Midpoint Ideas

**Data Source :**

[**https://www.epa.gov/ghgreporting/data-sets**](https://www.epa.gov/ghgreporting/data-sets)

* **EDA’s IDEAS**
  + Possible ideas - trending over time
  + By State
  + Electric quotient use
  + Look at facility ID’s.
  + Predicting Facility Emission Levels
  + Prediction of Emissions levels with new regulations being passed.
  + Forecasting emission is given by facility region or section
  + Industry type
  + Times series analysis by state to identify outliers for regulators by industry.

**he deliverable for the midpoint include the following:**

* **Executive Summary**
* **Problem Statement/Research objective(s)**
* **Exploratory Data Analysis**
* **Data Preparation/Feature Engineering**

Final project- draft

Practical Machine Learning Final Assignment - Midpoint

## 

## 

## **Executive Summary**

Our project targets to predict the annual direct point greenhouse gas (GHG) emissions from industrial facilities before they are officially reported. This prediction model will serve three primary business purposes: understanding industry trends, informing policy makers, and validating reported emissions. We will utilize EPA direct point GHG emissions data from industrial facilities spanning from 2011 to 2023 to build our machine learning model. Success will be measured by the accuracy of our emissions predictions, evaluated through cross-validation of test-train datasets.

**Group Members**

* David Van Dyke
* Ethan Norton
* Migus Wong
* Michael Yoo

## **Business Objective**

The goal is to predict the annual direct point GHG emissions from industrial facilities before they are officially reported. This prediction model will serve the following business purposes:

* Understand Industry Trends: Identify future changes in emissions levels to guide potential investment strategies.
* Inform Policy Makers: Deliver accurate, timely predictions that enable regulators to shape future environmental policies.
* Resource Optimization and Cost Reduction: Flag facilities that deviate from industry norms so that investments and regulatory focus can be directed to those areas, reducing unnecessary costs
* Validate and Forecast Emissions: Serve as an independent benchmark to assess the accuracy and reasonableness of reported data, thereby informing risk assessments and ESG evaluations.

**PROBLEM STATEMENT**

Industrial facilities must report their GHG emissions if they meet or exceed a threshold (typically ≥25,000 metric tons CO₂ per year). However, relying solely on end‑of‑year reported data can delay critical interventions. More importantly, there is a lack of regulation overall within the energy sector. Regulation is primarily focused on the emission of harmful emissions, but not all of them. Our task is to build predictive models that estimate facility emissions in advance, using historical data and engineered features. This will empower stakeholders—including investors and policymakers—to act proactively rather than reactively.

**RESEARCH OBJECTIVES**

Primarily, our goal is to develop a machine-learning pipeline that automates data ingestion, cleaning, feature engineering, and model training, as well as analyze trends and industry differences to detect anomalies and forecast future emissions. Ultimately, we want to provide actionable insights through comprehensive visualizations and analysis for policymakers and investors.

Success will be measured by predictive accuracy, validation, and compliance detection. R2 scores will assess how well the model captures the variability of emissions and RMSE to quantity prediction error. Validation will be evaluated by K-fold cross-validation to ensure consistency in model performance using different subsets of data. Lastly, compliance detection will be measured through precision and recall. High recall is key in catching potential non-compliant facilities, while high precision mitigates false alarms.

We will leverage annual greenhouse gas [(GHG) emissions](https://www.epa.gov/system/files/other-files/2024-10/2023_data_summary_spreadsheets.zip) data from industrial facilities—collected from 2011 through 2023—to build our machine-learning models. This dataset originates from the [EPA’s Greenhouse Gas Reporting Program (GHGRP)](https://www.epa.gov/ghgreporting/what-ghgrp), which requires facilities and suppliers to report their emissions if they exceed defined thresholds—typically 25,000 metric tons of CO₂ per year for most industries. Since 2010, the EPA has published annual spreadsheets summarizing these reported emissions, using a consistent Facility ID as the primary key to track each facility's data over time. We analyzed the multi-year summary of [EPA GHG emissions](https://ccdsupport.com/confluence/pages/viewpage.action?pageId=93290546) from direct point emitters (measured in metric tons). Our data sources include:

* Multi-Year Data Summary: High-level information for facilities over multiple years.
* Yearly Spreadsheets: Detailed yearly information, including reported emissions by greenhouse gas and process.

