STAT_410_Final

Akhil Ghosh and Anthony Orso

Introduction

Sentencing is a crucial role in the legal process, serving as the culmination of a series of judicial processes that determine the fate of defendants. For a visualization on the entire legal process defendants go through, see Figure 1. This dataset, accessible through the Cook County Data Catalog (**https://datacatalog.cookcountyil.gov/Courts/Sentencing/tg8v-tm6u/data**), provides information on the defendants following a guilty plea or a guilty verdict in a trial, containing details on their sentence type and the duration of their sentence. As an integral component of the criminal justice system, sentencing outcomes reflect not only the gravity of the offense committed but also the underlying values and biases of the legal framework in place. Through examination of the sentencing data, we can gain insight into the trends concerning defendants during the sentencing process, as well as uncover hidden biases.

Methods

Data Wrangling and Cleaning

The sentencing dataset was compiled from the Cook County data catalog, containing information on criminal cases from 1901 to 2023. The dataset was cleaned and pre-processed using the R programming language with the tidyverse and lubridate packages. Table 1 provides the final list of variables and their descriptions used in this analysis (specifically the model). The variables of focus include the updated offense category, race, gender, sentence type, and charge count. Updated offense category reflects the type of offense committed, while sentence type reflects the punishment defendants were subjected to.

The dataset was pre-processed to clean and organize the data. Date columns were converted from string to date format, and categorical variables such as race and gender were collapsed into more general categories. Offense categories were binned into broader groups, such as violent crimes, non-violent crimes, sex crimes, property crimes, firearm/weapon-related crimes, DUI and license-related crimes, fraud and deception, legal process crimes, and narcotics. Additionally, sentencing types were collapsed into categories like probation, prison, mental health services, death, and supervision.

Additionally, within the cleaned dataset cases with primary charge flags and current sentence flags were retained for further analysis. This was done in order to analyze the most severe and up-to-date sentencing outcomes for each case. The pre-processing steps resulted in a decrease from 280 thousand rows to 188 thousand.

Visualizations and Tables

The plots located in Appendix A highlight the distributions of sentencing cases across race, age, and gender demographics. Figure 2 in particular highlights how the raw frequencies of the people being sentenced skew heavily towards Black race in all categories, especially with regards to the Narcotics offense. This tracks with how harsh the African American community was targeted during the war on drugs.

Figure 3 and 4 breakdown the cases by gender and age during incident, respectively. Figure 3 highlights how the gender breakdown of crimes is not equal, with men being responsible for a significantly larger portion of violent crimes than women. This could be partially due to cultural and social factors, as women are socialized to not to be as aggressive as their male counterparts. Additionally, Figure 4 highlights how the age distribution for violent crimes is also skewed younger. This could be because adolescents and young adults are still developing their cognitive and emotional regulation skills, as the brain does not fully develop until age 25.

Figure 5 shows a CMH test that indicates that there exists a high degree of association between Race and Sentencing type, even when controlling for offense category. With a p-value of 2.2e-16, we reject the null that Race and Sentencing type are independent of each other. This is significant, because while there may be other factors at play like recidivism and charge counts that are not being factored in, still indicates a point of concern. This is looked into further in Figure 6, which breaks down the rates of each sentencing type by race for violent crimes.

Model Selection

Because of patterns observed in exploratory data analysis, our initial hunch was to build a binomial generalized linear mixed model (GLMM) to account for what we hypothesized would be a random effect for race. In doing so, we sought to provide persuasive evidence of discrepancies in sentencing across racial demographics. To our surprise, the random effect had a very small standard deviation, and thus intercluster differences were much smaller than we assumed. As a result we decided to pivot toward multinomial logistic regression. We experienced difficulty in quantifying the numerical differences between supervision, probation, and prison, so we opted toward nominal multinomial logistic regression. To account for a degree of order between the variables, we set supervision as the baseline so that the odds ratio speaks to the probability of observing a harsher sentence instead of supervision.

We chose AGE_AT_INCIDENT, GENDER, CHARGE_COUNT, RACE, and offenses_binned as our explanatory variables and collapsed_sentence_type as our response variable. The baseline for gender is female, and the baseline for the crime category is non-violent crime. After programming the model, we observed not a single level of the RACE factor achieving statistical significance. In addition, we were concerned it was causing other parameters to have high p-values, as observed with a statistically insignificant intercept. After removing the RACE variable, more variables achieved statistical significance. The reported equations are below.

