Health Outcomes in California with Warehouse Industry, Pollution, and Socioeconomic Factors from the California EnviroScreen Data

29 April 2022 | Group Project | GA DSIR 222

Giovanna Guevara, David Tersegno, Marshall Cyrus

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

OEHHA

California Office of Environmental
Health Hazard Assessment

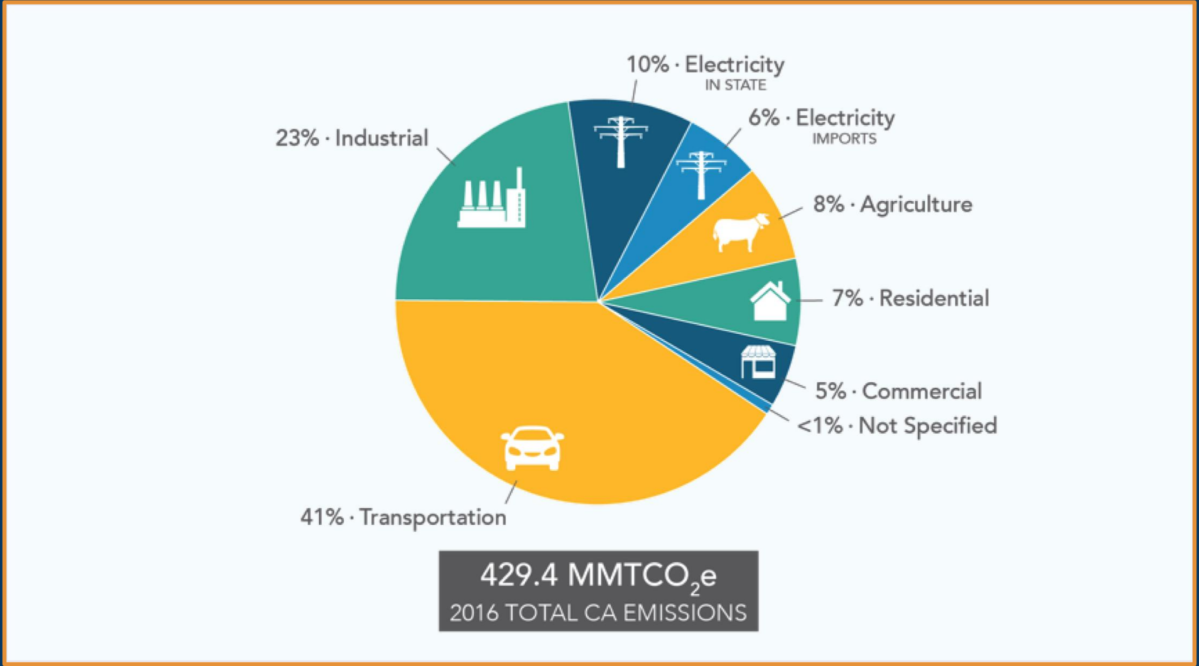
From the **California Office of Environmental Health Hazard Assessment**

https://oehha.ca.gov/calenviroscreen

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

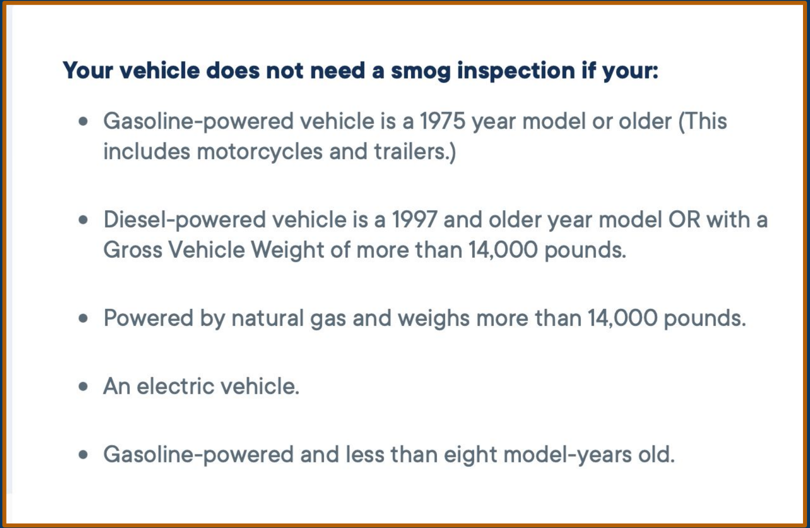
These measurements are compiled into a small number of summary scores, including a broad California EnviroScreen score indicating the regions with the most pressing needs.

Motivation: Transportation #1 Emission Source



SOURCE: CA Air Resources Board, 2018 GHG Emission Inventory, July 2018
https://www.arb.ca.gov/cc/inventory/data/data.htm

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 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, sites
    - Groundwater threats and impaired water body count

The CalEnviroScreen Model

EnviroScreen-specific “scores” are derived from measurements, also included in the dataset. Impact weights are determined by the CalEPA.

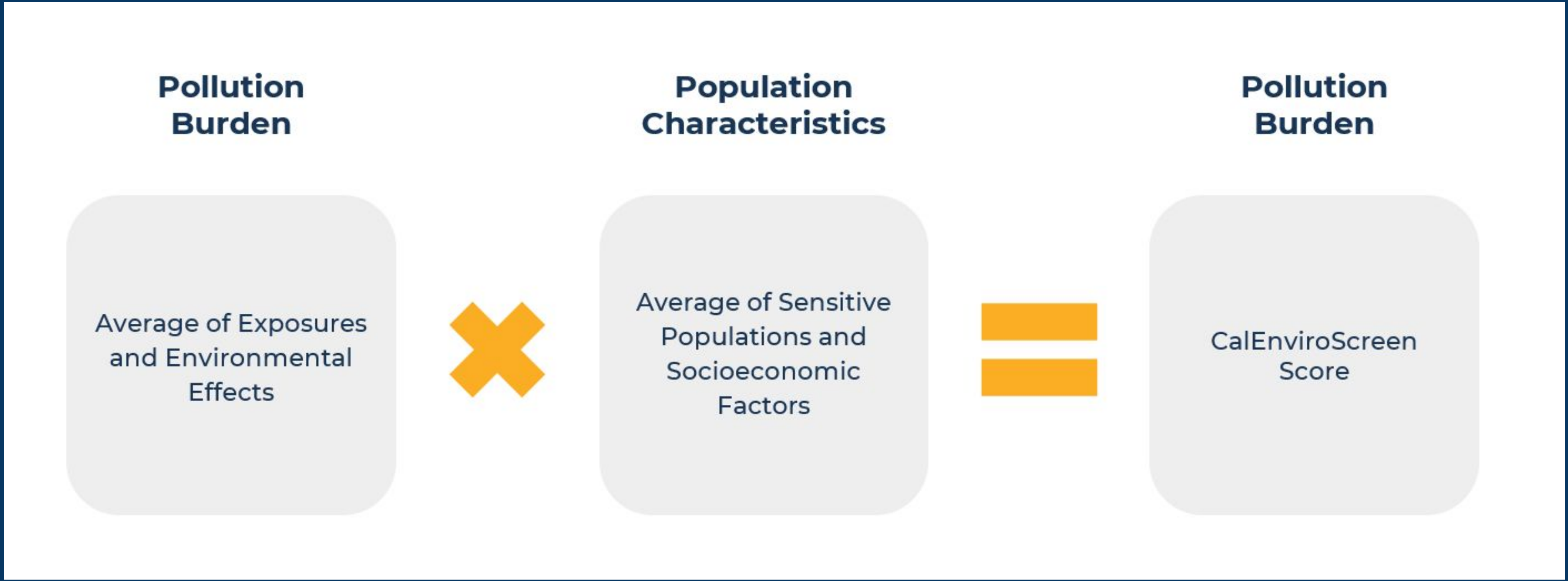
- Population characteristics
    - Sensitive population
        - Asthma
        - Cardiovascular disease
        - Low birth weight infants
    - Socioeconomic factors
        - Educational attainment
        - Housing burdened low income households
        - Linguistic isolation
        - Poverty
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

Targets:

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

- **Asthma:** ER visits per 10k population (double check)
- **Low birth weight:** number of low birth weight (<2000 g) infants born per ?????? (double check
- **Cardiovascular disease:** ER visits for heart attacks per 100 (double check)

