Final Report:  
Solar Energy Storage Project

**Problem Statement**

Find the best location of the three selected areas in the United States where we can best predict the GHI (Global Horizontal Irradiation), key indicator for the potential of energy storage, with weather indicators of the past 20 years to be able to install solar panels.

**Data Wrangling**

The raw dataset from NASA (hourly weather conditions tracked for the past 20 years) contained 175,320 rows with 16 columns because the data was collected every 60 minutes. After importing the data in different sections because the hourly data had the limitation of importing up to 10 years of data and only 5 variables at a time, so I needed to break up the data into different files and then needed to concatenate and merge the data to make 1 data frame. As I imported the three different locations to analyze, I needed to add a column for the state of the location. I looked at the first few rows, information description and statistical information like maximum, minimum, mean and standard deviation. I noticed I needed to create a new column for time by concatenating year, month, day, and hour. I also examined that the columns with data type int64 were correctly classified. I then checked if there was missing data and when I checked again the maximum and minimum information there were 4 columns with minimum values of -999 which seemed incorrect. Checked which rows where showing this value and it seemed they were N/A values. I then found and replaced -999 with N/A. Most of these values were concentrated in first month of data and the last 3 months so I decided to drop those months.

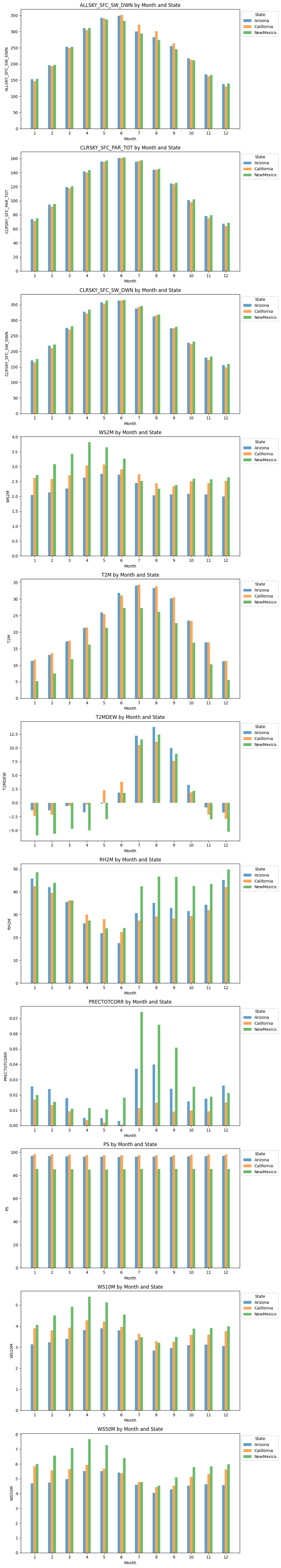
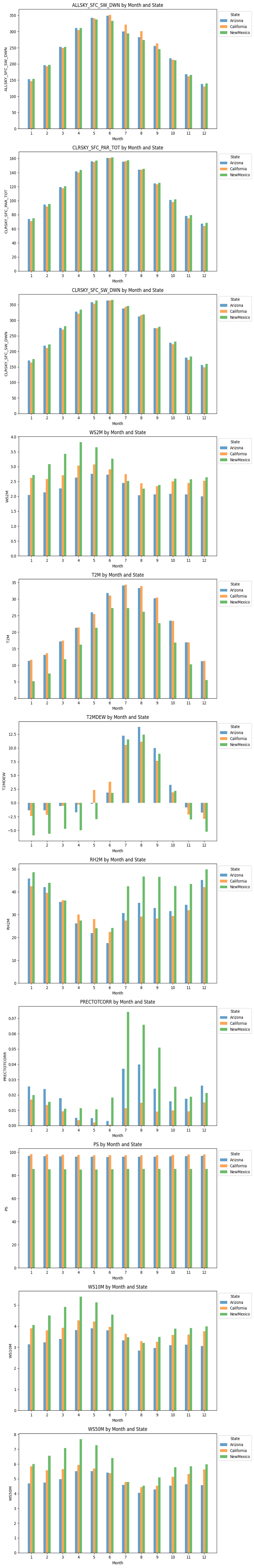
**Exploratory Data Analysis**

A group of green and orange lines

Description automatically generatedA group of green and orange lines

Description automatically generatedI first summarized the data visually of all columns with a rolling daily summary by location.

