Health Outcomes v. Warehouse Location

Something.

Intro: Diesel Trucks are Bad For Babies

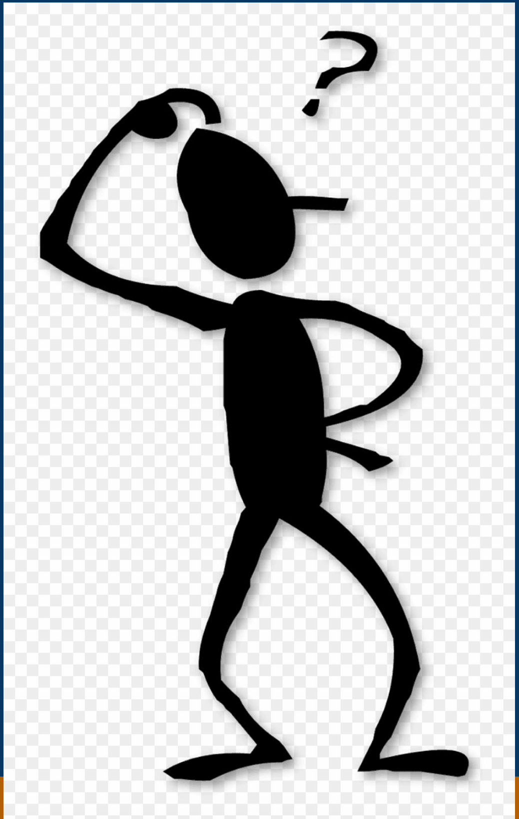
FEATURES DROPPED

SURPRISING FEATURES

BLURB

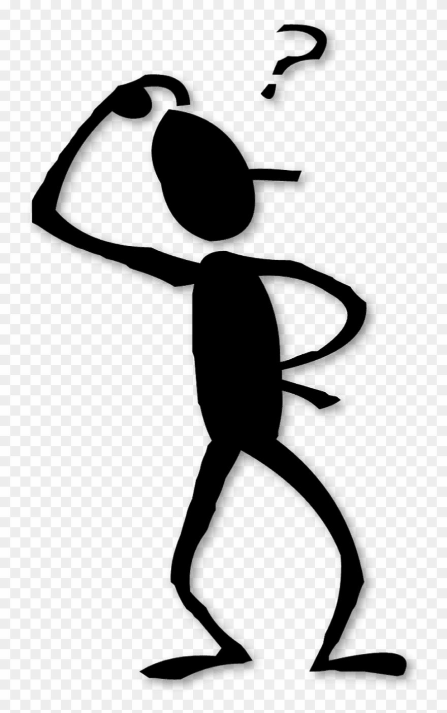
MANIPULATION

THE PROBLEM STATEMENT SLIDE



- What is the (quantifiable) effect of increased warehouse presence in California in the last decade on emergency healthcare?
- How well do the CalEnviroScreen scores reflect emergency healthcare counts?
- What indicators from the CalEnviroScreen dataset best determine the number of emergency healthcare visits?

Aggregated Modeling With Additional Features



As Data Scientists in *OEHHA*, we are tasked with developing models aggregating the four time-points from each report with additional information on warehouse density to assess primary mitigating factors addressing negative health outcomes.

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 20–, 20–, 20–, and 20–, 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.

The CalEnviroScreen Model

EnviroScreen-specific “scores” are derived from measurements.

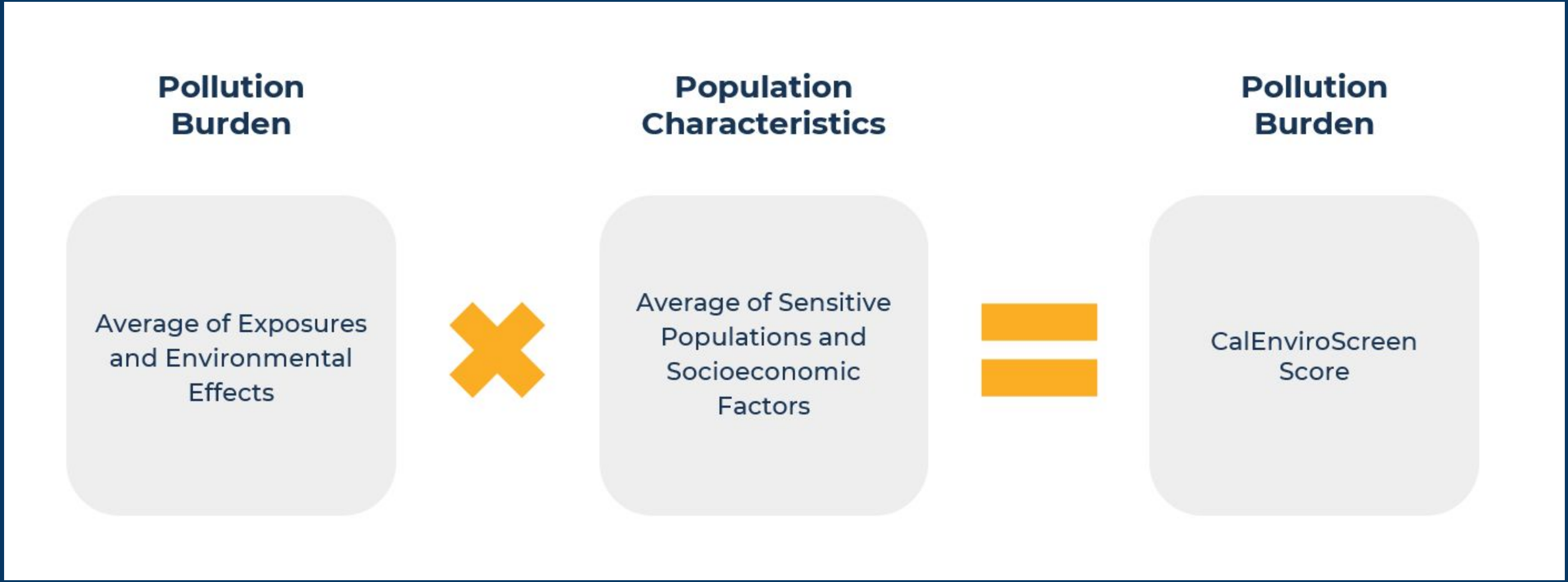
- Pollution Burden Score
  - Exposures
    - Ozone concentrations
    - Particulate matter emissions and concentrations (diesel, PM2.5)
    - Drinking water contaminants, lead risk
    - Toxic releases from facilities, pesticide use
    - Traffic density
  - Environmental Effects
    - Solid waste, sites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

