Springboard

Houston Air Quality

Drivers and Buffers of Bad Air Days

Springboard Data Science Career Track
Capstone Project

Anne Warren

THE STORY

Since 2008, there is an average of 129 Bad Air Days (BADs) each year in Houston. During BADs, the Air Quality Index (AQI) labels may be "moderate", "unhealthy", and a few times a year "very unhealthy".
This project aims at identifying drivers of air pollution to enable the prediction of the outdoor air quality in Houston for decades to come in function of measurable factors. A secondary aspect of the project is to measure the impact of outdoor air quality on indoor air quality.

THE HYPOTHESIS

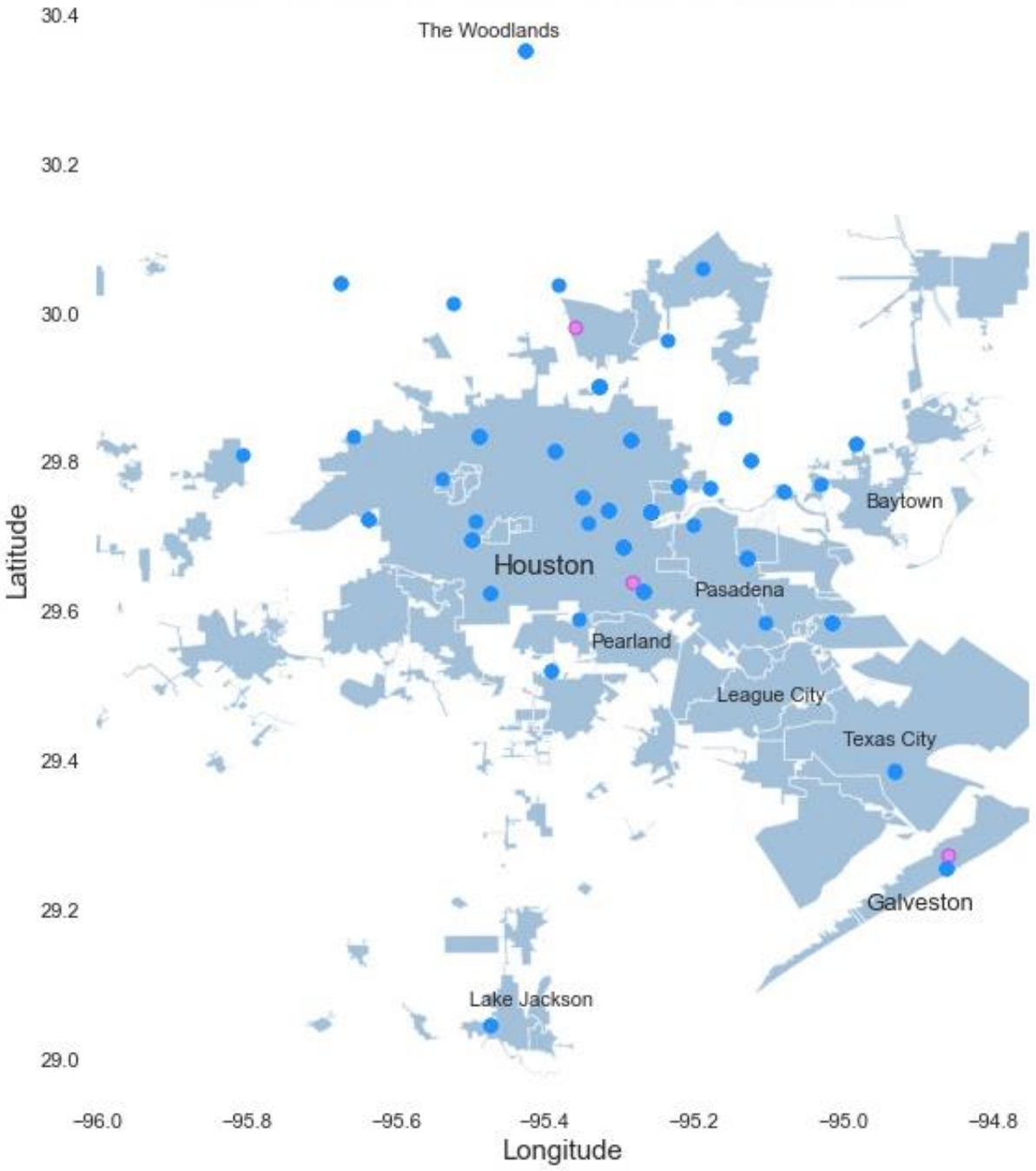
"More People, More Traffic, More Pollution"

Houston's traffic ranked 8th in the nation in 2020. To this daily pollution are added industrial emissions from refineries, plastic plants and other petro-chemical industries. Houston attracts an increasing number of workers (including the author of this report). According to the Houston-Galveston Area Council the population of Houston Metro will reach 9 millions in 2040.

