Racial Divergence of Aggravated Assault Arrests

Ivanshu Kaushik ik346
Mehak Malik mm3646
Jayant Chaudhari jc3303

1 Introduction

As we step into the 21st century, the United States stands out as the most racially diverse and affluent country globally. Despite notable economic progress, these gains are not distributed evenly throughout society, leaving certain segments of American communities marginalized and seemingly overlooked by the broader society.

Central to this narrative of marginalization is the pervasive and entrenched disparate treatment of individuals of color, a phenomenon that transcends the entire spectrum of the American criminal justice system. This pervasive issue is deeply intertwined with the broader fabric of structural racism, which permeates various domains within society, influencing social, economic, and political dynamics.

When the proportion of a racial/ethnic group within a certain ethnicity is greater than the proportion of such groups in the general population, it is said that there is "racial disparity" in the criminal justice system. The criminal justice system is riddled with ingrained racial biases, creating disparities at every juncture. These biases not only affect immediate decisions but also have a cascading impact, leading to adverse outcomes for people of color. To foster the development of fairer policies and research, it's crucial for researchers to delve into their work with a comprehensive awareness of how racist policies and implicit biases intertwine across various facets of the criminal justice system.

In this study, we aim to answer the following questions using statistical methods we learned in the course.

- Is there any relevance to the observed variations in the incidents of Aggravated Assault arrests among different racial groups in the United States during the period spanning 2017 to 2019?
- Is there a significant correlation between the degree of population density in regions (suburbs vs non-suburbs) and the incidence of reported crimes across various racial demographics?
- Is there a consistent pattern across diverse racial groups wherein rates of criminal activity exhibit fluctuations, with a tendency for increased incidence during the summer months as opposed to the winter season?

2 Dataset

2.1 Description

The data utilized for our analysis comes from the Uniform Crime Reporting Program, which includes arrest data for the United States for the years 2017, 2018, and 2019, broken down by age, sex, and race. The following criteria are used to further divide the data:
1.States: All 50 of the United States' states have crime report records available to us.
2. Race: There are five categories within the dataset: Asian, White, Black, Indian, and Hispanic.
3. Age: The population as a whole is split into two groups: adults (18–70) and juveniles (0–18).
4. Area Wise: Whether or not the city is located in a suburban area. A city with fewer than 50,000 residents is considered a suburban area.
5. Offense code: Out of all the offenses listed, code 04 denotes aggravated assaults.
6. Gender: Age is the dividing line between men and women.

Variable	Description
STATE	the full list of states included in the dataset
YEAR	It includes the year the arrest occurred, such as 2017/2018/2019
SUB	It indicates whether or not the area is suburban.
CORE	It indicates if the city is part of the core or not.
MONTH	It indicates which month of the year the arrest was made.
AW	the total count of cases for Adult Whites
AB	the total count of cases for Adult Blacks
AI	the total count of cases for Adult Indians
AA	the total count of cases for Adult Asians
AH	the total count of cases for Adult Hispanics
JW	the total count of cases for Juvenile Whites
JB	the total count of cases for Juvenile Blacks
JI	the total count of cases for Juvenile Indians
JA	the total count of cases for Juvenile Asians
JH	the total count of cases for Juvenile Hispanics

2.2 Data Pre-processing

1. Obtaining the Dataset: The NCVS website provided the dataset files for the years 2017, 2018, and 2019.
2. Bringing in all the libraries: To create high-level design graphs, ggplot was utilized, and libraries like tidyverse and dplyr were brought into it.
3. Assembling the dataset:

- The Jupyter notebook was populated with all three datasets.
- For the offense (04) of aggravated assaults, three subsets of the original datasets were constructed.
- One dataset, combined-df, was created by combining the three subsets.
- Over the course of the analysis, all NA values were changed to 0.
- A summer column was added, with values of '1' for the months (May, June, July, August, September, October) and '0' otherwise.
- There was encoding for the categorical columns SUB and CORE.
- The data was grouped according to STATES and the total count of all the races was summarized because the data included multiple row entries for a single state.
- There were 25 columns and 165681 rows in the final dataset.

