Final Project Writeup

Abstract

This study investigates the relationship between industrial facilities and greenhouse gas emissions, population density, and air quality in various US counties. Two models were constructed to address the research questions, which aimed to identify the facilities primarily linked to greenhouse gas emissions and predict air pollution levels in different counties. The findings demonstrate the impact of industrial activities and population density on air pollution. According to the findings, it is evident that both industrial establishments and human activities play a significant role in affecting the air quality. In addition, the emission of greenhouse gases also contributes to the deterioration of air quality. Establishment of this model provides valuable insights for policymakers to address air pollution caused by industrial activities.

Introduction

As the manufacturing industry experiences rapid growth, the detrimental consequences of industrial development have become increasingly evident. One notable impact is the degradation of air quality in surrounding areas, resulting from the emissions produced during manufacturing processes. This has established industrial activity as a main contributor to local air pollution. Consequently, it is crucial to examine the correlation between air quality levels in various counties and the distribution of facilities associated with greenhouse gas emissions.

Our investigation focuses on two primary research questions:

1. Which facilities are predominantly linked to greenhouse gas emissions (CO2, CH4, N2O, BIOCCO2, PFC, etc.), thereby posing a significant threat to the climate?
2. Is it possible to utilize information on these facilities to predict air pollution levels in different counties?

To address these questions, we constructed two models, utilizing available datasets encompassing air quality data, facilities data, and emissions data. Additionally, we hypothesized that population could influence air quality. Consequently, we incorporated an additional dataset to investigate the relationship between population density and air quality.

Literature Review

A multitude of scholarly investigations have recently expounded upon the correlation between industrial emissions and air quality. In 2019, Zhang et al. [1] disclosed that industrial operations were the principal drivers of air pollution in China, emphasizing the menace to public health engendered by industrial emissions. In 2020, Velders et al. [2] substantiated that diminishing gas emissions could appreciably reduce the yearly average PM2.5 concentration in the Netherlands. This decrease in emissions was predominantly ascribed to domestic and international industry, corroborating the idea that industrial activities can exert a significant influence on the environment and air quality.

Scholars have additionally observed that industrial endeavors can discharge various categories of greenhouse gases, encompassing nitrogen oxides (NOx) emanating from industrial boilers and power plants. This introduces the prospect that greenhouse gas emissions from industrial installations may contribute to both climate change and alterations in air quality [3]. Multiple research teams have undertaken exhaustive inquiries into how diverse kinds of gases affect air quality. For instance, the Tsai group scrutinized the air quality index trend based on greenhouse gas emissions in Taiwan [4]. Furthermore, certain studies have employed distinct greenhouse gases, such as NO, NO2, CO, and HONO, as markers of air quality [5], implying that greenhouse gases can be crucial determinants when forecasting local air quality.

It is crucial to note that human activities, including operating vehicles, combusting fuels, and manufacturing processes, have also been recognized as sources of air pollution, suggesting that the local populace can be another pivotal factor contributing to local air quality [6]. Nevertheless, the majority of studies thus far have solely furnished statistical analyses of air quality trends based on several determinants or indicated correlations with industrial activities. To our knowledge, no quantitative model has been devised to signify or predict air quality primarily by considering local industrial facilities, greenhouse gas emissions, and human population figures. Moreover, greenhouse gas emissions have been employed to denote the distribution of facilities that may impact the local environment. Consequently, we have devised models to uncover the potential association between these factors.

Datasets Introduction

Provided Datasets

- `us_greenhouse_gas_emissions_direct_emitter_facilities.csv` and `us_greenhouse_gas_emission_direct_emitter_gas_type.csv` contain data reported by EPA (Environment Protection Agency) on greenhouse gas emissions, detailing the specific types of gas reported by facilities and general information about the facilities themselves. The dataset is made available through EPA's GHGRP (Greenhouse Gas Reporting Program).
- `us_air_quality_measures.csv` contains data from the EPA's AQS (Air Quality System) that measures air quality on a county level from approximately 4000 monitoring stations around the country.

Additional Dataset

- `PopulationEstimates.csv` contains population estimates and Federal Information Processing Standards (FIPS) Code for the U.S. counties in 2010.

Data Cleaning and Wrangling

We performed data cleaning and wrangling on the DataFrames.

1. **Cleaning Column Names for Improved Data Readability and Analysis**

   For example, V_GHG_EMITTER_FACILITIES.ADDRESS1 becomes Address1 after cleaning, and we changed the column including state name to 'State' for each DataFrame for better readability.

2. **Selecting Useful Columns for Subsequent Analysis**

   We selected specific columns in the air_quality, emitter_gas_type, and emitter_facilities DataFrames that are useful for further analysis.

3. **Cleaning State and County Columns for Merging Purposes**

   We abbreviated state names and capitalized county names in each DataFrame for easier merging later.

4. **Filtering and Abbreviating Top Air Quality Measures**

   We identified five measure types appearing 34,199 times each in the MeasureName column, suggesting they are used by all counties to assess air quality. Therefore, we filtering and abbreviating top 5 air quality measures.

