Southern Methodist University

# GIDEON Project

Deason Criminal Justice Reform Center

Department of Statistics and Data Science

Léon Yuan & Hunter Schuler

Fall 2023 – Stat 6366

# Contents

# Executive Summary

Gauging Improvement in Defense Efforts and Outcomes in New York (GIDEON) investigates the impact of New York State's increased funding of public defense. The study explores whether reductions in indigent defense caseloads produce higher quality lawyering and considers how caseload reductions impact clients and their communities. This project is a first look at case data collected from one particular New York county and analyzes how interactions between defendants and attorneys changed from 2010 to 2022. The goals for this project included joining the various raw data sources for future analysis, analyzing the frequency of specific event types, and assessing how event type frequency has changed over time among misdemeanor and felony cases.

The source data for this project consists of three separate data sheets: a case query, an event query, and a name query. These data sheets were combined into one by the shared primary keys associated with each case. Analysis of the event codes in the data revealed that a large majority (78.6%) of the data is made up of just four event types. Various event types were plotted over time to look for meaningful changes in frequency. An R-Shiny application was also created to allow for additional exploration of the data. Time series graphs were used to analyze the change in event frequencies for felony and misdemeanor case types, using two different approaches for defining event frequency. For one frequency formula, most event types were relatively stable throughout the analysis period, fluctuating only mildly on an approximately monthly cycle. Only the Client Contact (CLCT) event type experienced noticeable change, increasing in April 2020 and fluctuating on similar monthly cycle thereafter. For the second frequency formula used, nearly all event types were stable across the analysis period, fluctuating mildly (with only a few isolated exceptions) over an approximately 2-week cycle. Analysis was also conducted on the duration of cases, which found that the frequency of events per-case-per-week did not exhibit many changes at all.

Overall, based on our analysis, the current data does not provide evidence of any significant changes in attorney-client interactions. Continued collection of data is recommended due to the possibility that systemic reforms may have a lagging effect on attorney-client interactions which may not yet be detectable in this data set.

## Introduction

The GIDEON Project investigates the impact of New York State's injection of $250 million in new funding to criminal defense systems. Long-term project goals include assessing whether reductions in public defender caseloads produce higher quality defense services, as well as other potential changes to defendant-attorney relationships. The project seeks to track how this reform changes interaction patterns between clients and their attorneys. In our role as consultants, we have limited prior knowledge in this domain and therefore cleaned and analyzed the data with little prior knowledge of the subject. When necessary, we consulted with the clients on domain-specific factual questions that affected the analysis (e.g., grouping certain case types as either felony or misdemeanor). The client goals for this project include linking each of the separate data sheets into one, uncovering frequency changes for event types of interest over time, and separately assessing the frequency changes of the various event types in felony and misdemeanor cases. To achieve the project goals, we will combine multiple datasets obtained from public defender case files. The integration of these separate data sheets will ensure a unified and cohesive analysis. Our methodology involves employing statistical models and data visualization tools to discern patterns and trends in event frequencies over time, with a specific focus on distinguishing changes in felony and misdemeanor cases.

## Statistical Methods

The R programming language and RStudio software were used for all analyses. The data consists of three data sheets of criminal case data from a particular county in New York from approximately 2015 to 2022 as follows Figure SM1.

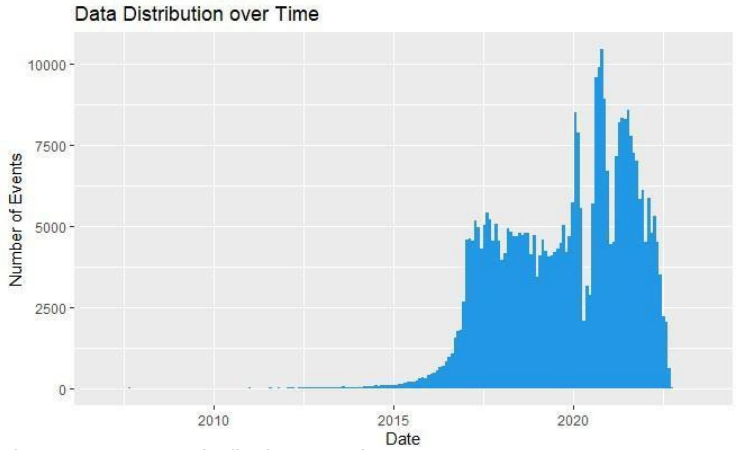



*Figure SM1 – Data Distribution over Time*

This data is limited to cases that were already closed when the data was queried. This restriction to closed-case data is important to the analysis in ways that will be discussed below. The Case Table Query is 81,256 rows across 144 variables and gives information about each case such as the offense charged, the name of the defendant, the name of the judge, etc. The Event Query Table is 389,030 rows across 31 variables and gives information about each particular event in every case, such as when the attorney met with the client, arraignment, court dates, etc. The Name Search Query is 45,139 rows across 53 variables and contains information about the defendants that appear in the Case Table Query. We used the "case file number" in Case Table Query to full join[1] the "event case number" in Event Query Table with the dplyrpackage in R. We then used "case client ID" in Case Table Query to fully join "name ID" in Name Search Query with the tidyversepackage in R. Finally, we combined these three datasets into one complete dataset with a dimensionality of 392,625 rows across 213 columns. Joining

