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 policymakers, and validating reported emissions. We will use EPA direct point GHG emissions data from industrial facilities 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 quantify 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 PREPARATION > 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:

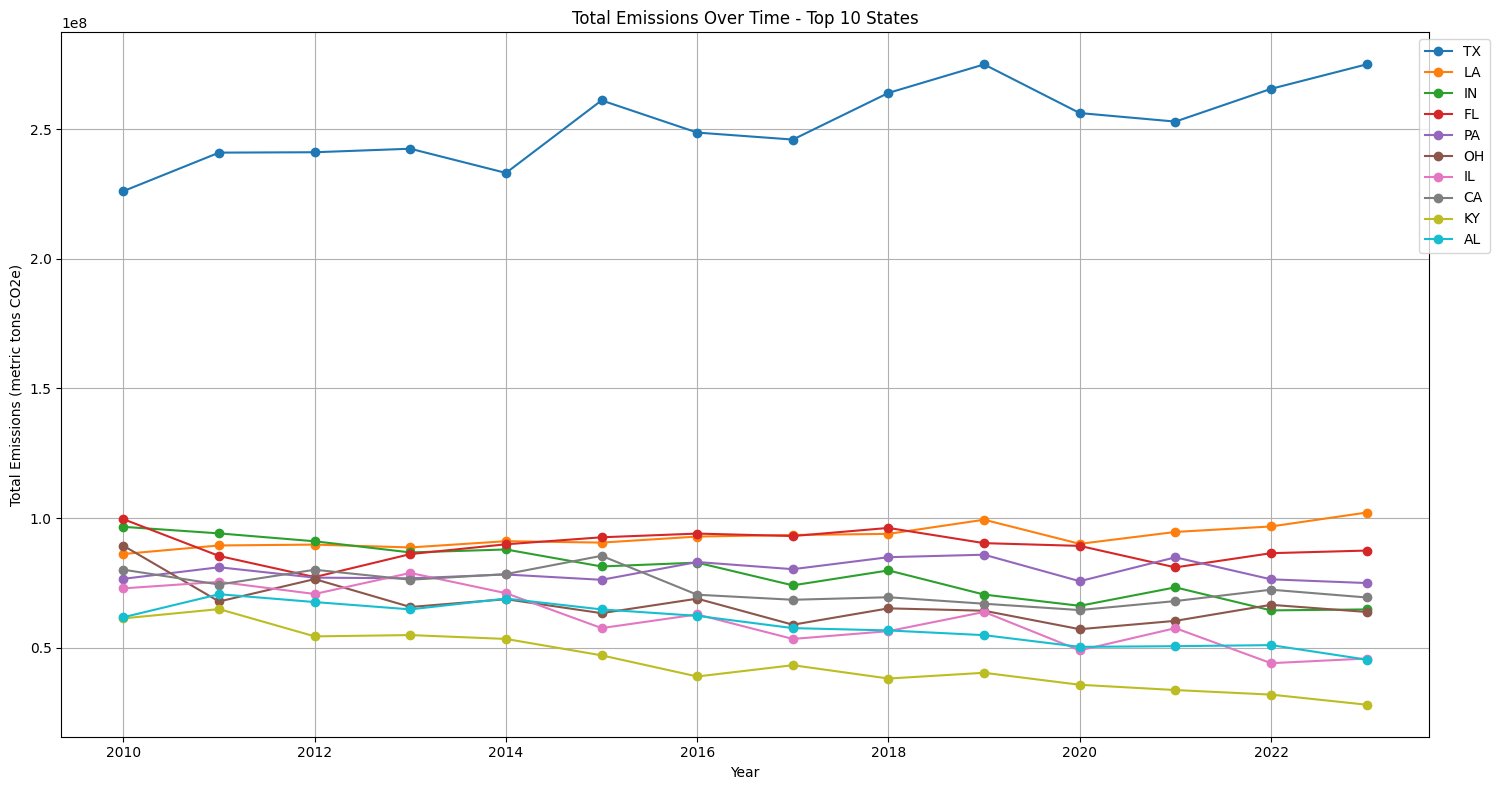
* Multi-Year Data Summary: High-level information for facilities over multiple years.
* Yearly Spreadsheets: Detailed yearly information, including reported emissions by greenhouse gas and process.

