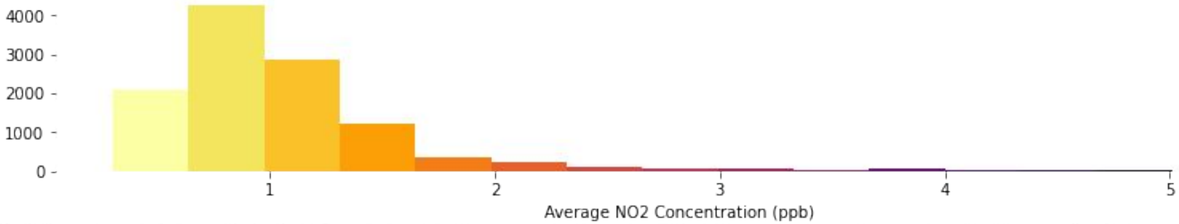
Hyperlocal Air-Quality Prediction in Houston, TX

Kaveri Chhikara, Carolyn Vilter, Vishal Vincent Joseph, Vignesh Venkatachalam

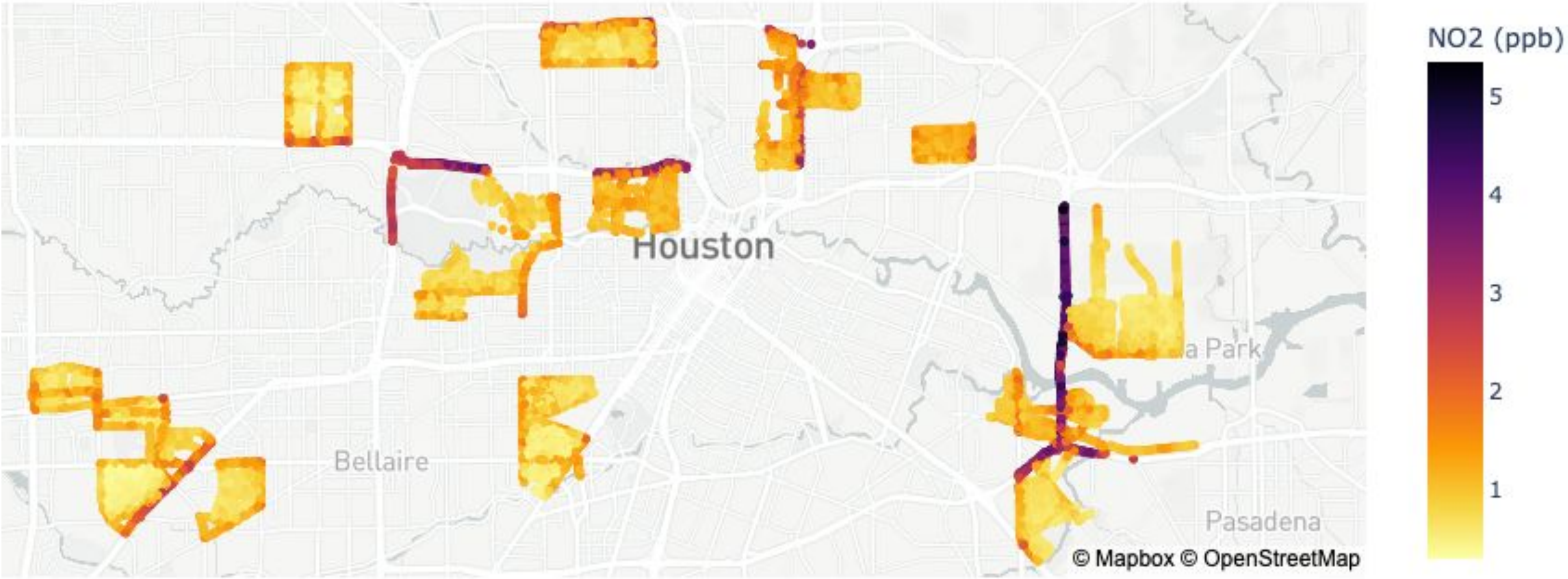
Background and Motivation

- Existing air monitoring network in the US today is shockingly sparse and the USEPA currently has < 2500 unique monitoring stations measuring criteria pollutants
- EDF and Google partnership to create hyperlocal sensor networks in multiple cities across the world using Google Street View Cars
- Our Focus:
  - <u>Location:</u> 22 neighborhoods of Houston
    - Energy Capital of US
    - Currently lacks zoning
    - Robust project methodology
  - <u>Target pollutant:</u> NO2
    - Better data availability
    - Appreciably higher levels closer to pollution sources
  - <u>Time Period:</u> Jul 2017 - Mar 2018

