# Project 1 - Analysis of American Bar Association data

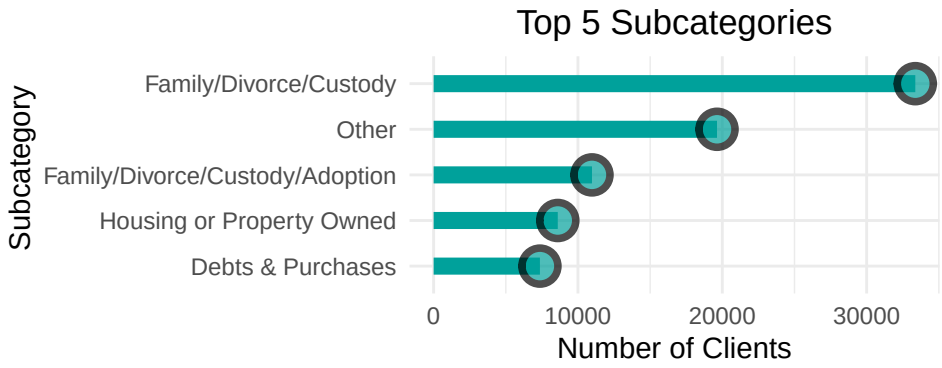
Daniel Fredin, Junhan Li, & Eric Chen

## Introduction:

The American Bar Association (ABA) offers free legal services throughout the United States through an online platform accessible in select states and territories. This platform enables eligible individuals to post legal inquiries and receive guidance from volunteer attorneys. The ABA aims to proactively anticipate the types of legal questions that arise in order to equip volunteers to address them effectively, identify the need for lawyers with specific expertise, and provide guidance to state partners based on prevailing trends.

Our goal is to provide guidance to the ABA regarding any recurring themes or emerging patterns observed in these interactions. This information would assist the ABA in advising its state partners, developing resources to address identified patterns, and formulating outreach strategies to effectively engage potential users and volunteers.

## Visualization 1: Investigating the Top 5 Subcategories of Asked Questions

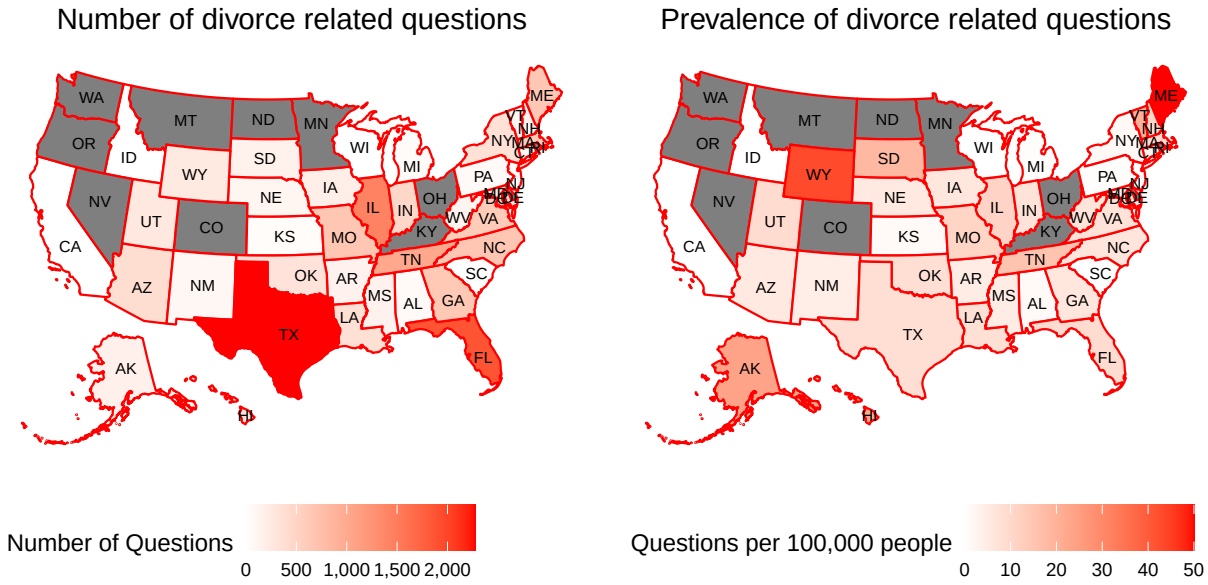



**Interpretation of visualization 1:**

Our project aims to begin by exploring the subcategories of legal questions that are frequently asked. This will enable us to identify the most common question type within the dataset we possess and use it as the central topic for our research inquiry.

