New Pollution

TEAM : DC20035
MEMBER: Zhenyang Wang, Yanzhi Shen

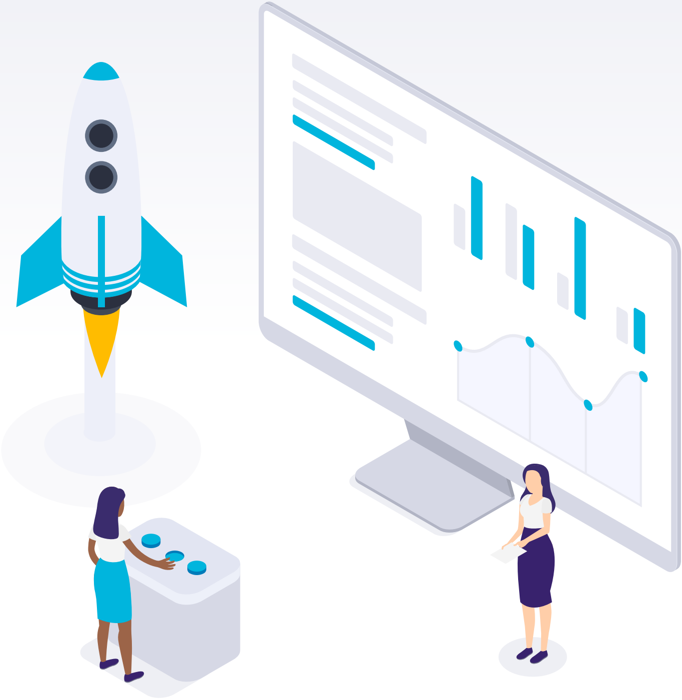
Content

- Issue: Air Pollution
- Data Sources
- Data Patterns: Panel
- Predictive Models:
  - GBDT Regression
  - GBDT Classification
- Main Findings
- Solutions
- Future Study
- Q&A

1

Sustainable Issue:

Air Pollution

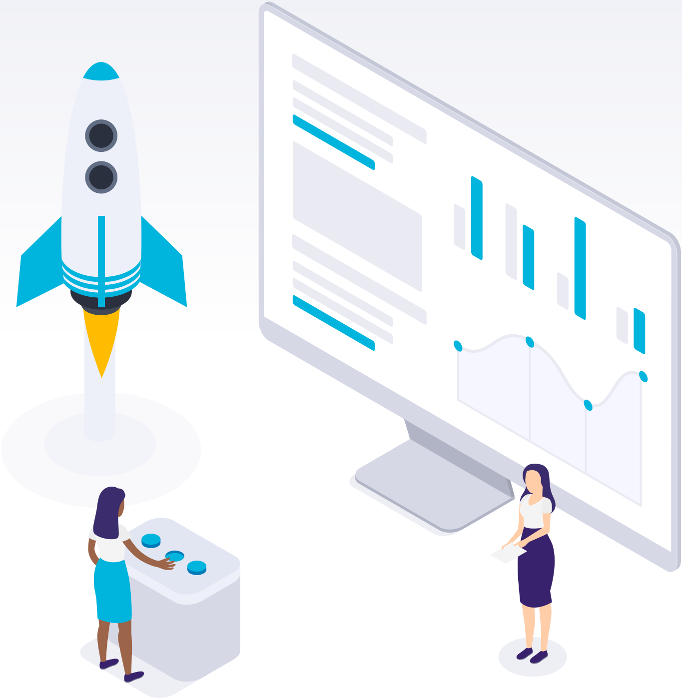


Air Pollution

Outdoor air quality affects public health both directly and indirectly, and it also affects natural and built resources.

2

Data Sources



Original Data

Date Local	NO2 Full Mean (stan	NO2 Full 1st Max Va	NO2 Full 1st Max Ho	O3 Full Mean (stand	O3 Full 1st Max Valu	O3 Full 1st Max Hou	SO2 Full Mean (stan	SO2 Full 1st Max Va	SO2 Full 1st Max Ho	CO Full Mean (stand	CO Full 1st Max Valu	CO Full 1st Max
1/1/2000	26.541667	39	17	0.004167	0.014	23	11.25	33	15	2.0625	4.1	
1/1/2000	26.541667	39	17	0.004167	0.014	23	11.25	33	15	2.005263	3.6	
1/1/2000	26.541667	39	17	0.004167	0.014	23	11.225	19	17	2.0625	4.1	
1/1/2000	26.541667	39	17	0.004167	0.014	23	11.225	19	17	2.005263	3.6	
1/2/2000	14.958333	22	0	0.023083	0.029	11	7.791667	11	0	0.658333	2.6	
1/2/2000	14.958333	22	0	0.023083	0.029	11	7.791667	11	0	0.883333	2.1	
1/2/2000	14.958333	22	0	0.023083	0.029	11	7.7625	11	2	0.658333	2.6	
1/2/2000	14.958333	22	0	0.023083	0.029	11	7.7625	11	2	0.883333	2.1	
1/3/2000	18.291667	32	17	0.014875	0.026	20	7.26087	11	11	0.6	2.4	1
1/3/2000	18.291667	32	17	0.014875	0.026	20	7.26087	11	11	0.566667	0.9	1
1/3/2000	18.291667	32	17	0.014875	0.026	20	6.971429	9.6	14	0.6	2.4	1
1/3/2000	18.291667	32	17	0.014875	0.026	20	6.971429	9.6	14	0.566667	0.9	1
1/4/2000	13.363636	25	18	0.01725	0.024	0	5.416667	10	19	0.504167	0.8	1
1/4/2000	13.363636	25	18	0.01725	0.024	0	5.416667	10	19	0.5125	0.9	
1/4/2000	13.363636	25	18	0.01725	0.024	0	5.3875	8.6	23	0.504167	0.8	1
1/4/2000	13.363636	25	18	0.01725	0.024	0	5.3875	8.6	23	0.5125	0.9	
1/5/2000	16.625	33	21	0.009583	0.025	3	8.375	12	11	0.6625	1.9	2
1/5/2000	16.625	33	21	0.009583	0.025	3	8.375	12	11	[illegible]70833	1	2

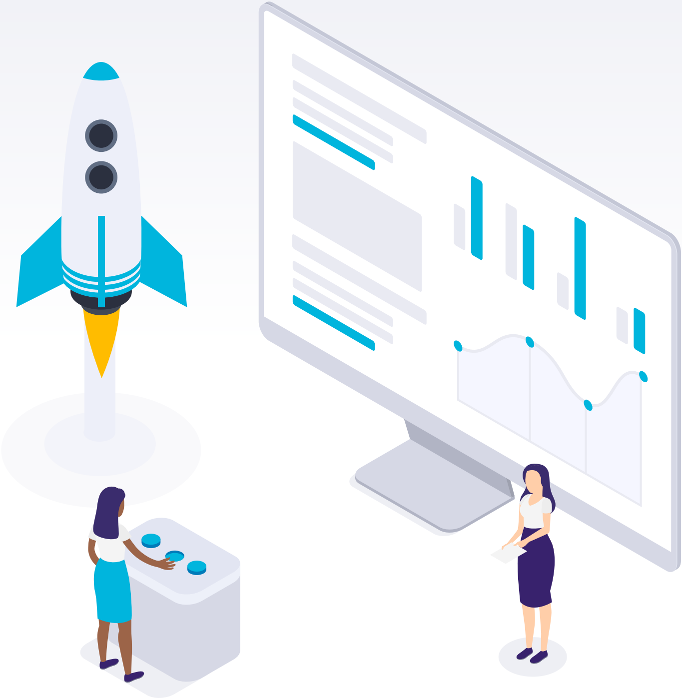
1. State
2. Date
3. NO2/SO2/03/C0 Units: Multiplier for NO2/SO2/03/C0
4. NO2/SO2/03/C0 Mean Full: Mean yield of the molecule Parts Per Billion or Million for the day --> Full Mean (standard)
5. NO2/SO2/03/C0 1st Max Value Full: Max value of the molecule Parts Per Billion or Million for the day --> 1st Max Value Full (standard)
6. NO2/SO2/03/C0 1st Max Hour Full: The hour that contains the max value of the molecule Parts Per Billion or Million for the day --> 1st Max Hour (standard)

