Air Quality Analysis and Simulation

Mar. 2020

Data Scientist
SJ Choi

Project Scope

Project Scope

- **Primary data : Air Quality**

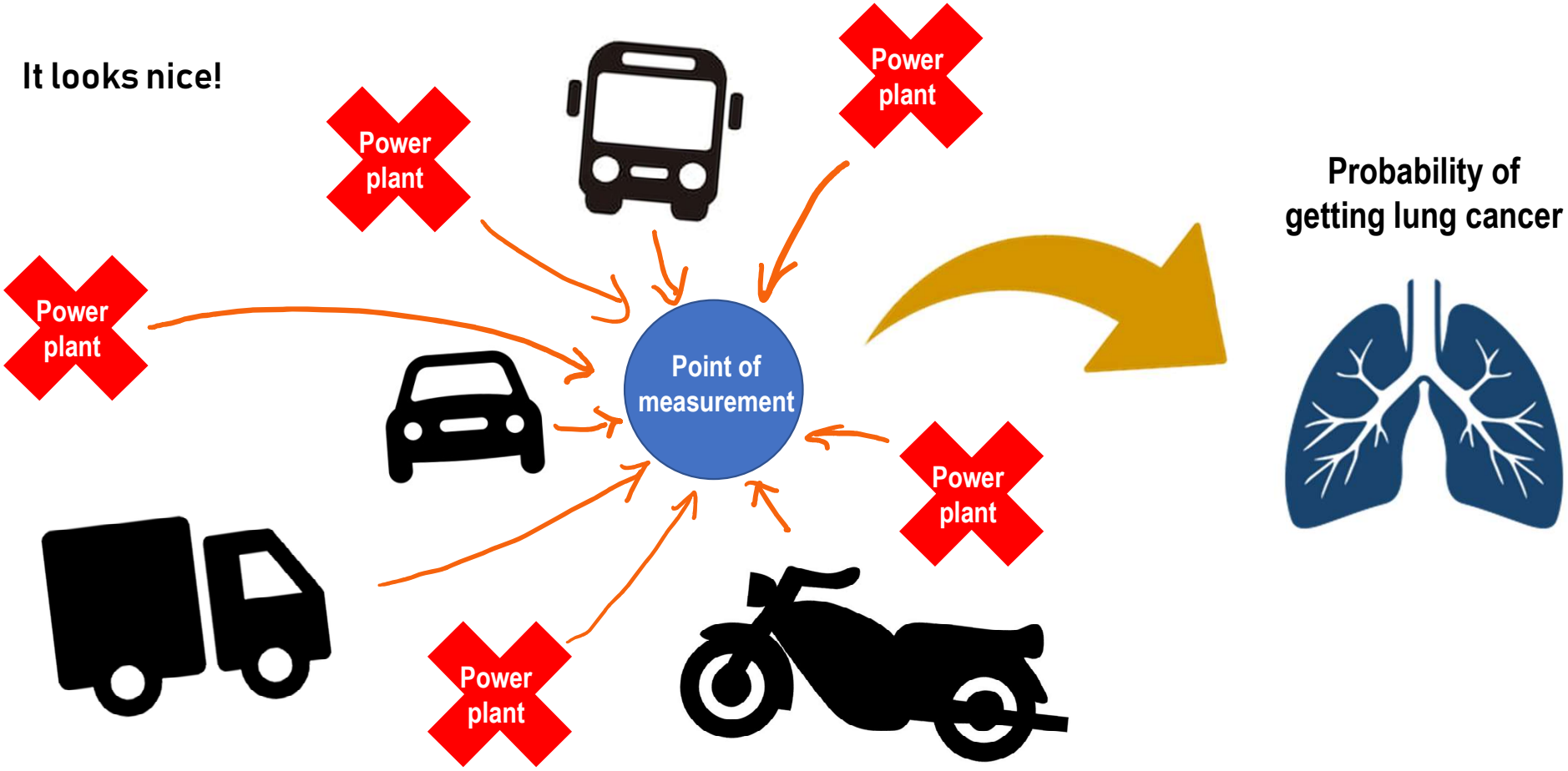
Air types	Ozone	PM 10	PM 2.5	CO	NO2	SO2
States	Alabama Alaska Arizona Arkansas California Colorado Connecticut Delaware Florida Georgia Hawaii Idaho Illinois Indiana Iowa Kansas Kentucky Louisiana Maine Maryland Massachusetts Michigan Minnesota Mississippi Missouri Montana Nebraska Nevada New Hampshire New Jersey New Mexico New York North Carolina North Dakota Ohio Oklahoma Oregon Pennsylvania Rhode Island South Carolina South Dakota Tennessee Texas Utah Vermont Virginia Washington West Virginia Wisconsin Wyoming					
Years	2003 ~ 2019					

- **Secondary data : Cause and effect of Air Quality**

Power plants	Power plants information in all states in 2018, 2016, 2014, 2012, 2010, 2009, 2007, 2005, 2004
Vehicles	Number of registered vehicles(cars, trucks, buses, motorcycles) in all states 2003 ~ 2017
Lung cancer	Number of incidence of lung cancer in All states 2003 ~ 2016

Project Scope

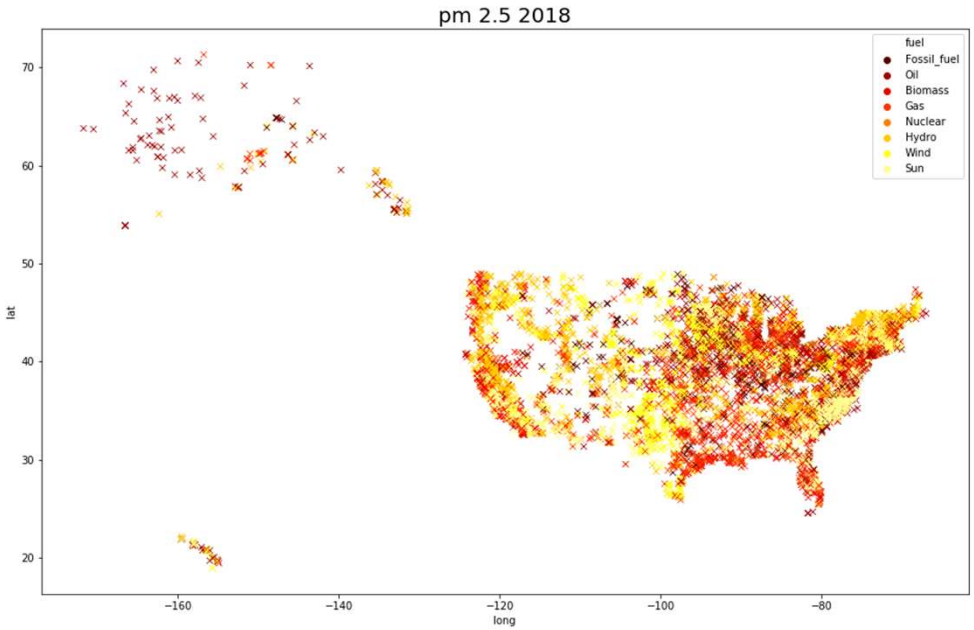
It looks nice!



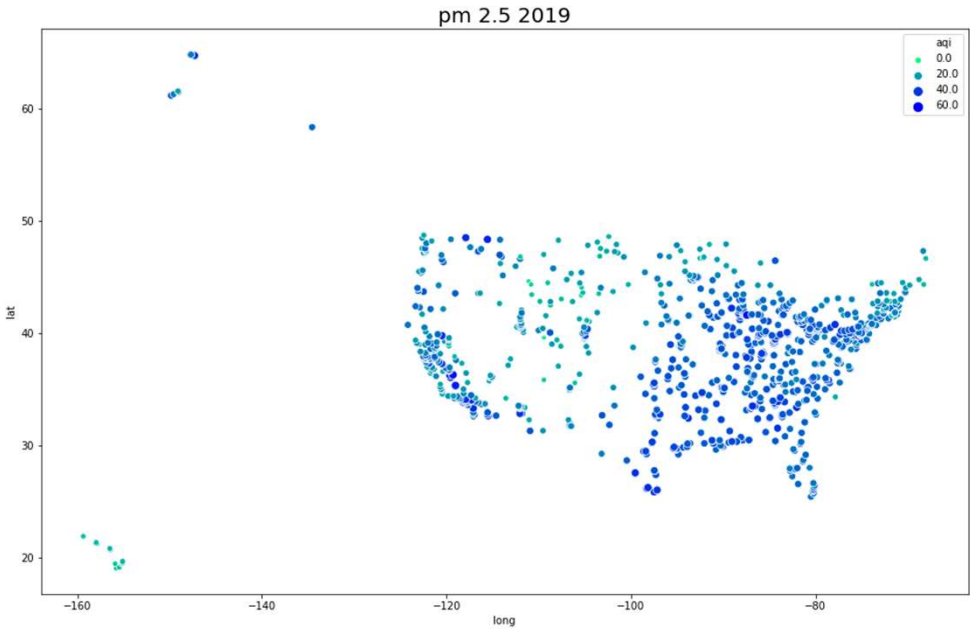
Project Scope

But in reality…

10,784 power plants x 1,278 measure points x 6 air types and other features to model



[Power plants]



[Measure points]

Project Scope – tuned !!

- **Primary data**

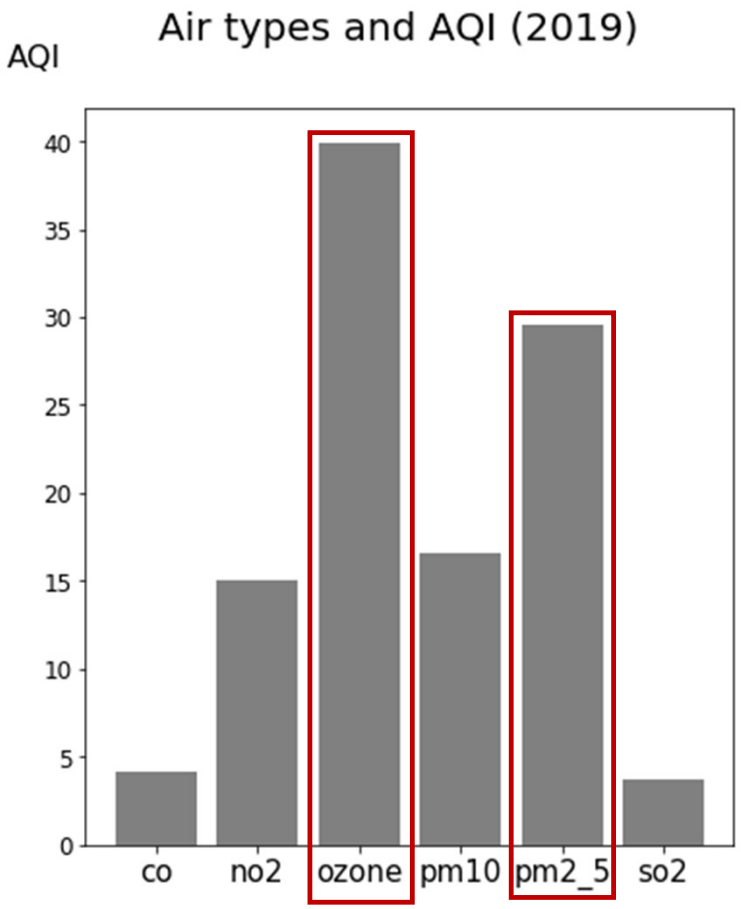
Air types	Ozone	PM 10	PM 2.5	CO	NO2	SO2
States	Alabama Alaska Arizona Arkansas California Colorado Connecticut Delaware Florida Georgia Hawaii Idaho Illinois Indiana Iowa Kansas Kentucky Louisiana Maine Maryland Massachusetts Michigan Minnesota Mississippi Missouri Montana Nebraska Nevada New Hampshire New Jersey New Mexico New York North Carolina North Dakota Ohio Oklahoma Oregon Pennsylvania Rhode Island South Carolina South Dakota Tennessee Texas Utah Vermont Virginia Washington West Virginia Wisconsin Wyoming					
Years	2003 ~ 2019					

- **Secondary data**

Power plants	All States in 2018, 2016, 2014, 2012, 2010, 2009, 2007, 2005, 2004
Vehicles	Number of registered vehicles(cars, trucks, buses, motocycles) in All states 2003 ~ 2017
Lung cancer	Number of incidence of lung cancer in All states 2003 ~ 2016

Air Quality

Why Ozone and PM 2.5 ?



- Ozone

How to be made: Bad ozone forms near the ground when pollutants (emitted by sources such as cars, power plants, industrial boilers, refineries, and chemical plants) react chemically in sunlight.
Health effects: It irritates the respiratory system, reduces lung function, inflames and damages the cells that line the lungs. It also makes the lungs more susceptible to infection. It causes aggravate asthma, other chronic lung diseases and permanent lung damage.

