ENV872 final project

https://github.com/Tonysflex/Xianhang-Xie_ENV872_EDA_FinalProject (https://github.com/Tonysflex/Xianhang-Xie_ENV872_EDA_FinalProject)

Xianhang Xie

1 Rationale and Research Questions

The release of toxic chemicals and greenhouse gases into the environment poses a significant threat to public health and the planet's sustainability. The EPA's Toxic Release Inventory (TRI) and Greenhouse Gas Reporting Inventory (GHG) collect emissions data from facilities across the United States. The TRI Program covers chemicals that cause cancer or other chronic human health effects, significant adverse acute human health effects, and significant adverse environmental effects.

The dataset combines geographical and industry-related data with facility-level emissions data from 2010 to 2014 to provide a comprehensive understanding of the sources and patterns of emissions across the country. Understanding the patterns of emissions is critical to designing effective mitigation and remediation strategies that can help reduce the harmful effects of these emissions on both the environment and human health. This analysis will provide valuable insights into the sources and patterns of emissions in different industries and regions, and these insights can be used to inform policy and regulatory decisions.

The central research questions for this analysis include identifying the States and Industries with the highest levels of toxic release and greenhouse gas emissions, as ranked by the Tri.Rank and GHG.Rank respectively, where 1 = most emission. Additionally, this analysis will explore the spatial distribution of emissions to identify if specific regions in the United States were more heavily impacted by toxic release and greenhouse gas emissions than others. Finally, this analysis will examine the impact of the current administration's decision to cut the EPA's budget on emissions data collection and analysis. The findings of this analysis will provide valuable insights into the sources and patterns of emissions in the United States and can inform policy decisions aimed at mitigating the impact of these emissions on both the environment and public health.

2 Research Questions

Question 1: How do the Green house gas and Toxic Release Inventory emission distributed over regions and industry types

Question 2: How do the Green house gas and Toxic Release Inventory emission change over time?

3 Dataset Information

The dataset provided combines the 2010-2014 “Facility-Level” emissions data with geographical and industry-related data. It is based on the EPA’s Toxic Release Inventory (TRI) and Greenhouse Gas Reporting Inventory (GHG), the national system of nomenclature that is used to describe industry-related emissions.

Chemicals covered by the TRI Program are those that cause:

- Cancer or other chronic human health effects
- Significant adverse acute human health effects
- Significant adverse environmental effects

The dataset contains 28 columnar variables, including UniqueID, Facility name, Rank TRI ’14, Rank GHG ’14, Latitude, Longitude, Location address, City, State, ZIP, County, FIPS code, Primary NAICS, Second primary NAICS, Third primary NAICS, Industry type, Parent companies 2014 (GHG), Parent companies 2014 (TRI), TRI air emissions 14 (in pounds), TRI air emissions 13 [and previous years], GHG direct emissions 14 (in metric tons), GHG direct emissions 13 [and previous years], GHG Facility Id, Second GHG Facility Id [and Third, Fourth, etc.], TRI Id, Second TRI Id [and Third, Fourth, etc.], FRS Id, Second FRS Id [and Third, Fourth, etc.]. The dataset was made available by the Center for Public Integrity. It can be downloaded from Kaggle: us-facilitylevel-air-pollution-20102014.

4 Data Wrangling

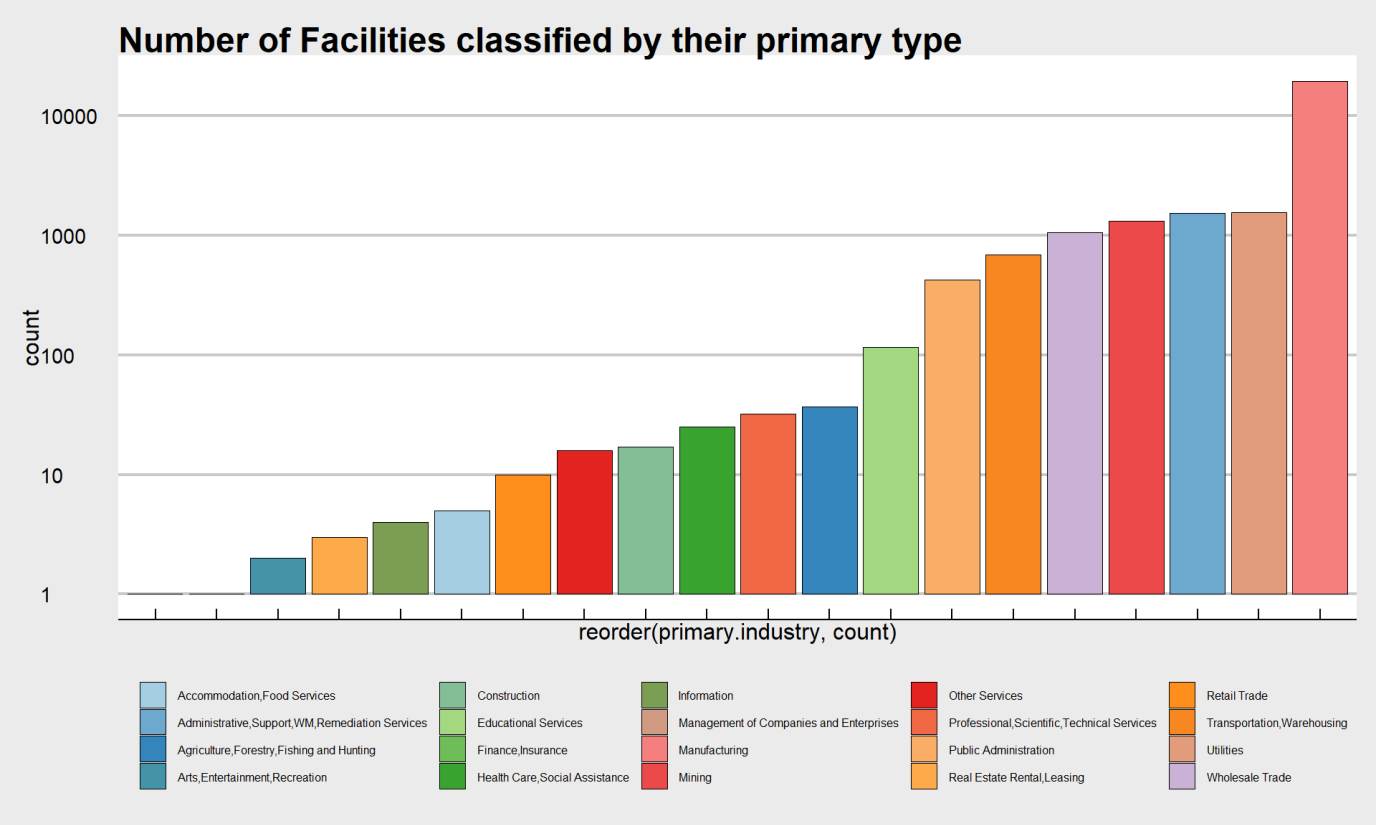
In this data wrangling process, we started by renaming some of the columns to improve their readability. This is because the original dataset had some columns with unclear or ambiguous names. We then converted some of the columns from character type to numeric type. Specifically, we converted columns 3-4 and 19-28 from character to numeric using a conversion function. We then focused on the North American Industry Classification System (NAICS) codes. The NAICS codes are used to classify business establishments for statistical purposes. We created two lookup tables for the main industries classification (2 and 3 digits), which we labeled as naics_first_level and naics_second_level. Using the lookup tables, we imputed the primary and second industry columns in the dataset. We used the first four digits of the Primary.NAICS column to match with naics_first_level and the first five digits to match with naics_second_level. These preprocessing steps will facilitate our data analysis by making the data more manageable and easier to work with. The new column names are more descriptive, and the data type conversion will allow us to perform calculations on the columns of interest.