3 Existing Analysis

- *Christopher D. Maxwella , Amanda L. Robinsonb, Lori A. Post* did a study on The impact of race on the adjudication of assaults

  From the study with 21,397 assault cases, whites constituted 32 percent, Hispanics 11 percent, Asians 3 percent, and blacks the highest at 54 percent. Another analysis on the same data showed that white offenders had approximately 22 percent higher odds of arrest for robbery, 13 percent higher for aggravated assault and 9 percent higher for simple assault compared to black offenders.

- *Aki Roberts Christopher J. Lyons usined a logistic regression model to predict Victim–Offender Racial and Clearance of Lethal Assault and found out the likelihood of the accused being acquitted*

  The findings suggest that incidents involving Whites have the highest probability of clearance, followed by interracial incidents. In contrast, aggravated assault incidents among non-Whites have the lowest likelihood of being resolved.

- *Casey T. Harris, Darrell Steffensmeier, Jeffrey T. Ulmer, and Noah Painter-Davis worked on Racial and Ethnic Disproportionality Between Arrest and Incarceration*

  It concluded that the proportion of blacks, whites, and Hispanics in the state prison admissions and prison population closely mirrors their representation in arrest statistics. The overall pattern reflects a low presence of whites, a substantial presence of blacks, and a relatively moderate representation of Hispanics at every step of the criminal justice system.

4 Methodology

4.1 EDA

Monthly and Yearly cases of aggravated assaults

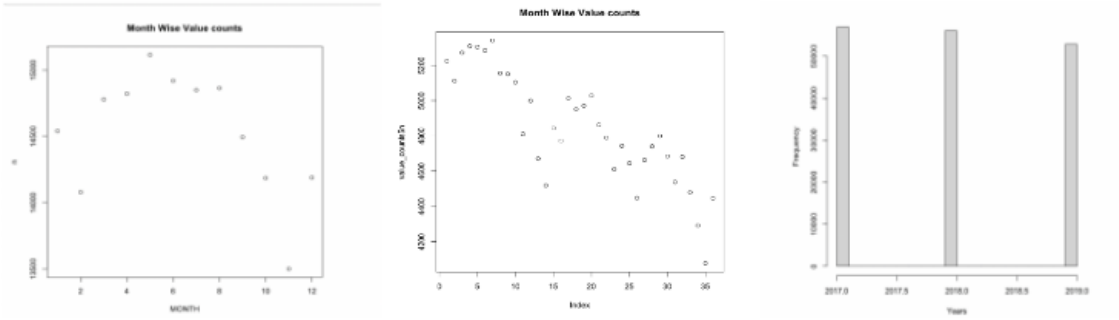


Figure 1: Monthly and Yearly cases of aggravated assaults

It represents the Monthly cases for the year 2019. Here we can observe that Most cases are reported in Summer It represents the Monthly cases for the year 2017-2019. Here We can see the assault cases decreasing through the Year

Top 10 states for aggravated assaults by Races compared to population of the states

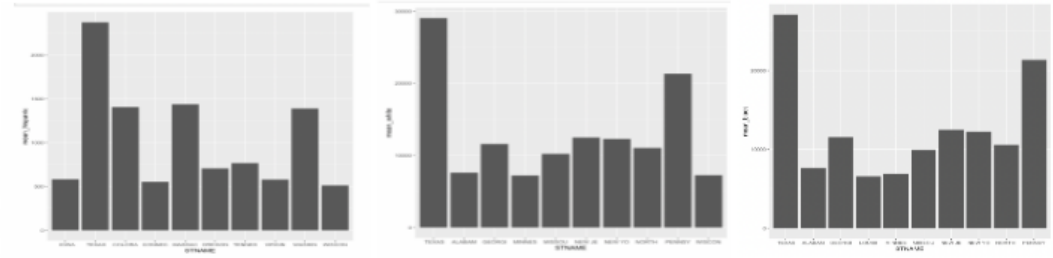


Figure 2: Top 10 states for aggravated assaults by Races compared to population of the states

From Fig 5 we infer that Texas has Most Assaults reported and the top 10 states are consistent Irrespective of Races While the racial distribution is approximately same for all races except Whites.