- PM 2.5 (particulate matter 2.5 micrometers or less)

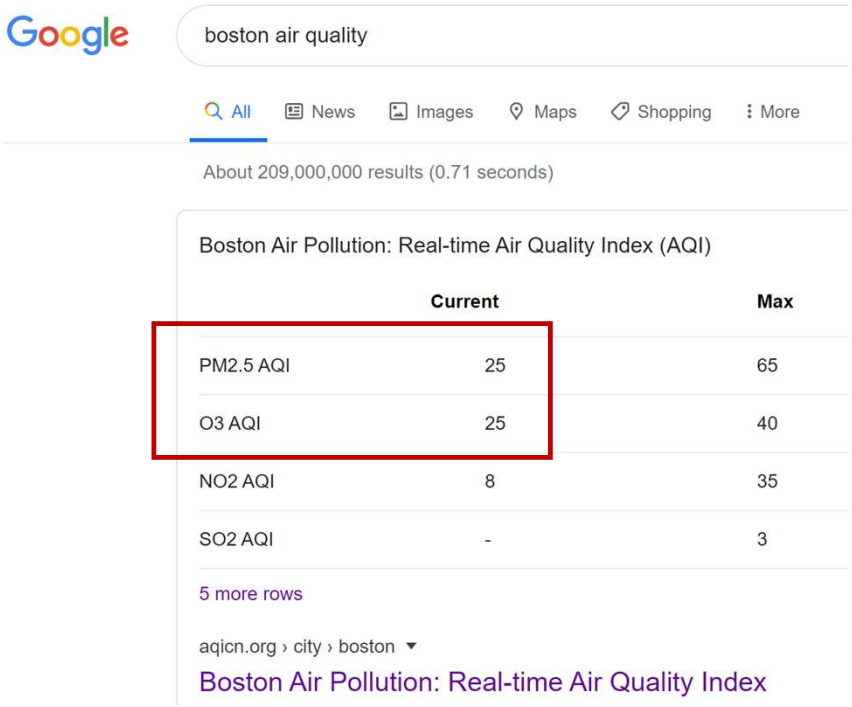
How to be made: Major sources of fine particles include motor vehicles, power plants, residential wood burning, forest fires, agricultural burning, some industrial processes, and other combustion processes.
Health effects: When exposed to particle pollution, people with heart or lung diseases and older adults are more likely to visit emergency rooms, be admitted to hospitals, or in some cases, even die. Particle pollution also can increase susceptibility to respiratory infections and can aggravate existing respiratory diseases, such as asthma and chronic bronchitis.

[Source: https://airnow.gov/index.cfm?action=aqi_brochure.index]

What is Air Quality Index (AQI)?

The AQI is an index for reporting daily air quality.

Air Quality Index (AQI) Values	Levels of Health Concern	Colors
When the AQI is in this range:	*...air quality conditions are:*	*...as symbolized by this color:*
0 - 50	Good	Green
51 - 100	Moderate	Yellow
101 - 150	Unhealthy for Sensitive Groups	Orange
151 - 200	Unhealthy	Red
201 - 300	Very Unhealthy	Purple
301 - 500	Hazardous	Maroon



[Source: https://airnow.gov/index.cfm?action=aqi_brochure.index]

Data

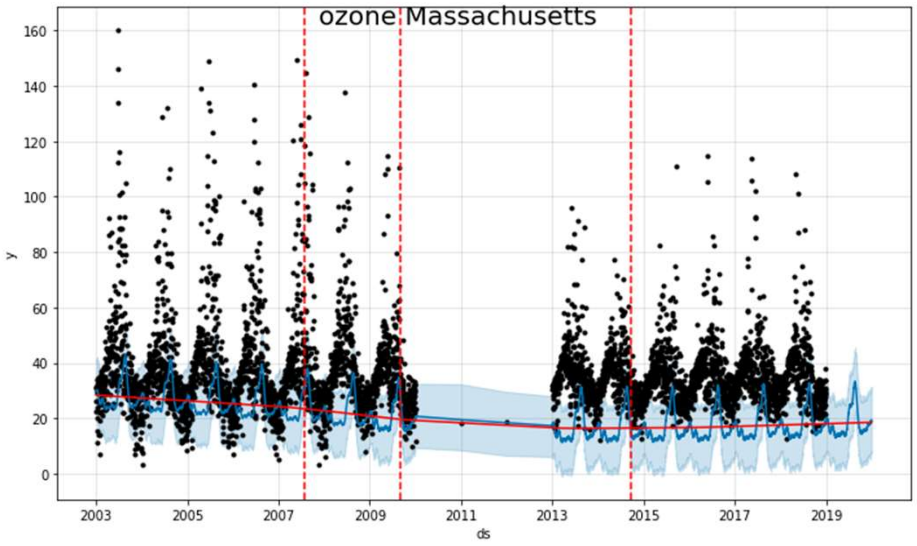
Data Overview

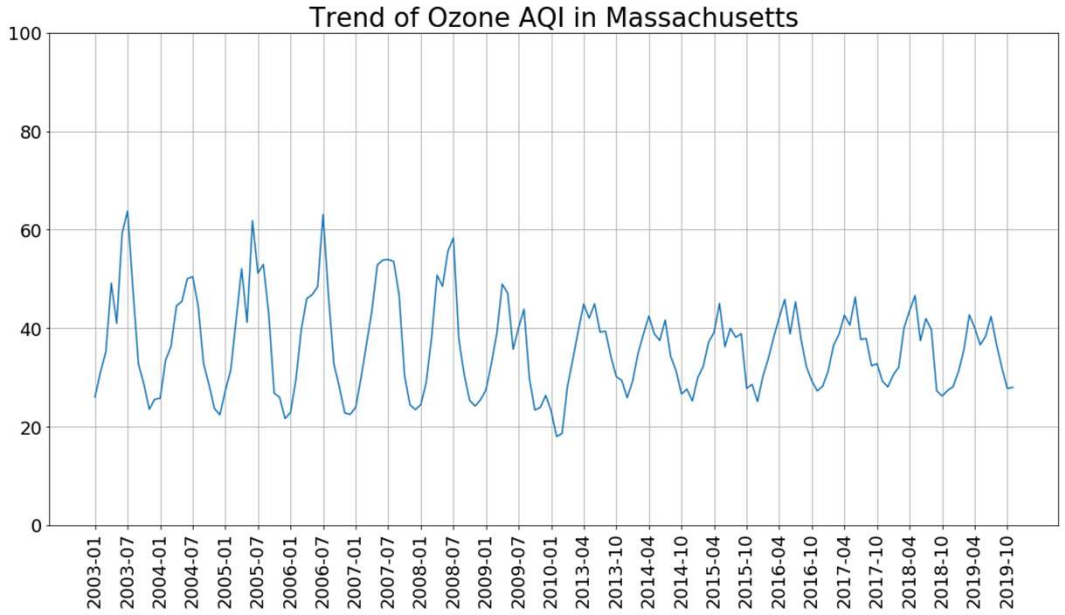
- Data overview

Data earned by scraping	- Air quality data (18 GB) - Geo data - Power plant data - Registered Vehicle data - Lung cancer data
After data cleaning	- Air quality data reduced to 4GB
After EDA	- Data was reproduced in certain format and 8GB was in use

AQI Trend - Ozone

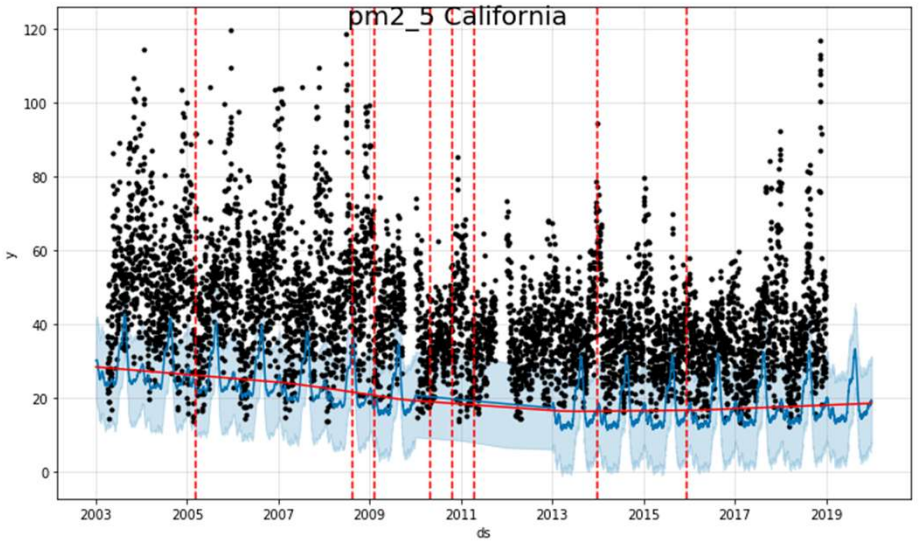
For ozone, we can group it before and after 2010
Ozone aqi is rather stable but the size of spike is descending

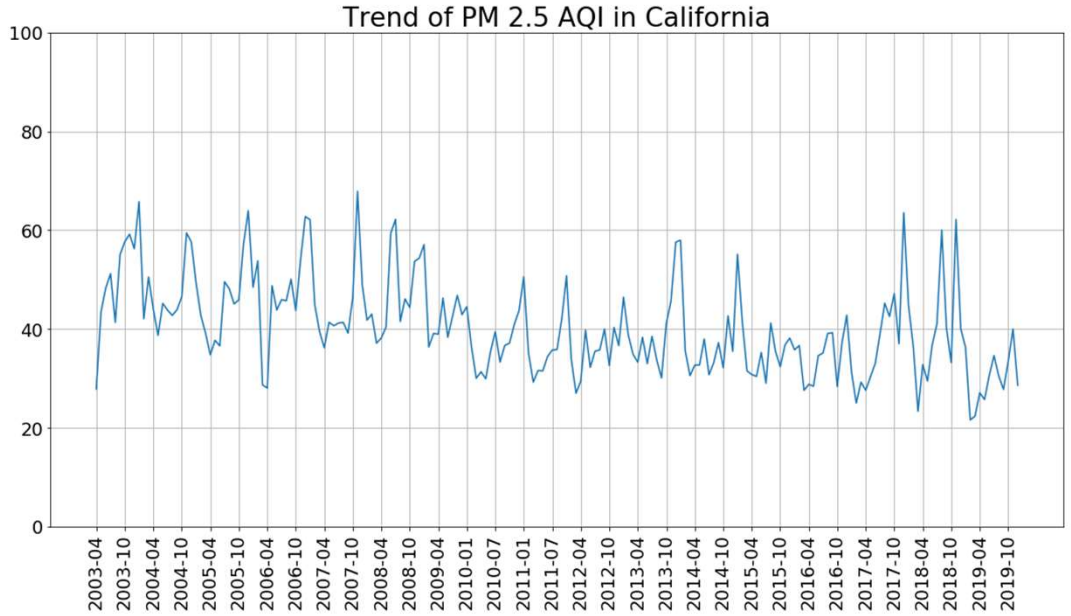




AQI Trend – PM 2.5

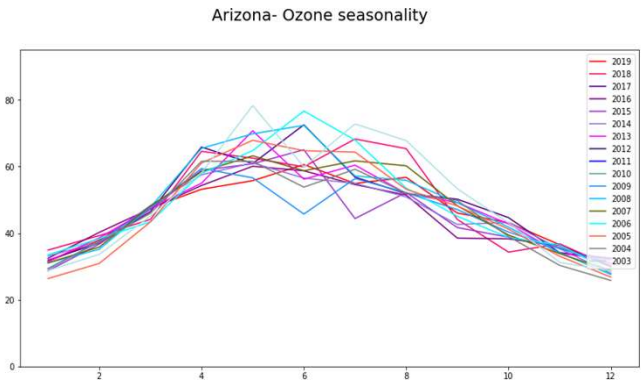
PM 2.5 shows gradual descending trend

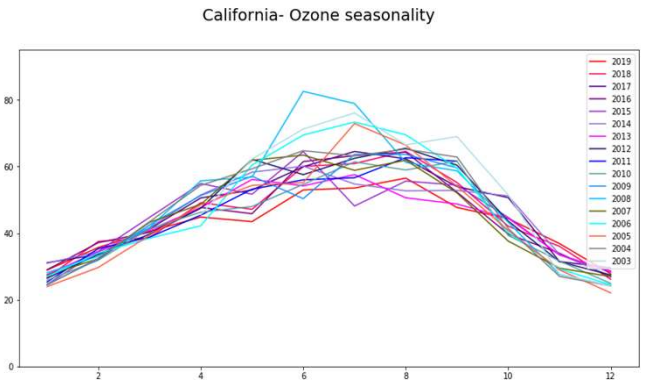


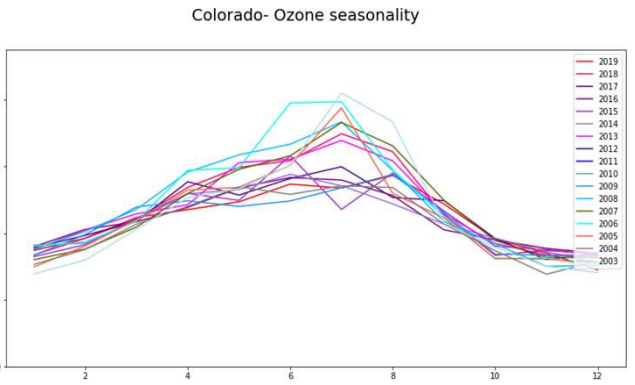


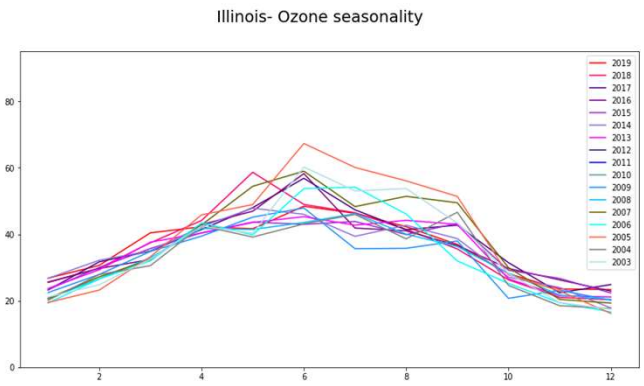
AQI Seasonality – Ozone

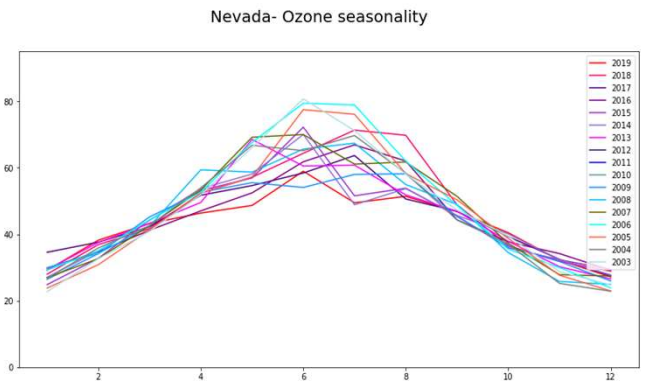
For ozone, the most common phenomenon is high in summer and low in winter.

