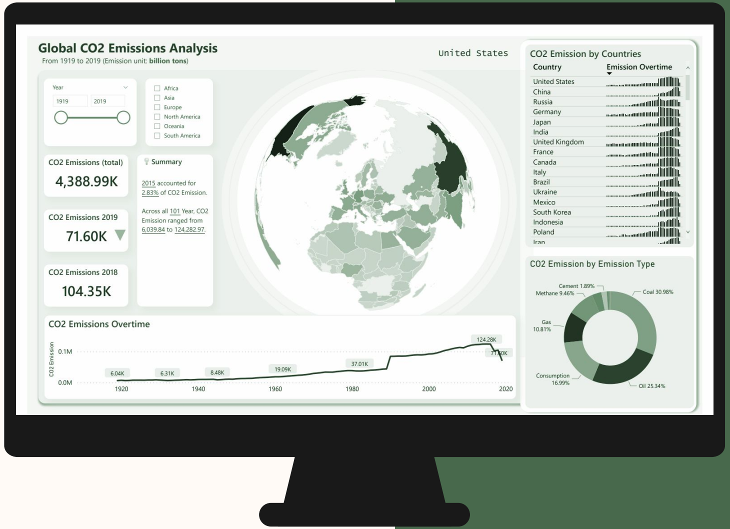
eiGO App

Environmental Info on the Go



Global CO2 Emissions Analysis

From 1919 to 2019 (Emission unit: **billion tons**)

United States

Year: 1919 – 2019

□ Africa
□ Asia
□ Europe
□ North America
□ Oceania
□ South America

CO2 Emissions (total): 4,388.99K

CO2 Emissions 2019: 71.60K

CO2 Emissions 2018: 104.35K

Summary

2015 accounted for 2.83% of CO2 Emission.

Across all 101 Year, CO2 Emission ranged from 6,039.84 to 124,282.97.

CO2 Emission by Countries

Country	Emission Overtime
United States	
China	
Russia	
Germany	
Japan	
India	
United Kingdom	
France	
Canada	
Italy	
Brazil	
Ukraine	
Mexico	
South Korea	
Indonesia	
Poland	
Iran	

Aayana Anand

Overview

The eiGO Missions

Manageably solve the Challenge of Harmful Emissions in the US

eiGO focuses are Greenhouse Gas and Toxic Substance Emissions, which vary across the country based on several varying factors, **making it difficult for an average American to keep track of their community's environmental footprint.** Additionally, **posing a large, recurring problem to an audience can be daunting.**

Despite this, reducing these types of emissions is a pressing problem that needs to be addressed. **Emissions collectively contribute to global warming and air pollution, which are detrimental to our health, agriculture, wellbeing, medical expenses, labor, and economy.**

Improve Users' Understandings of their State's Emissions

With the eiGO app's **data dashboard** personalized by location, any user can understand their community's emission reduction progress without needing to make generalizations or assumptions about their community's environmental standing.

To do this, we will begin by **showing the impacts of emissions for a select few states**, which will include **statistics for states overall as well as per state resident.**

Why does eiGO need data?

To Quantify the Impacts of Emissions

Data from the EPA, CDC, and US Census allow eiGO to quantify the impacts of emissions in a user's community by powering digestible metrics, visualizations, and insights. As it stands, information about the environment is difficult to access and fragmented – KPIs, trends, and visualizations powered by reliable data sources mitigate this problem to present emissions information tangibly for users. Having data from multiple sources with emissions AND demographics (ie. zip codes, states) will enable eiGO to draw connections and establish relationships specific to emissions in a user's own community. **Data empowers us with information, which we share with users to reinforce the importance of monitoring local emissions.**

To Aid Americans in their Sustainability Goals

A survey of 10,281 global consumers showed that 78% of individuals want to be more sustainable, and 63% of individuals have actually taken steps to do so. eiGO directly supports users in these goals to benefit their communities sustainably. **Providing valuable data and insights will aid the efforts of our existing audience (ie. by providing localized updates about estimated emissions overall and per state resident).**

BOTTOM LINE: **The eiGO app tackles the issue of reducing emissions and informing users about their local (state's) emissions by providing reliable insights on different levels (state and resident) – all driven by data sources including emissions and demographic information.**

What data will eiGO use?

Dataset	Description	Type	How Dataset will be Used
CDC National Toxic Substance Incidents Program (NTSIP) 2013-2014 Report (Dataset comes from 4th Link)	Data from the Hazardous Substance Emergency Events Surveillance (HSEES) program.	Number, Text	**To map data about States, Counties, etc. to information about Area Types and/or Substance Categories.** Connecting these pieces of information will allow us to understand which locations/communities have popular area types and/or commonly released substances. These insights will inform us and our users on what kinds of substances are commonly released in their communities and if land use/area type in their communities contribute to them.
EPA Greenhouse Gas Reporting Program (GHGRP) 2022 Data (Dataset comes from most recent file in the "2022 Data Summary Spreadsheets (zip)" download link	Data containing reported emissions by greenhouse gas and process.	Number, Text	**To map data about States, Counties, Cities, Zip codes, etc. to information about 2022 Direct Greenhouse Gas Emissions and Industry Types.** Connecting these pieces of information will allow us to understand which locations/communities have popular industry types and/or increased levels of greenhouse gas emissions. These insights will inform us and our users on what kinds of greenhouse gases are commonly released in their communities, how much of these greenhouse gases are released in their communities, and if land use/industry type in their communities contribute to them.
US Census Comparative Demographic Estimates (CP05)	US Census Data from 2018-2022 with Demographic Estimates Information.	Text	**To collect information about population demographics from several states.** We will connect these pieces of information to the information collected about substances and greenhouse gases, allowing us to further contextualize the environmental data based on the demographics of specific locations/communities. **Because all of this data is text representing percentages and population totals, this will need to be manipulated (ie. remove commas and percent signs) to be used in numeric calculations.**

The additional submitted PDFs (ghgp_faqs & ntsip_dictionary) will be used externally (not directly in the project analysis) as references, allowing us to keep track of what certain columns/fields mean in various contexts.

Chosen Dataset Information

Dataframe	Datasets used	Types of Data Collected	File Size	Dataframe Size
df_EPA_Greenhouse_Gases	EPA Dataset	Number, Text	2.1 MB	(8659, 25)
df_NTSIP_Substances	NTSIP Dataset	Number, Text	595 KB	(3131, 90)
df_Census	Census Dataset	Text	266 KB	(96, 469)
merged_environmental_df	EPA and NTSIP Datasets	Number, Text	327 MB	(894207, 60)

Original Analysis Plan

Distributions of Substance Categories

Using **merged_environmental_df,** we will see how the distribution of substance category varies across states and area types. To complete this, we will group the data frame by the **'State'** and '**AREATYP1**' columns respectively. Using these groupby objects, we will utilize the **'SUB_CAT'** column and count occurrences of the categories. To visualize these findings, we will develop pie charts reflecting the substance category distributions for each state and each area type. These substance categories are represented as numbers, so when we find the distribution of these "numbers," we will be able to reference back to the nstip_dictionary PDF to obtain further context.