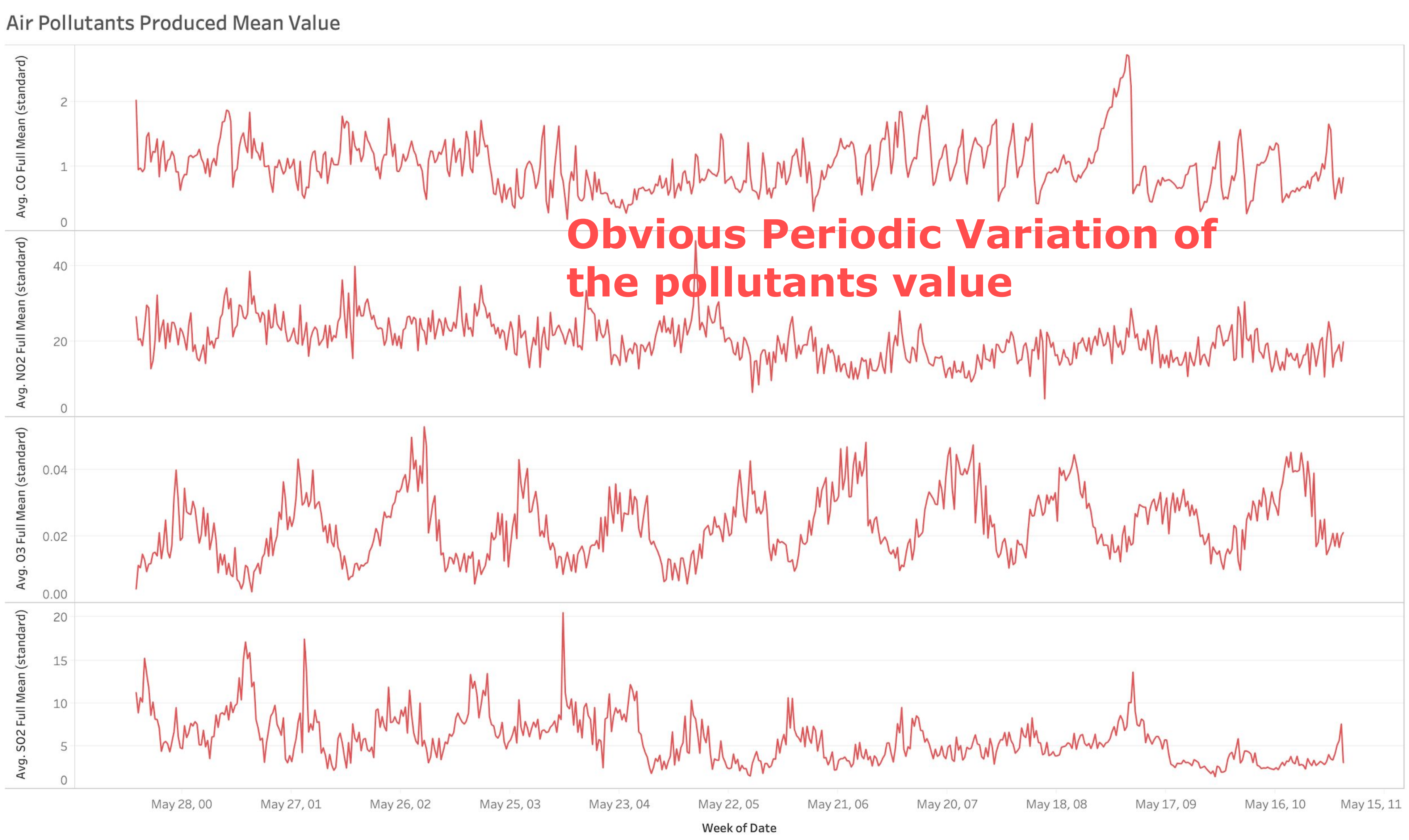
Supplementary Data

1. PRCP: Precipitation
2. SNOW: Snowfall
3. TMAX: Maximum temperature
4. TMIN: Minimum temperature
5. Year: from 2000-2010
6. Month: Jan-Dec(1-12)
7. Weekday: Monday - Sunday(0-6)
8. Holiday: US public holiday
9. before_holiday_7: Within 7 days before public holiday
10. after_holiday_7: Within 7 days after public holiday

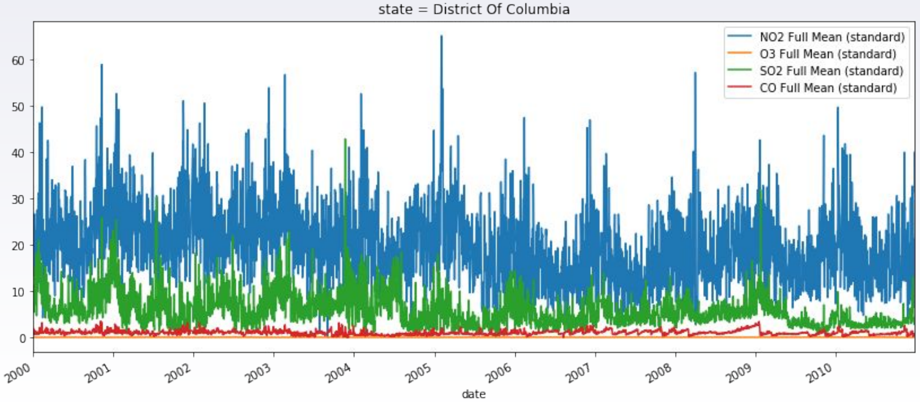
3

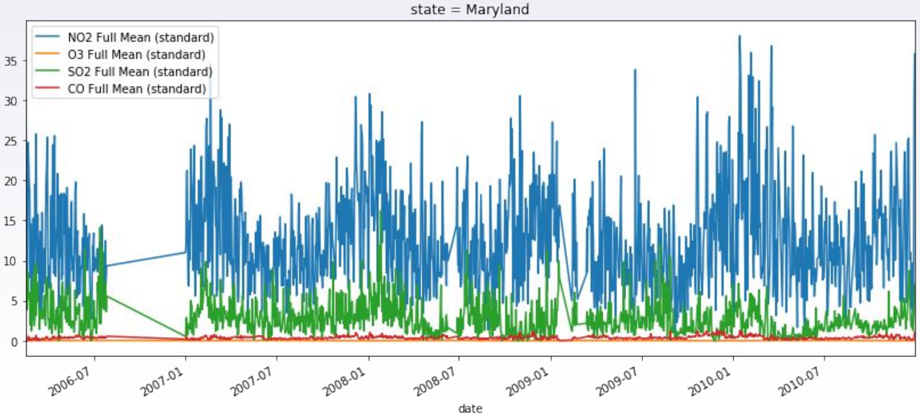
Data Patterns: Panel

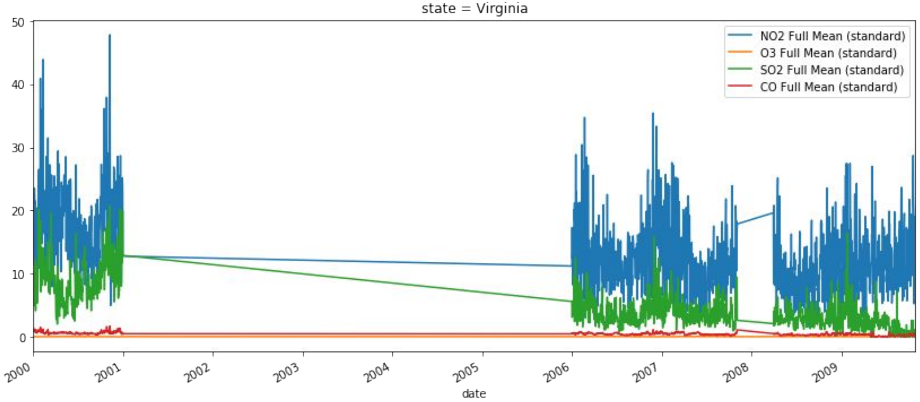




Air Pollutants Produced Mean Value
Avg. CO Full Mean (standard)
0
1
2
Avg. NO2 Full Mean (standard)
0
20
40
Avg. O3 Full Mean (standard)
0.00
0.02
0.04
Avg. SO2 Full Mean (standard)
0
5
10
15
20
May 28, 00
May 27, 01
May 26, 02
May 25, 03
May 23, 04
May 22, 05
May 21, 06
May 20, 07
May 18, 08
May 17, 09
May 16, 10
May 15, 11
Week of Date
Obvious Periodic Variation of the pollutants value







WHY WE ONLY CHOOSE DC?

Feature Engineering for Time-Series

1. **Date Time feature**

   (Year, Month,Week, Day)

2. **Lag feature**

   eg.predict the value at the next time (t+1) given the value at the previous time (t-1).

3. **Expanding Rolling Window features**

   These are a summary of values over a fixed window of prior time steps: eg. Mean, median, Max,Min

Features we can use

Use data today to predict tomorrow!