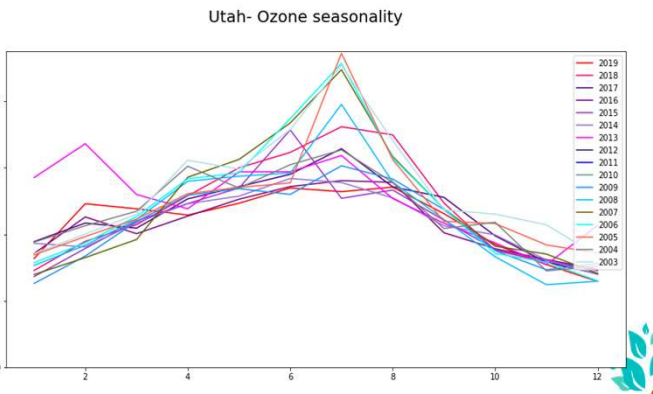






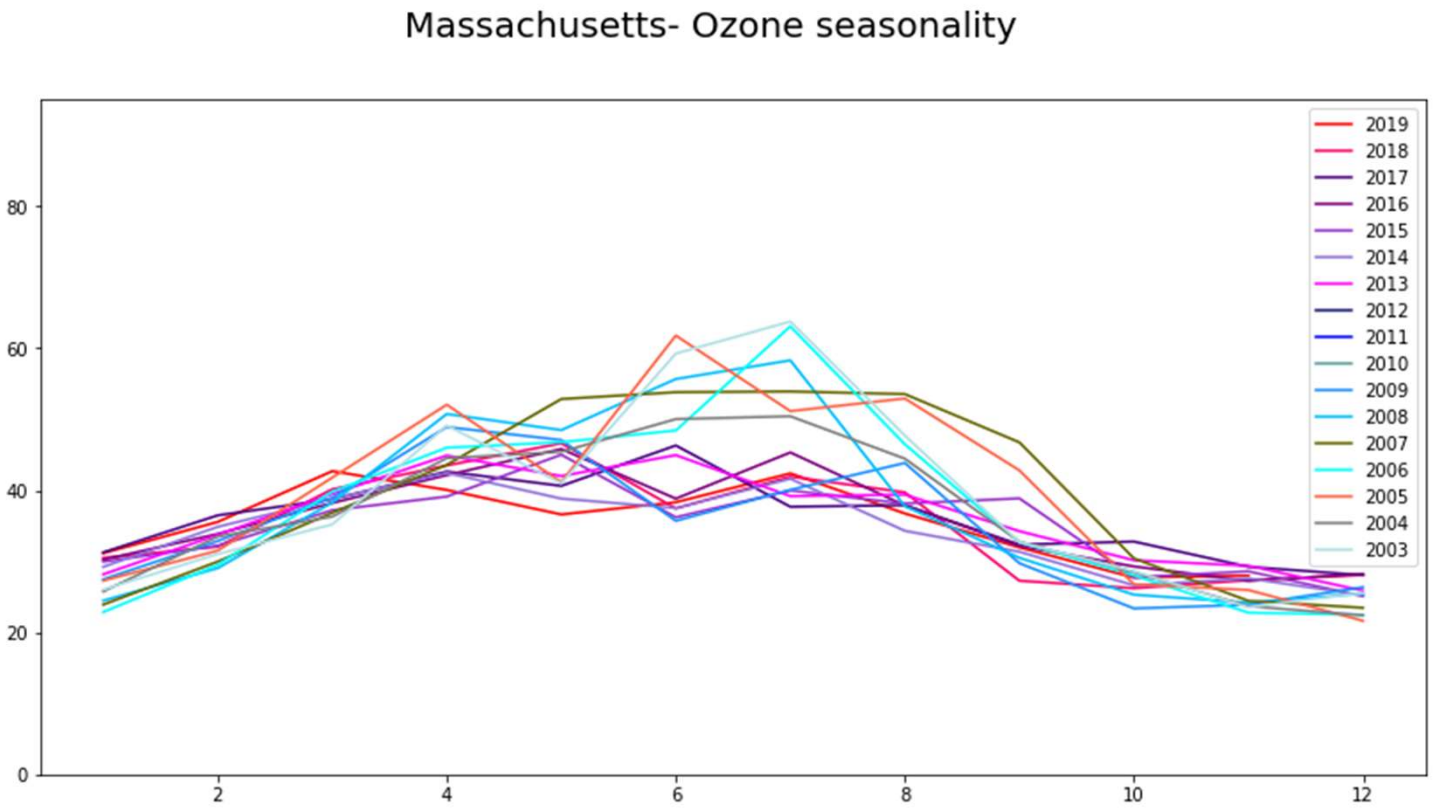






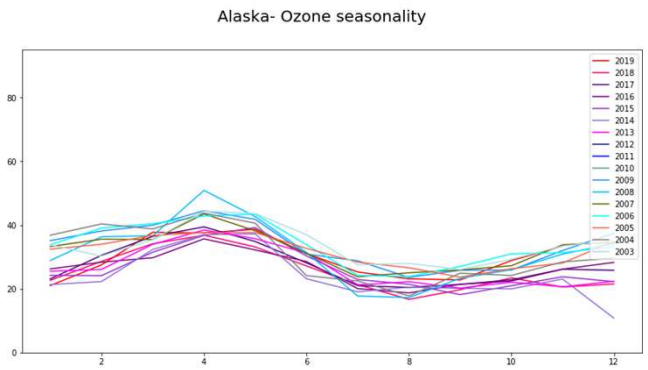
AQI Seasonality – Ozone

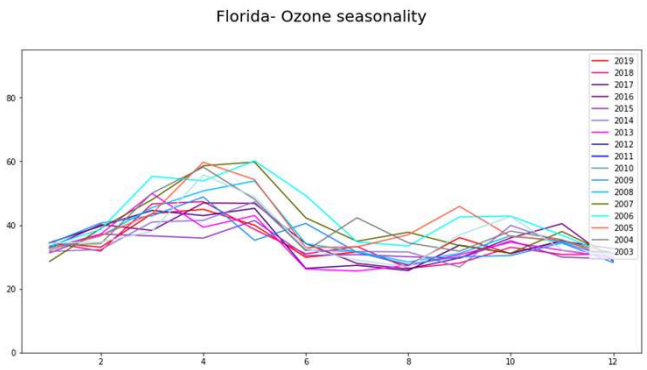
For most states, especially recently, the Ozone AQI has been stable over the years.

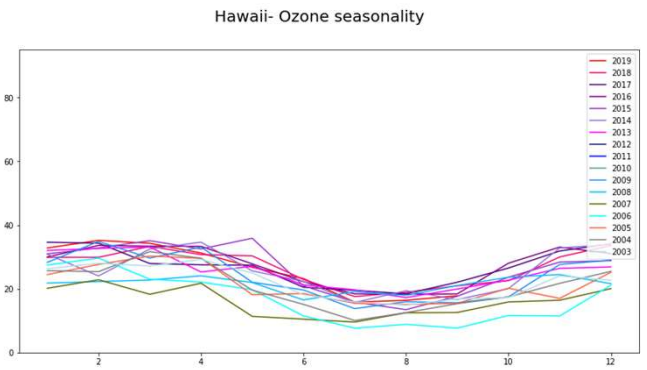


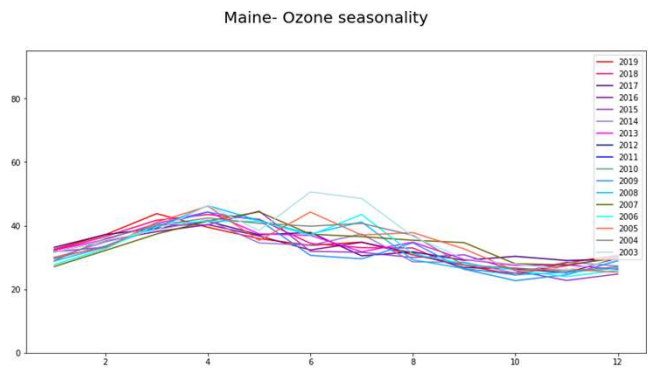
AQI Seasonality – Ozone

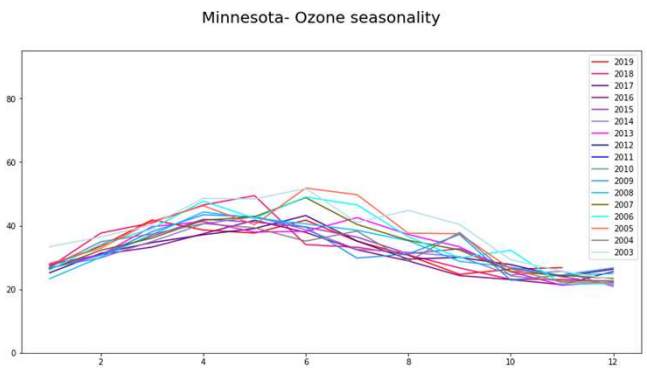
For some states, most of them are near the ocean, Ozone is high in spring.