[1] For an explanation on the full join function, see: https://dplyr.tidyverse.org/reference/mutate-joins.html

these separate datasets provides us with a complete data frame that better facilitates downstream analyses.

We combined the analysis for the second and third client goals. Before analyzing, we cleaned the linked dataset. First, we converted the event date string, open date string, and closed date string into a date object, and filtered out all cases before 2010 (less than 0.001% of the data). This filtering was done to exclude plainly erroneous data (e.g. years such as "1811"). Second, we reduced the data set to the event types of interest (Table A1) and grouped the data into two case types: felony and misdemeanor. We found that there were multiple felony categories used in the data, so we grouped all of the Violent Felony Offense (VFO) and Nonviolent Felony Offense (NVFO) events into the felony category as well. After these cleaning steps, we ended up with 94,971 felony cases and 251,164 misdemeanor cases.

Our analysis focused on using ggplot[2] visualizations to answer questions about how various aspects of the data changed over time:

- How do the event types of interest vary between felony and misdemeanor case types?
    - To get an overall sense of quantity, we created a bar plot to observe the total count of each event type grouped by felony and misdemeanor case type.
- How do the case types change over time?
    - We created a line plot to visualize the number of cases available each month for both felony and misdemeanor case types.
- Does case duration vary over time?

[2] For more information on the ggplotpackage and its various plotting functions, see: https://ggplot2.tidyverse.org/

    - We created a line plot to observe how the median duration for both felony and misdemeanor case types changed over time.
- How do the event types fluctuate over time among felony and misdemeanor cases?
    - We created a line plot to graph how the quantity of each event type changed each week for both felony and misdemeanor case types.

Focusing on visualizing the answers to these initial questions about the data improves accessibility to potential underlying trends in these features.

To analyze event frequency, we defined the frequency of events in two distinctive ways. The first definition is to calculate the number of each event type by the number of cases closed each week. This way is straightforward to understand and interpret but comes with certain limitations. For example, not all cases involve every event code and cases with extremely high numbers of specific event types skew such simple averages, potentially leading to a misrepresentation of the actual frequency. Moreover, this method does not account for the duration of cases which is an essential factor as, generally, more prolonged cases have more events. To address these limitations, we introduced a second way of measuring frequency: event rate. We calculated the medians of all event rates, each of which is determined by dividing the number of events in each case by that case's duration in weeks. This event rate measures how many events occur per-case-per-week. This method is not influenced by extreme numbers of events in one case and is more robust to variations in case durations, providing a balanced and realistic perspective.

After the data visualization analyses, time series models were to analyze frequency changes over time. Modeling can be a useful tool in this context to quantitatively analyze and interpret frequency changes over time. It allows us to assess the fit of various time series models which may provide insights into the patterns and dynamics of events in felony and misdemeanor cases. Such models may offer

quantitative assessments of how certain types of events vary weekly. Various models employing different model architectures were built using two types of time series. The first category of models looked at event type median count per week as the outcome variable. The second category of models used event rate (without respect to event type) each week as the outcome variable. Because the goal for these models is retrospective interpretations (instead of forecasting in time series), we mainly focused on how good our models fit to the data. A Poisson Regression model (Formula 1) was built which can incorporate multiple predictors while also taking the discreet nature of the outcome variable into consideration. The most important predictor included in the models was the durations of cases because, generally, longer cases have more events. For the second outcome, median of event rates, we built a grouped cross time series model with exponential trend smoothing (ETS[3]) as engines (Formula 2). The advantage of such a model is that time series for both felony and misdemeanor case types share similar trends across both time and event type. That is, this method can pool information across individual event types' time series and potentially reveal some hidden relationships between them (Diagram 1). Moreover, this method models all levels (Diagram SM1) of the time series at once to save time, as opposed to modeling them one by one. Additional models using SARIMA[4] and simple ETS were considered, but these methods proved to be relatively unfruitful and were therefore excluded.


