

Is effect of increased warehouse presence on health outcomes quantifiable ?



As Data Scientists in *OEHHA*, we are tasked with developing models aggregating additional information on warehouse density to assess primary mitigating factors addressing negative health outcomes.

- How well do the CalEnviroScreen scores reflect emergency healthcare counts?
- What indicators from the CalEnviroScreen dataset best determine the number of emergency healthcare visits?

Data Source: California EnviroScreen reports



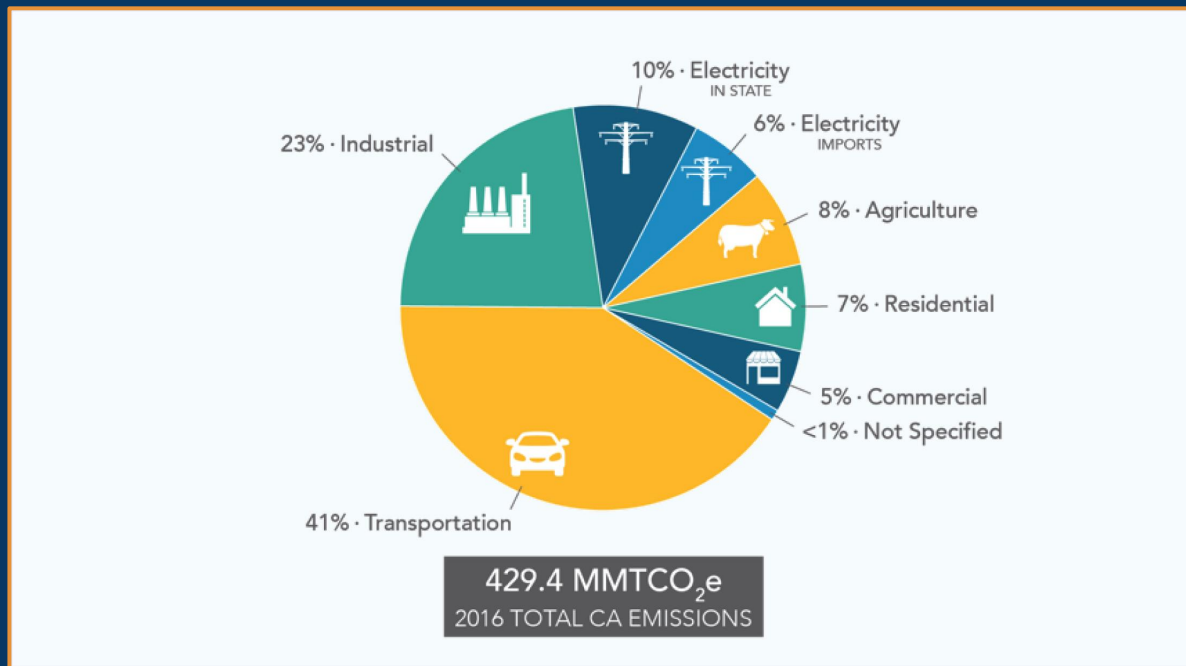
From the **California Office of Environmental Health Hazard Assessment**

<https://oehha.ca.gov/calenviroscreen>

A series of four datasets and reports, published 2013, 2014, 2018, and 2021, with pollution, health, and socioeconomic measurements for each of California's zip codes or census tracts.

Measurements compiled into a small number of “scores”, including California EnviroScreen (CAES) score.

Motivation: Transportation #1 Emission Source



Motivation: Emissions Exceptions for Freight Fleet

Your vehicle does not need a smog inspection if your:

- Gasoline-powered vehicle is a 1975 year model or older (This includes motorcycles and trailers.)
- Diesel-powered vehicle is a 1997 and older year model OR with a Gross Vehicle Weight of more than 14,000 pounds.
- Powered by natural gas and weighs more than 14,000 pounds.
- An electric vehicle.
- Gasoline-powered and less than eight model-years old.

SOURCE: CA DMV

dmv.ca.gov/portal/vehicle-registration/smog-inspections/



sanrafael-ca.americanlisted.com/trailers-mobile-homes/285001995-freightliner-classic-xl_22080645.html

*1995 Freightliner for Sale in San Rafael, Marin County,
San Francisco Bay Area, CA*

The CalEnviroScreen Model

EnviroScreen-specific “scores” are derived from meticulous measurements.

- Pollution Burden Score
 - Exposures
 - Ozone concentrations
 - Particulate matter emissions and concentrations (diesel, PM2.5)
 - Drinking water contaminants, lead risk
 - Toxic releases from facilities, pesticide use
 - Traffic density
 - Environmental Effects
 - Solid waste, hazardous waste sites
 - Groundwater threats and impaired water body count

The CalEnviroScreen Model

More EnviroScreen-specific “scores”

Impact weights are determined by the CalEPA.

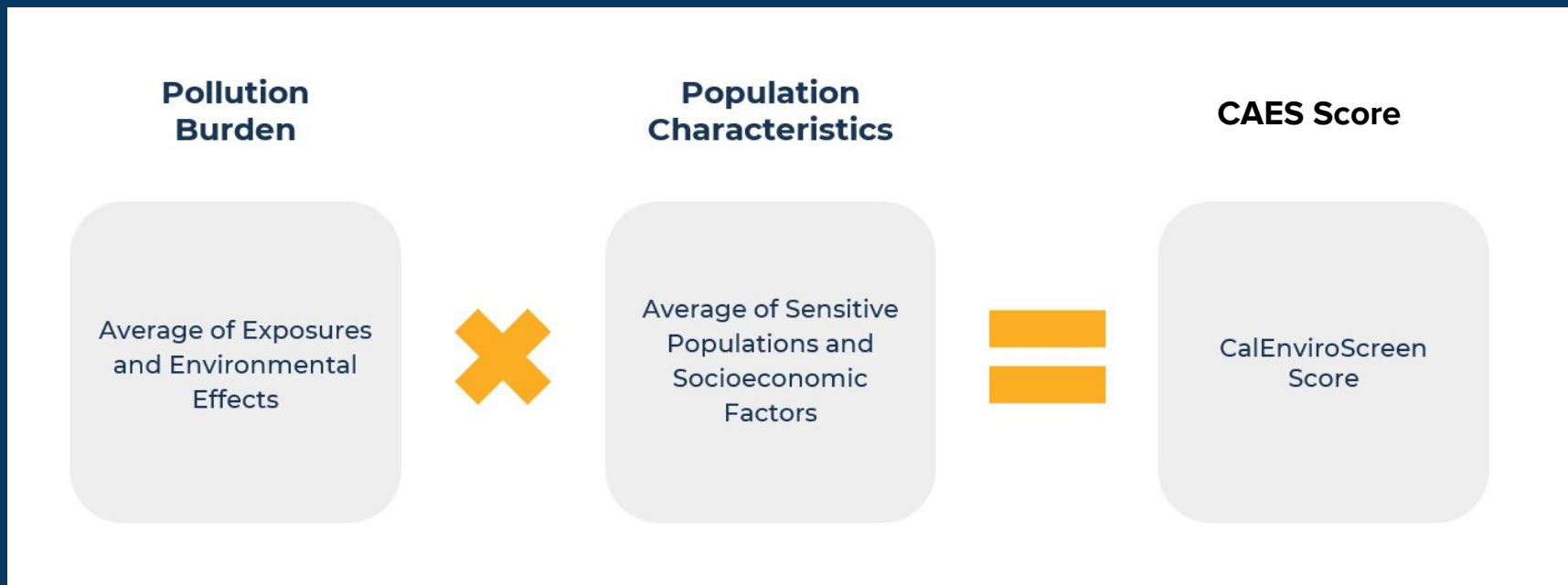
- Population characteristics
 - Sensitive population
 - Asthma
 - Cardiovascular disease
 - Low birth weight infants
 - Socioeconomic factors
 - Educational attainment
 - Housing burdened low income households
 - Linguistic isolation
 - Poverty
 - Unemployment

Targets

Target columns for models were counts of ER visits within a California zip code.

- **Asthma:** Age adjusted rate of ER visits per 10k population with Asthma
- **Low birth weight:** % of “live, singleton births” (<2.5 kg)
- **Cardiovascular disease:** ER visits for AMI per 10,000

