IE School of Human Sciences & Technology

MACHINE LEARNING II

Daniel Garcia Hernandez

SPANISH POWER PRICE PREDICTION

power_market.csv

Gianni Miller
Yannick van Dam
Maria Gonzalbez
Zohak Mirza
Göktuğ Aşcı
Gerardo Gandara
Joaquin Calderon

Section 1 - Group 5
06.03.2021

Section 1 - Group 5

08.03.2021

EXECUTIVE SUMMARY

In recent years electricity prices soar and have been changing continuously. This report aims at helping both companies and customers with future electricity price predictions using Machine Learning algorithms with past data. Information from the power_market.csv will be used for training the data and scoring.csv will be then imputed for testing the data as well as the model's accuracy. Assumptions taken for working on these goals were adjusting demand for both import and export of electricity, due to many interactions with for instance France or Spain. Limitations encountered were a relatively small, still sufficient, data set, a rather simple model (LogisticRegression) being the best predictor, the nature of the power resources (ie. coal, gas and wind as *Thermal Gap*) as well as them not being listed individually. However, it is important to note that the source data is in the quartiles normally distributed, hence the best predicting model (LogisticRegression) has conceptual, despite empirical, significance for predicting electricity prices.

Conducting exploratory data analysis it was found that the mean price of the data set is approximately 48, with a relatively large standard deviation of approximately 15€. This emphasizes the need for more accurate price predictions as well as creating better synergies between demand (consumers) and supply (the nuclear, wind and solar plants). Notable are the low presence of null values. In the correlation analysis it was identified that the variables of interest amongst others demand, supply, price or hour have significant correlated relationships, which benefits the purpose of this analysis. When engineering the features some interesting correlations were found and indicated which variables to select for the data engineering. The high correlation of the price with the thermal gap is due to the description of the thermal gap: the expensive powers. Price is negatively correlated with the net exchanged power because there are more imports than exports. Further the thermal gap is negatively correlated with the net exchange power because the exchange is with "cheap energy" and it is negatively correlated with the wind power forecast because the more wind power there is, the less "expensive powers" are used. This led to focusing on the variables *thermal_gap*, *net_exc_power, demand, Solar, nuclear and wind productions* and *price: target* for the model training as well as discarding *import* and *export*, as net_exc_power calculated by the difference of both.

Many models were trained and tested upon, as indicated above, a linear regression model was the winner. When cross validating this model showed the best results against amongst others XGBoost, Decision Tree or Random Forest. The LogisticRegression algorithm had the highest test scores and lowest mean squared error with approximately 12, which reduces the standard deviation by 20%. This model's predicted mean is approximately 43.50€ compared to approximately 48€ of the model as is.

In figure 6 *Time series graph showing predictions vs actual price of power in Spain*, it can be seen how accurate the predictions of this model are when compared to the prices of the scoring data set. It is visually visible that this model's predictive power for the two data sets at hand is significant. To conclude, the simplicity of this model is the beauty. With the help of a simple LinearRegression algorithm, this model can definitely predict the electricity prices of the following day for expectation management in the future.

Table of Contents

1. INTRODUCTION

The prices of power in Spain have fluctuated significantly over the last years and many have been trying to figure out why and how to reduce or at least predict their monthly payments. In 2018 the price was even at its highest in ten years at €74.58 per megawatt per hour[1]. The price for power is heavily influenced by the demand and supply. Therefore the problem of high prices has arisen from inaccurate prediction models which were unable to precise estimate the demand in the country therefore creating a deficit in the supply of power. With the time lags in production with multiple of the key variables such as nuclear, wind and solar, building an accurate prediction model will help the producers anticipate the demand of the power relieving the consumers of high prices. In this report there will be an explanation of how the final prediction model was created, which methods were used to get to our final prediction accuracy and our final results and conclusions. "The goal of this assignment is to build a Machine Learning model using what was learned in class."[2]

1.1. Assumptions & Limitations

Considering the above mentioned, one has to make assumptions and identify limitations when conducting a prediction analysis. The dataset has records for 3 years and 8 months from 2017 to 2020 although this is a lot, more data will always help a prediction algorithm get better. Additionally, the price of power is heavily influenced by the demand and the thermal gap. In this dataset there is only information about the cheaper power producers such as solar, wind and nuclear. This leads to making findings based on those variables and represents a limitation of the predictability of all electricity prices. Therefore to calculate the correlation between the expensive energy and the price, the demand minus the cheap energy sources had to be taken. This is a big assumption and limitation as it would be better to have the explicit column of each power resource such as coal and oil. Additionally there are significant imports and exports to France. This is because when France needs power or Spain needs power the countries are close geographically and therefore it is easiest to import or export, affecting the price due to additional cost items occurring. Hence a variable called *net exchange* had to be constructed, to account for the import and export factors. In the notebook there is a variable that gets created which is the net exchange which takes the imports minus the exports. The heatmap shows that there is a negative correlation for this variable but this is only because there are more imports than exports to France.