*Diagram SM1 – Modeling Levels*

[3] For a detailed explanation of ETS models, see: https://otexts.com/fpp3/ets.html
[4] For a detailed explanation of SARIMA models, see: https://otexts.com/fpp3/seasonal-arima.html

## Modeling Formulas

$$log(\lambda_i) = \beta_0 + \beta_1 * week_i + \beta_2 * \text{meadian\_diff}_i + \varepsilon_i$$

*Formula SM1 – Poisson Regression*

Where $\lambda_i$ is the event rate (number of events per case per week), all beta are coefficients, week is the time index of week number, median diff is the median of case durations of cases closed at week index, and $\varepsilon_i$ is the residuals of models.

$$Rate_t = Rate_{Felony,t} + Rate_{Mis,t}$$

*Formula SM2 – Grouped Cross Time Series Model*

$$Rate_{Felony,t} = Rate_{Felony,CLCT,t} + \cdots + Rate_{Felony,Other,t}$$
$$Rate_{Mis,t} = Rate_{Mis,CLCT,t} + \cdots + Rate_{Mis,Other,t}$$

Where Rate is the event rate for each event type (using the second definition discussed previously).

# Results

## Data Visualization

Our first bar plot Figure R1 shows all event types and is colored by case type. We found that client contact is the most common event type for both felony and misdemeanor case types. There are 68,446 client contact events for misdemeanors and 18,816 client contact events for felonies. Client meetings, investigator retained, interpreter retained, and pretrial hearing event types represent less than 0.04% of this data.

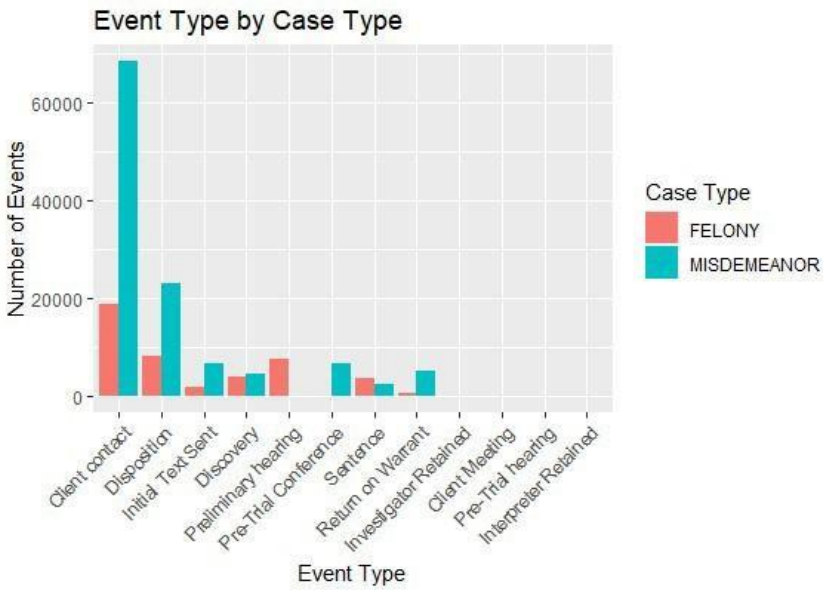



*Figure R1 – Event Type by Case Type*

Our Figure R2 for the number of felony and misdemeanor cases shows that generally, there are more misdemeanor cases than felony cases, with the exception of April, May, and June in 2020. Before 2020, there were approximately 900 misdemeanor cases each month and around 300 felony cases each month, and their time series lines appeared to remain stationary around these two levels. However, after 2020, the quantity of both case types decreased noticeably. This decrease is almost certainly an artifact resulting from the data being limited to cases that were already closed when the data was collected.

