Air Quality Analysis: 2000 - 2021

Air Quality Index

- Measure of air pollution
- Four major pollutants:
  - CO
  - NO2
  - SO2
  - O3
- Higher measures of AQI = more pollution = worse air quality

1. Data Structure

Data Structure

- Features:
  - Date (Year, Month, Day)
  - Location (Address, County, State)
  - Each Pollutant, Each Day:
    - Mean/Max
    - Time of Day of Max
    - AQI
- Observations:
  - Recorded daily in select locations for 2000 - 2021

Preprocessing

Data was already very tidy and required very little manipulation

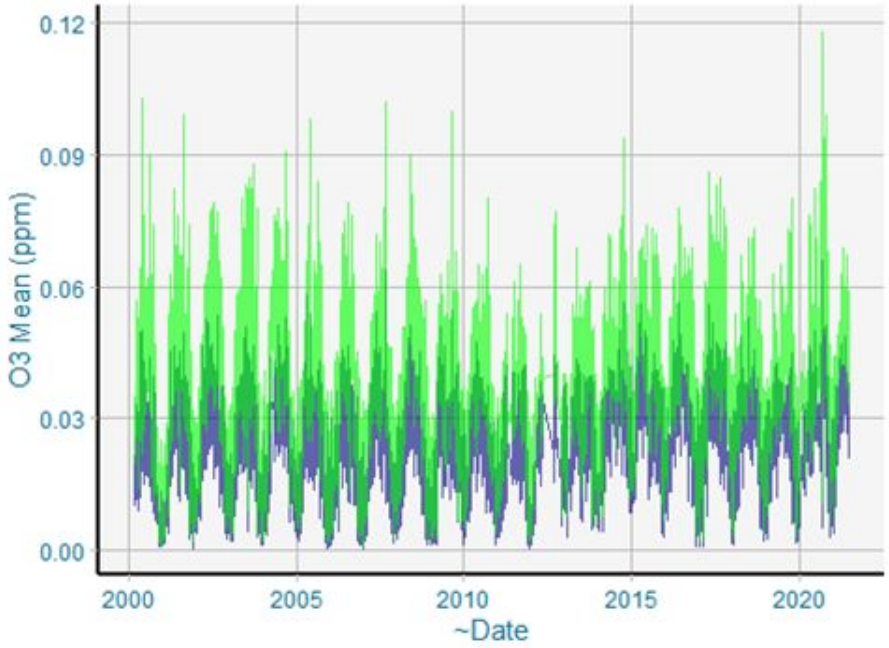
- Pollutant AQI -> Total AQI
- Binned AQI:
  - "Moderate", "At-Risk", "Unhealthy", "Very Unhealthy", "Hazard"
- Multiple daily readings (duplicates)
  - Aggregated means and consolidated to one obs

