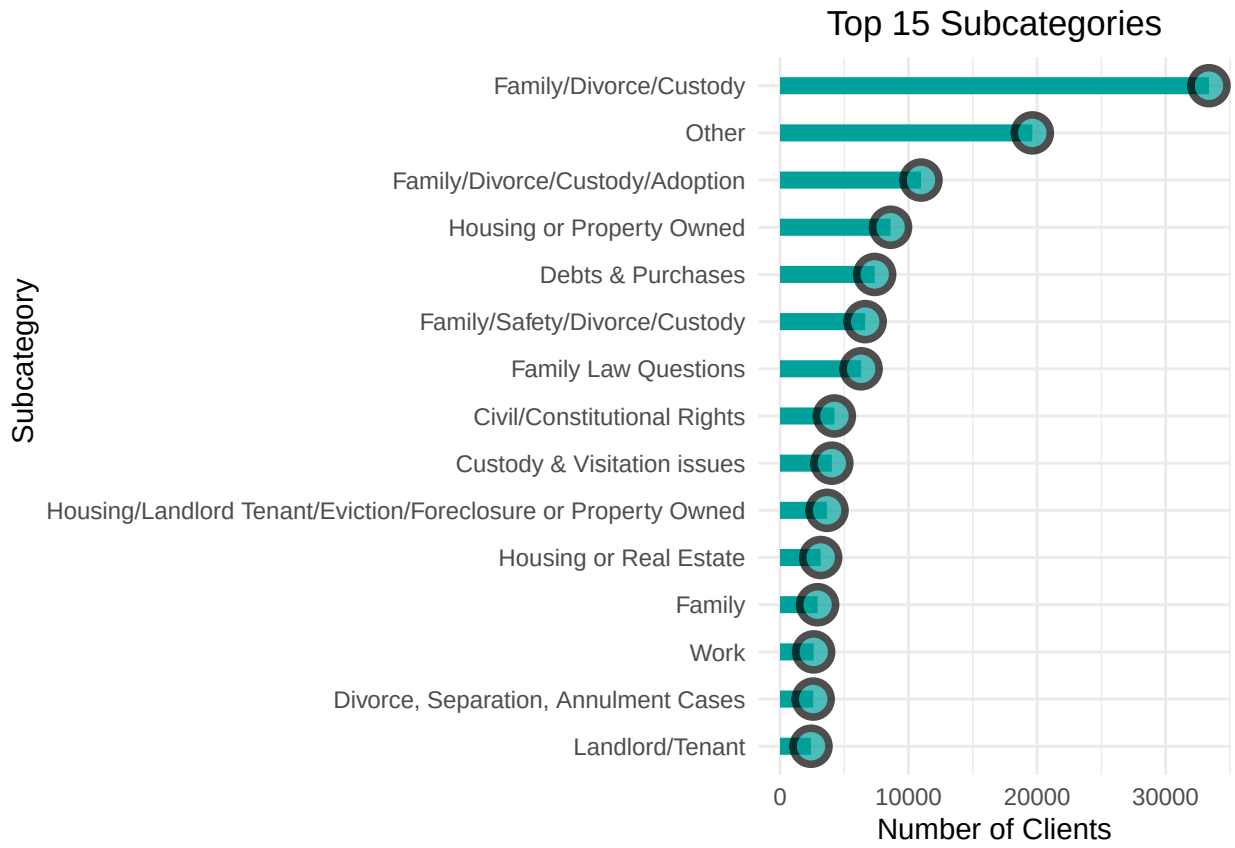
Project 1 - Analysis of American Bar Association data

