IST 707 Final Project

Zane Alderfer & Chris Snyder

December 10th, 2023

Introduction

-Why do people rely on different energy sources?

-Many countries are transitioning away from fossil fuels toward more sustainable sources

-Spain is noted for being a leader in renewable energy (renewables are Spain’s current primary energy source)

-Learning more about the conditions that favor renewable energy outputs.

-Understanding changes in Spain’s energy sources (could be applied to other countries).

-Creating models (using machine-learning algorithms) can help answer relevant questions related to energy usage in Spain.

Data Description

-The dataset for this project was obtained on Kaggle.com, called "Hourly energy demand generation and weather."

-This data contains 29 variables (and over 35,000 rows), including time, weather forecast, and metrics for outputs of different energy sources (ex. Geothermal, natural gas, biomass).

-The data collection time is 4 years (2015-2018).

Sample View

time	generation.biomass	generation.fossil.brown.coal.lignite	generation.fossil.coal.derived.gas	generation.fossil.gas	generation.fossil.hard.coal	generation.fossil.oil
Jan	447	329	0	4844	4821	162
Jan	449	328	0	5196	4755	158
Jan	448	323	0	4857	4581	157
Jan	438	254	0	4314	4131	160
Jan	428	187	0	4130	3840	156
Jan	410	178	0	4038	3590	156
Jan	401	172	0	4040	3368	158
Jan	408	172	0	4030	3208	160
Jan	413	177	0	4052	3335	161
Jan	419	177	0	4137	3437	163
Jan	422	173	0	4059	3516	167
Jan	421	226	0	3931	3845	166
Jan	428	303	0	3784	4220	167
Jan	425	288	0	3754	4404	167
Jan	423	260	0	3779	4256	166
Jan	421	183	0	3708	4038	160
Jan	422	256	0	3813	4191	163
Jan	426	322	0	3967	4707	165
Jan	427	282	0	4756	4756	164
Jan	442	303	0	4410	4918	147

Preprocessing Steps

-Downloaded data from Kaggle and loaded into RStudio

-Removed null values and columns

-Created subsets

-Aggregating **energy output types**

-Aggregating **weather data with output data**

-Changing **variable types** (ex. Transaction Data)

-Calculations from columns (ex. Forecasted outputs vs. actual outputs)

Preprocessing Steps (Cont.)

-The two primary subsets for experiments are:

Subset 1

-Aggregated output types (ex. low emissions outputs vs. high emission outputs)

Subset 2

-Energy outputs with combined weather data (ex. Output and weather data for Valencia, Spain (2015-2018).

Primary Investigative ("Problem") Questions

1) How have the outputs of different energy sources **evolved over the data-collection period (2015-2018)?**

2) Is there a relationship between **time of year** and output of difference energy sources?

3) Is there a **relationship between weather and output** of different energy sources? (Ex. Do overcast/dry/windless days tend to see higher energy outputs of non-renewable energy types?)

