An Analysis of Pro Bono Legal Services’ Response Time

Drew Pang, Meredith Alley, Anmol Sandhu, Richard Li, An Grocki

CONTEXT + OUR QUESTION

Understandably, “pro bono legal services" can be a scary concept for the people reaching out for help. On one hand, they have a hope that they’ll finally get their issues addressed. On the other hand, they have a complete lack of autonomy and just *hope* some kind attorney has the bandwidth to assist them. To us, *time to get a response* is a vital characteristic of attorney-client relationships that we wanted to investigate; particularly, we wanted to understand: **What features of the wording of a post asking for help correlate with increased response time?**

OUR DATASET

Our source for this project was the American Bar Association which, as an organization, is likely a reputable source for information. Furthermore, the data was proved by the American Statistical Association as the “large, rich, and complex data set” used in the 2023 ASA Datafest. The American Statistical Association is a professional organization for statisticians, therefore they have a reputation for providing accurate data. The set of data that we worked with, in particular, was from an online platform through which the ABA collected data about users who were searching for pro-bono legal help. As a result of this breadth of information and lack of human organization, there was a lot of data-cleaning that needed to occur before we could identify any trends (we’ll talk about this in more detail later). This data does not appear to be *missing* anything, per se, though it’s possible that there is data that was omitted for personal safety / privacy reasons. We know that the ABA redacted names, dates, and other identifiable information for this purpose, but don’t know to what extent it occurred.

Finally, this data is not a sample, at least to our knowledge. It appears to be a dataset describing the population of people seeking pro-bono legal help. As a result, we don’t need to make any confidence intervals. There’s no estimation occurring because we’re merely trying to identify trends in the population at large. Our results directly indicate which words DO result in an increased response time.