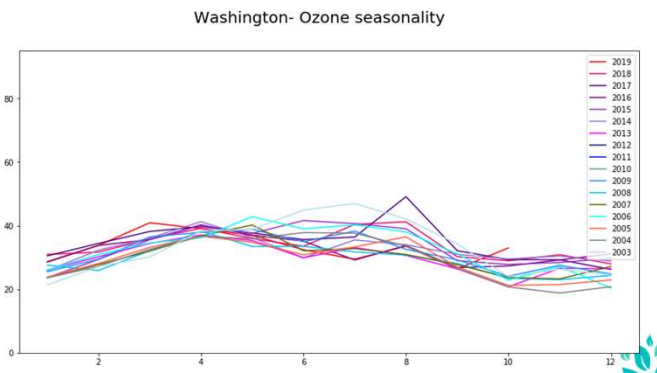






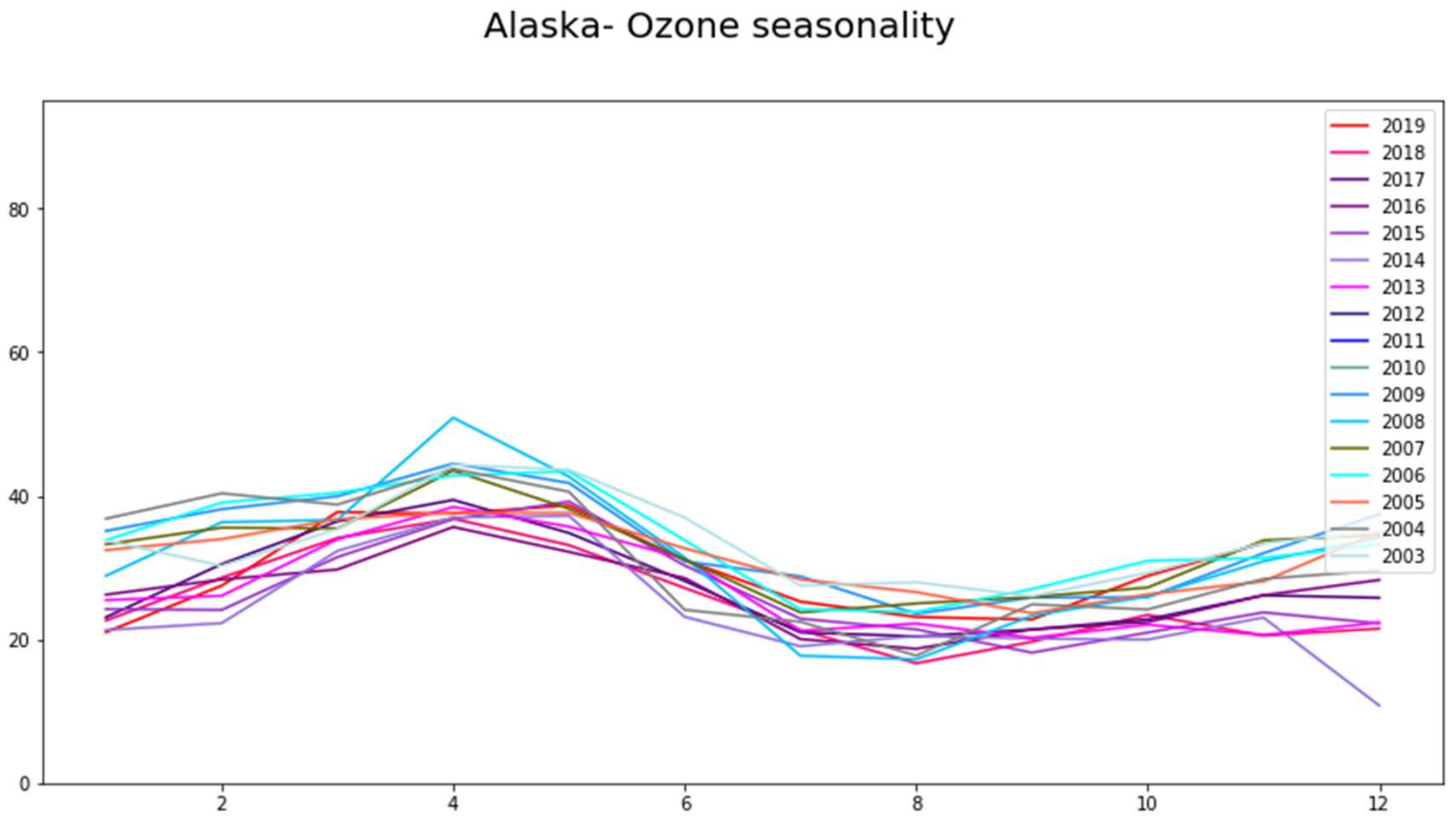






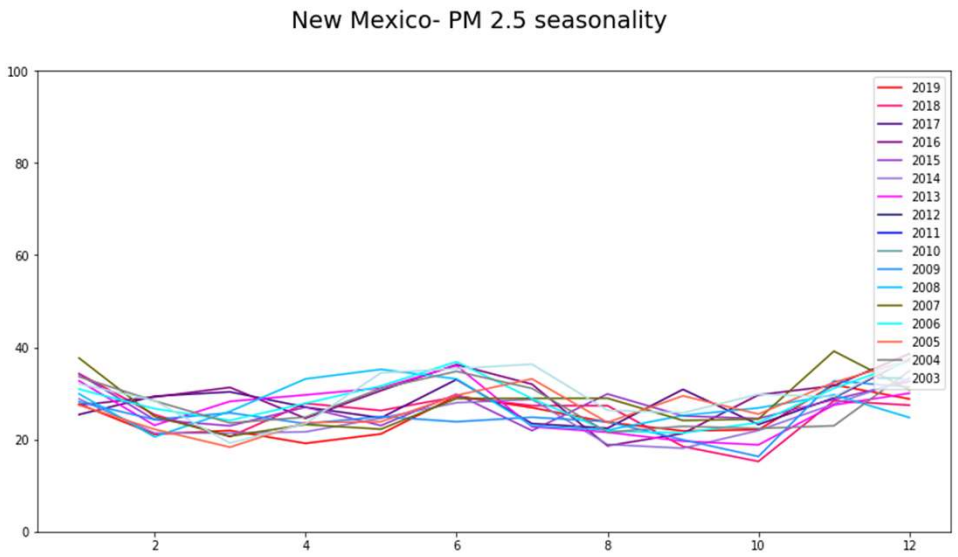
AQI Seasonality – Ozone

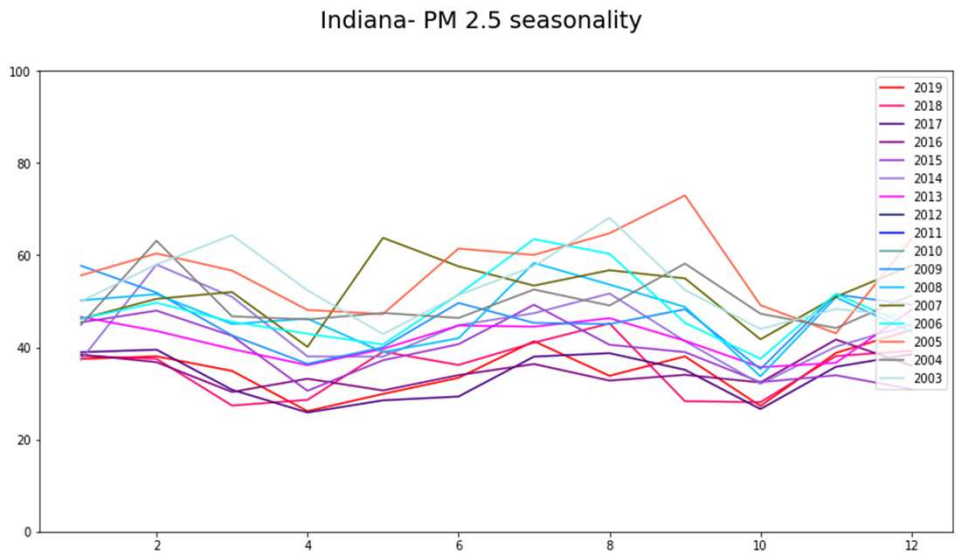
For some states, Ozone is high in spring.



