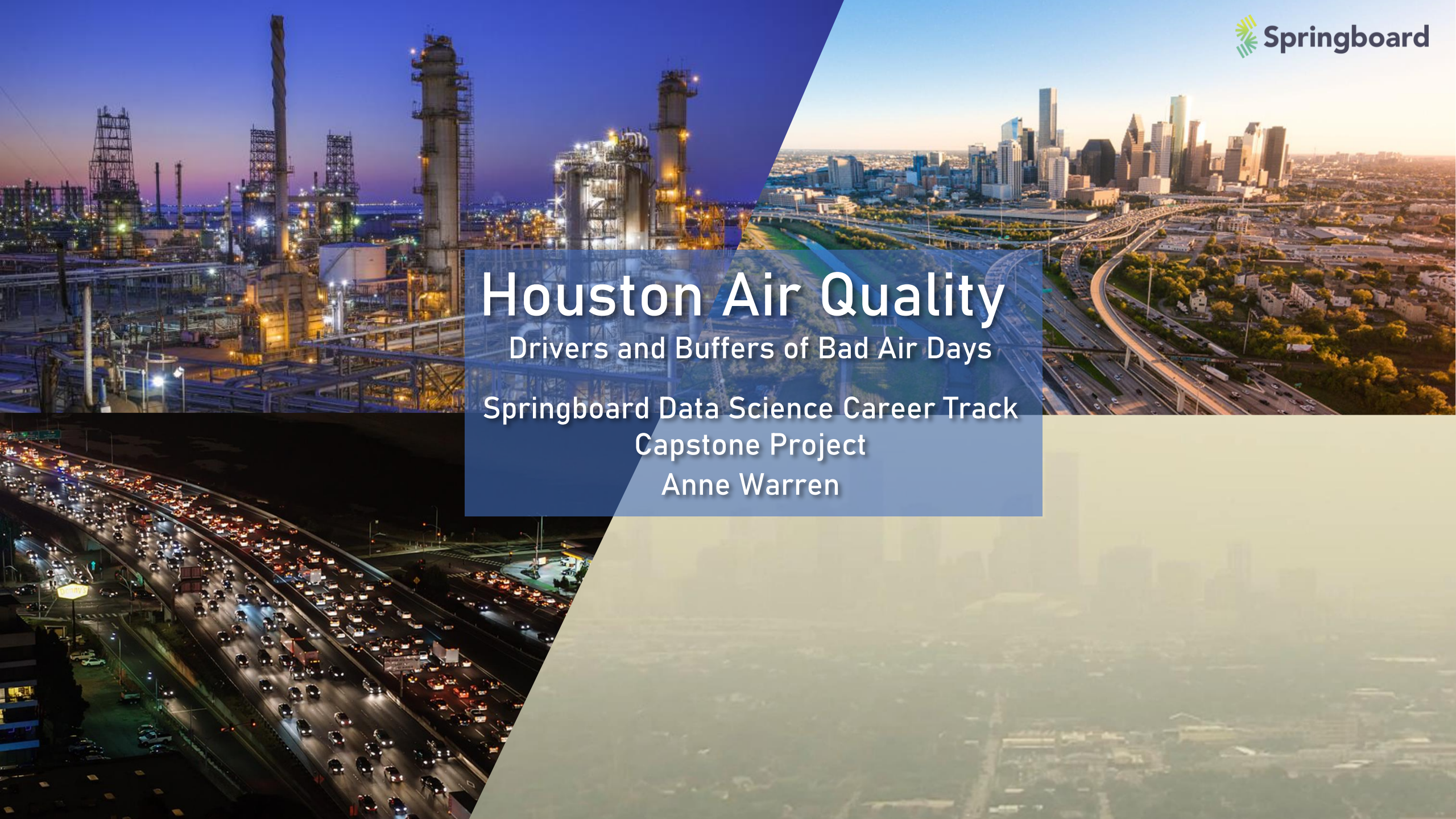


Houston Air Quality

Drivers and Buffers of Bad Air Days

Springboard Data Science Career Track
Capstone Project

Anne Warren

Bad Air Days (BADs) In The Daily Life Of Houstonians

https://cdns.abclocal.go.com/content/ktrk/images/cms/3768562_1280x720.jpg

The Problem…

Ozone Days

- Recurring Air Quality Alerts in Houston
- Higher Frequency of BADs Each Year
- Worsening of Air Quality during BADs

The Response…

CLIMATE MAYORS

Climate Mayors Announces New Chair, Houston Mayor Sylvester Turner

Mayor Sylvester Turner is a national example of a climate-proactive leader

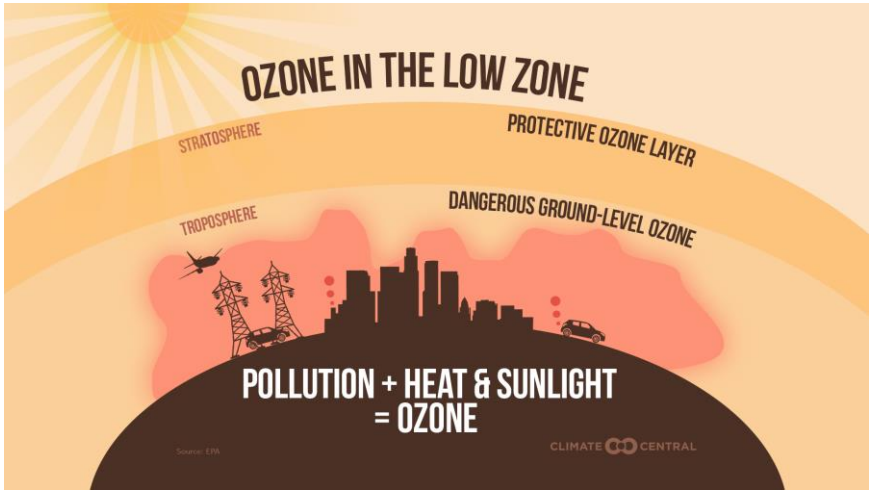
The Potential Factors…

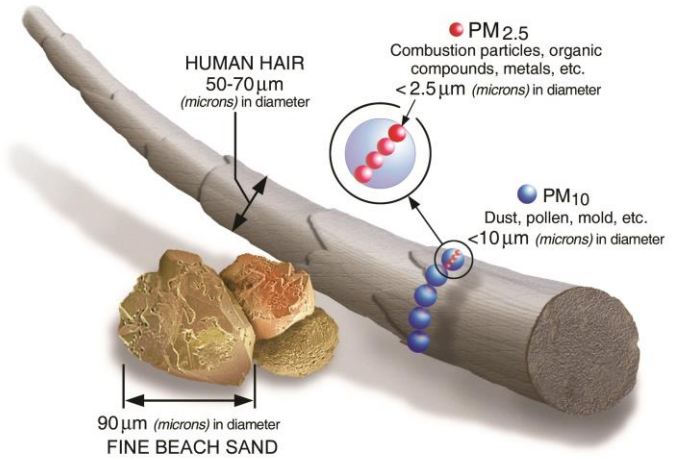
Increasing Population	9M by 2040 according to HGAC
Worsening Traffic	Ranked #8 in the nation in 2020
Increasing Pollution	From plastic plants, refineries
Fostering Climate	Sub-tropical conditions

https://d3ebobe8l15pwo.cloudfront.net/wp-content/uploads/2020/01/population-increase.jpg

Pollutants and Air Quality Index

Ozone, NO2, SO2, CO, PM 2.5, PM 10, Lead





AQI Calculation

$$I = \frac{I_{high} - I_{low}}{C_{high} - C_{low}}(C - C_{low}) + I_{low}$$

where:

I = the (Air Quality) index,
C = the pollutant concentration,
C_{low} = the concentration breakpoint that is ≤ C,
C_{high} = the concentration breakpoint that is ≥ C,
I_{low} = the index breakpoint corresponding to C_{low},
I_{high} = the index breakpoint corresponding to C_{high}.

