

# **Lead in the Water: The Effects of Blood Lead Levels on Incarceration Rates**

## **A Socio-Spatial Analysis**

staRstistics - Will Lieber, Wania Iftikhar Khan, AJ Tenser, Kami Akala

2025-04-30

### **Introduction**

Could exposure to lead increase one's likelihood of going to prison? After our team's systematic review of studies, we have explored the potential effects of lead exposure on brain development in children and adults. Various existing studies highlight the detrimental effects of lead on different brain regions, noticeable in a decrease in executive control and cognitive control, thereby affecting memory, mood, behavior and comprehension skills. Such exposure to lead during the developmental years of children causes irreversible damage, the effects of which can be seen later on in life.

Studies in the past, such as one conducted by Talayero et al. (2023)<sup>1</sup>, have highlighted a strong association between lead exposure during childhood and criminal tendencies during adulthood. One can be exposed to lead through different means, including water, which is the medium which we've chosen to investigate. Our research topic inquires about whether a relationship exists between a specified area's water lead levels and its incarceration rates, while also considering potential confounding effects of other demographic factors.

This research topic has important societal implications, namely the complicated intersection of crime, environmental racism, and more. It's an ever relevant question today and we hope to come to meaningful conclusions by the end of our analysis. Our initial hypothesis is that there is a positive relationship between water lead levels and the rate of incarceration with the existence of other interaction effects from things such as race and income. However, we acknowledge the intricate combination of social and institutional factors that increases one's likelihood of incarceration and understand the possibility for inconclusive findings with the focus on lead exposure.

1. [The association between lead exposure and crime: A systematic review](#)

## Our Data

We've chosen to create our dataframe from a variety of census data relating to California in 2020. Our data looks at different California census tracts and their respective statistics relating to blood lead levels, income, incarceration rates, and racial demographics. For our analysis, we are particularly focused on perc\_bll\_indicator, med\_income, our age and race variables and how well they can predict imprisonment\_rt. The demographic data was collected through the Census data collection process which involves online surveys, in-person questionnaires, and is self-reported. The blood lead level data was compiled using the mandatory reported data entered by laboratories and healthcare workers. The data was sorted into census tracts based on reported street address and observations deemed false positive or false negative by the California Department of Public Health (CDPH) have been omitted. To clean the data, we joined each dataset to each other based on census tract and renamed columns for clarity. In addition, we calculated new columns to simplify categories (i.e. age ranges, ethnicity) and omitted fields which would not be relevant for this investigation.

**Note:** All figures referenced throughout the report can be found in the Appendix section below.

## Univariate Data Exploration

**Blood Lead Levels (Figure 1):** There are 94 census tracts that didn't test blood lead levels in this data. These will likely need to be removed because these observations are not useful in the analysis. Additionally, there is a large concentration of census tracts that have a percent blood level indicator (bll) of 0 (1899 observations). While these should probably be included in the final analysis, removing these for data visualization produces a unimodal distribution that is heavily right skewed. Overall, the median blood level is 1.6891892 and the IQR is NA.

**Incarceration Rate (Figure 2):** After removing extreme outliers (the top 1% incarceration rates - some may ultimately be removed because the census tract has an extremely low population ex. ~3 people), the shape of the distribution is unimodal and right skewed. The median incarceration rate is 267 out of 100,000 with an IQR of 257. This doesn't appear surprising - there are fewer census tracts with particularly high imprisonment rates.

**Income (Figure 3):** The shape of the income distribution is also unimodal with a less extreme right skew and a median value of  $7.7225 \times 10^4$  and iqr of  $5.044 \times 10^4$ . This doesn't appear surprising - we'd expect median incomes of tracts to be concentrated towards the left.

**Spatial Distribution (Figure 4):** Given the spatial nature of this investigation, it is important to understand how the reported lead levels vary across the state. This map shows that there are contiguous series of tracts, especially in the Los Angeles region, where there may be spatial correlation. This is not surprising, as we expect water conditions across tracts to be similar.

## Bivariate Data Exploration

**Imprisonment vs BLL Indicator (Figure 5):** Imprisonment rate and the percentage of a census tract with a high bll has a weak positive correlation - at first glance, the bll indicator doesn't seem to have significant explanatory power for the variability in a census tract's imprisonment rate.

