Capstone Project

Component 1: Consultant's Approach and Internal Report

Jeffery Dittenber

April 30, 2026

Contents

1 Approach and Methodology

1.0.1 Process for Meeting a New Client

Suppose I am a statistician in a statistical consulting firm. I may be the principal or a general statistician assigned to case studies and consultations. Also suppose that in this firm, the initial contact is made either telephonically or via email and that by either of these two methods a screening process would take place. The screening process would provide information that would me to develop questions for the client before our initial meeting. The screening questions could be along the lines of: 1) What do you hope to accomplish by hiring a statistical consulting firm? 2) Do you now have or do you know what sort of data you have to support achieving your goals? 3) Who are the subject matter experts who will serve as a point of contact for our firm? 4) What sort of timeline are you expecting? 5) What is your budget for this work? This would all be prior to meeting with the client.

Upon meeting the client I would be sure to initiate a positive interaction by making eye contact, creating a welcoming and hospitable environment and establishing a rapport.

Once rapport has been established my primary goal is to understand the context of the problem. I would ask the client open-ended questions about the project, informed by the screening information, to determine if there is some other information or perspective missing.

1.0.2 Determine the Correct Analysis

To determine the correct analysis I will first ensure that I completely understand the client's goals. Next I would review the background of the problem in previous studies or establish methods. I will determine if the data is currently available, what the variables are, how it is collected and how it is stored. Using this information I will create a data dictionary that includes the variable information such as variable name, classification (numeric or categorical), the count, the range of values, and an example.

If the data is not collected, then I will determine what data is required, who the contact person is in their firm to help set up a data transfer or pipeline or other arrangements depending on the situation.

I would validate the data by ensuring the data was collected in a manner that is ethical and also supports the assumptions for statistical modeling and hypothesis testing. I would also determine if the sample size of the data is sufficient to answer the question by performing a power analysis.

Once it is certain that data is validated, collected in an ethical manner, contains sufficient observations and that there is no ambiguity for the values of the variables I would clean the data and prepare any subsets needed for my analysis using a statistical software, like R.

1.0.3 Ensure Data is Ready for Analysis

During this process I would complete an a a priori power analysis depending on my methods. I would then check for duplicates, missing values, anomalies in the data and check the data to see if it meets assumptions for any parametric tests I may use. For any ambiguous data or missing values, I may remove those observations rather than using a fill method.

2 Final Internal Report

2.1 Introduction

There is an old saying in America: ”You can only get as much justice as you can afford.” This reflects the concern that people with greater financial resources may be able to obtain more favorable outcomes in the justice system than people without similar resources.

This study examines publicly available felony court records from Cleveland County, Oklahoma, to explore whether attorney type is associated with sentencing outcomes. The study compares cases involving private counsel and appointed counsel. This analysis is observational and does not establish causality. Instead, it should be interpreted as a proof-of-concept study that may support a larger, more complete future investigation.

2.1.1 Significance of the Study

Although this study cannot determine whether attorney type causes sentencing differences, it can identify patterns that may justify further investigation. If sentencing outcomes differ substantially by representation type, then a larger study could examine whether those differences persist after accounting for charge severity, criminal history, plea decisions, bond status, and other legal factors. The goal is not to make causal claims from this sample, but to determine whether public court records can be used to study possible disparities in sentencing outcomes.

2.2 Research Questions

1. Is attorney type associated with case outcome: incarceration, suspended sentence, deferred sentence, or dismissal?

2. Among sentenced cases, is attorney type associated with incarceration versus suspended or deferred sentencing?

3. Is plea type associated with attorney type?

4. Among incarceration cases, does imposed sentence length differ by attorney type?

5. Which recorded case characteristics are most important for predicting incarceration outcome in exploratory models?

2.3 Methodology and Data Collection

2.3.1 Methodology

The data for this project were collected from public felony court records in Cleveland County, Oklahoma for the year 2018. Cleveland County had 1,554 felony cases on record during this period. The purpose of the data collection process was to construct a manageable random sample of felony cases that could be reviewed manually while still providing enough observations to compare sentencing outcomes by attorney type.

Each selected case was reviewed individually and coded for variables relevant to the research questions. These variables included attorney type, sentence type, plea, crime severity, other Oklahoma criminal records, number of counts, bond status, judge, day sentenced, case duration, imposed sentence length, and actual incarceration length when available. Because this project uses public administrative records rather than a pre-cleaned research dataset, the data collection process required careful interpretation of court entries, sentencing language, attorney appearances, and case dispositions.

2.3.2 Oklahoma State Courts Network (OSCN)

The Oklahoma State Courts Network (OSCN) was the primary source of data for this project. OSCN provides online access to Oklahoma court records by county. These records are public records and are available under the Oklahoma Open Records Act. For this study, OSCN was used to identify felony cases, attorney representation, plea information, sentence type, filing dates, disposition dates, sentencing terms, and other case characteristics.

Because OSCN records are organized case-by-case and were collected manually for this project, the data could not be collected as a single preformatted dataset. Each case had to be reviewed manually. This required reading docket entries and identifying the relevant information for coding. In particular, attorney type was coded based on the attorney of record near the time of disposition, rather than simply recording every attorney who appeared at any point in the case. This distinction is important because attorney changes can occur during a case, and the analysis focuses on representation connected to the sentencing outcome.

OSCN was also used to classify the case disposition. Cases were coded according to whether the outcome was incarceration, suspended sentence, deferred sentence, dismissal, amendment to misdemeanor, warrant, blank record, or deceased defendant. These distinctions were necessary because not all records represent final felony sentencing outcomes.

In addition, OSCN was used to determine whether each defendant had other criminal records in Oklahoma public court records, regardless of time frame. This variable was coded as a binary indicator of whether other Oklahoma criminal cases were present. Because the coding does not distinguish whether those cases occurred before or after the 2018 Cleveland County case, this variable is interpreted as a general indicator of other recorded Oklahoma criminal history, not as a strict prior-conviction measure.

2.3.3 Oklahoma Department of Corrections

For cases involving incarceration, the Oklahoma Department of Corrections offender lookup was reviewed as a supplemental source when actual incarceration information was available. This information was considered during data collection because actual incarceration time may differ from the sentence imposed by the court.

However, actual incarceration length was not used as a primary outcome in the final analysis. The main analysis focuses on sentencing decisions made by the court, especially sentence type, incarceration versus suspended or deferred sentencing, and imposed incarceration sentence length. Once a defendant enters the Department of Corrections system, actual time served may be affected by correctional policies, credits, release decisions, and other post-sentencing administrative factors outside the scope of this study.

2.3.4 Random Sampling

The sampling process began by identifying felony cases filed in Cleveland County during the 2018 calendar year. The cases were placed into a sequential list, and random numbers were generated to select cases for manual review. This process continued until 150 cases were collected.

The target sample size of 150 was selected to exceed the minimum sample size suggested by the a priori power analysis. This provided a buffer because some records were expected to be unusable for the main analysis due to ambiguity, missing information, amendment to misdemeanor, unresolved warrant status, blank records, or other issues that prevented clear classification of attorney type or sentencing outcome.

After collection, each record was reviewed for eligibility. Cases were retained for analysis only when they contained enough information to classify the relevant variables for the research questions and were not ambiguous with respect to attorney type or disposition. This step was necessary because the goal of the project was not merely to summarize all sampled cases, but to analyze cases that could meaningfully address the relationship between representation type and sentencing outcomes.

