Miranda Grisa, Jeff Metzel, Esin Onal, Chris Zhu
CMSC 21800 Data Science

Question: Were certain demographics more harshly sentenced in Cook County from 2000-2020?

Introduction

The topic of race has come up numerous times in the current political climate. With the BLM movement gaining traction, we wanted to figure out if their complaints about the justice system being unfair to African-Americans had a basis. We thought that using the Chicago sentencing dataset would be a good choice because it has significant black and white neighborhoods (Chicago is around 45% White and 30% Black).

We wanted to look at whether certain demographics were more harshly sentenced in Cook County from 2000-2020. We looked at the racial groups of black and white, and contrasted prison sentence length for six separate crimes: (1) first-degree murder, (2) possession of a controlled substance, (3) aggravated driving under the influence of alcohol, (4) armed robbery, (5) aggravated unlawful use of weapon, and (6) identity theft. These sentence lengths were compared with a Wilcoxon rank sum test because the distribution was not normal.

We found that, for (2) possession of a controlled substance, (3) aggravated driving under the influence of alcohol, (4) armed robbery, and (6) identity theft, one must reject the null hypothesis that there is no difference between sentence lengths. Of these crimes with statistically significant differences, only aggravated driving under the influence of alcohol had whites sentenced longer than blacks. The remaining three showed blacks being sentenced longer.

Literature Review

Much previous work has gone into the field of exploring race in criminal justice.

The United States Sentencing Commission wrote a report detailing the differences in sentence lengths throughout the country: they hoped to find out how sentencing changed after the historic Booker v. United States case, in which the Supreme Court ruled that sentence lengths could only be prescribed on the basis of evidence used in the case, and that any proof found afterwards was unable to change sentence severity.

However, “The Commission found that sentences of Black male offenders were longer than those of White male offenders for all periods studied,” and so the trend was there before the Supreme Court ruling (Schmitt, Reedt, and Blackwell 6). The commission accounted for “Prior Violence for Male Offenders” and found that, amongst a category of repeat offenders, blacks were sentenced around 20.7% longer than whites (Schmitt, Reedt, and Blackwell 16).

The study also did investigations on other demographics such as gender, finding that women were likely to be sentenced for less time than men. For example, white females were sentenced to 27.2% less time than white males, in the category having prior violences. Hispanic offenders, however, had similar sentencing lengths to whites up until the most recent time period examined by the commission, from 2011-2016, where they received a 5% longer sentence. Previous years were close to 0%, and the earliest periods from 1998 to 2003 and 2003-2004 had 3.6% and 4.4% less sentence time than whites, respectively.

The paper acknowledges that there could be biases, noting that “employment history or family circumstances” could affect a judge’s decision, but are unable to be accounted for (Schmitt, Reedt, and Blackwell 17). They state, clearly, that “results of the analyses presented in this report should be interpreted with caution and should not be taken to suggest discrimination on the part of judges” and can only signify correlation, not causation.

Other authors have investigated differences in legal outcomes for particular categories of crime tested in this project. For example, in the paper *Race, Gender, and Risk Perceptions of the*

Legal Consequences of Drinking and Driving, Frank A. Sloan, et al. investigate differences in arrest and legal outcomes for DUIs among black and white people. They compiled 2009 arrest records from eight cities in different U.S. states. They found with high confidence that black men had a higher probability of being fined, a higher average fine amount, and a higher probability of imprisonment if convicted of a DUI than white males. However, they also found with high confidence that black males received shorter sentences if imprisoned than white males. This is consistent with our results - as we found that only for the case of DUIs was sentencing length longer for whites than blacks - but the pre-sentencing findings in this study indicate that there are still underlying advantages for white convicts in the prosecution of a DUI.

There has also been research showing the inherent biases in criminal justice datasets. Criminological research shows that government databases are not a complete census of all criminal offenses, and also that the databases do not show a representative sample of all crimes. The research shows that police officers consider race in their evaluation of which neighborhoods to patrol and which individuals to search (Lum, Kristian, and William Isaac). Different races and racial neighborhoods may be policed differently to begin with, leading to inherent biases in the datasets.