5. **Synchronizing Datasets by Common Years and States**

   In order to examine the relationship between the distribution of air quality and the distribution of greenhouse-related gas/facilities across counties, we first checked whether data was available for the same years. We found that only 2010 and 2011 data were available in all three DataFrames, so we focused on these years.

6. **Merging the Dataset**

   a. **Merging DataFrames emitter_gas_type and emitter_facilities**

      We merged the DataFrames based on Facility_Id and Year using an inner join.

   b. **Aggregating and Merging Emission Data by Facility and Year**

      We created a pivot table, grouped the emitter_gas_type_facilities DataFrame, and merged the grouped DataFrame with the gas_emission DataFrame using their indices as keys.

   c. **Aggregating and Merging Greenhouse Gas Emissions with County Population Data**

      We grouped the data, reset the index, and merged the aggregated dataset with another dataset containing county population information using an inner join based on the 'State' and 'County' columns.

   d. **Merging Air Quality and Gas Emissions by County and Year**

      We created a pivot table from the air_quality dataset and merged it with the county_gas dataset using 'State', 'County', and 'Year' as common columns, performing an inner join to retain only matching rows from both datasets.

After data cleaning and wrangling, we got two useful DataFrames:

1. county_gas_air_quality: This DataFrame contains information on the sum of gas emissions from all facilities within a county, along with data on the major industry type of the county, air quality measurements using various methods, and population of the county. Each row in the DataFrame represents a single county.

   This DataFrame has a total of 3470 rows × 23 columns. The columns are ['State', 'County', 'Year', 'Annual Average PM2.5', 'Max 8h Ozone Days', 'Max 8h Ozone Person-Days', 'PM2.5 Percent Days', 'PM2.5 Person-days', 'Naics_Code', 'BIOCO2', 'CH4', 'CO2', 'HFC', 'HFE', 'N2O', 'NF3', 'Other', 'Other_Full', 'PFC', 'SF6', 'Very_Short', 'FIPS', 'Population'].

   - State: The state in which the county is located.
   - Year: The year of the data for the county, either 2010 or 2011.
   - Annual Average PM2.5: The annual average ambient concentrations of PM 2.5 in micrograms per cubic meter, calculated based on seasonal averages and daily measurements (monitor and modeled

data).

- Max 8h Ozone Days: The number of days with the maximum 8-hour average ozone concentration exceeding the National Ambient Air Quality Standard (monitor and modeled data).
- Max 8h Ozone Person-Days: The number of person-days with the maximum 8-hour average ozone concentration exceeding the National Ambient Air Quality Standard (monitor and modeled data).
- PM2.5 Percent Days: The percentage of days with PM2.5 levels exceeding the National Ambient Air Quality Standard (monitor and modeled data).
- PM2.5 Person-days: The number of person-days with PM2.5 exceeding the National Ambient Air Quality Standard (monitor and modeled data).
- Naics_Code: The North American Industry Classification System code of the industry with the highest number of industry in the county. It is a standardized coding system used to classify businesses and industries based on their economic activities.
- BIOCO2, CH4, CO2, HFC, HFE, N2O, NF3, Other, Other_Full, PFC, SF6, Very_Short: The total sum of carbon dioxide equivalent emissions from different gases in all facilities in the county.
- FIPS: Federal Information Processing Standards. It is a set of codes used to uniquely identify geographic areas such as states, counties, and cities in the United States. FIPS codes are commonly used in data analysis and research to identify and categorize locations.
- Population: The total number of individuals living in that county.

	State	County	Year	Annual Average PM2.5	Max 8h Ozone Days	Max 8h Ozone Person-Days	PM2.5 Percent Days	PM2.5 Person-days	Naics_Code	BIOCO2	...	HFE	N2O	NF3	Other	Other_Full	PFC	SF6	Very_Short	FIPS	Population
0	AL	Autauga	2010	11.093826	1.0	54613.0	0.000000	0.0	221112.0	1683572.3	...	0.0	25816.634	0.0	0.0	0.0	0.0	0.0	0.0	01001	54571
1	AL	Autauga	2011	11.106039	0.0	0.0	0.000000	0.0	221112.0	1683226.6	...	0.0	26252.608	0.0	0.0	0.0	0.0	0.0	0.0	01001	54571
2	AL	Baldwin	2010	9.687065	2.0	366446.0	0.000000	0.0	562212.0	0.0	...	0.0	0.000	0.0	0.0	0.0	0.0	0.0	0.0	01003	182265
3	AL	Baldwin	2011	9.599548	2.0	373454.0	0.000000	0.0	562212.0	0.0	...	0.0	0.000	0.0	0.0	0.0	0.0	0.0	0.0	01003	182265
4	AL	Barbour	2010	11.327428	0.0	0.0	0.273973	27341.0	212324.0	0.0	...	0.0	30.992	0.0	0.0	0.0	0.0	0.0	0.0	01005	27457

Figure 1. The figure displays the first five rows of the DataFrame county_gas_air_quality.