2.3.5 Power Analysis

Table 1: A Priori Power Analysis for Planned Chi-Square Tests

Planned Test	**Table**	**df**	**Effect Size**	**Power**	**Required N**
Attorney type by case outcome	2×4	3	$w = 0.30$	0.80	122
Attorney type by sentencing outcome among non-dismissed sentenced cases	2×3	2	$w = 0.30$	0.80	108
Attorney type by incarceration status	2×2	1	$w = 0.30$	0.80	88
Attorney type by plea type	2×2	1	$w = 0.30$	0.80	88

Note: Power analyses assume $\alpha = 0.05$, power = 0.80, and medium effect size $w = 0.30$.

The initial sample contained 150 cases, which provided a buffer above the minimum required sample sizes for the planned chi-square analyses. This buffer was important because some records were expected to be excluded after review due to ambiguous attorney classification, unresolved case status, amended misdemeanor outcomes, blank records, or other issues that prevented clear use in the primary analysis.

After review, 142 cases remained in the cleaned dataset. The sample size varied by research question because each analysis required a different subset of cases. For example, case-outcome analyses could include dismissed cases, while sentencing analyses focused only on cases resulting in incarceration, suspended sentence, or deferred sentence. This approach ensured that each statistical test used only the cases relevant to that specific research question.

2.3.6 Variable Refinement and Coding Logic

A sample of $N = 150$ cases was collected for manual review. After review, some cases were excluded because attorney type, disposition, or case status could not be clearly classified. The cleaned dataset contained multiple outcome categories, including amended misdemeanor outcomes, deceased defendants, dismissed cases, suspended sentences, deferred sentences, incarceration sentences, unresolved warrants, and blank records.

The sample varied by research question. Case-outcome analyses used cases with clear outcomes such as incarceration, suspended sentence, deferred sentence, and dismissal. Sentencing analyses excluded dismissed cases because dismissal is not a sentencing outcome. Binary incarceration analyses compared incarceration against suspended or deferred sentences only.

Table 2: Description of Variables Used in the Main Analysis

Variable	**Type**	**Definition / Coding Logic**
Attorney Type	Binary	Classified as private or appointed based on the attorney of record near disposition.
Sentence Type	Nominal	Case outcome category, including incarceration, suspended sentence, deferred sentence, or dismissal.
Time Sentenced	Ratio	Imposed incarceration sentence length, converted to days for analysis.
Plea	Nominal	Defendant's plea, primarily guilty or no contest.
Crime Severity	Binary	Coded as violent or nonviolent based on Oklahoma statutory classification.
Other Records	Binary	Indicates whether other Oklahoma criminal records were found in public records; not interpreted as strict prior history.
Counts	Nominal	Coded as one, more than one, or none where applicable.
Bond Posted	Nominal	Method of pretrial release, such as cash bond, personal recognizance, or none.

- **Temporal Scope of Disposition:** Only the initial disposition was considered. Post-adjudication events such as motions to accelerate or motions to revoke were not included in the main analysis.

- **Attorney Classification:** Attorney type was classified based on the attorney of record near the time of disposition. This was done because attorney changes can occur during a case.

- **Sentencing Lengths:** When a felony case included multiple counts, the maximum imposed incarceration sentence was recorded for the sentencing-length analysis.

- **Other Criminal Records:** The variable for other Oklahoma criminal records was coded from public records. It does not distinguish whether those records occurred before, during, or after the 2018 case.

- **Crime Severity:** Crime severity was coded as violent or nonviolent based on Oklahoma statutory classification.

- **Categorization of Non-Sentencing Outcomes:** Dismissed cases were retained for case-outcome analysis but excluded from sentencing analysis. Amended misdemeanor outcomes, unresolved warrants, deceased defendants, and blank records were described in the sample profile but excluded from the primary sentencing analyses.

2.4 Analysis

2.4.1 Construction of Analysis Datasets

Several analysis datasets were created because each research question required a different subset of cases. The full cleaned dataset was used to describe the overall sample. An attorney-eligible dataset was created by retaining only cases classified as private or appointed counsel. A reduced case-outcome dataset excluded amended misdemeanor outcomes, warrants, deceased defendants, and blank records.

For sentencing analyses, dismissed cases were excluded because dismissal is not a sentencing outcome. The sentenced dataset included only incarceration, suspended sentence, and deferred sentence outcomes. Finally, an incarcerated-only dataset was created for the sentence-length analysis because imposed incarceration time is only meaningful for cases that resulted in incarceration.

Table 3: Analysis Dataset Sizes

Dataset	**N**	**Use**
Full cleaned dataset	142	Overall sample description.
Attorney-eligible dataset	117	Cases with attorney type classified as appointed or private.
Reduced case-outcome dataset	117	Case-outcome analyses after excluding unusable administrative outcomes.
Sentenced dataset	100	Sentencing analyses; includes incarceration, suspended, and deferred outcomes.
Incarcerated-only dataset	39	Sentence-length analysis among cases resulting in incarceration.

The sample size varied by research question. This was expected because case-outcome, sentencing, binary incarceration, and sentence-length analyses require different eligibility rules.

2.4.2 Descriptive Summary of Case Characteristics

Table 4: Summary of Key Case Characteristics

Variable Group	Category	Count	Percent (%)
Sentence / Outcome	amended	15	10.56
Sentence / Outcome	blank	5	3.52
Sentence / Outcome	deceased	1	0.70
Sentence / Outcome	deferred	39	27.46
Sentence / Outcome	dismissed	17	11.97
Sentence / Outcome	incarceration	39	27.46
Sentence / Outcome	suspended	22	15.49
Sentence / Outcome	warrant	4	2.82
Attorney Type	appointed	63	53.85
Attorney Type	private	54	46.15
Number of Counts	more	72	50.70
Number of Counts	one	45	31.69
Number of Counts	missing/not applicable	25	17.61
Other Criminal Cases	no	17	14.53
Other Criminal Cases	yes	100	85.47
Crime Severity	nonviolent	74	74.00
Crime Severity	violent	26	26.00

The sample included a mix of sentencing and non-sentencing outcomes. The most common outcome categories were deferred sentence and incarceration, each representing 27.46% of the cleaned sample. Among attorney-eligible cases, appointed counsel represented 53.85% and private counsel represented 46.15%. Most cases had other Oklahoma criminal records identified in public records, and most classified cases were nonviolent.