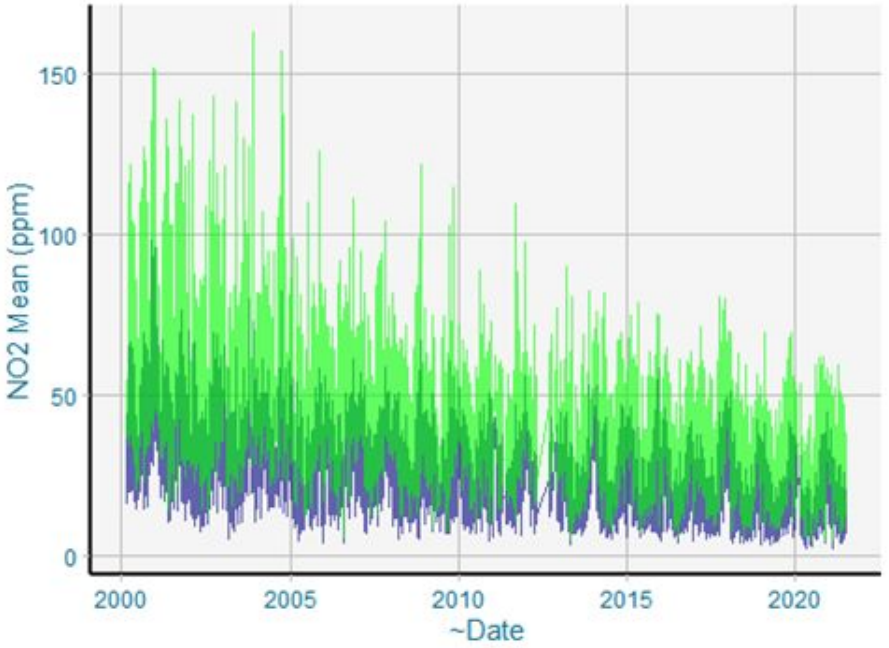
2. Visualizations

Changes of Pollutants, 2000-2021 in Los Angeles, CA

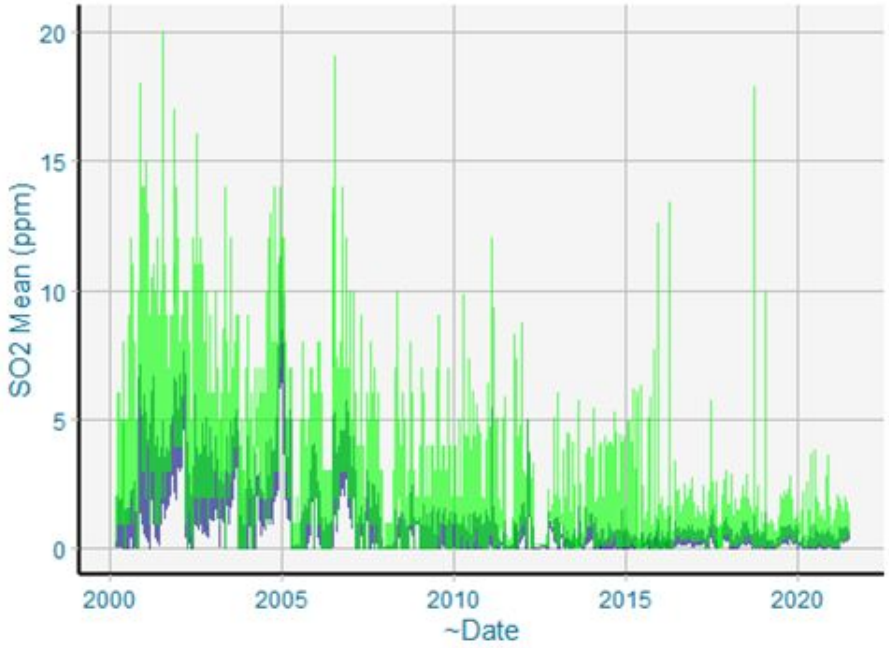
1. Each pollutant, 2000-2021
2. Each pollutant, 2020
    a. Yearly cycle

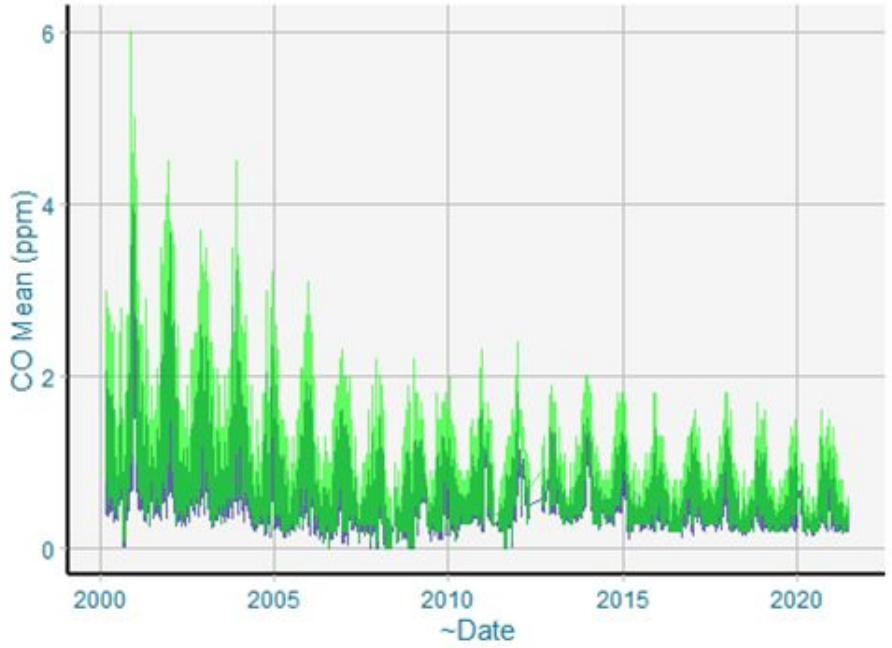
O3 / NO2


O3 Pollution 2000-2021
O3 Mean (ppm)
0.12
0.09
0.06
0.03
0.00
2000
2005
2010
2015
2020
~Date


