Project Proposal

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

Introduction

From our recent systematic review of studies we have explored the potential effects of lead exposure on brain development in children and adults. Various 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. Exposure to lead during the developmental years of children causes irreversible damage.

Studies in the past have shown a strong correlation between aggressive behavior, criminal tendencies and exposure to lead. Talayero et al. (2023)[1] highlights a strong association between lead exposure during childhood and criminal tendencies during adulthood. Exposure to lead can be through various means, including water, which is what we've chosen to investigate. Our research topic inquires about lead levels in water and if it relates to the incarceration rates of a specified area, while also considering other demographic considerations.

This research topic has important societal implications, namely the complicated intersection of crime, environmental racism, and potentially 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.

1. The association between lead exposure and crime: A systematic review

Data description

California incarceration rates comes from an organization called the prison policy initiative (https://www.prisonpolicy.org/origin/ca/2020/zipcode.html). PPI reports incarceration rate by zip code via data collected by the US Census Bureau. This data is publicly available because of new laws in California on prison gerrymandering. This data was redistricted to display number of people incarcerated in California State prisons by Peter Horton of the redistricting data hub. They publish all of their methodologies on their website. This dataset has 1803 observations. Since their data is derived from the census bureau, the data was collected through the official collection and reporting procedures of the US government. A limitation is that it does not disclose people incarcerated in federal prisons, but there are only around 12,000

California residents incarcerated in federal facilities. Comparatively, in California, 100,000 are incarcerated in state facilities (https://www.prisonpolicy.org/profiles/CA.html). All variables include zip code, city, number of people in state prison from that zip code, census population in 2020, total population 2020, and imprisonment rate per 100,000.

Zip code lead data comes from the California Department of Public Health (CDPH). We selected "Percent of Children of with a Blood Lead Level of 3.5 or higher in descending order-2022". This dataset has 1777 observations. Federal guidelines require that children served by Medicaid be screened for lead poisoning with a blood lead level (BLL) test at ages 12 and 24 months, and up to age 6 years if not previously tested. The CDPH collects and reports this data. All variables include zip code, postal district name, number of children under 6 with a blood lead level of 3.5 milligrams/dL or more, percent of children under 6 with a blood lead level of 3.5 milligrams/dL or more, and total number of children under 6 with a BLL level. One limitation is that this data and the other data are not from the same year. Another is that we are using blood lead level rates in children as a proxy for susceptibility to lead exposure for everyone in that zip code.

We also added demographic data sourced from census.gov. We added 16 columns pertaining to race, income, gender, and age in each zip code. Since these data are derived from the census bureau, the data were collected through the official collection and reporting procedures of the US government. One limitation is that these data are from the 2023 census report, which is not aligned with the lead data, which is from 2022, and the incarceration rate data, which is from 2020.

Data processing

Cleaning Process

This analysis requires the joining of several different datasets, namely the incarceration, lead, demographic, and income datasets for California. Almost all of the datasets contain the full range of CA zip codes, but the lead data has fewer observations than the other sets, so we conducted a series of left joins to the lead data to ensure that each observation has a complete range of data (which we checked and holds true). The formatting of the demographic and income data (sourced from census.gov) was, for all intents and purposes, unusable with over 360 columns in the demographic set and containing both estimates and margins of error. We pared down the sets to select only the columns that would be most relevant for this investigation, as well as created age variables that split the range in three to simplify the data. We also had to rename the columns in the census data to be intelligible and columns in the lead and incarceration sets to eliminate the capital letters, spaces, and symbols which would complicate the analysis. Our final dataset has all columns in tidy formatting and are all joined together by zip codes, each with full data for every column.