THE DATA

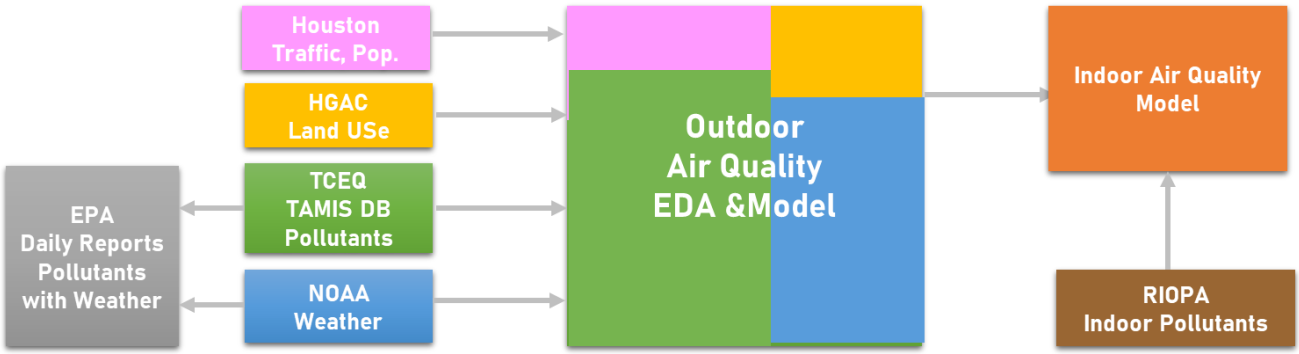
The data was collected from different sources:

- Pollutant Concentrations: TCEQ's Tamis DB (Texas Commission on Environmental Quality),
- Weather Data: NOAA (National Oceanic and Atmospheric Administration),
- Pollutant and Weather Data: US EPA (Environmental Protection Agency),
- Land Use: HGAC (Houston Galveston Area Council),
- Traffic and Population Data: City of Houston data website,
- Indoor/Outdoor Data: RIOPA study (Relationship of Indoor, Outdoor, and Personal Air).



DATA WRANGLING

The idea is to merge the data into a coherent data set by taking into account geographical locations, and by covering as much territory possible from The Woodlands to the North all the way down to Galveston to the Southeast on the coastline, including Baytown and Angleton (plants).

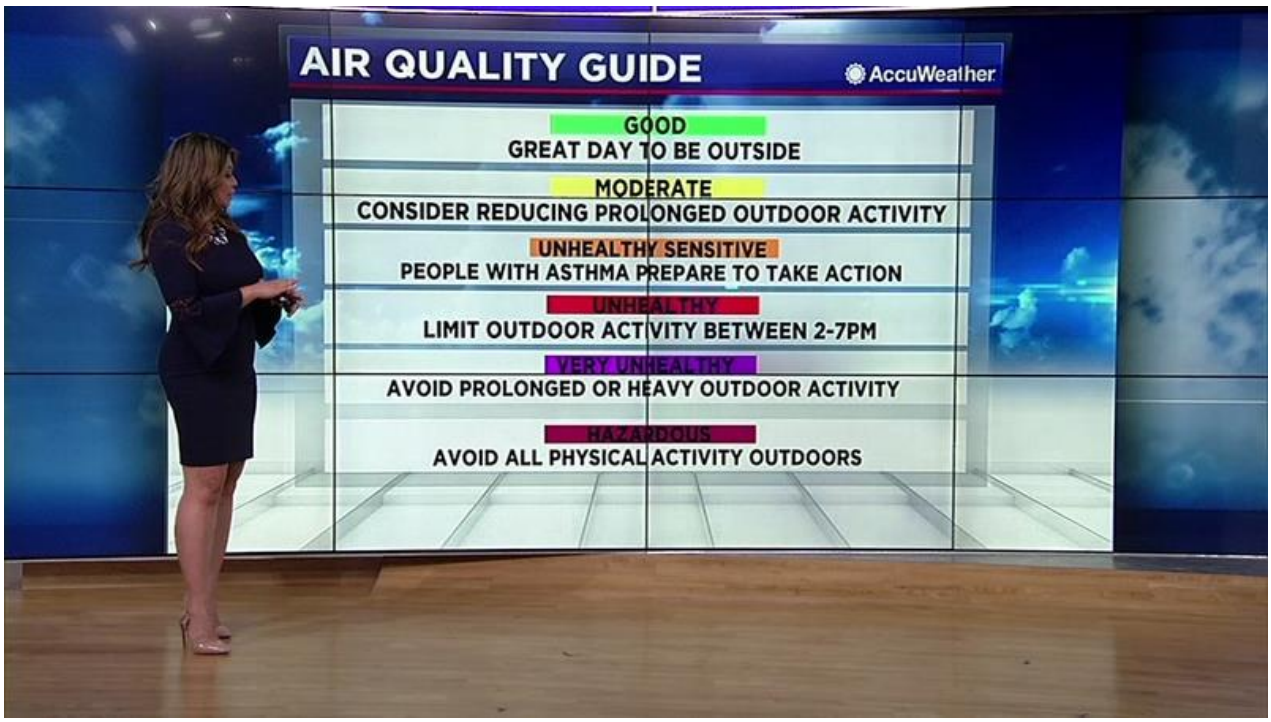


The main issues encountered during the wrangling of the data were:

