Final Project: California Incarceration and Socioeconomic Variables

Author: Debbie Olorunisola

Discussants: Data queried from the following sources

- $USDAPovertybyCounty$
- $CaliforniaDepartmentofCorrectionsandRehabilitationAdultInmateCharacteristics$
- $U.S.CensusBureauCountyPopulationDemographicEstimates(RaceandSex)2000-2009$
- $U.S.CensusBureauCountyPopulationDemographicEstimates(RaceandSex)2010-2019$

Introduction

For my project, I looked at California prison data. I had three main questions: * Is there a relationship between incarceration and socioeconomic factors, specifically race and poverty? * Do areas with high minority (specifically Black and Hispanic) populations tend to have higher incarcerated populations? Do they tend to be charged with certain crimes, relative to areas with wealthier and whiter populations? * Was there any significant change in incarceration levels between 2006 and 2012?

Importantly, I looked at geographic and time series data from before the legalization of marijuana in California. I wanted to see if there was any consistency over the years in arrests and if there were any years where there were significant changes in prison populations that was not necessarily explainable by an increase in overall county population.

The data set about the prison population does not disaggregate race information beyond Hispanic, Black and white, which limits analysis for other minority groups that may be disproportionately affected by policing and incarceration. It also does not contain person-specific information about racial demographics and their county of origin, which is why I want to cross-reference census data and data from the California Department of Corrections.

I found this topic interesting because California has one of the largest prison populations in the U.S. Although it has passed legislation to decriminalize marijuana, much of the prison population is still incarcerated for possession. A more interesting analysis would likely span 2000 to 2022, looking at how (1) the decriminalization of marijuana affected prison intake and the demographics of the prison population and (2) potentially analyzing the effects of COVID-19 on the prison population.

Results

Based on the data, it seems that the Hispanic and Black populations make up a disproportionate amount of the prison population in California, with the Hispanic population in California prisons being about 5.58% higher than the overall population proportion and the black population being about 19.5% higher in prisons. Due to this, I wanted to see whether these factors, or others, could explain the changes in incarceration over the years.

I found that looking at changes in the Hispanic population helped explain most of the variation in the data, relative to other factors. However, it still explained very little and there wasn't a clearly linear relationship between changes in incarceration rates and changes in the Hispanic population.

Data wrangling: California Census Data and Proportions of Incarcerated People by Year, County and Race.

I found my data about incarceration by year and county from the California Department of Corrections and Rehabilitation Adult Inmate Characteristics (CDCR). I converted the PDFs to CSVs using Excel and the cleaned them below (see "ca_prison_race" and "ca_prison_location"). There was missing information about Alpine County in 2012 because the report from 2012 omitted counties where the incarceration proportion was exactly or nearly 0. I also joined map data to the prison location dataframe so I could map the data points.

After this, I used the California census information to get data about the proportion of people by race and ethnicity in each California county (see: "ca_population_locations"). I combined two large data sets, one with population estimates between 2000 to 2010 and 2010 to 2020. Before bringing the CSV into R I queried for race and year. I extracted the years 2006, 2009 and 2012, based on their year codes in each data set. I then specifically extracted information on Non-Hispanic White people, Non-Hispanic Black people, and Hispanic people, so that I would not accidentally double-count anything. I then joined this with map data from California to visualize this later on. I turned the raw population estimates into proportions by dividing the total estimate (by race) by the total population estimate (by county).

The dataframe "ca_incarceration_proportion" contains the proportion of the population of each county that is incarcerated. to find this, I joined the prisoner location data from the CDCR with population location data and divided the raw prison population in each county by the raw population data by county.

Finally, I created "ca_poverty" which contains information about poverty by county in 2006, 2009, and 2012. I converted the TXT file into CSV using Excel and then used R to rename the columns to be simpler and match the way the columns in the other dataframes were named. I also corrected my typos for San Bernandino County.

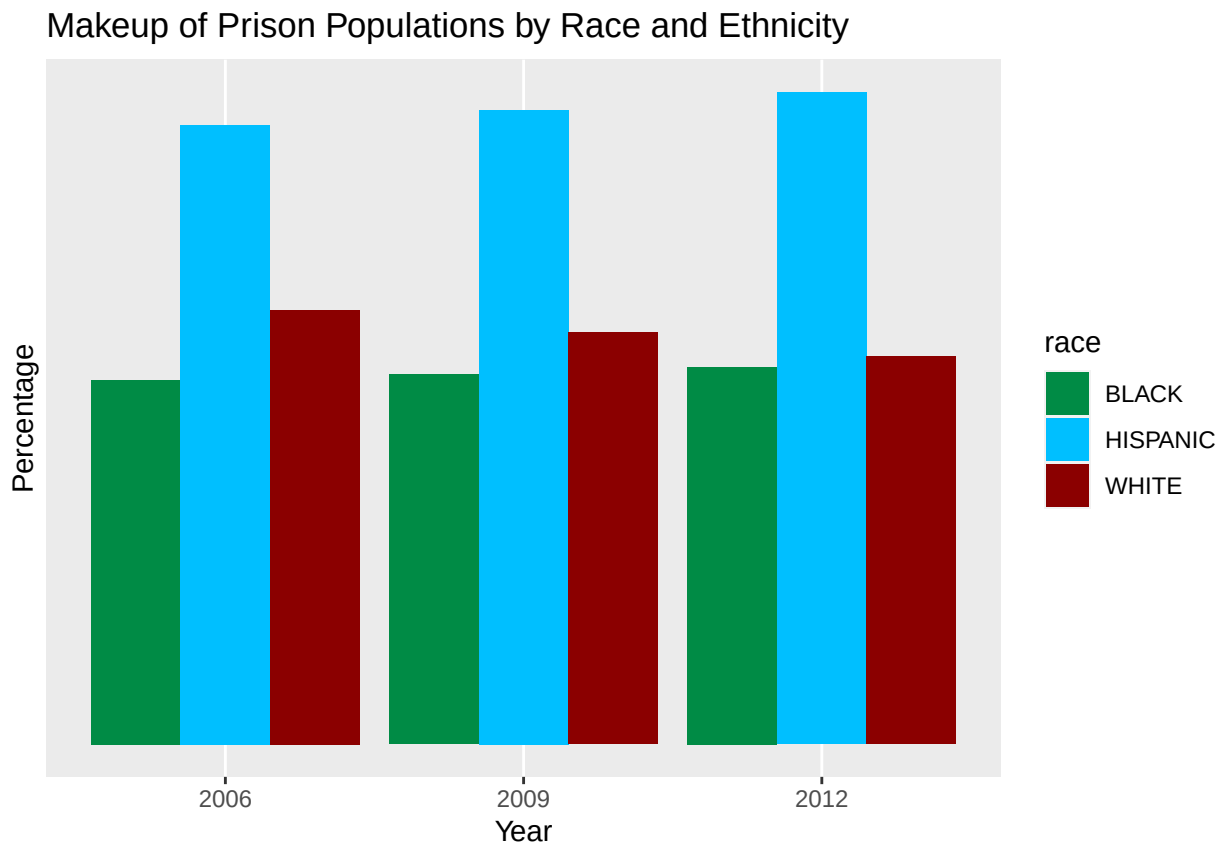
```
#CA data for prisoners by race per year
ca_prison_race <- ca_prison_race |>
   rename(race = "RACIAL.ETHNIC.GROUP", tot_percent = "TOTAL.PERCENT") |>
  filter(Year == 2006 | Year == 2009 | Year == 2012) |>
  filter(race == "BLACK" | race == "HISPANIC" | race == "WHITE") |>
  select(Year, race, tot_percent) |>
  group_by(race)

#locations of prisons
ca_prison_location <- ca_prison_location |>
  rename(tot_percent = "TOTAL.PERCENT", tot_number = "TOTAL.NUMBER", county = "COUNTY",
         year = "Year") |>
  filter((year == 2006 | year == 2009 | year == 2012) & (county != "ALL") & (county != "SUB TOTAL")) |>
```

```
  mutate(county = tolower(county)) |>
  mutate(state = "california") |>
  arrange(desc(county)) |> #alphabatized
  select(year, county, state, tot_percent, tot_number) |>
  mutate(tot_number = as.numeric(gsub(",","",tot_number)))

#0s for the missing 2012 alpine county
alpine_add <- data.frame(2012, "alpine", "california", 0, 0)
colnames(alpine_add) <- c("year", "county", "state", "tot_percent", "tot_number")

ca_prison_location <- rbind(ca_prison_location[1:20, ], alpine_add, ca_prison_location[21:173, ])

#making a big location and time-series dataset for the whole population
ca_race_demographic <- rbind(ca_race_demographic_pop0609, ca_race_demographic_pop2012)

ca_population_locations <- ca_race_demographic |>
  rename(county = CTYNAME, state = "STNAME", year = YEAR, tot_pop = TOT_POP) |>
  mutate(county = tolower(county), state = tolower(state)) |>
  mutate(year = replace(year, year == 8, 2006)) |>
  mutate(year = replace(year, year == 11, 2009)) |>
  mutate(year = replace(year, year == 5, 2012)) |>
  arrange(desc(county)) |> #alphabetized
  select(-c("AGEGRP"))

#subsetting for racial/ethnic population proportions
ca_race <- ca_population_locations |>
  group_by(year) |>
  mutate(prop_black = (NHBA_FEMALE + NHBA_MALE)/tot_pop, #NonHispanic Black
            prop_hispanic = (H_MALE + H_FEMALE)/tot_pop, #Hispanic
            prop_white = (NHWA_FEMALE + NHWA_MALE)/tot_pop) |> #NonHispanic White
  select(year, county, state, prop_black, prop_hispanic, prop_white)

#for mapping
ca_incarceration_proportion <- left_join(ca_prison_location,
                                         ca_population_locations,
                                         by = c("county" = "county",
                                                "state" = "state",
                                                "year" = "year")) |>
  mutate(prop_population = tot_number/tot_pop) #proportion of the population of each region that is inc

#also for mapping
ca_map_data <- map_data("county") |>
  filter(region == "california")

#poverty info

ca_poverty <- ca_overall_poverty |>
  select("Year", "Estimated.percent.of.people.of.all.ages.in.poverty", "County", "State") |>
  rename("year" = "Year", "percent_poverty" = "Estimated.percent.of.people.of.all.ages.in.poverty", "co
  filter(!(county == "")) |>
  arrange(desc(county)) |>
  mutate(state = "california", county = tolower(county),
         percent_poverty = percent_poverty/100)
```

```
  #change percents to actual percents

ca_poverty$county[ca_poverty$county == "san bernandino"] <- "san bernardino"
ca_poverty$county[ca_poverty$county == "san bernardino "] <- "san bernardino"
#misspelled san bernandino ._.
```

Visualize the data: Change the subtitle here to describe what you are plotting etc.

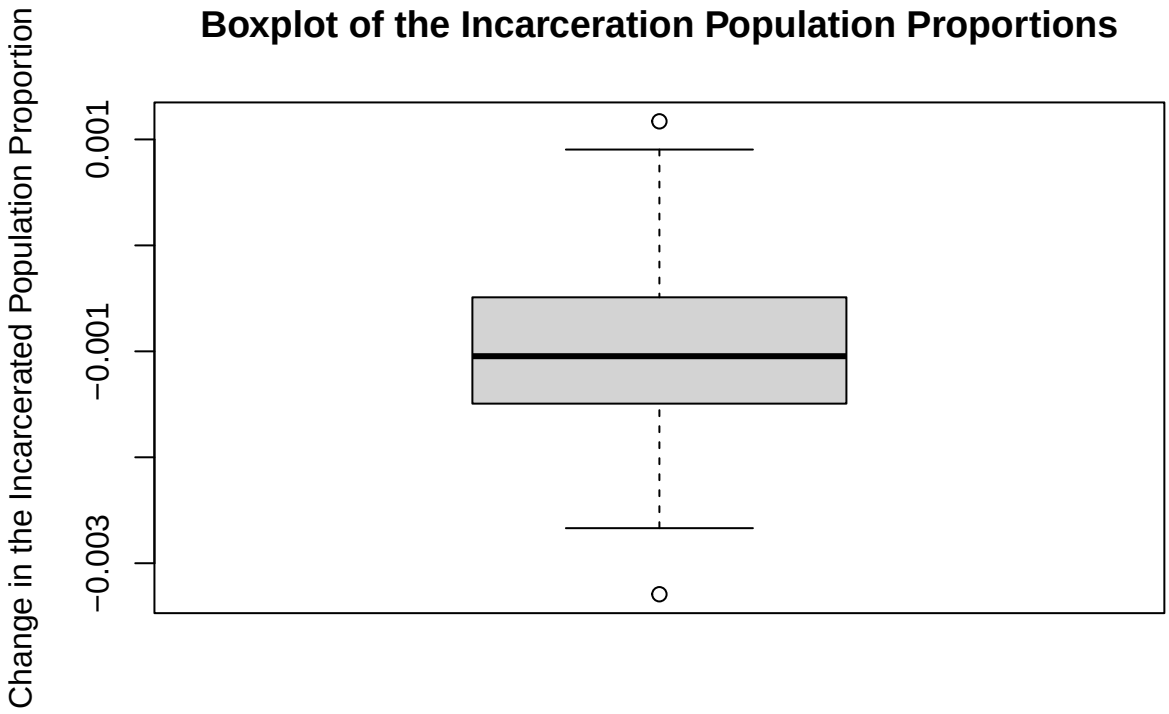
```
# graph of imprisonment by race
ca_prison_race |>
    ggplot(aes(x = as.factor(Year), y = tot_percent, fill = race)) +
    geom_bar(stat = "identity", position = position_dodge()) +
    scale_y_continuous(breaks = c(2006, 2009, 2012)) + scale_fill_manual(values = c("springgreen4",
    "deepskyblue", "darkred")) + ggtitle("Makeup of Prison Populations by Race and Ethnicity") +
    xlab("Year") + ylab("Percentage")
```




```
# Maps can be viewed in the appendix maps of proportion
# incarcerated by county and year 2006
ca_06 <- ca_incarceration_proportion |>
    filter(year == 2006) |>
```

```
    select("county", "state", "prop_population")

# 2009
ca_09 <- ca_incarceration_proportion |>
    filter(year == 2009) |>
    select("county", "state", "prop_population")

# 2012
ca_12 <- ca_incarceration_proportion |>
    filter(year == 2012) |>
    select("county", "state", "prop_population")

# maps of proportions of race/ethnicity in each place by
# year

# BLACK, 2006
ca_black_06 <- ca_race |>
    filter(year == 2006) |>
    select("county", "state", "prop_black")

# BLACK, 2009
ca_black_09 <- ca_race |>
    filter(year == 2009) |>
    select("county", "state", "prop_black")

# BLACK, 2012
ca_black_12 <- ca_race |>
    filter(year == 2012) |>
    select("county", "state", "prop_black")

# WHITE, 2006
ca_white_06 <- ca_race |>
    filter(year == 2006) |>
    select("county", "state", "prop_white")

# WHITE, 2009
ca_white_09 <- ca_race |>
    filter(year == 2009) |>
    select("county", "state", "prop_white")

# WHITE, 2012
ca_white_12 <- ca_race |>
    filter(year == 2012) |>
    select("county", "state", "prop_white")

# HISPANIC, 2006
ca_hispanic_06 <- ca_race |>
    filter(year == 2006) |>
    select("county", "state", "prop_hispanic")

# HISPANIC, 2009
ca_hispanic_09 <- ca_race |>
    filter(year == 2009) |>
```