Data Cleaning

```
library(tidyverse)
library(tidymodels)
library(patchwork)
library(ggplot2)
install.packages("readxl")
library(readxl)
library(dplyr)
library(GGally)
```

```
ca_incarceration <- read_excel("data/census_tract_prison.xlsx")
ca_lead <- read_excel("data/New_census_tract_lead_level.xlsx")


ca_demographic <- read_csv("data/demographic.csv", skip = 1)
ca_income <- read_csv("data/income.csv", skip = 1)
```

```
#transforming the data

#demographics
ca_demographic <- ca_demographic |>
  select(c(1:3, 27, 75, 164:176, 240:245)) |>
  rename_with(~ str_replace(., ".*!!", "")) |>
  filter(Geography != "0400000US06")

ca_demographic <- ca_demographic |>
  mutate(`Geography` = substr(`Geography`, 10, 20)) |>
  mutate(across(!(1:2), as.numeric)) |>
  mutate(`0_to_19` =
           `Under 5 years` + `5 to 9 years` +
           `10 to 14 years` + `15 to 19 years`,

         `20_to_44` =
           `20 to 24 years` + `25 to 29 years` +
           `30 to 34 years` + `35 to 39 years` +
           `40 to 44 years`,

         `45_to_64` =
           `45 to 49 years` + `50 to 54 years` +
           `55 to 59 years` + `60 to 64 years`
```

```
        ) |>
  mutate(perc_male = `Male population` / `Total population` * 100) |>
  select(!6:18) |>
  rename(census_tract = "Geography") |>
  rename(median_age = "Both sexes")


#income
ca_income <- ca_income |>
  select(c(1, 2, 25)) |>
  rename_with(~ str_replace(., ".*!!", "")) |>
  filter(Geography != "0400000US06")

ca_income <- ca_income |>
  mutate(`Geography` = substr(`Geography`, 10, 20)) |>
  mutate(across(!(1:2), as.numeric)) |>
  rename(census_tract = "Geography")


#incarceration
ca_incarceration <- ca_incarceration |>
  select(c(1, 2, 6)) |>
  mutate(fips = paste0("0", fips)) |>
  rename(census_tract = `fips`) |>
  mutate(across(!(1:2), as.numeric))


#lead
ca_lead <- ca_lead |>
  select(!1) |>
  mutate(`Census Tract` = paste0("0", `Census Tract`)) |>
  rename(census_tract = `Census Tract`) |>
  mutate(across(!1, as.numeric)) |>
  mutate(`Percent of children under 6 with a BLL of 3.5 µg/dL or greater` =
           `Percent of children under 6 with a BLL of 3.5 µg/dL or greater` * 100)
```

```
#left_join to pare down to least observation dataset
#lot of data cleaning

joined_data1 <- left_join(ca_lead, ca_incarceration, by = "census_tract")
joined_data2 <- left_join(joined_data1, ca_demographic, by = "census_tract")
```

```r
joined_data <- left_join(joined_data2, ca_income, by = "census_tract")

#post-join cleaning
bll_data <- joined_data |>
  janitor::clean_names() |>
  select(!c(geographic_area_name_x, geographic_area_name_y)) |>
  rename(num_bll_indicator =
           number_of_children_under_6_with_a_bll_of_3_5_mg_d_l_or_greater) |>
  rename(perc_bll_indicator =
           percent_of_children_under_6_with_a_bll_of_3_5_mg_d_l_or_greater) |>
  rename(num_bll =
           total_number_of_children_under_6_with_a_bll) |>
  rename(imprisonment_rt = imprisonment_rate_per_100_000) |>
  rename(tract_name = ca_census_tracts) |>
  rename(other_race = some_other_race) |>
  rename(black = black_or_african_american) |>
  rename(native_am = american_indian_and_alaska_native) |>
  rename(pac_islander = native_hawaiian_and_other_pacific_islander) |>
  rename(age_0_to_19 = x0_to_19) |>
  rename(age_20_to_44 = x20_to_44) |>
  rename(age_45_to_64 = x45_to_64) |>
  rename(med_income = median_income_dollars) |>
  rename(total_pop_2020 = total_population)

#DROP ANY ADDITIONAL BELOW HERE
```

```r
nrow(bll_data)
```

```
[1] 9114
```

```r
ncol(bll_data)
```

```
[1] 20
```

Univariate Response Variable EDA

```r
bll_data |>
  filter(imprisonment_rt <= quantile(imprisonment_rt, 0.98, na.rm = TRUE)) |>
  ggplot(aes(x = imprisonment_rt)) +
```

```
  geom_histogram() +
  labs(
    title = "Distribution of imprisonment rate per 100,000 people",
    subtitle = "by California Zip Codes",
    x = "Imprisonment rate (per 100,000 residents)",
    y = "Count"
  )
```