**Imprisonment vs Median Income (Figure 6):** Imprisonment rate and median income has a generally negative correlation. This isn't surprising as it is generally understood that a lack of resources, including financial, can be a driving force towards engaging in illegal activities.

**Imprisonment vs Population Aged 15-29 (Figure 7):** Imprisonment rate and percentage of population aged 15-29 in the census tract has a generally positive correlation. This isn't surprising as it's generally understood that one's likelihood to engage in illegal activities is higher as an adolescent.

### Potential Interactions (Figure 8):

There appears to be an interaction effect between race and income, as the relationship between median income and imprisonment rate differs by race. We created a variable, POC\_other, indicating the percentage of a census tract population that is non-White. In this figure we categorized tracts with POC\_other values that were above and below the median in the data of this variable to create a categorical split. The relationship between imprisonment rate and income appears more negatively correlated when the percentage of non-white people in the census tract is above the median.

## Methodology and Results

For our analysis process, we iterated through different models until we were satisfied with its conditions and performance. As an initial model, we did simple linear regression to predict a census tract's imprisonment rate using bll indicator variable, racial makeup, median income, median age, the percentage of males, and the interaction between income and race. The model output and conditions are below:

### Checking Initial Modelling Conditions

term	estimate	std.error	statistic	p.value
(Intercept)	-357.488	162.353	-2.202	0.028
perc_bll_indicator	-1.802	2.848	-0.633	0.527
POC_other	6.274	0.989	6.345	0.000
med_income	-0.001	0.000	-4.522	0.000
median_age	-3.145	1.379	-2.281	0.023

term	estimate	std.error	statistic	p.value
perc_male	17.270	2.859	6.041	0.000
POC_other:med_income	0.000	0.000	-3.974	0.000

Using the residual vs fitted plot (**Figure 9**), we can assess the model conditions. From this graph, we find

1. Normality condition can be relaxed.
2. Linearity condition not satisfied. Points not randomly scattered around the  $x = 0$  line.
3. Constant variance condition not satisfied. Points fan out across the  $x = 0$  line.
4. Independence condition also may not be satisfied. Census tracts next to each other could be more likely to have similar incarceration rates or blood lead levels.

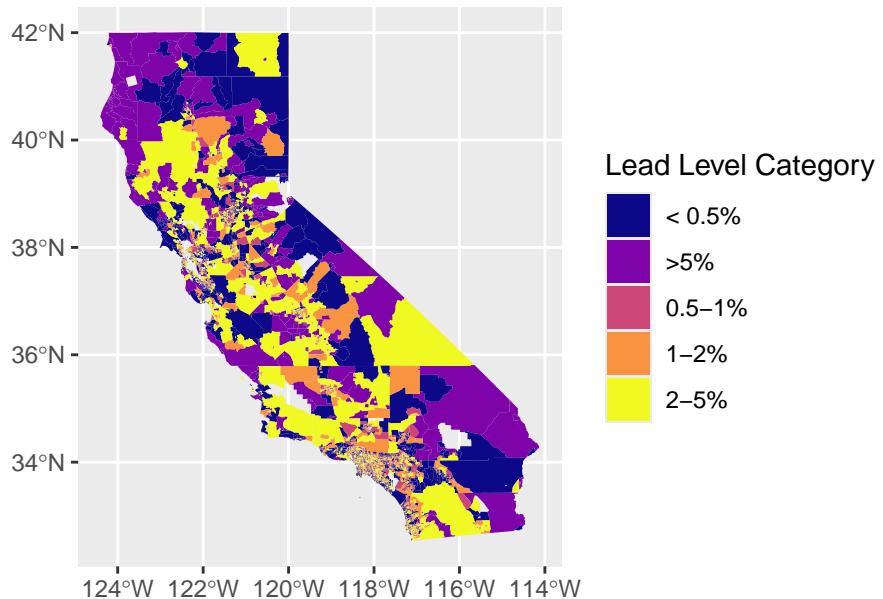
Based on these initial conditions, we decided it would make sense to log transform some of our variables, specifically our median income and imprisonment rate variables