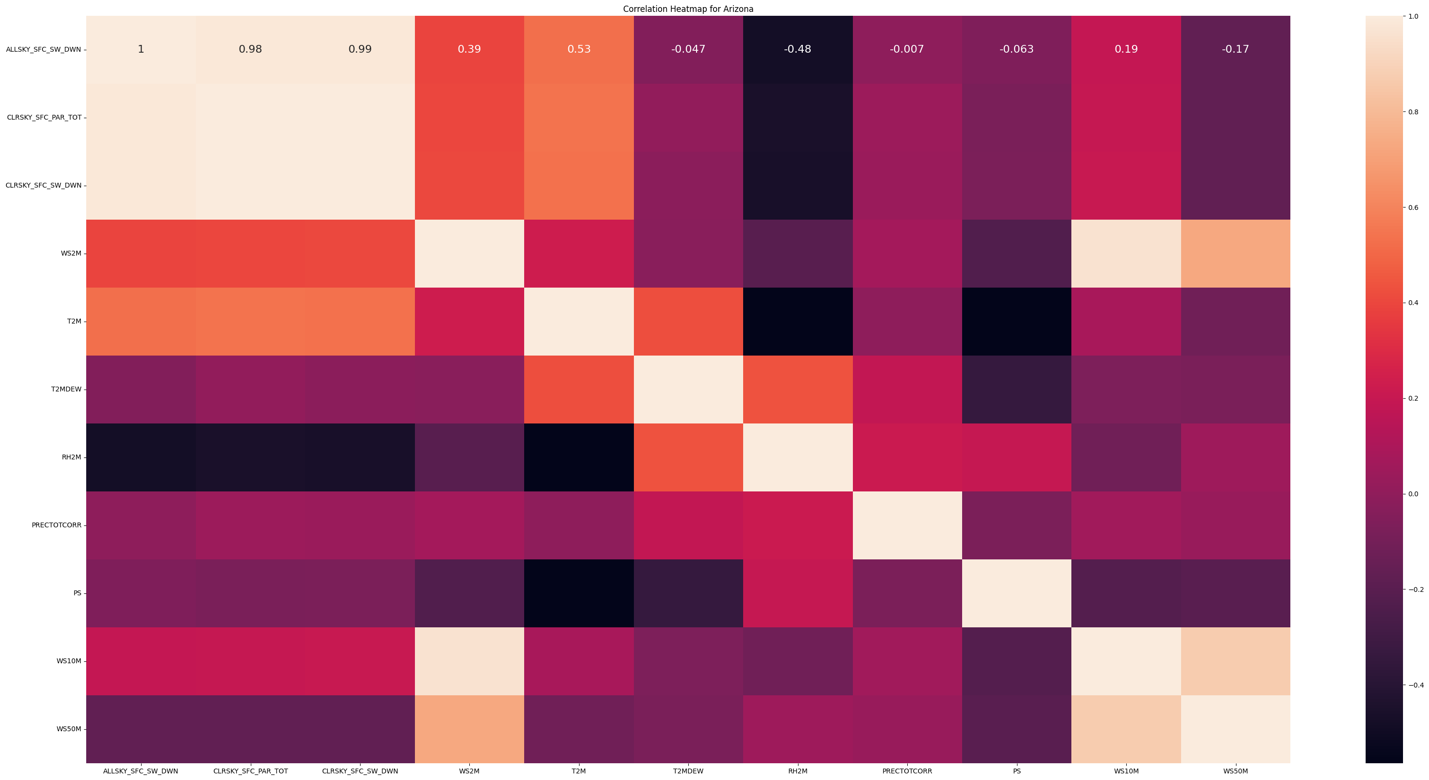
There is a clear seasonality in the data, where the global horizontal irradiation (GHI ) index [represented in this data by All Sky Surface Shortwave Downward Irradiance] peaks mid-year (summer) and troughs of each year (winter). The peaks and troughs in the data are in line with the data since the GHI index is higher during the summer months and troughs in the winter months.

Then I looked at the data by month:

Summarizing these charts:

* In terms of the GHI indices represented by the first 3 variables, the 3 states have a very similar pattern in the past 20 years.
* New Mexico is by far the windiest state of the 3.
* Arizona and California are roughly 5 or more Celsius degrees hotter on average than New Mexico.
* New Mexico has the lowest dew/frost point year-round apart from the months of July, August, September and October where California has the lowest dew/frost point.
* New Mexico is also the most humid every month on average except for May and June where California is the most humid.
* New Mexico also received the most rain in most months of the year, however Arizona receives more rain than New Mexico from December through March.
* Surface pressure is highest for California followed very closely by Arizona.

Then I extended the analysis to all the columns and created a heatmap of correlations by state:



A colorful squares with black and white squares

Description automatically generated

A colorful squares with numbers

Description automatically generated

There lighter colors show a strong positive correlation, and the darker colors show a negative correlation. Other than the 3 first features that are different variations of the GHI index that show a very correlation, almost 100%; the highest positive correlation across all three locations is the GHI index with temperature with correlations in the range of 60%.

