Exploring Sentence Length Bias in Cook County, IL Sentencing Data

Adin McAuliffe[a,1], **Jordan Miller**[a,1], **Samson O'Donell**[a,1], **and Sam Winchester**[a,1]

[a]Dartmouth College

This manuscript was compiled on March 14, 2023

In this report, we analyze Cook County, IL, sentencing data and search across defendant data, including age, gender, race, plea type, and judge experience, to investigate disparities in sentence lengths handed down by Illinois judges. In the United States Judicial System, judges, not the jury or prosecutor, have ultimate discretion when deciding what period of time a convicted defendant will spend in prison if any. Judges often follow sentencing guidelines that offer guidance to judges on punishing particular offenses. Aside from following mandatory minimum sentencing rules, it is up to the judge whether or not he or she adheres to sentencing guidelines, making sentence lengths potentially vulnerable to implicit or explicit biases. We search for disparities in sentence length severity to investigate this vulnerability and help inform the conversation around equal justice initiatives. In Cook County, our analysis finds that while plea type does appear to influence sentence severity, age, gender, race, and judicial experience do not significantly influence sentence severity. Furthermore, there does not appear to be a significant causal relationship between age, gender, race, and judicial experience and sentence severity, and these variables also appear to have poor predictive ability for sentence severity.

1. Introduction

The United States criminal justice system is unique in that our country's democratic values shape it. In the same way that American citizens vote to elect political leaders, jury trials rely on a jury composed of laypeople to vote to convict a defendant of an accused crime. After a suspect is arrested and charged with a crime, he or she becomes a defendant and enters into a plea of guilty or not guilty. If the defendant pleads guilty, he or she waives the right to a trial by peers and agrees to accept the sentence handed down by the sentencing judge. If the defendant pleads not guilty, he or she often participates in a jury trial, in which case the jury will find a verdict of guilty or not guilty. If the verdict is guilty, the judge similarly sentences the defendant to a term of imprisonment loosely based on accompanying sentencing guidelines and other factors.

We were motivated to examine sentence length disparities across demographics because of the discretion judges have in passing judgments. The presence of discretion led us to believe we might see biases within sentence lengths because behaviors differ between judges and between the same judges over time. We believe our report's investigation is important when considering criminal justice reform and, more specifically, when considering reelecting judges. In 2018 Cook County Circuit Judge Maura Slattery Boyle sought retention, meaning she ran for a subsequent term as judge unopposed by another candidate. According to Injustice Watch (1), "thirty-four times in the past six years, the Illinois Appellate Court has upended decisions by Cook County Circuit Judge Maura Slattery Boyle based on her errors, overturning her decisions at a pace far higher than that of other judges." It additionally reports that "Slattery Boyle [...] issues the most severe sentences compared to the 23 other Criminal Division judges who have presided over 1000 or more cases in the past six years." Inspired by this case and its window into the justice system, we intend to explore sentence length trends as related to the following variables:

- 1. Defendant Plea Type
- 2. Defendant Gender
- 3. Defendant Race
- 4. Defendant Age
- 5. Judge Experience

All authors contributed equally

[1] To whom correspondence should be addressed. E-mail: adin.r.mcauliffe.24@dartmouth.edu, jordan.a.miller.25@dartmouth.edu, samson.s.odonnell.24@dartmouth.edu, samuel.c.winchester.24@dartmouth.edu

Using cleaned Cook County Sentencing Data, we will create a uniform sentence length column for each convicted defendant and assign them a sentence severity score. We will then use the demographic data in the data set to conduct an inferential, causal, and predictive analysis of those five factors and the sentence severity score. To begin the analysis, we review the literature on criminal justice reforms, including studies on federal sentencing guidelines, interjudge sentence disparities, and sentence disparities across races. Inferential analyses will allow us to draw general conclusions on relationships between severity scores and various demographic variables. We will examine correlation coefficients and statistical significance across variables, identifying various strengths of relationships between variables. Our causal analysis will show us if any variables share a statistically significant causal relationship with sentence severity. Causality between a given variable and sentence severity would imply a cause-and-effect relationship, which we believe is very important when considering sentence reform policies. Similarly, the lack of a cause-and-effect relationship may point researchers and policymakers in other directions to identify potential unfair sentencing practices. Finally, our predictive analysis will allow us to test the hypothesis we reach from the causal and inferential analysis. In the context of our report, we are interested in which factors are most strongly associated with an increase in sentence severity.

2. Related Work

The literature surrounding sentence length disparity carries immense weight in the discussion of criminal justice reform. At both a state and federal level, sentence length disparities help inform the impact of preexisting sentencing guidelines or the need for updated ones. A study published in the Journal of Economics (2) titled “Measuring Interjudge Sentencing Disparity: Before and After the Federal Sentencing Guidelines” is concerned with the congressional objective, as evident from legislative history such as the 1984 sentencing reform act, to address the issue of sentencing disparities that may result from judicial discretion. Anderson et al. find that "the behavior of judges appears to be fairly consistent over time prior to the [implementation of federal sentencing] guidelines.” This report informs our research because it suggests an inherent consistency across sentencing over time. This suggestion may help us identify judges or specific defendant demographics that deviate from this baseline assumption. The Anderson et al. report also encourages further research into Cook County’s sentencing reforms to identify any particular laws or policies that may be more effectively correlated to sentence consistency than broad guideline implementations.

