# Springboard Capstone Project 1 – Statistical Data Analysis

# Predicting Short Term Solar Energy Production

**Connor McAnuff**

**August 15, 2019**

# 1. Overview

This report discusses statistical analysis performed on a subset of the daily energy availability and weather variable data. For the sake of brevity, many of the visualizations created are not included in this document. This report should be read in conjunction with the accompanying Jupyter Notebook.

# 2. Results and Discussion

## 2.1 Correlation matrix

The Pearson correlation coefficient has been calculated for between each weather variable and energy availability (the eventual machine learning target variable) and between each weather variable and all other weather variables. The resulting values can be visualized in a correlation matrix (Figure 1).

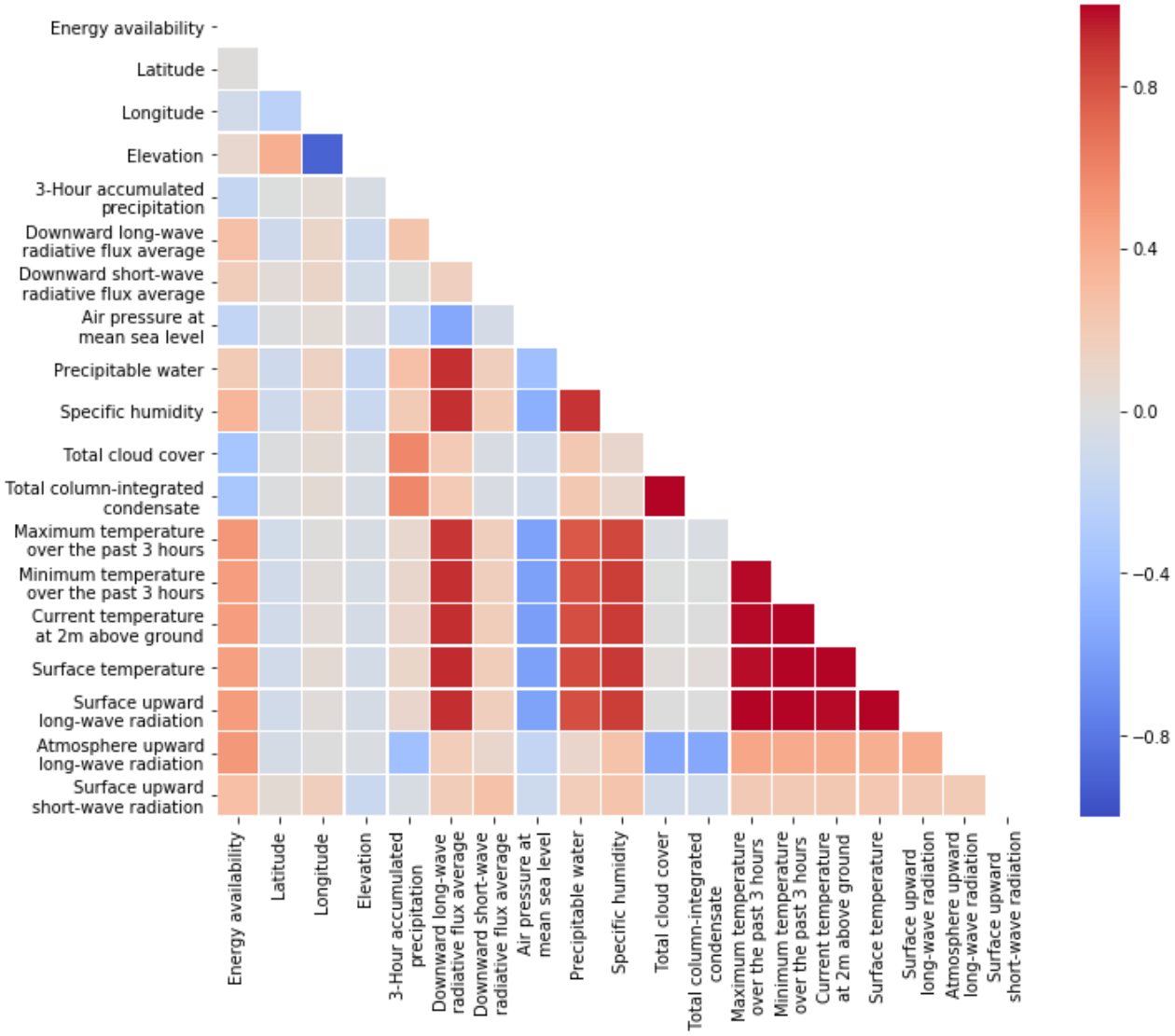


**Figure 1: Correlation matrix (Pearson correlation coefficient) for energy availability and weather variables at forecast hour 0.**

Energy availability is correlated with almost all variables. The weakest correlations are with latitude and elevation. The strong correlations indicate that the variables can be used in a machine learning model to predict energy availability. Between weather variables there are many strong correlations. These correlations will be considered during feature selection.

