Predicting Air Toxics Cancer Risk

Project Proposal

John Guo
Vishnu Kodicherla
Mya Carrizosa
General Assembly DSI-822

Table of contents

Problem Statement

- What is the best regression model to predict air toxics cancer rate using environmental and/or demographic features from the Environmental Protection Agency's Environmental Justice Screen data?

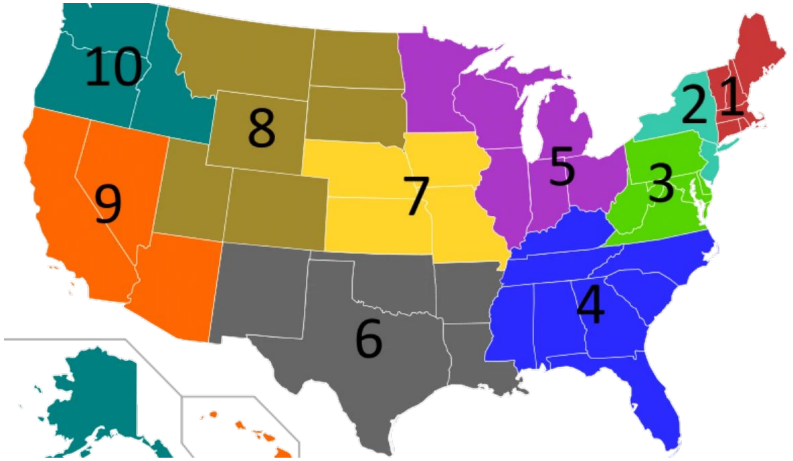
01

Datasets

Environmental Justice Screening Tract 2021
It has 25 features in total

Regions

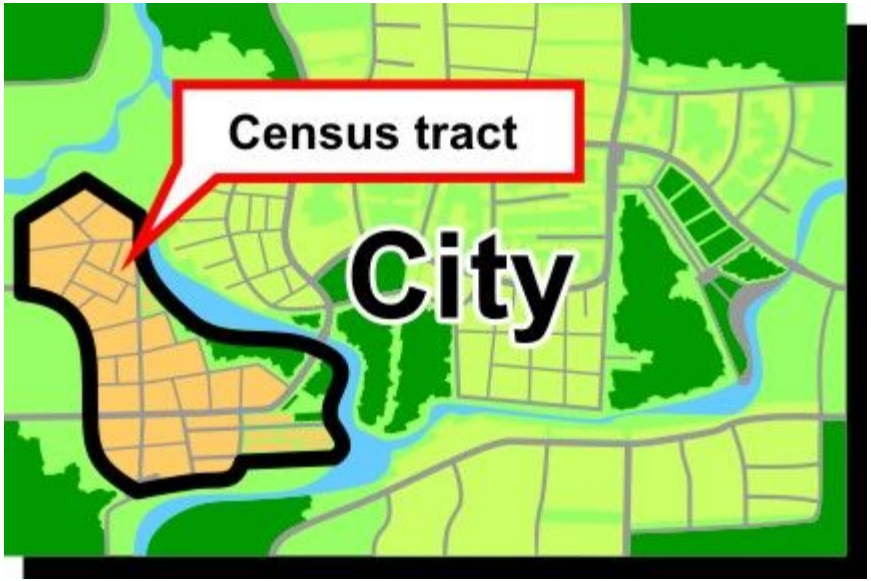
The United States is split into 10 regions
Puerto Rico is part of the dataset and was incorporated in region 2



What are Tracts?

Small subdivisions of county that has a population of 1200 -8000 people, but varies based off of density of settlement

- Exceptions: lower population sizes are known to be tribal areas
- Large population sizes are signs of dense urban areas.



Demographic Features

Population

People of Color Percentage

Low-income percentage

Less than high school Education

Linguistic isolation

Under the age 5

Over the age of 64

Unemployment percentage

Environmental Features

Pre 1960 percentage (Lead indicator)

Diesel particulate matter & pm25

Respiratory Hazard Index

Traffic proximity

Waste water discharge

Proximity to RMP sites

Proximity to NPL sites

Underground Storage

ozone

02

Cleaning

Cleaning

Nulls

Cleared out rows filled with zeros and NaNs/ too many zeros.