O_3 (ppb)	O_3 (ppb)	$PM_{2.5}$ (µg/m^3)	PM_{10} (µg/m^3)	CO (ppm)	SO_2 (ppb)	NO_2 (ppb)	AQI	AQI
C_{low} - C_{high} (avg)	C_{low} - C_{high} (avg)	C_{low}- C_{high} (avg)	C_{low} - C_{high} (avg)	C_{low} - C_{high} (avg)	C_{low} - C_{high} (avg)	C_{low} - C_{high} (avg)	I_{low} - I_{high}	Category
0-54 (8-hr)	-	0.0-12.0 (24-hr)	0-54 (24-hr)	0.0-4.4 (8-hr)	0-35 (1-hr)	0-53 (1-hr)	0-50	Good
55-70 (8-hr)	-	12.1-35.4 (24-hr)	55-154 (24-hr)	4.5-9.4 (8-hr)	36-75 (1-hr)	54-100 (1-hr)	51-100	Moderate
71-85 (8-hr)	125-164 (1-hr)	35.5-55.4 (24-hr)	155-254 (24-hr)	9.5-12.4 (8-hr)	76-185 (1-hr)	101-360 (1-hr)	101-150	Unhealthy for Sensitive Groups
86-105 (8-hr)	165-204 (1-hr)	55.5-150.4 (24-hr)	255-354 (24-hr)	12.5-15.4 (8-hr)	186-304 (1-hr)	361-649 (1-hr)	151-200	Unhealthy
106–200 (8-hr)	205–404 (1-hr)	150.5–250.4 (24-hr)	355–424 (24-hr)	15.5–30.4 (8-hr)	305–604 (24-hr)	650–1249 (1-hr)	201–300	Very Unhealthy
-	405-504 (1-hr)	250.5-350.4 (24-hr)	425-504 (24-hr)	30.5-40.4 (8-hr)	605-804 (24-hr)	1250-1649 (1-hr)	301-400	Hazardous
-	505-604 (1-hr)	350.5-500.4 (24-hr)	505-604 (24-hr)	40.5-50.4 (8-hr)	805-1004 (24-hr)	1650-2049 (1-hr)	401-500	

https://assets.climatecentral.org/images/uploads/gallery/2019AirQuality2_OzoneExplainer_en_title_lg.jpg https://www.epa.gov/sites/production/files/2016-09/pm2.5_scale_graphic-color_2.jpg

The Hypothesis, The Data, The Plan

The Hypothesis…

More People, More Traffic, More Pollution

public transportation stigma, one person means one vehicle on the road

The Data…

- Pollutant Concentrations from TCEQ's Tamis DB — link to TCEQ
- Pollutant & Weather Data from US EPA — link to EPA
- Land Use from Houston-Galveston Area Council (HGAC) — link to HGAC
- Weather Data from NOAA — link to NOAA
- Traffic and Population Data from Houston — link to COHGIS
- Indoor/Outdoor Data from the RIOPA Team — link to RIOPA

The Plan…

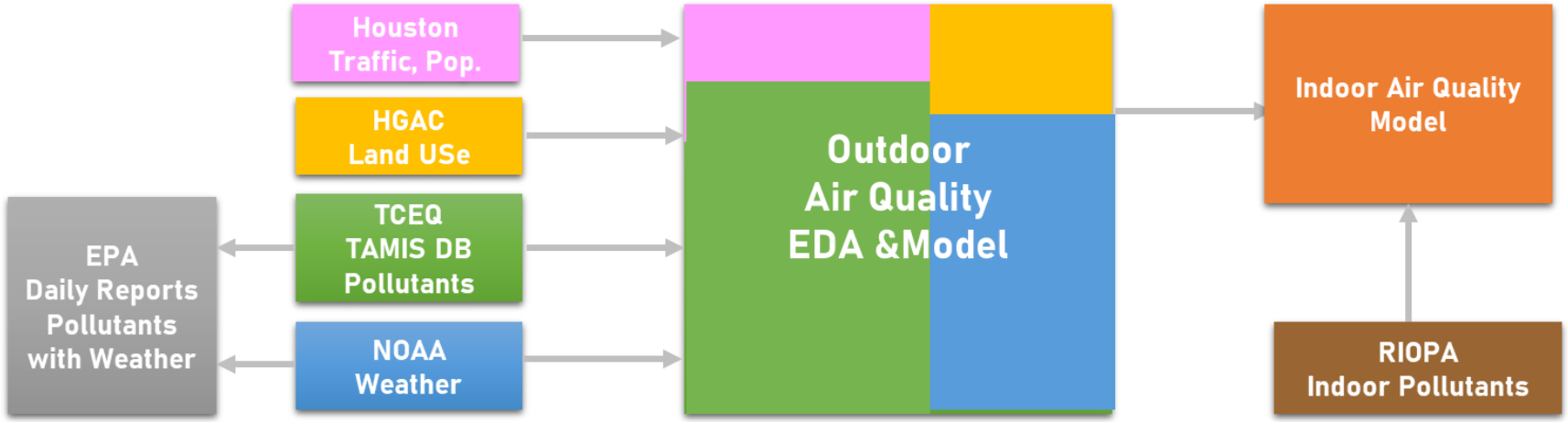
- Merge the data into a coherent dataset… taking into account geographical location
- Cover territory from The Woodlands (N, greener area) all the way down to Galveston (SE, coastline) including Baytown (E, plants) and Angleton (S, plants & coastline).
- Investigate the relationship between indoor and outdoor pollution to predict indoor pollution from modeled outdoor pollution.

Thinking the Dataset And Data Wrangling

A lot of bumps on the road…

- Matching sampling rate (1hr, 8hr, 24hr) — need one, full daily measurement
- Connect local weather to location — Houston is a large city
- Dealing with missing data — fill like a time series
- Linking indoor and outdoor with location — missing location data in RIOPA

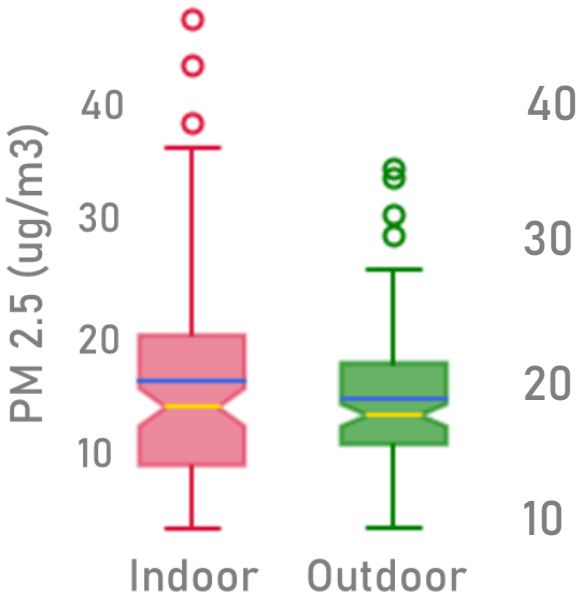
The Idea…

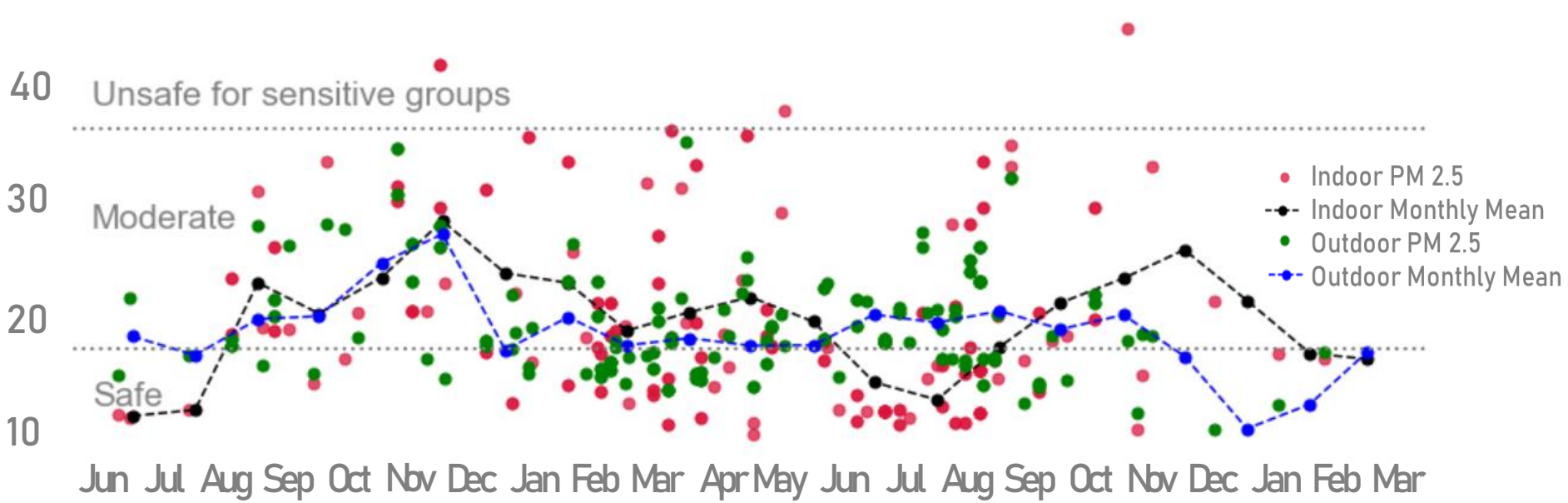


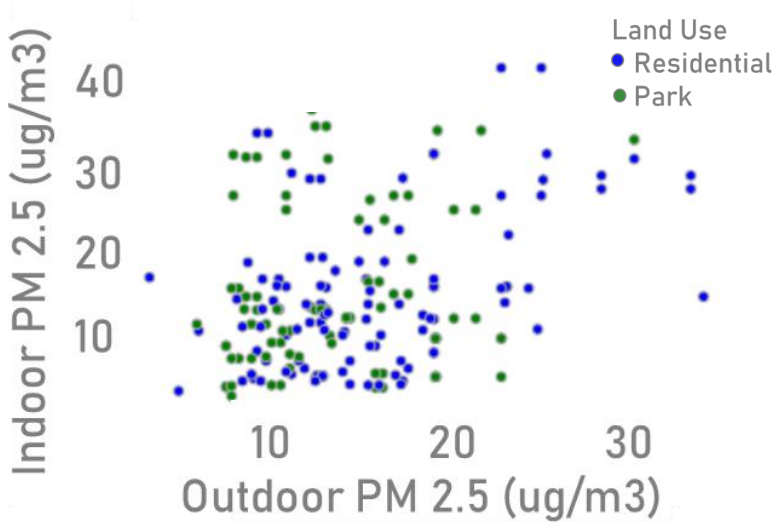
Wrangling…

- Convert pollutant concentration unis to AQI standards and labels
- Convert data to daily sampling, keep only full daily records (skip PM 10, Lead)
- Fill missing data using the appropriate imputation methods
- Bin traffic data, population data and land use data and apply using Manhattan distance
- Merge all data with the closest weather station