In the field of artificial intelligence and criminal justice, Kristen Lum, a current Professor at Penn, works on examining machine learning in the criminal justice system. Her work is on machine-learning based predictive policing, which is used to try to predict crime before it even happens. These algorithms are increasingly used by law systems, and depend upon the datasets that are inherently racially biased - as described above. Kristen's work has found that, as a result, these systems reinforce, and can amplify, historical racial biases (Lum, Kristian, and Isaac).

Algorithms that specifically detect recidivism are also being developed, which has also raised ethical concerns. COMPAS is software used by US courts to evaluate the probability of a defendant in becoming a recidivist, in other words repeating a crime. A 2016 study found that this algorithm incorrectly judged black defendants to be high risk, while incorrectly judging white defendants to be low risk. Black defendants who were not recidivists were twice as likely to be judged as high risk to their white counterparts (45% compared to 23%), while white defendants who did reoffend were labelled as low-risk twice as frequently (48% to 28%) (Larson and Angwin). These algorithms are beginning to be considered by many to be an unethical misuse of machine learning, which highlights the importance of explainability and fairness in machine learning.

Background Knowledge

In order to find any difference in sentence length for crimes committed by various demographics in the Chicago area's courts, we reference the sentencing dataset for Cook County, IL (Sentencing). In the county are 6 court districts: Chicago, Skokie, Rolling Meadows, Maywood, Bridgeview, and Markham, and we will use all of them. The dataset contains 243,006 rows with 41 columns, where each row represents a charge, and an entry is added every time a defendant makes a guilty plea or the court finds a person guilty in a trial. A sentencing hearing is then held to decide on the type and length of punishment, with the most common sentences for felony cases being prison, probation, jail, conditional discharge, supervision, or Cook County Boot Camp. It contains data from 1984-present and is updated yearly, with the most recent update being September 30, 2020. We will be using all of the data from this century, so it spans 20 years of cases (2000-2020). While we initially tried to only take data from 2015 to 2020, we preferred to extend the timeframe as it was not enough data - possibly adding some bias into our analysis. It only contains adult criminal felony cases, so juvenile cases are not included. The sentencing dataset is a part of four separate datasets provided by Cook County courts, which include “Intake,” “Initiation,” “Disposition,” and “Sentencing.”

Here is an example of the sentencing dataset, with unused columns removed:

Sentencing (Case 108890012037)

CASE_ID	DISPOSITION_CHARGED_OFFENSE_TITLE	SENTENCE_TYPE	CURRENT_SENTENCE	CASE_PARTICIPANT_ID	COMMITMENT_TERM	COMMITMENT_UNIT	RACE
108890012037	HOME INVASION	Conversion	True	145298548873	030	Year(s)	Black

Process & Decisions

We were specifically interested in how race plays a factor when facing punishment for a crime in Cook County, so we decided to focus specifically on the relationship between race and initial sentence for six crimes: first degree murder, possession of a controlled substance, aggravated driving under the influence of alcohol, armed robbery, aggravated unlawful use of a weapon, and identity theft. Before any analysis could be done, however, we had to refine our dataset.

First, we control for gender. This is important as gender might have an effect on the test. As we found in our literature review, females tend to be sentenced for shorter periods. As a result, we only consider males in our analysis.

We then control for resentencing, which is when someone who has already been sentenced gains a new sentence for the same crime. This happens when a sentence is appealed, if a set parole is broken, or any other number of situations, and it shows up in the dataset as two separate rows. We were interested in the relationship between race and the severity of punishment in Cook County, not the behavior of those already convicted, so we decided to control for this by taking only the first sentence declared in each case - namely by only taking the rows where the SENTENCE_PHASE flag was set to “Original Sentencing.” This allowed us to get the initial court reactions of each case without muddying the data with post-sentence behavior, such as breaking probation or bargaining for a lighter sentence.

Another confounding factor is the number of charges per case - a single case can have multiple charges against the defendant, each with its own sentence length, where the judge determines whether the offender must serve the charges consecutively (where the total prison time is the length of each charge summed up) or concurrently (where the total prison time is the length of the most severe charge). Unfortunately, this dataset did not provide a way to tell whether the charges were served concurrently or consecutively, so we instead went by primary charge sentence length - the length of the sentence for the most severe crime committed in a given case. We therefore kept only the rows where the boolean PRIMARY_CHARGE_FLAG was true.