Greenhouse Gas Emission Trends

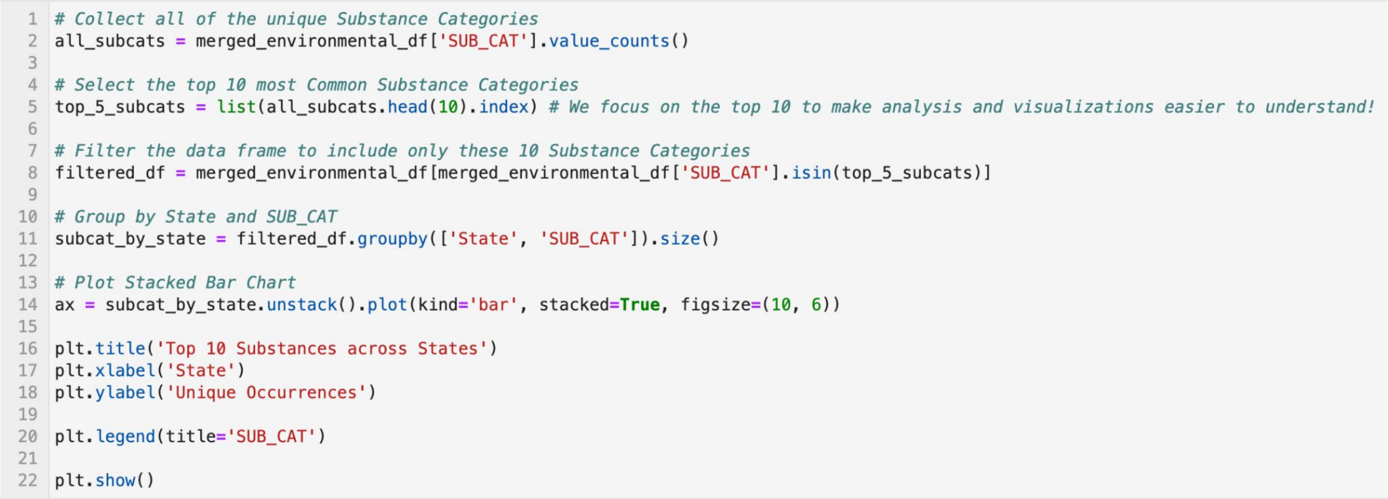
Using **merged_environmental_df**, we will see how mean greenhouse gas emissions vary across states from the years 2011-2022. To complete this, we will group the dataframe by the **'State'** and **'Year'** columns. Using this groupby object, we will utilize the '**2022 Total reported direct emissions**' column and aggregate the respective means for each year across each state. With the x-axis representing years and the y-axis representing mean greenhouse gas emissions, we will plot these findings using a multi-line plot, visualizing each state with it's own line on the same graph. Additionally, we will use **df_Census** to see how mean greenhouse gas emissions per state resident vary across states from the years 2011-2022. To do this, we will utilize the same findings from our first analysis, except we will now divide our results based on populations in those respective states. Our visualization format will also be identical, except for the y-axis now representing mean greenhouse gas emissions per resident.

Distributions of Area Types

Using **merged_environmental_df,** we will see how the distribution of area types varies across states. To complete this, we will group the dataframe by the **'State'** column. Using this groupby object, we will utilize the **'AREATYP1'** column and count occurrences of the area types. To visualize these findings, we will develop pie charts reflecting the area type distributions for each state. These area types are represented as numbers/letters, so when we find the distribution of these "numbers/letters," we will be able to reference back to the nstip_dictionary PDF to obtain further context.

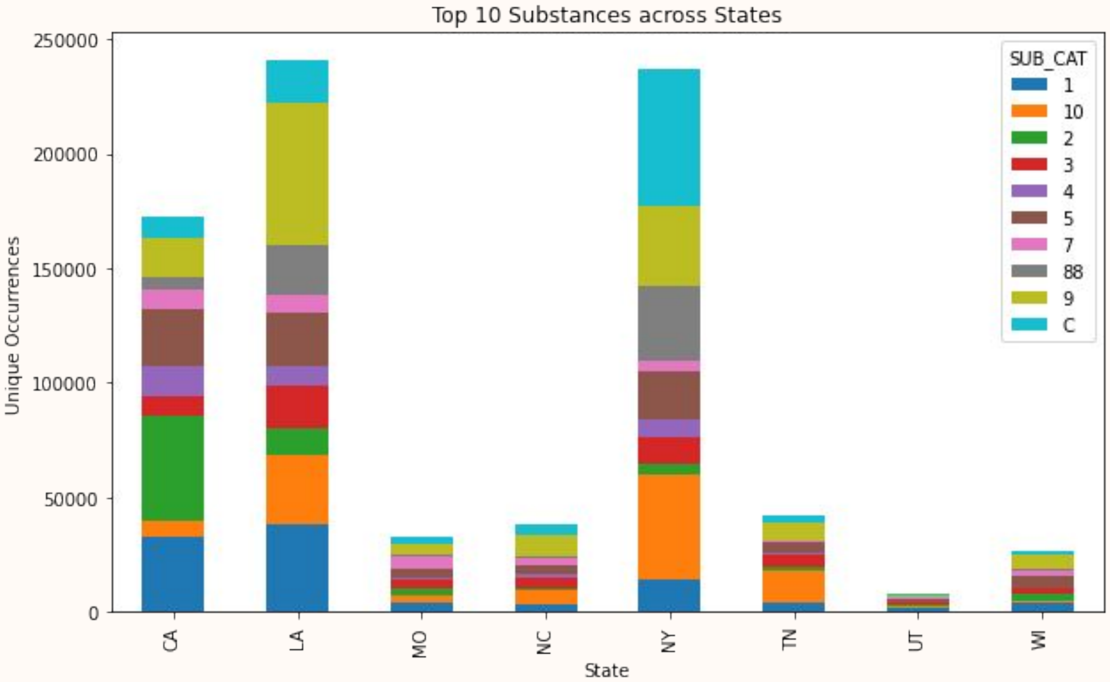
Completed Analysis & Takeaways

Distributions of Substance Categories across States (*Process*)



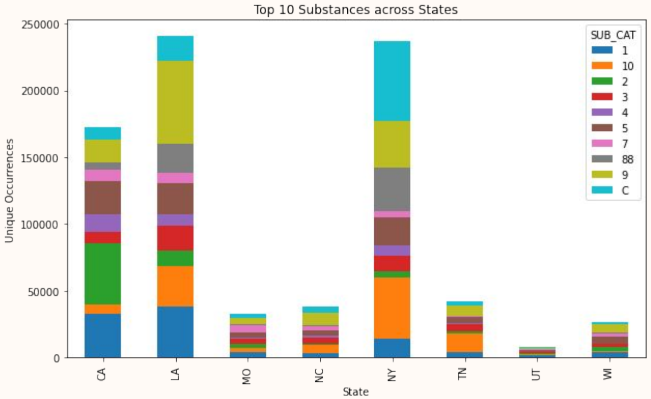
```python
# Collect all of the unique Substance Categories
all_subcats = merged_environmental_df['SUB_CAT'].value_counts()

# Select the top 10 most Common Substance Categories
top_5_subcats = list(all_subcats.head(10).index) # We focus on the top 10 to make analysis and visualizations easier to understand!

# Filter the data frame to include only these 10 Substance Categories
filtered_df = merged_environmental_df[merged_environmental_df['SUB_CAT'].isin(top_5_subcats)]

# Group by State and SUB_CAT
subcat_by_state = filtered_df.groupby(['State', 'SUB_CAT']).size()

# Plot Stacked Bar Chart
ax = subcat_by_state.unstack().plot(kind='bar', stacked=True, figsize=(10, 6))

plt.title('Top 10 Substances across States')
plt.xlabel('State')
plt.ylabel('Unique Occurrences')

plt.legend(title='SUB_CAT')

plt.show()
```

Distributions of Substance Categories across States (*Visualization*)



