Incarceration in the U.S.

Ryu Parish

2024-05-27

Introduction

Incarceration in the United States is a pressing social issue, marked by significant racial disparities. This report analyzes U.S. prison population data, with a primary focus on the Black racial group compared to other racial groups, including White, Latinx, Native, and Asian American/Pacific Islander (AAPI). Key variables include total prison populations and specific counts for these racial groups. By examining these variables, we aim to highlight trends over time and geographic variations, emphasizing the disproportionate incarceration rates of Black individuals relative to other groups. Understanding these disparities is essential for addressing systemic inequities in the criminal justice system. Data Collection Information

Who collected the data?

The dataset was collected by the Vera Institute of Justice, a non-profit organization focused on criminal justice reform.

How was the data collected or generated?

The data was collected from various sources, including state departments of corrections and other governmental agencies. The collection process likely involved aggregating publicly available data, reports, and records regarding prison populations across different states and years.

Why was the data collected?

The data was collected to provide a comprehensive and detailed view of the prison population in the United States, with an emphasis on racial demographics. The goal is to inform research, policy-making, and advocacy efforts aimed at addressing and reforming issues within the criminal justice system, particularly those related to racial disparities and mass incarceration.

How many observations (rows) are in the data?

After cleaning the data to remove rows with missing values, the number of observations (rows) is r complete_cases_count.

```r
# Calculate the number of rows with no missing values
complete_cases_count <- nrow(prison_data_clean)
complete_cases_count
```

```
## [1] 10155
```

How many features (columns) are in the data?

The dataset contains 14 features (columns) which include the year, state, and prison population counts for different racial groups.

```r
# Calculate the number of features (columns)
num_features <- ncol(prison_data_clean)
num_features
```

```
## [1] 39
```

What, if any, ethical questions or questions of power do you need to consider when working with this data?

```
When working with this dataset, several ethical questions and issues of power must be considered:
```

Privacy and Confidentiality:

Although the dataset is aggregated and anonymized, it's important to ensure that no individual-level data can be inferred, which could lead to privacy violations.

Representation and Bias:

The data may reflect underlying biases in the criminal justice system, such as racial profiling, discriminatory policing practices, and sentencing disparities. Researchers must be cautious in interpreting the data and avoid perpetuating stereotypes or biased conclusions.

Use of Data:

The data should be used responsibly to advocate for fair and just policies rather than punitive measures. Misuse of the data could reinforce harmful practices and further marginalize affected communities.

What are possible limitations or problems with this data?

```
The dataset has several limitations and potential issues:
```

Incomplete Data:

The dataset may have missing values for certain years or states, leading to gaps in the analysis. This incompleteness can affect the accuracy and reliability of trends and comparisons.

Inconsistencies in Reporting:

Different states may have varying definitions, reporting standards, and data collection methodologies, resulting in inconsistencies. These discrepancies can complicate cross-state comparisons and overall analysis.

Historical Context:

The data does not provide contextual information about changes in laws, policies, or socio-economic conditions that may impact incarceration rates. Understanding the broader context is crucial for accurate interpretation.

Aggregate Data:

The dataset is aggregated at the state level, which masks local variations and individual-level details. This aggregation can obscure important nuances and trends at more granular levels.

Potential Bias:

As the data reflects the operations of the criminal justice system, it inherently includes biases present in the system, such as racial profiling and discriminatory practices. Analysts must account for these biases in their interpretations and recommendations.

Load necessary libraries

```
library(dplyr)
library(ggplot2)
```

Load the dataset

```
# Extract race-related variables
race_vars <- prison_data %>%
  select(contains("white"), contains("black"), contains("aapi"), contains("latinx"), contains("native")

# Calculate summary statistics for race-related variables
race_summary <- race_vars %>%
  summarise(across(everything(), list(total = ~sum(.x, na.rm = TRUE),
                                      avg = ~mean(.x, na.rm = TRUE),
                                      max = ~max(.x, na.rm = TRUE),
                                      min = ~min(.x, na.rm = TRUE))))

race_summary
```

```
## # A tibble: 1 x 92
##   white_prison_pop_total white_prison_pop_avg white_prison_pop_max
##                    <dbl>                <dbl>                <dbl>
## 1                9903343                 163.                 9945
## # i 89 more variables: white_prison_pop_min <dbl>,
## #   white_female_prison_pop_total <dbl>, white_female_prison_pop_avg <dbl>,
## #   white_female_prison_pop_max <dbl>, white_female_prison_pop_min <dbl>,
## #   white_male_prison_pop_total <dbl>, white_male_prison_pop_avg <dbl>,
## #   white_male_prison_pop_max <dbl>, white_male_prison_pop_min <dbl>,
## #   white_pop_15to64_total <dbl>, white_pop_15to64_avg <dbl>,
## #   white_pop_15to64_max <dbl>, white_pop_15to64_min <dbl>, ...
```