Finally, it was observable that there are more imports than exports as well as the price being highly influenced by the power. For the second, initially three date intervals have been set up, but decreased the predictability of this model. Hence, this further limits the detailed prediction this report aimed for. Also, one could consider the model that "won" (LinearRegression) a limitation for the rather simple approach compared to other models like XGBoost and so on. However, in class it was said that in ML sometimes the beauty is the simplicity of algorithms for solving the problem at hand.

[1] murciatoday.com/-price-of-electricity-in-spain-soars-27-per-cent-in-the-midst-of-the-cold-snap_1547878-a.html.

[2] group_assignment_description.pdf

2. DATA INFORMATION

When conducting an analysis, it is important to have domain knowledge and understand the data at hand. The following data definitions have to be taken into account, provided by the PM:

- "The price of the power market is heavily affected by demand and also by those power producers with higher costs of production like coal, gas, etc. Those forecasts are not available, but you can infer how much "expensive production" is needed by subtracting the demand for the energy provided by nuclear, wind and solar. This is called Thermal Gap
- Analyze each time series by itself: look for outliers, NaNs, and see their distributions. Look for skewed distributions and think if you can convert the values into categories
- When dealing with feature engineering, a good start point can be checking the differ- ences, ratios, absolute differences between demand and each power production time- series.
- Create new features out of what you already know like the seasonality of solar pro- duction and the different periods of the day. Think that demand is heavily affected by human behavior and human behavior heavily depends on the sun!
- After you're done with Feature Engineering, look for correlations between features and the target, maybe your new variables are not correlated at all and therefore might only include noise in your data."[3]

2.1 Data Sources

Data sources provided for this analysis are two csv files with the following content:

- power_market.csv: dataset for training the model. Some of it was saved for testing the performance of the model, so that the model can be corrected for under- /overfitting.
- scoring.csv: once the model is final, this dataset was used to predict and the predictions were saved in a separate CSV file named prediction.csv.

2.2. Data Description

Further, the data at hand also had to be described, as follows:

- "Supply: power generation plants (nuclear, solar, wind, etc). Each type of production is different and should be analyzed independently before making assumptions. Analyzing seasonality, distributions and correlations is advised.
- Demand: all the power consumption in the country, from huge factories to small house- hold consumers. Very seasonal and heavily affected by temperature."[2]

2.2. Data Description

The following represent the main columns in the dataset and what they describe:

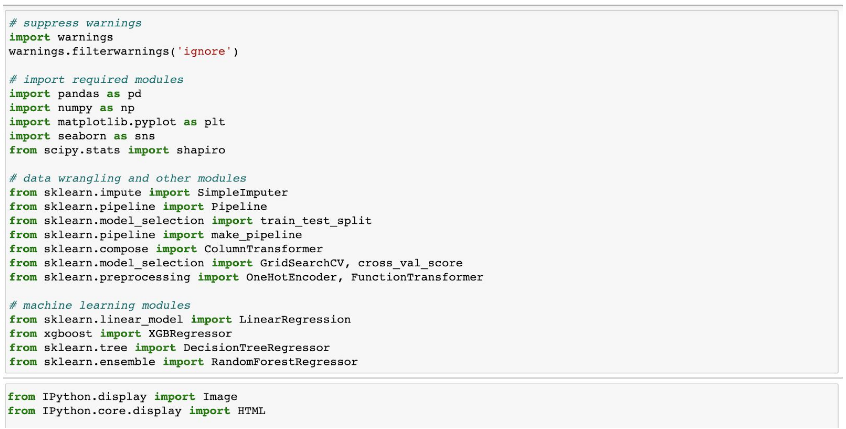
- **"date**: date of the observation "%Y-%m-%d"
- **hour**: hour of the observation, [0 - 23]
- **fc demand**: forecast of demand in MWh
- **fc nuclear**: forecast of nuclear power production in MWh

[3] group_assignment_description.pdf

- **import FR**: forecast of the importing capacity from France to Spain in MWh
- **export FR:** forecast of the exporting capacity from Spain to France in MWh
- **fc wind:** forecast of wind power production in MWh
- **fc solar pv**: forecast of PV solar (solar panels) power production in MWh
- **fc solar th:** forecast of thermal solar power production in MWh
- **price:** power price for each hour in €/MWh. This is the target we want you to predict."[2]

3. SETTING UP THE ENVIRONMENT

To be able to do the analysis one had to load and import multiple libraries into the notebook. Figure 1 shows all the libraries that were used.