2. facility_gas: Each row in this DataFrame represents a single facility, including information on its gas emissions and industry type.

   This DataFrame has a total of 13674 rows × 17 columns. The column names are as follows: ['Facility_Id', 'Year', 'State', 'County', 'Naics_Code', 'BIOCO2', 'CH4', 'CO2', 'HFC', 'HFE', 'N2O', 'NF3', 'Other', 'Other_Full', 'PFC', 'SF6', 'Very_Short'].

   - Facility_Id: Facility Identity number.
   - Year: The year of the data for the county, either 2010 or 2011.
   - State: The state in which the facility is located.
   - County: The county in which the facility is located.
   - Naics_Code: The North American Industry Classification System code of that facility. It is a standardized coding system used to classify businesses and industries based on their economic activities.
   - BIOCO2, CH4, CO2, HFC, HFE, N2O, NF3, Other, Other_Full, PFC, SF6, Very_Short: The total carbon dioxide equivalent emissions from different gases in that facility.

	Facility_Id	Year	State	County	Naics_Code	BIOCO2	CH4	CO2	HFC	HFE	N2O	NF3	Other	Other_Full	PFC	SF6	Very_Short
0	1000001	2010	WA	Whatcom	221112.0	0.0	138.25	292987.9	0.0	0.0	164.794	0.0	0.0	0.0	0.0	0.0	0.0
1	1000001	2011	WA	Whatcom	221112.0	0.0	17.00	35840.9	0.0	0.0	20.264	0.0	0.0	0.0	0.0	0.0	0.0
2	1000002	2010	IN	Jay	327213.0	0.0	37.00	108013.0	0.0	0.0	44.104	0.0	0.0	0.0	0.0	0.0	0.0
3	1000002	2011	IN	Jay	327213.0	0.0	37.50	109781.4	0.0	0.0	44.700	0.0	0.0	0.0	0.0	0.0	0.0
4	1000003	2010	NC	Vance	327213.0	0.0	27.75	78343.2	0.0	0.0	37.250	0.0	0.0	0.0	0.0	0.0	0.0

Figure 2. The figure displays the first five rows of the DataFrame facility_gas.

Exploratory Data Analysis

The Counties with the Worst Air Quality

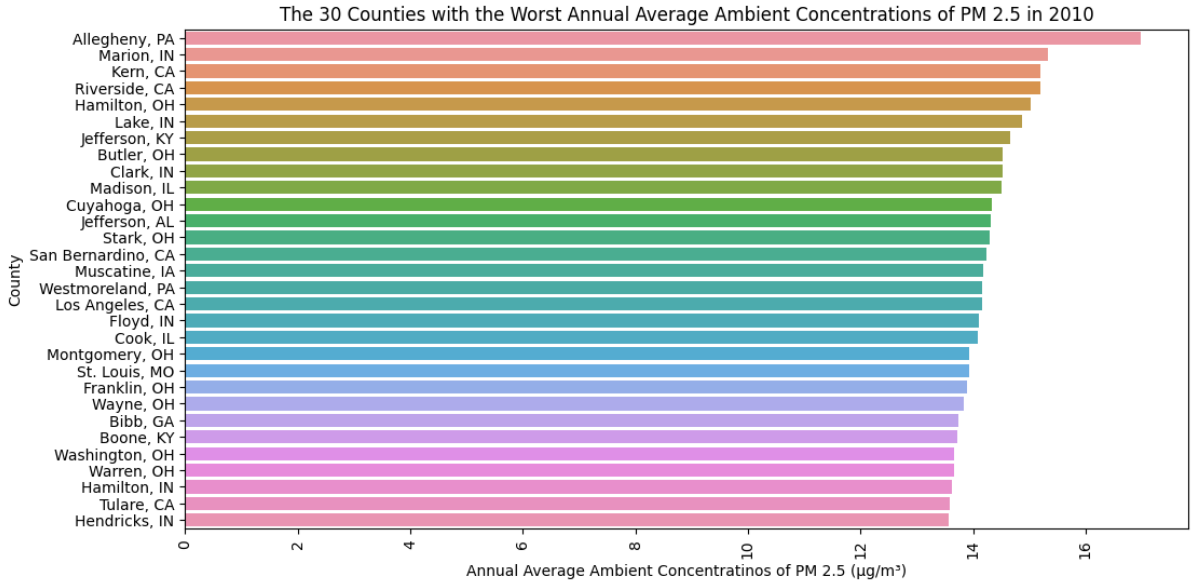

Figure 3. The figure displays the 30 counties that had the highest ambient concentration values of PM2.5 in the year 2010.