4

Predictive Model:

Gradient Boosting Regression

Gradient Boosting Regression

Training set: 0.6

Validation set: 0.2

Test set: 0.2

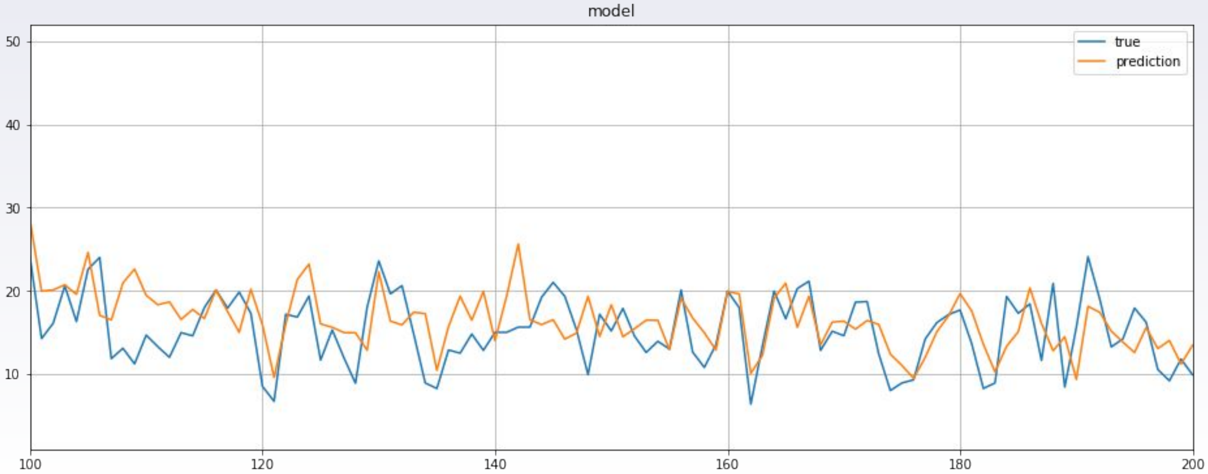
Features:

Date: the current date

+ Label: the predicted date for the next day
+ 64 columns of derived features

Sample NO2

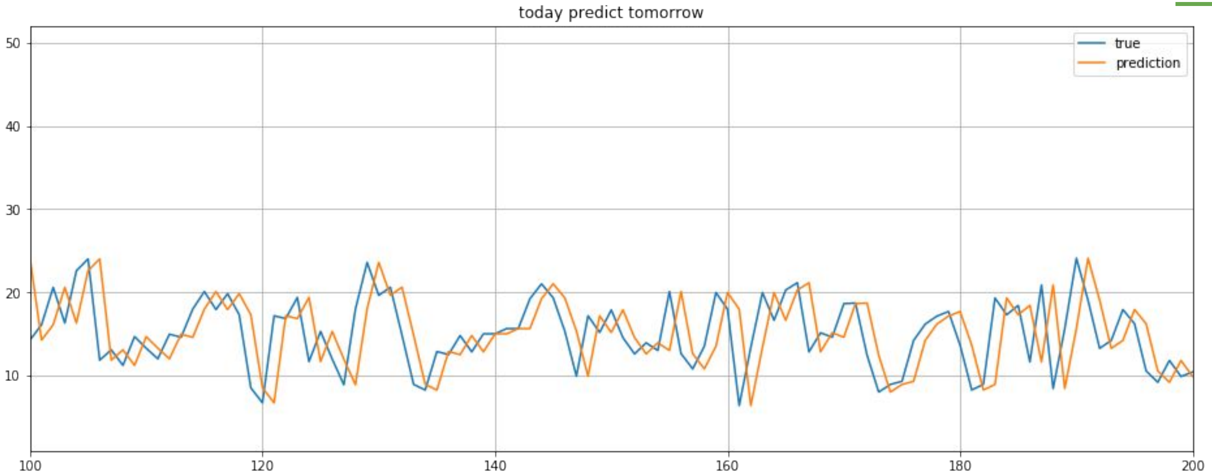
--> choose the top 33 features to predict the test set



RMSE in test set without derived features is:

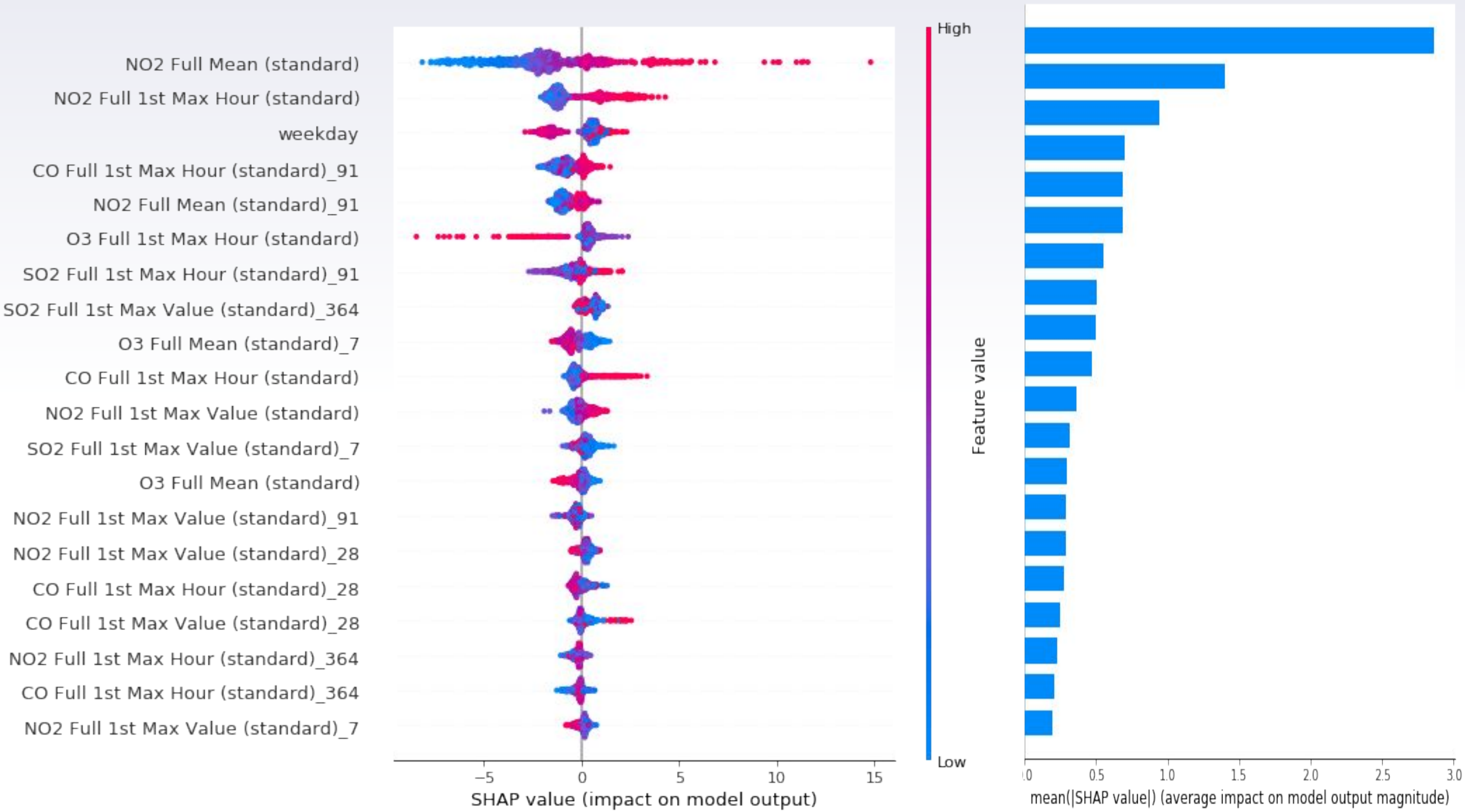
44.30579387724452

Today predicts tomorrow



RMSE in test set with derived features is:

28.06952499617282



NO2 Full Mean (standard)
NO2 Full 1st Max Hour (standard)
weekday
CO Full 1st Max Hour (standard)_91
NO2 Full Mean (standard)_91
O3 Full 1st Max Hour (standard)
SO2 Full 1st Max Hour (standard)_91
SO2 Full 1st Max Value (standard)_364
O3 Full Mean (standard)_7
CO Full 1st Max Hour (standard)
NO2 Full 1st Max Value (standard)
SO2 Full 1st Max Value (standard)_7
O3 Full Mean (standard)
NO2 Full 1st Max Value (standard)_91
NO2 Full 1st Max Value (standard)_28
CO Full 1st Max Hour (standard)_28
CO Full 1st Max Value (standard)_28
NO2 Full 1st Max Hour (standard)_364
CO Full 1st Max Hour (standard)_364
NO2 Full 1st Max Value (standard)_7
−5
0
5
10
15
SHAP value (impact on model output)
High
Feature value
Low
0.5
1.0
1.5
2.0
2.5
3.0
mean(|SHAP value|) (average impact on model output magnitude)

Use data today to predict tomorrow!

