Springboard
Capstone 02 Project Report
Kera Lagios
October 19, 2024

Predicting Energy Prices based on energy and weather data in Spain from 2015-2018

Introduction

Energy markets and weather are fundamentally related. First, weather events like heat waves and cold snaps cause increases in energy demand. Secondly, factors like clear sunny skies and high wind speeds increase renewable production of solar and wind energy, and the opposite causes that production to decrease. Both of these things can affect the price of energy. There are many entities that benefit from understanding and predicting everything from anticipated loads (demand on the system), price and even anticipated generation of renewables. Among those who could benefit from accurate models of energy and weather data are the general public, energy suppliers, cities and municipalities, and finance professionals who trade on the energy markets. This study examined four years (2015-2018) of hourly energy and weather data in Spain and created a model to predict price.

1.0 Data

Data was downloaded from Kaggle in the form of two CSV files. The data was compiled from several independent sources.

1.1 Citation

Kolasniwash, 2019: “Hourly energy demand generation and weather”. Kaggle, accessed August 17, 2024, https://www.kaggle.com/datasets/nicholasjhana/energy-consumption-generation-prices-and-weather

Cited sources of data:

- https://transparency.entsoe.eu/
- https://www.esios.ree.es/en/market-and-prices
- https://openweathermap.org/api

1.2 Description

The dataset is comprised of two CSV files: “energy_dataset.csv” and “weather_features.csv”. The data was cleaned and some features were added. At the end the data was grouped according to day and the maximum value was taken.

1.2.1 Energy data

The energy dataset contains hourly energy data from Spain from the midnight preceding January 1, 2015 until midnight December 31, 2018. The weather dataset contains hourly weather data from the midnight preceding January 1, 2015 until midnight December 31, 2018, for five cities in Spain: Barcelona, Bilbao, Madrid, Seville and Valencia.

The energy data set contains generation data for various energy sources such as fossil hard coal, fossil oil, fossil gas, solar, wind onshore in megawatts (MW). It also contains price and price day ahead,

forecast solar day ahead, forecast wind onshore day ahead, forecast wind offshore day ahead, total load forecast and total load actual. Price is represented in Euros. It should be noted that “load” is the measure of demand.

1.2.1.1 Missing energy data

The energy data set had several features that were either missing data (entries were NaNs) or had only zeros for entries. Those were as follows:

1. generation fossil coal-derived gas
2. generation fossil oil shale
3. generation fossil peat
4. generation geothermal
5. generation marine
6. generation wind offshore
7. generation hydro pumped storage aggregated
8. forecast wind offshore day ahead

These features were removed from the dataset.

1.2.2 Weather data

The weather data had numerous features including temperature, humidity, wind speed and wind direction. Units vary depending on the feature. Temperature is represented in Kelvin, but was converted to Celcius. For temperature, there were actually three features: temp, temp_min and temp_max. Experiments were conducted to determine whether “temp” was the average of “temp_min” and “temp_max” which it was not. After consulting the Open Weather API documentation, “temp_min” and “temp_max” represented the minimum and maximum temperature across a large city or megalopolis, and it was advised to use them optionally.[1] Ultimately, “temp_min” and “temp_max” were dropped.

There were also several categorical features: weather_id, weather_main, weather_description, and weather_icon. Due to the quantity of unique values for each of these features, and the added time required to process them, it was decided to remove these features.

When the weather data was grouped and aggregated by day using the maximum value, the wind direction was removed because it would not be accurate to take the maximum of the wind direction as the summarized value.

1.2.2.1 Missing weather data

The weather data set had no missing data

1.3 Feature engineering

From the date information, a feature for the month and the day of the week were added. Eventually these were converted to dummy/indicator variables. In addition, a total generation (“gen_total”) was created by summing all of the generation features, and the difference between total generation and load was also added, although the latter was eventually dropped.

The two datasets were joined on the date column.

[1] https://openweathermap.org/current#list

2.0 Methodology

One of the first issues to present itself was the fact that the energy data set had one observation per hour, while the weather dataset had five observations per hour, one for each of the five cities. This means that when the datasets were joined, the energy data was repeated for each of the five cities.

In part to simplify the problem, the data was grouped by day or by day and city_name, in the case of the weather data, and the max value for each day was derived. Experiments were conducted using the mean and the minimum as the aggregating function, but it did not appear to have a noticeable effect on correlations of the features.

It was also determined that although the data has a date/date-time column, this was not a time-series problem. For this study, the goal was to predict price based on "relationships between features", rather than involving "temporal dependencies that can handle trends and seasonality."[2]

It was also determined that this would be a regression problem.

3.0 Analysis

3.1 Trends in weather

In order to gain some basic insight into the weather data in general, and also how it compares across cities, several plotting methods were tried including lineplots, windroses and aggregated box plots. One challenge was just that the quantity of data made plots hard to read, and therefore hard to visually derive trends from. Eventually, two sets of plots were produced (below). The first is the average month-year data for each of the weather features (i.e. the average value for each feature, by city, for the month-year was plotted). Even at this scale, it is still hard to discern overall patterns and trends. If the process were to be repeated, it would be better to give more space to each plot to expand the data.

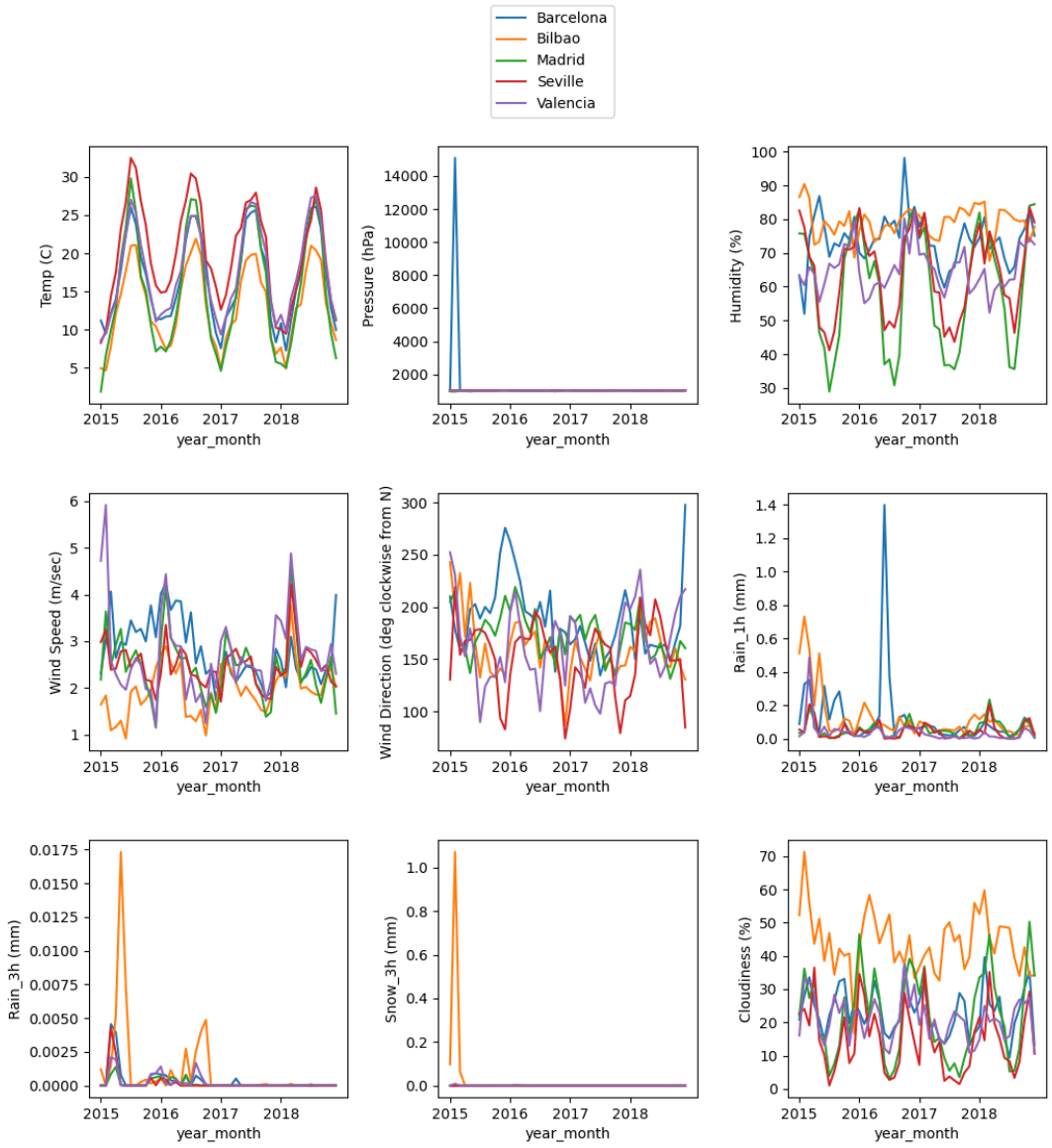
As a result of the low-legibility of the month-year plots, a second, similar plot was created which simply shows the average per year for each feature, by city. This is much easier to read, but obviously lacks a lot of the important data.