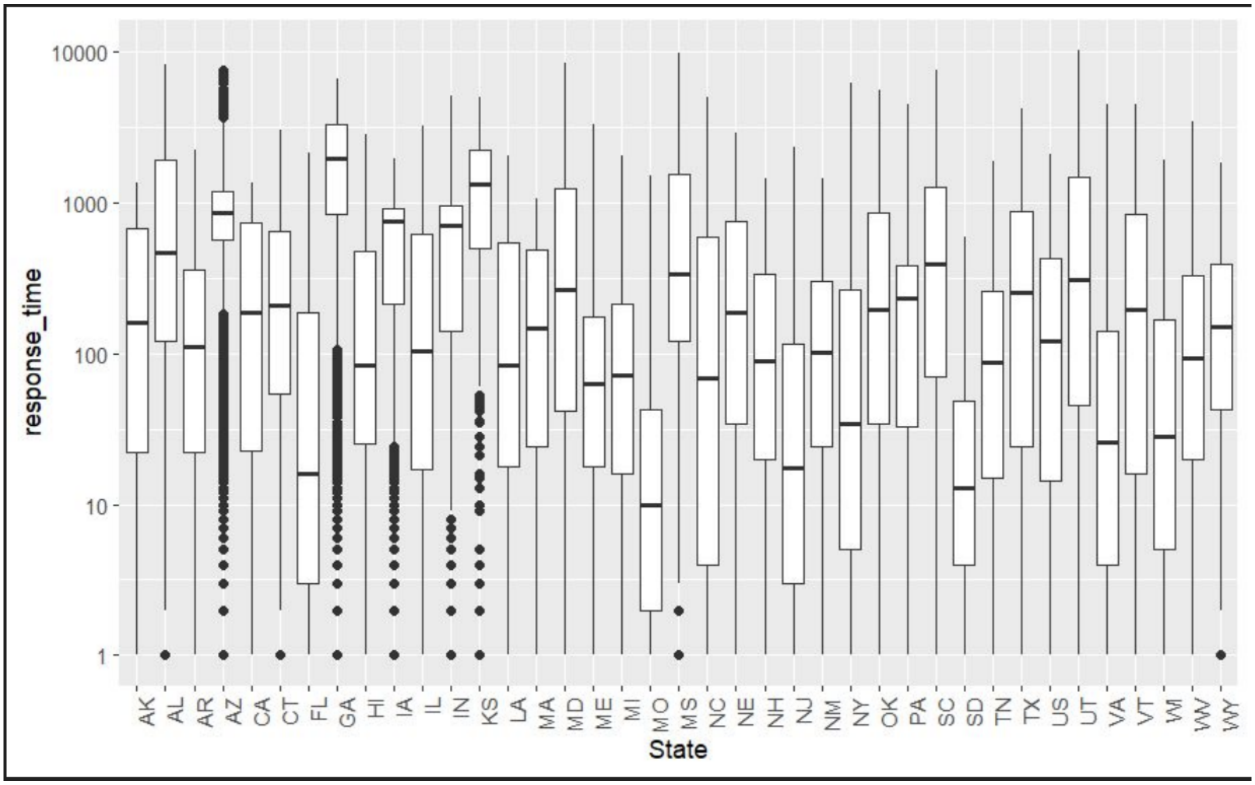
PROCEDURE

We began by looking at how long it took for an inquiry to get answered by-state. While working, we realized that when we visualized response time, we simply dropped unanswered cases. *But arguably, a perpetually unanswered case is worse than a case answered after a long time.* We handled this by defining “response time” as follows, to put a proper amount of weight on an unanswered case:

1. If an attorney took on the case, find out how much time passed between the question being asked and *the lawyer taking the case.*
2. If an attorney did not take on the case, find out how much time passed between the question being asked and the *case being closed*

From here, we visualized correlations between response time and many things, like state, ethnicity, and income.

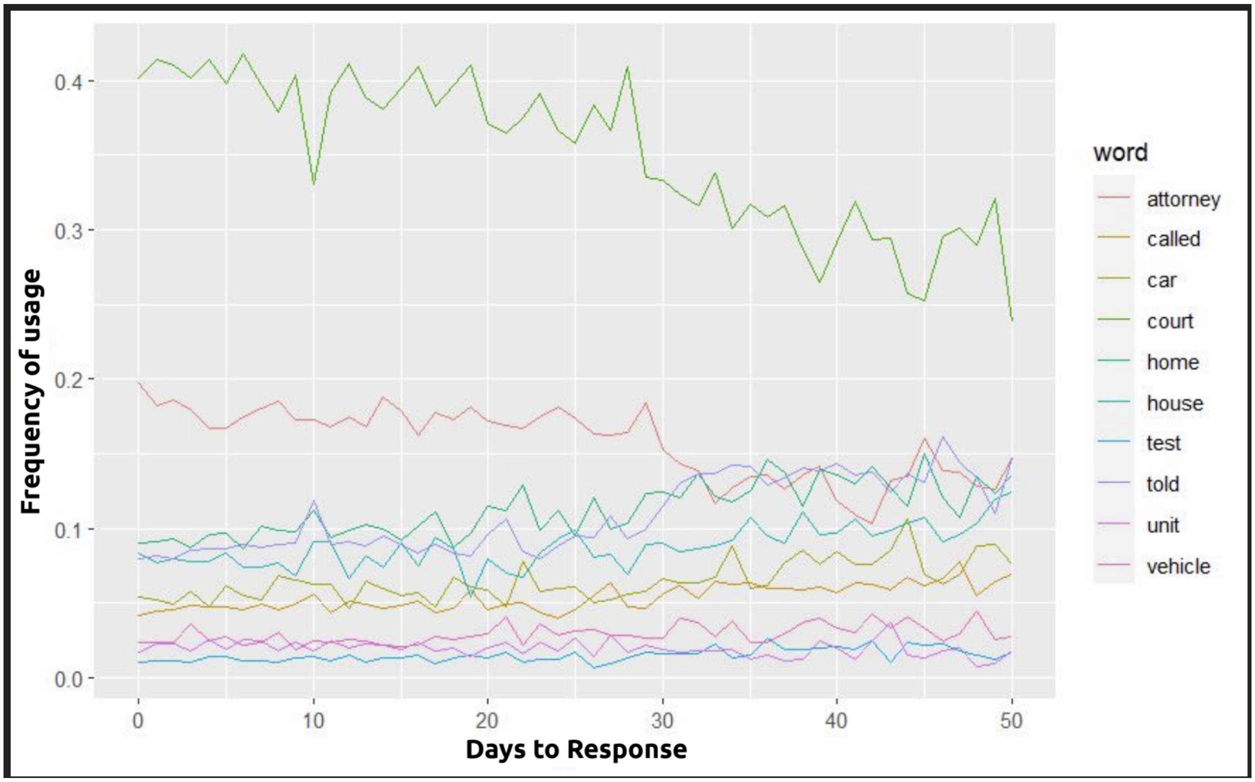
Below is a plot of the response times for each state.



