Leo Shi, Alberto Valencia, Nick Swetlin
Prof. Vishwanath
MATH189
June 9, 2024

Final Report

I. Problem Statement

In our proposal, we hypothesized that the various counties of California (CA) differed significantly in terms of public health indicators (air pollution, water cleanliness, environmental factors, etc.). We originally suspected these indicators of certain counties were correlated with poverty levels, and we sought to investigate whether the connection between poverty and these indicators had significant change across different regions of CA; that is, could we fit a model to predict poverty levels across the entire population of california, or would the differences in regions be severe enough to warrant such a model void?

We took a different direction than the methods listed in our proposal. Upon conducting some exploratory data analysis, we noticed that the "Education" variable ranged from 0 to 100. This variable measures the "percentage of the population over age 25 with less than a high school education (5-year estimate, 2015-2019)." As a result, a larger value in the education column suggests educational achievement is less compared to smaller values. Exploring this variable piqued our interest in examining how much poverty levels explain these fluctuations in education. So compared to our original goal, we pivoted to predicting the education variable using a few covariates, instead of poverty levels, and seeing if this type of model could generalize to census districts across the entire state of California.

Another difference between our actual analysis and our proposal was the specific methods used to investigate our question: we refrained from conducting permutation tests (suggested from our proposal), and we instead attempted to fit a linear regression model, since we believed there was a linear relationship between the most meaningful covariates and education level. All of the regression models we have created fall in one of two categories: the first is performed only on census districts restricted to the regions of San Diego and Los Angeles (SD & LA), and the second is on all census districts.

II. Relevance

Although there was no specific motivation for choosing to measure education levels using other covariates, education inequality (like income inequality), has been an ongoing issue in the United States. These problems seem to be inextricably linked, which made it just as interesting to explore.

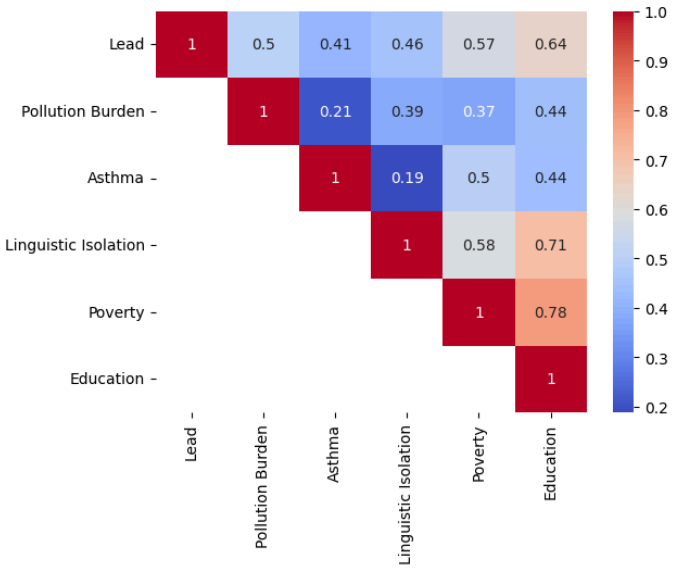
III. Data Sources

We examined data provided by CALEnviroScreen (CES), an online, government-backed tool for evaluating public health by region; the current version, CES 4.0, has a publicly available dataset on its website. This dataset has over 8000 observations (one per census district) and over 50 variables (different public health indicators) describing each observation, updated as recently as 2021. Interestingly, the dataset also has information pertaining to underlying health conditions, which could bring out interesting connections with proper exploration. Given the great number of variables, we observed a subset of the dataframe to begin EDA. We focused on:

- California County
- Education
- Poverty
- Lead
- Pollution Burden
- Asthma
- Linguistic Isolation

IV. Data Explanation

These features were selected based on a preliminary data analysis using a heatmap of the correlation between the covariates and the response variable. We chose these solely based on the correlation coefficient strength between the response and predictor variables.



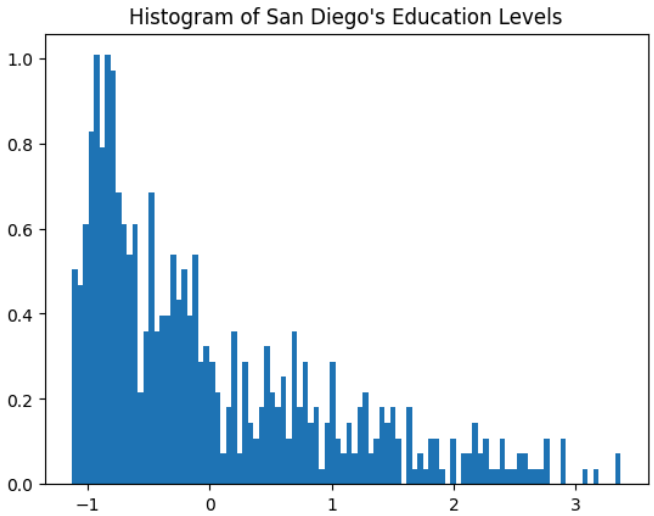
The following are the descriptions for each relevant feature provided by CalEnviroScreen:

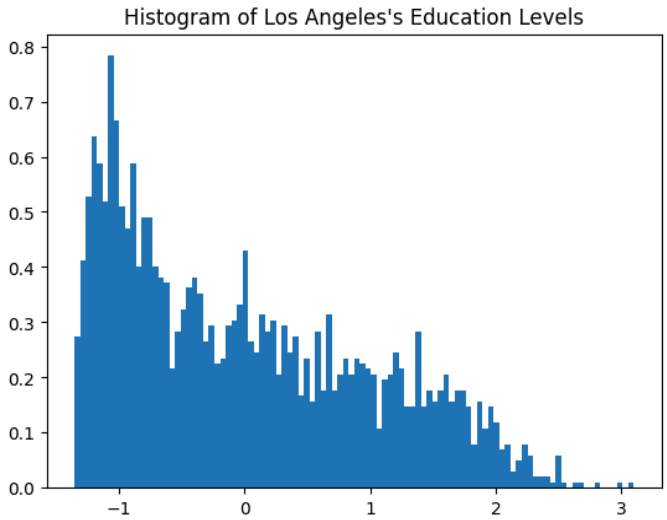
- **California County**: "California county that the census tract falls within."
- **Education**: "Percentage of the population over age 25 with less than a high school education (5-year estimate, 2015-2019)."
- **Poverty**: "Percent of the population living below two times the federal poverty level (5-year estimate, 2015-2019)"

- **Lead**: "Percentage of households within a census tract with likelihood of lead-based paint hazards from the age of housing combined with the percentage of households that are both low-income and have children under 6 years old (5-year estimates 2013-2017)."
- **Pollution Burden**: "Average percentiles of the seven exposures indicators (PM2.5 emission, diesel PM emission, drinking water contamination, lead risk, pesticide use, toxic releases from facilities, and traffic density)."
- **Asthma**: "Spatially modeled, age-adjusted rate of ED visits for asthma per 10,000 (averaged over 2015-2017)."
- **Linguistic Isolation**: "Percentage of limited English-speaking households, (2015-2019)."