Introductory and Summary information questions

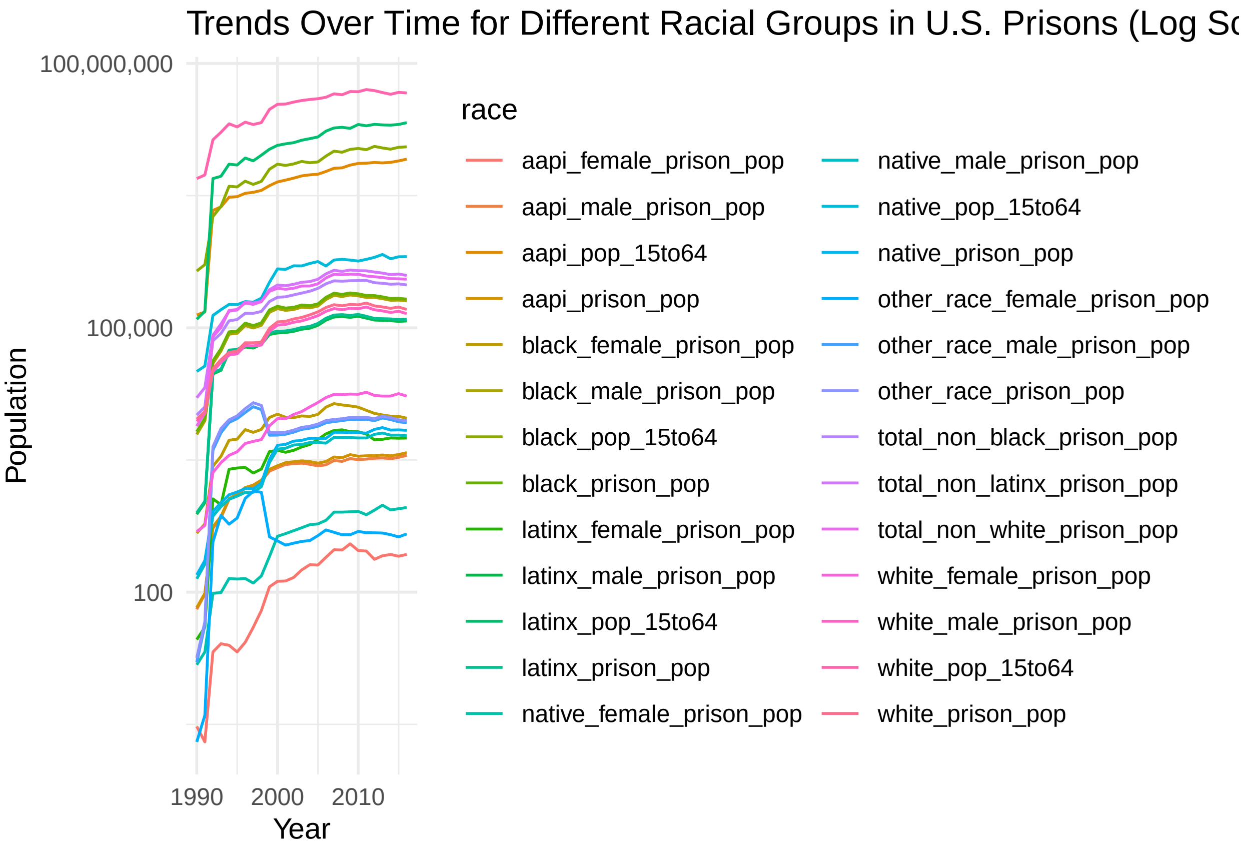
```
# Summary calculations
current_year <- max(prison_data_clean$year)
average_black_prison_pop <- prison_data_clean %>%
  filter(year == current_year) %>%
  summarise(avg_black_prison_pop = mean(black_prison_pop, na.rm = TRUE)) %>%
  pull(avg_black_prison_pop)

highest_black_prison_pop_state <- prison_data_clean %>%
  filter(year == current_year) %>%
  arrange(desc(black_prison_pop)) %>%
  slice(1) %>%
  pull(state)

lowest_black_prison_pop_state <- prison_data_clean %>%
  filter(year == current_year) %>%
  arrange(black_prison_pop) %>%
  slice(1) %>%
  pull(state)
```

```r
change_black_prison_pop_last_10_years <- prison_data_clean %>%
  filter(year %in% c(current_year, current_year - 10)) %>%
  group_by(year) %>%
  summarise(total_black_prison_pop = sum(black_prison_pop, na.rm = TRUE)) %>%
  summarise(change = diff(total_black_prison_pop))

average_non_black_prison_pop <- prison_data_clean %>%
  filter(year == current_year) %>%
  summarise(avg_non_black_prison_pop = mean(total_non_black_prison_pop, na.rm = TRUE)) %>%
  pull(avg_non_black_prison_pop)

# Create a summary paragraph with the statistics
summary_paragraph <- paste(
  "In the current year, the average Black jail population across all counties is", average_black_prison_
  ". The state with the highest Black jail population is", highest_black_prison_pop_state,
  ", while the state with the lowest Black jail population is", lowest_black_prison_pop_state,
  ". Over the last ten years, the total Black jail population has changed by", change_black_prison_pop_l
  ". Additionally, the average non-Black jail population across all counties in the current year is", av
  "."
)
summary_paragraph
```

```
## [1] "In the current year, the average Black jail population across all counties is 384.794545454545
```

