STAT 302: Project 1

Jaiden Atterbury, Ty Mellish, Qinyi Zhong

Due on 05-12-2023

Part 1: Context:

The main topic for this project is information surrounding the American Bar Association (ABA), and how this data can be used to help the ABA improve their services as a whole. For some context, in some states/territories in the United States, the ABA provides free of charge legal services via an online platform. This platform allows low-income individuals who qualify based on the low income laws of the state they reside in to post legal questions and receive advice from volunteer lawyers.

Although the ABA, which was founded in 1878, has been steadily increasing the amount of clients on their website since the early 2010s, they would like to anticipate the types of legal questions that arise so that they can prepare their volunteer lawyers to: address the questions at hand, better know what types of lawyers to recruit, and lastly know how to advise state partners on the general trends they are seeing. Furthermore, the ABA would also like to know what kinds of clients they are dealing with to better understand the cultural, societal and emotional context of the clients' messages, and thus be able to better understand how to listen and engage with clients.

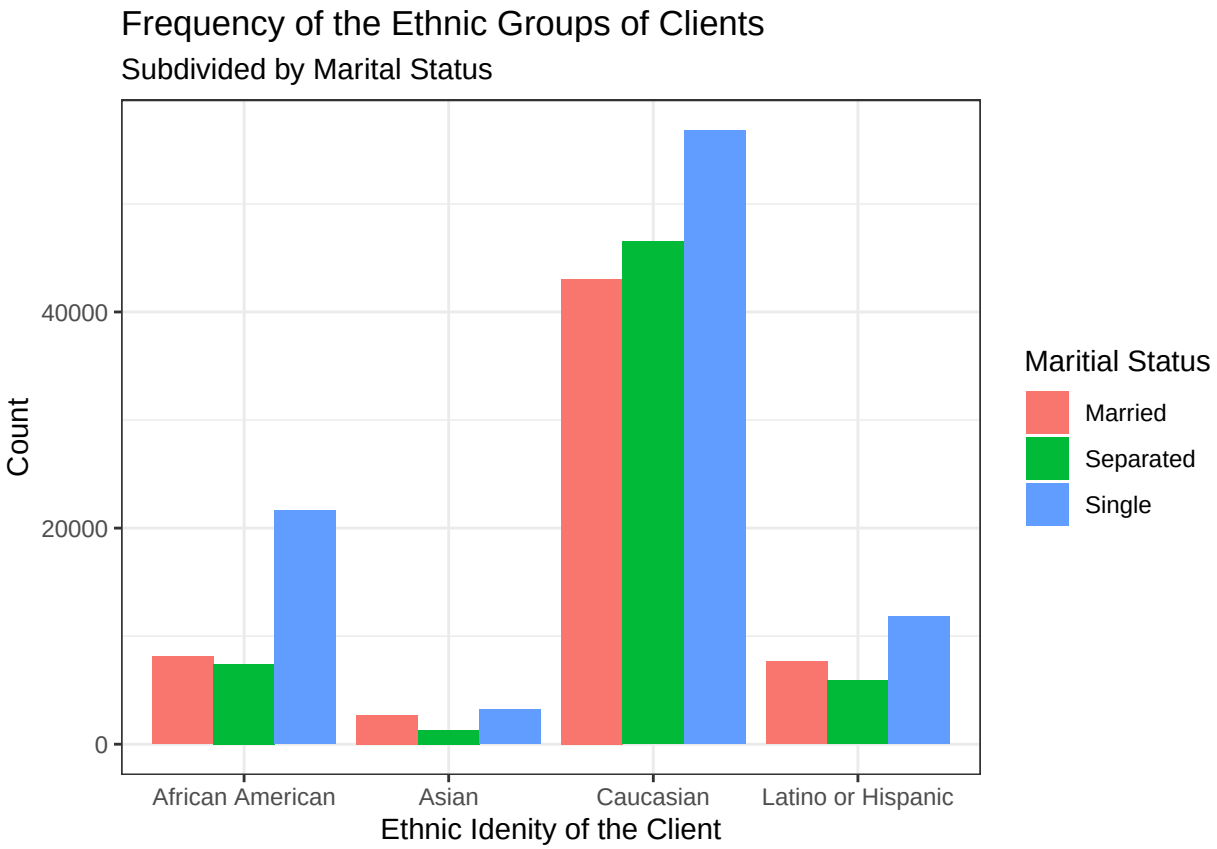
With this new information the ABA and their volunteer attorneys will be able to better adapt to: changing laws, cultural and societal changes, natural disasters, and other events that will cause demand shocks in the number of clients siting their legal advice website.

In this project, our team will analyze 8 datasets containing information from the ABA about legal questions, lawyers, clients, and state restrictions. With this data we will be able to provide advice to the ABA about certain trends we see in the conversations that will help the ABA advise their volunteer lawyers. Through these trends we will be able to advise the ABA on how they should allocate resources to respond to these given trends.

In order to be able to find such trends and report them to the ABA we will create three visualizations, pose and answer a research question, and lastly summarize our findings and give some suggestions to the ABA.