The next major confounding factor was recidivism, the tendency of a convicted criminal to reoffend. Whereas in resentencing the same person gets multiple sentences for one crime, in recidivism the same person gets multiple sentences for multiple cases over the span of their life. In the dataset, this shows up as the same CASE_PARTICIPANT_ID being tied to multiple cases over time. These repeat offenders skew the data because courts punish repeated offenses more harshly than initial offenses, so we controlled for this by only taking the earliest offense by date of each CASE_PARTICIPANT_ID. This allowed us to more closely capture the initial reactions of the court when presented with a case, without the interference of each individual's history.

We then split the data into groups based on the six different crimes we were investigating, found in the DISPOSITION_CHARGED_OFFENSE_TITLE column, and used the SENTENCE_TYPE (e.g. prison, conversion), COMMITMENT_TERM (length of sentence), and COMMITMENT_UNIT (units of length of sentence, e.g. years) in order to find the severity of each sentencing. Finally, the RACE column was used to separate each crime into two separate samples, Black and White, which we then used for statistical analysis.

Mathematical Model

To analyze this data we relied on hypothesis testing, a method where one assumes that some null hypothesis is correct and then leverages statistics to refute that assumption. The null hypothesis is by design a claim that there are no differences between two sampled populations, and that any measured difference in the two samples results exclusively from sampling error. In our case, this null hypothesis was that black and white individuals convicted of the six different crimes had to face the same sentence lengths.

We initially intended to rely on a series of two-sample Z tests to analyze the data, but our initial results showed that the data had a skewed distribution, and that some of the sample sizes were very small (n = 29). This complicated our plan, because two-sample Z tests implicitly make an assumption that the data being compared is normally distributed. To sidestep this issue, we pivoted to Wilcoxon rank-sum tests, which does not assume a normal distribution, and traditionally does well with comparing small samples.

On a high level, rank-sum tests compare the medians of data instead of the means, so they are more resistant to skewed data like ours. On a more mathematical level, the rank-sum test works by sorting all of the data and then assigning each data point a rank based on where in the order it falls. The data is then divided back into two the two samples, and the mean rank of each sample is computed. If the null hypothesis were true, and both samples were drawn from identical distributions, then the expected mean rank $r_{expected}$ for each sample would be $r_{expected} = \frac{n_1+n_2+1}{2}$, where n_1 and n_2 are the respective sizes of the samples. Finally, a p-value is generated by comparing the expected and measured averages. Should the p-value be below a predetermined acceptance threshold α, the null hypothesis would traditionally be rejected.

However, because we tested multiple hypotheses, we could not simply accept this at face value; instead, we applied the Bonferroni correction before rejecting the null hypothesis as a way to control for running multiple tests. To do this, we simply divided the α value by the number of hypotheses being tested, and only rejected the null hypothesis in cases where the p-value was below that reduced value: $p \leq \alpha/R$. We took $\alpha = 0.05$ and ran six tests, one for each of crimes in question, so our final acceptance threshold was $p \leq 0.05/6 = 0.0083$.

Assumptions

For us to apply the Wilcoxon rank-sum test, there is one assumption that needs to be true: the two samples have to be independent of each other. We can assume that this is satisfied, as the racial groups (black and white) are completely separate from each other.

The major benefit of running a Wilcoxon test instead of a z-test are how few assumptions are made in the process of the test; the variances between the two groups are not required to be the same, the sample sizes are also not required to be the same, and a normal distribution is not required. This lack of assumptions make the Wilcoxon rank sum test easier to use.

We will also separately be making a closed-world assumption that this dataset contains all the people that were sentenced in this time-frame. Since this is a government dataset showing all court cases, this is appropriate to assume. However, this assumption does not take into account the previously discussed inherent bias in criminal justice datasets towards the different races - the fact that if police are considering race in which neighborhoods to patrol, the data that has been collected by the police is already biased. There is nothing we could have done about this, but it is further discussed in the Confounding Factors section.