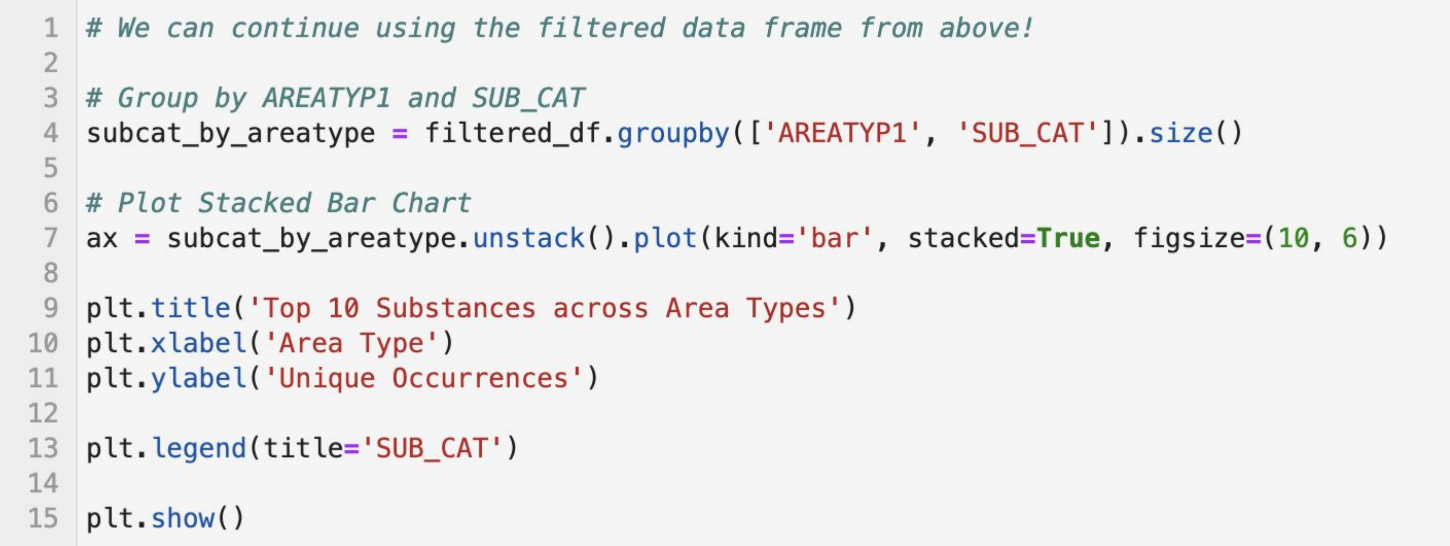
Distributions of Substance Categories across States (*Takeaways*)

- States with the highest frequency of substance-related incidents were California, Louisiana, and New York
  - These frequencies were significantly higher than the other 5 states in the data
- Each state reflects a wide variety of substance-related incidents
  - There is too much variation to make a conclusion about whether a certain substance is more prevalent in a given state/region or not



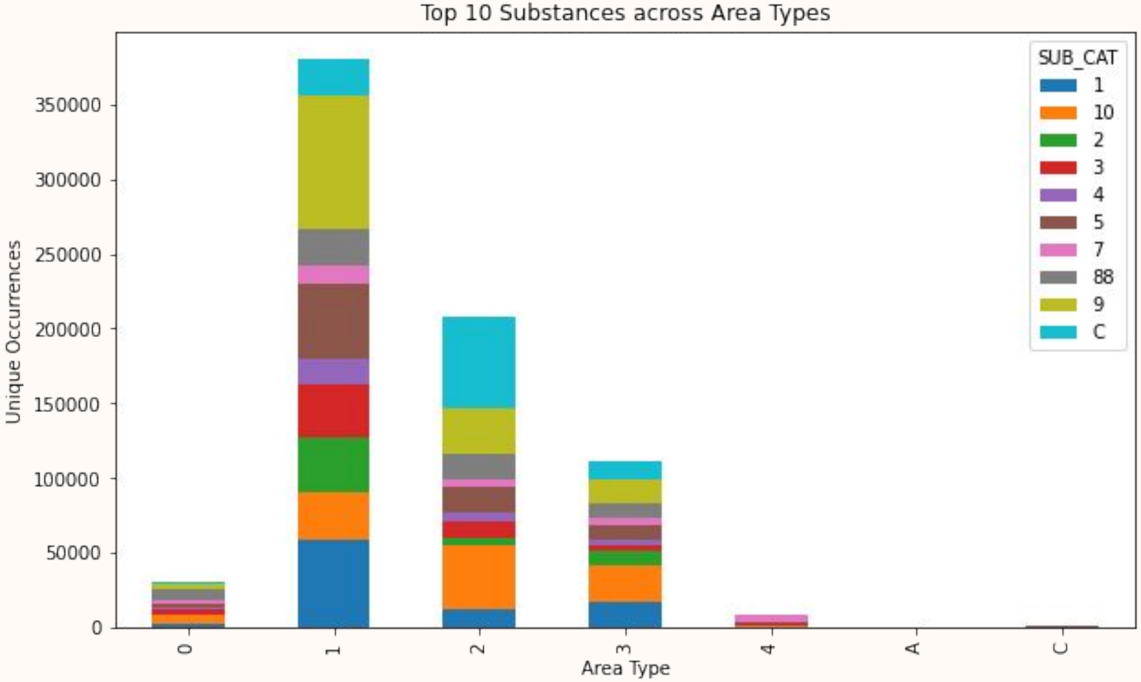
Based on these findings, we conclude that there does not appear to be a significant relationship between substance categories and states. We should conduct more analysis in the future to confirm this.

Distributions of Substance Categories across Area Types (*Process*)



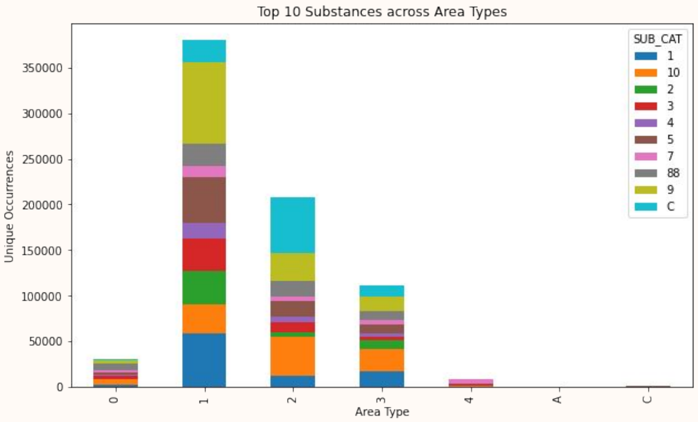
```python
# We can continue using the filtered data frame from above!

# Group by AREATYP1 and SUB_CAT
subcat_by_areatype = filtered_df.groupby(['AREATYP1', 'SUB_CAT']).size()

# Plot Stacked Bar Chart
ax = subcat_by_areatype.unstack().plot(kind='bar', stacked=True, figsize=(10, 6))

plt.title('Top 10 Substances across Area Types')
plt.xlabel('Area Type')
plt.ylabel('Unique Occurrences')

plt.legend(title='SUB_CAT')

plt.show()
```

Distributions of Substance Categories across Area Types (*Visualization*)



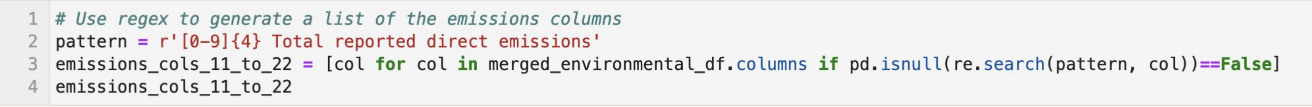
Distributions of Substance Categories across Area Types (*Takeaways*)

- The most common area type reported was type 1 (Industrial), followed by type 2 (Commercial) and type 3 (Residential)
    - In other words, these are the 3 area types where substance-related incidents occurred the most, regardless of the states they occurred in
- Each area type reflects a wide variety of substance-related incidents
    - There is too much variation to make a conclusion about whether a certain substance is more prevalent in a given area type or not

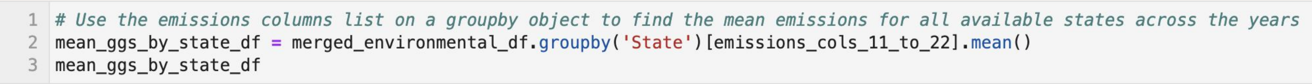


Based on these findings, we conclude that the most prevalent area types with substance-related incidents were Industrial, Commercial, and Residential. There does not appear to be a significant relationship between substance categories and area types. We should conduct more analysis in the future to confirm this.