From Fig 5 it becomes apparent that the average number of reported cases against White individuals is greater than that for other racial groups. The limited occurrence of cases involving juveniles can be attributed to a scarcity of reported incidents, largely because schools primarily handle and resolve such matters.

In Fig 4 we can see that more cases are reported in non-suburban area. One of the reasons can be differences in policing policies and lack of job opportunities. This trend is consistent with the data for Past two Years(2017-2018)

The plots in Fig 5 indicate that crime cases are generally lower in suburban areas compared to non-suburban areas across all races, suggesting higher crime rates in non-suburban regions. Categorization of states into core and non-core is based on the number of reported crime cases in each state.

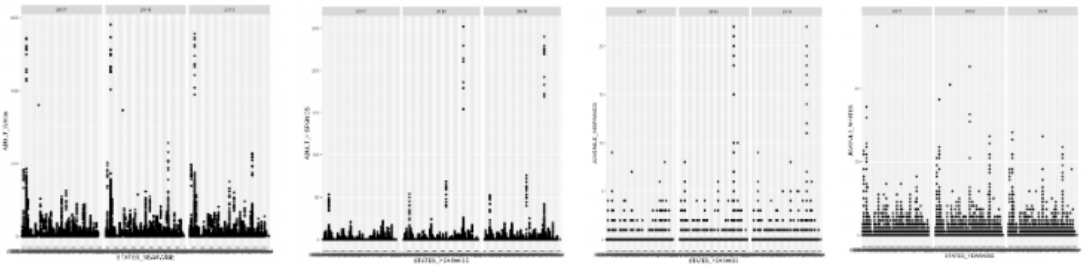


Figure 3: ADULT and Juvenile–Year wise Distribution of assault cases by States and Filters By races