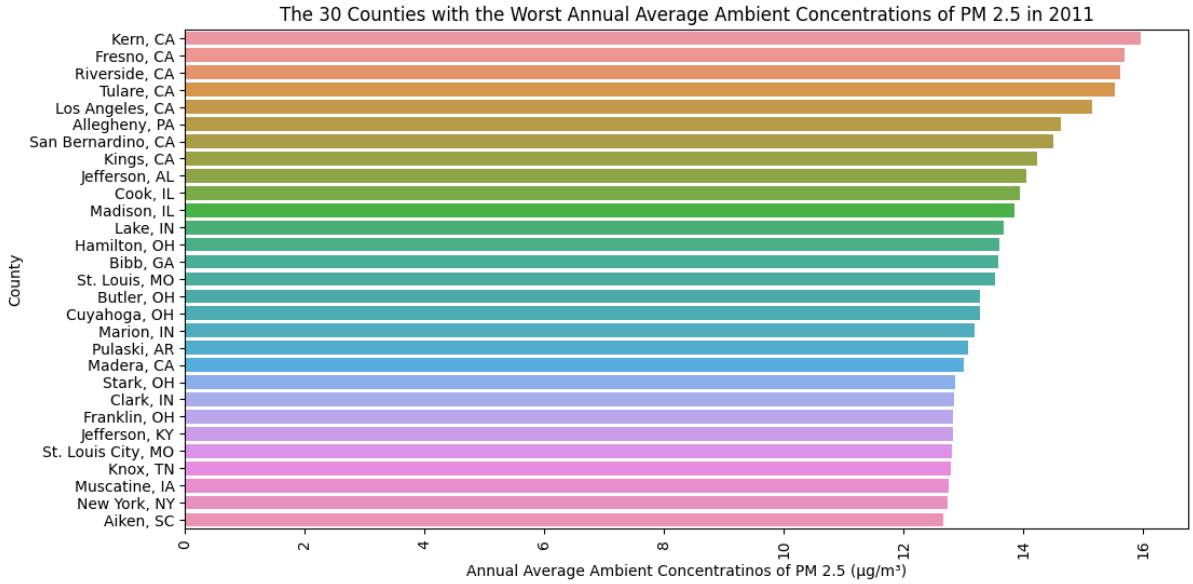

Figure 4. The figure displays the 30 counties that had the highest ambient concentration values of PM2.5 in the year 2011.

To begin with, we sorted the values of "Annual Average Ambient Concentrations of PM 2.5 in Micrograms Per Cubic Meter" for 2010 and identified the 30 counties with the highest values (i.e., the worst air quality in 2010), as shown in Figure 3. Similarly, we identified the 30 counties with the highest values in 2011 (Figure 4). It should be highlighted that in 2011, California faced a significant air pollution issue, as 7 out of the 10 counties with the highest PM2.5 index were situated within the state.

Distribution of Air Quality in Each County

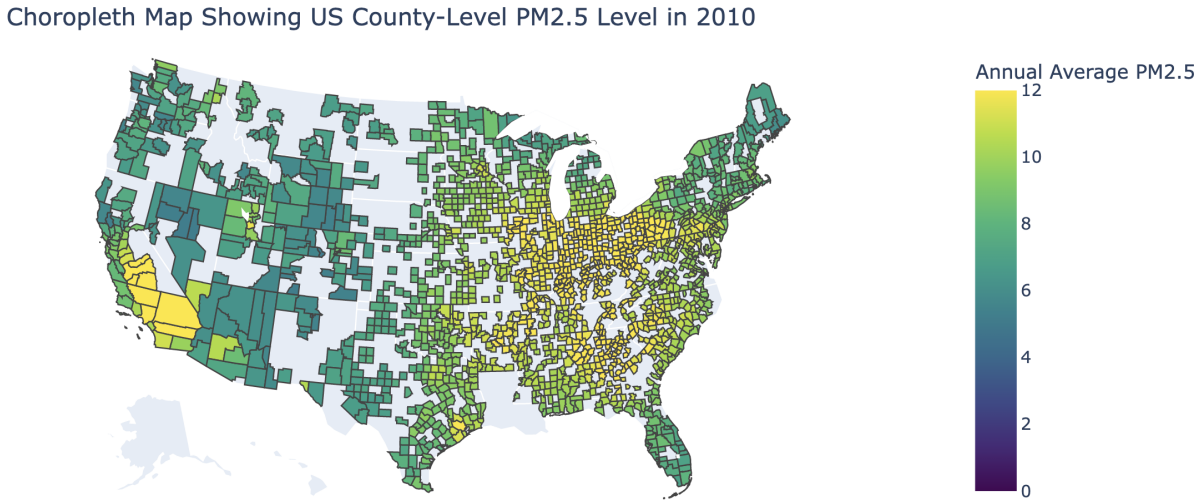

Figure 5. The figure displays the distribution of air quality (Annual average ambient concentrations of PM 2.5) in each county in 2010.

