Springboard Capstone Project 1

Prediction of Unhealthy Air Pollutant Levels In Order To Better Manage Patient Care

Molly McNamara

Air Pollution and Health - A Costly Relationship

- High air pollution levels have been consistently documented as a major environmental risk to health.
- Not only is air pollution a documented health risk, it is associated with an increased utilization of health care services.
- A RAND Corporation study found that "failing to meet air quality standards resulted in overall spending on hospital care in California of slightly more than $193 million over the period 2005–2007."

Air Pollution and Health –
Can We Help Kaiser Permanente Predict and Manage Patient Illness?

- Given the health risks involved in elevated levels of air pollutants, it would be useful for health care providers and hospitals to predict the uptick of services during periods of poor air quality.
- The theoretical client is Kaiser Permanente (KP), the largest managed care organization in the United States. KP runs a number of hospitals and clinics with an emphasis on preventive care and population health management and wishes to find a way to predict hospital and clinic visits to better manage patient care.

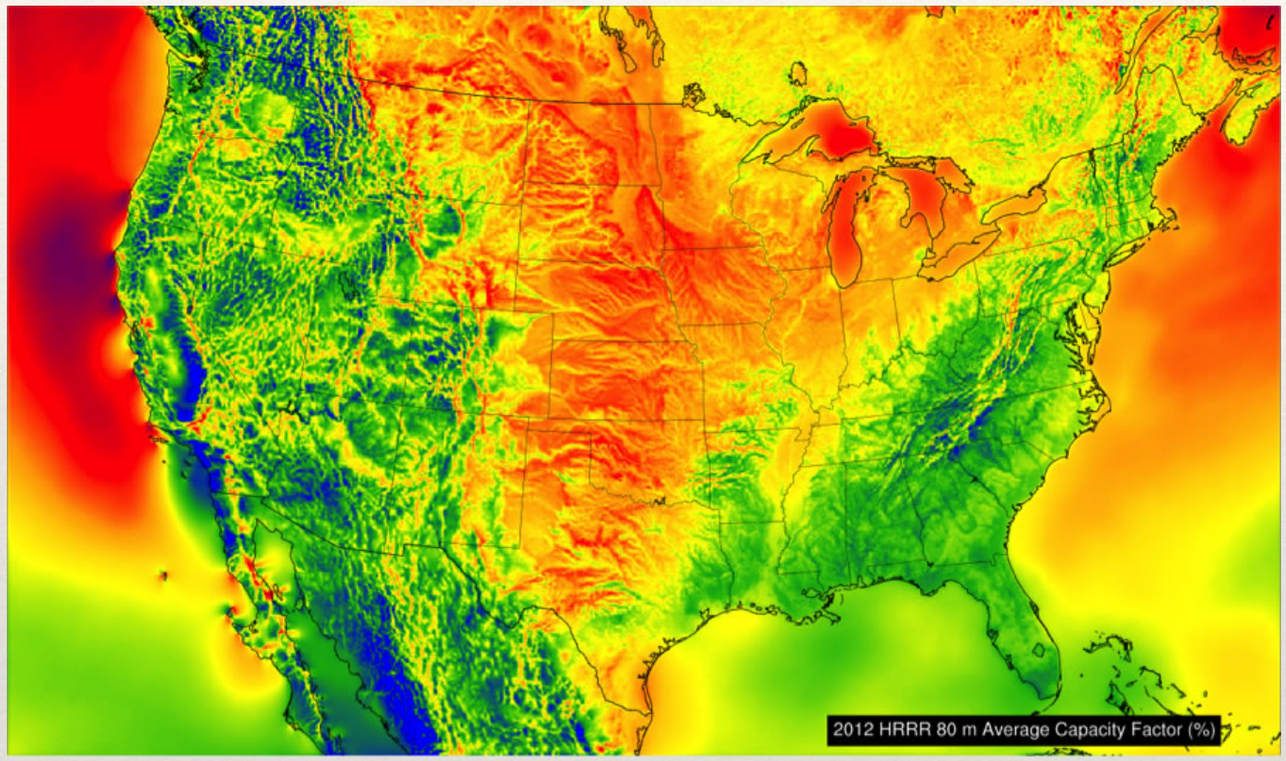
Pollution Data

Number of Particles	Name	Color	Health Implications
0 to 50	Good	Green	None
51 to 100	Moderate	Yellow	Unusually sensitive individuals should consider limiting prolonged outdoor exertion
101 to 150	Unhealthy for Sensitive Groups	Orange	Children, active adults, and people with respiratory disease, such as asthma, should limit prolonged outdoor exertion
151 to 200	Unhealthy	Red	Children, active adults, and people with respiratory disease, such as asthma, should avoid prolonged outdoor exertion; everyone else should limit prolonged outdoor exertion
201 to 300	Very Unhealthy	Purple	Children, active adults, and people with respiratory disease, such as asthma, should avoid outdoor exertion; everyone else should limit outdoor exertion
301-500	Hazardous	Maroon	Everyone should avoid all physical activity outdoors. DON'T GO OUTSIDE!