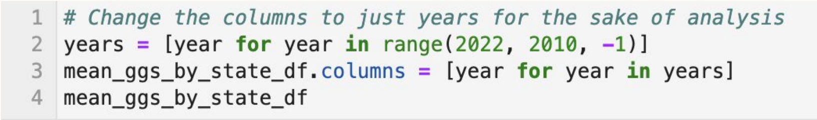
Mean Greenhouse Gas Emissions across States from 2011-2022 (*Process*)



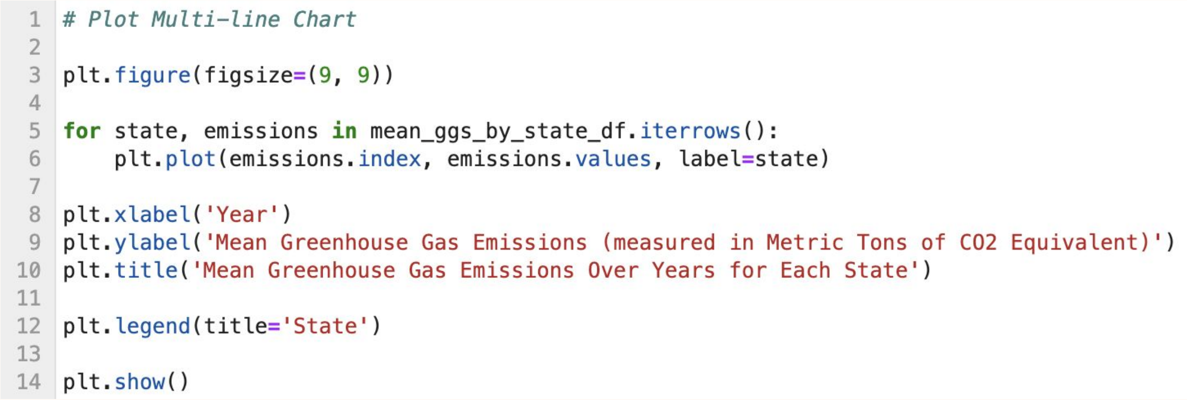
```python
# Use regex to generate a list of the emissions columns
pattern = r'[0-9]{4} Total reported direct emissions'
emissions_cols_11_to_22 = [col for col in merged_environmental_df.columns if pd.isnull(re.search(pattern, col))==False]
emissions_cols_11_to_22
```



```python
# Use the emissions columns list on a groupby object to find the mean emissions for all available states across the years
mean_ggs_by_state_df = merged_environmental_df.groupby('State')[emissions_cols_11_to_22].mean()
mean_ggs_by_state_df
```

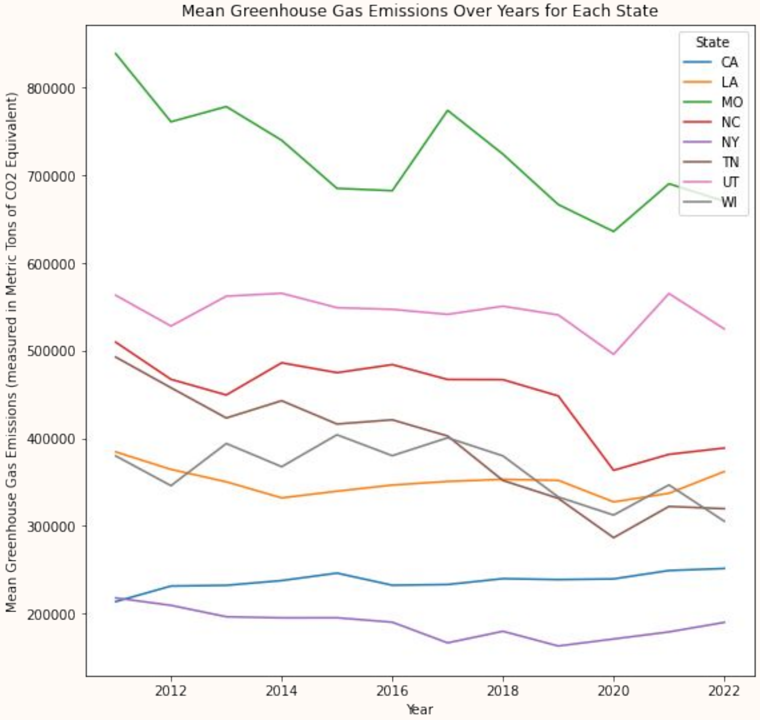


```python
# Change the columns to just years for the sake of analysis
years = [year for year in range(2022, 2010, -1)]
mean_ggs_by_state_df.columns = [year for year in years]
mean_ggs_by_state_df
```



```python
# Plot Multi-line Chart

plt.figure(figsize=(9, 9))

for state, emissions in mean_ggs_by_state_df.iterrows():
    plt.plot(emissions.index, emissions.values, label=state)

plt.xlabel('Year')
plt.ylabel('Mean Greenhouse Gas Emissions (measured in Metric Tons of CO2 Equivalent)')
plt.title('Mean Greenhouse Gas Emissions Over Years for Each State')

plt.legend(title='State')

plt.show()
```

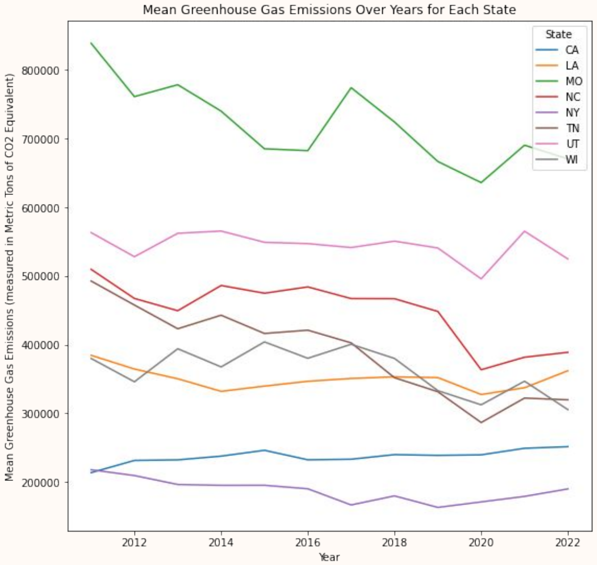
Mean Greenhouse Gas Emissions across States from 2011-2022 (*Visualization*)



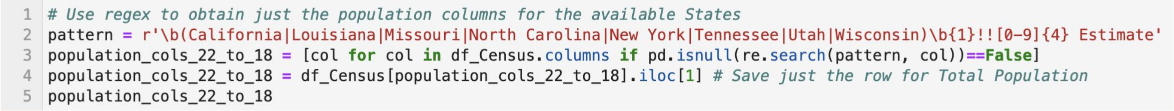
Mean Greenhouse Gas Emissions across States from 2011-2022 (*Takeaways*)

- States with the highest mean greenhouse gas emissions were Missouri and Utah, with the lowest mean greenhouse emissions being New York and California
- All states experienced a general decrease in mean greenhouse gas emissions over a 10 year period, with some states having slight increases between 2021 and 2022

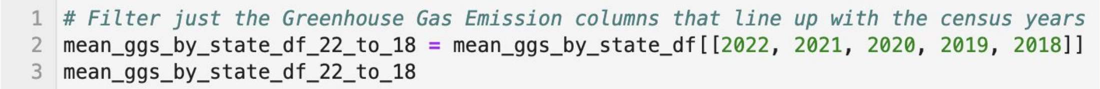
Based on these findings, we conclude that Missouri and Utah maintained the highest greenhouse gas emissions on average, even over a 10 year period. Conversely, New York and California had the lowest emissions. Despite the decrease in greenhouse gas emissions over time, North Carolina, Louisiana, California, and Wisconsin had slight increases in emissions between 2021 and 2022. Some states experienced slight increases in 2021, including Missouri, Utah, Wisconsin, and Tennessee.