Another study titled “American Indians and sentencing disparity: An Arizona test” (3) uses regression analyses to predict the sentences received by American Indian and White defendants convicted of six crimes, including homicide, sexual assault, robbery, assault, burglary, and larceny. In their report, Alvarez et al. find that robbery and burglary were the only crimes in which American Indians received longer sentences than Caucasians, and indeed white defendants received significantly longer sentences than American Indians for cases of homicide. This study influenced our choice of race as a critical demographic variable to consider in our analysis. Our analysis extends the work of these studies by analyzing sentences with a severity score rather than just comparing the sentence lengths. It also extends the work of these studies by modeling a range of demographic descriptive variables at once rather. Other models that only assess one of these variables are less realistic because judges make sentencing decisions with more than just one binary variable in mind.

3. Data

The public data for our study comes from the Cook County Government and the corresponding codebook is available online at the Cook County government’s website. Our analysis focuses on the sentencing data within the codebook. We retrieve a cleaned version of the public sentencing dataset from this course’s materials that includes the following additions to the original dataset: **simplified_offense_derived** column that simplifies the multiple offense categories into broader charges, **sentenceymd_derived** column that is a version of **SENTENCING_DATE** converted to date-time format, a **judgeid_derived** column that assigns a unique number to each named judge, and **is_(RACE)_derived** columns that assign boolean values of TRUE or FALSE to the respective columns for each defendant, depending on his or her race. These added variables allowed us an easy way to test our hypothesis about which descriptive variables might impact sentence severity. For example, we could use **is_male_derived** to investigate gender, **is_(race)_defined** to investigate age, **age_derived** to investigate age, and **CHARGE_DISPOSIITION** to investigate plea type. We chose this data for those descriptive variables, as well as for the fact that the data contains three variables (**COMMITMENT_TYPE**, **COMMITMENT_TERM**, and **COMMITMENT_UNIT**) that can be used to calculate each defendant’s sentence length. By calculating that sentence length, we knew we could then perform a basic statistical analysis to estimate each defendant’s sentence severity. A brief overview of the cleaned dataset we imported is below, in Table 1.

The full **sentencing_cleaned** dataset contains 135,165 entries. The unit of analysis is **CASE_ID**, because each row of the dataset corresponds to a unique court case in Cook County. The time window which the

Table 1. Characterizing Analysis Sample

Variable	Value
Number of Observations	135,165
Range of Dates	2012-01-01 to 2021-03-31
Range of Age	17 to 81 (Average is 33.38)
Number of Unique Charge Types	1,163

dataset covers spans from 2012-2021. We used this entire time window in our analysis. The unit of analysis we used was a column that denotes sentence severity (**sentence__length__zscore**), which we created using three existing variables in the dataset: **COMMITMENT__TERM**, **COMMITMENT__UNIT**, and **DISPOSITION__CHARGED__OFFENSE__TITLE**. For this analysis, we subsetted the data just to defendants who either pled guilty or not-guilty and later received a prison sentence. In other words, we do not use the data for defendants who 1) plead not-guilty and are subsequently exonerated, 2) were deemed mentally ill, 3) whose cases were dismissed, or 4) who pled guilty and received punishments such as probation, diversion programs, or community service. Even after fully cleaning and subsetting our data, we had well over 60,000 remaining unique court cases to investigate, so we are not concerned about our sample size. We did not merge this data with any other dataset(s) during our analysis, as the initial **sentencing__cleaned** data had enough information to model all of our hypotheses.

For the purposes of this study, the primary limitation of the data is that there is no indicator of whether defendants are repeat offenders. Having data on defendants' past offenses could explain harsher sentence lengths, especially when sentencing guidelines differ for repeat offenders. A further limitation is the errors in the dataset from improper input. For example, the **COMMITMENT__UNIT** column contains values like "pounds" that do not correspond with known criminal sentences. For the following analysis, **COMMITMENT__UNIT** values that did not correspond with an obvious unit of time were omitted to avoid clouding the data. Overall, however, because we were able to start with a version of the data that had already been cleaned, there were not many rows with missing data.

4. Methods

A. Data Investigation. The initial data investigation was completed in the **00__inspecting__data.ipynb** Jupyter Notebook. In looking at the data's attributes, we noticed that some variables in the data were cast as improper data types. For example, **CASE__ID** was cast as a numeric object, while **sentenceymd__derived** was cast as a string object. Before we could continue investigating the data with summary statistics, we fixed these type errors in the columns relevant to our investigation, changing **CASE__ID** and **CASE__PARTICIPANT__ID** to string objects and **sentenceymd__derived** to a datetime object. Next, we created a simplified summary table that included all of the variables we intended to use in our analysis. This simplified summary table contains characteristic information about the sample (in this case, the cleaned sentencing dataset), including the number of observations, the period of time the data entries in the sample span, the range in ages of defendants in the sample, and the number of unique judges in the sample.