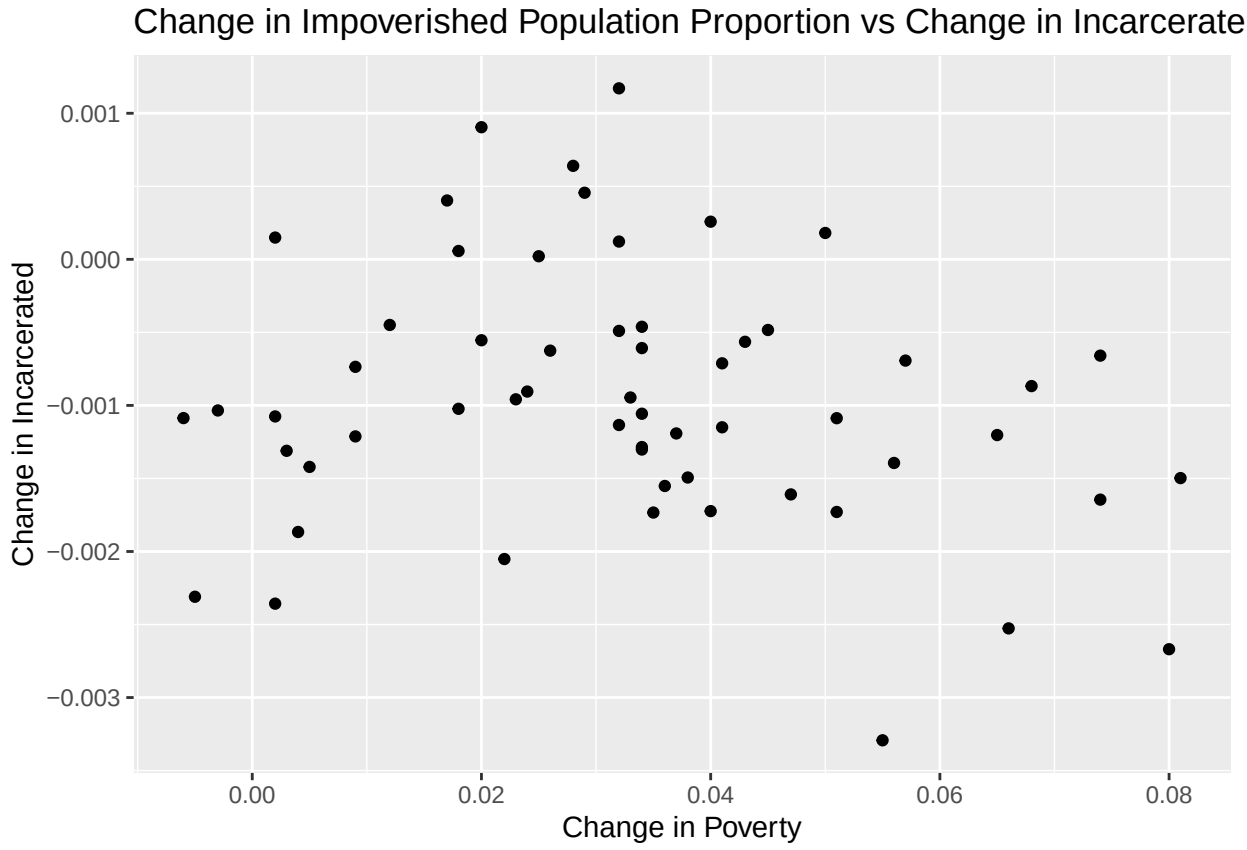
```
    select("county", "state", "prop_hispanic")

# HISPANIC, 2012
ca_hispanic_12 <- ca_race |>
    filter(year == 2012) |>
    select("county", "state", "prop_hispanic")

# looking at poverty by location 2006
ca_poverty_06 <- ca_poverty |>
    filter(year == 2006) |>
    select("county", "state", "percent_poverty")

# 2009
ca_poverty_09 <- ca_poverty |>
    filter(year == 2009) |>
    select("county", "state", "percent_poverty")

# 2012
ca_poverty_12 <- ca_poverty |>
    filter(year == 2012) |>
    select("county", "state", "percent_poverty")

# i want to look at change over time i verified that the
# counties were lined up properly with ca_*_12$county ==
# ca_*_06$county i also alphabetized each data set by
# county just in case!
change_in_poverty <- ca_poverty_12$percent_poverty - ca_poverty_06$percent_poverty

change_in_black <- ca_black_12$prop_black - ca_black_06$prop_black

change_in_hispanic <- ca_hispanic_12$prop_hispanic - ca_hispanic_06$prop_hispanic

change_in_white <- ca_white_12$prop_white - ca_white_06$prop_white

change_in_incarcerated <- ca_12$prop_population - ca_06$prop_population

variable_changes <- data.frame(unique(ca_map_data$subregion),
    change_in_incarcerated, change_in_poverty, change_in_white,
    change_in_hispanic, change_in_black)

# i want to see if there are any outliers in terms of
# change
boxplot(variable_changes$change_in_incarcerated, ylab = "Change in the Incarcerated Population Proportio
    main = "Boxplot of the Incarceration Population Proportions")
```



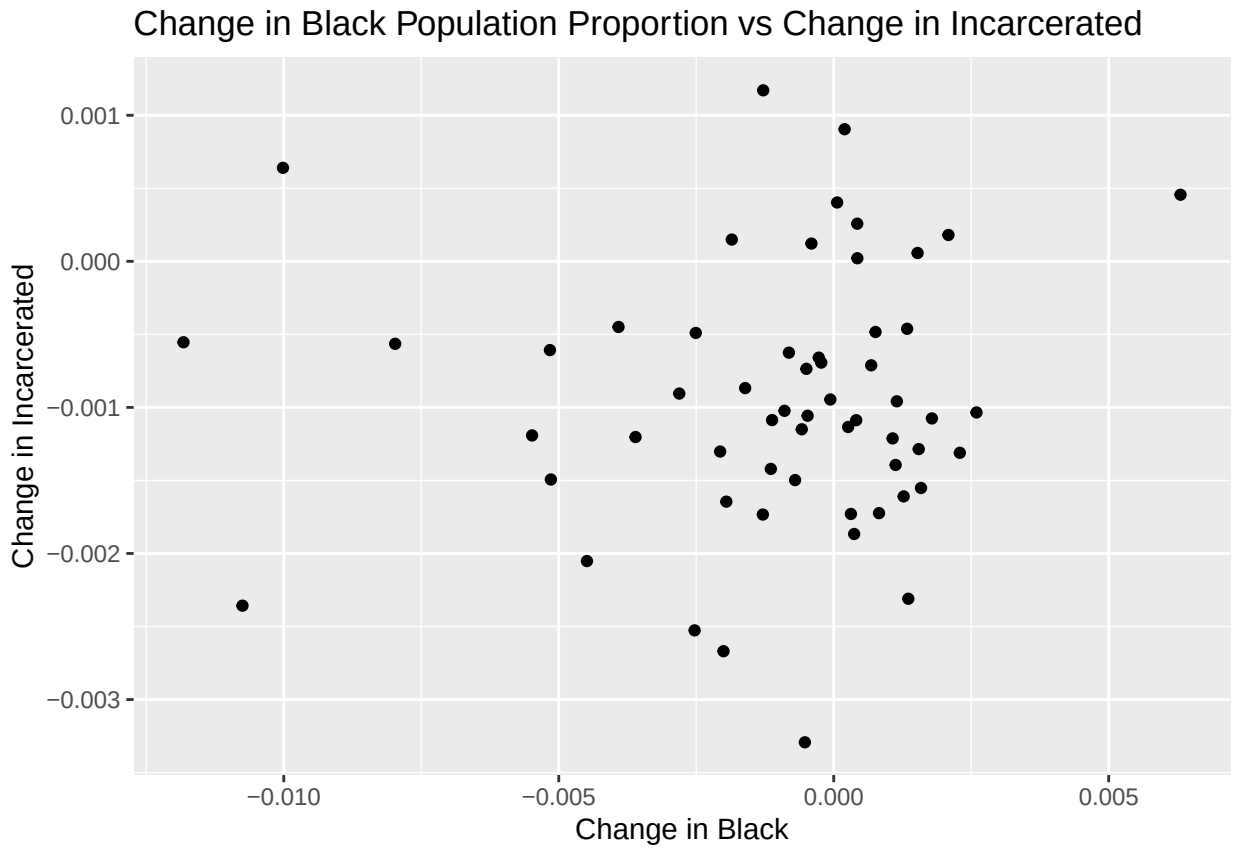

```
# Which ones are the outliers
outliers_county_incarc <- boxplot.stats(variable_changes$change_in_incarcerated)$out
indices <- which(variable_changes$change_in_incarcerated %in%
    outliers_county_incarc)
variable_changes[indices, ]
```

```
##    unique.ca_map_data.subregion. change_in_incarcerated change_in_poverty
## 8                      del norte           -0.003292651             0.055
## 52                        tehama            0.001170678             0.032
##    change_in_white change_in_hispanic change_in_black
## 8      -0.04638937         0.02547394   -0.0005246022
## 52     -0.04645632         0.02408634   -0.0012821185
```

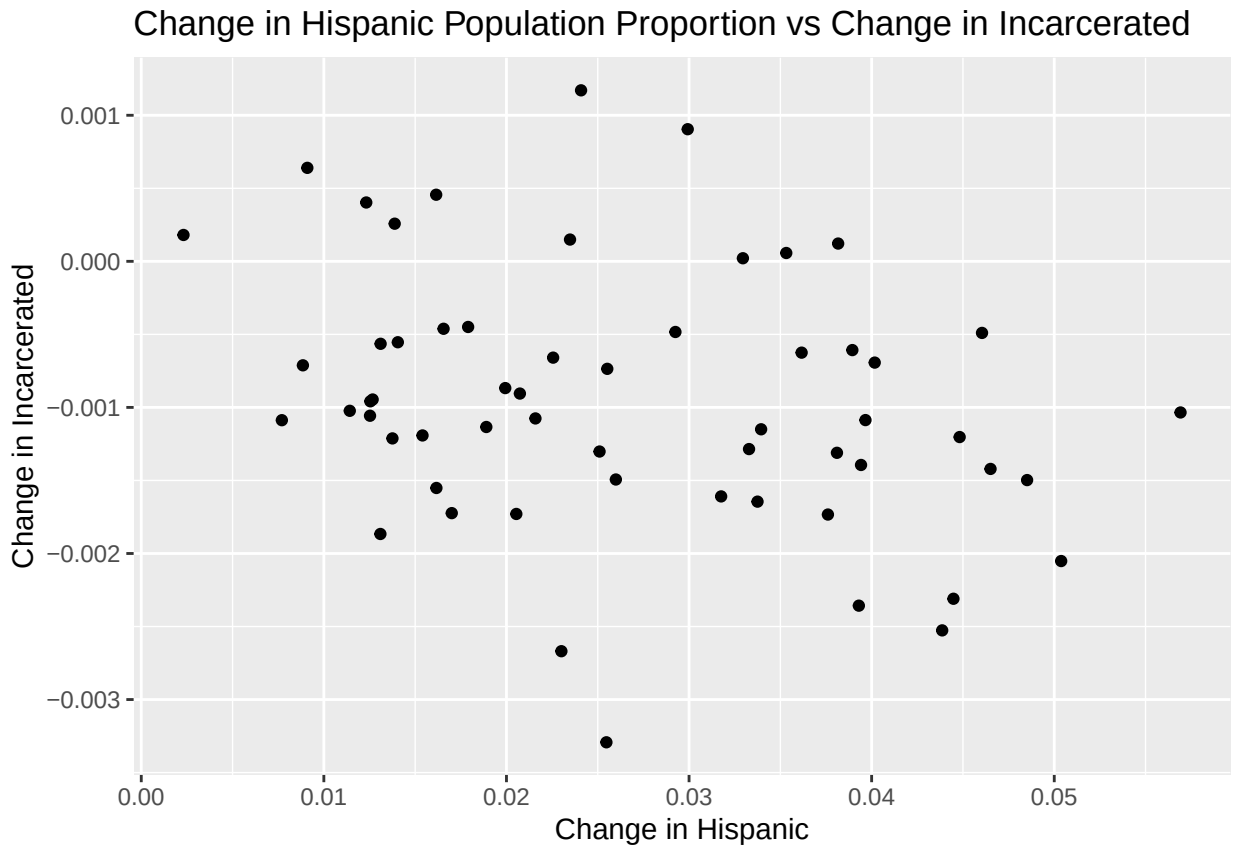
```
# plotting variables relative to changes in incarceration

variable_changes |>
    ggplot(aes(x = change_in_poverty, y = change_in_incarcerated)) +
    geom_point() + ggtitle("Change in Impoverished Population Proportion vs Change in Incarcerated") +
    xlab("Change in Poverty") + ylab("Change in Incarcerated")
```



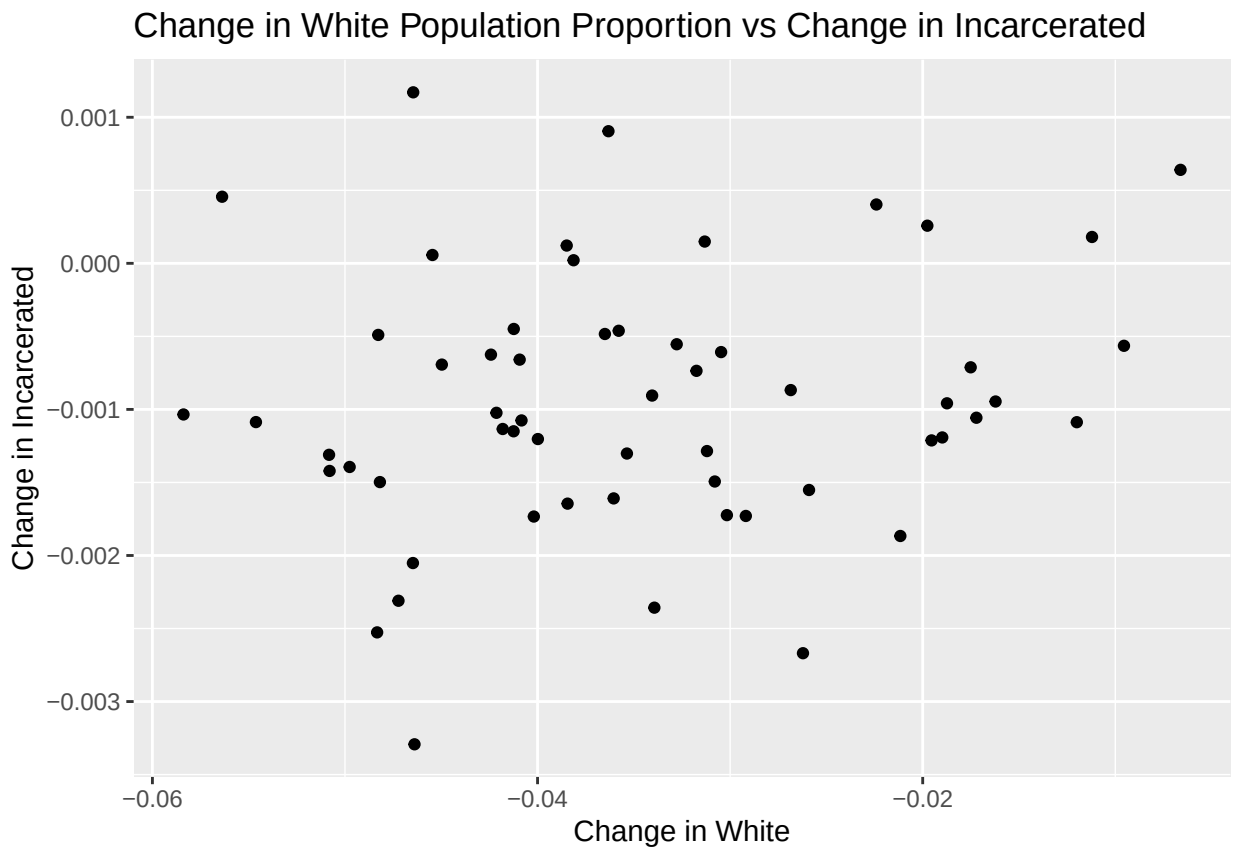

```
variable_changes |>
    ggplot(aes(x = change_in_black, y = change_in_incarcerated)) +
    geom_point() + ggtitle("Change in Black Population Proportion vs Change in Incarcerated") +
    xlab("Change in Black") + ylab("Change in Incarcerated")
```




```
variable_changes |>
    ggplot(aes(x = change_in_hispanic, y = change_in_incarcerated)) +
    geom_point() + ggtitle("Change in Hispanic Population Proportion vs Change in Incarcerated") +
    xlab("Change in Hispanic") + ylab("Change in Incarcerated")
```




```
# note that only the white proportion decreased in this
# time frame, so the plot starts with negative values
variable_changes |>
    ggplot(aes(x = change_in_white, y = change_in_incarcerated)) +
    geom_point() + ggtitle("Change in White Population Proportion vs Change in Incarcerated") +
    xlab("Change in White") + ylab("Change in Incarcerated")
```




```
# no clear relationship for any variable!
```

Discussing the Visualizations

From the visualizations, I came away with many questions. However, the lack of disaggregated data limits the analyses I can conduct. I will assess the following question, based on the scatterplot: is there a relationship between socioeconomic factors (specifically race, ethnicity, and poverty) and incarceration in California?

