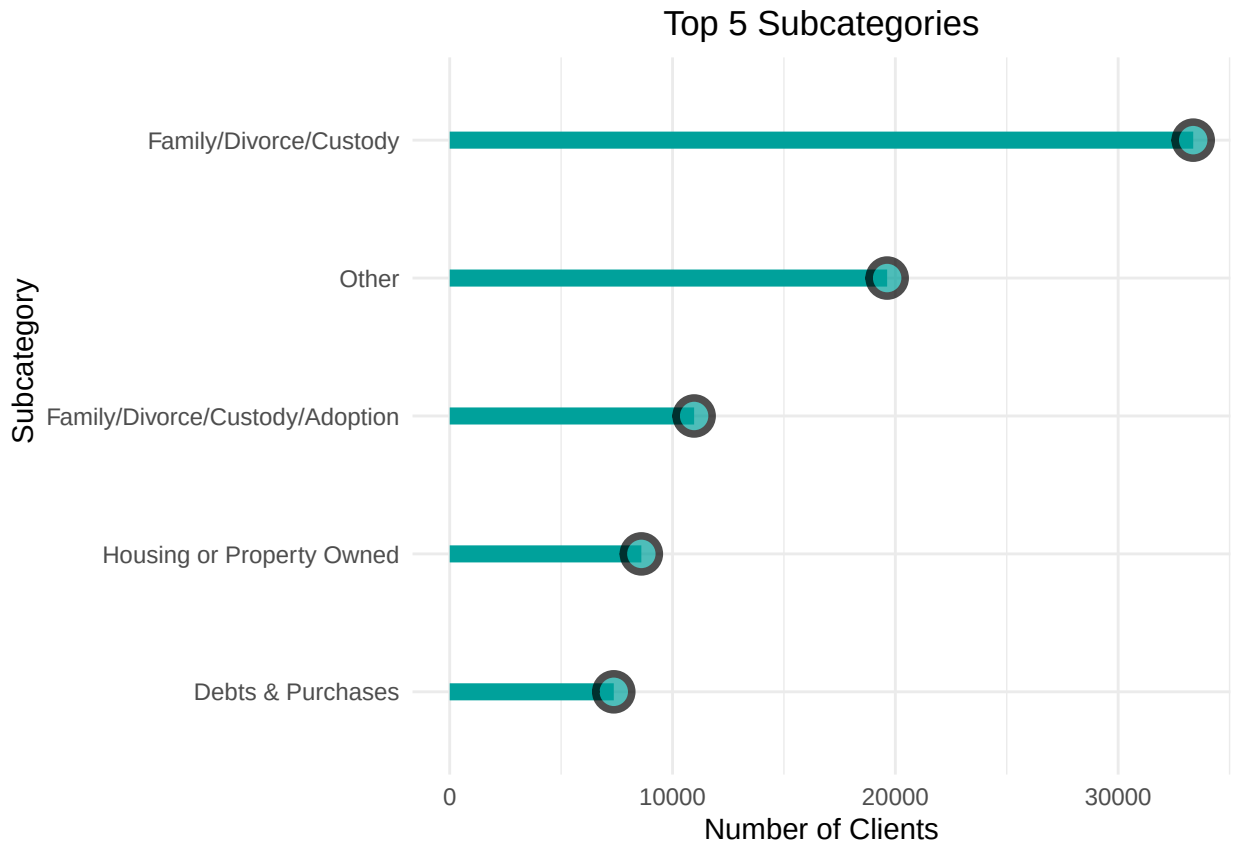
Project 1 - Analysis of American Bar Association data