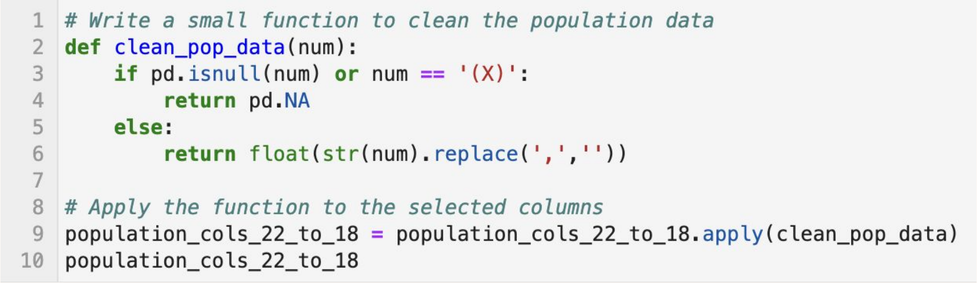
Mean Greenhouse Gas Emissions per State Resident across States from 2011-2022 (*Process pt. 1*)



```python
# Use regex to obtain just the population columns for the available States
pattern = r'\b(California|Louisiana|Missouri|North Carolina|New York|Tennessee|Utah|Wisconsin)\b{1}!![0-9]{4} Estimate'
population_cols_22_to_18 = [col for col in df_Census.columns if pd.isnull(re.search(pattern, col))==False]
population_cols_22_to_18 = df_Census[population_cols_22_to_18].iloc[1] # Save just the row for Total Population
population_cols_22_to_18
```

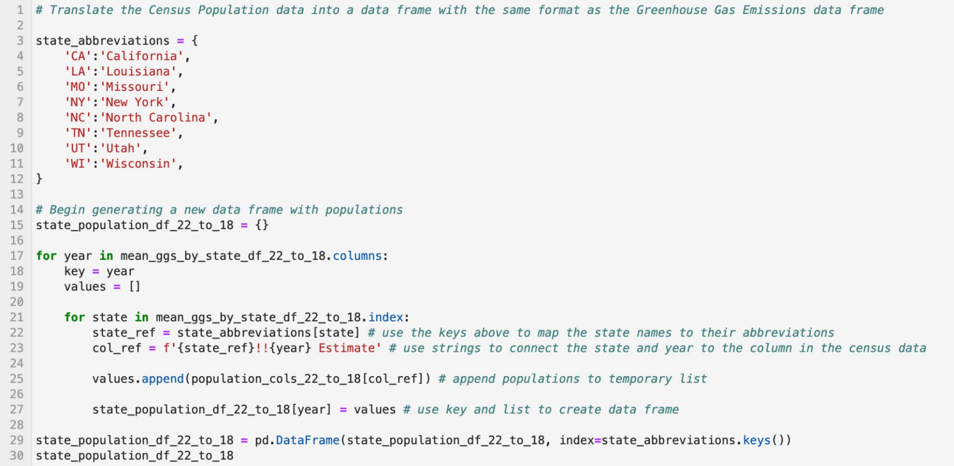


```python
# Filter just the Greenhouse Gas Emission columns that line up with the census years
mean_ggs_by_state_df_22_to_18 = mean_ggs_by_state_df[[2022, 2021, 2020, 2019, 2018]]
mean_ggs_by_state_df_22_to_18
```

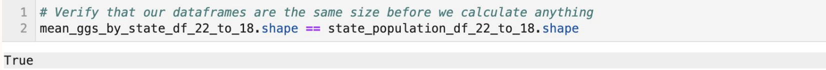


```python
# Write a small function to clean the population data
def clean_pop_data(num):
    if pd.isnull(num) or num == '(X)':
        return pd.NA
    else:
        return float(str(num).replace(',',''))

# Apply the function to the selected columns
population_cols_22_to_18 = population_cols_22_to_18.apply(clean_pop_data)
population_cols_22_to_18
```

Mean Greenhouse Gas Emissions per State Resident across States from 2011-2022 (*Process pt. 2*)

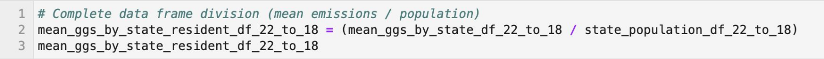


```python
# Translate the Census Population data into a data frame with the same format as the Greenhouse Gas Emissions data frame

state_abbreviations = {
    'CA':'California',
    'LA':'Louisiana',
    'MO':'Missouri',
    'NY':'New York',
    'NC':'North Carolina',
    'TN':'Tennessee',
    'UT':'Utah',
    'WI':'Wisconsin',
}

# Begin generating a new data frame with populations
state_population_df_22_to_18 = {}

for year in mean_ggs_by_state_df_22_to_18.columns:
    key = year
    values = []

    for state in mean_ggs_by_state_df_22_to_18.index:
        state_ref = state_abbreviations[state] # use the keys above to map the state names to their abbreviations
        col_ref = f'{state_ref}!!{year} Estimate' # use strings to connect the state and year to the column in the census data

        values.append(population_cols_22_to_18[col_ref]) # append populations to temporary list

        state_population_df_22_to_18[year] = values # use key and list to create data frame

state_population_df_22_to_18 = pd.DataFrame(state_population_df_22_to_18, index=state_abbreviations.keys())
state_population_df_22_to_18
```



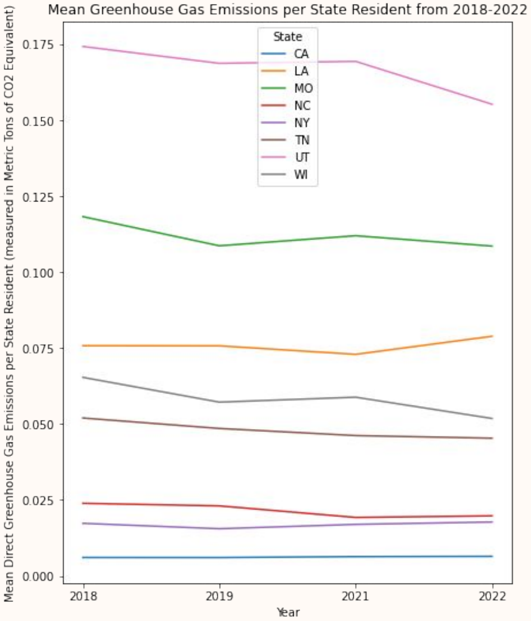
```python
# Verify that our dataframes are the same size before we calculate anything
mean_ggs_by_state_df_22_to_18.shape == state_population_df_22_to_18.shape
```

```
True
```



```python
# Complete data frame division (mean emissions / population)
mean_ggs_by_state_resident_df_22_to_18 = (mean_ggs_by_state_df_22_to_18 / state_population_df_22_to_18)
mean_ggs_by_state_resident_df_22_to_18
```