Analysis: Linear Models for Different Variables

Here I will answer the following question: Is there a relationship between socioeconomic factors (specifically race, ethnicity, and poverty) and incarceration?

To answer this, I generated linear models. All the data points are normally-distributed around the model, so it seems that a linear model was the appropriate choice. (See appendix for QQPlots.)

To evaluate the correlation, I compared the adjusted R-squared value when different socioeconomic factors are considered to see which one seems to explain the most of the difference in the data. I also looked at the p-value to see which factors are the most statistically significant. I looked for a p-value of $\alpha \leq 0.05$.

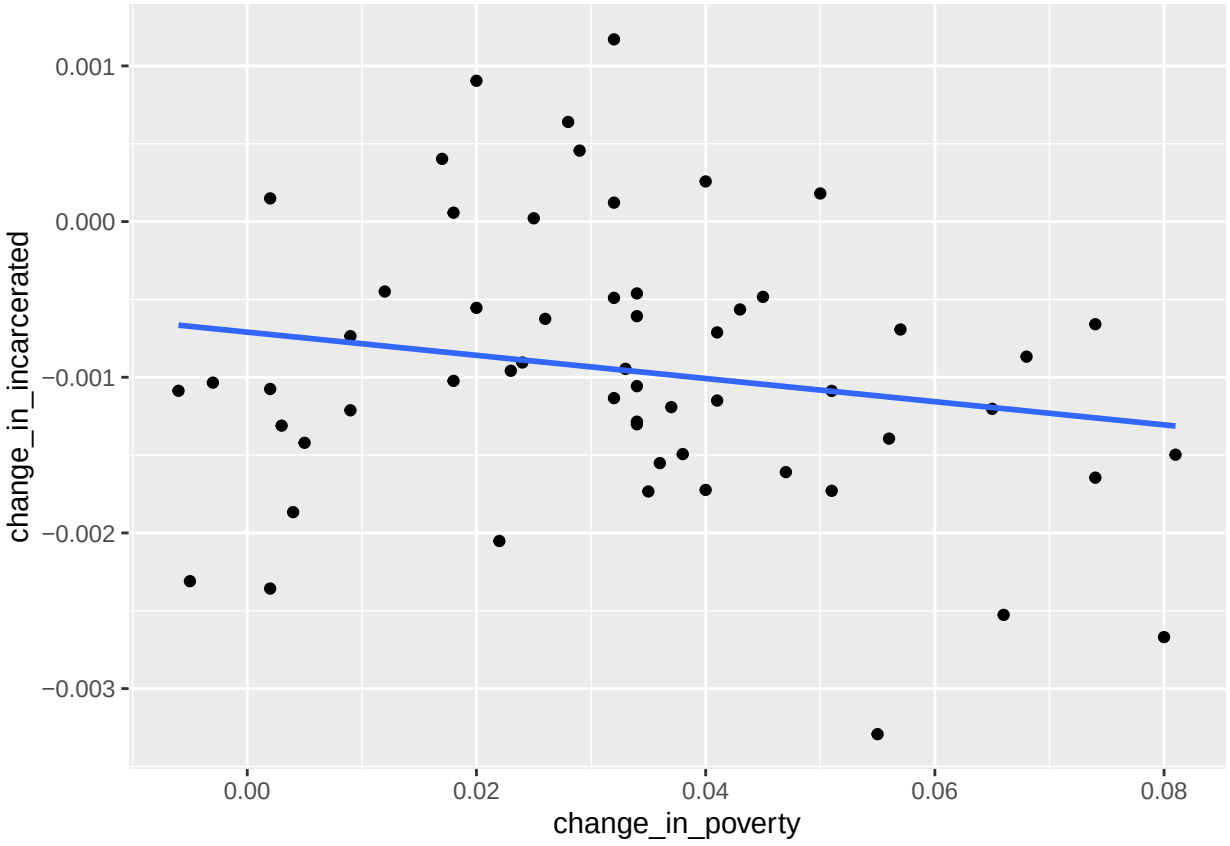
Most of the socioeconomic variables did not show a significant correlation between change in incarceration and change in the variable of interest. The only variable with that met the condition for the alpha level was the "change in Hispanic" variable. The adjusted R-squared was 0.06537, which, though small, was the

highest amount of variation explained for all variables. Since the R-squared value is positive, it seems that, when the Hispanic population increases, there is an increase in incarceration. The p-value was 0.02955, which is less than the alpha level of 0.05. Due to this, there is a low probability that this relationship is only observable due to random chance.

```
# Linear Model for Change in Poverty versus Incarceration
lm_poverty <- lm(change_in_incarcerated ~ change_in_poverty,
    data = variable_changes)
summary(lm_poverty)
```

```
##
## Call:
## lm(formula = change_in_incarcerated ~ change_in_poverty, data = variable_changes)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -2.173e-03 -4.841e-04 -8.538e-05  4.639e-04  2.119e-03
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -0.0007101  0.0002094  -3.391  0.00128 **
## change_in_poverty -0.0074449  0.0053200  -1.399  0.16720
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0008779 on 56 degrees of freedom
## Multiple R-squared:  0.03379,    Adjusted R-squared:  0.01654
## F-statistic: 1.958 on 1 and 56 DF,  p-value: 0.1672
```

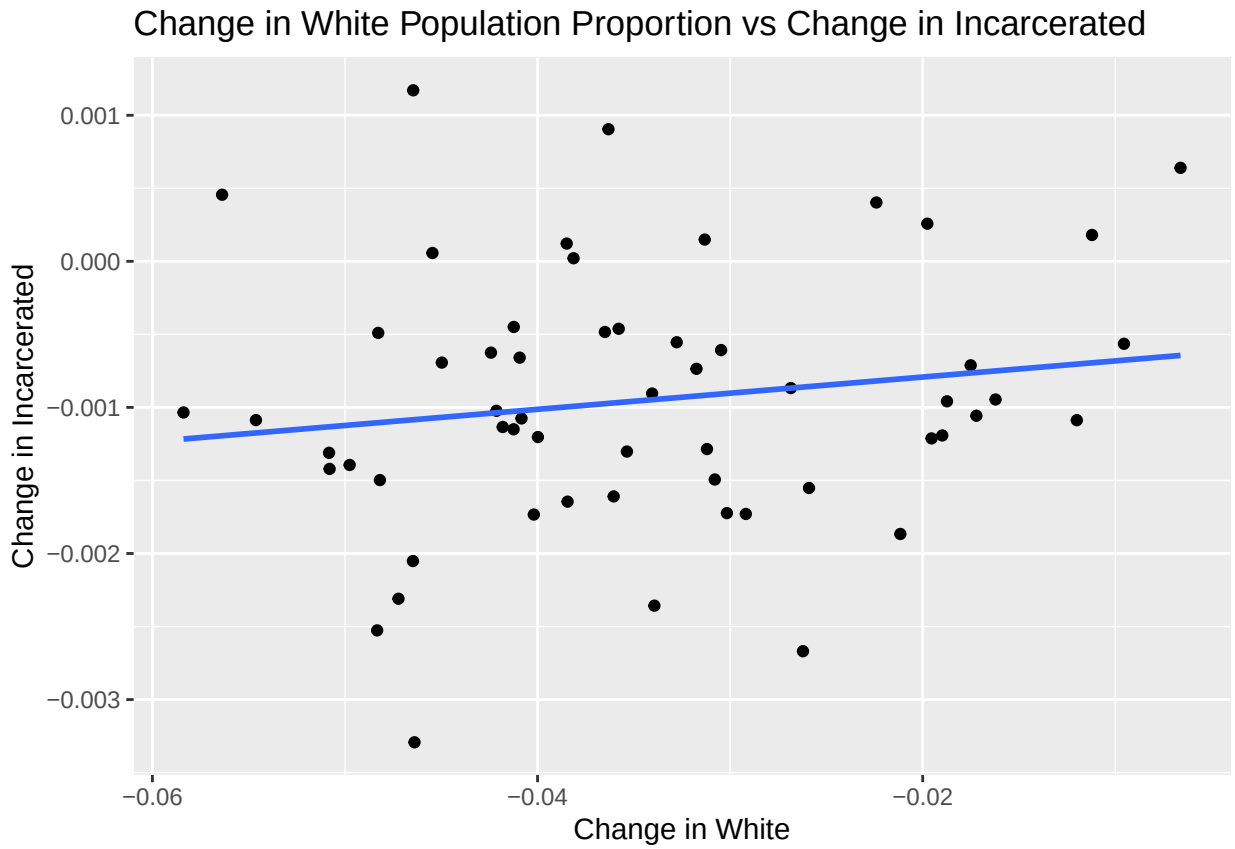
```
ggplot(variable_changes, aes(x = change_in_poverty, y = change_in_incarcerated)) +
    geom_point() + geom_smooth(method = "lm", se = FALSE)
```




```
# Linear Model for Changes in Race and Ethnicity as
# Predictors white
lm_white <- lm(change_in_incarcerated ~ change_in_white, data = variable_changes)
summary(lm_white)
```

```
##
## Call:
## lm(formula = change_in_incarcerated ~ change_in_white, data = variable_changes)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -2.209e-03 -4.217e-04 -7.703e-05  4.718e-04  2.255e-03
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -0.0005711  0.0003429  -1.665    0.101
## change_in_white  0.0110510  0.0092947   1.189    0.239
##
## Residual standard error: 0.000882 on 56 degrees of freedom
## Multiple R-squared:  0.02462,    Adjusted R-squared:  0.007204
## F-statistic: 1.414 on 1 and 56 DF,  p-value: 0.2395
```

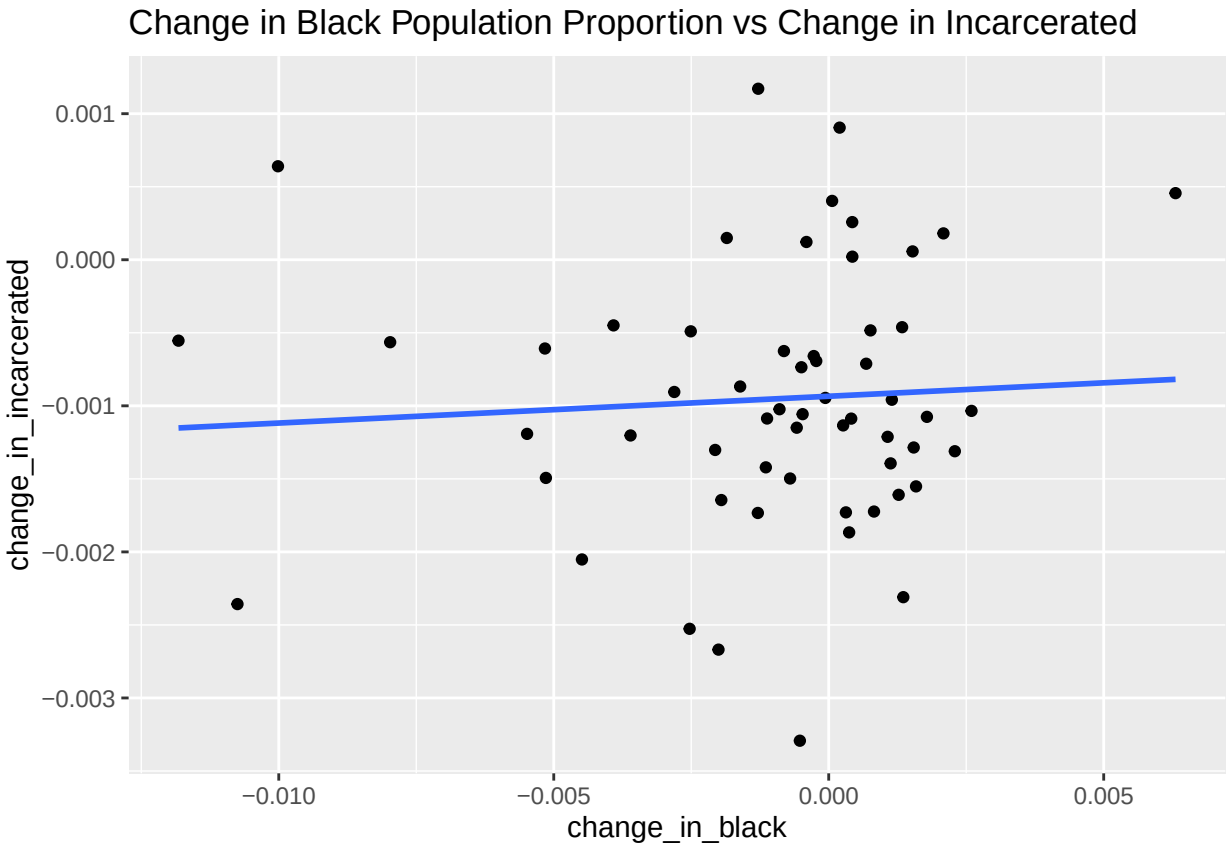
```
ggplot(variable_changes, aes(x = change_in_white, y = change_in_incarcerated)) +
    geom_point() + geom_smooth(method = "lm", se = FALSE) + ggtitle("Change in White Population Proporti
    xlab("Change in White") + ylab("Change in Incarcerated")
```




```
# black
lm_black <- lm(change_in_incarcerated ~ change_in_black, data = variable_changes)
summary(lm_black)
```

```
##
## Call:
## lm(formula = change_in_incarcerated ~ change_in_black, data = variable_changes)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0023486 -0.0004767 -0.0001229  0.0004799  0.0021287
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -0.0009345  0.0001237  -7.556 4.17e-10 ***
## change_in_black  0.0183537  0.0362023   0.507    0.614
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.000891 on 56 degrees of freedom
## Multiple R-squared:  0.004569,   Adjusted R-squared:  -0.01321
## F-statistic: 0.257 on 1 and 56 DF,  p-value: 0.6142
```