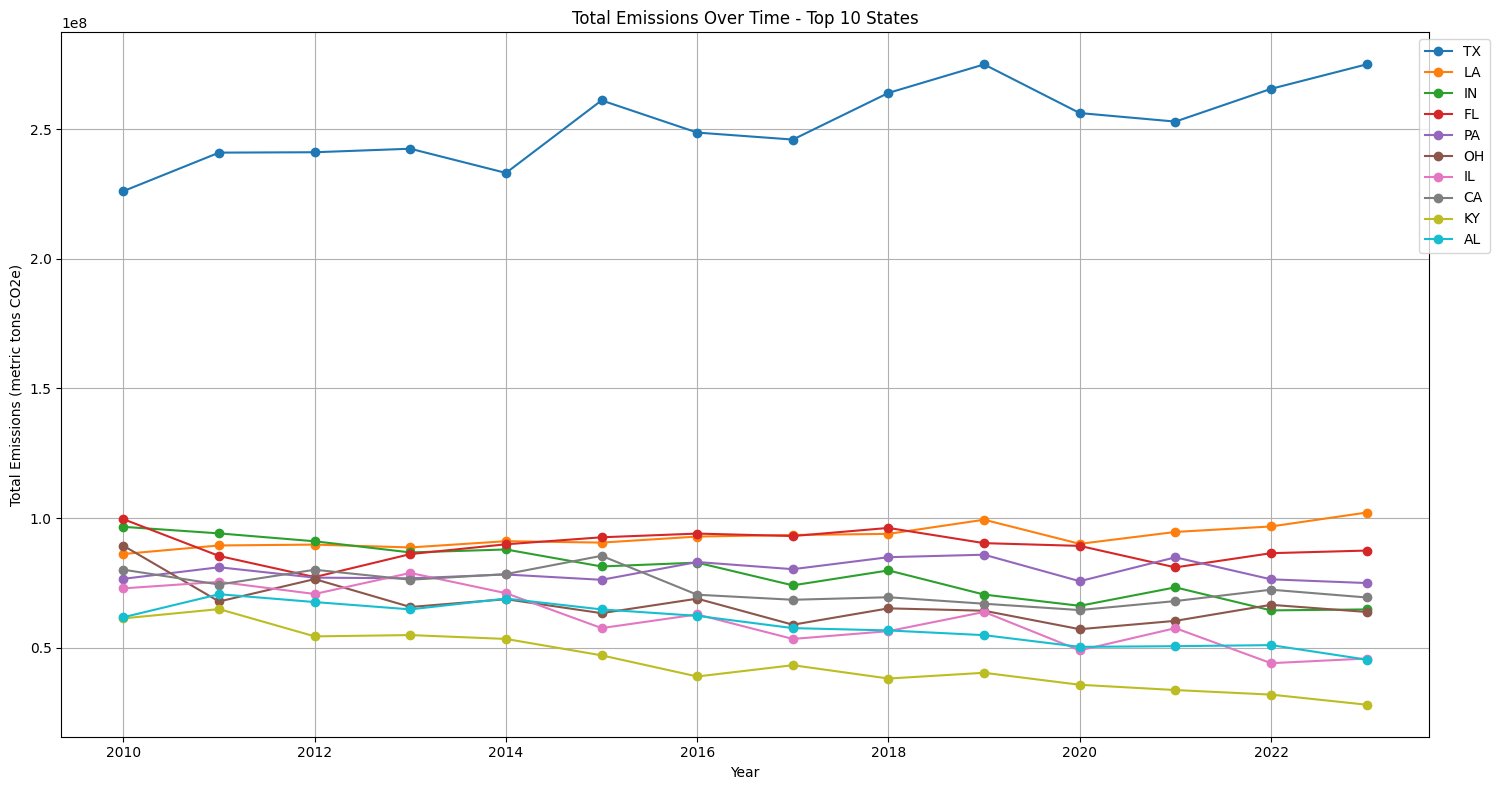
## **EDA/DATA PREPERATION/FEATURE ENGINEERING/**