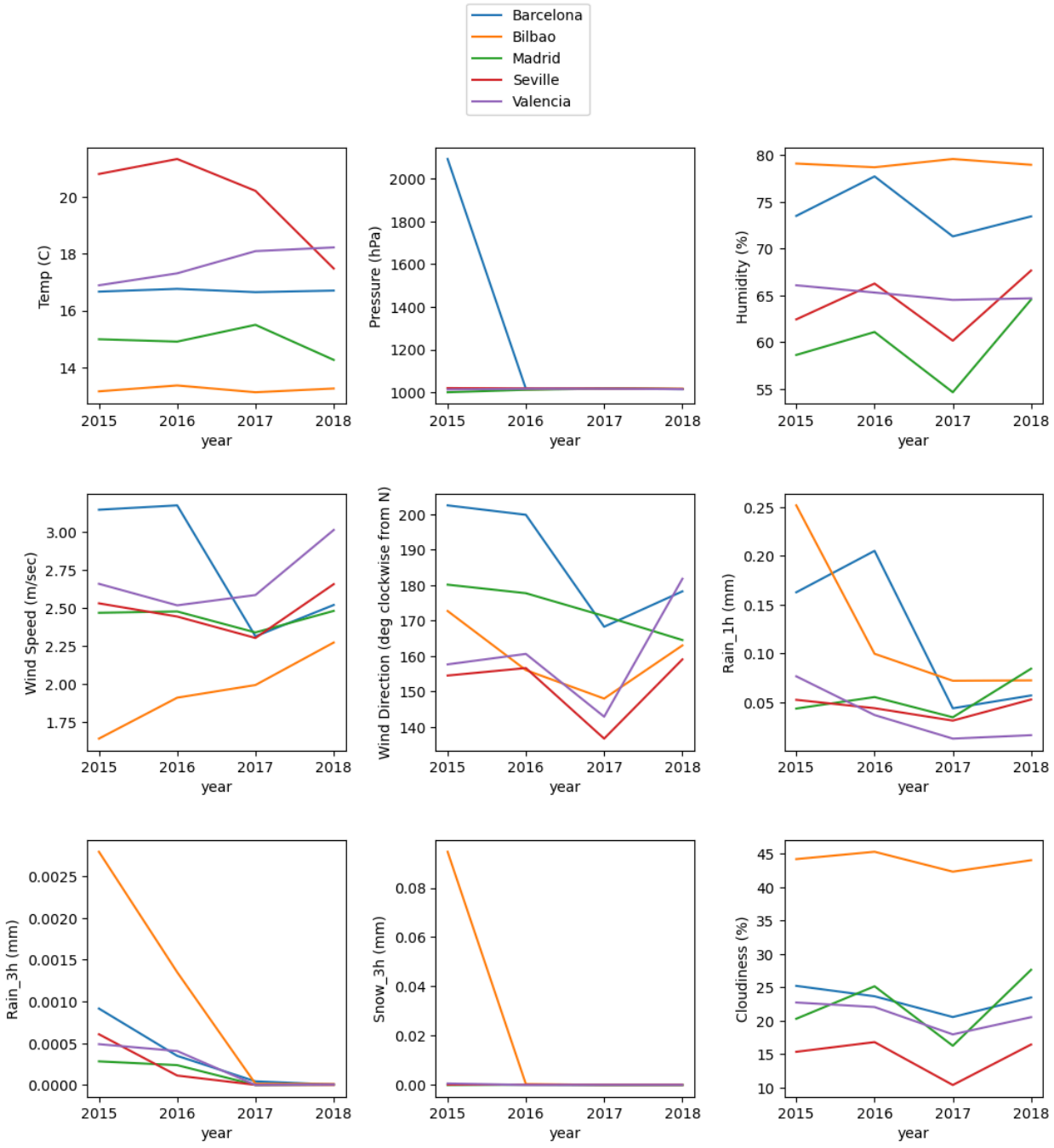
The data is consistent with general knowledge of the climate in Spain. Seville has the highest temperatures and lowest humidity and cloudiness, which makes sense for the sunny climes of Andalusia. Bilbao has the lowest temperatures, highest humidity and highest cloudiness which is also characteristic of the northwest of Spain. The pressure feature appears to possibly have an outlier or an error in the Barcelona data. And it appears that 2017 and 2018 were very dry.

It was interesting, given what we know about climate change, that there is not an increase in temperatures over the four years. In reality, it appears to have decreased from 2015-2018. Of course, these plots only show averages, so heatwaves and cold snaps are likely not visible.

---

[2] Dishant Salunke, "Understanding Time Series vs. Non-Time Series Problems," Medium, August 6, 2024. https://medium.com/@dishant.salunke9/understanding-time-series-vs-non-time-series-problems-2681e5656305#:~:text=Understanding%20whether%20your%20problem%20is,and%20achieve%20more%20accurate%20predictions.


Barcelona
Bilbao
Madrid
Seville
Valencia
Temp (C)
Pressure (hPa)
Humidity (%)
Wind Speed (m/sec)
Wind Direction (deg clockwise from N)
Rain_1h (mm)
Rain_3h (mm)
Snow_3h (mm)
Cloudiness (%)
year_month
2015
2016
2017
2018

