Civil Rights in States: Exploring U.S. Hate Crime Data in 2019-2020

3Lattes: Deontavius Harris, Adam Blumoff, and Helen Mak

Introduction

"Civil rights investigations are at the heart of what we do at the FBI for the simple reason that civil liberties and civil rights are the very heart of who we are as Americans." FBI Associate Deputy Director Jeffrey Sallet said during an FBI conference in Denver on June 2021[1].

In 2020, it seemed like there were sensational headlines in the news everyday. From the question of the origins of Covid-19 to the death of George Floyd followed by the Black Lives Matter movement, tensions between individuals and communities arose within the U.S. For our group, the onset of Hate Crimes reports on the news caught our attention. In social media, there are many claims that there have been more anti-Asian and anti-African-American hate crimes over the past years. In fact, Sallet said that "the FBI's Criminal Investigative Division has elevated civil rights violations to its highest-level national threat priority—a measure of how the FBI allocates money and resources"[2]. From this, we want to see what the FBI data says about Hate Crimes to determine where the money and resources should be allocated to. If Civil Rights is at the heart of who Americans are, then the FBI will want to address the following questions to figure out which states and racial groups need the most attention:

- Which states have seen the most significant change in the count of race-based hate crimes between 2019 and 2020?
- Which specific race-based hate crimes have significantly changed during 2019-2020?
- What is the relationship between race-based hate crime with population or income? How can we group states up by these relationships and are there significant differences between each group?

To answer these questions, we created two datasets. Our first dataset has the count of specific hate crimes for 2019 and 2020 as well as the percent change in total race-based hate crimes. Our second dataset includes the 2020 and percent change data from our first dataset and state based data on income, population, and population by race. To see which states have seen the most significant change in hate crimes between 2019-2020, we visualize the numbers on a map as well as output the values on tables. We create a side-by-side bar plot to see which specific race-based hate crimes have significantly changed. Finally, we compare two clustering methods, kmeans & mclust, on the relationship between hate crimes and income as well as hate crimes and population. Applying the mclust method is our way of going beyond in this project. This is because it is clustering method we haven't looked at and requires us to use a new package, mclust[3]. Further explanation for our data will be in the 'Data' section and descriptions of our methods, reasonings, and findings in our 'Analysis' section.

[1] https://www.fbi.gov/news/stories/hate-crimes-and-civil-rights-elevated-to-top-national-threat-priority-063021
[2] Ibid.
[3] https://cran.r-project.org/web/packages/mclust/vignettes/mclust.html

Data

Data Sources

We found our data on race-based hate crimes through the FBI Hate Crime Statistics[4]. The report of these statistics provides a good explanation of what determines a hate crime. The FBI's Hate Crime Data Collection Guidelines and Training Manual[5] explains: "Because motivation is subjective, it is sometimes difficult to know with certainty whether a crime resulted from the offender's bias. Moreover, the presence of bias alone does not necessarily mean that a crime can be considered a hate crime. Only when a law enforcement investigation reveals sufficient evidence to lead a reasonable and prudent person to conclude that the offender's actions were motivated, in whole or in part, by his or her bias, should an agency report an incident as a hate crime." Our data consists of single-bias data, an incident in which one or more offense types are motivated by the same bias.

In regards to how the FBI organizes the reporting of these hate crimes, law enforcement agencies report hate crimes every month or quarter to the FBI. They do so either through their state FBI Uniform Crime Reporting (UCR) Programs or directly. Either way, they submit the hate crime data electronically in a National Incident-Based Reporting System (NIBRS) submission, hate crime record layout, or a Microsoft Excel Workbook tool. The advantage of using these NIBRS submissions is that a more detailed report can be made on specific crimes, capturing more information than Excel reporting can. This information creates a more comprehensive analysis of each incident.

It is also important to note that we did not include data from Alabama since there were only 2 out of 430 agencies that reported in 2019 as well as Rhode Island. This means that the data was either lost somewhere or agencies didn't report on purpose. The percent change isn't accounted for because there were no observations to find the change for in 2019.

We took data from KFF.org to find information for Population[6], Race-specific population[7], and Median Annual Household Income[8] on a state-by-state basis. Since the most recent, complete data was for 2019, we used data for 2019 only. The 2019 data is based on an analysis of the U.S Census Bureau's March Supplement to the Current Population Survey (the CPS Annual Social and Economic Supplement or ASEC) from 2016, 2018, and 2020 (due to the challenges of the pandemic).

We will discuss the limitations of our data and analysis further in our conclusion.

Data Wrangling Process

The data wrangling code is under the folder "data-wrangling". We outputted the wrangled datasets and copied it to the "Report/data" file. The rscript for that wrangles the first dataset is "wrangle-hate-crimes.R" and the second dataset is "wrangle-income-race.R".

Hate Crime Dataset

From the FBI Hate Crimes dataset, we first filtered for single biases. Since the observations from start from 1991, we also filter for 2019 and 2020. We then use pivot wider to count specific hate crimes by state in 2019 and 2020, changing all the N/A values to 0. We also use mutate to find percent change over 2019-2020. The final variables for the first dataset that answer the first two questions (named "hate-crime-all.csv" in the data folder) are:

[4]https://crime-data-explorer.fr.cloud.gov/pages/explorer/crime/hate-crime

[5]https://www.fbi.gov/file-repository/ucr/ucr-hate-crime-data-collection-guidelines-training-manual-030122.pdf/view

[6](https://www.kff.org/other/state-indicator/total-number-of-residents-cps/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D

[7]https://www.kff.org/other/state-indicator/distribution-by-raceethnicity/?dataView=1&currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D

[8]https://www.kff.org/other/state-indicator/median-annual-income/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D

- State
- Percent Change in total race-based hate crimes (2019-2020)
- Count of total hate crimes by specific biases (2019 & 2020) including:
  - Anti-Black or African American
  - Anti-White
  - Anti-Hispanic or Latino
  - Anti-Asian
  - Anti-Arab
  - Anti-Native Hawaiian or Other Pacific Islander
  - Anti-American Indian or Alaska Native
  - Anti-Multiple Races
  - Anti-Other Race/Ethnicity/Ancestry