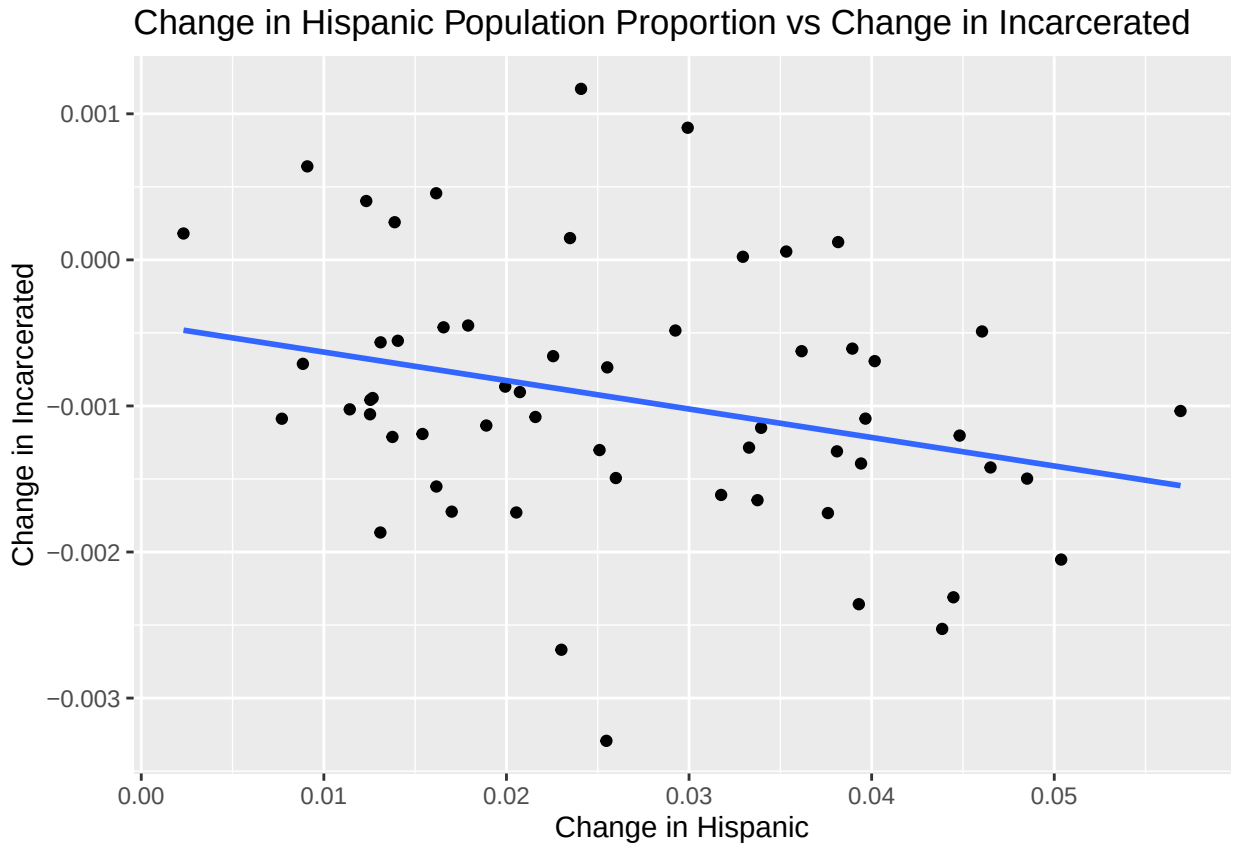
```
ggplot(variable_changes, aes(x = change_in_black, y = change_in_incarcerated)) +
    geom_point() + geom_smooth(method = "lm", se = FALSE) + labs(title = "Change in Black Population Pr
    xlab = "Change in Black", ylab = "Change in Incarcerated")
```




```
# Hispanic
lm_hispanic <- lm(change_in_incarcerated ~ change_in_hispanic,
    data = variable_changes)
summary(lm_hispanic)
```

```
##
## Call:
## lm(formula = change_in_incarcerated ~ change_in_hispanic, data = variable_changes)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -2.360e-03 -5.061e-04 -9.061e-05  5.212e-04  2.077e-03
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -0.0004364  0.0002579  -1.692   0.0962 .
## change_in_hispanic -0.0194903  0.0087277  -2.233   0.0296 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0008558 on 56 degrees of freedom
## Multiple R-squared:  0.08177,    Adjusted R-squared:  0.06537
## F-statistic: 4.987 on 1 and 56 DF,  p-value: 0.02955
```

```
ggplot(variable_changes, aes(x = change_in_hispanic, y = change_in_incarcerated)) +
    geom_point() + geom_smooth(method = "lm", se = FALSE) + ggtitle("Change in Hispanic Population Prop
    xlab("Change in Hispanic") + ylab("Change in Incarcerated")
```



Conclusion

Through this project, I found that an increase in Hispanic populations in California explained the most variance in the data (with the factors considered being race/ethncity and poverty). An increase in the Hispanic population was associated with increases in rates of incarceration. More thorough analysis, especially with data about the types of crimes that different populations are arrested for and differences in police presence in certain areas would likely help explain why some populations are overrepresented in prisons.

There is more recent data available on the CDCR website in Tableau that could guide future inquiry. I would have to request in order to download the data, but I think it would be interesting to see larger data points about change over time. I don't think there is any disaggregated data on the website, but I also think *that* would lead to valuable analysis because it would give me information on the county of origin, violation type, income level, and race of each specific person in jail/prison. This would allow me to conduct logistic regression analysis for some of the categorical factors and find more specific associations between variables.

Reflection

I found it interesting to visualize the socioeconomic variables of interest and think that I would enjoy turning them into gifs that used data from more years that I was able to in this project. I also found the table that Professor Meyers gave us from the midterm review about what test to use and when very helpful throughout the process. In terms of things that went wrong, once again, the lack of disaggregated data made it difficult to answer many of the questions I had. For example, I attempted to calculate whether there was a disproportionate amount of Black and Hispanic people in prisons/jail, using a parametric method, but I think a permutation method would have worked better. I also wanted to use observations from the maps to answer more questions, like:

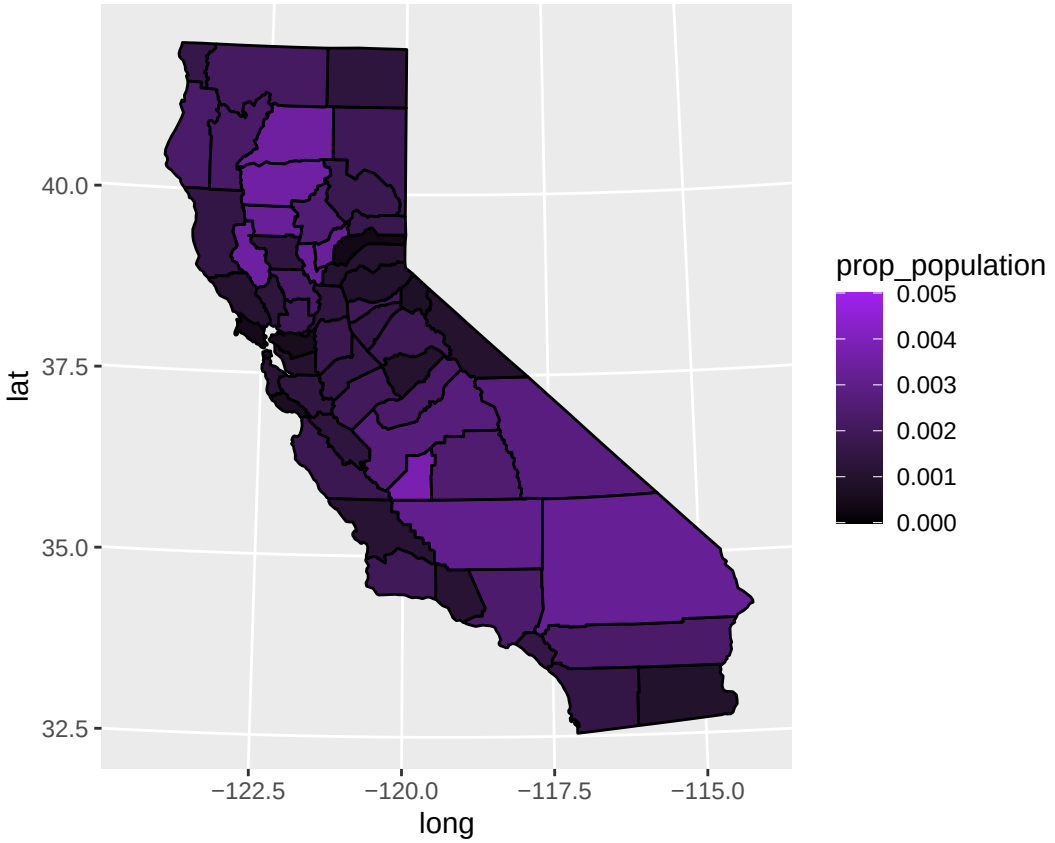
- Is Kings County's incarceration rate significantly higher than other California counties?
- Since Del Norte and Tehama counties are outliers in terms of change in incarceration, is this change in incarceration rate explained by any other variable?

The years I observed were chosen kind of arbitrarily, since I had to clean the data myself. If I had more time, I would use more years that the CDCR covers, in order to potentially conduct a more thorough analysis.

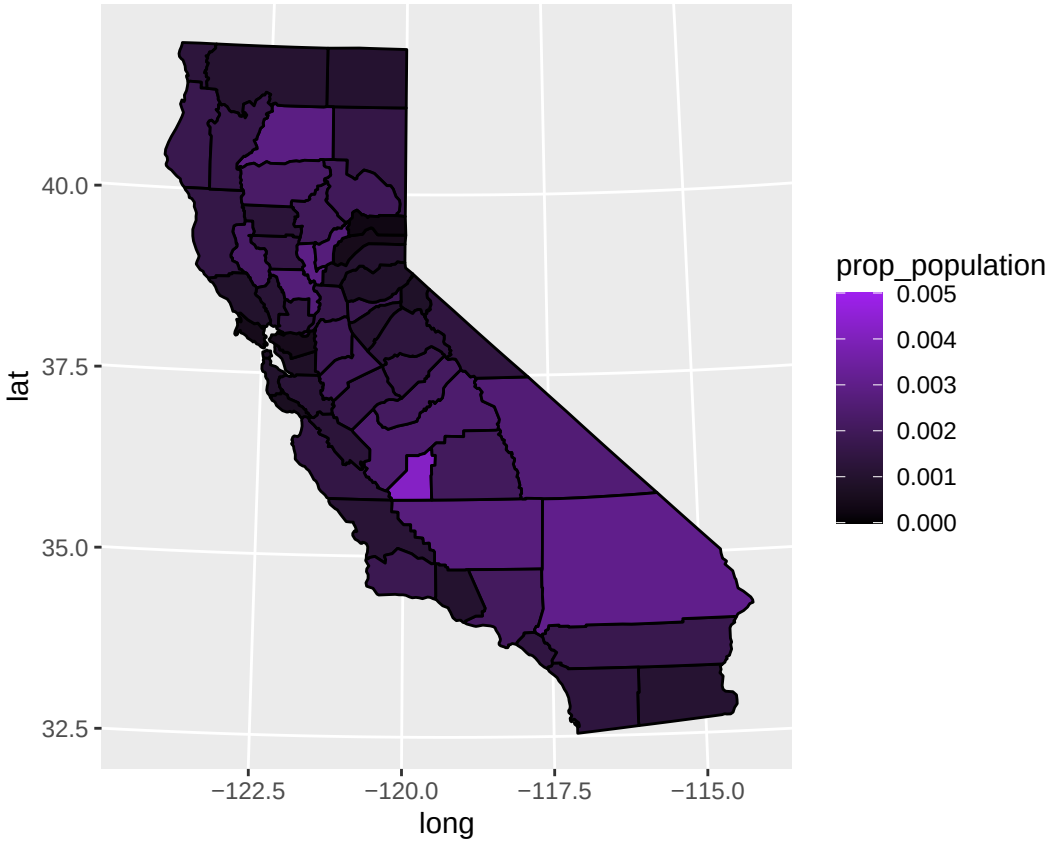
I spent about 15-20 hours on this project.

Appendix: Maps

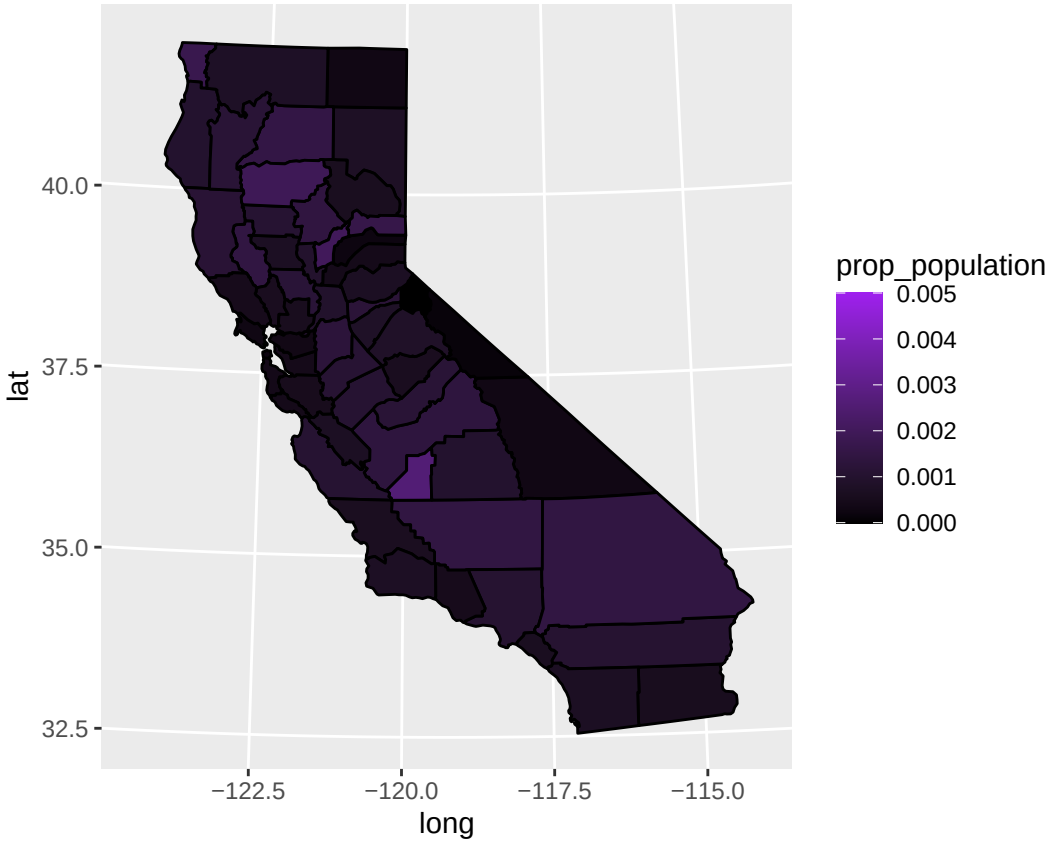
```
# maps of proportion incarcerated by county and year 2006

ca_map_pop_06 <- left_join(ca_map_data, ca_06, by = c(subregion = "county",
    region = "state"))

ggplot(ca_map_pop_06, aes(x = long, y = lat, group = group, fill = prop_population)) +
    geom_polygon(color = "black") + coord_map("polyconic") +
    scale_fill_gradient(low = "black", high = "purple", limits = c(0,
        0.005), oob = scales::squish) + scale_color_gradient(low = "black",
    high = "purple", limits = c(0, 0.005), oob = scales::squish)
```



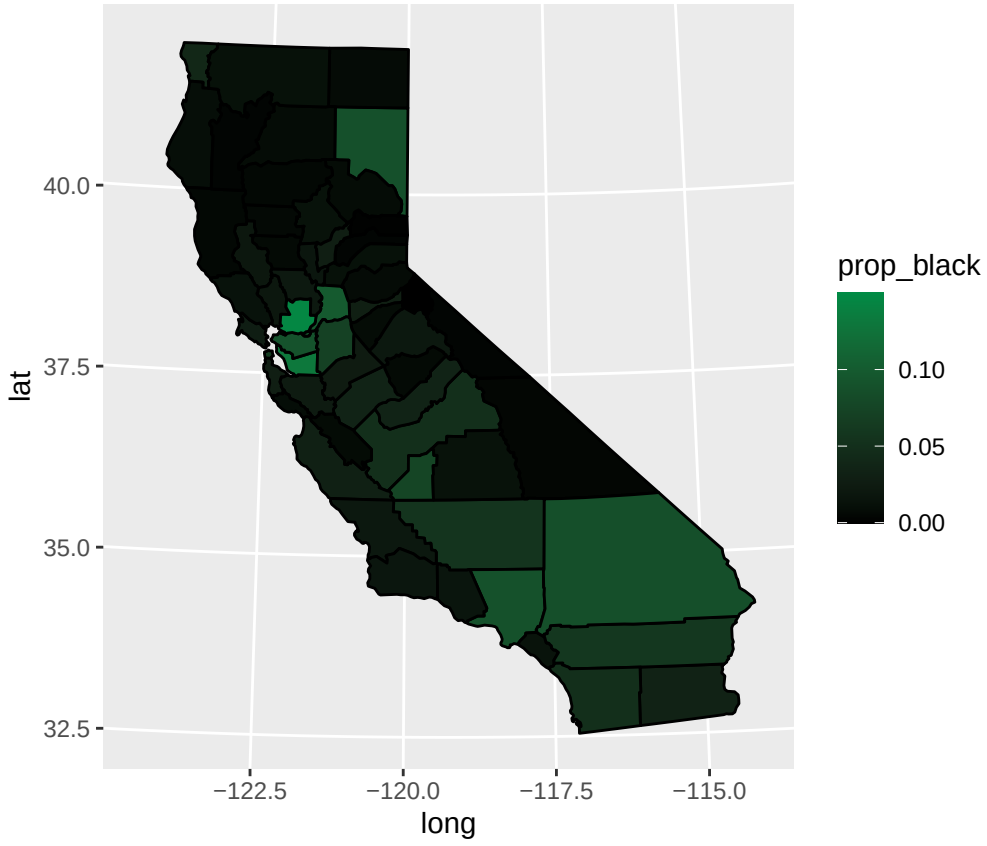

```
# 2009

ca_map_pop_09 <- left_join(ca_map_data, ca_09, by = c(subregion = "county",
    region = "state"))

ggplot(ca_map_pop_09, aes(x = long, y = lat, group = group, fill = prop_population)) +
    geom_polygon(color = "black") + coord_map("polyconic") +
    scale_fill_gradient(low = "black", high = "purple", limits = c(0,
        0.005), oob = scales::squish) + scale_color_gradient(low = "black",
    high = "purple", limits = c(0, 0.005), oob = scales::squish)
```