According to our horizontal lollipop chart, it is evident that within the top 5 subcategories, two of the highest-ranking categories of inquiries made by clients on the online platform pertain to divorce. The subcategory "Family/Divorce/Custody" holds the highest occurrence among the top 5, with nearly double the number of clients asking questions compared to the second-ranking subcategory, "Other." This highlights the importance of adequately preparing volunteers to handle divorce-related queries.

# Visualization 2: US Map of Divorce Related Questions Distribution



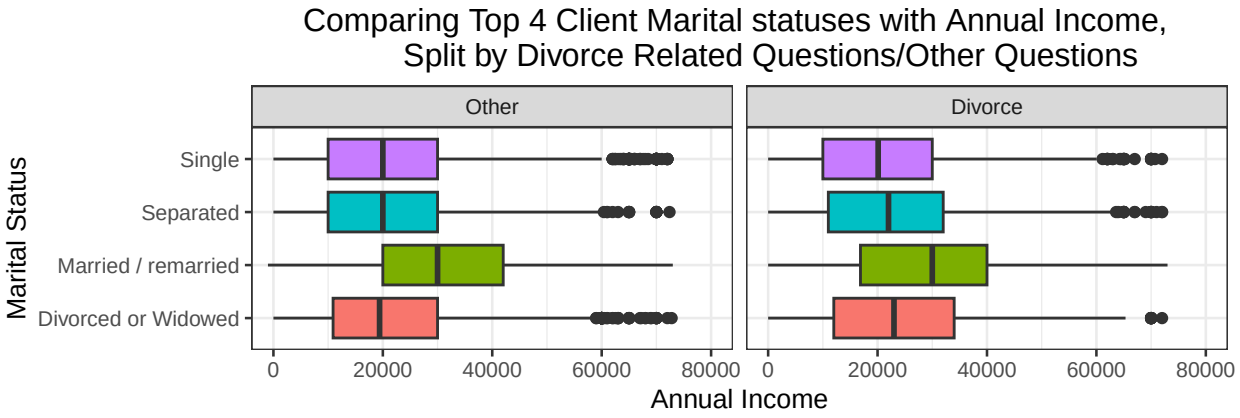

**Interpretation of visualization 2:**

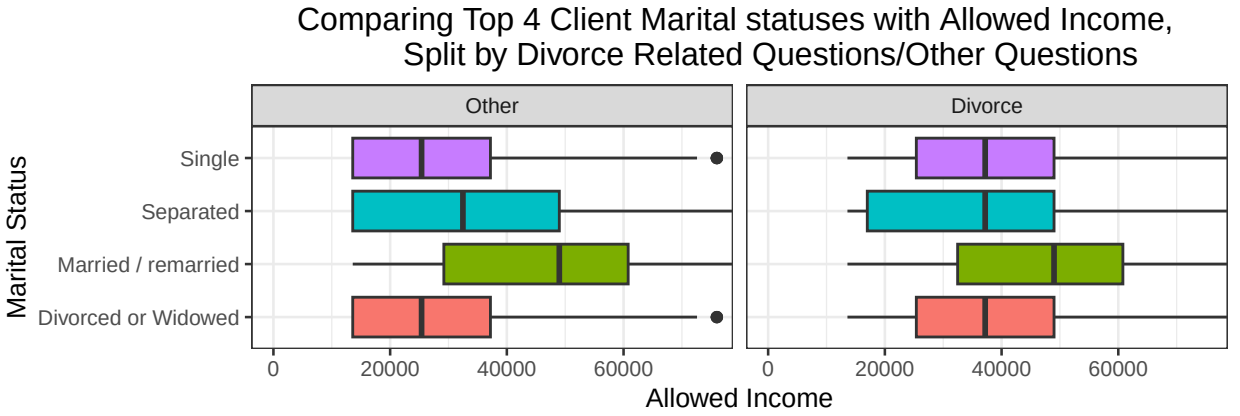
By analyzing the quantity of divorce-related inquiries and their distribution across states, we can gain insights into the clients' backgrounds and identify the regions with the highest occurrence of divorce-related questions.States like Washington, Oregon, Nevada, Montana, Colorado, North Dakota, Minnesota, Ohio, and Kentucky are depicted in gray on the chart due to legal requirements and confidentiality obligations. These states are prohibited from disclosing clients' information, including the specific category of legal questions they ask.