## 2.2 Bootstrap inference CHER Station July 1994-2007

As shown during exploratory data analysis and data storytelling, the energy availability varies greatly by month. It may be desired by a utility company to estimate the daily and monthly expected energy availability at a certain station for long term planning. Bootstrap inference can be used to determine confidence intervals on the mean daily energy availability for a given month. The distribution of daily energy for all days in July from 1994-2007 at CHER Station is shown in Figure 2.

The distribution is certainly non-normal. There are many values concentrated near the maximum of the distribution, and a long tail of values from the concentration down to 0. 10,000 sets of samples of the same size as the distribution have been taken and the mean of those sets of samples calculated as bootstrap replicates.

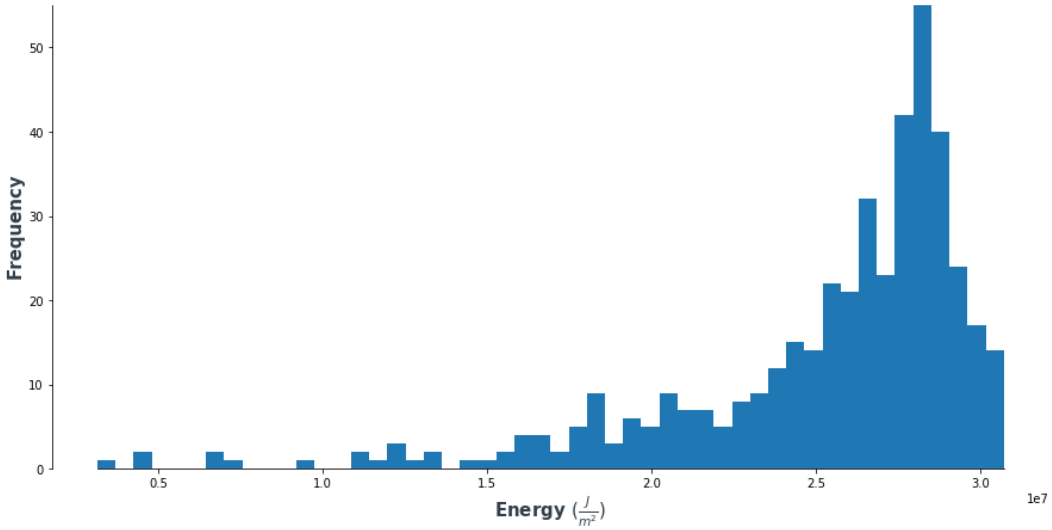


**Figure 2: CHER station July daily solar energy availability 1994-2007**

The distribution of bootstrap replicates of the mean is shown in Figure 3. The 95% confidence intervals represent the boundaries within which 95% of means will fall when July daily energy values at CHER station are sampled with replacement. The confidence intervals could be used be a utility company when forecasting expected daily values and monthly totals.