NO2 Pollution 2000-2021
NO2 Mean (ppm)
150
100
50
0
2000
2005
2010
2015
2020
~Date

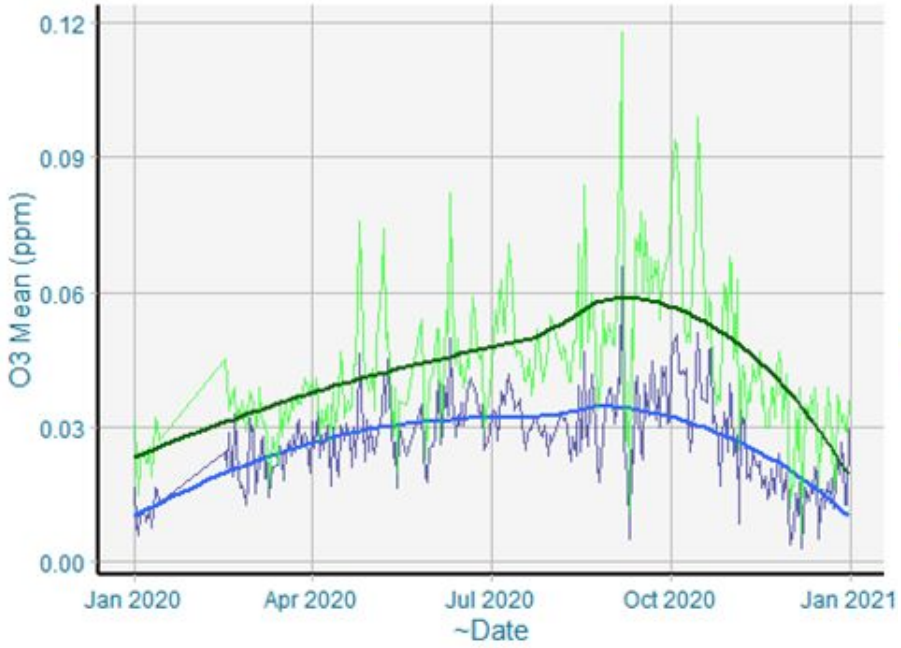
SO2 / CO

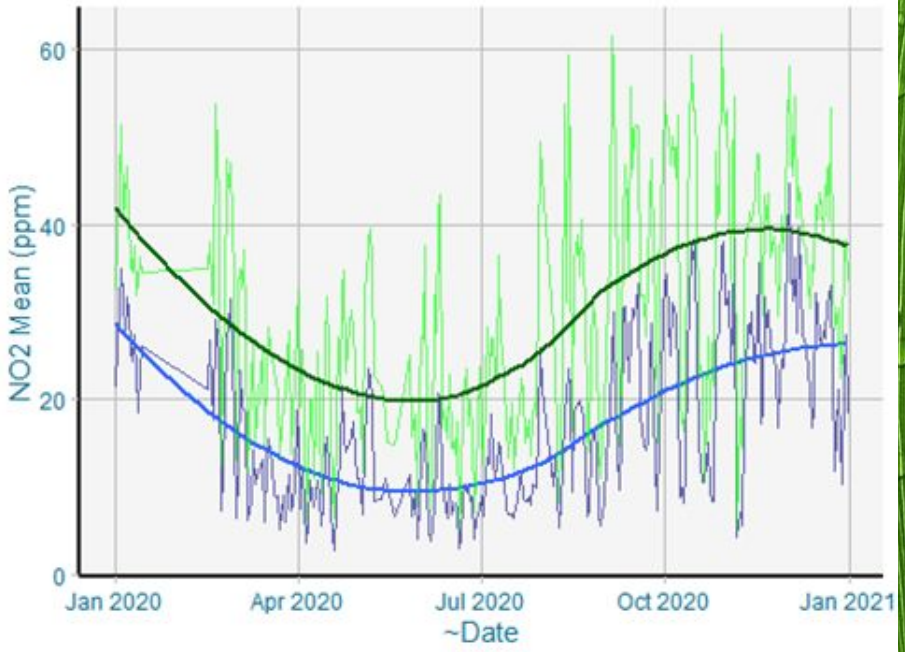

SO2 Pollution 2000-2021
SO2 Mean (ppm)
20
15
10
5
0
2000
2005
2010
2015
2020
~Date