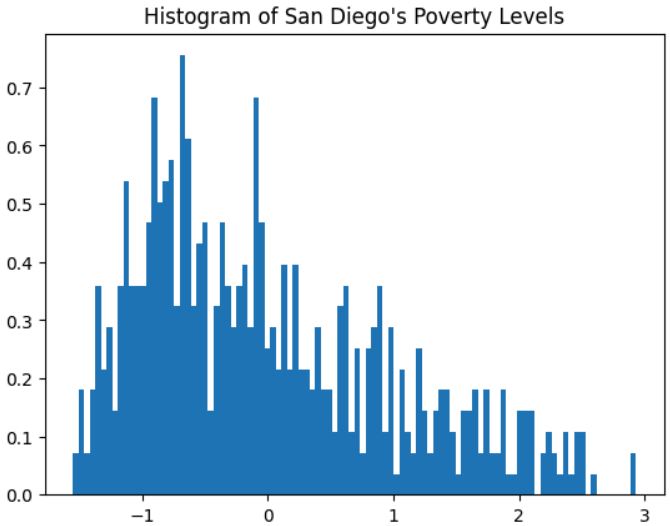
V. Exploratory Data Analysis

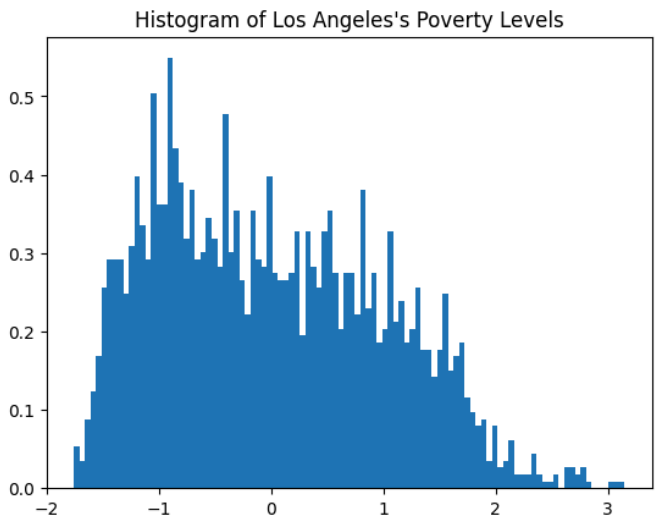
Since the variable we were interested in predicting was education, we generated a qq-plot and histogram to get a sense of the distribution of the educational achievement between counties in San Diego (SD) and Los Angeles (LA).



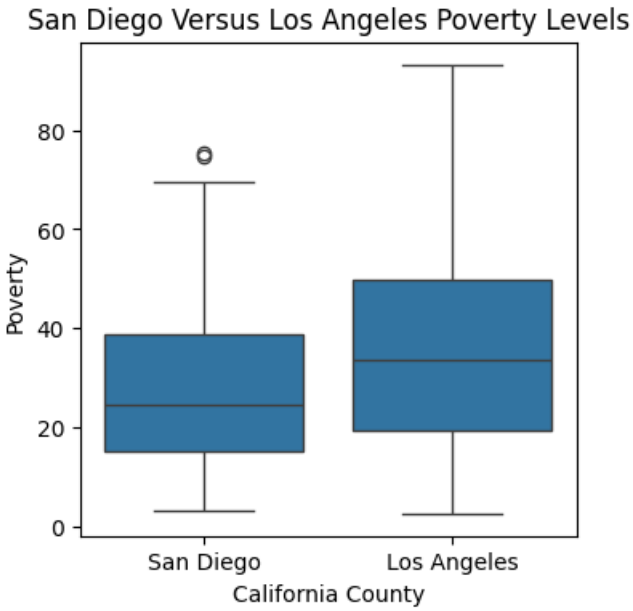


Similarly, we performed the same procedure for the poverty variable of both counties.



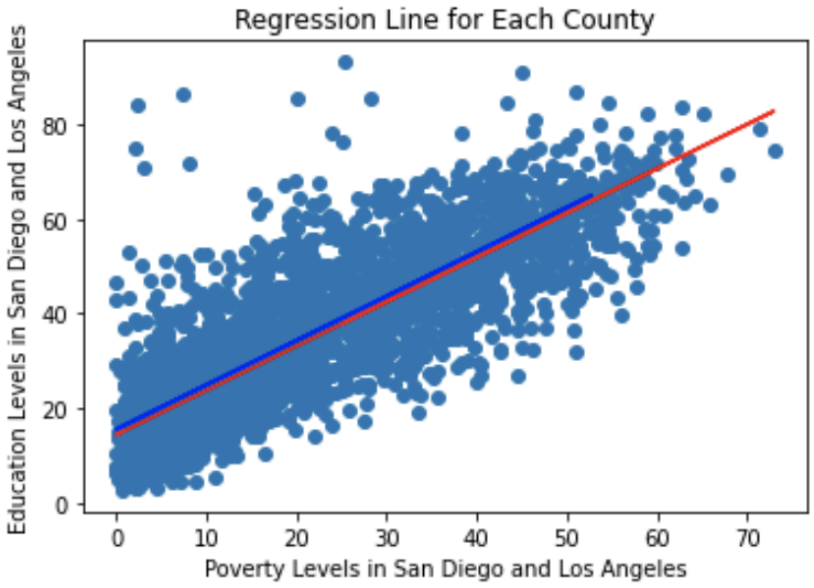


To easily compare the distribution of these variables across the two counties, we plotted a boxplot, which illustrated that LA had higher median levels of poverty for its census districts than SD did.

