707 Final Project

IST 707

12/12/2023

Zane Alderfer, Chris Snyder

Introduction

Across many disciplines, machine-learning algorithms offer strategies to analyze relationships in datasets. With artificial intelligence, computers can apply statistical models and specified algorithms to find patterns within and make predictions about data. Some of these tests are considered supervised learning models, such as Association-Rule mining and Clustering techniques. In these models, outputs can be measured against an expected result. Other models are unsupervised, meaning that there is no expected result from the analysis. These models include Decision Trees, Random Forest, Naive-Bayes, and kNN. These findings can be applied to extensive data sets more rapidly than by using manual processes, allowing findings to be communicated quickly and with a high degree of accuracy. In our data-driven world, the increase in data availability offers financial opportunities to various industries and new research capabilities.

In the context of energy data, these tools are valuable for several reasons. First, these algorithms can organize and predict patterns in energy consumption or output. This capability is a critical step to understand current usage (as well as future projections) of energy sources. For example, SVMs can be used to forecast the demand for energy, which directly benefits companies in the energy sector. Utility companies, for example, can be better informed to efficiently manage power generation/distribution. The resulting increase in efficient energy usage benefits the environment as well, as less resources are needlessly used. Focusing on renewable energy, these predictions benefit energy efficiency by predicting the highest energy output given a variety of conditions. For example, machine-learning algorithms can evaluate wind energy generation data from a variety of locations to determine the most attractive location for a wind

turbine farm. By applying these tools to energy data, researchers are better equipped to fulfill the growing demand for cleaner, cost-effective energy sources.

Analysis and Models

1. About the Data

The data being analyzed is collected from Kaggle, titled "Hourly Energy Demand Generation and Weather." This data contains 29 variables and over 35,000 rows. Columns in this dataset include time (Hour and Month), metrics for outputs of different energy sources (solar, petroleum, wind, biomass, etc.), and forecasts for weather and outputs. Energy outputs are recorded as numeric values in MW (Megawatts). A single column contains character inputs ("Month") and two columns contain logical values ("True/False"); the remaining variables are all numeric. The data collection period for this data starts at the beginning of 2015 and is collected continuously until the end of 2018 (four years). The initial dataset contains over 70,000 null values, mainly the result of two variables containing an overwhelming majority of null inputs. However, there are also gaps in the collection time present for entire rows in the dataset. To effectively experiment with machine-learning algorithms, several preprocessing steps are necessary.

Preprocessing Steps

The data was downloaded from Kaggle and loaded into RStudio. Libraries necessary for research steps were loaded and subsets of the original source data were created. The first subset types (ex. "df4") were created to eliminate any null values in the data; this was accomplished using the na.omit() function. Next, a subset that combined weather data with energy outputs from one the collection cities (Valencia) was created. Other subsets included changes to column variable types (ex. "character" to "numeric"), combined (summed) adjacent energy source

outputs, and calculations from other columns. One of these subsets ("EnergyAG2") includes aggregated columns that use high emission ("fossil") energy sources (ex. fossil coal, fossil peat, etc.) as well as combined outputs for wind/hydro generation sites. In another subset, the values from the forecasted total load (outputs) are divided by the actual total load to assess accuracy. The last subsets aggregate all energy outputs from sources that generate high emissions, separated from the aggregated outputs of low/zero emission sources.

Research Process/Steps

To accomplish all the intended aims of the project, tests were conducted on a variety of the created subsets. In the first round of testing, outputs from different energy sources were assessed by month/time. In these tests, weather data was not included, as this data was only available for one of the collection sites; including this data would have drastically reduced the total observations to be compared. In this stage, energy output types were compared throughout the collection period and by month. From these figures, patterns involving target variables were detected and helped provide direction for future tests. Next, Decision Trees were constructed for the aggregated subset, with the target variables "Fossil_High_Emissions" and "Low_Zero_Emissions." Association Rules were next constructed with 5 rules singled out with implications that could have some insight on how the weather impacts energy output. Then Naive Bayes predictions were created. Using 8 folds and the weather and energy data, ideally the Naive Bayes model would be able to predict the month that the data is from with some consistency.

Results & Interpretations

A collection of plots was created prior to running the algorithms to illustrate the data **(Figures 1-6)**. From these figures, there are observable patterns that can guide future testing. For

example, **Figure 1** shows a comparison in Spain's energy output forecasts versus actual output (amount in Megawatts). From this figure, there is a clear linear relationship between the forecasted output and the actual total output. This indicates a relatively high degree of accuracy for Spain's predictive models, which is helpful to predict changing outputs. The few outliers (less than 0.5% of the data) have a maximum difference of ~30% and tend to be the result of overestimating forecasted output. **Figure 2** shows the output of energy source types with high emissions (ex. fossil fuels) throughout the collection period of the data. **Figure 3** shows outputs of sources with low/zero emissions. Examples of these source types are nuclear, wind, solar, and hydroelectric. Both figures include a red trendline and show energy outputs rise and fall each year in a cyclical pattern, indicating the energy output demand fluctuates throughout the year. Even outputs within the same month vary significantly by each day. These figures do not appear to show a clear pattern (increasing/decreasing) in the trend of outputs for either source.

