Air Pollution + Crop Yields

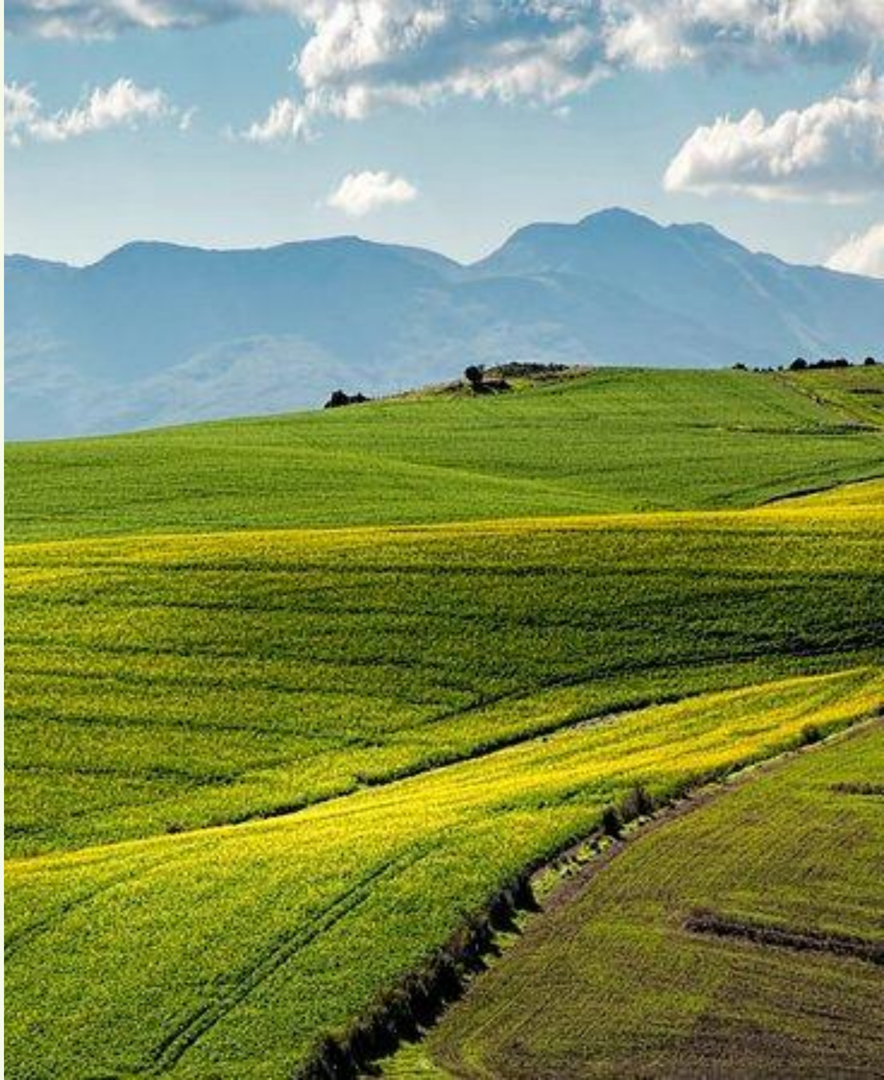
Project 3 - Data Engineering

Presented By The Data Ninjas

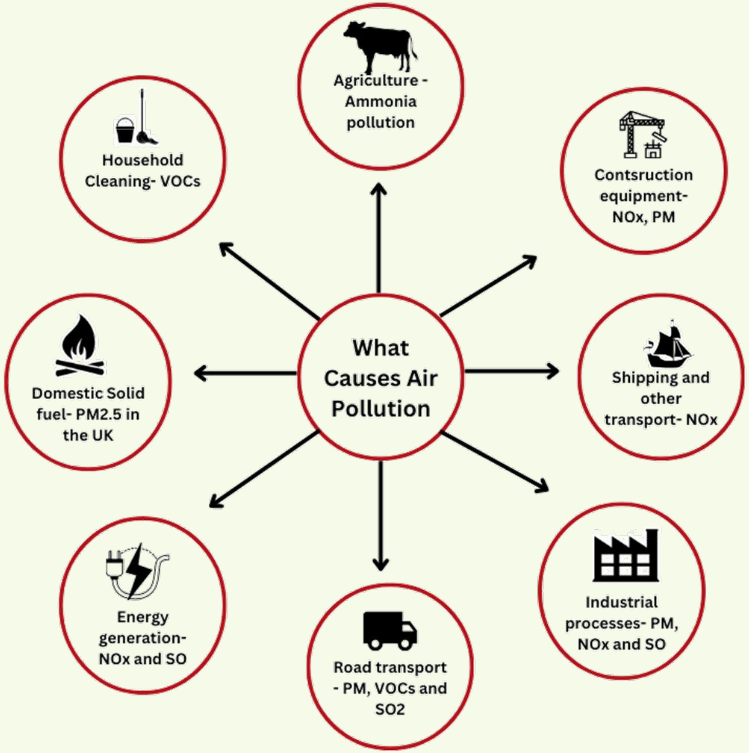
4.18.24

Crop Yields

Agriculture crops can be injured when exposed to high concentrations of various pollutants. Pollutants show visible markings on foliage reducing growth and ultimately damaging the crop.



Air Pollution Sources



What Causes Air Pollution

Agriculture - Ammonia pollution

Contsruction equipment- NOx, PM

Shipping and other transport- NOx

Industrial processes- PM, NOx and SO

Road transport - PM, VOCs and SO2

Energy generation- NOx and SO

Domestic Solid fuel- PM2.5 in the UK

Household Cleaning- VOCs

ClearIAS.com

Nitrogen Oxide

“Research led by David Lobell finds that areas around the globe with high amounts of nitrogen oxides pollution see significant declines in crop yields. This shows that reducing nitrogen oxides pollution is not only good for the climate and for health but also for food security.” -Standford News