Data Sources

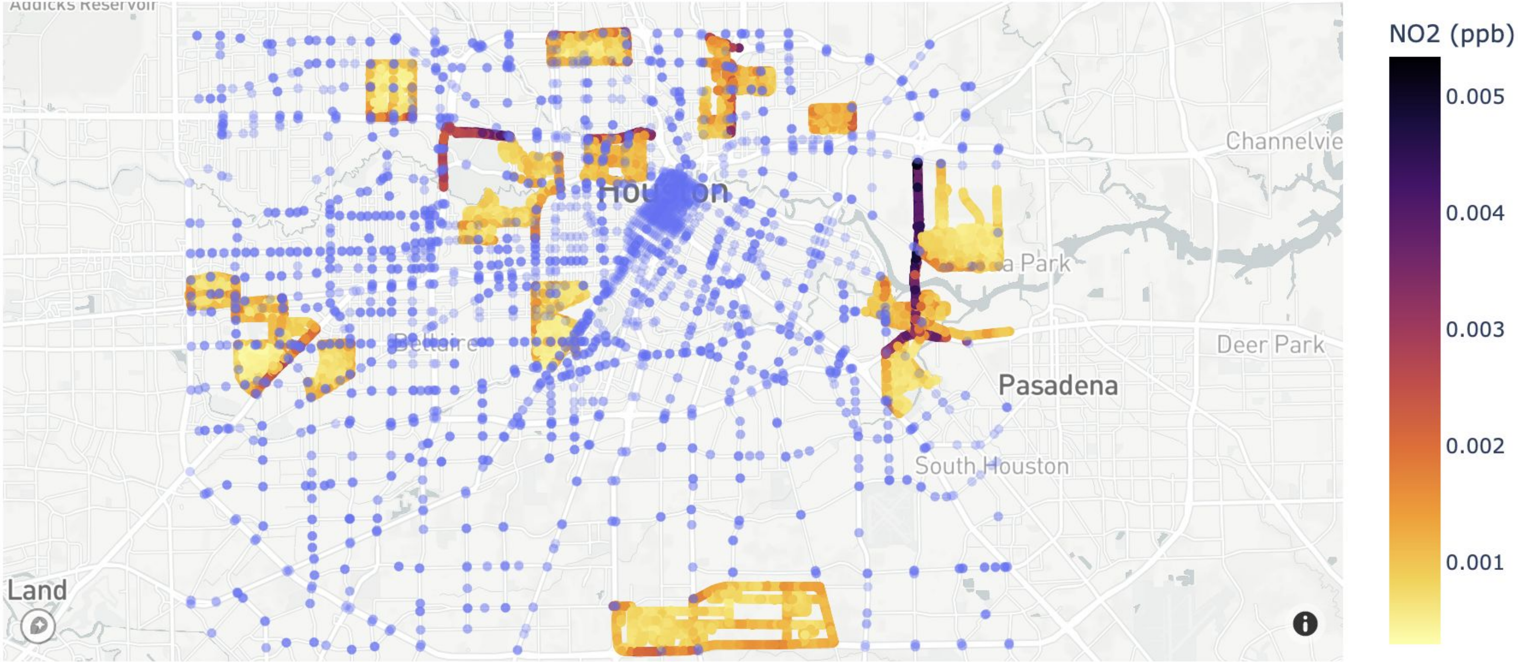


1. Google-EDF hyperlocal NO_2 monitoring data



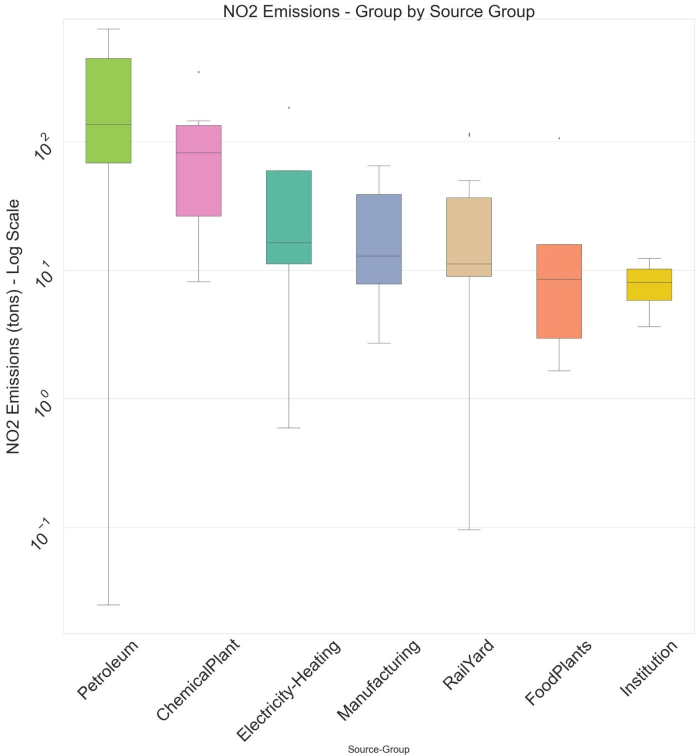
Data Sources

2. Traffic Intersections data

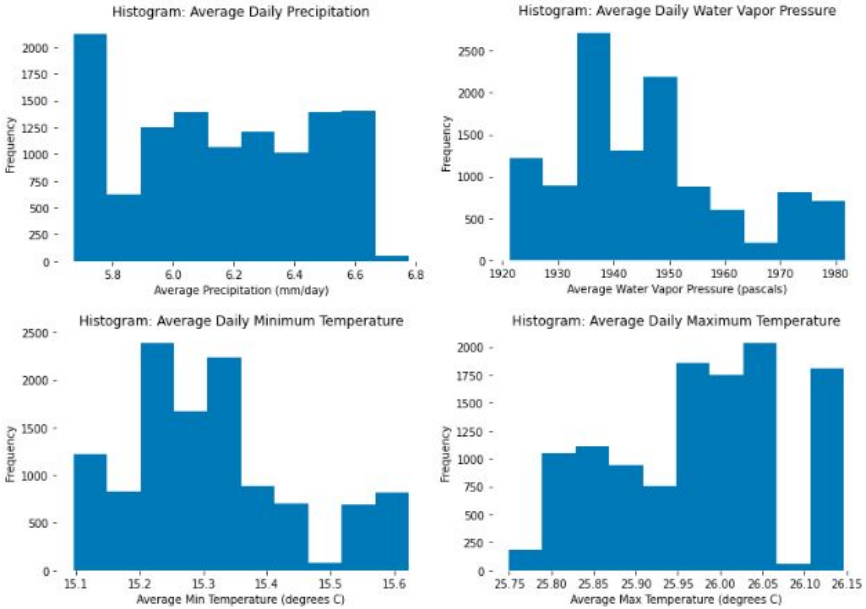


Data Sources

3. EDF Pollution Facilities data



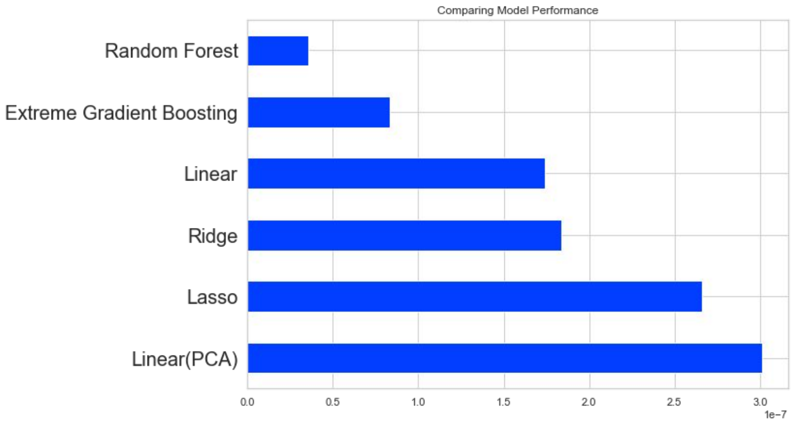
4. DAYMET 1km-grid weather data

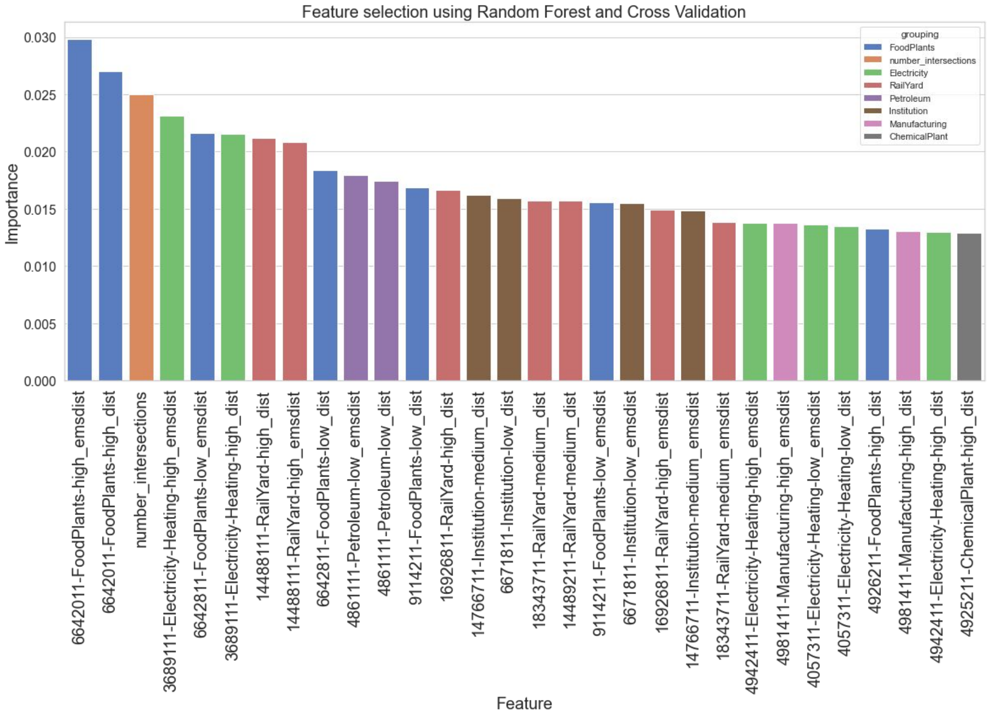


Models: Feature selection, CV & hyperparameters