Part 2: Visualizations:

Visualization 1: Ethnicity Trends:



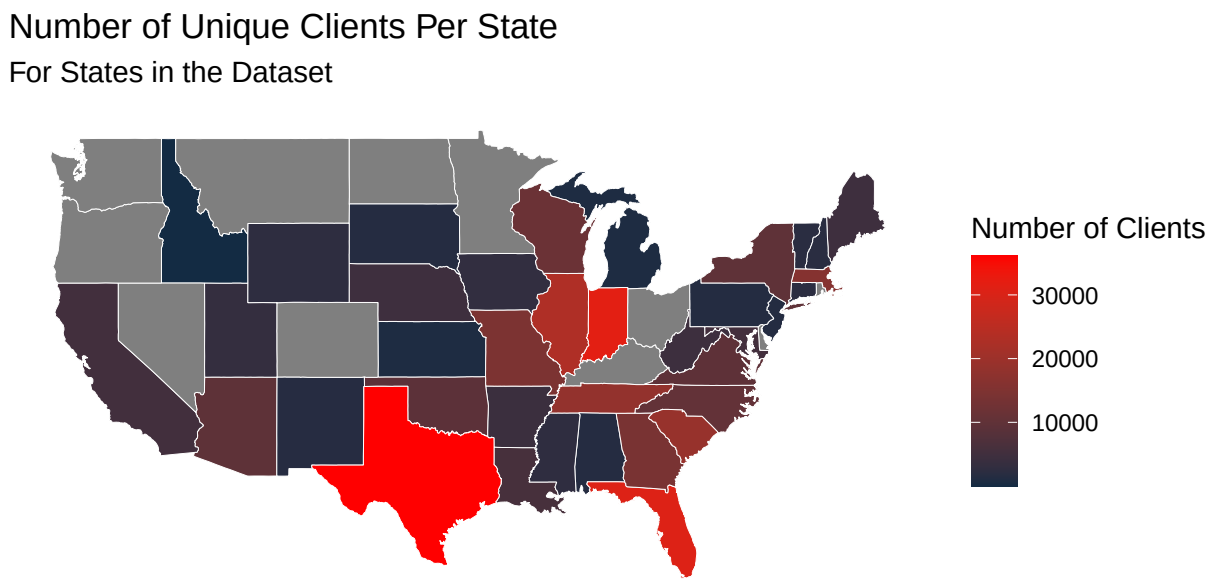
CAPTION: This visualization is a bar chart of the frequency of the four most prevalent ethnic identities of the clients using the ABA legal services across the three levels of marital status. The y-axis shows the frequencies of each ethnic identity based on each level of marital status, while the x-axis shows the particular ethnic identities, all of which are additionally split up by marital status which is sorted by color, as seen in the legend to the right. The ethnic identities chosen to be represented are the top four most represented ethnic identities for the clients in the data.

ANALYSIS: As can be seen in the above visualization, there are several ethnic communities that are more likely to be involved in the pro-bono legal system, the largest being Caucasian, with African American and Latino/Hispanic communities being quite smaller, and the Asian community being the smallest of the four. Within each community, it can be seen that the largest proportion of individuals consider themselves to be single. For Caucasian and Latino demographics, however, married and separated proportions are on a very similar level, to the point that if you were to consider the proportion of clients who either were previously married or are currently married, this population would greatly exceed the single population. This means that these communities are likely to have more questions regarding family, whereas the African American population has a significantly larger single population than married and separated, and will likely have less questions regarding family and marriage.

It is important to note that in the above visualization the number of levels of marital status were reduced significantly for the sake of simplicity. In particular, we started off by removing individuals who refused to identify their marital status as their inclusion gives us no information, the “Married” label is comprised of clients who considered themselves to be balled as “Married” or “Married or Remarried”, the “Separated” label is comprised of clients who considered themselves to be balled as “Separated”, “Divorced”, “Widowed”, or “Divorced or Widowed”, and lastly the “Single” label is simply comprised of individuals who considered themselves to be balled as “Single”. This visualization allows us to inform the ABA on the kinds of people they will be dealing with and the cultures that most represent these people. By doing this the ABA will be able to pinpoint their recruitment to math these given trends, and will be able to better communicate to

these certain individuals in the most represented groups in the dataset.

Visualization 2: Client Frequency Trends:

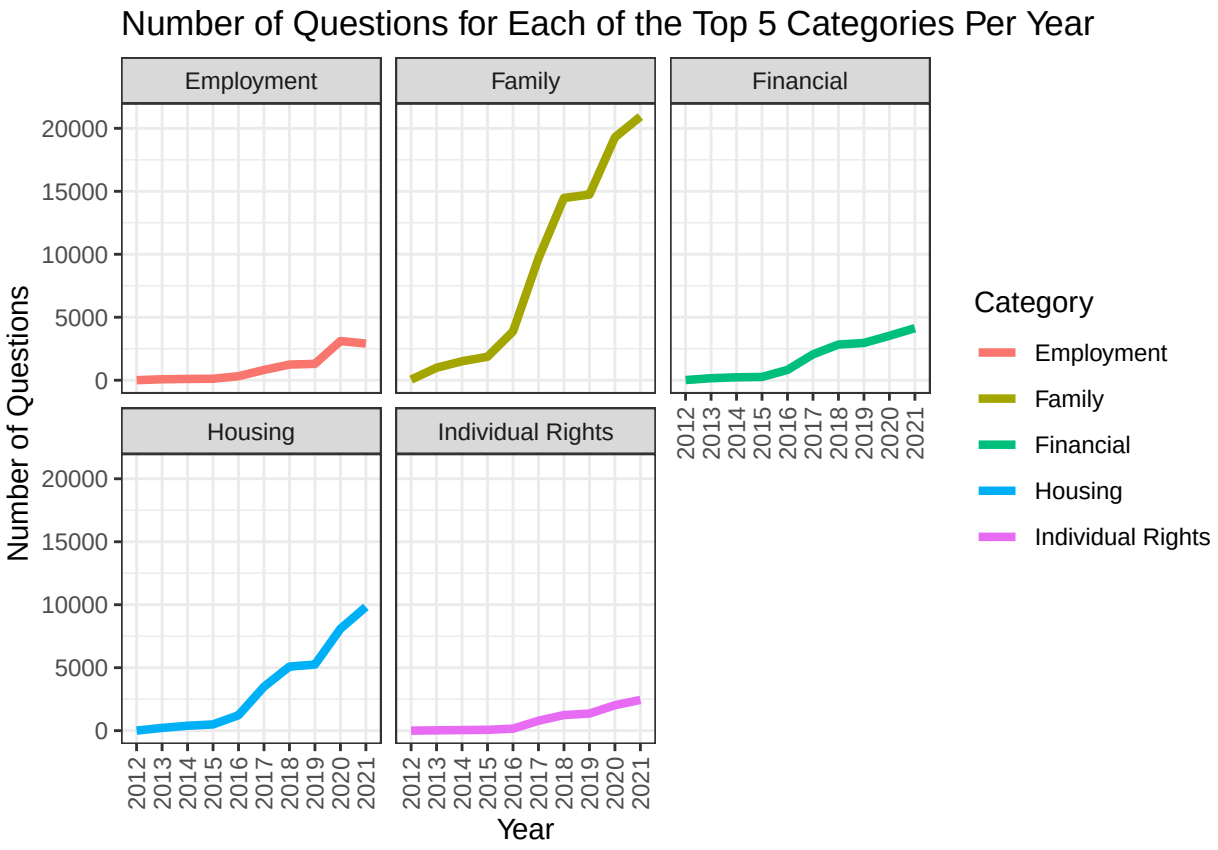


CAPTION: This visualization is a heat map of the United States based on the number of unique clients in the ABA dataset that reside in each state. In particular for the states that are represented in this dataset. The legend on the right shows which shades of red correlate to different values for the amount of clients in that state. Notice that the darker/the more blue the fill color is, the smaller the number of clients that live in that state. On the contrary, the brighter/the more red the fill color is, the higher the number of clients that live in that state.

ANALYSIS: First of all, it should be noted that the upper middle parts of America, also known as the Great Plains, as well as the Pacific Northwest have a majority of their states grayed out. This happens because, there were only 40 states/territories in which data was collected, and it happens to be that for these state there was no data collected on clients. However it is important to note that this doesn't mean that a pro-bono legal system does not exist in these states, but instead, means that data was not collected in these states for this survey. Select other states in middle America, and some in the east, exhibit similar areas with no data collected. Otherwise, as can be seen in the visualization, the highest concentration of clients in the pro-bono legal system is in Texas with a count of 36,050. There are several other areas with high concentrations, such as Florida, Illinois, and Indiana, all having counts of above 23,000. It should be noted that the visualization displays the number of clients without taking population into account, meaning that for Texas and Florida, having populations of over 29 and 21 million, this may be a smaller portion of the population than Illinois and Indiana, which have populations of under 13 and 7 million. However, these states would all still obviously have a need for a larger number of legal volunteers and services than the other areas with fewer clients. Lastly, Alaska and ABA federal do have data, however obviously those weren't able to be mapped, these "states" had 1567 and 1054 clients respectively. These numbers would put these "states" in the lower end of state counts.

By mapping the number of registered clients per each state, we can help advise ABA to train/recruit volunteer lawyers who specialize in the laws of the states with the most clients. This is important because each state has its own set of laws that differ from state-to-state, thus we will want to have a dense amount of volunteers from the states with the most people in them in order to ensure that response times are as quick as possible.

Visualization 3: Question Trends:



CAPTION: This visualization shows the number of questions asked in a year based on category of the question. In particular the visualization focuses on the top five types of questions in terms of frequency. On the y-axis, the number of questions for each category in each year is shown. On the x-axis, the years going from 2012 to 2021 are shown, and each line is sorted by a color corresponding to a category, which is seen in the legend to the right of the graph. It is important to note that the year 2022 was excluded from analysis since the data was collected before the year had concluded, thus it showed trends that don't represent the true nature of the clients.