Figure 1

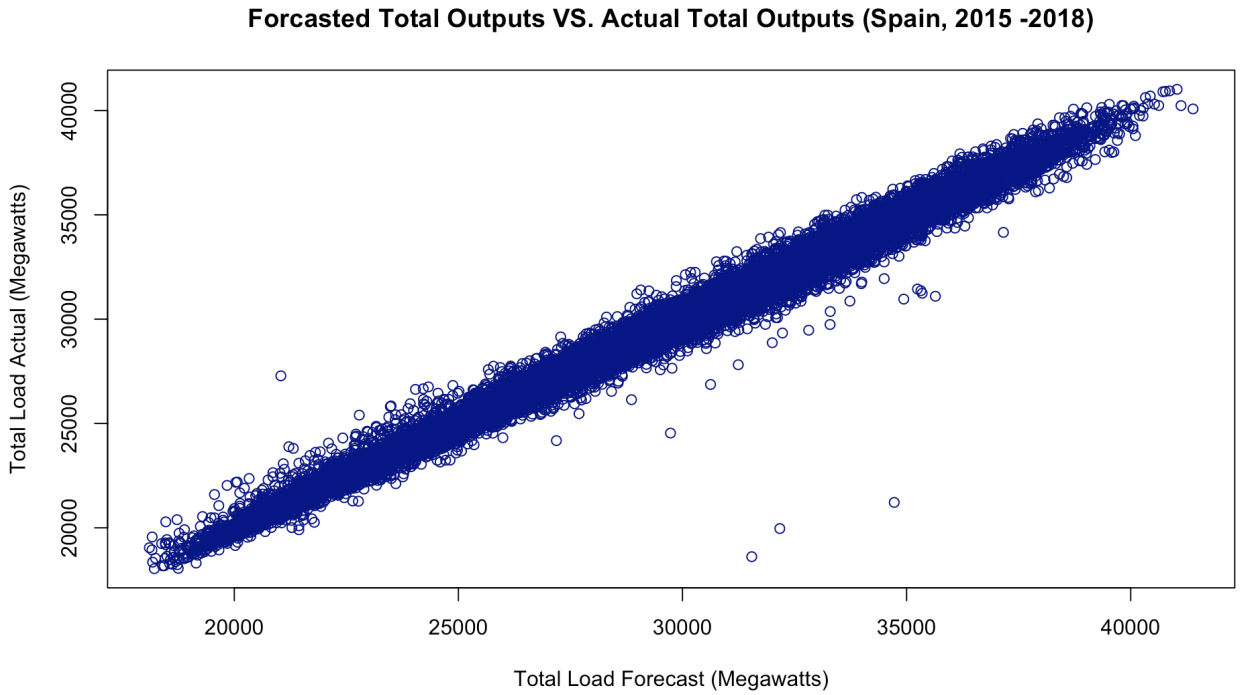


Figure 2

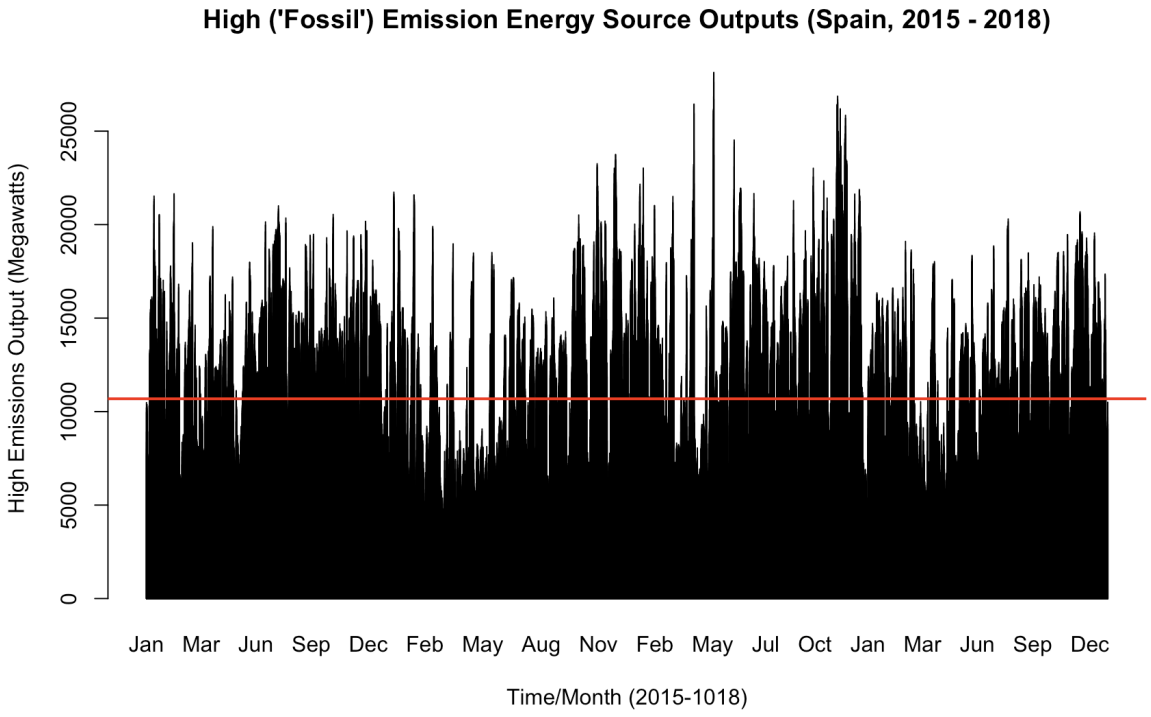
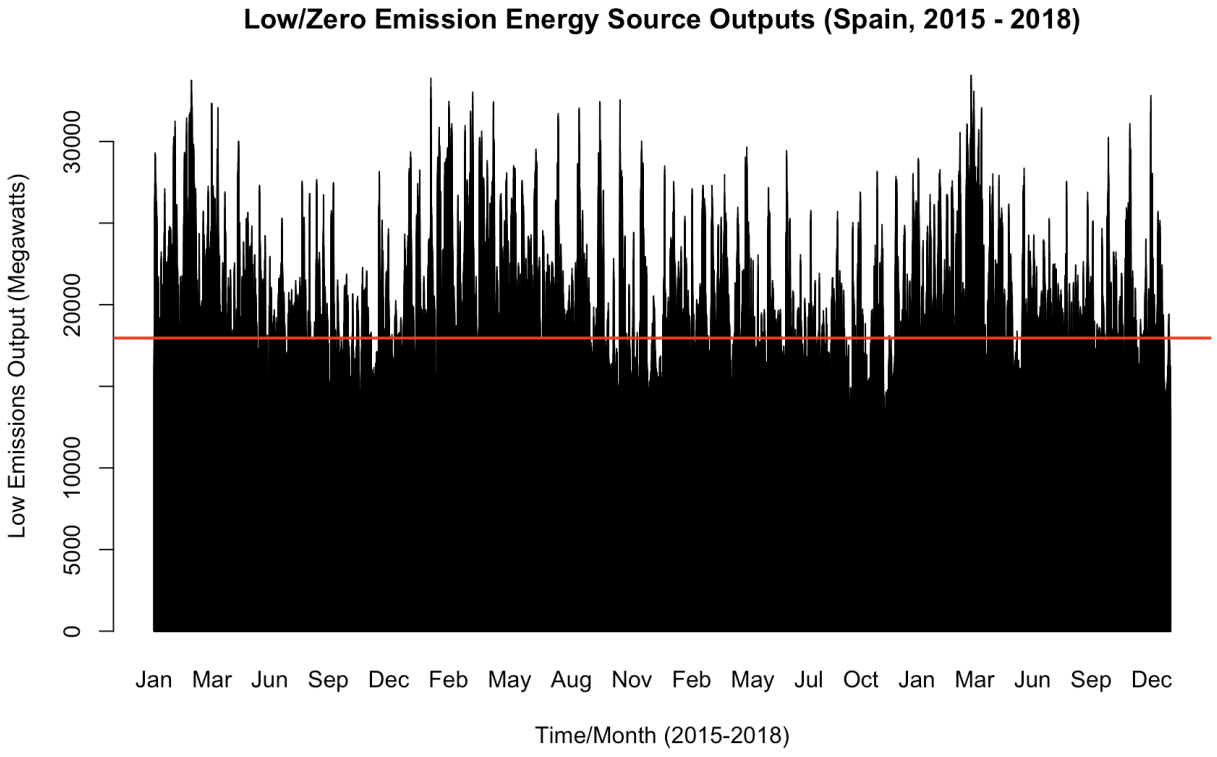