## EDA/FEATURE ENGINEERING

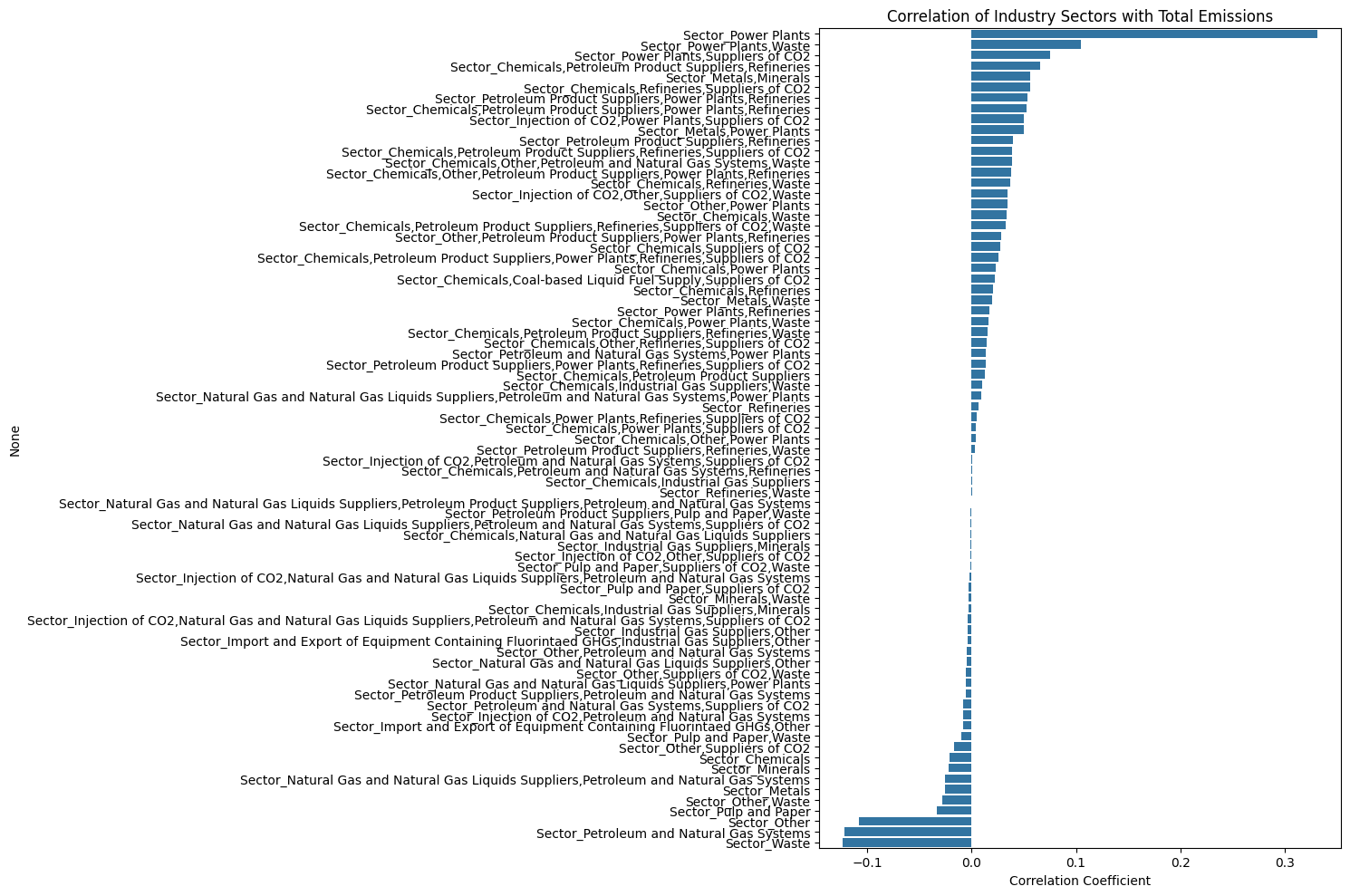
Below is a table summary of the team's EDA and data preparation. Please reference our Python notebook if you’d like more details. However, we also highlight some key findings that caught our attention below.