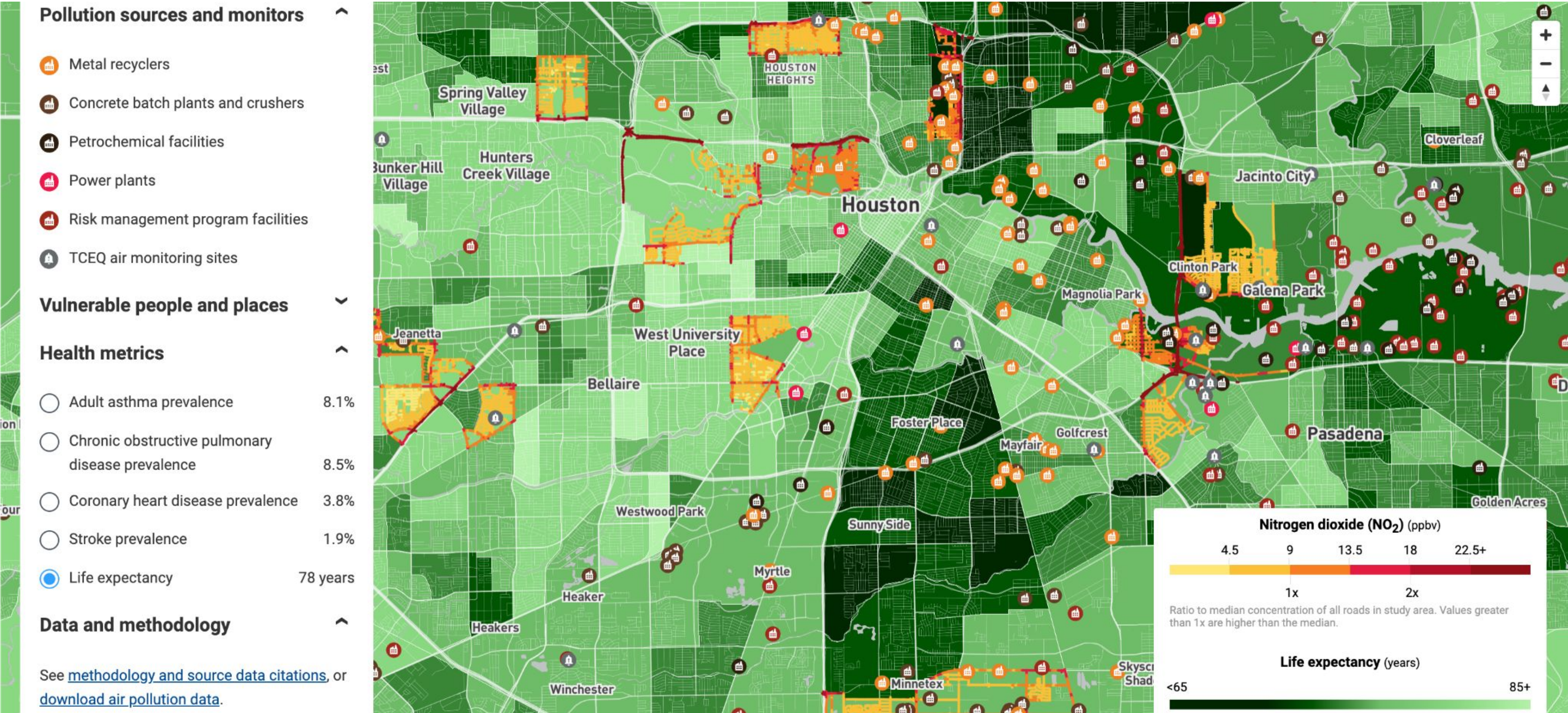
Class	Model	Feature Selection	Cross Validation	Hyperparameters
Linear	OLS (Baseline)	Univariate Screening (Pearson's Correlation)		
	Linear Regression	w/ & w/o PCA		
	Lasso Regression	Embedded	GridSearchCV	`alpha`: [$1e^{-5}$, $1e^{-4}$, $1e^{-3}$, $1e^{-2}$, $1e^{-1}$, 1]
	Ridge Regression	Embedded	GridSearchCV	`alpha`: [$1e^{-5}$, $1e^{-4}$, $1e^{-3}$, $1e^{-2}$, $1e^{-1}$, 1] `solver`: ['auto', 'svd', 'cholesky', 'sag']
Tree Based	Random Forest	Embedded	RandomizedSearchCV	`n_estimators`: [100, 300, 500] `max_features`: ['auto', 'sqrt', 'log2'] `max_depth`: range(10, 50, 10)
	XGBoost	Embedded	RandomizedSearchCV	`n_estimators`: range(50, 500, 50) `max_depth`: range(5, 50, 5)
NN				

Model Evaluation and Feature Importance





Results: Selecting regions for NO_2 prediction



Pollution sources and monitors ˄

- ● Metal recyclers
- ● Concrete batch plants and crushers
- ● Petrochemical facilities
- ● Power plants
- ● Risk management program facilities
- ● TCEQ air monitoring sites

Vulnerable people and places ˅

Health metrics ˄

○ Adult asthma prevalence	8.1%
○ Chronic obstructive pulmonary disease prevalence	8.5%
○ Coronary heart disease prevalence	3.8%
○ Stroke prevalence	1.9%
● Life expectancy	78 years

Data and methodology ˄

See methodology and source data citations, or download air pollution data.