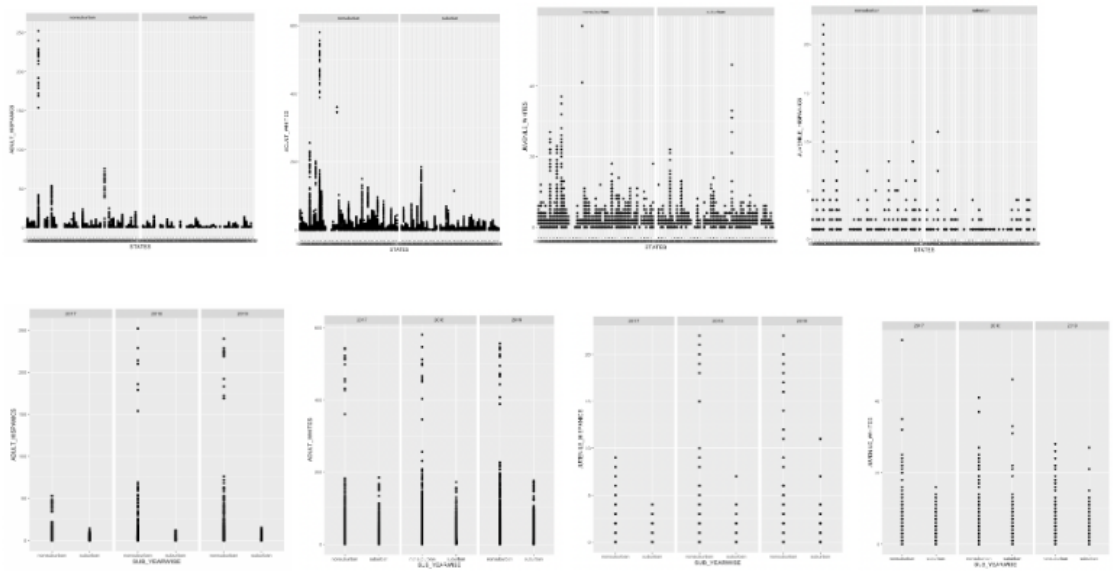


Figure 4: ADULT and Juvenile-Sub/Nonsub Distribution of assault cases by States and Filters By races

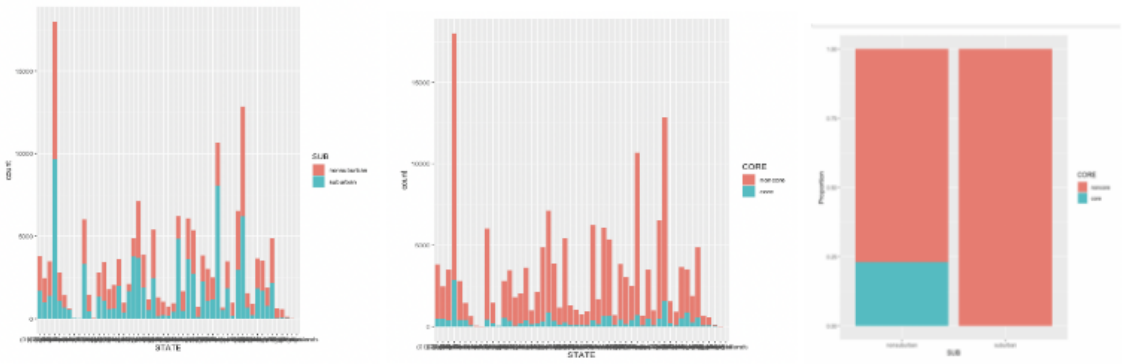


Figure 5: No. of Reports per States Categorized By Sub/Non-sub, Core/Non-core, Core/Sub

4.2 Regression

In the given dataset, we try to fit several regression models to predict the total number of arrest cases using several independent variables. Before we begin modelling the problem, we need to convert several columns into factors as they only take a limited number of values. As described in the dataset description, State is a column that can only take 53 values, so we convert it to a categorical variable. Similarly, Suburban is also converted into a categorical variable. A variable *summer* is generated which

indicates whether a given month is in the warm months or in the comparatively colder months.

Initially a model to predict the total number of arrest cases is formulated using independent variables as

$$Total = \beta_0 + \beta_1 * STATE + \beta_2 * YEAR + \beta_3 * SUB + \beta_4 * CORE + \beta_5 * MONTH + \beta_6 * avg\_total + \beta_7 * summer$$

This is not the final model, and just an experiment to see how much variance of the total arrests we are capable of capturing without any improvements on the base model. An adjusted R^2 score of 0.143 suggests that this is not a very capable model. However, as we add the number of adult whites as an extra variable, we are able to explain the variance of the total number of arrests very well. The adjusted R^2 score increases to 0.705 which suggests that there is a very high correlation between the total number of arrests and number of adult whites arrested. A similar observation is apparent when we add the number of adult blacks arrested in place of the adult whites arrested, while on adding Adult Asians and Adult Indians, no significant change is observed. This suggests a very high correlation between total number of arrests and the number of adult white and black arrests. We will explore these individually in the next section.

4.2.1 Base Model

Now, we shift our focus on the analysis of a single racial category (Adult whites). The base model is defined as

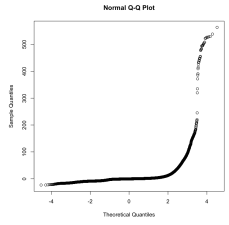
$AW = \sum_{i=1}^{8} \beta_i * X_i$ where X consists of State, Year, Suburban, Core city, Month, avg_total, avg_adult, summer

The result of this model is shown below.

