Theodore DeFuria
9 February 2026
EPA Emissions Data Analysis

I ran a random forest regressor on 5 predictors and 1 target variable. The data accounts for state and geographic location, which has a risk of multicollinearity to impact model performance. Both variables provide predictive utility to the model. Feature importances served as a starting point for further analysis. Unsupervised algorithms such as DBSCAN clustering was performed on the facilities in each state/territory. Their location relative to the largest population center in each locality, the facility count in each cluster, and the sum of emissions per cluster are displayed for exploration.

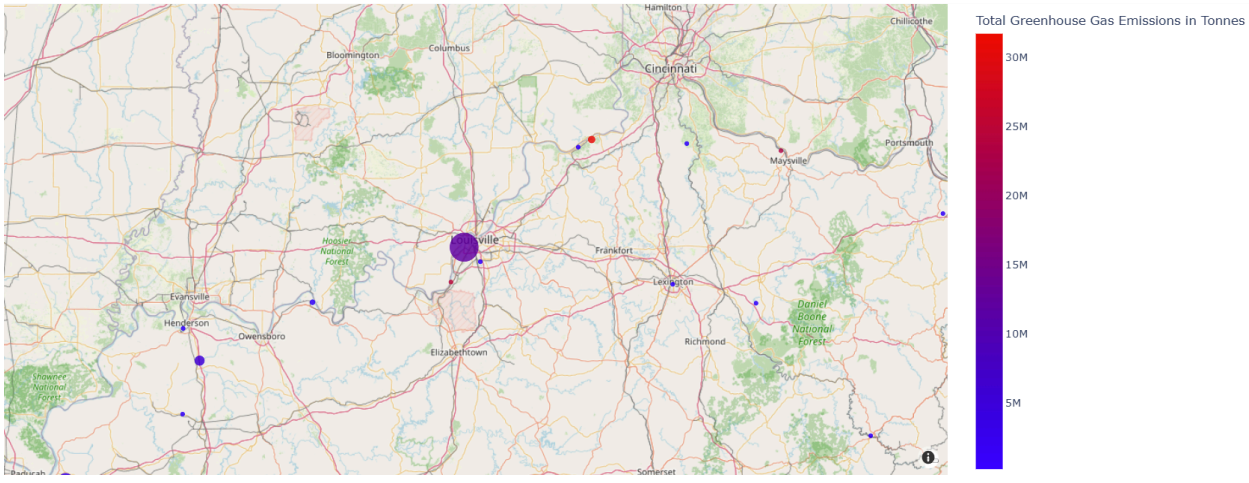
Each cluster represents a collection of industrial facilities in that state. The clusters are counted and display the number of facilities they represent as their size. The color scale corresponds to the sum of greenhouse gas emissions in tonnes. The scale spans from the highest emissions cluster in the state to the lowest, with red being the highest emissions cluster and blue being the lowest emissions cluster.

Power plants serve as the greatest industry sector predictor. Power plants represent 16% feature importance vs. 19% total across all industries. Therefore, knowing a facility's status as a power plant is five times more important to predicting the emissions output than knowing it is a part of any of the other 10 industry categories.

