Visualizing Two Decades of U.S. Air Pollution (2000–2023)

Course: STAT 663 – Statistical Graphics and Data Visualization

Team: Siyao Huang & Minxi Li

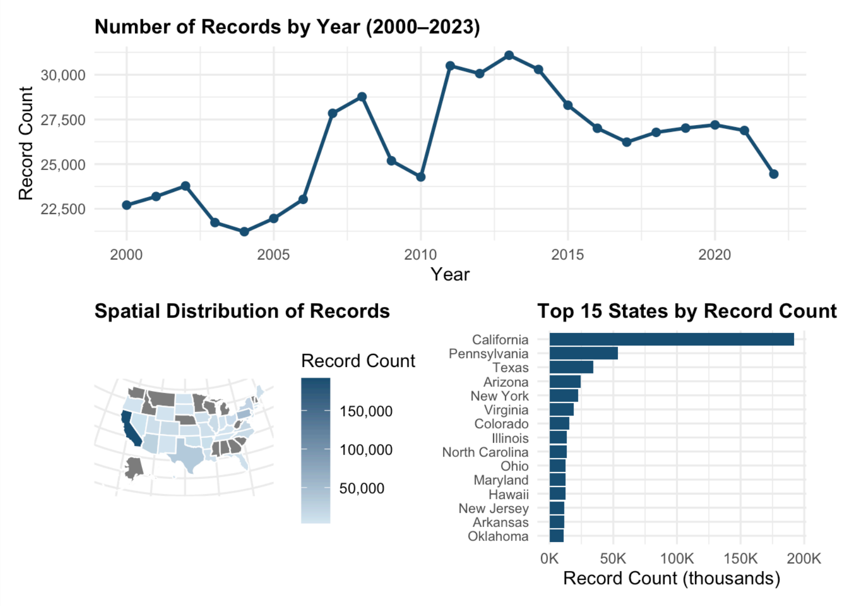
Dataset

- Data come from the U.S. EPA Air Quality System (AQS).

  AQS data are collected by state and local air monitoring stations across the United States. Each monitoring station continuously measures ambient concentrations of O_3, NO_2, SO_2, and CO. Daily pollutant measurements are reported to AQS and aggregated into state-level annual averages for this study.

- Includes four major pollutants:
  - **O_3** (Ozone)
  - **CO** (Carbon Monoxide)
  - **SO_2** (Sulfur Dioxide)
  - **NO_2** (Nitrogen Dioxide)

Dataset



Number of Records by Year (2000–2023)