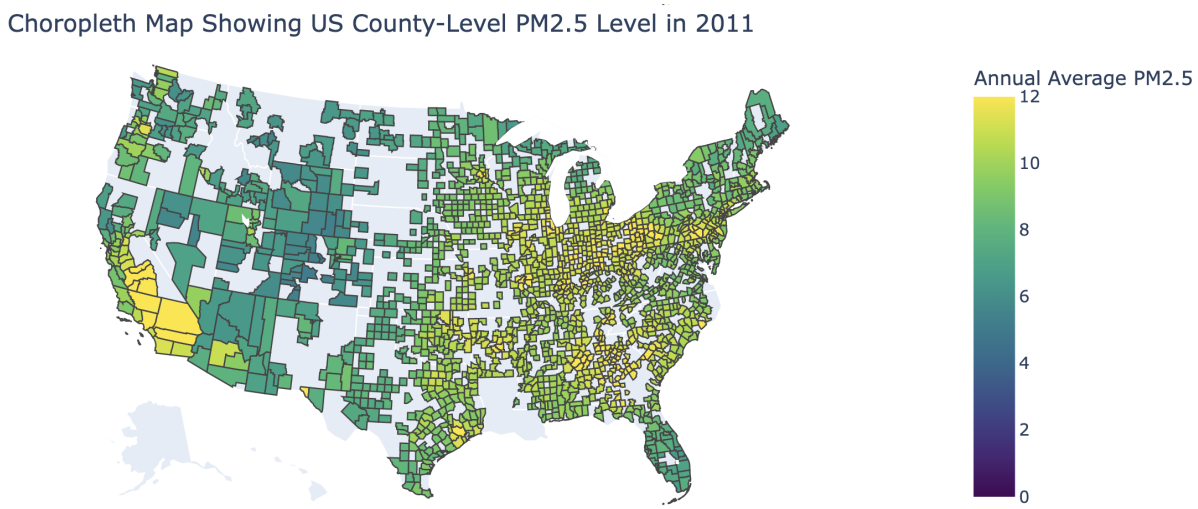

Figure 6. The figure displays the distribution of air quality (Annual average ambient concentrations of PM 2.5) in each county in 2011

To better understand air quality in the US, we created a choropleth map of the Annual Average PM2.5 level for each county in 2010 (Figure 5) and 2011 (Figure 6). The resulting visualization provides valuable insights into the distribution of air quality across the country. It is worth noting that the air quality in the eastern and western regions of the central United States was poor in 2010 and 2011.

