.190989	0.193920	0.162477	0.203352	0.157427	0.199940	0.215451	0.187275	0.179191	0.167531	0.162106
std	0.106693	0.099723	0.075321	0.095084	0.070655	0.085849	0.093345	0.079506	0.072005	0.071138	0.067917
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.133108	0.157391	0.142667	0.167123	0.138095	0.177333	0.183333	0.160511	0.158182	0.134896	0.128054
50%	0.199167	0.203000	0.177333	0.216438	0.166250	0.208315	0.221917	0.190682	0.184416	0.168067	0.163100
75%	0.262500	0.247333	0.203733	0.263836	0.192857	0.243643	0.261084	0.230202	0.211400	0.206349	0.202190
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

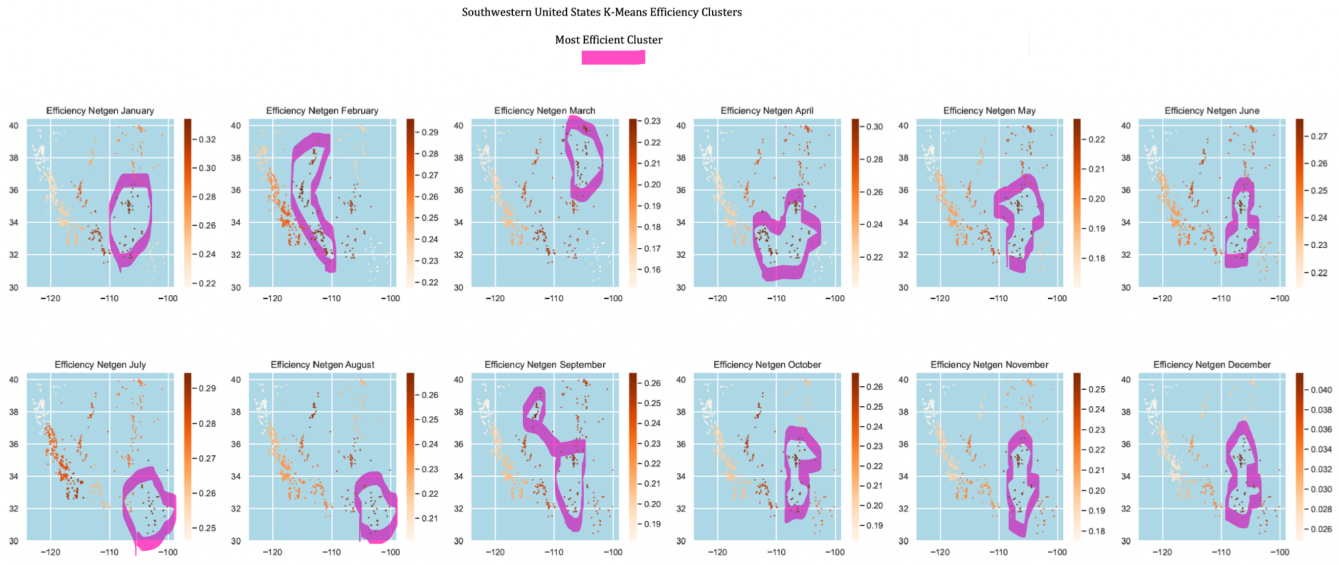
Logistic Regression to Predict American States:
In predicting the states based on efficiency, I was able to obtain an accuracy of 0.464, which I interpret as very good. The geographical configuration of each state is irrelevant and arbitrary to the data my model is built on. Given that I can predict the state of a solar power plant provided only with its monthly efficiency for a year with nearly 50% accuracy, with 48 possibilities (Hawaii and Alaska disincluded), it is very accurate.

K-Means Clustering for main-land United States:



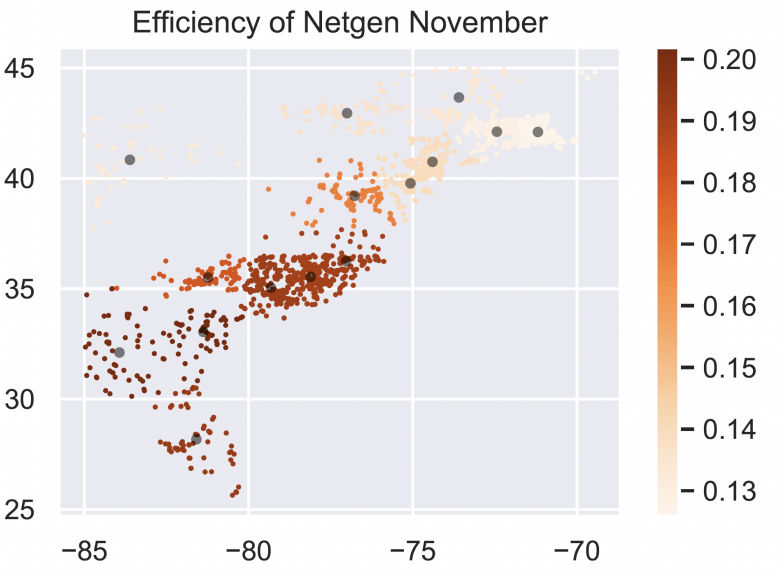
On the left is one of 12 main-land United States K-Means Clustering I created and on the right is a heat map of the mean temperature in July. The most efficient location for solar panels during July appears to be in the Northwestern United States where it is also one of the coolest locations during this month. From comparisons like these it is apparent that temperatures do not directly correlate with solar efficiency.

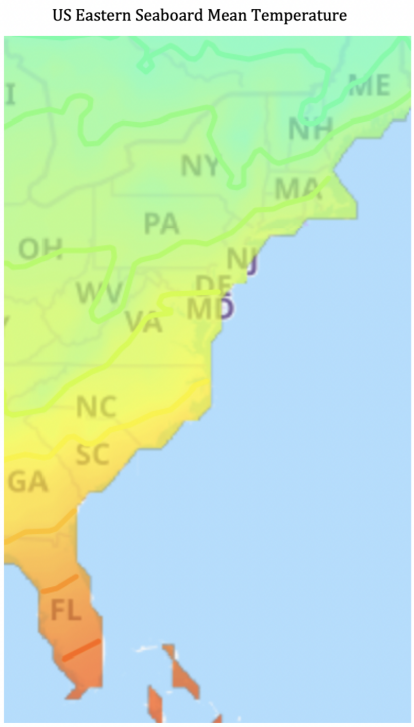
K-Means Clustering for Southwest United States:



K-Means Clustering for Eastern Seaboard:

When applying the model to the eastern seaboard of the U.S. I observed very reasonable projections that mapped exactly to to the temperature map, except for Florida, this is likely due to both Georgia and Florida receiving the same amount of sunlight but since Georgia is cooler, its solar power plants perform for effectively.