Eric Chen, Junhan Li, & Daniel Fredin

Visualization 1: Investigating the Top 15 Subcategories of Asked Questions



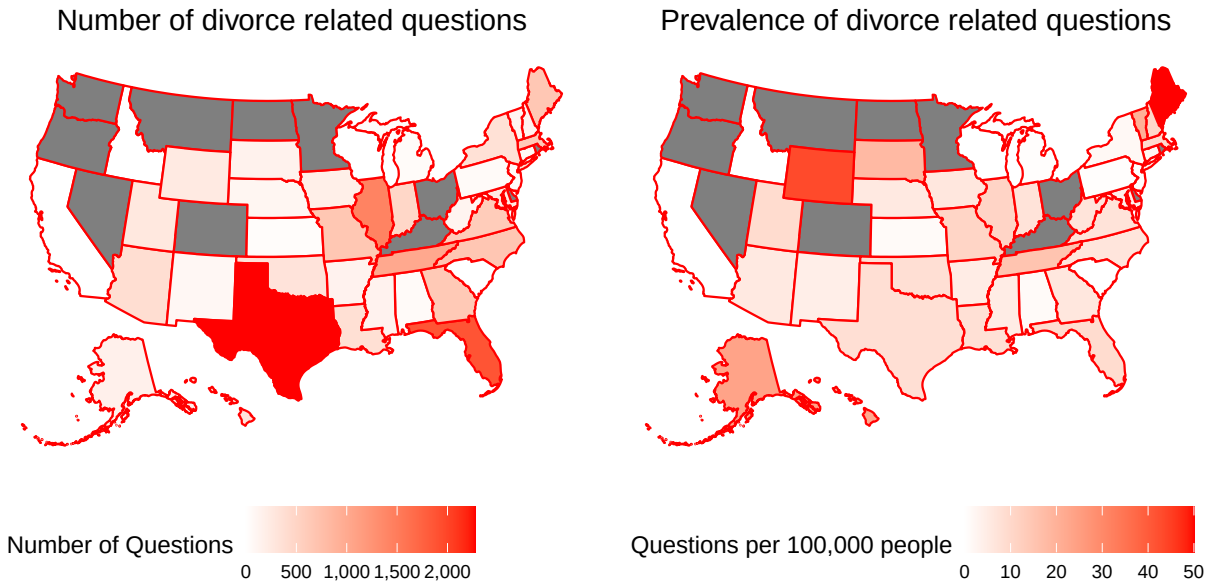
Interpretation of visualization 1:

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



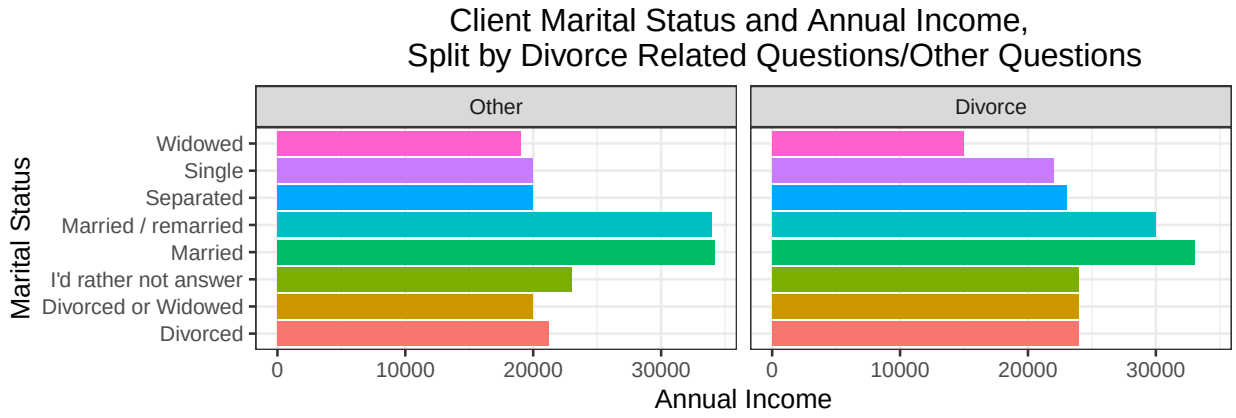
Interpretation of visualization 2:

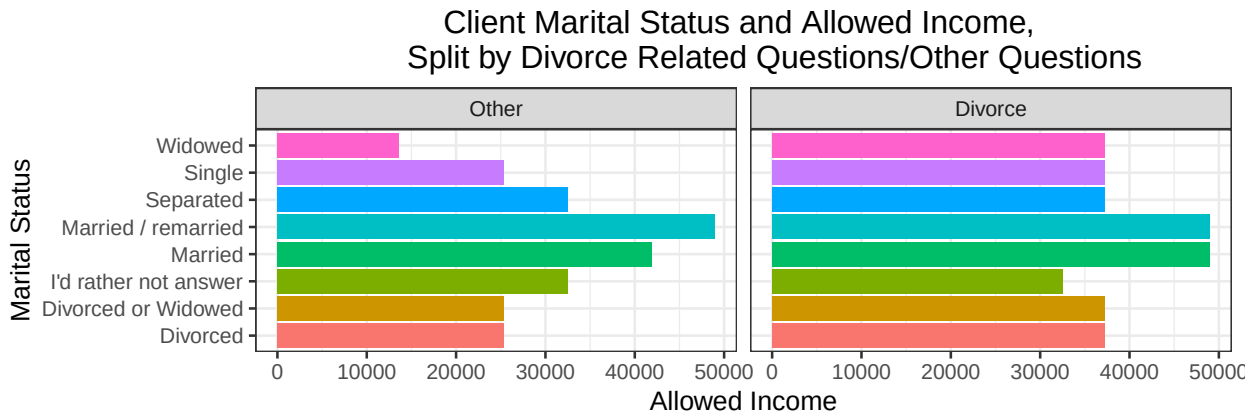
By analyzing the quantity of divorce-related inquiries and their distribution across states, we can gain insights into the clients' backgrounds and identify the regions with the highest occurrence of divorce-related questions.

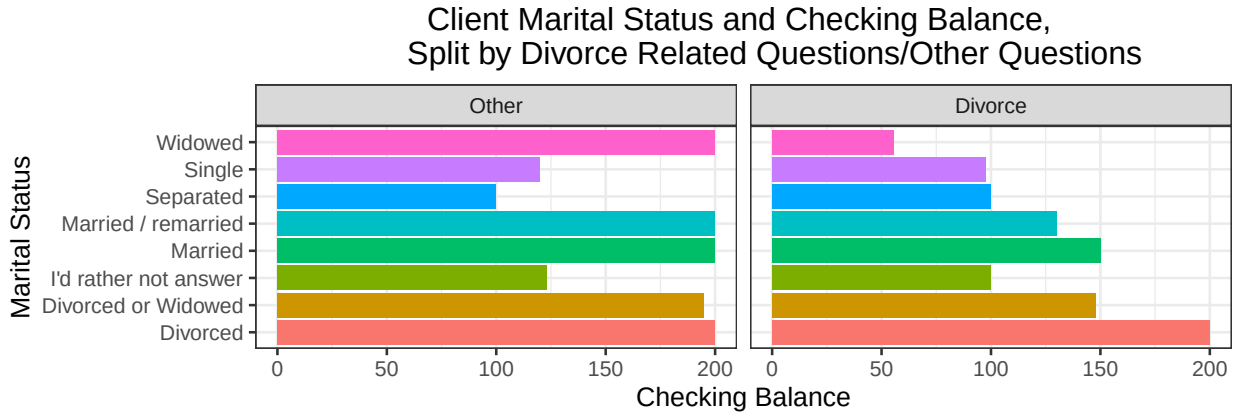
States like Washington, Oregon, Nevada, Montana, Colorado, North Dakota, Minnesota, Ohio, and Kentucky are depicted in gray on the chart due to legal requirements and confidentiality obligations. These states are prohibited from disclosing clients' information, including the specific category of legal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