| **EDA** | **TASK** | **VISUALS** |
| --- | --- | --- |
| Missing Data Analysis | •Checked Null values and dropped columns w/ >70%  •Created table.summary logs for missing counts | Python print statement |
| Yearly Emissions Stats | • Calculated facilities count per/yr  • Mean + total emissions  • Observed trends | Table/Summary Logs |
| State Level Analysis | • Grouped data by state to compare total & mean emissions  • Compared facility counts and emission | Line plots |
| Outlier Detection | • Identified top 100  • Filtered data for deeper review | Box/Violin plot |
| Top 100 Emitters | • Ranked Facilities by total emissions  • Summarized by state and industry sector | Bar Charts  Heatmaps |
| Monitoring Status Exploration | • Compared emissions for continuous monitoring  • Conducted t-tests for emission differences | Box and Violin Plots  t-test summaries |
| Industry/Sector Analysis | • Examined Industry Type (sector) for top contributors  • Created dummy variables for correlation check | Bar Charts  Correlation Heatmaps |
| Modeling (Random Forrest  & Decision Tree) | • Encoded categorial columns  • Trained ML models to predict emissions  • Checked feature importance | Feature Importance Bar Plot  Actual vs Predicted Scatter |

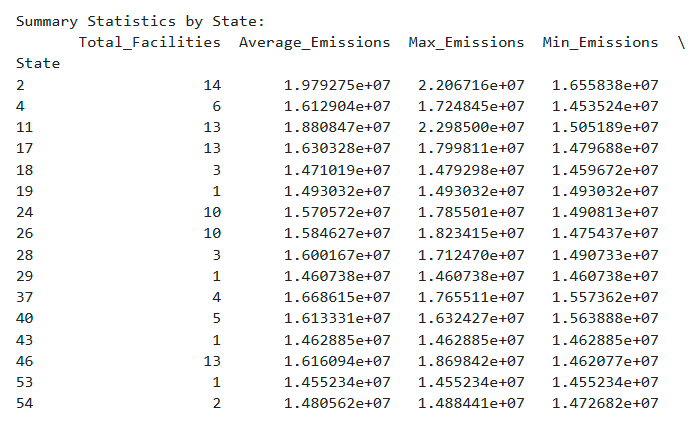
Looking at the top 10 states by total emissions over time helped us understand which states should be flagged for further review.



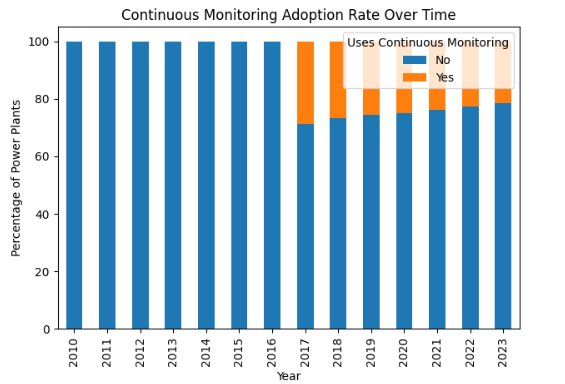
We also used advanced visual tools like folium to understand the severity of emissions across the US.

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.