- The Environmental Protection Agency collects and tracks pollution data
- Pollutant levels are rated on the Air Quality Index (AQI) scale

- Data was gathered for 4 major air pollutants (Ozone, Carbon Monoxide, Nitrogen Dioxide, and Sulfur Dioxide) for collection sites across the United States from 2000 through 2016

Weather Data



- Weather conditions can affect air quality and could be used to predict pollutant levels
- The National Oceanic and Atmospheric Administration compiles weather data
- Data was gathered for some of the largest cities in the United States from 2000 through 2015

Data Acquisition and Compilation

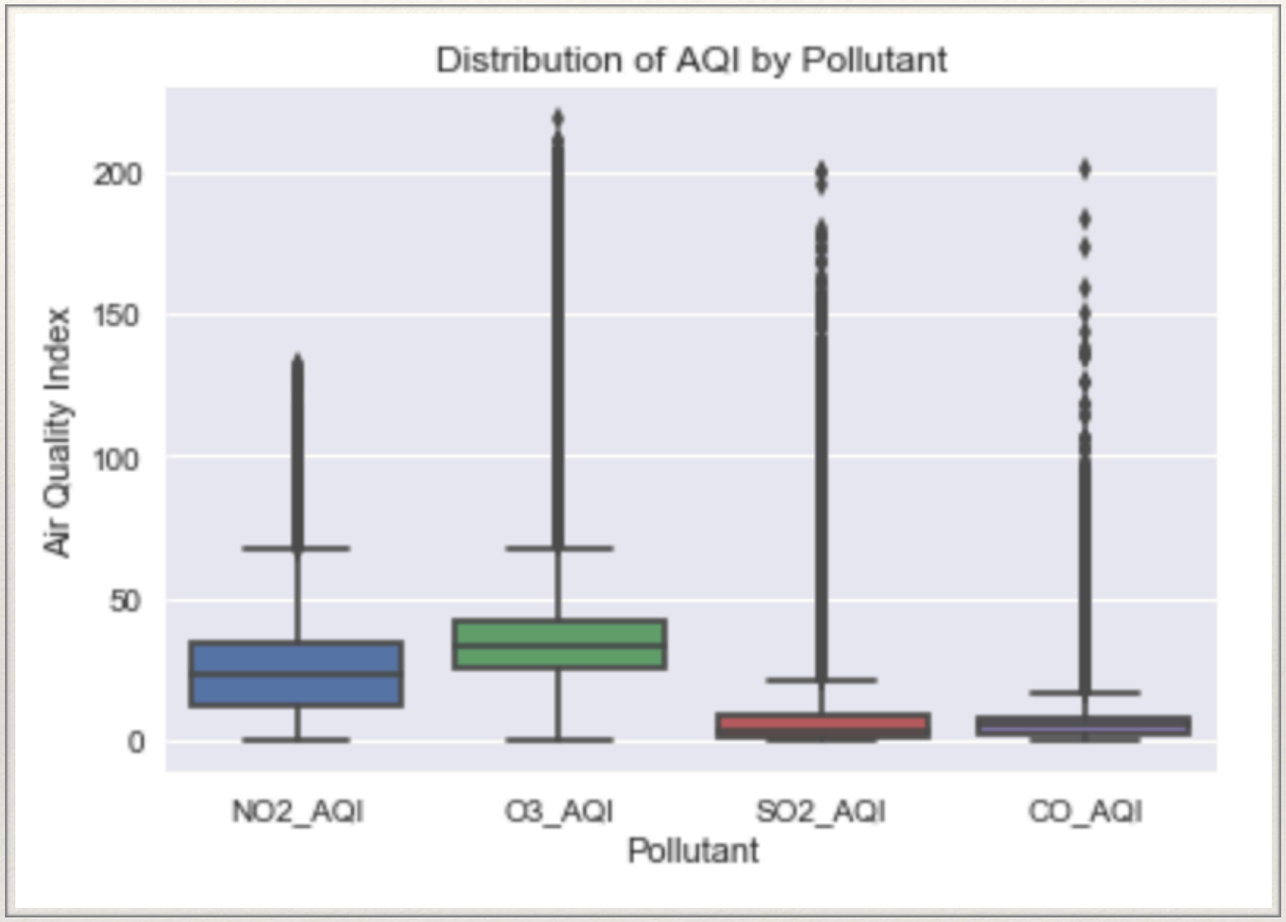
- Pollution Data:
  - 412,856 observations (1,746,661 observations prior to removal of duplicate data)
  - 28 features
  - 141 cities in 46 US states (plus 3 cities in Mexico)
  - 1 CSV file download, then extensive removal of duplicate/triplicate data