### Splitting the Data

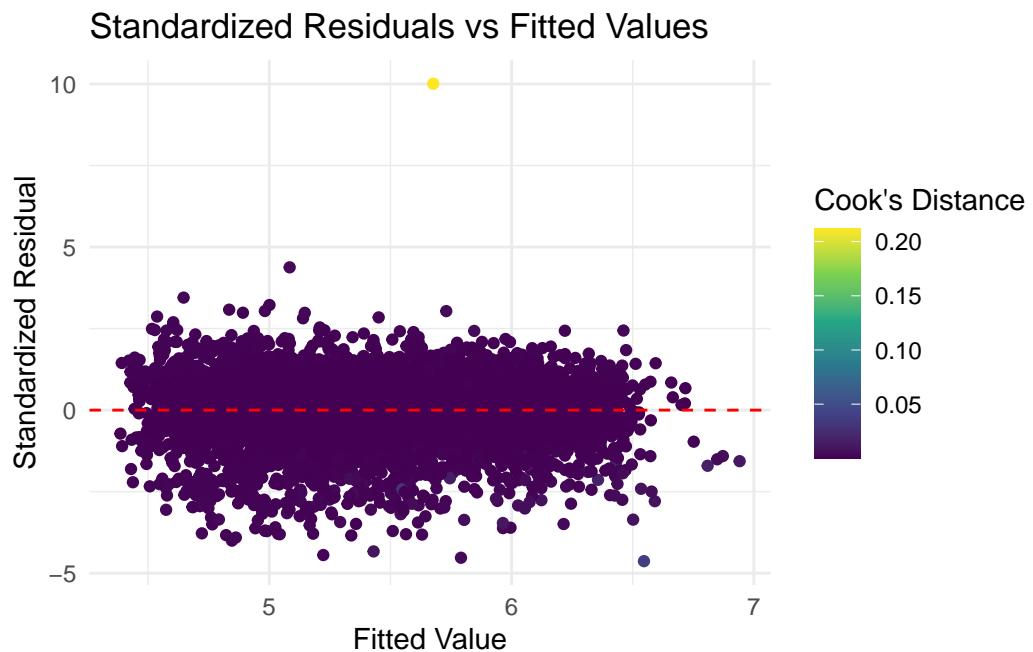
We were also concerned about spatial correlation amongst census tracts especially since there can be multiple within a single county meaning they would share a lot of features. So we originally split the data with 20% in training and 80% testing to minimize this effect. We received abysmal R squared values when doing this. Using ChatGPT for suggestions, it recommended spatial cross validation, which we researched further to ensure it made sense for our analysis<sup>2</sup>. It appeared to be a feasible method to address model overfitting and independence issues within our data. Because of that, we were then able to do a 80% training, 20% testing split.

### 2. Geocomputation with R

## Blood Lead Level Categories by Census Tract



## Final Model Selection



term	estimate	std.error	statistic	p.value
(Intercept)	11.971	0.330	36.315	0.000
log_med_income	-0.663	0.026	-25.624	0.000
perc_bll_indicator	-0.003	0.003	-1.045	0.296
POC_other	-0.027	0.009	-2.929	0.003
median_age	0.005	0.001	4.476	0.000
perc_male	0.008	0.003	2.915	0.004
log_med_income:POC_other	0.004	0.001	4.522	0.000

Start: AIC=-7536.27

```
log_imprisonment_rt ~ log_med_income + perc_bll_indicator + POC_other +
median_age + perc_male + POC_other * log_med_income
```

	Df	Sum of Sq	RSS	AIC
<none>		1925.7	-7536.3	
- perc_male	1	2.5847	1928.3	-7529.8
- median_age	1	6.0945	1931.8	-7518.2
- log_med_income:POC_other	1	6.2204	1931.9	-7517.8

Call:

```
lm(formula = log_imprisonment_rt ~ log_med_income + perc_bll_indicator +
POC_other + median_age + perc_male + POC_other * log_med_income,
data = bll_train)
```

Residuals:

Min	1Q	Median	3Q	Max
-2.5390	-0.2884	0.0555	0.3615	5.4796

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	11.9709416	0.3296440	36.315	< 2e-16 ***
log_med_income	-0.6631090	0.0258782	-25.624	< 2e-16 ***
perc_bll_indicator	-0.0026630	0.0025478	-1.045	0.29596
POC_other	-0.0274528	0.0093714	-2.929	0.00341 **
median_age	0.0053888	0.0012039	4.476	7.73e-06 ***
perc_male	0.0079678	0.0027333	2.915	0.00357 **
log_med_income:POC_other	0.0038170	0.0008441	4.522	6.23e-06 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

```

Residual standard error: 0.5515 on 6331 degrees of freedom
Multiple R-squared:  0.4201,    Adjusted R-squared:  0.4195
F-statistic: 764.3 on 6 and 6331 DF,  p-value: < 2.2e-16