B. Data Cleaning. The data cleaning process was completed in the **00__inspecting__data.ipynb** Notebook. In the data cleaning phase, we cleaned the dataset so that its variables could be modeled against sentence severity. This data cleaning process included importing a function that calculates z-score, creating a function that removes outliers, subsetting the data to just include defendants who plead innocent or guilty and received prison sentences in either the Illinois or Cook County Department of Corrections, creating variables for sentence length and for sentence length severity, and finally removing rows that contain outlying severe sentences from the data. We also subsetted the data to only include rows in which the charge against the defendant represented the most severe charge the defendant faced. We began this phase by importing the original **sentencing__data__cleaned** dataset, and then subsetting the data as outlined above. This subset removed 65000 rows from the dataset, leaving us with 69346 observations before we created the sentence length and sentence severity variables.

B.1. Cleaning Functions. We used two functions to clean our data, both defined at the top of the data cleaning script. The first of these functions was zscore(), which we imported from the scipy.stats python library. This zscore() function calculates a zscore() for each value it is passed. Z-scores represent the number of standard deviations an observation is from the sample mean. The second of these functions was **remove__outliers**, which takes in a dataframe and a column name that references the column from which to remove outliers, and outputs a dataframe from which outliers have been removed. To actually remove the outlying data, the function uses the basic IQR * 1.5 method, computing the IQR of the column passed and then removing all rows from the dataframe for which the column is either less than (Q1 - 1.5) * IQR or greater than (Q3 + 1.5) * IQR (In which Q1 and Q3 represent the first and third quartiles of the column values). This method of removing outliers is very common and works well for our relatively simple dataset.

B.2. Creating Sentence Length Variable. Before we could create a variable to represent the severity of each sentence in the dataset, we needed to create a standardized sentence length variable because the **sentencing_data_cleaned** dataset stored sentence length in three different columns: **COMMITMENT_TYPE** (what kind of sentence, **COMMITMENT_UNIT** (unit for sentence term), and **COMMITMENT_TERM** (number that when combined with unit gives a sentence length). To create a universal sentence length variable, we used **COMMITMENT_UNIT** and **COMMITMENT_TERM** to standardize all of the **COMMITMENT_TERM**values into a number in years. First, we converted **COMMITMENT_TERM** to a numeric, because the original dataset stored the column as string objects. We then create a new column, called **sentence_length**, that standardizes **COMMITMENT_UNIT** into years with a lambda function applied to each value in the column. This lambda function uses a series of if-else statements to divide **COMMITMENT_TERM** by 12 if **COMMITMENT_UNIT** is equal to "Months", divide **COMMITMENT_TERM** by 365 if **COMMITMENT_UNIT** is equal to "Days", and so on and so forth for each of the unique C**COMMITMENT_UNIT** values. For this calculation, we make the assumption that 1 life sentence is equal to 100 years. We used a lambda function rather than writing a generalizable function in this case because the operation needed was too specific to warrant a defined function.

B.3. Creating Sentence Severity Score Variable. Now that we had created a standard sentence length column (**sentence_length**), we could use the **zscore()** function to create a standardized sentence severity column. To do so, we began by grouping the data by **DISPOSITION_CHARGED_OFFENSE_TITLE**, a column in the dataset that indicates the unique charge a defendant faces at the disposition phase. The disposition phase is the last phase in criminal proceedings before defendants are sentenced. Grouping this variable grouped the data by identical charge type, such that groupings only contained rows in which the defendants faced identical charges at the latest phase in the criminal proceedings. Having grouped rows into each individual charge type, we created the severity score column **sentence_length_zscore** by applying the **zscore()** function to each value in the new **sentence_length** column, with the groupings intact. Thus, each value in the **sentence_length_zscore** column represents the number of standard deviations that the defendant's sentence was from the mean sentence length (**sentence_length**) for the exact charge type that he or she faced. By giving this distance in terms of standard deviation, these severity scores can be compared across the entire dataset, no matter what defendants have been charged with. Positive values for **sentence_length_zscore** indicate that a **sentence_length** is more severe than the average, while negative values indicate that a **sentence_length** is less severe than the average.

B.4. Removing Outliers. Finally, before we could proceed with our analysis, we needed to remove outliers from the dataset. Z-scores are sensitive to extreme values for the same reason that an average is sensitive to extreme values. In the context of sentence severity, z-scores are especially sensitive and outliers are particularly problematic, because sentence lengths can skew extremely high but they cannot skew extremely low (a judge can sentence you to 400 years for a crime with normal sentencing of 10 years, but they cannot sentence -390 years). Thus, in the dataset, the magnitudes of the most severe **sentence_length_zscore** are much larger than the magnitudes of the least severe **sentence_length_zscore**. In our sample, the maximum severity score was 98, while the minimum was -3.48. Without removing these outliers such that these magnitudes are closer together, it would be very difficult to normalize the data accurately. To fix this problem, we applied the **remove_outliers()** function to the s**sentence_length_zscore** column, removing all outlying values with the IQR * 1.5 method. After applying this function to the column, we quickly checked the maximum and minimum values to make sure the magnitudes were closer together. This check showed that the method worked, with the new maximum **sentence_length_zscore** equalling 1.54 and the new minimum **sentence_length_zscore** equalling -1.74. In the end, applying the **removing_outliers()** function removed about 4000 rows from the dataset, leaving us with 65857 rows at the end of the cleaning process. In total, the cleaning process removed 69308 observations from the sample, and created two new columns: sentence$_lengthand$**sentence_length_zscore**.