5

Predictive Model:

Gradient Boosting Classification

GBDT Binary Classification

Training set: 0.6

Validation set: 0.2

Test set: 0.2

Features:

Date: the current date

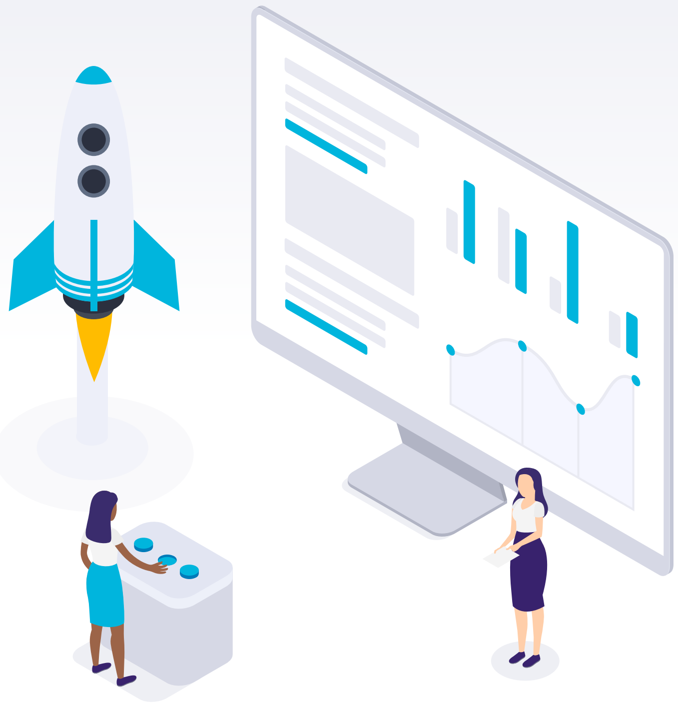
+ Label: the predicted date for the next day
+ 64 columns of derived features

--> choose the top 20 features to predict the test set
--> forecast whether the value will be going up or down on tomorrow using today data

--> **RESULT(Sample NO2):**

	precision	recall	f1-score	support
False	0.68	0.77	0.72	369
True	0.71	0.61	0.66	349
avg / total	0.69	0.69	0.69	718