EDA: Indoor Data 1999-2001 (RIOPA)



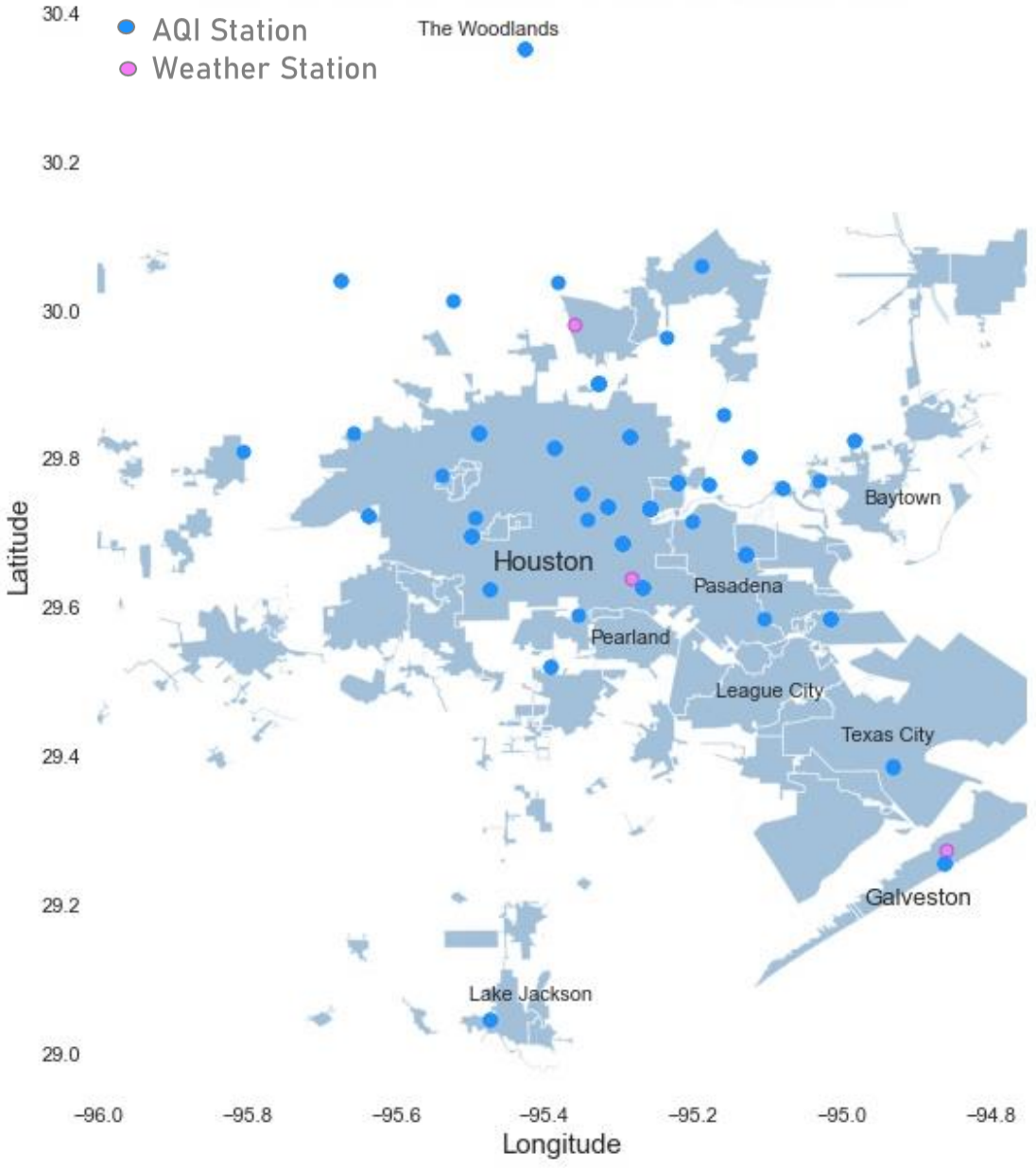




Why RIOPA data did not work out …

- No geographical data — cannot locate and tie to model
- Lack of land use diversity — no industrial or near industrial
- Lack of relevant trends — the meta data is incomplete
- Little dataset — less than 200 rows

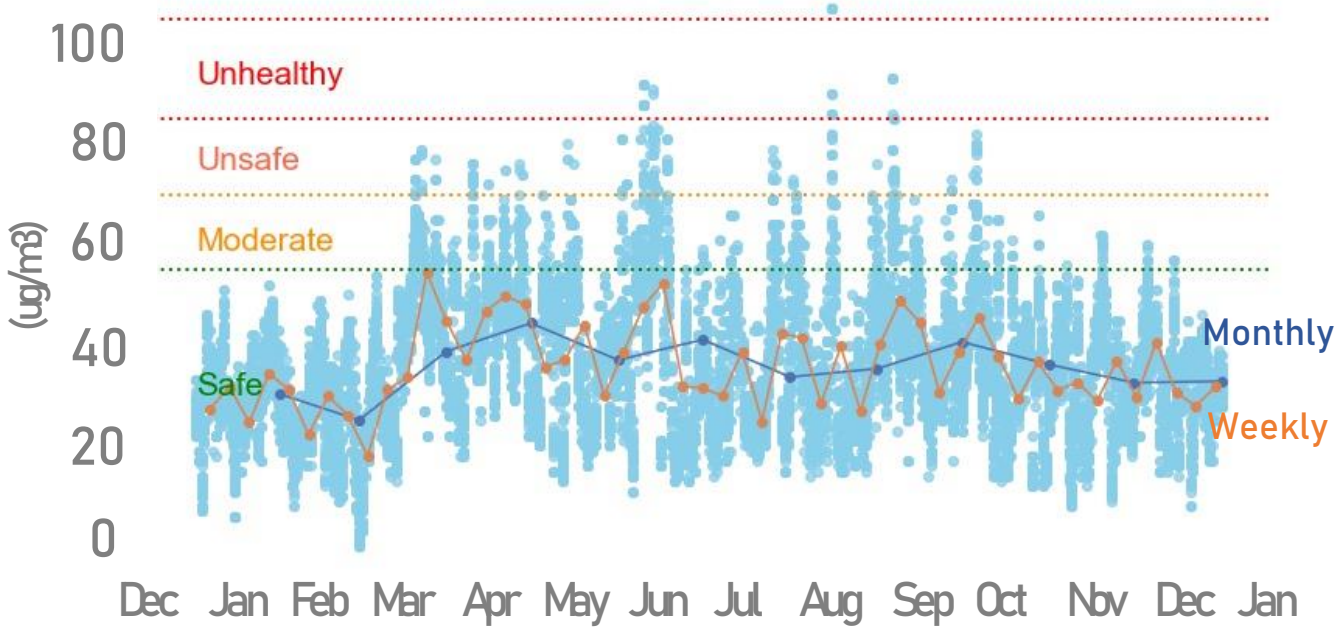
EDA: The Most Concerning Pollutants in Outdoor Data