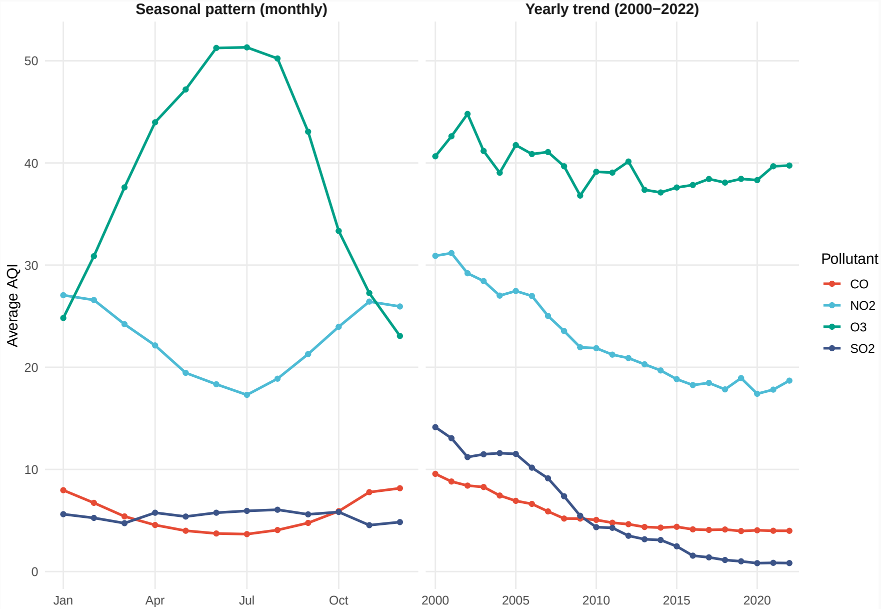
Spatial Distribution of Records

Top 15 States by Record Count

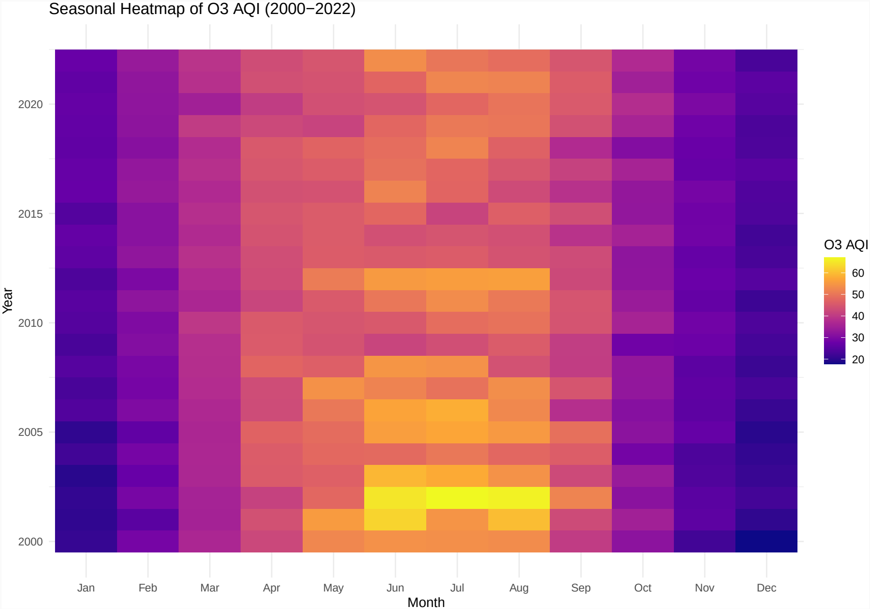
Research Question & Why It Matters

- **Main Research Question**:
  - How has U.S. air quality changed from 2000 to 2023 across time, regions, and pollutants?
- **Sub-questions**:
  - How do pollutant levels change over time?
  - Which states are more polluted or cleaner?
  - How do the four pollutants differ in behavior?
- **Why This Problem Matters**
  - Air pollution is directly linked to respiratory and cardiovascular risks.
  - Long-term pollution trends help evaluate: Environmental policies (e.g., Clean Air Act), Regional inequalities and Public-health impacts.
  - Understanding historical patterns supports better future planning and policy design.

Temporal analysis

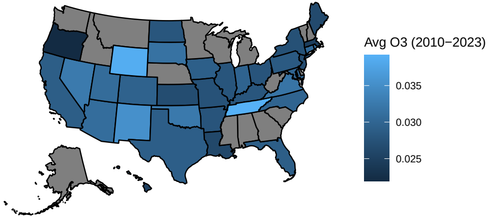


o3

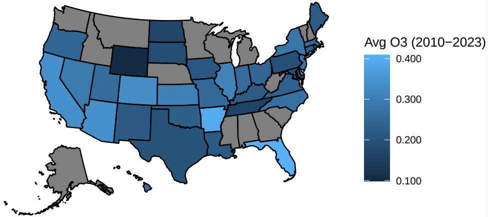

Seasonal Heatmap of O3 AQI (2000−2022)
Year
2000
2005
2010
2015
2020
Month
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
O3 AQI
60
50
40
30
20

State-Level Pollution Patterns (2010–2023)

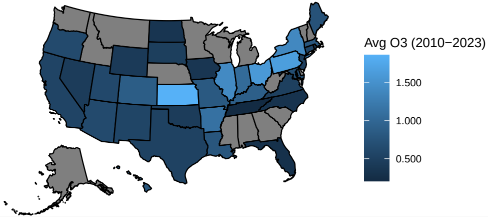
Average O3 Levels by State (2010−2023)



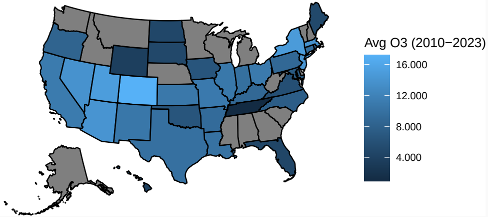
Average CO Levels by State (2010−2023)



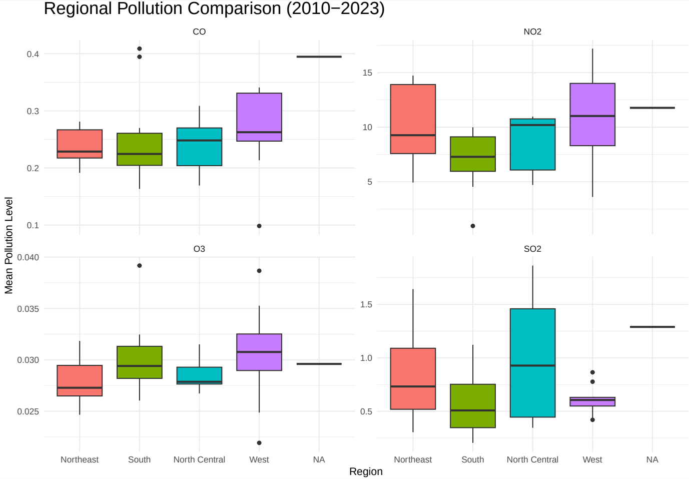
Average SO2 Levels by State (2010−2023)