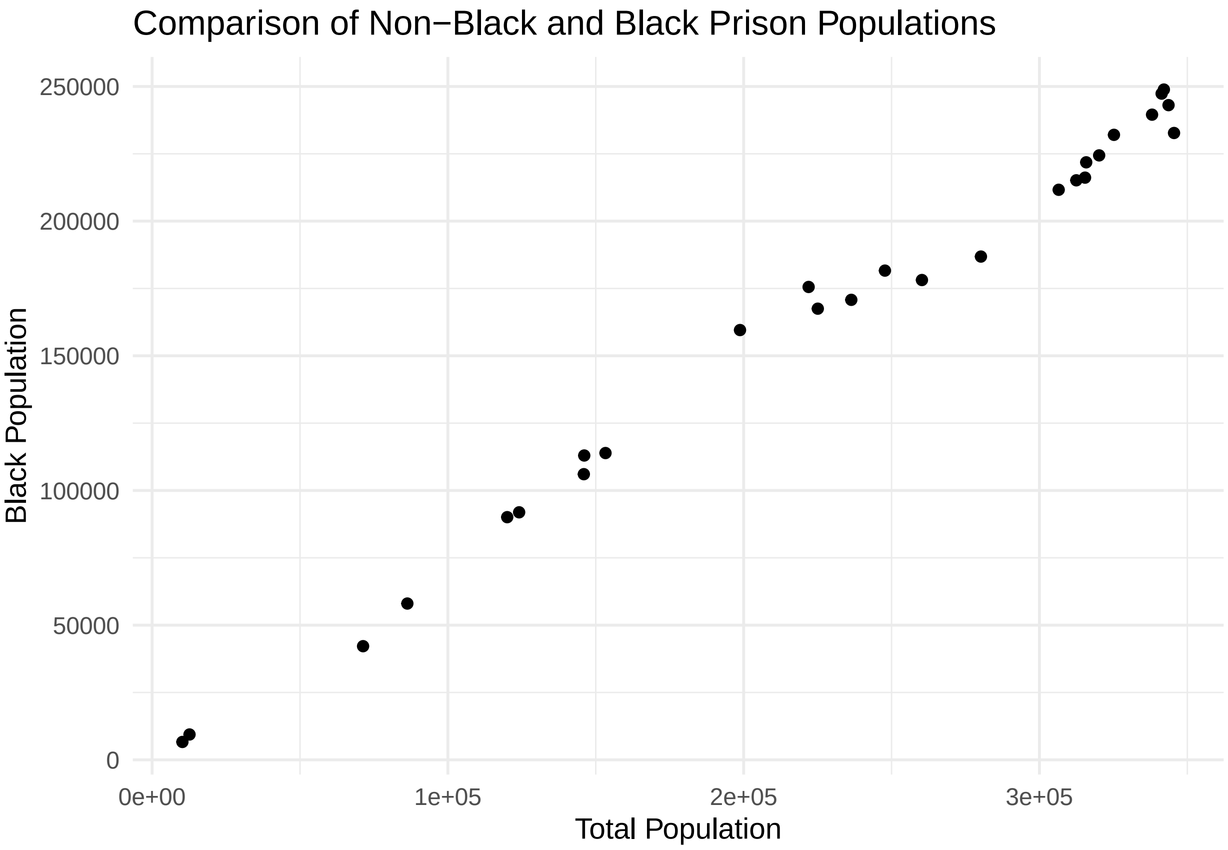
Trend over time with the select variables

```r
# Plot trends over time for race-related variables
race_time_trends <- prison_data_clean %>%
  group_by(year) %>%
  summarise(across(contains("white"), sum, na.rm = TRUE),
            across(contains("black"), sum, na.rm = TRUE),
            across(contains("aapi"), sum, na.rm = TRUE),
            across(contains("latinx"), sum, na.rm = TRUE),
            across(contains("native"), sum, na.rm = TRUE),
            across(contains("other"), sum, na.rm = TRUE)) %>%
  pivot_longer(cols = -year, names_to = "race", values_to = "population")

ggplot(race_time_trends, aes(x = year, y = population, color = race)) +
  geom_line() +
  scale_y_log10(labels = comma) +
  labs(title = "Trends Over Time for Different Racial Groups in U.S. Prisons (Log Scale)",
       x = "Year",
       y = "Population") +
  theme_minimal()
```