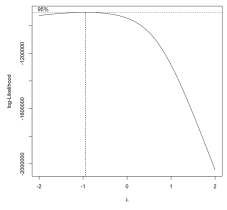
```
Residual standard error: 10.48 on 165621 degrees of freedom
Multiple R-squared:  0.1189,    Adjusted R-squared:  0.1186
F-statistic: 378.7 on 59 and 165621 DF,  p-value: < 2.2e-16
```

Figure 6: Output of the base model.

The output suggests that the initial model does not perform as we expect it to. This in turn suggests that our base model needs to be improved. A closer look on the normality assumption of the given regression model suggests that the given data should be transformed before proceeding as the error assumptions are not satisfied. We proceed to apply a box-cox transformation on the given data.



(a) Quantile-Quantile Plot

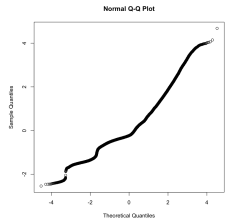


(b) Likelihood function of the Box-cox transformation

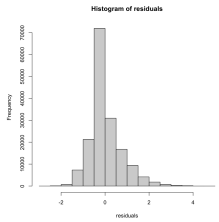
Figure 7

The Box-cox transformation suggests an optimal value of -1 to transform the regressor variable. Even though the output of this shows an increased R^2, on careful observation we notice that most of the variables are not significant anymore while the test is overall significant. Further, the y values have all become equal to $\overline{Y}$. We need to apply some other transformation on the given data as the output of Box-cox transformation is redundant. Applying log transformation on the given data improves the model performance and is the transformation we proceeded with for the improved model.

As is evident in Figure 8, the data is much less skewed now, even if not perfect. The residual plot also shows a more normal distribution of the residuals after the transformation. We proceed with this data to get an improved model in the next section.



(a) Q-Q Plot of the transformed data



(b) Histogram showing the residuals.

Figure 8

4.2.2 Improved Model

As we begin with the analysis, it is imperative to begin with the formulation of the model before we can proceed. The model is currently formulated as

$\log(AW) = \sum_{i=1}^{8} \beta_i * X_i$ where X consists of State, Year, Suburban, Core, Month, log(avg_total), log(avg_adult), summer

And the model performance is shown below 9.

```
Residual standard error: 0.7339 on 165621 degrees of freedom
Multiple R-squared:  0.3085,    Adjusted R-squared:  0.3083
F-statistic:  1252 on 59 and 165621 DF,  p-value: < 2.2e-16
```

Figure 9: Output of the improved model.

Now, we need to remove the outliers (8,773 points), leverage points (17,764 points) and influential points from the dataset to further improve our model. The final adjusted R^2 score obtained from our improved model is 0.364.

A view at the correlation plot 10 suggests that the number of white adults arrested is directly correlated to the number of people reported in a crime. We can make use of this to further improve our model.