- Matching sampling rates: the sampling rate (i.e. how many times a day or how long the sampling is) differs between pollutants. Only full daily records were kept following the way TCEQ reports to EPA (i.e. maximum values),
- Connect local weather to each air quality station: the weather data comes from three weather stations with N, SE and central locations,
- Deal with missing data: using appropriate imputation (time series driven) or by discarding the "day". Another problem was that the RIOPA data did not contain the location of the homes where indoor air sampling occurred. Consequently, the RIOPA model could not be connected to the outdoor air quality model,
- Traffic: Yearly traffic count was linked to the air quality stations within Manhattan distance of the traffic data point, subsequently summed and binned,
- Population: the data was integrated to the data set in term of "% growth" per year per neighborhood. The data was binned linked to the air quality stations within Manhattan distance.
- Land use: Houston is a mozaique of land use (i.e. no zonation) and therefore the notion of land use "near" another land use was added to the land use at the location (e.g. residential near land use). For instance, a given air quality station can be "Residential" and "Residential near Industrial".

AIR QUALITY INDEX (AQI)

The AQI describes the impact of pollution on the quality of life and daily activities. When applied to ozone concentrations the effects are as described in the picture below. The AQI is calculated using pollutant concentrations.



AQI Calculation and Reference Table for the US (Wikipedia)

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

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}.

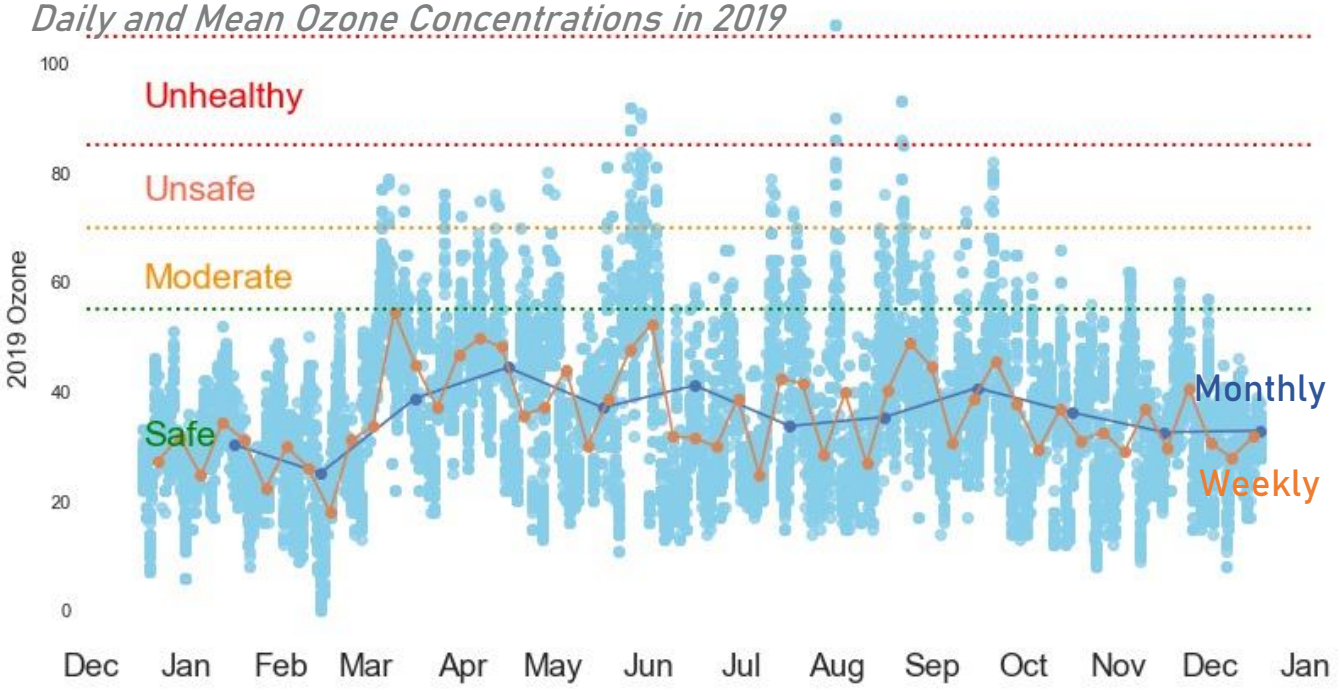
O_3 (ppb)	O_3 (ppb)	$PM_{2.5}$ (µg/m^3)	PM_{10} (µg/m^3)	CO (ppm)	SO_2 (ppb)	NO_2 (ppb)	AQI	AQI
C_{low} - C_{high} (avg)	C_{low} - C_{high} (avg)	C_{low}- C_{high} (avg)	C_{low} - C_{high} (avg)	C_{low} - C_{high} (avg)	C_{low} - C_{high} (avg)	C_{low} - C_{high} (avg)	I_{low} - I_{high}	Category
0-54 (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	