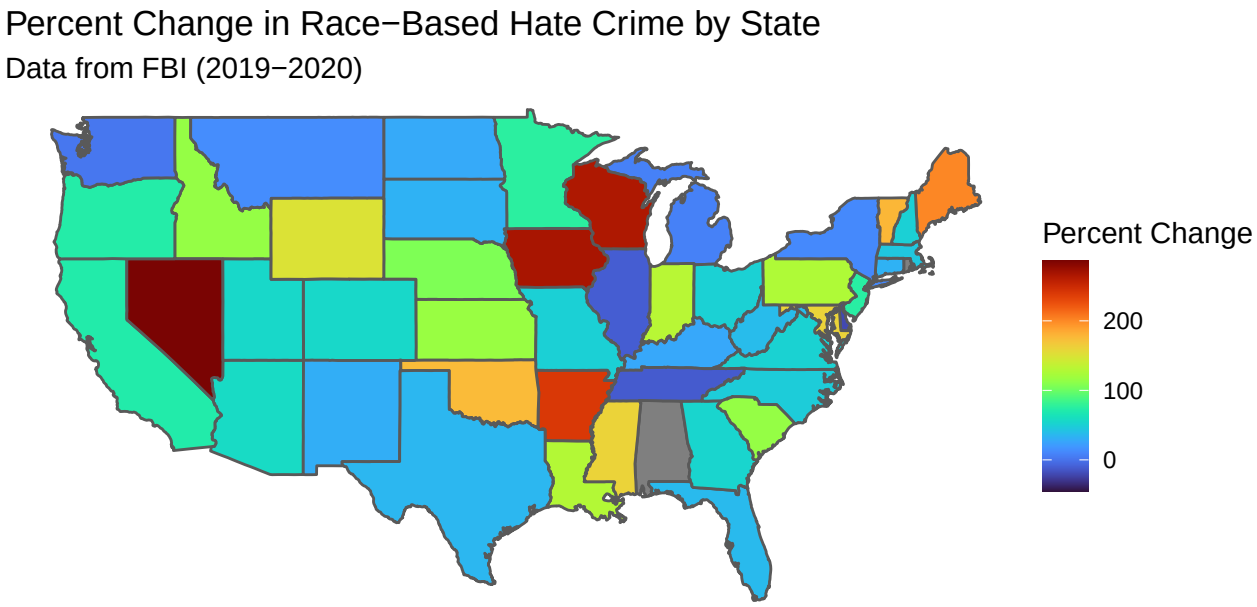
Population & Income Dataset

For the second dataset that answers the third question (named "state_race_population.csv" in the data folder), we joined the variables for income, population, and specific race population by state. We then added the hate crime variables for 2020 and the change in hate crimes between 2019-2020 from the first dataset to this one. The final variables for the second dataset are:

- State
- Change in race-based hate crimes (2019-2020)
- Count of total hate crimes by specific biases (2020) for all biases noted in the previous dataset's variables
- Median household income per state (2019)
- Population per state (2019)
- Race-specific population per state (2019)
  - White
  - Black
  - Hispanic
  - Asian
  - American Indian or Alaska Native
  - Native Hawaiian or Other Pacific Islander
  - Multiple Race

Analysis

Change in Race-Based Hate Crimes Between States



States Greater than 200% Change
Arkansas
Iowa
Maine
Nevada
Wisconsin

States With Negative Percent Change
Alaska
Delaware
District Of Columbia
Hawaii
Illinois
Tennessee

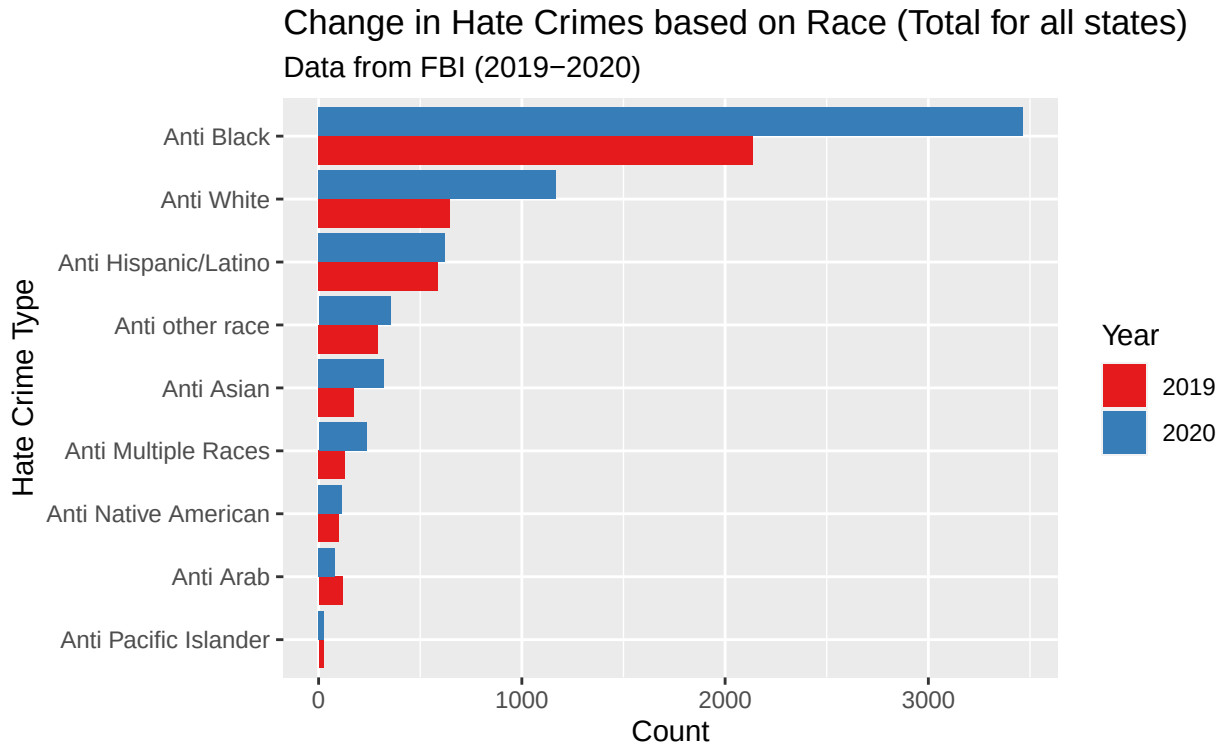
To answer our first question regarding which states have seen the most significant change in the count of race-based hate crimes between 2019 and 2020, we produced a map of the U.S. to determine the states with the most change. We also decided to use this visualization to see if there was any spatial patterns for percent change in race-based hate crimes.

The map above shows the percent change in race-based hate crimes from 2019-2020 in the USA. We learned from this visual that most states seemed to have an increase in hate crime from 2019-2020. In particular, the

states with a change greater than 200% stand out. Based on the table, the 5 states with a change greater than 200% between 2019-2020 were Arkansas, Iowa, Maine, Nevada, and Wisconsin. This means that in these states, hate crimes overall more than doubled. Three out of the five states are in the Midwest, but there seems to be no geographical correlation between each of these states.

On the other hand, there are a handful of states that improved their hate crime frequency. Based on the map, these states are also scattered all over the country, and there is no geographical correlation between the states. The table for states with a negative change in hate crime count shows that Alaska, Delaware, D.C, Hawaii, Illinois, and Tennessee all saw a decrease in hate crimes between 2019-2020. Also, Hawaii and Alaska do not show up on the visual map because the template that we used only includes the continental United States.

