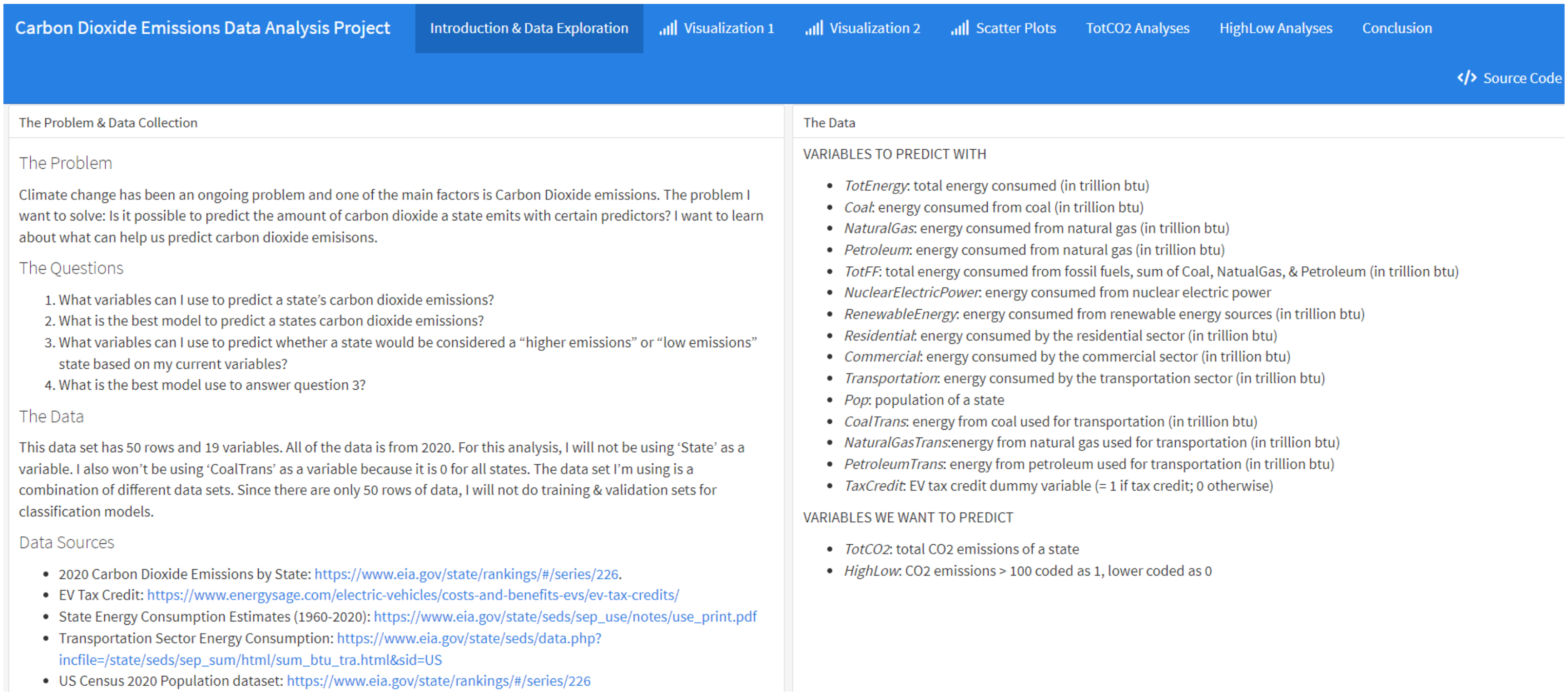


The Problem & Data Collection

The Problem

Climate change has been an ongoing problem and one of the main factors is Carbon Dioxide emissions. The problem I want to solve: Is it possible to predict the amount of carbon dioxide a state emits with certain predictors? I want to learn about what can help us predict carbon dioxide emisisons.