EDA: POLLUTANTS

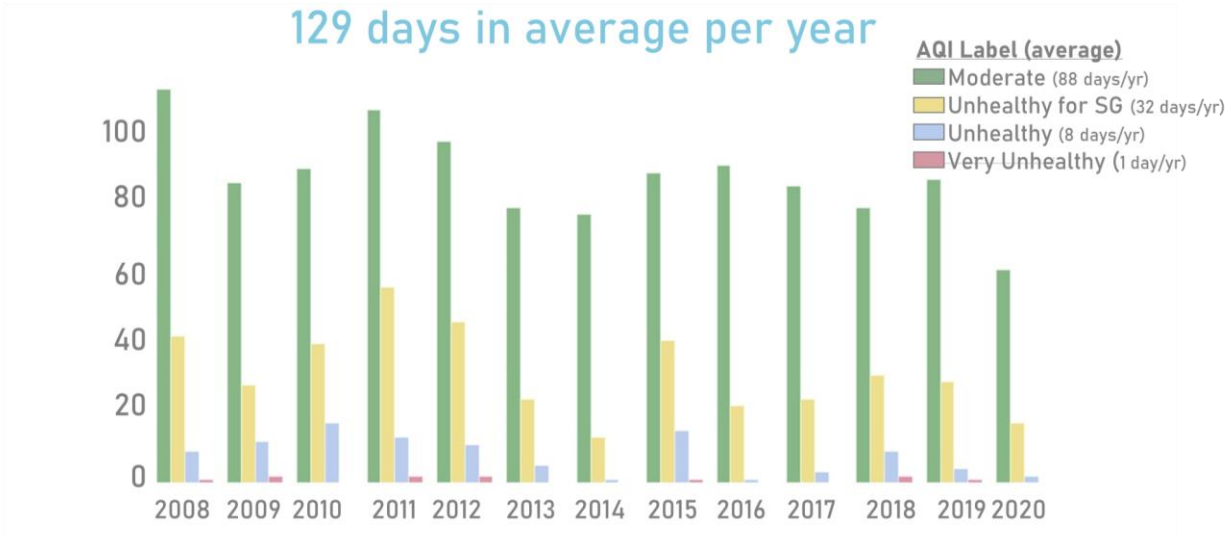
Ozone, NO2, SO2, CO, PM 2.5, PM 10, Lead

The Exploratory Data Analysis shows that between 2008 and 2020 Ozone is the problematic pollutant in Houston. The other pollutants remain in the "safe" zone and rarely go up to moderate if ever. Despite BADs, the weekly and monthly average ozone concentrations remain mostly in the "safe" zone. There is an average of 129 BADs per year which may be as low as "Moderate" (80-110 days per year) up to "Very Unhealthy" (1 to 2 days per year). Ozone levels are lower in Winter and Spring.

Daily and Mean Ozone Concentrations in 2019



BADs Due To Ozone (2008-2020)

