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

Kera Lagios
Springboard Capstone 02
October 20, 2024

https://www.evwind.es/2019/06/17/the-extension-of-life-of-wind-farms-a-new-challenge-for-wind-energy-in-spain/67616

Problem Statement

How does weather and energy generation/load information affect price of energy in Spain?

Objective

Develop a model that can predict price (Euros)/megawatt (MW) based on energy and weather data.

Scope

Weather information is included for five cities in Spain, however, these cities are not necessarily, or even likely, located where the solar, wind and hydro generation takes place.

Other economic data that could affect price, including information how much is sold or bought from neighboring countries is not included.

Data sources & other important information

Kolasniwash, 2019: “Hourly energy demand generation and weather”. Kaggle, accessed August 17, 2024,

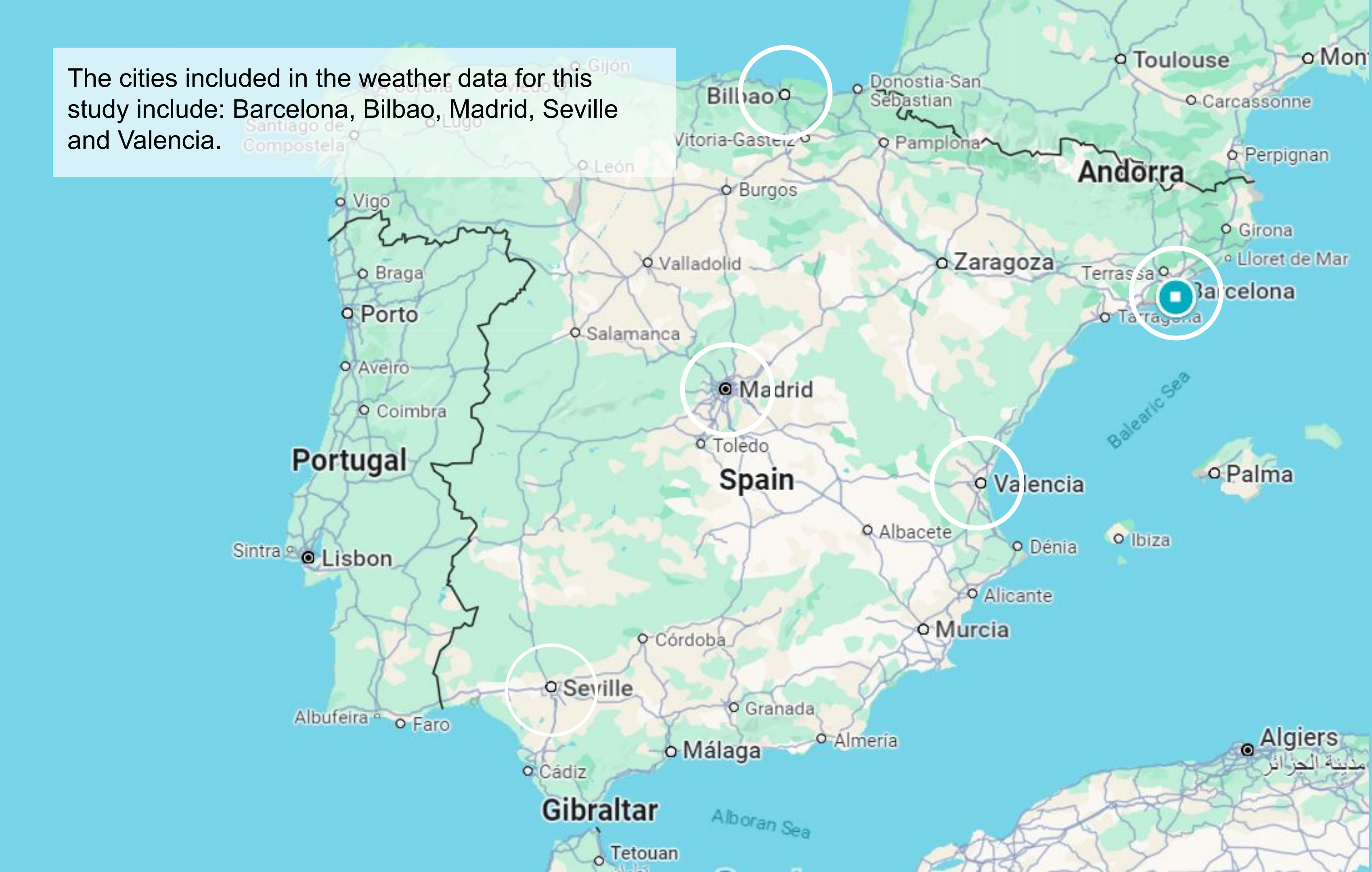
https://www.kaggle.com/datasets/nicholasjhana/energy-consumption-generation-prices-and-weather

https://transparency.entsoe.eu/

https://www.esios.ree.es/en/market-and-prices

https://openweathermap.org/api

The cities included in the weather data for this study include: Barcelona, Bilbao, Madrid, Seville and Valencia.



Data Description

- Two data sets: energy and weather
- Data is hourly
- Each dataset is four years from January 1, 2015 to December 31, 2018.
- Weather data set: numerical and categorical features
  - Example numeric features: temperature, humidity, pressure, wind speed, wind direction
- Energy dataset included numerical features
  - Example features: fossil hard coal, fossil gas, wind onshore, and nuclear
  - Additionally the following features were included: price actual, price day ahead, forecast wind onshore day ahead, forecast wind offshore day ahead, total load forecast, total load actual
- Data was aggregated by day, and the maximum value was taken.