The CalEnviroScreen Model



Pollution Burden

Average of Exposures and Environmental Effects

Population Characteristics

Average of Sensitive Populations and Socioeconomic Factors

Pollution Burden

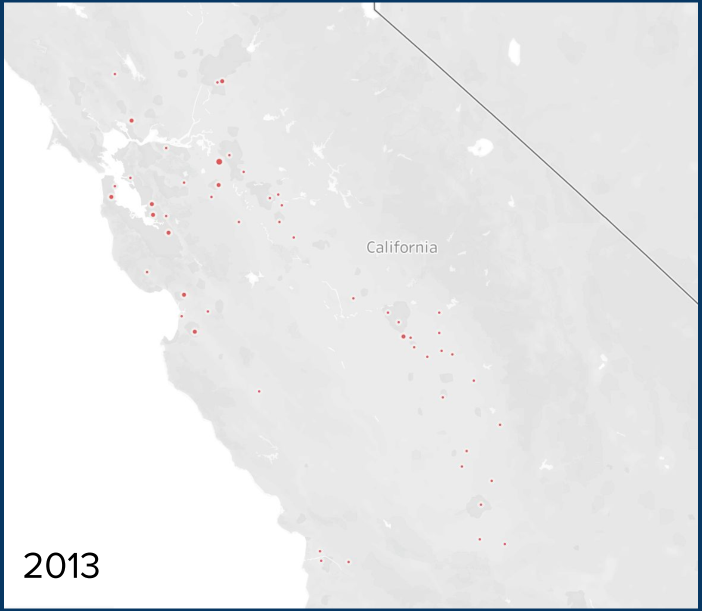
CalEnviroScreen Score

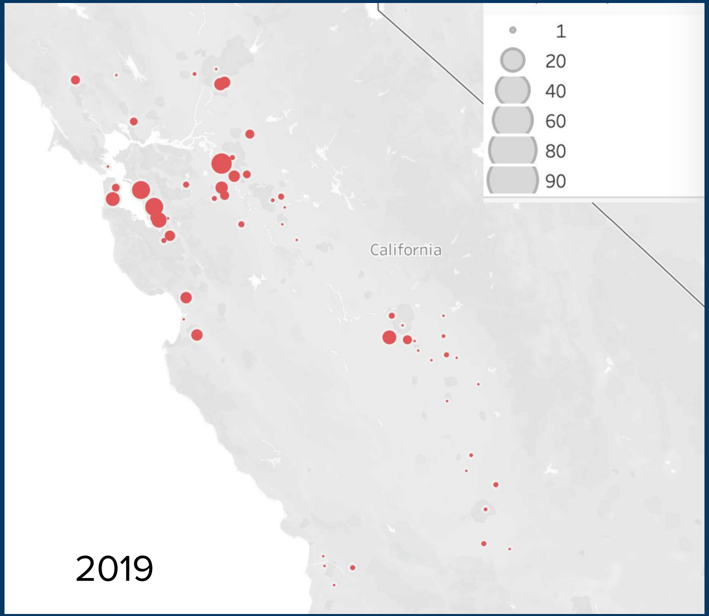
https://oehha.ca.gov/calenviroscreen/scoring-model

Data Source: US Census business counts by NAICS

- Some info here about it.

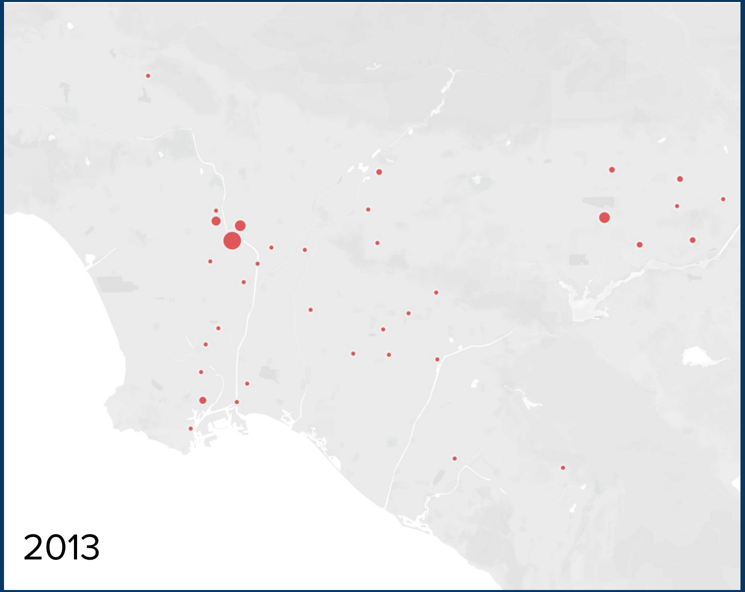
Time by zip — board warehouse business changes.

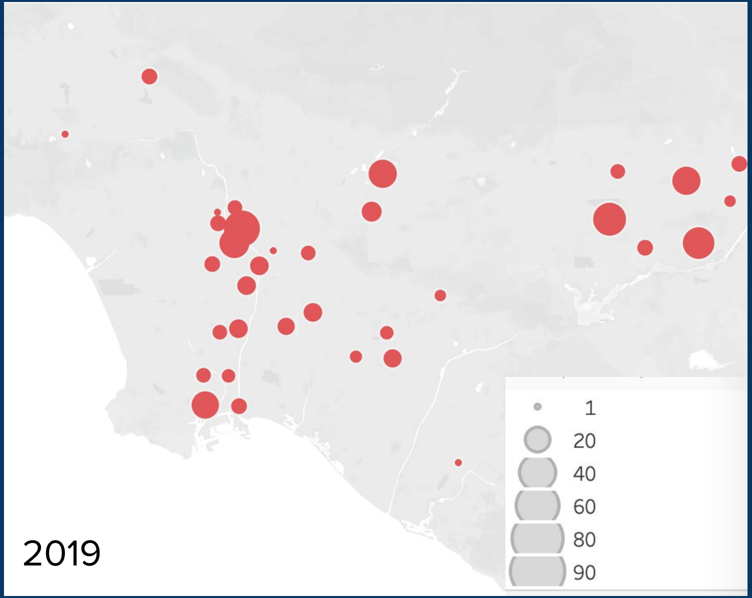




Central California: Estimated Total General Warehouses

Time by zip — board warehouse business changes.





Los Angeles, Inland Empire: Estimated. Total General Warehouses

Time by zip — board warehouse business changes.

Or, just time with california as a whole.

MAP

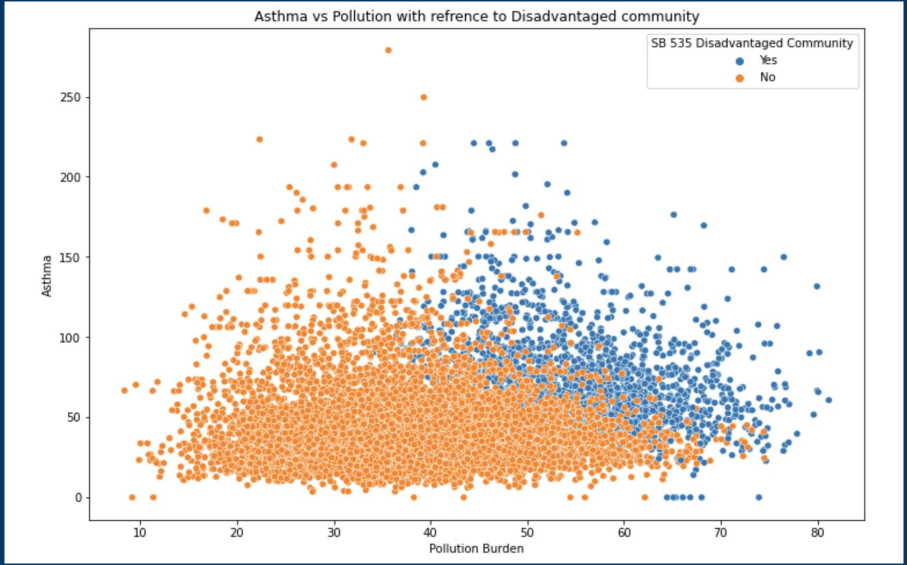
Time by zip — what are the biggest changers? Health or pollution

Only fitting four values for each county: caes 1, 2, 3, 4 years.