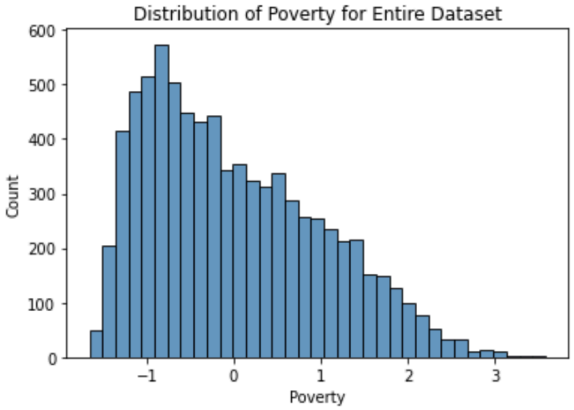


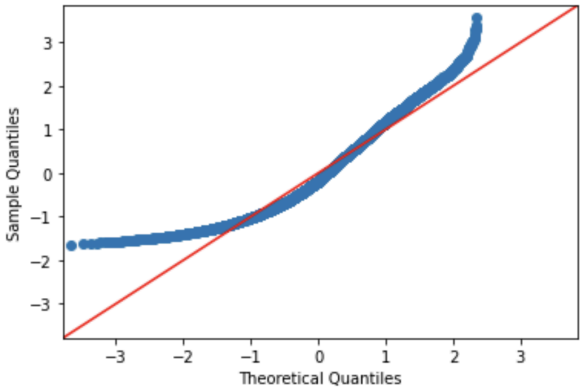
However, we were still unsure that these differences between these two boxplots constituted two different distributions, so we pivoted to using the Shapiro-Wilk test for distributions. For the distributions of poverty of both counties, we failed to reject the null hypothesis at a 0.05 significance level when comparing each to a normal distribution, suggesting that poverty levels in both countries were normally distributed. However, comparing the distributions of poverty across SD and LA under the Wald-Wolfowitz test for distributions yielded a rejection of the null hypothesis at the 0.05 significance level, suggesting that the poverty distributions of LA and SD were not the same. Based on our previous boxplots, we concluded that the distributions of poverty are both potentially normal distributions with different parameters – that is, a different mean and variance.

After investigating education and poverty across census districts only in SD or LA, we decided to explore the entirety of the observations in our dataset. The goal of generating the following visualizations was to get some rough understanding about the entirety of the data we were dealing with, identifying if there are any noticeable differences between analyzing exclusively SD and LA versus all census districts. Given that we wanted to fit a model for regression, we constructed a scatterplot between education and poverty to see if there was a linear relationship. There were many data points, leading to a mildly dense scatterplot, but we could notice a correlation between the two variables; that is, as poverty levels increase, education levels increase. Below, blue represents the line for SD while red represents the line for LA.

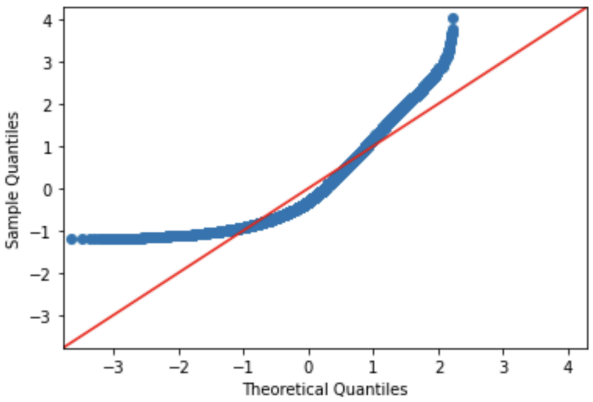


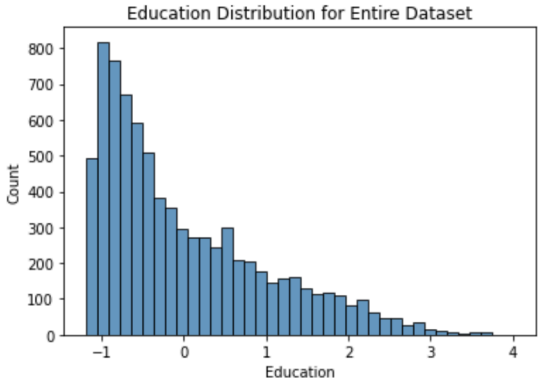
Recall that our education variable measures "Percentage of the population over age 25 with less than a high school education (5-year estimate, 2015-2019)", so this upward trend is indicative that as poverty levels increase, education performance decreases; this quirk of our "education" variable should be considered when interpreting all future predictions of education. To see whether or not education and poverty resembled a normal distribution for the entirety of our dataset, and not just in the census districts of SD & LA, we plotted both a histogram and qq-plot and those respective variables with their values standardized. Here are the figures for poverty:





And below are the figures for education:





Clearly, the distributions of education and poverty across all rows in the dataset were not normal. However, we were still curious if we could manufacture a linear regression model that incorporated these variables.

VI. Data Analysis Comparison

The primary goal of CalEnviroScreen, our data source, was to collect data and develop a screening tool in the form of a choropleth and interactive maps. Descriptions and rationale are proposed for why certain trends appear in their choropleths, such as low birth weight rates appearing in areas with higher rates of air pollutants; this tool helps policy-makers to "identify California communities that are disproportionately burdened by multiple sources of pollution" and make respective decisions. Hence, most of the work they have done are solely data transformations, exploratory data analysis, and choropleth mapping. This can be seen here (August et. al, 2021).

On the other hand, the data analysis and research question we propose are entirely different from the purpose of CalEnviroScreen. We use the data collected through the CalEnviroScreen project to facilitate our mission to discover the strength of connections *between* environmental factors and the education level one may receive in the state of California. However, the result of this project can also be used as a guidance in policymaking, specifically when government agencies are looking into improvements of general education in underserved communities and areas of California. We discuss this more in our limitations section.

VII. Regression Analysis - Districts in SD & LA

Initial Regression Model

As stated in our original proposal, the initial intended research question was to perform permutation tests on different poverty levels of regions in SD & LA to determine whether they were drawn from the same population pool, which would give insight on if our model predictions could or could not generalize to all of California. Despite having drifted away from this idea and instead gravitating

towards predicting education, we continued to explore using this region restriction for our initial regression analysis.

Of the ~2950 census districts that landed in SD & LA, 54 had missing values for education. It should be noted that we decided this was a negligible portion of the dataset and dropped these observations from our model entirely.