5 Exploratory Analysis

The exploratory data analysis (EDA) performed on our dataset involves the utilization of visualizations and summarization methods to gain insight into the dataset’s characteristics and distribution. We first begin with a visualization of facility counts by their primary industry, using a log scale to better represent the vast differences in facility counts among industries. The majority of the facilities are concentrated in industries such as Manufacturing, Utilities, Administrative and Support and Waste Management and Remediation Services, Mining, and Wholesale Trade.

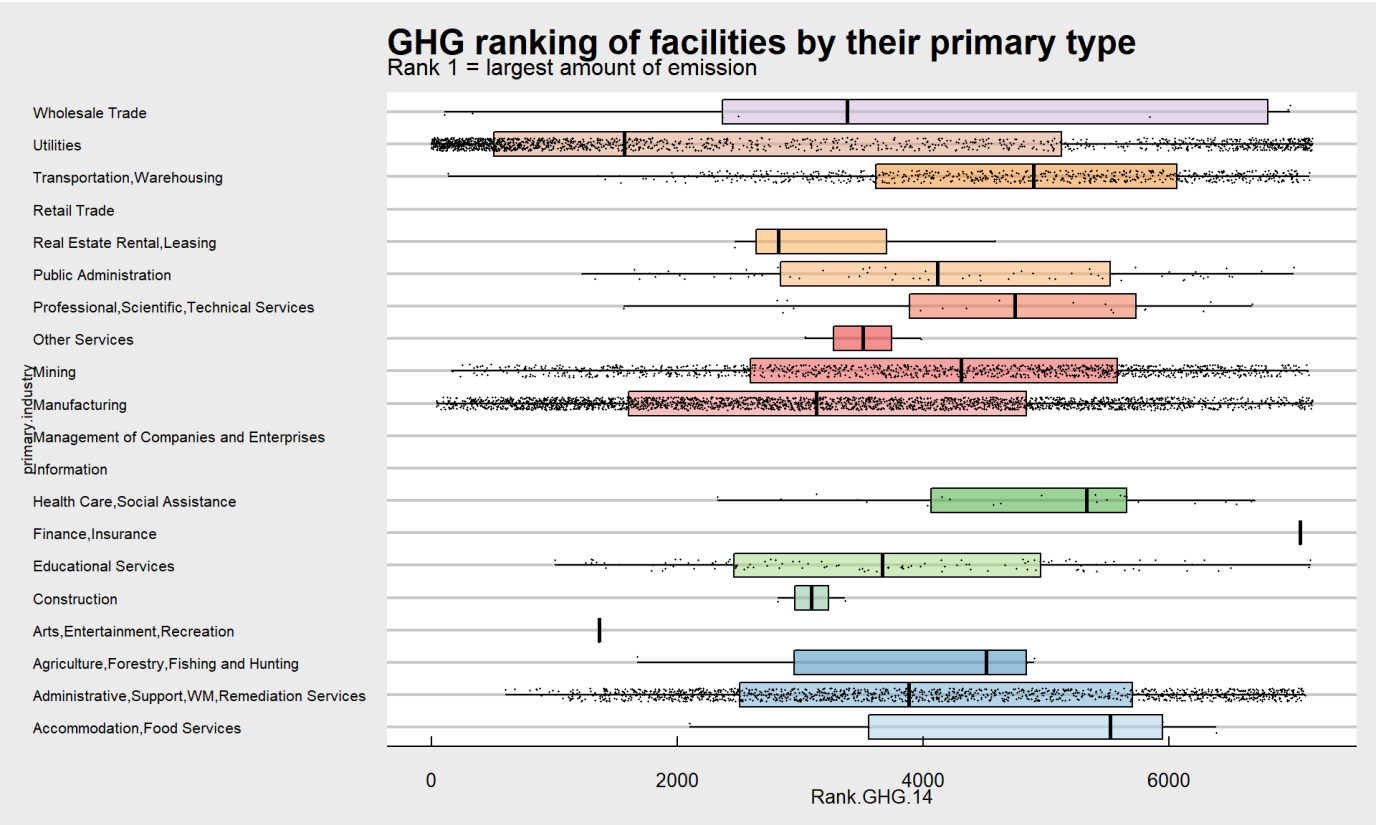
Then we make the next visualizations concentrate on environmental impact rankings of facilities, specifically the greenhouse gas (GHG) and Toxic Release Inventory (TRI) rankings in 2014. These rankings are plotted by the primary industry in a boxplot with jittered points that illustrate the distribution of rankings across industries. A summary table is also included, detailing the count of facilities and the mean rank for each industry in descending order of facility count. The table and plots help identify which industries have the highest environmental impact and those with the greatest variance in environmental impact rankings.

- Below is the visualization of facility counts by their primary industry, using a log scale to better represent the vast differences in facility counts among industries.



The plot shows that the distribution of facilities across primary industries is highly uneven. The majority of facilities belong to a few industries, such as Manufacturing, Utilities, Administrative and Support and Waste Management and Remediation Services, Mining, and Wholesale Trade. Conversely, the Accommodation, Food Services, Arts, Entertainment, Recreation, Finance, Insurance, and Management of Companies and Enterprises industries have relatively few observations, with less than 10 facilities in each.

- Below is the boxplot of greenhouse gas (GHG) 2014 ranking by the industry caterogy.