Average NO2 Levels by State (2010−2023)



Regional Pollution Comparison (2010–2023)



- Northeast
  - Highest NO□ (traffic & urban areas)
  - Elevated SO□ (older industrial zones)
- North Central
  - High NO□ and SO□
  - Strong industrial influence
- West
  - Highest O□ (sunlight + elevation effects)
  - Generally lower CO/NO□/SO□
- South
  - Moderate overall
  - Lower NO□ & SO□

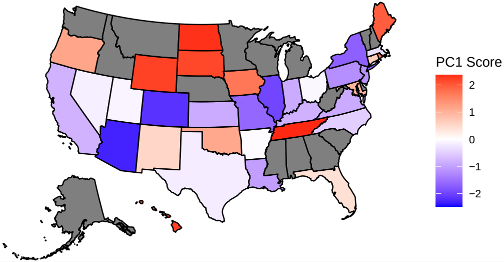
Why PCA? What it reveals beyond single-pollutant maps

- Single-pollutant maps tell where each pollutant is high/low
- But pollution types often co-occur (traffic → NO□ + CO□industry → SO□ + NO□)
- PCA summarizes overall pollution burden and pollution composition
- Helps identify:
    - ”Which states are overall the most polluted?”
    - ”Which states are ozone-dominated vs combustion-dominated?”

PCA

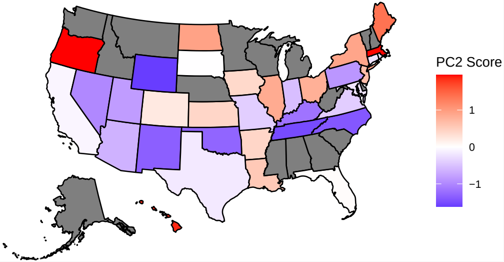
PC1: Overall Pollution Burden (AQI−based)

Higher = more polluted overall | Lower = cleaner

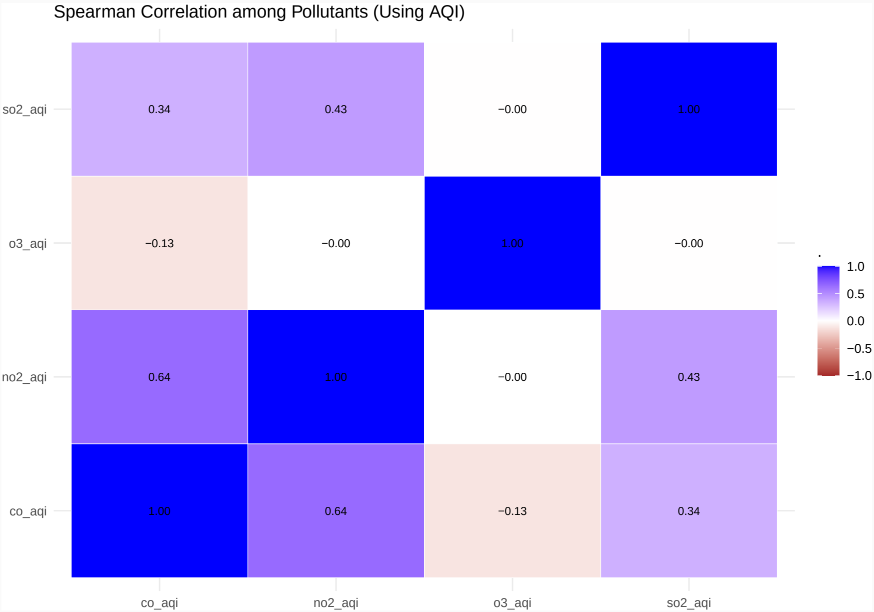


PC2: Pollution Composition (Ozone vs NO2/CO/SO2)