Clean Air Reform

American Farm Bureau Federation is requesting a database that contains nitrogen oxide levels by county in the U.S. dating back 30 years.

1. What are the air pollution trends by each U.S. state and county over the last 30 years?
2. What are the nitrogen oxide level trends for each U.S. state and county over the last 30 years?
3. Which U.S. counties have a high level of nitrogen oxide in 2023?

Understanding trends and where the highest concentration of nitrogen oxide levels exist will empower the American Farm Bureau Federation and the local farmers to advocate clean air reform.

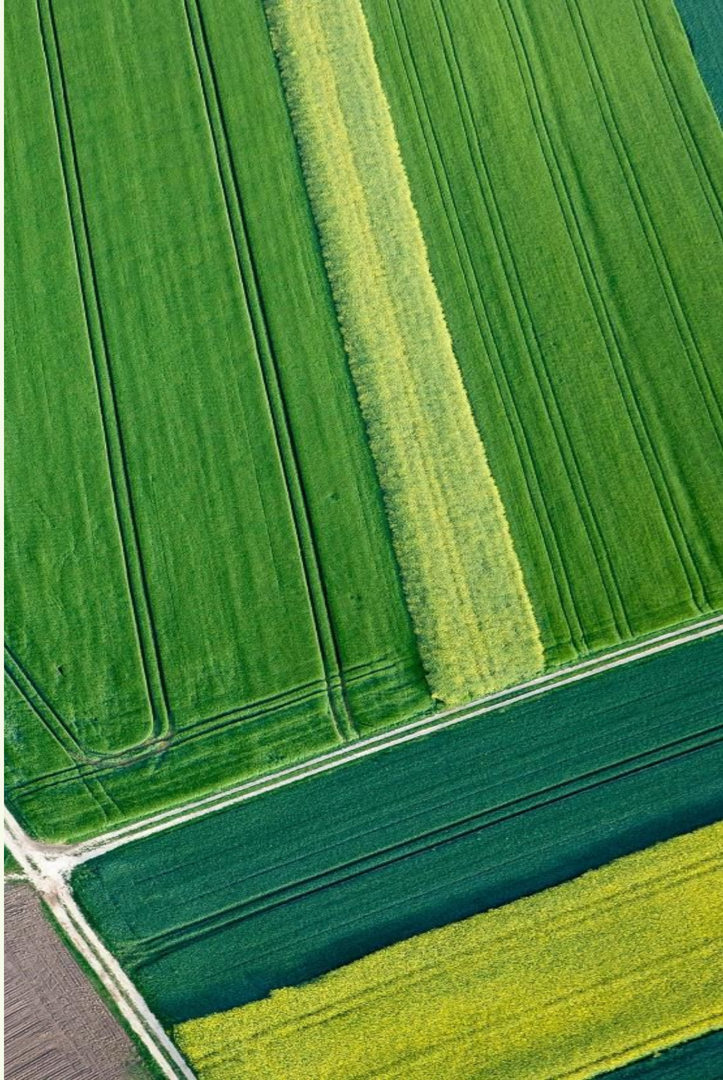
Ethical Considerations

Transparency

Clearly convey how and where the data was collected, how it will be used and ultimately shared.