Eric Chen, Junhan Li, & Daniel Fredin

Visualization 1: Investigating the Top 15 Subcategories of Asked Questions



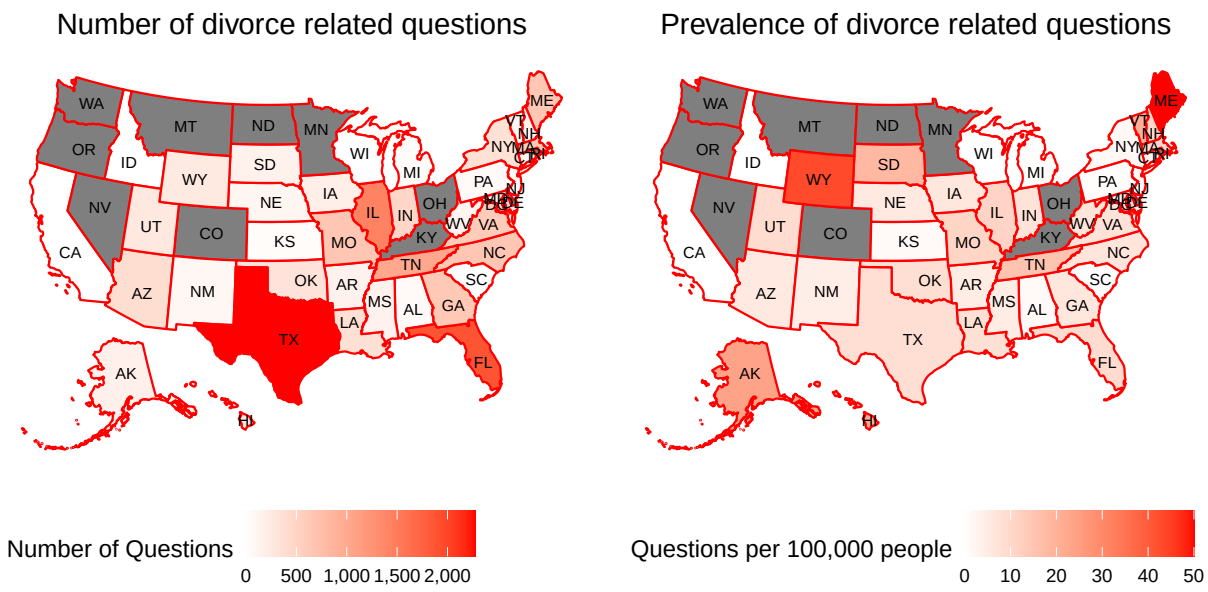
Interpretation of visualization 1:

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 15 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 15, 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.

We find this situation fascinating as it prompts us to seek answers to inquiries like: "What are the key factors influencing divorce?", "Does clients' background influence their inclination to ask divorce-related questions on the ABA online platform?", and, importantly, "How can we adequately train our volunteers to handle these divorce-related inquiries?"

Visualization 2: US Map of Divorce Related Questions Distribution



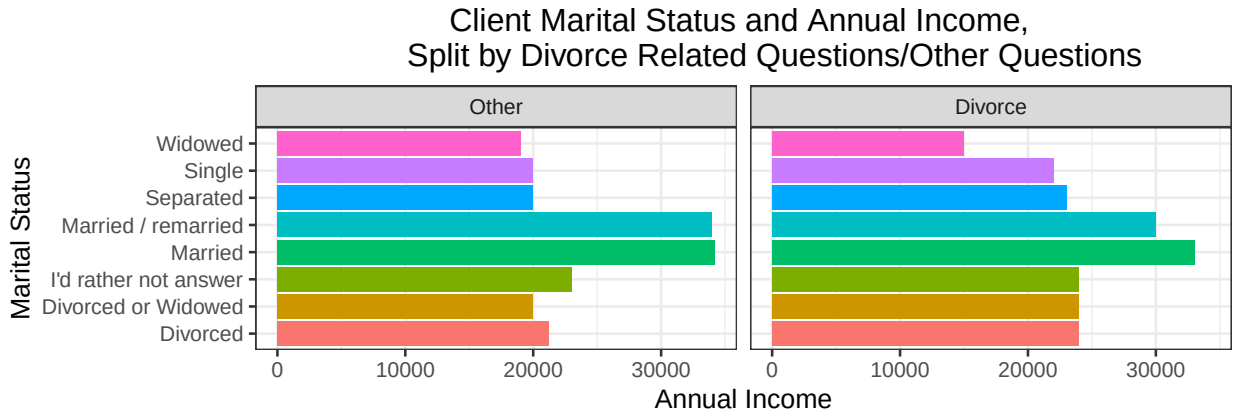
Interpretation of visualization 2:

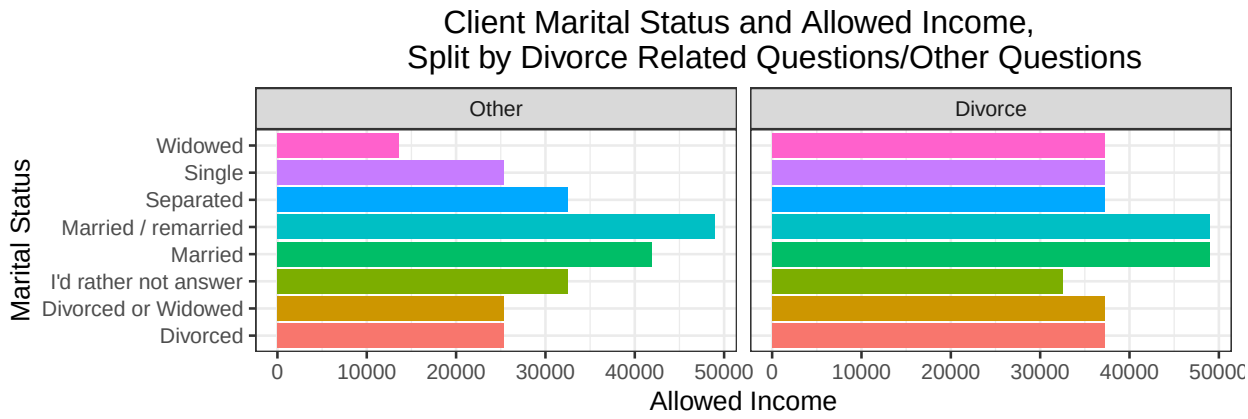
By analyzing the quantity of divorce-related inquiries and their distribution across states, we can gain insights into the clients' backgrounds and identify the regions with the highest occurrence of divorce-related questions.

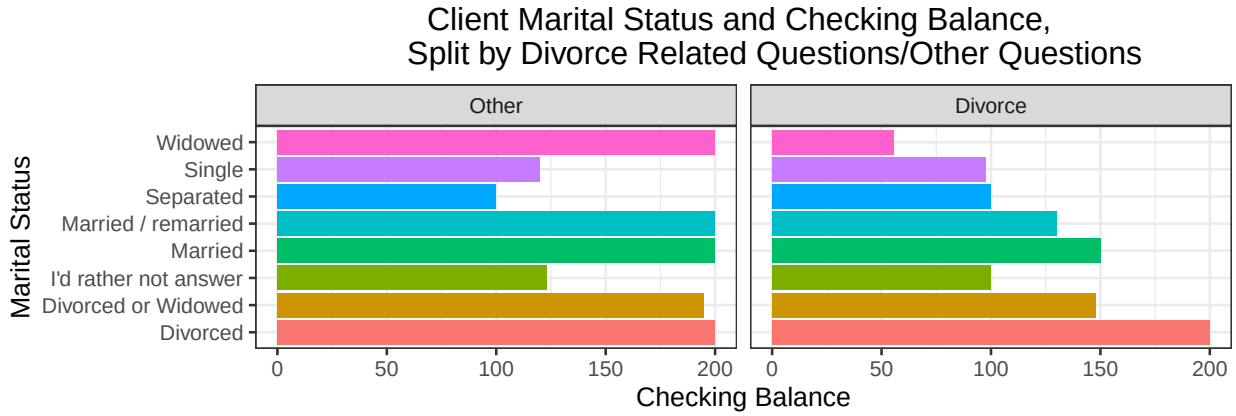
States like Washington, Oregon, Nevada, Montana, Colorado, North Dakota, Minnesota, Ohio, and Kentucky are depicted in gray on the chart due to legal requirements and confidentiality obligations. These states are prohibited from disclosing clients' information, including the specific category of legal questions they ask.

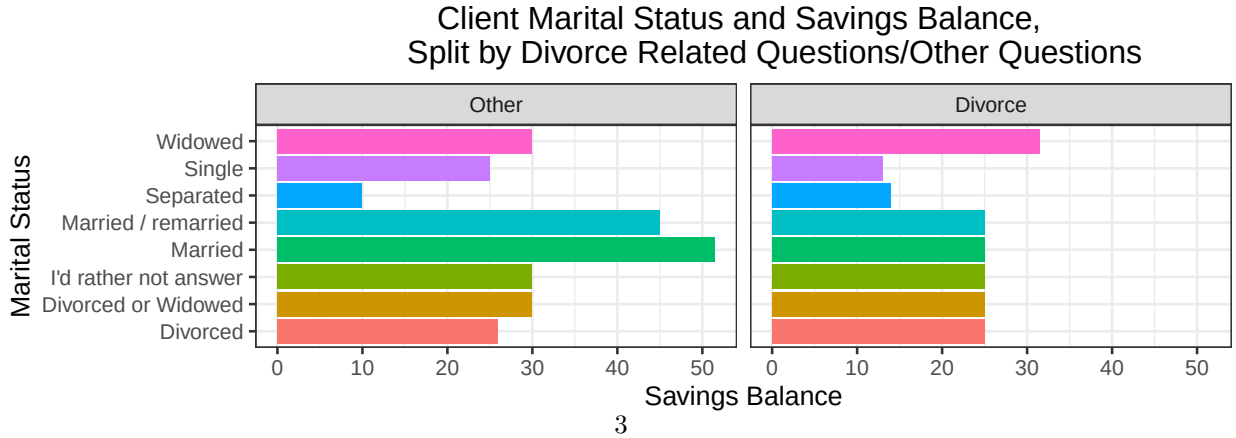
Among the 42 states that permit the revealing of client details, Texas and Florida demonstrate the greatest volume of inquiries concerning divorce law. Nonetheless, the prevalence of divorce-related questions takes on a distinct pattern when analyzed differently. When considering the number of queries per 100,000 residents, Texas and Florida 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