Having chosen our covariates by observation and distilling a subset of these covariates with a correlation heatmap (see Data Explanation), we fitted our first regression model on the SD & LA dataset. The model predicted education using the following covariates: poverty, lead, asthma, linguistic isolation, and pollution burden. The resulting model formula had this structure:

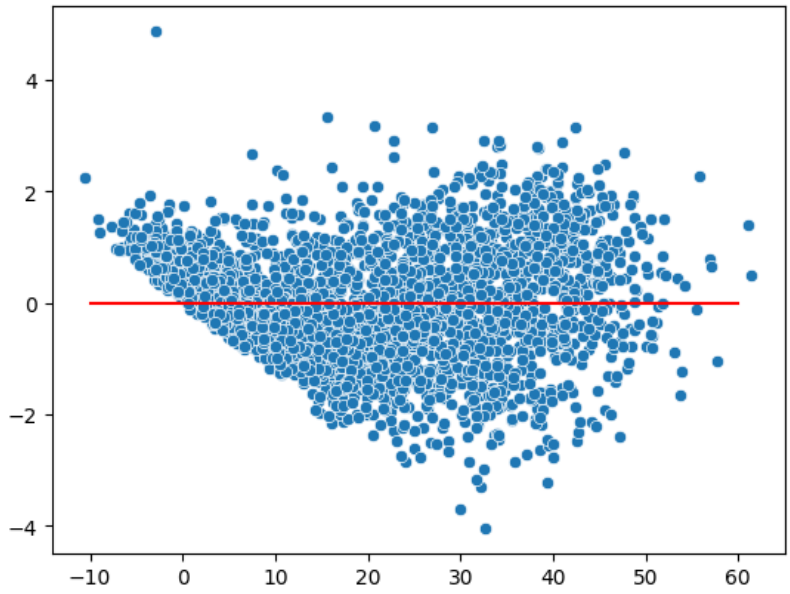
$$\textbf{Education} \sim \beta_0 + \beta_1 \textbf{Lead} + \beta_2 \textbf{Pollution Burden} + \beta_3 \textbf{Asthma} + \beta_4 \textbf{Linguistic Isolation} + \beta_5 \textbf{Poverty}$$

The model summary is as follows:

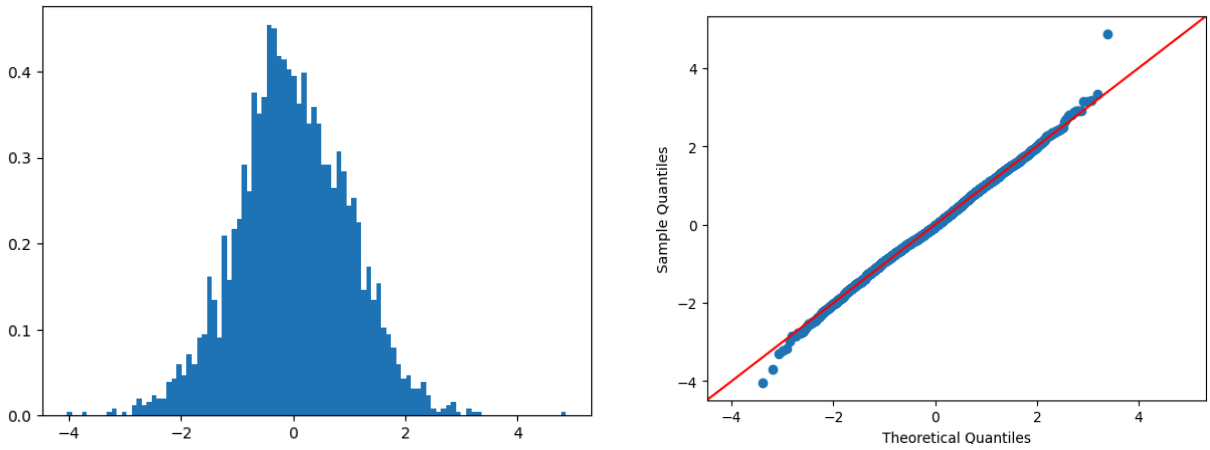
Dep. Variable:	Education	**R-squared:**	0.777
Model:	OLS	**Adj. R-squared:**	0.776
Method:	Least Squares	**F-statistic:**	1976.
Date:	Sun, 09 Jun 2024	**Prob (F-statistic):**	0.00
Time:	19:13:12	**Log-Likelihood:**	-9795.6
No. Observations:	2845	**AIC:**	1.960e+04
Df Residuals:	2839	**BIC:**	1.964e+04
Df Model:	5		
Covariance Type:	nonrobust		

- ACCEPTED MODEL ASSUMPTIONS: Linearity, Residual Independence, Completeness
- VIOLATED MODEL ASSUMPTIONS: Homoscedasticity, Residual Normality

We obtained a model with an R-squared value of 0.777, which, for a baseline model, suggested to us that we explained the variance of education to a reasonable degree. The five assumptions related to the general linear regression model needed to be investigated for conclusions to be drawn: linearity of data, independence of residuals, normality of residuals, heteroscedasticity of residuals, and completeness.



At first glance, the residual plot may seem odd: the residuals are bounded by a line on the bottom-left side. One possible explanation for this is the structure of the raw data we used to obtain this model. Note that some covariates in this dataset are continuous quantitative values ranging from 0 to 100; these values are either scores (like Pollution Burden) or the percentiles for each census district in their respective categories. Since these covariates are never negative, when compared to the negative intercept of our model, the first few residuals would naturally always be above the regression line we predicted, explaining the slanted "boundary" on the lower left side. A second possible explanation is that the distributions of certain variables across SD & LA are, in fact, different, as shown above with boxplots for poverty; this could lead to anomalous residual plot behavior by having a single model needing to account for data from multiple normal distributions with different parameters, resulting in weird fluxes in variance of prediction error. We reasonably conclude that this anomaly in the residual plot does not indicate a violation of the linearity assumption of our data. However, the heteroscedasticity assumption is violated.



An interesting finding came up when inspecting the normality of residuals (shown above). Again, at first glance this residual distribution seems highly normal as it resembles the bell curve quite well, except for a slight right skew at its peak; inspecting the qq-plot provides a similar intuition, with the

addition of mild outliers at each end. Testing the normality of this distribution using a Shapiro-Wilk test revealed a surprising result – the null hypothesis is rejected with a significance level of 0.05 with a p-value of 0.002; our residuals do not follow a normal distribution! After further inspection of independence and completeness of the model, we are confident that no other assumptions are violated. Despite having a non-normal, heteroscedastic residual distribution, the rest of our model's assumptions are still generally well met. Hence, we agreed to salvage this model; we sought out different correction strategies we could apply to our model so that linear regression assumptions could be met and conclusions could be drawn.