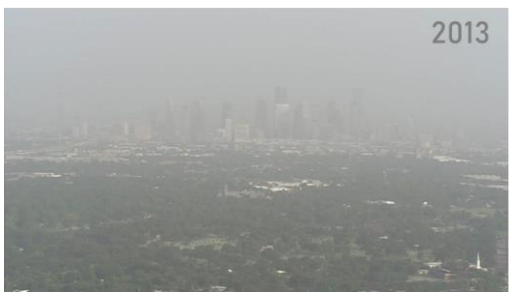


EDA: Not BAD

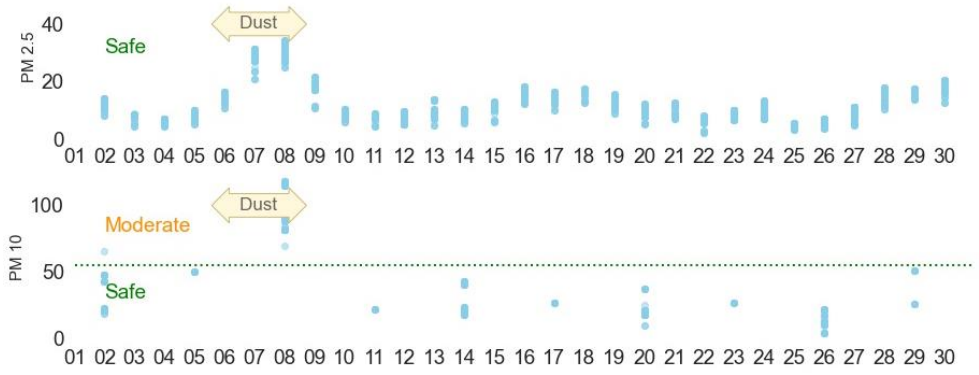
It looks bad but it is not BAD

Interestingly the EDA showed that severe weather or special events as bad as they look did not lead to a BAD.

- Hurricanes: Harvey (August 25-29th, 2017 - cat 4) and Ike (September 13th, 2008 - cat 2) lead to very little records during the event because the air quality stations are turned off and sheltered. There was no rebound of or decrease in pollution after the event.
- Flooding: Tropical Storm Imelda (September 19th, 2019) and Memorial Day Flood (May 25-26th, 2015 - 11 inches of rain in 9 hours) do not disturb the records.
- Chemical Plant Fire: Deerpark ITC fire (March 17-20th 2019) lasted several days and released a large plume of smoke that covered Houston but the PM pollutant records (VOCs might have spiked).
- Pandemic: 'Stay Home' order during the Covid-19 outbreak (March 16th - April 30th 2020) helped ozone to reach very low levels in March until March 24th which was the day after the media announced that Judge Hidalgo would declare a "Stay at Home" order to replace the "Shelter in Place" order.
- Saharan Dust Storm: The Dust Storms always make the news especially in August 6th-8th 2013 and in June 2020 when the dust haze was highly visible (see picture below). The particulate matter PM 10 barely reached "Moderate" while PM 2.5 remains in the safe zone.



PM Data During The Saharan Dust Storm (August 2013)



MODEL AND METRICS

XGB Regressor with MAE and RMSE

The model could be built focusing on ozone concentration values or AQI labels:

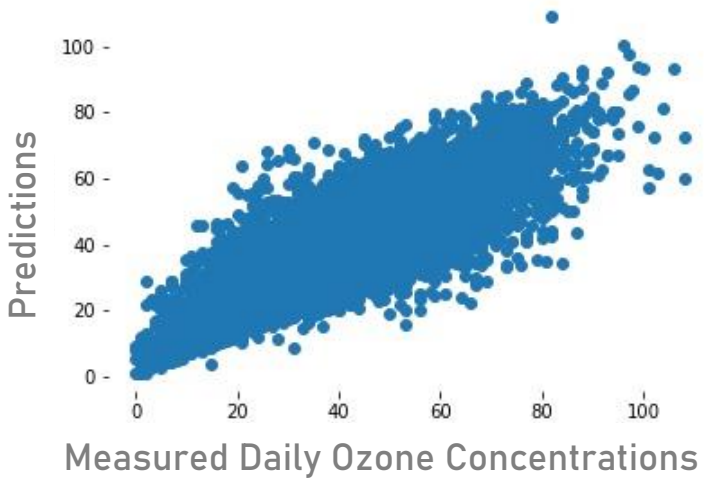
- In practice the dataset was too imbalanced to allow the detection of the rarer labels "Unhealthy" and "Very Unhealthy" by algorithms like Logistic Regression, SVM, KNN and XGB Classifier. Applying SMOTE to the dataset did not help the classification models because it forced the model to overfit the minority classes.
- Linear and multi-linear regression do not fit the problem.
- SVR with RBF kernel was performing well with the training set but performed poorly with the testing test.
- XGB Regressor provided the best cross-validation results.

The XGB Regressor was tuned using the following parameters:

- reg:Squarrederror objective with gbtree booster,
- colsample_bytree and subsample set to 1,
- Eta at 1.3 and 20 for maxdepth.

MAE is the primary metric because the aim of the model is to show trends rather than predicting daily data. RMSE is used to keep an eye on the distance between daily data and predicted data.
The metrics of the final model are: MAE = 4.3 and RMSE = 6.2.