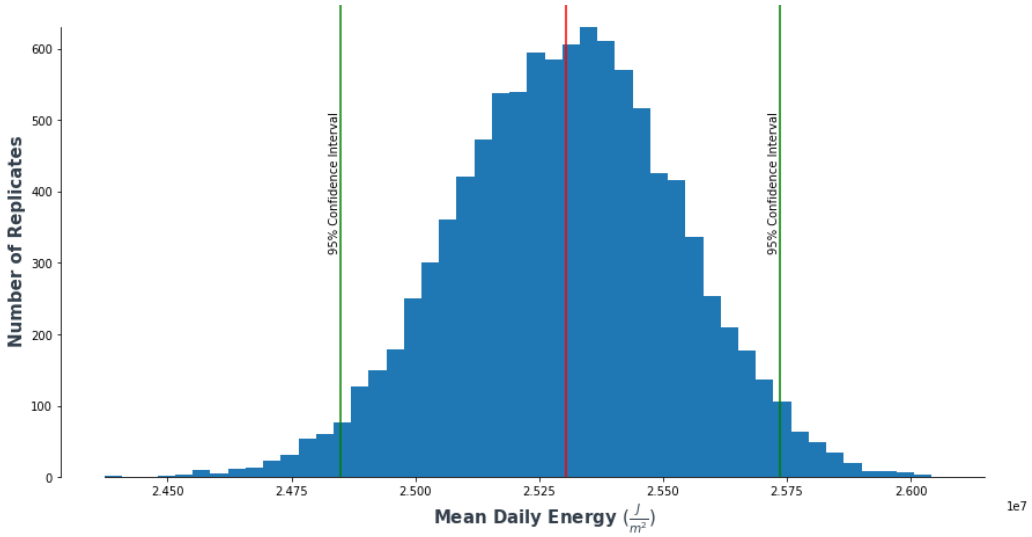


**Figure 3: CHER station July daily solar energy availability bootstrap replicates of mean.**

## 2.3 Bootstrap inference CHER Station January 1994-2007

The same bootstrap inference performed on the CHER station July energy data has been performed on the CHER station January energy data. The distribution of the data is again non-normal (Figure 4).

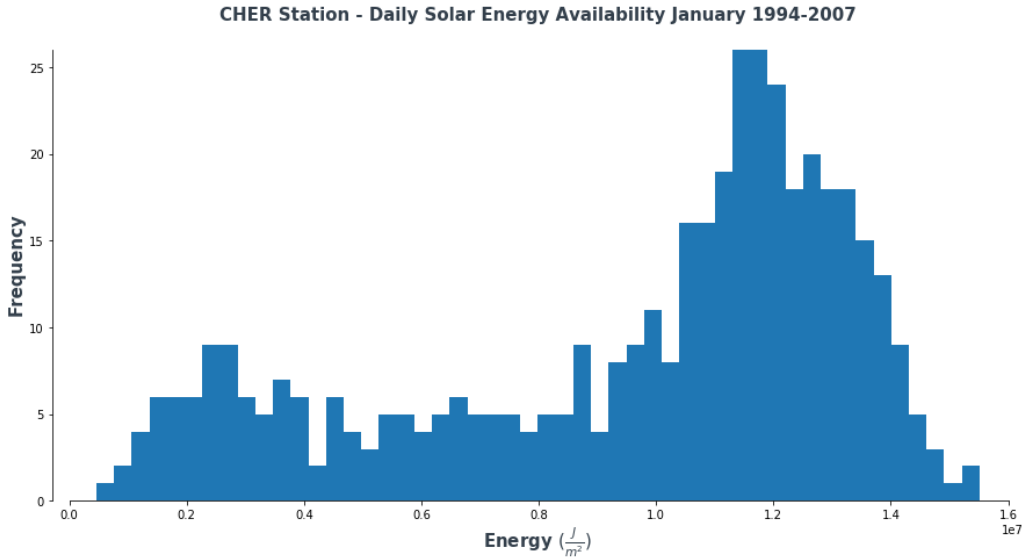


**Figure 4: CHER station January daily solar energy availability 1994-2007**

Figure 5 compares the bootstrap replicate means of July and January. The July means are much greater than the January means, as expected. Repeating this process for all months of the year would provide a utility company with confidence intervals on the expected values of daily energy for each month. Additionally, an estimate of the population standard deviation can be made, which could provide a utility company with information on which months will have the most variation from the mean.