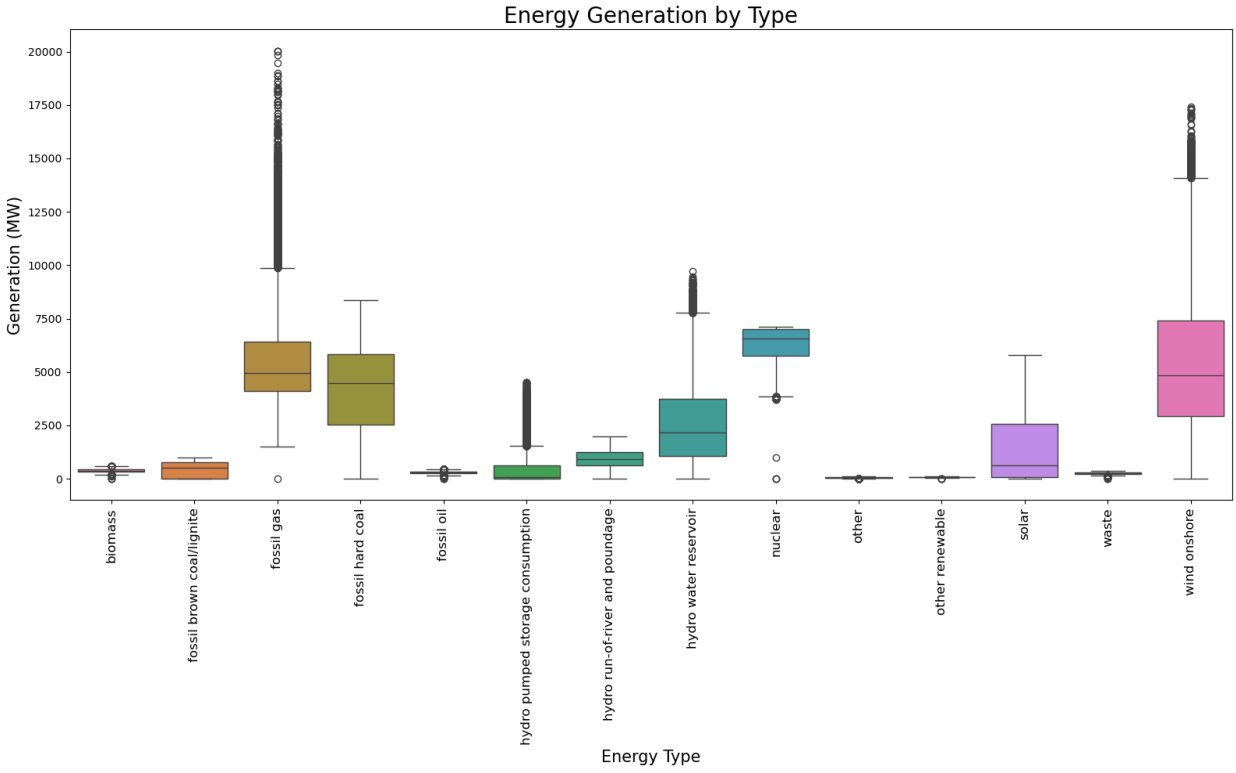


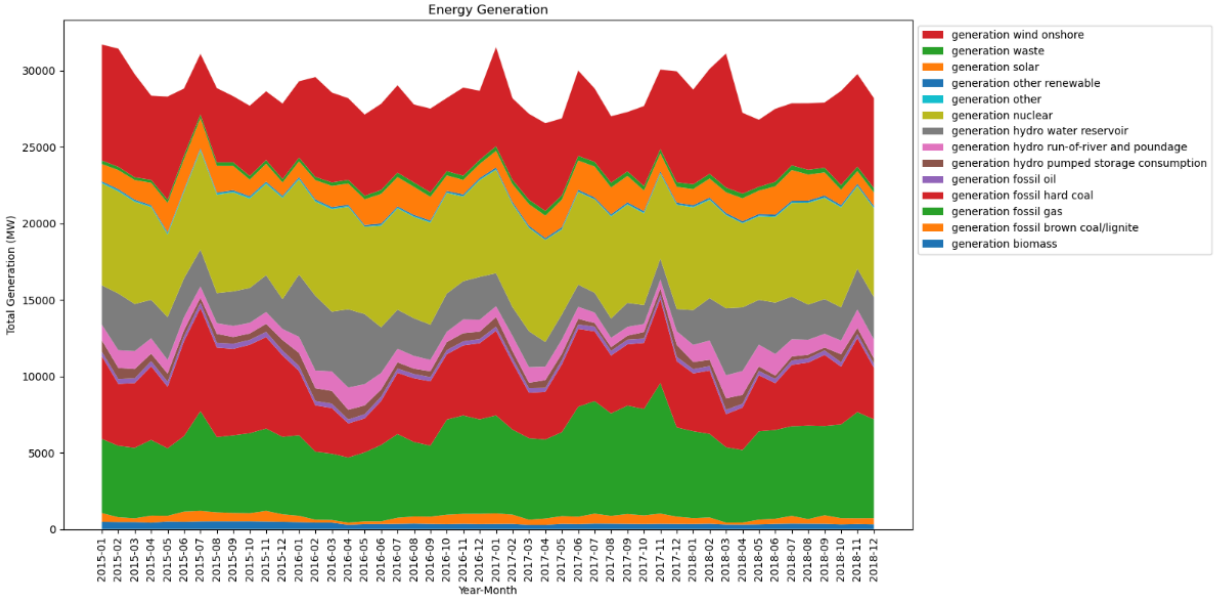
3.2 Trends in energy

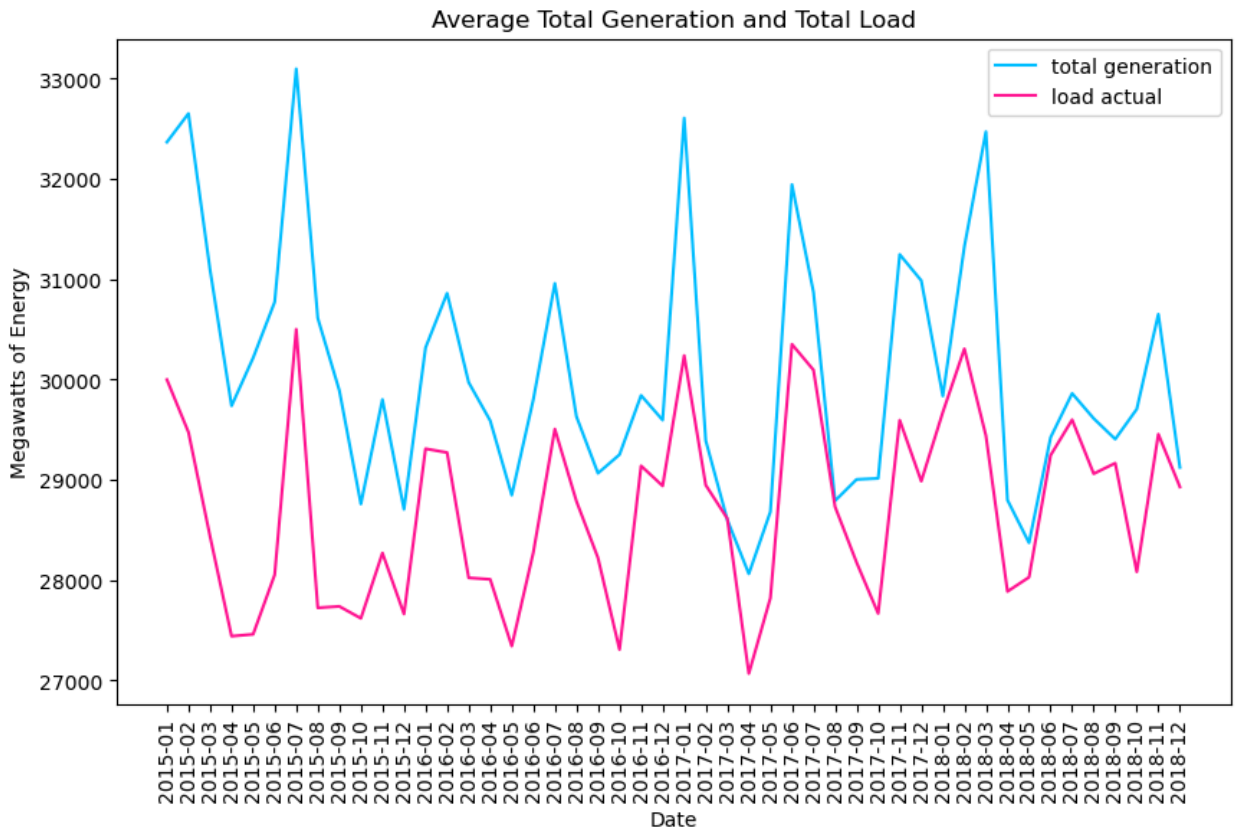
One of the first questions to answer about the energy data is: what are the biggest sources of energy in Spain. From the boxplot below, the five highest median production values for energy (in megawatts) are: nuclear, fossil gas, wind onshore, fossil hard coal, and hydro water reservoir.

By creating a stacked bar plot for the average energy generation by energy type for the four years, the same five energy sources are the greatest contributors to the overall energy generation. Although there are peaks and valleys, the generation and mix of energy sources stays roughly constant.

A plot of average load and average total generation. This shows, as it should that generation is greater than the load. The opposite would mean power outages and blackouts. There are a few times when the two are very close to each other (March 2017, August 2017 and June 2018), but because these are averages, it is not possible to know from this plot whether there was an actual problem during these times - one would have to look more closely at the monthly data during those times.







3.3 Price

Another feature of interest is “price actual”, as well as “price day ahead”. According to the website ISO New England, “The Day-Ahead Energy Market lets market participants commit to buy or sell wholesale electricity one day before the operating day, to help avoid price volatility.”[3] A plot of the average monthly values for each feature was plotted.