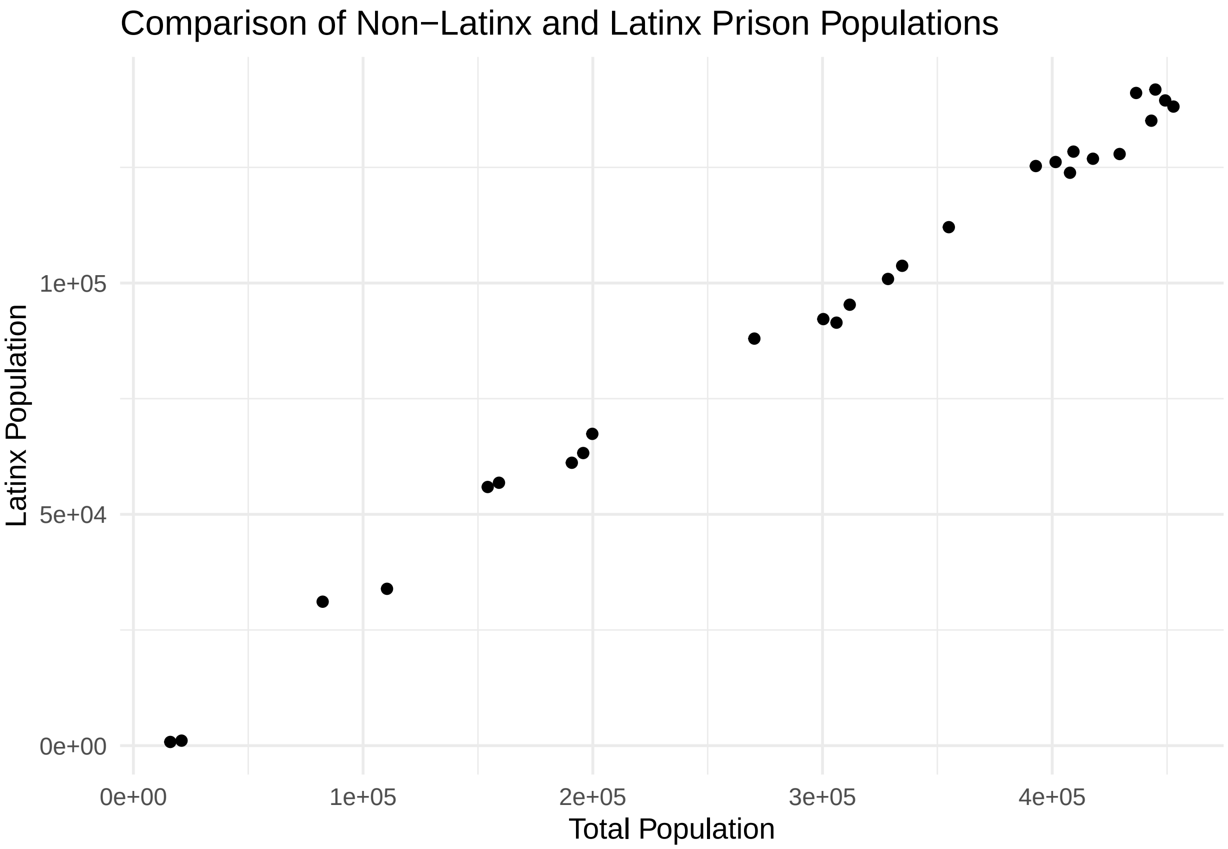
Comparison of Variables

```
# Compare total population with black prison population
prison_data_clean %>%
  group_by(year) %>%
  summarise(total_population = sum(total_non_black_prison_pop, na.rm = TRUE),
            black_population = sum(black_prison_pop, na.rm = TRUE)) %>%
  ggplot(aes(x = total_population, y = black_population)) +
  geom_point() +
  labs(title = "Comparison of Non-Black and Black Prison Populations",
       x = "Total Population",
       y = "Black Population") +
  theme_minimal()
```