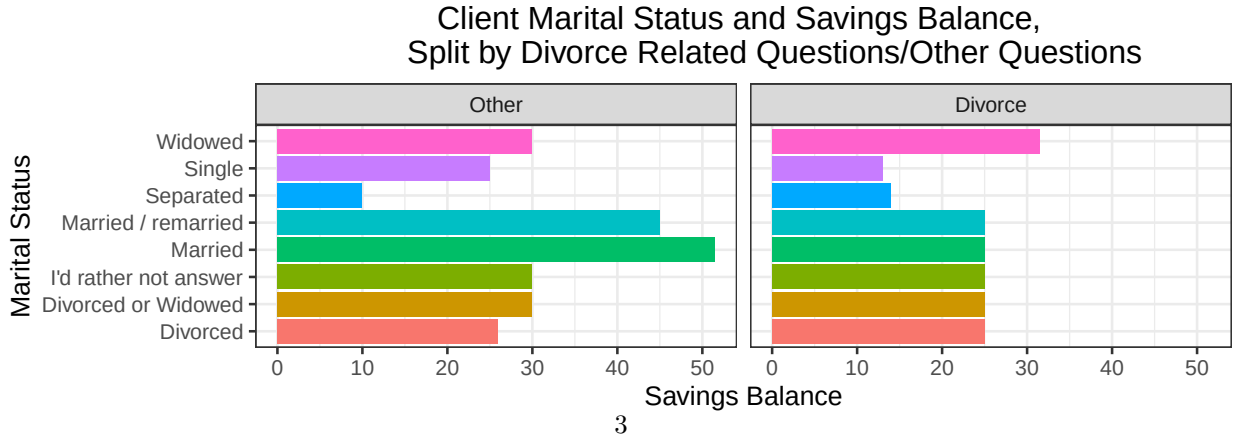
Among the 42 states that permit the revealing of client details, Texas and Florida demonstrate the greatest volume of inquiries concerning divorce law. Nonetheless, the prevalence of divorce-related questions takes on a distinct pattern when analyzed differently. When considering the number of queries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









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: Can we derive significant predictive ability by assessing clients' background, including their financial circumstances, relationship status, and state of residence, 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. Meanwhile, the independent or predictor variables encompass the client's age, marital status, state of residence, household size, annual income, as well as checking and savings balances.

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))
## StateAbbr           2.420537 39        1.011398
## Age                 1.311363  1        1.145148
## NumberInHousehold  27.909688  1        5.282962
## MaritalStatus       1.530729  7        1.030877
## AnnualIncome        1.166979  1        1.080268
## SavingsBalance      1.093005  1        1.045469
## CheckingBalance     1.094893  1        1.046372
## AllowedIncome      28.751815  1        5.362072
```

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))
## StateAbbr           1.088084 39        1.001083
## Age                 1.311192  1        1.145073
## NumberInHousehold   1.236508  1        1.111984
## MaritalStatus       1.530012  7        1.030843
## AnnualIncome        1.166916  1        1.080239
## SavingsBalance      1.092866  1        1.045402
## CheckingBalance     1.094757  1        1.046307
```

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 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.

```
## $Models
##   Formula
## 1 "Subcategory ~ (StateAbbr + Age + NumberInHousehold + MaritalStatus + AnnualIncome + SavingsBalanc
## 2 "Subcategory ~ MaritalStatus + SavingsBalance"
```

```
## 3 "Subcategory ~ SavingsBalance + CheckingBalance + MaritalStatus + NumberInHousehold"
## 4 "Subcategory ~ AllowedIncome"
## 5 "Subcategory ~ StateAbbr"

## $Fit.criteria
##   Rank Df.res   AIC  AICc   BIC McFadden Cox.and.Snell Nagelkerke   p.value
## 1   52  50570 49540 49540 50010 0.204400      0.221900    0.31380 0.000e+00
## 2    9  50610 58660 58660 58750 0.056150      0.066600    0.09421 0.000e+00
## 3   11  50610 58220 58220 58320 0.063350      0.074810    0.10580 0.000e+00
## 4    2  50620 61710 61710 61730 0.006917      0.008454    0.01196 9.225e-96
## 5   40  50580 55650 55650 56010 0.105600      0.121600    0.17200 0.000e+00
```