Privacy

Ensure participants have guaranteed anonymity.



ETL PROCESS

Data Collected and Sources

Extract

Transform

Load

United States Environmental Protection Agency (EPA)

Air Data: Air Quality Data Collected at Outdoor Monitors Across the US

30 years of air pollution data across all US counties

https://aqs.epa.gov/aqsweb/airdata/download_files.html

Farm Service Agency (FSA) U.S. Department Of Agriculture

Crop Acreage Data

11 years of farm count within each US county (2013 - 2023)

https://www.fsa.usda.gov/news-room/efoia/electronic-reading-room/frequently-requested-information/crop-acreage-data/index

Assumptions

Extract

- 30 years of data used from Environmental Protection Agency
- 11 years of farm count data was applied to the database from 2013-2023 due to lack of data availability

Transform

- Overlapping datasets are assumed to be accurate, e.g. Farm Service Agency
- Data for Puerto Rico and Virgin Islands excluded from DB
- PostgreSQL utilized to construct DB

Load

-No assumptions to consider

ERD

61 CSVs

3 Tables

2,366,774 Total Number of Rows

Extract

Transform

Load

Annual AQI By County

PK: County, State, Year

Annual AQI By Monitor

PK: Index

FK: County, State, Year

Total Farms By County

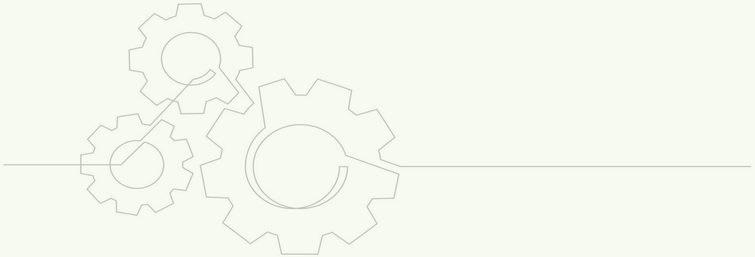
PK: County, State, Year

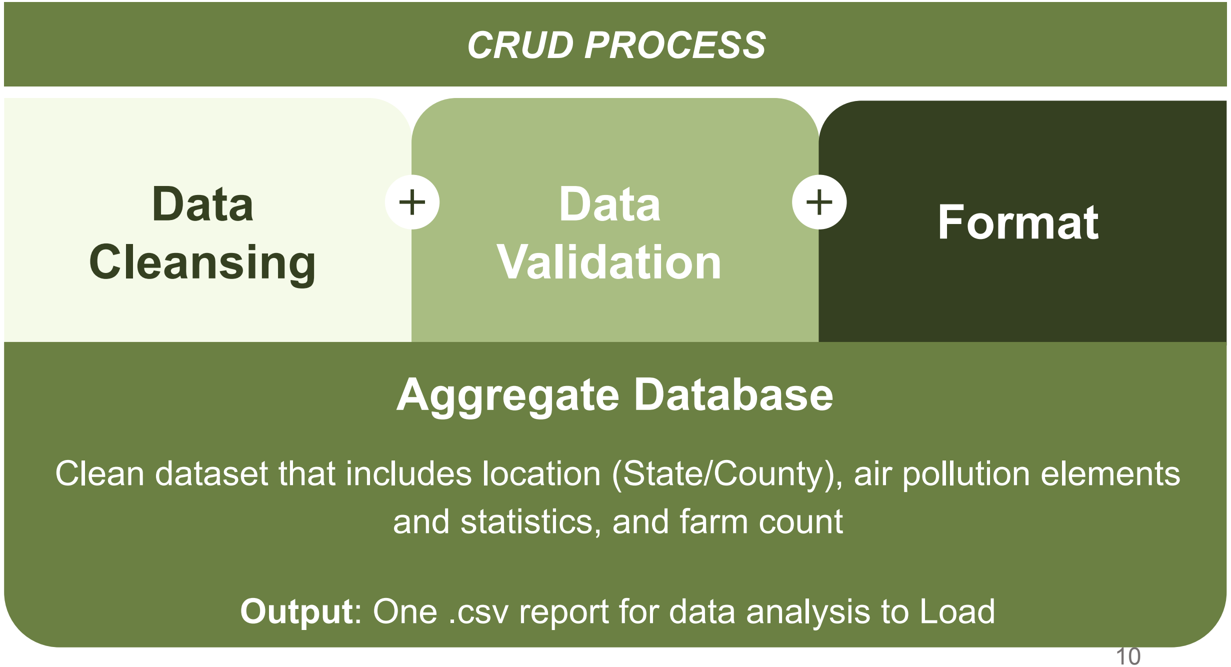
SQL

Extract

Transform

Load





CRUD PROCESS