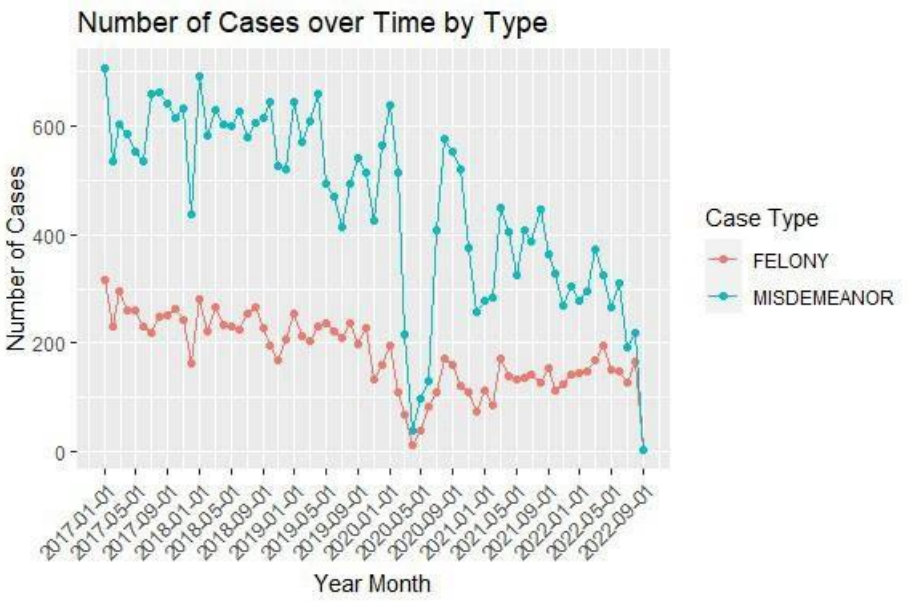



*Figure R2 – Number of Cases over Time by Type*

Our Figure R3 displays the case duration for felony and misdemeanor case types. We didn't find any apparent trend or something interesting in the graph below. However, we found felony cases have zero durations in 2020 on average. We assume this is because of the impacts of the COVID-19 pandemic. Overall, the median duration of felony cases is 7.1 weeks and the median duration of misdemeanor cases is 8 weeks.

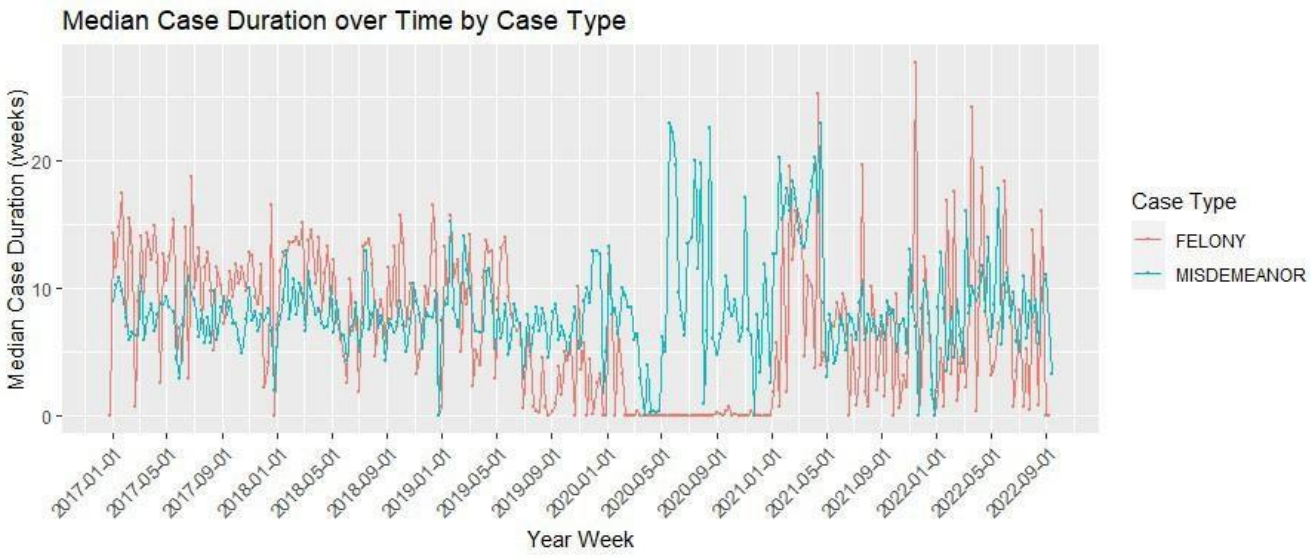



*Figure R3 – Median Case Duration over Time by Case Type*

The Figure R4 shows how the frequency of each event type changes each week for both felony and misdemeanor case types. The frequency used here is the first definition of frequency that was outlined in the methods section above. We find that beginning from May 2019, the frequency of client contacts for felony cases slowly increases from 1 client contact per felony case to 3.25 client contacts per felony case. We also find that beginning in February 2018, the frequency of client contacts for misdemeanor cases slowly increases from 1 client contact per misdemeanor case to 5 client contacts per misdemeanor case. All other event codes were stationary across time. On average, there are 0.5 events per felony case and 0.15 events per misdemeanor case.