- Weather Data:
  - 56,073 observations (1,247,655 observations prior to removal of duplicate data)
  - 90 features
  - 7 cities in the US
  - ~21 CSV file downloads, then concatenated all files into one

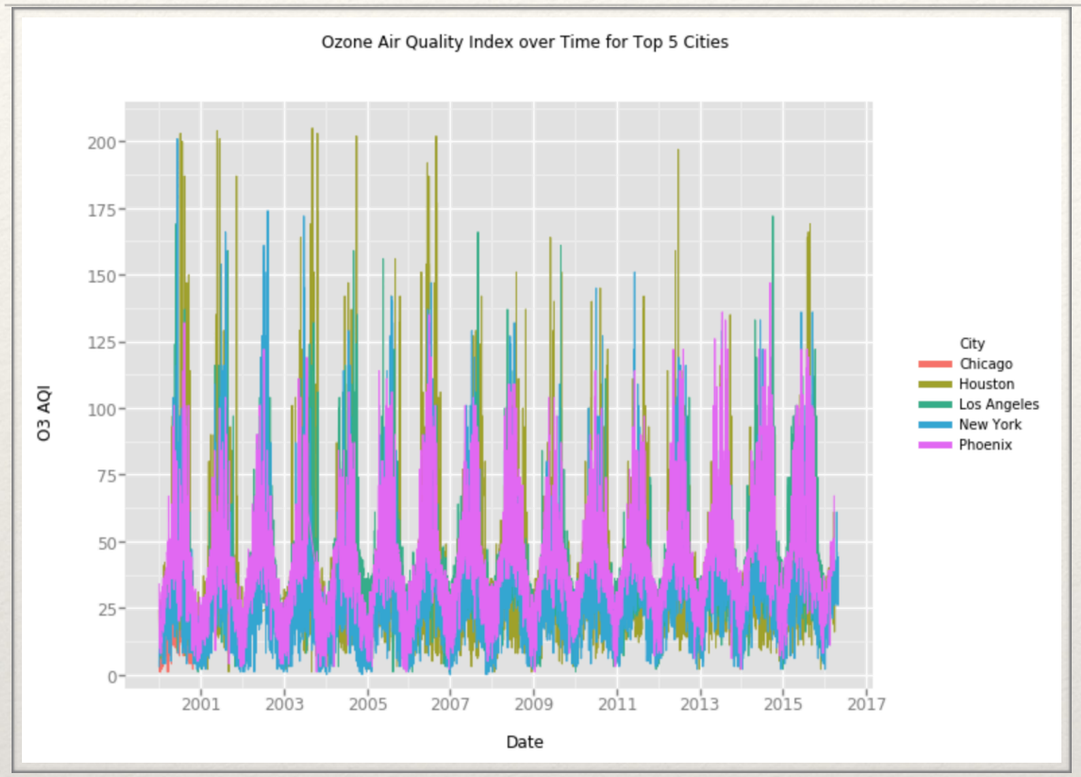
- Date columns converted to datetime objects
- Columns renamed to match
- Null values addressed
- Files merged on a left-join using City and Date