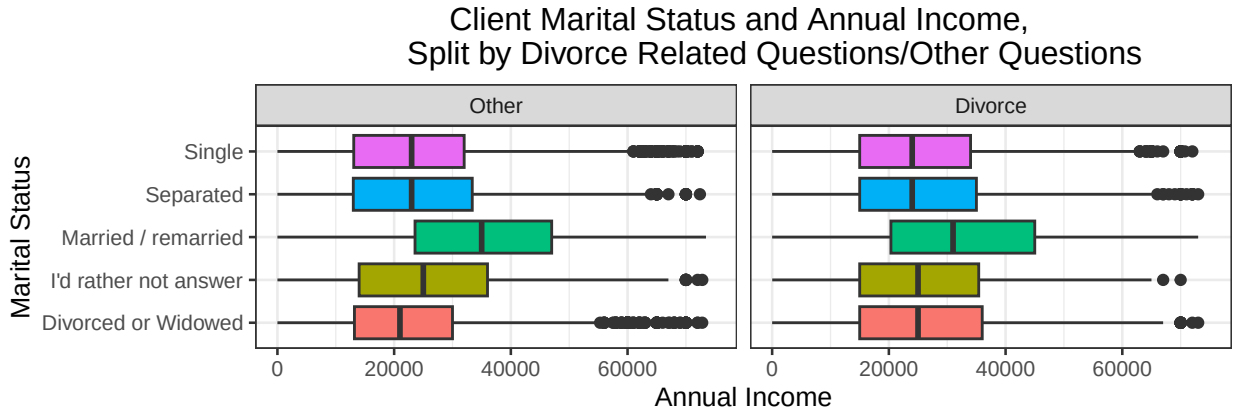


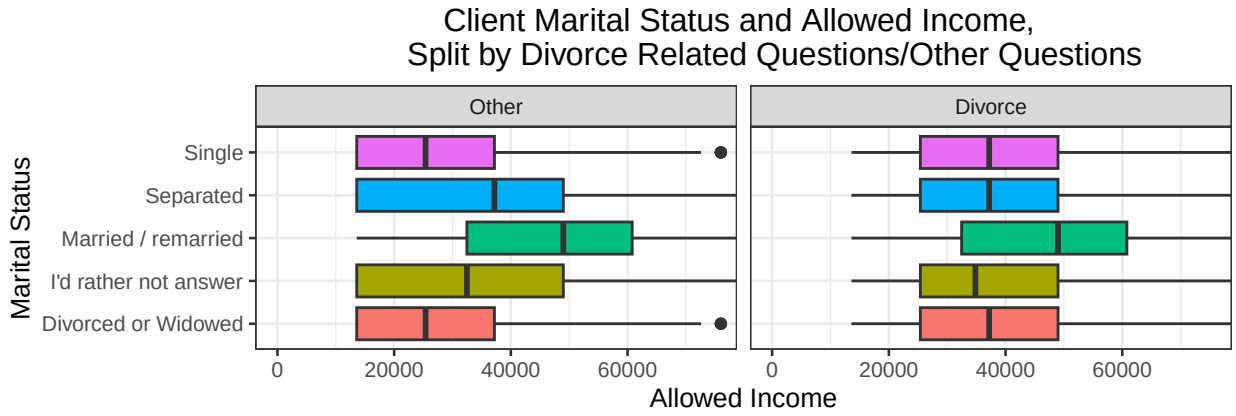


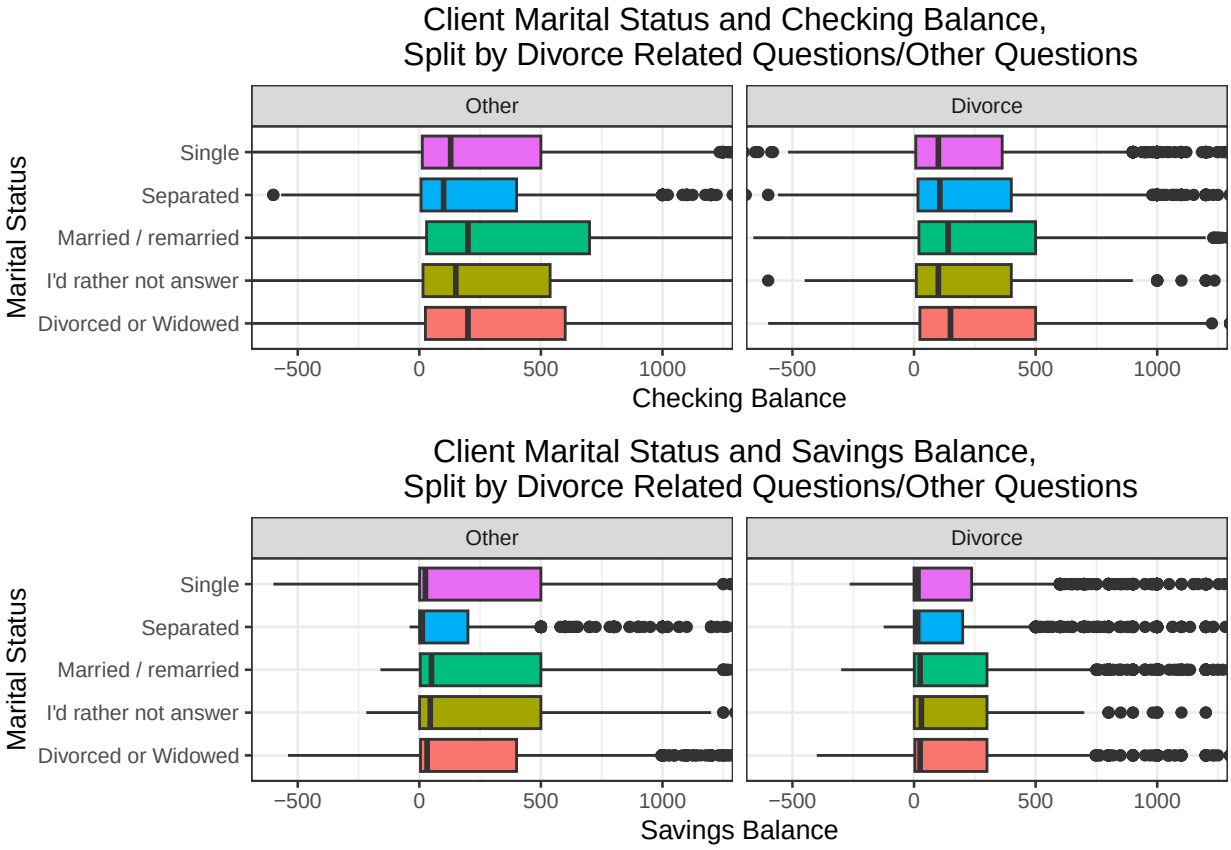




###############TESTING VISUALIZATION 3







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.

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 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 widowed have the lowest. This observation is intriguing as both statuses imply that the person does not currently have a partner. Clients with similar statuses, such as single or separated, have checking balances that fall between those of widowed and divorced individuals. However, regardless of the legal advice they seek, single and separated clients generally do not maintain high balances in their 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, except for the savings balance of clients seeking non-divorce-related legal advice.

Research question:

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

The objective of this research inquiry is to accurately forecast whether a client will seek legal assistance related to divorce or another area based solely on their background information, encompassing finances, relationship status, and residential location. In this study, we propose that the dependent variable is binary, representing whether the client poses a divorce-related question or not.