Nitrogen dioxide (NO_2) (ppbv)

4.5	9	13.5	18	22.5+

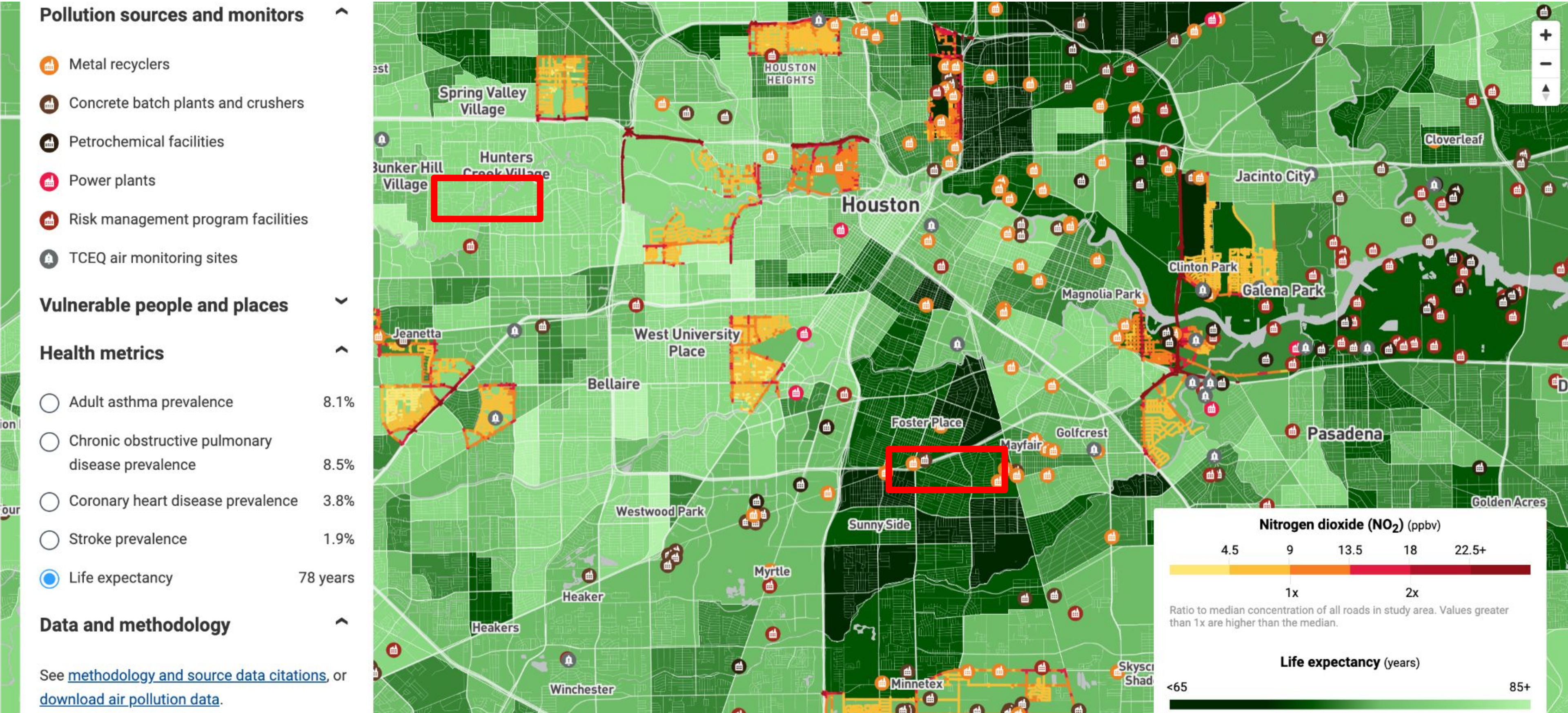
1x 2x

Ratio to median concentration of all roads in study area. Values greater than 1x are higher than the median.