Experiment Design

-Created subsets to answer our target questions

-Compared output sources throughout recording period and by month

-Created **Decision Trees** to compare sources from different output values

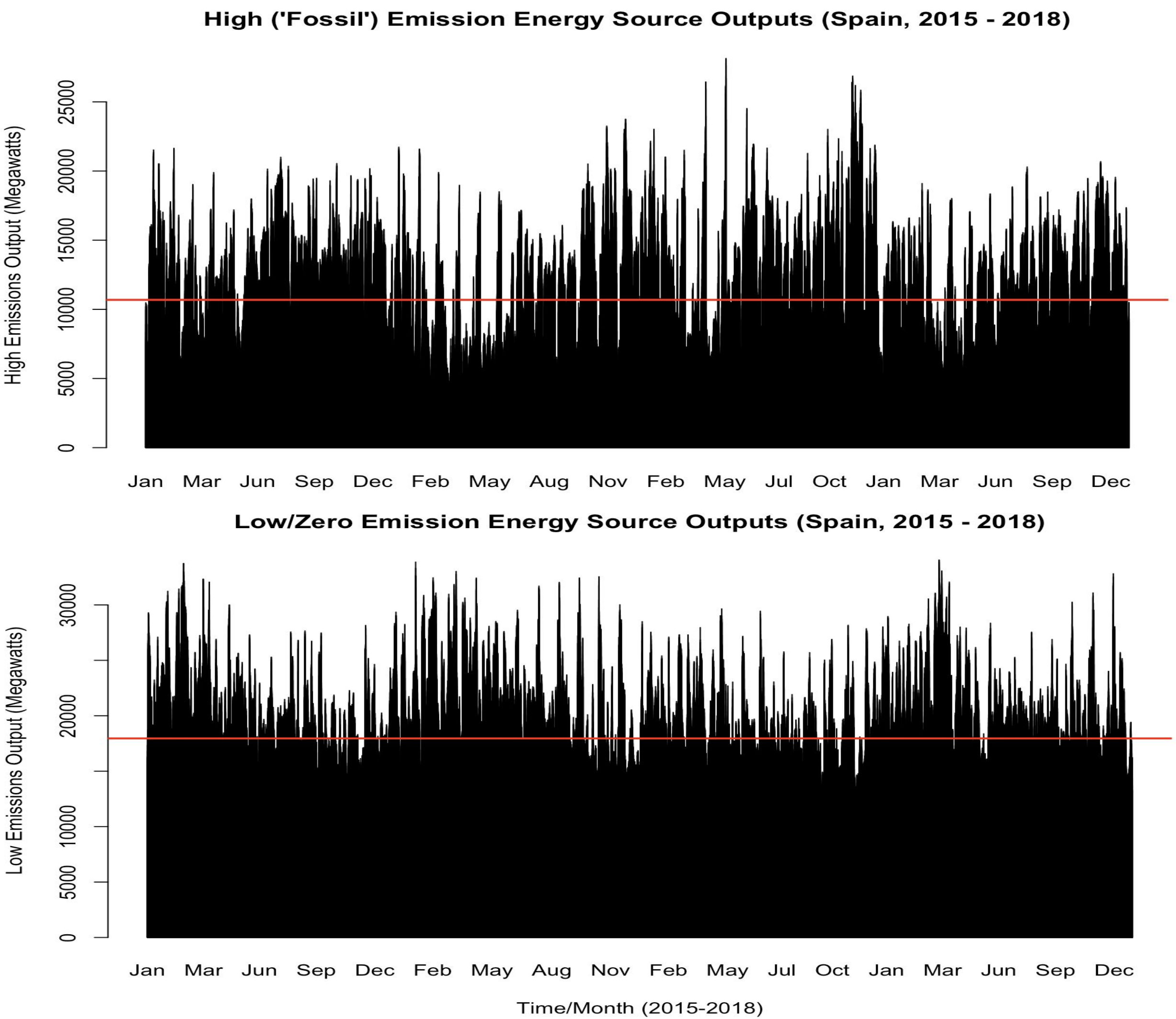
-Performed **AR mining** to find rules linking weather to output source(s)

-Performed **Naive-Bayes** to predict month, provided input data

-Created visualizations with subsets to answer questions

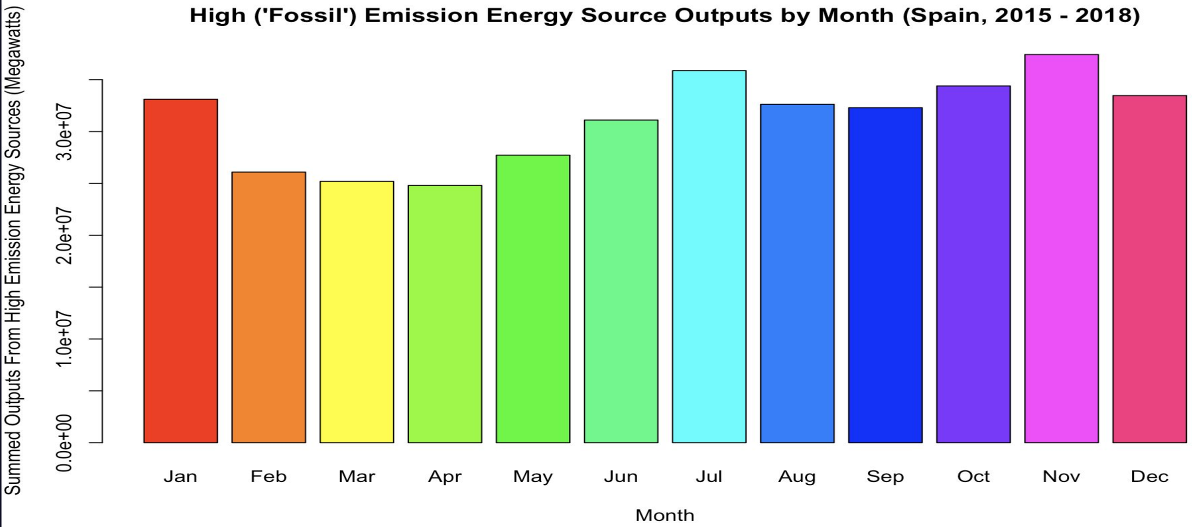
Results 1:

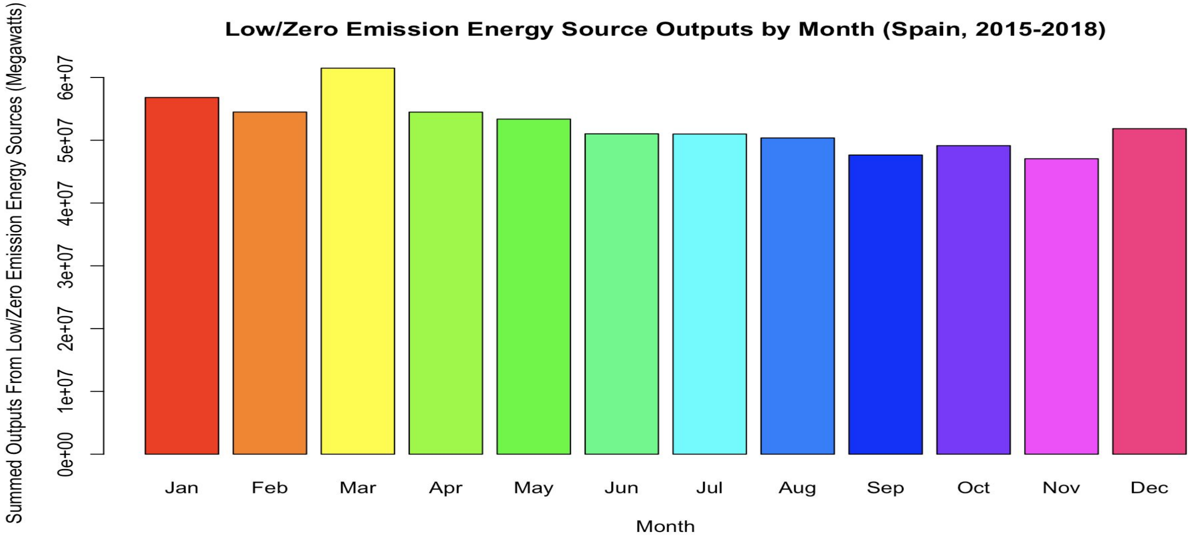
Output Source Type from 2015-2018



Results 2:

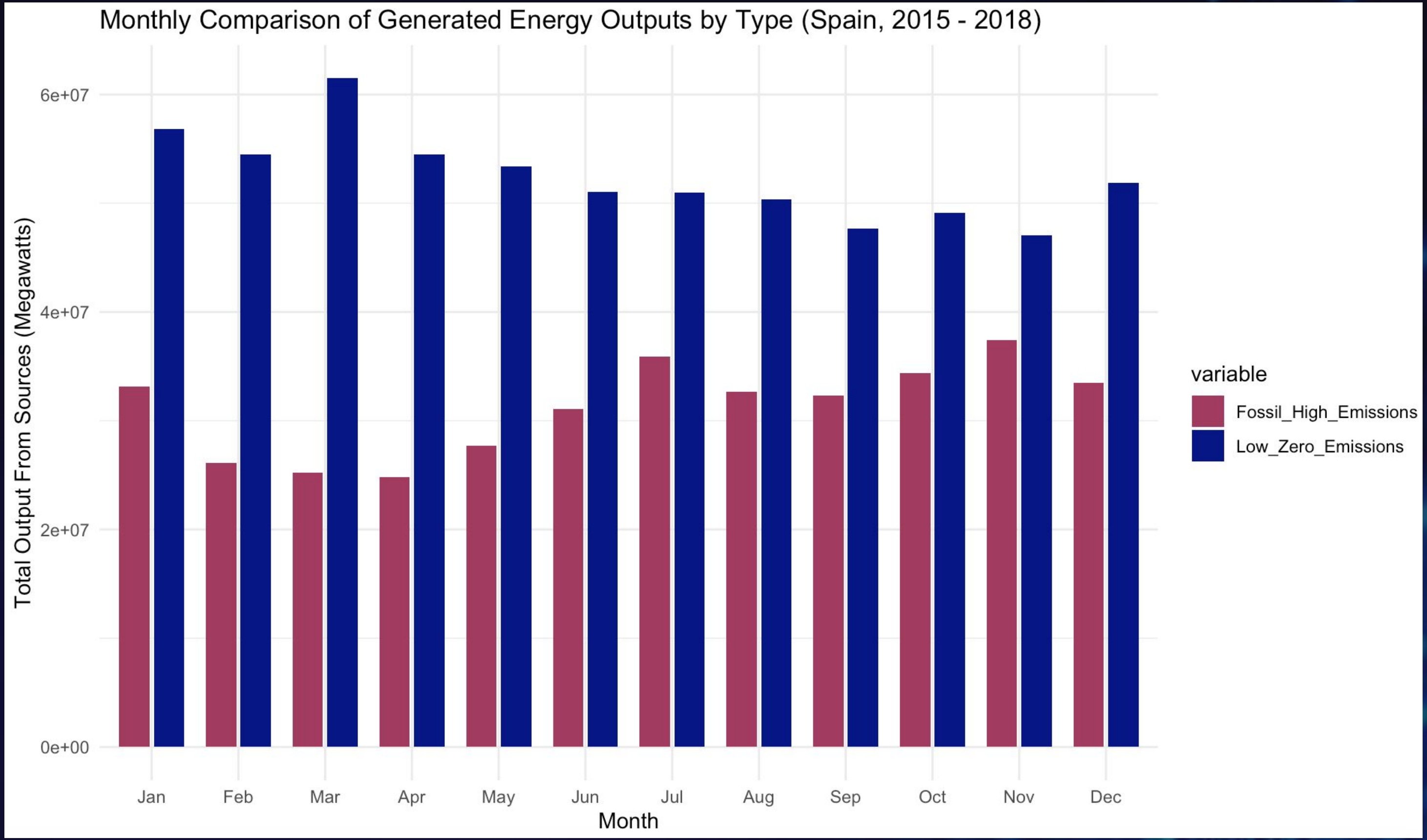
Time of Year vs. Energy Output Source





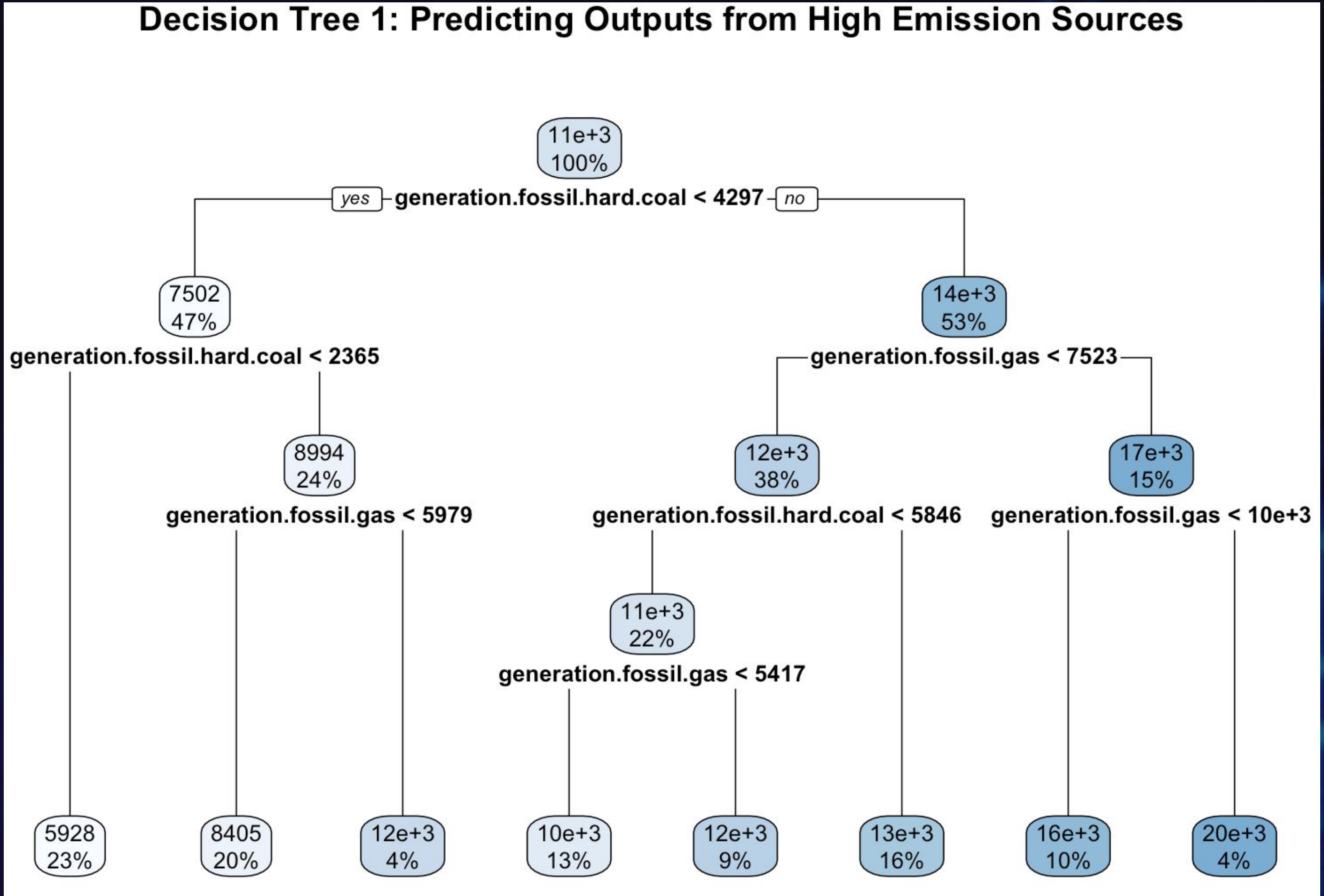
Results 3:

Time of Year vs. Energy Output Source (Combined)



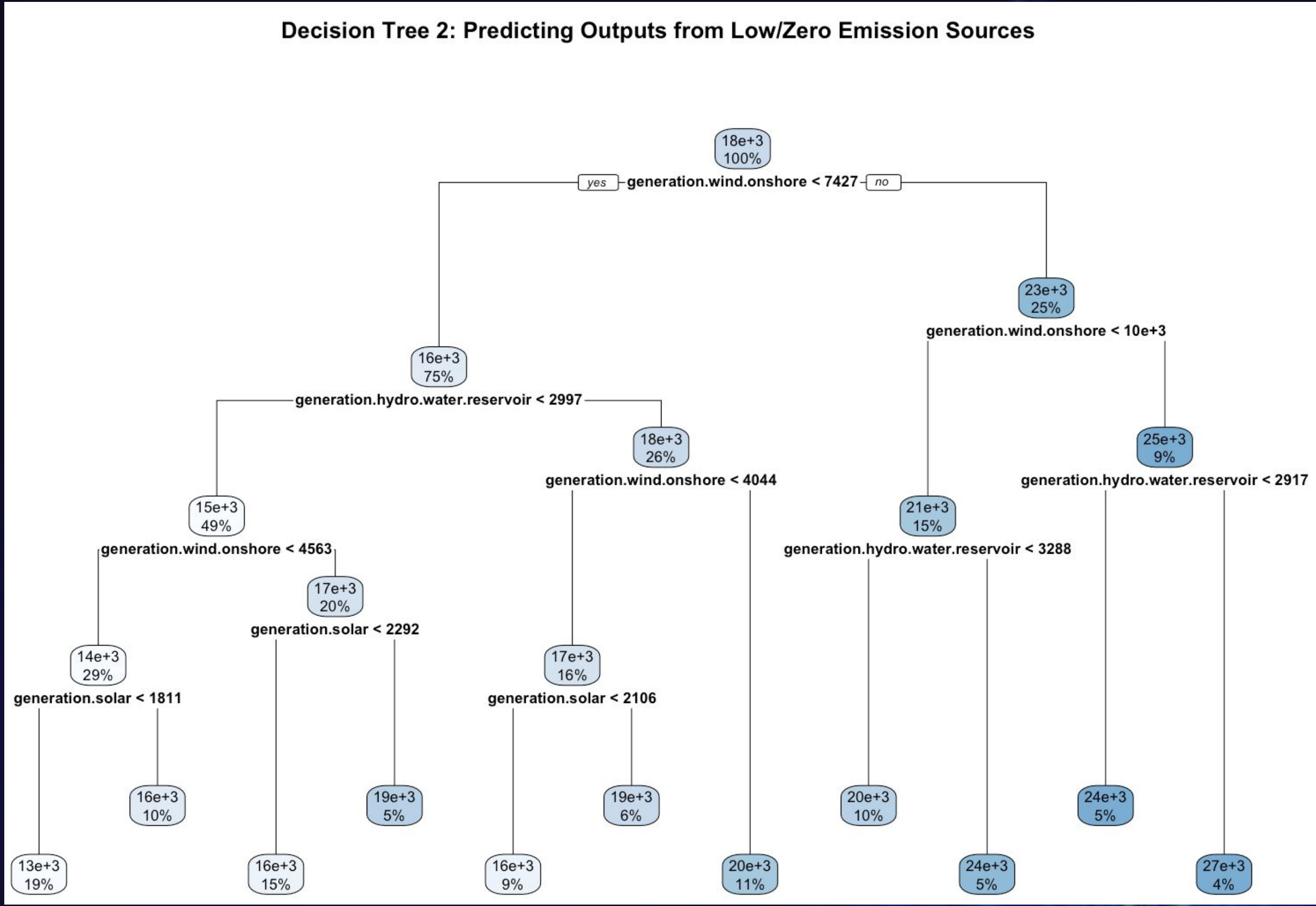
Results 4:

Decision Tree for high-emission output source types



Results 5:

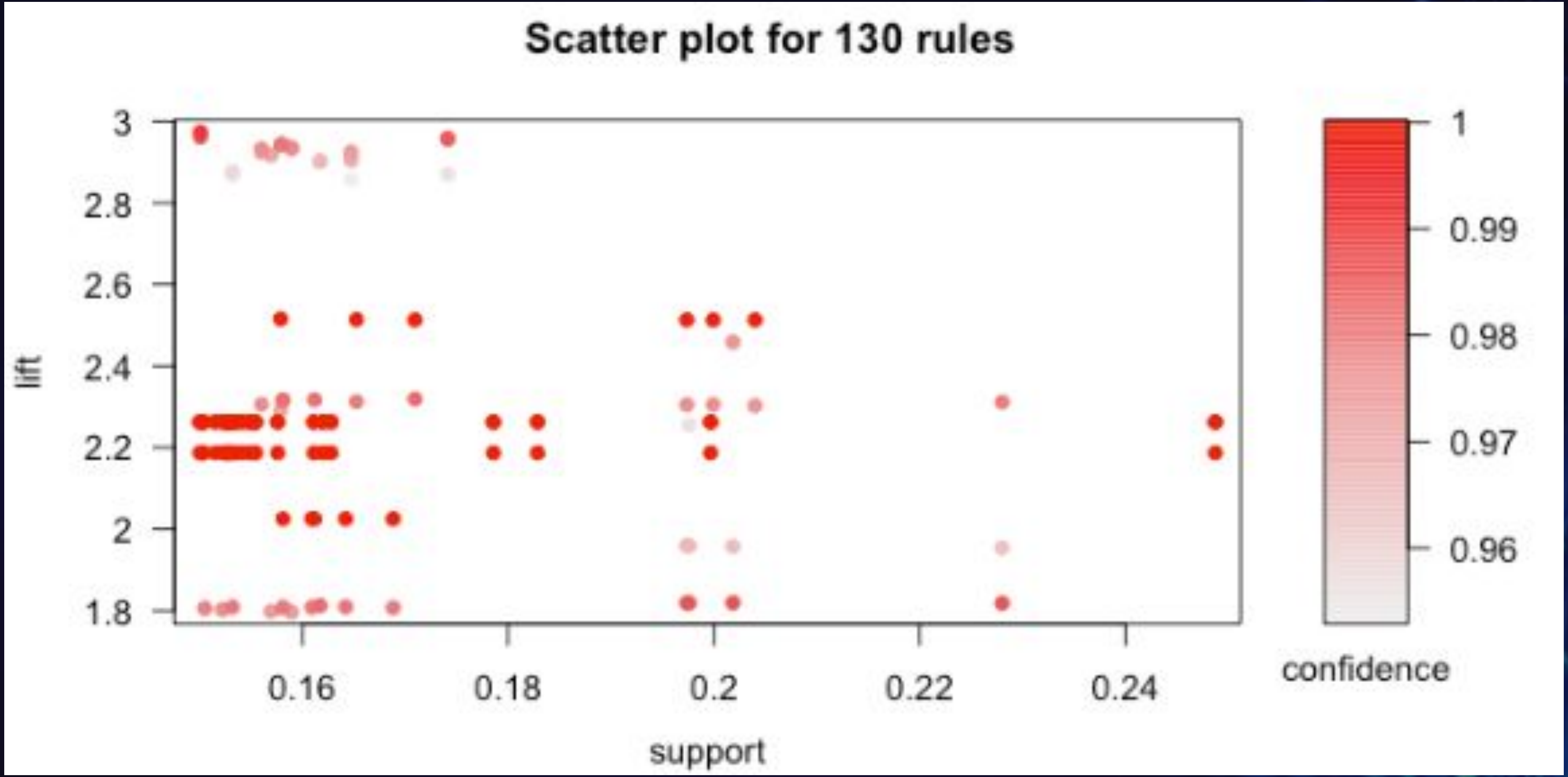
Decision Tree for low-emission output source types