INSIGHTS: From the visualization, we can see that ever since 2012, Family and Children has been the most commonly asked question category, with the difference only growing larger over time. The next most popular question category is Housing and Homelessness which is on a similar inclining pattern as to Family and Children but on a smaller degree of magnitude. The other three categories exhibit smaller growth on generally similar levels. Between 2019 and 2020, there was a small spike in Work, Employment, and Unemployment questions, which seems to be returning to a normal level in 2021. This may be due to the unemployment changes during the beginning of the COVID-19 pandemic and the shutdown/quarantine of many areas in the U.S. All the graphs seem to plateau somewhat between 2018 and 2019, then exhibiting the normal growth patterns from 2019 on, with the exception of the Work, Employment, and Unemployment spike. On a general level, all of the categories seem to exhibit slow growth from 2012 to 2015, and then exhibit an increase in the speed of growth from then on. This could be due to an increase in low income population, or due to an increase in need for legal questions in general.

This visualization is very important because it meets the ABAs goal of understanding what types of questions the clients using their website ask most frequently. By seeing the types of questions that occur most frequently, we can advise the ABA on how they should train/recruit their volunteer lawyers, as different types of questions require different types of expertise and areas of knowledge. Furthermore, this visualization also shows us that, in general, their services are being used more and more often as the years go on. This is very important because it allows the ABA to know that they should plan on expanding their services/recruit more lawyers.

Part 3: Research Question + Model:

Research Question + Context:
As discussed above, the ABA wants to know what types of questions they receive most often in order to allocate their resources/acquire certain lawyers with expertise in these ares most efficiently. Hence, in order to fulfill this need for the ABA our group has proposed the following research question: which variables significantly predict the type of question being asked by a given client? The steps to how we will answer this research question are mapped below.

Since we are interested in the type of question a client will ask given other aspects of the client, this task is a classification task. As seen in visualization 3, the Family and Children question type is the most frequent question type. In particular, it is almost as frequent as every other category combined. With that being said, for the sake of simplicity, we will split the Category variable into two categories: "Family and Children" and "Other." By doing this we unlock the ability to use logistic regression to model this phenomenon, and that is exactly what we will do.

In particular we will run three different models, each model adding one more variable than the last. Once we have created these three models we will compare the models and choose the best one given a certain criterion. Once we have selected the best model we will run through the model assumptions, test the model accuracy, and lastly present the relevant conclusions that we can find about the model.

The variables that we will be looking at are: age, number of people in the household, gender, and lastly marital status. The marital status variable is broken up exactly how it is in visualization 1. The gender variable is also combined for simplicity, for example, the three levels of the gender variable we will be focusing on are: male, female, and other.

One important aspect of our data is that we will be getting rid of any client that doesn't have values for all of these variables, this is because we can't accurately make a prediction if we don't have all of the required data for an individual. Furthermore, when creating our three models and testing the best one, we will split our entire dataset into a training set, which will contain 70% of the data, and a testing set, which will include 30% of the data.

Model Building:
Now that we have posed the research question and formulated what kind of model we are using, and why we are using that kind of model, we will now build three different models in order to answer said question. The first logistic regression model will be using gender, marital status, and age in order to predict which type of question the client asked. The second logistic regression model will be using gender, marital status, age, and the number of individuals in a household in order to predict which type of question the client asked. Lastly, the third logistic regression model will be using gender, marital status, age, and the number of individuals in a household in order to predict which type of question the client asked. We will run and analyze each of the three models below.

Model Summaries:
Below we will build and analyze the first model.

```
##
## Call:
```

```
## glm(formula = as.factor(Category) ~ as.factor(Gender) + as.factor(MaritalStatus) +
##     Age, family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7597  -1.0452  -0.7028   1.1525   2.6307
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                        1.6342566  0.0252166   64.81   <2e-16 ***
## as.factor(Gender)Male             -0.3366514  0.0139311  -24.16   <2e-16 ***
## as.factor(Gender)Other            -0.7256024  0.0671502  -10.81   <2e-16 ***
## as.factor(MaritalStatus)Separated  0.4299036  0.0166870   25.76   <2e-16 ***
## as.factor(MaritalStatus)Single    -0.8012644  0.0158224  -50.64   <2e-16 ***
## Age                               -0.0397283  0.0005458  -72.79   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 156700  on 114324  degrees of freedom
## Residual deviance: 146363  on 114319  degrees of freedom
## AIC: 146375
##
## Number of Fisher Scoring iterations: 4
```

As can be seen from the above output, the intercept and the parameters for gender, marital status, and age are all extremely significant at the 5% level. The first model also has an Akeike Information Criterion of 146375, which will come into play in the model comparison section below.