```

term	estimate	std.error	statistic	p.value
(Intercept)	11.971	0.330	36.315	0.000
log_med_income	-0.663	0.026	-25.624	0.000
perc_bll_indicator	-0.003	0.003	-1.045	0.296
POC_other	-0.027	0.009	-2.929	0.003
median_age	0.005	0.001	4.476	0.000
perc_male	0.008	0.003	2.915	0.004
log_med_income:POC_other	0.004	0.001	4.522	0.000

Table 4: Model Performance (RMSE and R-squared)

.metric	.estimator	.estimate
rmse	standard	184.971
rsq	standard	0.495

```

Start:  AIC=-9001.12
log_imprisonment_rt ~ log_med_income + perc_bll_indicator + POC_other +
median_age + perc_male + POC_other * log_med_income

Df Sum of Sq    RSS      AIC
<none>                      1983.0 -9001.1
- perc_male                  1     1.6252 1984.6 -8997.3
- median_age                 1    12.1954 1995.2 -8959.7
- log_med_income:POC_other   1   15.4345 1998.4 -8948.2

```

Table 5: Random-Split Model Performance (RMSE and R-squared)

.metric	.estimator	.estimate
rmse	standard	158.152
rsq	standard	0.442

```
[1] 0.4200831
```

```
[1] 0.47341
```

