

**Team Name:** UR SUSS

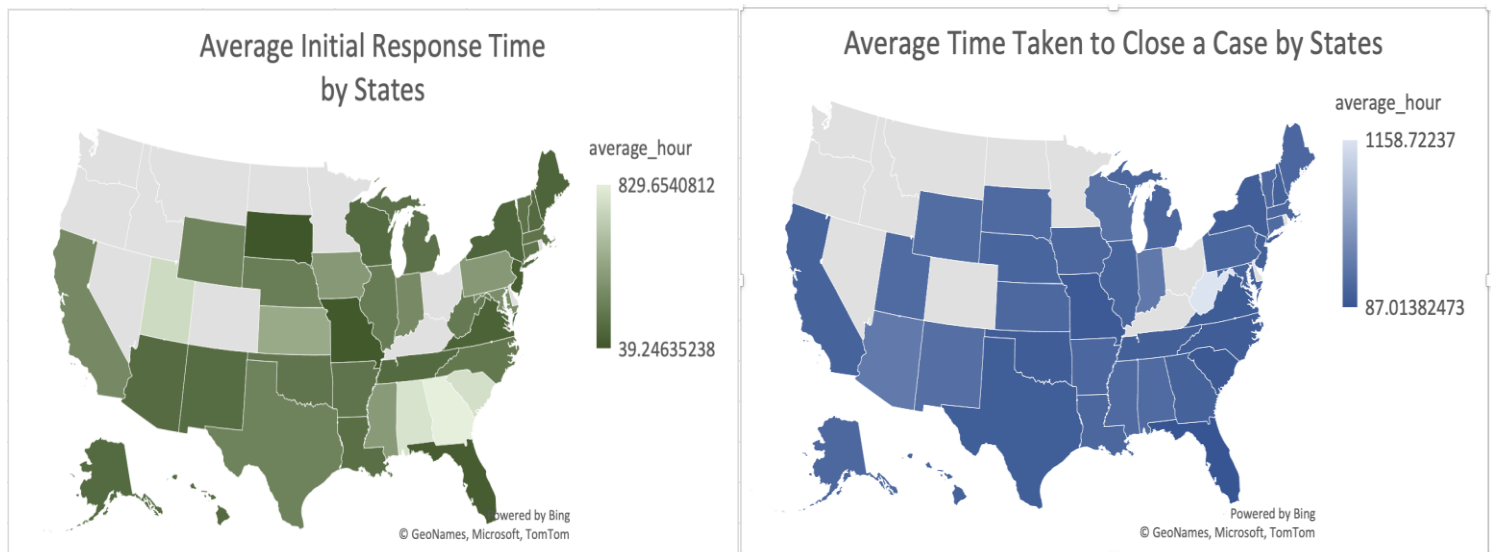
**Member Names:** George Daher, Carter Forbes, Xinying Lu, Tianyi Zhang, Runtao Zhou

