Final Report:  
Renewable Energy and Weather Conditions

**Problem Statement**

How does the changing weather conditions affect energy consumption? Do the changing weather patterns help us predict the renewable energy demand?

**Data Wrangling**

The raw dataset from Kaggle (energy consumption and weather conditions in Poland from 2017 to 2022) contained 196,776 rows with 17 columns because the data was collected every 15 minutes. After importing the data, I looked at the first few rows, information description and statistical information like maximum, minimum, mean and standard deviation. I noticed I needed to adjust the time because the data type was object. I also examined that the columns with data type int64 were correctly classified. I then checked if there was missing data and surprisingly there was not any missing data in any of the columns.

**Exploratory Data Analysis**

I first summarized the data visually, first looking at daily energy usage over time.

A green and blue line

Description automatically generated

There is a clear seasonality in the data, where the energy consumption peaks mid-year (summer) and troughs of each year (winter). The peaks seem to be getting slightly lower each year since 2018, this is demonstrated by the rolling moving average in the chart, the blue line which is the 90-day rolling average. The peaks and troughs in the data are probably due to the fact the electricity is used for frequently in summer for cooling whereas gas is probably used more in winter, to generate heat, probably do to the difference in prices.

Then I looked at the data by month:

A green and white bar graph

Description automatically generated

This reenforcing the seasonality of the data where the lowest consumption is in winter months and the highest consumption is in the summer months. This is very positive to be able to use renewable energy to replace the traditionally sourced electricity.

Next, I looked at the column GHI (Global Horizontal Irradiance) which calculated the amount of solar irradiance received from the sun on a horizontal surface. This is important since it will be the source of energy for the renewable energy receptors, they could be solar panels or mirrors, depending on further analysis of the GHI levels and prices.

A green and blue line

Description automatically generated

Very similar patterns to the energy consumption chart observed above where the highest GHI levels are received during the summer and the lowest during the winter months.

A black and orange bar graph

Description automatically generated with medium confidence

Here is the comparison of the energy usage versus the sun irradiation clearly showing a correlation.

A blue bar graph with red lines

Description automatically generated

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

A colorful squares with white and red squares

Description automatically generated

There lighter colors show a strong positive correlation and the darker colors show a negative correlation. There isn't any other clear high correlations with Energy usage other than Solar Irradiance (GHI) with a 91%.

Here a scatter plot of all the data points available of energy usage and sun irradiation:

A blue and white dotted background

Description automatically generated with medium confidence

The positive correlation is clearly illustrated in this chart.

Conclusions: Looking through the different information and graphs created, I would use GHI (Solar Irradiance) to try to forecast the energy usage to be able investigate the opportunity of producing renewable energy based on weather patterns and expectation of weather patterns. When testing different models, I would also include and try these different variables: temperature, wind speed, snow, rain, and day length because of the results from looking at the different charts and information above.

**Model Selection**

I first split the data into training (80%) and testing (20%) sets. Then I standardized the magnitude of the numeric features using a scaler.