## Discussion

## Appendix

### Univariate EDA

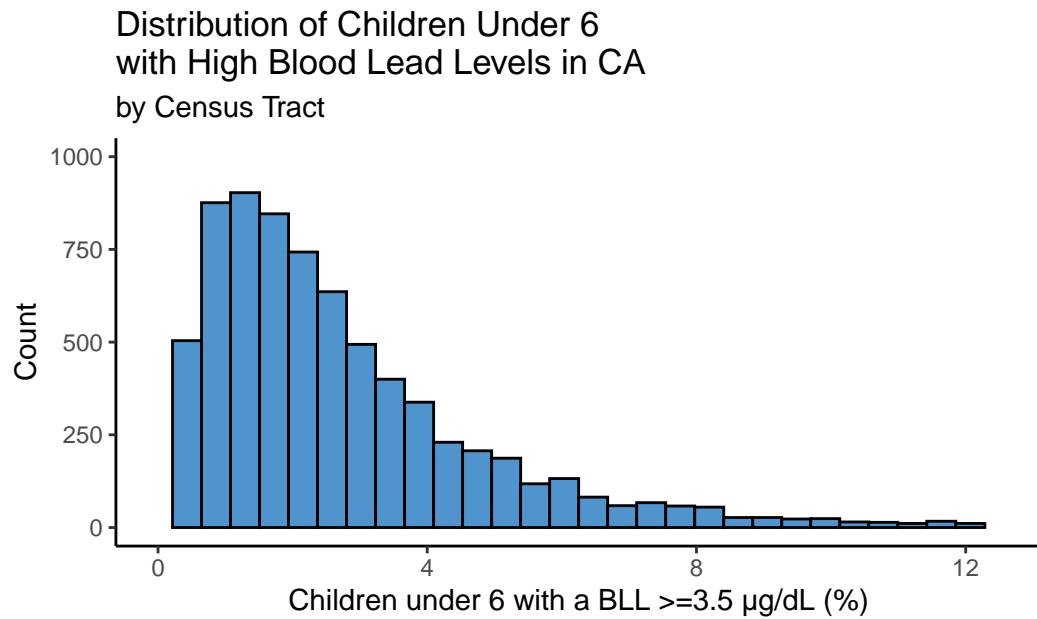


Figure 1

Figure 1: Distribution of lead in children

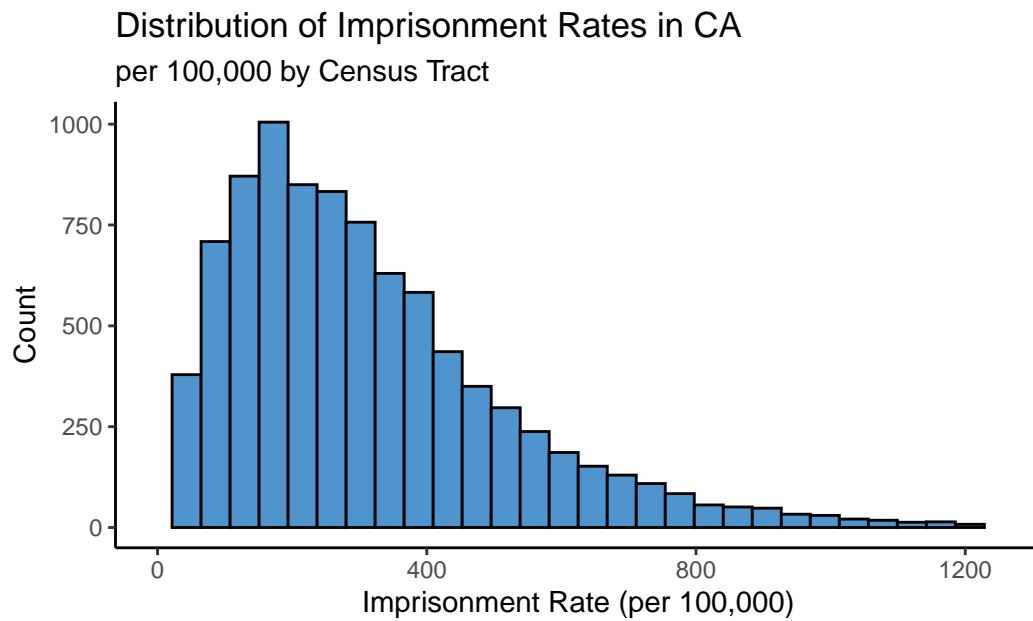


Figure 2