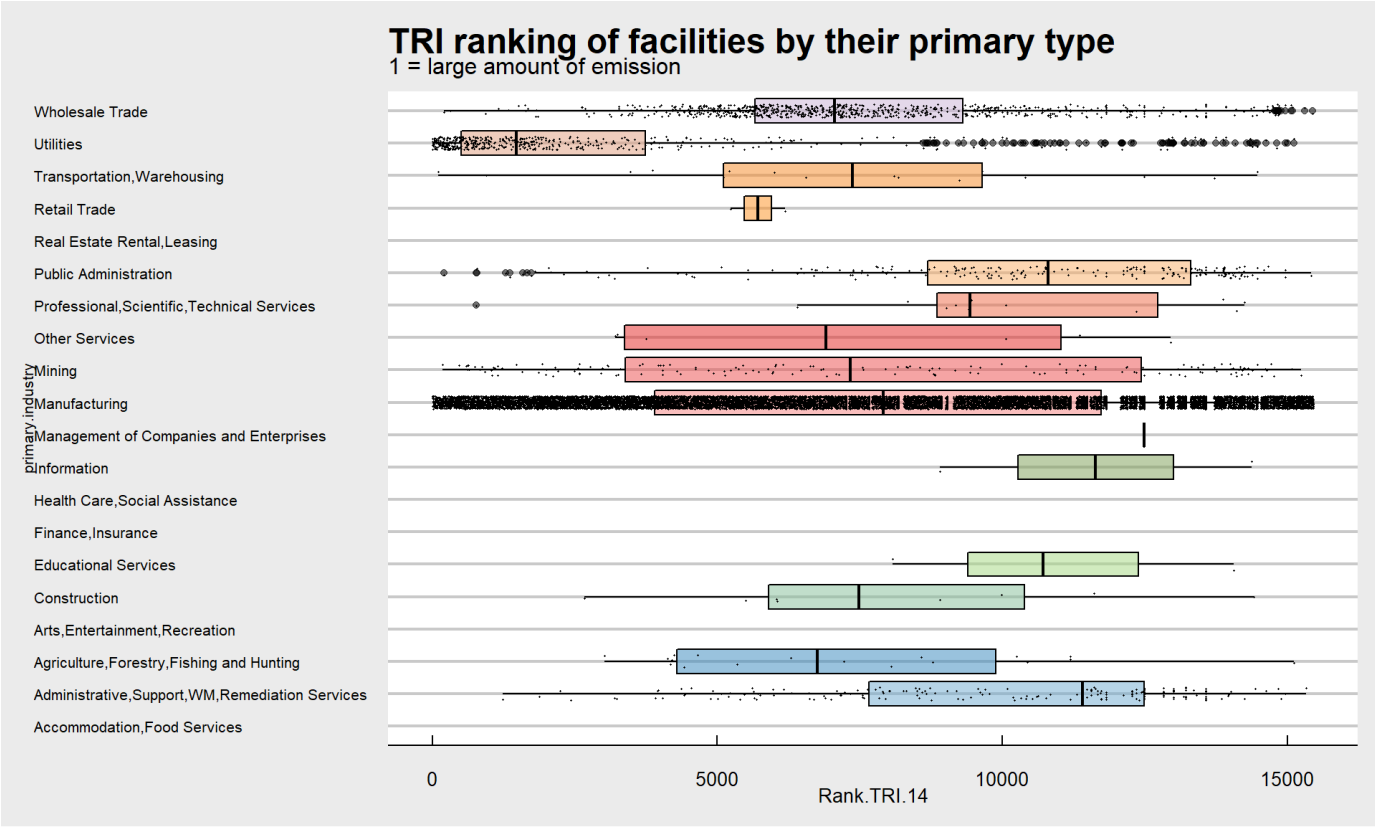


- Below is the table showing the first 6 industry with most counts and their mean rank of green house gas.

primary.industry <chr>	count <int>	meanGHG <dbl>
Manufacturing	2320	3266.039
Utilities	1503	2701.919
Administrative,Support,WM,Remediation Services	1307	4045.892
Mining	1133	4073.612
Transportation,Warehousing	653	4795.291
Educational Services	112	3705.384

6 rows

- Below is the boxplot of Toxic Release Inventory (TRI) 2014 ranking by the industry caterogy.



- Below is the table showing the first 6 industry with most counts and their mean rank of Toxic Release Inventory (TRI).

primary.industry <chr>	count <int>	meanGHG <dbl>
Manufacturing	13476	7838.367
Wholesale Trade	803	7586.588
Utilities	546	3135.068
Public Administration	219	10385.018
Administrative,Support,WM,Remediation Services	190	10075.368

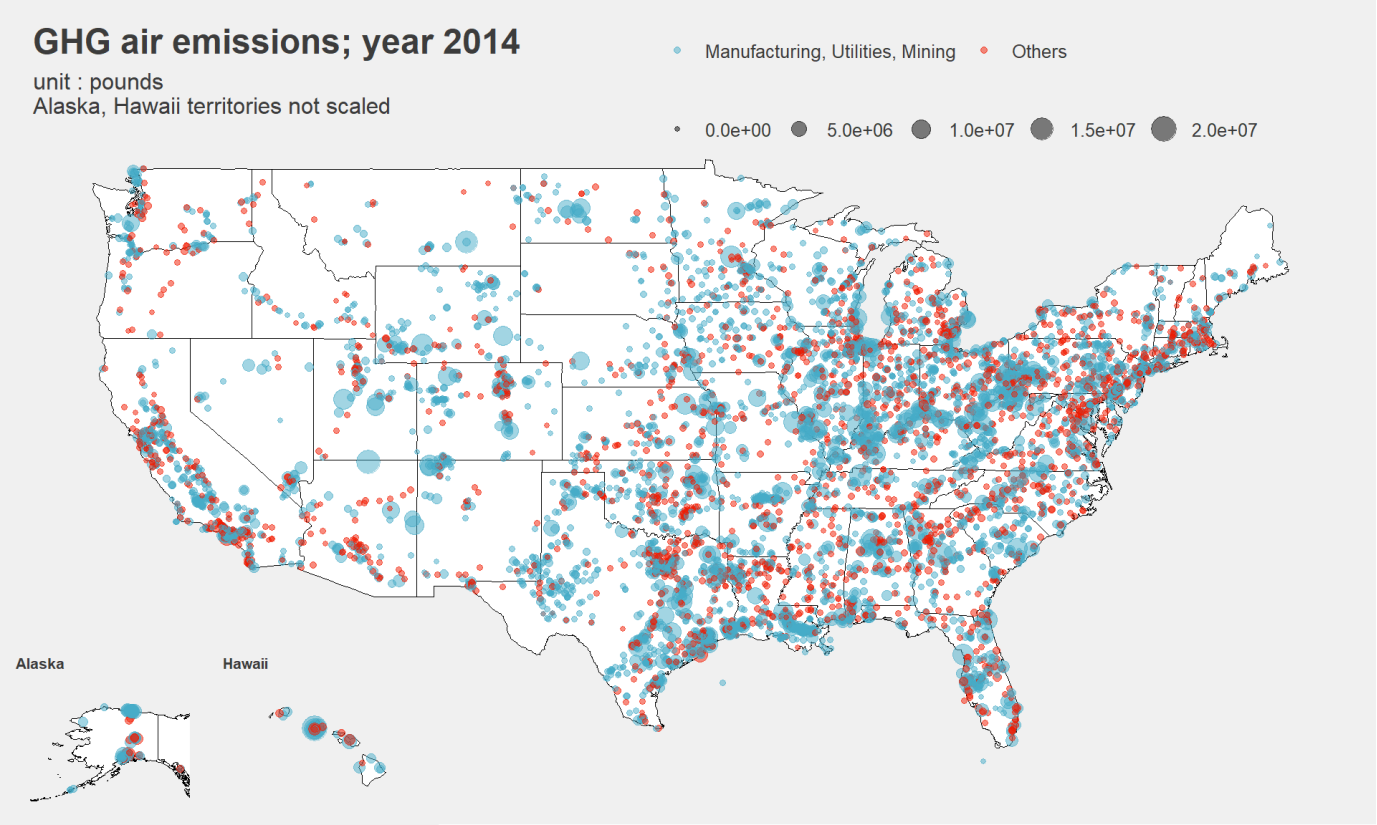
primary.industry <chr>	count <int>	meanGHG <dbl>
Mining	158	7740.076