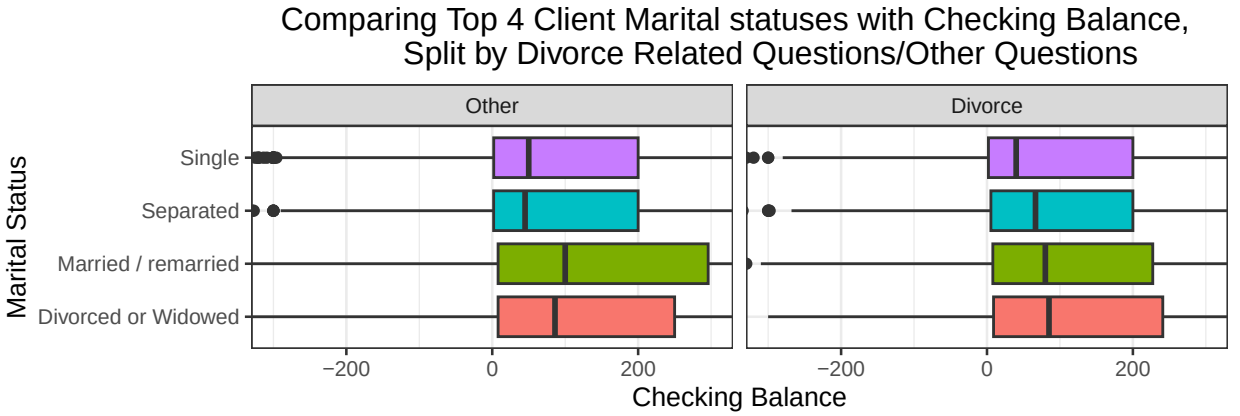
We recognize that the absence of data points for certain states poses a challenge to our model. We categorize these missing data points as Missing at Random (MAR) and suggest a solution by employing listwise deletion or available-case analysis. Using this approach, we exclusively consider cases with complete data for our analysis and omit the missing data.

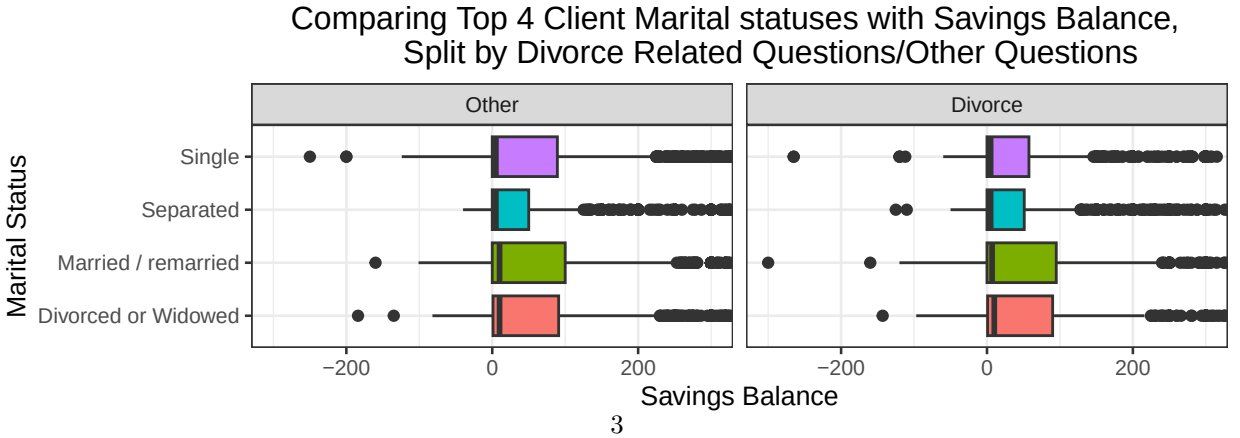
Among the 42 states that permit the revealing of client details, Texas, Florida and Indiana demonstrate the greatest volume of inquiries concerning divorce law. Nonetheless, the prevalence of divorce-related questions takes on a distinct pattern when analyzed differently. When considering the number of queries per 100,000 residents, Texas, Florida and Indiana no longer appear exceptional. Instead, it is Wyoming and Maine that emerge as prominent locations where the rate of individuals frequently seeking online guidance regarding divorce matters are the highest.