Figure 3



Focusing on month, **Figure 4** and **Figure 5** show the aggregated energy outputs of high and low emission sources, respectively. Both plots show fluctuations between energy output sources, with high emission source output varying up to 30% and low/zero emission sources varying 17% on average. November (Winter months) outputs are highest for high emissions sources, while March (Spring months) outputs are highest for renewable sources. Use of these source types is generally consistent with Spain's seasonality and climate; while Spain's climate is incredibly diverse, the summers are generally dry and hot, with higher rainfall the rest of the year. **Figure 6** shows this split between the aggregated source outcomes. On average, the output of low/zero emissions sources was ~60% higher than outputs from high emissions sources.

Figure 4

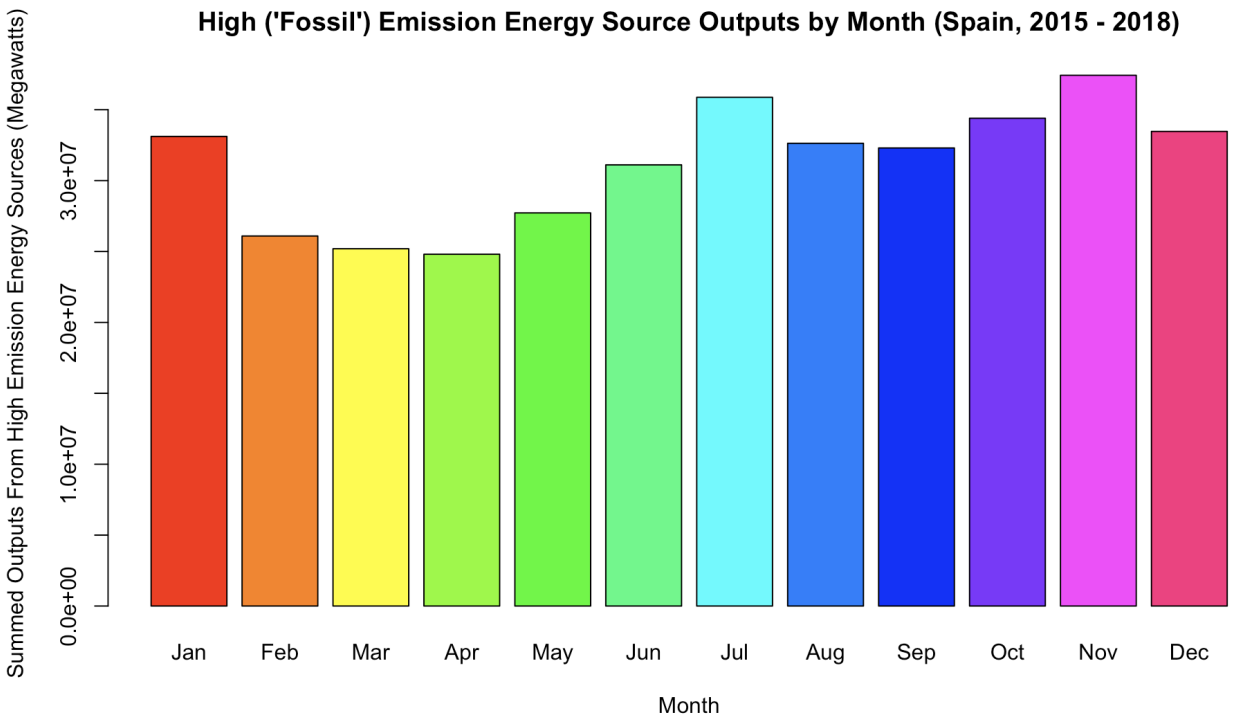


Figure 5

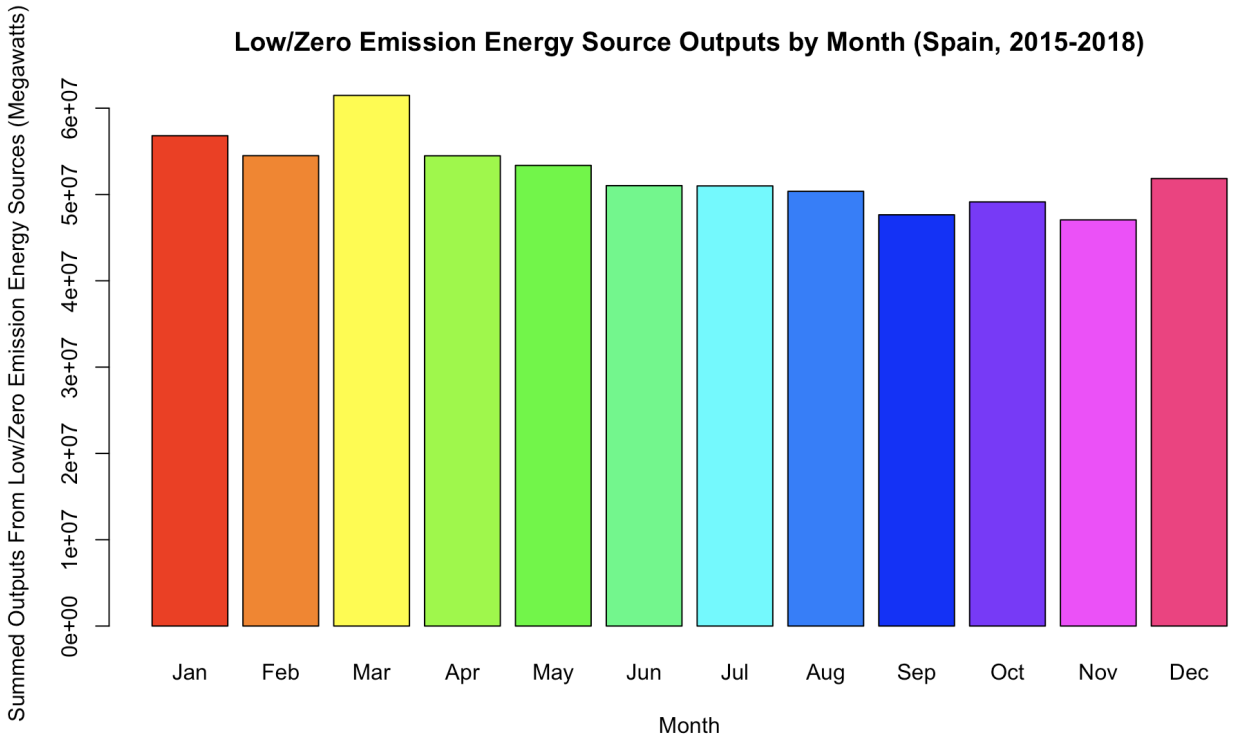
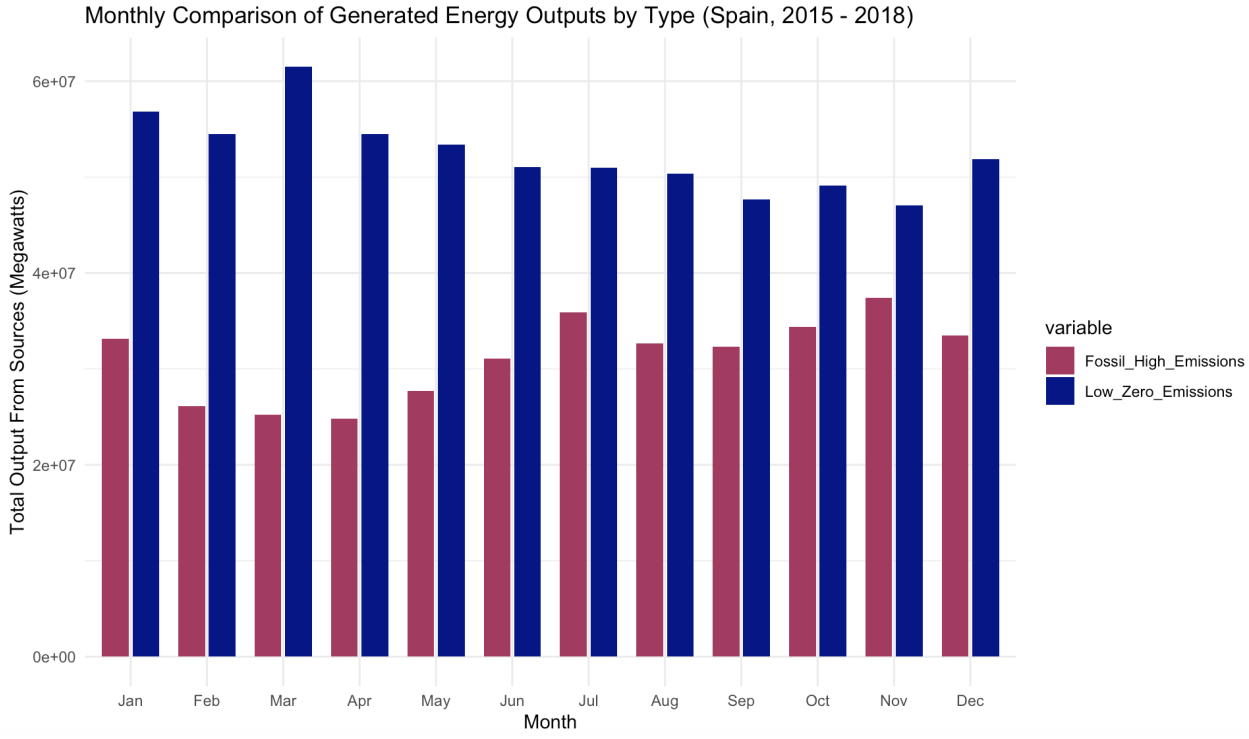