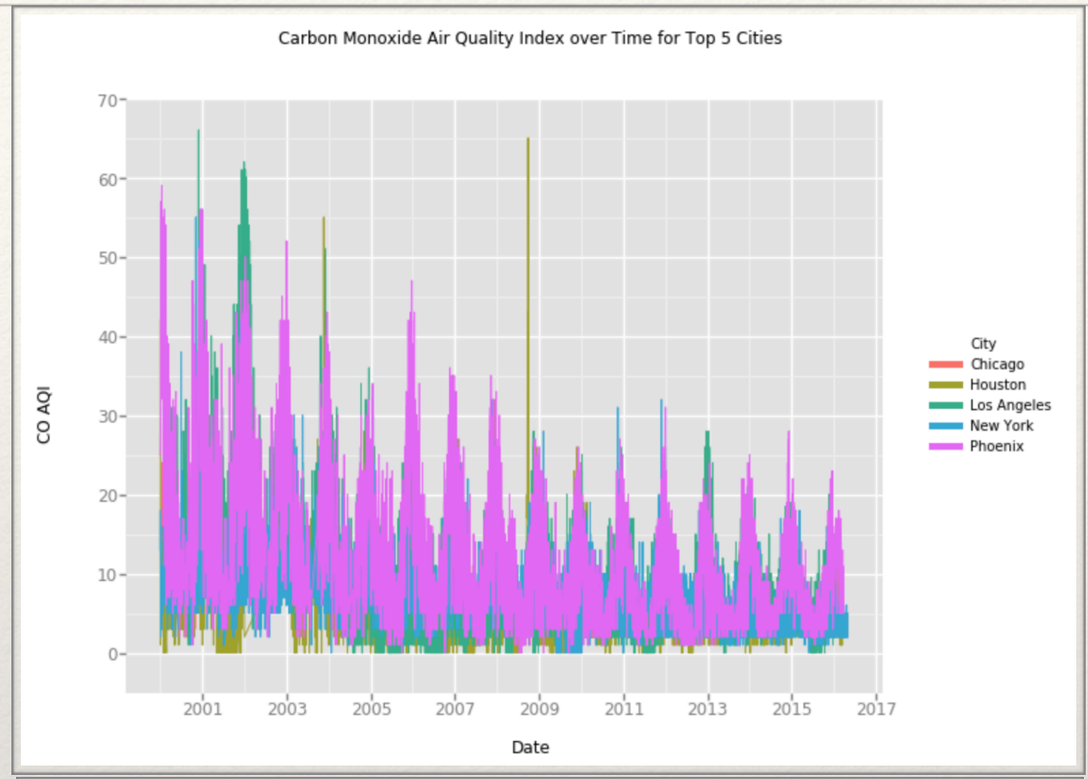
Data Exploration - Pollutant Distribution

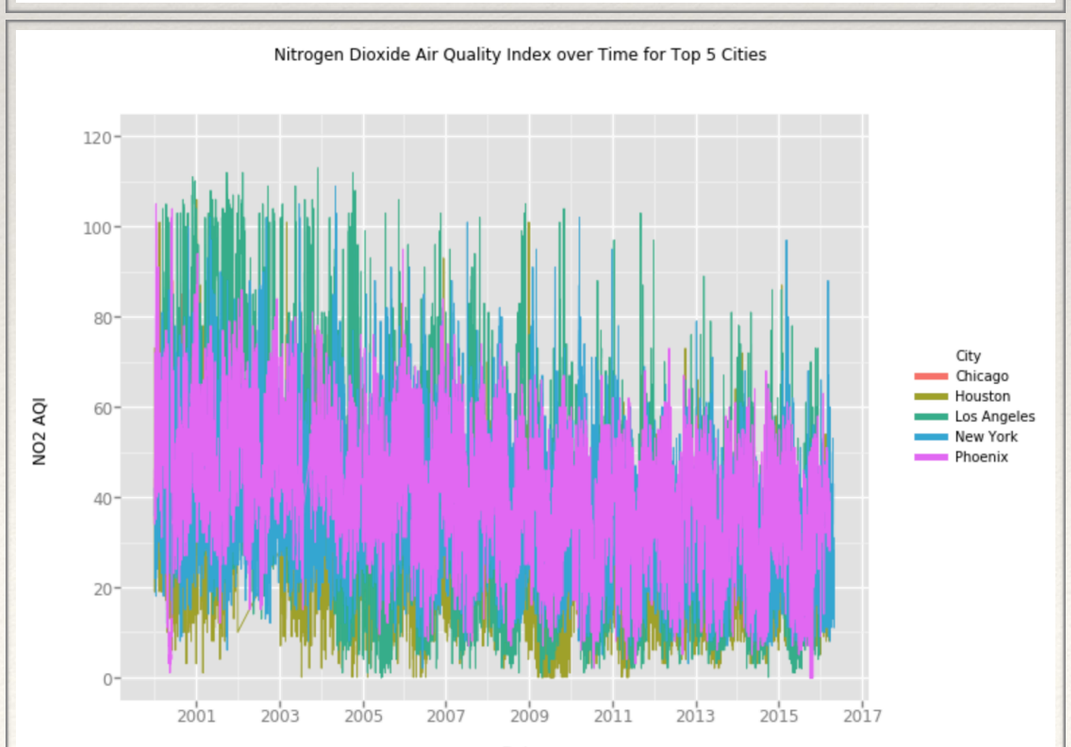
- AQI ranges were inspected for each pollutant
- Each pollutant was concentrated in the Good (0-50) zone but all demonstrated excursions to unsafe levels