Colored map

These highlight regions that may be trouble soon, or rapidly improving. (eda should show bad/good *currently*)

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

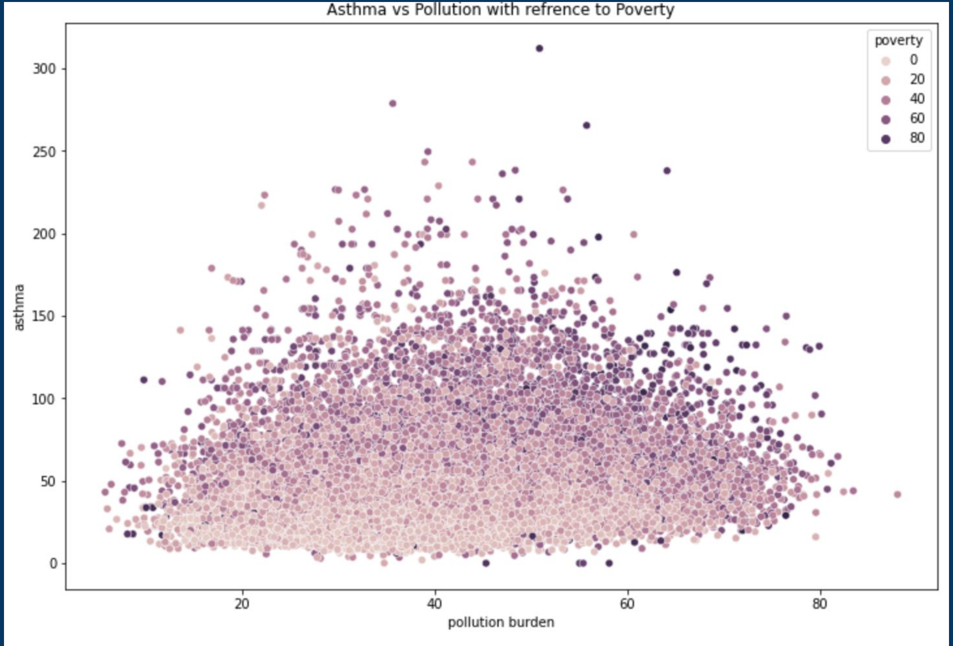
- Asthma and disadvantageous communities show to be connected as more of the blue dots are higher up on the asthma scale. ALso the pollution burden is higher for these aforementioned communities as well.

BLURB

MANIPULATION

- **Deal with missing values Fill with median**
-

EDA with with health, pollution, and Poverty



FEATURES DROPPED

–

–

–

–

–

–

–

–

–

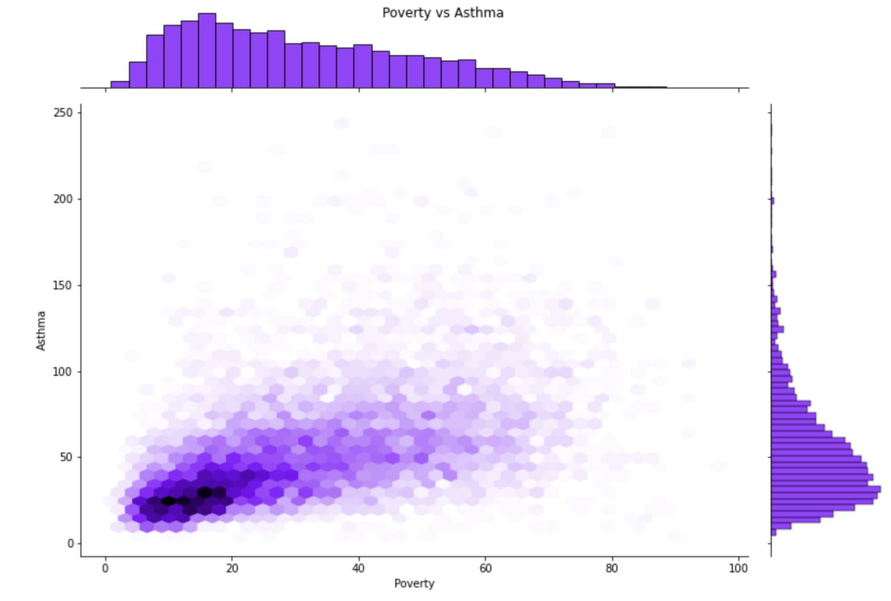
SURPRISING FEATURES

–

BLURB

MANIPULATION

EDA visuals cont. (marshall)



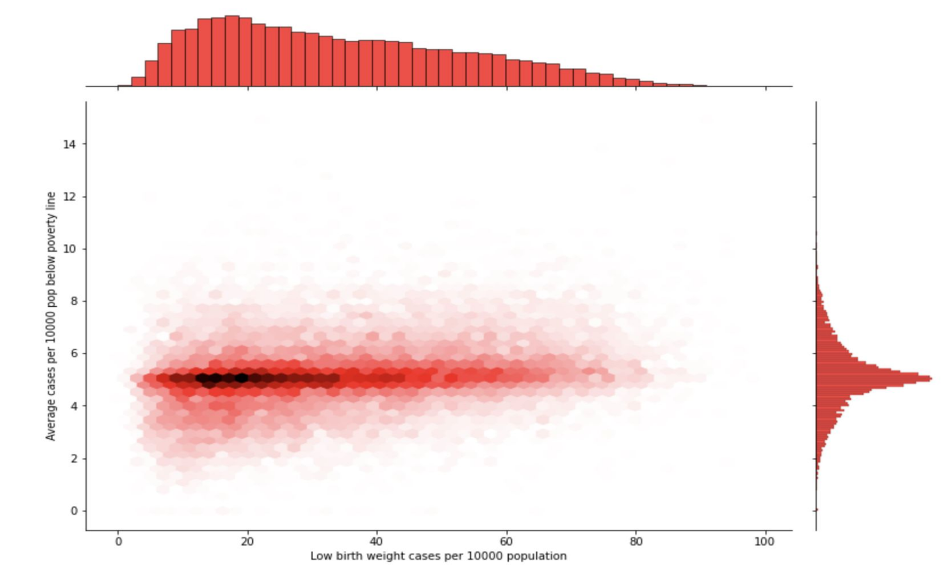
FEATURES DROPPED

SURPRISING FEATURES

BLURB

MANIPULATION

Poverty & Low Birth Weight



FEATURES DROPPED

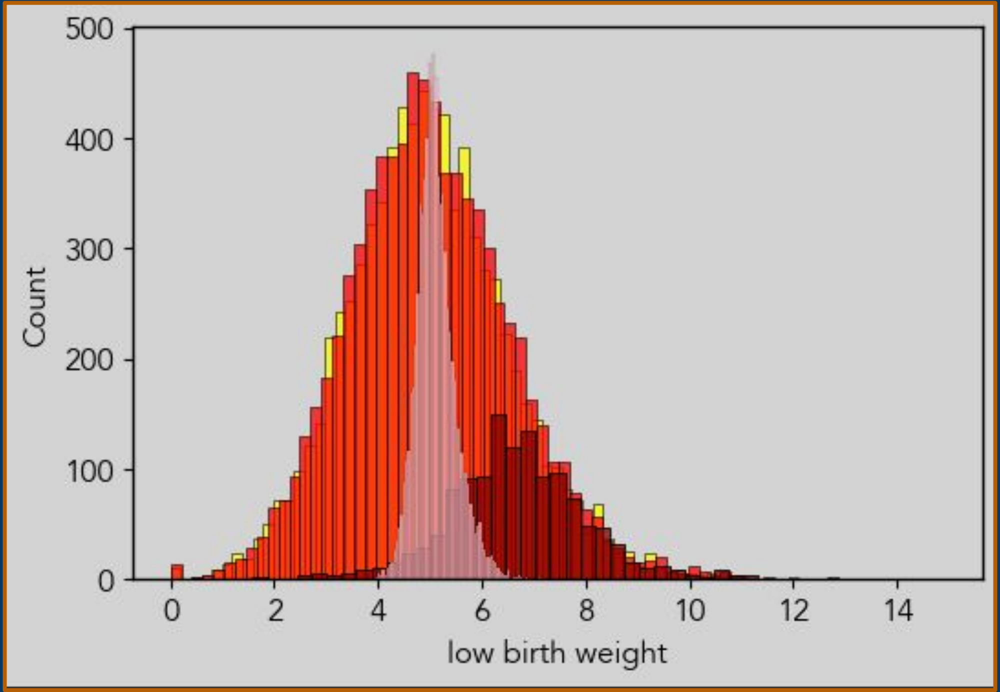
- Suspiciously straight line
- How could poverty not