6 rows

6 Analysis

6.1 Geographical distribution of green house gas emission of year 2014

In this section, new categories are defined to differentiate between the three primary industries with the highest number of facilities in the dataset (Manufacturing, Utilities, and Mining) and the remaining industries, for the purpose of plotting. This will allow for easier visualization and comparison between these groups.

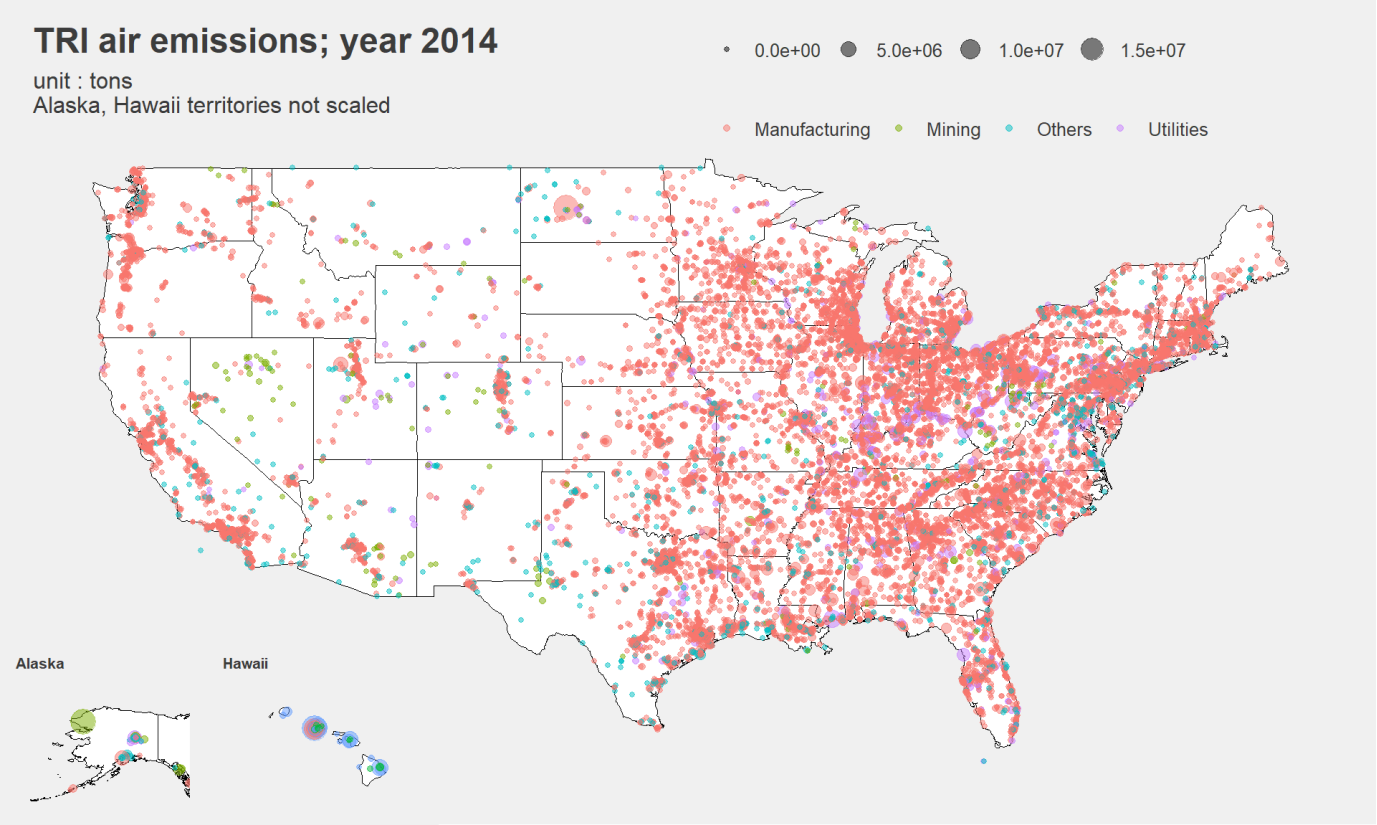


Here we display the geographical distribution of facilities in the US with their greenhouse gas (GHG) direct emissions in 2014, categorized by primary industry. The map is divided into two parts: the mainland US and the Hawaii and Alaska territories. The facilities in Manufacturing, Utilities, and Mining industries are categorized together, while the remaining industries are grouped into the “Others” category. The size of the points represents the magnitude of GHG direct emissions, and the color represents the primary industry category. The legend indicates which color represents each category. Alaska and Hawaii territories are shown on separate maps, as they are not scaled to the mainland US.