CO Pollution 2000-2021
CO Mean (ppm)
6
4
2
0
2000
2005
2010
2015
2020
~Date

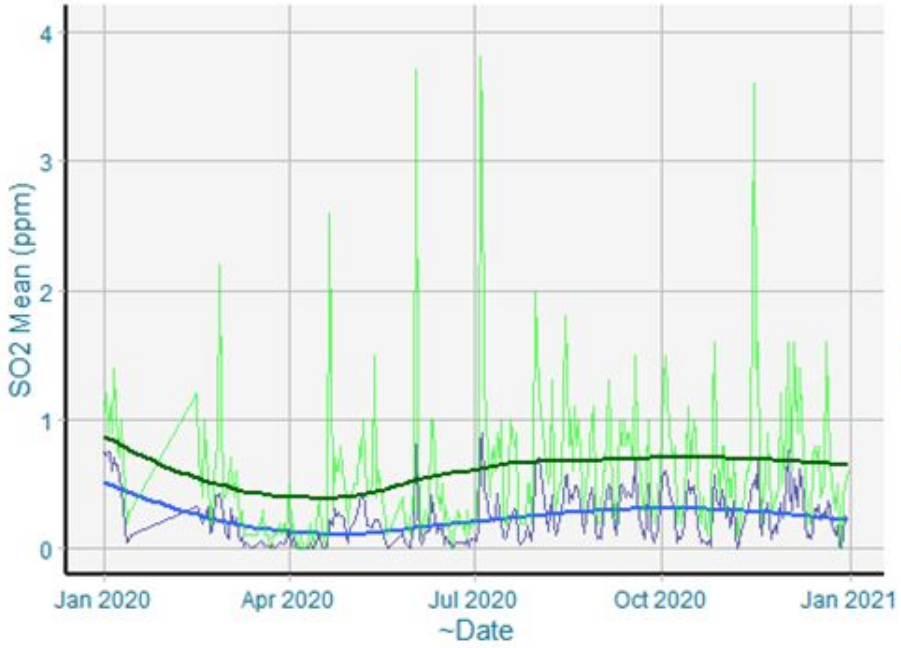
O3 / NO2

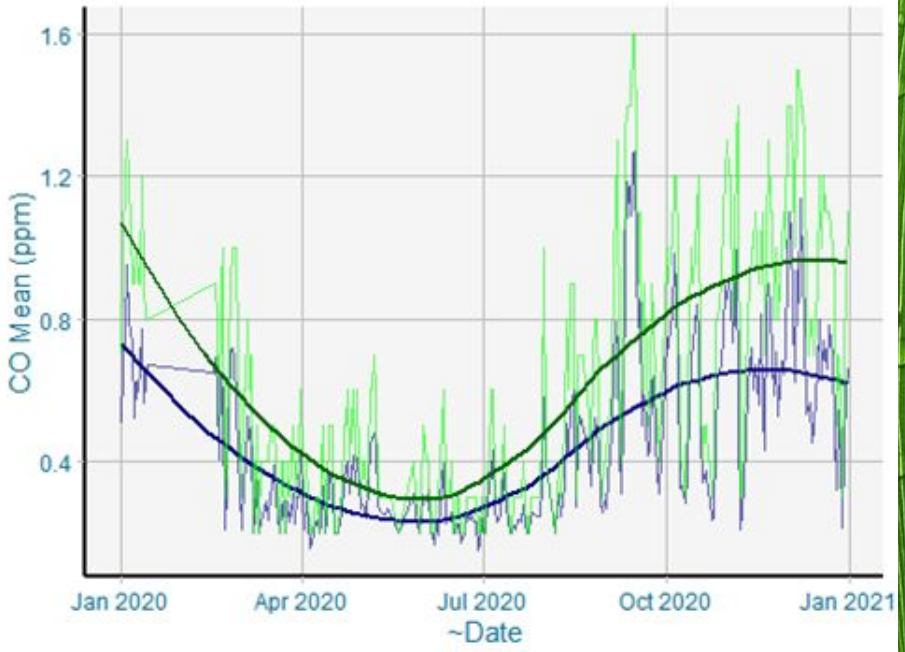
2020


O3 Pollution 2020
0.12
0.09
0.06
0.03
0.00
O3 Mean (ppm)
Jan 2020
Apr 2020
Jul 2020
Oct 2020
Jan 2021
~Date


NO2 Pollution 2020
60
40
20
0
NO2 Mean (ppm)
Jan 2020
Apr 2020
Jul 2020
Oct 2020
Jan 2021
~Date