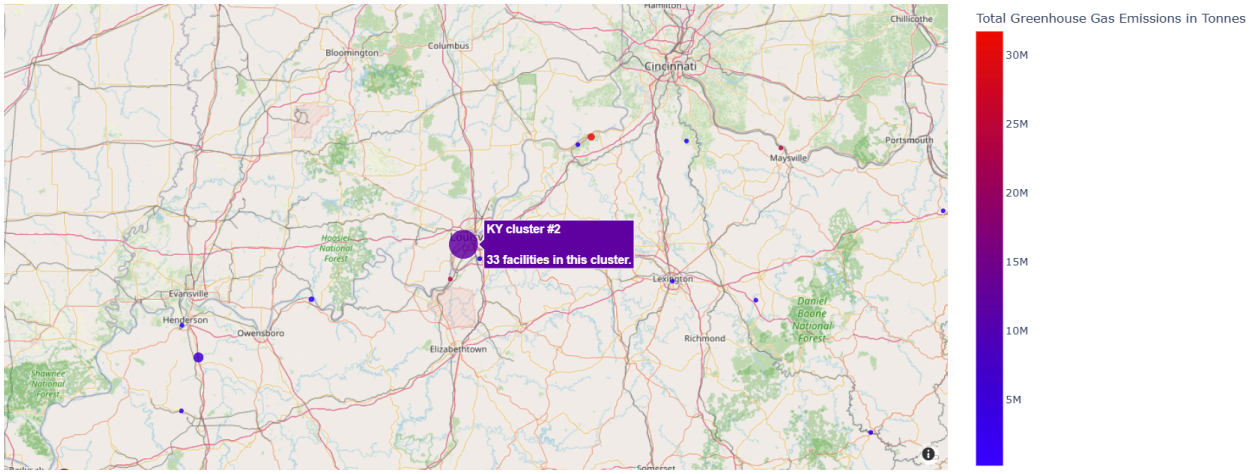
The distance of all states/territories largest emissions clusters to their respective major population center is a mean of 94.2 mi, which is likely to pass buffer zone regulations. In accordance with developed countries research on pollution and relatively high standard of living, these largest emissions facilities are located far away from urban centers, possibly to prioritize mitigating public health concerns.

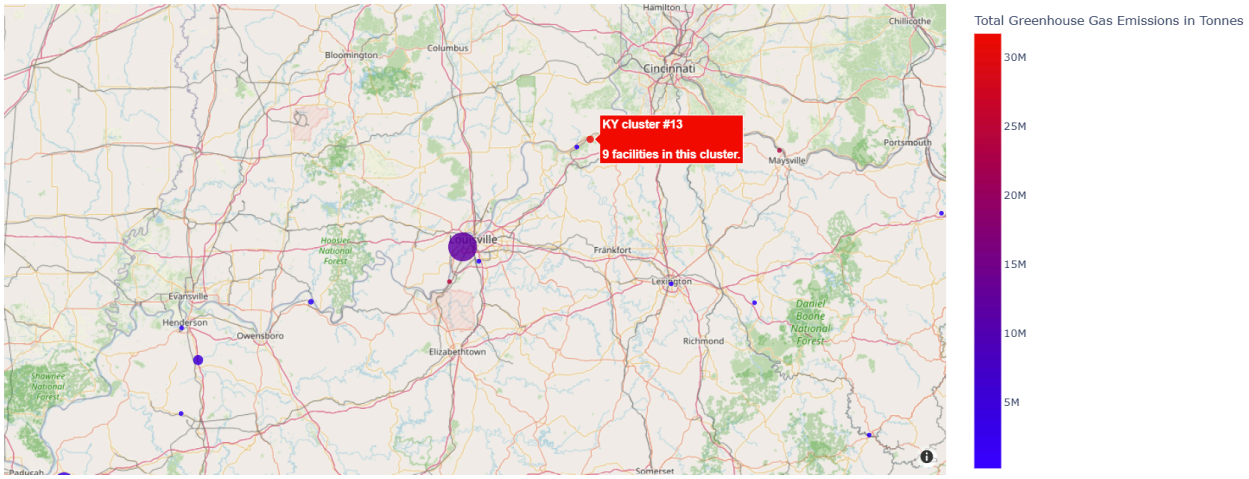
WV and KY indicated as top 2 states influencing emissions prediction. WV's mean emissions is 5th overall while KY is 9th, which implies presence represents positive correlation with GHG emissions in tonnes.

Along with Wyoming and Indiana, which make up the top 4 most impactful states on emissions prediction, these states maintain some of the highest reliance on coal to generate power.