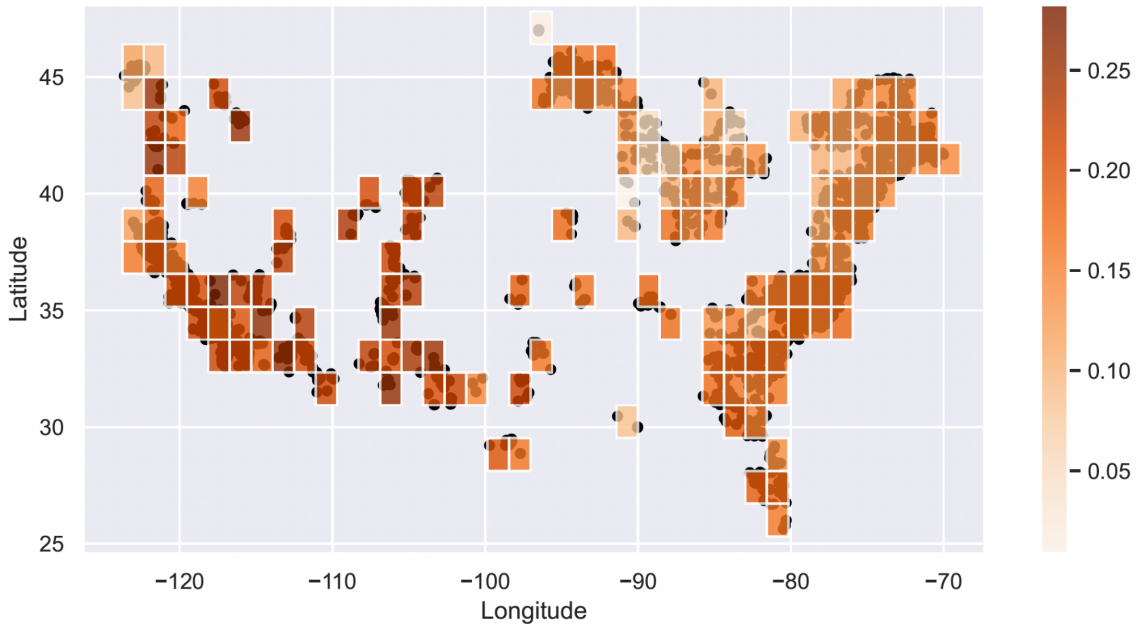


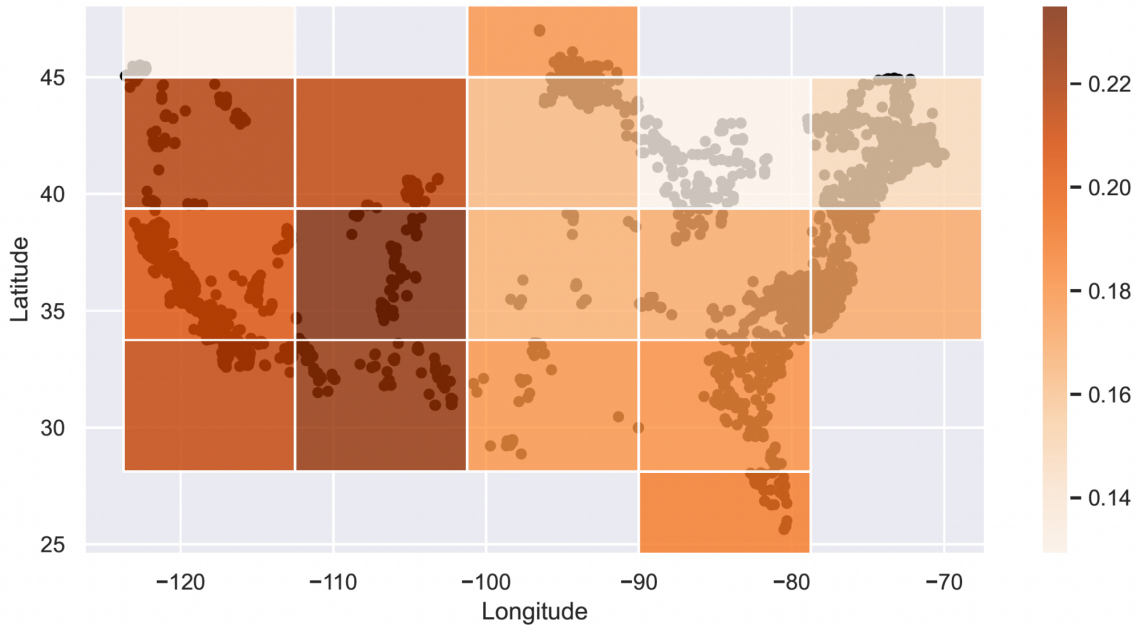
K-Means Clustering Western Seaboard:

The findings are very similar to the previous examples.

Main-Land United States Heat-Map of Efficiencies:
To provide an accurate image, I dropped the samples of all bounding boxes who's box frequency occurred less than 4 times to prevent misinformative boxes. Whild K-Means Clustering is not used here, this provides insight into the efficiencies of nearly all solar power plants.

After seeing this visualization, it is apparent that my formulation of efficiency doesn't capture the variables that make it very common to place solar power plants in the Midwest. The midwest is regarded as a good region due to its cold air and higher solar intensity. I used nameplate capacity as a measure for maximum generation, while in reality, the Midwest's efficiency lies in the fact that it more regularly reaches or surpasses its nameplate capacity, and this variable's use in my formula is thus skewed for this region.





Conclusion:
I aimed my project to provide more data-based insights to climatic effects on solar efficiency. Due to time constraints and coding experience, I was unable to effectively implement weather data into my models. But regardless, my model accurately describes the change in solar efficiency throughout seasons and can accurately predict efficiency.

Acknowledgments:

Provided temperature heat maps for two diagrams in the results section:
https://climatemaps.romgens.com

Provided the geohash framework that I implemented in my code:
https://github.com/vinsci/geohash

Provided functionality to the geohash framework:
https://pypi.org/project/pygeohash/

Power Output Data:
Energy Information Agency Power Output Data:
https://www.eia.gov/electricity/data/eia923/

Energy Information Agency Power Plant Data:
https://www.eia.gov/electricity/data/eia860/
(Generation sheet contains locations, and plant sheet contains power plant attributes)