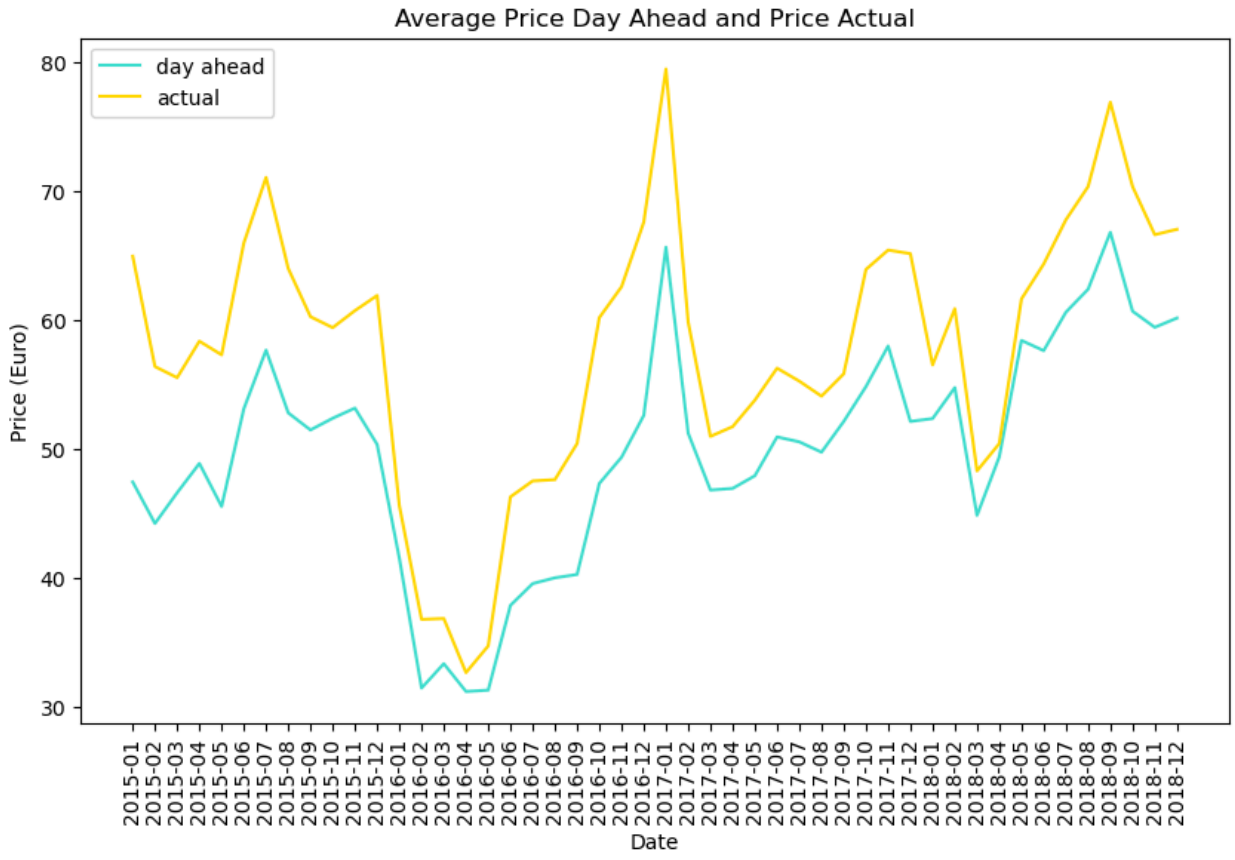
Actual price per megawatt ranges from 9.33 to 116.80 Euros, with a mean of 57.88. Price day ahead is generally lower than price actual, ranging from 2.06 to 101.99 Euros, with a mean of 49.87.

Summary statistics for Price Actual and Price Day Ahead (Euros/MW)				
	Mean	Stand. dev.	Min	Max
Price actual	57.88	14.20	9.33	116.80
Price day ahead	49.87	14.62	2.06	101.99

Energy prices appear to have a general upward trend over the course of the four years. The end of 2015 and the first half of 2016 show a noticeable lower price. According to the website Aleasoft, this was, “mainly due to the increase in electricity production with less expensive technologies such as

[3] ISO New England, “Day-Ahead and Real-Time Energy Markets,” 2024. https://www.iso-ne.com/markets-operations/markets/da-rt-energy-markets

hydroelectric, wind and nuclear power, as well as the decrease during that period in the prices of fossil fuels used in the electricity generation."[4]



3.3 Correlations

Several plots were created to evaluate correlations between features. A heatmap was generated of the correlation coefficients, and individual plots were created for several features, showing the sorted correlation coefficients for each feature.

Cities are loosely correlated with weather data, but not with anything else. Saturday and Sunday are negatively correlated with the load (forecast, actual), price (day ahead, actual), and total generation, as well as a few of the generation features.

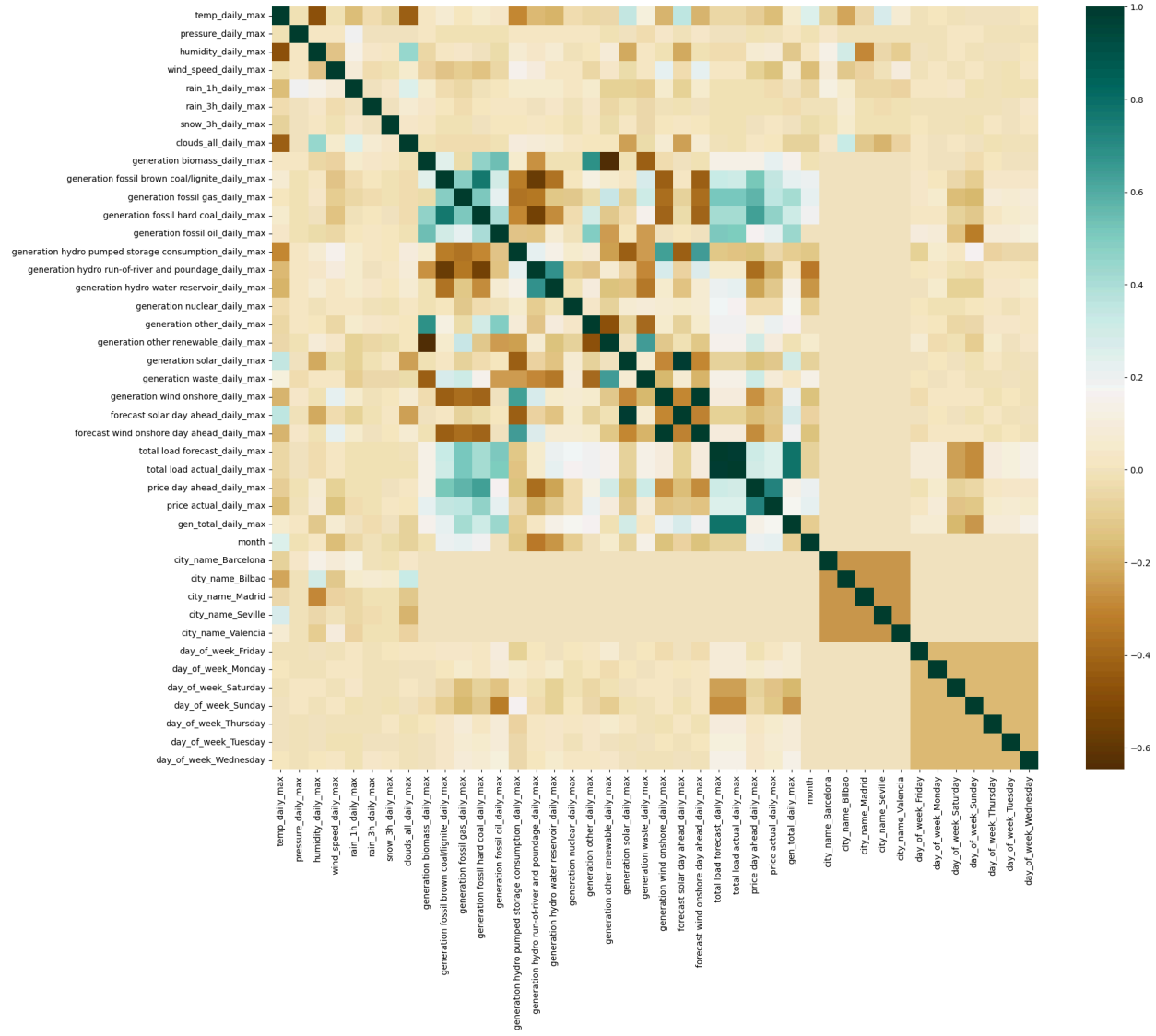
Price day ahead and total load forecast are highly correlated which stands to reason. Price actual is most strongly positively correlated with fossil brown coal/lignite, fossil gas and fossil hard coal, and most negatively correlated with generation hydro run-of-river poundage. Price day ahead has slightly higher correlations.