Figure 6



When the target variable is “Fossil High Emissions,” applying the Decision Tree algorithm to this subset reveals the key splitting variables are hard coal and fossil gas (**Figure 7**). This suggests that, out of the seven sources outlined in the dataset, these two sources offer the highest insight into the output amount. When both gas and hard coal outputs are high, these are the primary determiners/contributors of those high outputs. Focusing on **Figure 8**, the same splitting variable binary is not present. Instead, the pattern for high output of “Low/Zero Emissions” source appears to rely on wind (onshore) and hydropower (reservoir) energy generation. In this tree, it appears that solar outputs are relatively consistent and total outputs do not rely as heavily on this source. Other sources with high consistent outputs, like nuclear generation, do not appear in the model. This absence indicates less fluctuation of this source’s use throughout the year. These models also help explain fluctuations in energy outputs throughout the year through seasonality. Spain has wildly different seasons, and these can look different depending on topography, latitude, and proximity to water. However, energy sources in these models can be explained (in part) by the seasonal links to temperature, wind, and rainfall. For example, Spain typically experiences wetter late-Autumn, Winter, and Spring months, accounting for higher outputs from renewable sources from the available storms, rainfall, and snowmelt. Comparatively, nonrenewable outputs may be highest during Summer and early-Autumn months when rainfall (water) is more scarce.

Figure 7

Decision Tree 1: Predicting Outputs from High Emission Sources

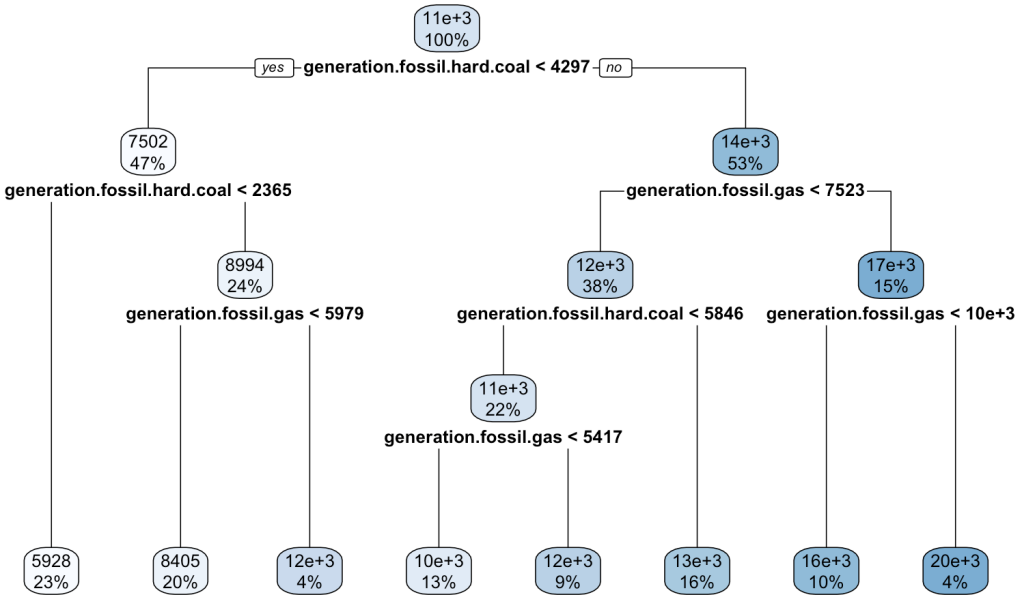
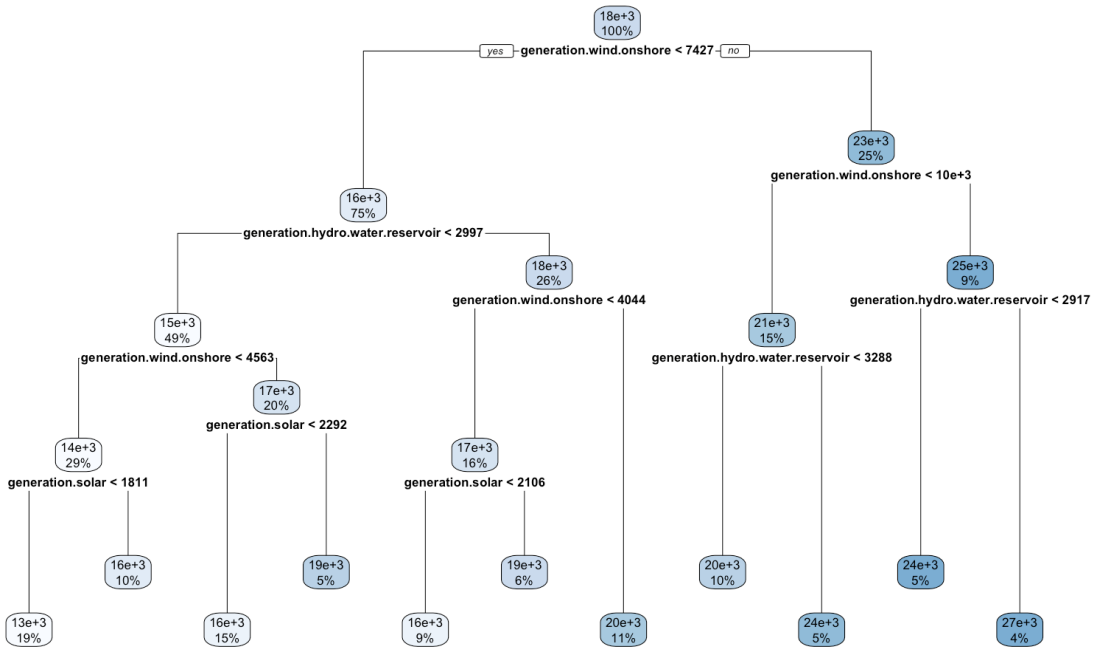