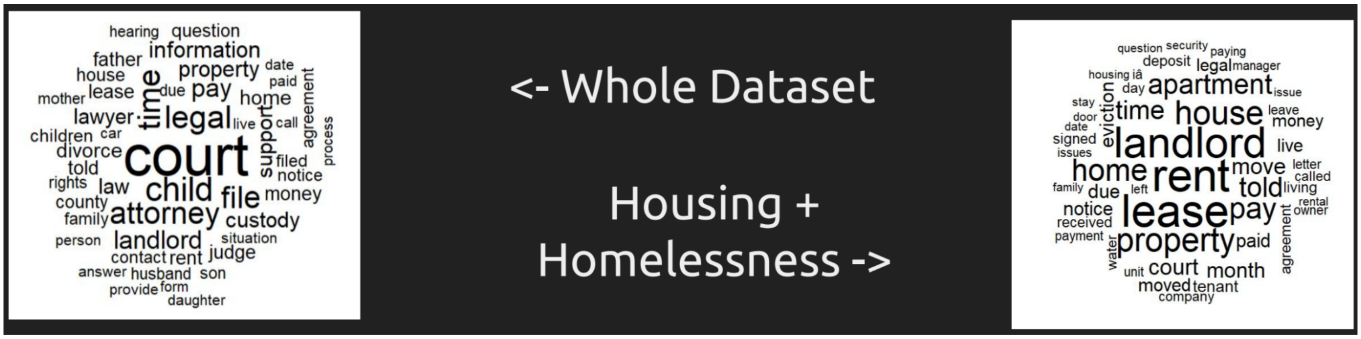
Then, we started looking at the transcripts to do word processing. Upon reading client-lawyer transcripts into R, we ran into a few issues. Every few data points, the transcripts bled into the time created column, because of a specific pattern of quotation marks and commas in the transcripts. While we could have removed all observations with the problematic transcripts, we would have lost almost half of the data. Therefore, to effectively use the transcripts, we fixed the data using a combination of tidyverse functions like *str_sub()*, *unite()*, and *str_replace()*.

To do real data processing on the rate at which questions were answered, we isolated the first post to each question ID, meaning we were only looking at the initial question posted by a client, and removed an official list of stop words, so that words like “and” and “the” were not in our final dataset.

For our final graph, we sorted all of the posts by the number of days it took to get a response, before totalling the words in that group of posts, and then comparing. The frequency of each of the 10 words that have the most impact on the response time are shown on the final graph(with reference to the number of words total for that time period).



In addition to making this frequency graph, we also created simple words clouds representing the frequency of certain words in the responses. Below are examples for the whole dataset and cases categorized as Housing + Homelessness only.



CONFIDENCE IN OUR RESULTS

As we mentioned earlier, there's no way for us to create confidence intervals for our work because we were making and answering claims with the entire dataset. As a result, we can be certain of our results (we don't extrapolate anything).

CONCLUSION

In our investigation, we noticed that using the words “court” or “attorney” trended positively with an earlier response time. This indicates that an emphasis on the *mechanics* of the law will make it more likely that a lawyer will take up the case. This suggests to us that “going to court” being involved in your case may encourage lawyers to treat the case more seriously and thus answer it more quickly.

QUESTIONS REMAINING

1. What solutions would these results suggest?
2. What could the ABA do to take into account the possible systemic bias towards posters that use certain phrases or words?
3. Could we abstract our computation to take phrases into account, instead of just words?
4. Could we abstract our computation to get a general read on the “emotions” of the words?
5. How do these results actually show systemic bias?
6. What would our results look like if we isolate different words, like “Please” or “Thank you”, that would show a sentiment to the posts?