The plot reveals that the Manufacturing, Utilities, and Mining industries emit significantly more greenhouse gas compared to other industries, which is consistent with our expectations. Large greenhouse gas emissions are widespread across the United States and could potentially represent power plants. Additionally, regions such

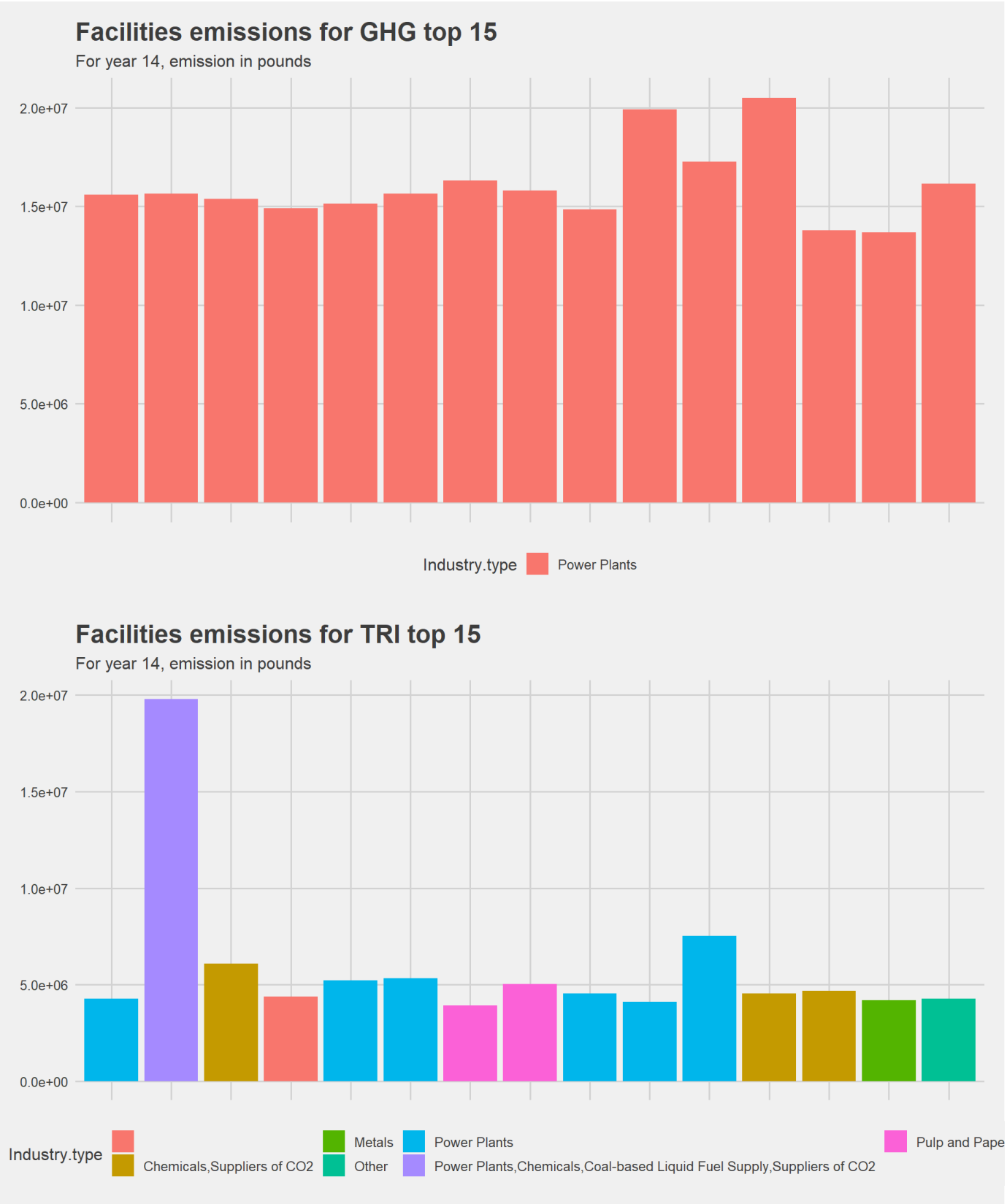
as the lake area, Texas, Florida, and the East Coast have higher emissions due to the high concentration of manufacturing industries in these areas.

6.2 Geographical distribution of Toxic Release Inventory emission of year 2014



The plot indicates that manufacturing is responsible for the largest Toxic Release Inventory (TRI) air emissions, followed by utilities. Similar to the previous plot, the regions with more industrial activity, such as the Lake area, Texas, Florida, and the East Coast, have higher TRI emissions.

6.3 The top 10 emissions



The analysis of the dataset reveals interesting findings regarding the top 15 emitters of greenhouse gas (GHG) and Toxic Release Inventory (TRI). The plot shows that all of the top 15 emitters of GHG are power plants, indicating that the energy sector has a significant impact on GHG emissions. On the other hand, the top 15 emitters of TRI consist of mostly power plants, as well as metals, paper, and other industries. This suggests that a wider range of industries are responsible for TRI emissions, with power plants still being a major contributor. These findings highlight the importance of monitoring and regulating the emissions of power plants in the US, as well as implementing policies to reduce emissions from a wider range of industries.

6.4 Hypothesis test for emissions across all industries

In order to determine if there is a significant difference in greenhouse gas (GHG) and Toxic Release Inventory (TRI) emissions across all industries, an ANOVA (Analysis of Variance) test can be conducted. The null hypothesis would be that there is no significant difference in GHG or TRI emissions across industries, while the alternative hypothesis would be that at least one industry has a different mean GHG emission than the others. The ANOVA test would require calculating the sum of squares between and within groups to calculate the F-statistic and corresponding p-value. If the p-value is less than the predetermined significance level (such as 0.05), we would reject the null hypothesis and conclude that at least one industry has a different mean GHG emission than the others.

- Green house gas

```
##                    Df    Sum Sq   Mean Sq F value Pr(>F)
## primary.industry   16 1.754e+15 1.096e+14   60.75 <2e-16 ***
## Residuals        7218 1.302e+16 1.804e+12
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 因为不存在，19129个观察量被删除了
```

The resulting P value is smaller than 0.05, thus we can conclude at least one industry type is different from others in GHG emission.