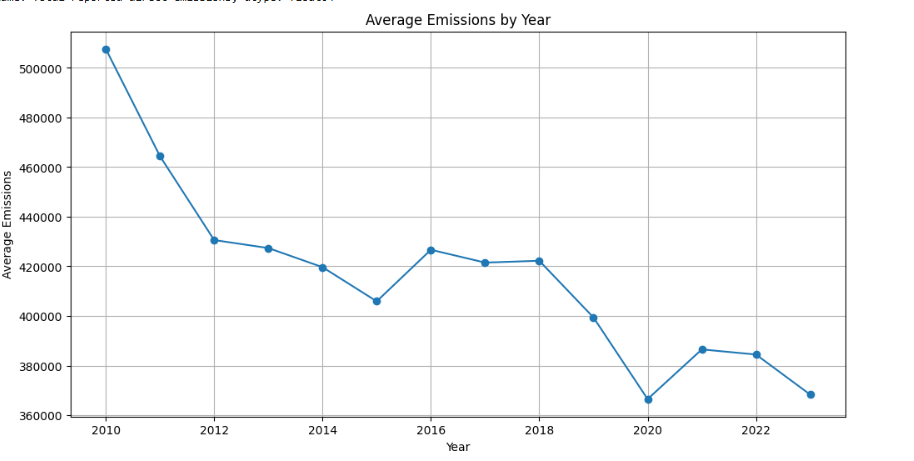
We also examined the mean emissions by state to create potential benchmarking metrics.

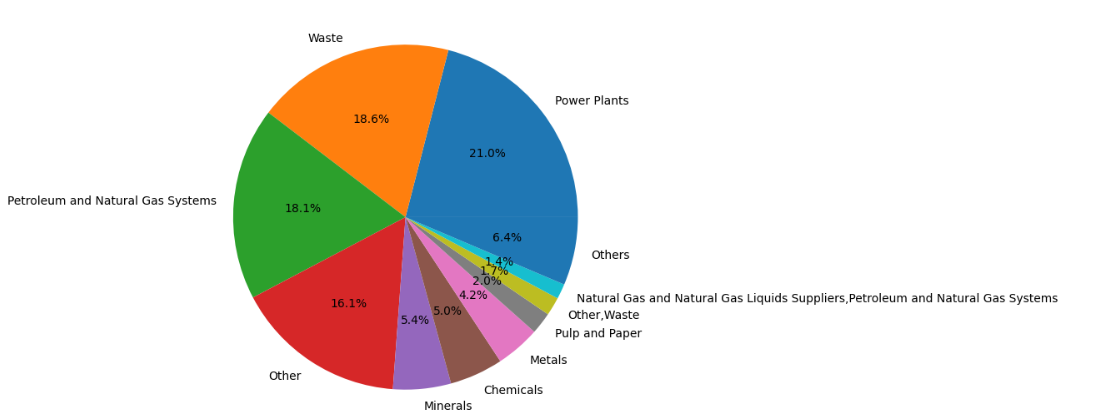


Moreover, we performed one‑hot coding for the CO₂‑related survey questions regarding continuous monitoring—including whether CO₂ is collected on‑site. By comparing facilities that employ continuous monitoring against those that do not, we observed an interesting trend: the adoption of continuous monitoring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



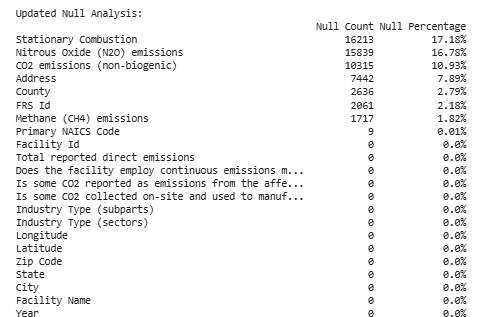
An observation where continuous monitoring may have been a factor in the decrease in overall emissions YoY.



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.



Below is a summary of the feature engineering/data preparation the group executed. There may be some carry over from the EDA summary we shared earlier, as they are interchangeable.

* 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 and 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/Normalization - 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.
* Outlier Filtering - removed extreme outliers or separated top emitters into a separate set for further analysis.

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 to our model.