The CalEnviroScreen Model



Pollution Burden

Average of Exposures and Environmental Effects

Population Characteristics

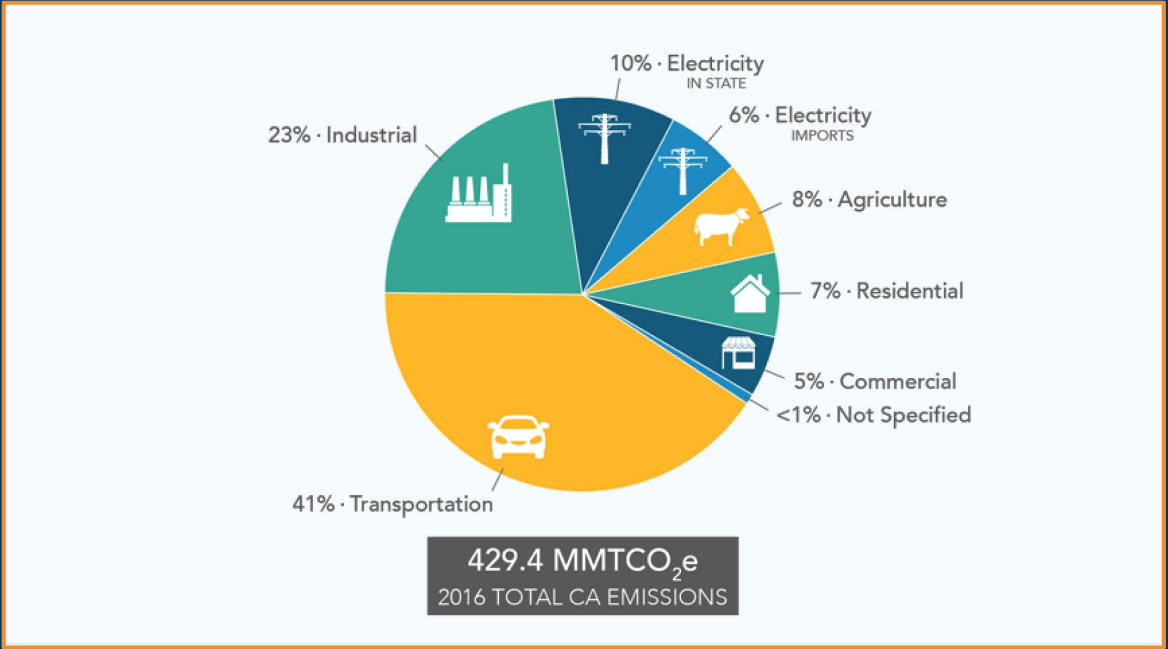
Average of Sensitive Populations and Socioeconomic Factors

Pollution Burden

CalEnviroScreen Score

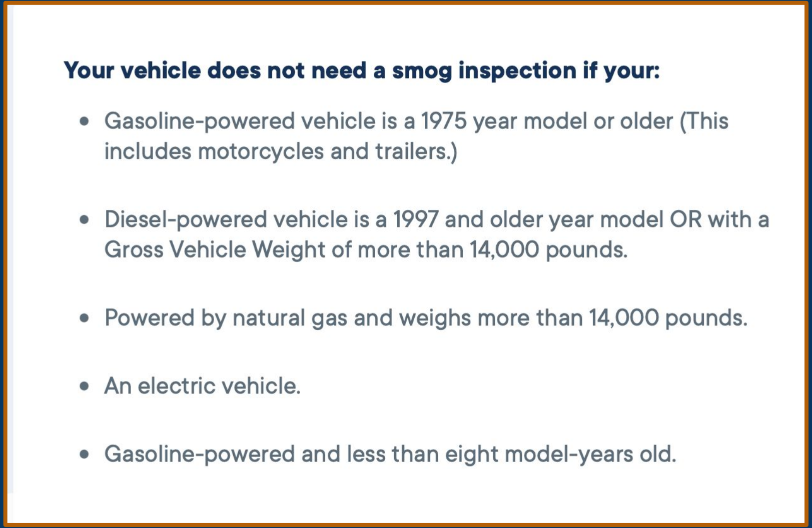
https://oehha.ca.gov/calenviroscreen/scoring-model

Motivation: Assessing Effect of Emission Sources



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

Motivation: Emissions Exceptions



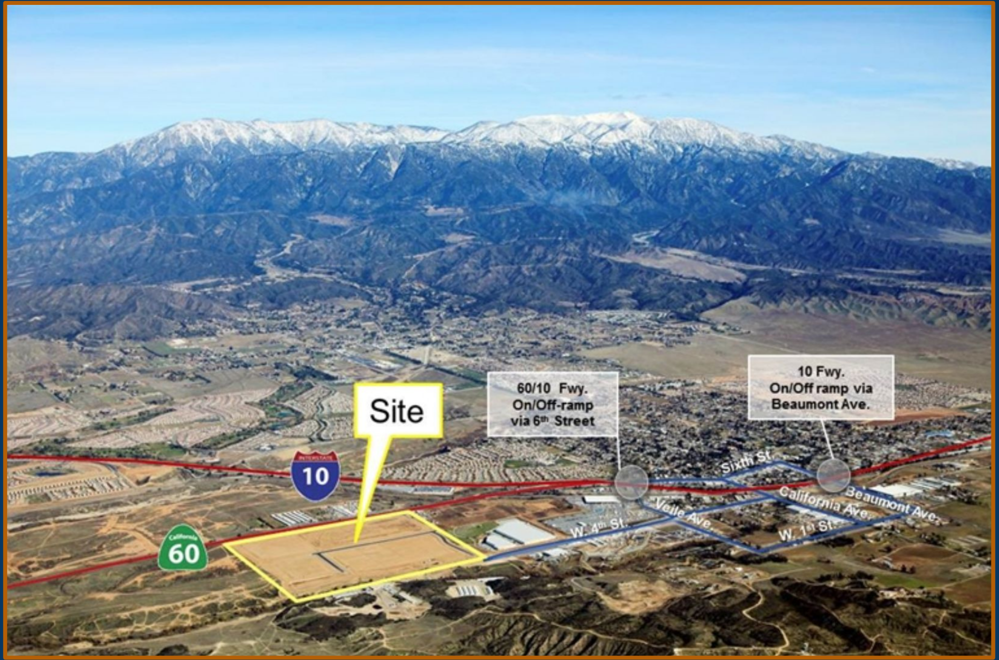
Your vehicle does not need a smog inspection if your:

- 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/

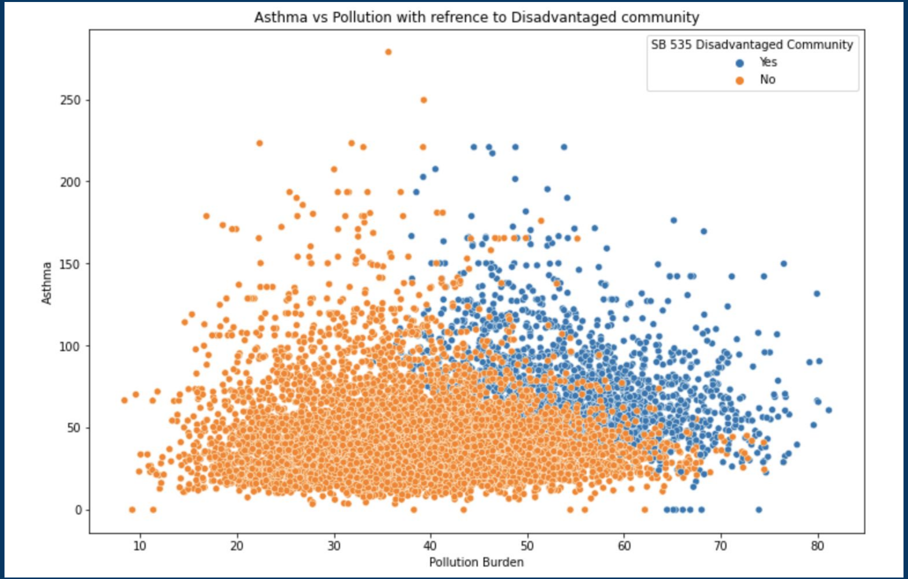
1995 Freightliner for Sale in San Rafael, Marin County, San Francisco Bay Area, CA

Motivation: Landscape Changes


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

SOURCE:
riversidewarehouses.com/listings/1020-prosperity-way-beaumont-ca-92223

EDA with with health, pollution, and Poverty(SB)



BLURB

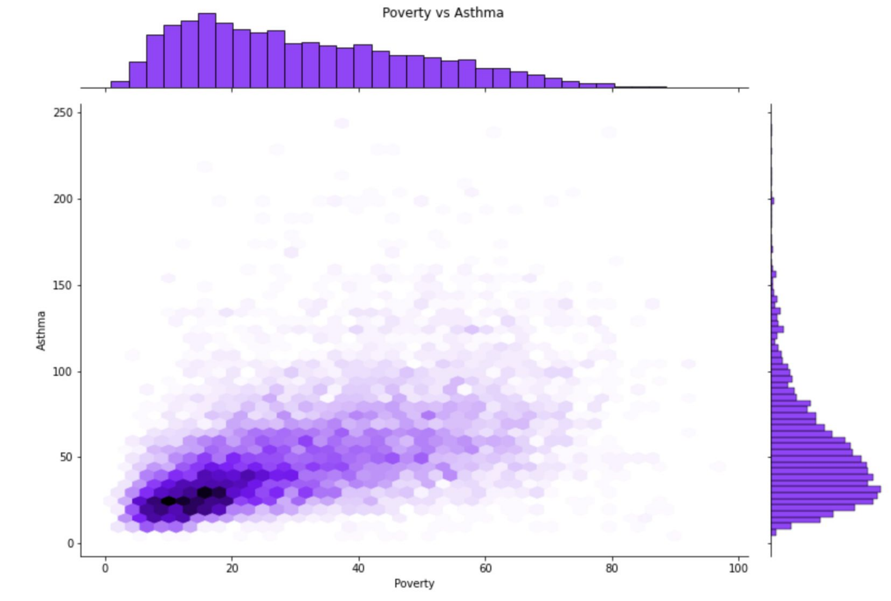
FEATURES DROPPED

-
-
-
-
-
-
-
-
-

SURPRISING FEATURES

MANIPULATION

EDA visuals cont. (marshall)



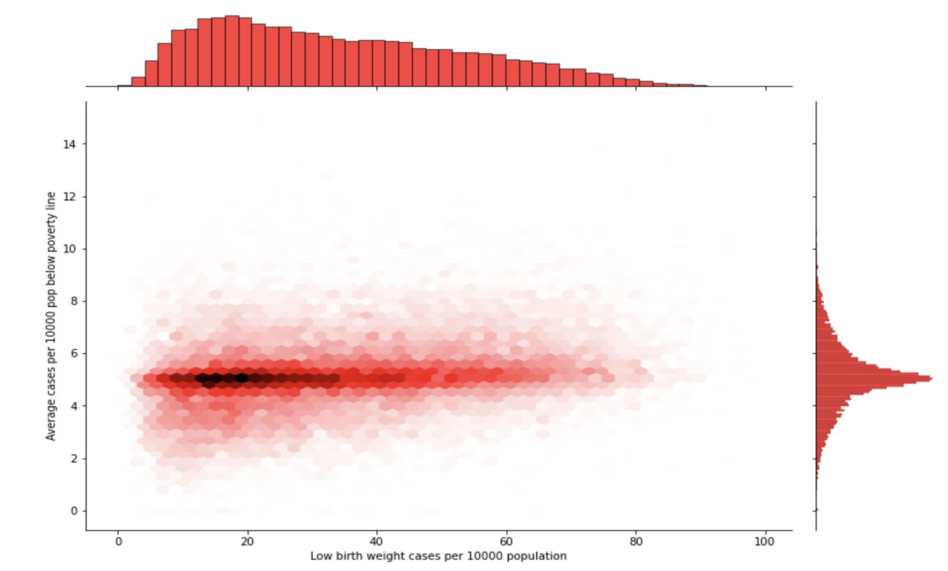
FEATURES DROPPED

SURPRISING FEATURES

BLURB

MANIPULATION

Most important features for health cont.