Summary of best model

With reference to the state of residence of Florida and marital status of Single, we can identify the positive predictor and negative predictor variables to be:

```
##                          (Intercept)                          StateAbbrAK
##                            0.02986586                           0.25024042
##                          StateAbbrAR                          StateAbbrAZ
##                            0.25393083                           0.13836205
##                          StateAbbrHI                          StateAbbrIA
##                            0.79314201                           0.03391055
##                          StateAbbrIL                          StateAbbrLA
##                            0.53154672                           0.43324428
##                          StateAbbrME                          StateAbbrMS
##                            0.72376774                           0.49783329
##                          StateAbbrNM                          StateAbbrSD
##                            0.18767264                           0.71144934
##                          StateAbbrTN                          StateAbbrTX
##                            0.01548571                           0.25953100
##                          StateAbbrVT                          StateAbbrWV
##                            0.12957102                           0.03588021
##                          StateAbbrWY                    NumberInHousehold
##                            0.55996379                           0.08183178
##              MaritalStatusDivorced   MaritalStatusDivorced or Widowed
##                            1.00152493                           1.19987276
## MaritalStatusI'd rather not answer               MaritalStatusMarried
##                            0.38659736                           0.76600319
##   MaritalStatusMarried / remarried             MaritalStatusSeparated
##                            0.94604110                           2.56085734

##          StateAbbrAL          StateAbbrCA          StateAbbrCT
##        -3.141950e-01        -1.791980e+01        -1.233012e-01
##          StateAbbrGA          StateAbbrID          StateAbbrIN
##        -5.654169e-02        -1.598573e+01        -1.527054e+00
##          StateAbbrKS          StateAbbrMA          StateAbbrMD
##        -3.974465e-02        -2.566881e-01        -2.719432e-02
##          StateAbbrMI          StateAbbrMO          StateAbbrNC
##        -9.960714e-02        -5.070567e-02        -2.007188e-01
##          StateAbbrNE          StateAbbrNH          StateAbbrNJ
##        -1.571948e+00        -1.689645e-01        -5.939755e-01
##          StateAbbrNY          StateAbbrOK          StateAbbrPA
```

```
##        -5.237575e-01        -3.035060e-01        -2.878429e-01
##          StateAbbrSC          StateAbbrUS          StateAbbrUT
##        -1.828112e+01        -1.809097e+01        -1.569470e-02
##          StateAbbrVA          StateAbbrWI                  Age
##        -1.384314e-01        -1.807015e+01        -4.102995e-02
## MaritalStatusWidowed         AnnualIncome       SavingsBalance
##        -5.755300e-01        -1.213875e-07        -1.848884e-05
##      CheckingBalance
##        -3.932488e-05
```