$$\begin{aligned}\log\left(\frac{\hat{\Pi}_{\text{prison}}}{\hat{\Pi}_{\text{supervision}}}\right) = {} & 0.774774 + 0.055244 \cdot \text{DUI/License} - 0.099799 \cdot \text{Firearm} + 0.312678 \cdot \text{Fraud} - 0.20746 \cdot \text{Legal} \\ & + 1.565658 \cdot \text{Narcotics} + 0.469823 \cdot \text{Property} + 1.814065 \cdot \text{SexCrime} + 0.873807 \cdot \text{Violent} \\ & + 1.640396 \cdot \text{Male} + 0.031624 \cdot \text{charge\_Count} + 0.025235 \cdot \text{age}\end{aligned}$$

$$\begin{aligned}\log\left(\frac{\hat{\Pi}_{\text{probation}}}{\hat{\Pi}_{\text{supervision}}}\right) = {} & 1.959141 + 0.718436 \cdot \text{DUI/License} - 0.421404 \cdot \text{firearm} + 1.105677 \cdot \text{Fraud} \\ & + 0.035705 \cdot \text{Legal} + 1.816690 \cdot \text{Narcotics} + 0.594715 \cdot \text{Property} + 1.632676 \cdot \text{SexCrime} \\ & + 0.606325 \cdot \text{Violent} + 0.541416 \cdot \text{Male} + 0.059268 \cdot \text{charge\_Count} + 0.007725 \cdot \text{age}\end{aligned}$$

A likelihood ratio test was conducted to compare the deviances between the null and full models, and the reported test statistic was 18512 on 32 degrees of freedom for a chi-squared distribution. This reports an incredibly small p-value of < 2.2 x 10^-16. We, thus, reject the null hypothesis that all coefficients are zero and conclude at least one parameter is statistically significant.

Model Interpretations

The coefficients of these parameters provide compelling insights around crime and reflect much of the EDA we conducted. Overall, the odds of observing probation or prison over supervision for most offenses increase compared to non-violent crime. Although the idea of this is obvious, our parameters provide evidence to this. The magnitude of increasing the odds of observing prison or probation instead of supervision are highest for sexual assault and narcotics. When a sexual assault has been proven beyond reasonable doubt and the offender has a guilty verdict, the punishment is usually harsh due to the heinous nature of rape.

In addition, the War on Drugs that has imprisoned so many people is reflected in the equally high and positive coefficients for the narcotics parameter. Assuming a drug offense is non-violent, it's shocking to see a multiplicative impact on the odds ratio similar to a charge as heinous as rape. Thus, although the race variable failed to achieve statistical significance, latent representations with variables such as narcotics do provide evidence for systemic racism in the criminal justice system. In fact, the odds of going to jail instead of supervision increase by a factor of exp(1.565658) (4.786) for a drug offense compared to non-violent crime. In addition, the odds of going to jail instead of supervision increase by a factor of exp(1.814065) (6.135) when there is a sex offense compared to non-violent crime.

Other notable parameter estimates occurred with age, gender, and the number of charges an offender carried during the proceedings. The multiplicative impact of male gender relative to female gender was positive (especially for prison compared to supervision), suggesting men go to prison over receiving supervision at a much higher rate than women. This tracks with our EDA where we saw the vast majority of violent crimes, narcotics charges, and sex offenses are given to men. Based on the aforementioned parameters around violent offense categories, it makes sense that male gender would be associated with a higher likelihood of going to prison instead of supervision. In addition, the odds of observing probation over supervision multiplies by a factor of 1.06106 for a one unit increase in charge count. Although that sounds small, offenders can carry many charges in their case. To illustrate, the odds of someone with 10 charges getting probation instead of supervision is 64% higher than someone who only has one charge. Finally, the odds of observing prison over supervision multiplies by a factor of 1.025556 for a one-unit increase in age. To illustrate, the odds of a 30-year-old going to prison instead of supervision are 28% higher than those of a 20-year old given other levels are fixed. This could be due to criminal sentencing being mitigated by the fact that younger people have less developed brains, are less capable of reasoning, and are more prone to emotional instability. Thus, judges may be more lenient on younger defendants.

Model Evaluation

To assess our model's performance, we used a confusion matrix. Our model had 61% accuracy, and that was mainly due to the model not once predicting supervision. This is likely attributed to our data set having very few supervision entries compared to probation and prison, which could have led to the model lacking information to learn the differences between the three levels. The most obvious solution is to revisit our data cleaning to ensure we properly binned sentencing levels accurately. Another solution is to add latent variables to account for racial discrepancies in sentencing. The best way to do this is to bin law enforcement districts into geographic regions, such as the South Side and the Loop. Independence testing on contingency tables showed clear discrepancies according to neighborhoods, so predominantly white neighborhoods would have observed more supervision, potentially providing additional information for the model to learn. This also could have overcome the statistical insignificance of the race variable which resulted in lost information.