Below we will build and analyze the second model.

```
##
## Call:
## glm(formula = as.factor(Category) ~ as.factor(Gender) + as.factor(MaritalStatus) +
##     Age + NumberInHousehold, family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4713  -1.0407  -0.7002   1.1511   2.6214
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                        1.3379787  0.0314396   42.56   <2e-16 ***
## as.factor(Gender)Male             -0.3253760  0.0139673  -23.30   <2e-16 ***
## as.factor(Gender)Other            -0.7039379  0.0672479  -10.47   <2e-16 ***
## as.factor(MaritalStatus)Separated  0.4923940  0.0171928   28.64   <2e-16 ***
## as.factor(MaritalStatus)Single    -0.7234704  0.0165616  -43.68   <2e-16 ***
## Age                               -0.0379084  0.0005585  -67.88   <2e-16 ***
## NumberInHousehold                  0.0601168  0.0038346   15.68   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 156700  on 114324  degrees of freedom
## Residual deviance: 146092  on 114318  degrees of freedom
## AIC: 146106
##
## Number of Fisher Scoring iterations: 4
```

As can be seen from the above output, the intercept and the parameters for gender, marital status, age, and number of individuals in the household are extremely significant at the 5% level. The second model also has an Akeike Information Criterion of 146106, which will come into play in the model comparison section below.

Below we will build and analyze the third model.

```
##
## Call:
## glm(formula = as.factor(Category) ~ as.factor(Gender) + as.factor(MaritalStatus) +
##     Age + NumberInHousehold + AnnualIncome, family = "binomial",
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.5781  -1.0400  -0.7014   1.1497   2.6228
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       1.364e+00  3.218e-02  42.396  < 2e-16 ***
## as.factor(Gender)Male            -3.222e-01  1.399e-02 -23.030  < 2e-16 ***
## as.factor(Gender)Other           -7.037e-01  6.726e-02 -10.463  < 2e-16 ***
## as.factor(MaritalStatus)Separated  4.831e-01  1.735e-02  27.853  < 2e-16 ***
## as.factor(MaritalStatus)Single   -7.340e-01  1.677e-02 -43.758  < 2e-16 ***
## Age                              -3.791e-02  5.588e-04 -67.832  < 2e-16 ***
## NumberInHousehold                 6.372e-02  3.966e-03  16.067  < 2e-16 ***
## AnnualIncome                     -1.393e-06  3.408e-07  -4.088 4.35e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 156700  on 114324  degrees of freedom
## Residual deviance: 146074  on 114317  degrees of freedom
## AIC: 146090
##
## Number of Fisher Scoring iterations: 4
```

As can be seen from the above output,the intercept and the parameters for gender, marital status, age, number of individuals in the household, and annual income are extremely significant at the 5% level. The third model also has an Akeike Information Criterion of 146090, which will come into play in the model comparison section below.

Choosing the Best Model:
For the above models, we will be using their corresponding AIC scores in order to determine which model to select. In the field of machine learning, the AIC is used to compare different models in order to determine which one is the best fit for the given data. This is possible because the AIC is calculated using the likelihood function of the model. Furthermore, AIC penalizes models for having more variables.

Since the AIC is good at determining which model fits the data the best, relative to other models, we will use this statistic to assess which model we will select which in turn answers our research question. In general, if the AIC is big, the fit is worse; if the AIC is smaller; the fit is better. Furthermore, if two models fit the data similarly on the basis of AIC, the simplest model is better.

With that said, the AIC of model 1 was 146375, the AIC of model 2 was 146106, and lastly the AIC for model 3 was 146090. Given that models 2 and 3 have noticeably lower AICs than model 1, we will start by eliminating model 1. However, even though model 3 has a lower AIC than model 2, since there AICs are relatively the same, we will defer to selecting model 2 since it is the simpler model.

Analyzing the best model:

Now that we have decided on using model 2, we will now report and interpret the parameter values in the context of the ABA data set. For a reminder the dependent variable is category, and the independent variables are gender, marital status, age, and number of individuals in the household. For all of the confidence intervals we will run them at the 95% level. Furthermore, since we are using AIC which involves using the log-likelihood function, it makes sense that we use the profiled log-likelihood intervals instead of the standard error intervals.

To start us off, we will start with the numerical variables and the intercept. The intercept took on a value of 1.3379787 with a z value 42.56, a standard error of 0.0314396, and a 95% confidence interval of [1.27632539, 1.39956278], hence the intercept is significant at the 5% level.

Next, we see that the age variable took on a slope value of -0.0379084 with a z value of -67.88, a standard error of 0.0005585, and a 95% confidence interval of [-0.03900442, -0.03681513], hence the age variable is significant at the 5% level. Hence for every one year increase in age, the log odds of the category decrease by -0.0379084, holding all other variables constant.

For our last numerical variable, we see that the age variable took on a slope value of 0.0601168 with a z value of 15.68, a standard error of 0.0038346, and a 95% confidence interval of [0.05264566, 0.06767348], hence the age variable is significant at the 5% level. Hence for every one member increase in the number of people in the household, the log odds of the category being Family and Children decrease by -0.0601168, holding all other variables constant.

Now we move onto our categorical variables, which have a different interpretation. We will start off with our gender variable. First off, the male level has a slope value of -0.3253760 with a z value of -23.30, a standard error of 0.0139673, and a 95% confidence interval of [-0.35276246, -0.29801123]. Now looking at the other level of gender, we see that it takes on a slope value of -0.7039379 , with a z value of -10.47, a standard error of 0.0672479, and a 95% confidence interval of [-0.83679237, -0.57310185]. Hence both of these levels are significant at the 5% level. Furthermore, being of the gender male and other, as compared to female, changes the log odds of category by -0.3253760 and -0.7039379, respectively.