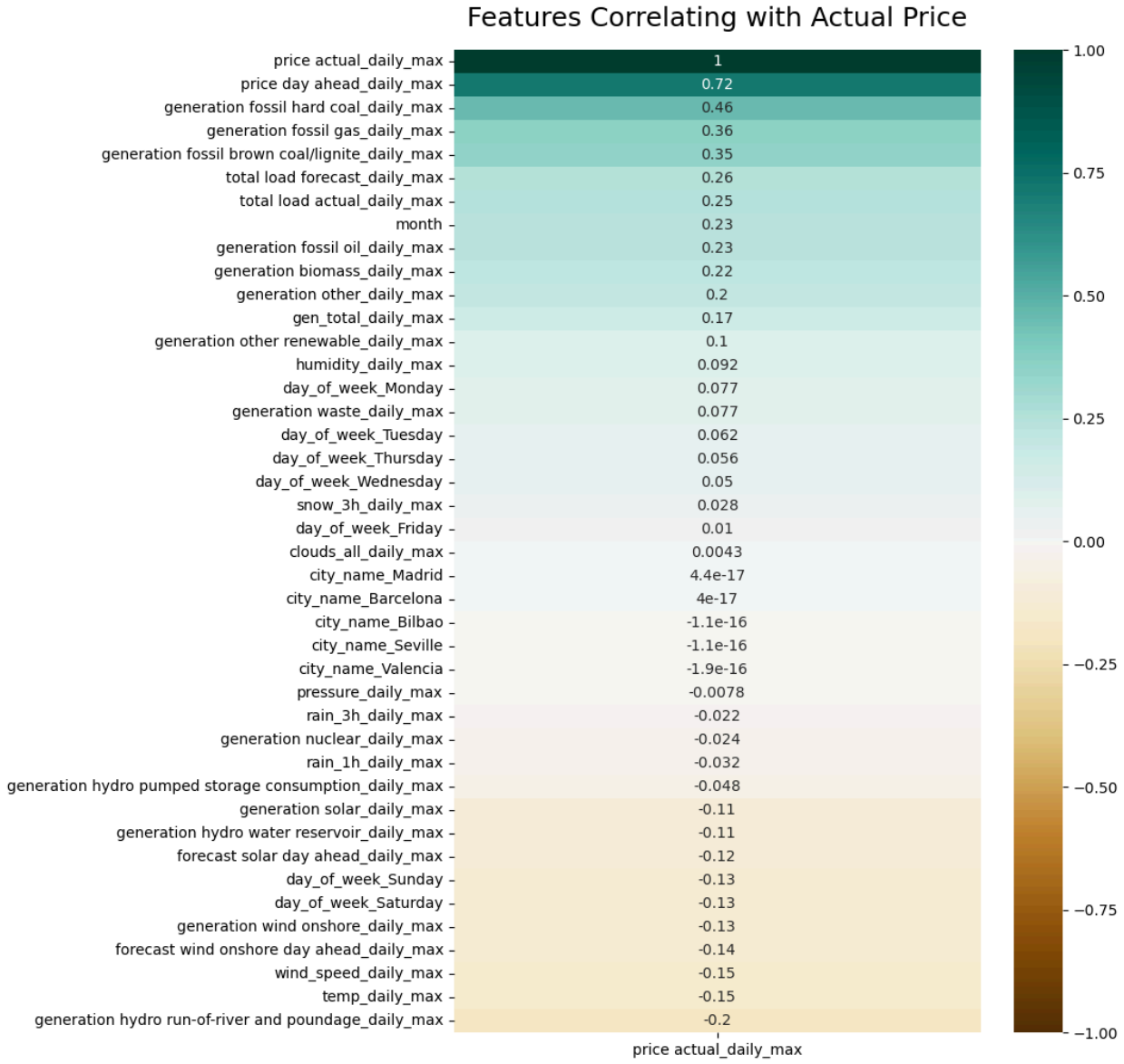
Wind speed is correlated, but not strongly with generation wind onshore, and forecast wind onshore day ahead. Temperature is correlated, but not strongly, with generation solar and forecast solar day ahead. Both of these things were very surprising and merit further investigation.