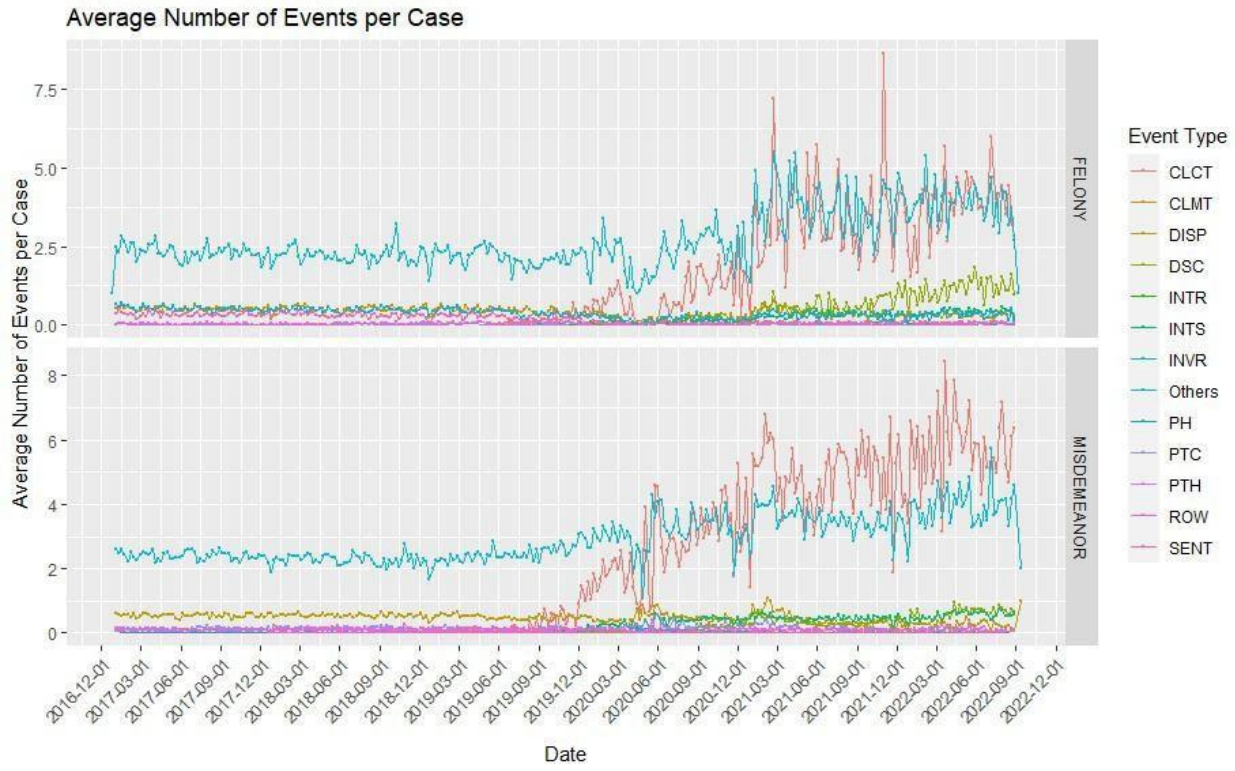



*Figure R4 – Average Number of Events*

The Figure R5 shows how the median event rate for both case types changes each week. This analysis uses the second definition of event rate that was discussed in the method section above. We find that there is no major or clear trend for event types or case types. There are only a few sudden

bumps up for some event types, but they all returned to their previous baseline level. We observe that once the durations of cases are considered, there are no notably important changes in event rate.