Adding Categorical Covariates: "Is_La"

One such attempt at model correction was including categorical covariates that considered differences in the distributions between SD & LA; this would eliminate the issue of having our model being "confused" by the different distributions of SD & LA. We began by introducing a new categorical feature "Is_La", where each row contains the boolean information of whether an observation is from an LA census district or not. We extended our model to include interaction terms between "Is_La" and each of the five aforementioned covariates. The resulting model formula had this structure:

$$\textbf{Education} \sim \beta_0 + \beta_1\,\textbf{Lead} + \beta_2\,\textbf{Pollution Burden} + \beta_3\,\textbf{Asthma} + \beta_4\,\textbf{Linguistic Isolation} + \beta_5\,\textbf{Poverty} + \beta_6\,\textbf{C(Is\_La)} + \beta_7\,\textbf{C(Is\_La):Lead} + \beta_8\,\textbf{C(Is\_La):Pollution Burden} + \beta_9\,\textbf{C(Is\_La):Asthma} + \beta_{10}\,\textbf{C(Is\_La):Linguistic Isolation} + \beta_{11}\,\textbf{C(Is\_La):Poverty}$$

The rationale behind constructing "Is_La" was to make our model behave similarly to having two separate regression planes, one for each of LA and SD. By treating census districts from LA differently from SD, perhaps our model would see a more stable distribution of variance across a more normal distribution of our residuals. Heteroscedasticity and normality would no longer be a thorn in our side! If this new model passed the tests for our regression assumptions, we could verify that the differences in education distributions across SD and LA were mucking up the quality of the model. If this new model did not pass, more investigation of the underlying issues in the data would be needed. Fitting the model produced the following result:

Dep. Variable:	Education	**R-squared:**	0.794
Model:	OLS	**Adj. R-squared:**	0.793
Method:	Least Squares	**F-statistic:**	994.2
Date:	Sun, 09 Jun 2024	**Prob (F-statistic):**	0.00
Time:	19:13:13	**Log-Likelihood:**	-9679.9
No. Observations:	2845	**AIC:**	1.938e+04
Df Residuals:	2833	**BIC:**	1.946e+04
Df Model:	11		
Covariance Type:	nonrobust		

- ACCEPTED MODEL ASSUMPTIONS: Linearity, Residual Independence, Completeness
- VIOLATED MODEL ASSUMPTIONS: Homoscedasticity, Residual Normality

At surface level, we were quite pleased with the mild uptick in the R-squared value (0.794), representing a "better" explanation of the variance in education. This was shortly undermined, however, when assessing regression assumptions for this "new-and-improved" model: nothing had changed. Linearity, independence, and completeness assumptions were still met, yet the heteroscedasticity and normality assumptions were still violated. If adding these additional categorical interaction terms did not affect the outcome of our model assumption tests, then it is clear that accounting for the differences in means and variances between the two different normal distributions of education for SD & LA is not enough, alone, to justify a linear model of our SD & LA dataset.

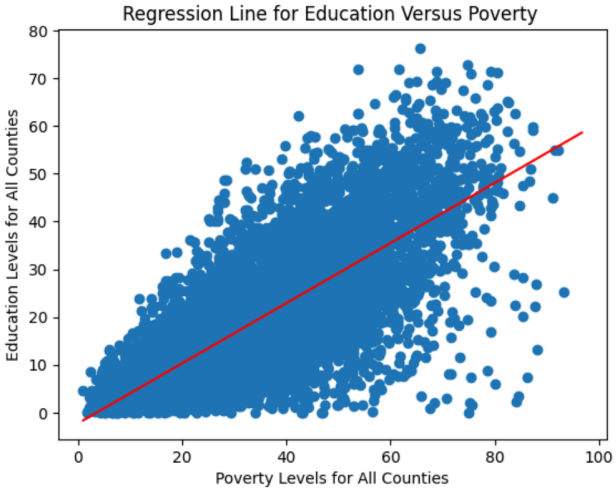
Clearly, there were still underlying issues with the data, but we just did not know where to look; we had inklings that either: certain variables were mucking up our model, or that the action of taking a subset of rows from our original dataset was negatively influencing our clarity of these distributions. We were leaning towards the latter. Could it be that fitting a model on only census districts from SD & LA, rather than all California census districts, lead to anomalous patterns of our residuals? What if we zoomed out and fit a regression model on *all* census districts in California?

VIII. Regression Analysis - All California Districts

Simple Regression Model

We needed a fresh start, so we looked at the whole dataset; we had a lingering suspicion that taking a subset of our data led to a fault in our model quality.

For the model containing all of California's census districts, we started off simple: poverty predicts education. A scatterplot between education and poverty and a line of best fit showed that poverty and education had strong correlation, which was more than enough justification to make this connection:



Again, note that education here is actually thought of as "lack of education", given the CES education variable description. Tinkering with the following model would illuminate if taking a subset of rows influenced our results.

$$\textbf{Education} \sim \beta_0 + \beta_1 \textbf{Poverty}$$

This model is quite simple and only somewhat explains the variance in education (R-squared equals 0.61), yet it served as a baseline to get our work started. We now knew that zooming out to all rows alone was not enough to rectify issues in our residual plots, nor issues with our regression assumptions. A fanning effect is still seen in the residual plot, indicating that perhaps choice of poverty as a variable is affecting our predictions.

Applying SD & LA Regression Model

As a curious comparison, we also tested our previous five covariate model from the SD & LA dataset on the entire dataset. As a reminder, the formula for this model is given by:

$$\textbf{Education} \sim \beta_0 + \beta_1 \textbf{Lead} + \beta_2 \textbf{Pollution Burden} + \beta_3 \textbf{Asthma} + \beta_4 \textbf{Linguistic Isolation} + \beta_5 \textbf{Poverty}$$

Below, we show the model summary:

Dep. Variable:	Education	**R-squared:**	0.751
Model:	OLS	**Adj. R-squared:**	0.751
Method:	Least Squares	**F-statistic:**	4626.
Date:	Sun, 09 Jun 2024	**Prob (F-statistic):**	0.00
Time:	19:13:19	**Log-Likelihood:**	-26177.
No. Observations:	7672	**AIC:**	5.237e+04
Df Residuals:	7666	**BIC:**	5.241e+04
Df Model:	5		
Covariance Type:	nonrobust		

- ACCEPTED MODEL ASSUMPTIONS: Linearity, Residual Independence, Completeness
- VIOLATED MODEL ASSUMPTIONS: Homoscedasticity, Residual Normality

Fair evidence suggests that though our model does an adequate job at predicting education at the surface level, it performs worse overall on the greater dataset than the SD & LA data subset; regression assumptions have not changed, which gives complete confirmation that stratifying data by county was not the main issue. Perhaps it was variable selection that was the issue; at this point, we realized "pollution burden" accounted for a combination of other variables in our dataset.