The CalEnviroScreen Model

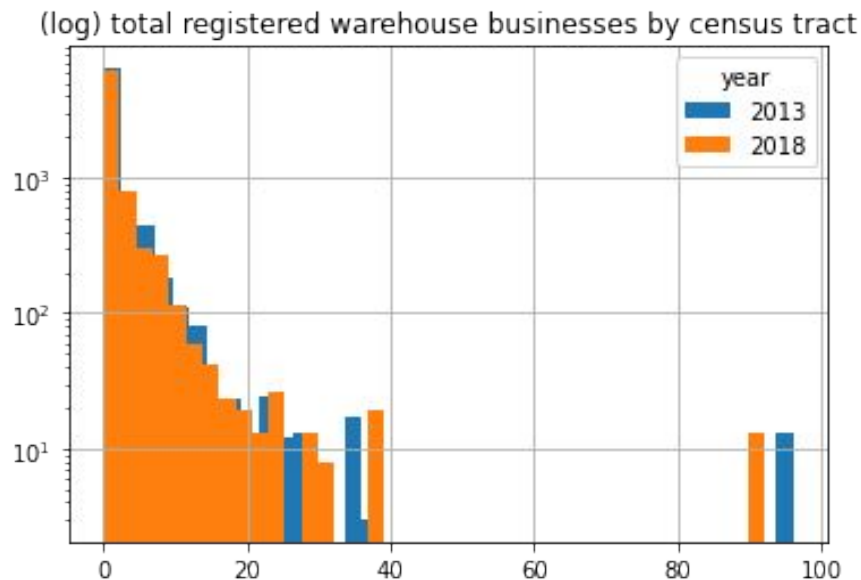


Data Source: US Census Bureau

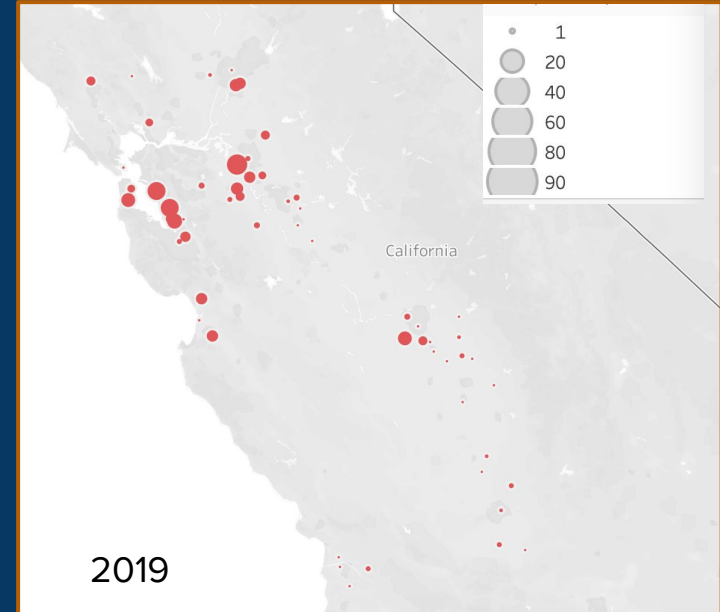
County Business Patterns



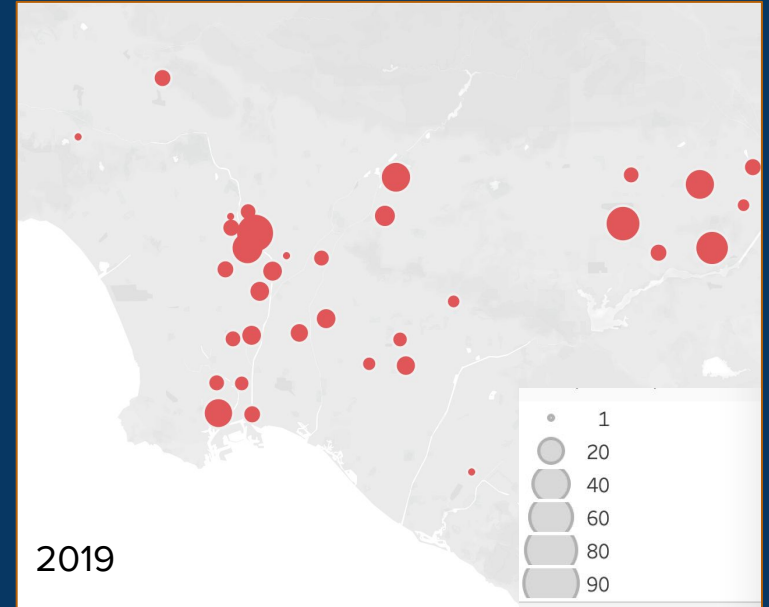
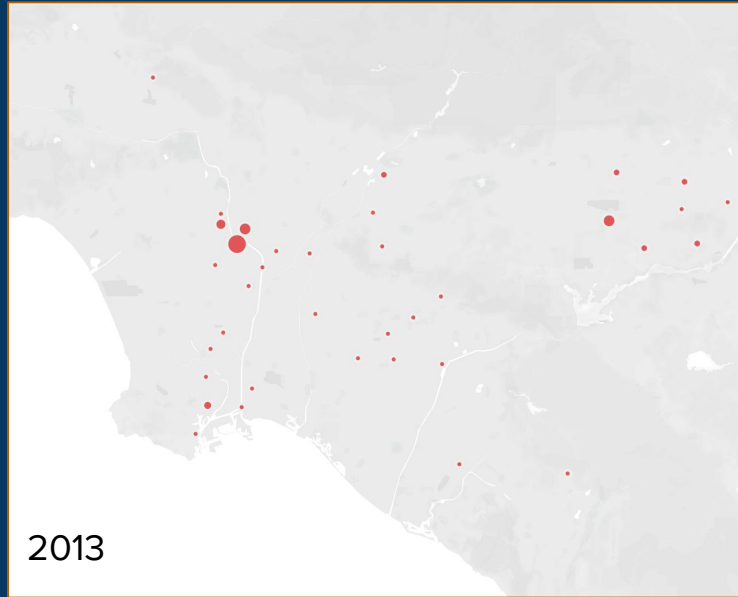
- Annual economic data by industry
 - Number of registered businesses
 - Employee count ranges
 - Payroll
- Warehouse data divided by zip code and North American Industrial Code System class (NAICS)
- Four NAICS warehouse types:
 - “General”
 - Farming
 - Refrigerated
 - “Other”
- The “heavy hitters” in the southeast:
 - Orlando (San Bernardino County)
 - “Inland empire” (Riverside County)



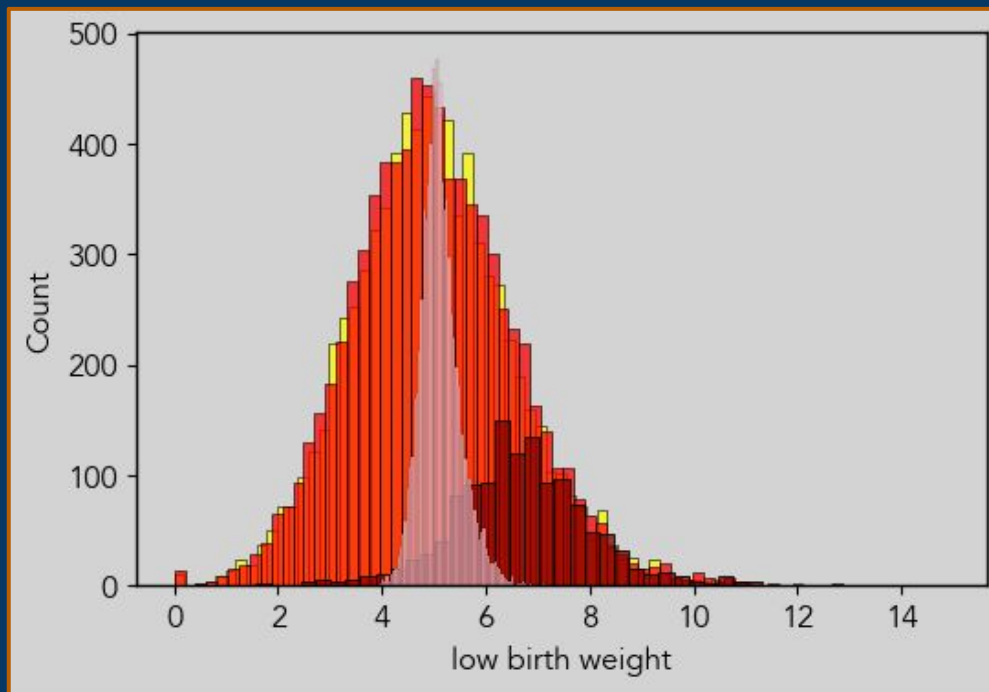
Warehouse Business Counts, 2013 - 2019



Warehouse Business Counts, 2013 - 2019

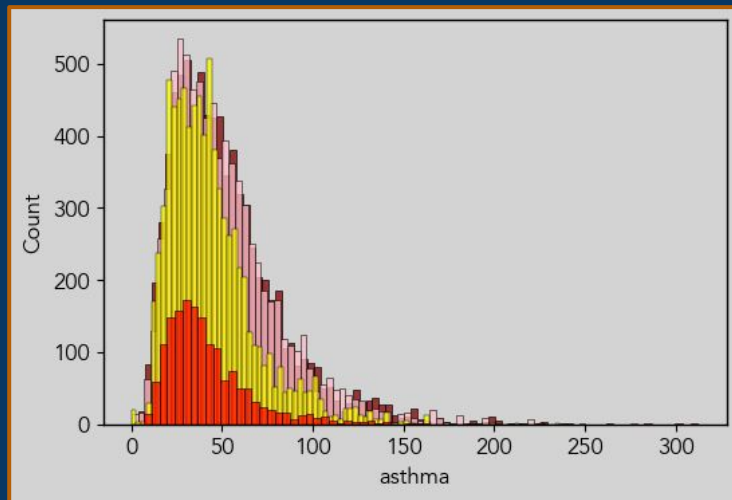


EDA: Low-Birth Weight

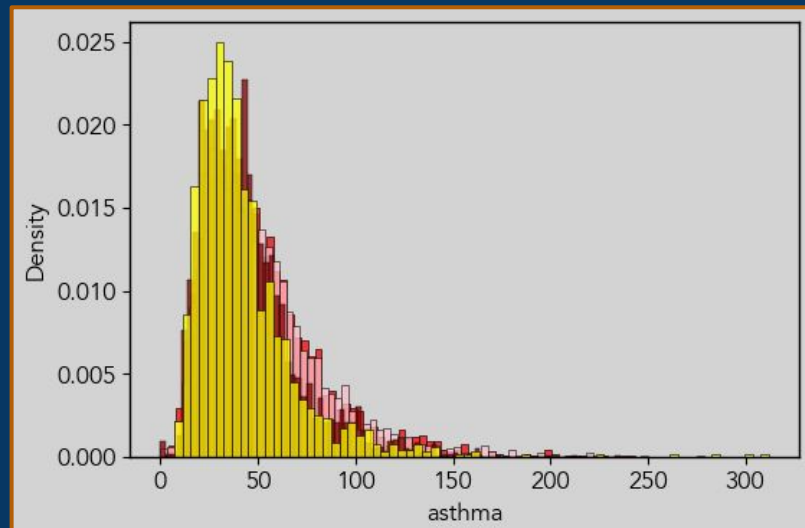


- % of newborns < 2500 g (5.5lb) in hospital for given ZIP
- all health metrics from CA reporting agency.
- Pink Peak: reporting used spatialized metrics vs. strict %