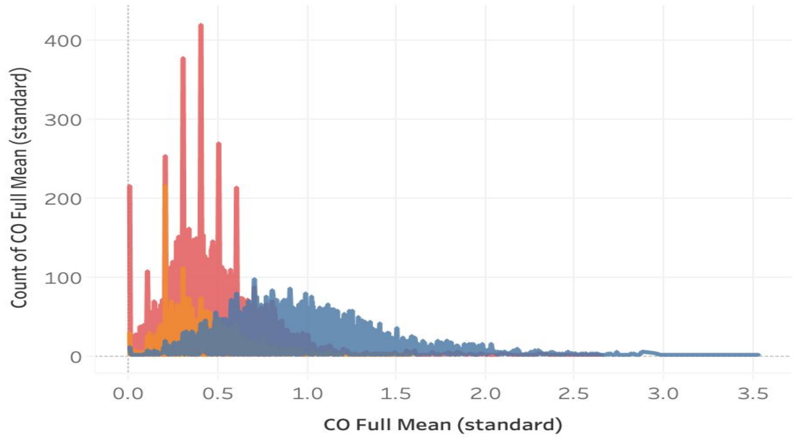
6 Main Findings



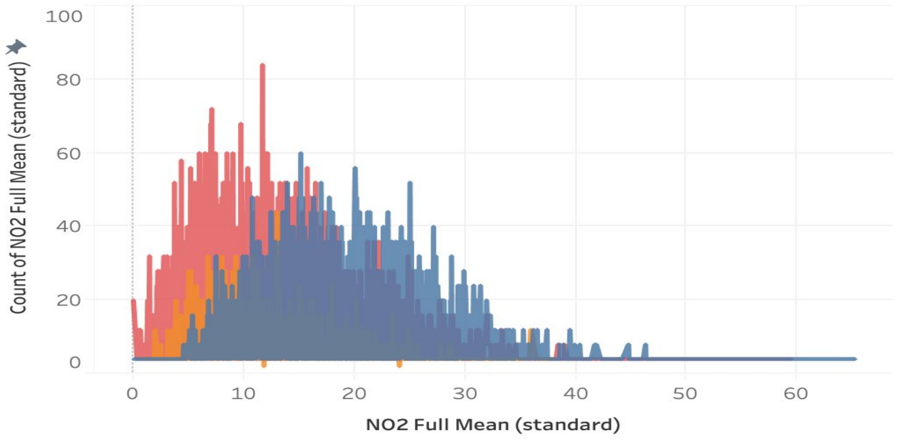
Compare

-- DC, MD and VA

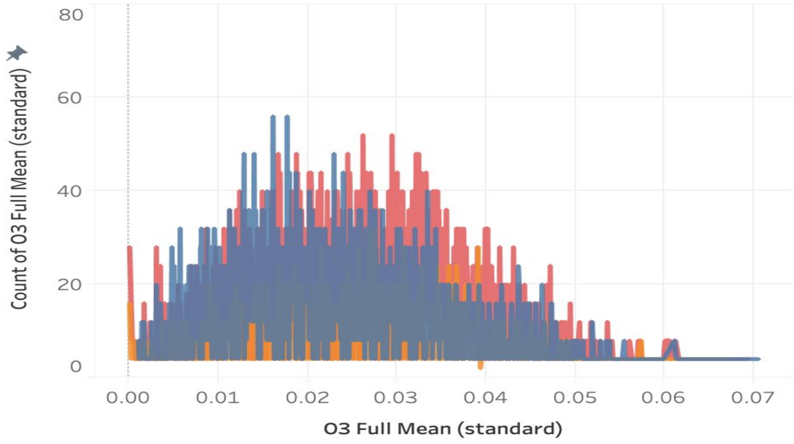
CO Full value Distribution



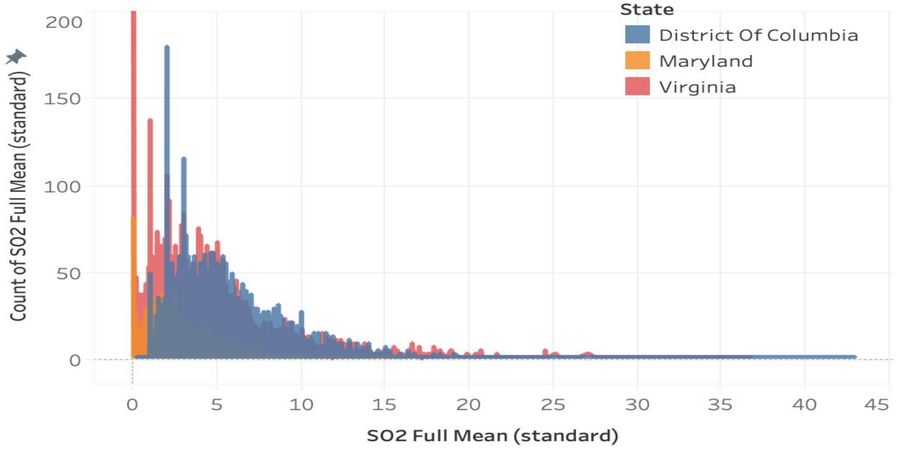
NO2 Full value Distribution



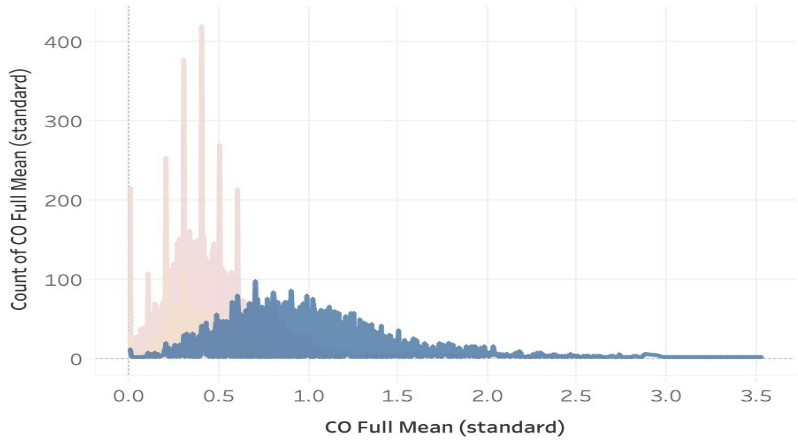
O3 Full value Distribution



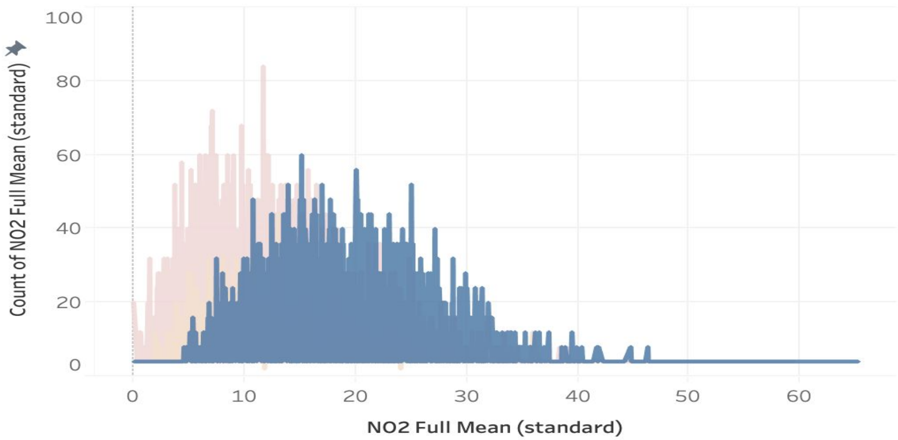
S02 Full value Distribution



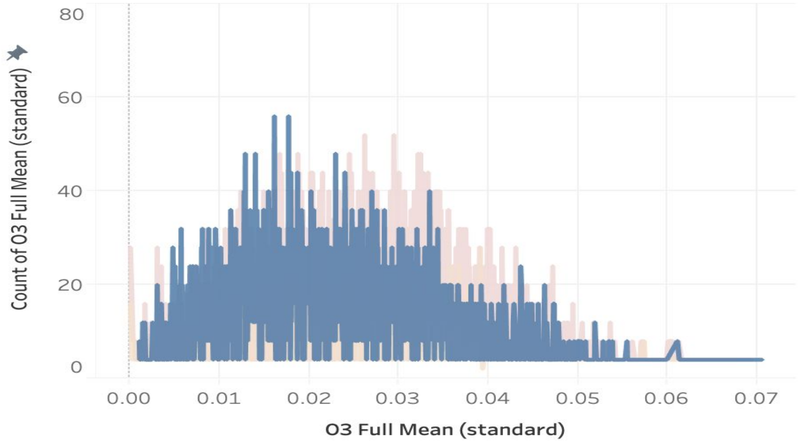
CO Full value Distribution



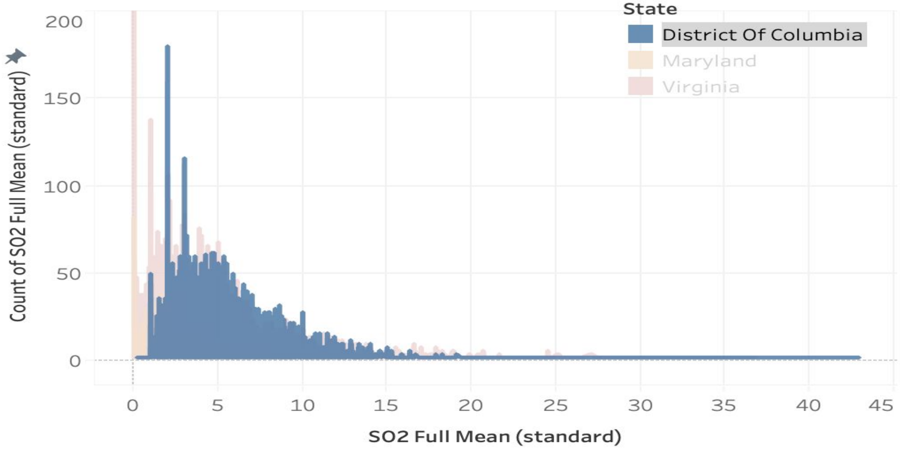
NO2 Full value Distribution



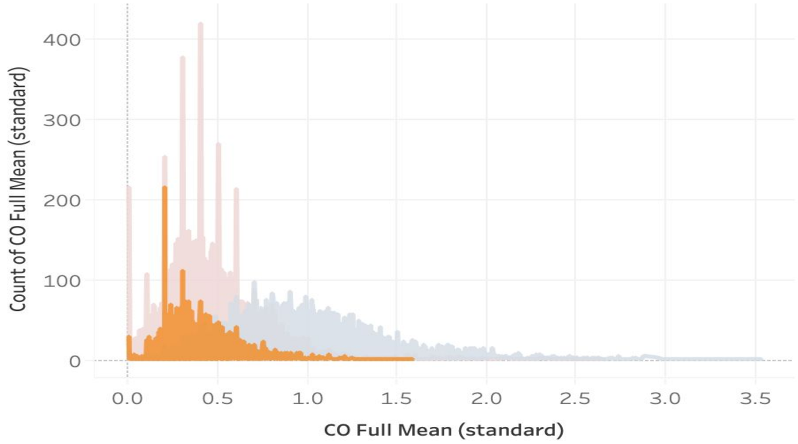