Pollutant in Houston's air 2008-2020

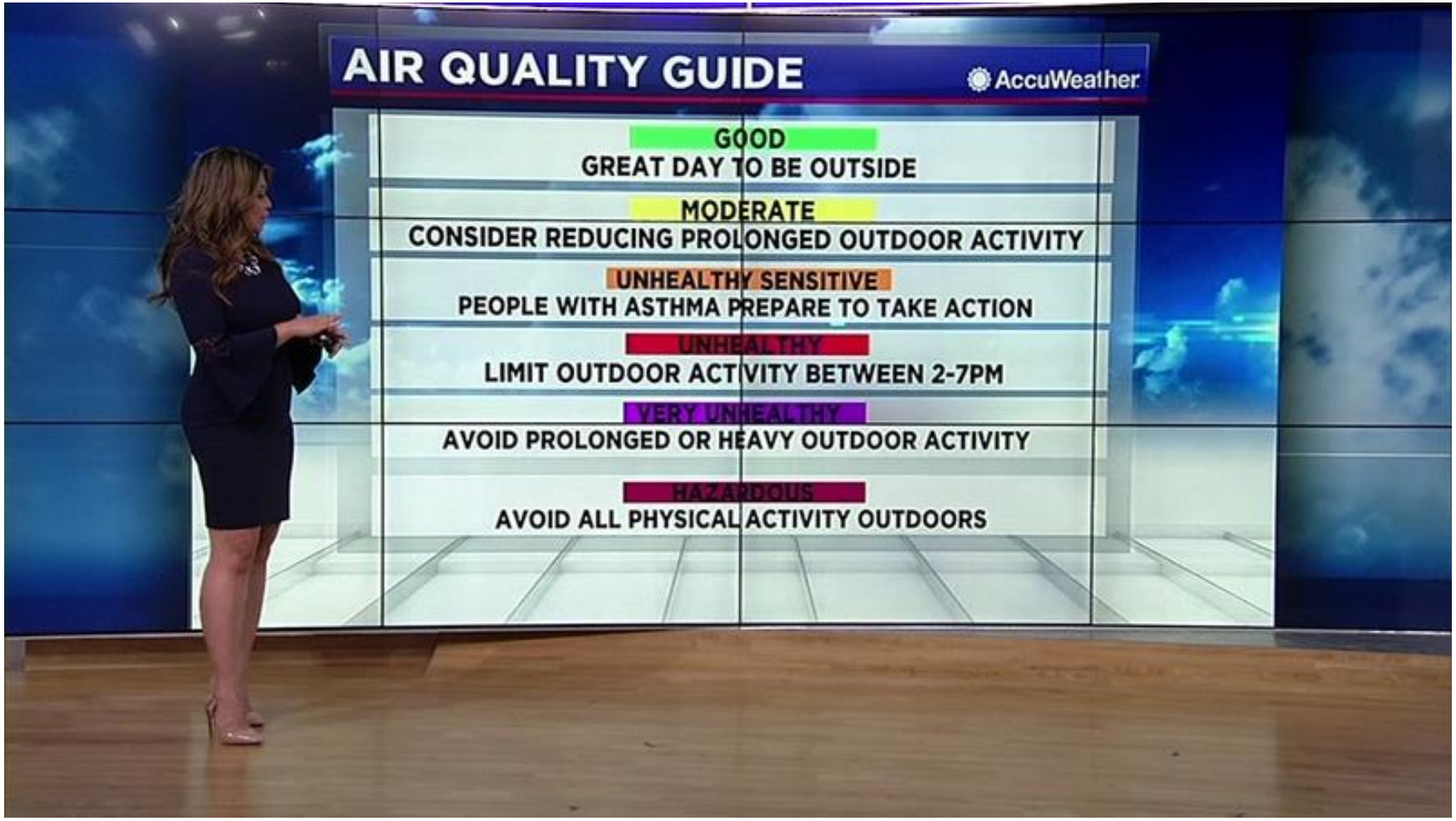
- Ozone is the only issue — reaches very unhealthy
- PM, SO2, NO2, CO Safe — rarely moderate
- Low levels of Lead — not enough data
- Average ozone within safe — despite BADs
- Seasonal ozone cyclicity — lower in Winter/Spring

Daily and Mean Ozone Concentrations in 2019



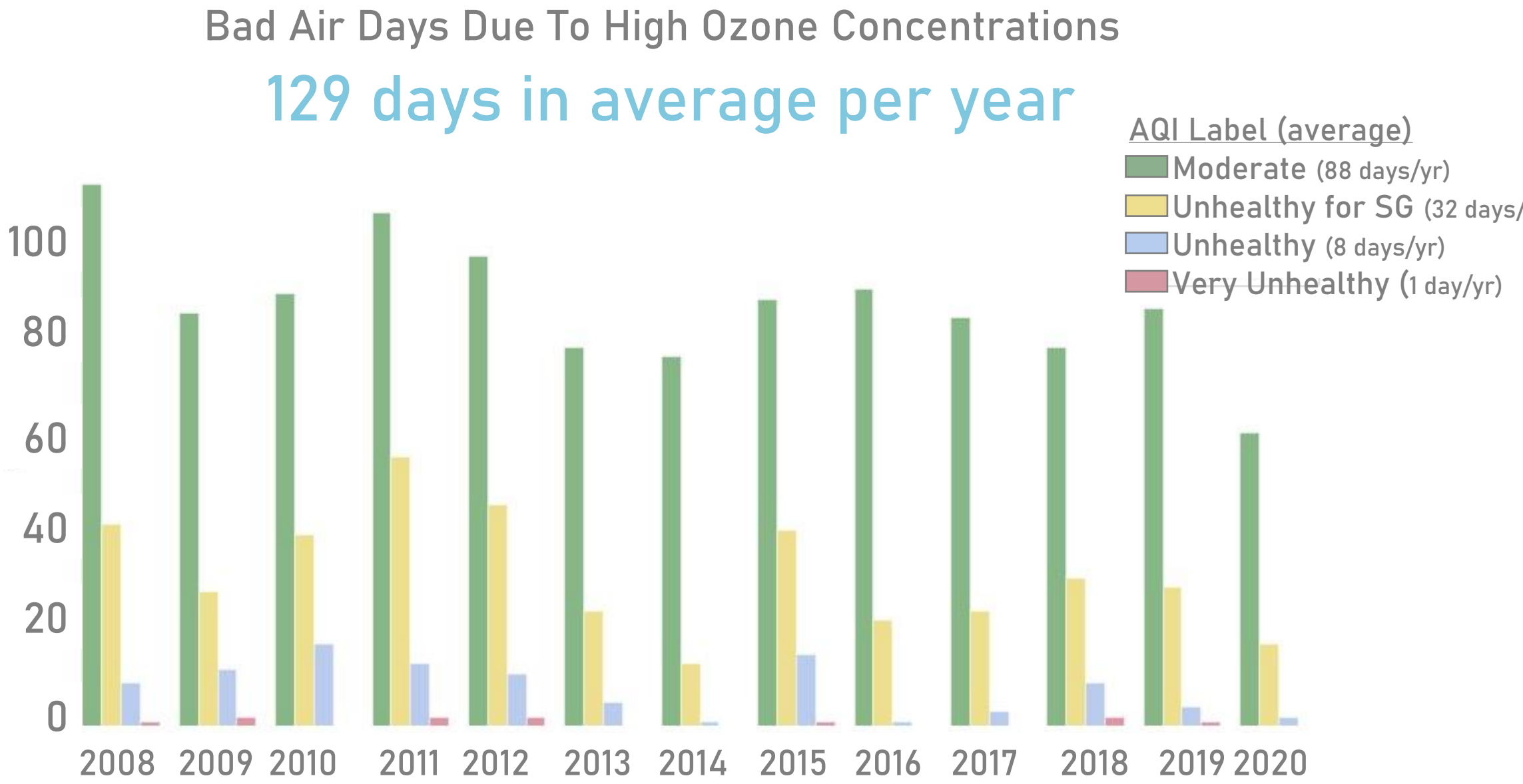
What does the Air Quality Index Means In Everyday Life?

AQI describes the impact of pollution on the quality of life and daily activities. When applied to Ozone concentration the effects are as described below.



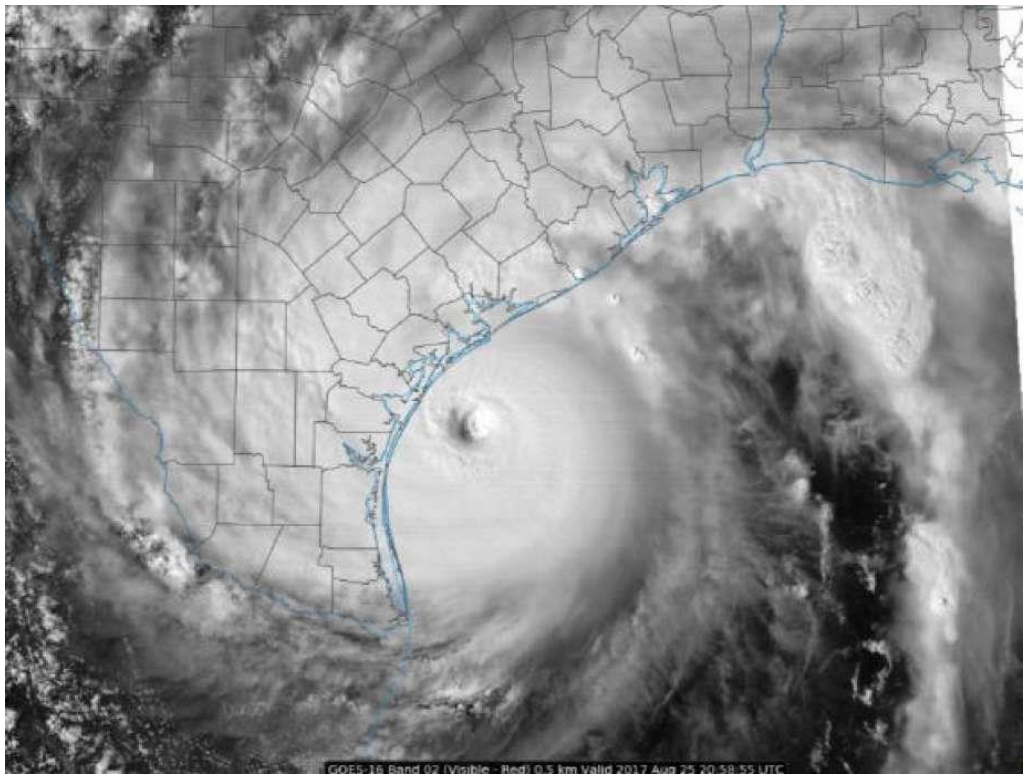
https://cdn.abcotvs.com/dip/images/5203091_031819-wx-air-quality-guide-vid_1_09-23-48,27.jpg?w=800&r=16%3A9