Change in Hate Crimes based on Race



For our second question about which specific race-based hate crimes have significantly changed during 2019-2020, we created a side-by-side bar plot to see the change for each type of race-based hate crime with each bar color-coded as the specific year. The data for this visualization was first gathered from the raw all hate crimes data set, which contains counts of the number and type of hate crime by state. We created a new observation where the state is the United States. This observation holds the sum of all the hate crimes for the country excluding Alabama which lacked hate crime data for one of the two years.

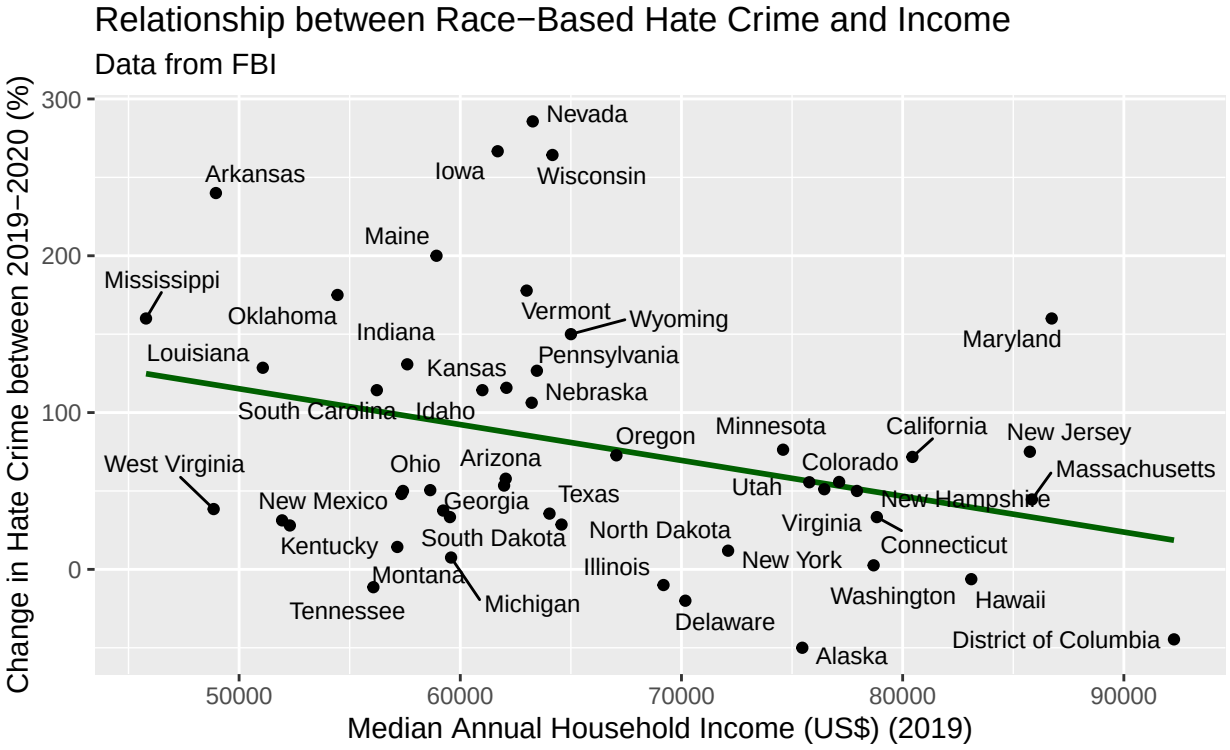
There seems to be a significant increase in hate crimes between the two years. In 2019, there were only 2,109 cases, but in 2020, there were 3,195 total cases. This is a staggering increase of 51.4936 percent. Most groups experienced roughly the same or more hate crimes except the Arab community which dropped from 59 cases to 40.

The largest absolute increase of crimes was amongst Anti-Black or African American hate crimes which increased by over 650 cases from 1,069 to 1732. On the other hand, the largest percentage increase was anti-Asian hate crimes which increased by 82.95 percent, followed by anti-multi racial crimes and anti-White crimes.

These results indicated that the social media was correct in saying that there was an increase of hate crimes in amongst these two groups who lead in absolute values and percentages. However, we must also acknowledge that almost every group experienced an increase in hate crimes which might mean that 2020 was just a hateful year. This could be hypothesized by the pandemic which restricted everyone in many ways. This restrictions could have led to greater hate crimes combined with the medias everpresent indications of more hate crimes.

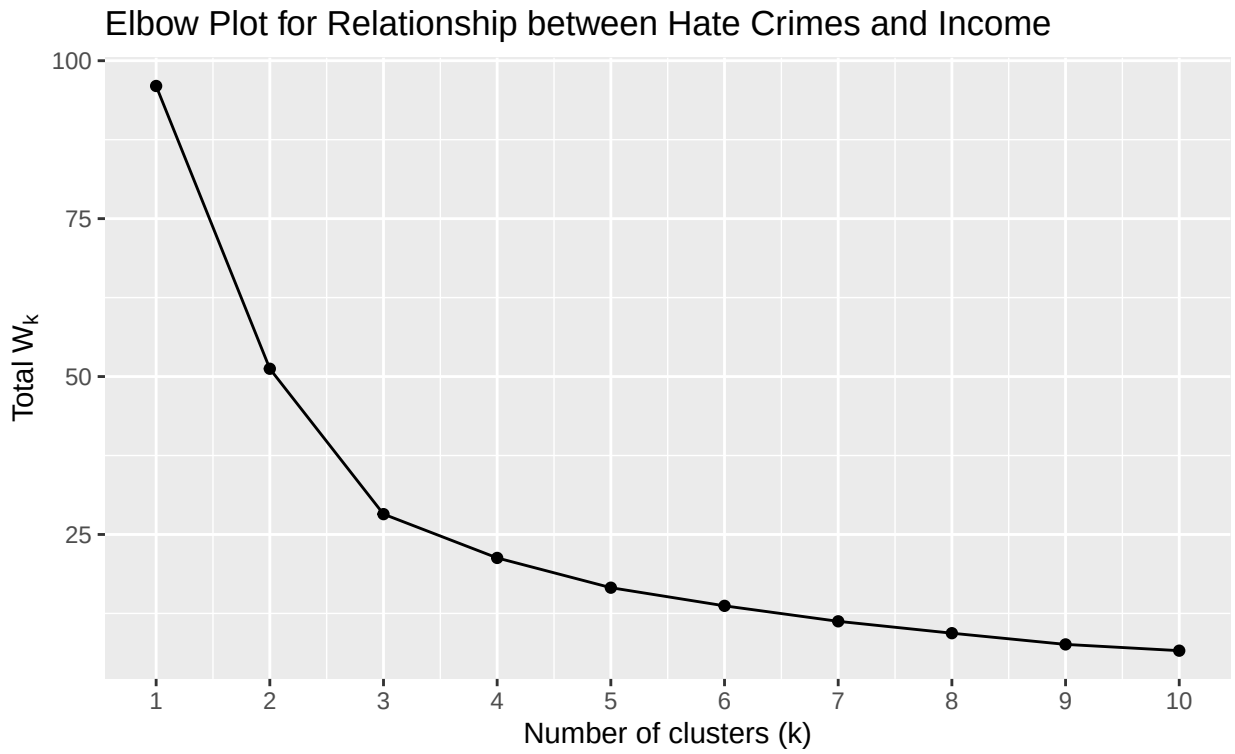
Relationship between Hate Crime & Income