Appendix A

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0     v purrr   1.0.1
## v tibble  3.2.1     v dplyr   1.1.4
## v tidyr   1.3.0     v stringr 1.5.0
## v readr   2.1.2     v forcats 1.0.0

## Warning: package 'dplyr' was built under R version 4.2.3

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

## Warning in !is.null(rmarkdown::metadata$output) && rmarkdown::metadata$output
## %in% : 'length(x) = 2 > 1' in coercion to 'logical(1)'
```

```
##
## Attaching package: 'kableExtra'
##
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

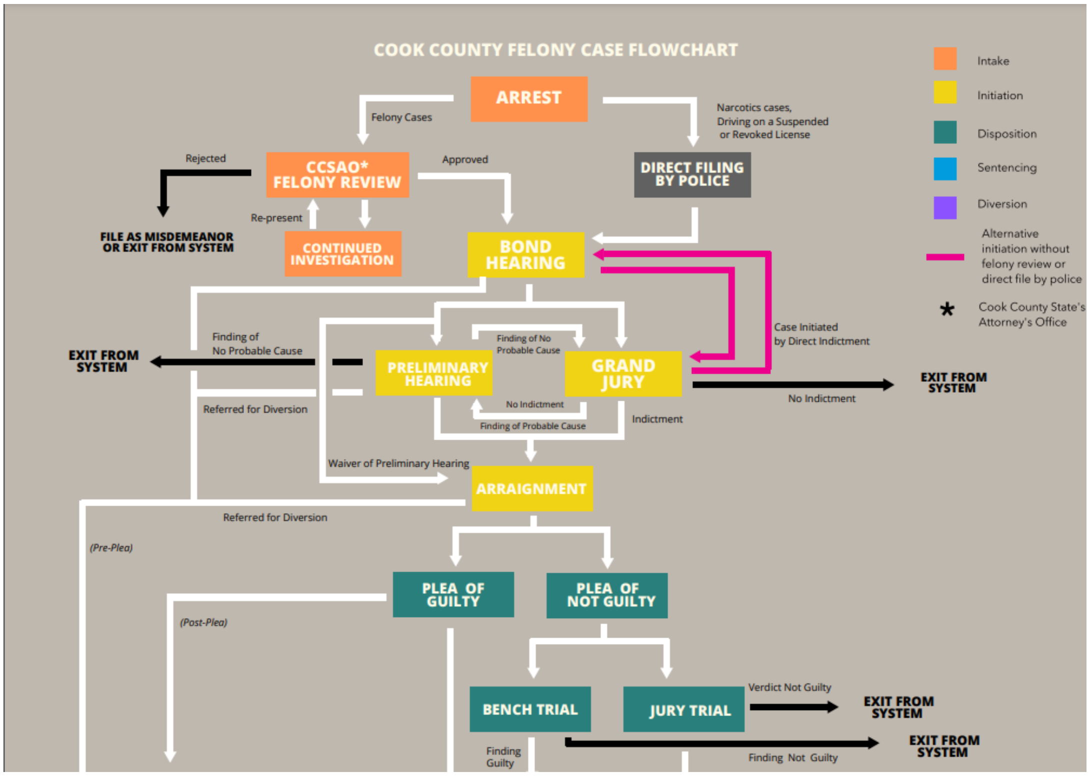

Figure 1: Legal Process Flow

Table 1: Description of Variables in Modelling Process

Variable	Description
CASE_ID	Internal unique identifier for each case
CASE_PARTICIPANT_ID	Internal unique identifier for each defendant associated with a case. Note that this is specific to each unique case id.
CHARGE_COUNT	Number of charges associated with one defendant in one case
offenses_binned	Collapsed offense category based upon the primary charge. Binned from UPDATED_OFFENSE_CATEGORY and CHARGE_DISPOSITION_TITLE, with 10 different categories: 'Violent Crimes', 'Non-violent Crimes', 'DUI and License-related Crimes', 'Firearm/Weapon-related Crime', 'Fraud and Deception', 'Legal Process Crime', 'Narcotics', 'Sex Crimes', Property Crimes', 'Other Offenses'
collapsed_sentence_type	Sentence type binned from SENTENCE_TYPE and COMMITMENT_TYPE if SENTENCE_TYPE was N/A. Types are: 'Prison', 'Probation', 'Death', 'Supervision'. Death indicates person passed away during process, and Supervision indicates probation with opportunity to get the charge erased from permanent record at the end.
age_at_incident	Age of defendant at date of incident, as recorded by law enforcement or self-reported by defendant