EDA: Asthma

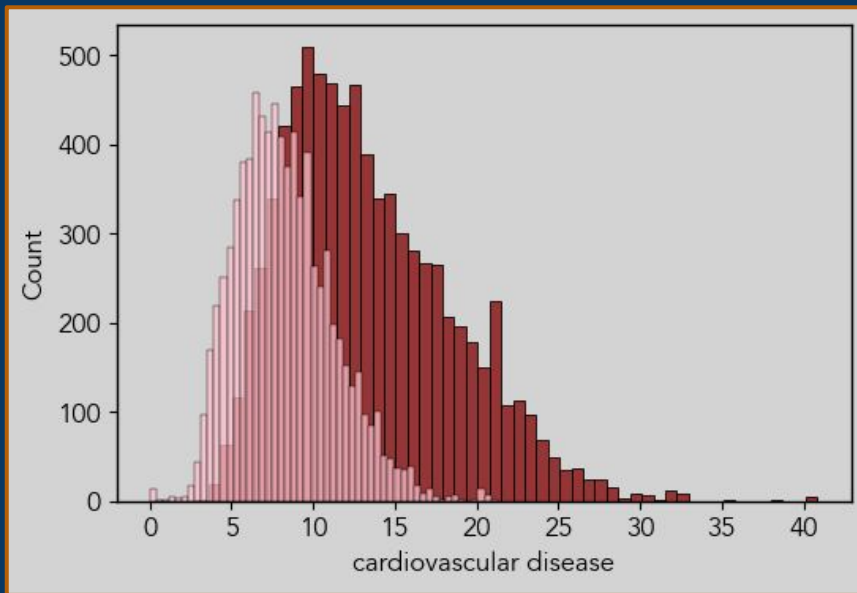


counts: more data in ES 2, 3, 4

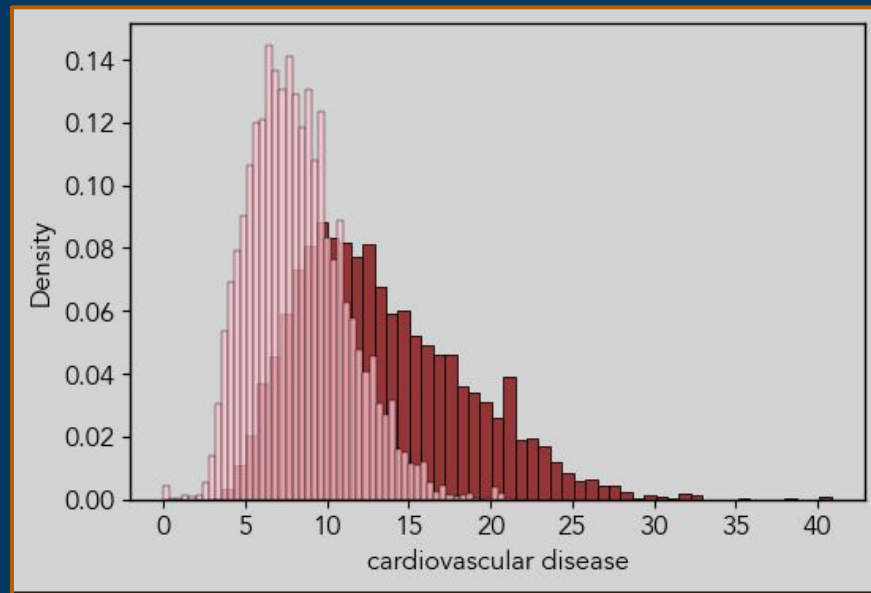


density: distribution relatively the same

EDA: Cardiovascular

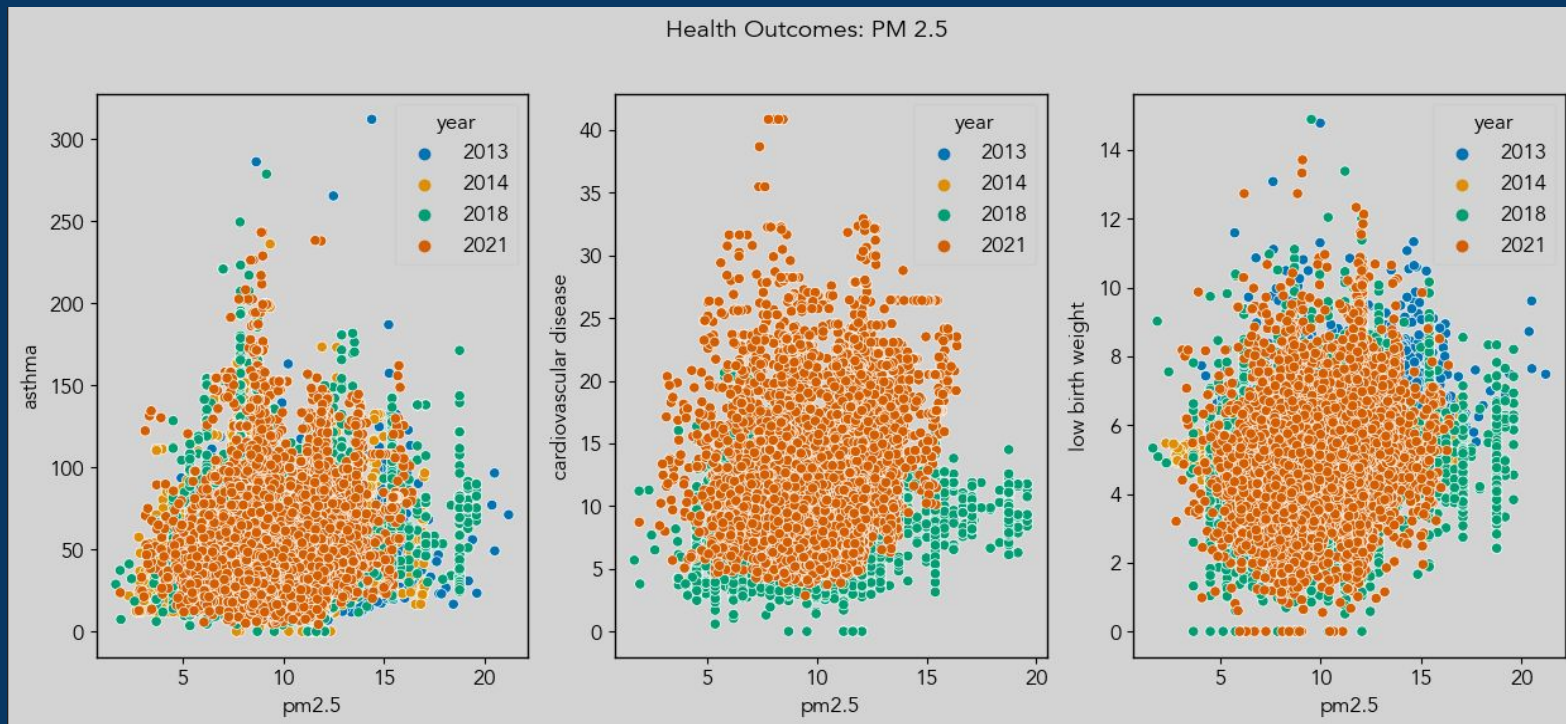


- data in 2018, 2021 reports only

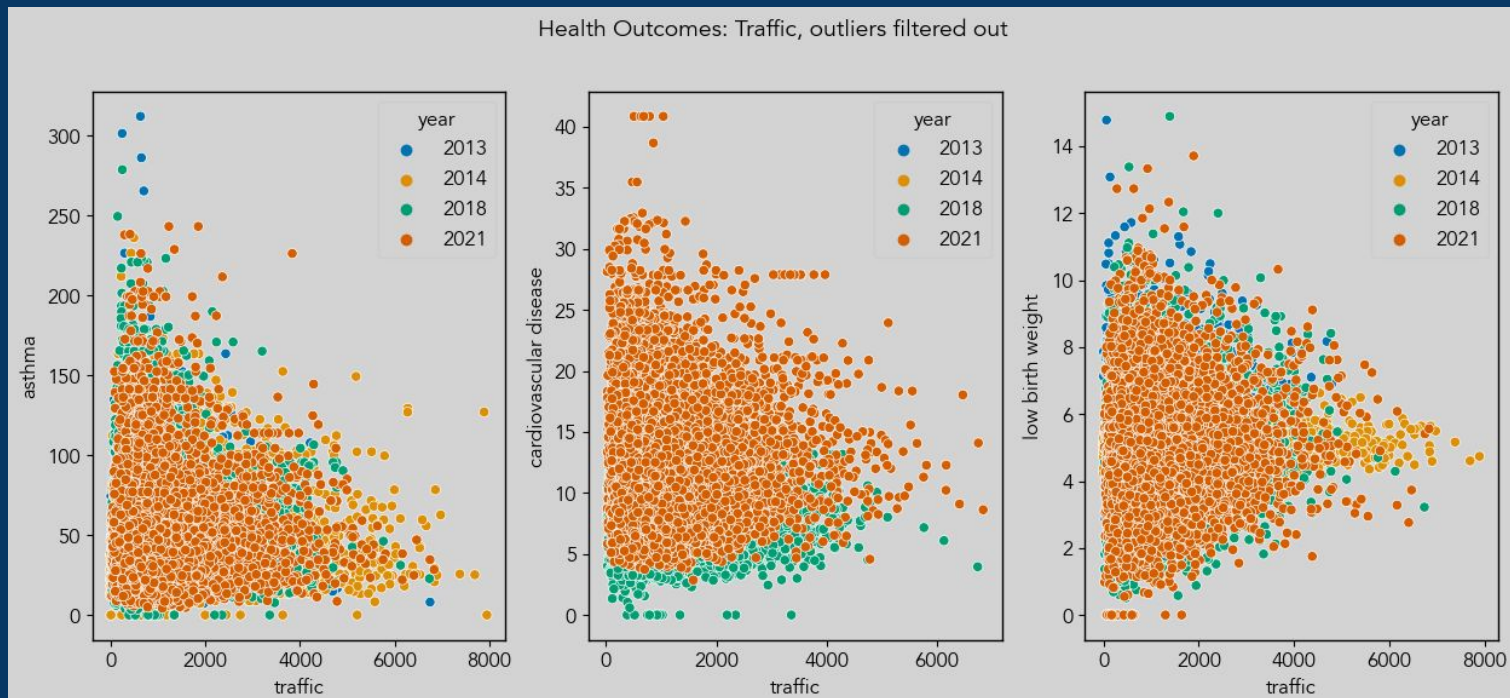


- increase over 3-year period

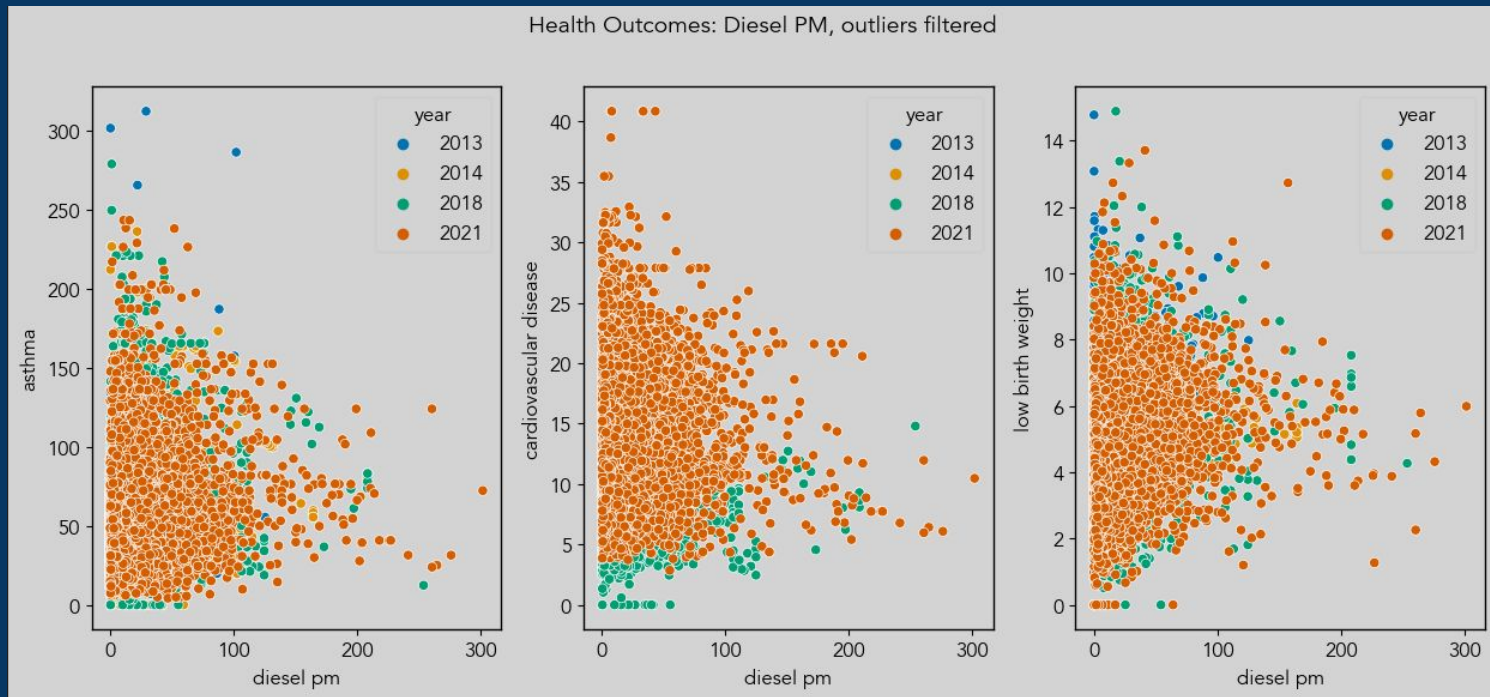
EDA: Health Outcomes, PM 2.5



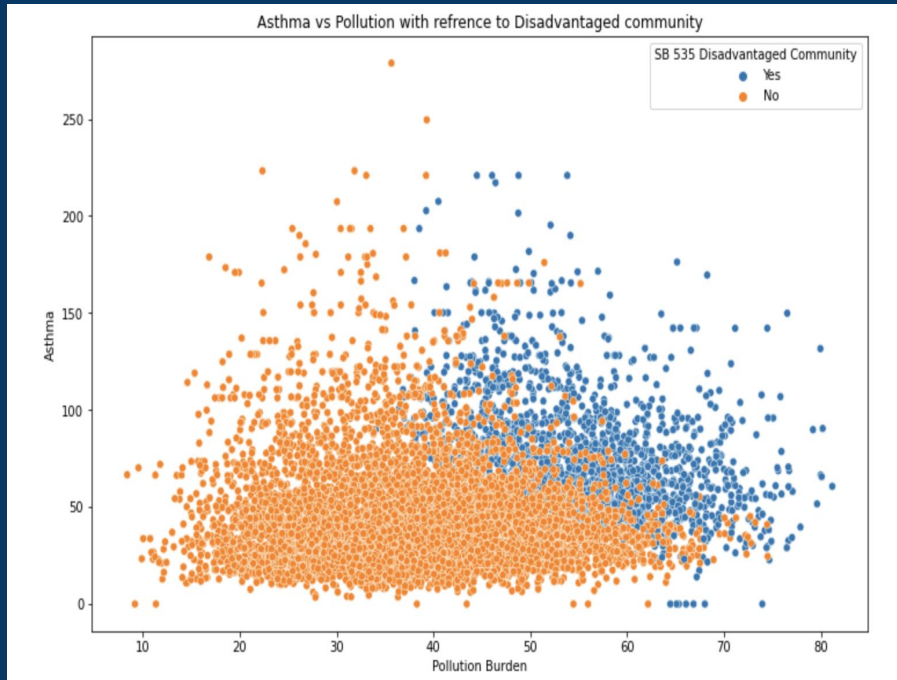