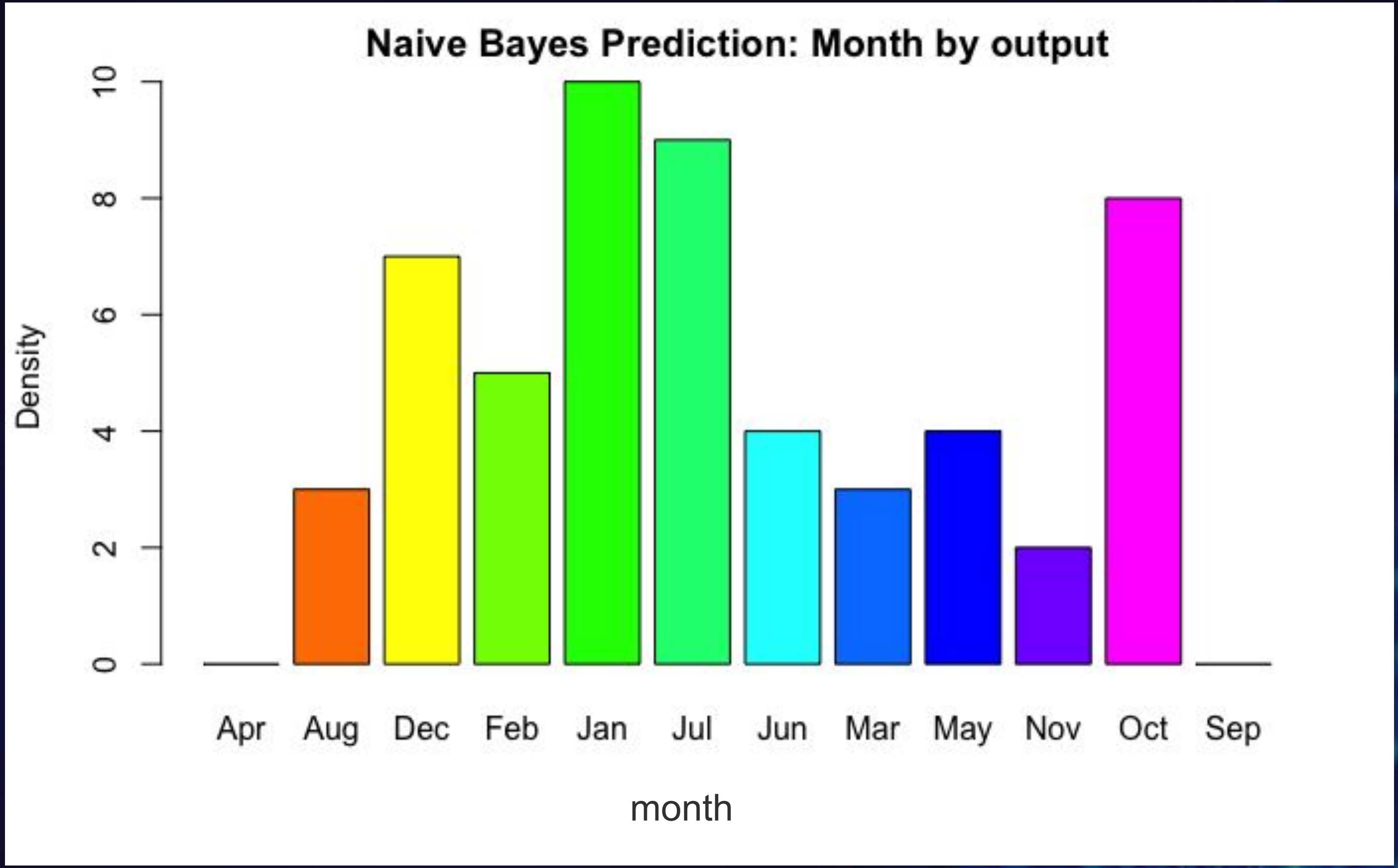
Results 6: AR Mining

```
    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
```