SO2 / CO
2020

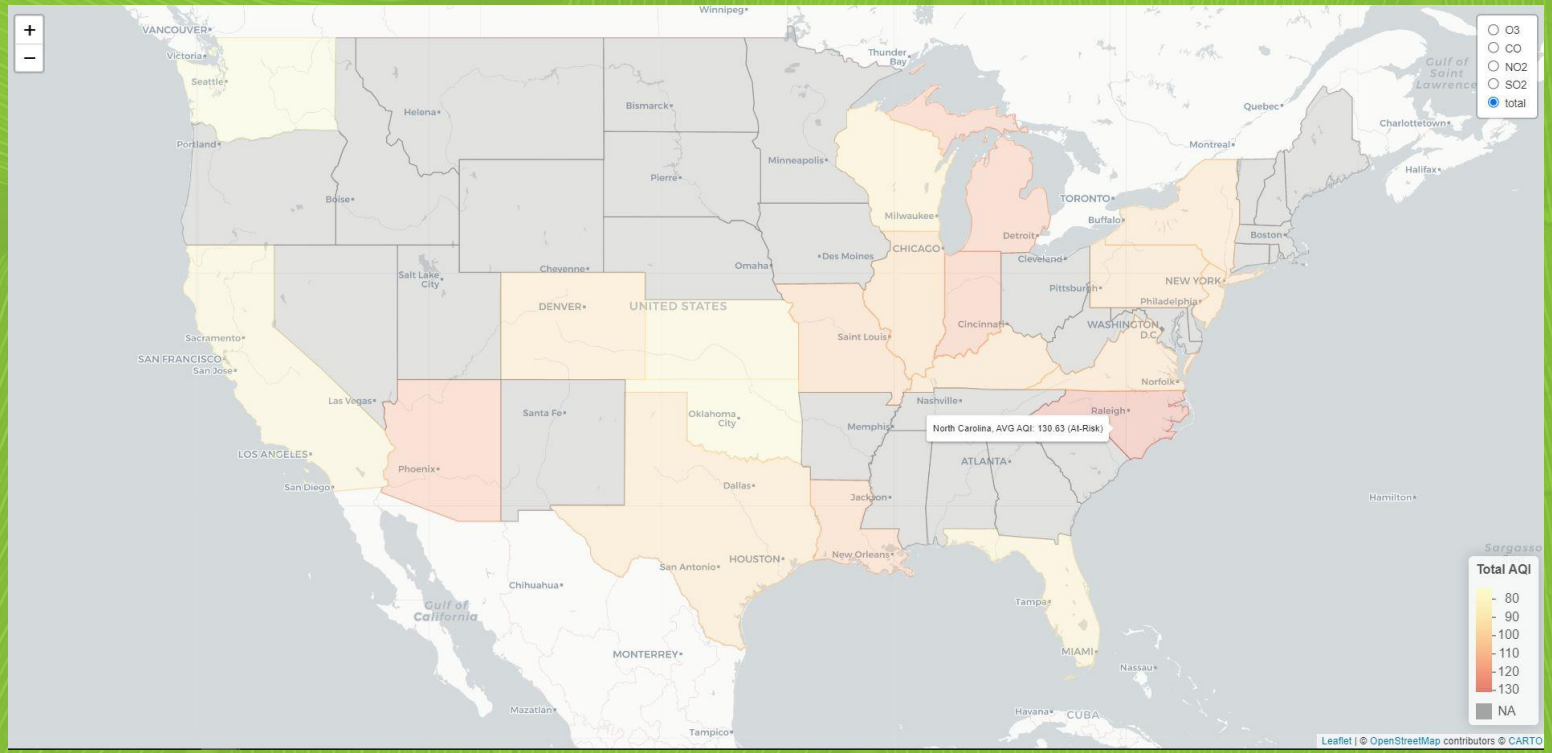

SO2 Pollution 2020
SO2 Mean (ppm)
0
1
2
3
4
Jan 2020
Apr 2020
Jul 2020
Oct 2020
Jan 2021
~Date


CO Pollution 2020
CO Mean (ppm)
0.4
0.8
1.2
1.6
Jan 2020
Apr 2020
Jul 2020
Oct 2020
Jan 2021
~Date

Mapping Air Quality in US

- Process
  - Simultaneously create leaflet maps for each year
  - Average AQI per state per year
  - AQI measurements (by pollutant, total)
    - Base group control
    - Auto-updating legend
  - Indicator of total AQI (“mild”, “hazard”, etc)