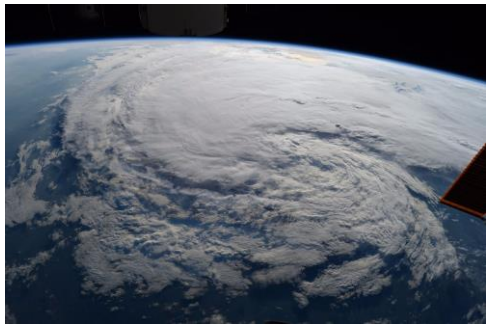
EDA: BADs At A Glance (2008-2020)

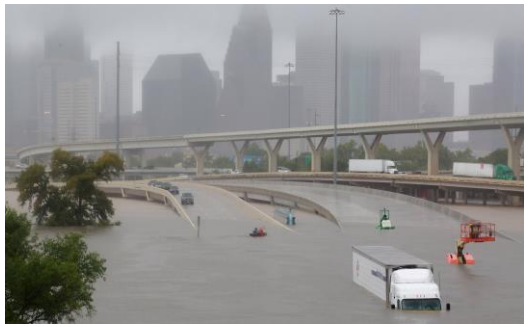


EDA: Sub-Tropical Weather and Ozone

Hurricane Harvey (August 2017)

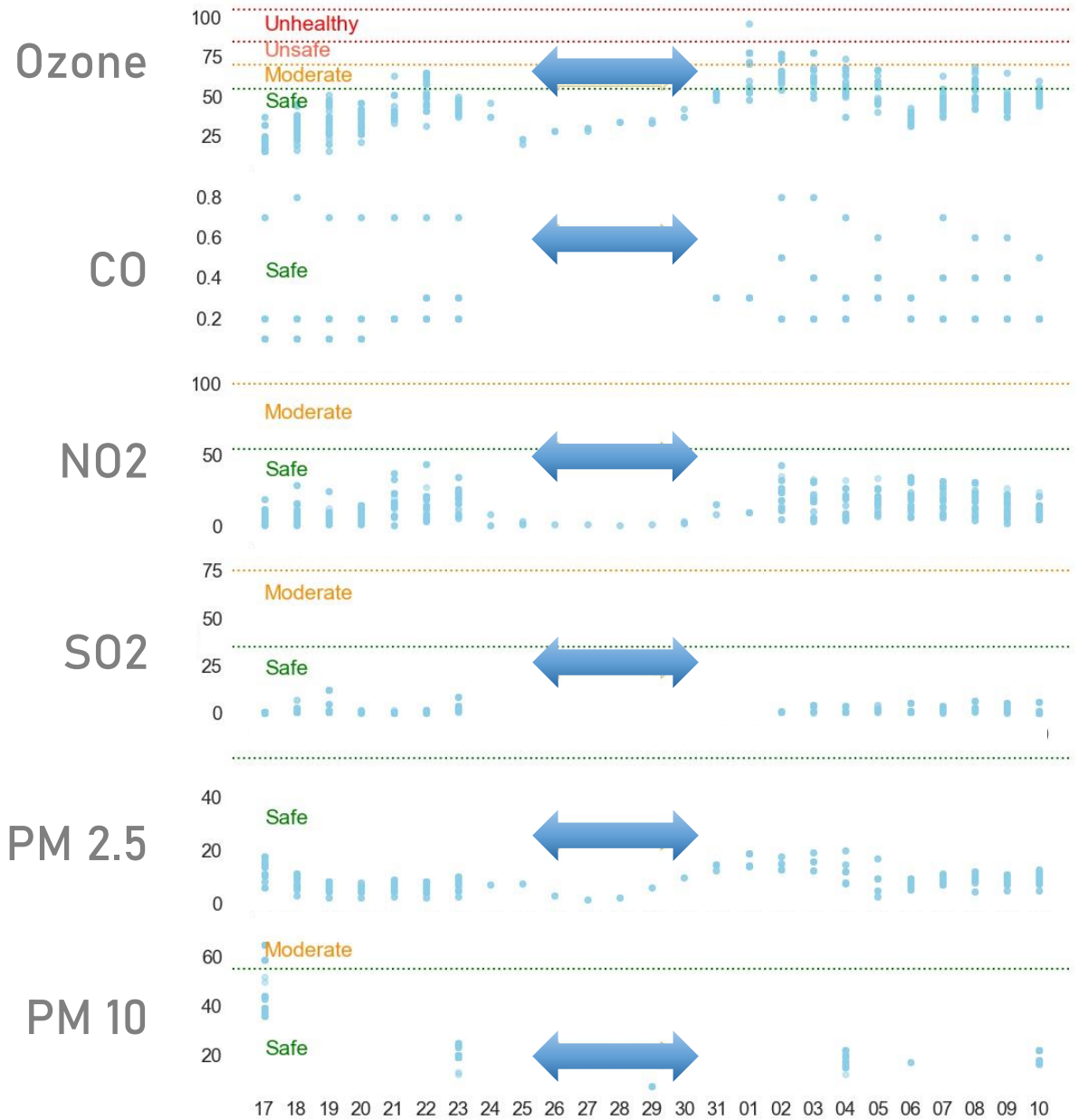






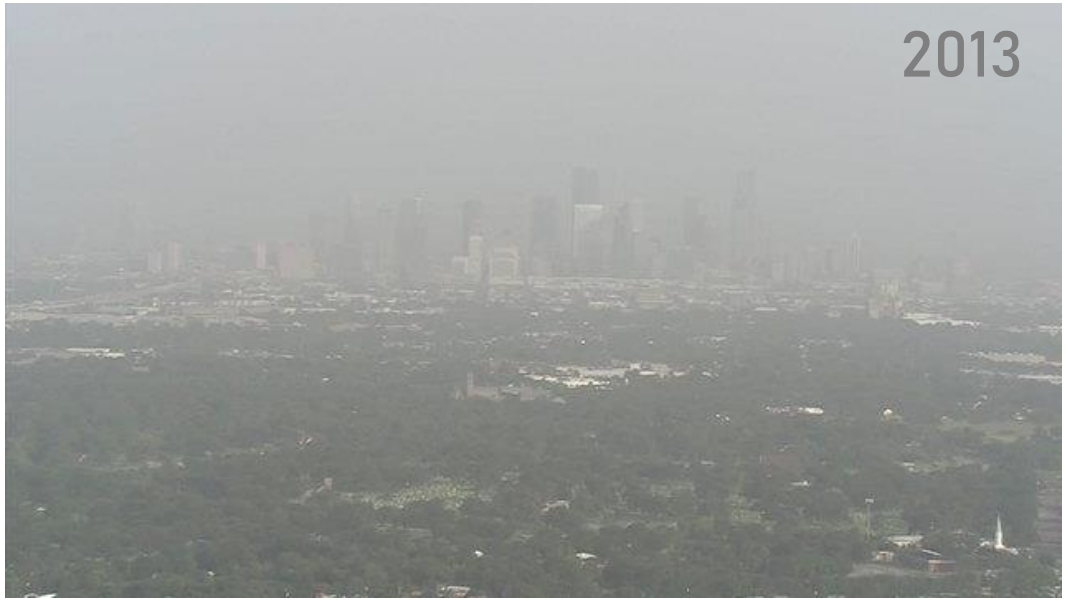
https://cdn.mos.cms.futurecdn.net/EtDugten6ufEikJDxm5JnS.jpg