Cancer

Dropped missing values in the target cancer column

Waste Water

Contained too many missing values to salvage

Cleaning

Outlier

Dropped an observation with twice the rate of “Cancer Alley”

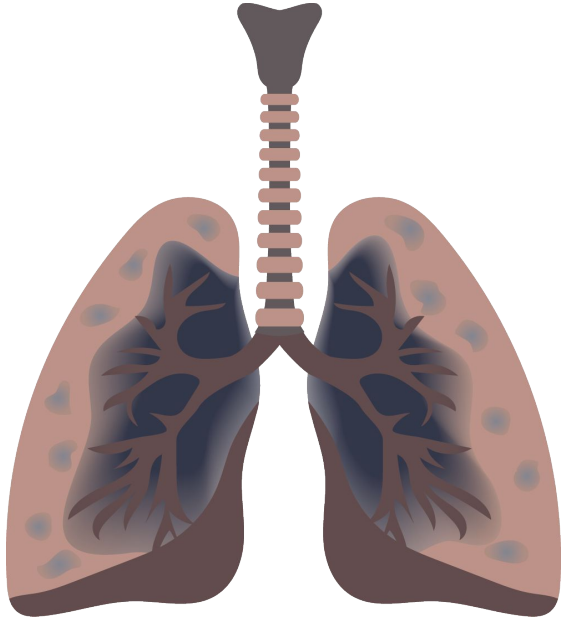
Population

Demographic data primarily %. Dropped population < 30

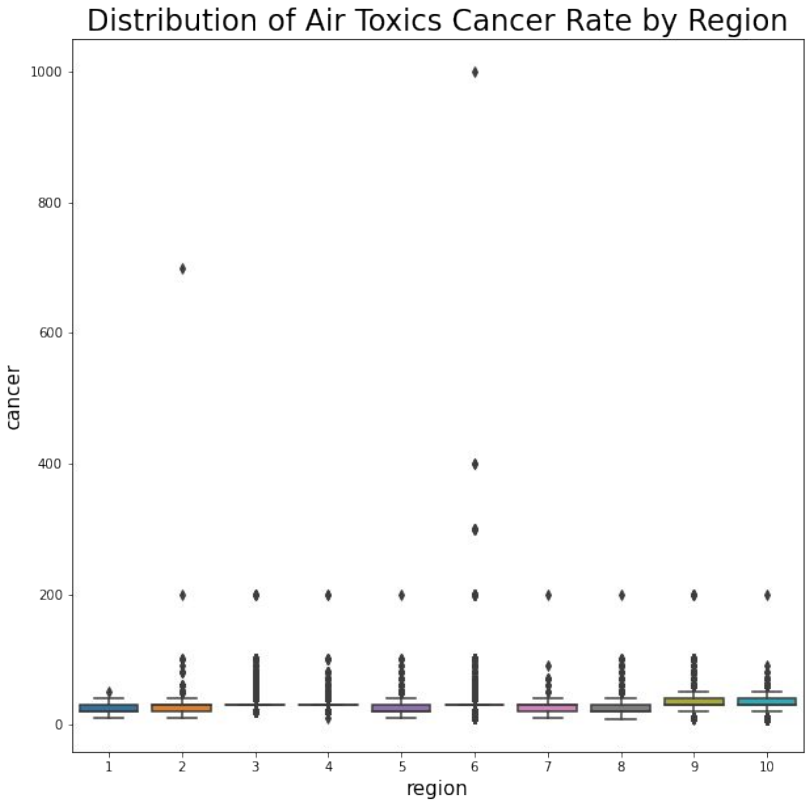
Impute

Imputed traffic proximity, ozone, and particulate matter data

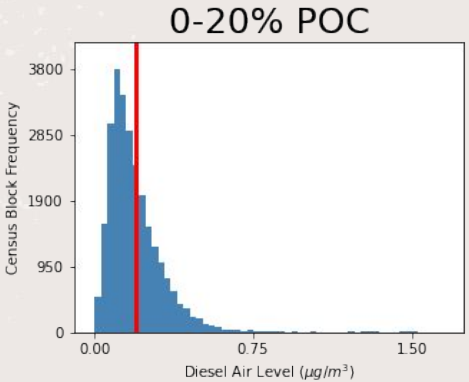
EDA

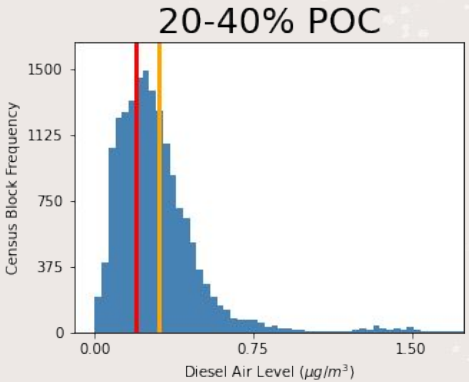