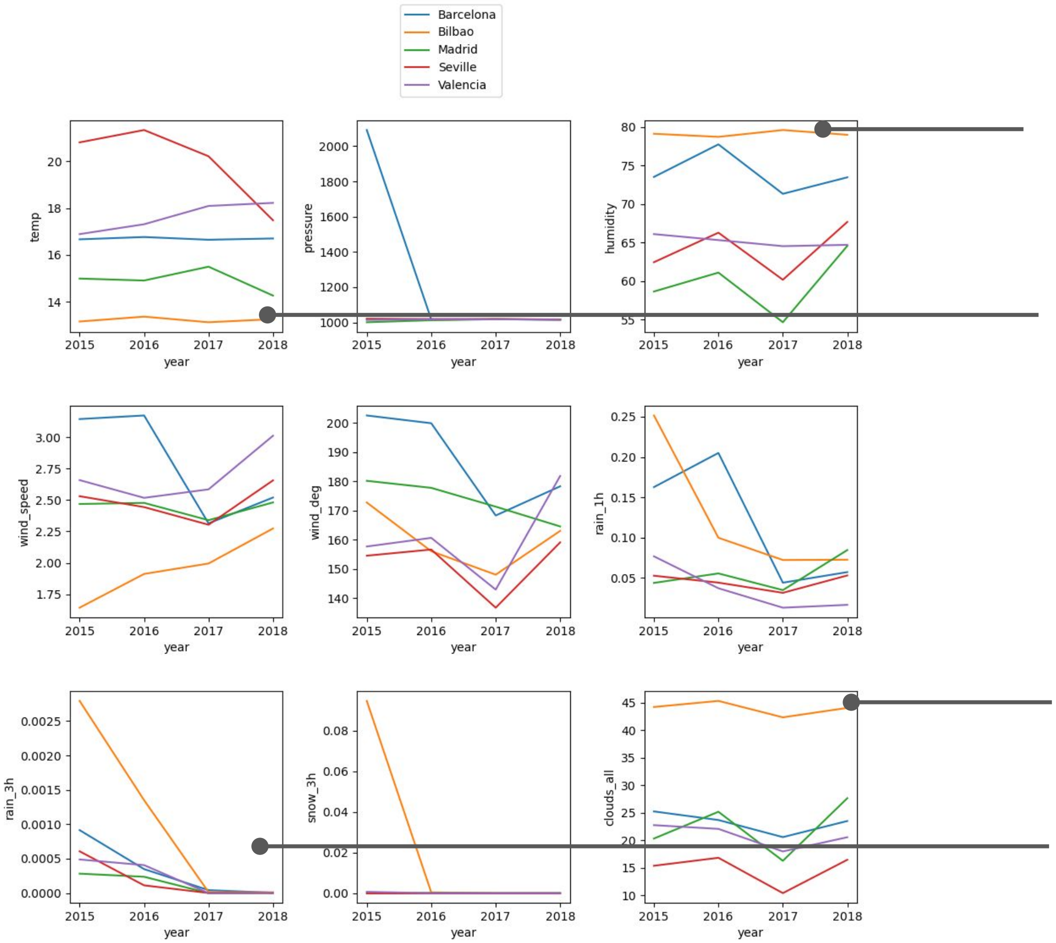
Missing Data & Additional Features

No missing weather data

Features dropped from weather data:
Categorical features
Wind direction

Energy generation features without data:
Fossil coal-derived gas
Fossil oil shale
Fossil peat
Geothermal
Marine
Wind offshore
Hydro pumped storage agg
Forecast wind offshore day ahead

Added features:
month
day of week
total generation
difference between load and generation

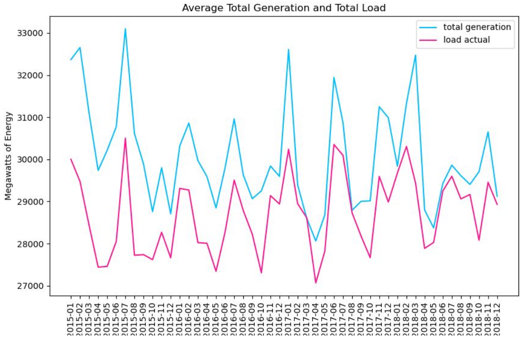


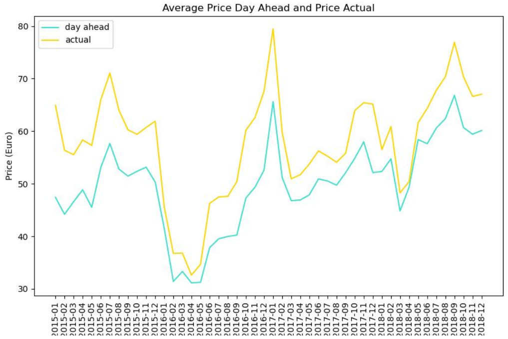
Bilbao and Barcelona are the most humid

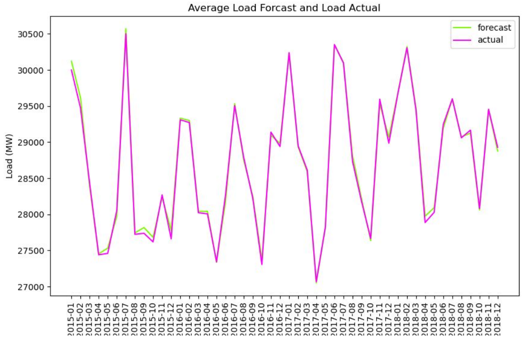
Seville is the warmest, Bilbao the coolest