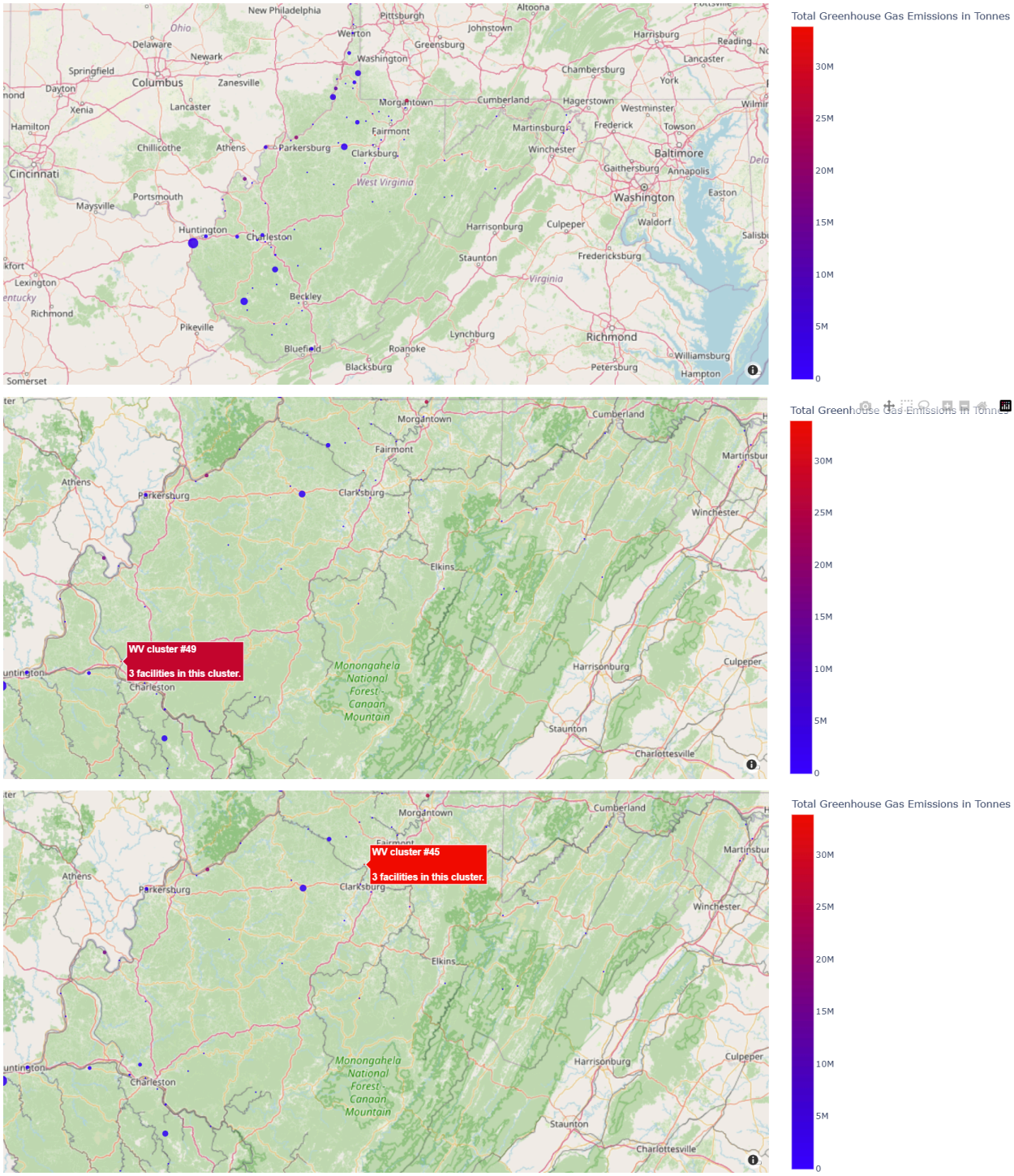
Visualizations produced with Plotly.js.





Ghent Power Plant, powered by coal, accounts for the largest emissions in Kentucky. It is 52.75 miles from the population center in Louisville, KY.

Visualizations produced with Plotly.js.



Harrison Power Station, powered by coal, accounts for the largest emissions in West Virginia. It is 100.43 miles from the population center in Charleston, WV.

Visualizations produced with Plotly.js.