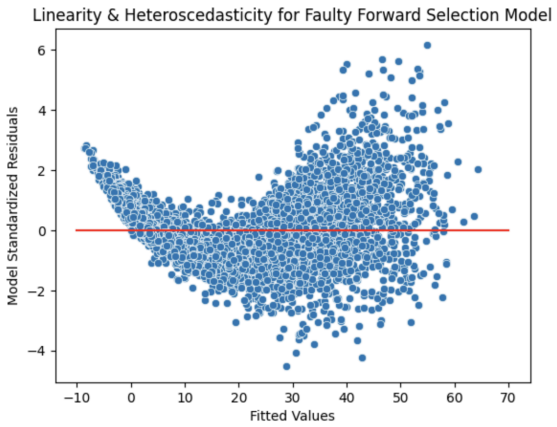
However, it is worth noting that from the above summary table it is seen that the p-value for all of these covariates is less than an alpha level of 0.05; in fact, they are barely non-zero. We should clarify: the relevance of discussing the p-value here is not to brashly claim statistical significance of a linear relationship between education and poverty. The true relevance of discussing the p-value here

is that there is evidence that some sort of connection between poverty and education exists, be it linear or nonlinear. So, despite apparent anomalies in the data, we were hesitant to let poverty go (and other variables for that matter) as a predictor of education because of the sheer proximity it has to education; perhaps some sort of feature transformation could be constructed that finally outed the underlying connection, and we would have a successful linear model. Still, before we could consider transformations, we needed to pin down exactly which other variables were prime contenders to include in the next version of our model. This led us to consider advanced variable selection techniques.

Forward Selection Modeling

In further exploration of our data, we sought to explore if we could create a better performing model using variable selection techniques. We used forward stepwise selection, as defined in class. The results yielded the following model summary:

Dep. Variable:	Education	**R-squared:**	0.945
Model:	OLS	**Adj. R-squared:**	0.944
Method:	Least Squares	**F-statistic:**	1302.
Date:	Sun, 09 Jun 2024	**Prob (F-statistic):**	0.00
Time:	13:08:37	**Log-Likelihood:**	-19614.
No. Observations:	7356	**AIC:**	3.942e+04
Df Residuals:	7260	**BIC:**	4.008e+04
Df Model:	95		
Covariance Type:	nonrobust		



Compared to our baseline, this model had a much better performance with respect to the R-squared and adjusted R-squared metrics, almost suspiciously well. With the exception of two categorical variables, all of the variables were technically statistically significant indicators of predicting education. However, an issue was that we fed forward selection dangerous variables: variables such as “CES_40_Percentile_Range” and “CES_40_Score”, which assign an overall score based on the other variables, namely our response variable of education (multicollinearity); we subsequently removed these since they carried implicit information about other variables. See the above residual plot for visual proof of these odd trends. We also removed percentile versions of other covariates, since we found them redundant. Again, we performed forward selection, but without these pesky variables. Examining the model summary below, we can see that exclusion of those variables significantly impacted model performance. Nonetheless, we see that the model does fit the data better than our baseline based on the R-squared metric of 0.803:

OLS Regression Results

Dep. Variable:	Education	**R-squared:**	0.803
Model:	OLS	**Adj. R-squared:**	0.801
Method:	Least Squares	**F-statistic:**	419.2
Date:	Sun, 09 Jun 2024	**Prob (F-statistic):**	0.00
Time:	12:55:02	**Log-Likelihood:**	-24267.
No. Observations:	7355	**AIC:**	4.868e+04
Df Residuals:	7283	**BIC:**	4.918e+04
Df Model:	71		
Covariance Type:	nonrobust		

- ACCEPTED MODEL ASSUMPTIONS: Linearity, Residual Independence, Completeness
- VIOLATED MODEL ASSUMPTIONS: Homoscedasticity, Residual Normality

Ultimately, forward selection's choice of variables performed no better at satisfying regression assumptions than other models, so conclusions can still not be drawn from such a model – though, this was not the purpose of forward selection. The purpose of using forward selection was to identify candidate variables for prediction; of the 28 variables fed to forward selection, 8 were excluded from the model, meaning we should likely exclude them from our model. These eight variables were:

- "Approximate_Location"
- "Asthma"
- "Cleanup_Sites"
- "Diesel_PM"
- "Education" (force-removed from forward selection)
- "Imp_Water_Bodies"
- "PM25"
- "ZIP"

LASSO and Elastic Net Regression

In an attempt to exhaust all of our tools for variable selection, we performed both lasso and elastic net regression. We performed lasso regression on all variables (excluding "dangerous" variables, such as 'CES_40_Score', 'Pollution Burden', and percentile variables). Using the variables selected by LASSO, we created a new model formula and fit our data on this formula. We repeated this exact procedure using Elastic Net regression, with a weight of 0.5. We compared the model formulas returned by LASSO regression and Elastic Net regression against our original for the sake of being thorough; of course, we also conducted regression model assumptions for each. For both models, we noticed that while the assumptions were unsatisfied, they were not severely violated. Below are the result from the LASSO regression model and the Elastic Net regression model, respectively:

Dep. Variable:	Education	**R-squared:**	0.754
Model:	OLS	**Adj. R-squared:**	0.754
Method:	Least Squares	**F-statistic:**	2939.
Date:	Sun, 09 Jun 2024	**Prob (F-statistic):**	0.00
Time:	13:51:13	**Log-Likelihood:**	-26129.
No. Observations:	7672	**AIC:**	5.228e+04
Df Residuals:	7663	**BIC:**	5.234e+04
Df Model:	8		
Covariance Type:	nonrobust		

Dep. Variable:	Education	**R-squared:**	0.772
Model:	OLS	**Adj. R-squared:**	0.771
Method:	Least Squares	**F-statistic:**	2554.
Date:	Sun, 09 Jun 2024	**Prob (F-statistic):**	0.00
Time:	20:44:55	**Log-Likelihood:**	-25530.
No. Observations:	7573	**AIC:**	5.108e+04
Df Residuals:	7562	**BIC:**	5.116e+04
Df Model:	10		
Covariance Type:	nonrobust		

- ACCEPTED MODEL ASSUMPTIONS: Linearity, Residual independence, Completeness
- VIOLATED MODEL ASSUMPTIONS: Homoscedasticity, Residual Normality

LASSO and Elastic Net produced nearly identical lists after their first run each, which, for simplicity, are the runs we are considering for the remainder of this report. Of the LASSO and Elastic Net Models, seven variables were found in common: Hazardous Waste, PM25, Latitude, Linguistic Isolation, Poverty, Asthma, Lead, and Longitude. Interestingly, this explains why the R-squared and adjusted R-squared metrics of the two models are similar in terms of performance.