Bilbao is cloudiest

Dry 2017-2018



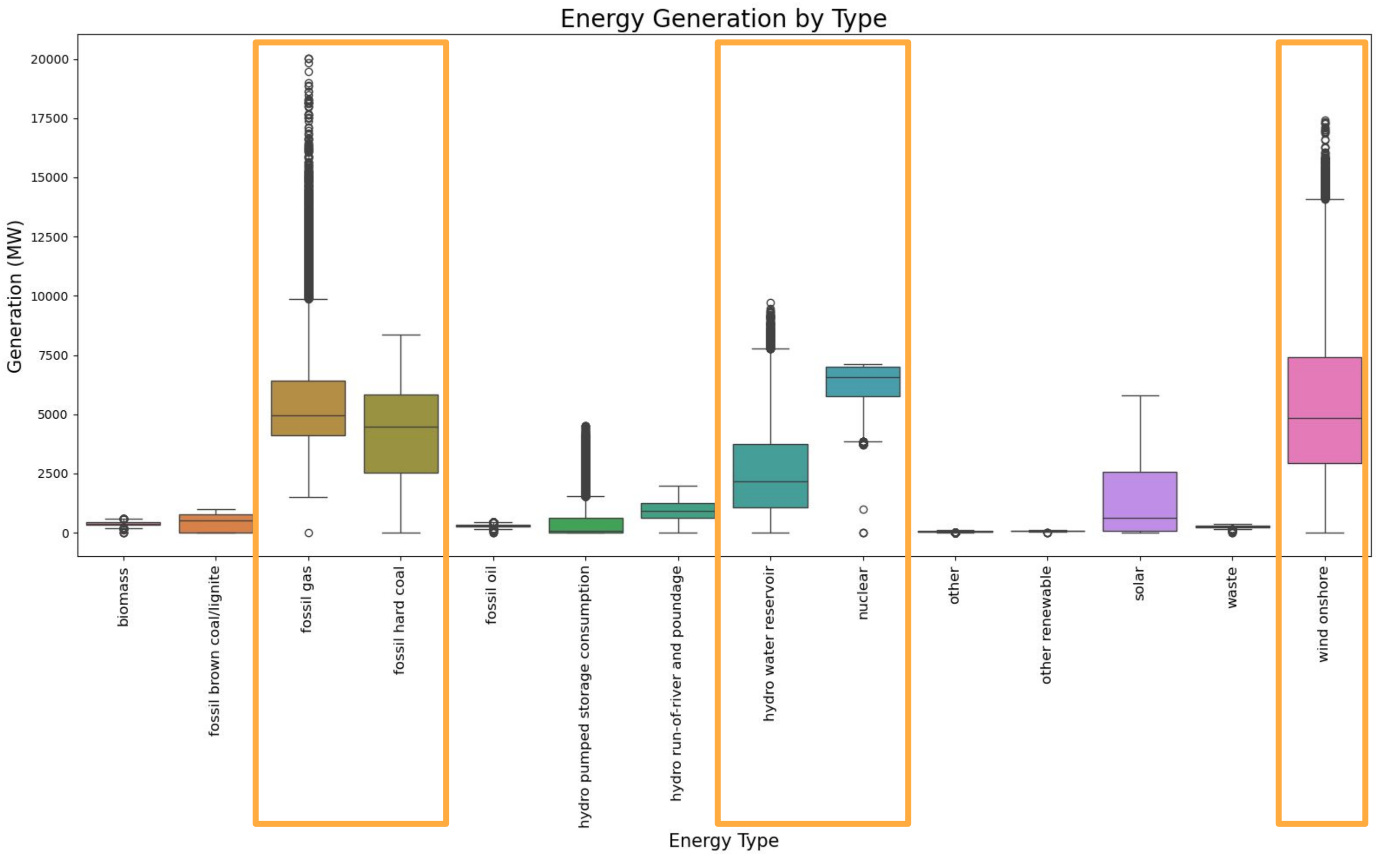




Generation is higher than load, meaning there is generally enough energy and no power outages or blackouts.

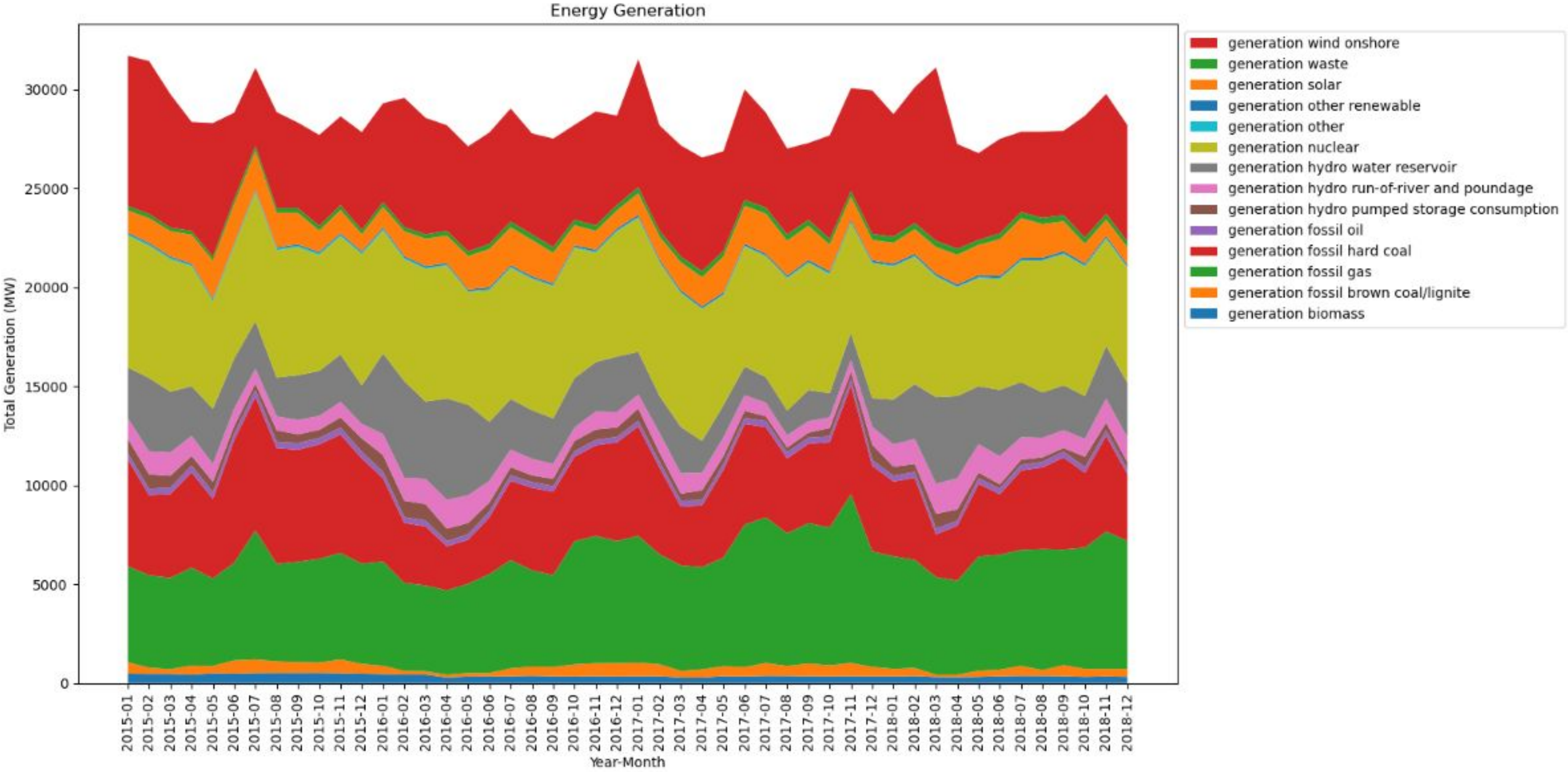
Price actual is higher than price day ahead.

In general, Spain is very good at predicting load, as the two curves are almost identical.