To determine the relationship between Hate Crime & Income, we will run a kmeans clustering method and compare its outcome to the mclust clustering method. The difference between the two methods are that kmeans requires standardization and uses an elbow plot to determine the number of clusters we can create based on minimizing the sum of squares. The mclust method picks the best model based on the Baysian Information Criteria, a goodness of fit statistic. mclust uses mixture models (m is for model) and assumes that the clusters (as they exist) are all multivariate normal distributions, so we don't need to standardize the variables. Thus, the existence of an underlying model in mclust makes it different from kmeans. We will also determine if there are significant differences between each groups by noting the difference between the centers of the clusters.



The scatterplot that shows the relationship between change in race-based hate crimes (2019-2020) and median annual household income (2019) shows a slightly negative correlation between the two variables. This means that as income goes up, the change in hate crimes goes down.

k-means



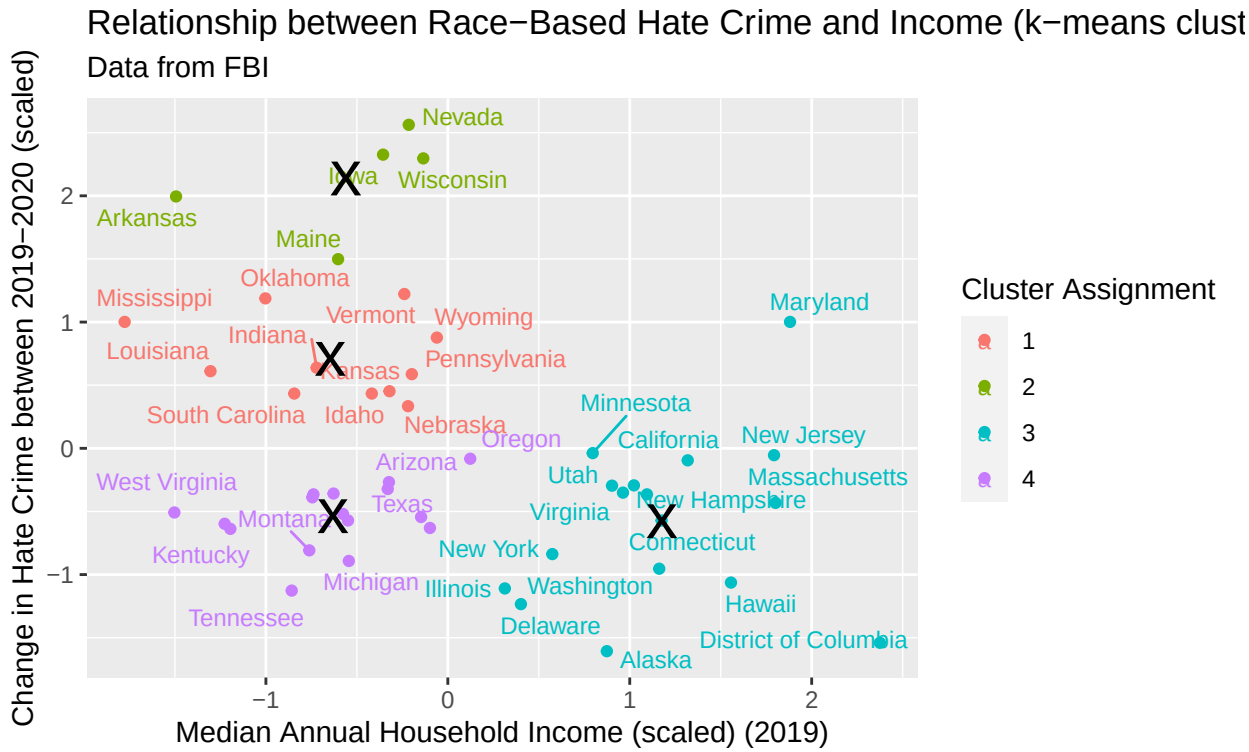


Table 3: kmeans Center of Clusters (Hate Crimes and Income)

Cluster	Median Annual Household Income (x)	Percent Change in Race-Based Hate Crimes (y)
1	-0.6456974	0.7071808

2	-0.5604540	2.1355925
3	1.1767567	-0.5787070
4	-0.6312451	-0.5386833

For the kmeans analysis, we initially created an elbow plot to determine the number of clusters to group up the observations in. We decided to use 4 clusters as it sufficiently minimizes the sum of squares and is a reasonable amount of clusters. Before plotting the observations, we standardized them. In the scatterplot above, the clusters are visually indicated by color and the 'x' marks indicate the centers of each cluster. Based on the graph, we see that the centers of clusters 1, 3, and 4 have similar values for median annual household income. The table tells us that, compared to the other clusters, the center for cluster 2 is around .6 units higher in income. In terms of change in hate crimes, cluster 3 has the highest values, followed by cluster 4, then clusters 1 and 2 are close (based on the table).

mclust

```
# Code for mclust analysis
incomesub <- income_sample %>%
  select(median_annual_household_income, percent_change_all)
mclustsol <- mclustBIC(incomesub)
income_mclust <- Mclust(incomesub, x = mclustsol)

# Plots mclust clusters using ggplot
income_sample_mclust <- income_sample %>%
  mutate(mclust_class = as.character(income_mclust$classification))

# Find centers for each cluster
income_center_x <- income_mclust$parameters$mean[c(1,3)]
income_center_y <- income_mclust$parameters$mean[c(2,4)]
income_center <- data.frame(income_center_x, income_center_y)
```