FEATURES DROPPED

SURPRISING FEATURES

BLURB

MANIPULATION

EDA : Census data info – warehouse counts

BLURB

Broad statistics on warehouse counts

MANIPULATION

SURPRISING FEATURES

Joined onto CAES data by tract or zip

Changes with time

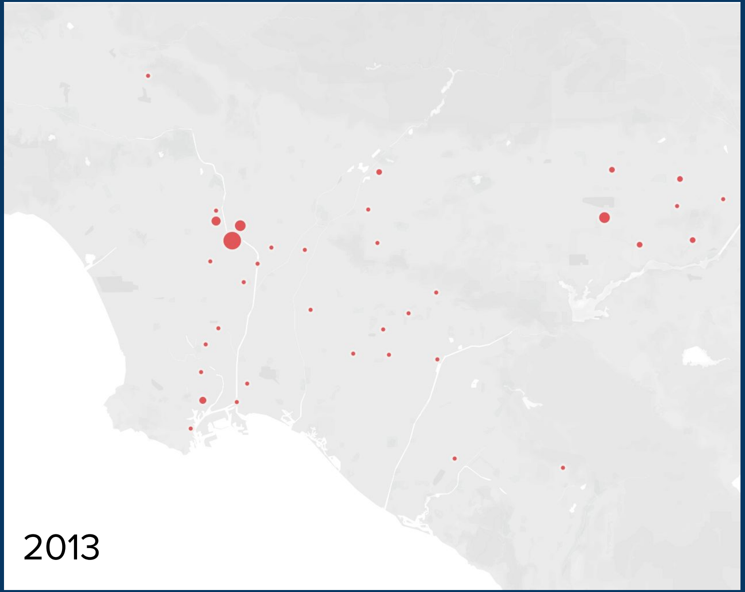
Joined onto CAES data by tract or zip

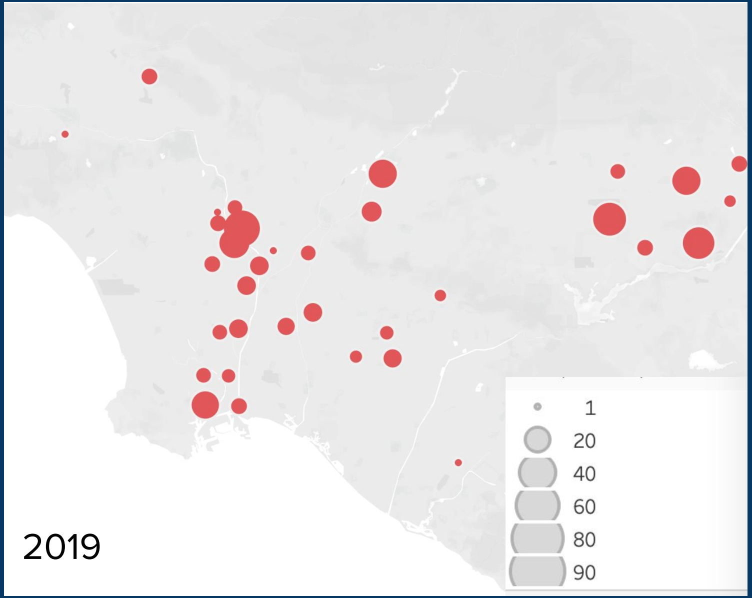
Time by zip — board warehouse business changes.

Or, just time with california as a whole.

MAP

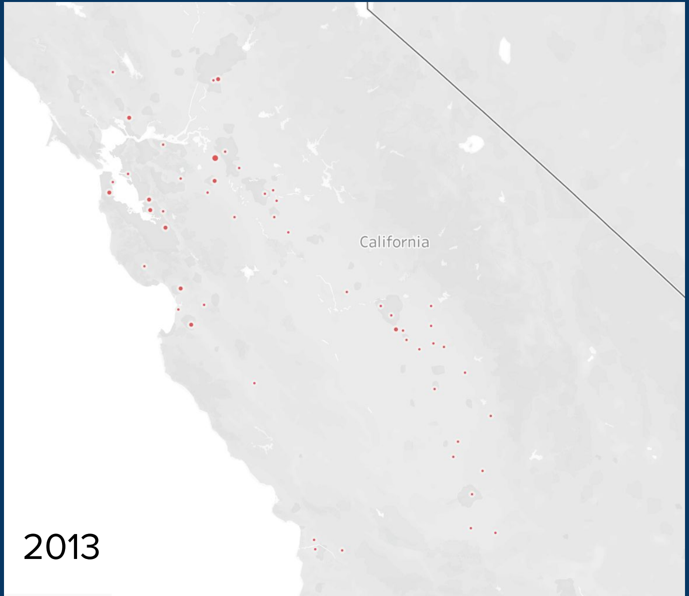
Time by zip — board warehouse business changes.

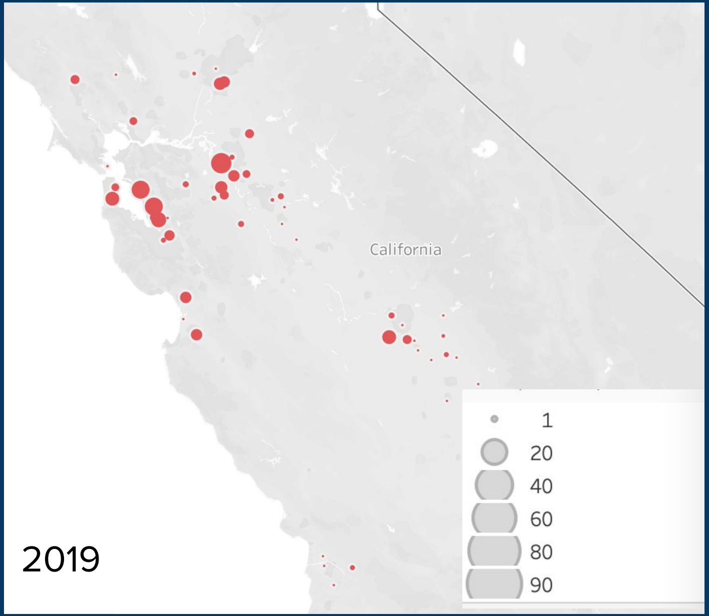




Los Angeles - Inland Empire, Est. Total General Warehouses

Time by zip — board warehouse business changes.





Central California - Estimated Total General Warehouses

Time by zip — w ? Health or pollution

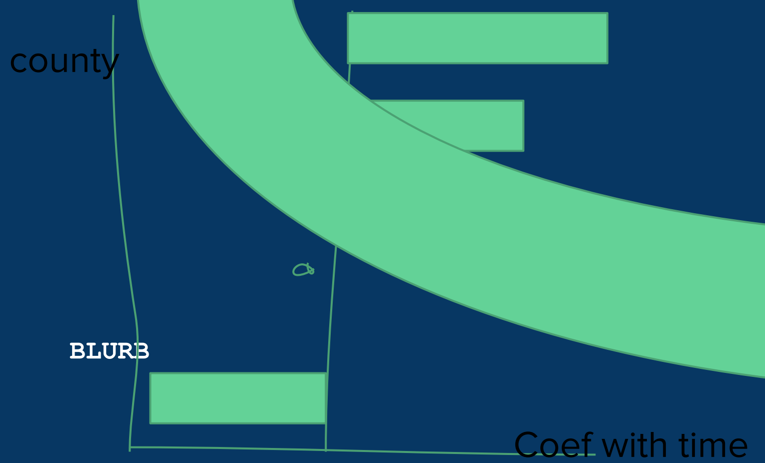
each county?