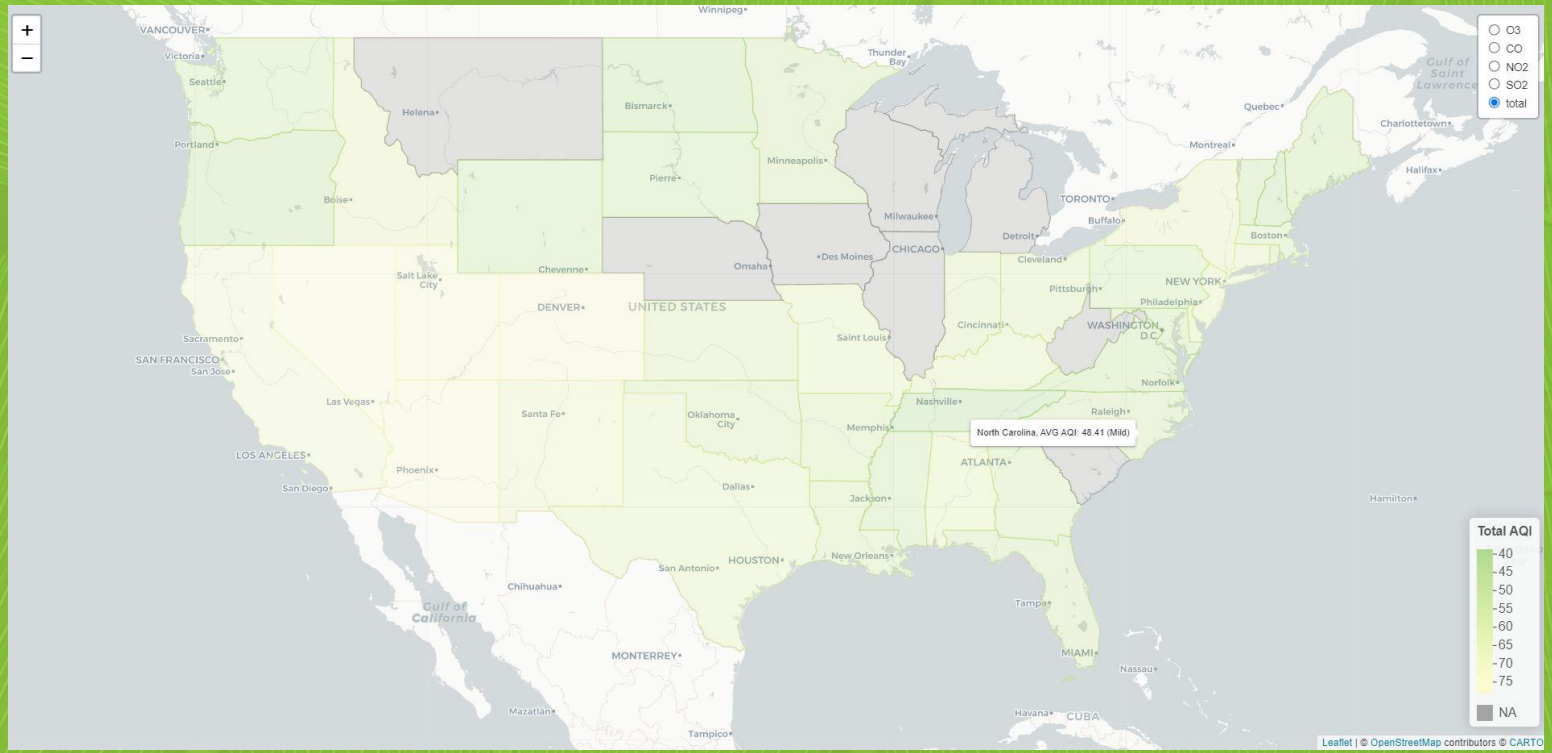
2000


O3
CO
NO2
SO2
total
North Carolina, AVG AQI: 130.63 (At-Risk)
Total AQI
80
90
100
110
120
130
NA
Leaflet | © OpenStreetMap contributors © CARTO

2010


O3
CO
NO2
SO2
total
North Carolina, AVG AQI: 80.35 (Moderate)
Total AQI
50
55
60
65
70
75
80
NA
Leaflet | © OpenStreetMap contributors © CARTO