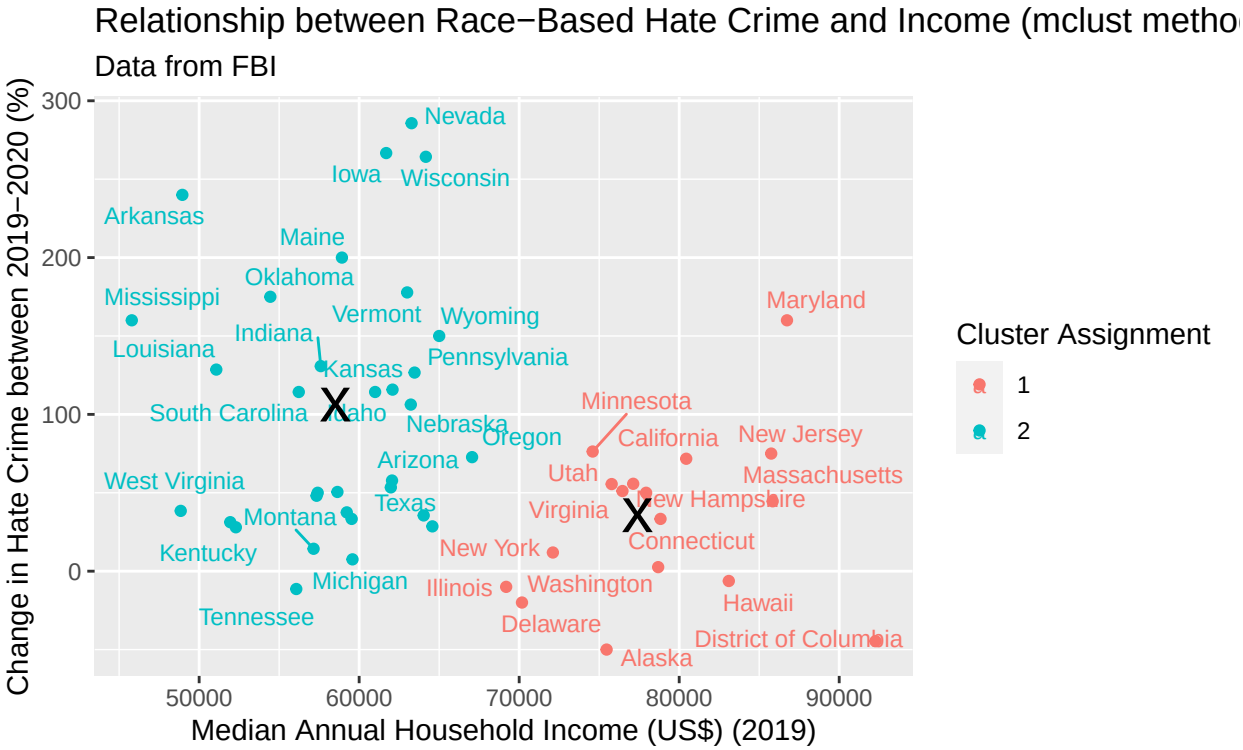


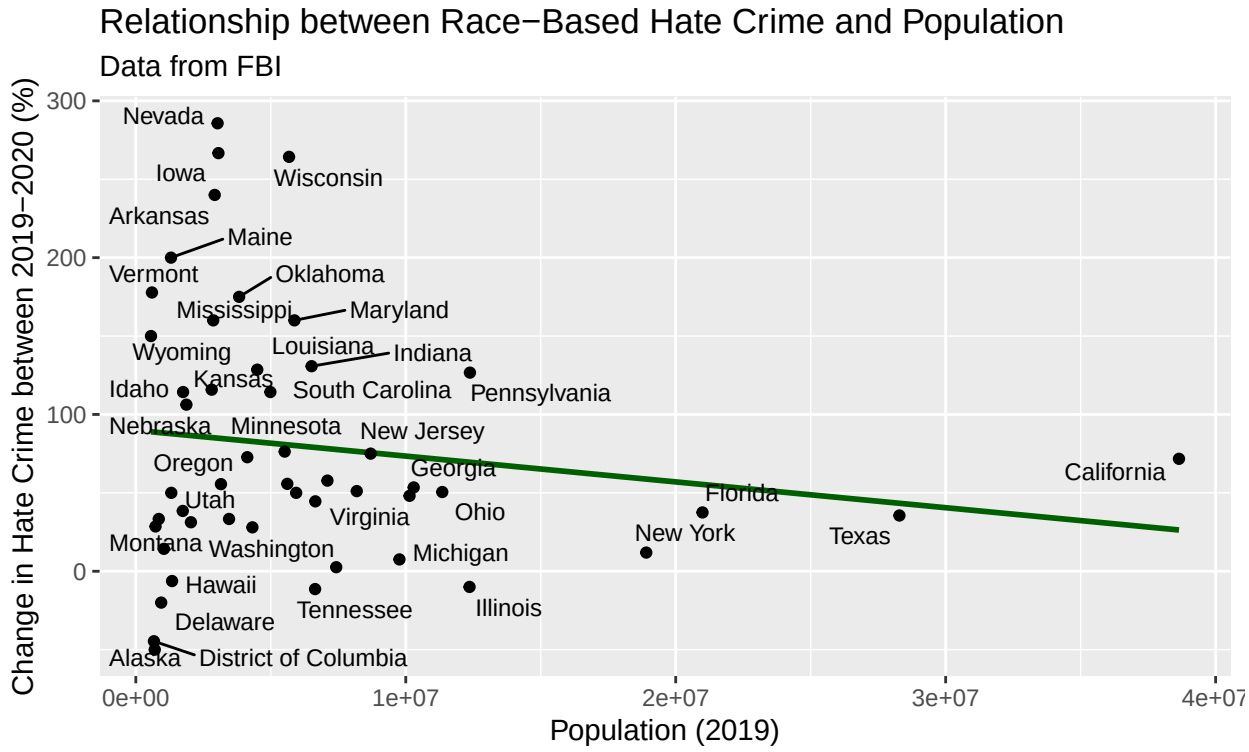
Table 4: mclust Center of Clusters (Hate Crimes and Income)

Cluster	Median Annual Household Income (x)	Percent Change in Race-Based Hate Crimes (y)
1	77391.45	35.57645
2	58514.44	106.13410

Compared to the kmeans method, the best model that our code gives us when running the mclust method has two instead of four clusters. Based on the graph and table, cluster 1 has a lower change in hate crimes on average and higher values of income. The states in cluster 2 all have lower incomes compared to the states in cluster 1; however, the change in hate crimes has a larger range, but on average it is around 71 units higher than cluster 1's average.