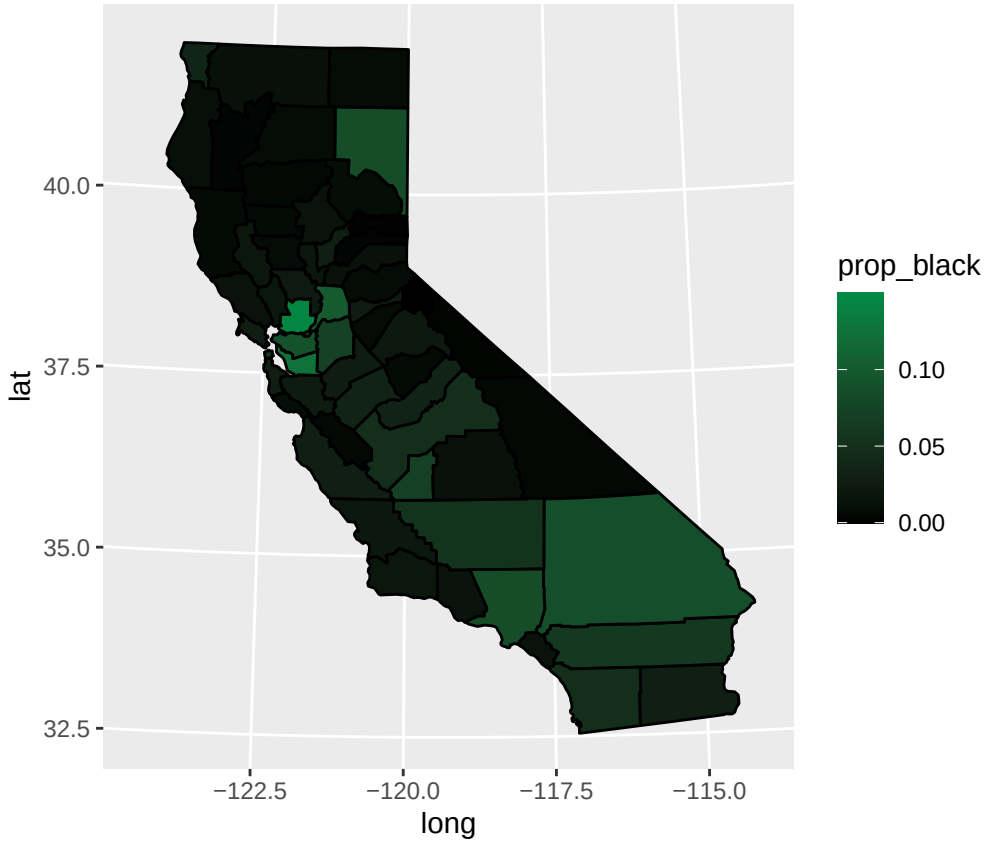

```
# 2012

ca_map_pop_12 <- left_join(ca_map_data, ca_12, by = c(subregion = "county",
    region = "state"))

ggplot(ca_map_pop_12, aes(x = long, y = lat, group = group, fill = prop_population)) +
    geom_polygon(color = "black") + coord_map("polyconic") +
    scale_fill_gradient(low = "black", high = "purple", limits = c(0,
        0.005), oob = scales::squish) + scale_color_gradient(low = "black",
    high = "purple", limits = c(0, 0.005), oob = scales::squish)
```



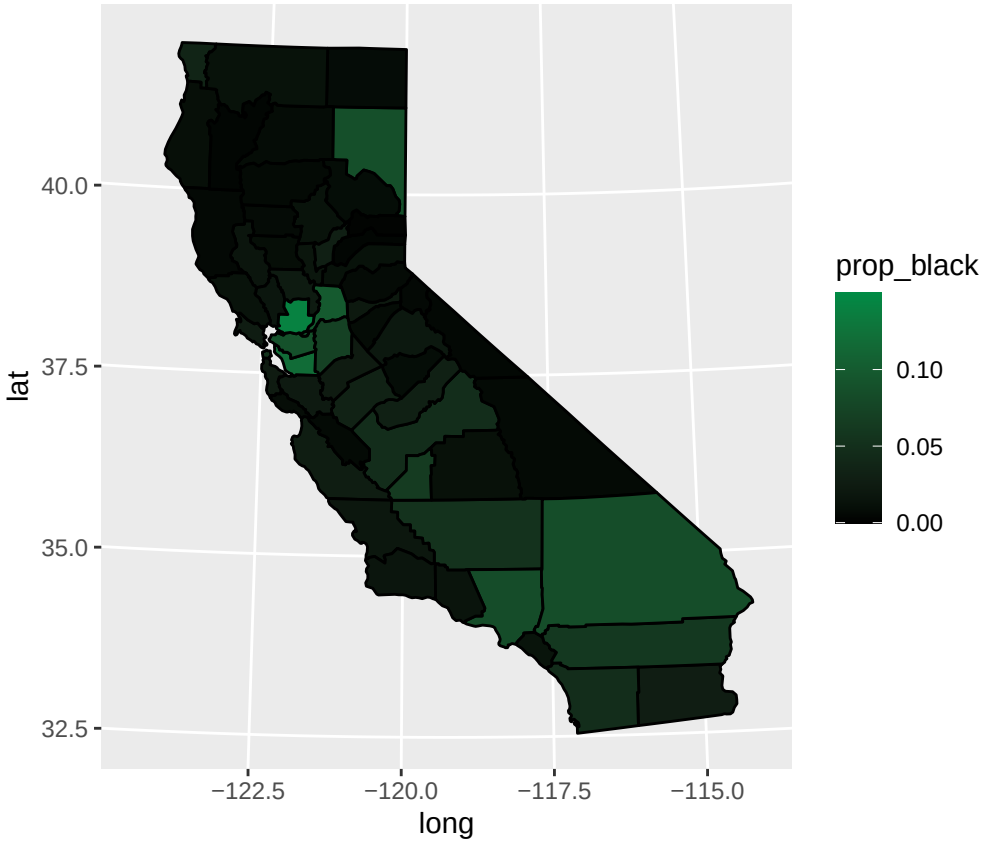

```
# maps of proportions of race/ethnicity in each place by
# year

# BLACK, 2006
ca_map_black_06 <- left_join(ca_map_data, ca_black_06, by = c(subregion = "county",
    region = "state"))

ggplot(ca_map_black_06, aes(x = long, y = lat, group = group,
    fill = prop_black)) + geom_polygon(color = "black") + coord_map("polyconic") +
    scale_fill_gradient(low = "black", high = "springgreen4",
        limits = c(0, 0.15), oob = scales::squish) + scale_color_gradient(low = "black",
    high = "springgreen4", limits = c(0, 0.15), oob = scales::squish)
```



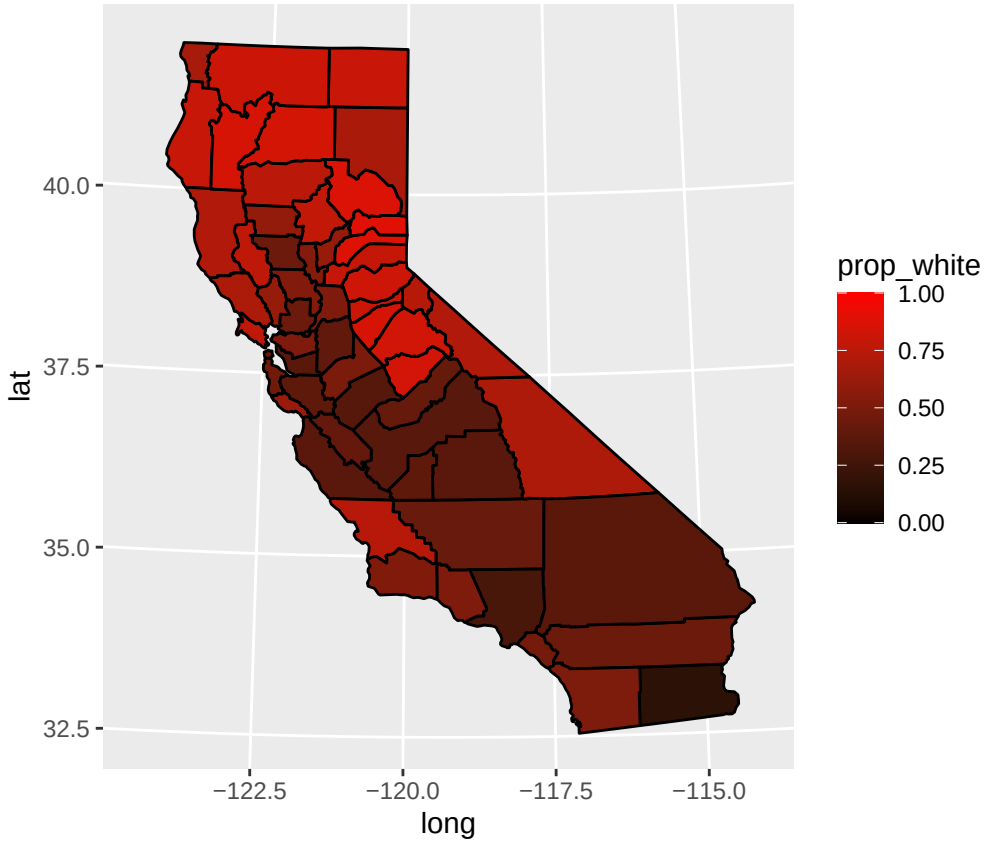

```
# BLACK, 2009

ca_map_black_09 <- left_join(ca_map_data, ca_black_09, by = c(subregion = "county",
    region = "state"))

ggplot(ca_map_black_09, aes(x = long, y = lat, group = group,
    fill = prop_black)) + geom_polygon(color = "black") + coord_map("polyconic") +
    scale_fill_gradient(low = "black", high = "springgreen4",
        limits = c(0, 0.15), oob = scales::squish) + scale_color_gradient(low = "black",
    high = "springgreen4", limits = c(0, 0.15), oob = scales::squish)
```



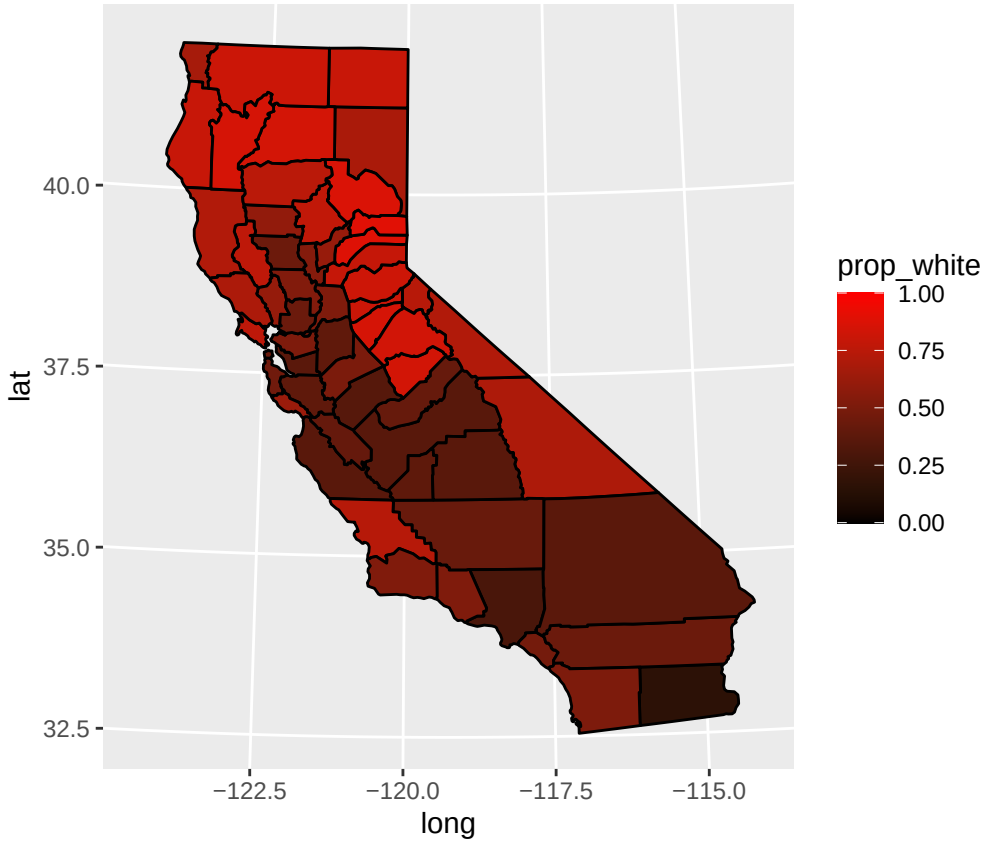

```
# BLACK, 2012

ca_map_black_12 <- left_join(ca_map_data, ca_black_12, by = c(subregion = "county",
    region = "state"))

ggplot(ca_map_black_12, aes(x = long, y = lat, group = group,
    fill = prop_black)) + geom_polygon(color = "black") + coord_map("polyconic") +
    scale_fill_gradient(low = "black", high = "springgreen4",
        limits = c(0, 0.15), oob = scales::squish) + scale_color_gradient(low = "black",
    high = "springgreen4", limits = c(0, 0.15), oob = scales::squish)
```



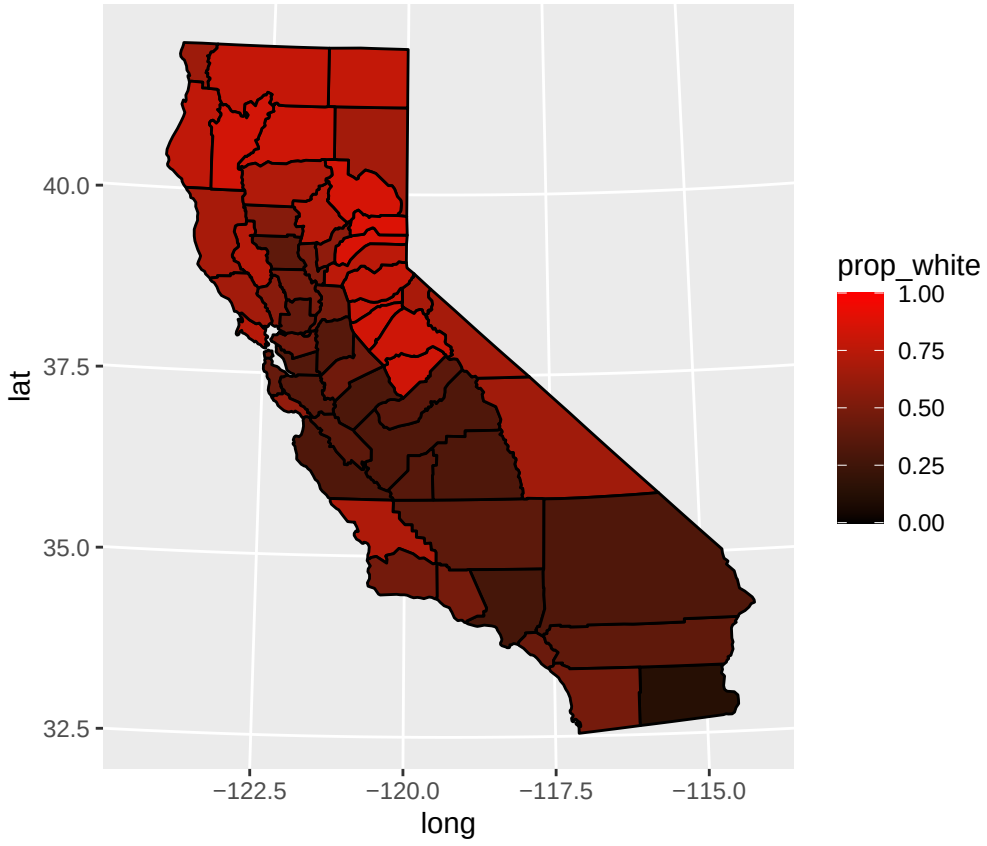

```
# WHITE, 2006

ca_map_white_06 <- left_join(ca_map_data, ca_white_06, by = c(subregion = "county",
    region = "state"))

ggplot(ca_map_white_06, aes(x = long, y = lat, group = group,
    fill = prop_white)) + geom_polygon(color = "black") + coord_map("polyconic") +
    scale_fill_gradient(low = "black", high = "red", limits = c(0,
        1), oob = scales::squish) + scale_color_gradient(low = "black",
    high = "red", limits = c(0, 1), oob = scales::squish)
```