Model creation & evaluation

```
# 2nd Model 2: Using LASSO

y <- model_data$Subcategory
x <- data.matrix(model_data[, c("Age", "NumberInHousehold", "MaritalStatus", "AnnualIncome", "SavingsBa

library(glmnet)
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

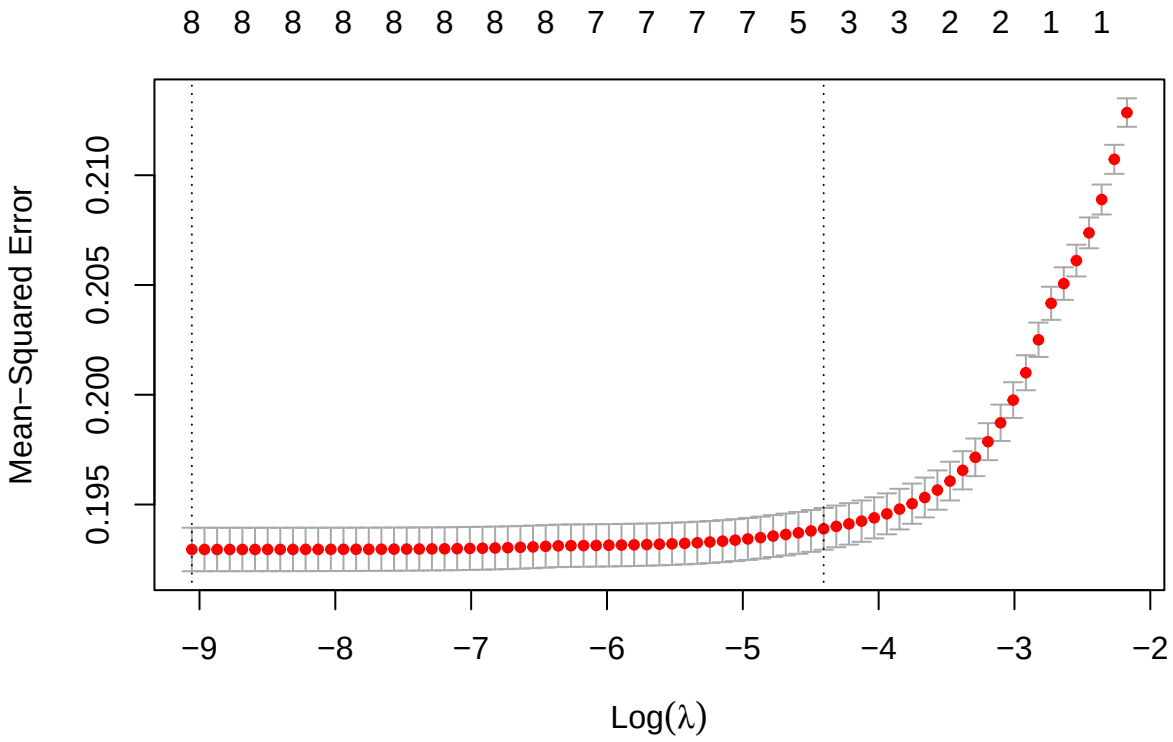
```
## Loaded glmnet 4.1-7
```

```
#perform k-fold cross-validation to find optimal lambda value
cv_model <- cv.glmnet(x, y, alpha = 1)

#find optimal lambda value that minimizes test MSE
best_lambda <- cv_model$lambda.min
best_lambda
```

```
## [1] 0.000116644
```

```
#produce plot of test MSE by lambda value
plot(cv_model)
```




```
#find coefficients of best model
best_model <- glmnet(x, y, alpha = 1, lambda = best_lambda)
coef(best_model)
```

```
## 9 x 1 sparse Matrix of class "dgCMatrix"
##                                s0
## (Intercept)         2.961272e-01
## Age                -5.681465e-03
## NumberInHousehold   3.136890e-02
## MaritalStatus       1.209292e-01
## AnnualIncome       -5.512443e-07
## SavingsBalance     -1.939704e-06
## CheckingBalance    -4.886939e-06
## AllowedIncome      -1.677883e-06
## GroupedStateAbbr   -8.165531e-03
```

Comparison of Models

To identify the optimal model, we assessed the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values associated with the five 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 initial model, which included all the independent variables

discussed in our research question. Therefore we selected model 1 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 five we created is the best, we can safely choose model 1 moving forward.