O3 Full value Distribution



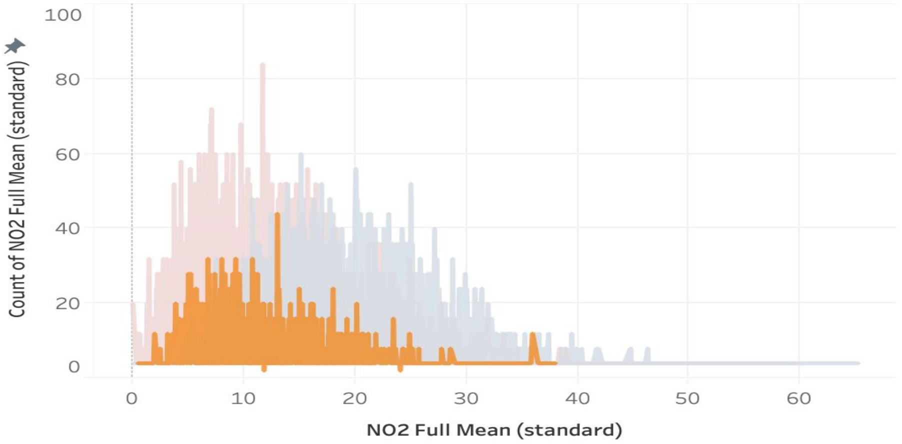
S02 Full value Distribution



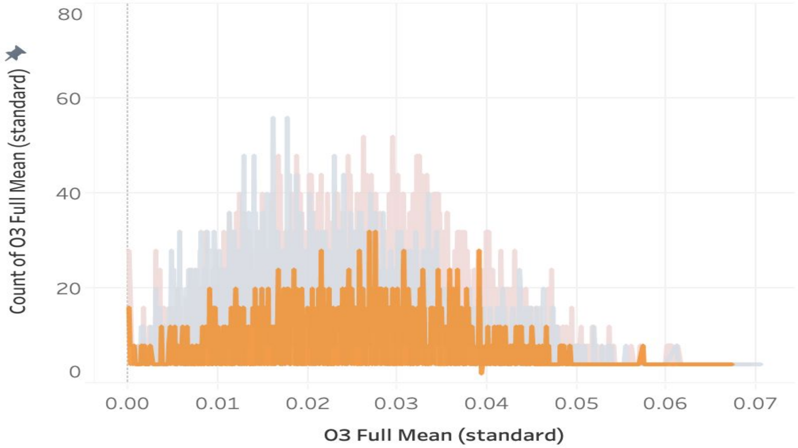
CO Full value Distribution



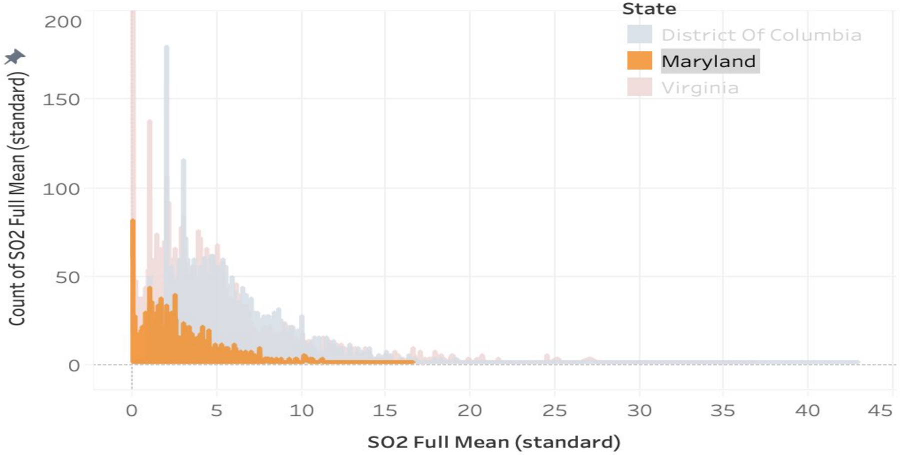
NO2 Full value Distribution



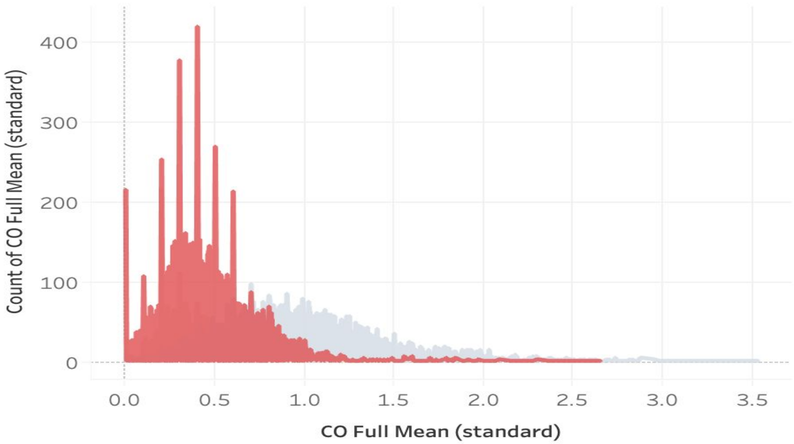
O3 Full value Distribution



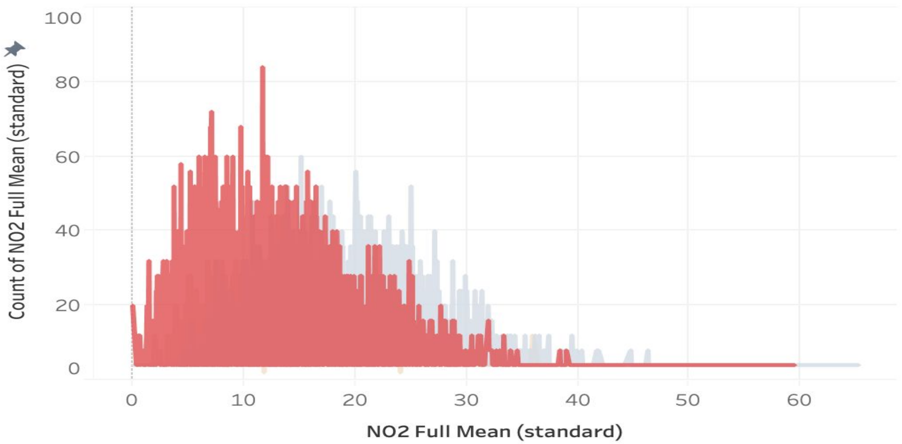
SO2 Full value Distribution



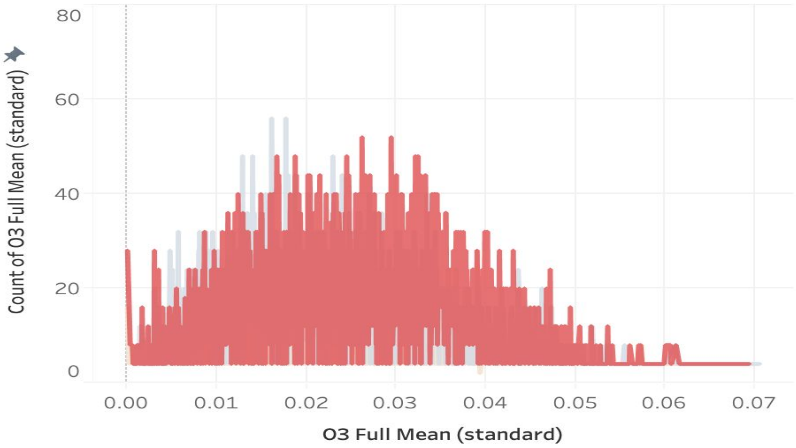
CO Full value Distribution



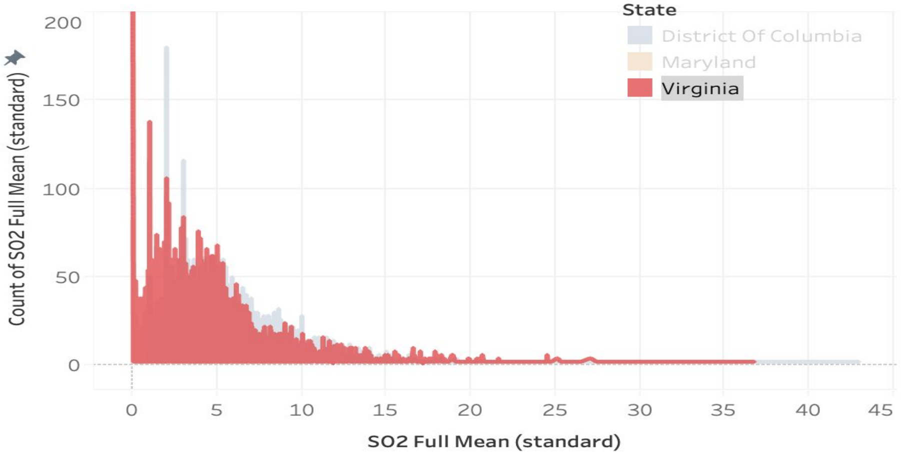
NO2 Full value Distribution



O3 Full value Distribution



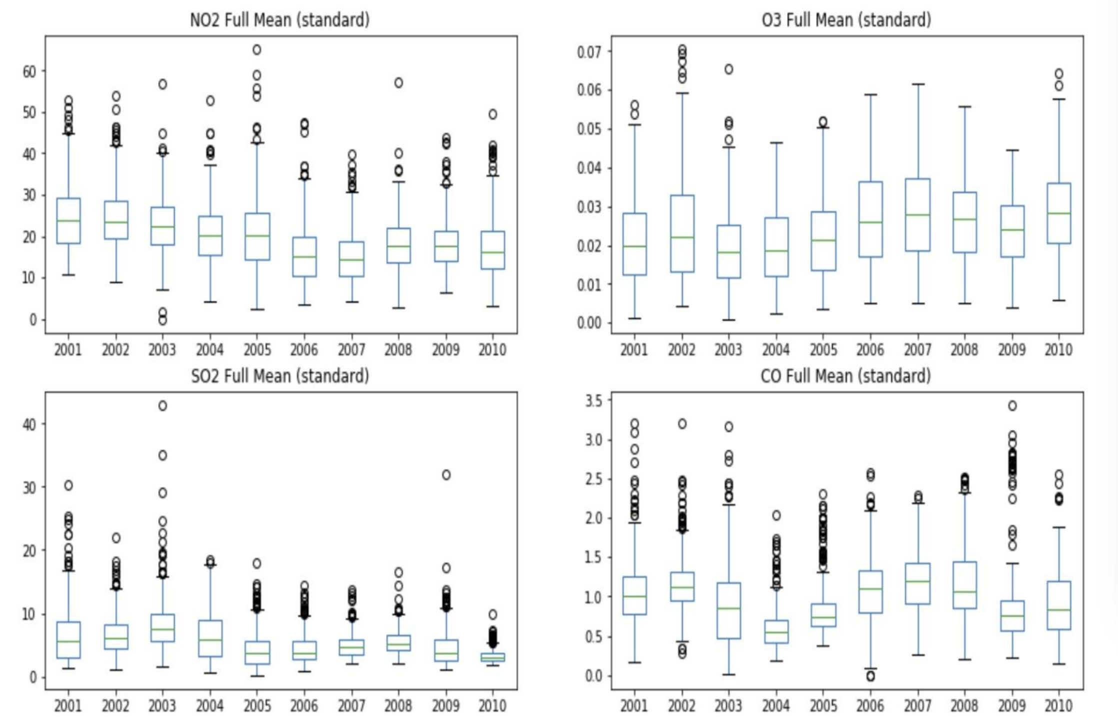