B.5. Identifying and Creating Modeling Variables. Following the cleaning of the dataset, we identified five descriptive variables we wanted to analyze in relation to sentence severity. Some of the values needed to explore these variables already existed in the dataset, and some we needed to create. All of these variables were to be modeled as independent variables against the dependent variable sentence severity (**sentence_length_zscore**). The identification and creation of these variables and columns was completed in the **02_creating_regression_variables.ipynb** Jupyter Notebook. The variables we planned to explore are as follows, in Table 2:

Table 2. Descriptor: Column(s) in Cleaned Data

Defendant Plea Type	is_guilty_plea*, is_innocent_plea*
Defendant Age	age_derived
Defendant Gender	is_male_derived, is_female_derived
Defendant Race	is_black_derived, is_white_derived, is_hisp_derived_, is_other_derived
Judge's Experience	nth_case*

As shown above in Table 2, some of the descriptors we wanted to explore already existed in the dataset, and some we needed to create. First, we recast all of the values in the existing **is_(descriptor)_derived** columns from boolean objects to binary numeric objects, so that they could be input into a regression model. Using the existing variable **CHARGE_DISPOSITION** which stores a defendant's plea type, we then created the binary variables **is_guilty_plea** and **is_innocent_plea**to take the value of 1 (indicating "yes") and 0 (indicating "no"). To create these variables, we used the built-in **numpy.where()** function. Next, we used a similar methodology to create the binary variable **is_female_derived**, complimentary to the existing **is_female_derived**, that takes the value of 1 when the column GENDER is equal to "female" and 0 when the column **GENDER** is not equal to "female".

Finally, we needed to create a variable to represent judicial experience. Because of how the original dataset uses unique cases as the unit of analysis, this variable needed to exist at the case level. As such, we created a variable for each case that indicates how many cases the judge assigned to that row's case had been assigned to before in the dataset. Obviously, this variable is not perfect, because we do not have data on judges' experience prior to when the data began being collected. However, for the purposes of our regression, the variable works, because we only need to be able to estimate how much experience the judges have relative to the other judges in the dataset at that time. To create this Nth case variable, we first recast **sentenceymd_derived** to a datetime object, because for some reason the .csv file format will not store this column properly despite fixing this issue in an earlier script. Next, sorted the rows in the dataset by date (**sentenceymd_derived**) and judge (**SENTENCE_JUDGE**). This sorting arranged the dataset so that each judge's case(s) were stacked together, in chronological order. We then grouped the dataset by **SENTENCE_JUDGE** and used the pandas function **cumcount()** to add indices to each object in the grouping. Finally, we added 1 to each of these indices so that each index would accurately represent a judge's nth case. We assigned these values to the new column **nth_case**. Now, all of the variables we needed, as outlined in Figure 2, had been created, and the dataset was ready for modeling and visualizing.

C. Research Design.

C.1. Independent Variables. As explained above, before conducting any analysis, we identified five descriptive variables that we hypothesized could influence the severity of the sentence a defendant receives. In the data cleaning process, we created columns to be used as indicators for all of these descriptive variables, so that the dataset is organized in a way that allows us to conduct all of our analysis. In the following sections of our research design, we will outline how we conducted an inferential and causal analysis of these five descriptive variables to create a hypothesis on how these descriptive variables impact sentence severity, as well as how we will use a predictive analysis to test our hypothesis. For our inferential analysis, we used visualizations to look for trends between the descriptors and sentence severity. For our causal analysis and predictive, we used regression models because **sentence_length_zscore**, the variable we want to 'predict', is a continuous variable (making our causal analysis a regression problem and not a classification problem). These five descriptive variables are:

- 1. Defendant Plea Type
- 2. Defendant Gender
- 3. Defendant Race
- 4. Defendant Age
- 5. Judicial Experience

C.2. Visualizations. To conduct an inferential analysis, we created visualizations for each of the descriptive variables identified above. Each of these visualizations were created in the **03_visualizations** R Script, and use normalization to plot the variables against sentence severity (**sentence_length_zscore**). To visualize the impact of plea type, race, and gender on sentence severity, we used R's ggplot package to plot the mean sentence severity in a barplot. Each bar in the plot represented a different category of the descriptive variable. To visualize the impact of age and judicial experience (using **nth_case**) on sentence severity, we used R's ggplot package, alongside the ggplot function **geom_smooth()**, to create a smoothed curve showing the impact of these continuous variables on sentence severity.

The visualizations of the binary variables were normalized using means, and the visualizations of the continuous variables are normalized by creating a smoothed scatter plot and plotting a line of best fit. We then exported all of these plots, to be used for visual analysis.