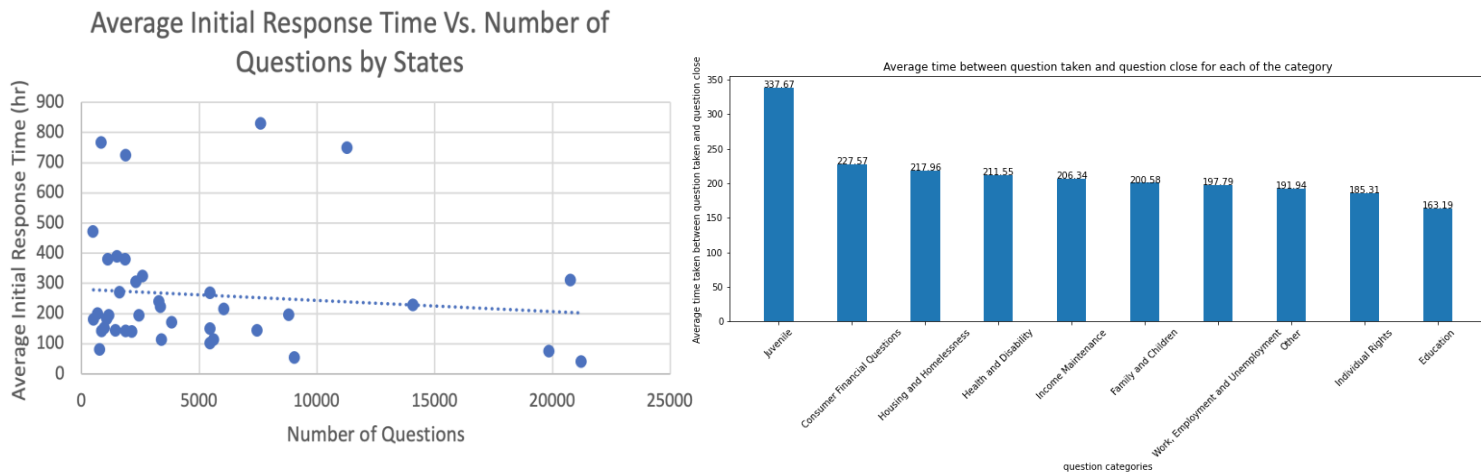
### **Introduction:**

In the data-set there were a number of variables of interest but we decided to focus our analysis on the questions data-set. Specifically, we wanted to analyze the trends in the number of questions asked over time, while also considering the category of the question.

There are a couple reasons that we decided this type of analysis could yield important results. One reason, by looking at the number of questions over time without factoring category, was to predict the overall growth of the platform in terms of the number of questions received per month. This could be important for a number of logistical reasons over the operation of the website, including the amount of administrator work needed to moderate the platform or the number of attorneys needed to address a reasonable amount of questions that were received.

Additionally, such a problem could be pursued further by isolating the categories of question and by considering other factors as possible ways to explain the trends in data.





## Summary of EDA

We started exploring the dataset by looking at the number of questions asked on this online forum by states. Texas has the highest number of questions asked (21230), followed by Indiana(20767) and Florida(19864), and there are 13 states that have not participated or participated very little in this online pro bono legal help website.

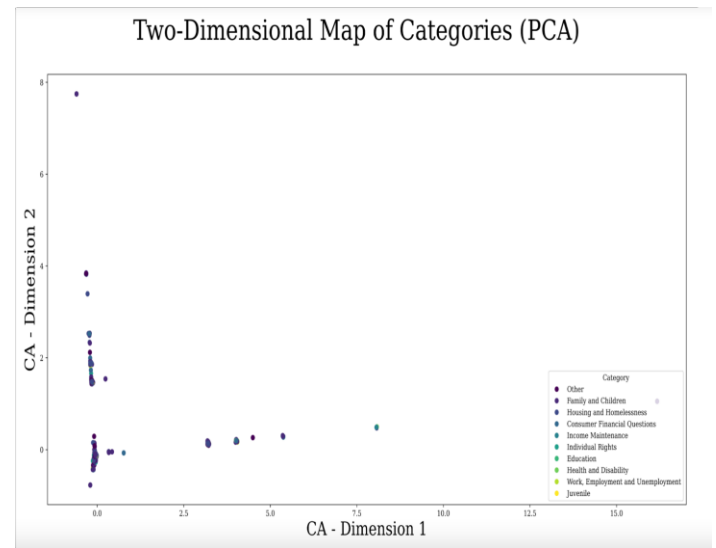
Figure one and two above evaluate efficiency per state. The first graph measures initial response time by state, calculated from the client.csv file's TakenOnUTC and AskedOnUTC columns. Florida and New York have low initial response times, while Utah and Georgia have high ones. The second graph shows the average time taken to close a case, calculated by subtracting the TakenOnUTC column from the ClosedOnUTC column in the client.csv file. These metrics enable comparisons of free lawyer consulting service efficiency between states and identify areas for improvement in reducing response times and speeding up case resolution.