Lastly, we will look at our marital status variable. First off, the separated level has a slope value of 0.4923940 with a z value of 28.64, a standard error of 0.0171928, and a 95% confidence interval of [0.45871331, 0.52610806]. Now looking at the single level of marital status, we see that it takes on a slope value of -0.7234704, with a z value of -43.68, a standard error of 0.0165616, and a 95% confidence interval of [-0.75593916, -0.69101859]. Hence both of these levels are significant at the 5% level. Furthermore, being of the marital status separated or single, as compared to married, changes the log odds of category by 0.4923940 and -0.7234704, respectively.

Thus, the final logistic regression equation for our model is: log odds of category $= 1.338 - 0.038 \cdot \text{Age} + 0.06 \cdot \text{NumberInHousehold} - 0.325 \cdot \text{Male} - 0.705 \cdot \text{Other} + 0.492 \cdot \text{Separated} - 0.723 \cdot \text{Single}$.

Model diagnostics:

Next, we will now use our test set created above to test the accuracy of our created model. Based on a simple algorithm that labels a prediction 1 if it is greater than or equal to zero, and zero if it is less than zero, we see that the accuracy of our model is roughly 64.2%.

Model Assumptions:
Our last step in answering this research question is to analyze our model assumptions to see if our model is of any use "in the real world." In particular we will follow the six main assumptions for logistic regression:

Assumption 1: The response variable is binary:
Since we purposely recodified our response variable category to only take two level: Family and Children and Other, we satisfy this assumption.

Assumption 2: The observations are independent:
Since the observations of the client were taken individually without any consideration of the other clients, it is safe to assume that our observations are independent.

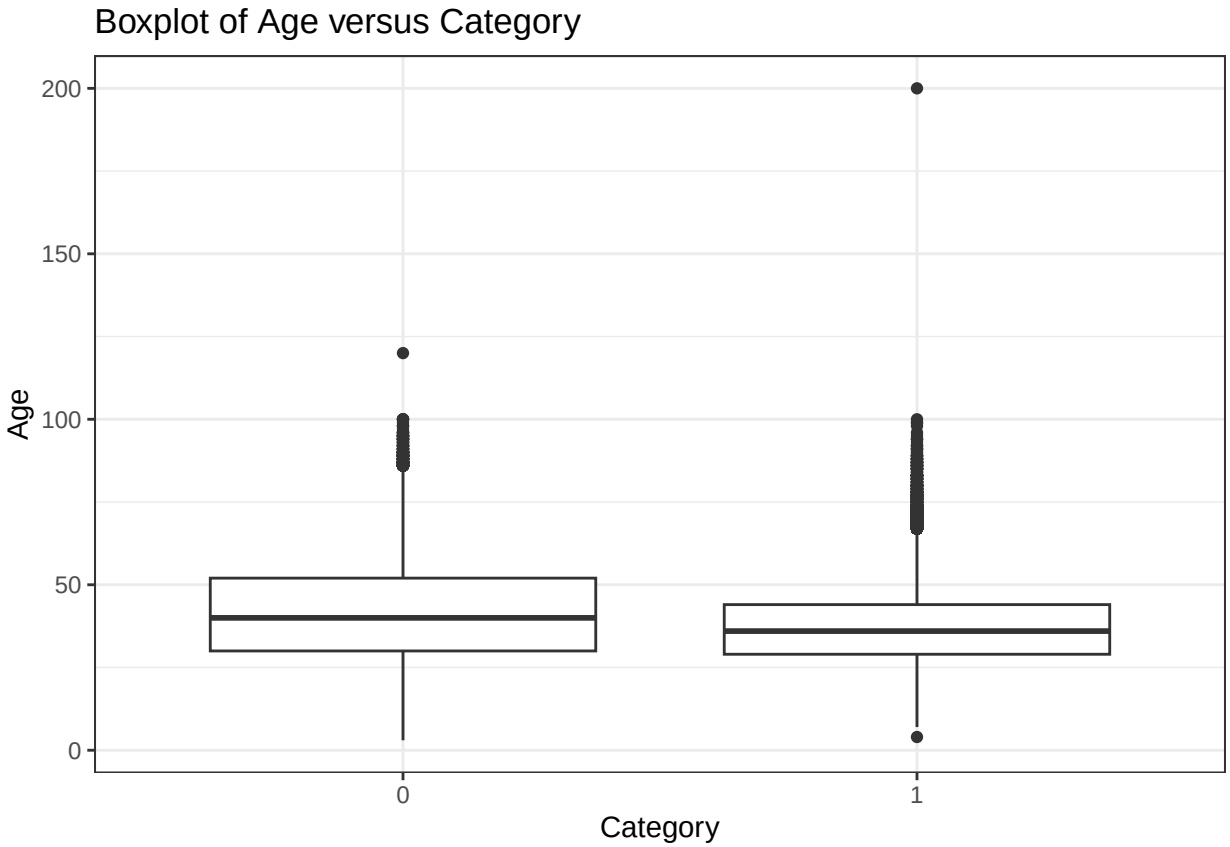
Assumption 3: There is no multicollinearity among the independent variables:
To asses this assumption we will run the variance inflation factor (VIF) on the model and see if any of these are greater than 5. Note that when running VIF on the logistic regression, we will instead get the generalized variance inflation factor (GVIF). To get VIF from GVIF we will use the following transformation: $\text{VIF} = \text{GVIF}^{2 \cdot \text{df}}$. Where *df* corresponds to the degrees of freedom of the corresponding variable.

```
##                                vif
## as.factor(Gender)        2.069591
## as.factor(MaritalStatus) 3.659636
## Age                      1.618510
## NumberInHousehold        1.303569
```