C.3. OLS Model. To begin our causal analysis, we created a simple Ordinary Least Squares model, in the **04_ols_model.ipynb** Jupyter Notebook. This model looked for linear relationships between any of the independent variables from above and sentence severity (**sentence_length_zscore**). The model was created using the **OLS()**function from the statsmodels package in python. First, we removed all rows from the dataset with NA values in the independent variables columns, then fit a model using the independent variable columns in our dataset as our x data and the severity score column as our y data. The coefficients created by the OLS model each indicate the linear relationship between that specific independent variable and sentence severity, and can be interpreted as a slope. The p value statistics is used to indicate statistical significance ($p < .05$ = statistically significant).

C.4. Decision Tree Model. To test our hypothesis with predictive analysis, we created a Decision Tree Regression model, using sklearn's **DecisionTreeRegressor()** function, to further assess if our selected independent variables are actually good predictors of sentence severity. This model can be found in the **05_Decision_Tree_model.ipynb** Jupyter Notebook. The goal with this model (and the following Random Forest Model) was simply to see how accurately we can predict sentence severity with the independent variables we identified above. A Decision Tree Regression model uses a tree structure to solve complex regression problems, and is used in our predictive analysis to look for non-linear relationships between the independent variables and sentence severity. We used the same x and y values as in the OLS model above, and separated these x and y values into a simple 80/20 train test split using the **train_test_split()** function from sklearn (80 percent training, 20 percent test data is the most common train test split). We then create a Decision Tree Regressor with a random state of 10 and a max depth of 10 (to hold constant across the entire model), and fit it to the training data. We then use the sklearn **predict()** function on the model with our x test data, and compare these predictions with the y test data by calculating the mean squared error, mean absolute error, and r-squared values. These values are used to determine how accurate the model's predictions were. More specifically, an r-squared value close to 1 would indicate an accurate model, while an r-squared value close to 0 or even below 0 would indicate an inaccurate model. We also use sklearn's function **feature_importances_** to assess which independent variables the model is relying on the most. Larger feature importances indicate that the model is relying heavily on that independent variable, while feature importances closer to zero indicate that the model is not reliant on the independent variable. These feature importances also indicate the quality of the model's predictions.

C.5. Random Forest Model. To further test our hypothesis with predictive analysis with a model that could be more accurate than the decision tree model, we created a Random Forest Regression model using sklearn's **RandomForestRegressor()** function. This model can be found in the **06_Random_Forest_model.ipynb** Jupyter Notebook. Random Forest regression models function exactly the same as Decision Tree regression models, except that a random forest model combines many decision trees into one model. This distinction means random forest models can be more accurate, but also more prone to overfitting. To create this model, we use the exact same methodology as above, except that instead of creating a decision tree regressor before we fit the model, we create a random forest regressor with a number of trees the model tries to combine equal to 100 (with the same random state and max depth). We assess the results of this Random Forest model using the same accuracy metrics, as well as look at the same feature importances. As in the Decision Tree model, we use r-squared to determine the model's accuracy.

5. Results

Results:

A. Visualizations. In our initial analysis of our selected features, we decided to create a series of simple visualizations modeling mean z-scores as a function of each factor. The subsequent figures show our analysis of the data based on plea decision, gender, race, age, and judge experience. We then use a series of models to determine the significance of each factor, following our visualization section.

A.1. Plea Decision. Plea Decision:

Our inferential analysis suggests that plea decisions have the most substantial impact on sentencing severity outcomes. Because we intuitively expect to see this difference in leniency, this finding serves as good evidence that our approach to tracking sentence severity between crimes is effective. This finding in Cook county is consistent with data across the United States. Research from InjusticeWatch, reports similar findings in their own study, concluding

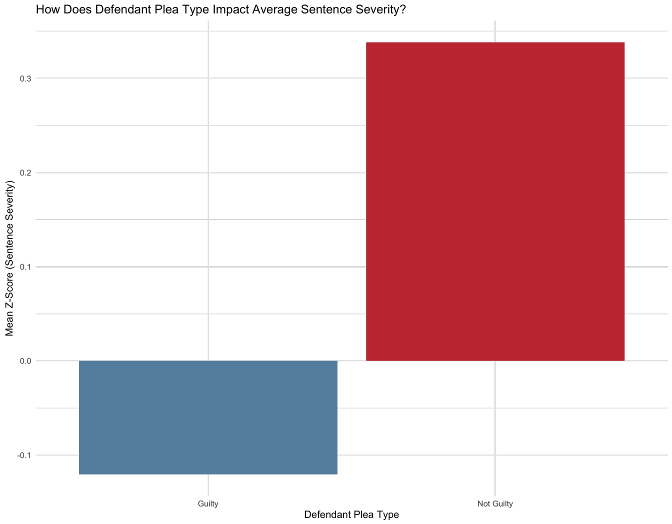
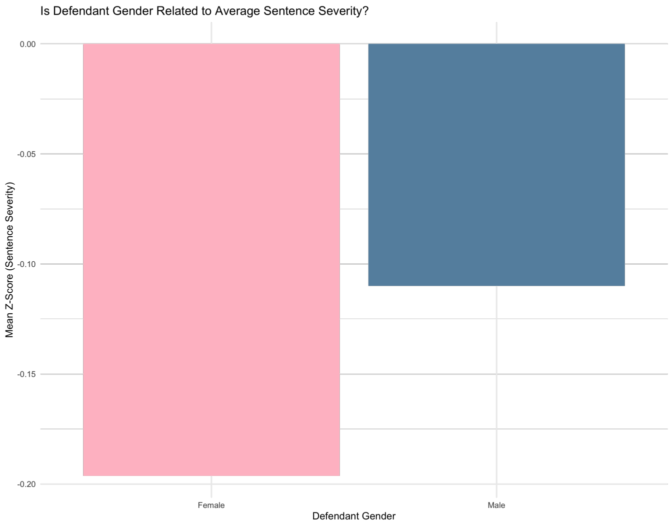