In addition we analyzed response times by category. Which is important as legal issues may require varying levels of attention and urgency. Criminal law or domestic violence cases may need immediate attention, while civil disputes or contract law cases may have longer response times. This helps identify areas where free lawyer consulting services may need to prioritize certain cases and allocate more resources to ensure timely and effective support for clients. From figure three we can see Juvenile cases take the longest and education cases take the shortest

We also looked at the correlation between the average initial response time and the number of questions received by states. According to the scatterplot, we can see that there is a slightly downward trend between the average response time and number of questions. This indicates that as the number of questions increases, the average waiting time for a case to be picked up by volunteers decreases. This could be because the high traffic of legal needs of a certain state is paired with a better established help system, and makes the whole process more efficient.

### **Correspondence Analysis on Client Demographic:**

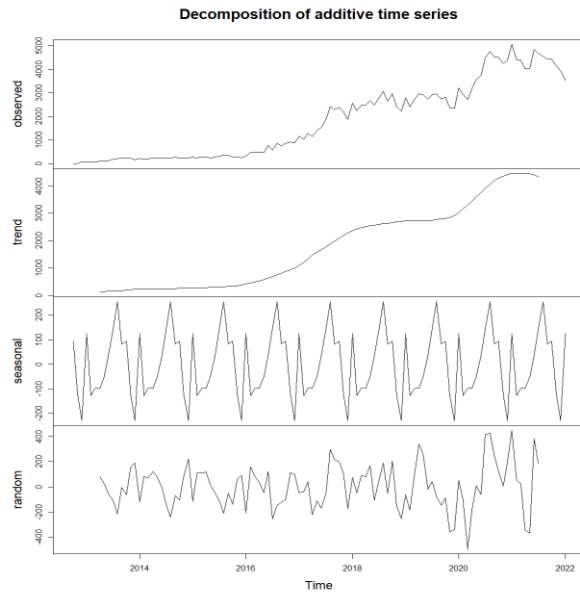
Correspondence Analysis is an extension of Principal Component Analysis (PCA) used for analyzing categorical data. It transforms a contingency table into a graphical representation that visualizes the relationships between categories in a low-dimensional space. The algorithm works by calculating row and column profiles, measuring their dissimilarity, and then reducing the dimensionality of the dissimilarity matrix using PCA. The resulting data is plotted in a low-dimensional space, where clustered categories are similar and occur together frequently in the data. The diagram shows that juvenile, family and children, and housing and houseless categories are similar and occur together frequently, indicating shared legal needs. Volunteer attorneys can tailor their language and communication style to ensure effective communication with clients from different socio-economic and cultural backgrounds, leading to better legal outcomes.



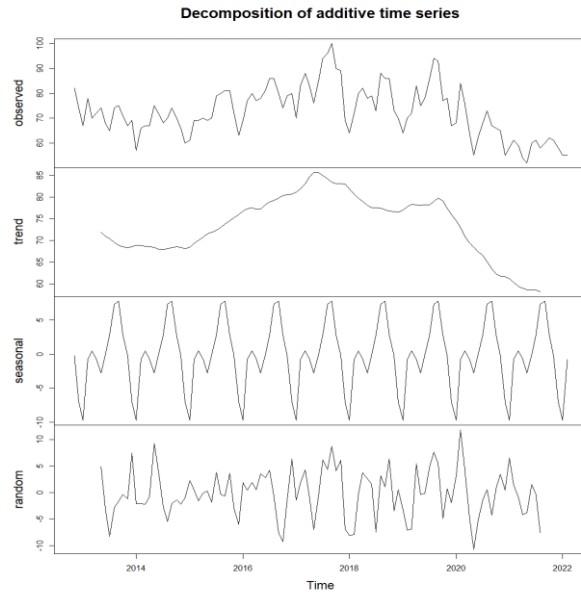
### **Time Series Analysis:**

We analyzed the data as a time series with the frequencies of questions asked per month over the time range in the data set (from October 2012 to January 2022). By formatting the data in this way, a number of techniques can be used to analyze the trends in the data by separating seasonal trends of year by year with the overall trend in the data and then also forecasting the volume of questions on the website in the future. In this section we used external data not in the dataset provided, by getting data on the search popularity of the term