Table 1: Results of Rank Sum Tests Comparing Sentence Lengths by Race in Each Crime Severity Group

Crime	Test Statistic	P-Value	Conclusion
First Degree Murder	-1.09	0.278	Fail To Reject Null Hypothesis
Possession of a Controlled Substance	-14.48	$1.71 * 10^{-47}$	Reject Null Hypothesis
Aggravated Driving Under the Influence of Alcohol	12.98	$1.54 * 10^{-38}$	Reject Null Hypothesis
Armed Robbery	-4.65	$3.35 * 10^{-6}$	Reject Null Hypothesis
Aggravated Unlawful Use of Weaponry	-0.0755	0.940	Fail To Reject Null Hypothesis
Identity Theft	-2.85	0.00439	Reject Null Hypothesis

Figure 1: Plot of Sentence Lengths by Race for First Degree Murder:

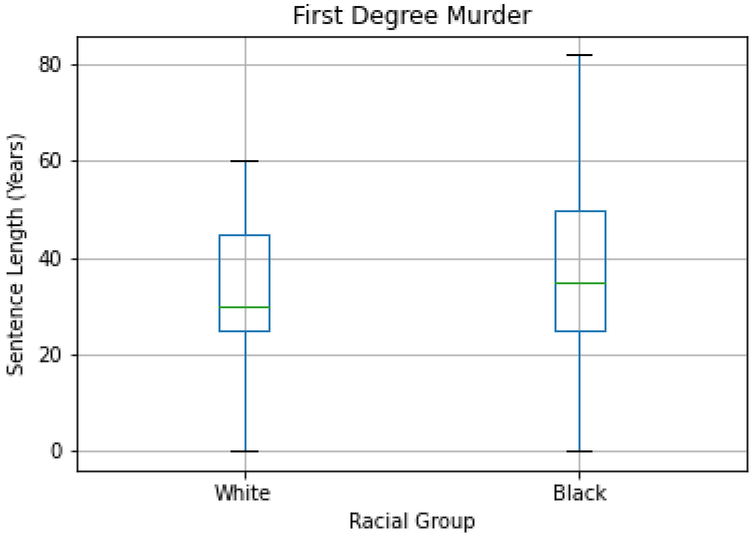


Figure 2: Plot of Sentence Lengths by Race for Possession of a Controlled Substance:

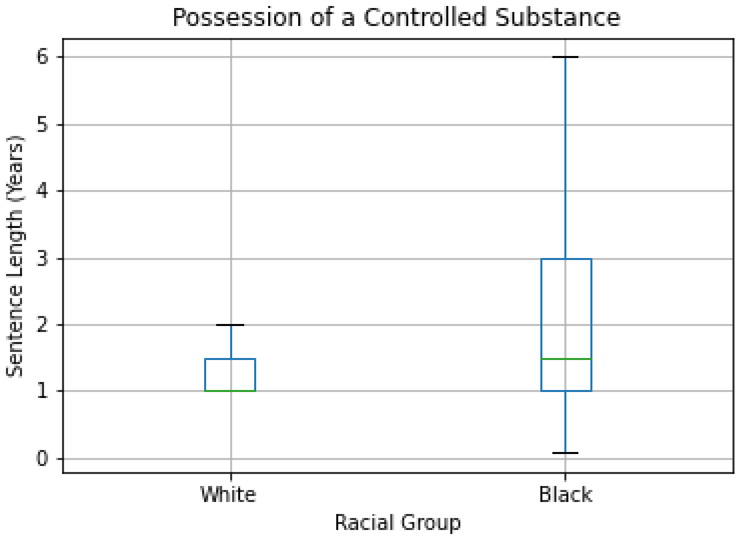


Figure 3: Plot of Sentence Lengths by Race for Aggravated Driving Under the Influence of Alcohol:

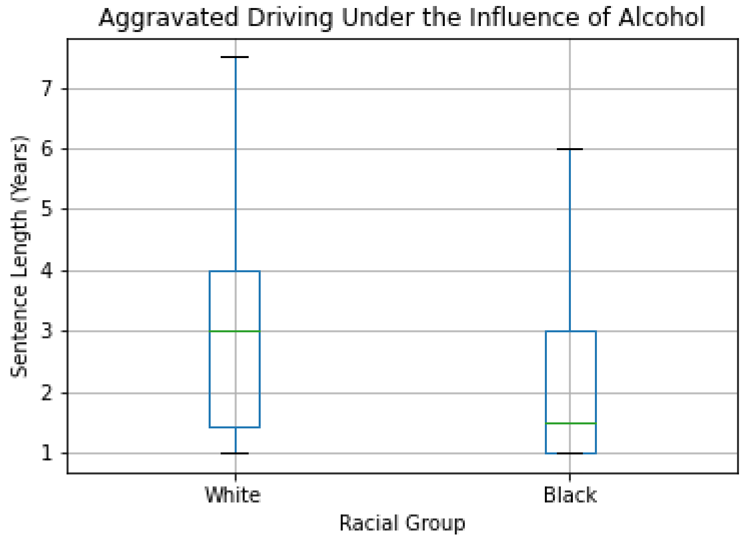


Figure 4: Plot of Sentence Lengths by Race for Armed Robbery:

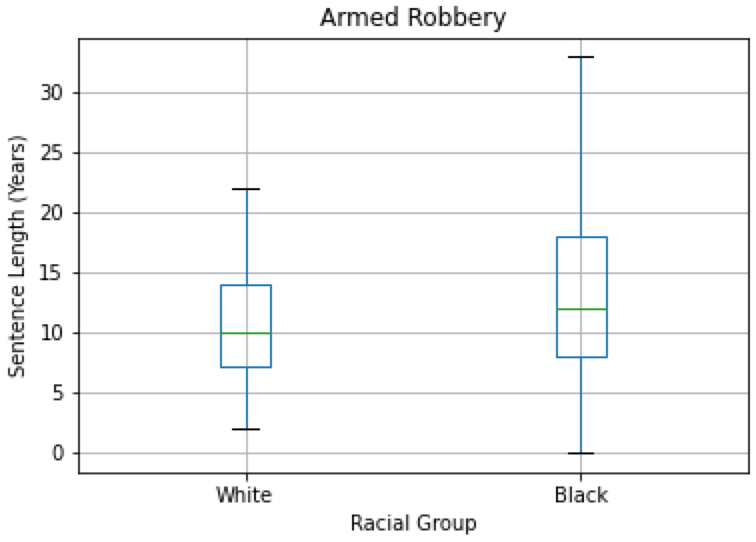


Figure 5: Plot of Sentence Lengths by Race for Aggravated Unlawful Use of Weaponry:

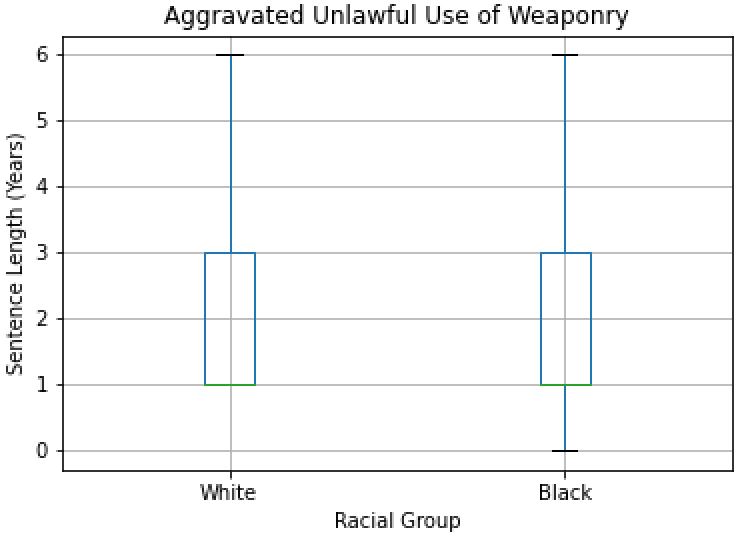
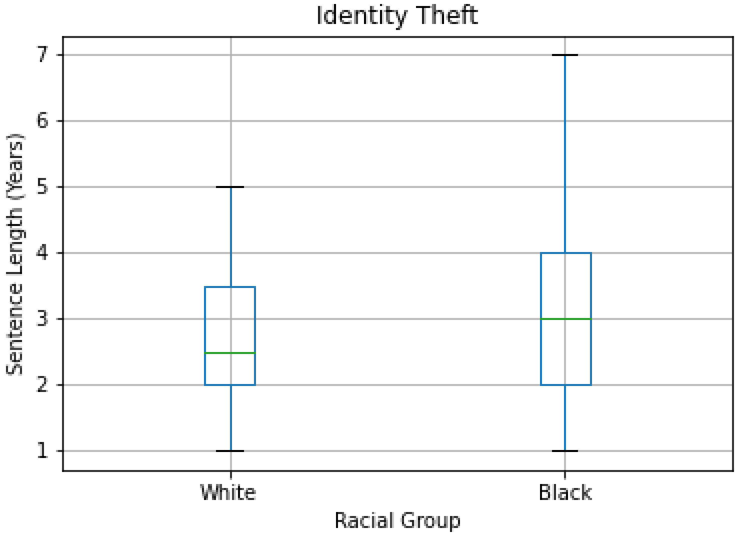