Fig. 1. Defendant Plea

that defendants who plead not guilty and were later convicted of unlawful possession of drugs were sentenced, on average, seven times as long as those who plead guilty. It is important to note the substantial difference in sentencing outcomes between those who plead guilty vs. not guilty is not entirely quantified in our data. In many situations, the prosecution will pursue lesser charges against those who plead guilty which is not quantified in our severity score. In other cases, the prosecution will recommend more leniency from the judge, but the judge is not obligated to do so. Our analysis shows that in Cook County, judges consistently grant leniency to defendants who plead guilty, even when prosecutors reduce the severity of their charges. This finding is the most significant finding in our research.

A.2. Gender. Gender:

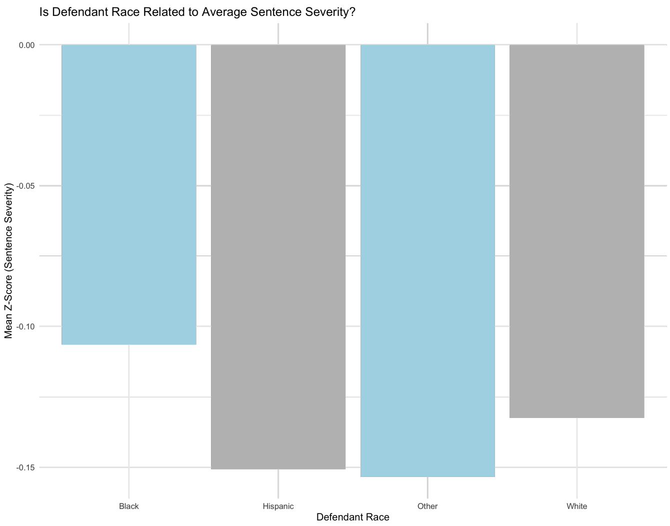
Fig. 2. Defendant Gender



Our inferential analysis also suggests that female defendants are granted greater leniency than male defendants. It is important to note that while this finding is consistent with other research, the magnitude of this disparity is impacted by the number of female defendants vs. male defendants. The small number of female defendants in our dataset means that our observations should be interpreted through a critical lens. Women are also much more likely to accept plea deals than men, which could explain the leniency they were granted. Our discussion of our regression findings casts more light on gender disparities.

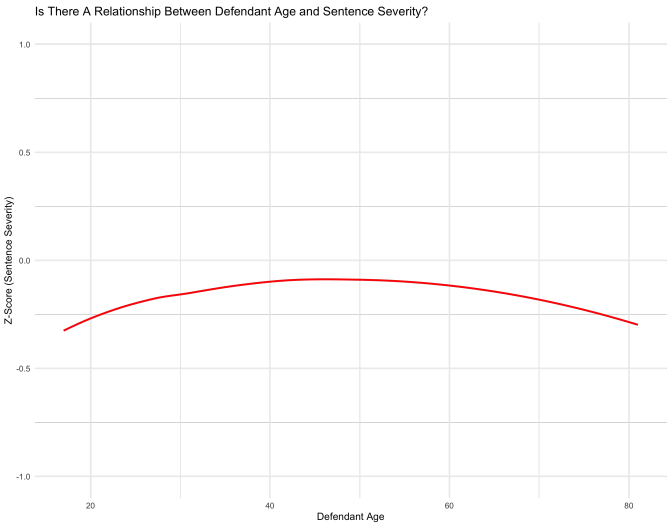
A.3. Race. Our analysis of racial sentencing disparities suggests limited disparities between defendant race and severity of sentencing outcome. We compared racial factors and found that while Black defendants received the harshest

Fig. 3. Defendant Race



sentences compared to other groups, the differences in outcomes were far less pronounced than other factors. This finding in Cook county is encouraging because it suggests that judges do not substantially racially discriminate in sentencing decisions. It is important to note that racial bias plays a large role in the criminal justice system, however, in Cook county our analysis suggests that at the sentencing level, these biases are surprisingly insignificant.