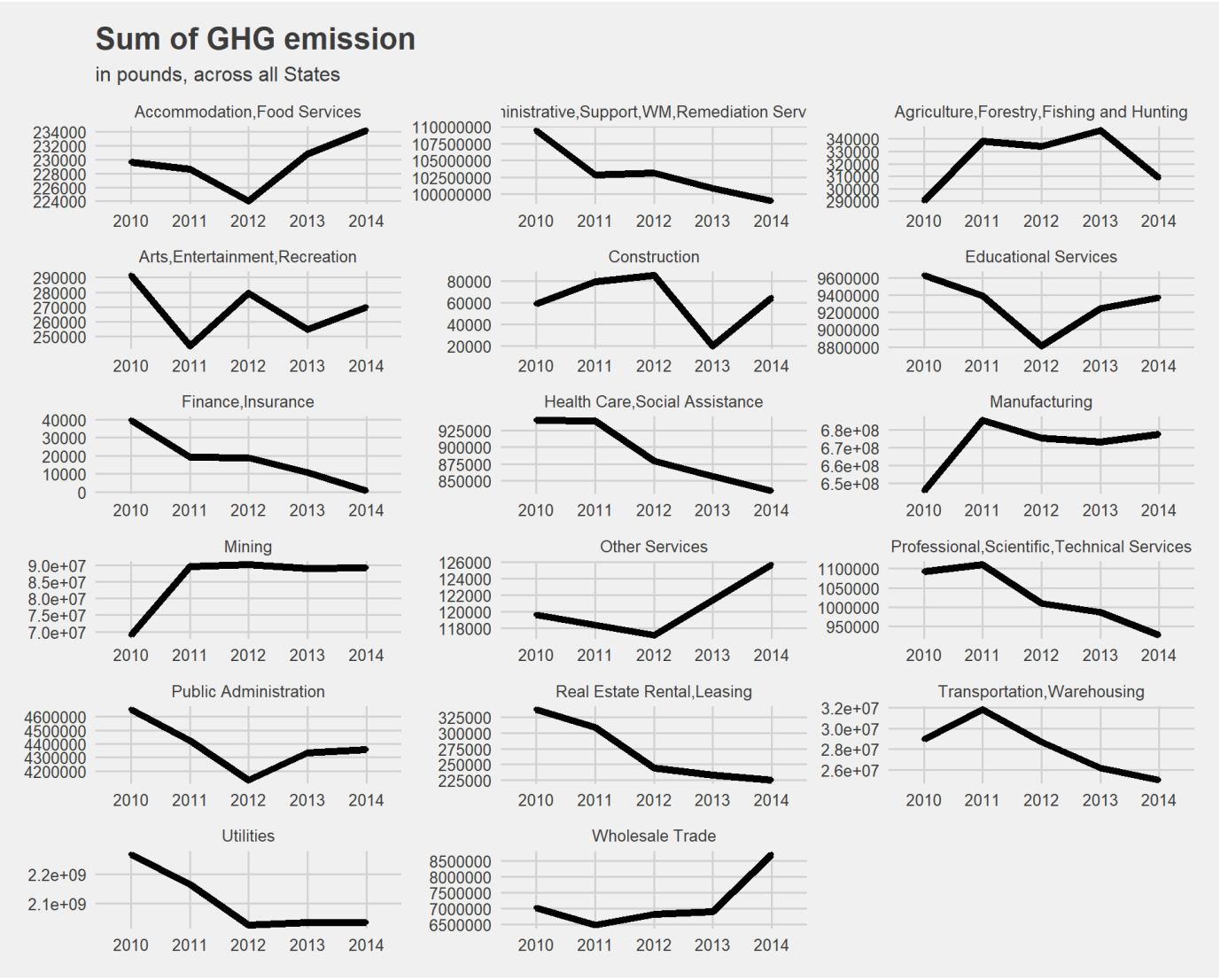
- Toxic Release Inventory

```
##                     Df    Sum Sq   Mean Sq F value Pr(>F)
## primary.industry    15 4.802e+13 3.201e+12   52.67 <2e-16 ***
## Residuals        21709 1.319e+15 6.078e+10
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 因为不存在，4639个观察量被删除了
```

The resulting P value is smaller than 0.05, thus we can conclude at least one industry type is different from others in TRI emission.