## Visualization 3: Financial Status Correlation with Marital Status



**Interpretation of visualization 3:**

By analyzing the financial status of clients in relation to their marital status, we can gain insights into whether their likelihood of asking divorce-related questions is influenced by their financial situation and relationship status. We filtered the data to contain only the 4 largest groups of marital status for comparison.

Based on the aforementioned visualization, it becomes apparent that clients seeking divorce-related advice generally have a higher average annual income compared to other clients. However, an exception arises within the married/remarried category, which exhibits even higher annual income than individuals in other marital statuses across both question categories. The states also acknowledge this difference and typically grant higher-income married individuals the opportunity to ask pro bono questions. In other words, the income threshold for asking questions free of charge is noticeably higher for those who are married or remarried in comparison to individuals in other marital statuses. Additionally, it is worth noting that if someone prefers not to disclose their marital status when asking a divorce-related question, their permitted income to inquire without charges decreases.

On average, clients seeking divorce-related guidance tend to have slightly lower checking and savings balances compared to clients with different types of inquiries. Among all marital statuses seeking divorce advice, individuals who are divorced possess the highest average checking balance, while those who are single or separated have the lowest. Regardless of the legal advice they seek, single and separated clients generally do not maintain high balances in their checking or savings accounts. In contrast to the data on annual/allowed income, married clients do not exhibit significantly higher checking or savings balances compared to other clients.

# Research question:

Is there significant predictive ability by assessing clients' sociodemographics to determine whether their legal question will be related to divorce?

The aim of this research investigation is to effectively predict whether a client will require legal support concerning divorce based solely on their sociodemographic factors. In this study, we suggest that the outcome variable is binary, indicating whether the client presents a question related to divorce or not, and our predictor variables will be determined through model fitting. Therefore, we intend to employ binary logistic regression for our analysis.

## Model creation & evaluation

**Comparison of Models**

To identify the optimal model, we assessed the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values associated with the three logistic regression models we constructed.

Based on the principle that model selection criteria strike a balance between goodness of fit and model complexity, it is known that AIC tends to favor more complex models, whereas BIC penalizes complexity more rigorously. Lower values of AIC or BIC indicate a better fit. By examining the chart, we note that the lowest AIC and BIC values correspond to our second model which included only the significant independent variables that were present in the initial model that contained all of the independent variables.

Therefore we selected model 2 as it had both the lowest AIC and BIC scores. AIC is a better indicator to use in answering our research question, as it maximizes the predictive power of the data for any future data. However, since both the AIC and BIC values agree on which model out of the three we created is the best, we can safely choose model 2 moving forward.

```
## The 3 models formulae:
```

```
## Model 1: Subcategory ~ Age + NumberInHousehold + MaritalStatus + AnnualIncome
##   + SavingsBalance + CheckingBalance + AllowedIncome + Ethnicity
##   + Gender + States

## Model 2: Subcategory ~ Age + NumberInHousehold + MaritalStatus + SavingsBalance
##   + CheckingBalance + Gender + States + Ethnicity

## Model 3: Subcategory ~ AnnualIncome + SavingsBalance + CheckingBalance
```

The three model's fit criteria:

```
## $Fit.criteria
##   Rank Df.res   AIC  AICc   BIC McFadden Cox.and.Snell Nagelkerke   p.value
## 1   17  50600 56260 56260 56420 0.095090      0.110200   0.155800 0.000e+00
## 2   14  45720 51250 51250 51380 0.093270      0.108800   0.153400 0.000e+00
## 3    4  50620 61940 61940 61990 0.003204      0.003924   0.005551 3.379e-43
```