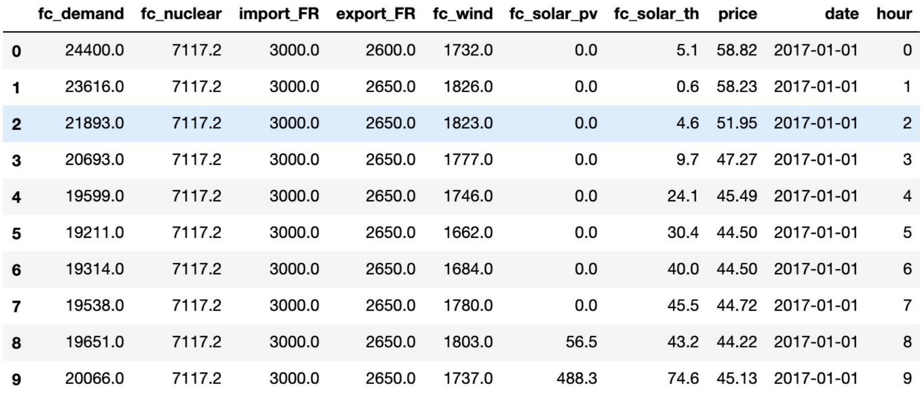
```
# suppress warnings
import warnings
warnings.filterwarnings('ignore')

# import required modules
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import shapiro

# data wrangling and other modules
from sklearn.impute import SimpleImputer
from sklearn.pipeline import Pipeline
from sklearn.model_selection import train_test_split
from sklearn.pipeline import make_pipeline
from sklearn.compose import ColumnTransformer
from sklearn.model_selection import GridSearchCV, cross_val_score
from sklearn.preprocessing import OneHotEncoder, FunctionTransformer

# machine learning modules
from sklearn.linear_model import LinearRegression
from xgboost import XGBRegressor
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor
```

```
from IPython.display import Image
from IPython.core.display import HTML
```

Figure 1: Environment Setup

4. EXPLORATORY DATA ANALYSIS

In the Data Analysis section the data frame raw data will be examined and a better understanding of it will be fostered. After downloading the relevant library, the data frame's schema was printed.



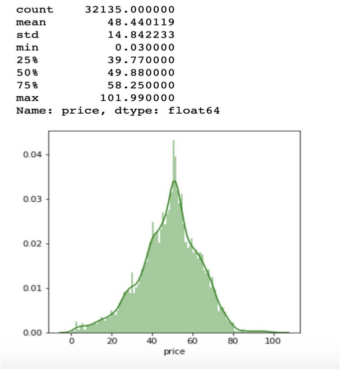
	fc_demand	fc_nuclear	import_FR	export_FR	fc_wind	fc_solar_pv	fc_solar_th	price	date	hour
0	24400.0	7117.2	3000.0	2600.0	1732.0	0.0	5.1	58.82	2017-01-01	0
1	23616.0	7117.2	3000.0	2650.0	1826.0	0.0	0.6	58.23	2017-01-01	1
2	21893.0	7117.2	3000.0	2650.0	1823.0	0.0	4.6	51.95	2017-01-01	2
3	20693.0	7117.2	3000.0	2650.0	1777.0	0.0	9.7	47.27	2017-01-01	3
4	19599.0	7117.2	3000.0	2650.0	1746.0	0.0	24.1	45.49	2017-01-01	4
5	19211.0	7117.2	3000.0	2650.0	1662.0	0.0	30.4	44.50	2017-01-01	5
6	19314.0	7117.2	3000.0	2650.0	1684.0	0.0	40.0	44.50	2017-01-01	6
7	19538.0	7117.2	3000.0	2650.0	1780.0	0.0	45.5	44.72	2017-01-01	7
8	19651.0	7117.2	3000.0	2650.0	1803.0	56.5	43.2	44.22	2017-01-01	8
9	20066.0	7117.2	3000.0	2650.0	1737.0	488.3	74.6	45.13	2017-01-01	9