The summary of our best model is displayed below.

```
##
## Call:
## glm(formula = Subcategory ~ . - AllowedIncome - Category, family = "binomial",
##     data = data_overall)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0922  -0.8226  -0.4700   0.9032   2.8305
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     2.987e-02  5.207e-02   0.574 0.566265
## StateAbbrAK                     2.502e-01  1.132e-01   2.211 0.027022 *
## StateAbbrAL                    -3.142e-01  1.576e-01  -1.993 0.046261 *
## StateAbbrAR                     2.539e-01  1.214e-01   2.092 0.036432 *
## StateAbbrAZ                     1.384e-01  7.336e-02   1.886 0.059299 .
## StateAbbrCA                    -1.792e+01  2.396e+02  -0.075 0.940379
## StateAbbrCT                    -1.233e-01  1.354e-01  -0.911 0.362487
## StateAbbrGA                    -5.654e-02  6.068e-02  -0.932 0.351437
## StateAbbrHI                     7.931e-01  9.090e-02   8.725  < 2e-16 ***
## StateAbbrIA                     3.391e-02  9.948e-02   0.341 0.733192
## StateAbbrID                    -1.599e+01  6.523e+03  -0.002 0.998045
## StateAbbrIL                     5.315e-01  4.922e-02  10.800  < 2e-16 ***
## StateAbbrIN                    -1.527e+00  6.087e-02 -25.087  < 2e-16 ***
## StateAbbrKS                    -3.974e-02  2.043e-01  -0.195 0.845765
## StateAbbrLA                     4.332e-01  8.170e-02   5.303 1.14e-07 ***
## StateAbbrMA                    -2.567e-01  5.774e-02  -4.445 8.77e-06 ***
## StateAbbrMD                    -2.719e-02  8.723e-02  -0.312 0.755218
## StateAbbrME                     7.238e-01  6.603e-02  10.961  < 2e-16 ***
## StateAbbrMI                    -9.961e-02  1.884e-01  -0.529 0.596925
## StateAbbrMO                    -5.071e-02  5.900e-02  -0.859 0.390072
## StateAbbrMS                     4.978e-01  1.300e-01   3.829 0.000128 ***
## StateAbbrNC                    -2.007e-01  5.912e-02  -3.395 0.000686 ***
## StateAbbrNE                    -1.572e+00  1.103e-01 -14.246  < 2e-16 ***
## StateAbbrNH                    -1.690e-01  1.101e-01  -1.534 0.125022
## StateAbbrNJ                    -5.940e-01  1.806e-01  -3.289 0.001005 **
## StateAbbrNM                     1.877e-01  1.410e-01   1.331 0.183125
## StateAbbrNY                    -5.238e-01  7.052e-02  -7.427 1.11e-13 ***
## StateAbbrOK                    -3.035e-01  7.828e-02  -3.877 0.000106 ***
## StateAbbrPA                    -2.878e-01  1.659e-01  -1.735 0.082713 .
## StateAbbrSC                    -1.828e+01  1.151e+02  -0.159 0.873843
## StateAbbrSD                     7.114e-01  1.288e-01   5.526 3.28e-08 ***
```

```
## StateAbbrTN                          1.549e-02  5.317e-02   0.291 0.770848
## StateAbbrTX                          2.595e-01  4.276e-02   6.070 1.28e-09 ***
## StateAbbrUS                         -1.809e+01  6.268e+02  -0.029 0.976973
## StateAbbrUT                         -1.569e-02  8.834e-02  -0.178 0.858995
## StateAbbrVA                         -1.384e-01  6.251e-02  -2.215 0.026781 *
## StateAbbrVT                          1.296e-01  1.207e-01   1.073 0.283199
## StateAbbrWI                         -1.807e+01  1.240e+02  -0.146 0.884132
## StateAbbrWV                          3.588e-02  1.210e-01   0.296 0.766888
## StateAbbrWY                          5.600e-01  1.010e-01   5.543 2.98e-08 ***
## Age                                 -4.103e-02  9.917e-04 -41.373  < 2e-16 ***
## NumberInHousehold                    8.183e-02  7.008e-03  11.678  < 2e-16 ***
## MaritalStatusDivorced                1.002e+00  8.047e-02  12.446  < 2e-16 ***
## MaritalStatusDivorced or Widowed     1.200e+00  3.470e-02  34.578  < 2e-16 ***
## MaritalStatusI'd rather not answer   3.866e-01  6.686e-02   5.782 7.37e-09 ***
## MaritalStatusMarried                 7.660e-01  6.683e-02  11.462  < 2e-16 ***
## MaritalStatusMarried / remarried     9.460e-01  3.064e-02  30.875  < 2e-16 ***
## MaritalStatusSeparated               2.561e+00  4.320e-02  59.279  < 2e-16 ***
## MaritalStatusWidowed                -5.755e-01  3.352e-01  -1.717 0.085949 .
## AnnualIncome                        -1.214e-07  5.323e-07  -0.228 0.819619
## SavingsBalance                      -1.849e-05  4.501e-06  -4.108 4.00e-05 ***
## CheckingBalance                     -3.932e-05  8.669e-06  -4.536 5.73e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 62131  on 50619  degrees of freedom
## Residual deviance: 49433  on 50568  degrees of freedom
## AIC: 49537
##
## Number of Fisher Scoring iterations: 17
```