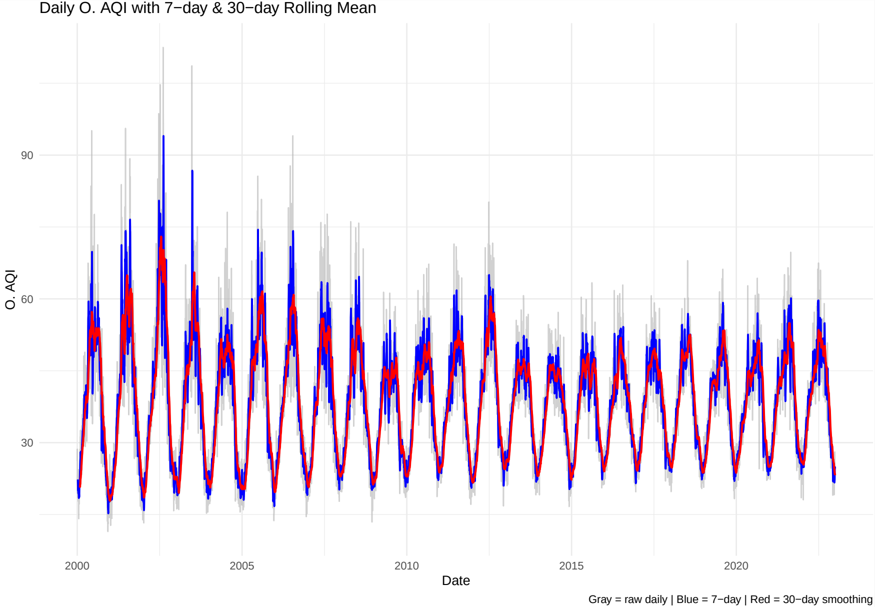
High = Ozone−dominated | Low = Combustion−pollution dominated



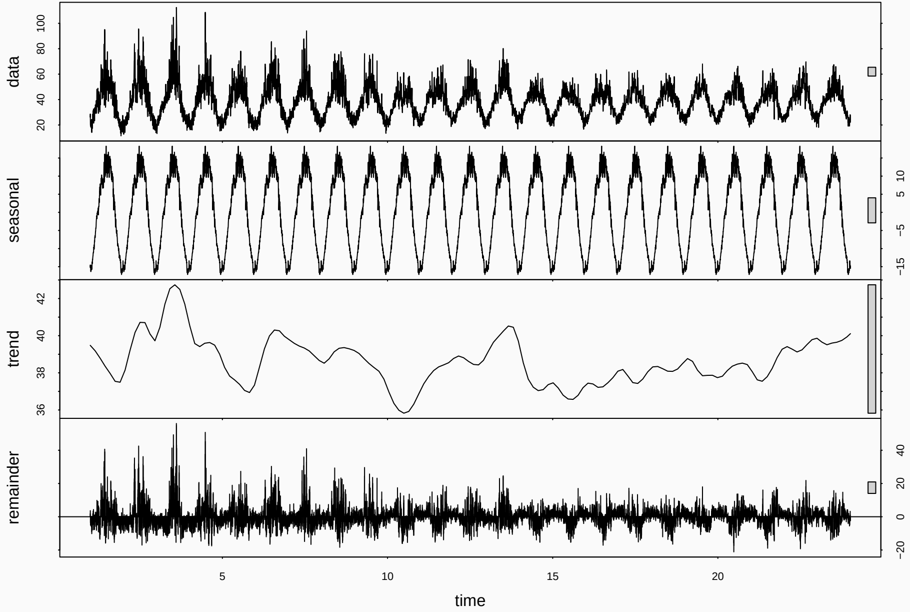
Correlation Among Pollutants



Rolling Mean Trend (7-day & 30-day)



STL Decomposition



AI-Assisted Workflow Reflection

- How I Used AI
  - Described the types of visual patterns I wanted to explore
  - Asked for suggestions on plot types (animation, PCA, correlation, seasonal trends, etc.)
  - Used AI-generated R code as a starting point
  - Improved slide layout, narrative, and storytelling with AI guidance
- Benefits
  - Accelerated exploratory data analysis
  - Suggested visualization ideas I hadn’t considered
  - Helped automate repetitive code (pivoting, summarizing, smoothing, faceting)
  - Improved the clarity and appearance of plots
  - Strengthened narrative structure across slides
- Limitations
  - Sometimes suggested variables not in the dataset
  - Produced repetitive or overly long code that required rewriting
  - Needed human verification for statistical interpretation (PCA meaning, correlations, etc.)
  - Lacked full context awareness; required manual refinement for reproducibility

Final Product Demonstration

Conclusions and Next Steps

- Key Findings**
  - Clear regional patterns: Northeast & industrial Midwest show higher NO□/SO□, while Western states show higher O□.
  - Strong pollutant relationships: NO□–CO–SO□ are positively correlated; O□ behaves independently with strong seasonality.
  - PCA effectively summarized pollution structure:
    - **PC1 = overall pollution burden**
    - **PC2 = ozone vs combustion pollution**
  - Temporal animation (2000–2023) reveals long-term O□ seasonality and declining NO□ trends.
- Limitations**
  - AQI values simplify true pollutant concentrations, losing granularity.
  - Monitoring station coverage is uneven across states and years.
  - PCA assumes linear structure; may miss nonlinear atmospheric interactions.
  - Some visualizations rely on aggregated means, which reduce local variability.
- Next Steps (with more time or better data)**
  - Build a full **time series forecasting model** comparing ARIMA, ETS, and Prophet across