2.4.3 Sentence Outcome by Attorney Type

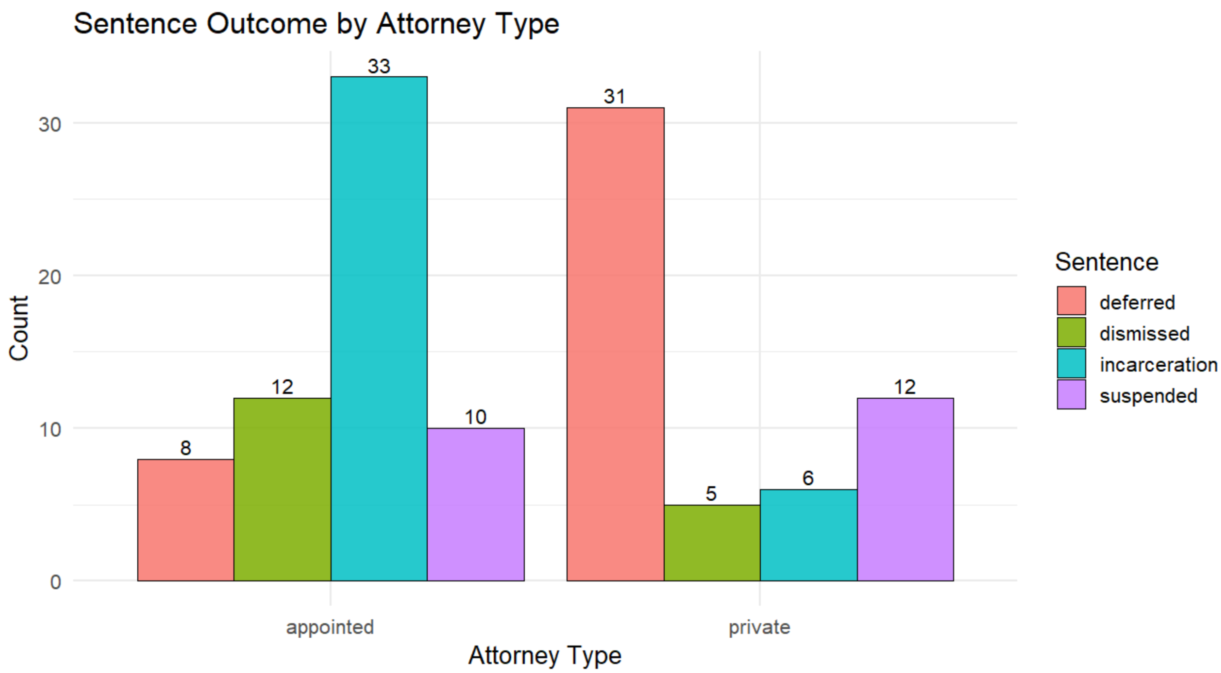

Figure 1: Sentence Outcome by Attorney Type

Figure 1 compares sentencing outcomes by attorney type. The visual pattern suggests that incarceration was more common among appointed-counsel cases, while deferred outcomes were more common among private-counsel cases.

2.4.4 Incarceration Status by Attorney Type

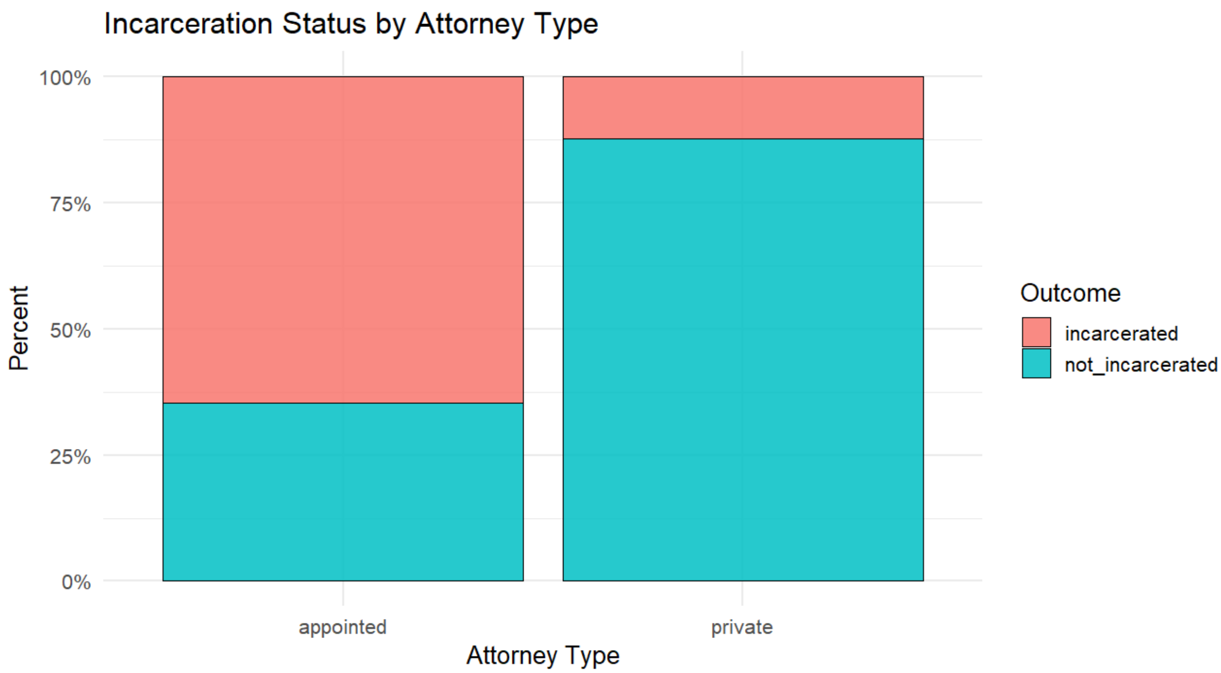

Figure 2: Incarceration Status by Attorney Type

Figure 2 compares incarceration versus suspended or deferred sentencing by attorney type. This visualization supports the binary sentencing-outcome comparison used later in the inferential analysis.

2.4.5 Plea Type by Attorney Type

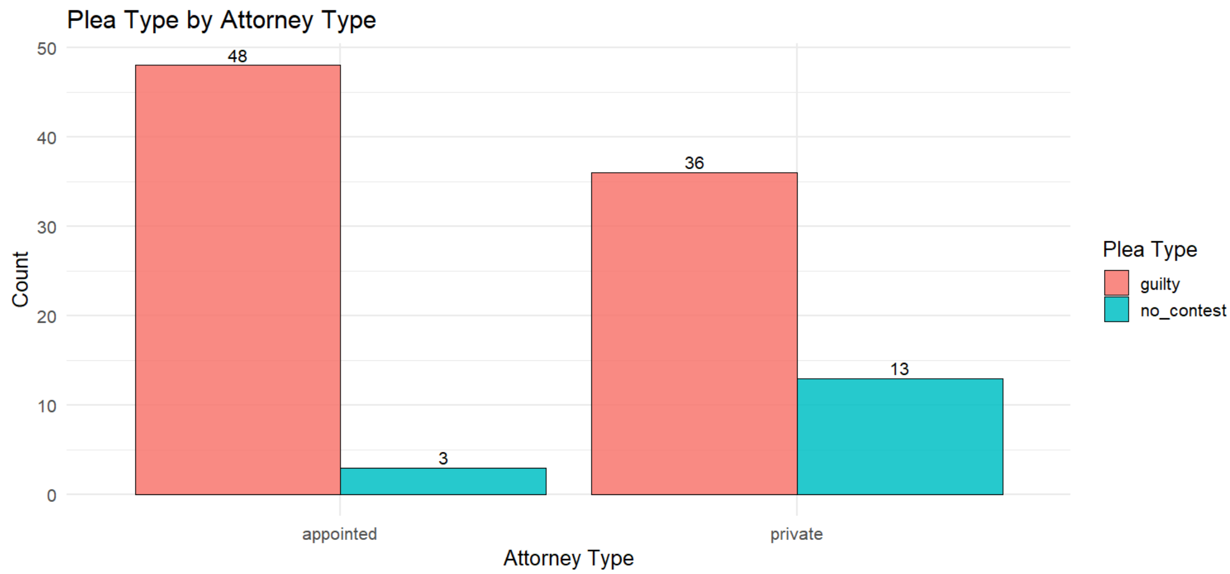

Figure 3: Plea Type by Attorney Type

Figure 3 compares plea type by attorney type. Plea type was included because plea behavior may help explain differences in sentencing outcomes.