Figure 8

Decision Tree 2: Predicting Outputs from Low/Zero Emission Sources



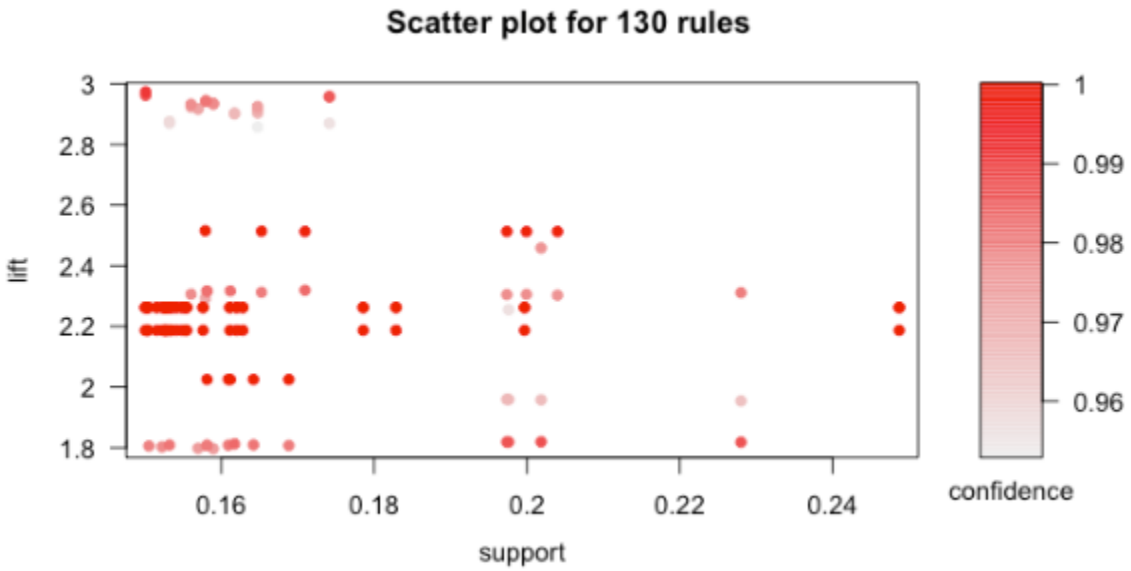
The weather seems to not just impact the total amount of energy output per day but also the type of energy that is being yielded the most depending on the weather. The 5 rules below in **Figure 9** were specifically picked due to their significance of relevancy. The criteria for all the rules were a .15 support and a .95 confidence. Rule 1 and 2 are somewhat in the same boat as they both are related to solar energy. When the temperature is high, solar generation is high and although the second rule would suggest that the weather is cloudy, you can assume this is during summer days as the solar generation is still high. Looking at the third rule with a 0.958 confidence, a cloudy day seems to drive a larger energy output. Finally looking at the last two rules, a clearer day seems to lead to a higher output of onshore wind energy versus a cloudy day tends to result in less output of onshore wind energy. **Figure 10** then shows a scatter plot indicating that it's rare when a set of rules has support above .20 as well as very few rules resulting in a lift of close to 3.

Figure 9

5 Association Rules: Interesting Implications

```
    lhs                                                rhs                                    support confidence  coverage     lift count
[1] {forecast.solar.day.ahead=[1.7e+03,5.84e+03],
     temp=[293,316],
     temp_min=[291,315]}                        => {generation.solar=[1.62e+03,5.79e+03]} 0.1560385  0.9748439 0.1600651 2.923864  5464
    lhs                                                rhs                                    support confidence  coverage     lift count
[1] {forecast.solar.day.ahead=[1.7e+03,5.84e+03],
     clouds_all=[20,100],
     weather_main=clouds}                       => {generation.solar=[1.62e+03,5.79e+03]} 0.1569523  0.9725712 0.1613788 2.917047  5496
    lhs                                          rhs                                        support confidence  coverage     lift count
[1] {total.load.actual=[3.11e+04,4.1e+04],
     clouds_all=[20,100],
     weather_main=clouds}                  => {total.load.forecast=[3.12e+04,4.14e+04]} 0.1532113  0.9587205 0.1598081 2.875997  5365
    lhs                                                rhs                    support confidence  coverage     lift count
[1] {forecast.wind.onshore.day.ahead=[6.42e+03,1.74e+04],
     weather_main=clear,
     weather_description=sky is clear}          => {clouds_all=[0,20)} 0.1504127          1 0.1504127 2.186922  5267
    lhs                                                rhs                     support confidence  coverage     lift count
[1] {generation.wind.onshore=[3.54e+03,6.44e+03),
     forecast.wind.onshore.day.ahead=[3.57e+03,6.42e+03),
     weather_main=clouds}                       => {clouds_all=[20,100]} 0.1522403  0.9783447 0.1556101 1.802615  5331
```