SURPRISING FEATURES

BLURB

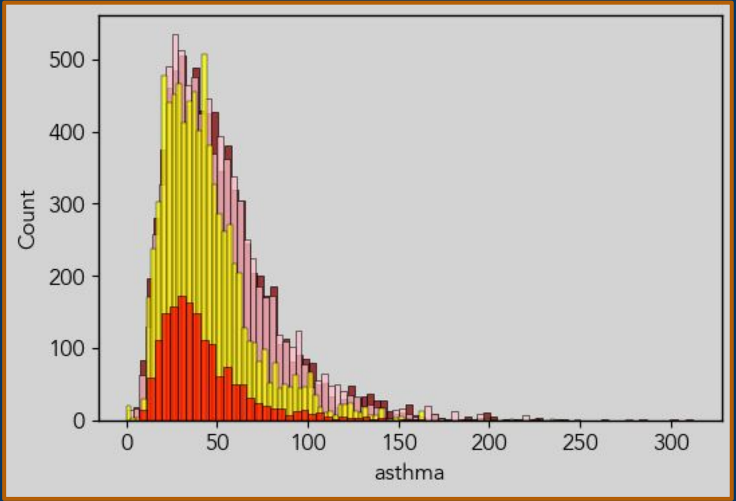
MANIPULATION

EDA: Low-Birth Weight

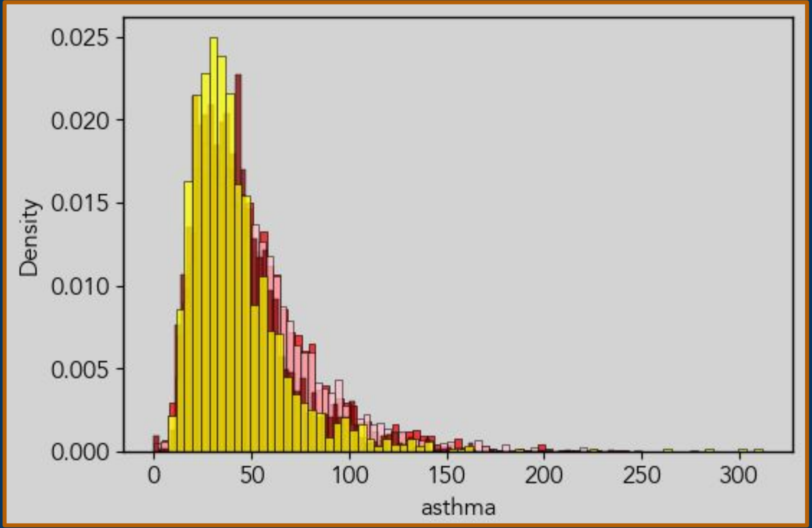


- % of newborns < 2500 g (5.5lb) in hospital for given ZIP
- all health metrics from CA reporting agency.
- Pink Peak: reporting used spatialized metrics vs. strict %