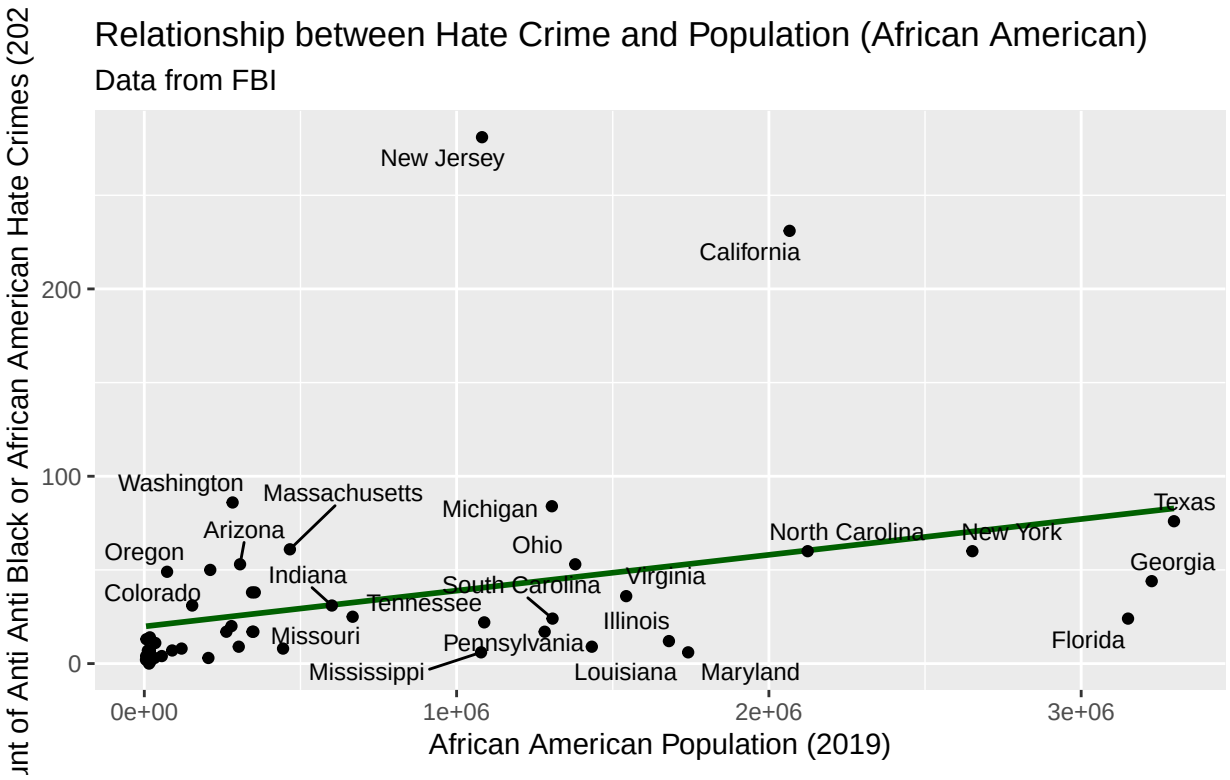
Relationship between Hate Crime & Population

To determine the relationship between change in race-based hate crimes and population as well as how we can group up states to see if there are any significant differences, we also compare the results of running a kmeans clustering analysis and mclust analysis like we did in the previous section.

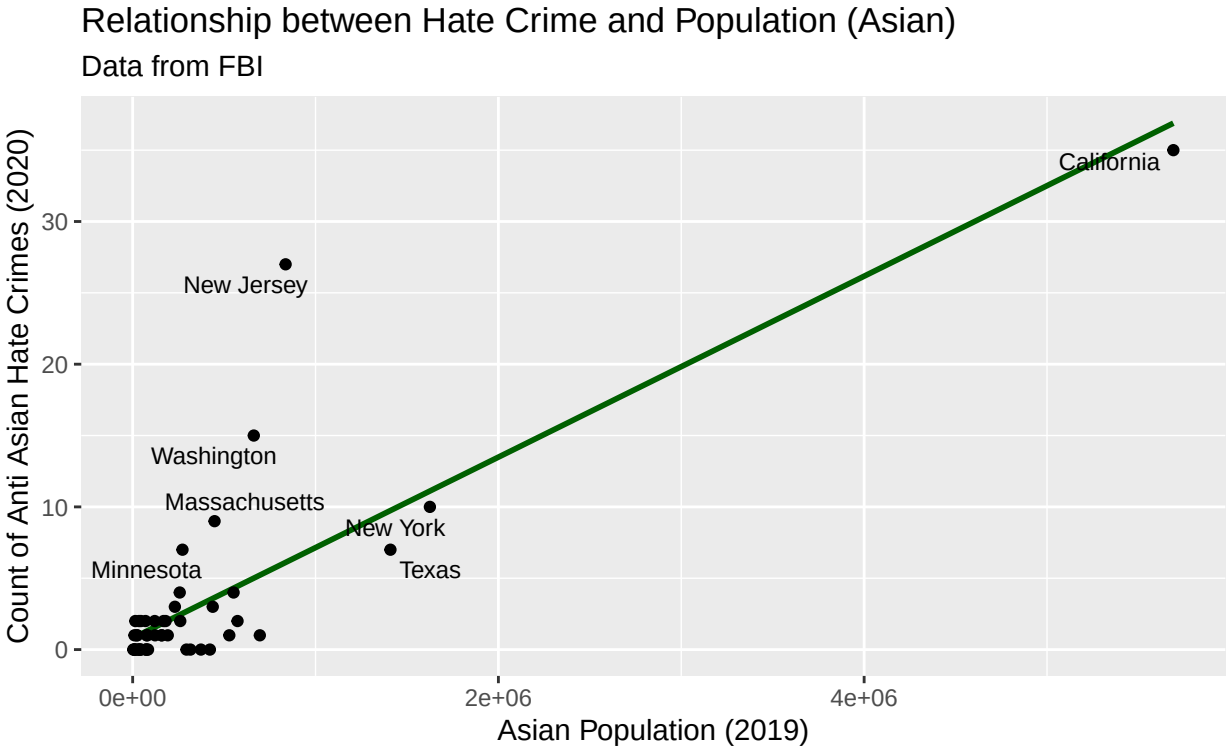


The scatterplot above shows the relationship between Hate Crime and Population. Based on the regression line, there seems to be an overall slight negative relationship between the two variables. However, most of the states are concentrated on the left side (where population is smaller). Thus, we should be weary of the large populations and potential outliers of California, Texas, Florida, and New York.

As noted in the analysis of our second question (change in specific race-based hate crimes), Anti-Black or African American had the largest positive change between 2019-2020, followed by Anti Asian-American. Thus, we plot the relationship between these race groups and their corresponding populations for each state.