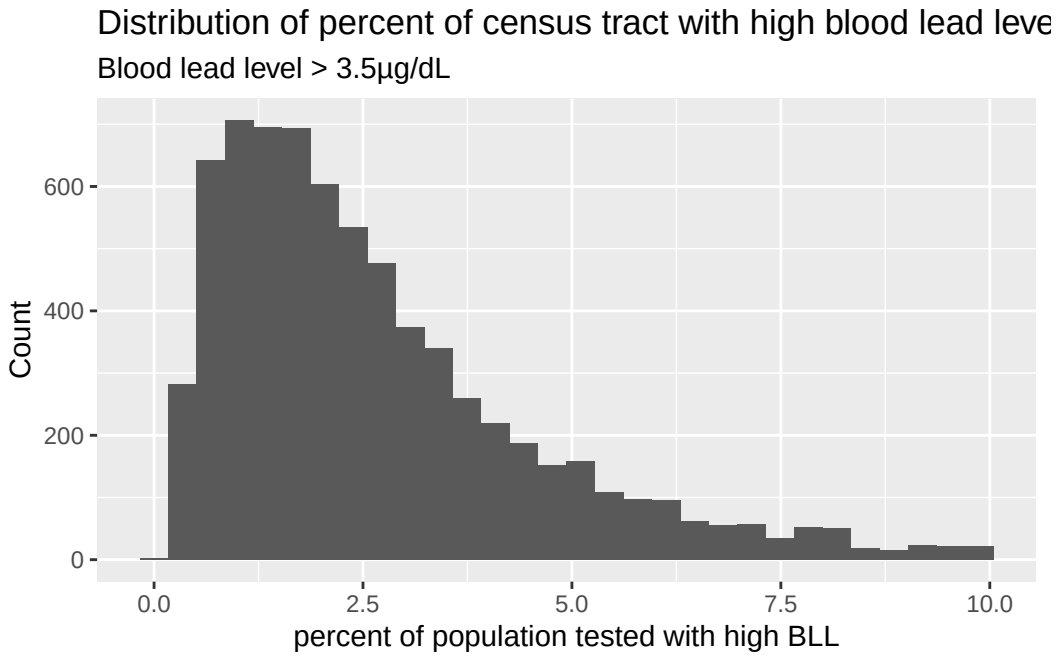

```
bll_data |>
  summarize(mean = mean(imprisonment_rt, na.rm = TRUE),
            median = median(imprisonment_rt, na.rm = TRUE),
            IQR = IQR(imprisonment_rt, na.rm = TRUE),
            sd = sd(imprisonment_rt, na.rm = TRUE),
            min = min(imprisonment_rt, na.rm = TRUE),
            max = max(imprisonment_rt, na.rm = TRUE))
```

```
# A tibble: 1 x 6
   mean median   IQR    sd   min    max
  <dbl>  <dbl> <dbl> <dbl> <dbl>  <dbl>
1  376.    267   257 2101.     0 100000
```

After removing extreme outliers (the top 2% incarceration rates- some of these 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. There are a few possible outliers in the data - primarily in zip codes with low populations (such as 93262 which has an imprisonment rate of 4,762 per 100,000). There are 21 zip codes with imprisonment rates over 1,000 which may be possible outliers. The median of the data is 251 imprisonments per 100,000 residents, while the IQR is 245.5.

Univariate Predictor Variable EDA

```
bll_data |>
  filter(perc_bll_indicator <= quantile(perc_bll_indicator, 0.98, na.rm = TRUE)) |>
  filter(num_bll != 0) %>%
  filter(perc_bll_indicator != 0) %>%
  ggplot(aes(x = perc_bll_indicator)) +
  geom_histogram() +
  labs(
    title = "Distribution of percent of census tract with high blood lead levels",
    subtitle = "Blood lead level > 3.5µg/dL",
    x = "percent of population tested with high BLL",
    y = "Count"
  )
```



```r
bll_data |>
  filter(num_bll != 0) %>%
  summarize(mean = mean(perc_bll_indicator, na.rm = TRUE),
            median = median(perc_bll_indicator, na.rm = TRUE),
            IQR = IQR(perc_bll_indicator, na.rm = TRUE),
            sd = sd(perc_bll_indicator, na.rm = TRUE),
            min = min(perc_bll_indicator, na.rm = TRUE),
            max = max(perc_bll_indicator, na.rm = TRUE))
```

```
# A tibble: 1 x 6
   mean median   IQR    sd   min   max
  <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl>
1  2.35   1.71  2.58  2.87     0   100
```

```r
#percent blood level over 1.5 the IQR
bll_data %>%
  filter(perc_bll_indicator > 1.5*2.599239)
```