EDA: Asthma

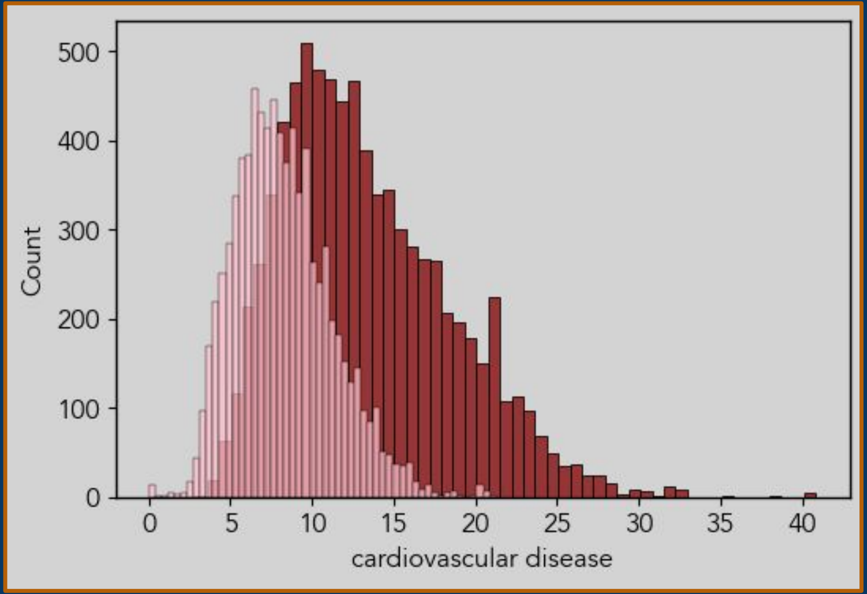


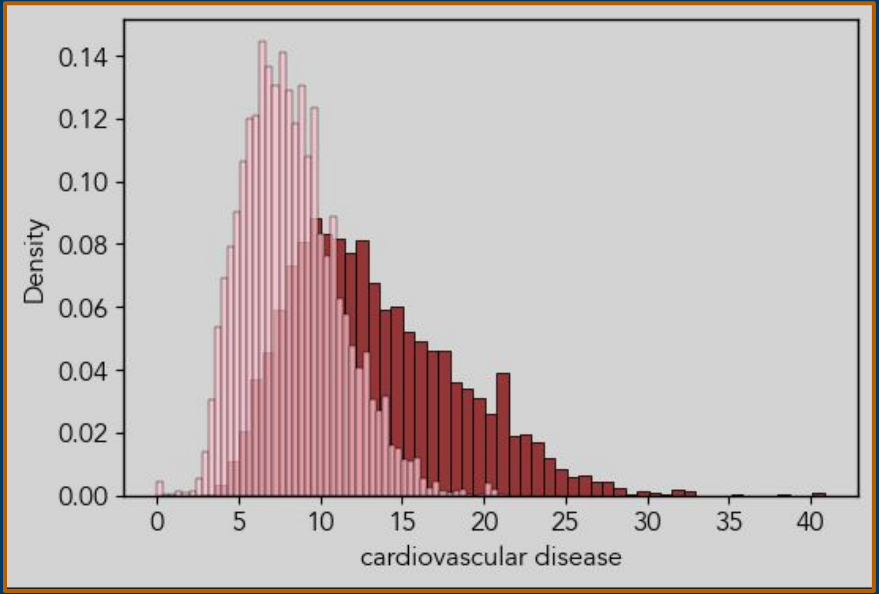
counts: more data in ES 2, 3, 4



density: distribution relatively the same

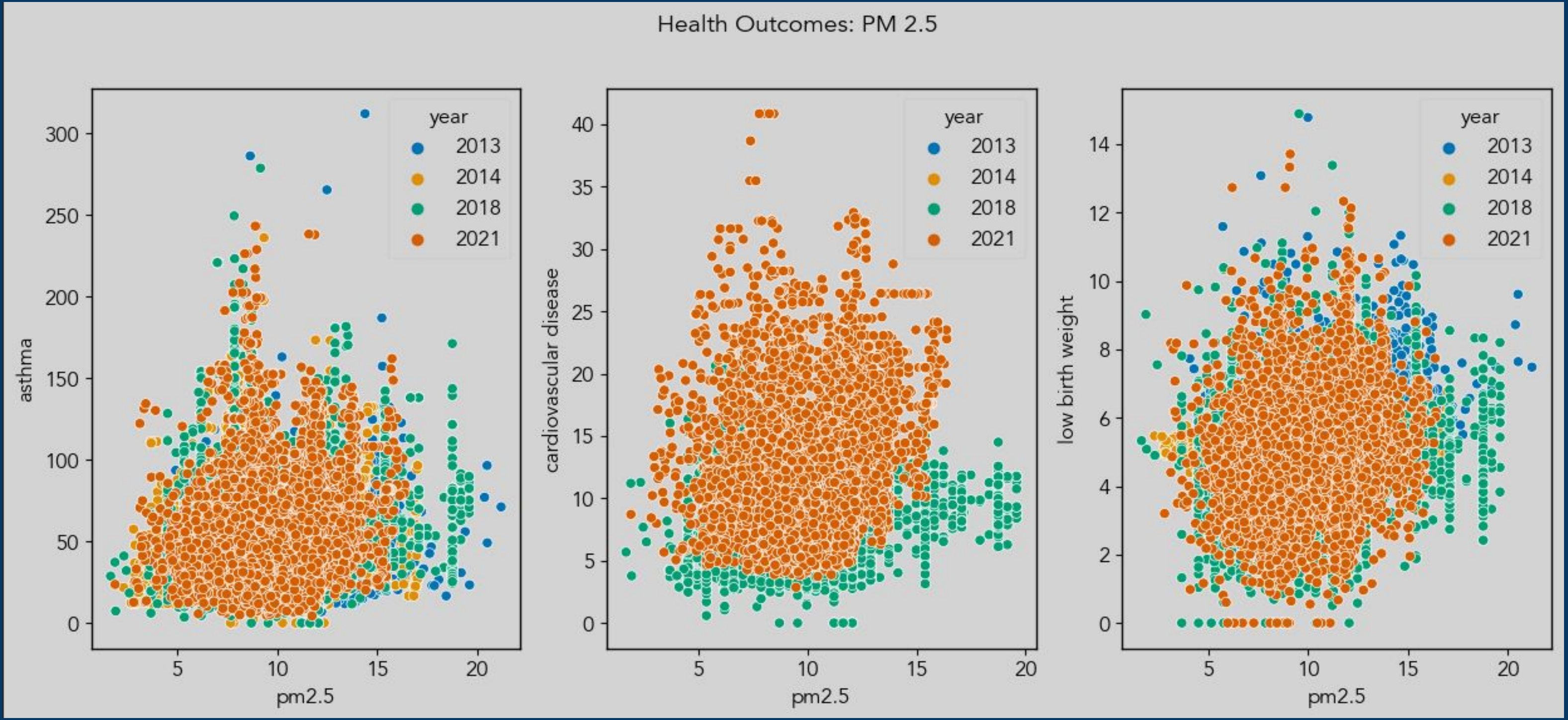
EDA: Cardiovascular



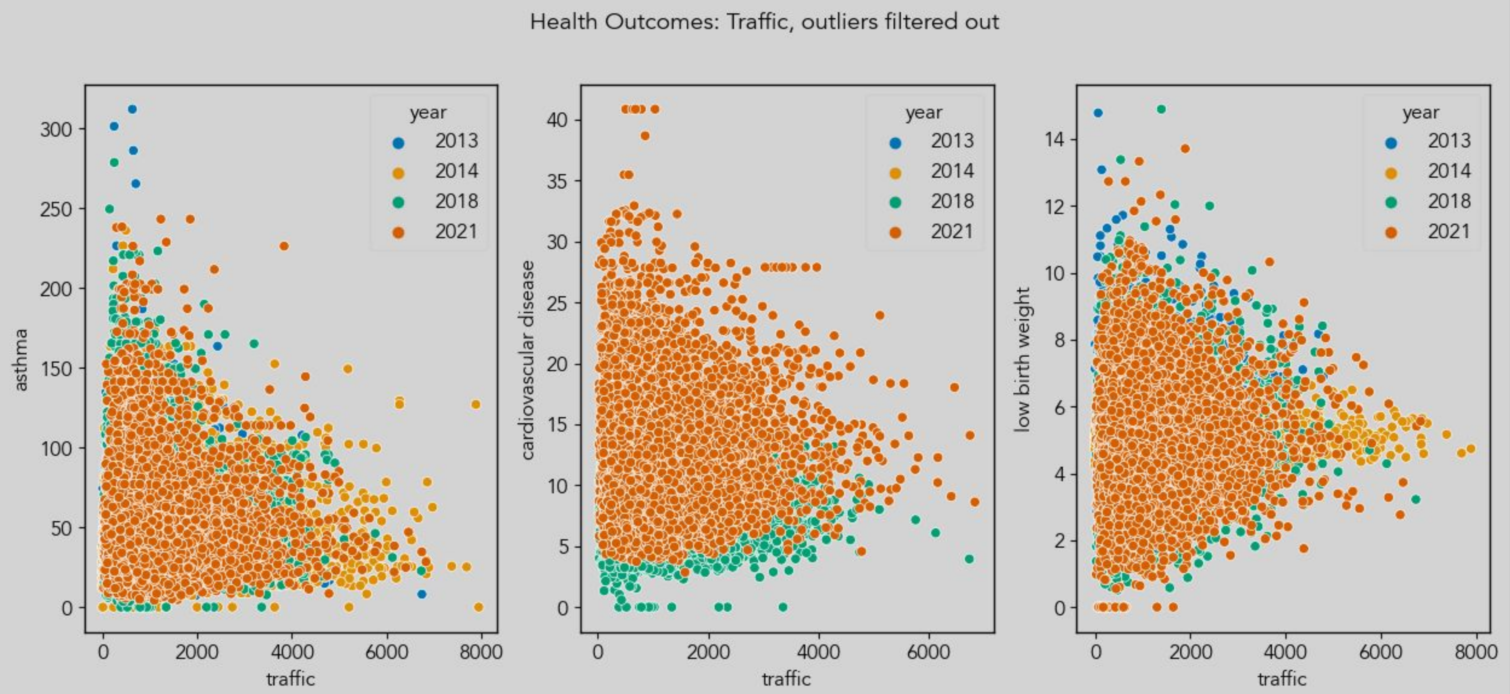


- data in 2018, 2021 reports only
- increase over 3-year period

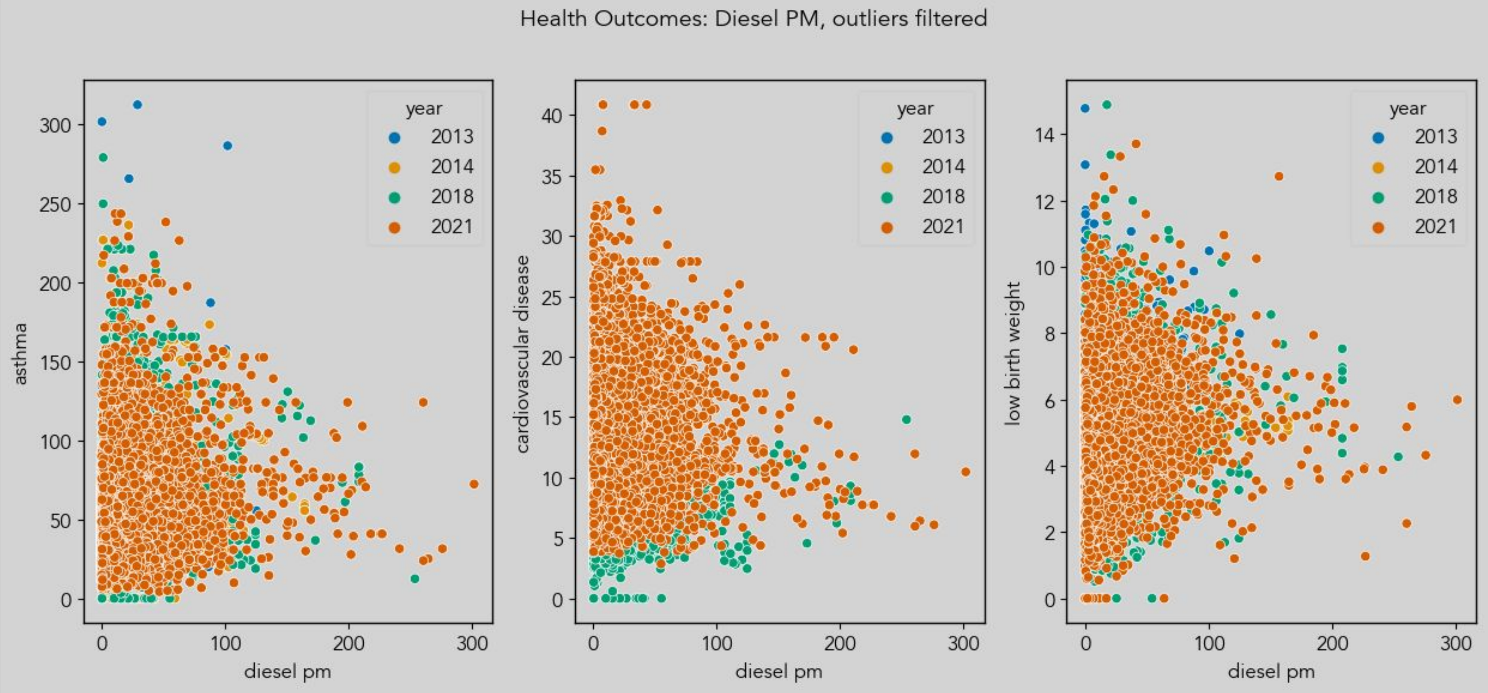
EDA: Health Outcomes, PM 2.5



EDA: Health Traffic Volume



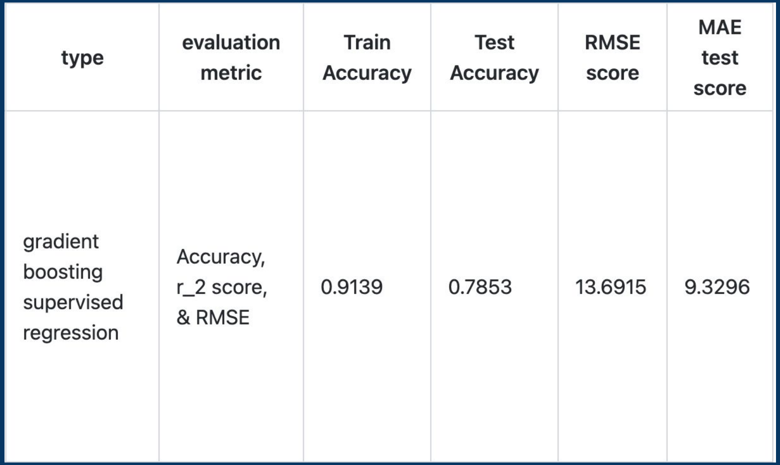
EDA: Health Outcomes, Diesel PM



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	Accuracy, r_2 score, & RMSE	0.9139	0.7853	13.6915	9.3296