NUMERIC

As t/best (with time)

weight

Cardiovascular disease worst/bes



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

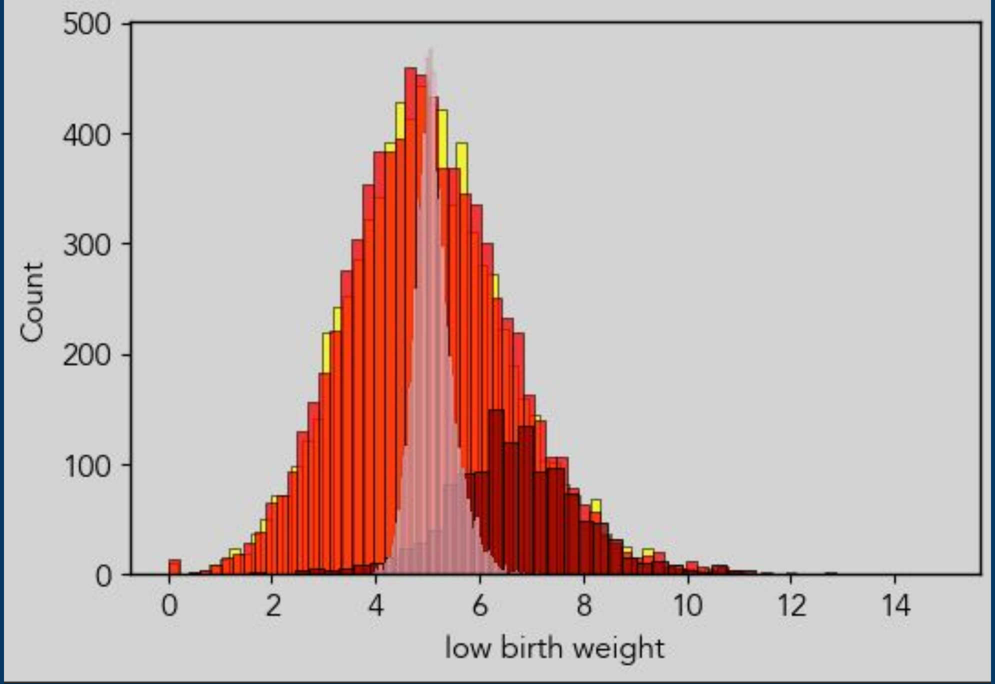
Time by zip — what are the biggest changers? Health or pollution

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

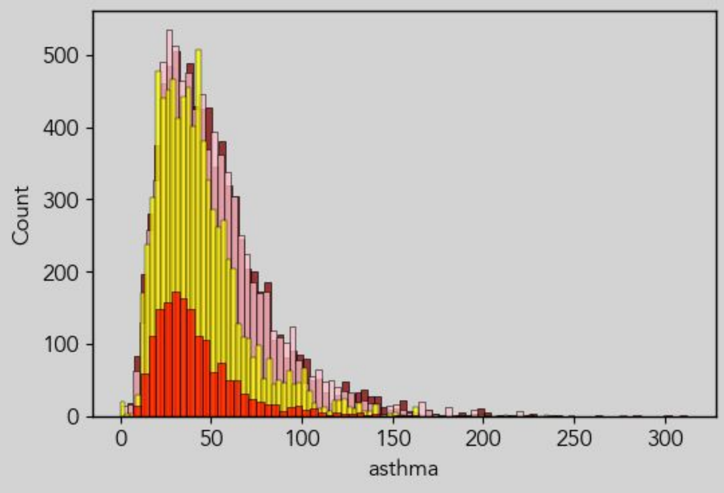
Colored map

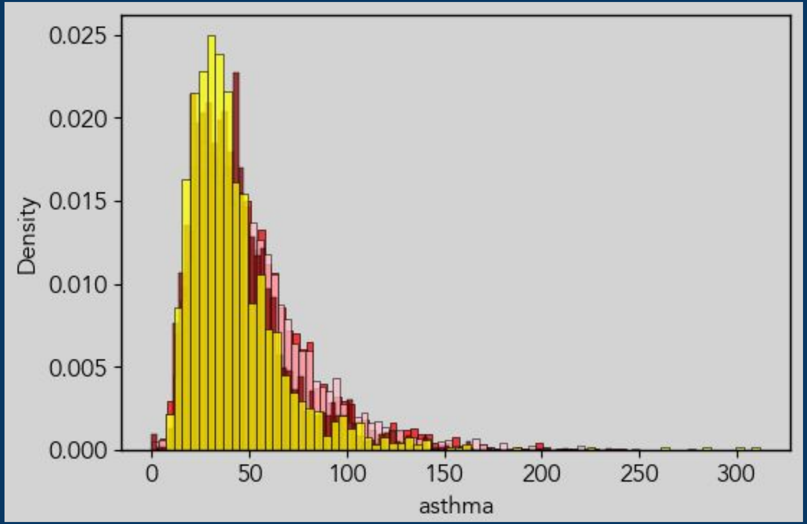
These highlight regions that may be trouble soon, or rapidly improving. (eda should show bad/good *currently*)