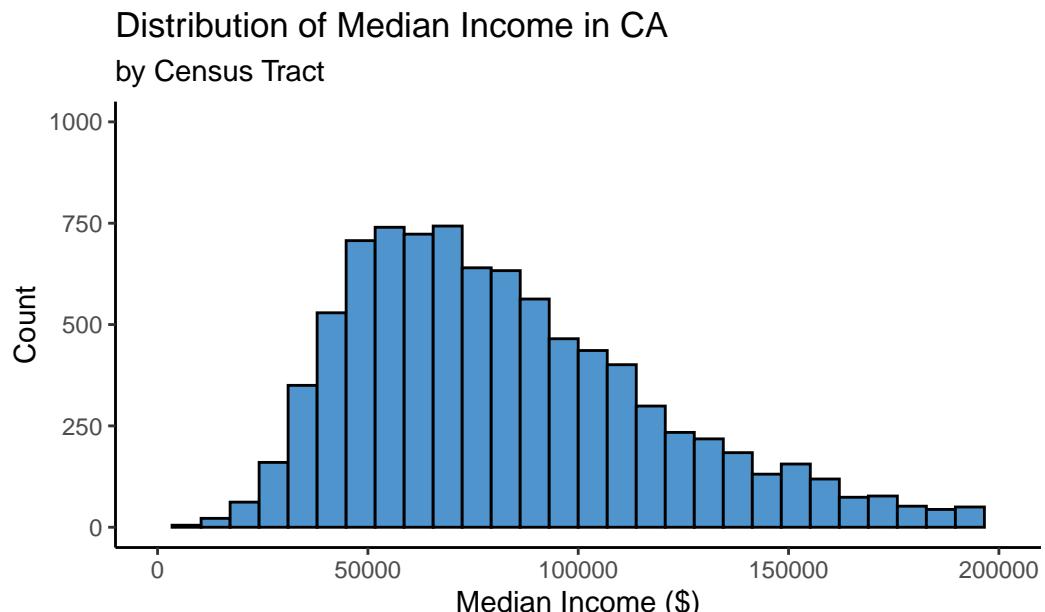


Figure 3

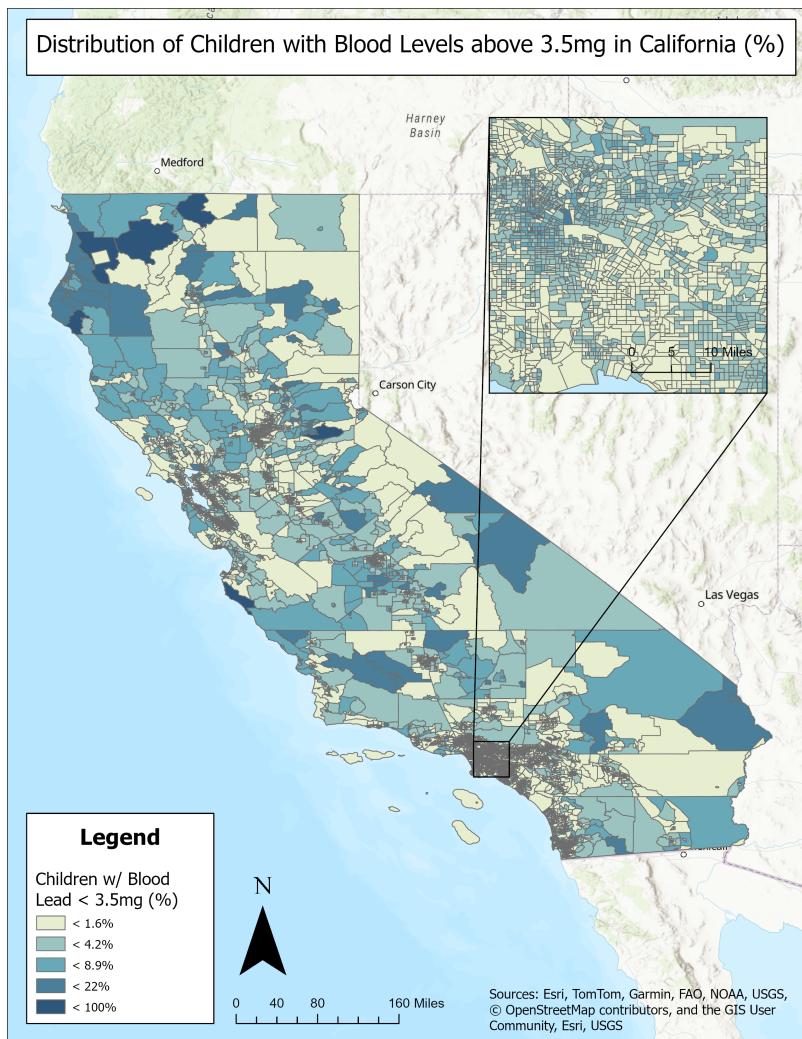


Figure 2: Figure 4: Spatial Distribution of BLL in California

## Bivariate EDA

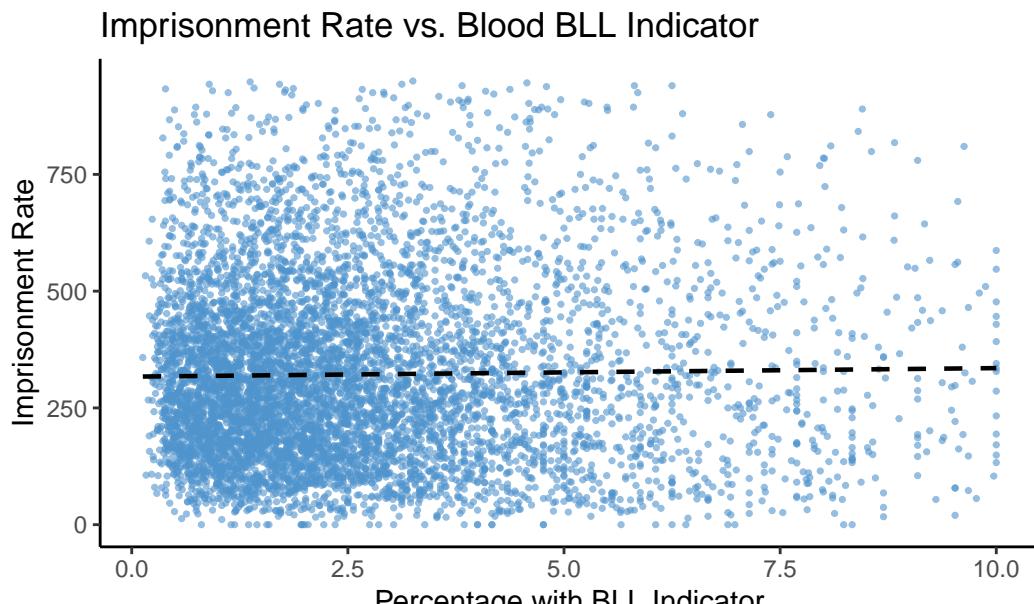