Figure 6: Plot of Sentence Lengths by Race for Identity Theft:



Results

The results of running this series of Wilcoxon Rank-Sum tests are shown in Table 1; of the six tests we ran, the p-values were small enough that we rejected the null hypothesis for four of the crimes and failed to reject the null hypothesis for the other two.

Specifically, the distributions of sentence lengths for black possession of a controlled substance, aggravated driving under the influence of alcohol, armed robbery, and identity theft are not equivalent to their white counterparts. These are alarming results, especially given that an ideal justice system would consistently follow the null hypothesis - that the distributions for various crimes would be approximately the same regardless of ethnicity.

Figures 1-6 provide additional information about the distributions of sentence lengths for each race and crime. In general, the boxplots look similar: they are skewed towards short sentences, with multiple outlying long sentences, though we removed outliers from these graphs to make the bulk of the data more easily interpretable. In fact, it was this non-normality that motivated our use of the rank-sum test. One notable observation is that for every crime, there were more black convicts with outlying long sentences than white convicts, and the maximum sentence was generally significantly longer for black convicts. However, this is most likely because there were many more data points for black convicts than white convicts, and the discrepancy is not necessarily true in general.

Of the four crime categories where the null hypothesis was rejected, three were found to be sentenced more harshly for black convicts than white convicts: possession of a controlled substance, armed robbery, and identity theft. Only one was found to be sentenced more harshly for white convicts, aggravated driving under the influence of alcohol. The literature review backs up our findings on this about DUI sentences being longer for whites than for blacks.

Confounding Factors

Certain biases and artifacts of data collection could affect our conclusion. One assumption we make is that all instances of the same crime have the same severity. If different groups tend to commit different forms of the same crime, this could skew the sentencing results. However, this is likely not an issue as the crime groups that were selected had very specific labels. The severities are likely to be equal.

Another factor is that areas with different demographics could have different rates of policing, and thus different rates of criminals being caught, which could be a factor in average sentence length. We know that this is likely to be true based on our literature review: it has been shown that police take race into account when considering which neighborhoods to patrol. This may mean that crimes for racial minorities are caught more frequently, and crimes for white demographics may go amiss, contributing to systematic error in sampling. These errors in sampling could not be addressed in our project, but are an interesting avenue for further research.

Another factor to consider is differences in sample size between racial groups within each crime. If certain racial groups have significantly less data within each crime, this could make results less certain - however, this is not a serious problem, as differences in sample size only affect the power of the test, not the validity, for the Wilcoxon rank sum test.

Additionally, while we believe we made the right choice by considering specific crimes for our test and thereby avoiding aggregating groups, this opened up two potential problems: (1) less data, and (2) potential bias by introducing data from earlier groups. The first problem, again, is likely not too significant an issue, as the rank sum test does well with small sample sizes. The second, however, could be a confounding factor in our results. Time series analysis accounts for the fact that data taken over a period of time may have internal structure, such as autocorrelation, seasonal varieties, and susceptibility to current events. The world has changed a lot over the past 20 years, so assuming that we can simply aggregate all of this data may not be a valid assumption. For example, the Supreme Court may have done landmark cases that significantly changed the sentence lengths within those years. However, none of us had familiarity with time-series data and it may have been outside the scope of this class. But a future analysis of the same dataset with regards to analysis over time could prove interesting results.