The model is not built for predicting daily values

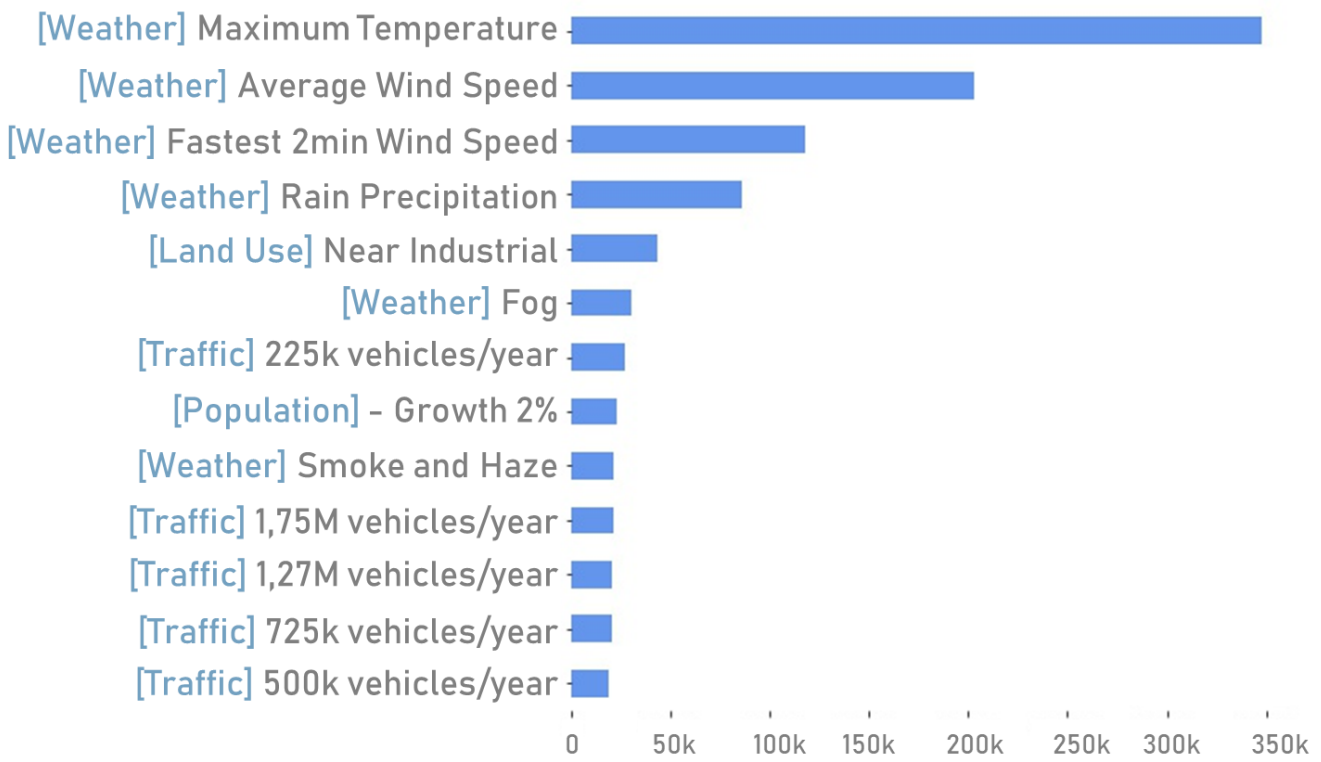


MODEL RESULTS

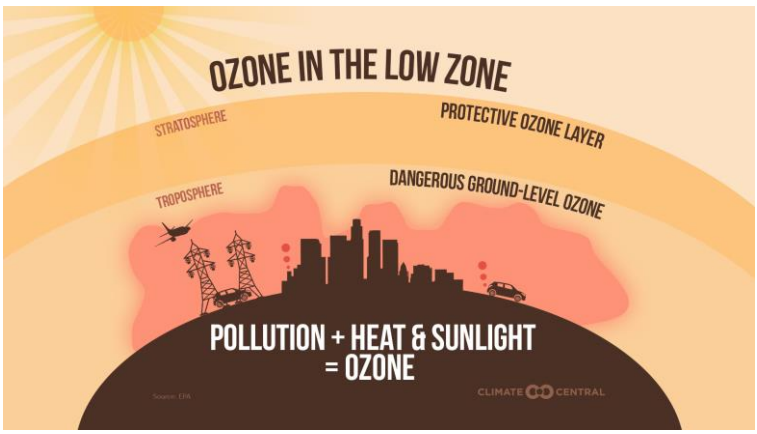
The Main Driver is Weather

The features of importance show that maximum temperature, wind speed and daily precipitation are the main drivers behind ozone concentrations in the air. Being near an industrial zone seems to play a role thought minor. Population and traffic have little influence.
The model shows that the hypothesis “More people, more traffic, more pollution” is wrong.

Features of importance



This outcome is not surprising because temperature is a catalyst in the formation of ozone. Regarding the wind, it may play a role of transportation, concentration or dispersion of ozone or of the chemicals leading its formation. The model can be used to see how this works in practice.



MODEL APPLICATION

The model is used to see what would happen to ozone concentrations levels (low, average and BADs) when the major drivers increase or decrease. For this the degrees are added or removed from the maximum temperature while the wind speed and total precipitation are multiplied by a factor.