Data Cleansing + **Data Validation** + **Format**

Aggregate Database

Clean dataset that includes location (State/County), air pollution elements and statistics, and farm count

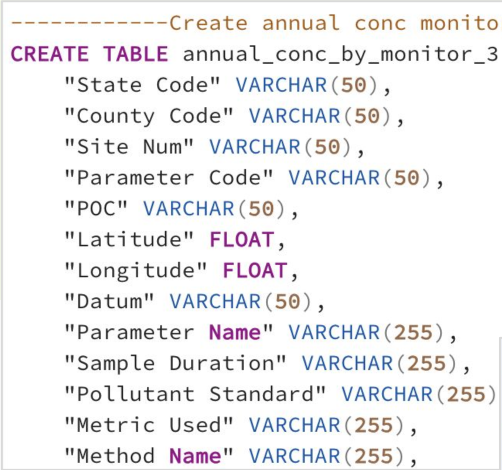
Output: One .csv report for data analysis to Load

SQL

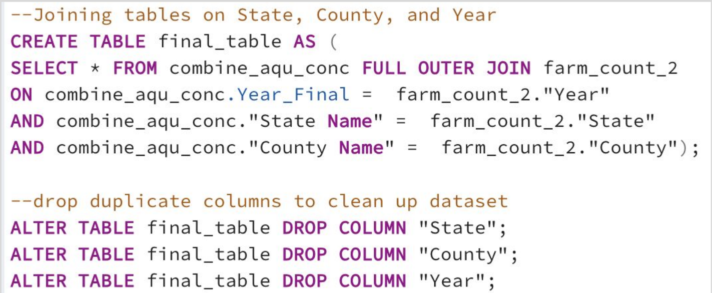
Extract

Transform

Load

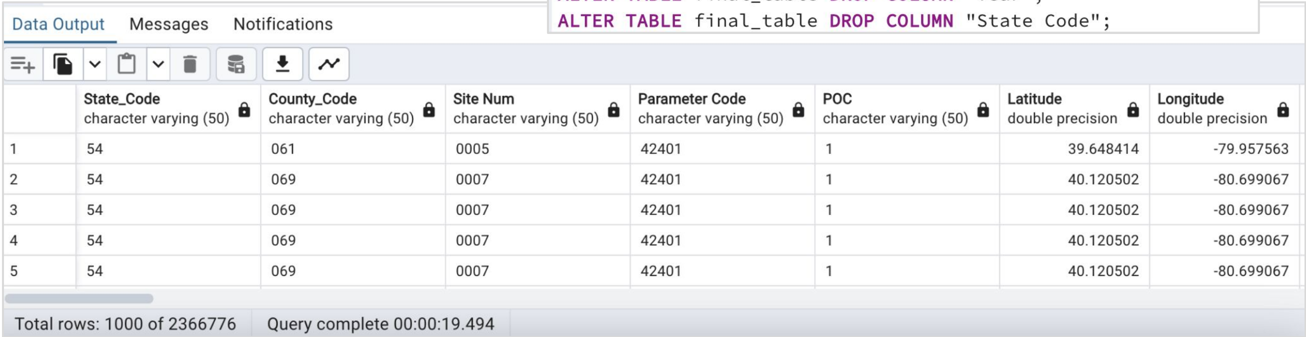


```
------------Create annual conc monitor table,
CREATE TABLE annual_conc_by_monitor_3 (
    "State Code" VARCHAR(50),
    "County Code" VARCHAR(50),
    "Site Num" VARCHAR(50),
    "Parameter Code" VARCHAR(50),
    "POC" VARCHAR(50),
    "Latitude" FLOAT,
    "Longitude" FLOAT,
    "Datum" VARCHAR(50),
    "Parameter Name" VARCHAR(255),
    "Sample Duration" VARCHAR(255),
    "Pollutant Standard" VARCHAR(255)
    "Metric Used" VARCHAR(255),
    "Method Name" VARCHAR(255),
```



```
--Joining tables on State, County, and Year
CREATE TABLE final_table AS (
SELECT * FROM combine_aqu_conc FULL OUTER JOIN farm_count_2
ON combine_aqu_conc.Year_Final =  farm_count_2."Year"
AND combine_aqu_conc."State Name" =  farm_count_2."State"
AND combine_aqu_conc."County Name" =  farm_count_2."County");

--drop duplicate columns to clean up dataset
ALTER TABLE final_table DROP COLUMN "State";
ALTER TABLE final_table DROP COLUMN "County";
ALTER TABLE final_table DROP COLUMN "Year";
ALTER TABLE final_table DROP COLUMN "State Code";
```