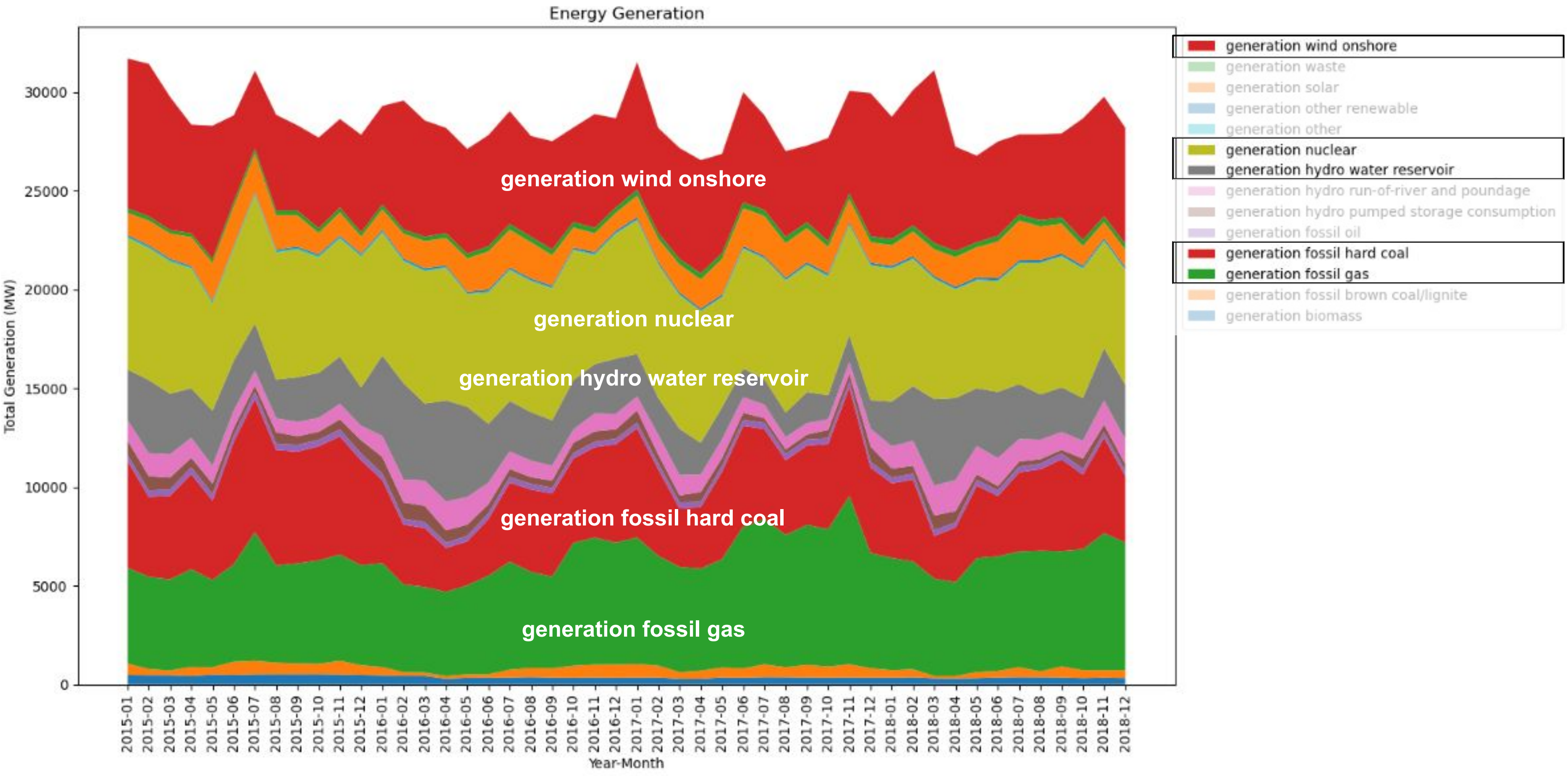


Highest median hourly generation:

1. Nuclear
2. Fossil gas
3. Fossil hard coal
4. Wind onshore
5. Hydro water reservoir



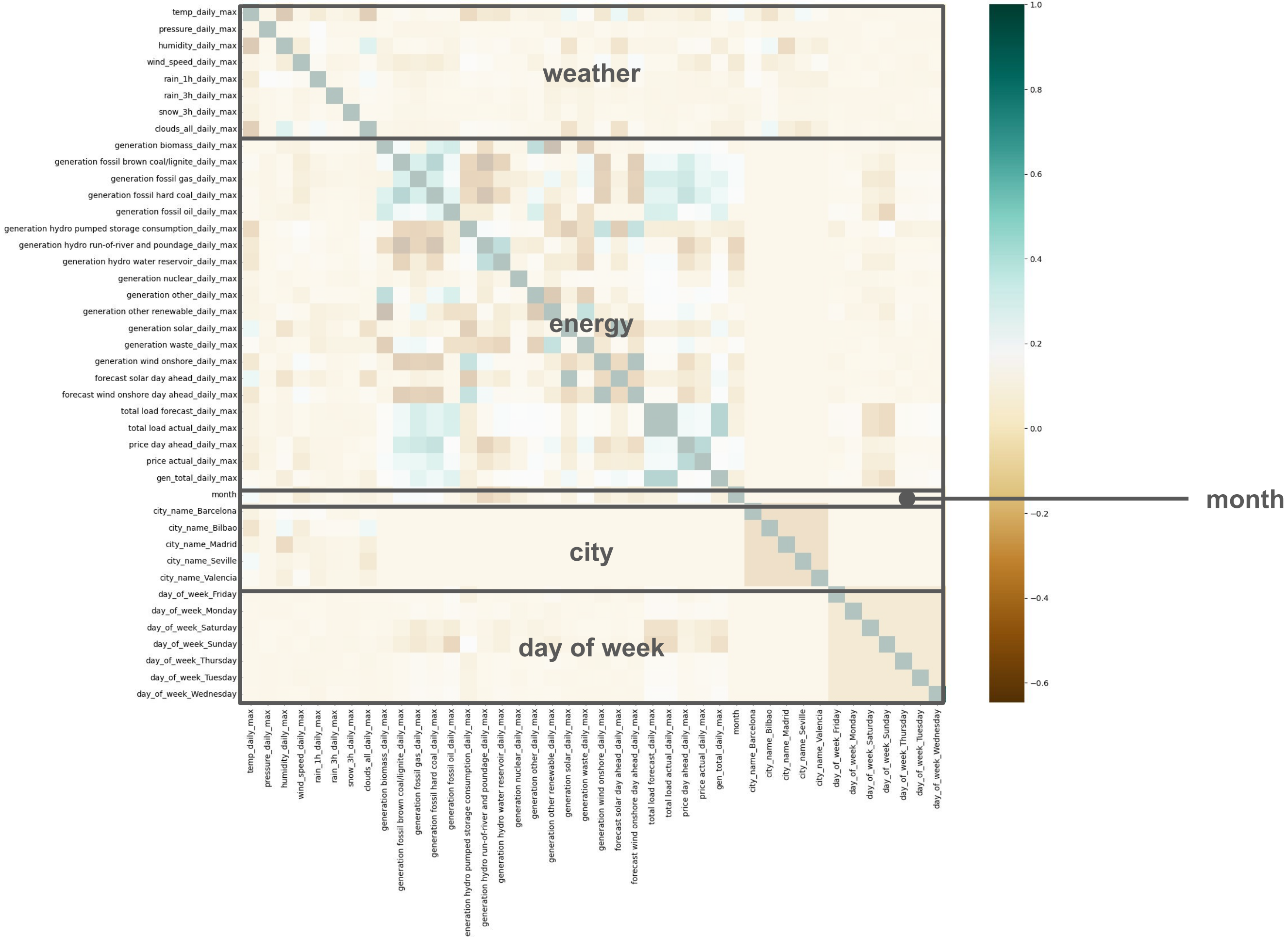
Energy Generation
Total Generation (MW)
Year-Month
0
5000
10000
15000
20000
25000
30000
2015-01
2015-02
2015-03
2015-04
2015-05
2015-06
2015-07
2015-08
2015-09
2015-10
2015-11
2015-12
2016-01
2016-02
2016-03
2016-04
2016-05
2016-06
2016-07
2016-08
2016-09
2016-10
2016-11
2016-12
2017-01
2017-02
2017-03
2017-04
2017-05
2017-06
2017-07
2017-08
2017-09
2017-10
2017-11
2017-12
2018-01
2018-02
2018-03
2018-04
2018-05
2018-06
2018-07
2018-08
2018-09
2018-10
2018-11
2018-12
generation wind onshore
generation waste
generation solar
generation other renewable
generation other
generation nuclear
generation hydro water reservoir
generation hydro run-of-river and poundage
generation hydro pumped storage consumption
generation fossil oil
generation fossil hard coal
generation fossil gas
generation fossil brown coal/lignite
generation biomass