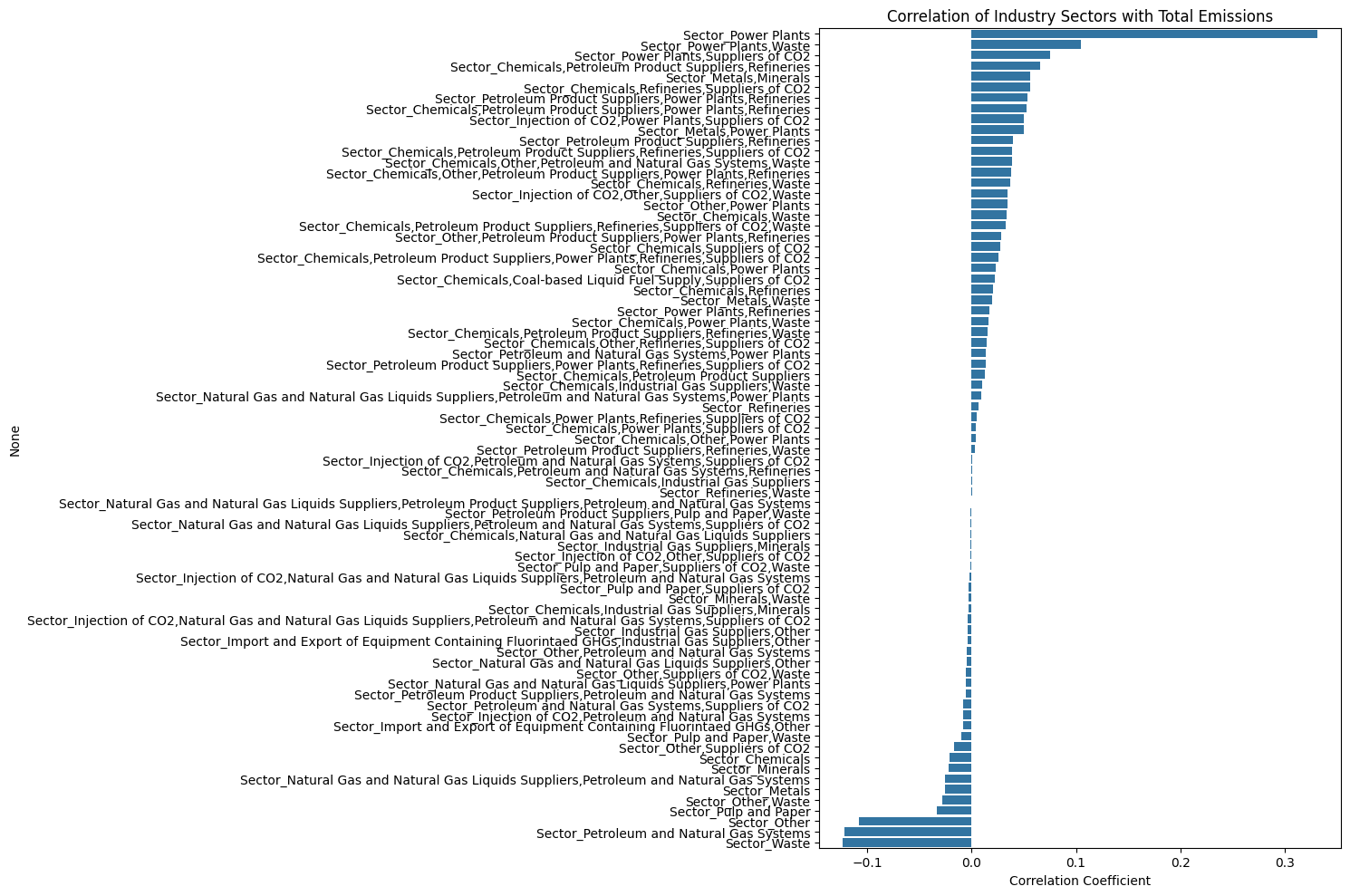
We will leverage annual greenhouse gas [(GHG) emissions](https://www.epa.gov/system/files/other-files/2024-10/2023_data_summary_spreadsheets.zip) data from industrial facilities—collected from 2011 through 2023—to build our machine-learning models. This dataset originates from the [EPA’s Greenhouse Gas Reporting Program (GHGRP)](https://www.epa.gov/ghgreporting/what-ghgrp), which requires facilities and suppliers to report their emissions if they exceed defined thresholds—typically 25,000 metric tons of CO₂ per year for most industries. Since 2010, the EPA has published annual spreadsheets summarizing these reported emissions, using a consistent Facility ID as the primary key to track each facility's data over time. We evaluated the multi-year summary of [EPA GHG emissions](https://ccdsupport.com/confluence/pages/viewpage.action?pageId=93290546) from direct point emitters (measured in metric tons) for our analysis. Our data sources include:

* Multi-Year Data Summary: High-level information for facilities over multiple years.
* Yearly Spreadsheets: Detailed yearly information, including reported emissions by greenhouse gas and process.

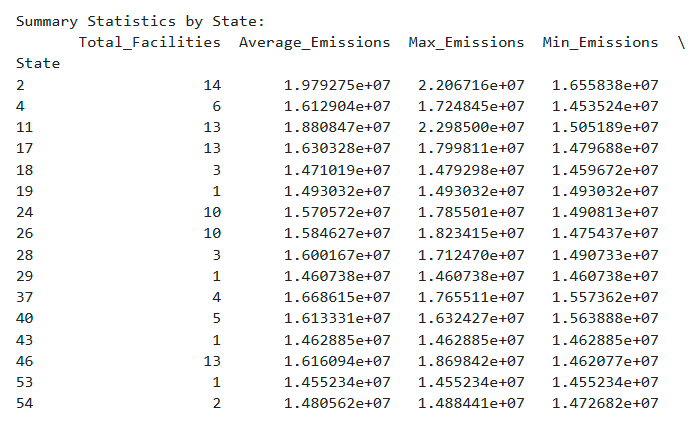
## **EDA/FEATURE ENGINEERING**

Below are some key highlights of our EDA. Please reference our Python notebook for the comprehensive details. We began by visualizing the top 10 states by total emissions over time to help us understand which states should be flagged for further review.

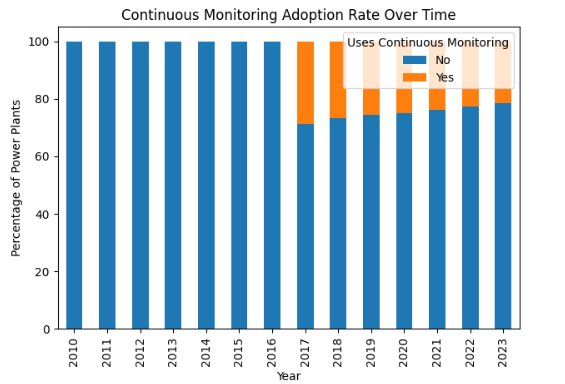


In addition, we ranked the correlation between industries using the encoded primary NAICS code and industry type. This ranking supports management and regulators by highlighting the most significant sources of emissions, thereby helping prioritize initiatives by sector.