**Summary of best model**

```
##
## Call:
## glm(formula = Subcategory ~ Age + NumberInHousehold + MaritalStatus +
##     SavingsBalance + CheckingBalance + Gender + States + Ethnicity,
##     family = "binomial", data = refined_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7269  -0.8312  -0.6448   1.1348   2.8943
##
## Coefficients:
##                                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)                             -5.758e-01  5.041e-02 -11.422  < 2e-16
## Age                                     -3.419e-02  9.727e-04 -35.147  < 2e-16
## NumberInHousehold                        6.792e-02  6.486e-03  10.473  < 2e-16
## MaritalStatusAll Other Marital Statuses  5.342e-01  6.201e-02   8.615  < 2e-16
## MaritalStatusDivorced or Widowed         1.049e+00  3.365e-02  31.181  < 2e-16
## MaritalStatusMarried / remarried         8.568e-01  2.929e-02  29.255  < 2e-16
## MaritalStatusSeparated                   2.163e+00  3.933e-02  54.998  < 2e-16
## SavingsBalance                          -1.457e-05  4.487e-06  -3.248 0.001160
## CheckingBalance                         -3.520e-05  8.697e-06  -4.048 5.17e-05
## GenderAll Other Genders                 -3.196e-01  8.328e-02  -3.837 0.000125
## GenderFemale                             3.834e-01  2.510e-02  15.275  < 2e-16
## StatesAll Other States                  -7.098e-02  2.364e-02  -3.003 0.002670
## EthnicityAfrican American               -1.021e-01  3.351e-02  -3.048 0.002303
## EthnicityAll Other Ethnicities           9.744e-02  2.466e-02   3.951 7.77e-05
##
## (Intercept)                             ***
## Age                                     ***
## NumberInHousehold                       ***
## MaritalStatusAll Other Marital Statuses ***
## MaritalStatusDivorced or Widowed        ***
## MaritalStatusMarried / remarried        ***
```

```
## MaritalStatusSeparated                  ***
## SavingsBalance                          **
## CheckingBalance                         ***
## GenderAll Other Genders                 ***
## GenderFemale                            ***
## StatesAll Other States                  **
## EthnicityAfrican American               **
## EthnicityAll Other Ethnicities          ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 56486  on 45730  degrees of freedom
## Residual deviance: 51218  on 45717  degrees of freedom
## AIC: 51246
##
## Number of Fisher Scoring iterations: 4
```

**Log-Odds regression equation**

$$\begin{aligned} logit(P(Divorce = 1)) &= \ln\left(\frac{P(Divorce = 1)}{1 - P(Divorce = 1)}\right) \\ &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 ... \beta_n X_n \end{aligned}$$

**Logistic Regression equation for the optimum model (Model #2):**

$$\begin{aligned} logit(P(Divorce = 1)) = &- 0.576 - (0.034 \cdot Age) + (0.068 \cdot NumberInHousehold) + (0.534 \cdot AllOtherMaritalStatuses) \\ &+ (1.049 \cdot DivorcedOrWidowed) + (0.857 \cdot Married/remarried) + (2.163 \cdot Separated) \\ &- (1.457 \times 10^{-5} \cdot SavingsBalance) - (3.520 \times 10^{-5} \cdot CheckingBalance) \\ &- (0.320 \cdot AllOtherGenders) + (0.383 \cdot Female) - (0.071 \cdot AllOtherStates) \\ &- (0.102 \cdot AfricanAmerican) + (0.097 \cdot AllOtherEthinicities) \end{aligned}$$

**Example of prediction:** A widowed African American female age 29, with 3 in her household from the state of Washington with a checking balance of $500 and a savings balance of $250 can be predicted as the following:

$$\begin{aligned} logit(P(Divorce = 1)) &= -0.576 - (0.034 \cdot 29) + (0.068 \cdot 3) + (0.534 \cdot 0) + (1.049 \cdot 1) + (0.857 \cdot 0) + (2.163 \cdot 0) \\ &- (1.457 \times 10^{-5} \cdot 250) - (3.520 \times 10^{-5} \cdot 500) - (0.320 \cdot 0) + (0.383 \cdot 1) - (0.071 \cdot 1) \\ &- (0.102 \cdot 1) + (0.097 \cdot 0) \\ &= -0.120 \end{aligned}$$

$$P(Divorce = 1) = \frac{e^{-0.12}}{1 + e^{-0.12}} = 0.47$$

```
## For the example, the probability for a divorce related to be asked is: 0.4348747
```

## Assumptions

Given that logistic regression was employed for our research inquiry, we assessed the presence of significant multicollinearity among the predictor variables while making the following assumptions:

- The dependent variable exhibits two distinct outcomes, namely divorce-related questions or non-divorce-related questions.
- Each observation in the dataset is independent and not a repeated measurement of the same client.
- A linear association exists between each predictor variable and the logit of the dependent variable.
- The dataset comprises a substantial number of samples.
- There are no extreme outliers or influential observations within the dataset.

