The Paranormal Distributions — DataFest 2023 Write-Up

Team Members: Eli Card, Noah Cline, Carissa French, Brendan Pinkerton

The intent of our statistical analysis was to identify strengths and weaknesses within the knowledge base of the attorneys answering client questions. We did this by examining response time from when a question was asked to when it was taken on by an attorney, i.e. the time from AskedOnUtc to TakenOnUtc. We found this time varies from category to category and from state to state.

It can be inferred that the American Bar Association was wanting to do a textual analysis on the conversations in order to help attorneys better speak with clients, however this has a few issues. First, in order to perform textual analysis on the questions and responses, more time would be necessary. Second, over 38,000 or the 200,000+ questions do not have associated question posts. Third, the formatting of the question posts, varying grammar rule violations, and byte code failures would require large amounts of cleaning to be read.

When cleaning the data, we noticed several quirks in the data. Most notably, there were conversations that were left open more than ten days after their last response; there were some data points that were left open for more than 300 days. We judged it to be important to remove any instances of this happening. We were tempted to identify these as outliers, but the sheer quantity of times this happened indicated something may be incorrect with the automatic closing described in the competition handout. Another quirk was the existence of HTML code within the questionposts.csv document, ZIP codes with only four digits instead of five, and the difficulty of having quotes and commas in strings in a comma-separated values format. Only some of these issues impeded our progress, but we thought it was interesting to note.

Instead of going the route of statistical tests, we looked at various summary statistics to give us a wider idea of some prevalent issues in the system. Investigation of box-and-whisker plots and Q-Q plots demonstrated clear non-normality of the data, which precluded us from pursuing some statistical tests. As the data is not normal, we used the median instead of the average to get an overall view of the trend in each state. The smaller the median, the better the response time. It's easy to understand why a quicker response time is preferable in legal cases, so the map included in the presentation can be used to see which states are in dire need of better response times. The ABA could look into the reason for these slower response times, which could be the result of not enough attorneys in the state, a lack of attorney confidence in answering certain questions, or perhaps something else we haven't considered. There appeared to be a significant difference in the percentage of questions asked for certain categories versus the percentage of those questions that were taken on. This could be another indication of a lack of expertise in some legal areas.

In order for attorneys to better connect with their clients, they must first possess the requisite knowledge and ability to assist those clients. Housing and Homelessness is a repeat offender as one of the largest differences in the two percentages mentioned above, which could be evidence of a topic needing focus. Investigation as to why such a large portion of questions are not taken on in the first place could also make the pro bono system more efficient.