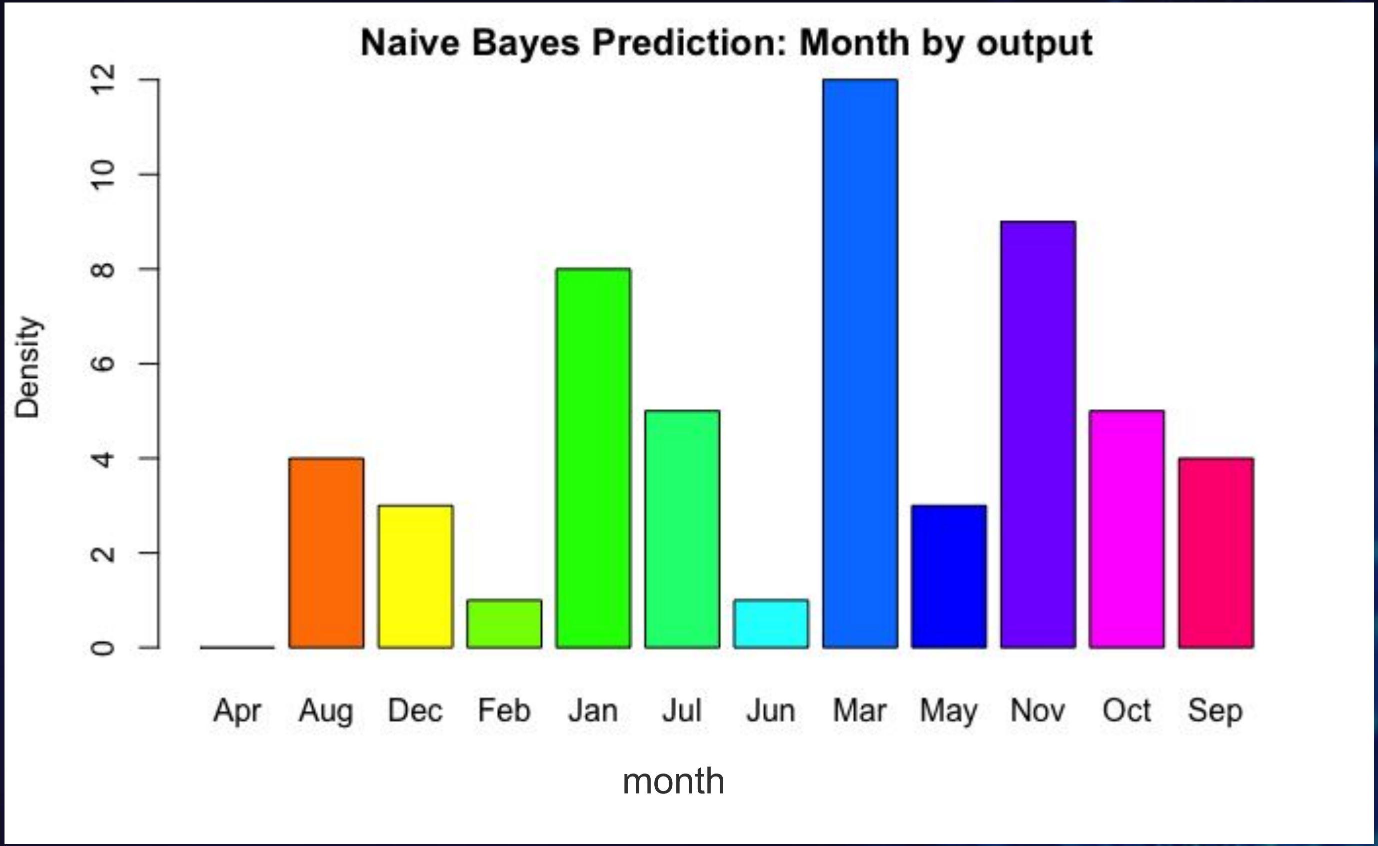
Results 7: AR Mining (pt. 2)



Results 8: Naive Bayes



Results 9: Naive Bayes



Interpretations

-In this dataset, **low/zero emission energy source outputs are consistently higher** than high emission source outputs

-**Weather and energy outputs (type) are linked**. For example, solar outputs are positively linked with sunny weather; hydro power for precipitation, etc.

-Energy source **outputs fluctuate throughout the year**; high emission outputs are highest July-January; low emission outputs are highest December-March.

-There is no definitive pattern showing an increase/decrease in outputs of all types throughout the data collection period.

Next Steps

- Produce visualizations to compare data from more years.
  - This will help us more confidently assess relationships between outputs, weather, and time.

- Perform these same tests (ex. AR mining, Naive-Bayes) and compare results.
  - This will increase confidence of our output results for Spain.

- Perform these tests and compare results from energy data in other countries.
  - This will allow us to experiment with even more factors that impact energy sources used.
  - Particularly countries with more erratic weather + increased renewable sources/outputs

Thank You

Questions?