BLURB

- **Maybe adding a graph above of predictions vs actual and and histogram of the residuals on next slide.**

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

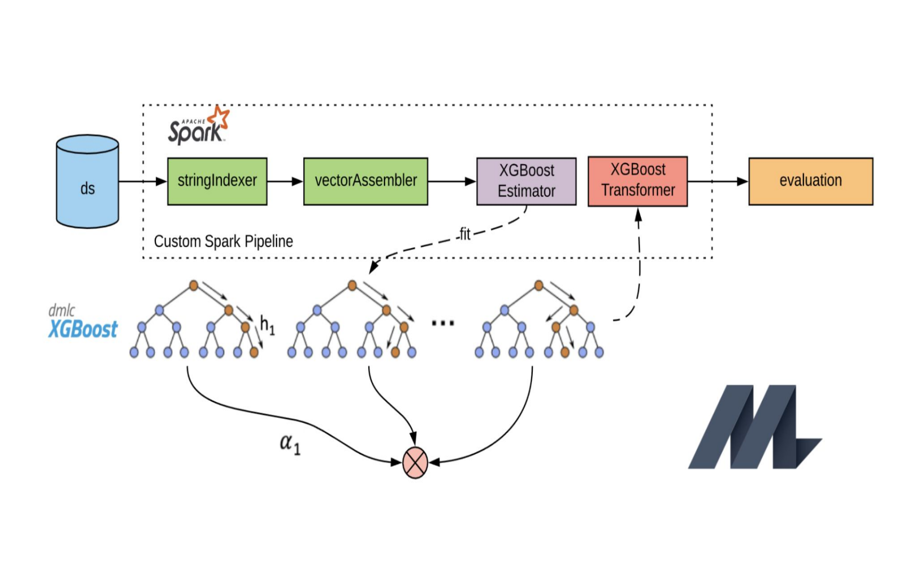
<u>Interpretation</u>

-

FINAL METRICS(old)

Train Accuracy:0.9472151826696329
Test Accuracy:0.7639090300436409
RMSE score:14.376141

XGboost GS CV fit to best params Asthma



FEATURES

-

Interpretation

-

BLURB

-

FINAL METRICS

Train Accuracy:0.9472151826696329
Test Accuracy:0.7639090300436409
RMSE score:14.376141

Model 2(marshall) Random Forest Reg

NUMERIC

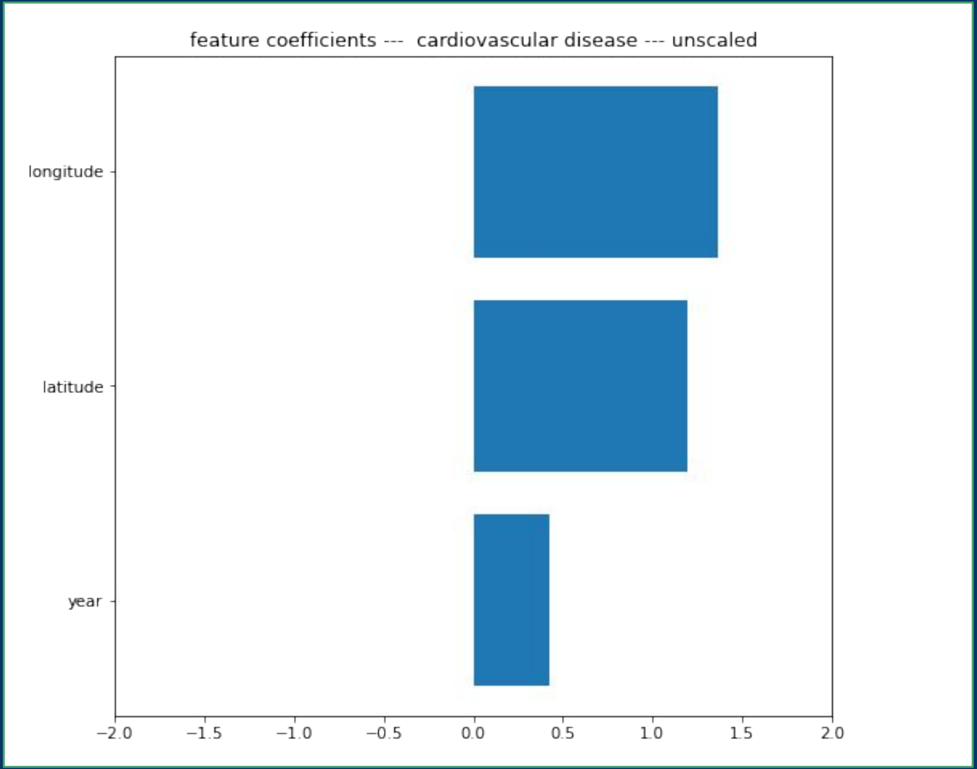
CATEGORICAL

BLURB

FINAL METRICS

Linear models

Linear model: health targets with year and location



NUMERIC FEATURES

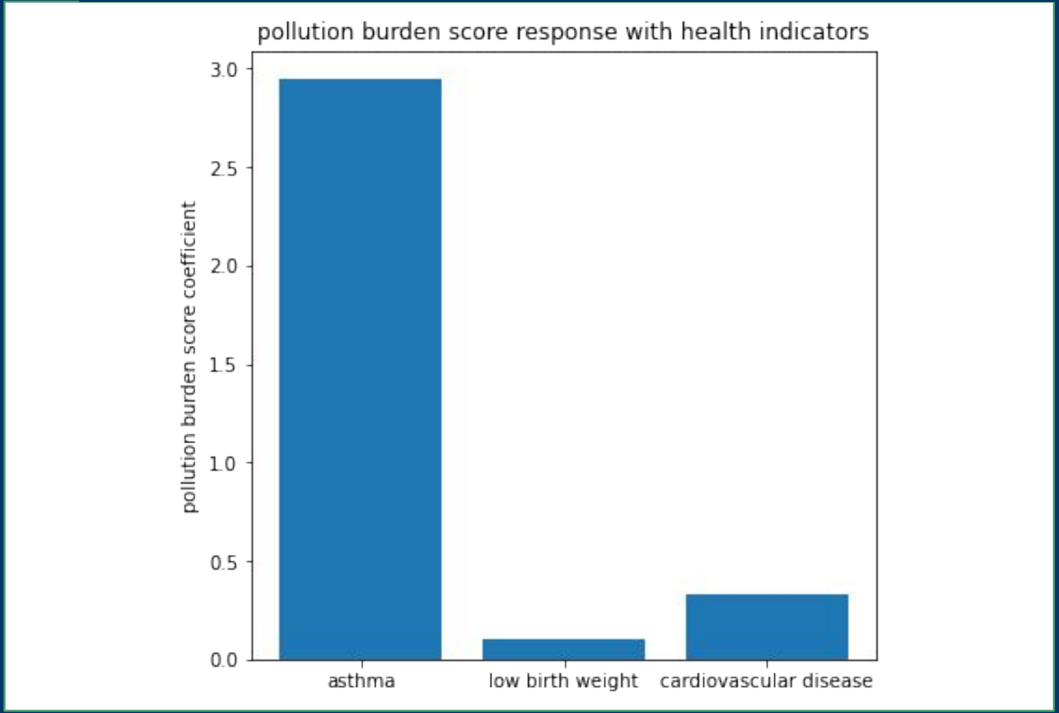
Year
Latitude
longitude

FINAL METRICS: R^2	
Asthma	0.054
Low birth weight	0.023
Cardiovascular disease	**0.17**