Correlations Between Different Features

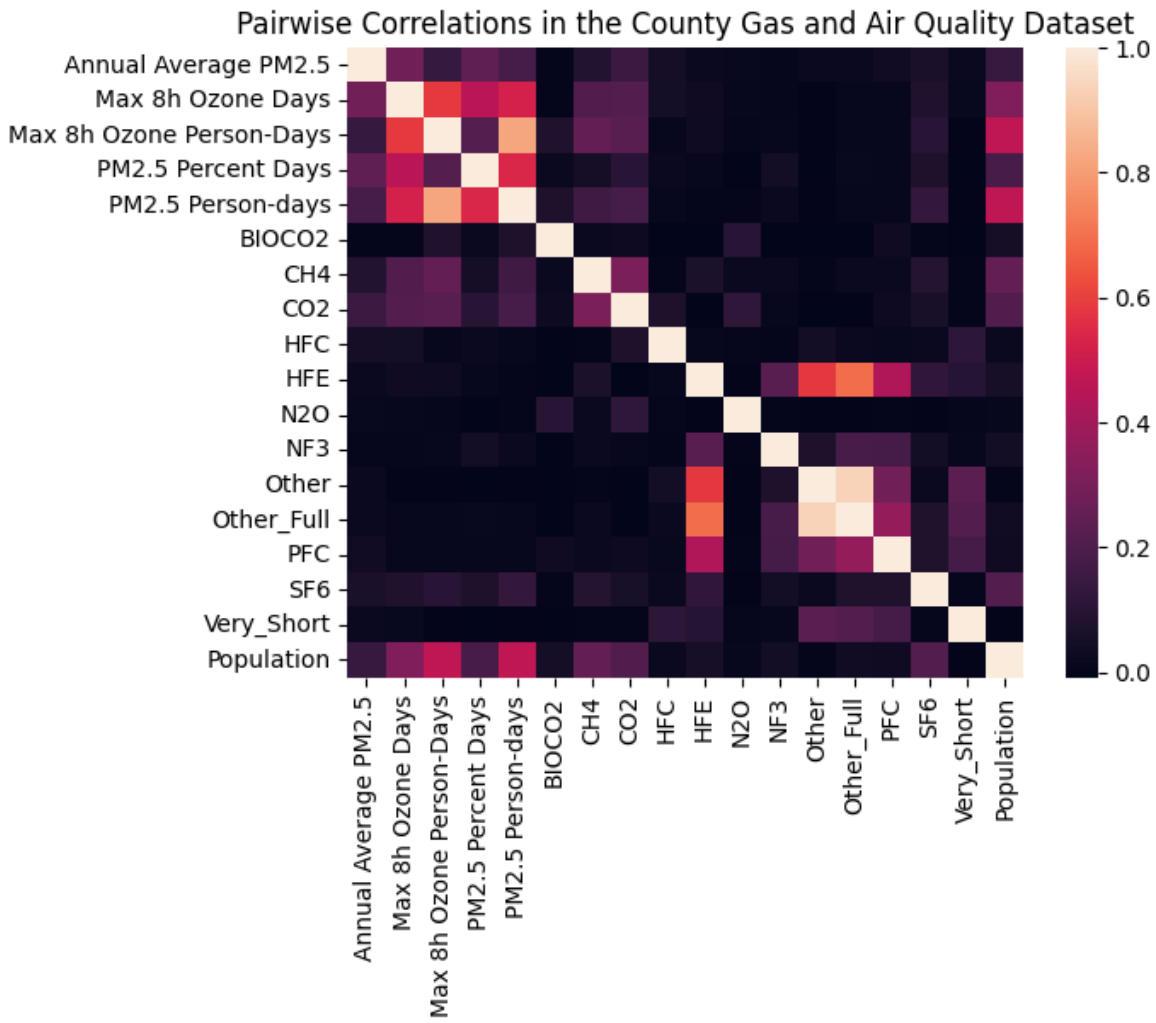

Figure 7. The figure displays the pairwise correlations in county_gas_air_quality.

We explored the correlations between different features related to air quality and gas emissions using a correlation heatmap (Figure 7), which allowed us to visualize the relationships between variables in the dataset, providing us with a comprehensive understanding of the data. There is a certain correlation between air quality measurements, and there is also a certain correlation between population and air quality indicators.

Data Modeling

Data Modeling 1 - Predicting Industry Type Based on Facility Gas Emission

We filtered the facility_gas DataFrame to retain only the rows with the top 3 most frequent industry types, which we intend to use as our predicted label (Crude Petroleum and Natural Gas Extraction, Fossil Fuel Electric Power Generation, Solid Waste Landfill).

Next, we used the gas and emission data from each facility as features to predict the industry type to which the facility belongs.

The following steps outline the procedure:

1. Select the features for analysis, which include the gas emissions of each facility (4495 rows × 11 columns).

	CH4	CO2	HFC	HFE	N2O	NF3	Other	Other_Full	PFC	SF6	Very_Short
12347	11.475000	24751.6	0.0	0.0	13.708	0.0	0.0	0.0	0.0	0.0	0.0
6213	28736.596375	0.0	0.0	0.0	0.000	0.0	0.0	0.0	0.0	0.0	0.0
5799	3105.000000	44882.6	0.0	0.0	23.542	0.0	0.0	0.0	0.0	0.0	0.0
5268	67280.000000	0.0	0.0	0.0	0.000	0.0	0.0	0.0	0.0	0.0	0.0
9835	2052.500000	0.0	0.0	0.0	0.000	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...
11071	9292.750000	3163167.2	0.0	0.0	16109.582	0.0	0.0	0.0	0.0	0.0	0.0
6389	98.500000	203099.6	0.0	0.0	116.220	0.0	0.0	0.0	0.0	0.0	0.0
2248	417.500000	899562.0	0.0	0.0	497.660	0.0	0.0	0.0	0.0	0.0	0.0
4451	4525.000000	25231.6	0.0	0.0	17.880	0.0	0.0	0.0	0.0	0.0	0.0
4973	21.000000	44672.0	0.0	0.0	23.840	0.0	0.0	0.0	0.0	0.0	0.0

4495 rows × 11 columns

Figure 8. The figure displays the features of training set data modeling 1.

2. Split the data into training and validation sets.
3. Standardize the features by subtracting the mean and scaling to unit variance.
4. Build and train a logistic regression model using the training data.
5. Evaluate the performance of the model on the validation set.

Data Modeling 2 - Predicting County Air Quality Based on Facility Gas Emission Sum and Major Industry Type

We first filtered county_gas_air_quality to include only the counties whose major industry type is among the top 10 most frequent industry types. Next, we performed one-hot encoding on these industry types to create new features for prediction.

To create an air quality index, we categorized the "Percent of days with PM2.5 levels over the National Ambient Air Quality Standard" variable into three severity levels based on their values, with the highest level indicating the worst air quality. It was observed that the original data for PM2.5 Percent Days was primarily composed of 0 and 0.273973 values. Consequently, the data was categorized into three levels (Good (0), Moderate (<0.3), and Poor (≥0.3)).

Using the sum of gas and emission data from facilities in each county, as well as population and other relevant features, our goal is to predict the severity levels of air quality.

The following steps outline the procedure:

1. Categorize the percentage of days with PM2.5 levels into three severity levels (Good, Moderate, Poor).
2. Perform one-hot encoding on industry types to create new features.
3. Select the features for analysis, which include the sum of gas emissions of all facilities in each county, population, and the one-hot encoded major industry type of the county (1803 rows × 25 columns).