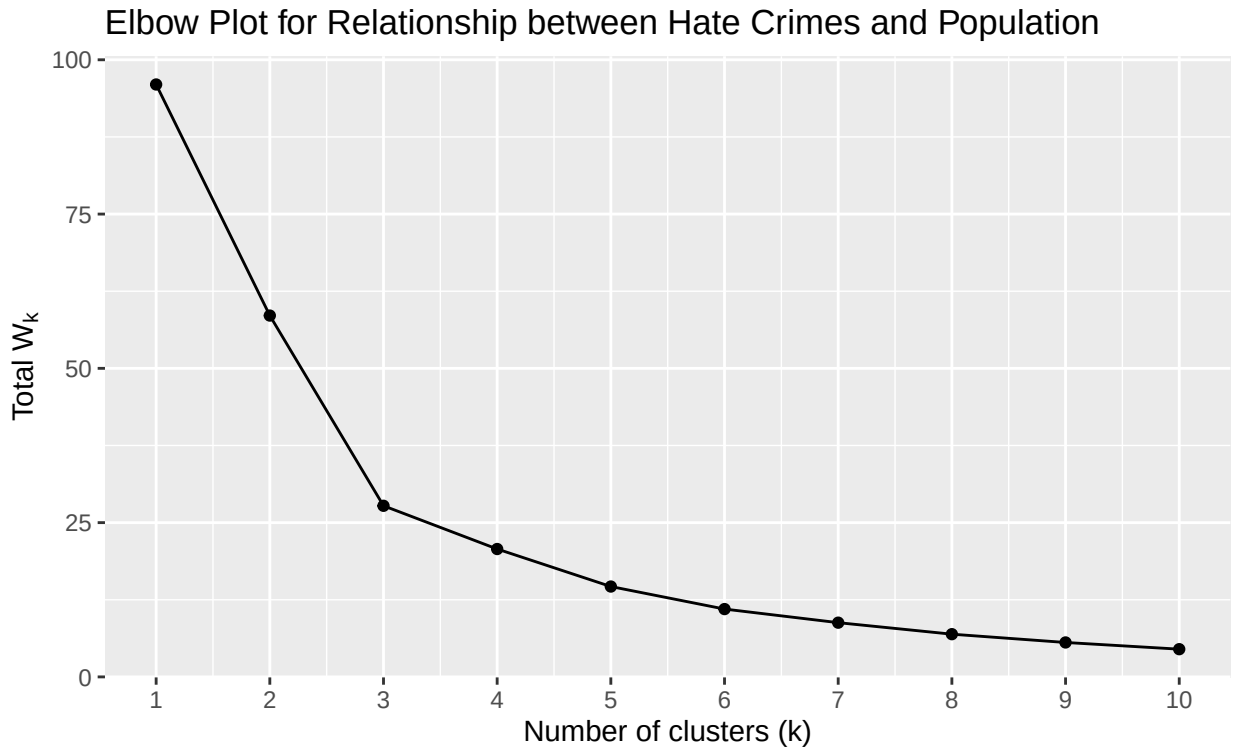


The scatterplot that shows the relationship between the count of anti-black or african american hate crimes in 2020 and black population for each state in 2019 shows a slightly positive correlation. This means that as the black population of a state increases, the count of hate crimes against black people increase. New Jersey and California look like potential outliers as they are significantly above the regression line in regards to the count of anti-black hate crimes.



The scatterplot that shows the relationship between the count of anti-asian hate crimes in 2020 and the asian population of a state shows a pretty positive correlation between the two variables. However, this correlation could be influenced by California, which has a considerably higher value for both variables compared to other states. New Jersey also has a higher amount of anti-asian hate crimes compared to other states with a similar asian population size.

kmeans



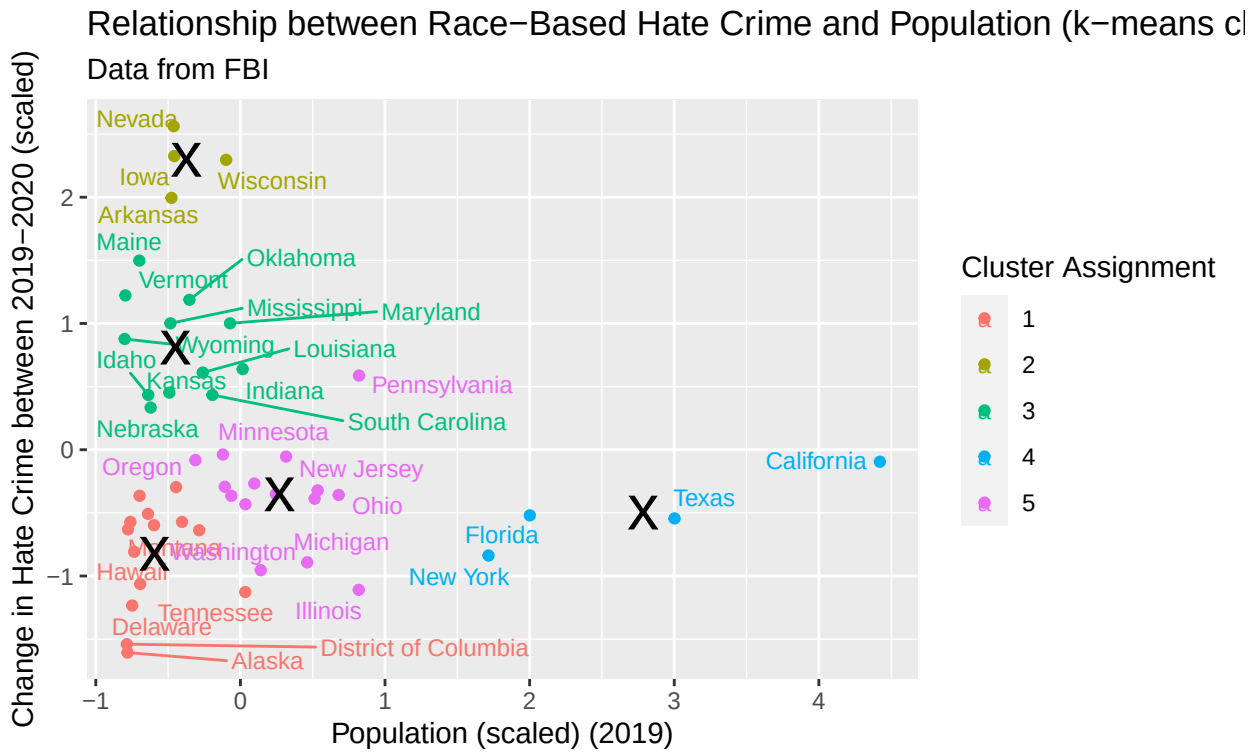


Table 5: kmeans Center of Clusters (Hate Crimes and Population)

Cluster	Population (x)	Percent Change in Race-Based Hate Crimes (y)
1	-0.5937970	-0.8254057
2	-0.3736677	2.2949608
3	-0.4491545	0.8075877
4	2.7851669	-0.4991387
5	0.2704677	-0.3545774

From the elbow plot, we chose to use 5 clusters as the number of clusters since it minimizes the sum of squares is still a reasonable amount of clusters. We also standardized both variables for the kmeans analysis. The scatterplot visually indicates the clusters by color and the 'x' marks indicate the centers for each cluster. From the table, the centers of clusters 1, 2, and 3 are pretty close in regards to population. However, in terms of change in hate crimes, clusters 2 has the highest values followed by cluster 3. Cluster 4 have the largest values for population but a relatively low change in hate crimes value similar (0.4 unit difference) to cluster 1.

mclust

```
# Code for mclust analysis
populationsub <- population_sample %>% select(total, percent_change_all)
mclustsol <- mclustBIC(populationsub)
population_mclust <- Mclust(populationsub, x = mclustsol)

# Plots mclust clusters using ggplot
population_sample_mclust <- population_sample %>%
  mutate(mclust_class = as.character(population_mclust$classification))
```

```
# Find centers for each cluster
population_center_x <- population_mclust$parameters$mean[c(1,3,5)]
population_center_y <- population_mclust$parameters$mean[c(2,4,6)]
population_center <- data.frame(population_center_x, population_center_y)
```