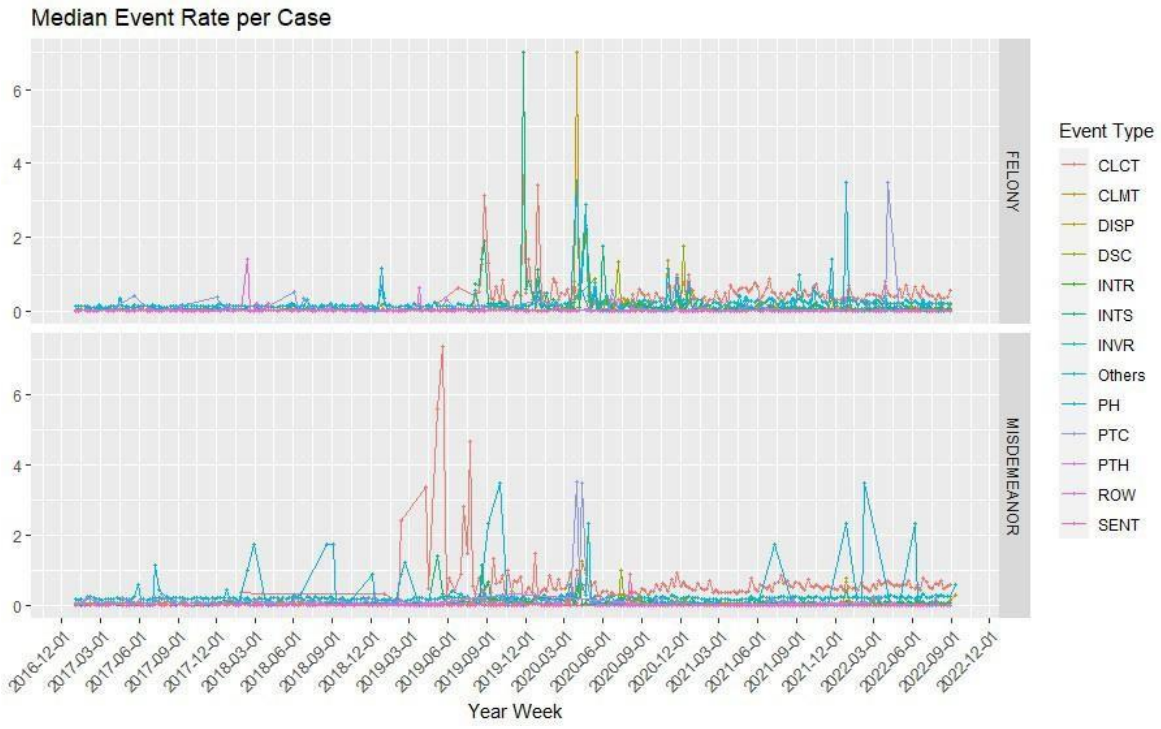



*Figure R5 – Median Event Rate*

The Figure R6 looks at how attorney-specific case loads may have changed over time. This graph first groups the data by the case closed date (rounded to the nearest week) to avoid a false decrease in count over time associated with the data only containing closed cases. Next, the data is grouped by attorney and a sum of all the events in each case is calculated. The number of events is then divided by the case duration (in weeks) to account for case length. This number is averaged for all attorneys in each week and the result is plotted. We observe what appears to be a modest decrease in the number of events per attorney per case per week. Further analysis would be required before drawing more rigorous conclusions about this apparent trend.

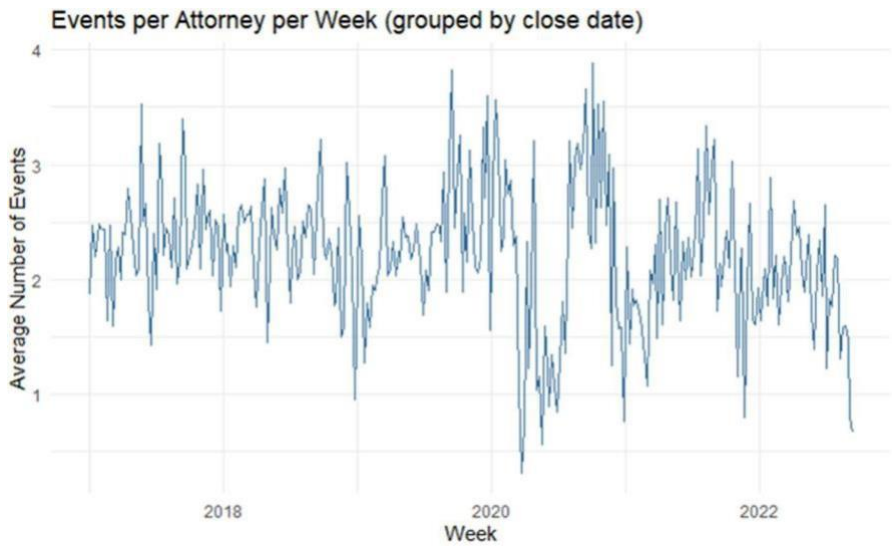



*Figure R6 – Events per Attorney per Week*

## Modeling

We apply a Poisson Regression model (Table R1). The outcome is the event rate (number of events per case per week). We find that the dispersion parameter for the Poisson family is 1. We further find that the Pearson residuals are consistent with model assumptions. There are no overdispersion issues. As a result, this model's assumptions are met, and this model performs well. If the week predictor increases by 1, the expected client contacts per case per week multiplies a factor 1, which means there are no changes at all. If the median case duration