The Questions

1. What variables can I use to predict a state's carbon dioxide emissions?
2. What is the best model to predict a states carbon dioxide emissions?
3. What variables can I use to predict whether a state would be considered a "higher emissions" or "low emissions" state based on my current variables?
4. What is the best model use to answer question 3?

The Data

This data set has 50 rows and 19 variables. All of the data is from 2020. For this analysis, I will not be using 'State' as a variable. I also won't be using 'CoalTrans' as a variable because it is 0 for all states. The data set I'm using is a combination of different data sets. Since there are only 50 rows of data, I will not do training & validation sets for classification models.

Data Sources

- 2020 Carbon Dioxide Emissions by State: https://www.eia.gov/state/rankings/#/series/226.
- EV Tax Credit: https://www.energysage.com/electric-vehicles/costs-and-benefits-evs/ev-tax-credits/
- State Energy Consumption Estimates (1960-2020): https://www.eia.gov/state/seds/sep_use/notes/use_print.pdf
- Transportation Sector Energy Consumption: https://www.eia.gov/state/seds/data.php?incfile=/state/seds/sep_sum/html/sum_btu_tra.html&sid=US
- US Census 2020 Population dataset: https://www.eia.gov/state/rankings/#/series/226

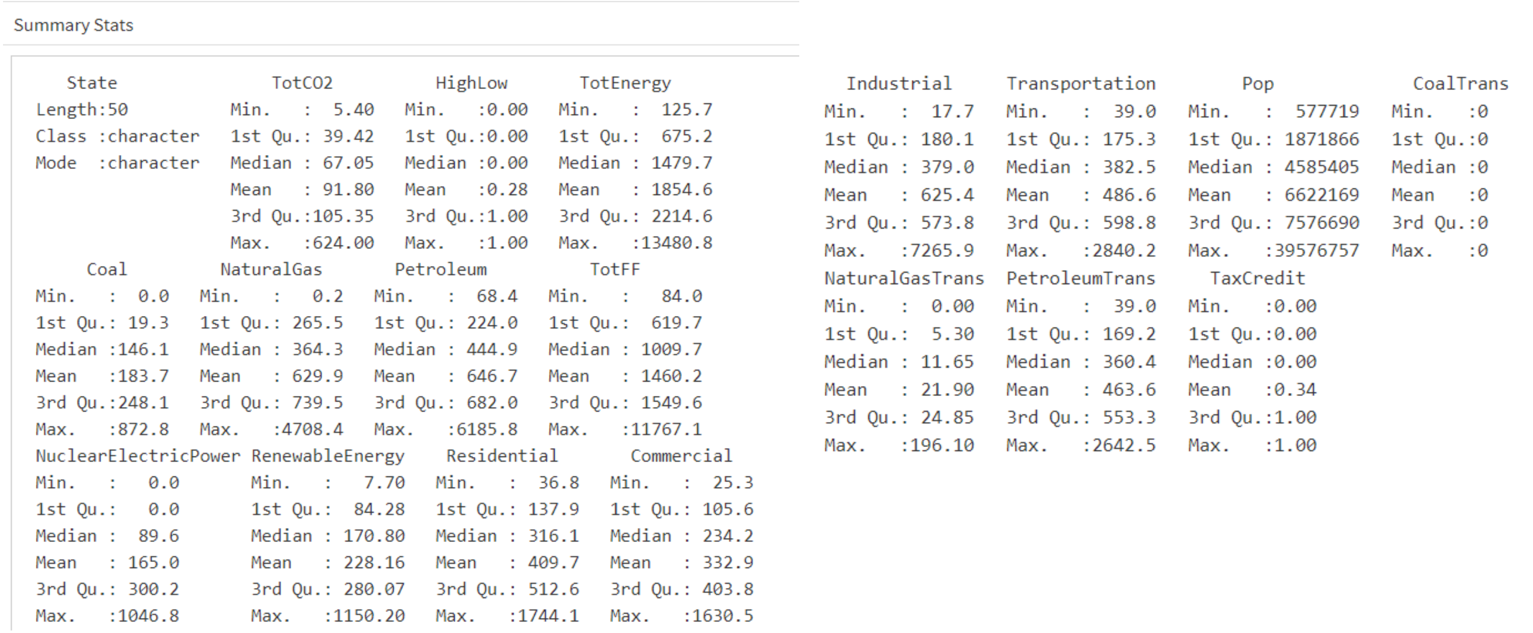
The Data

VARIABLES TO PREDICT WITH

- *TotEnergy*: total energy consumed (in trillion btu)
- *Coal*: energy consumed from coal (in trillion btu)
- *NaturalGas*: energy consumed from natural gas (in trillion btu)
- *Petroleum*: energy consumed from natural gas (in trillion btu)
- *TotFF*: total energy consumed from fossil fuels, sum of Coal, NatualGas, & Petroleum (in trillion btu)
- *NuclearElectricPower*: energy consumed from nuclear electric power
- *RenewableEnergy*: energy consumed from renewable energy sources (in trillion btu)
- *Residential*: energy consumed by the residential sector (in trillion btu)
- *Commercial*: energy consumed by the commercial sector (in trillion btu)
- *Transportation*: energy consumed by the transportation sector (in trillion btu)
- *Pop*: population of a state
- *CoalTrans*: energy from coal used for transportation (in trillion btu)
- *NaturalGasTrans*:energy from natural gas used for transportation (in trillion btu)
- *PetroleumTrans*: energy from petroleum used for transportation (in trillion btu)
- *TaxCredit*: EV tax credit dummy variable (= 1 if tax credit; 0 otherwise)

VARIABLES WE WANT TO PREDICT

- *TotCO2*: total CO2 emissions of a state
- *HighLow*: CO2 emissions > 100 coded as 1, lower coded as 0