As can be seen from the above vif test, gender has a vif of 2.07, marital status has a vif of 3.66, age has a vif of 1.62, and lastly the number of individuals in the household has a vif of 1.3. Since all of these values are less than 5 we can see that this assumption is not violated.

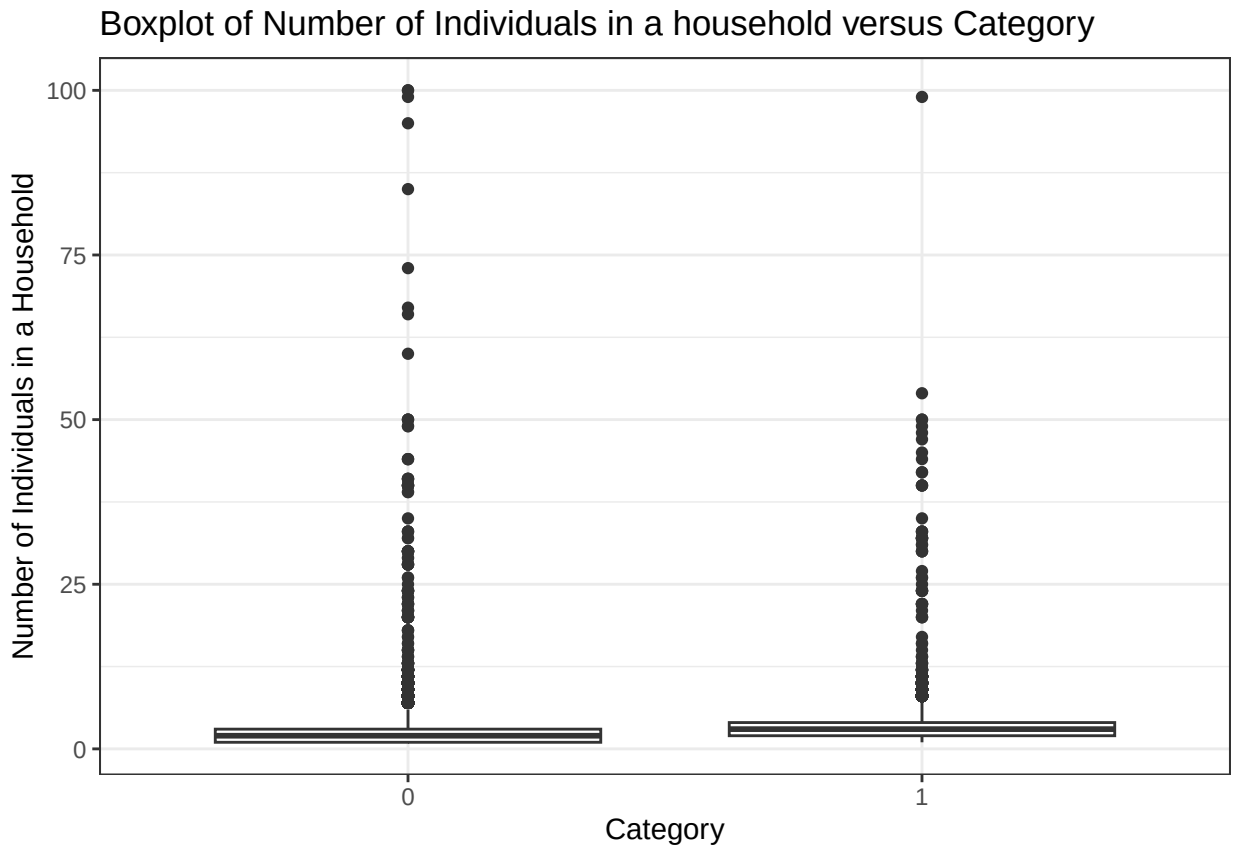
Assumption 4: There are no significant outliers:
To test this assumption we will create a boxplot of both of our numerical variables plotted against our dependent variable:

Below we will plot age versus category:



As can be seen there are plenty of influential points/outliers in the age variable. In particular these seem to be caused by incorrect/faulty data. Since no one is 200 years old.

Below we will plot the number of individuals in a household versus category:



As can be seen there are a seriously concerning amount of influential points/outliers in the number of individuals in a household variable. There are so many that it is nearly impossible to actually see the box in the above boxplot. In particular these seem to be caused by incorrect/faulty data.

Assumption 5: The sample size is large:
To test this assumption we will see if we have a minimum of 10 cases for the least frequent outcome for our categorical variables we will do this through the use of frequency tables.