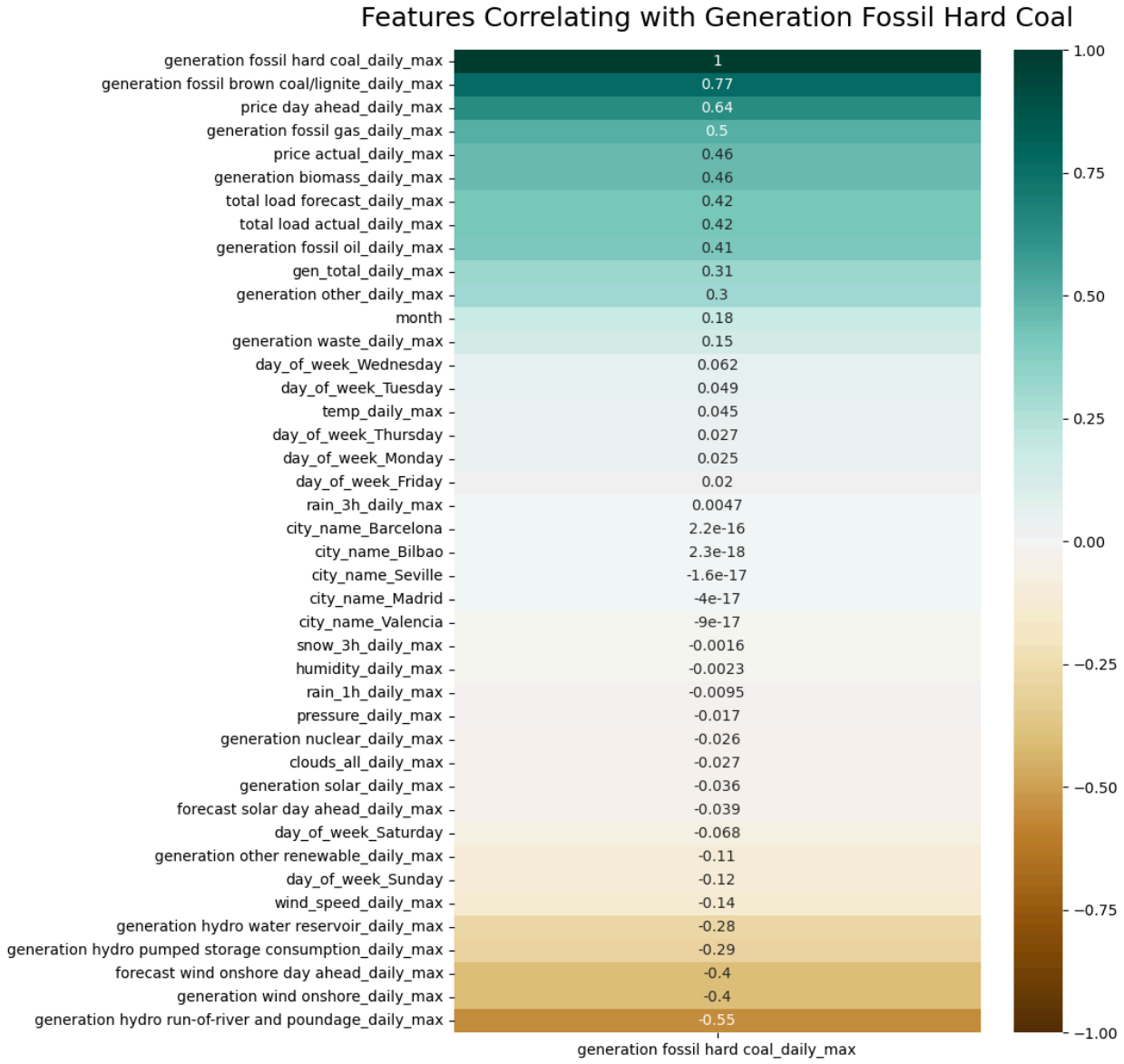
---

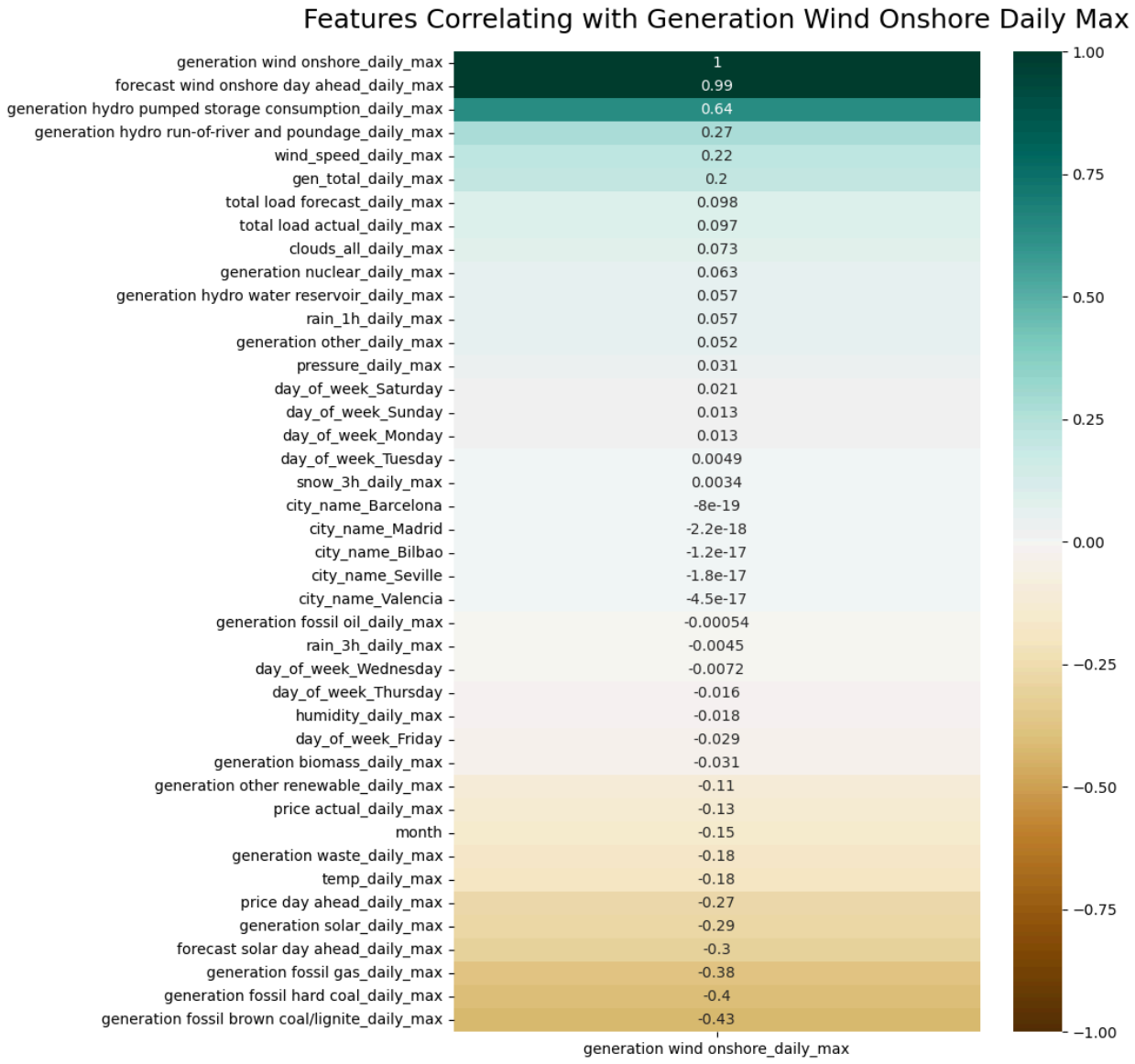
[4] Aleasoft, "Analysis Of The Spanish Wholesale Electricity Market During 2016," January 11, 2017, https://aleasoft.com/analysis-spanish-wholesale-electricity-market-2016/

In general, correlations were relatively weak.

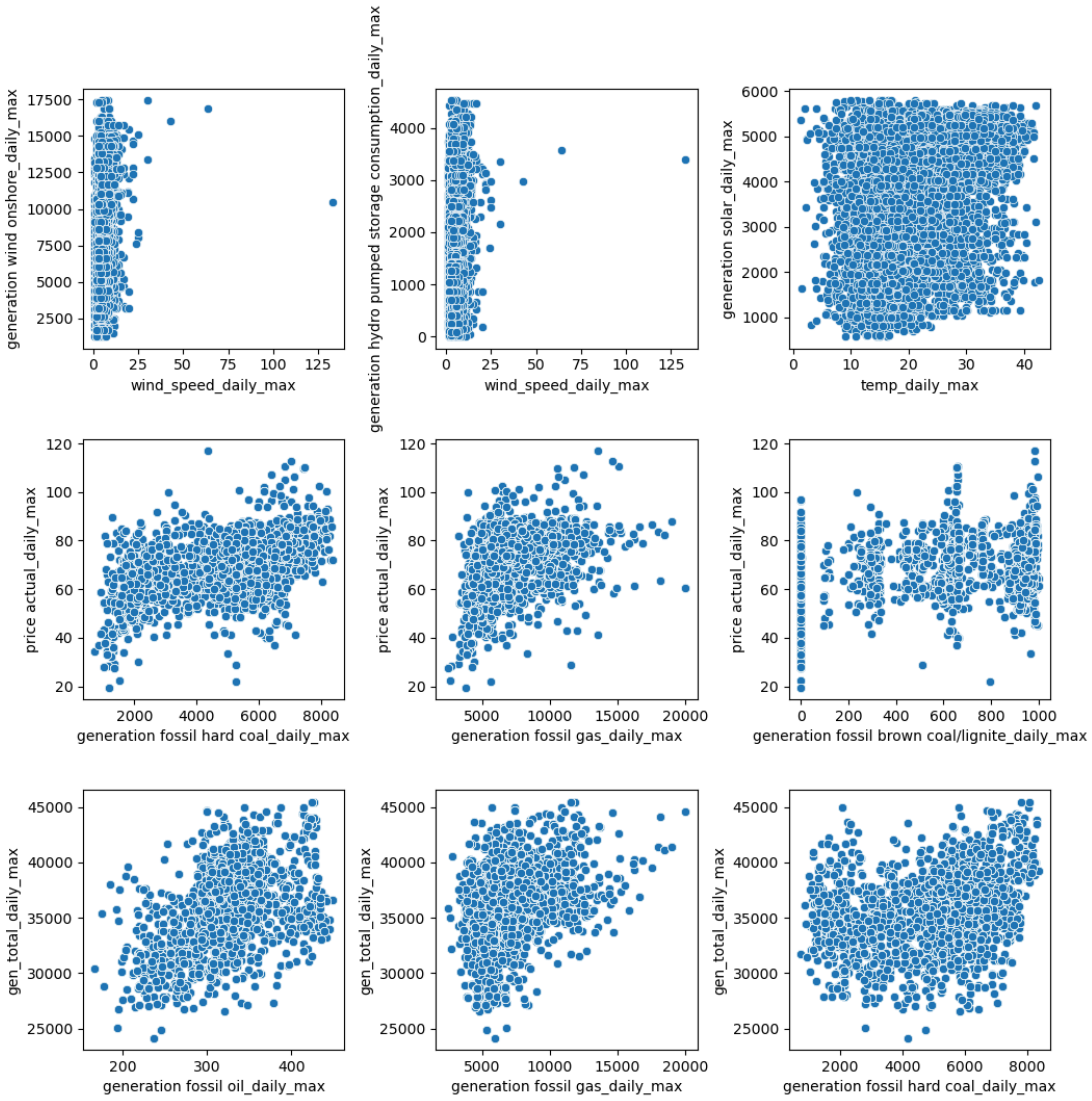



Features Correlating with Actual Price
price actual_daily_max 1
price day ahead_daily_max 0.72
generation fossil hard coal_daily_max 0.46
generation fossil gas_daily_max 0.36
generation fossil brown coal/lignite_daily_max 0.35
total load forecast_daily_max 0.26
total load actual_daily_max 0.25
month 0.23
generation fossil oil_daily_max 0.23
generation biomass_daily_max 0.22
generation other_daily_max 0.2
gen_total_daily_max 0.17
generation other renewable_daily_max 0.1
humidity_daily_max 0.092
day_of_week_Monday 0.077
generation waste_daily_max 0.077
day_of_week_Tuesday 0.062
day_of_week_Thursday 0.056
day_of_week_Wednesday 0.05
snow_3h_daily_max 0.028
day_of_week_Friday 0.01
clouds_all_daily_max 0.0043
city_name_Madrid 4.4e-17
city_name_Barcelona 4e-17
city_name_Bilbao -1.1e-16
city_name_Seville -1.1e-16
city_name_Valencia -1.9e-16
pressure_daily_max -0.0078
rain_3h_daily_max -0.022
generation nuclear_daily_max -0.024
rain_1h_daily_max -0.032
generation hydro pumped storage consumption_daily_max -0.048
generation solar_daily_max -0.11
generation hydro water reservoir_daily_max -0.11
forecast solar day ahead_daily_max -0.12
day_of_week_Sunday -0.13
day_of_week_Saturday -0.13
generation wind onshore_daily_max -0.13
forecast wind onshore day ahead_daily_max -0.14
wind_speed_daily_max -0.15
temp_daily_max -0.15
generation hydro run-of-river and poundage_daily_max -0.2
price actual_daily_max
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00