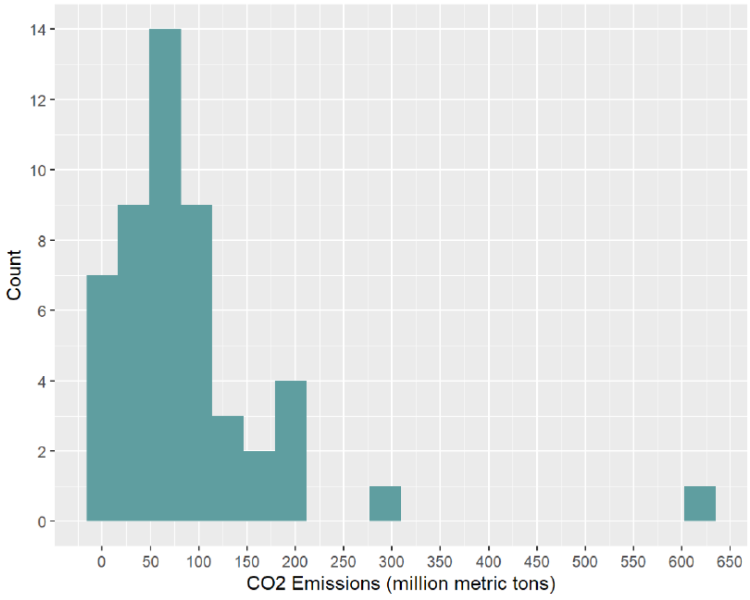
Summary Stats

```
    State               TotCO2          HighLow        TotEnergy      
 Length:50          Min.   :  5.40   Min.   :0.00   Min.   :  125.7  
 Class :character   1st Qu.: 39.42   1st Qu.:0.00   1st Qu.:  675.2  
 Mode  :character   Median : 67.05   Median :0.00   Median : 1479.7  
                    Mean   : 91.80   Mean   :0.28   Mean   : 1854.6  
                    3rd Qu.:105.35   3rd Qu.:1.00   3rd Qu.: 2214.6  
                    Max.   :624.00   Max.   :1.00   Max.   :13480.8  
      Coal         NaturalGas       Petroleum          TotFF        
 Min.   :  0.0   Min.   :   0.2   Min.   :  68.4   Min.   :   84.0  
 1st Qu.: 19.3   1st Qu.: 265.5   1st Qu.: 224.0   1st Qu.:  619.7  
 Median :146.1   Median : 364.3   Median : 444.9   Median : 1009.7  
 Mean   :183.7   Mean   : 629.9   Mean   : 646.7   Mean   : 1460.2  
 3rd Qu.:248.1   3rd Qu.: 739.5   3rd Qu.: 682.0   3rd Qu.: 1549.6  
 Max.   :872.8   Max.   :4708.4   Max.   :6185.8   Max.   :11767.1  
 NuclearElectricPower RenewableEnergy   Residential       Commercial    
 Min.   :   0.0       Min.   :   7.70   Min.   :  36.8   Min.   :  25.3  
 1st Qu.:   0.0       1st Qu.:  84.28   1st Qu.: 137.9   1st Qu.: 105.6  
 Median :  89.6       Median : 170.80   Median : 316.1   Median : 234.2  