Energy Generation
generation wind onshore
generation nuclear
generation hydro water reservoir
generation fossil hard coal
generation fossil gas
Total Generation (MW)
0
5000
10000
15000
20000
25000
30000
Year-Month
2015-01 … 2018-12
generation wind onshore
generation waste
generation solar
generation other renewable
generation other
generation nuclear
generation hydro water reservoir
generation hydro run-of-river and poundage
generation hydro pumped storage consumption
generation fossil oil
generation fossil hard coal
generation fossil gas
generation fossil brown coal/lignite
generation biomass

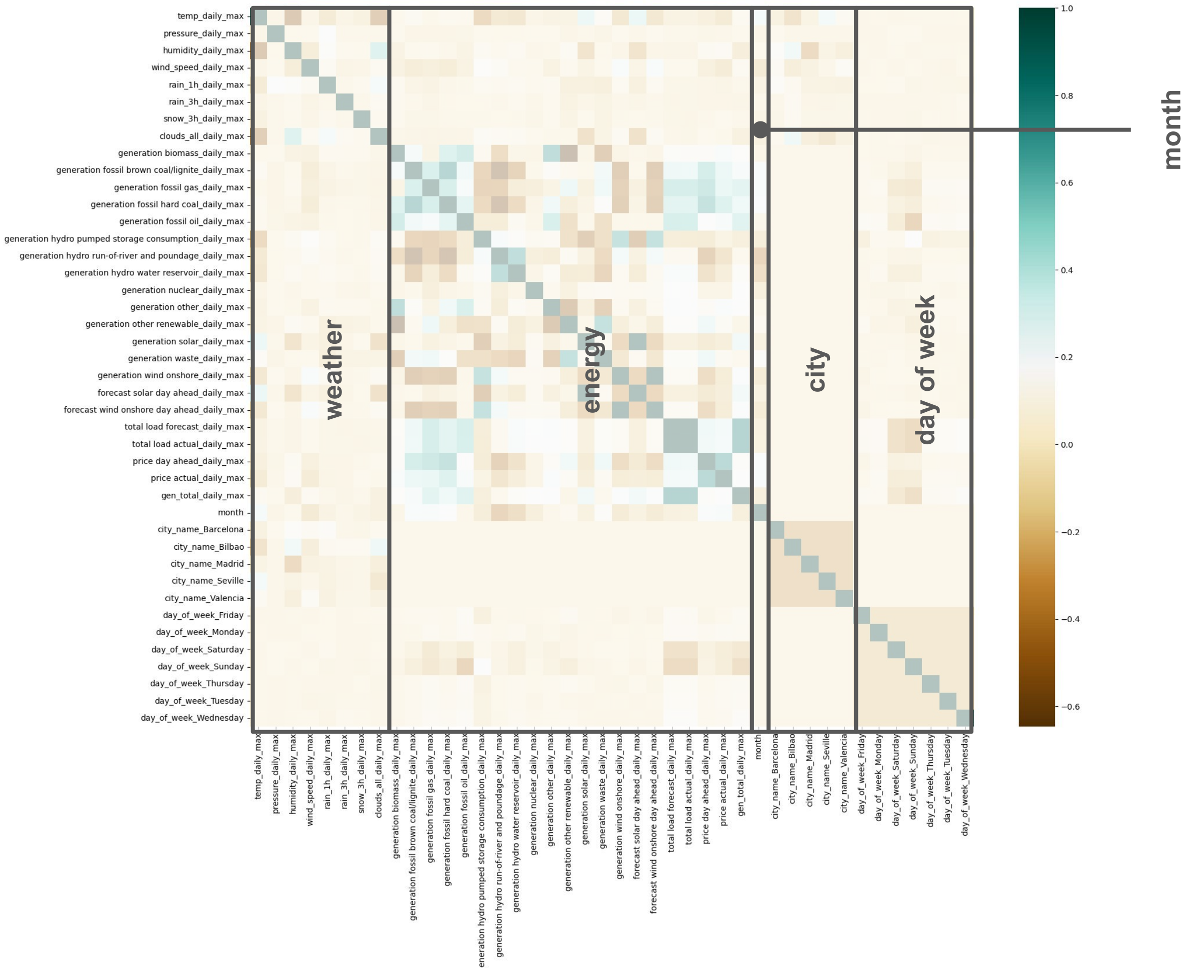
Correlation heatmap

- Dark green: positive correlation (as one feature increases the other increases)
- Dark brown: negative correlation (as one feature increases the other decreases)
- Map is “mirrored” along the diagonal
- The large amounts of beige reflects a zero correlation for many of the features.
- The few dark areas mean there are not a lot of strong correlations