AQI Seasonality – PM 2.5

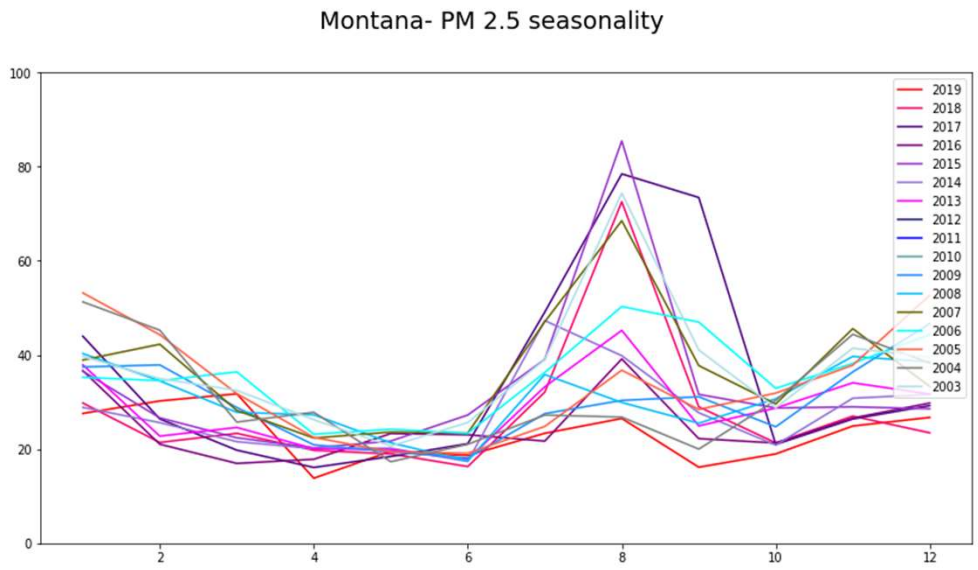
Various: Stable

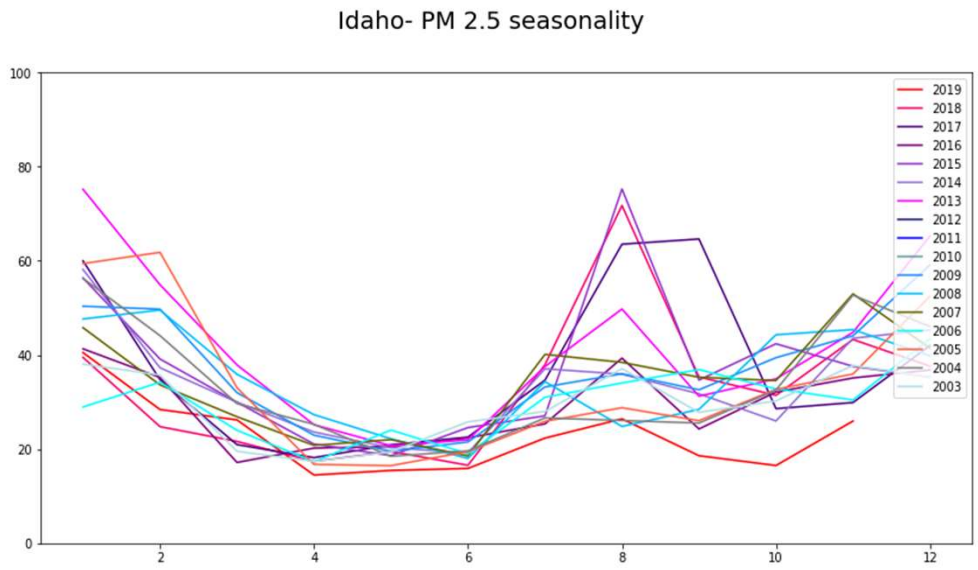




AQI Seasonality – PM 2.5

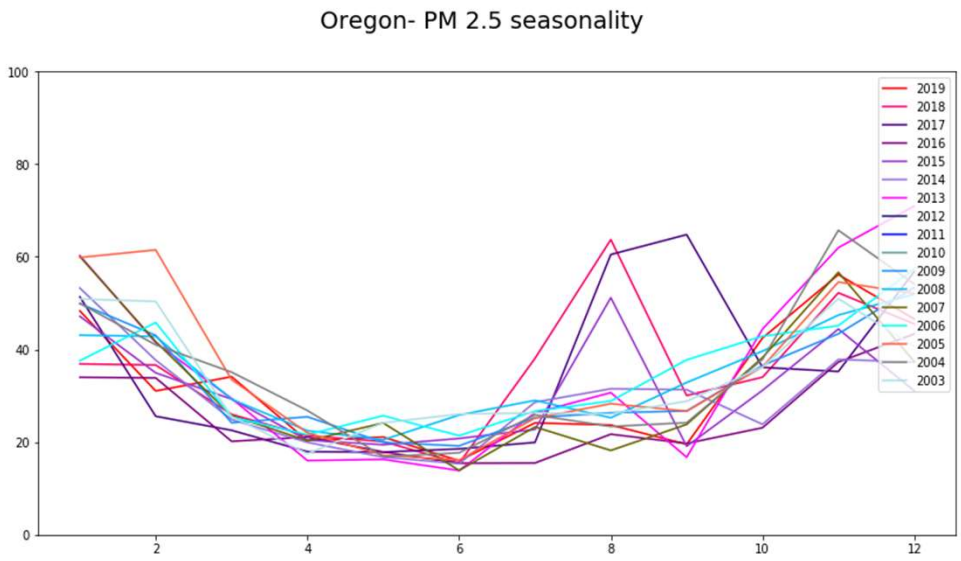
Various: High in summer

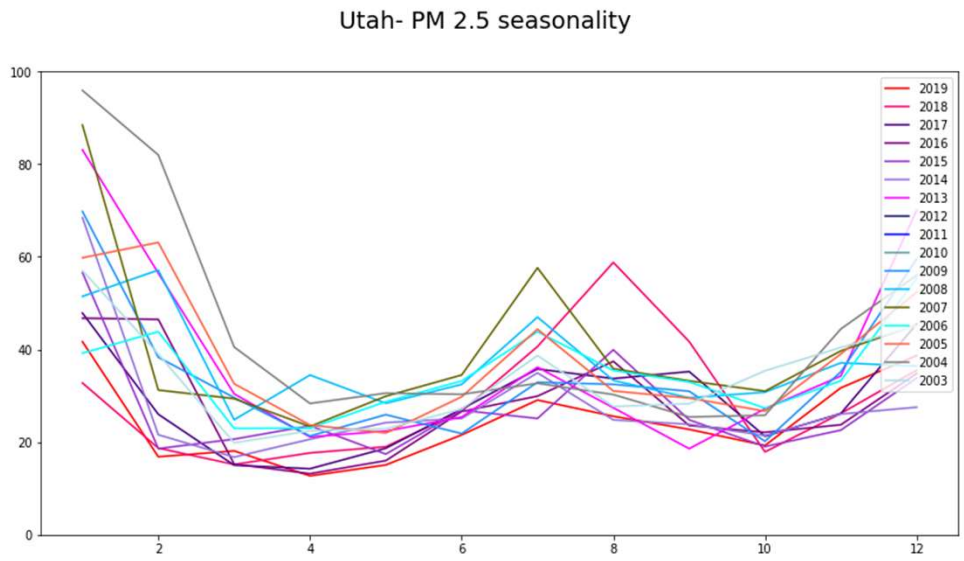




AQI Seasonality – PM 2.5

Various: High in winter

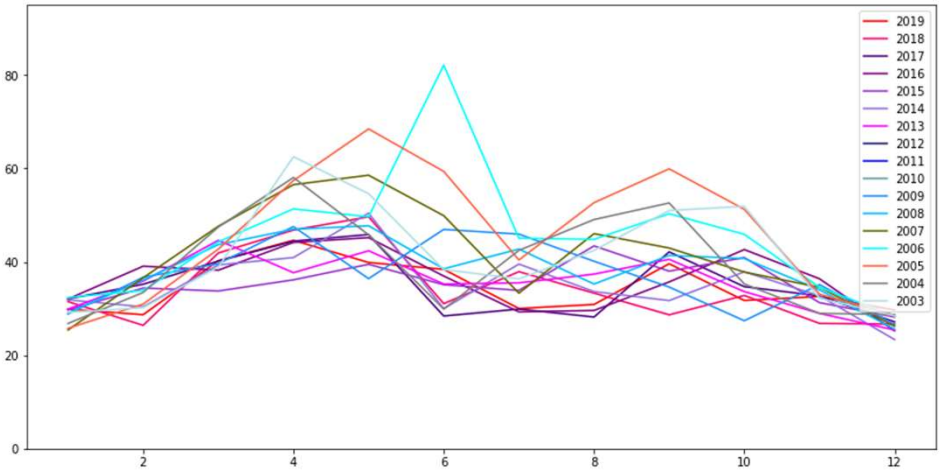




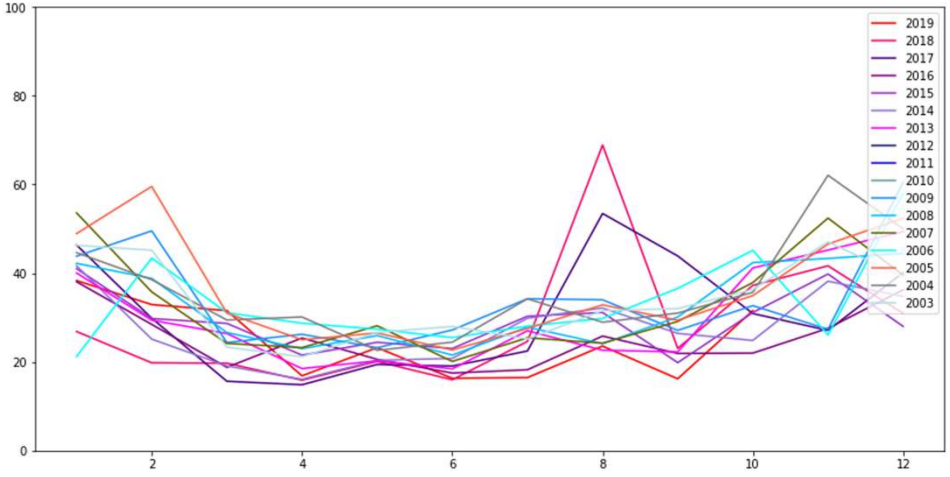
AQI Spikes

In some states, we can spot some cases out of trend or seasonality

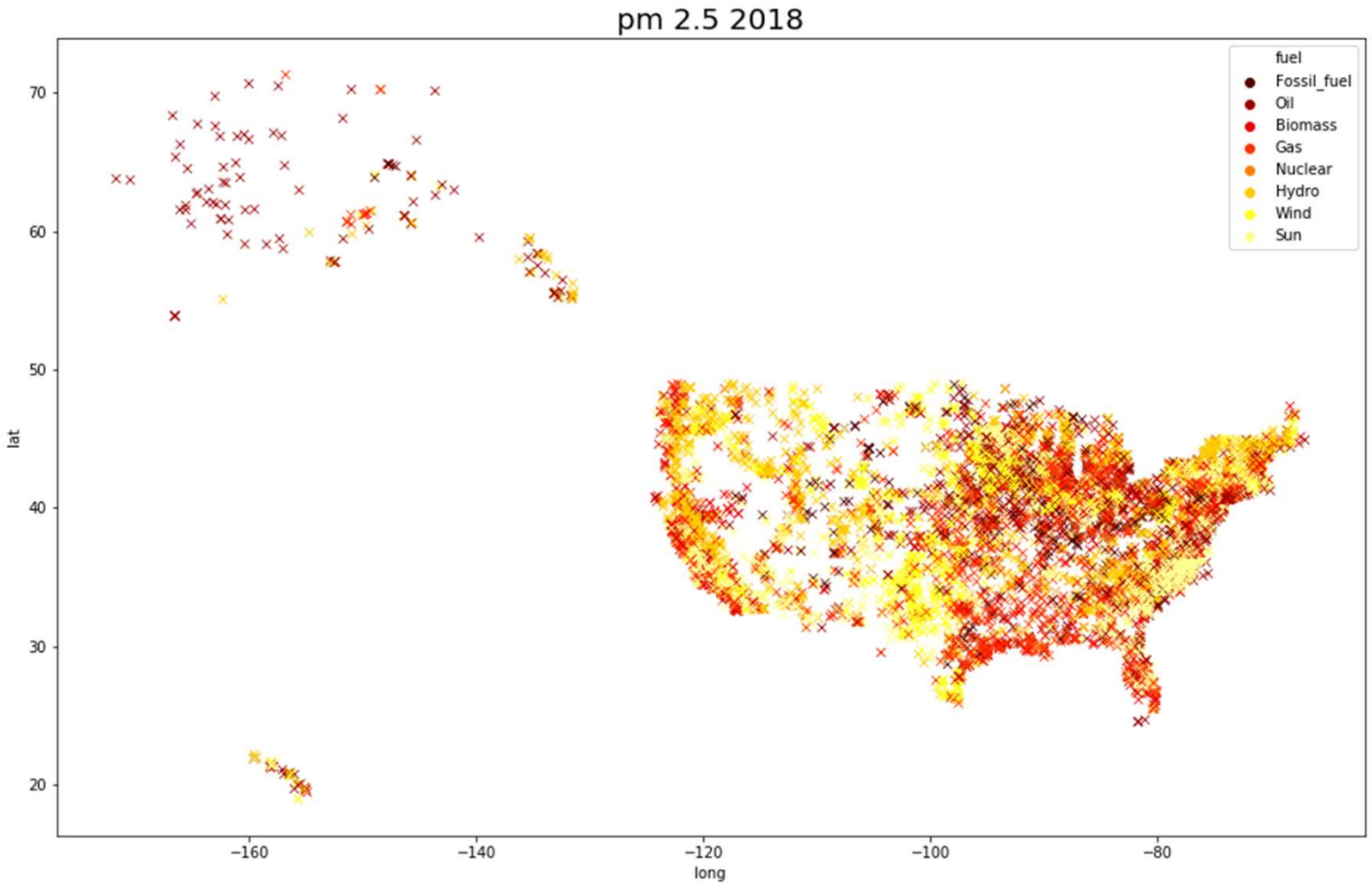
Louisiana- Ozone seasonality



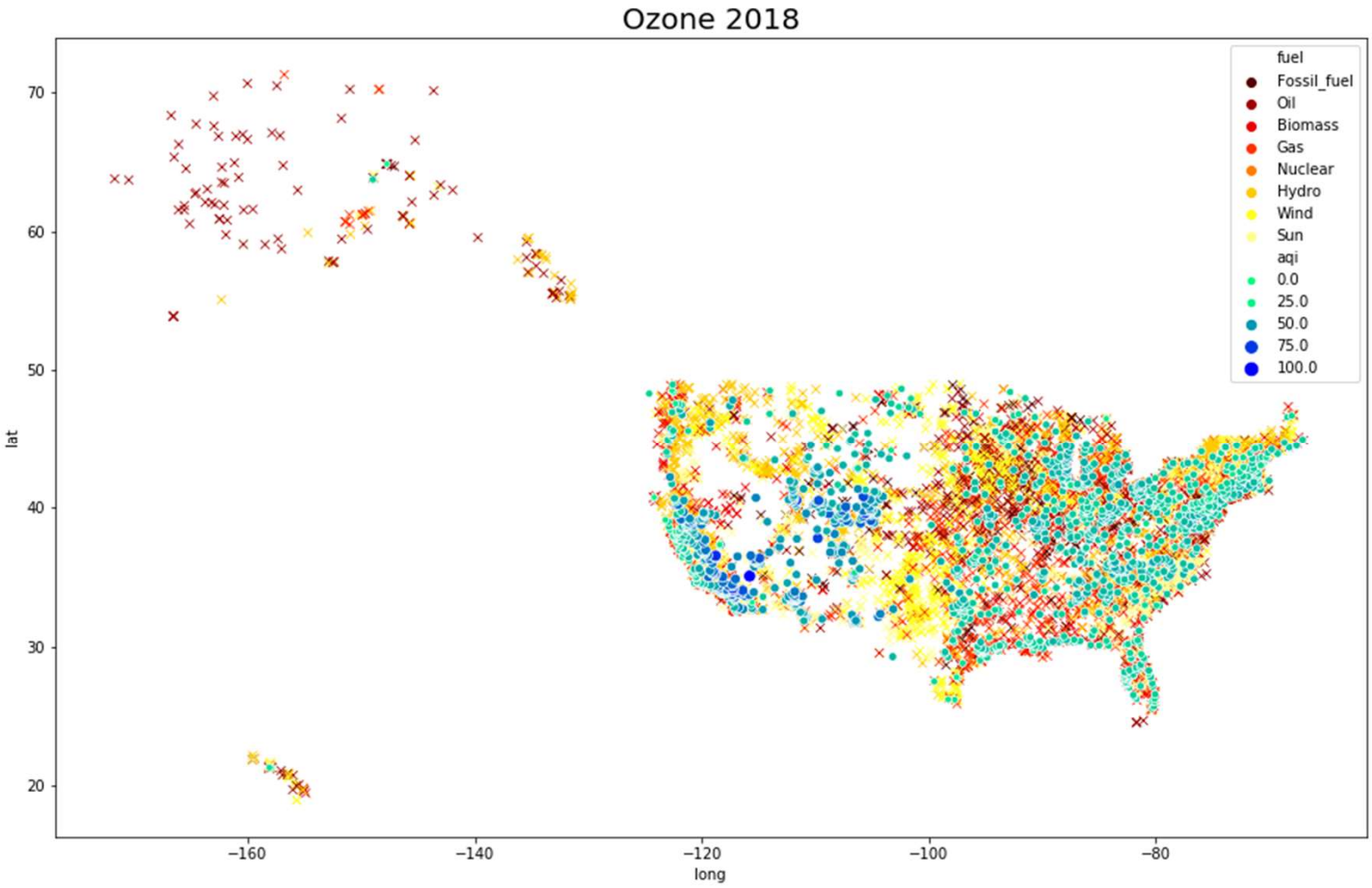
Washington- PM 2.5 seasonality



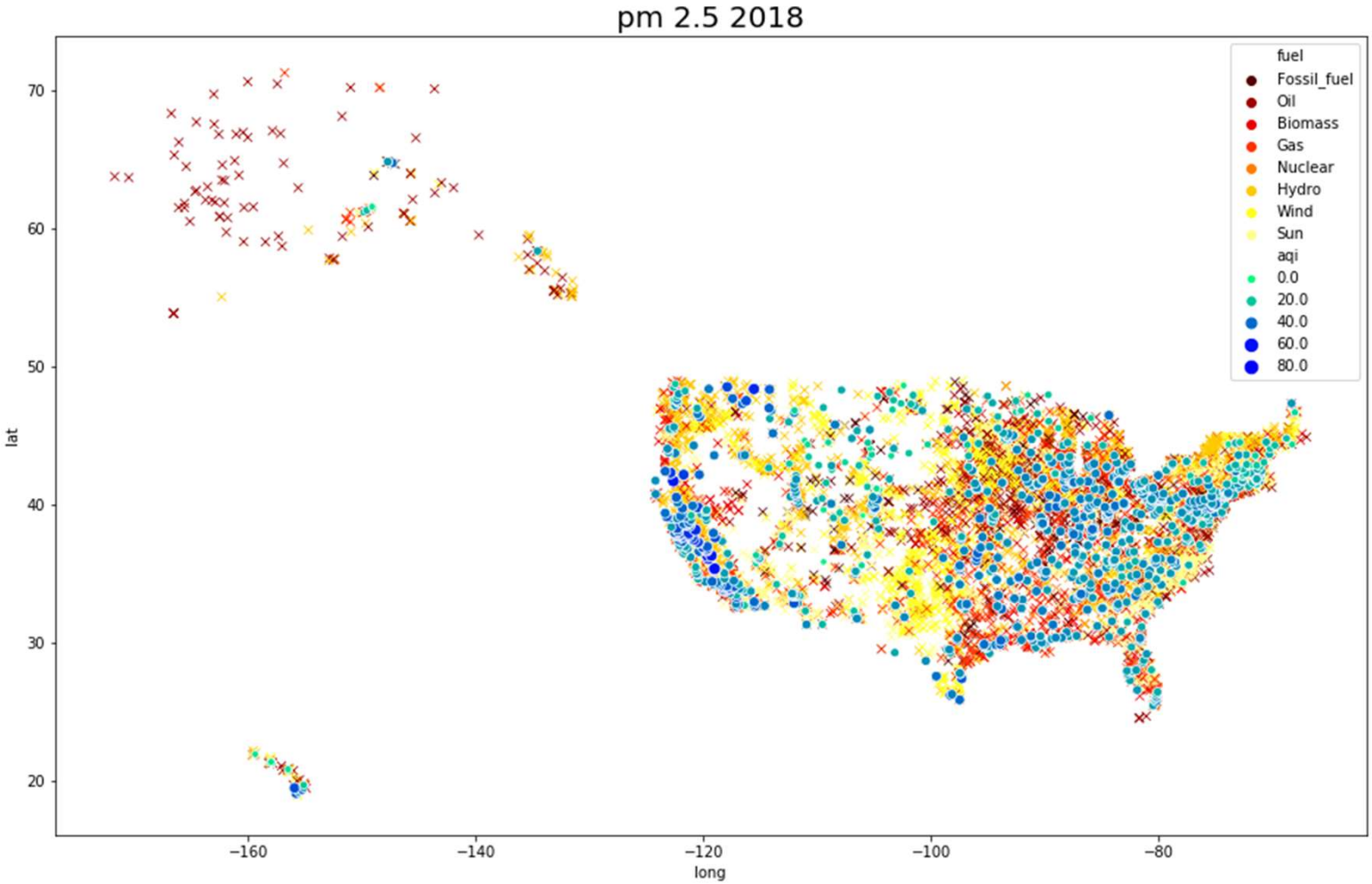
Data : Power plants



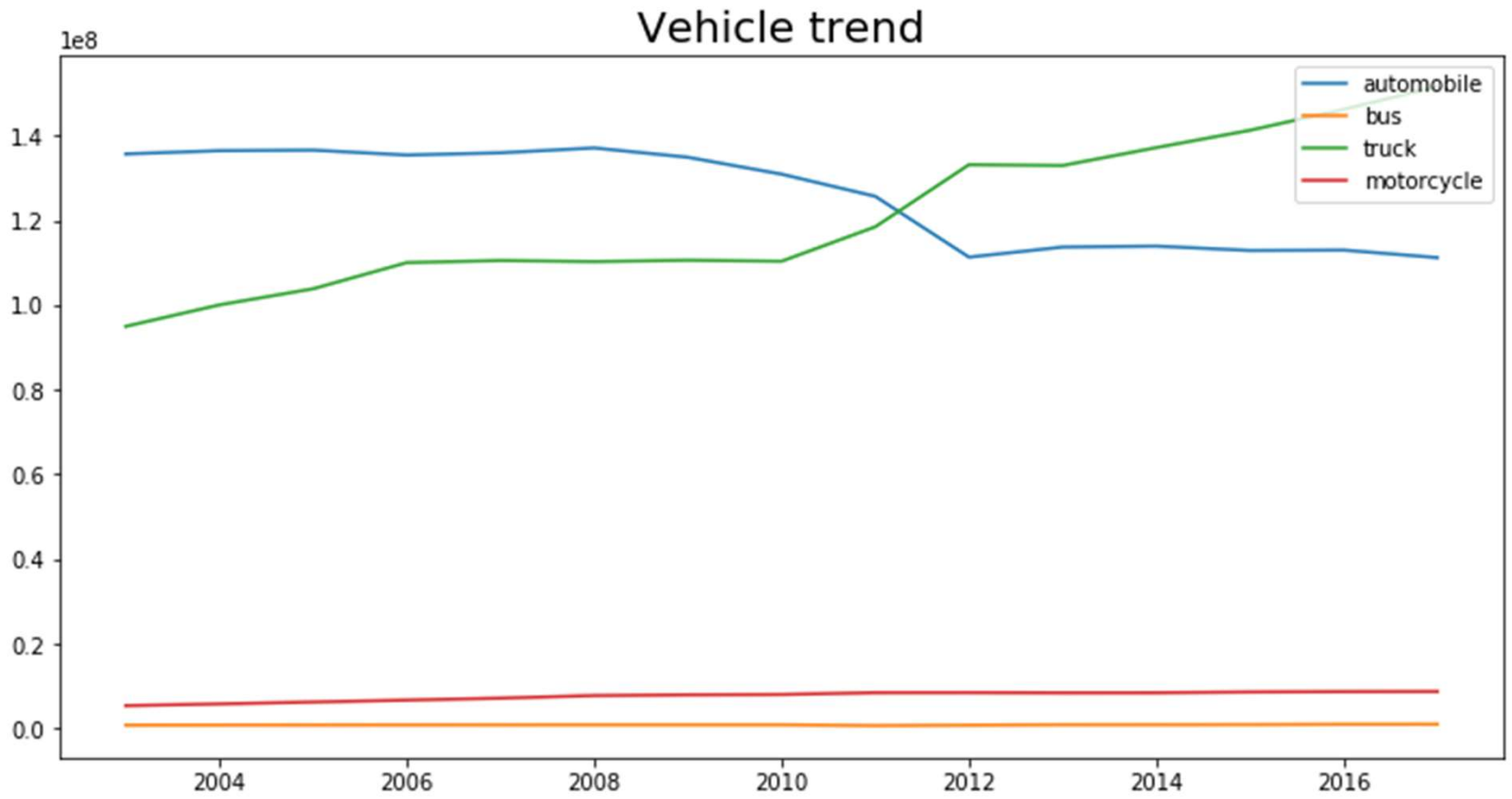
Data : Power plants



Data : Power plants



Data : Number of cars



Modeling

Modeling – Scope : Air type and Region

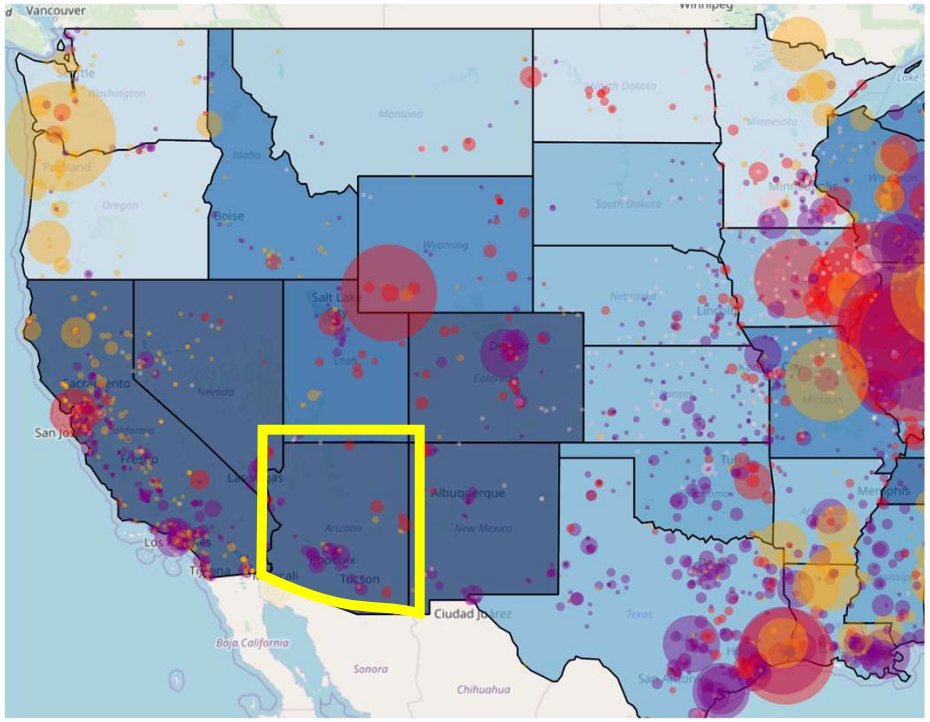
Why Airzona? Consistently bad.

[year 2003]

state	aqi
Utah	67.688889
California	55.474801
Arizona	54.173650
Wyoming	53.561369
Indiana	53.206009