EDA: Low-Birth Weight

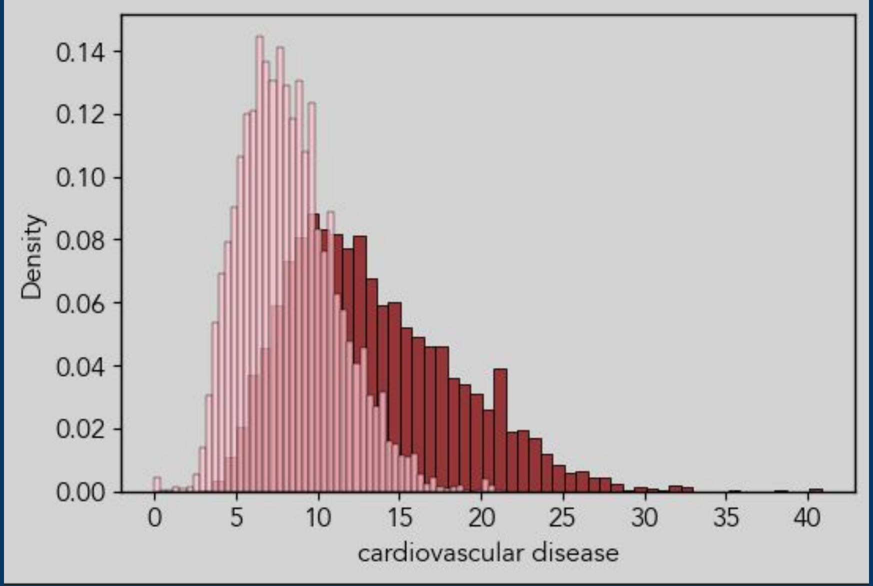


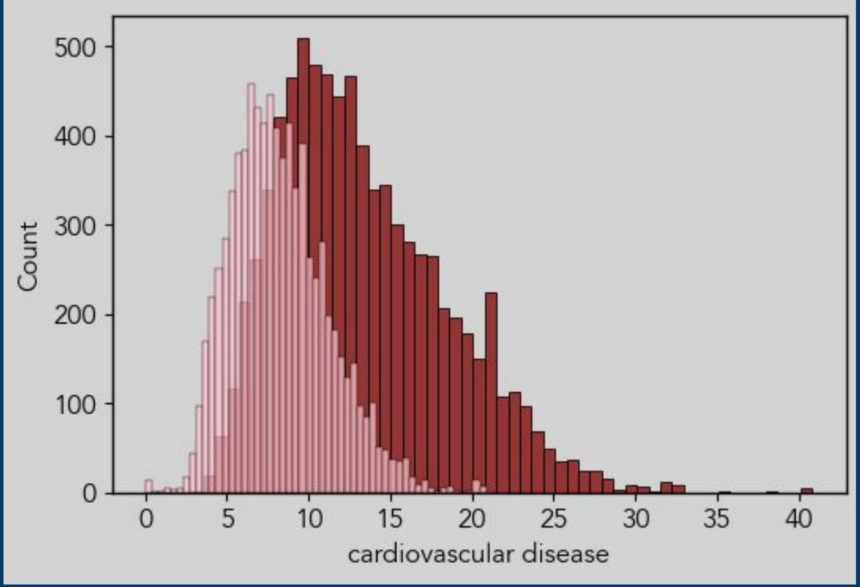
EDA: Asthma


Count
500
400
300
200
100
0
0
50
100
150
200
250
300
asthma

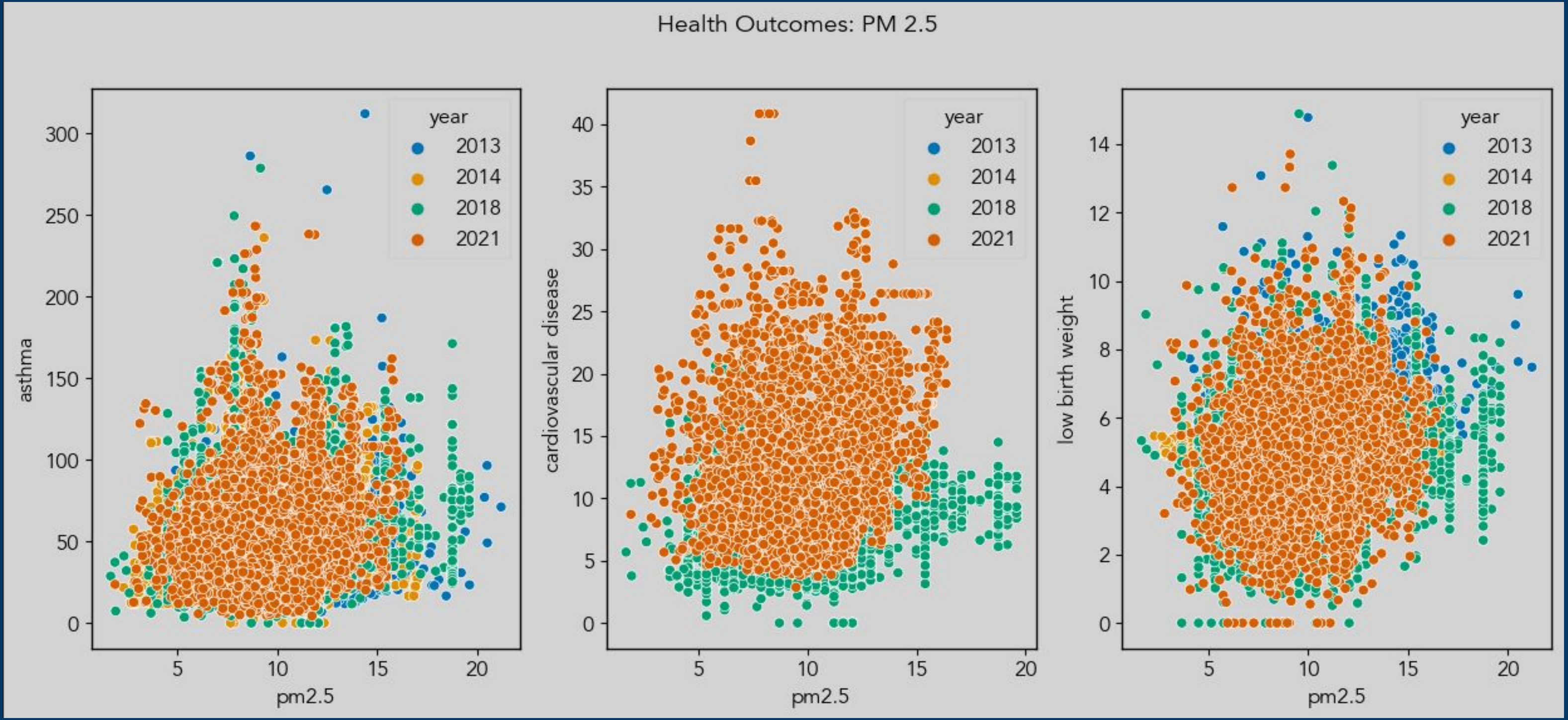

Density
0.025
0.020
0.015
0.010
0.005
0.000
0
50
100
150
200
250
300
asthma

EDA: Cardiovascular

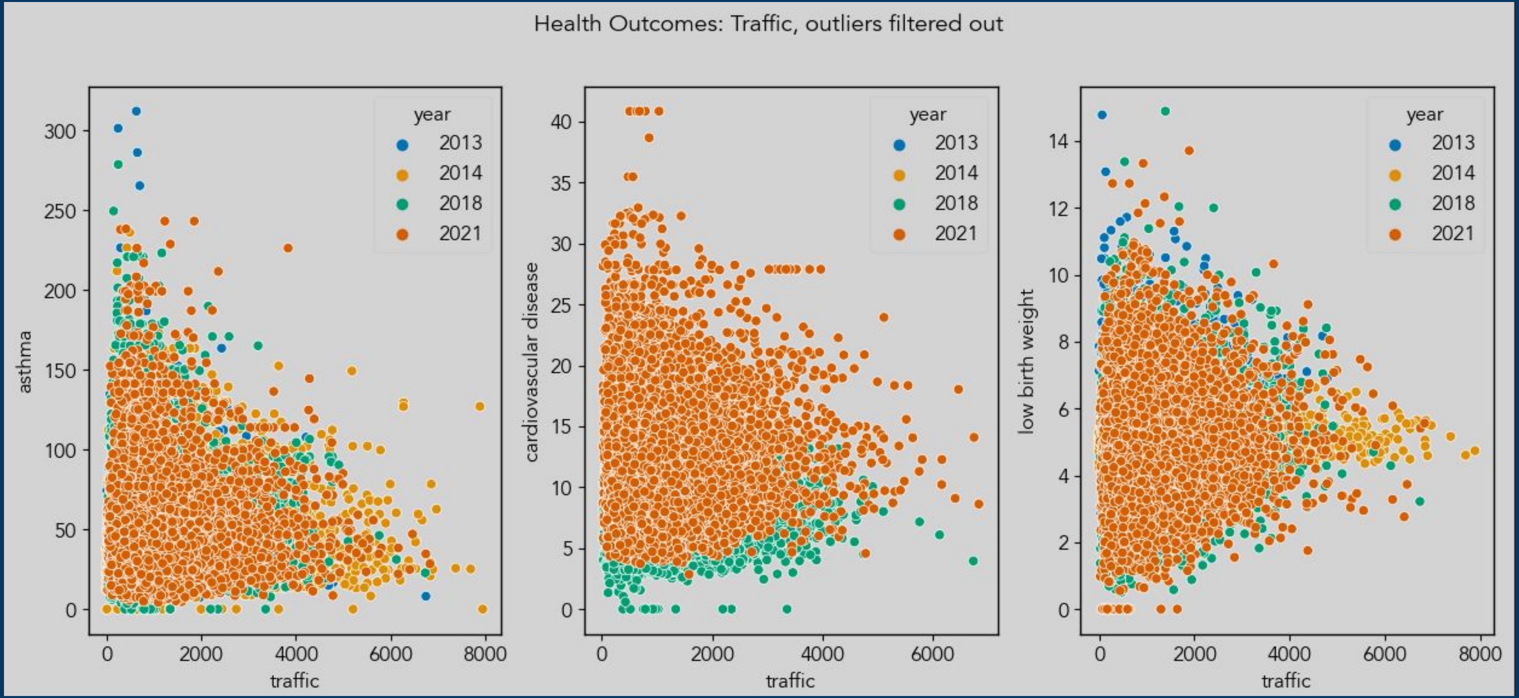




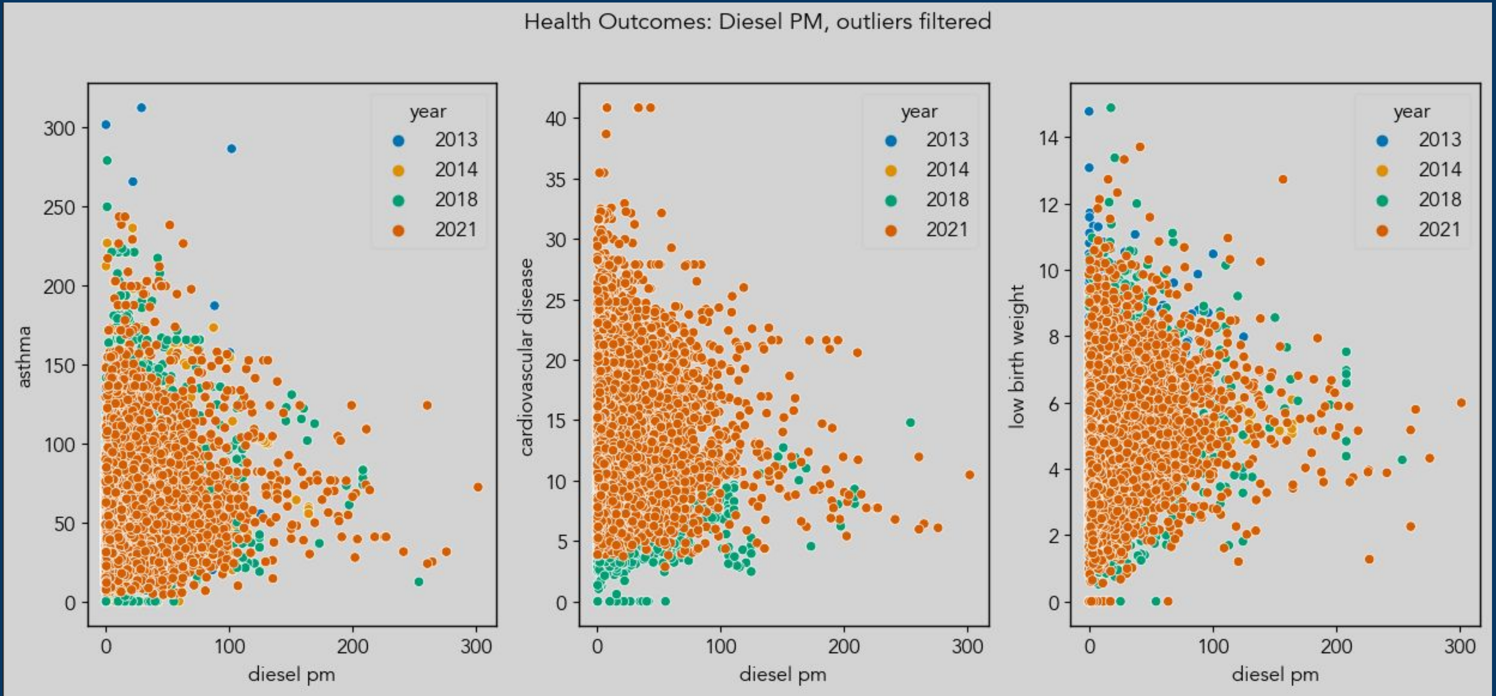
EDA: Health Outcomes, PM 2.5



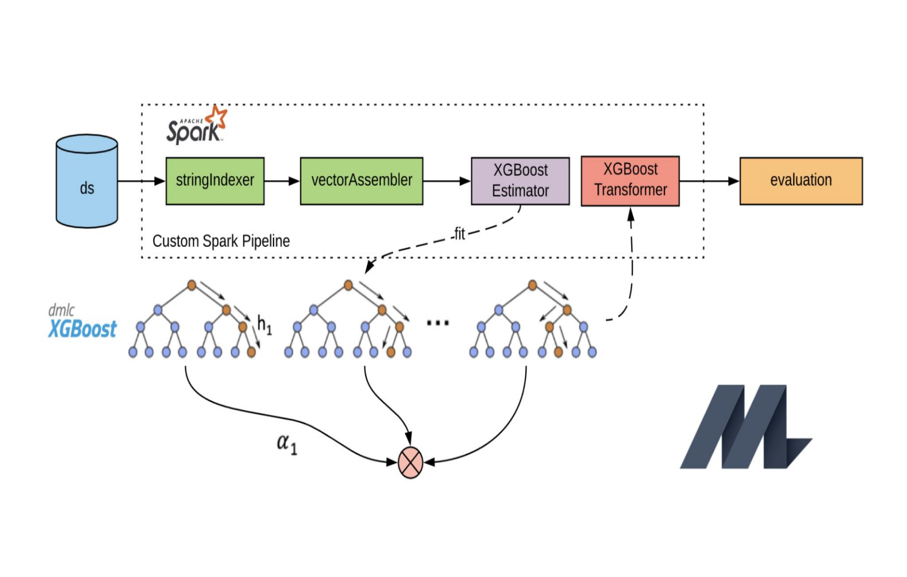
EDA: Health Traffic Volume



EDA: Health Outcomes, Diesel PM



XGboost, scaled, highest correlated features for Asthma



FEATURES

-

Interpretation

-

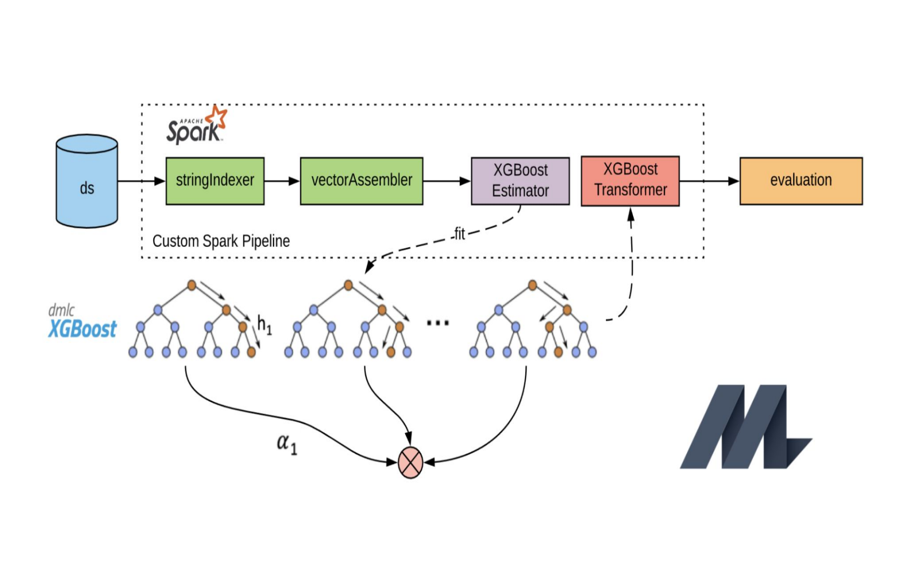
BLURB

- **XGBoost** can be used directly for **regression predictive modeling**.