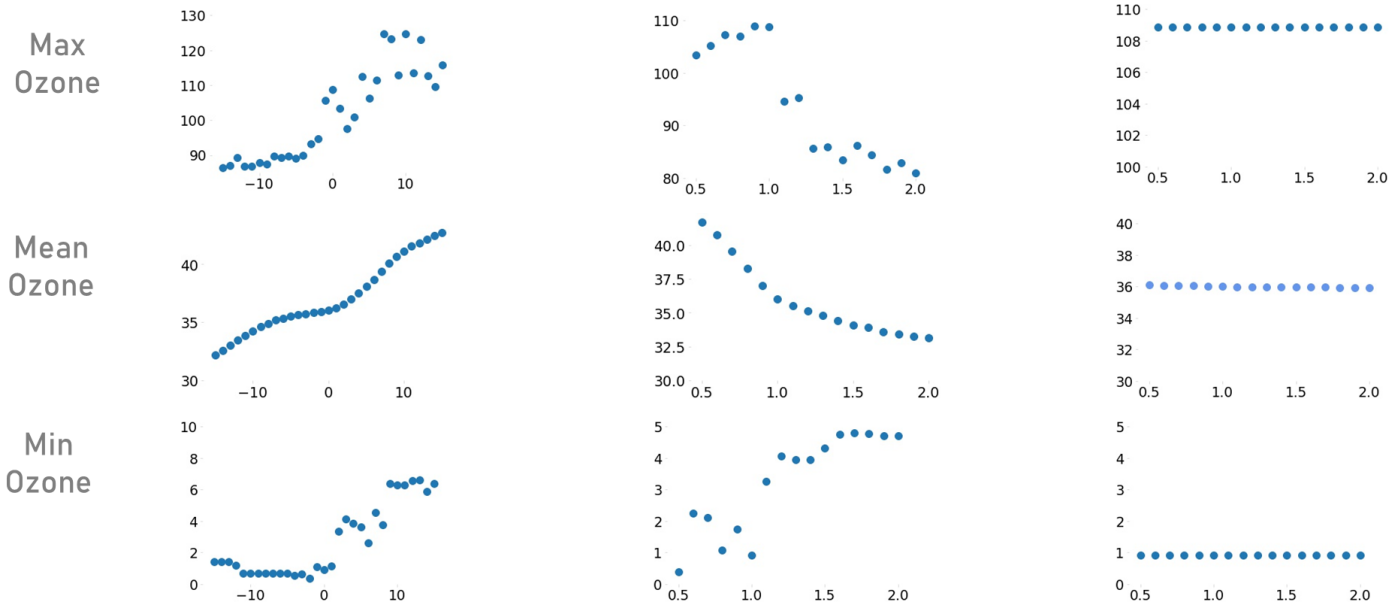
- Temperature is a catalyzer on all levels of ozone though,
- High winds play a dual role of dispersion on high level of ozone but tends to catalyze ozone formation in zones where ozone is low.
- Rain seems to have no influence by itself.

Individual contribution of drivers

+/- Maximum Temperature
the higher, the more ozone, until a threshold is reached

x Average and Fastest Wind
when high pushes ozone away, when low ozone is stagnant