Life expectancy (years)

<65 85+

Results: Selecting regions for NO_2 prediction



Pollution sources and monitors ˄

- Metal recyclers
- Concrete batch plants and crushers
- Petrochemical facilities
- Power plants
- Risk management program facilities
- TCEQ air monitoring sites

Vulnerable people and places ˅

Health metrics ˄

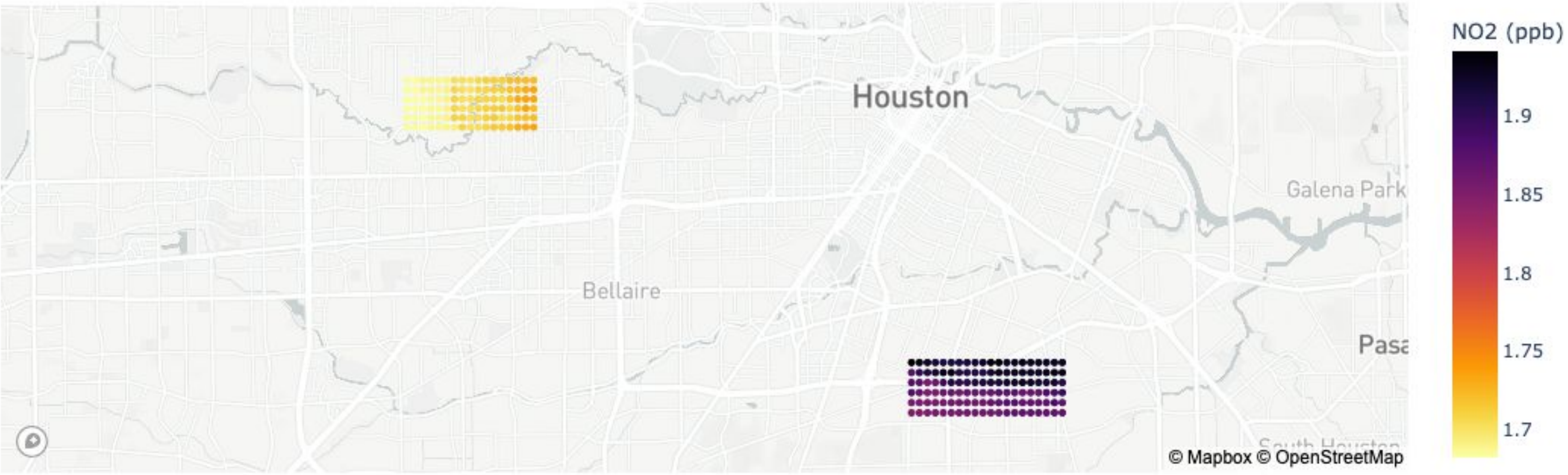
○ Adult asthma prevalence	8.1%
○ Chronic obstructive pulmonary disease prevalence	8.5%
○ Coronary heart disease prevalence	3.8%
○ Stroke prevalence	1.9%
● Life expectancy	78 years

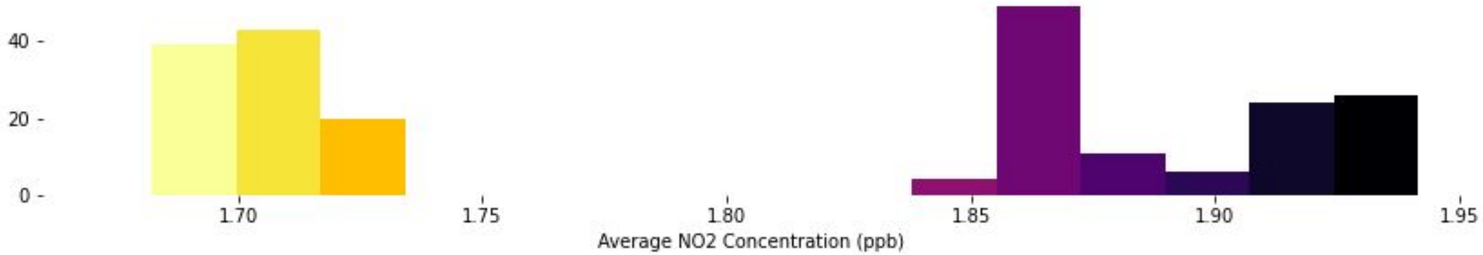
Data and methodology ˄

See methodology and source data citations, or download air pollution data.