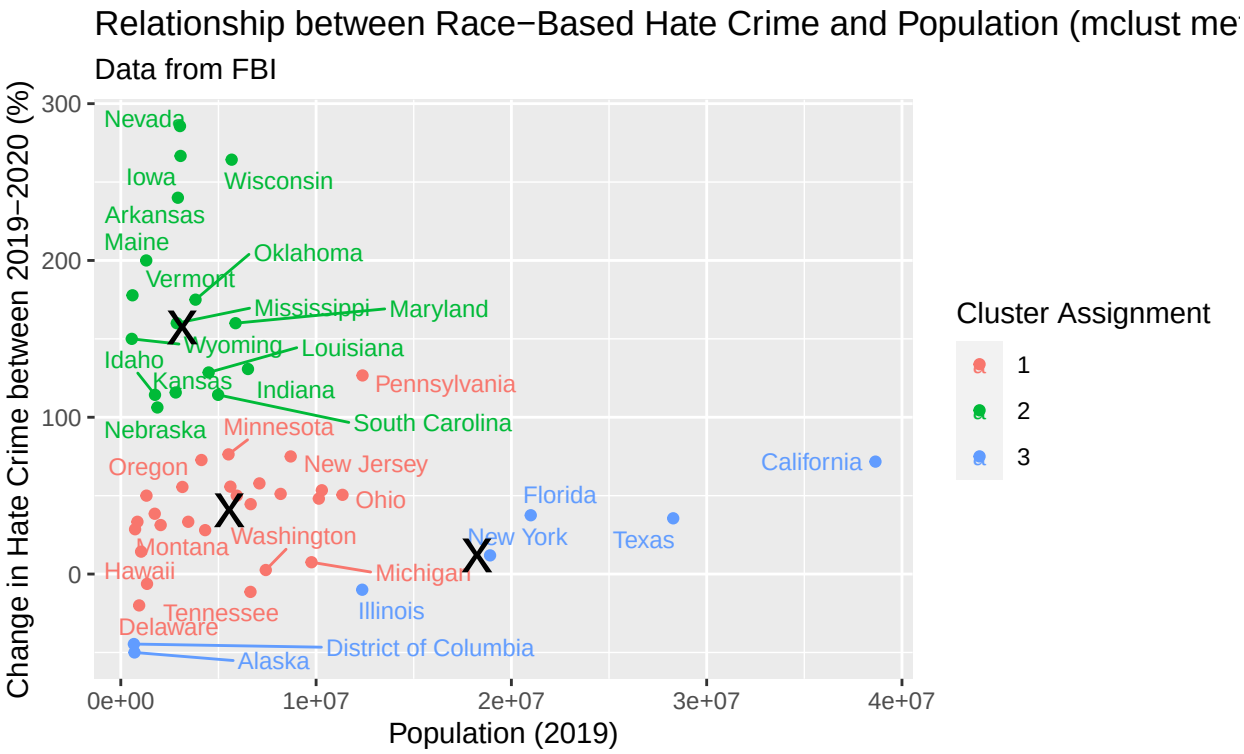


Table 6: mclust Center of Clusters (Hate Crimes and Population)

Cluster	Population (x)	Percent Change in Race-Based Hate Crimes (y)
1	5554288	40.56439
2	3146042	157.16873
3	18237784	11.77005

Compared to the kmeans analysis, the most optimal model that the mclust function gives has three rather than five clusters. The centers for the clusters show that cluster 2 has the highest values for change in hate crime and relatively lower values for population compared to the other clusters. Cluster 3 has larger population values on average and lower change in hate crime values on average compared to the other clusters. Cluster 3 has the least amount of states and seems to be grouped together based on low change in hate crime values (since the populations of the states have the largest range). From the table, the center of cluster 1 indicates that cluster's average is in between the two clusters for both variables.

Conclusion

Regarding the data, we discovered that the two states with the most significant change in count of race based hate crimes between 2019 and 2020 are California and New Jersey. They had a positive increase of 180 and 171 cases respectively. This is significant because all other states had less than a 50 count change. The race based hate crimes with the most significant change over the years are Anti-Black or African American, Anti-White, Anti-Asian, and Anti-Multiracial. From the map, we see that the states with the highest increase in hate crime percentage of at least 200 percent are Nevada (286%), Iowa (267%), Wisconsin (264%), Arkansas (240%), and Maine (200%).

Looking at the scatterplots, we see that the relationship between change in race-based hate crimes (2019-2020) and population (2019) is slightly negative. This means that when income goes us up, hate crimes go down. However, we have to be careful analyzing this result because the four more highly populated states California, Texas, Florida, and New York might be potential outliers. For the relationship between change in race-based hate crimes (2019-2020) and median income (2019), there is a negative correlation which means that as income increases, hate crimes decrease.

Through our clustering analysis, the significant differences between groups was determined through the differences in the cluster centers. We saw that the differences between cluster centers depended on which method we ran. Also, we ran 2 bi-variate models because it is easier to visualize the relationship between the two variables rather than three or more. It would be interesting to see the relationship between Hate Crime, Income, and Population together.

When comparing the two clustering methods we also wanted to determine which method gave us more information about the groups of states. In our opinion, we prefer to use the kmeans clustering as the additional number of clusters provides more information about the differences between each cluster. With just two or three clusters with mclust, it seems too general of a divide between the states in income or change in race based hate crimes (i.e the divide seems more dependent on one variable rather than a mixture of the both variables). In addition, it makes more sense to standardize the variables we used since the units for change in hate crimes (%), income ($), and population (# of people) are inherently different.

Limitations

As our data for each state is an aggregate of every city and county, the results of our analysis are very generalized for each state and hate crime. For further, more accurate research, finding a city or county-based statistics would be better; however, we were limited in the resources we could find. As noted for the population & income data, we are limited to 2019, so using data for 2020 would also help make our analysis more accurate since the count of hate crimes we use will be from 2020 (although this is only for two visualizations). In addition, whether an event is considered a "hate crime" and what race the victim identifies as is subjective and depends on each agency and individual. Another limitation is that we could not gather data from all 50 states. We could not find information on hate crimes in Alabama and Rhode Island for 2020 because either the data was lost or wasn't reported. Thus, the change in Alabama and Rhode Island can't be included since there is no change when only one year is given.

Regarding our clustering analysis, when plotting the relationship between Hate Crime and Population, we see that there could be potential outliers that could influence the correlation between the two variables. Our analysis focused on clustering the states to see their relationship with one another and identifying groups of states based on the two variables. However, we should be wary of checking conditions if we want to determine the statistical significance of the data.

Also, we did not find data on the perpetrator's race of these hate crimes. This data would have been a fascinating insight into the relationship between races, and we would have run a network analysis that would have explored the relationship between races.