Missouri	52.292199
Rhode Island	52.178523
North Carolina	50.388363
New Mexico	49.798825
Ohio	48.976968
Tennessee	48.962620

[year 2019]

state	aqi
New Mexico	49.022079
Arizona	47.054842
Colorado	45.722382
Wyoming	44.197719
Utah	44.079600
Nevada	43.985640
California	43.846660
Connecticut	43.492317
Maryland	43.211863
Tennessee	42.168118
South Carolina	42.005270



Modeling – Scope : Air type and Region (cont.)

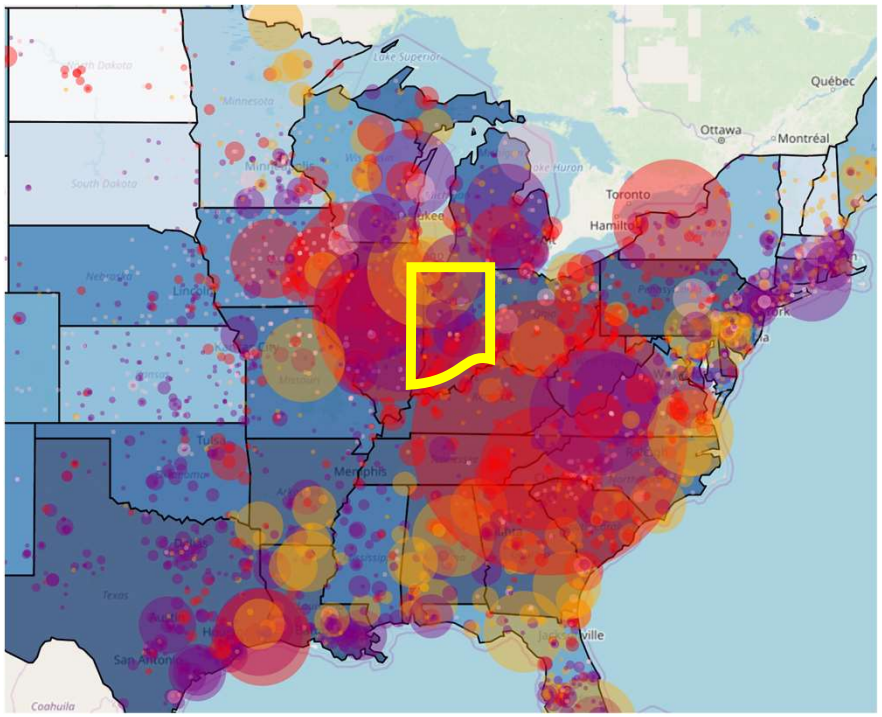
Why Indiana? Consistently bad and has power plants in all directions

[year 2003]

state	aqi
Ohio	54.397482
Georgia	53.552855
West Virginia	53.454940
Indiana	53.090788
Pennsylvania	52.167517
District Of Columbia	52.160822
Kentucky	52.000348
Alabama	51.951200
Illinois	51.896351
Tennessee	51.658869

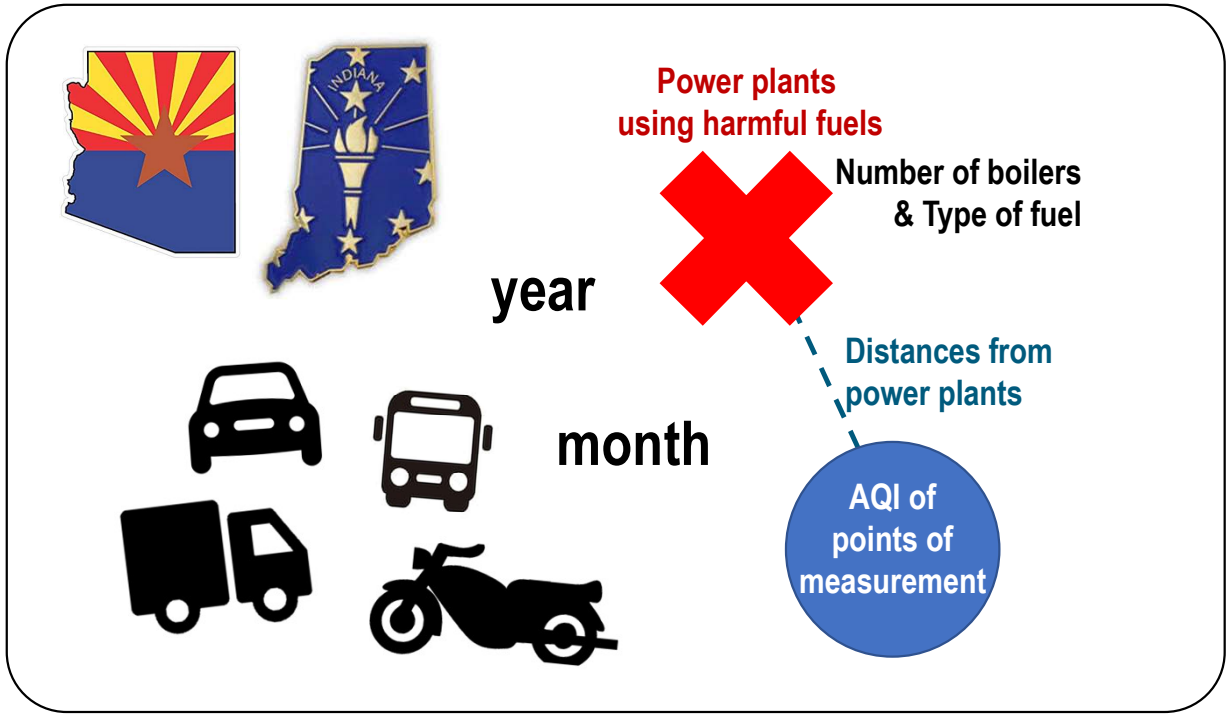
[year 2019]

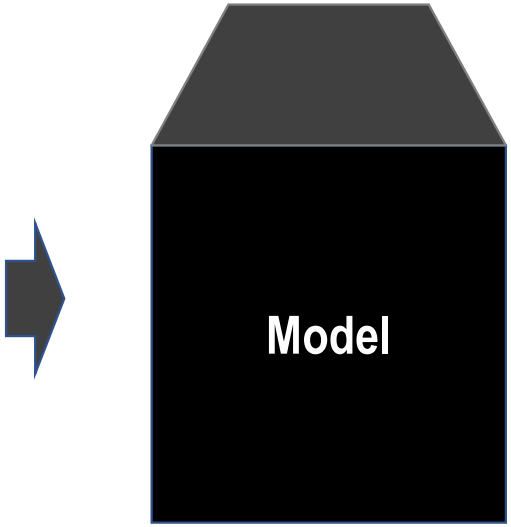
state	aqi
Georgia	37.320868
Alabama	36.318116
Illinois	35.893172
Mississippi	35.665475
Ohio	35.644599
Indiana	35.317011
Arkansas	35.301587
Pennsylvania	35.012357
Texas	34.922466
Louisiana	34.462083