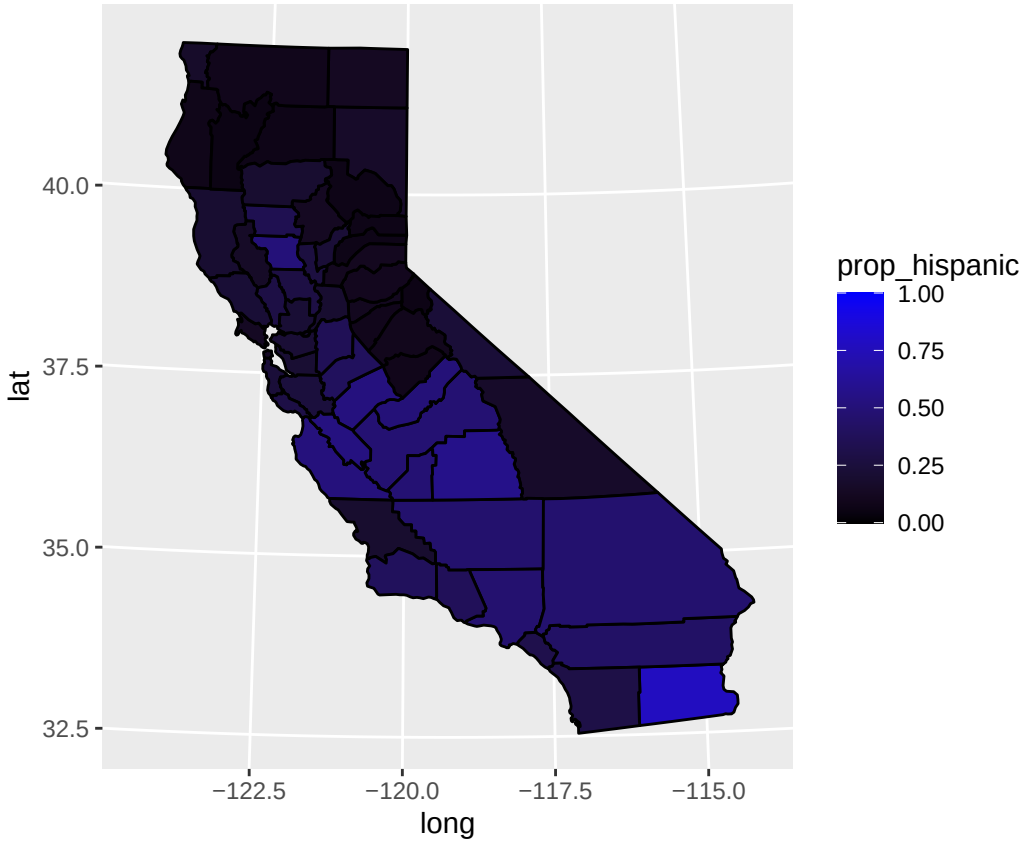

```
# WHITE, 2009

ca_map_white_09 <- left_join(ca_map_data, ca_white_09, by = c(subregion = "county",
    region = "state"))

ggplot(ca_map_white_06, aes(x = long, y = lat, group = group,
    fill = prop_white)) + geom_polygon(color = "black") + coord_map("polyconic") +
    scale_fill_gradient(low = "black", high = "red", limits = c(0,
        1), oob = scales::squish) + scale_color_gradient(low = "black",
    high = "red", limits = c(0, 1), oob = scales::squish)
```



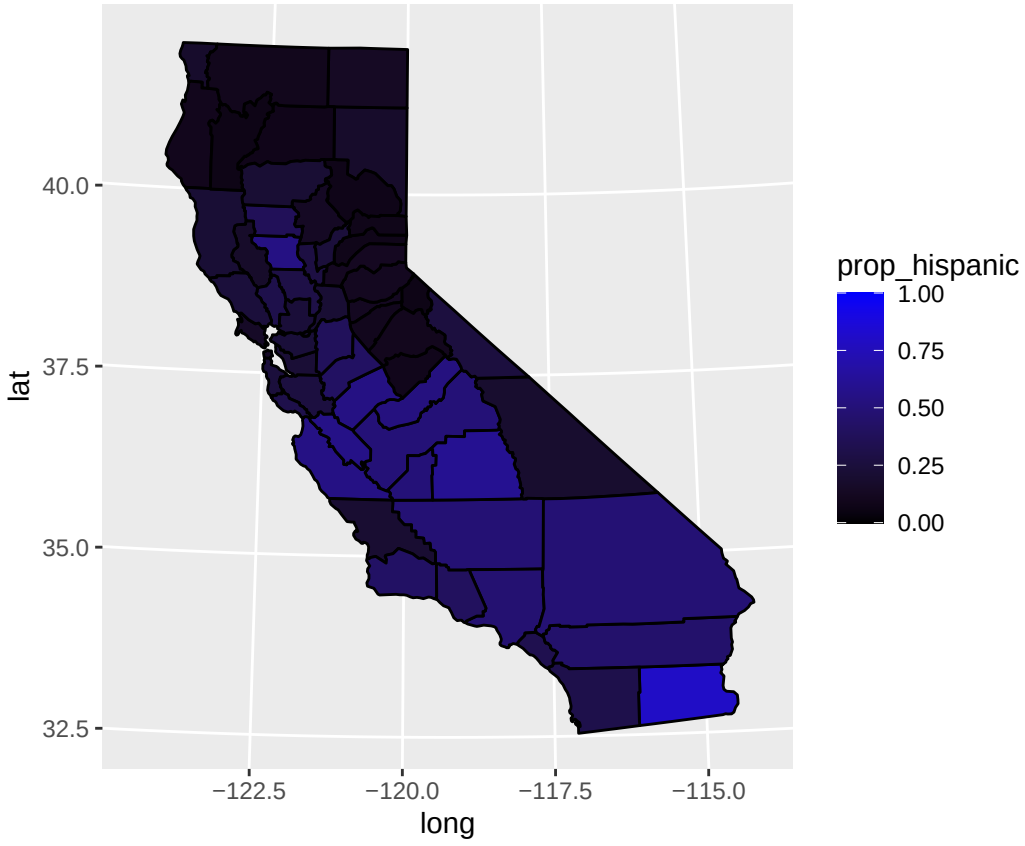

```
# WHITE, 2012

ca_map_white_12 <- left_join(ca_map_data, ca_white_12, by = c(subregion = "county",
    region = "state"))

ggplot(ca_map_white_12, aes(x = long, y = lat, group = group,
    fill = prop_white)) + geom_polygon(color = "black") + coord_map("polyconic") +
    scale_fill_gradient(low = "black", high = "red", limits = c(0,
        1), oob = scales::squish) + scale_color_gradient(low = "black",
    high = "red", limits = c(0, 1), oob = scales::squish)
```



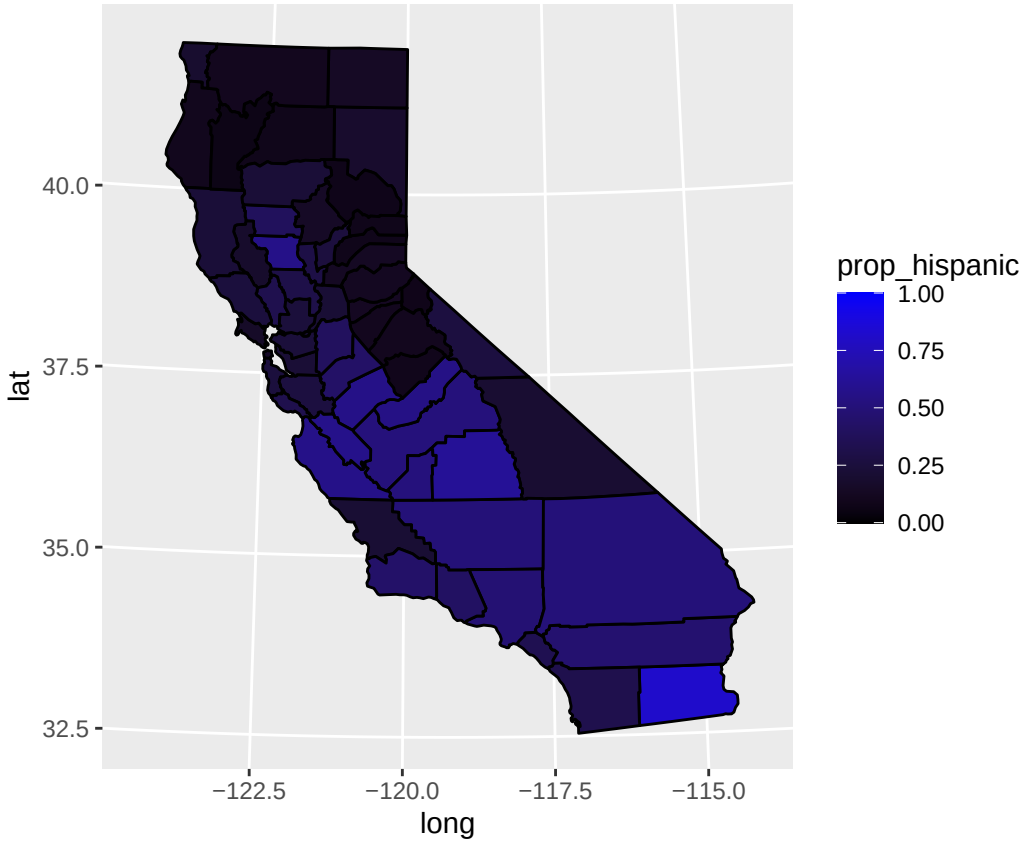

```
# HISPANIC, 2006

ca_map_hispanic_06 <- left_join(ca_map_data, ca_hispanic_06,
    by = c(subregion = "county", region = "state"))

ggplot(ca_map_hispanic_06, aes(x = long, y = lat, group = group,
    fill = prop_hispanic)) + geom_polygon(color = "black") +
    coord_map("polyconic") + scale_fill_gradient(low = "black",
    high = "blue", limits = c(0, 1), oob = scales::squish) +
    scale_color_gradient(low = "black", high = "blue", limits = c(0,
        1), oob = scales::squish)
```



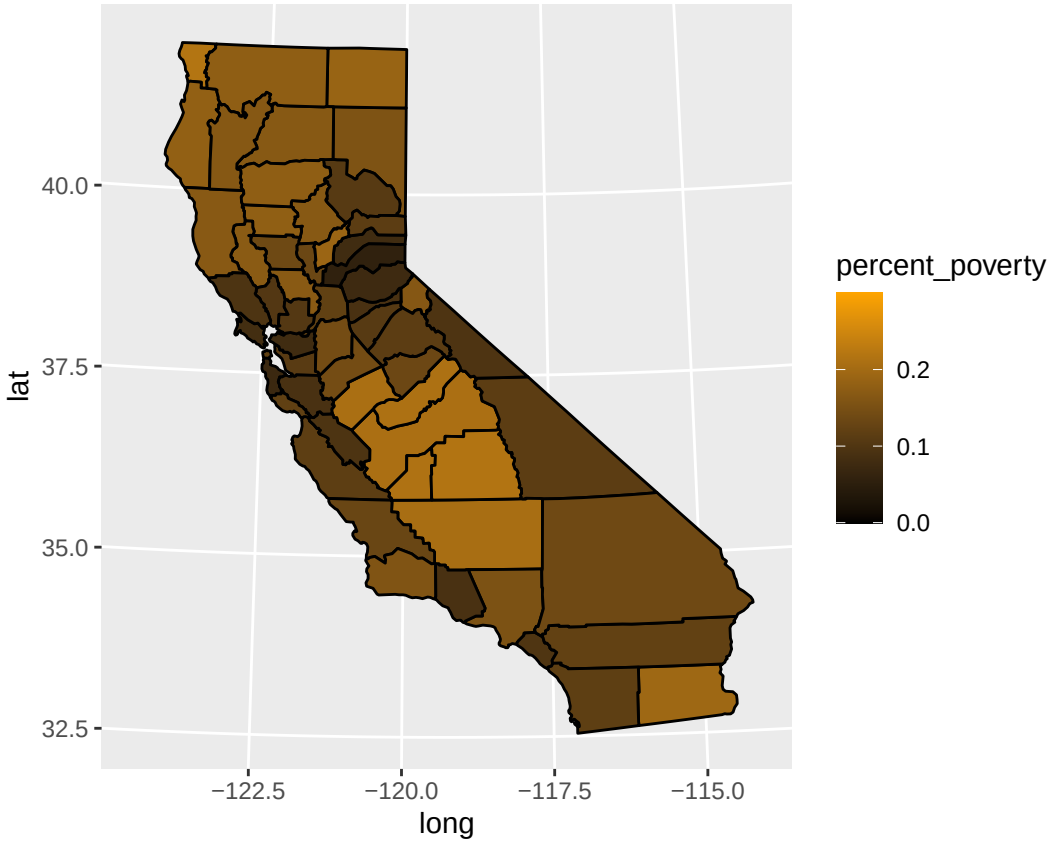

```
# HISPANIC, 2009

ca_map_hispanic_09 <- left_join(ca_map_data, ca_hispanic_09,
    by = c(subregion = "county", region = "state"))

ggplot(ca_map_hispanic_09, aes(x = long, y = lat, group = group,
    fill = prop_hispanic)) + geom_polygon(color = "black") +
    coord_map("polyconic") + scale_fill_gradient(low = "black",
    high = "blue", limits = c(0, 1), oob = scales::squish) +
    scale_color_gradient(low = "black", high = "blue", limits = c(0,
        1), oob = scales::squish)
```