2.4.6 Sentence Length Among Incarcerated Cases

Table 5: Sentence Length Among Incarcerated Cases

Attorney	N	Mean	SD	Median	IQR	Min	Max
Appointed	33	2005	1866	1460	2190	45	8760
Private	6	2008	1280	1825	821	730	4380

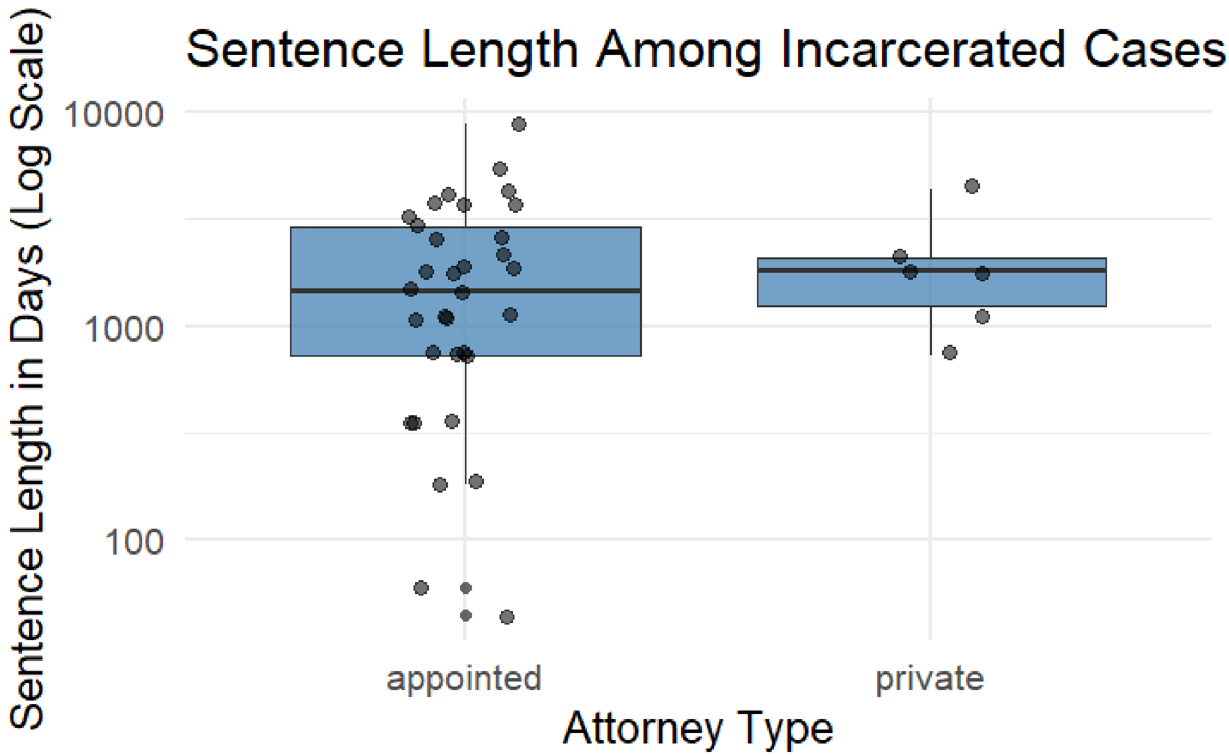

Figure 4: Sentence Length Among Incarcerated Cases by Attorney Type

Figure 4 compares imposed sentence length among cases that resulted in incarceration. Mean imposed sentence length was nearly identical by attorney type. The private-counsel group contained only six incarceration cases, so sentence-length comparisons should be interpreted cautiously.

2.4.7 Assumption and Diagnostic Checks

Table 6: Expected Cell Count Checks for Chi-Square Tests

Test	Minimum Expected	Maximum Expected	Cells Below 5
Attorney type by sentencing outcome	10.78	19.89	0
Attorney type by incarceration status	19.11	31.11	0
Attorney type by plea type	7.84	42.84	0
Attorney type by crime severity	12.74	37.74	0

Chi-Square Expected Cell Counts All expected cell counts were above 5, so the chi-square approximation was appropriate for these tests.

Table 7: Normality Check for Imposed Sentence Length

Attorney Type	N	Shapiro-Wilk p-value
Appointed	33	0.0003
Private	6	0.1886

Sentence-Length Normality Check Because the appointed-counsel group failed the normality check, the Wilcoxon rank-sum test was used for the sentence-length comparison.

Table 8: Cell Counts for Logistic Regression Outcome

Predictor	Category	Incarcerated	Not Incarcerated
Attorney Type	Private	6	43
Attorney Type	Appointed	33	18
Crime Severity	Nonviolent	29	45
Crime Severity	Violent	10	16
Other Records	No	3	14
Other Records	Yes	36	47
Counts	One	12	22
Counts	More	27	39

Table 9: Variance Inflation Factors for Logistic Regression Predictors

Predictor	VIF
Attorney Type	1.013
Crime Severity	1.010
Other Records	1.015
Counts	1.019

Logistic Regression Diagnostics The VIF values were close to 1, indicating no problematic multicollinearity among the predictors.

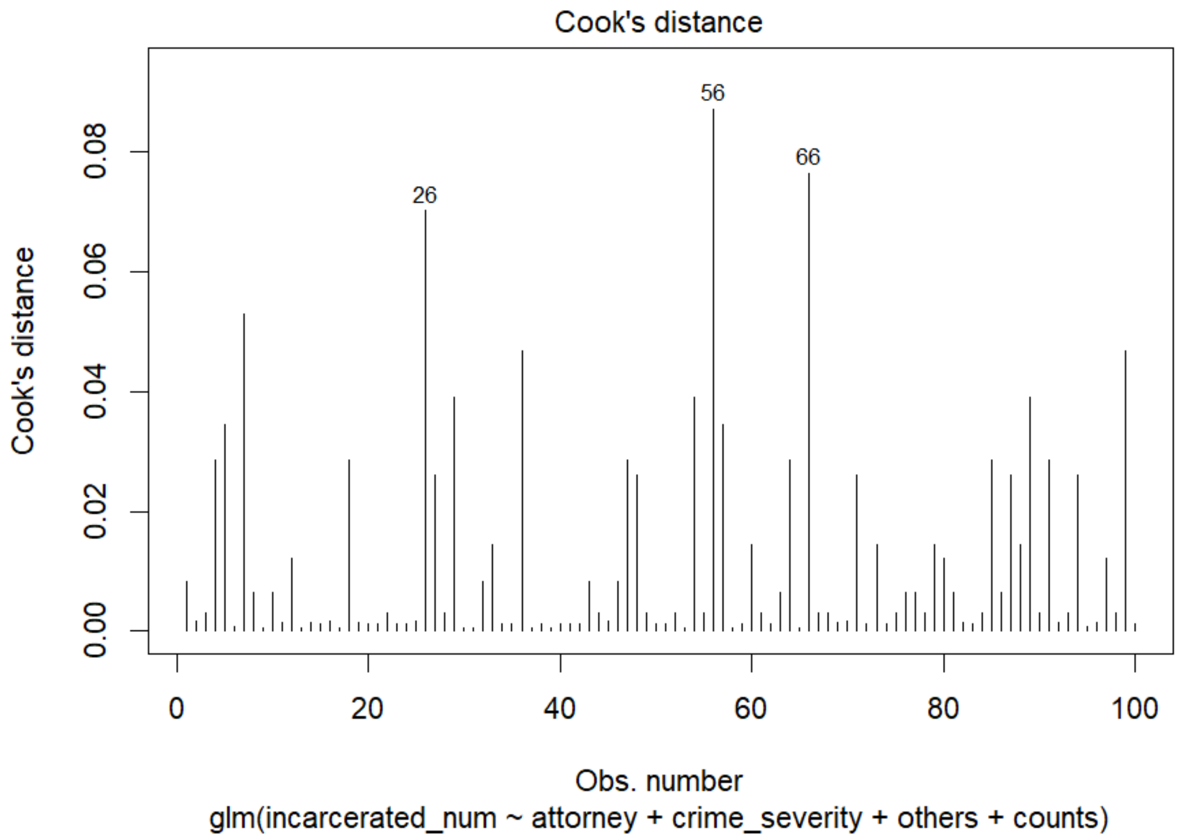

Figure 5: Cook's Distance for the Logistic Regression Model

Figure 5 shows Cook's distance values for the logistic regression model. No observation appears to have an extreme Cook's distance large enough to dominate the model, although observations 26, 56, and 66 are relatively more influential than the rest.