Figure 5

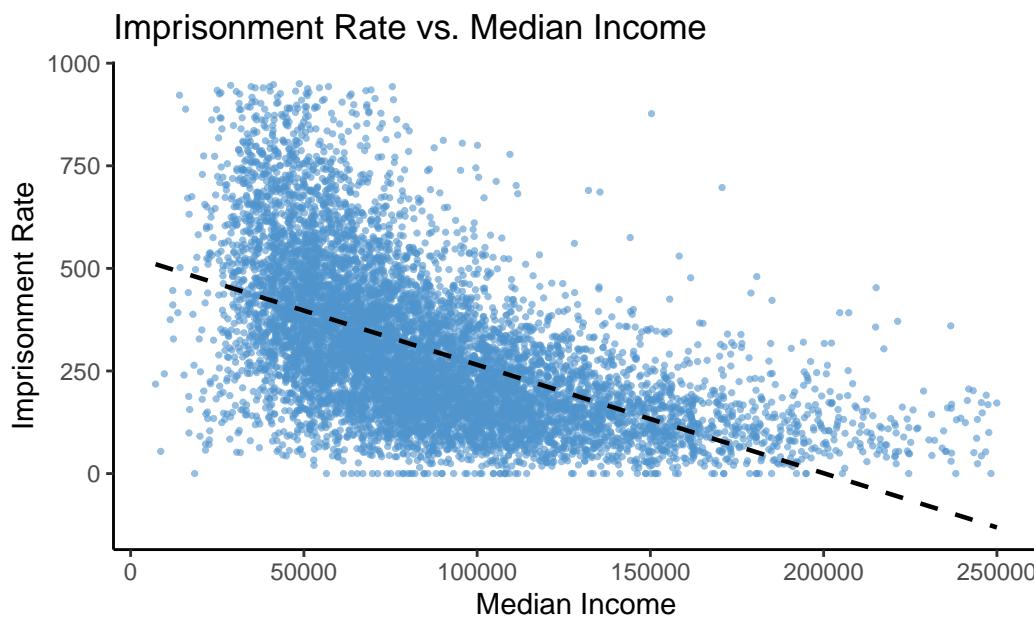


Figure 6

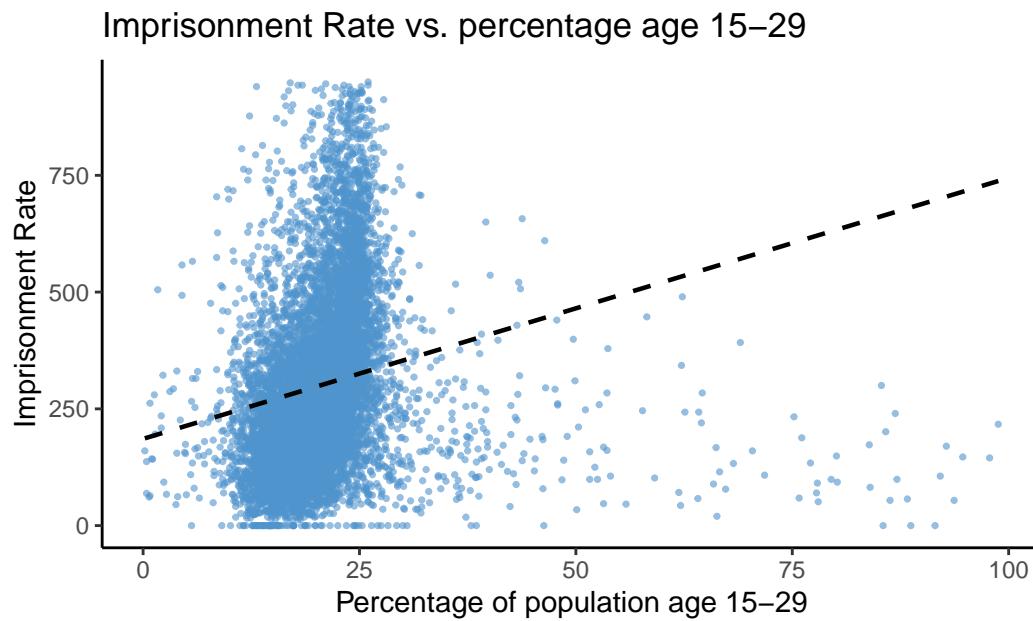


Figure 7

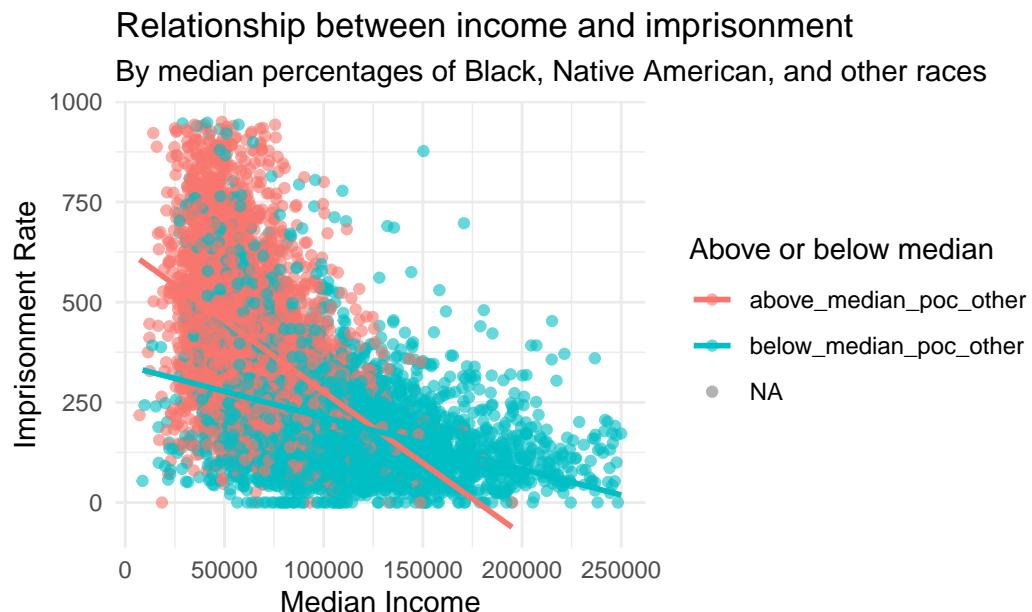


Figure 8