Figure 10



Using the energy output and weather indicators, it would be interesting to see if a valid prediction could be made for each month. Using the Naive Bayes prediction model, 3 of the 8 folds that were produced are shown below in **Figure 11**, **Figure 12**, and **Figure 13**. There are similarities across the fold outputs in terms of months with the most predictions. January and July are the two months that seem to be predicted the most based on the energy output and weathers indicators. This could be for a number of reasons such as January having similar weather to December, February or even March, so the predictor could associate those months' data with January. This could happen similarly for July with Summer months. Energy output could also vary more than weather so the varying energy output could also lead to the predictions being skewed somewhat. It is also interesting to note that April and September are rarely the model's prediction. This could be that these months have the most varying weather and energy output leading to difficulty creating enough support to predict those months. **Figure 11** and **Figure 12** have very similar predictions for each month but **Figure 13** looks a little different from the other two which could indicate that the data in this dataset may not be ideal for

predicting the month. Finally, **Figure 14** reiterates what the Naive Bayes bar graphs showed. Once again, January and July seem to be the common picks of the prediction model.

Figure 11

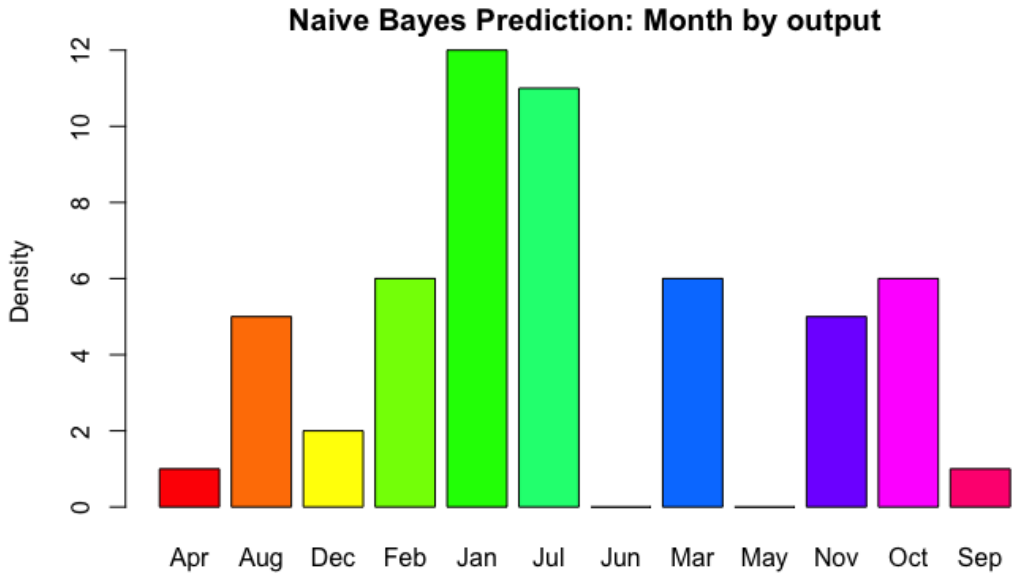


Figure 12

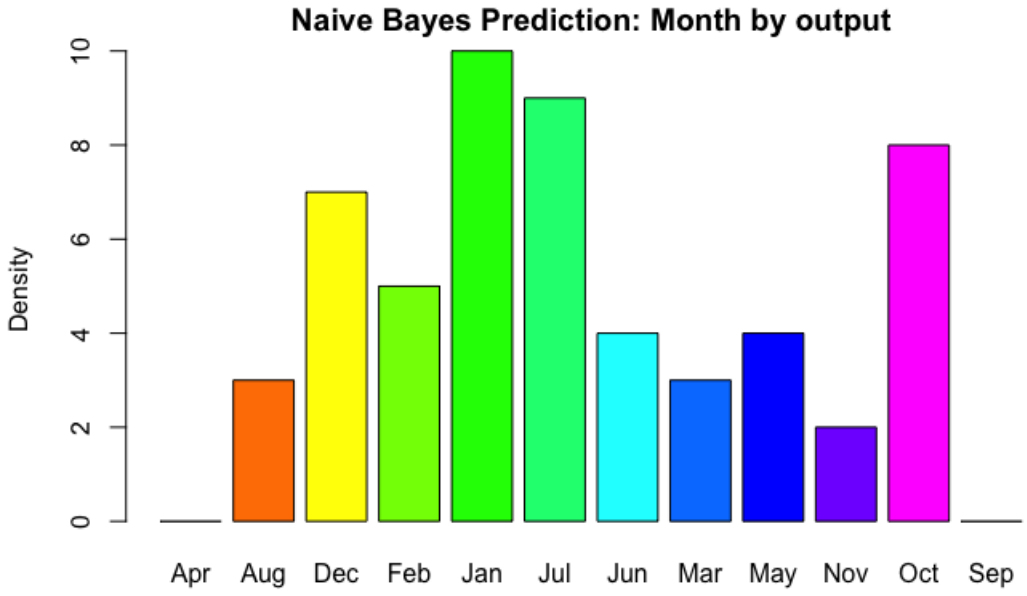


Figure 13

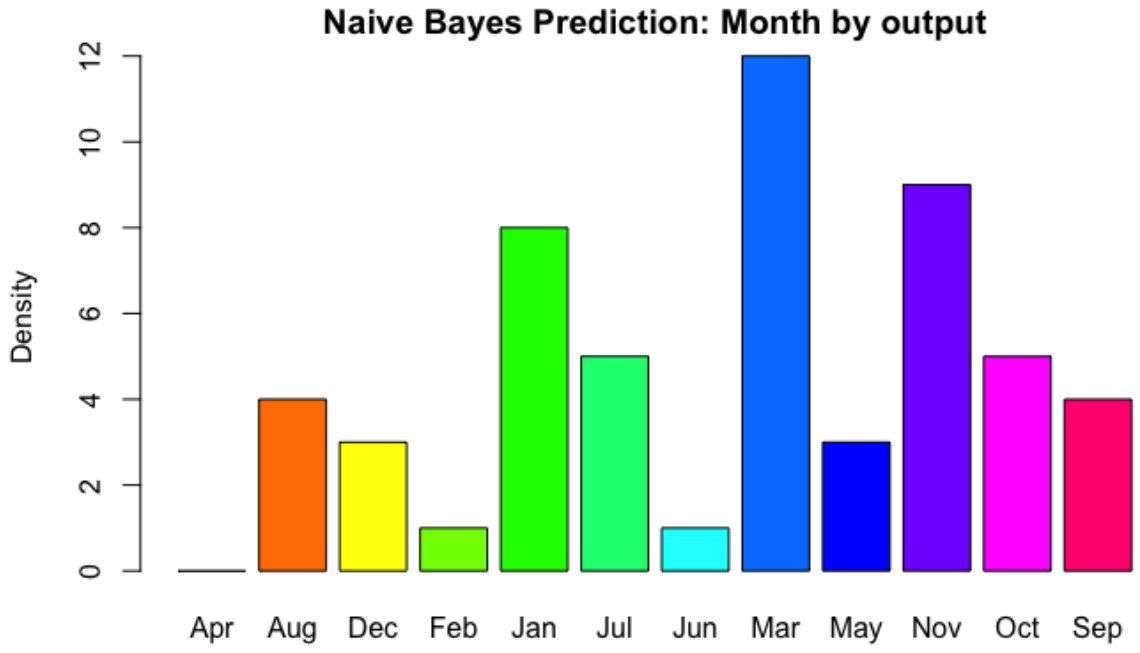


Figure 14

```
    1  2  3  4  5  6  7  8  9 10 11 12
1   4  0  1  0  0  0  0  1  0  0  0  0
2   0  6  0  1  1  3  4  3  4  4  1  5
3   2  0  2  2  1  1  1  3  3  1  4  1
4   1  0  0  2  5  2  3  6  1  3  0  0
5   3  4 11 13 11  4  3  7  6  7 11  8
6   3 10  4  2  2  8  6  3  3  7  8 11
7   0  3  0  1  0  1  5  0  2  0  0  3
8   5  0  4 11  6  1  4 17  2  1  1  0
9   1  3  1  0  0  1  5  1 10  0  0  1
10  1  0  7  1  1  5  0  2  2 14  5  4
11  4  2  6  1  3  2  4  0  6  5  8  4
12  2  4  2  1  3  1  3  2  2  0  2  2
 [1] Jul Jan Nov Jan Aug May Nov Jan Jan May Mar Jun Jan Mar Feb Jul Mar Aug Nov Oct Jul Sep Jan