### Testing for Multicollinearity

While conducting the multicollinearity assessment, it was noticed that the GVIF (Generalized Variance Inflation Factor) values for all our predictor variables were below 5. This suggests that we can safely assume the independence of observations in the dataset.

```
##                       GVIF Df GVIF^(1/(2*Df))
## Age                1.276382  1        1.129771
## NumberInHousehold  1.146192  1        1.070604
## MaritalStatus      1.652371  4        1.064789
## SavingsBalance     1.083504  1        1.040915
## CheckingBalance    1.090930  1        1.044476
## Gender             1.147914  2        1.035088
## States             1.007072  1        1.003530
## Ethnicity          1.079461  2        1.019299
```

### Independent and Dependent variables

We transformed the categorical variables into subsets to decrease the number of parameters utilized in our regression models. This transformation involved selecting a few factors with the highest frequency and grouping the remaining factors as a distinct category.

The MaritalStatus variable was modified to encompass 5 levels: “Single,” “Divorced or Widowed,” “Married / remarried,” “Separated,” and “All Other Marital Statuses.”

The StateAbbr variable was modified to have 2 levels: “Top 3 States” (consisting of FL, TX, and IN) and “All Other States.”

The Ethnicity variable was modified to have 4 levels: “Caucasian,” “African American,” “Not Hispanic or Latino,” and “All Other Ethnicities.”

The Gender variable was modified to have 3 levels: “Male,” “Female,” and “All Other Genders.”

In our optimal model, the independent variables that serve as the most significant predictors of the binary dependent variable “subcategory” (whether it pertains to divorce or not) are utilized. We considered “Single,” “Caucasian,” “Male,” and “Top 3 States” as our reference levels for the categorical variables.

The predictors deemed statistically significant with a confidence level of 5% are:

```
##  Age
##  NumberInHousehold
##  MaritalStatus(All Other Marital Statuses)
##  MaritalStatus(Divorced or Widowed)
```

```
##  MaritalStatus(Married / remarried)
##  MaritalStatus(Separated)
##  SavingsBalance
##  CheckingBalance
##  Gender(All Other Genders)
##  Gender(Female)
##  States(All Other States)
##  Ethnicity(African American)
##  Ethnicity(All Other Ethnicities)
```

## Odds Ratio and Confidence Intervals for optimum model

```
##                                                OR     2.5 %    97.5 %
## (Intercept)                                0.5622554 0.5093565 0.6206482
## Age                                        0.9663918 0.9645513 0.9682359
## NumberInHousehold                          1.0702849 1.0567658 1.0839770
## MaritalStatusAll Other Marital Statuses    1.7060930 1.5108512 1.9265652
## MaritalStatusDivorced or Widowed           2.8551300 2.6729220 3.0497589
## MaritalStatusMarried / remarried           2.3556060 2.2241952 2.4947808
## MaritalStatusSeparated                     8.6992505 8.0538128 9.3964141
## SavingsBalance                             0.9999854 0.9999766 0.9999942
## CheckingBalance                            0.9999648 0.9999478 0.9999818
## GenderAll Other Genders                    0.7264650 0.6170530 0.8552772
## GenderFemale                               1.4673025 1.3968627 1.5412943
## StatesAll Other States                     0.9314761 0.8893108 0.9756406
## EthnicityAfrican American                  0.9028999 0.8455040 0.9641921
## EthnicityAll Other Ethnicities             1.1023401 1.0503307 1.1569249
```

By considering "Top 3 States" as our reference states, "Single" as our reference marital status, "Caucasian" as our reference ethnicity and "Male" as our reference gender we can discern significant associations between the response variable and the corresponding covariates while keeping all other variables constant.