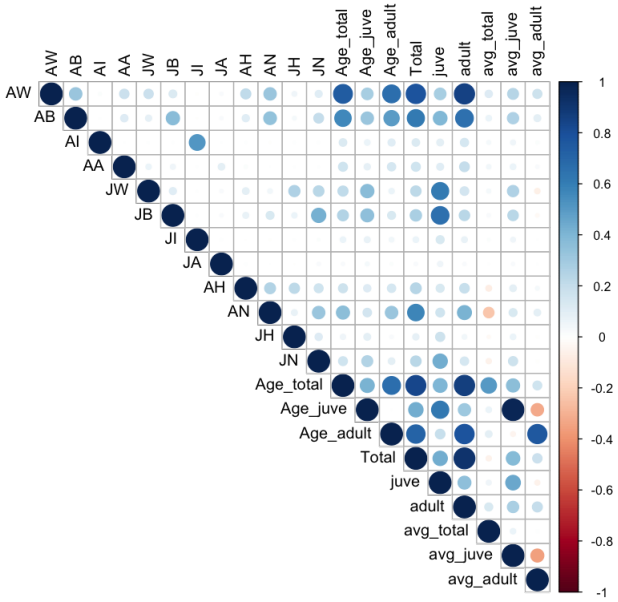


Figure 10: Correlation plot

As discussed earlier too, the number of people reported is shown to have a very high correlation with the number of white people arrested. We can use this fact to our advantage in our analysis of the given

dataset. We can add one more predictor variable and check the model performance. The new model shows a significant increase in the adjusted R^2 score. This further supports our initial assumption that the number of adult whites arrested is dependent on the total number of people reported. The adjusted R^2 score obtained from this new model is 0.774. Next, we remove the columns which inflate the variance of the given model. A look at the VIF of the predictor variables shows that State is inflating the variance significantly and we can create another model without using STATE. Even though the adjusted R^2 score is less for this model, this is a much simpler model than the previous one and simpler models are preferred over more complex models due to the *Theory of Parsimony*. Finally, we use stepwise regression to find a model with the lowest AIC score. The model is given by

$$\log(AW) = \beta_0 + \beta_1 * SUB + \beta_2 * CORE + \beta_3 * Month + \beta_4 * summer + \beta_5 * log(adult)$$

And the performance of the final model is given below.

```
Call:
lm(formula = AW ~ SUB + CORE + MONTH + summer + adult, data = combined_df)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5722 -0.0001  0.0212  0.2027  1.6079

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.0103261  0.0026079  -3.960 7.51e-05 ***
SUB1        -0.0198676  0.0020952  -9.482  < 2e-16 ***
COREY       -0.0575788  0.0038751 -14.859  < 2e-16 ***
MONTH        0.0007321  0.0003021   2.424   0.0154 *
summer1      0.0045539  0.0020579   2.213   0.0269 *
adult        0.7192277  0.0012783 562.635  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3706 on 143460 degrees of freedom
Multiple R-squared:  0.7331,    Adjusted R-squared:  0.7331
F-statistic: 7.88e+04 on 5 and 143460 DF,  p-value: < 2.2e-16
```

Figure 11: Performance of the Final Model

4.2.3 Hypothesis Testing

In this section we delve deep into the fundamental questions that we want to answer through this analysis.

Racial Divergence: Once we have our final model, we can go ahead and think about some of our initial problem statements. Ideally, the number of people reported should have no effect on the arrest rate of any of the races. To test our hypothesis, our null hypothesis is that the coefficient of *adult* $\neq 0$ while the null hypothesis is that the coefficient of *adult* is 0

$$H_o : \beta_{adults\_reported} = 0 \; vs \; H_a : \beta_{adult\_reported} \neq 0$$

Upon evaluation using the t-test, we obtain the p-value as 0 which indicated that we must reject our null hypothesis. Thus, the total number of people reported has an impact on the number of white arrests. Further, the value of $\beta_{adult} = 0.71$ suggests a significant dependence of log(AW) on log(Total adults reported). This means that as we change the number of adults reported by 1 (log(number of adults)), the number of whites arrested is increased by 0.71 given all other variables are constant. This implies a strong evidence of racial bias in the number of white people arrested. However, one key thing to note is that our analysis does not account for the overall population of whites.

A similar analysis when conducted for blacks also shows the coefficient to be 0.392. Ideally, this coefficient should be close to 0 as we should not be able to predict the number of arrests of any race given the total number of people arrested. For Asians and Indians, this number is very small (in the order of 10^{-2}) indicating less significant dependence on the number of people arrested.