 Mean   : 165.0       Mean   : 228.16   Mean   : 409.7   Mean   : 332.9  
 3rd Qu.: 300.2       3rd Qu.: 280.07   3rd Qu.: 512.6   3rd Qu.: 403.8  
 Max.   :1046.8       Max.   :1150.20   Max.   :1744.1   Max.   :1630.5  
```

```
   Industrial      Transportation       Pop             CoalTrans
 Min.   :  17.7   Min.   :  39.0   Min.   :  577719   Min.   :0  
 1st Qu.: 180.1   1st Qu.: 175.3   1st Qu.: 1871866   1st Qu.:0  
 Median : 379.0   Median : 382.5   Median : 4585405   Median :0  
 Mean   : 625.4   Mean   : 486.6   Mean   : 6622169   Mean   :0  
 3rd Qu.: 573.8   3rd Qu.: 598.8   3rd Qu.: 7576690   3rd Qu.:0  
 Max.   :7265.9   Max.   :2840.2   Max.   :39576757   Max.   :0  
 NaturalGasTrans  PetroleumTrans     TaxCredit   
 Min.   :  0.00   Min.   :  39.0   Min.   :0.00  
 1st Qu.:  5.30   1st Qu.: 169.2   1st Qu.:0.00  
 Median : 11.65   Median : 360.4   Median :0.00  
 Mean   : 21.90   Mean   : 463.6   Mean   :0.34  
 3rd Qu.: 24.85   3rd Qu.: 553.3   3rd Qu.:1.00  
 Max.   :196.10   Max.   :2642.5   Max.   :1.00  
```

Response Variables:

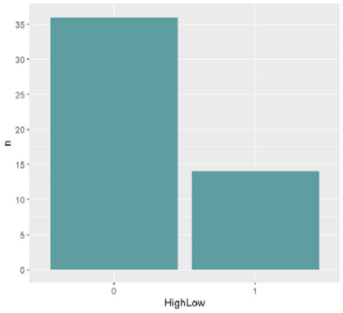
Total CO2 Emissions (in million metric tons)



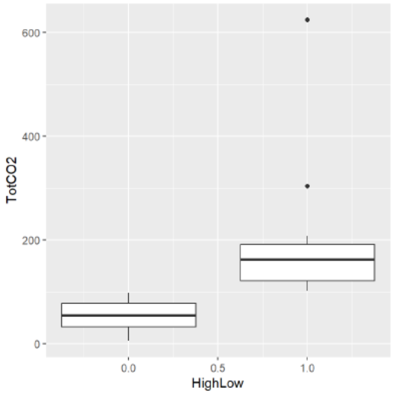
This is a histogram of the variable TotCO2 emissions which is the total CO2 emissions from a state. Most of the states fall between 0 & 200 million metric tons of carbon dioxide.