2020

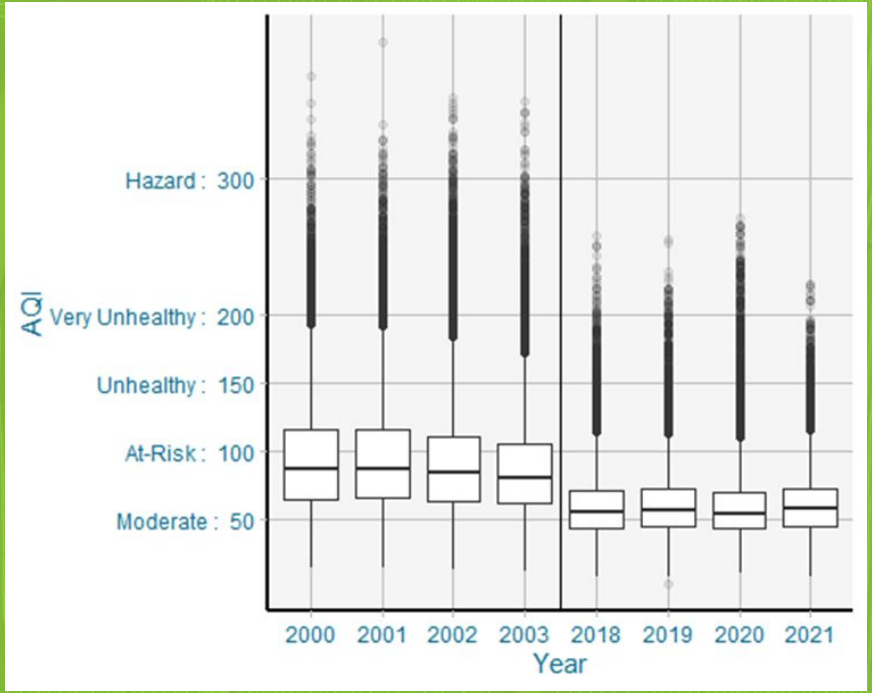

O3
CO
NO2
SO2
total
North Carolina, AVG AQI: 48.41 (Mild)
Total AQI
40
45
50
55
60
65
70
75
NA
Leaflet | © OpenStreetMap contributors © CARTO

TAKEAWAY

The air quality in the United States seems to have improved, indicating the possibility of climate initiative success

3. Statistical Analysis

AQI Distribution Change Over Time



Noticeable change in AQI over time

2003 median AQI > 2018 3q

Does NY have worse SO2 AQI than TX?

- Texas has the most fossil fuel power plants in US
- Adirondacks in NY known to have acid rain (caused by SO2 pollution)
  - Thanks BIO-101
- Non-parametric test used:
  - Mann Whitney U-test
    - 1-tailed
  - 99% Confidence

Null: $\mu_{NY} == \mu_{TX}$

Alt: $\mu_{NY} > \mu_{TX}$

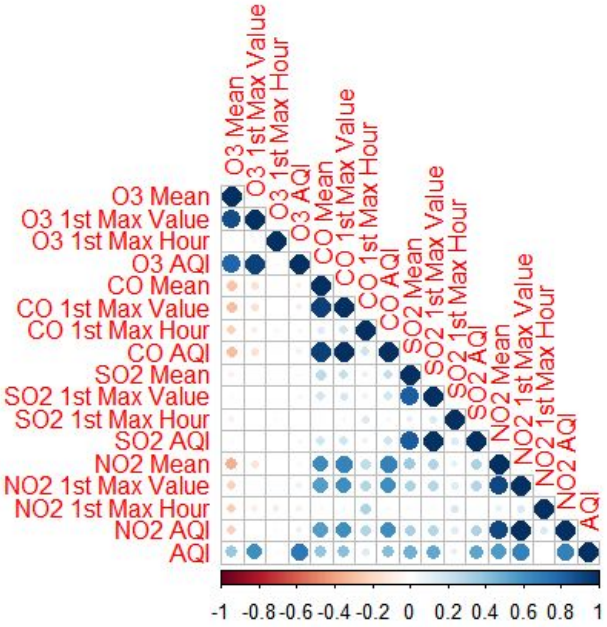
W-statistic: 5.7×10^{8}

P-Value: 2.2×10^{-16}

Reject the Null Hypothesis

New York has a higher average SO2 than Texas

Correlation Between Pollutants

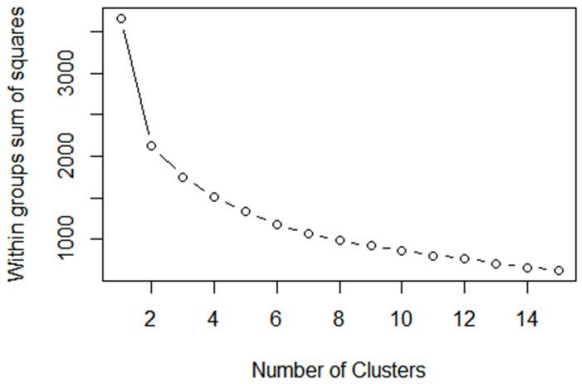


- Few interactions between pollutants
  - NO2 and CO moderate positive
  - O3 and NO2 larger cor with AQI