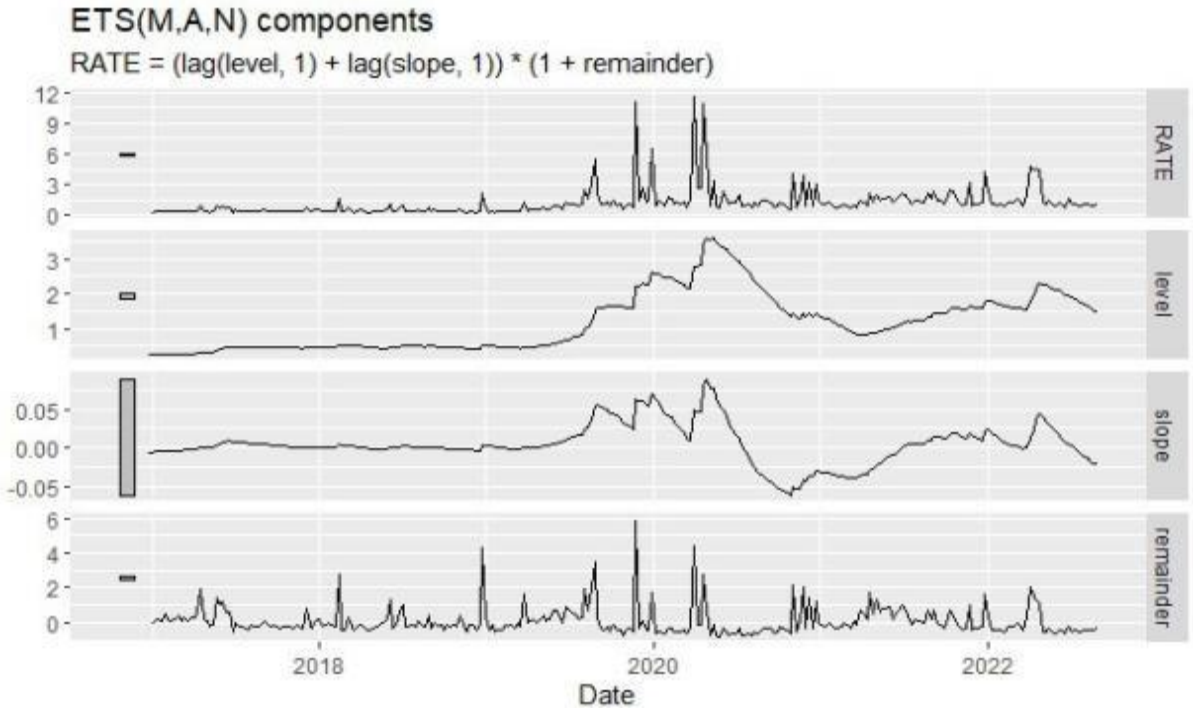
| Coefficients | Estimation | Std. Error | Z Value | Pr(>\|z\|) |
|---|---|---|---|---|
| Intercept | -2.55 | 1.61 | -1.58 | 0.11 |
| Week | 0.00024 | 0.00009 | 2.56 | 0.01 |
| Case Duration | 0.019 | 0.0045 | 4.33 | 0.000014 |

*Table R1 – Poisson Regression*

increases by 1, the expected client contacts per case per week increases by a factor of 1.019, which means there is only a slight increase by 1.9%. The p-values for two coefficients are very small, so there is a statistically significance. However, they are not practically significant. Overall, this model shows us there is not many changes of client contacts in terms of event rates.

We now build grouped cross time series models for event rates. At the individual level, such as the event rate for felony and client contacts, most event rates are slightly less than 1 event per case per week on median, and stable throughout the time. The top level has some sharp fluctuations around 2019 and the beginning of 2020, but is otherwise stable and slowly increases to a median of 2.5 events per case per week. The top-level graph is shown in Figure R7. The top row 'RATE' is the summation of all kinds of event rates. The 2nd row 'level' is the baseline level. The 3rd row 'slope' is the how frequent the total event rate changes based on the 'level'. The final row 'remainder' is the residuals after all decompositions.

*Figure R7 – Grouped Cross Time Series Model*



## Shiny App

We built an R Shiny Dashboard (Figure SA1) to display our data visualization dynamically. This dashboard is mainly used for observing descriptive statistics that answer initial questions about the data. A data dictionary tab called “Purpose Codes of Interest” displays a table with event types of interest as well as their descriptions. The second tab, “Bar Plot,” shows the total number of each event purpose code for felony and misdemeanor cases. The third tab, “Case Number,” shows how the total number of cases for felony and misdemeanors change by month. The fourth tab, “Case Duration,” shows how the median of case durations in weeks changes by each week for felony and misdemeanor case types. The fifth tab is called “Event Number” that shows how the total number of each purpose by case type changes by each week. The sixth tab, “Event Frequency,” shows how the average number of each type of purpose per case changes by each week. The seventh tab, “Event Rate,” shows how the median of event rate changes for felony and misdemeanor cases by week. The final tabs simply introduce the two clients and two consultants. Each plot is dynamic (using plotly) and answers our clients’ questions clearly. Conclusions and inferences were not included on our R Shiny app as the app’s purpose is to simply provide an objective data visualization for our clients to draw their own conclusions.