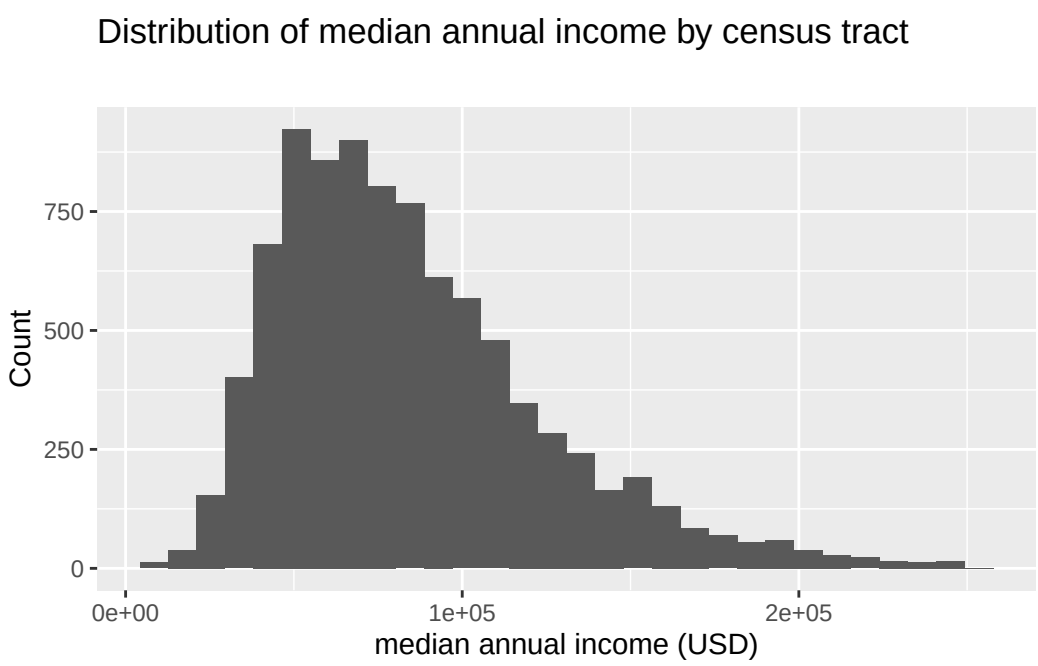
```
# A tibble: 1,614 x 20
   census_tract num_bll_indicator perc_bll_indicator num_bll tract_name
   <chr>                    <dbl>              <dbl>   <dbl> <chr>
 1 06001400300                  9               6.67     135 Alameda County, Ce~
 2 06001400700                 12               7.19     167 Alameda County, Ce~
 3 06001400800                  5               4.76     105 Alameda County, Ce~
 4 06001400900                 10              11.8       85 Alameda County, Ce~
 5 06001401000                 11               5.07     217 Alameda County, Ce~
 6 06001401100                 12               9.23     130 Alameda County, Ce~
 7 06001401200                  6               5.36     112 Alameda County, Ce~
 8 06001401400                 18               9.68     186 Alameda County, Ce~
 9 06001401500                  6               7.23      83 Alameda County, Ce~
10 06001401600                 10              10        100 Alameda County, Ce~
# i 1,604 more rows
# i 15 more variables: imprisonment_rt <dbl>, total_pop_2020 <dbl>,
#   male_population <dbl>, median_age <dbl>, white <dbl>, black <dbl>,
#   native_am <dbl>, asian <dbl>, pac_islander <dbl>, other_race <dbl>,
#   age_0_to_19 <dbl>, age_20_to_44 <dbl>, age_45_to_64 <dbl>, perc_male <dbl>,
#   med_income <dbl>
```

```r
bll_data %>%
  filter(perc_bll_indicator == 0) %>%
  nrow()
```

```
[1] 1899
```