Figure 2: Raw Data Frame

The data frame in figure 2 gives an overview of all the columns that are in the dataset and their values. From this the mean power price can be calculated which for this dataset is around €48.44. Although lower than the high in 2018 of €74.58 it can be found that this number is still very high as an average. This is relevant for the analysis as in the prediction the goal will be to see if one can predict the price in order to create better synergies between the demand (consumers) and the supply (the nuclear, wind and solar plants). Additionally it was found that the dataset has 32,135 records.

4.1 Null Values:

In this dataset there were only 13 Null values for both the import and export column. For this analysis these Null values were imputed with the median using SimpleImputer, as in this case import and exports stay relatively constant through the time period.



```
count    32135.000000
mean        48.440119
std         14.842233
min          0.030000
25%         39.770000
50%         49.880000
75%         58.250000
max        101.990000
Name: price, dtype: float64
```

Figure 3: Price Distribution of raw data

To gain a better understanding of the data, figure 3 illustrates the different summary statistics and graphs the distribution of price. It can be seen again that the mean is at €48.44 but more interestingly the standard deviation has a value of €14.84. This is significant as this would show that the price of power fluctuates often and there is a large discrepancy between the demand and supply at times. This can be further confirmed by looking at the minimum price in the dataset of €0.03 and the maximum price of €101.99. With an accurate model that can predict demand and supply it will help both parties to achieve a more stable price. After looking at Figure 3 the distribution for price would seem like it is normally distributed but after conducting a shapiro test it can be concluded that the data does not follow normal distribution. Also, it can be seen that the model is slightly skewed to the right and almost normally distributed, hence the LogisticRegression algorithm's confidence intervals can definitely increase the predictability of this model.

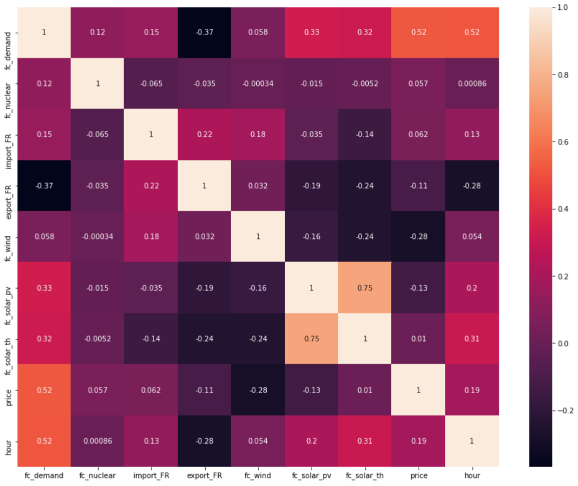


Figure 4: Correlation Analysis of Raw Data

To draw conclusions and make predictions about the price of power in Spain, one has to analyze how the different columns are correlated with each other. From this heat map we can see:

- forecast of demand in MWh has a correlation of **0.52** with **price and hour**
- forecast of demand in MWh has a correlation of **0.12** with **fc_nuclear**
- forecast of demand in MWh has a correlation of **0.32** with **fc_solar_pv** and **fc_solar_th**
- forecast of demand in MWh has a correlation of **0.12** with **import_FR**
- forecast of demand in MWh has a negative correlation of **-0.37** with **export_FR**

- price of the power production in €/MWh has a negative correlation of **-0.28** with **fc_wind**
- price of the power production in €/MWh has a negative correlation of **-0.13** with **fc_solar_pv**
- price of the power production in €/MWh has a correlation of **0.19** with **hour**
- price of the power production in €/MWh has a correlation of **-0.11** with **export_FR**

This first resume of the heatmap is recapping the most important correlations with the demand and price:

- We can appreciate that the demand is highly correlated with price and hour.
- It is negatively correlated with export, which makes sense because the exporting occurs when there is more production than demand.
- The two types of power which are more correlated to the demand are both solar thermal and photovoltaic.
- The price is negatively correlated with wind and solar photovoltaic power, which means that when its production increases, it lowers the price, which makes sense as they are natural energy.
- The price is also negatively correlated with export for the same reason as the demand. When this occurs there is an excess and lowers the price.
- The price is highly correlated to hour and demand.