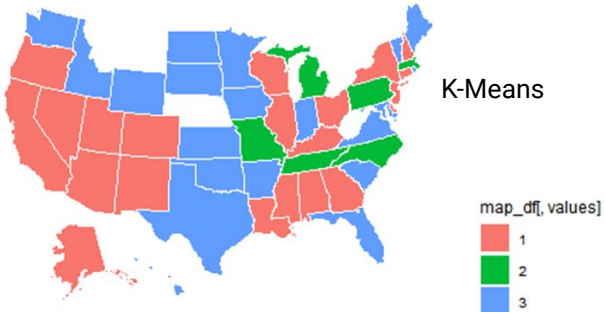
4. Machine Learning

Clustering

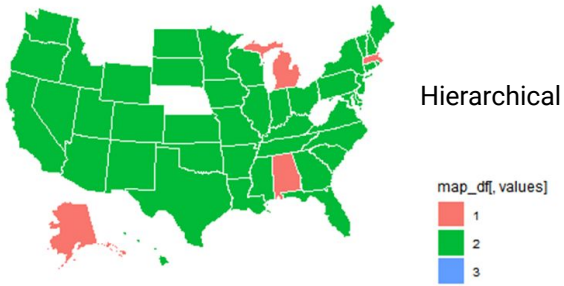
- Which states are most similar
  - Considering all pollutants
- K-means & Hierarchical
  - k=3



Results



K-Means



Hierarchical

- No obvious patterns
- K-means produced more balanced clusters

Regression: Mean or Max to predict?

- Method: Multiple Linear Regression
  - Predictors: Avgs/Maxs of each pollutant
  - **Max value** of pollutants is a better predictor of overall AQI than the average value
- Max Value Model: Adj R^2 = 0.9376
- Mean Value Model: Adj R^2 = 0.7678

- Hypothesis: Max will outperform min when used to predict

Regression Prediction Results

- Maximum AQI a great predictor
  - Using 80/20 split
  - 98.9% of predictions within 1 s.d. of AQI
  - 96.8% of predictions within ½ s.d. of AQI
  - 82.2% of predictions within ¼ s.d. of AQI
- Mean AQI an okay predictor
  - Using 80/20 split
  - 96.0% of predictions within 1 s.d. of AQI
  - 78.6% of predictions within ½ s.d. of AQI
  - 47.4% of predictions within ¼ s.d. of AQI

Classification: Rural or Urban?

- Urban state: >70% urban population
- Logistic Regression

% Split	Classification Accuracy
50/50	62.3%
75/25	61.2%
80/20	62.1%
90/10	***62.4%***

Future Work

- Deploy shiny server with time slider instead of 21 separate leaflet maps
- Examine relationship with geographic power plant distribution
- Examine on a city level instead of state
  - Is it a city or a town based on pollution?

Tools

R Programming language

- Tidyverse (https://www.tidyverse.org/)
- Usmap (https://cran.r-project.org/web/packages/usmap/usmap.pdf)
- Sf (https://cran.r-project.org/web/packages/sf/sf.pdf)
- Spdata (https://www.rdocumentation.org/packages/spData/versions/2.0.1)
- Ggmap (https://cran.r-project.org/web/packages/ggmap/ggmap.pdf)
- Corrplot (https://www.rdocumentation.org/packages/corrplot/versions/0.92/topics/corrplot)
- Wilcox test

Refs

Data:

https://www.kaggle.com/datasets/alpacanonymous/us-pollution-20002021

https://www2.census.gov/geo/docs/reference/ua/PctUrbanRural_State.xls

Other:

https://www.epa.gov/so2-pollution/sulfur-dioxide-basics

https://www.statista.com/statistics/1248106/so2-most-polluting-power-plants-united-states/

https://statsandr.com/blog/wilcoxon-test-in-r-how-to-compare-2-groups-under-the-non-normality-assumption/

https://www2.census.gov/geo/docs/reference/ua/PctUrbanRural_State.xls