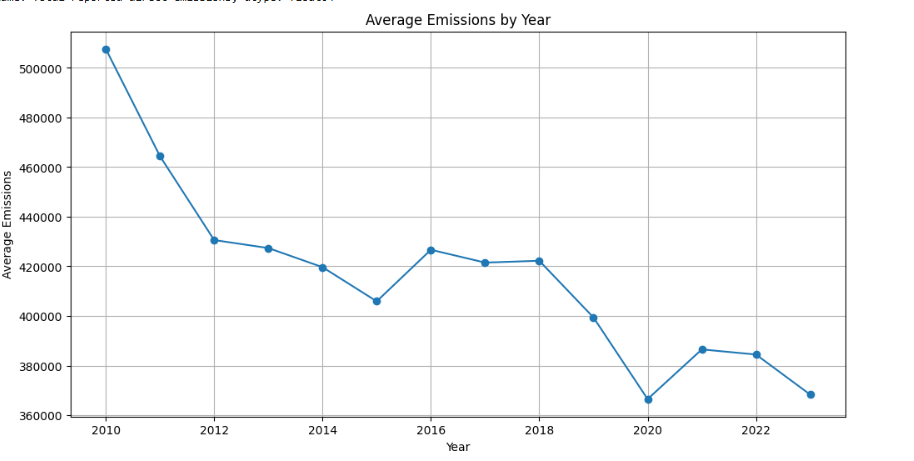
We also examined the mean emissions by state to create potential benchmarking metrics.

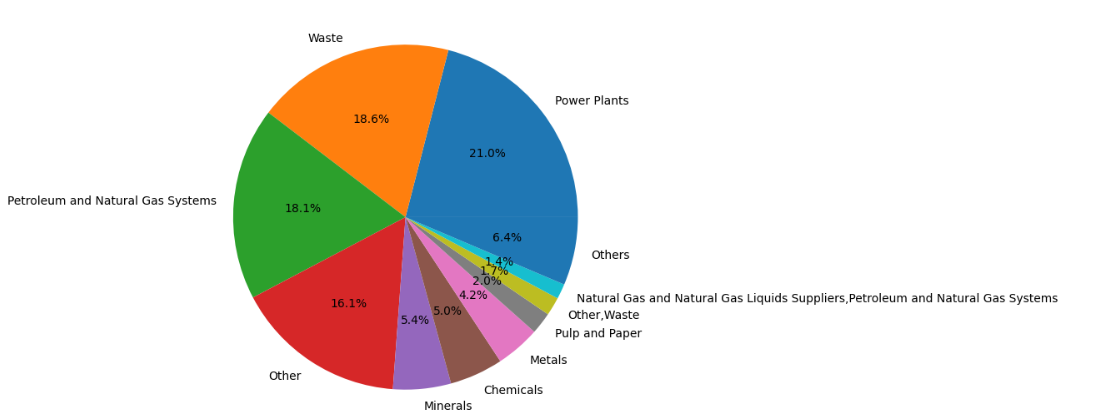


Moreover, we performed one‑hot coding for the CO₂‑related survey questions regarding continuous monitoring—including whether CO₂ is collected on‑site. By comparing facilities that employ continuous monitoring against those that do not, we observed an interesting trend: the adoption of continuous monitoring appears to have begun around 2017, and this timing may correlate with a decrease in overall year‑over‑year emissions—an observation where continuous monitoring may have been a factor in the reduction of overall emissions YoY.



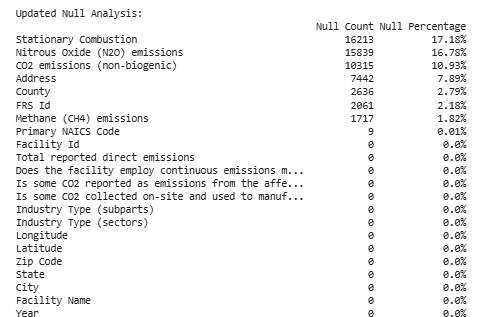
An observation where continuous monitoring may have been a factor in the decrease in overall emissions YoY.



Additionally, our analysis revealed that although power plants dominate the top 20 emission sites among the top 10 plants, they represent only about 21% of the total facility population.