5. FEATURE ENGINEERING

Once the data has been analysed, it can be prepared. In the data preparation process the data will be cleaned through dropping unnecessary columns, renaming some for clarity and reformatting respective columns. The first thing that can be found is that the last date is 2020-08-31. This means that one has 3 years and 8 months of data. More data is always better in order to create a prediction algorithm but for this dataset it will be sufficient.

To improve the model, it was identified (s. figure 4) that the price is heavily influenced by the demand and expensive products. This column did not exit but can be calculated by subtracting the sum of the cheap products by the demand. By doing this the prediction score was significantly improved, as one was able to separate the expensive products it was possible to drop the fc_solar and fc_demand column. The following together with the imputation of median into import and export created the new correlation matrix in Figure 5.

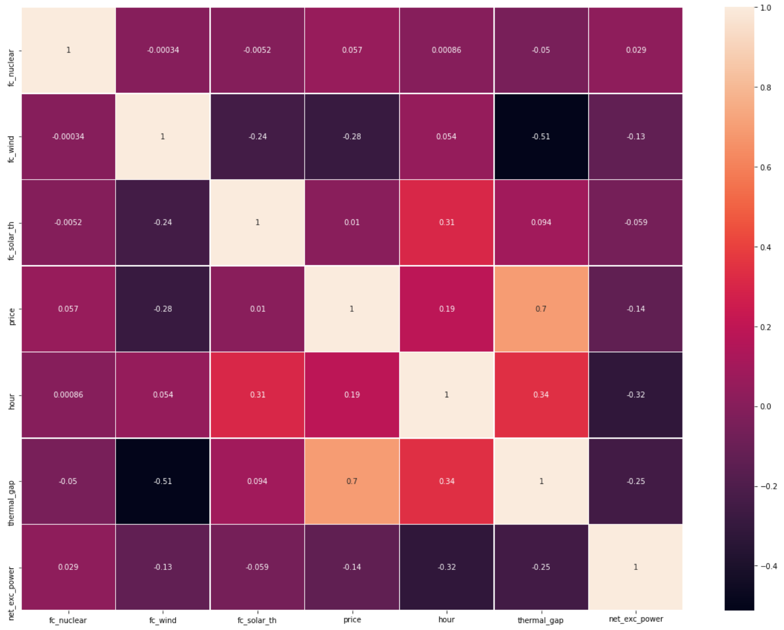


Figure 5: Correlation Matrix with Cleaned Data

To draw conclusions and make predictions about the price of power in Spain, one has to analyze how the different columns are correlated with each other. From this heat map we can see:

- price of the power production in €/MWh has a correlation of **0.7** with **thermal_gap**
- price of the power production in €/MWh has a negative correlation of **-0.14** with **net_exc_power**
- forecast of the 'expensive power' production in MWh has a negative correlation of **-0.25** with **net_exc_power**
- forecast of the 'expensive power' production in MWh has a negative correlation of **-0.51** with **fc_wind**

This resume of the heatmap is recapping the most important correlations with the demand and price after the feature engineering:

- The high correlation of the price with the thermal gap is due to the description of the thermal gap: the expensive powers.
- The price is negatively correlated with the net exchanged power because there are more imports than exports.
- The thermal gap is negatively correlated with the net exchange power because the exchange is with "cheap energy".
- Finally the thermal gap is negatively correlated with the wind power forecast because the more wind power there is, the less "expensive powers" are used.

The rest of the correlations are the same as before.
With the creation of new variables we need to select the ones we are going to use in our algorithm in order to avoid overfitting and noise.

5.1 Data engineering

After the previous correlation analysis we were able to find which variables were significant and relevant for our prediction model as well as which were not.

Used Variables:

1. Thermal_gap
2. Net_exc_power.
3. Demand
4. Solar, nuclear and wind productions
5. Price: target

Removed Variables:

Import and export were removed, as net_exc_power calculated by the difference of both.