There are 94 census tracts that tested zero blood lead levels in this data. These will likely need to be removed because these observations are not useful in the analysis. Additionally, outside of census tracts that tested zero people, there is a large concentration of census tracts that have a percent blood level indicator of 0 (1899 observations). While these should probably be included in the final analysis, removing these produces for data visualization produces a unimodal distributon that is heavily right skewed. Overall, the median blood level is 1.7 and the IQR is 2.57.

```
bll_data |>
  ggplot(aes(x = med_income)) +
  geom_histogram() +
  labs(
    title = "Distribution of median annual income by census tract",
    subtitle = "",
    x = "median annual income (USD)",
    y = "Count"
  )
```




```
bll_data |>
  summarize(mean = mean(med_income, na.rm = TRUE),
```

```
            median = median(med_income, na.rm = TRUE),
            IQR = IQR(med_income, na.rm = TRUE),
            sd = sd(med_income, na.rm = TRUE),
            min = min(med_income, na.rm = TRUE),
            max = max(med_income, na.rm = TRUE))
```

```
# A tibble: 1 x 6
    mean median   IQR     sd   min    max
   <dbl>  <dbl> <dbl>  <dbl> <dbl>  <dbl>
1 84844.  77225 50440 39676.  4918 250000
```

```
bll_data %>%
  filter(med_income > (1.5*50440)) %>%
  nrow()
```

```
[1] 4622
```

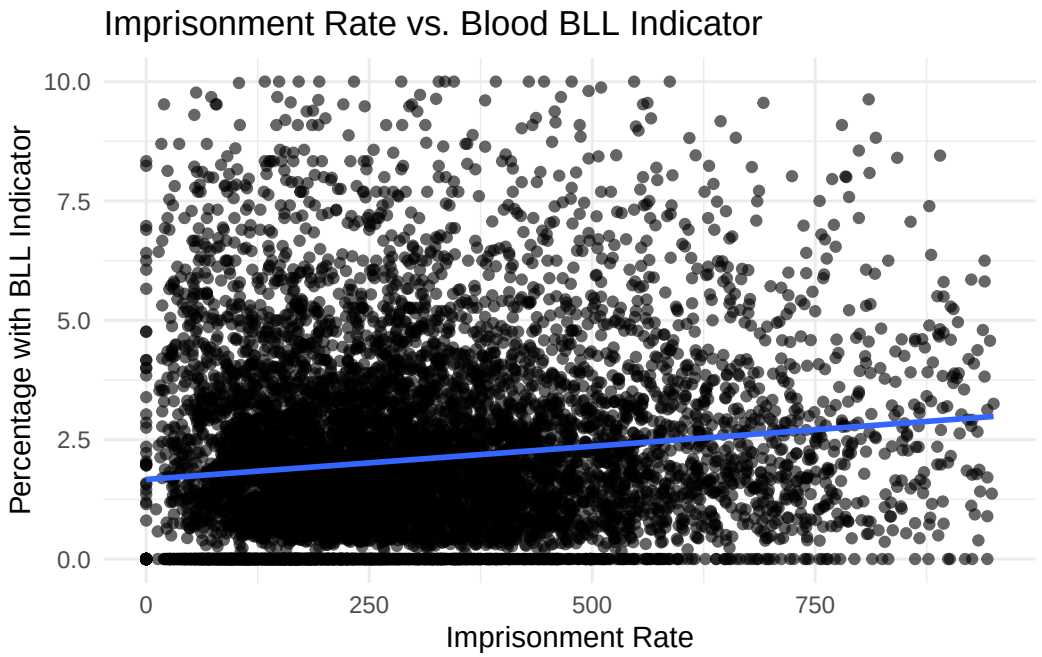
```
bll_data %>%
  filter(med_income < (1.5*50440)) %>%
  nrow()
```

```
[1] 4341
```

The distribution of median annual income is right skewed and centered at 77225.