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.

```
##                                          OR         2.5 %         97.5 %
## (Intercept)                    1.030316e+00  9.303528e-01  1.141021e+00
## StateAbbrAK                    1.284334e+00  1.028840e+00  1.603275e+00
## StateAbbrAL                    7.303766e-01  5.362360e-01  9.948044e-01
## StateAbbrAR                    1.289083e+00  1.016165e+00  1.635300e+00
## StateAbbrAZ                    1.148391e+00  9.945863e-01  1.325981e+00
## StateAbbrCA                    1.650170e-08 1.896240e-212 1.436032e+196
## StateAbbrCT                    8.839974e-01  6.779495e-01  1.152669e+00
## StateAbbrGA                    9.450271e-01  8.390611e-01  1.064376e+00
## StateAbbrHI                    2.210330e+00  1.849611e+00  2.641399e+00
## StateAbbrIA                    1.034492e+00  8.512373e-01  1.257198e+00
## StateAbbrID                    1.141528e-07  0.000000e+00           Inf
```

```
## StateAbbrIL                          1.701562e+00  1.545091e+00  1.873879e+00
## StateAbbrIN                          2.171746e-01  1.927503e-01  2.446938e-01
## StateAbbrKS                          9.610348e-01  6.439082e-01  1.434347e+00
## StateAbbrLA                          1.542253e+00  1.314041e+00  1.810099e+00
## StateAbbrMA                          7.736094e-01  6.908312e-01  8.663065e-01
## StateAbbrMD                          9.731721e-01  8.202423e-01  1.154615e+00
## StateAbbrME                          2.062188e+00  1.811850e+00  2.347116e+00
## StateAbbrMI                          9.051930e-01  6.257682e-01  1.309389e+00
## StateAbbrMO                          9.505584e-01  8.467632e-01  1.067077e+00
## StateAbbrMS                          1.645153e+00  1.275112e+00  2.122580e+00
## StateAbbrNC                          8.181425e-01  7.286266e-01  9.186559e-01
## StateAbbrNE                          2.076403e-01  1.672584e-01  2.577718e-01
## StateAbbrNH                          8.445389e-01  6.805576e-01  1.048032e+00
## StateAbbrNJ                          5.521280e-01  3.875405e-01  7.866153e-01
## StateAbbrNM                          1.206439e+00  9.151711e-01  1.590406e+00
## StateAbbrNY                          5.922908e-01  5.158334e-01  6.800808e-01
## StateAbbrOK                          7.382254e-01  6.332205e-01  8.606429e-01
## StateAbbrPA                          7.498794e-01  5.417347e-01  1.037997e+00
## StateAbbrSC                          1.149773e-08 1.140259e-106  1.159367e+90
## StateAbbrSD                          2.036941e+00  1.582655e+00  2.621626e+00
## StateAbbrTN                          1.015606e+00  9.151021e-01  1.127149e+00
## StateAbbrTX                          1.296322e+00  1.192112e+00  1.409642e+00
## StateAbbrUS                          1.390564e-08  0.000000e+00           Inf
## StateAbbrUT                          9.844278e-01  8.279138e-01  1.170530e+00
## StateAbbrVA                          8.707229e-01  7.703267e-01  9.842037e-01
## StateAbbrVT                          1.138340e+00  8.984621e-01  1.442262e+00
## StateAbbrWI                          1.419826e-08 4.050958e-114  4.976369e+97
## StateAbbrWV                          1.036532e+00  8.176329e-01  1.314035e+00
## StateAbbrWY                          1.750609e+00  1.436142e+00  2.133933e+00
## Age                                  9.598004e-01  9.579366e-01  9.616678e-01
## NumberInHousehold                    1.085273e+00  1.070469e+00  1.100282e+00
## MaritalStatusDivorced                2.722430e+00  2.325196e+00  3.187527e+00
## MaritalStatusDivorced or Widowed     3.319694e+00  3.101422e+00  3.553328e+00
## MaritalStatusI'd rather not answer   1.471964e+00  1.291181e+00  1.678058e+00
## MaritalStatusMarried                 2.151151e+00  1.887051e+00  2.452214e+00
## MaritalStatusMarried / remarried     2.575493e+00  2.425374e+00  2.734904e+00
## MaritalStatusSeparated               1.294691e+01  1.189582e+01  1.409088e+01
## MaritalStatusWidowed                 5.624067e-01  2.915808e-01  1.084781e+00
## AnnualIncome                         9.999999e-01  9.999988e-01  1.000001e+00
## SavingsBalance                       9.999815e-01  9.999727e-01  9.999903e-01
## CheckingBalance                      9.999607e-01  9.999437e-01  9.999777e-01
```

Accuracy of Best Model

Based on the classification table below, our optimal model accurately predicted 37,840 out of 50,620 total observations.

```
##
##     FALSE  TRUE
##   0 32076  3187
##   1  9593  5764
```

```
## [1] "The accuracy of our best model was: 74.75%."
```

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.

COMMENTS

WHAT WE NEED TO FINISH:

- Write the summary
  - Suggestion based on limitations , missing data = limitation, states didn't provide data, also some states didn't have Divorce as a subcategory (such as California and the states that are white in viz 2)

FROM PRINCE'S PROJECT EXPECTATIONS:

- Executive summary of your results, your main takeaways and what you learned completing this project.
- Findings consistent with literature (if any).
- Any limitations.
- Suggestions (usually based on limitations) and/or recommendations.