6. MODEL TRAINING

Now that all the data preprocessing has been done, the optimal model will be found through trying various reitterations of different models. The first model that was investigated was a decision tree regressor. Here the maximum depth was changed various times from 5-25 but the optimal number that was found was 15. Even with this number the decision tree regressor did not give the optimal prediction. Then the random forest regressor was used with maximum depth of 5-25. The optimal parameters for this would be 10,15. For the n_estimators(total number of trees) the optimal was (300, 400), the numbers between 5-500 were all tried. Finally for the max features 0.5 was the optimal and 0.1-0.9 was tried, but even with all the optimal parameters there were still other models which were better. Before starting the assignment the group all thought that the XGBoost would create the best model for this dataset. We found the optimal parameters for this too but to our surprise it was still not better than the linear regression model. Therefore for the final model the Linear Regression was used with normalization false. To be able to confirm that the model of Linear Regression a cross validation was done to ensure that the model was actually the best in different contexts. Then the root mean squared error was taken but since it was negative it had to be multiplied by -1 to get the results of 11.849.

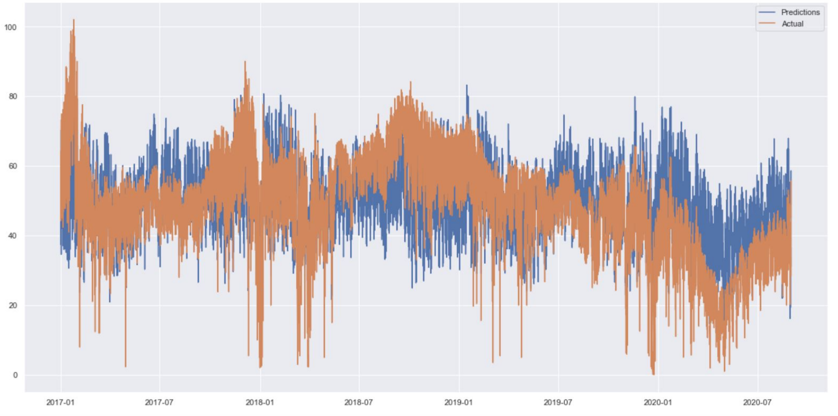


Figure 6: Time series graph showing predictions vs actual price of power in Spain

The final model was scored in a testing context where it turned out that it predicted the mean 43.56€ while the actual mean of the dataset was 48.44€. To show the accuracy of the model can be shown by the time series graph in figure 6. Here it can be seen that the model very accurately predicts the data with the orange line being the actual time series and the blue being the prediction.

6. CONCLUSION

This report focuses on finding a good prediction for the next day's electricity price. For this several steps were conducted. In the exploratory data analysis it was discovered that the mean power price per hour is around 48€ and that the distribution is not totally normally distributed. In the correlation analysis it was identified that demand is highly correlated with price and hour, negatively correlated with export as well as the price being negatively correlated with renewable energies. These findings laid the foundation for the feature engineering, where that led to creating a thermal gap function where demand was adjusted for the two energy types and the irrelevant columns have been dropped.

Hence, thermal_gap, net_exc_power, demand, production types and the price target were used in the model training. After preprocessing it was found out that surprisingly LogisticRegression was the best algorithm for the purpose of this study. Then grid search cross validation was run to identify the best mean square error (MSE). A MSE of approximately 11.8 highlights the significance of this model. Finally when comparing the scoring.csv mean of approximately 48€ with the prediction's mean of approximately 44€, it further emphasises the validity of this model. Finally in the time series it can be obtained how close this model's price predictions are to the actual prices.

To conclude, the simplicity of this model is the beauty. With the help of a simple LinearRegression algorithm, this model can definitely predict the electricity prices of the following day for expectation management in the future.

Sources:

- Price of Electricity in Spain Soars 27 per Cent in the Midst of the Cold Snap." *! Murcia Today - Archived - Price Of Electricity In Spain Soars 27 Per Cent In The Midst Of The Cold Snap*, murciatoday.com/-price-of-electricity-in-spain-soars-27-per-cent-in-the-midst-of-the-cold-snap_1547878-a.html.

- "Day-Ahead Hourly Price." *OMIE*, www.omie.es/en/market-results/daily/daily-market/daily-hourly-price.

- "Electricity Prices In Spain." *Electricity In Spain*, 15 Jan. 2021, electricityinspain.com/electricity-prices-in-spain/#:~:text=Costs%20Of%20Electricity%20In%20Spain,28.74%20per%20100%20kilowatt%2Dhour.

- "Spain Closes 2019 with 10% More Installed Renewable Power Capacity: Red Eléctrica De España." *Spain Closes 2019 with 10% More Installed Renewable Power Capacity | Red Eléctrica De España*, 19 Dec. 2019, www.ree.es/en/press-office/news/press-release/2019/12/spain-closes-2019-10-more-installed-renewable-power-capacity.