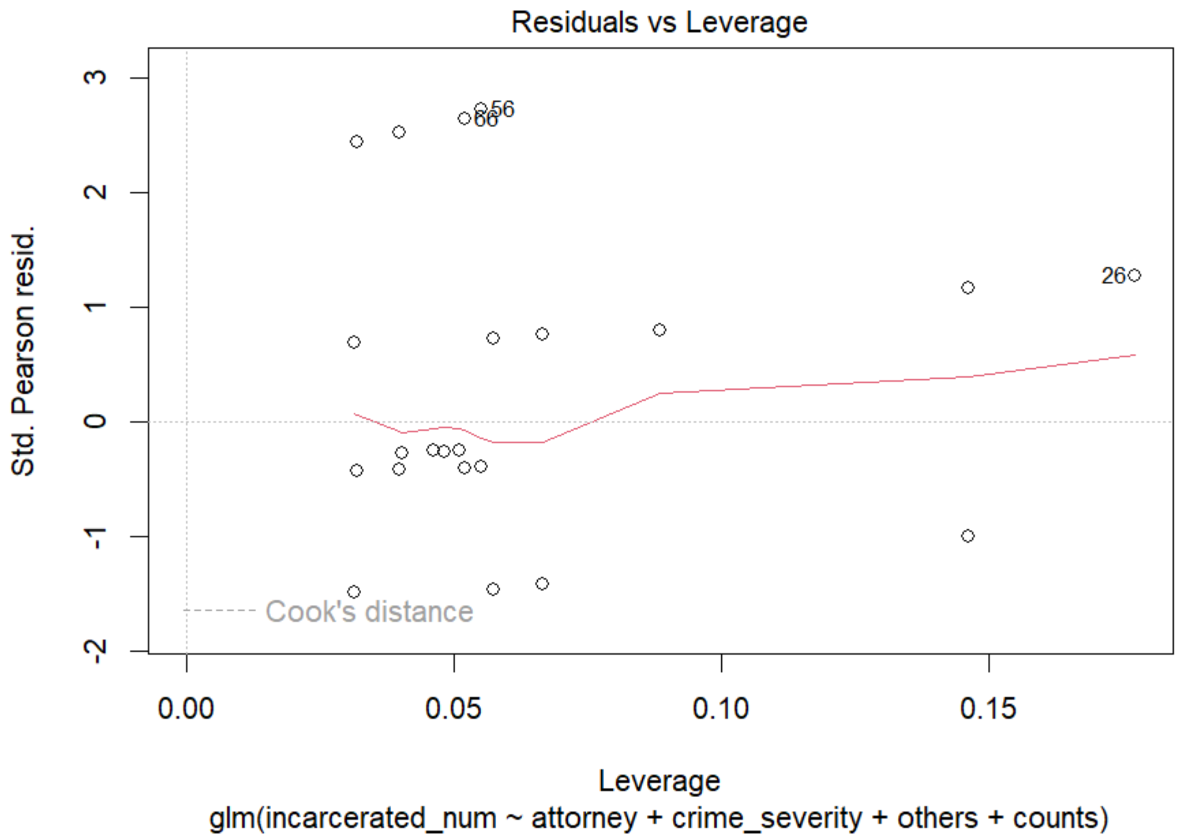

Figure 6: Residuals Versus Leverage for the Logistic Regression Model

Figure 6 shows residuals versus leverage for the logistic regression model. A few observations had higher leverage or larger residuals, but the plot did not indicate a single observation that clearly invalidated the model.

2.4.8 Categorical Association Tests

Several chi-square tests were used to examine associations among attorney type, sentence outcome, incarceration status, plea type, and crime severity.

Table 10: Detailed Chi-Square Test Results

Test	N	χ^2	df	p-value	Cramer's V	Conclusion
Attorney type by sentencing outcome	100	32.411	2	$< .001$	0.569	Strong association between attorney type and sentencing outcome.
Attorney type by incarceration status	100	26.747	1	$< .001$	0.517	Incarceration status differed substantially by attorney type.
Attorney type by plea type	100	6.466	1	0.011	0.254	Plea type was associated with attorney type.
Attorney type by crime severity	100	0.012	1	0.913	0.011	No evidence that violent/nonviolent case mix differed by attorney type.

Note: The first test used sentencing outcomes only: incarceration, suspended sentence, and deferred sentence. Dismissed cases were excluded from sentencing-specific analyses. Yates' continuity correction was applied by R for the 2×2 chi-square tests.

The strongest categorical result was the association between attorney type and incarceration status. Among sentenced cases, appointed-counsel cases had 33 incarceration outcomes and 18 non-incarceration outcomes, while private-counsel cases had 6 incarceration outcomes and 43 non-incarceration outcomes. This suggests that incarceration was much more common among appointed-counsel cases in this sample.

2.4.9 Sentence-Length Hypothesis Test

Sentence length was analyzed only among cases that resulted in incarceration. Because imposed sentence length was highly skewed, a Shapiro-Wilk normality check was conducted by attorney group before comparing sentence lengths.

Table 11: Normality Check for Imposed Sentence Length

Attorney Type	N	Shapiro-Wilk p-value
Appointed	33	0.0003
Private	6	0.1886

The appointed-counsel group failed the normality check, so the Wilcoxon rank-sum test was used as the primary comparison.

Table 12: Wilcoxon Rank-Sum Test for Imposed Sentence Length

Outcome	**Test Statistic**	**p-value**	**Conclusion**
Time sentenced in days	$W = 86$	0.625	No detectable difference by attorney type.

Among incarceration cases, there was no statistically significant difference in imposed sentence length by attorney type. This suggests that the stronger difference in the data concerns whether incarceration occurred, not the length of the imposed incarceration sentence once incarceration was ordered.

2.4.10 Logistic Regression Model

A logistic regression model was fit to predict incarceration among sentenced cases. The outcome variable was binary, coded as incarceration versus suspended or deferred sentencing. The predictors were attorney type, crime severity, other Oklahoma criminal records, and number of counts.

Table 13: Logistic Regression Predicting Incarceration

Predictor	**Log Odds**	**Odds Ratio**	**Lower 95% CI**	**Upper 95% CI**	**p-value**
Intercept	-2.723	0.066	0.011	0.394	0.003
Appointed attorney	2.507	12.263	4.319	34.814	$< .001$
Violent offense	-0.138	0.871	0.288	2.635	0.806
Other criminal records	0.912	2.489	0.547	11.325	0.238
More than one count	0.064	1.066	0.374	3.037	0.905

The model showed that attorney type was strongly associated with incarceration status. Holding crime severity, other Oklahoma criminal records, and number of counts constant, appointed-counsel cases had estimated odds of incarceration approximately 12.26 times the odds for private-counsel cases. Crime severity, other Oklahoma criminal records, and number of counts were not statistically significant in this model.

The likelihood-ratio test comparing the fitted model to an intercept-only model was statistically significant, $\chi^2(4) = 32.671$, $p = 1.395 \times 10^{-6}$. The AIC decreased from 135.75 for the null model to 111.08 for the fitted model. The McFadden pseudo-R^2 was 0.244.

2.4.11 Predicted Probability of Incarceration

Predicted probabilities were calculated from the logistic regression model to make the model results more interpretable. The comparison below holds crime severity fixed as nonviolent, other Oklahoma criminal records fixed as yes, and number of counts fixed as more than one.