### Residuals vs Fitted Values

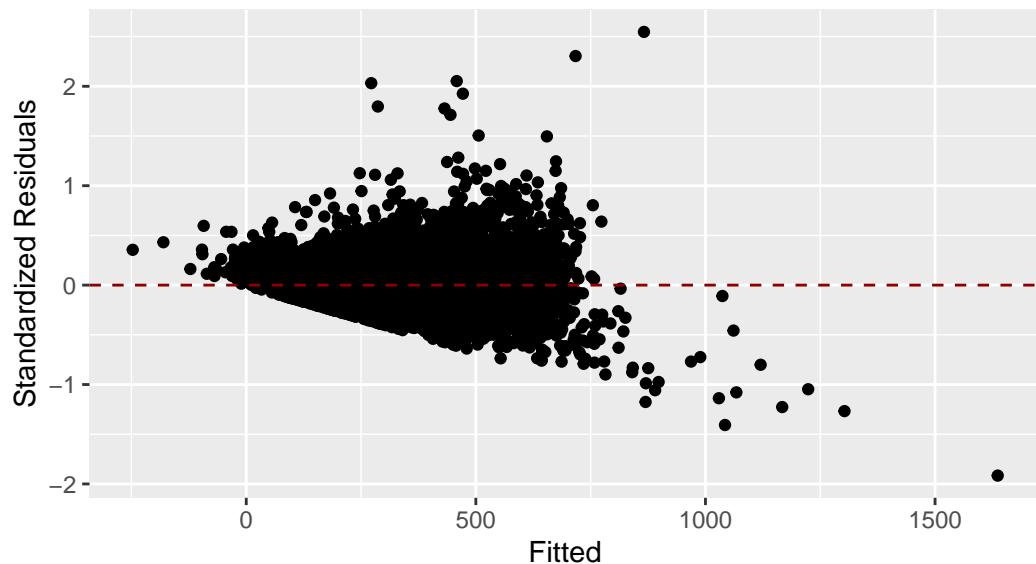


Figure 9

perc_bll_indicator	POC_other	med_income
1.047454	6.401602	2.310705
median_age	perc_male POC_other:med_income	
1.631162	1.030406	3.966140