Suburban vs Non Suburban: Ideally, the number of people reported should depend on the the population of the particular region. However, it has also been observed that the reliance of number of people arrested is very less. To test these theories, we conduct a hypothesis testing where we try to understand if White adults are reported more often in suburban areas when compared to non suburban

areas. The null hypothesis is given by coefficient of suburban ¡= 0 vs the alternative hypothesis is that the coefficient of suburban ¿ 0.

$$H_o : \beta_{SUB} \leq 0 \; vs \; H_a : \beta_{SUB} > 0$$

The p-value obtained for this one sided test is 1 which is more than alpha = 0.05. So we fail to reject the null hypothesis and conclude that white people are reported more in non suburban areas. This means that population has practically no impact on the number of people reported and arrested.

Summer effect A very interesting effect that is observed is that there are more violent crimes during summer months than winter months. In this analysis, we also see the racial dependence of this effect. In particular we will see how important relevant summer months are to the number of people arrested of different races.

$$H_o : \beta_{summer} = 0 \; vs \; H_a : \beta_{summer} \neq 0$$

The p-value obtained for this one sided test is 0.01 which is less than alpha = 0.05. So we reject the null hypothesis and conclude that summer plays an important role in the number of adult white arrests. This effect however is not significant for the other races.

5 Conclusion

These findings raise some important questions and identify some patterns that suggests the existence of racial disparities in our current justice system which can be resolved to make the law more fair. We found that given the number of people arrested, it is much easier to estimate the number of white people arrested than any other race. This indicates that there are more white arrests compared to any of the other races. We should keep in mind that we have not accounted for relative population sizes in this study which can be done to have a more extensive answer to the questions we proposed. However, our results are in line with what Christopher D. Maxwella , Amanda L. Robinsonb, Lori A. Post proposed in their study. Further, we also found evidence of racial disparity in the number of people arrested in different weather conditions. This is another strong point of contention and can be researched further to produce more relevant results. Finally, we analysed if more people are reported in the suburbs vs non-suburbs and reported that non suburban area has more crimes reported.

References

https://www.urban.org/sites/default/files/publication/104687/racial-and-ethnic-disparities-throughout
pdf

https://judicature.duke.edu/articles/getting-explicit-about-implicitbias/#:~:text=
Implicit%20bias%20can%20play%20a,pursue%20cases%20against%20Black%20defendants.

https://naacp.org/resources/criminal-justice-fact-sheet#:~:text=A%20trial%20usually%
20begins%20with,of%20the%20court's%20due%20process.

https://www.sentencingproject.org/reports/one-in-five-ending-racial-inequity-in-incarceration/

https://web.archive.org/web/20060601225204/http://www.sentencingproject.org/pdfs/5079.
pdf

https://www.researchgate.net/profile/Aki-Roberts/publication/240700101_Victim-Offender_
Racial_Dyads_and_Clearance_of_Lethal_and_Nonlethal_Assault/links/00b4952bb51660ba0d000000/
Victim-Offender-Racial-Dyads-and-Clearance-of-Lethal-and-Nonlethal-Assault.pdf

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4111266/

Delgado,R.,&Snchez-Delgado,H.(2023).Theeffectofseasonalityinpredictingthelevelofcrime.
Aspatialperspective.PloSone,18(5),e0285727.https://doi.org/10.1371/journal.pone.0285727

https://www.researchgate.net/profile/Lisa-Stolzenberg/publication/236778539_Race_and_
the_Probability_of_Arrest/links/5fffd0d5299bf1408893db76/Race-and-the-Probability-of-Arrest.
pdf

https://eprints.lse.ac.uk/84621/1/Owusu-Bempah_prosecuting_hate_crime.pdf
http://ndl.ethernet.edu.et/bitstream/123456789/8670/1/104.pdf#page=194

https://www.researchgate.net/profile/Franklin-Zimring-2/publication/249188376_American_
Youth_Violence_Issues_and_Trends/links/564d21ae08aefe619b0dce6f/American-Youth-Violence-Issues-and-T
pdf