Modeling - stage 1

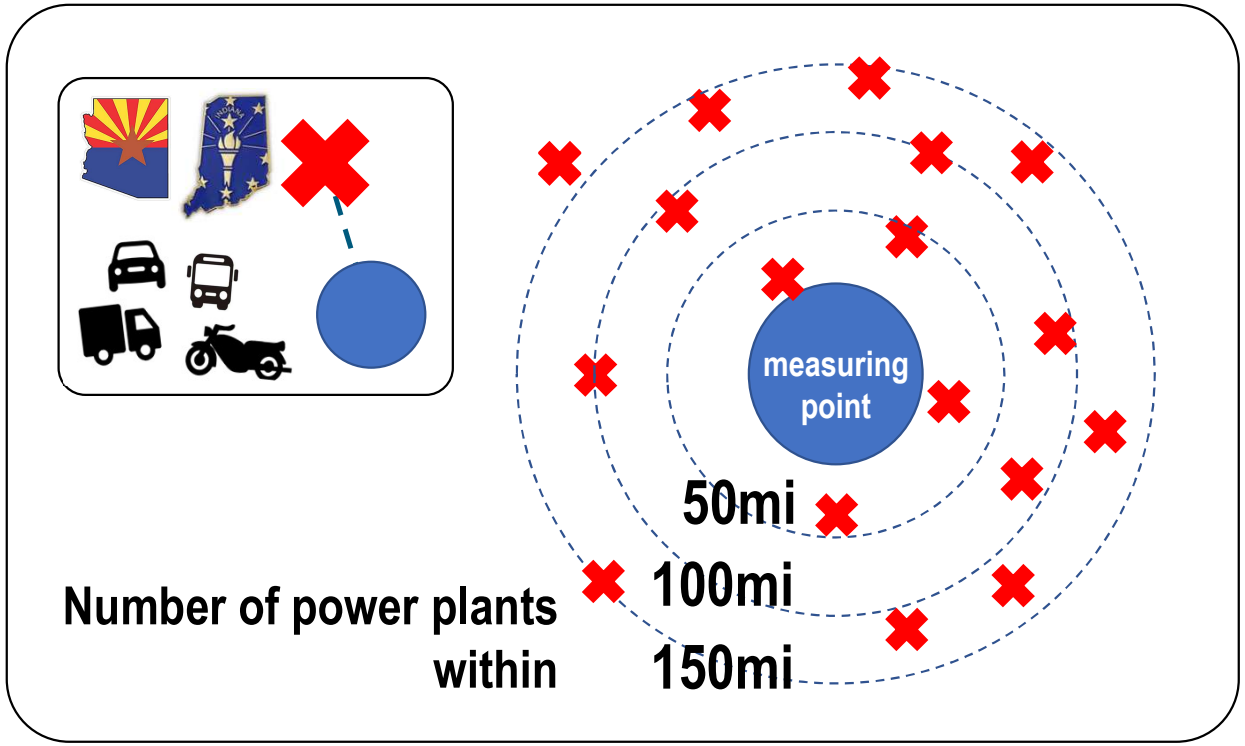
- Add every feature as raw and try every model available : score was not satisfactory





Modeling – stage 2

- Did more science – got the test score as 0.9998



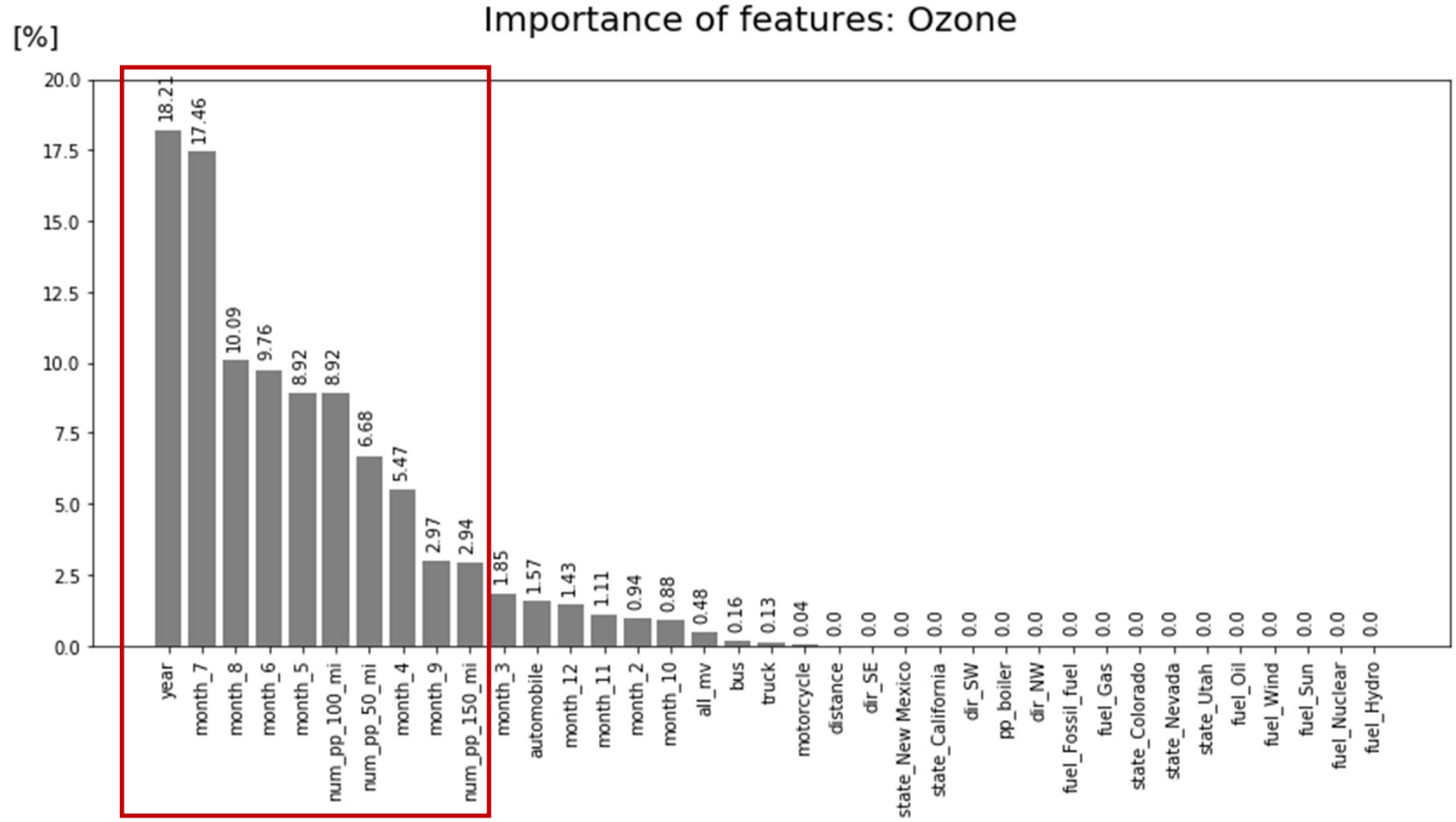
Modeling Result

- **Used modeling methods**

Model	Train score	Test score
Linear Regression	0.6041810022696907	0.6089746235876808
RidgeCV	0.5750836706873368	0.5803638683822172
LassoCV	0.4955812124391689	0.4991269593577929
Decision Tree Regressor	0.9998205259084376	0.9998171539806865
Random Forest Regressor	0.9998209339494756	0.9998174591436612
PCA	0.5750851470121187	0.5803588451446455

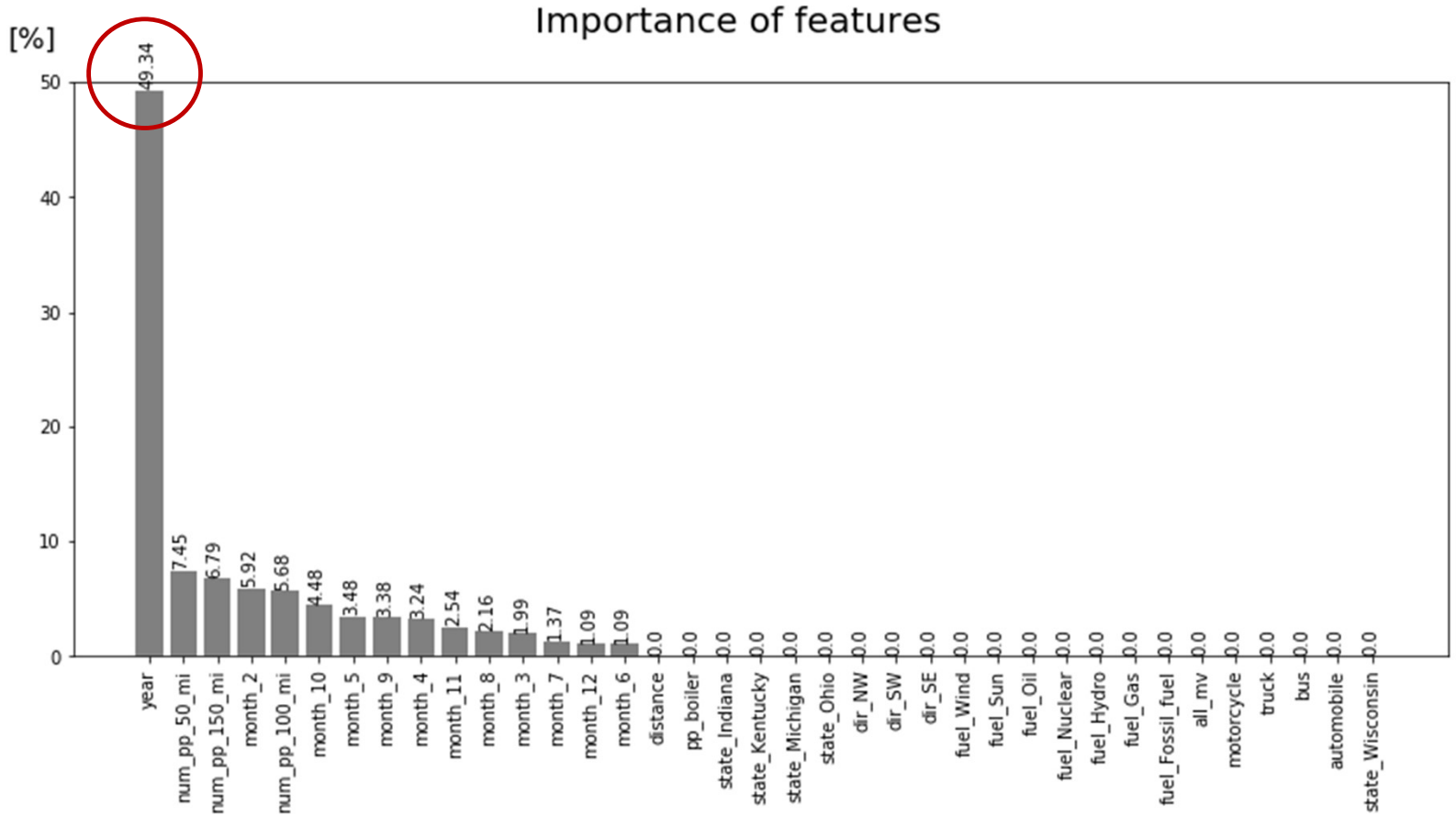
Modeling Result – Ozone

- **The Ozone AQI showed strong seasonality and it's reflected on the model**



Modeling Result – PM 2.5

- **The importance of 'year' feature is the highest → evidence of regulations of States**



Modeling Result – outside research

REPORTS & MULTIMEDIA / EXPLAINER

The Clean Air Act

Published Nov 18, 2010 | Updated Feb 1, 2012

The Clean Air Act—whose basic structure was established in 1970, and then amended in 1977 and 1990—is a United States federal law designed to protect human health and the environment from the effects of air pollution.

Under the Clean Air Act, the Environmental Protection Agency (EPA) is required to regulate emission of pollutants that "endanger public health and welfare." State and local governments also monitor and enforce Clean Air Act regulations, with oversight by the EPA.

This legislation has a 40-year track record of cutting dangerous pollution and has prevented more than 400,000 premature deaths and hundreds of millions of cases of respiratory and cardiovascular disease.

- **Regulations of government**

The Clean Air Act has helped:

- cut ground-level ozone, a dangerous component of smog, by more than 25 percent since 1980;
- reduce mercury emissions by 45 percent since 1990;
- reduce the main pollutants that contribute to acid rain, sulfur dioxide and nitrogen dioxide, by 71 percent and 46 percent, respectively since 1980;
- phase out the production and use of chemicals that contribute to the hole in the ozone layer; and
- reduce the lead content in gasoline, which has cut lead air pollution by 92 percent since 1980.

[Source: https://www.ucsusa.org/resources/clean-air-act]

Modeling Result – outside research

For more than forty-five years the Clean Air Act has cut pollution as the U.S. economy has grown.

- Experience with the Clean Air Act since 1970 has shown that protecting public health and building the economy can go hand in hand.
- Clean Air Act programs have lowered levels of six common pollutants -- particles, ozone, lead, carbon monoxide, nitrogen dioxide and sulfur dioxide -- as well as numerous toxic pollutants.

New cars, trucks, and nonroad engines use state-of-the-art emission control technologies.