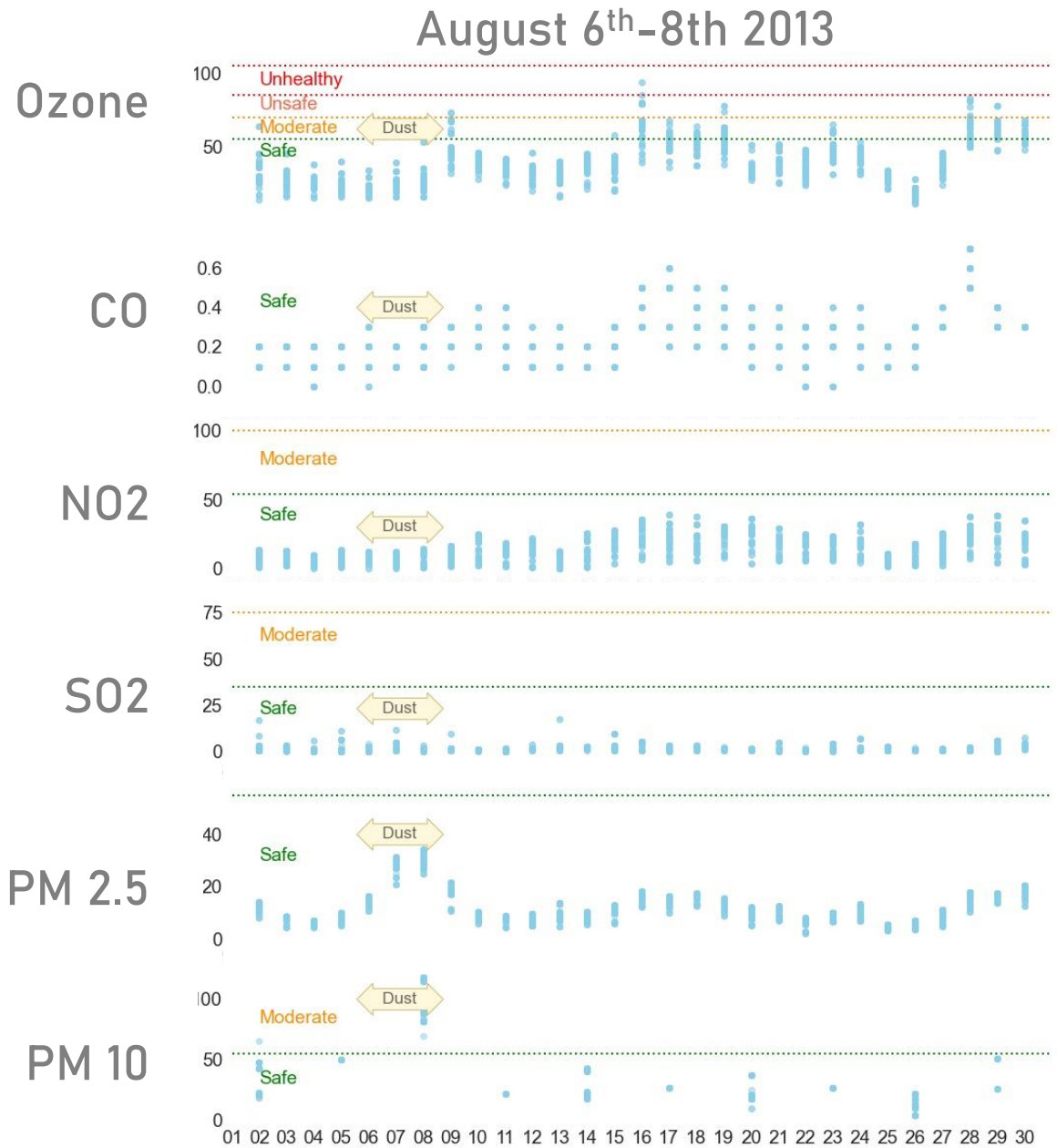
https://s.hdnux.com/photos/65/02/14/13902615/3/920x920.jpg



http://prod-upp-image-read.ft.com/ce281e52-8ba1-11e7-9580-c651950d3672

EDA: Saharan Dust and Particulate Matters





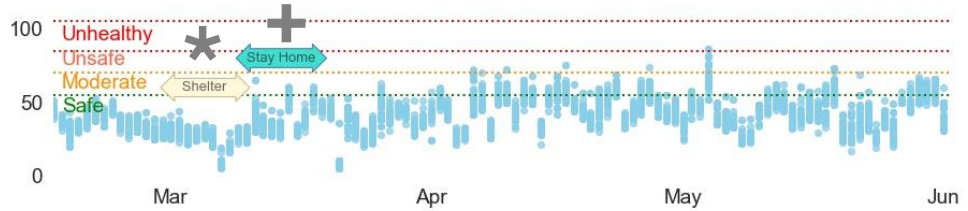
https://arc-anglerfish-arc2-prod-gmg.s3.amazonaws.com/public/5LL7DHMLDVAHHLFE42THFAVDLQ.jpg

https://cdn.abcotvs.com/dip/images/6272062_062620-ktrk-houston-skyline-saharan-dust-img.jpg

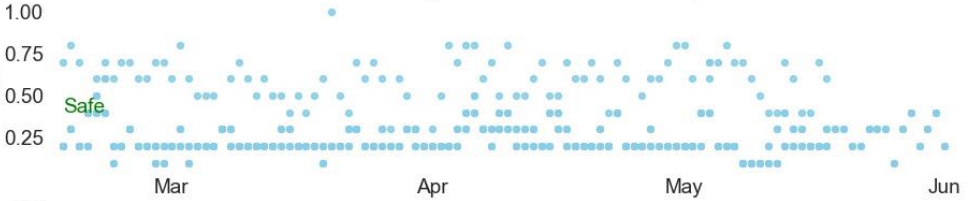
EDA: Pandemic Stay In Shelter (*) and Stay Home (+) Orders

March 13th – April 3rd 2020

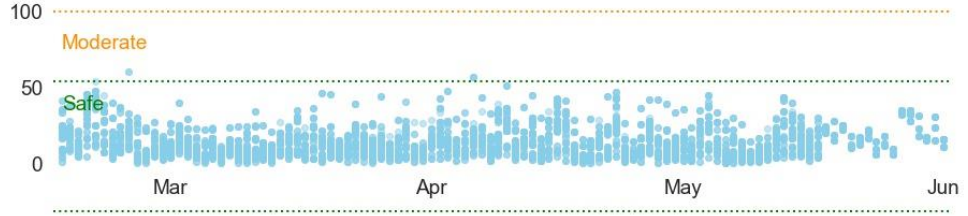
Ozone



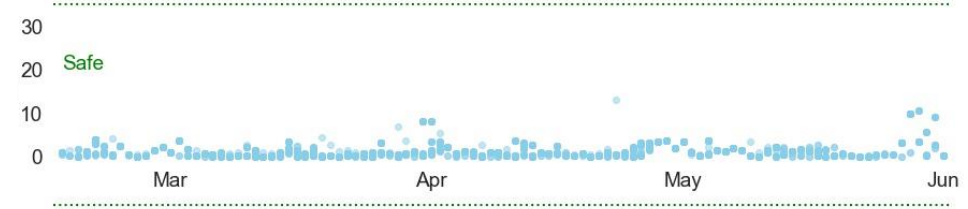
CO



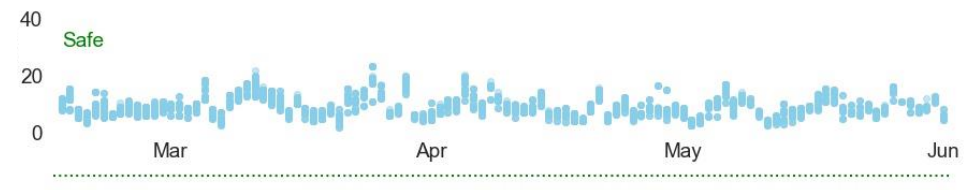
NO2



SO2



PM 2.5



PM 10

https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcTiJVQmHlM2g_cdCbQWysA5htrCU9W1anXNzA&usqp=CAU
https://www.khou.com/article/news/local/new-stay-home-work-safe-order-to-be-announced-as-harris-county-moves-to-covid-19-level-1/285-778b3e40-ef57-4de2-9e6c-cc4fddd10359

Choosing The Appropriate Model and Metrics

Two Possible Routes…

- Pollutant concentrations — difficult to predict a daily number
- AQI labels — huge imbalance in the dataset (1/14,000)