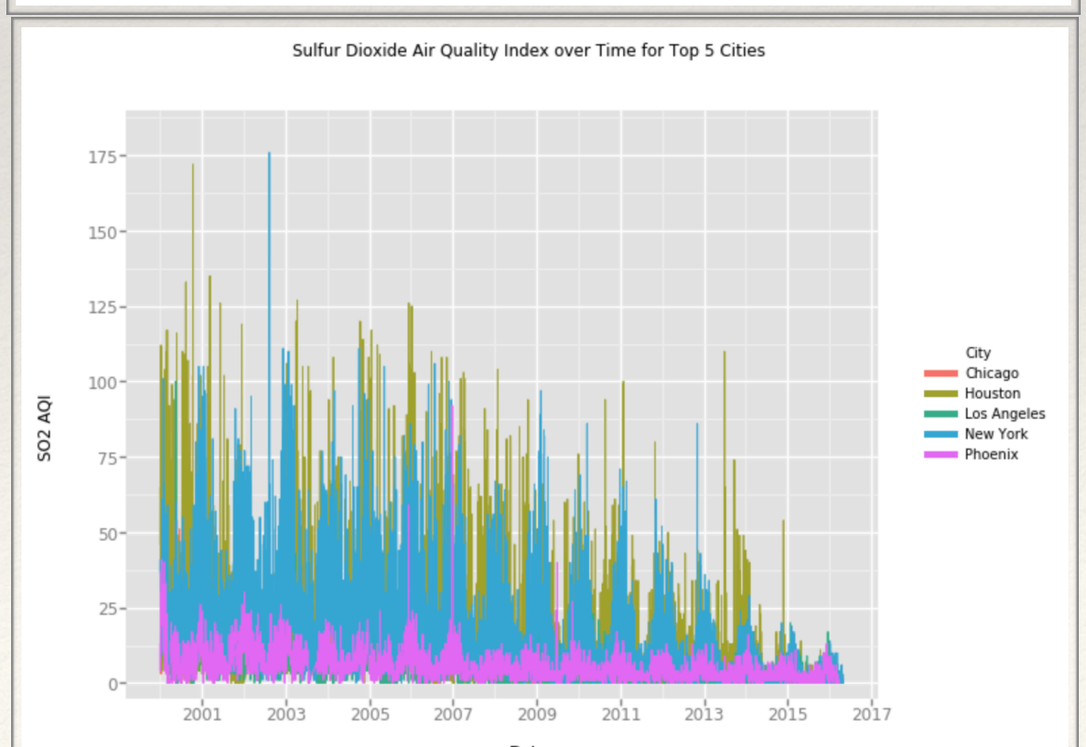


Data Exploration - Pollutant Trends Over Time for the 5 largest US Cities









Relationships Between Pollution and Weather Data

- Assessment of the correlation between pollutant levels and weather features showed that relationships did exist and were worth investigating further