Table 14: Predicted Probability of Incarceration by Attorney Type

Attorney Type	Log Odds	Predicted Probability	Lower 95%	Upper 95%
Private	-1.747	0.148	0.061	0.318
Appointed	0.760	0.681	0.503	0.818

For this model-based comparison, the predicted probability of incarceration was approximately 14.8% for private-counsel cases and 68.1% for appointed-counsel cases. This is a large estimated difference, but it should still be interpreted as an association rather than a causal effect.

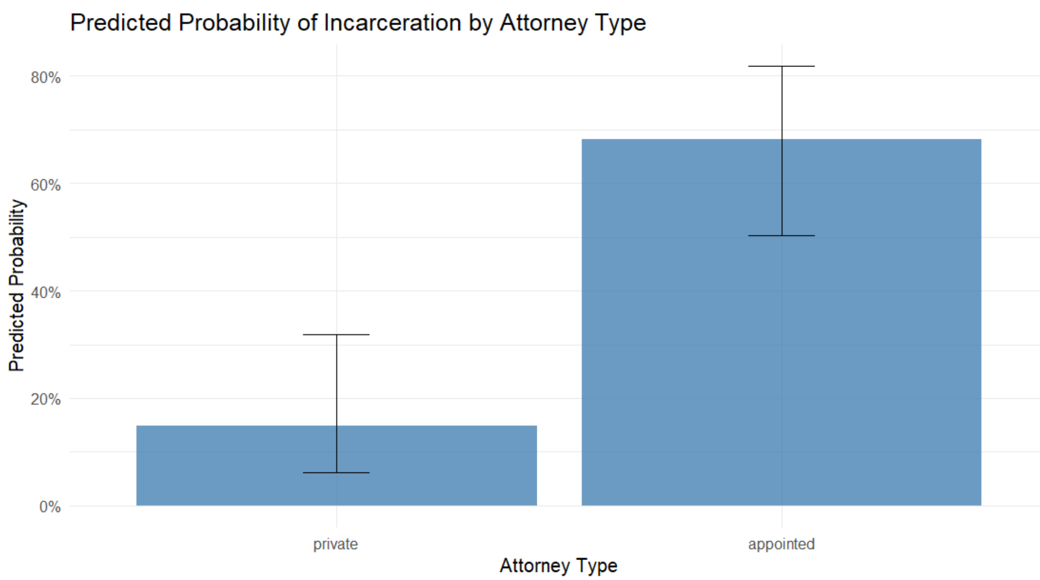

Figure 7: Predicted Probability of Incarceration by Attorney Type

Figure 7 displays the model-based predicted probability of incarceration by attorney type, holding crime severity, other Oklahoma criminal records, and number of counts fixed. The model estimated a substantially higher predicted probability of incarceration for appointed-counsel cases than for private-counsel cases.

2.4.12 Exploratory Random Forest Feature Importance

A random forest model was fit as an exploratory predictive model for incarceration outcome. The purpose of this model was not causal inference, but to examine which recorded variables were most important for predicting incarceration in this sample.

Table 15: Random Forest Feature Importance

Predictor	Permutation Importance
Attorney type	0.124
Crime severity	0.007
Counts	0.004
Other criminal records	0.003

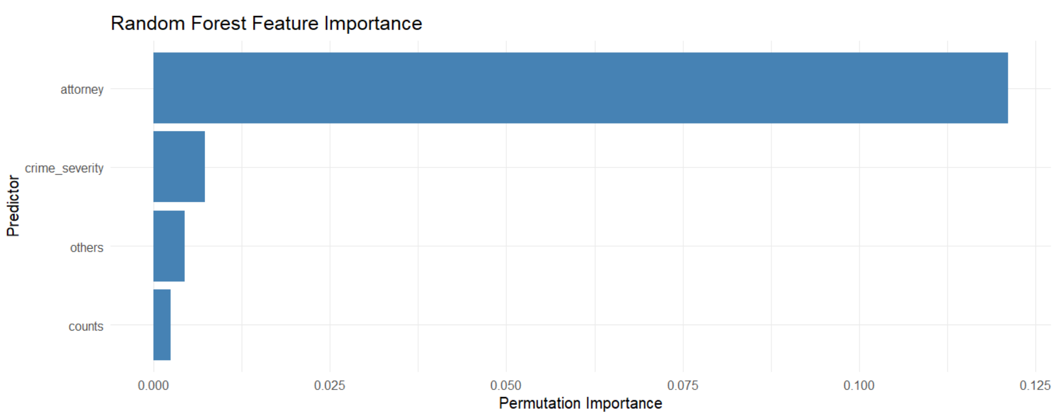

Figure 8: Random Forest Feature Importance

Figure 8 shows that attorney type had the largest permutation importance in the exploratory random forest model.

The exploratory random forest results were consistent with the logistic regression model. Attorney type had the largest permutation importance value, while the other recorded predictors contributed comparatively little to prediction in this model. Because the sample size is limited, these results should be interpreted as supplemental predictive evidence.

2.5 Discussion

2.5.1 Summary of Key Findings

The main finding is that attorney type was strongly associated with incarceration outcome in this sample. Among sentenced cases, appointed-counsel cases were more likely to result in incarceration, while private-counsel cases were more likely to result in suspended or deferred outcomes.

The chi-square tests showed statistically significant associations between attorney type and sentencing outcome, attorney type and incarceration status, and attorney type and plea type. The balance check for crime severity was not statistically significant, suggesting that the violent/nonviolent classification did not differ by attorney type in this sample.

The logistic regression model also identified attorney type as the strongest predictor of incarceration after accounting for crime severity, other Oklahoma criminal records, and number of counts. However, among cases that resulted in incarceration, imposed sentence length did not differ by attorney type.

2.5.2 Limitations

This study should be interpreted as an exploratory, proof-of-concept analysis. The data come from one county and one year, so the results should not be generalized beyond this sample without additional study.

Because the study uses observational public-record data, the results show associations rather than causation. Attorney type may be related to other case characteristics not fully captured in the dataset.

Some variables also have limits. The other-criminal-records variable indicates whether other Oklahoma records were found, but it does not establish timing. Crime severity was coded as violent or nonviolent, which simplifies more complex legal differences.

Despite these limitations, the analysis demonstrates that public court records can be systematically collected, coded, and analyzed to identify sentencing patterns worthy of further study.

2.5.3 Recommendations for Future Study

The results of this project provide enough evidence to justify a broader follow-up study. The patterns observed here suggest that representation type and sentencing outcomes can be studied systematically using court-record data.

A larger funded study could expand the data collection process across multiple counties, years, judges, attorneys, and offense categories. This would allow the use of hierarchical models to account for court-level, judge-level, county-level, and case-level variation.

Additional funding would also support more complete access to structured legal data, reducing the burden of manual coding and allowing more detailed variables to be included, such as charge type, plea agreement details, bond amount, attorney changes, and timing of other criminal records.

This project therefore serves as a foundation for a more comprehensive study of sentencing outcomes and representation type using advanced statistical modeling.

2.6 Author's Note

In transitioning from the internal report to the client-facing report, I shifted the focus toward the main highlights of the analysis using plain-language explanations and less statistical jargon. I removed detailed discussion of specific tests, assumptions, and diagnostic statistics so the client-facing report could focus on what the results mean rather than the mechanics of how each result was produced.

The visual style did not require major revision because I generally try to make visuals accessible from the beginning. However, I excluded diagnostic visuals from the client-facing report because those figures are more useful for internal statistical review than for a nontechnical audience. The client-facing visuals focus on the main comparisons and patterns relevant to the client.