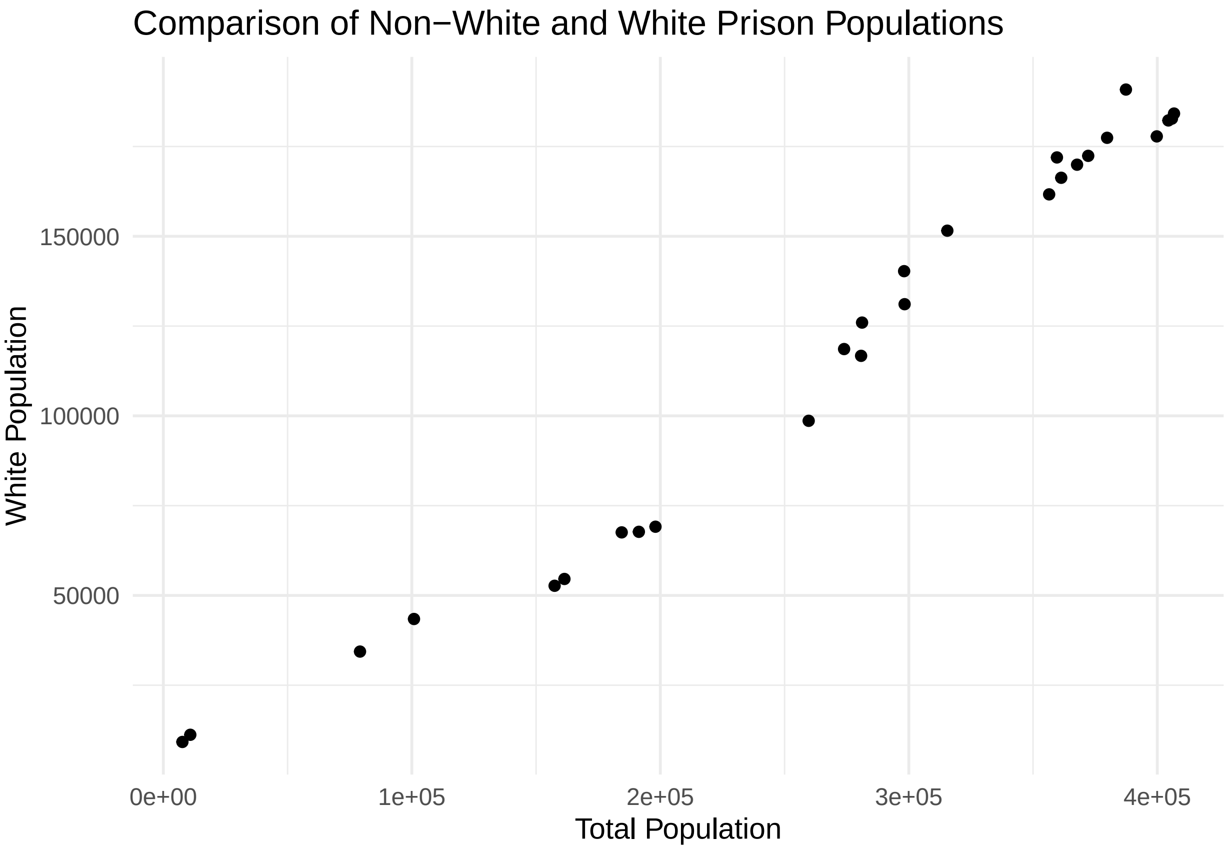
Comparison of Variables

```
# Compare total population with black prison population
prison_data_clean %>%
  group_by(year) %>%
  summarise(total_population = sum(total_non_latinx_prison_pop, na.rm = TRUE),
            latinx_population = sum(latinx_prison_pop, na.rm = TRUE)) %>%
  ggplot(aes(x = total_population, y = latinx_population)) +
  geom_point() +
  labs(title = "Comparison of Non-Latinx and Latinx Prison Populations",
       x = "Total Population",
       y = "Latinx Population") +
  theme_minimal()
```