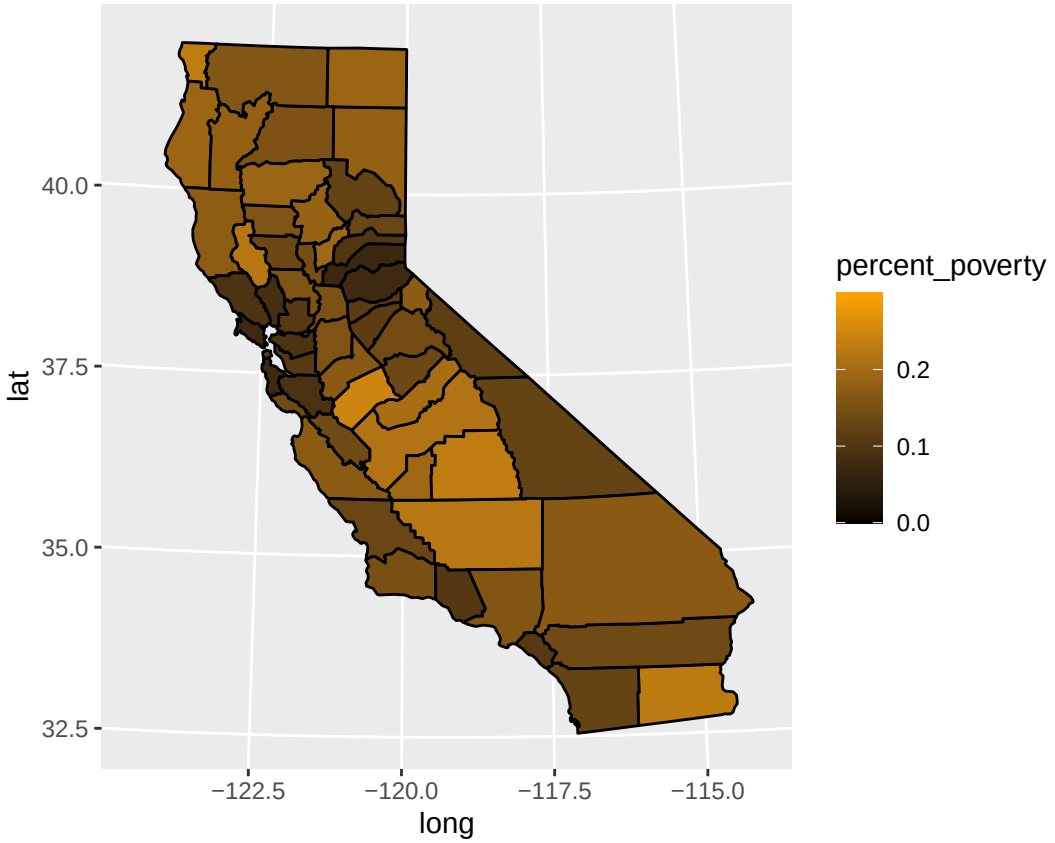

```
# HISPANIC, 2012
ca_map_hispanic_12 <- left_join(ca_map_data, ca_hispanic_12,
    by = c(subregion = "county", region = "state"))

ggplot(ca_map_hispanic_12, aes(x = long, y = lat, group = group,
    fill = prop_hispanic)) + geom_polygon(color = "black") +
    coord_map("polyconic") + scale_fill_gradient(low = "black",
    high = "blue", limits = c(0, 1), oob = scales::squish) +
    scale_color_gradient(low = "black", high = "blue", limits = c(0,
        1), oob = scales::squish)
```



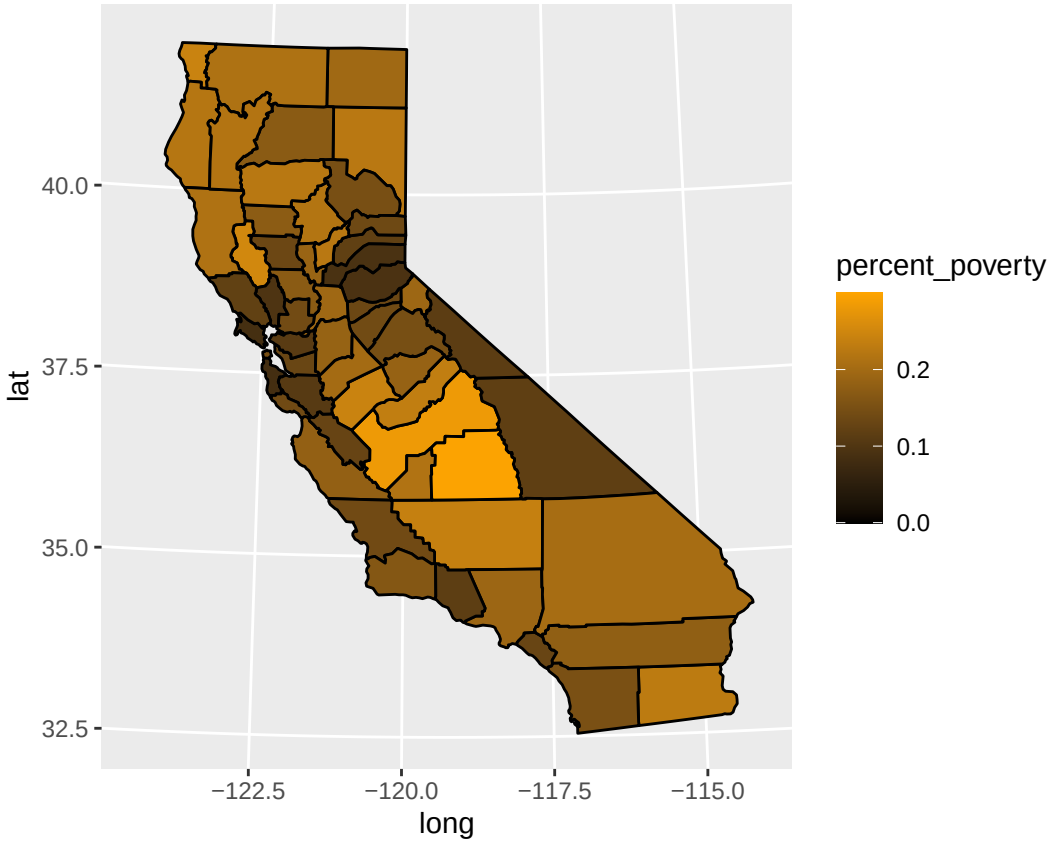

```
# looking at poverty by location 2006

ca_map_poverty_06 <- left_join(ca_map_data, ca_poverty_06, by = c(subregion = "county",
    region = "state"))

ggplot(ca_map_poverty_06, aes(x = long, y = lat, group = group,
    fill = percent_poverty)) + geom_polygon(color = "black") +
    coord_map("polyconic") + scale_fill_gradient(low = "black",
    high = "orange", limits = c(0, 0.3), oob = scales::squish) +
    scale_color_gradient(low = "black", high = "orange", limits = c(0,
        0.3), oob = scales::squish)
```




```
# 2009

ca_map_poverty_09 <- left_join(ca_map_data, ca_poverty_09, by = c(subregion = "county",
    region = "state"))

ggplot(ca_map_poverty_09, aes(x = long, y = lat, group = group,
    fill = percent_poverty)) + geom_polygon(color = "black") +
    coord_map("polyconic") + scale_fill_gradient(low = "black",
    high = "orange", limits = c(0, 0.3), oob = scales::squish) +
    scale_color_gradient(low = "black", high = "orange", limits = c(0,
        0.3), oob = scales::squish)
```

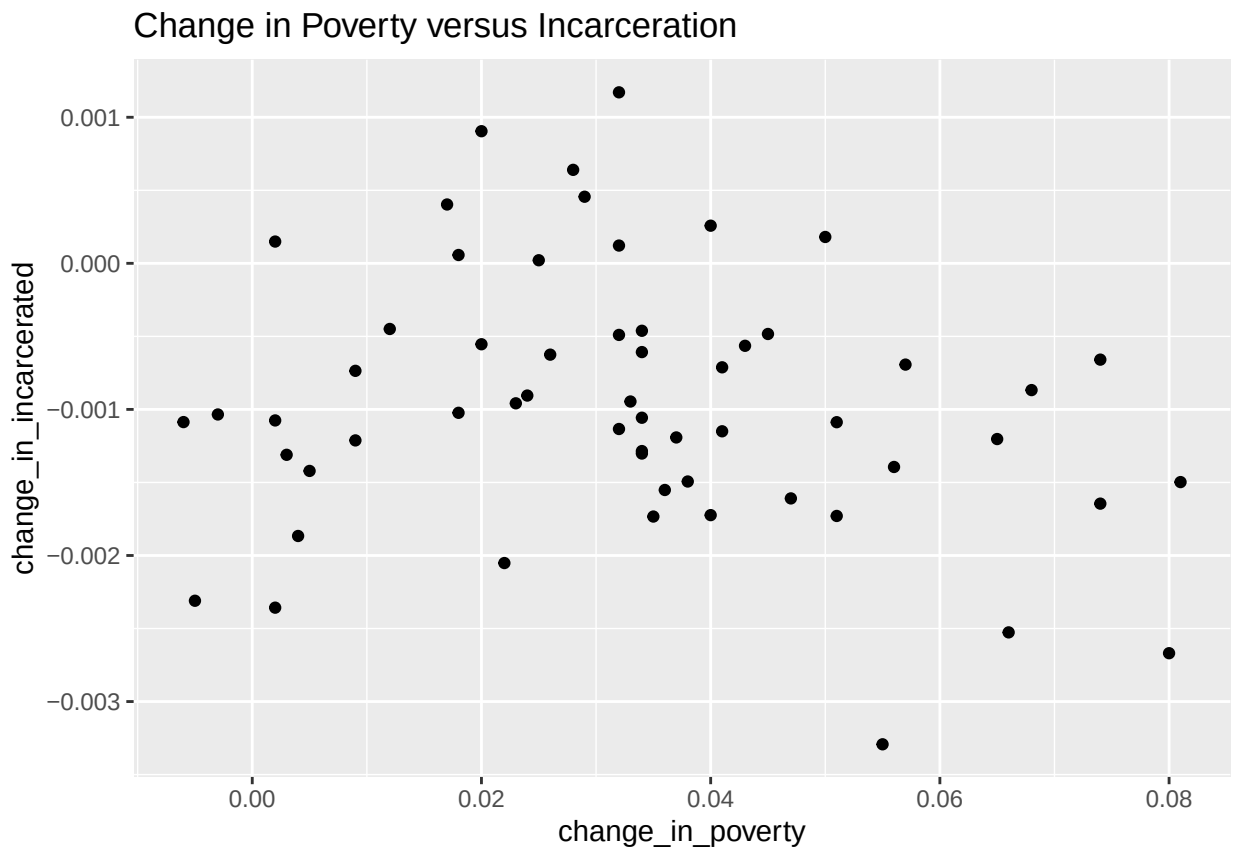



```
# 2012
ca_map_poverty_12 <- left_join(ca_map_data, ca_poverty_12, by = c(subregion = "county",
    region = "state"))

ggplot(ca_map_poverty_12, aes(x = long, y = lat, group = group,
    fill = percent_poverty)) + geom_polygon(color = "black") +
    coord_map("polyconic") + scale_fill_gradient(low = "black",
    high = "orange", limits = c(0, 0.3), oob = scales::squish) +
    scale_color_gradient(low = "black", high = "orange", limits = c(0,
        0.3), oob = scales::squish)
```

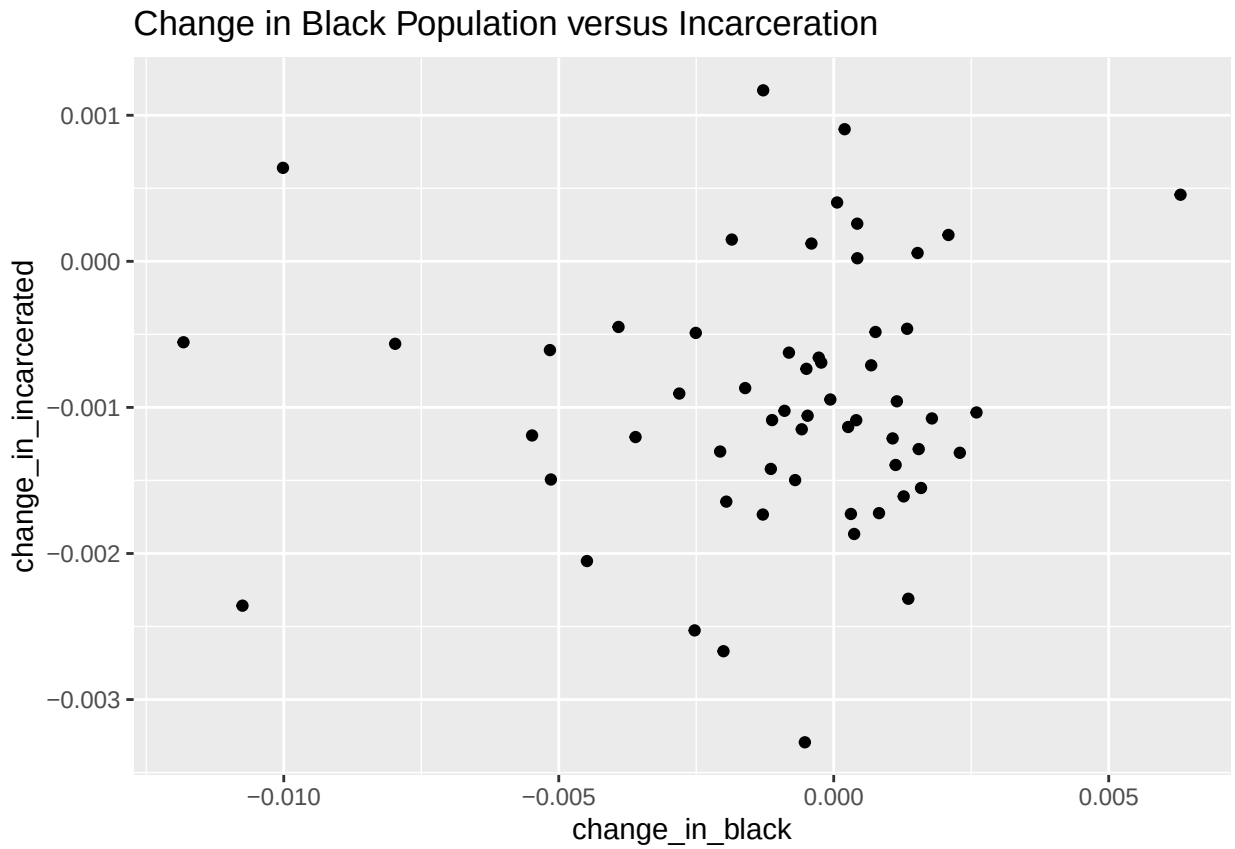


Appendix: Exploratory Plots

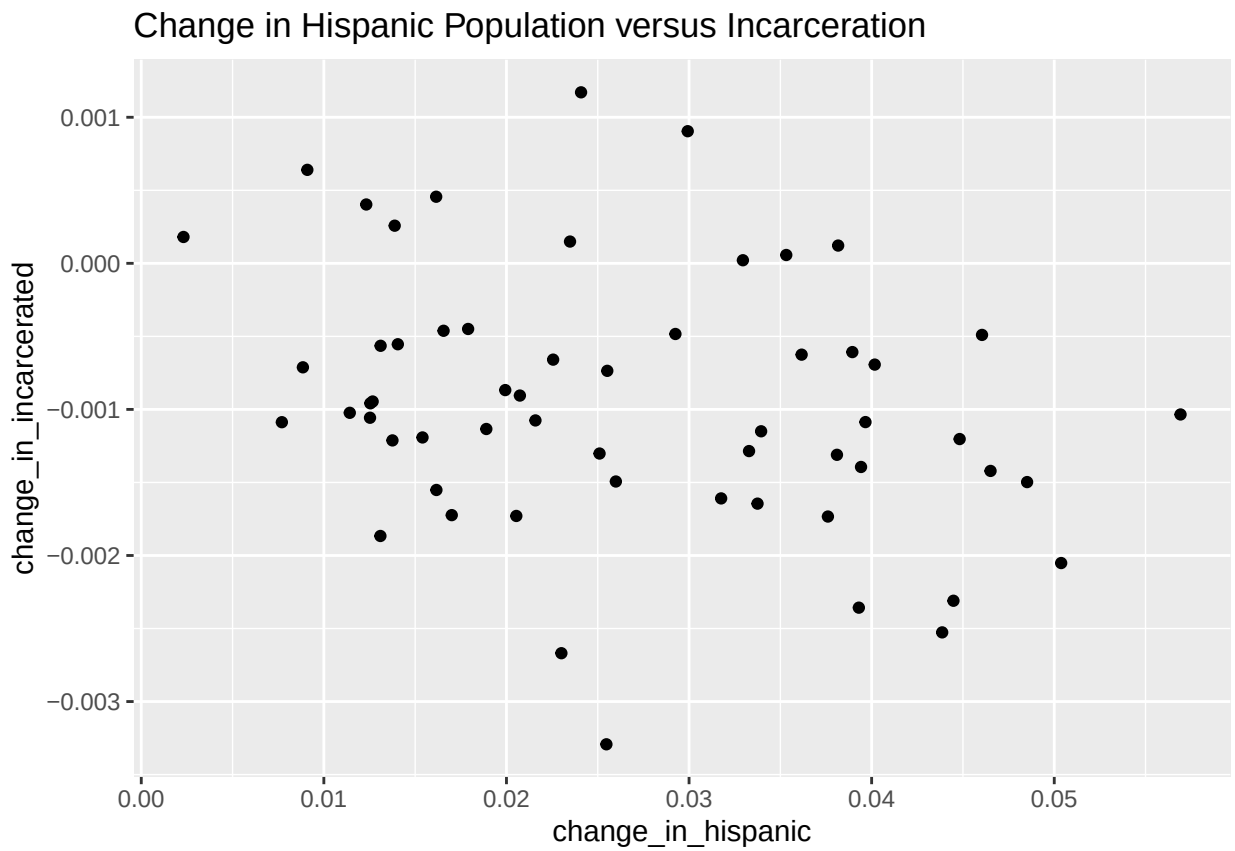
```
# plotting variables relative to changes in incarceration

variable_changes |>
    ggplot(aes(x = change_in_poverty, y = change_in_incarcerated)) +
    geom_point() + labs(title = "Change in Poverty versus Incarceration",
    xlab = "Proportion of People in Poverty", ylab = "Proportion of People Incarcerated")
```