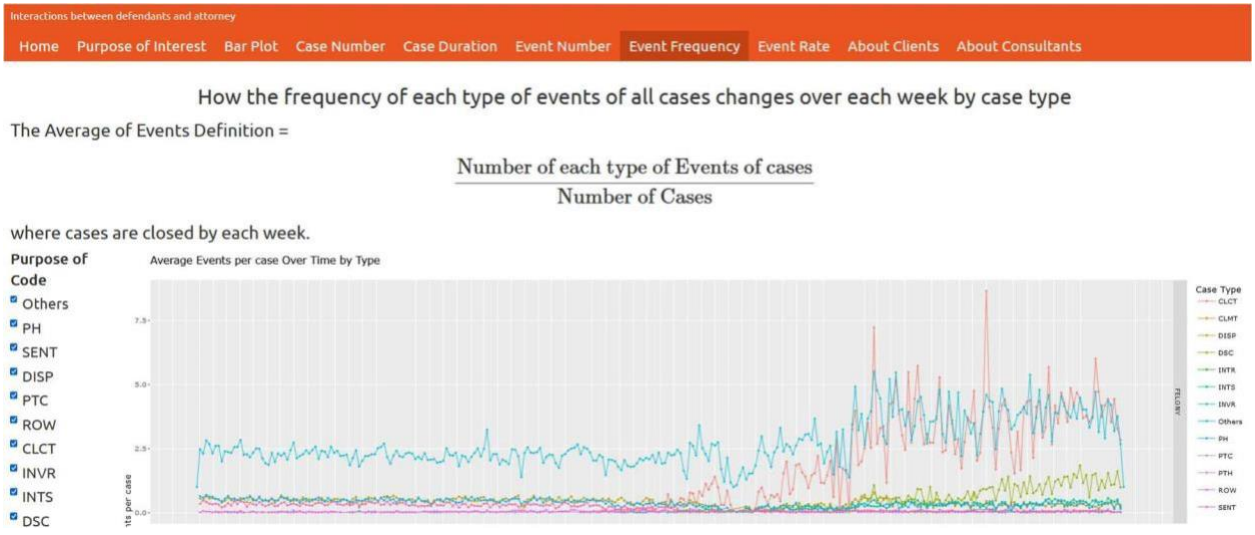



*Figure SA1 – Shiny App*

## Discussion, Conclusions, and Future Analyses

We conclude that after taking case durations and case numbers into consideration, we do not find any practically meaningful changes in using our defined event rate for felony and misdemeanor cases, with the exception of a few isolated sharp increases. This conclusion is supported by both data visualization and time series modeling. Client Contacts are the most common purpose and there are more client contact events after 2020 than pre-2020, but the rate remains unchanged per-case-per-week. Overall, based on our analysis, the current data does not provide evidence of any significant changes in attorney-client interactions. Changes observed after 2020 show more client contact events, but the event rates of cases after 2020 did not change. All other even purpose codes remain very stable across the entire analysis period.

There are some limitations in our analysis. We used medians to summarize event rate and case durations. There may be better ways to define this frequency for this data. There are only very limited events before 2010, so we dropped all cases before 2010. More data pre-2010 data may offer a more complete analysis. For future study, we would recommend the continued tracking of event frequency and event rate. It is possible that there are trends that are not yet detectable in the current data as it's possible that systemic reforms may have a lagging effect of attorney-client interactions which may not be observable for a few years. Additionally, further analysis of changes related specifically to public defenders may uncover additional insights. Lastly, we recommend more consideration of how to most appropriately define the frequency of events of interest, as different definitions may result in different inferences. Additional constructions of frequency metrics may allow for a more comprehensive analysis of potential effects from systemic reforms.

# References


- Hyndman, R., & Athanasopoulos, G. (2021a). 8.5 Innovations state space models for exponential smoothing | Forecasting: Principles and Practice (3rd ed). In *otexts.com.* https://otexts.com/fpp3/ets.html
- Hyndman, R., & Athanasopoulos, G. (2021b). 9.9 Seasonal ARIMA models | Forecasting: Principles and Practice (3rd ed). In *otexts.com.* https://otexts.com/fpp3/seasonal-arima.html
- Posit. (n.d.). *Mutating joins — mutate-joins.* Dplyr.tidyverse.org. https://dplyr.tidyverse.org/reference/mutate-joins.html
- Wickham, H. (2019). *Create Elegant Data Visualisations Using the Grammar of Graphics.* Tidyverse.org. https://ggplot2.tidyverse.org/

## Appendix – Table 1A

| Event Type Code | Description |
|---|---|
| **CLCT** | Client Contact |
| **CLMT** | Client Meeting |
| **DISP** | Disposition |
| **EXPR** | Expert Retained |
| **INTR** | Interpreter Retained |
| **INVR** | Investigator Retained |
| INTS | Initial Text Sent |
| DSC | Discovery |
| PH | Preliminary Hearing |
| PTH | Pre-Trial Hearing |
| PTC | Pre-Trial Conference |
| SENT | Sentence |
| ROW | Return on Warrant |

*Table A1 – Event Types of Interest*