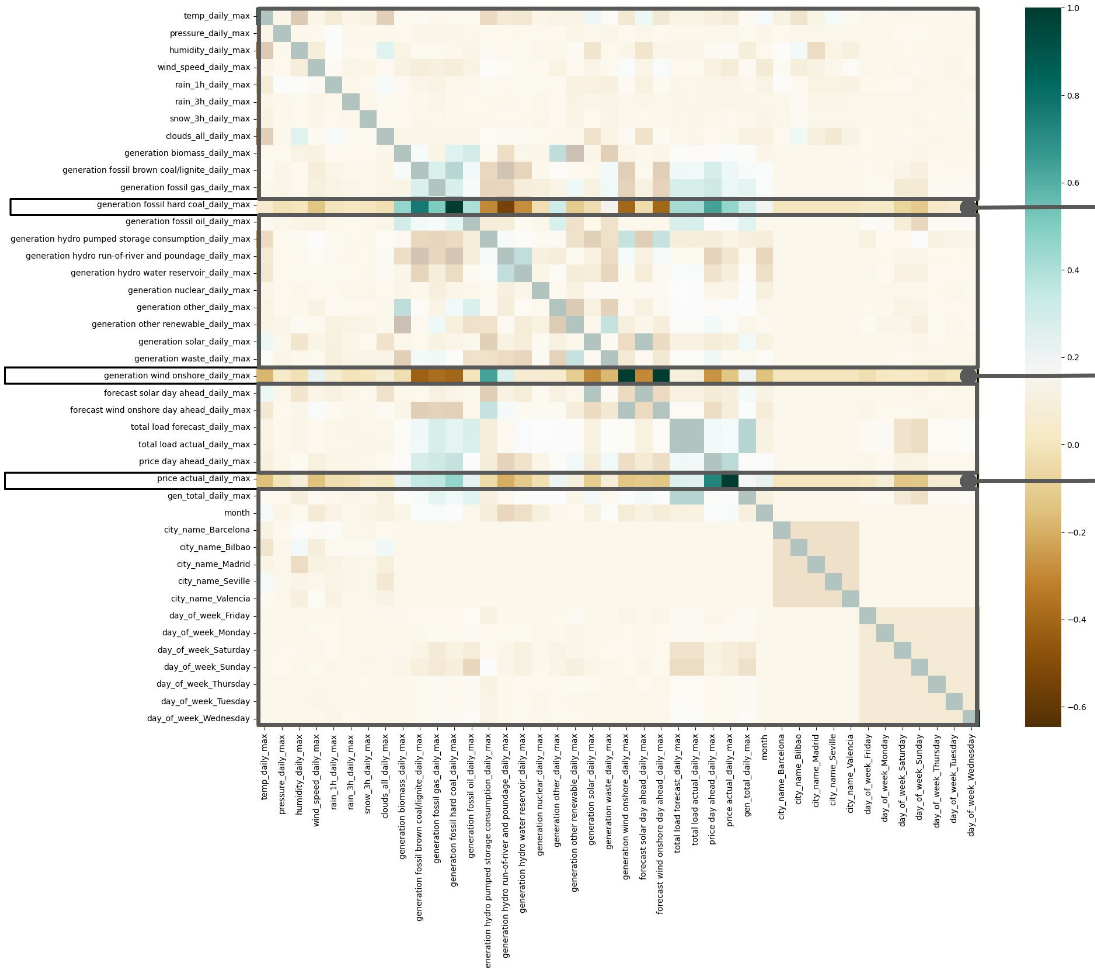
Correlation heatmap



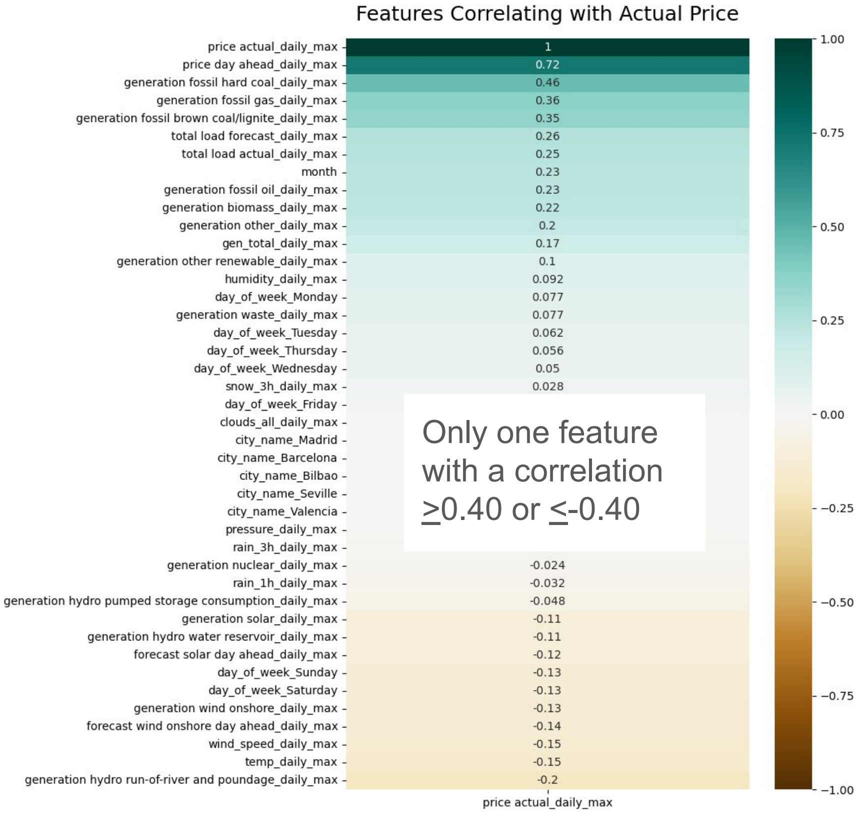
Correlation heatmap

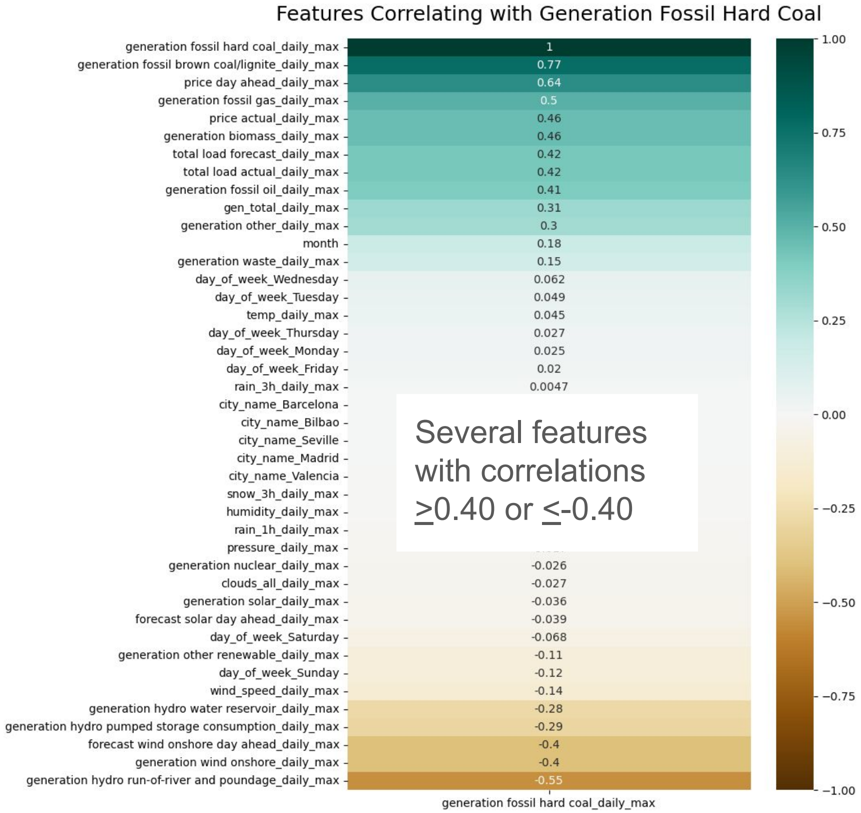


Correlation heatmap

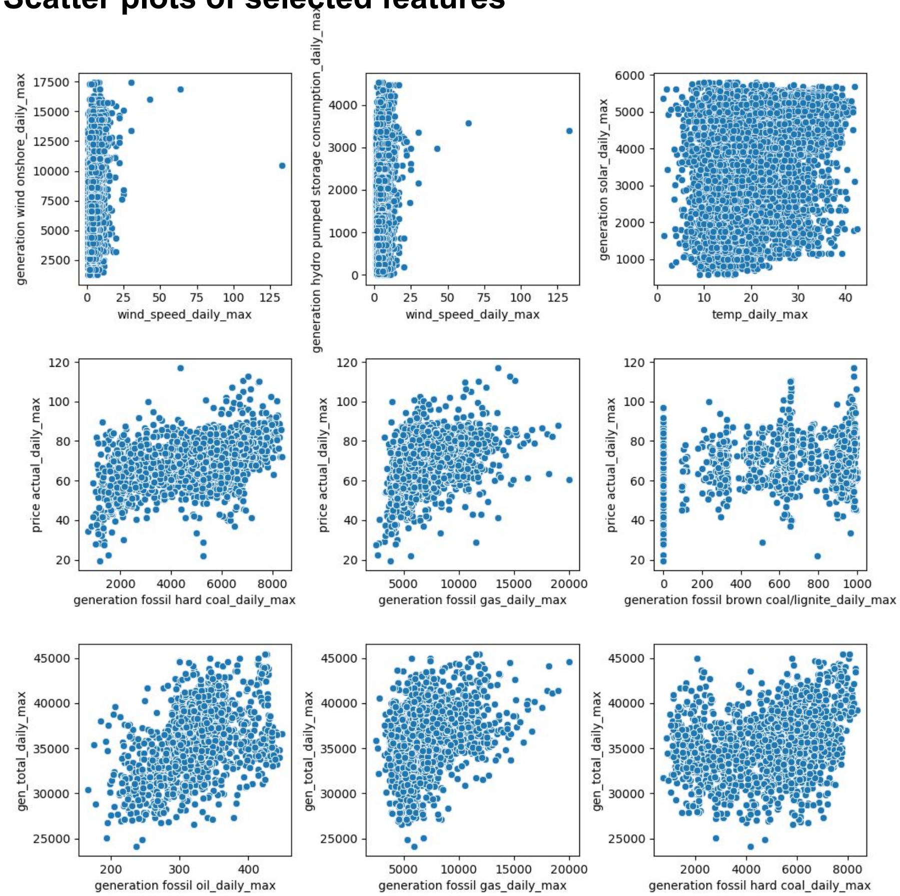


Ranked correlations by feature





Scatter plots of selected features



- Strong linear relationships in scatter plots would appear as the dots clustered along a diagonal line
- Only weak linear relationships are visible in some of these plots

Key Modeling Steps

Joining the energy and weather data sets.
Shifting features one day ahead to prevent the model from “knowing” information that would not be available until the next day.
Linear regression modeling including cross-validation and hyperparameter tuning.
Random forest regression modeling including cross-validation and hyperparameter tuning.