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

Cardiovascular disease ER visits

Linear model: CAES score features only



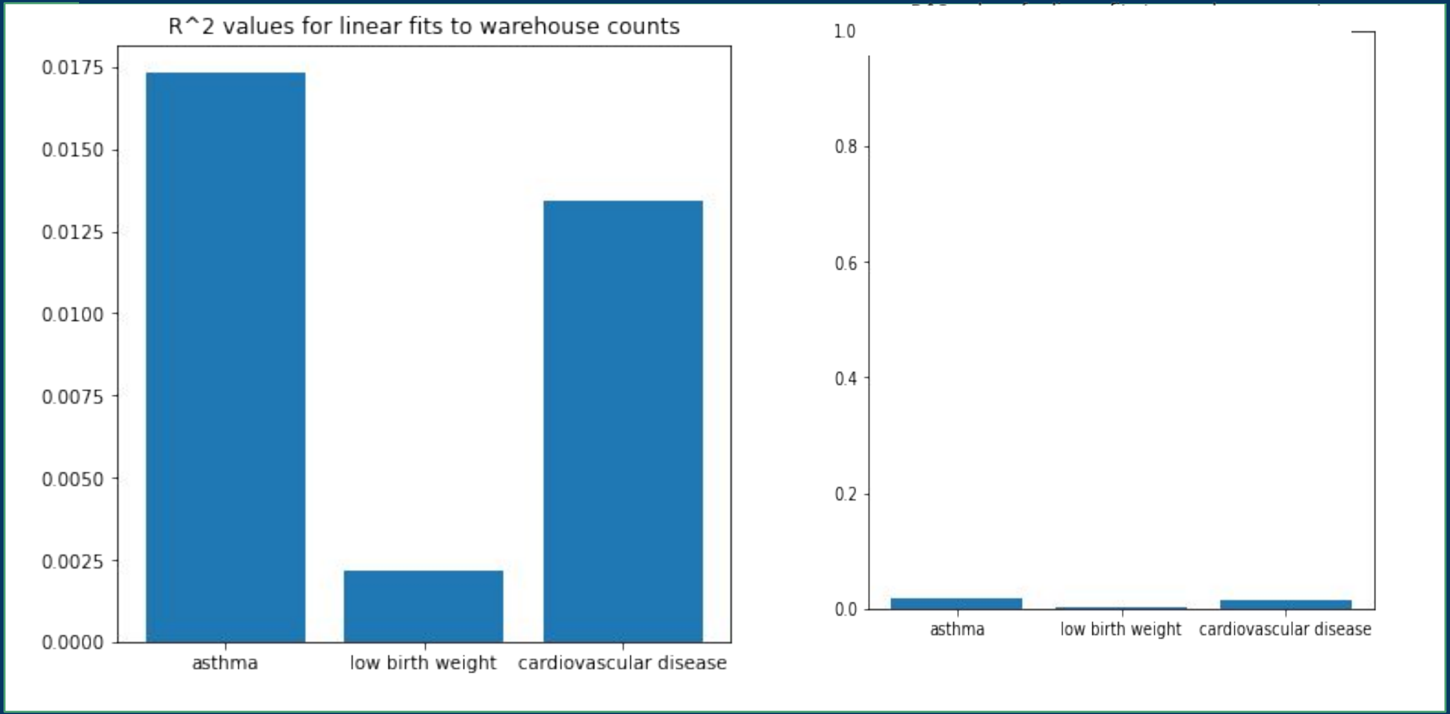
Evaluating the impact of CAES scores: Pollution burden