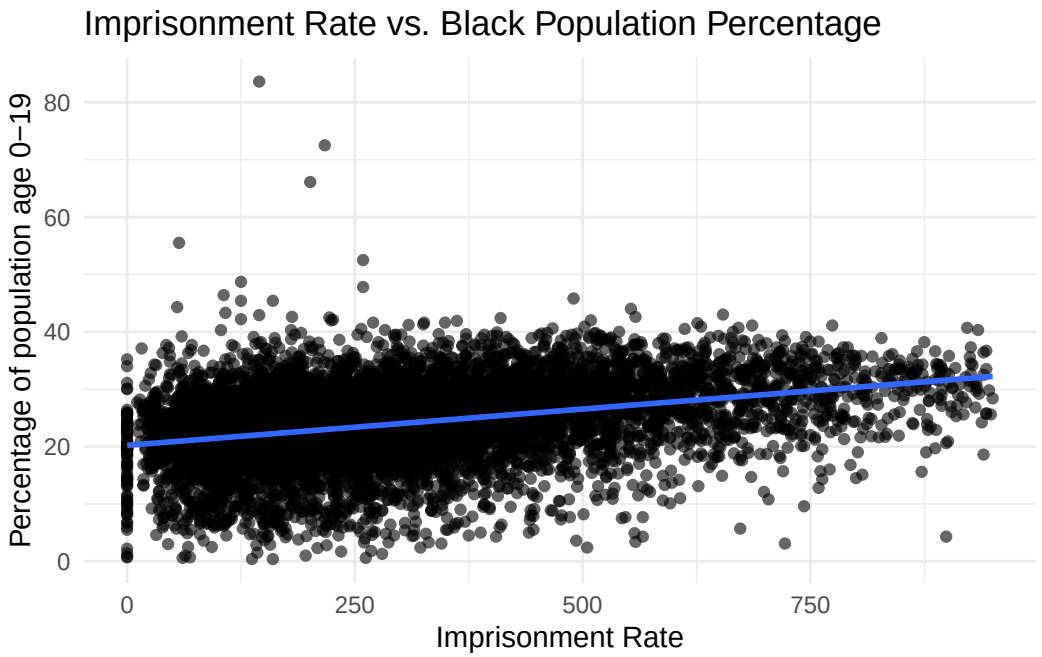
Bivariate Response Variable EDA

```
bll_data %>%
filter(perc_bll_indicator <= quantile(perc_bll_indicator, 0.98, na.rm = TRUE)) %>%
filter(imprisonment_rt <= quantile(imprisonment_rt, 0.98, na.rm = TRUE)) %>%
filter(num_bll != 0) %>%
ggplot(aes(x = imprisonment_rt, y = perc_bll_indicator)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Imprisonment Rate vs. Blood BLL Indicator",
       x = "Imprisonment Rate",
       y = "Percentage with BLL Indicator") +
  theme_minimal()
```




```r
bll_data %>%
filter(perc_bll_indicator <= quantile(perc_bll_indicator, 0.98, na.rm = TRUE)) %>%
filter(imprisonment_rt <= quantile(imprisonment_rt, 0.98, na.rm = TRUE)) %>%
filter(num_bll != 0) %>%
ggplot(aes(x = imprisonment_rt, y = med_income)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Imprisonment Rate vs. Median Income",
       x = "Imprisonment Rate",
       y = "Median Income") +
  theme_minimal()
```


```r
bll_data %>%
filter(perc_bll_indicator <= quantile(perc_bll_indicator, 0.98, na.rm = TRUE)) %>%
filter(imprisonment_rt <= quantile(imprisonment_rt, 0.98, na.rm = TRUE)) %>%
filter(num_bll != 0) %>%
ggplot(aes(x = imprisonment_rt, y = age_0_to_19)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Imprisonment Rate vs. Black Population Percentage",
       x = "Imprisonment Rate",
       y = "Percentage of population age 0-19") +
  theme_minimal()
```