EDA: Health Traffic Volume



EDA: Health Outcomes, Diesel PM



Asthma vs Pollution with reference to Disadvantaged Community



FEATURES DROPPED

- Percentile columns.
- Location columns.
- Features that are functions of other metrics in dataset.
- No additional warehouse data.

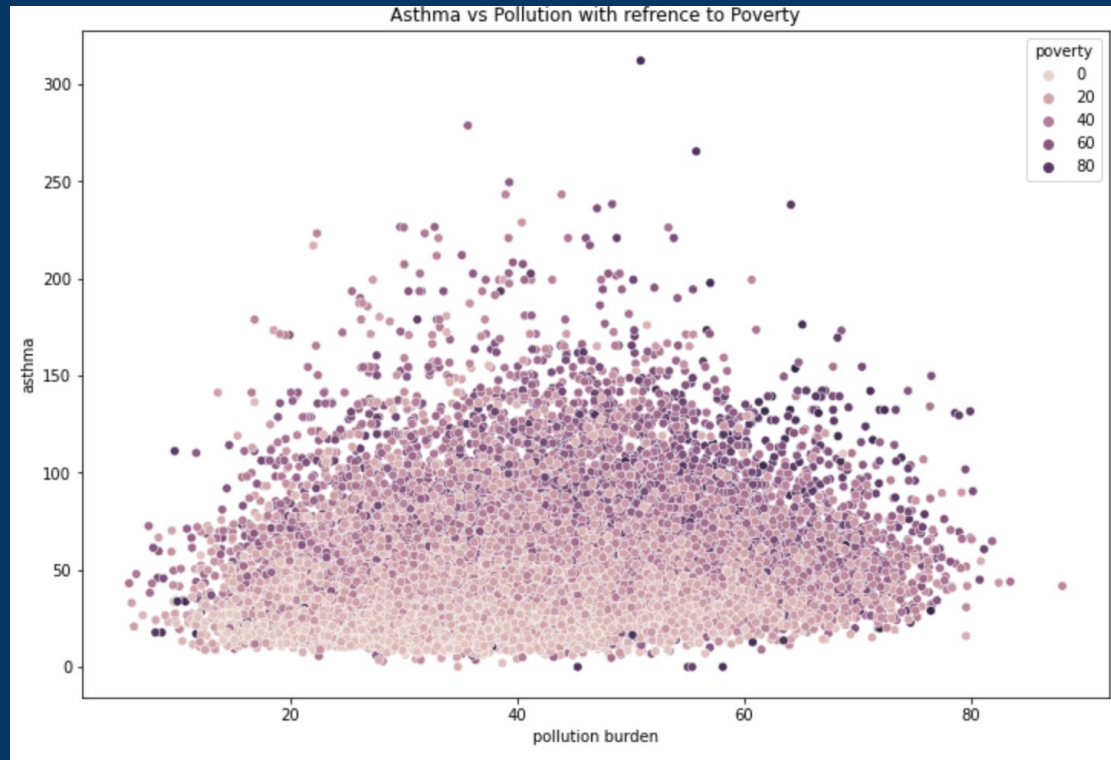
SURPRISING FEATURES

- Asthma/Pollution show to be connected to disadvantageous communities
- Socioeconomic factors correlate