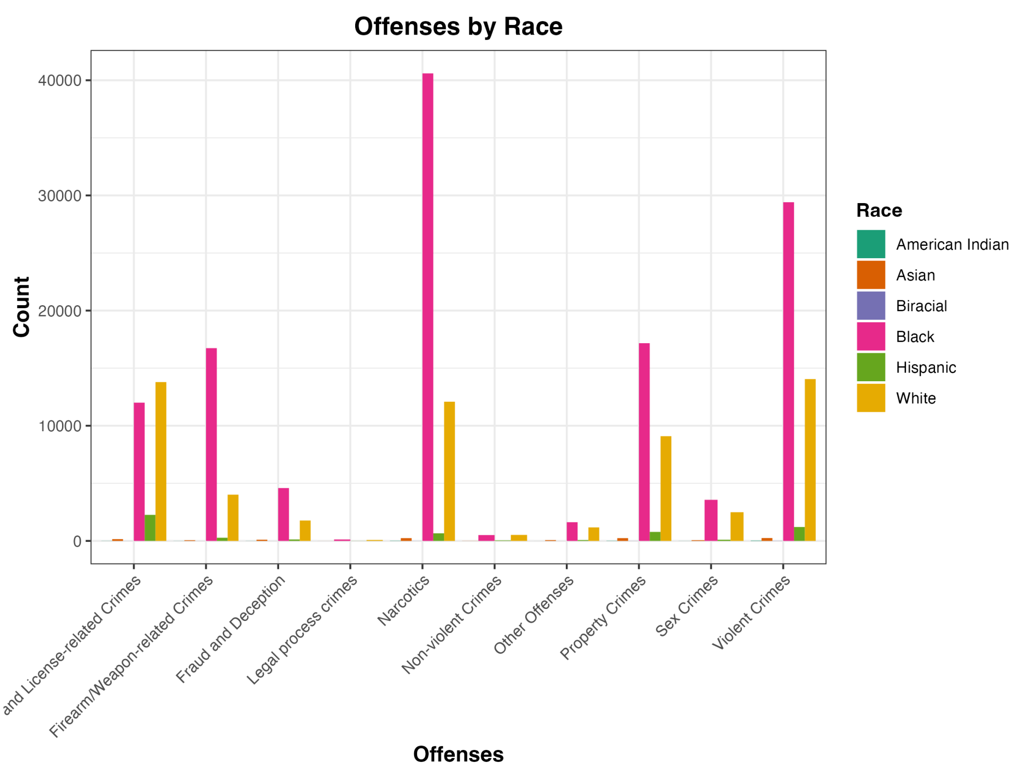

Figure 2: Race Distribution of Sentencing Outcomes

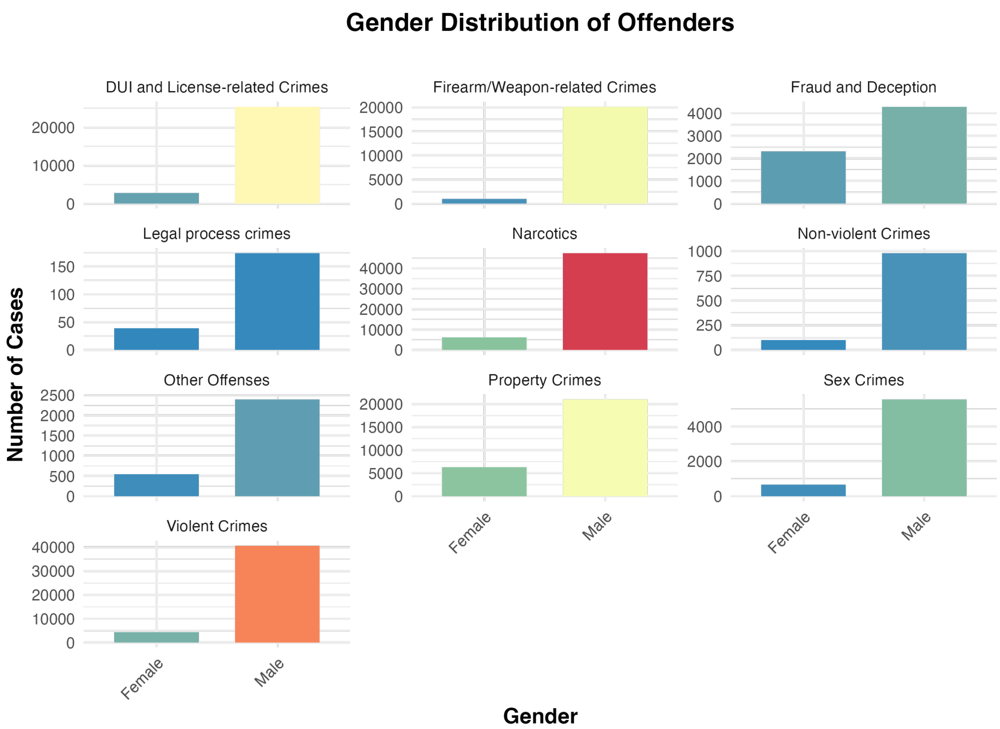

Figure 3: Gender Distribution of Offenders

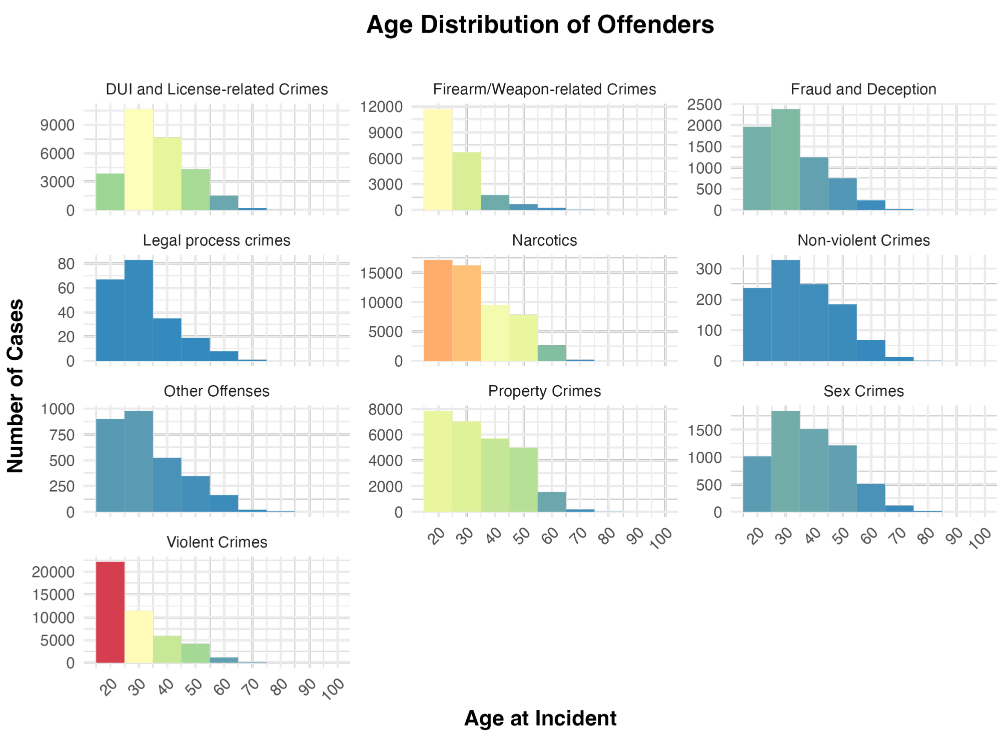

Figure 4: Age Distribution of Offenders

```
	Cochran-Mantel-Haenszel test

data:  contingency_table_3way
Cochran-Mantel-Haenszel M^2 = 5477.5, df = 10, p-value < 2.2e-16
```

Figure 5: Three-Way Contingency Table

Contingency table with percentages for Violent crimes

	Prison Count	Probation Count	Supervision Count	Prison Percentage	Probation Percentage	Supervision Percentage
American Indian	51	39	3	54.84	41.94	3.23
Asian	395	736	50	33.45	62.32	4.23
Biracial	9	20	1	30.00	66.67	3.33
Black	76423	48481	1424	60.50	38.38	1.13
Hispanic	2247	3159	128	40.60	57.08	2.31
White	24974	32792	1305	42.28	55.51	2.21