Linear Regression

31
Best value of k

7.17€
Mean absolute error

0.4739
r-squared

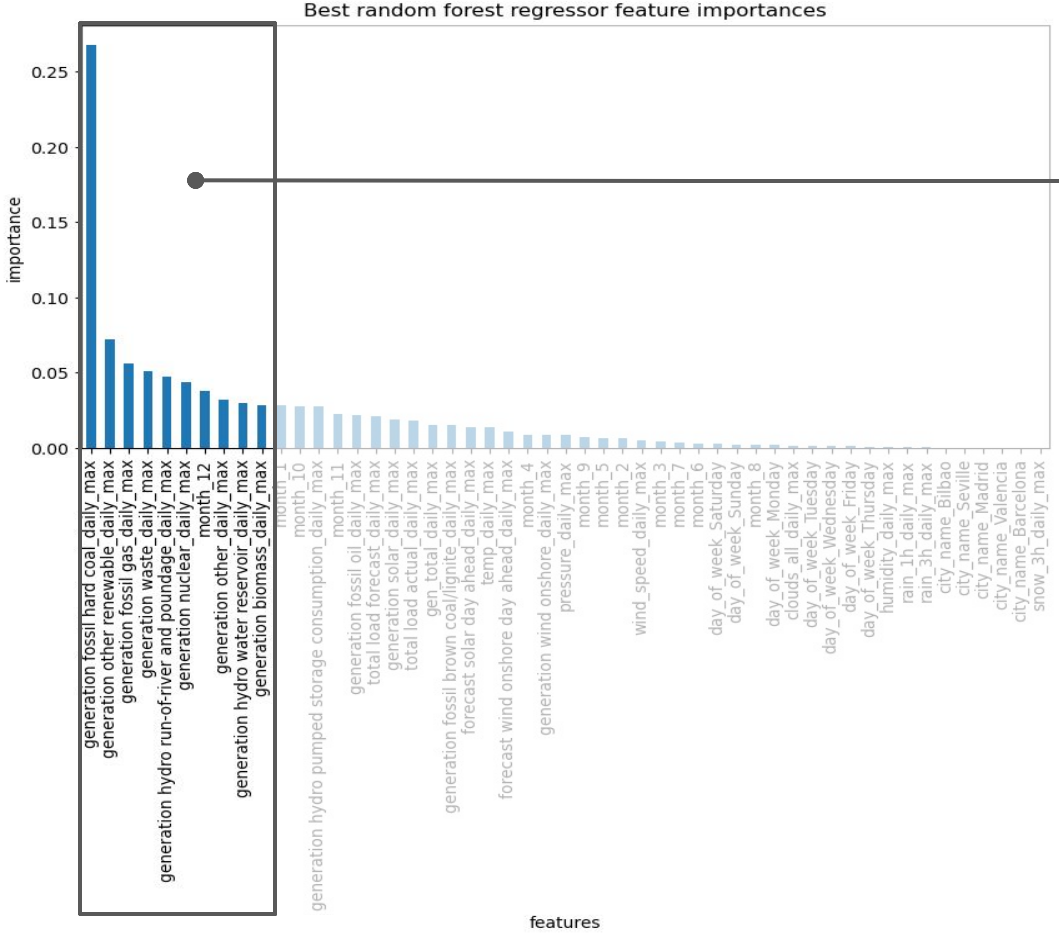
Random Forest Regression

0.76€
Mean absolute error

1000
Best n_estimators (number of trees)

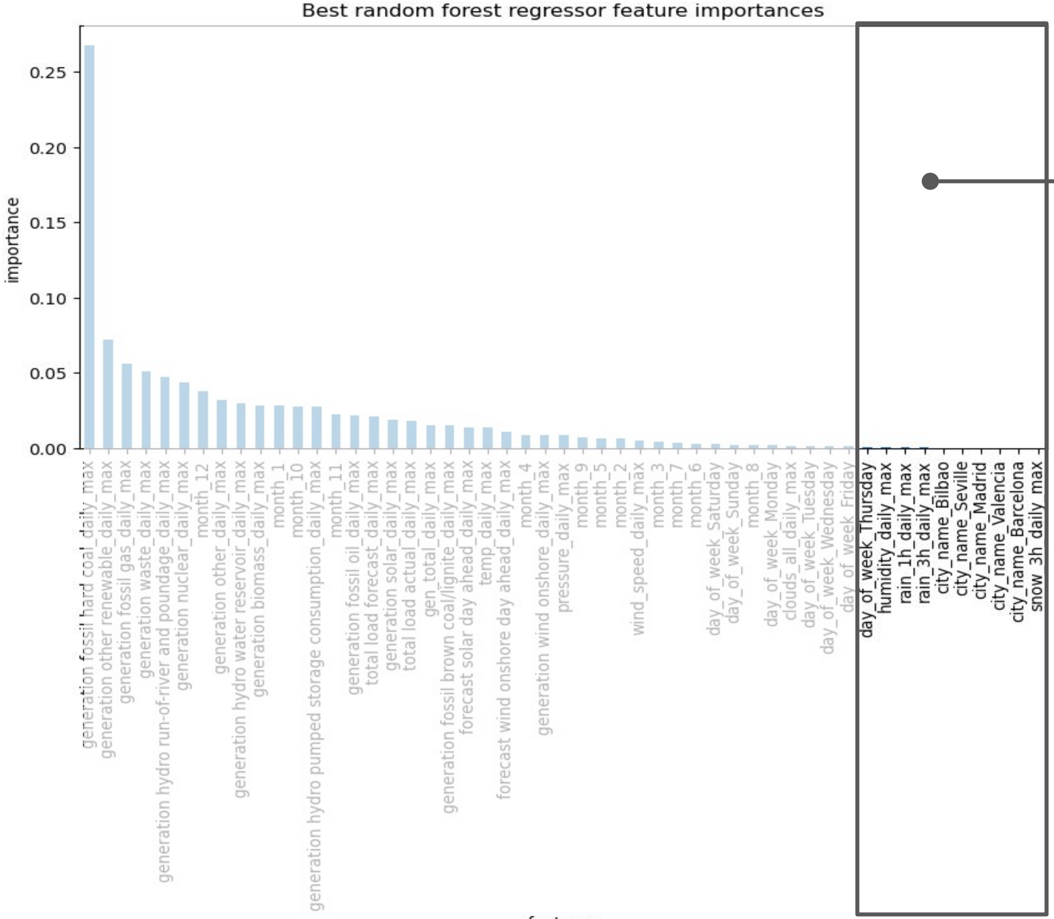
0.9624
Mean cross-validation score

Random Forest Regression Modeling



Most important features

Random Forest Regression Modeling



Least important features

Recommendations and Key Findings

In general this model and results provide several uses and insights:

1. First the model can be used to predict future prices of energy in Euros/MW given weather and energy data in Spain, although predicting beyond 2018 would ideally include more up-to-date data..
2. The most important features include:
    a. Hard coal
    b. Other renewables
    c. Gas
    d. Waste
    e. Hydro run-of-river and poundage
    f. Nuclear
    g. Fall and winter months (October through January)

    Ideally better understanding the generation of these sources and the influence of fall-winter months, would help understand price fluctuations.
3. **Hard coal**:
    a. Significant factor in price and in its correlation map
    b. The feature with the highest correlations with other features in the dataset
    c. 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.

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.