MANIPULATION

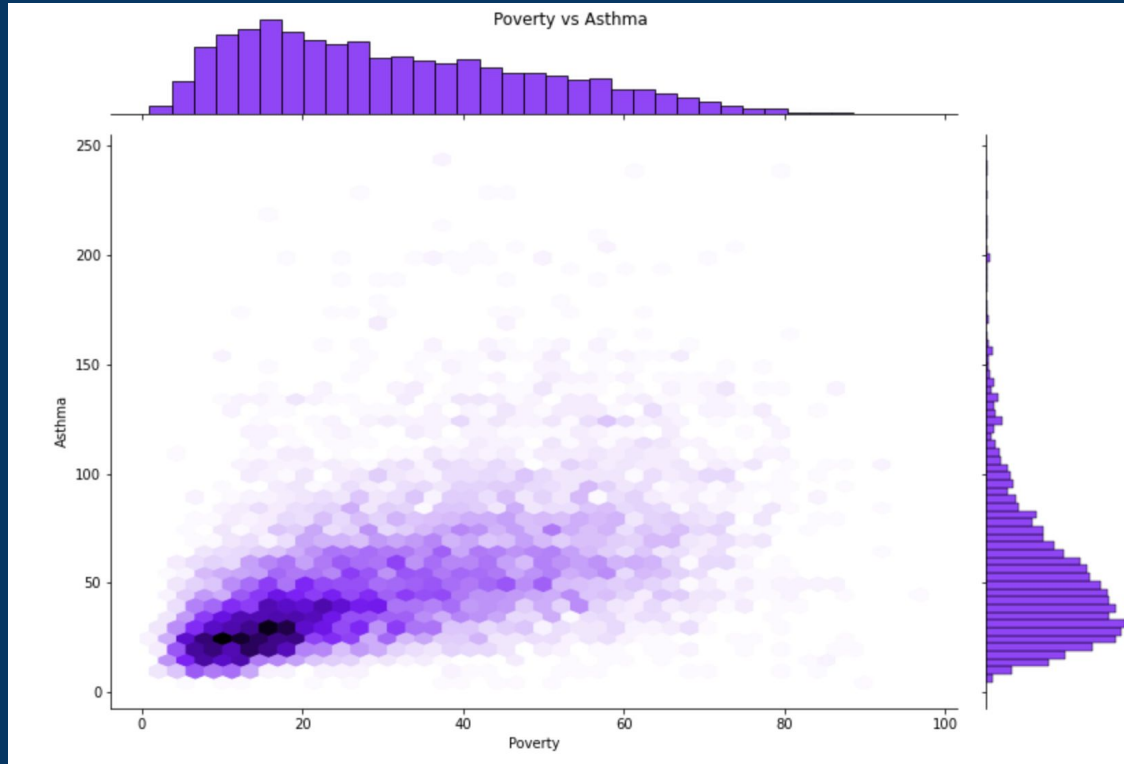
- Deal with missing values Fill with median
- Only 2018 dataset

Asthma vs Pollution with reference to Poverty



- With all 4 Enviroscreen datasets
- Many columns needed to be filled
- Increasing trend left to right

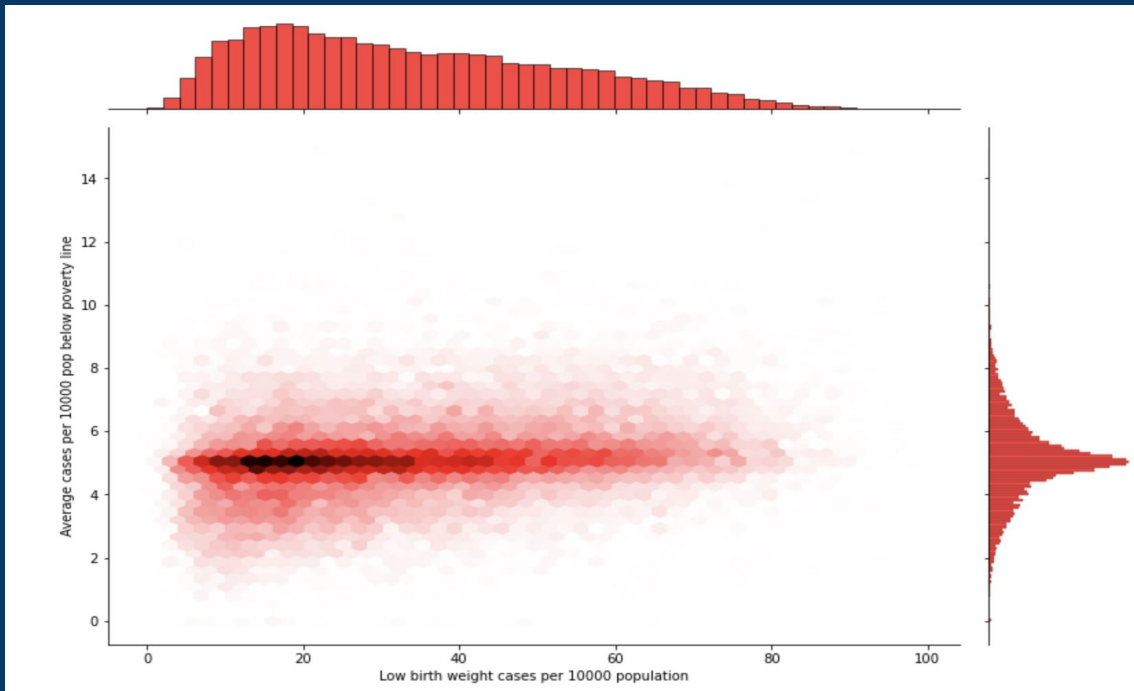
Joint plot showing Poverty vs asthma



Notable features

- Increasing trend
- Shows SB 535 is correlated with Poverty

Poverty & Low Birth Weight



Notable features

- Suspiciously straight line
- How could poverty be so correlated with everything before, but not LBW
- ER visits

XGBoost model

XGboost, scaled, & GS CV for Asthma target

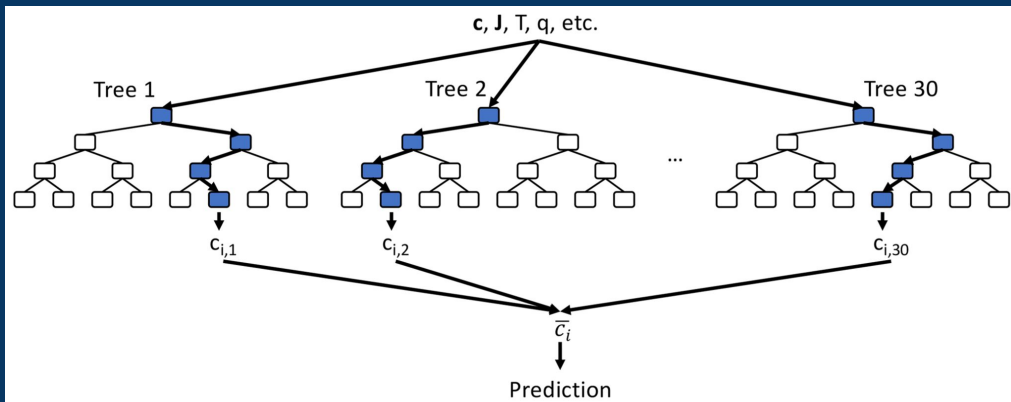
Colsample_bytree:0.4 Max_depth:8
Gamma: 0.1 Min_child_weight:7
Learning_rate:0.15 nthread:4

type	evaluation metric	Train Accuracy	Test Accuracy	RMSE score	MAE test score
gradient boosting supervised regression	R2, RMSE, & MAE	0.9139	0.7853	13.6915	9.3296

FEATURES

- Total population
- Ozone
- pm2.5
- Diesel pm
- Pesticides
- Traffic
- Cleanup sites
- Groundwater threats
- Haz. waste
- Imp. water bodies
- Solid waste
- Pollution burden
- Education
- Linguistic isolation
- Poverty
- Pop. char.
- Drinking water
- Tox. release
- Unemployment
- Ces_per
- Housing burden
- Est gen
- Est cold
- Est farm
- Est other

Model 2: Random Forest Regression



FEATURES

- Total population
- Ozone
- pm2.5
- Diesel pm
- Pesticides
- Traffic
- Cleanup sites
- Groundwater threats
- Haz. waste
- Imp. water bodies
- Solid waste
- Pollution burden
- Education
- Linguistic isolation
- Poverty
- Pop. char.
- Drinking water
- Tox. release
- Unemployment
- Ces_per
- Housing burden
- Est gen
- Est cold
- Est farm
- Est other