Figure 6: Table of rates of sentencning

Appendix B

```r
library(tidyverse)
library(lubridate)
url <- "https://datacatalog.cookcountyil.gov/api/views/tg8v-tm6u/rows.csv?accessType=DOWNLOAD"
sentencing_original <- read_csv(url)

#Converting date columns from string to dates
sentencing <- sentencing_original %>%
  mutate(RECEIVED_DATE = mdy_hms(RECEIVED_DATE) %>% as.Date(),
         DISPOSITION_DATE = mdy_hms(DISPOSITION_DATE) %>% as.Date(),
         SENTENCE_DATE = mdy_hms(SENTENCE_DATE) %>% as.Date(),
         INCIDENT_BEGIN_DATE = mdy_hms(INCIDENT_BEGIN_DATE) %>% as.Date(),
         INCIDENT_END_DATE = mdy_hms(INCIDENT_END_DATE) %>% as.Date(),
         ARREST_DATE = mdy_hms(ARREST_DATE) %>% as.Date(),
         FELONY_REVIEW_DATE = mdy_hms(FELONY_REVIEW_DATE) %>% as.Date(),
         ARRAIGNMENT_DATE = mdy_hms(ARRAIGNMENT_DATE) %>% as.Date(),
         RACE = str_to_title(RACE),
         LENGTH_OF_CASE_IN_DAYS = as.numeric(DISPOSITION_DATE - RECEIVED_DATE))

#collapsing the sentences categories
sentencing <- sentencing %>%
  mutate(SENTENCE_TYPE = case_when(
    SENTENCE_TYPE == "Conversion" ~ COMMITMENT_TYPE,
    TRUE ~ SENTENCE_TYPE
  ),collapsed_sentence_type = case_when(
    SENTENCE_TYPE %in% c("Cook County Boot Camp","2nd Chance Probation",
                         "Conditional Discharge", "Conditional Release",
                         "Probation", "Probation Terminated Instanter",
```

```r
                         "Probation Terminated Satisfactorily",
                         "Probation Terminated Unsatisfactorily") ~ "Probation",

    SENTENCE_TYPE %in% c("Jail", "Prison", "Illinois Department of Corrections") ~ "Prison",
    SENTENCE_TYPE == "Inpatient Mental Health Services" ~ "Mental Health Services",
    SENTENCE_TYPE %in% c("Natural Life", "Death") ~ "Death",
    SENTENCE_TYPE %in% c("Vocational Rehabilitation Impact Center(VRIC)","Supervision",
                         "Cook County Impact Incarceration Program") ~ "Supervision",
    TRUE ~ SENTENCE_TYPE
  ))


#collapsing the gender categories

sentencing <- sentencing %>%
  mutate(RACE = case_when(
    RACE %in% c("White", "White [Hispanic Or Latino]") ~ "White",
    RACE %in% c("Asian") ~ "Asian",
    RACE %in% c("Black") ~ "Black",
    RACE %in% c("Hispanic", "White/Black [Hispanic Or Latino]") ~ "Hispanic",
    RACE %in% c("American Indian") ~ "American Indian",
    RACE %in% c("Biracial") ~ "Biracial",
    TRUE ~ NA_character_
  ), GENDER = case_when(
    GENDER %in% c("Male", "Male name, no gender given") ~ "Male",
    GENDER %in% c("Female") ~ "Female",
    TRUE ~ NA_character_
  )) %>%
  filter(!is.na(RACE) | !is.na(GENDER))

#binning offense categories

violent_categories = c("Aggravated Assault Police Officer", "Aggravated Battery", "Aggravated Robbery",

nonviolent_categories <- c('Official Misconduct', 'Dog Fighting', 'Possession of Explosives', 'Possessi

sex_categories <- c("Attempt Sex Crimes", "Sex Crimes", 'Failure to Register as a Sex Offender','Violat

property_categories <- c('Criminal Trespass To Residence',"Criminal Damage to Property", "Home Invasion

firearm_categories <- c("Aggravated Assault Police Officer Firearm", "Aggravated Battery With A Firearm

dui_categories <- c("Aggravated DUI", "Driving With Suspended Or Revoked License", "DUI", "Escape - Fai

fraud_categories <- c("Benefit Recipient Fraud", "Forgery", "Fraud", "Fraudulent ID", "Identity Theft",

legal_categories <- c('Perjury', 'Bribery','Communicating With Witness','Tampering', 'Perjury','Obstruc

narc_categories <-  c("Narcotics", 'Possession of Contraband in Penal Institution')

sentencing <- sentencing %>% mutate(offenses_binned =
  case_when(
    UPDATED_OFFENSE_CATEGORY %in% violent_categories ~ "Violent Crimes",
```

```r
    UPDATED_OFFENSE_CATEGORY %in% nonviolent_categories ~ "Non-violent Crimes",
     UPDATED_OFFENSE_CATEGORY %in% sex_categories ~ "Sex Crimes",
    UPDATED_OFFENSE_CATEGORY %in% property_categories ~ "Property Crimes",
    UPDATED_OFFENSE_CATEGORY %in%  firearm_categories ~ "Firearm/Weapon-related Crimes",
    UPDATED_OFFENSE_CATEGORY %in% dui_categories ~ "DUI and License-related Crimes",
    UPDATED_OFFENSE_CATEGORY %in%  fraud_categories ~ "Fraud and Deception",
    UPDATED_OFFENSE_CATEGORY %in% legal_categories ~ "Legal process crimes",
    UPDATED_OFFENSE_CATEGORY %in% narc_categories ~ "Narcotics",