Features Correlating with Generation Fossil Hard Coal
generation fossil hard coal_daily_max
generation fossil brown coal/lignite_daily_max
price day ahead_daily_max
generation fossil gas_daily_max
price actual_daily_max
generation biomass_daily_max
total load forecast_daily_max
total load actual_daily_max
generation fossil oil_daily_max
gen_total_daily_max
generation other_daily_max
month
generation waste_daily_max
day_of_week_Wednesday
day_of_week_Tuesday
temp_daily_max
day_of_week_Thursday
day_of_week_Monday
day_of_week_Friday
rain_3h_daily_max
city_name_Barcelona
city_name_Bilbao
city_name_Seville
city_name_Madrid
city_name_Valencia
snow_3h_daily_max
humidity_daily_max
rain_1h_daily_max
pressure_daily_max
generation nuclear_daily_max
clouds_all_daily_max
generation solar_daily_max
forecast solar day ahead_daily_max
day_of_week_Saturday
generation other renewable_daily_max
day_of_week_Sunday
wind_speed_daily_max
generation hydro water reservoir_daily_max
generation hydro pumped storage consumption_daily_max
forecast wind onshore day ahead_daily_max
generation wind onshore_daily_max
generation hydro run-of-river and poundage_daily_max
1
0.77
0.64
0.5
0.46
0.46
0.42
0.42
0.41
0.31
0.3
0.18
0.15
0.062
0.049
0.045
0.027
0.025
0.02
0.0047
2.2e-16
2.3e-18
-1.6e-17
-4e-17
-9e-17
-0.0016
-0.0023
-0.0095
-0.017
-0.026
-0.027
-0.036
-0.039
-0.068
-0.11
-0.12
-0.14
-0.28
-0.29
-0.4
-0.4
-0.55
generation fossil hard coal_daily_max
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00



The seaborn pairplot took a very long time to run, and the text ended up being so small it was hard to find it useful. Instead a few of the features that seemed more promising based on the correlation heatmap were chosen for scatter plots. The result is below. These plots reflect the relatively weak correlations represented in the heatmap. In some cases where a strong linear relationship would be expected, such as the relationship between wind speed and wind generation or temperature and solar generation, no linear relationship is found. In some other cases, such as price actual versus hard coal or gas, a weak linear relationship is observed.



3.4 Shifting data

For each observation of the data, there was information about the weather, the generation and the actual load. In reality, for a given day, these features would not be known at the time in which the price is set. In order to compensate for this, the data had to be processed so that the weather and generation data for a given day, would shift to the next day. The process was undertaken as a separate additional step before the data was split.

4.0 Results

Two regression models were created: linear regression and random forest. Pipelines were used for imputing scaling and modeling.

4.1 Linear Regression

For the linear regression, the model was run first as simple linear regression. The second time, it was run using the “SelectBestk” method, and cross validation was run on that model. In the table below, metrics for the first two models are represented as the metric for ytrain and y-train-predict first, and y-test and y-test-pred second.

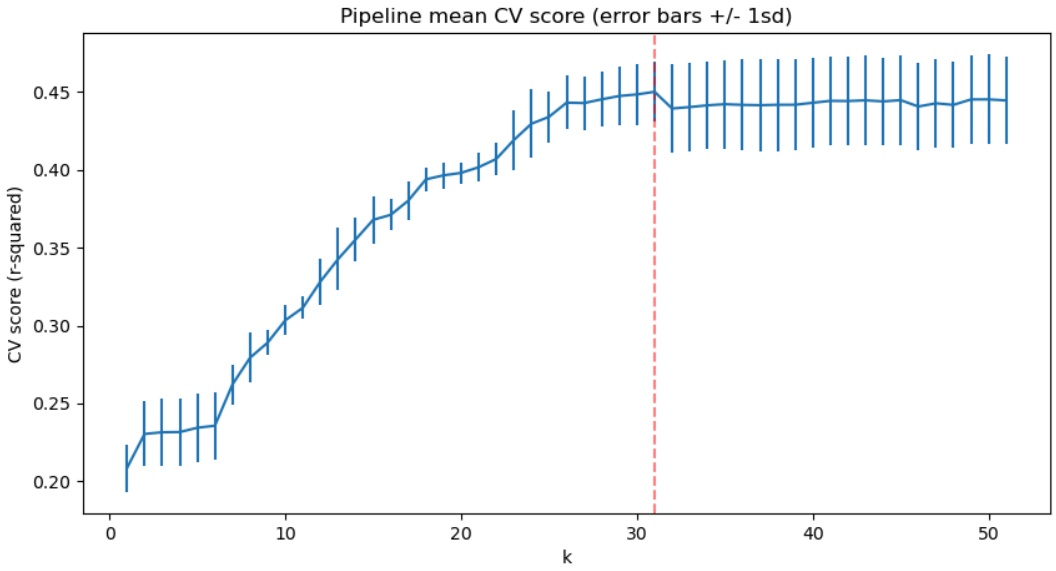
The linear regression model did not perform well. For the base linear regression, the R-squared for the test data was 0.4739, and the root mean squared error was 9.1200. Using SelectBestk, with k=10, did not improve performance, and actually worsened it to an R-squared of 0.3162 for the test data and 10.3978 for the root mean squared error.

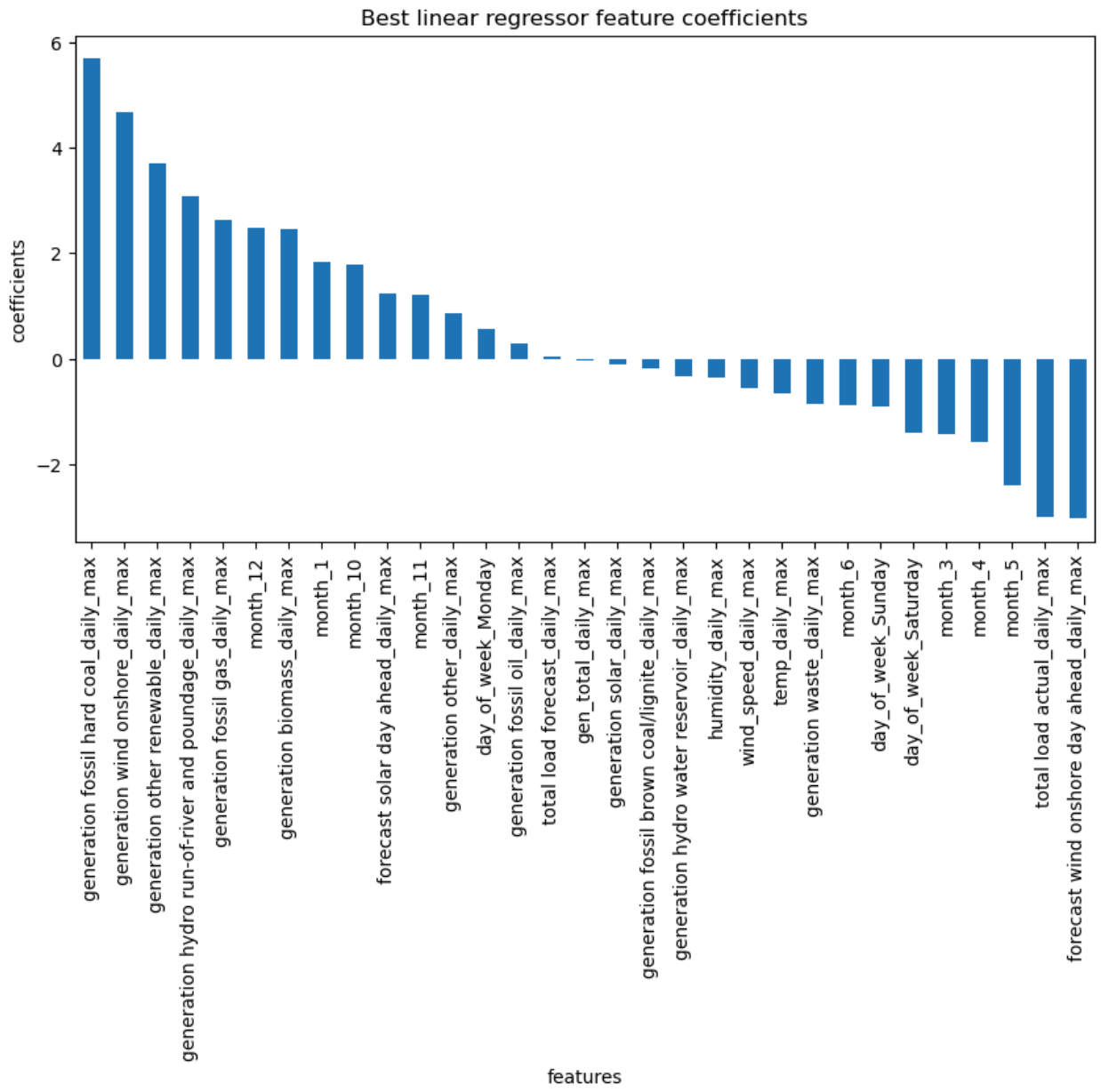
The mean of the scores from a 5-fold cross validation on the set was 0.3033, with a standard deviation of 0.0096, effectively showing the model would perform even worse on unseen data.