6.5 The 2010-2014 greenhouse gas (GHG)

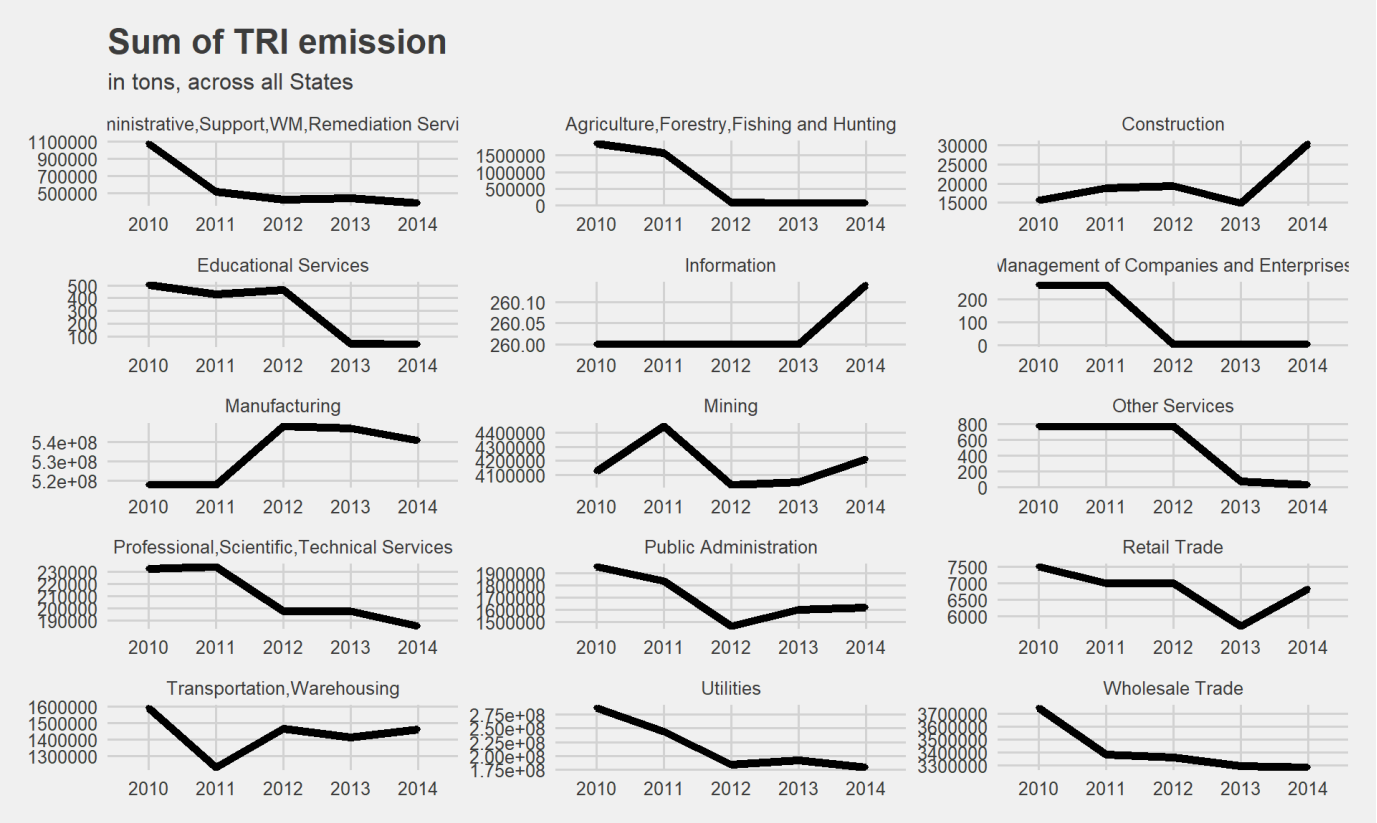
emission analysis



The trends in greenhouse gas (GHG) emissions for different industries from 2010 to 2014 were examined, and it was found that Manufacturing and Utilities continued to be the two main contributors with no indication of a decrease. Finance showed significant improvement by reducing its GHG emissions to less than 1000 pounds. Agriculture initially saw a 20% increase in GHG emissions but eventually decreased to a 10% increase compared to 2010 by 2014. Mining experienced a 30% increase in GHG emissions in 2011 and remained stable from 2011 to 2014. Finally, Wholesale Trade showed a 17% increase in GHG emissions in 2014 compared to 2013. These trends highlight the different paths that industries are taking in terms of managing their GHG emissions and provide insight into which industries are making progress and which ones require further attention.

6.6 The 2010-2014 Toxic Release Inventory (TRI)

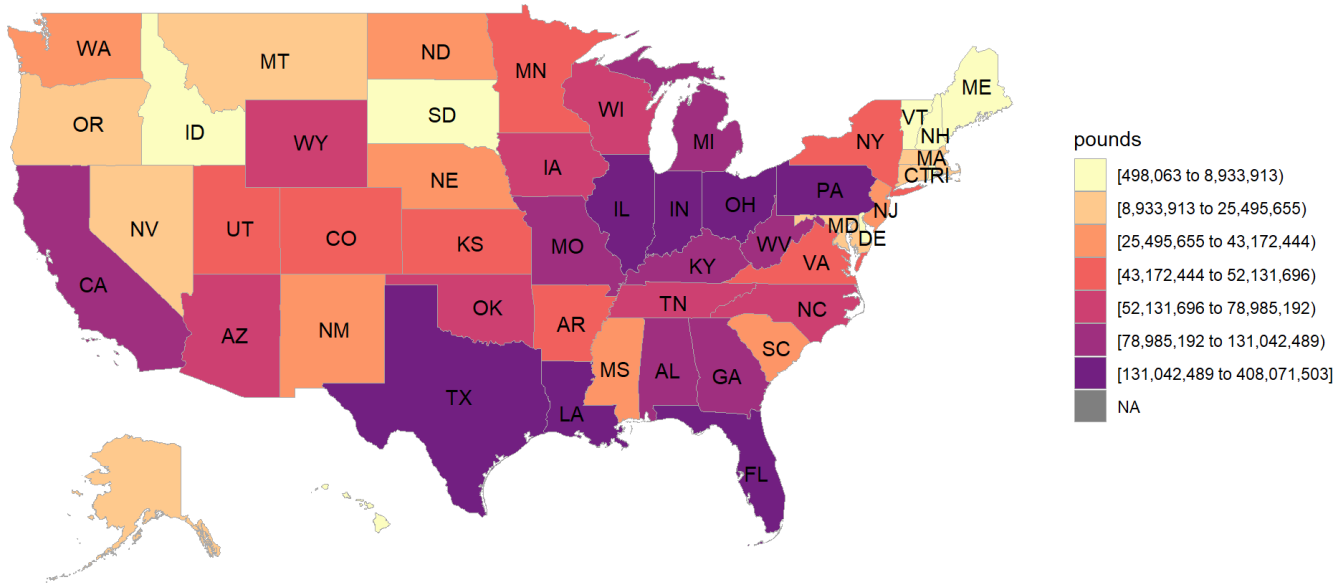
emission analysis



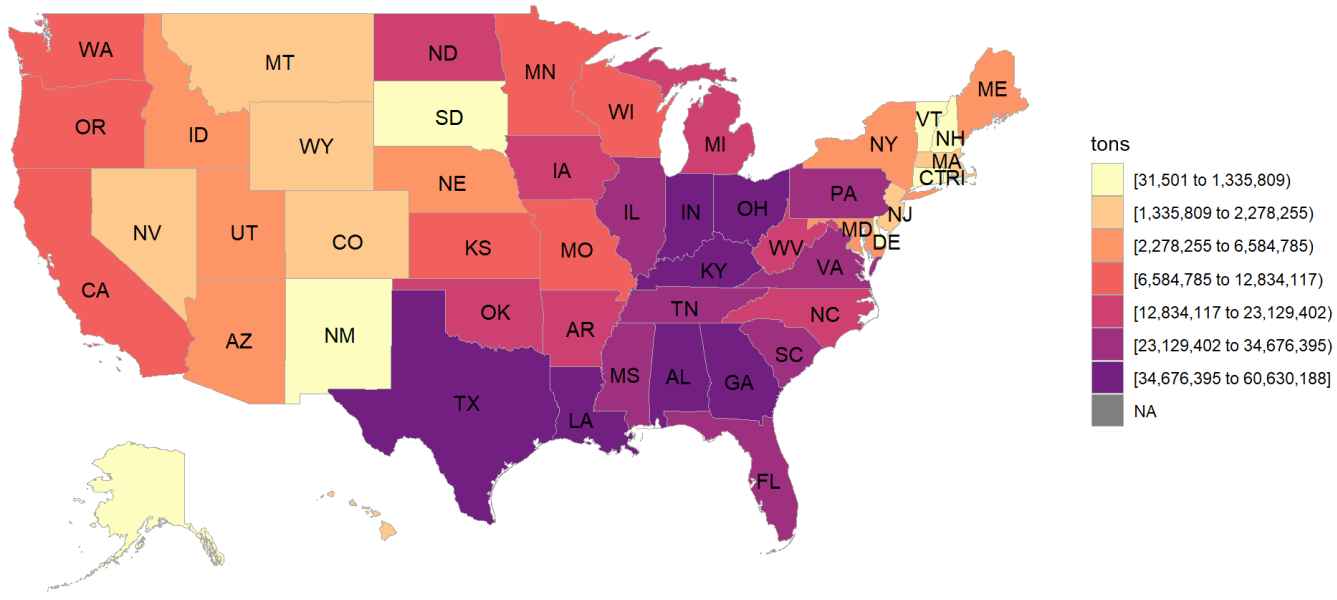
The Toxic Release Inventory (TRI) analysis reveals that Manufacturing and Utilities are the two primary industries with the highest environmental impact from 2010 to 2014. There is no sign of decrease in their impact during this period. Agriculture saw a significant decrease in its environmental impact, with its emissions dropping from 70k tons to around 2800 tons over the five years. Interestingly, Management was able to decrease its emission almost close to zero. In contrast, Construction almost doubled its emission between 2013 and 2014, indicating a significant increase in its environmental impact.