Response Variables: CO2 Emissions: High (1)/Low(0)

Bar Chart

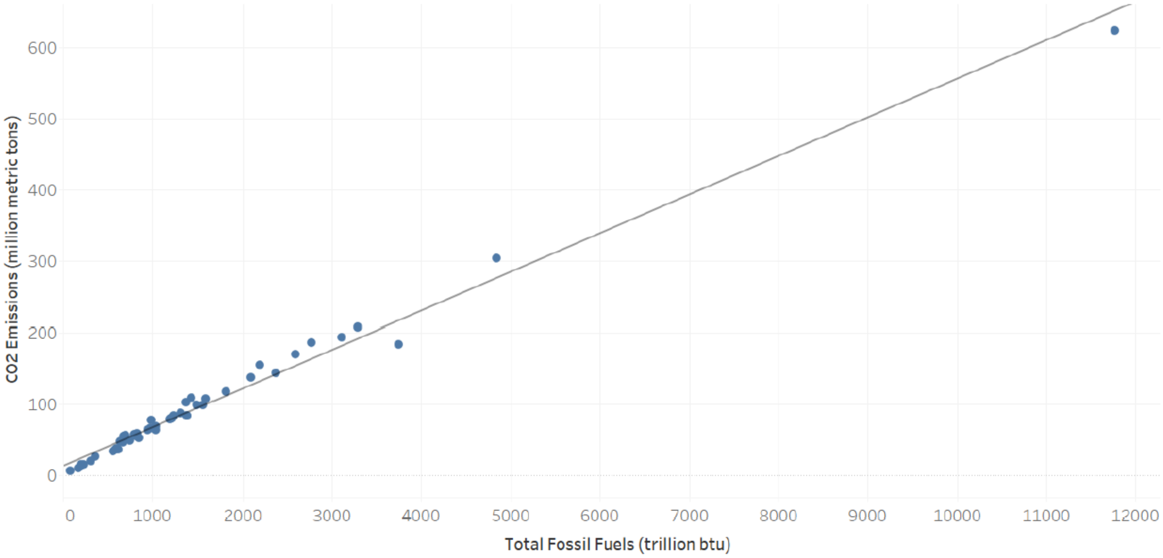


Box Plot



We can see that the majority of states are considered "Low Emission" states, meaning that they produce less than 100 million metric tons of carbon dioxide. We can also see that "High Emission" states have a larger range of values and has 2 outliers.

Total Fossil Fuels Consumed & Total CO2 Emissions



This is a scatter plot showing Total Fossil Fuels Consumed & Total Carbon Dioxide Emissions. Total Fossil Fuels Consumed is the sum of Coal, Natural Gas, & Petroleum consumed. Each point is a different state, but there seems to be a positive correlation between CO2 emissions & fossil fuels consumed.

Predict CO2 Emissions Models

I will be using prediction/estimation molding techniques to predict the amount of CO2 emissions for a state. The first technique I will use is a Multiple Linear Regression. Then I will run a Decision Tree Model and compare them.

Predict CO2 Emissions Model 1 (M1)

For this model my predictors were: Coal, NaturalGas, Petroleum, Transportation, Pop, & TaxCredit

99.98 %

Adjusted R-Squared

1.47

RMSE

Regression Output

	Estimate	Std. Error	t value	Pr(>|t|)
Coal	0.097	0.001	72.110	0.000
NaturalGas	0.052	0.001	37.045	0.000
Petroleum	0.028	0.001	18.949	0.000
PetroleumTrans	0.032	0.004	7.939	0.000
Pop	0.000	0.000	4.682	0.000
NaturalGasTrans	0.044	0.016	2.755	0.009
(Intercept)	1.143	0.452	2.527	0.016
RenewableEnergy	0.004	0.001	2.347	0.024
NuclearElectricPower	0.001	0.001	0.661	0.512
TaxCredit	-0.375	0.572	-0.655	0.516

Analysis Summary

After examining this model, there are some predictors that are not important in predicting CO2, so a pruned version of the model is created by removing predictors that are not significant.