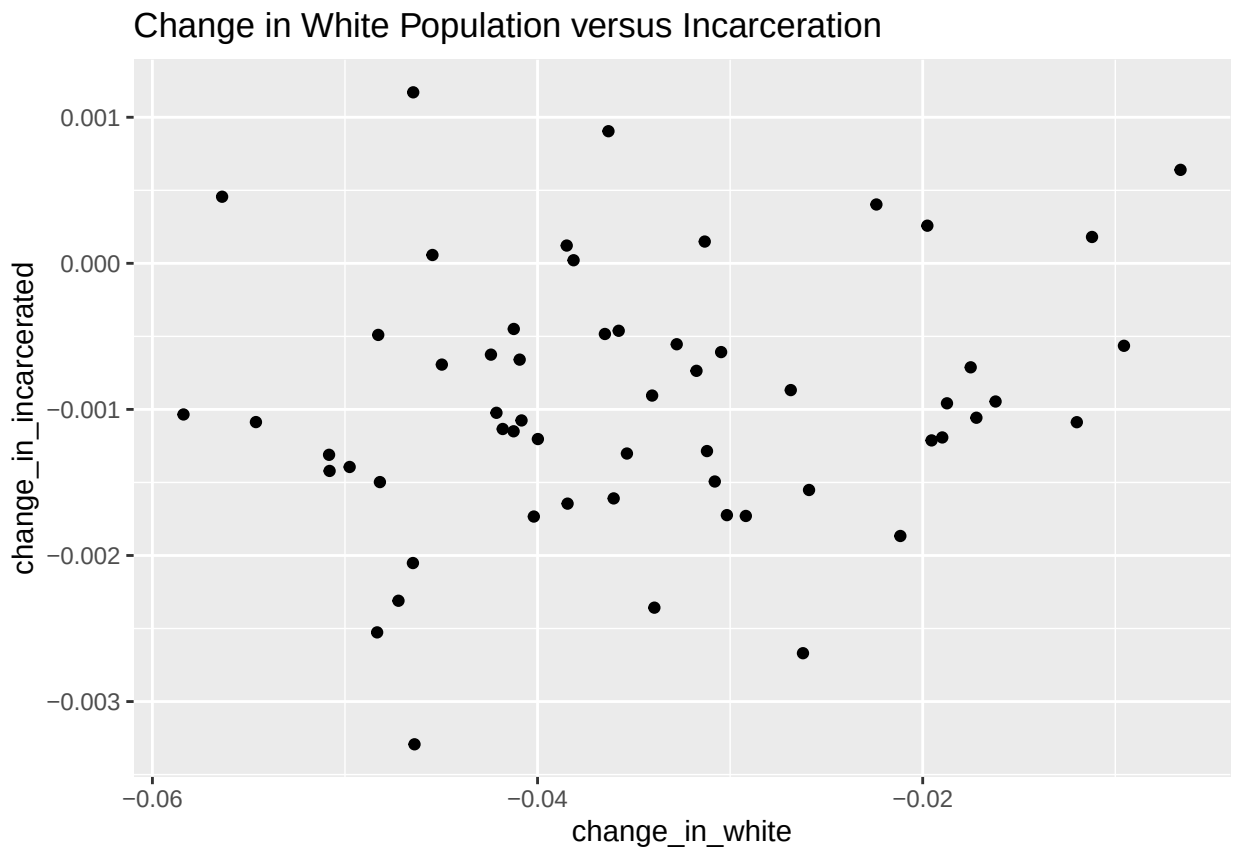

```
variable_changes |>
    ggplot(aes(x = change_in_black, y = change_in_incarcerated)) +
    geom_point() + labs(title = "Change in Black Population versus Incarceration",
    xlab = "Proportion of Black People", ylab = "Proportion of People Incarcerated")
```



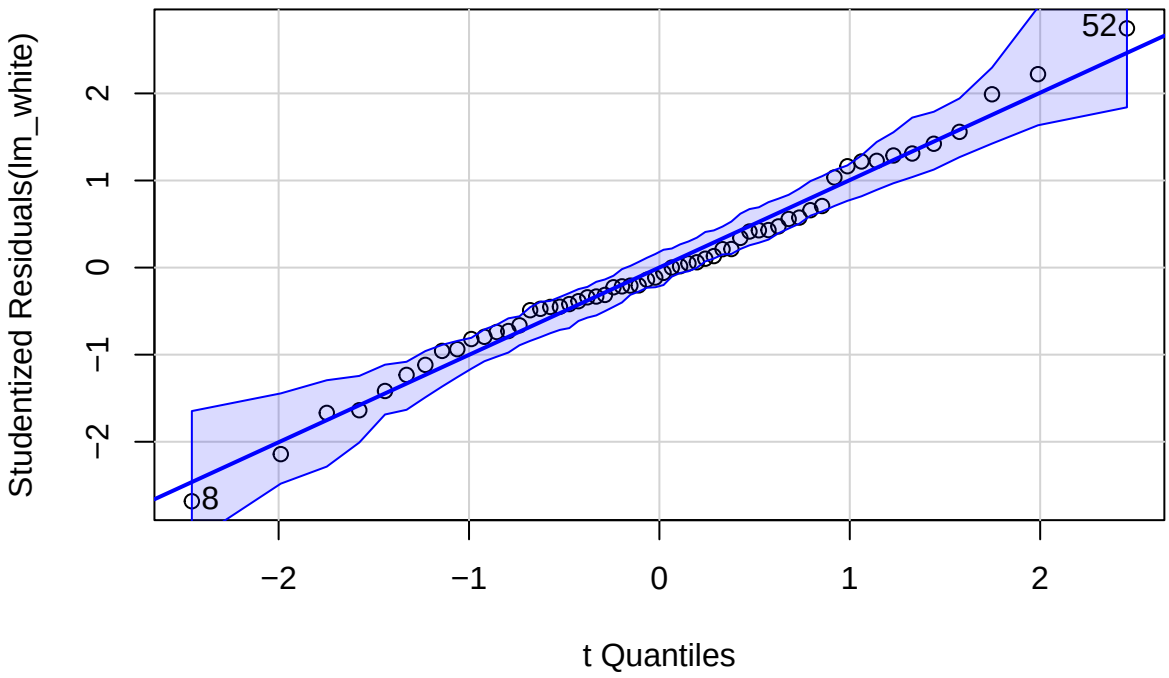

```
variable_changes |>
    ggplot(aes(x = change_in_hispanic, y = change_in_incarcerated)) +
    geom_point() + labs(title = "Change in Hispanic Population versus Incarceration",
    xlab = "Proportion of Black People", ylab = "Proportion of People Incarcerated")
```




```
# note that only the white proportion decreased in this
# time frame, so the plot starts with negative values
variable_changes |>
    ggplot(aes(x = change_in_white, y = change_in_incarcerated)) +
    geom_point() + labs(title = "Change in White Population versus Incarceration",
    xlab = "Proportion of Black People", ylab = "Proportion of People Incarcerated")
```

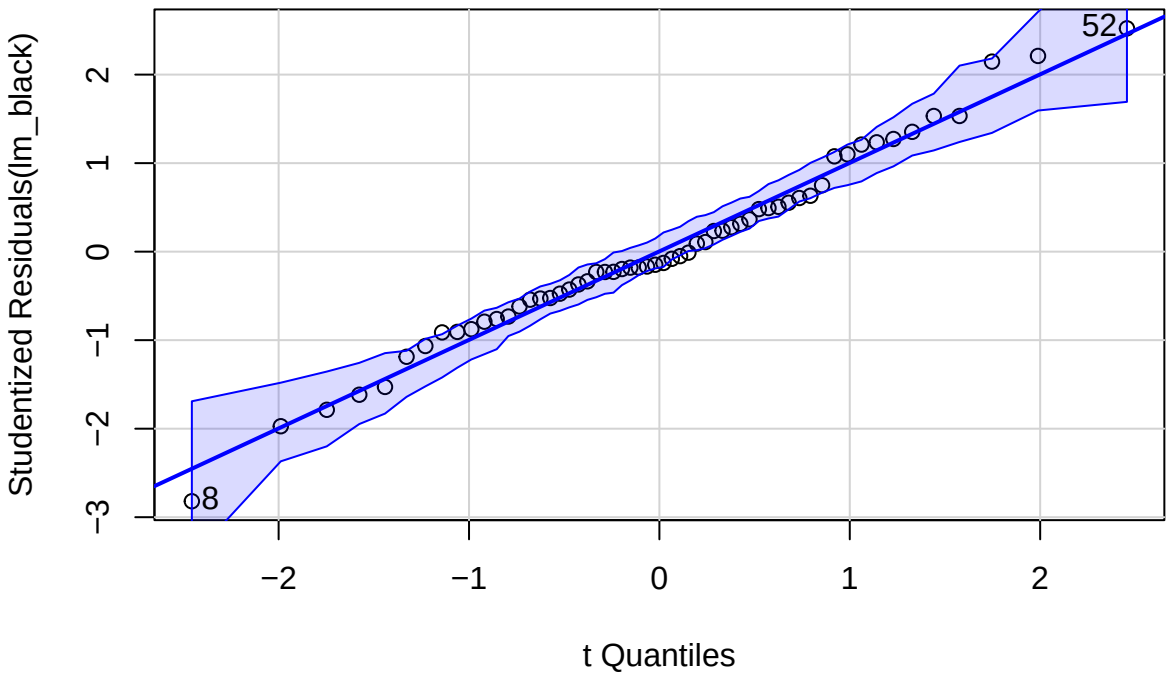



```
# no clear relationship for any variable!

# Checking normal distribution
qqPlot(lm_white)
```



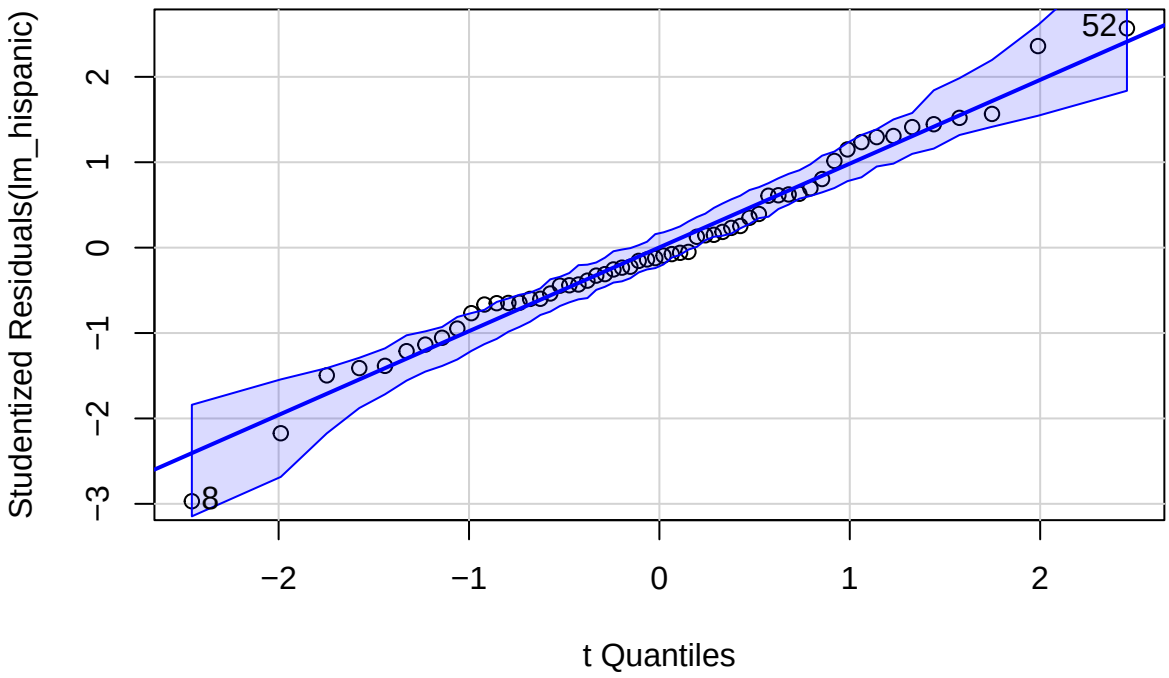

```
## [1]  8 52
```

```
qqPlot(lm_black)
```



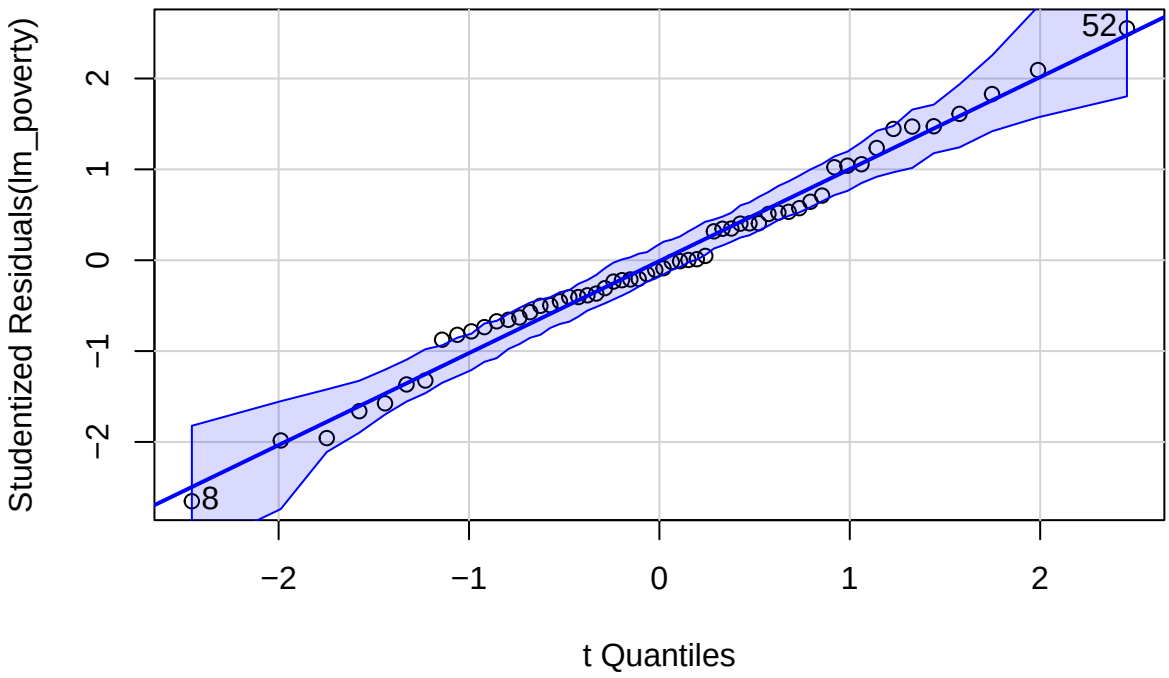

```
## [1]  8 52
```

```
qqPlot(lm_hispanic)
```




```
## [1]  8 52
```

```
qqPlot(lm_poverty)
```




```
## [1]  8 52
```