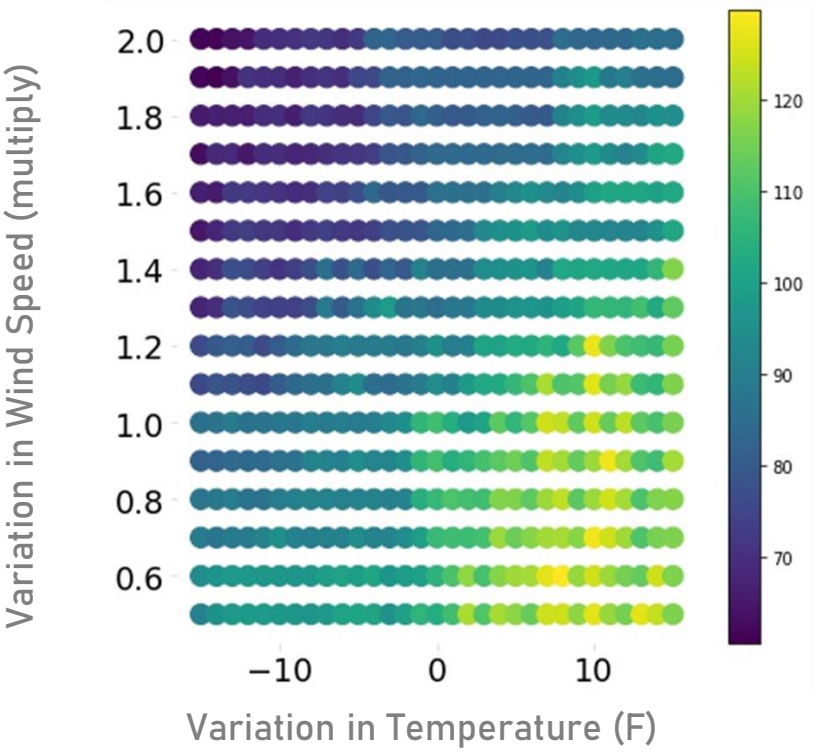
x Daily Precipitation
no influence



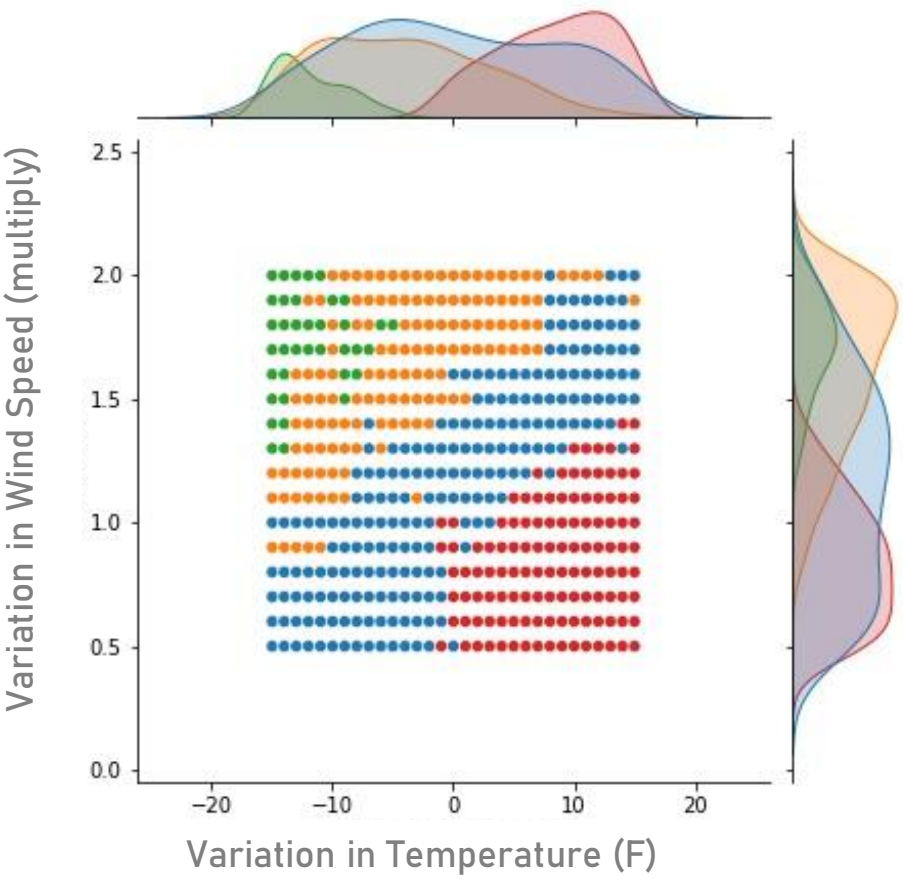
There is a very interesting game of buffer and driver between temperature and wind speed. High wind speeds tend to buffer the action of higher temperature which can be explained by the role of dispersion of the wind (i.e. dispersion of the ozone and also the dispersion of the chemicals involved in its formation). For instance, at the highest temperature (+15F) increasing the wind speed by 50% helps shifting the air quality from “Very Unhealthy” to “Unhealthy”.

DRIVERS AND BUFFERS

Ozone during BADs when temperature and wind speed vary



Ozone Concentration (ug/m3)



AQI Labels

Very Unhealthy
Unhealthy
Unhealthy for SG
Moderate

CONCLUSION

The most concerning pollutant in Houston is ozone. The hypothesis saying that more people in Houston would lead to more traffic and hence more pollution was wrong. The main driver of ozone pollution is the maximum outdoor temperature followed by wind speed.
The model is built to show trends as opposed to predict daily values. The MAE metric works for this purpose.

The model can help explore the influence of the drivers of ozone concentrations and also how they can catalyze or buffer each other. The model also can be used to predict the number and intensity of BADs due to ozone.

Increasing maximum temperatures tend to increase ozone concentrations while increasing average and fastest wind act as a buffer. The wind seems to disperse the ozone or the chemicals leading to ozone.

There are many ways the model can be improved:

- Traffic and population could be integrated differently,
- The surface and geographical frequency of land use could be added.
- VOCs emission could be added, which is part of the ozone formation equation.
- Industrial emission could be added, if available with geographical details.
- The imbalance of the AQI labels could be overcome using a Deep Learning model.