	NO2_AQI	NO2_Mean	O3_AQI	O3_Mean	SO2_AQI	SO2_Mean	CO_AQI	CO_Mean
NO2_AQI	1.000000	0.881274	0.045223	-0.255967	0.291974	0.285725	0.658486	0.626278
NO2_Mean	0.881274	1.000000	-0.157326	-0.464975	0.363908	0.399538	0.705560	0.710177
O3_AQI	0.045223	-0.157326	1.000000	0.803566	-0.139647	-0.194459	-0.150828	-0.160482
O3_Mean	-0.255967	-0.464975	0.803566	1.000000	-0.237105	-0.276438	-0.353471	-0.343462
SO2_AQI	0.291974	0.363908	-0.139647	-0.237105	1.000000	0.869305	0.212616	0.227979
SO2_Mean	0.285725	0.399538	-0.194459	-0.276438	0.869305	1.000000	0.246987	0.281580
CO_AQI	0.658486	0.705560	-0.150828	-0.353471	0.212616	0.246987	1.000000	0.945793
CO_Mean	0.626278	0.710177	-0.160482	-0.343462	0.227979	0.281580	0.945793	1.000000
Elevation	0.097733	0.001738	0.214930	0.105806	-0.198358	-0.206659	0.116639	0.036277
Latitude	0.120812	0.213737	-0.144872	-0.121196	0.236486	0.330124	-0.028794	0.012734
Longitude	-0.086619	-0.015832	-0.179328	-0.172209	0.309057	0.312947	-0.273969	-0.252403
TempAvg	-0.098196	-0.214020	0.303252	0.264058	-0.200744	-0.272387	-0.174159	-0.210898
TempMax	-0.045117	-0.196188	0.506922	0.416066	-0.221477	-0.314835	-0.101933	-0.133931
TempMin	-0.201180	-0.325204	0.460371	0.456067	-0.268449	-0.332588	-0.201287	-0.195400
AvgRelHumid	0.016907	0.072413	-0.122052	-0.064120	-0.043220	-0.012250	0.190745	0.220047
Sunrise	0.175859	0.269940	-0.401370	-0.544975	0.147641	0.179789	0.265103	0.207267
Sunset	-0.245383	-0.385103	0.476539	0.559262	-0.189297	-0.271986	-0.362961	-0.349550
AvgStationPressure	-0.062964	0.039089	-0.277608	-0.200617	0.245461	0.251817	-0.082720	-0.013924
AvgSeaLevelPressure	0.195285	0.231071	-0.138308	-0.198968	0.174850	0.203780	0.220634	0.202539

Machine Learning Approach and Challenges

- Supervised Learning: Logistic Regression
- Prediction of the AQI category as a binary choice
  - Initially used all 6 AQI categories but reduced to 2 categories (Good versus Elevated pollutant levels) due to class imbalances
- Imbalanced Class Distribution
  - As noted previously, the bulk of the observations were in the Good AQI category
  - Due to this imbalance, AQI data were not normally distributed

Machine Learning Steps

- Initial models of logistic regression with/without cross-validation and balanced class weighting
- Feature data was normalized and AQI outcomes recategorized to address class imbalances (even with re-categorization, Carbon Monoxide AQI was too underpowered in Elevated levels to pursue further - 0.1% of the observations)
- Confusion matrices were used to assess how well the models predicted the less populated class(es)
- Hyperparameter tuning was performed to optimize the models
- Feature importance was assessed via recursive feature elimination and random forest classification feature importance
- Final features for each pollutant model were selected and reduced models were generated