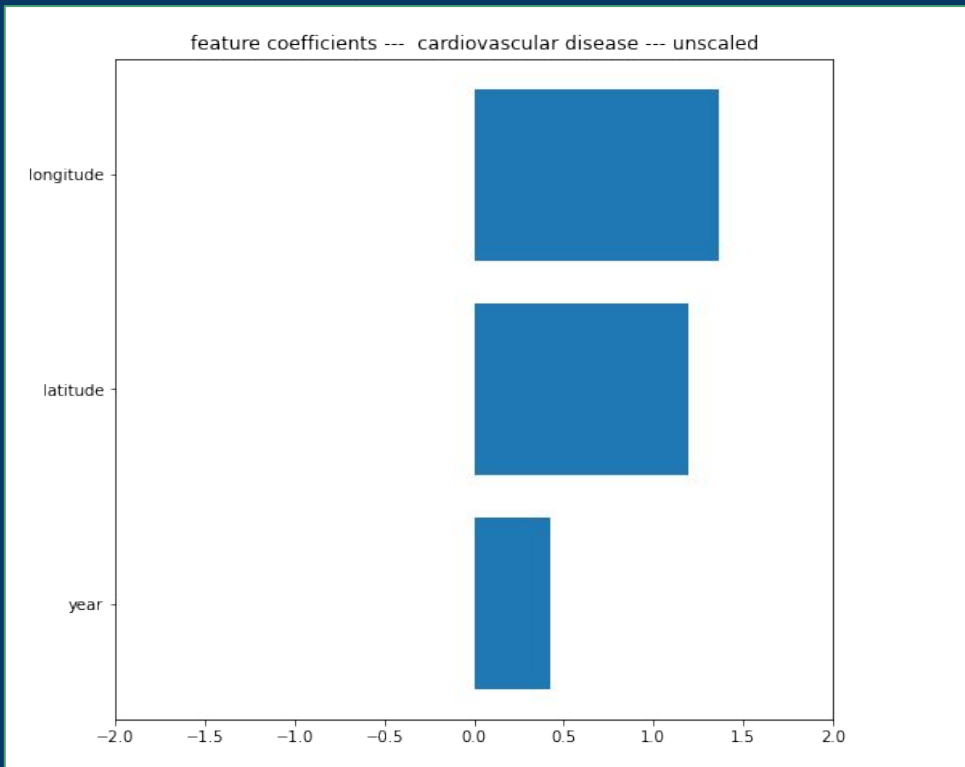
Params:

- Bootstrap = False
- Max_feat = sqrt
- n_est = 30
- min_sam_split = 4

meta estimator regression	R2, RMSE, & MAE	0.9634	0.7503	14.7307	9.9952

Linear models

Linear model: health targets with year and location



Linear models for each health outcome fit to year, latitude and longitude.

Cardiovascular disease ER visits

FEATURES

Year

Latitude

longitude

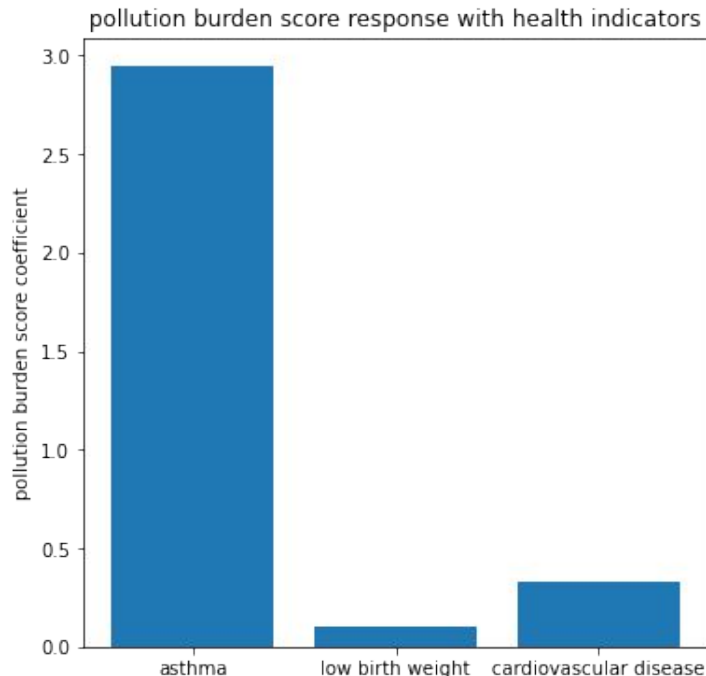
FINAL METRICS: R^2

Asthma 0.054

Low birth weight 0.023

Cardiovascular disease 0.170

Linear model: coefficients for CAES scores



FINAL METRICS:

Asthma

0.028

Low birth weight

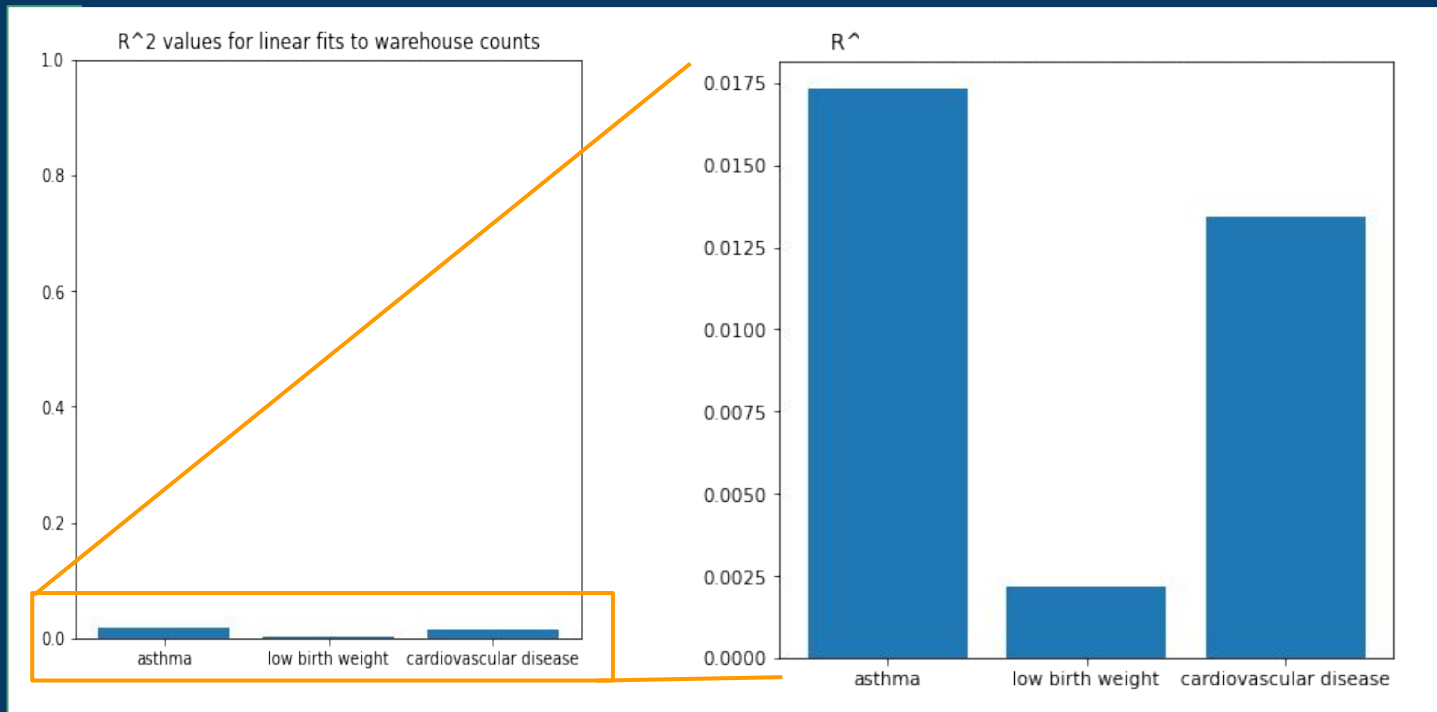
0.015

Cardiovascular
disease

0.017

Evaluating the impact of CAES score: Pollution burden

Linear model: Metrics for warehouse counts



FINAL METRICS:

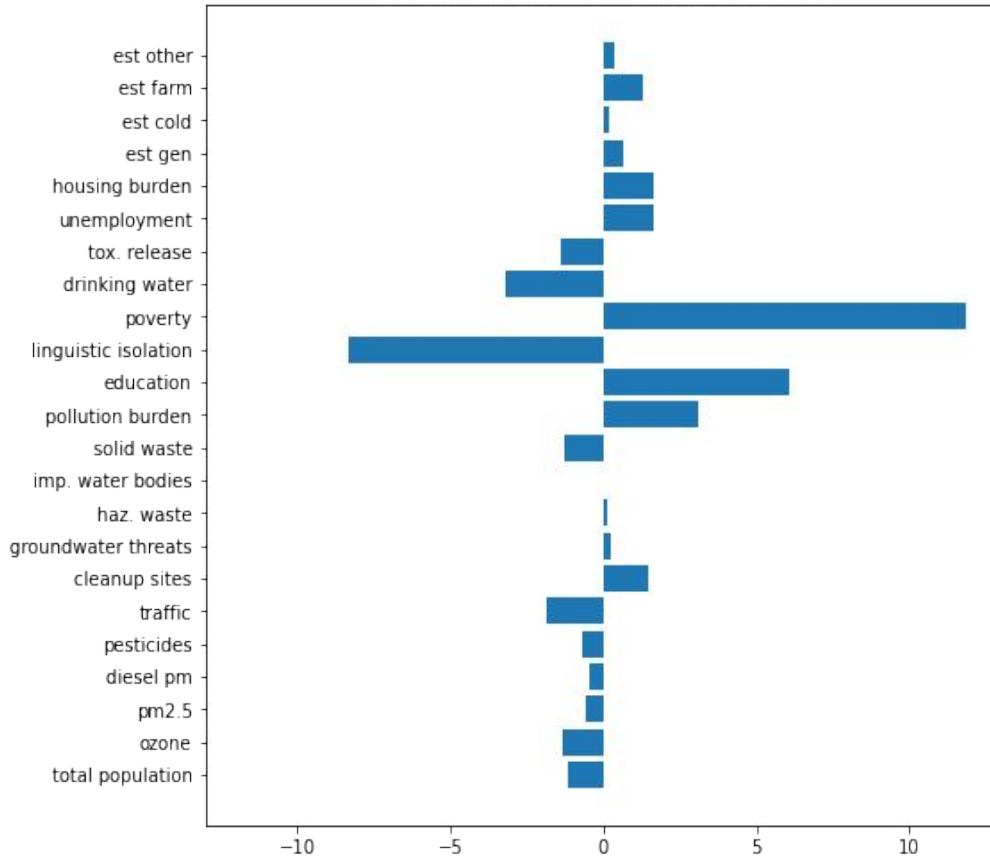
Asthma
0.017

Low birth weight
0.002

Cardiovascular
disease
0.013

Evaluating the impact of warehouse business types

feature coefficients --- asthma --- scaled



Linear model: “Selected columns” with Asthma

FEATURES

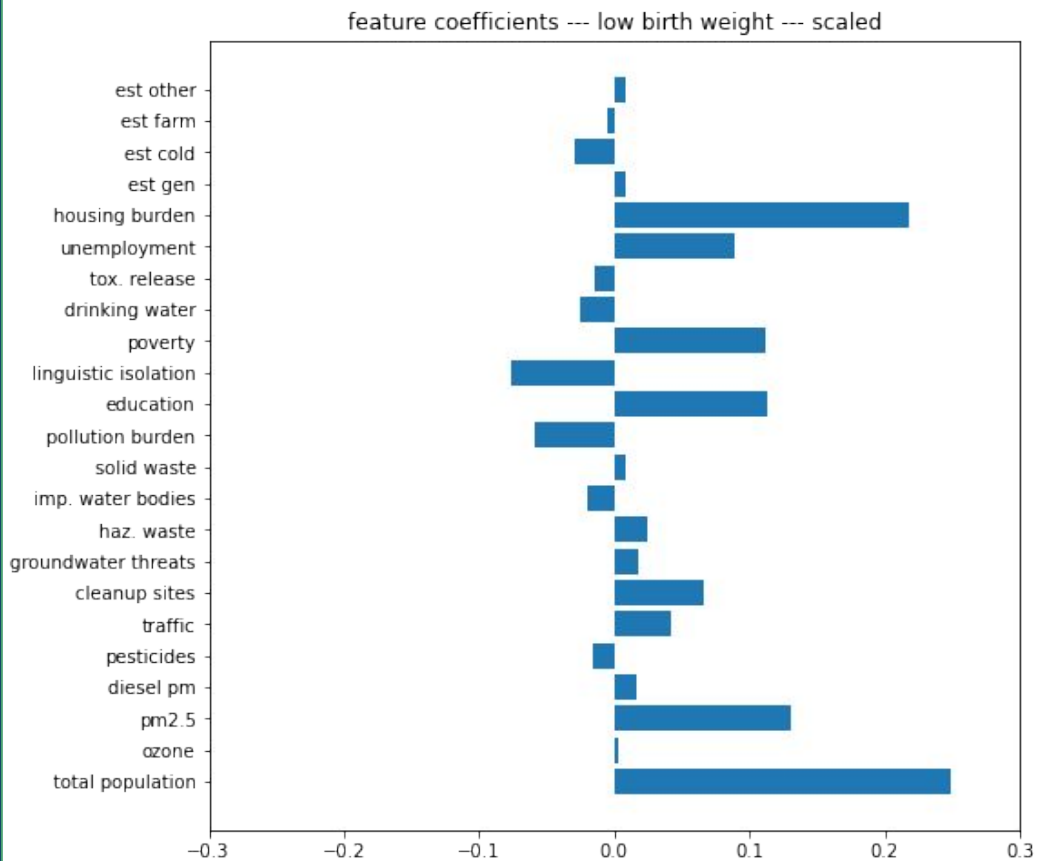
'total population',	'imp. water bodies',	'unemployment',
'ozone',	'solid waste',	'housing burden',
'pm2.5',	'pollution burden',	'est gen',
'diesel pm',	'education',	'est cold',
'pesticides',	'linguistic isolation',	'est farm',
'traffic',	'poverty',	'est other'
'cleanup sites',	'drinking water',	
'groundwater threats',	'tox. release',	
'haz. waste',		

FINAL METRICS: R²

Asthma - 0.29

Low birth weight - 0.14

Cardiovascular disease - 0.23



Linear model: “Selected columns” with Low birth weight

FEATURES

'total population',	'imp. water bodies',	'unemployment',
'ozone',	'solid waste',	'housing burden',
'pm2.5',	'pollution burden',	'est gen',
'diesel pm',	'education',	'est cold',
'pesticides',	'linguistic isolation',	'est farm',
'traffic',	'poverty',	'est other'
'cleanup sites',	'drinking water',	
'groundwater threats',	'tox. release',	
'haz. waste',		

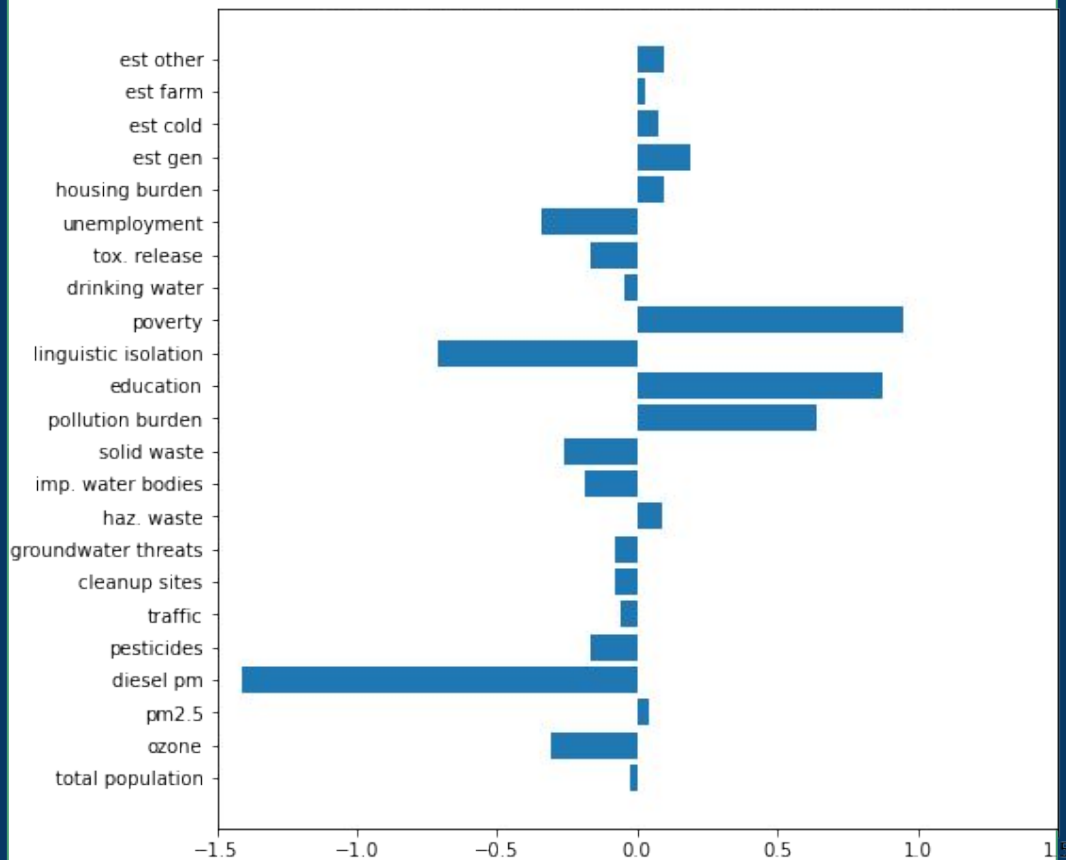
FINAL METRICS: R^2

Asthma - 0.29

Low birth weight - 0.14

Cardiovascular disease - 0.23

feature coefficients --- cardiovascular disease --- scaled



Linear model: “Selected columns” with Cardiovascular disease

FEATURES

'total population',	'imp. water bodies',	'unemployment',
'ozone',	'solid waste',	'housing burden',
'pm2.5',	'pollution burden',	'est gen',
'diesel pm',	'education',	'est cold',
'pesticides',	'linguistic isolation',	'est farm',
'traffic',	'poverty',	'est other'
'cleanup sites',	'drinking water',	
'groundwater threats',	'tox. release',	
'haz. waste',		