For example, in comparison to single clients, those who are separated (OR = 8.699, 95% CI = 8.054-9.397), Married / remarried (OR = 2.356, 95% CI = 2.224-2.495), and Divorced or Widowed (OR = 2.855, 95% CI = 2.673-3.050) all exhibited higher odds of inquiring about divorce-related matters. For instance, in terms of percentage change, the likelihood of a separated client posing divorce-related questions is approximately 770% greater than that of a single client.

The odds ratio (OR) for Age was found to be 0.966 (95% CI = 0.965-0.968). The confidence interval, which does not overlap with 1, indicates a significant variation in the likelihood of clients asking questions about divorce based on their age. Specifically, for each unit increase in a client's age, the odds of them inquiring about divorce decrease by 3.40%.

In comparison to male clients, those who are female exhibited higher odds (OR = 1.467, 95% CI = 1.397-1.541) of inquiring about divorce-related matters. In terms of percentage change, the likelihood of a female client posing divorce-related questions is approximately 46.7% greater than that of a male client.

The odds ratio (OR) for Number in Household was found to be 1.070 (95% CI = 1.057-1.084). The confidence interval, which does not overlap with 1, indicates a significant variation in the likelihood of clients asking questions about divorce based on number of persons in their household. Specifically, for each unit increase in a client's Number in Household, the odds of them inquiring about divorce increase by 7%.

In comparison to Caucasian clients, those who are African American exhibited lower odds (OR = 0.903, 95% CI = 0.846-0.964) of inquiring about divorce-related matters. In terms of percentage change, the likelihood of a African American client posing divorce-related questions is approximately 9.7% less than that of an Caucasian client.

The odds ratio (OR) for savings balance (OR = 0.999, 95% CI = 0.999-0.999) and checking balance (OR = 0.999, 95% CI = 0.999-0.999) were found to be nearly identically close to 1. Therefore it does not indicate a significant variation in the likelihood of clients asking questions about divorce based on their savings and checking balances.

In comparison to clients from the top 3 states (FL, TX, or IN), those who are located in any other State exhibited lower odds (OR = 0.931, 95% CI = 0.889-0.976) of inquiring about divorce-related matters. In terms of percentage change, the likelihood of a client located in the other states posing divorce-related questions is approximately 6.90% less than that of a client located in one of the top 3 states (FL, TX, or IN).

**Accuracy of Best Model**

```
##
##      FALSE  TRUE
##   0 29730  1910
##   1 10973  3118
```

```
##
## The accuracy of our best model was: 71.83 %.
```

The accuracy classification table presented indicates that around 72% of the fitted values were accurately classified, as calculated by adding the number of correctly classified values (29730) to the number of false negatives (3118), and dividing it by the total number of observations (45731). In terms of accuracy, this implies that our model performs significantly better than a random guess. Randomly guessing the correct classification for divorce-related questions would yield a 50% chance of being accurate. However, our model surpasses this baseline, demonstrating a higher level of accuracy.

# Summary & Conclusions

Based on the odds ratios obtained from our top-performing model, it is evident that clients hailing from states other than the Top 3 States, Florida, Texas, or Indiana, exhibit slightly less likelihood of inquiring about divorce-related matters in comparison to our chosen reference state groups, Top 3 States. The odds ratio for All Other States, slightly exceeding 0.93, indicates that clients from states other than the top 3 have approximately a slightly lower likelihood of asking a divorce-related question compared to clients from Top 3 states, holding all other variables constant.

Apart from the state of origin, the variables of age, family size, gender, ethnicity, and marital status exhibit significant predictive power in determining whether a client will inquire about divorce-related matters. According to the odds ratio, for each passing year, the likelihood of a client's question being related to divorce decreases by approximately 0.96, holding all other variables constant. All marital statuses exhibit odds ratios greater than 1, indicating an increased likelihood of divorce-related questions compared to our reference group, "single." Notably, being "separated" raises the odds by nearly a factor of 8.7, while holding all other variables constant. This aligns with the expectation that clients seeking divorce advice are more likely to be in strained relationships, while being single suggests an absence of a relationship altogether.