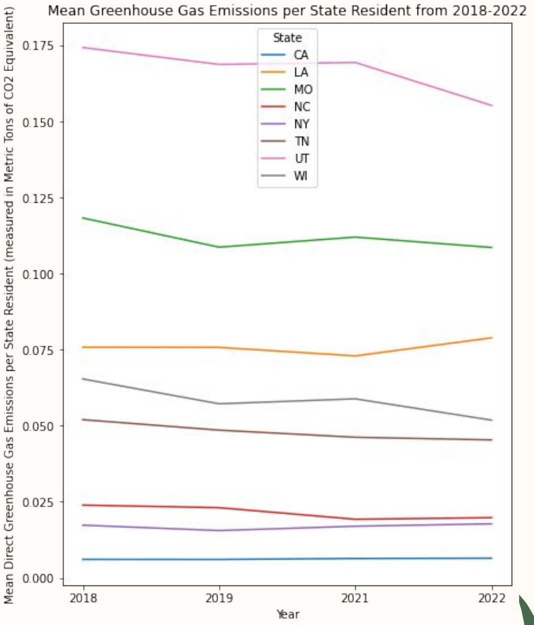
Mean Greenhouse Gas Emissions per State Resident across States from 2011-2022 (*Visualization*)



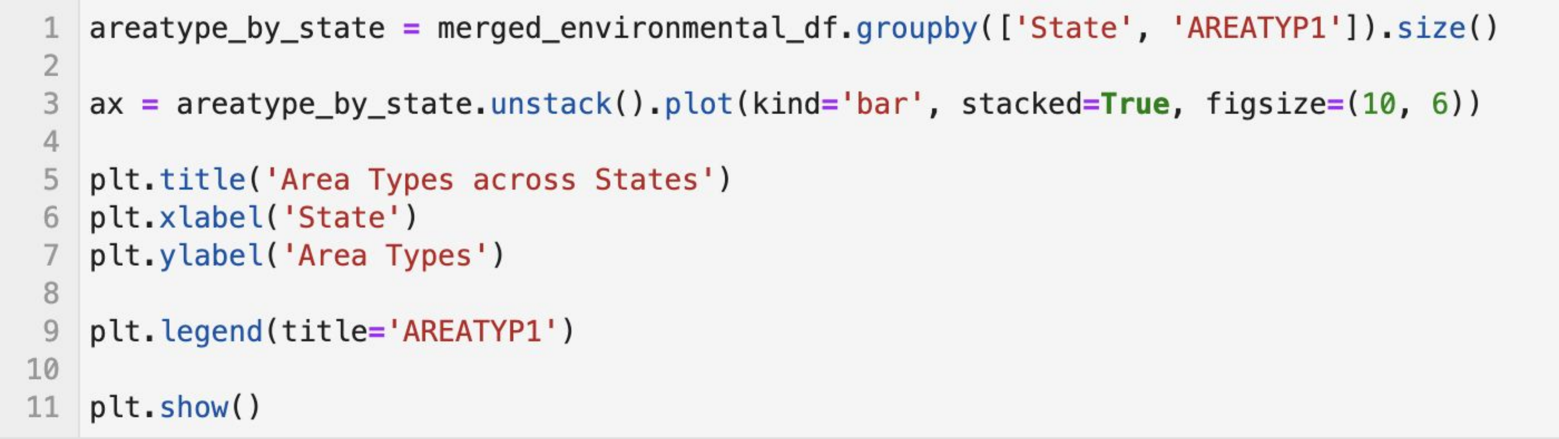
Mean Greenhouse Gas Emissions per State Resident across States from 2011-2022 (*Takeaways*)

- States with the highest mean greenhouse gas emissions per resident were Utah and Missouri, with the lowest mean greenhouse emissions per resident being California and New York
- All states experienced a slight decrease in mean greenhouse gas emissions per resident over a 10 year period, with Louisiana having a slight increase around 2022

Based on these findings, we conclude that Utah and Missouri maintained the highest greenhouse gas emissions per resident on average, even over a 10 year period. Conversely, California and New York had the lowest emissions per resident. Despite the slight decrease in greenhouse gas emissions per resident over time, almost all states experienced little to no change in the impact of greenhouse gas emissions on the resident/individual level, even though emissions overall for each state had significant downward changes.

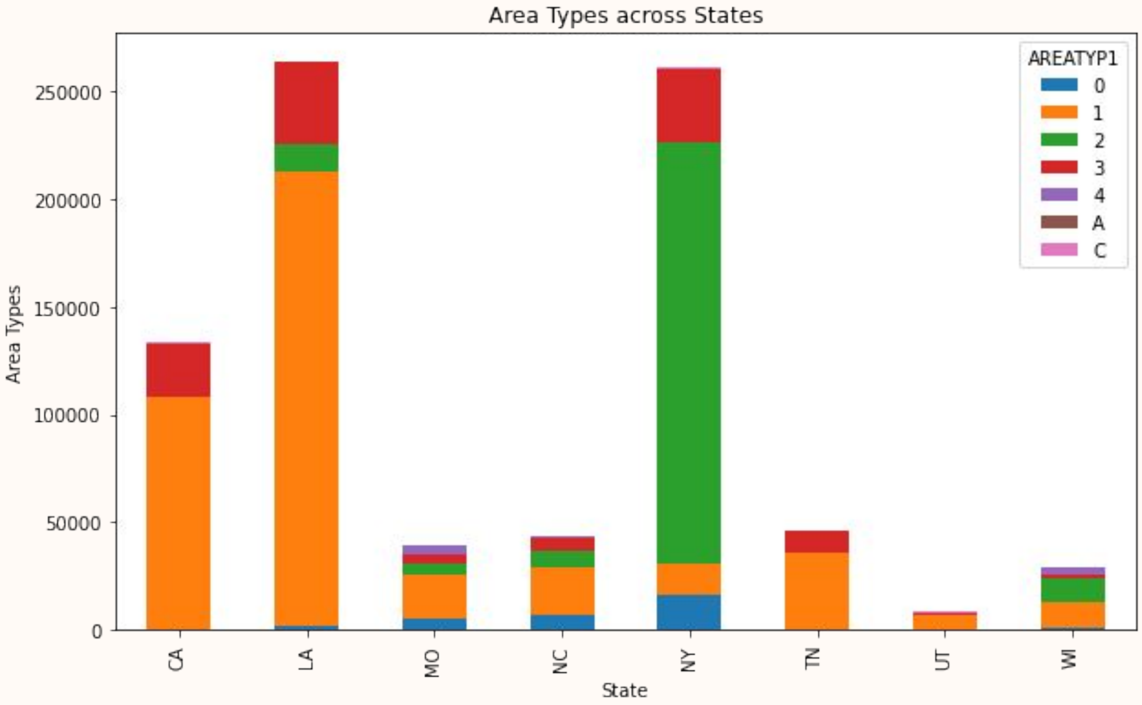


Distributions of Area Types across States (*Process*)



```python
areatype_by_state = merged_environmental_df.groupby(['State', 'AREATYP1']).size()

ax = areatype_by_state.unstack().plot(kind='bar', stacked=True, figsize=(10, 6))

plt.title('Area Types across States')
plt.xlabel('State')
plt.ylabel('Area Types')

plt.legend(title='AREATYP1')

plt.show()
```