FINAL METRICS: R²

Asthma - 0.29

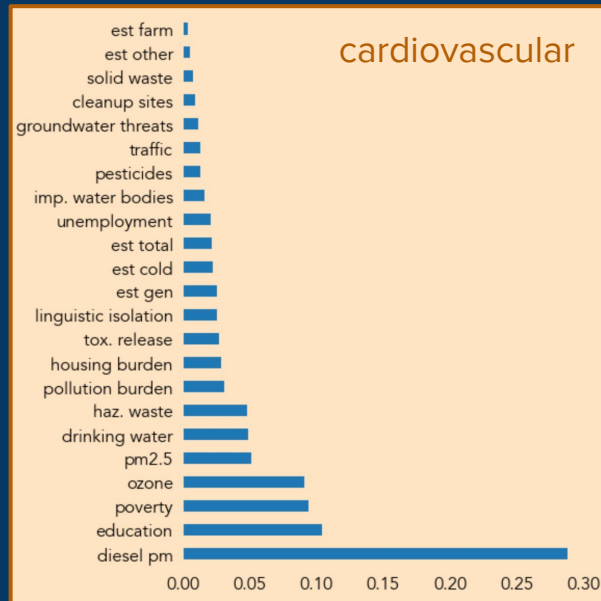
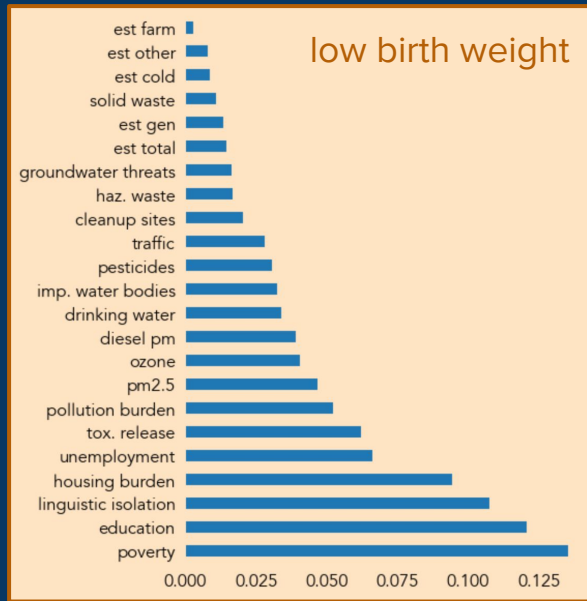
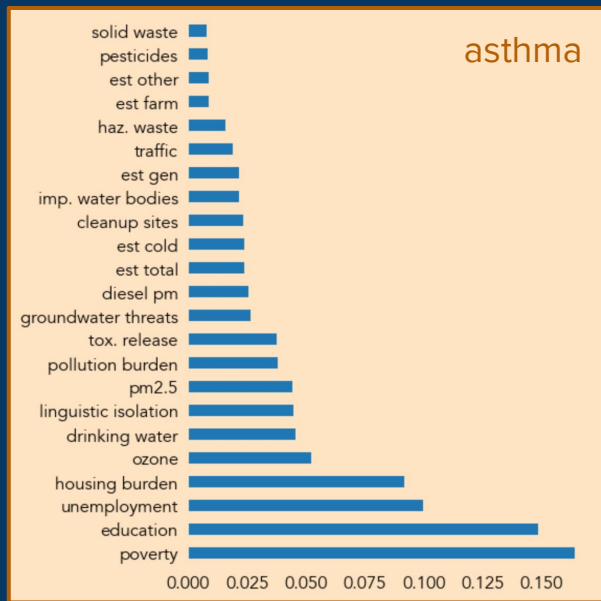
Low birth weight - 0.14

Cardiovascular disease - 0.23

Random Forest Regression: Feature Importances

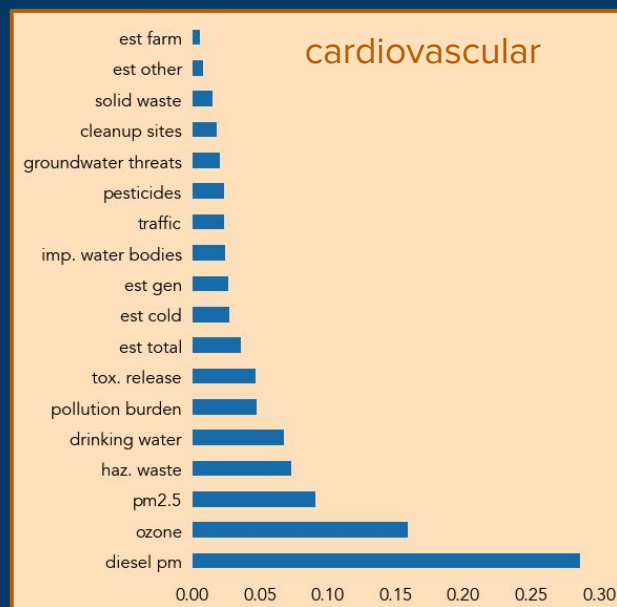
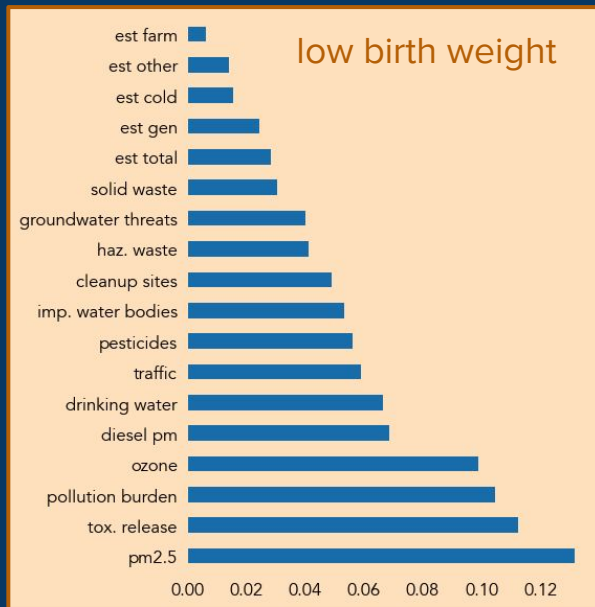
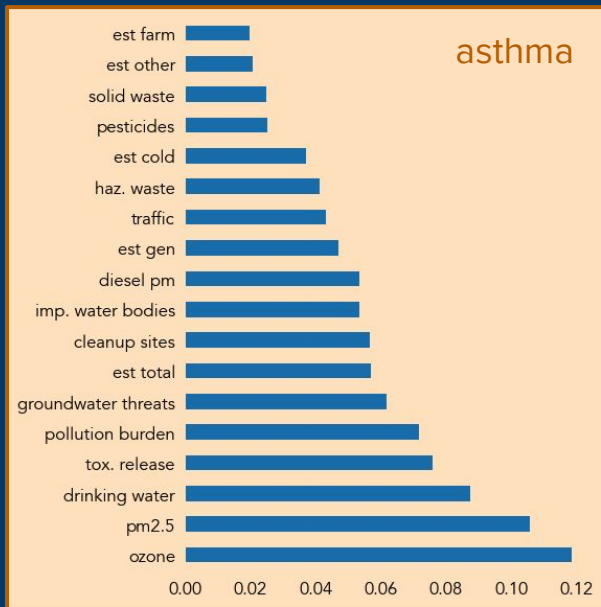
```
'max_depth' : [2, 5, 10, 20],  
'max_features' : [1]
```

```
'max_leaf_nodes' : [15, 30, 50]  
'n_estimators' : [100, 200]
```



RFR: Feature Importances, Non-Socio-Economic

```
'max_depth' : [2, 5, 10, 20], 'max_leaf_nodes' : [15, 30]
'max_features' : [1] 'n_estimators' : [100, 200]
```



We saw saw no meaningful relationship to create robust models for health outcomes using our warehouse-aggregated data.



- CalEnviroScreen scores highly reflect ASTHMA and POLLUTION BURDEN but not hospitalization incidence.
- Socioeconomic factors aggregated in CalEnviroScreen built best predictive models for negative health outcomes.

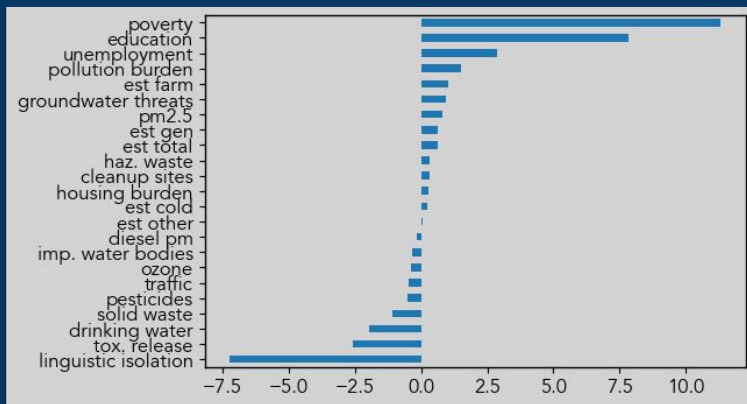
Model: SVR

Epsilon-Support Vector Regression

regularization: L2, C = 1

NUMERIC

CATEGORICAL



Feature Importances: really highlights

FINAL METRICS

DOGS

keep this slide

Twenty

breeds

of

dogs

DOGS

DOGS

20 breeds of dogs

Labrador

German Shepherd

Golden Retriever

Bulldog

Yorkshire Terrier

Boxer

Australian Shepherd

Rottweiler

Dachshund

Siberian Husky

French Bulldog

Great Dane

Irish Cocker

Chihuahua

Shih Tzu

100-year Mountain

American Staffordshire Terrier

Jack Russell

Doberman Pinscher