So Many Models, So Little Time…

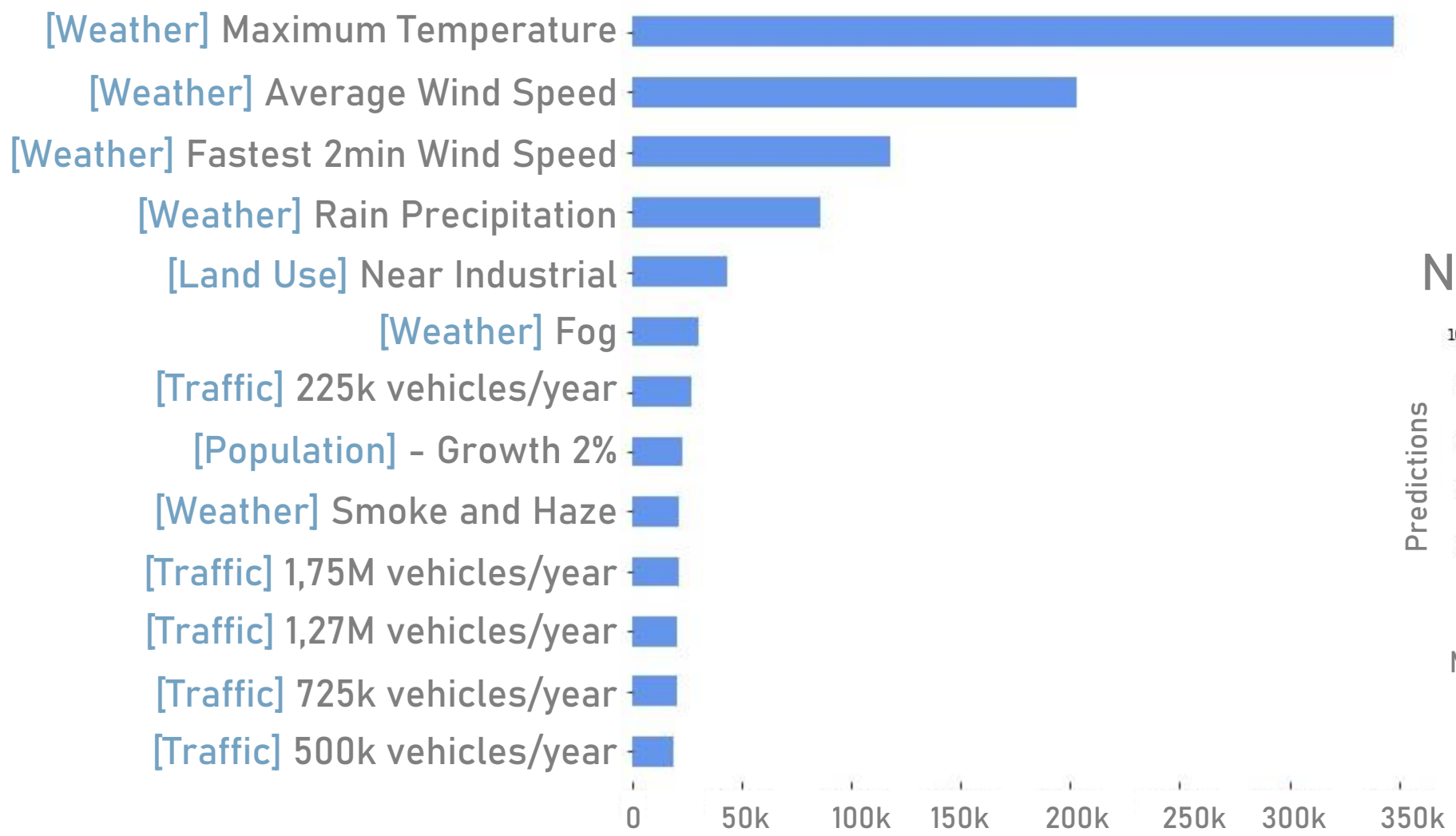
- Linear, multi-linear regression — this is not a linear problem
- SVR with RBF kernel — great training, fail testing
- Logistic Regression, SVM, KNN — do not capture imbalance, SMOTE does not help
- XGB Classifier — does not capture imbalance testing

The Chosen Model and Metrics…

- Model: XGB Regressor — objective: reg:squarederror, booster: gbtree, colsample_bytree = 1, subsample =1, eta = 1.3, maxdepth=20
- Metrics: MAE, RMSE — MAE to control the mean trend, as the aim of the model is to show trends rather than predicting daily data. RMSE to keep an eye on the distance between measured daily data and predicted daily data.

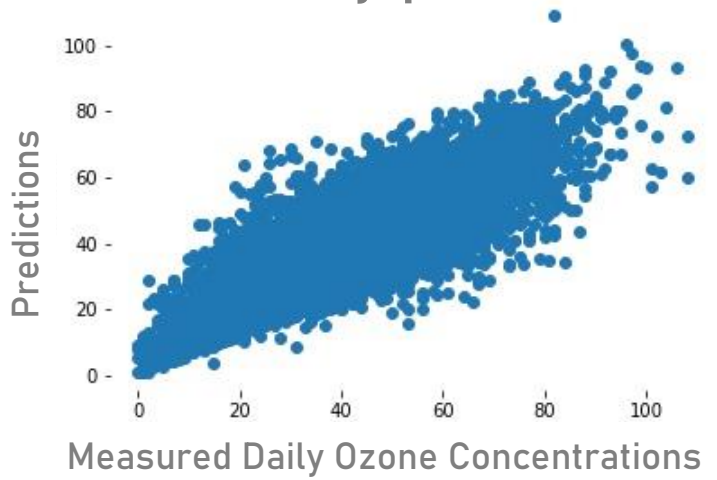
XGBoost Cross Validation Results: Hypothesis Is Wrong