The "Smart" Model

Given that LASSO and Elastic Net models involve penalty terms, they shrink the complexity of models and provide only the most "important" variables, making both techniques fitting for variable selection. Of the LASSO and Elastic Net Models produced, seven variables were found in common; we took these seven variables and constructed a linear model from them to see how much our model had either improved or stayed the same. Ironically, forward selection disagreed with choosing two of these variables (PM25, Asthma), but we took the advice of forward selection with a grain of salt given it is liable not to find the true optimal set of covariates. The formula for this model was given by:

$$\textbf{Education} \sim \beta_0 + \beta_1\,\textbf{Hazardous Waste} + \beta_2\,\textbf{PM25} + \beta_3\,\textbf{Asthma} + \beta_4\,\textbf{Linguistic Isolation} + \beta_5\,\textbf{Poverty} + \beta_6\,\textbf{Latitude} + \beta_7\,\textbf{Longitude} + \beta_8\,\textbf{Lead}$$

The model summary was, surprisingly, as follows:

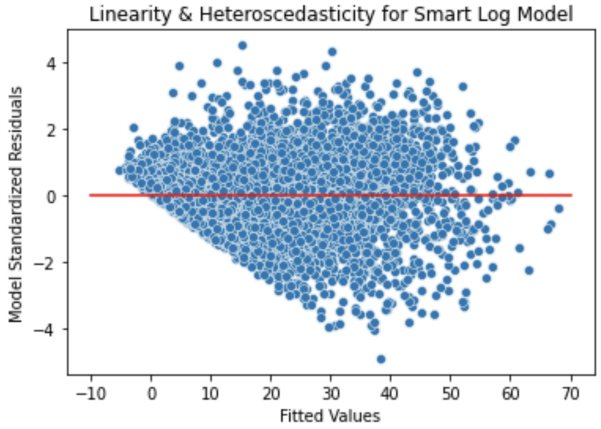
OLS Regression Results

Dep. Variable:	Education	**R-squared:**	0.749
Model:	OLS	**Adj. R-squared:**	0.749
Method:	Least Squares	**F-statistic:**	3267.
Date:	Sat, 08 Jun 2024	**Prob (F-statistic):**	0.00
Time:	20:48:41	**Log-Likelihood:**	-26209.
No. Observations:	7672	**AIC:**	5.243e+04
Df Residuals:	7664	**BIC:**	5.249e+04
Df Model:	7		
Covariance Type:	nonrobust		

- ACCEPTED MODEL ASSUMPTIONS: Linearity, Residual Independence, Completeness
- VIOLATED MODEL ASSUMPTIONS: Homoscedasticity, Residual Normality

Surprisingly, the R-squared metric (0.749) was actually worse off than that of our SD & LA model (0.751). Although it is expected that the inclusion of more covariates would explain the variance in education better, we believed our methods of variable selection would ultimately find an optimal "true" model. So, it was surprising when we saw that these variable selection techniques resulted in a model that had a similar quality to our previous model at best. We also expected that the outcomes of model assumptions would not change given we had not conducted feature transformation yet.

The first form of feature transformation we attempted was taking the log of our response variable. Perhaps the percentage of the population over age 25 with less than a high school education may not further increase after a certain degree, given the need for a district to sustain itself with a sizable working population with a reasonable education. If this is the case, then taking the log of our response variable, education, would be the proper way to transform our data. Sadly, this was not the case, as reflected by the subsequent residual plot:

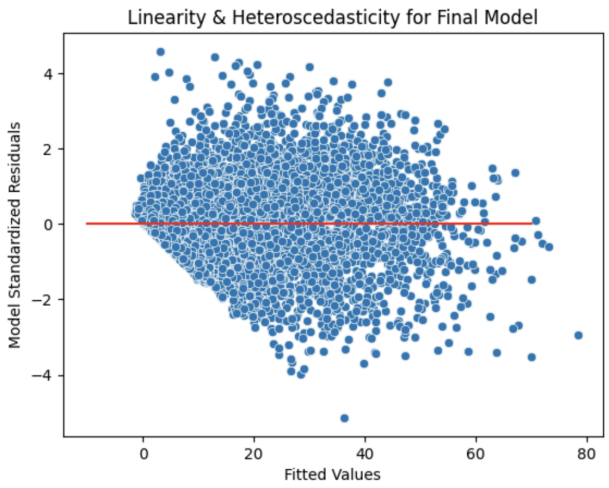


Final Model

Another form of feature engineering we included, which led to mildly successful results was the inclusion of all possible binomial interaction terms, producible from the seven variables of our smart model. Doing so allowed us to explain an additional reasonable portion of the variance of "education" with our model:

Dep. Variable:	Education	**R-squared:**	0.796
Model:	OLS	**Adj. R-squared:**	0.795
Method:	Least Squares	**F-statistic:**	1063.
Date:	Sat, 08 Jun 2024	**Prob (F-statistic):**	0.00
Time:	20:49:00	**Log-Likelihood:**	-25319.
No. Observations:	7646	**AIC:**	5.070e+04
Df Residuals:	7617	**BIC:**	5.090e+04
Df Model:	28		
Covariance Type:	nonrobust		

- ACCEPTED MODEL ASSUMPTIONS: Linearity, Residual Independence, Completeness
- VIOLATED MODEL ASSUMPTIONS: Homoscedasticity, Residual Normality



Of note here is, again, the uptick in the R-squared metric (0.796) compared to our SD & LA model (0.751). Arguably more important, however, was that the residual plot for our final model had softened the steepness of that bottom-left "boundary line"; our residual data, though still heteroscedastic, was less heteroscedastic under this model compared to other models! Clearly, the inclusion of more interaction terms will be helpful in future modeling of this dataset.