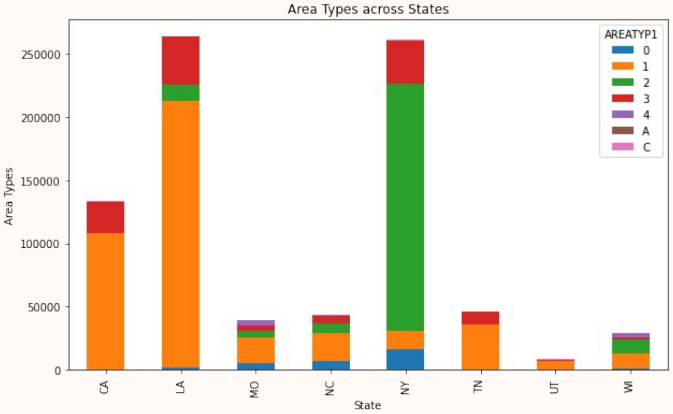
Distributions of Area Types across States (*Visualization*)



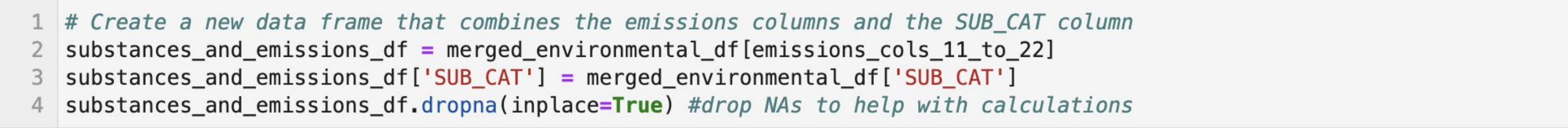
Distributions of Area Types across States (*Takeaways*)

- All of the states contain majority type 1 (Industrial) area type, with the exception of New York which contains a majority type 2 (Commercial) area type
- Louisiana, New York, and California have the highest frequency of area types

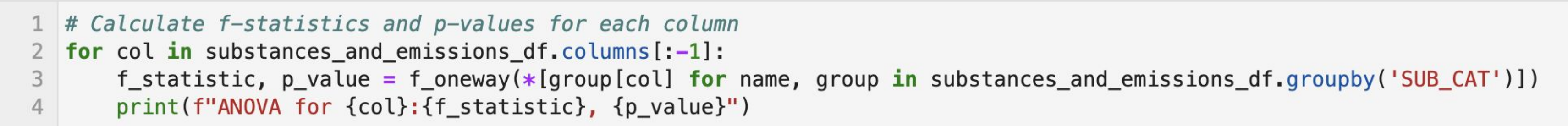
Based on these findings, we conclude that Industrial area types are most common across most states, with the exception of New York (Commercial). We can infer a correlation/attribution between these area types and greenhouse gas/toxic substance emissions for each state respectively.



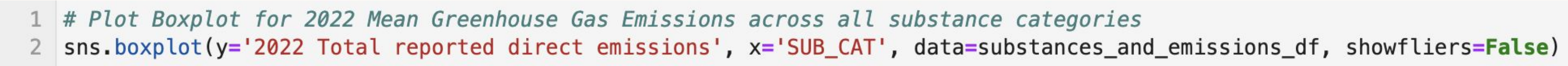
Relationship between Substances and Greenhouse Gas Emissions (*Process*)



```python
# Create a new data frame that combines the emissions columns and the SUB_CAT column
substances_and_emissions_df = merged_environmental_df[emissions_cols_11_to_22]
substances_and_emissions_df['SUB_CAT'] = merged_environmental_df['SUB_CAT']
substances_and_emissions_df.dropna(inplace=True) #drop NAs to help with calculations
```

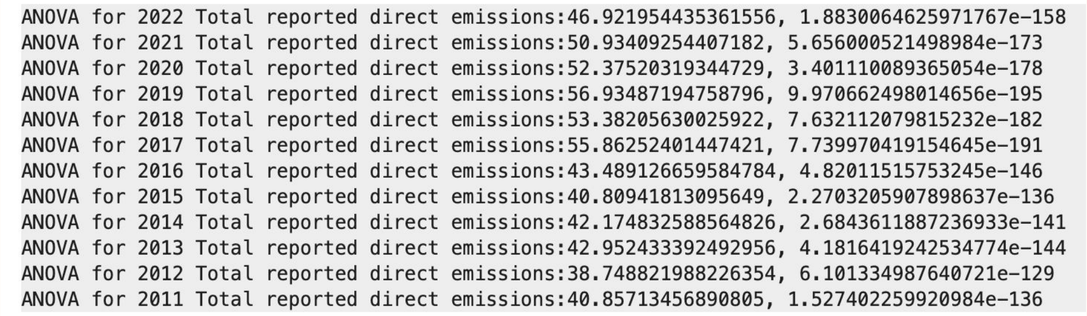


```python
# Calculate f-statistics and p-values for each column
for col in substances_and_emissions_df.columns[:-1]:
    f_statistic, p_value = f_oneway(*[group[col] for name, group in substances_and_emissions_df.groupby('SUB_CAT')])
    print(f"ANOVA for {col}:{f_statistic}, {p_value}")
```

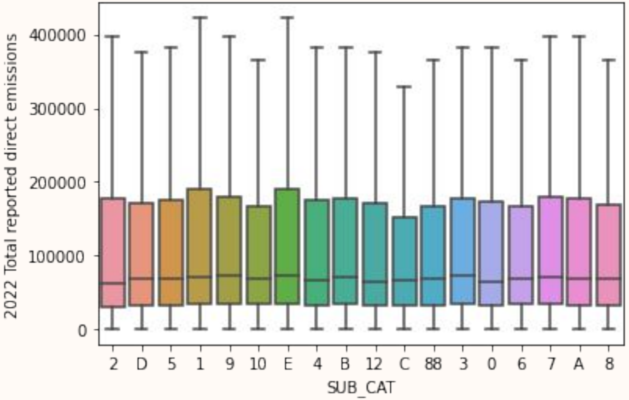


```python
# Plot Boxplot for 2022 Mean Greenhouse Gas Emissions across all substance categories
sns.boxplot(y='2022 Total reported direct emissions', x='SUB_CAT', data=substances_and_emissions_df, showfliers=False)
```

Relationship between Substances and Greenhouse Gas Emissions (*Visualization*)



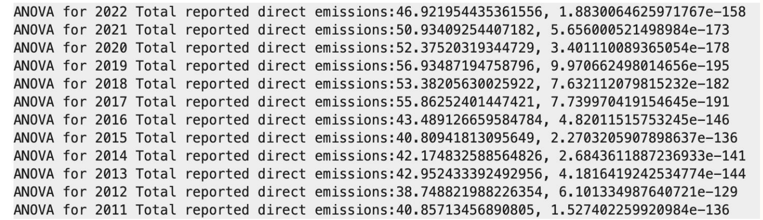
```
ANOVA for 2022 Total reported direct emissions:46.921954435361556, 1.8830064625971767e-158
ANOVA for 2021 Total reported direct emissions:50.93409254407182, 5.656000521498984e-173
ANOVA for 2020 Total reported direct emissions:52.37520319344729, 3.401110089365054e-178
ANOVA for 2019 Total reported direct emissions:56.93487194758796, 9.970662498014656e-195
ANOVA for 2018 Total reported direct emissions:53.38205630025922, 7.632112079815232e-182
ANOVA for 2017 Total reported direct emissions:55.86252401447421, 7.739970419154645e-191
ANOVA for 2016 Total reported direct emissions:43.489126659584784, 4.82011515753245e-146
ANOVA for 2015 Total reported direct emissions:40.80941813095649, 2.2703205907898637e-136
ANOVA for 2014 Total reported direct emissions:42.174832588564826, 2.6843611887236933e-141
ANOVA for 2013 Total reported direct emissions:42.952433392492956, 4.1816419242534774e-144
ANOVA for 2012 Total reported direct emissions:38.748821988226354, 6.101334987640721e-129
ANOVA for 2011 Total reported direct emissions:40.85713456890805, 1.527402259920984e-136
```