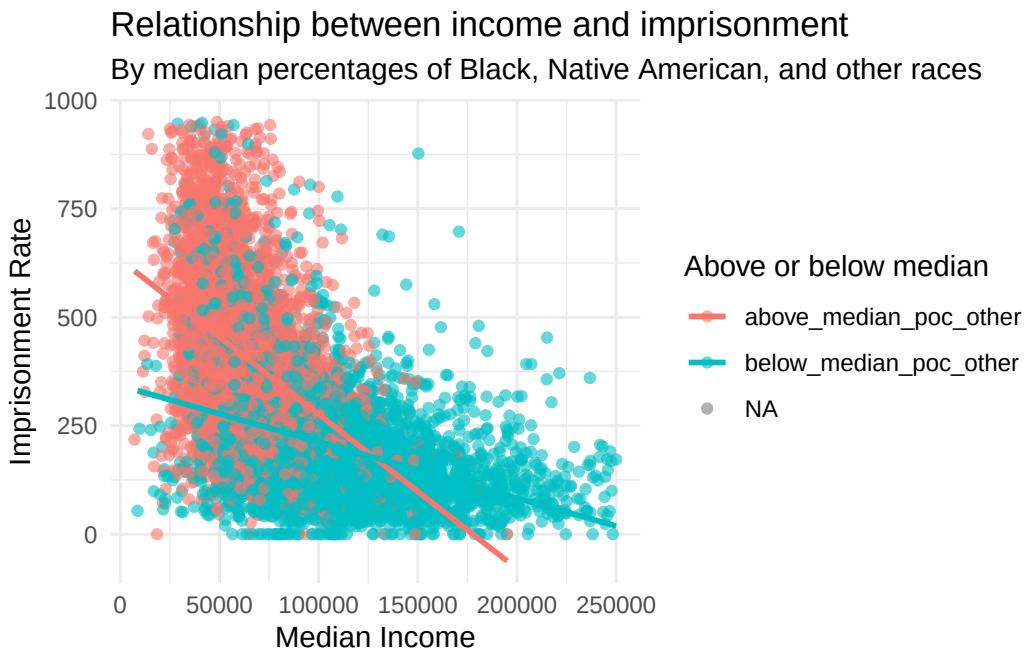


Interaction Effect EDA

```
bll_data <- bll_data %>%
  mutate(POC_other = black + native_am + other_race,
         race_categorical = ifelse(POC_other > median(POC_other, na.rm = TRUE),
                                   "above_median_poc_other",
                                   "below_median_poc_other"))

bll_data %>%
filter(perc_bll_indicator <= quantile(perc_bll_indicator, 0.98, na.rm = TRUE)) %>%
filter(imprisonment_rt <= quantile(imprisonment_rt, 0.98, na.rm = TRUE)) %>%
filter(num_bll != 0) %>%
ggplot(aes(x = med_income, y = imprisonment_rt, color = race_categorical)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(
    title = "Relationship between income and imprisonment",
    subtitle = "By median percentages of Black, Native American, and other races",
    x = "Median Income",
    y = "Imprisonment Rate",
    color = "Above or below median"
```

```
  ) +
  theme_minimal()
```



When comparing the relationship between median income and imprisonment rate, it appears that generally they have a negative correlation. This graph suggests there could be an interaction effect between race and income, as the relationship between median income and imprisonment rate differs by race. We created a categorical variable for the percentage of the census tract population that is black, native american, or "other race" that is categorically above or below the median in the data. The relationship between imprisonment rate and median income appears more negatively correlated when categorically above the median census tract population percentage of black, native american, and other_race. This supports that there could be an interaction effect between race and income.

Analysis approach

For our project, since we are looking to make conclusions about whether water lead levels and other societal factors correspond to higher imprisonment rates, we have identified a few particular variables of interest from our dataset to analyze. For predictors, we are looking at California zip codes (categorical), race (categorical), percent of children with a significant blood lead level (numerical), and income (numerical). As for our response variable, we are looking at the imprisonment rate within California zip codes (numerical). Given this research question has significant societal implications, we are interested in looking at a variety of predictors that

could give insight into variability in imprisonment rates. As such, we will likely rely on multiple linear regression to conduct our analysis. We plan to create different models and evaluate them to determine the best combination of predictors and their interactions to explain our response variable and hopefully make interesting conclusions.

Data dictionary

The data dictionary can be found here as well as in the README.