IX. Model & Coefficient Interpretation

There are two models whose coefficients we find worthwhile discussing, since they most closely relate to our project goal.

San Diego & Los Angeles Interaction Terms Model

Modeling education from other public health indicators for specifically the regions of SD & LA does not meet assumptions for linear regression. It is implied that further exploration and feature transformation needs to be done to construct such a linear model, or that there is fair evidence that one cannot construct a sound, interpretable linear regression model from given data. Our research

question is answered; if a linear model could not be reasonably applied to the regions of SD & LA alone, then such a model could not generalize to the whole of California. That being said, it is worth investigating what interpretations of the coefficients of the below model could have been, given that the assumptions for linear regression were met:

$$\textbf{Education} \sim \beta_0 + \beta_1 \textbf{ Lead} + \beta_2 \textbf{ Pollution Burden} + \beta_3 \textbf{ Asthma} + \beta_4 \textbf{ Linguistic Isolation} + \beta_5 \textbf{ Poverty} + \beta_6 \textbf{C(Is\_La)} + \beta_7 \textbf{ C(Is\_La):Lead} + \beta_8 \textbf{ C(Is\_La):Pollution Burden} + \beta_9 \textbf{ C(Is\_La):Asthma} + \beta_{10} \textbf{ C(Is\_La):Linguistic Isolation} + \beta_{11} \textbf{ C(Is\_La):Poverty}$$

Had the assumptions been met, our model would have been interpretable as, for census districts in SD & LA:

- β_0: …the default percentage of the population over the age of 25 with less than a high school degree is given that all other covariates are zero is -2.26%.
- β_1: …given all other covariates are held still, the percentage of the population over the age of 25 with less than a high school degree increases by 0.1041% for each unit increase in the "Lead" variable (the percentage of households within a census tract with likelihood of lead-based paint hazards from the age of housing combined with the percentage of households that are both low-income and have children under 6 years old).
- β_2: …given all other covariates are held still, the percentage of the population over the age of 25 with less than a high school degree decreases by 0.069% for each unit increase in the "Pollution Burden" Variable (average percentiles of the seven exposures indicators: PM2.5 emission, diesel PM emission, drinking water contamination, lead risk, pesticide use, toxic releases from facilities, and traffic density). Note that the behavior of this variable is uncertain given its high p-value of 0.048 and liability to introduce multicollinearity.
- β_3: …given all other covariates are held still, for census districts in SD & LA, the percentage of the population over the age of 25 with less than a high school degree increases by 0.075% for each unit increase in the "Asthma" Variable (spatially modeled, age-adjusted rate of ED visits for asthma per 10,000, averaged over 2015-2017).
- β_4: …given all other covariates are held still, for census districts in SD & LA, the percentage of the population over the age of 25 with less than a high school degree increases by 0.781% for each unit increase in the "Linguistic Isolation" Variable (Percentage of limited English-speaking households, 2015-2019).
- β_5: …given all other covariates are held still, for census districts in SD & LA, the percentage of the population over the age of 25 with less than a high school degree increases by 0.212% for each unit increase in the "Poverty" Variable average percentiles of the seven exposures indicators (PM2.5 emission, diesel PM emission, drinking water contamination, lead risk, pesticide use, toxic releases from facilities, and traffic density).
- β_6: …the additional percentage of the population over the age of 25 with less than a high school degree is given that all other covariates are zero, exclusively if a census district falls within the county of LA, is -15.25%.

- β_7 through β_{11} : the additional change in percentage of the population over the age of 25 with less than a high school degree seen with respect to the variables “Lead”, “Pollution Burden”, “Asthma”, “Linguistic Isolation” and “Poverty”, exclusively if a census district falls within the county of LA, are, respectively; 0.159%, 0.114%, -0.02%, -0.46%, and 0.18%.

Final Model From Entire Dataset

Our final linear model also did not meet regression assumptions. As stated before, it is implied that further exploration and feature transformation needs to be done to construct such a linear model, or that there is fair evidence that one cannot construct a sound, interpretable linear regression model from given data. For the sake of space, the final model summary table is included in our Jupyter Notebook.

Had the assumptions been met, our model would have been interpretable as, for census districts in all of California:

- Intercept: …the default percentage of the population over the age of 25 with less than a high school degree is given that all other covariates are zero is -1292%.
- Standalone Term Coefficients: …given that all other covariates are held still, and that all other covariates are zero, the default percentage of the population over the age of 25 with less than a high school degree is changed by the corresponding amount in the table.
- Interaction Term Coefficients: … the rate of change for the default percentage of the population over the age of 25 with less than a high school degree with respect to a particular covariate is changed by the corresponding amount in the table.

X. Limitations & Discussion

Interactive choropleths for the SD & LA Model, Smart Model, Final Model can be viewed here.

Prior to beginning our work on predicting education, there were some missing values in both the education and poverty columns. For this reason, we dropped those observations, since the dataset was relatively large. Given the size of the dataset, dropping the rows was appropriate because the rows that were dropped did not constitute a large portion of the data. Having had access to more rows does not necessarily mean our models would have changed significantly, but this could influence model results and lead to inaccuracies in interpretation if there are large enough anomalies present in missing data – especially if missing values are not missing at random.

A technical note: depending on the machine we ran our notebook on – specifically for the faulty forward selection – that we would often get the error message: “SVD did not converge”. In particular, we notice this on datahub.

As mentioned in the regression analysis, each of our models failed to satisfy the necessary assumptions. In particular, both the heteroscedasticity and normality assumptions were violated, making it difficult to conclude that there exists a linear relationship in the data. Accordingly, we answer our research question by stating that we cannot confidently conclude that our model

accurately represents the relationship between the covariates and education, and therefore that a linear model may be insufficient in generalizing the prediction of education for across census districts in California. We also cannot conclude that the relationship between these variables is causal. Since this dataset has use in sparking policy decisions, our model could not be used in practice. Further work must be conducted exploring why that is, which in turn may help spark conversations surrounding policy-making.

References

- August, Laura, et al. Edited by Vince Cogliano et al., *CalEnviroScreen 4.0*, California Office of Environmental Health Hazard Assessment (OEHHA), Oct. 2021, oehha.ca.gov/media/downloads/calenviroscreen/report/calenviroscreen40reportf2021.pdf.

For references used in our coding, see our FinalProject.ipynb.