One possible confounding factor is that some rows were dropped as a result of data entry errors. However, this is unlikely to be a very significant problem as the rows dropped were insignificant with regards to the overall size of the dataset. We were hoping to correct this by speaking to the dataset owners on these rows to understand what these rows should have contained, rather than just dropping them, but unfortunately we never got a response.

Another important confounding factor might be that certain races could be poorer or less educated on legal procedures. If this is true, it might mean that even if the sentence lengths are significantly different, this might not be due to a judge sentencing them for a longer period due to racial bias. It could mean that they approached the legal proceedings in an unideal way due to lack of knowledge or finances. This is certainly a valid possibility.

Another possible important consideration is that further analysis could go into what else happens before sentencing. Not every case is sentenced, and by ignoring unsentenced cases, we are neglecting a potentially important subset of the data. In this study, we wanted to specifically focus on sentencing data, but there are a lot of other considerations that must be made about how racial bias could occur, including the whole process from arrest to sentencing - as well as a broad range of domains in which racial bias could creep in before arrest. This domain is large and requires much further work.

Conclusion

The results of this analysis should be interpreted with caution. While they show that the sentencing lengths are different for four out of the six crimes we tested, with three out of four of them being skewed towards blacks being sentenced for longer, the tests we performed cannot fundamentally explain the reasons why such differences are observed. We have shown that there is a relationship between race and sentence length, but we cannot, for example, state that the reason for this is necessarily racism. Notably, though, the literature review did show that blacks were sentenced more harshly than whites (Schmitt, Reedt, and Blackwell 16), which we have found to be resoundingly true in our own analysis.

The results, however, are interesting and do request further research. There are many areas of exploration for the future. It could be interesting to approach this problem from a spatial lens, possibly running this same analysis for all the counties throughout the US. Global and local spatial autocorrelation statistics could be run to assess clustering in the US, to possibly analyze which areas may have higher racial bias in criminal justice.

Certain judges or courts could be Cook County could be more racially biased than others. We could get the sentencing lengths for different races for one crime from different judges, and likewise compare them with a rank sum test using the Bonferroni correction. We could extend this analysis to check if the race of the judge plays a role.

We could check if education and finance levels impact sentencing length, regardless of race, and also within each race. It might be difficult to get this information though, as we would require income and education levels for each person on the sentencing dataset, which might require personally identifiable information.

A time series analysis could be done in order to show how events over time such as social movements might have impacted sentencing length within racial groups. Various racially interesting moments in time over the past 20 years could be found, and the sentence lengths near these moments could be checked.

Also, further research could go into the different rates of policing in different areas, and how this impacts criminal justice datasets, in order to have a better understanding of what conclusions can be made from these datasets and just how large the error and uncertainty actually is.

References

"Sentencing." *Cook County Data Catalog*, Cook County State's Attorney Office, February 13, 2018, (https://datacatalog.cookcountyil.gov/Courts/Sentencing/tg8v-tm6u).

Frank A Sloan, Lindsey M Chepke, Dontrell V Davis. "Race, Gender, and Risk Perceptions of the Legal Consequences of Drinking and Driving." *PubMed,* Jun 2013, https://pubmed.ncbi.nlm.nih.gov/23708483/

Jeff Larson, Julia Angwin. "How We Analyzed the COMPAS Recidivism Algorithm." *ProPublica*, 23 May 2016, www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

United States, Supreme Court, United States Sentencing Commission. Demographic Differences in Sentencing: An Update to the 2012 Booker Report. United States Sentencing Commission, 14 Nov. 2017, https://www.ussc.gov/sites/default/files/pdf/research-and-publications/research-publications/2017/20171114_Demographics.pdf

Lum, Kristian, and William Isaac. "To Predict and Serve?" *Royal Statistical Society*, John Wiley & Sons, Ltd, 7 Oct. 2016, rss.onlinelibrary.wiley.com/doi/full/10.1111/j.1740-9713.2016.00960.x

Appendix

Glossary of terms used in the sentencing dataset:
https://www.cookcountystatesattorney.org/sites/default/files/files/documents/column_by_dataset_glossary_final_1.pdf

See Our Code:

https://mit.cs.uchicago.edu/chz/cmsc21800project