    UPDATED_OFFENSE_CATEGORY == "PROMIS Conversion" ~ NA_character_,
    TRUE ~ "Other Offenses"
  )
)

sentencing %>% filter(PRIMARY_CHARGE_FLAG == T & CURRENT_SENTENCE_FLAG==T& is.na(offenses_binned)) %>% g

#Still ~10k other offenses that are missing due to promise conversion. Can we find a way to utilize the
#Since I'm only really interested in the analysis of primary charges, only doing it for them. In case a


violent_pattern_regex <- str_c(c("AGG", "MURDER", "BURGLARY", "ROBBERY", "ARMED","BATTERY",'BATTERY/','I

sentencing <- sentencing %>%
  mutate(offenses_binned =
           if_else((PRIMARY_CHARGE_FLAG == T & CURRENT_SENTENCE_FLAG == T & is.na(offenses_binned)),
                   case_when(
                     str_detect(DISPOSITION_CHARGED_OFFENSE_TITLE, violent_pattern_regex) ~ "Violent Cri
                     str_detect(DISPOSITION_CHARGED_OFFENSE_TITLE, str_c("SEX")) ~ "Sex Crimes",
                     str_detect(DISPOSITION_CHARGED_OFFENSE_TITLE, str_c("BAIL|DISORDERLY|PCS|THREATEN|I
                     str_detect(DISPOSITION_CHARGED_OFFENSE_TITLE, str_c("UUW|FIREARM|WEAPON")) ~ "Firea
                     str_detect(DISPOSITION_CHARGED_OFFENSE_TITLE, str_c("DUI|DRIVING|VEHICLES|DRIVING|I
                     str_detect(DISPOSITION_CHARGED_OFFENSE_TITLE, str_c("SUBSTANCE|POSS|MANU|DRUG|NARC
                     str_detect(DISPOSITION_CHARGED_OFFENSE_TITLE, str_c("THEFT|DISP MERCH|INVASION|DAMA
                     str_detect(DISPOSITION_CHARGED_OFFENSE_TITLE, "FORGERY|CREDIT CARD") ~ "Fraud and I
                     T ~ offenses_binned),
                   offenses_binned)
  )


#Visualizing Missing Data

sentencing %>%
  summarise_all(funs(sum(is.na(.)))) %>%
  gather(key = "Variable", value = "Number_of_NAs") %>%
  arrange(desc(Number_of_NAs))

sentencing_imputed<- sentencing
mean_age <- mean(sentencing_imputed$AGE_AT_INCIDENT, na.rm = TRUE)
sentencing_imputed$AGE_AT_INCIDENT <- ifelse(is.na(sentencing_imputed$AGE_AT_INCIDENT), mean_age, senter



sentencing_subsetted <- sentencing_imputed %>% filter(PRIMARY_CHARGE_FLAG & CURRENT_SENTENCE_FLAG) %>% s
```

```r
# sentencing %>% filter(PRIMARY_CHARGE_FLAG&CURRENT_SENTENCE_FLAG) %>% group_by(SENTENCE_JUDGE) %>% sum
# sentencing_original
# sentencing_subsetted %>% filter(PRIMARY_CHARGE_FLAG&CURRENT_SENTENCE_FLAG) %>% pull(CHARGE_COUNT) %>%
# sentencing_subsetted %>% filter(PRIMARY_CHARGE_FLAG&CURRENT_SENTENCE_FLAG) %>% group_by(SENTENCE_JUDG


#Checking for NAs
# sentencing_subsetted %>%
#   summarise_all(funs(sum(is.na(.)))) %>%
#   gather(key = "Variable", value = "Number_of_NAs") %>%
#   arrange(desc(Number_of_NAs))

sentence_final <- sentencing_subsetted %>%
  filter(!is.na(GENDER) & !is.na(RACE) & collapsed_sentence_type %in% c("Prison","Probation","Supervisi

sentence_final %>%
  summarise_all(funs(sum(is.na(.)))) %>%
  gather(key = "Variable", value = "Number_of_NAs") %>%
  arrange(desc(Number_of_NAs))
```

```r
# Visualizing the sentencing cases broken down by race and offense type

race_offense <- ggplot(sentence_final, aes(x=offenses_binned, fill=RACE)) +
  geom_bar(position = "dodge") +
  scale_fill_brewer(palette="Dark2") + # You can try other palettes like "Set1", "Dark2", "Pastel1", et
  theme_bw() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        axis.title = element_text(size = 12, face = "bold"),
        plot.title = element_text(size = 14, face = "bold", hjust = 0.5),
        legend.title = element_text(face = "bold")) +
  labs(x="Offenses",
       y="Count",
       title="Offenses by Race",
       fill="Race")
#ggsave("race_sentence.png", race_offense, width = 8, height = 6, dpi = 300)

#Visualizing cases by age and offense type

age_offense <- sentence_final %>%
  ggplot(aes(x = AGE_AT_INCIDENT)) +
  geom_histogram(aes(fill = ..count..), binwidth = 10)  +
  scale_x_continuous(limits = c(15, 100), breaks = seq(0, 100, by = 10)) + # Add limits and breaks for
  scale_fill_distiller(palette = "Spectral", direction = -1, guide = "none") +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    axis.title = element_text(size = 12, face = "bold"),
    plot.title = element_text(size = 14, face = "bold", hjust = 0.5, margin = margin(b = 20)),
    legend.title = element_text(face = "bold"),
    plot.margin = margin(20, 10, 10, 10)
  ) +
  labs(
    title = "Age Distribution of Offenders",
```

```r
    x = "Age at Incident",