Features of Importance Show Main Driver is Weather



MAE = 4.3
RMSE = 6.2

Not for daily predictions

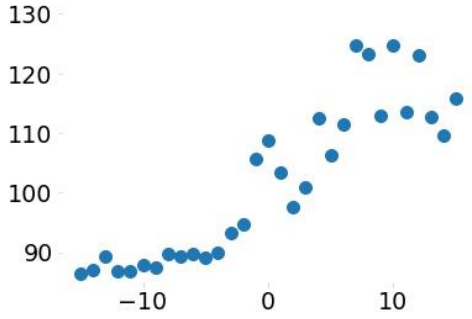


Model Application: Individual Contribution of Factors

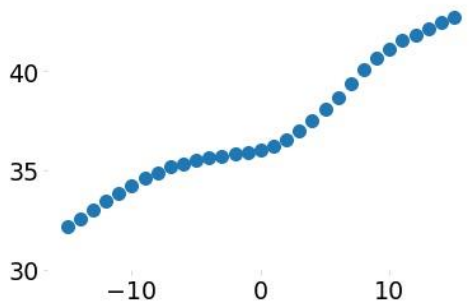
+/- Maximum Temperature

the higher, the more ozone, until a threshold is reached

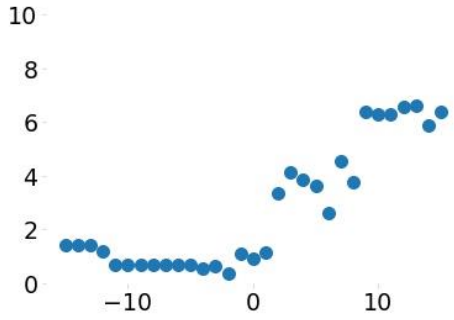
Max Ozone



Mean Ozone

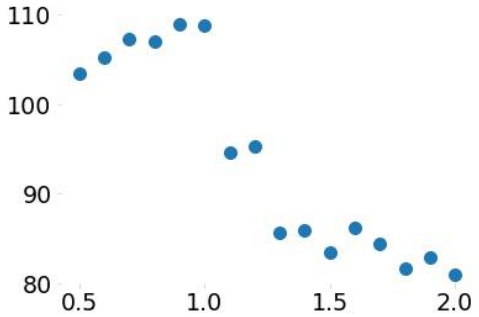


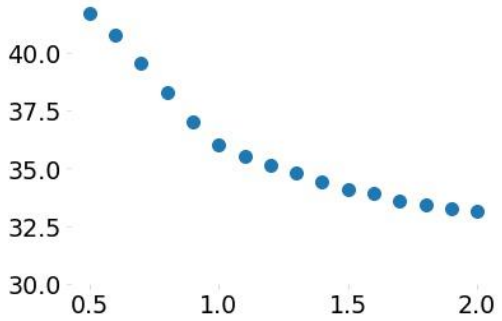
Min Ozone

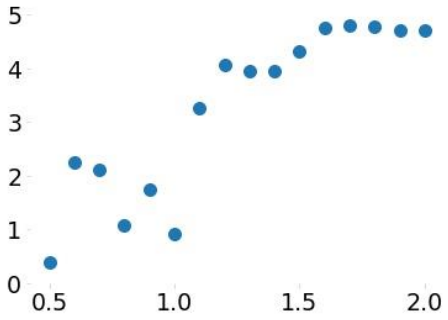


x Average and Fastest Wind

when high pushes ozone away, when low ozone is stagnant

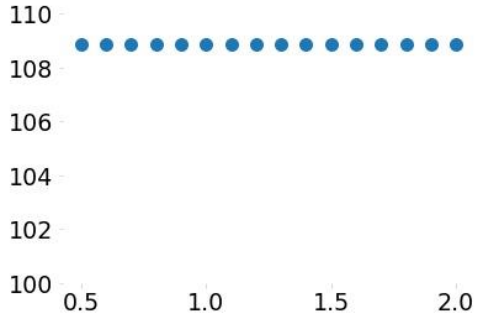


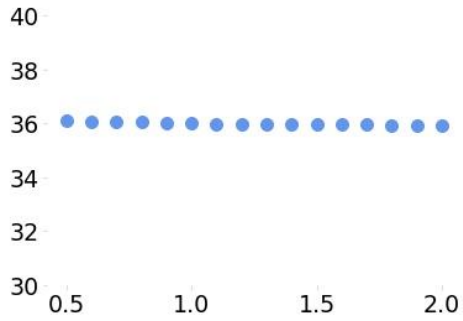


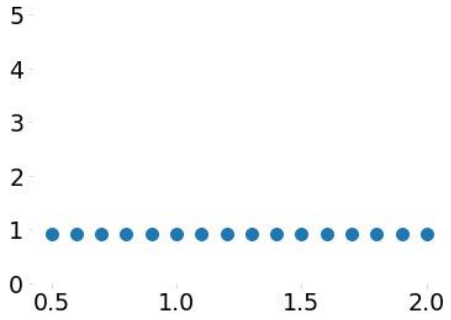


x Daily Precipitation

no influence

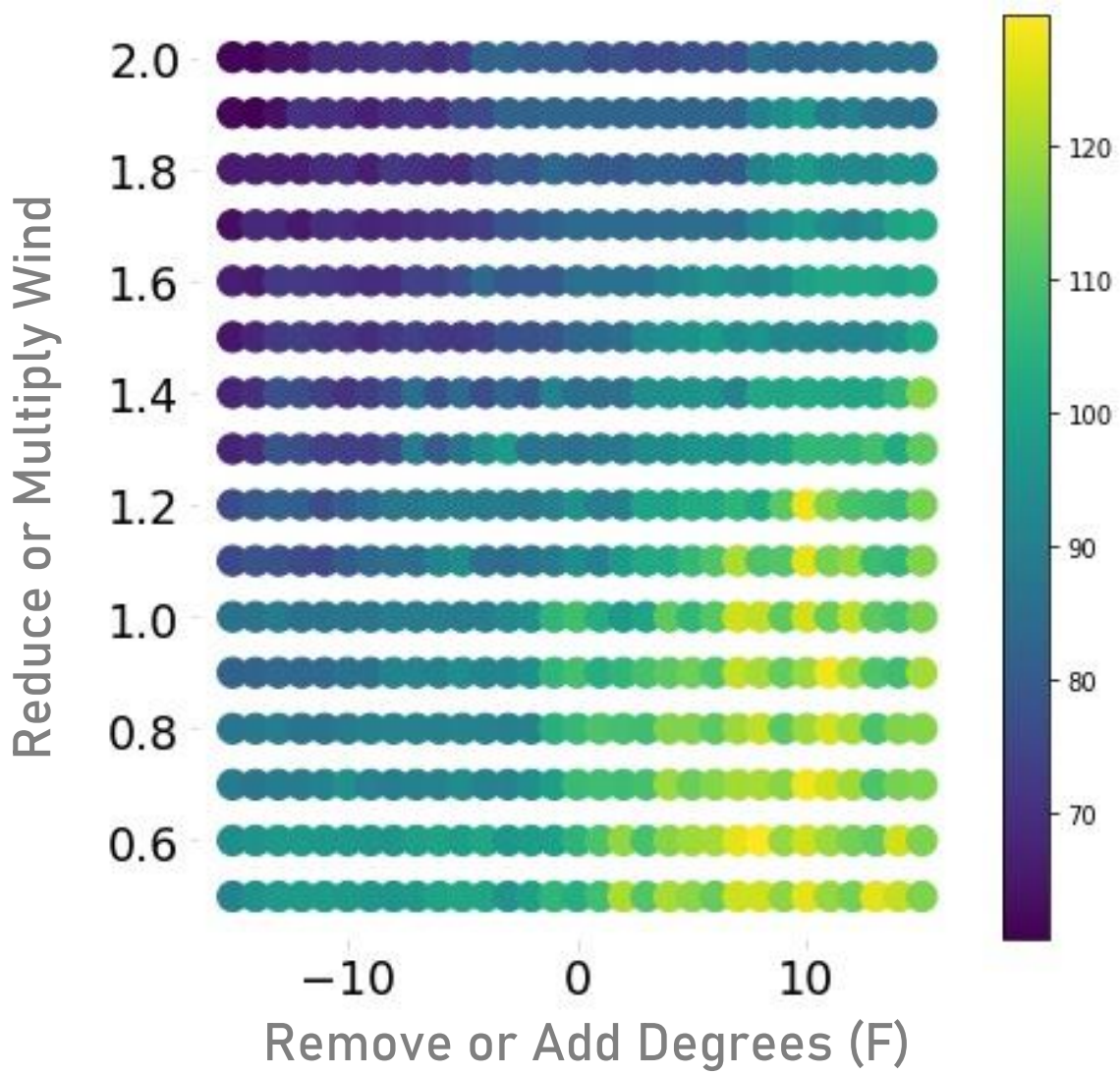




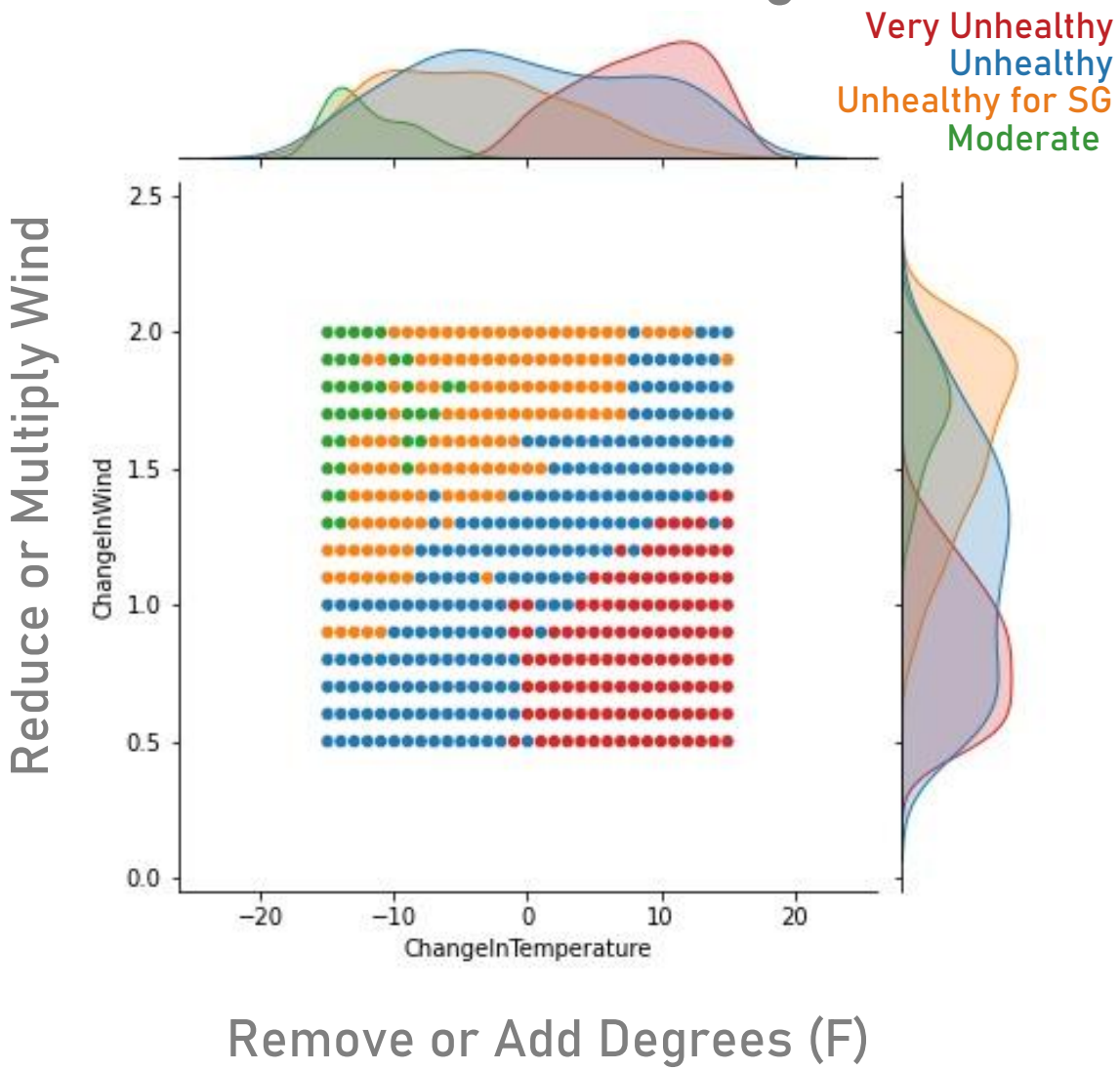


Model Application Results: Drivers and Buffers

Ozone Concentrations during BADs



Ozone AQI Label during BADs



Summary and Conclusion

The hypothesis was wrong	~~More People, More Traffic, More Pollution~~
The data showed	Ozone is the most concerning pollutant
The model showed	Maximum temperature and wind speed are the main drivers
The model can be applied to	Explore how the drivers influence ozone concentrations Explore how the drivers may catalyze or buffer each others Predict number and intensity of BADs
The model can be improved	Integrate traffic and population data differently (involving emissions?) Use land use surfaces and frequency Add VOC and additional NOx to model ozone formation Add geographically defined industrial emission (the data set has none) Beat the imbalance of AQI labels using Deep Learning
The interesting bits	The data contradicted the bias of the data scientist (too many hours spent in traffic) The model is built to show trends, not predict daily occurrences Thinking about climate change, hopefully Houston will get windier