	State_Code character varying (50)	County_Code character varying (50)	Site Num character varying (50)	Parameter Code character varying (50)	POC character varying (50)	Latitude double precision	Longitude double precision
1	54	061	0005	42401	1	39.648414	-79.957563
2	54	069	0007	42401	1	40.120502	-80.699067
3	54	069	0007	42401	1	40.120502	-80.699067
4	54	069	0007	42401	1	40.120502	-80.699067
5	54	069	0007	42401	1	40.120502	-80.699067

Total rows: 1000 of 2366776 Query complete 00:00:19.494

Python

Extract

Transform

Load

MODIN

Pandas Based Library

for large datasets

Three Tables | *2,366,774 Total Rows*

Annual AQI By County

Annual AQI By Monitor

Total Farms By County

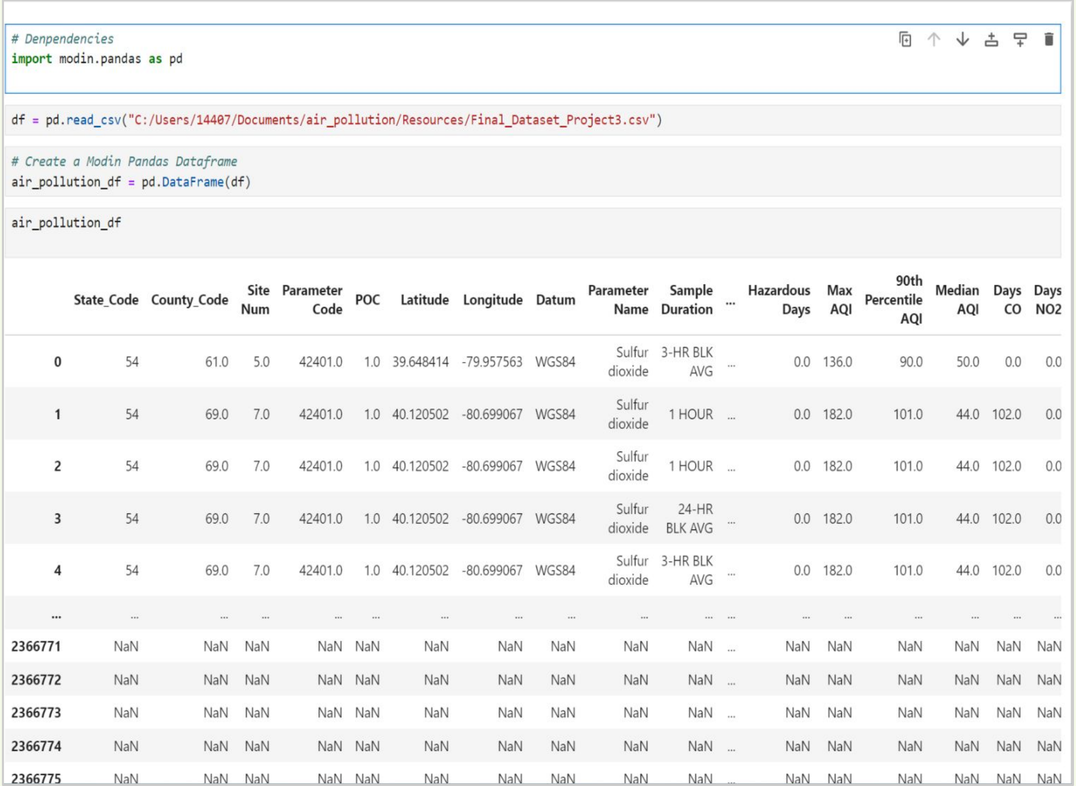
Python

Extract

Transform

Load

DataFrame



```
# Denpendencies
import modin.pandas as pd
```

```
df = pd.read_csv("C:/Users/14407/Documents/air_pollution/Resources/Final_Dataset_Project3.csv")
```

```
# Create a Modin Pandas Dataframe
air_pollution_df = pd.DataFrame(df)
```

```
air_pollution_df
```

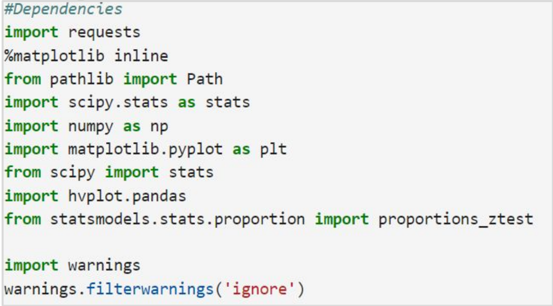
	State_Code	County_Code	Site Num	Parameter Code	POC	Latitude	Longitude	Datum	Parameter Name	Sample Duration	...	Hazardous Days	Max AQI	90th Percentile AQI	Median AQI	Days CO	Days NO2
0	54	61.0	5.0	42401.0	1.0	39.648414	-79.957563	WGS84	Sulfur dioxide	3-HR BLK AVG	...	0.0	136.0	90.0	50.0	0.0	0.0
1	54	69.0	7.0	42401.0	1.0	40.120502	-80.699067	WGS84	Sulfur dioxide	1 HOUR	...	0.0	182.0	101.0	44.0	102.0	0.0
2	54	69.0	7.0	42401.0	1.0	40.120502	-80.699067	WGS84	Sulfur dioxide	1 HOUR	...	0.0	182.0	101.0	44.0	102.0	0.0
3	54	69.0	7.0	42401.0	1.0	40.120502	-80.699067	WGS84	Sulfur dioxide	24-HR BLK AVG	...	0.0	182.0	101.0	44.0	102.0	0.0