Comparison of Variables

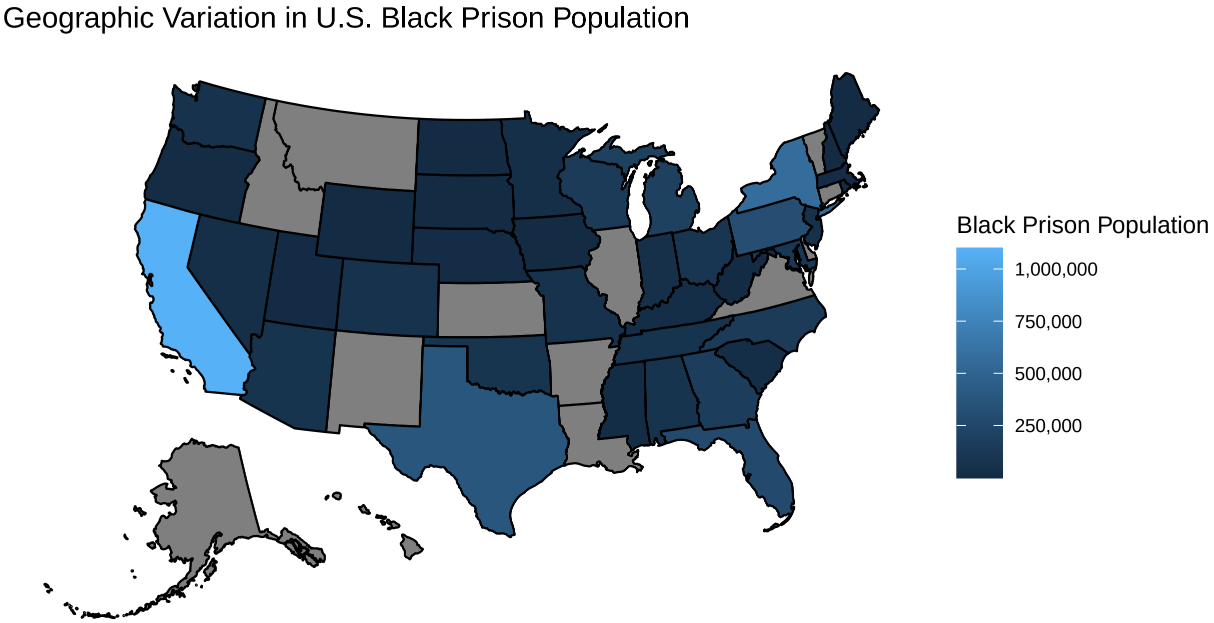
```
# Compare total population with black prison population
prison_data_clean %>%
  group_by(year) %>%
  summarise(total_population = sum(total_non_white_prison_pop, na.rm = TRUE),
            white_population = sum(white_prison_pop, na.rm = TRUE)) %>%
  ggplot(aes(x = total_population, y = white_population)) +
  geom_point() +
  labs(title = "Comparison of Non-White and White Prison Populations",
       x = "Total Population",
       y = "White Population") +
  theme_minimal()
```



So it looks like the proportional change between these three races and the rest of the races has not changed much. The pattern of incarceration has not changed much in the last century.

Geographic data mapping

```
# Summarize data by state for black prison population
state_data_black <- prison_data_clean %>%
  group_by(state) %>%
  summarise(black_population = sum(black_prison_pop, na.rm = TRUE))

plot_usmap(data = state_data_black, values = "black_population", lines = "black") +
  scale_fill_continuous(name = "Black Prison Population", label = scales::comma) +
  labs(title = "Geographic Variation in U.S. Black Prison Population") +
  theme(legend.position = "right")
```


Geographic Variation in U.S. Black Prison Population
Black Prison Population
1,000,000
750,000
500,000
250,000