S02 Full value Distribution



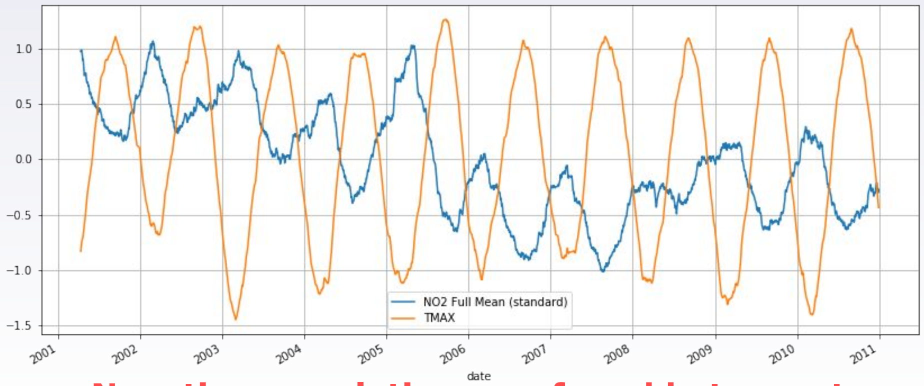
Focus on

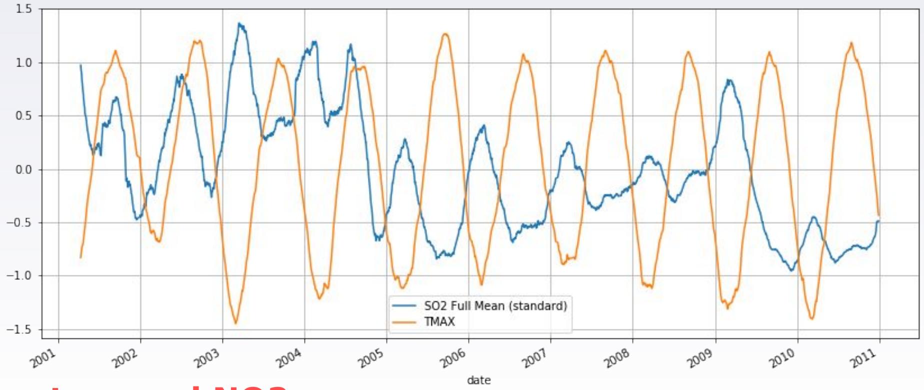
-- District of Columbia

Cyclical Pattern



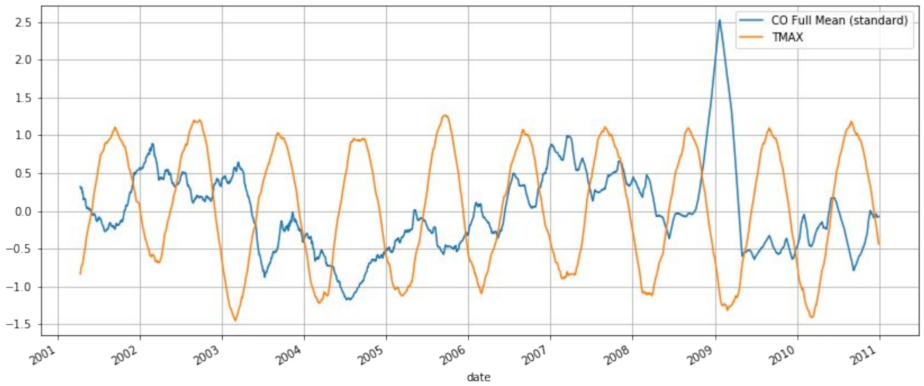
NO2 has an overall downward trend, O3 has fewer outliers than other pollutants, CO has so many outliers, and may have a longer time pattern.

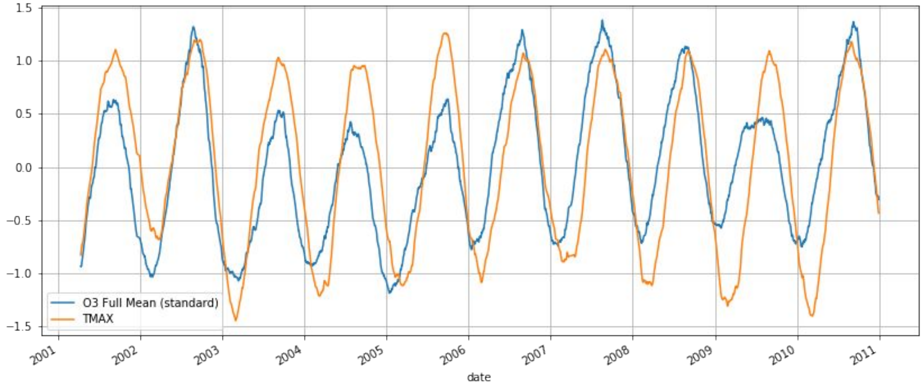




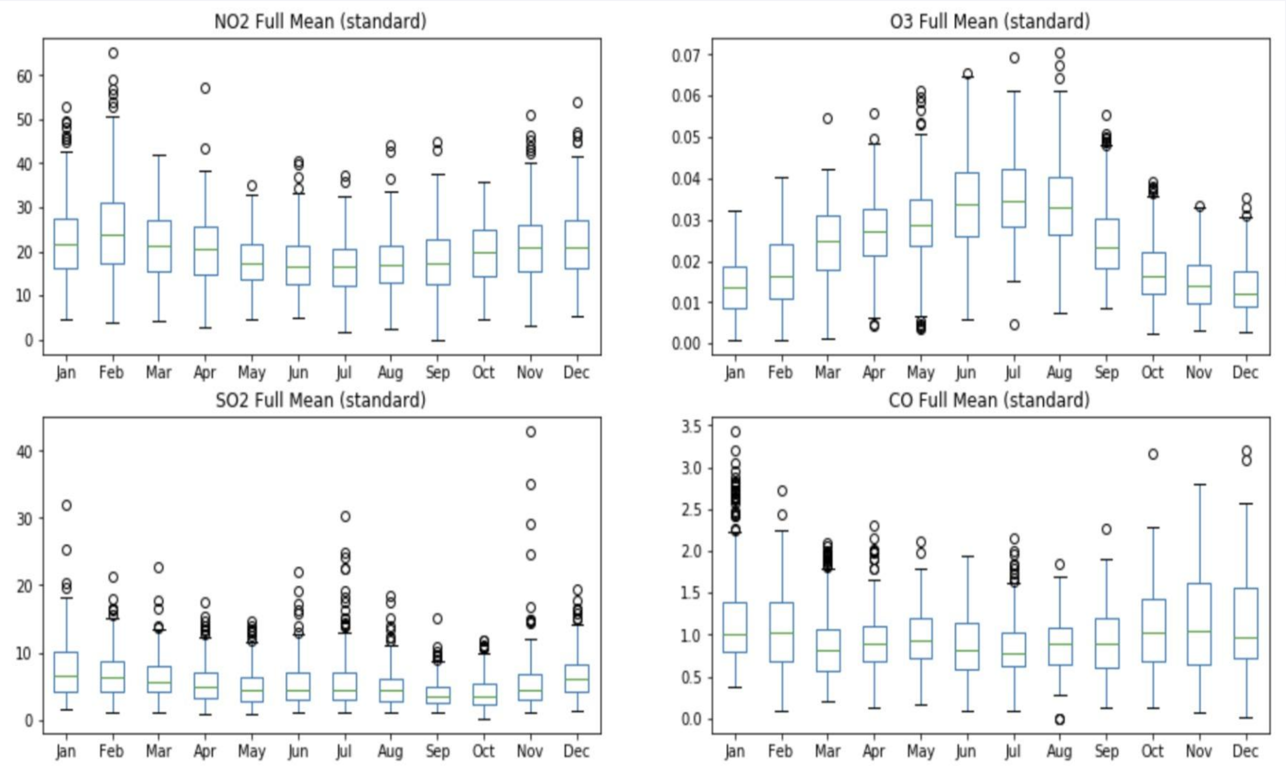
Negative correlation was found between temperature and NO2.

Positive correlation was found between temperature and O3. → seasonal pattern?