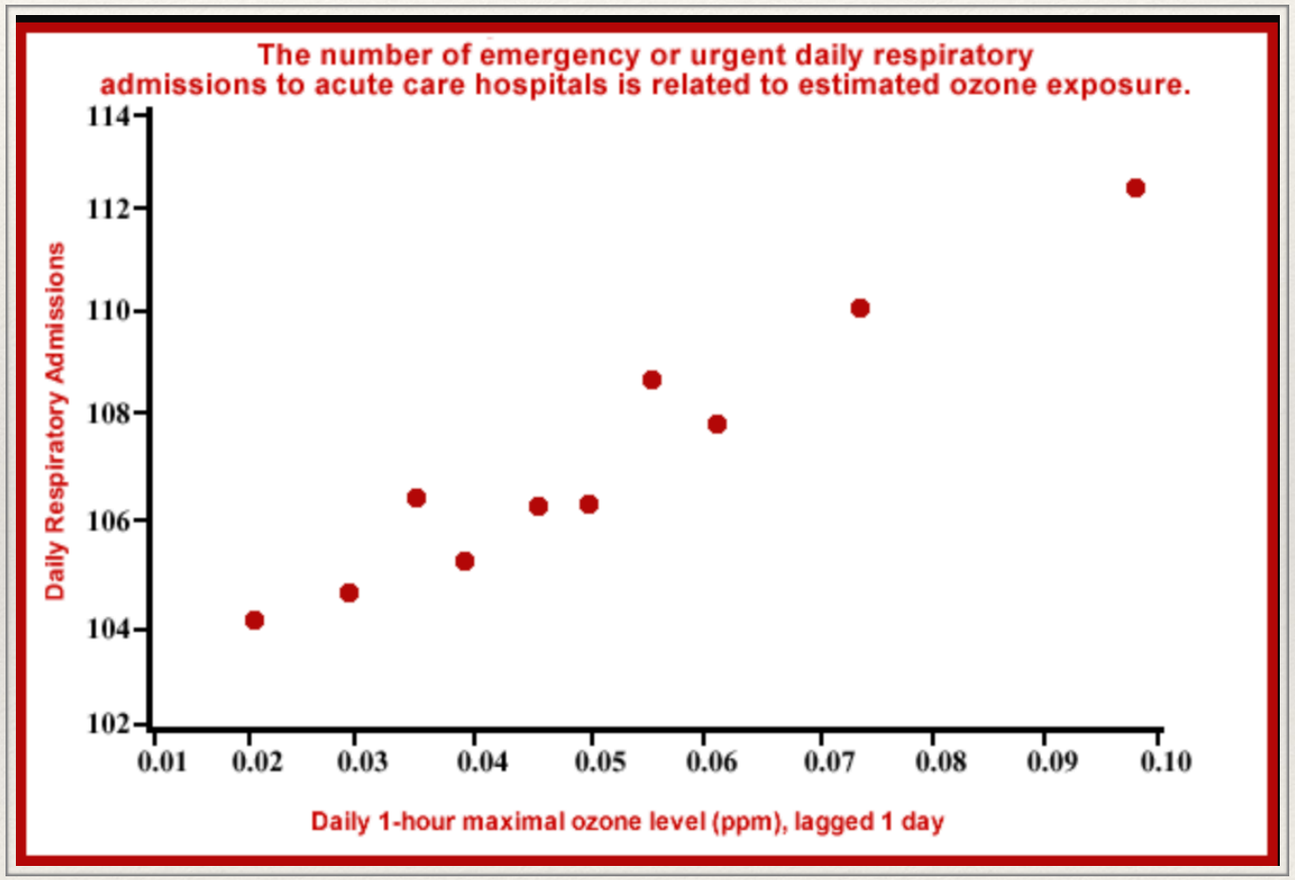
Final Pollutant Predictive Models

- Ozone:
  - Maximum Temperature
  - Average Temperature
  - Average Station Pressure
  - Sustained Wind Speed
- Nitrogen Dioxide
  - Maximum Temperature
  - Average Temperature
  - Average Station Pressure
  - Sustained Wind Speed
- Sulfur Dioxide
  - Average Relative Humidity
  - Elevation
  - Maximum Temperature
  - Minimum Temperature
  - Sustained Wind Speed

Pollutant	Accuracy	AUC
Ozone	0.76	0.87
Nitrogen Dioxide	0.70	0.75
Sulfur Dioxide	0.74	0.73