Linear Regression Model

Mean Absolute Error: 243.72

Mean Squared Error: 183462.87

R-squared: 0.83

A screen shot of a graph

Description automatically generated

Here is the residual plot:  
A screen shot of a graph

Description automatically generated

Here are the coefficients (importance) of each variable used:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **const** | -2055.0750 | 142.868 | -14.384 | 0.000 | -2335.092 | -1775.057 |
| **GHI** | 19.5370 | 0.035 | 559.285 | 0.000 | 19.468 | 19.605 |
| **temp** | -3.2259 | 0.278 | -11.619 | 0.000 | -3.770 | -2.682 |
| **pressure** | 1.0074 | 0.116 | 8.662 | 0.000 | 0.779 | 1.235 |
| **humidity** | -3.5781 | 0.110 | -32.584 | 0.000 | -3.793 | -3.363 |
| **wind\_speed** | -5.8647 | 0.629 | -9.324 | 0.000 | -7.097 | -4.632 |
| **rain\_1h** | 26.4366 | 4.121 | 6.415 | 0.000 | 18.360 | 34.513 |
| **snow\_1h** | -120.3970 | 14.122 | -8.525 | 0.000 | -148.076 | -92.718 |
| **clouds\_all** | -0.6507 | 0.071 | -9.179 | 0.000 | -0.790 | -0.512 |
| **isSun** | -103.0316 | 5.255 | -19.608 | 0.000 | -113.330 | -92.733 |
| **sunlightTime** | -1.2730 | 0.017 | -74.455 | 0.000 | -1.306 | -1.239 |
| **dayLength** | 1.9275 | 0.103 | 18.790 | 0.000 | 1.726 | 2.129 |
| **SunlightTime/daylength** | 862.3723 | 14.646 | 58.883 | 0.000 | 833.667 | 891.077 |
| **weather\_type** | 8.9970 | 2.134 | 4.216 | 0.000 | 4.815 | 13.179 |
| **day sin** | 27.4009 | 2.179 | 12.576 | 0.000 | 23.130 | 31.671 |
| **day cos** | -49.4983 | 3.156 | -15.683 | 0.000 | -55.684 | -43.312 |
| **year sin** | -106.5517 | 5.319 | -20.031 | 0.000 | -116.977 | -96.126 |
| **year cos** | 578.5802 | 28.059 | 20.620 | 0.000 | 523.586 | 633.575 |

By far the most important variable is GHI (solar irradiation), in a next step of importance the variables of sunlight time and humidity were also significant compared to the other variables studied. The r squared was 0.83 with MAE 244 and MSE 182,463.

Random Forest Model

Mean Absolute Error: 104.77

Mean Squared Error: 68848.37

R-squared (R2) Score: 0.94

A green line with dots

Description automatically generated

A white and blue rectangle with black border

Description automatically generated

The graph of the actual versus the estimated values of the Random Forest model clearly shows a tighter regression result. This is corroborated with a higher r squared of 0.94 and lower MAE of 105 and MSE of 68,848 compared to the linear regression model above. The feature importance is once again repeated where GHI is the most significant variable followed but a fraction of importance the length of the day.

LSTM (Long Short-Term Memory) model

MSE 167313.45

MAE 172.22

R2 0.88

A graph of a graph

Description automatically generated with medium confidence

The LSTM model provided better results than the linear model but worse than the Random Forest model with an r squared of 0.88, MAE 172 and MSE 167,313.

**Takeaways**

The best model for predicting the energy consumption was the Random Forest Model considering the MSE, MAE and R2 results with GHI (solar irradiation) having the strongest importance (this was a common result of all the models trained and tested). This is very positive because the renewable energy industry utilizes exactly the information of the amount of solar irradiation of a place to verify that a renewable energy plant is possible. This is because they tend to want to have the storage of the energy close to where the energy will be used so that the energy loss in minimal. If the energy stored must travel a long distance, there is always a probability of losing a percentage of that energy.

**Future Research**

Next steps would be to look at other countries or areas and corroborate the model results where there is a very strong relationship between energy consumption and solar irradiation or GHI. Because it is a very positive outcome (almost too good to be true), where the indicator of GHI is a very important indicator that is used by renewable energy companies to decide if solar panels or mirrors would work in certain areas of the world. It provides the information of the capability of energy storage that later can be used as electricity or heat, as needed.

Also, I'd like to explore further how the data is affected by looking at the subgroups of weather type and try and get more information from the data source provider as I was not able to figure out the difference between the five weather types provided in the dataset.

Although the Random Forest model had the best performance, I think that given the strong seasonality of the data the best model could be a SARIMA model that I wasn't able to run successfully. Below is a chart that reinforces my opinion of the strength of the seasonality. We would need to strip away the seasonality to find the true trend.

A group of blue lines

Description automatically generated