Seasonal Pattern



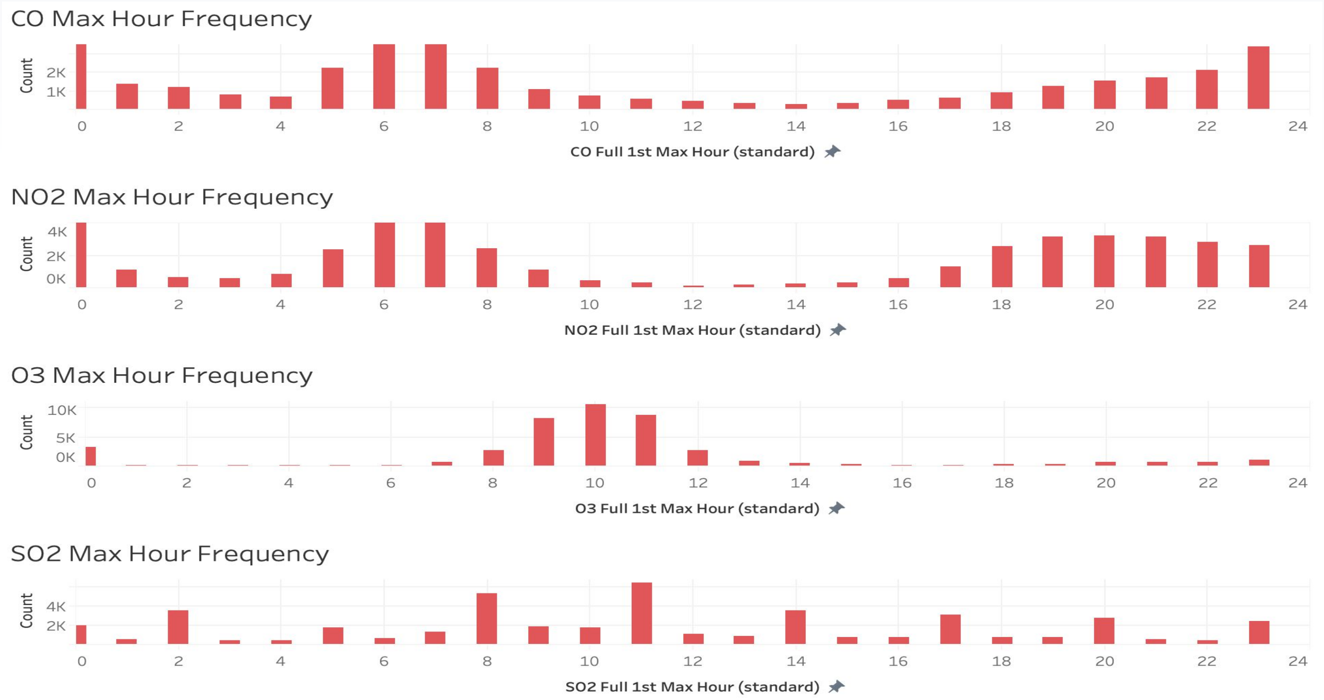
O3 arrives peak in summer due to photo-oxidation reaction, which results from high intensity of sunshine, while it shows a valley for NO2 in summer.

SO2 nearly has no seasonal patterns because the main source of SO2 is combustion of all sulfur-containing fuels, which is human behavior, it has weak correlation with season.

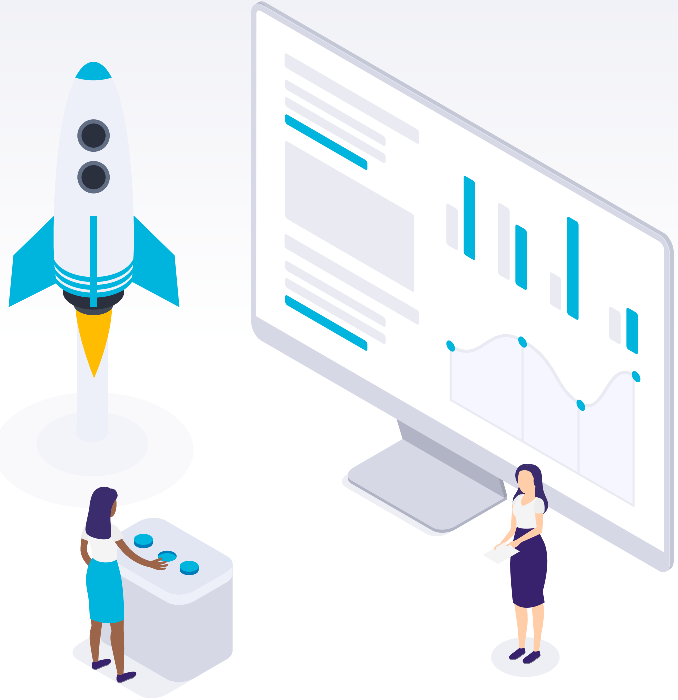
- **NO2&CO:** the diurnal cycles show two peaks during morning and evening traffic hours and valley during the afternoon hours. This phenomenon can be attributed to the day-night differences in the chemical removal of NO2 and CO.

- **O3:** NOx and CO are the main precursors of O3. So after the NO2’s peak in the morning, it decreases due to photo-chemical oxidation, and at the same time, O3 is produced, so we can see O3 increase rapidly. But in the evening, without sunlight, O3 cannot be produced

Daily Pattern



7 Solutions



Main Source of Pollutants

- **NO2 <--** Increased **fossil- and biofuels** combustion, prominent **energy** demand and higher **agricultural and cultivation**
- **SO2 <--** Combustion of all **sulfur-containing fuels (oil, coal and diesel)**
- **O3 <-- Photo-oxidation reactions of carbon-like compounds** such as CO, CH4 and NOx
- **CO <--** Emission from fossil- and biofuel combustion, biomass burning, and oxidation of methane (CH4) and non-methane hydrocarbon **<-- Coal, natural gas and oil**

Traffic-related Air Pollution

Transportation agencies and local jurisdictions can reduce traffic-related air pollution and improve air quality in these ways:

➔ Develop cleaner travel options:

- Expand public transportation systems
- Improve public transportation service
- Develop or improve bicycling and pedestrian infrastructure

➔ Reduce the distance between key destinations:

- Satisfy daily transportation needs through more efficient land use planning and zoning
- Make it more attractive and convenient to walk or bicycle instead of using using motor vehicles for transportation

Traffic-related Air Pollution

Transportation agencies and local jurisdictions can reduce traffic-related air pollution and improve air quality in these ways:

➔ **Create or support clean fueling infrastructure:**

- **Electric vehicle charging and hydrogen fueling stations**

➔ **Manage the transportation system:**

- **Increase vehicle and system operation efficiency through measures such as anti-idling policies, improved incident response, real-time travel information for public transportation**
- **Make it more attractive and convenient to walk or bicycle instead of using using motor vehicles for transportation**

Traffic-related Air Pollution

Transportation agencies and local jurisdictions can reduce traffic-related air pollution and improve air quality in these ways:

➔ **Encourage to buy green fleet vehicles and equipment:**

- ◆ **Fuel efficiency vehicles that use less oil**
- ◆ **Equipment that runs on cleaner fuels which produce fewer emissions**
- ◆ **Hybrid electric vehicles**
- ◆ **Electric vehicles that entirely removes tailpipe emissions**

➔ **Build up more strict vehicle emission standards:**

- ◆ **Especially reduce emissions from trucks and other freight sources**

EPA Three Sustainability Pillars

A sustainable approach is a systems-based approach that seeks to understand the interactions which exist among the three pillars (environment, social, and economic) in an effort to better understand the consequences of our actions.

➔ Environmental:

- **Air Quality:** Attain and maintain air-quality standards and reduce the risk from toxic air Pollutants

➔ Social:

- **Resource Security:** Protect, maintain, and restore access to basic resources (e.g. food, land, and energy)
- Example: Encourage energy reuse and recycling
- **Human Health:** Protect, sustain, and improve human health
- Example: Increasing supply of green energy sources to reduce need for fossil fuels

EPA Three Sustainability Pillars

A sustainable approach is a systems-based approach that seeks to understand the interactions which exist among the three pillars (environment, social, and economic) in an effort to better understand the consequences of our actions.

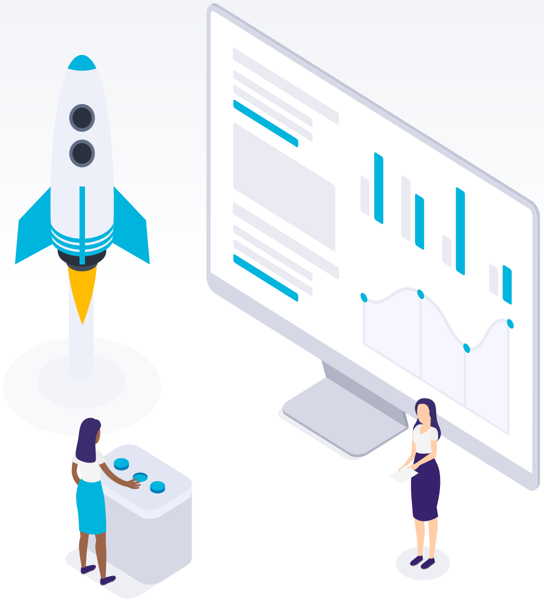
➔ Economic:

- **Costs:** Positively impact costs of processes, services, and products
- Air pollution takes its toll on the economy in several ways: it costs human lives, it affects vital products like food, it reduces the ability of ecosystems to perform functions societies need
- **Supply and Demand:** Promote price or quantity changes that alter economic growth, environmental health and social prosperity.
  Example: Increasing supply of green energy sources to reduce need for fossil fuels

8

Future Study

- Gather data with a longer time range
- Gather more daily data from industry, manufactary, agriculture and highway vehicle



THANKS!

Any Questions?

Contact us:

- zhenyang@terpmail.umd.edu
- yshen666@umd.edu