NUMERIC

CATEGORICAL

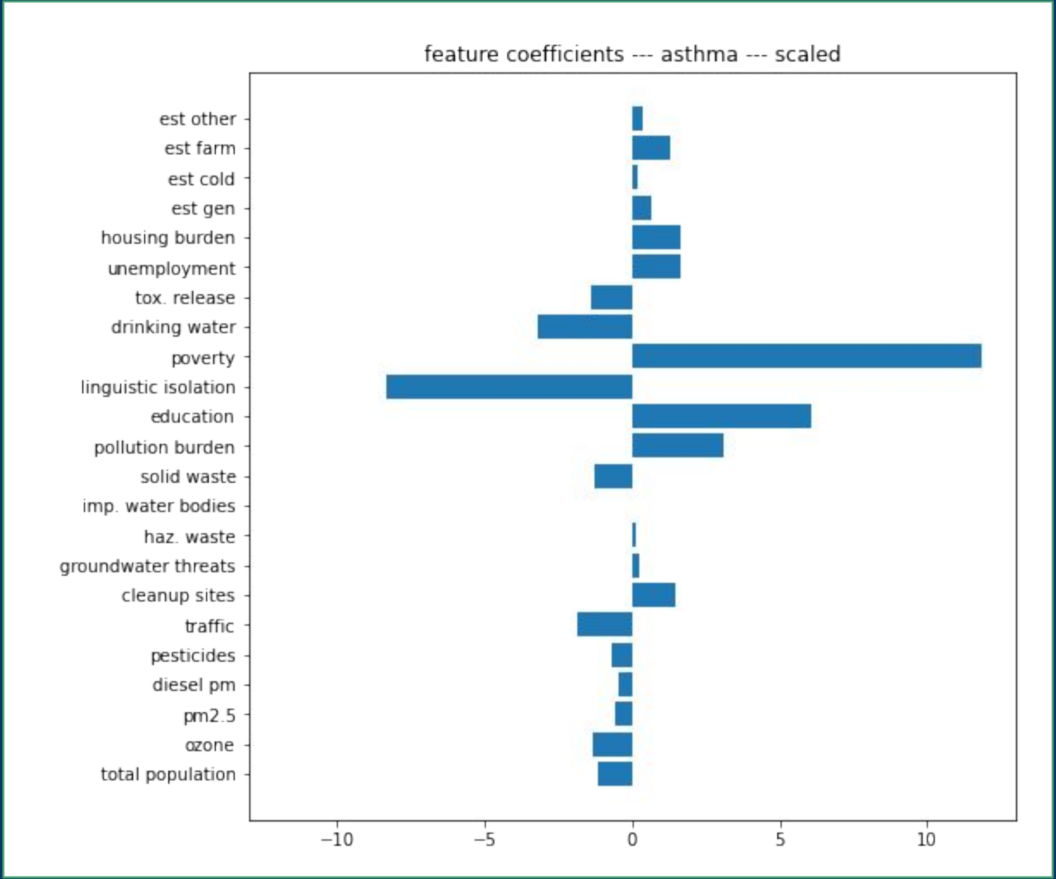
FINAL METRICS

Linear model: Warehouse counts



Evaluating the impact of warehouse business types

Linear model: “Selected columns” — Asthma



NUMERIC

'total population',

'ozone',

'pm2.5',

'diesel pm',

'pesticides',

'traffic',

'cleanup sites',

'groundwater threats',

'haz. waste',

'imp. water bodies',

'solid waste',

'pollution burden',

CATEGORICAL

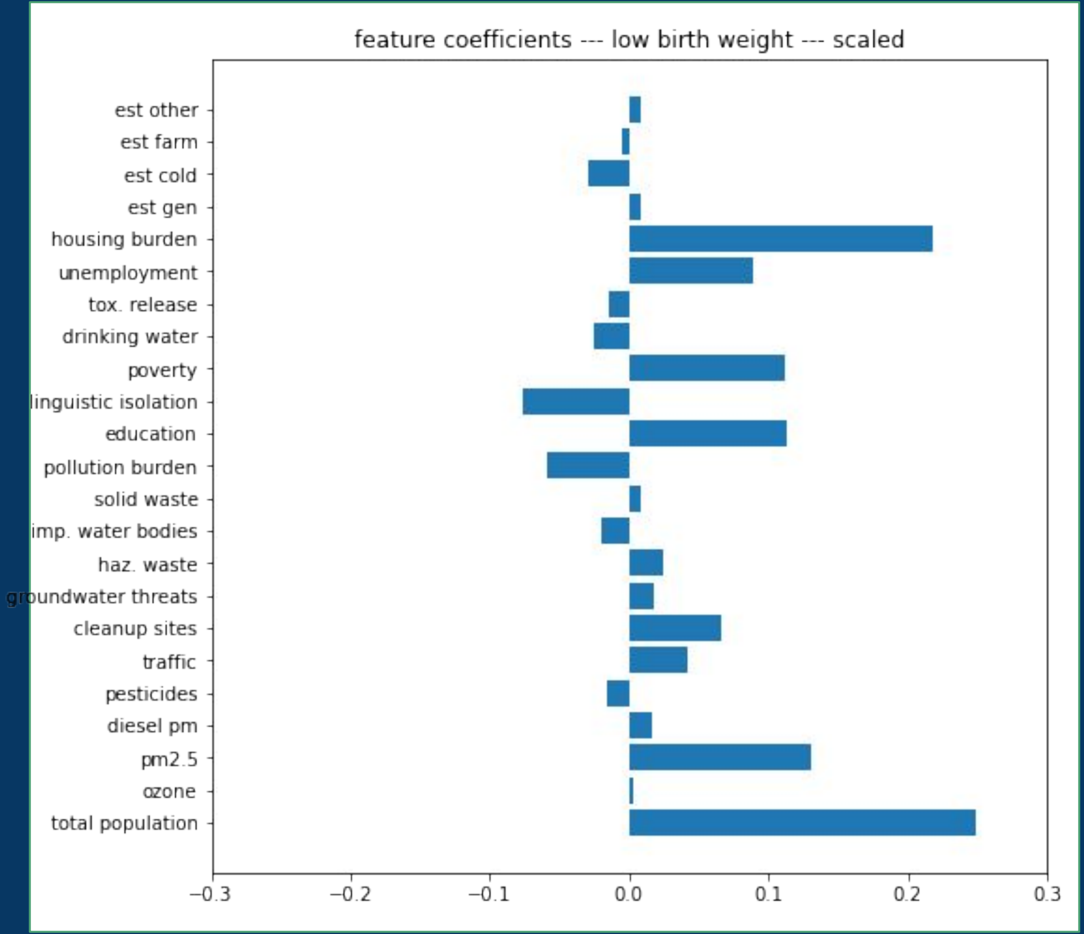
FINAL METRICS

R^2

Asthma - 0.29

Low birth weight - 0.14

Cardiovascular disease - 0.23



Linear model: “Selected columns” — Low birth weight

NUMERIC

'total population',

'ozone',

'pm2.5',

'diesel pm',

'pesticides',

'traffic',

'cleanup sites',

'groundwater threats',

'haz. waste',

'imp. water bodies',

'solid waste',

'pollution burden',

CATEGORICAL

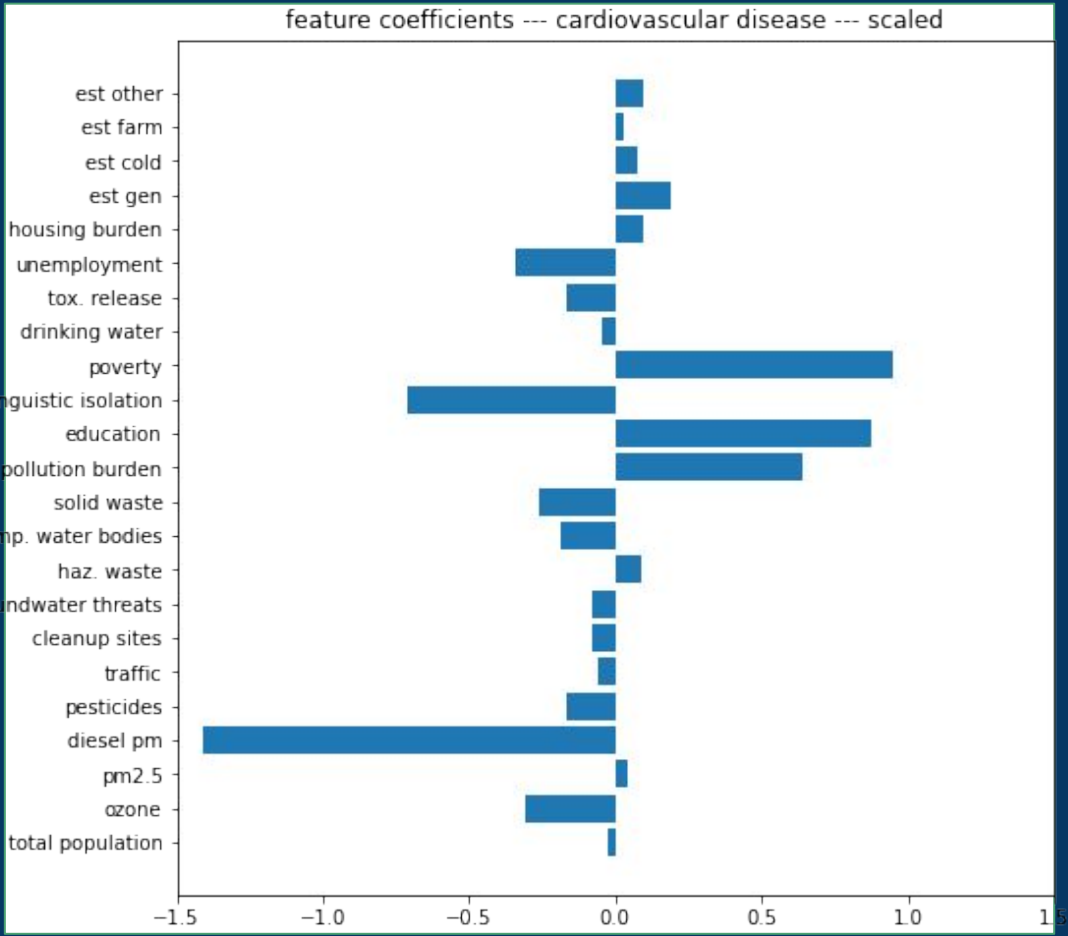
FINAL METRICS

R^2

Asthma - 0.29

Low birth weight - 0.14

Cardiovascular disease - 0.23



Linear model: “Selected columns” — Low birth weight

NUMERIC

'total population',

'ozone',

'pm2.5',

'diesel pm',

'pesticides',

'traffic',

'cleanup sites',

'groundwater threats',

'haz. waste',

'imp. water bodies',

'solid waste',

'pollution burden',

CATEGORICAL

FINAL METRICS

R^2

Asthma - 0.29

Low birth weight - 0.14

Cardiovascular disease - 0.23

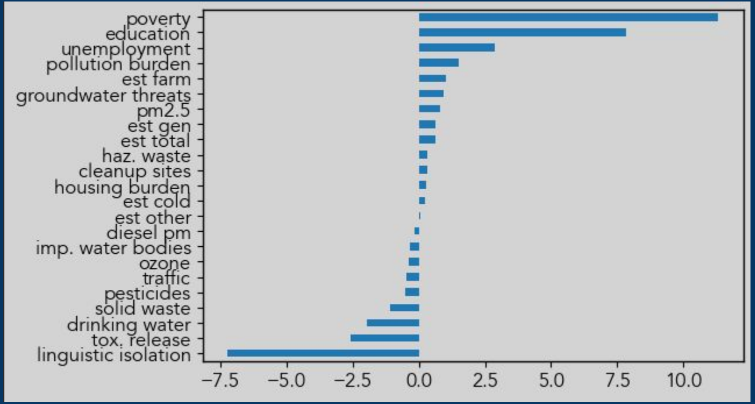
Model: SVR

Epsilon-Support Vector Regression

regularization: L2, C = 1

NUMERIC

CATEGORICAL



Feature Importances: really highlights

FINAL METRICS

Model: Random Forest Regression

n-estimators = 100

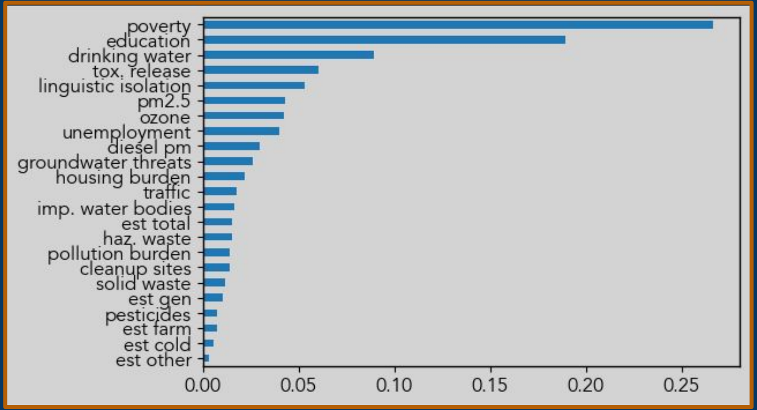
max_depth = 10

max_leaf_nodes = 10

max_features : auto

NUMERIC

CATEGORICAL



different importances:

FINAL METRICS

Reliable models of warehouse effect on health-outcomes were unattainable (???) with this data.

Site
60/10 Fwy. On/Off-ramp via 6th Street
10 Fwy. On/Off ramp via Beaumont Ave.
10
60
Sixth St.
Veile Ave.
W. 4th St.
California Ave.
Beaumont Ave.
W. 1st St.

- 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.

Twenty breeds of dogs

keep this slide

DOGS

DOGS

DOGS