	Annual Average PM2.5	Max 8h Ozone Days	BIOCO2	CH4	CO2	HFC	HFE	N2O	NF3	Other	Other_Full	PFC	SF6	Very_Short	Population
1008	1.634645	-0.407757	-0.251288	0.171169	0.672550	-0.024351	-0.041149	-0.019764	-0.052187	-0.026529	-0.048762	-0.085156	-0.097641	-0.035936	-0.138602
3126	0.010471	-0.407757	-0.251288	-0.130113	-0.458462	-0.024351	-0.041149	-0.075408	-0.052187	-0.026529	-0.048762	-0.085156	-0.097641	-0.035936	-0.239632
2890	0.041977	0.735611	-0.246686	-0.202582	0.097151	-0.024351	-0.041149	-0.065062	-0.052187	-0.026529	-0.048762	-0.085156	-0.097641	-0.035936	-0.044714
863	0.170236	-0.026634	-0.251288	-0.168116	-0.235011	-0.024351	-0.041149	-0.067481	-0.052187	-0.026529	-0.048762	-0.085156	-0.097641	-0.035936	-0.332418
2861	-1.455494	-0.407757	-0.251288	-0.205959	-0.363800	-0.024351	-0.041149	-0.074629	-0.052187	-0.026529	-0.048762	-0.085156	-0.097641	-0.035936	-0.310756
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1394	0.057459	0.545050	-0.246501	-0.153902	-0.087220	-0.024351	-0.041149	-0.047729	-0.052187	-0.026529	-0.048762	-0.085156	-0.097641	-0.035936	-0.015969
2367	-0.386432	-0.407757	-0.251288	-0.206237	-0.432264	-0.024351	-0.041149	-0.075191	-0.052187	-0.026529	-0.048762	-0.085156	-0.097641	-0.035936	-0.315506
1004	1.933264	-0.217196	-0.208504	-0.145321	-0.461593	-0.024351	-0.041149	-0.075272	-0.052187	-0.026529	-0.048762	-0.085156	-0.097641	-0.035936	-0.001646
1091	-0.422604	-0.407757	-0.251288	-0.036177	-0.454657	-0.024351	-0.041149	-0.075386	-0.052187	-0.026529	-0.048762	-0.085156	-0.097641	-0.035936	-0.238175
2255	1.340213	-0.217196	-0.251288	0.176727	0.202540	-0.024351	-0.041149	-0.025016	-0.052187	-0.026529	-0.048762	-0.085156	-0.097641	-0.035936	0.230226

1803 rows × 15 columns

Figure 9. The figure displays the first part features of training set data modeling 2, including the sum of gas emissions of all facilities in each county, population, two measurement results.

	Industry_Cement Manufacturing	Industry_Crude Petroleum and Natural Gas Extraction	Industry_Ethyl Alcohol Manufacturing	Industry_Fossil Fuel Electric Power Generation	Industry_Iron and Steel Mills	Industry_Natural Gas Liquid Extraction	Industry_Paper (except Newsprint) Mills	Industry_Paperboard Mills	Industry_Pipeline Transportation of Natural Gas	Industry_Solid Waste Landfill
1008	0	0	0	1	0	0	0	0	0	0
3126	0	0	0	1	0	0	0	0	0	0
2890	0	0	0	1	0	0	0	0	0	0
863	1	0	0	0	0	0	0	0	0	0
2861	0	0	0	0	0	1	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...
1394	0	0	0	1	0	0	0	0	0	0
2367	0	0	0	0	0	1	0	0	0	0
1004	0	0	0	1	0	0	0	0	0	0
1091	0	0	0	0	0	0	0	0	0	1
2255	0	0	0	1	0	0	0	0	0	0

1803 rows × 10 columns

Figure 10. The figure displays the second part features of training set data modeling 2, including the one-hot encoded major industry type of the county

4. Split the data into training and validation sets.
5. Standardize the non one-hot encoded features by subtracting the mean and scaling to unit variance.
6. Build and train a logistic regression model using the training data.
7. Evaluate the performance of the model on the validation set.

Analysis of Results

Data Modeling 1 Result

Model 1 Training Accuracy is 0.8133481646273637. Model 1 Validation Accuracy is 0.7612869745718733.

	Precision	Recall
Crude Petroleum and Natural Gas Extraction	0.91	0.06
Fossil Fuel Electric Power Generation	0.70	1.00
Solid Waste Landfill	0.86	0.83

Table 1. The table displays the data modeling 1 result.

The first model aims to predict facility industry type, by utilizing the gas and emission data from each facility as features. A logistic regression model was established. The data was divided into training and test datasets at an 70:30 ratio, yielding a training accuracy of 0.8133 and a test accuracy of 0.7613. The discrepancy

between the training and testing datasets indicates that the model is slightly overfitting. An accuracy exceeding 75% suggests that appropriate features were selected to predict facility industry type.

As accuracy alone cannot comprehensively represent a model's performance, Precision and Recall metrics were also employed to analyze Model 1's performance. The model demonstrated high precision (0.70) and recall (1.0) for the "Fossil Fuel Electric Power Generation" category and high precision (0.86) and recall (0.83) for the "Solid Waste Landfill" category, reflecting its efficiency in identifying counties belonging to this group. In contrast, the "Crude Petroleum and Natural Gas Extraction" category exhibited high precision (0.91) and low recall (0.06), suggesting that the model's insufficient features to distinguish "Crude Petroleum and Natural Gas Extraction" from other categories. These findings indicate a potential need for additional features with strong correlations to facility to further enhance the model's performance.