```
model_comparison[1]
```

```
## $Models
##   Formula
## 1 "Subcategory ~ Age + NumberInHousehold + MaritalStatus + AnnualIncome + SavingsBalance + CheckingBa
## 2 "Subcategory ~ Age + NumberInHousehold + MaritalStatus + SavingsBalance + CheckingBalance + Allowe
## 3 "Subcategory ~ AnnualIncome + SavingsBalance + CheckingBalance"
```

```
cat("Models Formulas:\n\n")
```

```
## Models Formulas:
```

```
cat("Model 1: Subcategory ~ Age + NumberInHousehold + MaritalStatus + AnnualIncome + SavingsBalance + C
```

```
## Model 1: Subcategory ~ Age + NumberInHousehold + MaritalStatus + AnnualIncome + SavingsBalance + Che
```

```
cat("Model 2: Subcategory ~ \n")
```

```
## Model 2: Subcategory ~
```

```
cat("Model 3: Subcategory ~ \n")
```

```
## Model 3: Subcategory ~
```

```
## $Fit.criteria
##   Rank Df.res   AIC  AICc   BIC McFadden Cox.and.Snell Nagelkerke   p.value
## 1   11  46400 52020 52020 52130 0.092180      0.107500   0.151700 0.000e+00
## 2    9  46410 52020 52020 52110 0.092130      0.107500   0.151600 0.000e+00
## 3    4  46410 57090 57090 57140 0.003439      0.004235   0.005974 9.405e-43
```

Logistic Regression for our best model: Log odds of asking question related to divorce = -0.04103 * Age - 1.849e-05 * SavingsBalance - 1.214e-07 * AnnualIncome - 3.932e-05 * CheckingBalance + 2.987e-02 + (coefficient of StateAbbr) + (coefficient of MaritalStatus)

Note: There are multiple cases of regression depend on the reference group of the factored categorical variables StateAbbr and MaritalStatus. ###############

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

During the multicollinearity assessment, it was observed that two of our independent variables, namely NumberInHousehold and AllowedIncome, exhibited a strong correlation. The GVIF values for these variables exceeded 5, indicating the need for their exclusion. This finding aligns with expectations, as the state's permitted income typically depends on the number of dependents residing in a household.

```
##                         GVIF Df GVIF^(1/(2*Df))
## Age                 1.281621  1        1.132087
## NumberInHousehold 12.715208  1        3.565839
## MaritalStatus       1.458208  3        1.064886
## AnnualIncome        1.155224  1        1.074814
## SavingsBalance      1.087567  1        1.042865
## CheckingBalance     1.091217  1        1.044613
## AllowedIncome      12.567627  1        3.545085
## GroupedStateAbbr    1.101278  1        1.049418
```

Fixing Multicollinearity

Upon exclusion of the independent variable "AllowedIncome," we observe a multicollinearity level of less than 5, indicating the absence of multicollinearity within our model.

```
##                         GVIF Df GVIF^(1/(2*Df))
## Age                 1.280657  1        1.131661
## NumberInHousehold 1.223605  1        1.106167
## MaritalStatus      1.455736  3        1.064585
## AnnualIncome       1.154768  1        1.074601
## SavingsBalance     1.087311  1        1.042742
## CheckingBalance    1.091455  1        1.044727
## GroupedStateAbbr   1.002813  1        1.001405
```

Summary of best model

By considering Florida as the state of residence reference and being single as the marital status reference, we can determine the predictors that positively and negatively influence the model.

```
##                NumberInHousehold MaritalStatusDivorced or Widowed
##                        0.07870526                       1.10794988
## MaritalStatusMarried / remarried           MaritalStatusSeparated
##                        0.85700785                       2.18503955

##                       (Intercept)                              Age
##                     -2.897039e-01                    -3.494773e-02
##                      AnnualIncome                    SavingsBalance
##                     -7.471693e-07                     -1.388220e-05
##                   CheckingBalance GroupedStateAbbrAll Other States
##                     -3.810806e-05                    -7.413201e-02
```

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