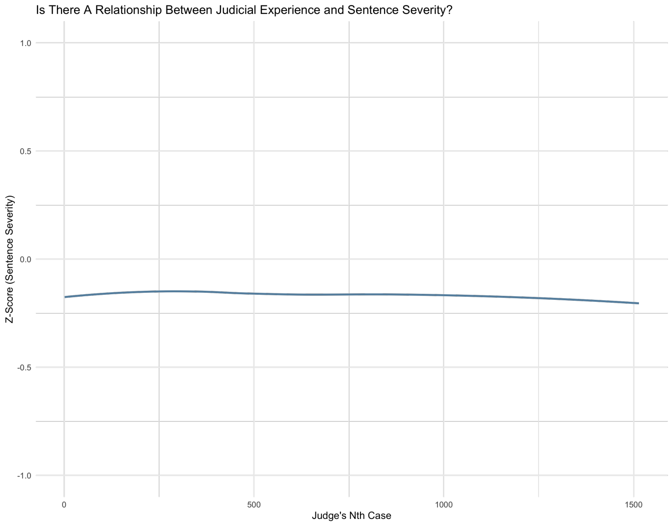
Fig. 4. Defendant Age



A.4. Age. Our analysis of defendant age suggests that those in the age range of 40-50 tend to receive the harshest relative sentences. These results are consistent with other research findings. It is important to note that this severity graph resembles the bell curve of our raw observation numbers, meaning that extensive interpretation could be skewed. It also resembles the bell curve of the general population. We saw greater observance of crimes in the middle age, and far fewer in elderly defendants. This finding is consistent with other research and is understandable. Juveniles and elderly defendants are favored in sentencing for different reasons. Younger offenders are more likely to have a clean criminal record and be tried as juveniles resulting in more favorable outcomes. Elderly defendants suffer health conditions and have fewer years of their lives left and are often granted forms of clemency as a result.

A.5. Judge Experience. Our research suggests no relationship between judge experience and deviation in sentence severity. Our results suggest that judges are extremely consistent in sentencing over the course of their careers, which aligns with our assumption going into the paper based on a comprehensive study cited earlier in our related work section.

Fig. 5. Judge Experience



Sentencing consistency is an incredibly important finding, as defendants do not have a choice in which judge they appear before. If there was a great sentencing disparity between judges with less or more experience, defendants would not be treated equally across the board. Consistency suggests each judge treats each new defendant as a new individual with his or her own story and potential extenuating circumstances. This is a critical finding in our research as it reinforces the integrity of the judicial system.

Table 3. Results: Linear Regression

Variable	Coefficient	Standard Error
const	-0.1326***	-0.0324
age_derived	0.0055***	-0.0002
is_guilty_plea	-0.2956***	-0.0180
nth_case	-0.0000***	0.0000
is_female_derived	0.0535	-0.0552
is_innocent_plea	0.1629***	-0.0219
is_male_derived	0.1567***	-0.0546
is_black_derived	-0.0092	-0.0107
is_white_derived	-0.0393***	-0.0117
is_hisp_derived	-0.0467***	-0.0115
is_other_derived	-0.0375	-0.0256
R-squared	0.0215	
R-squared Adj.	0.0214	

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

A.6. OLS Model. The results of our linear regression model suggest that the most significant factor in sentencing severity is a defendant's plea decision. This regression model corroborates our findings in the visualization section but quantifies the weighting and significance of each factor. We find that judge experience and age of defendant are essentially meaningless predictors of severity outcome, having a weighting of 0 and 0.006 respectively, meaning that judges decision making remains constant throughout their careers and do not discriminate based on a defendant's age. We also find that the race of a defendant does not have a meaningful impact on sentencing outcome. The only statistically significant racial findings meaning a p-value of less than 0.05, were in white and hispanic defendants and the results were essentially identical with a difference in weight of 0.007. This finding is encouraging because it supports the notion that judges do not racially discriminate in their sentencing decisions. The only statistically significant finding with regards to defendant profile was gender, where male defendants received much harsher sentences than female defendants where being male meant a mean z-score increase of 0.15, with a high probability of accuracy. While this is the only problematic example of judge bias in our analysis, it is important to note that this bias exists substantially across regions and districts in the United States. Additionally, considering the majority of defendants were male in our dataset, these biases had a relatively minimal impact on the average defendant. Leading into our predictive analysis, we hypothesize that these variables have little predictive power for sentencing severity at the sentencing stage.

B. Predictive Analysis: Testing the Hypothesis that these variables don't really matter. Based on the findings of our inferential and causal analysis of age, race, gender, plea type, and judicial experience in sentence severity, we hypothesize that these variables have little influence on sentence severity at the sentencing level. To test this hypothesis, we ran two predictive models below using age, race, gender, plea type, and judicial experience as independent variables. If these models are not accurate, the hypothesis will be confirmed.