Below is the frequency table of marital status and the category:

```
##
##     Married Separated Single
##   0   22732     22410  47117
##   1   20549     26685  24339
```

Below is the frequency table of gender and the category:

```
##
##     Female  Male Other
##   0  59738 31389  1132
##   1  53416 17689   468
```

Since both of these tables have well over 10 observations for every cell, it is safe to say that our sample size is sufficiently large.

Furthermore, we must run a chi-square test for each of these variables to see if the relationship is actually significant.

Below we will run a chi square test to check the relationship between marital status and the dependent variable category:

```
##
##  Pearson's Chi-squared test
##
## data:  regression_set$Category and regression_set$MaritalStatu
## X-squared = 5214.5, df = 2, p-value < 2.2e-16
```

As can be seen from the above chi-square test, since the p-value is practically zero we have significant evidence at the 5% level that there is a significant relationship between the marital status and the category variable.

Below we will run a chi square test to check the relationship between gender and the dependent variable category:

```
##
##  Pearson's Chi-squared test
##
## data:  regression_set$Category and regression_set$Gender
## X-squared = 1871, df = 2, p-value < 2.2e-16
```

As can be seen from the above chi-square test, since the p-value is practically zero we have significant evidence at the 5% level that there is a significant relationship between gender and the category variable.

Assumption 6: Linear relationship between the logit of the response and the independent variables:
The last assumption is that there needs to be a linear relationship between the logit of the response and the continuous independent variables. However, due to the fact that we don't yet have the machinery to check this assumption, we will skip it for now.

Conclusions:
As can be seen from the above assumptions, the only one that was violated was the no extreme outliers assumption. However, this assumption was severely violated to the point that we should question the validity of our model. With that being said, in the real world we would either have to use the model knowing its flawed, or try and fit a new model/refit the model with the outliers removed/dealt with.

Part 4: Conclusion:

Limitations:
The analysis of the dataset was subject to several limitations. First, the presence of missing values, especially data about the clients, was greater than expected. As a result, this could negatively impact analyzing the background of clients, which in turn can lead to misleading/incorrect generalization about certain clients as a whole. Removing missing values was one of the approaches for analyzing. However, this could result in reducing the scope of available information. Also, a lot of the time there can be a "story" to the missing values, and without questioning why the data points are missing in the first place you are missing out on key aspects of the data and limiting your analysis. Second, the dataset also contains unrealistic data, such as people who are 200 years old, negative income, and an unreasonable number of people in a household, etc. This unrealistic information may influence the accuracy of our findings and removing them will further reduce the available data. Furthermore, there is no way to be one hundred percent sure that a data point is incorrect, for example the household with 100 people could in fact be something like a foster home, etc. Lastly, several similar types of value under columns were combined to reduce redundancy during data analysis. For instance, there were 9 types of answers for marital status and 13 types of answers for gender.

To simplify the analyzing process, similar groups were combined and groups with small amounts of values or uninformative labels were dropped. The final limitation of our dataset/analysis is that not all of the model assumptions for our given model passed, which in turn puts in question the reliability of the model itself for prediction. Furthermore, since a large number of missing values and unrealistic values are removed, and some similar values are combined in groups, the current model proposed may not provide the most precise prediction. More assumptions need to be run to improve the model and prediction results.

Suggestions:
There are several suggestions we would make to improve the functionality of the data collection and our models. As mentioned above, based on the data collected, it was noticed that missing values are very common, which indicates that this data collection method is not the most effective one. First off, we would suggest improving the data collection process by providing clearer guidelines for labeling the data and instructions for conducting the survey. This can help reduce missing values and unrealistic data by ensuring that respondents provide accurate and reasonable entries. They could offer incentives to respondents who provide accurate and valid information that can motivate them to pay more attention to their responses. By doing these steps, the ABA will effectively reduce missing or unrealistic values in the data they collect. Furthermore, it would help to implement clear labels and documentation to the data For example, by clearly defining the range for value entries in every variable can help reduce unrealistic values before they are implemented into the data set and used in subsequent analysis.

Recommendations:
As seen in the first visualization, there are several areas in which a high concentration of clients exist (Texas, Florida, Illinois, and Indiana). These areas likely have a higher demand for pro-bono legal services, so perhaps the ABA should consider allocating more volunteers and resources into these regions. Also, as seen in the second visualization, several categories of legal questions are growing at rates higher than the others, the main one being questions about Family and Children, with a second growing category being Housing and Homelessness. This means that it may be worth considering weighing knowledge and skills related to these in-need categories in the training and hiring of law professionals and volunteers now and in the future to deal with these growing questions. Lastly, as seen in the third visualization, there are several ethnic communities that are more likely to get involved in the pro-bono system, the largest being Caucasian, with African American and Latino/Hispanic communities being somewhat smaller. It may be helpful to have volunteers that are familiar with these ethnic identities, so that they could better serve and understand their clients. This visualization is also a good way to understand some of the makeup of these communities, with the information that the largest part of each community is single, but most have similarly large proportions of both married and separated clients. It may be worth having volunteers who understand both the single and current or formerly married demographics and unique questions that come with either.