```
## [1] "(Intercept)"                       "Age"
## [3] "NumberInHousehold"                 "MaritalStatusDivorced or Widowed"
## [5] "MaritalStatusMarried / remarried" "MaritalStatusSeparated"
## [7] "SavingsBalance"                    "CheckingBalance"
## [9] "GroupedStateAbbrAll Other States"
```

Odds ratio and confidence intervals for the best model

By considering Florida as the reference state and Single as the reference marital status, 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 exhibited considerably higher odds (OR = 12.9, 95% CI = 11.9-14.1) of inquiring about divorce-related matters. In terms of percentage change, the likelihood of a separated client posing divorce-related questions is approximately 1190% greater than that of a single client.

The odds ratio (OR) for Age was found to be 0.956 (95% CI = 0.958-0.962). 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 4.40%.

For Number in Household, the odd ratio is 1.09, with 95% CI being 1.07 and 1.10. The confidence interval did not cross 1, which means there is a difference between clients with different number of people in household, so it's significant for our prediction. This means that the odds of client asking question related to divorce will increase by 9% for every unit increases of number of people in client's household with true population effect between 7% and 10%.

For Annual Income, the odd ratio is 0.999, with 95% CI being 0.999 and 1.00. The confidence interval crossed 1, which means there is no difference between clients with different annual income so it's not significant for our prediction.

For checking balance, the odd ratio is 0.99, with 95% CI being 0.99 and 0.99. The confidence interval did not cross 1, which means there is difference between clients of different checking balance when it comes to the odds of asking questions related to divorce. The influence of checking balance on the odds of clients asking divorce related question is equally weak as that of saving balance.

For state, let's take Hawaii(StateAbbrHI) as an example, the odds ratio is 2.21 with 95% CI being 1.85 to 2.64. The confidence interval did not cross 1, which means there is difference between clients from Hawaii and different states when it comes to the odds of asking questions related to divorce. This means that the odds for clients from the state of Hawaii to ask question related divorce is 2.21 times more likely than the clients from other states, with the true population effect between 1.85 to 2.64.

On the other hand, the state Georgia(StateAbbrGA), the odds ratio is 0.95 with 95% CI being 0.84 to 1.06. The confidence interval did cross 1, which means there is no difference between clients from Georgia and other different states when it comes to the odds of asking questions related to divorce.

```
##                                        OR     2.5 %    97.5 %
## (Intercept)                     0.7484851 0.6851132 0.8177188
## Age                             0.9656559 0.9638288 0.9674864
## NumberInHousehold               1.0818854 1.0677051 1.0962541
## MaritalStatusDivorced or Widowed 3.0281440 2.8397799 3.2290024
## MaritalStatusMarried / remarried 2.3561003 2.2265553 2.4931826
## MaritalStatusSeparated          8.8910002 8.2641107 9.5654435
## AnnualIncome                    0.9999993 0.9999982 1.0000003
## SavingsBalance                  0.9999861 0.9999775 0.9999948
## CheckingBalance                 0.9999619 0.9999444 0.9999793
## GroupedStateAbbrAll Other States 0.9285491 0.8867705 0.9722960
```

Accuracy of Best Model

Based on the classification table below, our optimal model accurately predicted 32,840 out of 50,620 total observations.

```
##
##     FALSE  TRUE
##   0 30376  1771
##   1 11220  3049
```

```
## [1] "The accuracy of our best model was: 72.01%."
```

Summary & Conclusions

Based on the odds ratios obtained from our top-performing model, it is evident that clients hailing from South Dakota, Maine, and Hawaii exhibit approximately double the likelihood of inquiring about divorce-related matters in comparison to our chosen reference state, Florida. We selected Florida as our reference state due to its significant population, social and economic influence, while having a relatively modest frequency of divorce-related queries. By excluding states without any recorded divorce-related questions, it becomes apparent that Indiana and Nebraska have notably lower rates of such inquiries compared to other states. The odds ratio, slightly exceeding 0.2, indicates that clients from these two states have approximately one-fifth the likelihood of asking a divorce-related question compared to clients from Florida, holding all other variables constant. It is important to note that although numerous states provide client data, many lack sufficient information to draw statistically significant conclusions. However, none of the aforementioned states fall into this category.

Apart from the state of origin, the variables of age, family size, 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, except for "widow," 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 13, 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. In conclusion, the balances of both checking and savings accounts have a strong predictive influence on the category of questions a client may ask. Conversely, we did not observe any predictive capability regarding a client's annual income.

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. While our data was limited to lower-income individuals and pro bono attorneys, we suspect that the trends we identified are indicative of other income brackets and paid attorneys on a broader scale. 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.

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 datasets contain far more information than what is necessary for drawing significant conclusions. In the dataset obtained from the American Bar Association pro bono service, we discovered that a select few states, precisely ten, lacked the provision of clients' demographic information. Nevertheless, since we have

excluded those states from our dataset, we can confidently affirm that we possess significant predictive power for clients residing in states that did provide client demographics.

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. As a result, we couldn't determine the extent to which clients in those states sought divorce-related information. However, since these states didn't provide any divorce-related data, similar to the previous limitation, we can assert with confidence that our model has significant predictive capability for clients in states where categorical data on divorces was available. Overall, we are confident that our model can accurately predict outcomes with approximately 75% accuracy for around 72% of the United States.

A 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 crossvalidation 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 74.75% success rate in predicting divorce-related queries from clients, it is important to conduct thorough crossvalidation 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. PMID: 17066126; PMCID: PMC1622921.