I continued to emphasize that this study is observational and exploratory, so no causal claim can be made from these results. For example, concluding that appointed attorneys cause worse outcomes would be similar to concluding that doctors working late-night emergency room shifts cause higher mortality simply because they see more serious cases. The pattern detected in this study does not prove causation, but it does warrant further consideration.

3 Video Presentation Link

The video presentation for this capstone project is available at the following link:

Capstone Video Presentation

4 Appendix: R Code

```
library(dplyr)

library(tidyr)

library(ggplot2)
```

```r
library(knitr)
library(scales)
library(ranger)
library(car)

set.seed(123)


# import and clean
################################################

df_raw <- read.csv("reduced_capstone4_26.csv", header = TRUE)

names(df_raw) <- names(df_raw) %>%
  tolower() %>%
  gsub("[^a-z0-9]+", "_", .) %>%
  gsub("_$", "", .)

df <- df_raw %>%
  mutate(across(where(is.character), ~ trimws(tolower(.x)))) %>%
  mutate(across(where(is.character), ~ na_if(.x, "na"))) %>%
  mutate(
    sentence = na_if(sentence, ""),
    attorney = na_if(attorney, ""),
    plea = na_if(plea, ""),
    crime_severity = na_if(crime_severity, ""),
    others = na_if(others, ""),
    counts = na_if(counts, ""),
    counts = case_when(
```

```
      counts %in% c("more", "more␣than␣one", "multiple") ~ "more",
      counts %in% c("one", "1") ~ "one",
      counts %in% c("none", "none␣dismissed", "none␣(dismissed)") ~
         "none",
      TRUE ~ counts
    ),
    time_sentenced_yr = as.numeric(time_sentenced_in_yr),
    time_sentenced_day = time_sentenced_yr * 365
  )



# analysis datasets
##########################################################

df_attorney <- df %>%
  filter(attorney %in% c("appointed", "private"))

df_reduced <- df_attorney %>%
  filter(!sentence %in% c("amended", "warrant", "deceased", "blank")
     )

df_sentenced <- df_reduced %>%
  filter(sentence %in% c("incarceration", "suspended", "deferred"))
     %>%
  mutate(
    incarcerated = ifelse(sentence == "incarceration",
                          "incarcerated",
                          "not_incarcerated"),
    incarcerated_num = ifelse(sentence == "incarceration", 1, 0)
```

```
  )

df_incarcerated <- df_sentenced %>%
  filter(sentence == "incarceration",
         !is.na(time_sentenced_day),
         time_sentenced_day > 0)

dataset_sizes <- tibble(
  dataset = c("full cleaned data", "attorney eligible", "reduced 
     outcome data",
              "sentenced data", "incarcerated only"),
  n = c(nrow(df), nrow(df_attorney), nrow(df_reduced),
        nrow(df_sentenced), nrow(df_incarcerated))
)

kable(dataset_sizes, caption = "analysis dataset sizes")



# summary tables
##############################################################

summary_table <- bind_rows(
  df %>%
    count(sentence, name = "count") %>%
    mutate(section = "sentence / outcome", category = sentence),

  df_attorney %>%
    count(attorney, name = "count") %>%
    mutate(section = "attorney type", category = attorney),
```

```
  df %>%
    count(counts, name = "count") %>%
    mutate(section = "number␣of␣counts", category = counts),

  df %>%
    filter(!is.na(others)) %>%
    count(others, name = "count") %>%
    mutate(section = "other␣criminal␣cases", category = others),

  df %>%
    filter(crime_severity %in% c("violent", "nonviolent")) %>%
    count(crime_severity, name = "count") %>%
    mutate(section = "crime␣severity", category = crime_severity)
) %>%
  select(section, category, count) %>%
  group_by(section) %>%
  mutate(percent = round(100 * count / sum(count), 2)) %>%
  ungroup()

kable(summary_table, caption = "summary␣of␣key␣case␣characteristics"
   )

sentence_length_summary <- df_incarcerated %>%
  group_by(attorney) %>%
  summarise(
    n = n(),
    mean = round(mean(time_sentenced_day), 0),
    sd = round(sd(time_sentenced_day), 0),
```

```
    median = round(median(time_sentenced_day), 0),
    iqr = round(IQR(time_sentenced_day), 0),
    min = round(min(time_sentenced_day), 0),
    max = round(max(time_sentenced_day), 0),
    .groups = "drop"
  )

kable(sentence_length_summary,
      caption = "sentence␣length␣among␣incarcerated␣cases")

# visualizations
##########################################################

ggplot(df_reduced, aes(x = attorney, fill = sentence)) +
  geom_bar(position = position_dodge(width = 0.9),
           color = "black", alpha = 0.85) +
  geom_text(
    stat = "count",
    aes(label = after_stat(count)),
    position = position_dodge(width = 0.9),
    vjust = -0.3,
    size = 4
  ) +
  labs(
    title = "Sentence␣Outcome␣by␣Attorney␣Type",
    x = "Attorney␣Type",
    y = "count",
    fill = "sentence"
  ) +
```

```
  theme_minimal()

ggplot(df_sentenced, aes(x = attorney, fill = incarcerated)) +
  geom_bar(position = "fill", color = "black", alpha = 0.85) +
  scale_y_continuous(labels = percent_format()) +
  labs(
    title = "Incarceration Status by Attorney Type",
    x = "Attorney Type",
    y = "Percent",
    fill = "outcome"
  ) +
  theme_minimal()

ggplot(df_sentenced %>% filter(!is.na(plea), plea != "none"),
       aes(x = attorney, fill = plea)) +
  geom_bar(position = position_dodge(width = 0.9),
           color = "black", alpha = 0.85) +
  geom_text(
    stat = "count",
    aes(label = after_stat(count)),
    position = position_dodge(width = 0.9),
    vjust = -0.3,
    size = 4
  ) +
  labs(
    title = "Plea Type by Attorney Type",
    x = "Attorney Type",
    y = "Count",
    fill = "plea type"
```

```
  ) +
  theme_minimal()

ggplot(df_incarcerated, aes(x = attorney, y = time_sentenced_day)) +
  geom_boxplot(fill = "steelblue", alpha = 0.75) +
  geom_jitter(width = 0.12, alpha = 0.55, size = 2) +
  scale_y_log10() +
  labs(
    title = "Sentence Length Among Incarcerated Cases",
    x = "Attorney Type",
    y = "Sentence Length in Days (Log Scale)"
  ) +
  theme_minimal()


# chi-square tests
##########################################################

run_cat_test <- function(data, x, y) {
  tab <- table(data[[x]], data[[y]])
  chi <- suppressWarnings(chisq.test(tab))
  list(table = tab, expected = round(chi$expected, 2), test = chi)
}

cramers_v <- function(tab) {
  chi <- chisq.test(tab, correct = FALSE)
  n <- sum(tab)
  r <- nrow(tab)
  c <- ncol(tab)
```

```r
  sqrt(as.numeric(chi$statistic) / (n * min(r - 1, c - 1)))
}

sentence_test <- run_cat_test(df_sentenced, "attorney", "sentence")
incarceration_test <- run_cat_test(df_sentenced, "attorney", "incarcerated")

plea_test <- df_sentenced %>%
  filter(!is.na(plea), plea != "none") %>%
  run_cat_test("attorney", "plea")

severity_test <- df_attorney %>%
  filter(crime_severity %in% c("violent", "nonviolent")) %>%
  run_cat_test("attorney", "crime_severity")

chisq_results <- tibble(
  test = c("attorney by sentencing outcome",
           "attorney by incarceration status",
           "attorney by plea type",
           "attorney by crime severity"),
  n = c(sum(sentence_test$table),
        sum(incarceration_test$table),
        sum(plea_test$table),
        sum(severity_test$table)),
  chi_square = round(c(sentence_test$test$statistic,