Predict CO2 Emissions Model 2 (M2)

For this analysis we will use a pruned Multiple Linear Regression Model. I also removed predictors that had I didn’t think were important in predicting CO2 emissions. The 3 predictors involving transportation were significant (based on their p-values) so I decided to use the Transportation variable as a predictor because it’s the sum of CoalTrans, NaturalGasTrans, & PetroleumTrans. These are the predictors in the final model: Coal, NaturalGas, Petroleum, & Transportation.

99.96 %

Adjusted R-Squared

1.84

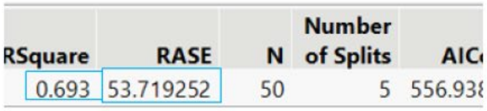
RMSE

Regression Output

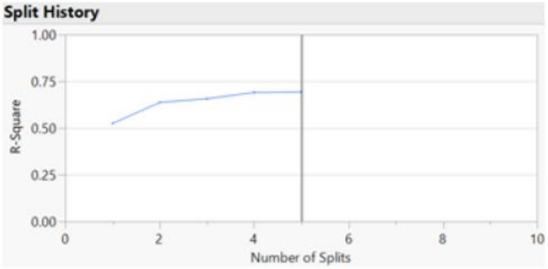
	Estimate	Std. Error	t value	Pr(>|t|)
Coal	0.094	0.001	63.235	0.000
NaturalGas	0.053	0.001	43.755	0.000
Transportation	0.050	0.001	35.613	0.000
Petroleum	0.024	0.001	21.954	0.000
(Intercept)	1.005	0.429	2.344	0.024

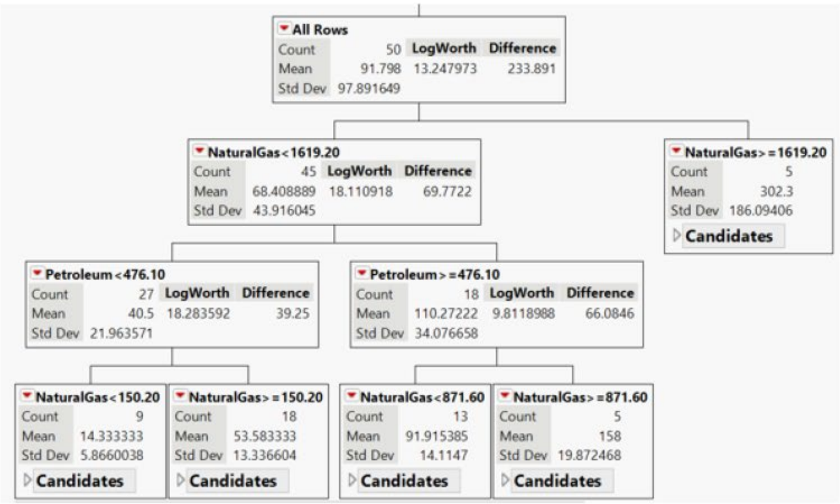
Decision Tree Model (M3)

The second prediction/estimation technique I will use to predict CO2 Emissions is a decision tree model. For this model my predictors were: Coal, NaturalGas, Petroleum, Transportation, Pop, & TaxCredit.



RSquare	RASE	N	Number of Splits	AIC
0.693	53.719252	50	5	556.93





After evaluating different decision tree models with different numbers of splits, the final model has 4 splits. It has the highest R-square value before it plateaus (as shown in the output on the left).

Model Comparison:

Out of the 3 models I created to predict the amount of CO2 emissions for a state, Model 2, a multiple linear regression, was the best. M2 was the better multiple regression model because it was simpler than M1. M2 used less predictors and R-square value didn't decrease very much from M1. M2 was better than M3 because the R-square value is much higher and can predict much better than M3.

Model	R-Square
Multiple Linear Regression 1 (M1)	0.9998
Multiple Linear Regression 2 (M2)	0.9996
Decision Tree (M3)	0.693