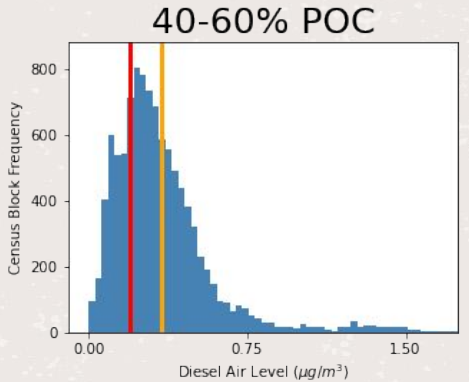
Box Plots By Region

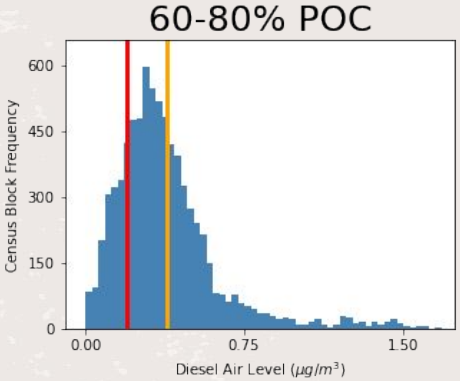


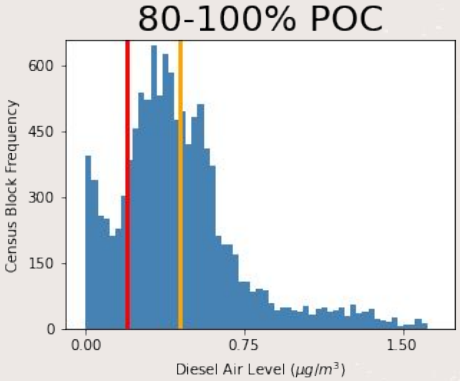
Diesel Air Levels Are Worse In Tracts With More POC







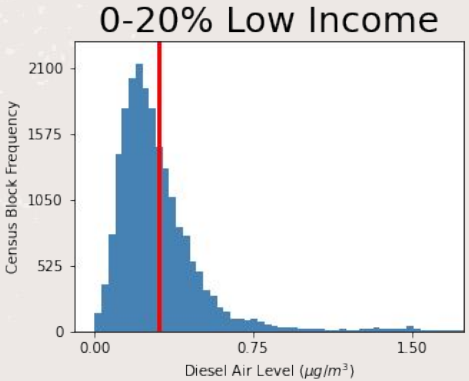


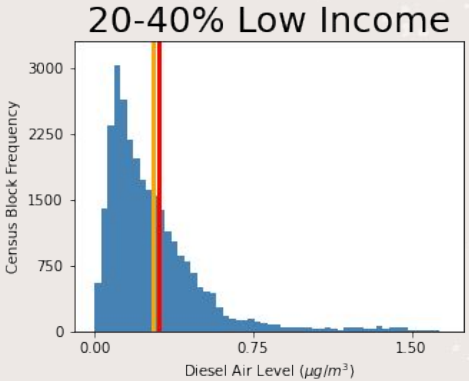


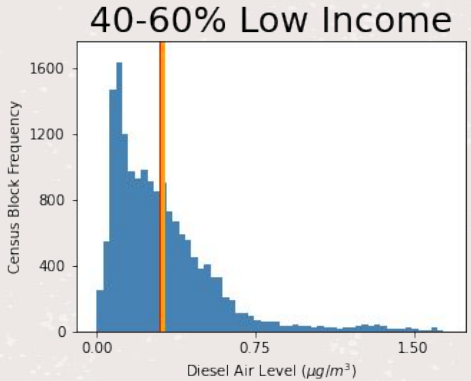
Diesel Mean for 0-20% POC

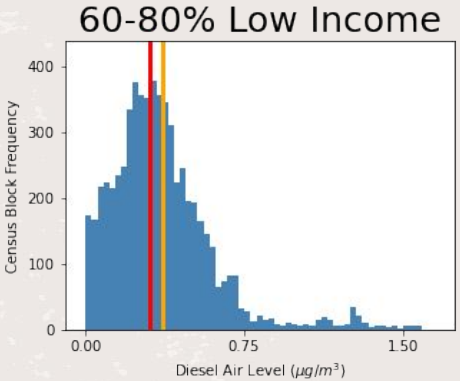
Diesel Mean for Plot

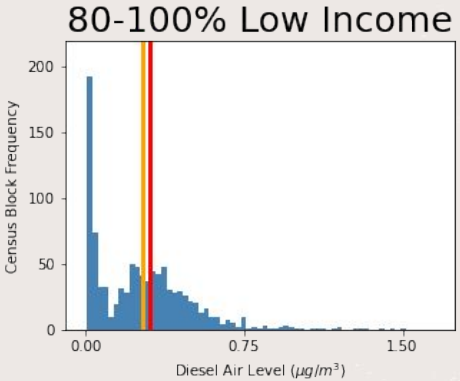
Diesel Exposure Is Not Strongly Related to % of Low Income Residents