When comparing the gender of clients, we can confirm that female clients, in relation to our reference gender "Male," have a slightly higher odds ratio of 1.46. This indicates an increase in the likelihood of asking a divorce-related question by a factor of approximately 1.5. Regarding the ethnicity of clients, African American clients have a slightly lower odds ratio of 0.9 compared to our reference ethnicity "Caucasian," while clients of all other ethnicities have a slightly higher odds ratio of 1.1. Neither ethnic group shows a significant difference from an odds ratio of 1, indicating minimal variation between the selected variable and the corresponding outcome. In conclusion, we did not observe any predictive capability concerning a client's

checking and savings accounts, as both their odds ratios were very close to 1, specifically 0.99. This suggests that there is no discernible difference between the selected variable and the corresponding outcome.

The key lesson we learned from this assignment is that statistical computing has the potential to be applied in professional domains for predicting trends and effectively allocating resources. Going beyond the project's requirements and outcomes, we believe the American Bar Association could employ the models we developed to determine the states requiring more divorce lawyers and collaborate with state governments accordingly. Another crucial lesson we learned from this project is the significance of establishing a clear objective when dealing with unorganized data. Given the extensive volume of data at our disposal, it becomes even more crucial to formulate specific research questions early on and filter the data accordingly. The majority of the datasets contain far more information than what is necessary for drawing significant conclusions.

However, the most significant constraint we faced in exploring our research question was the insufficient data from certain states and the apparent disparity in data availability among other states. In the dataset obtained from the American Bar Association pro bono service, we discovered that a select few states, precisely nine, lacked the provision of clients' demographic information. We encountered another limitation during our research. We found that the states of California, Idaho, South Carolina, and Wisconsin did not classify the client's legal questions related to divorce into specific subcategories. Instead, they grouped them under broader categories like Family and Children. Consequently, the level of interest among clients in those states regarding divorce-related information remains indeterminable. Nonetheless, considering the absence of client demographics or divorce-related data from these states, we opted for listwise deletion as a method to handle the missing data.

We acknowledge the significant challenge posed by the absence of data points for certain states in our model. These missing data points were classified as Missing at Random (MAR), and we proposed two solutions: listwise deletion or available-case analysis. With these approaches, we considered only cases with complete data, excluding the missing data from our analysis. An alternative approach would be imputation, which involves replacing missing data with substituted values. If we had implemented imputation, regression imputation would have been a sensible choice. This method involves running a regression using the available data to estimate the missing values, thereby increasing the sample size and reducing the standard error. Subsequently, the entire dataset would be reanalyzed. Despite the limitations imposed by the missing data, we are confident in the significant predictive capability of our model for clients residing in states where client demographics and categorical divorce data were available. Overall, our model exhibits approximately 72% accuracy in predicting outcomes for approximately 74% of the United States, providing us with confidence in its reliability.

Another potential strategy to address the constraints mentioned earlier involves incorporating predictor variables that encompass the entirety of the United States, instead of solely relying on data from particular states. This approach would enable us to accurately predict the client's legal inquiries nationwide, without the need to specify the states for which we possess accurate predictions. Another option could be for the American Bar Association (ABA) to mandate that states provide their data within a nationwide framework, guaranteeing consistent consideration of demographics and legal query categories.

After examining various scientific literature, we were unable to find any studies that reported similar findings related to our research question. However, we did successfully explore different sets of literature concerning the factors that contribute to accurately predicting divorce. This exploration is relevant to our goal of understanding our predictive power in divorce-related legal matters. Specifically, the National Institutes of Health (NIH) published an article discussing the risks of predicting divorce without conducting proper cross-validation analyses and sensitivity tests. They concluded that "exceptional initial predictive results can assist us in enhancing models by identifying significant risk factors" [1]. This conclusion allows us to better evaluate the accuracy of our predictive model. Although we achieved a 71.83% success rate in predicting divorce-related queries from clients, it is important to conduct thorough cross-validation and sensitivity testing to avoid overestimating the model's predictive capabilities. Instead, we should focus on the model's ability to identify crucial and meaningful predictors.

[1] Heyman RE, Smith Slep AM. The Hazards of Predicting Divorce Without Crossvalidation. J Marriage Fam. 2001 May;63(2):473-479. doi: 10.1111/j.1741-3737.2001.00473.x.