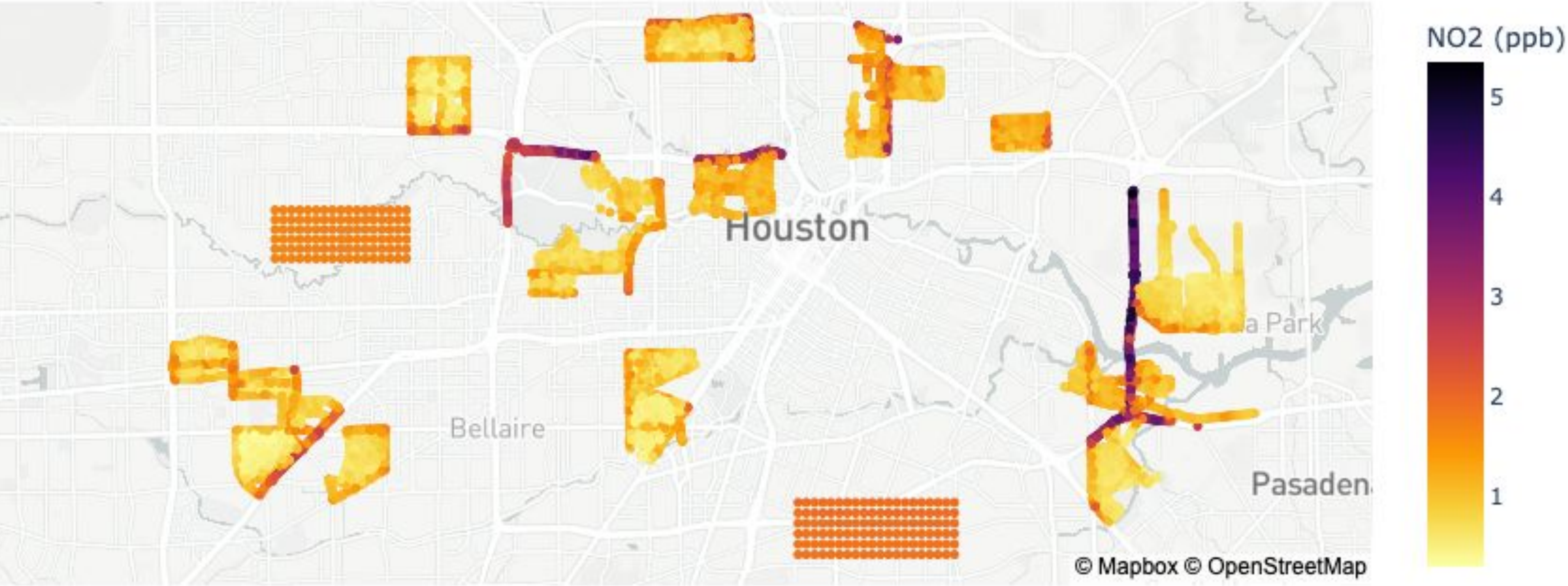
Results: NO_2 predictions for the selected regions

We observe relatively higher NO_2 concentrations in one of the regions





Results: NO_2 predictions in context of all Houston



In the larger context of all of Houston's NO_2 concentration, however, the two identified regions do not show much variation.

Future Improvements

- Resolving Missing Data:
  - Facilities classified as “unknown” are being dropped for now and will require manual work/identification

- Additional Target Variables with Policy Significance:
  - Further explore the relationship between Air Quality and key policy relevant factors like Life Expectancy, Asthma prevalence and Poverty

- Model Tuning:
  - Improve on existing Neural Network architecture by experimenting with additional layers, neurons and activation functions
  - Include more hyperparameters in sklearn models and tune using Grid Search instead of Randomized Search for tree based models