    y = "Number of Cases"
  ) +
  facet_wrap(~ offenses_binned, scales = "free_y", nrow = 4)


#ggsave("gender_age.png", age_offense, width = 8, height = 6, dpi = 300)

Gender_offense <- sentence_final %>%
  ggplot(aes(x = GENDER)) + # Update the x variable to GENDER
  geom_bar(aes(fill = ..count..), width = 0.7)  +
  scale_fill_distiller(palette = "Spectral", direction = -1, guide = "none") +
  scale_x_discrete() + # Use scale_x_discrete() for categorical variables
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    axis.title = element_text(size = 12, face = "bold"),
    plot.title = element_text(size = 14, face = "bold", hjust = 0.5, margin = margin(b = 20)),
    legend.title = element_text(face = "bold"),
    plot.margin = margin(20, 10, 10, 10)
  ) +
  labs(
    title = "Gender Distribution of Offenders",
    x = "Gender",
    y = "Number of Cases"
  ) +
  facet_wrap(~ offenses_binned, scales = "free_y", nrow = 4) # Increase the number of rows

#ggsave("gender_offenders.png", Gender_offense, width = 8, height = 6, dpi = 300)
```

```r
#Code to look at the contingency table b/w RACE and sentencing type, controlling for offense type
library(vcdExtra)
library(coin)
contingency_table_3way <- xtabs(~ RACE + collapsed_sentence_type + offenses_binned, data = sentence_fina

# Perform the CMH test
cmh_test <- mantelhaen.test(contingency_table_3way, conf.level = 0.95)

#Getting the contingency table for sentencing in violent crimes broken down by race

library(kableExtra)
violent_crimes_data <- filter(sentence_final, offenses_binned == "Violent Crimes")
contingency_table_violent <- xtabs(~ RACE + collapsed_sentence_type, data = sentence_final)

# Convert the contingency table to a data frame
contingency_df_violent <- as.data.frame.matrix(contingency_table_violent)
row_percentages_violent <- round(prop.table(contingency_table_violent, margin = 1) * 100, 2)

contingency_df_violent <- as.data.frame.matrix(contingency_table_violent)
percentages_df_violent <- as.data.frame.matrix(row_percentages_violent)

# Combine the contingency table and row percentages into a single data frame
combined_df_violent <- data.frame(contingency_df_violent, percentages_df_violent)
```

```r
rownames(combined_df_violent) <- rownames(contingency_df_violent)

# Rename columns to include "Count" and "Percentage" labels
colnames(combined_df_violent) <- c(outer(colnames(contingency_df_violent), c("Count", "Percentage"), pas

# Create a styled kable
styled_kable <- kable(combined_df_violent, caption = paste("Contingency table with percentages for", vio
  kable_styling(bootstrap_options = c("striped", "hover", "condensed", "responsive"))
```

```r
# table(sentence_final$offenses_binned)
# table(sentence_final$collapsed_sentence_type)

sentence_final$collapsed_sentence_type = factor(sentence_final$collapsed_sentence_type, levels = c("Pris
sentence_final$offenses_binned = factor(sentence_final$offenses_binned, levels = c("Non-violent Crimes"
                                                                       "Fraud and Deception","Legal pro
                                                                       "Property Crimes","Sex Crimes","

##Build VGLM
table(sentence_final$collapsed_sentence_type)
multi = vglm(collapsed_sentence_type ~ as.factor(offenses_binned) + as.factor(GENDER)+CHARGE_COUNT+AGE_A
               as.factor(RACE), family = multinomial, data = sentence_final)
multi2 = vglm(collapsed_sentence_type ~ as.factor(offenses_binned) + as.factor(GENDER)+CHARGE_COUNT+AGE_
              family = multinomial, data = sentence_final)
summary(multi)
summary(multi2)


predicted_probs <- fitted(multi2)
predicted_classes <- max.col(predicted_probs)
predicted_classes = as.factor(predicted_classes)

predicted_classes = factor(predicted_classes, levels = c(1,2,3))
predicted_classes

table(predicted_classes)
table(actual)

actual = as.factor(as.numeric(sentence_final$collapsed_sentence_type))
actual = factor(actual,levels = c(1,2,3))
actual

confusion_matrix <- confusionMatrix(predicted_classes, actual)
confusion_matrix

table(sentence_final$collapsed_sentence_type)

fit_null <- vglm(collapsed_sentence_type ~ 1, family = multinomial, data = forVGLM)
lrtest(multi,fit_null)

new = data.frame(offenses_binned = c(as.factor("Sex Crimes"),as.factor("Sex Crimes")),GENDER = c(as.fact
new

predict(multi2,new,type='response')
```