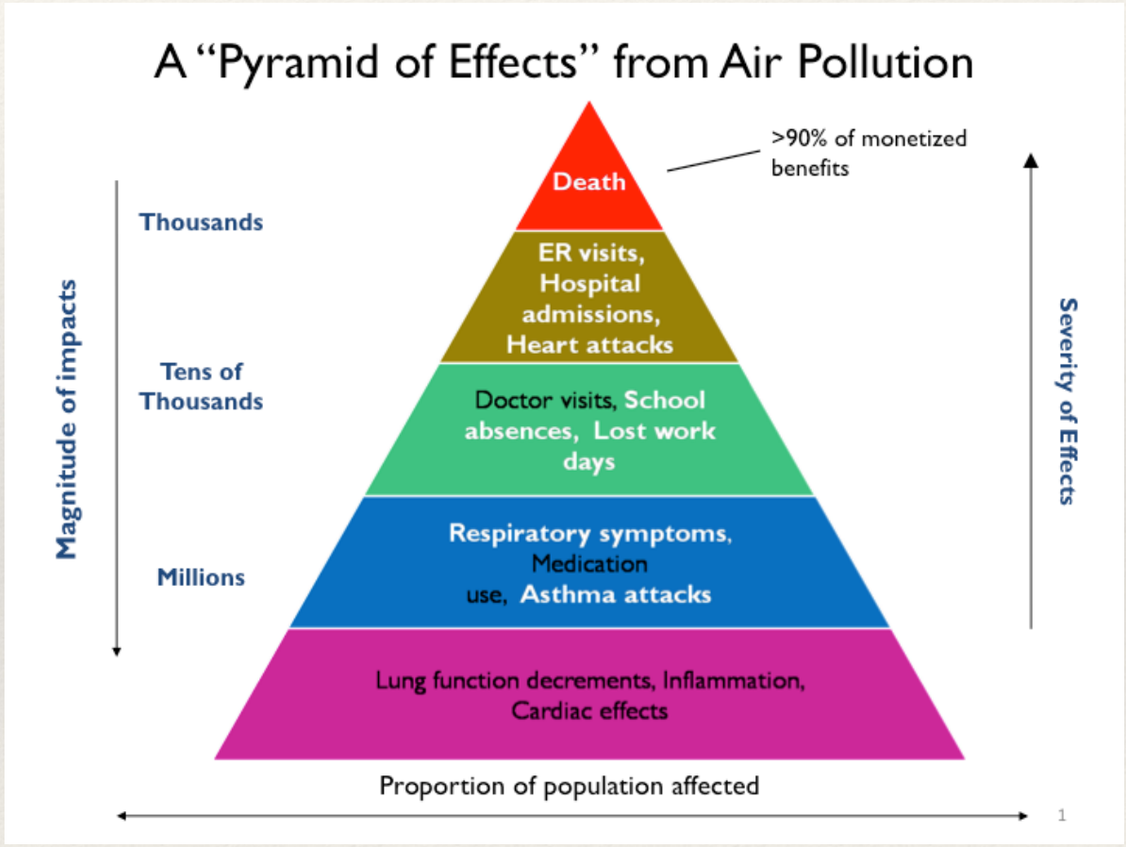
Conclusions

- The literature strongly supports a relationship between elevated pollutant levels and increased hospitalizations, emergency department visits and mortality.
- Certain features of the weather have been demonstrated here to predict pollution levels as normal/good versus elevated. This could prove useful in managing at-risk populations during periods of elevated pollution levels.



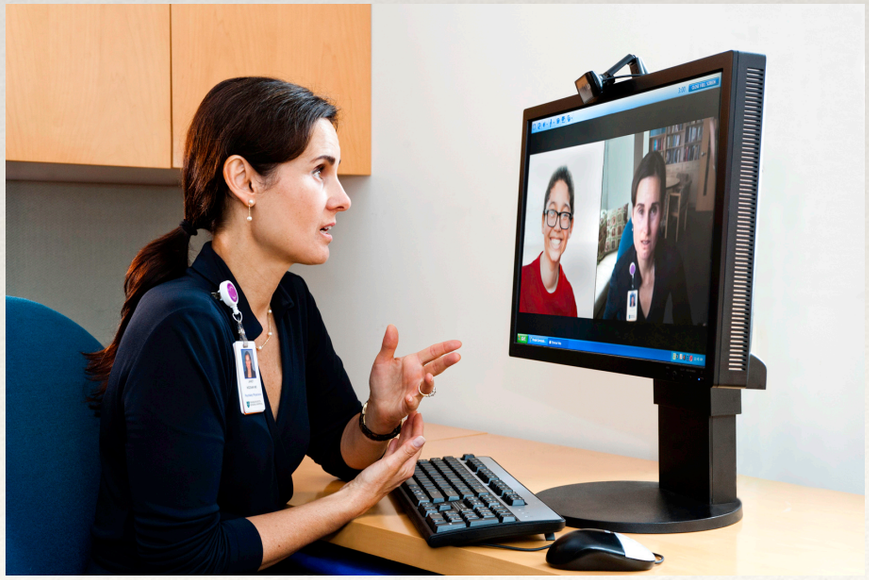
Recommendations for KP - Prevention

- Population Health Management
  - Use pollutant forecasting to message members by region with recommendations for elevated pollutant days. In line with EPA recommendations, this may include advisories to avoid time or exercise outside.
- Management of high risk patients
  - Use email and robocall messaging to provide information and recommended actions to patients already diagnosed with respiratory diseases, cardiovascular diseases, the elderly and the very young.
  - Use the existing KP online patient visit system to reach out to the highest-risk patients to discuss any further medical recommendations, such as changes in medication or more aggressive case management.



Recommendations for KP - Staffing

- Hospital and Emergency Rooms
  - Consider the addition of medical staff to outpatient clinics and emergency rooms on days with an expected increase in pollution-caused illness to allow improved management of the spike in visits and perhaps reduce hospital admissions if care could be provided sooner in an outpatient facility.
- TeleHealth
  - Increase staffing of telehealth resources to communicate with patients by phone or the internet in periods of higher pollutant levels to allow management of some patients without utilization of in-person resources.



Recommended Future Work

- Inclusion of data for emergency room visits and hospital admissions or respiratory diseases in order to better predict the actual increase in illness caused by pollutant levels. Unfortunately, the only available data through free sources was monthly/annual rates of hospital/ER visits, which do not provide the granularity necessary to make accurate predictions.
- Analysis of weather data leading up to any given day to determine the window of influence on pollutant levels (example: a 3-day rolling average of weather conditions)

Thank You

Special thanks to my incredibly helpful mentor, Liang Kuang!