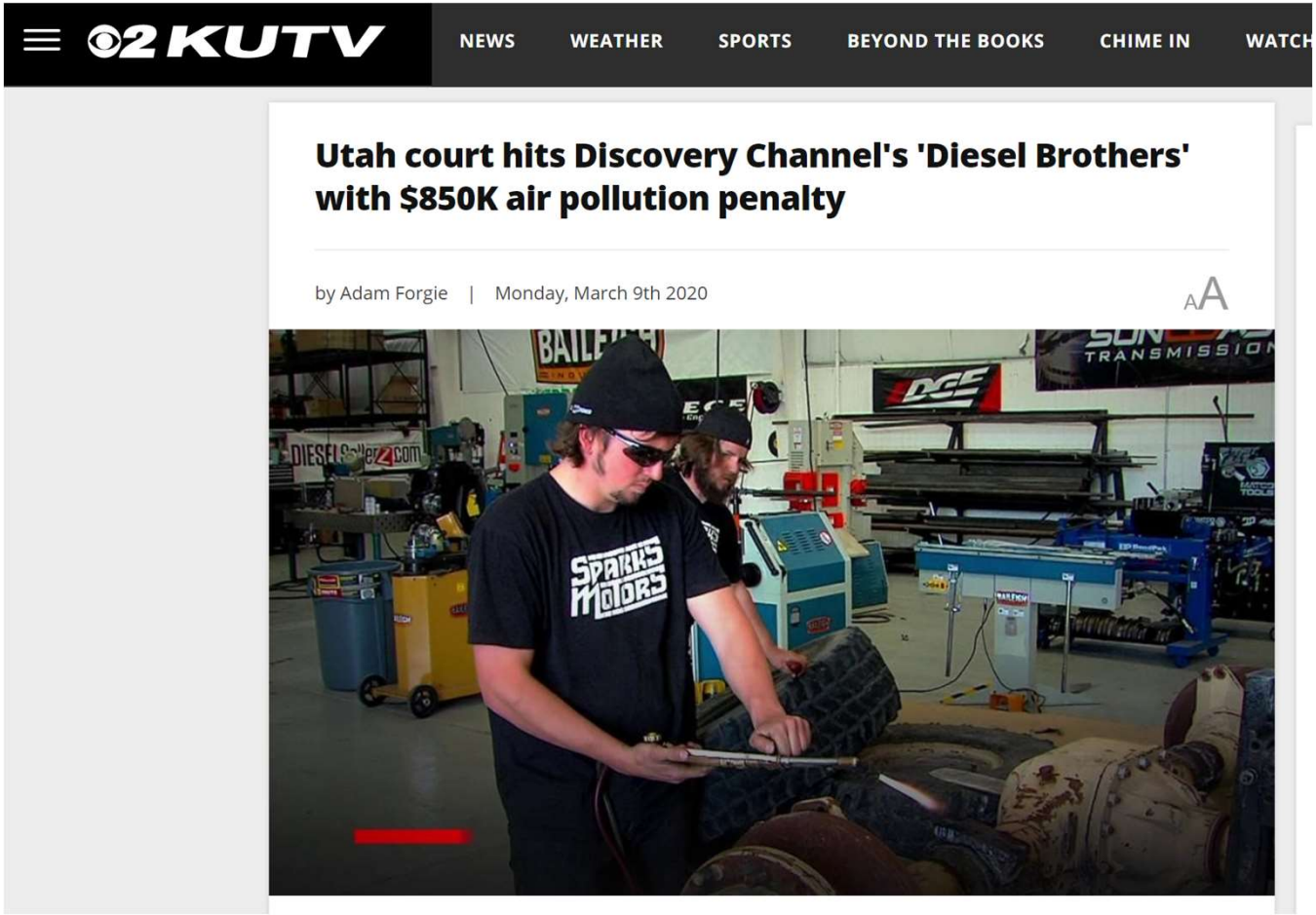
EPA has required dramatic reductions in emissions from new motor vehicles and non-road engines - such as those used in construction, agriculture, industry, trains and marine vessels -- through standards that require a combination of cleaner engine technologies and cleaner fuels. In 2013, EPA estimated the benefits of five key standards to cut emissions from vehicles, engines and fuel to 2030.

New power plants and factories use modern pollution control technology.

- The Act requires that when new industrial facilities are designed and built, good pollution control must be part of the design. This means that as new, cleaner facilities are built, the country's industrial base becomes cleaner overall. Public health is protected as economic growth proceeds.

[Source: https://https://www.epa.gov/clean-air-act-overview/progress-cleaning-air-and-improving-peoples-health]

Modeling Result – outside research



[Source: https://kutv.com/news/local/utah-court-hits-discovery-channels-diesel-brothers-with-850k-air-pollution-penalty]

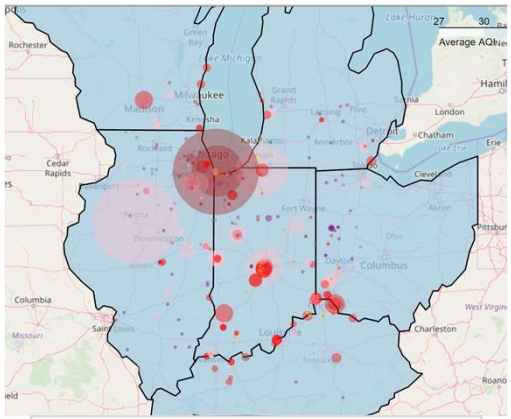
Simulation

How it works

1. Input conditions

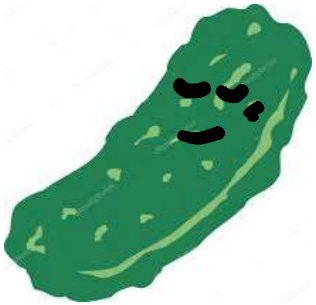
Air type
State
Year
Month
Number of power plants within 50 miles
Number of power plants between 50 & 100 miles
Number of power plants between 100 &150 miles

2. Resample data for the measuring point



Power plants within 50 miles
Power plants between 50 & 100 miles
Power plants between 100 &150 miles

3. Simulate with a pickled model

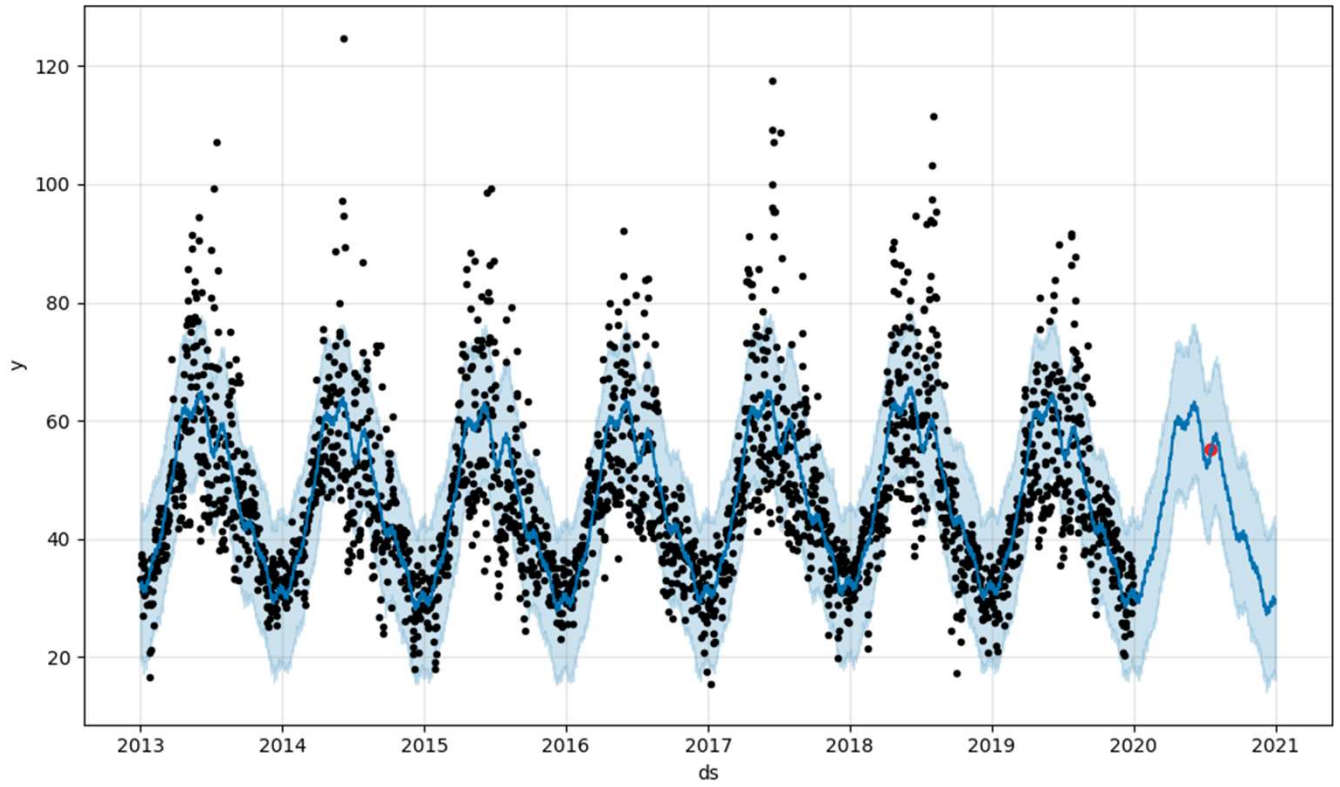


Predicted

AQI

Missed from simulation

- Flask displays only preloaded images and doesn't allow to show dynamically created images.



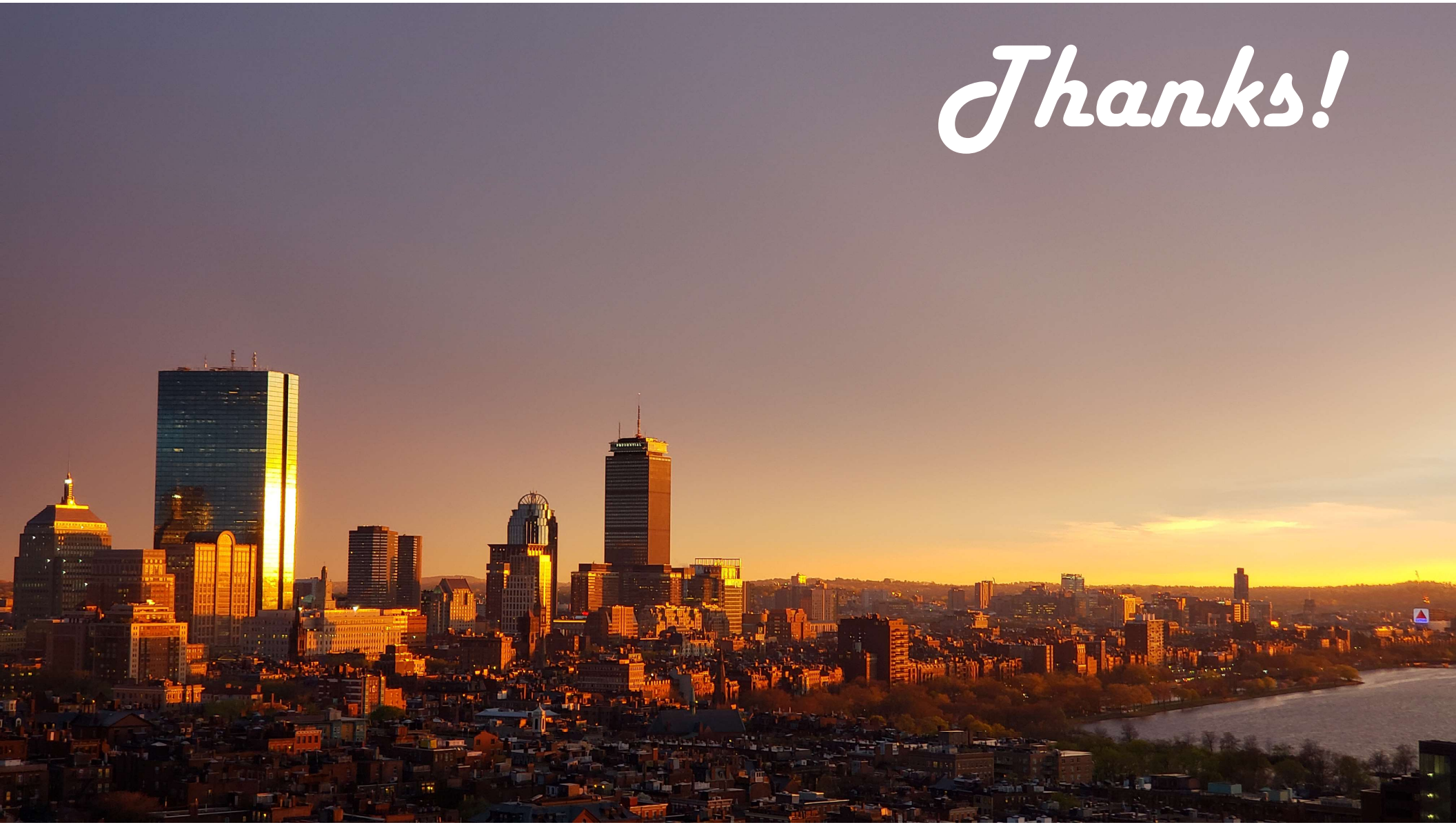
Next step

- Expand the scope to all states
- In simulation result, add guideline for AQI application
- Do more researches
- Upload simulator to AWS
- Try ‘django’ instead of ‘Flask’

References

- Special thanks to Joe Fitzgerald – introduced the theme and helped getting data!!!
- AQI data: https://aqs.epa.gov/aqsweb/documents/data_api.html
- Air quality index and explanation : https://airnow.gov/index.cfm?action=aqi_brochure.index
- Population data: https://www.census.gov/data/tables/time-series/demo/popest/2010s-state-total.html
- Number of registered cars : https://www.fhwa.dot.gov/policyinformation/quickfinddata/qfvehicles.cfm
- Lung cancer data: https://wonder.cdc.gov/cancer-v2016.HTML
- Power plant data:

  https://www.epa.gov/energy/emissions-generation-resource-integrated-database-egrid


Thanks!