	R-squared	Mean absolute error	Mean squared error	Root mean squared error
Linear regression	0.4676, 0.4739	7.3890, 7.1319	89.2992, 83.1747	9.4500, 9.1200
Linear regression w/SelectBestk (w/ default of 10)	0.3149, 0.3162	8.2879, 8.0977	114.9224, 108.1151	10.7202, 10.3978

Finally GridSearchCV was run with k equal to values from 1 to 51, representing the total number of features in the X_train set. The process selected a k of 31 as the best parameter, representing that it found that 31 features was the best number of features to include in order to optimize the performance.

The plots below show the R-squared scores for the various values of k, highlighting the best value of k at 31. The second plot below shows the best linear regression feature coefficients. Since the scores are so low for this model, it’s likely not useful to try to glean much from these plots, however, it is noticeable that the months of the year appear to be more important than they appeared to be in the initial correlation maps (winter months 10, 12, 1 positive, spring moths 3, 4, 5, negative).


Pipeline mean CV score (error bars +/- 1sd)
CV score (r-squared)
0.45
0.40
0.35
0.30
0.25
0.20
0
10
20
30
40
50
k



4.2 Random Forest

Given that these scores are very low, another model was tested: random forest regressor. Using a pipeline and cross validation, the median was used for the imputing method, and the standard scaler was used. Performance of the model jumped dramatically. The mean cross-validation score is 0.9596, with a standard deviation of 0.0147.

The next step was to run the GridSearchCV method with the random forest model. The parameters selected were:

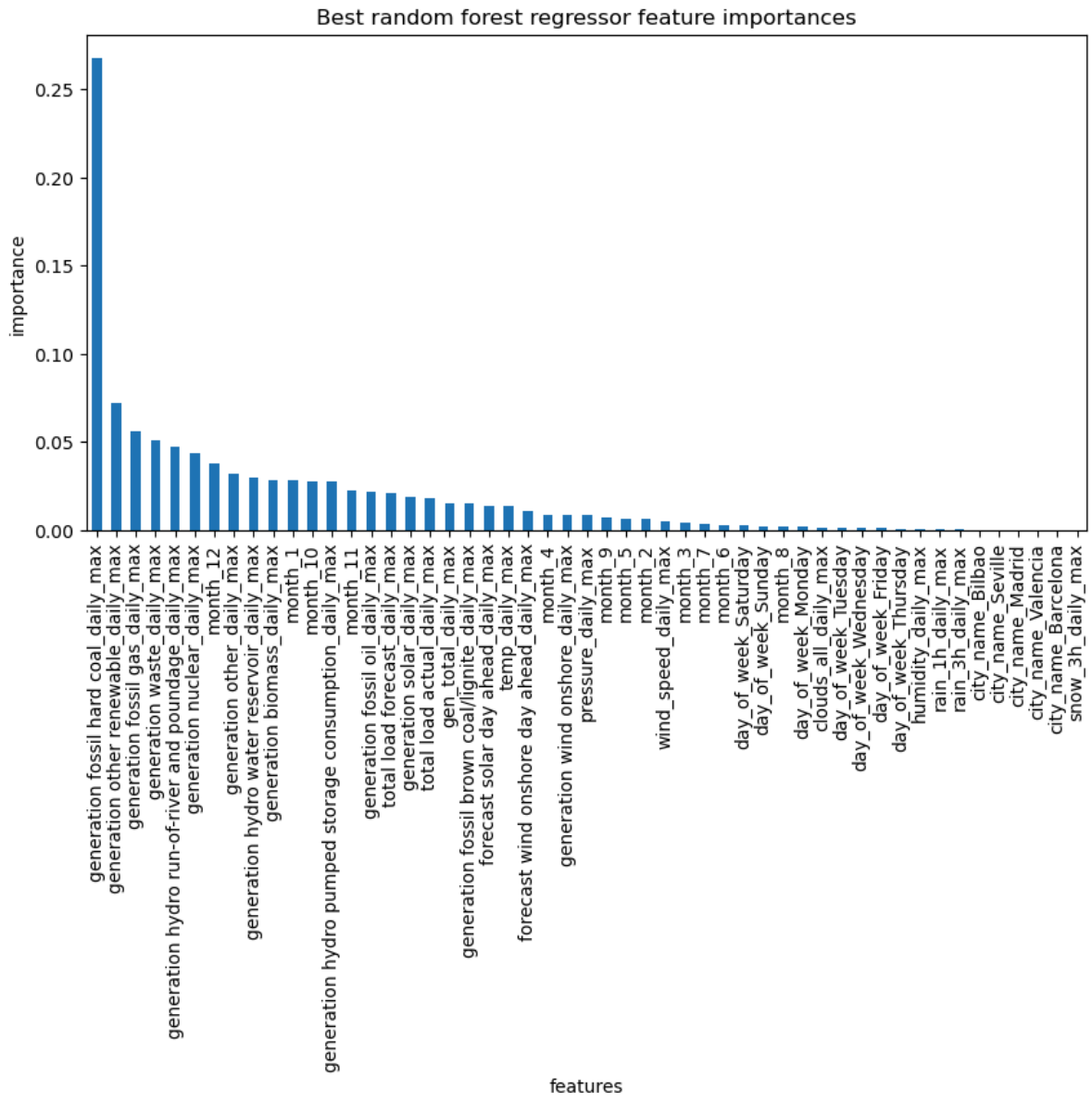
- Number of decision trees (n_estimators): 10, 31, 100, 316, 1000
- Either the Standard Scaler or no scaling
- Either impute with mean or median

This process determined that 1000 was the best of the selected number of decision trees, that mean was the best imputer and that the Standard Scaler should be used. The mean score for the cross validation

this time was 0.9624 with a standard deviation of 0.0137, slightly better than the initial random forest model.

Random forest significantly outperformed linear regression, although the Random Forest models took much longer to run. For example, the hyperparameter search with GridSearchCV took 15.21 seconds to fit the model, whereas the similar process used for the random forest model took 1434.19 seconds (almost 24 minutes).

The plot below shows the best random forest regressor feature importances. The original heat map showing features correlated with price showed hard coal, brown coal/lignite, gas and load as the most highly correlated (excluding price day ahead). Although hard coal still is the most important feature, it also selected other renewables, gas, waste, hydro run-of-river and poundage and nuclear as the most important. Interestingly, wind onshore and hydro water reservoir, which were in the top five of energy sources, do not appear in the top features here. Also interesting is that fall-winter months October, November, December and January appear more important here than they did in the correlation plots. The highest ranking weather feature (temperature) appears as 22nd in importance.



4.3 Final Evaluation

The final step is using the model on the X_test data and evaluating the results.

The mean absolute error of the linear regression model with k=31, is 7.169667026316342 meaning the model will predict price within 7.17 Euros of the actual price.

The mean absolute error of the random forest model using the “best estimators” from the GridSearchCV step, when trained on the X_train data, is 0.7576039680365213, meaning the model will predict price within 0.76 Euros of the actual price.

Of these two models the random forest model is superior even though it takes much longer to run.

5.0 Conclusions

While it seems logical that factors like the weather, the day of the week, the month of the year, and the generation of renewable energy sources like wind and solar would have a big impact on the price of energy, this dataset does not make those relationships clear. Correlation heatmaps and scatter plots of selected features confirmed this. In general there were weak correlations between price and most other features, and the strongest correlations were with non-renewable sources like hard coal.

Modeling reinforced this. For the linear model, which generally performed quite poorly, it selected 31 features as the the best number to include to minimize the error in predicting price, which is roughly 60% of the features. The random forest model, which performed much better, also optimized on 1000 decision trees in order to make the best predictions. In short, at least as far as this dataset is concerned, the problem is complex and defies simple analysis.

In the end, the optimized random forest model predicted prices fairly well, within 0.76 Euro, although it took a significant amount of time longer than the linear model to run.

5.1 Recommendations

In general this model and results can be utilized in several ways:

1. First the model can be used to predict future prices of energy in Euros/MW given weather and energy data.
2. The most important features in determining price for the best performing model include: hard coal, other renewables, gas, waste, hydro run-of-river and poundage and nuclear. Ideally better understanding the generation of these sources, would help understand price fluctuations. Additionally, the fall-winter months October through January also carry importance, and should factor into decisions made about price.
3. Hard coal appears to be a significant factor in price and in its correlation map, was the feature with the highest correlations with other features in the dataset, and it also was the most important factor in the best performing model. It is worthwhile to understand hard coal generation in order to better understand both price and other movements in energy generation.

5.2 Future work

Future work could include using other algorithms and potentially non-linear options to see if they improve performance by reducing the complexity of the models. Additionally, examining the data without summarizing by the maximum value per day would be useful to gauge if it has any impact on the results. Lastly, it would be interesting to do more research into the general lack of correlation between weather and price or generation, potentially looking more closely at the wind and solar weather data and price and wind and solar generation to better understand that relationship.