6.7 State total emission analysis

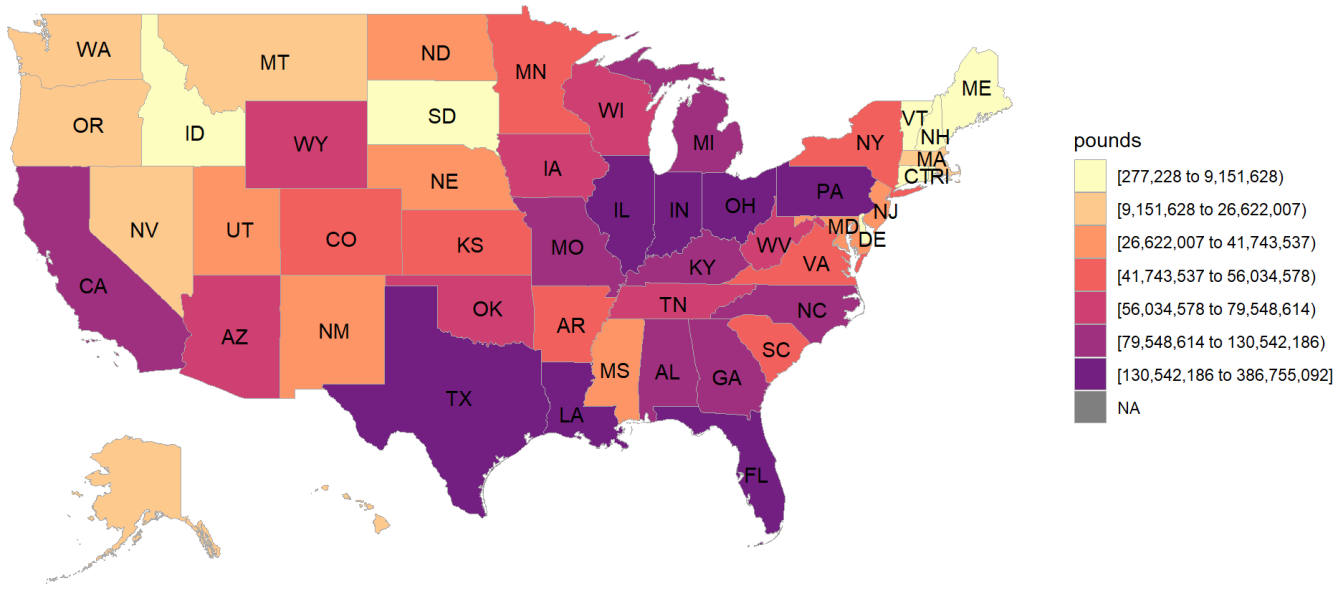
2014 GHG emission



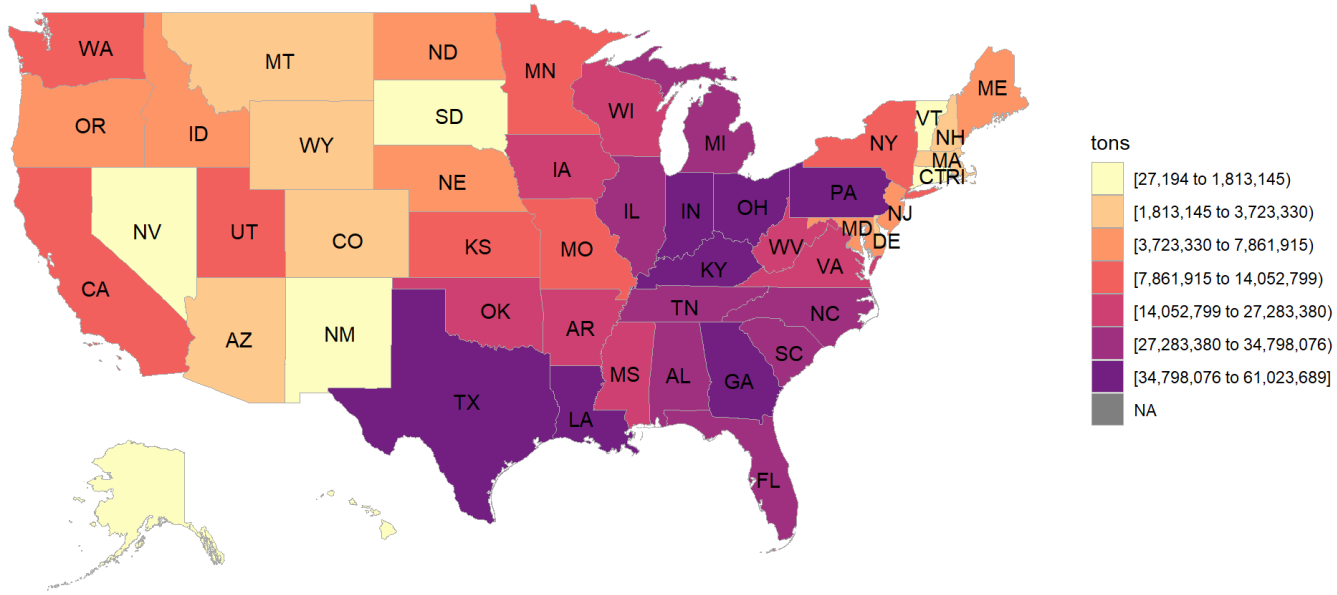
2014 TRI emission



2010 GHG emission



2010 TRI emission



7 Summary and Conclusions

In this analysis, we explored the greenhouse gas (GHG) and Toxic Release Inventory (TRI) emissions data for different industries across various US states from 2010 to 2014. The analysis revealed that Manufacturing and Utilities were the two primary industries with the highest GHG and TRI emissions, they are significantly having more emissions, and there was no sign of a decrease in their impact during this period. Agriculture saw a significant decrease in its environmental impact, with its emissions dropping from 70k tons to around 2800 tons

over the five years. Management was able to decrease its emission almost close to zero. In contrast, Construction almost doubled its emission between 2013 and 2014, indicating a significant increase in its environmental impact.

The state-level analysis showed that California and Texas were the states with the highest GHG emissions, while Ohio and Texas were the states with the highest TRI emissions in 2014. However, the analysis also revealed that some states, such as New York and Massachusetts, made significant progress in reducing their emissions. Overall, the results provide valuable insights into which industries and states are making progress and which ones require further attention to reduce their environmental impact.