In the graphs below, the graphs for the calculated components decomposed from two time series are shown, one of which is the trend in number of questions per month and the other of which uses data from Google about the search popularity of the topic of Legal Advice. The important thing to observe in both of these sets of graphs, is that the isolated irregularity is relatively constant across either time series. Additionally, the weight of the seasonal component for each of the time series is a significant value relative to the observed data for both datasets. This means that seasonal components show a significant trend in the data.

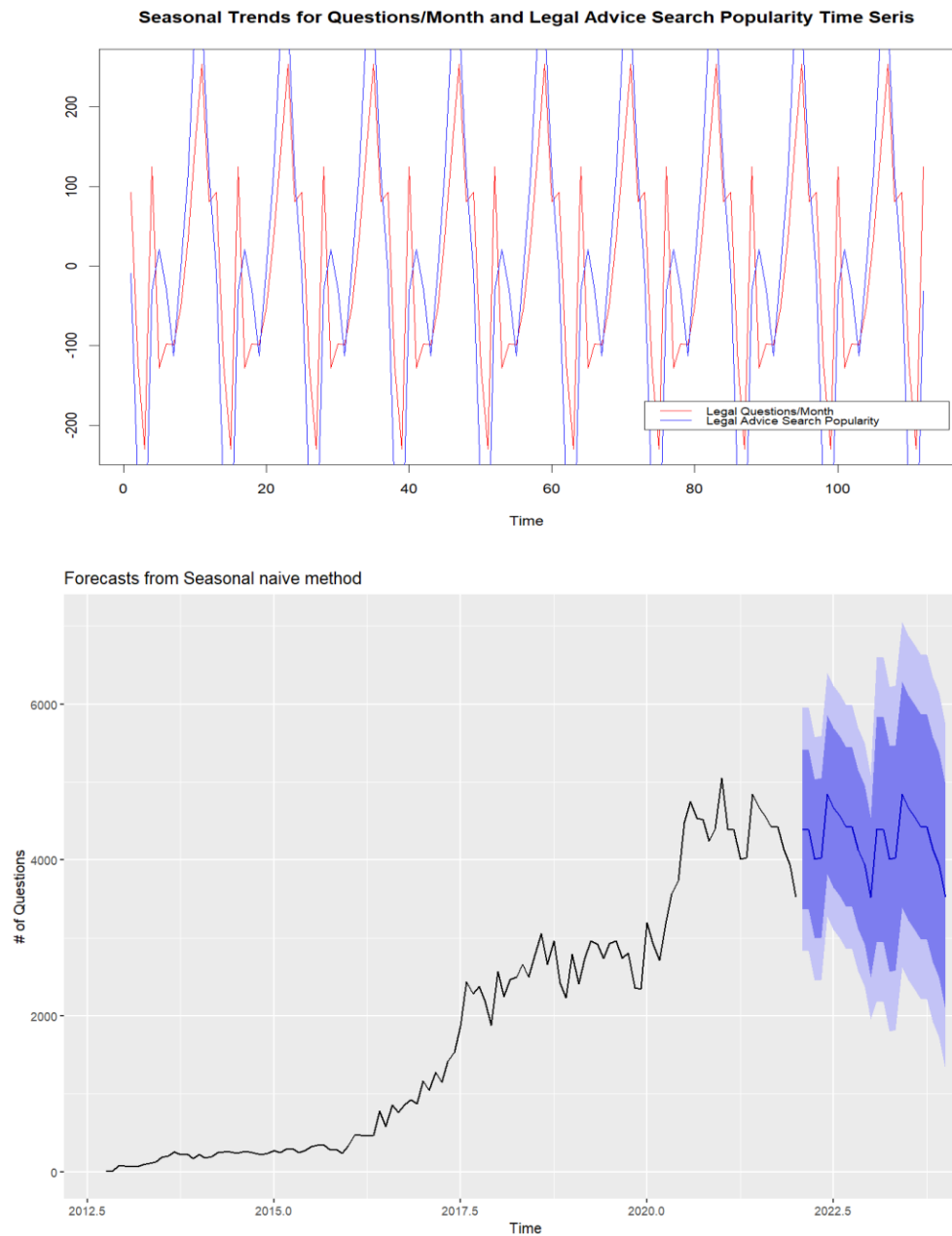


Decomposition of the Number of Questions/Month Time Series



Decomposition of the Legal Advice Topic Search Popularity Time Series

When comparing the seasonal trends of both time series, an interesting result can be observed as the general seasonal trends align over time between both data sets. In the figure to the left below, this relationship is illustrated with the scale of the trends normalized for visual clarity (note that the scale of the units for the variables measured by either time series is very different which is why this adjustment is needed). Together, these discoveries are important because they explain a statistically significant portion of the difference in traffic on the website from month to month as seasonal variation. Furthermore, the comparison to Google search data, allows us to reasonably conclude that the trends in the number of questions asked follow a more general trend about the periods of the year in which people seek the most legal advice online. Additionally, based on this, it would be reasonable to hypothesize that many people who search for legal advice online are able to find the resource and ask the questions they have. However to come up with a stronger conclusion on this, more search data would need to be referenced.



Another benefit of analyzing the data as a time-series is that a number of forecasting methods can be used to predict the possible values for a variable in the future based on the general and seasonal trends in the past data. Out of these methods we selected a seasonal naive forecasting method. We did this for two main reasons, first: the general non-seasonal trend in the number of questions appears to be plateauing somewhat, and seasonality of the data, as explained previously, was significant. Plotting our forecast with a prediction interval yielded the resulting graph to the right above.