Conclusions: Looking through the different information and graphs created, I would like to predict the All Sky Surface Shortwave Downward Irradiance (ALLSKY\_SFC\_SW\_DWN) and drop the other 2 variables that have very high correlations that basically show the same information, those variables are CLRSKY\_SFC\_PAR\_TOT and CLRSKY\_SFC\_SW\_DWN. Then interesting indices arrived from analyzing the charts would be temperature (T2M), wind speed (WS2M) and humidity (RH2M).

**Pre-Processing and Training Data Development**

In this step, I plotted histograms with 10 bins of each feature per state to check the distribution of values. The precipitation feature caught my attention since the majority of the time the feature had values equal to zero so I decided to create a categorical variable for precipitation, it either rains or it doesn’t. Then I split the data into training and testing sets with the test data representing 30%. The following step was to standardize the magnitude of the numeric features using a scaler.

**Model Selection**

I trained and analyzed 2- and 3-layer neural network models for each of the 3 locations: Arizona, California, and New Mexico. Within the Keras package and within each location, I would test using different units (refers to the number of neurons or nodes in that layer) in the dense model and also tweaking the epochs (refers to one complete pass through the entire training dataset).

**Arizona:**

Neural Network 2 layers: [units=64,units=1 & epochs=10]

Train MSE for Arizona (2 layers): 31766.3

Test MSE for Arizona (2 layers): 31935.11

Root Mean Squared Error (RMSE) for Arizona (2 layers): 178.7

r2 score or Arizona (2 layers): 0.7

Neural Network 3 layers: [Units=64, 4, 1 & epochs=20]

Train MSE for Arizona (3 layers): 32673.16

Test MSE for Arizona (3 layers): 32921.79

Root Mean Squared Error (RMSE) for Arizona (3 layers): 181.44

r2 score or Arizona (3 layers): 0.69

**California:**

Neural Network 2 layers: [units=64,units=1 & epochs=30]

Train MSE for California (2 layers): 30612.33

Test MSE for California (2 layers): 30797.69

Root Mean Squared Error (RMSE) for California (2 layers): 175.49

r2 score or California (2 layers): 0.71

Neural Network 3 layers: [Units=64, 4, 1 & epochs=20]

Train MSE for California (3 layers): 35290.72

Test MSE for California (3 layers): 35539.16

Root Mean Squared Error (RMSE) for California (3 layers): 188.52

r2 score or California (3 layers): 0.66

**New Mexico:**

Neural Network 2 layers: [units=64, units=1 & epochs=30]

Train MSE for New Mexico (2 layers): 30397.05

Test MSE for New Mexico (2 layers): 30197.81

Root Mean Squared Error (RMSE) for New Mexico (2 layers): 173.78

r2 score or New Mexico (2 layers): 0.71

Neural Network 3 layers: [Units=64, 4, 1 & epochs=20]

Train MSE for New Mexico (3 layers): 31488.28

Test MSE for New Mexico (3 layers): 31327.87

Root Mean Squared Error (RMSE) for New Mexico (3 layers): 177.0

r2 score or New Mexico (3 layers): 0.7

Below is a slice of the actual data of the GHI of New Mexico versus the predicted values of the 2-layer model. As observed, the values we are trying to predict have high volatility so getting a result of in the range of 0.70 r squared is a pretty good model. And this pattern was observed across all three locations chosen.

A blue and orange lines

Description automatically generated

**Takeaways**

The 2-layer models for all three locations gave very similar results with R squared ranging from 0.70 and 0.71 and RMSE 173.78 to 178.7. New Mexico having the lowest RMSE in both the 2-layer models and the 3-layer models. However, overall, the 2-layer models worked best for this data. A 0.71 for R squared for this hourly data with a high volatility as observed in the charts comparing the predicted and actual values.

The best model for predicting the GHI with weather conditions for a new location was the neural network model of 2-layer of New Mexico considering the MSE and R squared.

Although this model had the best performance, I wanted to do a quick comparison to a simpler, much less complex model as the linear multivariate model. The results below were interesting that the linear model had a slightly better result on one try compared to the neural network model that is more complex and takes longer to train. Although the result is interesting, these results were probably expected since it was a regression problem. I think the neural network model could perform better as we get more data.

**New Mexico:** Linear Multivariate model

Mean Squared Error: 27564.24

r2 score: 0.73

**Future Research**

One interesting feature to include in the neural network model would be to place a constraint for non-negative predicted values. A step to investigate is to place a constraint to ensure that the weights (kernels) of the neurons in the layer are non-negative. This might help to enforce the constraint on the output values but is not guaranteed.