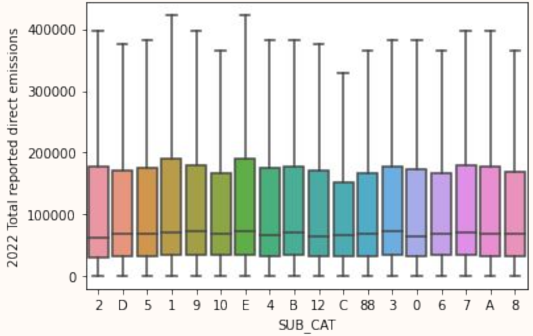
Relationship between Substances and Greenhouse Gas Emissions (*Takeaways*)

- The F-Statistics for all years (evaluating across all substance categories) were high
- The P-Values for all years (evaluating across all substance categories) were low
- The box plots representing mean greenhouse gas emissions stayed relatively consistent across substance categories, with respect to the middle 50% of the data
    - ○ Higher maxima corresponded to substance categories of 1 (Acid) and 4 (Chlorine)

Based on these findings, we conclude that there is a significant correlation between Mean Total Direct Emissions and Substance Categories for all Years. We know this because of the supporting F-Statistics and P-Values, which reject the null hypothesis (ie. support the notion that a valid relationship exists between these variables). This informs us that toxic substances and greenhouse gases go hand-in-hand when informing users about their state's emission contributions.



```
ANOVA for 2022 Total reported direct emissions:46.921954435361556, 1.8830064625971767e-158
ANOVA for 2021 Total reported direct emissions:50.93409254407182, 5.656000521498984e-173
ANOVA for 2020 Total reported direct emissions:52.37520319344729, 3.401110089365054e-178
ANOVA for 2019 Total reported direct emissions:56.93487194758796, 9.970662498014656e-195
ANOVA for 2018 Total reported direct emissions:53.38205630025922, 7.632112079815232e-182
ANOVA for 2017 Total reported direct emissions:55.86252401447421, 7.739970419154645e-191
ANOVA for 2016 Total reported direct emissions:43.489126659584784, 4.82011515753245e-146
ANOVA for 2015 Total reported direct emissions:40.80941813095649, 2.2703205907898637e-136
ANOVA for 2014 Total reported direct emissions:42.174832588564826, 2.6843611887236933e-141
ANOVA for 2013 Total reported direct emissions:42.952433392492956, 4.1816419242534774e-144
ANOVA for 2012 Total reported direct emissions:38.748821988226354, 6.101334987640721e-129
ANOVA for 2011 Total reported direct emissions:40.85713456890805, 1.527402259920984e-136
```



Conclusion

Major Analysis Takeaways

- **No significant relationship between substances and area types or between substances and states were detected.** With this being said, toxic substance emissions and greenhouse gas emissions were proven to have a strong relationship and correlation.
- **Certain area types (Industrial, Commercial, Residential) are more susceptible to substance-related incidents in general.**
- **For all states captured by the data, mean greenhouse gas emissions reduced overall in the last 10 years.** Interestingly however, mean greenhouse gas emissions per resident (with respect to each state) did not seen much change over the same period.
- **Certain states have consistently high mean greenhouse gas emissions overall and per resident**, including Utah and Missouri. On the other hand, certain states have consistently low mean greenhouse gas emissions overall and per resident, including California and New York.
- **All states in the data are dominated by Industrial area types, with the exception of New York and Commercial area types.** This means that for these states respectively, we can infer a correlation (not necessarily causation) between the predominant area type of a state and its greenhouse gas/toxic substance emission source(s).
- **Mean greenhouse gas emissions stayed relatively consistent across substance categories**, although higher-reaching ranges of greenhouse gas emissions corresponded to acid and chlorine substances.

Why You Should Care about this Analysis

Room for Growth with Individual Impacts

The stagnation of greenhouse gas emissions per state resident reflect a gap in the efforts of individual residents to reduce emissions in their respective states. Presenting these findings on our app demonstrates the room for growth in this area, showing that the downward trend in overall greenhouse gas emissions is not as attributed to individual changes/efforts. **This finding should motivate you and our users to take action and begin a downward trend in individual impact on emissions.**

Relevant, Localized Insights

Current information about the environment is general and fragmented. However, This analysis communicates takeaways and visualizations to make all of this information more tangible and applicable (currently on the State level). Newly accessible and digestible, **this analysis should encourage you and our users to understand where your state falls in terms of emissions and if you as a state resident are helping continue the downward trend of emissions.**

Future Analysis

Capture More State Data

- Expanding data sources to include more states across the country
- Equalizing data so that each state has more similar counts of data points
- Collecting geospatial data and experimenting with geographic visualizations

Continue Investigating Substance Data

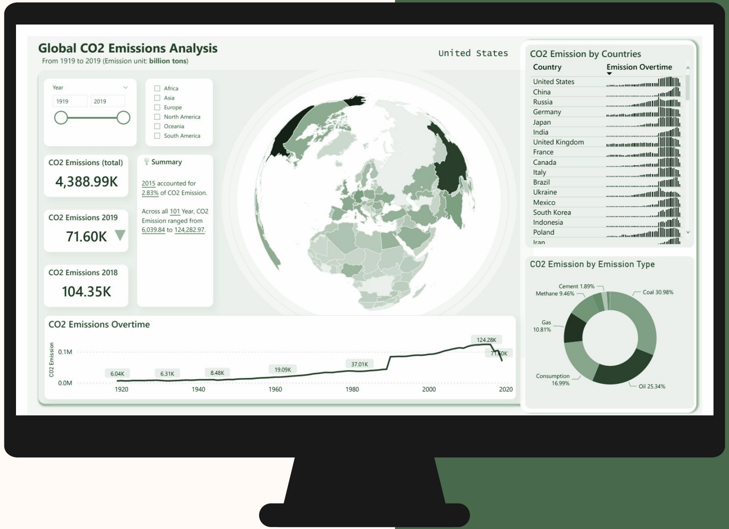
- Our analysis didn't find as many relationships/correlations corresponding to toxic substance emissions
- We could include new variables to see if they provide more context to the changes in toxic substance emissions over time

Continue Investigating Demographic Data

- The CDC Data contains several demographic pieces including race/ethnicity, age, sex, etc.
- Examining this information could allow us to establish more meaningful relationships to emissions.

eiGO App

Environmental Info on the Go



Global CO2 Emissions Analysis

From 1919 to 2019 (Emission unit: **billion tons**)

United States

Year: 1919 – 2019

□ Africa
□ Asia
□ Europe
□ North America
□ Oceania
□ South America

CO2 Emissions (total): 4,388.99K

CO2 Emissions 2019: 71.60K ▼

CO2 Emissions 2018: 104.35K

Summary

2015 accounted for 2.83% of CO2 Emission.

Across all 101 Year, CO2 Emission ranged from 6,039.84 to 124,282.97.

CO2 Emissions Overtime

CO2 Emission by Countries

Country	Emission Overtime
United States	
China	
Russia	
Germany	
Japan	
India	
United Kingdom	
France	
Canada	
Italy	
Brazil	
Ukraine	
Mexico	
South Korea	
Indonesia	
Poland	
Iran	

CO2 Emission by Emission Type

Thank You