**FEATURE ENGINEERING**

In preparing our data, we first evaluated the null values, examined the raw dataset, and assessed the completeness of each variable. We set a threshold whereby any column with more than 70% null values (i.e., less than 30% available data) was discarded, ensuring that only “usable” variables remained.



* One-hot encoding - was applied to categorical fields to prepare categorical data for learning models. The State and NAICS codes were one‑hot coded to observe geographical differences as well as further clean our data.
* Binary one-hot encoding - used transform survey questions regarding CO₂ collection and continuous monitoring, converting responses to 1 for “Yes” and 0 for “No.”
* Feature engineering for Year - “Year” variable was used as a predictor and to aggregate data for time‑series visualizations
* Location Aggregation - Location information, such as state, city, and zip code, was aggregated, with some “State” fields one‑hot coded.
* Composite and Derived Variable Creation - created by generating a new numerical variable for total emissions, allowing us to rank facilities nationally and by state.
* Industry Subpart Splitting - commas split of the ‘Industry Type” field to generate binary indicator columns for each unique subpart, providing richer variables to improve model accuracy and support more precise recommendations.
* Feature Scaling - used StandardScaler to ensure that features with different scales did not disproportionately influence the model.
* Encoding combination + Continuous Features - creating sparse matrix when applicable—to create a final feature matrix for modeling.

Overall, the feature engineering process eliminated columns with high missing rates and identified key variables to build a solid foundation for our models. Encoding states and NAICS codes allowed us to explore deeper insights and test datasets interchangeably. Our early hypothesis—that continuous monitoring may reduce emissions—was supported by our preliminary findings of the binary conversion of monitoring questions, potentially facilitating technology adoption in certain facilities to adopt continuous monitoring. Most importantly, scaling, encoding, and creating composite features have prepared the data for success when applied in the models we create.

**Initial model testing**

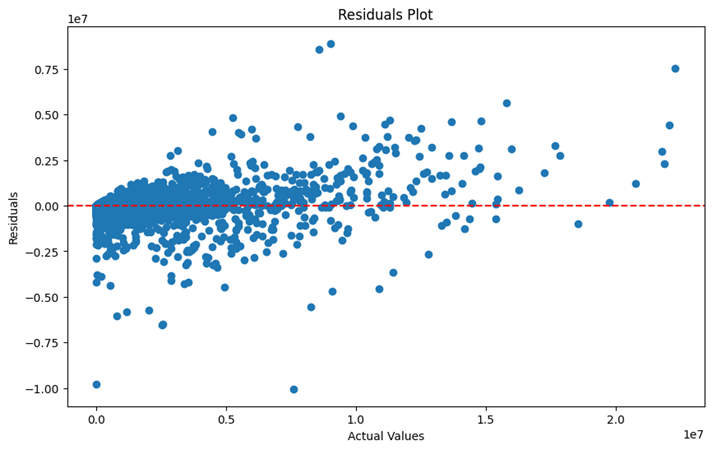
For starting model testing and development, we found the tree based models to be most successful. We believe this makes sense based on the type of data in the EPA spreadsheets. The data source does not provide much detail about the specific configuration of the industrial plants. There is no information on the number of heaters, types burners, types of heaters, type of fuel, amount of fuel consumed, product production levels, ect. These types of details are needed to make linear-based models successful.

The data source mainly consists of information about the industry category, the plant location and the emission year. Often, multiple pieces of information cover these categories. (**location-** city, state, zip, latitude, longitude) (**industry-** NAICS code, Industry Type (subparts), Industry Type (sectors))

For the initial model development, we checked only using the latitude and longitude for location and separating these industry types into one hot encoded features. The tree based models were able to give some reasonable prediction results. Linear and SVM models did not look reasonable on the early testing.

Additional tuning on the model type, engineered features, and hyperparameters is required. We’ll continue this in the second phase of the model testing.

**Residuals decision trees**

****

**Residuals linear model**