4	54	69.0	7.0	42401.0	1.0	40.120502	-80.699067	WGS84	Sulfur dioxide	3-HR BLK AVG	...	0.0	182.0	101.0	44.0	102.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2366771	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN
2366772	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN
2366773	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN
2366774	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN
2366775	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN

Graphs

Maps

Charts

Stats



```
#Dependencies
import requests
%matplotlib inline
from pathlib import Path
import scipy.stats as stats
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats
import hvplot.pandas
from statsmodels.stats.proportion import proportions_ztest

import warnings
warnings.filterwarnings('ignore')
```

What are the air pollution trends by each U.S. state and county over the last 30 years?

What are the nitrogen oxide level trends for each U.S. state and county over the last 30 years?

Which U.S. counties have a high level of nitrogen oxide in 2023?

Resources

CLEARIAS. (2024). Air Pollution: Types, Causes, and Effects. Retrieved from https://www.clearias.com/air-pollution/

Farm Service Agency (FSA) U.S. Department Of Agriculture. Crop Acreage Data. Retrieved from https://www.fsa.usda.gov/news-room/efoia/electronic-reading-room/frequently-requested-information/crop-acreage-data/index

Jordan, R. (2022). Stanford News. Less air pollution leads to higher crop yields, Stanford-led study shows. Retrieved from https://news.stanford.edu/2022/06/01/pollution-and-crops/#:~:text=Research%20led%20by%20David%20Lobell,but%20also%20for%20food%20security

United States Environmental Protection Agency (EPA). Air Data: Air Quality Data Collected at Outdoor Monitors Across the US. Retrieved from https://aqs.epa.gov/aqsweb/airdata/download_files.html

The End