                       incarceration_test$test$statistic,
                       plea_test$test$statistic,
                       severity_test$test$statistic), 3),
  df = c(sentence_test$test$parameter,
```

```
        incarceration_test$test$parameter,
        plea_test$test$parameter,
        severity_test$test$parameter),
  p_value = c(sentence_test$test$p.value,
              incarceration_test$test$p.value,
              plea_test$test$p.value,
              severity_test$test$p.value),
  cramers_v = round(c(cramers_v(sentence_test$table),
                      cramers_v(incarceration_test$table),
                      cramers_v(plea_test$table),
                      cramers_v(severity_test$table)), 3)
)

kable(chisq_results,
      caption = "chi-square test results")

expected_counts <- tibble(
  test = chisq_results$test,
  min_expected = c(min(sentence_test$expected),
                   min(incarceration_test$expected),
                   min(plea_test$expected),
                   min(severity_test$expected)),
  max_expected = c(max(sentence_test$expected),
                   max(incarceration_test$expected),
                   max(plea_test$expected),
                   max(severity_test$expected)),
  cells_below_5 = c(sum(sentence_test$expected < 5),
                    sum(incarceration_test$expected < 5),
                    sum(plea_test$expected < 5),
```

```
                    sum(severity_test$expected < 5))
)

kable(expected_counts,
      caption = "expected cell count checks")

# sentence length test
#########################################################

normality_table <- df_incarcerated %>%
  group_by(attorney) %>%
  summarise(
    n = n(),
    shapiro_p = round(shapiro.test(time_sentenced_day)$p.value, 4),
    .groups = "drop"
  )

kable(normality_table,
      caption = "normality check for sentence length")

wilcox_sentence <- wilcox.test(
  time_sentenced_day ~ attorney,
  data = df_incarcerated,
  exact = FALSE
)

sentence_length_test <- tibble(
  test = "wilcoxon rank-sum test",
  statistic = wilcox_sentence$statistic,
```

```r
  p_value = round(wilcox_sentence$p.value, 3)
)

kable(sentence_length_test,
      caption = "sentence length test among incarcerated cases")



# logistic regression
##########################################################

df_model <- df_sentenced %>%
  filter(
    attorney %in% c("appointed", "private"),
    crime_severity %in% c("violent", "nonviolent"),
    others %in% c("yes", "no"),
    counts %in% c("one", "more")
  ) %>%
  mutate(
    attorney = relevel(factor(attorney), ref = "private"),
    crime_severity = relevel(factor(crime_severity), ref = "
       nonviolent"),
    others = relevel(factor(others), ref = "no"),
    counts = relevel(factor(counts), ref = "one")
  ) %>%
  droplevels()

log_model <- glm(
  incarcerated_num ~ attorney + crime_severity + others + counts,
  data = df_model,
```

```
  family = binomial
)

logit_table <- data.frame(
  predictor = names(coef(log_model)),
  log_odds = coef(log_model),
  odds_ratio = exp(coef(log_model)),
  lower_95_ci = exp(confint.default(log_model)[, 1]),
  upper_95_ci = exp(confint.default(log_model)[, 2]),
  p_value = coef(summary(log_model))[, 4]
) %>%
  mutate(across(where(is.numeric), ~ round(.x, 3)))

kable(logit_table,
      caption = "logistic regression predicting incarceration")

null_model <- glm(incarcerated_num ~ 1, data = df_model, family =
   binomial)

model_fit <- tibble(
  statistic = c("likelihood ratio p-value", "null aic", "model aic",
                "mcfadden pseudo r-squared"),
  value = c(
    anova(null_model, log_model, test = "Chisq")$`Pr(>Chi)`[2],
    AIC(null_model),
    AIC(log_model),
    1 - as.numeric(logLik(log_model) / logLik(null_model))
  )
)
```

```
kable(model_fit,
      caption = "logistic␣regression␣model␣fit")

vif_table <- tibble(
  predictor = names(car::vif(log_model)),
  vif = round(as.numeric(car::vif(log_model)), 3)
)

kable(vif_table,
      caption = "variance␣inflation␣factors")

# predicted probabilities
##############################################

pred_data <- data.frame(
  attorney = factor(c("private", "appointed"), levels = levels(df_
     model$attorney)),
  crime_severity = factor("nonviolent", levels = levels(df_model$
     crime_severity)),
  others = factor("yes", levels = levels(df_model$others)),
  counts = factor("more", levels = levels(df_model$counts))
)

pred_link <- predict(log_model, newdata = pred_data,
                     type = "link", se.fit = TRUE)

pred_table <- pred_data %>%
  mutate(
```

```
    log_odds = pred_link$fit,
    predicted_probability = plogis(pred_link$fit),
    lower_95 = plogis(pred_link$fit - 1.96 * pred_link$se.fit),
    upper_95 = plogis(pred_link$fit + 1.96 * pred_link$se.fit)
  ) %>%
  mutate(across(where(is.numeric), ~ round(.x, 3)))

kable(pred_table,
      caption = "predicted probability of incarceration")

ggplot(pred_table, aes(x = attorney, y = predicted_probability)) +
  geom_col(fill = "steelblue", alpha = 0.8) +
  geom_errorbar(aes(ymin = lower_95, ymax = upper_95), width = 0.15)
      +
  scale_y_continuous(labels = percent_format()) +
  labs(
    title = "Predicted Probability of Incarceration by Attorney Type
       ",
    x = "Attorney Type",
    y = "Predicted Probability"
  ) +
  theme_minimal()


# diagnostics
###########################################

logit_cells <- bind_rows(
```

```
  as.data.frame.matrix(table(df_model$attorney, df_model$incarcerated)) %>%
    tibble::rownames_to_column("category") %>%
    mutate(predictor = "attorney type"),

  as.data.frame.matrix(table(df_model$crime_severity, df_model$incarcerated)) %>%
    tibble::rownames_to_column("category") %>%
    mutate(predictor = "crime severity"),

  as.data.frame.matrix(table(df_model$others, df_model$incarcerated)) %>%
    tibble::rownames_to_column("category") %>%
    mutate(predictor = "other records"),

  as.data.frame.matrix(table(df_model$counts, df_model$incarcerated)) %>%
    tibble::rownames_to_column("category") %>%
    mutate(predictor = "counts")
)

kable(logit_cells,
      caption = "cell counts for logistic regression")

plot(log_model, which = 4)
plot(log_model, which = 5)



# random forest
```

```
##################################################

rf_data <- df_model %>%
  mutate(
    incarcerated = factor(
      ifelse(incarcerated_num == 1,
             "incarcerated",
             "not_incarcerated")
    )
  ) %>%
  select(incarcerated, attorney, crime_severity, others, counts)

rf_model <- ranger(
  incarcerated ~ .,
  data = rf_data,
  classification = TRUE,
  probability = TRUE,
  num.trees = 1000,
  mtry = 2,
  min.node.size = 5,
  importance = "permutation"
)

importance_df <- data.frame(
  predictor = names(rf_model$variable.importance),
  importance = as.numeric(rf_model$variable.importance)
) %>%
  arrange(desc(importance))
```

```r
kable(importance_df,
      caption = "random forest feature importance")


ggplot(importance_df,
       aes(x = reorder(predictor, importance), y = importance)) +
  geom_col(fill = "steelblue") +
  coord_flip() +
  labs(
    title = "Random Forest Feature Importance",
    x = "Predictor",
    y = "Permutation Importance"
  ) +
  theme_minimal()
```