## **Conclusion & Suggestions**

It is important to measure the efficiency of free lawyer consulting services across different states and identify areas where improvements can be made to reduce response times and increase the speed of case resolution. The average initial response time and the time taken to close a case are two important metrics that can be used to compare the efficiency of free lawyer consulting services in each state. It is also important to analyze response times by category to identify areas where the free lawyer consulting service may need to allocate more resources or prioritize certain cases. By doing so, clients can receive the assistance they need in a timely manner and ensure that the service is providing effective and equitable support to those who need it.

The Correspondence Analysis graph helps in visualizing the relationships between the categories in a low-dimensional space. The categories juvenile, family and children, and housing and houseless are clustered together, indicating that there may be a correlation or shared legal needs among these categories. It is important to take into account the socio-economic and cultural background of their clients, and volunteer attorneys can tailor their language and communication style to ensure effective communication and avoid any language barriers.

Analyzing the data as a time series with the frequencies of questions asked per month over time is important to predict the overall growth of the platform in terms of the number of questions received per month. This could be important for a number of logistical reasons over the operation of the website, including the amount of administrator work needed to moderate the platform or the number of attorneys needed to address a reasonable amount of questions that were received.

Finally, our time series analysis demonstrated the importance of analyzing trends in the volume of questions asked over time in order to predict the growth of the platform and plan for logistical and staffing needs. This project highlights the importance of data-driven approach to better understand and address the legal needs of the underserved communities.

In conclusion, our analysis of the pro bono legal help website data set provided valuable insights into the trends and patterns of legal needs across different states and categories. The exploratory data analysis helps to identify areas where improvements can be made to ensure that clients receive timely and effective legal assistance. The Correspondence Analysis can help volunteer attorneys to tailor their language and communication style to better serve their clients.