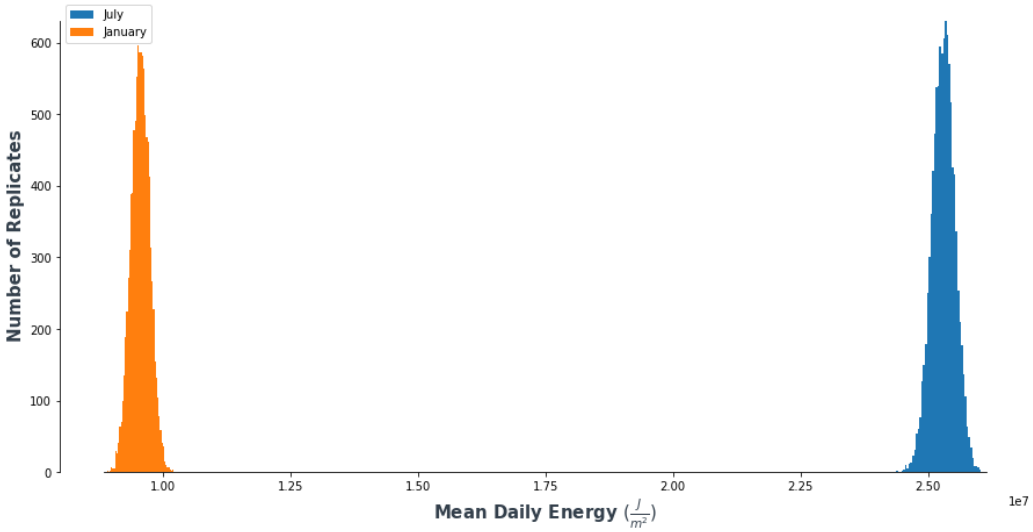


**Figure 5: CHER station January and July daily solar energy availability bootstrap replicates of mean.**

## 2.4 Bayesian inference CHER Station January 1994-2007

Bayesian inference can be used to model distributions of data and determine constraints on parameters of the model. An attempt at modelling the CHER station January energy data (Figure 4) has been made using Bayesian inference. It is proposed that the data could be constructed by sampling from 3 normal distributions with centers at approximately 0.25e7 J/m$^2$, 0.7e7 J/m$^2$, and 1.2e7 J/m$^2$. Pymc3 has been used to model the assignment of data into 3 normal distributions, and the mean and standard deviation of those distributions.

The parameter traces show good convergence for distributions 1 and 3, however they show poor convergence for distribution 2. This result is understandable, as distribution 2 is in between the other two distributions, and thus many means and standard deviations could be possible. Nonetheless, the distributions of posteriors show that distribution 2 was still delimited.

The posterior-mean parameters have been used to visualize the clusters (Figure 6) and also simulate data (Figure 7). It appears that the posterior-mean parameters construct a model that can simulate the data well. This model could be of use to utility companies looking to determine the frequency of differing amounts of energy availability for January at CHER station.