-

FINAL METRICS

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 model : time and space only

NUMERIC	CATEGORICAL
Year	
Latitude	
longitude	

BLURB

FINAL METRICS
Asthma - (0.042, 0.060)
Low birth weight - (0.007, 0.010)
Cardiovascular disease - (0.38, 0.38)

Linear model: CAES score features only

NUMERIC

CATEGORICAL

BLURB

Do this to evaluate the CAES scores. "Have they done the feature engineering for us already?"

FINAL METRICS

Linear model: “Selected columns” (pollution and industry info. Most of the features.)

NUMERIC

CATEGORICAL

BLURB
No year. No space. No CAES specific scores.

FINAL METRICS
r^2
Asthma - 0.593
Low birth weight - 0.385
Cardiovascular disease - 0.497

More on linear “selected features.”

NUMERIC

CATEGORICAL

Which features were most influential on the linear scale? (not svd/PCR, but just relative to scaled data. This is coefficient relative to scaled data.)

BLURB
No year. No space. No CAES specific scores.

FINAL METRICS, R^2:
Asthma - 0.593
Low birth weight - 0.385
Cardiovascular disease - 0.497

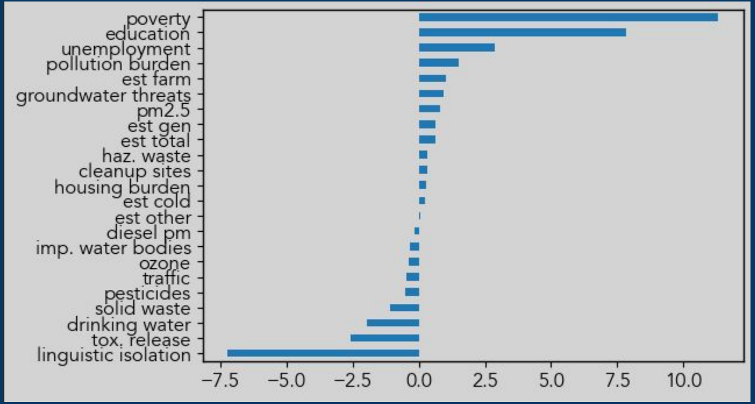
Model: SVR

Epsilon-Support Vector Regression

regularization: L2, C = 1

NUMERIC

CATEGORICAL



Feature Importances: really highlights

FINAL METRICS

Model: Random Forest Regression

n-estimators = 100

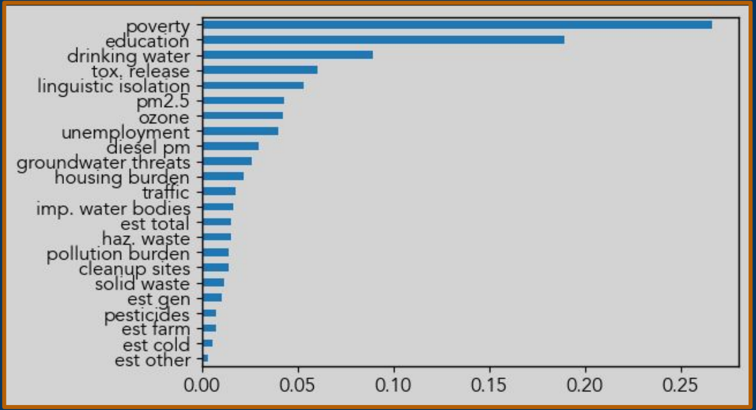
max_depth = 10

max_leaf_nodes = 10

max_features : auto

NUMERIC

CATEGORICAL



different importances:

FINAL METRICS

Twenty breeds of dogs

keep this slide

DOGS

DOGS

DOGS