Project Proposal

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

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

Introduction

Data description

```
ca_incarceration <- read_excel("data/California incarceration rate data.xlsx")
ca_lead <- read_excel("data/ZIPCodeData2022 (1).xlsx")

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

```
#transforming the data

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

ca_demographic <- ca_demographic |>
  select(!((contains("Estimate") | contains("Margin")) & !37)) |>
  select(c(1:6, 8:17, 40, 41, 42, 47, 55, 60)) |>
  rename_with(~ str_replace(., ".*!!", ""))
```

```r
ca_demographic <- ca_demographic |>
  mutate(`Geographic Area Name` = substr(`Geographic Area Name`, 7, 11)) |>
  mutate(across(!1, 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 34 years` + `35 to 44 years`,
         `45_to_64` = `45 to 54 years` + `55 to 59 years` + `60 to 64 years`) |>
  select(!7:16) |>
  rename(zip_code = `Geographic Area Name`)

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

ca_income <- ca_income |>
  select(c(1, 2, 25, 27)) |>
  rename_with(~ str_replace(., ".*!!", "")) |>
  mutate(`Geographic Area Name` = substr(`Geographic Area Name`, 7, 11)) |>
  mutate(across(!1, as.numeric)) |>
  rename(zip_code = `Geographic Area Name`)
```

```r
#renaming the columns
ca_incarceration <- ca_incarceration |>
  rename(zip_code = `California ZIP codes    `) |>
  mutate(zip_code = as.numeric(zip_code))


ca_lead <- ca_lead |>
  rename(zip_code = `ZIP Code`) |>
  mutate(zip_code = as.numeric(zip_code))
```

```r
#left_join to pare down to least observation dataset

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

joined_data <- left_join(joined_data2, ca_income, by = "zip_code")

joined_data_clean <- joined_data |>
  janitor::clean_names() |>
  select(!c(geography_y, city, census_population_2020)) |>
  rename(city = postal_district_name) |>
  rename(geography = geography_x)
```

```
joined_data_clean %>%
  nrow()
```

```
[1] 1777
```

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. ## Exploratory data analysis

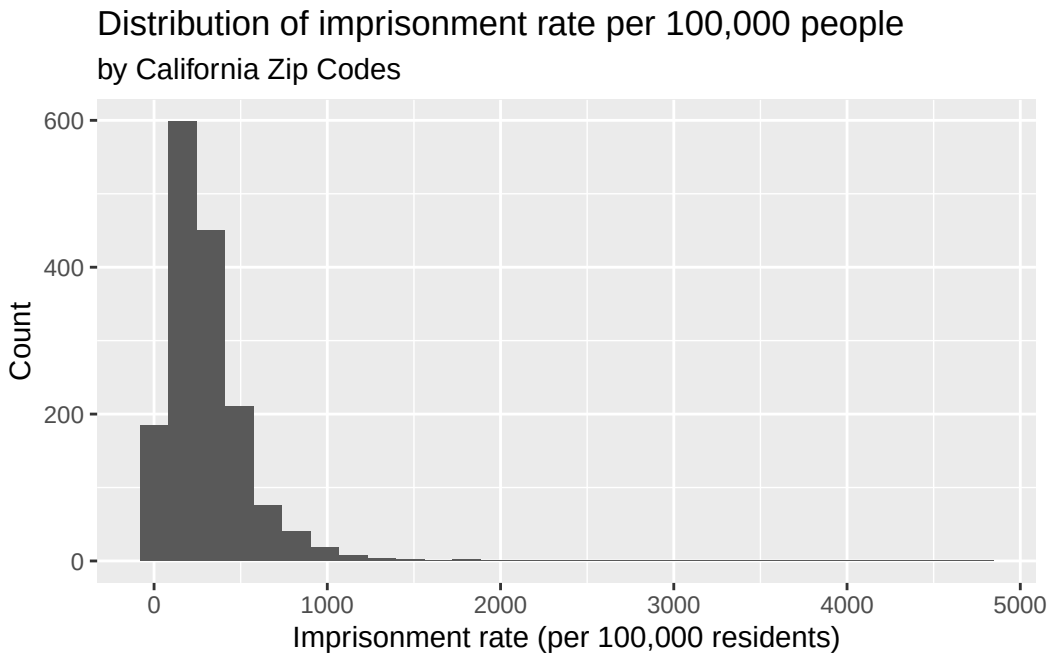
Data processing

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.

Response Variable EDA

```
joined_data_clean |>
  ggplot(aes(x = imprisonment_rate_per_100_000)) +
  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"
  )
```




```
joined_data_clean |>
  summarize(mean = mean(imprisonment_rate_per_100_000, na.rm = TRUE),
            median = median(imprisonment_rate_per_100_000, na.rm = TRUE),
            IQR = IQR(imprisonment_rate_per_100_000, na.rm = TRUE),
            sd = sd(imprisonment_rate_per_100_000, na.rm = TRUE),
            range = range(imprisonment_rate_per_100_000, na.rm = TRUE),
            min = min(imprisonment_rate_per_100_000, na.rm = TRUE),
            max = max(imprisonment_rate_per_100_000, na.rm = TRUE))
```

```
# A tibble: 2 x 7
   mean median   IQR    sd range   min   max
  <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1  300.    251  246.  252.     0     0  4762
2  300.    251  246.  252.  4762     0  4762
```

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.

Analysis approach

...

Data dictionary

The data dictionary can be found here [Update the link and remove this note!]