Apr Nov Oct Jan Jul Dec Aug Jan Nov Jan Jul Jul Jan
[37] Oct Mar Oct Jan Jan Mar Nov Jan Nov Dec Jul Mar Jul Jul Dec Mar Mar Jan
Levels: Apr Aug Dec Feb Jan Jul Jun Mar May Nov Oct Sep
```

Conclusion

The biggest takeaway from these models is that the correlation between weather and energy output is readily apparent but the same can't be said for the correlation between the combination of weather and energy with month. When just looking at the total output for both low and high emission energy, the amount of output is relatively consistent throughout the year. This is seen further in the Naive Bayes prediction model as the model wasn't particularly great at predicting the correct month. When looking at the decision tree model and the association rules created, not only can the breakdown of energy output on a given day be predicted accurately but a claim for the type of weather and amount of certain types of energy output can be reasonably made based on the data. This data has proven adequate in answering questions relating to correlations between weather and output while also proving a relative lack of correlation between month of year and energy output with a few exceptions such as temperature and solar energy but not enough of an impact to be noticed in overall energy output.

This study could be improved if it were to be extended to countries or cities with more erratic weather than Spain. That isn't to say that the years studied in this report may have been relatively mellow in terms of fluctuating weather in Spain. As the world transitions to more renewable/low emission forms of energy, it would be important to take into account the reduction in high emission energies if this study were performed again with updated data. Keeping all of that in mind, the evolution of types of energy sources and how weather may or may not impact them will be an important topic for years to come.

Works Cited

Jhana, N. (2019, October 10). *Hourly Energy Demand Generation and weather*. Kaggle. https://www.kaggle.com/datasets/nicholasjhana/energy-consumption-generation-prices-and-weather

Weatheronline.co.uk, W. (2023). *Spain*. https://www.weatheronline.co.uk/reports/climate/Spain.htm#