We utilized the gas and emission information of various facilities as features to predict the industrial sector to which each facility belongs. This model can be applied to identify the specific type of engineering involved in a facility. If gases are detected near a factory that do not correspond to its expected emission patterns, it could indicate that the facility is not operating according to its designated industry type, and may be engaging in fraudulent activities.

Data Modeling 2 Result

Model 2 Training Accuracy is 0.7881308929561841. Model 2 Validation Accuracy is 0.7868217054263565.

	Precision	Recall
Good	0.80	0.98
Moderate	0.25	0.03
Poor	0.57	0.24

Table 2. The table displays the data modeling 2 result.

The second model aims to predict air quality, as indicated by PM2.5 Percent Days, by utilizing Annual Average PM2.5, Max 8h Ozone Days, population distribution, greenhouse gas distribution, and major industry distribution across various counties. A logistic regression model was established. The data was divided into training and test datasets at an 70:30 ratio, yielding a training accuracy of 0.7881 and a validation accuracy of 0.7868. The minor discrepancy between the training and validation datasets indicates that the model is not overfitting. An accuracy exceeding 75% suggests that appropriate features were selected to predict PM2.5 Percent Days levels.

As accuracy alone cannot comprehensively represent a model's performance, particularly given the data distribution's heavy skew towards the "Good" category, Precision and Recall metrics were also employed to analyze Model 2's performance. The model demonstrated high precision (0.80) and recall (0.98) for the "Good" category, reflecting its efficiency in identifying counties belonging to this group. This is consistent with the dominant presence of the "Good" category in the dataset. In contrast, the "Moderate" category exhibited low precision (0.25) and recall (0.03), suggesting that the model's efficiency in identifying this group may be hindered by an underrepresentation of data points or insufficient features to distinguish "Moderate" from other categories. The "Poor" category's precision (0.57) and recall (0.24) fell between those of the "Good" and "Moderate" groups. This outcome implies that the selected features exhibit a discernible differential distribution between the "Poor" category and others, but the differences may not be significant or the features may be inadequate. Interestingly, despite having fewer data points than the "Moderate" group, the

model performed better in identifying the "Poor" category, suggesting that the number of data points is not the sole determinant of model performance. These findings indicate a potential need for additional features with strong correlations to air quality to further enhance the model's performance.

The establishment of this model revealed that industrial facilities and population significantly augment the model's predictive power, indicating that these factors are primary contributors to air pollution. Moreover, the model's suggested link between greenhouse gas and facilities, as well as the connection between facilities and PM2.5 Percent Days, implies that the impact of greenhouse gas on air quality may be intermediate but cannot be disregarded.

Conclusion and Future Work

We comprehensively investigated our inquiries through a meticulous examination of the data analysis. This process entail combining data from various datasets, identifying possible correlations between features, and presenting the results via tables and figures. Our models reveal the interrelationships among industries that emit harmful gases, greenhouse gas, population distribution, and air quality. This model is anticipated to be a dependable reference point for any prospective legislation being proposed.

References

1. Zhang, Q., Zheng, Y., Tong, D., et al. (2019). Drivers of improved PM2.5 air quality in China from 2013 to 2017. Proceedings of the National Academy of Sciences, 116(49), 24463-24469.
2. Velders, G.J.M.; Maas, R.J.M.; Geilenkirchen, G.P.; de Leeuw, F.A.A.M.; Ligterink, N.E.; Ruyssenaars, P.; de Vries, W.J.; Wesseling, J. Effects of European emission reductions on air quality in the Netherlands and the associated health effects. *Atmospheric Environment* **2020**, *221*, 117109, doi:https://doi.org/10.1016/j.atmosenv.2019.117109
3. Yerramilli, A.; Dodla, V.B.; Desamsetti, S.; Challa, S.V.; Young, J.H.; Patrick, C.; Baham, J.M.; Hughes, R.L.; Yerramilli, S.; Tuluri, F.; et al. Air quality modeling for the urban Jackson, Mississippi Region using a high resolution WRF/Chem model. *Int J Environ Res Public Health* **2011,** *8*, 2470-2490, doi:10.3390/ijerph8062470.
4. Tsai, W.-T.; Lin, Y.-Q. Trend Analysis of Air Quality Index (AQI) and Greenhouse Gas (GHG) Emissions in Taiwan and Their Regulatory Countermeasures. *Environments* **2021**, *8*, 29. https://doi.org/10.3390/environments8040029
5. Yerramilli, A.; Dodla, V.B.; Desamsetti, S.; Challa, S.V.; Young, J.H.; Patrick, C.; Baham, J.M.; Hughes, R.L.; Yerramilli, S.; Tuluri, F.; Hardy, M.G.; Swanier, S.J. Air Quality Modeling for the Urban Jackson, Mississippi Region Using a High Resolution WRF/Chem Model. *Int. J. Environ. Res. Public Health* **2011**, *8*, 2470-2490. https://doi.org/10.3390/ijerph8062470
6. Health, C.S.D.o.P. Air Pollution. Available online: (accessed on April 27, 2023).