Table 4. Results (Decision Tree Model)

Feature	Importance
is_guilty_plea	0.09350083212895030
nth_case	0.38682680733778600
is_female_derived	0.009375931675132370
is_innocent_plea	0.0
is_male_derived	0.04955010414505560
is_black_derived	0.0346409823984445
is_white_derived	0.024112595309624300
is_hisp_derived	0.007185224914617120
is_other_derived	0.009067589454828740
age_derived	0.3857399326355610
R-squared	0.003
MSE	0.311
MAE	0.426

B.1. Decision Tree Model. To extend our analysis of our prior linear regression model we created a decision tree model that iterated over the factors we identified earlier in our analysis, to get a better understanding of the root causes of sentencing disparities. It is important to note the magnitude of the MSE and MAE values. Considering the spread of z-scores for each factor rarely exceeded 0.4, and the difference between plea decisions was approximately 0.4, our decision tree model suggests that the features we identified have a very limited impact on z-score. The R-Squared value of 0.003 suggests that this model does not reliably predict z-score as a function of any of the features we identified. An R-Squared value of 0 would indicate that the model is equal to a model that returns the sample mean. As such, our model is barely better than a model that predicts the sample mean every time. These results when interpreted alongside our earlier regression corroborate our hypothesis that judge bias cannot be easily interpreted based on defendant profile features that influence disparate outcomes in the justice system. In many ways this is encouraging because it suggests judges do not substantially discriminate against defendants based on immutable characteristics.

B.2. Random Forest Model. Results (Random Forest Model): Our random forest model is quite similar to our decision tree model. Despite having a higher relative R-Squared value, the absolute value of R-Squared is still so close to zero

Table 5. Results (Random Forest Model):

Feature	Importance
is_guilty_plea	0.028987331737240000
nth_case	0.445452107065452
is_female_derived	0.02345879881929660
is_innocent_plea	0.0493848835967879
is_male_derived	0.02347301903482910
is_black_derived	0.0326038133695236
is_white_derived	0.019243223271158800
is_hisp_derived	0.012399326465675600
is_other_derived	0.009970710838068860
age_derived	0.35502678580196800
R-squared	0.024
MSE	0.305
MAE	0.423

that we can interpret minimal correlation between our selected features and resultant z-score. Again, this model is barely better than a model that just returns the sample mean. It is important to note that because plea deals are binary, both the forest and decision tree model did not weigh either factor heavily and as a result the accuracy of both models suffered. A holistic analysis of our linear regression and subsequent models suggests that plea decisions are the most significant feature in determining sentencing outcome for a defendant. All models reinforce the notion that judges in Cook County do not discriminate based on race, age, or experience, and thus the integrity of sentencing outcomes in the county are consistent at the sentencing level.

6. Conclusion

While our visualizations uncovered some interesting trends in the data, our linear regressions and statistical analysis found comprehensively that the features we analyzed, outside of plea decisions, were a poor predictor of sentence severity outcomes. Our analysis finds that in Cook County immutable characteristics such as race, age, gender, play a limited role in sentencing outcomes and reinforce the notion that sentencing outcomes vary based on case-by-case factors that are not related to general defendant characteristics. These findings are in many ways encouraging. In theory, sentence outcomes should not vary significantly based on a defendant's profile but instead be determined by the specifics of their offense. Our analysis suggests that judges in Cook County are professionals who set aside biases in the categories that we analyzed. Additionally, the finding that judge experience does not influence sentencing severity is notable. While the distribution of sentencing outcomes was more pronounced in novice judges, their mean severity score as a group mimicked their more experienced counterparts. This means that defendants in theory do not receive preferential treatment based on a judge's experience.

7. Limitations and Implications

Our analysis was limited by multiple factors. Because we created binary variables that subsetted across multiple features, we were limited in accuracy by the size of our dataset. The distribution of observable features was skewed in certain instances, notably in regard to gender, causing our regression to have a lower level of certainty and significance. Additionally, we could not account for other factors that likely played a substantial role in sentencing decision making such as criminal history and financial means that have been documented to significantly impact sentence outcomes.

Additionally, it is important to note that this analysis is only based on sentencing decisions and does not account for a plethora of other factors that determine a defendant's outcome in the judicial system. Access to quality attorneys, prosecutors pursuing lesser charges, and defendants receiving parole or even appealing sentences after their conviction were not quantified in our analysis. Thus, future research and efforts at reform should aim to tackle those issues and larger trends within the justice system. One way to look more broadly at the justice system would be to look into subsetting the data to examine specific crime types (for example, drug crimes) could make it more amenable to examine multiple phases of the justice system and better inform civil justice reform moving forward. Our analysis is valuable because it shows that disparate outcomes in Cook County are not likely to be a result of judge bias. As a result, future efforts at reform should target the root causes of bias in the justice system that begin far before a case goes before a judge and perpetuate themselves far after a defendant is sentenced. Future research should aim to quantify the significance of racial, gender, and economic characteristics over the course of an individual's entry and exit from the justice system, examining attorney access, prosecutorial decision making, and parole grants to truly assess the scope of characteristic bias and discrimination in Cook County's justice system.

8. References

1. JT Gosselin, Chicago's harshest judge seeks retention with 34 decisions undone in 6 years. *Injustice Watch.* (2018).
2. J Anderson, J Kling, K Stith, Sentencing disparity: Before and after the federal sentencing guidelines. *The J. Law Econ.* **42**, 271–306 (1999).
3. A Alvarez, RD Bachman, American indians and sentencing disparity: An arizona test. *J. Crim. Justice* **24** (year?).