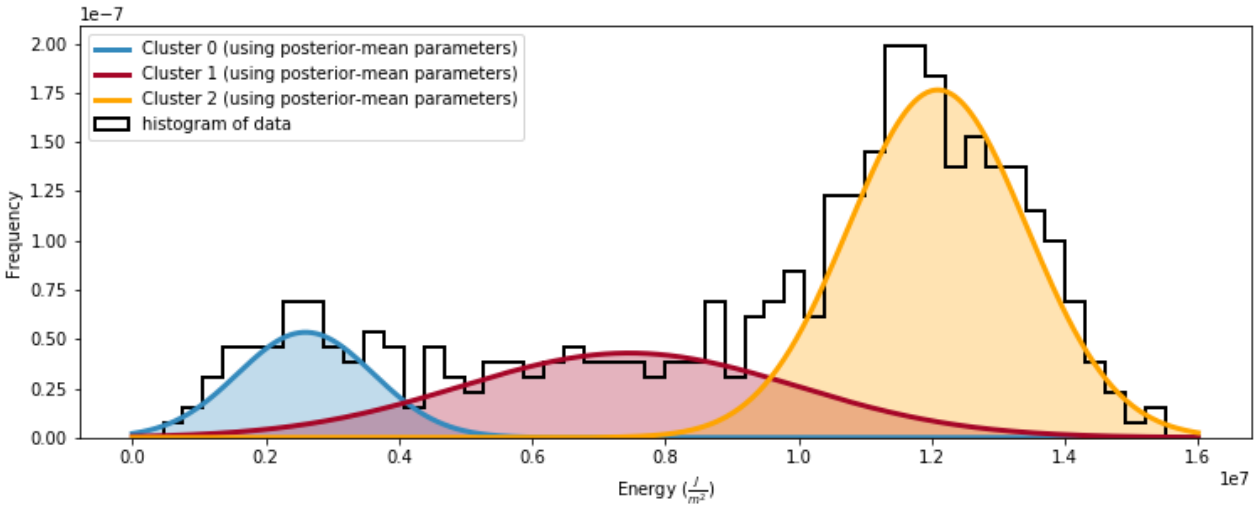


**Figure 6: Posterior-mean parameter clusters visualized alongside a histogram of the data.**

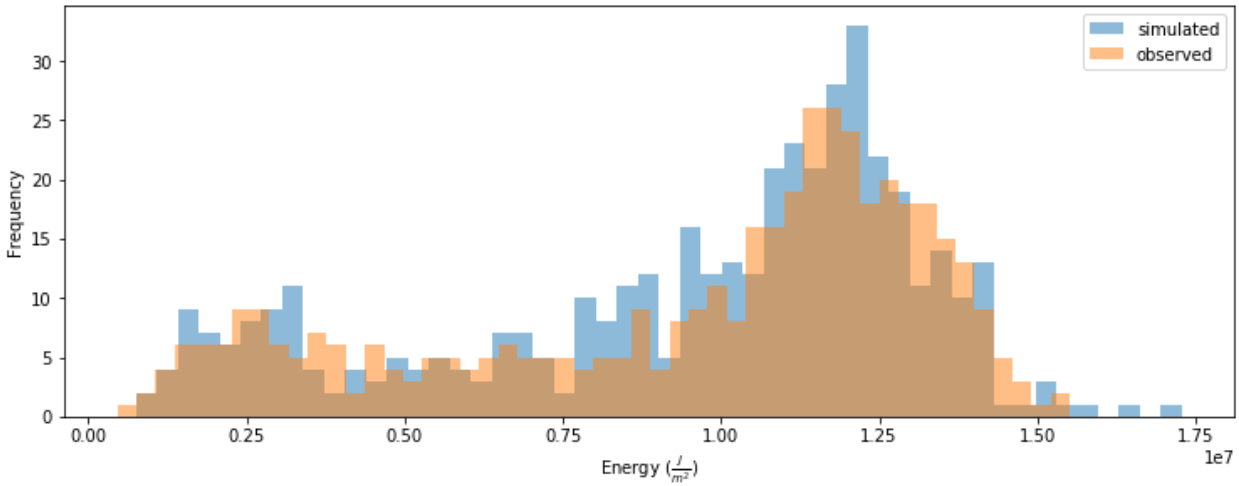


**Figure 7: Data simulated using posterior-mean parameters compared with observed data.**