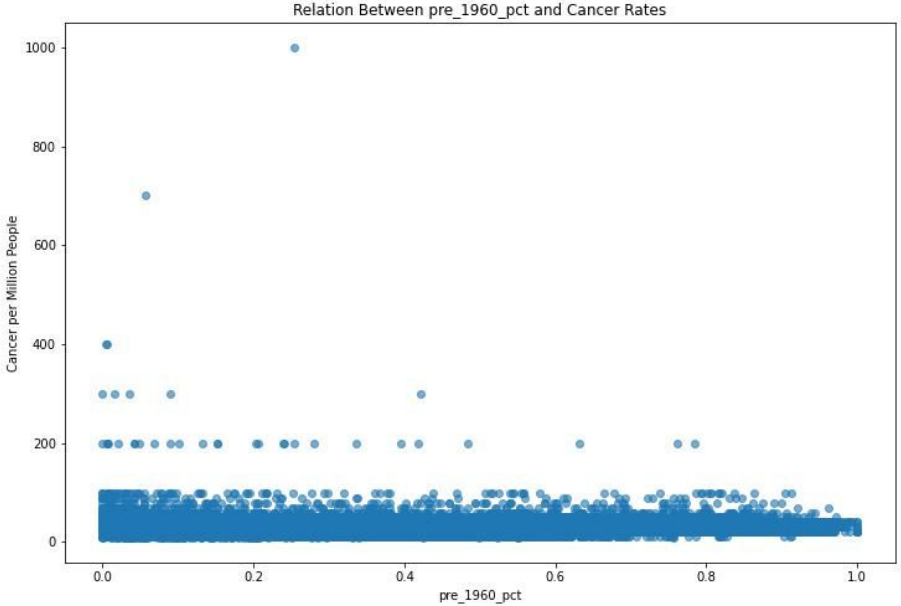




Diesel Mean for 0-20% Low Income

Diesel Mean for Plot

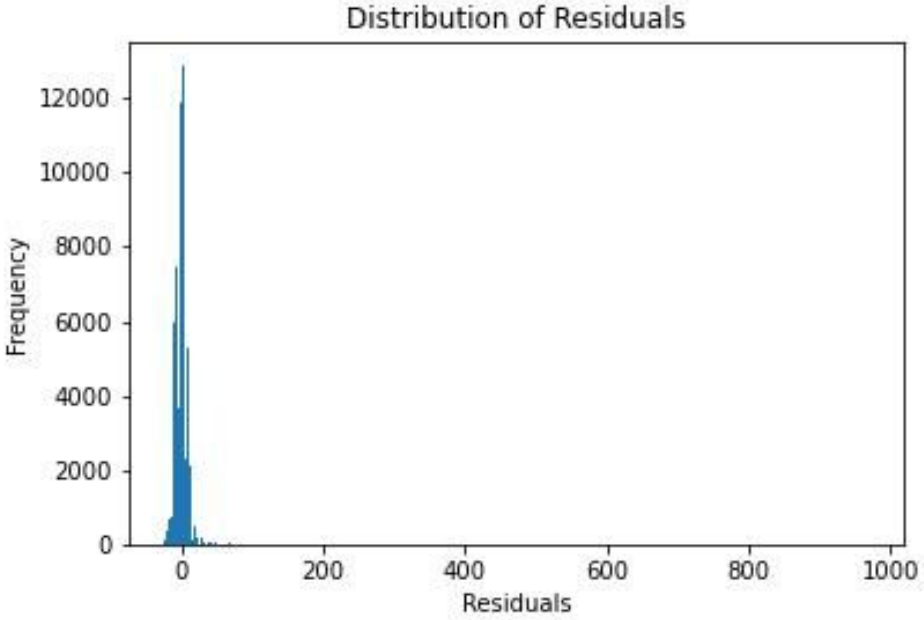
LINE Assumptions



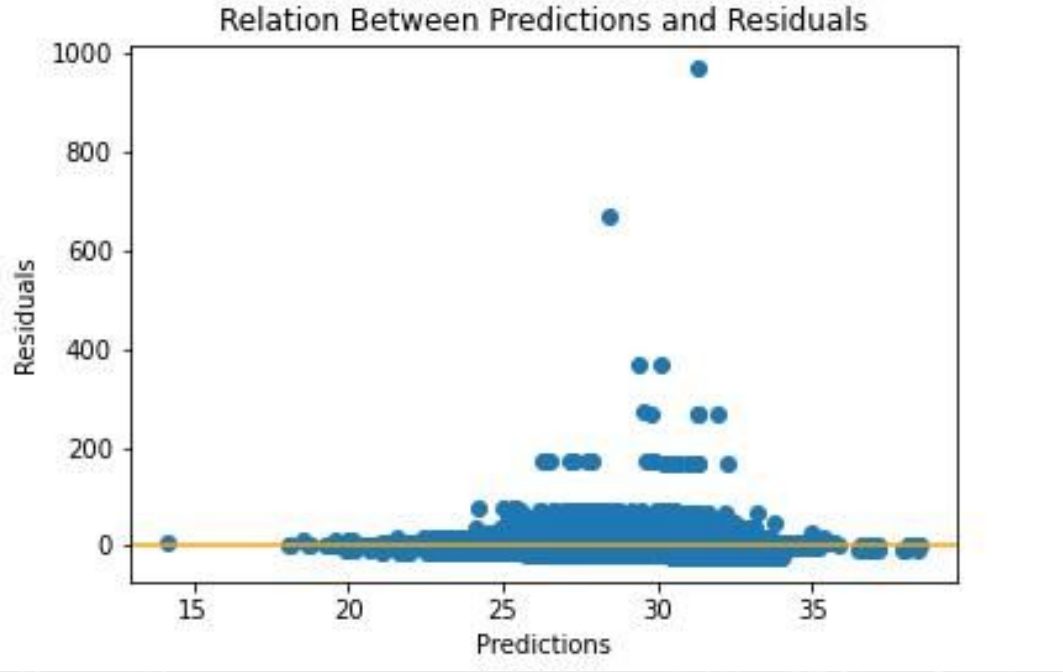
LINE Assumptions

	vif
pm_25	33.513364
ozone	28.157033
resp_index	12.481041
low_inc_pct	10.789563
under_5_pct	7.273332
diesel_pm	7.222390
poc_pct	6.733128
sub_hs_pct	6.265682
over_64_pct	5.431590
pop	5.274813

LINE Assumptions



LINE Assumptions



Modeling

With Demographic Data

Model	Train R^2	Val R^2
Linear Regression	0.707	0.703
Lasso	0.68	0.677
Ridge	0.68	0.675
Elastic Net	0.316	0.311
KNN	0.625	0.375
Decision Tree	0.425	0.255
Bagging	0.918	0.471
Random Forest	0.66	0.356
AdaBoost	0.294	0.133
Gradient Boosting	0.53	0.316

Environmental only

Model	Train R^2	Val R^2
Linear Regression	0.631	0.632
Lasso	0.624	0.623
Ridge	0.629	0.629
Elastic Net	0.317	0.311
KNN	0.625	0.371
Decision Tree	0.459	0.259
Bagging	0.898	0.394
Random Forest	0.533	0.296
AdaBoost	0.013	-0.0001
Gradient Boosting	0.494	0.285

Demonstration of solution

Some Tract Information:

- Region: 5
- Population: 3,511
- POC %: 0.83
- Low Income %: 0.38
- Diesel pm: 0.826
- Traffic Proximity: 3,897
- Waste Proximity: 5
- Ozone: 45

Predicted Cancer Rate:
34.67 cases per million people

True Cancer Rate:
30 cases per million people

Conclusions/ Recommendations

- Linear regression without regularization, predicting the log of cancer rates explained 70% of the variation in log of cancer rates
- Models that included demographic features performed equally well, or more often better, than models with just environmental features
- We recommend that the US government use the linear model to predict air toxics cancer rates to inform allocation of resources when budgeting for Medicare & Medicaid expenses.

Next Steps

- Collect more accurate measures of target construct (ex: lead exposure)
- Separate out percent POC into more specific categories to introduce more nuance into the model

Citations

- Environmental Protection Agency
  - https://www.epa.gov/ejscreen/ej-and-supplemental-indexes-ejscreen
- General Assembly Data Science Immersive Bootcamp Lessons 3.01, 3.02, 4.01, 4.06, 6.01, 6.02, 6.03, 6.04